Probability and Mathematical Statistics in Data Science

Lecture 09: Section 4.1: Cumulative Distribution Functions

Probability Mass Function (PMF) - Review

- The probability model for a discrete random variable X, lists its possible values and their probabilities.

Value of X	x_1	x_2	……..	x_k
Probability	p_1	p_2	……..	p_k

- Every probability, p_i, is a number between 0 and 1.
  $p_1+p_2+\ldots+p_k=1$

- The probability distribution or probability mass function (pmf) of a discrete random variable is defined for every number x by $p(x) = P(X=x)$.

Example - Review

Ex. Flip three fair coins. (*Binomial*)

S = {HHH, HHT, HTH, HTT, THT, THH, TTH, TTT}. Let's define random variable X to be the number of heads in the experiment, i.e., X(HHH)=3, X(THT)=1, etc.

X

0 TTT

1 TTH THT HTT

2 THH HTH HHT

3 HHH

Value of X	0	1	2	3
Probability	0.125	0.375	0.375	0.125

One can calculate the probability of an event by adding the probabilities p_i of the particular values of x_i that make up the event. For example, if we want to know the probability of getting less than 2 heads, we can use

$$P(X<2) = P(X=0) + P(X=1) = 0.125 + 0.375 = 0.5$$

Note: $P(X\leq 2) = P(X=0) + P(X=1) + P(X=2) = 0.875$

Random variables and their Distributions

- X, f(x)=P(X=x)

- Consider X = number of H in 3 tosses, then X ~ Bin(3,1/2)

- We can also define a new function F, called the ***cumulative distribution function,*** that, for each real number x, tells us how much mass has been accumulated by the time X reaches x.

- $\boldsymbol{F(x) = P(X \le x) = \sum_{k \le x} \binom{3}{k} p^k (1-p)^{n-k}}$.

x	0	1	2	3
$f(x) = P(X = x)$	1/8	3/8	3/8	1/8
$F(x)$				

Example

The PMF for a random variable X is as follows:

$$p(x) = \begin{cases} .500 & x = 0 \\ .167 & x = 1 \\ .333 & x = 2 \\ 0 & \text{otherwise} \end{cases}$$

The probability that *X* is at most 1 is then

In this example, $X \le 1.5$ if and only if $X \le 1$, so

$$P(X \le 1.5) = P(X \le 1) = .667$$

Similarly,

$$P(X \le 0) = P(X = 0) = .5, \quad P(X \le .75) = .5$$

Example continued

$$p(x) = \begin{cases} .500 & x = 0 \\ .167 & x = 1 \\ .333 & x = 2 \\ 0 & \text{otherwise} \end{cases}$$

And in fact for any x satisfying $0 \le x < 1$, $P(X \le x) = .5$. The largest possible X value is 2, so

$$P(X \le 2) = 1, \quad P(X \le 3.7) = 1, \quad P(X \le 20.5) = 1$$

and so on. Notice that $P(X < 1) < P(X \le 1)$ since the latter includes the probability of the X value 1, whereas the former does not. More generally, when X is discrete and x is a possible value of the variable, $P(X < x) < P(X \le x)$.

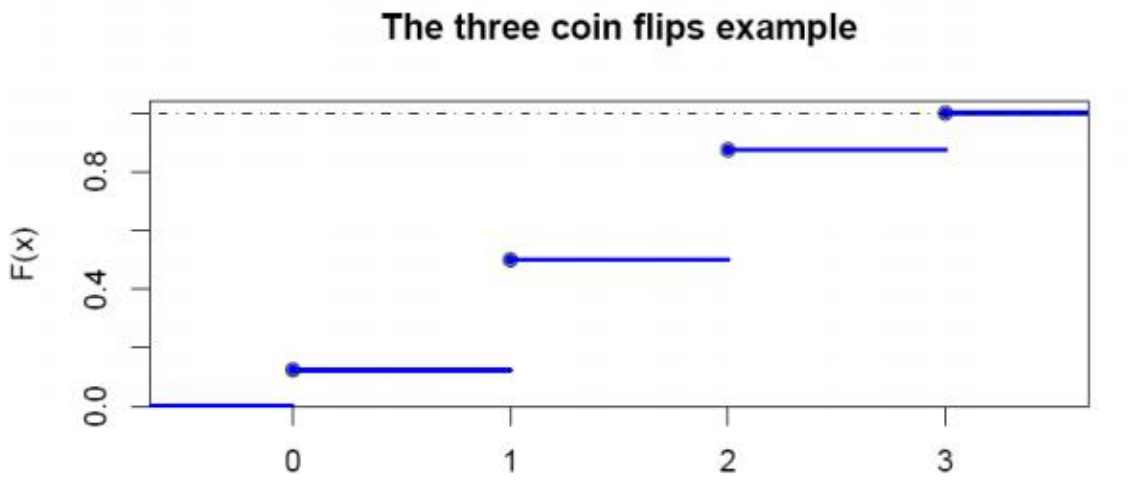
Cumulative Distribution Function (CDF)

- The cumulative distribution function (cdf) F(*x*) of a discrete rv variable X with pmf *p*(*x*) is defined for every number *x* by

$$\mathrm{F}(x) = \mathrm{P}(\mathrm{X} \le x) = \sum_{y:y \le x} p(y).$$

  For any number *x*, F(*x*) is the probability that the observed value of X will be at most *x*.

- For X a discrete rv, the graph of F(*x*) will have a jump at every possible value of X and will be flat between possible values. Such a graph is called a step function.



Exercise 4.5.2

- Cumulative distribution function (**cdf = F(x)**) of a random variable X is another way of describing the distribution of the probability.

- $F(x) = P(X \leq x)$

- $f(x) = P(X = x) = P(X \leq x) - P(X \leq x - 1) = F(x) - F(x - 1)$

w	-2	-1	0	1	3
$P(W = w)$	0.1	0.3	0.25	0.2	0.15

- A random variable W has the distribution shown in the table above. Sketch a graph of the cdf of W.

CDF and PMF

- A random variable W has the distribution shown in the table below. Sketch a graph of the pmf of W, and shade in F(1)

w	-2	-1	0	1	3
$P(W = w)$	0.1	0.3	0.25	0.2	0.15

CDFs are very useful because we often need sums of probabilities.

Example

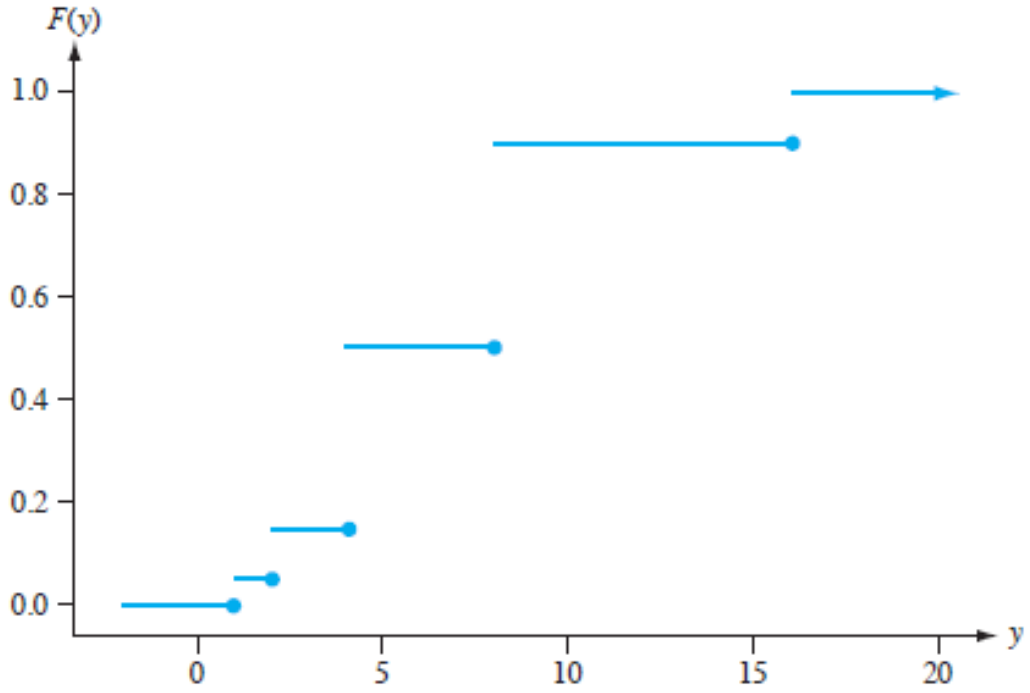
- A store carries flash drives with either 1 GB, 2 GB, 4 GB, 8 GB, or 16 GB of memory. The accompanying table gives the distribution of Y = the amount of memory in a purchased drive:

y	1	2	4	8	16
$p(y)$	.05	.10	.35	.40	.10

$$F(1) = P(Y \le 1) = P(Y = 1) = p(1) = .05$$
$$F(2) = P(Y \le 2) = P(Y = 1 \text{ or } 2) = p(1) + p(2) = .15$$
$$F(4) = P(Y \le 4) = P(Y = 1 \text{ or } 2 \text{ or } 4) = p(1) + p(2) + p(4) = .50$$
$$F(8) = P(Y \le 8) = p(1) + p(2) + p(4) + p(8) = .90$$
$$F(16) = P(Y \le 16) = 1$$

If y is less than 1, $F(y) = 0$ [e.g. $F(.58) = 0$], and if y is at least 16, $F(y) = 1$ [e.g. $F(25) = 1$]. The cdf is thus

$$F(y) = \begin{cases} 0 & y < 1 \\ .05 & 1 \le y < 2 \\ .15 & 2 \le y < 4 \\ .50 & 4 \le y < 8 \\ .90 & 8 \le y < 16 \\ 1 & 16 \le y \end{cases}$$



Example

- Many manufacturers have quality control programs that include inspection of incoming materials for defects. Suppose a computer manufacturer receives computer boards in lots of five. Two boards are selected from each lot for inspection. We can represent possible outcomes of the selection process by pairs. For example, the pair (1, 2) represents the selection of boards 1 and 2 for inspection.
- **a.** List the ten different possible outcomes.
- **b.** Suppose that boards 1 and 2 are the only defective boards in a lot of five. Two boards are to be chosen at random. Define X to be the number of defective boards observed among those inspected. Find the probability distribution of X.
- **c.** Let $F(x)$ denote the cdf of X. First determine F(0), $F(1)$, and $F(2)$; then obtain $F(x)$ for all other x.