Curtin College

DIPLOMA OF INFORMATION TECHNOLOGY

IPDA1005 INTRODUCTION TO PROBABILITY AND DATA ANALYSIS

Acknowledgement

We respectfully acknowledge the Elders and custodians of the Whadjuk Nyungar nation, past and present, their descendants and kin. Curtin College Bentley Campus enjoys the privilege of being located in Whadjuk / Nyungar Boodjar (country) on the site where the Derbal Yerrigan (Swan River) and the Djarlgarra (Canning River) meet. The area is of great cultural significance and sustains the life and well being of the traditional custodians past and present.

Table of Contents

Axioms of Probability

- The naıve definition of probability is limited. It assumes equally likely outcomes, and it can't handle infinite sample spaces.
- The axioms of probability were set out by Andrey Kolmogorov in 1933.

A **probability** is a function *P* defined on a set of subsets of the sample space S satisfying the following axioms:

1. For any event A, $P(A) \geq 0$.
2. $P(S) = 1$
3. If $A_1, A_2, A_3, \ldots$ is an infinite collection of mutually disjoint events (that is, $A_i \cap A_j = \emptyset$ for all $i \neq j$), then

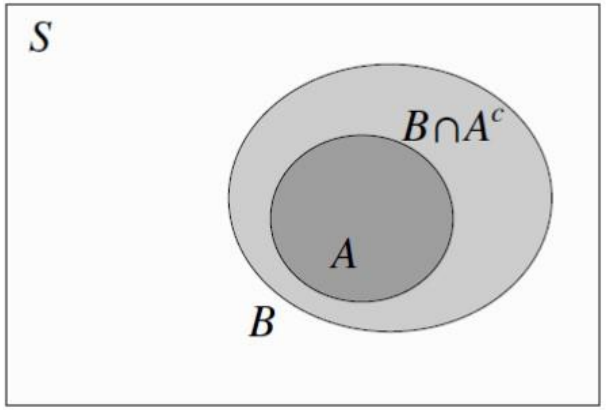
$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i)$$

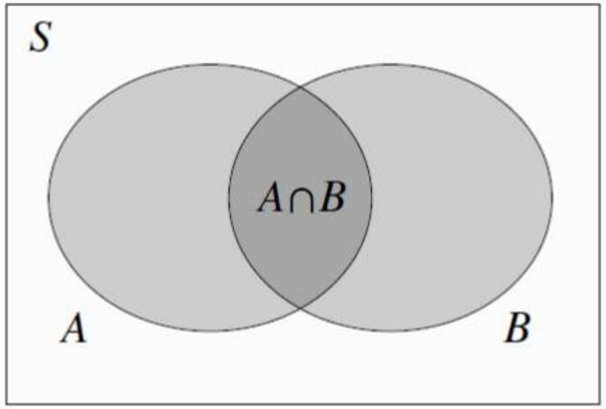
Derived Properties of Probability

- $P(A^c) = 1 - P(A)$
- $P(\emptyset) = 0.$
- For any event A, $P(A) \leq 1$
- If $A \subseteq B$, then $P(A) \leq P(B)$.
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

See DBC3 for proofs.

The last two results are illustrated in these Venn diagrams





An Example

In a residential suburb, cable company TVK provides internet service to 60% of households, television service to 80% of households, and both services to 50% of households. For a randomly selected household:

1. What is the probability that it gets at least one of these services from TVK?
2. What is the probability that it gets exactly one of these services from TVK?
    - Let $A = \{Internet\}$, and $B = \{TV\}$
    - $P(A) = 0.6$, $P(B) = 0.8$ and $P(A \cap B) = 0.5$.

1 $P(A \cup B) = 0.6 + 0.8 - 0.5 = 0.9$

2 $P(A \cap B^c) + P(A^c \cap B) = 0.1 + 0.3 = 0.4$

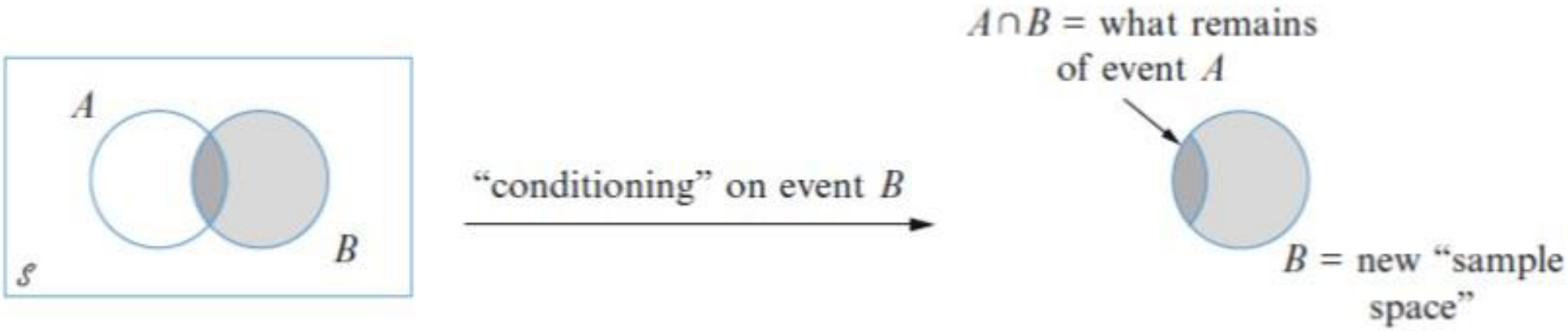
A visual or structural approach may help: Venn diagram, tree diagram, or contingency table

Conditional Probability

- In the previous example, the two events {Internet, TV} intersected. That is, $P(Internet \cap TV) > 0.$ We now focus on the interaction between events.
- The probability of an event A might change if we know that another event B has occurred.
- Let D represent the event that a randomly chosen person has a particular disease, and let T represent the event that a diagnostic test gives a positive result.
- Diagnostic tests are not perfect, so $P(test\ result\ is\ wrong) > 0$
- P(D), the chance that a randomly chosen person has the disease, will differ from P(D|T), the chance of disease in a person who had a positive test result. At least, they certainly should differ!
- Note the conditional probability notation, in which the event given to have occurred comes after the vertical bar.

An Example

- What is the probability of randomly selecting an ace from a standard deck of cards? Denote this event by A.
- What is the probability of randomly selecting a spade from a standard deck of cards? Denote this event by B.
- What is the probability of randomly selecting the ace of spades from a standard deck of cards? This is $P(A \cap B)$.
- If you are told that the selected card is a spade, what now is the probability that it is an ace of spades? This is $P(A|B)$.
- The original sample space has now shrunk!



Definition of Conditional Probability

- For any two events A and B with P(B) > 0, the conditional probability
- of A given B is defined as:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

- Conditional probability addresses the fundamental question: how should we update our beliefs in the light of evidence that we have observed?
- Note that $P(A|B)$ is not the same as $P(B|A)$ unless $P(A) = P(B)$.

Conditional Probability

- In card selection example, $P(Ace|Spade) = P(Ace)$, i.e., knowing that a card is a spade doesn't change the chance of it being an ace.
- The two events are then said to be independent. More on this later.

Exercise:

- A family has two children. Assume only males and females, and that the possible birth sequences are equally likely. What is the probability that both children are girls?
- How does your answer change if we know that -
  - one of the children is a girl?
  - the elder child is a girl?

Conditional Probability

- We can rearrange the expression for conditional probability to obtain
- the multiplication rule:

$$P(A \cap B) = P(A|B)P(B)$$
$$P(A \cap B) = P(B|A)P(A)$$

Example:

Four individuals have responded to a request by a blood bank for blood donations. None of them has donated before, so their blood

types are unknown. Suppose only type O+ is desired and only one of the

four actually has this type. If the potential donors are selected in random

order for typing, what is the probability that at least three individuals must

be tested to obtain the desired type?

The Blood Bank Example

- $\{at\ least\ 3\ must\ be\ tested\} \equiv \{first\ 2\ tested\ are\ not\ O+\}$.
- Let A_i denote the event that the ith individual tested is not O+.
- Then $P(A_1) = \frac{3}{4}$ and $P(A_2|A_1) = \frac{2}{3}$
- Then the probability that at least three are tested is

$$P(A1\ \cap\ A2) = P(A_1)P(A_2|A_1) = \left(\frac{3}{4}\right)\left(\frac{2}{3}\right)$$

- The multiplication rule for three events can be written as

$$P\ (A1\ \cap\ A2\ \cap\ A3) = P(A_1)P(A_2|A_1)P(A_3|A_2 \cap A_1)$$

This can be extended (laboriously) to any number of events.

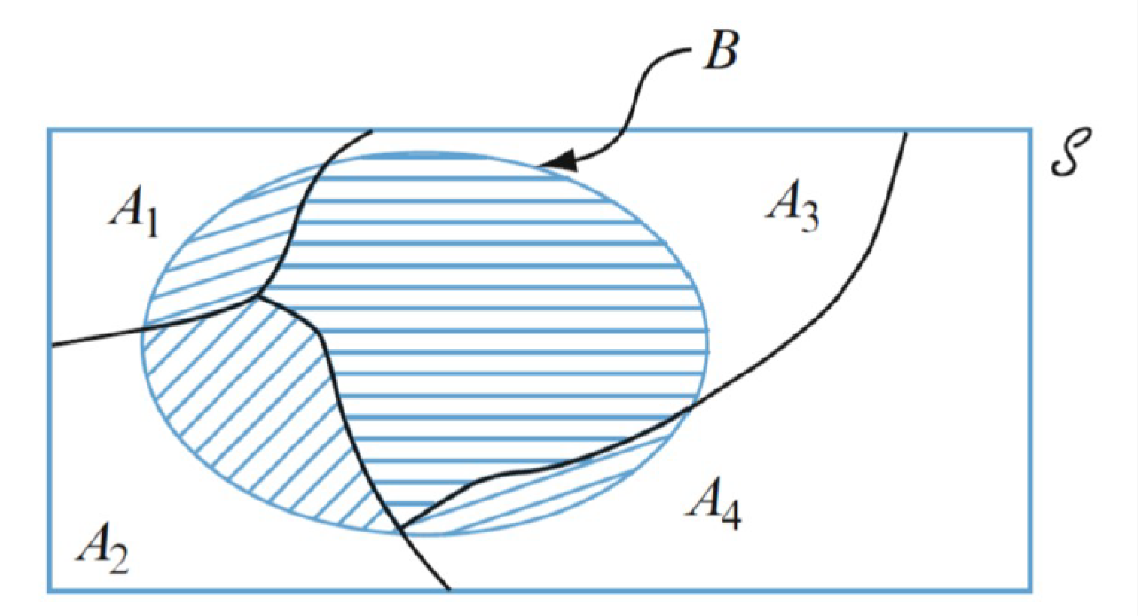
Exercise:

- Find the probability that exactly three individuals must be typed.

The Law of Total Probability

Let $A_1, A_2, \ldots A_k$ be a partition of the sample space. Then

$$\begin{aligned} P(B) &= P(B \cap A_1) + P(B \cap A_2) + \cdots + P(B \cap A_k) \\ &= P(A_1)P(B|A_1) + \cdots + P(A_k)P(B|A_k) \\ &= \sum_{i=1}^{k} P(A_i)P(B|A_i) \end{aligned}$$



An Example

A message packet sent from a computer to another takes one of the three internet routes, R_1, R_2 and R_3. The routing probabilities for R_1, R_2 and R_3 are 0.2, 0.1 and 0.7, respectively. 3% percent of the message packets delivered by route R_1 has some errors. The error rates for R_2 and R_3 are 5% and 0.8%, respectively.

- Find the probability that a message packet has an error. $P(E) = P(R_1)P(E|R_1) + P(R_2)P(E|R_2) + P(R_3)P(E|R_3) = (0.2)(0.03) + (0.1)(0.05) + (0.7)(0.008) = 0.0166$
- Suppose the received message packet has an error. What is the probability that it was delivered through R_2?

$$P(R_2|E) = \frac{P(R_2 \cap E)}{P(E)} = \frac{P(R_2)P(E|R_2)}{P(E)}$$

$$= \frac{(.1)(.05)}{.0166} = .3012$$

Bayes' Theorem

If we generalise the working we just used, we arrive at Bayes' Theorem:

Let $A_1, A_2, \dots A_k$ be a partition of the sample space and let B be an even such that $P(B) > 0$. Then

$$P(A_j|B) = \frac{P(A_j \cap B)}{P(B)} = \frac{P(A_j)P(B|A_j)}{\sum_{i=1}^{k} P(A_i)P(B|A_i)}, \quad j = 1, 2 \dots k$$

In Bayes' Theorem:

- We use the Law of Total Probability (in the denominator).
- The left hand side is conditioned on event B.
- The right hand side is conditioned on the set of events A_i.

A classic application of Bayes' Theorem is in the common experience of diagnostic testing ...

An Example

- **Solution:** Applying Bayes' theorem, with C = cancer and P = positive test result.

$$P(C|P) = \frac{P(C)P(P|C)}{P(C)P(P|C) + P(C^c)P(P|C^c)}$$
$$= \frac{(.0001)(.9)}{(.0001)(.9) + (.9999)(.001)} = 0.083$$

- Even after testing positive, Michelle's chance of cancer is only 8.3%.
- If the test was applied to many people, only 8.3% of those getting a positive test result would actually have cancer. This is mainly due to the very low prevalence of the cancer in the population for this example.
- Therefore screening programs should not target diseases with low prevalence, especially if the diagnostic procedure poses some risk. For example, the use of chest X-rays to screen for tuberculosis is no longer advocated.

Dependence of $P(C|P)$ on Disease Prevalence and Specificity

Definitions:

- Sensitivity $=$ chance of positive test result when diseased $= P(P|C)$
- Specificity $=$ chance of negative test result when healthy $= P(\bar{P}|\bar{C})$
- Prevalence $=$ proportion of population that have the disease $= P(C)$

We can use $P(C|P)$ as a measure of the reliability of the diagnostic test. The table shows $P(C|P)$ assuming sensitivity is held constant at 0.9.

	Prevalence		
Specificity	.01	.001	.0001
.9	.083	.009	.001
.99	.476	.083	.009
.999	.901	.474	.083

Independence

- Conditional probability means that we revise $P(A)$ in light of some other event B that has already occurred.
- If $P(A|B)$ differs from $P(A)$, then B's occurrence changed the chance of A occurring.
- But if $P(A|B) = P(A)$, then B's occurrence has no influence on the chance that A occurs. In this case we say that A and B are independent; the occurrence or non-occurrence of one event has no effecty on the chance that the other will occur.

Two events A and B are said to be *independent* if and only if

$$\begin{aligned} P(A|B) &= P(A) \\ P(B|A) &= P(B), \text{which implies} \\ P(A \cap B) &= P(A)P(B) \end{aligned}$$

Independence

- If A and B are independent,

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A)P(B)}{P(B)} = P(A).$$

- In fact,

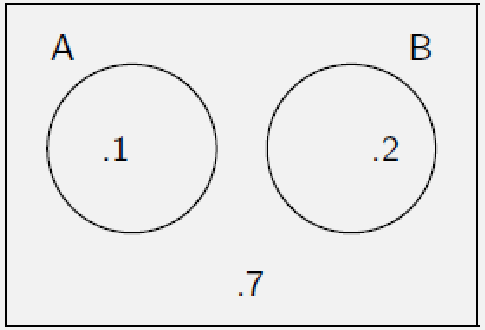
$$\begin{aligned} A \text{ and } B \text{ are independent } &\Leftrightarrow P(A \cap B) = P(A)P(B) \\ &\Leftrightarrow P(A|B) = P(A) \\ &\Leftrightarrow P(B|A) = P(B) \end{aligned}$$

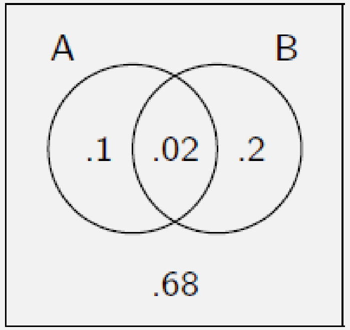
- In the example of randomly selecting a card from a 52-card deck, the event A of getting an ace and the event B of getting a spade are independent.

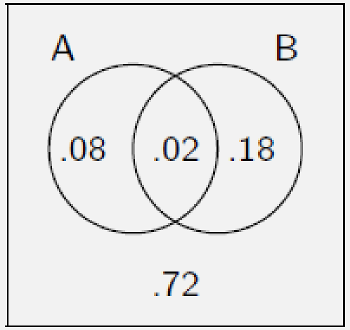
  Here $P(A \cap B) = \frac{1}{52}$ and $P(A)P(B) = \left(\frac{4}{52}\right)\left(\frac{13}{52}\right) = \frac{1}{52}$

Independence

- If A and B are independent, so are A and B^c, A^c and B, and A^c and B^c.
- Decide whether A and B are independent in the following cases:







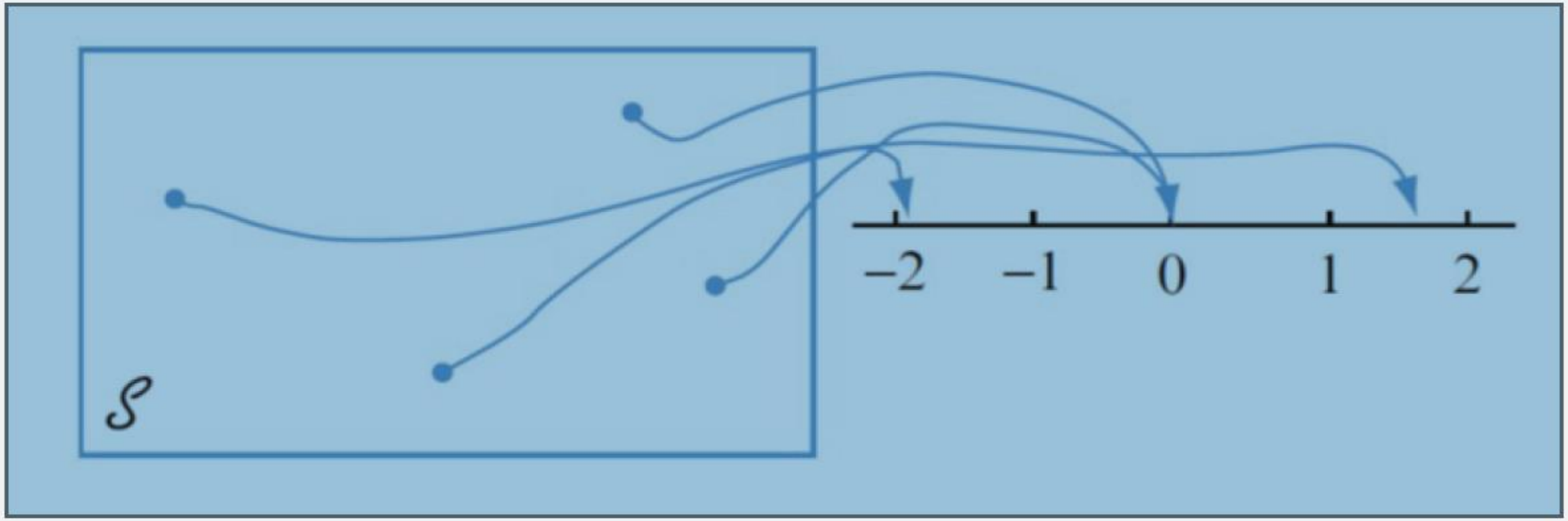
- In each case, also check if A^c and B^c are independent.

Random Variables

- A *random variable* is a function from the sample space S to $\mathbb{R}$, the set of real numbers.

  That is, a random variable is a numerical outcome from a random experiment.
- **Example:** A coin is tossed twice. Let X represent the number of heads obtained. Then X is a random variable.
  - Here X is a numerical outcome of the experiment.
  - The sample space is $\{HH, HT, TH, TT\}$ and X can be thought of as a real-valued function defined on S with
    $X(HH) = 2, \quad X(HT) = 1, \quad X(TH) = 1, \quad X(TT) = 0$
- Observing whether H or T is the outcome of a coin toss is not an example of a random variable, as the outcomes are not numerical.

Random Variables



- **Example:** A pair of dice is rolled. Let X be the sum and Y be the absolute difference of the numbers that come up.
  - X and Y are random variables defined on the same sample space.
  - $X + Y$ and $X - 2Y$ are also random variables.

Discrete and Continuous Random Variables

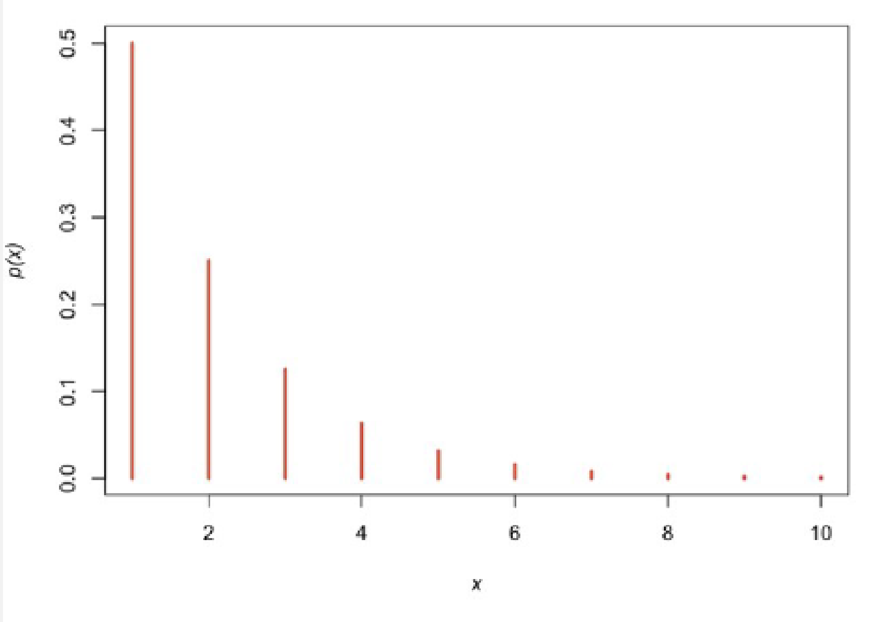
- A *discrete* random variable is one whose number of possible values is countable. This means the range of a discrete random variable is finite or countably infinite.
- Winning margin in a basket ball game, the number of heads when a coin is tossed 100 times, the number of rolls needed to get a double-six if a pair of dice is rolled repeatedly, the number of accidents on a highway, etc. are examples of discrete random variables.
- The values a discrete random variable takes don't have to be integers or non-negative.
- A *continuous* random variable is one that can take on uncountably infinitely many values, or a continuous spectrum of real numbers.
- The amount of rainfall, height of people, weight of newborns, temperature, etc. are examples of continuous random variables.

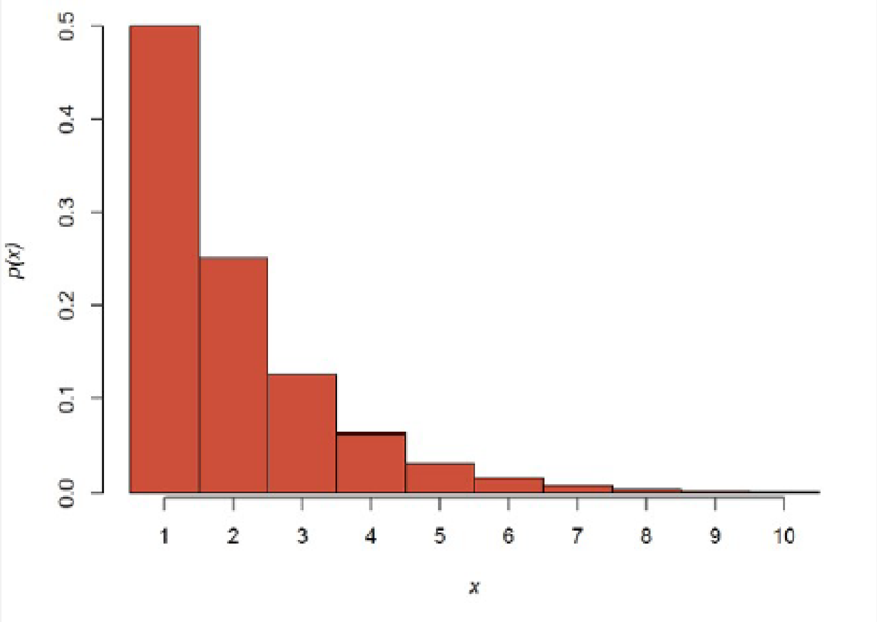
Discrete Probability Distributions

- The probability distribution of a discrete random variable X lists the possible values it takes along with the corresponding probabilities.
- This can be in the form of a table or a formula. If there is a formula, there is often a standard compact notation.
- The probability distribution of X determines how the total probability of 1 is distributed among the possible values of X.
- It is often preferable to express the probabilities associated with the values of a random variable by means of a function p such that $p(x) = P(X = x)$ for each x in the range of X.
- The function p is called the *probability mass function (PMF)* or the *probability function* of X and satisfies the following conditions:
    1. $p(x) \geq 0$ for all x.
    2. $\sum_x p(x) = 1$.

Graphing PMF

- A fair coin is tossed until a head comes up for the first time. Let X be the number of tosses needed.
- This is an infinite-valued random variable taking values $1, 2, 3 \ldots$, with $p(x) = P(X = x) = \left(\frac{1}{2}\right)^x$.
- We can use a vertical line chart or a histogram to depict a PMF graphically.





- **Exercise:** Sketch the probability mass function of the sum T of the numbers that come up when two dice are tossed.

Cumulative Distribution Function

- The *cumulative distribution function (CDF)* of a random variable X is given by

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

- **Example:** The distribution of the number X of tropical cyclones that will make landfall this season is given below:

x	0	1	2	3	4	5
$p(x)$	.7334	.2274	.0352	.0036	.0003	.0001

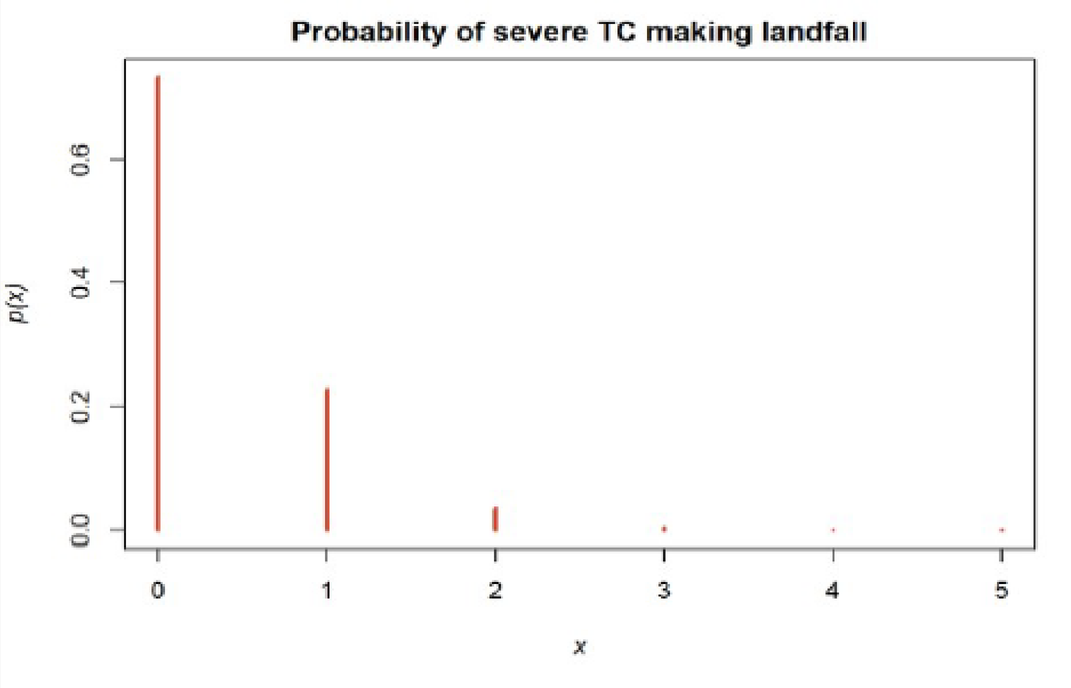
Then the CDF of X is given as follows:

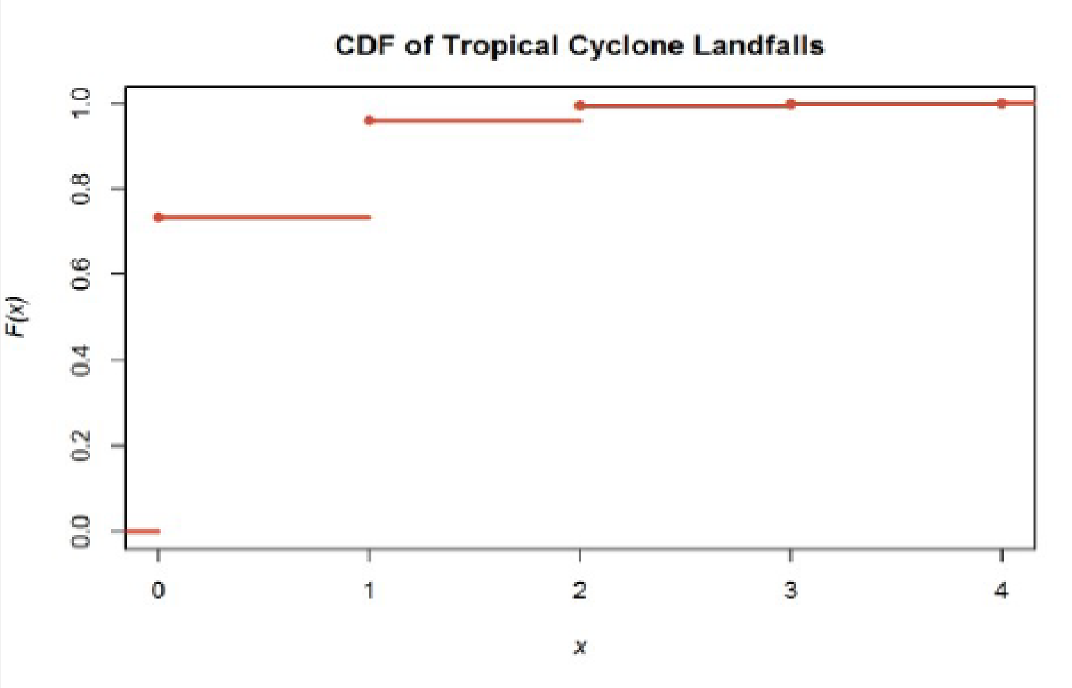
x	0	1	2	3	4	5
$F(x)$	.7334	.9608	.996	.9996	.9999	1

For instance, the probability of two or fewer cyclones is given by $P(X \leq 2) = F(2) = .996$.

Cumulative Distribution Function

- For a discrete random variable X, the graph of the CDF is a step function: it jumps at every possible value of X and is flat between possible values.





- We can calculate individual probabilities and probabilities of intervals from the CDF.
- $P(X = 2) = F(2) - F(1) = .996 - .9608 = .0352.$

Expectation of Random Variables

Let X be a discrete random variable with PMF $p(x)$. The *expectation* (or the expected value or the mean) of X, denoted by $E(X) = \mu_X = \mu$, is given by

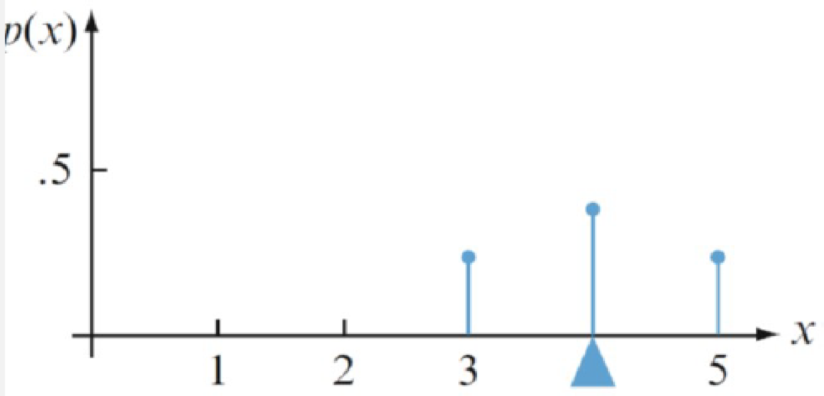
$$\mu = \sum_x x p(x)$$

- The expected value of X gives us an average value for X weighted by its probabilities. μ does not have to be a possible value for X.
- Geometrically, it is the point where the centre of gravity lies.

Visual Interpretation of $E(X)$

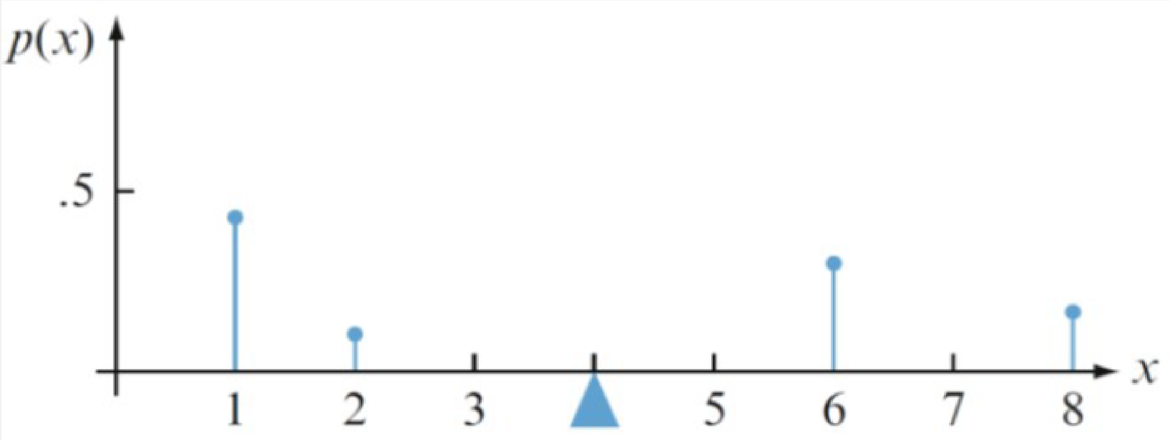
- **Example 1:**

x	$p(x)$	$xp(x)$
3	.3	.9
4	.4	1.6
5	.3	1.5
$\sum$	1	4



- **Example 2:**

x	$p(x)$	$xp(x)$
1	.4	.4
2	.1	.2
6	.3	1.8
8	.2	1.6
$\sum$	1	4



Expectation of a Function of a Random Variable

Let X be a discrete random variable with PMF $p(x)$ and let h be a function defined on the range of X. Then the expectation of $h(X)$ is calculated by

$$E[h(X)] = \sum_x h(x)p(x)$$

- For the random variable X in Example 1, compute $E(X^2)$ and $E(X^3+1)$.

x	$p(x)$	$xp(x)$	$x^2p(x)$	$(x^3+1)p(x)$
3	.3	.9	2.7	8.4
4	.4	1.6	6.4	26
5	.3	1.5	7.5	37.8
$\sum$	1	4	16.6	72.2

Hence $E(X^2) = 16.6$ and $E(X^3+1) =$ 72.2

Variance and Standard Deviation

- The distributions in the previous two examples have the same mean, but clearly one has a greater spread or variability than the other. The principal numerical measures of spread are variance and standard deviation.

For a discrete random variable X with PMF $p(x)$, the *variance* of X, denoted by $Var(X) = \sigma_X^2 = \sigma^2$, is given by

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

The *standard deviation* of X, denoted by $SD(X) = \sigma_X = \sigma$, is the square root of the variance of X.

$$\sigma = \sqrt{E\left[(X-\mu)^2\right]}$$

- The variance and standard deviation are always non-negative. They are zero if and only if the random variable takes a single value with probability 1, i.e., when variability is zero.

Properties of Expectation and Variance

Some important properties of expectation and variance.

For constants a and b,

1. $E(aX + b) = aE(X) + b$
2. $Var(aX + b) = a^2 Var(X)$
3. $SD(aX + b) = |a| SD(X)$
4. $Var(X) = E\left[X^2\right] - (E\left[X\right])^2 = E\left[X^2\right] - \mu^2$

- $E[h(X)] = h(E[X])$ *when h is a linear function*. The equality is not generally true.
- The last identity provides a more convenient way of computing variances.

Examples

We now compute variances and standard deviations for the two examples given earlier.

- **Example 1:**

x	$p(x)$	$xp(x)$	$x^2p(x)$
3	.3	.9	2.7
4	.4	1.6	6.4
5	.3	1.5	7.5
$\sum$	1	4	16.6

Hence,
$\sigma^2 = 16.6 - 4^2 = 0.6$
$\sigma = \sqrt{0.6} = 0.7746$

- **Example 2:**

x	$p(x)$	$xp(x)$	$x^2p(x)$
1	.4	.4	.4
2	.1	.2	.4
6	.3	1.8	10.8
8	.2	1.6	12.8
$\sum$	1	4	24.4

Hence,
$\sigma^2 = 24.4 - 4^2 = 8.4$
$\sigma = \sqrt{8.4} = 2.8983$

An Example

Example: A computer store purchased three computers wholesale at \$500 each and will retail them for \$1000 each. The supplier will repurchase any unsold stock at \$200 each. Let X be the number of computers sold, with probabilities $p(0) = 0.1$, $p(1) = 0.2$, $p(2) = 0.3$, $p(3) = 0.4$. Let $h(X)$ be the profit from selling X units.

Calculate:

- the expected value and variance of X.
- the expected profit, using $E[h(X)] = \sum_x h(x)p(x)$.
- the expected profit, using the expression for the expected value of a linear function.
- the variance of $h(X)$ using $Var[h(X)] = E\left[(h(X))^2\right] - [E(h(X))]^2$.
- the variance of $h(X)$ using the expression for the variance of a linear function.
- the standard deviation of $h(X)$.

Solution

The profit function is $h(X) = 500X - 300(3 - X) = 800X - 900$

x	$p(x)$	$xp(x)$	$x^2p(x)$	$h(x)$	$h(x)p(x)$	$h(x)^2p(x)$
0	.1	0	0	-900	-90	81,000
1	.2	.2	.2	-100	-20	2,000
2	.3	.6	1.2	700	210	147,000
3	.4	1.2	3.6	1,500	600	900,000
$\sum$	1	2	5	1,200	700	1,130,000

From which:

- $E(X) = 2,\ Var(X) = 5 - 2^2 = 1$
- $E[h(x)] = 700$
- $E[h(x)] = 800 \times 2 - 900 = 700$
- $Var[h(x)] = 1,130,000 - 700^2 = 640,000$
- $Var[h(x)] = 800^2 \times 1 = 640,000$
- $SD[h(X)] = \sqrt{640,000} = 800 \times Var(X) = 800$