Math 170A: Probability Theory

2023-24
Winter 24

Instructor: Rowan Killip (+ James Hogan)

Textbook: G. Grimmett - Probability

Topics: Probability spaces, probabilities, conditional probability, discrete random variables, expectation & variance, continuous random variables, multiple random variables

Table of Contents

Sample Spaces

1/18/24
Lecture 1

Def: Given an experiment, we define an elementary outcome to be any possible result of the experiment - even "wacky" ones (e.g., "No response", "experiment caught fire", etc.).

Sample Spaces

Def. Given an experiment, we define its sample space Ω to be the set of all elementary outcomes.

Ex: Rolling a dice $\rightarrow \Omega = \{1, 2, 3, 4, 5, 6\}$

Flipping a coin $\rightarrow \Omega = \{\text{Heads}, \text{Tails}\}$

Lifetime of a nucleus $\rightarrow \Omega = [0, +\infty)$ ↳ used to generate random #s

(Notation: Denote an outcome as $\omega \in \Omega$.)

Given an experiment, there may be many valid sample spaces.

Ex: Height of a person: i) $\Omega = [0, +\infty)$, simple case (even if ∄ a person w/ height 0, can still include in Ω w/o issue)

ii) $\Omega = \{1/8, 1/4, 3/8, \ldots\}$, e.g. if measurements are of finite precision

iii) $\Omega = [0, +\infty) \cup \{\text{"No response"}\}$, may appear in real-world statistics

iv) Ω = population [modern approach]

Observe: In (iv), the source of randomness is not the height of a person; the source of randomness is choosing a person out of the population. (Height may be det. by age, genetics, etc.)

→ Represent height as a random variable: a function from Ω / with domain Ω

Ex: Height as a function $H: \Omega \rightarrow [0, +\infty)$

Random Variables

- Random variables need not be numbers - can find random variables for names, birthdays, etc.
- Transformations of random variables can themselves be random variables
  - Ex: $H': \omega \in \Omega \mapsto H(\omega)^2 + 4$ is a random variable

Probabilities

1/8/24

Lecture 1 (cont.)

May want to find/assign probabilities; 2 approaches:

1) Naïve: Assign probabilities to individual outcomes $\omega \in \Omega$

→ Problem: Only works for countable Ωs, i.e. finite or countably infinite sample spaces

Ex: Taking $\Omega = [0, +\infty)$ for lifetime of a nucleus [random process]

→ Ω has infinite elements, but probability for any particular $\omega \in [0, +\infty)$ is 0

2) Better: Assign probabilities to events - subsets of Ω (subject to restrictions)

(*) Need to place restrictions on what subsets can be events, since assigning probabilities to all subsets would cause issues

Def: A collection $\mathcal{F}$ of subsets of Ω is called a σ-algebra if the following hold:

i) $\emptyset \in \mathcal{F}$ and $\Omega \in \mathcal{F}$ — note: "σ" usually implies "countable" in math

ii) If $A \in \mathcal{F}$, $A^c \in \mathcal{F}$ $[A^c = \Omega \setminus A]$

iii) If $A_n \in \mathcal{F}$ for $n \in \mathbb{N}$, $\bigcup_{n=1}^{\infty} A_n \in \mathcal{F}$ [countable unions]

↓

Def: Probabilities

Given sample space Ω and σ-algebra $\mathcal{F}$, a probability [measure] is an assignment $\mathbb{P}: \mathcal{F} \to \mathbb{R}$ subject to:

i) $\mathbb{P}(A) \geq 0 \ \forall A \in \mathcal{F}$

ii) $\mathbb{P}(\emptyset) = 0$; $\mathbb{P}(\Omega) = 1$

iii) If $A_n \in \mathcal{F}$ are disjoint for $n \in \mathbb{N}$, $\mathbb{P}(\bigcup_{n=1}^{\infty} A_n) = \sum_{n=1}^{\infty} \mathbb{P}(A_n)$. [countable/σ-additivity]

Note: The infinite sum $\sum_{n=1}^{\infty} \mathbb{P}(A_n)$ is guaranteed to converge because all terms $\mathbb{P}(A_n) \geq 0$ and sum is bounded above by $\mathbb{P}(\Omega) = 1$.

Probabilities (cont.)

1/10/24

Lecture 2

Remark: The countable union property of a σ-algebra also applies to finite unions.

(*) Ex: Let $A_1, A_2 \in \mathcal{F}$. Set $A_3 = A_4 = \dots = A_n = \dots = \emptyset$; then $\bigcup_{n=1}^{\infty} A_n = A_1 \cup A_2 \in \mathcal{F}$. (Prove for all $A_1, \dots, A_n$ by induction.)

Probability Measures

(*) Ex: If Ω is finite & non-empty, $\mathbb{P}(A) = \frac{|A|}{|\Omega|}$ defines a probability measure on all subsets of Ω

(cannot take uncountable sums unless all but countably many terms are 0.)

(*) Ex: If Ω is countable and $\exists\ p: \Omega \to [0,1]$ so that $\sum_{\omega \in \Omega} p(\omega) = 1$, then $\mathbb{P}(A) = \sum_{\omega \in A} p(\omega)$ defines a probability measure on all subsets of Ω.

(*) Theorem (Lebesgue): There exists a σ-algebra of subsets of $[0,1]$ containing all intervals $I \subset [0,1]$, and corresponding probability $\mathbb{P}$ satisfying $\mathbb{P}([a,b]) = b-a$ $\left[= \frac{b-a}{1-0}\right]$.

(*) Borel σ-algebra: smallest σ-algebra of $[0,1]$ containing all intervals

Observations

Recall (DeMorgan's Laws): i) $(A \cup B)^c = A^c \cap B^c$ ~ in event terminology: $(A \cup B)$ = "A happens or B happens" [A or B]; $(A^c \cap B^c)$ = "A doesn't happen and B doesn't happen"

ii) $(A \cap B)^c = A^c \cup B^c$ ~ $(A \cap B)$ = "A happens and B happens" [A and B]

→ Observe: i) per $(A \in \mathcal{F} \Rightarrow A^c \in \mathcal{F})$ and [σ-additivity], $\mathbb{P}(A) + \mathbb{P}(A^c) = \mathbb{P}(A \cup A^c) = 1$; in particular, $(\mathbb{P}(A) \geq 0\ \forall A \in \mathcal{F})$ and $(\mathbb{P}(A) = 1 - \mathbb{P}(A^c)) \Rightarrow 0 \leq \mathbb{P}(A) \leq 1$.

ii) per $(A, B \in \mathcal{F} \Rightarrow (A \cup B) \in \mathcal{F})$ and $(A \in \mathcal{F} \Rightarrow A^c \in \mathcal{F})$, $A^c \cup B^c = (A \cap B)^c \in \mathcal{F} \Rightarrow (A \cap B) \in \mathcal{F}$.

iii) if $A, B \in \mathcal{F}$, then $A \setminus B = A \cap B^c \in \mathcal{F}$ [$(A \setminus B)$ = "A but not B"]

iv) if $A, B \in \mathcal{F}$, then $\mathbb{P}(A \cup B) = \mathbb{P}(A \underset{\text{disj}}{\cup} [B \setminus A]) = \mathbb{P}(A) + \mathbb{P}(B \setminus A) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$

(*) $\mathbb{P}(B) = \mathbb{P}([B \cap A] \underset{\text{disj}}{\cup} [B \setminus A])$

(*) $\mathbb{P}(A \cup B \cup (\cup \dots)) \to$ use inclusion-exclusion

v) if $A \subseteq B \in \mathcal{F}$, then $\mathbb{P}(B) = \mathbb{P}(A \cup [B \setminus A]) \geq \mathbb{P}(A)$ [$A \subseteq B$ = "A implies B"]

vi) if $A_n \in \mathcal{F}$ for $1 \leq n \leq N$, then $\bigcup_{n=1}^{N} A_n \in \mathcal{F}$ and $\mathbb{P}(\bigcup_{n=1}^{N} A_n) = \mathbb{P}(A_1) + \mathbb{P}(\bigcup_{n=2}^{N} A_n) \leq \dots \leq \mathbb{P}(A_1) + \mathbb{P}(A_2) + \dots$ (by induction)

(*) Lady Tasting Tea

1/11/24
Disc 1

Probabilities (Review)

- State space: set of all outcomes of an "experiment" [a procedure w/ multiple possible outcomes]
- Event: a subset $E \subseteq \Omega$, which may be given a probability [additional condition: $E \in \mathcal{F}$]
  - i) Finite sets Ω: all subsets $E \subseteq \Omega$ can be events
  - ii) Infinite sets Ω: only measurable subsets $E \subseteq \Omega$ can be events [preview: measure theory]
    - Probabilities can only be aggregated using countable sums

(*) Ex: Lady Tasting Tea

Origin: Test to determine whether a person could distinguish between 2 types of prepping tea (A & B)

Format: Ask a person to identify the 4 "A" cups and 4 "B" cups out of 8 randomized cups

→ Observed results: All 8 cups were guessed correctly

Want to assess the null hypothesis: assuming the claim is false (i.e. no ability to distinguish A, B), what would be the probability of obtaining the observed results?

Defining the State Space: i) Ω = # correct [i.e. $\Omega = \{0, 1, \dots, 8\}$]

- Problem: Loses information (e.g. which cups, specifically, were right/wrong?)

ii) Ω = selections of A, B cups [e.g. $(1,2,4,7) \times (3,5,6,8)$]

- Observe: Contains redundant info - B selections are already implied by A selections, so we can reduce Ω w/o losing information

→ iii) Ω = selections of A cups [e.g. $(1,2,4,7)$]

# A correct	# ways to guess
0	$\binom{4}{0} \cdot \binom{4}{4} = 1$
1	$\binom{4}{1} \cdot \binom{4}{3} = 16$
2	$\binom{4}{2} \cdot \binom{4}{2} = 36$
3	$\binom{4}{3} \cdot \binom{4}{1} = 16$
4	$\binom{4}{4} \cdot \binom{4}{0} = 1$

($\binom{4}{1}$ from A, $\binom{4}{3}$ from B)

→ total # possible choices: $\binom{8}{4} = 70$ possible 4-subsets of $\{1, \dots, 8\}$

null hypothesis: all choices have equal probability (random guessing)

→ $\mathbb{P}(\text{all 4 correct}) = \frac{\#\text{ choices w/ 4 correct}}{\text{total \# choices}} = \frac{1}{70} \approx 1.7\%$ → reject null hypothesis

(implicitly: $\{\omega \in \Omega : \omega \text{ has all 4 correct}\}$)

Conditional Probability

1/12/24
Lecture 3

Theorem: Continuity/Monotone Convergence of Events

i) Given events $A_1 \subseteq A_2 \subseteq \dots \subseteq A_n \subseteq \dots$ for $n \in \mathbb{N}$, $\mathbb{P}(\bigcup_{n=1}^{\infty} A_n) = \lim_{n\to\infty} \mathbb{P}(A_n)$.

ii) Given events $A_1 \supseteq A_2 \supseteq \dots \supseteq A_n \supseteq \dots$ for $n \in \mathbb{N}$, $\mathbb{P}(\bigcap_{n=1}^{\infty} A_n) = \lim_{n\to\infty} \mathbb{P}(A_n)$.

(*) **Proof:** i) Define sets $D_1 = A_1$, $D_n = A_n \setminus A_{n-1}$ for $n \geq 2$. Since D_n's are disjoint, then per σ-additivity: $\mathbb{P}(\bigcup_{n=1}^{\infty} D_n) = \sum_{n=1}^{\infty} \mathbb{P}(D_n) = \lim_{N\to\infty} \sum_{n=1}^{N} \mathbb{P}(D_n)$

Since $\bigcup_{n=1}^{N} D_n = A_n$, then $\mathbb{P}(\bigcup_{n=1}^{\infty} D_n) = \mathbb{P}(\bigcup_{n=1}^{\infty} A_n) = \lim_{N\to\infty} \sum_{n=1}^{N} \mathbb{P}(D_n) = \lim_{n\to\infty} \mathbb{P}(A_n)$.

ii) Observe: $A_1^c \subseteq A_2^c \subseteq \dots \subseteq A_n^c \subseteq \dots$. Then per i), $\mathbb{P}(\bigcup_{n=1}^{\infty} A_n^c) = \lim_{N\to\infty} \mathbb{P}(A_N^c)$.

Per axioms, $\mathbb{P}(\bigcup_{n=1}^{\infty} A_n^c) = \mathbb{P}([\bigcap_{n=1}^{\infty} A_n]^c) = 1 - \mathbb{P}(\bigcap_{n=1}^{\infty} A_n)$. Similarly, $\lim_{n\to\infty}(A_N^c) = \lim_{n\to\infty}(1 - \mathbb{P}(A_n))$

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

(*) **Observation:** This implies that limits of probabilities exist.

Theorem: Conditional Probability

Given a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and an event B $[\in \mathcal{F}]$ with $\mathbb{P}(B) > 0$, then the function defined by:

$$A \longmapsto \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} \qquad \left[\text{“}\frac{\mathbb{P}(A \text{ and } B)}{\mathbb{P}(B)}\text{”}\right]$$

defines a probability measure on $(\Omega, \mathcal{F})$.

→ This measure, denoted $\mathbb{P}(A \mid B)$ [probability of A given B], is called the conditional probability given B.

↳ **Observation:** $\mathbb{P}(A \cap B) = \mathbb{P}(A \mid B) \cdot \mathbb{P}(B)$

Conditional Probability (cont.)

1/12/24
Lecture 3 (cont.)

(*) **Proof (Conditional Probability):** Want to show that $\mathbb{P}(\cdot|B)$ is a probability measure.

i) $\mathbb{P}(\Omega|B) = \frac{\mathbb{P}(\Omega \cap B)}{\mathbb{P}(B)} = \frac{\mathbb{P}(B)}{\mathbb{P}(B)} = 1$; $\mathbb{P}(\emptyset|B) = \frac{\mathbb{P}(\emptyset \cap B)}{\mathbb{P}(B)} = \mathbb{P}(\emptyset)/\mathbb{P}(B) = 0$. [trivial]

ii) $\mathbb{P}(A \cap B) \geq 0$, $\mathbb{P}(B) > 0 \Rightarrow \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} \geq 0$. [trivial]

(*) Do not take $\mathbb{P}(A|B)$ for events B with $\mathbb{P}(B) = 0$!!

iii) If $A_1, \dots, A_n, \dots$ disjoint for $n \in \mathbb{N}$:

$$\frac{\mathbb{P}\left(\left[\bigcup_{n=1}^{\infty} A_n\right] \cap B\right)}{\mathbb{P}(B)} = \frac{\mathbb{P}\left(\bigcup_{n=1}^{\infty} [A_n \cap B]\right)}{\mathbb{P}(B)} = \sum_{n=1}^{\infty} \frac{\mathbb{P}(A_n \cap B)}{\mathbb{P}(B)} = \sum_{n=1}^{\infty} \mathbb{P}(A_n|B)$$

Remarks: Probability can be interpreted as a mathematical description of ignorance [Bayes], e.g. ignorance & initial conditions. In this setting, $\mathbb{P}(\cdot|B)$ represents a new state of ignorance, after observing B happening. Namely, $\mathbb{P}(\cdot|B)$ does not "re-weight" events, i.e.: for events $A_1, A_2 \subseteq B$, $\frac{\mathbb{P}(A_1)}{\mathbb{P}(A_2)} = \frac{\mathbb{P}(A_1|B)}{\mathbb{P}(A_2|B)}$.

(*) "Prior distribution" - Bayesian notion of a state of complete ignorance [of the world]

→ Q: What happens if there are multiple observations?

$$\mathbb{P}((A|B_1)|B_2) = \frac{\mathbb{P}(A \cap B_2|B_1)}{\mathbb{P}(B_2|B_1)} = \frac{\mathbb{P}(A \cap B_1 \cap B_2)}{\mathbb{P}(B_1)} \cdot \frac{\mathbb{P}(B_1)}{\mathbb{P}(B_1 \cap B_2)} = \underline{\mathbb{P}(A|B_1, B_2)}$$

→ Observation: The order in which B_1, B_2 occur is irrelevant. [(*) "$\mathbb{P}((\cdot|B_1)|B_2)$" notation doesn't exist!]

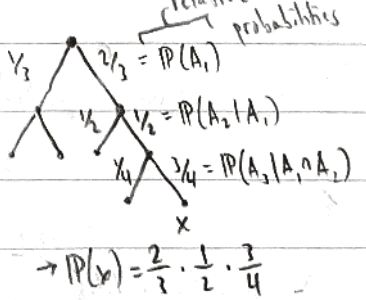
In some cases, conditional/relative probabilities can be used to understand the overall distribution [absolute probabilities].

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

$\rightarrow \mathbb{P}(A_1 \cap A_2 \cap \dots \cap A_n) = \mathbb{P}(A_n|A_1 \cap \dots \cap A_{n-1}) \cdot \mathbb{P}(A_1 \cap \dots \cap A_{n-1})$

$$\rightarrow \boxed{\mathbb{P}(A_1 \cap \dots \cap A_n) = \mathbb{P}(A_n|A_1 \cap \dots \cap A_{n-1}) \cdot \dots \cdot \mathbb{P}(A_2|A_1) \cdot \mathbb{P}(A_1)}$$

("multiplication rule" - by induction)



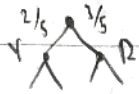
Bayes' Rule

1/17/24

Lecture 4

(*) Ex: Probability

AMC8: Given 2 yellow & 3 red counters [distinguishable] drawn without replacement, what is the probability of drawing both yellows without exhausting the reds? (or not)

i) Naïve soln.: Draw a tree 

↝ sum ℙs of tree leaves: $\frac{3}{5}$ [simple method]

ii) Alt.: Observe that "draw 5 randomly" equivalent to "shuffle & look at resulting order"

→ redefine Ω = set of possible permutations after shuffling ($|\Omega| = 5!$)

→ {Y exhausted} = {the last element is red} → $\mathbb{P}(\text{red at bottom}) = \frac{4!+4!+4!}{5!} = \frac{3}{5}$.

Partitions

Def: A partition of Ω is a countable collection of disjoint subsets $B_j \subseteq \Omega$ such that $\bigcup_j B_j = \Omega$.

Theorem: If $\{B_j\}$ form a partition of Ω with $\mathbb{P}(B_j) > 0$ for all j, then $\forall A \in \mathcal{F}$:

$$\mathbb{P}(A) = \sum_j \mathbb{P}(A \cap B_j) = \sum_j \mathbb{P}(A|B_j)\mathbb{P}(B_j)$$

Def: Bayes' Rule

Given events A, B with $\mathbb{P}(A), \mathbb{P}(B) > 0$:

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

Idea: We can imagine that B, B^c represent an underlying truth, and that A represents an attempt to measure it. If $\mathbb{P}(A|B), \mathbb{P}(A|B^c)$ can be found (via science, e.g.), then Bayes' Rule indicates how to update our knowledge of the truth given a measurement.

(Relies on knowing prior probabilities of B and/or B^c)

(*) The Boy-Girl Paradox

1/18/24

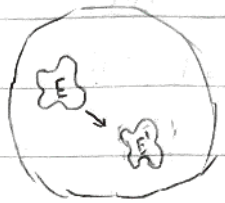
Disc 2

(*) Measure (Motivation)

Let Ω be the unit disk in $\mathbb{R}^2$, w/ $\mathbb{P}$ distributed uniformly across Ω.

Let $E \subseteq \Omega$ be an event; we want $\mathbb{P}(E)$ to be translationally invariant,

i.e $\mathbb{P}(E) = \mathbb{P}(E')$ for any translation of E, since $\mathbb{P}$ is uniform.



Problem: If we say this is true for all sets, we can find "bad subsets" s.t. translational invariance breaks infinite sums (and vice versa)

→ Solution: Restrict subsets under consideration to only "good subsets"

(*) σ-Algebra Example

Let $f: \Omega_1 \to \Omega_2$, and define σ-algebra Σ on Ω_2; show that $f^{-1}(\Sigma)$ is a σ-algebra on Ω_1.

Check axioms: i) $f^{-1}(\emptyset) = \emptyset$, $f^{-1}(\Omega_2) = \Omega_1 \to \emptyset, \Omega_1 \in f^{-1}(\Sigma)$

ii) Let $A \in f^{-1}(\Sigma)$. Per definition, $A = f^{-1}(B)$ for some B; $f^{-1}(B^c) = A^c \to A^c \in f^{-1}(\Sigma)$.

iii) Let $A_n \in f^{-1}(\Sigma)$. Define B_n s.t. $f^{-1}(B_n) = A_n$; then $B = \bigcup_n B_n \in \Sigma \to \bigcup_n A_n = \bigcup_n f^{-1}(\bigcup B_n) = f^{-1}(B)$. □

(*) The Boy-Girl Paradox

Say Ms. Smith has two children (Ω genders: M/F, $\mathbb{P}$ 50% each). What is $\mathbb{P}$(both are boys) given that:

i) the youngest is a boy? → $\mathbb{P}$(both boys) = $\mathbb{P}$(oldest is a boy) = $\frac{1}{2} = 50\%$

ii) at least one is a boy?

$\Omega = \{BB, BG, GB, GG\}$

	B	G
B	BB	BG
G	GB	GG

→ $\mathbb{P}(\text{both boys} \mid \geq 1 \text{ boy}) = \frac{1/4}{3/4} = \frac{1}{3} \approx 33\%$

iii) at least one is a boy born on a Tuesday?

$\Omega = \{\{B, G\} \times \{SMTWRFS\} \times \{B, G\} \times \{SMTWRFS\}\}$

→ $\mathbb{P}$(both boys | at least one born on a Tuesday)

$= \mathbb{P}(B, T, B, \cdot) + \mathbb{P}(B, \cdot, B, T) - \mathbb{P}(B, T, B, T)$ [Inclusion-Exclusion]

$= \frac{13}{27} \approx 48\%$

(*) Intuition:

$\mathbb{P}$(one boy on Tues | one boy) = $\frac{1}{7} \approx 14\%$

$\mathbb{P}$(one boy on Tues | two boys) = $\frac{13}{49} \approx 27\%$

Likelihoods

1/19/24

Lecture 5

Bayes' Rule (ii):

Given a [countable] partition $\{B_j\}$ of Ω with $\mathbb{P}(B_j)>0\ \forall j$ and event A with $\mathbb{P}(A)>0$:

$$\mathbb{P}(B_j|A)=\frac{\mathbb{P}(B_j\cap A)}{\mathbb{P}(A)}=\frac{\mathbb{P}(A|B_j)\mathbb{P}(B_j)}{\sum_k \mathbb{P}(A|B_k)\mathbb{P}(B_k)} \quad \text{[General case]}$$

Likelihoods

(*) Ex: Let a set of lions be partitioned into J(uvenile), M(ale adult), and F(emale adult).

Let $\mathbb{P}(\text{weighs over 200 kg}|J)=1/100$, $\mathbb{P}(\cdot|F)=1/50$, $\mathbb{P}(\cdot|M)=1/10$.

→ Observation: The $\mathbb{P}$'s do not add up to 1 → $\mathbb{P}(B|\cdot)$ is not a probability measure.

→ Call $\mathbb{P}(B|\cdot)$ a likelihood (vs. $\mathbb{P}(\cdot|A)$, which is a probability measure).

(*) Ex: Rolling Dice

Roll 2 dice independently, such that all 36 outcomes are equally likely; let 1 die be red, 1 blue.

Let $A_j=\{\text{red dice rolls a } j\}$, $B=\{\text{sum of dice is 6}\}$, $C=\{\text{blue dice rolls an even \#}\}$.

			← via Bayes' Rule →		
j	$\mathbb{P}(A_j)$	$\mathbb{P}(B\|A_j)$	$\mathbb{P}(A_j\|B)$	$\mathbb{P}(B\cap C\|A_j)$	$\mathbb{P}(A_j\|B\cap C)$
1	1/6	1/6	1/5	0	0
2	1/6	1/6	1/5	1/6	1/2
3	1/6	1/6	1/5	0	0
4	1/6	1/6	1/5	1/6	1/2
5	1/6	1/6	1/5	0	0
6	1/6	0	0	0	0

"prior probabilities" — likelihoods — "posterior probabilities"

(*) Estimation: "Given that B occurred, find the A_j that maximizes $\mathbb{P}(A_j|B)$."

Independence

1/19/24
Lecture 5 (cont.)

Def: Independence

We say that two events A, B are [statistically] independent if $\boxed{\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)}$

Remarks: i) Independence is a probabilistic statement, not a real-life "mechanism".

(*) Ex: Rolling a dice - $A = \{\text{even}\}$, $B = \{1 \text{ or } 2\}$ are independent, even if the dice roll itself is the same roll for both

ii) Independence is symmetric [a symmetrical relation].

iii) If A and B are independent, then so are A and B^c; A^c & B; A^c & B^c.

(*) Proof: $\mathbb{P}(A \cap B^c) = \mathbb{P}(A) - \mathbb{P}(A \cap B) = \mathbb{P}(A)(1 - \mathbb{P}(B)) = \mathbb{P}(A)\mathbb{P}(B^c)$. □

iv) If $\mathbb{P}(B) = 0$ or $\mathbb{P}(B) = 1$, then A, B are independent $\forall$ events A.

(*) Proof: Let $\mathbb{P}(B) = 0$; then $\mathbb{P}(A \cap B) \leq \mathbb{P}(B) = 0 \rightarrow \mathbb{P}(A \cap B) = 0 = \mathbb{P}(A)\mathbb{P}(B)$.

Let $\mathbb{P}(B) = 1$; then $\mathbb{P}(B^c) = 0$. □

v) If $\mathbb{P}(B) > 0$, then A, B are independent iff $\mathbb{P}(A|B) = \mathbb{P}(A)$.

(*) Proof: ($\Rightarrow$) $\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} = \frac{\mathbb{P}(A)\mathbb{P}(B)}{\mathbb{P}(B)} = \mathbb{P}(A)$.

($\Leftarrow$) $\mathbb{P}(A \cap B) = \mathbb{P}(A|B)\mathbb{P}(B) = \mathbb{P}(A)\mathbb{P}(B)$. □

Def: Conditional Independence

Let $\mathbb{P}(C) > 0$; we say that two events A, B are conditionally independent given C if:

$$\boxed{\mathbb{P}(A \cap B|C) = \mathbb{P}(A|C)\mathbb{P}(B|C)}$$ [Alt.: $\mathbb{P}(A|B, C) = \mathbb{P}(A|C)$]

(*) Ex: Let $A = \{\text{person} \leq 4 \text{ ft.}\}$, $B = \{\text{person doesn't know } 3+5=8\}$, $C = \{\text{person} \leq 6 \text{ years old}\}$

We might expect $\mathbb{P}(A \cap B) \approx \mathbb{P}(A) \approx \mathbb{P}(B)$, i.e. A and B are not independent (≤ 4 ft. ~ bad at math).

But we may find that $\mathbb{P}(A \cap B|C) = \mathbb{P}(A|C)\mathbb{P}(B|C)$.

└ Idea: height, math ability correlation disappears if "≤ 6 y.o." is known

Independence (cont.)

1/22/24
Lecture 6

Independence (cont.)

(*) Ex: Flipping 2 coins: $\mathbb{P}(\text{heads on } 1^{st}) = p$, $\mathbb{P}(\text{heads on } 2^{nd}) = q$

Hypothesis: the flips are independent ⇝

1^{st} \ 2^{nd}	H	T
H	pq	$p(1-q)$
T	$(1-p)q$	$(1-p)(1-q)$

↳ independence is a hypothesis; vs no independence

(*) Ex: Coins glued together; not independent ⇝

	H	T
H	p	0
T	0	$1-p$

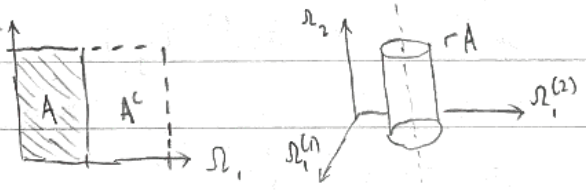
Independence of Multiple Experiments

Idea: Want to model carrying out 2 experiments $(\Omega_1, \mathcal{F}_1, \mathbb{P}_1)$, $(\Omega_2, \mathcal{F}_2, \mathbb{P}_2)$ independently

→ Combine state spaces: new state space $\Omega_1 \times \Omega_2$.

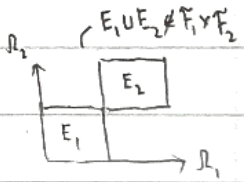
Want to model various events $[A \in \Omega_1] \times [B \in \Omega_2]$; ex: $A^c \times \Omega_2$ = "A doesn't happen"

(*) Geometrically:



Issue: $\mathcal{F}_1 \times \mathcal{F}_2$ is not guaranteed to be a σ-algebra $[\mathcal{F}_1 \times \mathcal{F}_2 = \{A \times B : A \in \mathcal{F}_1, B \in \mathcal{F}_2\}]$

→ Define tensor product: $\mathcal{F}_1 \otimes \mathcal{F}_2 :=$ the σ-algebra generated by $\mathcal{F}_1 \times \mathcal{F}_2$



Theorem. There is a unique probability $\mathbb{P}$ on $\mathcal{F}_1 \otimes \mathcal{F}_2$ with $\boxed{\mathbb{P}(A \times B) = \mathbb{P}_1(A) \cdot \mathbb{P}_2(B)}$.

Def. We say that events E_α indexed $\alpha \in A$ are independent if $\mathbb{P}(\bigcap_{\alpha \in B} E_\alpha) = \prod_{\alpha \in B} \mathbb{P}(E_\alpha)$ for every finite subset $B \subseteq A$.

Rmks: i) E's indexed by α (not n) to indicate E_α's need not be countable [index: function $\alpha \mapsto E_\alpha$]

ii) Definition holds for any finite subset, not just pairs!

iii) Consequence: Given $E_1, \dots, E_n$: need to check $P(\{E_i\})$ w/ size 2^n (can reduce to $2^n - 1 - n$)

(*) Misc. Notes

1/22/24
Lecture 6 (cont.) + HW 1-3

(*) Ex: Independence

Rolling 2 fair dice independently: let $A=\{1^{st} \text{ is even}\}$, $B=\{2^{nd} \text{ is even}\}$, $C=\{\text{sum is even}\}$

→ Observe: A, B, C all pairwise independent, but $\mathbb{P}(A\cap B\cap C)=\frac{1}{4}\neq \mathbb{P}(A)\mathbb{P}(B)\mathbb{P}(C)$

→ A, B, C not independent. [Independence fails at higher levels]

σ-Algebras (HW2)

Any intersection of σ-algebras is itself a σ-algebra.

(*) Given sample space Ω: intersection of all σ-algebras is $\{\emptyset, \Omega\}$.

⇝ Given collection A of subsets of Ω, $\exists$ smallest σ-algebra containing A [called σ-algebra generated by A, $\sigma(A)$] given by intersection of all σ-algebras containing A.

Given sample space Ω, define the Borel σ-algebra as the σ-algebra generated by open sets in Ω. (Identical to σ-algebra generated by all closed intervals, closed sets.)

Independence (cont.)

1/24/24

Lecture 7

Independence of Infinite Experiments

Given infinite experiments, we describe them via the infinite product of probability spaces.

(*) Ex: Infinitely many independent coin tosses

→ Given probability spaces $(\Omega_\alpha, \mathcal{F}_\alpha, \mathbb{P}_\alpha)$ for $\alpha \in A$, take the sample space $\prod_{\alpha \in A} \Omega_\alpha$.

Def. For such a scenario, define a cylinder set as follows: Let $S \subseteq A$ be a finite set, and for each $\alpha \in S$ let $E_\alpha \in \mathcal{F}_\alpha$. Then the associated cylinder set is the set:

$$\left(\prod_{\alpha \in S} E_\alpha\right) \times \left(\prod_{\alpha \in S^c} \Omega_\alpha\right)$$

(*) E_α may be interpreted as the events "of interest"

We want to be able to represent independence of experiments: $\mathbb{P}\left(\left(\prod_{\alpha \in S} E_\alpha\right) \times \left(\prod_{\alpha \in S^c} \Omega_\alpha\right)\right) = \prod_{\alpha \in S} \mathbb{P}_\alpha(E_\alpha)$

(*) We ignore the uncountable product $\prod_{\alpha \in S^c} \mathbb{P}_\alpha(\Omega_\alpha) = 1$

→ Natural σ-algebra: the σ-algebra generated by cylinder sets $\bigotimes_{\alpha \in A} \mathcal{F}_\alpha$ ["cylinder σ-algebra"]

Theorem Given probability spaces $(\Omega_\alpha, \mathcal{F}_\alpha, \mathbb{P}_\alpha)$, there is a unique probability law on $\left(\prod_{\alpha \in A} \Omega_\alpha, \bigotimes_{\alpha \in A} \mathcal{F}_\alpha, \mathbb{P}\right)$ satisfying ($\forall$ finite $S \subseteq A$):

$$\mathbb{P}\left(\left(\prod_{\alpha \in S} E_\alpha\right) \times \left(\prod_{\alpha \in S^c} \Omega_\alpha\right)\right) = \prod_{\alpha \in S} \mathbb{P}_\alpha(E_\alpha)$$

(*) Ex: Infinite Independent Coin Tosses

Probability spaces: $(\Omega_\alpha, \mathcal{F}_\alpha, \mathbb{P}_\alpha) = (\{H,T\}, \mathcal{P}\{H,T\}, \mathbb{P}(\omega) = \tfrac{1}{2}\ \forall \omega)$

→ can make a sequence: take indices $n \in \mathbb{N}$ ⇝ sample space: {functions $\mathbb{N} \to \{H,T\}$}

(*) Ex. Cylinder Set: $\{f : \mathbb{N} \to \{H,T\} \mid f(1) = H, f(2) = T, \ldots, f(5) \in \{H,T\}\}$ (e.g.)

Discrete Random Variables

1/24/24
Lecture 7 (cont.)

Def: Discrete Random Variables

A discrete [real-valued] random variable on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ is a function

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

such that $X(\Omega)$ is countable and measurable: for each $n \in X(\Omega)$, $\{\omega \in \Omega : X(\omega) = n\} \in \mathcal{F}$.

Remarks: i) $X(\Omega)$ can be any countable subset of $\mathbb{R}$, e.g. $\mathbb{Z}$. [May also refer to as $X: \Omega \to \mathbb{Z}$]

ii) In theory, a discrete random variable can map to any countable subset (not just subsets of $\mathbb{R}$); the advantage of real-valued variables [numbers] is that they can be added.

iii) For $n \in X(\Omega)$, the set $\{\omega \in \Omega : X(\omega) = n\}$ [called the "inverse image"] may also be denoted: $X^{-1}(\{n\})$, $\{X = n\}$, $\{X \text{ is } n\}$; for $A \subseteq X(\Omega)$, $X^{-1}(A) = \{\omega \in \Omega : X(\omega) \in A\}$.

(*) "Inverse image" does not require that an inverse of X exist

Def: Probability Mass Function (PMF)

Given a discrete random variable X, its probability mass function [PMF] is the function $p_X: \mathbb{R} \to [0,1]$ [$X(\Omega) \to [0,1]$] defined by:

$$p_X(n) = \mathbb{P}(\{X = n\})$$

Remark: The PMF determines the law/distribution of a r.v. [random variable]

Def: Given an event $E \in \mathcal{F}$, its indicator r.v. is the function 1_E def. by: $1_E(\omega) = \begin{cases} 1 & \text{if } \omega \in E \\ 0 & \text{if } \omega \notin E \end{cases}$

(*) Ex: Tossing Coins

Let $\Omega = \{H, T\}^{\mathbb{N}}$, and define indicators $X_1, \dots, X_n$: $X_k = 1_{\{k^{\text{th}} \text{ throw is heads}\}}$ [this is itself a r.v.]

→ Can use to count heads: # heads $N = X_1 + X_2 + \dots + X_n$ [# heads in first n trials]

(*) Multi-Armed Bandit

1/25/24

Disc 3

(*) Frequentist vs Bayesian Probability

Two interpretations of probability: i) Frequentist: probabilities are defined only as number of successes across repeated assignments; no notion of "prior distributions"

Results generally similar, except in the case of very heavily weighted prior beliefs.

ii) Bayesian: probabilities indicate degrees of belief in an event; update prior distributions in accordance with observations

(*) Pairwise vs Mutual Independence

Given a set of events, define: i) Events are pairwise independent if any two events are independent

ii) Events are mutually independent if, given any event, it is independent of any intersection of the other events

(*) Ex (Rock-Paper-Scissors): Define A={Alice beats Bob}, B={Bob beats Charlie}, C={Charlie beats Alice}.

→ Events are pairwise independent, but not mutually independent.

(*) Multi-Armed Bandit

Problem: Have a slot machine w/ two levers, each of which has a different probability of giving a reward.

i) [Easy ver.] Given $\mathbb{P}$'s p_1, p_2 (but don't know which is which); after n pulls (& successes) on one lever, determine if the lever was p_1 or p_2.

Want: $\mathbb{P}(\text{was } p_2 \mid n \text{ pulls}, k \text{ rewards}) \rightarrow = \dfrac{\mathbb{P}(n,k \mid p_1)\mathbb{P}(p_1)}{\mathbb{P}(n,k)}$

$\mathbb{P}(n,k \mid p_1) = (p_1)^k (1-p)^{n-k} \binom{n}{k}$; $\mathbb{P}(n,k) = \mathbb{P}(n,k \mid p_1)\mathbb{P}(p_1) + \mathbb{P}(n,k \mid p_2)\mathbb{P}(p_2)$

→ (eventually) $\mathbb{P}(p_1 \mid n,k) = \dfrac{1}{1 + (p_2/p_1)\left(\frac{1-p_2}{1-p_1}\right)^{n-k}}$

(*) ii) [Hard ver.] Same setup, can move between levers; want to maximize reward ~ solved in 2020

Discrete Probability Distributions

1/26/24
Lecture 8

Observation: All PMFs satisfy (by def.): i) $0 \le p_n(x) \le 1 \ \forall n$
ii) $\sum_{n \in \mathbb{Z}} p_X(n) = 1$

→ **Prop.** For any function $p_X: \mathbb{Z} \to \mathbb{R}$ and satisfying (i) and (ii), there is a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and r.v. $X: \Omega \to \mathbb{Z}$ with PMF given by p_X.

(*) **Pf:** Let $\Omega = \mathbb{Z}$ and $\mathcal{F} = P(\mathbb{Z})$. Define $\mathbb{P}(A) = \sum_{n \in A} p_X(n)$ and $X: \Omega \to \mathbb{Z}$ via $X(n) = n$.
Check: $\mathbb{P}(X^{-1}(\{m\})) = \mathbb{P}(\{m\}) = \sum_{n=m} p_X(n) = p_X(m)$. (Can also check $\mathbb{P}$-axioms.) □

Def. We say that a discrete r.v. $X: \Omega \to \mathbb{Z}$ is **Bernoulli(p) distributed** (and write that $X \sim$ Bernoulli(p)) if $0 \le p \le 1$ and the PMF of X is:

$$p_X(n) = \begin{cases} p & n = 1 \\ 1-p & n = 0 \\ 0 & \text{otherwise} \end{cases}$$ [Bernoulli distribution]

(*) **Ex:** The indicator r.v. 1_E is Bernoulli($\mathbb{P}(E)$) distributed.
Conversely, any $X \sim$ Bernoulli(p) is an indicator r.v. [namely, of $X^{-1}(\{1\})$].

Def. Given $0 \le p \le 1$ and $n \in \mathbb{N}$, we say that a r.v. X is **Binomial(n, p)** or that $X \sim$ Binomial(n, p) if its PMF is given by:

$$p_X(k) = \begin{cases} \binom{n}{k} p^k (1-p)^{n-k} & 0 \le k \le n \\ 0 & \text{otherwise} \end{cases}$$

(*) **Note:** We see immediately that $p_X(k) \ge 0$, to check that $\sum_{k \in \mathbb{Z}} p_X(k) = 1$, can use the Binomial Theorem: $\sum_{k \in \mathbb{Z}} \binom{n}{k} p^k (1-p)^{n-k} = (p + (1-p))^n = 1^n = 1$.

(*) p and n often represent a probability of success and number of trials, respectively

Discrete Probability Distributions (cont.)

1/22/24
Lecture 8

(*) Remark: Binomial Distributions

If $X_1, \dots, X_n \sim \text{Bernoulli}(p)$ are independent, then $[X_1 + \dots + X_n] \sim \text{Binomial}(n, p)$.

(*) Proof

Can consider sample space $\{S, F\}^n = \underbrace{\{S,F\} \times \dots \times \{S,F\}}_{n \text{ factors}}$:

$\{0 \text{ successes}\} = \{(F, \dots, F)\} = \{F\}^n \rightsquigarrow \mathbb{P}(\{0 \text{ successes}\}) = \mathbb{P}(F)^n = (1-p)^n$

$\{1 \text{ success}\} = \{(S, F, \dots, F), (F, S, F, \dots, F), \dots, (F, \dots, F, S)\}$ (i^{th} pos.)

$= \bigcup_{i=1}^{n} \{F\}^{i-1} \times \{S\} \times \{F\}^{n-i}$ [alt: $= \bigcup_{i=1}^{n} \{F, \dots, F, S, F, \dots, F\}$]

Observe: Events $\{\{F\}^{i-1} \times \{S\} \times \{F\}^{n-i}\}_i$ are disjoint

$\rightarrow \mathbb{P}\left(\bigcup_{i=1}^{n} \{F\}^{i-1} \times \{S\} \times \{F\}^{n-i}\right) = \sum_{i=1}^{n} \mathbb{P}\left(\{F\}^{n-1} \times \{S\}^i \times \{F\}^{n-i}\right)$

$= \sum_{i=1}^{n} (1-p)^{n-1} p = n(1-p)^{n-1} p \quad \left[= \binom{n}{1} (1-p)^{n-1} p\right]$

General case: $\mathbb{P}(\{k \text{ successes}\}) = \binom{n}{k} (1-p)^{n-k} p^k$ [same as $\text{Binomial}(n,p)$].

Def: Geometric Distribution

We say $X \sim \text{Geometric}(p)$ with $0 < p \leq 1$ if X has PMF given by:

$$p_X(n) = \begin{cases} (1-p)^{n-1} p & n = 1, 2, 3, \dots \\ 0 & \text{otherwise} \end{cases}$$

(*) Check PMF: $\sum_{n=1}^{\infty} (1-p)^{n-1} p = \frac{p}{1-(1-p)} = \frac{p}{p} = 1$ [using geometric series formula]

(!) Note: Strict $0 < p$ in $0 < p \leq 1$

Discrete Probability Distributions (cont.)

1/29/24
Lecture 9

2 Definitions of geometric distribution: i) # Bernoulli(p) trials before 1st success [X]
ii) # failures before success [Y = X − 1]

Can depict $\{X=k\}$ as set $(\{F\}^{k-1} \times \{S\} \times \Omega^{\mathbb{N}})$; $\Omega^{\mathbb{N}}$ signifies "ignore whatever happens afterward"

(*) Ex: Geometric Distribution

Given $X \sim \text{Geometric}(p>0)$, want to find $\mathbb{P}(X \geq n)$ for some n.

$$\mathbb{P}(X \geq n) = \sum_{k=n}^{\infty} \mathbb{P}(X=k) = \sum_{k=n}^{\infty} p(1-p)^{k-1} = \frac{p(1-p)^{n-1}}{1-(1-p)} = (1-p)^{n-1} \quad \text{[Geometric series formula]}$$

Observe: as $p>0$, $\mathbb{P}(X \geq n) \to 0$ as $n \to \infty$:

$\to \mathbb{P}(\text{succeed in finitely many trials}) = 1 - \mathbb{P}(\text{never succeed})$
$= 1 - \mathbb{P}(\bigcap_{n=1}^{\infty} X \geq n) = 1 - \lim_{n\to\infty} \mathbb{P}(X \geq n)$ [Continuity Thm.]
$= 1 - \mathbb{P}(\text{never succeed} = 0)$

(*) Def: We say that an event will almost surely occur if its complement has probability 0 [as opposed to surely occurring if its complement is empty].

Def: Poisson Distribution

We say that $X \sim \text{Poisson}(\lambda)$, where $\lambda \geq 0$ [$\lambda \in \mathbb{R}$], if its PMF is given by:

$$P_X(k) = \begin{cases} \frac{\lambda^k}{k!} e^{-\lambda} & k \in \mathbb{N} \\ 0 & \text{otherwise} \end{cases}$$

(*) Note: "Poisson" = "fish" [French]

Expected Value

1/29/24
Lecture 9 (cont.)

Check Poisson PMF: i) $p_X(k) \geq 0 \ \forall k$ ✓

ii) Recall: $\exp(x) := \sum_{k=0}^{\infty} \frac{x^k}{k!}$ [converges $\forall x \in \mathbb{C}$]; $\exp(x+y) = \exp(x)\exp(y)$

Define $e = \exp(1) \rightsquigarrow \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} e^{-\lambda} = \exp(\lambda)\exp(-\lambda) = \exp(0) = 1$. □

Def: Expected Value

Given discrete r.v. X, we define its <u>expected value</u> to be $\boxed{\mathbb{E}(X) = \sum_k k\, p_X(k)}$, provided that the sum <u>converges absolutely</u>, i.e. $\sum_k |k\, p_X(k)| < \infty$.

(•) Also called the "<u>mean</u>", "<u>average</u>", or "<u>expectation</u>" of X

(•) <u>Absolute convergence</u> gives us that the value of $\mathbb{E}(X)$ does not depend on the "order" of the k's

(*) <u>Thm (Riemann)</u>: If a sum is <u>conditionally convergent</u> but <u>not absolutely convergent</u> (i.e. $\sum_{n=1}^{\infty} a_n$ converges to finite $\mathbb{R}$, but not $\sum_{n=1}^{\infty} |a_n|$), then we can reorder its terms to create a new sum <u>converging to any arbitrary value in $\mathbb{R}$</u>.

(*) Ex: $\sum_{n=1}^{\infty} \frac{(-1)^n}{n} = \log(1/2)$, but we can rearrange its terms to get any value in $\mathbb{R}$ (47, e.g.: pick terms >0 until above 47, then terms <0 until below 47, etc.)

(•) Can view $\mathbb{E}$ like an integral:

$\rightsquigarrow$ area under graph $= \sum w_i \cdot h_i = \sum$ height · width $\left[= \sum_k k \cdot \mathbb{P}(X = k) \right]$

Expected Values of Distributions

i) $X \sim \text{Bernoulli}(p) \rightarrow \mathbb{E}(X) = 0 \cdot \mathbb{P}(X=0) + 1 \cdot \mathbb{P}(X=1) = 0(1-p) + 1(p) = \underline{p}$.

ii) $X \sim \text{Binomial}(p) \rightarrow \mathbb{E}(X) = \sum_{k=0}^{n} k \binom{n}{k} p^k (1-p)^{n-k}$.

- Recall: $\sum_{k=0}^{n} \binom{n}{k} x^k y^{n-k} = (x+y)^n$

$\xrightarrow{\partial/\partial x} \sum_{k=0}^{n} \binom{n}{k} k x^{k-1} y^{n-k} = n(x+y)^{n-1} \xrightarrow{\text{times } x} \sum_{k=0}^{n} \binom{n}{k} k x^k y^{n-k} = nx(x+y)^{n-1}$

$\rightarrow \mathbb{E}(X) = \sum_{k=0}^{n} \binom{n}{k} k p^k (1-p)^{n-k} = np(p + (1-p))^{n-1} = np(1)^{n-1} = \underline{np}$.

Expected Value (cont.)

1/31/24
Lecture 10

Rmk: If X_i are independent random variables, $\mathbb{E}(\sum X_i) = \sum \mathbb{E}(X_i)$ [$\mathbb{E}$ is linear].

Expected Values of Distributions (cont.)

iii) $X \sim \text{Geometric}(p) \to \mathbb{E}(X) = \sum_{k=1}^{\infty} k\,\mathbb{P}(X=k) = \sum_{k=1}^{\infty} k(1-p)^{k-1}p = \sum_{\ell=0}^{\infty}(\ell+1)p(1-p)^{\ell}$

- Recall: $\sum_{\ell=0}^{\infty} x^{\ell}$ [$|x|<1$] $= \frac{1}{1-x}$

$\xrightarrow{\partial/\partial x} \sum_{\ell=0}^{\infty} \ell x^{\ell-1} = \frac{1}{(1-x)^2}$

$\to \sum_{\ell=0}^{\infty}(\ell+1)(1-p)^{\ell}p = \frac{p}{(1-(1-p))^2} = \frac{1}{p}$. □

iv) $X \sim \text{Poisson}(\lambda) \to \mathbb{E}(X) = \sum_{k=0}^{\infty} k\frac{\lambda^k}{k!}e^{-\lambda}$.

Differentiate $e^{\lambda} = \sum_{\ell=0}^{\infty}\frac{\lambda^{\ell}}{\ell!} \rightsquigarrow \mathbb{E}(X) = \lambda$. □

Expected Values
i) Bernoulli(p): $\mathbb{E}(X) = p$
ii) Binomial(n,p): $\mathbb{E}(X) = np$
iii) Geometric(p): $\mathbb{E}(X) = 1/p$
iv) Poisson(λ): $\mathbb{E}(X) = \lambda$

Prop. Fix $\lambda \geq 0$; let $X_n \sim \text{Binomial}(n, \lambda/n)$ and $X \sim \text{Poisson}(\lambda)$.

Then, as $n \to \infty$, $\boxed{p_{X_n}(k) \to p_X(k)}$ [X_n converge in distribution to X]

(*) Proof: **Lemma**: $(1-\lambda/n)^n$ converges to $e^{-\lambda}$ as $n \to \infty$ for any $\lambda \in \mathbb{R}$.

(*) Proof: $(1-\lambda/n)^n = \exp(n\log(1-\lambda/n))$ for $n > \lambda$.

$\frac{\log(1-\lambda/n) - \log(1)}{1/n} = \frac{d}{dx}\log(1-\lambda x)\Big|_{x=0} = \frac{-\lambda}{(1-\lambda x)} = -\lambda \rightsquigarrow (1-\lambda/n)^n = \exp(-\lambda) = e^{-\lambda}$. □

$p_{X_n}(k) = \binom{n}{k}\cdot(\lambda/n)^k(1-\lambda/n)^{n-k} = \frac{n\cdot(n-1)\cdot\ldots\cdot(n-k+1)}{k!}\cdot\frac{\lambda^k}{n^k}\cdot\frac{(1-\lambda/n)^n}{(1-\lambda/n)^k}$.

$\rightsquigarrow$ Notice: $\frac{n\cdot(n-1)\cdot\ldots\cdot(n-k+1)}{n^k} \to 1$, $(1-\lambda/n)^k \to 1$ as $n \to \infty$.

$\to$ As $n \to \infty$, $p_{X_n}(k) \to \frac{\lambda^k}{k!}(1-\lambda/n)^n = \frac{\lambda^k}{k!}e^{-\lambda} = p_X(k)$. □

Observation: Let X a disc. r.v. $X: \Omega \to \mathbb{Z}$, $g: \mathbb{Z} \to \mathbb{Z}$; then $g \circ X: \Omega \to \mathbb{Z}$ is a random variable.

Prop: $\boxed{\mathbb{E}(g(X)) = \sum_k g(k)\,\mathbb{P}(X=k)}$

(*) Proof: $\mathbb{E}(g(X)) = \sum_{\ell}\ell\,\mathbb{P}(g(X)=\ell)$

$= \sum_{\ell}\ell \sum_{k:\,g(k)=\ell}\mathbb{P}(X=k)$

$= \sum_{\ell}\sum_{\substack{k \\ g(k)=\ell}} g(k)\,\mathbb{P}(X=k) = \sum_k g(k)\,\mathbb{P}(X=k)$. □

(*) Discrete Probability Distributions

2/1/24

Disc 4

Discrete Random Variables (Review)

A discrete random variable is a function $X: \Omega \to E$ [E countable] $\sim \Omega$ typically ignored during calculation

The distribution of an r.v. X is the $\mathbb{P}$-measure given by $\mathbb{Q}(F) = \mathbb{P}(X \in F)$; if $E = \mathbb{N}$, this is determined by the probability mass function (PMF) $k \mapsto \mathbb{P}(X = k)$.

(*) Ex: $\mathbb{P}(X=k) = C\frac{1}{k^2}$ for some normalizing constant C (s.t. $\sum_k \mathbb{P}(X=k) = 1$); want $\mathbb{P}(X \text{ even})$ (works even w/o knowing C)

$\to = \mathbb{P}(X=2) + \dots = \sum_{k \text{ even}} \mathbb{P}(k) = \sum_{k=1}^{\infty} \mathbb{P}(X=2k) = \sum_{k=1}^{\infty} \left(\frac{1}{(2k)^2}\right) = \frac{C}{4}\sum_{k=1}^{\infty} 1/k^2 = \frac{1}{4}\sum_{k=1}^{\infty}\left(\frac{1}{k^2}\right) = 1/4.$

Vandermonde's identity: $\boxed{\sum_{k=0}^{r}\binom{n}{k}\binom{m}{r-k} = \binom{n+m}{r}}$ 2 methods of proof: i) Counting: ① … ⓝ ⓜ … ⓝ+ⓜ (choose r: pick (k) of top, $(r-k)$ of bottom)

(*) Pmf (i): $(1+x)^{n+m} = \sum_{r=0}^{n+m}\binom{n+m}{r}x^r = (1+x)^n(1+x)^m = \left(\sum_{a=0}^{n}\binom{n}{a}x^a\right)\left(\sum_{b=0}^{m}\binom{m}{b}x^b\right)$

$\to = \sum_{r}^{n+m}\left(\sum_{a+b=r}\binom{n}{a}\binom{m}{b}\right)x^r \xrightarrow{1/x^r} \sum_{r=0}^{n+m}\binom{n+m}{r} = \sum_{r=0}^{n+m}\left(\sum_{k=0}^{r}\binom{n}{k}\binom{m}{r-k}\right)$. □

(*) Ex (Poisson)

Let $X \sim Poi(\lambda)$, $Y \sim Poi(\mu)$; want $\mathbb{P}(X+Y=k)$ [PMF of $X+Y$]

$\mathbb{P}(X+Y=k) = \sum_{l=0}^{k}\mathbb{P}(X=l \cap Y=k-l) = \sum_{l=0}^{k}\mathbb{P}(X=l)\mathbb{P}(Y=k-l) = \sum_{l=0}^{k} e^{-\lambda}\frac{\lambda^l}{l!}e^{-\mu}\frac{\mu^{k-l}}{(k-l)!}$

$\to = e^{-\lambda-\mu}\sum_{l=0}^{k}\binom{k}{l}\frac{1}{k!}\lambda^l\mu^{k-l} = e^{-\lambda-\mu}\frac{1}{k!}\sum_{l=0}^{k}\binom{k}{l}\lambda^l\mu^{k-l} = e^{-(\lambda+\mu)}\frac{(\lambda+\mu)^k}{k!}$

$\to (X+Y) \sim Poisson(\lambda+\mu)$.

(*) Ex (ii)

Poisson distribution represents idea of avg. # successes / interval; used for radioactive decay, e.g.

→ Q: Let $U \sim Poisson(\lambda=1)$, $P \sim Poisson(\mu=58)$ [decay / 1 sec]. If I measure 58 decay / 1 sec., what is $\mathbb{P}$ that the sample is of P? Assume initial $\mathbb{P}(U) = \mathbb{P}(P) = 1/2$.

→ Use Bayes' Rule $\mathbb{P}(P \mid N=58) = \dfrac{\mathbb{P}(N=58 \mid P)\mathbb{P}(P)}{\mathbb{P}(N=58 \mid P)\mathbb{P}(P) + \mathbb{P}(N=58 \mid U)\mathbb{P}(U)} = 0.99\ldots97$ (63 9's)

Variance & Standard Deviation

2/2/24
Lecture 11

Def: Variance & Standard Deviation

Defns: Given r.v. X:

i) We call $\mathbb{E}(X^m)$ the m^{th} moment of X. (typically, $m \in \mathbb{N}$ or $\mathbb{Z}$.)

ii) We define the variance of X as:

$$\boxed{Var(X) = \mathbb{E}\left[(X-\mu)^2\right]}, \text{ where } \mu = \mathbb{E}(X).$$

iii) We define the standard deviation of X (alt: root mean square / RMS deviation) as:

$$\boxed{\sigma(X) = \sqrt{Var(X)}}$$

Prop. i) $Var(aX) = a^2 Var(X)$.

(*) Proof: $\mathbb{E}(aX) = \sum_k ak\mathbb{P}(X=k) = a\mathbb{E}(X) = a\mu$

$\rightarrow Var(aX) = \mathbb{E}((aX-a\mu)^2) = \mathbb{E}(a^2(X-\mu)^2) = a^2\mathbb{E}((X-\mu)^2) = a^2 Var(X)$. □

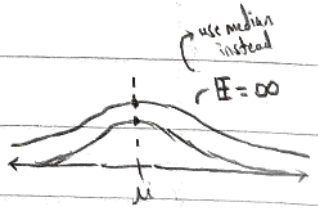
ii) $\boxed{Var(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2}$ ~ more useful definition for computation

(*) Proof: $\mathbb{E}((X-\mu)^2) = \mathbb{E}(X^2 - 2\mu X + \mu^2) = \mathbb{E}(X^2) - 2\mu\mathbb{E}(X) + \mu^2$

$\rightarrow = \mathbb{E}(X^2) - 2\mathbb{E}(X)^2 + \mathbb{E}(X)^2 = \mathbb{E}(X^2) - \mathbb{E}(X)^2$. □

(*) Notes on Variance & Standard Deviation

- Variance used more in mathematics; standard deviation used more in real science/statistics
  - Standard deviation converts squared units back to regular units, effectively
- Warning: Mean $\mathbb{E}(X)$, variance $Var(X)$ both very sensitive to outliers; vs. median (more stable, but more difficult mathematically)



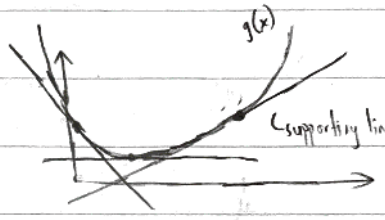
Convexity & Jensen's Inequality

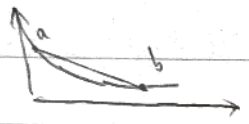
2/2/24
Lecture 11 (cont.)

Jensen's Inequality	If $g: \mathbb{R} \to \mathbb{R}$ is convex, then $\boxed{\mathbb{E}(g(X)) \geq g(\mathbb{E}(X))}$.

Def: Convex Functions

We consider a function $f: \mathbb{R}^n \to \mathbb{R}$ convex if, for every point x' in the domain, we can find a supporting line [plane, hyperplane, etc.] passing through x' that has value $\underbrace{\leq f(x)}_{\text{"below the graph"}}$ at all points x.

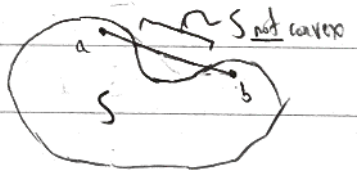




Equivalently ($\mathbb{R}^1$): for any points a, b in the domain, the secant line between a, b has value $\geq f(x)$ at all points x.

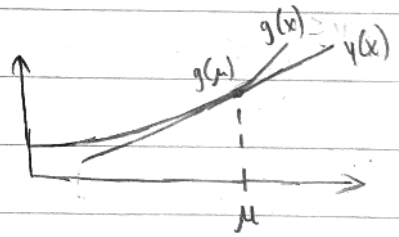
(*) Def: Convex Sets

A set is convex if, for any two points a, b in the set, the "line" between a, b is contained in the set.



(*) Proof (Jensen's inequality)

Let $\mu = \mathbb{E}(X)$ and consider supporting line $y = m[x-\mu] + g(\mu)$; by def., $g(x) \geq m[x-\mu] + g(\mu)$, with equality at $x = \mu$.



$$\mathbb{E}(g(X)) = \sum_k g(k)\mathbb{P}(X=k)$$
$$\geq \sum_k [m(k-\mu)+g(\mu)]\mathbb{P}(X=k) = m\sum_k k\mathbb{P}(X=k) + [g(\mu)-m\mu]\sum_k \mathbb{P}(X=k)$$
$$= m\mu - m\mu + g(\mu) = g(\mu) = g(\mathbb{E}(X)). \quad \square$$

Conditional PMF & Expectation

2/2/24
Lecture 11
+ Lecture 12

Def: Conditional PMF & Expectation

If X a r.v. and A an event with $\mathbb{P}(A) > 0$, we define the conditional PMF of X given A:

$$\boxed{p_{X|A}(k) = \mathbb{P}(X = k \mid A)}$$

Similarly, we define the conditional expectation of X given A:

$$\boxed{\mathbb{E}(X|A) = \sum_k k\, p_{X|A}(k)} \quad \text{[given the sum converges absolutely]}$$

Thm: Partition Theorem / Law of Total Expectation

Given r.v. X with finite expectation (i.e. $\mathbb{E}(|X|) < \infty$; converges absolutely) and partition $\{B_j\}$ of Ω with each $\mathbb{P}(B_j) > 0$, we have:

$$\boxed{\mathbb{E}(X) = \sum_j \mathbb{E}(X|B_j)\mathbb{P}(B_j)}$$

In particular, the expectations $\mathbb{E}(X|B_j)$ exist.

(*) Proof

By Partition Theorem (original), $\mathbb{P}(X=k) = \sum_j \mathbb{P}(X=k|B_j)\mathbb{P}(B_j)$.

$$\to \mathbb{E}(X) = \sum_k k\,\mathbb{P}(X=k) = \sum_k k\left[\sum_j \mathbb{P}(X=k|B_j)\mathbb{P}(B_j)\right]$$

$$\text{(i)} \quad \stackrel{?}{=} \sum_j \left[\sum_k k\,\mathbb{P}(X=k|B_j)\right]\mathbb{P}(B_j)$$

$$\text{(ii)} \quad \stackrel{?}{=} \sum_j \mathbb{E}(X|B_j)\mathbb{P}(B_j)$$

→ Need to show: (i) Can exchange order of sums

(ii) The sums $\mathbb{E}(X|B_j)$ exist [converge absolutely]

Conditional Expectation (cont.)

2/5/24

Lecture 12 (cont.)

(*) Proof (cont.)

(i): Lemma: Can exchange order of integration [summation] if either:

1) The sum is absolutely convergent (Fubini)

2) All terms are non-negative; may result in infinite sum (Tonelli) (alt: all terms have same sign)

Want to show: $\sum_{j,k} |k| \mathbb{P}(X=k|B_j)\mathbb{P}(B_j)$ absolutely convergent [$<\infty$]

$\to$ (all terms ≥ 0) $= \sum_k |k| \left(\sum_j \mathbb{P}(X=k|B_j)\mathbb{P}(B_j)\right) = \sum_k |k| \mathbb{P}(X=k) = \mathbb{E}(|X|) < \infty$ (by assump.)

(ii) Want to show: $\mathbb{E}(X|B_j) = \sum_k k \mathbb{P}(X=k|B_j)$ absolutely convergent

Know: $\sum_j \underbrace{[\sum_k |k| \mathbb{P}(X=k|B_j)]}_{\mathbb{E}(X)_+} \mathbb{P}(B_j) < \infty$ (from before) & $\mathbb{P}(B_j)>0 \ \forall j$

$\to$ freeze j: know $\sum_k |k| \mathbb{P}(X=k|B_j)\underbrace{\mathbb{P}(B_j)}_{>0} < \infty \ \forall j$.

□ □

(*) Ex: Partition Theorem

Given $N \sim \text{Poisson}(\lambda)$ many Bernoulli(p) trials (indep.), want avg. number of successes $\mathbb{E}(X)$.

$\to$ Intuitive answer: λp successes

Mathematically: Use partition $B_n = \{N=n\}$, $n=0,1,2,\ldots$

$\to \mathbb{E}(X|B_n) = np$

$\to \mathbb{E}(X) = \sum_n \mathbb{E}(X|B_n)\mathbb{P}(B_n)$

$= \sum_{n\geq 0} (np)\mathbb{P}(B_n)$

$= p \sum_{n\geq 0} n\mathbb{P}(B_n) = p\mathbb{E}(N) = \underline{p\lambda}$.

(*) Expected Value & Variance

2/8/24
Disc 5

Variance: $\text{Var}(X) = \mathbb{E}[(X-\mu)^2]$; used due to having "nice properties" (e.g. additivity)

↳ Variance/expected value only exist if sum is absolutely convergent!

(*) Ex: i) [$\mathbb{E}(X)$ diverges] $\mathbb{P}(X=k) = C/k^2 \to \mathbb{E}(X) = C\sum k \cdot 1/k = C\sum 1/k$, diverges

ii) [$\text{Var}(X)$ diverges] $\mathbb{P}(X=k) = C/k^3 \to \text{Var}(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2$

$\mathbb{E}(X^2) = C\sum k^2 \cdot 1/k^3 = C\sum 1/k$, diverges

(*) Conditional Convergence

Want to find r.v. X with $\mathbb{E}(X) = \sum \frac{(-1)^n}{n}$ [conditionally convergent]

↝ Can initialize X with some distribution ($X \sim \text{Geometric}(1/2)$, e.g.)

→ Take $\mathbb{E}(f(X)) = \sum_{n=1}^{\infty} f(n)\frac{1}{2^n}$ with $f(n) = 2^n \frac{(-1)^n}{n}$.

(*) Hat-Check Problem

Problem: There are n people, each with a distinguishable hat. Given that the hats are shuffled and then given back, what is $\mathbb{E}(N = \#\text{ of people with the correct hat})$?

Approaches: i) Via Partition Theorem & lots of calculation [bad]

ii) Observe: for any person i, $\mathbb{P}(i \text{ gets the correct hat}) = 1/n$

→ Define $X_i = \#$ of correct hats received by person i [$\mathbb{E}(X_i) = 1/n$]

→ $N = X_1 + \dots + X_n \to \mathbb{E}(N) = \mathbb{E}(X_1 + \dots + X_n) = \underline{1}$.

(expectation linear even when X_i's are dependent)

(*) Application: Unique IP Addresses

Problem: Given some large number of visitors to a website, want to track # unique visitors space-efficiently. (i.e. w/o storing all IP addresses)

Solution: Given IP address (e.g. 192.168.1.1), map to random binary string (0's and 1's); across visitors, keep track of largest number of leading 1's seen M.

→ Ex: Given n unique visitors, $\mathbb{E}(M) = \sum_{k=1}^{\infty} \mathbb{P}(M \geq k) = \sum_{k=1}^{\infty}\left(1 - \left(1-\frac{1}{2^k}\right)^n\right) \approx \log_2(n)$ (approximate via integral)

Parameter Estimation

2/9/24

Lecture 13 Parameter Estimation

Given a distribution with unknown parameter(s), we may want to estimate those parameter(s) based on sample(s) of that distribution.

(*) Ex: We know $X \sim \text{Geometric}(p)$; want to find p.

Approaches:

i) Assume we measure $X = k$; then we can estimate p using maximum likelihood estimation [MLE]: finding the value p that maximizes $\mathbb{P}(X=k \mid p)$. ← conditioning informally

Optimization - can differentiate to find a maximum

$$0 = \frac{\partial}{\partial p}\left(p(1-p)^{k-1}\right) = 1 - pk \longrightarrow p = 1/k. \quad [\text{if } 0 < p < 1]$$

ii) To model a random estimate from a random experiment, we can interpret an estimate for p as a random variable $\frac{1}{X}$.

→ Q: Is our estimate biased? i.e. $\mathbb{E}\left(\frac{1}{X} \mid p\right) \stackrel{?}{=} p$

Check: $\mathbb{E}\left(\frac{1}{X}\right) = \sum_{k=1}^{\infty} \frac{1}{k}(1-p)^{k-1} p$ ← (*) Recall: $-\log(1-x) = \sum_{\ell=1}^{\infty} \frac{1}{\ell} x^{\ell}$

→ $= \frac{p}{1-p} \log\left[\frac{1}{1-(1-p)}\right] = \frac{p}{1-p} \log\left[1/p\right]$.

→ A: Systematically overestimates p [$> p$ for $0 < p < 1$]

iii) Alternatively, we could try to estimate $\mu = 1/p$ [mean of average variable]

→ notice: MLE $\hat{\mu} = 1/\hat{p}$

Estimating $\mu = k = X$, get unbiased estimate $\mathbb{E}(X) = 1/p = \mu$.

Multiple Random Variables

2/9/24
Lecture 13 (cont.)

Multiple [Discrete] Random Variables

Given a sample space Ω, we may define multiple random variables $X, Y: \Omega \to \mathbb{Z}$ (e.g.); need a way to talk about multiple (possibly interdependent) random variables.

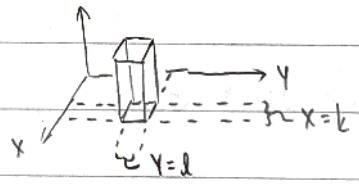
→ Recall: To describe a single random variable, we used PMFs.

Def: Joint PMF

Given two r.v.s $X, Y: \Omega \to \mathbb{Z}$, the joint PMF of X and Y is the function $P_{X,Y}: \mathbb{Z}^2 \to [0,1]$ defined by:

$$P_{X,Y}(k,\ell) = \mathbb{P}(X=k \ \& \ Y=\ell)$$

- $\{X=k \ \& \ Y=\ell\}$ also denoted: $\{\omega \in \Omega : X(\omega)=k \ \& \ Y(\omega)=\ell\}$, $X^{-1}(\{k\}) \cap Y^{-1}(\{\ell\})$
- Can visualize joint PMFs as 2D histograms:



Observation: Given the joint law [PMF] of X and Y, can recover the marginal distribution of X (i.e. P_X) via partition rule:

$$P_X(k) = \sum_{\ell} P_{X,Y}(k,\ell)$$ [Marginal distribution]

Prop: $\mathbb{E}(g(X,Y)) = \sum_{k,\ell} g(k,\ell) P_{X,Y}(k,\ell)$ [iff the sum is absolutely convergent]

In particular, if $\mathbb{E}(|X|) < \infty$ and $\mathbb{E}(|Y|) < \infty$, then $\mathbb{E}(|X+Y|) < \infty$ and:

$$\mathbb{E}(X+Y) = \mathbb{E}(X) + \mathbb{E}(Y)$$ [Linearity of expectation]

Multiple Random Variables (cont.)

2/12/24
Lecture 14

Claim: (i) $\mathbb{E}(g(X,Y)) = \sum_{k,\ell} g(k,\ell)\, p_{X,Y}(k,\ell)$.

(ii) If $\mathbb{E}(|X|), \mathbb{E}(|Y|) < \infty$, then $\mathbb{E}(|X+Y|) < \infty$ and $\mathbb{E}(X+Y) = \mathbb{E}(X) + \mathbb{E}(Y)$.

(*) Can extend to all finite sums via induction

(*) **Proof:** (i) Initially, assume g is non-negative (lets us use Tonelli).

$$\mathbb{E}(X+Y) = \sum_m m\,\mathbb{P}(g(X,Y)=m) = \sum_{m\geq 0} m \left(\sum_{k,\ell} 1_{g(k,\ell)=m}\,\mathbb{P}(X=k, Y=\ell)\right)$$

$$\to = \sum_{k,\ell}\left(\sum_{m\geq 0} m\, 1_{g(k,\ell)=m}\right)\mathbb{P}(X=k, Y=\ell)$$

$$= \sum_{k,\ell} g(k,\ell)\,\mathbb{P}(X=k, Y=\ell)$$

(*) equality: either both are finite, or neither is finite

For general g: write as $g = g\wedge 0 + g\vee 0$ [$\wedge$ = min, $\vee$ = max]

→ can apply previous result: converges for g iff converges for both $g\wedge 0$, $g\vee 0$ individually.

(ii) Assume $\mathbb{E}(|X|), \mathbb{E}(|Y|)$ exist.

$$\mathbb{E}(X) + \mathbb{E}(Y) = \left(\sum_k k\,\mathbb{P}(X=k)\right) + \left(\sum_\ell \ell\,\mathbb{P}(Y=\ell)\right)$$

$$= \sum_{k,\ell} (k+\ell)\,\mathbb{P}(X=k, Y=\ell)$$

Since this sum is composed of terms of $\mathbb{E}(X)$ and $\mathbb{E}(Y)$ [both absolutely convergent], then the new sum converges absolutely.

By (i), $\sum_{k,\ell}(k+\ell)\,\mathbb{P}(X=k, Y=\ell) = \mathbb{E}(X+Y)$. □

(*) Indicator Random Variables

- Indicator r.v.'s used similarly in vector calc. to integrate over hard surfaces: $\iint_B f(x,y) = \iint_{\mathbb{R}^2} 1_B f(x,y)$
- **Prop:** Given event A, $\mathbb{P}(A) = \mathbb{E}(1_A)$.

Independence of Random Variables

2/12/24

Lecture 14 (cont.)

(*) Hat-Check Problem (cont.)

Recall: n hats given back randomly to n people; want X = # of hats given to the right person.

(i) $\mathbb{E}(X)$: Define $X_i = 1_{i^{th} \text{ person received } i^{th} \text{ hat}}$ $\rightsquigarrow \mathbb{E}(X_i) = 1/n$

$X = \sum_{i=1}^{n} X_i \rightarrow \mathbb{E}(X) = \mathbb{E}(\sum_{i=1}^{n} X_i) = \sum_{i=1}^{n} \mathbb{E}(X_i) = n \cdot 1/n = \underline{1}$.

(ii) $\mathrm{Var}(X)$: Want to use $\mathrm{Var}(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2$; need to find $\mathbb{E}(X^2)$.

$\mathbb{E}(X^2) = \mathbb{E}\left([X_1 + \dots + X_n]^2\right)$; expand expression

$\rightarrow = \sum_j \mathbb{E}(X_j^2) + 2\sum_{i<j} \mathbb{E}(X_i X_j)$

Observe: (i) $X_j^2 = X_j$ [either 0 or 1]

(ii) The # of permutations fulfilling both X_i, X_j is $(n-2)!$, out of total $n!$

(iii) # choices of i, j is $\binom{n}{2}$

$\rightarrow \sum_j \mathbb{E}(X_j^2) + 2\sum_{i<j} \mathbb{E}(X_i X_j) = n \cdot 1/n + 2\binom{n}{2}\frac{(n-2)!}{n!} = 2$ [$= \mathbb{E}(X^2)$]

$\rightarrow \mathrm{Var}(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2 = 2 - 1^2 = \underline{1}$.

Def: Independent Random Variables

We say two r.v.'s X, Y are independent when, for every k & ℓ, the events $\{X = k\}, \{Y = \ell\}$ are independent; i.e. $p_{X,Y}(k,\ell) = p_X(k)\,p_Y(\ell)$.

Remarks: (i) Similarly, we say r.v.'s $X_1, \dots, X_i, \dots$ are independent if, for any choices of k_i, the events $\{X_1 = k_1\}, \dots, \{X_i = k_i\}, \dots$ are independent.

- Equivalently: any finite subset of the k_i's

(ii) Indicators $1_A, 1_B, \dots$ are independent only if the underlying events are independent

(iii) Lemma: If X and Y are independent, then so are $f(X), g(Y)$ for any X, Y.

- Alt: σ-algebras of all sets gen. by X, gen. by Y are independent

(*) Proof: $\mathbb{P}(f(X) = r \;\&\; g(Y) = \ell) = \sum_{k,\ell} 1_{f(k)=r \;\&\; g(\ell)=s}\, \mathbb{P}(X = k \;\&\; Y = \ell)$

$\rightarrow = \sum_{k,\ell} 1_{f(k)=r} 1_{g(\ell)=s} \mathbb{P}(X=k)\mathbb{P}(Y=\ell) = \left(\sum_k 1_{f(k)=r}\mathbb{P}(X=k)\right)\left(\sum_\ell 1_{g(\ell)=s}\mathbb{P}(Y=\ell)\right) = \mathbb{P}(f(X)=r)\,\mathbb{P}(g(Y)=\ell)$. □

Independence of Random Variables (cont.)

2/14/24
Lecture 15

Prop. If X, Y are independent and $\mathbb{E}(|X|), \mathbb{E}(|Y|) < \infty$, then XY has finite expectation given by:

$$\boxed{\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)} \quad [< \infty]$$

(*) **Proof:** Consider $\sum_{k,\ell} |k\ell| \mathbb{P}(X=k \,\&\, Y=\ell) = \sum_{k,\ell} |k| \cdot |\ell| \cdot \mathbb{P}(X=k)\mathbb{P}(Y=\ell)$

$$= \sum_k |k| \mathbb{P}(X=k) \left[\sum_\ell |\ell| \mathbb{P}(Y=\ell)\right]$$

$$\xrightarrow{\text{Tonelli}} = \left(\sum_k |k| \mathbb{P}(X=k)\right)\left(\sum_\ell |\ell| \mathbb{P}(Y=\ell)\right) = \mathbb{E}(|X|)\mathbb{E}(|Y|)$$

By assumption, the product $\mathbb{E}(|X|)\mathbb{E}(|Y|)$ is finite [converges absolutely]

$\rightsquigarrow$ can perform same process (using Fubini instead of Tonelli) to find $\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$. □

Corollary: If X, Y are independent and $f(X)$, $g(Y)$ have finite expectation, then $f(X)g(Y)$ has finite expectation and $\mathbb{E}[f(X)g(Y)] = \mathbb{E}(f(X))\mathbb{E}(g(Y))$.

Remark. If two r.v.'s X & Y have the property that $\mathbb{E}(f(X)g(Y)) = \mathbb{E}(f(X))\mathbb{E}(g(Y))$ for every pair of bounded functions f and g, then X, Y are independent.

(*) **Proof:** Assuming the $\mathbb{E}$ claim is true, we can take $f = \mathbb{1}_{\{k\}}(X)$, $g = \mathbb{1}_{\{\ell\}}(Y)$ bounded functions for any $k, \ell \in \mathbb{Z}$. Then:

$$\mathbb{P}(X=k \cap Y=\ell) = \mathbb{E}(f(X)g(Y))$$

$$= \mathbb{E}(f(X))\mathbb{E}(g(Y)) = \mathbb{P}(X=k)\mathbb{P}(Y=\ell) \text{ for arbitrary } k, \ell. \square$$

(*) **Note on Prop:** To see the role of independence, consider integrals $\int_0^1 x^{-1/2}\,dx$, $\int_0^1 y^{-1/2}\,dy = 2$

$\rightarrow$ can take $\int_0^1\int_0^1 x^{-1/2}y^{-1/2}\,dx\,dy$, but not $\int_0^1 x^{-1/2}x^{-1/2}\,dx$ [$= \infty$].

(*) Expectation of Multiple Variables

2/15/24

Disc 6

Recall (Linearity of Expectation): For any r.v.'s $X_1, \dots, X_n$: $\mathbb{E}(\sum_{i=1}^n X_i) = \sum_{i=1}^n \mathbb{E}(X_i)$, and exists if all $\mathbb{E}(X_i)$'s exist. [X_i's need not be independent!]

(*) **Ex:** Flip n coins; 1^{st} coin 50/50, $\mathbb{P}(\text{flip } k+1 \text{ is H} \mid \text{flip } k \text{ is H}) = \mathbb{P}(\text{flip } k+1 \text{ is T} \mid \text{flip } k \text{ is T}) = p$

$\rightsquigarrow$ **initial hypothesis:** $\mathbb{P}(i^{th} \text{ flip is H}) = \mathbb{P}(i^{th} \text{ flip is T}) = 1/2$ (by symmetry)

set N = # heads; want $\mathbb{E}(N)$, but hard to compute directly $\rightarrow$ use linearity of expectation

$N = \sum_{i=1}^n 1_{\{x_i = H\}} \rightarrow \mathbb{E}(N) = \sum_{i=1}^n \mathbb{E}(1_{H_i}) = \sum_{i=1}^n \mathbb{P}(H_i) = n \cdot 1/2 = n/2$ [alt: $N = \mathbb{E}(\#H) + \mathbb{E}(\#T)$]

(*) **Ex:** Shuffle n cards $1, \dots, n$ and deal 1 by 1 (keeping track of running max of cards dealed)

Define N = # of times this maximum changes; want $\mathbb{E}(N)$

$\rightarrow$ (i) **Naïve soln:** $X_i = 1_{i^{th} \text{ card a new max}} \rightsquigarrow N = \sum_{i=1}^n X_i$ [Issue: $\mathbb{E}(N) = \sum_{i=1}^n \mathbb{E}(X_i)$ hard to compute]

(ii) **Better:** Observe that if j cards have been drawn, $\mathbb{P}(j^{th} \text{ card was largest}) = \mathbb{P}(j^{th} \text{ was smallest})$

$= \mathbb{P}(j \text{ was } i^{th} \text{ largest})$ etc.; all orderings equally likely $\rightarrow \mathbb{P}(j^{th} \text{ is max of } 1^{st} j) = 1/j$

$\rightsquigarrow \mathbb{E}(N) = \sum_{i=1}^n \mathbb{E}(X_i) = \sum_{i=1}^n \mathbb{P}(i^{th} \text{ is max of } 1^{st} i) = \sum_{i=1}^n 1/i$ [$= n^{th}$ harmonic H_n; $\approx \log n$]

(*) **Ex (3-SAT):** Given Boolean variables $B_1, \dots, B_k$ and 3-clauses $C_1, \dots, C_n$, want to determine if $\exists$ an assignment to $B_1, \dots, B_k$ satisfying all clauses. [Assume each clause has 3 distinct variables]

$\rightsquigarrow$ **Soln:** Assign truth values to $B_1, \dots, B_k$ uniformly at random [truth values $\in \{T, F\}$]

$\rightsquigarrow$ interpreting clauses C_i as r.v.'s $C_i = 1_{C_i \text{ satisfied}}$, set N = # clauses satisfied $= \sum_{i=1}^n C_i$

Observe: For any clause $C_i = (B_j \vee B_k \vee B_\ell)$ [or negations of $B_{j,k,\ell}$], $\mathbb{P}(C_i \text{ satisfied}) = 7/8$

given random assignment of B_j, B_k, B_ℓ; true for all $i \rightarrow \mathbb{E}(N) = \sum_{i=1}^n \mathbb{E}(C_i) = \frac{7n}{8}$ ~ fairly high

In particular: if $n = 7$, $\mathbb{E}(N) = \frac{7 \cdot 7}{8} = 6.125 > 6 \rightarrow$ there exists at least 1 assignment with $N = 7$

[Doesn't work for larger n; ex: $n = 11 \rightarrow$ only guarantees 10, $n = 10^6 \rightarrow$ only guarantees $\frac{7}{8} \cdot 10^6$]

(*) **Ex (Coupon Collector):** n coupons drawn randomly & unif; want $\mathbb{E}$(total # drawn, incl. repeats, by the time n unique collected)

$\rightarrow$ Define $T_0 = 0, T_1, \dots, T_n$ [T_j = # drawn when j^{th} unique collected] $\rightsquigarrow \mathbb{E}(N) = \mathbb{E}(T_n - T_{n-1}) + \dots + \mathbb{E}(T_1 - T_0)$; $T_{j+1} - T_j$'s are $\sim \text{Geo}(\frac{n-j}{n})$

Covariance

2/16/23
Lecture 16

Thm. Let X, Y be independent r.v.'s and let $S = X+Y$:

(i) $p_S(n) = \sum_{k \in \mathbb{Z}} p_X(n-k)\, p_Y(k)$.

(ii) If $\mathbb{E}(|X|+|Y|)$ exists, then $\mathbb{E}(X+Y)$ exists and is $\mathbb{E}(X+Y) = \mathbb{E}(X) + \mathbb{E}(Y)$.

(iii) If $\mathbb{E}(|X|^2+|Y|^2) < \infty$, then $\operatorname{var}(S)$ exists and is $\operatorname{var}(S) = \operatorname{var}(X) + \operatorname{var}(Y)$.

(*) **Proof:** (i) $\{S=n\} = \bigcup_{k \in \mathbb{Z}} [\{X=n-k \cap Y=k\}]$ disjoint

$\rightarrow \mathbb{P}(S=n) = \sum_k \mathbb{P}(X=n-k \cap Y=k) = \sum_k \mathbb{P}(X=n-k)\,\mathbb{P}(Y=k)$

(ii) See p. 126

(iii) $\operatorname{var}(S) = \mathbb{E}\left[(S-\mathbb{E}(S))^2\right] = \mathbb{E}\left[(X+Y-\mathbb{E}(X)-\mathbb{E}(Y))^2\right]$ ~ $\operatorname{cov}(X,Y)$

$= \mathbb{E}\left[(X-\mathbb{E})^2\right] + \mathbb{E}\left[(Y-\mathbb{E}(Y))^2\right] + 2\mathbb{E}\left[(X-\mathbb{E}(X))(Y-\mathbb{E}(Y))\right]$

$= \operatorname{var}(X) + \operatorname{var}(Y) + 2\,\mathbb{E}(X-\mathbb{E}(X))\,\mathbb{E}(Y-\mathbb{E}(Y))$ [X, Y indep]

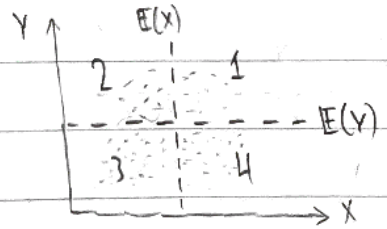
$= \operatorname{var}(X) + \operatorname{var}(Y)$. □

Def: Covariance

Given two r.v.'s X and Y, we define the covariance of X and Y to be:

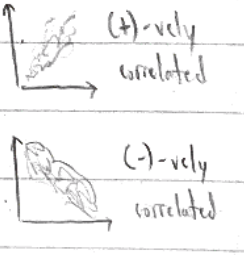
$$\operatorname{cov}(X,Y) = \mathbb{E}\left[(X-\mathbb{E}(X))(Y-\mathbb{E}(Y))\right]$$

(*) Covariance



Quadrants 1 & 3: contribute positively to $\operatorname{cov}(X,Y)$

Quadrants 2 & 4: contribute negatively to $\operatorname{cov}(X,Y)$



- X, Y independent $\rightarrow \operatorname{cov}(X,Y) = 0$; $X=Y \rightarrow \operatorname{cov}(X,Y) = \operatorname{var}(X)$
- Variance, covariance both sensitive to outliers

Cumulative Distribution Functions

2/16/23
Lecture 16 (cont.)

(*) Convolution

The function $p_S(n) = \sum_{k \in \mathbb{Z}} p_X(n-k) p_Y(k)$ is a discrete convolution of p_X and p_Y.

Def: Real-Valued Random Variable

A (real-valued) random variable on probability space $(\Omega, \mathcal{F}, \mathbb{P})$ is a function $X: \Omega \to \mathbb{R}$ that is measurable, i.e.: for all $x \in \mathbb{R}$, $\{X \leq x\}$ belongs to $\mathcal{F}$.

↝ Need to go beyond PMFs p_X [$p_X(n) = \mathbb{P}(X=n)$] in continuous case:

Def: Cumulative Distribution Function (CDF)

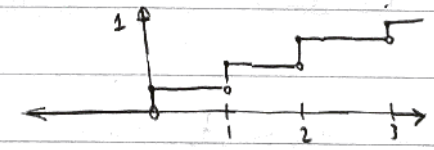
Given real-valued r.v. X, we define the cumulative distribution function [CDF] of X to be the function $F_X: \mathbb{R} \to [0,1]$ defined by:

$$F_X(x) = \mathbb{P}(X \leq x)$$

Notes on CDFs

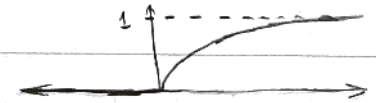
- CDFs are always increasing [non-decreasing]
- CDFs need not be continuous

(*) Ex: $X =$ (i) # heads in 3 tosses:



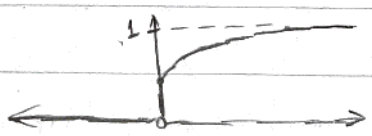
[Discrete]

(ii) bus waiting times:



[Continuous]

(iii) inches of rain:



[Neither cts. nor disc.]

Continuous Random Variables

2/21/24
Lecture 17

Prop. (i) $\lim_{x\to\infty} F_X(x) = 0$, $\lim_{x\to\infty} F_X(x) = 1$

(ii) F_X is <u>non-decreasing</u>: if $x \le y$, $F_X(x) \le F_X(y)$.

(iii) F_X is "<u>càdlàg</u>" [continuous from left, limit from right]: $\lim_{y \downarrow x} F_X(y) = \mathbb{P}(X \le x)$,
$\lim_{y \uparrow x} F_X(y) = F_X(x)$

Thm (Lebesgue). If F_X is a function satisfying the above, $\exists$ probability measure $\mathbb{P}$ on $\Omega = \mathbb{R}$ endowed with the Borel σ-algebra that satisfies $\mathbb{P}(\{y \in \mathbb{R} : y \le x\}) = F_X(x)$ $\forall\, x \in \mathbb{R}$, i.e. the r.v. $X : y \mapsto y$ has CDF given by F_X.

Rmk: Given $a < b$, $\mathbb{P}(a < X \le b) = \mathbb{P}(\{X \le b\} \setminus \{X \le a\})$
$= \mathbb{P}(X \le b) - \mathbb{P}(X \le a) = F_X(b) - F_X(a)$.

$\rightarrow$ $\mathbb{P}(a < X \le b) = F_X(b) - F_X(a)$.

Def: Continuous Random Variables

A r.v. $X : \Omega \to \mathbb{R}$ is called <u>continuous</u> if $\exists$ <u>integrable function</u> $f_X : \mathbb{R} \to \mathbb{R}$ such that:

$$F_X(x') = \int_{-\infty}^{x'} f_X(x)\, dx$$

The function f_X is called the <u>probability density function</u> [PDF] of X.

Rmks. (i) Integrability constraint ensures all sets have defined "length" under f_X (see also: Lebesgue-Stieltjes measure)

(ii) It is true that F_X is continuous for any continuous r.v. X; however, it is <u>not</u> true that every continuous CDF F_X admits a probability density (see: Cantor function)

$\rightsquigarrow$ Class of functions [r.v.'s] that do admit PDFs called <u>absolutely continuous</u>

Continuous Random Variables (cont.)

2/21/24
Lecture 17 (cont.)

Def: Absolute Continuity

We say a function F is absolutely continuous if, for every $\varepsilon > 0$, $\exists\, \delta > 0$ s.t. for any set of intervals (a_i, b_i) with $|b_1 - a_1| + |b_2 - a_2| + \dots + |b_n - a_n| < \delta$, then $\sum_{i=1}^{n} |F(b_i) - F(a_i)| < \varepsilon$.

Prop. If $X: \Omega \to \mathbb{R}$ is a cts. r.v., then $\boxed{\mathbb{P}(a < X < b) = \mathbb{P}(a \le X \le b) = \int_a^b f_X(x)\,dx}$ for every $a \le b$.

(vs Lebesgue integration: $\int_{(a,b)} \neq \int_{[a,b]}$)

→ Corollary: $\mathbb{P}(X = a) = 0 \;\; \forall a \in \mathbb{R}$.

(*) **Proof:** Want to show that $\mathbb{P}(a \le X \le b) \le \mathbb{P}(a < X \le b) \le \mathbb{P}(a \le X \le b)$.

Observe: $\{a < X < b\} = \bigcup_{n \in \mathbb{N}} \{a < X \le b - 1/n\}$

→ by continuity: $\mathbb{P}(a < X < b) = \lim_{n\to\infty} [F_X(b - 1/n) - F_X(a)] = F_X(b) - F_X(a)$.

Similarly: $\{a \le X \le b\} = \bigcap_{n \in \mathbb{N}} \{a - 1/n < X \le b\} \to \mathbb{P}(a \le X \le b) = \lim_{n\to\infty} [F_X(b) - F_X(a - 1/n)]$

$= F_X(b) - F_X(a)$. ∎

Remark: f_X is a density measured in units of reciprocal length (e.g.: probability/kg., $\mathbb{P}/m.$)

(*) Ex: Let $Y = aX$ [X, Y cts. r.v.'s], $a > 0$

→ $F_Y(y) = \mathbb{P}(Y \le y) = \mathbb{P}(aX \le y) = \mathbb{P}(X \le y/a) \rightsquigarrow \underline{F_Y(y) = F_X(y/a)}$.

Differentiating: $F_Y'(y) = f_Y(y) = \frac{1}{a} F_X'(y/a) = \underline{\frac{1}{a} f_X(y/a)}$.

Def: Uniform Distribution

We say that $X \sim \text{Uniform}([a,b])$ if it has PDF f_X defined by:

$$\boxed{f_X(x) = \begin{cases} \frac{1}{b-a} & a \le x \le b \\ 0 & \text{otherwise} \end{cases}}$$

(*) Covariance & Correlation

2/22/24
Disc. 7

Recall: (i) $\mathrm{Var}(X) = \mathbb{E}((X-\mu)^2) = \mathbb{E}(X^2) - \mathbb{E}(X)^2$
(ii) $\mathrm{Cov}(X,Y) = \mathbb{E}((X-\mu)(Y-\nu)) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)$.

~ Observe: $\mathrm{Cov}(X,X) = \mathrm{Var}(X)$.

~ Define correlation: $\boxed{\rho(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}} \in [-1,1]}$

(*) Correlation only captures linear relationships

Prop. (i) If X, Y are independent, then they are uncorrelated.
(ii) Even if X, Y are uncorrelated, they need not be independent.

(*) Proof: (i) If X, Y indep: $\mathrm{Cov}(X,Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y) = \mathbb{E}(X)\mathbb{E}(Y) - \mathbb{E}(X)\mathbb{E}(Y) = 0$.
(ii) Let X be a $\mathbb{Z}$-valued r.v. that is symetrically distributed: $\mathbb{P}(X=k) = \mathbb{P}(X=-k)$.
Let $Y = X^2$. [Dependent, unless $Y = 0$].
~ $\mathrm{Cov}(X,Y) = \mathbb{E}(X,Y) - \mathbb{E}(X)\mathbb{E}(Y) = \mathbb{E}(X^3) - \mathbb{E}(X)\mathbb{E}(X^2)$
→ [X symm. dist $\Rightarrow k p_X(k) = -(-k) p_X(-k)$] $= 0 - 0 \cdot \mathbb{E}(X^2) = 0$.

(*) Ex (Random Walk). Let $X_1, X_2, \ldots$ i.i.d. Bernoulli($\frac{1}{2}$) [$X_i \in \{\pm 1\}$], and let $S_n = X_1 + \cdots + X_n$.
Want $\mathrm{Cov}(S_n, S_m)$ for $n \leq m$.

~ Lemma. Covariance is bilinear & symmetric: $\mathrm{Cov}(a(X+Y), bZ) = ab\,\mathrm{Cov}(X,Z) + ab\,\mathrm{Cov}(Y,Z)$.
↳ similar to inner product

$\Rightarrow \mathrm{Cov}(S_n, S_m) = \mathrm{Cov}\left(\sum_{i=1}^n X_i, \sum_{j=1}^m X_j\right) = \sum_{i=1}^n \sum_{j=1}^m \mathrm{Cov}(X_i, X_j)$.
Since X_i's indep, $\mathrm{Cov}(X_i, X_j) = \delta_{ij}\mathrm{Var}(X_i)$ ~ $\sum_{i=1}^n \sum_{j=1}^m \mathrm{Cov}(X_i, X_j) = \sum_{i=1}^n \mathrm{Var}(X_i) = n$. [indep. of m, $m-n$]

(*) Prop. If X, Y uncorrelated, then $\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$.

(*) Proof: $\mathrm{Var}(X+Y) = \mathrm{Cov}(X+Y, X+Y) = \mathrm{Cov}(X,X) + \mathrm{Cov}(Y,Y) + \underbrace{2\mathrm{Cov}(X,Y)}_{=0,\text{ by assump.}} = \mathrm{Var}(X) + \mathrm{Var}(Y)$. ▨

Common Density Functions

2/23/24
Lecture 18

Def: Normal (Gaussian) Distribution

We say that $X \sim (\mu, \sigma^2)$ [is normally distributed with mean μ, variance σ^2] if it has PDF:

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad \text{for } x \in (-\infty, \infty)$$

→ **Def.** We say that $W \sim \text{LogNormal}(\mu, \sigma^2)$ if $\ln(W) \sim N(\mu, \sigma^2)$.

(*) **Ex:** Looking at CDF F_W: $F_W(w) = \mathbb{P}(W \le w) = \mathbb{P}(\ln(W) \le \ln(w))$ [$w > 0$]

→ (Letting $Y = \ln(W)$) $= F_Y(\ln(w)) = \int_{-\infty}^{\ln(w)} f_Y(x)\,dx$

Differentiating: $f_W(w) = F_W'(w) = \frac{1}{w} \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{-\frac{(\ln(w)-\mu)^2}{2\sigma^2}}$ ($\frac{1}{w}$ called the Jacobian factor)

Def: We say that X is Cauchy-Lorentz distributed if it has PDF $f_X(x) = \frac{1}{\pi(1+x^2)}$ [$\approx \frac{1}{\pi x^2}$ as $x \to \infty$]

Def: Exponential Distribution

We say that $X \sim \text{Exponential}(\lambda)$ [with rate λ] if it has PDF:

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x} & x \ge 0 \\ 0 & x < 0 \end{cases}$$

[Can be defined equiv. w/ $x > 0$, $x \le 0$]

↳ (*) Ex: Bus waiting time, nucleus disintegration time

Def: Gamma Function

For $\text{Re}(z) > 0$ [real part of z], define the gamma function $\Gamma(z) = \int_0^\infty x^{z-1} e^{-x}\,dx$

Observe: for $\text{Re}\, z > 0$, $z\Gamma(z) = \int_0^\infty \left(\frac{d}{dx} x^z\right) e^{-x}\,dx = 0 - \int_0^\infty x^z e^{-x}\,dx = \Gamma(z+1)$ [via Integration by Parts]

(from: $x^z e^{-x}\big|_0^\infty = 0$ at $x=0$, $\to 0$ as $x \to \infty$)

In particular: $\Gamma(1) = 1$, $\Gamma(2) = 1\Gamma(1)$, $\Gamma(3) = 2\Gamma(2)$, ..., $\Gamma(n) = (n-1)\Gamma(n-1) \to \Gamma(n+1) = n!$

Expectation for Continuous Random Variables

2/23/24
Lecture 18 (cont.)

Def: Gamma Distribution

We say that $X \sim \text{Gamma}(k, \lambda)$ with shape $k > 0$, rate $\lambda > 0$ if it has PDF:

$$f_X(x) = \begin{cases} \frac{1}{\Gamma(k)} \lambda^k x^{k-1} e^{-\lambda x} & x \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

(*) Ex: (i) If $X_1, \dots, X_n \sim \text{Exponential}(\lambda)$ are independent, then $\sum_{i=1}^n X_i \sim \text{Erlang}(n, \lambda) = \text{Gamma}(n, \lambda)$.

(ii) If $X_1, \dots, X_n \sim N(0,1)$ indep., then $\sum_{i=1}^n \sim \chi^2_n = \text{Gamma}(n/2, 1/2)$. [Chi-squared]

Recall (Expectation): For disc. r.v.'s, defined $\mathbb{E}(X) = \sum k p_X(k)$ if sum absolutely convergent.

Absolute convergence: if $|f(x)|$ integrable & $\int |f(x)| dx < \infty$, is $f(x)$ integrable with $\int f(x) dx < \infty$?

→ (i) Riemann int.: No (ex: take $f(x)$ w/ $f(x) = +1$ for $x \in \mathbb{Q}$, $f(x) = -1$ for $x \notin \mathbb{Q}$ over $[0,1]$)

(ii) Lebesgue int.: Yes, provided f is measurable ($\{x : f(x) > \lambda\} \in \mathcal{F}$)

Def: Expectation for Continuous R.V.s

Given cts r.v. X, we define its expectation to be:

$$\mathbb{E}(X) = \int_{-\infty}^{\infty} x f_X(x) dx$$ [Expected Value]

... provided that $\int_{-\infty}^{\infty} |x| f_X(x) dx < \infty$.

(*) Ex: Assuming the velocity of air particles given by $X \sim N(0, *\text{some T-dep. parameter}*)$

→ use X^2 to find $E = \frac{1}{2}mv^2$, e.g.

(*) Misc. Notes (HW 4-6)

2/26/24
HW 4-6

Sequences of Events (HW4)

Given events E_k for $k \in \mathbb{N}$, define: (i) $\{\text{infinitely many } E_k \text{ occur}\} = \limsup_{k\to\infty} E_k = \bigcap_{n=1}^{\infty}\left[\bigcup_{k=n}^{\infty} E_k\right] = \limsup_{k\to\infty} 1_{E_k}(\omega)$

(ii) $\{\text{finitely many } E_k \text{ occur}\} = \liminf_{k\to\infty} E_k = \bigcup_{n=1}^{\infty}\left[\bigcap_{k=n}^{\infty} E_k\right] = \liminf_{k\to\infty} 1_{E_k}(\omega)$

Expected Value (HW5)

If $X: \Omega \to \mathbb{Z}$ is a non-negative random variable, then $\mathbb{E}(X) = \sum_{n=1}^{\infty} \mathbb{P}(X \geq n)$.

More generally: if G, g are f'ns with $G(k) = \sum_{n=0}^{k} g(n)$, then $\mathbb{E}(G(X)) = \sum_{n=0}^{\infty} g(n)\,\mathbb{P}(X \geq n)$.

Sequences of Bernoulli Trials (HW5)

Given an infinite sequence of Bernoulli(p) trials, $(0 < p < 1)$ and Y_r = position of r^{th} success, Y_r has PMF: $\mathbb{P}(Y_r = k) = \begin{cases} \binom{k-1}{r-1} p^r (1-p)^{k-r} & k \in \mathbb{N},\ k \geq r \\ 0 & \text{otherwise} \end{cases}$

Poisson Random Variables (HW6)

Given $X \sim \text{Poisson}(\lambda)$ and $Y \sim \text{Poisson}(\mu)$ independent, $X + Y \sim \text{Poisson}(\lambda + \mu)$.

Sampling & Estimation (HW6)

Given values $x_1, \ldots, x_n$ for a random variable X with unknown distribution, the sample mean $\bar{X} = \frac{1}{n}(x_1 + \ldots + x_n)$ is an unbiased estimate for the true mean $\mathbb{E}(X)$.

The term $Z = \frac{1}{n}\left([x_1 - \bar{X}]^2 + \ldots + [x_n - \bar{X}]^2\right)$ is an underestimate of true variance: $\mathbb{E}(Z) = \frac{n-1}{n}\text{Var}(X)$.

Expectation for Continuous Random Variables (cont.)

2/25/24

Lecture 19

Recall (Expectation): $\mathbb{E}(X) = \int x f_X(x)\,dx$ is defined given 2 conditions for integrability:

(i) Measurability: If f_X is measurable, then xf_X is measurable

(ii) Finiteness: The integral is absolutely convergent

Theorem: If g is continuous and X a cts. r.v., then $g \circ X$ has expectation given by:

$$\mathbb{E}(g(X)) = \int_{-\infty}^{\infty} g(x) f_X(x)\,dx$$

... provided the RHS exists. [converges absolutely]

Remarks: (i) It is not necessarily true that $g \circ X$ is a continuous r.v.

(*) Ex: $g(x) = \lambda$ [constant] $\rightarrow g \circ X(\omega) = \lambda$, has CDF: 

(ii) It is true that $g \circ X$ is a random variable, i.e. $g \circ X$ is measurable [using g cts.]

(*) **Proof (ii):** Want to show: $g \circ X$ measurable, i.e. $\{\omega : g \circ X(\omega) \leq \lambda\} \in \mathcal{F}$

Know: g is continuous; X is measurable, i.e. $\{\omega : X(\omega) \leq \lambda\} \in \mathcal{F}$.

Want $\{\omega : g \circ X(\omega) \leq \lambda\} \in \mathcal{F}$; equivalently, want $\{\omega : g \circ X(\omega) > \lambda\} \in \mathcal{F}$.

Recall: "g cts." $\Leftrightarrow$ "inverse image of open sets is open", $\{\omega : g \circ X(\omega) > \lambda\} = \{\omega : X(\omega) = g^{-1}(\overbrace{(\lambda, \infty)}^{\text{open interval}})\}$

$\rightsquigarrow$ Need to show: For any open set $O \subseteq \mathbb{R}$, $\{\omega : X(\omega) \in O\} \in \mathcal{F}$. $[X^{-1}(O) \in \mathcal{F}]$

Since X measurable, know that $X^{-1}((a_n, b_n])$ true $\forall$ half-open intervals $(a_n, b_n]$

$\rightsquigarrow$ Only need to show that any open set O can be expressed as countable union of half-open intervals, i.e. $O = \bigcup_{n \in \mathbb{N}} (a_n, b_n]$.

Can construct O as such: $\bigcup_{q \in \mathbb{Q} \cap O} (q - \frac{1}{2}\text{dist}(q, O^c), q + \frac{1}{2}\text{dist}(q, O^c)]$ ← countable union

$\rightsquigarrow X^{-1}(O) = \bigcup_{n \in \mathbb{N}} (a_n, b_n] = \bigcup_{n \in \mathbb{N}} [\{X \leq b_n\} \cap \{X \leq a_n\}^c] \in \mathcal{F}$. ∎

(*) Generalized Quantile Functions

2/28/24
Lecture 20

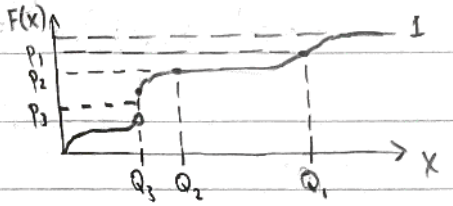
Def: Generalized Quantile Functions

Given a CDF F, the <u>generalized quantile function</u> associated to F is the function $Q:(0,1)\to\mathbb{R}$ def. by:

$$\boxed{Q(p)=\inf\{x: F(x)\geq p\}}$$

<u>Rmks:</u> (i) If F is invertible, <u>$Q=F^{-1}$</u>

(ii) Since F is non-decreasing, $\{x: F(x)\geq p\}$ is either of the form $[a(p),\infty)$ or $(a(p),\infty)$.



In particular, since F is <u>continuous from the right</u>, then if we take a decreasing sequence $(x_n)_n$ converging to $a(p)$ with $F(x_n)\geq p$, then per continuity, $F(\lim_{n\to\infty} x_n)=F(a)=\lim_{n\to\infty}F(x_n)\geq p$. Then <u>$\{x: F(x)\geq p\}=[a(p),\infty)$</u>.

In fact, $a(p)$ is either:

(i) The leftmost point with <u>$F(a)=p$</u>, if $p\in\text{range}(F)$

(ii) The leftmost point with <u>$F(a)=\min(\text{range}(F)\cap[p,\infty))$</u>, if $p\notin\text{range}(F)$.

(*) Since $\text{range}(F)\cap[p,\infty)$ is closed, it has a minimum element

→ Hence: <u>$Q(p)\leq x \iff F(x)\geq p$</u>.

(*) Proof: ($\Rightarrow$) If $x\geq Q(p)=\inf\{x: F(x)\geq p\}$, then $x\in\{y: F(y)\geq p\}\to F(x)\geq p$.

($\Leftarrow$) If $F(x)\geq p$, then $x\in\{y: F(y)\geq p\}\rightsquigarrow x\geq\inf\{y: F(y)\geq p\}=Q(p)$.

(iii) If <u>$p\leq q$</u>, then $\{x: F(x)\leq p\}\subseteq\{x: F(x)\geq q\}$

$\to \inf\{x: F(x)\leq p\}\leq\inf\{x: F(x)\geq q\}\to$ <u>$Q(p)\leq Q(q)$</u>.

(*) We can also define $Q:[0,1]\to\mathbb{R}$, at the risk of obtaining $Q=\infty$ at the ends

Joint CDF & Independence

2/28/24
Lecture 20 (cont.)

Prop. (Simulation of R.V.s) Let F be a CDF with associated quantile function Q. Then given $U \sim \text{Uniform}([0,1])$, the r.v. $Q(U)$ has $F_X(x) = F(x)$.

(*) Proof: $F_X(x) = \mathbb{P}(X \le x) = \mathbb{P}(\{\omega : Q(U(\omega)) \le x\}) = \mathbb{P}(\{\omega : F(x) \ge U(\omega)\})$

$\to = \mathbb{P}(U(\omega) \le F(x)) = \int_{-\infty}^{F(x)} f_U(u)\,du = F(x)$ for $F(x) \in [0,1]$. ▨

Def: Joint CDF

Given 2 R.V.s $X, Y : \Omega \to \mathbb{R}$, the joint CDF of X and Y is the function $F_{X,Y} : \mathbb{R}^2 \to [0,1]$ def. by:

$$F_{X,Y}(x,y) = \mathbb{P}(X \le x \ \&\ Y \le y)$$

More explicitly: $F_{X,Y}(x,y) = \mathbb{P}(\{\omega : X(\omega) \le x\} \cap \{\omega : Y(\omega) \le y\})$.

Def: Independence

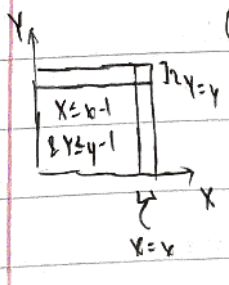
We say two R.V.s X, Y are independent if $\mathbb{P}(X \le x \ \&\ Y \le y) = \mathbb{P}(X \le x)\mathbb{P}(Y \le y)$ for all $x, y \in \mathbb{R}$. Alt: $F_{X,Y}(x,y) = F_X(x)F_Y(y)$.

Lemma. Discrete r.v.'s X, Y are independent in the above sense $[\mathbb{P}(X \le x \ \&\ Y \le y) = \mathbb{P}(X \le x)\mathbb{P}(Y \le y)]$ iff they are independent in the "old" sense $[\mathbb{P}(X = x \ \&\ Y = y) = \mathbb{P}(X = x)\mathbb{P}(Y = y)]$.

(*) Proof: ($\Leftarrow$) $\mathbb{P}(X \le x \ \&\ Y \le y) = \sum_{\substack{n \le x \\ m \le y}} \mathbb{P}(X = n \ \&\ Y = m) = \sum_{\substack{n \le x \\ m \le y}} \mathbb{P}(X = n)\mathbb{P}(Y = m)$

$\to = \left[\sum_{n \le x} \mathbb{P}(X = n)\right]\left[\sum_{m \le y} \mathbb{P}(Y = m)\right] = \mathbb{P}(X \le x)\mathbb{P}(Y \le y)$.

($\Rightarrow$) $\mathbb{P}(X = x \ \&\ Y = y) = \mathbb{P}(X \le x \ \&\ Y \le y) - \mathbb{P}(X \le x \ \&\ Y \le y-1) - \mathbb{P}(X \le x-1 \ \&\ Y \le y) + \mathbb{P}(X \le x-1 \ \&\ Y \le y-1)$



$\to = \mathbb{P}(X \le x)\mathbb{P}(Y \le y) - \mathbb{P}(X \le x)\mathbb{P}(Y \le y-1) - \mathbb{P}(X \le x-1)\mathbb{P}(Y \le y) + \mathbb{P}(X \le x-1)\mathbb{P}(Y \le y-1)$

$= [\mathbb{P}(X \le x) - \mathbb{P}(X \le x-1)][\mathbb{P}(Y \le y) - \mathbb{P}(Y \le y-1)] = \mathbb{P}(X = x)\mathbb{P}(Y = y)$. ▨

(*) Continuous Random Variables

2/29/24
Disc 8

PDFs & CDFs: $P_X(x)\,dx = \mathbb{P}(x \le X \le x+dx)$ (*) P_X, P_X equiv. to f_X, F_X

$\rightsquigarrow \int_a^b P_X(x)\,dx = \mathbb{P}(a \le X \le b)\ [= P_X(b) - P_X(a)]$

Relationships between CDFs

(*) Ex: Let X be a cts r.v. and let $Y = g(X)$ s.t. g strictly increasing & invertible

$\int_a^b P_Y(y)\,dy = \mathbb{P}(a \le Y \le b) = \mathbb{P}(a \le g(X) \le b)$

Since g strictly increasing & g invertible [preserves inequalities]; hence:

$\rightarrow = \mathbb{P}\left(g^{-1}(a) \le X \le g^{-1}(b)\right) = \int_{g^{-1}(a)}^{g^{-1}(b)} P_X(x)\,dx \quad \left[P_Y(b) - P_Y(a) = P_X(g^{-1}(b)) - P_X(g^{-1}(a))\right]$

In particular, since g strictly increasing & invertible, g^{-1} is differentiable

$\rightsquigarrow \int_a^b P_Y(y)\,dy = \int_{g^{-1}(a)}^{g^{-1}(b)} P_X(g^{-1}(y)) \cdot (g^{-1})'(y)\,dy$

(*) Ex: Let $X \sim N(0,1)$, $Y = |X|$

$\rightsquigarrow \int_a^b P_Y(y)\,dy = \mathbb{P}(a \le Y \le b) = \mathbb{P}(a \le X \le b) + \mathbb{P}(-b \le X \le -a) = 2\mathbb{P}(a \le X \le b) = \int_a^b 2P_X(x)\,dx$

$\Rightarrow \int_a^b P_Y(y)\,dy = \int_a^b 2P_X(x)\,dx \Rightarrow P_Y = 2P_X$.

(*) The Cauchy Distribution

Recall: A cts r.v. X is Cauchy-distributed if it has PDF $\boxed{P_X(x) = \frac{1}{\pi(1+x^2)}}$

(*) π a normalizing constant: $\frac{1}{1+x^2} = \frac{d}{dx}\arctan(x)$

$\rightarrow \int_{-\infty}^{\infty} \frac{dx}{1+x^2} = \arctan(x)\Big|_{-\infty}^{\infty} = \pi$

$\rightsquigarrow$ Expect: $X \sim$ Cauchy has $\mathbb{E}(X) = 0$ by symmetry

Reality: $\int_{-\infty}^{\infty} \frac{x\,dx}{\pi(1+x^2)} = \frac{1}{\pi}\left[\frac{1}{2}\log(1+x^2)\right]\Big|_{-\infty}^{\infty} = \text{DNE}$ ($\mathbb{E}(X)$ not defined)

(*) Intuition: $\frac{x}{1+x^2} \approx \frac{1}{x}$ for large x [not integrable]

(*) Can find X with $\mathbb{E}(X)$ but no $\mathrm{Var}(X)$, $\mathbb{E}(X^2)$ but no $\mathbb{E}(X^{n+1})$, etc. similarly

Joint Continuity

3/1/24

Lecture 21

(*) Simulating Random Variables

Recall: We can simulate any distribution [CDF F] using its gen. quantile function and an r.v. $X \sim \text{Uniform}(0,1)$.

$\rightsquigarrow$ Q: How can we simulate r.v.'s $X, Y \sim \text{Uniform}(0,1)$?

A: 1 approach - imagine $X_i \sim \text{Bernoulli}(1/2)$ independent for $n \in \mathbb{N}$ (infinite coin tosses)

$\rightsquigarrow$ set $\boxed{X = \sum_{i=1}^{\infty} \frac{1}{2^i} X_i}$ [exp. independent binary digits] $\rightsquigarrow$ For 2 r.v.s: $X = \sum \frac{1}{2^i} X_{2i}$ & $Y = \sum \frac{1}{2^i} X_{2i+1}$

(*) Coupling

Def: Given two r.v.'s X and Y on $\mathbb{P}$-spaces $(\Omega_1, \mathcal{F}_1, \mathbb{P}_1)$ and $(\Omega_2, \mathcal{F}_2, \mathbb{P}_2)$, respectively, a coupling of X and Y is a new $\mathbb{P}$-space $(\Omega, \mathcal{F}, \mathbb{P})$ with r.v.'s X' and Y' over it, such that X and X', Y and Y' share the same distribution.

(*) Given two r.v.'s X, Y & function f s.t. $F_X(x) = F_Y(f(x))$ $\forall x$, a Skorokhod coupling of X and Y creates r.v.'s X', Y' such that $Y' = f(X')$.

(*) Ex: $X \sim \text{Unif}([0,1])$, $Y \sim \text{Unif}([0,2]) \rightarrow$ coupling: $X' \sim \text{Unif}([0,1])$, $Y' = 2X'$
[generates perfectly correlated values - Y' max when X' max, min when X' min, etc.]

(*) Only works if f exists! If $X = 1$ [constant], e.g., impossible to "get" randomness from $X \rightarrow$ no f exists

Def: Joint Continuity

We say that two r.v.s X, Y are jointly continuous if there is an integrable function $f_{X,Y}: \mathbb{R}^2 \rightarrow [0, \infty)$ (called the joint PDF) such that:

$$\boxed{F_{X,Y}(x,y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f_{X,Y}(x', y')\, dy'\, dx'}$$

Joint Continuity (cont.)

3/1/24
Lecture 21 (cont.)

Prop. If X, Y are jointly continuous, then they are [individually] continuous with $f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dy$ [describing the marginal distribution of X] & likewise for Y.

(*) Proof: $F_X(x) = \mathbb{P}(X \le x) = \lim_{y\to\infty} \mathbb{P}(X \le x \ \&\ Y \le y) = \lim_{y\to\infty} F_{X,Y}(x,y)$ ~(*) Taking a limit since $F_Y(\infty)$ undefined

$= \lim_{y\to\infty} \int_{-\infty}^{x}\int_{-\infty}^{y} f_{X,Y}(x',y')\,dy'\,dx' \to F_X(x) = \int_{-\infty}^{x}\int_{-\infty}^{\infty} f_{X,Y}(x',y')\,dy'\,dx'$

$\rightsquigarrow f_X(x) = \frac{d}{dx}F_X(x) = \frac{d}{dx}\int_{-\infty}^{x}\int_{-\infty}^{\infty} f_{X,Y}(x',y')\,dy'\,dx' = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dy$. ▨

~(*) For Riemann $\int_{-\infty}^{\infty}$, compute as $\int_a^b$ & send a, b to $-\infty, \infty$ separately [Lebesgue - needs obs. convergence]

Prop. If X, Y are individually continuous and independent, then they are jointly continuous with joint PDF $f_{X,Y}(x,y) = f_X(x)\,f_Y(y)$.

(*) Proof: $F_{X,Y}(x,y) = F_X(x)F_Y(y) = \left[\int_{-\infty}^{x} f_X(x')\,dx'\right]\left[\int_{-\infty}^{y} f_Y(y')\,dy'\right] = \int_{-\infty}^{x}\int_{-\infty}^{y} f_X(x')f_Y(y')\,dy'\,dx'$. ▨

~(*) On convergence: as long as the individual $\int^x$, $\int^y$'s are conditionally convergent, then the $\int\int$ term is conditionally convergent provided that it is over "boxes" only.

(*) Note: Conversely, if $f_{X,Y}(x,y) = f_X(x)f_Y(y)$ for variables X, Y, then X, Y are independent

Theorem. If X, Y are jointly continuous, then: $\boxed{\mathbb{E}(g(X,Y)) = \iint_{xy} g(x,y)f_{X,Y}(x,y)\,dx\,dy}$ provided that the integral is absolutely convergent.

→ **Corollary:** Given measurable set $B \subseteq \mathbb{R}^2$ and X, Y jointly continuous, then:

$\mathbb{P}((X,Y) \in B) = \mathbb{E}(1_B(X,Y)) = \iint_{\mathbb{R}^2} 1_B(x,y)f_{X,Y}(x,y)\,dy\,dx$.

Functions of Random Variables

3/1/24

Lecture 21 + Lecture 22

Prop. If X, Y are continuous & independent, then $Z = X+Y$ is continuous and has PDF given by $f_Z(z) = \int_{-\infty}^{\infty} f_X(x) f_Y(z-x)\,dx$.

(*) **Proof:** $F_Z(z) = \mathbb{P}(Z \le z) = \mathbb{P}(X+Y \le z)$

Define $H(z) = \{(x,y) : x+y \le z\} \Rightarrow F_Z(z) = \mathbb{E}(1_{H(z)}(X,Y))$

$\rightsquigarrow = \iint_{\mathbb{R}^2} 1_{H(z)} f_{X,Y}(x,y)\,dy\,dx = \iint_{\mathbb{R}^2} 1_{H(z)} f_X(x) f_Y(y)\,dy\,dx = \int_{-\infty}^{\infty}\int_{-\infty}^{z-x} f_X(x) f_Y(y)\,dy\,dx$ [CDF]

For PDF: $f_Z(z) = \frac{d}{dz} F_Z(z) = \frac{d}{dz}\left[\int_{-\infty}^{\infty}\int_{-\infty}^{z-x} f_X(x) f_Y(y)\,dy\,dx\right] = \int_{-\infty}^{\infty} f_X(x) f_Y(z-x)\,dx$

(*) Check PDF: $\iint f_X(x) f_Y(z-x)\,dx\,dz = \int_{-\infty}^{\infty}\left(\int_{-\infty}^{\infty} f_Y(z-x)\,dz\right) f_X(x)\,dx = 1$. ▨

(*) **Note:** $\int_{-\infty}^{\infty} f_X(x) f_Y(z-x)\,dx$ is a convolution of X and Y: $\underbrace{\cdots\cdots}_{x \text{ points [stimulus]}} \Rightarrow \sim\sim\sim$ [convolution]

(*) **Mollification:** Can take Lebesgue integrals of rough functions by using convolution to approximate with smoother Riemann integrals

(*) **Ex:** Let $X, Y \sim \text{Gamma}(1,2)$ independent; want to find the PDF of $Z = X+Y$.

Solution: Per Prop., $f_Z(z) = \int_{-\infty}^{\infty} f_X(z-y) f_Y(y)\,dy = \int_0^{\infty} \frac{1}{\Gamma(2)}(z-y)e^{-(z-y)} \cdot \frac{1}{\Gamma(2)} y e^{-y}\,dy$ $[\cdot\, 1_{(0,\infty)}(y) 1_{(0,\infty)}(z-y)]$

$\rightsquigarrow = \int_0^z \frac{1}{\Gamma(2)^2}(z-y)y e^{-z}\,dy = \frac{z^3 e^{-z}}{\Gamma(2)^2}\int_0^z \frac{1}{z}\left(1-\frac{y}{z}\right)\left(\frac{y}{z}\right)dy \xrightarrow{u=y/z} = \frac{1}{\Gamma(2)^2} z^3 e^{-z}\int_0^1 (1-u)u\,du$

Rather than solve directly, define $C = \int_0^1 (1-u)u\,du \Rightarrow f_Z(z) = \frac{1}{\Gamma(2)^2} C z^3 e^{-z}$.

Compare to PDF of $\text{Gamma}(1,4)$: $f(x) = \frac{1}{\Gamma(4)} z^3 e^{-z} 1_{(0,\infty)}(z) \rightsquigarrow f_Z = \frac{C\Gamma(4)}{\Gamma(2)^2}[f(z)]$

Since both are PDFs, $\int_0^{\infty} f(x)\,dx = 1 = \int_0^{\infty} f_Z(x)\,dx = \int_0^{\infty} \frac{C\Gamma(4)}{\Gamma(2)^2}[f_Z(x)]\,dx \Rightarrow \frac{C\Gamma(4)}{\Gamma(2)^2} = 1$. [Euler's β function]

$\Rightarrow$ PDF of $X+Y$ is PDF of $W \sim \text{Gamma}(1,4)$.

Functions of Random Variables (cont.)

3/4/24

Lecture 22

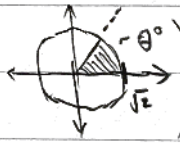
(*) Ex: Let $X, Y \sim N(0,1)$ indep.; want PDF of $Z = X^2 + Y^2$.

Soln: Let $B(z) = \{(x,y) : x^2+y^2 \le z\}$ [ball w/ radius z] $\to$ $\overbrace{\mathbb{P}(X^2+Y^2 \le z)}^{F_Z(z)} = \mathbb{P}((X,Y) \in B(z))$ $[z \ge 0]$

$$\to = \iint_{\mathbb{R}^2} 1_B(x,y) f_{X,Y}(x,y)\,dx\,dy = \iint_B \left[\frac{1}{\sqrt{2\pi}} e^{-x^2/2}\right]\left[\frac{1}{\sqrt{2\pi}} e^{-y^2/2}\right] dx\,dy = \frac{1}{2\pi}\int_0^{2\pi}\int_0^{\sqrt{z}} e^{-r^2/2}\, r\,dr\,d\theta \quad \text{[Polar coords.]}$$

$$\xrightarrow{u = r^2/2} = \int_0^{z/2} e^{-u}\,du \;;\; f_Z(z) = \frac{d}{dz}F_Z(z) = \frac{d}{dz}\int_0^{z/2} e^{-u}\,du \implies \underline{f_Z(z) = \frac{1}{2}e^{-z/2}} \text{ for } z \ge 0$$

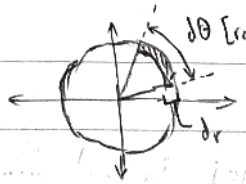
More generally, may want joint distribution of $Z = X^2+Y^2$ and 4-quadrant $\arctan(X,Y) = \Theta$.
[4-quadrant $\arctan(X,Y)$ more stable for X near 0 than $\arctan(Y/X)$; takes values $\in [0, 2\pi)$]

$\to$ As a diagram: $F_{Z,\Theta}(z,\theta) = \mathbb{P}\left((X,Y) \in \right.$  $\left.\right) = \int_0^{\theta}\int_0^{\sqrt{z}} e^{-r^2/2}\,\frac{1}{2\pi}\,r\,dr\,d\Theta$

$$\implies \underline{f_{Z,\Theta}(z,\theta) = \frac{\partial^2}{\partial z\,\partial\theta}F_{Z,\Theta}(z,\theta) = \frac{1}{2\pi}\cdot\frac{1}{2}e^{-1/2}e^{-z}\sqrt{z} = \frac{1}{4\pi}e^{-z/2}} \quad \text{for } z \ge 0,\ \theta \in [0, 2\pi)$$

(*) Obs: Z, Θ independent $[f_{Z,\Theta}(z,\theta) = f_Z(z) f_\Theta(\theta)]$

(*) Polar coordinates: $dx\,dy \mapsto r\,dr\,d\theta$



(*) Steradians: 3D area analogue to radians [2D, length]

(*) Tip: $F_{X,Y}(x,y) = \mathbb{E}\left(1_{\{X \le x, Y \le y\}}\right) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} 1_{\{x' \le x, y' \le y\}} f_{X,Y}(x',y')\,dx\,dy$

$\sim$ Can use indicator functions to "capture" difficult regions for integration

Conditional Density Functions

3/6/24

Lecture 23

Conditioning Continuous R.V.s

Given continuous r.v. X and event A with $\mathbb{P}(A)>0$, can find conditional CDF: $\boxed{\mathbb{P}(X\le x\,|\,A)=\frac{\mathbb{P}(X\le x,\,A)}{\mathbb{P}(A)}}$

(*) Ex: Let $X\sim Exp(\lambda)$ $[f_X(x)=1_{[0,\infty)}\lambda e^{-\lambda x}]$ $\rightsquigarrow \mathbb{E}(X)=1/\lambda$ $[1/\text{rate}]$

$\rightsquigarrow$ Want to find probability law of X given $X\ge a$ $[a>0]$.

Soln: $\mathbb{P}(X\le x\,|\,X\ge a)=\frac{\mathbb{P}(a\le X\le x)}{\mathbb{P}(a\le X)}=\frac{\int_0^\infty 1_{\{a\le x'\le x\}}(x')f_X(x')\,dx'}{\int_0^\infty 1_{\{a\le x'\}}(x')f_X(x')\,dx'}=\frac{\int_a^x f_X(x')\,dx'}{\int_a^\infty f_X(x')\,dx'}$

(*) For simplicity, define $[a,x]$ as empty interval if $x\le a$

$\rightarrow=\frac{e^{-\lambda a}-e^{-\lambda x}}{e^{-\lambda a}}=1-e^{-\lambda(x-a)}\Rightarrow F_X(x\,|\,X\ge a)=\begin{cases}1-e^{-\lambda(x-a)} & x\ge a\\ 0 & x<a\end{cases}$ same as $\mathbb{P}$ law for $Exp(\lambda)$, just "shifted" right by a

(*) Ex: Suppose X,Y jointly continuous; want conditional law of Y given that $a\le X\le b$. $[\&\ \mathbb{P}(a\le X\le b)>0]$

Soln.: $F_Y(y\,|\,a\le X\le b)=\mathbb{P}(Y\le y\,|\,a\le X\le b)=\frac{\mathbb{P}(Y\le y\cap a\le X\le b)}{\mathbb{P}(a\le X\le b)}=\frac{\int_{-\infty}^{y}\int_a^b f_{X,Y}(x',y')\,dx'\,dy'}{\int_a^b f_X(x')\,dx'}$

$\rightarrow$ conditional PDF: $f_Y(y\,|\,X\in[a,b])=\frac{d}{dy}F_Y(y\,|\,X\in[a,b])=\frac{\int_a^b f_{X,Y}(x',y)\,dx'}{\int_a^b f_X(x')\,dx'}$.

(*) Ex (Random Point on Dartboard). Let $U,V\sim Unif([-1,1])$ be independent. Want joint law of U,V conditioned on $X^2+Y^2\le 1$.

Soln.: Want $\mathbb{P}(1_{A\subseteq\mathbb{R}^2}(U,V)=1\,|\,U^2+V^2\le 1)=\frac{\mathbb{P}((U,V)\in[A\cap B])}{\mathbb{P}((U,V)\in B)}=\frac{\int_{-1}^{1}\int_{-1}^{1}1_A 1_B\,du\,dv}{\underbrace{\int_{-1}^{1}\int_{-1}^{1}1_B(u,v)\,du\,dv}_{\text{area of the unit ball}}}$

$\rightarrow=\frac{\int_B 1_A(u,v)\,du\,dv}{\pi\cdot 1^2}$. For $\{U\le x\ \&\ V\le y\}$, $\mathbb{P}(U\le x\ \&\ V\le y\,|\,(U,V)\in B)=\frac{1}{\pi}\int_{-\infty}^{x}\int_{-\infty}^{y}1_B(u,v)\,du\,dv$.

$\Rightarrow$ joint PDF given by $f_{U,V}(u,v\,|\,u^2+v^2\le 1)=\frac{1}{\pi}1_B(u,v)$. $\sim$ (*) Say that (U,V) is uniformly distributed by area (not by radius) on the ball

(*) Functions of Random Variables

3/7/24

Disc. 9

(*) Fundamental Theorem of Calculus

$$\frac{d}{dt}\int_{a(t)}^{b(t)} f(x,t)\,dx = f(b(t),t)\,b'(t) - f(a(t),t)\,a'(t) + \int_{a(t)}^{b(t)} \frac{\partial}{\partial t} f(x,t)\,dx$$

① ② ③ — 3 terms correspond to 3 locations where t appears in orig. integral

(*) Ex: Let X, Y be jointly continuous; want to find PDF of $Z = XY$.

Soln: $F_Z(z) = \mathbb{P}(Z \le z) = \mathbb{P}(XY = z) = \mathbb{P}(XY \le z \cap X \ge 0) + \mathbb{P}(XY \le 0 \cap X \le 0)$

events overlap at $X=0$, but $\mathbb{P}(X=0)=0$ → not an issue

$\rightsquigarrow = \mathbb{P}(Y \le z/x \cap X \ge 0) + \mathbb{P}(Y \ge z/x \cap X \le 0)$

$$= \int_0^\infty \int_{-\infty}^{z/x} f_{X,Y}(x,y)\,dy\,dx + \int_{-\infty}^{0} \int_{z/x}^{\infty} f_{X,Y}(x,y)\,dy\,dx$$

$$f_Z(z) = \frac{d}{dz}F_Z(z) = \int_0^\infty \frac{d}{dz}\left[\int_{-\infty}^{z/x} f_{X,Y}(x,y)\,dy\right]dx + \int_{-\infty}^{0} \frac{d}{dz}\left[\int_{z/x}^{\infty} f_{X,Y}(x,y)\,dy\right]dx$$

$$\rightsquigarrow = \int_0^\infty f_{X,Y}(x, z/x)\,\frac{d}{dz}(z/x)\,dx + \int_{-\infty}^{0} f_{X,Y}(x, z/x)\,\frac{d}{dz}(-z/x)\,dx$$

$$= \int_{-\infty}^{\infty} f_{X,Y}(x, z/x)\,\frac{1}{|x|}\,dx$$

(*) Notes: (i) Is okay to have R.V.s undefined on events with $\mathbb{P}(E) = 0$

(ii) Generally, to find any sort of probability law for continuous R.V.s, typically best to start by looking at CDF

(*) Integral Examples

Recall: $\Gamma(s) = \int_0^\infty x^{s-1} e^{-x}\,dx$, $B(s,t) = \int_0^1 x^{s-1}(1-x)^{t-1}\,dx \left[= \frac{\Gamma(s)\Gamma(t)}{\Gamma(s+t)}\right]$

(i) $I = \int_0^\infty x^7 e^{-4x^3}\,dx$

→ Define $u = 4x^3$, $du = 12x^2$ $\Rightarrow I = \int_0^\infty (u/4)^{5/3} e^{-u}\,\frac{du}{12} = \frac{1}{12 \cdot 4^{5/3}}\Gamma(8/3)$

(ii) $J = \int_0^{3z^2} x^5 (z^2 - x/3)^2\,dx$

→ Define $u = x/3z^2$ $\Rightarrow J(z) = 3^6 z^{16} B(6,3)$

(iii) $K = \int_0^1 x^2 \ln(1/x)^3\,dx$

→ Define $u = \ln(1/x)$ $\Rightarrow K = 1/4\,\Gamma(4)$

Conditional Density Functions (cont.)

3/8/24

Lecture 24

Conditional Density Functions

Recall: For discrete r.v.s, could define $p_{Y|X}(m|n) = \mathbb{P}(Y=m \mid X=n)$ as conditional PMF for Y given $X=n$ (and $\mathbb{P}(X=n)>0$).

In particular: (i) $\mathbb{E}(Y|X=n) = \sum_m m\, p_{Y|X}(m|n)$

(ii) $p_Y(m) = \sum_{n\in\mathbb{N}} p_{Y|X}(m|n)\, p_X(n)$ [Law of Total Probability]

(*) Warning: $p_{Y|X}(m|n)$ undef. if $p_X(n)=0$; can ignore such $p_X(n)$ factor here

From prev. lecture: saw $f_Y(y \mid a\le X\le b) = \dfrac{\int_a^b f_{X,Y}(x,y)\,dx}{\int_a^b f_X(x)\,dx}$

~ If x is a point with $f_X(x)>0$ (and "a little continuous"), can shrink intervals $[a,b]$ down to x (using a sequence $[x-\frac{1}{n}, x+\frac{1}{n}]$, e.g.) to get a limit:

$$\boxed{f_{Y|X}(y|x) = \frac{f_{X,Y}(x,y)}{f_X(x)}} \quad [\text{limit of } f_Y(y \mid a\le X\le b) \text{ as } a,b\to x]$$

=> Lebesgue: Above limit exists for all but an exceptional set E that is negligible w.r.t. X (i.e. $\int_E f_X(x')\,dx' = 0$; $\mathbb{P}(X\in E)=0$)

Similar to discrete case, can define conditional expectation (defined for all $x\notin E$):

$$\boxed{\mathbb{E}(Y|X=x) = \int_Y y\, f_{Y|X}(y|x)\,dy}$$

~ Theorems: (a) $f_Y(y) = \int_x f_{Y|X}(y|x) f_X(x)\,dx$ [Law of Total Probability]

(b) $\mathbb{E}(Y) = \int_x \mathbb{E}(Y|X=x) f_X(x)\,dx$ [Law of Total Expectation]

(c) $\mathbb{E}(g(Y)|X=x) = \int_Y g(y) f_{Y|X}(y|x)\,dy$

(d) $\mathbb{E}(g(Y)h(X)) = \int_x \mathbb{E}(g(Y)|X=x)\, h(x) f_X(x)\,dx$

(*) Proof: (a) RHS is $\int_x \frac{f_{X,Y}(x,y)}{f_X(x)} f_X(x)\,dx = \int_x f_{X,Y}(x,y)\,dx = f_Y(y)$.

(b) RHS is $\int_{-\infty}^{\infty}\left[\int_{-\infty}^{\infty} y f_{Y|X}(y|x)\,dy\right] f_X(x)\,dx = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} y f_{X,Y}(x,y)\,dy\,dx = \mathbb{E}(Y)$.

Conditional Density Functions (cont.)

3/8/24
Lecture 24
+Lecture 25

(*) Proof (cont.)

(c) Skipped; can be feasibly proven if g "nice", more difficult otherwise
[Ex: g constant $\Rightarrow$ $g(Y)$ may be continuous, discrete, or neither]

(d) Per (c), RHS given by:

$$\int_{-\infty}^{\infty}\left[\int_{-\infty}^{\infty} g(y) f_{Y|x}(y|x)\,dy\right] h(x) f_X(x)\,dx = \int_{-\infty}^{\infty} g(y)h(x) f_{X,Y}(x,y)\,dy\,dx = \mathbb{E}(g(Y)h(X)) \quad \blacksquare$$

(*) Note: In all cases, $f_X(x)$ term [i.e. possibly undefined factor] dropped out

(*) Ex: Imagine X, Y have joint PDF $f_{X,Y}(x,y) = e^{-y}\mathbb{1}_{0\le x\le y}$; what can be said about X, Y?

Soln.: (i) $f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dy = \mathbb{1}_{x\ge 0}\int_x^{\infty} e^{-y}\,dy = \mathbb{1}_{x\ge 0}e^{-x}$ $\Rightarrow$ $\underline{X \sim \text{Exp}(1)}$ [$=\text{Gamma}(1,1)$]

(ii) $f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dx = \mathbb{1}_{y\ge 0}\int_0^{y} e^{-y}\,dx = \mathbb{1}_{y\ge 0}ye^{-y}$ $\Rightarrow$ $\underline{Y \sim \text{Gamma}(2,1)}$

(iii) $f_{Y|x}(y|x) = \frac{\mathbb{1}_{0\le x\le y}e^{-y}}{\mathbb{1}_{0\le x}e^{-x}} = \mathbb{1}_{y\ge x}e^{-(y-x)}$ for $x \ge 0$

$\rightsquigarrow$ Define $\underline{Z = Y - X}$ $\Rightarrow$ $f_{Y|x} = f_Z$ [$Z \sim \text{Exp}(1)$]

Lemma. Jointly continuous r.v.s $\underline{X, Y}$ are independent iff $\underline{f_{Y|x}(y|x) = f_Y(y)}$ for all but a negligible set $x \in E$ s.t. $\int_E f_X(x)\,dx = 0$.

(*) Proof: Given by: (i) $f_{X,Y}(x,y) = f_{Y|x}(y|x) f_X(x)$

(ii) X, Y independent iff $f_{X,Y}(x,y) = f_X(x)f_Y(y)$ $\blacksquare$

$\rightsquigarrow$ Soln. (cont.): Z as above $\Rightarrow$ $f_{X,Z}(x,z) = f_{Z|x}(z|x)f_X(x) = (\mathbb{1}_{z\ge 0}e^{-z})(\mathbb{1}_{x\ge 0}e^{-x})$ $\Rightarrow$ X, Z independent.

X, Z indep. $\rightsquigarrow$ can express Y as $Y = X + Z$ [$X, Z \sim \text{Exp}(1)$, generated separately]

$\rightarrow$ (iv) $f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)} = \frac{\mathbb{1}_{0\le x\le y}e^{-y}}{\mathbb{1}_{0\le x}e^{-y}} = \frac{1}{y}\mathbb{1}_{0\le x\le y}$ $\rightsquigarrow$ given $Y = y$, X dist. unif. on $[0, y]$.

Multivariate Normal Distribution

31/11/24

Lecture 25

Notes: (i) We say two PDFs f_1, f_2 are equal if $\int_A f_1(x)\,dx = \int_A f_2(x)\,dx$ for every measurable set A, i.e. $f_1(x) = f_2(x)$ for all x except perhaps on a set E that is negligible [$\int_E f_X(x)\,dx = 0$].

(ii) In cases where a PDF/conditional PDF may or may not be defined, can use well-defined expression $\mathbb{P}(\dots \mid X = x)$ instead.

Def: Positive-Definite

A real matrix Σ is called positive-definite if it is square, symmetric, and has $\vec{\xi} \cdot \Sigma \vec{\xi} > 0$ $\forall\, \vec{\xi} \in \mathbb{R}^n \setminus \{\vec{0}\}$.

Equivalently: Σ is orthogonally diagonalizable, i.e. $\Sigma = ODO^T$ for some $D = \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{bmatrix}$ and has all eigenvalues > 0.

Def: Multivariate Normal Distribution

We say that r.v.s $X_1, \dots, X_n$ are jointly normal if there is a positive-definite matrix Σ (called the covariance matrix) and vector $\vec{\mu} \in \mathbb{R}^n$ (called the mean) s.t.:

$$f_{\vec{X}}(\vec{x}) = \frac{1}{\sqrt{\det(2\pi\Sigma)}} \exp\left(-\frac{1}{2}(\vec{x}-\vec{\mu}) \cdot \Sigma^{-1}(\vec{x}-\vec{\mu})\right)$$

(*) Note: $\det(2\pi\Sigma) = (2\pi)^n \Sigma$ (n = # columns)

Theorem: (a) $\int f_{\vec{X}}(\vec{x})\,d\vec{x}' \left[= \int_{-\infty}^{\infty} \cdots \int f_{\vec{X}}(\vec{x})\,dx_1 \cdots dx_n\right] = 1$

(b) $\int \vec{x} f_{\vec{X}}(\vec{x}')\,d\vec{x}' = \vec{\mu}$

(c) $\int (\vec{x}_i - \vec{\mu}_i)(\vec{x}_j - \vec{\mu}_j) f_{\vec{X}}(\vec{x}')\,dx' = \Sigma_{ij}$

$\Sigma_{ij} = \mathrm{Cov}(X_i, X_j)$

Multivariate Normal Distribution (cont.)

3/11/24
Lecture 25
+ Lecture 26

(*) **Proof:** (a) If $\vec{\mu} \neq 0$, perform change of vars. w/ $\vec{y} = \vec{x} - \vec{\mu}$ [Jacobian: $\det\left(\frac{\partial y_i}{\partial x_j}\right) = \det I_n = 1$]
↝ reduces to cases with $\vec{\mu} = \vec{0}$.

Canfactor $\Sigma = BB^T$ for some invertible B (proof later)
↝ can make change of variables $\vec{x} = B\vec{y}$ with $\frac{\partial x_i}{\partial y_j} = B_{ij}$ $\Rightarrow$ Jacobian $= \det(B)$

$$\Rightarrow \int f_{\vec{x}}(\vec{x})\, d\vec{x} = \int \frac{1}{\sqrt{\det(2\pi\Sigma)}} \exp\left\{-\frac{1}{2}\vec{x}\cdot\Sigma^{-1}\vec{x}\right\} = \frac{1}{\sqrt{\det(2\pi\Sigma)}} \int \exp\left\{-\tfrac{1}{2}B\vec{y}\cdot\Sigma^{-1}B\vec{y}\right\}\det(B)\, d\vec{y}$$

Observations: (i) $B\vec{y}\cdot\Sigma^{-1}B\vec{y} = B\vec{y}\cdot B^{-T}B^{-1}B\vec{y} = B^{-1}B\vec{y}\cdot B^{-1}B\vec{y} = \vec{y}\cdot\vec{y} = |\vec{y}|^2$

(ii) $\det(\Sigma) = \det(BB^T) = \det(B)\det(B^T) = \det(B)^2$

$$\Rightarrow \int f_{\vec{x}}(\vec{x})\, d\vec{x} = \frac{1}{\sqrt{2^n\pi^n \det(B)^2}} \int \exp\left\{-\tfrac{1}{2}|\vec{y}|^2\right\} d\vec{y} = (2\pi)^{-n/2} \int \exp\left\{-\frac{|\vec{y}|^2}{2}\right\} d\vec{y}$$

Recall: $\int e^{-y^2/2}\, dy = \sqrt{2\pi} \Rightarrow (2\pi)^{-n/2} \int \exp\left\{-\frac{|\vec{y}|^2}{2}\right\} d\vec{y} = (2\pi)^{-n/2} \int\cdots\int e^{-y_1^2/2}\cdots e^{-y_n^2/2}\, dy_1\cdots dy_n$
$= (2\pi)^{-n/2}(2\pi)^{n/2} = 1$. ▨

(b) To find mean, solve for each vector coordinate separately:

$$\left[\mathbb{E}(\vec{X})\right]_j = \left[\int \vec{x} f_{\vec{x}}(\vec{x})\, d\vec{x}\right]_j = (2\pi)^{-n/2} \int \left[B\vec{y}\right]_j \exp\left\{-\frac{|\vec{y}|^2}{2}\right\} d\vec{y}$$

(*) $B\vec{y} = (\vec{x} - \vec{\mu})$; if $\vec{\mu}_j \neq 0$, factors out

$\left[B\vec{y}\right]_j = B_{j1}y_1 + \cdots + B_{jn}y_n \Rightarrow = (2\pi)^{-n/2} \int\cdots\int (B_{j1}y_1 + \cdots + B_{jn}y_n)\left(e^{-y_1^2/2}\cdots e^{-y_n^2/2}\right) dy_1\cdots dy_n$

Recall: $\int y e^{-y^2/2}\, dy = 0 \Rightarrow \forall\, i$: $(2\pi)^{-n/2} \int\cdots\int e^{-y_1^2/2}\cdots y_i e^{-y_i^2/2}\cdots e^{-y_n^2/2}\, dy_1\cdots dy_n = 0$
$\Rightarrow \left[\mathbb{E}(\vec{X})\right]_i = 0\ \forall\, j \Rightarrow \mathbb{E}(\vec{X}) = \vec{0}$. ▨

↳ if $\vec{\mu} \neq \vec{0}$, $\mathbb{E}(\vec{X}) = \vec{\mu}$

Multivariate Normal Distribution (cont.)

3/13/24
Lecture 26 (cont.)

(*) Prof (cont.):

(c) Substituting $B\vec{y} = (\vec{x} - \vec{\mu})$:

$$\underline{\mathrm{Cov}(X_i, X_j)} = \mathbb{E}\left[(\vec{X}_i - \vec{\mu}_i)(\vec{X}_j - \vec{\mu}_j)\right] = \int (\vec{x}_i - \vec{\mu}_i)(\vec{x}_j - \vec{\mu}_j) f_{\vec{X}}(\vec{x})\, d\vec{x}$$

$$= (2\pi)^{-n/2} \int (B\vec{y})_i (B\vec{y})_j \exp\left\{-\tfrac{1}{2}|\vec{y}|^2\right\} d\vec{y}$$

$$= (2\pi)^{-n/2} \int \cdots \int \left(\sum_\ell B_{i\ell} y_\ell\right)\left(\sum_m B_{jm} y_m\right) \exp\left\{-\tfrac{1}{2}(y_1^2 + \cdots + y_n^2)\right\} dy_1 \cdots dy_n$$

$$= (2\pi)^{-n/2} \sum_{\ell,m} \left[B_{i\ell} B_{jm} \int_- \int y_\ell y_m e^{-\frac{1}{2}(y_1^2 + \cdots + y_n^2)} dy_1 \cdots dy_n \right]$$

→ Two cases: (i) $\underline{\ell \neq m}$: Integral $= 0$ $\left(\int y_\ell e^{-\frac{y_\ell^2}{2}} dy_\ell = 0\right)$

(ii) $\underline{\ell = m}$: $\int y_\ell^2 e^{-y_\ell^2/2} dy = \sqrt{2\pi}$ [Variance of $N(0,1)$]

$$\Rightarrow \mathrm{Cov}(X_i, X_j) = \sum_{\ell,m} B_{i\ell} B_{jm} \delta_{\ell m} = \sum_{\ell,m} B_{i\ell} I_{\ell m} B^T_{mj} = (BIB^T)_{ij}$$

$$= (BB^T)_{ij} = \underline{\Sigma_{ij}}. \quad \blacksquare$$

Factoring $\Sigma = BB^T$

Prop. For any positive-definite matrix Σ, $\exists$ invertible matrix B such that $\boxed{\Sigma = BB^T}$.

(*) **Proof:** Since Σ is symmetric & positive-definite, then it is orthogonally diagonalizable, i.e. $\exists\, O, D$ s.t. $\Sigma = ODO^T$, where the columns of O form an orthonormal basis of eigenvectors [importantly: $OO^T = I$] and D is a diagonal matrix of the form:

$$D = \begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{pmatrix}, \quad \lambda_i > 0 \ \forall\, i$$

(*) The Cholesky Factorization

3/13/24
Lecture 26 (cont.)

(*) Proof (cont.)

Then we can factor Σ as: $\Sigma = \underbrace{\left[O\begin{pmatrix}\sqrt{\lambda_1} & & \\ & \ddots & \\ & & \sqrt{\lambda_n}\end{pmatrix}O^T\right]}_{B}\underbrace{\left[O\begin{pmatrix}\sqrt{\lambda_1} & & \\ & \ddots & \\ & & \sqrt{\lambda_n}\end{pmatrix}O^T\right]^T}_{B^T}$

(*) Alt: $\Sigma = \left(O\begin{pmatrix}\sqrt{\lambda_1} & & \\ & \ddots & \\ & & \sqrt{\lambda_n}\end{pmatrix}\right)\left(O\begin{pmatrix}\sqrt{\lambda_1} & & \\ & \ddots & \\ & & \sqrt{\lambda_n}\end{pmatrix}\right)^T$ ↪ (*) Recall: $(AB)^T = B^T A^T$

The Cholesky Factorization

Prop. If C symmetric & positive-definite, then $\exists$ matrix L lower-triangular s.t. $C = LL^T$.

(*) **Proof:** Let B be the invertible matrix such that $C = BB^T$.

Then $\exists$ orthonormal matrix Q, upper triangular matrix R such that $B^T = QR$.
[QR decomposition; via Gram-Schmidt, e.g.]

Then $C = BB^T = (R^TQ^T)(QR) = R^TR$.

Setting $L = R^T$, then L is lower-triangular with $C = LL^T$. ▨

Multivariate Parameter Estimation

3/13/24
Lecture 26
*Lecture 27

(*) Estimation for Multivariate Normal Distributions

Setup: Imagine we have a machine that produces errors $X \sim N(0, a^2)$, but our view of these errors is impeded by measurement error $Z \sim N(0, \sigma^2)$, i.e. our "measured" error $Y = X + Z$ has conditional PDF given X:

$$f_{Y|X}(y|x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(y-x)^2}{2\sigma^2}} = f_Z(y-x)$$

Want to find joint law of X, Y: $f_{X,Y}(x,y) = f_{Y|X}(y|x) f_X(x) = \frac{1}{2\pi\sqrt{\sigma^2 a^2}} \exp\left\{-\frac{x^2}{2a^2} - \frac{(x-y)^2}{2\sigma^2}\right\}$

$\rightsquigarrow$ Interpreting as joint normal distribution: $-\frac{x^2}{2a^2} - \frac{(x-y)^2}{2\sigma^2} = -\frac{1}{2}\begin{bmatrix} x \\ y \end{bmatrix} \cdot \Sigma^{-1} \begin{bmatrix} x \\ y \end{bmatrix}$

$$\Rightarrow \Sigma^{-1} = \begin{pmatrix} \frac{1}{a^2} + \frac{1}{\sigma^2} & -\frac{1}{\sigma^2} \\ -\frac{1}{\sigma^2} & \frac{1}{\sigma^2} \end{pmatrix} \Rightarrow \Sigma = \begin{pmatrix} a^2 & a^2 \\ a^2 & a^2 + \sigma^2 \end{pmatrix}, \quad \det(\Sigma) = a^2\sigma^2$$

Maximum Likelihood Estimation

Problem: Given measurement $Y = y$, what is MLE $\hat{x}_{ML} = \operatorname{argmax}_x f_{Y|X}(y|x)$ for X?

Soln.: $f_{Y|X} = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y-x)^2}{2\sigma^2}\right) \rightsquigarrow \hat{x}_{ML}(y) = y$. [Unbiased estimate]

(*) **Issue:** Although the MLE is unbiased, it also ignores any knowledge about the prior distribution of X (is inefficient - unused knowledge).

Multivariate Parameter Estimation (cont.)

3/15/24
Lecture 27 (cont.)

(*) Estimation for Multivariate Normal Distributions (cont.)

Want to find a new estimator that leverages our prior knowledge about X, Y.

(i) Maximum Posterior Probability

Want to find posterior probability $f_{X|Y}(x|y) = \frac{f_{X,Y}}{f_Y}$.

Know: (i) X, Y jointly normal $\Rightarrow$ X, Y individually normal

(ii) $\mathbb{E}\left(\begin{bmatrix} X \\ Y \end{bmatrix}\right) = \vec{\mu} = 0 \Rightarrow \mathbb{E}(X), \mathbb{E}(Y) = 0$

(iii) $\mathrm{Var}(Y) = \mathrm{Cov}(Y,Y) = \Sigma_{22} = a^2 + \sigma^2$

$\Rightarrow Y \sim N(0, a^2+\sigma^2)$.

$$\rightarrow f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)} = \frac{\sqrt{2\pi(a^2+\sigma^2)}}{2\pi\sqrt{a^2\sigma^2}} \exp\left(-\frac{a^2+\sigma^2}{2a^2\sigma^2}\left[x - \frac{a^2}{a^2+\sigma^2}y\right]^2\right)$$

Want to find value for X maximizing $f_{X|Y}(x|y)$

$$\rightsquigarrow \hat{X}_{MAP} = \underset{x}{\mathrm{argmax}}\, f_{X|Y}(x|y) = \frac{a^2}{a^2+\sigma^2}y$$

(*) Is a biased estimator: underestimates y

(*) Observation: Is a "smarter" estimator than MLE: balances expected % of error from X vs. from Z based on relative magnitudes of a^2, σ^2

Multivariate Parameter Estimation (cont.)

3/15/24

Lecture 27 (cont.)

(*) Estimation for Multivariate Normal Distributions (cont.)

Alt. approach: take $\mathbb{E}(X|Y=y)$ directly

(ii) Minimum Mean Squared Error

Recall: $(X|Y=y) \sim N\left(\frac{a^2 y}{a^2+\sigma^2}, \frac{a^2\sigma^2}{a^2+\sigma^2}\right)$

$$\Rightarrow \mathbb{E}(X|Y=y) = \int x f_{X|Y}(x|y)\,dx = \frac{a^2}{a^2+\sigma^2} y = \hat{x}_{MSE}$$

(*) Def: Mean Squared Error

The mean squared error of an estimate $\hat{x}$ is defined as the value:

$$MSE = \mathbb{E}\left([X-\hat{x}]^2\right)$$

Prop. $\hat{x}_{MSE} = \mathbb{E}(X|Y=y)$ is the value for $\hat{x}$ that minimizes MSE for X given Y.

(*) Proof: $\mathbb{E}\left([X-\hat{x}]^2 | Y=y\right) = \mathbb{E}\left([X-\mathbb{E}(X|Y=y)+\mathbb{E}(X|Y=y)+\hat{x}(y)]^2 | Y=y\right)$

$$\rightarrow = \mathbb{E}\left([X-\mathbb{E}(X|Y=y)]^2 + [\hat{x}(y)-\mathbb{E}(X|Y=y)]^2 + \right.$$

$$\left. + 2\left[\mathbb{E}(X-\mathbb{E}(X|Y=y))\mathbb{E}(\mathbb{E}(X|Y=y)-\hat{x}(y))\right] \,\middle|\, Y=y\right)$$

$$= \underbrace{Var(X|Y=y)}_{\text{indep. of } \hat{x}(y)} + \underbrace{[\hat{x}(y)-\mathbb{E}(X|Y=y)]^2}_{\text{Not a r.v.; is a fixed \# given } y} + 0$$

$\rightsquigarrow$ MSE is minimized w/ $MSE_{min} = Var(X|Y=y)$ iff $\hat{x}(y) = \mathbb{E}(X|Y=y) = \hat{x}_{MSE}$. ▨

(*) Misc. Notes (HW 7-10)

3/15/24

HW 7-10

Independent Geometric Variables (HW7)

Given $X_1, X_2 \sim \text{Geometric}(p)$, $\mathbb{P}(X_1 \geq X_2)$ is, by symmetry, $\mathbb{P}(X_1 \geq X_2) = \frac{1}{2-p}$.

Indicator Random Variables (HW7)

Given events E_i for $1 \leq i \leq n$: $\mathbb{P}(\text{no } E_i \text{ occurs}) = \mathbb{E}\left\{\prod_{i=1}^{n}[1-1_{E_i}]\right\} = \sum_{k=0}^{n}(-1)^k \sum_{|S|=k} \mathbb{E}\left\{\prod_{i \in S} 1_{E_i}\right\}$

CDF Properties (HW7)

For any r.v. X, its CDF F_X has the properties:

(i) F_X is non-decreasing: $x \leq y \Rightarrow F_X(x) \leq F_X(y)$

(ii) $\lim_{x \to -\infty} F_X(x) = 0$, $\lim_{x \to \infty} F_X(x) = 1$

(iii) F_X is "càdlàg" [continuous from right, limit from left], i.e. for any $x \in \mathbb{R}$:

(a) $\lim_{y \uparrow x} F_X(y) = \mathbb{P}(X < x)$

(b) $\lim_{y \downarrow x} F_X(y) = F_X(x)$

Cosine (HW8)

Given $\Theta \sim \text{Unif}([0, 2\pi])$, $X = \cos(\Theta)$ has CDF $F_X(k) = 1 - \frac{\arccos(k)}{\pi} \Rightarrow$ PDF $f_X(k) = \frac{1}{\pi\sqrt{1-x^2}}$

Functions of Multiple R.V.s (HW8)

Given (real-valued) r.v.s X, Y and $g: \mathbb{R}^2 \to \mathbb{R}$, then $g(X, Y)$ is a random variable [is measurable].

The Gamma Distribution (HW8)

Given $X \sim \Gamma(k, \lambda)$, then for $n \in \mathbb{N}$ X has n^{th} moment $\boxed{\mathbb{E}(X^n) = \frac{\Gamma(n+k)}{\Gamma(k)}\lambda^{-n} = k(k+1)\cdots(n+k-1)\lambda^{-n}}$

If $Y \sim N(0, \sigma^2)$, then $Y^2 \sim \Gamma\left(\frac{1}{2}, \frac{1}{2\sigma^2}\right)$.

(*) Misc. Notes (HW7-10, cont.)

3/15/24
HW7-10 (cont.)

The Normal Distribution (HW8)

Given $Y \sim N(0, \sigma^2)$, then for $n \in \mathbb{N}$ Y has n^{th} moment $\mathbb{E}(Y^n) = \begin{cases} \sigma^n (n-1)!! & n \text{ even} \\ 0 & n \text{ odd} \end{cases}$

The Box-Muller Transform (HW9)

Given $U, V \sim \text{Unif}(0,1)$, the r.v.s $X = \sqrt{-2\ln(U)}\cos(2\pi V)$, $Y = \sqrt{-2\ln(U)}\sin(2\pi V)$ are independent and $\sim N(0,1)$.

The Chi-Squared Distribution (HW9)

An r.v. X is $X \sim \chi^2_\nu$ $[\nu > 0]$ iff it has PDF $f_X(x) = \begin{cases} \frac{1}{2^{\nu/2}\Gamma(\nu/2)} x^{\frac{\nu}{2}-1} e^{-x/2} & x \geq 0 \\ 0 & \text{otherwise} \end{cases}$

Given $X \sim \chi^2_\nu$, $Y \sim \chi^2_k$ independent, $X + Y \sim \chi^2_{\nu+k}$.

(*) Euler's Beta Integral: $\int_0^1 u^{\frac{\nu}{2}-1}(1-u)^{\frac{k}{2}-1}\,du = \frac{\Gamma(\nu/2)\Gamma(k/2)}{\Gamma\left(\frac{\nu+k}{2}\right)}$

The Student t-Distribution (HW9)

Given $X \sim \chi^2_\nu$ and $Z \sim N(0,1)$ indep., we say that $T = Z\sqrt{\frac{\nu}{X}}$ follows a Student t-distribution with ν degrees of freedom.

$$\text{PDF: } f_T(x) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\pi\nu}\,\Gamma(\nu/2)}\left(1 + \frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}} \quad [x \in \mathbb{R}]$$

(*) Misc. Notes (HW 7-10, cont.)

3/15/24
HW 7-10
(cont)

(*) Another Euler integral: $\boxed{\int_{-\infty}^{\infty} \left(1+\frac{1}{\nu}t\right)^{-\frac{\nu+1}{2}} dt = \frac{\sqrt{\nu\pi}\,\Gamma(\nu/2)}{\Gamma(\frac{\nu+1}{2})}}$

Rejection Sampling (HW10)

Given continuous f'n and PDF f_X with: (i) $0 \le f(x) \le M \;\forall\; x \in \mathbb{R}$

(ii) $f(x) = 0 \;\forall\; |x| \ge M$

$\rightsquigarrow$ Given $U_n \sim Unif([-M, M])$ and $V_n \sim Unif([0, M])$ independent, can sample from X by taking U_n, V_n until $f(U_n) \le V_n$

$\Rightarrow$ returned value U_n has PDF $f_{U_n}(u) = f_X(u)$.

Sums of Exponential R.V.s (HW10)

Given $X_1, \dots, X_n \sim Exp(\lambda)$, $Y_i = \sum_{j=1}^{i} X_j$ for $1 \le i \le n$, the variables $Y_1, \dots, Y_i$ have joint PDF given by:

$$\boxed{f_{Y_1,\dots,Y_i}(y_1,\dots,y_i) = \lambda^i e^{-\lambda y_i} \mathbb{1}_{0 \le y_1 \le \dots \le y_i}}$$

Convolution of Normals

Prop. If $X_1 \sim N(\mu_1, \sigma_1^2)$, $X_2 \sim N(\mu_2, \sigma_2^2)$ are independent, then $\boxed{X_1 + X_2 \sim N(\mu_1+\mu_2, \sigma_1^2+\sigma_2^2)}$.

(*) Proof: Defining $Z = X_1 + X_2$, have: $f_Z(z) = \int_{-\infty}^{\infty} f_{X_1}(z-y) f_{X_2}(y)\,dy = \frac{1}{2\pi\sigma_1\sigma_2} \int_{-\infty}^{\infty} \exp\left\{-\frac{(z-y-\mu_1)^2}{2\sigma_1^2} - \frac{(y-\mu_2)^2}{2\sigma_2^2}\right\} dy$

Simplifying the exponent, obtain: $\left[\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2}\right] y^2 - 2\left[\frac{z-\mu_1}{\sigma_1^2} + \frac{\mu_2}{\sigma_2^2}\right] y + \left[\frac{(z-\mu_1)^2}{\sigma_1^2} + \frac{(\mu_2)^2}{\sigma_2^2}\right]$

$\rightsquigarrow$ Can complete the square: $ay^2 + by + c = a\left(y + \frac{b}{2a}\right)^2 + c - \frac{b^2}{4a}$

$\rightarrow$ Eventually obtain $f_Z(z) = \frac{1}{\sqrt{2\pi(\sigma_1^2+\sigma_2^2)}} \exp\left\{-\frac{(z-\mu_1-\mu_2)^2}{2(\sigma_1^2+\sigma_2^2)}\right\}$.