LECTURE 1

Overview of basic concepts in probability theory

Axiomatic approach to probability by Kolmogorov (an axiom is a postulate assumed to be true without a proof required)

Axiomatization allows you to set a number of primitive that may be used to deductively derive more complex truths.

Example

Peano's axiomatization of natural numbers \((0, 1, 2, 3, 4, \ldots)\) (1889)

- A unary function ("successor")
  - There is a natural number \(0\)
  - Every natural number \(a\) has a successor \(S\).
  - No natural number has \(0\) as its successor
  - Distinct numbers have distinct successors: \(a \neq b \Rightarrow S\ a \neq S\ b\)
  - If a property is true of \(0\), and of the successor of every natural number that has that property, then the property is held by all natural numbers

Axioms should be broad enough to allow profound results and truths to be deduced, but they should not be redundant.

Axioms associated will on experiment that has random outcomes

Example

Take a die. Roll it. Observe numerical outcome

Q1 What can I possibly observe?
A new \(\{1, 2, 3, 4, 5, 6\} = \text{the set of all possible outcomes}
\}
\(\omega \in \Omega\) is called an outcome

Q2 What am I interested in? What if I care only if the # is even or odd?
A new Outcomes of interest are called events.
Say \(E = \{1, 3, 5\}, \Omega = \{2, 4, 6\}\). Events are subsets of \(\Omega\)
Events of interest have to have a special structure.
If we denote the set of events by \(\mathcal{F}\), we want \(\mathcal{F}\) to be a \(\sigma\)-algebra, that is
\[ A \in \mathcal{F} \]
\[ A \in \mathcal{F} \Rightarrow A^c \in \mathcal{F} \]
\[ A, B \in \mathcal{F} \Rightarrow A \cup B \in \mathcal{F} \]

For our example, \( \mathcal{F} = \{ \varnothing, \{1, 2, 3, 4, 6\}, \{1, 2, 5, 6\}, \{2, 4, 6\} \} \)

De Morgan's laws:
\[ (A \cup B)^c = A^c \cap B^c \]
\[ (A \cap B)^c = A^c \cup B^c \]

Hence, have redundant axiom that if \( A_1, A_2, \ldots \in \mathcal{F} \), then \( \bigcap_{i=1}^{\infty} A_i \in \mathcal{F} \)

Q3 How likely is the observation?

Answer: We need to assign a measure to every possible event that tells you how often you may expect to see that event.

That measure is called probability, \( P \)
\[ P(\{1\}) = P(1) = 1/6, \quad P(2) = 1/6, \quad \ldots, \quad P(6) = 1/6 \]

The measure also has to satisfy some properties, which are axiomatized:

1. \( A_i \) are mutually exclusive if \( A_i \cap A_j = \varnothing \) for \( i \neq j \)
2. \( P(A) > 0 \) for \( A \in \mathcal{F} \)
3. If \( A, B \in \mathcal{F} \), \( A \cap B = \varnothing \), then \( P(A \cup B) = P(A) + P(B) \)
4. If \( A_1, A_2, \ldots \) is a sequence of mutually exclusive events, then \( P(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i) \)
5. \( P(\Omega) = 1 \)

Things we can deduce:
\[ A \subseteq B \Rightarrow P(B) = P(A) + P(B \setminus A) \]
\[ P(A \cup B) = P(A) + P(B) - P(A \cap B) \]

The triple \( (\Omega, \mathcal{F}, P) \) is called a probability space.

Later

Bertrand's paradoxes (Joseph Bertrand)

Consider an equilateral triangle inscribed in a circle. Suppose a chord of the circle is chosen at random. What is the probability that the chord is longer than the side of the triangle?
Let $\Omega = \{ w : 0 \leq w \leq 1 \}$

Want $\mathcal{F}$ to include closed intervals, and want "measure" to be length

i.e. $[a, b] \in \mathcal{F} \quad \forall a, b \in [0, 1], \; a \leq b$

$$\mu([a, b]) = b - a$$

$\mathcal{F}$ has to include open intervals, singletons, etc. So why not take

$\mathcal{F} = \mathcal{P}(\Omega)$ = power set of $\Omega$ = set of all subsets of $\Omega > \aleph_0$

Cantor's theorem states that the power set of any set has cardinality $> \aleph_0$.

Lebesgue (measure) is tricky to assign to all sets in the power set $\mathcal{P}(\Omega)$

Proof of Cantor's Theorem

Take $S = \{ 1, 2, 3, \ldots \} = \mathbb{N}$

$2^S = \{ \emptyset, \{ 1 \}, \{ 2 \}, \{ 3 \}, \ldots \}$

same cardinality as

$\{ 1, 2, 3, \ldots \}$

Pairing

1 $\leftrightarrow \emptyset$

2 $\leftrightarrow \{ 1 \}$

3 $\leftrightarrow \{ 2 \}$

4 $\leftrightarrow \{ 3, 4, 5 \}$

etc.

Since $2^S$ is a power set it must contain the set of all unselfish numbers. What is this set associated with?

- if associated with selfish LHS, then contradiction
- as set is unselfish
- if associated with unselfish LHS, then contradiction
- as LHS should then be in set, which means number should be selfish

Simply, too many sets in the power set and "enlarged" contradictions arise: $\mathcal{P}(\mathbb{R})$

$$0 \leq \mu([a, b]) = b - a \quad a < b$$

$$\mu(A + y) = \mu(A) \quad \forall A \subseteq \mathbb{R}, \; y \in \mathbb{R}$$

if $A = \bigcup_{i=1}^{n} B_i \cup \emptyset, \quad \forall i, j \neq i, \; \cap_{i=1}^{n} B_i = 0$ \hspace{1cm} $A + y = \{ a + y : a \in A \}$

Then $\mu(A) = \sum_{i=1}^{n} \mu(B_i)$
Q - rational #s
Q is countably infinite
Say \( x, y \in \mathbb{R} \) are equivalent if \( x - y \in \mathbb{Q} \)
(equivalence: \( x \sim y \iff y \sim x \iff x - y \in \mathbb{Q} \))
Define sets \( Q_x = Q + x \), \( \forall x \in \mathbb{R} \). For all \( x, y \)
Either \( Q_x = Q_y \) or \( Q_x \cap Q_y = \emptyset \)
\( Q_x \cap [0,1] \neq \emptyset \), \( \forall x \in \mathbb{R} \)
\( V = \{ \tilde{q}_1, \tilde{q}_2, \ldots \} \) rationals in \([-1, 1]\)
\( V_i = V + \tilde{q}_i \), \( V_i \)'s are disjoint
\( [0,1] \subseteq \bigcup_i V_i \subseteq [-1, 2] \)
\( V_i \)'s translations of \( V \), all have to have the same
length
\[ \text{length}([0,1]) = 1 \leq \text{length} \left( \bigcup_i V_i \right) \leq \text{length}([-1, 2]) = 3 \]
\( \mathbb{F} = \) smallest \( \sigma \)-algebra containing all
the open subsets of \( \Omega \)
= intersection of all \( \sigma \)-algebras containing all the open subsets of \( \Omega \)
= Borel \( \sigma \)-algebra over \([0,1]\)
Not every subset of \( \Omega \) is a Borel set.
(Use Lebesgue measure on Borel sets as prob. measure to be "safe")

Cylinder sets: Cartesian product \( \times = \prod \mathbb{X}_x \) of topological spaces \( \mathbb{X}_x \). The canonical projection \( p_x : \times \to \mathbb{X}_x \). Given any \( \times \)-open set \( U \subset \mathbb{X}_x \), \( \prod U(x) \) is called an open cylinder.
The intersection of a finite \# of \( \times \)-open cylinders is a cylinder set.
\( S = \{ s_1, \ldots, n \} \)
\( S^2 = \{ x = (\ldots, x_{-1}, x_0, x_1, \ldots) : x_k \in S, \forall k \in \mathbb{Z} \} \)
Open cylinders \( C[\alpha] = \{ x \in S^2 : x_\alpha = \alpha \} \)
\( C[\alpha_0, \ldots, \alpha_m] = C[\alpha_0] \cap \ldots \cap C[\alpha_0, \ldots, \alpha_m] = \{ x \in S^2 : x_\alpha = \alpha_0, \ldots, x_{\alpha + m} = \alpha_m \} \)
Used for repeated trials: \( w = \) outcome = infinite sequence
events = cylindrical sets
Continuity of the probability measure

Suppose \( B_1, B_2, \ldots \) is a sequence of events.

a) If \( B_1 \subseteq B_2 \subseteq \cdots \), then
\[
\lim_{i \to \infty} P(B_i) = P(\bigcap_{i=1}^{\infty} B_i)
\]

b) If \( B_1 \supseteq B_2 \supseteq \cdots \), then
\[
\lim_{i \to \infty} P(B_i) = P(\bigcap_{i=1}^{\infty} B_i)
\]

\[
\lim_{n \to \infty} a_n = a \Leftrightarrow \exists \varepsilon > 0 \text{ s.t. } \forall n > n_0(\varepsilon), |a_n - a| < \varepsilon
\]

Let \( D_i = B_i \setminus D_{i-1} \), \( i \geq 2 \), and \( D_1 = B_1 \). The sets \( D_i \) are disjoint.

\[
\lim_{i \to \infty} P(B_i) = \lim_{i \to \infty} P(\bigcup_{i=1}^{\infty} D_i) = \lim_{i \to \infty} \sum_{i=1}^{\infty} P(D_i) = \sum_{i=1}^{\infty} P(D_i)
\]

\[
= P\left( \bigcup_{i=1}^{\infty} D_i \right) = P\left( \bigcup_{i=1}^{\infty} B_i \right)
\]

Example

\[
\begin{array}{c}
S = \{(x, y) \mid (x, y) \in S_1, S_2, S_3\}
\end{array}
\]

**Events**

\[
\mathcal{F} = \mathcal{B} \left\{ \text{subsets of } S \right\}
\]

\( T_1 > T_2 > T_3 \rightarrow T_4 \), \( P(T_i) = \frac{a_i}{n^i}, i = 1, 2, 3 \)

\( \mathcal{F} = \mathcal{B} \left\{ \text{cylinder sets of } S \right\} \)

\( T \in \mathcal{F} \), and \( P(T) = P(\bigcup_{i=1}^{\infty} T_i) = \lim_{i \to \infty} P(T_i) = \frac{1}{2} \)

**Independence and Conditional probability**

**Events** \( A_1, \ldots, A_n \) are independent if

\[
P(A_i \cap A_j) = P(A_i)P(A_j) \quad i \neq j
\]

\[
P(A_i \cap A_j \cap A_k) = P(A_i)P(A_j)P(A_k) \quad i \neq j \neq k
\]

\[
P(A_1 \cdots A_n) = P(A_1) \cdots P(A_n)
\]

**Pairwise independence**

Let \( B \) be an event s.t. \( P(B) > 0 \)

**Conditional prob**

\[
P(A \mid B) = \frac{P(A \cap B)}{P(B)}
\]
Total law of probability and Bayes formula

Let $E_1, \ldots, E_n$ be events that form a partition of $\Omega$ (i.e. $U E_i = \Omega$ and $E_i \cap E_j = \emptyset$ for $i \neq j$)

Then $P(A) = P(A|E_1) + \cdots + P(A|E_n)$

$$= P(A|E_1)P(E_1) + \cdots + P(A|E_n)P(E_n)$$

Since $P(E_i|A) = \frac{P(A|E_i)P(E_i)}{P(A)} = \frac{P(A|E_i)}{P(A)}$

we can get Bayes formula

$$P(E_i|A) = \frac{P(A|E_i)P(E_i)}{P(A|E_1)P(E_1) + \cdots + P(A|E_n)P(E_n)}$$

Borel-Cantelli Lemma

Consider a sequence of events $A_1, A_2, \ldots$

and define

$$A = \limsup_{n \to \infty} A_n = \bigcap_{k=1}^{\infty} \bigcup_{n=k}^{\infty} A_n$$

\(\forall k \in \mathbb{N}, \exists n > k \text{ s.t. a sample } \omega \text{ belongs to some } A_n\)

$A = \{ \omega : \omega \text{ belongs to infinitely many } A_i \text{'s} \}$

may also define

$$A = \bigcup_{k=1}^{\infty} \bigcap_{n=k}^{\infty} A_n = \liminf_{n \to \infty} A_n$$

\(\exists k \text{ s.t. a sample } \omega \text{ belongs to all } A_n \text{ for } n \geq k\)

$A = \{ \omega : \omega \text{ belongs to all but finitely many } A_i \text{'s} \}$

Recall classical def. of limsup and liminf for sequences

$$\limsup_{n \to \infty} a_n = \lim_{n \to \infty} \sup_{k \geq n} a_k$$

Let $p_n = P(A_n)$. If $\sum p_n = \infty$, then $P(A_n) \text{ infinitely often } \to 0$

If $\sum p_n = \infty$, $A_n$ are mutually independent. Then $P(A_1 \cap \cdots) = 1$

Proof. From continuity of probability

(as $U A_1$ is a sequence of nonincreasing sets)

$$P(A_1 \text{ infinitely often}) = \lim_{n \to \infty} P(U A_n)$$

$$B_n = \bigcup_{k=n}^{\infty} A_n \quad B_1 > B_2 > \cdots \quad \lim_{n \to \infty} P(B_n) = P(U A_1)$$
\[ P(\bigcup_{n=k}^{\infty} A_n) = \lim_{n \to \infty} P(\bigcup_{n=k}^{\infty} A_n) = \lim_{n \to \infty} \sum_{n=k}^{\infty} P(A_n) = \lim_{n \to \infty} \sum_{n=k}^{\infty} P_n = \sum_{n=k}^{\infty} P_n \quad (4) \]

If \[ \sum_{n=k}^{\infty} P_n < \infty \], then \[ \lim_{k \to \infty} \sum_{n=k}^{\infty} P_n = 0 \]

**For the proof of the second part de Morgan's law:**

\[ P(\bigcup_{n=k}^{\infty} A_n) = \lim_{n \to \infty} P(\bigcup_{n=k}^{\infty} A_n) = \lim_{n \to \infty} P(\bigcap_{n=k}^{\infty} A_n^c) \]

\[ = \lim_{n \to \infty} \left[ 1 - P(\bigcap_{n=k}^{\infty} A_n) \right] = \lim_{n \to \infty} \left[ 1 - \prod_{n=k}^{\infty} (1 - P(A_n)) \right] \]

Now observe that \[ \exp(-u) = 1 - u + \frac{u^2}{2} - \cdots \Rightarrow 1 - u, \forall u \]

\[ \geq \lim_{n \to \infty} \left[ 1 - \exp(-\sum_{n=k}^{\infty} P_n) \right] = 1 - \exp\left(-\sum_{n=k}^{\infty} P_n\right) = 1 - \exp(-\infty) = 1 \]

\[ B_m = \bigcup_{n=k}^{m} A_n \quad B_k \subseteq B_{k+1} \subseteq \ldots \]

\[ P(\bigcup_{n=k}^{m} A_n) = P(B_m) = \lim_{m \to \infty} P(B_m) = \lim_{m \to \infty} P(\bigcup_{n=k}^{m} A_n) \]

This result is a form of a 0-1 law of probability that often apply to tail events (i.e., \( \limsup A_n \), \( \liminf A_n \), tail events) only depend on the limiting behavior of the sequence.

Random variables \((\Omega, \mathcal{F}, P)\), \((\mathbb{R}, \mathcal{M})\)-measurable space

\( \forall B \in \mathcal{M}, \quad X^{-1}(B) \in \mathcal{F} \)

\( X: \Omega \to \mathbb{R} \) s.t. \( \{\omega : X(\omega) \leq c\} \in \mathcal{F} \) for any \( c \in \mathbb{R} \)

A random variable defined as above is said to be \( \mathcal{F} \)-measurable.

CDF: \( P\{\omega : X(\omega) \leq c\} = \mathcal{F}_X(c) \)

\[ \mathcal{F}_X(c) = P\{X \leq c\} = F_X(c) \]

\[ P\{X \leq c\} = \lim_{c \to \infty} P\{X < c\} = \lim_{c \to \infty} \mathcal{F}_X(c) = \lim_{c \to \infty} F_X(c) \]

\[ F_X(c) \leftarrow P\{X < c\} = F_X(c) - F_X(c^-) \]
A function F is a CDF of some RV if it has the following properties:

a) F is non-decreasing
b) \( \lim_{x \to -\infty} F(x) = 0 \)
   \( \lim_{x \to \infty} F(x) = 1 \)
c) F is right continuous

**Only if** : F is a CDF of some X

a) \( F(x) = P\{X \leq x\} = P\{X \leq x\} + P\{x < X \leq x + \epsilon\} \geq P\{X \leq x + \epsilon\} \)
b) Only show \( \lim_{x \to \infty} F(x) = 1 \)
   \( B_n = \{X \leq n\}, n \in \mathbb{N} \)
   \( B_1 \subseteq B_2 \subseteq \ldots \)
   \( \lim_{n \to \infty} F(n) = \lim_{n \to \infty} P\{B_n\} = P\{\bigcup_{n=1}^{\infty} B_n\} = P\{\Omega\} = 1 \)

What about the reals?

For any \( \epsilon > 0 \), \( \exists n(\epsilon) \) s.t. \( \forall x > n(\epsilon), \quad F(x) < 1 - \epsilon \)

Hence, \( \lim_{x \to \infty} F(x) = 1 \)

c) \( A_n = \{X \leq x + \frac{1}{n}\}, \quad n \geq 1 \)

\( A_1 \supset A_2 \supset A_3 \supset \ldots \)

\( \lim_{n \to \infty} F(x + \frac{1}{n}) = \lim_{n \to \infty} P(A_n) = P(\bigcap_{i=1}^{\infty} A_i) = P\{X \leq x\} = F(x) \)

Same as M, we have that this holds for all \( n \).

**It is difficult. Measure theory (see, for example, Th 4.1 of Billingsley, Probability and Measure)**

Say \( \Omega = (0,1) \)

\( S = \) Borel \( \sigma \)-algebra

\( P = \) Lebesgue measure

Hinges on property of Lebesgue measure, easiest to show if F(x) is monotonic, in which case there is a 1-1 mapping between \( \Omega = (0,1) \) and \( \mathbb{R} \), and hence a reverse mapping \( \mathcal{E} \)

\( P\{X \leq x\} = P\{\omega \in (0,1) : \mathcal{E}(\omega) \leq x\} \)

\( = P\{\omega \in (0,1) : \omega \leq F(x)\} = F(x) \)
Discrete RVs: Probability mass function
\[ p(x) = P(X = x) \]

Continuous RVs:

**Absolute continuity** (strong smoothness) stronger than continuity of uniform continuity
Allows rigorous use of the fundamental theorem of calculus
\( I \subset \mathbb{R} \), an interval
\( f: I \rightarrow \mathbb{R} \) AC on I, if \( \forall \varepsilon > 0, \exists \delta > 0 \)
s.t. whenever for a collection of int. \((x_k, y_k)\) \( k = 1, 2, \ldots \) s.t.,
\[ (x_k, y_k) \subset I \quad (x_k, y_k) \cap (x_m, y_m) = \emptyset \quad \forall k \neq m \]
It holds
\[ \sum_{k} (y_k - x_k) < \delta \]
does not fluctuate on a set of measure zero
\[ \sum_{k} |f(y_k) - f(x_k)| < \varepsilon \]
Maps sets of measure zero to sets of measure zero

If \( f \) absolutely continuous \( \Rightarrow f \) has a derivative \( f' \) almost everywhere
The derivative is Lebesgue integrable, and
\[ f(x) = f(a) + \int_{a}^{x} f'(t) dt \quad \forall x \in [a, b] \quad \text{(Lebesgue integral)} \]

Cantor sets - Cantor function

Lebesgue vs. Riemann integral
Later, as examples

If \( F \) is absolutely continuous, can define so called pdf \( f_x(x) \) s.t.
\[ F(x) = \int_{-\infty}^{x} f_x(u) du \quad \text{(why? since } F(-\infty) = 0) \]

If \( f(x) \) is continuous at \( u \), have
\[ f(u) = \lim_{\varepsilon \to 0} \frac{1}{\varepsilon} \int_{u-\varepsilon}^{u+\varepsilon} f(v) dv \]

We take the integrals to be Lebesgue integrals

Riemann integral

\([a, b]\) partition, say \( a = x_1 < x_2 < \ldots < x_n = b \) so that \([a, b] = \bigcup_{i=1}^{n} [x_i, x_{i+1}]\)
\( m = \max (x_{i+1} - x_i) \) \( \rightarrow \) want \( m \) to be very small
Tagged with \( t_i \in [x_i, x_{i+1}] \), \( i = 1, \ldots, n-1 \)

Riemann sum
\[ \sum_{i=1}^{n} f(t_i) (x_{i+1} - x_i) \rightarrow \text{upper sum (sup)} \]
\[ \sum_{i=1}^{n} f(t_i) (x_{i+1} - x_i) \rightarrow \text{lower sum (inf)} \]
take inf. over all partitions of the upper sum
take sup. over all partitions of the lower sum
if they are equal, call the sum the Riemann integral

functions that are not Riemann integrable: no

Lebesgue integral: uses Lebesgue sums instead, we partition the range, not
the domain
\[ \int_a^b f(x) \, dx = \lim_{n \to \infty} \sum_{i=1}^n f(y_i^*) (y_i - y_{i-1}) \]

**Expectation of a RV**
*Simple RV: Discrete *
\[ E[X] = \sum_{i=1}^n x_i P(X = x_i) \]

*Simple: \( X \) a finite set \( \{x_1, \ldots, x_n\} \) s.t. \( X(\omega) = x_\omega, \ \forall \omega \)

If \( X \) has a pdf, \( E[X] = \int_{-\infty}^{\infty} x f(x) \, dx \) in the Lebesgue sense (mention what well defined means)

Properties
- Linearity (follows from sum-derived definition of integrals, i.e., integration being a linear operator)

\[ E[X + Y] = E[X] + E[Y], \quad E[cX] = cE[X] \quad \text{for } c \text{ a constant} \]

- \[ E[X] = \int_{-\infty}^{\infty} (1 - F_X(x)) \, dx - \int_{-\infty}^{\infty} F_X(x) \, dx \]

/at least one of the two integrals is finite

**Function of a RV** \((\mathcal{O}, \mathfrak{S}, P)\)
\[ X : \mathcal{O} \to \mathbb{R} \text{ so that } \{\omega : X(\omega) \leq c \} \in \mathfrak{S} \quad \forall c \in \mathbb{R} \]

\[ g : \mathbb{R} \to \mathbb{R} \text{ such that } \{x \in \mathbb{R} : g(x) \leq c \} \text{ is a Borel subset of } \mathbb{R} \]
\[ X(\omega) = g(X(\omega)) \text{ composition of two functions} \]

We have \( E[g(X)] = \int_{-\infty}^{\infty} g(x) f_X(x) \, dx \) in the Lebesgue sense

**Higher order moments**
[\( E[X^n] \) for \( n \geq 1 \)]

\[ \operatorname{var}(X) = E[(X - E[X])^2] = E[X^2] - E[X]^2 \]

**Characteristic function** \( \phi_X(u) = E[e^{iuX}], \quad \phi_X^{(r)}(0) = i^r E[X^r] \)

if higher order moments exist.
Examples of distributions that are frequently used

**Bernoulli X**

\[ P(X = 0) = 1 - p, \quad P(X = 1) = p, \quad \text{where } p \in [0, 1] \]

\[ E[X] = p, \quad \text{var}(X) = p(1-p) \]

**Binomial X**

\[ n \geq 1, \quad p \in [0, 1] \]

\[ P(X = i) = \binom{n}{i} p^i (1-p)^{n-i} \]

\[ E[X] = np, \quad \text{var}(X) = np(1-p) \]

**Poisson X**

\[ \lambda > 0 \]

\[ P(X = i) = \frac{\lambda^i e^{-\lambda}}{i!} \]

\[ E[X] = \lambda, \quad \text{var}(X) = \lambda \]

**Geometric X**

\[ 0 \leq p \leq 1 \]

\[ P(X = i) = (1-p)^{i-1} p \]

\[ E[X] = \frac{1}{p}, \quad \text{var}(X) = \frac{1-p}{p^2} \]

**Continuous**

**Gaussian** \( N(\mu, \sigma^2) \), \( \mu \in \mathbb{R} \), \( \sigma > 0 \)

\[ f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left( - \frac{(x-\mu)^2}{2\sigma^2} \right) \]

\[ E[X] = \mu, \quad \text{var}(X) = \sigma^2 \]

**Exponential**

\[ f(x) = \lambda e^{-\lambda x}, \quad x > 0 \]

\[ E[X] = \frac{1}{\lambda}, \quad \text{var}(X) = \frac{1}{\lambda^2} \]

**Uniform**

\[ f(x) = \begin{cases} \frac{1}{b-a}, & \text{if } x \in [a, b] \\ 0, & \text{elsewhere} \end{cases} \]

\[ U(a, b) \]
Conditional densities and jointly distributed RVs

\[ \forall x_1, x_2, \ldots, x_m \text{ RVs over the same probability space } (\Omega, \mathcal{F}, \mathbb{P}) \]

\[ F_{x_1 \ldots x_m}(a_1, \ldots, a_m) = P(x_1 \leq a_1, \ldots, x_m \leq a_m) \]

\[ b' \]

\[ P \left\{ (x_1, x_2) \in S \right\} = F_{x_1 x_2}(b_1 b') - F_{x_1 x_2}(a_1 b') \]

\[ = -F_{x_1 x_2}(b_1 a') + F_{x_1 x_2}(a_1 a') \]

All properties/concepts carry over from 1-dim case: joint pdf for absolutely continuous CDFs.

In particular, we are interested in joint pdf's \( f_{x_1 x_2}(x_1 x_2) \) based on which we can define conditional pdf's

\[ f_{x_1 | x_2}(x_1 | y) = \frac{f_{x_1 x_2}(x_1 y)}{f_{x_2}(y)} \]

provided \( f_{x_2}(y) \)

and for \(-\infty < x < \infty\)

**Correlation and covariance**

two random variables over the same probability space will finite second moments

Correlation \( \text{corr} = \frac{\text{cov}(x, y)}{\sqrt{\text{var}(x)} \sqrt{\text{var}(y)}} \)

Covariance \( \text{var}(x, y) = \text{cov}(x - \text{E}[x], y - \text{E}[y]) \)

Correlation coefficient \( \rho_{xy} = \frac{\text{cov}(x, y)}{\sqrt{\text{var}(x) \text{var}(y)}} \)

A number of useful properties of correlations/covariances follows from Schwartz's inequality

\[ |\text{E}[xy]| \leq \sqrt{\text{E}[x^2] \text{E}[y^2]} \]

Note: we assumed \( \text{E}[x^2], \text{E}[y^2] \) are finite

**Hölder's inequality** Let \((S, \mathcal{S}, \mu)\) be a measure space, \( p, q \in [1, \infty] \) s.t. \( \frac{1}{p} + \frac{1}{q} = 1 \). Then, for all measurable IR-or C-valued functions \( f, g \) over \( S \), we have

\[ \|fg\|_1 \leq \|f\|_p \|g\|_q \]

where \( \|f\|_p = \left( \int_S |f|^p \, d\mu \right)^{1/p} \)

**Minkowski's inequality**

\[ \|f + g\|_p \leq \|f\|_p + \|g\|_p \]


Observe that 
\[(a+b)^2 = a^2 + b^2 + 2ab \leq a^2 + b^2 + \frac{2(a^2 + b^2)}{2} = 2a^2 + 2b^2\]

so that \(\bar{E}(x^2), \bar{E}(y^2)\) being finite implies \(\bar{E}((x+\alpha y)^2)\) being finite for some constant \(\alpha\).

Take \(\alpha = -\frac{\bar{E}(xy)}{\bar{E}(y^2)}\)

\[0 \leq \bar{E}\left[(x - \frac{\bar{E}(xy)}{\bar{E}(y^2)} y)^2\right] = \bar{E}\left[x^2 - 2\frac{\bar{E}(xy)^2}{\bar{E}(y^2)} + \frac{\bar{E}(xy)^2}{\bar{E}(y^2)}\right] = \bar{E}(x^2) - \bar{E}(xy)^2/\bar{E}(y^2) \geq \bar{E}(xy)^2 \leq \bar{E}(x^2)\bar{E}(y^2)\]

Cauchy–Schwarz gives
\[|\text{Cov}(x, y)| \leq \sqrt{\text{Var}(x)\text{Var}(y)}\]

and consequently, \(-1 \leq \text{Cov}(x, y) \leq 1\).

It is also useful to know that
\[\text{Cov}(x+y, u+v) = \text{Cov}(x, u) + \text{Cov}(x, v) + \text{Cov}(y, u) + \text{Cov}(y, v)\]

and \(\text{Cov}(ax+b, cy+d) = ac\text{Cov}(x, y)\)

Inequalities connecting probabilities and expectations:
Markov's inequality
If \(f\) is a monotonically increasing function on \(\mathbb{R}^+\), \(X\) is a RV, \(a > 0\), \(f(a) > 0\)

Then \(P(\{X\geq a\}) \leq \frac{\bar{E}[f(\lfloor X \rfloor)]}{f(a)}\)

Special case \(P(\{X \geq 0\}) = 1\), \(c > 0\)
\[P(\{X > c\}) \leq \frac{\bar{E}[X]}{c}\]

Chebyshev inequality: \(X\) has finite expectation \(\mu\), variance \(\sigma^2\)
\[P(\{|X - \mu| \geq \alpha\}) \leq \frac{\sigma^2}{\alpha^2}\]

Jensen's inequality
If \(X\) is a RV, \(f\) a convex function, then
\[\bar{E}[f(X)] \leq f(\bar{E}[X])\]
Some examples

Buffon's needle problem
See NSTE (Office for Hall, Science and Technology education)

Needle of length \(b=1\)
Regular grid of parallel lines at distance \(d=1\) from each other
You throw a needle randomly

\[
\begin{align*}
&\text{If } \frac{1}{2} \sin(x) \geq x, \text{ we will intersect the line, if } x > \frac{1}{2} \sin(x) \text{ we will not}
\\
&\text{Joint pdf of } (D, X)
\\
&f(x, d) = \frac{2}{\pi} \cdot \frac{4}{1/2} = \frac{4}{\pi}
\\
&P = \int_{-\pi/2}^{\pi/2} \int_{-\pi/2}^{\pi/2} \frac{4}{\pi} \sin(x) \, dx \, dd = \frac{4}{\pi} \int_{-\pi/2}^{\pi/2} \frac{1}{2} \sin(x) \, dx = \frac{2}{\pi} (-\cos(x)) \bigg|_{-\pi/2}^{\pi/2} = \frac{2}{\pi}
\end{align*}
\]
Additional discussion

There is no set cardinality strictly between that of the integers and reals. Infinite sets have different cardinalities, which are described by the so-called aleph numbers ($\aleph_0, \aleph_1, \aleph_2, \ldots$)

$\aleph_0$ = cardinality of integers, rational numbers, etc

$2^{\aleph_0}$ = cardinality of reals

$\aleph_1$ = aleph-one

Continuum hypothesis (Cantor) $2^{\aleph_0}$

We need an extension theorem from measure theory to ensure consistent definition of probability.

$f$ is a measure to accommodate the following:
if $B \subseteq A \in \mathcal{F}$ and $P(A) = 0 \implies B \in \mathcal{F}$ and $P(B) = 0$

Lebesgue measure

Given a set $E \subseteq \mathcal{F}$ and length of $I = [a, b]$ given by $l(I) = b - a$, the Lebesgue outer measure $\lambda^*(E)$ is defined as

$$\lambda^*(E) = \inf \left\{ \sum_{k=1}^{\infty} l(I_k) : (I_k)_{k \in \mathbb{N}} \text{ is a sequence of open intervals with } E \subseteq \bigcup_{k=1}^{\infty} I_k \right\}$$

Note: every Borel set is Lebesgue measurable (converse is not true)

Note on axiom of choice

"Choice" is associated with distinguishable characteristic. If no such characteristic is available and we have infinitely many choices to make, we need to invoke AC.

Note on Gödel's incompleteness theorem

Any statement expressed using axioms should be provable true or false, in case that the axioms are complete, Gödel's IT asserts (vaguely) that any consistent set of axioms based on which some form of arithmetic can be carried out is incomplete.
Mar/Tue
6 7  Hurd

Mon
27  Feb

April
3/4/5
M  W
April
10
Convergence of random variables

\[ X : \Omega \rightarrow \mathbb{R} \] over some prob. space \((\Omega, \mathcal{F}, P)\)

s.t. if \(A \subseteq \mathcal{F}^B \) over \(\mathbb{R} \), then \(X^{-1}(A) \in \mathcal{F}\)

What happens at each point?

Functions converge to another function: \(X_1, \ldots, X_n, \ldots \rightarrow X\)

Find the set of points \(w\) on which \(X_1(w), X_2(w), \ldots\) converges to some value \(X(w)\) (e.g., \(w^*\)). Denote this set of points an \(\mathcal{C}\), and let the values of the limits be summarized by a function \(X(w)\). There may be some points on \(\Omega\) (e.g., \(w^{**}\)) where there is no convergence. Then, let \(X(w)\) be arbitrary.

We say that \(X_n(w) \rightarrow X(w)\) a.s. (almost surely) if \(P(C) = 1\).

Alternatively, \(X(w)\) is the almost sure limit of a sequence of random variables \(X_n(w), n=1,2,\ldots\) defined over the same probability space if \(P\{ \lim_{n \to \infty} X_n = X \} = 1\).

### Example 1 \((X_n: n \geq 1)\)

\[
X_n(w) = \frac{w^n}{n} \quad w \in [0, 1] = \Omega
\]

\[
X_n(w) = \begin{cases} 
0, & 0 < w < 1 \\
1, & w = 1 
\end{cases}
\]

\[
P\{1 \neq 0\} = 0
\]

\[
\lim_{n \to \infty} X_n(w) = 0 \quad \text{or} \quad X_n(w) \rightarrow 0
\]

### Example 2

\(X_1, X_2, \ldots\) over \(\Omega = \{1, 2, \ldots\}\), same prob. space

\[
X_n(w) = \begin{cases} 
\frac{n}{n+w}, & w = 1 \\
(-1)^n, & w = T
\end{cases}
\]

if \(w = 1\) \(X_1(w) = \frac{1}{2}, X_2(w) = \frac{2}{3}, \ldots\) \(X_n(w) \rightarrow 1\)

if \(w = T\) \(X_1(w) = -1, X_2(w) = \frac{1}{2}, X_3(w) = -1, \ldots\) \(X_n(w)\) does not
So, if the coin is fair \( P(H) = P(T) = 1/2 \) then \( X_n(w) \) does not converge to any random variable.

A sequence of random variables \( X_n \to X \) almost surely if

\[
\sum_{n=1}^{\infty} P\{ \omega : |X_n(\omega) - X(\omega)| > \varepsilon \} < \infty
\]

but not if

\[
\sum_{n=1}^{\infty} P\{ \omega : |X_n(\omega)| > \varepsilon \} = \infty
\]

Ex: Show that \( X_n, n = 1, 2, \ldots \), defined according to

\[
X_n = \begin{cases} 
-\frac{4}{n}, & \text{prob. } \frac{1}{2} \\
\frac{4}{n}, & \text{prob. } \frac{1}{2}
\end{cases}
\]

converges a.s. to 0

\[
P\{ \omega : |X_n(\omega) - X(\omega)| > \varepsilon \} = P\{ \omega : |X_n(\omega)| > \varepsilon \}
\]

| \( X_n(\omega) > \varepsilon \) only happens if \( n < 1/\varepsilon \) (since \( |X_n(\omega)| = \frac{4}{n} \))
\[
\sum_{n=1}^{\infty} P\{ |X_n| > \varepsilon \} = \sum_{n=1}^{\left\lfloor \frac{1}{\varepsilon} \right\rfloor} P\{ |X_n| > \varepsilon \} \leq \left\lfloor \frac{4}{\varepsilon} \right\rfloor < \infty
\]

Convergence in probability

\( X_n \overset{\text{prob.}}{\to} X \) where \( X_1, X_2, \ldots, X \) are over the same prob. space if for any \( \varepsilon > 0 \)

\[
\lim_{n \to \infty} P\{ |X - X_n| > \varepsilon \} = 0
\]

\( X_n \overset{p}{\to} X \)

\[
\lim_{n \to \infty} P\{ \omega : |X(\omega) - X_n(\omega)| > \varepsilon \} = 0 \quad \text{Compare to (\ast \ast)}
\]

Analogy: Members of a club

Attendance of meetings of the club

Almost sure: almost all members have perfect attendance

Probability: almost all meetings when full

Compare \( P(\omega : \lim_{n \to \infty} X_n(\omega) = X(\omega)) = 1 \) vs.

\[
\lim_{n \to \infty} P(\omega : |X_n(\omega) - X(\omega)| > \varepsilon) = 0
\]
A sequence of RVs $X_1, X_2, \ldots$ converges in the L1 sense if all variables are over the same prob. space $E[|X_n|^1] < \infty$ and
$$
lm_{n \to \infty} E[(X_n - X)^2] = 0 \quad X_n \xrightarrow{m.s.} X
$$

From the definition, it follows that $X_n \xrightarrow{m.s.} X \Rightarrow E[X^2] < \infty$

$$
lm_{n \to \infty} E[(X_n - X)^2] = 0 \Rightarrow \exists \varepsilon s.t. \forall n \forall \varepsilon \quad E[(X_n - X)^2] < \varepsilon
$$

Using Markov's inequality:

$$
E[X^2] = E[(X - X_{n_0} + X_{n_0})^2]^{1/2} \leq E[(X - X_{n_0})^2]^{1/2} + E[(X_{n_0})^2]^{1/2} < \infty
$$

Def: A sequence of RVs $X_1, X_2, X_3, \ldots$ converges in distribution to a RV $X$

$$
\lim_{n \to \infty} F_n(x) = F(x) \text{ at all points of continuity of } F
$$

$$
X_n \xrightarrow{d} X \Rightarrow E[f(X_n)] = E[f(X)] \quad n \to \infty
$$

for all continuous bounded functions $f$

$$
\Phi_{X_n} \xrightarrow{p} \Phi_X \quad \Phi = \text{Characteristic function}
$$

pointwise convergence

Example

$X_1, X_2, \ldots, X_n$ iid uniform $[0, 1]$

$$
M_n = \max(X_1, \ldots, X_n)
$$

$N_n \xrightarrow{p} ?$

We show that $M_n \xrightarrow{p} 1$

The CDF of $M_n$ is $F_n(x) = x^n, x \in [0, 1]$

For $\varepsilon > 0$, $P(\frac{1}{n} \leq M_n - 1 < \varepsilon) = P(\frac{1}{n} \leq M_n < 1 + \varepsilon) = (1 - \varepsilon)^n$

Hence $\lim_{n \to \infty} P(\frac{1}{n} \leq M_n - 1 < \varepsilon) = 0$

Next, we show that $n(1 - M_n)$ has an interesting limiting distribution

$$
P(\sqrt{n}(1 - M_n) \leq \frac{x}{n} = P(M_n \geq 1 - \frac{x}{n}) = 1 - (1 - \frac{x}{n})^n
$$

$\Rightarrow 1 - \exp(-x)$ as $n \to \infty$
Relationships between different modes of convergence

\[ X_n \xrightarrow{a.s.} X \ \implies \ X_n \xrightarrow{p} X \ \quad \ \quad X_n \xrightarrow{m.s.} X \ \implies \ X_n \xrightarrow{d} X \]

\[ P\left\{ |X_n - Y| \leq \epsilon \right\} \xrightarrow{n \to \infty} 1 \quad \text{for some RV } Y \text{ with finite second moment then } X_n \xrightarrow{p} Y \implies X_n \xrightarrow{d} Y \]

A sequence of RVs can have only one limit, up to differences on a set of probability zero.

\[ X_n \xrightarrow{d} X, \ X_n \xrightarrow{d} Y \implies X, Y \text{ have the same distribution} \]

Suppose that \( X_n \) has density \( f_n, f_n \downarrow f \), and \( X \) has density \( f \).

Then \( f_n(x) \to f(x) \) for all but countably many \( x \)

\[ \implies X_n \xrightarrow{d} X \quad \text{(Scheffe's theorem)} \]

The converse is not true.

A sequence of discrete RVs can converge in distribution to a continuous RV and vice versa.

Suppose that \( X_1, X_2, \ldots \) are iid RVs with

\[ P\{X_1 = j/10^n\} = \frac{4}{40}, \quad j = 0, 1, \ldots, 9 \]

\[ U_n = \frac{n}{2} \frac{X_n}{10^n} \]

\[ U_n \xrightarrow{d} U \quad \text{where } U \text{ is uniform } [0, 1] \]

\[ P\{U_n = \frac{j}{40^n}\} = \frac{4}{40^n}, \quad j = 0, \ldots, 9 \cdot 10^n - 9 \]

For \( j/10^n \leq x < (j+1)/10^n \)

\[ P\{U_n \leq x\} = \frac{j+1}{40^n} \]

and \( |P\{U_n \leq x\} - x| \leq 10^{-n} \implies 0 \quad n \to \infty \)

\[ \implies P\{U_n \leq x\} \to x \quad \forall x \in [0, 1] \]
Continuous mapping theorem
Suppose that $g$ is $\mathbb{R}$-valued and continuous

$$X_n \xrightarrow{p} X \Rightarrow g(X_n) \xrightarrow{p} g(X)$$

$$X_n \xrightarrow{d} X \Rightarrow g(X_n) \xrightarrow{d} g(X)$$

 Slutsky's theorem

$$X_n \xrightarrow{d} X$$

$$Y_n \xrightarrow{d} \theta, \text{ a constant}$$

Then

$$X_n + Y_n \xrightarrow{d} X + \theta$$

$$X_n Y_n \xrightarrow{d} \theta X$$

Some proofs

$$X_n \xrightarrow{a.s.} X \Rightarrow X_n \xrightarrow{p} X$$

For some $\epsilon > 0$, let

$$A_n = \{\omega : |X_n(\omega) - X(\omega)| < \epsilon\}$$

Need to show that $a.s. \Rightarrow P(\bigcup_{n=1}^{\infty} A_n) = 1$

$$B_n = \{\omega : |X_k(\omega) - X(\omega)| < \epsilon, \forall k \leq n\}$$

$$B_n \subseteq A_n$$

$$B_1 \subseteq B_2 \subseteq B_2 \subseteq \ldots$$

$$\lim_{n \to \infty} P(B_n) = P(\bigcup_{n=1}^{\infty} B_n) = P(B)$$

and

$$\omega : \lim_{n \to \infty} X_n(\omega) = X(\omega) \subseteq B$$

$a.s. \Rightarrow P(B) = 1 = \lim_{n \to \infty} P(B_n) \Rightarrow \lim_{n \to \infty} P(A_n) = 1 \Rightarrow X_n \xrightarrow{p} X$

Suppose that $X_n \xrightarrow{\text{m.s.}} X$ and let $\epsilon > 0$

Markov's inequality $\forall \epsilon > 0$

$$P\left(\frac{1}{n} \sum_{j=1}^{n} (X_j - X) \geq \epsilon\right) \leq \frac{E\left[\sum_{j=1}^{n} (X_j - X)^2\right]}{\epsilon^2}$$

and

$$P(X_n - X \geq \epsilon) \to 0$$

$$P(1 | X_n - X | > \epsilon) \to 0$$

i.e.

$$X_n \xrightarrow{p} X$$
Notice that $A$ is an $MU$ code, since according to the Condition 2 $C$ is an $MU$ code and adding runs of ones does not violate the $MU$ property in $A$. Moreover, our construction guarantees that each substring of length $f$ of an element in $A$ contains $1^{f+1}$ and avoids $0^{f+1}$ as a subword. So, there is no $0^{f+1}$ to bond with $1^{f+1}$ and form a $PD$ product of length at least $f$ in $A$. Therefore, $A$ is a binary $f-APD$ and $MU$ code. In addition, one can verify that $A$ also inherits the Error-Correcting property from $C$, and both codes have the same minimum Hamming distance.

**Theorem 1.** Let $A_{APD\_MU}(n, f)$ denote the maximum size of a $f-APD\_MU$ code over a binary alphabet, for positive integers $n = pf$. Then there exist constants $0 < C_6 < C_7$ such that

$$C_6 \frac{2^n}{n^{p+1}} \leq A_{APD\_MU}(n, f) \leq C_7 \frac{2^n}{n^{p+1}}$$

**Proof.** To justify the lower bound we use the aforementioned construction. In this construction $|A| = |C|$. When $2l \leq n - 2$ the number of binary sequences of length $x - l - 2$ containing no $l$ consecutive zeros is at least

$$2^{x-1}-2 -(x - 2l - 1)2^{x-2l-2}$$
$$\geq 2^{x-1}-2 - n2^{x-2l-2}$$
$$= 2^{n-p-2-l(p+1)} - n2^{n-p-2-l(p+2)}$$
$$= 2^{n-p-2} \left[ 2^{-l(p+1)} - n2^{-l(p+2)} \right]$$

The function $2^{-l(p+1)} - n2^{-l(p+2)}$ is maximized when $l = \log_2 \left[ n \left( \frac{p+1}{p+2} \right) \right] + \delta$, where $\delta$ is chosen so that $|\delta| < 1$ and $l$ is an integer. In this case, the value of $2^{-l(p+1)} - n2^{-l(p+2)}$ is bounded below by $\frac{C_6}{n^{p+1}}$. $\square$