COMP9020

Foundations of Computer Science

Lecture 16: Statistics

Lecturers: Katie Clinch (LIC)
Paul Hunter
Simon Mackenzie

Course admin: Nicholas Tandiono

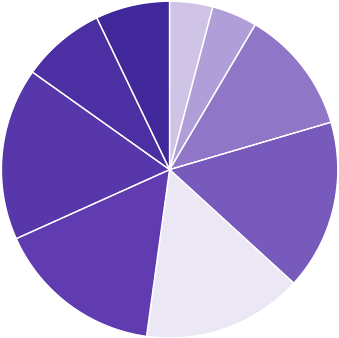
Course email: `cs9020@cse.unsw.edu.au`

Announcements

- **myExperience survey** - please fill this in
- **Only two deadlines remaining!**
  - Assignment 4 - deadline today (Thursday 18th April) at 6pm
  - Quiz 9 - deadline Wednesday 24th April at 6pm
- **Revision lecture tomorrow**
  - Vote for the topics you most want to review here.
  - Currently weeks 4, 5 and 7 have the most votes:

Legend (with response counts)

	Number theory (week 1)	18
	Sets and Languages (week 2)	19
	Graph theory (week 3)	52
	Relations (week 4)	71
	Functions (week 5)	67
	Algorithmic Analysis (week 5 video recording)	70
	Recursion and Induction (weeks 5 and 7)	72
	Logic (week 8)	35
	Counting, Probability and Statistics (weeks 9 and ...	31



Outline

Outline

Random Variables

Definition

An (integer-valued) **random variable** X is a function from Ω to $\mathbb{Z}$. In other words, it associates a number value with every outcome.

Random variables are often denoted by $X, Y, Z, \ldots$
We extend arithmetic to random variables in the natural way.

Definition

Given random variable $X : \Omega \to \mathbb{Z}$, random variable $Y : \Omega \to \mathbb{Z}$ and integer k, we can combine X, Y and k to obtain the following functions on all $\omega \in \Omega$:

Addition of variables:	$X + Y :$	$\omega \mapsto X(\omega) + Y(\omega)$
Multiplication of variables:	$X \cdot Y :$	$\omega \mapsto X(\omega) \cdot Y(\omega)$
Scalar addition:	$X - k :$	$\omega \mapsto X(\omega) - k$
Scalar multiplication:	$kX :$	$\omega \mapsto k \cdot X(\omega)$

Examples

Example

Random variable X: value of rolling one die

$\Omega = \{1, 2, 3, 4, 5, 6\}$

$X(i) = i$

Example

Random variable X_s: sum of rolling two dice

$\Omega = \{(1, 1), (1, 2), \ldots, (6, 6)\}$

$X_s((1, 1)) = 2 \qquad X_s((1, 2)) = 3 = X_s((2, 1)) \quad \ldots$

Question

Is $X_s = X + X$*?*

Examples

Example

Random variable X: value of rolling one die

$\Omega = \{1, 2, 3, 4, 5, 6\}$

$X(i) = i$

Example

Random variable X_s: sum of rolling two dice

$\Omega = \{(1, 1), (1, 2), \ldots, (6, 6)\}$

$X_s((1, 1)) = 2 \qquad X_s((1, 2)) = 3 = X_s((2, 1)) \ \ldots$

Question

Is $X_s = X + X$? No.

$X_s(\omega_1, \omega_2) = X(\omega_1) + Y(\omega_2)$. where $X : \omega_1 \mapsto \omega_1$ and $Y : \omega_2 \mapsto \omega_2$ are independent and identically distributed (i.i.d)

Expectation

Definition

The **expected value** (often called "expectation" or "average") of a random variable X is

$$E(X) = \sum_{k \in \mathbb{Z}} P(X = k) \cdot k$$

NB

Expectation is a truly universal concept; it is the basis of all decision making, of estimating gains and losses, in all actions under risk. Historically, a rudimentary concept of expected value arose long before the notion of probability.

Examples

Example

The expected value when rolling one die is:

$$E(X) = \frac{1}{6} \cdot 1 + \frac{1}{6} \cdot 2 + \ldots + \frac{1}{6} \cdot 6 = 3.5$$

Example

The expected sum when rolling two dice is

$$E(X_s) = \frac{1}{36} \cdot 2 + \frac{2}{36} \cdot 3 + \ldots + \frac{6}{36} \cdot 7 + \ldots + \frac{1}{36} \cdot 12 = 7$$

Examples

Example

$\boxed{\text{RW: 9.3.3}}$ **Question.** Buy one lottery ticket for \$1. The only prize is \$1M. Each ticket has probability $6 \cdot 10^{-7}$ of winning. What is the expected value of the lottery ticket?

Answer. There are two types of ticket, winning tickets and losing tickets, so we let $\Omega = \{win, lose\}$, and $X_L : \Omega \to \mathbb{Z}$ such that $X_L(win) = \$999,999$ and $X_L(lose) = -\$1$. Then

$$E(X_L) = 6 \cdot 10^{-7} \cdot \$999,999 + (1 - 6 \cdot 10^{-7}) \cdot -\$1 = -\$0.4.$$

Outline

Linearity of expectation

Theorem (linearity of expected value)

For any random variables X, Y and integer k:

- $E(X+Y) = E(X) + E(Y)$
- $E(k \cdot X) = k \cdot E(X)$

Example

The expected sum when rolling two dice can be computed as

$$E(X_s) = E(X) + E(Y) = 3.5 + 3.5 = 7$$

Example

Example

Question. Calculate $E(S_n)$, where

$$S_n \stackrel{\text{def}}{=} \text{no. of } \textsc{heads} \text{ in } n \text{ coin tosses}$$

Example

Example

Question. Calculate $E(S_n)$, where

$$S_n \stackrel{\text{def}}{=} \text{no. of HEADS in } n \text{ coin tosses}$$

Answer 1. (the 'hard way')
Using the definition of the expectation, we have

$$E(S_n) = \sum_{k=0}^{n} P(S_n = k) \cdot k = \sum_{k=0}^{n} \frac{1}{2^n} \binom{n}{k} \cdot k$$

Since there are $\binom{n}{k}$ sequences of n coin tosses with k HEADS, and each sequence has the probability $\frac{1}{2^n}$, this gives

$$E(S_n) = \frac{1}{2^n} \sum_{k=1}^{n} \frac{n}{k} \binom{n-1}{k-1} k = \frac{n}{2^n} \sum_{k=0}^{n-1} \binom{n-1}{k} = \frac{n}{2^n} \cdot 2^{n-1} = \frac{n}{2}$$

where we used the 'binomial identity' $\sum_{k=0}^{n} \binom{n}{k} = 2^n$ to simplify.

Example

Example

Question. Calculate $E(S_n)$, where

$$S_n \stackrel{\text{def}}{=} \text{no. of } \textsc{heads} \text{ in } n \text{ coin tosses}$$

Answer 2. (the 'easy way')
When $n = 1$, we have that $S_1 : \{\textsc{tails}, \textsc{heads}\} \to \{0, 1\}$ denotes the number of HEADS obtained in one coin toss.
For general $n > 1$, we can think of S_n as being equivalent to repeating a single S_1 coin toss, n distinct (and independent) times.
Let S_{1_i} denote coin toss number i in this sequence. Then

$$E(S_n) = E(S_{1_1} + S_{1_2} + \ldots + S_{1_n}) = \sum_{i=1}^{n} E(S_{1_i}) = n \cdot E(S_1) = n \cdot \frac{1}{2}$$

Observations

Fact

If $X_1, X_2, \ldots, X_n$ are independent, identically distributed random variables, then $E(X_1 + X_2 + \ldots + X_n) = E(nX_1) = nE(X_1)$.

NB

$X_1 + X_2 + \ldots + X_n$ and nX_1 are very different random variables.

Exercises

Exercise

Question. You face a quiz consisting of six true/false questions, and your plan is to guess the answer to each question (randomly, with probability 0.5 of being right). There are no negative marks, and answering four or more questions correctly suffices to pass.

1. What is the probability of passing?
2. What is the expected score?

Exercise

Question. You face a quiz consisting of six true/false questions, and your plan is to guess the answer to each question (randomly, with probability 0.5 of being right). There are no negative marks, and answering four or more questions correctly suffices to pass.

1. What is the probability of passing?
2. What is the expected score?

Answer (1). To pass you would need four, five or six correct guesses. Therefore,

$$P(\text{pass}) = \frac{\binom{6}{4} + \binom{6}{5} + \binom{6}{6}}{2^6} = \frac{15 + 6 + 1}{64} \approx 34\%.$$

Exercises

Exercise

Question. You face a quiz consisting of six true/false questions, and your plan is to guess the answer to each question (randomly, with probability 0.5 of being right). There are no negative marks, and answering four or more questions correctly suffices to pass.

1. What is the probability of passing?
2. What is the expected score?

Answer (1). To pass you would need four, five or six correct guesses. Therefore,

$$P(\text{pass}) = \frac{\binom{6}{4} + \binom{6}{5} + \binom{6}{6}}{2^6} = \frac{15 + 6 + 1}{64} \approx 34\%.$$

Answer (2). Let $X_i : \{\text{True}, \text{False}\} \to \{0, 1\}$ denote the possible points we can score on each question $i \in \{1, 2, 3, 4, 5, 6\}$. Our expected score for question i is $E(X_i) = 0.5(0) + 0.5(1) = 0.5$. So our expected score for the quiz, denoted $E(X)$ is

$$E(X) = E(X_1 + X_2 + \cdot + X_6) = 0.5(6) = 3.$$

Exercises

Exercise

RW: 9.3.7

Question. An urn has $m + n = 10$ marbles, $m \geq 0$ red and $n \geq 0$ blue. 7 marbles selected at random without replacement. What is the expected number of red marbles drawn?

Exercises

Exercise

RW: 9.3.7

Question. An urn has $m + n = 10$ marbles, $m \geq 0$ red and $n \geq 0$ blue. 7 marbles selected at random without replacement. What is the expected number of red marbles drawn?

Answer. Let R and B denote the number of red and blue marbles selected, such that $R + B = 7$.

Then

$$E(R) = \sum_{r=0}^{7} P(R = r, B = 7 - r) \cdot r$$

Observe that if $r > m$ or $7 - r > n$ then $P(R = r, B = 7 - r) = 0$.

Otherwise, we have that $P(R = r, B = 7 - r) = \frac{\binom{m}{r}\binom{n}{n-r}}{\binom{10}{7}}$, and so the above can be rewritten as:

$$E(R) = \frac{\binom{m}{0}\binom{n}{7}}{\binom{10}{7}} \cdot 0 + \frac{\binom{m}{1}\binom{n}{6}}{\binom{10}{7}} \cdot 1 + \frac{\binom{m}{2}\binom{n}{5}}{\binom{10}{7}} \cdot 2 + \ldots + \frac{\binom{m}{7}\binom{n}{0}}{\binom{10}{7}} \cdot 7$$

Outline

Example

Example

Question. Find the average waiting time for the first HEAD, with no upper bound on the 'duration' (one allows for all possible sequences of tosses, regardless of how many times TAILS occur initially).

Answer. Let Ω be the sample space of all possible sequences of H and T.

Let $X_w : \Omega \to \mathbb{N}$ such that $X_w(\omega)$ is the first location in ω containing an H (i.e. the waiting time for ω). For example $X_w(TTHTH \cdots) = 3$.

Then the average waiting time is

$$\begin{aligned} A = E(X_w) &= \sum_{k=1}^{\infty} k \cdot P(X_w = k) = \sum_{k=1}^{\infty} k \frac{1}{2^k} \\ &= \frac{1}{2^1} + \frac{2}{2^2} + \frac{3}{2^3} + \cdots \end{aligned}$$

Expected time to success

How can we calculate $A = \frac{1}{2^1} + \frac{2}{2^2} + \frac{3}{2^3} + \cdots$?

Method 1: Geometric progressions

$$
\begin{aligned}
A &= \frac{1}{2} + \frac{2}{2^2} + \frac{3}{2^3} + \cdots \\
&= \left(\frac{1}{2} + \frac{1}{2^2} + \frac{1}{2^3} + \cdots\right) + \left(\frac{1}{2^2} + \frac{1}{2^3} + \cdots\right) + \left(\frac{1}{2^3} + \cdots\right) + \cdots \\
&= 1 + \frac{1}{2} + \frac{1}{2^2} + \cdots = 2.
\end{aligned}
$$

Method 2: Recursion

$$
A = \sum_{k=1}^{\infty} \frac{k}{2^k} = \sum_{k=1}^{\infty} \frac{k-1}{2^k} + \sum_{k=1}^{\infty} \frac{1}{2^k} = \frac{1}{2} \sum_{k=1}^{\infty} \frac{k-1}{2^{k-1}} + 1 = \frac{1}{2}A + 1
$$

So $A = \frac{A}{2} + 1$, which gives $A = 2$.

NB

A much simpler but equally valid argument is that you expect 'half' a HEAD *in 1 toss, so you ought to get a 'whole'* HEAD *in 2 tosses.*

Exercise

Exercise

$\boxed{\text{RW: 9.4.12}}$ **Question.** A die is rolled until the first 4 appears. What is the expected waiting time?

Exercise

Exercise

$\boxed{\text{RW: 9.4.12}}$ **Question.** A die is rolled until the first 4 appears. What is the expected waiting time?

Answer. $P(\text{roll } 4) = \frac{1}{6}$ hence $E(\text{no. of rolls until first } 4) = 6$.

Example

Question. To find an object $\mathcal{X}$ in an unsorted list L of n elements, one needs to search linearly through L. Let the probability of $\mathcal{X} \in L$ be p. Then there is $1 - p$ likelihood of $\mathcal{X}$ being absent altogether. Find the expected number of comparison operations.

Example

Question. To find an object $\mathcal{X}$ in an unsorted list L of n elements, one needs to search linearly through L. Let the probability of $\mathcal{X} \in L$ be p. Then there is $1-p$ likelihood of $\mathcal{X}$ being absent altogether. Find the expected number of comparison operations.

Answer. If the element $\mathcal{X}$ is in the list, then the number of comparisons averages to $\frac{1}{n}(1+\ldots+n)$. If $\mathcal{X}$ is absent, we need n comparisons.
The first case has probability p, the second $1-p$. Combining these we find

$$\begin{aligned} E_n &= p\left(\frac{1+\ldots+n}{n}\right)+(1-p)n \\ &= p\left(\frac{n+1}{2}\right)+(1-p)n \\ &= \left(1-\frac{p}{2}\right)n+\frac{p}{2} \end{aligned}$$

As one would expect, increasing p leads to a lower E_n.

Success vs Expected value

Question

Does high probability of success lead to a high expected value?

Success vs Expected value

Question

Does high probability of success lead to a high expected value?

Generally, no.

Example

Buying more tickets in the lottery increases your chances of winning, but the expected value of winnings *decreases*.

Example

Example

Roulette (outcomes $0, 1, \ldots, 36$). Win: $35 \times$ bet

Strategy 1: Bet $\$1$ on a single number

- Probability of winning: $\frac{1}{37}$
- Expected winnings: $\frac{1}{37}.(\$35) + \frac{36}{37}(-\$1) \approx -2.7c$

Example

Example

Roulette (outcomes $0, 1, \ldots, 36$). Win: $35 \times \text{bet}$

Strategy 2: Place $\$1$ bets on 24 numbers, selected from among 0 to 36.

- Probability of winning: $\frac{24}{37} \approx 65\%$
- Expected winnings:
  - If one of the numbers comes up, win \$35 from the bet on that number and lose \$23 from the bets on the remaining numbers, thus collecting \$12.
    This happens with probability $p = \frac{24}{37}$.
  - With probability $q = \frac{13}{37}$ none of the numbers appear, leading to loss of \$24.

So expected winnings are:
$p \cdot \$12 - q \cdot \$24 = \$12\frac{24}{37} - \$24\frac{13}{37} = -\$\frac{24}{37} \approx -65c = 24 \times -2.7c$

Gambler's ruin

Many so-called 'winning systems' that purport to offer a winning strategy do something akin — they provide a scheme for frequent relatively moderate wins, but at the cost of an occasional very big loss.

It turns out (it is a formal theorem) that there can be *no system* that converts an 'unfair' game into a 'fair' one. In the language of decision theory, 'unfair' denotes a game whose individual bets have negative expectation.

It can be easily checked that any individual bets on roulette, on lottery tickets or on just about any commercially offered game have negative expected value.

Outline

Standard Deviation and Variance

Definition

For random variable X with expected value (or: **mean**) $\mu = E(X)$, the **standard deviation** of X is

$$\sigma = \sqrt{E((X-\mu)^2)}$$

and the **variance** of X is

$$\sigma^2$$

Standard deviation and variance measure how spread out the values of a random variable are. The smaller σ^2 the more confident we can be that $X(\omega)$ is close to $E(X)$, for a randomly selected ω.

NB

The variance can be calculated as $E((X-\mu)^2) = E(X^2) - \mu^2$

Example

Example

Question. Let the random variable $X_d \stackrel{\text{def}}{=}$ value of a rolled die. What is (a) the mean, (b) the variance and (c) the standard deviation?

(a)

$$\mu = E(X_d) = 3.5$$

(b)

$$E(X_d^2) = \frac{1}{6} \cdot 1 + \frac{1}{6} \cdot 4 + \frac{1}{6} \cdot 9 + \frac{1}{6} \cdot 16 + \frac{1}{6} \cdot 25 + \frac{1}{6} \cdot 36 = \frac{91}{6}$$

$$\text{Hence } \sigma^2 = E(X_d^2) - \mu^2 = \frac{35}{12}$$

(c)

$$\sigma = \sqrt{\frac{35}{12}} \approx 1.71$$

Exercises

Exercises

Question (Supp). Two independent experiments are performed.
$p_1 = P(\text{1st experiment succeeds}) = 0.7$
$p_2 = P(\text{2nd experiment succeeds}) = 0.2$ Random variable X counts the number of successful experiments.

a. Expected value of X?
b. Probability of exactly one success?
c. Probability of at most one success?
d. Variance of X?

Exercises

Question (Supp). Two independent experiments are performed.
$p_1 = P(\text{1st experiment succeeds}) = 0.7$
$p_2 = P(\text{2nd experiment succeeds}) = 0.2$ Random variable X counts the number of successful experiments.

- (a) Expected value of X?
- (b) Probability of exactly one success?
- (c) Probability of at most one success?
- (d) Variance of X?

Answer. For experiment i, let $X_i = 1$ if the experiment is successful and $X_i = 0$ otherwise.
(a) $E(X) = E(X_1) + E(X_2) = 1 \cdot p_1 + 1 \cdot p_2 = 0.7 + 0.2 = 0.9.$
(b) $P(X = 1) = P(X_1 = 1, X_2 = 0) + P(X_1 = 0, X_2 = 1)$
$$= 0.7 \cdot 0.8 + 0.3 \cdot 0.2 = 0.62.$$
(c) $P(X = 1) + P(X_1 = X_2 = 0) = 0.62 + 0.3 \cdot 0.8 = 0.86.$
(d) $\sigma^2 = E(X^2) - E(X)^2 = (0.62 \cdot 1 + 0.14 \cdot 4) - 0.9^2 = 0.37.$