Foundations/Stats

STATS 743A

Cameron Roopnarine*

11th February 2023

LECTURE 1
7th September

- Textbook: **Statistical Inference** by *George Casella + Roger L. Berger*
- Office hours: Monday 1:30–2:30 in HH210.

Set Theory

DEFINITION 1: Containment

$$A \subset B \iff x \in A \implies x \in B.$$

DEFINITION 2: Equality

$$A = B \iff A \subset B \text{ and } B \subset A.$$

DEFINITION 3: Union

The **union** of A and B, written $A \cup B$, is the set of elements that belong to either A or B or both:

$$A \cup B = \{x : x \in A \text{ or } x \in B\}.$$

For example, if $A = \{0, 2, 4, 6, 8\}$ and $B = \{0, 3, 6, 9\}$, then

$$A \cup B = \{0, 2, 3, 4, 6, 8, 9\}.$$

DEFINITION 4: Intersection

The **intersection** of A and B, written $A \cap B$, is the set of elements that belong to both A and B:

$$A \cap B = \{x : x \in A \text{ and } x \in B\}.$$

DEFINITION 5: Complementation

The **complement** of A, written A^c, is the set of all elements that are not in A:

$$A^c = \{x : x \notin A\}.$$

*LaTeXer

DEFINITION 6: Relative Complement

The **relative complement** of A in B, written $B \setminus A$, is the set of all elements that are in B and not in A:

$$B \setminus A = \{x : x \in B \text{ and } x \notin A\} = B \cap A^c.$$

THEOREM 1: De Morgan's Laws

For any events A and B defined on a sample space S,

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

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

Proof:

(i) Let $x \in (A \cup B)^c$. We know that $x \notin (A \cup B)$, so $x \notin A$ and $x \notin B$. Hence, $x \in A^c$ and $x \in B^c$, which means $x \in (A^c \cap B^c)$. Therefore, $(A \cup B)^c \subset (A^c \cap B^c)$.

Let $y \in (A^c \cap B^c)$. We know that $y \in A^c$ and $y \in B^c$, so $y \notin A$ and $y \notin B$. Hence, $y \notin (A \cup B)$, which means $y \in (A \cup B)^c$. Therefore, $(A^c \cap B^c) \subset (A \cap B)^c$.

(ii) Let $x \in (A \cap B)^c$. We know that $x \notin (A \cap B)$, so $x \notin A$ or $x \notin B$. Hence, $x \in A^c$ or $x \in B^c$, which means $x \in (A^c \cup B^c)$. Therefore, $(A \cap B)^c \subset (A^c \cup B^c)$.

Let $y \in (A^c \cup B^c)$. We know that $y \in A^c$ or $y \in B^c$, so $y \notin A$ or $y \notin B$. Hence, $y \notin (A \cap B)$, which means $y \in (A \cap B)^c$. Therefore, $(A^c \cup B^c) \subset (A \cap B)^c$.

THEOREM 2: Distributive Laws

(i) $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$.

(ii) $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$.

DEFINITION 7: Injective and Surjective

Let A and B be sets and let $f\colon A \to B$.

- We say f is **injective** (or **one-to-one**, written as $1\colon 1$) when for all $x, y \in A$, if $f(x) = f(y)$, then $x = y$.
- We say f is **surjective** (or **onto**) when for every $y \in B$, there exists at least one $x \in A$ such that $f(x) = y$.

DEFINITION 8: Countability

A set S is **countable** if there exists an injective function $f\colon S \to \mathbf{N}$.

EXAMPLE 1

The set $\mathbf{Z}$ of all integers is countable. First, match 0 with 1. Then, for $n > 0$, match n with $2n$ and match

$-n$ with $2n+1$.

1	0
2	1
3	-1
4	2
5	-2
6	3
7	-3

THEOREM 3

The unit interval $[0, 1]$ *is not countable.*

Proof (Cantor's diagonalization argument): Assume for a contradiction that there is some bijection $f\colon \mathbf{N} \to [0, 1]$.

1	$f(1) = 0.5000\cdots$
2	$f(2) = 0.14152\cdots$
3	$f(3) = 0.33333\cdots$
4	$f(4) = 0.110100100010000\cdots$
5	$f(5) = 0.12345\cdots$

Denote

$$\begin{aligned} f(1) &= 0.a_{11}a_{12}a_{13}a_{14}\cdots \\ f(2) &= 0.a_{21}a_{22}a_{23}a_{24}\cdots \\ &\vdots \\ f(n) &= 0.a_{n1}a_{n2}a_{n3}a_{n4}\cdots \end{aligned}$$

For example, $a_{24} = 5$. Let

$$\begin{aligned} b_1 &= 9 - a_{11}\cdots \\ b_2 &= 9 - a_{22}\cdots \\ b_3 &= 9 - a_{33}\cdots \\ &\vdots \\ b_n &= 9 - a_{nn}\cdots \end{aligned}$$

Then, $0.b_1b_2b_3\cdots$ does not appear anywhere in my list, since for every $n \geq 1$, the n^{th} digit of this number is different from the n^{th} digit of the n^{th} number on my list. This contradicts my assumption that f is a bijection.

DEFINITION 9

A **probability space** is an ordered triple $(\Omega, \mathcal{F}, \mathbb{P})$ where

- Ω is a non-empty set, called the *sample space* (where elements $\omega \in \Omega$ are called "events"),
- $\mathcal{F}$ is a collection of subsets of Ω, called the *σ-algebra* (where elements $A \in \mathcal{F}$ are called "events") with the following properties:

  S1 $\Omega \in \mathcal{F}$,

  S2 $\forall A \in \mathcal{F}$, $(\Omega \setminus A) = A^c \in \mathcal{F}$ (closed under complements),

  S3 For any sequence $A_1, A_2, A_3, \ldots \in \mathcal{F}$, we get $\bigcup_i A_i \in \mathcal{F}$ (closed under countable unions),

- $\mathbb{P}\colon \mathcal{F} \to [0,1]$ with

  P1 $\mathbb{P}(\Omega) = 1$,

  P2 $\mathbb{P}(A) \geq 0$ for all A, and

  P3 if $A_1, A_2, \ldots,$ are disjoint elements of $\mathcal{F}$, then

$$\mathbb{P}\left(\bigcup_i A_i\right) = \sum_i \mathbb{P}(A_i) \quad \text{(countable additivity)}.$$

Lecture 2

9th September

EXAMPLE 2

Flip a fair coin.

- Sample space: $\Omega = \{\text{H}, \text{T}\}$; that is, $|\Omega| = 2$.
- $\mathcal{F} = \{\emptyset, \{\text{H}\}, \{\text{T}\}, \{\text{H}, \text{T}\}\}$; that is, $|\mathcal{F}| = 2^{|\Omega|} = 4$.

Whenever Ω is countable, we define $\mathcal{F}$ to be the set of all subsets of Ω, $\mathcal{F} = 2^\Omega$ (we can always choose the power set of Ω as our discrete σ-algebra).

- H is an outcome, $\text{H} \in \Omega$.
- $\emptyset$ is an event, $\emptyset \in \mathcal{F}$.
- $\{\text{H}\}$ is an event, but H is not an event, and $\{\text{H}\}$ is not an outcome.
- $\mathbb{P}(\emptyset) = 0$.
- $\mathbb{P}(\{\text{H}\}) = 1/2$.
- $\mathbb{P}(\{\text{T}\}) = 1/2$.
- $\mathbb{P}(\{\text{H}, \text{T}\}) = 1$.
- $\mathbb{P}(\text{H}) = \text{undefined}$.

EXAMPLE 3

Let $\Omega = \{(x,y) \in \mathbf{R}^2 : x^2 + y^2 \leq 25\}$ (disc of radius 5). Suppose that we have a bullseye of radius 1, the probability of hitting the bullseye is $1/25$.

$$\begin{aligned} \text{Bullseye} &= \{(x,y) \in \Omega : x^2 + y^2 \leq 1\}. \\ \mathbb{P}(\text{Bullseye}) &= \frac{\text{Area(Bullseye)}}{\text{Area}(\Omega)} \\ &= \frac{\pi \cdot 1^2}{\pi \cdot 5^2} \\ &= \frac{1}{25}. \end{aligned}$$

My σ-algebra $\mathcal{F}$ for dart-throwing will be the smallest σ-algebra that includes all sets of the form

$$\big((a,b] \times (c,d]\big) \cap \Omega,\ a < b,\ c < d,\ a,b,c,d \in \mathbf{R}.$$

- $|\mathbf{N}| = |\mathbf{Z}| = |\mathbf{Q}| = \aleph_0$.
- $|\mathbf{R}| = \big|[0,1]\big| = |\mathbf{R}^n| = 2^{\mathbf{N}} = 2^{\aleph_0}$.

- $|2^{\mathbf{R}^2}| = 2^{2^{\aleph_0}} \gg 2^{\aleph_0}$.

PROPOSITION 1

Given a probability space $(\Omega, \mathcal{F}, \mathbb{P})$*,*

(i) *For all* $A \in \mathcal{F}$*,* $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$*.*

(ii) $\mathbb{P}(\emptyset) = 0$*.*

(iii) $\forall A \in \mathcal{F}$*,* $\mathbb{P}(A) \leq 1$*.*

(iv) $\forall A, B \in \mathcal{F}$*,* $\mathbb{P}(B \cap A^c) = \mathbb{P}(B) - \mathbb{P}(A \cap B)$*.*

Proof:

(i) By (S2), $A^c \in \mathcal{F}$. Since $A^c \cap A = \emptyset$,

$$\begin{aligned} \mathbb{P}(A^c) + \mathbb{P}(A) &= \mathbb{P}(A^c \cup A) && \text{by (P3)} \\ &= \mathbb{P}(\Omega) \\ &= 1 && \text{by (P1).} \end{aligned}$$

(ii) $\mathbb{P}(\emptyset) = \mathbb{P}(\Omega^c) = 1 - \mathbb{P}(\Omega) = 0$ by (i).

(iii) $\mathbb{P}(A) = 1 - \mathbb{P}(A^c) \leq 1$ since $\mathbb{P}(A^c) \geq 0$.

(iv) $(A \cap B) \subseteq A$, and $(A^c \cap B) \subseteq A^c$, so $\big((A \cap B) \cap (A^c \cap B)\big) \subseteq (A \cap A^c) = \emptyset$. Thus,

$$\begin{aligned} \mathbb{P}(A \cap B) + \mathbb{P}(A^c \cap B) &= \mathbb{P}\big((A \cap B) \cup (A^c \cap B)\big) \\ &= \mathbb{P}\big(B \cap (A \cup A^c)\big) \\ &= \mathbb{P}(B \cap \Omega) \\ &= \mathbb{P}(B). \end{aligned}$$

THEOREM 4: Inclusion-exclusion for two events

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

Proof:

$$\begin{aligned} A \cup B &= A \cup (B \cap \Omega) \\ &= A \cup (B \cap (A \cup A^c)) \\ &= (A \cup (B \cap A)) \cup (B \cap A^c) \\ &= A \cup (B \cap A^c). \end{aligned}$$

Therefore, A is disjoint from $B \cap A^c$. Thus,

$$\begin{aligned} \mathbb{P}(A \cup B) &= \mathbb{P}(A) + \mathbb{P}(B \cap A^c) \\ &= \mathbb{P}(A) + (\mathbb{P}(B) - \mathbb{P}(A \cap B)). \end{aligned}$$

THEOREM 5: Inclusion-exclusion principle for probabilities

For any $A_1, A_2, \ldots, A_n \in \mathcal{F}$,

$$\begin{aligned}\mathbb{P}\left(\bigcup_{i=1}^{n} A_i\right) &= \underbrace{\mathbb{P}(A_1) + \mathbb{P}(A_2) + \cdots + \mathbb{P}(A_n)}_{n \text{ terms}} \\ &\quad \underbrace{- \mathbb{P}(A_1 \cap A_2) - \cdots - \mathbb{P}(A_{n-1} \cap A_n)}_{\binom{n}{2} \text{ terms}} \\ &\quad \underbrace{+ \mathbb{P}(A_1 \cap A_2 \cap A_3) + \cdots}_{\binom{n}{3} \text{ terms}} \\ &\quad - \mathbb{P}(A_1 \cap A_2 \cap A_3 \cap A_4) - \cdots \\ &\quad \vdots \\ &= \sum_{J \subseteq \{1,2,\ldots,n\}, J \neq \emptyset} (-1)^{|J|+1}\, \mathbb{P}\left(\bigcap_{i \in J} A_i\right).\end{aligned}$$

PROPOSITION 2: Bonferroni's Inequality

$$\mathbb{P}(A \cap B) \geq \mathbb{P}(A) + \mathbb{P}(B) - 1.$$

Proof: Using the inclusion-exclusion theorem, we have

$$\begin{aligned}\mathbb{P}(A \cap B) &= \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cup B) \\ &\geq \mathbb{P}(A) + \mathbb{P}(B) - 1 \qquad \text{since } \mathbb{P}(A \cup B) \leq 1.\end{aligned}$$

Lecture 3

14th September

EXAMPLE 4

Suppose we have 4 shirts, 3 pairs of blue jeans, 2 pairs of shorts, and 2 pairs of shoes. How many outfits can we make?

Solution: $4 \times (3 + 2) \times 2$.

EXAMPLE 5

Suppose we have a bag of 7 marbles numbered $1, 2, \ldots, 7$. We pick one marble uniformly (equal probability) at random, then put it back in the bag. Repeat this process three more times. We care about the order.

i. How many outcomes are in this experiment?

ii. What is $\mathbb{P}\big((2, 4, 2, 7)\big)$?

iii. What is $\mathbb{P}\big(\{\text{all } 4 \text{ picks are even numbers}\}\big)$.

Solution:

i. This is known as **sampling with replacement**. In our example, $|\Omega| = 7^4$. We can represent our sample space as the set of ordered quadruples.

$$\begin{aligned}\Omega &= \big\{(a,b,c,d) : a,b,c,d \in \{1,2,3,4,5,6,7\}\big\} \\ &= \{1,2,3,4,5,6,7\}^4.\end{aligned}$$

The set of ordered quadruples (or 4-tuples) of numbers 1 to 7.

ii. $1/7^4$.

iii. $(3/7)^4 = 3^4/7^4$.

DEFINITION 10

The **Cartesian product** of two sets A and B is

$$A \times B = \big\{(a,b) : a \in A, b \in B\big\}.$$

For example, $\{x,y\} \times \{1,2,3\} = \big\{(x,1),(x,2),(x,3),(y,1),(y,2),(y,3)\big\}$. That is,

$$A^n = \underbrace{A \times \cdots \times A}_{n \text{ times}}.$$

PROPOSITION 3

If Ω is countable and $\mathcal{F} = 2^\Omega$, then

$$\forall A \subseteq \Omega,\ \mathbb{P}(A) = \sum_{i \in A} \mathbb{P}(\{i\}).$$

Proof: Follows from countable additivity.

PROPOSITION 4

If Ω is finite and all outcomes are equally likely (i.e., $\forall x, y \in \Omega$, $\mathbb{P}(\{x\}) = \mathbb{P}(\{y\})$), then

$$\forall A \subseteq \Omega,\ \mathbb{P}(A) = \frac{|A|}{|\Omega|}.$$

Sampling without Replacement (Ordered)

EXAMPLE 6

Suppose we do the same experiment as Example 5, but we don't pick marbles back after picking them. Then, $|\Omega| = 7 \times 6 \times 5 \times 4 = \frac{7!}{3!}$, and

$$\Omega = \big\{(a,b,c,d) \in \{1,2,\ldots,7\}^4 : a \neq b \neq c \neq d \neq a \neq c, b \neq d\big\}.$$

These 4-tuples without repeats are called 4-arrangements.

Sampling without Replacement (Unordered)

EXAMPLE 7

We reach in and grab 4 marbles all at once.

$$\Omega = \big\{A \subseteq \{1, 2, \ldots, 7\} : |A| = 4\big\}.$$

Hence,

$$|\Omega| = \frac{7 \times 6 \times 5 \times 4}{4!} = \binom{7}{4}.$$

These are called 4-combinations. Every 4-combination can be matched up with $4!$ 4-arrangements. So,

$$\big|\{\text{4-arrangements}\}\big| = 4! \times \big|\{\text{4-combinations}\}\big|,$$

and we can re-arrange the equation above to get the number of 4-combinations.

EXAMPLE 8

Suppose we have a standard deck of cards (52 cards where there are 13 ranks and 4 suits).

i. Number of events that we get a full house (3 cards of one rank, and 2 cards of another rank)?

ii. Number of events that we get two pairs (2 cards of one rank, 2 cards of another rank, and one last card of a different rank).

Solution:

i. Number of events:

$$\binom{13}{1}\binom{4}{3}\binom{12}{1}\binom{4}{2} = 13 \times 4 \times 12 \times 6.$$

ii. Number of events:

$$\binom{13}{2}\binom{4}{2}\binom{4}{2} \times 44.$$

Conditional Probability

Idea: Revising your estimate based on partial information.

EXAMPLE 9

- 38.0M Canadians.
- 4.23M positive COVID-19 tests in Canada (pretend all distinct people).

$$\mathbb{P}(\{\text{positive}\}) = \frac{4.23 \times 10^6}{3.8 \times 10^6} \approx 11.1\%$$

Now, suppose we have further data for Quebec.

- 8.49M people in Quebec.
- 1.19M positive tests in Quebec.

$$\mathbb{P}(\{\text{positive}\} \mid \{\text{QC}\}) = \frac{1.19}{8.49} \approx 14.0\%.$$

DEFINITION 11

If A and B are events, and $\mathbb{P}(B) > 0$, then the **conditional probability of A given B** is

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

DEFINITION 12

Events A and B are **independent** if

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

Equivalently, A and B are **independent** if either $\mathbb{P}(B) = 0$ or $\mathbb{P}(A \mid B) = \mathbb{P}(A)$.

EXAMPLE 10

Roll a fair 6-sided die. Let $A = \{1, 2\}$, $B = \{1, 3, 5\}$, $C = \{2, 4, 6\}$.

$$\begin{aligned}
\mathbb{P}(A) &= \frac{2}{6} = \frac{1}{3}. \\
\mathbb{P}(B) &= \frac{3}{6} = \frac{1}{2}. \\
\mathbb{P}(A \cap B) &= \mathbb{P}(\{1\}) \\
&= \frac{1}{6} \\
&= \mathbb{P}(A)\,\mathbb{P}(B) \\
&= \frac{1}{3} \times \frac{1}{2}. \\
\mathbb{P}(C) &= \frac{3}{6} = \frac{1}{2}. \\
\mathbb{P}(B \cap C) &= \mathbb{P}(\varnothing) = 0 \neq \frac{1}{2} \times \frac{1}{2}.
\end{aligned}$$

Therefore, B and C are not independent, but they are disjoint events. In probability theory, *disjoint events* are also called **mutually exclusive events**.

DEFINITION 13

If A and B are **disjoint**, then

- $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$;
- $\mathbb{P}(A \cap B) = 0$.

If A and B are **independent**, then

- $\mathbb{P}(A \cap B) = \mathbb{P}(A)\,\mathbb{P}(B)$;
-

$$\begin{aligned}
\mathbb{P}(A \cup B) &= \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A)\,\mathbb{P}(B) \\
&= (\mathbb{P}(A) - 1)(1 - \mathbb{P}(B)) + 1 \\
&= 1 - (1 - \mathbb{P}(A))(1 - \mathbb{P}(B)) \\
&= 1 - \mathbb{P}(A^c)\,\mathbb{P}(B^c) \\
&= 1 - \mathbb{P}((A \cup B)^c) \\
&= 1 - \mathbb{P}(A^c \cap B^c).
\end{aligned}$$

This proves that A^c is independent of B^c.

EXAMPLE 11

Suppose we have a standard deck of cards. What is the probability that we have four aces if we select four cards?

$$\frac{4}{52} \times \frac{3}{1} \times \frac{2}{50} \times \frac{1}{49} = \frac{1}{\binom{52}{4}}.$$

Hence,

$$\mathbb{P}(A_1 \cap A_2 \cap A_3 \cap A_4) = \mathbb{P}(A_1)\,\mathbb{P}(A_2 \mid A_1)\,\mathbb{P}(A_3 \mid A_1 \cap A_2)\,\mathbb{P}(A_4 \mid A_1 \cap A_2 \cap A_3).$$

Lecture 4

16th September

DEFINITION 14

A **partition** of a set S is a collection of subsets of $A_1, \ldots, A_n \subseteq S$ with the properties

(i) $A_1 \cup A_2 \cup \cdots \cup A_n = S$

(ii) $\forall 1 \leq i < j \leq n$, $A_i \cap A_j = \emptyset$.

THEOREM 6: Law of Total Probability

If $A_1, \ldots, A_n$ is a partition of Ω into events and $B \in \mathcal{F}$, then

$$\mathbb{P}(B) = \mathbb{P}(A_1)\,\mathbb{P}(B \mid A_1) + \mathbb{P}(A_2)\,\mathbb{P}(B \mid A_2) + \cdots + \mathbb{P}(A_n)\,\mathbb{P}(B \mid A_n) = \sum_{i=1}^{n} \mathbb{P}(A_i)\,\mathbb{P}(B \mid A_i).$$

EXAMPLE 12

- 20% of students in STATS 2D are first years, 45% are second years, and 35% are third years.
- 25% of first years are getting an A, along with 35% of second years, and 50% of third years.

What's the overall percentage who are getting an A?

$$A_n = \{n^{\text{th}} \text{ year students}\}.$$

$\{A_1, A_2, A_3\}$ is a partition of any class Ω.

$$B = \{\text{students getting an A}\}.$$

$$\mathbb{P}(B) = 20\% \cdot 25\% + 45\% \cdot 35\% + 35\% \cdot 50\%.$$

Bayes Rule allows us to flip the direction of conditioning.

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

$$\begin{aligned}\mathbb{P}(\{\text{third year}\} \mid \{\text{getting an A}\}) &= \mathbb{P}(A_3 \mid B) \\ &= \frac{\mathbb{P}(A_3)\,\mathbb{P}(B \mid A_3)}{\sum_{i=1}^{3} \mathbb{P}(A_i)\,\mathbb{P}(B \mid A_i)} \\ &= \end{aligned}$$

EXAMPLE 13: Monty Hall Problem

- Let $A_j = \{\text{car behind door } j\}$ for $j = 1, 2, 3$.
- Let $G_2 = \{\text{Monty reveals goat behind door } 2\}$.
- For simplicity, assume we choose door 1 first.

$$
\begin{aligned}
\mathbb{P}(A_1 \mid G_2) &= \frac{\mathbb{P}(A_1)\,\mathbb{P}(G_2 \mid A_1)}{\sum_{i=1}^{3} \mathbb{P}(A_j)\,\mathbb{P}(G_2 \mid A_j).} \\
&= \frac{\frac{1}{3}\cdot\frac{1}{2}}{\frac{1}{3}\cdot\frac{1}{2} + \frac{1}{3}\cdot 0 + \frac{1}{3}\cdot 1} \\
&= \frac{1/6}{1/6 + 1/3} \\
&= \frac{1}{1+2} \\
&= \frac{1}{3}.
\end{aligned}
$$

DEFINITION 15

A **random variable** is a (measurable) function on a probability space. A real-valued radom variable is a function $X \colon \Omega \to \mathbf{R}$.

EXAMPLE 14

Suppose we flip a fair coin three times. (We care about the order because we want each event to be equally likely to occur.) There are 8 possible outcomes:

$$
\begin{aligned}
\Omega &= \{\mathrm{H}, \mathrm{T}\}^3 \\
&= \{\mathrm{HHH}, \mathrm{HHT}, \mathrm{HTH}, \cdots, \mathrm{TTT}\}.
\end{aligned}
$$

If X is the number of heads tossed, then it is the function

- $\mathrm{HHH} \to 3$;
- $\mathrm{HHT}, \mathrm{HTH}, \mathrm{THH} \to 2$;
- $\mathrm{HTT}, \mathrm{THT}, \mathrm{TTH} \to 1$;
- $\mathrm{TTT} \to 0$.

Therefore, $X(\mathrm{HTT}) = 1$.

DEFINITION 16

A random variable is **discrete** if it only has countably many possible values, meaning range(X) is countable.

DEFINITION 17

If X is discrete, then it has a **probability function** (PMF)

$$P_X \colon \text{codomain}(X) \to [0, 1].$$

$$P_X(k) = \mathbb{P}(X = k).$$

EXAMPLE 15

In our coin tossing example,

$$\begin{aligned} P_X(0) &= \frac{1}{8}. \\ P_X(1) &= \mathbb{P}(X = 1) \\ &= \mathbb{P}(\{\text{TTH}, \text{THT}, \text{HTT}\}) \\ &= \frac{3}{8}. \\ P_X(2) &= \frac{3}{8}. \\ P_X(3) &= \frac{1}{8}. \\ P_X(k) &= 0 \text{ for } k \notin \{0, 1, 2, 3\}. \end{aligned}$$

DEFINITION 18

Given a real-valued random variable $X : \Omega \to \mathbf{R}$, the **probability distribution of** X is the probability measure

$$\mathcal{L}_X(A) = \mathbb{P}(\{X \in A\}) = \mathbb{P}(\{x \in \Omega : X(w) \in A\})$$

for any reasonably nice (Borel) subset $A \subseteq \mathbf{R}$.

EXAMPLE 16

In our coin tossing example,

$$\mathcal{L}_X(A) = \frac{1}{8}\mathbb{I}\{0 \in A\} + \frac{3}{8}\mathbb{I}\{1 \in A\} + \frac{3}{8}\mathbb{I}\{2 \in A\} + \frac{1}{8}\mathbb{I}\{3 \in A\}.$$

$$\mathcal{L}_X([1/2, 2 \cdot 1/2]) = \frac{3}{8} + \frac{3}{8} = \frac{3}{4}.$$

DEFINITION 19

For any real-valued random variable $X : \Omega \to \mathbf{R}$, the **cumulative distribution function** (CDF) of X is the function

$$F_X(t) = \mathbb{P}(\{X \le t\}) = \mathbb{P}(\{w \in \Omega : X(w) \le t\}).$$

EXAMPLE 17

In our coin tossing example,

- $F_X(-1) = 0$.
- $F_X(1) = \frac{1}{8} + \frac{3}{8} = \frac{1}{2}$.
- $F_X(1.5) = \frac{1}{2}$.

$$F_X(t) = \begin{cases} 0 & t < 0 \\ 1/8 & 0 \le t < 1 \\ 1/2 & 1 \le t < 2 \\ 7/8 & 2 \le t < 3 \\ 1 & t \ge 3 \end{cases}.$$

THEOREM 7

Two real-valued random variables have the same distribution if and only if their CDFs are equal.

DEFINITION 20

A random variable has a Uniform distribution on $[0, 1]$ if it has CDF

$$F_X(t) = \begin{cases} t & 0 \le t \le 1 \\ 0 & t < 0 \\ 1 & t > 1 \end{cases}$$

THEOREM 8

$F : \mathbf{R} \to [0, 1]$ *is a CDF for some random variable if and only if*

(i) $\lim_{t\to\infty} F(t) = 1$;

(ii) $\lim_{t\to-\infty} F(t) = 0$;

(iii) F *is non-decreasing; that is,* $F(s) \le F(t)$ *for all* $-\infty < s \le t < \infty$.

EXAMPLE 18

Suppose we have a dart board with radius 1 ft.

$$\Omega = \{(x, y) \in \mathbf{R}^2 : x^2 + y^2 \le 1\}.$$

$$\begin{aligned} F_R(t) &= \mathbb{P}(\{R \le t\}) \\ &= \mathbb{P}(\{(x, y) \in \Omega : x^2 + y^2 \le t^2\}) \\ &= \frac{\text{Area(radius } t \text{ circle)}}{\text{Area(unit circle)}} \\ &= \frac{\pi t^2}{\pi \cdot 1^2} \\ &= t^2. \end{aligned}$$

$$F_R(t) = \begin{cases} 0 & t < 0 \\ t^2 & 0 \le t \le 1 \\ 1 & t > 1 \end{cases}$$

DEFINITION 21

A random variable X is continuous if its CDF F_X is continuous. In that case, it has a probability density function (PDF)

$$f_X = \frac{\mathrm{d}F_X}{\mathrm{d}t}.$$

LECTURE 5
21st September

DEFINITION 22

Fix some event $B \in \mathcal{F}$ with $\mathbb{P}(B) > 0$. Define $\mu\colon \mathcal{F} \to \mathbf{R}$ by

$$\mu(A) = \mathbb{P}(A \mid B).$$

THEOREM 9

μ is a probability measure on $(\Omega, \mathcal{F})$. Conditional probabilities are a probability measure.

Proof: We need to check properties (i)–(iii) for μ.

(i)

$$\begin{aligned} \mathbb{P}(\Omega) &= \mathbb{P}(\Omega \mid B) \\ &= \frac{\mathbb{P}(\Omega \cap B)}{\mathbb{P}(B)} \\ &= \frac{\mathbb{P}(B)}{\mathbb{P}(B)} \\ &= 1. \end{aligned}$$

(ii) $\forall A \in \mathcal{F}$, $\mu(A) = \dfrac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} \geq 0$.

(iii) Suppose $A_1, A_2, \ldots$ are disjoint events. Then,

$$\begin{aligned} \mu(A_1 \cup A_2 \cup \cdots) &= \mathbb{P}(A_1 \cup A_2 \cdots \mid B) \\ &= \frac{\mathbb{P}((A_1 \cup A_2 \cup \cdots) \cap B)}{\mathbb{P}(B)} \\ &= \frac{\mathbb{P}((A_1 \cap B) \cup (A_2 \cap B) \cup \cdots)}{\mathbb{P}(B)}. \end{aligned}$$

Note that for all $1 \leq i < j$, $(A_i \cap B) \cap (A_j \cap B) = A_i \cap A_j \cap B = \emptyset \cap B = \emptyset$, so the events $(A_1 \cap B), (A_2 \cap B), \ldots$ are pairwise disjoint. Thus, by the countable additivity of $\mathbb{P}$,

$$\begin{aligned} \mu(A_1 \cup A_2 \cup \cdots) &= \frac{\mathbb{P}(A_1 \cap B) + \mathbb{P}(A_2 \cap B) + \cdots}{\mathbb{P}(B)} \\ &= \mu(A_1) + \mu(A_2) + \cdots, \end{aligned}$$

as desired.

REMARK 1: Expected value of Geometric Series

$$\begin{aligned}
\frac{\mathbb{E}[X]-(1-p)\,\mathbb{E}[X]}{p} &= \sum_{k=1}^{\infty} k(1-p)^{k-1} - \sum_{j=1}^{\infty} j(1-p)^j \\
&= \sum_{j=0}^{\infty}(j+1)(1-p)^j - \sum_{j=1}^{\infty} j(1-p)^j \qquad \text{sum index } j=k+1 \\
&= 1\cdot(1-p)^0 + \sum_{j=1}^{\infty}(1-p)^j[(j+1)-j] \\
&= 1 + \frac{(1-p)^1}{1-(1-p)} \\
&= 1 + \frac{1-p}{p} \\
&= 1 + \frac{1}{p} - \frac{p}{p} \\
&= \frac{1}{p}.
\end{aligned}$$

Therefore, we have

$$\begin{aligned}
\mathbb{E}[X]\frac{1-(1-p)}{p} &= \frac{1}{p} \\
\mathbb{E}[X]\frac{p}{p} &= \frac{p}{p} \\
\mathbb{E}[X] &= \frac{1}{p}.
\end{aligned}$$

EXAMPLE 19

Roll a 6-sided die until we get a 6. Let $X =$ the number of rolls. Let $B = \{\text{all rolls are even numbers}\}$. What is $\mathbb{E}[X \mid B]$?

Solution:

$$\mathbb{P}(B \mid X=k) = \left(\frac{2}{5}\right)^{k-1}.$$

Hence,

$$\begin{aligned}
\mathbb{P}(B) &= \sum_{k=1}^{\infty} \mathbb{P}(\{X = k\})\, \mathbb{P}(B \mid X = k) \\
&= \sum_{k=1}^{\infty} \left(\frac{5}{6}\right)^{k-1} \frac{1}{6} \left(\frac{2}{5}\right)^{k-1} \\
&= \frac{1}{6} \sum_{k=1}^{\infty} \left(\frac{2}{6}\right)^{k-1} \\
&= \frac{1}{6} \cdot \frac{1}{1 - 1/3} \\
&= \frac{1}{6} \cdot \frac{3}{2} \\
&= \frac{1}{4}.
\end{aligned}$$

Hence,

$$\begin{aligned}
\mathbb{E}[X \mid B] &= \sum_{k=1}^{\infty} k\, \mathbb{P}(X = k \mid B) \\
&= \sum_{k=1}^{\infty} k \frac{\mathbb{P}(\{X = k\} \cap B)}{\mathbb{P}(B)} \\
&= \frac{1}{\mathbb{P}(B)} \sum_{k=1}^{\infty} k\, \mathbb{P}(\{X = k\})\, \mathbb{P}(B \mid X = k) \\
&= \frac{1}{1/4} \sum_{k=1}^{\infty} k \frac{1}{6} \left(\frac{1}{3}\right)^{k-1} \\
&= 4 \cdot \frac{1}{6} \cdot \frac{3}{2} \underbrace{\sum_{k=1}^{\infty} k \left(\frac{1}{3}\right)^{k-1} \frac{2}{3}}_{\text{EV of GEO}\left(\frac{2}{3}\right)} \\
&= 4 \cdot \frac{1}{6} \cdot \frac{3}{2} \cdot \frac{3}{2} \\
&= \frac{3}{2}.
\end{aligned}$$

THEOREM 10

$p\colon S \to \mathbf{R}$ is a PMF for some RV if and only if

(i) $p(v) \geq 0$ for all $v \in S$, and

(ii) $\sum_{v \in S} p(v) = 1$.

<u>*Remark*</u>*: This implies that $\{v \in S : p(s) > 0$ is countable$\}$.*

EXERCISE 1

The sum of uncountably infinitely many positive numbers always diverges to infinity.

THEOREM 11

$f\colon \mathbf{R} \to \mathbf{R}$ is a PDF for some RV if and only if

(i) $f(x) \geq 0$ for all $x \in \mathbf{R}$, and

(ii) $\int_{-\infty}^{\infty} f(x)\,\mathrm{d}x = 1$.

EXAMPLE 20

Let $U \sim \text{Uniform}[0,1]$. The PDF is

$$f_U(t) = \begin{cases} 1, & 0 \leq t \leq 1, \\ 0, & \text{otherwise.} \end{cases}.$$

Technically, the derivative doesn't exist at 0 since there's a change of direction, but it doesn't matter since we only integrate PDFs.

EXAMPLE 21

Let $U \sim \text{Uniform}[0, 1/2]$. The PDF is

$$f_U(t) = \begin{cases} 2, & 0 \leq t \leq 1/2, \\ 0, & \text{otherwise.} \end{cases}$$

DEFINITION 23

A standard logistic distribution is defined by the CDF

$$F(t) = \frac{1}{1+e^{-t}}.$$

The PDF is

$$f_X(t) = \frac{\mathrm{d}F}{\mathrm{d}x} = (-1)\frac{1}{(1+e^{-t})^2}(-e^{-t}) = \frac{e^{-t}}{(1+e^{-t})^2}.$$

The PDF looks like a bell curve, but with heavier tails.

EXAMPLE 22

Calculate $\mathbb{P}(\{-1 \leq X \leq 1\})$ for the standard logistic distribution.

Solution:

- Method 1:
$$\mathbb{P}(\{-1 \leq X \leq 1\}) = F(1) - F(-1).$$
- Method 2:
$$\mathbb{P}(\{-1 \leq X \leq 1\}) = \int_{-1}^{1} f(t)\,\mathrm{d}t.$$

EXAMPLE 23

Calculate $\mathbb{E}[X]$ for the standard logistic distribution.

Solution:

$$\begin{aligned}\mathbb{E}[X] &= \int_{-\infty}^{\infty} tf(t)\,\mathrm{d}t \\ &= \int_{-\infty}^{\infty} \frac{te^{-t}}{(1+e^{-t})^2}\,\mathrm{d}t \\ &= \text{IBP}.\end{aligned}$$

Lecture 6

23rd September

DEFINITION 24

If f is a function $f\colon A \to B$, then the **pre-image** of a set $C \subseteq B$ under f is

$$f^{-1}(C) = \{x \in A : f(x) \in C\}.$$

The **image** of a set $D \subseteq A$ under f is

$$f(D) = \{f(x) : x \in D\}.$$

EXAMPLE 24

If $f\colon \mathbf{R} \to \mathbf{R}$ is the function $f(x) = x^2$, then the pre-image

$$\begin{aligned} f^{-1}\big([0,4]\big) &= [-2,2]. \\ f^{-1}\big([1,9]\big) &= [-3,-1] \cup [1,3]. \\ f^{-1}\big([-5,-2]\big) &= \emptyset. \end{aligned}$$

REMARK 2

$\forall C, D \subseteq A$,

$$f(C \cup D) = f(C) \cup f(D).$$

It is not always the case (for non-injective functions) that

$$f(C \cap D) = f(C) \cap f(D).$$

In 24, if we consider $C = [-2,-1]$ and $D = [1,2]$, then $C \cap D = \emptyset$, $f(C \cap D) = \emptyset$, and $f(C) \cap f(D) = [1,4] \cap [1,4] = [1,4]$.

PROPOSITION 5

$$\begin{aligned} f^{-1}(C \cup D) &= f^{-1}(C) \cup f^{-1}(D). \\ f^{-1}(C \cap D) &= f^{-1}(C) \cap f^{-1}(D). \end{aligned}$$

Proof: Exercise.

Suppose $X : \Omega \to \mathbf{R}$ is a discrete random variable and $Y = g(X)$ for some $g : \mathbf{R} \to \mathbf{R}$. How would we find the PMF of Y using the PMF p_X of X?

$$\begin{aligned} p_Y(9) &= \mathbb{P}(\{Y = 9\}) \\ &= \mathbb{P}(\{X \in g^{-1}(\{9\})\}) \\ &= \sum_{j \in g^{-1}(\{9\})} p_X(j). \end{aligned}$$

In general,

$$p_Y(k) = \sum_{j \in g^{-1}(\{k\})} p_X(j).$$

THEOREM 12

If $Y = g(X)$ for some random variable X and some (measurable) function $g : \mathbf{R} \to \mathbf{R}$, then for any set of $A \subseteq \mathbf{R}$,

$$\mathbb{P}(\{Y \in A\}) = \mathbb{P}(\{X \in g^{-1}(A)\}).$$

EXAMPLE 25

Suppose $g(x) = \sqrt{x}$, X is a non-negative random variable, and $Y = \sqrt{X}$. Find the CDF of Y.

Solution:

$$\begin{aligned} F_Y(v) &= \mathbb{P}(\{Y \leq v\}) && v \geq 0 \\ &= \mathbb{P}(\{\sqrt{X} \leq v\}) \\ &= \mathbb{P}(\{X \leq v^2\}) \\ &= F_X(v^2). \end{aligned}$$

$\sqrt{x}$ was a monotone increasing function, so it preserved the inequality.

THEOREM 13

Let $Y = g(X)$.

- *If $g(X)$ is a strictly increasing function, then*

$$F_Y(v) = F_X(g^{-1}(v)).$$

- *If $g(X)$ is a strictly decreasing function, then*

$$F_Y(v) = 1 - F_X(g^{-1}(v)).$$

EXAMPLE 26

Let $X \sim \text{Uniform}[0, 1]$.

$$f_X(t) = \begin{cases} 1, & t \in [0,1], \\ 0, & t \notin [0,1], \end{cases} \qquad F_X(t) = \begin{cases} 0, & t < 0, \\ t, & t \in [0,1], \\ 1, & t > 1. \end{cases}$$

If $Y = -\log(X)$. Note that $\log(1) = 0$ and $\lim_{t \to 0} \log(t) = -\infty$. Also,

$$g(x) = -\log(x) \iff -g(x) = \log(x) \iff x = e^{-g(x)},$$

so $g^{-1}(v) = e^{-v}$. For $v \geq 0$,

$$\begin{aligned} F_Y(v) &= 1 - F_X(g^{-1}(v)) \\ &= 1 - F_X(e^{-v}) \\ &= 1 - e^{-v}. \end{aligned}$$

Hence, Y is a continuous RV with PDF

$$f_Y(v) = \frac{\mathrm{d}}{\mathrm{d}v} \begin{cases} 0, & v < 0, \\ 1 - e^{-v}, & v \geq 0 \end{cases} = \begin{cases} 0, & v < 0, \\ e^{-v}, & v \geq 0. \end{cases}$$

Thus, $Y \sim$ EXP(1).

DEFINITION 25

The **quantile function** of a random variable X is the right-continuous (almost) left-inverse of the CDF of X,

$$Q_X(v) = \inf\{t \in \mathbf{R} : F_X(t) > v\}.$$

Hence, if F_X is strictly increasing at t, then

$$Q_X(F_X(t)) = t.$$

$Q_X(90\%)$ is the 90^{th} percentile of the value of X — the value that X is less than 90% of the time.

THEOREM 14

If $U \sim$ Uniform$[0,1]$ and F is a continuous CDF that is strictly increasing, then $F^{-1}(U)$ is a random variable whose CDF is F.

REMARK 3

Suppose X is a continuous random variable, g is a differentiable and strictly increasing. Before, we had $Y = g(X)$, $F_Y(v) = F_X \circ g^{-1}(v)$, so the PDF of Y is

$$\begin{aligned} f_Y(v) &= \frac{\mathrm{d}}{\mathrm{d}v} F_X(g^{-1}(v)) \\ &= f_X(g^{-1}(v))(g^{-1})'(v). \\ &= f_X(g^{-1}(v)) \frac{1}{g'(g^{-1}(v))} \\ &= \frac{f_X \circ g^{-1}(v)}{g' \circ g^{-1}(v)}. \end{aligned}$$

You can think of it as taking the reflection along the line $y = x$ for g.
If g is differentiable and strictly decreasing, then

$$f_Y(v) = -\frac{f_X \circ g^{-1}(v)}{g' \circ g^{-1}(v)}.$$

We can simplify these formulas for any differentiable function g (strictly increasing or decreasing) as

$$f_Y(v) = \frac{f_X \circ g^{-1}(v)}{|g' \circ g^{-1}(v)|}.$$

EXAMPLE 27

What if our function is neither strictly increasing nor decreasing? In general,

$$f_Y(v) = \sum_{f \in g^{-1}(\{v\})} \frac{f_X(t)}{|g'(t)|},$$

for all t such that $g(t) = v$. We require that g is differentiable, and g is not constant on any interval. Let $X \sim \text{Uniform}[0, 2\pi)$, and $Y = \sin^2(X)$. Find $\mathbb{P}(\{Y \leq t\})$.

DEFINITION 26: Expectation

If X is discrete, then

$$\mathbb{E}[X] = \sum_v v\, \mathbb{P}(\{X = v\}) = \sum_v p_X(v).$$

If X is continuous, then

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x f_X(x)\,\mathrm{d}x.$$

Furthermore, if X is discrete then

$$\mathbb{E}[g(X)] = \sum_v g(v) p_X(v),$$

or if X is continuous then

$$\mathbb{E}[g(X)] = \int_{-\infty}^{\infty} g(x) f_X(x)\,\mathrm{d}x.$$

DEFINITION 27: Variance

The variance (or 2$^{\text{nd}}$ central moment) of a random variable X is

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

DEFINITION 28: Moments

For an integer $p \geq 1$, the p^{th} moment of X is $\mathbb{E}[X^p]$. The p^{th} central moment of X is $\mathbb{E}[(X - \mathbb{E}[X])^p]$.

DEFINITION 29: Moment Generating Function (MGF)

The **moment generating function** (MGF) of a random variable X is the function

$$M_X(t) = \mathbb{E}[e^{tX}].$$

REMARK 4

For each value of t that we plug in, we're calculating a different expected value. Why?

(1) The MGF uniquely specifies the probability distribution.

(2) Grants easy access to all moments.

(3) Easy to handle sums of independent random variables.

THEOREM 15

Suppose X and Y are random variables and their MGFs are both defined (integrals exist) in some interval $(-\delta, \delta)$ for some $\delta > 0$. If $M_X(t) = M_Y(t)$ for all $-\delta < t < \delta$, then $X \stackrel{d}{=} Y$.

THEOREM 16

Suppose $X, X_1, X_2 \ldots$ all have MGFs that are defined on $(-\delta, \delta)$ for some $\delta > 0$. If $M_{X_n}(t) \to M_X(t)$ as $n \to \infty$ for all $-\delta < t < \delta$, then $F_{X_n}(x) \to F_X(x)$ as $n \to \infty$ for all $x \in \mathbf{R}$.

THEOREM 17

For $p \geq 1$, if the MGF of X is differentiable p times at $t = 0$, then

$$\mathbb{E}[X^p] = M_X^{(p)}(0).$$

(Rough) Proof:

$$\left(\frac{d}{dt}\right)^p M_X(t) = \left(\frac{d}{dt}\right)^p \mathbb{E}[e^{tX}] \underbrace{=}_{\text{next lecture}} \mathbb{E}\left[\left(\frac{d}{dt}\right)^p e^{tX}\right] = \mathbb{E}[X^p e^{tX}].$$

At $t = 0$, this is $\mathbb{E}[X^p \cdot 1] = \mathbb{E}[X^p]$.

EXAMPLE 28

Let $G \sim \text{GEO}(p)$. Find the MGF of G, and then calculate the first moment.

Solution: The MGF is given by

$$\begin{aligned}
M_G(t) &= \mathbb{E}[e^{tG}] \\
&= \sum_{k=1}^{\infty} e^{tk}(1-p)^{k-1}p \\
&= \sum_{k=1}^{\infty} (e^t)^k (1-p)^{k-1}p \\
&= pe^t \sum_{k=1}^{\infty} ((1-p)e^t)^{k-1} \\
&= pe^t \frac{1}{1-(1-p)e^t} \\
&= \frac{pe^t}{1-(1-p)e^t} \\
&= \frac{p}{e^{-t}-1+p}.
\end{aligned}$$

We can calculate the first moment (expected value) as follows:

$$M_G'(t) = (-1)\frac{p}{(e^{-t}-1+p)^2}(-e^{-t}) \implies M_G'(0) = \frac{p}{(1-1+p)^2}(1) = \frac{p}{p^2} = \frac{1}{p}.$$

THEOREM 18

If $X_1, \ldots, X_n$ are jointly independent random variables, $S = X_1 + X_2 + \cdots + X_n$, and these random variables' MGFs are all defined at some value t, then

$$M_S(t) = M_{X_1}(t) M_{X_2}(t) \cdots M_{X_n}(t).$$

Proof: Since $X_1, \dots, X_n$ are jointly independent, we have

$$\mathbb{E}[e^{tS}] = \mathbb{E}[e^{t(X_1+\cdots+X_n)}] = \mathbb{E}[e^{tX_1+\cdots+tX_n}] = \mathbb{E}[e^{tX_1}\cdots e^{tX_n}] = \mathbb{E}[e^{tX_1}]\cdots\mathbb{E}[e^{tX_n}] = \prod_{i=1}^{n} M_{X_i}(t).$$

EXAMPLE 29

Suppose $I_1, I_2, \dots$ are a sequence of independent and identically distributed (IID) BERN(p) trials $p_{I_j}(0) = 1-p$, $p_{I_j}(1) = p$. Find the MGF of I_j, and then find the MGF of BIN(n,p).

Solution: For a single Bernoulli RV,

$$M_{I_j}(t) = (1-p)\cdot 1 + p \cdot e^{1t} = (1-p) + pe^t.$$

Now, note that the Binomial RV is the sum of n IID Bernoulli trials, so

$$M_S(t) = (1-p+pe^t)^n.$$

EXAMPLE 30

Suppose $N \sim$ POI(λ); that is,

$$p_N(k) = e^{-\lambda}\frac{\lambda^k}{k!},\ k \geq 0.$$

Find the MGF of N and calculate $\mathbb{E}[N]$ using the MGF.

Solution: The MGF of N is

$$\begin{aligned} M_N(t) &= \sum_{k=0}^{\infty} e^{tk} e^{-\lambda}\frac{\lambda^k}{k!} \\ &= e^{-\lambda}\sum_{k=0}^{\infty}\frac{(e^t\lambda)^k}{k!} \\ &= e^{-\lambda}e^{e^t\lambda} \\ &= e^{\lambda(e^t-1)}. \end{aligned}$$

Therefore, the expected value is

$$M_N'(t) = \lambda e^{\lambda(e^t-1)}e^t \implies M_N'(0) = \mathbb{E}[N] = \lambda e^{\lambda(1-1)}e^0 = \lambda.$$

PROPOSITION 6

If $S_k \sim$ BIN$(k, \lambda/k)$ and $N \sim$ POI(λ), then $M_{S_k}(t) \to M_N(t)$ for all $t \in \mathbf{R}$.

Proof: Note that

$$\left(1+\frac{a}{n}\right)^{bn} \xrightarrow{n\to\infty} e^{ab}.$$

Hence,

$$\begin{aligned}
M_{S_k}(t) &= \left(1 - \frac{\lambda}{k} + \frac{\lambda}{k}e^t\right)^k \\
&= \left(1 - \frac{\lambda(e^t-1)}{k}\right)^k \\
&\xrightarrow{k\to\infty} e^{\lambda(e^t-1)} = M_N(t),
\end{aligned}$$

which is the MGF for N, as desired.

PROPOSITION 7

In the same setup as Proposition 6, $p_{S_k}(j) \to p_N(j)$ as $k \to \infty$ for all $j \geq 0$.

Proof:

$$\begin{aligned}
\binom{k}{j}\left(\frac{\lambda}{k}\right)^j\left(1-\frac{\lambda}{k}\right)^{k-j} &= \frac{k(k-1)\cdots(k-j+1)}{j!}\frac{\lambda^j}{k^j}\underbrace{\left(1-\frac{\lambda}{k}\right)^k}_{\xrightarrow{k\to\infty} e^{-\lambda}}\underbrace{\left(1-\frac{\lambda}{k}\right)^{-j}}_{\xrightarrow{k\to\infty} 1} \\
&\xrightarrow{k\to\infty} \frac{k(k-1)\cdots(k-j+1)}{k^j}\frac{\lambda^j}{j!}e^{-\lambda}(1) \\
&\xrightarrow{k\to\infty} \underbrace{\frac{k}{k^j}\cdot\frac{k-1}{k^j}\cdots\frac{k-j+1}{k^j}}_{\xrightarrow{k\to\infty} 1}\frac{\lambda^j}{j!}e^{-\lambda} \\
&\xrightarrow{k\to\infty} \frac{\lambda^j}{j!}e^{-\lambda}
\end{aligned}$$

Lecture 7

5th October

REMARK 5: Algebraic Properties of Expectation and Variance

- $\mathbb{E}[aX+b] = a\,\mathbb{E}[X] + b$.
- $\operatorname{Var}(aX+b) = a^2\operatorname{Var}(X)$.
- $\mathbb{E}[X+Y] = \mathbb{E}[X] + \mathbb{E}[Y]$.
- If X and Y are independent, then $\mathbb{E}[XY] = \mathbb{E}[X]\,\mathbb{E}[Y]$, and we say X and Y are **uncorrelated**.
- To calculate $\operatorname{Var}(X+Y)$, we have

  $$\begin{aligned}
  \operatorname{Var}(X+Y) &= \mathbb{E}[(X+Y)^2] - \mathbb{E}[X+Y]^2 \\
  &= \mathbb{E}[X^2] + 2\,\mathbb{E}[XY] + \mathbb{E}[Y^2] - \mathbb{E}[X]^2 - 2\,\mathbb{E}[X]\,\mathbb{E}[Y] + \mathbb{E}[Y]^2 \\
  &= \operatorname{Var}(X) + \operatorname{Var}(Y) + 2\operatorname{Cov}(X,Y).
  \end{aligned}$$

  If X and Y are uncorrelated, then $\operatorname{Var}(X+Y) = \operatorname{Var}(X) + \operatorname{Var}(Y)$.

PROPOSITION 8

$$\mathbb{P}\left(\left\{\sum_{i=1}^{\infty}|X_i| < \infty\right\}\right) = 1 \implies \mathbb{E}\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{\infty} \mathbb{E}[X_i].$$

EXAMPLE 31

If $S \sim \text{BIN}(n, p)$, then $S = I_1 + \cdots + I_n$, where $I_1, \ldots, I_n \overset{\text{iid}}{\sim} \text{BERN}(p)$ trials; that is,

$$\mathbb{P}(\{I_j = 0\}) = 1 - p,$$
$$\mathbb{P}(\{I_j = 1\}) = p.$$

Hence,

$$\mathbb{E}[I_j] = (0)(1-p) + (1)(p) = p.$$

Therefore,

$$\mathbb{E}[S] = \mathbb{E}\left[\sum_{j=1}^{n} I_j\right] = \sum_{j=1}^{n} \mathbb{E}[I_j] = np.$$

EXAMPLE 32

Suppose I have 100 people at a party. They drop their coats in a pile (all have coats). When they leave, each take a uniform random coat. Let X denote the number of people who get back their own coat.

(a) $\mathbb{P}(\{X = 0\})$,

(b) $\mathbb{E}[X]$,

(c) $\text{Var}(X)$.

Solution: Let $X = I_1 + \cdots + I_n$, where $I_j \sim \text{BERN}(1/100)$. Note that the I_j's are not independent.

(a) Inclusion-exclusion.

(b) $\mathbb{E}[X] = \sum_{i=1}^{100} \mathbb{E}[I_j] = (100)(1/100) = 1.$

(c)

$$\begin{aligned}
\mathbb{E}[X^2] &= \mathbb{E}\left[\left(\sum_{j=1}^{n} I_j\right)^2\right] \\
&= \mathbb{E}\left[\sum_{j=1}^{n} I_j^2 + 2 \sum_{1 \leq k < j \leq n} I_j I_k\right] \\
&= \sum_{j=1}^{n} \mathbb{E}[I_j^2] + 2\binom{100}{2} \mathbb{E}[I_1 I_2] \\
&= \sum_{j=1}^{n} \left[(0)^2(1 - 1/100) + (1)^2(1/100)\right] + 2\frac{100 \cdot 99}{2} \frac{1}{100} \frac{1}{99} \\
&= 2.
\end{aligned}$$

Thus,

$$\text{Var}(X) = 2 - 1 = 1.$$

Converges to $\text{POI}(1)$ as $n \to \infty$.

EXAMPLE 33

Every box of Sugar Bombs cereal has a toy inside. There are 100 different toys and each box contains an i.i.d. uniform random toy. Let X be the number of boxes purchased in order to complete a set of at least one of each toy. Find $\mathbb{E}[X]$.

Solution: Let Y_j be the number of additional trials to get $(j+1)^{\text{st}}$ toy after first j toys. For each j, $Y_j \sim \text{GEO}\big(\frac{100-j}{100}\big)$. Hence,

$$X = Y_0 + Y_1 + \cdots + Y_{99}.$$

Therefore,

$$\begin{aligned}
\mathbb{E}[X] &= \sum_{j=0}^{99} \mathbb{E}[Y_j] \\
&= \sum_{j=0}^{99} \frac{100}{100-j} \\
&= (100) \sum_{j=0}^{99} \frac{1}{100-j} \\
&= 100\left(1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{100}\right) \\
&\approx 100 \ln(100).
\end{aligned}$$

REMARK 6: Measure-Theoretic Integration

Recall Theorem 17. In general, for a random variable $X : \Omega \to \mathbf{R}$ defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$,

$$\mathbb{E}[X] = \int_\Omega X(\omega)\,\mathrm{d}\mathbb{P}(\omega).$$

Recall that for Riemann sums, we draw vertical bars under the function. However, for Lebesgue (measure) integral, we draw horizontal bars, which implies that we do not need a continuous function.
Idea:

$$\begin{aligned}
\mathbb{E}[X] &= \int_\Omega X(\omega)\,\mathrm{d}\mathbb{P}(\omega) \\
&= \lim_{n\to\infty} \sum_{j=-\infty}^{\infty} \mathbb{P}\left(\left\{\frac{j}{n} < X \le \frac{j+1}{n}\right\}\right) \cdot \frac{j}{n}.
\end{aligned}$$

REMARK 7

$$\lim_{x\to\infty} \lim_{y\to\infty} \left(\frac{1}{x}\right)^{1/y} = \lim_{x\to\infty} 1 = 1.$$

$$\lim_{y\to\infty} \lim_{x\to\infty} \left(\frac{1}{x}\right)^{1/y} = \lim_{y\to\infty} 0 = 0.$$

THEOREM 19: Lebesgue Dominated Convergence Theorem

Suppose X is a measurable function (random variable) on a measure probability space $(\Omega, \mathcal{F}, \Omega)$, and $X_1, X_2, X_3, \ldots$ is a sequence of real-valued measurable functions on this space that converge pointwise to X;

that is,

$$\forall \omega \in \Omega,\ \lim_{n\to\infty} X_n(\omega) = X(\omega).$$

Suppose there is some non-negative measurable function Y such that for all $n \geq 1$ and for all $\omega \in \Omega$,

$$|X_n(\omega)| \leq Y(\omega),$$

and $\int_\Omega Y(\omega)\,\mathrm{d}\mathbb{P}(\omega) < \infty$. Then, we conclude that

$$\lim_{n\to\infty} \int_\Omega |X_n(\omega) - X(\omega)|\,\mathrm{d}\mathbb{P}(\omega) = 0.$$

Moreover, $\int_\Omega X(\omega)\,\mathrm{d}\mathbb{P}(\omega)$ exists (is finite) and equals

$$\lim_{n\to\infty} \int_\Omega X(\omega)\,\mathrm{d}\mathbb{P}(\omega).$$

This theorem also holds for infinite measure spaces.

Lecture 8

7th October

Cancelled.

Lecture 9

17th October

THEOREM 20: Dominated Convergence Theorem

Suppose $f_1, f_2, \ldots$ is a sequence of functions mapping some measure space S to $\mathbf{R}$ $(S, \mathcal{A}, \mu)$ is a measure space, and suppose $\forall x \in S$ $\lim_{n\to\infty} f_n(x)$ converges. Let $f(x)$ denote this limit (pointwise convergence). Additionally, suppose there is a function $g\colon S \to \mathbf{R}$ such that

(1) For all $n \geq 1$, for all $x \in S$ $|f_n(x)| \leq g(x)$.

(2) $\int_S g(x)\,\mathrm{d}\mu(x) < \infty$.

Then,

$$\lim_{n\to\infty} \int_S |f_n(x) - f(x)|\,\mathrm{d}\mu(x) = 0.$$

PROPOSITION 9

$\dfrac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}[e^{tX}] = \mathbb{E}[Xe^{tX}]$ *for t near 0.*

For x and t fixed,

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}e^{tx} &= \lim_{\delta\to 0} \frac{e^{(t+\delta)x} - e^{tx}}{\delta} \\ &= \lim_{\delta\to 0} e^{tx}\left(\frac{e^{\delta x} - 1}{\delta}\right).\end{aligned}$$

We want to find $g(x)$

$$\lim_{\delta\to 0} \frac{e^{\delta x} - 1}{\delta} \overset{\text{LHR}}{=} \lim_{\delta\to 0} \frac{xe^{\delta x}}{1} = x.$$

Therefore,

$$g(t,x) = e^{tx}(|x|+1) \implies \left| e^{tx}\frac{e^{\delta x}-1}{\delta} \right| \le g(t,x), \text{ sufficiently small } \delta.$$

Need $\mathbb{E}[g(t,X)] < \infty$: If

$$\mathbb{E}\big[e^{tX}(|X|+1)\big] < \infty,$$

then by the DCT,

$$\mathbb{E}[\lim(\,\cdot\,)] = \lim \mathbb{E}[\,\cdot\,] \iff \mathbb{E}\big[Xe^{tX}\big] = M_X'(t).$$

DEFINITION 30: Hypergeometric Distribution

Suppose we have a bag with N blue balls and M red. We sample k times without replacement and count the number of blue balls picked. Then,

$$H \sim \text{HG}(k; M, N).$$

For $0 \le j \le k$ and $k - M \le j \le N$,

$$p_H(j) = \frac{\binom{N}{j}\binom{M}{k-j}}{\binom{N+M}{k}}.$$

Expectation: Let

$$I_m = \begin{cases} 1, & \text{if } m^{\text{th}} \text{ pick blue,} \\ 0, & \text{otherwise.} \end{cases}$$

Then, $H = I_1 + \cdots + I_k$ so

$$\mathbb{E}[I_m] = 1\,\mathbb{P}(\{I_m = 1\}) = \frac{N}{N+M},\ 1 \le m \le n.$$

Therefore,

$$\mathbb{E}[H] = k\frac{N}{N+M}.$$

Variance:

$$\begin{aligned}
\mathbb{E}[H^2] &= \mathbb{E}\left[\left(\sum_{j=1}^{k} I_j\right)^2\right] \\
&= \mathbb{E}\left[\sum_{j=1}^{k} I_j^2\right] + 2\,\mathbb{E}\left[\sum_{1\le j<i\le n} I_j I_i\right] \\
&= k\,\mathbb{E}[I_1^2] + 2\binom{k}{2}\mathbb{E}[I_1 I_2] \\
&= k\left(\frac{N}{N+M}\right) + k(k-1)(0 + 1\,\mathbb{P}(\{I_1 = I_2 = 1\})) \\
&= \frac{kN}{N+M} + k(k-1)\frac{N}{N+M}\frac{N-1}{N+M-1}.
\end{aligned}$$

Therefore,

$$\text{Var}(H) = \cdots.$$

DEFINITION 31: Negative Binomial Distribution

Suppose we have a coin with probability p of flipping heads. We flip repeatedly until we have r heads ($r \geq 1$). If Y is the number of tosses, then

$$Y \sim \text{NB}(r, p).$$

For $j \geq r$,

$$p_Y(j) = \binom{j-1}{r-1}(1-p)^{j-r}p^r.$$

EXAMPLE 34

If $G_1, \ldots, G_r \overset{\text{iid}}{\sim} \text{GEO}(p)$, then $G_1 + \cdots + G_r \sim \text{NB}(r, p)$.

$$M_{G_1}(t) = \frac{p}{e^{-t} + p - 1} \implies M_Y(t) = \left(\frac{p}{e^{-t} + p - 1}\right)^r.$$

DEFINITION 32: Gamma Function

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1}e^{-x}\,\mathrm{d}x,\ \Re(\alpha) > 0.$$

PROPOSITION 10

1. $\Gamma(1) = 1$.
2. $\Gamma(\alpha + 1) = \alpha\Gamma(\alpha)$ *for* $\alpha > 0$.
3. $\Gamma(1/2) = \sqrt{\pi}$.

1. Simply,

$$\Gamma(1) = \int_0^\infty 1e^{-x}\,\mathrm{d}x = \left[-e^{-x}\right]_0^\infty = 0 - (-1) = 1.$$

2. Integration by parts: let $u = x^\alpha$, $dv = e^{-x}dx$, $du = \alpha x^{\alpha-1}dx$, $v = -e^{-x}$,

$$\begin{aligned}\Gamma(\alpha+1) &= \int_0^\infty x^\alpha e^{-x}\,\mathrm{d}x \\ &= \left[-x^\alpha e^{-x}\right]_0^\infty + \int_0^\infty e^{-x}\alpha x^{\alpha-1}\,\mathrm{d}x \\ &= 0 + \alpha\Gamma(\alpha).\end{aligned}$$

DEFINITION 33

We say $G \sim \text{GAM}(\alpha, \lambda)$ with shape parameter $\alpha > 0$ and rate parameter $\lambda > 0$ if it has pdf

$$f_G(t) = \frac{\lambda^\alpha}{\Gamma(\alpha)}t^{\alpha-1}e^{-\lambda t},\ t > 0.$$

Lecture 10

21st October

Author's Note: Missed lecture due to convocation. The following lecture is typed after the lecture, and the notes were sourced from **Zhang Yiran** (main source), Baek Inwook, and Zhang Zhiyue.

DEFINITION 34: Polya's Urn

Start with one red ($\mathcal{R}$) and one blue ($\mathcal{B}$) ball. At each step, select a ball at random, then put it back into the urn along with an additional ball of the same colour.

Question: Does the percentage of $\mathcal{B}$ converge? If so, to what number?

EXAMPLE 35: Order in Polya's Urn is Irrelevant

$$\begin{aligned}\mathbb{P}\{\mathcal{RBBBRBR}\} &= \frac{1}{2}\frac{1}{3}\frac{2}{4}\frac{3}{5}\frac{2}{6}\frac{4}{7}\frac{3}{8} \\ &= \frac{4!3!}{8!}. \\ \mathbb{P}\{\mathcal{RRRBBBB}\} &= \frac{1}{2}\frac{2}{3}\frac{3}{4}\frac{1}{5}\frac{2}{6}\frac{3}{7}\frac{4}{8} \\ &= \frac{4!3!}{8!}.\end{aligned}$$

Therefore, the two sequences are exchangeable; that is, order of $\mathcal{R}$ and $\mathcal{B}$ is irrelevant.

EXAMPLE 36

$$\mathbb{P}\{3\mathcal{R} + 4\mathcal{B} \text{ in the first 7 picks}\} = \frac{4!3!}{8!}\binom{7}{3} = \frac{1}{8},$$

where we multiplied by $\binom{7}{3}$ because this is the number of ways to make a sequence of $3\mathcal{R}4\mathcal{B}$. Also,

$$\mathbb{P}\{1\mathcal{R} + 6\mathcal{B}\} = \frac{1!6!}{8!}\binom{7}{1} = \frac{1}{8}.$$

EXAMPLE 37: Random Spinner Game

Suppose $X_i \overset{\text{iid}}{\sim} \text{Uniform}[0, 1]$ is independent of Y for $i \geq 1$, and define

$$C_i = \begin{cases} \mathcal{B}, & X_i < Y, \\ \mathcal{R}, & X_i \geq Y. \end{cases}$$

Remarks:

- $\mathbb{P}\{1^{\text{st}} \text{ is } \mathcal{B}\} = \mathbb{P}\{C_1 = \mathcal{B}\} = \mathbb{P}\{X_1 < Y\}$.
- Since X_1 and Y are iid,
$$\mathbb{P}\{X_1 < Y\} = \mathbb{P}\{Y < X_1\}.$$
- Since X_1 and Y are continuous and independent,
$$\mathbb{P}\{X_1 = Y\} = 0.$$

Using these facts, we have

$$2\,\mathbb{P}\{X_1 < Y\} = \mathbb{P}\{X_1 < Y\} + \mathbb{P}\{Y < X_1\} = 1 - \mathbb{P}\{X_1 = Y\} = 1.$$

Therefore, $\mathbb{P}\{X_1 < Y\} = 1/2$, that is $\mathbb{P}(C_1 = \mathcal{B}) = 1/2$.

$$\begin{aligned}\mathbb{P}(\{C_2 = \mathcal{B}\} \mid \{C_1 = \mathcal{B}\}) &= \frac{\mathbb{P}\{C_1 = \mathcal{B}, C_2 = \mathcal{B}\}}{\mathbb{P}\{C_1 = \mathcal{B}\}} \\ &= \frac{\mathbb{P}\{C_1 = C_2 = \mathcal{B}\}}{\mathbb{P}\{C_1 = \mathcal{B}\}} \\ &= \frac{1/3}{1/2} \\ &= \frac{2}{3},\end{aligned}$$

where the numerator was calculated via

$$\begin{aligned}\mathbb{P}\{C_1 = C_2 = \mathcal{B}\} &= \mathbb{P}\{X_1 < Y, X_2 < Y\} \\ &= \int_0^1 \int_0^y \int_0^y 1 \, \mathrm{d}x_1 \, \mathrm{d}x_2 \, \mathrm{d}y \\ &= \frac{1}{3}.\end{aligned}$$

Therefore, $\mathbb{P}(\{C_2 = \mathcal{B}\} \mid \{C_1 = \mathcal{B}\}) = \frac{1/3}{1/2} = \frac{2}{3}$.
For the same reason as before,

$$\mathbb{P}\{X_1 < X_2 < Y\} = \mathbb{P}\{X_2 < X_1 < Y\} = \cdots = \mathbb{P}\{Y < X_1 < X_2\} = \frac{1}{3!} = \frac{1}{6}.$$

$$\begin{aligned}\mathbb{P}(\{C_9 = \mathcal{B}\} \mid \{\mathcal{B}\mathcal{B}\mathcal{R}\mathcal{B}\mathcal{R}\mathcal{R}\mathcal{B}\mathcal{B}\}) &= \mathbb{P}\{X_1, X_2, X_4, X_7, X_8 < Y,\ X_3, X_5, X_6 > Y\} \\ &= \frac{5!3!}{9!}.\end{aligned}$$

The random spinner game is the same process as Polya's urn.

- Conditionally given Y, the C_i's are independent each with probability Y of being $\mathcal{B}$.
- By the law of large numbers, the percentage of $\mathcal{B}$ picks converges to Y.

THEOREM 21: De Finetti's Theorem for Polya's Urn

The percentage of $\mathcal{B}$ picks converges almost surely (100% probability to converge). Let Y denote the limit,

- $Y \sim$ *Uniform*$[0, 1]$.
- *Given Y, the picks are conditionally independent each with probability Y of being $\mathcal{B}$.*

EXAMPLE 38

The conditional cdf of Y given $C_1 = \mathcal{B}$ is

$$\begin{aligned}F_{Y|C_1=\mathcal{B}} &= \mathbb{P}(\{Y \le t\} \mid \{C_1 = \mathcal{B}\}) \\ &= \frac{\mathbb{P}(\{Y \le t\} \cap \{C_1 = \mathcal{B}\})}{\mathbb{P}\{C_1 = \mathcal{B}\}} \\ &= \frac{t^2/2}{1/2} \\ &= t^2,\end{aligned}$$

where numerator is calculated via

$$\mathbb{P}(\{Y \leq t\} \cap \{C_1 = \mathcal{B}\}) = \int_0^t \int_0^y 1 \, \mathrm{d}x_1 \, \mathrm{d}y = \frac{t^2}{2}.$$

The conditional pdf of Y given $C_1 = \mathcal{B}$ is

$$f_{Y|C_1=\mathcal{B}}(t) = \begin{cases} 2t, & t \in [0,1], \\ 0, & \text{otherwise.} \end{cases}$$

THEOREM 22: Law of Total Probability (Continuous)

If Y is a continuous random variable with pdf f_Y, then for any event A,

$$\mathbb{P}(A) = \int_{-\infty}^{\infty} \mathbb{P}(A \mid \{Y = y\}) f_Y(y) \, \mathrm{d}y,$$

given we can make sense of the conditional probability.

EXAMPLE 39

Using the Law of Total Probability, the conditional cdf of Y given $\mathcal{BBRBRRBB}$

$$\begin{aligned} F_{Y|\mathcal{BBRBRRBB}}(t) &= \mathbb{P}(\{Y \leq t\} \cap \{\mathcal{BBRBRRBB}\}) \\ &= \int_0^t \mathbb{P}(\{\mathcal{BBRBRRBB}\} \mid \{Y = y\}) \, \mathrm{d}y \\ &= \int_0^t \frac{y^5(1-y)^3}{(5!3!)/(9!)} \, \mathrm{d}y \\ &= \frac{9!}{5!3!} \int_0^t y^5(1-y)^3 \, \mathrm{d}y. \end{aligned}$$

The conditional pdf of Y given $\mathcal{BBRBRRBB}$ is

$$f_{Y|\mathcal{BBRBRRBB}}(t) = \begin{cases} \frac{9!}{5!3!} t^5(1-t)^3, & t \in [0,1], \\ 0, & \text{otherwise,} \end{cases}$$

which is the $\mathrm{Beta}(6,4)$ distribution.

Lecture 11
26th October

DEFINITION 35

The **joint probability mass function** (joint pmf) of a sequence $X_1, \ldots, X_n$ of discrete random variables is a function $p \colon \mathbf{R}^n \to [0,1]$ with

$$p(a_1, \ldots, a_n) = \mathbb{P}\big(\{X_1 = a_1\} \cap \cdots \cap \{X_n = a_n\}\big).$$

EXAMPLE 40

Suppose we are rolling two 4-sided die independently. The joint pmf is

$$p(a,b) = \begin{cases} \frac{1}{16}, & a,b \in \{1,2,3,4\}, \\ 0, & \text{otherwise.} \end{cases}$$

EXAMPLE 41

Suppose we roll a die and flip a coin. Let X be a die roll and

$$Y = \begin{cases} X, & \text{if H,} \\ 5 - X, & \text{if T.} \end{cases}$$

b \ a	1	2	3	4
1	1/8	0	0	1/8
2	0	1/8	1/8	0
3	0	1/8	1/8	0
4	1/8	0	0	1/8

Note that

$$\mathbb{P}(\{Y = 3\}) = \frac{1}{8} + \frac{1}{8} = \frac{1}{4}.$$

REMARK 8

If p is the joint pmf of (X, Y), then

$$\mathbb{P}(\{X = k\}) = \sum_j \underbrace{p(k,j)}_{\mathbb{P}(\{X=k,Y=j\})} .$$

$$\mathbb{P}(\{Y = k\}) = \sum_j p(j,k).$$

In this context of starting with a joint distribution, distribution of components are called "marginal distributions."

DEFINITION 36

If p is the joint pmf of $X_1, \dots, X_n$, then the marginal distribution of X_k for any $k \in \{1, 2, \dots, n\}$ is

$$\mathbb{P}(\{X_k = a\}) = \sum_{b_1,\dots,b_{k-1},b_{k+1},\dots,b_n} p(b_1, \dots, b_{k-1}, a, b_{k+1}, \dots, b_n).$$

THEOREM 23

$X_1, \dots, X_n$ (discrete) are jointly independent if and only if their joint pmf is the product of their individual pmfs; that is,

$$p_{X_1,\dots,X_n}(b_1, \dots, b_n) = p_{X_1}(b_1) \cdots p_{X_n}(b_n).$$

EXAMPLE 42

Let X and Y be independent with pmfs

$$p_X(-1) = \frac{1}{2},$$
$$p_X(0) = \frac{1}{4},$$
$$p_X(1) = \frac{1}{4},$$
$$p_Y(0) = \frac{1}{3},$$
$$p_Y(1) = \frac{2}{3}.$$

They have joint pmf

		X		
		-1	0	1
Y	0	$1/6$	$1/12$	$1/12$
	1	$1/3$	$1/6$	$1/6$

DEFINITION 37

If $X_1, \ldots, X_n$ are continuous random variables and $f\colon \mathbf{R}^n \to [0, \infty)$ $(A \subseteq \mathbf{R}^n)$ that satisfies

$$\int \cdots \int_A f(x_1, \ldots, x_n) \, \mathrm{d}x_1 \cdots \mathrm{d}x_n$$

then f is a joint pdf for these variables, and they are said to be jointly continuous.

EXAMPLE 43

Suppose we have two continuous random variables X and Y.

$$\mathbb{P}(\{(X, Y) \in A\}) = \iint_A f(x, y) \, \mathrm{d}x \, \mathrm{d}y.$$

If A is a rectangle, then $A = [a, b] \times [c, d]$, which implies

$$\mathbb{P}(\{a \leq X \leq b, c \leq Y \leq d\}) = \int_c^d \int_a^b f(x, y) \, \mathrm{d}x \, \mathrm{d}y.$$

THEOREM 24

$X_1, \ldots, X_n$ (continuous) are jointly independent if and only if they are jointly continuous with joint pdf

$$f_{X_1, \ldots, X_n}(a_1, \ldots, a_n) = f_{X_1}(a_1) \cdots f_{X_n}(a_n).$$

EXAMPLE 44

$$f(x, y) = \begin{cases} 2x^2, & x \in [0, 1], \ |y| \leq x, \\ 0, & \text{otherwise.} \end{cases}$$

Verifying we have a probability density function:

$$\begin{aligned}\int_0^1 \int_{-x}^{x} 2x^2 \, \mathrm{d}y \, \mathrm{d}x &= \int_0^1 \Big[2x^2 y \Big]_{y=-x}^{y=x} \mathrm{d}x \\ &= \int_0^1 2x^2 (x - (-x)) \, \mathrm{d}x \\ &= \int_0^1 4x^3 \, \mathrm{d}x \\ &= \Big[x^4 \Big]_{x=0}^{x=1} \\ &= 1.\end{aligned}$$

Calculating Probabilities: To calculate $\mathbb{P}(\{Y \geq 1/2\})$, we could work out the system of inequalities: $0 \leq x \leq 1$, $-x \leq y \leq x$, and $1/2 \leq y$ yields

$$1/2 \leq y \leq x \leq 1.$$

Or we can work it out graphically.

$$\begin{aligned}\mathbb{P}\left(\left\{ Y \geq \frac{1}{2} \right\}\right) &= \int_{1/2}^{1} \int_{1/2}^{x} 2x^2 \, \mathrm{d}y \, \mathrm{d}x \\ &= \int_{1/2}^{1} \Big[2x^2 y \Big]_{y=1/2}^{y=x} \mathrm{d}x \\ &= \int_{1/2}^{1} (2x^3 - x^2) \, \mathrm{d}x \\ &= \left[\frac{x^4}{2} - \frac{x^3}{3} \right]_{x=1/2}^{x=1} \\ &= \frac{1}{2} - \frac{1}{32} - \frac{1}{3} + \frac{1}{24}.\end{aligned}$$

DEFINITION 38

The marginal density of X is

$$f_X(t) = \int_{-\infty}^{\infty} f_{X,Y}(t,u) \, \mathrm{d}u.$$

DEFINITION 39: Expectation (Continuous)

$$\mathbb{E}\big[g(X,Y)\big] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x,y) f_{X,Y}(x,y) \, \mathrm{d}x \, \mathrm{d}y.$$

For example, to calculate $\mathbb{E}[XY]$ we use $g(x,y) = xy$.

EXAMPLE 45: Polya Urn

$$\mathbb{P}(\{3^{\text{rd}} \text{ pick R}\} \mid \{\text{BB}\}) = \frac{1}{4}.$$

If Y is the limiting percentage of blue, then

$$\begin{aligned}
\mathbb{P}\left(\left\{Y \leq \frac{1}{2}\right\} \,\middle|\, \{\text{BB}\}\right) &= \mathbb{P}\left(X_1, X_2, Y \leq \frac{1}{2}\right) \\
&= \int_0^{1/2} \int_0^{1/2} \int_0^{1/2} 1 \,\mathrm{d}x_1 \,\mathrm{d}x_2 \,\mathrm{d}x_3 \\
&= \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} \\
&= \frac{1}{8}.
\end{aligned}$$

$$\mathbb{P}(\{Y \leq t\}) = t^3.$$

$$f_Y(t) = \begin{cases} 3t^2, & t \in [0,1] \\ 0, & \text{otherwise}, \end{cases}$$

which is a $\text{Beta}(3,1)$ distribution.

$$\frac{1}{B(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1},\ x \in [0,1].$$

Lecture 12

28th October

Discussion on gamma function when $\alpha = 0$.

EXAMPLE 46

Suppose $X \sim \text{GAM}(\alpha, \lambda)$. Find $M_X(t)$.

Solution:

$$\begin{aligned}
M_X(t) &= \mathbb{E}[e^{tX}] \\
&= \int_0^\infty e^{tx} \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x} \,\mathrm{d}x \\
&= \frac{\lambda^\alpha}{\Gamma(\alpha)} \int_0^\infty x^{\alpha-1} e^{-\lambda x(1-t/\lambda)} \,\mathrm{d}x \\
&= \frac{\lambda^\alpha}{\Gamma(\alpha)} \int_0^\infty \frac{u^{\alpha-1}}{(\lambda-t)^{\alpha-1}} e^{-u} \frac{1}{\lambda-t} \,\mathrm{d}u \qquad u = x(\lambda-t) \iff \mathrm{d}u = (\lambda-t)\,\mathrm{d}x \\
&= \frac{\lambda^\alpha}{\Gamma(\alpha)(\lambda-t)^\alpha} \int_0^\infty u^{\alpha-1} e^{-u} \,\mathrm{d}u \\
&= \frac{\lambda^\alpha}{\Gamma(\alpha)(\lambda-t)^\alpha} \Gamma(\alpha) \\
&= \left(\frac{\lambda}{\lambda-t}\right)^\alpha.
\end{aligned}$$

EXAMPLE 47

Suppose $X_1 \sim \text{GAM}(1/2, 2)$ and $X_2 \sim \text{GAM}(3, 2)$ are independent. Find $M_Y(t)$ where $Y = X_1 + X_2$.

Solution: Since X_1 and X_2 are independent,

$$\begin{aligned} M_Y(t) &= M_{X_1}(t)M_{X_2}(t) \\ &= \left(\frac{2}{2-t}\right)^{1/2}\left(\frac{2}{2-t}\right)^{3} \\ &= \left(\frac{2}{2-t}\right)^{7/2}. \end{aligned}$$

Therefore, $Y \sim \text{GAM}(3.5, 2)$.

EXAMPLE 48

The pdf for $\text{GAM}(1, \lambda)$ is

$$f_X(t) = \frac{\lambda^1}{\Gamma(1)} t^0 e^{-\lambda t} = \lambda e^{-\lambda t},$$

which is $\text{EXP}(1)$.

REMARK 9

$$\text{BIN}\left(n, \frac{\lambda}{n}\right) \xrightarrow{n\to\infty} \text{POI}(\lambda).$$

EXAMPLE 49

Suppose Chocolat gets 1 customer every 10 minutes, on average (discrete time).

(i) Model level 1:

- Every minute there is an independent $1/10$ chance for a customer to enter (0 chance for multiple customers in the same minute).
- Let T_1 be the waiting time for the first customer in minutes,

$$T_1 = \text{waiting time for the first customer in minutes} \sim \text{GEO}\left(\frac{1}{10}\right),$$

and $\mathbb{E}[T_1] = 10$.

$$N_{60} = \text{number of customers in the first hour} \sim \text{BIN}\left(60, \frac{1}{10}\right).$$

(ii) Model level 2:

- Every second there is a $1/600$ chance for a customer to enter, independently.

$$T_1 = \text{waiting time in minutes} = \frac{\tilde{T}_1}{60}, \text{ where } \tilde{T}_1 \sim \text{GEO}\left(\frac{1}{600}\right),$$

and $\mathbb{E}[T_1] = 600/60 = 10$.

$$N_{60} = \text{number of customers in the first hour} \sim \text{BIN}\left(3600, \frac{1}{600}\right).$$

As we approach continuity,

$$N_{60} \xrightarrow{d} \text{POI}\left(\frac{60}{10}\right),\ T_1 \xrightarrow{d} \text{EXP}\left(\frac{1}{10}\right).$$

For $t \geq 0$,

$$N(t) = \text{number of arrivals in the first } t \text{ minutes} \sim \text{POI}\left(\frac{1}{10}t\right).$$

DEFINITION 40

A **Poisson process** ($N(t)$ for $t \geq 0$) with rate λ is a stochastic process with the properties:

(1) For $0 \leq t_1 < t_2$,

$$(N(t_2) - N(t_1)) \sim \text{POI}(\lambda(t_2 - t_1)).$$

(2) For $0 \leq t_1 < t_2 < \cdots < t_n$, the variables

$$(N(t_2) - N(t_1)), (N(t_3) - N(t_2)), \ldots (N(t_n) - N(t_{n-1}))$$

are jointly independent.

DEFINITION 41

$$T_n = \inf \{t \geq 0 : N(t) \geq n\}$$

is the arrival time of the n^{th} customer.

THEOREM 25: Interarrival Times

$\Delta_1 = T_1$, and $\Delta_n = T_n - T_{n-1}$ for $n \geq 2$ are known as ***interarrival times****. Then, $\Delta_1, \ldots, \Delta_n \overset{iid}{\sim}$ EXP(λ) variables.*

COROLLARY 1

For $0 \leq n_1 < n_2 < \cdots < n_k$,

$$T_{n_2} - T_{n_1}, T_{n_3} - T_{n_2}, \ldots, T_{n_k} - T_{n_{k-1}}$$

are jointly independent with respective probability distributions

$$(T_{n_{j+1}} - T_{n_j}) \sim \text{GAM}(n_{j+1} - n_j, \lambda).$$

EXAMPLE 50

$T_3 \sim \text{GAM}(3, \lambda)$ and $T_5 - T_3 \sim \text{GAM}(2, \lambda)$ are independent.

EXAMPLE 51

Suppose $X \sim \text{GAM}(\alpha, 1)$ and $Y \sim \text{GAM}(\beta, 1)$ are independent (rate doesn't matter, set it equal to 1 for simplicity).

EXAMPLE 52

$\alpha = 3$, $\beta = 5$, $X = T_3$, $Y = T_8 - T_3$. What is the distribution of T_3/T_8? If it took two hours for 8 people to arrive, what is the conditional distribution of how long it took for three people to arrive?

That is, find the distribution of

$$Z = \frac{X}{X + Y},\ 0 \leq Z \leq 1.$$

For $t \in [0,1]$,

$$\frac{x}{x+y} \leq t \implies x \leq \frac{ty}{1-t}.$$

Thus, noting that X and Y are independent,

$$\begin{aligned}\mathbb{P}(\{Z \leq t\}) &= \int_0^\infty \int_0^{ty/(1-t)} \frac{1}{\Gamma(\alpha)} x^{\alpha-1} e^{-x} \frac{1}{\Gamma(\beta)} y^{\beta-1} e^{-y} \, \mathrm{d}x \, \mathrm{d}y \\ &= \frac{1}{\Gamma(\alpha)\Gamma(\beta)} \int_0^\infty \int_0^{ty/(1-t)} x^{\alpha-1} y^{\beta-1} e^{-(x+y)} \, \mathrm{d}x \, \mathrm{d}y.\end{aligned}$$

Multivariable substitution:

$$u = \frac{x}{x+y}, \; v = x+y \implies x = uv, \; y = v - uv = v(1-u).$$

$$J = \frac{\partial(x,y)}{\partial(u,v)} = \begin{vmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \end{vmatrix} = \begin{vmatrix} v & u \\ -v & 1-u \end{vmatrix} = |(v)(1-u) - (u)(-v)| = v.$$

Note that $u \leq t$ and $0 \leq v < \infty$, which implies

$$\begin{aligned}&= \frac{1}{\Gamma(\alpha)\Gamma(\beta)} \int_0^\infty \int_0^1 (uv)^{\alpha-1} \big(v(1-u)\big)^{\beta-1} e^{-v} v \, \mathrm{d}u \, \mathrm{d}v \\ &= \frac{1}{\Gamma(\alpha)\Gamma(\beta)} \int_0^\infty \underbrace{v^{\alpha-1} v^{\beta-1}}_{v^{\alpha+\beta-1}} e^{-v} \, \mathrm{d}v \int_0^1 u^{\alpha-1}(1-u)^{\beta-1} \, \mathrm{d}u \\ &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \int_0^1 u^{\alpha-1}(1-u)^{\beta-1} \, \mathrm{d}u.\end{aligned}$$

DEFINITION 42: Beta Distribution

We say $X \sim \text{Beta}(\alpha, \beta)$ with shape parameters $0 < \alpha \in \mathbf{R}$ and $0 < \beta \in \mathbf{R}$ if it has pdf

$$f_X(t \mid \alpha, \beta) = \frac{1}{B(\alpha,\beta)} t^{\alpha-1}(1-t)^{\beta-1}, \; t \in [0,1]$$

where $B(\alpha, \beta)$ denotes the beta function,

$$B(\alpha,\beta) = \int_0^1 x^{\alpha-1}(1-x)^{\beta-1} \, \mathrm{d}x = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}.$$

THEOREM 26

If $X \sim GAM(\alpha, 1)$ and $Y \sim GAM(\beta, 1)$, then

$$Z = \frac{X}{X+Y} \sim \textit{Beta}(\alpha, \beta)$$

and is independent of

$$X + Y \sim GAM(\alpha+\beta, 1).$$

Lecture 13

2nd November

EXAMPLE 53

Suppose $X \sim \mathcal{N}(\mu, \sigma^2)$. Find $M_X(t)$.

Solution: Recall that

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}.$$

Hence,

$$\begin{aligned}
M_X(t) &= \int_{-\infty}^{\infty} \exp\{tx\} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} \mathrm{d}x \\
&= \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} \exp\left\{-\frac{1}{2\sigma^2}(x^2 - 2\mu x + \mu^2 - 2\sigma^2 tx)\right\} \mathrm{d}x \\
&= \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} \exp\left\{-\frac{1}{2\sigma^2}(x^2 - 2(\mu+\sigma^2 t)x + (\mu+\sigma^2 t)^2 - (\mu+\sigma^2 t)^2 + \mu^2)\right\} \mathrm{d}x \\
&= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{\frac{(\mu+\sigma^2 t)^2 - \mu^2}{2\sigma^2}\right\} \int_{-\infty}^{\infty} \exp\left\{-\frac{(x-(\mu+\sigma^2 t)^2)}{2\sigma^2}\right\} \mathrm{d}x \\
&= \exp\left\{\frac{2\mu\sigma^2 t + \sigma^4 t^2}{2\sigma^2}\right\} \\
&= \exp\left\{\mu t + \frac{\sigma^2 t^2}{2}\right\}.
\end{aligned}$$

EXAMPLE 54

If $X_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$, $X_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$ (independent), and $Y = X_1 + X_2$, then

$$\begin{aligned}
M_Y(t) &= M_{X_1}(t) M_{X_2}(t) \\
&= \exp\left\{\mu_1 t + \frac{\sigma_1^2 t^2}{2} + \mu_2 t + \frac{\sigma_2^2 t^2}{2}\right\} \\
&= \exp\left\{(\mu_1+\mu_2)t + \frac{(\sqrt{\sigma_1^2+\sigma_2^2})^2 t^2}{2}\right\}.
\end{aligned}$$

Therefore, $Y \sim \mathcal{N}(\mu_1+\mu_2, \sigma_1^2+\sigma_2^2)$,

Recall: If X is a continuous random variable with pdf f_X and g is a differentiable function $g\colon \mathbf{R} \to \mathbf{R}$ whose derivative only equals 0 at countably many points then the pdf of $Y = g(X)$ is

$$f_Y(y) = \sum_{x\colon\, g(x)=y,\; g'(x)\neq 0} \frac{f_X(x)}{|g'(x)|}.$$

EXAMPLE 55

Suppose $X \sim$ EXP(3) and $Y = 10X$. Find $f_Y(y)$.

Solution: Aside:

$$g(x) = 10x \implies g^{-1}(y) = \frac{x}{10} \implies g'(y) = 10.$$

Hence,

$$f_X(x) = 3e^{-3x},\ x \geq 0.$$
$$f_Y(y) = \frac{f_X(y/10)}{g'(y/10)} = \frac{3e^{-3y/10}}{10},\ y \geq 0.$$

EXAMPLE 56

Suppose $Z \sim \mathcal{N}(0,1)$ and $X = Z^2$.

Solution:

$$f_Z(t) = \phi(t) = \frac{1}{\sqrt{2\pi}} \exp\{-\frac{t^2}{2}\}.$$

Aside:

$$g(t) = t^2 \implies g'(t) = 2t.$$

Hence,

$$\begin{aligned} f_X(v) &= \sum_{t:\ t^2=v} \frac{\phi(t)}{|g'(t)|} \\ &= \frac{\phi(\sqrt{v})}{|g'(\sqrt{v})|} + \frac{\phi(-\sqrt{v})}{|g'(-\sqrt{v})|},\ v > 0 \\ &= \frac{1}{2\sqrt{v}} \frac{1}{\sqrt{2\pi}} e^{-v/2} + \frac{1}{2\sqrt{v}} \frac{1}{\sqrt{2\pi}} e^{-v/2} \\ &= \frac{1}{\sqrt{2\pi}} v^{-1/2} e^{-v/2}, \end{aligned}$$

which is GAM$(1/2, 1/2)$.

EXAMPLE 57

If $Y_1, Y_2, \ldots, Y_p \overset{\text{iid}}{\sim}$ GAM$(1/2, 1/2)$, then

$$Y_1 + \cdots + Y_p \sim \text{GAM}\left(\frac{p}{2}, \frac{1}{2}\right).$$

DEFINITION 43: Chi-squared

The Gamma distribution with shape $p/2$ and rate $1/2$ for any positive integer p is also called the **Chi-squared** distribution with p degrees of freedom, and we write $Y \sim \chi^2_p$. This is the distribution of the sum of squares of p independent standard normal variables. It has pdf

$$f(x) = \frac{(1/2)^{p/2}}{\Gamma(p/2)} x^{p/2-1} e^{-x/2}.$$

THEOREM 27: Markov's Inequality

If X is a non-negative random variable, then

$$\mathbb{P}(X \geq a) \leq \frac{\mathbb{E}[X]}{a}.$$

Proof: Suppose X is a non-negative random variable. Then,

$$\mathbb{E}[X] = \int_0^\infty t\,\mathbb{P}(X \in \mathrm{d}t).$$

Fix $a > 0$,

$$\begin{aligned}\mathbb{E}[X] &= \int_0^a t\,\mathbb{P}(X \in \mathrm{d}t) + \int_a^\infty t\,\mathbb{P}(X \in \mathrm{d}t) \\ &\geq \int_0^a 0\,\mathbb{P}(X \in \mathrm{d}t) + \int_a^\infty a\,\mathbb{P}(X \in \mathrm{d}t) \\ &= a\big(1 - F_X(a)\big) \\ &= a\,\mathbb{P}(X \geq a).\end{aligned}$$

Therefore,

$$\mathbb{E}[X] \geq a\,\mathbb{P}(X \geq a).$$

REMARK 10: Triangle Flip Trick

Suppose X is non-negative and discrete.

$$\begin{aligned}\mathbb{E}[X] &= 0\,\mathbb{P}(X = 0) + 1\,\mathbb{P}(X = 1) + 2\,\mathbb{P}(X = 2) + \cdots \\ &= \mathbb{P}(X = 1) + \mathbb{P}(X = 2) + \mathbb{P}(X = 2) + \mathbb{P}(X = 3) + \mathbb{P}(X = 3) + \mathbb{P}(X = 3) + \cdots \\ &= \mathbb{P}(X \geq 1) + \mathbb{P}(X \geq 2) + \mathbb{P}(X \geq 3) + \cdots \\ &= \sum_{k=1}^\infty \mathbb{P}(X \geq k).\end{aligned}$$

Suppose X is non-negative and continuous with pdf f_X.

$$\begin{aligned}\mathbb{E}[X] &= \int_0^\infty t f_X(t)\,\mathrm{d}t \\ &= \int_0^\infty f_X(t) \int_0^t 1\,\mathrm{d}s\,\mathrm{d}t \\ &= \int_0^\infty \int_s^\infty f_X(t)\,\mathrm{d}t\,\mathrm{d}s \\ &= \int_0^\infty 1 - F_X(t)\,\mathrm{d}s.\end{aligned}$$

THEOREM 28: Chebyshev's Inequality

For any random variable Y,

$$\mathbb{P}\Big(\big|Y - \mathbb{E}[Y]\big| \geq a\Big) \leq \frac{\mathrm{Var}(Y)}{a^2}.$$

Proof: Consider a random variable Y (does not have to be non-negative).

$$\begin{aligned}\mathbb{P}\big(|Y-\mathbb{E}[Y]|\ge a\big) &= \mathbb{P}\big((Y-\mathbb{E}[Y])^2\ge a^2\big)\\ &\le \frac{\mathbb{E}\big[(Y-\mathbb{E}[Y])^2\big]}{a^2} \text{by Markov's inequality}\\ &= \frac{\mathrm{Var}(Y)}{a^2}.\end{aligned}$$

EXAMPLE 58

If $X\sim \mathrm{EXP}(3)$, then $\mathbb{E}[X]=1/3$. Using Markov's inequality,

$$\mathbb{P}(X\ge 5)\le \frac{1/3}{5}=\frac{1}{15}.$$

$$\mathbb{P}(X\ge 5)\le \mathbb{P}\big(|X-\mathbb{E}[X]|\ge 14/3\big)\le \frac{\mathrm{Var}(X)}{(14/3)^2}=\frac{(1/3)^2}{(14/3)^2}=\frac{1}{14^2}=\frac{1}{196}.$$

Lecture 14

4th November

DEFINITION 44

If X is a discrete random variable and A is an event with $\mathbb{P}(A)>0$, then the **conditional pmf** of X given A is

$$p_{X|A}(k)=\frac{\mathbb{P}(\{X=k\}\cap A)}{\mathbb{P}(A)}.$$

This is another probability mass function:

- Non-negative (ratio of probabilities);
- Sums to 1:

$$\begin{aligned}\sum_k p_{X|A}(k) &= \frac{\mathbb{P}(\{X=k\}\cap A)}{\mathbb{P}(A)}\\ &= \frac{1}{\mathbb{P}(A)}\sum_k \mathbb{P}(\{X=k\}\cap A)\\ &= \frac{1}{\mathbb{P}(A)}\,\mathbb{P}\left(\bigcup_k(\{X=k\}\cap A)\right)\\ &= \frac{1}{\mathbb{P}(A)}\,\mathbb{P}\left(A\cap\bigcup_k\{X=k\}\right)\\ &= \frac{1}{\mathbb{P}(A)}\,\mathbb{P}(A\cap\Omega)\\ &= 1.\end{aligned}$$

DEFINITION 45

If Y is another discrete RV then we can define the conditional pmf of X given $Y=y$ in the same manner:

$$p_{X|Y}(k\mid y)=\frac{\mathbb{P}(\{X=k\}\cap\{Y=y\})}{\mathbb{P}(\{Y=y\})}.$$

If y is fixed, then this is a pmf over different values of k.

DEFINITION 46

The **conditional expectation** of X given $Y = y$ is:

$$\mathbb{E}[X \mid Y = y] = \sum_k k p_{X|Y}(k \mid y).$$

THEOREM 29

If $g\colon \mathbf{R}^2 \to \mathbf{R}$,

$$\mathbb{E}[g(X,Y) \mid Y = y] = \sum_k g(k,y) p_{X|Y}(k \mid y),$$

then

$$\sum_y \mathbb{E}[g(X,Y) \mid Y = y] p_Y(y) = \mathbb{E}[g(X,Y)].$$

That is,

$$\mathbb{E}[\mathbb{E}[g(X,Y) \mid Y]] = \mathbb{E}[g(X \mid Y)].$$

The expectation of the conditional expectation equals the expectation.

EXAMPLE 59

$$f_{X,Y}(x,y) = \begin{cases} 2x^2, & x \in [0,1],\ y \in [-x,x], \\ 0, & \text{otherwise.} \end{cases}$$

Recall: The marginal density for X is

$$f_X(x) = \int_{-\infty}^{\infty} f(x,y)\,\mathrm{d}y.$$

DEFINITION 47

In this setting, for $y \in \mathbf{R}$ with $f_Y(y) > 0$, the conditional pdf for X given $Y = y$ is

$$f_{X|Y} = \frac{f_{X,Y}(x,y)}{f_Y(y)}.$$

For y fixed, we can check if this is a pdf:

- Non-negative;
- Integrate to 1:

$$\begin{aligned} \int_{-\infty}^{\infty} f_{X|Y}(x \mid y)\,\mathrm{d}x &= \int_{-\infty}^{\infty} \frac{f_{X,Y}(x,y)}{f_Y(y)}\,\mathrm{d}x \\ &= \frac{\int_{-\infty}^{\infty} f_{X,Y}(x,y)\,\mathrm{d}x}{f_Y(y)} \\ &= \frac{f_Y(y)}{f_Y(y)} \\ &= 1. \end{aligned}$$

In our example, $-1 \le -x \le y \le x \le 1$, so

$$\begin{aligned} f_Y(y) &= \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,\mathrm{d}x \\ &= \int_{|y|}^{1} 2x^2\,\mathrm{d}x \\ &= \left[\frac{2}{3}x^3\right]_{x=|y|}^{x=1} \\ &= \frac{2}{3}(1-|y|^3),\ -1 \le y \le 1. \end{aligned}$$

Check:

$$\begin{aligned} \int_{-1}^{1} \frac{2}{3}(1-|y|^3)\,\mathrm{d}y &= \int_{-1}^{0} \frac{2}{3}(1+y^3)\,\mathrm{d}y + \int_{0}^{1} \frac{2}{3}(1-y^3)\,\mathrm{d}y \\ &= \left[\frac{2}{3}y + \frac{1}{6}y^4\right]_{y=-1}^{y=0} + \left[\frac{2}{3}y - \frac{1}{6}y^4\right]_{y=0}^{y=1} \\ &= -\left(-\frac{2}{3}+\frac{1}{6}\right) + \left(\frac{2}{3}-\frac{1}{6}\right) \\ &= \frac{1}{2}+\frac{1}{2} \\ &= 1. \end{aligned}$$

Thus,

$$f_{X|Y}(x \mid y) = \begin{cases} 3\dfrac{x^2}{1-|y|^3},\ y \le x \le 1, \\ 0, & \text{otherwise,} \end{cases} \qquad y \in [-1,1]$$

For example, if $y = -1/2$, then

$$f_{X|Y}(x \mid -1/2) = \begin{cases} \dfrac{24}{7}x^2,\ \dfrac{1}{2} \le x \le 1, \\ 0, & \text{otherwise.} \end{cases}$$

We can compute

$$\begin{aligned} \mathbb{E}[X \mid Y = y] &= \int_{-\infty}^{\infty} x f_{X|Y}(x \mid y)\,\mathrm{d}x \\ &= \int_{|y|}^{1} 3\frac{x^3}{1-|y|^3}\,\mathrm{d}x \\ &= \frac{3}{1-|y|^3}\int_{|y|}^{1} x^3\,\mathrm{d}x \\ &= \frac{3}{1-|y|^3}\left[\frac{1}{4}x^4\right]_{x=|y|}^{x=1} \\ &= \frac{3}{4(1-|y|^3)}(1-|y|^4). \end{aligned}$$

So,

$$\mathbb{E}\left[X \,\middle|\, Y = -\frac{1}{2}\right] = \frac{3}{4(7/8)}\frac{15}{16} = \frac{45}{56}.$$

If y was not fixed, we would have

$$\mathbb{E}[X \mid Y] = \frac{3}{4(1-|Y|^3)}(1-Y^4),$$

which is a random variable.

REMARK 11: Why do we care about $\mathbb{E}[X \mid Y]$?

$\mathbb{E}[X \mid Y]$ is the best guess for the value of X, based on Y, in the sense that it minimizes

$$\mathbb{E}\big[(X - \mathbb{E}[X \mid Y])^2\big].$$

That is, $\mathbb{E}[Y \mid X]$ in statistics is the **true regression function**.

$$\mathbb{E}[g(X,Y) \mid Y = y] = \int_{-\infty}^{\infty} g(x,y) f_{X|Y}(x \mid y)\,\mathrm{d}x.$$

Thus, $\mathbb{E}[g(X,Y) \mid Y]$ is the best guess for $g(X,Y)$, based on Y. It has the property that

$$\mathbb{E}[\mathbb{E}[g(X,Y) \mid Y]] = \mathbb{E}[g(X,Y)]$$

EXAMPLE 60

Suppose X is the first arrival time for a Poisson process with rate $\lambda = 2$, and Y is the second arrival time. So, the joint pdf is

$$f_{X,Y}(x,y) = \begin{cases} 4e^{-2y}, & 0 \le x < y, \\ 0, & \text{otherwise.} \end{cases}$$

Check:

$$\begin{aligned} \int_0^\infty \int_x^\infty 4e^{-2y}\,\mathrm{d}y\,\mathrm{d}x &= \int_0^\infty \big[-2e^{-2y}\big]_{y=x}^{y\to\infty}\,\mathrm{d}x \\ &= \int_0^\infty 2e^{-2x}\,\mathrm{d}x \\ &= \big[-e^{-2x}\big]_{x=0}^{x\to\infty} \\ &= 0 - (-1) \\ &= 1. \end{aligned}$$

The marginal pdf for Y is

$$\begin{aligned} f_Y(y) &= \int_0^\infty f_{X,Y}(x,y)\,\mathrm{d}x \\ &= \int_0^y 4e^{-2y}\,\mathrm{d}x \\ &= \big[x4e^{-2y}\big]_{x=0}^{x=y} \\ &= 4ye^{-2y}, \ y \ge 0, \end{aligned}$$

which is GAM$(2,2)$. The conditional pdf of X given Y is

$$f_{X|Y}(x \mid y) = \begin{cases} \dfrac{4e^{-2y}}{4ye^{-2y}}, & x \in [0,y], \\ 0, & \text{otherwise} \end{cases} = \begin{cases} \dfrac{1}{y}, & x \in [0,y], \\ 0, & \text{otherwise.} \end{cases}$$

For a fixed Y-value $Y = y$, X is conditionally Uniform$[0, y]$. That is,

$$\frac{X}{Y} \sim \text{Uniform}[0,1],$$

and is independent of Y.

Lecture 14

9th November

Suppose X and Y are jointly continuous with joint pdf $f_{X,Y} \colon \mathbf{R}^2 \to [0,\infty)$. Suppose $U = g_1(X,Y)$, $V = g_2(X,Y)$. Any region $S \subseteq \mathbf{R}^2$ for which $\mathbb{P}\big((X,Y) \in S\big)$ must have $\operatorname{Area}(S) > 0$, which fails in the example $(X, 1-X)$ or $(X, g(X))$ generally.

Let $A = \big\{(x,y) : f_{X,Y}(x,y) > 0\big\}$. Suppose g_1 and g_2 satisfy the property that $\forall S \subseteq A$, if $\operatorname{Area}(S) > 0$, then

$$\Big\{\big(g_1(x,y), g_2(x,y)\big) : (x,y) \in S\Big\}$$

has positive area. If there exists differentiable functions $h_1, h_2 \colon \mathbf{R}^2 \to \mathbf{R}$ such that

$$\begin{aligned} h_1\big(g_1(x,y), g_2(x,y)\big) &= x, \\ h_2\big(g_1(x,y), g_2(x,y)\big) &= y, \end{aligned}$$

then U, V have joint pdf

$$f_{U,V}(u,v) = f_{X,Y}\big(h_1(u,v), h_2(u,v)\big)\mathbf{J}$$

where

$$\mathbf{J} = \left|\frac{\partial(x,y)}{\partial(u,v)}\right| = \begin{vmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \end{vmatrix}.$$

EXAMPLE 61: Distribution of the Product of Beta Variables

Suppose $X \sim \text{Beta}(2,3)$ and $Y \sim \text{Beta}(5,9)$ are independent. Let $U = XY$ and $V = X$. Hence,

$$g_1(x,y) = xy, \quad g_2(x,y) = x,$$

and

$$h_1(u,v) = v, \quad h_2(x,y) = \frac{u}{v}.$$

Since the range of (X,Y) is $[0,1]^2$, the range of (U,V) is also $[0,1]^2$. Furthermore, $0 \le u \le v \le 1$. By independence,

$$\begin{aligned} f_{X,Y}(x,y) &= \frac{\Gamma(5)}{\Gamma(2)\Gamma(3)} x^{2-1}(1-x)^{3-1} \frac{\Gamma(14)}{\Gamma(5)\Gamma(9)} y^{5-1}(1-y)^{9-1} \\ &= \frac{\Gamma(14)}{\Gamma(2)\Gamma(3)\Gamma(9)} x(1-x)^2 y^4 (1-y)^8. \end{aligned}$$

For $0 \le u \le v \le 1$,

$$f_{U,V}(u,v) = \underbrace{\frac{\Gamma(14)}{\Gamma(2)\Gamma(3)\Gamma(9)}}_{C} v(1-v)^2 \left(\frac{u}{v}\right)^4 \left(1 - \frac{u}{v}\right)^8 \mathbf{J},$$

where

$$\mathbf{J} = \begin{vmatrix} 0 & 1 \\ 1/v & -u/v^2 \end{vmatrix} = \left|-\frac{1}{v}\right| = \frac{1}{v}.$$

The marginal pdf of U is

$$f_U(u) = \int_u^1 Cv(1-v)^2 \left(\frac{u}{v}\right)^4 \left(1 - \frac{u}{v}\right)^8 \frac{1}{v} \,\mathrm{d}v.$$

See textbook [Casella Example 4.3.3] for calculation, not done in class.
Start a Polya's urn with 2 red, 3 blue, and 9 green:

$$X = \lim \frac{\text{red}}{\text{red} + \text{blue}}, \quad Y = \lim \frac{\text{red} + \text{blue}}{\text{all}}.$$

Hence, $XY \sim \text{Beta}(2, 12)$.

EXAMPLE 62

Suppose $X \sim \mathcal{N}(1, 4)$ and $Y \sim \mathcal{N}(2, 1)$ are independent.

$$f_{X,Y}(x, y) = \frac{1}{2\pi\sqrt{4 \cdot 1}} \exp\left\{-\frac{(x-1)^2}{2 \cdot 4}\right\} \exp\left\{-\frac{(y-2)^2}{2 \cdot 1}\right\}.$$

Let $U = X + Y$ and $V = X - Y$. Are U and V independent? (U, V) can have any values in $\mathbf{R}^2$.

$$g_1(x, y) = xy, \quad g_2(x, y) = x - y,$$

and

$$h_1(u, v) = \frac{u+v}{2}, \quad h_2(u, v) = \frac{u-v}{2}.$$

The Jacobian is

$$\mathbf{J} = \begin{vmatrix} 1/2 & 1/2 \\ 1/2 & -1/2 \end{vmatrix} = \left| -\frac{1}{4} - \frac{1}{4} \right| = \frac{1}{2}.$$

Thus,

$$\begin{aligned}
f_{U,V}(u, v) &= \frac{1}{4\pi} \exp\left\{-\frac{(\frac{u+v}{2} - 1)^2}{8} - \frac{(\frac{u-v}{2} - 2)^2}{2}\right\} \frac{1}{2} \\
&= \frac{1}{8\pi} \exp\left\{-\frac{(u+v-2)^2}{32} - \frac{(u-v-4)^2}{8}\right\} \\
&= \frac{1}{8\pi} \exp\left\{-\frac{(u+v)^2}{32} - \frac{(u-v)^2}{8} + \frac{4(u+v)}{32} + \frac{8(u-v)}{8} - \frac{4}{32} - \frac{16}{8}\right\} \\
&= \frac{1}{8\pi} \exp\left\{-\frac{(u+v)^2}{32} - \frac{4(u-v)^2}{32} + \frac{4(u+v)}{32} + \frac{32(u-v)}{32} - \frac{4}{32} - \frac{64}{32}\right\} \\
&= \frac{1}{8\pi} \exp\left\{-\frac{2uv}{32} + \frac{8uv}{32} + u \text{ terms} + v \text{ terms} + \text{constant}\right\}.
\end{aligned}$$

We have uv terms, so U and V are not independent.

DEFINITION 48: Convolution

If X and Y are jointly continuous, $U = X + Y$, then the **convolution** is defined by

$$f_U(u) = \int_{-\infty}^{\infty} f_{X,Y}(t, u-t)\, \mathrm{d}t$$

EXAMPLE 63

Suppose X and Y are independent Uniform$[0, 1]$ where $U = X + Y$.

$$
\begin{aligned}
f_U(u) &= \int_{-\infty}^{\infty} f_{X,Y}(t, u-t)\,\mathrm{d}t \\
&= \begin{cases} \int_{u-1}^{1} f_{X,Y}(t, u-t)\,\mathrm{d}t & 0 < u < 1 \\ \int_0^u 1\,\mathrm{d}t & 1 \le u < 2 \end{cases} \\
&= \begin{cases} u & 0 < u < 1 \\ 2-u & 1 \le u < 2 \\ 0 & \text{otherwise.} \end{cases}
\end{aligned}
$$

Lecture 15

11th November

- A **sample** from a probability distribution is a sequence, independent, identically distributed (iid) variables with that distribution.
- A **sample with replacement** from a finite population (meaning a finite set S) is a sequence of iid random variables chosen from the uniform distribution on S.
- A **sample without replacement** from a finite population S is a sequence of random variables each chosen from the uniform distribution on S, but conditioned to all having distinct values.

For the rest of this lecture, we assume all samples are iid.

DEFINITION 49: Statistic

Given a sample $X_1, X_2, \ldots, X_n$, a **statistic** of the sample is a real- or vector-valued function $T(X_1, X_2, \ldots, X_n)$.

In our probabilistic model, a statistic is another random variable.

EXAMPLE 64

Some examples of statistics include:

- Order Statistics: highest value, 2^{nd} highest;
- Percentiles: 90^{th} percentile, median, 1^{st} quantile.

DEFINITION 50: Sample Mean (Average), Sample Variance, Sample Standard Deviation

Given a sample $X_1, \ldots, X_n$, the **sample mean** or **average** of the sample is

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

The **sample variance** is

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

The **sample standard deviation** is $S = \sqrt{S^2}$.

THEOREM 30

$$\sum_{i=1}^{n}(X_i - a)^2 = \arg\min_{a\in\mathbf{R}} \sum_{i=1}^{n}(X_i - a)^2.$$

Proof: Fix $a \in \mathbf{R}$,

$$\begin{aligned}
\sum_{i=1}^{n}(X_i - a)^2 &= \sum_{i=1}^{n}((X_i - \bar{X}) + (\bar{X} - a))^2 \\
&= \sum_{i=1}^{n}(X_i - \bar{X})^2 + 2\underbrace{\sum_{i=1}^{n}(X_i - \bar{X})(\bar{X} - a)}_{0} + \sum_{i=1}^{n}(\bar{X} - a)^2 \\
&= \sum_{i=1}^{n}(X_i - \bar{X})^2 + \sum_{i=1}^{n}(\bar{X} - a)^2 \\
&> \sum_{i=1}^{n}(X_i - \bar{X})^2,
\end{aligned}$$

unless $a = \bar{X}$, in which case they are equal. The middle term is 0 since

$$\begin{aligned}
\sum_{i=1}^{n}(X_i - \bar{X})(\bar{X} - a) &= (\bar{X} - a)\sum_{i=1}^{n}(X_i - \bar{X}) \\
&= (\bar{X} - a)(0) \\
&= 0.
\end{aligned}$$

THEOREM 31

$$(n-1)S^2 = \sum_{i=1}^{n} X_i^2 - n\bar{X}^2.$$

Proof: By the previous argument with $a = 0$,

$$\begin{aligned}
\sum_{i=1}^{n} X_i^2 &= \sum_{i=1}^{n}(X_i - \bar{X})^2 + \sum_{i=1}^{n}\bar{X}^2 \\
&= (n-1)S^2 + n\bar{X}^2.
\end{aligned}$$

THEOREM 32

Suppose $X_1, X_2, \ldots, X_n$ is an iid sample from a probability distribution with mean $\mu \in \mathbf{R}$ and variance $\sigma^2 < \infty$. Then,

(i) $\mathbb{E}[\bar{X}] = \mu$;

(ii) $\mathrm{Var}(\bar{X}) = \sigma^2/n$;

(iii) $\mathbb{E}[S^2] = \sigma^2$.

Proof:

(i) Easy: $\mathbb{E}[\bar{X}] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n} n\,\mathbb{E}[X_1] = \mu$.

(ii) Still easy:

$$
\begin{aligned}
\operatorname{Var}(\bar{X}) &= \operatorname{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) \\
&= \frac{1}{n^2}\operatorname{Var}\left(\sum_{i=1}^{n} X_i\right) \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\operatorname{Var}(X_i) && X_i \perp\!\!\!\perp X_j \text{ for } i \neq j \\
&= \frac{1}{n^2} n \operatorname{Var}(X_1) && X_i \text{ iid} \\
&= \frac{\sigma^2}{n}.
\end{aligned}
$$

(iii) Still easy, but long

$$
\begin{aligned}
\mathbb{E}[S^2] &= \mathbb{E}\left[\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2\right] \\
&= \frac{1}{n-1}\sum_{i=1}^{n}\mathbb{E}\big[(X_i - \bar{X})^2\big] \\
&= \frac{n}{n-1}\,\mathbb{E}[(X_1 - \bar{X})^2] \\
&= \frac{n}{n-1}\,\mathbb{E}\left[\left(X_1 - \sum_{i=1}^{n}\frac{X_i}{n}\right)^2\right] \\
&= \frac{n}{n-1}\left\{\mathbb{E}[X_1^2] - 2\,\mathbb{E}\left[X_1\sum_{i=1}^{n}\frac{X_i}{n}\right] + \mathbb{E}[\bar{X}^2]\right\} \\
&= \frac{n}{n-1}\left\{(\sigma^2+\mu^2) + \left(\frac{\sigma^2}{n}+\mu^2\right) - 2\left(\mathbb{E}\left[X_1\frac{X_1}{n}\right] + \mathbb{E}\left[X_1\sum_{i=2}^{n}\frac{X_i}{n}\right]\right)\right\} \\
&= \frac{n}{n-1}\left\{(\sigma^2+\mu^2) + \left(\frac{\sigma^2}{n}+\mu^2\right) - 2\left(\frac{1}{n}\,\mathbb{E}[X_1^2] + \frac{1}{n}\,\mathbb{E}[X_1X_2]\right)\right\} \\
&= \frac{n}{n-1}\left\{(\sigma^2+\mu^2) + \left(\frac{\sigma^2}{n}+\mu^2\right) - 2\left(\frac{1}{n}(\sigma^2+\mu^2) + \frac{n-1}{n}\,\mathbb{E}[X_1]\,\mathbb{E}[X_2]\right)\right\} \\
&= \frac{n}{n-1}\left\{(\sigma^2+\mu^2) + \left(\frac{\sigma^2}{n}+\mu^2\right) - 2\left(\frac{1}{n}(\sigma^2+\mu^2) + \frac{n-1}{n}\mu^2\right)\right\} \\
&= \frac{n}{n-1}\left\{\sigma^2+\mu^2+\frac{\sigma^2}{n}+\mu^2-\frac{2}{n}\sigma^2-\frac{2}{n}\mu^2-2\frac{n-1}{n}\mu^2\right\} \\
&= \frac{n}{n-1}\left\{\frac{(n-1)\sigma^2}{n}\right\} \\
&= \sigma^2.
\end{aligned}
$$

DEFINITION 51: Exponential Family

A family of pdfs (continuous) or pmfs (discrete) form an **exponential family** if it has the form

$$f(x \mid \boldsymbol{\theta}) = h(x)c(\boldsymbol{\theta}) \exp\left\{\sum_{i=1}^{k} w_i(\boldsymbol{\theta})t_i(x)\right\},$$

where $h(x) \geq 0$, $c(\boldsymbol{\theta}) \geq 0$. All real-valued functions h and $t_1, \ldots, t_k$ cannot depend on θ; c and $w_1, \ldots, w_k$ cannot depend on x.

EXAMPLE 65

For n fixed, the family of Binomial distributions BIN(n, p) for $0 < p < 1$ form an exponential family.

Solution: First,

$$f(j \mid p) = \binom{n}{j} p^j (1-p)^{n-j} = \binom{n}{j} \left(\frac{p}{1-p}\right)^j (1-p)^n,$$

where we define

$$h(j) = \begin{cases} \binom{n}{j}, & 0 \leq j \leq n, \\ 0, & \text{otherwise}, \end{cases}$$

$$c(p) = \begin{cases} (1-p)^n, & 0 < p < 1 \\ 0, & \text{otherwise}. \end{cases}$$

We want

$$\left(\frac{p}{1-p}\right)^j = \exp\{w_1(p)t_1(j)\},$$

so if we set $t_1(j) = j$, we get

$$\exp\{w_1(p)j\} = \left(\frac{p}{1-p}\right)^j \implies w_1(p)j = j \ln\left(\frac{p}{1-p}\right) \implies w_1(p) = \ln\left(\frac{p}{1-p}\right).$$

Therefore,

$$f(j \mid p) = h(j)c(p) \exp\{w_1(p)t_1(j)\} = \binom{n}{j}(1-p)^n \exp\left\{\ln\left(\frac{p}{1-p}\right)j\right\}.$$

EXAMPLE 66

The normal distribution $\mathcal{N}(\mu, \sigma^2)$ form an exponential family for $\mu \in \mathbf{R}$, $0 < \sigma^2 < \infty$.

Solution: First,

$$\begin{aligned} f(x \mid \mu, \sigma^2) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{x^2}{2\sigma^2} - \frac{\mu^2}{2\sigma^2} + \frac{\mu x}{\sigma^2}\right\} \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{\mu^2}{2\sigma^2}\right\} \exp\left\{-\frac{x^2}{2\sigma^2} + \frac{\mu x}{\sigma^2}\right\}. \end{aligned}$$

Define the following functions:

$$\begin{aligned} h(x) &= 1, \\ c(\mu, \sigma^2) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{\mu^2}{2\sigma^2}\right\}, \\ w_1(\mu, \sigma^2) &= -\frac{1}{2\sigma^2}, & t_1(x) &= x^2, \\ w_2(\mu, \sigma^2) &= \frac{\mu}{\sigma^2}, & t_2(x) &= x. \end{aligned}$$

We try to fit this representation with as few w_i's and t_i's as possible. If the number of terms in the sum k (number of w_i's and t_i's) equals the number of parameters for the family of distributions, then this is a **full exponential family**.
If k is greater than the number of parameters, then this is a **curved exponential family**.

THEOREM 33

If X is a random variable whose distribution comes from an exponential family,

$$f(x \mid \boldsymbol{\theta}) = h(x)c(\boldsymbol{\theta}) \exp\left\{\sum_{i=1}^{k} w_i(\boldsymbol{\theta}) t_i(x)\right\},$$

then for any parameter θ,

(i) $\mathbb{E}\left[\sum_{i=1}^{k} \frac{\partial w_i(\boldsymbol{\theta})}{\partial \theta_j} t_i(X)\right] = -\frac{\partial}{\partial \theta_j} \ln(c(\boldsymbol{\theta}))$*;*

(ii) $\operatorname{Var}\left(\sum_{i=1}^{k} \frac{\partial w_i(\boldsymbol{\theta})}{\partial \theta_j} t_i(X)\right) = -\frac{\partial^2}{\partial \theta_j^2} \ln(c(\boldsymbol{\theta})) - \mathbb{E}\left[\sum_{i=1}^{k} \frac{\partial^2 w_i(\boldsymbol{\theta})}{\partial \theta_j^2} t_i(X)\right].$

EXAMPLE 67

Let $X \sim \text{BIN}(n, p)$. From Example 65, we know that

$$\begin{aligned} h(j) &= \begin{cases} \binom{n}{j}, & 0 \le j \le n, \\ 0, & \text{otherwise}, \end{cases} \\ c(p) &= \begin{cases} (1-p)^n, & 0 < p < 1 \\ 0, & \text{otherwise}, \end{cases} \\ t_1(j) &= j, \\ w_1(p) &= \ln\left(\frac{p}{1-p}\right). \end{aligned}$$

To use $\mathbb{E}\left[\sum_{i=1}^{k} \frac{\partial w_i(p)}{\partial p} t_i(X)\right] = -\frac{\partial}{\partial p} \ln(c(p))$, we compute

$$\begin{aligned} \frac{\partial w_i(p)}{\partial p} &= \frac{\partial}{\partial p} \ln\left(\frac{p}{1-p}\right) = \frac{1}{p/(1-p)} \frac{(1-p)\cdot 1 - p(-1)}{(1-p)^2} = \frac{1}{p(1-p)}, \\ -\frac{\partial \ln(c(p))}{\partial p} &= -\frac{\partial}{\partial p} \ln((1-p)^n) = -n\frac{\partial}{\partial p} \ln(1-p) = -n\frac{1}{1-p}(-1) = \frac{n}{1-p}. \end{aligned}$$

Hence,

$$\mathbb{E}\left[\frac{1}{p(1-p)} X\right] = \frac{n}{1-p} \implies \mathbb{E}[X] = np.$$

THEOREM 34

Suppose $X_1, \ldots, X_n$ *are iid samples from a distribution that comes from an exponential family,*

$$f(x \mid \boldsymbol{\theta}) = h(x)c(\boldsymbol{\theta}) \exp\left\{\sum_{i=1}^{k} w_i(\boldsymbol{\theta}) t_i(x)\right\}.$$

Define statistics $T_1, T_2, \ldots, T_k$ *by*

$$T_i(X_1, \ldots, X_n) = \sum_{j=1}^{n} t_i(X_j),\ 1 \le i \le k.$$

If the set

$$\big\{(w_1(\boldsymbol{\theta}), w_2(\boldsymbol{\theta}), \ldots, w_k(\boldsymbol{\theta})) : \boldsymbol{\theta} \text{ is an allowed value for the parameter}\big\}$$

contains an open subset of $\mathbf{R}^k$ *(usually true for full exponential families), then the distribution of the vector* $(T_1, \ldots, T_k) = \boldsymbol{T}$ *is itself an exponential family of the form*

$$f_{\boldsymbol{T}}(u_1, \ldots, u_k \mid \boldsymbol{\theta}) = H(u_1, \ldots, u_k)c(\boldsymbol{\theta})^n \exp\left\{\sum_{i=1}^{k} w_i(\boldsymbol{\theta}) u_i\right\}.$$

LECTURE 16

16th November

DEFINITION 52: Order Statistic

Given a sample $X_1, X_2, \ldots, X_n$, let $X_{(1)}$ denote the lowest value in the sample,

$$X_{(j)} = \min\{x \in \mathbf{R} : \big|\{i \in [n] : X_i \le x\}\big| \ge j\},\ 1 \le j \le n.$$

So $X_{(1)} \le X_{(2)} \le \cdots \le X_{(n)}$ is a decreasing re-ordering of our sample. We call $X_{(j)}$ the j^{th} **order statistic** of the sample.

EXAMPLE 68

If our sample is

X_1	X_2	X_3	X_4	X_5	X_6	X_7
5	3	6	2	9	1	2

Then,

$X_{(1)}$	$X_{(2)}$	$X_{(3)}$	$X_{(4)}$	$X_{(5)}$	$X_{(6)}$	$X_{(7)}$
1	2	2	3	5	6	9

DEFINITION 53: Median

The **median** of a sample of size n is

$$\begin{cases} X_{((n+1)/2)}, & n \text{ odd}, \\ \dfrac{X_{(n/2)} + X_{((n/2)+1)}}{2}, & n \text{ even}. \end{cases}$$

DEFINITION 54: Percentile

For $\frac{1}{2n} < p < 1 - \frac{1}{2n}$, we can define the p^{th} **percentile** as $X_{([np])}$, where $[x]$ denotes rounding to the nearest integer.

If we wanted to be more precise, we could define the p^{th} **percentile** as

$$(\lfloor np \rfloor + 1 - np) X_{(\lfloor np \rfloor)} + (np - \lfloor np \rfloor) X_{(\lfloor np \rfloor + 1)},$$

where $\lfloor x \rfloor$ is the floor of x.

THEOREM 35

Suppose $X_1, \ldots, X_n$ is a sample from a discrete distribution with possible values $a_1 < a_2 < a_3 < \cdots$. Define $p_i = \mathbb{P}\{X = a_i\}$ and $P_i = \sum_{j=1}^{i} p_i = \mathbb{P}\{X \le a_i\}$. Then,

$$\mathbb{P}\{X_{(j)} \le a_i\} = \sum_{m=j}^{n} \binom{n}{m} P_i^m (1 - P_i)^{n-m}.$$

$$\mathbb{P}\{X_{(j)} = a_i\} = \sum_{m=j}^{n} \binom{n}{m} \left[P_i^m (1 - P_i)^{n-m} - P_{i-1}^m (1 - P_{i-1})^{n-m} \right].$$

Proof: For $1 \le k \le n$, let

$$I_k = \begin{cases} 1, & X_k \le a_i, \\ 0, & \text{otherwise.} \end{cases}$$

Since the X_k are independent, the $(I_k,\ k = 1, 2, \ldots, n)$ are independent. Thus,

$$S = \sum_{k=1}^{n} I_k \sim \text{BIN}(n, q),$$

where $q = \mathbb{P}\{X_1 \le a_i\} = P_i$. Hence,

$$\begin{aligned} \mathbb{P}\{X_{(j)} \le a_i\} &= \mathbb{P}\{S \ge j\} \\ &= \sum_{m=j}^{n} \mathbb{P}\{S = m\} \\ &= \sum_{m=j}^{n} \binom{n}{m} P_i^m (1 - P_i)^{n-m}. \end{aligned}$$

The second formula is

$$\mathbb{P}\{X_{(j)} = a_i\} = \mathbb{P}\{X_{(j)} \le a_i\} - \mathbb{P}\{X_{(j)} \le a_{i-1}\}.$$

EXAMPLE 69

Suppose $G_1, \ldots, G_9 \overset{\text{iid}}{\sim} \text{GEO}(1/6)$.

$$p_i = \left(\frac{5}{6}\right)^{i-1} \frac{1}{6} = \mathbb{P}\{G = 1\},$$

$$P_i = \mathbb{P}\{G \le i\} = 1 - \mathbb{P}\{G > i\} = 1 - \left(\frac{5}{6}\right)^{i}.$$

Median:

$$\begin{aligned}\mathbb{P}\{G_{(5)} = k\} &= \sum_{m=5}^{9} \binom{9}{m} \{P_k^m (1-P_k)^{9-m} - P_{i-1}^m (1-P_k)^{9-m}\} \\ &= \sum_{m=5}^{9} \binom{9}{m} \left\{ \left[1 - \left(\frac{5}{6}\right)^k\right]^m \left(\frac{5}{6}\right)^{k(9-m)} - \left[1 - \left(\frac{5}{6}\right)^{k-1}\right]^m \left(\frac{5}{6}\right)^{(k-1)(9-m)} \right\}\end{aligned}$$

THEOREM 36

For any sample of size n, from any discrete distribution, for any possible value of a of the variables,

$$\mathbb{P}\{X_{(j)} = a\} = \sum_{m=j}^{n} \binom{n}{m} \{\mathbb{P}\{X \le a\}^m \, \mathbb{P}\{X > a\}^{n-m} - \mathbb{P}\{X < a\}^m \, \mathbb{P}\{X \ge a\}^{n-m}\}.$$

THEOREM 37

Suppose $X_1, \ldots, X_n$ is a sample from a continuous distribution on $\mathbf{R}$ with pdf f and cdf of F. Then,

$$\mathbb{P}\{X_{(j)} \le t\} = \sum_{m=j}^{n} \binom{n}{m} F(t)^m (1 - F(t))^{n-m}$$

is the cdf of $X_{(j)}$. The pdf of $X_{(j)}$ is

$$\begin{aligned}\mathbb{P}(X_{(j)} \in \mathrm{d}t) &= f_{X_{(j)}}(t)\,\mathrm{d}t \\ &= \binom{n}{j-1, 1, n-j} F(t)^{j-1} (1 - F(t))^{n-j} f(t)\,\mathrm{d}t,\end{aligned}$$

noting that

$$\binom{n}{j-1, 1, n-j} = j\binom{n}{j}.$$

Proof: The argument for the first formula is the same as in the discrete case. To get the second, we will differentiate. Define $g_m(x) = x^m (1-x)^{n-m}$, so

$$\begin{aligned}g_m'(x) &= m x^{m-1} (1-x)^{n-m} + x^m (n-m)(1-x)^{n-m-1}(-1) \\ &= \big(m(1-x) - (n-m)x\big) x^{m-1} (1-x)^{n-m-1} \\ &= (m - nx) x^{m-1} (1-x)^{n-m-1}.\end{aligned}$$

Also,

$$\begin{aligned}f_{X_{(j)}}(t) &= \frac{\mathrm{d}}{\mathrm{d}t} F_X(t) \\ &= \sum_{m=j}^{n} \frac{\mathrm{d}}{\mathrm{d}t} \binom{n}{m} g_m(F(t)) \\ &= \sum_{m=j}^{n} \binom{n}{m} (m - nF(t)) F(t)^{m-1} (1 - F(t))^{n-m-1} f(t),\end{aligned}$$

RIP.

EXAMPLE 70

Suppose $U_1, \ldots, U_n \overset{\text{iid}}{\sim} \mathcal{U}[0,1]$.

$$\begin{aligned} f_{U_{(j)}}(t) &= j\binom{n}{j} t^{j-1}(1-t)^{n-j} \cdot 1 \\ &= \frac{n!}{(j-1)!(n-j)!} t^{j-1}(1-t)^{n-j}. \end{aligned}$$

That is, $U_{(j)} \sim \text{Beta}(j, n+1-j)$. Also,

$$\mathbb{E}[U_{(j)}] = \frac{j}{j+n+1-j} = \frac{j}{n+1}.$$

Lecture 17

18th November

DEFINITION 55: Convergence in Probability

Given a sequence of random variables $X_1, X_2 \ldots$, and a random variable Y, we say the sequence **converges in probability** to Y, denoted

$$X_n \xrightarrow{p} Y$$

if

$$\forall \varepsilon > 0, \lim_{n\to\infty} \mathbb{P}\{|Y - X_n| \geq \varepsilon\} = 0.$$

THEOREM 38: Weak Law of Large Numbers (WLLN)

If $X_1, X_2, \ldots$ is a sequence of independent random variables with

$$\text{Var}(X_n) \leq \sigma^2 < \infty, \; \forall n,$$

which implies all X_n have finite expectation, then

$$\frac{S_n - \mathbb{E}[S_n]}{n} \xrightarrow{p} 0, \; \forall n,$$

where $S_n = \sum_{j=1}^n X_j$.

Proof: Let $\varepsilon > 0$.

$$\begin{aligned} \mathbb{P}\left(\left|\frac{S_n - \mathbb{E}[S_n]}{n}\right| \geq \varepsilon\right) &= \mathbb{P}\left(\left|\frac{S_n}{n} - \mathbb{E}\left[\frac{S_n}{n}\right]\right| \geq \varepsilon\right) \\ &\leq \frac{\text{Var}\left(\frac{S_n}{n}\right)}{\varepsilon^2} \\ &\leq \frac{\frac{1}{n^2}\text{Var}(S_n)}{\varepsilon^2} \\ &\leq \frac{(n\sigma^2)/n^2}{\varepsilon^2} \\ &= \frac{\sigma^2}{n\varepsilon^2} \\ &\xrightarrow{n\to\infty} 0. \end{aligned}$$

COROLLARY 2: Weak Law of Large Numbers

If the $\{X_n\}_{n\geq 1}$ *are iid, then*

$$\frac{S_n}{n} \xrightarrow{p} \mathbb{E}[X_1].$$

THEOREM 39

Fix $0 < q < p < \infty$. *For a random variable* X, *if* $\mathbb{E}\big[|X|^p\big] < \infty$, *then* $\mathbb{E}\big[|X|^q\big] < \infty$.

Proof: Suppose $\mathbb{E}\big[|X|^p\big] < \infty$.

$$\begin{aligned}\mathbb{E}\big[|X|^q\big] &= \mathbb{E}\Big[|X|^q\,\mathbb{I}\big\{|X| < 1\big\}\Big] + \mathbb{E}\Big[|X|^q\,\mathbb{I}\big\{|X| \geq 1\big\}\Big] \\ &\leq \mathbb{P}\big\{|X| < 1\big\} + \mathbb{E}\Big[|X|^p\,\mathbb{I}\big\{|X| \geq 1\big\}\Big] \\ &< \infty.\end{aligned}$$

THEOREM 40

Suppose $X_1, X_2, \ldots, X_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$ *and*

$$S_n = \sqrt{\frac{1}{n-1}\sum_{j=1}^{n}(X_j - \bar{X})^2}.$$

Then,

$$(n-1)\frac{S_n^2}{\sigma^2} \sim \chi^2(n-1) = GAM\left(\frac{n-1}{2}, \frac{1}{2}\right).$$

Proof: Casella Section 5.3.

EXAMPLE 71

By Theorem 40,

$$\operatorname{Var}\left((n-1)\frac{S_n^2}{\sigma^2}\right) = 2n-2,$$

hence

$$\operatorname{Var}(S_n^2) = \frac{\sigma^4(2n-2)}{(n-1)^2} = \frac{2\sigma^4}{n-1}.$$

Using Chebyshev's inequality,

$$\mathbb{P}\big\{|S_n^2 - \sigma^2| \geq \varepsilon\big\} \leq \frac{2\sigma^4/(n-1)}{\varepsilon^2} \xrightarrow{n\to\infty} 0.$$

Hence,

$$S_n^2 \xrightarrow{p} \sigma^2.$$

THEOREM 41

For any continuous function $g\colon \mathbf{R} \to \mathbf{R}$, *if*

$$X_n \xrightarrow{p} Y,$$

then

$$g(X_n) \xrightarrow{p} g(Y).$$

COROLLARY 3

$S_n \xrightarrow{p} \sigma^2$ for sample standard deviation of $\mathcal{N}(\mu, \sigma^2)$ samples.

Proof: Take square roots.

DEFINITION 56: Almost Sure Convergence

Given a sequence of random variables $X_1, X_2, \ldots$, and a random variable Y, we say the sequence **converges almost surely** (a.s.) to Y, denoted

$$X_n \xrightarrow{a.s.} Y$$

if

$$\forall \varepsilon > 0,\ \mathbb{P}\Big\{\lim_{n\to\infty} |Y - X_n| \geq \varepsilon\Big\} = 0.$$

Equivalently,

$$\mathbb{P}\Big\{\lim_{n\to\infty} |Y - X_n| = 0\Big\} = 1.$$

THEOREM 42: Almost Sure Convergence $\implies$ Convergence in Probability

If $X_n \xrightarrow{a.s.} Y$, then $X_n \xrightarrow{p} Y$.

Proof: Assume $X_n \xrightarrow{a.s.} Y$. Fix $\varepsilon > 0$.

$$N = \max\big\{\{n \in \mathbf{N} : \forall m > n,\ |Y - X_n| \leq \varepsilon\} \cup \{1\}\big\}$$

(the last time that $Y - X_n > \varepsilon$). Since $X_n \xrightarrow{a.s.} Y$, N is a.s. finite so $\mathbb{P}\{N > n\} \xrightarrow{n\to\infty} 0$.

$$\mathbb{P}\big(|Y - X_n| \geq \varepsilon\big) \leq \mathbb{P}\{n \leq N\} \xrightarrow{n\to\infty} 0.$$

LEMMA 1: Borel-Cantelli Lemma

For a sequence of events $A_1, A_2, \ldots$, if

$$\sum_{n=1}^{\infty} \mathbb{P}(A_n) < \infty,$$

then

$$\mathbb{P}(\textit{infinitely many of the } A_n \textit{ happen}) = \mathbb{P}\left(\bigcap_{n=1}^{\infty} \bigcup_{j=n}^{\infty} A_j\right) = 0.$$

If the $(A_n,\ n \geq 1)$ are independent, then the converse of this is true.

EXAMPLE 72: Convergence in Probability $\not\Longrightarrow$ Almost Sure Convergence

Suppose for all $n \geq 1$, $X_n \overset{\text{iid}}{\sim} \text{BERN}(1/n)$. Note that $X_n \xrightarrow{p} 0$, but

$$\sum_{n=1}^{\infty} \mathbb{P}\{X_n = 1\} = \sum_{n=1}^{\infty} \frac{1}{n} = \infty.$$

By the Borel-Cantelli lemma, there are a.s. finitely many n for which $X_n = 1$.

Lecture 17

30th November

DEFINITION 57

We say a sequence of events $(A_n)_{n\geq 1}$ happens **infinitely often** on an outcome ω if for all N, there exists $n > N$ such that $\omega \in A_n$ where

$$\{(A_n)_{n\geq 1} \text{ i.o}\} = \bigcap_{N\geq 1} \bigcup_{n\geq N} .$$

THEOREM 43: Borel-Cantelli Lemma

If $\sum_{n=1}^{\infty} \mathbb{P}(A_n) < \infty$, then

$$\mathbb{P}\{(A_n)_{n\geq 1} \text{ i.o}\} = 0.$$

Proof: Let $Y = \sum_{n=1}^{\infty} \mathbb{I}\{A_n\}$, so $Y \in \mathbf{N} \cup \{\infty\}$.

$$\begin{aligned}
\mathbb{E}[Y] &= \mathbb{E}\left[\sum_{n=1}^{\infty} \mathbb{I}\{A_n\}\right] \\
&= \sum_{n=1}^{\infty} \mathbb{E}[\mathbb{I}\{A_n\}] \\
&= \sum_{n=1}^{\infty} (1\,\mathbb{P}(A_n) + 0\,\mathbb{P}(A_n^c)) \\
&= \sum_{n=1}^{\infty} \mathbb{P}(A_n) \\
&< \infty.
\end{aligned}$$

Thus, $\mathbb{P}\{Y = \infty\} = 0$. Alternatively,

$$\mathbb{P}\{Y > n\} \leq \frac{\mathbb{E}[Y]}{n} \xrightarrow{n\to\infty} 0,$$

so $\mathbb{P}\{Y = \infty\} = 0$.

COROLLARY 4

If the events A_n are independent, then the converse of Theorem 43 is also true.

Proof: Suppose the $(A_n)_{n\geq 1}$ are independent and that $\sum_{n=1}^{\infty} \mathbb{P}(A_n) = \infty$ and we will show $\mathbb{P}\{(A_n) \text{ i.o.}\} = 1$. For all $N \in \mathbf{N}$,

$$\begin{aligned}
\mathbb{P}\left(\bigcup_{n\geq N} A_n\right) &= 1 - \mathbb{P}\left(\bigcup_{n\geq N}^{\infty} A_n^c\right) \\
&= 1 - \prod_{n\geq N} (1 - \mathbb{P}(A_n)) \\
&\geq 1 - \prod_{n\geq N} e^{-\mathbb{P}(A_n)} \\
&= 1 - e^{-\sum_{n\geq N} \mathbb{P}(A_n)} \\
&= 1 - e^{-\infty} \\
&= 1.
\end{aligned}$$

EXAMPLE 73: Convergence in Probability $\not\Longrightarrow$ Almost Sure Convergence

For all $n \geq 1$, $X_n \overset{\text{iid}}{\sim}$ BERN$(1/n)$. Then,

$$\mathbb{P}\{|X_n - 0| > \varepsilon\} = \frac{1}{n} \to 0.$$

But,

$$\sum_{n=1}^{\infty} \mathbb{P}\{X_n = 1\} = \sum_{n=1}^{\infty} \frac{1}{n} = \infty.$$

Let $Y_n = nX_n$, so

$$Y_n = \begin{cases} n, & \text{w.p. } \frac{1}{n}, \\ 0, & \text{w.p. } 1 - \frac{1}{n}. \end{cases}$$

$\mathbb{E}[Y_n] = 1$ for every n,

$$Y_n \xrightarrow{p} 0,$$

but $\mathbb{E}[Y_n] \to 1$, so $Y_n \overset{a.s.}{\not\longrightarrow} 0$.

LEMMA 2: Kronecker's Lemma

For a sequence $(X_n)_{n\geq 1} \in (0, \infty)^{\mathbf{N}}$, *if* $\sum_{n=1}^{\infty} \frac{X_n}{n} < \infty$, *then* $\lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} X_n = 0$.

Proof: Suppose $S = \sum_{n=1}^{\infty} \frac{X_n}{n} < \infty$, then

$$\sum_{n=1}^{N} \frac{X_n}{n} - \sum_{n=1}^{N} \frac{X_n}{N} = \sum_{n=1}^{N} \frac{X_n}{n}\left(1 - \frac{n}{N}\right) \xrightarrow{n\to\infty} S.$$

Fix $\varepsilon > 0$. Let N_1 be sufficiently large such that

$$\sum_{n=N_1}^{\infty} \frac{X_n}{n} < \frac{\varepsilon}{2},$$

and let $N_2 > N_1$ be sufficiently large so that

$$\frac{N_1}{N_2} < \frac{\varepsilon}{2S},$$

that is, $N_2 = \lceil \frac{2N_1 S}{\varepsilon} \rceil$. Then,

$$\begin{aligned}
\sum_{n=1}^{N_2} \frac{X_n}{n}\left(1 - \frac{n}{N_2}\right) &\geq \sum_{n=1}^{N_1} \frac{X_n}{n}\left(1 - \frac{N_1}{N_2}\right) \\
&\geq \left(1 - \frac{\varepsilon}{2S}\right) \sum_{n=1}^{N_1} \frac{X_n}{n} \\
&\geq \left(1 - \frac{\varepsilon}{2S}\right)\left(S - \frac{\varepsilon}{2}\right) \\
&= S - S\frac{\varepsilon}{2S} - \frac{\varepsilon}{2} + \frac{\varepsilon^2}{4S} \\
&\geq S - \frac{\varepsilon}{2} - \frac{\varepsilon}{2} \\
&= S - \varepsilon.
\end{aligned}$$

We conclude that

$$\sum_{n=1}^{N} \frac{X_n}{n} - \sum_{n=1}^{N} \frac{X_n}{N} \xrightarrow{n\to\infty} S.$$

Therefore,

$$\sum_{n=1}^{N} \frac{X_n}{N} \xrightarrow{N\to\infty} S - S = 0.$$

THEOREM 44: Strong Law of Large Numbers

If $X_1, X_2, \ldots$ is a sequence of IID random variables with $\mathbb{E}[|X_i|] < \infty$, then

$$\frac{1}{n}\sum_{j=1}^{n} X_j \xrightarrow{a.s.} \mathbb{E}[X_1]$$

as $n \to \infty$.

First Step of Proof: Let $Y_n = X_n \, \mathbb{I}\{|X_n| \le n\}$.

Lecture 18

2nd December

THEOREM 45: Kolmogorov's Inequality

Suppose $X_1, \ldots, X_n$ are independent random variables with finite expectation. For $1 \le j \le n$, let $S_j = X_1 + \cdots + X_j$. Then for any $\varepsilon > 0$,

$$\mathbb{P}\left(\max_{1\le j\le n} \left|S_j - \mathbb{E}[S_j]\right| \ge \varepsilon\right) \le \frac{\operatorname{Var}(S_n)}{\varepsilon^2}.$$

Proof: Assume WLOG, $\mathbb{E}[X_j] = 0$ for $j = 1, \ldots, n$, so $\mathbb{E}[S_j] = 0$ as well. Let

$$A_j = \begin{cases} |S_j| < \varepsilon, & 1 \le j < k, \\ |S_k| \ge \varepsilon, & \text{otherwise.} \end{cases}$$

$$A = \bigcup_{k=1}^{n} A_k = \{\max_{1\le j\le n} |S_j| \ge \varepsilon\}.$$

Let $\mathbb{I}\{A\} = 1$ if A happens, and $\mathbb{I}\{A\} = 0$ otherwise. Now,

$$\operatorname{Var}(S_n) = \mathbb{E}[S_n^2] \ge \mathbb{E}[S_n^2 \, \mathbb{I}\{A\}] = \mathbb{E}\left[S_n^2\left(\sum_{k=1}^{n} \mathbb{I}\{A_k\}\right)\right] = \sum_{k=1}^{n} \mathbb{E}[S_n^2 \, \mathbb{I}\{A_k\}].$$

For $1 \le k \le n$, define $Y_k = X_{k+1} + X_{k+2} + \cdots + X_n$ so that

$$S_n = S_k + Y_k.$$

$$\begin{aligned}
\mathbb{E}[S_n^2\,\mathbb{I}\{A_k\}] &= \mathbb{E}\big[(S_k+Y_k)^2\,\mathbb{I}\{A_k\}\big] \\
&= \mathbb{E}[S_k^2\,\mathbb{I}\{A_k\}] + 2\,\mathbb{E}[S_kY_k\,\mathbb{I}\{A_k\}] + \mathbb{E}[Y_k^2\,\mathbb{I}\{A_k\}] \\
&= \mathbb{E}[S_k^2\,\mathbb{I}\{A_k\}] + 2\,\mathbb{E}[S_k\,\mathbb{I}\{A_k\}]\underbrace{\mathbb{E}[Y_k]}_{0} + \mathbb{E}[Y_k^2\,\mathbb{I}\{A_k\}] \\
&= \mathbb{E}[S_k^2\,\mathbb{I}\{A_k\}] + \underbrace{\mathbb{E}[Y_k^2\,\mathbb{I}\{A_k\}]}_{\geq 0} \\
&\geq \mathbb{E}[S_k^2\,\mathbb{I}\{A_k\}] \\
&\geq \mathbb{E}[\varepsilon^2\,\mathbb{I}\{A_k\}] \\
&= \varepsilon^2\,\mathbb{P}(A_k).
\end{aligned}$$

Plugging this back in,

$$\mathrm{Var}(S_n) \geq \sum_{k=1}^{n} \mathbb{E}[S_n^2\,\mathbb{I}\{A_k\}] \geq \sum_{k=1}^{n} \varepsilon^2\,\mathbb{P}(A_k) = \varepsilon^2\,\mathbb{P}(A_k).$$

Thus,

$$\mathbb{P}(A) \leq \frac{\mathrm{Var}(S_n)}{\varepsilon^2}.$$

THEOREM 46: Kolmogorov's Criterion

Suppose $X_1, X_2, \ldots$ are independent random variables with

$$\sum_{k=1}^{\infty} \frac{\mathrm{Var}(X_n)}{k^2} < \infty.$$

Then,

$$\frac{S_n - \mathbb{E}[S_n]}{n} \xrightarrow{a.s.} 0 \text{ as } n \to \infty,$$

where $S_n = \sum_{k=1}^{n} X_k$ for $n \geq 1$.

Proof: Assume WLOG that $\mathbb{E}[S_k] = 0$ for $k \geq 1$. Fix $\varepsilon > 0$. Let

$$A_k = \frac{|S_n|}{n} \geq \varepsilon, \quad \text{for some } n \in (2^{k-1}, 2^k].$$

We want to show

$$\mathbb{P}\{(A_k)_{k\geq 1} \text{ i.o.}\} = 0.$$

Using the Borel-Cantelli lemma, we want to show $\sum_{k=1}^{\infty} \mathbb{P}(A_k) < \infty$.

$$\begin{aligned}
\mathbb{P}(A_k) &\leq \mathbb{P}\{|S_n| \geq 2^{k-1}\varepsilon\} && \text{for some } n \leq 2^k \\
&\leq \frac{\mathrm{Var}(S_{2^k})}{(2^{k-1}\varepsilon)^2} && \text{by Kolmogorov's Inequality} \\
&= \frac{4}{\varepsilon^2}\frac{\mathrm{Var}(S_{2^k})}{2^{2k}}.
\end{aligned}$$

Therefore,

$$\begin{aligned}
\sum_{k=1}^{\infty} \mathbb{P}(A_k) &\le \frac{4}{\varepsilon^2} \sum_{k=1}^{\infty} \frac{\operatorname{Var}(S_{2^k})}{2^{2k}} \\
&= \frac{4}{\varepsilon^2} \sum_{k=1}^{\infty} 2^{-2k} \sum_{j=1}^{2^k} \operatorname{Var}(X_j) \\
&= \frac{4}{\varepsilon^2} \sum_{\substack{1 \le k < \infty \\ 1 \le j \le 2^k}} 2^{-2k} \operatorname{Var}(X_j) \\
&= \frac{4}{\varepsilon^2} \sum_{j=1}^{\infty} \operatorname{Var}(X_j) \sum_{k=\lceil \log_2(j) \rceil}^{\infty} (2^{-2})^k \\
&= \frac{4}{\varepsilon^2} \sum_{j=1}^{\infty} \operatorname{Var}(X_j) \frac{(2^{-2})^{\lceil \log_2(j) \rceil}}{1 - 2^{-2}} \\
&\le \frac{4}{\varepsilon^2} \frac{4}{3} \sum_{j=1}^{\infty} \operatorname{Var}(X_k) (2^{-2})^{\log_2(j)} \\
&= \frac{16}{3\varepsilon^2} \sum_{j=1}^{\infty} \operatorname{Var}(X_j) j^{-2} \\
&< \infty
\end{aligned}$$

by our hypothesis. It's worth noting that to change the sums we have $j \le 2^k$, $2^k \ge j$, $k \ge \log_2(j)$ so $k \ge \lceil \log_2(j) \rceil$. Therefore, by Borel-Cantelli lemma,

$$\mathbb{P}\{(A_k)_{k \ge 1} \text{ i.o.}\} = 0.$$

Since this holds for every $\varepsilon > 0$,

$$\mathbb{P}\left\{ \lim_{n \to \infty} \frac{|S_n|}{n} = 0 \right\} = 1.$$

THEOREM 47: Strong Law of Large Numbers (IID)

If $X_1, X_2, \ldots$ are IID variables with finite expectation and $S_n = \sum_{j=1}^{n} X_j$ for $n \ge 1$, then

$$\frac{S_n}{n} \xrightarrow{a.s.} \mathbb{E}[X_1] \text{ as } n \to \infty.$$

Proof: For $n \ge 1$, let $Y_n = X_n \, \mathbb{I}\{|X_n| \le n\}$.

$$\begin{aligned}
\sum_{n=1}^{\infty} \mathbb{P}\{|X_n| > n\} &= \sum_{n=1}^{\infty} \sum_{k=n}^{\infty} \mathbb{P}\{k < |X_1| \le k+1\} \\
&= \sum_{k=1}^{\infty} \sum_{n=1}^{k} \mathbb{P}\{k < |X_1| \le k+1\} \\
&= \sum_{k=1}^{\infty} k \, \mathbb{P}\{k \le |X_1| \le k+1\} \\
&\le \mathbb{E}\big[|X_1|\big] \\
&< \infty.
\end{aligned}$$

Thus, by Borel-Cantelli,

$$\mathbb{P}\{X_n \neq Y_n \text{ i.o.}\} = 0.$$

Hence, it suffices to prove

$$\frac{S_n'}{n} \xrightarrow{a.s.} \mathbb{E}[X_1] \text{ as } n \to \infty,$$

where $S_n' = \sum_{j=1}^n Y_j$. We can also assume WLOG $\mathbb{E}[X_1] = 0$.

$$\mathbb{E}[Y_n] = \mathbb{E}[X_1 \, \mathbb{I}\{|X_1| \leq n\}] \to \mathbb{E}[X_1] \text{ as } n \to \infty$$

(Application of the Dominated Convergence Theorem). Therefore,

$$\frac{1}{n}\sum_{j=1}^n \mathbb{E}[Y_j] \to 0 \text{ as } n \to \infty.$$

It would suffice to prove

$$\frac{1}{n}\sum_{j=1}^n \underbrace{(Y_j - \mathbb{E}[Y_j])}_{Z_j} \xrightarrow{a.s.} 0 \text{ as } n \to \infty.$$

Note that $\mathbb{E}[Z_j] = 0 \implies \mathrm{Var}(Z_j) = \mathrm{Var}(Y_j)$. By Kolmogorov's Criterion, it would be sufficient to show

$$\sum_{k=1}^\infty \frac{\mathrm{Var}(Z_j)}{j^2} < \infty.$$

$$\begin{aligned}
\sum_{k=1}^\infty \frac{\mathrm{Var}(Y_k)}{k^2} &\leq \sum_{k=1}^\infty \frac{\mathbb{E}[Y_k^2]}{k^2} \\
&= \sum_{k=1}^\infty \frac{\mathbb{E}[X_k^2 \, \mathbb{I}\{X_k\} < k]}{k^2} \\
&= \sum_{k=1}^\infty \frac{1}{k^2} \sum_{j=1}^k \mathbb{E}[X_1^2 \, \mathbb{I}\{j-1 < |X_1| < j\}] \\
&= \sum_{j=1}^\infty \mathbb{E}[X_j^2 \, \mathbb{I}\{j-1 < |X_j| \leq j\}] \sum_{k=j}^\infty \frac{1}{k^2} \\
&\leq \sum_{j=1}^\infty \mathbb{E}[X_j^2 \, \mathbb{I}\{j-1 < |X_j| \leq j\}] \frac{C}{j} && \text{for some } C > 0 \\
&\leq \sum_{j=1}^\infty \mathbb{E}[j|X_1| \, \mathbb{I}\{j-1 < |X_1| \leq j\}] \frac{C}{j} \\
&= C \sum_{j=1}^\infty \mathbb{E}[|X_1| \, \mathbb{I}\{j-1 < |X_1| \leq j\}] \\
&= C \, \mathbb{E}\left[|X_1| \sum_{j=1}^\infty \mathbb{I}\{j-1 < |X_1| \leq j\} \right] \\
&= C \, \mathbb{E}[|X_1|] \\
&< \infty.
\end{aligned}$$

DEFINITION 58: Statistic

Recall, given a sample $(X_1, X_2, \ldots, X_n) = \boldsymbol{X}$, a **statistic** of the sample is some function $T(\boldsymbol{X}) \in \mathbf{R}^d$.

DEFINITION 59

We say T is a **sufficient statistic** for a parameter θ of the distribution of the sample if any inference about θ based on the sample should depend only on $T(\boldsymbol{X})$.

DEFINITION 60

$T(\boldsymbol{X})$ is a **sufficient statistic** for θ if the conditional distribution of $\boldsymbol{X}$ given $T(\boldsymbol{X})$ does not depend on θ.

$$\theta \leftrightarrow T(\boldsymbol{X}) \leftrightarrow \boldsymbol{X}.$$

- In a Bayesian framework, we would say θ and $\boldsymbol{X}$ are conditionally independent given $T(\boldsymbol{X})$.
- If $T(\boldsymbol{x}) = T(\boldsymbol{y})$, then our inferences about θ should be the same in the sample.

THEOREM 48

If $p(\boldsymbol{x} \mid \theta)$ is the joint pmf or pdf of the sample $\boldsymbol{X}$ and $q(t \mid \theta)$ is the pmf or (joint) pdf of $T(\boldsymbol{X})$, then $T(\boldsymbol{X})$ is a sufficient statistic for θ if and only if the ratio

$$\frac{p(\boldsymbol{x} \mid \theta)}{q(T(\boldsymbol{x}) \mid \theta)} = \mathbb{P}(\{\boldsymbol{X} = \boldsymbol{x}\} \mid \{T(\boldsymbol{X}) = T(\boldsymbol{x})\})$$

does not depend on θ; that is,

$$\forall \boldsymbol{x}\ \exists C \in [0, \infty) \text{ such that } \forall \theta,\ \frac{p(\boldsymbol{x} \mid \theta)}{q(T(\boldsymbol{x}) \mid \theta)} = C.$$

EXAMPLE 74

$X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{BERN}(\theta)$ for $0 < \theta < 1$. Let $T(\boldsymbol{X}) = \sum_{j=1}^n X_j$, so $T(\boldsymbol{X}) \sim \text{BIN}(n, \theta)$. Fix $\boldsymbol{x} \in \{0,1\}^n$. Let $t = \sum_{i=1}^n x_i = T(\boldsymbol{x})$. First,

$$p(\boldsymbol{x} \mid \theta) = \prod_{i=1}^n \begin{cases} \theta, & x_i = 1, \\ 1-\theta, & x_i = 0 \end{cases} = \theta^{T(\boldsymbol{x})} = (1-\theta)^{n - T(\boldsymbol{x})}.$$

REMARK 12

$$\mathbb{P}\{11010 \mid \theta\} = \theta \cdot \theta \cdot (1-\theta) \cdot \theta \cdot (1-\theta) = \theta^3 (1-\theta)^2.$$

Second,

$$q(t \mid \theta) = \binom{n}{t} \theta^t (1-\theta)^{n-t}.$$

REMARK 13

$$\mathbb{P}\{\text{three 1's and two 0's} \mid \theta\} = \mathbb{P}\{T(\boldsymbol{X}) = 3 \mid \theta\}.$$

Using these two facts,

$$\frac{p(\boldsymbol{x} \mid \theta)}{q(T(\boldsymbol{x}) \mid \theta)} = \frac{1}{\binom{n}{T(\boldsymbol{x})}},$$

which does not depend on θ, so T is sufficient for θ.

EXAMPLE 75

$\boldsymbol{X} = (X_1, \ldots, X_n) \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$. Suppose σ^2 is known and μ is unknown.

$$T(\boldsymbol{X}) = \bar{X} = \frac{X_1 + \cdots + X_n}{n} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right).$$

Recall the following trick:

$$\begin{aligned}
\sum_{i=1}^{n}(x_i - \mu)^2 &= \sum_{i=1}^{n}\big((x_i - \bar{x}) + (\bar{x} - \mu)\big) \\
&= n(\bar{x} - \mu)^2 + \sum_{i=1}^{n}(x_i - \bar{x}) + 2(\bar{x} - \mu)\underbrace{\sum_{i=1}^{n}(x_i - \bar{x})}_{0} \\
&= n(\bar{x} - \mu)^2 + \sum_{i=1}^{n}(x_i - \bar{x}).
\end{aligned}$$

First,

$$p(\boldsymbol{x} \mid \mu, \sigma^2) = (2\pi\sigma^2)^{-n/2} \exp\left\{-\frac{\sum_{i=1}^{n}(x_i - \mu)^2}{2\sigma^2}\right\}.$$

Second,

$$q(t(\boldsymbol{x}) \mid \mu, \sigma^2) = \left(2\pi\frac{\sigma^2}{n}\right)^{-1/2} \exp\left\{-\frac{(\bar{x} - \mu)^2}{2(\sigma^2/n)}\right\}.$$

Hence,

$$\begin{aligned}
\frac{p}{q} &= \sqrt{n}(2\pi\sigma^2)^{(1-n)/2} \exp\left\{-\frac{1}{2\sigma^2}\left(n(\bar{x} - \mu)^2 + \sum_{i=1}^{n}(x_i - \bar{x})^2\right) + \frac{n(\bar{x} - \mu)^2}{2\sigma^2}\right\} \\
&= \sqrt{n}(2\pi\sigma^2)^{-(n-1)/2} \exp\left\{-\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{2\sigma^2}\right\},
\end{aligned}$$

which does not depend on μ, so T is a sufficient statistic for μ.

The vector of order statistics of a sample $X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}$ is sufficient for everything.

THEOREM 49: Factorization Theorem (Halmos + Savage, 1949)/(Neyman 1935)

T is sufficient for θ if and only if there exists functions g and h such that

$$\forall \boldsymbol{x}\ \forall \theta,\ p(\boldsymbol{x} \mid \theta) = g(T(\boldsymbol{x} \mid \theta))h(\boldsymbol{x})$$

Proof (Discrete Setting): Assume $\boldsymbol{X}$ is a sample from a discrete distribution.
($\implies$) Assume T is sufficient. Choose $g(t, \theta) = q(t \mid \theta)$ and $h(\boldsymbol{x}) = \mathbb{P}\{\boldsymbol{X} = \boldsymbol{x} | T(\boldsymbol{X}) = T(\boldsymbol{x})\}$ (sufficiency

was used to define $h(\boldsymbol{x})$). Hence,

$$\begin{aligned}
g(T(\boldsymbol{x}),\theta)h(\boldsymbol{x}) &= q(T(\boldsymbol{x}) \mid \theta)\frac{\mathbb{P}\{\boldsymbol{X}=\boldsymbol{x} \mid \theta\}}{\mathbb{P}\{T(\boldsymbol{X})=T(\boldsymbol{x}) \mid \theta\}} \\
&= q(T(\boldsymbol{x}) \mid \theta)\frac{p(\boldsymbol{x} \mid \theta)}{q(T(\boldsymbol{x}) \mid \theta)} \\
&= p(\boldsymbol{x} \mid \theta).
\end{aligned}$$

($\Longleftarrow$) Assume $p(\boldsymbol{x} \mid \theta) = g(T(\boldsymbol{x}),\theta)h(\boldsymbol{x})$ for some g and h. Then,

$$q(t \mid \theta) = \sum_{\boldsymbol{x}\colon T(\boldsymbol{x})=t} p(\boldsymbol{x} \mid \theta).$$

Therefore,

$$\begin{aligned}
\frac{p(\boldsymbol{x} \mid \theta)}{q(T(\boldsymbol{x}) \mid \theta)} &= \frac{g(T(\boldsymbol{x}),\theta)h(\boldsymbol{x})}{\sum_{\boldsymbol{y}\colon T(\boldsymbol{y})=T(\boldsymbol{x})} g(T(\boldsymbol{y}),\theta)h(\boldsymbol{y})} \\
&= \frac{g(T(\boldsymbol{x}),\theta)h(\boldsymbol{x})}{\sum_{\boldsymbol{y}\colon T(\boldsymbol{y})=T(\boldsymbol{x})} g(T(\boldsymbol{x}),\theta)h(\boldsymbol{y})} \\
&= \frac{h(\boldsymbol{x})}{\sum_{\boldsymbol{y}\colon T(\boldsymbol{y})=T(\boldsymbol{x})} h(\boldsymbol{y})},
\end{aligned}$$

which does not depend on θ, so T is sufficient.

EXAMPLE 76

In our $\mathcal{N}(\mu,\sigma^2)$ example with $T(\boldsymbol{X}) = \bar{X}$, we have

$$h(\boldsymbol{x}) = \sqrt{n}(2\pi\sigma^2)^{-(n-1)/2}\exp\left\{-\frac{\sum_{i=1}^n (x_i-\bar{x})^2}{2\sigma^2}\right\},$$

and

$$q(t,\mu) = \left(2\pi\frac{\sigma^2}{n}\right)^{-1/2}\exp\left\{-\frac{(t-\mu)^2}{2(\sigma^2/n)}\right\}.$$

EXAMPLE 77

IID Uniform$\{1,2,\dots,\theta\}$ for $\theta \in \mathbf{N}$.

$$p(\boldsymbol{x} \mid \theta) = \begin{cases} \dfrac{1}{\theta^n}, & \max\limits_{1\le i\le n} x_i \le \theta, \\ 0, & \text{otherwise.} \end{cases}$$

So, $T(\boldsymbol{x}) = \max_{1\le i\le n} x_i$. Take

$$g(t,\theta) = \begin{cases} \dfrac{1}{\theta^n}, & t_i \le \theta, \\ 0, & \text{otherwise,} \end{cases}$$

and

$$h(\boldsymbol{x}) = \begin{cases} 1, & x_i \in \mathbf{N}\ \forall i, \\ 0, & \text{otherwise.} \end{cases}$$