Computer Science

CSC196: Analyzing Data

Discrete Random Variables

Jason Pacheco and Cesim Erten

Outline

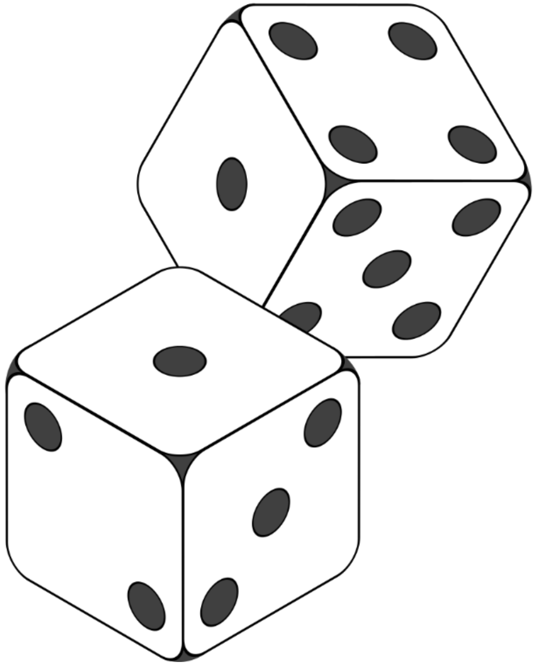
- Random Variables
- Discrete Probability Distributions
- Fundamental Rules of Probability

Outline

- Random Variables
- Discrete Probability Distributions
- Fundamental Rules of Probability

Random Events and Probability

Suppose we roll two fair dice…



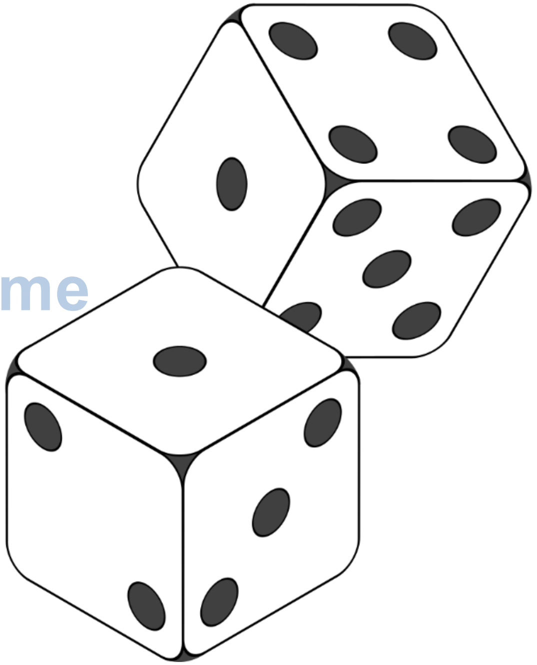
- What are the possible outcomes?
- What is the *probability* of rolling **even** numbers?
- What is the *probability* of rolling **odd** numbers?

…probability theory *gives a mathematical formalism to addressing such questions…*

Definition An **experiment** or **trial** is any process that can be repeated with well-defined outcomes. It is *random* if more than one outcome is possible.

Random Events and Probability

Definition An **outcome** is a possible result of an experiment or trial, and the collection of all possible outcomes is the **sample space** of the experiment,



Outcome

Example (1,1), (1,2), …, (6,1), (6,2), …, (6,6)

Sample Space

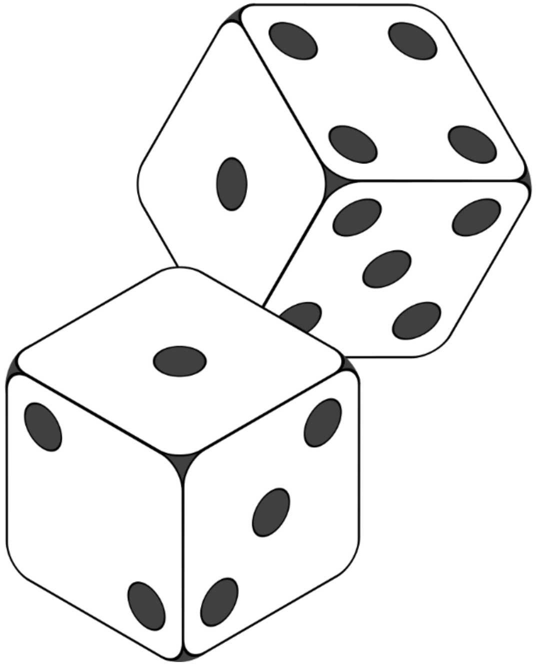
Definition An **event** is a *set* of outcomes (a subset of the sample space),

Example Event Roll at least a single 1

{(1,1), (1,2), (1,3), …, (1,6), …, (6,1)}

Random Variables

Suppose we are interested in a distribution over the sum of dice…



Option 1 Let E_i be event that the sum equals *i*

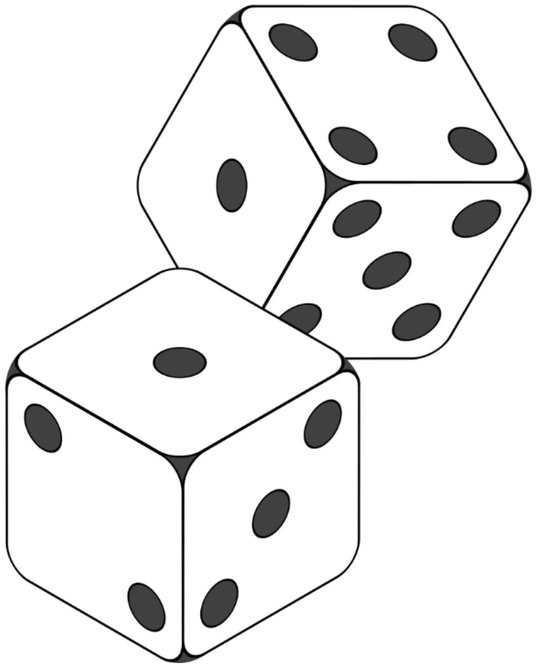
Two dice example:

$$E_2 = \{(1,1)\} \quad E_3 = \{(1,2),(2,1)\} \quad E_4 = \{(1,3),(2,2),(3,1)\}$$

$$E_5 = \{(1,4),(2,3),(3,2),(4,1)\} \quad E_6 = \{(1,5),(2,4),(3,3),(4,2),(5,1)\}$$

Enumerate all possible means of obtaining desired sum. Gets cumbersome for N>2 dice…

Random Variables



<u>Option 2</u> Use a function of sample space…

(Informally) A random variable is a function that maps events to numeric values.

Example X is the *sum of two dice* with values,

$$X \in \{2, 3, 4, \ldots, 12\}$$

Example Flip a coin and let random variable Y represent the outcome,

$$Y \in \{\text{Heads}, \text{Tails}\}$$

Example

Example 3.1: Two balls are drawn in succession without replacement from an urn containing 4 red balls and 3 black balls. The possible outcomes and the values y of the random variable Y, where Y is the number of red balls, are

Sample Space	y
RR	2
RB	1
BR	1
BB	0

Random Variables and Probability

Capitol letters represent random variables

Lowercase letters are realized *values*

$$X = x$$

$X = x$ is the **event** that X takes the value x

Example Let X be the random variable (RV) representing the sum of two dice with values,

$$X \in \{2, 3, 4, \ldots, 12\}$$

X=5 is the *event* that the dice sum to 5.

Example: Bernoulli Random Variables

Example 3.3: Consider the simple condition in which components are arriving from the production line and they are stipulated to be defective or not defective. Define the random variable X by

$$X = \begin{cases} 1, & \text{if the component is defective,} \\ 0, & \text{if the component is not defective.} \end{cases}$$

Clearly the assignment of 1 or 0 is arbitrary though quite convenient. This will become clear in later chapters. The random variable for which 0 and 1 are chosen to describe the two possible values is called a **Bernoulli random variable**.

Discrete vs. Continuous Probability

Discrete RVs take on a finite or countably infinite set of values

Continuous RVs take an uncountably infinite set of values

- Representing / interpreting / computing probabilities becomes more complicated in the continuous setting
- We will focus on discrete RVs for the moment…

Example

3.1 Classify the following random variables as discrete or continuous:

X: the number of automobile accidents per year in Virginia.

Y: the length of time to play 18 holes of golf.

M: the amount of milk produced yearly by a particular cow.

N: the number of eggs laid each month by a hen.

P: the number of building permits issued each month in a certain city.

Q: the weight of grain produced per acre.

Outline

- Random Variables
- Discrete Probability Distributions
- Fundamental Rules of Probability

Probability Mass Function

A function $P(X)$ is a **probability mass function (PMF)** of a discrete random variable X if the following conditions hold:

(a) It is nonnegative for all values in the support,

$$P(X = x) \geq 0$$

(b) The sum over all values in the support is 1,

$$\sum_x P(X = x) = 1$$

Intuition Probability mass is conserved, just as in physical mass. Reducing probability mass of one event must increase probability mass of other events so that the definition holds...

Probability Mass Function

Example Let X be the outcome of a single fair die. It has the PMF,

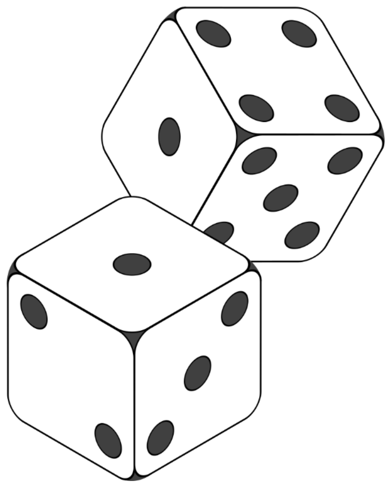
$$P(X = x) = \frac{1}{6} \qquad \text{for } x = 1, \ldots, 6$$

Uniform Distribution

Example We can often represent the PMF as a vector. Let S be an RV that is the *sum of two fair dice*. The PMF is then,

Observe that S does <u>not</u> follow a uniform distribution

$$P(S = s) = \begin{pmatrix} p(S=2) \\ p(S=3) \\ p(S=4) \\ \vdots \\ p(S=12) \end{pmatrix} = \begin{pmatrix} 1/36 \\ 1/18 \\ 1/2 \\ \vdots \\ 1/36 \end{pmatrix}$$



Example

Determine the value c so that each of the following functions can serve as a probability distribution of the discrete random variable X:

$$P(X = x) = c(x^2 + 4), \text{for } x = 0, 1, 2, 3;$$

Example

A shipment of 20 similar laptop computers to a retail outlet contains 3 that are defective. If a school makes a random purchase of 2 of these computers, find the probability distribution for the number of defectives.

Hints:

- How many ways of selecting 2 laptops from a total of 20?
- How many ways of selecting 0, 1, or 2 defective computers out of 3?
- How many ways of selecting 0, 1, or 2 non-defective computers out of the remaining 17?

Graphical Representation

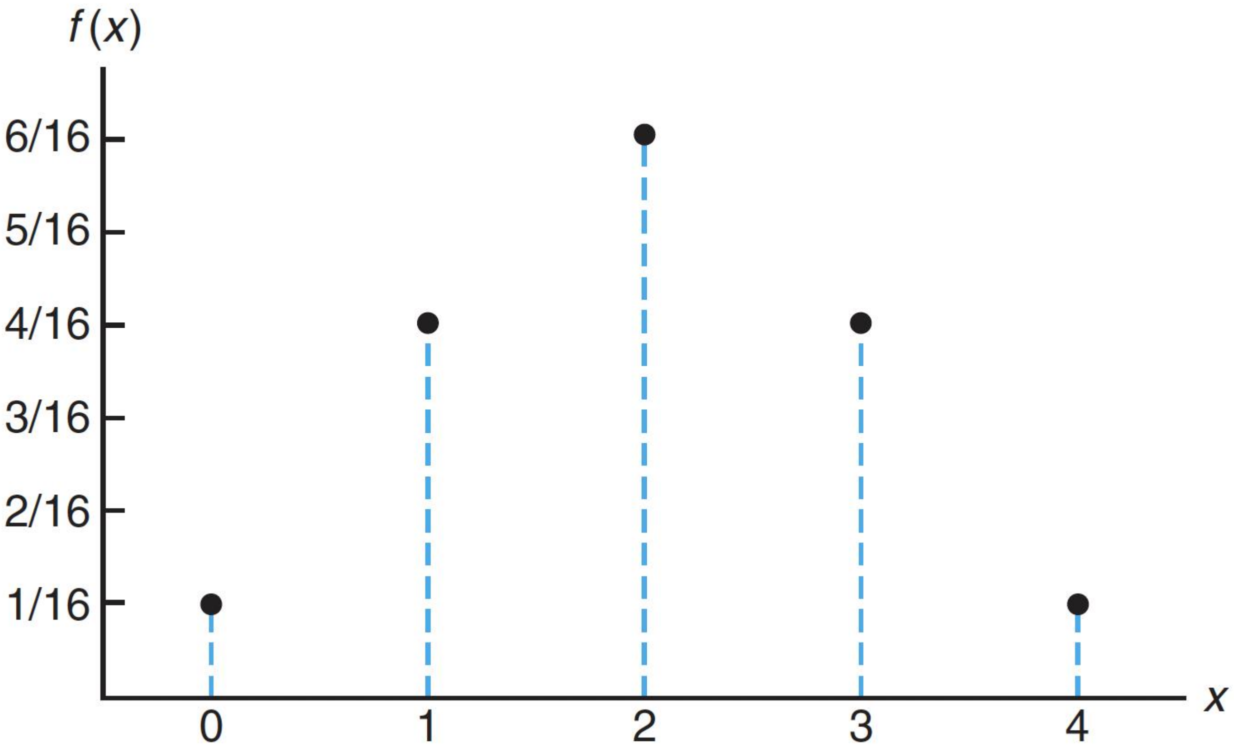


Figure 3.1: Probability mass function plot.

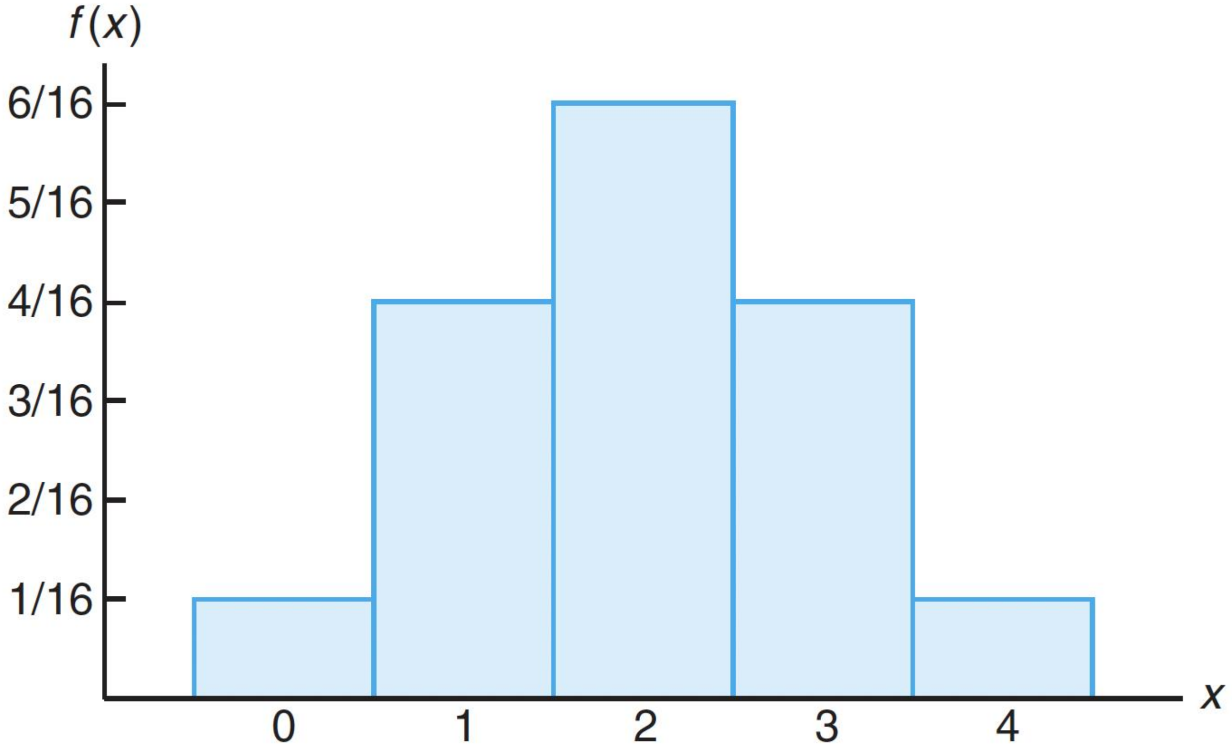


Figure 3.2: Probability histogram.

Cumulative Distribution Function

The **cumulative distribution function (CDF)** of a discrete random variable is defined as:

$$P(X \le x) = \sum_{t \le x} P(X = t)$$

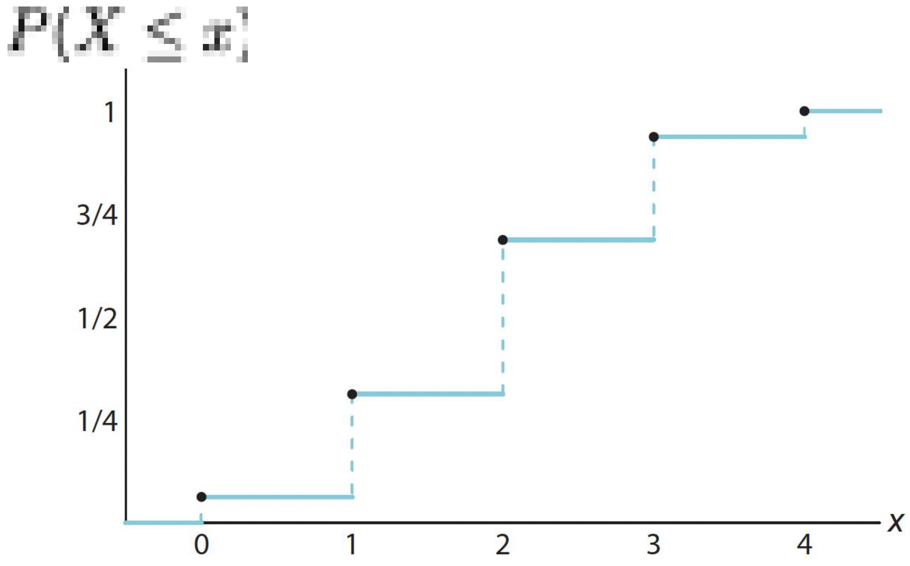


Figure 3.3: Discrete cumulative distribution function.

Joint Probability

Definition Two (discrete) RVs X and Y have a *joint PMF* denoted by $P(X, Y)$ and the probability of the event X=x and Y=y denoted by $P(X = x, Y = y)$ where,

(a) It is nonnegative for all values in the support,

$$P(X = x, Y = y) \geq 0$$

(b) The sum over all values in the support is 1,

$$\sum_x \sum_y P(X = x, Y = y) = 1$$

Joint Probability

Let X and Y be *binary RVs.* We can represent the joint PMF p(X,Y) as a 2x2 array (table):

		Y = 0	Y = 1
X	0	0.04	0.36
	1	0.30	0.30

All values are nonnegative

Joint Probability

Let X and Y be *binary RVs.* We can represent the joint PMF p(X,Y) as a 2x2 array (table):

		Y = 0	Y = 1
X	0	0.04	0.36
	1	0.30	0.30

The sum over all values is 1:
0.04 + 0.36 + 0.30 + 0.30 = 1

Joint Probability

Let X and Y be *binary RVs.* We can represent the joint PMF p(X,Y) as a 2x2 array (table):

		Y = 0	Y = 1
X	0	0.04	0.36
	1	0.30	0.30

P(X=1, Y=0) = 0.30

Example

Two ballpoint pens are selected at random from a box that contains 3 blue pens, 2 red pens, and 3 green pens. If X is the number of blue pens selected and Y is the number of red pens selected, find the joint probability function P(X,Y).

Hint:

- How many ways are there of selecting 2 pens out of the total 8?
- How many ways of selecting X blue pens from 3?
- How many ways of selecting Y red pens from 2?
- How many ways of selecting the remaining number of green pens from 3?

Example

Determine the values of c so that the following functions represent joint probability distributions of the random variables X and Y:

(a) $f(x, y) = cxy$, for $x = 1, 2, 3$; $y = 1, 2, 3$;

Example

From a sack of fruit containing 3 oranges, 2 apples, and 3 bananas, a random sample of 4 pieces of fruit is selected. If X is the number of oranges and Y is the number of apples in the sample, find

the joint probability P(X=1,Y=1).

Outline

- Random Variables
- Discrete Probability Distributions
- Fundamental Rules of Probability

Fundamental Rules of Probability

Law of total probability

$$P(Y) = \sum_x P(Y, X = x)$$

- P(y) is a **marginal** distribution
- This is called **marginalization**

Proof

$$\sum_x P(Y, X = x) = \sum_x P(Y) P(X = x \mid Y)$$ (chain rule)

$$= P(Y) \sum_x P(X = x \mid Y)$$ (distributive property)

$$= P(Y)$$ (PMF sums to 1)

Fundamental Rules of Probability

Given two RVs X and Y the **conditional distribution** is:

$$P(X \mid Y) = \frac{P(X,Y)}{P(Y)}$$

By the law of total probability, we also have the definition:

$$P(X \mid Y) = \frac{P(X,Y)}{\sum_x P(X=x,Y)}$$

Note *This definition of the conditional is largely consistent with what you have seen in terms of random events.*

Tabular Method

Let X, Y be binary RVs with the joint probability table

For Binomial use K-by-K probability table.

		Y: y_1	Y: y_2	$P(x)$
X	x_1	0.04	0.36	0.4 ($P(x_1)$)
X	x_2	0.30	0.30	0.6 ($P(x_2)$)
$P(y)$		0.34 ($P(y_1)$)	0.66 ($P(y_2)$)	

$P(y_1)=P(x_1,y_1)+P(x_2,y_1)$
$P(y_2)=P(x_1,y_2)+P(x_2,y_2)$
[i.e., sum down columns]

$P(x_1)=P(x_1,y_1)+P(x_1,y_2)$
$P(x_2)=P(x_2,y_1)+P(x_2,y_2)$
[i.e., sum across rows]

Tabular Method

We don’t care about event $Y=y_2$

		Y	
		y_1	y_2
X	x_1	0.04	Censored!
	x_2	0.30	
		0.34	

0.34 ← $P(y_1)$

$P(x|y_1)=?$

Tabular Method

X		$Y=y_1$
	x_1	0.04
	x_2	0.30
		0.34

0.34 ← $P(y_1)$

⇒

$P(x \mid y_1)$
0.04 / 0.34
0.30 / 0.34

These sum to one:
A conditional probability distribution is still a probability distribution

Independence of RVs

Definition *Two random variables* X *and* Y *are* *independent* *if and only if,*

$$P(X = x, Y = y) = P(X = x)P(Y = y)$$

- This must hold for all values x and y.
- If for any values x and y,

$$P(X = x, Y = y) \neq P(X = x)P(Y = y)$$

  then X and Y are **dependent.**
- Example: Rolling two dice, each die is independent of the other
- Independence is *symmetric:* if X is independent of Y then Y is independent of X
- Equivalent definition of independence: $P(X \mid Y) = P(X)$

Independence of RVs

Definition RVs $X_1, X_2, \ldots, X_N$ are mutually independent if and only if,

$$P(X_1 = x_1, \ldots, X_N = x_N) = \prod_{i=1}^{N} P(X_i = x_i)$$

In words: If a set of random variables is independent, then their joint probability is a product of their marginals.

Independence of RVs

Definition *Two random variables* X *and* Y *are* *<u>conditionally independent</u>* *given* Z *if and only if,*

$$P(X = x, Y = y \mid Z = z) = P(X = x \mid Z = z)P(Y = y \mid Z = z)$$

for all values x, y, and z.

- N RVs conditionally independent, given Z, if and only if:

$$P(X_1 = x_1, \ldots, X_N = X_N \mid Z = z) = \prod_{i=1}^{N} P(X_i = x_i \mid Z = z)$$

- Equivalent def'n of conditional independence: $P(X \mid Y, Z) = P(X \mid Z)$
- Conditional independence is symmetric