Applied Probability and Statistics I - STAT400

Tom Mitchell

Mestiyage - Fall 2024

Syllabus

Grading

- Homework — 28% (4% each)
- R Projects — 12% (4% each)
- Two exams — 30% (15% each)
- Final exam — 30%

Office Hours

- Tuesday: 1:00 PM - 1:50 PM (in person, MTH 4106)
- Wednesday: 11:00 AM - 11:50 AM (online)

Exams

- 2 midterms and a final exam

Lecture 1: Tuesday 8/27/2024

Course Overview: STAT400

We study:

1. Probability
2. Descriptive Statistics
3. Inferential Statistics

1. Probability

Probability is the mathematical study of uncertainty.

2. Descriptive Statistics

Descriptive Statistics involves methods for summarizing and describing the important characteristics of a dataset.

3. Inferential Statistics

Inferential Statistics involves methods for using data from a subset (sample) of a larger group (population) to make meaningful conclusions. Note that each time you pick a subset, you lose out on certain information, leading to uncertainty.

Setting: Conducting an Experiment

In this course, we will often assume we are about to conduct an experiment. The possible outcomes of the experiment are known, but the exact outcome is not.

Examples:

- Tossing a coin
- Rolling a die

To study these types of situations, we introduce a *model for probability*.

Ordered Triples $(\Omega, \mathcal{F}, P)$

- Ω is the **sample space**: The set of all possible outcomes of the experiment.
- $\mathcal{F}$ is the **event space**:
  - An **event** is a subset of Ω.
  - An event captures the idea that in many situations, we care about collections of outcomes rather than a single outcome.
  - The event space contains collections of outcomes for which we can assign probabilities.

- P is the **probability measure**:
$$P : \mathcal{F} \to \mathbb{R}$$
where P assigns probabilities to events.

Example 1: Tossing a Fair Coin

Consider the simple experiment of tossing a **fair** coin.

$$\Omega = \{H, T\}$$

$$\mathcal{F} = \{\emptyset, \{H\}, \{T\}, \{H, T\}\}$$

- $P(\emptyset) = 0$
("A probability of 0 represents an outcome/collection of outcomes that never takes place.")
- $P(\{H, T\}) = 1$
("A probability of 1 represents an outcome/collection of outcomes that always takes place.")
- $P(\{H\}) = \frac{1}{2}$
(Requires the coin to be fair.)
- $P(\{T\}) = \frac{1}{2}$
(Requires the coin to be fair.)

Example 2: Rolling a Fair Die

Consider the experiment of rolling a fair die. Note: ($\mathcal{P}(\Omega)$ is the Powerset of Ω)

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

$$\mathcal{F} = \mathcal{P}(\Omega) =$$
$$\{\emptyset, \{1\}, \{2\}, \ldots, \{6\}, \{1, 2\}, \ldots, \{1, 6\}, \{2, 3\}, \ldots,$$
$$\{1, 2, 3\}, \{2, 3, 4\}, \ldots, \{1, 2, 3, 4\}, \{1, 2, 3, 4, 5\}, \ldots, \Omega = \{1, 2, 3, 4, 5, 6\}\}$$

- $P(\{i\}) = \frac{1}{6}$ for $i = 1, 2, 3, 4, 5, 6$
- $P(\emptyset) = 0$
- $P(\Omega) = 1$
- $P(\{1, 2\}) =?$
(Listing out probabilities for each event is **not** feasible.)

We would want a way to calculate the probability of a particular event by using the "base" probabilities.

We will look at these rules for computing probabilities of complicated events in terms of basic ones in a bit.

Example 3: Tossing a Fair Coin Until We See Heads for the First Time

Consider the experiment of tossing a fair coin repeatedly until we observe heads for the first time, then we stop.

$$\Omega = \{(H), (T,H), (T,T,H), (T,T,T,H), \ldots\}$$

$$\mathcal{F} = \mathcal{P}(\Omega)$$

Let A_i be the event of seeing heads on the ith toss.

$$P(A_i) = \frac{1}{2^i}$$

Question: What is the probability of seeing heads on an even-numbered toss?

$$Q = P(\text{seeing heads on an even-numbered toss}) = ?$$

Example 4: Picking a Number from the Interval [0, 1]

Consider picking a number from the interval $[0,1]$ such that the probabilities of selecting numbers from two subintervals I_1 and I_2 of $[0,1]$ are the same whenever I_1 and I_2 have the same length.

$$\Omega = [0,1]$$

$$\mathcal{F} = \mathcal{P}(\Omega)$$

$$P(\text{selecting a number from a subinterval } I \text{ of } [0,1]) = \text{length}(I)$$

$$P([0, \frac{1}{4}]) = \frac{1}{4} - 0 = \frac{1}{4}$$

$$P([0, \frac{1}{2}]) = \frac{1}{2} - 0 = \frac{1}{2}$$

$$P\left(\left\{\frac{1}{2}\right\}\right) = P\left(\left[\frac{1}{2}, \frac{1}{2}\right]\right) = \frac{1}{2} - \frac{1}{2} = 0$$

"The probability of 0 indicates an event that does not take place or an event that is so unlikely that the only reasonable value to assign to it is 0, i.e., an event that almost surely does not take place."

On Thursday will learn how to compute problems similar to $P(\mathbb{Q} \cap [0,1]) = ?$

Lecture 2: Thursday 8/29/2024

Probability

- Mathematical study of uncertainty.

Probability Space

$$(\Omega, F, P)$$

- Ω - All possible outcomes of the experiment of interest.
- F - Contains events, i.e., collections of outcomes we would like to assign probabilities to.
- P - Probability measure.

$$P : F \to \mathbb{R}$$

On Tuesday, we discussed examples and introduced some formulas to compute probabilities of "basic" events.

Goal for Today

Explore how to compute probabilities of "complicated" events in terms of basic events.

Two Questions

1. How do we express a complicated event in terms of basic ones?
2. What rules can be used to compute probabilities once Question 1 has been answered?

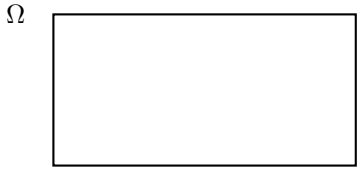
- Question 1 is best answered by using tools from set theory.

$\to$ We treat Ω as the list of all possible outcomes.

$\to$ Events are sublists of Ω.

Since we are only interested in outcomes (or combinations thereof) of the experiment, we treat Ω as the universal set (i.e., we pretend that nothing outside of Ω exists).

Pictorial Representation

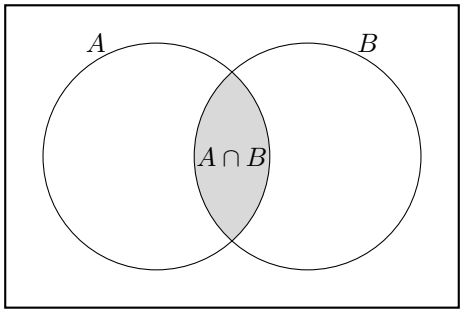


We can manipulate collections of events to obtain new events in various ways:

- Intersections
- Unions
- Relative complements

Given events A and B, the intersection is the event that contains all outcomes common to both A and B. We denote this by $A \cap B$.

Intersection of Two Events



Example

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

Let $A = \{1, 3, 5\}$ and $B = \{4, 5\}$. Then, the intersection of A and B is:

$$A \cap B = \{5\}$$

Extending the Intersection to Families of Events

The idea of the intersection of events can be extended to families of events.

$$\bigcap_{i=1}^{3} A_i = (A_1 \cap A_2) \cap A_3$$

For an infinite family of events:

$$\bigcap_{i=1}^{\infty} A_i = A_1 \cap A_2 \cap A_3 \cap A_4 \cap A_5 \cap \ldots$$

Disjoint Events

Two events are said to be disjoint if $A \cap B = \emptyset$.

- Events $A_1, \ldots, A_n$ are said to be disjoint if $\bigcap_{i=1}^{n} A_i = \emptyset$.
- Events $A_1, \ldots, A_n, \ldots$ are said to be disjoint if $\bigcap_{i=1}^{\infty} A_i = \emptyset$.
- Events $A_1, A_2, \ldots$ are said to be pairwise disjoint if $A_i \cap A_j = \emptyset$ for all $i \neq j$.

Example

Consider the sample space $\Omega = \{1, 2, 3, 4, 5, 6\}$ with the following events:

$$A = \{1, 3, 5\}, \quad B = \{4, 5\}, \quad C = \{6\}$$

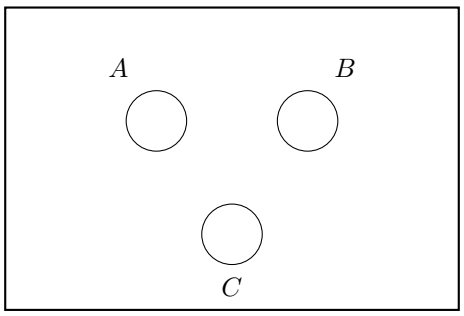
Here, A, B, and C are disjoint events, meaning:

$$A \cap B \cap C = \emptyset$$

However, they are not pairwise disjoint, because:

$$A \cap B = \{5\} \neq \emptyset$$

Pairwise Disjoint Sets

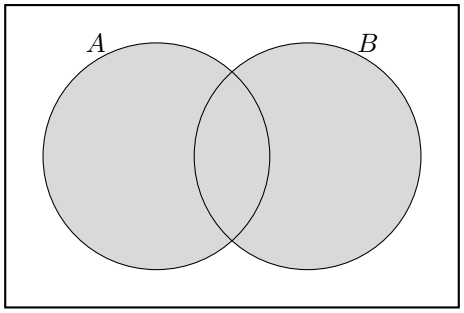


The word "and" usually translates to intersection when discussing sets or events. Conversely, the word "or" translates to a union.

Unions

The union of two events A and B, denoted by $A \cup B$, is the event that contains all outcomes that are in A, B, or both.

Venn Diagram: Union of Two Sets



$A \cup B$ (shaded region)

Example

Consider the sample space $\Omega = \{1, 2, 3, 4, 5, 6\}$ with the following events:

$$A = \{1, 3, 5\}, \quad B = \{2, 3, 5\}$$

The union of A and B is:

$$A \cup B = \{1, 2, 3, 5\}$$

Infinite and Finite Unions

Infinite versions of unions and finite versions of unions with more than two sets can be similarly defined.

$$\bigcup_{i=1}^{3} A_i = A_1 \cup A_2 \cup A_3$$

Similarly, the union of more than two finite sets can be defined as:

$$\bigcup_{i=1}^{n} A_i = A_1 \cup A_2 \cup \ldots \cup A_n$$

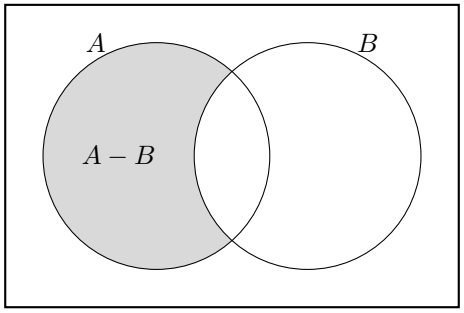
Unions can also be extended to infinite collections of sets. For example, the union of an infinite sequence of sets $A_1, A_2, \ldots$ is denoted as:

$$\bigcup_{i=1}^{\infty} A_i = A_1 \cup A_2 \cup \ldots$$

Relative Complements

Given events A and B, the relative complement $A - B$ is the event that contains only the outcomes that are **unique** to A.

Venn Diagram: Unique to A ($A - B$)



Example

Consider the sample space $\Omega = \{1, 2, 3, 4, 5, 6\}$ with the following events:

$$A = \{1, 2, 3\}, \quad B = \{2, 4, 5\}$$

The relative complements are:

$$A - B = \{1, 3\}, \quad B - A = \{4, 5\}$$

Complement of A

We use A^c to denote the complement of A within the universal set Ω, which is the set of outcomes that are not in A:

$$A^c = \Omega - A = \{4, 5, 6\}$$

The word **not** is associated with complements and the phrases "and not", "but not" are associated with relative compliments

Aside (for the HW): Verifying Set Identities with Venn Diagrams

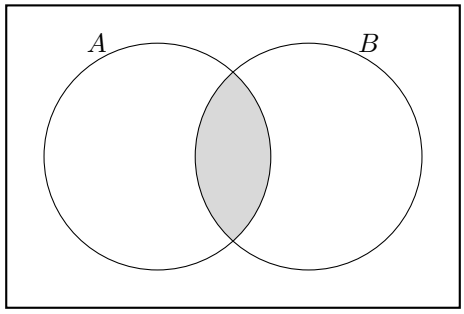
We can use Venn diagrams to verify set identities. For example, let's verify the identity:

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

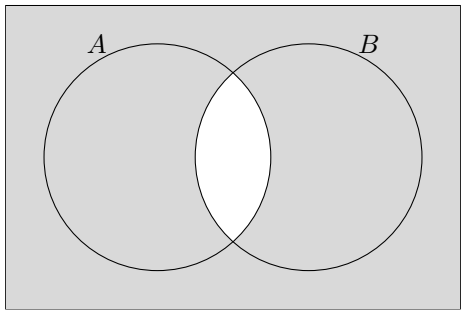
Venn diagrams representing set operations of (left hand side) LHS and (right hand side) RHS respectively are shown below:

LHS

$A \cap B$

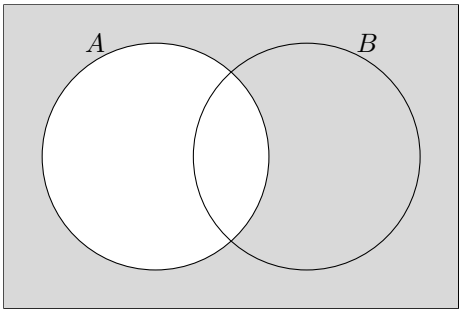


$(A \cap B)^c$

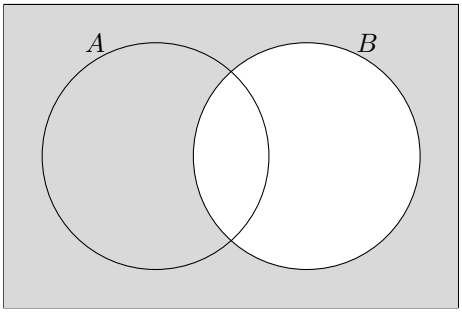


RHS

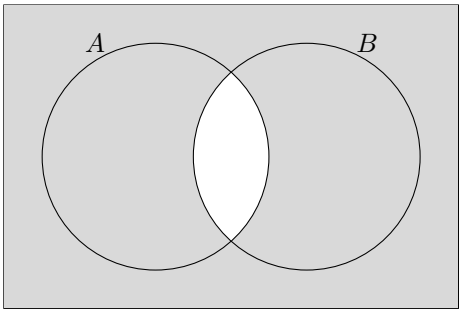
A^c



B^c



$A^c \cup B^c$



Conclusion

Both sides of the identity are visually represented by the same shaded region, thus confirming that:

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

The shaded regions for $(A \cap B)^c$ and $A^c \cup B^c$ match. Therefore, $(A \cap B)^c = A^c \cup B^c$.

Note: (we are not worrying about edges cases, for example, where A and B are disjoint but the identity still holds true)

Now that the terminology is in place, we start to answer Q2.

Axioms

1. $P(\Omega) = 1$, $P(\emptyset) = 0$
2. $P(A) \geq 0 \quad$ for all $A \in \mathcal{F}$
3. If $A_1, A_2, \ldots, A_n, \ldots \in \mathcal{F}$ are pairwise disjoint, then

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

Example: Consider the problem of tossing a fair coin until we observe heads. We want to find P(heads on an even numbered toss)

Let A be the event that heads appears for the first time on an even-numbered toss.

Thus, A can be expressed as:

$$A = \{(T, H), (T, T, T, H), (T, T, T, T, T, H), \ldots\}$$

Recall, A_i is heads on the ith toss.

$$P(A_i) = \frac{1}{2^i}$$

Notice that we can describe A as the union of disjoint events where heads occurs for the first time on the $2i$-th toss:

$$A = A_2 \cup A_4 \cup A_6 \cup \cdots = \bigcup_{i=1}^{\infty} A_{2i}$$

where A_{2i} is the event that heads first appears on the $2i$-th toss. The probability of A_{2i} can be computed as:

$$P(A_{2i}) = \left(\frac{1}{2}\right)^{2i}$$

Since the events A_{2i} are pairwise disjoint, we can use axiom 3 to see the probability of A is given by:

$$P(A) = P\left(\bigcup_{i=1}^{\infty} A_{2i}\right) = \sum_{i=1}^{\infty} P(A_{2i}) = \sum_{i=1}^{\infty} \left(\frac{1}{2}\right)^{2i} = \sum_{i=1}^{\infty} \left(\frac{1}{4}\right)^{i}$$

This series is an infinite geometric series with the first term $a = \frac{1}{4}$ and common ratio $r = \frac{1}{4}$:

Recall, that a the sum of a converging infinite geometric series where $|r| < 1$ is:

$$S = a + ar + ar^2 + ... = \sum_{i=0}^{\infty} (ar^n) = \frac{a}{1-r}$$

Therefore:

$$\sum_{i=1}^{\infty} \left(\frac{1}{4}\right)^{i} = \frac{\frac{1}{4}}{1-\frac{1}{4}} = \frac{\frac{1}{4}}{\frac{3}{4}} = \frac{1}{3}$$

Thus, the probability that the first occurrence of heads is on an even-numbered toss is $\boxed{\frac{1}{3}}$.

Derived rule: If $A_1, \ldots, A_N \in \mathcal{F}$ are pairwise disjoint, then

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

Example: Consider a fair die roll. What is the probability of rolling a 1, 3, or 5?

$$P(\{i\}) = \frac{1}{6} \quad \text{for } i = 1, \ldots, 6$$

$$P(\{1,3,5\}) = P(\{1\}) + P(\{3\}) + P(\{5\}) = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{1}{2}$$

Lecture 3: Thursday 9/3/2024

1 Probability

Definition: Probability is the mathematical framework for quantifying and analyzing uncertainty. A probability space is defined by the triple $(\Omega, \mathcal{F}, P)$, where:

- Ω is the sample space (set of all possible outcomes)
- $\mathcal{F}$ is the event space (sigma-algebra of subsets of Ω)
- P is the probability measure (function assigning probabilities to events)

Axioms of Probability

The probability measure P satisfies the following axioms:

1. $P(\Omega) = 1$ and $P(\emptyset) = 0$
2. For all $A \in \mathcal{F}$, $P(A) \geq 0$
3. (Countable additivity) If $A_1, A_2, \ldots \in \mathcal{F}$ are pairwise disjoint events (i.e $A_i \cap A_j = \emptyset$ for all $i \neq j$):

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

Note: Understanding and being able to derive properties from these axioms is crucial for the first midterm and for building a solid foundation in probability theory.

Important Results

1. **Decomposition of Probability:** For any two events $A, B \in \mathcal{F}$:

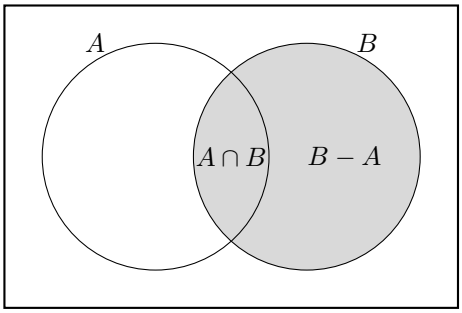
$$P(B) = P(B - A) + P(A \cap B)$$

Proof:

- Observe that B can be partitioned into two disjoint sets: $B = (B - A) \cup (A \cap B)$
- Note that $(B - A) \cap (A \cap B) = \emptyset$ (these sets are mutually exclusive)
- By the axiom of countable additivity (which implies finite additivity for disjoint events):

$$P(B) = P((B - A) \cup (A \cap B)) = P(B - A) + P(A \cap B)$$

This result is fundamental in probability theory and is often used in solving complex probability problems. It allows us to break down the probability of an event into mutually exclusive parts, which can be easier to calculate individually.



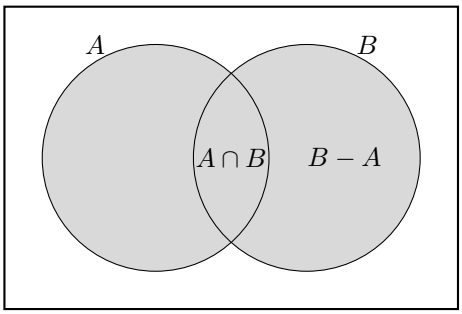
2. **Finite Additivity:** For any two events $A, B \in \mathcal{F}$:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

Proof:

- Observe that $A \cup B = A \cup (B - A)$ and $A \cap (B - A) = \emptyset$
- Since A and $B - A$ are disjoint, $P(A \cup B) = P(A) + P(B - A)$ (finite additivity)
- From the decomposition result: $P(B) = P(B - A) + P(A \cap B)$
- Rearranging: $P(B - A) = P(B) - P(A \cap B)$
- Substituting into our earlier equation:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

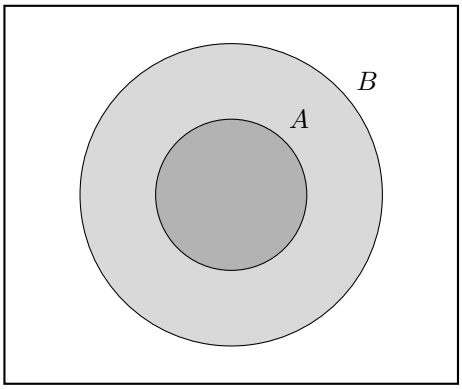


3. **Monotonicity:** If $A \subseteq B$, then $P(A) \leq P(B)$

Proof:

- If $A \subseteq B$, then:
  - $B = A \cup (B - A)$
  - $A \cap (B - A) = \emptyset$
- By finite additivity: $P(B) = P(A \cup (B - A)) = P(A) + P(B - A)$
- Since probabilities are non-negative, $P(B - A) \geq 0$

- Therefore, $P(B) \geq P(A)$

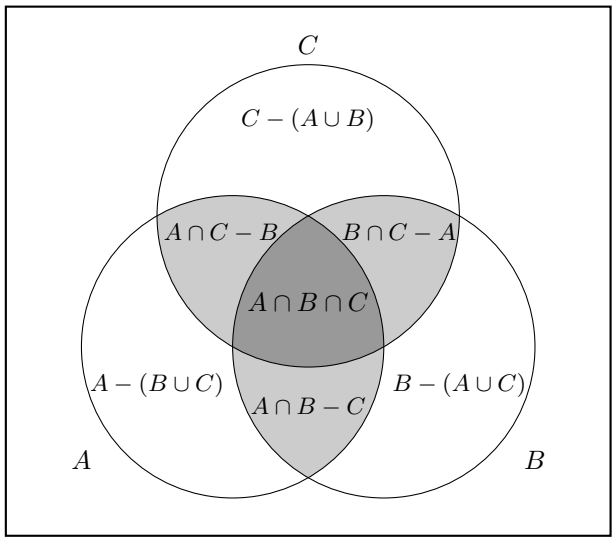


4. **Union of Three Sets:** For $A, B, C \in \mathcal{F}$:

$$P(A \cup B \cup C)$$

How to compute:

$$\begin{aligned}
P(A \cup B \cup C) &= P(A) + P(B \cup C) - P(A \cap (B \cup C)) \\
&= P(A) + [P(B) + P(C) - P(B \cap C)] - P(A \cap (B \cup C)) \quad \text{(decompose the union)} \\
&= P(A) + P(B) + P(C) - P(B \cap C) - P((A \cap B) \cup (A \cap C)) \quad \text{(by distributive law)} \\
&= P(A) + P(B) + P(C) - P(B \cap C) - (P(A \cap B) + P(A \cap C) - P((A \cap B) \cap (A \cap C))) \\
&= P(A) + P(B) + P(C) - P(B \cap C) - (P(A \cap B) + P(A \cap C) - P(A \cap B \cap C)) \\
&= P(A) + P(B) + P(C) - P(B \cap C) - P(A \cap B) - P(A \cap C) + P(A \cap B \cap C)
\end{aligned}$$



Example: Suppose that A and B are events with $P(A) = 0.5$ and $P(B) = 0.6$. Find numbers a and b such that $a \leq P(A \cup B) \leq b$ and $a \leq P(A \cap B) \leq b$.

1. Using the addition rule of probability:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) = 0.5 + 0.6 - P(A \cap B) = 1.1 - P(A \cap B)$$

2. Since $A \cup B \subseteq \Omega$ and $P(\Omega) = 1$, we have:

$$P(A \cup B) \leq P(\Omega) \leq 1$$

3. From steps 1 and 2:

$$1.1 - P(A \cap B) \leq 1 \implies P(A \cap B) \geq 0.1$$

4. The maximum value for $P(A \cap B)$ is $\min(P(A), P(B)) = 0.5$, because:

   - $A \cap B \subseteq A$ and $A \cap B \subseteq B$
   - Therefore, $P(A \cap B) \leq P(A)$ and $P(A \cap B) \leq P(B)$

5. We can conclude:

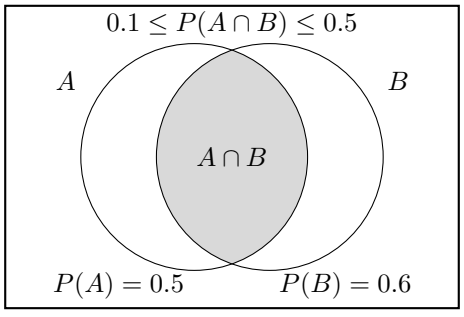
$$0.1 \leq P(A \cap B) \leq 0.5$$
$$0.6 \leq P(A \cup B) \leq 1.0$$

6. To verify the lower bound for $P(A \cup B)$:

$$\begin{aligned} P(A \cup B) &= P(A) + P(B) - P(A \cap B) \\ &\geq P(A) + P(B) - \min(P(A), P(B)) \\ &= 0.5 + 0.6 - 0.5 = 0.6 \end{aligned}$$

Conclusion: We have found that $a = 0.1$ and $b = 0.5$ satisfy $a \leq P(A \cap B) \leq b$, while $a = 0.6$ and $b = 1.0$ satisfy $a \leq P(A \cup B) \leq b$.

Visualization: To better understand the relationship between A and B, consider the following Venn diagram:



This visualization illustrates why $0.1 \leq P(A \cap B) \leq 0.5$ and $0.6 \leq P(A \cup B) \leq 1.0$.

Example: When Alice visits a certain grocery store, she:

- buys apples with probability 0.4
- buys bananas with probability 0.7
- buys both with probability 0.25

Compute the probability that on her next visit to the store, Alice:

1. buys either apples or bananas
2. buys bananas but not apples
3. buys neither apples nor bananas

Step 1: Define notation and given probabilities

$$A : \text{Alice buys apples}$$
$$B : \text{Alice buys bananas}$$

$$P(A) = 0.4$$
$$P(B) = 0.7$$
$$P(A \cap B) = 0.25$$

Note: Symbol correspondences

- $\cup$ corresponds to "or" (union)
- $\cap$ corresponds to "and" (intersection)
- $\complement$ corresponds to "not" (complement)
- $-$ corresponds to "and not", or "but not" (relative complement)

Step 2: Calculate probabilities

1. $P(A \cup B) = P(A) + P(B) - P(A \cap B) = 0.4 + 0.7 - 0.25 = 0.85$
2. $P(B - A) = P(B) - P(A \cap B) = 0.7 - 0.25 = 0.45$
3. $P(A^c \cap B^c) = P((A \cup B)^c) = 1 - P(A \cup B) = 1 - 0.85 = 0.15$

Note: Important probability relationships:

$$P(A) = 1 - P(A^c)$$
$$\Omega = A \cup A^c$$
$$P(\Omega) = P(A) + P(A^c) = 1$$
$$P(A^c) = 1 - P(A)$$

Computing probabilities for finite, equally likely outcomes: For any event $A \subseteq \Omega$, where Ω is a finite sample space with equally likely outcomes, the probability of A is given by:

$$P(A) = \frac{\text{number of elements in } A}{\text{number of elements in } \Omega} = \frac{|A|}{|\Omega|}$$

where $|A|$ denotes the cardinality (number of elements) of set A.

Lecture 4: Thursday 9/5/2024

Probability for Finite, Equally Likely Outcomes

Consider an experiment with finitely many outcomes, each of which is equally likely. This scenario is often referred to as "fair" or "uniformly at random" (with finite outcomes).

In this situation:

- $\mathcal{F} = \mathcal{P}(\Omega)$
- For any event $A \in \mathcal{F}$, $P(A) = \frac{|A|}{|\Omega|}$

Proof that $P(A) = \frac{|A|}{|\Omega|}$ satisfies the axioms of probability:

1. For any outcome $\omega \in \Omega$, we show $P(\{\omega\}) = \frac{1}{|\Omega|}$

   Let $\Omega = \{\omega_1, \omega_2, \ldots, \omega_n\}$. Then:

   - $P(\{\omega_1\}) = P(\{\omega_2\}) = \cdots = P(\{\omega_n\})$ (equally likely)
   - $\{\omega_i\}$ are pairwise disjoint for $i = 1, \ldots, n$
   - $\Omega = \bigcup_{i=1}^{n} \{\omega_i\}$
   - $P(\Omega) = P(\bigcup_{i=1}^{n} \{\omega_i\}) = \sum_{i=1}^{n} P(\{\omega_i\})$ (finite additivity)

   Let $x = P(\{\omega_i\})$ for any i. Then:

   $$1 = P(\Omega) = nx$$
   $$x = \frac{1}{n} = \frac{1}{|\Omega|}$$

   Therefore, $P(\{\omega_1\}) = P(\{\omega_2\}) = \cdots = P(\{\omega_n\}) = \frac{1}{|\Omega|}$

2. For any event $A \subseteq \Omega$, we prove $P(A) = \frac{|A|}{|\Omega|}$

   Let $A = \{\omega_{i_1}, \omega_{i_2}, \ldots, \omega_{i_k}\}$ where $k = |A|$. Then:

   $$\begin{aligned} P(A) &= P(\{\omega_{i_1}\} \cup \{\omega_{i_2}\} \cup \cdots \cup \{\omega_{i_k}\}) \\ &= \sum_{j=1}^{k} P(\{\omega_{i_j}\}) \quad \text{(finite additivity)} \\ &= \sum_{j=1}^{k} \frac{1}{n} \\ &= \frac{k}{n} = \frac{|A|}{|\Omega|} \end{aligned}$$

To properly use the probability formula, we need to be able to count effectively. There are two fundamental principles of counting:

- Additive principle of counting

- Multiplicative principle of counting

Additive Principle of Counting: If we have n ways of doing one thing and m ways of doing another thing, where both actions cannot be performed simultaneously, then there are $n + m$ ways of selecting an action.

Example (Additive Principle): Suppose you own:

- 6 pairs of jeans
- 5 pairs of slacks

Then there are $6 + 5 = 11$ ways of selecting an article of clothing.

Multiplicative Principle of Counting: If we have n ways of doing one thing and m ways of doing another thing, then there are $n \times m$ ways of performing both actions.

Example (Multiplicative Principle): Continuing from the previous example, if you are packing for a trip and want to pack:

- 1 pair of jeans
- 1 pair of slacks

There are $6 \times 5 = 30$ ways of selecting a pair of jeans and a pair of slacks.

Problems we study in this class can be categorized according to two main features:

1. Replacement
2. Order

Examples:

- Passwords: Replacement allowed, order matters
- Card games: Usually no replacement, order may or may not matter (depends on the situation)

We can visualize these categories in a 2x2 grid:

	Replacement	**No Replacement**
Order Matters	✓	✓
Order Doesn't Matter	×	✓

This grid helps us classify different probability problems based on their characteristics. The ✓ and × are used to indicate what we will study in this class.

Problem: Suppose that a password consists of 10 characters. The character set contains 256 possible characters. If characters are chosen uniformly at random, what is the probability of picking a password that starts with an 'A'?

Analysis:

- Replacement: ✓
- Order matters: ✓

Visualization:

- All possible passwords:

$$256, 256, 256, \ldots, 256$$

$$\_\_\_\_\_$$

(256 choices for each of the 10 slots)

- Passwords starting with 'A':

$$A, 256, 256, \ldots, 256$$

$$A\_\_\_\_\_$$

(1 choice for first slot ('A'), 256 choices for each of the remaining 9 slots)

Calculation:

- Total number of possible passwords: 256^{10} (256 choices for each of 10 slots)
- Number of passwords starting with 'A': 1×256^9 (1 choice for first slot, 256 choices for each of 9 slots)
- Probability = $\frac{\text{Favorable outcomes}}{\text{Total outcomes}} = \frac{1 \times 256^9}{256^{10}} = \frac{1}{256}$

Conclusion: The probability of picking a password that starts with an 'A' is $\frac{1}{256}$.

General Rule: If you have a list of n things to choose from, replacement is allowed, order matters, and you choose k times, then there are n^k possible choices.

New Problem: Using the previous example (where $n = 256$ and $k = 10$), what is the probability of randomly picking a password that repeats at least one character?

Strategy: It is sometimes easier to calculate the probability of the complement of an event.

Approach: We will count the number of passwords where each character is unique.

Let A be the event that the randomly chosen password has no repeated characters.

$P(A^c) = 1 - P(A)$

Number of passwords with unique characters:

- Visualization:

$$\underbrace{256 \cdot 255 \cdot 254 \cdot 253 \cdot 252 \cdot 251 \cdot 250 \cdot 249 \cdot 248 \cdot 247}_{\text{10 positions}}$$

- Mathematical representation: $\prod_{i=0}^{9}(256 - i)$
- Probability calculation: $P(A^c) = 1 - \frac{\prod_{i=0}^{9}(256-i)}{256^{10}}$

General Formula: When picking k things out of n things while keeping track of the order (where $k \leq n$):

Number of different ways:

$$\begin{aligned}
&= n \cdot (n-1) \cdot (n-2) \cdot \ldots \cdot (n-k+1) \\
&= n \cdot (n-1) \cdot (n-2) \cdot \ldots \cdot (n-k+1) \cdot \frac{(n-k) \cdot (n-k-1) \cdot \ldots \cdot 1}{(n-k) \cdot (n-k-1) \cdot \ldots \cdot 1} \\
&= \frac{n!}{(n-k)!}
\end{aligned}$$

Where $n!$ is defined as:

$$n! = \prod_{i=1}^{n} i \quad \text{and} \quad 0! = 1$$

$\frac{n!}{(n-k)!}$ is known as "n permute k", denoted as ${}^nP_k$.

No replacement, order does not matter:
The trick is to solve the problem assuming order matters, then adjust by dividing by the number of ways to order the k things.

Consider the following slots:

$$\overbrace{\text{- - - - - - - - - -}}^{k \text{ slots}}$$

First, we count the number of ways to fill these slots while keeping track of the order:

- This is given by ${}^nP_k = \frac{n!}{(n-k)!}$
- It represents the number of ways to choose and arrange k items from n options

However, this count includes permutations of the same k items, which we don't want to distinguish. To adjust for this overcounting:

- We divide by the number of ways to order k items, which is $k!$
  - $k!$ represents all possible permutations of the k chosen items
- This adjustment gives us the following formula:

$$\frac{{}^nP_k}{k!} = \frac{n!}{k!(n-k)!} = {}^nC_k = \binom{n}{k}$$

The result, $\binom{n}{k}$, is known as "n choose k". It represents the number of ways to choose k items from n options when the order doesn't matter.

Lecture 5: Tuesday 9/10/2024

No Replacement, Order Does Not Matter

A common source of examples for this scenario is card games like poker.

Example: Suppose you are dealt 5 cards from a well-shuffled deck of cards. What is the probability that you are dealt a straight (5 cards in sequence)?

- A well-shuffled deck means all 52 cards are equally likely to be drawn (uniformly at random).
- A straight has 5 consecutive ranks but not all of the same suit.

Common Strategy: Undercount and Adjust

- We count a specific case and adjust for the number of cases.
- Recall: For $\binom{n}{k} = \frac{{}^{n}P_k}{k!}$, we overcounted and adjusted by dividing by $k!$.

Let's consider a specific type of straight: one containing the ranks 9, 10, J, Q, K.

- There are 4 different suits, so there are four ways to choose each rank.
- Total number of ways for a specific straight: $4^5 = 1024$
- Subtract straight flushes: $1024 - 4 = 1020$ (4 straight flushes, one for each suit)

The number of all straights is equal to $(4^5 - 4)\times$ number of options for the starting rank.

- Ace can be considered the highest or lowest rank, but not both:
  - A, 2, 3, 4, 5 ✓
  - 10, J, Q, K, A ✓
  - J, Q, K, A, 2 ×
- There are 10 possible starting ranks (Ace(1) through 10)

Therefore, the total number of straights is:

$$10 \times (4^5 - 4) = 10 \times 1020 = 10,200$$

The required probability is:

$$P(\text{Straight}) = \frac{\text{Number of straights}}{\text{Number of possible hands}}$$

$$= \frac{10(4^5 - 4)}{\binom{52}{5}}$$

Conditional Probability

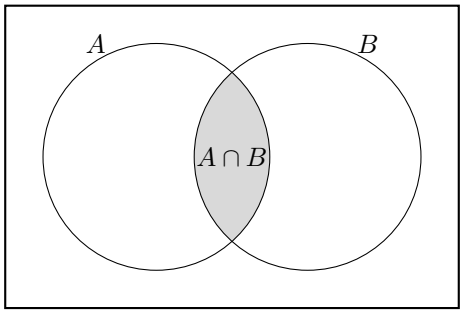
Conditional probability measures the likelihood of an event occurring given that another event has already occurred. It allows us to update our probability estimates based on new information.

Definition: Given two events A and B, where $P(A) > 0$, the conditional probability of B given that A has occurred is denoted by $P(B|A)$. This probability is defined as:

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

Example: Suppose you buy a lottery ticket and the first four numbers drawn match the numbers on your ticket. Your chances of winning have increased because we now have additional information.

The formula for conditional probability can be interpreted as the proportion of outcomes in A that also belong to B. This is visually represented in the Venn diagram below:



In this diagram, $P(B|A)$ represents the proportion of the area of A that overlaps with B.

Independence

Two events A and B are independent if:

$$P(A \cap B) = P(A) \cdot P(B)$$

Note: This definition holds for all values of $P(A)$ and $P(B)$, including when either or both equal zero.

In the case where $P(A) > 0$, independence implies:

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

Which is equivalent to $P(B|A) = P(B)$, meaning the occurrence of A does not affect the probability of B.

Lecture 6: Tuesday 9/12/2024

Review: Conditional Probability and Independence

On Tuesday, we discussed conditional probability and independence.

- Conditional Probability: $P(B|A) = \frac{P(A \cap B)}{P(A)}$, where $P(A) > 0$
- We added the requirement $P(B) > 0$, not necessary for the equation to hold true, but useful for computing $P(A|B)$ based on $P(B|A)$

Independence

Events A and B are independent if $P(A \cap B) = P(A)P(B)$

If we assume $P(A) > 0$, then $P(B) = \frac{P(A \cap B)}{P(A)} = P(B|A)$

Example: Rolling Two Fair Dice

Consider an experiment of rolling two fair dice. Alice and Bob are given the following question:

> If Alice rolls the dice and tells Bob that the sum of the dice was 4, does this give Bob new information about whether the number 2 was rolled on the first die?

Analysis

- Sample Space: $\Omega = \{(i, j) | i, j = 1, 2, 3, 4, 5, 6\}$
- $P(\{i, j\}) = \frac{1}{6} \cdot \frac{1}{6} = \frac{1}{36}$ (assuming independence)
- Alternate approach: All outcomes are equally likely

Let's define events:

- A: The sum of the numbers is 4
- B: The first die roll results in a 2

$$
\begin{aligned}
A &= \{(1,3), (2,2), (3,1)\} \\
B &= \{(2,1), (2,2), (2,3), (2,4), (2,5), (2,6)\} \\
A \cap B &= \{(2,2)\}
\end{aligned}
$$

$$
\begin{aligned}
P(B) &= \frac{6}{36} = \frac{1}{6} \\
P(A \cap B) &= \frac{1}{36} \\
P(B|A) &= \frac{P(A \cap B)}{P(A)} = \frac{\frac{1}{36}}{\frac{3}{36}} = \frac{1}{3} \neq \frac{1}{6} = P(B)
\end{aligned}
$$

Therefore, Alice's information does give Bob new information about the probability of rolling a 2 on the first die.

Independence as an Assumption

It's important to note that independence is often an assumption, and it may not always be justified. Consider the following example:

Sally Clark Case and SIDS

In the Sally Clark case, a misuse of probability led to a wrongful conviction:

- Probability of SIDS (Sudden Infant Death Syndrome) was estimated as $\frac{1}{7000}$
- Probability of two SIDS deaths was calculated as $(\frac{1}{7000})^2 = \frac{1}{49\times10^6}$
- This low probability was incorrectly interpreted as indicating guilt

However, this calculation had two major flaws:

1. The assumption of independence was not justified
2. Low probability events do occur

Frequentist Interpretation of Probability

The frequentist interpretation of probability is defined as:

$$P(A) = \lim_{n\to\infty} \frac{\text{number of times A occurred}}{n}$$

where n is the number of independent repetitions of the experiment.

Example: Fair Coin Toss

For a fair coin toss:

$$P(\text{H}) = \frac{1}{2} = \lim_{n\to\infty} \frac{\text{number of heads}}{n}$$

In practice, we approximate this as:

$$P(\text{H}) \approx \frac{\text{number of heads observed}}{\text{total number of tosses}}$$

Note: R Project 2 includes exercises related to these concepts.

Bayes' Theorem - Required Terminology & Proving Intermediate Results

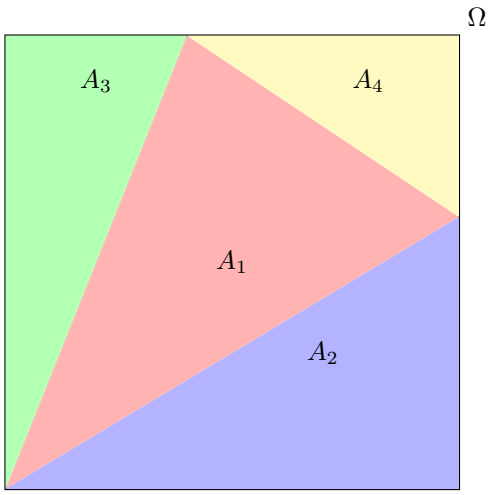
Note: This will be on the list of proofs for exams.

Partition

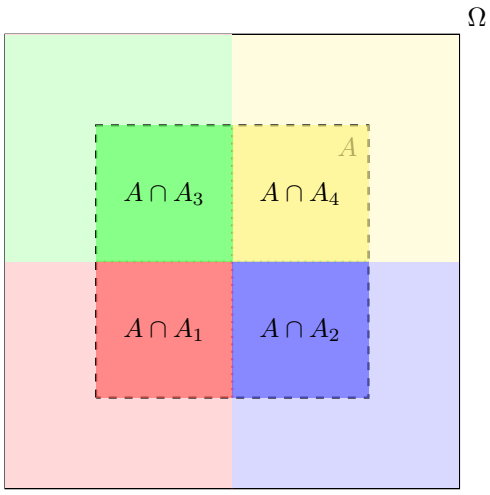
A partition of Ω is a collection of sets $\{A_\alpha : \alpha \in A\}$ such that:

1. $A_\alpha \cap A_\beta = \emptyset$ for all $\alpha, \beta \in A$ and $\alpha \neq \beta$
2. $\bigcup_{\alpha \in A} A_\alpha = \Omega$

Here is a visual example of a partition:



A partition breaks up the sample space into distinct, non-overlapping pieces.
Note:
1) Given $A \subseteq \Omega$, the intersections $\{A \cap A_\alpha : \alpha \in \mathcal{A}\}$ form a partition of A.
Here is a visual example:



We typically consider problems where the partition $\{A_i\}$ is finite. That is, the partition splits up the sample space into finitely many non-overlapping pieces. For example, we might have $|\{A_i\}| = 2$ or $|\{A_i\}| = 3$.

Law of Total Probability
Let $\{A_1, \ldots, A_n\}$ be a partition of the sample space Ω with $P(A_i) > 0$ for all i, and let B be an event with $P(B) > 0$. Then:

$$P(B) = \sum_{i=1}^{n} P(A_i \cap B) = \sum_{i=1}^{n} P(B|A_i)P(A_i)$$

This law allows us to calculate the probability of an event B by considering its interaction with each part of the partition.

Lecture 7: Tuesday 9/17/2024

Conditional Probability

- **Definition of Conditional Probability**
- **Independence**
- **Bayes' Theorem**

Partition

A **partition** of the sample space Ω is a collection $\{B_1, \ldots, B_n\}$ of events such that:

$$B_i \cap B_j = \emptyset \quad \text{for all } i \neq j,$$
$$\bigcup_{i=1}^{n} B_i = \Omega.$$

The diagrams above illustrate how the sample space is split into non-overlapping regions.

Law of Total Probability

Suppose that $\{B_1, \ldots, B_n\}$ is a partition of the sample space Ω with $P(B_i) > 0$ for all i. Then, for any event A with $P(A) > 0$, we have:

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

Consider the event A, where

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

Note that the events $A \cap B_i$ are pairwise disjoint (as shown in the above diagram).

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

Additionally, the conditional probability is given by:

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

Therefore:

$$P(A \cap B_i) = P(A \mid B_i)P(B_i)$$

$$P(A) = \sum_{i=1}^{n} P(A \mid B_i)P(B_i)$$

Bayes' Theorem

Suppose $B_1, B_2, \ldots, B_n$ is a partition of the sample space Ω with $P(B_i) > 0$ for all i. Let A be an event with $P(A) > 0$. Then, for any k where $1 \leq k \leq n$,

$$P(B_k \mid A) = \frac{P(A \mid B_k)P(B_k)}{\sum_{j=1}^{n} P(A \mid B_j)P(B_j)}$$

Proof: By the definition of conditional probability,

$$P(B_k \mid A) = \frac{P(A \cap B_k)}{P(A)} = \frac{P(A \mid B_k)P(B_k)}{P(A)}$$

Using the Law of Total Probability,

$$P(A) = \sum_{j=1}^{n} P(A \mid B_j)P(B_j)$$

Therefore,

$$P(B_k \mid A) = \frac{P(A \mid B_k)P(B_k)}{\sum_{j=1}^{n} P(A \mid B_j)P(B_j)}$$

The Law of Total Probability allows us to change the order of conditioning.

Example: Suppose an experiment consists of tossing a fair coin followed by rolling a fair die if the coin comes up heads, and rolling a fair four-sided die if it comes up tails.

What is the probability that we see heads given that we see a one?

Let H be the event that the coin comes up heads, and F be the event that the die rolls a one. Then,

$$P(F) = P(F \mid H)P(H) + P(F \mid T)P(T)$$

Solution:

We are to find $P(H \mid F)$.

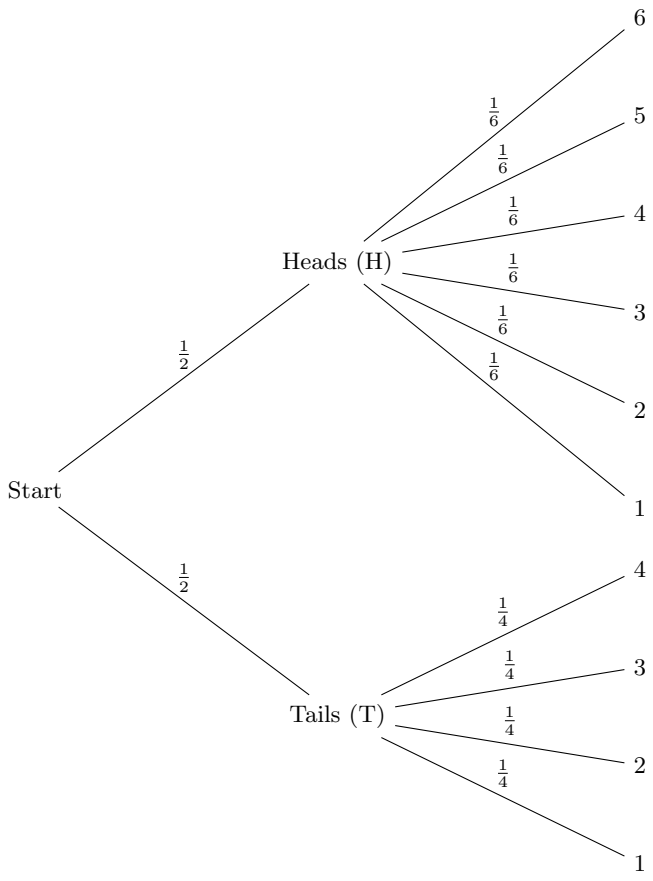


Figure 1: Experiment Diagram: Coin Toss and Die Roll

$$\Omega = \{(H,1),(H,2),(H,3),(H,4),(H,5),(H,6),(T,1),(T,2),(T,3),(T,4)\}$$

Bayes' Theorem: Let H be the event that the coin comes up heads, and T be the event that it comes up tails.

- $P(H) = \frac{1}{2} = P(T)$
- $H \cup T = \Omega$
- $H \cap T = \emptyset$

Thus, H and T form a partition of Ω.

$$P(i \mid H) = \frac{1}{6} \quad \text{for all } i \in \{1, 2, 3, 4, 5, 6\}$$

$$P(i \mid T) = \frac{1}{4} \quad \text{for all } i \in \{1, 2, 3, 4\}$$

Bayes' Theorem:

$$P(\text{Heads} \mid \text{One}) = \frac{P(\text{One} \mid \text{Heads}) \cdot P(\text{Heads})}{P(\text{One} \mid \text{Heads}) \cdot P(\text{Heads}) + P(\text{One} \mid \text{Tails}) \cdot P(\text{Tails})}$$

Using Bayes' Theorem,

$$P(H \mid F) = \frac{P(F \mid H) \cdot P(H)}{P(F \mid H) \cdot P(H) + P(F \mid T) \cdot P(T)}$$

Plugging in the values,

$$P(H \mid F) = \frac{\left(\frac{1}{6}\right) \cdot \left(\frac{1}{2}\right)}{\left(\frac{1}{6} \cdot \frac{1}{2}\right) + \left(\frac{1}{4} \cdot \frac{1}{2}\right)} = \frac{\frac{1}{12}}{\frac{1}{12} + \frac{1}{8}} = \frac{\frac{1}{12}}{\frac{5}{24}} = \frac{2}{5}$$

Therefore, the probability that the coin was heads given that we saw a one is $\boxed{\frac{2}{5}}$.

Problem Statement:

A doctor is called to see a sick child in a particular neighborhood. The doctor has prior information that sick children in the neighborhood are suffering from either measles or the flu, and no child has both. Additionally, the doctor knows that:

- 90% of sick children have measles.
- The probability that a child with measles has a rash is 0.95.
- The probability that a child with the flu has a rash is 0.08.

The child that the doctor sees has a rash. What is the probability that the child has measles?

Definitions:

- Ω: Sick children from the neighborhood.
- M: The child has measles.
- F: The child has the flu.
- R: The child has a rash.

M and F form a partition of the sample space Ω.

Given:

$$P(M) = 0.1, \quad P(F) = 0.9$$
$$P(R \mid M) = 0.95, \quad P(R \mid F) = 0.08$$

Applying Bayes' Theorem:

$$P(M \mid R) = \frac{P(R \mid M) \cdot P(M)}{P(R \mid M) \cdot P(M) + P(R \mid F) \cdot P(F)}$$

$$= \frac{0.95 \times 0.1}{0.95 \times 0.1 + 0.08 \times 0.9}$$

$$= \frac{0.095}{0.095 + 0.072} = \frac{0.095}{0.167} \approx 0.569$$

Conclusion: The probability that the child has measles given that the child has a rash is approximately $\boxed{0.569}$.

Problem Statement: Suppose that you have 3 cards identical in every way except for their color. One card is red on both sides, one card is blue on both sides, and one card is red on one side and blue on the other. You put the cards into a hat and mix them up. You pick one and put it on a desk. You see the color red. What is the probability that the other side is blue?

Let RR denote card 1, BB denote card 2, and RB denote card 3.
Let R be the event that we see red when we put the card on the desk.
RB, BB, and RR form a partition of ω.

$$P(RB) = P(BB) = P(RR) = \frac{1}{3}$$

$$P(R \mid RB) = \frac{1}{2}, \quad P(R \mid BB) = 0, \quad P(R \mid RR) = 1$$

$$P(RB \mid R) = \frac{P(R \mid RB) \cdot P(RB)}{P(R \mid RR) \cdot P(RR) + P(R \mid RB) \cdot P(RB) + P(R \mid BB) \cdot P(BB)}$$

$$P(RB \mid R) = \frac{\left(\frac{1}{2}\right) \cdot \left(\frac{1}{3}\right)}{\left(\frac{1}{2}\right) \cdot \left(\frac{1}{3}\right) + 0 \cdot \left(\frac{1}{3}\right) + 1 \cdot \left(\frac{1}{3}\right)} = \frac{\frac{1}{6}}{\frac{1}{6} + 0 + \frac{1}{3}} = \frac{\frac{1}{6}}{\frac{1}{2}} = \boxed{\frac{1}{3}}$$

Lecture 8: Thursday 9/19/2024

Random Variables

A **random variable** is a function from the sample space to the real numbers.

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

Notes:

- The actual definition has a lot more technicalities to rule out pathologies. We will not worry about these.

Random variables are based on the fact that in many situations we care about a value associated with our pick from the sample space rather than the pick itself.

Example: A study involving the effect of a new medication on blood pressure.

A random variable is a function. Thinking in terms of examples may better highlight why the terminology is what it is.

In order for a random variable to be useful, we should be able to compute

$$P(X \in A) = P(\{\omega : X(\omega) \in A\}) \quad \text{for } A \subseteq \mathbb{R}.$$

$P(X \in A)$ is the probability of running across an outcome that causes the function to take a value inside of A.

Example (cont):

$$P(X \in [119, 125]) = P(\{\omega \mid X(\omega) \in [119, 125]\})$$

How do we convey this information?

Distribution function for a random variable X:

$$D_X : \mathcal{P}(\mathbb{R}) \to \mathbb{R}$$

$$D_X(A) = P(X \in A)$$

Since $\mathcal{P}(\mathbb{R})$ contains very complicated subsets of real numbers, it is unlikely that we are able to express D_X in a form that is suitable for computation.

For a random variable X, the **cumulative distribution function** (CDF) is defined as:

$$F_X : \mathbb{R} \to \mathbb{R}$$

$$F_X(a) = P(X \leq a) = P(X \in (-\infty, a])$$

Suppose that an experiment consists of tossing a fair coin. Let X track the outcome of the coin toss in the natural way. ($\Omega = \{H, T\}$) Define $X(T) = 0$ and $X(H) = 1$. Compute the CDF and the distribution function of X.

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

Suppose $a = -1.32975$:

$$F_X(a) = P(X \leq a) = 0$$

since no outcome of X can have a value less than a.

Note, this is true for any $a < 0$:

$$F_X(a) = 0 \quad \text{for all } a < 0$$

Suppose $a = 0$:

$$F_X(0) = P(X \leq 0) = \frac{1}{2}$$

Suppose $a = 0.5$:

$$F_X(a) = P(X \leq 0.5) = P(X = 0) = \frac{1}{2}$$

Suppose $0 \leq a < 1$:

$$F_X(a) = \frac{1}{2}$$

Suppose $a = 1$:

$$F_X(a) = P(X \leq a) = P(X \leq 1) = P(X = 0 \text{ or } X = 1) = 1$$

Suppose $a = 1.5$ so $a > 1$:

$$F_X(a) = P(X \leq a) = P(X \leq 1.5) = 1$$

since any outcome of X must be less than or equal to 1.

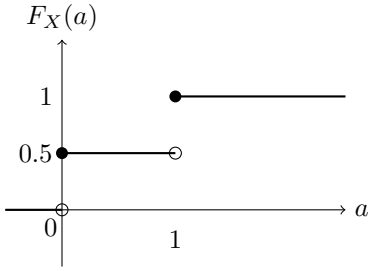


Figure 2: CDF of X

$$F_X(a) = \begin{cases} 0 & \text{if } a < 0 \\ 0.5 & \text{if } 0 \leq a < 1 \\ 1 & \text{if } a \geq 1 \end{cases}$$

1. $\lim_{a \to -\infty} F_X(a) = 0$
2. $\lim_{a \to \infty} F_X(a) = 1$
3. F_X is monotonically increasing, i.e., if $a < b$, $F_X(a) \leq F_X(b)$
4. F_X is right continuous, i.e., $\lim_{a \to a^+} F_X(a) = F_X(a)$ (there are no points of discontinuity when approaching from the right side)

Definition: X and Y are said to have the same distribution if $D_X(A) = D_Y(A)$ for all $A \subseteq \mathbb{R}$. (i.e., X and Y cannot be distinguished by looking at the probabilities alone.)

Example:
Generate the numbers 0 and 1 so that both are equally likely.

- Toss a fair coin.
- Pick a number uniformly at random from the interval $[0, 1]$.

$Y : [0, 1] \to \mathbb{R}$

$$Y(\omega) = \begin{cases} 0 & \text{if } 0 < \omega < \frac{1}{2} \\ 1 & \text{if } \frac{1}{2} < \omega < 1 \end{cases}$$

We cannot tell X and Y apart by looking at probabilities alone.
Fact: $D_X = D_Y$ if and only if $F_X = F_Y$.

$$D_X(A) = \begin{cases} 0 & \text{if } 0, 1 \notin A \\ 0.5 & \text{if } \{0\}, \{1\} \subseteq A \text{ but not } \{0, 1\} \subseteq A \\ 1 & \text{if } \{0, 1\} \subseteq A \end{cases}$$

$D_X(A) = P(X \in A)$
Example: Suppose you roll a fair die.
Let $Y : \Omega \to \mathbb{R}$ be given by $Y(\omega) = \omega$, where $\Omega = \{1, 2, 3, 4, 5, 6\}$.
What is F_Y?
Case 1: $a < 1$

$$F_Y(a) = 0$$

Case 2: $1 \leq a < 2$

$$F_Y(a) = P(Y \leq a) = P(\{1\}) = \frac{1}{6}$$

Case 3: $2 \leq a < 3$

$$F_Y(a) = P(Y \leq a) = P(\{1, 2\}) = \frac{2}{6} = \frac{1}{3}$$

Case 4: $3 \leq a < 4$

$$F_Y(a) = P(Y \leq a) = P(\{1, 2, 3\}) = \frac{3}{6} = \frac{1}{2}$$

Case 5: $4 \leq a < 5$

$$F_Y(a) = P(Y \leq a) = P(\{1, 2, 3, 4\}) = \frac{4}{6} = \frac{2}{3}$$

Case 6: $5 \leq a < 6$

$$F_Y(a) = P(Y \leq a) = P(\{1, 2, 3, 4, 5\}) = \frac{5}{6}$$

Case 7: $a \geq 6$

$$F_Y(a) = P(Y \leq a) = P(\{1, 2, 3, 4, 5, 6\}) = 1$$

Lecture 9: Tuesday 9/24/2024

Probability Distributions and Random Variables

Discrete Random Variables

The **Cumulative Distribution Function** (CDF) for a random variable X is defined as:

$$F_X(a) = P(X \leq a) = P(X \in (-\infty, a])$$

Properties of CDF:

1. F_X is non-decreasing
2. $\lim_{x \to -\infty} F_X(x) = 0$ and $\lim_{x \to \infty} F_X(x) = 1$
3. F_X is right-continuous

Example: Fair Coin Toss

Let X be a random variable representing the outcome of a fair coin toss, where 0 represents tails and 1 represents heads.

The CDF for this scenario is:

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

Example: Fair Die Roll

Let X be a random variable representing the outcome of rolling a fair six-sided die.

Sample space: $\Omega = \{1, 2, 3, 4, 5, 6\}$

The random variable X is defined as: $X(\omega) = \omega$ for $\omega \in \Omega$

The CDF for X is:

$$F_X(a) = \begin{cases} 0 & \text{if } a < 1 \\ \frac{1}{6} & \text{if } 1 \leq a < 2 \\ \frac{1}{3} & \text{if } 2 \leq a < 3 \\ \frac{1}{2} & \text{if } 3 \leq a < 4 \\ \frac{2}{3} & \text{if } 4 \leq a < 5 \\ \frac{5}{6} & \text{if } 5 \leq a < 6 \\ 1 & \text{if } a \geq 6 \end{cases}$$

Note: If X and Y are random variables, then their distributions D_X and D_Y are equal if and only if their CDFs are equal, i.e., $F_X = F_Y$.

Example: Uniform Distribution on $[0, 1]$

Let X be a random variable representing a number picked uniformly at random from the interval $[0, 1]$.

We define $X : [0, 1] \to \mathbb{R}$ as $X(\omega) = \omega$ for $\omega \in [0, 1]$.

To find the cumulative distribution function (CDF) of X, we calculate $P(X \leq a)$ for any real number a:

$$P(X \leq a) = P(X \in (-\infty, a]) = P(\{\omega \mid X(\omega) \in (-\infty, a]\})$$

We consider three cases:

1. If $a < 0$: $P(X \leq a) = 0$, since $X \in [0, 1]$.
2. If $0 \leq a \leq 1$: $P(X \leq a) = P(\{\omega \mid X(\omega) \in [0, a]\}) = a$, due to the uniform distribution on $[0, 1]$.
3. If $a > 1$: $P(X \leq a) = 1$, since $X \leq 1$ always.

Therefore, the CDF of X is:

$$F_X(a) = \begin{cases} 0, & a < 0 \\ a, & 0 \leq a \leq 1 \\ 1, & a \geq 1 \end{cases}$$

This piecewise function represents the uniform distribution on $[0, 1]$.

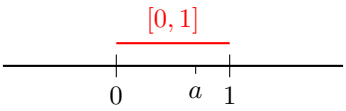


Figure 3: Number line showing a between 0 and 1 for Uniform Distribution

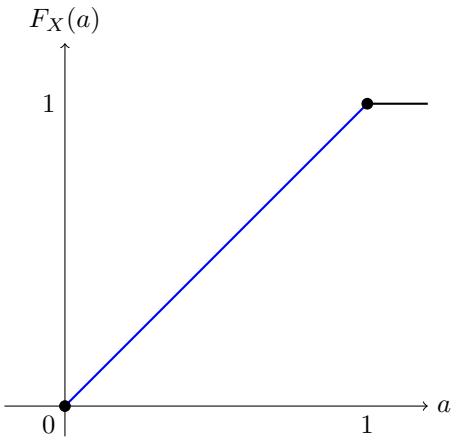


Figure 4: CDF of Uniform Distribution on $[0, 1]$

Discrete Random Variables

A discrete random variable X is one that takes countably many values. This means it is either finite or can be put into a sequence (countably infinite).

Note: Countably many values refers to outputs, not inputs.

Examples:

- Die roll: $X = \{1, 2, 3, 4, 5, 6\}$
- Coin toss: $X = \{H, T\}$

A discrete random variable comes with a **probability mass function** (PMF or pmf):

$$P_X(x) = P(X = x)$$

Example: Fair coin toss

$$X(H) = 1, \quad X(T) = 0$$

$$P_X(x) = \begin{cases} 0.5 & \text{if } x = 0 \text{ or } x = 1 \\ 0 & \text{otherwise} \end{cases}$$

Example: Fair die roll

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

$$P_X(x) = \begin{cases} \frac{1}{6} & \text{if } x \in \{1, 2, 3, 4, 5, 6\} \\ 0 & \text{otherwise} \end{cases}$$

Probabilities for a discrete random variable can be computed as sums of the pmf:

$$P(X \in A) = \sum_{x \in A} P_X(x)$$

Note: For all x, $P_X(x) \geq 0$

Example: Die roll

$$\begin{aligned} P(X \in (2,6)) &= P(\{\omega \mid X(\omega) \in (2,6)\}) \\ &= \sum_{x \in (2,6)} P_X(x) \quad \text{where } P_X(x) > 0 \\ &= P_X(3) + P_X(4) + P_X(5) \\ &= \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{3}{6} = \frac{1}{2} \end{aligned}$$

Important properties of PMF:

- $P_X(x) \geq 0$ for all x
- $\sum_{x \in (-\infty, \infty)} P_X(x) = 1$, where the sum is taken over all x such that $P_X(x) > 0$

Properties of Discrete Random Variables:
Suppose X and Y are discrete random variables. Then:

- $D_X = D_Y$ if and only if $F_X = F_Y$ if and only if $P_X(x) = P_Y(x)$ for all $x \in \mathbb{R}$
- If the CDF of X is piecewise constant, then X is a discrete random variable
- For a discrete random variable X, $P_X(x) = F_X(x) - F_X(x^-)$

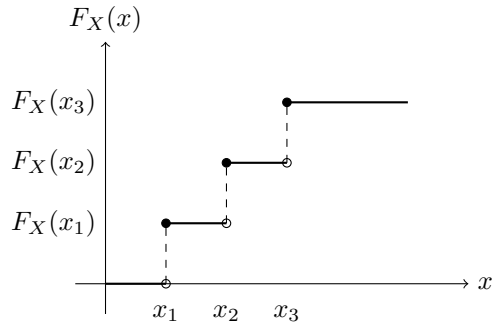


Figure 5: Illustration of CDF for a discrete random variable

Note: X is a point of discontinuity where x^- is the previous point of discontinuity.

$$F_X(x^-) = \begin{cases} \text{previous point of discontinuity} & \text{if it exists} \\ 0 & \text{otherwise} \end{cases}$$

Example: For a fair die roll,

$$P_X(5) = F_X(5) - F_X(4) = \frac{5}{6} - \frac{4}{6} = \frac{1}{6}$$

Lecture 10: Thursday 9/26/2024

Discrete Random Variables

Example: Suppose that you toss a fair coin until you see heads for the first time.

Let X be a random variable that counts the number of tosses.

Explain why X is a discrete random variable and compute its **probability mass function** (PMF).

X can only take countably many values.

X takes values from the set $\{*, 1, 2, 3, \ldots\}$.

This set is countable, therefore X is discrete.

Note: The asterisk (*) is used to denote all tails.

$$P(X = n) = \frac{1}{2^n}$$

$$P_X(n) = \begin{cases} \frac{1}{2^n} & \text{if } n \in \{1, 2, 3, \ldots\} \\ 0 & \text{otherwise} \end{cases}$$

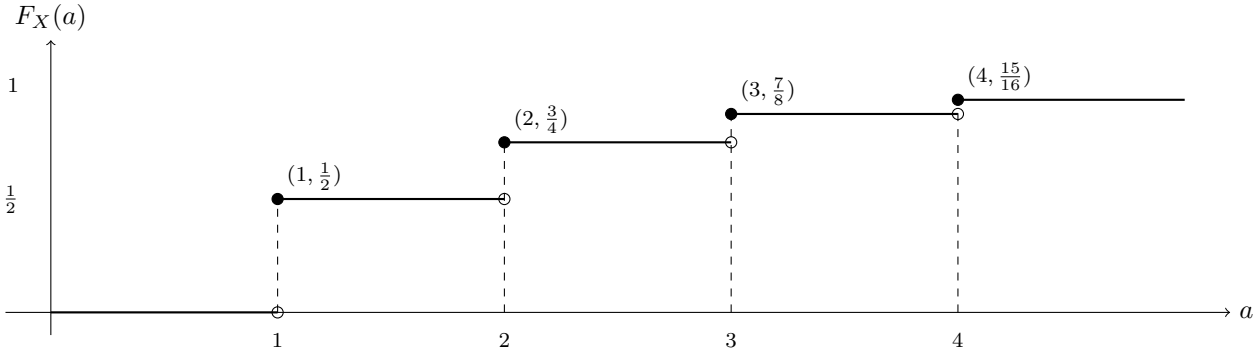


Figure 6: CDF of X

CDF of X

$$F_X(a) = \begin{cases} 0 & \text{if } a < 1 \\ \frac{1}{2} & \text{if } 1 \leq a < 2 \\ \frac{3}{4} & \text{if } 2 \leq a < 3 \\ \frac{7}{8} & \text{if } 3 \leq a < 4 \\ \frac{15}{16} & \text{if } 4 \leq a < 5 \\ \vdots & \end{cases}$$

Recall, $\lim_{a \to \infty} F_X(a) = 1$.

Continuous Random Variables

A random variable X is said to be (absolutely) continuous if there exists a function f_X such that

$$f_X(x) \geq 0 \quad \text{and} \quad P(a \leq X \leq b) = \int_a^b f_X(x)\,dx$$

$f_X(x)$ is called the **probability density function** (PDF) of X.

Note: The definition of continuous random variables is a little bit broader. We say a random variable X is continuous if its cdf is continuous. But we will not encounter them in this course, so we take absolutely continuous random variables to mean continuous random variables.

Example

Suppose that X is a continuous random variable with a pdf

$$f_X(x) = \begin{cases} 1 & \text{if } x \in [0,1] \\ 0 & \text{otherwise} \end{cases}$$

1. Verify that $f_X(x)$ is a pdf.
2. Compute $P\left(\frac{1}{2} < X < \frac{3}{4}\right)$.
3. Compute $P\left(X > \frac{3}{4} \mid X > \frac{1}{2}\right)$.
4. Compute the cdf of X.

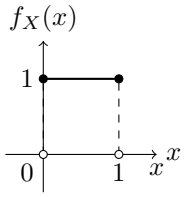
Solution

1) A pdf has to satisfy the following properties:

- $f_X(x) \geq 0$
- $\int_{-\infty}^{\infty} f_X(x)\,dx = 1$

For the given pdf:

$$f_X(x) = \begin{cases} 1 & \text{if } x \in [0,1] \\ 0 & \text{otherwise} \end{cases}$$



$$\int_{-\infty}^{0} f_X(x)\,dx = 0$$

$$\int_{0}^{1} f_X(x)\,dx = 1$$

$$\int_{1}^{\infty} f_X(x)\,dx = 0$$

Therefore:

$$\int_{-\infty}^{\infty} f_X(x)\,dx = 1$$

2) Compute $P\left(\frac{1}{2} < X < \frac{3}{4}\right)$:

Things to notice: Suppose Y is a continuous random variable with pdf $f_Y(y)$. Then for any $a \in \mathbb{R}$, $P(Y = a) = 0$.

$$P(Y = a) = P(a \leq Y \leq a) = \int_{a}^{a} f_Y(y)\,dy = 0$$

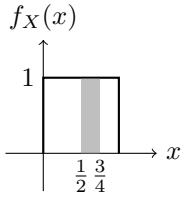
Note:

- "Possible but not probable": $P(X = \frac{1}{2}) = 0$
- "Impossible": $P(X = -3) = 0$

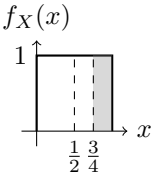
For our problem:

$$P\left(\frac{1}{2} < X < \frac{3}{4}\right) = \int_{\frac{1}{2}}^{\frac{3}{4}} f_X(x)\,dx$$

$$P\left(\frac{1}{2} < X < \frac{3}{4}\right) = \int_{\frac{1}{2}}^{\frac{3}{4}} 1\,dx = x\Big|_{\frac{1}{2}}^{\frac{3}{4}} = \frac{3}{4} - \frac{1}{2} = \frac{1}{4}$$



3) Compute $P\left(X > \frac{3}{4} \mid X > \frac{1}{2}\right)$:



$$P\left(X > \frac{3}{4} \mid X > \frac{1}{2}\right) = \frac{P\left(X > \frac{3}{4} \cap X > \frac{1}{2}\right)}{P\left(X > \frac{1}{2}\right)}$$

$$P\left(X > \frac{3}{4}\right) = \int_{\frac{3}{4}}^{1} 1\,dx = 1 - \frac{3}{4} = \frac{1}{4}$$

$$P\left(X > \frac{1}{2}\right) = \int_{\frac{1}{2}}^{1} 1\,dx = 1 - \frac{1}{2} = \frac{1}{2}$$

$$P\left(X > \frac{3}{4} \mid X > \frac{1}{2}\right) = \frac{\frac{1}{4}}{\frac{1}{2}} = \frac{1}{2}$$

4) Compute the cdf of X:

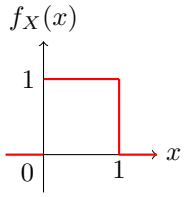
$$F_X(a) = P(x \le a) = \int_{-\infty}^{a} f_X(x)\,dx$$

Case 1: $a < 0$

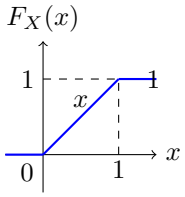
$$F_X(a) = \int_{-\infty}^{a} 0\,dx = 0$$

$$F_X(a) = \begin{cases} 0 & \text{if } a < 0 \\ a & \text{if } 0 \le a < 1 \\ 1 & \text{if } a \ge 1 \end{cases}$$

PDF of X



CDF of X



Facts: Let X and Y be two continuous random variables with pdfs $f_X(x)$ and $f_Y(y)$. Then:

$D_X = D_Y$ if and only if $F_X = F_Y$ if and only if $f_X(x) = f_Y(y)$ almost everywhere.

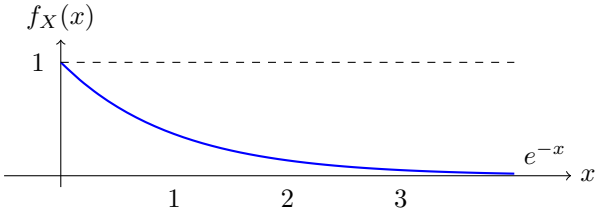
Example: Suppose X is a continuous random variable with pdf:

$$f_X(x) = \begin{cases} e^{-x} & \text{if } x \ge 0 \\ 0 & \text{otherwise} \end{cases}$$

Compute $P(X > \frac{1}{2})$:

$$\begin{aligned}
P(X > \frac{1}{2}) &= \int_{\frac{1}{2}}^{\infty} e^{-x}\,dx \\
&= \lim_{b\to\infty} \int_{\frac{1}{2}}^{b} e^{-x}\,dx \\
&= \lim_{b\to\infty} \left[-e^{-x}\right]_{\frac{1}{2}}^{b} \\
&= \lim_{b\to\infty} \left[(-e^{-b}) - (-e^{-\frac{1}{2}})\right] \\
&= \lim_{b\to\infty} \left[-e^{-b} + e^{-\frac{1}{2}}\right] \\
&= 0 + e^{-\frac{1}{2}} \\
&= e^{-\frac{1}{2}} \\
&= \frac{1}{\sqrt{e}}
\end{aligned}$$

PDF of X: $f_X(x) = e^{-x}$ for $x \geq 0$

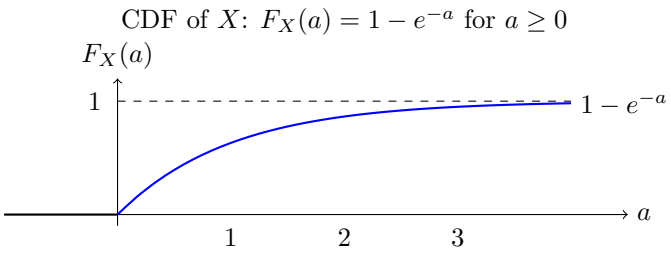


Compute the cdf of X:

$$\begin{aligned}
F_X(a) &= P(X \leq a) \\
&= \int_{-\infty}^{a} f_X(t)\,dt \\
&= \begin{cases} 0 & \text{if } a < 0 \\ \int_0^a e^{-t}\,dt & \text{if } a \geq 0 \end{cases} \\
&= \begin{cases} 0 & \text{if } a < 0 \\ [-e^{-t}]_0^a & \text{if } a \geq 0 \end{cases} \\
&= \begin{cases} 0 & \text{if } a < 0 \\ -e^{-a} - (-e^0) = 1 - e^{-a} & \text{if } a \geq 0 \end{cases}
\end{aligned}$$

Therefore, the cdf of X is:

$$F_X(a) = \begin{cases} 0 & \text{if } a < 0 \\ 1 - e^{-a} & \text{if } a \geq 0 \end{cases}$$



Properties of CDF:

1. Limits:

$$\lim_{a \to -\infty} F_X(a) = 0$$
$$\lim_{a \to \infty} F_X(a) = 1$$

2. Right-continuity:

$$\lim_{a \to x^+} F_X(a) = F_X(x)$$

3. Monotonicity:

$$\text{If } a < b, \text{ then } F_X(a) \leq F_X(b)$$

Probability calculations using CDF:

$$P(X \leq a) = F_X(a)$$
$$P(X > a) = 1 - F_X(a)$$
$$P(a < X \leq b) = F_X(b) - F_X(a)$$

Example: For our exponential distribution with $F_X(a) = 1 - e^{-a}$ for $a \geq 0$:

$$\begin{aligned} P(X > \frac{1}{2}) &= 1 - P(X \leq \frac{1}{2}) = 1 - F_X(\frac{1}{2}) \\ &= 1 - (1 - e^{-\frac{1}{2}}) = e^{-\frac{1}{2}} \end{aligned}$$

$$\begin{aligned} P(\frac{1}{2} < X < 10) &= F_X(10) - F_X(\frac{1}{2}) \\ &= (1 - e^{-10}) - (1 - e^{-\frac{1}{2}}) \\ &= e^{-\frac{1}{2}} - e^{-10} \end{aligned}$$

Fact: If F_X is differentiable, with possibly finitely many exceptions, then

$$F_X'(x) = f_X(x) \quad \text{(except where } F_X \text{ is not differentiable)}$$

Expected Value

The expected value of a random variable X, denoted as E(X), is defined as:

$$E(X) = \begin{cases} \sum_i x_i P(X = x_i) & \text{if X is discrete} \\ \int_{-\infty}^{\infty} x f_X(x)\, dx & \text{if X is continuous} \end{cases}$$

Properties of Expected Value:

- E(X) is a "weighted sum" or "weighted average" of possible values of X.
- E(X) provides a single value summary of the random variable X.
- E(X) represents the long-run average of the random variable over many trials.

Important Inequalities

Markov's Inequality: For a non-negative random variable X and any positive number a,

$$P(X \geq a) \leq \frac{E(X)}{a}$$

Chebyshev's Inequality: For any random variable X with finite expected value μ and finite non-zero variance σ^2,

$$P(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}$$

Law of Large Numbers

Strong Law of Large Numbers (SLLN): If we generate a large number of independent and identically distributed random variables $X_1, \ldots, X_n$, and compute their average, then this average will converge to the expected value E(X) as n approaches infinity:

$$\frac{1}{n} \sum_{i=1}^{n} X_i \approx E(X)$$

This law provides a theoretical foundation for the empirical average to approximate the expected value for a large number of trials.

Lecture 11: Tuesday 10/1/2024

Expected Value

The expected value of a random variable X, denoted as E(X), is defined as:

$$E(X) = \begin{cases} \sum_i x_i P(X = x_i) & \text{if X is discrete} \\ \int_{-\infty}^{\infty} x f_X(x)\, dx & \text{if X is continuous} \end{cases}$$

Properties of Expected Value:

- E(X) is a "weighted sum"
- E(X) provides a single value summary of the random variable X
- E(X) represents the long-run average of the random variable over many trials

Important Inequalities

- **Markov's Inequality**
- **Chebyshev's Inequality**

Strong Law of Large Numbers (SLLN)

Suppose that you repeat the experiment associated with X in an independent manner and write out the measurements obtained as $X_1, X_2, \ldots, X_n$.

For large n, the average approximates the expected value:

$$\frac{1}{n} \sum_{i=1}^{n} X_i \approx E(X)$$

Example: Rolling a Fair Die

Let $X(\omega) = \omega$ report the outcome of the die roll.

Probability Mass Function:

$$P_X(x) = \begin{cases} \frac{1}{6} & \text{if } x = 1, 2, 3, 4, 5, 6 \\ 0 & \text{otherwise} \end{cases}$$

Expected Value Calculation:

$$\begin{aligned} E(X) &= \sum_{i=1}^{6} iP_X(i) \\ &= 1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + 3 \cdot \frac{1}{6} + 4 \cdot \frac{1}{6} + 5 \cdot \frac{1}{6} + 6 \cdot \frac{1}{6} \\ &= \frac{1}{6}(1+2+3+4+5+6) \\ &= \frac{1}{6}(\frac{6 \cdot 7}{2}) \\ &= \frac{7}{2} = 3.5 \end{aligned}$$

Experiment:

- Roll the die n = 10000 times and record the outcomes.
- Compute the average of the outcomes.
- Notice that the average $\frac{1}{n}\sum_{i=1}^{n} X_i$ is close to 3.5.

Example: Suppose X is a continuous random variable with cumulative distribution function:

$$F_X(x) = \begin{cases} e^{-x} & \text{for } x \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

Find $E(X)$.

$$\begin{aligned} E(X) &= \int_{-\infty}^{\infty} x f_X(x)\, dx \\ &= \int_{0}^{\infty} x e^{-x}\, dx \quad (\text{since } f_X(x) = 0 \text{ for } x < 0) \\ &= \lim_{b \to \infty} \int_{0}^{b} x e^{-x}\, dx \end{aligned}$$

To solve this integral, we use integration by parts:

- Let $u = x$ and $dv = e^{-x}\, dx$
- Then $du = dx$ and $v = -e^{-x}$

Applying integration by parts:

$$\begin{aligned}
E(X) &= \lim_{b\to\infty}\left[-xe^{-x}\Big|_0^b + \int_0^b e^{-x}\,dx\right] \\
&= \lim_{b\to\infty}\left[(-be^{-b} - (-0\cdot e^0)) + \left(-e^{-x}\Big|_0^b\right)\right] \\
&= \lim_{b\to\infty}\left[-be^{-b} + \left(-e^{-b} - (-e^0)\right)\right] \\
&= \lim_{b\to\infty}\left[-be^{-b} - e^{-b} + 1\right] \\
&= 1 - \lim_{b\to\infty}(be^{-b} + e^{-b}) \\
&= 1 - 0 - 0 \quad (\text{since } \lim_{b\to\infty} be^{-b} = \lim_{b\to\infty} e^{-b} = 0) \\
&= 1
\end{aligned}$$

Therefore, $E(X) = 1$.

Law of the Unconscious Statistician (LOTUS)

Let $g : \mathbb{R} \to \mathbb{R}$ be a real-valued function of the random variable X. Then:

$$E[g(X)] = \begin{cases} \sum_x g(x)P_X(x) & \text{if } X \text{ is discrete} \\ \int_{-\infty}^{\infty} g(x)f_X(x)\,dx & \text{if } X \text{ is continuous} \end{cases}$$

Note: $E[g(X)] = E[Z]$ where $Z = g(X)$. However, computing $E[Z]$ directly as

$$E[Z] = \int_{-\infty}^{\infty} z f_Z(z)\,dz$$

can be challenging and is often unnecessary due to LOTUS.

Example: Suppose X is a discrete random variable with PMF:

$$P_X(x) = \begin{cases} \frac{1}{3} & \text{if } x = -1, 0, 1 \\ 0 & \text{otherwise} \end{cases}$$

What is the PMF of $Y = X^2$?

Solution:

$$P_Y(y) = \begin{cases} \frac{1}{3} & \text{if } y = 0 \\ \frac{2}{3} & \text{if } y = 1 \\ 0 & \text{otherwise} \end{cases}$$

Using the definition:

$$\begin{aligned} E(X^2) &= \sum_{y=0}^{1} y P_Y(y) \\ &= 0 \cdot \frac{1}{3} + 1 \cdot \frac{2}{3} = \frac{2}{3} \end{aligned}$$

Using LOTUS:

$$\begin{aligned} E(X^2) &= \sum_{x=-1}^{1} x^2 P_X(x) \\ &= (-1)^2 \cdot \frac{1}{3} + 0^2 \cdot \frac{1}{3} + 1^2 \cdot \frac{1}{3} = \frac{2}{3} \end{aligned}$$

Fact: For constants a and b,

$$E(aX + b) = aE(X) + b$$

Suppose X is a continuous random variable with PDF $f_X(x)$.

Proof using LOTUS:

$$\begin{aligned} E(aX + b) &= \int_{-\infty}^{\infty} (ax + b) f_X(x)\, dx \\ &= a \int_{-\infty}^{\infty} x f_X(x)\, dx + b \int_{-\infty}^{\infty} f_X(x)\, dx \\ &= aE(X) + b \cdot 1 \\ &= aE(X) + b \end{aligned}$$

Variance

Variance measures the spread of observations around the mean of a random variable.

For a random variable X with mean $\mu = E(X)$, the variance is defined as:

$$V(X) = E((X - \mu)^2)$$

For any random variable X, the variance can be computed as:

$$V(X) = E(X^2) - (E(X))^2$$

Proof: Let X be a continuous random variable with probability density function $f_X(x)$ and mean

$\mu = E(X)$. Using the Law of the Unconscious Statistician (LOTUS):

$$\begin{aligned}
V(X) &= E((X-\mu)^2) \\
&= \int_{-\infty}^{\infty} (x-\mu)^2 f_X(x)\,dx \\
&= \int_{-\infty}^{\infty} (x^2 - 2\mu x + \mu^2) f_X(x)\,dx \\
&= \int_{-\infty}^{\infty} x^2 f_X(x)\,dx - 2\mu \int_{-\infty}^{\infty} x f_X(x)\,dx + \mu^2 \int_{-\infty}^{\infty} f_X(x)\,dx \\
&= E(X^2) - 2\mu E(X) + \mu^2 \\
&= E(X^2) - 2\mu^2 + \mu^2 \\
&= E(X^2) - \mu^2 \\
&= E(X^2) - (E(X))^2
\end{aligned}$$

Note: The proof for discrete random variables follows a similar structure, replacing integrals with sums.

Visualization of Low and High Variance

To illustrate the concept of low and high variance, consider two normal distributions:

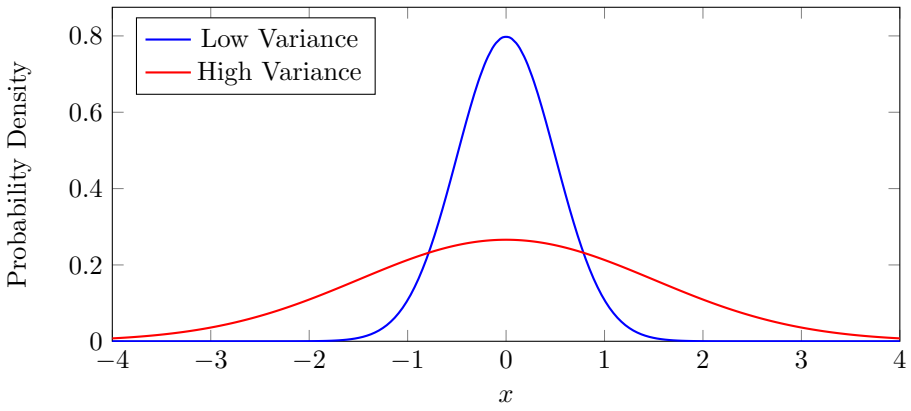


Figure 7: Comparison of normal distributions with low and high variance

- The blue curve represents a distribution with low variance, more concentrated around its mean.
- The red curve represents a distribution with high variance, more spread out from its mean.
- Both distributions have the same mean (center) but differ in their spread.

Markov's Inequality

For a non-negative random variable X and any positive constant a, if $E(X)$ exists and is finite:

$$P(X \geq a) \leq \frac{E(X)}{a}$$

Example: Suppose X is a non-negative random variable with $E(X) = 5$. Find upper bounds for:

- $P(X \geq 5)$
- $P(X \geq 500)$

Solution:

$$P(X \geq 5) \leq \frac{E(X)}{5} = \frac{5}{5} = 1 \quad \text{(not a useful bound)}$$
$$P(X \geq 500) \leq \frac{E(X)}{500} = \frac{5}{500} = 0.01$$

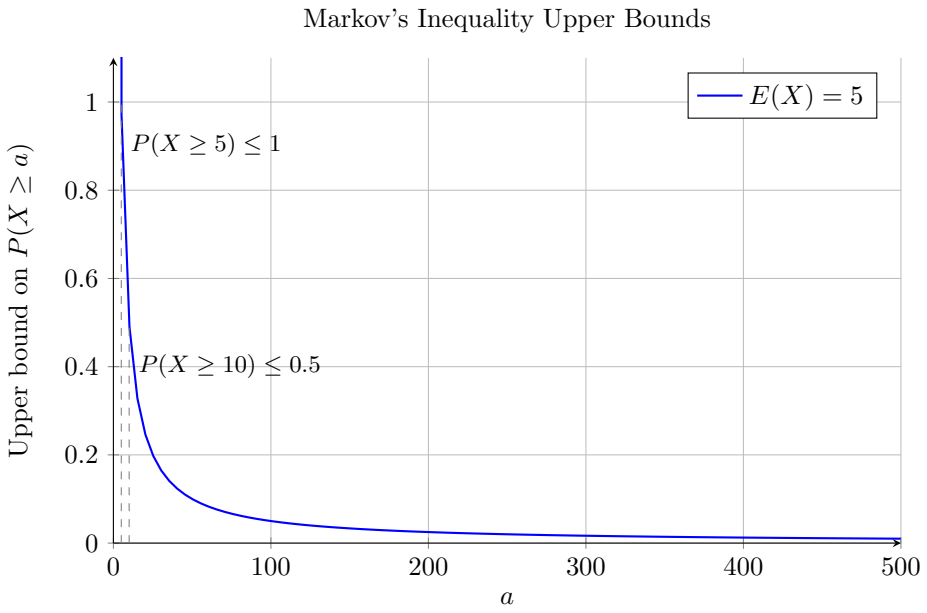


Figure 8: Visualization of Markov's Inequality upper bound for $E(X) = 5$, extended to $a = 500$

The graph above illustrates Markov's Inequality for different $E(X)$ values. The curves represent the upper bound $\frac{E(X)}{a}$ as a varies. As a increases, the upper bound decreases, providing tighter bounds for larger values of a.

Chebyshev's Inequality

For a random variable X with finite mean μ and variance σ^2, and for any positive constant k:

$$P(|X-\mu| \geq k) \leq \frac{\sigma^2}{k^2}$$

Example: Let X be a random variable with $E(X) = 4$ and $V(X) = 16$. Find an upper bound for $P(|X-4| \geq 1600)$.
Solution:

$$\begin{aligned} P(|X-4| \geq 1600) &= P(|X-\mu| \geq 1600) \\ &\leq \frac{V(X)}{1600^2} \\ &= \frac{16}{1600^2} \\ &= \frac{1}{10000} \approx 0.0001 \end{aligned}$$

Note: $\sigma = \sqrt{V(X)} = 4$, and $k = \frac{1600}{\sigma} = \frac{1600}{4} = 400$

Strong Law of Large Numbers (SLLN)

Example: Approximate $\int_0^1 \sqrt{1-x^2}dx$
Solution: The exact value of this integral is $\frac{\pi}{4}$.

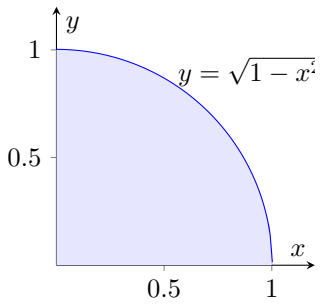


Figure 9: Graph of $y = \sqrt{1-x^2}$ (quarter circle)

The graph above illustrates the region under the curve $y = \sqrt{1-x^2}$ from $x = 0$ to $x = 1$, which forms a quarter circle. The area of this region is equal to $\frac{\pi}{4}$, which is the exact value of the integral $\int_0^1 \sqrt{1-x^2}dx$.

Let Y be a continuous random variable with PDF:

$$f_Y(y) = \begin{cases} 1 & 0 \leq y \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

Using the Law of the Unconscious Statistician (LOTUS), we can calculate:

$$\begin{aligned} E(\sqrt{1-Y^2}) &= \int_{-\infty}^{\infty} \sqrt{1-y^2} f_Y(y) dy \\ &= \int_0^1 \sqrt{1-y^2} dy \end{aligned}$$

To evaluate this integral, we can use the substitution $u = 1 - y^2$, $du = -2ydy$:

$$\begin{aligned} E(\sqrt{1-Y^2}) &= \int_0^1 \sqrt{1-y^2} dy \\ &= -\frac{1}{2} \int_1^0 \frac{\sqrt{u}}{\sqrt{1-u}} du \\ &= \frac{1}{2} \int_0^1 \frac{\sqrt{u}}{\sqrt{1-u}} du \\ &= \frac{\pi}{4} \end{aligned}$$

The last step involves recognizing this as a standard integral that evaluates to $\frac{\pi}{2}$.

By the Strong Law of Large Numbers, if we generate independent samples $Y_1, Y_2, \ldots, Y_n$ from the distribution of Y, then:

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \sqrt{1-Y_i^2} = E(\sqrt{1-Y^2}) = \frac{\pi}{4} \quad \text{(almost surely)}$$

This provides a method to approximate $\frac{\pi}{4}$ using random sampling.

Lecture 12: Thursday 10/3/2024

Special Distributions (Discrete)

- Bernoulli Distribution
- Geometric Distribution
- Binomial Distribution
- Hypergeometric Distribution

Bernoulli Distribution

- A Bernoulli trial is an experiment with only two possible outcomes: success (S) or failure (F).
- A Bernoulli random variable X is a discrete random variable with probability mass function:

$$P_X(x) = \begin{cases} p & \text{if } x = 1 \text{ (success)} \\ q = 1 - p & \text{if } x = 0 \text{ (failure)} \end{cases}$$

- The probability of success is p.
- We denote this as $X \sim \text{Ber}(p)$.
- p is a parameter that keeps track of the probability of success, which may differ between experiments modeled by Bernoulli trials.

Properties:

$$E(X) = \sum xP_X(x) = 0 \cdot (1-p) + 1 \cdot p = p$$
$$V(X) = E(X^2) - (E(X))^2 = \sum x^2 P_X(x) - p^2 = p - p^2 = p(1-p)$$

Geometric Distribution

Consider an experiment where you repeat independent Bernoulli trials until you observe the first success, at which point you stop. Let X be the number of trials needed to see the first success.

- X is a discrete random variable.
- Let p be the probability of success for an individual Bernoulli trial, where $0 < p < 1$.
- We want to find the probability mass function (PMF) of X.

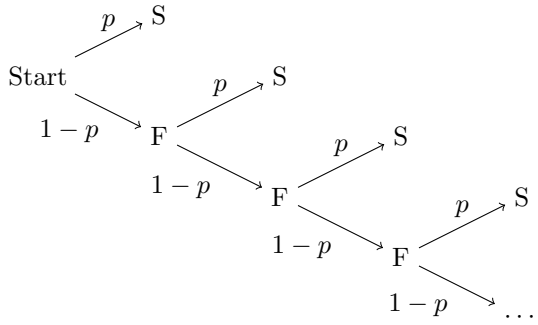
Probability Mass Function:
For $k = 1, 2, 3, \ldots$, the PMF of X is given by:

$$P(X = k) = (1-p)^{k-1}p$$

This can be interpreted as:

- $(1-p)^{k-1}$: probability of $k-1$ consecutive failures
- p: probability of success on the k-th trial

Visual Representation:



Where F represents failure and S represents success.

Note: The geometric distribution models the number of trials until the first success, including the successful trial.

The probability mass function $P_X(x)$ for the geometric distribution is:

$$P_X(x) = \begin{cases} (1-p)^{x-1}p & \text{for } x = 1, 2, 3, \ldots \\ 0 & \text{otherwise} \end{cases}$$

where p is the probability of success on each trial.

We denote this as $X \sim \text{Geom}(p)$

The expected value is:

$$E(X) = \frac{1}{p}$$

Example: Lottery

Suppose there is a lottery where the chances of winning are $\frac{1}{10^6}$. What does $X \sim \text{Geom}(\frac{1}{10^6})$ and $E(X) = 10^6$ tell you about this? Say you always buy the ticket with your collection of lucky numbers.

- X represents the number of times the lottery is played until you win.
- On average ($E(X)$), you need to play the lottery 10^6 times to see your first win.

Example: Probability Calculation

Suppose that $X \sim \text{Geom}(\frac{1}{3})$. What is the probability that $X > 2$ (i.e., $P(X > 2)$)?

$$
\begin{aligned}
P(X > 2) &= 1 - P(X \leq 2) \\
&= 1 - \sum_{x=1}^{2} \frac{1}{3} \left(\frac{2}{3}\right)^{x-1} \\
&= 1 - \left[\frac{1}{3} + \frac{1}{3} \cdot \frac{2}{3}\right] \\
&= 1 - \frac{5}{9} = \frac{4}{9}
\end{aligned}
$$

Note: $p = \frac{1}{3}$ and $1 - p = \frac{2}{3}$

Binomial Distribution

Suppose that a fixed number of Bernoulli trials n, each trial with a probability of success p, are conducted in an independent manner. Let X be the total number of successes. X is a discrete random variable and we say that $X \sim \text{Bin}(n, p)$. In other words, X has a binomial distribution with parameters n and p.

Note: $X \sim \text{Bin}(1, p)$ is the same as $X \sim \text{Ber}(p)$.

To derive $P_X(x)$:

- Consider a sequence: $\underbrace{S \ldots S}_{(k \text{ successes})} \underbrace{F \ldots F}_{(n-k \text{ failures})}$
- Probability of this sequence: $p^k(1-p)^{n-k}$
- Number of possible rearrangements: $\binom{n}{k}$

Note: recall that $\binom{n}{k} = \frac{n!}{k!(n-k)!}$

Therefore, the probability mass function is:

$$
P_X(x) = \begin{cases} \binom{n}{x} p^x (1-p)^{n-x}, & x = 0, 1, 2, \ldots, n \\ 0, & \text{otherwise} \end{cases}
$$

Properties of Binomial Distribution:

$$
\begin{aligned}
E(X) &= np \\
V(X) &= npq \quad \text{where } q = 1 - p
\end{aligned}
$$

Consider a biased coin that comes up heads with a probability of $\frac{1}{4}$. We flip this coin 32 times in succession. Calculate the following:

1. What is the probability we see exactly 12 heads?
2. What is the probability we see between 7 and 12 heads?
3. What is the probability we see 20 tails?

4. What is the average number of heads (expected number of heads)?

If we define X as the random variable representing the number of heads observed, then X follows a binomial distribution with parameters $n = 32$ and $p = \frac{1}{4}$, i.e., $X \sim \text{Bin}(32, \frac{1}{4})$

1. $P(X = 12)$:
$$P_X(x) = \begin{cases} \binom{32}{12} \left(\frac{1}{4}\right)^{12} \left(\frac{3}{4}\right)^{20} & \text{if } x = 12 \\ 0 & \text{otherwise} \end{cases}$$
So $P(X = 12) = \binom{32}{12} \left(\frac{1}{4}\right)^{12} \left(\frac{3}{4}\right)^{20}$
2. $P(7 < X < 12)$ (There is some ambiguity in whether 7 and 12 are included but unless indicated otherwise between **excludes** endpoints):
$$P_X(x) = \begin{cases} \binom{32}{x} \left(\frac{1}{4}\right)^{x} \left(\frac{3}{4}\right)^{32-x} & \text{if } x = 7, 8, 9, 10, 11, 12 \\ 0 & \text{otherwise} \end{cases}$$
So $P(7 < X < 12) = \sum_{x=8}^{11} P_X(x) = \sum_{x=8}^{11} \binom{32}{x} \left(\frac{1}{4}\right)^{x} \left(\frac{3}{4}\right)^{32-x}$
3. $P(Y = 20)$ (equivalent to 20 tails):
$$P_Y(y) = \begin{cases} \binom{32}{20} \left(\frac{3}{4}\right)^{20} \left(\frac{1}{4}\right)^{12} & \text{if } y = 20 \\ 0 & \text{otherwise} \end{cases}$$
So $P(Y = 20) = \binom{32}{20} \left(\frac{3}{4}\right)^{20} \left(\frac{1}{4}\right)^{12}$
4. $E(X) = np = 32 \cdot \frac{1}{4} = 8$
5. $V(X) = npq = 32 \cdot \frac{1}{4} \cdot \frac{3}{4} = 6$

Lecture 13: Tuesday 10/8/2024

(went over study guide for exam 1 − I don't see a reason to put this here)

Lecture 14: Thursday 10/10/2024

(Exam 1)

Lecture 15: Tuesday 10/15/2024

Special Discrete Distributions

- Bernoulli Distribution
- Geometric Distribution
- Binomial Distribution
- Hypergeometric Distribution
- Poisson Distribution
- Poisson Process

Hypergeometric Distribution

Suppose an experiment involves randomly selecting n items from a population of size N that contains M successes.

Let X be a random variable that counts the number of successes in the sample.

Then X follows a hypergeometric distribution with parameters N, M, and n, denoted as:

$X \sim \text{Hypergeometric}(N, M, n)$

Where:

- N is the total population size
- M is the number of successes in the population
- n is the number of items drawn from the population

The probability mass function is given by:

$$P(X = x) = \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}}$$

for $x = 0, 1, 2, ..., \min(n, M)$

Note: The binomial distribution can approximate the hypergeometric distribution under suitable circumstances.

Poisson Distribution

Let $\lambda > 0$ be fixed. A random variable X has a Poisson distribution with parameter λ, written as $X \sim \text{Poisson}(\lambda)$, if:

$$P_X(x) = P(X = x) = \frac{\lambda^x e^{-\lambda}}{x!} \quad \text{for } x = 0, 1, 2, ...$$

Where λ is the average number of successes per unit of time or space.

Verify that $P_X(x)$ is a valid PMF:

Verification that $P_X(x)$ is a valid PMF:

1. $P_X(x) \geq 0$ for all x:
   - By convention, $P_X(x) = 0$ for $x \notin \{0, 1, 2, \ldots\}$
   - For $x \in \mathbb{N} \cup \{0\}$, $P_X(x) > 0$ because:
     - $e^{-\lambda} > 0$
     - $\lambda^x > 0$
     - $x! > 0$

   Therefore, $P_X(x)$ is non-negative for all x.

2. $\sum_{x=0}^{\infty} P_X(x) = 1$:

$$\begin{aligned}\sum_{x=0}^{\infty} P_X(x) &= \sum_{x=0}^{\infty} \frac{\lambda^x e^{-\lambda}}{x!} \\ &= e^{-\lambda} \sum_{x=0}^{\infty} \frac{\lambda^x}{x!} \\ &= e^{-\lambda} \cdot e^{\lambda} \\ &= 1\end{aligned}$$

Therefore, $P_X(x)$ is a valid PMF.
Expected Value and Variance:

- Expected Value:

$$\begin{aligned}E(X) &= \sum_{x=0}^{\infty} x \cdot P_X(x) \\ &= \sum_{x=0}^{\infty} x \cdot \frac{\lambda^x e^{-\lambda}}{x!} \\ &= \lambda e^{-\lambda} \sum_{x=1}^{\infty} \frac{\lambda^{x-1}}{(x-1)!} \\ &= \lambda e^{-\lambda} \cdot e^{\lambda} \\ &= \lambda\end{aligned}$$

- Variance:

$$\begin{aligned}V(X) &= E(X^2) - (E(X))^2 \\ &= \lambda + \lambda^2 - \lambda^2 \\ &= \lambda\end{aligned}$$

Poisson Approximation to Binomial:
Let $X \sim \text{Bin}(n, p)$ and $Y \sim \text{Poisson}(np)$. Then for any subset $A \subseteq \{0, 1, 2, \ldots\}$:

$$|P(X \in A) - P(Y \in A)| \leq np^2$$

Note: $E(X) = E(Y) = np$
Examples:

- For $n = 100$, $p = \frac{1}{2}$:

$$|P(X \in A) - P(Y \in A)| \leq 100 \cdot \left(\frac{1}{2}\right)^2 = 25$$

- For $n = 100$, $p = \frac{1}{10^4}$:

$$|P(X \in A) - P(Y \in A)| \leq 100 \cdot \left(\frac{1}{10^4}\right)^2 = \frac{1}{10^6}$$

The Poisson distribution can be used to approximate probabilities of rare events (when np is small). This is known as the law of rare events.

Example: Suppose a large factory with many workers experiences 3 accidents each month on average. Approximate the probability that a particular month has 2 accidents.
Solution:

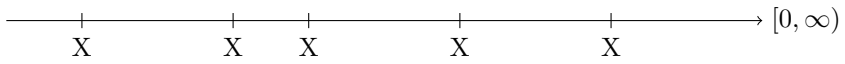
- Let X be the number of accidents in a fixed month
- Assume $X \sim \text{Poisson}(3)$ since $E(X) = 3$
- $P(X = 2) = e^{-3}\frac{3^2}{2!} \approx 0.2240$

Poisson Process:
A Poisson process models occurrences of events over time. A Poisson Process with intensity or rate $\lambda > 0$ is a collection of random points on $[0, \infty)$ with the following properties:

1. The points are distinct.
2. The number of points in a bounded interval $I \subseteq [0, \infty)$, denoted by $N(I)$, follows a Poisson distribution with parameter $\lambda \cdot \text{length}(I)$.
3. If $I_1, I_2, \ldots, I_n$ are non-overlapping intervals in $[0, \infty)$, then $N(I_1), N(I_2), \ldots, N(I_n)$ are independent random variables.

Lecture 16: Thursday 10/17/2024

Poisson Process



(X is a event that takes place)

A Poisson process with intensity (rate) $\lambda > 0$ is a collection of random points on $[0, \infty)$ with the following properties:

- The points are distinct.
- The number of points in any interval $I \subseteq [0, \infty)$, denoted by $N(I)$, follows a Poisson distribution with parameter $\lambda \cdot \text{length}(I)$.
- If $I_1, I_2, \ldots, I_n$ are non-overlapping intervals in $[0, \infty)$, then $N(I_1), N(I_2), \ldots, N(I_n)$ are independent random variables.

Example: Suppose customers arrive at a store according to a Poisson process with intensity 5/hr. The store is open from 9am to 6pm.

i) Find the probability that no customer comes to the store within one hour of opening.

$N([9, 10]) \sim \text{Poisson}(5)$

$P(N([9, 10]) = 0) = e^{-5}\frac{5^0}{0!} = e^{-5}$

ii) Find the probability that there are two customers between 9-10, three customers between 10-10:30, and five between 2-3:30.

$P(N([9, 10]) = 2, N([10, 10:30]) = 3, N([14, 15:30]) = 5)$

$= P(N([9, 10]) = 2) \cdot P(N([10, 10:30]) = 3) \cdot P(N([14, 15:30]) = 5)$

Where:

$$N([9, 10]) \sim \text{Poisson}(5)$$
$$N([10, 10:30]) \sim \text{Poisson}(2.5)$$
$$N([14, 15:30]) \sim \text{Poisson}(7.5)$$

Required probability:

$$= \frac{5^2}{2!}e^{-5} \cdot \frac{2.5^3}{3!}e^{-2.5} \cdot \frac{7.5^5}{5!}e^{-7.5}$$
$$= e^{-15} \cdot \frac{5^2}{2!} \cdot \frac{2.5^3}{3!} \cdot \frac{7.5^5}{5!}$$

iii) Find the probability that three customers arrive between 10-10:30 given 12 customers showed up between 10-12.

$$
\begin{aligned}
P(N([10,10.5]) = 3 \mid N([10,12]) = 12) &= \frac{P(N([10,10.5]) = 3, N([10.5,12]) = 9)}{P(N([10,12]) = 12)} \\
&= \frac{P(N([10,10.5]) = 3) \cdot P(N([10.5,12]) = 9)}{P(N([10,12]) = 12)} \\
&= \frac{e^{-2.5}\frac{2.5^3}{3!} \cdot e^{-7.5}\frac{7.5^9}{9!}}{e^{-10}\frac{10^{12}}{12!}} \\
&= \frac{2.5^3 \cdot 7.5^9 \cdot 12!}{3! \cdot 9! \cdot 10^{12}}
\end{aligned}
$$

Special Distributions

Continuous Distributions

- Uniform Distribution
- Exponential Distribution
- Normal Distribution

Uniform Distribution

In the discrete case, uniform meant that each individual outcome was equally likely.

For any continuous random variable X, $P(X = a) = \int_a^a f(x)dx = 0$ for any specific value a.

So directly extending this idea doesn't work. For a continuous random variable, a reasonable way to extend this is to say that no one region should have a higher probability of producing an outcome as any other as long as both regions have the same length.

If $X \sim U([a,b])$, then:

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

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

$$
E(X) = \int_{-\infty}^{\infty} x f_X(x)dx = \int_a^b x\frac{1}{b-a}dx = \frac{a+b}{2}
$$

$$
V(X) = E(X^2) - (E(X))^2 = \int_{-\infty}^{\infty} x^2 f_X(x)dx - (\frac{a+b}{2})^2 = \frac{(b-a)^2}{12}
$$

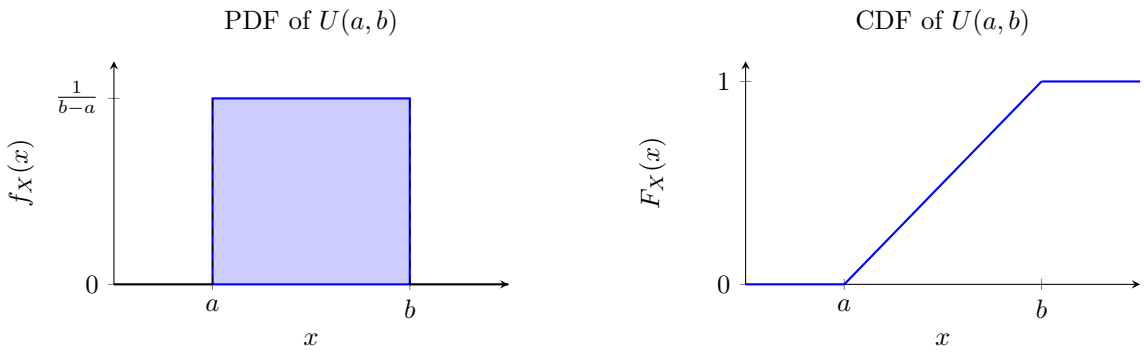


Figure 10: PDF and CDF of Uniform Distribution $U(a,b)$

Exponential Distribution

The exponential distribution is related to the Poisson distribution.

We say that $X \sim \text{Exp}(\lambda)$ if:

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

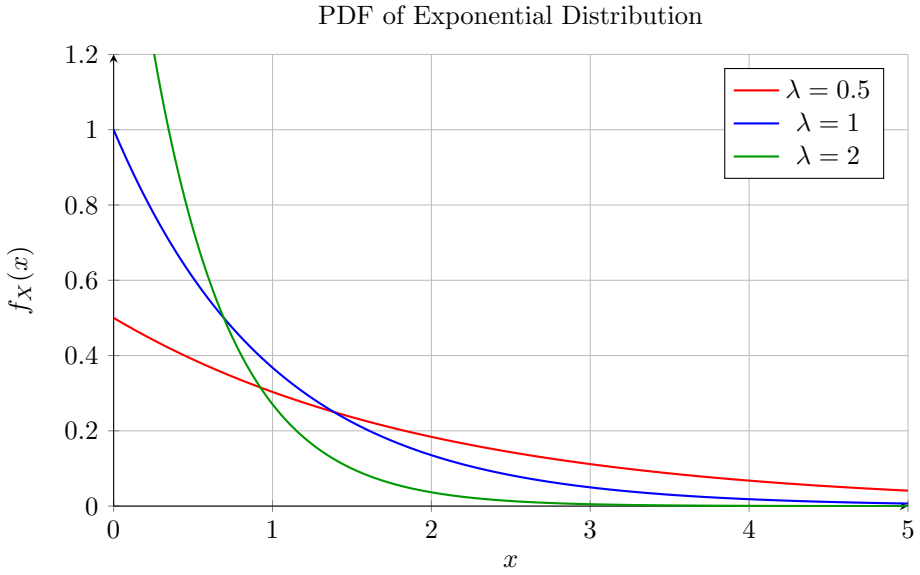


Figure 11: Probability Density Function (PDF) of Exponential Distribution for different λ values

The exponential distribution is often used to model wait times, with $\lambda = \frac{1}{\text{average wait}}$.

$$E(X) = \frac{1}{\lambda}$$

$$V(X) = \frac{1}{\lambda^2}$$

$$F_X(x) = \begin{cases} 0 & \text{if } x < 0 \\ 1 - e^{-\lambda x} & \text{if } x \geq 0 \end{cases}$$

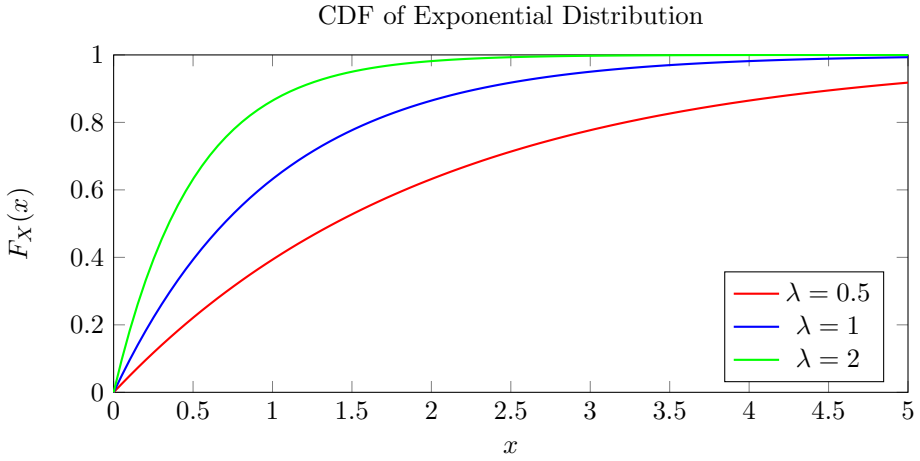


Figure 12: Cumulative Distribution Function (CDF) of Exponential Distribution for different λ values

$$F_X(x) = \int_{-\infty}^{x} f_X(t)dt = 1 - e^{-\lambda x}$$

$$F_X(x) = \int_{0}^{x} \lambda e^{-\lambda t}dt = 1 - e^{-\lambda x}$$

Example: Restaurant Wait Times

The average wait time at a restaurant to be seated is 2 minutes. Two customers walk into the restaurant at different times of the day. What is the probability they both were seated within 4 minutes of their arrival? (Assume that the wait times are exponentially distributed)

Let $X_1 \sim \text{Exp}(0.5)$ be the wait time for the 1st customer and $X_2 \sim \text{Exp}(0.5)$ be the wait time for the 2nd customer.

$$\begin{aligned} P(X_1 \leq 4, X_2 \leq 4) &= P(X_1 \leq 4) \cdot P(X_2 \leq 4) \quad \text{(Assuming independence)} \\ &= (1 - e^{-0.5 \cdot 4}) \cdot (1 - e^{-0.5 \cdot 4}) \\ &= (1 - e^{-2})^2 \end{aligned}$$

Note: It's often faster to use the CDF instead of computing integrals in this case.

Lecture 17: Tuesday 10/22/2024

Normal Distribution

The random variable X has a normal distribution with mean μ and variance σ^2 (denoted as $X \sim N(\mu, \sigma^2)$) if:

X has a pdf given by:

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \text{ where } x \in \mathbb{R}$$

We begin by examining the standard normal distribution, where $\mu = 0$ and $\sigma^2 = 1$.

For $X \sim N(0, 1)$, the probability density function (pdf) is given by:

$$f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}, \quad x \in \mathbb{R}$$

Since there is no straightforward antiderivative for this function, evaluating the integral directly poses challenges.

To overcome this issue, we utilize pre-calculated values (typically provided in tables) for $P(Z \leq z)$, where $Z \sim N(0, 1)$.

It is conventional to denote $P(Z \leq z)$ as $\Phi(z)$. In other words:

$$\Phi(z) = P(Z \leq z) = F_Z(z) \quad \text{for } Z \sim N(0, 1)$$

$\Phi(z)$ is referred to as the cumulative distribution function (cdf) of the standard normal distribution.

Examples:

- $P(Z \leq 1.19) = \Phi(1.19) = 0.8830$
- $P(-1.23 \leq Z \leq 1.43)$

$$\begin{aligned} &= P(Z \leq 1.43) - P(Z \leq -1.23) \\ &= \Phi(1.43) - \Phi(-1.23) \\ &= 0.9236 - 0.1093 \\ &= 0.8143 \end{aligned}$$

Important property: The standard normal distribution is symmetric about 0, i.e., $f_X(x) = f_X(-x)$ for all x so $f_X(x)$ is even.

Handling Non-Standard Normal Distributions:

For non-standard normal distributions, we use integration by substitution.

Let $Z = \frac{X-\mu}{\sigma}$, where $X \sim N(\mu, \sigma^2)$.

Exercise: Show that:

$$\int_a^b \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2} dx = \int_{\frac{a-\mu}{\sigma}}^{\frac{b-\mu}{\sigma}} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} dz$$

We compute probabilities for $X \sim N(\mu, \sigma^2)$ in terms of $Z \sim N(0, 1)$ using the integration substitution above.

If $X \sim N(\mu, \sigma^2)$, then $Z = \frac{X-\mu}{\sigma} \sim N(0,1)$, and:

$$P(a < X < b) = P\left(\frac{a-\mu}{\sigma} < Z < \frac{b-\mu}{\sigma}\right) = \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)$$

Example: Suppose that the heights of females on a particular island can be modeled using a normal distribution with mean 150 cm and variance 25 cm^2.

What is the probability that a randomly selected female has a height that falls between 145 cm and 160 cm?

Given: $X \sim N(150, 25)$

We want to find: $P(145 < X < 160)$

Solution: Let $Z = \frac{X-150}{5}$, then:

$$\begin{aligned} P(145 < X < 160) &= P\left(\frac{145-150}{5} < Z < \frac{160-150}{5}\right) \\ &= P(-1 < Z < 2) \\ &= \Phi(2) - \Phi(-1) \end{aligned}$$

= plug in values from the table.

Example: Suppose $X \sim N(\mu, \sigma^2)$. What is the probability that observation falls within a standard deviation of the mean?

$X \sim N(\mu, \sigma^2)$

$$\begin{aligned} P(|X-\mu| < \sigma) &= P(-\sigma < X - \mu < \sigma) \\ &= P(\mu - \sigma < X < \mu + \sigma) \end{aligned}$$

Let $Z = \frac{X-\mu}{\sigma}$, then $Z \sim N(0,1)$

$$\begin{aligned} P(|X-\mu| < \sigma) &= P\left(\frac{\mu-\sigma-\mu}{\sigma} < Z < \frac{\mu+\sigma-\mu}{\sigma}\right) \\ &= P(-1 < Z < 1) \\ &= \Phi(1) - \Phi(-1) \\ &= 0.8413 - 0.1587 \\ &= 0.6826 \end{aligned}$$

Random Vectors / Joint Distributions

A random vector $\underline{X} : \Omega \to \mathbb{R}^n$ is given by:

$$\underline{X} = \begin{pmatrix} X_1 \\ \vdots \\ X_n \end{pmatrix}$$

where $X_1, \ldots, X_n : \Omega \to \mathbb{R}$.

Example: Let $\Omega = \{$people in a particular hospital at a given fixed point in time$\}$.

- $X_1(\omega)$ = blood pressure (systolic) of ω
- $X_2(\omega)$ = weight of ω
- $X_3(\omega)$ = height of ω

Then, $\underline{X} = \begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix}$

Example: Cat pictures in machine learning algorithms
Let $\Omega = \{\text{set of cat pictures of a fixed type}\}$
For a black and white image with 32x32 pixels:

- Each pixel is represented by a random variable
- Total number of random variables $= 32 \times 32 = 1024$

Thus, we have a random vector $\underline{X} : \Omega \to \mathbb{R}^{1024}$

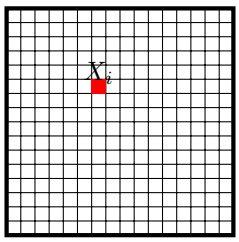


Figure 13: Representation of a 32x32 pixel image with highlighted pixel X_i

Let $\underline{X} = (X_1, X_2, \ldots, X_{1024})$ be a random vector where:

- $X_i(\omega)$ represents the brightness/darkness of the i-th pixel for picture ω
- ω denotes a specific picture in the sample space

In image classification algorithms, the goal is to compute $P(\text{label} \mid \text{picture})$. For example:

- label $\in \{\text{cat}, \text{not cat}\}$
- picture is represented by the random vector $\underline{X}$

Thus, we can express this probability as:

$$P(\text{label} = \text{cat} \mid \text{picture}) = P(\text{label} = \text{cat} \mid \underline{X} = \underline{x})$$

Key Ideas:

- A single random variable is often insufficient to describe complex problems.
- Many real-world applications require analyzing how multiple random variables interact when considered together.

Lecture 18: Thursday 10/24/2024

Random Vectors

In many situations, it is not sufficient to keep track of just one random quantity - we need to track several random quantities simultaneously. This leads us to the concept of random vectors.

Definition: A random vector is an ordered collection of random variables that maps from a sample space Ω to a multi-dimensional real space:

$$\begin{aligned} \text{Individual components:} \quad & X_1, \ldots, X_n : \Omega \to \mathbb{R} \\ \text{Vector notation:} \quad & \underline{X} = (X_1, \ldots, X_n)^T \\ \text{Overall mapping:} \quad & \underline{X} : \Omega \to \mathbb{R}^n \end{aligned}$$

Computing Probabilities for Random Vectors

To understand how to compute probabilities for random vectors, let's first review the simpler case of a single random variable:

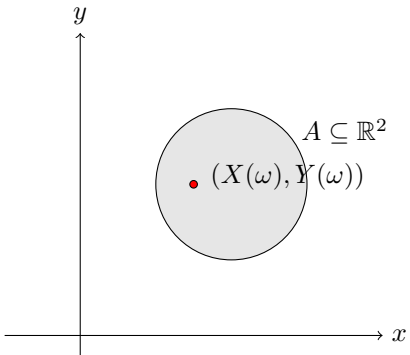
For a single random variable X:

- X maps from Ω to $\mathbb{R}$
- For any subset $A \subseteq \mathbb{R}$
- The probability is: $P(X \in A) = P(\{\omega \in \Omega : X(\omega) \in A\})$

This concept extends naturally to random vectors. For a two-dimensional random vector $\underline{X} = \begin{pmatrix} X \\ Y \end{pmatrix}$:

$$P(\underline{X} \in A) = P(\{\omega \in \Omega : \underline{X}(\omega) \in A\}) = P\left(\left\{\omega : \begin{pmatrix} X(\omega) \\ Y(\omega) \end{pmatrix} \in A\right\}\right)$$

where A is now a subset of $\mathbb{R}^2$. For example, A could be a circle, rectangle, or any other two-dimensional region:



Example: A is a circle in $\mathbb{R}^2$ containing point $(X(\omega), Y(\omega))$

We would like to do the same thing with a random vector $\underline{X} = \begin{pmatrix} X \\ Y \end{pmatrix}$.

We introduce the notions of joint probability mass functions and joint probability density functions for this purpose.

Joint Probability Mass Function (PMF)

The joint PMF for a random vector $\begin{pmatrix} X \\ Y \end{pmatrix}$ is the function $P_{X,Y}(x, y)$ such that:

$$P_{X,Y}(x, y) = P(X = x, Y = y)$$

Note: A discrete random vector is one that takes countably many values.
Example: Consider an experiment consisting of rolling two fair dice.

- Let X_1 be a random variable tracking the first die's outcome
- Let X_2 be a random variable tracking the second die's outcome

The sample space is:

$$\Omega = \{(i, j) : i, j \in \{1, 2, 3, 4, 5, 6\}\}$$

And the random variables are defined as:

$$X_1(i, j) = i \quad \text{(first die outcome)}$$
$$X_2(i, j) = j \quad \text{(second die outcome)}$$

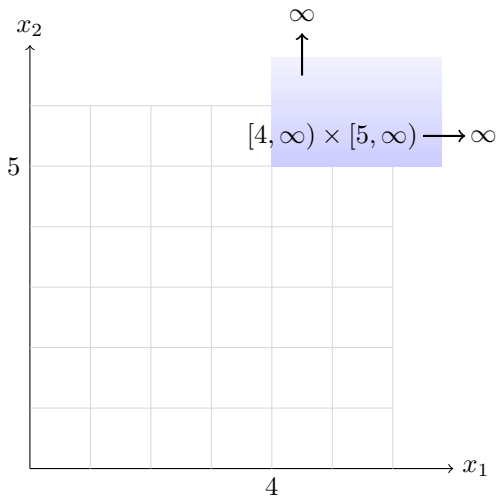
The joint PMF of X_1 and X_2 is:

$$P_{X_1,X_2}(x_1, x_2) = \begin{cases} \frac{1}{36} & \text{if } x_1, x_2 \in \{1, 2, 3, 4, 5, 6\} \\ 0 & \text{otherwise} \end{cases}$$

This reflects that each possible outcome (x_1, x_2) has equal probability $\frac{1}{36}$ since the dice are fair.

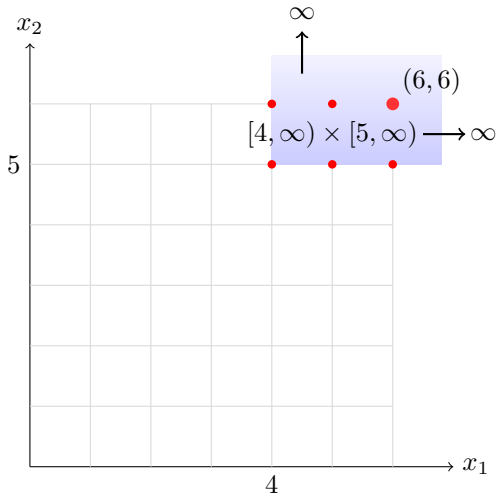
Let $\underline{X} = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$ be our random vector.

We want to compute $P(\underline{X} \in [4, \infty) \times [5, \infty))$, which represents the probability that $X_1 \geq 4$ and $X_2 \geq 5$ simultaneously.



Key equation:

$$P(\underline{X} \in A) = \sum_{(\underline{x} \in A)} P_{\underline{X}}(\underline{x}) where P_{\underline{X}} > 0$$



For region $A = [4, \infty) \times [5, \infty)$, we need to find the probability that our random variable falls in this region.

$$P(\underline{X} \in A) = \sum_{(\underline{x} \in A)} P_{\underline{X}}(\underline{x})$$

Looking at the graph above, we can see that region A contains these specific points:

$$\underline{X} \in \{(4, 5), (4, 6), (5, 5), (5, 6), (6, 5), (6, 6)\}$$

Since we're rolling two fair dice, each possible outcome has an equal probability of $\frac{1}{36}$. Let's add up the probabilities:

$$\begin{aligned}
P(\underline{X} \in A) &= \sum_{(x_1,x_2)\in A} P_{X_1,X_2}(x_1, x_2) \\
&= P_{X_1,X_2}(4,5) + P_{X_1,X_2}(4,6) + P_{X_1,X_2}(5,5) \\
&\quad + P_{X_1,X_2}(5,6) + P_{X_1,X_2}(6,5) + P_{X_1,X_2}(6,6) \\
&= \underbrace{\frac{1}{36} + \frac{1}{36} + \frac{1}{36} + \frac{1}{36} + \frac{1}{36} + \frac{1}{36}}_{\text{We have 6 points in region A, each with probability } \frac{1}{36}} \\
&= \frac{6}{36} = \frac{1}{6}
\end{aligned}$$

Important Notes:

- Just like with a regular probability mass function (PMF), we can find probabilities by adding up values from a joint PMF
- For any joint PMF, all probabilities must be non-negative ($P_{X,Y}(x, y) \geq 0$) and sum to 1: $\sum_{(x,y)\in\mathbb{R}^2} P_{X,Y}(x, y) = 1$

Example: Coin Toss and Die Roll

Let's work through a clear example. Consider this two-step experiment:

1. First, we toss a fair coin
2. Then, based on the coin result:
    - If we get heads (H): We roll a regular fair 6-sided die
    - If we get tails (T): We roll a fair 4-sided die

Let's define our variables:

- Let X be the coin toss result (H or T)
- Let Y be the number we get from the die roll

All possible outcomes are:

$$\Omega = \{(H,1), (H,2), (H,3), (H,4), (H,5), (H,6), (T,1), (T,2), (T,3), (T,4)\}$$

Let's find the joint PMF $P_{X,Y}(x, y)$ step by step:
To find this, we need two pieces:

- The probability of getting each coin result ($P_X(x)$)
- The probability of getting each die roll given the coin result ($P_{Y|X}(y|x)$)

Step 1: Probability of coin results ($P_X(x)$) Since we're using a fair coin:

$$P_X(x) = \begin{cases} \frac{1}{2} & \text{if } x \text{ is H or T} \\ 0 & \text{for any other value} \end{cases}$$

Step 2: Probability of die rolls given coin result ($P_{Y|X}(y|x)$)

$$P_{Y|X}(y|x) = \begin{cases} \frac{1}{6} & \text{if we got heads (H) and } y \text{ is 1,2,3,4,5, or 6} \\ \frac{1}{4} & \text{if we got tails (T) and } y \text{ is 1,2,3, or 4} \\ 0 & \text{for any other combination} \end{cases}$$

Note: This only becomes a proper PMF after you specify whether you got heads or tails.

Step 3: Putting it together - The Joint PMF We multiply the probabilities from steps 1 and 2: $P_{X,Y}(x,y) = P_X(x) \cdot P_{Y|X}(y|x)$

$$P_{X,Y}(x,y) = \begin{cases} \frac{1}{12} & \text{if we got heads (H) and } y \text{ is 1,2,3,4,5, or 6} \\ \frac{1}{8} & \text{if we got tails (T) and } y \text{ is 1,2,3, or 4} \\ 0 & \text{for any other combination} \end{cases}$$

Let's verify this is correct: All probabilities should sum to 1:

$$\underbrace{6 \cdot \frac{1}{12}}_{\text{6 possible outcomes with H}} + \underbrace{4 \cdot \frac{1}{8}}_{\text{4 possible outcomes with T}} = \frac{1}{2} + \frac{1}{2} = 1$$

Understanding Conditional Probabilities with Multiple Variables:

When we have more than two variables, we can condition on multiple events:

- $P_{X_1|X_2,X_3}(x_1|x_2,x_3)$ means "probability of X_1 given we know both X_2 and X_3"
- $P_{X_1,X_3|X_2}(x_1,x_3|x_2)$ means "joint probability of X_1 and X_3 given we know X_2"

Example: Finding the Probability of Rolling a 1 Let's find $P_Y(1)$ (the probability of rolling a 1 regardless of the coin flip):

$$P_Y(y_0) = \sum_x P_{X,Y}(x,y_0) \quad \text{(Add up probabilities for all possible coin results)}$$

$$P_Y(1) = \sum_x P_{X,Y}(x,1)$$

$$= P_{X,Y}(H,1) + P_{X,Y}(T,1) \quad \text{(Probability with heads + Probability with tails)}$$

$$= \frac{1}{12} + \frac{1}{8} = \frac{5}{24}$$

Understanding Independent Random Variables:

Two random variables X and Y are independent if knowing one tells us nothing about the other. Mathematically, this means their joint PMF equals the product of their individual PMFs:

$$P_{X,Y}(x,y) = P_X(x)P_Y(y)$$

In our coin and die example, X and Y are NOT independent because:

- Knowing we got tails (X = T) tells us Y can't be 5 or 6
- Knowing we got Y = 6 tells us we must have gotten heads

Lecture 19: Tuesday 10/29/2024

Random Vectors

Question: How do we assign probabilities for random vectors (those that take countably many values)?

For discrete random vectors, we use the *joint PMF*.
For continuous random vectors, we use the *joint probability density function*.

Key Equations:
Let $P_{X,Y}(x,y)$ be the joint PMF for x, y, and let $\underline{x} = (x,y)$. Then:

$$P(\underline{x} \in A) = \sum_{(\underline{x} \in A)} P_{X,Y}(\underline{x}) \quad \text{where } P_{X,Y}(\underline{x}) > 0$$

$$P_{X,Y}(x,y) = P_{Y|X}(y|x)P(x) = P_{X|Y}(x|y)P(y)$$

For independent X and Y:

$$P_{X,Y}(x,y) = P_X(x)P_Y(y)$$

Law of Total Probability:

$$P_X(x) = \sum_y P_{X,Y}(x,y) \quad \text{where } P_{X,Y}(x,y) > 0$$

Joint Continuity:
Random variables X, Y are jointly continuous if there exists a function $f_{X,Y}(x,y)$ (joint probability density function) such that:

- $f_{X,Y}(x,y) \geq 0$ for all (x,y)
- $\int_A f_{X,Y}(x,y)\, d\mu = P((X,Y) \in A)$ for all "nice" $A \subseteq \mathbb{R}^2$

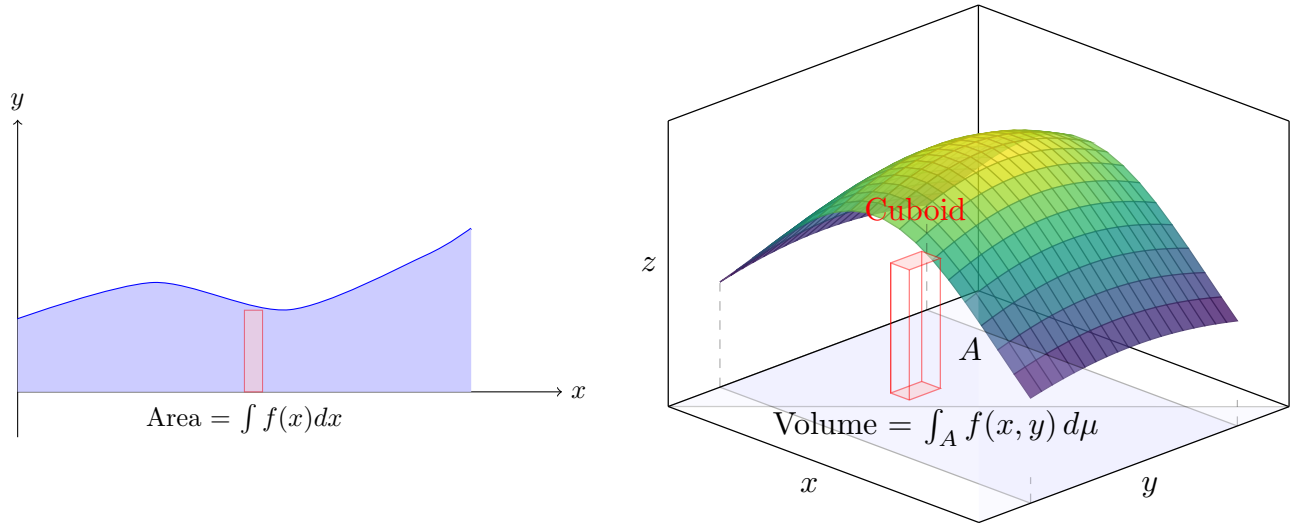
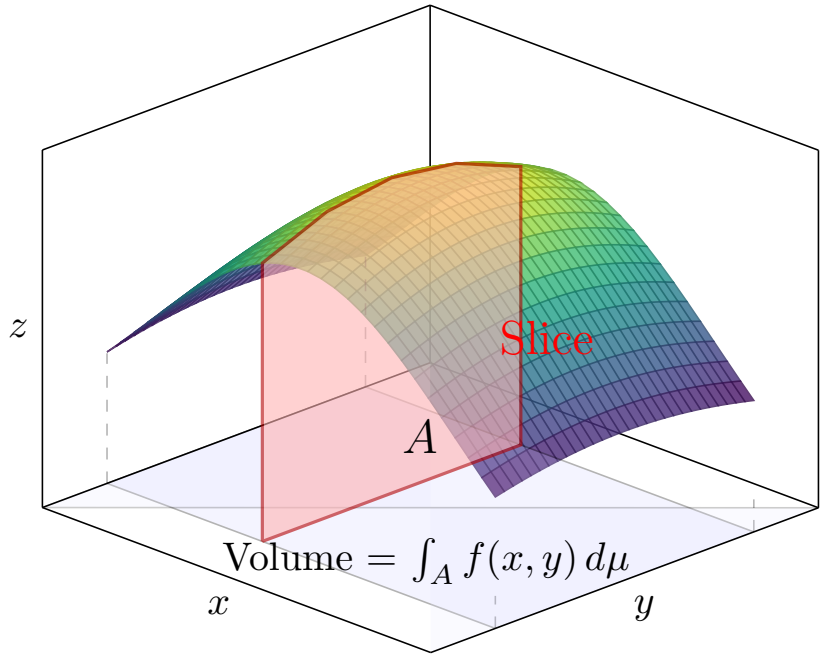


Figure 14: Comparison of single and double integrals: area under a curve (left) vs volume under a surface (right)

This figure demonstrates the Riemann sum approach to finding the volume under a surface.

While this approach is conceptually helpful, it temporarily prevents us from using the Fundamental Theorem of Calculus.

To overcome this limitation, we can use iterated integrals. Iterated integrals allow us to compute the volume by adding up infinitesimally thin slices of the surface. To understand iterated integrals, consider slicing the volume vertically:



The volume of a single slice can be approximated by:

$$\Delta x \int_{a(x)}^{b(x)} f(x,y)\,dy$$

The total volume is approximately the sum of all slices:

$$\sum_{i=1}^{n} \Delta x \int_{a(x)}^{b(x)} f(x,y)\,dy$$

Taking the limit as $\Delta x \to 0$, the exact volume is:

$$\text{Volume} = \int_c^d \int_{a(x)}^{b(x)} f(x,y)\,dy\,dx$$

Alternatively, we could slice along the y-axis instead of the x-axis, yielding:

$$\int_e^f \int_{g(y)}^{h(y)} f(x,y)\,dx\,dy$$

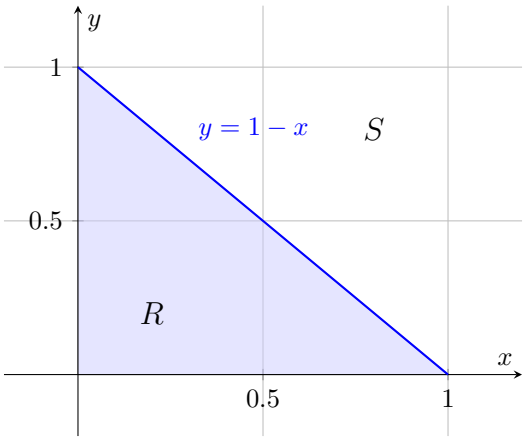
For all examples we will encounter in this class (and most examples in general), the order of integration will yield identical results.

Note: The drawings above are not to scale nor perfect representations, but are meant to illustrate the general concept of integration over a region in 3D space.

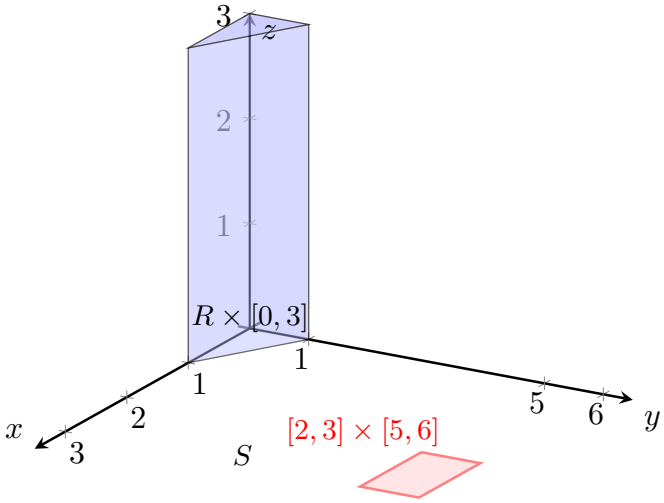
Example: Suppose that $f_{X,Y}(x,y) = \begin{cases} k & \text{for } (x,y) \in R \\ 0 & \text{otherwise} \end{cases}$
is a joint probability density function, where R is the triangular region bounded by:

$$0 \leq x \leq 1, \quad 0 \leq y \leq 1 - x$$

1. Compute the value of k
2. Compute $P(Y > X)$



For example what is $P\big((X,Y) \in [2,3] \times [5,6]\big)$



$P\big((X,Y) \in [2,3] \times [5,6]\big) = 0$ because the region does not overlap with R.
The region R for which $f(x,y)$ is non-zero contains the possible outcomes for X and Y. Refer

to the diagrams above for the following:

i) Verify that $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)\,dx\,dy = 1$

Since $f(x,y)$ is a joint pdf:

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)\,dx\,dy = 1$$

Split into regions R and complement S:

$$= \int_R f(x,y)\,dx\,dy + \int_S f(x,y)\,dx\,dy$$

Since $f(x,y) = 0$ for all $(x,y) \in S$:

$$= \int_R f(x,y)\,dx\,dy + 0$$

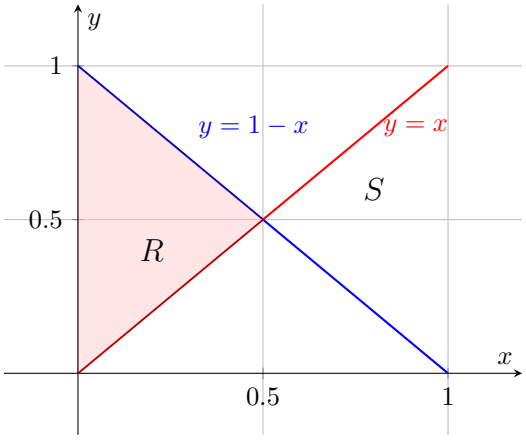
Therefore:

$$= \int_R f(x,y)\,dx\,dy = 1$$

Because we have dx, dy in the integrand, we can integrate with respect to x first, then y.

$$\int_0^1\int_0^{1-y} k\,dx\,dy = 1$$
$$k\int_0^1\int_0^{1-y} 1\,dx\,dy = 1$$
$$k\int_0^1 (1-y)\,dy = 1$$
$$k\left(\int_0^1 1\,dy - \int_0^1 y\,dy\right) = 1$$
$$k\left([y]_0^1 - \left[\frac{y^2}{2}\right]_0^1\right) = 1$$
$$k\left(1 - \frac{1}{2}\right) = 1$$
$$k\left(\frac{1}{2}\right) = 1$$
$$k = 2$$

ii) Compute $P(Y > X)$



$$y = x$$

$$y = 1 - x \quad \text{where } x = \frac{1}{2}$$

Find $P(Y > x)$:

$$P(Y > x) = \int_0^{1/2} \int_x^{1-x} 2\, dy\, dx$$

$$= \int_0^{1/2} 2(1 - x - x)\, dx$$
$$= \int_0^{1/2} 2(1 - 2x)\, dx$$
$$= 2\left([x]_0^{\frac{1}{2}} - \left[x^2\right]_0^{\frac{1}{2}} \right)$$
$$= 2\left(\frac{1}{2} - \frac{1}{4} \right)$$
$$= 2\left(\frac{1}{4} \right)$$
$$= \frac{1}{2}$$

Lecture 20: Thusday 10/31/2024

Jointly Continuous Random Variables X and Y

Let $f(x, y)$ be the joint probability density function (PDF) of X and Y:

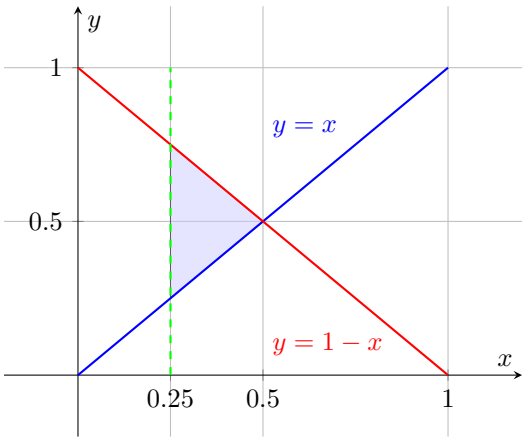
$$f(x,y) = \begin{cases} k & \text{for } (x,y) \in R \\ 0 & \text{otherwise} \end{cases}$$

Important Properties:

- $P(X = \frac{1}{2}, Y = \frac{1}{2}) = 0$
- $P(0 < X < \frac{1}{2}, 0 < Y < \frac{1}{2}) > 0$
- $P(X > 1, Y > 1) = 0$

Key Calculations:

1. Total probability: $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x,y)\, dx\, dy = 1$
2. $P(Y > X)$ was calculated above
3. $P(X > \frac{1}{4} | Y > X) = \frac{P(X > \frac{1}{4}, Y > X)}{P(Y > X)}$



The probability can be calculated as:

$$P(X > \frac{1}{4}, Y > X) = \int_{\frac{1}{4}}^{\frac{1}{2}} \int_{x}^{1-x} 2\, dy\, dx$$

$$= 2\int_{\frac{1}{4}}^{\frac{1}{2}} [y]_x^{1-x}\, dx$$

$$= 2\int_{\frac{1}{4}}^{\frac{1}{2}} [(1-x)-x]\, dx$$

$$= 2\int_{\frac{1}{4}}^{\frac{1}{2}} (1-2x)\, dx$$

$$= 2\left[x - x^2\right]_{\frac{1}{4}}^{\frac{1}{2}}$$

$$= 2\left[\left(\frac{1}{2}-\frac{1}{4}\right)-\left(\frac{1}{4}-\frac{1}{16}\right)\right]$$

$$= 2\left[\frac{1}{4}-\frac{3}{16}\right]$$

$$= 2\left[\frac{4}{16}-\frac{3}{16}\right]$$

$$= 2\left[\frac{1}{16}\right]$$

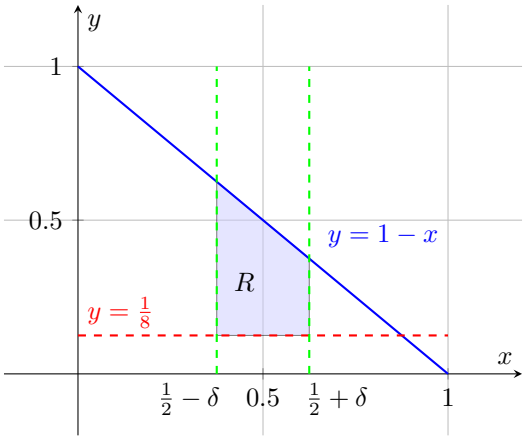
$$= \frac{1}{8}$$

Therefore:

$$P(X > \frac{1}{4} \mid Y > X) = \frac{P(X > \frac{1}{4}, Y > X)}{P(Y > X)} = \frac{\frac{1}{8}}{\frac{1}{2}} = \frac{1}{4}$$

iv) $P(X > \frac{1}{8} \mid Y = \frac{1}{2})$
The conditional probability $P(X > \frac{1}{8} \mid Y = \frac{1}{2})$ is undefined because:

1. $P(X > \frac{1}{8} \mid Y = \frac{1}{2}) = \frac{P(X > \frac{1}{8}, Y = \frac{1}{2})}{P(Y = \frac{1}{2})}$
2. $P(Y = \frac{1}{2}) = 0$
3. $P(X > \frac{1}{8}, Y = \frac{1}{2}) = 0$



$$\lim_{\delta \to 0} P(X > \frac{1}{8} \mid Y = \frac{1}{2}) = \frac{P(X > \frac{1}{8}, Y = \frac{1}{2})}{P(Y = \frac{1}{2})}$$

$$f_{x|y}(x|y) = \frac{f_{x,y}(x,y)}{f_y(y)}$$

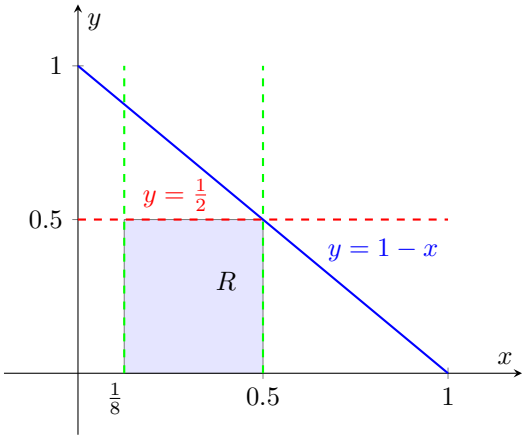
$$f_{x|y}(x|y)$$

is the conditional probability density function of X given $Y = y$

$$\frac{f_{x,y}(x,y)}{f_y(y)}$$

represents the ratio of joint PDF to marginal PDF of Y

$$P(X > \frac{1}{8} | Y = \frac{1}{2}) = \int_{\frac{1}{2}-\delta}^{\frac{1}{2}+\delta} f_{x|y}(x|\frac{1}{2})\, dx$$



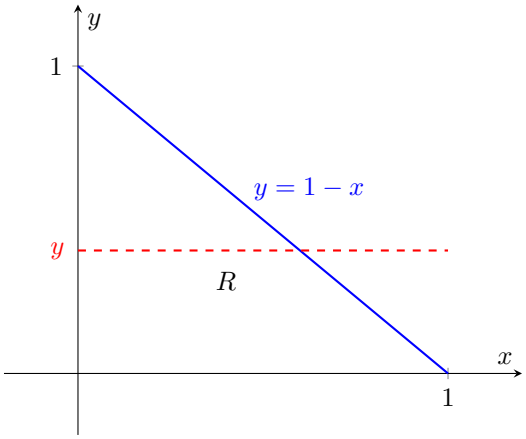
$$P(X > \frac{1}{8} | Y = \frac{1}{2}) = \int_{\frac{1}{8}}^{\frac{1}{2}} f_{x|y}(x|\frac{1}{2})\, dx$$

$$f_{x|y}(x|\frac{1}{2}) = \frac{f_{x,y}(x, \frac{1}{2})}{f_y(\frac{1}{2})}$$

Key equation:

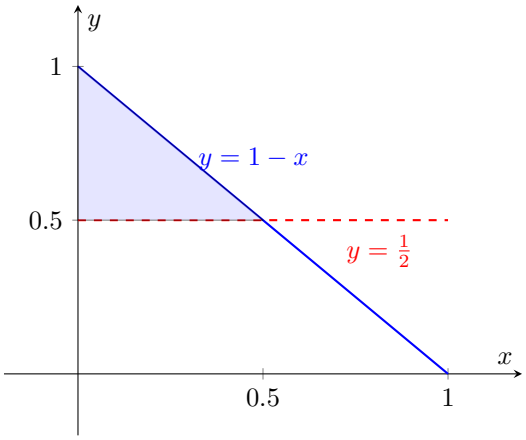
$$f_y(y) = \int_{-\infty}^{\infty} f_{x,y}(x, y)\, dx$$

$$\int_0^{1-y} 2\, dx = 2(1-y) \quad \text{for } 0 \le y \le 1$$



The marginal pdf of Y is: $2(1-y)$
Example: Find $P(Y > \frac{1}{2})$

$$P(Y > \frac{1}{2}) = \int_{\frac{1}{2}}^{1} \int_0^{1-y} f_{x,y}(x, y)\, dx\, dy$$
$$= \int_{\frac{1}{2}}^{1} f_y(y)\, dy$$



We can compute probabilities for expressions with just y (or just x) by using the marginal PDF of y (or x).

$$\begin{aligned}
P(X > \frac{1}{8} | y = \frac{1}{2}) &= \int_{\frac{1}{8}}^{\frac{1}{2}} \frac{f_{x,y}(x, \frac{1}{2})}{f_y(\frac{1}{2})}\, dx \\
&= \int_{\frac{1}{8}}^{\frac{1}{2}} \frac{2}{2(1 - \frac{1}{2})}\, dx \\
&= \int_{\frac{1}{8}}^{\frac{1}{2}} 2\, dx \\
&= 2\,[x]_{\frac{1}{8}}^{\frac{1}{2}} \\
&= 2\left[\frac{1}{2} - \frac{1}{8}\right] \\
&= 2\left[\frac{3}{8}\right] \\
&= \frac{3}{4}
\end{aligned}$$

Testing Independence of X and Y

To determine if X and Y are independent, we use the key equation:

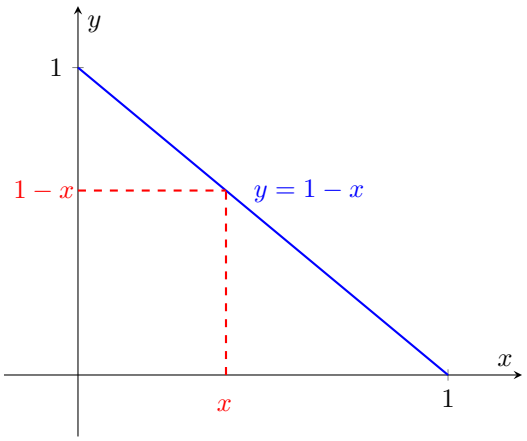
$$f_{x,y}(x, y) = f_x(x) f_y(y) \quad \text{if and only if X and Y are independent}$$

We already know:

- $f_{x,y}(x, y) = 2$ for points in R
- $f_y(y) = 2(1 - y)$ for $0 \leq y \leq 1$

Let's find $f_x(x)$:

$$\begin{aligned} f_x(x) &= \int_{-\infty}^{\infty} f_{x,y}(x,y)\,dy \\ &= \int_0^{1-x} 2\,dy \\ &= 2(1-x) \quad \text{for } 0 \le x \le 1 \end{aligned}$$



Conclusion: X and Y are not independent because:

$$f_{x,y}(x,y) \neq f_x(x) f_y(y)$$

$$2 \neq 2(1-x)2(1-y)$$

If we write the general expression for the conditional probability density function:

$$f_{x|y}(x|y) = \frac{f_{x,y}(x,y)}{f_y(y)} = \frac{2}{2(1-y)} = \frac{1}{1-y}$$

for $0 \le x \le 1-y$ and $0 \le y \le 1$

For the double integral:

$$-\int_1^2 \int_0^{1-y} (x^2 y + y x^3)\,dx\,dy$$

You can continue this process to find the final answer but here are the first few steps.

$$= -\int_1^2 \int_0^{1-y} (x^2 y + y x^3)\,dx\,dy$$

$$= -\int_1^2 \left[\frac{x^3 y}{3} + \frac{y x^4}{4}\right]_0^{1-y} dy$$

$$= -\int_1^2 \left[\frac{y(1-y)^3}{3} + \frac{y(1-y)^4}{4}\right] dy$$

Expected Value of a Function of Two Random Variables:

$$\mathbb{E}[g(X,Y)] = \begin{cases} \displaystyle\sum_x \sum_y g(x,y) p_{x,y}(x,y) & \text{if } X \text{ and } Y \text{ are discrete} \\ \displaystyle\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x,y) f_{x,y}(x,y)\, dx\, dy & \text{if } X \text{ and } Y \text{ are continuous} \end{cases}$$

where $p_{x,y}(x,y)$ is the joint probability mass function for discrete random variables and $f_{x,y}(x,y)$ is the joint probability density function for continuous random variables.

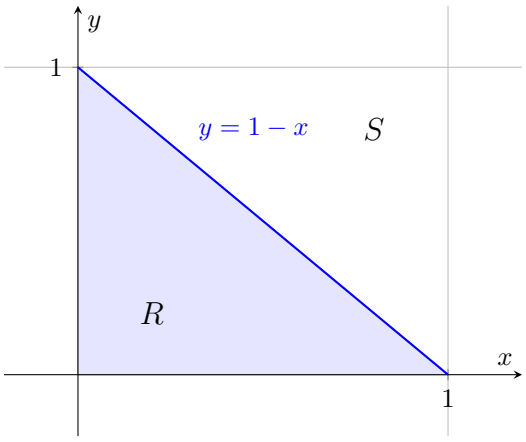
Example: Let $f_{X,Y}(x,y)$ be the joint probability density function given by:

$$f_{X,Y}(x,y) = \begin{cases} 2 & \text{if } (x,y) \in R \\ 0 & \text{otherwise} \end{cases}$$

where R is the region shown in the figure below.

Find:

1. $\mathbb{E}(XY)$
2. $\mathbb{E}(X^2Y)$



$$\begin{aligned} \mathbb{E}(XY) &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xy f_{X,Y}(x,y)\, dx\, dy \\ &= \int_0^1 \int_0^{1-y} xy \cdot 2\, dx\, dy \\ &= \int_0^1 \left[x^2 y\right]_0^{1-y}\, dy \\ &= \int_0^1 y(1-y)^2\, dy \end{aligned}$$

$$= \int_0^1 (y - 2y^2 + y^3)\, dy$$

$$= \left[\frac{y^2}{2} - \frac{2y^3}{3} + \frac{y^4}{4}\right]_0^1$$

$$= \left(\frac{1}{2} - \frac{2}{3} + \frac{1}{4}\right) - (0) = \frac{12 - 16 + 6}{24} = \frac{2}{24} = \boxed{\frac{1}{12}}$$

$$\mathbb{E}(X^2Y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x^2 y f_{X,Y}(x, y)\, dx\, dy$$

$$= \int_0^1 \int_0^{1-y} x^2 y \cdot 2\, dx\, dy$$

$$= \int_0^1 \left[\frac{2x^3 y}{3}\right]_0^{1-y} dy$$

$$= \int_0^1 \frac{2y(1-y)^3}{3}\, dy$$

$$= \frac{2}{3} \int_0^1 (y - 3y^2 + 3y^3 - y^4)\, dy$$

$$= \frac{2}{3} \left[\frac{y^2}{2} - y^3 + \frac{3y^4}{4} - \frac{y^5}{5}\right]_0^1$$

$$= \frac{2}{3} \left(\frac{1}{2} - 1 + \frac{3}{4} - \frac{1}{5}\right)$$

$$= \frac{2}{3} \cdot \frac{10 - 20 + 15 - 4}{20} = \frac{2}{3} \cdot \frac{1}{20} = \boxed{\frac{1}{30}}$$

Lecture 21: Tuesday 11/5/2024

Expected Values for Functions of Random Vectors

$$\mathbb{E}[g(x_1, x_2, \ldots, x_n)] = \begin{cases} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} g(\underline{x}) f_{\underline{x}}(\underline{x})\, dx_1 \cdots dx_n & \text{for continuous } \underline{x} \text{ with joint pdf } f_{\underline{x}}(\underline{x}) \\ \sum_{x_1} \cdots \sum_{x_n} g(\underline{x}) p_{\underline{x}}(\underline{x}) & \text{for discrete } \underline{x} \text{ with joint pmf } p_{\underline{x}}(\underline{x}) \end{cases}$$

Consider the example where we discussed an experiment that involves tossing a fair coin followed by rolling a fair six or four sided die depending on the outcome of the coin toss.

Let X and Y be defined as in this example

Compute $\mathbb{E}(X)$ and $\mathbb{E}(Y)$

$$P_{X,Y}(x,y) = \begin{cases} \frac{1}{12} & \text{if } x = 1, y \in \{1, 2, \ldots, 6\} \\ \frac{1}{8} & \text{if } x = 0, y \in \{1, 2, \ldots, 4\} \\ 0 & \text{otherwise} \end{cases}$$

$$\mathbb{E}(Y) = \sum_y \sum_x y P_{X,Y}(x,y)$$

$$= \sum_y \left(y P_{X,Y}(1,y) + y P_{X,Y}(0,y) \right)$$

$$= \sum_y y P_{X,Y}(0,y) + \sum_y y P_{X,Y}(1,y)$$

$$= \sum_{y=1}^{4} y \cdot \frac{1}{8} + \sum_{y=1}^{6} y \cdot \frac{1}{12}$$

$$= \frac{1}{8}(1+2+3+4) + \frac{1}{12}(1+2+3+4+5+6)$$

$$= \frac{10}{8} + \frac{21}{12}$$

$$= \frac{10}{8} + \frac{14}{8}$$

$$= \frac{24}{8} = \boxed{3}$$

$$\mathbb{E}(X) = \sum_x \sum_y x P_{X,Y}(x,y) = \sum_y y P_{x,y}(1,y). = \frac{1}{12}(1+2+3+4+5+6)$$

$$\mathbb{E}[g(x)] = \sum_x g(x) P_X(x)$$

Suppose X, Y are jointly continuous with joint pdf $f_{X,Y}(x,y)$. Then:

$$\mathbb{E}[aX + bY] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(ax + by)f_{X,Y}(x,y)\,dx\,dy$$

$$= a\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}xf_{X,Y}(x,y)\,dx\,dy + b\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}yf_{X,Y}(x,y)\,dx\,dy$$

$$= a\mathbb{E}[X] + b\mathbb{E}[Y]$$

Note:

$$\mathbb{E}[X] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}xf_{X,Y}(x,y)\,dx\,dy$$

$$\mathbb{E}[X] = \int_{-\infty}^{\infty}xf_X(x)\,dx$$

Verify that these are the same.
If X and Y are independent, then:

$$f_{X,Y}(x,y) = f_X(x)f_Y(y)$$

$$\mathbb{E}[XY] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}xyf_{X,Y}(x,y)\,dx\,dy$$

$$= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}xyf_X(x)f_Y(y)\,dx\,dy$$

$$= \left(\int_{-\infty}^{\infty}xf_X(x)\,dx\right)\left(\int_{-\infty}^{\infty}yf_Y(y)\,dy\right)$$

$$= \mathbb{E}[X]\mathbb{E}[Y]$$

$$\mathbb{E}[XY] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}xyf_{X,Y}(x,y)\,dx\,dy$$

Key Fact: For jointly continuous random variables X and Y, they are independent if and only if their joint probability density function can be written as the product of their marginal densities:

$$f_{X,Y}(x,y) = f_X(x)f_Y(y)$$

$$= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}xyf_X(x)f_Y(y)\,dx\,dy$$

$$= \left(\int_{-\infty}^{\infty}xf_X(x)\,dx\right)\left(\int_{-\infty}^{\infty}yf_Y(y)\,dy\right)$$

$$= \mathbb{E}[X]\mathbb{E}[Y]$$

Expected Value of Functions

Why do we care about $\mathbb{E}[g(X)]$ and $\mathbb{E}[g(X,Y)]$?

These quantities are fundamental in statistics because:

- They appear in key theorems like the Strong Law of Large Numbers (SLLN) and Weak Law of Large Numbers (WLLN)
- They allow us to work with weighted sums and transformations of random variables
- Most importantly, they model real statistical scenarios

In practice, a sample is modeled as a collection of random variables (a random vector), and statistics are modeled as functions defined on this sample. For example:

The sample mean can be written as:

$$\bar{X} = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

where X_i represents the ith observation in the sample.

Example: Consider the probability $P\left(\frac{X_1+X_2+\cdots+X_n}{n} > 165\right)$

To understand statistics thoroughly, we need to understand functions of random variables. We have taken the first step by learning how to compute $\mathbb{E}[g(X,Y)]$.

This naturally leads to the question: Can we compute probability mass functions (PMF) or probability density functions (PDF) for functions of random variables?

Example Problem: Consider an experiment consisting of:

- Tossing a fair coin
- Rolling a fair die

Let X represent the coin toss outcome (0 for heads, 1 for tails) and Y represent the die roll outcome.

What is the PMF of $X + Y$?

Note that:

- $X + Y$ is a discrete random variable (takes finitely many values)
- X and Y are independent, so $P_{X,Y}(x,y) = P_X(x)P_Y(y)$

The joint probability distribution $P_{X,Y}(x,y)$ can be represented in the following table:

$X\backslash Y$	1	2	3	4	5	6
0	$p_{X,Y}(0,1) = \frac{1}{12}$	$p_{X,Y}(0,2) = \frac{1}{12}$	$p_{X,Y}(0,3) = \frac{1}{12}$	$p_{X,Y}(0,4) = \frac{1}{12}$	$p_{X,Y}(0,5) = \frac{1}{12}$	$p_{X,Y}(0,6) = \frac{1}{12}$
1	$p_{X,Y}(1,1) = \frac{1}{12}$	$p_{X,Y}(1,2) = \frac{1}{12}$	$p_{X,Y}(1,3) = \frac{1}{12}$	$p_{X,Y}(1,4) = \frac{1}{12}$	$p_{X,Y}(1,5) = \frac{1}{12}$	$p_{X,Y}(1,6) = \frac{1}{12}$

Let $Z = X + Y$. The probability mass function of Z is:

$$P_Z(z) = \begin{cases} \frac{1}{12} & \text{if } z = 1 \\ \frac{1}{6} & \text{if } z = 2,3,4,5,6 \\ \frac{1}{12} & \text{if } z = 7 \\ 0 & \text{otherwise} \end{cases}$$

To verify this, let's compute $P_Z(1)$ as an example:

$$\begin{aligned} P_Z(1) &= P(Z = 1) \\ &= P(X = 0, Y = 1) \\ &= \frac{1}{12} \end{aligned}$$

To verify this, let's compute $P_Z(2)$ as an example:

$$\begin{aligned} P_Z(2) &= P(Z = 2) \\ &= P(X = 0, Y = 2) + P(X = 1, Y = 1) \\ &= \frac{1}{12} + \frac{1}{12} = \frac{2}{12} = \frac{1}{6} \end{aligned}$$

Key Features:

- X and Y are independent random variables
- We are working with their sum $Z = X + Y$

Important Facts:

1. **Linearity of Expectation:**

$$E\left(\sum_{i=1}^{n} a_i X_i\right) = \sum_{i=1}^{n} a_i E(X_i)$$

2. **Variance of Independent Sum:**

$$V\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} a_i^2 V(X_i) \quad \text{when } X_i\text{'s are independent}$$

Lecture 22: Thursday 11/7/2024

Random Vectors

Many situations require keeping track of multiple quantities simultaneously. For example:

- Analyzing pictures (one random variable per pixel)
- Tossing a coin twice:
  - X_1 - first coin toss
  - X_2 - second coin toss

We have learned how to do computations related to probabilities of random vectors.

Types of Random Vectors

1. **Discrete Random Vectors**
   - Joint PMF: quantities of interest are computed via sums
2. **Jointly Continuous Random Variables**
   - Joint PDF: quantities of interest are computed via integrals

We will consider functions of random vectors. Samples (before observations are made) and statistics can be modeled by random vectors and functions of random vectors.

Random Samples and Statistics

- **Random Sample:** $X_1, \ldots, X_n$ with special properties where X_i reflects the ith observation
- **Statistics:** A function defined on the sample
  - Example: Sample average $= \frac{X_1+\cdots+X_n}{n}$

Important Note: Statistics are random variables too! This should be understood in the context discussed above.

Summary of What We've Learned

We have learned how to compute averages for functions of random vectors:

- $E(XY)$, $E(X)$, $E(Y)$, $E(X^2Y)$, $E(X^3Y)$
- $E(\sum_{i=1}^n a_i X_i) = \sum_{i=1}^n a_i E(X_i)$
- If X and Y are independent, then $E(XY) = E(X)E(Y)$

Computing PMF for Functions of Random Vectors

Consider X_1 and X_2 as independent identically distributed discrete random variables with PMF $p_{X_1}(x)$ and $p_{X_2}(x)$, where:

$$P_X(x) = \begin{cases} 0.3, & x = 0 \\ 0.35, & x = 1 \\ 0.25, & x = 2 \\ 0.1, & x = 3 \end{cases}$$

Problem: Compute the PMF of $X_1 + X_2$
Solution: Since X_1 and X_2 are independent:

$$P_{X_1,X_2}(x_1, x_2) = P_{X_1}(x_1)P_{X_2}(x_2)$$

Let's compute the joint PMF using a table:

$X_1 \backslash X_2$	0	1	2	3
0	0.09	0.105	0.075	0.03
1	0.105	0.1225	0.0875	0.035
2	0.075	0.0875	0.0625	0.025
3	0.03	0.035	0.025	0.01

Then $P_{X_1+X_2}(k) = \sum_{x_1+x_2=k} P_{X_1,X_2}(x_1, x_2)$

$$P_{X_1+X_2}(x) = \begin{cases} 0.09, & x = 0 \\ 0.21, & x = 1 \\ 0.285, & x = 2 \\ 0.24, & x = 3 \\ 0.125, & x = 4 \\ 0.05, & x = 5 \\ 0.01, & x = 6 \end{cases}$$

For example:

$$\begin{aligned} P_{X_1+X_2}(0) &= P_{X_1,X_2}(0,0) = 0.09 \\ P_{X_1+X_2}(1) &= P_{X_1,X_2}(0,1) + P_{X_1,X_2}(1,0) \\ &= 0.105 + 0.105 = 0.21 \end{aligned}$$

$$\begin{aligned} E(X_1 + X_2) &= \sum_{t=0}^{6} t P_{X_1+X_2}(t) \\ &= E(X_1) + E(X_2) \quad \text{(by linearity)} \\ &= 2E(X_1) \quad \text{(since } X_1, X_2 \text{ are identically distributed)} \end{aligned}$$

$$\begin{aligned} V(X_1 + X_2) &= E((X_1 + X_2 - E(X_1 + X_2))^2) \\ &= E((X_1 - E(X_1) + (X_2 - E(X_2)))^2) \\ &= E((X_1 - E(X_1))^2) + E((X_2 - E(X_2))^2) + 2E((X_1 - E(X_1))(X_2 - E(X_2))) \\ &= V(X_1) + V(X_2) + 2\text{Cov}(X_1, X_2) \end{aligned}$$

Covariance

Definition: For random variables X_1 and X_2, their covariance is:

$$\text{Cov}(X_1, X_2) = E(X_1 X_2) - E(X_1)E(X_2)$$

Important Property: If X_1 and X_2 are independent, then:

$$E(X_1 X_2) = E(X_1)E(X_2) \quad \text{and} \quad \text{Cov}(X_1, X_2) = 0$$

Properties of IID Random Variables:

Let $X_1, X_2, \ldots, X_n$ be independent identically distributed random variables with mean μ and variance σ^2. Then:

1. Mean of sum:
$$E(X_1 + X_2 + \cdots + X_n) = n\mu$$

2. Variance of sum:
$$V(X_1 + X_2 + \cdots + X_n) = n\sigma^2$$

3. Mean of sample mean:
$$E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \mu$$

4. Variance of sample mean:
$$V\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{\sigma^2}{n}$$

Proof of Variance Properties:

For independent random variables, the variance of their sum equals the sum of their variances:

$$\begin{aligned} V\left(\sum_{i=1}^{n} X_i\right) &= V(X_1) + V\left(\sum_{i=2}^{n} X_i\right) \\ &= \sigma^2 + V\left(\sum_{i=2}^{n} X_i\right) \end{aligned}$$

By induction, this gives us:

$$V\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} V(X_i) = n\sigma^2$$

For the variance of the sample mean, we use the property that $V(aX) = a^2V(X)$:

$$V\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}V\left(\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2} \cdot n\sigma^2 = \frac{\sigma^2}{n}$$

Central Limit Theorem:

Suppose that $X_1, \ldots, X_n$ are IID random variables with $E(X_1) = \mu$ and $V(X_1) = \sigma^2$.
Let $S_n = \sum_{i=1}^{n} X_i$ and $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$.
Then as $n \to \infty$:

$$P\left(a < \frac{S_n - n\mu}{\sigma\sqrt{n}} < b\right) = \Phi(b) - \Phi(a)$$

for $-\infty \leq a \leq b \leq \infty$, where Φ is the standard normal CDF.

Therefore, for large n:

$$P\left(a < \frac{S_n - n\mu}{\sigma\sqrt{n}} < b\right) \approx \Phi(b) - \Phi(a)$$

$$P\left(a < \frac{S_n - n\mu}{\sigma\sqrt{n}} < b\right) = P\left(a < \frac{\frac{1}{n}S_n - \mu}{\frac{\sqrt{n}}{n}\sigma} < b\right)$$

$$= P\left(a < \frac{\bar{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}} < b\right)$$

Note: A common rule of thumb is that $n > 30$ is generally considered "large enough" for the Central Limit Theorem to apply.

Example: Suppose that a fair die is rolled 100 times. What is the approximate probability that the sum is between 340 and 360?

Let X_i represent the outcome of the ith roll, where X_i are IID random variables.
Let $S_{100} = \sum_{i=1}^{100} X_i$ be the sum of all rolls.
We want to find: $P(340 \leq S_{100} \leq 360)$
For a fair die:

$$E(X_i) = 3.5$$
$$V(X_i) = 2.92$$

By the Central Limit Theorem:

$$S_{100} \sim N(350, 100 \cdot 2.92)$$

Therefore:

$$P(340 \leq S_{100} \leq 360) = P\left(\frac{340 - 350}{\sqrt{292}} \leq Z \leq \frac{360 - 350}{\sqrt{292}}\right)$$

Lecture 23: Tuesday 11/12/2024

(went over study guide for exam 2 − I don't see a reason to put this here, but u can find the answers on canvas)

Lecture 24: Thursday 11/14/2024

(Exam 2)

Lecture 25: Tuesday 11/19/2024

Inferential Statistics

Probability vs. Inferential Statistics

- **Probability:** Given that data comes from a particular distribution, how likely is a particular collection of outcomes?
- **Inferential Statistics:** Given observed data, what can we infer about the distribution or process that generated it?

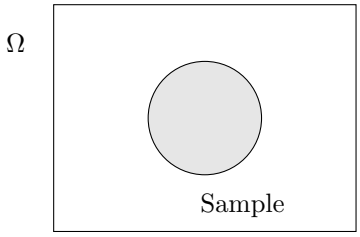
Examples

- A coin is flipped multiple times. Is it fair?
- A company claims their batteries last 5 years. Is this claim reasonable based on sample data?

Key Concepts

Sample A subset drawn from the population (the entire collection under study)

Statistic A numerical summary calculated from the sample data



A sample provides information about the underlying population through a subset of observations.

When we move from studying a population to analyzing a sample, we introduce uncertainty into our conclusions. This uncertainty is what allows us to apply probability theory to statistical inference.

Definition: A *random sample* of size n is a collection $x_1, x_2, \ldots, x_n$ of independent, identically distributed random variables (i.i.d. RVs).

Note: The term "sample" has multiple meanings:

- In everyday usage, it refers to a collection of actual observations
- In statistics, it can refer to either the random variables or their observed values
- When we need to be precise, we use "realization of a sample" or "realized sample" to refer specifically to the observed values

Definition: A *statistic* is any function calculated from the sample data.
Example: Let $X_1, \ldots, X_n$ be a sample of size n.
Sample Mean: $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i = T(X_1, \ldots, X_n)$
Sample Variance: $S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2 = U(X_1, \ldots, X_n)$
General Statistic: For any function V,

$$V(X_1, \ldots, X_n) = f(X_1, \ldots, X_n)$$

Important Notes:

1. A statistic is a random variable and thus has its own probability distribution.
2. The Central Limit Theorem describes the approximate distribution of $\bar{X}_n$ for large n.
3. For precise definitions:
   - A sample must consist of observable quantities
   - A statistic must be calculable from the sample data

Observation: Based on experience, we know that $\bar{X}$ and S^2 are useful statistics, while general functions $V(X_1, \ldots, X_n)$ may be less meaningful.
Question: What makes a statistic "good"? To answer this, we need to establish some terminology about estimation.

Point Estimators

Let θ be some parameter associated with the population. This parameter is fixed but unknown. Our goal is to use the data (realized sample/observations) to determine the value of θ.
Definition: An *estimator* for θ, usually denoted by $\hat{\theta}$, is a statistic used to approximate θ.
This naturally leads to the question:

What makes $\hat{\theta}$ a good estimator for θ?

Example:

- $\bar{X}$ is a good estimator for μ but not for σ^2
- S^2 is a good estimator for σ^2 but not for μ

Definition: An estimator $\hat{\theta}$ for θ is said to be *unbiased* if $E(\hat{\theta}) = \theta$ for all possible values of θ.
Example 1: $\bar{X}$ is an unbiased estimator for μ

$$\begin{aligned} E(\bar{X}) &= E\left(\frac{X_1 + X_2 + \cdots + X_n}{n}\right) \\ &= \frac{1}{n}[E(X_1) + E(X_2) + \cdots + E(X_n)] \\ &= \frac{1}{n}[n\mu] = \mu \end{aligned}$$

Example 2: S^2 is an unbiased estimator for σ^2

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

Example 3: $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$ is a biased estimator of σ^2
Definition: The bias of an estimator $\hat{\theta}$ for θ is given by $E(\hat{\theta}) - \theta$

- An estimator is unbiased if and only if its bias equals 0

Computing the bias of $\hat{\sigma}^2$ as an estimator for σ^2:
We want to find $E(\hat{\sigma}^2) - \sigma^2$
Note that $\hat{\sigma}^2 = \frac{n-1}{n}S^2$
Therefore:

$$\begin{aligned} E(\hat{\sigma}^2) &= E\left(\frac{n-1}{n}S^2\right) \\ &= \frac{n-1}{n}E(S^2) \\ &= \frac{n-1}{n}\sigma^2 \end{aligned}$$

Thus, the bias is:

$$\begin{aligned} \text{bias}(\hat{\sigma}^2) &= \frac{n-1}{n}\sigma^2 - \sigma^2 \\ &= -\frac{\sigma^2}{n} \end{aligned}$$

Definition: A sequence of estimators $\{\hat{\theta}_n\}$ for θ is said to be consistent if for every $\epsilon > 0$:

$$\lim_{n\to\infty} P(|\hat{\theta}_n - \theta| > \epsilon) = 0$$

Lecture 26: Thursday 11/21/2024

Review of Inferential Statistics

Inferential statistics allows us to make meaningful conclusions about a population based on a sample.

Key Concepts Covered:

- Defined random samples and statistics in terms of random variables
- Explored making inferences about μ based on $\bar{X}$
- Studied making inferences about σ^2 based on S^2

Point Estimation

Point estimation involves finding our best guess for a population parameter θ (e.g., $\theta = \mu$ or $\theta = \sigma^2$) given a sample.

Point estimation is one of three main components of statistical inference, alongside:

- Confidence intervals
- Hypothesis testing

Properties of Estimators

An estimator $\hat{\theta}$ is unbiased if for all possible values of θ:

$$E(\hat{\theta}) = \theta$$

Examples:

- $\bar{X}$ is an unbiased estimator for μ
- S^2 is an unbiased estimator for σ^2
- $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$ is a biased estimator for σ^2

Definition: The bias of an estimator $\hat{\theta}$ is:

$$\text{bias}(\hat{\theta}) = E(\hat{\theta} - \theta) = E(\hat{\theta}) - \theta$$

Example: Let's find the bias of $\hat{\sigma}^2$:

$$\begin{aligned}
\text{bias}(\hat{\sigma}^2) &= E(\hat{\sigma}^2 - \sigma^2) = E(\hat{\sigma}^2) - \sigma^2 \\
&= E(\frac{n-1}{n}S^2) - \sigma^2 \\
&= \frac{n-1}{n}E(S^2) - \sigma^2 \\
&= \frac{n-1}{n}\sigma^2 - \sigma^2 \\
&= -\frac{\sigma^2}{n}
\end{aligned}$$

Consistency (Applies to a sequence of estimators)

Definition: A sequence of estimators $\{\hat{\theta}_n\}_{n\in\mathbb{N}}$ is said to be consistent if for all $\epsilon > 0$:

$$\lim_{n\to\infty} P(|\hat{\theta}_n - \theta| > \epsilon) = 0$$

Example: The sequence $(\bar{X}_n)_{n\in\mathbb{N}}$ is consistent for μ. This is known as the Weak Law of Large Numbers.

Exercise: Suppose you have three estimators for μ:

- X_1
- $\bar{X}_n$

- $\bar{X}_n^* = \frac{1}{n-2}\sum_{i=2}^{n-2} X_i$

Which estimator would you use if the sample $X_1, \ldots, X_n$ has $E(X_i) = \mu$ and $\text{Var}(X_i) = \sigma^2$?

Solution: Let's analyze each estimator's expected value:

- $E(X_1) = \mu$
- $E(\bar{X}_n) = \mu$
- $E(\bar{X}_n^*) = E\left(\frac{1}{n-2}\sum_{i=2}^{n-2} X_i\right) = \frac{1}{n-2}\sum_{i=2}^{n-2} E(X_i) = \frac{1}{n-2}\sum_{i=2}^{n-2} \mu = \mu$

So X_1, $\bar{X}_n$, and $\bar{X}_n^*$ are all unbiased estimators for μ.

Unbiasedness by itself is not good enough to give us the best estimator, even if we agree that unbiased estimators are preferable, as we do in this class.

Analyzing Consistency:

- X_1 (where $\hat{\theta} = X_1$ for all n) is not consistent
- $\bar{X}_n$ is consistent - this follows from the Weak Law of Large Numbers (WLLN)
- $\bar{X}_n^*$ is consistent

Definition: Let $\hat{\theta}$ be an estimator. The Mean Square Error (MSE) of $\hat{\theta}$ is:

$$MSE(\hat{\theta}) = E((\hat{\theta} - \theta)^2)$$

This represents the average squared deviation from θ.

Comparing Estimators: Given two estimators $\hat{\theta}_1$ and $\hat{\theta}_2$ for θ, we say $\hat{\theta}_1$ is better than $\hat{\theta}_2$ if:

$$MSE(\hat{\theta}_1) < MSE(\hat{\theta}_2)$$

Important Facts:

- $MSE(\hat{\theta}) = (bias(\hat{\theta}))^2 + Var(\hat{\theta})$
- For unbiased estimators, $MSE(\hat{\theta}) = Var(\hat{\theta})$ since $bias(\hat{\theta}) = 0$

When choosing between two unbiased estimators, we choose the one with lower variance (this is known as the Minimum Variance Unbiased Estimator principle, or MVUE).

Variance Calculations: For $\bar{X}_n$:

$$\begin{aligned} Var(\bar{X}_n) &= Var\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) \\ &= \frac{1}{n^2} Var\left(\sum_{i=1}^{n} X_i\right) \\ &= \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i) \text{ (since } X_i\text{'s are independent)} \\ &= \frac{\sigma^2}{n} \end{aligned}$$

Similarly, for $\bar{X}_n^*$:

$$Var(\bar{X}_n^*) = \frac{\sigma^2}{n-2}$$

Since $Var(\bar{X}_n) < Var(\bar{X}_n^*)$, and both estimators are unbiased, the MVUE principle tells us that $\bar{X}_n$ is the better estimator for μ.

Example: Consider a coin that comes up heads with unknown probability p. How can we construct an estimator $\hat{p}$ for p?

Solution:

- Toss the coin n times
- Let $\hat{p} = \frac{\text{number of heads}}{n}$

General Question: How do we construct estimators when we are not familiar with the parameter or there is no intuitive answer?

The key is to construct estimators with desirable statistical properties such as:

- Maximum likelihood estimators (MLE)
- Method of moments estimators (MME)

Let $X_1, \ldots, X_n$ be a random sample of size n with $X_i \sim \text{Ber}(p)$, where $0 < p < 1$ is fixed but unknown.

The sample proportion $\hat{p} = \frac{1}{n}\sum_{i=1}^n X_i$ is an estimator for p, noting that $E(X_1) = p$.

We can derive this same estimator using the maximum likelihood method:

- Maximum likelihood works by finding the parameter value that maximizes the probability of observing the given sample
- Let $f(x;\theta)$ be the pdf/pmf for the population distribution
- The likelihood function is defined as:

$$L(\theta) = \prod_{i=1}^n f(X_i;\theta)$$

- For a pmf, the likelihood represents the joint probability:

$$P(X_1 = x_1, \ldots, X_n = x_n) = \prod_{i=1}^n P(X_i = x_i;\theta)$$

Therefore, the likelihood function represents the joint probability mass/density viewed as a function of the parameter θ.

Lecture 27: Tuesday 11/26/2024

Maximum Likelihood Estimators (MLE)

The goal is to find the parameter value θ that maximizes the likelihood of observing our sample data.

Definition: The likelihood function is given by:

$$L(\theta) = \prod_{i=1}^{n} f(X_i; \theta)$$

where $f(X_i; \theta)$ is the joint pmf/pdf.

Steps to find MLE:

1. Calculate the likelihood function $L(\theta)$
2. Find the value θ_0 such that $L(\theta_0) \geq L(\theta)$ for all possible values of θ

Example: Let $X_1, \ldots, X_n$ be a random sample where $X_i \sim \text{Ber}(p)$ with unknown fixed p. The pmf is given by:

$$f(x_i; p) = \begin{cases} p^x(1-p)^{1-x} & \text{if } x = 0, 1 \\ 0 & \text{otherwise} \end{cases}$$

Therefore, the likelihood function is:

$$\begin{aligned} L(p) &= \prod_{i=1}^{n} f(x_i; p) \\ &= \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} \end{aligned}$$

To maximize, we find critical points using calculus:

Let $l(p) = \ln(L(p))$ be the log-likelihood function.

$$\begin{aligned} l(p) &= \ln\left(\prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i}\right) \\ &= \sum_{i=1}^{n} \ln(p^{x_i}(1-p)^{1-x_i}) \\ &= \sum_{i=1}^{n} \left(x_i \ln(p) + (1-x_i)\ln(1-p)\right) \end{aligned}$$

Taking the derivative:

$$\begin{aligned} l'(p) &= \sum_{i=1}^{n} \left(\frac{x_i}{p} - \frac{1-x_i}{1-p} \right) \\ &= \sum_{i=1}^{n} \frac{x_i - x_i p - p + x_i p}{p(1-p)} \\ &= \frac{1}{p(1-p)} \sum_{i=1}^{n} (x_i - p) \\ &= \frac{1}{p(1-p)} \left(\sum_{i=1}^{n} x_i - np \right) \end{aligned}$$

Setting $l'(p) = 0$:

$$\frac{1}{p(1-p)} \left(\sum_{i=1}^{n} x_i - np \right) = 0$$

Solving for p:

$$p = \frac{\sum_{i=1}^{n} x_i}{n}$$

Therefore, the maximum likelihood estimator is:

$$\hat{p}_{MLE} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Example Application: Maximum Likelihood Estimation for Bernoulli Trials

We have already shown that:

$$\hat{p}_{MLE} = \frac{\sum_{i=1}^{n} x_i}{n}$$

In practice, this means:

- Toss the coin n times
- Count the number of heads
- Calculate $\hat{p}_{MLE} = \frac{\text{number of heads}}{n}$

Specific Example: Let X_1, X_2, X_3 be a sample of size 3 where $X_i \sim \text{Ber}(p)$ with unknown p. Then:

$$\hat{p}_{MLE} = \frac{\sum_{i=1}^{3} x_i}{3}$$

For the realization $x_1 = 0, x_2 = 1, x_3 = 1$:

$$\hat{p}_{MLE} = \frac{0+1+1}{3} = \frac{2}{3}$$

Example: Let $f(x;\theta)$ be the exponential distribution with parameter θ:

$$f(x;\theta) = \begin{cases} \theta e^{-\theta x} & \text{if } x \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

Let $\theta > 0$ be fixed but unknown. Given a random sample $X_1, \ldots, X_n$ of size n where each X_i has pdf $f(x;\theta)$, find $\hat{\theta}_{MLE}$.

Solution:

$$\begin{aligned} L(\theta) &= \prod_{i=1}^{n} f(x_i;\theta) \\ &= \prod_{i=1}^{n} \theta e^{-\theta x_i} \end{aligned}$$

Taking the natural logarithm:

$$\begin{aligned} l(\theta) &= \ln(L(\theta)) \\ &= \ln\left(\prod_{i=1}^{n} \theta e^{-\theta x_i}\right) \\ &= \sum_{i=1}^{n} \ln(\theta e^{-\theta x_i}) \\ &= \sum_{i=1}^{n} (\ln(\theta) - \theta x_i) \\ &= n\ln(\theta) - \theta \sum_{i=1}^{n} x_i \end{aligned}$$

Taking the derivative:

$$l'(\theta) = \frac{n}{\theta} - \sum_{i=1}^{n} x_i$$

Setting $l'(\theta) = 0$ and solving:

$$\begin{aligned} \frac{n}{\theta} - \sum_{i=1}^{n} x_i &= 0 \\ \theta &= \frac{n}{\sum_{i=1}^{n} x_i} \end{aligned}$$

Therefore:

$$\hat{\theta}_{MLE} = \frac{n}{\sum_{i=1}^{n} x_i}$$

To verify this is a maximum:

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

Since $l''(\theta) < 0$ for all $\theta > 0$, by the second derivative test, $l(\theta)$ achieves a maximum at $\hat{\theta}_{MLE}$.

Multiple Parameters:

What happens if we have two or more parameters? Let's look at an example with two parameters.

Example: Let $X_1, \ldots, X_n$ be a random sample where $X_i \sim \mathcal{N}(\mu, \sigma^2)$ with both μ and σ^2 unknown.

Can we use MLE? Yes! The process is similar, but we'll need to take partial derivatives. The pdf is:

$$f(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

The likelihood function:

$$L(\mu,\sigma) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$$

Taking the natural logarithm:

$$\begin{aligned} l(\mu,\sigma) &= \ln\left(\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}\right) \\ &= \sum_{i=1}^{n} \ln\left(\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}\right) \\ &= \sum_{i=1}^{n} \left(-\ln(\sqrt{2\pi}) - \ln(\sigma) - \frac{(x_i-\mu)^2}{2\sigma^2}\right) \\ &= -\frac{n}{2}\ln(2\pi) - n\ln(\sigma) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2 \end{aligned}$$

To find the maximum likelihood estimators, we need to find the critical points by setting both partial derivatives of $f(x,y) = x^2 + y^2$ equal to zero:

$$\frac{\partial l}{\partial \mu} = 0 \quad \text{and} \quad \frac{\partial l}{\partial \sigma} = 0$$

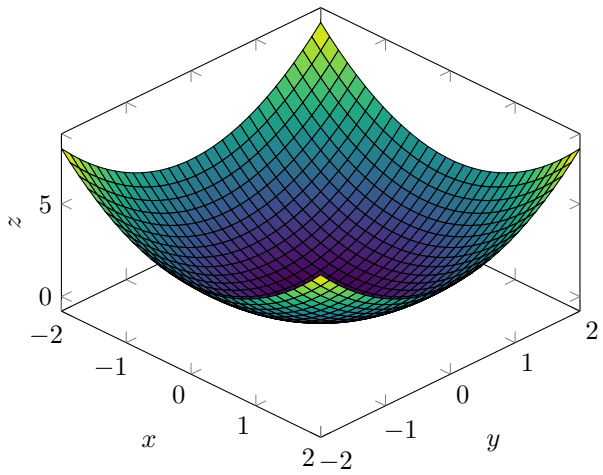


Figure 15: 3D surface plot of $f(x,y) = x^2 + y^2$

Taking the partial derivatives:

$$\begin{aligned}\frac{\partial l}{\partial \mu} &= \frac{\partial}{\partial \mu}\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right)\\ &= -\frac{1}{2\sigma^2}\sum_{i=1}^{n}\frac{\partial}{\partial \mu}(x_i-\mu)^2\\ &= -\frac{1}{2\sigma^2}\sum_{i=1}^{n}2(x_i-\mu)(-1)\\ &= \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu)\end{aligned}$$

Setting $\frac{\partial l}{\partial \mu} = 0$:

$$\begin{aligned}\frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu) &= 0\\ \sum_{i=1}^{n}x_i - n\mu &= 0\\ \hat{\mu}_{MLE} &= \frac{1}{n}\sum_{i=1}^{n}x_i\end{aligned}$$

$$\frac{\partial l}{\partial \sigma} = -\frac{n}{\sigma} + \frac{1}{\sigma^3}\sum_{i=1}^{n}(x_i-\mu)^2$$

Setting $\frac{\partial l}{\partial \sigma} = 0$:

$$\begin{aligned}-\frac{n}{\sigma} + \frac{1}{\sigma^3}\sum_{i=1}^{n}(x_i-\mu)^2 &= 0\\ \sigma^2 &= \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2\\ \sigma^2 &= \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{X})^2\\ \hat{\sigma}_{MLE} &= \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{X})^2}\end{aligned}$$