# COMP9020 Week 10
# Term 3, 2019
# Statistics

- Textbook (R & W) - Ch. 5, Sec. 5.1–5.3; Ch. 9
- Supplementary Exercises Ch. 5, 9 (R & W)

# Applications to CS

Statistics let you extract useful information about random processes.

Applications in CS include

- Sampling from large data sets
- Identifying anomolies
- Making predictions

# Summary of topics

- Random variables and expectation
- Linearity of expectation
- Expected time to success
- Standard deviation and variance

# Summary of topics

- Random variables and expectation
- Linearity of expectation
- Expected time to success
- Standard deviation and variance

# Random Variables

## Definition

An (integer) **random variable** is a function from $\Omega$ 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 variables $X$, $Y$ and integer $k$:

$$\begin{aligned} X+Y: &\quad \omega \mapsto X(\omega)+Y(\omega) \\ X.Y: &\quad \omega \mapsto X(\omega).Y(\omega) \\ X-k: &\quad \omega \mapsto X(\omega)-k \\ kX: &\quad \omega \mapsto k.X(\omega) \end{aligned}$$

# 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.*

$X_s = X + Y$ where $X$ and $Y$ 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

$\boxed{9.3.3}$ Buy one lottery ticket for \$1. The only prize is \$1M. Each ticket has probability $6 \cdot 10^{-7}$ of winning.

$\Omega = \{win, lose\} \qquad X_L(win) = \$999,999 \qquad X_L(lose) = -\$1$

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

# Summary of topics

- Random variables and expectation
- Linearity of expectation
- Expected time to success
- Standard deviation and variance

# Linearity of expectation

**Theorem (linearity of expected value)**

*For any random variables $X, Y$ and integer $k$:*

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

**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

$E(S_n)$, where $S_n \stackrel{\text{def}}{=} |\text{no. of HEADS in } n \text{ tosses}|$

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

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

# Example

## Example

$E(S_n)$, where $S_n \stackrel{\text{def}}{=} |\text{no. of HEADS in } n \text{ tosses}|$

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

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

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

  using the 'binomial identity' $\sum_{k=0}^{n} \binom{n}{k} = 2^n$

# Example

## Example

$E(S_n)$, where $S_n \stackrel{\text{def}}{=} |\text{no. of } \textsc{heads} \text{ in } n \text{ tosses}|$

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

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

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

  using the 'binomial identity' $\sum_{k=0}^{n} \binom{n}{k} = 2^n$
- 'easy way'
  $E(S_n) = E(S_1^1 + \ldots + S_1^n) = \sum_{i=1\ldots n} E(S_1^i) = nE(S_1) = n \cdot \frac{1}{2}$

Note: $S_n \stackrel{\text{def}}{=} |\textsc{heads} \text{ in } n \text{ tosses}|$ while each $S_1^i \stackrel{\text{def}}{=} |\textsc{heads} \text{ in } 1 \text{ toss}|$

# 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.*

# Exercise

## Exercise

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.
What is the probability of passing and what is the expected score?

# Exercise

## Exercise

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.

What is the probability of passing and what is the expected score?

To pass you would need four, five or six correct guesses. Therefore,

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

The expected score from a single question is 0.5, as there is no penalty for errors. For six questions the expected value is $6 \cdot 0.5 = 3$

# Exercise

$\boxed{9.3.7}$

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?

# Exercise

$\boxed{9.3.7}$

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?

$$\frac{\binom{m}{0}\binom{n}{7}}{\binom{10}{7}} \cdot 0 + \frac{\binom{m}{1}\binom{n}{6}}{\binom{10}{7}} \cdot 1 + \frac{\binom{m}{2}\binom{n}{5}}{\binom{10}{7}} \cdot 2 + \ldots + \frac{\binom{m}{7}\binom{n}{0}}{\binom{10}{7}} \cdot 7$$

e.g.

$$\frac{\binom{5}{2}\binom{5}{5}}{\binom{10}{7}} \cdot 2 + \frac{\binom{5}{3}\binom{5}{4}}{\binom{10}{7}} \cdot 3 + \frac{\binom{5}{4}\binom{5}{3}}{\binom{10}{7}} \cdot 4 + \frac{\binom{5}{5}\binom{5}{2}}{\binom{10}{7}} \cdot 5$$

$$= \frac{10}{120} \cdot 2 + \frac{50}{120} \cdot 3 + \frac{50}{120} \cdot 4 + \frac{10}{120} \cdot 5 = \frac{420}{120} = 3.5$$

## Summary of topics

- Random variables and expectation
- Linearity of expectation
- Expected time to success
- Standard deviation and variance

# Example

## Example

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).

$$A = E(X_w) = \sum_{k=1}^{\infty} k \cdot P(X_w = k) = \sum_{k=1}^{\infty} k \frac{1}{2^k}$$
$$= \frac{1}{2^1} + \frac{2}{2^2} + \frac{3}{2^3} + \dots$$

# Example

## Example

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).

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

This can be evaluated by breaking the sum into a sequence of geometric progressions

$$\frac{1}{2} + \frac{2}{2^2} + \frac{3}{2^3} + \ldots$$

$$= \left(\frac{1}{2} + \frac{1}{2^2} + \frac{1}{2^3} + \ldots\right) + \left(\frac{1}{2^2} + \frac{1}{2^3} + \ldots\right) + \left(\frac{1}{2^3} + \ldots\right) + \ldots$$

$$= 1 + \frac{1}{2} + \frac{1}{2^2} + \ldots = 2$$

# Expected time to success

There is also a recursive 'trick' for solving the sum

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

Now $A = \frac{A}{2} + 1$ and $A = 2$

**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.*

**Theorem**

*If the probability of success is $p$ then:*

- *The expected number of (indep.) trials before 1 success is $\frac{1}{p}$*
- *The expected number of (indep.) trials before $k$ successes is $\frac{k}{p}$*

# Exercise

## Exercise

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

# Exercise

**Exercise**

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

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

## Example

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

## Example

To find an object $\mathcal{X}$ in an unsorted list $L$ of elements, one needs to search linearly through $L$. Let the probability of $\mathcal{X} \in L$ be $p$, hence there is $1 - p$ likelihood of $\mathcal{X}$ being absent altogether. Find the expected number of comparison operations.
If the element is in the list, then the number of comparisons averages to $\frac{1}{n}(1 + \ldots + n)$; if absent we need $n$ comparisons. The first case has probability $p$, the second $1 - p$. Combining these we find

$$E_n = p\frac{1 + \ldots + n}{n} + (1 - p)n = p\frac{n+1}{2} + (1 - p)n = (1 - \frac{p}{2})n + \frac{p}{2}$$

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: $35 \times \text{bet}$

**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.

# Summary of topics

- Random variables and expectation
- Linearity of expectation
- Expected time to success
- Standard deviation and variance

# 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 $\sigma^2$ the more confident we can be that $X(\omega)$ is close to $E(X)$, for a randomly selected $\omega$.

## NB

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

# Example

## Example

Random variable $X_d \stackrel{\text{def}}{=}$ value of a rolled die

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

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

$$\text{Hence,} \quad \sigma^2 = E(X_d^2) - \mu^2 = \frac{35}{12} \quad \to \quad \sigma \approx 1.71$$

# Exercises

**Exercises**

$\boxed{9.5.10}$ (Supp) Two independent experiments are performed.
$P(\text{1st experiment succeeds}) = 0.7$
$P(\text{2nd experiment succeeds}) = 0.2$
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?

(e) Variance of $X$?

# Exercises

## Exercises

$\boxed{9.5.10}$ (Supp) Two independent experiments are performed.
$P(\text{1st experiment succeeds}) = 0.7$
$P(\text{2nd experiment succeeds}) = 0.2$
Random variable $X$ counts the number of successful experiments.

(a) Expected value of $X$? $E(X) = 0.7 + 0.2 = 0.9$

(b) Probability of exactly one success? $0.7 \cdot 0.8 + 0.3 \cdot 0.2 = 0.62$

(c) Probability of at most one success? (b)$+0.3 \cdot 0.8 = 0.86$

(e) Variance of $X$? $\sigma^2 = (0.62 \cdot 1 + 0.14 \cdot 4) - 0.9^2 = 0.37$