Honors Statistical Inference I - APMA 1655

Milan Capoor

Spring 2023

Probability

1 Lecture 1, Jan 25: Random outcomes & sample spaces

Features of random events:

- There is more than one possible outcome.
- Before doing or observing the experiment of interest, you do not know which outcome you will see.
- Some possible outcomes are likely, and some other outcomes are quite unlikely e.g., winning one billion dollars by buying a lottery ticket.

Sample Spaces

Sample space: the set of all possible outcomes or results of that experiment (usually denoted Ω) Examples:

- Coin toss ($\Omega = \{H, T\}$)
- Schrodinger's cat ($\Omega = \{\text{alive}, \text{dead}\}$)
- The lifespan of a tree ($\Omega = \{1, 2, ...100, ...\} = \mathbb{Z}_+$)

Also note that not all elements/subsets of Ω are equally likely - hence, "probability"

2 Lecture 2, Jan 27: Events, event operations, and infinite operations

Suppose Ω is a sample space. Then:

- *Event:* each subset E of Ω
- *Impossible event:* the empty set $\emptyset$

Example: Final exam scores

1. The sample space: $\Omega = \{0, 1, 2, ..., 100\}$
2. The event "score is higher than 50": $E = \{51, 52, ..., 100\} \subset \Omega$
3. The event "score is a negative number": $\emptyset$

Set/Event Operations

Suppose Ω is a sample space and A and B are events $\{\omega \in \Omega : \omega \in A \text{ and } \omega \in B\}$:

- *Intersection:* both A and B occur; the collection of elements that are in sets A AND B

$$A \cap B$$

- *Union:* either A or B occurs; the collection of elements in A or B

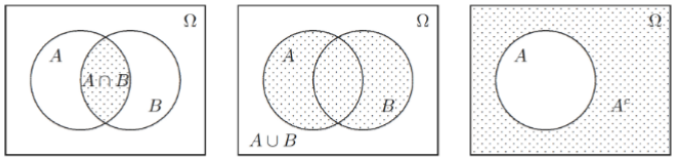
$$A \cup B$$

- *Complement:* the collection of elements that are not in A; the opposite event of A

$$A^c$$

Note that

$$\Omega^c = \emptyset$$
$$\emptyset^c = \Omega$$



Intersection $A \cap B$ Union $A \cup B$ Complement A^c

De Morgan's Laws: for any two events A and B we have the following

$$(A \cup B)^c = A^c \cap B^c \tag{1}$$
$$(A \cap B)^c = A^c \cup B^c \tag{2}$$

Infinite Sets

Suppose $A_1, A_2, A_3, ...A_n, A_{n+1}, ...$ are events. Some of them may be identical and some of them may be empty

Infinite Operations:

- *Infinite intersection:* the collection of events that are in ALL the sets $A_1, ..., A_n$; i.e. "all the events A_n for $n = 1, 2, ...$ happen"

$$\bigcap_{n=1}^{\infty} A_n = \{\omega \in \Omega : \omega \in A_n \forall n = 1, 2, 3, ...\}$$

- *Infinite union:* the collection of elements in at least one of the sets; "at least one of these events happen"

$$\bigcup_{n=1}^{\infty} A_n = \{\omega \in \Omega : \exists n' | \omega \in A_{n'}\}$$

("there exists at least one n' such that omega is in the set)

3 Lecture 3, Jan 30: Probability space & properties of probability

Disjoint: for two events A and B, they are disjoint if $A \cap B = \emptyset$

Mutually disjoint: if all pairwise intersections of $A_1, A_2, ..., A_n$ are empty ($A_n \cup A_m = \emptyset$ if $n \neq m$)

Definition of Probability: from the following definition we can derive everything in probability theory.

Let Ω be a sample space. Suppose $\mathbb{P}$ is a real-valued function of subsets of Ω

$$\mathbb{P} : \{\text{subsets of } \Omega\} \to \mathbb{R}, \quad A \mapsto \mathbb{P}\{A\}$$

where A is an input and $\mathbb{P}A$ is the corresponding output. If $\mathbb{P}$ satisfies the following three axioms, the pair $(\Omega, \mathbb{P})$ is a *probability space*

1. $\mathbb{P}(A) \geq 0$ for any subset $A \subset \Omega$ (the probability of an event must be non-negative)
2. $\mathbb{P}(\Omega) = 1$
3. For any sequence of disjoint subsets $\{A_i\}_{i=1}^{\infty}$ (i.e. $A_i \cap A_j) = \emptyset$ we have

$$\mathbb{P}\{\bigcup_{i=1}^{\infty} A_i\} = \sum_{i=1}^{\infty} \mathbb{P}(A_i)$$

The map $\mathbb{P}$ is called a *probability*

We can define this as a specific function

$$\mathbb{P}(A) := \frac{\#A}{n} \quad A \subset \Omega$$

where #A is the number of elements in A and $\Omega = \{1, 2, ..., n\}$ with a large n.

Lecture 4, Feb 1: Properties of Probability

Let $(\Omega, \mathbb{P})$ be a probability space. Then,

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

   Note: while this implies that the probability of an impossible even is 0, there can be zero-probability events which are not themselves impossible

2. if two events E_1 and E_2 satisfy $E_1 \cap E_2 = \emptyset$, then

$$\mathbb{P}(E_1 \cup E_2) = \mathbb{P}(E_1) + \mathbb{P}(E_2)$$

3. if $A, B \subset \Omega$ and $A \subset B$, then $\mathbb{P}(A) \leq \mathbb{P}(B)$

   (Intuitively, if A happens, B must also happen so B is more likely)

4. $0 \leq \mathbb{P}(A) \leq 1$ for $A \subset \Omega$
5. $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$

6. for any $A, B \subset \Omega$

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

7. for any countable collection of subsets

$$\mathbb{P}\{\bigcup_{n=1}^{\infty} A_n\} \leq \sum_{n=1}^{\infty} \mathbb{P}\{A_n\}$$

Note: the equality is obvious from axiom three in the case where all events are mutually disjoint. in the case of intersections, though, rule 6 must be generalized to account for overlap, hence the less than or equal to

4 Lecture 5, Feb 3: Conditional Probability

Part I - Motivating Problem

We know that a family has two children.

$$\Omega = \{(g,g),(b,b),(g,b),(b,g)\}$$
$$\mathbb{P}(A) = \frac{\#A}{\#\Omega} = \frac{\#A}{4} \quad A \subset \Omega$$

Event 1: $A = \{(g,g),(b,g),(g,b)\}$ ("at least one is girl")

$$\mathbb{P}(A) = \frac{3}{4}$$

Now suppose we get further information that the family has at least one boy:
$B = \{(b,g),(b,b),(g,b)\}$:

$$\mathbb{P}(A|B) = \frac{\#(A \cap B)}{\#B} = \frac{\#\{(b,g),(g,b)\}}{\#\{(b,g),(b,b),(g,b)\}} = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} = \frac{1}{2}$$

("knowing that event B occurs, what is the updated likelihood of A?")

Part II - A more rigorous definition

Let $A, B \subset \Omega$ such that

1. if $\mathbb{P}(B) > 0$ we call the following "the *conditional probability of A given B*"
$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$$
2. otherwise, if $\mathbb{P}(B) = 0$ then $\mathbb{P}(A|B)$ is not well defined in the scope of this course (see real analysis)

Theorem: Let $(\Omega, \mathbb{P})$ be a probability space. Suppose $B \subset \Omega$ and $\mathbb{P}(B) > 0$. Then let
$$\tilde{\mathbb{P}}(A) := \mathbb{P}(A|B) \quad \forall A \subset \Omega$$
Then, $\tilde{\mathbb{P}}$ is another probability defined on Ω such that $\tilde{\mathbb{P}}$ also satisfies the 3 axioms

Part III - Properties of Conditional Probabilities

Assuming $\mathbb{P}(B) > 0$, and $A, B \subset \Omega$:

1. Multiplication Law
$$\mathbb{P}(A \cap B) = \mathbb{P}(A|B) \cdot \mathbb{P}(B)$$
2. Let $B_1, B_2, ..., B_n \subset \Omega$. We say the events provide a *partition* of Ω if they are mutually disjoint and they satisfy[1]
$$\bigcup_{i=1}^{n} B_i = \Omega$$
3. *The law of total probability*

   Let $B_1, B_2, ..., B_n$ be events and they provide a partition of Ω and $\mathbb{P}(B_i) > 0$. Then,
$$\mathbb{P}(A) = \sum_{i=1}^{n} \mathbb{P}(A|B_i) \cdot \mathbb{P}(B_i)$$

[1]Note that for $B_i = \{B_1, ...B_n\}$,
$$A \cap \left(\bigcup_{i=1}^{n} B_i \right) = A \cap (B_1 \cup B_2 \cup ... \cup B_n) = \bigcup_{i=1}^{n} A \cap B_i$$

Proof:

$$\Omega = \bigcup_{i=1}^{n} B_o \implies A = A \cap \Omega$$

Then, by the laws above, $A \cap B_1, A \cap B_2, ...A \cap B_n$ is mutually disjoint. So

$$\mathbb{P}(A) = \mathbb{P}(\bigcup_{i=1}^{n} A \cap B_i) = \sum_{i=1}^{n} \mathbb{P}(A \cap B_i)$$

Then from the multiplication law, $\mathbb{P}(A) = \sum_{i=1}^{n} \mathbb{P}(A|B_i) \cdot \mathbb{P}(B_i)$ and the desired result follows

4. Corollary: Let B be an event with $0 < \mathbb{P}(B) < 1$. Then we have

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

Proof: Let $B_1 = B,\ B_2 = B^c,\ n = 2.$ B_1 and B_2 obviously partition Ω and

$$P(B_1) = \mathbb{P}(B) > 0, \quad \mathbb{P}(B_2) = \mathbb{P}(B^c) = 1 - \mathbb{P}(B) > 0$$

Then the corollary follows from the law of total probability

5 Lecture 6, Feb 6: Bayes' formula

Part I - Bayes' Rule

Suppose $B_1, ..., B_n$ provide a partition of Ω. In addition, $\mathbb{P}(B_i) > 0 \quad \forall i \in [1, n]$. Let A be any event such that $\mathbb{P}(A) > 0$. Then,

$$\mathbb{P}(B_i|A) = \frac{\mathbb{P}(A|B_i) \cdot \mathbb{P}(B_i)}{\sum_{j=1}^{n} \mathbb{P}(A|B_j) \cdot \mathbb{P}(B_j)} \quad i = 1, 2, ..., n$$

Proof: The definition of conditional probability implies

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

The multiplication law implies

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

The law of total probability implies

$$\mathbb{P}(A) = \sum_{i=1}^{n} \mathbb{P}(A|B_i) \cdot \mathbb{P}(B_i)$$

Then the desired result follows.

Though this proof is trivial, the formula is quite meaningful in that it allows an exchange between the conditions and results. Note that this is not the general Bayes formula.

6 Lecture 7, Feb 8: Independence

Independent events do not affect each other's outcomes (e.g. flipping two coins).That is

$$\begin{cases} \mathbb{P}(A|B) = \mathbb{P}(A) \\ \mathbb{P}(B|A) = \mathbb{P}(B) \end{cases}$$

In other words, knowing A does not help predict B.

Together with those equations and the multiplication law, we have

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

(for independent events)

Definition: Let $(\Omega, \mathbb{P})$ be a probability space.

1. Suppose A and B are two events. They are *independent* if

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

2. Suppose $A_1, A_2, ...A_n$ are a sequence of events. The sequence is *mutually independent* if

$$\mathbb{P}(A_m \cap A_n) = \mathbb{P}(A_m) \cdot \mathbb{P}(A_n) \quad m \neq n$$

Theorem: Let $(\Omega, \mathbb{P})$ be a probability space with $A, B \subset \Omega$ such that $\mathbb{P}(A) > 0$ and $\mathbb{P}(B) > 0$ (this is necessary to have the conditional probabilities well-defined). Then the following three equations are equivalent:

1. $\mathbb{P}(A|B) = \mathbb{P}(A)$

2. $\mathbb{P}(B|A) = \mathbb{P}(B)$
3. $\mathbb{P}(A \cap B) = \mathbb{P}(A) \cdot \mathbb{P}(B)$

Note that we choose this third equation to be the definition of independence stated above specifically because it does not defend on the positive probability condition of the other two

Example: Suppose a family has 3 children of unknown gender

$$\Omega = \{(g,g,g),(g,g,b),(g,b,g),(b,g,g),(g,b,b),(b,g,b),(b,b,g),(b,b,b)\}$$

$$\#\Omega = 2^3 = 8 \Longrightarrow \mathbb{P}(A) := \frac{\#A}{8}$$

Consider the events A ("the family has boys and girls") and B ("the family has at most one girl")

Question: Are A and B independent?

$$\begin{aligned} \mathbb{P}(A) &= \frac{6}{8} \\ \mathbb{P}(B) &= \frac{4}{8} \\ \mathbb{P}(A) \cdot \mathbb{P}(B) &= \frac{3}{8} \\ \mathbb{P}(A \cap B) &= \frac{3}{8} = \mathbb{P}(A) \cdot \mathbb{P}(B) \end{aligned}$$

So the events are independent

7 Lecture 8, Feb 10: Random Variable

Probability theory has 3 building blocks:

1. Sample space (Ω)
2. Probability ($\mathbb{P}$) Together, these two give us probability space $(\Omega, \mathbb{P})$
3. Random Variable

Motivating Example: Let Ω be the collection of all undergraduate students at Brown. Then $\mathbb{P}(A) := \frac{\#A}{\#\Omega}$. And, for each student $\omega \in \Omega$,

$$X(\omega) = \text{the SAT score of the given student} \quad \{X : \Omega \to \mathbb{R},\ \omega \mapsto X(\omega)\}$$

but where ω is unknown (say to protect anonymity) so it can be any value in Ω. Note, however, that if ω is unknown, then X is also uncertain.

Definition: Let $(\Omega, \mathbb{P})$ be a probability space. Suppose X is a real-valued function defined on Ω,

$$\begin{aligned} X : \Omega &\to \mathbb{R} \\ \omega &\mapsto X(\omega) \end{aligned}$$

We thus call X a *random variable*

8 Cumulative Distribution Functions

Motivating Example: Ω is all the undergrads at Brown, $\omega \in \Omega$ is a student at Brown, $X(\omega)$ is a random variable denoting the SAT score of a Brown student

Let A_{100} be the event "the SAT score of a Brown student is ≤ 100. Then,

$$A_{100} = \{\omega \in \Omega : X(\omega) \leq 100\}$$

Definition: Let $(\Omega, \mathbb{P})$ be a probability space and X be a random event. Then for any real number $X \in \mathbb{R}$, we define the event A_x by

$$A_x = \{\omega \in \Omega : X(\omega \leq x)\}$$

We define a real-valued function F on $\mathbb{R}$ by

$$F(x) := \mathbb{P}(A_x) = \mathbb{P}(\{\omega \in \Omega : X(\omega) \leq x\})$$

This function F is the *cumulative distribution function (CDF)* of the random variable X, usually written F_X

9 Lecture 9, Feb 13: Cumulative distribution function

Part I - Review

A random variable X is a function defined on a sample space Ω

For each real number x, we define an event

$$A_x = \{\omega \in \Omega : X(\omega) \leq x\}$$

We then define a function by

$$F_X : \begin{array}{c} \mathbb{R} \to [0,1] \\ x \mapsto \mathbb{P}(A_x) \end{array}$$

Note that $F_X = \mathbb{P}(X \leq x)$ and this function is called *the cumulative distribution function*

Part II - The CDF

Example 1: Flipping a coin

- $\Omega = \{H, T\}$
- $\mathbb{P}(A) := \frac{\#A}{\#\Omega} = \frac{\#A}{2} \quad (A \subset \Omega)$
- $$\begin{cases} X(H) = 1 \\ X(T) = 0 \end{cases}$$

Claim: the CDF of X is

$$F_X(x) = \begin{cases} 0 & x < 0 \\ \frac{1}{2} & 0 \leq x < 1 \\ 1 & x \geq 1 \end{cases}$$

Proof:

1. When $x < 0$, then
$$A_x = \{\omega \in \Omega : X(\omega) \leq x\} = \emptyset$$
(this is impossible because X is only 0 or 1 so not negative). So $F_X(x) = \mathbb{P}(A_x) = 0$
2. When $0 \leq x < 1$,
$$A_x = \{\omega \in \Omega : X(\omega) \leq x\} = \{\omega \in \Omega : X(\omega) = 0\} = \{T\}$$
so $F_X(x) = \mathbb{P}(\{T\}) = \frac{1}{2}$

3. When $x \geq 1$

$$A_x = \{\omega \in \Omega : X(\omega) \leq x\} = \{\omega \in \Omega : X(\omega) \leq 1\} = \{T, H\}$$

so $P(A_x) = 1$

Example 2: Flipping a biased coin

- $\Omega = \{H, T\}$
- $$\mathbb{P}(A) := \begin{cases} p & A = \{H\} \\ 1-p & A = \{T\} \end{cases}$$

  for some $p \in [0, 1]$
- $X(H) = 1, X(T) = 0$

Claim: the CDF here is

$$F_X(x) = \begin{cases} 0 & x < 0 \\ 1-p & 0 \leq x < 1 \\ 1 & x \geq 1 \end{cases}$$

Proof: same as above

Note: this example actually refers to the Bernoulli Distribution **Definition:** Let X be a random variable on $(\Omega, \mathbb{P})$ If the CDF of X is

$$F_X(x) = \begin{cases} 0 & x < 0 \\ 1-p & 0 \leq x < 1 \\ 1 & x \geq 1 \end{cases}$$

for some $o \in [0, 1]$, we say "X follows the Bernoulli Distribution with success probability p" and denote $X \sim \text{Bernoulli}(p)$

Theorem: Let $(\Omega, \mathbb{P})$ be a probability space. Suppose X is a random variable defined on Ω and its CDF is F_X

F_X is non-decreasing ($F_X(x_1) \leq F_X(x_2) \quad x_1 \leq x_2$) for any CDF.

Proof:

$$A_{x_1} = \{\omega \in \Omega : X(\omega) \leq x_1\}$$

$$A_{x_2} = \{\omega \in \Omega : X(\omega) \leq x_2\}$$
$$x_1 \leq x_1 \implies Ax_1 \subset Ax_2$$
$$F_X(x_1) = \mathbb{P}(A_{x_1}) \leq \mathbb{P}(A_{x_2}) \leq F_X(x_2)$$

Lecture 10, Feb 15:

Part I - Review

Let X be a random variable defined on a probability space $(\Omega, \mathbb{P})$. For any real number x,

$$A_x = \{\omega \in \Omega : X(\omega) \leq x\}$$

Then the cumulative distribution function of X is

$$F_X(x) = \mathbb{P}(A_x)$$

Part II - Properties of the CDF

Let X be any random variable on any probability space $(\Omega, \mathbb{P})$ with $F_X(x)$ as the corresponding CDF

1. Any CDF is non-decreasing $(F_X(x_1) \leq F_X(x_2) \quad x_1 \leq x_2)$
2. $$\lim_{x \to -\infty} F_X(x) = 0$$
3. $$\lim_{x \to \infty} F_X(x) = 1$$
   (Note the rigorous proof is needs real analysis)
4. F_X is right-continuous, i.e.
   $$\forall x_0 \in \mathbb{R}, \quad F_X(x_0) = \lim_{x \to x_0^+} F_X(x)$$
   Note that the "right-continuous" property is implied from the $\leq$ sign. With a strict less-than inequality, the CDF becomes left-continuous
5. For any $x_0 \in \mathbb{R}$,
   $$\mathbb{P}(\{\omega \in \Omega : X(\omega) = x_0\}) = F_X(x_0) - \lim_{x \to x_0^-} F_X(x)$$
   Note that this is zero if the CDF is continuous

Lecture 11, Feb 17

Part I - Review

The building blocks of probability together define the CDF:

$$\begin{cases} \Omega \\ \mathbb{P} \\ X \end{cases} \implies F_X(x) = \mathbb{P}(\{\omega \in \Omega : X(\omega) \leq x\})$$

Part II - Moving backwards

Question: Given a function F (satisfying some conditions), do there exist a sample space, a probability, and a random variable corresponding to the given F?

Theorem: Suppose we have a $F : \mathbb{R} \to [0, 1]$ satisfying

- F is non-decreasing
- With end behavior

$$\lim_{x \to -\infty} F(x) = 0, \quad \lim_{x \to \infty} F(x) = 1$$

- F is right-continuous

Then, there exist a sample space, a probability, and a random variable such that

$$F(x) = \mathbb{P}(\{\omega \in \Omega : X(\omega) \leq x\})$$

Proof: far beyond the scope of this course

Part III - Classification of Random Variables

Types of random variables:

- Continuous
- Discrete
- Neither continuous nor discrete

Definition: a function F is continuous if

$$\lim_{x \to x_0^-} F(x) = \lim_{x \to x_0^+} F(x) = F(x_0) \quad \forall\, x_0$$

Definition: a random variable X is a continuous random variable if the CDF $F_X : \mathbb{R} \to [0, 1]$ is a continuous function

Theorem: Let X be a random variable on $(\Omega, \mathbb{P})$. If X is a continuous random variable,

$$\mathbb{P}(\{\omega \in \Omega : X(\omega) = x_0\}) = 0 \quad \forall\, x_0 \in \mathbb{R}$$

Proof:

$$\mathbb{P}(\{\omega \in \Omega : X(\omega) = x_0\}) = F_X(x_0) - \lim_{x \to x_0^-} F_X(x) = F_X(x_0) - F_X(x_0) = 0$$

Lecture 12, Feb 22: Continuous Random Variables

Part I - A "theorem"

"Theorem": Let F_X be the CDF of a continuous random variable X. Then, F_X is differentiable.

Remark: the true and rigorous version of this "theorem" requires lots of pure math so this version does have some edge cases such that "differentiable" really means "piecewise differentiable"

Example:

$$F(x) = \begin{cases} 0 & x < 0 \\ x & 0 \leq x < 1 \\ 1 & x \geq 1 \end{cases}$$

This is not technically differentiable because of the sharp points at $x = \{0, 1\}$ but we can use a generalized derivative to cheat:

$$F'(x) = \begin{cases} 0 & x < 0 \text{ or } x > 1 \\ 1 & 0 < x < 1 \\ k & x = 0 \text{ or } x = 1 \end{cases}$$

where k is any value whatsoever.

Definition: Let F_X be the CDF of a continuous random variable X.

$$p_X(x) = F'_X(x)$$

We call $p_X(x)$ the *probability density function (PDF)* of the random variable X

Part II - The Rigorous Treatment (optional)

Definition: X is a continuous random variable if F_X is absolutely continuous

Theorem: X is a continuous random variable. Then its CDF F_X is differentiable "almost everywhere with respect to the Lebesgue measure"

Definition: $p_X(x) = F'_X(x)$ is the probability density function for a continuous random variable X

Part III - The Probability Density Function

Less rigorously,

$$p_X(x) = F'_X(x)$$

is the probability density function and is the piecewise derivative of F_X for the continuous random variable X

Theorem: Let X be a continuous RV with CDF $F_X(x)$ and PDF $p_X(x)$. Then,

$$F_X(x) = \int_{-\infty}^{x} p_X(t)\, dt$$

Remarks:

- $p_X(x) := F'_X(x)$ so CDF determines PDF
- $F_X(x) = \int_{-\infty}^{\infty} p_X(t)\, dt$ so PDF determines CDF

Part IV - Examples of continuous random variables

Definition: Let X be a RV. If the CDF od X is

$$F_X(x) = \begin{cases} 0 & x < a \\ \frac{x-a}{x-b} & a \leq x \leq b \\ 1 & x > b \end{cases}$$

(for $a < b$), then X follows the uniform distribution between a and b.

This is continuous so we can take the piecewise derivative as follows:

$$p_X(x) = \begin{cases} 0 & x < a \text{ or } x > b \\ \frac{1}{b-a} & a \leq x \leq b \end{cases}$$

Notice that the graph of this function is a rectangle with base $b - a$ and height $1/(b - a)$.

Experiment: Randomly select a number between 0 and 1

- $\Omega = (0, 1)$
- $X : \Omega \to \mathbb{R}, \quad X(\omega) = \omega, \quad \omega \in \Omega = (0, 1)$
- Because we randomly select numbers, we assume $X \sim \text{Unif}(0, 1)$
- Let $E = \{0.5\} = \{\omega \in \Omega : X(\omega) = 0.5\} = \emptyset$ be the event "the selected number is exactly 0.5"

So,

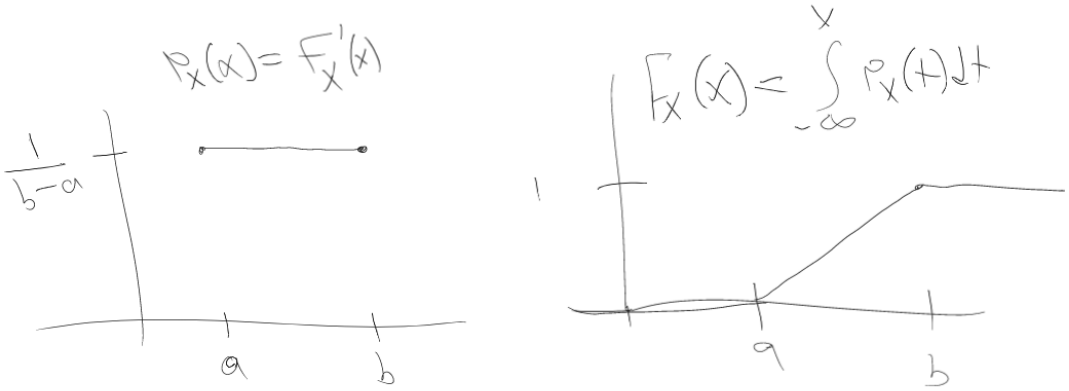
$$\mathbb{P}(E) = \mathbb{P}(X = 0.5) = F_X(0.5) - \lim_{x \to 0.5^-} F_X(x) = 0$$

because F_X is continuous so the limit is equal to the value

Lecture 13, Feb 24:

Part I - Review

- X is a continuous random variable iff F_X is a continuous function
- $p_X(x) = F_X'(x)$ is the "probability density function" of X
- The following are the graphs of the PDF and CDF of the random variable $X \sim \text{Unif}(a, b) \quad a < b$



- "Impossible" implies "zero probability" but zero probability does not imply impossible

Part II - The Probability Density Function

Theorem: Let X be a continuous random variable and its PDF is $P_X(x)$. Then

$$\int_{-\infty}^{\infty} p_X(t)\,dt = 1$$

Proof:

$$\begin{aligned}\int_{-\infty}^{\infty} p_X(t)\,dt &= \lim_{x\to+\infty}\int_{-\infty}^{x} p_X(t)\,dt \\ &= \lim_{x\to+\infty} F_X(x) = 1 \quad \blacksquare\end{aligned}$$

Interlude: Applications of the above law

- Bayesian statistics

$$f(\theta|x) = \frac{f(x,\theta)}{\int_{-\infty}^{\infty} f(x,\theta)\,d\theta}$$

- Quantum mechanics For each fixed t, $|\psi(x,t)|^2$ is the PDF of x so

$$\int_{-\infty}^{\infty} |\psi(x,t)|^2\,dx = 1$$

which is the "normalization condition od the Schrodinger equation"

Part III - Normal Distributions

Definition: Let X be a RV. We say "X follows the normal distribution with mean μ and variance σ^2" (denoted $X \sim N(\mu, \sigma^2)$), if

$$F_X(x) = \int_{-\infty}^{x} \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{(t-\mu)^2}{2\sigma^2}}$$

additionally,

$$p_X(x) = F_X'(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(t-\mu)^2}{2\sigma^2}}$$

Lecture 14, Feb 27: Discrete Random Variables

Part I - Preparation

Definition: Let A be a subset of $\mathbb{R}$. Then the *indicator function of A* is

$$\mathbf{1}_A(x) = \begin{cases} 1 & x \in A \\ 0 & x \notin A \end{cases}$$

Example 1: If $A = [0, +\infty)$, then

$$\mathbf{1}_A(x) = \mathbf{1}_{[0,+\infty)}(x) = \begin{cases} 1 & x \geq 0 \\ 0 & x < 0 \end{cases}$$

Example 2: If $A = [\sqrt{2}, +\infty)$, then

$$\mathbf{1}_A(x) = \mathbf{1}_{[\sqrt{2},+\infty)}(x) = \not\Vdash_{[0,+\infty)}(x - \sqrt{2}) = \begin{cases} 1 & x \geq \sqrt{2} \\ 0 & x < \sqrt{2}0 \end{cases}$$

Also note that for $X \sim$ Bernoulli($\frac{1}{2}$),

$$\begin{aligned} F_X(x) &= \begin{cases} 0 & x < 0 \\ \frac{1}{2} & 0 \leq x < 1 \\ 1 & x \geq 1 \end{cases} \\ &= \frac{1}{2}\mathbf{1}_{[0,+\infty)}(x) + \mathbf{1}_{[1,+\infty)}(x) \\ &= \mathbb{P}(X = 0) + \mathbb{P}(X = 1) \end{aligned}$$

Remember that the random variable X that conforms to the bernoulli distribution can only take 2 values which thus map to the two left endpoints 0 and 1 of the indicator functions.

More generally, say X takes values in $\{x_k\}_{k=1}^K = \{x_1, x_2, ..., x_K\}$ so the CDF can also be written

$$F_X(x) = \sum_{k=1}^{K} \mathbb{P}(X = x_k) \cdot \mathbf{1}_{[x_k, +\infty)}(x)$$

Heuristically, this is the sum of probabilities at each point x_k

Lecture 15, March 1: Discrete Random Variable Definition

Definition: Let X be a RV. X is a discrete random variable if its CDF can be written in the form

$$F_X(x) = \sum_{k_1}^{K} p_k \cdot \mathbf{1}_{[x_k, +\infty)}(x)$$

where $p_k \geq 0$ and $\sum_{k=1}^K p_k = 1$ as p_k represents a probability. Also, x_k is a real number, and K is allowed to be $+\infty$.

Finally, the ordered sequence $\{p_k\}_{k=1}^K$ is the *probability mass function* of X and represents the probability of each endpoint event

Remark: that the sequence $\{p_k\}_{k=1}^K$ is referred to as a "function" is simply a convention as the PMF is the counterpoint to the CDF. It may also refer to any function

$$k \mapsto p_k, \quad k \in \mathbb{Z}$$

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

$$F_X(x) = (1-p) \cdot \mathbf{1}_{[0,+\infty)}(x) + p \cdot \mathbf{1}_{[1,+\infty)}(x)$$

Remark: From the CDF

$$F_X(x) = \sum_{k=1}^{K} p_k \cdot \mathbf{1}_{[x_k, +\infty)}(x)$$

we can derive the PMF $\{p_k\}_{k=1}^K$. And if we know $\{x_k\}_{k=1}^K$ we can go from PMF to CDF

Theorem: Let X be a RV defined on $(\Omega, \mathbb{P})$ and its CDF is

$$F_X(x) = \sum_{k=1}^{K} p_k \cdot \mathbf{1}_{[x_k, +\infty)}(x)$$

then we have

$$\mathbb{P}(\{\omega \in \Omega : X(\omega) = x_k\}) = p_k$$

Proof: Omitted. The theorem is cheating. To make it true, make some topological assumptions such that $\{x_k\}_{k=1}^K$ as a subset of the real line is locally finite (or its derived set is $\emptyset$)

Poisson Distribution ($X \sim \mathbf{Pois}(\lambda)$)

"X follows the Poisson distribution with rate parameter λ" if $K = \infty$, $x_k = k \quad \forall k \in [0, \infty)$,

$$p_k = \frac{\lambda^k e^{-\lambda}}{k!}$$

where that quotient represents the probability that "k events occur in a fixed time interval"

Then,

$$\begin{aligned}
\sum_{k=0}^{\infty} p_k &= \sum_{k=0}^{\infty} \frac{\lambda^k e^{-\lambda}}{k!} \\
&= e^{-\lambda} \sum_{k=0}^{\infty} \frac{\lambda^k e^{-\lambda}}{k!} \\
&= e^{-\lambda} e^{\lambda} \quad \text{(taylor series for } e^x\text{)} \\
&= 1
\end{aligned}$$

Note: the starting index must be 0 for the taylor series to work

Random Variables that are neither continuous nor discrete

Part I - Preparation

Let Y and Z be random variables. They are independent if for *any*subsets $A, B \subset \mathbb{R}$,

$\mathbb{P}(\{\omega \in \Omega : (Y(\omega) \in A) \cap (Z(\omega) \in B)\}) = \mathbb{P}(\{\omega \in \Omega : Y(\omega) \in A\}) \cdot \mathbb{P}(\{\omega \in \Omega : Z(\omega) \in A\})$

Part II - Definition

Example: Let Y, Z be independent random variables such that

$$\begin{cases} Y \sim \text{Bernoulli}(\frac{1}{2}) \\ Z \sim N(0,1) \end{cases}$$

Then let

$$X(\omega) = Y(\omega) + (1 - Y(\omega)) \cdot Z(\omega)$$

Claim:

$$F_X(x) = \frac{1}{2}\mathbf{1}_{[0,+\infty)}(x) + \frac{1}{2}F_N(x) = \frac{1}{2}\mathbf{1}_{[0,+\infty)}(x) + \frac{1}{2}\int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}\, dt$$

Then the graph of F_X is continuous for all x except at $x = 0$ where there is a jump discontinuity.

Proof:

$$\begin{cases} A_X = \{\omega \in \Omega : X(\omega \leq x)\} \\ B = \{\omega \in \Omega : Y(\omega) = 1\} \end{cases}$$

Then by definition of CDF,

$$F_X(x) = \mathbb{P}(A_x) = \mathbb{P}(A_X|B) \cdot \mathbb{P}(B) + \mathbb{P}(A_X|B^x) \cdot \mathbb{P}(B^c)$$

(by law of total probability)

Lecture March 3: Independence

Part I - Independence between two events (Review)

Definition: Let $(\Omega, \mathbb{P})$ be a probability space. $\tilde{A}, \tilde{B} \subset \Omega$ are two events that are independent if

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

Part II - Independence between two random variables

Definition Version 1: Let Y and Z be RVs defined on $(\Omega, \mathbb{P})$. Y ans Z are independent if *for any* subsets $A, B \subset \mathbb{R}$ we define events via Y and Z by

$$\tilde{A} = \{\omega \in \Omega : Y(\omega) \in A\}$$
$$\tilde{B} = \{\omega \in \Omega : Z(\omega) \in B\}$$

where A and B are independent

Definition Version 2: Let Y and Z be random variables defined on $(\Omega, \mathbb{P})$. Y and Z are independent if

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

for any $A, B \in R$

Note that

$$\mathbb{P}((Y \in A) \cap (Z \in B)) = \mathbb{P}(\tilde{A} \cap \tilde{B})$$

showing that the two definitions are equivalent

Part III - Examples

Example 1 (Intuition):

$$Y = \text{outcome of Michael flipping a coin}$$
$$Z = \text{outcome of Taylor Swift flipping a coin}$$

Obviously, these are independent.

Example 2 (with math):

Y and Z are two independent RVs defined on $(\Omega, \mathbb{P})$. Let

$$Y \sim \text{Bernoulli}(\frac{1}{2})$$
$$Z \sim N(0, 1)$$
$$X(\omega) := Y(\omega) + (1 - Y(\omega)) \cdot Z(\omega)$$

Then for a fixed real number x,

$$\mathbb{P}((X \leq x) \cap (Y = 0)) = \mathbb{P}((Z \leq x) \cap (Y = 0))$$

by definition of X and the fixed value of Y. The above is then equal to

$$\mathbb{P}((Y \in \{0\}) \cap (Z \in (-\infty, x]))$$

where if $A = \{0\}$ and $B = (-\infty, x]$,

$$\begin{aligned} &= \mathbb{P}(Y \in \{0\}) \cdot \mathbb{P}(Z \in (-\infty, x]) \\ &= \mathbb{P}(Y = 0) \cdot \mathbb{P}(Z \leq x) \\ &= \frac{1}{2} F_Z(x) \end{aligned}$$

10 Lecture March 6: Mean/Expected Value

Part I - Motivation

Example 1: Flip a fair coin 1000 times. For the i-th flip

$$X_i = \begin{cases} 1 & \text{if Heads} \\ 0 & \text{if Tails} \end{cases}$$

Then intuitively, the average value (when the number of flips is large) will be

$$\bar{X}_{1000} = \frac{\sum_{i=0}^{1000} X_i}{1000} \approx \frac{1}{2}$$

or

$$\lim_{n \to \infty} \bar{X}_n = \frac{1}{2}$$

Then, in general,

$$\bar{X}_n = \frac{X_1 + X_2 + ... + X_n}{n} \approx \sum_{k=0}^{K} x_k \cdot p_k$$

Example 2: Office hours are 2-4pm on Mondays. On the i-th Monday, X_i students come. This can be modelled as

$$X_i \sim \text{Pois}(\lambda)$$

so

$$\bar{X}_n \approx \sum_{k=0}^{\infty} k \cdot \frac{\lambda^k e^{-k}}{k!} = \lambda$$

Part II - Conjecture

The law of large numbers: for almost all discrete CDFs and n sufficiently large

$$\boxed{\bar{X}_n = \frac{X_1 + X_2 + ... + X_n}{n} \approx \sum_{k=0}^{K} x_k \cdot p_k}$$

Part III - Definition

Discrete Definition: Let X be a discrete random variable with CDF

$$F_X(x) = \sum_{k=0}^{K} p_k \cdot \mathbf{1}_{[x_k, +\infty)}(x)$$

If

$$\sum_{k=0}^{K} |x_t| \cdot p_k < +\infty$$

then we call the following sum *the expected value/mean of X*:

$$\mathbb{E}(X) = \sum_{k=0}^{K} x_k \cdot p_k$$

Continuous Definition: Let X be a continuous random variable with PDF $p_X(x)$. If

$$\int_{-\infty}^{\infty} |x| \cdot p_X(x)\, dx < +\infty$$

then we call the following the *expected value/mean of X*:

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

11 Lecture March 8

Part I - Review

Definition: Let X be a discrete random variable with CDF

$$F_X(x) = \sum_{k=0}^{K} p_k \cdot \mathbf{1}_{[x_k, +\infty)}(x)$$

1. If $\sum_{k=0}^{K} |x_k| \cdot p_k < \infty$, then we call $\sum_{k=0}^{k} x_k \cdot p_k$ the expected value of X denoted $\mathbb{E}(X)$
2. If $\sum_{k=0}^{k} |x_k| \cdot p_k = \infty$ we say $\mathbb{E}(X)$ does not exist

Part II - Example

$$\begin{aligned} K &:= \infty \\ x_k &:= (-1)^{k+1} \cdot k \quad \forall k = 0, 1, 2, ... \\ p_0 &:= 0 \\ p_k &:= \frac{6}{\pi^2} \frac{1}{k^2} \quad \forall k = 1, 2, ... \end{aligned}$$

Then

$$\sum_{k=0}^{K} x_k \cdot p_k = \sum_{k=1}^{\infty} (-1)^{k+1} \cdot k \cdot \frac{6}{\pi^2} \cdot \frac{1}{k^2} = \frac{6}{\pi^2} \sum_{k=1}^{\infty} \frac{(-1)^{k+1}}{k}$$

from calculus, that series is equal to $\log 2$ so

$$= \frac{6}{\pi^2} \log 2$$

Question: is $\mathbb{E}(X) = \frac{6}{\pi^2} \log 2$? **Answer:** No. The expected value does not exist.

$$\sum_{k=0}^{K} |x_k| \cdot p_k = \sum_{k=1}^{\infty} k \cdot \frac{6}{\pi^2} \frac{1}{k} = \infty$$

because the sum diverges

Part III - Theory of Series

Definition: Let $\{a_k\}_{k=0}^{\infty}$ be an ordered series

1. If $\lim_{n \to \infty} \sum_{k=0}^{n} a_k$ exists, we say $\sum_{k=0}^{\infty} a_k$ is convergent

Theorem: If $\sum_{k=0}^{\infty} |a_k|$ is convergent, then so is

$$\sum_{k=0}^{\infty} a_k$$

Proof: Follows from Cauchy's convergence test

Definition:

1. If $\sum_{k=0}^{\infty} |a_k|$ is convergent, $\sum_{k=0}^{\infty} a_k$ is *absolutely convergent*
2. If $\sum_{k=0}^{\infty} a_k$ is convergent but $\sum_{k=0}^{\infty} |a_k|$ is divergent, then $\sum_{k=0}^{\infty} a_k$ is *conditionally convergent*

Remark: This reduces the condition of the expected value definition to "The expected value exists if the sum is absolutely convergent"

Definition: Suppose $\mathbb{N} = \{0, 1, 2, ...\}$. Any bijective map (i.e. the map is one-to-one and onto; $\forall i \in \mathbb{N}\ \exists$ a unique $j : \sigma(j) = i$) σ from $\mathbb{N} \mapsto \mathbb{N}$ is called a *permutation*

Theorem:

1. If $\sum_{k=0}^{\infty} a_k$ is absolutely convergent,

$$\sum_{k=0}^{\infty} a_{\sigma(k)} = \sum_{k=0}^{\infty} a_k$$

   for all permutations σ

2. (Riemann series theorem) If $\sum_{k=0}^{\infty} a_k$ is conditionally convergent, for any $A \in \mathbb{R} \cup \{-\infty, +\infty\}$, there exists a permutation σ such that

$$\sum_{k=0}^{\infty} a_{\sigma(k)} = A$$

Interpretation: For a conditionally convergent sum, you can select a permutation such that the sum equals any value you want. This also offers a way to define the reals via results of sums of permutations.

Application to Expected Value:

$$\sum_{k=0}^{\infty} p_k \cdot \mathbf{1}_{[x_k, +\infty)}(x)$$

where "all x_k's are created equal" (x_k is unordered). Then for an absolutely convergent sum

$$\mathbb{E}(X) = \sum_{k=0}^{\infty} x_k \cdot p_k = \sum_{k=0}^{\infty} x_{\sigma(k)} \cdot p_{\sigma(k)}$$

Lecture 19, March 10:

Part I - Review

Definition: Let X be a discrete RV with CDF

$$\sum_{k=0}^{K} p_k \cdot \mathbf{1}_{[x_k,+\infty)}(x)$$

1. If $\sum_{k=0}^{K} |x_k| \cdot p_k < \infty$ then $\mathbb{E}(X) = \sum_{k=0}^{K} x_k \cdot p_k$
2. If $\sum_{k=0}^{|} x_k| \cdot p_k = \infty$ then $\mathbb{E}(X)$ does not exist

Question: In the summation $\sum_{k=0}^{\infty} x_k \cdot p_k = x_0p_0 + x_1p_1 + ...$, does the order matter?

Answer: No! So long as

$$\sum_{k=0}^{\infty} x_k \cdot p_k = \sum_{k=0}^{\infty} x_{\sigma(k)} \cdot p_{\sigma(k)}$$

for all permutations σ

1. If $\sum_{k=0}^{\infty} |x_k| \cdot p_k < \infty$ then the permutation invariance is true and $\sum_{k=0}^{K} x_k \cdot p_k < \infty$ is *absolutely convergent*
2. If $\sum_{k=0}^{\infty} |x_k| \cdot p_k = \infty$ then the permutation invariance is false

Definition: Let X be a continuous RV with PDF $p_X(x)$

1. If $\int_{-\infty}^{\infty} |x| \cdot p_X(x)\, dx < \infty$, then $\mathbb{E}X = \int_{-\infty}^{\infty} x \cdot p_X(x)\, dx$
2. If $\int_{-\infty}^{\infty} |x| \cdot p_X(x)\, dx < \infty$, then $\mathbb{E}X$ does not exist

Remark: the condition "$\int_{-\infty}^{\infty} |x| \cdot p_X(x)\ dx < \infty$" is more subtle and requires Lebesgue integrals

Example: Let X be a continuous RV with PDF of the Cauchy-Lorentz distribution

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

Claim: $\mathbb{E}X$ does not exist Proof: HW 5

Transformations of RV

Motivation:

Let Ω be the studetns taking APMA 1655. $X(\omega)$ is the score student ω gets in the final exam. The professor wants to curve the scores such that

$$Y(\omega) = \min\{X(\omega)^2, 100\}$$

$$g(x) = \min\{X^2, 100\} \implies Y(\omega) = g(X(\omega))$$

More rigorously: Suppose we are given

- RV X: $\Omega \to \mathbb{R}$
- a function g: $\mathbb{R} \to \mathbb{R}$
- $\Omega \xrightarrow{X} \mathbb{R} \xrightarrow{g} \mathbb{R}$
- $\omega \mapsto X(\omega) \mapsto g(X(\omega))$
- $g(X)$ is a random variables $\omega \mapsto g(X(\omega))$

What is $\mathbb{E}[g(X)]$?

Definition: Suppose X and g are given.

1. Suppose X is discrete and has CDF $\sum_{k=0}^{K} p_k \cdot \mathbf{1}_{[x_k,\infty)}$. If $\sum_{k=0}^{K} |g(x_k)| \cdot p_k < \infty$ then
$$\mathbb{E}[g(X)] = \sum_{k=0}^{K} g(x_k) \cdot p_k$$
Otherwise, $\mathbb{E}[g(X)]$ does not exist.
2. Suppose X is continuous and has PDF $p_X(x)$. If $\int_{-\infty}^{\infty} |g(x)| \cdot p_X(x)\ dx < \infty$ then $\mathbb{E}[g(x)] = \int_{-\infty}^{\infty} g(x) \cdot p_X(x)\ dx$ Otherwise, $\mathbb{E}[g(x)] does not exist$

Definition: Let X be a RV with CDF

$$F_X(x) = p \cdot F_Z(x) + (1-p) \cdot F_W(x)$$

where

1. $0 \le 0 \le 1$
2. Z is a discrete RV with CDF F_Z

3. W is a continuous RV with CDF F_W

Then

$$\mathbb{E}X = p \cdot \mathbb{E}Z + (1-p)\mathbb{E}W$$

(assuming $\mathbb{E}Z$ and $\mathbb{E}W$ exist)

Example:

$$X = YZ + (1-Y)W$$

where

$$X \sim \text{Bernoulli}(\frac{1}{3})$$
$$Y \sim \text{Pois}(\lambda)$$
$$Z \sim N(1000, 1)$$

and all three are independent.

Then

$$\mathbb{E}Z = \lambda \quad \mathbb{E}W = 1000$$

and

$$F_X(x) = \frac{1}{3}F_Z(x) + \frac{2}{3}F_W(x)$$

so

$$\mathbb{E}X = \frac{1}{3}\lambda + \frac{2}{3}(1000)$$

Lecture March 13: Variance

Part I - Motivation

- X follows a distribution
- The distribution generates numbers $X_1, ...X_n$
- $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \approx \mathbb{E}X$
- $e_n = |\bar{X}_n - \mathbb{E}X|$ (that is, the error is the distance between the average of n numbers and the expected value)

Part II - Exploring error

Simulation 1: Generate $X_1, ...X_n$ from Bernoulli(p) *Claim:*

$$e_n = |\bar{X}_n - p| \overset{\text{likely}}{<} \sqrt{p(1-p)\frac{2\log(\log n)}{n}}$$

Proof: Derive from Brownian Motion

Simulation 2: Generate $X_1, ...X_n$ from Pois(λ) *Claim:*

$$e_n = |\bar{X}_n - \lambda| \overset{\text{likely}}{<} \sqrt{\lambda \cdot \frac{2\log(\log n)}{n}}$$

Conjecture: Generate $X_1, ..., X_n$ from a distribution. Then

$$e_n = |\bar{X}_n - \mathbb{E}X| \overset{\text{likely}}{<} \sqrt{V \cdot \frac{2\log(\log n)}{n}}$$

where V is the variance and both the expected value and variance exist for the given distribution

Remark: this is known as the "law of the iterated logarithm"

Part III - Variance

Let X be a random variable whose expected value $\mathbb{E}X$ exists. We define a function

$$g(x) = (x - \mathbb{E}X)^2$$

Then the *variance of X* is

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

or the "expected squared deviation from $\mathbb{E}X$

Lecture March 15

Part I - Review of $\mathbb{E}[g(x)]$

If X is a discrete RV and has CDF $\sum_{k=0}^{K} p_k \cdot \mathbf{1}_{[x_k,\infty)}(x)$ then (if it exists),

$$\mathbb{E}[g(x)] = \sum_{k=0}^{K} g(x_k) \cdot p_k$$

If X is a continuous RV and has PDF $p_X(x)$ then (if it exists)

$$\int_{-\infty}^{\infty} g(x) \cdot p_X(x)\, dx$$

Part II - Review of Variance

Definition: Let X be a RV whose expected value exists. Then the variance of X (if it exists) is

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

Remark:

1. If $\mathbb{E}[(X - \mathbb{E}X)^2] = \infty$, we say the variance does not exist
2. IF X is discrete and has CDF $\sum_{k=0}^{K} p_k \cdot \mathbf{1}_{[x_k,\infty)}(x)$,

$$\mathrm{Var}(X) = \sum_{k=0}^{k} (x_k - \mathbb{E}X)^2 \cdot p_k$$

3. If X is continuous and has PDF $p_X(x)$,

$$\mathrm{Var}(X) = \int_{-\infty}^{\infty} (x - \mathbb{E}X)^2 \cdot p_X(x)\, dx$$

Part III - Properties of Expected Values

Let X be a continuous or discrete RV.

1. For a constant c, $\mathbb{E}c = c$
2. Let a and b be two constants, then

$$\mathbb{E}[aX + b] = a(\mathbb{E}X) + b$$

3. Let $g_1(x), g_2(x), ..g_J(x)$ be functions. Suppose $\mathbb{E}[g_j](x)$ exists for all $k = 1, 2, ...J$. Then

$$\mathbb{E}[g_1(X) + g_2(X) + ... + g_J(x)] = \mathbb{E}[\sum_{k=1}^{J} g_j(x)] = \sum_{j=1}^{J} \mathbb{E}[g_J(X)]$$

exists and shows that the expected value is linear

Proof: all the properties derive from the properties of the summation and the integral. Details are homework.

Part IV - Properties of Variance

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

$$\begin{aligned}\mathbb{E}[(X - \mathbb{E}X)^2] &= \mathbb{E}[X^2 - 2(\mathbb{E}X) \cdot X + (\mathbb{E}X)^2] \\ &= \mathbb{E}X^2 + \mathbb{E}[-2(\mathbb{E}X) \cdot X] + (\mathbb{E}X)^2 \\ &= \mathbb{E}X^2 - 2(\mathbb{E}X)(\mathbb{E}X) + (\mathbb{E}X)^2 \\ &= \mathbb{E}(X^2) - (\mathbb{E}X^2) \quad \blacksquare\end{aligned}$$

2. $\text{Var}(aX + b) = a^2\text{Var}(X)$

   Proof: HW

3. For any constant x, $\text{Var}(c) = 0$

   Proof: $\text{Var}(c) = \mathbb{E}(c^2) - (\mathbb{E}c)^2 = c^2 - c^2 = 0$

4. Let X be a RV. If $\text{Var}(X) = 0$ then there exists a constant c such that $\mathbb{P}(\{\omega \in \Omega : X(\omega) = c\}) = 1$

   Proof: See real analysis.

Lecture March 17:

Part I - Review:

If $X(\omega) = c$ for all $\omega \in \Omega$ for some random variable X and constant c,

$$\text{Var}(X) = 0$$

Additionally, if $\text{Var}(X) = 0$ then $\exists c$ such that

$$\mathbb{P}(\{\omega \in \Omega : X(\omega) - c\}) = 1$$

Part II - Law of Large Numbers (LLN)

A sloppy version: Let X be a RV that follows a distribution which generates random numbers $X_1, X_2, ..., X_n$. Then if $\mathbb{E}X$ exists

$$\bar{X_n} = \frac{X_1 + X_2 + ... + X_n}{n} \approx \mathbb{E}X$$

e.g. if

- $X \sim \text{Bernoulli}(p) \implies \mathbb{E}X = p$
- $X \sim \text{Pois}(\lambda) \implies \mathbb{E}X = \lambda$

However, this raises some questions: What does $\approx$ mean? In addition to the existence criteria, what other conditions are necessary?

Part III - Preparations for the Formal

Definition: Let $\{X_i\}_{i=1}^{\infty}$ be an infinitely long sequence of RVs defined on $(\Omega, \mathbb{P})$. We say $X_1, X_2, ...$ are independent if

$$\mathbb{P}(\{\omega \in \Omega : X_i \in A_i \quad \forall i = 1, 2, ..., n\}) = \prod_{i=1}^{n} \mathbb{P}(\{\omega \in \Omega : X_i(\omega \in A_i)\})$$

for any positive integer n, and any subsets $A_1, ...A_n \subset \mathbb{R}$

Definition: Let $\{X_i\}_{i=1}^{\infty}$ be an infinitely long sequence of RVs defined on $(\Omega, \mathbb{P})$. We say $X_1, X_2, ...$ are independently and identically distributed if

1. $X_1, X_2, ...$ are independent
2. $X_1, X_2, ...$ share the same CDF, i.e. $F_{X_1} = F_{X_2} = F_{X_3} = ...$

Part IV - The Law of Large Numbers

Theorem: Let $\{X_i\}_{i=1}^{\infty}$ be an infinitely long sequence of RVs defined on $(\Omega, \mathbb{P})$. Suppose $X_1, X_2, ...$ are independently and identically distributed and

$$A\{\omega \in \Omega : \lim_{n\to\infty} \frac{X_1(\omega) + X_2(\omega) + ... + X_n(\omega)}{n} = \lim_{n\to\infty} \bar{X}_n = \mathbb{E}X_1\}$$

where $\mathbb{E}X_1 = \mathbb{E}X_2 = ...$ because the CDFs are equal. Then

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

Example: Flip a fair coin infinitely many times.

- An outcome ω is then an infinitely long sequence of H and T.
- $\Omega = \{\omega = (\omega^{(1)}, \omega^{(2)}, \omega^{(3)}, ...) : \text{each } \omega^{(i)} \text{ is either H or T}\}$

- $$X_i(\omega) = \begin{cases} 1 & \omega^{(i)} = H \\ 0 & \omega^{(i)} = T \end{cases}$$

- $\exists \mathbb{P}$ such that

$$\mathbb{P}(\{\omega \in \Omega : X_i(\omega) = 1\}) = \mathbb{P}(\{\omega \in \Omega : X_i(\omega) = 0\}) = \frac{1}{2}$$

  and $X_1, X_2, ...$ are independent which implies $X_1, X_2, ...$ are identically and independently distributed.

Notice that

$$A = \{\omega \in \Omega : \lim_{n\to\infty} \bar{X}_n = \frac{1}{2}\}$$

And by LLN, $\mathbb{P}(A) = 1$ but $A \neq \Omega$. Note: To see why, imagine a sequence of infinite tails such that

$$X(\omega^*) = 0 \implies \lim_{n\to\infty} \bar{X}_n(\omega^*) = 0 \implies \omega^* \notin A$$

Lecture March 20

Part I - Review of the Law of Large Numbers

Theorem: Let $\{X_i\}_{i=1}^{\infty}$ be an infinitely long sequence of RVs defined on $(\Omega, \mathbb{P})$. Suppose the random variables are independently and identically distributed (they share the same CDF) and $\mathbb{E}X_1$ exists. Then

$$\mathbb{P}\left(\{\omega \in \Omega : \lim_{n\to\infty} \bar{X}_n = \mathbb{E}X_1\}\right) = 1$$

Note that because the CDFs are the same,

$$\mathbb{E}X_1 = \mathbb{E}X_2 = \mathbb{E}X_3 = ...$$

Additionally, denote

$$\bar{X}_n = \frac{X_1(\omega) + ... + X_n(\omega)}{n}$$

as the *sample average* which determines $\mathbb{E}X_1$ (the *population average*).

Then, the law of large numbers can be expressed "the sample average converges to the population average with probability 1"

Note that the independence condition is crucial and without it, the LLN is not necessarily true.

Example: X is a RV on $(\Omega, \mathbb{P})$ where $X \sim$ Bernoulli$(\frac{1}{2})$ For each $i = 1, 2, ...$

$$X_i(\omega) = X(\omega) \quad \forall \omega \in \Omega$$

$$X_1 = X_2 = X_3 = ... = X \implies X_1, X_2, ... \text{ are not independent}$$

so

$$A = \{\omega \in \Omega : \bar{X}_n = \frac{1}{2}\} = \{\omega \in \Omega : X(\omega) = \frac{1}{2}\}$$

The LLN anticipates $\mathbb{P}(A) = 1$ however

$$\mathbb{P}(A) = \mathbb{P}(X = \frac{1}{2}) = 0$$

so the LLN does not hold

Part II - Monte Carlo Integration

Motivation: Integrals are HARD.

$$I = \int_0^1 \cos^{-1}\left(\frac{\cos(\frac{\pi}{2}X)}{1 + 2\cos(\frac{\pi}{2}x)}\right) dx = \frac{5\pi}{12}$$

But we can approximate it with the LLN! First, let $U \sim$ Unif$(0, 1)$ whose PDF is $\mathbf{1}_{[0,1)}(x)$ Denote the crazy integrand $g(x)$ then

$$I = \mathbb{E}[g(U)]$$

Theorem: Generalized Law of Large Numbers Let $\{X_i\}_{i=1}^{\infty}$ be an infinitely long sequence of RVs defined on $(\Omega, \mathbb{P})$. Let $g(x)$ be a continuous function. Suppose the random variables are independently and identically distributed (they share the same CDF) and $\mathbb{E}[g(X_1)]$ exists. Then

$$\mathbb{P}\left(\{\omega \in \Omega : \lim_{n\to\infty} \frac{g(X_1(\omega)) + ... + g(X_n(\omega))}{n} = \mathbb{E}[g(X_1)]\}\right) = 1$$

So applied to the example above using this generalized LLN, we can generate $X_1(\omega), X_2(\omega), ... \overset{iid}{\sim}$ Unif$(0, 1)$. Then the sample average of $g(X_i)$ will approximate the value of the integral for enough random numbers.

Lecture March 22: Monte Carlo Integration

Part I - Introduction

Let X be a continuous RV with PDF $p_X(x)$ and $g(x)$ is a real-values function. Then

$$E[g(x)] = \int_{-\infty}^{\infty} g(x) \cdot p_X(x)\, dx$$

and

$$\text{Var}[g(x)] = \int_{-\infty}^{\infty} (g(x) - \mathbb{E}[g(x)])^2 \cdot p_X\, dx$$

(if the expected exists).

Part II - An Example

Let $X \sum \text{Unif}(0,1)$ and

$$g(x) = cos^{-1}\left(\frac{\cos(\frac{\pi}{2}X)}{1 + 2\cos(\frac{\pi}{2}x)}\right) \qquad 0 < x < 1$$

and the PDF of X is

$$p_X(x) = \mathbf{1}_{(0,1)}(x) = \begin{cases} 1 & 0 < x < 1 \\ 0 & \text{otherwise} \end{cases}$$

So

$$\mathbb{E}[g(X)] = \int_{-\infty}^{\infty} g(x) \cdot \mathbf{1}_{(0,1)}(x)\, dx = \int_0^1 g(x)\, dx = \int_0^1 cos^{-1}\left(\frac{\cos(\frac{\pi}{2}X)}{1 + 2\cos(\frac{\pi}{2}x)}\right)\, dx = \frac{5\pi}{12}$$

This is a VERY HARD integral to calculate but we can approximate it with Monte Carlo integration.

Using the Generalized Law of Large Numbers, we define an infinite sequence of independently and identically distributed RVs,

$$X_1, X_2, X_3, ... \sim \text{Unif}(0,1)$$

so

$$\frac{g(X_1) + g(X_2) + ...g(X_n)}{n} \approx \mathbb{E}[g(X_1)] = \int_0^1 \cos^{-1}\left(\frac{\cos(\frac{\pi}{2}X)}{1 + 2\cos(\frac{\pi}{2}x)}\right)\, dx$$

This is especially useful for higher dimensional integrals as it allows us to take the running sum of only n numbers rather than say a 100-dimensional Riemann sum.

Note that for other integrals with bounds (a, b) rather than $(0, 1)$ we can still use the same method by defining a new random variable from $U \sim \text{Unif}(0, 1)$, where:

$$X = a + (b - a) \cdot U \sim \text{Unif}(a, b)$$

and then calculating $\bar{X}_n$

Also note that there will be some error between the approximated value and the true value

$$e_n(\omega) = \frac{g(X_1(\omega)) + ... + g(X_n(\omega))}{n} - \mathbb{E}[g(X_1)]$$

and by *The Law of the Iterated Logarithm* $|e_n(\omega)|$ is likely to be smaller than

$$\sqrt{\text{Var}[g(X_1)] \cdot \frac{2\log(\log n)}{n}}$$

12 Lecture March 24:

Part I - Law of the Iterated Logarithm

Theorem: Let $X_1, X_2, X_3, ...$ be identically and independently distributed RVs defined on $(\Omega, \mathbb{P})$ Suppose $\mathbb{E}X_1$ and $\text{Var}(X_1)$ exist. Then,

$$\mathbb{P}\left(\{\omega \in \Omega : \lim_{m\to\infty}\left[\sup_{n\geq m}\left(\frac{e_n(\omega)}{\sqrt{\text{Var}X_i\frac{2\log(\log n)}{n}}}\right)\right] = 1\}\right) = 1$$

Heuristically, when n is large,

$$\boxed{\begin{aligned}&\mathbb{P}(\{\omega \in \Omega : |e_n(\omega)| \leq \sqrt{\text{Var}X_i\frac{2\log(\log(n))}{n}}\}) \approx 1\\ &\mathbb{P}(\{\omega \in \Omega : |e_n(\omega)| > \sqrt{\text{Var}X_i\frac{2\log(\log(n))}{n}}\}) \approx 0\end{aligned}}$$

Part II - Example

$$U_1, U_2, U_3, ... \overset{iid}{\sim} \text{Unif}(0,1)$$
$$g(x) = \cos^{-1}\left(\frac{\cos(\frac{\pi}{2}x)}{1+2\cos(\frac{\pi}{2}x)}\right)$$

By the generalized LLN,

$$\frac{g(U_1(\omega)) + ... + g(U_n(\omega))}{n} \approx \mathbb{E}[g(U)] = \int_0^1 g(x)\, dx = \frac{5\pi}{12}$$

(By Monte Carlo simulation) Then,

$$X_i(\omega) = g(U_i(\omega)) \quad \forall i, \omega$$
$$\implies \bar{X}_n(\omega) \approx \mathbb{E}X_1 = \mathbb{E}[g(U)]$$
$$e_n(\omega) = \bar{X}_n(\omega) - \mathbb{E}X_1$$

Let's say we want the error to be no higher than 10^{-5} so

$$|e_n(\omega)| \leq \sqrt{\text{Var}(X_i) \cdot \frac{2\log(\log n)}{n}} \leq 10^{-5}$$

Using another Monte carlo simulation to calculate $\text{Var}(X_i) \approx 0.007556$, we then may choose a large n such that the inequality is true.

Part III - Random Walks

Definition: Let $\{X_i\}_{i=1}^{\infty}$ be a sequence of RVs defined on $(\Omega, \mathbb{P})$ and $X_1, X_2, ...$ are iid. For each positive n, we define

$$S_n(\omega) = X_1(\omega) + ... + X_n(\omega)$$

The sequence $\{S_n(\omega)\}_{n=1}^{\infty}$ is a *random walk*

Example 1: Let $X_1, X_2, ...$ iid that satisfy

$$\mathbb{P}(X_1 = -1) = \mathbb{P}(X_2 = 1) = \frac{1}{2}$$

then $\{S_n(\omega)\}_{n=1}^{\infty}$ is the "1-dim simple RW"

Example 2: $X_1, X_2, ... \overset{iid}{\sim} N(0, \sigma^2)$ for some fixed σ. Then $\{S_n(\omega)\}$ is a discrete-time Brownian motion (introduced by Einstein in 1905).

Lecture April 3: The Central Limit Theorem (CLT)

Part I - Introduction

Let $\{X_i\}_{i=1}^{\infty}$ be a sequence of iid RV on $(\Omega, \mathbb{P})$. If for all n,

$$S_n(\omega) = X_1(\omega) + ... + X_n(\omega)$$

then $\{S_n(\omega)\}_{n=1}^{\infty}$ is a random walk. And

$$\bar{X}_n(\omega) = \frac{S_n(\omega)}{n} \approx \mathbb{E}X_1 \quad (n \to \infty)$$

So

$$e_n(\omega) = \bar{X}_n(\omega) - \mathbb{E}X_1$$

where

$$\mathbb{P}(\{\omega \in \Omega : |e_n(\omega)| \leq \sqrt{\mathrm{Var}X_1 \cdot \frac{2 \ln \ln n}{n}}\}) \approx 1$$

Preview of the CTL:

$$\sqrt{n} \cdot e_n(\omega) \dot{\sim} N(0, \mathrm{Var}X_1)$$

which means that the product "asymptotically follows" the normal distribution

Part II - Applications of the Central Limit Theorem

1. A different way of quantifying $e_n(\omega)$
2. Foundation of "hypothesis testing" and "confidence intervals"

Part III - The Theorem

Theorem: Let $\{X_i\}_{i=1}^{\infty}$ be a sequence of iid RVs on $(\Omega, \mathbb{P})$. Suppose $\mathbb{E}X_i$ and $\mathrm{Var}\, X_i$ exist. Define a sequence of random variables $\{G_n\}_{n=1}^{\infty}$ so that

$$G_n(\omega) := \sqrt{n} \cdot e_n(\omega) = \sqrt{n} \cdot (\bar{X}_n(\omega) - \mathbb{E}X_1)$$

Then, the CDF of G_n converges to the CDF of $N(0, \mathrm{Var}\ X_1)$ as $n \to \infty$, i.e.,

$$\lim_{n \to \infty} \mathbb{P}(\omega \in \Omega : G_n(\omega) \leq x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi \cdot \mathrm{Var}\ X_1}} \cdot \exp\left(-\frac{t^2}{2\mathrm{Var}\ X_1}\right) \, dt$$

Briefly,

$$G_n(\omega) = \sqrt{n} \cdot e_n(\omega) \dot{\sim} N(0, \text{Var } X_1)$$

Corollary: Suppose we are under the same conditions as the CLT. Then the CDF of $\frac{G_n(\omega)}{\sqrt{\text{Var } X_1}}$ converges to the CDF of $N(0,1)$, i.e.

$$\lim_{n\to\infty} \mathbb{P}(\omega \in \Omega : \frac{G_n(\omega)}{\sqrt{\text{Var } X_1}} \leq x) = \int_{-\infty}^{x} = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{t^2}{2}\right)\, dt$$

Briefly,

$$\sqrt{n} \cdot \frac{\bar{X}_n(\omega) - \mathbb{E}X_1}{\sqrt{\text{Var } X_1}} \dot{\sim} N(0,1)$$

Part IV - Error Bounds

Let $\{X_i\}_{i=1}^{\infty}$ be a sequence of iid RVs on $(\Omega, \mathbb{P})$. Suppose $\mathbb{E}X_i$ and $\text{Var } X_i$ exist.

The LIL implies

$$\lim_{n\to\infty} \mathbb{P}\left(\{\omega \in \Omega : |e_n(\omega)| \leq \sqrt{\text{Var } X_1 \cdot \frac{2log(\log n)}{n}}\}\right) = 1$$

or in other words, "with probability (confidence) around 100%, $|e_n(\omega)| \leq \sqrt{2\log(\log n)} \cdot \sqrt{\frac{\text{Var } X_1}{n}}$"

But from the CLT,

$$\begin{aligned}
&\mathbb{P}\left(\{\omega \in \Omega : |e_n(\omega)| \leq z \cdot \sqrt{\frac{\text{Var } X_1}{n}}\}\right) \\
&= \mathbb{P}\left(\{\omega \in \Omega : -z \leq \sqrt{n} \cdot \frac{e_n(\omega)}{\sqrt{\text{Var } X_1}} \leq z\}\right) \\
&= \mathbb{P}\left(\{\omega \in \Omega : \sqrt{n} \cdot \frac{e_n(\omega)}{\sqrt{\text{Var } X_1}} \leq z\}\right) - \mathbb{P}\left(\{\omega \in \Omega : \sqrt{n} \cdot \frac{e_n(\omega)}{\sqrt{\text{Var } X_1}} \leq -z\}\right) \\
&\approx \Phi(z) - \Phi(-z) \\
&= 2\Phi(z) - 1 \quad (z > 0)
\end{aligned}$$

Where Φ is the CDF of $N(0,1)$.

Now let z^* denote the positive real number such that $\Phi(z^*) = 0.975$ so

$$\mathbb{P}\left(\{\omega \in \Omega : |e_n(\omega)| \leq z^* \cdot \sqrt{\frac{\text{Var } X_1}{n}}\}\right) \approx 2\Phi(z^*) - 1 = 0.95$$

Note: Generally, you may choose z^* such that $\Phi(z^*) = 1-\alpha/2$ so $2\Phi(z^*)-1 = 1-\alpha$. Then z^* is called the $1-\alpha/2$ quantile of $N(0,1)$ and must be computed using a computer. For example, the 0.975 quantile is 1.959964.

This means that using the CLT we have "with probability (confidence) 95% we have

$$|e_n(\omega)| \leq z^* \cdot \sqrt{\frac{\text{Var } X_1}{n}} \approx 1.96 \cdot \sqrt{\frac{\text{Var } X_1}{n}}$$

"

Conclusion: Using the CLT we can establish much tighter error bounds for large n at the cost of only 5% confidence.

13 Lecture April 7

Definition: Ler $G_1, G_2, ...$ be a sequence of RVs. We say G_n *converge weakly* to a continuous RV G is

$$\lim_{n\to\infty} F_{G_n}(x) = F_G(x)$$

for all real x.

Lecture April 7: Proof of the CLT

Part I - Weak Convergence

Definition: A sequence $G_1, ...G_n$ of RVs *converges weakly* to a continuous RV G if

$$\lim_{n\to\infty} F_{G_n}(x) = F_G(x) \quad \forall x \in \mathbb{R}$$

This is briefly denoted $G_n \xrightarrow{w} G$.

Remarks:

1. G_n is sead to converge "stronly" to G if

$$\lim_{n\to\infty} G_n(\omega) = G(\omega) \quad \forall \omega \in \Omega$$

2. Let $X_1, ..., X_n$ be iid Rvs whose expected value and variances exist. Then

$$G_n(\omega) = \sqrt{n} \cdot e_n(\omega)$$

and

$$G \sim N(0, \text{Var } X_1)$$

Thus, the CLT stated

$$\lim_{n\to\infty} F_{G_n}(x) = \text{CDF of } N(0, \text{Var } X_1) = F_G(x)$$

can also be written

$$G_n \xrightarrow{w} G \sim N(0, \text{Var } X_1)$$

Part II - Moment Generating Functions

Definition: Let X be a RV. Then

$$M_X(t) = \mathbb{E}[e^{t \cdot X}]$$

is the *moment-generating function of X*, provided the expected value exists for all $t \in \mathbb{R}$

Note that the name "moment-generating" comes from the Taylor Expansion:

$$e^{tX} = 1 + tX + \frac{t^2X^2}{2!} + \frac{t^3X^3}{3!} + ...$$

$$\begin{aligned}
\mathbb{E}[e^{tX}] &= \mathbb{E}\left[\sum_{n=0}^{\infty} \frac{t^n X^n}{n!}\right] \\
&= \sum_{n=0}^{\infty} \mathbb{E}\left[\frac{t^n X^n}{n!}\right] \quad \text{(proved in homework)} \\
&= 1 + t\mathbb{E}X + \frac{t^2}{2!}\mathbb{E}X^2 + \frac{t^3}{3!} + ... \\
\frac{d}{dt}M_X(t) &= \mathbb{E}X + t\mathbb{E}X^2 + \frac{t^2}{2!}\mathbb{E}X^3 + ...
\end{aligned}$$

Then notice that

$$\frac{d}{dt}M_X(0) = \mathbb{E}X$$

which is called the "first moment of X".

Taking the second derivative, we get

$$\frac{d^2}{dt^2}M_X(0) = \mathbb{E}X^2$$

which is the "second moment of X"

Generally,

$$\frac{d^k}{dt^k}M_X(0) = \mathbb{E}X^k$$

gives the "k-th moment of X"

Theorem: Let $G, G_1, G_2, ...$ be a sequence of RVs. If

$$\lim_{n\to\infty} M_{G_n}(t) = M_G(t)$$

then

$$G_n \xrightarrow{w} G$$

Proof: Omitted

Theorem: Let $X, X_1, ..., X_n$ be RVs.

$$S_n(\omega) := X_1(\omega) + X_2(\omega) + ... + X_n(\omega) = \sum_{i=1}^{n} X_i(\omega)$$

1. If $X_1, ..., X_n$ are independent

$$M_{S_n}(t) = \prod_{i=1}^{n} M_{X_i}(t)$$

2. If $X_1, ..., X_n$ are also identically distributed

$$M_{S_n}(t) = (M_{X_1}(t))^n$$

(because they share the same moment generating function by iid-ness)

Part III - Some Lemmas

Lemma 1: Let $\{C_n\}_{n=1}^{\infty}$ be a sequence of real numbers satisfying

$$\lim_{n\to\infty} C_n = 0$$

If

$$\lim_{n\to\infty} n \cdot C_n = \lambda$$

then

$$\lim_{n\to\infty} (1 + C_n)^n = e^{\lambda}$$

Proof: See lecture notes

Lemma 2: If $G \sim N(0, \sigma^2)$ then

$$M_G(t) = e^{\frac{t^2\sigma^2}{2}}$$

Proof: Homework

Part IV - Proof of the CLT

$$\begin{aligned} G_n &= \sqrt{n}\left[\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) - \mathbb{E}X_1\right] \\ &= \sqrt{n}\cdot\left(\frac{1}{n}\sum_{n=1}^{n} X_i\right) - \frac{1}{n}n\mathbb{E}X_1 \\ \\ &= \sqrt{n}\cdot\frac{1}{n}\left[\sum_{i=1}^{n} X_1 - \sum_{i=1}^{n}\mathbb{E}X_1\right] \\ &= \sqrt{n}\cdot\frac{1}{n}\left[\sum_{i=1}^{n}(X_i - \mathbb{E}X_1)\right] \end{aligned}$$

then

$$\begin{aligned}
M_{G_n}(t) &= \mathbb{E}[e^{tG_n}] \\
&= \mathbb{E}[e^{\sum_{i=1}^n \left(\frac{t}{\sqrt{n}}(X_i - \mathbb{E}X_1)\right)}] \\
&= \left(\mathbb{E}[e^{\frac{t}{\sqrt{n}}}(X_1 - \mathbb{E}X_1)]\right)^n \\
&= \left(\mathbb{E}\left[1 + \frac{t}{\sqrt{n}}(X_1 - \mathbb{E}X_1) + \frac{t^2}{2n}(X_1 - \mathbb{E}X_1)^2 + \sum_{k=3}^{\infty} \frac{t^k}{k \cdot n^{k/2}}(X_1 - \mathbb{E}X)^k\right]\right)^n \\
&\approx (1 + \frac{t^2}{2n}\text{Var } X_1)^n
\end{aligned}$$

So

$$M_{G_n}(t) \approx (1 + c_n)^n \to e^{c_n} = \text{MGF of } N(0, \text{Var } X_1)$$

where $c_n = \frac{t^2}{2n}\text{Var } X_1$

Lecture April 10: Random Vectors

Definitions:

1. A (column) vector $\vec{X} = (X_1, X_2, ..., X_n)^T$ is called a random vector defined on $(\Omega, \mathbb{P})$ if each of its components is a RV defined on $(\Omega, \mathbb{P})$
2. The CDF of $\vec{X}$ is an n-variable function

$$\begin{aligned}
F_{\vec{X}}(x_1, x_2, ..., x_n) &= \mathbb{P}(\{\omega \in \Omega : X_1(\omega) \leq x_1, ..., X_n(\omega) \leq X_n\}) \\
&= \mathbb{P}\left(\bigcap_{i=1}^n \{\omega \in \Omega : X_i(\omega) \leq x_i\}\right)
\end{aligned}$$

3. $\vec{X} = (X_1, ..., X_n)^T$ is said to be a continuous vector if $F_{\vec{X}}$ is differentiable

The PDF of $\vec{X}$ is defined by

$$p_{\vec{X}}(x_1, ..., x_n) = \frac{\partial}{\partial x_1}\frac{\partial}{\partial x_2}...\frac{\partial}{\partial x_n}F_{\vec{X}}(x_1, x_2, ..., x_n)$$

Example: For $n = 2$,

$$p_{\vec{X}}(x_1, x_2) = \frac{\partial}{\partial x_1}\left(\frac{\partial}{\partial x_2}F_{\vec{X}}(x_1, x_2)\right)$$

Definition: Let $\vec{X} = (X_1, X_2, ..., X_n)$ be a continuous random vector with PDF $p_{\vec{X}}(x_1, x_2, ..., x_n)$. Suppose $g(x_1, x_2, ..., x_n) = g(\vec{x})$ is an n-variable function. Then

$$\mathbb{E}[g(\vec{X})] = \underbrace{\int_{-\infty}^{\infty} ... \int_{-\infty}^{\infty}}_{n \text{ integrals}} g(x_1, ..., x_n) \cdot p_{\vec{X}}(x_1, ..., x_n)\, dx_1 ... dx_n$$

if

$$\int_{\mathbb{R}^n} |g(x_1, ..., x_n)| \cdot p_{\vec{X}}(x_1, ..., x_n)\, dx_1 ... dx_n < \infty.$$

and $\mathbb{E}[g(\vec{X})]$ is a scalar.

Theorem: Let $\vec{X} = (X_1, ..., X_n)$. If $X_1, ..., X_n$ are independent, then

$$F_{\vec{X}}(x_1, ..., x_n) = \prod_{i=1}^{n} F_{X_i}(x_i).$$

Proof: Omitted

Furthermore, if $\vec{X}$ is also a continuous random vector.

$$p_{\vec{X}}(x_1, ..., x_n) = \prod_{i=1}^{n} p_{X_i}(x_i)$$

Proof:

$$\begin{aligned} p_{\vec{X}}(x_1, ..., x_n) &= \frac{\partial}{\partial x_1} ... \frac{\partial}{\partial x_n} F_X(x_1, ..., x_n) \\ &= \frac{\partial}{\partial x_1} ... \frac{\partial}{\partial x_n} \prod_{i=1}^{n} F_{X_i}(x_i) \\ &= \left(\prod_{i=1}^{n} \frac{\partial}{\partial x_i} F_{X_i}(x) \right) \\ &= \prod_{i=1}^{n} p_{X_i}(x_i) \quad \blacksquare \end{aligned}$$

Theorem: $X_1, ..., X_n$ are iid RVs. Define $S_n = \sum_{i=1}^{n} X_i$. Then

$$M_{S_n}(t) = (M_{X_1}(t))^n$$

Proof: $\vec{X} = (X_1, ..., X_n)$. To simplify the proof (though it is not necessary for the result), assume $\vec{X}$ is continuous.

$$\begin{aligned}
M_{S_n}(t) &= \mathbb{E}[e^{tS_n}] \\
&= \mathbb{E}\left[e^{tX_1+tX_2+...+tX_n}\right] \\
&= \mathbb{E}\left[\prod_{i=1}^{n} e^{tX_i}\right] \\
&= \int_{\mathbb{R}^n} e^{tX_i} \cdot p_X(x_1, ..., x_n)\, dx_1 ... dx_n \\
&= \int_{\mathbb{R}^n} \left(\prod_{i=1}^{n} e^{tX_i}\right)\left(\prod_{i=1}^{n} p_{X_i}(x_i)\right)\, dx_1 ... dx_n \quad \text{(by independence)} \\
&= \int_{\mathbb{R}^n} \prod_{i=1}^{n} \left(e^{tX_i} \cdot p_X(x_i)\right)\, dx_1 ... dx_n \\
&= \int_{\mathbb{R}^{n-1}} \prod_{i=2}^{n} e^{tX_i} \cdot p_{X_i}(x_i) \left(\int_{-\infty}^{\infty} e^{tX_1} \cdot p_{X_1}(x_1)\, dx_1\right)\, dx_2 ... dx_n \\
&= \int_{\mathbb{R}^{n-1}} \prod_{i=2}^{n} e^{tX_i} \cdot p_{X_i}(x_i)\mathbb{E}[e^{tX_1}]\, dx_2 ... dx_n \\
&= \int_{\mathbb{R}^{n-1}} \prod_{i=2}^{n} e^{tX_i} \cdot p_{X_i}(x_i) M_{X_1}(t)\, dx_2 ... dx_n \\
&= \prod_{n=1}^{n} M_{X_i}(t) \\
&= (M_{X_1}(t))^n \quad \text{by identical distribution} \quad \blacksquare
\end{aligned}$$

Lecture April 12: Proof of the CLT

Let $\{X_i\}_{i=1}^{\infty}$ be a sequence of iid RVs. Their expected values and variances exist.

$$G_{n,\alpha} = n^{\alpha} \cdot (\bar{X}_n - \mathbb{E}X_1) \dot{\sim} \begin{cases} ? & \alpha < \frac{1}{2} \\ N(0, \text{Var } X_2) & \alpha = \frac{1}{2} \\ ? & \alpha > \frac{1}{2} \end{cases}$$

Note: The missing cases will be resolved in Homework.

Then, with $G_n(\omega) = G_{n,\frac{1}{2}}(\omega)$ this gives us the statement of the CLT:

$$G_n \xrightarrow{w} = G \sim N(0, \text{Var } X_1)$$

Proof: It suffices to show that

$$\lim_{n\to\infty} M_{G_n}(t) = M_G(t) = \exp(\frac{t^2}{2} \cdot \text{Var } X_1)$$

Observe:

(**Note:** for different α all steps are the same except the coefficient of the exponent)

$$\begin{aligned}
G_n &= \sqrt{n} \cdot \left[\left(\frac{1}{n} \sum_{i=1}^{n} X_i \right) - \mathbb{E}X_1 \right] \\
&= \frac{1}{\sqrt{n}} \cdot \sum_{i=1}^{n} (X_i - \mathbb{E}X_1) \\
M_{G_n}(t) &= \mathbb{E}[e^{tG_n}] = \mathbb{E}[\exp(\frac{t}{\sqrt{n}} \sum_{i=1}^{n} (X_i - \mathbb{E}X_1))] \\
&= \mathbb{E}\left[\exp(\sum_{i=1}^{n} \frac{t}{\sqrt{n}} (X_i - \mathbb{E}X_1)) \right] \\
&= \mathbb{E}\left[\prod_{i=1}^{n} \exp\left(\frac{t}{\sqrt{n}} (X_i - \mathbb{E}X_1) \right) \right] \quad \text{(by iid)} \\
&= \left(\mathbb{E}\left[\exp\left(\frac{t}{\sqrt{n}} (X_i - \mathbb{E}X_1) \right) \right] \right)^n \\
&= \left(\mathbb{E}\left[1 + \frac{t}{\sqrt{n}} (X_1 - \mathbb{E}X_1) + \frac{t^2}{2n} (X_1 - \mathbb{E}X_1)^2 + \sum_{k=3}^{\infty} \frac{t^k}{k! n^{k/2}} (X_1 - \mathbb{E}X_1)^k \right] \right)^n \\
&= \left(1 + \underbrace{\frac{t^2}{2n} \text{Var } X_1 + \sum_{k=3}^{\infty} \frac{t^k}{k! n^{k/2}} \mathbb{E}[(...)^k]}_{C_n} \right)^n \quad \text{(because } \mathbb{E}[X_1 - \mathbb{E}X_1] = 0)
\end{aligned}$$

Lemma: Let $\{C_n\}_{n=1}^{\infty}$ be a sequence of real numbers. If $c_n \to 0$ and $n \cdot c_n \to \lambda$ when $n \to \infty$, then

$$\lim_{n\to\infty} (1 + C_n)^n = e^{\lambda}$$

In this problem,

- $c_n = 0$
- $n \cdot c_n = \frac{t^2}{2}\text{Var } X_1 + \sum_{k=3}^{\infty} \frac{t^k}{k!\, n^{\frac{k}{2}-1}} \mathbb{E}[(...)^l]$

When $k \geq 3$, $\left(\frac{k}{2} - 1\right) > 0$ so

$$\lim_{n\to\infty} n \cdot C_n = \frac{t^2}{2}\text{Var } X_1 = \lambda$$

Then using the lemma,

$$M_{G_n}(t) = (1 + c_n)^n \overset{n\to\infty}{\longrightarrow} e^{\lambda} = \exp(\frac{t^2}{2}\text{Var } X_1) = M_G(t)$$

This completes the proof.

Statistics

Lecture April 14: Statistical Models

Definition: Suppose we are interested in an experiment. Before performing the experiment we do not know the outcome.

Example:

Experiment: flip a coin n times

$$\Omega = \{\omega = (\omega^{(1),\ldots,\omega^{(n)}}) : \omega^{(i)} = \{H, T\}\}$$

Here, X_1, ..., X_n are functions on Ω

$$X_i = \begin{cases} 1 & \omega^{(i)} = H \\ 0 & \omega^{(i)} = T \end{cases}$$

Performing the experiment is equivalent to fixing an $\omega^* \in \Omega$.

For the fixed ω^*,

$$x_i = X_i(\omega^*) \implies \{X_i\}_{i=1}^n \quad \text{which is the "data"}$$

Lecture April 14: Data, Models, and Inference

Definition: Suppose we are interested in an experiment - trying to observe outcomes. (e.g. flip an unfair coin n times)

Part I - Data

Assume that before performing the experiment, we don't know the outcome so we consider Ω, the set of all possible outcomes.

Let X_1, ..., X_n be known functions defined on Ω. e.g.

$$X_i(\omega) = \begin{cases} 1 & \omega^{(i)} = H \\ 0 & \omega^{(i)} = T \end{cases}$$

Performing the experiment is equivalent to picking and fixing $\omega^* \in \Omega$.

For that fixed ω^*, we can define a collection $\{x_i\}_{i=1}^n$ of *deterministic* numbers by

$$x_i = X_i(\omega^*) \quad i = 1, ..., n$$

Thus, after the experiment, the value of x_i is known and nothing is random.

The collection $\{x_i\}_{i=1}^n$ is called *sample data* and the n is referred to as the *sample size.*

Part II -Models

Let $\mathfrak{F} = \{F_\theta\}_{\theta \in \Theta}$ be a family of real-valued functions satisfying the "CDF properties".

The $\mathfrak{F}$-based model is an assumption:

$\exists \theta^* \in \Theta$ and a probability $\mathbb{P}$ defined on Ω such that $X_1, X_2, ..., X_n \sim^{idd} F_{\theta^*} = \mathbb{P}(\{\omega \in \Omega : X_1(\omega) \leq x\})$

If the assumption is incorrect, we say the model is *misspecified*

Parametric:

1. If Θ is a subset of a finite-dimensional space, then the model is called a parametric model
2. If Θ is a subset of an infinite-dimensional space, then the model is called a nonparametric model

Examples of Parametric models:

- for $\Theta = \mathbb{R}$, $F_\theta = F_{\theta^* \sim N(\theta, 1)}$
- $\Theta = \mathbb{R} \times (0, +\infty)$, $\theta = (\mu, \sigma)$, $F_\theta = F_{X \sim N(\mu, \sigma^2)}$

Examples of nonparametric models: Θ the class of functions on $(0, 1)$ such that

$$\begin{cases} \int_0^1 |f(x)|^2 \, dx < +\infty \\ \int_0^1 |f'(x)|^2 \, dx < +\infty \\ \int_0^1 |f''(x)|^2 \, dx < +\infty \end{cases}$$

which is used in machine learning for "smoothing splines"

Part III - Statistical Inference

Statistical inference is the process of combining probability theory and $\{x_i\}_{i=1}^n$ to infer the value of θ^*

Lecture April 17

Part I - Review of Statistical Models

Let $\mathbb{P}$ be *the* underlying probability generating experimental outcomes.

Consider a family of CDFs $\mathfrak{F} = \{F_\theta\}_{\theta \in \Theta}$

A $\mathfrak{F}$-based model is an assumption that there exists a $\theta^* \in \Theta$ such that $X_1, ..., X_n \overset{iid}{\sim} F_{\theta^*}(x) = \mathbb{P}(X_1 \leq x)$

e.g. "We assume that flipping a coin (fair or unfair) will follow a Bernoulli distribution F with success rate p. $\mathfrak{F}$ is all the Bernoulli distributions and F_{θ^*} is a specific one I guess it to be. If $F_{\theta^*} = F$ my model is correct. Otherwise, it is unspecified."

The process of using data $\{X_i\}_{i=1}^n = \{X_i(\omega^*)\}_{i=1}^n$ and probability theory to infer the underlying $\mathbb{P}$ (or θ^*) is referred to as "statistical inference."

Part II - statistics

	Probability theory	Statistics
Experiment	Happens before (ω unknown)	Happens after (ω^* is fixed)
Underlying $\mathbb{P}$	Is known	Is not known

Thus the goal of statistics is to infer $\mathbb{P}$ using data.

Part III - Statistical Inference

Branches of Statistical Inference:

1. Hypothesis testing (HT)
2. Point estimation
3. Confidence intervals

Part IV - Hypothesis Testing

Let $\mathfrak{F} = \{F_\theta\}_{\theta \in \Theta}$ be the parameter space and assume that the $\mathfrak{F}$-based model is true (i.e. there exists a $\theta^* \in \Theta$ such that $F_{\theta^*}(x) = \mathbb{P}(X_1 \leq x)$)

Let $\Theta = \Theta_0 \cup \Theta_1$. Then we have two hypotheses as Θ_0 and Θ_1 partition Θ. Either:

1. $H_0 : \theta^* \in \Theta_0$ (which we call the null hypothesis)
2. $H_1 : \theta^* \in \Theta_1$ (the alternative hypothesis)

Only one of these can be true.

Example:

- $\Theta = \mathbb{R}$
- $F_\theta = F_{N(\theta,1)}$
- $\Theta_0 = \{0\}$
- $\Theta_1 = \mathbb{R} - \{0\}$

So our two hypotheses are $H_0 : \theta^* = 0$ vs. $H_1 : \theta^* \neq 0$

Definition: Suppose the sample size is n (i.e. $\{X_i\}_{i=1}^n$ is the data). Any function T mapping $\mathbb{R}^n \mapsto \{0, 1\}$ is called a *test*.

Explanation: IF $T(x_1, x_2, ..., x_n) = 1$, we reject H_0. If it is equal to 0, we accept H_0

Example: $T(\xi_1, \xi_2, ...\xi_n) = 1 \implies$ we always reject H_0 for all data. This is a test but not a good one.

Example: $\Theta = \mathbb{R}$ and $F_\theta = F_{N(\theta,1)}$ so $H_0 : \theta^* = 0$ and $H_1 : \theta^* \neq 0$. By the law of large numbers,

$$\frac{x_1 + x_2 + ... + x_n}{n} \approx \theta^*$$

so we can make a test

$$T(x_1, ..., x_n) = \mathbf{1}\left\{ |\frac{x_1 + x_2 + ... + x_n}{n}| \right\} > c$$

Lecture April 19: Tests

Question: WHat are the criteria for choosing good tests? First, $\mathfrak{F} = \{F_\theta\}_{\theta \in \Theta}$ and then we assume that $F_\theta(x)$ is piecewise differentiable (though this is not actually

necessary in a more rigorous formulation.)

$$p(x|\theta) = \frac{d}{dx} F_\theta(x)$$

with $X_1, X_2, ..., X_n \overset{iid}{\sim} F_theta$ and

$$\vec{X} = (X_1, X_2, ..., X_n)^T \sim F_{\vec{X}}(x_1, ..., x_n)$$

By iid-ness, this implies that

$$F_{\vec{X}}(x_1, x_2, ..., x_n) = \prod_{i=1}^{n} F_{\theta^*}(x_i)$$

which means that the PDF of $\vec{X}$ is

$$\prod_{i=1}^{n} p(x_i|\theta^*)$$

For a test

$$T(X_1(\omega), X_2(\omega), ..., X_n(\omega)) \sim \text{Bernoulli}(\lambda)$$

$$\lambda = \mathbb{E}[T(\vec{X})] = \int_{\mathbb{R}^n} T(\xi_1, ..., \xi_n) \cdot \prod_{i=1}^{n} p(\xi_i|\theta^*)\, d\xi_1 ... d\xi_n = \mathbb{P}(T = 1)$$

where $\mathbb{P}(T = 1)$ is "the expected rejection rate of the null hypothesis"

Note: this equation depends on θ^* but this is not actually known so the expected value can't actually be calculated as such.

So we must consider all $\theta \in \Theta$ so

$$\beta_T(\theta) = \int_{\mathbb{R}^n} T(\xi_1 ..., \xi_n) \cdot \prod_{i=1}^{n} p(\xi_i|\theta)\, d\xi_1 ... d\xi_n$$

so

$$\beta_T(\theta^*) = \mathbb{E}[T(\vec{X})]$$

T makes decisions which may be correct or incorrect. Hence there may be error:
Definitions: Error

1. *Type I Error:* the null hypothesis is true ($\theta^* \in \Theta_0$) but we reject it ($T = 1$)

2. *Type II Error:* the null hypothesis is false ($\theta^* \in \Theta_1$) but we fail to reject it ($T = 0$)

Scenario 1: If $\theta^* \in \Theta_0$,

$$\mathbb{P}(\text{Type 1 error}) = \mathbb{P}(T = 1) = \mathbb{E}[T(\vec{X})] = \beta_T(\theta^*)$$

so $\beta_T(\theta^*)$ must be small but we cannot minimize it as normal without knowing θ^*. BUT

$$\sup_{\theta \in \Theta_0} \beta_T(\theta)$$

has to be small.

This gives the first condition of a good test: type 1 error is small or

$$\boxed{\sup_{\theta \in \Theta_0} \beta_T(\theta) \text{ must be small}}$$

where

$$\beta_T(\theta) = \int_{\mathbb{R}^n} T(\xi_1 ..., \xi_n) \cdot \prod_{i=1}^{n} p(\xi_i|\theta) \; d\xi_1 ... d\xi_n$$

Scenario 2: If $\theta^* \notin \Theta_0$

$$\mathbb{P}(\text{type II error}) = \mathbb{P}(T = 0) = 1 - \mathbb{P}(T = 1) = 1 - \beta_T(\theta^*)$$

so $\beta_T(\theta^*)$ has to be large or

$$\boxed{\beta_T(\theta) \text{ has to be large for all } \theta \in \Theta_1}$$

Lecture April 21

Part I - Review

For a test $T(\vec{X})$ of an $\mathfrak{F}$-based model,

$$T(X_1(\omega), ..., X_n(\omega)) = R(\omega) \in \{0, 1\} \implies R \sim \text{Bernoulli}(r)$$

where $r = \mathbb{P}(R = 1) = \mathbb{E}R$. Then

$$\beta_T(\theta) = \underbrace{\int_{-\infty}^{\infty} ... \int_{-\infty}^{\infty}}_{n} T(\xi_1, ..., \xi_n) \cdot \prod_{i=1}^{n} p(\xi_i|\theta) \; d\xi_1 ... d\xi_n$$

where $p(\xi_i|\theta) = F'_\theta(\xi_i)$. Which means that $\beta_T(\theta^*) = \mathbb{E}R$.

This function can be interpreted "if F_θ is the true CDF, the probability of rejecting H_0 through T is $\beta_T(\theta)$.

This function gives two criteria for a good test relative to other tests:

1. To make the probability of a type 1 error (the null hypothesis is correct but we reject it) small, we want $\sup_{\theta\in\Theta} \beta_T(\theta)$ ("the significance of T") to be small.
2. To make the probability of a type 2 error (null hypothesis is false but we fail to reject it) we want $\beta_T(\theta)$ to be large for every $\theta \in \Theta_1$

Part II - Uniformly Most Powerful Test (UMP Test)

Definition: Let $\alpha \in (0, 1)$ be pre-specified. Suppose T^* is a test with significance α. (That is, $\sup_{\theta\in\Theta_0} \beta^*_T(\theta) = \alpha$). Then the T^* is said to be a UMP test with significance α if:

$$\forall T : \sup_{\theta\in\Theta_0} \beta_T(\theta) = \alpha$$

we have

$$\beta_T(\theta) \leq \beta_{T^*}(\theta) \quad \forall\theta \in \Theta_1$$

Neyman-Pearson Lemma (1930s): With $\Theta = \{\theta_0, \theta_1\}$, $\Theta_0 = \{\theta_0\}$, $\Theta_1 = \{\theta_1\}$. Let $p(\xi|\theta) = F'_\theta(\xi)$ for all $\theta \in \Theta$. For any $\alpha \in (0, 1)$, *the* UMP test with significance alpha is

$$T_\alpha(\xi_1, ..., \xi_n) = \mathbf{1}\left(\frac{\prod_{i=1}^n p(\xi_i|\theta_1)}{\prod_{i=1}^n p(\xi|\theta_0)} > C_\alpha\right)$$

where C_α is the solution to

$$\beta_{T_{NP,\alpha}}(\theta_0) = \alpha$$

Proof: Omitted

Example:

$\Theta = \{\theta_0, \theta_1\}$, $\Theta_0 = \{\theta_0\}$, $\Theta_1 = \{\theta_1\}$. F_θ is the CDF of $N(\theta, 1)$. So

$$p(\xi_i|\theta) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{(\xi_i - \theta)^2}{2})$$

$$\prod_{i=1}^n p(\xi_i|0) = (2\pi^{-n/2}) \exp(-\frac{1}{2}\sum_{i=1}^n \xi_i^2)$$

$$\prod_{i=1}^{n} p(\xi_i|1) = (2\pi^{-n/2}) \exp(-\frac{1}{2}\sum_{i=1}^{n}(\xi_i - 1)^2)$$
$$= (2\pi^{-n/2}) \exp(-\frac{1}{2}\sum_{i=1}^{n}(\xi_i)^2) \cdot \exp\left(\sum_{i=1}^{n}\xi_i\right) \cdot \exp(-\frac{n}{2})$$

So

$$T_{NP,\alpha}(\xi_1, ..., \xi_n) = \mathbf{1}\left(\exp(-\frac{n}{2}) \cdot \exp(n \cdot \bar{\xi}_n > C_\alpha)\right) = \mathbf{1}\left(\bar{\xi}_n > \frac{\log C_\alpha}{n} + \frac{1}{2}\right)$$

We have data $\{x_i\}_{i=1}^n$

Interpretation: We reject $H_0 : \theta^* = 0$ if $\bar{X}_n > \frac{1}{2} + \frac{\log C_a}{n}$

We fail to reject H_0 if $\bar{X}_n \leq \frac{1}{2} + \frac{\log C_a}{n}$

Lecture April 24: The Maximum Likelihood Estimator (MLE)

Part I - Framework

Let $\mathfrak{F} = \{F_\theta\}_{\theta \in \Theta}$ be a family of CDFs indexed by θ with $\mathbb{P}$ as the underlying probability. We then assume that the $\mathfrak{F}$-based model is correct, i.e.,

$$\exists \theta^* \in \Theta : X_1, ..., X_n \overset{iid}{\sim} F_{\theta^*}(x) = \mathbb{P}(X_1 \leq x)$$

The process of estimating θ^* is referred to as "point estimating" in statistics.

Point Estimating:

1. MLE
2. The method of moments
3. Mean squared estimation (MSE)/Least squares Estimator

Note: Only MLE is covered in APMA 1655.

Part II - Motivation from the Neyman-Pearson View

- $\Theta = \{\theta_0, \theta_1\}$
- $\Theta_0 = \{\theta_0\}$
- $\Theta_1 = \{\theta_1\}$
- $H_0 : \theta^* = \theta_0$
- $H_1 : \theta^* = \theta_1$
- $\mathfrak{F} = \{F_\theta\}_{\theta \in \Theta}$
- $p(\xi|\theta) = F'_\theta(\xi)$
- Data $\{x_i\}_{i=1}^n$ is given and fixed.

The Neyman-Pearson Test (the Uniformly Most-Powerful test) is

$$T(x_1 ..., x_n) = \mathbf{1}\left(\frac{\prod_{i=1}^n p(x_i|\theta_1)}{\prod_{i=1}^n p(x_i|\theta_0)} > c_a\right)$$

That is, if $T(\vec{x}) - 2$ we reject the null hypothesis. Otherwise, we accept the null hypothesis.

From that inequality, when the numerator tends to be large, we tend to believe that $\theta^* = \theta_1$ and vice versa. In other words, we believe that the true parameter θ^* is

$$\underset{\theta \in \{\theta_0, \theta_1\}}{\operatorname{argmax}} \left(\prod_{i=1}^n p(x_i|\theta)\right)$$

Part III - Definition of MLE

Definition: Let Θ be a subset of a finite dimensional space. Let $\mathfrak{F} = \{F_\theta\}_{\theta \in \Theta}$ be a family of CDFs and $p(\xi|\theta) = F'_\theta(\xi)$. Suppose we have data $\mathfrak{D} = \{x_i\}_{i=1}^n$ Then

$$L(\theta|\ \mathfrak{D}) = \prod_{i=1}^n p(x_i|\theta)$$

is the likelihood function. And then the maximum likelihood estimator is

$$\hat{\theta}_{MLE} = \underset{\theta \in \Theta}{\operatorname{argmax}}\ L(\theta|\ \mathfrak{D})$$

Part IV - Calculating MLE

The "log-likelihood" is

$$\log L(\theta|\ \mathfrak{D}) = l(\theta|\mathfrak{D}) = \sum_{i=1}^{n} \log p(x_i|\theta)$$

So because log is strictly increasing,

$$\underset{\theta\in\Theta}{\operatorname{argmax}}\ L(\theta|\ \mathfrak{D}) = \underset{\theta\in\Theta}{\operatorname{argmax}}\ l(\theta|\ \mathfrak{D})$$

Thus

$$\boxed{\frac{\partial}{\partial\theta} l(\theta|\mathfrak{D}) = \sum_{i=1}^{n} \frac{\partial}{\partial\theta} \log[p(x_i|\theta)] = 0}$$

Example: Say $\hat{\theta}$ is the solution to the above equation. If $\frac{\partial^2}{\partial\theta^2} l(\theta|\mathfrak{D})\big|_{\theta=\hat{\theta}} < 0$. Then

$$\hat{\theta} = \hat{\theta}_{MLE} = \underset{\theta\in\Theta}{\operatorname{argmax}}\ l(\theta|\ \mathfrak{D})$$

Example: (will be on final!)

- $\Theta = \mathbb{R}$
- F_θ is the CDF of $N(0,1)$ so $p(\xi|\theta) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{(\xi-\theta)^2}{2})$
- $\mathfrak{D} = \{x_i\}_{i=1}^n$

Then

$$\begin{aligned} l(\theta|\mathfrak{D}) &= \log\left[\prod_{i=1}^{n} p(x_i|\theta)\right] \\ &= \sum_{i=1}^{n} \log[p(x_i|\theta)] \\ &= \sum_{i=1}^{n} \left[-\frac{1}{2}\log(2\pi) - frac(x_i-\theta)^2 2\right] \\ \frac{\partial}{\partial\theta} l(\theta|\ \mathfrak{D}) &= \sum_{i=1}^{n}(x_i-\theta) \\ &= \left(\sum_{i=1}^{n}\right) - n\cdot\theta = 0 \\ &\implies \hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x}_n \end{aligned}$$

$$\frac{\partial^2}{\partial\theta^2} l(\theta|\mathfrak{D}) = -n < 0$$

Therefore

$$\boxed{\hat{\theta}_{MLE} = \bar{x}_n}$$

Lecture April 26:

Part I - MLE

Suppose the $\{F_\theta\}_{\theta\in\Theta}$-based model is correct. Thus the goal of the MLE is to estimate the "true parameter" θ^* for which $X_1, ..., X_n \overset{iid}{\sim} F_{\theta^*}(x) = \mathbb{P}(X_1 \leq x)$

We assume that F_θ is (piecewise differentiable). We have a collection of given, fixed, deterministic data $D = \{x_i\}_{i=1}^n$.

We define a "likelihood function"

$$L(\theta|D) = \prod_{i=1}^{n} p(x_i|\theta)$$

For different data, this function changes value so we select the theta which maximizes the function:

$$\hat{\theta}_{MLE} = \arg\max_{\theta \in \Theta} L(\theta|D)$$

Note: $-L(\theta|D)$ is usually referred to as a loss function

$$\arg\max_{\theta \in \Theta} L(\theta|D) = \arg\min_{\theta \in \Theta}(-L(\theta|D))$$

In most real-world applications, the argmax is very difficult to calculate. Two primary methods to approximate it are Gradient Descent and the Expectation-Maximization Algorithm.

We also define the "log-likelihood":

$$l(\theta|D) = \log L(\theta|D) = \sum_{i=1}^{n} \log p(x_i|\theta)$$

Then because log is strictly increasing,

$$\arg\max_{\theta \in \Theta} L(\theta|D) = \arg\max_{\theta \in \Theta} l(\theta|D)$$

Part II - An Example

Application: $\Theta = \mathbb{R}$ and

$$p(\xi|\theta) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{(\xi - \theta)^2}{2})$$

$$\begin{aligned} l(\theta|D) &= \sum_{i=1}^{n} logp(x_i|\theta) \\ &= \sum_{i=1}^{n} \log\left(\frac{1}{\sqrt{2\pi}} \exp(-\frac{(x_i - \theta)^2}{2})\right) \\ &= \sum_{i=1}^{n} \left(-\frac{1}{2}\log(2\pi) - \frac{1}{2}(x_i - \theta)^2\right) \\ &= -\frac{n}{2}\log(2\pi) - \sum_{i=1} \frac{1}{2}(x_i - \theta)^2 \end{aligned}$$

From calculus, we know the maximum will come at the critical point:

$$\begin{aligned}\frac{\partial}{\partial\theta}l(\theta|D) &= \frac{\partial}{\partial\theta}\left(\sum_{i=1}^{n} -\frac{1}{2}(x_i-\theta)^2\right)\\ &= \sum_{i=1}^{n}(x_i-\theta)\\ &= -n\cdot\theta + \sum_{i=1}^{n} x_i\end{aligned}$$

Thus

$$\hat{\theta} = \frac{\partial}{\partial\theta}l(\theta|D) = 0 \longrightarrow \hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x}_n$$

Then from the second derivative,

$$\frac{\partial^2}{\partial\theta^2}l(\theta|D) = -n < 0$$

so this is in fact the maximum!

Thus,

$$\hat{\theta} = \bar{x}_n = \hat{\theta}_{MLE}$$

As we will see, $\hat{\theta}_{MLE} \approx \theta^*$ and we have found our function!

In this example, our model is $\{N(\theta,1)\}_{\theta\in\mathbb{R}}$ (the collection of normal distributions). We want the calculated MLE ($\hat{\theta}_{MLE}$) to be close to the true parameter θ^*.

Before the experiment we only have potential data $\{X_i\}_{i=1}^n$. So before the experiment the MLE is

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

The Law of Large Numbers implies that

$$\hat{\theta}_{MLE} = \bar{X}_n(\omega) \approx \mathbb{E}X_1 = \theta^*$$

Note: the last equal sign is only true if the model is correct.

But the Central Limit Theorem implies that $\sqrt{n}(\bar{X}_n - \theta^*) \dot{\sim} N(0,1)$

Note: these results only hold for this specific example! Other models and problem will have different implications but the results of the CLT and LLN are true for all MLE.

Part III - The Final Theorem

Theorem: Suppose on $(\mathbb{P}, \Omega)$ the $\{F_\theta\}_{\theta\in\Theta}$-based model is correct.

After the experiment,

$$\hat{\theta}_{MLE}(\vec{x}) = \arg\max_{\theta\in\Theta} L(\theta|\ \vec{x})$$

But before the experiment

$$\hat{\theta}_{MLE}(X_1(\omega),\ ...,\ X_n(\omega)) \overset{def}{=} \hat{\theta}_n(\omega)$$

Under some "regularity conditions" (omitted),

- (Consistency)

$$\forall \varepsilon > 0, \qquad \lim_{n\to\infty} \mathbb{P}(\{\omega \in \Omega : |\hat{\theta}(\omega) - \theta^*| < \varepsilon\}) = 1$$

- (Aymptotic Normality)

$$\sqrt{n}(\hat{\theta}_n - \theta^*) \dot{\sim} N(0, \frac{1}{I(\theta^*)})$$

  where the "Fisher Information" $I(\theta)$ is

$$I(\theta) = \int_{-\infty}^{\infty} \left(\frac{\partial}{\partial\theta} \log p(\xi|\theta)\right)^2 \cdot p(\xi|\ \theta)\ d\xi$$

Proof: don't be silly. look at the topology of the parameter space

Lecture May 1: Confidence Sets (Optional)

Part I - Setup

Let $\mathbb{P}$ be the underlying probability generating experimental outcomes. We select a specific family of CDFs $\{F_\theta\}_{\theta\in\Theta}$ and assume

$$\exists \theta^* \in \Theta : X_1, X_2,\ ...,\ X_n \overset{iid}{\sim} F_{\theta^*}(x) = \mathbb{P}(X_1 \leq x)$$

The goal of confidence sets is to find subsets of Θ that cover the true parameter θ^* with a high probability.

Part II - Definition

Suppose C_n is a set-value function $C_n : \mathbb{R}^n \to \{\text{subsets of } \Theta\}$. Let $\alpha \in (0,1)$ be pre-specified. Then C_n is a *confidence set* with confidence $1-\alpha$ if

$$\mathbb{P}(\{\omega \in \Omega : \theta^* \in C_n(X_1(\omega), ..., X_n(\omega))\}) = 1 - \alpha$$

If the values of C_n are intervals the corresponding confidence sets are referred to as *confidence intervals*. Note that this implies the dimensionality of Θ is 1.

Lrem