COMP1311 Discrete probability & statistics

Andersen Ang
U.Southampton UK
angms.science

December 16, 2024

1st draft
May 22, 2023

Pre-course information

- What is probability & statistics: modelling *uncertainty*
  $\Longrightarrow$ important for CS

- We study **discrete (classical) probability**
  - We study probability using combinatorics

- ~~We study **continuous statistics** using calculus~~

- Material: lecture slides + $\underbrace{\text{workbook + reading + online video}}_{\text{self learning}}$

- Book
  - *Discrete Mathematics and Its Applications* by Kenneth Rosen enough for this course
  - *Concrete mathematics: a foundation for computer science* by Graham, Knuth & Patashnik classic
  - *Schaum's Outline of probability and statistics* for practise

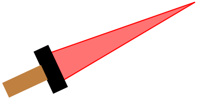
- Outcome: become *less ignorant* in probability & statistics

Prerequisite

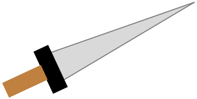
- Set theory: probability is defined by set
  - Notation of set
  - Membership, subset
  - Complement, cardinality
  - Union, intersection, set minus / relative complement

- Combinatorics: techniques carry to probability
  - Sum rule, incl-excl principle, complement, product rule, division rule
  - Permutation, combination, binomial, multinomial
  - Generating function

Which weapon is better?



The Epic Excalibur of Externality Covered by Prismatic Dragon-blood
Physical Damage: 6-13.2
Attacks Per Second: 1.45
Critical Strike Chance: 8%
Every Third Strike Deals Triple Damage



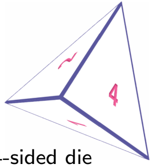
Just a big sword
Physical Damage: 10-41
Attacks Per Second: 1

Contents

Sample space, event and classical probability

- **Def (Sample space)**
  The set of all possible outcome is called the **sample space** Ω
  - $\Omega \neq \varnothing$ (non-triviality)
- **E.g.** (Tossing a coin)
  - Possible output = Head H or Tail T
  - Ω(tossing a coin once) $= \{H, T\}$
  - Ω(tossing a coin twice) $= \{HH, HT, TH, TT\}$
  - Ω(tossing a coin thrice) $= \{HHH, HHT, HTH, THH, HTT, THT, TTH, TTT\}$
- **Def (Event)**
  Any subset of Ω is called an **event** E.
- By set theory we have $E \subset \Omega$ and $\Omega = \bigcup E$
- **Def (Classical probability)**
  The **classical probability** of an event E is $\mathbb{P}(E) := \dfrac{|E|}{|\Omega|}$.
  Mathematics does not allow divided-by-zero $\iff$ $\Omega \neq \varnothing$ (non-triviality)
  However $\begin{cases} E \text{ is possibly empty} \\ \Omega \text{ is possibly infinite} \end{cases}$

E.g. : 4-sided die in Dungeons & Dragon



- 1d4 = roll one 4-sided die
- $\Omega(\text{1d4}) = \{1, 2, 3, 4\}$
- $E_1 :=$ "less than or equal to 3"
  $$\mathbb{P}(E_1) = \frac{|E_1|}{|\Omega|} = \frac{|\{1,2,3\}|}{|\{1,2,3,4\}|} = \frac{3}{4} = 0.75$$
- $E_2 :=$ "even number"
  $$\mathbb{P}(E_2) = \frac{|E_2|}{|\Omega|} = \frac{|\{2,4\}|}{|\{1,2,3,4\}|} = \frac{2}{4} = 0.5$$
- $E_3 :=$ "larger than zero"
  $$\mathbb{P}(E_3) = \frac{|E_3|}{|\Omega|} = \frac{|\{1,2,3,4\}|}{|\{1,2,3,4\}|} = \frac{4}{4} = 1$$
- $E_4 :=$ "less than -2"
  $$\mathbb{P}(E_4) = \frac{|E_4|}{|\Omega|} = \frac{|\varnothing|}{|\{1,2,3,4\}|} = \frac{0}{4} = 0$$
  Remark: $\varnothing$ is always a subset of any set

- 2d4 = roll two 4-sided dies
  $$\Omega(\text{2d4}) = \begin{Bmatrix} (1,1) & (1,2) & (1,3) & (1,4) \\ (2,1) & (2,2) & (2,3) & (2,4) \\ (3,1) & (3,2) & (3,3) & (3,4) \\ (4,1) & (4,2) & (4,3) & (4,4) \end{Bmatrix}, \qquad |\Omega| = 16$$
- $E_5 := \{(i,j) \mid i + j > 6\}$
  $$\mathbb{P}(E_5) = \frac{|E_5|}{|\Omega|} = \frac{|\{(3,3),(3,4),(4,3),(4,4)\}|}{16} = 0.25$$
- $E_6 := \{(i,j) \mid i < j\}$
  $$\mathbb{P}(E_6) = \frac{|E_6|}{|\Omega|} = \frac{|\{(1,2),(1,3),(1,4),(2,3),(2,4),(3,4)\}|}{16} = 0.375$$
- $E_7 := \{(i,j) \mid i = j + 1\}$ $\qquad \mathbb{P}(E_7) = \frac{3}{16}$
- $E_8 := \{(i,j) \mid i + j \text{ is a prime number}\}$ $\qquad \mathbb{P}(E_8) = \frac{9}{16}$

Three probability axioms

- **Axiom 0** (non-triviality) $\Omega \neq \varnothing$
- **Axiom 1** (nonnegativity) $\mathbb{P}(E) \;\geq\; 0$
- **Axiom 2** (sample space has probability 1) $\mathbb{P}(\Omega) \;\equiv\; 1$
- **Axiom 3** (σ-additivity) If $E_1, E_2, \ldots$ are disjoint, then

$$\mathbb{P}\Big(\bigcup_i E_i\Big) \;=\; \sum_i \mathbb{P}(E_i). \qquad (\sigma\text{-additivity})$$

  - In set: two sets A, B are disjoint $\iff A \cap B = \varnothing \iff$ they share nothing common
  - In combinatorics: we do not allow cross-terms in the inclusion-exclusion principle
  - In probability: two events E, F are *mutually exclusive* $\iff$ they can't occur at the same time

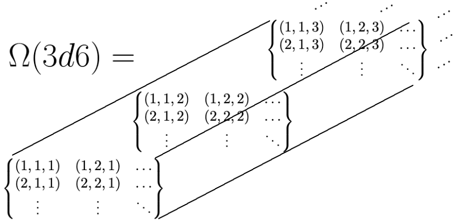
- These axioms imply
  - $\mathbb{P}(E) \leq 1 \; \forall E$
  - $\mathbb{P}(\varnothing) = 0.$
  - If $E \subset F$, then $\mathbb{P}(E) \leq \mathbb{P}(F)$
    Proof $F \overset{E \subseteq F}{=} E \cup (E^c \cap F)$, so $\mathbb{P}(F) = \mathbb{P}\big(E \cup (E^c \cap F)\big) \overset{\text{Axiom3}}{=} \mathbb{P}(E) + \underbrace{\mathbb{P}\big(E^c \cap F\big)}_{\overset{\text{Axiom1}}{\geq} 0} \geq \mathbb{P}(E).$

Complementary event

- **Definition** The **complementary event** of E in Ω, denoted as E^c, is defined as $E^c := \Omega \setminus E$.
- **Theorem** $\mathbb{P}(E^c) = 1 - \mathbb{P}(E)$.
Proof: $1 \overset{\text{Axiom 1}}{=} \mathbb{P}(\Omega) = \mathbb{P}(E \cup E^c) \overset{\text{Axiom 3}}{=} \mathbb{P}(E) + \mathbb{P}(E^c)$.
- **E.g.** $\Omega(\text{tossing a coin twice}) = \{HH, HT, TH, TT\}$
  - $E :=$ "at least one H" $= \{HH, HT, TH\}$
  - $\mathbb{P}(E) = \dfrac{|E|}{|\Omega|} = \dfrac{|\{HH, HT, TH\}|}{|\{HH, HT, TH, TT\}|} = \dfrac{3}{4}$
  - $E^c = \Omega \setminus E = \{TT\}$
  - $\mathbb{P}(E^c) = \dfrac{|E^c|}{|\Omega|} = \dfrac{|\{TT\}|}{|\{HH, HT, TH, TT\}|} = \dfrac{1}{4}$
  - $\mathbb{P}(E^c) = 1 - \mathbb{P}(E)$ is true

- E^c is useful when counting E is tedious
- 3d6 = roll three 6-sided die thrice

$\Omega(3d6) =$



- Let $E_9 := \{(i,j,k) \mid i+j+k < 18\}$, then

$$\begin{aligned}
\mathbb{P}(E_9) &= 1 - \mathbb{P}(E_9^c) \\
&= 1 - \mathbb{P}(\{(i,j,k) \mid i+j+k = 18\}) \\
&= 1 - \frac{\left|\{(i,j,k) \mid i+j+k = 18\}\right|}{|\Omega|} \\
&= 1 - \frac{1}{6^3} \approx 0.9953
\end{aligned}$$

(Two) Mutually exclusive events $\equiv$ disjoint $:=$ can't occur at the same time

- Complementary vs mutually exclusive
  - Complementary $\Longrightarrow$ mutually exclusive
  - Complementary $\not\Longleftarrow$ mutually exclusive

  E.g. : $\Omega = \{1,2,3\}, E = \{1\}, F = \{2,3\}, G = \{3\}$
  - E, F are mutually exclusive ($\because E \cap F = \varnothing$)
  - E, G are also mutually exclusive ($\because E \cap G = \varnothing$)
  - F, G are not mutually exclusive ($\because F \cap G = G \neq \varnothing$)
  - $F = E^c = \Omega \setminus E$
  - $G \neq E^c = \Omega \setminus E$.

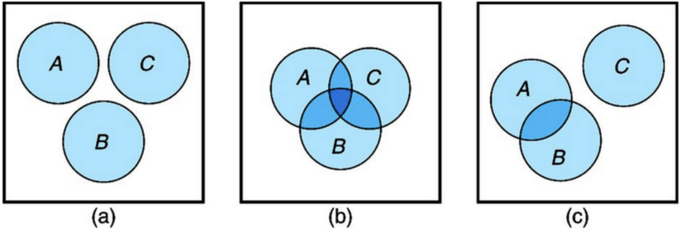
- E, F mutually exclusive $\iff \mathbb{P}(E \cup F) = \mathbb{P}(E) + \mathbb{P}(F)$ sum rule (prob. ver.)

- E, F not mutually exclusive $\iff \mathbb{P}(E \cup F) = \mathbb{P}(E) + \mathbb{P}(F) - \mathbb{P}(E \cap F)$ incl-excl principle (prob. ver.)

- **Theorem** $\mathbb{P}(E \cup F) = \mathbb{P}(E) + \mathbb{P}(F) - \mathbb{P}(E \cap F)$
  Proof
  $$\mathbb{P}(E \cup F) = \mathbb{P}(E \cup (E^c \cap F)) \overset{\text{Axiom3}}{=} \mathbb{P}(E) + \mathbb{P}(E^c \cap F) \qquad (*)$$
  Since $F = (E \cap F) \cup (E^c \cap F)$, so
  $$\mathbb{P}(F) = \mathbb{P}(E \cap F) + \mathbb{P}(E^c \cap F) \quad \Longrightarrow \quad \mathbb{P}(E^c \cap F) = \mathbb{P}(F) - \mathbb{P}(E \cap F) \qquad (**)$$
  Put $(**)$ into $(*)$ gives $\mathbb{P}(E \cup F) = \mathbb{P}(E) + \mathbb{P}(F) - \mathbb{P}(E \cap F)$.

Multiple mutually exclusive events

- If A, B, C are mutually exclusive, that means
  - If A occurs, B and C do not occur
  - If B occurs, A and C do not occur
  - If C occurs, A and B do not occur
- If $E_1, E_2, E_3, \ldots$ are mutually exclusive, that means if E_j occurs, all $E_{\neq j}$ do not occur
- **E.g.** $\Omega = \{1,2,3,4,5,6,7,8,9\}, E_1 = \{1,2\}, E_2 = \{3,4,5\}, E_3 = \{6,7,8,9\}$
  - E_1, E_2, E_3 are mutually exclusive to each other
- **E.g.** $\Omega = \{1,2,3,4,5,6,7,8,9\}, E_1 = \{1,2\}, E_2 = \{2,3,4,5,6,7\}, E_3 = \{6,7,8,9\}$
  - E_1, E_2, E_3 are not mutually exclusive to each other, because $E_2 \cap E_3 \neq \varnothing$



Mutually exclusive? a: yes b: no c: no

Probability of multiple events: Inclusion-exclusion principle (probability ver.)

- Inclusion-exclusion principle (probability ver.)
  - A, B mutually exclusive $\iff \mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$ $\quad (*)$
  - A, B not mutually exclusive $\iff \mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$ $\quad (\#)$
- **E.g.** $\Omega = \{1,2,3,4,5,6\}$, $E_1 = \{1\}, E_2 = \{2,3\}, E_3 = \{3,4\}, E_4 = \{4,5,6\}, E_5 = \{6\}$

$$\begin{aligned}
\mathbb{P}\Big(\bigcup_{i=1}^{5} E_i\Big) &= \mathbb{P}(E_1 \cup E_2 \cup E_3 \cup E_4 \cup E_5) \\
&\overset{(*)}{=} \mathbb{P}(E_1) + \mathbb{P}(E_2 \cup E_3 \cup E_4 \cup E_5) \\
&\overset{(\#)}{=} \mathbb{P}(E_1) + \mathbb{P}(E_2) + \mathbb{P}(E_3 \cup E_4 \cup E_5) - \mathbb{P}\Big(E_2 \cap (E_3 \cup E_4 \cup E_5)\Big) \\
&= \mathbb{P}(E_1) + \mathbb{P}(E_2) + \mathbb{P}(E_3 \cup E_4 \cup E_5) - \mathbb{P}\big(\{3\}\big) \\
&\overset{(\#)}{=} \mathbb{P}(E_1) + \mathbb{P}(E_2) + \mathbb{P}(E_3 \cup E_4) + \mathbb{P}(E_5) - \mathbb{P}\Big((E_3 \cup E_4 \cup) \cap E_5\Big) - \mathbb{P}\big(\{3\}\big) \\
&= \mathbb{P}(E_1) + \mathbb{P}(E_2) + \mathbb{P}(E_3 \cup E_4) + \mathbb{P}(E_5) - \mathbb{P}\big(\{6\}\big) - \mathbb{P}\big(\{3\}\big) \\
&\overset{(\#)}{=} \mathbb{P}(E_1) + \mathbb{P}(E_2) + \mathbb{P}(E_3) + \mathbb{P}(E_4) - \mathbb{P}(E_3 \cap E_4) + \mathbb{P}(E_5) - \mathbb{P}\big(\{6\}\big) - \mathbb{P}\big(\{3\}\big) \\
&= \mathbb{P}(E_1) + \mathbb{P}(E_2) + \mathbb{P}(E_3) + \mathbb{P}(E_4) - \mathbb{P}(\{4\}) + \mathbb{P}(E_5) - \mathbb{P}\big(\{6\}\big) - \mathbb{P}\big(\{3\}\big) \\
&= \frac{1}{6} + \frac{2}{6} + \frac{2}{6} + \frac{3}{6} - \frac{1}{6} + \frac{1}{6} - \frac{1}{6} - \frac{1}{6} = \frac{6}{6} = 1 = \mathbb{P}(\Omega) = \mathbb{P}\Big(\bigcup_{i=1}^{5} E_i\Big)
\end{aligned}$$

Exercise: find $\mathbb{P}(E_2 \cup E_3 \cup E_4 \cup E_5)$ without using complement.

Non-trivial / advanced topics Not in exam

- What if we toss a coin infinitely many times?
- Zero probability $\neq$ impossibility / never happens
- Probability 1 $\neq$ absolute / always happens
- Actually, what is probability?
  - Classical interpretation $\leftarrow$ we focus **discrete probability**
  - Frequentist interpretation **continuous probability**
  - Bayesian interpretation **continuous probability**

  Bayesian epistemology is a foundation of modern philosophy of science.
- Measure theory: formalize continuous probability
  - Measure
  - σ-algebra

  stackexchange.com: Why do we need sigma-algebras to define probability spaces?

Section summary

1. Probability is about three things $(\Omega, E, \mathbb{P})$
   - Sample space Ω: the set of all possible outcome $\Omega \neq 0$ (non-triviality)
   - Event E: a set of possible outcomes in the sample space
   - Classical definition of probability $\mathbb{P}(E) = \frac{|E|}{|\Omega|}$

2. Three axioms:

2.1 $\mathbb{P}(E) \geq 0$
2.2 $\mathbb{P}(\Omega) \equiv 1$
2.3 $\mathbb{P}\Big(\bigcup_i E_i\Big) = \sum_i \mathbb{P}(E_i)$ if E_i are disjoint

3. Complementary event $E^c := \Omega \setminus E$ and $\mathbb{P}(E^c) = 1 - \mathbb{P}(E)$

4. Disjoint / Mutually exclusive event
   - A, B mutually exclusive $\iff \mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$ sum rule (prob. ver.)
   - A, B not mutually exclusive $\iff \mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$ incl-excl principle (prob. ver.)

Contents

All combinatorics techniques carry over to probability.

- $k!$
- $k^{\underline{n}}$
- $\binom{n}{k}$
- $\binom{n}{k_1, k_2, \ldots, k_r}$
- $\left\langle {n \atop r} \right\rangle$
- Generating function & techniques
- Inclusion-Exclusion Principle

Hi we are back !

You will see them later

- Bernoulli $p(x|\theta) = \theta^x (1-\theta)^{1-x}, \theta \in [0,1], x \in \mathbb{N}$ — toss coin 1 times, 1 success
- Binomial $p(k|n,\theta) = \begin{pmatrix} n \\ k \end{pmatrix} \theta^k (1-\theta)^{n-k}, n, k \in \mathbb{N}$ — toss coin n times, k success
- Geometric $p(k|\theta) = (1-\theta)^{k-1}\theta, k \in \{1, 2, ...\}$ — toss coin k times, first success at the kth time
- Hypergeometric $p(k|N,K,n) = \dfrac{\begin{pmatrix} K \\ k \end{pmatrix}\begin{pmatrix} N-K \\ n-k \end{pmatrix}}{\begin{pmatrix} N \\ n \end{pmatrix}}$ — k succeed of n draw with no replacement in N-choose-K
- Negative binomial $p(m|\theta) = \left\langle \begin{matrix} n \\ k \end{matrix} \right\rangle (1-\theta)^m \theta^n$ — toss coin n times, k fails
- Trinomial and multinomial $p(k_1, k_2, k_3|n, \theta_1, \theta_2, \theta_3) = \begin{pmatrix} n \\ k_1, k_2, k_3 \end{pmatrix} \theta_1^{k_1}\theta_2^{k_2}\theta_3^{k_3}$ — generalized binomial
- Poisson $p(k|\lambda) = \dfrac{\lambda^k e^{-\lambda}}{k!}, \lambda \geq 0, k \in \mathbb{N}$ — probability of k events occur during an interval

Probability Generating Function

- Power series

$$F(x) = f_0 + f_1 x + f_2 x^2 + \ldots + f_k x^k + \ldots$$

- Probability generating function

$$G(x) = p_0 + p_1 z + p_2 z^2 + \ldots + p_k z^k + \ldots$$

- Probability mass function

$$p_0, p_1, p_2, p_3, \ldots$$

$$p_k = \mathbb{P}(\text{event occur } k \text{ times}) = \frac{1}{k!} \frac{\mathrm{d}}{\mathrm{d}x^k} G(x) \Big|_{x=0}$$

- Probability exponential generating function

$$E(x) = p_0 + p_1 z + p_2 \frac{z^2}{2!} + \ldots + p_k \frac{z^k}{k!} + \ldots$$

E.g. 2d6: toss a six-sided die twice, find the probability that the sum is 4?

- Ω(6-sided die) $= \{1, 2, 3, 4, 5, 6\}$ with probability $\{p_1, p_2, p_3, p_4, p_5, p_6\}$, the GF of 1d6 is

$$G_{1d6}(x) = p_1 x + p_2 x^2 + p_3 x^3 + p_4 x^4 + p_5 x^5 + p_6 x^6.$$

  A die cannot give outcome 0 so there is no $1x^0$ in $G_{1d6}(x)$

- The GF corresponds to all possible outcome of $2d6$ is $G_{2d6}(x) = G_{1d6}(x) \cdot G_{1d6}(x)$

$$G_{2d6}(x) = G_{1d6}(x) \cdot G_{1d6}(x) = p_1 p_1 x^2 + (p1 p_2 + p_2 p_1) x^3 + (p_1 p_3 + p_2 p_2 + p_3 p_1) x^4 + \cdots$$

- Recall in a polynomial of x, the notation $[x^n]$ refers to the coefficient of x^n in the polynomial.

- $\mathbb{P}($ sum is 4 $) = [x^4] G_{2d6} = p_1 p_3 + p_2 p_2 + p_3 p_1$.

- If the die is fair, $p_i = \dfrac{1}{6}$, then the probability is $\dfrac{3}{36} = \dfrac{1}{12}$.

E.g. (source) 3d6: toss a fair six-sided die thrice, what is the probability that the sum is 13?

- $G_{1d6}(x) = \frac{1}{6}x + \frac{1}{6}x^2 + \frac{1}{6}x^3 + \frac{1}{6}x^4 + \frac{1}{6}x^5 + \frac{1}{6}x^6$. The GF $G_{3d6} = G_{1d6}^3(x)$. The answer is $[x^{13}]G_{1d6}^3(x)$, i.e.,

$$[x^{13}]\frac{(x + x^2 + x^3 + ... + x^6)^3}{6^3} = \frac{1}{6^3}[x^{13}](x + x^2 + ... + x^5 + x^6)^3 = \frac{1}{6^3}[x^{10}](1 + x + ... + x^5)^3.$$

- So we look for $\frac{1}{6^3}[x^{10}](1 + x + ... + x^5)^3$.

$$\begin{aligned}
[x^{10}](1 + x + ... + x^5)^3 &= [x^{10}]\left(\frac{1 - x^6}{1 - x}\right)^3 && \text{geometric sum} \\
&= [x^{10}]\left(1 - x^6\right)^3\left(\frac{1}{1 - x}\right)^3 \\
&= [x^{10}]\sum_{k=0}^{3}\binom{3}{k}(-x^6)^k 1^{3-k}\left(1 + x + x^2 + ...\right)^3 && \text{binomial theorem, geometric series} \\
&= [x^{10}]\sum_{k=0}^{3}\binom{3}{k}((-1)x^6)^k\sum_{r=0}^{\infty}\binom{r + 3 - 1}{r}x^r && \text{expansion of geometric series} \\
&= [x^{10}]\sum_{k=0}^{3}\binom{3}{k}(-1)^k x^{6k}\sum_{r=0}^{\infty}\binom{2 + r}{r}x^r \\
&= [x^{10}]\sum_{k=0}^{3}\binom{3}{k}(-1)^k x^{6k}\sum_{r=0}^{\infty}\binom{2 + r}{2}x^r && \binom{n}{k} = \binom{n}{n - k}
\end{aligned}$$

Combine the x term gives

$$[x^{10}](1+x+...+x^5)^3 \;=\; [x^{10}]\sum_{k=0}^{3}\binom{3}{k}(-1)^k\sum_{r=0}^{\infty}\binom{2+r}{2}x^{6k+r}.$$

We look for coefficient of x^{10}, let $10 =: s = 6k + r$ so $r = s - 6k$, and

$$[x^s]\sum_{k=0}^{3}\binom{3}{k}(-1)^k\sum_{s-6k=0}^{\infty}\binom{2+s-6k}{2}x^s \quad \overset{s=10}{=} \quad [x^{10}]\sum_{k=0}^{3}\binom{3}{k}(-1)^k\sum_{10-6k=0}^{\infty}\binom{12-6k}{2}x^{10}.$$

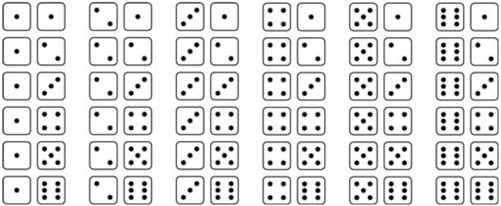
$\binom{12-6k}{2}$ is nonzero only for $k = 0, 1$, hence

$$\binom{3}{0}(-1)^0\binom{12-6(0)}{2}+\binom{3}{1}(-1)^1\binom{12-6(1)}{2} \;=\; 1\cdot\binom{12}{2}-3\cdot\binom{6}{2} \;=\; 21.$$

The probability is $\frac{21}{6^3} = \frac{21}{216} \approx 0.1$.

Six-sided die

- All the possible outcome of 2d6 (toss a six-sided die twice)



$$
\begin{array}{cccccc}
x^1x^1 & x^2x^1 & x^3x^1 & x^4x^1 & x^5x^1 & x^6x^1 \\
x^1x^2 & x^2x^2 & x^3x^2 & x^4x^2 & x^5x^2 & x^6x^2 \\
x^1x^3 & x^2x^3 & x^3x^3 & x^4x^3 & x^5x^3 & x^6x^3 \\
x^1x^4 & x^2x^4 & x^3x^4 & x^4x^4 & x^5x^4 & x^6x^4 \\
x^1x^5 & x^2x^5 & x^3x^5 & x^4x^5 & x^5x^5 & x^6x^5 \\
x^1x^6 & x^2x^6 & x^3x^6 & x^4x^6 & x^5x^6 & x^6x^6
\end{array}
\Longrightarrow
\begin{array}{cccccc}
x^2 & x^3 & x^4 & x^5 & x^6 & x^7 \\
x^3 & x^4 & x^5 & x^6 & x^7 & x^8 \\
x^4 & x^5 & x^6 & x^7 & x^8 & x^9 \\
x^5 & x^6 & x^7 & x^8 & x^9 & x^{10} \\
x^6 & x^7 & x^8 & x^9 & x^{10} & x^{11} \\
x^7 & x^8 & x^9 & x^{10} & x^{11} & x^{12}
\end{array}
$$

$$1x^2 + 2x^3 + 3x^4 + 4x^5 + 5x^6 + 6x^7 + 5x^8 + 4x^9 + 3x^{10} + 2x^{11} + 1x^{12}$$

Six-sided die

- By product rule: 3d6 = 2d65 $\times$ 1d6 All the possible outcome

$$(1x^2 + 2x^3 + 3x^4 + 4x^5 + 5x^6 + 6x^7 + 5x^8 + 4x^9 + 3x^{10} + 2x^{11} + 1x^{12}) \times (x^1 + x^2 + x^3 + x^4 + x^5 + x^6)$$

$$= \begin{matrix} x^3 & x^4 & x^5 & x^6 & x^7 & x^8 \\ 2x^4 & 2x^5 & 2x^6 & 2x^7 & 2x^8 & 2x^9 \\ 3x^5 & 3x^6 & 3x^7 & 3x^8 & 3x^9 & 3x^{10} \\ 4x^6 & 4x^7 & 4x^8 & 4x^9 & 4x^{10} & 4x^{11} \\ 5x^7 & 5x^8 & 5x^9 & 5x^{10} & 5x^{11} & 5x^{12} \\ 6x^8 & 6x^9 & 6x^{10} & 6x^{11} & 6x^{12} & 6x^{13} \\ 5x^9 & 5x^{10} & 5x^{11} & 5x^{12} & 5x^{13} & 5x^{14} \\ 4x^{10} & 4x^{11} & 4x^{12} & 4x^{13} & 4x^{14} & 4x^{15} \\ 3x^{11} & 3x^{12} & 3x^{13} & 3x^{14} & 3x^{15} & 3x^{16} \\ 2x^{12} & 2x^{13} & 2x^{14} & 2x^{15} & 2x^{16} & 2x^{17} \\ x^{13} & x^{14} & x^{15} & x^{16} & x^{17} & x^{18} \end{matrix}$$

$$[x^{13}] = 6 + 5 + 4 + 3 + 2 + 1 = 21$$

21 out of the 6^3 possible ways $= \dfrac{21}{6^3}$

Different die

- 1d4 and 1d6: you toss a 4-sided die and a 6-side die
  What is the probability that the sum is 5?

- Ans: $[x^5](x^1 + x^2 + x^3 + x^4)(x^1 + x^2 + x^3 + x^4 + x^5 + x^6)$

$$\begin{array}{cccc} x^1x^1 & x^2x^1 & x^3x^1 & x^4x^1 \\ x^1x^2 & x^2x^2 & x^3x^2 & x^4x^2 \\ x^1x^3 & x^2x^3 & x^3x^3 & x^4x^3 \\ x^1x^4 & x^2x^4 & x^3x^4 & x^4x^4 \\ x^1x^5 & x^2x^5 & x^3x^5 & x^4x^5 \\ x^1x^6 & x^2x^6 & x^3x^6 & x^4x^6 \end{array} \Longrightarrow \begin{array}{cccc} x^2 & x^3 & x^4 & x^5 \\ x^3 & x^4 & x^5 & x^6 \\ x^4 & x^5 & x^6 & x^7 \\ x^5 & x^6 & x^7 & x^8 \\ x^6 & x^7 & x^8 & x^9 \\ x^7 & x^8 & x^9 & x^{10} \end{array}$$

$$(x^1 + x^2 + x^3 + x^4)(x^1 + x^2 + x^3 + x^4 + x^5 + x^5) = x^2 + 2x^3 + 3x^4 + 4x^5 + 4x^6 + 4x^7 + 3x^8 + 2x^9 + x^{10}$$

$$[x^5](x^1 + x^2 + x^3 + x^4)(x^1 + x^2 + x^3 + x^4 + x^5 + x^6) = 4$$

4 out of the $|1d4| \cdot |1d6|$ possible ways $= \dfrac{4}{4 \cdot 6} = \dfrac{1}{6}$

Coin and die

- You toss a coin and a 4-sided die
  If the coin gives 0 (tail), we take value of zero
  If the coin gives 1 (head), we take the value of the 4-side die
  What is the probability you get a value 3?
- By brute force $\Omega = \Big\{(0,1),(0,2),(0,3),(0,4),(1,1),(1,2),(1,3),(1,4)\Big\}, \quad |\Omega| = 8$
  $E = \{(1,3)\}, \quad |E| = 1 \qquad \mathbb{P}(E) = \dfrac{1}{8}$
- By product rule $\underbrace{1/2}_{\text{chance of getting 1 in coin}} \times \underbrace{1/4}_{\text{chance of getting 3 in die}} = 1/8.$
- By generating function (advanced)
  - Coin $C(x) = \dfrac{1}{2} + \dfrac{1}{2}x$
  - Die $D(x) = \dfrac{1}{4}x + \dfrac{1}{4}x^2 + \dfrac{1}{4}x^3 + \dfrac{1}{4}x^4$
  - The value depends on the coin toss: $G(x) = C\Big(D(x)\Big) = \dfrac{1}{2} + \dfrac{1}{2}D(x) = \dfrac{1}{2} + \dfrac{1}{8}x + \dfrac{1}{8}x^2 + \dfrac{1}{8}x^3 + \dfrac{1}{8}x^4$
  - Answer is $[x^3]G(x) = \dfrac{1}{8}$
  - Dependent random variables $\iff$ generating function composition

Probability is just combinatorics

- You flip a fair coin 5 times. Find the probability of getting **exactly** 3 heads.

$$\#\text{ways get 3 head} = \binom{5}{3} = 10$$

$$|\Omega| = 2^5 = 32$$

$$\mathbb{P}(\text{toss 5 get 3 heads}) = \frac{\binom{5}{3}}{2^5} = \frac{10}{32}$$

- Find the probability of getting **at least** 3 heads

$$\begin{aligned}\mathbb{P}(\text{toss 5 get} \geq 3 \text{ heads}) &= \mathbb{P}\Big((\text{toss 5 get 3 heads}) \text{ OR } (\text{toss 5 get 4 heads}) \text{ OR } (\text{toss 5 get 5 heads})\Big) \\ &= \frac{\binom{5}{3} + \binom{5}{4} + \binom{5}{5}}{2^5} = \frac{16}{32}\end{aligned}$$

- Find the probability of getting even number of heads

$$\begin{aligned}\mathbb{P}(\text{even number of heads}) &= \mathbb{P}\Big((\text{toss 5 get 2 heads}) \text{ OR } (\text{toss 5 get 4 heads})\Big) \\ &= \frac{\binom{5}{2} + \binom{5}{4}}{2^5}\end{aligned}$$

Football match is a trinomial

- Football match has 3 outcome: win (W), lose (L) and draw (D)
- Suppose Manchester City F.C. has a constant win chance 0.5, lose change 0.2 and a draw change 0.3, regardless of what team they play against.
- Now Manchester City F.C. plays 20 games.
- Find the probability of getting 10 W, 4 L and 6D.

$$\mathbb{P}(W = 10, L = 4, D = 6) = \binom{20}{10, 4, 6} 0.5^{10} 0.2^4 0.3^6 = \frac{20!}{10!4!6!} 0.5^{10} 0.2^4 0.3^6 = 0.044.$$

- Find the probability of getting at least 19 W

$$\mathbb{P}(19, 1, 0) + \mathbb{P}(19, 0, 1) + \mathbb{P}(20, 0, 0) = \binom{20}{19, 1, 0} 0.5^{19} 0.2^1 0.3^0 + \binom{20}{19, 0, 1} 0.5^{19} 0.2^0 0.3^1 + \binom{20}{20, 0, 0} 0.5^{20} 0.2^0 0.3^0$$

- Find the probability of getting at least 15 W

W	15	15	15	15	15	15	16	16		20
L	5	4	3	2	1	0	4	3	$\cdots$	0
D	0	1	2	3	4	5	0	1		0

Gene is a quadrinomial — Not in exam

- Human genome has four type: A, T, C, G

$$\binom{n}{n_A, n_T, n_C, n_G} p_A^{n_A} p_T^{n_T} p_C^{n_C} p_G^{n_G} = \frac{n!}{n_A! n_T! n_C! n_G!} p_A^{n_A} p_T^{n_T} p_C^{n_C} p_G^{n_G}$$

- Human has $n = 20000$ genes

$$\binom{20000}{n_A, n_T, n_C, n_G} p_A^{n_A} p_T^{n_T} p_C^{n_C} p_G^{n_G} = \frac{20000!}{n_A! n_T! n_C! n_G!} p_A^{n_A} p_T^{n_T} p_C^{n_C} p_G^{n_G}$$

- Suppose X-men is possible and has a specific gene

$$\cdots CTACGTGCCCGCCGAGGAG \cdots$$

  The chance you are X-men

$$\mathbb{P}\Big(\text{your gene has the same string as X-men gene}\Big)$$

- Actually this is how you calculate $\mathbb{P}(\text{you get cancer})$

Contents

Random variable (RV)

- Let $\Omega = \{1, 2, 3\}$, let X be a random variable over Ω with $X = \begin{cases} 1 & \text{with probability } 1/2 \\ 2 & \text{with probability } 1/4 \\ 3 & \text{with probability } 1/4 \end{cases}$

- A *realisation* is a particular value from Ω drawn at random
  For example, a 22 sample realisation

$$3, 3, 1, 3, 2, 1, 1, 1, 2, 3, 3, 2, 1, 3, 3, 2, 1, 2, 1, 2, 1, 1$$

  There are nine 1s, six 2s and seven 3s. We expect 1s to appear more frequently the more realisations we take

- RV notation $\mathbb{P}(X = x),\ x \in \Omega$
  It means "the probability of random variable X takes the value x in the space Ω"

- For X we have

$$\mathbb{P}(X = 1) = 1/2, \qquad \mathbb{P}(X = 2) = 1/4, \qquad \mathbb{P}(X = 3) = 1/4,$$

  What about $\mathbb{P}(X = 5)$? Zero or undefined.

Random variable and event

- $X = x$ and E are the same thing: $X = x$ can be seen as "an event that X takes the value x"
- Recall the probability axioms, we have

$$
\begin{aligned}
\mathbb{P}(E) &\geq 0 \quad &\iff \quad \mathbb{P}(X = x) &\geq 0 && \text{(Axiom 1)} \\
\mathbb{P}(\Omega) &\equiv 1 \quad &\iff \quad \sum_{x \in \Omega} \mathbb{P}(X = x) &= 1. && \text{(Axiom 2)}
\end{aligned}
$$

$$
\mathbb{P}\Big(\bigcup_i E_i\Big) \overset{E_i \text{ disjoint}}{=} \sum_i \mathbb{P}(E_i) \quad \iff \quad \mathbb{P}\Big(X \in \bigcup_i A_i\Big) \overset{A_i \text{ disjoint}}{=} \sum_i \mathbb{P}(X \in A_i). \qquad \text{(Axiom 3)}
$$

- **E.g.**

$$
X = \begin{cases} 1 & \text{with probability } 1/2 \\ 2 & \text{with probability } 1/4 \\ 3 & \text{with probability } 1/4 \end{cases} \iff \mathbb{P}(X = 1) = 1/2, \quad \mathbb{P}(X = 2) = 1/4, \quad \mathbb{P}(X = 3) = 1/4,
$$

- Then $\mathbb{P}(X \geq 2)$ is

$$
\begin{aligned}
\mathbb{P}(X \in \{2\} \cup \{3\}) &\overset{\text{Axiom 3}}{=} \mathbb{P}(X \in \{2\}) + \mathbb{P}(X \in \{3\}) \\
&= \mathbb{P}(X = 2) + \mathbb{P}(X = 3) \\
&= 1/4 + 1/4 \\
&= 1/2
\end{aligned}
$$

Example: Tossing a fair coin thrice

- Toss a fair coin thrice.
  Let X be the r.v. of the number of heads obtained, find $\mathbb{P}(X=2)$ and $\mathbb{P}(X<2)$, are the events $(X=2)$ and $(X<2)$ complementary? mutually exclusive?

- Answer: let $\Omega = \{HHH, HHT, HTH, THH, HTT, THT, TTH, TTT\}$ and $E \subset \Omega$ be the event of $X = 2$.

$$\mathbb{P}(E) = \mathbb{P}(X=2) = \frac{\big|\{HHT, HTH, THH\}\big|}{|\Omega|} = \frac{3}{8}$$

  Let F be the event of $(X < 2)$

$$\mathbb{P}(F) = \mathbb{P}(X<2) = \frac{\big|\{HTT, THT, TTH, TTT\}\big|}{|\Omega|} = \frac{4}{8} = \frac{1}{2}$$

- E, F are mutually exclusive since $E \cap F = \varnothing$

- E, F are not complementary ($F \neq E^c$) because $\mathbb{P}(F) = \frac{1}{2} \neq \frac{5}{8} = \mathbb{P}(E^c) = 1 - \mathbb{P}(E)$

Contents

Bi-variate / two random variables

- Let $\mathcal{X} = \{1, 2, 3\}, \mathcal{Y} = \{1, 2\}$ be the sample spaces of two RVs $X \in \mathcal{X}$ and $Y \in \mathcal{Y}$.

- The Cartesian product $\mathcal{X} \times \mathcal{Y}$ is the sample space Ω for the pair (i, j)

$$\mathcal{X} \times \mathcal{Y} \;=\; \left\{ \begin{array}{ccc} (1,1), & (2,1), & (3,1), \\ (1,2), & (2,2), & (3,2) \end{array} \right\}$$

- An example of distribution over $\Omega = \mathcal{X} \times \mathcal{Y}$

	X=1	X=2	X=3
Y=1	0.05	0.15	0.1
Y=2	0.25	0.15	0.3

Hence $\mathbb{P}(X = 1, Y = 1) = 0.05$ and $\mathbb{P}(X = 3, Y = 2) = 0.3$.

- **Definition** $\mathbb{P}(X = x, Y = y)$ is called the *joint probability* of $X = x$ and $Y = y$.

Example of joint probability

	Wearing glasses (G)	Not wearing glasses (N)
Wear hat (H)	0.05	0.15
Not wearing hat (N)	0.45	0.35

- $\mathcal{X} = \{$wearing glasses, not wearing glasses$\}$
- $\mathcal{Y} = \{$wearing hat, not wearing hat$\}$

$$\mathcal{X} \times \mathcal{Y} = \Big\{(G, H), (G, N), (N, H), (N, N)\Big\}$$

- $\mathbb{P}(X = G, Y = N) = 0.45$
- Axiom of probability has to hold, so
  - $\mathbb{P}(X = x, Y = y) \geq 0$ axiom 1
  - $\sum_{x \in \mathcal{X}, y \in \mathcal{Y}} \mathbb{P}(X = x, Y = y) = 1$ axiom 2
  - $\mathbb{P}\Big(X \in \bigcup_i A_i, Y \in \bigcup_j B_j\Big) \overset{\text{if disjoint}}{=} \sum_{x \in \mathcal{X}, y \in \mathcal{Y}} \mathbb{P}(X \in A_i, Y \in B_j)$ axiom 3

Marginal probability

	Wearing glasses (G)	Not wearing glasses (N)
Wear hat (H)	0.05	0.15
Not wearing hat (N)	0.45	0.35

- $\mathbb{P}(X = x) = \sum_{y \in \mathcal{Y}} \mathbb{P}(X = x, Y = y)$ means only looking at $X = x$ regardless of Y

- $\mathbb{P}(\text{wearing glasses}) = \mathbb{P}(X = G) = 0.5 = \mathbb{P}(X = G, Y = H) + \mathbb{P}(X = G, Y = N)$

- $\mathbb{P}(\text{not wearing hat}) = \mathbb{P}(Y = N) = 0.8 = \mathbb{P}(X = G, Y = N) + \mathbb{P}(X = N, Y = N)$

- **Definition** $\mathbb{P}(X = x) = \sum_{y \in \mathcal{Y}} \mathbb{P}(X = x, Y = y)$ is called marginal probability

Joint probability and marginal probability table

Input table

	$X = x_1$	$X = x_2$	$\cdots$	$X = x_N$	
$Y = y_1$	$\mathbb{P}(X = x_1, Y = y_1)$	$\mathbb{P}(X = x_2, Y = y_1)$	$\cdots$	$\mathbb{P}(X = x_N, Y = y_1)$	
$Y = y_2$	$\mathbb{P}(X = x_1, Y = y_2)$	$\mathbb{P}(X = x_2, Y = y_2)$	$\cdots$	$\mathbb{P}(X = x_N, Y = y_2)$	
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
$Y = y_M$	$\mathbb{P}(X = x_1, Y = y_M)$	$\mathbb{P}(X = x_2, Y = y_M)$	$\cdots$	$\mathbb{P}(X = x_N, Y = y_M)$	

Augmented table

	$X = x_1$	$X = x_2$	$\cdots$	$X = x_N$	
$Y = y_1$	$\mathbb{P}(X = x_1, Y = y_1)$	$\mathbb{P}(X = x_2, Y = y_1)$	$\cdots$	$\mathbb{P}(X = x_N, Y = y_1)$	$\mathbb{P}(Y = y_1)$
$Y = y_2$	$\mathbb{P}(X = x_1, Y = y_2)$	$\mathbb{P}(X = x_2, Y = y_2)$	$\cdots$	$\mathbb{P}(X = x_N, Y = y_2)$	$\mathbb{P}(Y = y_2)$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
$Y = y_M$	$\mathbb{P}(X = x_1, Y = y_M)$	$\mathbb{P}(X = x_2, Y = y_M)$	$\cdots$	$\mathbb{P}(X = x_N, Y = y_M)$	$\mathbb{P}(Y = y_M)$
	$\mathbb{P}(X = x_1)$	$\mathbb{P}(X = x_2)$	$\cdots$	$\mathbb{P}(X = x_N)$	

Conditional probability

- **Definition** $\mathbb{P}(X = x \,|\, Y = y)$ is called conditional probability, meaning the probability of $X = x$ conditional on $Y = y$, defined as

$$\mathbb{P}(X = x \,|\, Y = y) \;=\; \frac{\mathbb{P}(X = x, Y = y)}{\mathbb{P}(Y = y)} \;=\; \frac{\text{joint on } X, Y}{\text{marginal on } Y}$$

- Example

	X=1	X=2	X=3
Y=1	0.05	0.15	0.1
Y=2	0.25	0.15	0.3

$$\mathbb{P}(X = 1 \,|\, Y = 1) \;=\; \frac{\mathbb{P}(X = 1, Y = 1)}{\mathbb{P}(Y = 1)} \;=\; \frac{0.05}{0.3} \;\approx\; 0.1667$$

$$\mathbb{P}(X = 1 \,|\, Y = 2) \;=\; \frac{\mathbb{P}(X = 1, Y = 2)}{\mathbb{P}(Y = 2)} \;=\; \frac{0.25}{0.7}$$

- Can we have $\mathbb{P}(Y = y) = 0$? No.

Independent random variables

- **Definition** X, Y are independent if

$$\mathbb{P}(X = x\,,\, Y = y) \;=\; \mathbb{P}(X = x)\mathbb{P}(Y = y) \qquad \forall x \in \mathcal{X}, \forall y \in \mathcal{Y}$$

- This implies conditional = marginal

$$\begin{aligned} \mathbb{P}(X = x \,|\, Y = y) &= \frac{\mathbb{P}(X = x, Y = y)}{\mathbb{P}(Y = y)} = \frac{\text{joint on } X, Y}{\text{marginal on } Y} \\ &= \frac{\mathbb{P}(X = x)\mathbb{P}(Y = y)}{\mathbb{P}(Y = y)} \\ &= \mathbb{P}(X = x) \end{aligned}$$

- Information on Y tells nothing about X

i.i.d. (independent and identically distributed)

- **Definition** X, Y are i.i.d. random variables mean they are independent and identically distributed, i.e.,

$$\begin{aligned}\mathbb{P}(X = x, Y = y) &= \mathbb{P}(X = x)\mathbb{P}(Y = y) \quad \forall x \in \mathcal{X}, y \in \mathcal{Y} \\ \mathbb{P}(X = x) &= \mathbb{P}(Y = x) \quad \forall x \in \mathcal{X}\end{aligned}$$

- **Definition** $X_1, X_2, X_3, \ldots$ are independent and identically distributed random variable if all of them are mutually independent and

$$\mathbb{P}(X_1 = x) = \mathbb{P}(X_2 = x) = \mathbb{P}(X_3 = x) = \cdots \quad \forall x \in \mathcal{X}$$

- **E.g.** (10d6) toss one six-sided die 10 times
  - Independent: the outcome of the die will not affect other, all the 10 results are independent from each other

$$\mathbb{P}(X_1 = x_1, X_2 = x_2, ..., X_{10} = x_{10}) = \mathbb{P}(X_1 = x_1)\mathbb{P}(X_2 = x_2)...\mathbb{P}(X_{10} = x_{10})$$

  - Identically distributed: same six-sided die

$$\mathbb{P}(X_1 = x) = \mathbb{P}(X_2 = x) = \mathbb{P}(X_{10} = x)$$

  - Hence if I am looking for the probability of rolling 10 six

$$\mathbb{P}(X_1 = 6, X_2 = 6, ..., X_{10} = 6) = \mathbb{P}(X_1 = 6)\mathbb{P}(X_2 = 6)...\mathbb{P}(X_{10} = 6) = \Big(\mathbb{P}(X = 6)\Big)^{10}$$

  - If the die is fair $\mathbb{P}(X = 6) = \frac{1}{6}$, then the chace of rollinig 10 six is $\frac{1}{6^{10}}$

Bayes' theorem (Not in exam)

- Conditional probability $\mathbb{P}(S=s|T=t) = \frac{\mathbb{P}(S=s,T=t)}{\mathbb{P}(T=t)} \iff \text{Conditional} = \frac{\text{Joint}}{\text{Magrinal}}$

$$
\begin{aligned}
\text{Conditional} = \frac{\text{Joint}}{\text{Magrinal}} &\iff \text{Conditional} \cdot \text{Magrinal} = \text{Joint} \\
&\iff \text{Joint} = \boxed{\text{Conditional} \cdot \text{Magrinal}} \\
&\iff \mathbb{P}(S=s,T=t) = \mathbb{P}(T=t,S=s) \\
&\iff \mathbb{P}(T=t,S=s) = \boxed{\mathbb{P}(T=t|S=s)\mathbb{P}(S=s)}
\end{aligned}
$$

- Now we have

$$\mathbb{P}(S=s|T=t) = \frac{\mathbb{P}(S=s,T=t)}{\mathbb{P}(T=t)} = \frac{\boxed{\mathbb{P}(T=t|S=s)\mathbb{P}(S=s)}}{\mathbb{P}(T=t)}$$

i.e.,

$$\mathbb{P}(S=s|T=t) = \frac{\mathbb{P}(T=t|S=s)\mathbb{P}(S=s)}{\mathbb{P}(T=t)}$$

Football example: sport analytic

- In sport, teams play at their own venue ("at home") and at other team's venues ("away").

- Consider the home and away performance for the team Southampton.
  The information regarding the total number of home $(H = 1)$, away $(H = 0)$, wins $(R = 2)$, draws $(R = 1)$ and losses $(R = 0)$ for the 20XX seasons is:
  - 12 home games won
  - 2 home games drawn
  - 5 home games lost
  - 9 away games won
  - 8 away games drawn
  - 2 away games lost

- First we construct the table

	Lose $R = 0$	Draw $R = 1$	Win $R = 2$
away $H = 0$	2	8	9
home $H = 1$	5	2	12

Football example: sport analytic

	Lose $R=0$	Draw $R=1$	Win $R=2$
away $H=0$	2	8	9
home $H=1$	5	2	12

- What is the marginal probability of Southampton will win a game, regardless of whether it is played at home or away?

$$\mathbb{P}(R=2) \;=\; \frac{9+12}{2+5+8+2+9+12} \;=\; \frac{21}{38}.$$

- What is the conditional probability of Southampton will win a game, given that they are playing at home?

$$\mathbb{P}(R=2|H=1) \;=\; \frac{\mathbb{P}(R=2, H=1)}{\mathbb{P}(H=1)} \;=\; \frac{\frac{12}{2+8+9+5+2+12}}{\frac{5+2+12}{2+8+9+5+2+12}} \;=\; \frac{\frac{12}{38}}{\frac{19}{38}} \;=\; \frac{12}{19}.$$

- What is the conditional probability of Southampton will win a game, given that they are playing away?

$$\mathbb{P}(R=2|H=0) = 1 - \mathbb{P}(R=2|H=1).$$

- Do you believe that Southampton is more likely to win when at home versus when they play away?

$$\mathbb{P}(R=2|H=1) > \mathbb{P}(R=2|H=0)$$

Football example: sport analytic – not lose two out of three games ... 1/2

	Lose $R = 0$	Draw $R = 1$	Win $R = 2$
away $H = 0$	2	8	9
home $H = 1$	5	2	12

Suppose Southampton will play an away game, then a home game, and then an away game in their next three games. What is the probability that they will not lose two out of three of these games?

First we simplify:

$$\{\text{NOT lose}\} = \{\text{win}\} \text{ OR } \{\text{draw}\}$$

Then we have the table

	Lose $R = 0$	Not lose $R \neq 0$
away $H = 0$	2	17
home $H = 1$	5	14

The numbers in the table are not probability (Probability Axiom 1: $\mathbb{P}(\Omega) = 1$), we need to normalize them

	Lose $R = 0$	Not lose $R \neq 0$
away $H = 0$	2/38	17/38
home $H = 1$	5/38	14/38

Now we see that the numbers in the table sum to 1, so Probability Axiom 1 is true.

Football example: sport analytic – not lose two out of three games ... 2/2

	Lose $R=0$	Not lose $R\neq 0$
away $H=0$	$2/38$	$17/38$
home $H=1$	$5/38$	$14/38$

Suppose Southampton will play an away game, then a home game, and then an away game in their next three games. What is the probability that they will not lose two out of three of these games?
All the 8 possibilities of the 3 games

$$\Big\{\underbrace{LLL}_{\text{3 loses}},\ \underbrace{LLN,\ LNL,\ NLL}_{\text{2 lose 1 not lose}},\ \underbrace{LNN,\ NLN,\ NNL}_{\text{1 lose 2 not lose}},\ \underbrace{NNN}_{\text{3 not lose}}\Big\}$$

Then

$$\text{NOT}\{\text{2 loses}\} = \Big\{\underbrace{LLL}_{\text{3 loses}},\ \underbrace{LNN,\ NLN,\ NNL}_{\text{1 lose 2 not lose}},\ \underbrace{NNN}_{\text{3 not lose}}\Big\}$$

$$\mathbb{P}\Big(\text{NOT}\{\text{2 loses}\}\Big) \overset{\text{sum rule}}{=} \frac{2}{38}\frac{5}{38}\frac{2}{38}+\frac{2}{38}\frac{14}{38}\frac{17}{38}+\frac{17}{38}\frac{5}{38}\frac{17}{38}+\frac{17}{38}\frac{14}{38}\frac{2}{38}+\frac{17}{38}\frac{14}{38}\frac{17}{38}$$

$$= \frac{(2)(5)(2)+(2)(14)(17)+(17)(5)(17)+(17)(14)(2)+(17)(14)(17)}{38^3}$$

$$\approx 11\%$$

Section summary

- $\mathbb{P}(X = x, Y = y)$ Joint probability

- $\mathbb{P}(X = x) = \sum_{y \in \mathcal{Y}} \mathbb{P}(X = x, Y = y)$ Marginal probability

- $\mathbb{P}(X = x|Y = y) = \frac{\mathbb{P}(X = x, Y = y)}{\mathbb{P}(Y = y)}, \quad \mathbb{P}(Y = y) > 0$ Conditional probability

- Conditional $= \frac{\text{Joint}}{\text{Magrinal}}$

- Their calculation / operation

False positive / false alarm and false negative

- False positive / false alarm

$$\mathbb{P}\Big(\text{ diagnosed pregnant} \mid \text{ not pregnant}\Big)$$

- 1983 Soviet nuclear false alarm incident
- ChatGPT makeup bullshit
- Issue of false promise

- False negative

$$\mathbb{P}\Big(\text{ diagnosed not pregnant} \mid \text{ pregnant}\Big)$$

- False negative can be more dangerious
  "You have cancer but diagnosed no cancer"
  vs
  "You have no cancer but diagnosed with cancer"

About your future

		$H = 0$ (not study hard)	$H = 1$ (study hard)
Your future	$F = 0$ (bad future)	$\mathbb{P}(F = 0 \vert H = 0)$	$\mathbb{P}(F = 0 \vert H = 1)$
	$F = 1$ (good future)	$\mathbb{P}(F = 1 \vert H = 0)$	$\mathbb{P}(F = 1 \vert H = 1)$

- Common sense: $\mathbb{P}(F = 1|H = 0)$ is low.
- Common sense: $\mathbb{P}(F = 1|H = 1)$ is NOT 1 but statistically high.
- What is life

$$\mathbb{P}\Big(\text{Tomorrow} \Big| \ (\text{Yesterday} \mid \text{two days ago} \,)\Big)$$

Contents

Descriptive statistics

Distribution	Measure of centrality	Measure of spread	Measure of symmetry	Measure of tailedness
	mean (average)	range	skewness	kurtosis
	median (robust average)	**variance**		
	mode (minmax average)	standard deviation		
		interquartile range		

- **What's the point of statistics**: how do you know a bag of 1kg rice is good quality?
  - check each grain one by one — but you have to check 30000 grains
  - check 20 grains and use these 20 grains to summarize the bag — this is statistics

- Issues of statistics
  - Is statistics absolutely correct?
  - Issue of outlier / robust statistics
  - Issue of imbalanced Data
  - Misuse of statistics
  - Reliability of statistics: Anscombe's quartet

Pick one

Option A

50% chance you win 1 million, 50% chance you lose 1 million, only allowed to gamble once

Option B

50% chance you win $\frac{1}{100}$ million, 50% chance you lose $\frac{1}{100}$ million, allowed to gamble 100 times

Probability Distribution function

- Writing $\mathbb{P}(X = x)$ is too clumsy, just write $p(x) := \mathbb{P}(X = x)$

- **Definition** $p(x)$ is called a *probability distribution function*
  - **Definition** $p(x)$ is called a *probability density function* if X is a continuous random variable
  - **Definition** $p(x)$ is called a *probability mass function* if X is a discrete random variable

- Similarly, we write
  - $p(x, y) = \mathbb{P}(X = x, Y = y)$
  - $p(x \,|\, y) = \mathbb{P}(X = x \,|\, Y = y)$
  - $p(x \,|\, y) = \dfrac{p(x, y)}{p(y)}, \; p(y) > 0$

Example of discrete probability distribution = probability mass function

- **E.g.** The probability mass function (PMF) of a discrete random variable X is
$\mathbb{P}(X=x) = \begin{cases} \frac{1}{12} & x \in \{1,2,\dots,12\} \\ 0 & \text{else} \end{cases}$, find $\mathbb{P}(X+2 < 3X-4 \leq 2X+7)$

- **Solution** First we work on simplifying the expression

$$
\begin{aligned}
& \mathbb{P}(X+2 < 3X-4 \leq 2X+7) \\
= & \mathbb{P}(X+2-X < 3X-4-X \leq 2X+7-X) \\
= & \mathbb{P}(2 < 2X-4 \leq X+7) \\
= & \mathbb{P}(2+4 < 2X-4+4 \leq X+7+4) \\
= & \mathbb{P}(6 < 2X \leq X+11) \\
= & \mathbb{P}(3 < X \leq 11) \\
= & \mathbb{P}(X \in \{4,5,\dots,11\}) \\
= & \frac{8}{12} = \frac{2}{3}
\end{aligned}
$$

$6 < 2X \leq X+11$ eq. to $6 < 2X$ AND $2X \leq X+11$, eq. to $3 < X$ AND $X \leq 11$

- This PMF is known as uniform distribution (later)

Mean / Expected Value

- **Definition** Given a distribution $p(x) = \mathbb{P}(X = x)$, we define the expected value of the RV X as

$$\mathbb{E}[X] = \begin{cases} \sum_{x \in \mathcal{X}} xp(x) & \text{discrete RV} \\ \int_{x \in \mathcal{X}} xp(x)dx & \text{continuous RV} \end{cases}$$

- Example $\mathbb{P}(X = 1) = 0.5, \mathbb{P}(X = 2) = 0.4, \mathbb{P}(X = 3) = 0.1$

$$\mathbb{E}[X] \;=\; 1 \cdot 0.5 + 2 \cdot 0.4 + 3 \cdot 0.1 \;=\; 1.6$$

- Example. $\mathbb{P}(X = x) = p(x|\mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\Big(-\frac{(x-\mu)^2}{2\sigma^2}\Big)$, then $\mathbb{E}[X] = \mu$, the key in the proof

$$\int_{-\infty}^{\infty} e^{-x^2} dx = \sqrt{\pi}$$

- **Average** is a special case of expected value

  Other measures of centrality: median, mode, geometric mean, harmonic mean

Example of discrete expected value ... 1/2

- **E.g.** Data from 100 epileptic people sampled at random in one year.

Number of seizures	number of people
0	34
2	21
4	18
6	11
8	16

- To find the *sample mean* (observed average), we first identity the sample space

$$\mathcal{X} = \{0, 2, 4, 6, 8\}.$$

  i.e., $x = 1$ is an impossible event.

- Then we construct the table

x	$p(x)$
0	34/100
2	21/100
4	18/100
6	11/100
8	16/100

$$\text{sample mean } \overline{x} = \sum_{x \in \mathcal{X}} x p(x) = 0 \cdot \frac{34}{100} + 2 \cdot \frac{21}{100} + \cdots + 8 \cdot \frac{16}{100} = 3.08$$

- **Very important**: sample mean $\neq$ expectation. We are using sample mean to **estimate** expectation. It is possible that sample mean is a **bad estimate** of expectation

Example of discrete expected value ... 2/2

$\mathcal{X} = \{0, 2, 4, 6, 8\}$.

x	$p(x)$
0	34/100
2	21/100
4	18/100
6	11/100
8	16/100

$\bar{x} = 3.08$

- **E.g.** What is the probability of selecting a person from this 100 people that the person has more than 3.08 seizures in one year?

$$\mathbb{P}(x \geq \bar{x}) = \frac{|x \in \{4, 6, 8\}|}{100} = \frac{18 + 11 + 16}{100} = 0.45.$$

- **E.g.** Find $\mathbb{P}\big(|x - \bar{x}| > 1\big)$

$$\mathbb{P}\big(|x - \bar{x}| > 1\big) = \frac{|x \in \{0, 2, 6, 8\}|}{100} = \frac{|x \in \mathcal{X} \setminus \{4\}|}{100} = 1 - \frac{18}{100} = 0.82$$

- **E.g.** Find $\mathbb{P}\big(|x - \bar{x}| < 2\big)$

$$\mathbb{P}\big(|x - \bar{x}| < 2\big) = \frac{|x \in \{2, 4\}|}{100} = 0.39$$

Expected value under transformation

$$\mathbb{E}\big[f(X)\big] \;=\; \begin{cases} \displaystyle\sum_{x\in\mathcal{X}} f(x)p(x) & \text{discrete RV} \\ \displaystyle\int_{x\in\mathcal{X}} f(x)p(x)dx & \text{continuous RV} \end{cases}$$

- Example $\mathbb{P}(X=1)=0.5, \mathbb{P}(X=2)=0.4, \mathbb{P}(X=3)=0.1$

$$\mathbb{E}\big[\ln(X)\big] \;=\; \ln(1)\cdot 0.5 + \ln(2)\cdot 0.4 + \ln(3)\cdot 0.1 \;=\; 0.3871$$

- $\mathbb{E}\big[\ln(X)\big]$ is used in *maximum likelihood estimator* (not in exam)

- **E.g.** Let X be the random variable of tossing a fair 4-sided die once, find $\mathbb{E}[X^2]$

$$\mathbb{E}\big[X^2\big] = \sum_{x\in\mathcal{X}=\{1,2,3,4\}} x^2 p(x) \;=\; (1)^2\cdot p(1) + (2)^2\cdot p(2) + (3)^2\cdot p(3) + (4)^2\cdot p(4)$$

$$= \frac{1^2+2^2+3^2+4^2}{4} = \frac{4(5)(9)}{4(6)} = \frac{15}{2} = 7.5$$

Remark: sum of squares of natural numbers $1^2+2^2+\cdots+n^2 = \dfrac{n(n+1)(2n+1)}{6}$.

Expected value is linear: $\mathbb{E}\big[aX + bY + c\big] \;=\; a\mathbb{E}\big[X\big] + b\mathbb{E}\big[Y\big] + c$

$$
\begin{aligned}
\mathbb{E}\big[aX + bY + c\big] = \mathbb{E}\big[f(X,Y)\big] = \sum_{x\in\mathcal{X},y\in\mathcal{Y}} f(x,y)p(x,y) &= \sum_{x\in\mathcal{X},y\in\mathcal{Y}} (ax + by + c)p(x,y) \\
\text{[expand } (ax+by+c)p(x,y)\text{]} &= \sum_{x\in\mathcal{X},y\in\mathcal{Y}} axp(x,y) + byp(x,y) + cp(x,y) \\
\text{[distribute summation sign]} &= a\sum_{x\in\mathcal{X},y\in\mathcal{Y}} xp(x,y) + b\sum_{x\in\mathcal{X},y\in\mathcal{Y}} yp(x,y) + c\sum_{x\in\mathcal{X},y\in\mathcal{Y}} p(x,y) \\
\text{[rewrite summation sign]} &= a\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} xp(x,y) + b\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} yp(x,y) + c\sum_{(x,y)\in\Omega} p(x,y) \\
\text{[rearrange summation sign, Axiom of probability]} &= a\sum_{x\in\mathcal{X}} x\sum_{y\in\mathcal{Y}} p(x,y) + b\sum_{y\in\mathcal{Y}} y\sum_{x\in\mathcal{X}} p(x,y) + c \\
\text{[rewrite } p(x,y) = \mathbb{P}(X=x,Y=y)\text{]} &= a\sum_{x\in\mathcal{X}} x\sum_{y\in\mathcal{Y}} \mathbb{P}(X=x,Y=y) + b\sum_{y\in\mathcal{Y}} y\sum_{x\in\mathcal{X}} \mathbb{P}(X=x,Y=y) + c \\
\text{[relationship between joint and marginal probability]} &= a\sum_{x\in\mathcal{X}} x\,\mathbb{P}(X=x) + b\sum_{y\in\mathcal{Y}} y\,\mathbb{P}(Y=y) + c \\
\text{[rewrite } \mathbb{P}(X=x,Y=y) = p(x,y)\text{]} &= a\sum_{x\in\mathcal{X}} xp(x) + b\sum_{y\in\mathcal{Y}} yp(y) + c \\
\text{[definition of expectation]} &= a\mathbb{E}[X] + b\mathbb{E}[Y] + c
\end{aligned}
$$

Expected value of independent product: $\mathbb{E}\big[XY\big] = \mathbb{E}\big[X\big]\mathbb{E}\big[Y\big]$

$$
\begin{aligned}
\mathbb{E}\big[XY\big] = \mathbb{E}\big[f(X,Y)\big] = \sum_{x\in\mathcal{X},y\in\mathcal{Y}} f(x,y)p(x,y) \quad &= \quad \sum_{x\in\mathcal{X},y\in\mathcal{Y}} xyp(x,y) \\
[X,Y \text{ independent so } p(x,y) = p(x)p(y)] \quad &= \quad \sum_{x\in\mathcal{X},y\in\mathcal{Y}} xyp(x)p(y) \\
\text{[split the summation]} \quad &= \quad \Big(\sum_{x\in\mathcal{X}} xp(x)\Big)\Big(\sum_{y\in\mathcal{Y}} yp(y)\Big) \\
&= \quad \mathbb{E}[X]\mathbb{E}[Y]
\end{aligned}
$$

Similarly, $\mathbb{E}\big[X_1X_2\cdots X_n\big] = \mathbb{E}[X_1]\mathbb{E}[X_2]\cdots\mathbb{E}[X_n]$ if all X_i are independent

What if X,Y not independent? Then just the first line $\mathbb{E}\big[XY\big] = \sum_{x\in\mathcal{X},y\in\mathcal{Y}} f(x,y)p(x,y)$

A long example of $\mathbb{E}\big[f(X)\big]$ ... 1/2

- Find $\mathbb{E}[X+Y]$, where $\begin{cases} X \text{ denotes the random variable of tossing a fair 4-sided die once} \\ Y \text{ denotes the random variable of tossing a fair 6-sided die once} \end{cases}$

- How to solve $\mathbb{E}\big[f(X)\big]$
  - **Method 1. Using shortcut formula** $\mathbb{E}\big[X+Y\big] \;=\; \mathbb{E}\big[X\big] + \mathbb{E}\big[Y\big]$
  - **Method 2. Using definition**
    - Let $S = f(X)$ be a new random variable, i.e., $s = f(x)$ step 1. Find all possible $s \in S$
    - By definition of expected value, $\mathbb{E}[S] = \sum sp(s)$
    - As f do not change probability, so $p(s) = p(x)$ step 2. Find all probability $p(s)$
    - So $\mathbb{E}[f(X)] = \mathbb{E}[S] = \sum sp(s) = \sum f(x)p(x)$

- **Method 1. Using expected value is linear**

$$\mathbb{E}\big[X+Y\big] \;=\; \mathbb{E}\big[X\big] + \mathbb{E}\big[Y\big] \;=\; \frac{1+2+3+4}{4} + \frac{1+2+3+4+5+6}{6} \;=\; \frac{4(5)}{4(2)} + \frac{6(7)}{2(6)} = 2.5 + 3.5 = 6$$

Using shortcut save you from lots of workload.

A long example of $\mathbb{E}\big[f(X)\big]$ ... 2/2

- Find $\mathbb{E}[X+Y]$, where $\begin{cases} X \text{ denotes the random variable of tossing a fair 4-sided dice once} \\ Y \text{ denotes the random variable of tossing a fair 6-sided dice once} \end{cases}$
- Let $S = X + Y$, we need to identify the sample space of S
- The sample space of (X, Y), which is NOT the same as $X + Y$, is ordered pair $\neq$ sum

$$\text{sample space of } (x,y) = \begin{bmatrix} (1,1) & (1,2) & (1,3) & (1,4) \\ (2,1) & (2,2) & (2,3) & (2,4) \\ (3,1) & (3,2) & (3,3) & (3,4) \\ (4,1) & (4,2) & (4,3) & (4,4) \\ (5,1) & (5,2) & (5,3) & (5,4) \\ (6,1) & (6,2) & (6,3) & (6,4) \end{bmatrix}, \quad \text{probability of } (x,y) = \begin{bmatrix} \frac{1}{24} & \frac{1}{24} & \frac{1}{24} & \frac{1}{24} \\ \frac{1}{24} & \frac{1}{24} & \frac{1}{24} & \frac{1}{24} \\ \frac{1}{24} & \frac{1}{24} & \frac{1}{24} & \frac{1}{24} \\ \frac{1}{24} & \frac{1}{24} & \frac{1}{24} & \frac{1}{24} \\ \frac{1}{24} & \frac{1}{24} & \frac{1}{24} & \frac{1}{24} \\ \frac{1}{24} & \frac{1}{24} & \frac{1}{24} & \frac{1}{24} \end{bmatrix}$$

- Now $S = X + Y$ has the sample space

$$S = X + Y = \begin{bmatrix} 2 & 3 & 4 & 5 \\ 3 & 4 & 5 & 6 \\ 4 & 5 & 6 & 7 \\ 5 & 6 & 7 & 8 \\ 6 & 7 & 8 & 9 \\ 7 & 8 & 9 & 10 \end{bmatrix} \quad \text{probability of } (x,y) = \begin{bmatrix} \frac{1}{24} & \frac{1}{24} & \frac{1}{24} & \frac{1}{24} \\ \frac{1}{24} & \frac{1}{24} & \frac{1}{24} & \frac{1}{24} \\ \frac{1}{24} & \frac{1}{24} & \frac{1}{24} & \frac{1}{24} \\ \frac{1}{24} & \frac{1}{24} & \frac{1}{24} & \frac{1}{24} \\ \frac{1}{24} & \frac{1}{24} & \frac{1}{24} & \frac{1}{24} \\ \frac{1}{24} & \frac{1}{24} & \frac{1}{24} & \frac{1}{24} \end{bmatrix}$$

- We can now construct a table

s	2	3	4	5	6	7	8	9	10
$p(s)$	$\frac{1}{24}$	$\frac{2}{24}$	$\frac{3}{24}$	$\frac{4}{24}$	$\frac{4}{24}$	$\frac{4}{24}$	$\frac{3}{24}$	$\frac{2}{24}$	$\frac{1}{24}$

Hence

$$\mathbb{E}[S] = \sum sp(s) = 2(\frac{1}{24}) + 3(\frac{2}{24}) + 4(\frac{3}{24}) + 5(\frac{4}{24}) + 6(\frac{4}{24}) + 7(\frac{4}{24}) + 8(\frac{3}{24}) + 9(\frac{2}{24}) + 10(\frac{1}{24}) = \frac{144}{24} = 6$$

Practise $\mathbb{E}\big[f(X)\big] \;=\; \sum f(x)p(x)$

Find $\mathbb{E}[(X-2Y)^2]$, where

$$\begin{cases} X \text{ denotes the random variable of tossing a fair 2-sided die once} \\ Y \text{ denotes the random variable of tossing a fair 4-sided die once} \end{cases}$$

solve this using method by definition and also using shortcut formula

Solution next page.

Practise $\mathbb{E}\big[f(X)\big] \;=\; \sum f(x)p(x)$, solution

- Method 1

$$\mathbb{E}[S] = \mathbb{E}[(X-2Y)^2] = \mathbb{E}[X^2 - 4XY + 4Y^2] = \mathbb{E}[X^2] - 4\mathbb{E}[XY] + 4\mathbb{E}[Y^2] = \mathbb{E}[X^2] - 4\mathbb{E}[X]\mathbb{E}[Y] + 4\mathbb{E}[Y^2]$$

  - $\mathbb{E}[X] = \dfrac{1+2}{2} = 1.5$
  - $\mathbb{E}[X^2] = \dfrac{1^2+2^2}{2} = 2.5$
  - $\mathbb{E}[Y] = \dfrac{1+2+3+4}{4} = \dfrac{4(5)}{4(2)} = 2.5$
  - $\mathbb{E}[Y^2] = \dfrac{1^2+2^2+3^2+4^2}{4} = \dfrac{4(5)(9)}{4(6)} = 7.5$ $\qquad 1^2 + \cdots + n^2 = \dfrac{n(n+1)(2n+1)}{6}$

$$\mathbb{E}[S] = \mathbb{E}[X^2] - 4\mathbb{E}[X]\mathbb{E}[Y] + 4\mathbb{E}[Y^2] = 2.5 - 4(1.5)(2.5) + 4(7.5) = 17.5$$

- Method 2
  - The set of all possible (x,y) is $\begin{bmatrix}(1,1) & (1,2) & (1,3) & (1,4)\\ (2,1) & (2,2) & (2,3) & (2,4)\end{bmatrix}$
  - Let $S=(X-2Y)^2$, we have $X-2Y = \begin{bmatrix}-1 & -3 & -5 & -7\\ 0 & -2 & -4 & -6\end{bmatrix}$ and hence for $S=(X-2Y)^2$ we have

$$S = \begin{bmatrix}1 & 9 & 25 & 49\\ 0 & 4 & 16 & 36\end{bmatrix}, \quad S = \Big\{0,1,4,9,16,25,36,49\Big\} \text{ with all } p(s) = \frac{1}{8},$$

$$\text{thus } \mathbb{E}[S] = \frac{0+1+4+9+16+25+36+49}{8} \;=\; 17.5$$

Moment and moment-generating function **Not in exam**

- $\mathbb{E}[X] = \mathbb{E}[X^1] = \sum_{x\in\mathcal{X}} x^1 p(x)$ is 1st-order moment

- $\mathbb{E}[X^2] = \sum_{x\in\mathcal{X}} x^2 p(x)$ is 2st-order moment

- k-th moment: $\mathbb{E}[X^k] = \sum_{x\in\mathcal{X}} x^k p(x)$ i.e., $\mathbb{E}[f(X)] = \sum_{x\in\mathcal{X}} f(x)p(x)$ with $f(x) = x^k$

- Moments are terms in the Taylor series of moment-generating function

$$e^{tX} \;=\; 1 + tX + \frac{1}{2!}t^2X^2 + \frac{1}{3!}t^3X^3 + \cdots + \frac{1}{n!}t^nX^n + \dots \qquad \text{(Taylor series)}$$

  Moment-generating function

$$\mathbb{M}_X(t) = \mathbb{E}[e^{tX}] \;=\; 1 + t\mathbb{E}[X] + \frac{1}{2!}t^2\mathbb{E}[X^2] + \frac{1}{3!}t^3\mathbb{E}[X^3] + \cdots + \frac{1}{n!}t^n\mathbb{E}[X^n] + \dots$$

- If X is a continuous RV, then $\mathbb{M}_X(t)$ is the Laplace transform of p_X on $-x$: $\mathbb{M}_X(t) = \mathcal{L}\{p_X\}(-t)$

Practise (Madbook 3.3 Q1)

	X=1	X=2	X=3
Y=1	0.1	0.1	0.2
Y=2	0.2	a	0.1

Find

- a
- $\mathbb{E}[X]$
- $\mathbb{E}[Y]$
- $\mathbb{E}[2X]$
- $\mathbb{E}[-3Y]$
- $\mathbb{E}[X^2]$
- $\mathbb{E}[Y^2]$
- $\mathbb{E}[X+Y]$
- $\mathbb{E}[XY]$
- $\mathbb{E}[(X,Y)]$
- $\mathbb{P}(X=1|Y=0)$
- $\mathbb{P}(Y=0|X=1)$

Section summary

- We write $\mathbb{P}(X=x)=p(x)$
- Expectation $\mathbb{E}[X] \coloneqq \sum_{x\in\mathcal{X}} xp(x)$
- $\mathbb{E}[f(X)] = \sum_{x\in\mathcal{X}} f(x)p(x)$
- $\mathbb{E}[aX+bY+c] = a\mathbb{E}[X] + b\mathbb{E}[Y] + c$
- $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$ if X, Y independent
- Conditional expectation $\mathbb{E}[X|Y]$
- Marginal expectation $\mathbb{E}[X]$
- Joint expectation $\mathbb{E}[X,Y]$

Contents

Variance

$$\mathbb{V}[X] = \mathbb{E}\Big[\big(X - \mathbb{E}[X]\big)^2\Big] \iff \mathbb{E}[f(X)] \quad \text{where} \quad f(\cdot) = (\,\cdot\, - \mathbb{E}[\,\cdot\,])^2$$

- Variance = standard deviation^2 , standard deviation = $\sqrt{\text{variance}}$

- **E.g.** $\mathbb{P}(X=1) = 0.5, \mathbb{P}(X=2) = 0.4, \mathbb{P}(X=3) = 0.1$, recall $\mathbb{E}[X] = 1.6$, so
$$\mathbb{V}[X] = (1-1.6)^2 \cdot 0.5 + (2-1.6)^2 \cdot 0.4 + (3-1.6)^2 \cdot 0.1 = 0.44$$

- Recall $\mathbb{E}[aX + bY + c] = a\mathbb{E}[X] + b\mathbb{E}[Y] + c$, we have
$$\begin{aligned}
\mathbb{V}[X] &= \mathbb{E}\Big[\big(X - \mathbb{E}[X]\big)^2\Big] \\
&= \mathbb{E}\Big[X^2 - 2X\mathbb{E}[X] + \big(\mathbb{E}[X]\big)^2\Big] \\
&= \mathbb{E}\Big[X^2\Big] - \mathbb{E}\Big[2X\mathbb{E}[X]\Big] + \mathbb{E}\Big[\big(\mathbb{E}[X]\big)^2\Big] \\
&= \mathbb{E}\Big[X^2\Big] - 2\mathbb{E}[X]\mathbb{E}[X] + \big(\mathbb{E}[X]\big)^2 \qquad \mathbb{E}\Big[2X\mathbb{E}[X]\Big] = 2\mathbb{E}[X]\mathbb{E}[X] \text{ since } \mathbb{E}[X] \text{ is a number} \\
&= \mathbb{E}[X^2] - (\mathbb{E}[X])^2
\end{aligned}$$

Covariance and correlation

- Variance: seeing the variable as a whole entity
  Covariance: seeing the variable part by part

- **Definition** Given two RVs X, Y, covariance is defined as

$$\text{cov}(X,Y) \;:=\; \mathbb{E}\Big[(X-\mathbb{E}[X])(Y-\mathbb{E}[Y])\Big] \;=\; \mathbb{E}[XY]-\mathbb{E}[X]\mathbb{E}[Y]$$

  How to remember: recall variance $\mathbb{V}[X] := \mathbb{E}\Big[(X-\mathbb{E}[X])^2\Big] = \mathbb{E}\Big[(X-\mathbb{E}[X])(X-\mathbb{E}[X])\Big]$
  Cov $=$ var with one X replaced by Y

- **Definition** Given two RVs X, Y, the **Pearson correlation coefficient** is defined as

$$\text{corr}(X,Y) \;:=\; \frac{\text{cov}(X,Y)}{\sqrt{\mathbb{V}[X]}\sqrt{\mathbb{V}[Y]}}. \qquad \text{(correlation)}$$

- $-\infty \le \text{cov}(X,Y) \le \infty$ and $-1 \le \text{corr}(X,Y) \le 1$ correlation $=$ normalized covariance
- If X, Y independent, then $\text{cov}(X,Y) = \text{corr}(X,Y) = 0$ converse is not true

Example of cov$(X,Y) = \mathbb{E}\Big[(X-\mathbb{E}[X])(Y-\mathbb{E}[Y])\Big]$ of two RVs

- An example of distribution over $\Omega = \mathcal{X} \times \mathcal{Y}$

	X=1	X=2	X=3
Y=1	0.05	0.15	0.1
Y=2	0.25	0.15	0.3

- Step 1. Get $\mathbb{E}[X]$
  - $\mathbb{E}[X]$ is X only
  - marginal probability on X means we "collapse Y" and get

X=1	X=2	X=3
0.3	0.3	0.4

and so $\mathbb{E}[X] = 2.1$

- Step 2. Get $\mathbb{E}[Y]$
  - $\mathbb{E}[Y]$ is Y only
  - marginal probability on Y means we "collapse X" and get

Y=1	0.3
Y=2	0.7

and so $\mathbb{E}[Y] = 1.7$

- Step 3.

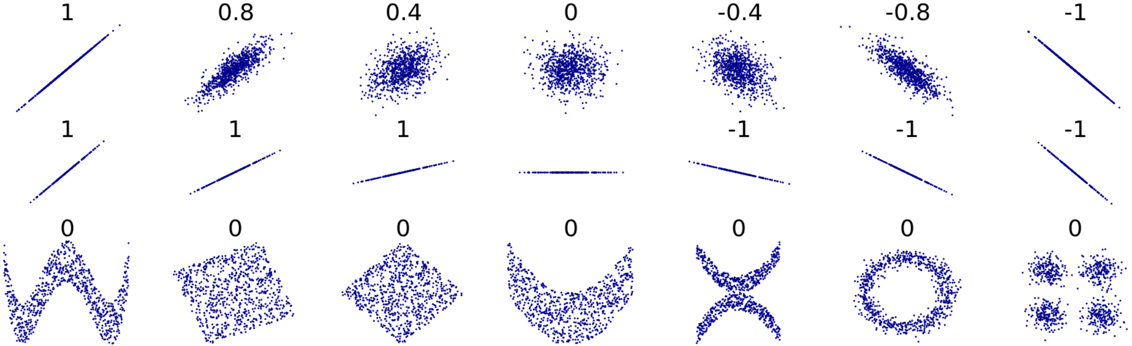
	$X-\mathbb{E}[X] = -1.1$	$X-\mathbb{E}[X] = -0.1$	$X-\mathbb{E}[X] = 0.9$
$Y-\mathbb{E}[Y] = -0.7$	0.05	0.15	0.1
$Y-\mathbb{E}[Y] = 0.3$	0.25	0.15	0.3

$$\mathrm{cov}(X,Y) = (-1.1)(-0.7)(0.05) + (-0.1)(-0.7)(0.15) + \cdots$$

- Another method: $\mathbb{E}\Big[(X-\mathbb{E}[X])(Y-\mathbb{E}[Y])\Big] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$.

Why cov and corr: study the probabilistic relationship between X and Y

- Positive covariance/correlation
  if X is greater than $\mathbb{E}[X]$ then *likely* Y is *greater* than $\mathbb{E}[Y]$
- Negative covariance/correlation
  if X is greater than $\mathbb{E}[X]$ then *likely* Y is *less* than $\mathbb{E}[Y]$



- Correlation is not causation
  - "The lack of pirates is causing global warming"
  - "Fireman causing fire"
  - "cholesterol is bad"

Properties of cov

$$\begin{aligned}
\mathrm{cov}(X,X) &= \mathbb{V}[X] \\
\mathrm{cov}(aX,Y) &= a\mathrm{cov}(X,Y) \\
\mathrm{cov}(X+c,Y) &= \mathrm{cov}(X,Y) \\
\mathrm{cov}(X+Z,Y) &= \mathrm{cov}(X,Y)+\mathrm{cov}(Z,Y)
\end{aligned}$$

Generalization

$$\mathrm{cov}\left(a_1X_1+a_2X_2+...+a_mX_m,\ b_1Y_1+b_2Y_2+...+a_nY_n\right) = \sum_{i=1}^{m}\sum_{j=1}^{n} a_ib_j\mathrm{cov}(X_i,Y_j)$$

We will not go too deep into these.

Quadratic formula of variance

- If X, Y are two random variables, then

$$\mathbb{V}[aX + bY + c] = a^2\mathbb{V}[X] + 2ab\mathrm{cov}(X, Y) + b^2\mathbb{V}[Y]. \qquad \text{(important)}$$

  Corollary: if X, Y are independent: $\mathrm{cov}(X, Y) = 0$, so

$$\mathbb{V}[aX + bY + c] = a^2\mathbb{V}[X] + b^2\mathbb{V}[Y]$$

- Think of this as

$$\begin{aligned}(aX + bY)^2 &= (aX)^2 + 2(aX)(bY) + (bY)^2 \\ &= a^2X^2 + 2abXY + b^2Y^2\end{aligned}$$

- Generalization

$$\mathbb{V}[aX + bY + cZ + d] = a^2\mathbb{V}[X] + 2ab\mathrm{cov}(X, Y) + 2ac\mathrm{cov}(X, Z) + b^2\mathbb{V}[Y] + 2bc\mathrm{cov}(Y, Z) + c^2\mathbb{V}[Z]$$

  Similar to

$$(aX + bY + cZ)^2 = a^2X^2 + 2abXY + 2acXZ + Y^2 + 2bcYZ + Z^2$$

Contents

Example of conditional expectation and conditional variance

- Context: you live next to the sea and you want to see dolphin
- $\mathcal{X} = \{ \text{ no dolphin}, \text{has dolphin}\}$
- $\mathcal{Y} = \{ \text{ bad weather day}, \text{good weather day}\}$
- Consider $X|Y$
- $\mathbb{P}(X|Y = \text{ bad weather day})$
- $\mathbb{E}[X|Y = \text{ bad weather day}]$
- $\mathbb{E}[X|Y = \text{good weather day}]$
- $\mathbb{E}[X|Y]$
- $\mathbb{V}[X|Y = \text{ bad weather day}]$
- $\mathbb{V}[X|Y = \text{good weather day}]$
- $\mathbb{V}[X|Y]$

Conditional Expectation

- **Definition** $\mathbb{E}[X|Y=y]$ is the conditional expectation of X given $Y=y$

$$\mathbb{E}(X|Y=y) = \sum xp(x|y) = \sum x \frac{p(x,y)}{p(y)}$$

or equivalently, a random variable $Z(y) = \mathbb{E}[X|Y=y]$ defined as

$$Z(y) = \begin{cases} \mathbb{E}[X|Y=y_1] & \text{with probability } \mathbb{P}(Y=y_1) \\ \mathbb{E}[X|Y=y_2] & \text{with probability } \mathbb{P}(Y=y_2) \\ \vdots & \end{cases}$$

Z is a function of y. I.e., Z depends on y.

- **E.g.**

	X=1 (lived 30yr)	X=2 (lived 60yr)	X=3 (lived 90yr)
Y=1 (no cancer)	a	b	c
Y=2 (cancer)	d	e	f

The point is, if we are focusing on $Y=1$, then we ignore the information of $Y \neq 1$ when we do the calculation

Example

	X=1 (lived 30yr)	**X=2 (lived 60yr)**	**X=3 (lived 90yr)**
Y=1 (no cancer)	a	b	c
Y=2 (cancer)	d	e	f

- Obtain the marginal probabilities

	X=1 (lived 30yr)	X=2 (lived 60yr)	X=3 (lived 90yr)	
Y=1 (no cancer)	a	b	c	$\mathbb{P}(\text{no cancer}) = \mathbb{P}(Y=1) = a+b+c$
Y=2 (cancer)	d	e	f	$\mathbb{P}(\text{cancer}) = \mathbb{P}(Y=2) = d+e+f$
	$\mathbb{P}(X=1) = a+d$	$\mathbb{P}(X=2) = b+e$	$\mathbb{P}(X=3) = c+f$	

- $X|Y=1$

	X=1 (lived 30yr)	X=2 (lived 60yr)	X=3 (lived 90yr)	
Y=1 (no cancer)	a	b	c	$\mathbb{P}(\text{no cancer}) = \mathbb{P}(Y=1) = a+b+c$

Meaning of $X|Y=1$: the summary of "if no cancer", what are the chance you lived short / mid / long

- $\mathbb{E}[X|Y=1]$
The a,b,c are NOT probability for $X|Y=1$, because $a+b+c \neq 1$.
To make a,b,c probability for $X|Y=1$, we normalize

$$(a,b,c) \;\mapsto\; \Big(\frac{a}{a+b+c}, \frac{b}{a+b+c}, \frac{c}{a+b+c}\Big) \;=\; \Big(\frac{a}{\mathbb{P}(Y=1)}, \frac{b}{\mathbb{P}(Y=1)}, \frac{c}{\mathbb{P}(Y=1)}\Big)$$

Now

$$\mathbb{E}[X|Y=1] \;=\; \underbrace{1}_{X=1} \underbrace{\frac{a}{a+b+c}}_{\mathbb{P}(X=1|Y=1)} + \underbrace{2}_{X=2} \underbrace{\frac{b}{a+b+c}}_{\mathbb{P}(X=2|Y=1)} + \underbrace{3}_{X=3} \underbrace{\frac{c}{a+b+c}}_{\mathbb{P}(X=3|Y=1)}$$

Example

	X=1	X=2	X=3	
Y=1	**0.05**	**0.15**	**0.1**	$\mathbb{P}(Y=1)=0.3$
Y=2	**0.25**	**0.15**	**0.3**	$\mathbb{P}(Y=2)=0.7$

- Let $Z = \mathbb{E}[X|Y=y]$ $\quad Z = \begin{cases} \mathbb{E}[X|Y=1] & \text{with probability } \mathbb{P}(Y=1)=0.3 \\ \mathbb{E}[X|Y=2] & \text{with probability } \mathbb{P}(Y=2)=0.7 \end{cases}$

$$\mathbb{E}[X|Y=1] = 1\cdot\frac{0.05}{0.3} + 2\cdot\frac{0.15}{0.3} + 3\cdot\frac{0.1}{0.3} = 2.16666666667$$

$$\mathbb{E}[X|Y=2] = 1\cdot\frac{0.25}{0.7} + 2\cdot\frac{0.15}{0.7} + 3\cdot\frac{0.3}{0.7} = 2.07142857143$$

$$Z = \begin{cases} 2.16666666667 & \text{with prob } 0.3 \\ 2.07142857143 & \text{with prob } 0.7 \end{cases} \iff \mathbb{E}[Z] = 2.16666666667\,\cdot\,0.3 + 2.07142857143\,\cdot\,0.7 = 0.65 + 1.45 = 2.1$$

- Short-cut (be cautious)

$$\begin{aligned}
\mathbb{E}[Z] &= \underbrace{1\cdot\frac{0.05}{0.3} + 2\cdot\frac{0.15}{0.3} + 3\cdot\frac{0.1}{0.3}}_{\mathbb{E}[X|Y=1]} \cdot \underbrace{0.3}_{\mathbb{P}(Y=1)} + \underbrace{1\cdot\frac{0.25}{0.7} + 2\cdot\frac{0.15}{0.7} + 3\cdot\frac{0.3}{0.7}}_{\mathbb{E}[X|Y=2]} \cdot \underbrace{0.7}_{\mathbb{P}(Y=2)} \\
&= 1\cdot 0.05 + 2\cdot 0.15 + 3\cdot 0.1 \; + \; 1\cdot 0.25 + 2\cdot 0.15 + 3\cdot 0.3 \\
&= 1\cdot\underbrace{(0.05+0.25)}_{\mathbb{P}(X=1)} + 2\cdot\underbrace{(0.15+0.15)}_{\mathbb{P}(X=2)} + 3\cdot\underbrace{(0.1+0.3)}_{\mathbb{P}(X=3)} \\
&= \mathbb{E}[X] \text{ the } \textbf{unconditional} \text{ expectation of } X, \text{ this is because } \mathbb{E}[X] = \mathbb{E}_Y\Big[\mathbb{E}[X|Y]\Big]
\end{aligned}$$

- Practise: find $W(x) = \mathbb{E}[Y|X=x]$ and also $\mathbb{E}[W]$.

This is incorrect

	X=1	X=2	X=3	
Y=1	0.05	0.15	0.1	$\mathbb{P}(Y=1)=0.3$
Y=2	0.25	0.15	0.3	$\mathbb{P}(Y=2)=0.7$

- The following expression is nonsense

$$1\cdot\frac{0.05}{0.3}+2\cdot\frac{0.15}{0.3}+3\cdot\frac{0.1}{0.3}+1\cdot\frac{0.25}{0.7}+2\cdot\frac{0.15}{0.7}+3\cdot\frac{0.3}{0.7}$$

- Why: it violates the probability axiom "the probability of sample space is 1"

$$\begin{aligned}
&1\cdot\frac{0.05}{0.3}+2\cdot\frac{0.15}{0.3}+3\cdot\frac{0.1}{0.3}+1\cdot\frac{0.25}{0.7}+2\cdot\frac{0.15}{0.7}+3\cdot\frac{0.3}{0.7}\\
=\;&1\Big(\frac{0.05}{0.3}+\frac{0.25}{0.7}\Big)+2\Big(\frac{0.15}{0.3}+\frac{0.15}{0.7}\Big)+3\Big(\frac{0.1}{0.3}+\frac{0.3}{0.7}\Big)\\
=\;&1\Big(0.52\Big)+2\Big(0.71\Big)+3\Big(0.76\Big)
\end{aligned}$$

The values $(0.52, 0.71, 0.76)$ do not sum to 1 $\implies$ they are not probability.

Conditional variance $\mathbb{V}[X|Y=y]$

- Example

	X=1	X=2	X=3	
Y=1	0.05	0.15	0.1	$\mathbb{P}(Y=1)=0.3$
Y=2	0.25	0.15	0.3	$\mathbb{P}(Y=2)=0.7$

- This is wrong, because the big bracket terms in the second line are not probability

$$(1-2.1)^2\frac{0.05}{0.3}+(2-2.1)^2\frac{0.15}{0.3}+(3-2.1)^2\frac{0.1}{0.3}+(1-2.1)^2\frac{0.25}{0.7}+(2-2.1)^2\frac{0.15}{0.7}+(3-2.1)^2\frac{0.3}{0.7}$$

$$=\quad(1-2.1)^2\left(\frac{0.05}{0.3}+\frac{0.25}{0.7}\right)+(2-2.1)^2\left(\frac{0.15}{0.3}+\frac{0.15}{0.7}\right)+(3-2.1)^2\left(\frac{0.1}{0.3}+\frac{0.3}{0.7}\right)$$

- Suggested approach: calculate one-by-one
  - What is $X|Y=1$

	X=1	X=2	X=3	
Y=1	0.05	0.15	0.1	$\mathbb{P}(Y=1)=0.3$

$\xrightarrow{\text{normalisation}}$

X=1	X=2	X=3
0.05/0.3	0.15/0.3	0.1/0.3

$$\mathbb{E}[X|Y=1]=2.16\ldots,\quad \mathbb{V}[X|Y=1]=(1-2.16\ldots)^2\frac{0.05}{0.3}+(2-2.16\ldots)^2\frac{0.15}{0.3}+(3-2.16\ldots)^2\frac{0.1}{0.3}\;=\;0.25$$

  - What is $X|Y=2$

	X=1	X=2	X=3	
Y=2	0.25	0.15	0.3	$\mathbb{P}(Y=2)=0.7$

$\xrightarrow{\text{normalisation}}$

X=1	X=2	X=3
0.25/0.7	0.15/0.7	0.3/0.7

$$\mathbb{E}[X|Y=2]=2.07\ldots\quad \mathbb{V}[X|Y=1]=(1-2.07\ldots)^2\frac{0.25}{0.7}+(2-2.07\ldots)^2\frac{0.15}{0.7}+(3-2.07\ldots)^2\frac{0.3}{0.7}\;=\;0.41$$

Conditional variance $\mathbb{V}[X|Y=y]$

- Example

	X=1	X=2	X=3	
Y=1	0.05	0.15	0.1	$\mathbb{P}(Y=1)=0.3$
Y=2	0.25	0.15	0.3	$\mathbb{P}(Y=2)=0.7$

- Let $W(y)=\mathbb{V}[X|Y=y]$, then

$$W=\begin{cases}0.25 & \text{with probability } \mathbb{P}(Y=1)=0.3\\ 0.41 & \text{with probability } \mathbb{P}(Y=2)=0.7\end{cases}$$

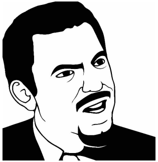
- Then you can compute $\mathbb{E}[W]$ and $\mathbb{V}[W]$

Advanced topic Not in exam

- You can keep going on ...
  $\mathbb{V}\Big[\mathbb{E}\big[\mathbb{V}[X|Y]\big]\Big|Y\Big]$

  $\mathbb{E}\Big[\mathbb{V}\big[f(X)|Y\big]\Big]$

  $\mathbb{V}\Big[g\Big(\mathbb{E}\Big[\mathbb{V}\big[f(X)|h(Y)\big]\Big]\Big|Y\Big)\Big]$



- Therefore we need tools:
  - let $\mu = \mathbb{E}[X]$ and $\sigma^2 = \mathbb{V}[X]$
  - $f(x)$ is twice differentiable at x

$$\mathbb{E}[f(X)] \approx f(\mu) + \frac{\sigma^2}{2}\frac{\partial^2 f(x)}{\partial x^2}\Big|_{x=\mu}$$

$$\mathbb{V}[f(X)] \approx \sigma^2\left[\frac{\partial^2 f(x)}{\partial x^2}\Big|_{x=\mu}\right]^2$$

  If $f(x) = g(h(x))$, use chain rule in calculus.

- Or conditional over two random variables...

Section summary

- Variance $\mathbb{V}[X] = \mathbb{E}\Big[\big(X - \mathbb{E}[X]\big)^2\Big] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$
- $\mathbb{V}[X|Y]$
- cov$(X, Y) = \mathbb{E}\Big[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])\Big] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$
- corr$(X, Y) = \dfrac{\text{cov}(X, Y)}{\sigma_X \sigma_Y}$, where σ_X^2= Var of X
- Meaning of cov and corr

Short-cut formula: Fundamental theorems of poker   Not in exam

- Law of total expectation

$$\mathbb{E}[X] = \mathbb{E}_Y\Big[\mathbb{E}[X|Y=y]\Big]$$

- Law of total variance

$$\mathbb{V}[X] = \mathbb{E}_Y\Big[\mathbb{V}[X|Y=y]\Big] + \mathbb{V}_Y\Big[\mathbb{E}[X|Y=y]\Big]$$

- Law of total probability

$$\mathbb{P}(X) = \sum_{y\in\mathcal{Y}} \mathbb{P}(X|Y=y)\mathbb{P}(Y=y)$$

- Law of total covariance

$$\mathrm{cov}(X,Y) = \mathbb{E}_Z\Big[\mathrm{cov}(X,Y|Z=z)\Big] + \mathrm{cov}\Big(\mathbb{E}[X|Z=z], \mathbb{E}[Y|Z=z]\Big)$$

"The probability laws for decision-making when dealing with incomplete information"

Ultimate example

	Y=0	Y=1
X=0	0.2	0.4
X=1	0.4	0

1. $\mathbb{P}(X=0) = \mathbb{P}(X=0, Y=0) + \mathbb{P}(X=0, Y=1) = 0.2 + 0.4 = 0.6$ $\quad \mathbb{P}(X=1) = \mathbb{P}(X=1, Y=0) + \mathbb{P}(X=1, Y=1) = 0.4 + 0 = 0.4$
2. $\mathbb{P}(Y=0) = \mathbb{P}(X=0, Y=0) + \mathbb{P}(X=1, Y=0) = 0.2 + 0.4 = 0.6$ $\quad \mathbb{P}(Y=1) = \mathbb{P}(X=0, Y=1) + \mathbb{P}(X=1, Y=1) = 0.4 + 0 = 0.4$
3. $\mathbb{E}[X] = 0 \cdot \mathbb{P}(X=0) + 1 \cdot \mathbb{P}(X=1) = 0.4$ $\quad \mathbb{E}[Y] = 0 \cdot \mathbb{P}(Y=0) + 1 \cdot \mathbb{P}(Y=1) = 0.4$
4. $\mathbb{E}[XY] = (0 \cdot 0)\mathbb{P}(X=0, Y=0) + (0 \cdot 1)\mathbb{P}(X=0, Y=1) + (1 \cdot 0)\mathbb{P}(X=1, Y=0) + (1 \cdot 1)\mathbb{P}(X=1, Y=1) = 0$
5. $\text{cov}(X,Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] = 0 - 0.4 \cdot 0.4 = 0.16 \neq 0 \implies X, Y$ are not independent
6. $\mathbb{V}[X] = (0 - \mathbb{E}[X])^2\mathbb{P}(X=0) + (1 - \mathbb{E}[X])^2\mathbb{P}(X=1) = 0.4^2 \cdot 0.6 + 0.6^2 \cdot 0.4 = 0.24$
7. $\mathbb{V}[Y] = (0 - \mathbb{E}[Y])^2\mathbb{P}(Y=0) + (1 - \mathbb{E}[Y])^2\mathbb{P}(Y=1) = 0.4^2 \cdot 0.6 + 0.6^2 \cdot 0.4 = 0.24$
8. $\text{corr}(X,Y) = \dfrac{\text{cov}(X,Y)}{\sqrt{\mathbb{V}[X]}\sqrt{\mathbb{V}[Y]}} = \dfrac{0.16}{0.24} \approx 0.66$
9. $\mathbb{P}(X=0|Y=0) = \dfrac{\mathbb{P}(X=0, Y=0)}{\mathbb{P}(Y=0)} = \dfrac{0.2}{0.6} \approx 0.33$ $\quad \mathbb{P}(X=1|Y=0) = 1 - \mathbb{P}(X=0|Y=0) \approx 0.66$
10. $\mathbb{E}[X|Y=0] = 0 \cdot \mathbb{P}(X=0|Y=0) + 1 \cdot \mathbb{P}(X=1|Y=0) = 0 + 0.66 = 0.66$
11. $\mathbb{P}(X=0|Y=1) = \dfrac{\mathbb{P}(X=0, Y=1)}{\mathbb{P}(Y=1)} = \dfrac{0.4}{0.4} = 1$ $\quad \mathbb{P}(X=1|Y=1) = 1 - \mathbb{P}(X=0|Y=1) = 0$
12. $\mathbb{E}[X|Y=1] = 0 \cdot \mathbb{P}(X=0|Y=1) + 1 \cdot \mathbb{P}(X=1|Y=1) = 0 \cdot 1 + 1 \cdot 0 = 0$
13. Let $Z = \mathbb{E}[X|Y] = \begin{cases} \mathbb{E}[X|Y=0] & Y=0 \\ \mathbb{E}[X|Y=1] & Y=1 \end{cases}$. We have $Z = \begin{cases} 0.66 & \text{with probability } 0.6 \\ 0 & \text{with probability } 0.4 \end{cases}$, so the PMF of Z is $p(z) = \begin{cases} 0.6 & z = 0.66 \\ 0.4 & z = 0 \\ 0 & \text{otherwise} \end{cases}$
14. $\mathbb{E}[Z] = 0.66 \cdot 0.6 + 0 \cdot 0.4 = 0.4$ $\quad$ note that $\mathbb{E}[X] = 0.4$ so we have $\mathbb{E}[X] = \mathbb{E}[Z] = \mathbb{E}[\mathbb{E}[X|Y]]$
15. $\mathbb{V}[Z] = \mathbb{E}[Z^2] - (\mathbb{E}[Z])^2 = 0.66^2 \cdot 0.6 + 0^2 \cdot 0.4 - 0.4^2 \approx 0.106$
16. $\mathbb{V}[X|Y=0] = (0 - \mathbb{E}[X|Y=0])^2\mathbb{P}(X=0|Y=0) + (1 - \mathbb{E}[X|Y=0])^2\mathbb{P}(X=1|Y=0) = 0.66^2 \cdot 0.33 + 0.34^2 \cdot 0.66 \approx 0.22$
17. $\mathbb{V}[X|Y=1] = (0 - \mathbb{E}[X|Y=1])^2\mathbb{P}(X=0|Y=1) + (1 - \mathbb{E}[X|Y=1])^2\mathbb{P}(X=1|Y=1) = 0^2 \cdot 1 + 1^2 \cdot 0 = 0$
18. Let $V = \mathbb{V}[X|Y] = \begin{cases} \mathbb{V}[X|Y=0] & Y=0 \\ \mathbb{V}[X|Y=1] & Y=1 \end{cases}$. We have $V = \begin{cases} 0.22 & \text{with probability } 0.6 \\ 0 & \text{with probability } 0.4 \end{cases}$

Contents

Discrete Distributions, overview

- We denote $p(x|\theta), x \in \mathcal{X}, \theta \in \Theta$
  - θ: parameter
  - Θ: set of valid parameter
  - by changing θ we change the distribution
- Bernoulli — toss coin 1 times, 1 success
- Binomial — toss coin n times, k success
- Trinomial and multinomial — generalized binomial
- Uniform $\frac{1}{n}$ — evenly distributed
- Geometric — toss coin k times, first success at the kth time
- Hypergeometric — k succeed of n draw with no replacement in N-choose-K
- Negative binomial — toss coin n times, k fails
- Poisson — probability of k events occur during an interval

Discrete Distributions, overview

- Bernoulli $p(x|\theta) = \theta^x(1-\theta)^{1-x}, \theta \in [0,1], x \in \mathbb{N}$ — toss coin 1 times, 1 success
- Binomial $p(k|n,\theta) = \binom{n}{k}\theta^k(1-\theta)^{n-k}, n,k \in \mathbb{N}$ — toss coin n times, k success
- Trinomial and multinomial $p(k_1,k_2,k_3|n,\theta_1,\theta_2,\theta_3) = \binom{n}{k_1,k_2,k_3}\theta_1^{k_1}\theta_2^{k_2}\theta_3^{k_3}$ — generalized binomial
- Uniform $\frac{1}{n}$ — evenly distributed
- Geometric $p(k|\theta) = (1-\theta)^{k-1}\theta, k \in \{1,2,...\}$ — toss coin k times, first success at the kth time
- Hypergeometric $p(k|N,K,n) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}$ — k succeed of n draw with no replacement in N-choose-K
- Negative binomial $p(k|\theta) = \left\langle {n \atop k} \right\rangle (1-\theta)^n\theta^k$ — toss coin n times, k fails
- Poisson $p(k|\lambda) = \frac{\lambda^k e^{-\lambda}}{k!}, \lambda \geq 0, k \in \mathbb{N}$ — probability of k events occur during an interval

Bernoulli distribution - single binary event (e.g. toss a coin)

- $\Omega = \{0, 1\}$ (i.e., H or T, success or fail)

- $\mathbb{P}(X = 1|\theta) = \theta$
  $\theta \in [0, 1]$: probability of success
  $1 - \theta$: probability of fail
  Here 1 means success

- Probability mass function

$$p(x|\theta) = \theta^x (1 - \theta)^{1-x}$$

  - $p(1|\theta) = \mathbb{P}(X = 1|\theta) = \theta = \theta^1 (1 - 1)^{1-1}$ = probability of success
  - $p(0|\theta) = \mathbb{P}(X = 0|\theta) = 1 - \theta = \theta^0 (1 - 0)^{1-0}$ = probability of fail

- If RV X follows a Bernoulli distribution under parameter θ, we write $X \sim \text{Ber}(\theta)$

Bernoulli distribution $X \sim \textbf{Ber}(\theta)$

Sample space	$\Omega = \{0, 1\}$
Parameter	$\theta \in [0, 1]$
Meaning of the parameter	chance of success
PMF	$p(x\|\theta) = \theta^x (1-\theta)^{1-x}$
$\mathbb{E}[X]$	θ
$\mathbb{V}[X]$	$\theta(1-\theta)$

- $\mathbb{E}[X] = \underbrace{(1-\theta)}_{x=0} \cdot 0 + \underbrace{\theta}_{x=1} \cdot 1 = 0 + \theta = \theta$

- $X^2 = X$ for $X \sim \text{Ber}(\theta)$
  - This means $X^3 = X^4 = \ldots = X^k = X$

- $\mathbb{E}[X^2] = (1-\theta)0^2 + \theta 1^2 = \theta$

- $\mathbb{V}[X] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2 = \theta - \theta^2 = \theta(1-\theta)$

Binomial distribution - multiple binary events

- Binomial = out of n independent Bernoulli trials, exactly k success

$$p(k|n,\theta) = \binom{n}{k} \prod_{i=1}^{k} p(x_i|\theta) = \binom{n}{k} \theta^k (1-\theta)^{n-k}.$$

- Example. 4d2 (flip a coin four times), considering having $k = 2$ success, we have 6 possible cases

$$\begin{gathered} 1,1,0,0 \\ 1,0,1,0 \\ 1,0,0,1 \\ 0,1,1,0 \\ 0,1,0,1 \\ 0,0,1,1 \end{gathered}$$

The probability

$$p(k=2|\theta) \;=\; \binom{n=4}{k=2} \theta^2 (1-\theta)^{4-2} \;=\; \frac{4!}{2!2!} \theta^2 (1-\theta)^2 \;=\; 6 \underbrace{\theta^2}_{\text{2 success}} \underbrace{(1-\theta)^2}_{\text{2 fail}}$$

Binomial distribution $X \sim \textsf{Bin}(\theta)$

Sample space	$\Omega = \{0, 1, \ldots, n\}$ for number of success
Parameter	$\theta \in [0, 1]$
Meaning of the parameter	θ chance of success
Meaning of n	n independent Bernoulli trials
Meaning of k	exactly k success
PMF	$p(n, k \mid \theta) = \binom{n}{k} \theta^k (1-\theta)^{n-k}$

- Binomial = out of n independent Bernoulli trials, exactly k success
- Recall $Y \sim \text{Ber}(\theta)$ then $\mathbb{E}[Y] = \theta$
- $\mathbb{E}[X]$

$$\underbrace{X = \sum_{i=1}^{n} X_i}_{X_i \sim \text{Ber}(\theta)} \implies \mathbb{E}[X] = \mathbb{E}\Big[\sum_{i=1}^{n} X_i\Big] \overset{\text{expectation is linear}}{=} \sum_{i=1}^{n} \mathbb{E}[X_i] = \sum_{i=1}^{n} \underbrace{\theta}_{\mathbb{E}[X_i]=\theta} = n\theta$$

- Physical meaning: suppose success change $\theta = 0.5$, then for $n = 12$ trials, you expect to see half $= 6 = 12 \cdot 0.5$ of them success

Binomial distribution $X \sim \textbf{Bin}(\theta)$, find $\mathbb{E}[X^2]$

- We have $\underbrace{X = \sum_{i=1}^{n} X_i}_{X_i \sim \text{Ber}(\theta)}$

- Then $\mathbb{E}[X^2] \;=\; \mathbb{E}\left[\left(\sum_{i=1}^{n} X_i\right)^2\right] \;=\; \mathbb{E}\left[\sum_{i=1}^{n} X_i^2 + 2\sum_{i<j} X_i X_j\right]$

$$\left(\sum_{i=1}^{n} X_i\right)^2 = \sum_{i=1}^{n} X_i^2 + 2\sum_{i<j} X_i X_j$$

E.g. Take $n = 4$

$$(X_1 + X_2 + X_3 + X_4)^2 = \begin{matrix} X_1X_1 & X_1X_2 & X_1X_3 & X_1X_4 \\ X_2X_1 & X_2X_2 & X_2X_3 & X_2X_4 \\ X_3X_1 & X_3X_2 & X_3X_3 & X_3X_4 \\ X_4X_1 & X_4X_2 & X_4X_3 & X_4X_4 \end{matrix} = \begin{matrix} X_1^2 & X_1X_2 & X_1X_3 & X_1X_4 \\ X_2X_1 & X_2^2 & X_2X_3 & X_2X_4 \\ X_3X_1 & X_3X_2 & X_3^3 & X_3X_4 \\ X_4X_1 & X_4X_2 & X_4X_3 & X_4^4 \end{matrix}$$

$$= \sum_{i=1}^{4} X_i^2 + \begin{array}{c|c|c|c|} & X_1X_2 & X_1X_3 & X_1X_4 \\ X_2X_1 & & X_2X_3 & X_2X_4 \\ X_3X_1 & X_3X_2 & & X_3X_4 \\ X_4X_1 & X_4X_2 & X_4X_3 & \end{array}$$

$$= \sum_{i=1}^{4} X_i^2 + 2 \cdot \begin{array}{|c|c|c|} X_1X_2 & X_1X_3 & X_1X_4 \\ & X_2X_3 & X_2X_4 \\ & & X_3X_4 \\ & & \end{array} = \sum_{i=1}^{4} X_i^2 + \sum_{i<j} X_i X_j$$

Binomial distribution $X \sim$ Bin(θ), find $\mathbb{E}[X^2]$ and $\mathbb{V}[X]$

- We have $\mathbb{E}[X^2] = \mathbb{E}\left[\sum_{i=1}^{n} X_i^2 + 2\sum_{i<j} X_i X_j\right] \overset{\text{expectation is linear}}{=} \sum_{i=1}^{n} \mathbb{E}[X_i^2] + 2\sum_{i<j} \mathbb{E}[X_i X_j]$

- Recall $Y \sim$ Ber(θ) then $Y^2 = Y$. Therefore $\mathbb{E}[X^2] = \sum_{i=1}^{n} \mathbb{E}[X_i^1] + 2\sum_{i<j} \mathbb{E}[X_i X_j]$

- Recall that Binomial random variable is defined as the sum of n independent Bernoulli variable, so X_i, X_j are independent, therefore $\mathbb{E}[X^2] = \sum_{i=1}^{n} \mathbb{E}[X_i] + 2\sum_{i<j} \mathbb{E}[X_i]\mathbb{E}[X_j]$

- Recall $Y \sim$ Ber(θ) then $\mathbb{E}[Y] = \theta$. Therefore $\mathbb{E}[X^2] = \sum_{i=1}^{n} \theta + 2\sum_{i<j} \theta\theta = \sum_{i=1}^{n} \theta + 2\sum_{i<j} \theta^2$

- first sum has n terms, second sum has $\frac{n(n-1)}{2}$ terms: $\mathbb{E}[X^2] = n\theta + 2\frac{n(n-1)}{2}\theta^2 = n\theta + n(n-1)\theta^2$

- Recall $\mathbb{V}[X] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$
  $\mathbb{V}[X] = n\theta + n(n-1)\theta^2 - (n\theta)^2 = n\theta(1-\theta)$

E.g. Gambling

- Bernoulli: either win (W) with probability θ or lose (L) with probability $1-\theta$.

- Binomial: you gamble 5 times, find the probability of winning at most 2 times (including 2)

$$\binom{5}{0}\theta^0(1-\theta)^5 + \binom{5}{1}\theta^1(1-\theta)^4 + \binom{5}{2}\theta^2(1-\theta)^3$$

- Suppose $\begin{cases} +40 & W \\ -10 & L \end{cases}$ and $\theta = 0.5$ (fair), find the expected return for at most 2 W in the 5 gambling bets.
  - Recall that $\mathbb{E}[X] = \sum_{x\in\Omega} \mathbb{P}(X = x|\theta)x$
  - For 5L, $x = -50$
  - For 1W 4L, $x = 50 + 4(-10) = 10$
  - For 2W 3L, $x = 2(50) + 3(-10) = 70$

$$\binom{5}{0}0.5^0 0.5^5(-50) + \binom{5}{1}0.5^1 0.5^4(10) + \binom{5}{2}0.5^2 0.5^3(70)$$

Trinomial and multinomial distribution

- **Trinomial distribution**
  Possible out come: $\{1, 2, 3\}$ with probability $\{p_1, p_2, p_3\}$

$$p(n_1, n_2, n_3 \,|\, p_1, p_2, p_3) = \binom{n_1 + n_2 + n_3}{n_1, n_2, n_3} p_1^{n_1} p_2^{n_2} p_3^{n_3}.$$

- **E.g.** Human have four gene types $\{A, T, C, G\}$ with occurrence probability p_A, p_T, p_C, p_G.
  In a length-5 string, what is the probability the string is $ATCGA$?

$$\binom{5}{2, 1, 1, 1} p_A^2 p_T p_C p_G.$$

E.g. Multinomial in gambling

- Outcome: big win (+50), win (+10), lose (-10), super big lose (-100)

- Probability: $p_{\text{BigWin}} = 0.1, p_{\text{Win}} = 0.4, p_{\text{Lose}} = 0.45, p_{\text{superBigLose}} = 0.05$

- Find the expected return of at most 2 win (big win or win both ok) if you bet 5 times.

- Let $X=$ num big win, $Y =$ num win, $Z =$ num lose, $W =$ num super big lose

$$\mathbb{P}(X, Y, Z, W) = \binom{n}{X, Y, Z, W} p_{\text{BigWin}}^X p_{\text{Win}}^Y p_{\text{Lose}}^Z p_{\text{superBigLose}}^W = \binom{5}{X, Y, Z, W} 0.1^X 0.4^Y 0.45^Z 0.05^W$$

$$x = X \cdot 50 + Y \cdot 10 + Z \cdot (-10) + W \cdot (-100) = 10(5X + Y - Z - 10W)$$

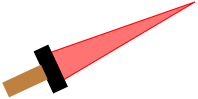
- Expected return is the sum of $\binom{5}{X,Y,Z,W} 0.1^X 0.4^Y 0.45^Z 0.05^W 10(5X + Y - Z - 10W)$ for all "2 wins"

Uniform distribution $X \sim U(a, b)$

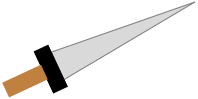
- Multiple outcomes but all are equally likely
- Uniform = evenly distributed
- $\mathbb{P}(X = k) = \dfrac{1}{n}$ for all $k \in \{a, a+1, \ldots, b\}$
- $\mathbb{E}[X] = \dfrac{a+b}{2}$
- $\mathbb{V}[X] = \dfrac{(b-a+1)^2 - 1}{12}$
- **E.g.** A fair die follows uniform distribution
  - Do you remember that for a 6-sided fair die, its expected value is 3.5
  - Here we have $\mathbb{E}[X] = \dfrac{n+1}{2} = \dfrac{1+6}{2} = \dfrac{7}{2}$
- **E.g.** Random permutation follows uniform distribution.

RPG (Role Playing Game) follows uniform distribution $X \sim U(n)$

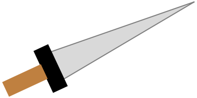
Which weapon is better?



The Epic Excalibur of Externality Covered by Prismatic Dragon-blood
Physical Damage: 6-13.2
Attacks Per Second: 1.45
Critical Strike Chance: 8%
Every Third Strike Deals Triple Damage

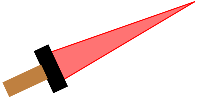


Normal Sword
Physical Damage: 10-41
Attacks Per Second: 1



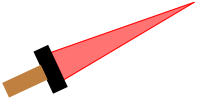
Physical Damage: 10-41
Attacks Per Second: 1

- Attacks Per Second is 1 so we can ignore that
- Let X denotes Damage Per Second
- $X \sim U([10, 41])$, meaning it is possible to take value $\{10, 11, \ldots, 41\}$
- To find the expected damage per second: $\mathbb{E}[X] = \dfrac{10 + 41}{2} = 25.5$
- Variance $\mathbb{V}[X] = \dfrac{(41 - 10 + 1)^2 - 1}{12} = \dfrac{41^2 - 1}{12} = 85.25$
  It means the fluctuation is huge: sometimes you get high number, sometimes you get small number



Physical Damage: 6-13.2
Attacks Per Second: 1.45
Critical Strike Chance: 8%
Every Third Strike Deals Triple Damage

- Critical Strike means double damage
- Ignore Triple Damage for a moment



Damage 8.7-19.14, 92%
Damage 17.4-38.28, 8%

- Let X be the damage per second
- $X = 0.92U([8.7, 19.14]) + 0.08U([17.4, 38.28])$
- $\mathbb{E}[X] = 0.92\frac{19.14 + 8.7}{2} + 0.08\frac{38.28 + 17.4}{2} = 15.0336$
- Now consider the "every 3rd strike deals triple damage", then

$$\frac{\mathbb{E}[X] + \mathbb{E}[X] + 3\mathbb{E}[X]}{3} \approx \frac{15 + 15 + 45}{3} = 25$$

Geometric distribution

$$p(k|\theta) = (1-\theta)^{k-1}\theta$$

- Geometric = out of k independent Bernoulli trials, all fails except the last (the kth) trail
- Memoryless: the probability of success in future trials is independent of the number of past failures
- $\mathbb{P}(X = k|\theta) = (1-\theta)^{k-1}\theta$ for all $k \in \{1, 2, \ldots, \infty\}$
- $\mathbb{E}[X] = \dfrac{1}{\theta}$
- $\mathbb{V}[X] = \dfrac{1-\theta}{\theta^2}$
- **E.g.** Gaming k times until seeing a success follows geometric distribution
- **E.g.** Phone-call for customer service: calling k times until a staff takes your call follows geometric distribution
- **E.g.** Einstein in school: seeing many students until you find an Einstein in school follows geometric distribution

Geometric distribution with $p(k|\theta) = (1-\theta)^{k-1}\theta$ has $\mathbb{E}[X] = \frac{1}{\theta}$

- Let X be a geometrically distributed random variable with PMF $p(k|\theta)$ under parameter θ.

- The expected value is

$$\mathbb{E}[X] \; = \; \sum_{k=1}^{\infty} k \cdot P(X=k) \; = \; \sum_{k=1}^{\infty} k \cdot (1-\theta)^{k-1}\theta \; = \; \theta \sum_{k=1}^{\infty} k(1-\theta)^{k-1}.$$

- (This step is probably WTF for some of you)
  Do you remember generating function $1 + 2x + 3x^2 + \ldots = \frac{\mathrm{d}}{\mathrm{d}x} \frac{1}{1-x} = \frac{1}{(1-x)^2}$ from combinatorics class?

$$\mathbb{E}[X] \; = \; \theta \sum_{k=1}^{\infty} k(1-\theta)^{k-1} \; = \; \theta \frac{1}{(1-(1-\theta))^2} \; = \; \theta \cdot \frac{1}{\theta^2} = \frac{1}{\theta}$$

Geometric distribution with $p(k|\theta) = (1-\theta)^{k-1}\theta$ has $\mathbb{V}[X] = \frac{1-\theta}{\theta^2}$

- Let X be a geometrically distributed random variable with PMF $p(k|\theta)$ under parameter θ.
- The variance $\mathbb{V}[X] = \mathbb{E}[X^2] - \Big(\mathbb{E}[X]\Big)^2 = \mathbb{E}[X^2] - \frac{1}{\theta^2}$
- Find $\mathbb{E}[X^2]$:

$$\mathbb{E}[X^2] \;=\; \sum_{k=1}^{\infty} k^2 (1-\theta)^{k-1}\theta \;=\; \theta \sum_{k=1}^{\infty} k^2 (1-\theta)^{k-1}$$

- Same trick in the same logic $\sum_{k=1}^{\infty} k^2 x^{k-1} = \frac{1+x}{(1-x)^3}$

$$\mathbb{E}[X^2] \;=\; \theta \sum_{k=1}^{\infty} k^2 (1-\theta)^{k-1} \;=\; \theta \frac{1+(1-\theta)}{\theta^3} = \frac{2-\theta}{\theta^3}$$

- Last step

$$\mathbb{V}[X] = \frac{2-\theta}{\theta^2} - \frac{1}{\theta^2} = \frac{1-\theta}{\theta^2}$$

Example of Geometric distribution with $p(k|\theta) = (1-\theta)^{k-1}\theta$

- Consider a fair coin with $\theta = 0.5$ for success

- **E.g.** Flip a fair coin $k = 1$ time and success at $k = 1$
  Success at $k = 1 \iff$ zero fail
  $\mathbb{P} = (1-0.5)^{1-1}(0.5) = 0.5$

- **E.g.** Flip a fair coin $k = 2$ times and success at $k = 2$
  $\mathbb{P} = (1-0.5)^{2-1}(0.5) = 0.25$
  All the possible outcome of 2 toss (S=success, F = fail)

$$\Omega(\text{2 tosses}) \;=\; \{FF, FS, SF, SS\}, \qquad \mathbb{P}(\text{success at 2}) = \frac{|\{FS\}|}{|\Omega|} = \frac{1}{4} = 0.25$$

- **E.g.** Flip a fair coin $k = 20$ times, consecutively failed first 19 times and success at $k = 20$

$$\mathbb{P} = (1-0.5)^{19} 0.5$$

  "keep failing is very unlikely" $\implies$ working hard is useful

Negative binomial distribution

$$p(k|\theta) = \left\langle \begin{matrix} n \\ k \end{matrix} \right\rangle (1-\theta)^n \theta^k$$

- Negative binomial = out of n independent Bernoulli events, exactly k fails
- $\left\langle \begin{matrix} n \\ k \end{matrix} \right\rangle$:
  - number of ways to achieve k fails in n trials, where the order of successes and failures matters
  - this is to find the number of ways to place k fail in n trials, with the remaining trials being success
- Negative binomial distribution = the sum of k independent geometric random variables with parameter θ. Therefore
  - $\mathbb{E}[X] = k\dfrac{1-\theta}{\theta}$
  - $\mathbb{V}[X] = k\dfrac{1-\theta}{\theta^2}$
- Binomial VS Negative Binomial next slide

Binomial VS Negative Binomial

- Both
  - are based on independent Bernoulli trials, let's call it i.B.t
  - have two numbers: n, k

- Binomial:
  - What you fixed: n i.B.t, the number of i.B.t
  - What you are looking for: the $\mathbb{P}$ of a certain number of success

- Negative Binomial:
  - What you fixed: k number of success
  - What you are looking for: the $\mathbb{P}$ of the number n of i.B.t needed so that you get k success

- On tossing fair coin
  - Binomial
    - nature: toss 6 times for you
    - you: want to know the $\mathbb{P}$ of getting exactly 4 heads
  - Negative Binomial:
    - nature: toss until you get 4 success
    - you: want to know the $\mathbb{P}$ of needing exactly 6 tosses to get those 4 success

Example of Negative binomial distribution

$$p(k|\theta) = \left\langle \begin{matrix} n \\ k \end{matrix} \right\rangle (1-\theta)^n \theta^k$$

- **E.g.** Toss a fair coin ($\theta = 0.5$ for success). Find the probability of 6 toss is need so that you get exactly 4 fails.

$$\left\langle \begin{matrix} 6 \\ 4 \end{matrix} \right\rangle (1-0.5)^4 0.5^2 = \begin{pmatrix} 6+4-1 \\ 4 \end{pmatrix} 0.5^4 0.5^2 = 15.6\%$$

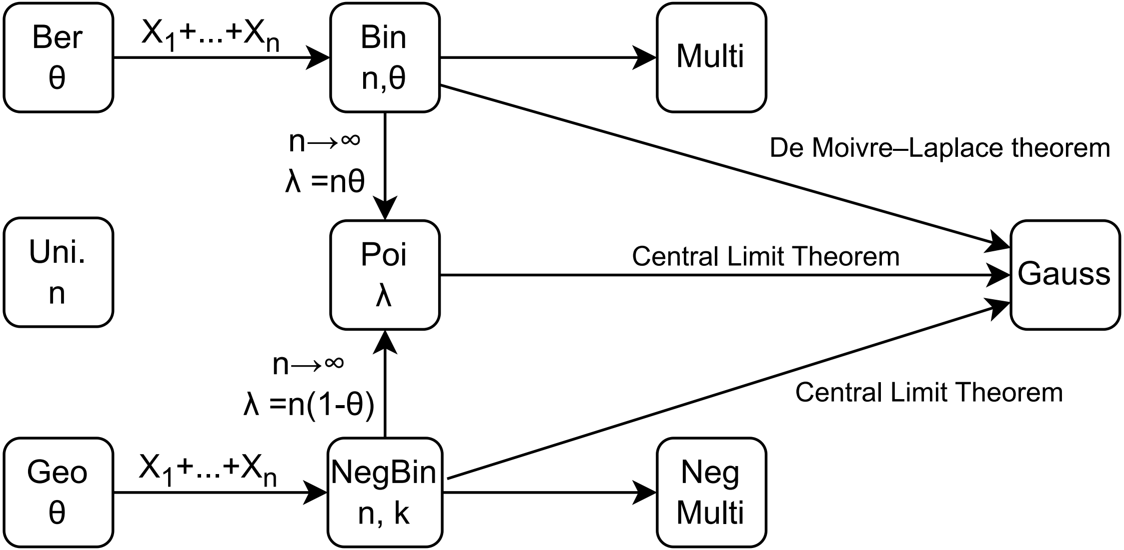
- **E.g.** Toss a fair coin ($\theta = 0.5$ for success). Find the probability of 6 toss is need so that you get at least 4 fails.

$$\left\langle \begin{matrix} 6 \\ 4 \end{matrix} \right\rangle (1-0.5)^4 0.5^2 + \left\langle \begin{matrix} 6 \\ 5 \end{matrix} \right\rangle (1-0.5)^5 0.5^1 + \left\langle \begin{matrix} 6 \\ 6 \end{matrix} \right\rangle (1-0.5)^6 0.5^0$$

Poisson distribution

$$\mathbb{P}(X = k|\lambda) = \frac{\lambda^k e^{-\lambda}}{k!}$$

- Poisson = number of events in a fixed interval of time or space
- Events occur with a known constant average rate λ and independently of the time since the last event
- $\mathbb{P}(X = k|\lambda) = \frac{\lambda^k e^{-\lambda}}{k!}$ for all $k \in 0, 1, 2, \ldots$
- $\mathbb{E}[X] = \mathbb{V}[X] = \lambda$
- **E.g.** Number of emails received in an hour follows a Poisson distribution
- **E.g.** Number of bus arrival at a station in a day follows a Poisson distribution
- **E.g.** Number of decay events per unit time from a radioactive source follows a Poisson distribution


Ber
θ
X1+...+Xn
Bin
n,θ
Multi
n→∞
λ =nθ
De Moivre–Laplace theorem
Uni.
n
Poi
λ
Central Limit Theorem
Gauss
n→∞
λ =n(1-θ)
Central Limit Theorem
Geo
θ
X1+...+Xn
NegBin
n, k
Neg
Multi

Summary

- $(\Omega, E, \mathbb{P})$ and three axioms: $\mathbb{P}(E) \geq 0$, $\mathbb{P}(\Omega) \equiv 1$ and $\mathbb{P}\Big(\bigcup_i E_i\Big) = \sum_i \mathbb{P}(E_i)$ if E_i are disjoint
- Complementary event $E^c \coloneqq \Omega \setminus E$ and $\mathbb{P}(E^c) = 1 - \mathbb{P}(E)$
- Disjoint / Mutually exclusive event
  - A, B mutually exclusive $\iff \mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$
  - A, B not mutually exclusive $\iff \mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$
- $\mathbb{P}(X = x, Y = y)$ Joint probability
- $\mathbb{P}(X = x) = \sum_{y \in \mathcal{Y}} \mathbb{P}(X = x, Y = y)$ Marginal probability
- $\mathbb{P}(X = x | Y = y) = \dfrac{\mathbb{P}(X = x, Y = y)}{\mathbb{P}(Y = y)}, \quad \mathbb{P}(Y = y) > 0$ Conditional probability
- For laziness we write $\mathbb{P}(X = x) = p(x)$
- Expectation $\mathbb{E}[f(X)] \coloneqq \sum_{x \in \mathcal{X}} f(x)p(x)$
  - $\mathbb{E}[aX + bY + c] = a\mathbb{E}[X] + b\mathbb{E}[Y] + c$
  - Conditional expectation $\mathbb{E}[X|Y]$
  - Marginal expectation $\mathbb{E}[X]$
  - Joint expectation $\mathbb{E}[X, Y]$
- $\mathbb{V}[X] = \mathbb{E}\Big[\big(X - \mathbb{E}[X]\big)^2\Big] = \mathbb{V}[X] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$
- cov$(X, Y) = \mathbb{E}\Big[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])\Big] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$
- Ber, Bin, Uni, Geo, NegBin, Poi

Contents

Continuous parametric distributions Not in exam

- We denote $p(x|\theta), x \in \mathcal{X}, \theta \in \Theta$
  - θ is the parameter
  - Θ is the set of valid parameter
  - by changing θ we change the distribution

- Gaussian distribution $X \sim \mathcal{N}(\mu, \sigma),\ p(x|\mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp\Big(-\frac{(x-\mu)^2}{2\sigma^2}\Big),\ \sigma > 0$
  - Standard normal distribution $\mu = 0, \sigma = 1$, we call such X standard score, denoted as Z

- Uniform distribution $p(k|a, b) = \frac{1}{b-a+1}, b \geq a$

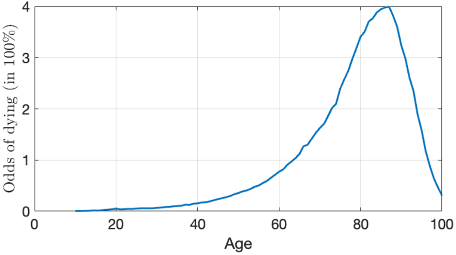
- Central limit theorem

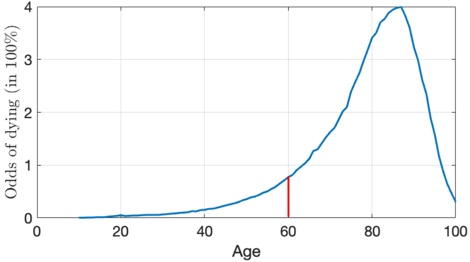
- Beta distribution $\frac{x^{\alpha-1}(1-x)^{\beta-1}}{\int_0^1 u^{\alpha-1}(1-u)^{\beta-1}du}$

- Marchenko-Pastur distribution

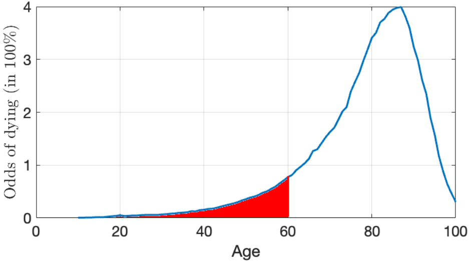
Distributions and cumulative function

Not in exam

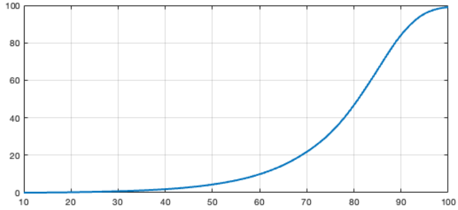




$\mathbb{P}(X = 60)$ probability you die exactly at age 60



$\mathbb{P}(X \leq 60)$ probability you die before (including) 60



Cumulative distribution $\int_{-\infty}^{x} p(x)dx$ or $\sum_{-\infty}^{x} p(x)$

Application: max likelihood estimation of Poisson model of covid Not in exam

- Poisson distribution

$$p(k|\lambda) = \frac{\lambda^k e^{-\lambda}}{k!}, \lambda \geq 0, k \in \mathbb{N}$$

- $p(0|\lambda) =$ probability of recovery at the same day of getting covid

- $p(1|\lambda) =$ probability of recovery after 1 day of getting covid

- $p(2|\lambda) =$ probability of recovery after 2 days of getting covid

- How do we now the model λ? We learn it from data by fitting.

- Suppose we are giving a record of days people recover as $[15, 11, 28, 38, 18, ...]$, i.e.,
  - 1st subject recovered after 15 days, $k_1 = 15$
  - 2nd subject recovered after 11 days, $k_2 = 11$
  - and so on

  So you are now given

$$\frac{\lambda^{15} e^{-\lambda}}{15!}, \frac{\lambda^{11} e^{-\lambda}}{11!}, \frac{\lambda^{28} e^{-\lambda}}{28!}, \ldots$$

  and you want to find λ that maximize these probabilities

Maximum likelihood estimation of Poisson model of covid **Not in exam**

- Poisson distribution
$$p(k|\lambda) = \frac{\lambda^k e^{-\lambda}}{k!}, \lambda \geq 0, k \in \mathbb{N}$$

- Given n observation / data / measurement of $k_1, k_2, \ldots, k_n$.

- The probability of all these event occur under a parameter λ is
$$\frac{\lambda_1^k e^{-\lambda}}{k_1!} \cdot \frac{\lambda_2^k e^{-\lambda}}{k_2!} \cdots \frac{\lambda_n^k e^{-\lambda}}{k_n!} =: \prod_{i=1}^{n} \frac{\lambda_i^k e^{-\lambda}}{k_i!} = L(\lambda|k_1, k_2, \cdots, k_N)$$
and you want to find λ that maximize this probability L known as likelihood.

- The λ that makes such likelihood most likely to occur
$$\max L(\lambda|k_1, k_2, \cdots, k_N) = \max \prod_{i=1}^{n} \frac{\lambda_i^k e^{-\lambda}}{k_i!}$$
where $\max$ stands for "maximize"

- **Due to mathematical reason**, we work on the negative log of L
$$\max \prod_{i=1}^{n} \frac{\lambda_i^k e^{-\lambda}}{k_i!} = \min -\log \prod_{i=1}^{n} \frac{\lambda_i^k e^{-\lambda}}{k_i!}$$

Not in exam

$$f(\lambda) := -\log \prod_{i=1}^{n} \frac{\lambda_i^k e^{-\lambda}}{k_i!} \quad = \quad -\log \frac{\lambda_1^k \lambda_2^k \cdots \lambda_n^k \underbrace{e^{-\lambda} e^{-\lambda} \cdots e^{-\lambda}}_{n \text{ times}}}{k_1! k_2! \cdots k_n!}$$

$$= \quad -\log \frac{\lambda^{k_1+k_2+\cdots+k_n} e^{-n\lambda}}{k_1! k_2! \cdots k_n!}$$

$$= \quad -\log\left(\lambda^{k_1+k_2+\cdots+k_n}\right) - \log\left(e^{-n\lambda}\right) + \log\left(k_1! k_2! \cdots k_n!\right)$$

$$= \quad -\Big(k_1 + k_2 + \cdots + k_n\Big) \log(\lambda) + n\lambda + \Big(\log k_1! + \log k_2! + \cdots + \log k_n!\Big)$$

Calculus 101: to find the extreme point of a function f, take derivative to zero

$$\frac{df}{d\lambda} \; = \; -\frac{k_1 + k_2 + \cdots + k_n}{\lambda} + n + 0 \; = \; 0 \quad \Longrightarrow \quad \lambda = \frac{k_1 + k_2 + \cdots + k_n}{n}$$

We usually denote such λ as $\hat{\lambda}_{\text{MLE}}$, stands for maximum likelihood estimate

Summary of MLE Poisson model of covid

Not in exam

- Giving a record of days n people recover as $[k_1, k_2, k_3, ...] = [15, 11, 28, ...]$
- You assume the recovery follows a Poisson model $p(k|\lambda) = \dfrac{\lambda^k e^{-\lambda}}{k!}, \lambda \geq 0, k \in \mathbb{N}$
- We need to estimate the parameter λ to use the model
  - How: we take $\hat{\lambda}_{\text{MLE}} = \dfrac{k_1 + k_2 + \cdots + k_n}{n}$
- Now we have $p(k|\hat{\lambda}_{\text{MLE}}) = \dfrac{\hat{\lambda}_{\text{MLE}}^k e^{-\hat{\lambda}_{\text{MLE}}}}{k!}$
- Now suppose a person get covid,
  - he wants to know the probability that he will recover after 1 day, he calculate $p(1|\hat{\lambda}_{\text{MLE}}) = \dfrac{\hat{\lambda}_{\text{MLE}}^1 e^{-\hat{\lambda}_{\text{MLE}}}}{1!}$
  - he wants to know the probability that he will recover after 10 days, he calculate $p(10|\hat{\lambda}_{\text{MLE}}) = \dfrac{\hat{\lambda}_{\text{MLE}}^{10} e^{-\hat{\lambda}_{\text{MLE}}}}{10!}$
- Same model for bus-waiting

Anscombe's quartet: 4 sets of data

Not in exam

Francis Anscombe (1918 - 2001)
An English statistician.

I		II		III		IV	
x	y	x	y	x	y	x	y
10	8,04	10	9,14	10	7,46	8	6,58
8	6,95	8	8,14	8	6,77	8	5,76
13	7,58	13	8,74	13	12,74	8	7,71
9	8,81	9	8,77	9	7,11	8	8,84
11	8,33	11	9,26	11	7,81	8	8,47
14	9,96	14	8,1	14	8,84	8	7,04
6	7,24	6	6,13	6	6,08	8	5,25
4	4,26	4	3,1	4	5,39	19	12,5
12	10,84	12	9,13	12	8,15	8	5,56
7	4,82	7	7,26	7	6,42	8	7,91
5	5,68	5	4,74	5	5,73	8	6,89

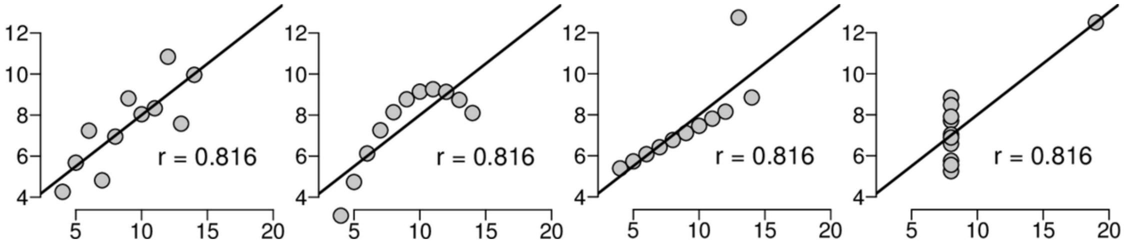
- Same number of data points: $n = 11$
- Same sum: $\sum x = 99$, $\sum y = 82.51$
- Same mean: $\mathbb{E}[X] = 90$, $\mathbb{E}[Y] = 7.5$
- Same variance: $\mathbb{V}[X] = 11.0224$, $\mathbb{V}[Y] = 4.1209$
- Same std: $\sqrt{\mathbb{V}[X]} = 3.32$, $\sqrt{\mathbb{V}[Y]} = 2.03$

- Same equation of regression $Y = 3 + 0.5X$
- Same standard error of estimate of slope $= 0.118$
- Same sum of squares $X - \bar{X} = 110$
- Same residual sum of squares of $Y = 13.75$
- Same correlation coefficient $= 0.82$
- Same $r^2 = 0.67$

Can you tell they are the same distribution?

Anscombe's quartet: statistics is not enough

Not in exam



ok | wrong model | effect of outlier | wrong model & outlier

This is why { data visualization / machine learning } are important. They can avoid these.

Information Theory and Communication Theory **Not in exam**

- How do we measure the "amount of information" in a sentence?
- Entropy $\mathbb{E}[-\log p(x)]$
- Source entropy
- Channel capacity
- Fundamental limit of data compression
- Fundamental limit of communication
- Fundamental limit of cryptography

Random graph = Graph theory $\oplus$ probability **Not in exam**

- Erdos–Renyi Model: Each edge included with probability p
  Applications: Transition from sparse to dense graphs

- Stochastic Block Models: Nodes partitioned into blocks with edge probabilities based on block membership
  Applications: Community detection in networks

- Turing Award 2023