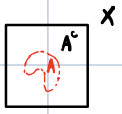
Measure Theory

Let X be a set. Let $\mathcal{A} \subseteq \mathcal{P}(X)$ be a subset of the power set.

We call $\mathcal{A}$ a σ-algebra if the following is true: and we call an element of $\mathcal{A}$ an $\mathcal{A}$-measurable set.

- $\emptyset, X \in \mathcal{A}$;
- $A \in \mathcal{A} \Rightarrow A^c := X \setminus A \in \mathcal{A}$;
- $A_i \in \mathcal{A},\ i \in \mathbb{N} \Rightarrow \bigcup_{i=1}^{\infty} A_i \in \mathcal{A}$.



($i \in \mathbb{N}$: countably many)

Examples: (1) $\mathcal{A} = \{\emptyset, X\}$ (smallest σ-algebra).

(2) $\mathcal{A} = \mathcal{P}(X)$ (largest σ-algebra).

It is "easy to show" the following. Let $\mathcal{A}_i$ be σ-algebras on X, where $i \in \Lambda$ (Λ: arbitrary index set). Then $\bigcap_{i \in \Lambda} \mathcal{A}_i$ is a σ-algebra.

For $M \subseteq \mathcal{P}(X)$, there is a smallest σ-algebra containing M: $\bigcap_{M \subseteq \mathcal{A}} \mathcal{A}$ where $\mathcal{A}$ are σ-algebras.

We define $\sigma(M) := \bigcap_{\substack{M \subseteq \mathcal{A} \\ \mathcal{A}\ \sigma\text{-algebra}}} \mathcal{A}$ often called the σ-algebra generated by M.

Let $X = \{a, b, c, d\}$ and $M = \{\{a\}, \{b\}\}$. Then $\sigma(M) = \{\emptyset, X, \{a\}, \{b\}, \{a, b\}, \{b, c, d\}, \{a, c, d\}, \{c, d\}\}$.

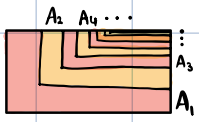
Let X be a topological space. We define the Borel σ-algebra $\mathcal{B}(X)$ to be the σ-algebra generated by open sets, i.e. $\mathcal{B}(X) := \sigma(\tau)$ on some space (X, τ).

A set X and σ-algebra $\mathcal{A}$ form a measurable space $(X, \mathcal{A})$.

A map $\mu : \mathcal{A} \to [0, \infty]$ ($= [0, \infty) \cup \{\infty\}$; volume) is called a measure if the following is fulfilled:

- $\mu(\emptyset) = 0$ (empty set has no volume);
- $\mu\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} \mu(A_i)$ with $A_i \cap A_j = \emptyset,\ \forall A_i \in \mathcal{A}$,

  (additivity of μ).

but we also want the notion of volume approximations. Consider the picture 

with countably many A_i.

We construct the sequence $\{A_1, A_2, \ldots\}$ and can still take unions.

Therefore
- $\mu\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mu(A_i)$ with $A_i \cap A_j = \emptyset,\ \forall A_i \in \mathcal{A}$,

  (σ-additivity).

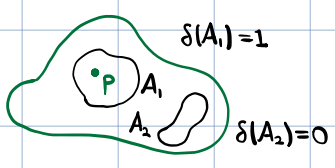
Then we call $(X, \mathcal{A}, \mu)$ a measure space.

Let X be a set and $\mathcal{A}$ be a given σ-algebra on X.

(a) We can define the counting measure for $A \in \mathcal{A}$ by $\mu(A) := \begin{cases} |A| & \text{if } A \text{ has finitely many elements;} \\ \infty & \text{otherwise.} \end{cases}$

We then have the rules $x + \infty := \infty$, $x \cdot \infty := \infty$, $0 \cdot \infty := 0$ (only applicable in measure theory).

(b) We can define the Dirac measure for $p \in X$ by $\delta_p(A) := \begin{cases} 1, & \text{if } p \in A; \\ 0, & \text{if } p \notin A. \end{cases}$



(c) We want a measure on $X = \mathbb{R}^n$ to have the following properties:

the Lebesgue measure
- $\mu([0, 1]^n) = 1$ (the unit cube has "volume" 1);
- $\mu(x + A) = \mu(A)$ for all $x \in \mathbb{R}^n$ ("volume" measurement is translation invariant).

We shall define the concept of a probability model formally.

A **probability model** is a set Ω called the **sample space** together with a collection of subsets $\mathcal{A}$ of Ω called **events**, and a real-valued function $P: \mathcal{A} \to [0,1] \subset \mathbb{R}$ called a **probability measure**.

Probability model:
- Ω sample space
- $\mathcal{A}$ events (collection of subsets of Ω)
- P probability measure

The collection $\mathcal{A}$ of events satisfies the following:

$\mathcal{A}$
- $\Omega \in \mathcal{A}$;
- If $A \in \mathcal{A}$, then $\Omega \setminus A \in \mathcal{A}$; ($\Rightarrow \emptyset \in \mathcal{A}$)
- If $\{A_i\}_{i \in \Lambda} \subseteq \mathcal{A}$ is a (possibly infinite) sequence of events in $\mathcal{A}$, then $\bigcup_{i \in \Lambda} A_i \in \mathcal{A}$. ($\Lambda$ index set)

The probability measure P satisfies the following:

P
- $P(A) \geq 0$ for all $A \in \mathcal{A}$;
- $P(\Omega) = 1$;
- The probability measure is countably additive, i.e.
$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i)$$ for any infinite sequence of pairwise disjoint A_i.

Exercise: Check whether the following $\mathcal{A}$ satisfy the axioms above.

- $\mathcal{A} = \mathcal{P}(\Omega)$ the power set. Yes, trivially.
- $\Omega = \mathbb{R}$ and $\mathcal{A} =$ (all finite $S \subset \mathbb{R}$ and complements).

  1 $\emptyset \in \mathcal{A} \Rightarrow \mathbb{R} = \Omega \in \mathcal{A}$. ✓

  2 $A \in \mathcal{A} \Rightarrow \Omega \setminus A \in \mathcal{A}$ by definition ($\Omega = \mathbb{R}$) (either A finite $\Rightarrow \mathbb{R} \setminus A \in \mathcal{A}$, or A infinite $\Rightarrow A = \mathbb{R} \setminus B$ for some $B \in \mathcal{A}$ $\Rightarrow \mathbb{R} \setminus A = B \in \mathcal{A}$.) ✓

  3 Not satisfied. Take $A_i = \{i\}$ for $i \in \mathbb{N}$ so $\{A_i\}_{i \in \mathbb{N}} \subseteq \mathcal{A}$, but $\bigcup_{i \in \mathbb{N}} A_i \notin \mathcal{A}$. (countable union)

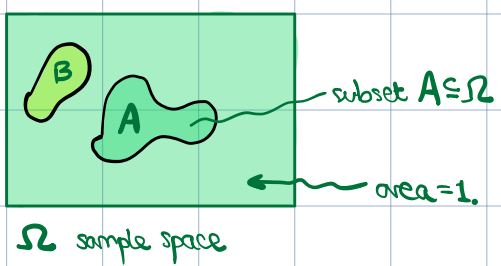
- $\Omega = \mathbb{R}$ and $\mathcal{A} =$ (all countable $S \subset \mathbb{R}$ and complements).

  1, 2 See above. ✓

  3 Should be true since countable unions of countable sets are countable. ✓

- $\mathcal{A} = \{\emptyset, \Omega\}$ trivially.
- $\Omega = \{1, 2, 3, \ldots\}$ and $\mathcal{A}$ is all the sets $A \subset \Omega$ such that $\lim_{n \to \infty} \frac{|A \cap \{1, 2, \ldots, n\}|}{n}$ exists.

From measure theory, probability measures are measures with total mass = 1.



$P: \mathcal{A} \to [0,1]$ ($\mathcal{A}$: collection of subsets of Ω) with:

- $P(\Omega) = 1$;
- $P(A) \in [0,1]$ for $A \in \mathcal{A}$;
- $P(A \cup B) = P(A) + P(B)$ whenever $A \cap B = \emptyset$; (disjoint A and B)
- $P(\bigcup_{j=1}^{\infty} A_j) = \sum_{j=1}^{\infty} P(A_j)$. ($A_j$ disjoint)

Let Ω be a set. A collection of subsets $\mathcal{A} \subseteq \mathcal{P}(\Omega)$ is called a σ-algebra if:

- $\emptyset, \Omega \in \mathcal{A}$;
- If $A \in \mathcal{A}$ then $\Omega \setminus A \in \mathcal{A}$;
- If $A_1, A_2, \ldots \in \mathcal{A}$ then $\bigcup_{j=1}^{\infty} A_j \in \mathcal{A}$. (countable)

} the elements of a σ-algebra are called events in probability theory.

Example: Take a die, roll it once. Then $\Omega = \{1,2,3,4,5,6\}$ (sample space).

Say $\mathcal{A} = \mathcal{P}(\Omega)$. Since every side has the same probability, we have $P: \mathcal{A} \to [0,1]$ given by $P(A) = \frac{\#A}{\#\Omega}$.

(e.g. $P(\{2\}) = \frac{1}{6}$, or $P(\{2,4,6\}) = \frac{3}{6} = \frac{1}{2}$).

Exercise: Prove that $P(\Omega \setminus A) = 1 - P(A)$ for $A \in \mathcal{A}$. ($\mathcal{A}$: σ-algebra)

Probability problems can be categorized into discrete, (absolutely) continuous, and mixed.

Discrete	Continuous
sample space Ω finite or countable	sample space $\Omega \subseteq \mathbb{R}^n$ uncountable with $\Omega \in \mathcal{B}(\mathbb{R}^n)$ a Borel set
σ-algebra: could take $\mathcal{A} = \mathcal{P}(\Omega)$ (standard)	σ-algebra: could take $\mathcal{A} = \mathcal{B}(\Omega)$ the Borel σ-algebra of Ω.
$P: \mathcal{A} \to [0,1]$ determined by $P(\{\omega\})$ for all $\omega \in \Omega$ (singletons) ⇓ define a probability mass function $(p_\omega)_{\omega \in \Omega}$ with $p_\omega \geq 0$, $\sum_{\omega \in \Omega} p_\omega = 1$ so we can write $P(A) := \sum_{\omega \in A} p_\omega$.	$P: \mathcal{A} \to [0,1]$ determined by a probability density function $f: \Omega \to \mathbb{R}$ with $f(x) \geq 0$ (measurable), $\int_\Omega f(x)\,dx = 1$ so ⇒ define $P(A) := \int_A f(x)\,dx$.

Examples: Roll a biased die, so $\Omega = \{1,2,3,4,5,6\}$ and set $p_1 = p_2 = p_3 = p_4 = p_5 = \frac{1}{10}$ and $p_6 = \frac{1}{2}$.

Then $P(\{1,2,3,4,5\}) = \sum_{i=1}^{5} p_i = 5 \cdot \frac{1}{10} = \frac{1}{2}$.

Consider $\Omega = [0,2]$ and $f: \Omega \to \mathbb{R}$ given by $f(x) = \frac{1}{2}$.

Then $\int_\Omega f(x)\,dx = \int_{[0,2]} \frac{1}{2}\,dx = 1$.

Here in general, we have

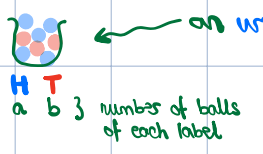
$\int_{A \subseteq \Omega} f(x)\,dx = \frac{1}{2} \int_A 1\,dx = \frac{1}{2} \cdot (\text{Lebesgue measure } (A))$.

Consider tossing a coin with possible outcomes H, T. Then $P(H) = \frac{a}{a+b} \in \mathbb{Q} \cap [0,1]$ for some $a, b \in \mathbb{N}$ (assume)

(a fair coin has $a = b$).

Now consider drawing a ball from a box, where balls are either labelled H or T.

Then $P(H) = \frac{a}{a+b} \in \mathbb{Q} \cap [0,1]$ again. (assume)



In both cases above, we have $\Omega = \{H, T\}$ and $P(\{H\}) = \underbrace{\frac{a}{a+b}}_{p}$ and $P(\{T\}) = \underbrace{\frac{b}{a+b}}_{1-p}$. (probability mass functions) P_H and P_T define P in this discrete case

Both represent binomial distributions:

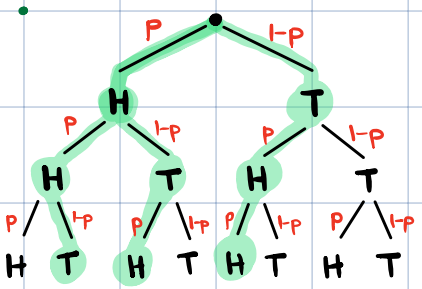
- n tosses of the same coin, counting the heads; (→ without order)
- draw n balls with replacement, counting the heads;
- generally sample size n, unordered, with replacement counting problem.

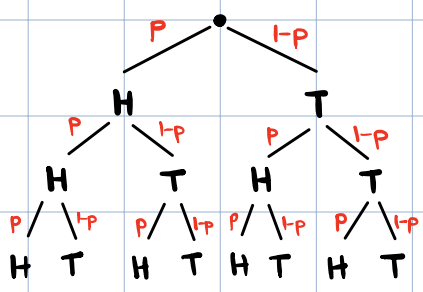
Then we would have $\Omega = \{0, 1, 2, \dots, n\}$ (e.g. heads count) and $P(\{k\}) = \binom{n}{k} p^k (1-p)^{n-k}$ (probability of k heads) with parameters (n, p).

Consider arbitrary $p \in [0,1]$ and $n=3$ (shown on the decision tree on the right).

Then the probability $P(\{k\})$ of choosing k heads H is given by

$$P(\{k\}) = \binom{n}{k} p^k (1-p)^{n-k} = \binom{3}{2} p^2 (1-p)^1.$$





Recall that a probability space is given by $(\Omega, \mathcal{A}, P)$. (Ω: sample space; $\mathcal{A} \subseteq \mathcal{P}(\Omega)$ σ-algebra "events"; P: probability measure $P: \mathcal{A} \to [0,1]$)

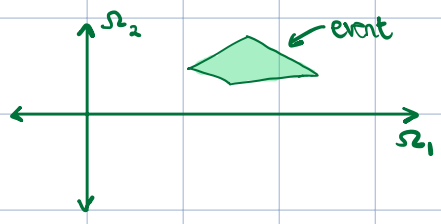
How can we take products of probability spaces? (and why should we?)

Why? Consider the procedure:
1 Throw a die;
2 Throw a point into $[-1,1] \subset \mathbb{R}$. A possible outcome is $(3, \frac{1}{4})$. What is its probability?

First probability space P_1: $(\Omega_1, \mathcal{A}_1, P_1)$ (discrete)
$\{1,\dots,6\}$ $\mathcal{P}(\Omega)$ $P_1(A) = \sum_{k \in A} \frac{1}{6}$

Second probability space P_2: $(\Omega_2, \mathcal{A}_2, P_2)$ (continuous)
$[-1,1]$ $\mathcal{B}(\Omega)$ $P_2(A) = \int_A \frac{1}{2} dx$

We have a new product probability space:
$(\Omega_1 \times \Omega_2, \sigma(\mathcal{A}_1 \times \mathcal{A}_2), P)$
Cartesian product; product σ-algebra; product measure



Consider a collection $(\Omega_n, \mathcal{A}_n, P_n)$ of probability spaces with $n \in \mathbb{N}_{>0}$. Take the case $n=2$.

The product (probability) space is given by $(\Omega_1 \times \Omega_2, \sigma(\mathcal{A}_1 \times \mathcal{A}_2), P)$ where P is a product measure satisfying for $A_1 \in \mathcal{A}_1$ and $A_2 \in \mathcal{A}_2$ we have $P(A_1 \times A_2) = P_1(A_1) \cdot P_2(A_2)$.
($A_1 \times A_2$ lives in product σ-algebra)

Consider the event of ((rolling a 2 or 3), (hitting $[-1,0]$)).

Then $P(\{2,3\} \times [-1,0]) = P_1(\{2,3\}) \cdot P_2([-1,0]) = \sum_{k \in \{2,3\}} \frac{1}{6} \cdot \int_{[-1,0]} \frac{1}{2} dx = \frac{2}{6} \cdot \frac{1}{2} = \frac{1}{6}$.

Let us provide a general definition. Consider a collection $(\Omega_n, \mathcal{A}_n, P_n)$ of probability spaces with $n \in \mathbb{N}_{>0}$.

The product (probability) space $\mathbb{P}$ given by $(\Omega, \mathcal{A}, P)$ is defined by:

- $\Omega = \Omega_1 \times \Omega_2 \times \dots = \prod_{j \in \mathbb{N}_{>0}} \Omega_j$ having elements sequences $(\omega_1, \omega_2, \omega_3, \dots)$ $\in \Omega_1$ $\in \Omega_2$ $\in \Omega_3$
- $\mathcal{A} = \sigma(\text{"cylinder sets"})$ the product σ-algebra.
  - $A_1 \times \Omega_2 \times \Omega_3 \times \dots$
  - $\Omega_1 \times A_2 \times \Omega_3 \times \dots$
  } take the smallest σ-algebra containing all these subsets
- P given by $P(\underbrace{A_1 \times \dots \times A_m}_{\text{change finitely many sets}} \times \Omega_{m+1} \times \Omega_{m+2} \times \dots) = \prod_{1 \le j \le m} P_j(A_j)$ the product measure.

Example: Throw a dice infinitely many times. Since each probability space is the same, denote each as $(\Omega_0, \mathcal{A}_0, P_0)$.
$\{1,\dots,6\}$ $\mathcal{P}(\Omega)$ $P_0(A) = \sum_{k \in A} \frac{1}{6}$

Then the produce space is $(\Omega, \mathcal{A}, P)$ by the definition above.
$\Omega_0 \times \Omega_0 \times \dots$; $\mathcal{A}$: product σ-algebra; P: product measure

Consider the event $A \in \mathcal{A}$ given by A: "at the 100th throw, we roll the first 6."

Thus $A = \underbrace{\{6\}^c \times \dots \times \{6\}^c}_{99 \text{ times}} \times \{6\} \times \underbrace{\Omega_0 \times \Omega_0 \times \dots}_{\text{we don't care what happens after}}$ where $\{6\}^c = \Omega_0 \setminus \{6\}$ the complement.

$$\Rightarrow P(A) = P_0(\{6\}^c) \cdots P_0(\{6\}^c) \cdot P_0(\{6\}) = P_0(\{6\}^c)^{99} \cdot P_0(\{6\}) = \left(\frac{5}{6}\right)^{99} \cdot \frac{1}{6}.$$

The Hypergeometric Distribution

Keep 3 things in mind: size $n \in \mathbb{N}$, sample unordered, no replacement. The urn model provides a nice intuition again.

Consider drawing n balls from a jar containing ball colours C (some finite set). $C = \{0, 1, 2, 3\}$.

One possible outcome for $n=3$ may be so we essentially have some function $f: C \to \mathbb{N}_{\geq 0}$, $(1, 1, 1)$ (adding to $n=3$). (Indexing colours can allow us to write outcomes as tuples)

Then Ω = (sample space) set of all functions (or tuples) of that form $= \{(k_c)_{c \in C} \in \mathbb{N}_{\geq 0}^C \mid \sum_{c \in C} k_c = n\}$.

As all discrete cases, we can take our σ-algebra $\mathcal{A} = \mathcal{P}(\Omega)$.

For our example, we have $\Omega = \{(k_0, k_1, k_2) \in \mathbb{N}_{\geq 0}^3 \mid k_1 + k_2 + k_3 = n\}$.

To know the probability of an outcome, we must know the initial counts of all colours. We introduce N_c = number of balls of colour c in the urn with $N := \sum_{c \in C} N_c$ total count.

Then we have $P(\{k_0, k_1, k_2\}) = \dfrac{\binom{N_0}{k_0} \cdot \binom{N_1}{k_1} \cdot \binom{N_2}{k_2}}{\binom{N}{n}}$. ($k_0$: exactly k_0 red balls; $\binom{N}{n}$: all possibilities to take n balls from N balls.)

We can use probability mass functions in this case. (discrete)

The (multivariant) hypergeometric distribution is described by the probability mass function

$$P(\{(k_c)_{c \in C}\}) = \frac{\prod_{c \in C} \binom{N_c}{k_c}}{\binom{N}{n}}$$ where n are chosen.

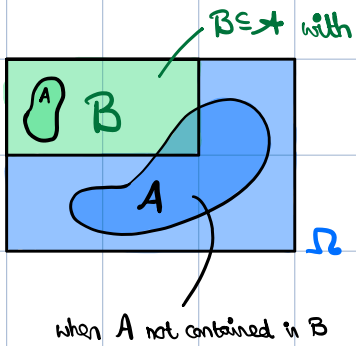
Hypergeometric distribution for 2 colours: $C = \{0, 1\}$ (colours) and $N_0 + N_1 = N$ (counts). Since only two colours exist, we can simply count for one colour to get the other.

Count the 1s: $\Omega = \{0, 1, \dots, n\}$ since when n balls are picked, these are our possibilities.

Then we can define $P: \mathcal{P}(\Omega) \to [0, 1]$ by $P(\{k\}) = \dfrac{\binom{N_1}{k}\binom{N_0}{n-k}}{\binom{N}{n}}$.

Conditional probability is a special type of probability measure defined relative to another probability measure.

Let $(\Omega, \mathcal{A}, P)$ define a probability space, and let $B \subseteq \mathcal{A}$ be a subset of our σ-algebra (of events) with $P(B) \neq 0$.



We can define new probability space(s)

$(B, \tilde{\mathcal{A}}, \tilde{P})$ ($\tilde{\mathcal{A}}$: only consisting $A \in \mathcal{A}$ such that $A \subseteq B$.) $\tilde{P}(A) = \dfrac{P(A)}{P(B)}$ } normalized

but what if we don't want to change our sample space Ω and σ-algebra $\mathcal{A}$?

$\rightsquigarrow (\Omega, \mathcal{A}, P_B)$ where $P_B(A) = \dfrac{P(A \cap B)}{P(B)}$. ($A \in \mathcal{A}$ may not be contained in B!)

Now we define conditional probability.

Let $(\Omega, \mathcal{A}, P)$ define a probability space, and $B \in \mathcal{A}$ an element with $P(B) \neq 0$.

We define $P_B(A) = P(A|B) := \frac{P(A \cap B)}{P(B)}$ the conditional probability of A under B.

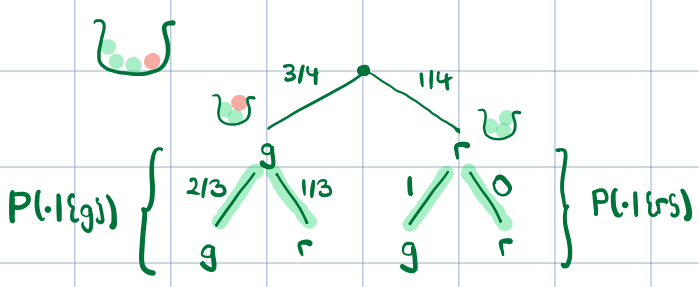
The conditional probability (measure) given B is given by $P(\cdot|B): \mathcal{A} \to [0,1]$. Notice $P(B|B) = 1$ by definition.

If $P(B) = 0$ we define $P(A|B) := 0$. Here, we don't have a well-defined probability measure.

Example: Urn model: ordered, without replacement. Define colour set $C := \{r, g\}$ so $\Omega = C \times C$ when 2 are picked.

$\mathcal{A} = \mathcal{P}(\Omega)$

P probability mass function



$P(\{g,g\}) = 1/2$
$P(\{g,r\}) = 1/4$
$P(\{r,g\}) = 1/4$
$P(\{r,r\}) = 0$

Suppose B = "first ball green". Then $B = \{(g,g), (g,r)\}$ and if $A = \{(g,r)\} \in \mathcal{A}$ then

$$P(A|B) = P(\{(g,r)\}|B) = \frac{P(\{(g,r)\} \cap B)}{P(B)} = \frac{P(\{(g,r)\})}{P(B)} = \frac{\frac{1}{4}}{\frac{3}{4}} = \frac{1}{3}.$$

Counting Principles (Combinatorics)

Consider a bag with n elements, (possibly balls of various colours) where we select k of these elements at random. We want to figure out:

# Possibilities	We replace	We don't replace
Order matters	$n(n-1)\cdots(n-(k-1)) = \frac{n!}{(n-k)!}$	n^k
Order doesn't matter	$\frac{n!}{k!(n-k)!} = \binom{n}{k} = C_k^n$	$\frac{(k+(n-1))!}{k!(n-1)!}$

how much have we overcounted by? (e.g. $213 \equiv 321 \equiv 123$)
how many ways to rearrange k things?

non-trivial case: we have n things, pick k, order doesn't matter, we replace.

n buckets with k tennis balls.

Q: How many ways to put k tennis balls in n buckets?

$\equiv$ k balls and $n-1$ barriers: $\underbrace{\circ\circ|\circ|\circ\cdots\circ|\circ}_{k \text{ balls}} \equiv aababa\cdots aba$

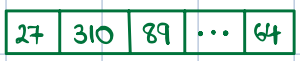
permute 'a's in $k!$ ways
permute 'b's in $(n-1)!$ ways

$$= \frac{(k+(n-1))!}{k!(n-1)!}$$

arrange $(k+(n-1))$ objects, overcounting by $k!(n-1)!$

The Birthday Problem: We have k people, and we want to know the probability of a birthday match.

Assume 365 days and a uniform distribution.



so $|\Omega| = 365^k$

$\Rightarrow \Omega = \{(a_1, a_2, \ldots, a_k) \mid a_i \in \mathbb{Z}/k\mathbb{Z}\}$ and $\mathcal{A} = \mathcal{P}(\Omega)$ because discrete problem.

Say $A \in \mathcal{A}$ represents events with a match. Then $P(A) = 1 - P(\Omega|A) = 1 - P(A^c)$. ($A^c$: no matches)

A^c: $\bullet_{365}$ $\bullet_{364}$ $\cdots$ $\bullet_{365-(k-1)}$ ← entry choices

$$\Rightarrow |A^c| = \frac{365!}{(365-k)!} \quad \text{so} \quad P(A^c) = \frac{|A^c|}{|\Omega|} = \frac{365!}{(365-k)!}\,365^{-k}$$

$$\Rightarrow P(A) = 1 - \frac{365!}{(365-k)!}\,365^{-k}.$$

Theorem (Bayes): Let $(\Omega, \mathcal{A}, P)$ define a probability space with $A, B \in \mathcal{A}$ of nonzero probability.

Then $P(A|B)\,P(B) = P(B|A)\,P(A)$.

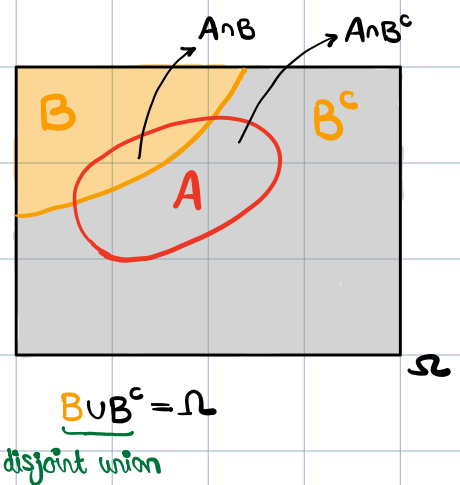
Proof: Simple. We have the conditional probability measure defined above, so

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \quad \text{and} \quad P(B|A) = \frac{P(B \cap A)}{P(A)} = \frac{P(A \cap B)}{P(A)} \Rightarrow P(B|A)\,P(A) = P(A|B)\,P(B). \quad \square$$

($P(A \cap B)$ on both sides)

Law of Total Probability

Consider a probability space $(\Omega, \mathcal{A}, P)$. Say $A \in \mathcal{A}$ and $B \in \mathcal{A}$.



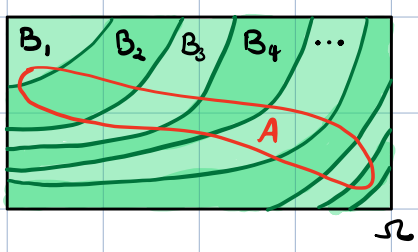
We have $P(A) = P((A \cap B) \cup (A \cap B^c))$ ← disjoint because $B \cap B^c = \emptyset$, so we add

$= P(A \cap B) + P(A \cap B^c)$

$= P(A|B)P(B) + P(A|B^c)P(B^c)$.

$B \cup B^c = \Omega$ (disjoint union)

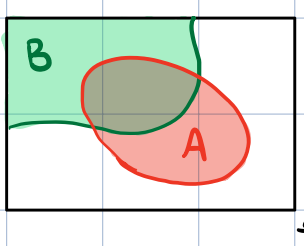
Consider the case with countably many $B_i \in \mathcal{A}$ for $i \in \Lambda \subseteq \mathbb{N}$ (index set) with $\bigcup_{i \in \Lambda} B_i = \Omega$ (disjoint union).



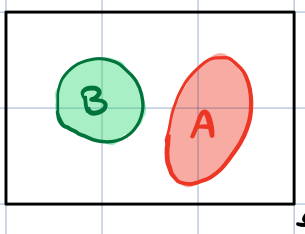
Then $P(A) = P\left(\bigcup_{i \in \Lambda} (A \cap B_i)\right) = \sum_{i \in \Lambda} P(A \cap B_i) = \sum_{i \in \Lambda} P(A|B_i)\,P(B_i)$ (disjoint union)

Example: Monty Hall Problem.

Consider a space $(\Omega, \mathcal{A}, P)$ and events $A, B \in \mathcal{A}$. We want to deduce something about independence.



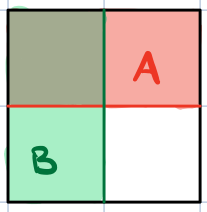
are they independent?



they are mutually exclusive!
are they independent?

We want $P(A|B) = P(A)$ and $P(B|A) = P(B)$.

Notice then $P(A \cap B) = P(A)\,P(B)$.



$P(A) = \frac{1}{2}$, $P(B) = \frac{1}{2}$ $\Rightarrow$ $P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{(\frac{1}{4})}{(\frac{1}{2})} = \frac{1}{2}$, and $P(B|A) = \frac{1}{2}$.

$\Rightarrow$ independent!

Let $(\Omega, \mathcal{A}, P)$ be a probability space. Two events $A, B \in \mathcal{A}$ are independent if $P(A \cap B) = P(A)\,P(B)$.

A family $(A_i)_{i \in \Lambda}$ with $A_i \in \mathcal{A}$ is called independent if $P\left(\bigcap_{j \in J} A_j\right) = \prod_{j \in J} P(A_j)$ for all finite $J \subseteq \Lambda$ (index set).

independence and disjoint are NOT the same thing!

Example: 2 dice rolls with order: $(\Omega, \mathcal{A}, P)$ and A = "first throw gives 6", B = "sum of throws = 7."

$\Omega = \{1, 2, \dots, 6\}^2$, $\mathcal{A} = P(\Omega)$, P = uniform distribution $P(\{\omega_1, \omega_2\}) = \frac{1}{36}$

Then $P(A) = \frac{1}{6}$ and $P(B) = P(\{(1,6), (2,5), (3,4), (4,3), (5,2), (6,1)\}) = \frac{6}{36} = \frac{1}{6}$.

$\Rightarrow P(A \cap B) = \frac{1}{6} P(\{1,6\}) = \frac{1}{36} = P(A)\,P(B) \Rightarrow A$ and B are independent.

Example: throw a point into $[0,1]$. Then $(\Omega, \mathcal{A}, P)$

$\Omega = [0,1]$, $\mathcal{A} = \mathcal{B}(\Omega)$, P = uniform density distribution $f: \Omega \to \mathbb{R}$, $x \mapsto 1$ the indicator function $\mathbb{1}_\Omega(x) = \begin{cases} 1 & \text{if } x \in \Omega; \\ 0 & \text{otherwise.} \end{cases}$

For independent $A, B \in \mathcal{A}$ we have:

$$\int_{A \cap B} \mathbb{1}_\Omega(x)\,dx = P(A \cap B) = P(A)\,P(B) = \int_A \mathbb{1}_\Omega(x)\,dx \int_B \mathbb{1}_\Omega(x)\,dx$$

$\|$ $\|$

$$\int_{[0,1]} \mathbb{1}_{A\cap B}(x)\,dx = \int_{[0,1]} \mathbb{1}_A(x)\,dx \int_{[0,1]} \mathbb{1}_B(x)\,dx$$

Theorem (Bayes' Formula): Let $(\Omega, \mathcal{A}, P)$ define a probability space. Let $B_1, \dots, B_k \in \mathcal{A}$ be events that partition Ω, so $B_i \cap B_j = \emptyset$ and $\bigcup_{i=1}^{k} B_i = \Omega$. Say $P(B_i) > 0$ for all i. Let $A \in \mathcal{A}$.

Then $$P(B_i|A) = \frac{P(A|B_i)\,P(B_i)}{\sum_{j=1}^{k} P(A|B_j)\,P(B_j)}.$$

Proof: For each i, we have $P(B_i|A) = \frac{P(A\cap B_i)}{P(A)} = \frac{P(A|B_i)\,P(B_i)}{P(A)}$ (by definition of the conditional probability measure)

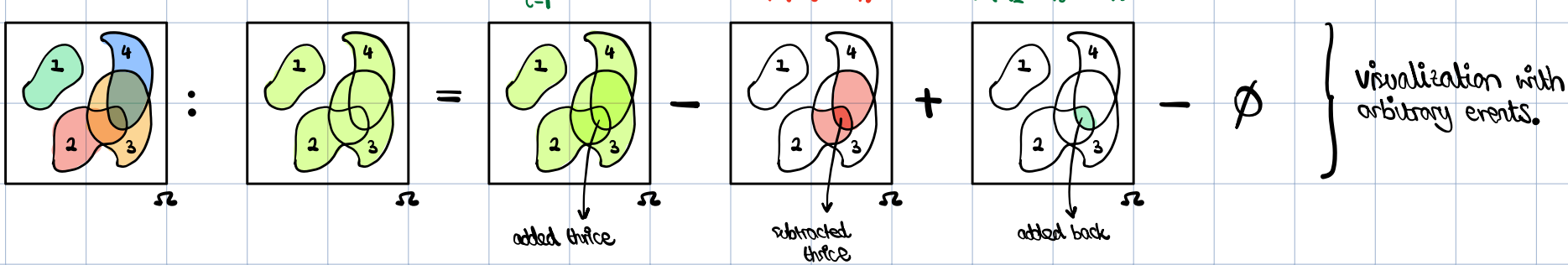
$= \frac{P(A|B_i)\,P(B_i)}{\sum_{j=1}^{k} P(A|B_j)\,P(B_j)}$. (by Law of Total Probability) Done. □

A family $(A_i)_{i\in I}$ with $A_i \in \mathcal{A}$ is independent if $P(\bigcap_{j\in J} A_j) = \prod_{j\in J} P(A_j)$ for any finite $J \subseteq I$, $J \neq \emptyset$.

We say $A_1, \dots, A_n \in \mathcal{A}$ are mutually (or simply) independent if the equation $P(A_{i_1} \cap \dots \cap A_{i_k}) = P(A_{i_1}) \cdots P(A_{i_k})$ holds for all combinations of indices $i_1, \dots, i_k$ distinct.

Proposition (Indusion-Exclusion): Let $(\Omega, \mathcal{A}, P)$ define a space. For arbitrary $A_1, \dots, A_n \in \mathcal{A}$, we have

$$P\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} P(A_i) - \sum_{i<j} P(A_i \cap A_j) + \sum_{i<j<k} P(A_i \cap A_j \cap A_k) - \dots + (-1)^{n-1} P\left(\bigcap_{i=1}^{n} A_i\right).$$



Exercise: There are n letters and n envelopes. Letters are randomly assigned an envelope. What is the probability that no letter is assigned to the envelope it belongs to?

Let $A_i :=$ "envelope i is assigned to letter i" so we want $P(\bigcup_{i=1}^{n} A_i)$. ($P(A_1 \cup A_2 \cup \dots \cup A_n) \approx A_1$ OR A_2 OR $\dots$ OR A_n)

Notice that $P(A_i) = \frac{1}{n}$ for all i.

$$P(A_i \cap A_j) = P(A_j|A_i)\,P(A_i) = \frac{1}{n-1}\cdot\frac{1}{n} \quad \text{for } i<j.$$ (there are $n-1$ choices left given A_i was successful)

$$P(A_i \cap A_j \cap A_k) = P(A_j \cap A_k | A_i)\,P(A_i)$$

$$\vdots \qquad = \frac{1}{(n-2)(n-1)}\cdot\frac{1}{n} \quad \text{for } i<j<k.$$

$$P(A_1 \cap \dots \cap A_n) = \frac{1}{n!}.$$

$$\Rightarrow P\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} \frac{1}{n} - \sum_{i<j} \frac{(n-2)!}{n!} + \sum_{i<j<k} \frac{(n-3)!}{n!} - \dots + (-1)^{n-1}\frac{1}{n!}$$ and now we count:

$$= n\cdot\frac{1}{n} - \binom{n}{2}\frac{(n-2)!}{n!} + \binom{n}{3}\frac{(n-3)!}{n!} - \dots + (-1)^{n-1}\frac{1}{n!}$$

$$= 1 - \sum_{r=2}^{n}(-1)^{r-1}\binom{n}{r}\frac{(n-r)!}{n!} = 1 - \sum_{r=2}^{n}(-1)^{r-1}\frac{n!}{r!(n-r)!}\frac{(n-r)!}{n!} = 1 - \sum_{r=2}^{n}(-1)^{r-1}\frac{1}{r!}.$$

Example: Randomly take 13 cards from a deck of 52. Define $B :=$ "get exactly 6 spades, or exactly one ace." (exclusive-or) Find $P(B)$.

Let $A_1 =$ "getting exactly 6 spades" and $A_2 =$ "getting exactly 1 ace".

Then $P(B) = P(A_1 \cup A_2) = P(A_1) + P(A_2) - P(A_1 \cap A_2)$.

$$P(A_1) = \frac{\binom{13}{6}\binom{39}{7}}{\binom{52}{13}}$$

pick 6 cards from 13 spades; pick 7 cards from 39 non-spades; $|\Omega|$ pick 13 cards from 52 cards

$$P(A_2) = \frac{\binom{4}{1}\binom{48}{12}}{\binom{52}{13}}$$

$A_1 \cap A_2$: getting 6 spades and 1 ace.

$$\binom{12}{6}\binom{3}{1}\binom{36}{5} + \binom{12}{5}\binom{1}{1}\binom{36}{7}$$

6 non-ace spades; 1 non-spade ace; 5 non-ace spades; 1 spade ace; 7 non-ace (-3) non-spade (-13)

getting a non-spade ace; getting a spade ace

$$\rightsquigarrow \frac{\binom{12}{6}\binom{3}{1}\binom{36}{5}}{\binom{52}{13}} + \frac{\binom{12}{5}\binom{1}{1}\binom{36}{7}}{\binom{52}{13}}$$

Let $(\Omega, \mathcal{A}, P)$ be a probability space. We say $A, B \in \mathcal{A}$ are conditionally independent given $C \in \mathcal{A}$ if $P(A \cap B \mid C) = P(A \mid C)\,P(B \mid C)$.

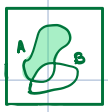
Exercise: Say $A, B \in \mathcal{A}$ with $P(A) = 0.4$ and $P(B) = 0.7$, and $P(A \cup B) = 0.8$.

a) Find $P(A \mid B)$. We have $P(A \mid B) = \frac{P(A \cap B)}{P(B)}$ so we need $P(A \cap B)$. Notice $P(A \cap B) = P(A) + P(B) - P(A \cup B)$.

so we are done (use calculator).

b) Find $P(A \mid B^c)$. We have $P(A \mid B^c) = \frac{P(A \cap B^c)}{P(B^c)}$ ($P(B^c) = 1 - P(B)$) $\rightsquigarrow P(A \cap B^c) = P(A) - P(A \cap B)$. Done.

($P(A \cap B)$ from (a))



c) Find $P(A^c \mid B^c)$. We have $P(A^c \mid B^c) = \frac{P(A^c \cap B^c)}{P(B^c)}$ ($P(B^c) = 1 - P(B)$) $\rightsquigarrow P(A^c \cap B^c) = P((A \cup B)^c) = 1 - P(A \cup B)$. Done.

space $(\Omega, \mathcal{A}, P)$ — Ω: sample space; E: measurable space (usually $\mathbb{R}$)

A random variable X is a measurable function $X: \Omega \to E$.

The probability that X takes on a value in a measurable $S \subseteq E$ is denoted $P(X \in S) = P(\{\omega \in \Omega \mid X(\omega) \in S\})$.
(P: probability measure; $\{\omega \in \Omega \mid X(\omega) \in S\} \in \mathcal{A}$)

When $|\text{Image}(X)| \leq \aleph_0$ we call X a discrete random variable and its distribution discrete.
⇒ defined by probability mass functions

When $\text{Image}(X) \subseteq E$ is uncountable we call X a continuous random variable.

In the special case that X is absolutely continuous, its distribution can be defined by a probability density function.
(in particular, each point must have probability zero)

Here's a more intuitive (yet still accurate) explanation.

Suppose we roll 2 dice, so we have a space $(\Omega, \mathcal{A}, P)$ with $\Omega = \{1,2,\dots,6\}^2$, $\mathcal{A} = \mathcal{P}(\Omega)$, P uniform discrete. We may be interested in probabilities of their sums.

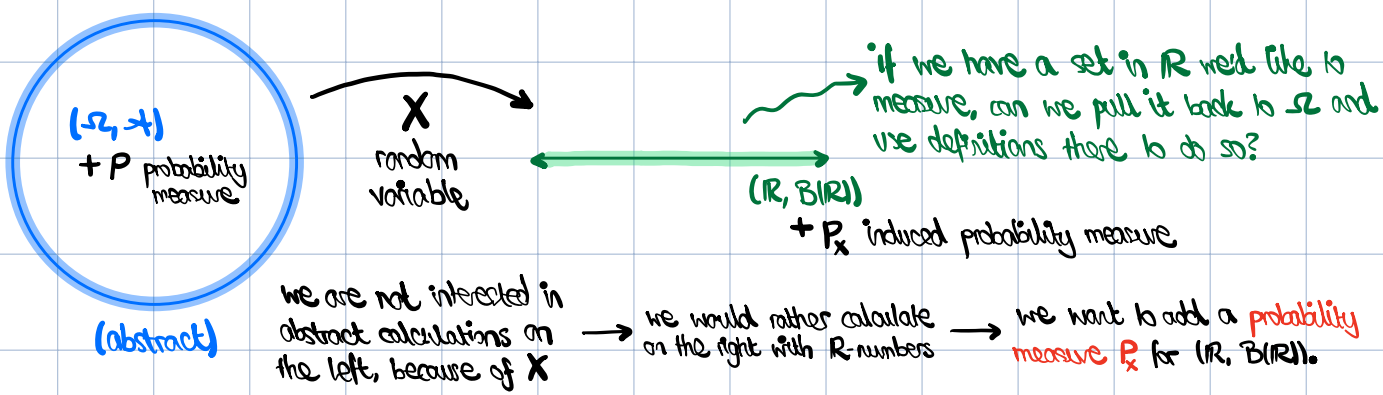
Define a function $X: \Omega \to \mathbb{R}$ by
$X((x_1, x_2)) = x_1 + x_2$ ($(x_1,x_2) \in \mathcal{A}$, $x_1+x_2 \in \mathbb{R}$)
and this is a random variable.

Let $(\Omega, \mathcal{A})$ and $(\tilde{\Omega}, \tilde{\mathcal{A}})$ be measurable spaces (event spaces). A (measurable) map $X: \Omega \to \tilde{\Omega}$ is called a random variable if $X^{-1}(\tilde{A}) \in \mathcal{A}$ ($X^{-1}(\tilde{A}) \in \mathcal{A}$) for all $\tilde{A} \in \tilde{\mathcal{A}}$.

Define $(\Omega, \mathcal{A})$ with $\Omega = \{1,2,\dots,6\}^2$, $\mathcal{A} = \mathcal{P}(\Omega)$ and $(\tilde{\Omega}, \tilde{\mathcal{A}})$ with $\tilde{\Omega} = \mathbb{R}$, $\tilde{\mathcal{A}} = \mathcal{B}(\mathbb{R})$ for rolling two dice. Define $X: \mathcal{A} \to \tilde{\mathcal{A}} = \mathbb{R}$, $(w_1, w_2) \mapsto w_1 + w_2$.

Then $X^{-1}(\tilde{A}) \in \mathcal{A} = \mathcal{P}(\Omega)$ is trivially fulfilled. ← Why? I don't fully understand here.

Let $(\Omega, \mathcal{A})$ and $(\tilde{\Omega}, \tilde{\mathcal{A}})$ with $\tilde{\Omega} = \mathbb{R}$, $\tilde{\mathcal{A}} = \mathcal{B}(\mathbb{R})$ be measurable spaces (event spaces). On one hand, we have an abstract event space, and on the other, a very concrete one.



Let $(\Omega, \mathcal{A}, P)$ be a probability space, and $X: \Omega \to \mathbb{R}$ (with σ-algebra $\mathcal{B}(\mathbb{R})$) a random variable.

We define $P_X: \mathcal{B}(\mathbb{R}) \to [0,1] \subseteq \mathbb{R}$ with $P_X(B) = P(X^{-1}(B)) = P(X \in B)$ to be the probability distribution of X.
($B \in \mathcal{B}(\mathbb{R})$; P: probability in initial space; $X^{-1}(B)$: pre-image of $B \in \mathcal{B}(\mathbb{R})$ under random var. X; another notation)

Proposition (P_X): Let $(\Omega, \mathcal{A}, P)$ be a probability space and $X: \Omega \to \mathbb{R}$ a random variable.

Then P_X defines a probability measure on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$.

Proof: Notice $X^{-1}(\mathbb{R}) = \Omega$, and $X^{-1}(\emptyset) = \emptyset$. We proceed by definition.

1 $P_X(\mathbb{R}) = P(X \in \mathbb{R}) = P(X^{-1}(\mathbb{R})) = P(\Omega) = 1.$ Δ

2 $P_X(\emptyset) = P(X \in \emptyset) = P(X^{-1}(\emptyset)) = P(\emptyset) = 0.$ (since P probability measure) Δ

3 σ-additivity: Say $\{A_i\}_{i \in \Lambda}$ is a pairwise disjoint countable collection with $A_i \in \mathcal{B}(\mathbb{R})$.

Thus by the stability of pre-images under $\cap$, we have $i \neq j \Rightarrow X^{-1}(A_i) \cap X^{-1}(A_j) = X^{-1}(\underbrace{A_i \cap A_j}_{\emptyset}) = \emptyset$.

$\Rightarrow \{\underbrace{X^{-1}(A_i)}_{\in \mathcal{A}}\}_{i \in \Lambda}$ pairwise disjoint.

We have $P_X(\bigcup_{i \in \Lambda} A_i) = P(X \in \bigcup_{i \in \Lambda} A_i) = P(X^{-1}(\bigcup_{i \in \Lambda} A_i))$ but pre-images are stable under unions, so

$= P(\bigcup_{i \in \Lambda} X^{-1}(A_i)) = \sum_{i \in \Lambda} P(X^{-1}(A_i)) = \sum_{i \in \Lambda} P_X(A_i).$ Δ

(σ-additivity of P since $X^{-1}(A_i) \in \mathcal{A}$ and $\{X^{-1}(A_i)\}_{i \in \Lambda}$ pairwise disjoint)

4 Trivially $P_X(A) \geq 0$ for $A \in \mathcal{B}(\mathbb{R})$ since $P_X(A) = P(\underbrace{X^{-1}(A)}_{\in \mathcal{A}}) \geq 0.$ Δ $\square$

If $\tilde{P}$ is a probability measure on $\mathbb{R}$, and $P_X = \tilde{P}$, then we write $X \sim \tilde{P}$ and say X is distributed as $\tilde{P}$.

Example: Consider n tosses of a coin with probability p_h of tossing a heads. Define $X: \Omega \to \mathbb{R}$ by $X(\omega) = \#1$s in ω.
$\Rightarrow X \sim Bin(n, p)$

Then $(\underbrace{\Omega}_{\{0,1\}^n}, \underbrace{\mathcal{A}}_{P(\Omega)}, \underbrace{P}_{\text{BERNOULLI}})$ $\xrightarrow{X}$ $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$.

$\{0,1\}^n$: T H; $P(\Omega)$; BERNOULLI: $P(\{\omega\}) = p_h^{\#1}(1-p_h)^{\#0}$ (#1 = number of 1s/Hs in ω); X = random variable; $X \sim Bin(n, p)$

Example: Roll two 4-sided dice such that $(\underbrace{\Omega}_{\{1,2,3,4\}^2}, \mathcal{A}, P)$ is defined as standard. Define $X: \Omega \to \mathbb{R}$, $(\omega_1, \omega_2) \mapsto \omega_1 + \omega_2$ the sum of the rolls.

Then $X^{-1}(\{4\}) = \{\omega \in \Omega \mid X(\omega) = 3\}.$

$= \{\underbrace{(1,3)}_{\in \Omega}, \underbrace{(2,2)}_{\in \Omega}, \underbrace{(3,1)}_{\in \Omega}\} \subseteq \mathcal{A}.$

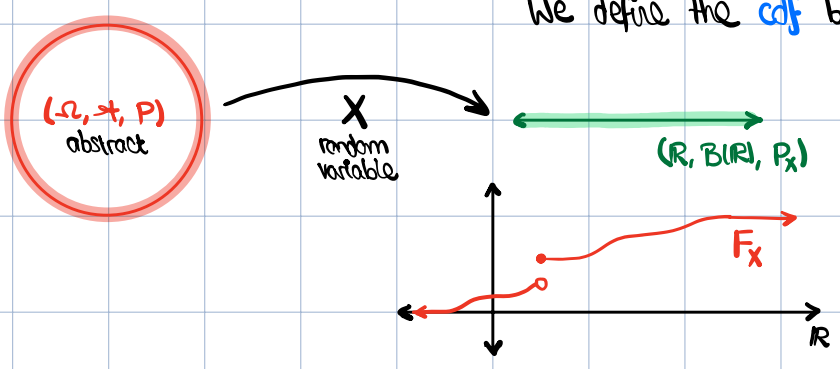
Then $P_X(\{4\}) = P(X \in \{4\}) = P(X^{-1}(\{4\})) = P(\{(1,3), (2,2), (3,1)\}).$

Remark: An alternate common notation for $P_X(\{a_1, \ldots, a_n\})$ is $P(\underbrace{X = a_1, \ldots, a_n}_{X \in \{a_1, \ldots, a_n\}})$.

$P_X([a,b])$ is $P(\underbrace{a \leq X \leq b}_{X \in [a,b]})$.

We define a cumulative distribution function (cdf) and will show that every random variable X has a cdf.

We define the cdf by $F_X: \mathbb{R} \to [0,1]$ with $F_X(x) = P_X((-\infty, x])$ i.e.
$= P(X \in (-\infty, x])$ i.e.
$= P(X \leq x).$



- nondecreasing (monotonically increasing)
- $\lim_{x \to -\infty} F_X = 0$
- $\lim_{x \to \infty} F_X = 1$
- F_X is right-continuous ($\lim_{x \to x_0^+} F_X(x) = F(x_0)$) reason: x included in interval $\Rightarrow$ singleton gives nonempty set

Let $(\Omega, \mathcal{A}, P)$ be a discrete probability space, and $X: \Omega \to \mathbb{R}$ a random variable.

We define the expected value (or expectation) of X as the weighted sum $E[X] = \sum_{\substack{x \in \mathbb{R} \\ P_X(x) > 0}} P_X(\{x\}) \cdot x$.

($P_X(\{x\}) \in \mathbb{R}_{>0}$; $x \in \mathbb{R}$; multiplication in $\mathbb{R}$; countable sum in the discrete case.)

Proposition (Linearity of Expectation): Let X be a discrete random variable. Then for $a, b \in \mathbb{R}$, we have $E[aX+b] = aE[X]+b$ (linearity).

Proof: Computation: $E[aX+b] = \sum_x (ax+b) P_X(x) = a\sum_x x P_X(x) + b \underbrace{\sum_x P_X(x)}_{1} = aE[X]+b$ by definition. $\square$

(see Theorem below!)

Theorem (Law of Unconscious Statistician): Let $(\Omega, \mathcal{A}, P)$ be a discrete probability space. Let $X: \Omega \to \mathbb{R}$ be a random variable, and say $g: \mathbb{R} \to \mathbb{R}$ is measurable. Then $E[g(X)] = \sum_{\substack{x \in \mathbb{R} \\ \text{countable}}} g(x) P_X(x)$.

Proof: By definition, we have $X: \Omega \to \mathbb{R}$ a measurable function. Main idea: Define a composition random variable $\tilde{X} = g(X) = X \circ g$.

We have $\tilde{X}: \Omega \xrightarrow{X} \mathbb{R} \xrightarrow{g} \mathbb{R}$ and $\tilde{X}$ measurable since X and g measurable.

Consider the induced probability measure $P_{g(X)} = P_{\tilde{X}}$ defined by $P_{\tilde{X}}(A) = P(\tilde{X} \in A) = P(\tilde{X}^{-1}(A))$.

(pullback of A under $\tilde{X}$)

For $A \in \Omega$ we have $A \mapsto X(A) \mapsto g(X(A))$, so $\tilde{X}^{-1}(A) = (X \circ g)^{-1}(A) = (g^{-1} \circ X^{-1})(A) = X^{-1}(g^{-1}(A))$.

We have $E[g(X)] = E[\tilde{X}] = \sum_x x \cdot P_{\tilde{X}}(x)$

$= \sum_x x \cdot P(X^{-1}(g^{-1}(x)))$ so consider $y = g^{-1}(x) \Rightarrow x = g(y)$:

$= \sum_y g(y) \cdot P(X^{-1}(y)) = \sum_y g(y) P_X(y)$. We are done. $\square$

Elementary Distributions For this section, we do computation. Implicitly assume $(\Omega, \mathcal{A}, P)$ defines our abstract space.

Proposition (E[Binomial]): Let $X \sim Bin(n,p)$, so $P_X(x) = \binom{n}{x} p^x (1-p)^x$ ($x \in \mathbb{R}$) where p is the probability of event success.

Then $E[X] = np$.

Proof: By definition, we have $E[X] = \sum_{\substack{x \in \mathbb{R} \\ P_X(x) > 0}} P_X(x) \cdot x$ (n trials so at most $x=n$ successes, at least $x=0$ successes)

(recall we have $P(\Omega) = P_X(\mathbb{R}) = \sum_{x=0}^{n} \binom{n}{x} p^x (1-p)^{n-x} = 1$ by definition of a probability measure)

$= \sum_{x=0}^{n} x \cdot \binom{n}{x} p^x (1-p)^{n-x}$ (values of x that make sense)

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

($x=0$ term gives 0)

$= \sum_{x=1}^{n} x \cdot \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x}$

$= np \sum_{x=1}^{n} \frac{(n-1)!}{(x-1)!(n-x)!} p^{x-1} (1-p)^{n-x}$ and now we re-index:

$= np \sum_{x+1}^{n} \frac{(n-1)!}{(x-1)!((n-1)-(x-1))!} p^{x-1} (1-p)^{n-1-(x-1)}$

$= np \underbrace{\sum_{x=0}^{m} \frac{m!}{x!(m-x)!} p^x (1-p)^{m-x}}_{1} = np$ and we are done. $\square$

We define the Poisson distribution to be $\frac{e^{-\mu} \mu^x}{x!}$ (defined by a Poisson process) where μ is an intensity parameter. We revisit this when looking at Stochastic processes.

Proposition (E[Poisson]): Let $X \sim \text{Poisson}(\lambda)$. Then $E[X] = \lambda$.

Proof: By definitions. We have $E[X] = \sum_x x \cdot P_X(x) = \sum_{x=0}^{\infty} x \frac{e^{-\lambda}\lambda^x}{x!} = \sum_{x=1}^{\infty} x \frac{e^{-\lambda}\lambda^x}{x!} = \sum_{x=1}^{\infty} \frac{e^{-\lambda}\lambda^x}{(x-1)!} = \lambda \underbrace{\sum_{x=0}^{\infty} \frac{e^{-\lambda}\lambda^{x-1}}{(x-1)!}}_{1} = \lambda$ and we are done. □

Let $(\Omega, \mathcal{A}, P)$ be a probability space, and $X: \Omega \to \mathbb{R}$ a random variable.

We define the variance by $\text{Var}(X) = E[(X-\mu)^2]$ where $\mu = E[X]$ (so $\text{Var}[X] = E[(X-E[X])^2]$).

"measure of spread" — measuring how much X deviates from mean

Proposition (Properties of Variance): Let X be a random variable. Then:

1. $\text{Var}(aX+b) = a^2\text{Var}(X)$ for $a, b \in \mathbb{R}$;
2. $\text{Var}(X) = E[X^2] - \mu^2$ where $\mu = E[X]$.

Proof (1): Definitions. $\text{Var}(X) = E[(X-E[X])^2]$

$\Rightarrow \text{Var}(aX+b) = E[(aX+b-E[aX+b])^2] \overset{\text{linearity of } E[X]}{=} E[(aX+b-aE[X]-b)^2]$

$= E[a^2(X-E[X])^2] \overset{\text{linearity}}{=} a^2 E[(X-E[X])^2] = a^2\text{Var}(X)$. △

(2): Definitions. $\text{Var}(X) = E[(X-E[X])^2] = E[X^2 - 2X\underbrace{E[X]}_{\mu} + \underbrace{E[X]^2}_{\mu^2}]$

$\overset{\text{linearity}}{=} E[X^2] - 2\mu\underbrace{E[X]}_{\mu} + \mu^2 = E[X^2] - \mu^2$. △□

Proposition (Poisson Approximation for Binomial Distribution): Let $X \sim \text{Binomial}(n, p)$ and $Y \sim \text{Poisson}(\lambda)$ where $\lambda = np$.

Then $X \approx Y$ for large n, and $X \xrightarrow{n\to\infty} Y$.

Proof (Rough): Notice $E[X] = np = E[Y]$. This justifies our choice $\lambda = np$. Thus $p = \frac{\lambda}{n}$.

We have $P_X(x) = \binom{n}{x} p^x (1-p)^{n-x}$

$= \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x}$

$= \frac{1}{x!} \underbrace{n(n-1)\cdots(n-x+1)}_{\downarrow n\to\infty} p^x (1-p)^{n-x} \xrightarrow{n\to\infty} \frac{1}{x!} n^x \left(\frac{\lambda}{n}\right)^x \left(1-\frac{\lambda}{n}\right)^n = \frac{1}{x!}\lambda^x e^{-\lambda} = P_Y(x)$. □

($n \cdots n = n^x$, x times)

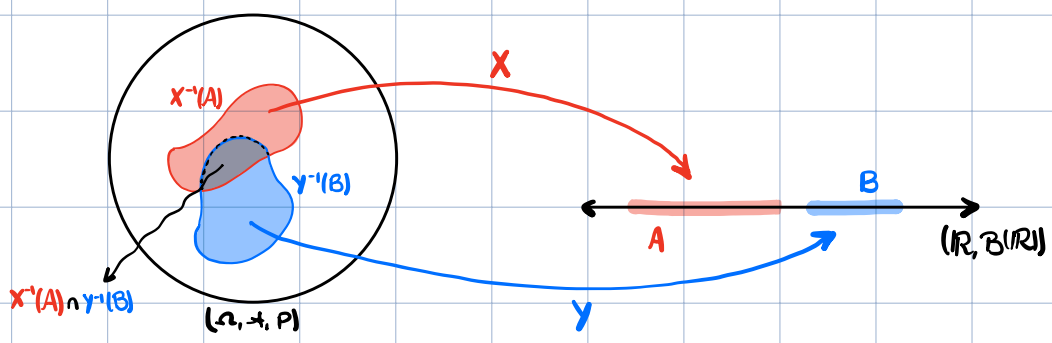
Multivariate Distributions

Let $(\Omega, \mathcal{A}, P)$ be a probability space, and let $X: \Omega \to \mathbb{R}$, $Y: \Omega \to \mathbb{R}$ be random variables on Ω.

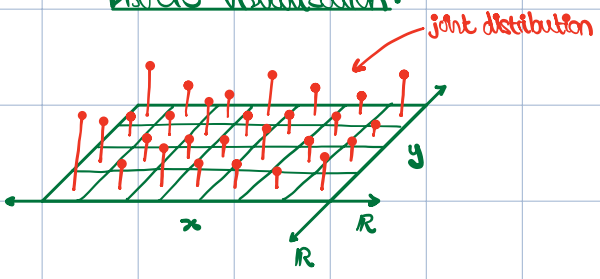
Q: How do X and Y relate/interact?

We can define a bivariate probability function $P_{X,Y}(x)$ induced by X and Y, by $P_{X,Y}(A, B) = P(X^{-1}(A) \cap Y^{-1}(B))$, $A, B \in \mathcal{B}(\mathbb{R})$.

$P_{X,Y}: \mathcal{B}(\mathbb{R})^2 \to [0,1] \subseteq \mathbb{R}$



Discrete visualization:



Let $(\Omega, \mathcal{A}, P)$ be a probability space and $X, Y: \Omega \to \mathbb{R}$ be random variables. Then X and Y are independent random variables

if $X^{-1}((-\infty, x])$ and $Y^{-1}((-\infty, x])$ are independent events for any $x, y \in \mathbb{R}$.

$$\Leftrightarrow P(X^{-1}((-\infty,x]) \cap Y^{-1}((-\infty,x])) = P(X^{-1}((-\infty,x]))\, P(Y^{-1}((-\infty,x])) \Leftrightarrow \underbrace{P(X \le x,\ Y \le y)}_{F_{(X,Y)}(x,y)} = F_X(x)\, F_Y(y).$$

(notation; cumulative distribution function)

We define $F_{(X,Y)}(x,y) = P(X \le x,\ Y \le y) = P(X^{-1}((-\infty,x]) \cap Y^{-1}((-\infty,x]))$ the joint cumulative distribution function

$F_{(X,Y)}: \Omega \to \mathbb{R}^2$. Thus random variables X and Y are independent $\Leftrightarrow F_{(X,Y)}(x,y) = F_X(x)\, F_Y(y)$.

Example: Consider Ω a product space $\Omega = \Omega_1 \times \Omega_2$. Then any random variables of the form

$X: \Omega \to \mathbb{R}$, $(w_1, w_2) \mapsto f(w_1)$ and $Y: \Omega \to \mathbb{R}$, $(w_1, w_2) \mapsto g(w_2)$ are independent.

A family $(X_i)_{i \in \Lambda}$ of random variables is independent if $P((X_j \le x_j)_{j \in J}) = \prod_{j \in J} P(X_j \le x_j)$ for all finite $J \subseteq \Lambda$, $J \neq \emptyset$, for all $x_j \in \mathbb{R}$.

Let $(\Omega, \mathcal{A}, P)$ be a probability space, and $X: \Omega \to \mathbb{R}$ a random variable. We define the (general) expectation

$$E[X] = \int_\Omega X \, dP.$$

(measure P; Lebesgue integral)

Say we have a measurable function $g: \mathbb{R} \to \mathbb{R}$. We can define a new random variable $g(X): \Omega \to \mathbb{R}$.

How does $E[X]$ relate to $E[g(X)]$?

Consider $\int_A g(X)\, dP$ for $A \in \mathcal{A}$. We have

$$\int_A g(X)\, dP = \int_A g(\underbrace{X(w)}_{x})\, dP(w) = \int_{X(A)} g(x)\, d(P \circ X^{-1})(x) = \int_{X(A)} g(x)\, dP_X(x).$$

(integration in some abstract space; change of variables; P_X new measure; integration in $\mathbb{R}$; x so $X^{-1}(x) = w$)

Discrete case: $\sum_{x \in X(A)} g(x)\, P_X(\{x\})$. Continuous case: $\int_{X(A)} g(x)\, f_X(x)\, dx$.

Proposition (Monotonicity of Expectation): Let $(\Omega, \mathcal{A}, P)$ be a probability space, and $X, Y: \Omega \to \mathbb{R}$ random variables.

If $X \le Y$ almost surely (so $P(\{\omega \in \Omega \mid X(\omega) \le Y(\omega)\}) = 1$), then $E[X] \le E[Y]$.

We define the standard deviation $\sigma(X)$ of a random variable X by $\sigma(X) = \sqrt{\mathrm{Var}(X)}$.

Proposition: Let X and Y be independent random variables. Then $E[XY] = E[X]E[Y]$ and $\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$.

Proof: By definitions and algebra. Not very insightful. □

Proposition: Let X and Y be independent random variables, and $g, h: \mathbb{R} \to \mathbb{R}$ be measurable. Then $E[g(X)\,h(Y)] = E[g(X)]\,E[h(Y)]$.

Proof: Denote our probability space $(\Omega, \mathcal{A}, P)$. We have $E[g(X)] = \int_\Omega g(X)\, dP = \int_{\mathbb{R}} g(x)\, dP_X(x)$.

$$E[h(Y)] = \int_\Omega h(Y)\, dP = \int_{\mathbb{R}} h(y)\, dP_Y(y).$$

We prove the following Theorem instead:

Theorem (Functions Preserve Independence): Let X and Y be independent random variables $X, Y: \Omega \to \mathbb{R}$.

Let $g, h: \mathbb{R} \to \mathbb{R}$ be functions. Then $g(X)$ and $h(Y)$ are independent random variables.

Proof: Let $A, B \subseteq \mathbb{R}$. Notice $g^{-1}(A), h^{-1}(B) \subseteq \mathbb{R}$. Then we have

$$P(g(X)\in A,\ h(Y)\in B) = P(X\in g^{-1}(A),\ Y\in h^{-1}(B)) \overset{\text{independence}}{=} P(X\in g^{-1}(A))\,P(Y\in h^{-1}(B)) = P(g(X)\in A)\,P(h(Y)\in B).$$

Done by definitions. □

We showed $g(X)$ and $h(Y)$ are independent, so we are done. □

Let $(\Omega, \mathcal{A}, P)$ be a probability space, and $X, Y: \Omega \to \mathbb{R}$ random variables. We define the covariance of X and Y by

$$\mathrm{Cov}(X,Y) = E[(\underbrace{X-E[X]}_{X-\mu_X})(\underbrace{Y-E[Y]}_{Y-\mu_Y})] \quad \text{(comparing how } (X-\mu_X) \text{ and } (Y-\mu_Y) \text{ relate).}$$

Corollary: We have $\mathrm{Cov}(X,Y) = E[XY] - E[X]E[Y]$.

Corollary: Let X and Y be independent random variables. Then $\mathrm{Cov}(X,Y) = 0$.

Proposition (Cauchy-Schwarz): Let X and Y be random variables. Then $\mathrm{Cov}(X,Y)^2 \le \mathrm{Cov}(X,X)\,\mathrm{Cov}(Y,Y)$.

Let X and Y be random variables. We define the correlation coefficient ρ by $\rho = \dfrac{\mathrm{Cov}(X,Y)}{\sigma(X)\sigma(Y)} = \dfrac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}}$.

Proposition (Correlation): Let X, Y be random variables. Then $|\rho| \le 1$ and $|\rho| = 1 \iff Y = aX+b$ for some $a, b \in \mathbb{R}$.

Proof: We have $|\rho| \le 1$ immediate by Cauchy-Schwarz. □

Example: Suppose 15 televisions are to be purchased by a local bar from a large production run of 500.

Say 450 will last at least 5 years without needing repair.

a) Calculate the probability that ≥ 12 last ≥ 5 years.

We have a Binomial distribution with $p_{success} = 0.9$. Then say $X \sim \mathrm{Bin}(15, 0.9)$ (15: number of trials)

so we want $P(X \ge 12) = P(X=12) + P(X=13) + P(X=14) + P(X=15)$

$$= \sum_{x=12}^{n} \binom{n}{x} p^x (1-p)^{n-x} = \binom{15}{12}(0.9)^{12}(0.1)^3 + \binom{15}{13}(0.9)^{13}(0.1)^2 \approx 0.944.$$
$$+ \binom{15}{14}(0.9)^{14}(0.1)^1 + \binom{15}{15}(0.9)^{15}(0.1)^0$$

We can also approximate this using a Poisson process, where on average 0.02 per television dies annually, or $X \sim \mathrm{Poisson}(0.02)$. With 15 televisions and 5 years, we'd take $X \sim \mathrm{Poisson}(0.10)$ (per television) and

$$P_X(x) = \frac{e^{-\lambda}\lambda^x}{x!} = \frac{e^{-1.5}1.5^x}{x!}$$ describes the probability of seeing x such events over 5 years.

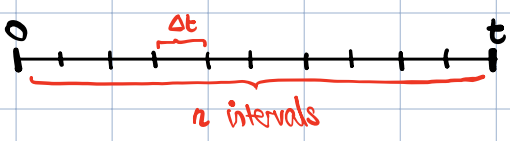
We want $\sum_{x=0}^{3} P_X(x) = (e^{-1.5})\left(\frac{1.5^0}{0!} + \frac{1.5^1}{1!} + \frac{1.5^2}{2!} + \frac{1.5^3}{3!}\right) \approx 0.934.$

Consider wanting to model the distribution of events over time, with the following assumptions:

1. **Independence.** The probability of events occuring in non-overlapping intervals are independent.
2. **Individuality.** Probability of simultaneous events occuring is 0, so $P[\ge 2 \text{ events in } [t, t+\Delta t)] = o(\Delta t)$ ($\lim_{\Delta t \to 0} \frac{o(\Delta t)}{\Delta t} = 0$)
3. **Homogeneity.** Regardless of time interval $[t+\Delta t]$, we have $P(X \text{ occurs in } [t, t+\Delta t]) = \lambda \Delta t + o(\Delta t)$.

error term going to 0 as $\Delta t \to 0$

Consider a unit time interval (length t), and slice it uniformly into n intervals:



and define $\Delta t = \frac{t}{n}$.

Notice $P[1 \text{ event in interval of size } \Delta t] \xrightarrow{\Delta t \to 0} \lambda \Delta t = \lambda \frac{t}{n}$. We can model this binomially:

$$P[1 \text{ event in } \Delta t] = \lambda \tfrac{t}{n} \Rightarrow P[k \text{ events in } t] = \underbrace{\binom{n}{k}}_{n \text{ trials of } \lambda \frac{t}{n}} \left(\lambda \tfrac{t}{n}\right)^k \left(1 - \lambda \tfrac{t}{n}\right)^{n-k} = \frac{n!}{k!(n-k)!} \lambda^k \left(\tfrac{t}{n}\right)^k \left(1 - \lambda \tfrac{t}{n}\right)^{n-k}$$

$\downarrow n \to \infty$

$$\lim_{n\to\infty} \frac{1}{k!} \underbrace{\frac{n \cdot (n-1) \cdot \ldots \cdot (n-(k+1))}{n^k}}_{\to 1} (\lambda t)^k \underbrace{\left(1 - \tfrac{\lambda t}{n}\right)^n}_{\to e^{-\lambda t}} \underbrace{\left(1 - \tfrac{\lambda t}{n}\right)^{-k}}_{\to 1}$$

$$= \frac{(\lambda t)^k e^{-\lambda t}}{k!}.$$

We call the process above a Poisson process, and for $X \sim \text{Poisson}(\lambda)$ (intensity parameter), define $P_X(x) = \frac{e^{-\lambda} \lambda^x}{x!}$ the Poisson distribution.

Let X be a random variable with $\mu_X = E[X]$. Let $k \in \mathbb{N} = \{1, 2, \ldots\}$ be positive.

We define the k^{th} moment of X to be $E[X^k]$. When $k=1$ this is expectation.

k^{th} central moment of X to be $E[(X-\mu_X)^k]$. When $k=2$ this is variance.

Since $Var(X) = E[X^2] - E[X]^2$ we see that calculating the 2^{nd} moment of X can be useful. In general, how can we compute these?

Let X be a random variable with a distribution function F_X. The moment generating function of F_X is defined by

$M(t) = E[e^{tX}]$ provided this exists for $t \in B(0, \varepsilon)$ for some $\varepsilon \in \mathbb{R}_{>0}$.

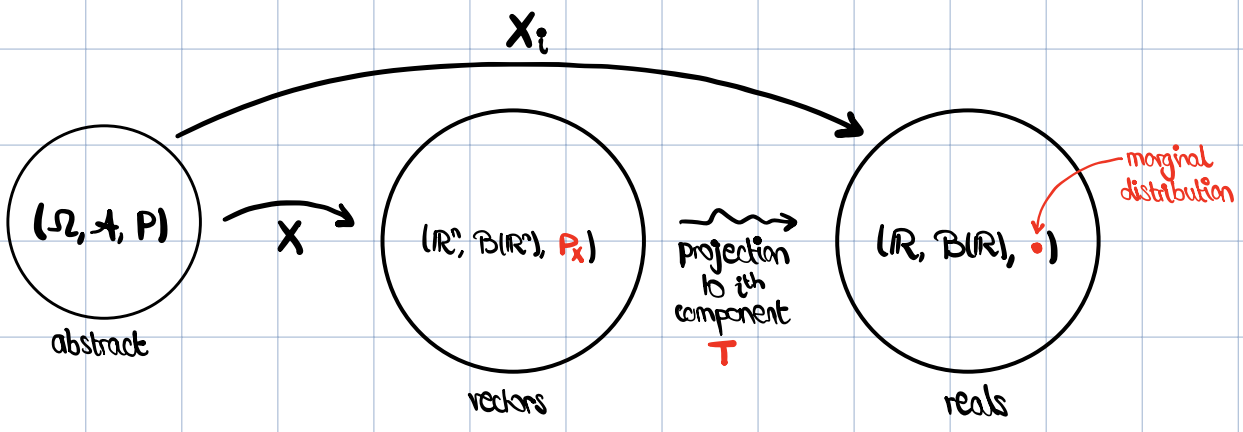
Theorem (Moment Generating Functions): Say X is a random variable with moment generating function $M(t)$.

Then $E[X^n] = M^{(n)}(0)$ (the n^{th} moment of X is the n^{th} derivative of $M(t)$ at $t=0$).

Let $(\Omega, \mathcal{A}, P)$ be a probability space, and $X: \Omega \to \mathbb{R}^n$ be a random variable (or a random vector).

We can write X in terms of component random variables, so $X(\omega) = \begin{pmatrix} X_1(\omega) \\ \vdots \\ X_n(\omega) \end{pmatrix}$ where $X_i: \Omega \to \mathbb{R}$.

If we know the distribution of X, what can we find out about the distribution of X_i?



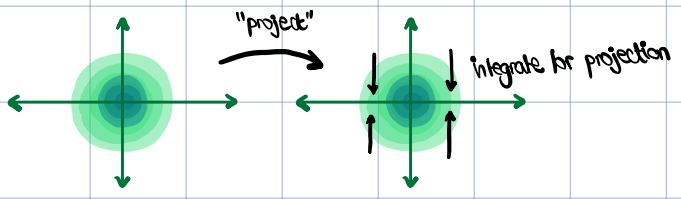
The function $P_{X_i} = (P_X)_T$ is called the marginal distribution of X with respect to component i.

We can calculate the marginal cumulative distribution function $F_{X_i}(t) = P_{X_i}((-\infty, t]) = P_X(\mathbb{R} \times \ldots \times (-\infty, t] \times \mathbb{R} \times \ldots \times \mathbb{R})$ (i^{th} component)

$= P_X(X_i \leq t, \ X_j \in \mathbb{R})$. ($j \neq i$, $j = 1, 2, \ldots, n$)

Continuous Case: P_X has a probability density function $f_X: \mathbb{R}^n \to \mathbb{R}_{\geq 0}$.

(or using contours)



Then the distribution $f_{X_i}(t)$ can be defined $f_{X_i}(t) = \int_{\mathbb{R}^{n-1}} f_X(x_1, \ldots, t, \ldots x_n)\, d(x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n)$. ($i^{th}$ component; all except i; $n-1$ integrations)

This is the marginal probability density function with respect to the i^{th} variable.

Discrete Case: P_X has a probability mass function $(p_x)_{x \in \mathbb{R}^n}$.

Then the marginal probability mass function $(p_t)_{t \in \mathbb{R}}$ is $p_t = \sum_{x_2, \ldots \in \mathbb{R}} p(t, x_2, \ldots, x_n)$.

Recall that, for a space $(\Omega, \mathcal{A}, P)$ and $B \in \mathcal{A}$, $P(B)>0$, we have an induced conditional space $(\Omega, \mathcal{A}, P(\cdot|B))$.

Notice a random variable $X:\Omega \to \mathbb{R}$ naturally lives on both spaces. However, its behaviour may change.

For X, we define $E[X] = \int_\Omega X \, dP \rightsquigarrow E[X|B] = \int_\Omega X \, dP(\cdot|B)$ the conditional expectation (given events) of X given B.

If we write $P(A|B) = \frac{P(A\cap B)}{P(B)} = \frac{1}{P(B)} \int_A \mathbb{1}_B \, dP$ then $E[X|B] = \frac{1}{P(B)} \int_\Omega X \mathbb{1}_B \, dP = \frac{1}{P(B)} E[\mathbb{1}_B X]$.

($\mathbb{1}_B$: indicator function; $\frac{1}{P(B)}$: scaling factor; $\mathbb{1}_B$: indicator; X: random variable)

Example: Let $X \sim N(0,1^2)$ so $f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}$. Say $B = \mathbb{R}_{>0} = \{x \in \mathbb{R} \mid x>0\}$.

Then $E[X|B] = \frac{1}{P(B)} \int_\Omega X(\omega) \mathbb{1}_B(\omega) \, dP(\omega) = \frac{1}{P(B)} \int_{\mathbb{R}} x \underbrace{\mathbb{1}_B(X^{-1}(x))}_{=\begin{cases}1 \text{ for } x>0,\\ 0 \text{ for } x\le 0.\end{cases}} f_X(x) \, dx$

$$= \frac{1}{P(B)} \int_0^\infty x f_X(x) \, dx = 2 \cdot \frac{1}{\sqrt{2\pi}} \int_0^\infty x e^{-\frac{1}{2}x^2} dx = 1.$$

Notice $E[\mathbb{1}_A|B] = \int_\Omega \mathbb{1}_A \, dP(\cdot|B) = \int_A dP(\cdot|B) = P(A|B)$ so we can write conditional probability as conditional expectation of indicator functions.

Example: Consider a container of N balls, M of which are red, and n are picked without replacement.

Say $X:\Omega \to \mathbb{R}$ is the random variable counting the number of red balls picked. What is $E[X]$?

(Use indicator functions, define $Y:\Omega \to \mathbb{R}^n$ a random variable and consider marginals).

Define $Y:\Omega \to \mathbb{R}^n$ by $Y((\omega_1,\dots,\omega_n)) = \begin{pmatrix} Y_1(\omega_1) \\ \vdots \\ Y_n(\omega_n)\end{pmatrix}$ where $Y_i(\omega_i) = \begin{cases}1 \text{ if } \omega_i \text{ red;}\\ 0 \text{ otherwise.}\end{cases}$ Notice $(Y_i)_{i=1}^n$ is independent. (is it?)

Therefore, by $X = \sum_{i=1}^n Y_i$, we have $E[X] = E[\sum_{i=1}^n Y_i] = \sum_{i=1}^n E[Y_i]$.

Consider the marginal $P_i(\omega_i) = \sum_{x=1}^n P(x_1,\dots,x_i,\dots,x_n)$. Notice $P_i(\omega_i) = P_j(\omega_j)$ for any i,j by this.

$\Rightarrow E[Y_i] = E[Y_j]$ for all pairs i,j. In particular, $\sum_{i=1}^n E[Y_i] = nE[Y_1] = n(P(Y_1=1)\cdot 1 + 0)$

$$= n\left(\frac{M}{N}\right).$$

What's $Var(X)$? Well, $Var(X) = E[X^2] - E[X]^2 = E[X^2] - \left(n^2 \frac{M^2}{N^2}\right)$.

We have $E[X^2] = E[(\sum_{i=1}^n Y_i)^2]$. By independence, $E[Y_iY_j] = E[Y_i]E[Y_j]$ for any $i \neq j$ (are they really independent?)

$$= E[\sum_{i=1}^n Y_i^2] + \sum_{i<j} E[Y_i]E[Y_j] = E[\sum_{i=1}^n Y_i^2] + \sum_{i<j} E[Y_1]^2 \quad \text{because } E[Y_i] = E[Y_j].$$

$$= \sum_{i=1}^n E[Y_i^2] + \binom{n}{2} E[Y_1]^2.$$

Notice $E[Y_i^2] = \sum_{x=0}^1 x^2 P(Y_i = x) = 1^2 P(Y_i = 1) = \frac{M}{N}$.

Say $X_1 \sim \text{Poisson}(\mu_1)$ and $X_2 \sim \text{Poisson}(\mu_2)$ are independent random variables. Find the conditional probability mass function of X_1 given $X_1 + X_2 = t$ for some $t \in \mathbb{N}$.

We have $P[X_1 = x_1 | X_1 + X_2 = t] = \frac{P[X_1 = x_1,\ X_1+X_2 = t]}{P[X_1+X_2=t]} = \frac{P[X_1 = x_1,\ X_2 = t-x_1]}{P[X_1+X_2=t]} \overset{\text{independence}}{=} \frac{P[X_1=x_1]\,P[X_2=t-x_1]}{P[X_1+X_2=t]}$ (definition)

$$\Rightarrow P[X_1 = x_1 | X_1+X_2=t] = \left(\frac{e^{-\mu_1}\mu_1^{x_1}}{x_1!}\right)\left(\frac{e^{-\mu_2}\mu_2^{t-x_1}}{(t-x_1)!}\right)\left(P[X_1+X_2=t]\right).$$

$$= \frac{\left(\frac{e^{-\mu_1}\mu_1^{x_1}}{x_1!}\right)\left(\frac{e^{-\mu_2}\mu_2^{t-x_1}}{(t-x_1)!}\right)}{\text{Poisson}(\mu_1+\mu_2)}$$

$$= \sum_{x_1=0}^t P[X_1 = x_1,\ X_2 = t-x_1]$$

$$\overset{\text{independence}}{=} \sum_{x_1=0}^t P[X_1=x_1]\,P[X_2=t-x_1].$$

$$= \sum_{x_1=0}^t \frac{e^{-\mu_1}\mu_1^{x_1}}{x_1!}\frac{e^{-\mu_2}\mu_2^{t-x_1}}{(t-x_1)!}$$

$$= e^{-(\mu_1+\mu_2)}\mu_2^t \sum^n \frac{1}{x_1!(t-x_1)!}\left(\frac{\mu_1}{\mu_2}\right)^{x_1}$$

$$= \binom{t}{x_1}\left(\frac{\mu_1}{\mu_1+\mu_2}\right)^{x_1}\left(1-\frac{\mu_1}{\mu_1+\mu_2}\right)^{t-x_1}$$

$$= \text{Binomial}\left(t, \frac{\mu_1}{\mu_1+\mu_2}\right).$$

Therefore $P[X_1=x_1 \mid X_1+X_2=t] = \text{Binomial}\left(t, \frac{\mu_1}{\mu_1+\mu_2}\right)$.

$$= \frac{e^{-(\mu_1+\mu_2)}}{t!}\mu_2^t \underbrace{\sum_{x_1=0}^{n} \frac{t!}{x_1!(t-x_1)!}\left(\frac{\mu_1}{\mu_2}\right)^{x_1}1^{t-x_1}}_{\text{binomial}}$$

$$= \frac{e^{-(\mu_1+\mu_2)}}{t!}\mu_2^t\left(1+\frac{\mu_1}{\mu_2}\right)^t$$

$$= \frac{e^{-(\mu_1+\mu_2)}}{t!}(\mu_1+\mu_2)^t$$

$$= \text{Poisson}(\mu_1+\mu_2).$$

People arrive at a house from 5:30pm-9:00pm according to a Poisson process, averaging 12 people per hour.

What number of people is most likely to arrive?

We want $\underset{x\in\mathbb{N}}{\text{argmax}}\, P(M=x)$ where M is the total number of people who arrive. Notice $M \sim \text{Poisson}(\lambda)$ ($\lambda=42$ intensity parameter)

$$= \underset{x\in\mathbb{N}}{\text{argmax}}\, \frac{e^{-\lambda}\lambda^x}{x!}$$ but we cannot compute this. How can we proceed?

↝ Consider $\frac{P(M=x)}{P(M=x+1)}$ and see how this behaves.

Suppose $X \sim \text{Hypergeometric}(N, M, n)$. Find $E[X]$.

We have $P_X(x) = \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}}$ so $E[X] = \sum_{x=0}^{n} x P_X(x) = \sum_{x=0}^{n} x \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}}$.

$$\sum_{x=0}^{n} x \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}} = \sum_{x=1}^{n} x \frac{M!}{x!(M-x)!}\frac{(N-M)!}{(n-x)!((N-M)-(n-x))!}\cdot\frac{n!(N-n)!}{N!}$$

$$= \frac{Mn}{N}\sum_{x=1}^{n}\frac{(M-1)!}{(x-1)!((M-1)-(x-1))!}\frac{(N-M)!}{(n-x)!((N-M)-(n-x))!}\frac{(n-1)!((N-1)-(n-1))!}{(N-1)!}$$

($(n-x)!((N-M)-(n-x))! \to (n-1)-(x-1)$)

$$= \frac{Mn}{N}\sum_{x^*=0}^{n^*}\frac{M^*}{x^*!(M^*-x^*)!}\frac{(N-M)!}{(n^*-x^*)!((N-M)-(n^*-x^*))!}\frac{n^*!(N^*-n^*)!}{N^*!}$$

$$= \frac{Mn}{N}\,1 = \frac{Mn}{N}.$$

Consider n Bernoulli trials, and let $X_i : \Omega \to \mathbb{R}$ be the indicator that the i^{th} trial is a success.

In other words, define $X : \Omega \to \mathbb{R}^n$ by $X((\omega_1,\dots,\omega_n)) = \begin{pmatrix} X_1(\omega_1) \\ \vdots \\ X_n(\omega_n)\end{pmatrix}$ and notice all the X_i are independent.

Let $Y \sim \text{Bernoulli}(n, p)$ counting the number of successes, so $Y = \sum_{i=1}^{n} X_i$.

Then $\text{Var}(Y) = \text{Var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n}\text{Var}(X_i) + 2\sum_{i<j}\text{Cov}(X_i, X_j)$ but since $X_i \perp X_j$, we have $\text{Cov}(X_i, X_j) = 0$.

$$= \sum_{i=1}^{n}\text{Var}(X_i)$$ but each X_i is "identical" so $\text{Var}(X_i) = \text{Var}(X_j)$

$$= n\text{Var}(X_1) = n(E[X_1^2] - E[X_1]^2)$$

$$= n\left(\sum_{x=0}^{1} x^2 P_{X_1}(x) - p^2\right) = n\left(1^2 P(X_1=1) - p^2\right) = n(p-p^2) = np(1-p).$$

Say $X \sim \text{Hypegeometric}(N, M, n)$. What's $\text{Var}(X)$?

Well, we have X counting the number of successes from n picks out of N samples.

Consider $Y : \Omega \to \mathbb{R}^n$ given by $Y\begin{pmatrix}\omega_1 \\ \vdots \\ \omega_n\end{pmatrix} = \begin{pmatrix} Y_1(\omega_1) \\ \vdots \\ Y_n(\omega_n)\end{pmatrix}$ where $Y_i : \Omega \to \mathbb{R}$ are indicator random variables.

Notice $X = \sum Y_i$ so $\text{Var}(X) = \text{Var}\left(\sum Y_i\right) = \sum \text{Var}(X_i) + 2\sum \text{Cov}(X_i, X_j)$.

We have $E(X_i) = \sum_{x=0}^{1} x P_{X_i}(x) = P(X_i = 1) = \frac{M}{N}$. Notice this is invariant of $i = 1, \dots, n$.

Therefore $\mathrm{Var}(X_i) = E[X_i^2] - E[X_i]^2 = \sum_{x=0}^{1} x^2 P_{X_i}(x) - \frac{M^2}{N^2} = 1^2 \frac{M}{N} = \frac{M}{N} - \frac{M^2}{N^2}$

We also have $\mathrm{Cov}(X_i, X_j) = E[X_i X_j] - E[X_i]E[X_j]$

$$= \underbrace{\sum_{x_1=0}^{1} \sum_{x_2=0}^{1} x_1 x_2 P(X_i = x_1, X_j = x_2)}_{\text{↳ } 1 \cdot 1 \, P(X_i = 1, X_j = 1) \text{ but this is just } P(X_2 = 1 \mid X_1 = 1) P(X_1 = 1) = \frac{M-1}{N-1}\left(\frac{M}{N}\right)} - \left(\frac{M}{N}\right)^2$$

$$\Rightarrow \mathrm{Cov}(X_i, X_j) = \frac{M-1}{N-1}\left(\frac{M}{N}\right) - \left(\frac{M}{N}\right)^2$$

$$\Rightarrow \mathrm{Var}(X) = \sum_{i=1}^{n} \mathrm{Var}(X_i) + 2\sum_{i<j} \mathrm{Cov}(X_i, X_j) = n\left(\frac{M}{N} - \frac{M^2}{N^2}\right) + 2\binom{n}{2}\left(\frac{M-1}{N-1}\left(\frac{M}{N}\right) - \left(\frac{M}{N}\right)^2\right).$$

Theorem (Multivariate LOTUS): Say X_1 and X_2 are random variables $X_1, X_2 : \Omega \to \mathbb{R}$ such that $(X_1, X_2) \sim P(x_1, x_2)$.

Let $g : \mathbb{R}^2 \to \mathbb{R}$ be measurable. Then $E[g(X_1, X_2)] = \sum_{x_1} \sum_{x_2} g(x_1, x_2) P(X_1 = x_1, X_2 = x_2)$.

Proof: Recall $P(X_1 \in A, X_2 \in B)$ is defined by $P(X_1^{-1}(A) \cap X_2^{-1}(B)) = P_{X_1, X_2}(A, B)$ the joint (bivariate) distribution.

Notice $(X_1, X_2) : \Omega \to \mathbb{R}^2$ defines a random variable pair. Define a new random variable by

$\tilde{Z} = g((X_1, X_2)) : \Omega \xrightarrow{(X_1, X_2)} \mathbb{R}^2 \xrightarrow{g} \mathbb{R}$. Then notice $\tilde{Z}^{-1}(A) = (X_1, X_2)^{-1}(g^{-1}(A))$ for $A \in \mathcal{B}(\mathbb{R})$.

We have $E[\tilde{Z}] = \sum_{z} z P_{\tilde{Z}}(z)$

$$= \sum_{z} z P(\tilde{Z}^{-1}(\{z\})) = \sum_{z} z P((X_1, X_2)^{-1}(g^{-1}(\{z\})))$$

and let $S \subseteq \mathbb{R}^2$ be defined by $S = g^{-1}(\{z\}) \Rightarrow z = g(S)$.

$$= \sum_{S} g(S) P((X_1, X_2)^{-1}(S))$$

$$= \sum_{x_1} \sum_{x_2} g(x_1, x_2) P(\underbrace{(X_1, X_2)^{-1}((x_1, x_2))}_{X_1^{-1}(\{x_1\}) \cap X_2^{-1}(\{x_2\})})$$

$$= \sum_{x_1} \sum_{x_2} g(x_1, x_2) P(X_1 = x_1, X_2 = x_2) \quad \text{as required.} \quad \square$$

Some of the examples above showed the usefulness of indicator functions in computing expectation and variance.

Let's look at some more.

Consider a chain of N beads composed of two colours: A (probability p) and B (probability $1-p$). Let $X : \Omega \to \mathbb{R}$ count the number of "alternates".

Alternates are defined by a sequence "AB" or "BA", so "ABAB" has 3 alternates. Find $E[X]$ and $\mathrm{Var}(X)$.

Define a random variable $Y : \Omega \to \mathbb{R}^N$ by $Y((w_1, \dots, w_N)) = (Y_1(w_1, w_2), \dots, Y_{N-1}(w_{N-1}, w_N), 0)$.

Each Y_i is an indicator random variable $Y_i(w_i, w_{i+1}) = \begin{cases} 1 & \text{if } w_i \neq w_{i+1} \text{ (alternates)} \\ 0 & \text{otherwise} \end{cases}$.

Notice $X = \sum_{i=1}^{N} Y_i$ so $E[X] = E[\sum_{i=1}^{N} Y_i] \overset{\text{linearity}}{=} \sum_{i=1}^{N} E[Y_i] = \sum_{i=1}^{N-1} E[Y_i] + \underbrace{E[Y_N]}_{0}$.

We have $E[Y_i] = \sum_{x} x P(Y_i = x) = \sum_{x=0}^{1} x P[Y_i = x] = P[Y_i = 1]$

where $P[Y_i] = P[Y_j]$ for all i, j, and $P[Y_i = 1] = P[W_{i+1} = A \mid W_i = B] P[W_i = B] + P[W_{i+1} = B \mid W_i = A] P[W_i = A] = p(1-p) + (1-p)p$.

$$\Rightarrow E[X] = \sum_{i=1}^{N-1} E[Y_i] = (N-1)(2p(1-p))$$

We have $\mathrm{Var}(X) = \mathrm{Var}(\sum_{i=1}^{N} Y_i) = \sum_{i=1}^{N} \mathrm{Var}(Y_i) + 2\sum_{i<j} \mathrm{Cov}(Y_i, Y_j)$.

Here, we have $\mathrm{Var}(Y_i) = E[Y_i^2] - E[Y_i]^2 \overset{\text{LOTUS}}{=} \sum_{x=0}^{1} x^2 P[Y_i = x] - (2p(1-p))^2 = P[Y_i = 1] - (2p(1-p))^2$

$$= 2p(1-p) - (2p(1-p))^2.$$

and $\mathrm{Cov}(Y_i, Y_j) = E[Y_iY_j] - E[Y_i]E[Y_j]$. Notice that $E[Y_iY_j] = \sum_x \sum_y xy\, P[Y_i=x, Y_j=y]$

$= P(Y_i=1, Y_j=1)$

$= P(w_i \neq w_{i+1},\ w_j \neq w_{j+1})$ ← independent if $i+1 \neq j$, otherwise dependent

Some of the above examples rely on the following notion.

Consider discrete random variables $X, Y : \Omega \to \mathbb{R}$. Recall for an event $B \in \mathcal{A}$, we had $E(X|B) = \frac{1}{P(B)} E[\mathbb{1}_B(X)]$. ← indicator function

Similarly, we have $E[X|Y=y] = \sum_x x P(X=x|Y=y)$

$= \sum_x x \frac{P(X=x, Y=y)}{P(Y=y)}$. ← joint probability mass function; ← some function of y

Notice $E[X|Y=y] = f(y) : \mathbb{R} \to \mathbb{R}$ and $Y : \Omega \to \mathbb{R}$ so $f(Y) : \Omega \xrightarrow{Y} \mathbb{R} \xrightarrow{f} \mathbb{R}$ defines a new random variable.

Then $f(Y)$ is called the conditional expectation of X given Y denoted $E[X|Y]$.

Remark: Notice $E[X|Y]$ is a random variable, not a value.

Example: Consider a die throw, and a random variable $X : \Omega \to \mathbb{R}$ checking if the roll is even, so $X(\omega) = \begin{cases} 1 \text{ if } \omega \in \{2,4,6\}, \\ 0 \text{ otherwise.} \end{cases}$

$Y : \Omega \to \mathbb{R}$ checking if we rolled a six, so $Y(\omega) = \begin{cases} 1 \text{ if } \omega = 6, \\ 0 \text{ otherwise.} \end{cases}$

Then $E[X|Y] : \Omega \to \mathbb{R}$ is a random variable and $E[X|Y](\omega) = \begin{cases} E[X|Y=0], & \omega \in \{1,\dots,5\}; \\ E[X|Y=1], & \omega \in \{6\}. \end{cases}$

Midterm Practice:

1a) $P[X|Y=1] = \frac{P[X, Y=1]}{P[Y=1]}$ where $P[Y=1] = P[X=1, Y=1] + P[X=2, Y=1]$.
$= 2/8 = 1/4$

For $x=1$: $f(1|1)$ $\frac{P[X=1, Y=1]}{P[Y=1]} = 4 \cdot \frac{1}{8}$ and for $x=2$: $f(2|1)$ $\frac{P[X=2, Y=1]}{P[Y=1]} = 4 \cdot \frac{1}{8}$.

$P[X|Y=2] = \frac{P[X, Y=1]}{P[Y=1]}$ where $P[Y=2] = 1/2 + 1/4 = 3/4$

For $x=1$: $f(1|2)$ $\frac{P[X=1, Y=2]}{P[Y=2]} = \frac{4}{3} \cdot \frac{1}{4} = \frac{1}{3}$ and for $x=2$: $f(2|2)$ $\frac{P[X=2, Y=2]}{P[Y=2]} = \frac{4}{3} \cdot \frac{1}{2} = \frac{2}{3}$.

b) X and Y are not independent because $f(2|1) = \frac{1}{2} \neq f(2) = \frac{5}{8}$.

c) $P[XY \leq 3]$. Define $S = XY$ so $S \in \{1, 2, 4\}$. Notice $P[S \leq 3] = 1 - P[S=4]$

$= 1 - P[X=2, Y=2] = 1 - \frac{1}{2} = \frac{1}{2}$

$P[X/Y > 1]$. Define $S = X/Y$ so $S \in \{1/2, 1, 2\}$. We have $P[S>1] = P[S=2] = P[X=2, Y=1] = \frac{1}{8}$.

$P[X+Y>2] = 1 - P[X+Y \leq 2] = 1 - P[X=1, Y=1] = 1 - \frac{1}{8} = \frac{7}{8}$.

3 a) Roll a fair die 100 times independently. Denote X_i = the number of i's, for $i \in \{1,\dots,6\}$.

Find the p.m.f. of $T_o = X_1 + X_3 + X_5$.

We have T_o the count of odd die rolls, so each roll has $p_{odd} = \frac{1}{2}$.

Thus we have $T_o \sim \text{Binomial}(n, p_{odd}) = \text{Binomial}(100, \frac{1}{2})$ so $P_{T_o}(x) = \binom{100}{x}(\frac{1}{2})^x(\frac{1}{2})^{100-x}$

b) Find the p.m.f. of X_5 given $T_o = t_o$.

We have to find $P[X_5 \mid X_1 + X_3 + X_5 = t_o]$. Consider $S = X_1 + X_3$ so $X_5 + S = t_o$.

so $S \sim \text{Binomial}(n, \dots)$.

$\Rightarrow$ find $P[X_5 \mid X_5 + S = t_o]$.

Continuous Distributions

We have developed a theory of probability above. We will now introduce several useful distributions.

1 Exponential: $X \sim \text{Exp}(\lambda)$ with $f(x) = \lambda e^{-\lambda x}$.

a) Show that X follows cdf $F(x) = \begin{cases} 1-e^{-\lambda x} & \text{for } x>0; \\ 0 & \text{for } x \le 0. \end{cases}$

- models lifetime and/or decay.
- models "wait time" by considering λ to be a Poisson parameter (so X represents waiting time between events).

so $F_X(x) = P[X \le x] = 1 - P[X > x] = 1 - P[\underbrace{\text{no events in } [0,x)}_{Y \sim \text{Poisson}(\lambda x)}] = 1 - \frac{e^{-\lambda x}(\lambda x)^0}{0!} = 1 - e^{-\lambda x}$ for all $x > 0$.

- "memoryless" property: $P[X > t+s \mid X > s] = P[X > t]$ (probability of a "leap" only depends on "leap" size, not its location).

★ Exercise

2 Gamma function: $\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x}\,dx$, and $X \sim \text{Gamma}(\alpha, \beta)$ with $f(x) = \frac{1}{\beta^\alpha \Gamma(\alpha)} x^{\alpha-1} e^{-x/\beta}$.

$\Gamma(n+1) = n!$ for $n \in \mathbb{N}$. $n\Gamma(n) = \Gamma(n+1)$. ★ Exercise (integration by parts)

$\frac{1}{\beta^\alpha \Gamma(\alpha)}$: normalization constant

- extension of $\text{Exp}(\lambda)$. When $\alpha = 1$, we have $\text{Gamma}(1, \beta) = \text{Exp}(\frac{1}{\beta}) \Rightarrow \text{Gamma}(1, \frac{1}{\lambda}) = \text{Exp}(\lambda)$.

3 Uniform distribution.

4 β-distribution: $\beta(\alpha_1, \alpha_2) = \int_0^1 x^{\alpha_1 - 1}(1-x)^{\alpha_2 - 1}\,dx = \frac{\Gamma(\alpha_1)\Gamma(\alpha_2)}{\Gamma(\alpha_1+\alpha_2)}$. If $X \sim \text{Beta}(\alpha_1, \alpha_2)$ then

$f(x) = \frac{1}{\beta(\alpha_1,\alpha_2)} x^{\alpha_1-1}(1-x)^{\alpha_2-1}$ for $x \in [0,1]$.

5 Normal distribution: $X \sim N(0,1)$ with $f(x) = \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}}$ (standard normal).

- if $X \sim N(0,1)$ then $F_X(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}}\,dx$ (no closed form).

6 (General) Normal distribution $Z \sim N(\mu, \sigma^2)$ with $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}$ (general normal).

- if $X \sim N(\mu, \sigma^2)$ then $F_X(x) = \int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}\,dx$ (no closed form).

Proposition (Uniformity): Let X be a random variable such that $F_X(x)$ is strictly increasing (so F_X^{-1} exists).

Let $Y \sim \text{Unif}(0,1)$. Then $F_X^{-1}(Y) \overset{\text{cdf}}{\sim} F_X(x)$.

Proof (rough): We have $F_X(x) = P[X \leq x]$. Define $Y \sim \text{Unif}(0,1)$ so $X = F_X^{-1}(Y)$.

Then $P[X \leq x] = P[F_X^{-1}(Y) \leq x] = P[Y \leq \underbrace{F_X(x)}_{\in[0,1] \text{ since c.d.f.}}] = F_X^{-1}(Y)$. $\square$

Proposition ($N(0,1)$): Let $X \sim N(0,1)$. Then $E[X] = 0$ and $\text{Var}(X) = 1$.

Proof: By definitions. We have $E[X] = \int_{-\infty}^{\infty} x f(x)\,dx = \int_{-\infty}^{\infty} x \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}\,dx = \frac{1}{\sqrt{2\pi}}\left(\int_{-\infty}^{0} x e^{-\frac{x^2}{2}}\,dx + \int_{0}^{\infty} x e^{-\frac{x^2}{2}}\,dx\right) = 0$. (0 by symmetry)

and $\text{Var}(X) = \int_{-\infty}^{\infty} x^2 f(x)\,dx - \mu^2 = \int_{-\infty}^{\infty} x^2 \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}\,dx$

$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} x^2 e^{-\frac{x^2}{2}}\,dx \overset{*}{=} \frac{1}{\sqrt{2\pi}} \sqrt{2\pi} = 1$. $\square$ (* compute)

Proposition (Reparametrization Trick): Let $X \sim N(\mu, \sigma^2)$. Then $Z = \frac{X-\mu}{\sigma} \sim N(0,1)$.

Proof: Notice $\frac{x-\mu}{\sigma}$ defines a bijection (in particular, a homeomorphism), so the inverse transform exists and is unique.

We have $Z = \frac{X-\mu}{\sigma} \Rightarrow X = \sigma Z + \mu$ so $F_Z(z) = P(Z \leq z)$

$= P(\frac{X-\mu}{\sigma} \leq z) \overset{\text{inverse transform}}{=} P(X \leq \sigma z + \mu)$.

Therefore $f_Z(z) = \frac{d}{dz} F_Z(z) = \frac{d}{dz} P[X \leq \sigma z + \mu] = \frac{d}{dz} F_X(\sigma z + \mu) = f_X(\sigma z + \mu) \frac{d}{dz}(\sigma z + \mu)$

$\overset{\text{how? } *}{=} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{\sigma z + \mu - \mu}{\sigma}\right)^2} \cdot \sigma$

$= \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}$ a standard normal. $\square$

Corollary ($N(\mu, \sigma^2)$): Let $X \sim N(\mu, \sigma^2)$. Then $E[X] = \mu$ and $\text{Var}(X) = \sigma^2$.

Proof: Let $Y \sim N(0,1)$ and consider $X = \sigma Y + \mu$ by the reparametrization trick.

Then $E[X] = E[\sigma Y + \mu] = \sigma \cancel{E[Y]}^{0} + \mu = \mu$.

$\text{Var}(X) = \text{Var}(\sigma Y + \mu) = \sigma^2 \cancel{\text{Var}(Y)}^{1} = \sigma^2$. $\square$

Gamma Function Technique

Recall that if $X \sim \text{Gamma}(\alpha, \beta)$ where $\alpha, \beta > 0$, then $f_X(x) = \frac{1}{\beta^\alpha \Gamma(\alpha)} x^{\alpha-1} e^{-x/\beta}$ for $x \in [0, \infty)$.

Since f_X defines a probability density, we have

$\int_0^\infty \frac{1}{\beta^\alpha \Gamma(\alpha)} x^{\alpha-1} e^{-x/\beta}\,dx = 1 \Rightarrow \int_0^\infty x^{\alpha-1} e^{-x/\beta}\,dx = \beta^\alpha \Gamma(\alpha)$. This defines the Gamma function technique.

Corollary: Let $X \sim \text{Gamma}(\alpha, \beta)$. Then $E[X^p] = \beta^p \frac{\Gamma(p+\alpha)}{\Gamma(\alpha)}$.

Proof: We use the Γ-technique.

$E[X^p] = \int_0^\infty x^p \frac{1}{\beta^\alpha \Gamma(\alpha)} x^{\alpha-1} e^{-x/\beta}\,dx = \frac{1}{\beta^\alpha \Gamma(\alpha)} \int_0^\infty x^{p+\alpha-1} e^{-x/\beta}\,dx = \frac{1}{\beta^\alpha \Gamma(\alpha)} \beta^{p+\alpha} \Gamma(p+\alpha) = \frac{\beta^p \Gamma(p+\alpha)}{\Gamma(\alpha)}$. $\square$

Corollary: Let $X \sim \text{Gamma}(\alpha, \beta)$. Then $\text{Var}(X) = \alpha\beta^2$.

Proof: $\text{Var}(X) = E[X^2] - E[X]^2 = \beta^2 \frac{\Gamma(2+\alpha)}{\Gamma(\alpha)} - \left(\beta \frac{\Gamma(1+\alpha)}{\Gamma(\alpha)}\right)^2 = \beta^2(\alpha+1)(\alpha) - (\alpha\beta)^2 = \alpha\beta^2 + \alpha^2\beta^2 - \alpha^2\beta^2 = \alpha\beta^2$. □

* **Exercise:** Let $Z \sim N(\mu, \sigma^2)$. Use the reparametrization trick and Γ-technique to find $E[X]$ and $\text{Var}(X)$.
* **Exercise:** Use the Γ-technique to find $E[X]$ and $\text{Var}(X)$ where: 1) $X \sim \text{Unif}(a,b)$ 2) $X \sim \text{Beta}(a,b)$.

Moment Generating Functions

Previously, we defined the notion of a moment generating function for a random variable X.

We now develop a theory of moment generating functions.

* **Exercise:** Let $X \sim \text{Bin}(n,p)$. Find $M_X(t)$ and find the values of t such that $M_X(t)$ is well-defined.

Exercise: Let $X \sim \text{Gamma}(\alpha, \beta)$. Find $M_X(t)$ and find the values of t such that $M_X(t)$ is well-defined.

We have $M_X(t) = E[e^{tX}] = \int_0^\infty e^{tx} \frac{1}{\beta^\alpha \Gamma(\alpha)} x^{\alpha-1} e^{-x/\beta} dx = \frac{1}{\beta^\alpha \Gamma(\alpha)} \int_0^\infty x^{\alpha-1} e^{-x/\beta + tx} dx$ (let $k = \frac{1}{\beta^\alpha \Gamma(\alpha)}$)

$= k \int_0^\infty x^{\alpha-1} e^{-x(1/\beta - t)} dx$

$= k \int_0^\infty x^{\alpha-1} \exp\left(-\frac{x}{(\frac{1}{\beta}-t)^{-1}}\right) dx$. Define $\tilde{\beta} = (\frac{1}{\beta} - t)^{-1}$, so

$k \int_0^\infty x^{\alpha-1} \exp\left(-\frac{x}{(\frac{1}{\beta}-t)^{-1}}\right) dx = k \int_0^\infty x^{\alpha-1} e^{-x/\tilde{\beta}} dx$

$= k \tilde{\beta}^\alpha \Gamma(\alpha) = \frac{1}{\beta^\alpha \Gamma(\alpha)} (\frac{1}{\beta} - t)^{-\alpha} \Gamma(\alpha)$ ← Γ-technique on $\tilde{\beta}$

$= \beta^{-\alpha}(\beta^{-1} - t)^{-\alpha} = (\beta\beta^{-1} - \beta t)^{-\alpha} = (1 - \beta t)^{-\alpha}$.

We used the Γ-technique above, which restricts our values of t, since we require $\alpha > 0$ and $\tilde{\beta} > 0 \Rightarrow (\frac{1}{\beta} - t) > 0 \Rightarrow t < \frac{1}{\beta}$.

Would we get weaker constraints if we did not use the Γ function? We verify that our bounds for t are necessary.

Say $t \geq \frac{1}{\beta}$. If $t = \frac{1}{\beta}$, then $E[e^{tX}] = \int_0^\infty e^{tx} k x^{\alpha-1} e^{-x/\beta} dx$

$= k \int_0^\infty x^{\alpha-1} e^{-x/\beta + x/\beta} dx = k \int_0^\infty x^{\alpha-1} dx \to \infty$.

If $t > \frac{1}{\beta}$. Then $E[e^{tX}] = \int e^{tx} k x^{\alpha-1} e^{-x/\beta} dx = k \int_0^\infty x^{\alpha-1} e^{x(t - \frac{1}{\beta})} dx$

$\geq \underbrace{\min_{x \in [0,\infty]} e^{x(t-\frac{1}{\beta})}}_{1} \int_0^\infty x^{\alpha-1} dx \to \infty$.

Therefore $M_X(t) = (1-\beta t)^{-\alpha}$ for $t < \frac{1}{\beta}$ and since $\beta > 0$ is finite, we can find $\varepsilon > 0$ such that $M_X(t)$ is continuous on $B(0, \varepsilon) \Rightarrow E[X^n] = M_X^{(n)}(0)$ is well-defined.

Exercise: Let $X \sim N(0,1)$. Find $M_X(t)$.

We have $M_X(t) = \int_{-\infty}^\infty e^{tx} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^\infty e^{-\frac{1}{2}(x^2 - 2tx)} dx = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^\infty e^{-\frac{1}{2}(x-t)^2} e^{\frac{1}{2}t^2} dx = e^{\frac{t^2}{2}} \underbrace{\int_{-\infty}^\infty \frac{1}{\sqrt{2\pi}} e^{\frac{1}{2}(x-t)^2} dx}_{1} = e^{\frac{t^2}{2}}$.

since p.d.f. of $N(t,1)$

Notice the argument works for any $t \in \mathbb{R} \Rightarrow M_X(t)$ well-defined for all $t \in \mathbb{R}$.

Proposition (MGF of a linear function): Let X be a random variable and let $Y = aX + b$ for $a, b \in \mathbb{R}$.

Say $M_X(t)$ exists in $B(0, \varepsilon)$. Then $M_Y(t) = e^{bt} M_X(at)$ for $t \in B(0, \frac{\varepsilon}{|a|})$.

Proof: Follow the computation. $M_Y(t) = E[e^{tY}] = E[e^{t(aX+b)}] = E[e^{bt} e^{atX}] \overset{\text{homogeneity}}{=} e^{bt} E[e^{atX}] = e^{bt} M_X(at)$.

Then $M_Y(t)$ exists if $M_X(at)$ exists $\Rightarrow at \in B(0, \varepsilon) \Rightarrow t \in B(0, \frac{\varepsilon}{|a|})$. $\square$

Example: Say $X \sim N(\mu, \sigma^2)$. Find $M_X(t)$ using the reparametrization trick.

Notice $X = \sigma Z + \mu$ where $Z \sim N(0,1)$. By the above Proposition, we have $M_Z(t) = e^{-\frac{t^2}{2}} \Rightarrow M_X(t) = e^{t\mu} M_Z(\sigma t)$

$= e^{t\mu} e^{\frac{1}{2}(\sigma t)^2}$ for $t \in \mathbb{R}$.

Exercise: Say $X \sim \text{Gamma}(\alpha, \beta)$. Let $Y = X/\beta$. Find $M_Y(t)$ and state the values of t for which this is well-defined.

Let X and Y be random variables. We define the joint moment generating function $M_{(X,Y)}(t_1, t_2) = E[e^{(t_1, t_2)\binom{X}{Y}}] = E[e^{t_1 X + t_2 Y}]$. (vector multiplication)

Given a joint MGF $M_{(X,Y)}(t_1, t_2)$, how can we marginalize to find $M_X(t_1)$ or $M_Y(t_2)$?

Corollary: If $M_{(X,Y)}(t_1, t_2)$ is a joint MGF, then $M_X(t_1) = M_{(X,Y)}(t_1, 0)$ and $M_Y(t_2) = M_{(X,Y)}(0, t_2)$.

Proof: Obvious. Since $M_{(X,Y)} : \mathbb{R}^2 \to \mathbb{R}$ well-defined in $B(0, \varepsilon) \subseteq \mathbb{R}^2$, we have $M_{(X,Y)}(t_1, 0)$ and $M_{(X,Y)}(0, t_2)$ well-defined. $\square$

The definition of a joint MGF (and the above Corollary) generalize nicely to an arbitrary family $(X_i)_{i \in \Lambda}$ of random variables.

Theorem (Joint MGFs): Let X, Y be random variables with joint MGF $M_{(X,Y)}(t_1, t_2)$. Then $E[X^j Y^k] = \left. \frac{\partial^{j+k}}{\partial t_1^j \partial t_2^k} M_{(X,Y)}(t_1, t_2) \right|_{(t_1, t_2) = 0}$.

Proof: X, Y discrete. Then $M_{X,Y}(t_1, t_2) = E[e^{t_1 X + t_2 Y}] = \sum_x \sum_y e^{t_1 x + t_2 y} f(x, y)$

$$\Rightarrow \frac{\partial^{j+k}}{\partial t_1^j \partial t_2^k} M_{X,Y}(t_1, t_2) = \frac{\partial^{j+k}}{\partial t_1^j \partial t_2^k} \sum_x \sum_y e^{t_1 x + t_2 y} f(x, y) \overset{\text{finite sums commute}}{=} \sum_x \sum_y \frac{\partial^{j+k}}{\partial t_1^j \partial t_2^k} e^{t_1 x + t_2 y} f(x, y)$$

($f(x,y)$: independent of t_1 and t_2)

$$= \sum_x \sum_y x^j e^{t_1 x} y^k e^{t_2 y} f(x, y)$$

$$\overset{t_1 = 0}{\underset{t_2 = 0}{=}} \sum_x \sum_y x^j y^k f(x, y) = E[X^j Y^k]. \text{ Done. } \Delta$$

($= E[X^j Y^k]$: multivariate LOTUS)

X, Y continuous. Then $M_{X,Y}(t_1, t_2) = E[e^{t_1 X + t_2 Y}] = \iint_D e^{t_1 x} e^{t_2 y} f(x, y)\, dx\, dy$

$$\frac{\partial^{j+k}}{\partial t_1^j \partial t_2^k} M_{X,Y}(t_1, t_2) = \frac{\partial^{j+k}}{\partial t_1^j \partial t_2^k} \iint_D e^{t_1 x} e^{t_2 y} f(x, y)\, dx\, dy \overset{*}{=} \iint_D \frac{\partial^{j+k}}{\partial t_1^j \partial t_2^k} e^{t_1 x} e^{t_2 y} f(x, y)\, dx\, dy$$

$$\overset{t_1 = 0}{\underset{t_2 = 0}{=}} \iint_D x^j y^k f(x, y)\, dx\, dy = E[X^j Y^k]. \quad \square$$

Corollary: Let X, Y be random variables. Then $E[XY] = \left. \frac{\partial^2}{\partial t_1 \partial t_2} M_{X,Y}(t_1, t_2) \right|_{(t_1, t_2) = 0}$.

Theorem (MGF Independence): X and Y are independent $\Leftrightarrow M_{X,Y}(t_1,t_2) = M_X(t_1)M_Y(t_2)$ for all $(t_1,t_2) \in B(0,\varepsilon) \subseteq \mathbb{R}^2$.

Proof ($\Rightarrow$): Let X and Y be independent. Then $M_{X,Y}(t_1,t_2) = E[e^{t_1X}e^{t_2Y}] \overset{\text{independence}}{=} E[e^{t_1X}]\,E[e^{t_2Y}] = M_X(t_1)M_Y(t_2)$. Δ

($\Leftarrow$): Now say X, Y are random variables such that $M_{X,Y}(t_1,t_2) = M_X(t_1)M_Y(t_2)$ for all $(t_1,t_2) \in B(0,\varepsilon)$.

Then $M_{X,Y}(t_1,t_2) = E[e^{t_1X}e^{t_2Y}] = E[e^{t_1X}]E[e^{t_2Y}] = M_X(t_1)M_Y(t_2)$.

Denote $g(X) = e^{t_1X}$ and $h(Y) = e^{t_2Y}$.

X, Y discrete. Then $E[g(X)h(Y)] = \sum_x \sum_y g(x)h(y)f(x,y)$

$\overset{\text{by assumption}}{=} \left(\sum_x g(x)f(x)\right)\left(\sum_y h(y)f(y)\right)$. Suppose both sums are finite, so

$$\Rightarrow \sum_{i=1}^{n}\sum_{j=1}^{m} g(x_i)h(y_j)f(x_i,y_j) = \left(\sum_{i=1}^{n} g(x_i)f(x_i)\right)\left(\sum_{j=1}^{m} h(y_j)f(y_j)\right) = \sum_{i=1}^{n}\sum_{j=1}^{m} g(x_i)h(y_j)f(x_i)f(y_j)$$

$$\Rightarrow f(x_i,y_j) = f(x_i)f(y_j) \text{ for all } i,j.$$

(marginal distributions)

Therefore $F_X(x_i)F_Y(y_i) = F_{X,Y}(x_i,y_i)$ so X and Y are independent (by definition). Δ

The other cases are out of scope. $\square$

This gives us an (easier) way to check for independence: find $M_{X,Y}(t_1,t_2) \Rightarrow$ derive M_X and M_Y, check factorization.

Theorem (Uniqueness of MGF): If random variables X and Y satisfy $M_X = M_Y$, then $X, Y \sim f$ are identical.

Example: Recall $Z \sim \text{Gamma}(\alpha,\beta) \Rightarrow M_Z(t) = (1-\beta t)^{-\alpha}$. If $M_X(t_1) = (1-t_1)^{-1}$ for $t_1 < 1 \Rightarrow X \sim \text{Gamma}(1,1)$.

If $M_Y(t_2) = (1-t_2)^{-2}$ for $t_2 < 1 \Rightarrow Y \sim \text{Gamma}(2,1)$.

Notice $E[e^{tX}e^{tY}] = M_{X+Y}(t)$ (both some t) and $M_{X+Y}: \mathbb{R} \to \mathbb{R}$ is fundamentally different from $M_{X,Y}: \mathbb{R}^2 \to \mathbb{R}$.

Corollary: Let X, Y be independent random variables. Then $M_{X+Y}(t) = M_X(t)M_Y(t)$.

Proof: We have $M_{X+Y}(t) = E[e^{t(X+Y)}] = E[e^{tX}e^{tY}] = E[e^{tX}]E[e^{tY}] = M_X(t)M_Y(t)$. $\square$

Example: Say $X_1 \sim \text{Bin}(n,p)$ and $X_2 \sim \text{Bin}(m,p)$ are independent. Show $X_1 + X_2 \sim \text{Bin}(n+m,p)$ using MGF uniqueness.

We have $M_{X_1}(t_1) = (1+pe^{t_1}-p)^n$ and $M_{X_2}(t_2) = (1+pe^{t_2}-p)^m$

$$\Rightarrow M_{X_1+X_2}(t) = M_{X_1}(t)M_{X_2}(t) = (1+pe^t-p)^n(1+pe^t-p)^m = (1+pe^t-p)^{n+m} \Rightarrow X_1+X_2 \sim \text{Bin}(n+m,p).$$

(MGF of $\text{Bin}(n+m,p)$)

Example: (done below) Say $X_1 \sim \text{Gamma}(\alpha_1,\beta)$ and $X_2 \sim \text{Gamma}(\alpha_2,\beta)$. Show that $X_1+X_2 \sim \text{Gamma}(\alpha_1+\alpha_2,\beta)$.

Example: (done below) Say $X_i \sim N(\mu_i,\sigma_i^2)$ for $i = 1,\dots,n$ are all independent. Show that $c_0 + \sum_{i=1}^{n} c_iX_i \sim N\left(c_0 + \sum_{i=1}^{n} c_i\mu_i,\ \sum_{i=1}^{n} c_i^2\sigma_i^2\right)$.

Convergence

We introduce the notion of convergence in two ways.

Let $(X_i)_{i\in\mathbb{N}}$ be a sequence of random variables.

We say $(X_i)_{i\in\mathbb{N}}$ converges in probability to a random variable $X \iff \lim_{n\to\infty} P(|X_n - X| \geq \varepsilon) = 0$ } strong convergence

and write $X_n \xrightarrow{P} X$. An "unusual" outcome becomes less likely as the sequence progresses.

We say $(X_i)_{i\in\mathbb{N}}$ converges in distribution to a random variable $X \iff \lim_{n\to\infty} F_{X_n}(x) = F_X(x)$ at all continuity points $x\in\mathbb{R}$ of F_X } weak convergence (pointwise)

and write $X_n \xrightarrow{D} X$.

Let $T\subseteq\mathbb{R}$ be an arbitrary set (interpreted as the set of time points). For each $t\in T$, define $X_t : \Omega \to \mathbb{R}$ a random variable.

Then $(X_t)_{t\in T}$ is called a stochastic process.

For $\omega\in\Omega$, the map $\begin{matrix} T \to \mathbb{R} \\ t \mapsto X_t(\omega) \end{matrix}$ is called a path.

Let $(X_t)_{t\in T}$ be a stochastic process with $\overbrace{T\subseteq\mathbb{Z}}^{\text{discrete-time}}$ or $\overbrace{T\subseteq\mathbb{R}}^{\text{continuous-time}}$.

We call $(X_t)_{t\in T}$ a Markov process (or a Markov chain) if for all $n\in\mathbb{N}$, $t_1, t_2, \dots, t_n, t \in T$ with $t_1 < t_2 < \dots < t_n < t$ (set of time points; ordered in time)

and $x_1, \dots, x_n, x \in \mathbb{R}$, we have $P(X_t = x \mid X_{t_1} = x_1, \dots, X_{t_n} = x_n) = P(X_t = x \mid X_{t_n} = x_n)$. (Markov property)

In other words, the probability of the "future" only depends on the present.

If $X_n \overset{cdf}{\sim} F_n(x)$ and $X \overset{cdf}{\sim} F(x)$ with $X_n \xrightarrow{D} X$, then $F_n(x) \approx F(x)$ at all continuity points of F. Useful approximation.

For notation, we write $X_i \overset{IID}{\sim} F$ if X_i are independent and identically distributed random variables.

Example: Say $Y_i \overset{IID}{\sim} \text{Unif}(0,\theta)$. Let $X_n = \max\{Y_1, \dots, Y_n\}$. Find X such that $X_n \xrightarrow{D} X$.

Notice $X_1 = \max\{Y_1\}$, $X_2 = \max\{Y_1, Y_2\}$, $\vdots$. What's $\lim_{n\to\infty} F_n(x)$? Notice

$$F_n(x) \overset{\text{defn}}{=} P(X_n \leq x) = P(\max\{Y_1,\dots,Y_n\} \leq x)$$

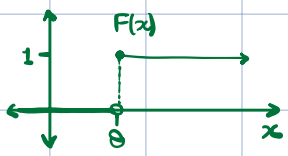
$$\overset{\max\{Y_i\}\leq x \Rightarrow \text{all } Y_i \leq x}{=} P(Y_1 \leq x, \dots, Y_n \leq x)$$

$$\overset{\text{independence}}{=} \prod_{i=1}^{n} P(Y_i \leq x)$$

$$\overset{\text{identically distributed}}{=} \prod_{i=1}^{n} F_i(x) = (F_i(x))^n = \begin{cases} 0^n & \text{if } x \leq 0 \\ \left(\frac{x}{\theta}\right)^n & \text{if } 0 < x < \theta \\ 1^n & \text{if } x \geq \theta \end{cases}.$$

$$\Rightarrow \lim_{n\to\infty} F_n(x) = \begin{cases} 0 & \text{for } x \leq 0 \\ 0 & \text{for } 0 < x < \theta \\ 1 & \text{for } x \geq \theta \end{cases}$$ ← this convergence is not uniform!

Therefore $f(x) \overset{\text{probability density}}{=} \frac{d}{dx}F(x) = \theta$ (constant):



and notice that $F_n(x) \to F(x)$ wherever F continuous $\Rightarrow F_n(x) \to F(x)$ on $\mathbb{R}\setminus\{\theta\}$. Thus $X_n \xrightarrow{D} X = \theta$.

Theorem: Let $(X_n)_{n\in\mathbb{N}}$ be a sequence/stochastic process. If $X_n \xrightarrow{P} X$ (strong convergence), then $X_n \xrightarrow{D} X$ (weak convergence).

Proof: We can only prove this for $X = a$ (constant) without heavy measure theory.

Let $X = a \in \mathbb{R}$. Notice $F_X(x)$ is a step function continuous on $\mathbb{R}\setminus\{a\}$, and $F_X(x)$ is constant on $\mathbb{R}\setminus\{a\}$.

$\lim F_n(x) = F(x)$ on $\mathbb{R}\setminus\{a\}$.

We show $\forall \varepsilon>0$, we have $\underbrace{F(x-\varepsilon)}_{\text{lower bound}} \le \lim_{n\to\infty} F_n(x) \le \underbrace{F(x+\varepsilon)}_{\text{upper bound}}$ for all $x\in\mathbb{R}\setminus\{a\}$.

(1) upper bound

$$F_n(x) \overset{\text{defn}}{=} P(X_n\le x) \overset{\text{total probability}}{=} P(X_n\le x, \overbrace{X\le x+\varepsilon}^{A}) + P(X_n\le x, \overbrace{X>x+\varepsilon}^{A^c})$$

$X_n\le x$ and $X-\varepsilon>x \Rightarrow X-X_n\ge\varepsilon$

$$\le P(X\le x+\varepsilon) + P(X-X_n\ge\varepsilon) \le P(X\le x+\varepsilon) + P(|X_n-X|\ge\varepsilon)$$

(superset)

$$= F(x+\varepsilon) + \underbrace{P(|X_n-X|\ge\varepsilon)}_{\to 0 \text{ since } X_n\xrightarrow{P}X} \xrightarrow{n\to\infty} F(x+\varepsilon).$$

(2) lower bound

$$F(x-\varepsilon) = P(X<x-\varepsilon, \overbrace{X_n\le x}^{B}) + P(X<x-\varepsilon, \overbrace{X_n>x}^{B^c})$$

$$\le P(X_n\le x) + P(X_n-X\ge\varepsilon) \le P(X_n\le x) + \underbrace{P(|X_n-X|\ge\varepsilon)}_{\to 0 \text{ since } X_n\xrightarrow{P}X} \xrightarrow{n\to\infty} \lim_{n\to\infty} F_n(x).$$

Thus $F(x-\varepsilon) \le \lim_{n\to\infty} F_n(x) \le F(x+\varepsilon)$ for any $\varepsilon>0$. Holds for all $x\in\mathbb{R}\setminus\{a\}$. □

Theorem: Let (X_n) be a sequence/stochastic process. If $X_n\xrightarrow{D}X=C$ (constant), then $X_n\xrightarrow{P}X=C$ at all continuity points of F_X.

Remark: We have $X_n\xrightarrow{P}X=C \Leftrightarrow X_n\xrightarrow{D}X=C$.

Proof: Notice $X=C\in\mathbb{R} \Rightarrow F_X$ continuous on $\mathbb{R}\setminus\{C\} \Rightarrow C$ not a continuity point of F_X.

Thus $\forall\varepsilon>0$, we have $P(|X_n-X|\ge\varepsilon) = P(|X_n-C|\ge\varepsilon)$

$$= P(X_n-C\ge\varepsilon) + P(X_n-C\le-\varepsilon)$$

$$= P(X_n\ge C+\varepsilon) + P(X_n\le C-\varepsilon) \quad \text{and we bound this above:}$$

$P(X_n\ge C+\varepsilon) = 1-P(X_n<C+\varepsilon) \le 1-P(X_n\le C+\frac{\varepsilon}{2})$

$P(X_n\le C-\varepsilon) = F_{X_n}(C-\varepsilon)$ by definition

$$\le 1-F_{X_n}(C+\tfrac{\varepsilon}{2}) + F_{X_n}(C-\varepsilon) \xrightarrow{n\to\infty} 1-F_X(C+\tfrac{\varepsilon}{2}) + F_X(C-\varepsilon)$$

($C+\frac{\varepsilon}{2}$ and $C-\varepsilon$ are continuity points of F_X and $X_n\xrightarrow{D}X$.)

$$= 1-1+0 = 0.$$

Therefore $0\le P(|X_n-X|\ge\varepsilon) \le 1-F_{X_n}(C+\frac{\varepsilon}{2}) + F_{X_n}(C-\varepsilon) \xrightarrow{n\to\infty} 0$

$\Rightarrow P(|X_n-X|\ge\varepsilon) \xrightarrow{n\to\infty} 0$. Thus $X_n\xrightarrow{P}X$. □

Remark: In general, note that $X_n\xrightarrow{D}X \not\Rightarrow X_n\xrightarrow{P}X$.

Example: Say $X\sim N(0,1)$ and the sequence $(X_i)_{i\in\mathbb{N}}$ is defined by $X_i=-X$. We show $X_n\xrightarrow{D}X$ but $X_n\not\xrightarrow{P}X$.

Notice by symmetry, $X_i=-X \Rightarrow f_{X_i}(x) = f_X(-x) = \frac{1}{\sqrt{2\pi}}e^{\frac{(-x)^2}{2}} = \frac{1}{\sqrt{2\pi}}e^{\frac{x^2}{2}} = f_X(x)$ (could alternatively use MGF uniqueness to prove that $X_i\sim N(0,1)$)

$\Rightarrow X_i\sim N(0,1)$. Thus obviously $X_n\xrightarrow{D}X$, since $f_{X_i}=f_X \Rightarrow F_{X_i}=F_X$

$\Rightarrow F_{X_n}(x)\xrightarrow{n\to\infty}F_X(x)$ for all $x\in\mathbb{R}$.

However, $X_n\xrightarrow{P}X \Leftrightarrow \forall\varepsilon>0$, we have $\lim_{n\to\infty} P(|X_n-X|\ge\varepsilon)=0$, but

$$= \lim_{n\to\infty} P(|-X-X|\ge\varepsilon) = P(2|X|\ge\varepsilon) \neq 0 \text{ for } \varepsilon=1 \text{ (for example).}$$

Therefore $X_n\not\xrightarrow{P}X$.

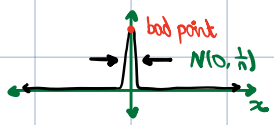
Notice that $X_n\xrightarrow{D}X$ restricts convergence to continuity points of F_X.

Example: Let $X_i\overset{IID}{\sim}N(0,1)$ and define $\bar{X}_n=\frac{1}{n}\sum_{i=1}^n X_i$. Then $\bar{X}_n\xrightarrow{D}X=0$ but $\lim_{n\to\infty}F_{\bar{X}_n}(0)\neq F_X(0)$ (since $0\in\mathbb{R}$ not a continuity point).

Let $F_n=F_{\bar{X}_n}$. Then $\lim_{n\to\infty}F_n(x) = \lim_{n\to\infty}P(\bar{X}_n\le x)$.

Notice $\bar{X}_n=\frac{1}{n}\sum_{i=1}^n X_i$

intuitively

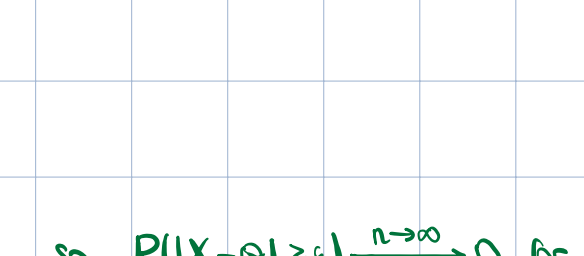


$\Rightarrow \bar{X}_n \sim N\left(\frac{1}{n}\sum_{i=1}^n 0,\ \frac{1}{n^2}\sum_{i=1}^n 1\right) = N\left(0, \frac{1}{n}\right)$. By the reparametrization trick, $\frac{\bar{X}_n - 0}{\sqrt{\frac{1}{n}}} \sim N(0,1)$, so

$$F_n(x) = P(\bar{X}_n \le x) \overset{\text{reparam.}}{=} P\left(\frac{\bar{X}_n - 0}{\sqrt{\frac{1}{n}}} \le \frac{x-0}{\sqrt{\frac{1}{n}}}\right) = P(Z \le \sqrt{n}x) \text{ where } Z \sim N(0,1)$$

$$= F_Z(\sqrt{n}x) \xrightarrow{n\to\infty} \begin{cases} 0 \text{ if } x<0 \\ \frac{1}{2} \text{ if } x=0 \\ 1 \text{ if } x>0 \end{cases} \text{ so } \lim_{n\to\infty} F_n(0) = \frac{1}{2} \neq 1 = F_X(0) \text{ where } X = 0.$$

(F_Z: standard normal cdf)



Notice $F_X(x) = \begin{cases} 0 \text{ if } x<0 \\ 1 \text{ if } x=0 \\ 1 \text{ if } x>0 \end{cases}$ so $F_n(x) \xrightarrow{n\to\infty} F_X(x)$ for $x \in \mathbb{R}\setminus\{0\}$ (continuity points).

Example: Let $Y_i \overset{IID}{\sim} \text{Unif}(0,\theta)$ and let $X_n = \max\{Y_1,\dots,Y_n\}$. Show $X_n \xrightarrow{D}$ and find the distribution limit.

Intuitively, $\max\{Y_1,\dots,Y_n\} \xrightarrow{n\to\infty} \max[0,\theta] = \theta$. So let $X = \theta$. We show $X_n \xrightarrow{D} X = \theta$.

It suffices to prove $X_n \xrightarrow{P} X = \theta$ (strong convergence).

By definition, $\forall \varepsilon > 0$, we want $\lim_{n\to\infty} P(|X_n - \theta| \ge \varepsilon) = 0$.

Notice $P(|X_n - \theta| \ge \varepsilon) = P(|\max\{Y_1,\dots,Y_n\} - \theta| \ge \varepsilon) = P(\theta - \max\{Y_1,\dots,Y_n\} \ge \varepsilon)$ (since $\theta \ge X_n$)

$$= P(\max\{Y_1,\dots,Y_n\} \le \theta - \varepsilon) = P(Y_1 \le \theta-\varepsilon, \dots, Y_n \le \theta - \varepsilon)$$

$$\overset{\text{independence}}{=} \prod_{i=1}^n P(Y_i \le \theta - \varepsilon) \overset{\text{identical}}{=} (P(Y_1 \le \theta - \varepsilon))^n = \begin{cases} 0 \text{ if } \theta - \varepsilon < 0 \\ \left(\frac{\theta-\varepsilon}{\theta}\right)^n \text{ if } 0 \le \theta - \varepsilon < \theta \end{cases}$$

$\downarrow n\to\infty$

$$\begin{cases} 0 \text{ if } \theta - \varepsilon < 0 \\ \lim_{n\to\infty}\left(1 - \frac{\varepsilon}{\theta}\right)^n = 0 \text{ for } 0 \le \theta - \varepsilon < \theta \end{cases}$$

so $P(|X_n - \theta| \ge \varepsilon) \xrightarrow{n\to\infty} 0$ for all $\varepsilon > 0$. Thus $X_n \xrightarrow{P} X = \theta \Rightarrow X_n \xrightarrow{D} X = \theta$. □

Proposition (e-limit): If $b \in \mathbb{R}$ and $\lim_{n\to\infty} \psi(n) = 0$, then $\lim_{n\to\infty}\left(1 + \frac{b}{n} + \frac{\psi(n)}{n}\right)^n = e^b$.

Example: Let $Y_i \overset{IID}{\sim} \text{Exp}(1)$. Let $X_n = \max\{Y_1,\dots,Y_n\} - \log n$. Show that $X_n \xrightarrow{D}$ and find the distribution limit.

Notice $f_{Y_i}(y) = e^{-y}$ for $y \ge 0$

$\Rightarrow F_{Y_i}(y) = \begin{cases} 0 \text{ for } y \le 0 \\ 1 - e^{-y} \text{ for } y > 0 \end{cases}$. Let $F_n(x) = F_{X_n}(x)$ be the c.d.f. of X_n.

Then $F_n(x) = P(X_n \le x) = P(\max\{Y_1,\dots,Y_n\} - \log n \le x)$

$$= P(\max\{Y_1,\dots,Y_n\} \le x + \log n) = P(Y_1 \le x + \log n, \dots, Y_n \le x + \log n)$$

$$\overset{\text{independence}}{=} \prod_{i=1}^n P(Y_i \le x + \log n) \overset{\text{identical}}{=} (P(Y_i \le x + \log n))^n$$

$$= (F_{Y_i}(x + \log n))^n = \begin{cases} 0^n \text{ for } x + \log n \le 0 \\ (1 - e^{-(x+\log n)})^n \text{ for } x + \log n > 0 \end{cases}$$

Notice $(1 - e^{-(x+\log n)})^n = (1 - e^{-x}e^{-\log n})^n = (1 - e^{-x}e^{\log n^{-1}})^n$

$$= \left(1 - \frac{e^{-x}}{n}\right)^n \xrightarrow[\text{e-limit}]{n\to\infty} e^{-e^{-x}}. \quad \text{Notice } \{x \in \mathbb{R} \mid x + \log n > 0\} \xrightarrow{n\to\infty} \mathbb{R}.$$

$\Rightarrow \lim_{n\to\infty} F_n(x) = e^{-e^{-x}} = F(x)$ for all $x \in \mathbb{R}$.

We must verify that $F(x)$ defines a valid cumulative distribution function:

$\lim_{x\to-\infty} F(x) = \lim_{x\to-\infty} e^{-e^{-x}} \longrightarrow e^{-e^{\infty}} = e^{-\infty} = 0.$

$\lim_{x\to\infty} F(x) = \lim_{x\to\infty} e^{-e^{-x}} \longrightarrow e^{-e^{-\infty}} = e^{-0} = 1.$

$F(x)$ non-decreasing: satisfied by $-e^{-x}$ increasing.

Therefore $F(x)$ is a well-defined c.d.f. $\Rightarrow X_n \xrightarrow{D} X \overset{cdf}{\sim} e^{-e^{-x}}$ for $x \in \mathbb{R}$. $\square$

Let $(X_i)_{i\in\mathbb{N}}$ be a sequence/stochastic process with $(M_{X_i}(t))_{i\in\mathbb{N}}$ their moment generating functions.

Notice, for fixed $t\in\mathbb{R}$, that $(M_{X_i}(t))_{i\in\mathbb{N}} \subseteq \mathbb{R}$ is a real sequence.

Theorem (MGF Convergence): Let $(X_i)_{i\in\mathbb{N}}$ be a sequence of random variables, and $(M_{X_i}(t))_{i\in\mathbb{N}}$ their MGFs, all well-defined in some common neighbourhood $B(0,\varepsilon)\subseteq\mathbb{R}$. Suppose $M_{X_n}(t) \to M_X(t)$ for $t\in B(0,\varepsilon)$. Then $X_n \xrightarrow{D} X$.

Proof: Heavy measure theory. $\square$

We will use this to prove the Central Limit Theorem.

Example: Let $X_n \sim Bin(n,p)$. Show that $X_n \xrightarrow{D}$ and find the limit such that $\lambda = np$ (λ: constant). Use the limit to approximate $F_{X_n}(x)$.

Since $np=\lambda$ is a constant, we have $X_n \sim Bin(n, \frac{\lambda}{n})$. We postulate that $X_n \xrightarrow{D} X$ where $X\sim Poi(\lambda)$.

Notice $M_{X_n}(t) = (1+pe^t-p)^n = (1+\frac{\lambda}{n}e^t-\frac{\lambda}{n})^n = (1+\frac{\lambda}{n}(e^t-1))^n \xrightarrow[\text{e-limit}]{n\to\infty} e^{\lambda(e^t-1)}$ for $t\in\mathbb{R}$

and notice that $M_X(t) = e^{\lambda(e^t-1)}$ for $t\in\mathbb{R}$

$\Rightarrow \lim_{n\to\infty} M_{X_n}(t) = M_X(t)$ for $t\in\mathbb{R} \Rightarrow X_n \xrightarrow{D} X \sim Poi(\lambda)$. $\square$

How can we approximate $F_{X_n}(x)$? Notice $\underset{Bin(n,p)}{X_n} \xrightarrow{D} \underset{Poi(\lambda)}{X}$ are both discrete, so F_X (step function on $\mathbb{Z}$) has continuity points $\mathbb{R}\setminus\mathbb{Z}_{\geq 0}$.

Therefore, we don't necessarily have $\lim_{n\to\infty} F_{X_n}(x) = F_X(x)$ for $x\in\mathbb{Z}_{\geq 0}$.

However, for $0<\varepsilon<1$ we have that $x-\varepsilon,\ x+\varepsilon \in \mathbb{R}\setminus\mathbb{Z}_{\geq 0}$ for $x\in\mathbb{Z}_{\geq 0}$.

$\Rightarrow \lim_{n\to\infty} F_{X_n}(x\pm\varepsilon) = F_X(x\pm\varepsilon)$ so $P(X_n=x) = F_{X_n}(x+\varepsilon) - F_{X_n}(x-\varepsilon) \xrightarrow{n\to\infty} F_X(x+\varepsilon) - F_X(x-\varepsilon) = P(X=x)$.

(continuity points)

Theorem (Central Limit): Let X_i be independent, identically distributed random variables with $E[X_i]=\mu$ and $Var(X_i)=\sigma^2<\infty$.

Then $\frac{\sqrt{n}(\bar{X}_n-\mu)}{\sigma} \xrightarrow{D} N(0,1)$ where $\bar{X}_n = \frac{1}{n}\sum X_i$ (sample average).

Proof: For this course, we must add one (weak) assumption: $M_{X_i}(t)$ exist in some $B(0,\varepsilon)$. We use MGF convergence.

Define $Z_n = \frac{\sqrt{n}(\bar{X}_n-\mu)}{\sigma}$. We want $Z_n \xrightarrow{D} Z\sim N(0,1)$. Denote $M_n = M_{Z_n}$ and $M = M_Z$.

Suffices to show $M_n(t) \to M(t)$ for $t\in B(0,\varepsilon^*)$.

We have $M(t) = e^{\frac{1}{2}t^2}$ for all $t\in\mathbb{R}$, so we want $\lim_{n\to\infty} M_n(t) = e^{\frac{1}{2}t^2}$.

Notice $Z_n = \frac{\sqrt{n}(\bar{X}_n-\mu)}{\sigma} = \frac{\sqrt{n}(\frac{1}{n}\sum_{i=1}^n X_i-\mu)}{\sigma} = \frac{\sqrt{n}}{n}\left(\sum_{i=1}^n X_i - n\mu\right)\sigma^{-1}$

$= \frac{1}{\sqrt{n}}\left(\sum_{i=1}^n \left(\frac{X_i-\mu}{\sigma}\right)\right)$ so Z_n is a linear combination of standardized X_i

$= \frac{1}{\sqrt{n}}\sum_{i=1}^n Y_i$ where $Y_i = \frac{X_i-\mu}{\sigma} \sim N(0,1)$.

Important: $Y_i\sim N(0,1) \Rightarrow E[Y_i]=0$ and $E[Y_i^2] = Var(Y_i) + \cancel{E[Y_i]^2}^{\,0} = Var(Y_i) = 1$.

$\Rightarrow M_{Y_i}'(0)=0$ and $M_{Y_i}''(0)=1$.

Computing $M_n(t)$.

We have $Z_n = \frac{1}{\sqrt{n}}\sum_{i=1}^n Y_i \Rightarrow M_n(t) = E[\exp(\frac{t}{\sqrt{n}}\sum_{i=1}^n Y_i)]$ where $X_i \overset{IID}{\sim} N(\mu,\sigma^2) \Rightarrow Y_i \overset{IID}{\sim} N(0,1)$.

$= E[\prod^n \exp(\frac{t}{\sqrt{n}}Y_i)]$

independence → $= \prod_{i=1}^{n} E[\exp(\frac{t}{\sqrt{n}}Y_i)] \overset{\text{identical}}{=} (M_{Y_t}(\frac{t}{\sqrt{n}}))^n$ where $M_{Y_t}(\frac{t}{\sqrt{n}}) = E[\exp(\frac{t}{\sqrt{n}}Y_t)] = E[\exp(\frac{t}{\sqrt{n}}(\frac{X-\mu}{\sigma}))]$

exists for $\frac{t}{\sigma\sqrt{n}} \in B(0,\varepsilon)$

$\Rightarrow t \in B(0, \varepsilon\sigma\sqrt{n}) \xrightarrow{n\to\infty} \mathbb{R}$.

Taylor Expansion of $M_{Y_t}(\frac{t}{\sqrt{n}})$ at $t=0$.

$$M_{Y_t}(\tfrac{t}{\sqrt{n}}) = \underbrace{M_{Y_t}(0)}_{1} + \underbrace{M'_{Y_t}(0)}_{0}(\tfrac{t}{\sqrt{n}}) + \frac{1}{2!}\underbrace{M''_{Y_t}(0)}_{1}(\tfrac{t}{\sqrt{n}})^2 + O(n^{-\frac{3}{2}}).$$

higher order terms

$$= 1 + \frac{1}{n}(\frac{1}{2}t^2) + O(n^{-\frac{3}{2}}) \Rightarrow M_n(t) = (M_{Y_t}(\tfrac{t}{\sqrt{n}}))^n = (1 + \frac{\frac{1}{2}t^2}{n} + O(n^{-\frac{3}{2}}))^n$$

around $t=0$; where $O(n^{-\frac{3}{2}}) = \frac{O(n^{-\frac{1}{2}})}{n}$ and $O(n^{-\frac{1}{2}}) \xrightarrow{n\to\infty} 0$.

e-limit ↓

$e^{\frac{t^2}{2}}$ for $t \in B(0,\varepsilon^*)$.

Therefore $M_n(t) \to M(t)$ for $t \in B(0,\varepsilon^*)$, so by MGF convergence, we have $Z_n \xrightarrow{D} Z \sim N(0,1)$. □

We define the χ distribution as follows: say $X_i \overset{IID}{\sim} \text{Gamma}(1,\beta)$. Define χ^2 as $\frac{2\sum_{i=1}^{n} X_i}{\beta} \sim \chi^2(2n)$.

If $X \sim \chi(n)$, then $f_X(x) = \frac{x^{k-1}e^{\frac{-x^2}{2}}}{2^{k/2-1}\Gamma(k/2)}$ for $x \geq 0$.

Example: Say $Y_n \sim \chi(n)$ for $n \in \mathbb{N}_{\geq 0}$. Show that $Z_n = \frac{Y_n - n}{\sqrt{2n}} \xrightarrow{D} Z \sim N(0,1)$.

We have $Y_n = \sum_{i=1}^{n} X_i$ where $X_i \overset{IID}{\sim} \text{Gamma}(\frac{1}{2}, 2)$.

Therefore since $E[X_i] = \frac{1}{2}\cdot 2 = 1 < \infty$ and $\text{Var}(X_i) = \frac{1}{2}\cdot 2^2 = 2 < \infty$.

$\overset{CLT}{\Rightarrow} \frac{\sqrt{n}(\bar{X}_n - 1)}{\sqrt{2}} \xrightarrow{D} N(0,1)$ so $\frac{\sqrt{n}}{\sqrt{2}}(\frac{1}{n}\sum_{i=1}^{n} X_i - 1) = \frac{1}{\sqrt{2n}}(\sum_{i=1}^{n} X_i - n) = \frac{Y_n - n}{\sqrt{2n}} \xrightarrow{D} N(0,1)$.

Example: Fires are reported to a station following a Poisson process, with an average of 1 fire per 4 hours. Find the probability that the 500^{th} fire is reported by the end of the 84^{th} day.

Let X_i be the waiting time between the $(i-1)^{th}$ and $(i)^{th}$ fire.

Then $X_i \overset{IID}{\sim} \text{Exp}(6)$ (6 per day) and $f_{X_i}(x) = f_i(x) = 6e^{-6x}$ for $x > 0$.

Notice $E[X_i] = \frac{1}{6} < \infty$ and $\text{Var}(X_i) = \frac{1}{6^2} < \infty$ so

$\overset{CLT}{\Rightarrow} \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \xrightarrow{D} N(0,1)$ so with $n=500$, $\mu=\frac{1}{6}$, $\sigma=\frac{1}{6}$, we want $\sum_{i=1}^{500} X_i \leq 84$. (waiting time for 84 fires)

We have $\sqrt{500}\left(\frac{1}{500}\sum_{i=1}^{500} X_i - \frac{1}{6}\right) / (\frac{1}{6}) = \sqrt{500}\left(\frac{\sum X_i}{500} - \frac{1}{6}\right)\cdot 6 \approx Z \sim N(0,1)$

and want $P\left(\sqrt{500}\left(\frac{\frac{\sum X_i}{500} - \frac{1}{6}}{\frac{1}{6}}\right) \leq \sqrt{500}\left(\frac{\frac{84}{500} - \frac{1}{6}}{\frac{1}{6}}\right)\right)$ by reparametrization.

$N(0,1)$; 0.18

$\approx P(Z \leq 0.18)$. Use a lookup table.

Proposition (Markov's Inequality): For any random variable X with $E[|X|^k] < \infty$, we have $P[|X| \geq c] \leq \frac{E[|X|^k]}{c^k}$ for all $c > 0$, for all $k \in \mathbb{Z}_{>0}$.

Proof: Say X_i are continuous. Similar proof for X_i discrete.

We have $\frac{E[|X|^k]}{c^k} = \frac{1}{c^k}\int_{-\infty}^{\infty} |x|^k f_X(x)\,dx$ and $\mathbb{R} = \{|x| \geq c\} \cup \{|x| < c\}$:

$$= \frac{1}{c^k}\left(\int_{\{|x|\geq c\}} |x|^k f_X(x)\,dx + \underbrace{\int_{\{|x|<c\}} |x|^k f_X(x)\,dx}_{\geq 0}\right)$$

$$\geq \frac{1}{c^k}\int_{\{|x|\geq c\}} |x|^k f_X(x)\,dx \geq \frac{1}{c^k}\underbrace{\min_{x\in\{|x|\geq c\}} |x|^k}_{x=c}\int_{\{|x|\geq c\}} f_X(x)\,dx = \int_{\{|x|\geq c\}} f_X(x)\,dx = P(|X|\geq c). \quad \square$$

(definition)

Example: If $Y_n = \bar{X}_n \sim N(\mu, \frac{\sigma^2}{n})$, then $Y_n = \bar{X}_n \xrightarrow{P} \mu$.

Notice by Markov's Inequality that $E[|Y_n|^k]\,c^{-k} \geq P[|Y_n| \geq c]$ for all $c>0$, $k \in \mathbb{Z}_{>0}$.

$\Rightarrow P(|Y_n - \mu| \geq c) \leq E[|Y_n-\mu|^k]\,c^{-k}$

$\Rightarrow P(|Y_n-\mu| \geq \varepsilon) \leq E[|Y_n-\mu|^k]\,\varepsilon^{-k} = M^{(k)}_{|Y_n-\mu|}(0)$ for all $k \in \mathbb{Z}_{>0}$. Let $k=2$:

$\Rightarrow P(|Y_n-\mu| \geq \varepsilon) \leq E[|Y_n-\mu|^2]\,\varepsilon^{-2} = E[(Y_n-\mu)^2]\,\varepsilon^{-2} = (\mathrm{Var}(Y_n-\mu) + \cancel{E[Y_n-\mu]^2}^{0})\,\varepsilon^{-2} = \mathrm{Var}(Y_n)\,\varepsilon^{-2} = \frac{\sigma^2}{n\varepsilon^2}$.

Thus $\underbrace{P(|Y_n-\mu| \geq \varepsilon) \leq \frac{\sigma^2}{n\varepsilon^2}}_{\text{Chebyshev's Inequality}} \xrightarrow{n\to\infty} 0$ so $Y_n \xrightarrow{P} \mu$ by definition.

Theorem (Weak Law of Large Numbers): Suppose X_i are IID with $E[X_i] = \mu$ and $\mathrm{Var}(X_i) = \sigma^2 < \infty$, then $\bar{X}_n \xrightarrow{P} \mu$

where $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$ (average).

Proof: By independence, we have $\mathrm{Var}(\bar{X}_n) = \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^n X_i\right) \overset{\text{independence}}{=} \frac{1}{n^2}\sum_{i=1}^n \mathrm{Var}(X_i) \overset{\mathrm{Var}(X_i)=\sigma}{=} \frac{1}{n^2}\sum_{i=1}^n \sigma^2 = \frac{\sigma^2}{n}$.

Using Chebyshev's inequality: $P(|\bar{X}_n-\mu| \geq \varepsilon) \leq \frac{\sigma^2}{n\varepsilon^2} \xrightarrow{n\to\infty} 0$

so $\bar{X}_n \xrightarrow{P} X = \mu$ by definition. Done. $\square$

Example: Say $X_1 \sim \text{Gamma}(\alpha_1, \beta) \perp X_2 \sim \text{Gamma}(\alpha_2, \beta)$. Show that $X_1 + X_2 \sim \text{Gamma}(\alpha_1+\alpha_2, \beta)$.

We have that $M_{X_1+X_2}(t) = M_{X_1}(t) M_{X_2}(t)$ by independence.

$\Rightarrow M_{X_1+X_2}(t) = M_{X_1}(t) M_{X_2}(t) = (1-\beta t)^{-\alpha_1}(1-\beta t)^{-\alpha_2} = (1-\beta t)^{-(\alpha_1+\alpha_2)} = M_Y(t)$ where $Y \sim \text{Gamma}(\alpha_1+\alpha_2, \beta)$

$\Rightarrow X_1 + X_2 \sim \text{Gamma}(\alpha_1+\alpha_2, \beta)$ by MGF uniqueness. $\square$

Example: Say $X_i \sim N(\mu_i, \sigma_i^2)$ for $i = 1, \dots, n$ are all independent. Show that $c_0 + \sum_{i=1}^n c_i X_i \sim N\left(c_0 + \sum_{i=1}^n c_i \mu_i, \sum_{i=1}^n c_i^2 \sigma_i^2\right)$.

We use MGF uniqueness. Notice that

$M_{c_0+\sum_{i=1}^n c_i X_i}(t) = M_{c_0}(t) \prod_{i=1}^n M_{c_i X_i}(t)$. We have $M_{c_0}(t) = E[e^{c_0 t}] = e^{c_0 t}$ and

$$M_{c_i X_i}(t) = E[e^{c_i X_i t}] = E[e^{X_i (c_i t)}] = M_{X_i}(c_i t).$$

Write $X_i = \sigma_i Z + \mu_i$ where $Z \sim N(0,1)$. Then $M_{X_i}(c_i t) = M_{\sigma_i Z + \mu_i}(c_i t) \overset{\text{Prop.}}{=} e^{\mu_i t} M_Z(\sigma_i c_i t) = e^{\mu_i t} e^{\frac{1}{2}(\sigma_i c_i t)^2}$.

$$\Rightarrow M_{c_0+\sum_{i=1}^n c_i X_i}(t) = M_{c_0}(t) \prod_{i=1}^n M_{c_i X_i}(t)$$
$$= e^{c_0 t} \prod_{i=1}^n e^{\mu_i t} e^{\frac{1}{2}(\sigma_i c_i t)^2} = e^{t(c_0 + \sum_{i=1}^n \mu_i)} e^{\frac{1}{2}\sum_{i=1}^n (\sigma_i c_i t)^2} = e^{t(c_0 + \sum_{i=1}^n \mu_i)} e^{\frac{1}{2}t^2 \sum_{i=1}^n (\sigma_i c_i)^2} \overset{\text{MGF uniqueness}}{\sim} N\left(c_0 + \sum_{i=1}^n \mu_i, \sum_{i=1}^n \sigma_i^2 c_i^2\right).$$

Notice if $Y \sim N(\mu, \sigma^2)$, then $Y = \sigma Z + \mu \Rightarrow M_Y(t) = M_{\sigma Z + \mu}(t) = e^{\mu t} M_Z(\sigma t) = e^{\mu t} e^{\frac{1}{2}(\sigma t)^2}$

Example: Say $Y_n \sim \text{Bin}(n, \theta)$. Show that $\frac{Y_n}{n} \xrightarrow{P} \theta$.

Method 1: Markov's Inequality. We show $P(|\frac{Y_n}{n} - \theta| \geq \varepsilon) \xrightarrow{n\to\infty} 0$.

Define $X_n = \frac{Y_n}{n}$. Then $P(|X_n - \theta| \geq c) \leq \frac{E[|X_n-\theta|^k]}{c^k}$ for any $c > 0$, $k \in \mathbb{Z}_{>0}$. Choose $c = \varepsilon$ and $k = 2$.

$$\Rightarrow P(|X_n - \theta| \geq \varepsilon) \leq \frac{E[|X_n-\theta|^2]}{\varepsilon^2} = E[(X_n-\theta)^2]\varepsilon^{-2} \overset{(E[X_n] = E[\frac{Y_n}{n}] = n\theta\frac{1}{n} = \theta)}{=} \text{Var}(X_n)\varepsilon^{-2} = \frac{\theta(1-\theta)}{n}\varepsilon^{-2} \xrightarrow{n\to\infty} 0.$$

$\text{Var}(X_n - \theta) = E[X_n-\theta]^2 - (E[X_n]-\theta)^2$, and $(E[X_n]-\theta)^2 \to 0$.

Thus $X_n \xrightarrow{P} \theta$.

Method 2: Indicator functions and WLLN.

Define $Y_n = \begin{pmatrix} X_1 \\ \vdots \\ X_n \end{pmatrix}$ where $X_i = \text{Bin}(1, \theta)$ are indicator functions, so $Y_n = \sum_{i=1}^n X_i$ and $\frac{Y_n}{n} = \frac{1}{n}\sum_{i=1}^n X_i$.

Notice X_i are IID with $E[X_i] = \theta$ and $\text{Var}(X_i) = \theta(1-\theta) < \infty$.

$\overset{\text{WLLN}}{\Rightarrow} \frac{1}{n}\sum_{i=1}^n X_i \xrightarrow{P} \theta$. Thus $\frac{Y_n}{n} = \frac{1}{n}\sum_{i=1}^n X_i \xrightarrow{P} \theta$.

Exercise: Say $X \sim \text{Gamma}(\alpha, \beta)$. Let $Y = X/\beta$. Find $M_Y(t)$ and state the values of t for which this is well-defined.

We have $M_Y(t) = M_{X/\beta}(t) = E[e^{tX\beta^{-1}}] = M_X(t/\beta) = (1-\beta(\frac{t}{\beta}))^{-\alpha} = (1-t)^{-\alpha}$.

$M_X(t)$ defined for $t < \frac{1}{\beta} \Rightarrow M_X(t/\beta)$ defined for $t/\beta < 1/\beta \Rightarrow t < 1$.

Exercise: Let $Z \sim N(\mu, \sigma^2)$. Use the reparametrization trick and Γ-technique to find $E[X]$ and $\text{Var}(X)$.

Easy. $Z = \sigma X + \mu$ where $X \sim N(0,1) \Rightarrow M_Z(t) = M_{\sigma X+\mu}(t) = e^{\mu t} M_X(\sigma t) = e^{\mu t} e^{\frac{1}{2}\sigma^2 t^2}$

Exercise: Use the Γ-technique to find $E[X]$ and $Var(X)$ where: 1) $X \sim Unif(a,b)$ 2) $X \sim Beta(a,b)$.

1) Let $X \sim Unif(a,b)$. Then $E[X] = \int_a^b x f_X(x)\,dx = \int_a^b x \frac{1}{b-a}\,dx = \frac{1}{b-a}\left(\frac{1}{2}x^2\right)_a^b$ = How could we have used Γ-technique?
$= \frac{1}{b-a}\frac{1}{2}(b^2-a^2) = \frac{b+a}{2}$ (as expected).

$Var(X) = \int_a^b x^2 f_X(x)\,dx - E[X]^2 = \frac{1}{b-a}\left(\frac{1}{3}x^3\right)_a^b$.

2) Let $X \sim Beta(a,b)$. Then $f_X(x) = \frac{1}{B(a,b)} x^{a-1}(1-x)^{b-1}$ for $x \in [0,1]$

$\Rightarrow E[X] = \int_0^1 x f_X(x)\,dx = \frac{1}{B(a,b)}\int_0^1 x^a (1-x)^{b-1}\,dx$. Let $\alpha_1 = a+1$ and $\alpha_2 = b$, so

$= \frac{1}{B(\alpha_1-1,\alpha_2)}\int_0^1 x^{\alpha_1-1}(1-x)^{\alpha_2-1}\,dx = \frac{\Gamma(\alpha_1-1+\alpha_2)}{\Gamma(\alpha_1-1)\Gamma(\alpha_2)}\frac{\Gamma(\alpha_1)\Gamma(\alpha_2)}{\Gamma(\alpha_1+\alpha_2)}$ (Γ-technique)

$= \frac{\Gamma(\alpha_1+\alpha_2-1)}{\Gamma(\alpha_1-1)}\frac{(\alpha_1-1)\Gamma(\alpha_1-1)}{(\alpha_1+\alpha_2-1)\Gamma(\alpha_1+\alpha_2-1)} = \frac{(\alpha_1-1)}{(\alpha_1+\alpha_2-1)} = \frac{a}{a+b}$.

$\Rightarrow Var(X) = \int_0^1 x^2 f_X(x)\,dx - E[X]^2$ where $\int_0^1 x^2 f_X(x)\,dx = \frac{1}{B(a,b)}\int_0^1 x^{a+1}(1-x)^b\,dx$

$= B(a,b)^{-1}\int_0^1 x^{\alpha_1-1}(1-x)^{\alpha_2}$ where $\alpha_1 = a+2$ and $\alpha_2 = b$.

$= B(\alpha_1-2,\alpha_2)^{-1}\frac{\Gamma(\alpha_1)\Gamma(\alpha_2)}{\Gamma(\alpha_1+\alpha_2)}$ by the Γ-technique, and so

$= \frac{\Gamma(\alpha_1-2+\alpha_2)}{\Gamma(\alpha_1-2)\Gamma(\alpha_2)}\frac{\Gamma(\alpha_1)\Gamma(\alpha_2)}{\Gamma(\alpha_1+\alpha_2)} = \frac{\Gamma(\alpha_1-2+\alpha_2)}{\Gamma(\alpha_1-2)}\frac{(\alpha_1-1)(\alpha_1-2)\Gamma(\alpha_1-2)}{(\alpha_1+\alpha_2-1)(\alpha_1+\alpha_2-2)\Gamma(\alpha_1+\alpha_2-2)}$

$= \frac{(\alpha_1-1)(\alpha_1-2)}{(\alpha_1+\alpha_2-1)(\alpha_1+\alpha_2-2)} = \frac{(a+1)(a)}{(a+b+1)(a+b)}$ *

So $Var(X) = \frac{(a+1)(a)}{(a+b+1)(a+b)} - \left(\frac{a}{a+b}\right)^2$.

*: Can we use MGF to find $E[X^2]$ instead? We have $E[X^2] = M_X^{(2)}(0)$

and $M_X(t) = E[e^{tX}] = \int_0^1 e^{tx} f_X(x)\,dx$

$\Rightarrow M_X^{(2)}(t) = \frac{d^2}{dt^2}\int_0^1 e^{tx} f_X(x)\,dx = \int_0^1 \frac{\partial^2}{\partial t^2} e^{tx} f_X(x)\,dx$ (Leibniz by continuity on $[0,1] \Rightarrow$ uniform continuity)

$= \int_0^1 x^2 e^{tx} f_X(x)\,dx$

$= B(a,b)^{-1}\int_0^1 e^{tx} x^{a+1}(1-x)^{b-1}\,dx$ and it's unclear how to continue.

Exercise: Let $X \sim Bin(n,p)$. Find $M_X(t)$ and find the values of t such that $M_X(t)$ is well-defined.

We have $M_X(t) = E[e^{tX}] = \sum_{x=1}^n e^{tx} P_X(x) = \sum_{x=1}^n e^{tx} \cdot \binom{n}{x} p^x (1-p)^{n-x}$

$= \sum_{x=1}^n \binom{n}{x} (pe^t)^x (1-p)^{n-x}$

$= ((pe^t) + (1-p))^n = (pe^t + 1 - p)^n$ by Binomial formula, for $t \in \mathbb{R}$.

10.4 12) Say X has a discrete uniform distribution on $\{a, a+1, \dots, b\}$ with $P_X(x) = \frac{1}{b-a+1}$ for $x \in \{a, \dots, b\}$.

a) Find $M_X(t)$.

We have $M_X(t) = E[e^{tX}] = \sum_{x=a}^b e^{tx} P_X(x) = \sum_{x=a}^b e^{tx}\frac{1}{b-a+1} = \frac{1}{b-a+1}\sum_{x=a}^b e^{tx} = \frac{1}{b-a+1}\frac{e^{at}(1-e^{(b-a+1)t})}{1-e^t}$

$= \frac{1}{b-a+1}\frac{e^{at}-e^{(b+1)t}}{1-e^t}$.

b) Use $M_X(t)$ to find $E[X]$ and $E[X^2]$.

$E[X] = M_X'(0) = \frac{d}{dt}\left(\frac{1}{b-a+1}\frac{e^{at}-e^{(b+1)t}}{1-e^t}\right)\Big|$ too lazy.

Similarly for $M_X''(0)$.

10.4 13) $\mathrm{Im}(X) = \{0, 1, 2\}$. $E[X] = 1$, $E[X^2] = 1.5$.

a) We have $E[X] = \sum_{x=0}^{2} x P_X(x) = P_X(1) + 2P_X(2) = 1$.

$E[X^2] = \sum_{x=0}^{2} x^2 P_X(x) = P_X(1) + 4P_X(2) = 1.5$. Thus $2P_X(2) = 0.5 \Rightarrow P_X(2) = \frac{1}{4}$ and $P_X(1) = \frac{1}{2}$

$\Rightarrow P_X(0) = \frac{1}{4}$.

Thus $M_X(t) = E[e^{tX}] = \sum_{x=0}^{2} e^{tx} P_X(x) = e^{0t}\cdot\frac{1}{4} + e^{t}\cdot\frac{1}{2} + e^{2t}\cdot\frac{1}{4}$.

b) $E[X^3] = M_X^{(3)}(0) = \frac{d^3}{dt^3}\left(\frac{1}{4}e^{2t} + \frac{1}{2}e^{t} + \frac{1}{4}\right)_{t=0} = \left(2e^{2t} + \frac{1}{2}e^{t}\right)_{t=0} = 2 + \frac{1}{2} = \frac{5}{2}$.

$E[X^4] = M_X^{(4)}(0) = \frac{d^4}{dt^4}\left(\frac{1}{4}e^{2t} + \frac{1}{2}e^{t} + \frac{1}{4}\right)_{t=0} = \left(4e^{2t} + \frac{1}{2}e^{t}\right)_{t=0} = \frac{9}{2}$.

c) If $E[X] = a$ and $E[X^2] = b$, then $E[X] \overset{=a}{=} 1\cdot P_X(1) + 2\cdot P_X(2)$ and $E[X^2] \overset{=b}{=} 1\cdot P_X(1) + 4\cdot P_X(2)$

$\Rightarrow P_X(2) = \frac{1}{2}(b-a)$ and $P_X(1) = a - (b-a) = 2a - b$ and $P_X(0) = 1 - \left(\frac{1}{2}(b-a) + 2(a-b)\right)$.

Therefore $P_X(x)$ fully determined.

10.4 16) Let $X_i \overset{\text{iid}}{\sim} N(0,1)$.

c) Let $S_n = \sum_{i=1}^{n} X_i$. Find $M_{S_n}(t)$, and thus the distribution.

We have $M_{S_n}(t) = E[e^{tS_n}] = E[e^{t\sum_{i=1}^{n} X_i}] = E[\prod_{i=1}^{n} e^{tX_i}]$ (indep.)

$\overset{\text{independence}}{=} \prod_{i=1}^{n} E[e^{tX_i}]$ where $E[e^{tX_i}] = M_{X_i}(t) = e^{\frac{t^2}{2}}$ by perious results

$\overset{\text{identical}}{=} (e^{\frac{t^2}{2}})^n = e^{\frac{nt^2}{2}}$.

Notice $Z \sim N(\mu, \sigma^2) \Rightarrow M_Z(t) = M_{\sigma X+\mu}(t) = e^{\mu t} M_X(\sigma t) = e^{\mu t} e^{\frac{\sigma^2 t^2}{2}}$ so $\mu = 0$ and $\sigma^2 = n$, so

by uniqueness of MGF, we have $S_n \sim N(0, n)$.

d) $Z = n^{-1/2}(S_n - n)$. Then $Z = \frac{1}{\sqrt{n}} S_n - \sqrt{n}$ and $S_n = nX$ where $X \sim N(0,1)$

$\Rightarrow Z = \sqrt{n}X - \sqrt{n}$. Therefore $M_Z(t) = M_{\sqrt{n}X - \sqrt{n}}(t) = e^{\sqrt{n}t} M_X(\sqrt{n}t) = e^{\sqrt{n}t} e^{\frac{nt^2}{2}}$ so $\mu = \sqrt{n}$ and $\sigma^2 = n$

by uniqueness of MGF, we have $Z \sim N(\sqrt{n}, n)$.

10.4 18) Say X continuous with $f_X(x) = \frac{1}{\theta^2} x e^{-x/\theta}$ for $x > 0$, $\theta > 0$.

a) Find $M_X(t)$.

We have $M_X(t) = E[e^{tX}] = \int_0^\infty e^{tx} \frac{1}{\theta^2} x e^{-x/\theta} dx = \frac{1}{\theta^2}\int_0^\infty x e^{tx} e^{-x/\theta} dx$

$= \theta^{-2}\int_0^\infty x e^{-x(\theta^{-1} - t)} dx$ (Γ-technique: $\int_0^\infty x^{\alpha-1} e^{-x/\beta} dx = \beta^\alpha \Gamma(\alpha)$)

$\overset{\alpha=2,\ \beta=(\theta^{-1}-t)^{-1}}{=} \theta^{-2}\int x^{\alpha-1} e^{-x/\beta} dx = \theta^{-2}\beta^\alpha \Gamma(\alpha)$ for $\beta > 0 \Rightarrow (\theta^{-1} - t)^{-1} > 0$

$\Rightarrow \theta^{-1} - t > 0$

$\Rightarrow t < \theta^{-1}$ and $\theta > 0$

$\Rightarrow M_X(t)$ exists in $B(0, \theta^{-1})$.

$= \theta^{-2}(\theta^{-1} - t)^{-2}\Gamma(2)$ ($\Gamma(2) = 1$)

$= \theta^{-2}(\theta^{-1} - t)^{-2}$.

b) Say $X \sim \mathrm{Exp}(\theta) \perp Y \sim \mathrm{Exp}(\theta)$. Find the distribution of $S = X + Y$.

We have $M_{X+Y}(t) = M_X(t) M_Y(t)$ by independence.

$$M_X(t) = E[e^{tX}] = \int_0^\infty e^{tx}\theta e^{-\theta x}dx$$

$$= \theta\int_0^\infty e^{-x(\theta-t)}dx$$

$$\overset{\substack{\alpha=1\\ \beta=(\theta-t)^{-1}\\ \rightsquigarrow t<\theta}}{=} \theta\int_0^\infty x^{\alpha-1}e^{-x/\beta}dx = \theta\beta^\alpha\Gamma(\alpha) = \theta(\theta-t)^{-1}\cancel{\Gamma(1)}^{\,1} = \theta(\theta-t)^{-1}$$

$$\Rightarrow M_{X+Y}(t) = \theta^2(\theta-t)^{-2}$$

From (a) we define $f(x;\theta) = \frac{1}{\theta^2}xe^{-x/\theta}$ with MGF $\theta^{-2}(\theta^{-1}-t)^{-2}$

$\Rightarrow f(x;\theta)$ defines $2\text{Exp}(\theta^{-1})$ since $2\text{Exp}(\theta)$ has MGF $\theta(\theta-t)^{-1}$

$\Rightarrow f_S(x) = \theta^2 xe^{-x\theta}$.

10.4 19) Recall the MGF of a Bernoulli random variable X is $M_X(t) = (1-p+pe^t)^n$ for $t\in\mathbb{R}$.

Find the MGF of $Z_n = \frac{X-np}{\sqrt{np(1-p)}}$ as $n\to\infty$ (where p fixed).

We have:

$$M_{Z_n}(t) = M_{\frac{X-np}{\sqrt{np(1-p)}}}(t) = e^{-\sqrt{np(1-p)}^{-1/2}t}M_X(\sqrt{np(1-p)}\,t)$$

$$= e^{-\sqrt{np(1-p)}^{-1/2}t}(1-p+pe^{\sqrt{np(1-p)}t})^n$$

$$= ((e^{-\frac{1}{n}\sqrt{np(1-p)}^{-1/2}t})(1-p+pe^{\sqrt{np(1-p)}t}))^n$$

10.4 20) The number of trades N of stock XXX in a day follows $\text{Poi}(1)$. At each trade (i^{th} trade), the change in stock price is X_i, where $X_i\sim N(0,\tau^2)$. Say X_i are independent of each other and of N. Find the MGF of the total change in stock price over the day.

We want $M_X(t)$ where $X = \sum_{i=1}^N X_i$. Thus $M_X(t) = E[e^{tX}] = E[e^{t\sum_{i=1}^N X_i}] \overset{\text{independence}}{=} \prod_{i=1}^N E[e^{tX_i}] \overset{\text{identical}}{=} (M_{X_i}(t))^n$

$$= (e^{\frac{\tau^2t^2}{2}})^n = e^{\frac{n\tau^2t^2}{2}}$$

so $X\sim N(0,n\tau^2)$ by MGF uniqueness.

9.7.1) In a row of 25 switches, each is considered to be "on" or "off". Define $P(\text{"on"}) = 0.6$ for a switch, independently of other switches. Find the mean and variance of #unlike pairs among the 24 adjacent switch pairs.

In the desired outcome, there are only two possible configurations: ON, OFF, ON, ..., ON so say U = all pairs unlike.
OFF, ON, OFF, ..., OFF

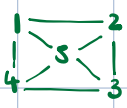
Define $X_i = \begin{cases} 1 \text{ if } i^{th} \text{ switch ON} \\ 0 \text{ if } i^{th} \text{ switch OFF} \end{cases}$. Say $Y_i = \begin{cases} 1 \text{ if } X_i \neq X_{i+1} \\ 0 \text{ if } X_i = X_{i+1} \end{cases}$.

Then $P(U) = P(X_1=1, X_2=0, \dots, X_{25}=1) + P(X_1=0, X_2=1, \dots, X_{25}=0)$

$= P(X_1=1)P(X_2=0)\cdots P(X_{25}=1) + P(X_1=0)P(X_2=1)\cdots P(X_{25}=0)$ by independence

$= 0.6^{13}0.4^{12} + 0.6^{12}0.4^{13}$.

9.7 36)



Define $X_i = 1$ if i^{th} island isolated. We have $X_i = 1$ with probability p^3 for $i=1,\dots,4$ (by independence), p^4 for $i=5$.

Let $Y = \sum_{i=1}^{5} X_i$ be the number of isolated islands.

Then $E[Y] = E[\sum_{i=1}^{5} X_i] = \sum_{i=1}^{5} E[X_i] = 4E[X_1] + E[X_5] = 4p^3 + p^4$.

$\text{Var}(Y) = \text{Var}(\sum_{i=1}^{5} X_i) = \sum_{i=1}^{5} \text{Var}(X_i) + 2\sum_{i<j}\sum \text{Cov}(X_i, X_j)$. Really annoying.

Exercise: Say X, Y are continuous random vars. with joint function $f_{X,Y}(x,y) = f(x,y) = e^{-x-y}$ for $x, y > 0$.

a) Find $M_{X,Y}(t_1, t_2) = M(t_1, t_2)$ and specify the ranges of t_1, t_2 for which it exists.

We have $M_{X,Y}(t_1,t_2) = E[e^{t_1X+t_2Y}] = \int_x\int_y e^{t_1x+t_2y} f_{X,Y}(x,y)\,dy\,dx = \int_0^\infty\int_0^\infty e^{t_1x}e^{t_2y}e^{-x}e^{-y}\,dy\,dx = \int_0^\infty e^{t_1x}e^{-x}\int_0^\infty e^{t_2y}e^{-y}\,dy\,dx$

$= \int_0^\infty e^{x(t_1-1)}\,dx \int_0^\infty e^{y(t_2-1)}\,dy$

$= \left[\frac{e^{x(t_1-1)}}{t_1-1}\right]_0^\infty \left[\frac{e^{y(t_2-1)}}{t_2-1}\right]_0^\infty$ $\quad t_1<1, t_2<1$

$= \frac{1}{(1-t_1)(1-t_2)}$.

b) What is $M_X(t)$ and what is its marginal distribution?

$M_X(t) = M_{X,Y}(t,0) = \frac{1}{1-t}$. Since Gamma$(\alpha,\beta)$ has MGF $\frac{1}{(1-\beta t)^\alpha}$ we have $X \sim$ Gamma$(1,1)$.

c) Are X and Y independent?

We have $X \perp Y \Leftrightarrow M_X M_Y = M_{X,Y}$. We have this property clearly, so $X \perp Y$.

Exercise: Say $X \sim N(0,1)$.

a) Find $M_X(t)$.

We have $M_X(t) = E[e^{tX}] = \int_{\mathbb{R}} e^{tx} f_X(x)\,dx$

$= \int_{-\infty}^{\infty} e^{tx}\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}\,dx$

$= \int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}}e^{tx-\frac{x^2}{2}}\,dx = \int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}}e^{\frac{t^2}{2}+\frac{x^2}{2}-tx-\frac{t^2}{2}}\,dx$

$= \int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}}e^{\frac{t^2}{2}}e^{-\frac{1}{2}(x-t)^2}\,dx = e^{\frac{t^2}{2}}\underbrace{\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}}e^{-\frac{(x-t)^2}{2}}\,dx}_{1} = e^{\frac{t^2}{2}}$ for all $t \in \mathbb{R}$.

since p.d.f. of $N(t,1)$

b) Say $Z \sim N(\mu, \sigma^2)$. Find M_Z using M_X.

$Z = \sigma X + \mu$ by reparametrization. Then $M_Z(t) = M_{\sigma X+\mu}(t) = e^{\mu t}M_X(\sigma t) = e^{\mu t}e^{\frac{(\sigma t)^2}{2}} = e^{\mu t}e^{\frac{\sigma^2 t^2}{2}}$.

Exercise: If $X_n \xrightarrow{p} a$ and $Y_n \xrightarrow{p} b$, then $X_nY_n \xrightarrow{p} ab$. (Hint: Show $Y_n \xrightarrow{p} b \Rightarrow \exists M>0$ such that $\lim_{n\to\infty} P(|Y_n|<M) = 1$.)

We prove the hint. Say $Y_n \xrightarrow{p} b$, so by definition, for all $\varepsilon>0$, we have $\lim_{n\to\infty} P(|Y_n - b| \geq \varepsilon) = 0$.

Notice $Y_n \xrightarrow{p} b \Rightarrow Y_n \xrightarrow{D} b$, so $F_{Y_n}(x) \to F_b(x)$ at all continuity points of $F_b(x)$ (so $x \in \mathbb{R}\setminus\{b\}$).

$$\Rightarrow \lim_{n\to\infty} F_{Y_n}(x) = F_b(x) \text{ for } x \in \mathbb{R}\setminus\{b\}.$$

Thus we have $F_{Y_n}(M) \to F_b(M) = \begin{cases} 0 & \text{for } M<b \\ 1 & \text{for } M \geq b \end{cases}$ so pick $M>b$. Satisfied.

$$F_{Y_n}(-M) \to F_b(-M) = \begin{cases} 0 & \text{for } -M<b \\ 1 & \text{for } -M \geq b \end{cases}$$

Therefore $\exists M, N$ such that $\lim_{n\to\infty} P(|Y_n|<N) = 1$ and $\lim_{n\to\infty} P(|X_n|<M) = 1$.

We want $\lim_{n\to\infty} P(|X_nY_n - ab| \geq \varepsilon) = 0$

We have $P(|X_nY_n - ab| \geq 0) = P(|X_nY_n + aY_n - aY_n - ab| \geq \varepsilon)$

$$\leq P(|(X_n-a)Y_n| + |a(Y_n-b)| \geq \varepsilon) \quad \text{since } |\alpha+\beta| \leq |\alpha|+|\beta|$$

$$\leq P(|(X_n-a)Y_n| \geq \tfrac{\varepsilon}{2}) + P(|a(Y_n-b)| \geq \tfrac{\varepsilon}{2})$$

$\cdots$ algebra.

Exercise Say $(X_i)_{i\in\mathbb{N}}$ is a sequence such that $E[X_n] = \mu_n$ and $Var(X_n) = \sigma_n^2$. If $\mu_n \to \mu$ and $\sigma_n^2 \to 0$, then $X_n \xrightarrow{p} \mu$.

By Markov's inequality we have $P(|X_n - \mu| \geq C) \leq \dfrac{E[|X_n-\mu|^k]}{C^k}$ for any $C, k>0$.

Therefore $P(|X_n-\mu| \geq \varepsilon) \leq \dfrac{E[|X_n-\mu|^2]}{\varepsilon^2} = \dfrac{Var(X_n)}{\varepsilon^2} = \dfrac{\sigma_n^2}{\varepsilon^2} \xrightarrow{n\to\infty} 0$. Therefore $X_n \xrightarrow{p} \mu$ by definition.

Exercise: Say $X \sim Poi(\mu)$ where $\mu \in \mathbb{Z}$ large. Show that $\dfrac{X-\mu}{\sqrt{\mu}} \xrightarrow{D} N(0,1)$ and use the approximation to find $P(X>12)$.

We use the CLT. Notice that if $X_1 \sim Poi(\lambda_1) \perp X_2 \sim Poi(\lambda_2)$ then $X_1 + X_2 \sim Poi(\lambda_1+\lambda_2)$.

Therefore since $\mu \in \mathbb{Z}$ large, consider $X_i \overset{iid}{\sim} Poi(1)$ so $\sum_{i=1}^{\mu} X_i \sim Poi(\sum_{i=1}^{\mu} 1) = Poi(\mu)$.

Since $E[X_i] = 1 < \infty$ and $Var(X_i) = 1 < \infty$ we have

$$\overset{CLT}{\Rightarrow} \frac{\sqrt{\mu}(\bar{X}_\mu - 1)}{1} \xrightarrow{D} N(0,1)$$

(mean; standard deviation)

$$= \sqrt{\mu}\left(\frac{1}{\mu}\sum_{i=1}^{\mu} X_i - 1\right) = \sqrt{\mu}\left(\frac{1}{\mu}X - 1\right) = \frac{1}{\sqrt{\mu}}(X-\mu) \xrightarrow{D} N(0,1).$$

We want $P(X>12)$. Notice $P(X>12) = 1 - P(X \leq 12) = 1 - F_X(12)$ and we have $F_X(x) \to F_Z(x)$ at all $x \in \mathbb{R}\setminus\mathbb{Z}$.

Thus by using a continuity correction, we have $1 - F_X(12) = 1 - F_X(12.5)$

$$\approx 1 - F_Z(12.5) \quad \text{where } \tfrac{1}{\sqrt{\mu}}(X-\mu) \xrightarrow{D} N(0,1) \Rightarrow X \xrightarrow{D} Z \sim N(\mu, \mu).$$

$$= 1 - F_Y\left(\frac{12.5-\mu}{\sqrt{\mu}}\right) \quad \text{where } Y \sim N(0,1).$$