

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

(1) Probability Spaces & Probabilities - 98	(3) Multiple Random Variables - 125
(i) Sample Spaces - 98	(i) Joint PMFs - 125
(ii) Probabilities - 99	(ii) Independent Random Variables - 127
(iii) Continuity of Probability - 102	(iii) Covariance - 130
(iv) Conditional Probability - 102	
(v) Likelihoods & Posterior Probabilities - 106	(4) Continuous Random Variables - 131
(vi) Independence - 107	(i) Real-Valued Random Variables - 131
	(ii) Continuous Random Variables - 132
(2) Discrete Random Variables - 111	(iii) Common Density Functions - 133
(i) Discrete Random Variables - 111	(iv) Expectation for Continuous R.V.s - 136
(ii) Discrete Probability Distributions - 113	
(iii) Expected Value - 116	(5) Multiple Continuous Random Variables - 142
(iv) Variance & Standard Deviation - 119	(i) Joint CDF & Independence - 142
(v) Conditional PMFs & Expectation - 121	(ii) Joint Continuity - 144
(vi) Parameter Estimation - 124	(iii) Conditional Density Functions - 148
	(iv) Conditional Expectation - 150
	(v) Multivariate Normal Distribution - 154

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 $\rightarrow \Omega = \{1, 2, 3, 4, 5, 6\}$

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

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

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

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

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

iv) $\Omega = \text{population}$ [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 dict. 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

11.8 / 24

Lecture 11

May want to find/assign probabilities; 2 approaches:

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

→ Problem: Only works for countable Ω 's, 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 $P: \mathcal{F} \rightarrow \mathbb{R}$

subject to:

- i) $P(A) \geq 0 \forall A \in \mathcal{F}$
- ii) $P(\emptyset) = 0$; $P(\Omega) = 1$
- iii) If $A_n \in \mathcal{F}$ are disjoint for $n \in \mathbb{N}$, $P\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} P(A_n)$. [countable/ σ -additivity]

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

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 $A_3 = A_4 = \dots = A_n = \emptyset$; 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, $P(A) = \frac{|A|}{|\Omega|}$ 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 $\exists p: \Omega \rightarrow [0, 1]$ so that $\sum_{\omega \in \Omega} p(\omega) = 1$, then $P(A) = \sum_{\omega \in A} p(\omega)$
 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 P satisfying $P([a, b]) = b - a$ [$= \frac{b-a}{1-0}$].

(*) 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: $(A^c \cap B^c)^c = A$ doesn't happen and B doesn't happen
 ii) $(A \cap B)^c = A^c \cup B^c$ ~ $(A \cap B)^c = "A$ happens and B happens" [A and B]

→ Observe: i) per $(A \in \mathcal{F} \Rightarrow A^c \in \mathcal{F})$ and ([σ -additivity]), $P(A) + P(A^c) = P(A \cup A^c) = 1$; in particular, $(P(A) \geq 0 \vee A \in \mathcal{F})$ and $(P(A) = 1 - P(A^c)) \Rightarrow 0 \leq P(A) \leq 1$.
 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 $A \setminus B = A \cap B^c \in \mathcal{F}$ [$(A \setminus B)^c = "A but not B"$]
 iv) if $A, B \in \mathcal{F}$, then $P(A \cup B) = P(A \bigcup_{i,j} [B \setminus A]) = P(A) + P(B \setminus A) = P(A) + P(B) - P(A \cap B)$
 (or) $P(B) = P([B \cap A] \bigcup_{i,j} [B \setminus A])$

(*) $P(A \cup B \cup C \dots) \rightarrow$ use inclusion-exclusion

v) if $A \subseteq B \in \mathcal{F}$, then $P(B) = P(A \cup [B \setminus A]) \geq P(A)$ [$A \subseteq B \Rightarrow A$ 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 $P(\bigcup_{n=1}^N A_n) = P(A_1) + P(\bigcup_{n=2}^N A_n) \leq \dots \leq P(A_1) + P(A_2) + \dots$
 (by induction)

(*) Lady Tasting Tea

1/11/24

Due 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 preparing 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) $\Omega = \# \text{ correct}$ [i.e. $\Omega = \{0, 1, \dots, 8\}$]

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

ii) $\Omega = \text{selections of A, B cups}$ [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) $\Omega = \text{selections of A cups}$ [e.g. $(1, 2, 4, 7)$]

A correct | # ways to guess → total # possible choices: $\binom{8}{4} = 70$ possible 4-subsets of $\{1, \dots, 8\}$

$$0 \quad \binom{4}{4} \cdot \binom{4}{0} = 16$$

↑ for A
↓ from B

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

$$1 \quad \binom{4}{3} \cdot \binom{4}{1} = 36$$

$$\rightarrow P(\text{all 4 correct}) = \frac{\# \text{ choices w/ 4 correct}}{\text{total # choices}} = \frac{1}{70} \approx 1.7\%$$

reject null hypothesis

$$2 \quad \binom{4}{2} \cdot \binom{4}{2} = 16$$

implies: $\{w \in \Omega : w \text{ has all 4 correct}\}$

$$3 \quad \binom{4}{1} \cdot \binom{4}{3} = 16$$

$$4 \quad \binom{4}{0} \cdot \binom{4}{4} = 1$$

Conditional Probability

17/12/24

Lecture 3

Theorem: Continuity/Monotone Convergence of Events

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

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

(*) 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,

$$\text{Then per } \sigma\text{-additivity: } \mathbb{P}\left(\bigcup_{n=1}^{\infty} D_n\right) = \sum_{n=1}^{\infty} \mathbb{P}(D_n) = \lim_{N \rightarrow \infty} \sum_{n=1}^N \mathbb{P}(D_n)$$

$$\text{Since } \bigcup_{n=1}^N D_n = A_N, \text{ then } \mathbb{P}\left(\bigcup_{n=1}^{\infty} D_n\right) = \mathbb{P}\left(\bigcup_{n=1}^{\infty} A_n\right) = \lim_{N \rightarrow \infty} \sum_{n=1}^N \mathbb{P}(D_n) = \lim_{N \rightarrow \infty} \mathbb{P}(A_N).$$

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

$$\text{Per axioms, } \mathbb{P}\left(\bigcup_{n=1}^{\infty} A_n^c\right) = \mathbb{P}\left(\left[\bigcup_{n=1}^{\infty} A_n\right]^c\right) = 1 - \mathbb{P}\left(\bigcup_{n=1}^{\infty} A_n\right). \text{ Similarly, } \lim_{N \rightarrow \infty} (A_N^c) = \lim_{N \rightarrow \infty} (1 - \mathbb{P}(A_N))$$

$$\rightarrow 1 - \lim_{N \rightarrow \infty} \mathbb{P}(A_N) \rightarrow 1 - \mathbb{P}\left(\bigcup_{n=1}^{\infty} A_n\right) = 1 - \lim_{n \rightarrow \infty} \mathbb{P}(A_n). \square$$

(*) 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:

$A \mapsto \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$	$\left[\frac{\text{"P(A and B)"} }{\mathbb{P}(B)} \right]$
--	---

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

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

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

Conditional Probability (cont.)

1/12/24

Lecture 3

(cont.)

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

$$\text{i)} P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(B)}{P(B)} = 1; P(\emptyset|B) = \frac{P(\emptyset \cap B)}{P(B)} = \frac{P(\emptyset)}{P(B)} = 0, [\text{trivial}]$$

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

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

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

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

Remark: Probability can be interpreted as a mathematical description of ignorance [Bayes], e.g.

ignorance & initial conditions. In this setting, $P(\cdot|B)$ represents a new state of

ignorance, after observing B happening. Namely, $P(\cdot|B)$ does not "re-weight" events, i.e.:
for events $A_1, A_2 \subseteq B$, $\frac{P(A_1)}{P(A_2)} = \frac{P(A_1|B)}{P(A_2|B)}$.

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

→ Q: What happens if there are multiple observations?

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

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

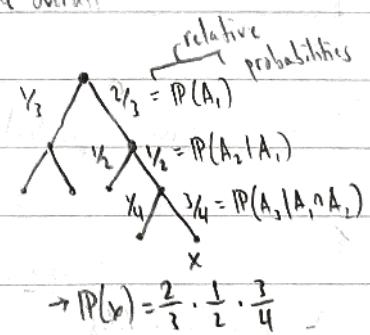
In some cases, conditional/relative probabilities can be used to understand the overall distribution [absolute probabilities].

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

$$\rightarrow P(A_1 \cap A_2 \cap \dots \cap A_n) = P(A_1|A_2 \cap \dots \cap A_n) \cdot P(A_2 \cap \dots \cap A_n)$$

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

("multiplication rule" - by induction)



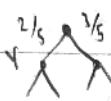
Bayes' Rule

1/17/24

Lecture 4 (* Ex: Probability

(for not)

AMC8: Given 2 yellow & 3 red counters [distinguishable] drawn without replacement, what is the probability of drawing both yellows without exhausting the reds?

i) Naïve soln.: Draw a tree  → sum PPs of tree leaves: $\frac{3}{5}$ [simple method]

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

→ redefine $\Omega = \text{set of possible permutations after shuffling } (\Omega = 5!)$

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

Partitions

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

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

$$P(A) = \sum_j P(A \cap B_j) = \sum_j P(A|B_j) P(B_j)$$

Def: Bayes' Rule

Given events A, B with $P(A), P(B) > 0$:

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

Idea: We can imagine that B, B^c represent an underlying truth, and that A represents an attempt to measure it. If $P(A|B), P(A|B^c)$ 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

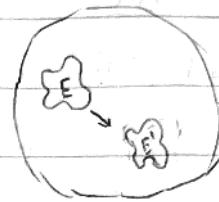
Disc 2

(*) Measure (Motivation)

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

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

i.e. $P(E) = P(E')$ for any translation of E , since 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 $f: \Omega_2 \rightarrow \Omega_1$, and define σ -algebra Σ on Ω_2 ; show that $f^{-1}(\Sigma)$ is a σ -algebra on Ω_1 .

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

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

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

(*) The Boy-Girl Paradox

Say Mr. Smith has two children (2 genders: M/F, $P=50\%$ each). What is $P(\text{both are boys})$ given that:

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

ii) at least one is a boy?

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

	B	G
B	BB	BG
G	GB	GG

$$\rightarrow P(\text{at least one boy}) = \frac{3}{4} = \frac{1}{3} \approx 33\%$$

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

$$\Omega = \{\text{B, G}\} \times \{\text{SMTURFS}\} \times \{\text{B, G}\} \times \{\text{SMTURFS}\}$$

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

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

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

(*) Intuition:

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

$$P(\text{one boy on Tues | two boys}) = \frac{12}{49} \approx 27\%$$

Likelihoods

1/19/24

Lecture 5

Bayes' Rule (ii):

Given a [countable] partition $\{B_i\}$ of Ω with $P(B_i) > 0 \forall i$, and event A with $P(A) > 0$:

$$P(B_i | A) = \frac{P(B_i \cap A)}{P(A)} = \frac{P(A | B_i)P(B_i)}{\sum_k P(A | B_k)P(B_k)}$$

[General case]

Likelihoods

(*) Ex: Let a set of lions be partitioned into J(juvenile), M(male adult), and F(female adult).

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

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

→ Call $P(B_i | \cdot)$ a likelihood (vs. $P(\cdot | A)$, 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 $A_j = \{\text{red die rolls a } j\}$, $B = \{\text{sum of dice is 6}\}$, $C = \{\text{blue die rolls an even \#}\}$.

via Bayes' Rule

j	$P(A_j)$	$P(B A_j)$	$P(A_j B)$	$P(B \cap C A_j)$	$P(A_j B \cap C)$
1	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{36}$	0	0
2	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{36}$	$\frac{1}{6}$	$\frac{1}{2}$
3	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{36}$	0	0
4	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{36}$	$\frac{1}{6}$	$\frac{1}{2}$
5	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{36}$	0	0
6	$\frac{1}{6}$	0	0	0	0

"prior probabilities" \uparrow "likelihoods" \uparrow "posterior probabilities" \uparrow

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

Independence

1/19/24

Lecture 5
(cont.)

Def: Independence

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

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 and B ; A^c and B^c .

(*) Proof: $P(A \cap B^c) = P(A) - P(A \cap B) = P(A)(1 - P(B)) = P(A)P(B^c)$. \square

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

(*) Proof: Let $P(B) = 0$; then $P(A \cap B) \leq P(B) = 0 \rightarrow P(A \cap B) = 0 = P(A)P(B)$.
Let $P(B) = 1$; then $P(B^c) = 0$. \square

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

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

Def: Conditional Independence

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

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

(*) 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 $P(A \cap B) \approx P(A) \approx P(B)$, i.e. A and B are not independent ($\leq 4 \text{ ft.} \sim \text{bad at math}$).

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

[Idea: height, math ability correlation disappears if " $\leq 6 \text{ y.o.}$ " is known]

Independence (cont.)

1/22/24

Lecture 6 Independence (cont.)

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

Hypothesis: the flips are independent

$\Omega^{(2)}$	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	$1-p$

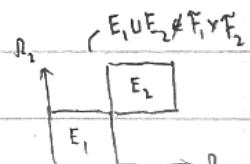
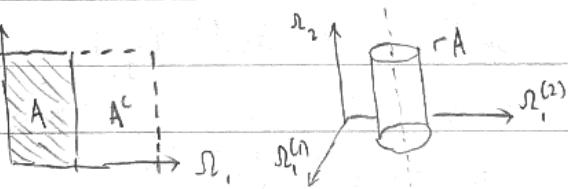
Independence of Multiple Experiments

Idea: Want to model carrying out 2 experiments $(\Omega_1, \mathcal{F}_1, P_1), (\Omega_2, \mathcal{F}_2, P_2)$ independently

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

Want to model various events $[A \in \mathcal{F}_1] \times [B \in \mathcal{F}_2]$; ex: $A^c \times \Omega_2 = "A \text{ doesn't happen}"$

(*) 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: $\mathcal{F}_1 \otimes \mathcal{F}_2 := \text{the } \sigma\text{-algebra generated by } \mathcal{F}_1 \times \mathcal{F}_2$.

Theorem. There is a unique probability P on $\mathcal{F}_1 \otimes \mathcal{F}_2$ with $P(A \times B) = P_1(A) \cdot P_2(B)$.

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

Remarks: 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 $P(\{E_j\})$ w/ size 2^n (can reduce to $2^n - 1 - n$)

(*) Misc. Notes

1/22/24

(*) Ex: Independence

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

(cont.)

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

+HW 1-3

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

σ -Algebras (HV2)

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 $(\Omega_\alpha, \mathcal{F}_\alpha, P_\alpha)$ 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:

$$(\prod_{\alpha \in S} E_\alpha) \times (\prod_{\alpha \in S^c} \Omega_\alpha) \quad (*) \text{ E_α may be interpreted as the events "of interest"}$$

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

(*) We ignore the uncountable product $\prod_{\alpha \in A} P_\alpha(\Omega_\alpha) = 1$

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

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

$$P((\prod_{\alpha \in S} E_\alpha) \times (\prod_{\alpha \in S^c} \Omega_\alpha)) = \prod_{\alpha \in S} P_\alpha(E_\alpha)$$

(*) Ex: Infinite Independent Coin Tosses

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

→ Can make a sequence: take indices $n \in \mathbb{N} \rightsquigarrow$ sample space: $\{\text{functions } \mathbb{N} \rightarrow \{H, T\}\}$

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

Discrete Random Variables

1/24/24

Lecture 7

Def: Discrete Random Variables

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

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

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 $X: \Omega \rightarrow \mathbb{Z}$]

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 $p_X: \mathbb{R} \rightarrow [0, 1]$ [$X(\Omega) \rightarrow [0, 1]$] defined by:

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

Remark: The PMF determines the law/distribution of a r.v.

Def: Given an event $E \in \mathcal{F}$, its indicator r.v. is the function $\mathbf{1}_E$ def. by

[random variable]

$$\mathbf{1}_E(\omega) = \begin{cases} 1 & \text{if } \omega \in E \\ 0 & \text{if } \omega \notin E \end{cases}$$

(*) Ex: Tossing Coins

Let $\Omega = \{H, T\}^{\mathbb{N}}$, and define indicators $X_1, \dots, X_n: X_k = \mathbf{1}_{\{\text{kth throw is heads}\}}$ / 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

Results generally similar, except in the case of very heavily weighted prior dists. \leftarrow repeated assignments; no notion of "prior distributions"
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 $A = \{\text{Alien beats Bob}\}$, $B = \{\text{Bob beats Charlie}\}$, $C = \{\text{Charlie beats Alien}\}$.

→ 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 P 's p_1, p_2 (but don't know which is which); after n pulls (k successes) on one lever, determine if the lever was p_1 or p_2 .

Want: $P(\text{was } p_2 \mid n \text{ pulls}, k \text{ rewards}) \rightarrow \frac{P(n, k \mid p_2) P(p_2)}{P(n, k)}$

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

$$\rightarrow (\text{eventually}) P(p_1 \mid n, k) = \frac{1}{1 + (p_2/p_1)(\frac{1-p_2}{1-p_1})^{n-k}}$$

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

Discrete Probability Distributions

1/26/24

Lecture 8

Observation: All PMFs satisfy (by def.): i) $0 \leq p_n(x) \leq 1 \forall n$

$$\text{ii)} \sum_{n \in \mathbb{Z}} p_x(n) = 1$$

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

(*) Pf: Let $\Omega = \mathbb{Z}$ and $\mathcal{F} = P(\mathbb{Z})$. Define $P(A) = \sum_{n \in A} p_x(n)$ and $X: \Omega \rightarrow \mathbb{Z}$ via $X(n) = n$.

(Check: $P(X^{-1}(\{m\})) = P(\{m\}) = \sum_{n=m} p_x(n) = p_x(m)$. (Can also check P-axioms.) \square)

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

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

(*) Ex: The indicator r.v. $\mathbf{1}_E$ is $\text{Bernoulli}(P(E))$ distributed.

Conversely, any $X \sim \text{Bernoulli}(p)$ is an indicator r.v. [Namely, of $X^{-1}(\{1\})$].

Def. Given $0 \leq p \leq 1$ and $n \in \mathbb{N}$, we say that a r.v. X is Binomial(n, p) or that

$X \sim \text{Binomial}(n, p)$ if its PMF is given by:

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

(*) Note: We see immediately that $p_x(k) \geq 0$, 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.)

1/20/24

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

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

$$\begin{aligned} \{\text{1 success}\} &= \{(S, F, \dots, F), (F, S, F, \dots, F), \dots, (F, \dots, F, S)\} \\ &= \bigcup_{i=1}^n \{\text{F}^{i-1} \times \{S\} \times \{\text{F}\}^{n-i}\} \quad [\text{alt: } = \bigcup_{i=1}^n \{F, \dots, F, S, F, \dots, F\}] \end{aligned}$$

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

$$\begin{aligned} \rightarrow P\left(\bigcup_{i=1}^n \{\{F\}^{i-1} \times \{S\} \times \{F\}^{n-i}\}\right) &= \sum_{i=1}^n P(\{\{F\}^{i-1} \times \{S\} \times \{F\}^{n-i}\}) \\ &= \sum_{i=1}^n (1-p)^{i-1} p = n(1-p)^{n-1} p \quad [= \binom{n}{1} (1-p)^{n-1} p] \end{aligned}$$

General case: $P(\{k \text{ successes}\}) = \binom{n}{k} (1-p)^{n-k} p^k$ [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]

alt) # failures before success [Y = X - 1]

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

(*) Ex: Geometric Distribution

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

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

Observe: as $p > 0$, $P(X \geq n) \rightarrow 0$ as $n \rightarrow \infty$:

$\rightarrow P(\text{success infinitely many trials}) = 1 - P(\text{never succeed})$

$$\begin{aligned} &= 1 - P(\bigcap_{n=1}^{\infty} X \geq n) = 1 - \lim_{n \rightarrow \infty} P(X \geq n) \quad [\text{Continuity Thm.}] \\ &= 1; \quad P(\text{never succeed} = 0) \end{aligned}$$

(*) 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 (Check Poisson PMF: i) $p_X(k) = 0 \forall k \in \mathbb{N}$

(cont.) 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 $e = \exp(1) \sim \sum_{k=0}^{\infty} \frac{1^k}{k!} e^k = \exp(1)\exp(-1) = \exp(0) = 1$. \square

Def: Expected Value

Given discrete r.v. X , we define its expected value to be $\boxed{\mathbb{E}(X) = \sum_k k p_X(k)}$, provided

that the sum converges absolutely, i.e. $\sum_k |k p_X(k)| < \infty$.

(*) Also called the "mean", "average", or "expectation" of X

(*) Absolute convergence gives us that the value of $\mathbb{E}(X)$ does not depend on the "order" of the k 's.

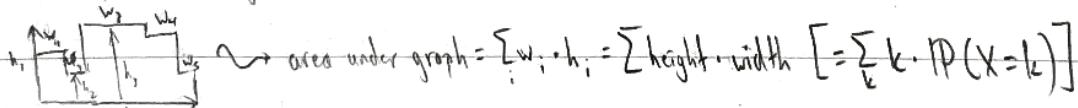
(*) Thom (Liemann): If a sum is conditionally convergent but not absolutely convergent

(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 converging to any arbitrary value in \mathbb{R} .

(*) 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}

(e.g., e.g.: pick terms > 0 until above 47, then terms < 0 until below 47, etc.)

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



Expected Values of Distributions

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

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

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

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

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

Expected Value (cont.)

1/31/24

Lecture 10

Prop: 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.)

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

$$\text{- Recall: } \sum_{k=0}^{\infty} k x^k [x \leq 1] = \frac{1}{(1-x)^2}$$

$$\rightarrow \sum_{k=0}^{\infty} (k+1) (1-p)^k p = \frac{p}{(1-(1-p))^2} = \frac{p}{p} = 1 \quad \square$$

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

$$\text{Differentiate } e^{\lambda} = \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} \rightarrow \mathbb{E}(X) = \lambda$$

Expected Values

$$\text{i) Bernoulli}(p): \mathbb{E}(X) = p$$

$$\text{ii) Binomial}(n, p): \mathbb{E}(X) = np$$

$$\text{iii) Geometric}(p): \mathbb{E}(X) = \frac{1}{p}$$

$$\text{iv) Poisson}(\lambda): \mathbb{E}(X) = \lambda$$

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

Then, as $n \rightarrow \infty$,

$$P_{X_n}(k) \rightarrow P_X(k) \quad [X_n \text{ converge in distribution to } X]$$

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

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

$$\frac{\log(1 - \lambda/n) - \log(1)}{\lambda/n} = \frac{1}{\lambda} \log(1 - \lambda/n) \Big|_{n=0} = \frac{-\lambda}{(1-\lambda)x} = -\lambda \rightarrow (1 - \lambda/n)^n = \exp(-\lambda) = e^{-\lambda}. \quad \square$$

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

$$\rightarrow \text{Notice: } \frac{n(n-1)\dots(n-k+1)}{n^k} \rightarrow 1, (1 - \lambda/n)^k \rightarrow 1 \text{ as } n \rightarrow \infty.$$

$$\rightarrow \text{As } n \rightarrow \infty, P_{X_n}(k) \rightarrow \frac{\lambda^k}{k!} (1 - \lambda/n)^n = \frac{\lambda^k}{k!} e^{-\lambda} = P_X(k). \quad \square$$

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

Prop:

$$\mathbb{E}(g(X)) = \sum_k g(k) P(X=k)$$

(*) Proof: $\mathbb{E}(g(X)) = \sum_k g(k) P(g(X)=k)$

$$= \sum_k \sum_{g(k)=l} P(X=k)$$

$$= \sum_k \sum_{g(k)=l} g(k) P(X=k) = \sum_k g(k) P(X=k) \quad \square$$

(*) Discrete Probability Distributions

2/1/24

Disc 4

Discrete Random Variables (Review)

A discrete random variable is a function $X: \Omega \rightarrow E$ [E countable]. $\sim n$ typically used during calculation

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

$$(*) \text{ Ex: } \mathbb{P}(X = k) = C \frac{1}{k!} \text{ for some normalizing constant } C; \text{ want } \mathbb{P}(X \text{ even})$$

$\left. \begin{array}{l} \sum_k \mathbb{P}(X = k) = 1 \\ \text{works even w/o knowing } C \end{array} \right\}$

$$\rightarrow \mathbb{P}(X = 2) = \sum_{k=0}^{\infty} \mathbb{P}(k) = \sum_{k=1}^{\infty} \mathbb{P}(X = 2k) = \sum_{k=1}^{\infty} \left(\frac{1}{(2k)!} \right) = \frac{1}{4} \sum_{k=1}^{\infty} \frac{1}{k!} = \frac{1}{4} e = \frac{1}{4}.$$

Vandermonde's identity: $\sum_{k=0}^r \binom{n}{k} \binom{m}{r-k} = \binom{n+m}{r}$

2 methods of proof: 1) Counting: choose r : $\begin{array}{c} \textcircled{1} \dots \textcircled{r} \\ \textcircled{1} \dots \textcircled{r} \end{array}$ pick \textcircled{k} of top, $\textcircled{r-k}$ of bottom

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

$$\rightarrow = \sum_{r=0}^{n+m} \left(\sum_{a+b=r} \binom{n}{a} \binom{m}{b} \right) x^r \xrightarrow{\text{choose } r} \sum_{r=0}^{n+m} \binom{n+m}{r} = \sum_{r=0}^{n+m} \left(\sum_{k=0}^r \binom{n}{k} \binom{m}{r-k} \right). \square$$

(*) Ex (Poisson)

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

$$\begin{aligned} \mathbb{P}(X+Y=k) &= \sum_{l=0}^k \mathbb{P}(X=l \wedge Y=k-l) = \sum_{l=0}^k \mathbb{P}(X=l) \mathbb{P}(Y=k-l) = \sum_{l=0}^k \frac{\lambda^l}{l!} e^{-\lambda} \frac{\mu^{k-l}}{(k-l)!} \\ &\rightarrow = e^{-\lambda-\mu} \sum_{l=0}^k \binom{k}{l} \frac{1}{l!} \lambda^l \mu^{k-l} = e^{-\lambda-\mu} \frac{1}{k!} \sum_{l=0}^k \binom{k}{l} \lambda^l \mu^{k-l} = e^{-(\lambda+\mu)} \frac{(\lambda+\mu)^k}{k!} \end{aligned}$$

$$\rightarrow (X+Y) \sim \text{Poisson}(\lambda+\mu).$$

(*) Ex (ii)

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

$\rightarrow Q$: Let $U \sim \text{Poisson}(\lambda=1)$, $P \sim \text{Poisson}(\mu=58)$ [decay/1 sec.]. If I measure 50 decay/1 sec.,

what is \mathbb{P} that the sample is of P? Assume initial $\mathbb{P}(U) = \mathbb{P}(P) = 1/2$.

$$\rightarrow \text{Use Bayes' Rule: } \mathbb{P}(P | N=50) = \frac{\mathbb{P}(N=50 | P) \mathbb{P}(P)}{\mathbb{P}(N=50 | P) \mathbb{P}(P) + \mathbb{P}(N=50 | U) \mathbb{P}(U)} = \frac{0.99 \dots 97}{63.95}$$

Variance & Standard Deviation

2/2/24

Lecture 11

Def: Variance & Standard Deviation

Defs: 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:

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

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

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

Prop. i) $\text{Var}(aX) = a^2 \text{Var}(X)$.

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

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

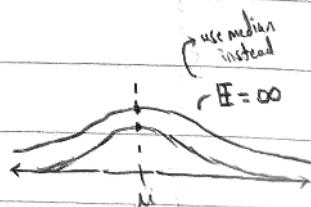
ii) $\text{Var}(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2$ ~ more useful definition for computation

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

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

(*) 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 $\text{Var}(X)$ both very sensitive to outliers; vs. median (more stable, but more difficult mathematically)



Convexity & Jensen's Inequality

2/1/24

Lecture 11
(cont.)

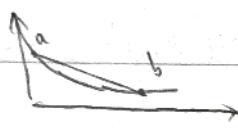
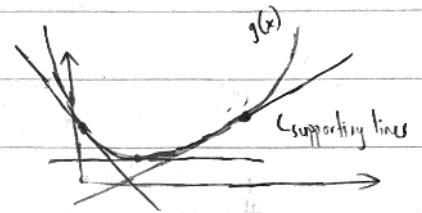
Jensen's
Inequality

If $g: \mathbb{R} \rightarrow \mathbb{R}$ is convex, then

$$\mathbb{E}(g(X)) \geq g(\mathbb{E}(X)).$$

Def: Convex Functions

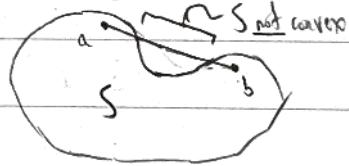
We consider a function $f: \mathbb{R}^n \rightarrow \mathbb{R}$ convex if, for every point x in the domain, we can find a supporting line [plane, hyperplane, etc.] passing through x that has value $\leq f(x)$ at all points x :
 $\underbrace{L \text{ below the graph}}$



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 .

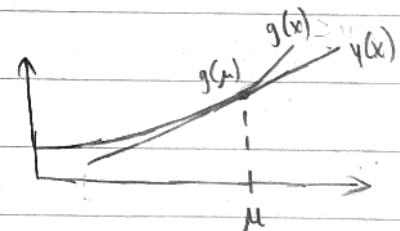
(*) Def: Convex Sets

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., $\geq g(x) \geq m[x - \mu] + g(\mu)$, with equality at $x = \mu$.



$$\mathbb{E}(g(X)) = \sum_k g(k) \mathbb{P}(X=k)$$

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

$$= m\mu - m\mu + g(\mu) = g(\mu) = g(\mathbb{E}(X)). \square$$

Conditional PMF & Expectation

2/2/24

Lecture 11

+ Lecture 12

Def: Conditional PMF & Expectation

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

$$P_{X|A}(k) = P(X=k | A)$$

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

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

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 $P(B_j) \geq 0$, we have:

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

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

(*) Proof

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

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

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

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

→ 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 (4) Proof (cont.)

- (cont.) (i): Lemma: Can exchange order of integration [summation] if either:
- 1) The sum is absolutely convergent (Fubini) (alt: all terms have same sign)
 - 2) All terms are non-negative; may result in infinite sum (Tonelli)

Want to show: $\sum_{j,k} |k| P(X=k|B_j) P(B_j)$ absolutely convergent [∞]
 \rightarrow (all terms ≥ 0) = $\sum_k |k| \left(\sum_j P(X=k|B_j) P(B_j) \right) = \sum_k |k| P(X=k) = E(|X|) < \infty$ (by assumption)

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

$E(X)$
 $\forall L_{\text{now}}: \sum_j \left[\sum_k |k| P(X=k|B_j) \right] P(B_j) < \infty$ (from before) & $P(B_j) > 0 \quad \forall j$
 \rightarrow freeze j : $\sum_k |k| P(X=k|B_j) P(B_j) \xrightarrow{L \rightarrow 0} \infty \quad \forall j$.

\square

(*) Ex: Partition Theorem

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

\rightarrow Intuitive answer: λp successes

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

$$\rightarrow E(X|B_n) = np$$

$$\rightarrow E(X) = \sum_n E(X|B_n) P(B_n)$$

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

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

(*) 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] $P(X=k) = \frac{c}{k^2} \rightarrow \mathbb{E}(X) = c \sum k \frac{1}{k^2} = c \sum \frac{1}{k}$, diverges

ii) [$\text{Var}(X)$ diverges] $P(X=k) = \frac{c}{k^3} \rightarrow \text{Var}(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2$

$\mathbb{E}(X^2) = c \sum k^2 \frac{1}{k^3} = c \sum \frac{1}{k} \text{, diverges}$

(*) 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 ($X \sim \text{Geometric}(\frac{1}{2})$, 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 $\mathbb{E}(N = \# \text{ of people with the correct hat})$?

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

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

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

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

Expectation linear even when X_i 's are dependent

(*) Application: Unique IP Addresses

i.e. via string of IP addresses

Problem: Given some large number of visitors to a website, want to track # unique visitors space-efficiently.

Solution: Given IP address (e.g. 192.168.1.1), map to random binary string (0's and 1's); across visitors,

keep track & largest number of leading 1's seen M .

→ Ex: Given n unique visitors, $\mathbb{E}(M) = \sum_{k=1}^{\infty} P(M \geq k) = \sum_{k=1}^{\infty} \left(1 - \left(1 - \frac{1}{2^k}\right)^n\right) \approx \log_2(n)$ 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 $P(X=k|p)$. 2 conditionally internally

Optimization - can differentiate to find a maximum

$$0 = \frac{\partial}{\partial p} (p(1-p)^{k-1}) = 1 - pk \longrightarrow p = \frac{1}{k}, \text{ if } 0 < p < 1$$

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. $E(\frac{1}{X}|p) \stackrel{?}{=} p$

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

$$\rightarrow = \frac{p}{1-p} \log\left[\frac{1}{1-p}\right] = \frac{p}{1-p} \log\left[\frac{1}{p}\right], \quad \text{(*)}$$

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

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

→ notice: MLE $\hat{\mu} = \frac{1}{p}$

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

Multiple Random Variables

2/9/24

Lecture 17

Multiple [Discrete] Random Variables

Given a sample space Ω , we may define multiple random variables $X, Y: \Omega \rightarrow \mathbb{Z}$ (e.g.) and 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 $X, Y: \Omega \rightarrow \mathbb{Z}$, the joint PMF of X and Y is the function $p_{X,Y}: \mathbb{Z}^2 \rightarrow [0, 1]$

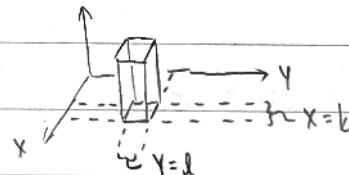
defined by:

$$p_{X,Y}(k, l) = P(X=k \wedge Y=l)$$

- $\{X=k \wedge Y=l\}$ also denoted: $\{\omega \in \Omega : X(\omega)=k \wedge Y(\omega)=l\}, X^{-1}(\{k\}) \cap Y^{-1}(\{l\})$

- 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:

$$p_X(k) = \sum_l p_{X,Y}(k, l) \quad [\text{Marginal distribution}]$$

Prop.: $\mathbb{E}(g(X, Y)) = \sum_{k,l} g(k, l) p_{X,Y}(k, l)$ [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:

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

Multiple Random Variables (cont.)

2/12/24

Lecture 14

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

(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).

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

$$= \sum_{k,l} \left(\sum_{m=0} m \mathbf{1}_{g(k, l) = m} \right) P(X=k, Y=l)$$

$$= \sum_{k,l} g(k, l) P(X=k, Y=l) \quad - (*) \text{ equality: either both are finite, or neither is finite}$$

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.

$$\begin{aligned} \mathbb{E}(X) + \mathbb{E}(Y) &= \left(\sum_k k P(X=k) \right) + \left(\sum_l l P(Y=l) \right), \\ &= \sum_{k,l} (k+l) P(X=k, Y=l). \end{aligned}$$

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

(*) Indicator Random Variables

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

Independence of Random Variables

2/12/24

Lecture 14

(*) 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 $X_i = \mathbb{1}_{\text{person receives } i^{\text{th}} \text{ hat}}$ $\leadsto \mathbb{E}(X_i) = 1/n$

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

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

$\mathbb{E}(X^2) = \mathbb{E}([X_1 + \dots + X_n]^2)$; 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 $\binom{n}{2}$

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

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

Def: Independent Random Variables

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

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

- Equivalently: any finite subset of the k_i 's

(ii) Indicators $\mathbb{1}_A, \mathbb{1}_B, \dots$ 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 f, g .

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

(*) Proof: $P(f(X)=r \& g(Y)=l) = \sum_k \mathbb{1}_{f(X)=r} \mathbb{1}_{g(Y)=l} P(X=k \& Y=l)$

$$\rightarrow = \sum_k \mathbb{1}_{f(X)=r} \mathbb{1}_{g(Y)=l} P(X=k)P(Y=l) = \left(\sum_k \mathbb{1}_{f(X)=r} P(X=k) \right) \left(\sum_l \mathbb{1}_{g(Y)=l} P(Y=l) \right) = P(f(X)=r)P(g(Y)=l). \square$$

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 $\sum_{k,l} |kl| P(X=k \& Y=l) = \sum_{k,l} |k| \cdot |l| \cdot P(X=k) P(Y=l)$
 $= \sum_k |k| P(X=k) \left[\sum_l |l| P(Y=l) \right]$
 $\xrightarrow{\text{Tonelli}} = \left(\sum_k |k| P(X=k) \right) \left(\sum_l |l| P(Y=l) \right) = \mathbb{E}(|X|)\mathbb{E}(|Y|)$

By assumption, the product $\mathbb{E}(|X|)\mathbb{E}(|Y|)$ is finite [converges absolutely].
 → can perform some 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 $f = \mathbb{1}_{\{X=k\}}(X)$, $g = \mathbb{1}_{\{Y=l\}}(Y)$ bounded functions for any $k, l \in \mathbb{Z}$. Then:

$$\begin{aligned} P(X=k \& Y=l) &= \mathbb{E}(f(X)g(Y)) \\ &= \mathbb{E}(f(X))\mathbb{E}(g(Y)) = P(X=k)P(Y=l) \text{ for arbitrary } k, l. \quad \square \end{aligned}$$

(*) Note on Prop: To see the role of independence, consider integrals $\int_0^1 x^{-1/2} dx, \int_0^1 y^{1/2} dy = 2$
 → can take $\int_0^1 \int_0^1 x^{-1/2} y^{1/2} dxdy$, but not $\int_0^1 x^{-1/2} y^{1/2} dx [=\infty]$.

(*) Expectation of Multiple Variables

2/15/24

Disc 6

Recall (Linearity of Expectation): For any r.v.'s X_1, \dots, X_n : $E(\sum_i X_i) = \sum_i E(X_i)$, and exists if all $E(X_i)$'s exist. [X_i 's need not be independent!]

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

→ initial hypothesis: $P(10^{\text{th}} \text{ flip is H}) = P(10^{\text{th}} \text{ flip is T}) = \frac{1}{2}$ (by symmetry)

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

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

(*) Ex: Shuffle n cards 1, ..., n and deal 1 by 1 (keeping track of running max of cards dealt)

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

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

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

→ $P(j^{\text{th}} \text{ was i}^{\text{th}} \text{ largest}) \text{ etc. } i \text{ all orderings equally likely} \rightarrow P(j^{\text{th}} \text{ is max of } 1^{\text{st}}, \dots, j^{\text{th}}) = \frac{1}{j}$

$$\rightarrow E(N) = \sum_{i=1}^n E(X_i) = \sum_{i=1}^n P(j^{\text{th}} \text{ is max of } 1^{\text{st}}, \dots, i^{\text{th}}) = \sum_{i=1}^n \frac{1}{i} \quad [= \text{nth harmonic } H_n; i \approx \log n]$$

(*) 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]

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

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

Observe: For any clause $C_i = (B_j \vee B_k \vee B_l)$ [or negations of B_j, k, l], $P(C_i \text{ satisfied}) = \frac{7}{8}$

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

In particular: if $n=7$, $E(N) = \frac{49}{8} = 6.125 > 6 \rightarrow$ 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 & uniformly; want $E(\text{total } \# \text{ drawn, incl. repeats, by the time } n \text{ unique collected})$

→ Define $T_0=0, T_1, \dots, T_n$ [$T_j = \# \text{ drawn when } j^{\text{th}} \text{ unique collected}$] → $E(N) = E(T_n - T_{n-1}) + \dots + E(T_1 - T_0)$; $T_{j+1} - T_j$ are $\sim \text{Geo}(\frac{n-j}{n})$ [129]

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 $\text{var}(S)$ exists and is $\text{var}(S) = \text{var}(X) + \text{var}(Y)$.

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

$$\rightarrow P(S=n) = \sum_k P(X=n-k \cap Y=k) = \sum_k P(X=n-k) P(Y=k)$$

(ii) See p. 126

$$(iii) \text{var}(S) = \mathbb{E}[(S - \mathbb{E}(S))^2] = \mathbb{E}[(X+Y - \mathbb{E}(X) - \mathbb{E}(Y))^2]$$

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

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

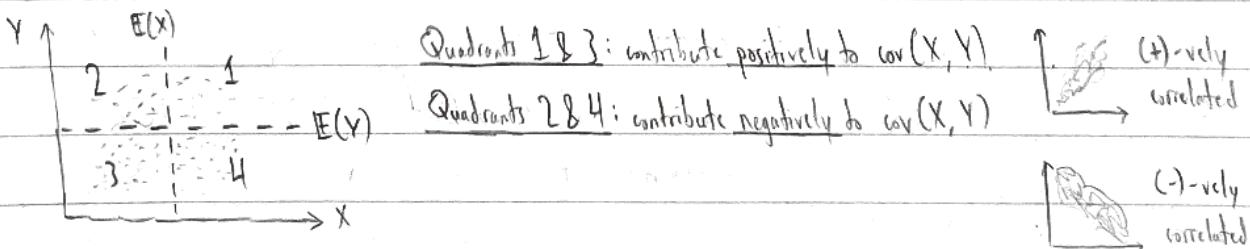
$$= \text{var}(X) + \text{var}(Y). \quad \square$$

Defn Covariance

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

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

(* Covariance



- X, Y independent $\rightarrow \text{cov}(X, Y) = 0$; $X = Y \rightarrow \text{cov}(X, Y) = \text{var}(X)$

- Variance, covariance both sensitive to outliers

Cumulative Distribution Functions

2/16/23

Lecture 16

(*) Convolution

The function $p_z(n) = \sum_{k \in \mathbb{Z}} p_x(n-k) p_y(k)$ is a discrete convolution of p_x and p_y .

(cont.)

Def: Real-Valued Random Variable

A (real-valued) random variable on probability space (Ω, \mathcal{F}, P) is a function

$X: \Omega \rightarrow \mathbb{R}$ that is measurable, i.e.: for all $x \in \mathbb{R}$, $\{X \leq x\}$ belongs to \mathcal{F} .

→ Need to go beyond PMFs, $p_x [p_x(n) = P(X=n)]$ in continuous case:

Def: Cumulative Distribution Function (CDF)

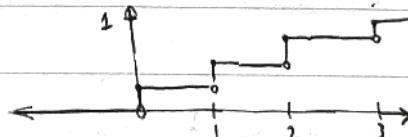
Given real-valued r.v. X , we define the cumulative distribution function [CDF] of X to be the function $F_X: \mathbb{R} \rightarrow [0, 1]$ defined by:

$$F_X(x) = P(X \leq x)$$

Notes on CDFs

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

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



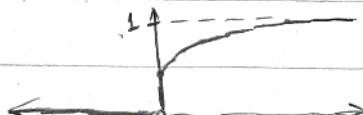
[Discrete]

(ii) bus waiting times:



[Continuous]

(iii) inches of rain:



[Neither cts. nor disc.]

Continuous Random Variables

2/21/24

Lecture 17 Prop. (i) $\lim_{x \rightarrow -\infty} F_X(x) = 0$, $\lim_{x \rightarrow \infty} F_X(x) = 1$

(ii) F_X is non-decreasing: if $x \leq y$, $F_X(x) \leq F_X(y)$.

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

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

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

Def: Continuous Random Variables

A r.v. $X: \Omega \rightarrow \mathbb{R}$ is called continuous if \exists integrable function $f_X: \mathbb{R} \rightarrow \mathbb{R}$ such that:

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

The function f_X is called the probability density function [PDF] of X .

Rmk. (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 not true that every continuous CDF F_X admits a probability density (see: Cantor function)

→ Class of functions [r.v.'s] that do admit PDFs called absolutely continuous

Continuous Random Variables (cont.)

2/21/24

Lecture 17

(cont.)

Def: Absolute Continuity

We say a function F is absolutely continuous if, for every $\epsilon > 0$, $\exists \delta > 0$ 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 $\sum_{i=1}^n |F(b_i) - F(a_i)| < \epsilon$,

Prop. If $X: \Omega \rightarrow \mathbb{R}$ is a cts. r.v., then

for every $a \leq b$,

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

$$\mathbb{P}(a \leq X < b) = \mathbb{P}(a \leq X \leq b) = \int_a^b f_X(x) dx$$

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

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

Observe: $\{a < X \leq b\} = \bigcup_{n \in \mathbb{N}} \{a < X \leq b - \frac{1}{n}\} \rightarrow \mathbb{P}(a < X \leq b) = \lim_{n \rightarrow \infty} \mathbb{P}(a < X \leq b - \frac{1}{n})$

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

Similarly: $\{a \leq X \leq b\} = \bigcap_{n \in \mathbb{N}} \{a - \frac{1}{n} < X \leq b\} \rightarrow \mathbb{P}(a \leq X \leq b) = \lim_{n \rightarrow \infty} [F_X(b) - F_X(a - \frac{1}{n})] = F_X(b) - F_X(a)$. ◻

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], $a > 0$

→ $F_Y(y) = \mathbb{P}(Y \leq y) = \mathbb{P}(aX \leq y) = \mathbb{P}(X \leq y/a) \rightarrow F_Y(y) = F_X(y/a)$.

Differentiating: $F'_Y(y) = f_Y(y) = \frac{1}{a} F'_X(y/a) = \frac{1}{a} f_X(y/a)$.

Def: Uniform Distribution

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

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

(*) Covariance & Correlation

2/22/24

Disc. 7

$$\text{Recall: (i) } \text{Var}(X) = \mathbb{E}((X-\mu)^2) = \mathbb{E}(X^2) - \mathbb{E}(X)^2$$

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

$$(ii) \text{Cov}(X, Y) = \mathbb{E}((X-\mu)(Y-\tau)) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y).$$

~ Define correlation:

$$\rho(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}}$$

$$\in [-1, 1]$$

(*) 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: $\text{Cov}(X, Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y) = \mathbb{E}(X)\mathbb{E}(Y) - \mathbb{E}(X)\mathbb{E}(Y) = 0$.

(ii) Let X be a \mathbb{Z} -valued r.v. that is symmetrically distributed: $P(X=k) = P(X=-k)$.

Let $Y = X^2$. [Dependent, unless $Y=0$].

$$\rightarrow \text{Cov}(X, Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y) = \mathbb{E}(X^3) - \mathbb{E}(X)\mathbb{E}(X^2)$$

$$\rightarrow [X \text{ symm. dist} \rightarrow kP_X(k) = -(-k)P_X(-k)] = 0 - 0 \cdot \mathbb{E}(X^2) = 0.$$

(*) Ex (Random Walk). Let X_1, X_2, \dots i.i.d. Bernoulli($\frac{1}{2}$), $[X_i \in \{\pm 1\}]$, and let $S_n = X_1 + \dots + X_n$.

Want $\text{Cov}(S_n, S_m)$ for $n \leq m$.

~ Lemma. Covariance is bilinear & symmetric: $\text{Cov}(a(X+Y), bZ) = ab(\text{Cov}(X, Z) + \text{Cov}(Y, Z))$.

^t similar to inner product

$$\Rightarrow \text{Cov}(S_n, S_m) = \text{Cov}\left(\sum_{i=1}^n X_i, \sum_{j=1}^m X_j\right) = \sum_{i=1}^n \sum_{j=1}^m \text{Cov}(X_i, X_j)$$

$$\text{Since } X_i \text{'s indep, } \text{Cov}(X_i, X_j) = \delta_{ij} \text{Var}(X_i) \sim \sum_{i=1}^n \sum_{j=1}^m \text{Cov}(X_i, X_j) = \sum_{i=1}^n \text{Var}(X_i) = n. [\text{Indep. of } m, m-n]$$

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

(*) Proof: $\text{Var}(X+Y) = \text{Cov}(X+Y, X+Y) = \text{Cov}(X, X) + \text{Cov}(Y, Y) + 2\text{Cov}(X, Y) = \text{Var}(X) + \text{Var}(Y)$. \square

$\underbrace{t=0}_{\text{by assump.}}$

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 : $F_W(u) = P(W \leq u) = P(\ln(W) \leq \ln(u))$ [$W > 0$]

$$\rightarrow (\text{Letting } X = \ln(W)) = F_X(\ln(u)) = \int_{-\infty}^{\ln(u)} f_X(x) dx$$

Differentiating: $f_W(u) = F'_W(u) = \frac{1}{u} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(\ln(u)-\mu)^2}{2\sigma^2}}$ ($\frac{1}{u}$ called the Jacobian factor).

Def: We say that X is Cauchy-Lorentz distributed if it has PDF

$$f_X(x) = \frac{1}{\pi(1+x^2)} \quad [\approx \frac{1}{\pi x} \text{ as } x \rightarrow \infty]$$

Def: Exponential Distribution

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

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0 \\ 0 & x < 0 \end{cases} \quad [\text{Can be defined equiv. w/ } x > 0, x \neq 0]$$

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

Def: Gamma Function

For $\operatorname{Re}(z) > 0$ [real part of z], define the gamma function

$$\Gamma(z) = \int_0^\infty x^{z-1} e^{-x} dx$$

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

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

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

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 $\sum_{i=1}^n X_i^2 \sim \chi^2_n = \text{Gamma}\left(\frac{n}{2}, \frac{1}{2}\right)$. [Chi-squared]

Recall (Expectation): For disc. r.v.'s, defined $E(X) = \sum k p_X(k)$ 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 $f(x) \sim 1$ 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:

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

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

(*) Ex: Assuming the velocity of air particles given by $X \sim N(0, \sigma^2 \text{ some T-dep. parameter}^*)$

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

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

2/26/24

Sequences of Events (HW4)

Given events E_k for $k \in \mathbb{N}$, define:

- (i) $\{\text{infinitely many } E_k \text{ occur}\} = \limsup_{k \rightarrow \infty} E_k = \bigcap_{n=1}^{\infty} \left[\bigcup_{k=n}^{\infty} E_k \right] = \limsup_{k \rightarrow \infty} \mathbb{1}_{E_k}(w)$
- (ii) $\{\text{finitely many } E_k \text{ occur}\} = \liminf_{k \rightarrow \infty} E_k = \bigcup_{n=1}^{\infty} \left[\bigcap_{k=n}^{\infty} E_k \right] = \liminf_{k \rightarrow \infty} \mathbb{1}_{E_k}(w)$

HW 4-6

Expected Value (HW5)

If $X: \Omega \rightarrow \mathbb{Z}_+$ 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's 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 $Y_r = \text{position of } r^{\text{th}} \text{ success}$,

$$Y_r \text{ has PMF: } \mathbb{P}(Y_r = k) = \begin{cases} \binom{k-1}{r-1} p^r (1-p)^{k-r} & k \in \mathbb{N}, k \geq r \\ 0 & \text{otherwise} \end{cases}$$

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

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

Expectation for Continuous Random Variables (cont.)

2/26/24

Lecture 19

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

(i) Measurability: If f_X is measurable, then $x f_X$ is measurable

(ii) Finiteness: The integral is absolutely convergent

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

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

... 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. $\{w : g \circ X(w) \leq \lambda\} \in \mathcal{F}$

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

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

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

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

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: $\bigcup_{q \in Q \cap O} (q - \frac{1}{2} \text{dist}(q, O^c), q + \frac{1}{2} \text{dist}(q, O^c))$ countable union

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

(*) Generalized Quantile Functions

2/28/24

Lecture 20

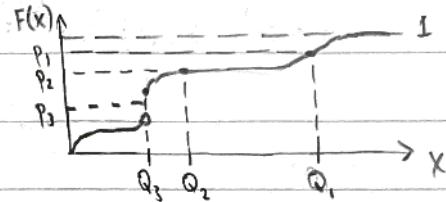
Def: Generalized Quantile Functions

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

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

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

(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 continuous from the right,

then if we take a decreasing sequence $(x_n)_n$ converging to $a(p)$ with $F(x_n) \geq p$,

then per continuity, $F(\lim_{n \rightarrow \infty} x_n) = F(a) = \lim_{n \rightarrow \infty} F(x_n) \geq p$. Then $\{x : F(x) \geq p\} = [a(p), \infty)$.

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

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

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

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

→ Hence: $Q(p) \leq x \Leftrightarrow F(x) \geq p$.

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

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

(iii) If $p \leq q$, then $\{x : F(x) \leq p\} \subseteq \{x : F(x) \geq q\}$.

$\Rightarrow \inf\{x : F(x) \leq p\} \leq \inf\{x : F(x) \geq q\} \Rightarrow Q(p) \leq Q(q)$.

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

Joint CDF & Independence

2/28/24

Lecture 20 Prop. (Simulation of R.V.s) Let F be a CDF with associated quantile function Q . Then given (cont.) $U \sim \text{Uniform}(0, 1)$, the r.v. $Q(U)$ has $F_X(x) = F(x)$.

$$(*) \text{ Proof: } F_X(x) = P(X \leq x) = P(\{\omega : Q(U(\omega)) \leq x\}) = P(\{\omega : F(x) \geq U(\omega)\}) \\ \rightarrow P(U(\omega) \leq F(x)) = \int_{-\infty}^{F(x)} f_U(u) du = F(x) \text{ for } F(x) \in [0, 1]. \quad \square$$

Def: Joint CDF

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

$$\boxed{F_{X,Y}(x, y) = P(X \leq x \wedge Y \leq y)}$$

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

Def: Independence

We say two R.V.s X, Y are independent if $P(X \leq x \wedge Y \leq y) = P(X \leq x)P(Y \leq y)$ 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 [$P(X \leq x \wedge Y \leq y) = P(X \leq x)P(Y \leq y)$] iff they are independent in the "old" sense [$P(X=x \wedge Y=y) = P(X=x)P(Y=y)$].

$$(*) \text{ Proof: } (\Leftarrow) P(X \leq x \wedge Y \leq y) = \sum_{n \leq x} \sum_{m \leq y} P(X=n \wedge Y=m) = \sum_{n \leq x} P(X=n) \sum_{m \leq y} P(Y=m) \\ \rightarrow \left[\sum_{n \leq x} P(X=n) \right] \left[\sum_{m \leq y} P(Y=m) \right] = P(X \leq x)P(Y \leq y).$$

$$(\Rightarrow) P(X=x \wedge Y=y) = P(X \leq x \wedge Y \leq y) - P(X \leq x \wedge Y \leq y-1) - P(X \leq x-1 \wedge Y \leq y) \\ + P(X \leq x-1 \wedge Y \leq y-1) \\ \rightarrow P(X \leq x)P(Y \leq y) - P(X \leq x)P(Y \leq y-1) - P(X \leq x-1)P(Y \leq y) + P(X \leq x-1)P(Y \leq y-1) \\ = [P(X \leq x) - P(X \leq x-1)][P(Y \leq y) - P(Y \leq y-1)] = P(X=x)P(Y=y). \quad \square$$

(*) Continuous Random Variables

2/29/24

PDFs & CDFs: $p_X(x)dx = P(X \leq x \leq x+dx)$ (*) p_X, P_X equiv. to f_X, F_X

$$\sim \int_a^b p_X(x)dx = P(a \leq X \leq b) [= P_X(b) - P_X(a)]$$

Disc 8

Relationships between CDFs

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

$$\int_a^b p_Y(y)dy = P(a \leq Y \leq b) = P(a \leq g(X) \leq b)$$

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

$$\rightarrow P(g^{-1}(a) \leq X \leq g^{-1}(b)) = \int_{g^{-1}(a)}^{g^{-1}(b)} p_X(x)dx \quad [P_Y(b) - P_Y(a) = P_X(g^{-1}(b)) - P_X(g^{-1}(a))]$$

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

$$\sim \int_a^b p_Y(y)dy = \int_{g^{-1}(a)}^{g^{-1}(b)} p_X(g^{-1}(y)) \cdot (g^{-1})'(y)dy$$

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

$$\sim \int_0^b p_Y(y)dy = P(0 \leq Y \leq b) = P(0 \leq X \leq b) + P(-b \leq X \leq 0) = 2P(0 \leq X \leq b) = \int_0^b 2p_X(x)dx$$

$$\Rightarrow \int_0^b p_Y(y)dy = \int_0^b 2p_X(x)dx \Rightarrow p_Y = 2p_X$$

The Cauchy Distribution

Recall: A cts r.v. X is Cauchy-distributed if it has PDF

$$p_X(x) = \frac{1}{\pi(1+x^2)}$$

(*) It's a normalizing constant: $\int_{-\infty}^{\infty} \frac{1}{\pi(1+x^2)} dx = \frac{1}{\pi} \arctan(x) \Big|_{-\infty}^{\infty} = \frac{1}{\pi}$

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

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

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

(*) Can find X with $E(X)$ but $\text{Var}(X)$, $E(X^2)$ but not $E(X^{n+1})$, etc., similarly

Joint Continuity

3/1/24

Lecture 21

(*) Simulating Random Variables

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

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 $i \in \mathbb{N}$ (infinite coin tosses)

$$\leadsto \text{set } X = \sum_{i=1}^{\infty} \frac{1}{2^i} X_i \quad [\text{Exp. independent binary digits}] \sim \text{Uniform}(0,1) \quad \begin{aligned} X &= \sum_{i=1}^{\infty} \frac{1}{2^i} X_i \\ Y &= \sum_{i=1}^{\infty} \frac{1}{2^i} X_{2i+1} \end{aligned}$$

(*) Coupling

Def: Given two r.v.'s X and Y in \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 $X \rightarrow m$ if f exists

Def: Joint Continuity

We say that two r.v.'s X, Y are jointly continuous if there is an integrable function $f_{X,Y}: \mathbb{R}^2 \rightarrow [0,1]$ (called the joint PDF) such that:

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

Joint Continuity (cont.)

3/1/24

Lecture 21

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 . (cont.)

$$\text{(*) Proof: } F_X(x) = P(X \leq x) = \lim_{y \rightarrow \infty} P(X \leq x \text{ & } Y \leq y) = \lim_{y \rightarrow \infty} F_{X,Y}(x, y) \quad \begin{matrix} \curvearrowleft \\ \text{(*) Taking a limit since } \\ F_Y(\infty) \text{ undefined} \end{matrix}$$

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

$$\therefore f_X(x) = \frac{d}{dx} F_X(x) = \frac{d}{dx} \int_{-\infty}^x \left(\int_{-\infty}^y f_{X,Y}(x', y') dy' \right) dx' = \int_{-\infty}^x f_{X,Y}(x, y) dy. \quad \blacksquare$$

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

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

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

(**) On convergence: as long as the individual \int_x^∞ , \int_y^∞ 's are conditionally convergent, then the $\int y^\infty$ term is conditionally convergent provided that it is over "borders" 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: $\boxed{\mathbb{E}(g(x,y)) = \iint_{\mathbb{R}^2} g(x,y) f_{X,Y}(x,y) dx dy}$ provided that the integral is absolutely convergent.

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

$$\mathbb{P}((x,y) \in B) = \mathbb{E}(I_B(x,y)) = \iint_B I_B(x,y) f_{X,Y}(x,y) dx dy.$$

Functions of Random Variables

3/1/24

Lecture 21 Prop. If X, Y are continuous & independent, then $Z = X + Y$ is continuous and has PDF given by $f_Z(z) = \int_{-\infty}^{\infty} f_X(x) f_Y(z-x) dx$.

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

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

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

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

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

(*) Note: $\int_{-\infty}^{\infty} f_X(x) f_Y(z-x) dx$ is a convolution of X and Y : $\underbrace{\dots}_{X \text{ points [stimulus]}} \dots \Rightarrow \text{XXXXXX}$ [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 $Z = X+Y$.

$$\text{Solution: Per Prop., } f_Z(z) = \int_{-\infty}^{\infty} f_X(z-y) f_Y(y) dy = \int_{-\infty}^{\infty} \frac{1}{\Gamma(2)} (z-y)^{-2} e^{-(z-y)} \cdot \frac{1}{\Gamma(2)} y e^y dy \left[1_{[0, \infty)}(y) 1_{[0, \infty)}(z-y) \right] \\ \rightarrow \int_0^z \frac{1}{\Gamma(2)} (z-y)^{-2} y e^y dy = \frac{z^3 e^z}{\Gamma(2)} \int_0^z \frac{1}{2} (1-y/2)(y/2) dy \stackrel{u=y/2}{=} \frac{1}{\Gamma(2)} z^3 e^{-z} \int_0^1 (1-u) u du$$

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

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

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

\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. i want PDF of $Z = X^2 + Y^2$.

Soln: Let $B(z) = \{(x, y) : x^2 + y^2 \leq z\}$ [ball w/ radius \sqrt{z}] $\rightarrow \underbrace{\Pr(X^2 + Y^2 \leq z)}_{F_Z(z)} = \Pr((X, Y) \in B(z))$ [$z \geq 0$].

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

$$\xrightarrow{u=\frac{r^2}{2}} = \int_0^{z/2} e^{-u} du; f_z(z) = \frac{1}{\delta z} F_z(z) = \frac{1}{\delta z} \int_0^{z/2} e^{-u} du \Rightarrow f_z(z) = \frac{1}{2} e^{-z/2} \text{ for } z \geq 0$$

More generally, may want joint distribution of $Z = X^2 + Y^2$ and 4-quadrant arctan(X, Y) = Θ .

[4-quadrant arctan(X, Y) more stable for X near 0 than arctan($X/2$); value values $\in [0, 2\pi]$]

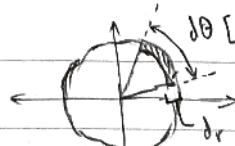
$$\rightarrow \text{As a diagram: } F_{Z,\Theta}(z, \theta) = \Pr((X, Y) \in \text{shaded region}) = \int_0^\theta \int_0^{\sqrt{z}} \frac{1}{2\pi} r dr d\theta$$

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

(*) 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]

$$(*) \text{ Tip: } F_{X,Y}(x, y) = \mathbb{E}(1_{\{X \leq x, Y \leq y\}}) = \iint_{-\infty}^{\infty} \iint_{-\infty}^{\infty} 1_{\{X \leq x, Y \leq y\}} f_{X,Y}(x', y') dx' dy'$$

~ 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 $P(A) > 0$, can find conditional CDF:

$$P(X \leq x | A) = \frac{P(X \leq x \cap A)}{P(A)}$$

(*) Ex: Let $X \sim \text{Exp}(\lambda)$ [$F_X(x) = \int_{[0, \infty)} \lambda e^{-\lambda x} dx$] $\rightarrow E(X) = \frac{1}{\lambda}$ [from]

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

(*) For simplicity,

define $[a, \infty]$ as
empty interval if
 $x \leq a$

$$\text{Sln: } P(X \leq x | X \geq a) = \frac{P(a \leq X \leq x)}{P(a \leq X)} = \frac{\int_a^x \int_{[a, \infty)} f_X(x) f_X(x') dx' dx}{\int_a^\infty \int_{[a, \infty)} f_X(x) f_X(x') dx' dx} = \frac{\int_a^x f_X(x') dx'}{\int_a^\infty f_X(x') dx'}$$

$$\rightarrow = \frac{e^{-\lambda a} - e^{-\lambda x}}{e^{-\lambda x}} = 1 - e^{-\lambda(x-a)} \Rightarrow F_X(x | X \geq a) = \begin{cases} 1 - e^{-\lambda(x-a)} & x \geq a \\ 0 & x < a \end{cases}$$

some as P law for $\text{Exp}(\lambda)$,
just "shifted" right by a

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

$$\text{Sln: } F_Y(y | a \leq X \leq b) = P(Y \leq y | a \leq X \leq b) = \frac{P(Y \leq y \cap a \leq X \leq b)}{P(a \leq X \leq b)} = \frac{\int_0^b \int_{[a, b]} f_{X,Y}(x, y) dx dy}{\int_a^b f_X(x) dx}$$

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

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

$$\text{Sln: } \text{Want } P(\mathbb{1}_{A \subseteq \mathbb{R}^2}(U, V) = 1 | U^2 + V^2 \leq 1) = \frac{P((U, V) \in A \cap B)}{P((U, V) \in B)} = \frac{\iint_A \mathbb{1}_B(u, v) du dv}{\iint_B \mathbb{1}_B(u, v) du dv}$$

$\underbrace{\quad \quad \quad}_{\text{area of the unit ball}}$

$$\rightarrow = \frac{\iint_A \mathbb{1}_A(u, v) du dv}{\pi \cdot 1^2}. \text{ For } \{U = u \& V = v\}, P(U = u \& V = v | (U, V) \in B) = \frac{1}{\pi} \iint_B \mathbb{1}_B(u, v) du dv.$$

\Rightarrow joint PDF given by $f_{U,V}(u, v | U^2 + V^2 \leq 1) = \frac{1}{\pi} \mathbb{1}_B(u, v)$. $\sim (*)$ Say that (U, V) is uniformly distributed by area (unit radius) on the ball

(*) Functions of Random Variables

3/7/24

Disc. 9

(*) Fundamental Theorem of Calculus

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

3 terms correspond to 3 locations where it appears in orig. integral

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

$$\text{Sln: } F_Z(z) = P(Z \leq z) = P(XY \leq z) = P(XY \leq z \cap X \geq 0) + P(XY \leq z \cap X \leq 0)$$

$$\sim = P(Y \leq z/X \cap X \geq 0) + P(Y \geq z/X \cap X \leq 0)$$

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

Events overlap at
 $X=0$, but $P(X=0)=0$
not an issue

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

$$\sim = \int_0^{\infty} f_{X,Y}(x, z/x) \frac{1}{x} dx + \int_{-\infty}^0 f_{X,Y}(x, z/x) \frac{1}{x} (-z/x) dx$$

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

(*) Notes: (i) It's okay to have R.V.s undefined on events with $P(E) = 0$

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

(*) Integral Examples

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

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

$$\rightarrow \text{Define } u = 4x^3, \quad du = 12x^2 \Rightarrow I = \int_0^{\infty} (u/4)^{7/3} e^{-u} \frac{du}{12} = \frac{1}{12} \frac{1}{4} \int_0^{\infty} u^{7/3} e^{-u} du = \frac{1}{12} \frac{1}{4} \Gamma(8/3)$$

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

$$\rightarrow \text{Define } u = z^2 - x^3 \Rightarrow J(z) = \int_0^{\sqrt{z}} u^5 B(6, 3) du$$

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

$$\rightarrow \text{Define } u = \ln(Y_x) \Rightarrow K = \frac{1}{4} \int_0^1 u^3 T(4) du$$

Conditional Density Functions (cont.)

3/8/24

Lecture 24 Conditional Density Functions

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

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

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

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

$$\text{From prev. lecture: saw } f_{Y|X}(y|a \leq X \leq b) = \frac{\int_a^b f_{X,Y}(x,y) dx}{\int_a^b f_X(x) dx}$$

~ 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 $[x - 1/n, x + 1/n]$, e.g.) to get a limit:

$$f_{Y|X}(y|x) = \frac{\int_{x-1/n}^{x+1/n} f_{X,Y}(x,y) dx}{\int_{x-1/n}^{x+1/n} f_X(x) dx} \quad [\text{limit of } f_{Y|X}(y|a \leq X \leq b) \text{ as } a, b \rightarrow x]$$

=> Lebesgue: Above limit exists for all but an exceptional set E that is negligible w.r.t. X (i.e. $\int_E f_X(x) dx = 0$; $P(X \in E) = 0$)

r defined for all $x \notin E$

Similar to discrete case, can define conditional expectation:

$$E(Y|X=x) = \int y f_{Y|X}(y|x) dy$$

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

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

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

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

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

$$(b) \text{LHS is } \int_x \left[\int_y y f_{Y|X}(y|x) dy \right] f_X(x) dx = \int_x \int_y y f_{X,Y}(x,y) dy dx = E(Y).$$

Conditional Density Functions (cont.)

3/8/24

Lecture 24

+ Lecture 25

(*) Proof (cont.)

(i) Skipped; can be feasibly proven if g "nice", more difficult otherwise

[Ex: g constant $\Rightarrow g(Y)$ may be continuous, discrete, or neither]

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

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

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

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

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

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

(iii) $f_{Y|X}(y|x) = \frac{\int_0^y e^{-x} dx}{\int_0^{\infty} e^{-x} dx} = \mathbb{1}_{x \leq y} e^{-(y-x)} \text{ for } x \geq 0$

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

Lemma. Jointly continuous r.v.s X, Y are independent iff $f_{Y|X}(y|x) = f_Y(y)$ 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

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

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

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

Multivariate Normal Distribution

3/11/24

- 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 [$\int_E f_1(x) dx = 0$].
 - (ii) In cases where a PDF/conditional PDF may or may not be defined, can use well-defined expression $P(\dots | X=x)$ 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 $O = [\vec{v}_1 \dots \vec{v}_n]$ 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: $\det(2\pi\Sigma) = (2\pi)^n \Sigma$ ($n = \# \text{ columns}$)

Theorem: (a) $\int f_{\vec{x}}(\vec{x}) d\vec{x} = \int \dots \int f_{\vec{x}}(\vec{x}) dx_1 \dots dx_n = 1$

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

(c) $\underbrace{\int (\vec{x}_i - \vec{\mu}_i)(\vec{x}_j - \vec{\mu}_j) f_{\vec{x}}(\vec{x}) d\vec{x}}_{\text{Cov}(X_i, X_j)} = \Sigma_{ij}$

Multivariate Normal Distribution, (cont.)

3/11/24

(*) 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 $\vec{\mu} = \vec{0}$.

Lecture 25

+ Lecture 26

Con factor $\Sigma = BB^T$ for some invertible B (proof later)

~ can make change of variables $\vec{x} = B\vec{y}$ with $\frac{\partial x_i}{\partial y_j} = B_{ij} \Rightarrow \text{Jacobian} = \det(B)$

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

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

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

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

$$\text{Recall: } \int e^{-y^2/2} dy = \sqrt{2\pi} \Rightarrow (2\pi)^{-n/2} \int \dots \int e^{-y_1^2/2} \dots e^{-y_n^2/2} dy_1 \dots dy_n = (2\pi)^{-n/2} (2\pi)^{n/2} = 1. \quad \square$$

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

$$[\mathbb{E}(\vec{x})]_j = \left[\int \vec{x} f_{\vec{x}}(\vec{x}) d\vec{x} \right]_j = (2\pi)^{-n/2} \int [B\vec{y}]_j \exp\left\{-\frac{1}{2} |\vec{y}|^2\right\} d\vec{y} \quad (*) \quad \begin{array}{l} B\vec{y} = (\vec{x} - \vec{\mu}); \\ \text{if } \vec{\mu} \neq \vec{0}, \text{ factors out} \end{array}$$

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

$$\text{Recall: } \int y_i e^{-y^2/2} dy = 0 \Rightarrow \forall j: (2\pi)^{-n/2} \int \dots \int e^{-y_1^2/2} \dots y_j e^{-y_1^2/2} \dots e^{-y_n^2/2} dy_1 \dots dy_n = 0 \\ \Rightarrow [\mathbb{E}(\vec{x})]_j = 0 \quad \forall j \Rightarrow \mathbb{E}(\vec{x}) = \vec{0}. \quad \square$$

L if $\vec{\mu} \neq \vec{0}$, $\mathbb{E}(\vec{x}) = \vec{\mu}$

Multivariate Normal Distribution (cont.)

3/13/24

Lecture 26 (* Proof (cont.):

(cont)

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

$$\begin{aligned}\text{Cov}(X_i, X_j) &= \mathbb{E}[(\vec{x}_i - \vec{\mu}_i)(\vec{x}_j - \vec{\mu}_j)] = \int (\vec{x}_i - \vec{\mu}_i)(\vec{x}_j - \vec{\mu}_j) f_{\vec{x}}(\vec{x}) d\vec{x} \\ &= (2\pi)^{-n/2} \int \dots \int (B\vec{y})_i (B\vec{y})_j \exp\left\{-\frac{1}{2}|\vec{y}|^2\right\} d\vec{y} \\ &= (2\pi)^{-n/2} \int \dots \int \left(\sum_k B_{ik} y_k \right) \left(\sum_m B_{jm} y_m \right) \exp\left\{-\frac{1}{2}(y_1^2 + \dots + y_n^2)\right\} dy_1 \dots dy_n \\ &= (2\pi)^{-n/2} \sum_{i,j} \left[B_{i1} B_{j1} \int \dots \int y_1 y_n e^{-\frac{1}{2}(y_1^2 + \dots + y_n^2)} dy_1 \dots dy_n \right]\end{aligned}$$

→ Two cases: (i) $i \neq j$: Integral = 0 ($\int y_k e^{-\frac{1}{2}y^2} dy_k = 0$)

(ii) $i = j$: $\int y_k^2 e^{-\frac{1}{2}y^2} dy_k = \sqrt{2\pi}$ [Variance of $N(0, 1)$]

$$\Rightarrow \text{Cov}(X_i, X_j) = \sum_{l,m} B_{il} B_{jm} S_{lm} = \sum_{l,m} B_{il} I_{ln} B_{mj}^T = (BB^T)_{ij} = (BB^T)_{ij} = \sum_{ij} \blacksquare$$

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:

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

(*) The Cholesky Factorization

3/13/24

Lecture 26
(cont.)

(*) Proof (cont.)

$$\text{Then we can factor } \Sigma \text{ as: } \Sigma = \underbrace{\begin{bmatrix} 0 & \\ & \ddots & 0^T \end{bmatrix}}_B \underbrace{\begin{bmatrix} 0 & \\ & \ddots & 0^T \end{bmatrix}}_{0^T}^T$$

$$(*) \text{ Alt: } \Sigma = \left(\begin{bmatrix} 0 & \\ & \ddots & 0^T \end{bmatrix} \right) \left(\begin{bmatrix} 0 & \\ & \ddots & 0^T \end{bmatrix} \right)^T \quad \text{(*) Recall: } (AB)^T = B^T A^T$$

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.]

$$\text{Then } C = BB^T = (R^T Q^T)(QR) = R^T R.$$

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

Multivariate Parameter Estimation

3/13/24

Lecture 26 (*) Estimation for Multivariate Normal Distributions

Lecture 27 Setup: Imagine we have a machine that produces errors $X \sim N(0, \sigma^2)$, 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 : $f_{X,Y}(x,y) = f_{Y|X}(y|x)f_X(x) = \frac{1}{2\sqrt{\sigma^2\alpha^2}} \exp\left\{-\frac{x^2}{2\sigma^2} - \frac{(x-y)^2}{2\sigma^2}\right\}$

→ Interpreting as joint normal distribution: $-\frac{x^2}{2\sigma^2} - \frac{(x-y)^2}{2\sigma^2} = -\frac{1}{2}[x] \cdot \Sigma^{-1}[x]$

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

Maximum Likelihood Estimation

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

Soln.: $f_{Y|X} = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y-x)^2}{2\sigma^2}\right) \rightsquigarrow \hat{x}_m(y) = y$. [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 $f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)}$

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

(ii) $E(XY) = \bar{\mu} = 0 \Rightarrow E(X), E(Y) = 0$

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

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

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

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

$$\hat{x}_{MAP} = \underset{x}{\operatorname{argmax}} f_{X|Y}(x|y) = \frac{\sigma^2}{\sigma^2 + \sigma^2} y \quad (*) \text{ Is a biased estimator: underestimates } y$$

(*) Observation: Is a "smarter" estimator than MLE: balances expected % of error

from X vs. from Z based on relative magnitudes of σ^2, σ^2

Multivariate Parameter Estimation (cont.)

3/15/24

Lecture 27 (♦) Estimation for Multivariate Normal Distributions (cont.)

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

(ii) Minimum Mean Squared Error

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

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

(♦) Def: Mean Squared Error

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

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

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

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

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

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

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

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

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

3/15/24

HW 7-10

Independent Geometric Variables (HW 7)

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

Indicator Random Variables (HW 7)

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

CDF Properties (HW 7)

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

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

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

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

Cosine (HW 8)

Given $\theta \sim \text{Unif}([0, 2\pi])$, $Y = \cos(\theta)$ has CDF $F_Y(y) = 1 - \frac{\arccos(y)}{\pi} \Rightarrow \text{PDF } f_Y(y) = \frac{1}{\pi \sqrt{1-y^2}}$

Functions of Multiple R.V.s (HW 8)

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

The Gamma Distribution (HW 8)

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

If $Y \sim N(0, \sigma^2)$, then $\boxed{Y^2 \sim T(\frac{1}{2}, \frac{1}{2\sigma^2})}$.

(*) 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 χ^2_v iff it has PDF

$$f_X(x) = \begin{cases} \frac{1}{2^{v/2} \Gamma(v/2)} x^{\frac{v}{2}-1} e^{-x/2} & x \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

Given $X \sim \chi^2_v$, $Y \sim \chi^2_k$ independent, $X+Y \sim \chi^2_{v+k}$.

(*) Euler's Beta Integral:

$$\int_0^1 u^{\frac{v}{2}-1} (1-u)^{\frac{k}{2}-1} du = \frac{\Gamma(\frac{v}{2}) \Gamma(\frac{k}{2})}{\Gamma(\frac{v+k}{2})}$$

The Student t-Distribution (HW9)

Given $X \sim \chi^2_v$ and $Z \sim N(0, 1)$ indep., we say that $T = Z \sqrt{\frac{v}{X}}$ follows a

Student t-distribution with v degrees of freedom.

$$\text{PDF: } f_T(x) = \frac{\Gamma(\frac{v+1}{2})}{\sqrt{\pi v} \Gamma(\frac{v}{2})} \left(1 + \frac{x^2}{v}\right)^{-\frac{v+1}{2}} \quad [x \in \mathbb{R}]$$

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

3/15/24

HW7-10

(cont.)

(*) Another Euler integral:

$$\int_{-\infty}^{\infty} \left(1 + \frac{1}{\nu} t\right)^{-\frac{\nu+1}{2}} dt = \frac{\sqrt{\nu}\pi^{\frac{1}{2}} \Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu+1}{2})}$$

Rejection Sampling (HW10)

Given continuous f 'n and PDF f_X with:

- (i) $0 \leq f(x) \leq M \quad \forall x \in \mathbb{R}$

$$(ii) f(x) = 0 \quad \text{if } |x| \geq M$$

Given $U_n \sim \text{Unif}(-M, M)$ and $V_n \sim \text{Unif}(0, M)$ independent, can sample from X by taking U_n, V_n until $f(U_n) \leq V_n$
 \Rightarrow returned value U_n has PDF $f_{U_n}(u) = f_X(u)$.

Sums of Exponential R.V.s (HW10)

Given $X_1, \dots, X_n \sim \text{Exp}(\lambda)$, $Y_i = \sum_{j=1}^i X_j$ for $1 \leq i \leq n$, the variables Y_1, \dots, Y_n have joint PDF given

by:

$$f_{Y_1, \dots, Y_n}(y_1, \dots, y_n) = \lambda^n e^{-\lambda y_n} \prod_{0 < y_1 < \dots < y_n}$$

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

$$X_1 + X_2 \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$$

(*) 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: $\left[\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2}\right] y^2 - 2\left[\frac{z-\mu_1}{\sigma_1^2} + \frac{\mu_2}{\sigma_2^2}\right] y + \left[\frac{(z-\mu_1)^2}{\sigma_1^2} + \frac{(\mu_2)^2}{\sigma_2^2}\right]$

\rightarrow Can complete the square: $ay^2 + by + c = a(y + \frac{b}{2a})^2 + c - \frac{b^2}{4a}$

\rightarrow Eventually obtain $f_Z(z) = \frac{1}{\sqrt{2\pi(\sigma_1^2 + \sigma_2^2)}} \exp\left\{-\frac{(z-\mu_1-\mu_2)^2}{2(\sigma_1^2 + \sigma_2^2)}\right\}$.