L1.1 **Probability** — Course notes: James Norris, Douglas Kennedy, Richard Weber
Website: vittoria silvestri

Let $\Omega = \{w_1, w_2, \dots\}$

Ω: sample space; $w_1, w_2, \dots$: possible outcomes

$A \subseteq \Omega$ is called an event

EX: roll a die $\Omega = \{1,2,3,4,5,6\}$ $\quad |\Omega| = 6$

$A_1 = \{1,2,3\} = \{$the outcome is at most 3$\}$

$A_2 = \{2,4,6\} = \{$the outcome is even$\}$

$A_3 = \Omega \setminus \{1\} = \{$the outcome is not 1$\}$

EX: draw a card from a standard deck

$\Omega = \{\text{cards}\}$ $\quad |\Omega| = 52$

$A_1 = \{$the card is a Jack$\} \Rightarrow |A_1| = 4$

$A_2 = \{$the card is a diamond$\} \Rightarrow |A_2| = 13$

$A_3 = \{$the card is not QS$\} \Rightarrow |A_3| = 51$

EX: pick a natural number

$\Omega = \mathbb{N}$

$A_1 = \{$the outcome is at most 5$\} = \{0,1,2,3,4,5\}$

$A_2 = \{$the outcome is odd$\}$

$A_3 = \{$the outcome is not 7$\} = \mathbb{N} \setminus \{7\}$

EX: pick a number in $[0,1]$

$\Omega = [0,1]$, uncountable

Events: $A_1 = \{x : x < \frac{1}{3}\} = [0, \frac{1}{3})$

$A_2 = \{x : x \neq 0.7\} = [0,1] \setminus \{0.7\}$

$A_3 = \{x : x = 2^{-n} \text{ for some } n \in \mathbb{N}\} = \{1, \frac{1}{2}, \frac{1}{4}, \dots\}$

L1.2

Remark: for first part of course, only looking at finite or countable sample spaces

Definition (Probability space)

Let Ω be any set.

Let $\mathcal{F}$ be a set of subsets of Ω.

$\mathcal{F}$ is a <u>σ-algebra</u> if:

1) $\Omega \in \mathcal{F}$
2) if $A \in \mathcal{F}$, then $A^c \in \mathcal{F}$
3) if $(A_n)_{n \geq 1}$ is a sequence of elements of $\mathcal{F}$, then $\bigcup_{n \geq 1} A_n \in \mathcal{F}$, i.e. "$\mathcal{F}$ closed under countable unions"

Let $\mathbb{P} : \mathcal{F} \longrightarrow [0,1]$.

$\mathbb{P}$ is a <u>probability measure</u> if

1) $\mathbb{P}(\Omega) = 1$
2) for any collection $(A_n)_{n \geq 1}$ of disjoint events in $\mathcal{F}$,
$$\mathbb{P}\Big(\bigcup_{n \geq 1} A_n\Big) = \sum_{n \geq 1} \mathbb{P}(A_n)$$

The triple $(\Omega, \mathcal{F}, \mathbb{P})$ is called <u>probability space</u>.

<u>Remark:</u> When Ω is countable, we will be taking $\mathcal{F}$ to be the set of all subsets of Ω, unless otherwise specified.

We will refer to $\mathbb{P}(A)$ as the probability of A.

<u>Equally likely outcomes</u>

NOT EMPTY!

Let Ω be a finite set $\Omega = \{\omega_1, \omega_2, \dots, \omega_{|\Omega|}\}$.

Let $\mathbb{P} : \mathcal{F} \to [0,1]$ be defined by $\mathbb{P}(A) = |A| / |\Omega|$ for all $A \in \mathcal{F} = \mathcal{P}\Omega$

In particular, $\mathbb{P}(\{\omega_i\}) = 1/|\Omega| \;\; \forall\, \omega_i \in \Omega$

L1.3 Note that:

1) $\mathbb{P}(\Omega) = |\Omega|/|\Omega| = 1$ ✓

2) Take $A_1, A_2, \dots, A_n$ disjoint subsets of Ω, then

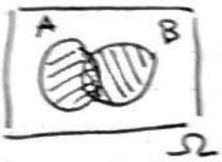
$$\mathbb{P}\left(\bigcup_{k=1}^{n} A_k\right) = \frac{\left|\bigcup_{k=1}^{n} A_k\right|}{|\Omega|} = \frac{\sum_{k=1}^{n} |A_k|}{|\Omega|} = \sum_{k=1}^{n} \underbrace{\frac{|A_k|}{|\Omega|}}_{\mathbb{P}(A_k)}$$

You can make the sequence infinite by adding $A_1, A_2, \dots, A_n, \emptyset, \emptyset, \dots$ ← indeed most general

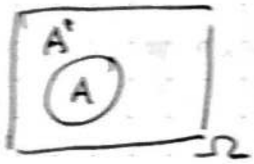
$$\mathbb{P}(\emptyset) = \frac{|\emptyset|}{|\Omega|} = \frac{0}{|\Omega|} = 0$$

if $A \subseteq B$ then $\mathbb{P}(A) = \dfrac{|A|}{|\Omega|} \leq \dfrac{|B|}{|\Omega|} = \mathbb{P}(B)$

$$\mathbb{P}(A \cup B) = \frac{|A \cup B|}{|\Omega|} = \frac{|A| + |B| - |A \cap B|}{|\Omega|} = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$$



$$\mathbb{P}(A^c) = \frac{|A^c|}{|\Omega|} = \frac{|\Omega| - |A|}{|\Omega|} = 1 - \mathbb{P}(A)$$



EX: roll a die, $\Omega = \{1,2,3,4,5,6\}$

$$\mathbb{P}\left(\begin{smallmatrix}\text{outcome}\\ \text{is even}\end{smallmatrix}\right) = \frac{|\{2,4,6\}|}{6} = \frac{1}{2}$$

$$\mathbb{P}\left(\begin{smallmatrix}\text{outcome}\\ \text{at most}\end{smallmatrix}\ 2\right) = \frac{|\{1,2\}|}{6} = \frac{1}{3}$$

EX: (Largest digit) Consider a string of random digits $0, 1, \dots, 9$ of length $n \geq 2$.

For $0 \leq k \leq 9$, what is the probability that the largest digit is at most k?

$\Omega = \{0,1,\dots,9\}^n \Rightarrow 10^n = |\Omega|$

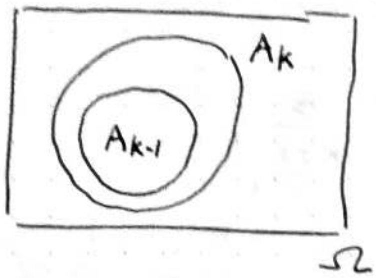
$A_k = \{$ strings with largest digit at mos $k \}$

$|A_k| = (k+1)^n$

$\mathbb{P}(A_k) = |A_k| / |\Omega| = (k+1/10)^n$

L1.4

$B_k = \{$ strings with largest digit exactly $k \}$



$$B_k = A_k \setminus A_{k-1}$$

$$|B_k| = |A_k| - |A_{k-1}|$$
$$= (k+1)^n - k^n$$

$$\mathbb{P}(B_k) = \frac{|B_k|}{|\Omega|} = \frac{(k+1)^n - k^n}{10^n}$$

RECAP

Probability space $(\Omega, \mathcal{F}, \mathbb{P})$

- Ω: set
- $\mathcal{F}$: σ-algebra
  1) $\Omega \in \mathcal{F}$
  2) $A \in \mathcal{F} \Rightarrow A^c \in \mathcal{F}$
  3) $(A_n)_{n \geq 1}$ seq in $\mathcal{F}$ $\Rightarrow \bigcup_{n \geq 1} A_n \in \mathcal{F}$
- $\mathbb{P}$: proba measure $\mathbb{P} : \mathcal{F} \to [0,1]$
  1) $\mathbb{P}(\Omega) = 1$
  2) $(A_n)_{n \geq 1}$ disjoint $\Rightarrow \mathbb{P}(\bigcup_{n \geq 1} A_n) = \sum_{n \geq 1} \mathbb{P}(A_n)$

1) Equally likely outcomes, $|\Omega| < \infty$, $\mathbb{P}(A) = \frac{|A|}{|\Omega|}$ $\forall A \in \mathcal{P}\Omega$

COMBINATORICS

Counting methods

1) **Multiplication rule**

Consider N finite sets $\Omega_1, \Omega_2, \dots, \Omega_N$ with cardinalities $n_1, n_2, \dots, n_N$.

Want to count $|\{(x_1, x_2, \dots, x_N) : x_k \in \Omega_k\}|$

$|\Omega_1| = n_1$

$$|\Omega_1 \times \Omega_2| = |\{(x_1, x_2) : x_1 \in \Omega_1, x_2 \in \Omega_2\}| = n_1 n_2$$

$$|\Omega_1 \times \Omega_2 \times \Omega_3| = n_1 n_2 n_3$$

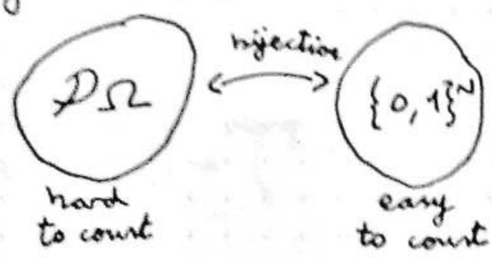
$\vdots$

$$|\Omega_1 \times \dots \times \Omega_N| = n_1 n_2 \cdots n_N$$ « Multiplication rule »

EX: Number of subsets of a finite set

Let $\Omega = \{\omega_1, \dots, \omega_N\}$, $|\Omega| = N < \infty$.

How many subsets?



If $A = \{\omega_1, \omega_N\}$, send to $(1, 0, 0, \dots, 0, 1)$.

Then since $|\{0,1\}^N| = 2^N$, we conclude $|\mathcal{P}\Omega| = 2^N$.

2) Permutations

Q: How many different orderings of n elements are there?

$$n(n-1)(n-2)\cdots 2 \cdot 1 = n!$$

EX: In how many different ways can we shuffle 52 cards? $52!$

3) Subsets of size k

How many different ways are there of picking k elements from n?

- With order: $n(n-1)\cdots(n-k+1) = n!/(n-k)!$

$$\underbrace{|\,|\,|\,|\,|\,|\,|}_{k}\;\underbrace{|\,|\,|}_{n-k} \qquad \frac{n!}{(n-k)!}$$

- Without order:

$$\underbrace{|\,|\,|\,|\,|\,|\,|}_{k}\;\underbrace{|\,|\,|\,|\,|}_{n-k} \qquad \frac{n!}{(n-k)!\,k!} = \binom{n}{k}$$ "Binomial coefficient"

EX: Show that $\sum_{k=0}^{n} \binom{n}{k} = 2^n$

$$\overbrace{\underbrace{|\,|\,|}_{n_1}\;\underbrace{|\,|\,|}_{n_2}\;\underbrace{|\,|\,|}_{n_k}}^{n}$$

Divide n elements into subsets of cardinalities $n_1, n_2, \dots, n_k$ with $n_1 + n_2 + \dots + n_k = n$.

$$\frac{n!}{n_1!\, n_2! \cdots n_k!} = \binom{n}{n_1 \cdots n_k}$$ "multinomial coefficient"

Subsets with repetition

Q: In how many different ways can we pick k elements from n, allowing for repetition?

- with order $\underbrace{n \cdot n \cdots n}_{k} = n^k$

L2.3

- <u>without order</u> Idea: build a bijection with a set of known size

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

$A = \{2, 1, 2\} \longmapsto * \mid * * \mid \mid$

$B = \{1, 2, 4\} \longmapsto * \mid * \mid \mid *$

$$\binom{k+n-1}{k}$$

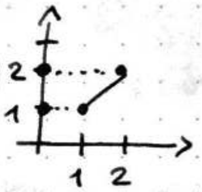
subsets with rep. without order $\longleftrightarrow$ strings of length $k+n-1$ with k stars

hard — easy

EX: How many strictly increasing functions are there from $\{1, \dots, k\}$ to $\{1, \dots, n\}$? $n \geq k$

$f : \{1, \dots, k\} \longrightarrow \{1, \dots, n\}$

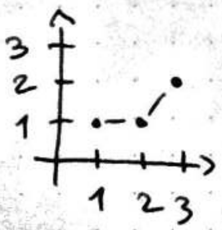
A: incr. functions $\longleftrightarrow$ B: subsets of size k of $\{1, \dots, n\}$



Now, $|B| = \binom{n}{k} \Rightarrow |A| = \binom{n}{k}$

Non-decreasing functions are harder:

$f : \{1, \dots, k\} \longrightarrow \{1, \dots, n\}$



A: non-dec. functions f $\longleftrightarrow$ B: inc functions $\{1, \dots, k\} \to \{1, \dots, n+k-1\}$ g

take f to $f+n$, i.e. $g(k) = f(k) + k - 1$

so we use $|B| = \binom{n+k-1}{k} \Rightarrow |A| = \binom{n+k-1}{k}$

§3.3 Asymptotic behaviour of $n!$

Recall that we write

$$a_n \sim b_n$$

if $\frac{a_n}{b_n} \longrightarrow 1$ as $n \longrightarrow \infty$.

Stirling's formula: $n! \sim \sqrt{2\pi n}\; n^n e^{-n}$ (1)

Note that this implies $\log(n!) \sim \log(\sqrt{2\pi n}\; n^n e^{-n})$ (2)

But the converse is harder.

Proof of (2) Write $l_n := \log(n!) = \sum_{i=2}^{n} \log i$

- Write $\lfloor x \rfloor$ for the lower integer part of x

$$\log \lfloor x \rfloor \leq \log x \leq \log(\lfloor x \rfloor + 1)$$

integrate over the interval $[1,n]$ to get

$$\underbrace{\int_1^n \log\lfloor x \rfloor \, dx}_{\substack{\log((n-1)!) \\ = l_{n-1}}} \leq \int_1^n \log x \, dx \leq \underbrace{\int_1^n \log(\lfloor x \rfloor + 1)\, dx}_{\substack{\log(n!) \\ = l_n}}$$

- This implies $\underbrace{\int_1^n \log x \, dx}_{n\log n - n + 1} \leq l_n \leq \underbrace{\int_1^{n+1} \log x \, dx}_{(n+1)\log(n+1) - n}$

$$\frac{n\log n - n + 1}{n \log n} \to 1 \qquad \frac{(n+1)\log(n+1) - n}{n\log n} \to 1$$

$$\therefore \frac{l_n}{n\log n} \longrightarrow 1$$

-

Proof of ①: Note that

$$\int_a^b f(x)\,dx = \frac{f(a)+f(b)}{2}(b-a) + \frac{1}{2}\int_a^b (x-a)(x-b) f''(x)\,dx$$

Take $f(x) = \log x$, to get that for each $k \geq 1$

$$\int_k^{k+1} \log x\,dx = \frac{\log k + \log(k+1)}{2} + \underbrace{\frac{1}{2}\int_k^{k+1} (x-k)(k+1-x)\frac{1}{x^2}\,dx}_{a_k = \frac{1}{2}\int_0^1 x(1-x)\cdot\frac{1}{(x+k)^2}\,dx}$$

Sum over $k=1$ to $n-1$ to get

$$\int_1^n \log x\,dx = \underbrace{\frac{1}{2}\log((n-1)!) + \frac{1}{2}\log(n!)}_{\log(n!) - \frac{1}{2}\log n} + \sum_{k=1}^{n-1} a_k$$

$$n\log n - n + 1 = \log(n!) - \frac{1}{2}\log n + \sum_{k=1}^{n-1} a_k$$

$$\log(n!) = n\log n - n + 1 + \frac{1}{2}\log n - \sum_{k=1}^{n-1} a_k$$

$$= \left(n+\frac{1}{2}\right)\log n - n + 1 - \sum_{k=1}^{n-1} a_k$$

apply exp to get

$$n! = n^{n+\frac{1}{2}} e^{-n} \cdot \exp\left(1 - \sum_{k=1}^{n-1} a_k\right)$$

Note that $\sum_{k=1}^{\infty} a_k < \infty$ since

$$a_k = \frac{1}{2}\int_0^1 x(1-x)\frac{1}{(x+k)^2}\,dx \leq \frac{1}{2}\cdot\frac{1}{k^2}\int_0^1 x(1-x)\,dx = \frac{1}{12k^2}$$

hence let $A =$

$\exp\left(1 - \sum_{n=1}^{\infty} a_n\right)$.

4.8.2

Therefore we are left with

$$n! = n^{n+\frac{1}{2}} e^{-n} \cdot A \cdot \exp\Big(\sum_{k=n}^{\infty} a_k \Big)$$

Therefore $$\frac{n!}{n^{n+\frac{1}{2}} e^{-n} A} = \exp\Big(\sum_{k=n}^{\infty} a_k \Big) \xrightarrow[n\to\infty]{} 1$$

and so $$\boxed{n! \sim n^{n+\frac{1}{2}} e^{-n} A}$$

It remains to show that $A = \sqrt{2\pi}$.

Observe that

$$2^{-2n} \binom{2n}{n} \sim \frac{\sqrt{2}}{A\sqrt{n}}.$$

Ineed, $$2^{-2n}\binom{2n}{n} = 2^{-2n} \frac{(2n)!}{n! \cdot n!} \sim 2^{-2n} \frac{2^{2n+\frac{1}{2}} \cancel{n^{2n+\frac{1}{2}}} \cancel{e^{-2n}} \cancel{A}}{A^{\cancel{2}} n^{\cancel{2n+1}\ 1/2} \cancel{e^{-2n}}}$$

$$= \sqrt{2}/A\sqrt{n}$$

We now use that expressions like LHS come up when computing integrals of powers of cosine.

Let $$I_n = \int_0^{\pi/2} \cos^n \vartheta \, d\vartheta = \sin\vartheta \cos^{n-1}\vartheta \Big|_0^{\pi/2}$$

	D	I
+	$\cos^{n-1}\vartheta$	$\cos\vartheta$
−	$-(n-1)\cos^{n-2}\vartheta \cdot \sin\vartheta$	$\sin\vartheta$
+		

$$+ \int_0^{\pi/2} \sin^2\vartheta \cos^{n-2}\vartheta \, d\vartheta$$

$$= I$$

$$I_n = \frac{n-1}{n} I_{n-2}, \quad I_0 = \frac{\pi}{2}, \quad I_1 = 1$$

$$I_{2n} = \frac{2n-1}{2n} \cdot \frac{2n-3}{2n-2} \cdots \frac{3}{4} \cdot \frac{1}{2} I_0 = \frac{(2n)!}{(2^n n!)^2} \cdot \frac{\pi}{2}$$

$$= 2^{-2}\binom{2n}{n} \cdot \frac{\pi}{2}$$

L3.4

$$I_{2n+1} = \frac{2n}{2n+1} \cdots \frac{4}{5} \cdot \frac{2}{3} \underbrace{I_1}_{1} = \left(2^{-2n}\binom{2n}{n}\right)^{-1} \cdot \frac{1}{2n+1}$$

But note that

$$\frac{I_{2n}}{I_{2n-2}} = \frac{2n-1}{2n} \xrightarrow[n\to\infty]{} 1$$

But $(I_n)_{n\geq 1}$ is decreasing sequence, therefore

$$\frac{I_{2n}}{I_{2n+1}} \xrightarrow[n\to\infty]{} 1 \quad \text{as well} \quad \text{via sandwich}$$

$$''$$

$$(2n+1)\left(2^{-2n}\binom{2n}{n}\right)^2 \cdot \frac{\pi}{2} \xrightarrow[n\to\infty]{} 1$$

$$\Rightarrow \left(2^{-2n}\binom{2n}{n}\right)^2 \sim \frac{2}{\pi(2n+1)} \sim \frac{1}{\pi n}$$

Therefore $2^{-2n}\binom{2n}{n} \sim \frac{1}{\sqrt{\pi n}}$.

But we knew that also

$$2^{-2n}\binom{2n}{n} \sim \frac{\sqrt{2}}{A\sqrt{n}}$$

$\therefore \quad A = \sqrt{2\pi} \quad$ as needed. $\square$

L 4.1

$(\Omega, \mathcal{F}, \mathbb{P})$

Ω: set, $\mathcal{F}$: σ-algebra, $\mathbb{P} : \mathcal{F} \to [0,1]$

1) $\mathbb{P}(\Omega) = 1$

2) If $(A_n)_{n \geq 1}$ is a sequence of disjoint events,

$$\mathbb{P}\left(\bigcup_{n=1}^{\infty} A_n \right) = \sum_{n=1}^{\infty} \mathbb{P}(A_n)$$

Stirling's formula

$$n! \sim \sqrt{2\pi n}\,(n/e)^n$$

Properties of probability measures

1) $\mathbb{P}(A) \in [0,1]$

2) if $A \cap B = \emptyset$ then $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$

indeed, take sequence $A, B, \emptyset, \ldots$ so that

$$\underbrace{\mathbb{P}\left(\bigcup_{n=1}^{\infty} A_n \right)}_{A \cup B} = \underbrace{\sum_{k=1}^{\infty} \mathbb{P}(A_k)}_{\mathbb{P}(A) + \mathbb{P}(B)}$$

3) take $\Omega, \emptyset$

$$\mathbb{P}(\Omega \cup \emptyset) = \underbrace{\mathbb{P}(\Omega)}_{1} + \sum_{n \geq 2} \mathbb{P}(\emptyset) \to 0$$

4) $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$

since $A \cup A^c = \Omega$

$A \cap A^c = \emptyset$

$$\underset{\substack{\| \\ 1}}{\mathbb{P}(\Omega)} = \mathbb{P}(A \cup A^c) = \mathbb{P}(A) + \mathbb{P}(A^c)$$

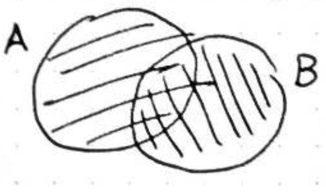
5) if $A \subseteq B$ then $\mathbb{P}(A) \leq \mathbb{P}(B)$

as $\mathbb{P}(B) = \mathbb{P}(A \cup (B \setminus A))$

$$= \mathbb{P}(A) + \underbrace{\mathbb{P}(B \setminus A)}_{\geq 0}$$

L4.2

6) $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$



In particular,
$\mathbb{P}(A \cup B) \leq \mathbb{P}(A) + \mathbb{P}(B)$

In general, $\mathbb{P}(A_1 \cup A_2 \cup \cdots \cup A_n) \leq \sum_{k=1}^{n} \mathbb{P}(A_k)$.

P1 Countable subadditivity

For any sequence of events $(A_n)_{n=1}^{\infty}$,

$$\mathbb{P}\left(\bigcup_{n=1}^{\infty} A_n \right) \leq \sum_{n=1}^{\infty} \mathbb{P}(A_n)$$

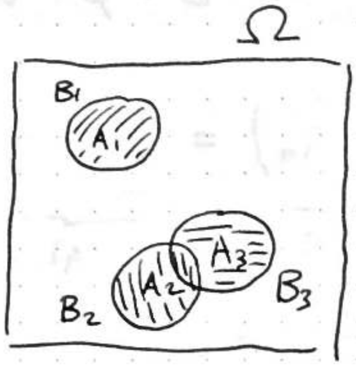
Indeed, let $B_1 = A_1$,

$B_2 = A_2 \setminus A_1$

$B_3 = A_3 \setminus (A_1 \cup A_2)$

$\vdots$

$B_n = A_n \setminus \left(\bigcup_{k=1}^{n-1} A_k \right)$.



Then the sets $(B_n)_{n \geq 1}$ are disjoint, and

$$\bigcup_{n=1}^{\infty} B_n = \bigcup_{n=1}^{\infty} A_n .$$

$B_n \subseteq A_n$

Then $\mathbb{P}\left(\bigcup_{n=1}^{\infty} A_n \right) = \mathbb{P}\left(\bigcup_{n=1}^{\infty} B_n \right) = \sum_{n=1}^{\infty} \mathbb{P}(B_n) \leq \sum_{n=1}^{\infty} \mathbb{P}(A_n)$.

P2 Continuity of probability measures

Let $(A_n)_{n \geq 1}$ be an increasing sequence of events

$A_1 \subseteq A_2 \subseteq \cdots$ Then:

$$\mathbb{P}\left(\bigcup_{n=1}^{\infty} A_n \right) = \lim_{n \to \infty} \mathbb{P}(A_n) = \lim_{n \to \infty} \mathbb{P}\left(\bigcup_{m=1}^{n} A_m \right)$$

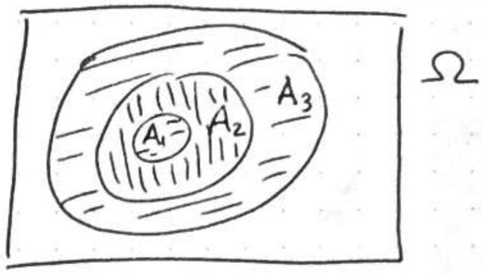
since $\bigcup_{m=1}^{n} A_m = A_n$.

Proof: Define $B_1 = A_1$

$B_2 = A_2 \setminus A_1$

$\vdots$

$B_n = A_n \setminus \left(\bigcup_{k=1}^{n-1} A_k \right)$



Then $(B_n)_{n \geq 1}$ is a sequence of disjoint events

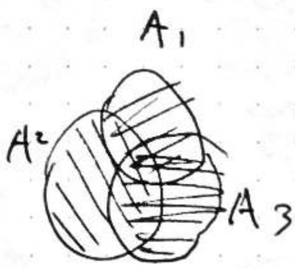
$$\begin{aligned}
\mathbb{P}\left(\bigcup_{n=1}^{\infty} A_n \right) &= \mathbb{P}\left(\bigcup_{n=1}^{\infty} B_n \right) \\
&= \sum_{n=1}^{\infty} \mathbb{P}(B_n) \\
&= \lim_{N \to \infty} \sum_{n=1}^{N} \mathbb{P}(B_n) \\
&= \lim_{N \to \infty} \mathbb{P}\left(\bigcup_{n=1}^{N} B_n \right) \\
&= \lim_{N \to \infty} \mathbb{P}(A_N).
\end{aligned}$$

$\square$

Similarly, if $(A_n)_{n \geq 1}$ is a sequence of decreasing events $A_1 \supseteq A_2 \supseteq \cdots$, then

$$\mathbb{P}\left(\bigcap_{n=1}^{\infty} A_n \right) = \lim_{n \to \infty} \mathbb{P}(A_n).$$

(P3) Inclusion-exclusion formula



We have seen that

$$\mathbb{P}(A_1 \cup \cdots \cup A_n) \leq \sum_{k=1}^{n} \mathbb{P}(A_k)$$

want equality.

$n=2$:

$$\mathbb{P}(A_1 \cup A_2) = \mathbb{P}(A_1) + \mathbb{P}(A_2) - \mathbb{P}(A_1 \cap A_2)$$

$n=3$:

$$\begin{aligned}
\mathbb{P}(A_1 \cup A_2 \cup A_3) = {} & \mathbb{P}(A_1) + \mathbb{P}(A_2) + \mathbb{P}(A_3) - \mathbb{P}(A_1 \cap A_2) \\
& - \mathbb{P}(A_2 \cap A_3) - \mathbb{P}(A_3 \cap A_1) + \mathbb{P}(A_1 \cap A_2 \cap A_3)
\end{aligned}$$

L4.4

Inclusion-exclusion formula:

$$\mathbb{P}(A_1 \cup \cdots \cup A_n) = \sum_{k=1}^{n} (-1)^{k+1} \sum_{1 \leq i_1 < \cdots < i_k \leq n} \mathbb{P}(A_{i_1} \cap \cdots \cap A_{i_k})$$

$$\begin{aligned} &= \mathbb{P}(A_1) + \cdots + \mathbb{P}(A_n) \\ &\quad - \mathbb{P}(A_1 \cap A_2) - \cdots - \mathbb{P}(A_{n-1} \cap A_n) \\ &\quad + \mathbb{P}(A_1 \cap A_2 \cap A_3) + \cdots + \mathbb{P}(A_{n-2} \cap A_{n-1} \cap A_n) \\ &\quad - \cdots \\ &\quad + (-1)^{n+1} \mathbb{P}(A_1 \cap \cdots \cap A_n) \end{aligned}$$

Note that by taking equally likely outcomes, this gives an identity for the cardinality of the union of n sets.

$$|A_1 \cup \cdots \cup A_n| = \sum_{k=1}^{n} (-1)^{k+1} \sum_{1 \leq i_1 < \cdots < i_k \leq n} |A_{i_1} \cap \cdots \cap A_{i_k}|$$

<u>Proof</u>: True for $n=2$. For general n, assume it's true for $n-1$, write

$$\mathbb{P}(A_1 \cup \cdots \cup A_n) = \mathbb{P}(A_1 \cup \cdots \cup A_{n-1}) + \mathbb{P}(A_n) - \mathbb{P}(I_1 \cup I_2 \cup \cdots \cup I_{n-1})$$

where $I_k =: A_k \cap A_n$ for $1 \leq k \leq n-1$.

Expand the terms with inductive hypothesis. $\square$

<u>EX</u>: <u>Derangements</u>

A permutation of $\{1, \dots, n\}$ is called a derangements if it doesn't have any fixed points.

<u>Q</u>: $\mathbb{P}(\{\text{derangements}\}) = ?$

Let $\Omega = \{\text{permutations of } \{1, \dots, n\}\} = \{\text{bijections } \omega : \{1, \dots, n\} \to \{1, \dots, n\}\}$

$|\Omega| = n!$ Let $A_i = \{\omega \in \Omega : \omega(i) = i\}$. Then $\{\text{derangements}\} = (A_1 \cup \cdots \cup A_n)^c$

$$\mathbb{P}(\text{der.}) = 1 - \sum_{k=1}^{n} (-1)^{k+1} \sum_{1 \leq i_1 < \cdots < i_k \leq n} \underbrace{\mathbb{P}(A_{i_1} \cap \cdots \cap A_{i_k})}_{\text{perms. fix } i_1, \dots, i_k,\ \left(\frac{(n-k)!}{n!}\right)} = 1 - \sum_{k=1}^{n} (-1)^{k+1} \left(\frac{(n-k)!}{n!}\right) \binom{n}{k}$$

$$= \sum_{k=0}^{n} (-1)^k \frac{1}{k!} \xrightarrow[n \to \infty]{} e^{-1}$$

L5.1

Properties of prob. measures

$$\mathbb{P}(\Omega)=1, \quad \mathbb{P}(A)+\mathbb{P}(A^c)=1, \quad \mathbb{P}(\emptyset)=0$$

1) Countable subadditivity $\mathbb{P}(\bigcup_{n\geq 1} A_n) \leq \sum_{n\geq 1} \mathbb{P}(A_n)$

2) Continuity $\mathbb{P}(A_n) \to \mathbb{P}(\bigcup_n A_n)$ if $A_1 \subseteq \cdots \subseteq A_k \subseteq \cdots$

$\mathbb{P}(A_n) \to \mathbb{P}(\bigcap_n A_n)$ if $A_1 \supseteq \cdots \supseteq A_k \supseteq \cdots$

3) Inclusion - exclusion

$$\mathbb{P}(A_1 \cup \cdots \cup A_n) = \sum_{k=1}^{n} (-1)^{k+1} \sum_{1\leq i_1 < \cdots < i_k \leq n} \mathbb{P}(A_{i_1} \cap \cdots \cap A_{i_k})$$

Applying IE

I) Derangements

Choose a random permutation ω of $\{1,\dots,n\}$

$\mathbb{P}(\omega$ is a derangement$)$

$A_i = \{\omega \in \Omega : \omega(i) = i\}$

$A = (\bigcup_{i=1}^{n} A_i)^c$

$$\mathbb{P}(A) = 1 - \sum_{k=1}^{n} (-1)^{k+1} \sum_{1\leq i_1<\cdots<i_k\leq n} \mathbb{P}(A_{i_1} \cap \cdots \cap A_{i_k})$$

$|A_{i_1} \cap \cdots \cap A_{i_k}| = (n-k)!$

So $\mathbb{P}(A) = 1 - \sum_{k=1}^{n} (-1)^{k+1} \binom{n}{k} \frac{(n-k)!}{n!} = \sum_{k=0}^{n} (-1)^k \frac{n!}{k!} / n! \to e^{-1}$

II) Surjective functions

$\Omega = \{f : \{1,\dots,n\} \to \{1,\dots,m\}\}$

$|\Omega| = m^n$

$\mathbb{P}(f\in\Omega : f$ is surjective$)$

Let $A_i = \{f \in \Omega : i \notin f(\{1,\dots,n\})\}$.

$A = (\bigcup_{i=1}^{m} A_i)^c$

$$\mathbb{P}(A) = 1 - \mathbb{P}(\bigcup_{i=1}^{m} A_i) = 1 - \sum_{k=1}^{m} (-1)^{k+1} \sum_{1\leq i_1<\cdots<i_k\leq m} \mathbb{P}(A_{i_1} \cap \cdots \cap A_{i_k})$$

$\frac{(m-k)^n}{m^n}$

$$= 1 - \sum_{k=1}^{m} (-1)^{k+1} \binom{m}{k} \left(1-\frac{k}{m}\right)^n$$

$$= \sum_{k=0}^{m} (-1)^k \binom{m}{k} \left(1-\frac{k}{m}\right)^n$$

L5.2 **Bonferroni inequalities**

$n=2$: $\mathbb{P}(A\cup B) = \mathbb{P}(A)+\mathbb{P}(B) - \mathbb{P}(A\cup B) \leq \mathbb{P}(A)+\mathbb{P}(B)$

$n=3$: $\mathbb{P}(A\cup B\cup C) = \mathbb{P}(A)+\mathbb{P}(B)+\mathbb{P}(C)$

$$- \mathbb{P}(A\cap B) - \mathbb{P}(A\cap C) - \mathbb{P}(B\cap C)$$

$$+ \mathbb{P}(A\cap B\cap C)$$

In general, truncating the IE formula at step k gives an upper bound if k is odd, and a lower bound if k is even

Proof: by induction on n

$n=2$: done

Assume for $n-1$, then

$$\mathbb{P}(A_1\cup\cdots\cup A_n) = \mathbb{P}(A_1\cup\cdots\cup A_{n-1}) + \mathbb{P}(A_n)$$

$$- \mathbb{P}(I_1\cup\cdots\cup I_{n-1}) \quad \text{where } I_k = A_k\cap A_n$$

truncate at step k

truncate at step $k-1$

k odd: upper bound − lower bound => upper bound

k even: lower bound − upper bound => lower bound □

Independence

Def.^{n} Two events $A, B\in\mathcal{F}$ are independent if

$$\mathbb{P}(A\cap B) = \mathbb{P}(A)\cdot\mathbb{P}(B)$$

Ex: Roll two dice. $\Omega = \{1,\dots,6\}\times\{1,\dots,6\}$

$$|\Omega| = 36$$

$A = \{\text{the sum is } 6\}$ $\quad B = \{\text{first die gives } 4\}$

$\mathbb{P}(A) = \frac{5}{36}$ $\quad \mathbb{P}(B) = \frac{6}{36} = \frac{1}{6}$ $\quad \mathbb{P}(A\cap B) = \frac{1}{36}$

$$\mathbb{P}(A\cap B) \neq \mathbb{P}(A)\cdot\mathbb{P}(B)$$

so A, B not independent.

Let $A' = \{\text{the sum is } 7\}$. Then $\mathbb{P}(A') = \frac{6}{36} = \frac{1}{36}$,

$$\mathbb{P}(A'\cap B) = \frac{1}{36} = \mathbb{P}(A')\cdot\mathbb{P}(B).$$

So A', B independent.

Note that if A is indep. of B then A is indep. of B^c.

Know $\mathbb{P}(A \cap B) = \mathbb{P}(A) \cdot \mathbb{P}(B)$.

Start from $\mathbb{P}(A) = \mathbb{P}(A \cap B) + \mathbb{P}(A \cap B^c)$ union of disjoint events

$$\therefore \quad \mathbb{P}(A \cap B^c) = \mathbb{P}(A) - \mathbb{P}(A)\mathbb{P}(B)$$
$$= \mathbb{P}(A)(1 - \mathbb{P}(B))$$
$$= \mathbb{P}(A)\,\mathbb{P}(B^c) \quad \text{as needed.}$$

<u>Def:</u> Events $A_1, \dots, A_n$ are independent if for any indices $1 \leq i_1 < i_2 < \dots < i_k \leq n$ it holds that

$$\mathbb{P}(A_{i_1} \cap A_{i_2} \cap \dots \cap A_{i_k}) = \mathbb{P}(A_{i_1}) \cdot \mathbb{P}(A_{i_2}) \cdots \mathbb{P}(A_{i_k})$$

<u>Remark:</u> Note that if $A_1, \dots, A_n$ are independent, then they are pairwise independent, i.e. $\forall\ 1 \leq i < j \leq n$, $\mathbb{P}(A_i \cap A_j) = \mathbb{P}(A_i)\mathbb{P}(A_j)$

We now show that independence is strictly stronger than pairwise independence: roll two coins $\Omega = \{H, T\} \times \{H, T\}$

$A = \{1^{st} \text{ coin } T\} = \{(T,H), (T,T)\}$

$B = \{2^{nd} \text{ coin } T\} = \{(H,T), (T,T)\}$

$C = \{\text{exactly one } T\} = \{(H,T), (T,H)\}$

$\mathbb{P}(A) = \mathbb{P}(B) = \mathbb{P}(C) = \frac{1}{2}$

$\mathbb{P}(A \cap B) = \mathbb{P}(B \cap C) = \mathbb{P}(C \cap A) = \frac{1}{4}$ ✓ so pairwise independent

But $\mathbb{P}(A \cap B \cap C) = 0 \neq \frac{1}{8}$ ✗ so not independent

<u>Product spaces with equally likely outcomes</u>

Let $\Omega = \Omega_1 \times \dots \times \Omega_n$, $|\Omega| = |\Omega_1| \times \dots \times |\Omega_n|$, with $\mathbb{P}(A) = |A|/|\Omega|$ for all $A \subseteq \Omega$.

Suppose $A_1, \dots, A_n$ are of the form

$$A_i = \{\omega = (\omega_1, \dots, \omega_n) : \omega_i \in B_i\} \text{ for some } B_i \subseteq \Omega_i$$
$$= \Omega_1 \times \Omega_2 \times \dots \times B_i \times \dots \times \Omega_n$$

Note that $\mathbb{P}(A_i) = |A_i|/|\Omega| = \dfrac{|\Omega_1| \cdots |\Omega_{i-1}||B_i| \cdots |\Omega_n|}{|\Omega_1| \cdots |\Omega_{i-1}||\Omega_i| \cdots |\Omega_n|} = \dfrac{|B_i|}{|\Omega_i|}$

L5.4

$$\mathbb{P}(A_{i_1} \cap \cdots \cap A_{i_k}) = \frac{|A_{i_1} \cap \cdots \cap A_{i_k}|}{|\Omega|} = \frac{|B_{i_1}|}{|\Omega_{i_1}|} \cdot \frac{|B_{i_2}|}{|\Omega_{i_2}|} \cdots \frac{|B_{i_k}|}{|\Omega_{i_k}|}$$

$$= \mathbb{P}(A_{i_1}) \cdots \mathbb{P}(A_{i_k})$$

Therefore $A_1, \dots, A_n$ are independent.

L6.1 <u>Properties of prob. measures</u>

1) Countable subadditivity
2) Continuity of probability
3) Inclusion - Exclusion formula
4) Bonferroni inequalities

Independence

A indep. of B if

$$\mathbb{P}(A \cap B) = \mathbb{P}(A) \cdot \mathbb{P}(B)$$

written $A \perp\!\!\!\perp B$

$A \perp\!\!\!\perp B \Rightarrow A \perp\!\!\!\perp B^c \Rightarrow A^c \perp\!\!\!\perp B^c \Rightarrow A^c \perp\!\!\!\perp B$

Also look at independence of more and independence $\not\Leftrightarrow$ pairwise indep.

Conditional probability

For any two events A, B with $\mathbb{P}(B) > 0$, we define the conditional probability of A given B by

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} \leftarrow \text{crucial } \mathbb{P}(B) \neq 0$$

Let $\tilde{\mathbb{P}}(A) := \mathbb{P}(A \mid B)$. Then $\tilde{\mathbb{P}} : \mathcal{F} \to [0,1]$ has the properties of a probability measure. Indeed

- $\tilde{\mathbb{P}}(\Omega) = \mathbb{P}(\Omega \mid B) = \dfrac{\mathbb{P}(\Omega \cap B)}{\mathbb{P}(B)} = \dfrac{\mathbb{P}(B)}{\mathbb{P}(B)} = 1$

- for any family of disjoint events $(A_n)_{n \geq 1}$, we have

$$\tilde{\mathbb{P}}\Big(\bigcup_{n\geq 1} A_n\Big) = \mathbb{P}\Big(\bigcup_{n \geq 1} A_n \,\Big|\, B\Big) = \frac{\mathbb{P}\big(\big(\bigcup_{n\geq 1} A_n\big) \cap B\big)}{\mathbb{P}(B)}$$

$$= \frac{\mathbb{P}\big(\bigcup_{n\geq 1}(A_n \cap B)\big)}{\mathbb{P}(B)} = \frac{\sum_{n \geq 1} \mathbb{P}(A_n \cap B)}{\mathbb{P}(B)} \quad \text{as } \mathbb{P} \text{ is a probability measure}$$

$$= \sum_{n\geq 1} \mathbb{P}(A_n \mid B) = \sum_{n \geq 1} \tilde{\mathbb{P}}(A_n).$$

Ex: Equally likely outcomes

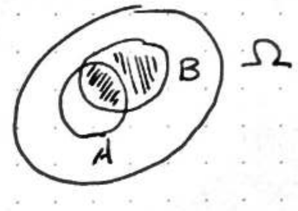
Take $|\Omega| < \infty$ with $\mathbb{P}(A) = \frac{|A|}{|\Omega|}$ for all $A \in \mathcal{F}$.

Pick B non-empty subset of Ω, then $\mathbb{P}(B) > 0$.

Then for all $A \in \mathcal{F}$

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(A\cap B)}{\mathbb{P}(B)} = \frac{|A\cap B| / \cancel{|\Omega|}}{|B| / \cancel{|\Omega|}} = \frac{|A\cap B|}{|B|}.$$

In this setting (equally likely outcomes) $\mathbb{P}(A \mid B)$ is the proportion of outcomes which are in A as well as B.



Remarks:

1) If $A \cap B = \emptyset$ (A, B) disjoint and $\mathbb{P}(B) > 0$, then $\mathbb{P}(A\cap B) = 0$

2) Let B, C be events with $\mathbb{P}(B\cap C) > 0$. Then for all $A \in \mathcal{F}$,

$$\mathbb{P}(A \mid B\cap C) = \frac{\mathbb{P}(A\cap B\cap C)}{\mathbb{P}(B\cap C)} = \frac{\mathbb{P}(A\cap B \mid C)\,\mathbb{P}(C)}{\mathbb{P}(B\mid C)\,\mathbb{P}(C)}$$

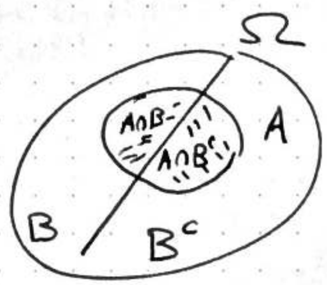
Here we have used that $\mathbb{P}(A\cap B) = \mathbb{P}(A\mid B)\mathbb{P}(B)$.

Hence
$$\mathbb{P}(A \mid B\cap C) = \frac{\mathbb{P}(A\cap B\mid C)}{\mathbb{P}(B\mid C)}.$$

Note that if $A \perp\!\!\!\perp B$ and $\mathbb{P}(B) > 0$, then
$$\mathbb{P}(A\mid B) = \mathbb{P}(A).$$

Indeed,
$$\mathbb{P}(A\mid B) = \frac{\mathbb{P}(A\cap B)}{\mathbb{P}(B)} = \frac{\mathbb{P}(A)\,\cancel{\mathbb{P}(B)}}{\cancel{\mathbb{P}(B)}} = \mathbb{P}(A).$$

Note that if $\mathbb{P}(B) > 0$ then
$$\mathbb{P}(A\cap B) = \mathbb{P}(A\mid B)\,\mathbb{P}(B).$$



If, moreover, $\mathbb{P}(B^c) > 0$, then
$$\mathbb{P}(A\cap B^c) = \mathbb{P}(A\mid B^c)\,\mathbb{P}(B^c).$$

But note that $B \cup B^c = \Omega$

so $A = (A\cap B) \cup (A\cap B^c)$.

It follows that

$$\mathbb{P}(A) = \mathbb{P}(A \cap B) + \mathbb{P}(A \cap B^c)$$
$$= \mathbb{P}(A \mid B)\,\mathbb{P}(B) + \mathbb{P}(A \mid B^c)\,\mathbb{P}(B^c).$$

Law of total probability

Let $(B_n)_{n \geq 1}$ be a sequence of disjoint events such that

$\bigcup_{n \geq 1} B_n = \Omega$, and $\mathbb{P}(B_n) > 0 \ \forall n$.

Then for all $A \in \mathcal{F}$, it holds that

« sequence can be finite »

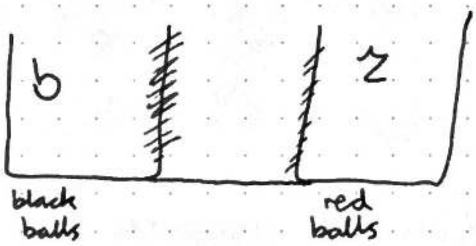
$$\mathbb{P}(A) = \sum_{n \geq 1} \mathbb{P}(A \mid B_n) \cdot \mathbb{P}(B_n).$$

Remark: can drop assumption $\mathbb{P}(B_n) > 0$ by defining $\mathbb{P}(A \mid B_n)\,\mathbb{P}(B_n) = 0$ if $\mathbb{P}(B_n) = 0$.

Proof: $\mathbb{P}(A) = \mathbb{P}(A \cap \Omega) = \mathbb{P}(A \cap (\bigcup_{n \geq 1} B_n))$

$$= \mathbb{P}\Big(\bigcup_{n \geq 1} A \cap B_n\Big)$$
$$= \sum_{n \geq 1} \mathbb{P}(A \cap B_n)$$
$$= \sum_{n \geq 1} \mathbb{P}(A \mid B_n)\,\mathbb{P}(B_n). \qquad \square$$

EX:



Pick 2 balls without replacement.

Q: What is the probability that the second ball is black?

Let A be "2nd ball black".

Let B be "1st ball black".

Then
$$\mathbb{P}(A) = \underbrace{\mathbb{P}(A \mid B)}\,\underbrace{\mathbb{P}(B)} + \underbrace{\mathbb{P}(A \mid B^c)}\,\underbrace{\mathbb{P}(B^c)}$$
$$= \frac{b-1}{b+r-1} \cdot \frac{b}{b+r} + \frac{b}{b+r-1} \cdot \frac{r}{b+r}$$

$$= \frac{b}{b+r} \quad \text{oh}$$

L6.4 Q: Can we switch the conditioning?

Let A, B be two events of positive probability.
Then $$\mathbb{P}(A|B) = \frac{\mathbb{P}(A\cap B)}{\mathbb{P}(B)} = \frac{\mathbb{P}(B|A)\,\mathbb{P}(A)}{\mathbb{P}(B)}.$$

Bayes's theorem generalises the above

Theorem (Bayes):
Let $(B_n)_{n\geq 1}$ be a sequence of disjoint events such that $\bigcup_{n\geq 1} B_n = \Omega$. Assume that $\mathbb{P}(B_n)>0$ $\forall n$.
Then, for all events A with $\mathbb{P}(A)>0$,
$$\mathbb{P}(B_n|A) = \frac{\mathbb{P}(A|B_n)\,\mathbb{P}(B_n)}{\sum_{k\geq 1}\mathbb{P}(A|B_k)\,\mathbb{P}(B_k)}.$$

Proof: Trivial. $\square$

Recap: Independence, Law of Total Probability, Bayes

Proof of Bayes: $$\mathbb{P}(B_n | A) = \frac{\mathbb{P}(B_n \cap A)}{\mathbb{P}(A)} = \frac{\mathbb{P}(A|B_n)\,\mathbb{P}(B_n)}{\underbrace{\sum_k \mathbb{P}(A|B_k)\,\mathbb{P}(B_k)}_{\text{TOTAL PROBABILITY}}} \qquad \square$$

EX:

$A = \{2^{nd} \text{ ball black}\}$

$B = \{1^{st} \text{ ball black}\}$

$$\mathbb{P}(B|A) = \frac{\mathbb{P}(A|B)\,\mathbb{P}(B)}{\mathbb{P}(A|B)\,\mathbb{P}(B) + \mathbb{P}(A|B^c)\,\mathbb{P}(B^c)} = \frac{\cancel{b}(b-1)}{\cancel{b}(b-1) + r\cancel{b}}.$$

Applications

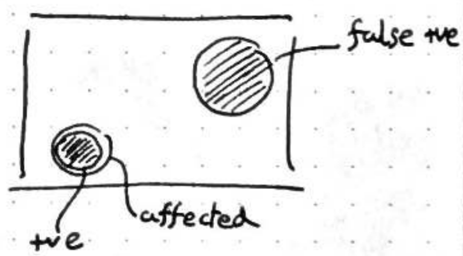
1) False positives for rare conditions

Spose a rare condition affects 0.001 of the people.
A randomly chosen individual is tested for the condition, and the outcome is positive (i.e. affected)

Q: How much can you trust the test?

Empirically, test known to give 98% positive response if the individual is affected, and 1% positive response if they're not.
Let $A = \{\text{individual is affected}\}$, $B = \{\text{outcome is positive}\}$.

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(B|A)\,\mathbb{P}(A)}{\mathbb{P}(B|A)\,\mathbb{P}(A) + \mathbb{P}(B|A^c)\,\mathbb{P}(A^c)}$$



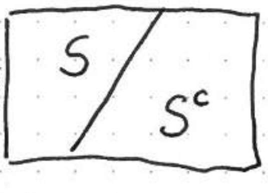
$$= \frac{(0.98)(0.001)}{(0.98)(0.001) + (0.01)(0.999)}$$

$$= 0.089 \quad \frown$$

2) Simpson's paradox

- For A, B events, we say that "B attracts A" if
  $\mathbb{P}(A|B) > \mathbb{P}(A)$.
- Given an event S, we say that "B attracts A on S" if
  $\mathbb{P}(A|B \cap S) > \mathbb{P}(A|S)$.

B attracts A on S



B attracts A on S^c

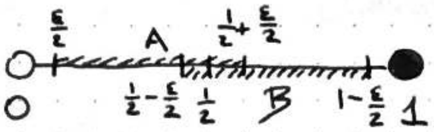
One might think that if B attracts A on both S and S^c, then B attracts A.

L7.2 This is in general false.

Let $\Omega = (0,1]$.

We can define a probability measure on Ω such that

$$\mathbb{P}\big((a,b]\big) = b-a, \quad \text{for all } 0 \leq a < b \leq 1. \quad \text{NOT TRIVIAL}$$



Fix $\varepsilon \in (0, \frac{1}{4})$, let $A = (\frac{\varepsilon}{2}, \frac{1}{2}+\frac{\varepsilon}{2}]$,
$B = (\frac{1}{2}-\frac{\varepsilon}{2}, 1-\frac{\varepsilon}{2}]$,

$S = (0, \frac{1}{2}]$, $\quad S^c = (\frac{1}{2}, 1]$.

I will show that
- B attracts A on S
- B attracts A on S^c
- B does not attract A.

$$\mathbb{P}(A \mid B\cap S) = \frac{\mathbb{P}(A\cap B\cap S)}{\mathbb{P}(B\cap S)} = \frac{\varepsilon/2}{\varepsilon/2} = 1$$

so B attracts A on S

$$\mathbb{P}(A \mid S) = \frac{\mathbb{P}(A\cap S)}{\mathbb{P}(S)} = \frac{\frac{1}{2}-\frac{\varepsilon}{2}}{\frac{1}{2}} = 1-\varepsilon$$

$$\mathbb{P}(A \mid B\cap S^c) = \frac{\mathbb{P}(A\cap B\cap S^c)}{\mathbb{P}(B\cap S^c)} = \frac{\varepsilon/2}{\frac{1}{2}-\frac{\varepsilon}{2}} = \frac{\varepsilon}{1-\varepsilon}$$

so B attracts A on S^c

$$\mathbb{P}(A \mid S^c) = \frac{\mathbb{P}(A\cap S^c)}{\mathbb{P}(S^c)} = \frac{\varepsilon/2}{\frac{1}{2}} = \varepsilon$$

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(A\cap B)}{\mathbb{P}(B)} = \frac{\varepsilon}{1/2} = 2\varepsilon$$

$$\mathbb{P}(A) = \frac{1}{2}$$

But $\varepsilon < \frac{1}{4}$ so $\mathbb{P}(A\mid B) < \mathbb{P}(A)$.
So B does not attract A.

- -

$\mathbb{P}(A) = \mathbb{P}(A\mid S)\mathbb{P}(S) + \mathbb{P}(A\mid S^c)\mathbb{P}(S^c) = (1-\varepsilon)\times\frac{1}{2} + \varepsilon\times\frac{1}{2} = \frac{1}{2}$

$\mathbb{P}(A\mid B) = \mathbb{P}(A\mid B\cap S)\mathbb{P}(S\mid B) + \mathbb{P}(A\mid B\cap S^c)\mathbb{P}(S^c\mid B) = 1\times\varepsilon + \frac{\varepsilon}{1-\varepsilon}\times(1-\varepsilon) = 2\varepsilon$

Important: $1-\varepsilon \gg \frac{\varepsilon}{1-\varepsilon}$ so even if $1 > 1-\varepsilon$ cannot make up.

"Whites in Wisconsin >> Blacks in Texas" so $\therefore$ even if "Whites in Texas > in Wisconsin"

3) Paradox of the two children

(1) I have two children, one of which is a boy

(2) Same as above, but the boy was born on a Tuesday

(1): $\Omega = \{ \underset{\frac{1}{4}}{BB}, \underset{\frac{1}{2}}{BG}, \underset{\frac{1}{4}}{GG} \}$ $\qquad \mathbb{P}(BB \mid BB \cup BG) = \dfrac{1/4}{1/4 + 1/2} = 1/3$

(2): $\Omega = \{ \underset{(\frac{1}{14})^2}{B^*B^*}, \underset{\frac{6}{14^2}\cdot 2}{B^*B}, \underset{\frac{7}{14^2}\cdot 2}{B^*G}, BG, GG, BB \}$

B^* = Tuesday boy

B = non-Tuesday boy $\qquad \mathbb{P}(\{B^*B^*, B^*B, \cancel{B^*}\ BB\}$ given $\{B^*B^*, B^*B, B^*G\})$

$$= \frac{\mathbb{P}(B^*B^*, B^*B)}{\mathbb{P}(B^*B^*, B^*B, B^*G)} = \frac{1 + 6\cdot 2}{1 + 6\cdot 2 + 7\cdot 2} = \frac{13}{27}$$

~ two boy family means more chance of Tuesday boy

or: get rid of BG so higher proportion of BB(*) and BB

or: B^*B^* pretty negligible

L8.1 <u>Equally likely outcomes</u>

$|\Omega| < \infty \qquad P(A) = \frac{|A|}{|\Omega|}$

<u>EX</u>: Toss a biased coin

$P(\{H\}) = p \qquad\qquad \Omega = \{H, T\}$

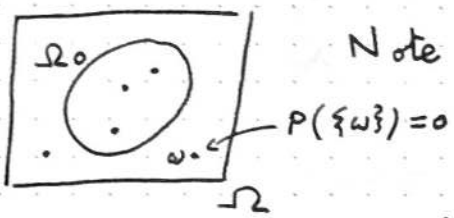
$P(\{T\}) = 1 - p$

Note that $P(\{H\}) + P(\{T\}) = P(\Omega) = 1$

Probability distributions

A probability distribution on $(\Omega, \mathcal{F})$ is a probability measure.

<u>Def</u>: A probability distribution is <u>discrete</u> if there exists a countable subset $\Omega_0 \subseteq \Omega$ such that $P(\Omega) = 1$.

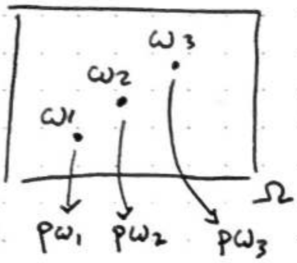


Note that in this case $\forall A \in \mathcal{F}$

$$P(A) = \sum_{\omega \in A \cap \Omega_0} P(\{\omega\})$$

assuming $\forall \omega \in \Omega_0$, $\{\omega\} \in \mathcal{F}$.

We will focus on countable sample spaces, i.e. $\Omega = \Omega_0$.



Note that a probability distribution is uniquely determined by $\{p_\omega := P(\{\omega\}),\ \omega \in \Omega = \Omega_0\}$, since then $P(A) = \sum_{\omega \in A} p_\omega$.

Refer to $\{p_\omega : \omega \in \Omega = \Omega_0\}$ as the <u>mass function</u> of the probability distribution P, and to the p_ω's as the <u>weights</u>.

Note that $p_\omega \in [0, 1]$ for all $\omega \in \Omega_0$ and that $\sum_{\omega \in \Omega_0} p_\omega = 1$ since $\bigcup_{\omega \in \Omega_0} \{\omega\}$.

> ALERT:
> $p_\omega : \Omega_0 \to [0, 1]$
> is a function

Some common probability distributions

1: Bernoulli distribution

$\Omega = \{0, 1\}$ with $p_1 = p$, $p_0 = 1 - p$

for some parameter $p \in [0, 1]$.

L8.2 Note that $p_1 + p_0 = 1$. This is referred to as the Bernoulli distribution of parameter p, denoted <u>Bernoulli(p)</u>.

2) Binomial distribution

(Number of heads in N tosses of a biased coin)

$\Omega = \{0, 1, 2, \dots, N\}$ where $N \geq 1$.

Consider the probability distribution given by

$$p_k = \binom{N}{k} p^k (1-p)^{N-k} \qquad 0 \leq k \leq N$$

And check that

$$\sum_{k=0}^{N} \binom{N}{k} p^k (1-p)^{N-k} = [p + (1-p)]^N = 1.$$

This is referred to as the Binomial distribution of parameters N, p denoted Binomial (N, p). Note that Binomial $(1, p)$ coincides with Bernoulli(p).

3) Geometric distribution

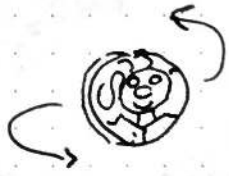
(Toss a coin until 1^{st} head)

Take $\Omega = \mathbb{N}$.

Consider the probability distribution on Ω given by the weights

$$p_k = p(1-p)^{k-1} \qquad k \geq 1$$

Note that $p_k \in [0, 1]$ and that



$$\sum_{k=1}^{\infty} p(1-p)^{k-1} = p \cdot \frac{1}{1-(1-p)} = 1$$

Note that if $p = 0$ computation invalid but e.z.

This is referred to as the Geometric distribution of parameter p, and it is denoted by Geometric(p).

<u>Warning</u>: the term geometric distribution also used with $\Omega = \mathbb{N} \cup \{0\}$ given by $p_k = p(1-p)^k$.

4) Poisson distribution

Take $\Omega = \mathbb{N} \cup \{0\}$ and consider the probability distribution determined by the weights

$$p_k = e^{-\lambda} \frac{\lambda^k}{k!} \quad \text{for } k \geq 0$$

where $\lambda \in (0, \infty)$ is a parameter.

Note that $p_k \in [0,1]$ for all $k \geq 0$, and that

$$\sum_{k=0}^{\infty} e^{-\lambda} \frac{\lambda^k}{k!} = e^{-\lambda} \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} = e^{-\lambda} e^{\lambda} = 1$$

This is referred to as the Poisson distribution of parameter λ, denoted by Poisson(λ).

We now show that the Poisson distribution arises as the limit of

$$\mathrm{Bin}\left(N, \frac{\lambda}{N}\right)$$

as $N \to \infty$. We will show $p_k^{\mathrm{Bin}} \xrightarrow[N \to \infty]{} p_k^{\mathrm{Poisson}}$ for $k \geq 0$.

$$p_k = \binom{N}{k} \left(\frac{\lambda}{N}\right)^k \left(1 - \frac{\lambda}{N}\right)^{N-k}$$

$$= \underbrace{\frac{N(N-1)\cdots(N-k+1)}{N^k}}_{\downarrow \atop 1} \cdot \frac{\lambda^k}{k!} \cdot \underbrace{\left(1 - \lambda/N\right)^N}_{\downarrow \atop e^{-\lambda}} \cdot \underbrace{\left(1 - \lambda/N\right)^{-k}}_{\downarrow \atop 1}$$

$$\xrightarrow[N \to \infty]{} e^{-\lambda} \frac{\lambda^k}{k!}, \text{ as needed.}$$

Probability distributions

Let Ω be countable. Specify distribution on Ω by specifying the weights $(p_\omega : \omega \in \Omega)$.

1) <u>Bernoulli distribution</u> (p) $p \in [0,1]$

$\Omega = \{0, 1\}$ $\quad p_1 = p \quad p_0 = 1-p \quad p_k = p^k(1-p)^{1-k}$

2) <u>Binomial distribution</u> (N, p)

$\Omega = \{0, 1, \dots, N\}$ $\quad p_k = \binom{N}{k} p^k (1-p)^{N-k}$

3) <u>Geometric</u> (p)

$\Omega = \mathbb{N}$ $\quad p_k = (1-p)^{k-1} p$

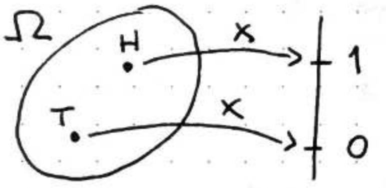
4) <u>Poisson</u> (λ)

$\Omega = \mathbb{N} \cup \{0\}$ $\quad p_k = \dfrac{e^{-\lambda}\lambda^k}{k!}$

Weights of binomial $(N, \frac{\lambda}{N}) \xrightarrow[N \to \infty]{}$ Weights of Poisson (λ)

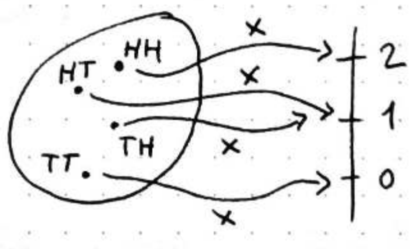
Random Variables

<u>Def.^n</u> let Ω be a countable set in a probability space $(\Omega, \mathcal{F}, P)$. A random variable on such a space taking values in S is a function $X : \Omega \to S$.

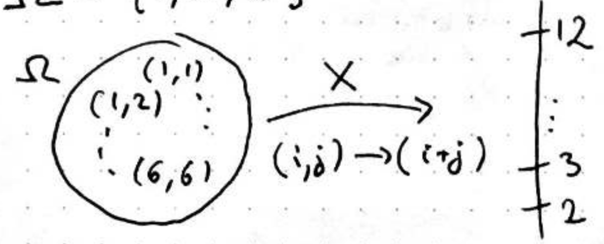


$X(H) = 1$
$X(T) = 0$

$\Omega = \{H, T\} \times \{H, T\}$



$\Omega = \{1, \dots, 6\}^2$



$Y(i,j) = \max\{i, j\}$
$Z(i,j) = i$

L9.2

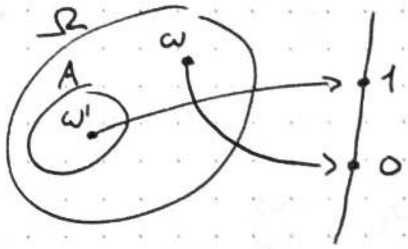
Typically take $S \subseteq \mathbb{R}^k$ for some $k \in \mathbb{N}$ in which case we say X is a <u>real-valued</u> random variable.

<u>EX</u>: (Indicator function)

Fix a probability space $(\Omega, \mathcal{F}, P)$. For some $A \in \mathcal{F}$ define

$$X = \mathbb{1}_A$$

by setting $X(\omega) = \begin{cases} 1 \text{ if } \omega \in A, \\ 0 \text{ if } \omega \notin A. \end{cases}$



Note that

(i) $\mathbb{1}_{A^c} = 1 - \mathbb{1}_A$

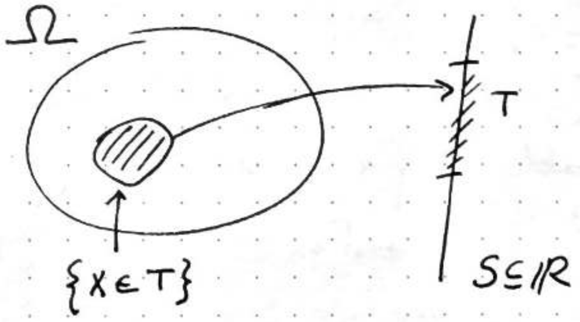
indeed if $\omega \notin A$, then $\omega \in A^c$ so we have $1 = 1 - 0$ ✓

and if $\omega \in A$, then $\omega \notin A^c$ and $0 = 1 - 1$ ✓

(ii) $\mathbb{1}_{A \cap B} = \mathbb{1}_A \mathbb{1}_B$

(iii) $\mathbb{1}_{A \cup B} = \mathbb{1}_A + \mathbb{1}_B - \mathbb{1}_{A \cap B} = 1 - (1 - \mathbb{1}_A)(1 - \mathbb{1}_B)$

For a subset $T \subseteq S$ we denote the set of all outcomes such that $X(\omega) \in T$ by $\{\omega \in \Omega : X(\omega) \in T\} = \{X \in T\}$



For all $x \in S$, define

$$p_x = P(X = x) = P(\{\omega \in \Omega : X(\omega) = x\}).$$

Then we call p_x the probability distribution of the random variable X. This defines a probability distribution on S.

restrict S to $X(\Omega)$ so <u>countable</u>

If the distribution of X is *Bernoulli*(p) say that X is a *Bernoulli random variable of parameter*, written

$$X \sim \text{Bernoulli}(p).$$

Similarly for *Binomial*, *Geometric*, *Poisson*.

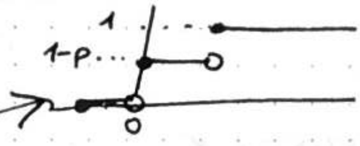
L9.3

Let X be a real valued random variable. Then define

$$F_X : \mathbb{R} \to [0,1]$$

$$x \to P(X \le x) \longrightarrow P(\{\omega \in \Omega : X(\omega) \le x\})$$

the distribution function of X.



Ex: Spose $X \sim$ Bernoulli(p). Then

$$F_X(x) = \begin{cases} 0 & \text{if } x < 0 \\ 1-p & \text{if } 0 \le x < 1 \\ 1 & \text{if } x \ge 1 \end{cases}$$

«Triggered»

not true always

Note that F_X is $\boxed{\text{piecewise constant,}}$ right continuous ($\lim_{y \to x^+} F_X(y) = F_X(x)$), $\lim_{x \to -\infty} F_X(x) = 0$, $\lim_{x \to \infty} F_X(x) = 1$, and non-decreasing.

true for all ~~discrete~~ random variables

Note that the distribution function uniquely determines the weights of the associated random variable. It follows that it uniquely determines the distribution itself.

Remark: For continuous random variables X, define them as functions $X : \Omega \to \mathbb{R}$ such that $\{X \le x\} \in \mathcal{F}$ $\forall x \in \mathbb{R}$. Will not discuss further (see Probability + Measure)

Def: Two random variables $X : \Omega \to S_X$, $Y : \Omega \to S_Y$ are said to be independent if $P(X = x, Y = y) = P(X = x)P(Y = y)$ $\forall x \in S_X, y \in S_Y$

In general, if $X_k : \Omega \to S_k$ for $1 \le k \le n$ for some $n \ge 2$, we say that $X_1, X_2, \ldots, X_n$ are independent if

$$P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) = P(X_1 = x_1) \cdots P(X_n = x_n)$$

$\forall (x_1, x_2, \ldots, x_n) \in S_1 \times S_2 \times \cdots \times S_n$.

Note that this implies that for all $j \in \{1,\dots,n\}$ and all distinct indices $1 \le i_1 < i_2 < \dots < i_j \le N$, it holds that

$$P(X_{i_1}=x_{i_1},\dots,X_{i_j}=x_{i_j}) = \prod_{k=1}^{j} P(X_{i_k}=x_{i_k})$$

for all $(x_{i_1},\dots,x_{i_j}) \in S_{i_1}\times\dots\times S_{i_k}$

For $n=3$, follows from the observation that if

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

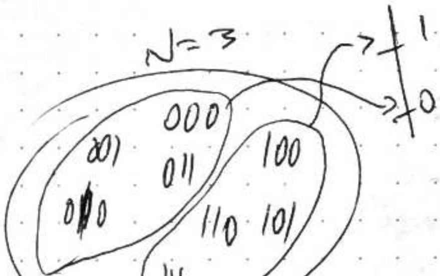
for all $(x,y,z) \in S_X \times S_Y \times S_Z$, then for any $x \in S_X$, $y \in S_Y$

we have $P(X=x, Y=y) = \sum_{z \in S_Z} P(X=x, Y=y, Z=z)$

$$= P(X=x)P(Y=y)\underbrace{\sum_{z\in S_Z} P(Z=z)}_{1}$$

<u>Example</u>: Fix $N \geq 1$, and let $\Omega = \{0,1\}^N$, so

Ω is the set of all sequences of 0s & 1s of length N.



$|\Omega| = 2^N$

For $1 \le k \le N$, let $X_k : \Omega \to \{0,1\}$ be defined by $X_k(\omega) = \omega_k$ i.e. k^{th} entry

On Ω, define the probability distribution $p_\omega = \prod_{k=1}^{n} p^{\omega_k}(1-p)^{1-\omega_k}$

$$= p^{\#1s}(1-p)^{\#0s}$$

Show that for all $1 \le k \le n$,

$$X_k \sim \text{Bernoulli}(p)$$

$$P(X_k = 1) = \sum_{\omega:\omega_1=1} P(\omega) = \sum_{\omega:\omega_1=1} \prod_{k=1}^{n} p^{\omega_k}(1-p)^{1-\omega_k} = p\underbrace{\sum_{\omega:\omega_1=1}\prod_{k=2}^{n} p^{\omega_k}(1-p)^{1-\omega_k}}_{1}$$

6.10.1

Ex: $\Omega = \{0,1\}^N$

For $\omega \in \Omega$ let $p_\omega := \prod_{k=1}^{N} p^{\omega_k}(1-p)^{1-\omega_k}$.

$\omega = (\omega_1, \omega_2, \dots, \omega_N)$

For $1 \leq k \leq N$ define $X_k(\omega) := \omega_k \in \{0,1\}$

(X_k reads the k^{th} digit of the sequence ω)

$$P(\{X_k = 1\}) = \sum_{\substack{\omega \in \Omega \\ \omega_k = 1}} p_\omega = \sum_{\substack{\omega \in \Omega \\ \omega_k = 1}} \prod_{j=1}^{N} p^{\omega_j}(1-p)^{1-\omega_j}$$

$$= p \underbrace{\sum_{\substack{\omega \in \Omega \\ \omega_k = 1}} \prod_{\substack{j=1 \\ j \neq k}}^{N} p^{\omega_j}(1-p)^{1-\omega_j}}_{1} = p$$

In fact, the X_k are independent, since

$$P(X_1 = \omega_1, \dots, X_n = \omega_n) = P(\{(\omega_1, \omega_2, \dots, \omega_n)\})$$

$$= \prod_{k=1}^{N} \underbrace{p^{\omega_k}(1-p)^{1-\omega_k}}_{P(X_k = \omega_k)} = \prod_{k=1}^{N} p^{\omega_k}(1-p)^{1-\omega_k}$$

$$= \prod_{k=1}^{N} P(X_k = \omega_k)$$

Let $S_n := \sum_{k=1}^{N} X_k$. <Number of 1s>

$S_n(\omega) \in \{0, \dots, N\}$

For $0 \leq k \leq N$,

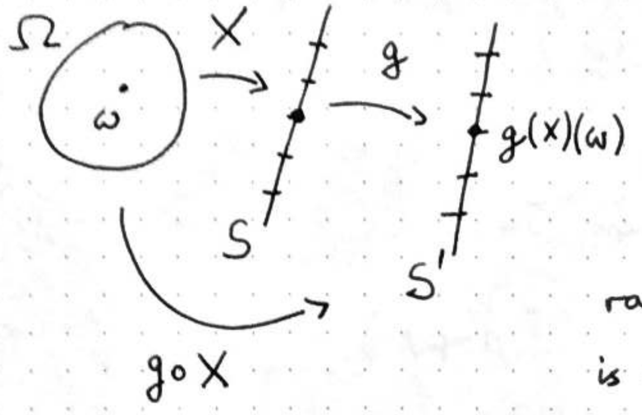
$$P(S_N = k) = \sum_{\omega : k\,1s} p_\omega$$

$$= \sum_{\omega : k1s} \underbrace{\prod_{j=1}^{N} p^{\omega_j}(1-p)^{1-\omega_j}}_{p^{\#1s}(1-p)^{\#0s}}$$

$$= \binom{n}{k} p^k (1-p)^{n-k}$$

Hence $S_N \sim B(N, p)$

<u>Def:</u> If X is a random variable taking values in S, and $g: S \to S'$, then $g(X)$ is itself a random variable taking values in S'.



Ω $\mathcal{F}$ Σpx $\int dx$

$\mathbb{E}$xpectation

From now on, only consider real-valued random variables. Say a random variable is non negative if $X(\omega) \geq 0 \;\forall\, \omega \in \Omega$.

<u>Def:</u> The expectation of a non-negative random variable X is defined as $E(X) = \sum_{x \in S_X} x \cdot P(X = x)$ where S_X is the image of Ω under X. Assume Ω countable.

Note sum may diverge. If it converges, it does absolutely. If it diverges, assign the value $+\infty$.

<u>EX:</u> $X \sim \text{Bernoulli}(p)$ then

$$E(X) = 1 \cdot \underbrace{p_1}_{p} + 0 \cdot \underbrace{p_0}_{1-p} = p$$

$X = \mathbb{1}_A$ gives $E(X) = 1 \cdot P(A) + 0 \cdot P(A^c) = P(A)$ ü

$X \sim \text{Binomial}(N, p)$

$$E(X) = \sum_{k=0}^{N} \binom{N}{k} p^k (1-p)^{N-k} \cdot k$$

or use linearity

$$= \sum_{k=0}^{N} \frac{N!}{(k-1)!(n-k)!} p^k (1-p)^{N-k}$$

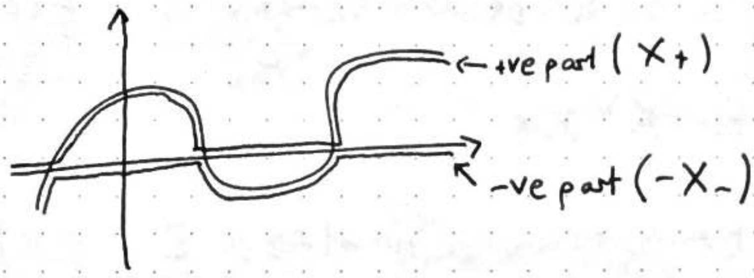
$$= Np \underbrace{\sum_{k=0}^{N} \frac{(N-1)!}{(k-1)!(N-k)!} p^{k-1} (1-p)^{N-k}}_{1} \quad ü$$

$X \sim \text{Poisson}(\lambda)$

$$E(X) = \sum_{k=0}^{\infty} k \cdot P(X=k) = \sum_{k=1}^{\infty} e^{-\lambda} \frac{k\lambda^k}{k!} = \lambda e^{-\lambda} \underbrace{\sum_{k=1}^{\infty} \frac{\lambda^{k-1}}{(k-1)!}}_{e^{\lambda}} = \lambda$$

For a random variable X, define

$$X_+ = \max\{X, 0\}$$
$$X_- = \max\{-X, 0\}$$



Note that X_+ and X_- are non-negative. At most one is non-zero.

Moreover, $X = X_+ - X_-$, $|X| = X_+ + X_-$

Def^n: If X is a random variable s.t. at least one of $E(X_+)$ and $E(X_-)$ is finite, define $E(X) = E(X_+) - E(X_-)$

$$= \sum_{x \in S_X} x \cdot P(X=x)$$

Ex: Take $X : \Omega \to \mathbb{Z} \setminus \{0\}$ with weights

$$p_k = \frac{3}{\pi^2 k^2} = p_{-k}$$

Then $E(X_+) = \sum_{k=1}^{\infty} k \cdot P(X=k) = \pi^2 \sum_{k=1}^{\infty} \frac{1}{k^2} \cdot k \longrightarrow \infty$

Both $E(X_+)$, $E(X_-)$ blow up.

Properties

1) If $X \geq 0$ (non-negative r.v.) then $E(X) \geq 0$.
If $X \geq 0$ then $E(X) = 0$ iff $P(X=0) = 1$.

2) For $c \in \mathbb{R}$ i.e. $c : \Omega \to \mathbb{R}$, $\omega \to c$ then $E(c) = c$ trivially.

$E(cX) = c\,E(X)$ because of sum thingys

3) For X, Y random variables, $E(X+Y) = E(X) + E(Y)$
so long as no $\infty - \infty$ shit.

Proof: $E(X+Y) = \sum_{z = x+y} z \cdot P(X+Y=z)$ assuming $E(X+Y)$ exists

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

$$= \sum_x x \boxed{\sum_y P(X=x, Y=y)} \longrightarrow P(X=x) \text{ Total probability}$$

$$+ \sum_y y \boxed{\sum_x P(X=x, Y=y)} \longrightarrow P(Y=y) \text{ Total probability}$$

$$= E(X) + E(Y)$$

2) and 3) combine for linear combinations.

$\forall c_1, \dots, c_N \quad E(c_1 X_1 + \dots + c_N X_N) = \sum_i c_i E(X_i)$

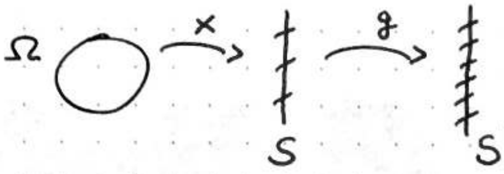
<u>Def</u>: A random variable X is said to be <u>integrable</u> if $E(|X|) < \infty$ i.e. $E(X_+), E(X_-) < \infty$

The expectation is <u>linear</u> i.e.

$$E(\lambda X + \mu Y) = \lambda E(X) + \mu E(Y)$$

as can be shown by dividing into +ve and -ve parts.

(4) Let $X : \Omega \to S$ and take a function $g : S \to S'$



Then $E(gX) = \sum_{x\in S} g(x) P(X=x)$ ← weights of distro on S

<u>Proof</u>: Let $Y = g(X) : \Omega \to S'$. Then

$$E(g(X)) = E(Y) = \sum_{y\in S'} y P(g(x)=y) = \sum_{y \in g(S)} y \underbrace{P(g(X)=y)}$$

$\{\omega : g(x)(\omega) = y\}$
$= \{\omega : \cancel{g}\, X(\omega) \in g^{-1}(y)\}$

$$= \underbrace{\sum_{y\in \cancel{\$}\, g(S)} \sum_{x \in g^{-1}(\cancel{\$}\, y)}}_{\sum_{x\in S}} g(x) P(X=x)$$

so get $\sum_{x\in S} g(x) P(X=x)$. □

<u>Remark</u>: Important examples are obtained by taking $g(x) = x^k$. Then $E(g(X)) = E(X^k)$ is called the <u>k^{th} moment</u> of the r.v. X, given by

$$E(X^k) = \sum_{x\in S} x^k P(X=x).$$

(5) Let X be a non-negative r.v. taking integer values, i.e. $X \cancel{\#} : \Omega \to S \subseteq \mathbb{N} \cup \{0\}$.

Then $E(X) = \sum_{k=1}^{\infty} P(X \geq k)$.

<u>Proof</u>:
$$E(X) = \sum_{k=0}^{\infty} k P(X=k) = \sum_{k=1}^{\infty} \cancel{\ \ } \Big(\sum_{j=1}^{k} 1\Big) P(X=k)$$
$$= \sum_{k=1}^{\infty} \sum_{j=1}^{k} P(X=k) = \sum_{j=1}^{\infty} \sum_{k=j}^{\infty} P(X=k)$$
$$= \sum_{j=1}^{\infty} P(X \geq j)$$
:) epic style □

EX: $X \sim Geo(p)$

("X counts coin tosses until 1^{st} head")

$X : \Omega \to \mathbb{N}$

$$E(X) = \sum_{k=1}^{\infty} \underbrace{P(X \geq k)}_{\substack{\text{toss} \\ \text{coin at} \\ \text{least } k \\ \text{times}}} = \sum_{k=1}^{\infty} (1-p)^{k-1} = \frac{1}{p}$$

Ö

Functions of independent random variables

Let X, Y be two independent r.v.s taking values in S_X, S_Y.

Let $f : S_X \to S_X'$, $g : S_Y \to S_Y'$.

Then $f(X), g(Y)$ are independent.

In general, if $X_1, X_2, \dots, X_N$ are independent, so are

$f_1(X_1), \dots, f_N(X_N)$ (*)

In particular, $E(XY) = E(X)E(Y)$ and in general

$E(f(X)g(Y)) = E(f(X))E(g(Y))$.

Proof: $$E(XY) = \sum_{x \in S_X} \sum_{y \in S_Y} xy \underbrace{P(X=x, Y=y)}_{\text{splits}}$$

$$= \sum_{x \in S_X} \sum_{y \in S_Y} xy\, P(X=x)P(Y=y)$$

$$= \left(\sum_{x \in S_X} x P(X=x) \right) \left(\sum_{y \in S_Y} y P(Y=y) \right)$$

$= E(X)E(Y)$ so done. $\square$

Proof of (*). Suppose $X_1, \dots, X_N$ are independent.

Then $P(f_1(X_1) = y_1, f_2(X_2) = y_2, \dots, f_N(X_N) = y_N)$

$$= P(X_1 \in f_1^{-1}(y_1), X_2 \in f_2^{-1}(y_2), \dots, X_N \in f_N^{-1}(y_N))$$

$$= P(X_1 \in f_1^{-1}(y_1)) P(X_2 \in f_2^{-1}(y_2)) \cdots P(X_N \in f_N^{-1}(y_N)) \quad \text{by indep}$$

$$= P(f_1(X_1) = y_1) P(f_2(X_2) = y_2) \dots P(f_N(X_N) = y_N)$$ $\square$

Application: inclusion-exclusion via expectation

Observe the identity

$$\prod_{i=1}^{n}(1-x_i) = \sum_{k=0}^{n}(-1)^k \sum_{1\le i_1<\cdots<i_k\le n} x_{i_1}x_{i_2}\cdots x_{i_k}.$$

For events $A_1, A_2, \ldots, A_n$ define the rvs $X_k = \mathbb{1}_{A_k}$. Then

$$\underbrace{\prod_{i=1}^{n}\underbrace{(1-X_k)}_{\mathbb{1}_{A_k^c}}} = \sum_{k=0}^{n}(-1)^k \sum_{1\le i_1<\cdots<i_k\le n} \underbrace{X_{i_1}X_{i_2}\cdots X_{i_k}}_{\mathbb{1}_{A_{i_1}\cap\cdots\cap A_{i_k}}}$$

$$\mathbb{1}_{A_1^c\cap\cdots\cap A_n^c} = 1 - \mathbb{1}_{A_1\cup A_2\cup\cdots\cup A_n}$$

Hence $\mathbb{1}_{A_1\cup\cdots\cup A_n} = \sum_{k=1}^{n}(-1)^{k-1}\sum_{1\le i_1<\cdots<i_k\le n}\mathbb{1}_{A_{i_1}\cap\cdots\cap A_{i_k}}$.

Take E of both sides, use linearity

$$P(A_1\cup\cdots\cup A_n) = \sum_{k=1}^{n}(-1)^{k-1}\sum_{1\le i_1<\cdots<i_k\le n} \cancel{E}\,P(A_{i_1}\cap\cdots\cap A_{i_k})$$

Variance

For X an integrable random variable, define

$$Var(X) = E((X-E(X))^2) \geq 0.$$

Note that $Var(X) = E(X^2) - E(X)^2$ since

$$(X-E(X))^2 = X^2 - 2XE(X) + E(X)^2$$

Note that $E(X^2) \geq [E(X)]^2$ since $Var(X)\geq 0$.

Properties:
- $Var(X)\geq 0$
- $Var(X) = 0 \iff P(X=E(X)) = 1$
- $Var(cX) = E((cX - cE(X))^2) = c^2 Var(X)$
- $Var(X+c) = Var(X)$
- X, Y indep $\Rightarrow Var(X+Y) = Var(X) + Var(Y)$ obvious generalisation

$cov(X,X) = Var(X)$

L12.1

<u>Ex:</u> $X \sim \text{Poisson}(\lambda)$

Know $E(X) = \lambda$.

$$E(X^2) = \sum_{k=0}^{\infty} k^2 P(X=k) = \sum_{k=1}^{\infty} k^2 \frac{e^{-\lambda}\lambda^k}{k!}$$

$$k(k-1)+k$$

$$= \sum_{k=1}^{\infty} \underbrace{\frac{k(k-1)e^{-\lambda}\lambda^k}{k!}}_{\frac{e^{-\lambda}\lambda^k}{(k-2)!}} + \underbrace{\sum_{k=1}^{\infty} \frac{k e^{-\lambda}\lambda^k}{k!}}_{E(X)}$$

$$= \sum_{k=2}^{\infty} \frac{e^{-\lambda}\lambda^2\lambda^{k-2}}{(k-2)!} + \lambda = \lambda^2 \underbrace{\sum_{k=0}^{\infty} \frac{e^{-\lambda}\lambda^k}{k!}}_{1} + \lambda$$

Hence $Var(X) = \cancel{\lambda^2} + \lambda - \cancel{\lambda^2} = \lambda$.

<u>Def^n:</u> The <u>standard deviation</u> of a random variable X is defined as $\sqrt{Var(X)}$.

<u>Covariance</u>

<u>Def^n:</u> The covariance of two random variables X, Y with finite mean is defined as $Cov(X,Y) = E[(X-E(X))(Y-E(Y))] = E(XY) - E(X)E(Y)$

<u>Properties:</u>
- $Cov(X,Y) = Cov(Y,X)$
- $Cov(X,X) = Var(X)$
- $Cov(X,c) = 0$
- $Cov(cX,Y) = c\,Cov(X,Y)$
- $Cov(X+Z,Y) = Cov(X,Y) + Cov(Z,Y)$

Therefore Covariance is <u>bilinear</u>, that is

$$Cov\left(\sum_{i=1}^{n} a_i X_i, \sum_{j=1}^{m} b_j Y_j\right) = \sum_{i=1}^{n}\sum_{j=1}^{n} a_i b_j\, Cov(X_i, Y_j).$$

- $Var(X+Y) = Var(X) + 2Cov(X,Y) + Var(Y)$

<u>Proof</u>:
$$\begin{aligned} Var(X+Y) &= E\left[(X+Y - E(X) - E(Y))^2\right] \\ &= E\left[\left((X-E(X)) + (Y-E(Y))\right)^2\right] \\ &= E\left[(X-E(X))^2 + 2(X-E(X))(Y-E(Y)) + (Y-E(Y))^2\right] \\ &= Var(X) + 2Cov(X,Y) + Var(Y). \quad \square \end{aligned}$$

- If X, Y independent, $E(XY) = E(X)E(Y) \Rightarrow Cov(X,Y) = 0$ so formula from before holds.

<u>Remark</u> X, Y independent $\not\Longleftarrow\Longrightarrow$ $Cov(X,Y) = 0$

<u>EX</u> Let X_1, X_2, X_3 be independent Bernoulli rvs of parameter $\frac{1}{2}$

Let $Y_1 = 2X_1 - 1$, $Y_2 = 2X_2 - 1$

$$Y_1 = \begin{cases} 1 & \text{wp } 1/2 \\ -1 & \text{wp } 1/2 \end{cases} \qquad Y_2 = \begin{cases} 1 & \text{wp } 1/2 \\ -1 & \text{wp } 1/2 \end{cases}$$

Then $E(Y_1) = E(Y_2) = 0$.

Let $Z_1 = Y_1 X_3$, $Z_2 = Y_2 X_3$.

Then
$$\begin{aligned} Cov(Z_1, Z_2) &= E(Z_1 Z_2) - E(Z_1)E(Z_2) \\ &= E(Y_1 Y_2 X_3^2) - E(Y_1 X_3)E(Y_2 X_3) \\ &\overset{\text{independence}}{=} E(\cancel{Y_1})E(\cancel{Y_2})E(X_3) - E(\cancel{Y_1})E(X_3)E(\cancel{Y_2})E(X_3) \\ &= 0. \end{aligned}$$

But Z_1, Z_2 are not independent, since

$P(Z_1 = 0, Z_2 = 0) = P(X_3 = 0) = \frac{1}{2}$

$P(Z_1 = 0) = P(X_3 = 0) = \frac{1}{2} = P(Z_2 = 0)$

and $\frac{1}{2} \neq \frac{1}{2} \cdot \frac{1}{2}$.

L12.3

Inequalities

1) Markov's Inequality

Let $X \geq 0$ be an r.v. Then for any $\lambda > 0$,

$$P(X \geq \lambda) \leq \frac{1}{\lambda} E(X).$$

Proof: Take $Y = \lambda \mathbb{1}_{\{X \geq \lambda\}}$.

Then $Y \leq X$ since

if $X < \lambda$ then $Y = 0 \leq X$, and

if $X \geq \lambda$ then $Y = \lambda \leq X$.

Dark af

Therefore $E(Y) \leq E(X)$

$$\Rightarrow \lambda P(X \geq \lambda) \leq E(X)$$

$$\Rightarrow P(X \geq \lambda) \leq \frac{1}{\lambda} E(X).$$

$\square$

2) Chebyshev's inequality

For X r.v. with finite mean $E(X)$, and $\lambda > 0$, then

$$P(|X - E(X)| \geq \lambda) \leq \frac{Var(X)}{\lambda^2}.$$

Proof: $P(|X - E(X)| \geq \lambda) = P(\underbrace{(X - E(X))^2}_{\text{non-negative r.v.}} \geq \lambda^2)$

Nice

$$\leq \frac{1}{\lambda^2} E((X - E(X))^2) \quad \text{by Markov}$$

$$= \frac{1}{\lambda^2} Var(X).$$

$\square$

3) Cauchy - Schwarz inequality

For X, Y random variables,

$$E(|XY|) \leq \sqrt{E(X^2)E(Y^2)}$$

Proof: It suffices to prove the above for X, Y non-negative r.v.s with $E(X^2)$ finite, $E(Y^2)$ finite.

L12.4

Start from $XY \leq \frac{1}{2}(X^2 + Y^2)$ AMGM

take expectations $E(XY) \leq \frac{1}{2}[E(X^2) + E(Y^2)]$

If $E(X^2) = E(Y^2) = 0$, inequality obvious.

So assume $P(Y > 0) > 0$ WLOG.

For $t \in \mathbb{R}$, look at

$$0 \leq (X - tY)^2 = X^2 - 2tXY + t^2 Y^2$$

Take expectations

$$0 \leq E(X^2) - 2t\,E(XY) + t^2 \underbrace{E(Y^2)}_{>0}.$$

discriminant $4E(XY)^2 - 4E(X^2)E(Y^2) \leq 0$ since at most one root

Alternatively, plug in $t = \frac{E(XY)}{E(Y^2)}$. □

<u>Note</u>: equality in CS when X is a multiple of Y.

More formally if $P(X = tY) = 1$.

<u>Def</u>

<u>Def^{n}</u> for X, Y random variables with positive variance, define the correlation coefficient

$$\mathrm{Corr}(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)}\sqrt{\mathrm{Var}(Y)}}$$

Then $\mathrm{Corr}(X,Y) \in [-1,1]$ ~~since~~ by applying CS to the rvs $\overline{X} = X - E(X)$, $\overline{Y} = Y - E(Y)$

! also need $E(|X|) \geq |E(X)|$

L13.1

Markov's inequality

$X \geq 0$, $\lambda \in (0,\infty)$

$$P(X \geq \lambda) \leq \frac{1}{\lambda} E(X)$$

Chebyshev's inequality

$$P(|X - E(X)| > \lambda) \leq \frac{1}{\lambda^2} Var(X)$$

Cauchy - Schwarz

$$E(|XY|) \leq \sqrt{E(X^2)E(Y^2)}$$

$$Corr(X,Y) := \frac{Cov(X,Y)}{\sqrt{Var(X)\,Var(Y)}} \in [-1,1]$$

Jensen's inequality

If $f: I \to \mathbb{R}$ is convex and X is an integrable r.v. on I, then $E(f(X)) \geq f(E(X))$.

Can remember sense of inequality via

$$E(X^2) \geq [E(X)]^2.$$

Recall that f is said to be convex on I if for all $x,y \in I$, $t \in [0,1]$,

$$f(tx + (1-t)y) \leq t f(x) + (1-t) f(y).$$

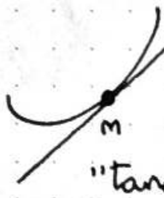
If f is twice differentiable, convexivity $\Leftrightarrow f'' \geq 0$ on I.

f can be convex without being diff'ble, see $f(x) = |x|$.

Observe the following:

for any $m \in I$ then there exist $a,b \in \mathbb{R}$ such that

$$f(m) = am+b, \quad f(x) \geq ax+b \text{ for all } x \in I$$



Indeed, given m, for all $x,y \in I$ such that $x < m < y$, there is a $t \in (0,1)$ with $m = tx + (1-t)y$.

Plugging this into def^n of convexity,

$$f(m) \leq t f(x) + (1-t) f(y)$$

$$\Rightarrow f(m) - f(x) \leq (1-t)(f(y) - f(x)) \qquad \text{since } 1-t = \frac{m-x}{y-x}$$

$$\Rightarrow \frac{f(m)-f(x)}{m-x} \leq \frac{f(y)-f(x)}{y-x}$$

$\Rightarrow \cancel{f(m) \leq}\ \dfrac{f(m)-f(x)}{m-x} \leq a \leq \dfrac{f(y)-f(x)}{y-x}$

no

$\Rightarrow f(m) \leq a(m-x) + f(x)$

$\Rightarrow f(x) \geq ax \underbrace{-am + f(m)}_{b}$.

"$\square$"

<u>Proof of Jensen</u>

Take $m = E(X)$. Then by the above property, $\exists\, a, b \in \mathbb{R}$ such that

$f(m) = am + b, \qquad f(X) \geq aX + b$

Take expectations to get

$$E(f(X)) \geq E(aX+b) = aE(X) + b = am + b$$
$$= f(m) = f(E(X)).$$

$\square$

<u>Application</u> AM GM inequality

AM: arithmetic mean; GM: geometric mean

Take X to be a random variable that takes values $x_1, \dots, x_n \in \mathbb{R}$ each with probability $\frac{1}{n}$.

If f is a convex function on I containing $x_1, \dots, x_n$, then

$$E(f(X)) \geq f(E(X))$$

$$E(f(X)) = \sum_{k=1}^{n} f(x_k) P(X = x_k) = \frac{1}{n}\sum_{k=1}^{n} f(x_k)$$

$$f(E(X)) = f\left(\sum_{k=1}^{n} x_k P(X = x_k)\right) = f\left(\frac{1}{n}\sum_{k=1}^{n} x_k\right)$$

Pick $f: \mathbb{R}^+ \to \mathbb{R}$, $x \to -\log x$. Convex by inspection. Get

$$\frac{1}{n}\sum_{k=1}^{n}(-\log x_k) \geq -\log\left(\frac{1}{n}\sum_{k=1}^{n} x_k\right)$$

$$\frac{1}{n}\sum_{k=1}^{n}(-\log x_k) = -\frac{1}{n}\log\left(\prod_{k=1}^{n} x_k\right) = -\log\left(\left(\prod_{k=1}^{n} x_k\right)^{1/n}\right)$$

L13.3

$$\Rightarrow \log\left(\underbrace{\left(\prod_{k=1}^{n} x_k\right)^{1/n}}_{GM}\right) \leq \log\left(\underbrace{\frac{1}{n}\sum_{k=1}^{n} x_k}_{AM}\right)$$

We get $$\boxed{\left(\prod_{k=1}^{n} x_k\right)^{1/n} \leq \frac{1}{n}\sum_{k=1}^{n} x_k}$$

Random walks

Def.^n A stochastic / random process is a sequence of random variables $(X_n)_{n\in\mathbb{N}}$. Think of n as discrete time, with X_n being the state of the process at time n.

An integer valued process is said to be a random walk if it is of the form

$$X_n = x + Y_1 + Y_2 + \cdots + Y_n$$

for Y_i identically distributed, independent random variables.

If $Y_i \in \{-1, 1\}$ for all i, then $(X_n)_{n\geq 0}$ is a simple random walk.

The Y_is are called the steps (or jumps, or increments) of the process.

In Ex Sheet 1:

$X_n = Y_1 + \cdots + Y_n$

where $Y_i = \begin{cases} +1 & \text{wp } 1/2 \\ -1 & \text{wp } 1/2 \end{cases}$ ← simple symmetric random walks

A gambler has fortune £$x \in \mathbb{N}$.

If $x > 0$, he bets £1 on heads:

- if he wins, his fortune increases by 1
- if he loses, his fortune decreases by 1

His goal is to get to fortune $a > x$.

Let $X_n =$ fortune of the gambler after n bets.

Then $X_n = x + Y_1 + \cdots + Y_n$ where $(Y_n)_{n\geq 1}$ are iid with $P(Y_i = 1) = p$, $P(Y_i = -1) = 1 - p = q$

L3.4

1) What is the probability that the gambler gets to fortune a before he runs out of money?

Let h_x be the probability of winning, starting from x.

Then $h_0 = 0$, $h_a = 1$.

For $1 \leq x \leq a-1$,

$$\begin{aligned} h_x &= P(\text{win} \mid Y_1 = 1)\, P(Y_1 = 1) + P(\text{win} \mid Y_1 = -1)\, P(Y_1 = -1) \\ &= h_{x+1} \cdot p + h_{x-1} \cdot (1-p). \end{aligned}$$

We have boundary conditions, so can solve recurrence relation.

- $\underline{p = q = \frac{1}{2}}$

$$2h_x = h_{x+1} + h_{x-1} \Rightarrow h_{x+1} - h_x = h_x - h_{x-1} \qquad \text{"linear"}$$

So solution must be linear.

Boundary conditions give $\boxed{h_x = \frac{x}{a}}$.

- $\underline{p \neq q}$

$h_x = h_{x+1} p + h_{x-1} q$

look for solution of form $h_x = \lambda^x$

$\Rightarrow \lambda = p\lambda^2 + q \Rightarrow \lambda = 1$ or $\lambda = q/p \neq 1$

Therefore general solution is

$h_x = A + B\left(\frac{q}{p}\right)^x$

Imposing $h_0 = 0 \Rightarrow A + B = 0 \Rightarrow B = -A$

$h_a = 1 \Rightarrow A + B\left(\frac{q}{p}\right)^a = 1 \Rightarrow A = \dfrac{1}{1 - \left(\frac{q}{p}\right)^a}$

Hence

$$h_x = \frac{1 - \left(\frac{q}{p}\right)^x}{1 - \left(\frac{q}{p}\right)^a}.$$

2) How long does the game last? Let

$T = \min\{n \geq 0 : X_n \in \{a, 0\}\}$

$\tau_x = E_x(T)$

Observe that $\tau_0 = \tau_a = 0$

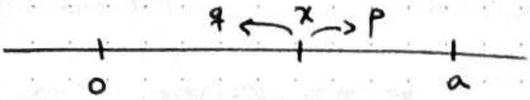
For $1 \leq x \leq a-1$, get

$\tau_x = \underline{\underline{1}} + p\tau_{x+1} + q\tau_{x-1}$

difference w/ previous relation

L14.1

Simple Random Walk on $\mathbb{Z}$



- $h_x = p h_{x+1} + q h_{x-1}$

$p = \frac{1}{2}$: $h_x = A + Bx$

$p \neq q$: $h_x = A + B\left(\frac{q}{p}\right)^x$

- $\tau_x = E_x(T)$

$\tau_x = 1 + p\tau_{x+1} + q\tau_{x-1}$

$p = \frac{1}{2}$

Look for p.i. Cx^2

$$Cx^2 = 1 + p \cdot C(x+1)^2 + q \cdot C(x-1)^2$$
$$= Cx^2 + C + 1$$

$\Rightarrow C = -1$

So general solution $\tau_x = -x^2 + A + Bx$

Impose $\tau_0 = \tau_a = 0$ to get

$\tau_x = x(a-x)$

$p \neq q$

Look for p.i. Cx

$$Cx = 1 + p \cdot C(x+1) + q \cdot C(x-1)$$
$$= 1 + Cx + C(p-q)$$

$\Rightarrow C = \dfrac{1}{q-p}$

So general solution is

$$\tau_x = \frac{1}{q-p}x + A + B\left(\frac{q}{p}\right)^x$$

Impose $\tau_0 = \tau_a = 0$ to get

$$\tau_x = \frac{x}{q-p} - \frac{a}{q-p}\,\frac{(q/p)^x - 1}{(q/p)^a - 1}$$

CONDITIONAL EXPECTATION

<u>Def</u>: For a random variable X and an event A with $P(A) > 0$, define

$$E(X \mid A) = \frac{E(X \cdot \mathbb{1}_A)}{P(A)}.$$

Note that if $X = \mathbb{1}_B$,

$$E(\mathbb{1}_B \mid A) = \frac{E(\mathbb{1}_A \cdot \mathbb{1}_B)}{P(A)} = \frac{P(A \cap B)}{P(A)}$$

$\parallel$

$P(B \mid A)$

Law of total expectation

For a collection of disjoint events $(B_n)_{n\geq 1}$ s.t. $\bigcup_{n\geq 1} B_n = \Omega$, with $P(B_n) > 0$ for all $n \geq 1$, and X a random variable, we have

$$E(X) = \sum_{n\geq 1} E(X \mid B_n)\, P(B_n)$$

This follows from countable additivity of the expectation:

take $(X_n)_{n\geq 1}$ random variables, and let

$$X = \sum_{n\geq 1} X_n$$

Then $E(X) = \sum_{n\geq 1} E(X_n)$.

Indeed, if the X_n are non-negative, we have

$$\begin{aligned} E(X) &= \sum_{\omega\in\Omega} X(\omega)\cdot P(\{\omega\}) \\ &= \sum_{\omega\in\Omega} \sum_{n\geq 1} X_n(\omega)\cdot P(\{\omega\}) \\ &= \sum_{n\geq 1} \underbrace{\sum_{\omega\in\Omega} X_n(\omega)\cdot P(\{\omega\})}_{E(X_n)}. \end{aligned}$$

can do since everything is non-negative

Proof of law of total expectation

Write

$$\begin{aligned} X &= X\cdot \mathbb{1}_\Omega \\ &= X\cdot \mathbb{1}_{(\bigcup_{n\geq 1} B_n)} \\ &= X\cdot \Big(\sum_{n\geq 1} \mathbb{1}_{B_n}\Big) \\ &= \sum_{n\geq 1} \underbrace{X\cdot \mathbb{1}_{B_n}}_{X_n}. \end{aligned}$$

B_n disjoint

Take expectations and use countable additivity

$$E(X) = E\Big(\sum_{n\geq 1} X_n\Big) = \sum_{n\geq 1} E(X_n)$$

$$\Rightarrow E(X) = \sum_{n\geq 1} E(X\cdot \mathbb{1}_{B_n}) = \sum_{n\geq 1} E(X\mid B_n)\cdot \mathbb{P}(B_n) \quad \square$$

PROBABILITY GENERATING FUNCTIONS

Let X be a random variable in $\mathbb{N}\cup\{0\}$.

Def: The probability generating function of X is defined by

$$G_X(t) = E(t^X) = \sum_{n\geq 0} t^n \cdot P(X=n)$$

Note that $G_X(1) = 1$, and the radius of convergence of power series is therefore at least one.

Remark: The pgf of X uniquely determines its distro, since $G_X(t)$ defines a smooth function on $[-1,1]$, and so $(D^k G_X)(0) = k!\cdot P(X=k)$. Recover weights.

EX: $X\sim \text{Bernoulli}(p)$

$$G_X(t) = E(t^X) = t^1 P(X=1) + t^0 P(X=0) = \cancel{t^1}\, tp + (1-p)$$

EX: $X\sim \text{Geo}(p)$

$$G_X(t) = E(t^X) = \sum_{k\geq 1} \cancel{p}\, t^k (1-p)^{k-1} p$$

$$= tp \sum_{k\geq 0} [t(1-p)]^k = \frac{tp}{1-t(1-p)} \quad \boxed{\text{provided } |t(1-p)| < 1}$$

EX: $X\sim \text{Poi}(\lambda)$

$$G_X(t) = E(t^X) = \sum_{k\geq 0} t^k \cdot \frac{e^{-\lambda}\lambda^k}{k!} = e^{-\lambda}\cdot e^{t\lambda} \quad \ddot{\smile}$$

Note that if $X_1, X_2, \ldots X_n$ are indep. rvs, then if we set

$$S = X_1 + X_2 + \cdots + X_n,$$

then
$$G_S(t) = E\left(t^{\sum_{k=1}^n X_k}\right) = E\left(t^{X_1}\cdot t^{X_2}\cdots t^{X_n}\right)$$
$$= E(t^{X_1})\, E(t^{X_2}) \cdots E(t^{X_n})$$
$$= G_{X_1}(t)\, G_{X_2}(t) \cdots G_{X_n}(t).$$

Moreover if the X_n are identically distributed,

$$G_S(t) = \left(G_X(t)\right)^n.$$

EX: $X \sim \text{Binomial}(N, p)$

Then $G_X(t) = E(t^X) = \sum_{k \geq 0} t^k P(X = k)$

$$= \sum_{k \geq 0} t^k \binom{N}{k} p^k (1-p)^{N-k}$$

$$= (tp + 1 - p)^N$$

which is same as Bernoulli to the N^{th} power.

Therefore sum of iid Bernoulli(p) is Binomial (N, p)

Remark: provided the pgf $G_X(t)$ of X is a power series with radius of convergence strictly larger than 1, we can differentiate at $t=1$ to get:

- $G_X(1) = E(1^X) = 1$

- $G_X'(1) = \frac{d}{dt}\left[\sum_{n \geq 0} t^n P(X=n)\right]_{t=1}$

$$= \left[\sum_{n \geq 0} n t^{n-1} P(X=n)\right]_{t=1}$$

$$= \sum_{n \geq 1} n\, P(X=n)$$

$$= E(X)$$

- $G_X''(1) = \sum_{n \geq 0} n(n-1) t^{n-2} P(X=n)$

$$= E(X(X-1))$$

in general $G^{(k)}(1) = E(X(X-1)\cdots(X-k+1))$.

If the radius of convergence is exactly 1, then the same result holds with $\lim_{t \to 1} G_X^{(k)}(t)$ in place of $G_X^{(k)}(1)$.

G_{X^n}?

E14.5

Applications

1) Random sums

Let $(X_n)_{n\geq 1}$ be iid rvs with common pgf $G_X(t)$.

Let N be a further rv independent of the X_n.

Define the rv

$$\underset{\overset{\uparrow}{\text{rv}}}{S} = \sum_{n=1}^{N \leftarrow \text{rv}} X_n \swarrow \text{rv}$$

Then
$$G_S(t) = E(t^S) = E\left(t^{\sum_{n=1}^{N} X_n(\omega)}\right)$$
$$= \sum_{n=\cancel{0}1}^{\infty} E\left(t^{\sum_{k=1}^{N} X_k(\omega)} \,\middle|\, N=n\right) \cdot P(N=n)$$
$$= \sum_{n=1}^{\infty} \underbrace{E\left(t^{\sum_{k=1}^{n} X_k}\right)}_{[G_X(t)]^n} \cdot P(N=n)$$
$$= \sum_{n=1}^{\infty} \left(G_X(t)\right)^n \cdot P(N=n)$$
$$= G_N(G_X(t))$$

epic style

Differentiate to get

$$\underset{\underset{E(S)}{\parallel}}{G_S'(1)} = G_N'(\underbrace{G_X(1)}_{1}) \cdot G_X'(1) = E(N)\,E(X)$$

L15.1

Prob. generating functions

The pgf of X is

$$G_X(t) = E(t^X) = \sum_{n=0}^{\infty} t^n P(X=n)$$

$$\Rightarrow G_X(1) = 1$$

Moreover, can recover

$$P(X=k) = \frac{1}{k!} G_X^{(k)}(0)$$

$$E(X) = G_X'(1)$$

$$E(X(X-1)) = G_X''(1)$$

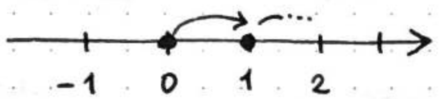
X, Y indep. $\Rightarrow G_{X+Y}(t) = G_X(t)\, G_Y(t)$

Counting non-negative paths

Let P_n be the set of paths

$$(x_0, x_1, \dots, x_{2n})$$

such that $x_0 = x_{2n} = 0$, $|x_i - x_{i-1}| = 1$, $x_i \geq 0 \ \forall i$.



Let $C_n = |P_n|$. Want to compute C_n.

Note that $x_1 = 1 = x_{2n-1}$

Let k be the first positive i such that $x_{2i} = 0$.

So $x_{2k-1} = 1$, and can decompose path into

$$\underbrace{\underset{0}{x_0}, \underset{1}{x_1}, \dots, \underset{1}{x_{2k-1}}}, \underbrace{\underset{0}{x_{2k}}, \dots, \underset{0}{x_{2n}}}_{\text{in } P_{n-k}}$$

$x_1 - 1, \dots, x_{2k-1} - 1 \in P_{k-1}$

By the multiplication rule,

$$C_n = \sum_{k=1}^{n} C_{k-1} C_{n-k}.$$

Introduce the power series $c(t) = \sum_{n=0}^{\infty} C_n t^n$.

Note that $c(0) = C_0 = 1$.

Note that radius of convergence of $c(t)$ is at least $1/4$,

since $C_n \leq 2^{2n}$ so use ratio, comparison tests

Moreover,

$$c(t) = \sum_{n=0}^{\infty} C_n t^n = \sum_{n=0}^{\infty} \sum_{k=1}^{n} C_{k-1} C_{n-k} t^n$$

$$= 1 + \sum_{n=1}^{\infty} \sum_{k=1}^{n} C_{k-1} C_{n-k} t^{k-1} \cdot t \cdot t^{n-k}$$

$$\cancel{= 1 + t \sum_{n=1}^{\infty} C_{k-1} t^{k-1} \sum_{k=1}^{n} C_{n-k} t^{n-k}}$$

$$= 1 + t \underbrace{\sum_{k=1}^{\infty} C_{k-1} t^{k-1}}_{c(t)} \underbrace{\sum_{n=k}^{\infty} C_{n-k} t^{n-k}}_{c(t)}$$

$$= 1 + t\,[c(t)]^2$$

Solve quadratic to get

$$c(t) = \frac{1 \pm \sqrt{1-4t}}{2t}$$

Know $c(0) = 1$ to exclude the + solution

$$c(t) = \frac{1 - \sqrt{1-4t}}{2t}$$

So $\sqrt{1-4t} = 1 - 2t\,c(t)$

$$\sqrt{1-4t} = 1 - 2\sum_{n=1}^{\infty} \frac{1}{n}\binom{2n-2}{n-1} t^n$$

$\Rightarrow$

$$\cancel{t}\,c(t) = \sum_{n=1}^{\infty} \frac{1}{n}\binom{2n-2}{n-1} t^{n-1}$$

$$C_n = \frac{1}{n+1}\binom{2n}{n}$$

Catalan Numbers

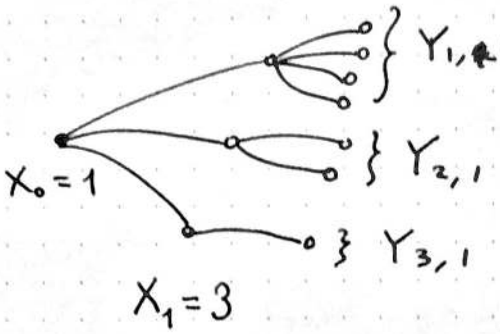
BRANCHING PROCESSES

Let $(X_n)_{n \geq 0}$ be a random process s.t.

$X_0 = 1$ (1 individual at time 0)

$$X_{n+1} = \begin{cases} Y_{1,n} + Y_{2,n} + \cdots + Y_{X_n,n} & \text{if } X_n \geq 1 \\ 0 & \text{if } X_n = 0 \end{cases}$$

where $(Y_{k,n})$ are iid rvs with same distro as X_1



It's branchin

$X_n = \#$ individuals alive at time n

$X_1 = Y_{1,0}$

We call the distribution of X_1 the offspring distro.

We say that the process is extinct at time n if $X_n = 0$.

Remark: Note that on the event $\{X_n = m\}$ $\underset{\uparrow 1?}{}$ yes

$$X_{n+1} = \sum_{j=1}^{m} X_n^{(j)}$$

where $X_n^{(1)}, \dots, X_n^{(m)}$ are iid branching processes with the same offspring distribution.

Mean population size

Let $E(X_1) = \mu \in (0, \infty)$.

Want to show that

$$E(X_n) = \mu^n, \qquad (**)$$

do it by induction.

Take $n=1$ to get $E(X_1) = \mu$.

Assume $(**)$ holds for some $n \geq 1$. Then

$$E(X_{n+1}) = \sum_{m=0}^{\infty} E(X_{n+1} \mid X_n = m)\, P(X_n = m) \qquad \text{Law of total exp}^n$$

$$= \sum_{m=0}^{\infty} E\Big(\sum_{k=1}^{m} Y_{k,n} \,\Big|\, X_n = m\Big)\, P(X_n = m)$$

$$= \sum_{m=0}^{\infty} E\Big(\sum_{k=1}^{m} Y_{k,n}\Big)\, P(X_n = m)$$

$$= \sum_{m=0}^{\infty} \sum_{j=1}^{m} E(X_1) \cdot P(X_n = m)$$

$$= \sum_{m=0}^{\infty} \mu m \, P(X_n = m) = \mu E(X_n) = \mu^{n+1} \quad \ddot{\smile}$$

<u>Probability generating function</u>

Let $G_n(t) = E(t^{X_n})$.

Write $G(t)$ in place of $G_1(t)$.

Note that $G_0(t) = E(t^{X_0}) = t$.

For $n \geq 0$,

$$\begin{aligned} G_{n+1}(t) &= E\left(t^{X_{n+1}}\right) \\ &= \sum_{m=0}^{\infty} E\left(t^{X_{n+1}} \mid X_n = m\right) P(X_n = m) \\ &= \sum_{m=0}^{\infty} E\left(t^{Y_{1,n} + \cdots + Y_{m,n}}\right) P(X_n = m) \\ &= \sum_{m=0}^{\infty} \Big(\underbrace{E(t^{Y_{1,n}})}_{G(t)} \Big)^m P(X_n = m) \\ &= G_n\big(G(t)\big). \end{aligned}$$

Then iterate to get

$$G_n(t) = \underbrace{G \circ G \circ \cdots \circ G}_{n \text{ times}}(t)$$

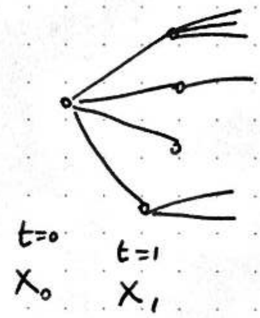
Can recover

$$\begin{aligned} E(X_n) = G_n'(1) &= \big(G_{n-1}(G(t)) \big)' \text{ at } t = 1 \\ &= G_{n-1}'(\underbrace{G(1)}_{1}) \cdot G'(1) \\ &= E(X_{n-1}) \cdot E(X_1) \end{aligned}$$

and do as before.

Branching Processes

$(X_n)_{n\geqslant 0} \qquad X_0 = 1$



$$G(t) = E(t^{X_1}) \quad \text{pgf of the offspring distro}$$

$$= \sum_{n=0}^{\infty} t^n P(X_1 = n)$$

$$G(0) = P(X_1 = 0)$$

Showed that $E(X_n) = \mu^n \quad \forall n \geqslant 0$

where $\mu = E(X_1)$.

We showed that if $G_n(t) = E(t^{X_n})$

then $G_n(t) = \underbrace{G\circ \cdots \circ G}_{n \text{ times}}(t)$.

Extinction probability

Assume that $P(X_1 = 0) > 0$.

For $n \geqslant 0$ let

$q_n = P(X_n = 0)$

the process is extinct at time n

$$q = P(X_n = 0 \text{ for some } n \geqslant 0)$$

Note that

$$\{X_n = 0\} \subseteq \{X_{n+1} = 0\} \subseteq \cdots$$

so $q_n \leqslant q_{n+1}$.

Moreover $\{X_n = 0 \text{ for some } n \geqslant 0\} = \bigcup_{n\geqslant 0} \{\overset{A_n}{X_n = 0}\}$.

So $q = P\left(\bigcup_{n\geqslant 0} A_n\right) = \lim_{n\to\infty} P(A_n) = \lim_{n\to\infty} q_n$ by continuity.

Moreover, $q_{n+1} = P(X_{n+1} = 0) = G_{n+1}(0) = G(G_n(0)) = G(q_n)$

Send n to infinity on both sides:

$q = G(q)$ using G continuous

The extinction prob. is a fixed point of G.

THEOREM

The extinction probability q of a Galton-Watson process $(X_n)_{n\geq 0}$ is the smallest non-negative solution to $q = G(q)$.

Moreover, provided $P(X_1 = 1) < 1$, we have

$$q < 1 \quad \text{iff} \quad \mu > 1$$

where $\mu := E(X_1)$.

Proof: We know that q is solution to $q = G(q)$ and $q \in [0,1]$.

Let s denote the $\underbrace{\text{smallest non-negative}}_{\text{(any)}}$? solution to $s = G(s)$.

Want to show that $q \leq s$.

Note that G defines a continuous function on $[0,1]$, which is non-decreasing since $G(t) = \sum_{n\geq 0}^{\infty} t^n \underbrace{P(X_1 = n)}_{\geq 0}$ and $G(1) = 1$.

Note that $q_0 = P(X_0 = 0) = 0 \leq s$.

Moreover, by induction if $q_n \leq s$ then

$$q_{n+1} = G(q_n) \underset{\substack{\uparrow \\ \text{non dec.}}}{\leq} G(s) = s$$

So we conclude that $q_n \leq s$ for all $n \geq 0$, and therefore

$$q = \lim_{n\to\infty} q_n = \sup\{q_n\} \leq s.$$

For the second part, let us start with $\mu > 1$. Then

$$\mu = E(X_1) = \lim_{t \nearrow 1} G'(t) > 1$$

This implies, together with $G(1) = 1$, implies that $G(t) < t$ for t sufficiently close to 1 (MVT).

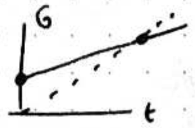
On the other hand, $G(0) = P(X_1 = 0) \geq 0$, so by the intermediate value theorem there exists $t \in [0,1)$ with $G(t) = t$. So $q < 1$.

Conversely, suppose $\mu \leq 1$.

If $P(X_1 \leq 1) = 1$, then $P(X_1 = 1) = \mu$, $P(X_0 = 1) = 1 - \mu$

$\Rightarrow G(t) = (1-\mu) + \mu t$ which if $\mu \neq 1$ $(P(X_1 = 1) < 1)$ has only

only one fixed point, $t=1$.



Finally, consider $P(X_1 \geq 2) > 0$.

Then for $t \in (0,1)$ differentiate $G(t)$ term by term to get

$$G'(t) = \sum_{n=0}^{\infty} n \underbrace{t^{n-1}}_{<1\ (n \geq 2)} P(X_1 = n) < \sum_{n=0}^{\infty} n P(X_1 = n) = \mu \leq 1.$$

Therefore the function $F(t) = G(t) - t$

is strictly decreasing on $[0,1]$. (MVT)

Moreover, $F(0) = G(0) = P(X_1 = 0) > 0$

$$F(1) = G(1) - 1 = 0.$$

Therefore 1 is the smallest t with $F(t) = 0$, i.e. $q = 1$. $\square$

<u>EX</u>: Let $(X_n)_{n \geq 0}$ be a Galton-Watson process with offspring distro

$$P(X_1 = 0) = \frac{1}{3} \qquad P(X_1 = 2) = \frac{2}{3}$$

Then $G(t) = E(t^{X_1}) = \frac{1}{3} + \frac{2}{3}t^2$.

Solve $G(t) = t$

i.e. $\frac{2}{3}t^2 - t + \frac{1}{3} = 0$

to get $(t-1)(\frac{2}{3}t - \frac{1}{3}) = 0$

$\Rightarrow q = \frac{1}{2}$.

<u>Continuous probability distributions</u>

<u>Def</u>: A function $f: \mathbb{R} \to \mathbb{R}$ s.t. $f(x) \geq 0$ for all $x \in \mathbb{R}$, and

$\int_{\mathbb{R}} f(x)\, dx = 1$, is called a probability density function.

From f we can define a probability measure on $\mathbb{R}$ by setting

$$\mu((-\infty, x]) = \int_{-\infty}^{x} f(x)\, dx \quad \text{for all } x \in \mathbb{R}.$$

Note that $\mu(\mathbb{R}) = \int_{-\infty}^{\infty} f(x)\, dx = 1$, and

$$\mu((a,b]) = \mu((-\infty, b]) - \mu((-\infty, a]) = \int_a^b f(x)\, dx$$

for all $a \leq b$.

L16.4 Remark:

To be precise we would have to introduce a σ-algebra on $\mathbb{R}$, called a <u>Borel σ-algebra</u>, and define μ on all sets A in the Borel σ-algebra by setting

$$\mu(A) = \int_A f(x)\,dx.$$

It suffices to know that it contains <u>all</u> intervals (in fact it is the smallest σ-algebra containing all intervals).

<u>Terminology</u>: Throughout the rest of the course, when we say "A Borel set" we mean $A \in$ Borel σ-algebra, and hence a set on which $\mu(A)$ is defined.

17.1

Continuous prob. distributions

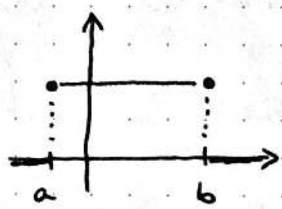
<u>Def</u>: $f : \mathbb{R} \to \mathbb{R}$ is a pdf if $f(x) \geq 0 \ \forall x \in \mathbb{R}$, and
$$\int_{\mathbb{R}} f(x)\,dx = 1.$$

Build a prob. measure μ on $\mathbb{R}$ by setting
$$\mu((-\infty, x]) = \int_{-\infty}^{x} f(x)\,dx$$

In fact
$$\mu(A) = \int_A f(x)\,dx$$
for all A in the Borel σ-algebra of $\mathbb{R}$.

<u>Notation</u>: We write
$$\mathbb{1}_A(x) = \begin{cases} 1 & \text{if } x \in A, \\ 0 & \text{otherwise.} \end{cases}$$

1) For $a < b$, the pdf
$$f(x) = \frac{1}{b-a} \mathbb{1}_{[a,b]}(x)$$
defines the <u>uniform distribution</u> "on" $[a,b]$.



Note that
$$\mu(A) = \int_A f(x)\,dx = \frac{\text{"length"}(A \cap [a,b])}{b-a}.$$

2) <u>Exponential distribution</u>, defined by the pdf
$$f(x) = \lambda e^{-\lambda} \mathbb{1}_{[0,\infty)}(x)$$
where $\lambda \in (0,\infty)$ is some parameter. Note $f(x) \geq 0$ and
$$\int_{-\infty}^{\infty} f(x)\,dx = \int_0^{\infty} \lambda e^{-\lambda}\,dx = \left[-e^{-\lambda}\right]_0^{\infty} = 1$$

3) <u>Gamma distribution</u>, given by the pdf
$$f(x) = \frac{x^{\alpha-1}\lambda^{\alpha} e^{-\lambda x}}{\Gamma(\alpha)} \mathbb{1}_{[0,\infty)}(x)$$
where $\alpha, \lambda \in (0,\infty)$ are parameters, and Γ is given by

$$\Gamma(\alpha) = \int_0^\infty y^{\alpha-1} e^{-y}\, dy.$$

Note that $f(x) \geq 0$ and

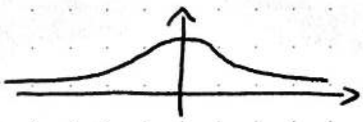
$$\int_{-\infty}^{\infty} f(x)\, dx = \int_0^\infty \frac{x^{\alpha-1}\lambda^\alpha e^{-\lambda x}}{\Gamma(\alpha)}\, dx \underset{y=\lambda x}{=} \frac{1}{\Gamma(\alpha)} \int_0^\infty \frac{y^{\alpha-1}\cancel{\lambda} e^{-y}}{\cancel{\lambda}}\, dy = 1.$$

Also, for $\alpha \in \mathbb{Z}^+$, we have $\Gamma(\alpha) = (\alpha-1)!$, which can be seen by doing integration by parts $\alpha-1$ times and noting $\Gamma(1) = 1$.

Note that for $\alpha = 1$, recover the exponential distribution of parameter λ. Cool story bro.

4) Cauchy distribution, given by the pdf

$$f(x) = \frac{1}{\pi(1+x^2)}.$$



Note $f(x) \geq 0$ and that

$$\int_{-\infty}^{\infty} f(x)\, dx = \frac{2}{\pi}\int_0^\infty \frac{dx}{1+x^2} = \frac{2}{\pi}\left[\arctan x\right]_0^\infty = 1$$

5) Gaussian / normal distribution in two flavours.

The pdf

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$$

defines the standard Gaussian distribution. For $\mu \in \mathbb{R}$, $\sigma > 0$, define the pdf $f_{\mu,\sigma}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$ for the Gaussian distribution of parameters $\underset{\text{mean}}{\mu}$ and $\underset{\text{variance}}{\sigma^2}$.

Note that $f(x) \geq 0$ and

$$\left(\int_{-\infty}^{\infty} f(x)\, dx\right)^2 = \frac{1}{2\pi}\int_0^\infty dx \int_0^\infty dy\, e^{-\frac{x^2+y^2}{2}} = \overbrace{\frac{1}{2\pi}\int_0^{2\pi} d\theta}^{1} \overbrace{\int_0^\infty dr \cdot r e^{-r^2/2}}^{1}$$

Continuous random variables

Recall that, to each random variable X we can associate a distribution function $F_X : \mathbb{R} \to [0,1]$ by setting

$$F_X(x) = P(X \leq x) \text{ for all } x \in \mathbb{R}.$$

Then $F_X(x) \longrightarrow 0$ as $x \to -\infty$

$F_X(x) \longrightarrow 1$ as $x \to \infty$

and F_X is right-continuous and non-decreasing.

<u>Def</u>: A random variable X is said to be <u>continuous</u> if its distribution function F_X is continuous. It is said to be <u>absolutely continuous</u> if there exists a pdf f_X such that

$$F_X(x) = P(X \leq x) = \int_{-\infty}^{x} f_X(y)\, dy$$

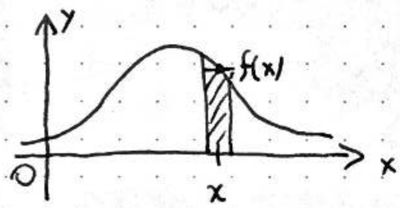
In this case, we say X has pdf f_X, and

$$P(X \in A) = \int_A f_X(y)\, dy.$$

<u>Remark</u>: Although f_X is not a probability, we have

$$P(x \leq X \leq x + \Delta x) = \int_x^{x+\Delta x} f_X(y)\, dy \underset{\substack{\uparrow \\ \text{MVT} \\ + \text{cts.}}}{\approx} f_X(x) \cdot \Delta x$$

so f_X does give the probability that X belongs to $[x, x+\Delta x]$ for small Δx.



Note that F_X is right continuous, and

$$F_X(x^-) = \lim_{n\to\infty} F_X\left(x - \tfrac{1}{n}\right)$$

$$= \lim_{n\to\infty} P\left(X \leq x - \tfrac{1}{n}\right) = P(X < x)$$

Hence $F_X(x) - F_X(x^-) = P(\{X = x\})$.

So F_X is continuous at x iff $P(\{X = x\}) = 0$.

Note that if F_X is *differentiable with piece-wise continuous derivative f, then for all $a < b$,

$$P(a < X \leq b) = F_X(b) - F_X(a) = \int_a^b f(x)\, dx$$

* piecewise?

It follows that f is the pdf of X.

On the other hand, suppose X has pdf f. Then for $h>0$

$$\left| \frac{F_X(x+h) - F_X(h)}{h} - f(x) \right| = \left| \frac{1}{h} \int_x^{x+h} f(y)\,dy - f(x) \right|$$

$$\leqslant \sup_{y \in [x, x+h]} |f(y) - f(x)| \quad \text{by triangle inequality}$$

So for $h>0$, send $h \to 0$. If f is continuous at x, this goes to 0, so ~~X~~ F_X is diff'ble, and $F_X'(x) = f(x)$.

<u>Def</u>: The <u>distribution</u> of a random variable X is the probability measure on $\mathbb{R}$ given by $\mu(A) = P(X \in A)$ for all Borel sets A.

<u>Properties of the exponential distribution</u>

1) <u>Memory-less property</u>

Let X be exponetial of parameter λ. For $s, t > 0$, we have

$$P(X > t+s \mid X > t) = \frac{P(X > t+s)}{P(X > t)} = \frac{e^{-\lambda(t+s)}}{e^{-\lambda}} = e^{-\lambda s}$$

Now $P(X>t) = \int_t^\infty \lambda e^{-\lambda x}\,dx = e^{-\lambda t}$, so get ⮍

So $P(X > t+s \mid X > t) = P(X > s)$, which is independent of t.

In fact, this meomry-less property characterises the exponential distribution. Indeed, assume X is memory-less. Then

$$P(X > t+s) = P(X > t+s \mid X > t) \cdot P(X > t) = P(X > s) P(X > t)$$

It follows that for integer $m \geqslant 1$,

$$P(X > m) = [P(X > 1)]^m$$

Therefore $\forall\, n, m \geqslant 1$

$$\left[P\left(X > \frac{m}{n}\right)\right]^n = P(X > m) = [P(X > 1)]^m$$

Therefore

$$P\left(X > \tfrac{m}{n}\right) = \left[P(X>1) \right]^{m/n}.$$

Assume $P(X>1) \in (0,1)$ so $P(X>1) = e^{-\lambda}$ for some $\lambda \in (0,\infty)$.

Hence $P(X>x) = e^{-\lambda x}$ for all $x \in \mathbb{Q}^+$.

Since both functions are non-increasing, deduce

$$P(X>x) = e^{-\lambda x} \text{ for all } x \in \mathbb{R}^+.$$

Hence $X \sim \mathrm{Exp}(\lambda)$.

2) <u>Exponential as limit of geometrics</u>

Let $T \sim \mathrm{Exp}(\lambda)$. For $n \geq 1$, define

$$T_n = \lfloor nT \rfloor.$$

Note that $P(T_n \geq k) = P(\lfloor nT \rfloor \geq k) = P\left(T \geq \frac{k}{n}\right) = e^{-\lambda k/n}$.

$\Rightarrow T_n \sim \text{Geometric}\,(\underbrace{1 - e^{-\lambda/n}}_{p_n})$

Clearly $\frac{T_n}{n} \to T$ as $n \to \infty$ by construction. link with Bin $\searrow$

So T is limit of Geometric $(1 - e^{-\lambda/n})$

$$\downarrow$$

$\approx \lambda/n$ for n large

L18.1

Expectation and variance

Def: For X non-negative rv we define

$$E(X) = \int_{\mathbb{R}} x f_X(x)\, dx = \int_0^\infty x f_X(x)\, dx$$

where f_X is the pdf of X.

In general, provided

$$E(X_+) = \int_0^\infty x f_X(x)\, dx < \infty$$

or $$E(X_-) = \int_{-\infty}^{0} x f_X(x)\, dx < \infty$$

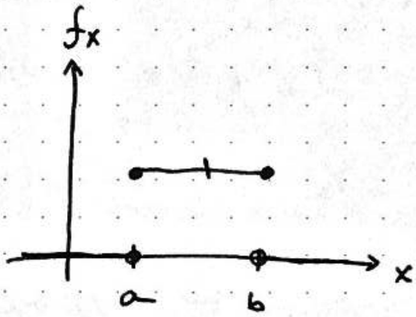
we define

$$E(X) = E(X_+) - E(X_-) = \int_{\mathbb{R}} x f(x)\, dx$$

EX: For $X \sim U[a,b]$ we have

$$E(X) = \int_{-\infty}^{\infty} x f_X(x)\, dx = \int_a^b x \cdot \frac{1}{b-a}\, dx$$

$$= \frac{1}{b-a} \cdot \frac{1}{2}(b^2 - a^2) = \frac{1}{2}(a+b)$$

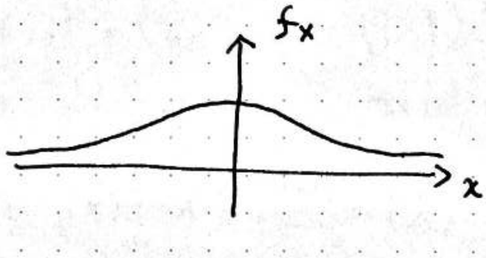


EX: $X \sim$ Cauchy

$$E(X_+) = \int_0^\infty x \cdot f_X(x)\, dx = \int_0^\infty \frac{x}{\pi(1+x^2)}\, dx = +\infty$$

$$E(X_-) = \int_{\infty}^{0} x \cdot f_X(x)\, dx = \int_0^\infty y \cdot f_X(y)\, dy = +\infty$$

Since both diverge, $E(X)$ undefined



EX: $X \sim N(0,1)$ $\quad f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$

Note that

$$E(X_+) = \int_0^\infty x \cdot \frac{1}{\sqrt{2\pi}} \cdot e^{-x^2/2}\, dx < \infty$$

and similarly for $E(X_-)$. Moreover $E(X_+) = E(X_-)$.

So $E(X) = E(X_+) - E(X_-) = 0$.

The expectation ~~st~~ satisfies all the properties seen in the discrete case.

In particular, it is <u>linear</u>.

Moreover, for any non-negative function $g : \mathbb{R} \to [0,\infty)$ we have

$$E(g(X)) = \int_{\mathbb{R}} g(x) \cdot f_X(x)\, dx$$ NO PROOF PROVIDED

In general,

$$E(g(X)) = \int_{\mathbb{R}} g(x) \cdot f_X(x)\, dx$$

for all g such that $E(|g(X)|) < \infty$.

Finally, if X is a non-negative random variable (its pdf f_X is supported in $[0,\infty)$) then

$$E(X) = \int_0^\infty P(X \geq x)\, dx = \int_0^\infty (1 - F_X(x))\, dx$$

where F_X is the distribution function of X.

<u>EX</u>: $X \sim \text{exponential}(\lambda) \Rightarrow f_X(x) = \lambda e^{-\lambda x} \mathbb{1}_{[0,\infty)}(x)$

$$E(X) = \int_0^\infty (1 - F_X(x))\, dx = \int_0^\infty (1 - \underbrace{F_X(x)}_{\int_0^x \lambda e^{-\lambda y} dy = 1 - e^{-\lambda x}})\, dx = \int_0^\infty e^{-\lambda x} dx = \frac{1}{\lambda}$$

<u>Variance</u>

Recall that X is said to be <u>integrable</u> if

$$E(|X|) = E(X_+) + E(X_-) < \infty$$

<u>Def</u>: For X integrable, define

$$\mathrm{Var}(X) = E\left[(X - E(X))^2\right] = E(X^2) - [E(X)]^2$$

$$= \int_{-\infty}^{\infty} x^2 f_X(x)\, dx - \left(\int_{-\infty}^{\infty} x\, f_X(x)\, dx \right)^2$$

18.3

EX: $X \sim U[a,b]$

$$E(X^2) = \int_a^b x^2 \cdot \frac{1}{b-a}\,dx = \frac{b^3-a^3}{3(b-a)} = \frac{b^2+ab+a^2}{3}$$

$$\therefore \operatorname{Var}(X) = E(X^2) - E(X)^2 = \frac{b^2+ab+a^2}{3} - \frac{b^2+2ab+a^2}{4} = \frac{(b-a)^2}{12}$$

EX: $X \sim N(0,1)$, $E(X) = 0$

$$E(X^2) = \int_{-\infty}^{\infty} x^2 \cdot e^{-x^2/2} \cdot \frac{dx}{\sqrt{2\pi}} = \underbrace{-x\frac{e^{-x^2/2}}{\sqrt{2\pi}}\Big|_{-\infty}^{\infty}}_{0} + \underbrace{\int_{-\infty}^{\infty} \frac{e^{-x^2/2}}{\sqrt{2\pi}}\,dx}_{1} = 1$$

EX: If $X \sim \exp(\lambda)$ then

$$\operatorname{Var}(X) = \frac{1}{\lambda^2}$$

Transformation of one-dimensional random variables

Let X be a r.v. with pdf f_X taking values in some $S \subseteq \mathbb{R}$.
Let $g: S \to \mathbb{R}$ be a function, and $Y = g(X)$.
Q: What is the pdf of Y?

Theorem: Let X be a random variable taking values in an open interval $S \subseteq \mathbb{R}$ with pdf f_X, with f_X piece-wise continuous.
Let $g: S \to \mathbb{R}$ be such that g is diff'ble on S with continuous derivative ~~&~~ g' such that g' never vanishes on S.
If $Y = g(X)$, then Y has piece-wise continuous pdf given by

$$f_Y(y) = f_X(g^{-1}(y)) \cdot \left| \frac{d}{dy} g^{-1}(y) \right|$$

where g^{-1} denotes the inverse function of g.

Proof Let F_X, F_Y denote the distribution functions of X and Y.
Then for all $y \in \mathbb{R}$, we have

$$F_Y(y) = P(Y \leq y) = P(g(X) \leq y)$$

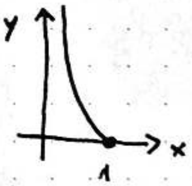
Assume first that g is strictly increasing on S. Then
$$F_Y(y) = P(X \leq g^{-1}(y)) = F_X(g^{-1}(y))$$
By the chain rule, F_Y has piece-wise continuous derivative given by

$$f_Y(y) = f_X(g^{-1}(y)) \cdot \underbrace{\frac{d}{dy} g^{-1}(y)}_{>0}$$

Note that Y takes values in $g(S)$.
If, on the other hand, g is decreasing on S, we have
$$F_Y(y) = P(Y \leq y) = P(g(X) \leq y) = P(X \geq g^{-1}(y)) = 1 - F_X(g^{-1}(y))$$
As before, differentiate to obtain

$$f_Y(y) = -f_X(g^{-1}(y)) \cdot \underbrace{\frac{d}{dy} g^{-1}(y)}_{<0} = \left| f_X(g^{-1}(y)) \frac{d}{dy} g^{-1}(y) \right|$$

EX: Let X be uniform $[0,1]$
Assume that X takes values in $(0,1]$ since $P(X=0)=0$
Take $g(x) = -\log x$, and set $Y = g(X) = -\log X$
Then $g^{-1}(y) = e^{-y}$ for $y \geq 0$
Recall $f_X(x) = \mathbb{1}_{[0,1]}(x)$.
Therefore Y has pdf



$$f_Y(y) = f_X(g^{-1}(y)) \left| \underbrace{\frac{d}{dy} g^{-1}(y)}_{-e^{-y}} \right| = e^{-y} \mathbb{1}_{[0,\infty)}(y)$$

Therefore $Y \sim$ exponential (1).

L18.5

EX: Let $X \sim N(\mu, \sigma^2)$

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

$$g^{-1}(y) = \sigma y + \mu$$

Take $g: \mathbb{R} \to \mathbb{R}$ given by $g(x) = \frac{x-\mu}{\sigma}$.

Let $Y = g(X)$. Then

$$f_Y(y) = f_X(g^{-1}(y)) \cdot \left| \frac{d}{dy} g^{-1}(y) \right| = f_X(\sigma y + \mu) \cdot \sigma$$

$$= \frac{1}{\sqrt{2\pi \cancel{\sigma^2}}} e^{-\frac{(\cancel{\sigma} y + \cancel{\mu} - \cancel{\mu})^2}{2 \cancel{\sigma^2}}} \cdot \cancel{\sigma} = \frac{1}{\sqrt{2\pi}} e^{-y^2/2}$$

So if $X \sim N(\mu, \sigma^2)$ then $Y \sim N(0, 1)$.

Similarly from $N(0,1)$ can do $Y \sim N(\mu, \sigma^2)$ by $Y = \sigma X + Y$.

Note that this can be used to see that if $X \sim N(\mu, \sigma^2)$ then

$\frac{X-\mu}{\sigma} \sim N(0,1)$ so $E\left(\frac{X-\mu}{\sigma}\right) = 0$, $Var\left(\frac{X-\mu}{\sigma}\right) = 1$

$$\frac{E(X)}{\sigma} - \frac{\mu}{\sigma} \qquad \frac{1}{\sigma^2} Var(X)$$

$$\Rightarrow Var(X) = \sigma^2$$

$$\Rightarrow E(X) = \mu$$

Remark: Could have obtained the same conclusion by observing that

$$F_Y(y) = P(Y \leq y) = P\left(\frac{X-\mu}{\sigma} \leq y\right) = P(X \leq \sigma y + \mu)$$

$$= \int_{-\infty}^{\sigma y + \mu} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx$$

Differentiate to get

$$f_Y(y) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-y^2/2} \cdot \sigma = \frac{1}{\sqrt{2\pi}} e^{-y^2/2} \text{ as before}$$

L18.6

Multi-variate distributions

<u>Def</u>: The joint distn function of two random variables X, Y is defined as $F_{X,Y}(x,y) = P(X \leq x, Y \leq y)$ for all $x, y \in \mathbb{R}$

Note that $F : \mathbb{R}^2 \to [0,1]$, it is non-decreasing in each coordinate, and
$$\lim_{y \to \infty} F_{X,Y}(x,y) = F_X(x)$$
$$\lim_{x \to \infty} F_{X,Y}(x,y) = F_Y(y).$$

<u>Def</u>: X, Y have joint probability density function $f_{X,Y}$ if their joint distn function is of the form
$$F_{X,Y}(x,y) = \int_{-\infty}^{x} du \int_{-\infty}^{y} dv \, f_{X,Y}(u,v)$$

In this case, we have
$$P((X,Y) \in A) = \iint_A f_{X,Y}(x,y) \, dA$$

True for all A Borel sets in $\mathbb{R}^2$, in particular open & closed sets.

Joint distribution of X, Y

$$F_{X,Y}(x,y) = P(X \leq x, Y \leq y) \quad \text{for } x, y \in \mathbb{R}.$$

Then

$$\lim_{y\to\infty} F(x,y) = F_X(x)$$

$$\lim_{x\to\infty} F(x,y) = F_Y(y).$$

X, Y have joint pdf $f_{X,Y}$ if their joint distribution is

$$F_{X,Y}(x,y) = \int_{-\infty}^{x} du \int_{-\infty}^{y} dv \; f_{X,Y}(u,v)$$

In this case

$$P((X,Y) \in A) = \iint_A f_{X,Y}(u,v)\,du\,dv$$

for all Borel sets A, and

$$f_{X,Y}(x,y) = \frac{\partial^2}{\partial x \partial y} F_{X,Y}(x,y).$$

Note that $f_{X,Y}(x,y) \geq 0$ and

$$\int_{-\infty}^{\infty} du \int_{-\infty}^{\infty} dv \; f_{X,Y}(u,v) = 1.$$

Remark The marginal distributions f_X, f_Y can be recovered from the joint pdf $f_{X,Y}$ via

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dy, \qquad f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dx.$$

For $g : \mathbb{R}^2 \to \mathbb{R}$ non-negative, we have

$$E(g(X,Y)) = \iint_{\mathbb{R}^2} g(x,y)\, f_{X,Y}(x,y)\,dx\,dy.$$

This extends to all functions g such that $E(|g(X,Y)|) < \infty$.

In particular, take $g(X,Y) = XY$ to get

$$\begin{aligned}\mathrm{Cov}(X,Y) &= E\big[(X-E(X))(Y-E(Y))\big] \\ &= E(XY) - E(X)E(Y) \\ &= \iint_{\mathbb{R}^2} xy\, f_{X,Y}(x,y)\,dx\,dy - \left(\int_{\mathbb{R}} x f_X(x)\,dx\right)\left(\int_{\mathbb{R}} y f_Y(y)\,dy\right)\end{aligned}$$

All of the above generalises to arbitrarily many rvs.
For example, the joint distribution of $X_1, \dots, X_n$ is

$$F_{X_1,\dots,X_n}(x_1,\dots,x_n) = P(X_1 \le x_1, \dots, X_n \le x_n)$$

$X_1, \dots, X_n$ have joint pdf $f_{X_1,\dots,X_n}$ if

$$F_{X_1,\dots,X_n}(x_1,\dots,x_n) = \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_n} f_{X_1,\dots,X_n}(u_1,\dots,u_n)\,du_1 \cdots du_n .$$

To recover the marginal pdf of X_k we use

$$f_k(x) = \int_{\mathbb{R}^{n-1}} f_{X_1,\dots,X_n}(x_1,\dots,x_{k-1},x,x_{k+1},\dots,x_n)\,dx_1 \cdots dx_{k-1}\,dx_{k+1} \cdots dx_n$$

<u>Independence</u>

The random variables $X_1, \dots, X_n$ are said to be independent if $P(X_1 \le x_1, \dots, X_n \le x_n) = \prod_{k=1}^{n} P(X_k \le x_k)$ for all $x_1, \dots, x_n$.

<u>Theorem</u>: 1) Let $X_1, \dots, X_n$ be independent rvs with pdfs $f_{X_1}, \dots, f_{X_n}$ respectively. Then their joint pdf is given by

$$f_{X_1,\dots,X_n}(x_1,\dots,x_n) = \prod_{k=1}^{n} f_{X_k}(x_k).$$

2) If $X_1, \dots, X_n$ have joint pdf of the form above

$$f_{X_1,\dots,X_n}(x_1,\dots,x_n) = \prod_{k=1}^{n} f_k(x_k)$$

for some functions f_k, then $X_1, \dots, X_n$ are independent and their pdfs are proportional to the f_k.

Remark: With this theorem in hand, it suffices to show that the joint pdf factorises into f_k to show independence. But the f_k need not be proper pdfs.

Proof of Theorem: 1) If $X_1, \dots, X_n$ are independent, then if we let $A = (-\infty, x_1] \times \cdots \times (-\infty, x_n]$ where the x_i are arbitrary, then $P((X_1, \dots, X_n) \in A) = \int \cdots \int_A f_{X_1, \dots, X_n}(x_1, \dots, x_n)\, dx_1 \cdots dx_n$.

But we know that, by independence, this is equal to

$$\prod_{k=1}^{n} P(X_k \leq x_k) = \prod_{k=1}^{n} \int_{-\infty}^{x_k} f_{X_k}(u_k)\, du_k.$$

$$\overset{\sim}{\equiv} \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_n} f_{X_1}(u_1) \cdots f_{X_n}(u_n)\, du_1 \cdots du_n.$$

From this, it is clear that the joint pdf factorises.

2) Now assume the $X_1, \dots, X_n$ have joint pdf

$$f_{X_1, \dots, X_n}(x_1, \dots, x_n) = \prod_{k=1}^{n} f_k(x_k)$$

Note that

$$\int \cdots \int_{\mathbb{R}^n} f_{X_1, \dots, X_n}(x_1, \dots, x_n)\, dx_1 \cdots dx_n = 1.$$

so we may rescale the f_k so they integrate to 1.

Take $A = A_1 \times \cdots \times A_n$ Borel set. Then

$$P((X_1, \dots, X_n) \in A) = \int \cdots \int_A \left[\prod_{k=1}^{n} f_k(x_k) \right] dx_1\, dx_2 \cdots dx_n$$

$$= \prod_{k=1}^{n} \left[\int_{A_k} f_k(x_k)\, dx_k \right].$$ So pick $1 \leq k \leq n$ and set

$$A = \mathbb{R} \times \cdots \times \mathbb{R} \times A_k \times \mathbb{R} \times \cdots \times \mathbb{R}.$$ Then

$P((X_1, \dots, X_n) \in A) = P(X_k \in A_k) = \int_{A_k} f_k(x_k)\, dx_k$ by above, since we rescaled the f_j. ~~So done~~ $\boxtimes$

L19.4

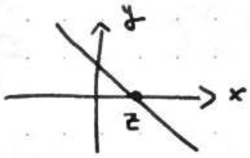
So X_k has pdf f_k. To see independence, note

$$P(X_1 \leq x_1, \dots X_n \leq x_n) = \prod_{k=1}^{n} \int_{-\infty}^{x_k} f_k(u_k)\, du_k = \prod_{k=1}^{n} P(X_k \leq u_k).$$

by the above

□

<u>Example</u> Let X, Y be independent rvs with pdfs f_X and f_Y.
By Theorem, they have joint pdf $f_{X,Y}(x,y) = f_X(x) f_Y(y)$.
Let $Z = X + Y$. Then for any $z \in \mathbb{R}$,



$$P(Z \leq z) = P(X + Y \leq z)$$

$$= \int_{-\infty}^{\infty} dx \int_{-\infty}^{z-x} dy\; f_{X,Y}(x,y)$$

$$= \int_{-\infty}^{\infty} dx \int_{-\infty}^{z} du\; f_Y(u-x) f_X(x)$$

$u = x + y$

$$= \int_{-\infty}^{z} du \left(\int_{-\infty}^{\infty} dx\; f_Y(u-x) f_X(x) \right).$$

Differentiate wrt z to get

$$f_{X+Y}(z) = \int_{-\infty}^{\infty} f_Y(z-x) f_X(x)\, dx = (f_X * f_Y)(z).$$

convolution

$\therefore$ The pdf of the sum of <u>independent</u> rvs is given by the convolution of the marginal pdfs.

If $X_1, \dots, X_n$ are independent rvs, and $g_1, \dots, g_n$ are functions on $\mathbb{R}$, then we have

$$E(g_1(X_1) \cdots g_n(X_n)) = \prod_{k=1}^{n} E(g_k(X_k))$$

and $g_1(X_1), \dots, g_n(X_n)$ are independent rvs.

In particular, if X, Y are independent, then $\mathrm{Cov}(X,Y) = 0$.

Transformation of 2D random variables

Theorem: Let X, Y be random variables with joint pdf $f_{X,Y}$ supported on a domain $D \subseteq \mathbb{R}^2$.

Let $g: D \to \mathbb{R}^2$ defined by $g(x,y) = (u(x,y), v(x,y))$ such that g is a bijection from D to $g(D)$ with continuous derivative ~~det~~ such that $\det g'(x,y) \neq 0$.

Write g^{-1} for the inverse of g, given by $g^{-1}(u,v) = (x(u,v), y(u,v))$.

Then the random variables defined by

$$(U,V) = g(X,Y)$$

have joint pdf

$$f_{U,V}(u,v) = f_{X,Y}(g^{-1}(u,v)) \cdot |J|$$

inverse of g

where

$$J = \det \begin{pmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \end{pmatrix}.$$

want $f_{u,v}\,du\,dv = f_{x,y}\,dx\,dy$ and $du\,dv = |J^{-1}|\,dx\,dy = \left|\frac{\partial(u,v)}{\partial(x,y)}\right| dx\,dy$

Example: Let X, Y be independent $\exp(\lambda)$

Then $f_{X,Y}(x,y) = \lambda^2 e^{-\lambda(x+y)} \mathbb{1}_{[0,\infty)\times[0,\infty)}$.

Define $U = X+Y$, $V = X-Y$.

Then $g(x,y) = (x+y, x-y) = (u,v)$

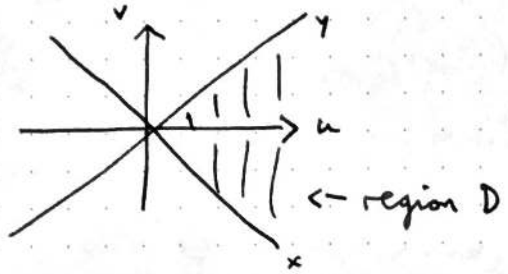
so $g^{-1}(u,v) = \left(\frac{u+v}{2}, \frac{u-v}{2}\right)$.

$$J = \det \begin{pmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & -\frac{1}{2} \end{pmatrix} = -\frac{1}{2}$$

and $f_{U,V}(u,v) = f_{X,Y}\left(\frac{u+v}{2}, \frac{u-v}{2}\right) \cdot \frac{1}{2} = \frac{\lambda^2}{2} e^{-\lambda\left(\frac{u+v}{2}\right) - \lambda\left(\frac{u-v}{2}\right)} \mathbb{1}_{U \times V}$

L19.6

$$f_{U,V}(u,v) = \frac{\lambda^2}{2\cancel{4}} e^{-\lambda u} \mathbb{1}_D(u,v)$$



$$\mathbb{1}_D(u,v) = \mathbb{1}_{[0,\infty)}(u)\ \mathbb{1}_{[-u,u]}(v)$$

L20.1

$X, Y \sim \exp(\lambda)$ indep.

$$f_{X,Y}(x,y) = \lambda^2 e^{-\lambda(x^{\cancel{2}}+y^{\cancel{2}})} \mathbb{1}_{[0,\infty)}(x)\, \mathbb{1}_{[0,\infty)}(y)$$

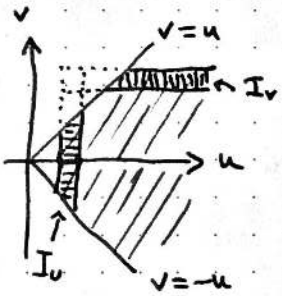
supported on $D = [0,\infty) \times [0,\infty)$

$$g: \begin{cases} U = X+Y \\ V = X-Y \end{cases} \qquad g^{-1}: \begin{cases} X = \frac{1}{2}(U+V) \\ Y = \frac{1}{2}(U-V) \end{cases}$$

$$J = \det\begin{pmatrix} 1/2 & 1/2 \\ 1/2 & -1/2 \end{pmatrix} = -\frac{1}{2}$$

$$\therefore f_{U,V}(u,v) = f_{X,Y}(g^{-1}(u,v)) \cdot |J|$$

$$= \frac{1}{2}\lambda^2 \underbrace{e^{-\lambda[(\frac{u+v}{2})+(\frac{u-v}{2})]}}_{e^{-\lambda u}} \underbrace{\mathbb{1}_{[0,\infty)}\left(\frac{u+v}{2}\right)}_{u+v \geq 0} \underbrace{\mathbb{1}_{[0,\infty)}\left(\frac{u-v}{2}\right)}_{u-v \geq 0}$$



$$= \frac{1}{2}\lambda^2 e^{-\lambda u} \mathbb{1}_{[0,\infty)}(u)\, \mathbb{1}_{[-u,u]}(v)$$

Remark: U, V are not independent, since the joint pdf is not supported on a rectangular region

To make this rigorous, look for intervals I_u, I_v such that $P(U \in I_u), P(V \in I_v) > 0$ and $P((U,V) \in I_U \times I_V) = 0$.

$f_{U,V}$ supported on non-rectangle $\Rightarrow$ U, V not indep

converse not necessarily true

EX (Polar coordinates)

Let X, Y be independent $N(0,1)$.

Set $g^{-1}: \begin{cases} X = R\cos\Theta \\ Y = R\sin\Theta \end{cases}$ on $(R,\Theta) \in [0,\infty) \times [0, 2\pi)$.

$$J = \det\begin{pmatrix} \cos\Theta & -R\sin\Theta \\ \sin\Theta & R\cos\Theta \end{pmatrix} = R$$

$$f_{R,\Theta}(r,\theta) = \cancel{R}_{r} \cdot f_{X,Y}(r\cos\theta, r\sin\theta) = \cancel{R}_{r} \cdot \frac{1}{2\pi} e^{-\frac{r^2}{2}} \mathbb{1}_{[0,\infty)}(r)\, \mathbb{1}_{[0,2\pi)}(\theta)$$

Note this means R and Θ are independent, Θ is uniform ☺

L20.2

Theorem Let $\underline{X} = (X_1, X_2, \dots, X_n)$ be a random variable in $\mathbb{R}^n$ with pdf $f_{\underline{X}}$ supported on a domain $D \subseteq \mathbb{R}^n$.

Let $g : D \to \mathbb{R}^n$ be a bijection between D and $g(D)$ and suppose g has continuous derivative s.t. $\det g'(\underline{x}) \neq 0 \quad \forall \underline{x} \in D$.

Set $\underline{Y} = g(\underline{X}) = (Y_1, Y_2, \dots, Y_n)$.

Then $\underline{Y}$ has pdf

$$f_{\underline{Y}}(\underline{y}) = f_{\underline{X}}(g^{-1}(\underline{y})) \cdot |J|$$

where J is the Jacobian of the inverse transformation g^{-1}.

Moment Generating Functions

Def: Let X be a random variable in $\mathbb{R}$. The moment generating function M_X of X is defined by

$$M_X(\theta) = E(e^{\theta X}) = \int_{-\infty}^{\infty} e^{\theta x} f_X(x)\, dx.$$

Note that the integral is well-defined, even though it may be $+\infty$.

EX $X \sim \exp(\lambda)$

$$M_X(\theta) = \int_0^{\infty} e^{\theta x} \cdot \lambda e^{-\lambda x} dx = \lambda \int_0^{\infty} e^{-(\lambda-\theta)x} dx$$

$$= \begin{cases} +\infty & \text{if } \theta \geq \lambda, \\ \dfrac{\lambda}{\lambda - \theta} & \text{if } \theta < \lambda. \end{cases}$$

EX $X \sim$ Cauchy

$$M_X(\theta) = \int_{-\infty}^{\infty} \frac{e^{\theta x}}{\pi(1+x^2)} dx = \begin{cases} +\infty & \text{if } \theta \neq 0 \\ 1 & \text{if } \theta = 0 \end{cases}$$

Note that $M_X(0)$ is always 1.

L 20.3

<u>EX</u>: $X \sim N(0,1)$

$$M_X(\theta) = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{\theta x - \frac{x^2}{2}} dx = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{(x^2 - 2\theta x)}{2}} dx.$$

Note that $x^2 - 2\theta x = (x-\theta)^2 - \theta^2$ so get

$$\int_{-\infty}^{\infty} e^{+\theta^2/2} \cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-\theta)^2}{2}} dx = e^{+\theta^2/2}$$

In general, if $X \sim N(\mu, \sigma^2)$ then we can write

$X \sim \mu + \sigma Z$, so

$$M_X(\theta) = E(e^{\theta X}) = E(e^{\theta\mu + \theta\sigma Z})$$
$$= e^{\theta\mu} M_Z(\theta\sigma) = e^{\theta\mu + \frac{\theta^2\sigma^2}{2}}$$

If X, Y are independent random variables,

$$M_{X+Y}(\theta) = M_X(\theta) M_Y(\theta)$$

Indeed,

$$M_{X+Y}(\theta) = E(\theta^{X+Y}) = E(\theta^X \theta^Y) \overset{\text{indep}}{=} E(\theta^X) E(\theta^Y)$$

In general, if $X_1, \dots, X_n$ are indep.

$$M_{X_1 + \dots + X_n}(\theta) = \prod_{k=1}^{n} M_{X_k}(\theta)$$

If they are identically distributed, then

$$M_{X_1 + \dots + X_n}(\theta) = [M_{X_1}(\theta)]^n.$$

<u>Theorem</u>: let X, Y be two random variables with mgf $M_X = M_Y$. Suppose $M_X(\theta)$ is bounded on some open neighbourhood of $\theta = 0$. Then X and Y have the same distibution.

Note that the assumption $M_X(\theta) < \infty$ around θ is necessary, since if $X \sim$ Cauchy, then X and $2X$ have different distribution but the same mgf.

EX: $X \sim N(\mu_X, \sigma_X^2)$ $Y \sim N(\mu_Y, \sigma_Y^2)$ be indep.

Then $M_{X+Y}(\theta) = M_X(\theta) M_Y(\theta)$

$$= e^{\theta\mu_X + \frac{\theta^2}{2}\sigma_X^2} \cdot e^{\theta\mu_Y + \frac{\theta^2\sigma_Y^2}{2}}$$

$$= e^{\theta(\mu_X+\mu_Y) + (\sigma_X^2+\sigma_Y^2)\theta^2/2}$$

So $X+Y \sim N(\mu_X+\mu_Y, \sigma_X^2+\sigma_Y^2)$.

EX: Let $X_1, \dots, X_n$ be independent $\exp(\lambda)$. Then

$M_{X_1}(\theta) = \lambda/_{\lambda-\theta}$ for $\theta < \lambda$ so

$M_{X_1+\dots+X_n}(\theta) = (\lambda/_{\lambda-\theta})^n$ for $\theta < \lambda$ and ∞ otherwise

By direct computation, check that this is the moment generating function of a Gamma (n, λ).

Assume that $M_X(\theta) < \infty$ in a neighbourhood of 0.

Note that $M_X(0) = 1$. Since

$$e^{\theta X} = 1 + \theta X + \tfrac{1}{2}\theta^2 X^2 + \tfrac{1}{3!}\theta^3 X^3 + \cdots$$

taking expectations gives

$$M_X(\theta) = 1 + \theta E(X) + \tfrac{1}{2}\theta^2 E(X^2) + \cdots$$

By differentiating above wrt θ recover moments of X.

$M_X^{(k)}(0) = E(X^k)$.

EX: $X \sim \exp(\lambda)$

$$M_X(\theta) = \frac{\lambda}{\lambda-\theta} \text{ for } \theta < \lambda$$

$$M_X'(\theta) = \frac{\lambda}{(\lambda-\theta)^2} \Rightarrow E(X) = M_X'(0) = \frac{1}{\lambda}$$

$$M_X''(\theta) = \frac{2\lambda}{(\lambda-\theta)^3} \Rightarrow E(X^2) = M_X''(0) = \frac{2}{\lambda^2}$$

Continuity Theorem for mgfs

Def: Let $(X_n)_{n\geq 1}$, ~~be~~, X be random variables with distribution functions $(F_{X_n})_{n\geq 1}$, F_X. We say that $X_n \longrightarrow X$ in distribution if

$F_{X_n}(x) \longrightarrow F_X(x)$ for all x at which F_X is continuous.

L20.5

<u>Theorem</u> Let $(X_n)_{n\geq 1}$, X be rvs with mgfs $(M_{X_n})_{n\geq 1}$, M_X respectively. If

$$M_{X_n}(\theta) \longrightarrow M_X(\theta)$$

for all $\theta \in \mathbb{R}$ and $M_X(\theta) < \infty$ in a neighbourhood of 0, then

$$X_n \longrightarrow X$$

in distribution.

Gaussian Random Variables

<u>Def</u>: A random variable X in $\mathbb{R}$ is said to be Gaussian if it can be written as $X = \mu + \sigma Z$

for some $\mu \in \mathbb{R}$, $\sigma \in [0, \infty)$ where $Z \sim N(0,1)$.

Note that we allow $\sigma = 0$.

If $\sigma > 0$ then X has pdf

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

<u>Def</u>: A random variable $\underline{X} = \begin{pmatrix} X_1 \\ \vdots \\ X_n \end{pmatrix}$ in $\mathbb{R}^n$ is said to be Gaussian if for all $\underline{u} \in \mathbb{R}^n$ the random variable

$$\underline{u}^T \underline{X} = \sum_{i=1}^{n} u_i X_i \quad \text{is Gaussian in } \mathbb{R}.$$

Note that $\underline{X}$ Gaussian in $\mathbb{R}^n$ implies each X_i Gaussian in $\mathbb{R}$

Moreover, for any matrix $A \in \mathbb{R}^{m \times n}$ and vector $\underline{b} \in \mathbb{R}^m$, the random variable

$$A\underline{X} + \underline{b}$$

is Gaussian in $\mathbb{R}^m$.

Indeed, for all $\underline{u} \in \mathbb{R}^m$, have that

$$\underline{u}^T (A\underline{X} + \underline{b}) = \underbrace{(\underline{u}^T A)}_{\text{vector in } \mathbb{R}^n} \underline{X} + \underbrace{\underline{u}^T \underline{b}}_{\text{constant}}$$

is a Gaussian in $\mathbb{R}^*$.

<u>Def</u>: For $\underline{X}$ a Gaussian random variable in $\mathbb{R}^n$ define

$$\underline{\mu} = E(\underline{X}) = \begin{pmatrix} E(X_1) \\ \vdots \\ E(X_n) \end{pmatrix}$$

and

$$\mathrm{Var}(\underline{X}) = E\Big[\underbrace{(\underline{X} - \underline{\mu})(\underline{X} - \underline{\mu})^T}_{\substack{n \times n \text{ matrix with} \\ ij \text{ entry } (X_i - \mu_i)(X_j - \mu_j)}}\Big]$$

So V is an $n \times n$ matrix with

$$V_{ij} = \mathrm{Cov}(X_i, X_j).$$

In particular

$$V_{ii} = \mathrm{Var}(X_i).$$

Refer to $\underline{\mu}$ as the <u>mean</u> of $\underline{X}$, and to V as the <u>covariance matrix</u>. Note that V is symmetric, and non-negative definite, since $\forall\, \underline{u} \in \mathbb{R}^n$

$$\begin{aligned} \underline{u}^T V \underline{u} &= \underline{u}^T E\left((\underline{X}-\underline{\mu})(\underline{X}-\underline{\mu})^T\right) \underline{u} \\ &= E\Big((\underline{u}^T\underline{X} - \underline{u}^T\underline{\mu})\underbrace{(\underline{X}-\underline{\mu})^T\underline{u}}_{(\underline{u}^T\underline{X}-\underline{u}^T\underline{\mu})^T \leftarrow \text{1d thingy}}\Big) \quad = \mathrm{Var}(\underline{u}^T\underline{X}) \end{aligned}$$

which is necessarily non-negative.

diagonalise

If $\underline{X}$ is Gaussian in $\mathbb{R}^n$ with mean $\underline{\mu}$ and cov. matrix V, then for any $\underline{u} \in \mathbb{R}^n$ we have that $\underline{u}^T\underline{X}$ is Gaussian in $\mathbb{R}$ with

$$E(\underline{u}^T\underline{X}) = \underline{u}^T E(\underline{X}) = \underline{u}^T\underline{\mu}$$

$$\begin{aligned} \mathrm{Var}(\underline{u}^T\underline{X}) &= E\left[(\underline{u}^T\underline{X} - \underline{u}^T\underline{\mu})(\underline{u}^T\underline{X} - \underline{u}^T\underline{\mu})^T\right] \\ &= \underline{u}^T V \underline{u} \quad \text{from above} \end{aligned}$$

We showed that

$$\underline{u}^T\underline{X} \sim N(\underline{u}^T\mu, \underline{u}^T V \underline{u}).$$

<u>Moment generating functions in $\mathbb{R}^n$</u>

Recall that if $Z \sim N(0,1)$ then $M_Z(\theta) = e^{\theta^2/2}$, and if $X \sim N(\mu, \sigma^2)$ then $M_X(\theta) = e^{\mu\theta + \sigma^2\theta^2/2}$.

<u>Def</u>: The mgf of a random variable $\underline{X}$ in $\mathbb{R}^n$ is given by the function $M_{\underline{X}} : \mathbb{R}^n \longrightarrow \mathbb{R} \cup \{\pm\infty\}$ given by

$$M_{\underline{X}}(\underline{\theta}) = E(e^{\underline{\theta}^T\underline{X}}) = E\left(\exp\left\{\sum_{i=1}^n \theta_i X_i\right\}\right).$$

Recall that $M_{\underline{X}}$ determines the distribution of $\underline{X}$ provided

$M_{\underline{X}}(\underline{\theta}) < \infty$ for all $\underline{\theta}$ in a neighbourhood of $\underline{0}$.

Let $\underline{X}$ be a Gaussian random variable in $\mathbb{R}^n$ with mean $\underline{\mu}$ and cov V.

Then for any $\underline{\theta} \in \mathbb{R}^n$,

$$\underline{\theta}^T \underline{X} \sim N(\underline{\theta}^T \underline{\mu}, \underline{\theta}^T V \underline{\theta})$$

so the mgf of $\underline{X}$ is

$$\begin{aligned} M_{\underline{X}}(\underline{\theta}) = E(e^{\underline{\theta}^T \underline{X}}) &= M_{\underline{\theta}^T \underline{X}}(1) \\ &= e^{E(\underline{\theta}^T \underline{X}) + \frac{1}{2} Var(\underline{\theta}^T \underline{X})} \\ &= \exp(\underline{\theta}^T \underline{\mu} + \tfrac{1}{2} \underline{\theta}^T V \underline{\theta}). \end{aligned}$$

Note that $M_{\underline{X}}$ only depends on $\underline{\mu}$ and V and it is finite.

It follows that $M_{\underline{X}}$ uniquely determines the distribution.

We conclude that the distribution of a Gaussian rv in $\mathbb{R}^n$ is uniquely determined by $\underline{\mu} = E(\underline{X})$ and $V = Var(\underline{X})$.

We write $\underline{X} \sim N(\underline{\mu}, V)$.

Construction

We've seen that if $\underline{X}$ is Gaussian in $\mathbb{R}^n$ then $E(\underline{X}) \in \mathbb{R}^n$ and $Var(\underline{X})$ is an $n \times n$ symmetric non-negative definite matrix.

But we'll now show that for any $\underline{\mu} \in \mathbb{R}^n$, V $n \times n$ symmetric non-negative definite matrix we can construct a Gaussian random variable $\underline{X} \in \mathbb{R}^n$ with

$E(\underline{X}) = \underline{\mu}$, $Var(\underline{X}) = V$.

To see this, take $Z_1, \dots, Z_n$ to be iid. $N(0,1)$ in $\mathbb{R}$, and set $\underline{Z} = \begin{pmatrix} Z_1 \\ \vdots \\ Z_n \end{pmatrix}$. Then

$$\begin{aligned} M_{\underline{Z}}(\underline{\theta}) = E(e^{\underline{\theta}^T \underline{Z}}) &= E(e^{\sum_{i=1}^n \theta_i Z_i}) \\ \xrightarrow{\text{indep.}} &= \prod_{i=1}^n E(e^{\theta_i Z_i}) = \prod_{i=1}^n e^{\theta_i^2/2} \end{aligned}$$

So get

$$M_{\underline{Z}}(\underline{\theta}) = e^{|\underline{\theta}|^2/2}.$$

$\underline{Z}$ is a Gaussian random variable in $\mathbb{R}^n$ with

$$E(\underline{Z}) = \underline{0} \quad \text{and} \quad Var(\underline{Z}) = I_n.$$

For any $\underline{\mu} \in \mathbb{R}^n$ and V non-neg. def. matrix, then there exists Σ $n \times n$ symmetric matrix such that $\overset{\text{sym.}}{\wedge}$

$$\Sigma\Sigma = V.$$

We call Σ the square root of V.

Define $\underline{X} = \underline{\mu} + \Sigma \underline{Z}$.

Then $E(\underline{X}) = E(\underline{\mu} + \Sigma\underline{Z}) = \underline{\mu} + \overset{\text{zero}}{\cancel{\Sigma E(\underline{Z})}}$,

and

$$\begin{aligned} Var(\underline{X}) &= E\left((\Sigma\underline{Z})(\Sigma\underline{Z})^T\right) \\ &= E(\Sigma \underline{Z}\underline{Z}^T \Sigma^T) \\ &= \Sigma E(\underline{Z}\underline{Z}^T)\Sigma^T \\ &= \Sigma I \Sigma^T \\ &= \Sigma\Sigma^T \\ &= V \end{aligned}$$

So we conclude that $\underline{X} = \underline{\mu} + \Sigma\underline{Z}$ is Gaussian in $\mathbb{R}^n$ with

$$E(\underline{X}) = \underline{\mu}, \quad Var(\underline{X}) = \Sigma^2 = V.$$

Probability density functions

Let $\underline{Z} = \begin{pmatrix} z_1 \\ \vdots \\ z_n \end{pmatrix}$ be Gaussian in $\mathbb{R}^n$ with $\underline{Z} \sim N(\underline{0}, I_n)$.

Then the probability density function of $\underline{Z}$ is given by

$$f_{\underline{Z}}(\underline{z}) = \prod_{i=1}^{n} f_{z_i}(z_i) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} e^{-z_i^2/2} = \frac{1}{(2\pi)^{n/2}} e^{-|\underline{z}|^2/2}$$

For any $\underline{X} \sim N(\underline{\mu}, V)$ we have $\underline{X} = \underline{\mu} + \sigma\underline{Z}$ where $\sigma\sigma = V$.

Assume that V is pos def.

Then set

$$g(\underline{Z}) = \underline{\mu} + \sigma\underline{Z}. \quad \text{Then } g \text{ is invertible with}$$

$$g^{-1}(\underline{X}) = \sigma^{-1}(\underline{X} - \underline{\mu})$$

This has Jacobian $J = \det(\sigma^{-1}) = (\det\sigma)^{-1} = (\det V)^{-1/2}$.

L21.5

So by the change of variables formula

$$f_{\underline{X}}(\underline{x}) = f_{\underline{Z}}(g^{-1}(\underline{x})) \cdot |J|$$

$$= \frac{1}{(2\pi)^{n/2}} e^{-|\sigma^{-1}(\underline{x}-\underline{\mu})|^2/2} \cdot (\det V)^{-1/2}$$

and since

$$|\sigma^{-1}(\underline{x}-\underline{\mu})|^2 = (\sigma^{-1}(\underline{x}-\underline{\mu}))^T(\sigma^{-1}(\underline{x}-\underline{\mu}))$$

$$= (\underline{x}-\underline{\mu})^T \sigma^{-1T}\sigma^{-1}(\underline{x}-\underline{\mu})$$

$$= (\underline{x}-\underline{\mu})^T V^{-1}(\underline{x}-\underline{\mu})$$

get

$$f_{\underline{X}}(\underline{x}) = \frac{1}{(2\pi)^{n/2}} e^{-\frac{(\underline{x}-\underline{\mu})^T V^{-1}(\underline{x}-\underline{\mu})}{2}} (\det V)^{-1/2}$$

L22.1

Gaussian rvs

$\underline{X}$ is Gaussian in $\mathbb{R}^n$ if $\forall\, \underline{u} \in \mathbb{R}^n$
$\underline{u}^T \underline{X}$ is Gaussian in $\mathbb{R}$

$$\underline{\mu} = \begin{pmatrix} E(X_1) \\ \vdots \\ E(X_n) \end{pmatrix} \qquad V = (V_{ij}) \text{ where } V_{ij} = \mathrm{Cov}(X_i, X_j)$$

If $\underline{X} \sim N(\underline{\mu}, V)$ then

$$M_{\underline{x}}(\underline{\theta}) = \exp\left\{ \underline{\theta}^T \underline{\mu} + \tfrac{1}{2} \underline{\theta}^T V \underline{\theta} \right\}.$$

Pdf: $\underline{Z} \sim N(\underline{0}, I_n)$

$$f_{\underline{z}}(\underline{z}) = \frac{1}{(2\pi)^{n/2}} e^{-|\underline{z}|^2/2}$$

Take $\underline{X} = \underline{\mu} + \sigma \underline{Z}$ for σ square root of V (pos def)

$$f_{\underline{x}}(\underline{x}) = \frac{1}{(2\pi)^{n/2} (\det V)^{1/2}} e^{-\frac{1}{2}(\underline{x}-\underline{\mu})^T V^{-1} (\underline{x}-\underline{\mu})}$$

If V is nonnegative definite, then in some Cartesian basis

$$V' = \begin{pmatrix} U & 0 \\ 0 & 0 \end{pmatrix} \quad \text{i.e.} \quad S V S^T = \begin{pmatrix} U & 0 \\ 0 & 0 \end{pmatrix}$$

for U positive definite. This is the covariance matrix of $S\underline{X}$, so it suffices to write down the pdf of $\underline{X}$ with V of the form $\begin{pmatrix} U & 0 \\ 0 & 0 \end{pmatrix}$.

For V $n \times n$ and U $m \times m$ with $m \leq n$ then $\underline{X}$ is of the form

$$\underline{X} = \begin{pmatrix} \underline{Y} \\ \underline{v} \end{pmatrix} \begin{matrix} \}\, m \\ \}\, n-m \end{matrix}$$

where $\underline{Y}$ is Gaussian with cov U and $\underline{v}$ is a constant vector.

The pdf of $\underline{Y}$ is simply

$$f_Y(\underline{y}) = \frac{1}{(2\pi)^{m/2}} e^{-\frac{1}{2}(\underline{y}-\underline{\lambda})^T U^{-1} (\underline{y}-\underline{\lambda})} (\det U)^{-\frac{1}{2}}.$$

Bivariate Gaussians ($\mathbb{R}^2$)

$\underline{X}$ is a bivariate Gaussian if it's a Gaussian in $\mathbb{R}^2$, $\underline{X} = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$

Let $E(X_1) = \mu_1$, $E(X_2) = \mu_2$

$Var(X_1) = \sigma_1^2$, $Var(X_2) = \sigma_2^2$

$\rho = Corr(X_1, X_2) = Cov(X_1, X_2) / \sqrt{Var(X_1) Var(X_2)}$

so long as $\sigma_1 \sigma_2 > 0$.

By Cauchy-Schwarz, $\rho \in [-1, 1]$.

Then $\underline{X}$ has mean and covariance matrix

$$\underline{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} \qquad V = \begin{pmatrix} \sigma_1^2 & \sigma_1\sigma_2\rho \\ \sigma_1\sigma_2\rho & \sigma_2^2 \end{pmatrix}.$$

Moreover, for any $\mu_1, \mu_2, \sigma_1, \sigma_2, \rho$ with appropriate values, can construct a bivariate Gaussian with covariance and mean as above.

Indeed, by previous work, just have to check that

$\begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$ is non-neg def for all σ_1, σ_2, ρ.

To see this, show that $\forall \underline{u} \in \mathbb{R}^2$ it holds that

$$\underline{u}^T V \underline{u} = (u_1, u_2) \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \end{pmatrix}$$

$$= (u_1^2\sigma_1^2 + u_2^2\sigma_2^2)(1+\rho) - \rho(u_1\sigma_1 - u_2\sigma_2)^2$$

which Checks Out somehow.

So $\mu_1, \mu_2, \sigma_1, \sigma_2, \rho$ parametrise all bivariate Gaussians.

Assume that $\sigma_1\sigma_2 > 0$ to exclude trivial cases. If so,

$\rho = 0$: $V = \begin{pmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{pmatrix}$ and $\underline{X}$ has pdf

$$f_{\underline{X}}(\underline{x}) = \frac{1}{2\pi\sigma_1\sigma_2} e^{-\frac{1}{2}(\underline{x}-\underline{\mu})^T V^{-1}(\underline{x}-\underline{\mu})}$$

$$= \frac{1}{2\pi\sigma_1\sigma_2} e^{-\left(\frac{(x_1-\mu_1)}{2\sigma_1^2} + \frac{(x_2-\mu_2)}{2\sigma_2^2}\right)}$$

which could have been derived from independent Gaussians.

L22.3

We conclude that X_1 and X_2 are independent Gaussians.

This shows that if X_1, X_2 are Gaussian

$$\text{Cov}(X_1, X_2) = 0 \iff X_1, X_2 \text{ indep.}$$

not true in general

If, however, $\rho \neq 0$, then take any $a \in \mathbb{R}$.

$$\text{Cov}(X_1, X_2 - aX_1) = \text{Cov}(X_1, X_2) - a\,\text{Var}(X_1)$$
$$= \rho\sigma_1\sigma_2 - a\sigma_1^2$$

We can make the covariance zero by taking $a = \frac{\rho\sigma_2}{\sigma_1}$.

Define $Y = X_2 - aX_1$ for this a. Then X_1 and Y are jointly Gaussian, since

$$\begin{pmatrix} X_1 \\ Y \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ -a & +1 \end{pmatrix} \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$$

so $\begin{pmatrix} X_1 \\ Y \end{pmatrix}$ is Gaussian, since obtained via linear function on $\begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$.

Hence X_1, Y are independent. ☺

Therefore, we have obtained the following decomposition for X_2

$$X_2 = aX_1 + \cancel{a}Y$$

DIRE

indep.

Note that ☺

Limit Theorems

Take a sequence $(X_n)_{n \geq 1}$ of iid. random variables in $\mathbb{R}$.

Let $S_n = X_1 + \cdots + X_n$.

We are going to show that, roughly speaking,

$$\frac{S_n}{n} \xrightarrow{n \to \infty} \mu \qquad \text{(Law of large numbers)}$$

$$\frac{S_n - n\mu}{\sigma\sqrt{n}} \xrightarrow{n \to \infty} N(0,1) \qquad \text{(CLT)}$$

L22.4

1) Weak Law of Large Numbers (WLLN)

Theorem Let $(X_n)_{n\geq 1}$ be a sequence of iid. random variables with finite common mean μ.

Let $S_n = X_1 + \cdots + X_n$. Then $\forall \varepsilon > 0$,

$$P\left(\left|\frac{S_n}{n} - \mu\right| > \varepsilon\right) \xrightarrow{n\to\infty} 0. \quad \text{<convergence in probability>}$$

Proof (Assuming finite variance σ^2)

Note that

$$E\left(\frac{S_n}{n}\right) = \mu \quad \text{and} \quad Var\left(\frac{S_n}{n}\right) = \frac{\sigma^2}{n}.$$

Then by Chebyshev,

$$P\left(\left|\frac{S_n}{n} - \mu\right| > \varepsilon\right) \leq \frac{1}{\varepsilon^2}\frac{\sigma^2}{n}.$$

The right hand side goes to 0 as $n \to \infty$. So done. $\square$

2) Strong Law of Large Numbers (SLLN) *non examinable*

Theorem $(X_n)_{n\geq 1}$ as above, $S_n = X_1 + \cdots + X_n$. Then

$$P\left(\lim_{n\to\infty} \frac{S_n}{n} = \mu\right) = 1.$$

To see that SLLN $\Rightarrow$ WLLN, assume $(Y_n)_{n\geq 1}$ are iid and $P(Y_n \to 0) = 1$. Want to deduce $P(|Y_n| > \varepsilon) \xrightarrow[n\to\infty]{} 0$.

Indeed, for any $\varepsilon > 0$,

$1 = P(|Y_n| \leq \varepsilon$ for all n large enough$)$.

Let $A_n = \{|Y_k| \leq \varepsilon \text{ for all } k \geq n\}$.

Then $A_n \subseteq A_{n+1}$ and $\bigcup_{n\geq 1} A_n = \{|Y_k| \leq \varepsilon \text{ for all } k \text{ large enough}\}$.

So

$$\lim_{n\to\infty} P(A_n) = P\left(\bigcup_{n\geq 1}^{\infty} A_n\right) = P(|Y_k| \leq \varepsilon \text{ for all } k \text{ large enough}) = 1$$

$$\Rightarrow \lim_{n\to\infty} P(A_n^c) = 0, \text{ as desired.}$$

L23.1

Limit Theorems

$(X_n)_{n\geq 1}$ iid. $E(X_1) = \mu$

$S_n = X_1 + \cdots + X_n$

$$\frac{S_n}{n} \xrightarrow{n\to\infty} \mu$$

WLLN: $\forall \varepsilon > 0$

$$P\left(\left|\frac{S_n}{n} - \mu\right| > \varepsilon\right) \xrightarrow{n\to\infty} 0$$

convergence in probability

SLLN:

$$P\left(\frac{S_n}{n} \xrightarrow{n\to\infty} \mu\right) = 1$$

almost sure convergence

$(Y_n)_{n\geq 1}$ rvs

$$P(Y_n \to 0) = 1 \implies \forall \varepsilon > 0,\ P(|Y_n| > \varepsilon) \longrightarrow 0$$

almost sure converge $\implies$ convergence in probability $\implies$ convergence in distribution

Cheese, Lettuce, Tomato

$\frac{S_n}{n} \approx \mu$ so $S_n \approx n\mu$ How does it differ?

<u>Theorem</u> Let $(X_n)_{n\geq 1}$ be a sequence of iid rvs with finite mean μ and finite variance σ^2. Set

$$S_n = X_1 + \cdots + X_n$$

Then if $Z \sim N(0,1)$, we have

$$\frac{S_n - n\mu}{\sigma\sqrt{n}} \xrightarrow{d} Z$$

$\int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-z^2/2}\, dz$

i.e. $\forall\ x \in \mathbb{R}$ it holds

$$P\left(\frac{S_n - n\mu}{\sigma\sqrt{n}} \leq x\right) \longrightarrow P(Z \leq x) = \Phi(x).$$

<u>Proof</u> Assume mgf finite in a neighbourhood $(-\delta, \delta)$ of 0.

We prove that

$$M_{\frac{S_n - n\mu}{\sigma\sqrt{n}}}(\theta) \longrightarrow M_Z(\theta)$$

and then use the continuity theorem for mgfs.

L23.2

Note that

$$\frac{S_n - n\mu}{\sigma\sqrt{n}} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \underbrace{\left(\frac{X_i - \mu}{\sigma} \right)}_{Y_i}$$

where the $(Y_i)_{i \geq 1}$ are iid with mean 0 and variance 1.

It suffices to prove theorem for $\mu = 0$, $\sigma = 1$. Need

$$M_{\frac{S_n}{\sqrt{n}}} \longrightarrow M_Z$$

Indeed,

$$\begin{aligned} E\left(e^{\frac{\theta}{\sqrt{n}} S_n}\right) &= E\left(e^{\frac{\theta}{\sqrt{n}}(X_1 + \cdots + X_n)}\right) \\ &= \prod_{i=1}^{n} E\left(e^{\frac{\theta}{\sqrt{n}} X_i}\right) \\ &= \left[E\left(e^{\frac{\theta}{\sqrt{n}} X_1}\right) \right]^n \\ &= \left[M_X\left(\frac{\theta}{\sqrt{n}}\right) \right]^n. \end{aligned}$$

↓ remainder

Let
$$R(x) = \frac{x^3}{2} \int_0^1 e^{tx} (1-t)^2 \, dt$$
and note that
$$e^x = 1 + x + x^2/2 + R(x)$$
via integral form of remainder.

Note that for $\theta \in (-\delta/2, \delta/2)$ we have

$$\begin{aligned} |R(\theta x)| &= \left| \frac{(\theta x)^3}{2} \int_0^1 e^{t\theta x} (1-t)^2 \, dt \right| \\ &\leq \frac{|\theta|^3 |x|^3}{2} \cdot e^{\frac{\delta}{2}|x|} \underbrace{\int_0^1 (1-t)^2 \, dt}_{1/3} \\ &= \underbrace{\left(\frac{\delta |x|}{2} \right)^3}\left(\frac{2|\theta|}{\delta} \right)^3 \cdot \underbrace{\frac{1}{3!}} e^{\frac{1}{2}\delta |x|} \end{aligned}$$

$$\leq e^{\frac{\delta |x|}{2}}$$

L23.3

So $$|R(\theta x)| \leq \left(\frac{2|\theta|}{\delta}\right)^3 e^{\delta|x|}$$
$$\leq \left(\frac{2|\theta|}{\delta}\right)^3 (e^{\delta x} + e^{-\delta x}).$$

Hence $$|R(\theta X_1)| \leq \left(\frac{2|\theta|}{\delta}\right)^3 (e^{\delta X_1} + e^{-\delta X_1})$$

and $$E(|R(\theta X_1)|) \leq \left(\frac{2|\theta|}{\delta}\right)^3 \underbrace{(M_{X_1}(\delta) + M_{X_1}(-\delta))}_{<\infty}$$
$$= o(|\theta|^2).$$

Note that, taking expectations in the identity
$$e^{\theta X_1} = 1 + \theta X_1 + \tfrac{1}{2}(\theta X_1)^2 + R(\theta X_1)$$
we get

remainder

$$M_X(\theta) = 1 + \theta E(X_1) + \tfrac{1}{2}\theta^2 E(X_1^2) + E(R(\theta(X_1)))$$
$$= 1 + \tfrac{1}{2}\theta^2 + o(|\theta|^2).$$

In all, we have
$$M_X\left(\frac{\theta}{\sqrt{n}}\right) = 1 + \frac{1}{2n}\theta^2 + o\left(\frac{|\theta|^2}{n}\right)$$

from which
$$M_{\frac{S_n}{\sqrt{n}}}(\theta) = \left[M_X\left(\frac{\theta}{\sqrt{n}}\right)\right]^n = \left(1 + \frac{\theta^2}{2n} + o\left(\frac{|\theta|^2}{n}\right)\right)^n$$
$$\longrightarrow e^{\frac{1}{2}\theta^2}, \text{ as desired.}$$ □

That whole opinion poll thingy.

Approximate $B(n,p)$ as $N(np, \sqrt{npq}^{\,2})$.

C L T

B

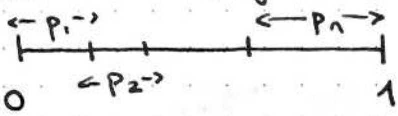
L 23.4

Simulation of RVs

∩ the matrix

Let X be a discrete rv taking values $x_1, \dots, x_n$ with probabilities $p_1, p_2, \dots, p_n$.

Assume can generate uniform numbers in $(0,1)$.



Take $U \sim U[0,1]$ and set

$$X = \sum_{k=1}^{n} x_k \mathbb{1}(U \in I_k)$$

where I_k is the one with length p_k, i.e.

$$I_k = \left[\sum_{j=1}^{k-1} p_j , \sum_{j=1}^{k} p_j \right).$$

Then X has the right distribution, since

$$P(X = x_k) = P(U \in I_k) = p_k \quad \ddot{\smile}$$

This generalises to random variables taking countably many values.

To generate a sequence of iid random variables $(X_n)_{n \geq 1}$ just take a sequence of iid rvs $U[0,1]$.

Continuous: Let X be a continuous rv with pdf f supported on I, with distribution function F. Then F is an increasing bijection from I to $(0,1)$ and the inverse is an increasing bijection $(0,1) \to I$.

Let $U \sim U(0,1)$ and set $X = F^{-1}(U)$. Then we happy, for X has distribution function

$$F_X(x) = P(X \leq x) = P(U \leq F(x)) = F(x) \quad \ddot{\smile}$$

EX: if $U \sim U(0,1]$ then $-\log U \sim \exp(1)$.

in general, $-\frac{1}{\lambda} \log U \sim \exp(\lambda)$.

L24.1

EX: (Box-Müller transform)

Want to simulate X, Y independent $N(0,1)$

Recall that if we define R, Θ by

$$X = R\cos\Theta \quad , \quad Y = R\sin\Theta$$

for $R \in [0,\infty)$, $\Theta \in [0,2\pi)$ then R, Θ are indep with

$$f_R(r) = re^{-r^2/2}\,\mathbb{1}_{[0,\infty)}(r)$$

$$f_\Theta(\theta) = \frac{1}{2\pi}\,\mathbb{1}_{[0,2\pi)}(\theta)$$

Try to generate R, Θ. Have U_1, U_2 indep with

$$U_1 \sim U(0,1] \qquad U_2 \sim U[0,1)$$

Then set $\Theta = 2\pi\,\Theta$ so we happy.

For R, note that $R^2 \sim \exp\left(\frac{1}{2}\right)$ so set $R^2 = -2\log U_1$.

This gives $\begin{cases} X = \sqrt{-2\log U_1}\,\cos(2\pi U_2) \\ Y = \sqrt{-2\log U_1}\,\sin(2\pi U_2) \end{cases}$

Rejection Sampling

Let $A \subset [0,1]^d$ have "positive volume" $|A|$.

Want to simulate $\underline{X}$ uniformly distributed on A, i.e.

for all Borel sets B, want

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

Start with a sequence of iid uniform rvs in $[0,1]^d$ denoted $(\underline{U}_n)_{n\geq 1}$

which can be built with a collection of uniform $[0,1]$ rvs $(U_{n,k})_{n=1,\ k=1}^{\infty\quad d}$

and set

$$\underline{U}_n = (U_{n,1}, \dots, U_{n,d})$$

for each n.

Define

$$N = \min\{n \geq 1 : \underline{U}_n \in A\}.$$

Set

$$\underline{X} = \underline{U}_N.$$

Then for a Borel set B,

$$P(\underline{X} \in B) = P(\underline{U}_N \in B)$$
$$\underset{\text{tot pr}}{=} \sum_{n=1}^{\infty} P(\underline{U}_n \in B \mid N = n)\, P(N = n)$$

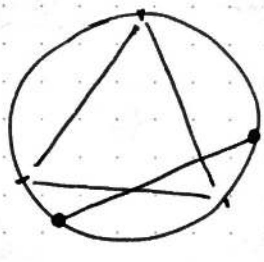
But

$$P(\underline{U}_N \in B \mid N = n) = \frac{P(\underline{U}_N \in B, N = n)}{P(N = n)}$$
$$= \frac{P(U_1 \notin A, \ldots, U_{n-1} \notin A, U_n \in A \cap B)}{P(U_1 \notin A, \ldots, U_{n-1} \notin A, U_n \in A)}$$
$$= \frac{P(U_n \in A \cap B)}{P(U_n \in A)} \quad \text{by independence}$$
$$= \frac{|A \cap B|}{|A|}.$$

Get right answer after distributing.

Geometric Probability

~ Bertrand's paradox



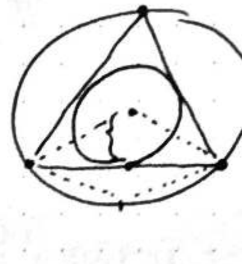
Q: draw a random chord on a circle.
What is the probability that its length exceeds the side of an inscribed equilateral triangle?

Let E be the event of interest.

1) Choose two endpoints uniformly at random & join them.
Then WLOG first at north pole, yaddah yaddah $\frac{1}{3}$

L24.3

2) Pick point uniformly inside disc, then draw chord whose midpoint is the generated point. By ez geometry, get $\frac{1}{4}$.

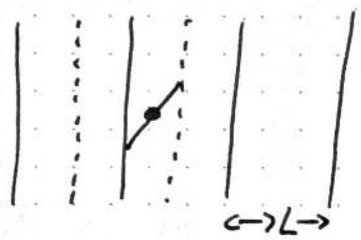


3) Draw a diameter at random, then a point uniformly on the diameter. Then chord with that point as its midpoint. Get $\frac{1}{2}$.

Problem was ill-posed: no information on which probabilities to assign to the chords, yaddah yaddah.

~ <u>Buffon's needle</u>

Roll a needle on a floor with lines separated by L.



Needle has length $\ell \leq L$.

What is probability that needle intersects some line if rolled randomly?

Let X denote the horizontal coordinate of centre.

Let Θ denote angle to horizontal, $\Theta \in [-\frac{\pi}{2}, \frac{\pi}{2})$

$$p = P\left(-\frac{\ell\cos\Theta}{2} \leq X \leq \frac{\ell\cos\Theta}{2}\right)$$

$$= \int_{-\pi/2}^{\pi/2} d\theta \int_{-\frac{\ell\cos\theta}{2}}^{\frac{\ell\cos\theta}{2}} dx \cdot \frac{1}{\pi L} \qquad \text{via pdfs (use } \ell \leq L) \quad X \sim U[-L/2, L/2]$$

$$= \frac{2\ell}{L\pi} \quad \text{can calculate } \pi$$

<u>Idea</u>: use this to estimate π

Take n needles (n large) and roll them independently.

Set $X_k = \{k^{th} \text{ needle crossed}\}$ and

$$\hat{p}_n = \frac{1}{n}\sum_{k=1}^{n} X_k.$$

Then (X_k) are Bernoulli(p) and indep.

Using CLT, get

$$\frac{\hat{p}_n - p}{\sqrt{pq}}\sqrt{n} \sim Z$$

so $$\hat{p}_n \sim p + \sqrt{\frac{pq}{n}}\, Z.$$

Let $f(x) = {}^{2\ell}/_{Lx}$ so $f(\pi) = p$, $f(p) = \pi$.

Taylor expand f

$$\hat{\pi}_n = f(\hat{p}_n) = \pi - \frac{\pi}{p}(\hat{p}_n - p)$$

$$= \pi - \pi\sqrt{\frac{q}{np}}\, Z.$$

If we want

$$P(|\hat{\pi}_n - \pi| > 0.001) \leqslant 0.01,$$

then take $\ell = L$ which gives

$$n \geqslant 3.75 \times 10^7$$ ☹