Math 3070/6070 Introduction to Probability

Mon/Wed/Fri 11:00am - 11:50am

Instructor: Dr. Xiang Ji, xji4@tulane.edu

Lecture 1:Aug 22

Today

- Introduction
- Introduce yourself
- Course logistics

What is this course about?

This course will provide a calculus-based introduction to probability theory. Material covered will include fundamental axioms of probability, combinatorics, discrete and continuous random variables, multivariate distributions, expectation, and limit theorems, generally following Chapters 1-5 of the textbook. This course is a critical prerequisite for more advanced work in statistical theory and analysis.

Prerequisite

- Calculus

Why learn probability

- The subject of probability theory is the foundation upon which all of statistics is built.
- It provides you a tool to model
  - populations
  - experiments
  - almost anything else that could be considered a random phenomenon
  - example topics in Data Analysis course
- Through these models, statisticians are able to draw inferences about populations based on examination of only a part of the whole.
- A must have for any Data Scientists.

What this course WILL NOT do for you

It will not help you:

- Beat the casino at blackjack (although it may convince you that it is better not to gamble, or that a casino is a great business).
- Answer your friends' silly questions such as "What are the chances it will rain tomorrow?" (although it might make you think of ways that you might model and compute it).

Syllabus

Check course website frequently for updates and announcements.

https://tulane-math-3070-2022.github.io/

HW submission

Students are required to submit hand-written homework in recitations to the TA. Homework assignments are expected every two weeks with 4-5 problems at a time.

Presentations

Do we want to have a 5 bonus point towards the final grade with a presentation?

Last year comments

Not really, this is my first time teaching this course. There will be an internal mid-term-ish evaluation for this course. Will remember to go over them.

Lecture 2:Aug 24

Last time

- Introduction
- Introduce yourself
- Course logistics

Today

- Set theory (1.1)
- Axiomatic Foundations (1.2)

Set Theory

One of the main objectives of a statistician is to draw conclusions about a population of objects by conducting an experiment. The first step in this endeavor is to identify the possible outcomes or, in statistical terminology, the *sample space*.

Definition The set, S, of all possible outcomes of a particular experiment is called the *sample space* for the experiment.

Example The sample space of

- tossing a coin just once, contains two outcomes, heads and tails

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

- observing reported SAT scores of randomly selected students at a certain university

$$S = \{200, 210, 220, \dots, 780, 790, 800\}$$

- an experiment where the observation is reaction time to a certain stimulus

$$S = (0, \infty)$$

Definition An *event* is any collection of possible outcomes of an experiment, that is , any subset of S (including S itself).

Let A be an event,

- A is a subset of S,
- event A occurs if the outcome of the experiment is in the set A,
- we generally speak of the probability of an event, rather than a set.

Set operations:

- Containment:
$$A \subset B \iff x \in A \implies x \in B$$

- Equality:
$$A = B \iff A \subset B \text{ and } B \subset A$$

- Union: the union of A and B, written as $A \cup B$, is the set of elements that belong to either A or B or both
$$A \cup B = \{x : x \in A \text{ or } x \in B\}.$$

- Intersection: the intersection of A and B, written $A \cap B$, is the set of elements that belong to both A and B:
$$A \cap B = \{x : x \in A \text{ and } x \in B\}.$$

- Complementation: the complement of A, written A^c, is the set of all elements that are not in A:
$$A^c = \{x : x \notin A\}.$$

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

1. Commutativity
$$A \cup B = B \cup A,$$
$$A \cap B = B \cap A;$$

2. Associativity
$$A \cup (B \cup C) = (A \cup B) \cup C,$$
$$A \cap (B \cap C) = (A \cap B) \cap C;$$

3. Distributive Laws
$$A \cap (B \cup C) = (A \cap B) \cup (A \cap C),$$
$$A \cup (B \cap C) = (A \cup B) \cap (A \cup C);$$

4. DeMorgan's Laws
$$(A \cup B)^c = A^c \cap B^c,$$
$$(A \cap B)^c = A^c \cup B^c;$$

We show the proof of $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$ in the distributive laws. Caution: Venn diagrams are helpful in visualization, but they do not constitute a formal proof. To prove that two sets are equal, we need to show that each set contains the other.
proof:

- $A \cap (B \cup C) \subset (A \cap B) \cup (A \cap C)$:
  Let $x \in (A \cap (B \cup C))$. By definition of intersection, $x \in (B \cup C)$ that is, either $x \in B$ or $x \in C$. Since x also must be in A, we have that either $x \in (A \cap B)$ or $x \in (A \cap C)$; therefore, $x \in ((A \cap B) \cup (A \cap C))$.

- $(A \cap B) \cup (A \cap C) \subset A \cap (B \cup C)$:
  Let $x \in ((A \cap B) \cup (A \cap C))$. This implies that $x \in (A \cap B)$ or $x \in (A \cap C)$. If $x \in (A \cap B)$, then x is in both A and B. Since $x \in B$, then $x \in (B \cup C)$and thus $x \in (A \cap (B \cup C))$. It follows the same argument when $x \in (A \cap C)$, we still have $x \in (A \cap (B \cup C))$.

Definition Two events A and B are *disjoint* (or *mutually exclusive*) if $A \cap B = \varnothing$. The events $A_1, A_2, \ldots$ are *pairwise disjoint* (or *mutually exclusive*) if $A_i \cap A_j = \varnothing$ for all $i \neq j$.

Definition If $A_1, A_2, \ldots$ are pairwise disjoint and $\cup_{i=1}^{\infty} A_i = A_1 \cup A_2 \cup \cdots = S$, then the collection of $A_1, A_2, \ldots$ forms a *partition* of S.

Example The sets $A_i = [i, i+1), i = 0, 1, 2, \ldots$ form a partition of $[0, \infty)$.

Basics of Probability Theory

When an experiment is performed, the realization of the experiment is an outcome in the sample space. If the experiment is performed a number of times, then

- different outcomes may occur each time
- some outcomes may repeat
- the "frequency of occurrence" of an outcome can be thought of as a probability

However, we **do not** define probabilities in terms of frequencies but instead take the mathematically simpler axiomatic approach. The axiomatic approach is not concerned with the interpretations of probabilities, but is concerned only that the probabilities are defined by a function satisfying the axioms. Interpretations of the probabilities are quite another matter:

- The "frequency of occurrence" of an event is one example of a particular interpretation of probability.
- Another possible interpretation is a subjective one, where we can think of the probability as a belief in the chance of an event occurring.

Axiomatic Foundations

For each event A in the sample space S, we want to associate with A a number between zero and one that will be called the probability of A, denoted by $\Pr(A)$. The domain of Pr is the set where the arguments of the function $\Pr(\cdot)$ are defined. It is natural to define the domain of Pr as all subsets of S, that is for each $A \subset S$, we define $\Pr(A)$ as the probability that A occurs. However, there are some technical difficulties to overcome which requires us to familiarize with the following.

Definition A collection of subsets of S is called a *sigma algebra* (or *Borel field*), denoted by $\mathcal{B}$, if it satisfies the following three properties:

1. $\varnothing \in \mathcal{B}$ (the empty set is an element of $\mathcal{B}$).
2. If $A \in \mathcal{B}$, then $A^c \in \mathcal{B}$ ($\mathcal{B}$ is closed under complementation).
3. If $A_1, A_2, \cdots \in \mathcal{B}$, then $\cup_{i=1}^{\infty} A_i \in \mathcal{B}$ ($\mathcal{B}$ is closed under countable unions).

From Property (1) and (2), we see that the empty set and its complement S (since $S = \varnothing^c$) are always in a sigma algebra. In fact, they construct the *trivial* algebra $\{\varnothing, S\}$ which is the smallest sigma algebra.

By DeMorgan's Law, (3) can be replaced by:

$$3'. \text{ if } A_1, A_2, \cdots \in \mathcal{B}, \text{ then } \cap_{i=1}^{\infty} A_i \in \mathcal{B}.$$

This is because:

$$(\cup_{i=1}^{\infty} A_i^c)^c = \cap_{i=1}^{\infty} A_i.$$

Example If S is finite or countable (where the elements of S can be put into $1-1$ correspondence with a subset of the integers), then these technicalities really do not arise, for we we define for a given sample space S,

$$\mathcal{B} = \{\text{all subsets of } S, \text{ including } S \text{ itself}\}.$$

If S has n elements, there are 2^n sets in $\mathcal{B}$ (why?).[hint: for each element, it is either in or out of a subset, so 2 choices].

Example Let $S = (-\infty, \infty)$, the real line. Then $\mathcal{B}$ is chosen to contain all sets of the form

$$[a, b], (a, b], (a, b), \text{ and } [a, b)$$

for all real numbers a and b. Also, from the properties of $\mathcal{B}$, it follows that $\mathcal{B}$ contains all sets that can be formed by taking (possibly countably infinite) unions and intersections of sets of the above varieties.

We now define a probability function.

Definition Given a sample space S and an associated sigma algebra $\mathcal{B}$, a *probability function* is a function Pr with domain $\mathcal{B}$ that satisfies

1. $\Pr(A) \geqslant 0$ for all $A \in \mathcal{B}$.
2. $\Pr(S) = 1$.

3. If $A_1, A_2, \cdots \in \mathcal{B}$ are pairwise disjoint, then $\Pr(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \Pr(A_i)$.

The above three properties are usually referred to as the Axioms of Probability (or the Kolmogorov Axioms, after A. Kolmogorov, one of the fathers of probability theory). Any function that satisfies the Axioms of Probability is called a probability function.

Example Consider the simple experiment of tossing a fair coin (just once), so $S = \{H, T\}$. A reasonable probability function is the one that assigns equal probabilities to heads and tails, that is,

$$\Pr(\{H\}) = \Pr(\{T\}).$$

Since $S = \{H\} \cup \{T\}$, we have , from Axiom 1, $\Pr(\{H\} \cup \{T\}) = 1$. Also, $\{H\}$ and $\{T\}$ are disjoint, so $\Pr(\{H\} \cup \{T\}) = \Pr(\{H\}) + \Pr(\{T\})$. Collectively, we have

$$\begin{aligned}
&\Pr(\{H\}) = \Pr(\{T\}) \\
&\Pr(\{H\} \cup \{T\}) = 1 \\
&\Pr(\{H\} \cup \{T\}) = \Pr(\{H\}) + \Pr(\{T\})
\end{aligned}$$

Therefore, $\Pr(\{H\}) = \Pr(\{T\}) = \frac{1}{2}$.

Lecture 3: Aug 26

Last time

- Set theory (1.1)
- Axiomatic Foundations (1.2)

Today

- 5 bonus point presentation results
- Axiomatic Foundations (1.2)
- Calculus of Probabilities (1.2)
- Conditional Probability (1.3)

Example   If S is finite or countable (where the elements of S can be put into $1-1$ correspondence with a subset of the integers), then these technicalities really do not arise, for we we define for a given sample space S,

$$\mathcal{B} = \{\text{all subsets of } S, \text{ including } S \text{ itself}\}.$$

If S has n elements, there are 2^n sets in $\mathcal{B}$ (why?).[hint: for each element, it is either in or out of a subset, so 2 choices].

Example   Let $S = (-\infty, \infty)$, the real line. Then $\mathcal{B}$ is chosen to contain all sets of the form

$$[a, b], (a, b], (a, b), \text{ and } [a, b)$$

for all real numbers a and b. Also, from the properties of $\mathcal{B}$, it follows that $\mathcal{B}$ contains all sets that can be formed by taking (possibly countably infinite) unions and intersections of sets of the above varieties.

We now define a probability function.

Definition   Given a sample space S and an associated sigma algebra $\mathcal{B}$, a *probability function* is a function Pr with domain $\mathcal{B}$ that satisfies

1. $\Pr(A) \geqslant 0$ for all $A \in \mathcal{B}$.
2. $\Pr(S) = 1$.
3. If $A_1, A_2, \cdots \in \mathcal{B}$ are pairwise disjoint, then $\Pr(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \Pr(A_i)$.

The above three properties are usually referred to as the Axioms of Probability (or the Kolmogorov Axioms, after A. Kolmogorov, one of the fathers of probability theory). Any function that satisfies the Axioms of Probability is called a probability function.

Example Consider the simple experiment of tossing a fair coin (just once), so $S = \{H, T\}$. A reasonable probability function is the one that assigns equal probabilities to heads and tails, that is,

$$\Pr(\{H\}) = \Pr(\{T\}).$$

Since $S = \{H\} \cup \{T\}$, we have , from Axiom 1, $\Pr(\{H\} \cup \{T\}) = 1$. Also, $\{H\}$ and $\{T\}$ are disjoint, so $\Pr(\{H\} \cup \{T\}) = \Pr(\{H\}) + \Pr(\{T\})$. Collectively, we have

$$\begin{aligned}
&\Pr(\{H\}) = \Pr(\{T\}) \\
&\Pr(\{H\} \cup \{T\}) = 1 \\
&\Pr(\{H\} \cup \{T\}) = \Pr(\{H\}) + \Pr(\{T\})
\end{aligned}$$

Therefore, $\Pr(\{H\}) = \Pr(\{T\}) = \frac{1}{2}$.

Caculus of Probabilities

We start with some fairly self-evident properties of the probability function when applied to a single event.

Theorem If Pr is a probability function and A is any set in $\mathcal{B}$, then

1. $\Pr(\varnothing) = 0$, where $\varnothing$ is the empty set;
2. $\Pr(A) \leqslant 1$;
3. $\Pr(A^c) = 1 - \Pr(A)$.

proof:

- It's easy to prove (3) first. Since
  - $\Pr(A \cup A^c) = \Pr(S) = 1$,
  - A and A^c are disjoint, by axiom (3), $\Pr(A \cup A^c) = \Pr(A) + \Pr(A^c)$.

  so that $\Pr(A) + \Pr(A^c) = \Pr(S) = 1$
- with (3) proved, (1) is simple. because we know that
  - $S \cup \varnothing = S$,
  - $S \cap \varnothing = \varnothing$, they are disjoint,

  so that $\Pr(\varnothing) + \Pr(S) = \Pr(\varnothing \cup S) = \Pr(S)$.
- now for (2), $\Pr(A) = 1 - \Pr(A^c) \leqslant 1$, by axiom (1).

Theorem If Pr is a probability function and A and B are any sets in $\mathcal{B}$, then

1. $\Pr(B \cap A^c) = \Pr(B) - \Pr(A \cap B)$;
2. $\Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(A \cap B)$;

3. If $A \subset B$, then $\Pr(A) \leqslant \Pr(B)$.

proof:

1. For (1), we have $B = \{B \cap A\} \cup \{B \cap A^c\}$ and $\{B \cap A\} \cap \{B \cap A^c\} = \varnothing$, therefore
$$\Pr(B) = \Pr(\{B \cap A\} \cup \{B \cap A^c\})$$

2. For (2), we plug in (1) first such that we only need to show $\Pr(A \cup B) = \Pr(A) + \Pr(B \cap A^c)$. Since $A \cap \{B \cap A^c\} = \varnothing$ and $A \cup B = A \cup \{B \cap A^c\}$ (use a Venn diagram, or see Exercise 1.2), we have $\Pr(A \cup B) = \Pr(A) + \Pr(B \cap A^c)$.

3. For (3), if $A \subset B$, then $A \cap B = A$. Then using (1), we have
$$0 \leqslant \Pr(B \cap A^c) = \Pr(B) - \Pr(A)$$

Formula (2) in the above theorem gives a useful inequality for the probability of an intersection (Bonferroni's Inequality):
$$\Pr(A \cap B) \geqslant \Pr(A) + \Pr(B) - 1.$$

Theorem If Pr is a probability function, then

1. $\Pr(A) = \sum_{i=1}^{\infty} \Pr(A \cap C_i)$ for any partition $C_1, C_2, \ldots$;
2. $\Pr(\cup_{i=1}^{\infty} A_i) \leqslant \sum_{i=1}^{\infty} \Pr(A_i)$ for any sets $A_1, A_2, \ldots$.

where (1) is also referred to as "Total probability" and (2) is Boole's inequality.
proof:
By definition, since $C_1, C_2, \ldots$ form a partition, we have $C_i \cap C_j = \varnothing$ for all $i \neq j$, and $S = \cup_{i=1}^{\infty} C_i$. Therefore,
$$A = A \cap S = A \cap (\cup_{i=1}^{\infty} C_i) = \cup_{i=1}^{\infty} (A \cap C_i),$$
where the last equality follows from the Distributive Law. Since $\{A \cap C_i\} \cap \{A \cap C_j\} = \varnothing$ (i.e. $A \cap C_i$ and $A \cap C_j$ are disjoint), we have
$$\Pr(A) = \Pr\left(\cup_{i=1}^{\infty} (A \cap C_i)\right) = \sum_{i=1}^{\infty} \Pr(A \cap C_i).$$

To establish Boole's Inequality, we first construct a disjoint collection $A_1^*, A_2^*, \ldots$, with the property that $\cup_{i=1}^{\infty} A_i^* = \cup_{i=1}^{\infty} A_i$. We define A_i^* by
$$A_1^* = A_1,\ A_i^* = A_i \backslash \left(\cup_{j=1}^{i-1} A_j\right),\ i = 2, 3, \ldots,$$
where the notation $A \backslash B$ denotes the part of A that does not intersect with B. In other words, $A \backslash B = A \cap B^c$. It's easy to see that $\cup_{i=1}^{\infty} A_i^* = \cup_{i=1}^{i} nfty A_i$, and we have
$$\Pr\left(\cup_{i=1}^{\infty} A_i\right) = \Pr\left(\cup_{i=1}^{\infty} A_i^*\right) = \sum_{i=1}^{\infty} \Pr(A_i^*)$$

where the last equality holds because A_i^* are disjoint. To see this, consider any pair of $A_i^* \cap A_k^*, i > k$, then

$$\begin{aligned} A_i^* \cap A_k^* &= \left\{A_i \backslash (\cup_{j=1}^{i-1} A_j)\right\} \cap \left\{A_k \backslash (\cup_{j=1}^{k-1} A_j)\right\} \\ &= \left\{A_i \cap (\cup_{j=1}^{i-1} A_j)^c\right\} \cap \left\{A_k \cap (\cup_{j=1}^{k-1} A_j)^c\right\} \\ &= \left\{A_i \cap \left(\cap_{j=1}^{i-1} A_j^c\right)\right\} \cap \left\{A_k \cap \left(\cap_{j=1}^{k-1} A_j^c\right)\right\} \\ &= \varnothing. \end{aligned}$$

Lastly, we have $\Pr(A_i^*) \leqslant \Pr(A_i)$.

Conditional Probability

All of the probabilities that we have dealt with thus far have been unconditional probabilities. A sample space was defined and all probabilities were calculated with respect to that sample space. In many instances, however, we are in a position to update the sample space based on new information. In such cases we want to be able to update probability calculations or to calculate *conditional probabilities*.

Definition If A and B are events in S, and $\Pr(B) > 0$, then the *conditional probability* of A given B, written $\Pr(A|B)$, is

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

Note that B becomes the sample space now: $\Pr(B|B) = 1$.

Example Four cards are dealt from the top of a well-shuffled deck. What is the probability that they are the four aces? (there are in total 52 cards)

solution:
We define two events first. Let A be the event $\{$4 aces on top$\}$, and B be the event $\{$the first card on top is an ace$\}$. For a well-shuffled deck, all groups of 4 cards are equally likely. In total, there are $\begin{pmatrix} 52 \\ 4 \end{pmatrix} = \frac{52!(52-4)!}{4!} = 270,725$ distinct groups. Therefore, the probability of event A is $\Pr(A) = \frac{1}{270,725}$.
Note, $\begin{pmatrix} n \\ m \end{pmatrix}$ reads "from n choose m" (for $m \leqslant n$) and calculates by $\begin{pmatrix} n \\ m \end{pmatrix} = \frac{n!(n-m)!}{m!}$ that gives the number of distinct combinations of choosing m elements from n total elements.
Now, let's calculate $\Pr(A|B)$. First of all, $A \subset B$, so that we have $\Pr(A \cap B) = \Pr(A)$. For $\Pr(B)$, having an ace on top instead of the other 12 kinds, $\Pr(B) = \frac{1}{13}$. Then $\Pr(A|B) = \frac{\Pr(A \cap B)}{\Pr(B)} = \frac{\Pr(A)}{\Pr(B)} = \frac{1}{20,825}$.

Theorem (Bayes' Rule) Let $A_1, A_2, \dots$ be a partition of the sample space, and let B be any set. Then, for each $i = 1, 2, \dots,$

$$\Pr(A_i|B) = \frac{\Pr(B|A_i)\Pr(A_i)}{\sum_{j=1}^{\infty}\Pr(B|A_j)\Pr(A_j)}.$$

proof:
By "Total probability", we have $\Pr(B) = \sum_{j=1}^{\infty}\Pr(B \cap A_i)$ which is the denominator. Therefore, $\Pr(A_i|B) = \frac{\Pr(A_i \cap B)}{\Pr(B)} = \frac{\Pr(B|A_i)\Pr(A_i)}{\sum_{j=1}^{\infty}\Pr(B \cap A_i)}$.

Lecture 4: Aug 29

Last time

- Axiomatic Foundations (1.2)
- Calculus of Probabilities (1.2)

Today

- HW1 due 09/02, submit in the following recitation
- Conditional Probability (1.3)
- Independence (1.3)

Theorem If Pr is a probability function and A and B are any sets in $\mathcal{B}$, then

1. $\Pr(B \cap A^c) = \Pr(B) - \Pr(A \cap B)$;
2. $\Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(A \cap B)$;
3. If $A \subset B$, then $\Pr(A) \leqslant \Pr(B)$.

proof:

1. For (1), we have $B = \{B \cap A\} \cup \{B \cap A^c\}$ and $\{B \cap A\} \cap \{B \cap A^c\} = \varnothing$, therefore
$$\Pr(B) = \Pr(\{B \cap A\} \cup \{B \cap A^c\})$$
2. For (2), we plug in (1) first such that we only need to show $\Pr(A \cup B) = \Pr(A) + \Pr(B \cap A^c)$. Since $A \cap \{B \cap A^c\} = \varnothing$ and $A \cup B = A \cup \{B \cap A^c\}$ (use a Venn diagram, or see Exercise 1.2), we have $\Pr(A \cup B) = \Pr(A) + \Pr(B \cap A^c)$.
3. For (3), if $A \subset B$, then $A \cap B = A$. Then using (1), we have
$$0 \leqslant \Pr(B \cap A^c) = \Pr(B) - \Pr(A)$$

Formula (2) in the above theorem gives a useful inequality for the probability of an intersection (Bonferroni's Inequality):

$$\Pr(A \cap B) \geqslant \Pr(A) + \Pr(B) - 1.$$

Theorem If Pr is a probability function, then

1. $\Pr(A) = \sum_{i=1}^{\infty} \Pr(A \cap C_i)$ for any partition $C_1, C_2, \ldots$;
2. $\Pr(\cup_{i=1}^{\infty} A_i) \leqslant \sum_{i=1}^{\infty} \Pr(A_i)$ for any sets $A_1, A_2, \ldots$.

where (1) is also referred to as "Total probability" and (2) is Boole's inequality.
proof:
By definition, since $C_1, C_2, \dots$ form a partition, we have $C_i \cap C_j = \varnothing$ for all $i \neq j$, and $S = \cup_{i=1}^{\infty} C_i$. Therefore,

$$A = A \cap S = A \cap (\cup_{i=1}^{\infty} C_i) = \cup_{i=1}^{\infty} (A \cap C_i),$$

where the last equality follows from the Distributive Law. Since $\{A \cap C_i\} \cap \{A \cap C_j\} = \varnothing$ (i.e. $A \cap C_i$ and $A \cap C_j$ are disjoint), we have

$$\Pr(A) = \Pr\left(\cup_{i=1}^{\infty} (A \cap C_i)\right) = \sum_{i=1}^{\infty} \Pr(A \cap C_i).$$

To establish Boole's Inequality, we first construct a disjoint collection $A_1^*, A_2^*, \dots$, with the property that $\cup_{i=1}^{\infty} A_i^* = \cup_{i=1}^{\infty} A_i$. We define A_i^* by

$$A_1^* = A_1,\ A_i^* = A_i \backslash \left(\cup_{j=1}^{i-1} A_j\right),\ i = 2, 3, \dots,$$

where the notation $A \backslash B$ denotes the part of A that does not intersect with B. In other words, $A \backslash B = A \cap B^c$. It's easy to see that $\cup_{i=1}^{\infty} A_i^* = \cup_{i=1}^{\infty} A_i$, and we have

$$\Pr\left(\cup_{i=1}^{\infty} A_i\right) = \Pr\left(\cup_{i=1}^{\infty} A_i^*\right) = \sum_{i=1}^{\infty} \Pr(A_i^*)$$

where the last equality holds because A_i^* are disjoint. To see this, consider any pair of $A_i^* \cap A_k^*, i > k$, then

$$\begin{aligned}
A_i^* \cap A_k^* &= \left\{A_i \backslash (\cup_{j=1}^{i-1} A_j)\right\} \cap \left\{A_k \backslash (\cup_{j=1}^{k-1} A_j)\right\} \\
&= \left\{A_i \cap (\cup_{j=1}^{i-1} A_j)^c\right\} \cap \left\{A_k \cap (\cup_{j=1}^{k-1} A_j)^c\right\} \\
&= \left\{A_i \cap \left(\cap_{j=1}^{i-1} A_j^c\right)\right\} \cap \left\{A_k \cap \left(\cap_{j=1}^{k-1} A_j^c\right)\right\} \\
&= \varnothing.
\end{aligned}$$

Lastly, we have $\Pr(A_i^*) \leqslant \Pr(A_i)$.

Conditional Probability

All of the probabilities that we have dealt with thus far have been unconditional probabilities. A sample space was defined and all probabilities were calculated with respect to that sample space. In many instances, however, we are in a position to update the sample space based on new information. In such cases we want to be able to update probability calculations or to calculate *conditional probabilities*.

Definition If A and B are events in S, and $\Pr(B) > 0$, then the *conditional probability* of A given B, written $\Pr(A|B)$, is

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

Note that B becomes the sample space now: $\Pr(B|B) = 1$. For disjoint events, if $A \cap B = \varnothing$, then $\Pr(A|B) = 0$ and $\Pr(B|A) = 0$.

Conditional probability satisfies the axioms of probability:

1. $\Pr(S|B) = 1$,
2. $\Pr(A|B) \geqslant 0$,
3. If $A_1, A_2, \ldots$ are mutually exclusive events, then $\Pr(\cup_{i=1}^{\infty} A_i|B) = \sum_{i=1}^{\infty} \Pr(A_i|B)$

Example Four cards are dealt from the top of a well-shuffled deck. What is the probability that they are the four aces? What is the probability of getting four aces at the top if knowing the first card is an ace? (there are in total 52 cards)

solution:
We define two events first. Let A be the event $\{4$ aces on top$\}$, and B be the event $\{$the first card on top is an ace$\}$. For a well-shuffled deck, all groups of 4 cards are equally likely. In total, there are $\begin{pmatrix} 52 \\ 4 \end{pmatrix} = \frac{52!(52-4)!}{4!} = 270,725$ distinct groups. Therefore, the probability of event A is $\Pr(A) = \frac{1}{270,725}$.
Note, $\begin{pmatrix} n \\ m \end{pmatrix}$ reads "from n choose m" (for $m \leqslant n$) and calculates by $\begin{pmatrix} n \\ m \end{pmatrix} = \frac{n!}{m!(n-m)!}$ that gives the number of distinct combinations of choosing m elements from n total elements.
Now, let's calculate $\Pr(A|B)$. First of all, $A \subset B$, so that we have $\Pr(A \cap B) = \Pr(A)$. For $\Pr(B)$, having an ace on top instead of the other 12 kinds, $\Pr(B) = \frac{1}{13}$. Then $\Pr(A|B) = \frac{\Pr(A \cap B)}{\Pr(B)} = \frac{\Pr(A)}{\Pr(B)} = \frac{1}{20,825}$.

Theorem (Bayes' Rule) Let $A_1, A_2, \ldots$ be a partition of the sample space, and let B be any set. Then, for each $i = 1, 2, \ldots,$

$$\Pr(A_i|B) = \frac{\Pr(B|A_i)\Pr(A_i)}{\sum_{j=1}^{\infty} \Pr(B|A_j)\Pr(A_j)}.$$

proof:
By "Total probability", we have $\Pr(B) = \sum_{j=1}^{\infty} \Pr(B \cap A_i)$ which is the denominator. Therefore, $\Pr(A_i|B) = \frac{\Pr(A_i \cap B)}{\Pr(B)} = \frac{\Pr(B|A_i)\Pr(A_i)}{\sum_{j=1}^{\infty} \Pr(B \cap A_i)}$.

Independence

Definition Two events, A and B, are *statistically independent* if

$$\Pr(A \cap B) = \Pr(A)\Pr(B)$$

Note that independence could have been defined using Bayes' rule by $\Pr(A|B) = \Pr(A)$ or $\Pr(B|A) = \Pr(B)$ as long as $\Pr(A) > 0$ or $\Pr(B) > 0$. More notation, often statisticians omit $\cap$ when writing intersection in a probability function which means $\Pr(AB) = \Pr(A \cap B)$. Sometime, statisticians use comma (,) to replace $\cap$ inside a probability function too, $\Pr(A, B) = \Pr(A \cap B)$.

Theorem If A and B are independent events, then the following pairs are also independent.

1. A and B^c,
2. A^c and B,
3. A^c and B^c.

proof:
For (1),

$$\begin{aligned}
\Pr(A, B^c) &= \Pr(A) - \Pr(A, B) \\
&= \Pr(A) - \Pr(A)\Pr(B) \\
&= \Pr(A)(1 - \Pr(B)) \\
&= \Pr(A)\Pr(B^c)
\end{aligned}$$

For (2), we just need to switch A and B.
For (3), we have A^c and B are independent, then we can treat A^c as A' and B as B', then A' and B'^c are independent which is A^c and B^c are independent.
Alternatively, for (2),

$$\begin{aligned}
\Pr(A^c, B) &= \Pr(A^c|B)\Pr(B) \\
&= [1 - \Pr(A|B)]\Pr(B) \\
&= [1 - \Pr(A)]\Pr(B) \\
&= \Pr(A^c)\Pr(B).
\end{aligned}$$

And for (3),

$$\begin{aligned}
\Pr(A^c, B^c) &= \Pr(A^c) - \Pr(A^c, B) \\
&= \Pr(A^c) - \Pr(A^c)\Pr(B) \\
&= \Pr(A^c)\Pr(B^c).
\end{aligned}$$

Example Let the sample space S consist of the 3! permutations of the letters a, b, and c along with the three triples of each letter. Thus,

$$S = \left\{ \begin{array}{ccc} \text{aaa} & \text{bbb} & \text{ccc} \\ \text{abc} & \text{bca} & \text{cba} \\ \text{acb} & \text{bac} & \text{cab} \end{array} \right\}.$$

Furthermore, let each element of S have probability $\frac{1}{9}$. Define

$$A_i = \{i^{th} \text{ place in the triple is occupied by } a\}.$$

What are the values for $\Pr(A_i), i = 1, 2, 3$? Are they pairwise independent?
solution
It is easy to count that

$$\Pr(A_i) = \frac{1}{3}, i = 1, 2, 3,$$

and

$$\Pr(A_1, A_2) = \Pr(A_1, A_3) = \Pr(A_2, A_3) = \frac{1}{9}$$

so that A_is are pairwise independent.

Definition* A collection of events $A_1, \ldots, A_n$ are *mutually independent* if for any subcollection $A_{i_1}, \ldots, A_{i_k}$, we have

$$\Pr\left(\cap_{j=1}^{k} A_{i_j}\right) = \prod_{j=1}^{k} \Pr(A_{i_j}).$$

Lecture 5: Aug 31

Last time

- Conditional Probability (1.3)
- Independence (1.3)

Today

- HW1 due 09/02
- Random variables
- Distribution Functions

Random Variables

In many experiments, it is easier to deal with a summary variable than with the original probability structure.

Example consider an opinion poll, we might decide to ask 50 people whether they agree or disagree with a certain issue. If we record a "1" for agree and "0" for disagree, the sample space for this experiment has 2^{50} elements (all length 50 strings consist of 1s and 0s). However, if we are only interested in the number of people who agree, we may define a variable $X =$ number of 1s recorded out of 50. Then, the sample space for X is the set of integers $\{0, 1, 2, \ldots, 50\}$.

Definition A *random variable* (r.v.) is a function from a sample space S into the real numbers.

Example In some experiments random variables are implicitly used

Examples of random variables

Experiment	Random variable
Toss two dice	X = sum of numbers
Toss a coin 25 times	X = number of heads in 25 tosses
Apply different amounts of fertilizer to corn plants	X = yield / acre

In defining a random variable, we have also defined a new sample space (the range of the random variable).

Induced probability function Suppose we have a sample space $S = \{s_1, s_2, \ldots, s_n\}$ with a probability function Pr defined on the original sample space. We define a random variable X with range $\mathcal{X} = \{x_1, \ldots, x_m\}$. We can define a probability function $\Pr_X$ on $\mathcal{X}$ in the following way. We will observe $X = x_i$ if an only if the outcome of the random experiment is an $s_j \in S$ such that $X(s_j) = x_i$. Therefore,

$$\Pr_X(X = x_i) = \Pr(\{s_j \in S : X(s_j) = x_i\}),$$

defines an *induced* probability function on $\mathcal{X}$, defined in terms of the original function Pr.

We will write $\Pr(X = x_i)$ rather than $\Pr_X(X = x_i)$ for simplicity. Note on notation: random variables will always be denoted with uppercase leeters and the realized values of the variable (or its range) will be denoted by the corresponding lowercase letters.

Example Consider the experiment of tossing a fair coin three times. Define the random variable X to be the number of heads obtained in the three tosses. A complete enumeration of the value of X for each point in the sample space is

s	HHH	HHT	HTH	THH	TTH	THT	HTT	TTT
$X(s)$	4	2	2	2	1	1	1	0

What is the range of X? What is the induced probability function $\Pr_X$?
solution:
The range for the random variable X is $\mathcal{X} = \{0, 1, 2, 3\}$. Assuming all 8 points in S has probability $\frac{1}{8}$. By simply counting, we see that the induced probability function on $\mathcal{X}$ is

x	0	1	2	3
$\Pr_X(X = x)$	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$

So far, we have seen finite S and finite $\mathcal{X}$, and the definition of $\Pr_X$ is straightforward. If $\mathcal{X}$ is uncountable, we define the induced probability function, $\Pr_X$ for anyset $A \subset \mathcal{X}$,

$$\Pr_X(X \in A) = \Pr(\{s \in S : X(s) \in A\}).$$

This defines a legitimate probability function for which the Kolmogorov Axioms can be verified.

Distribution Functions

Distribution Functions are used to describe the behavior of a r.v.

Cumulative distribution function

Definition The *cumulative distribution function* or *cdf* of a random variable X, denoted by $F_X(x)$, is defined by

$$F_X(x) = \Pr_X(X \leqslant x), \text{ for all } x.$$

Definition The *survival function* of a random variable X, is defined by

$$S_X(x) = 1 - F_X(x) = \Pr_X(X > x).$$

Example Consider the experiment of tossing three fair coins, and let X = number of heads observed. The cdf of X is

$$F_X(x) = \begin{cases} 0 & \text{if } -\infty < x < 0 \\ \frac{1}{8} & \text{if } 0 \leqslant x < 1 \\ \frac{1}{2} & \text{if } 1 \leqslant x < 2 \\ \frac{7}{8} & \text{if } 2 \leqslant x < 3 \\ 1 & \text{if } 3 \leqslant x < \infty \end{cases}$$

Some properties of the cdf:
Let $F(x)$ be a cdf. Then

1. $0 \leqslant F(x) \leqslant 1$
2. $\lim_{x \to -\infty} F(x) = 0$
3. $\lim_{x \to \infty} F(x) = 1$
4. F is nondecreasing: if $a < b$, then $F(a) \leqslant F(b)$
5. F is right-continuous: $\lim_{x \downarrow b} F(x) = F(b)$, or $\lim_{x \to b^+} F(x) = F(b)$
6. $\Pr(a < X \leqslant B) = F(b) - F(a)$

Theorem The function $F(x)$ is a cdf if and only if the following three conditions hold:

1. $\lim_{x \to -\infty} F(x) = 0$ and $\lim_{x \to \infty} F(x) = 1$
2. F is nondecreasing: if $a < b$, then $F(a) \leqslant F(b)$
3. F is right-continuous: $\lim_{x \downarrow b} F(x) = F(b)$, or $\lim_{x \to b^+} F(x) = F(b)$

The cdf does not contain information about the original sample space.

Definition Two random variables X and Y are identically distributed if, for every Borel set $A \subset \mathbb{R}$, $\Pr(X \in A) = \Pr(Y \in A)$.

Example Toss a fair coin n times. The number of heads and the number of tails have the same distribution.

Theorem The following two statements are equivalent:

1. The random variables X and Y are *identically distributed*.
2. $F_X(x) = F_Y(x)$ for every x.