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

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

- 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 $\Pr(\dot{})$ 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 $\cup_{i=1}^{\infty} A_i^* = \cup_{i=1}^{i} nftyA_i$, 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 $\binom{52}{4} = \frac{52!(52-4)!}{4!} = 270,725$ distinct groups. Therefore, the probability of event A is $\Pr(A) = \frac{1}{270,725}$.
Note, $\binom{n}{m}$ reads "from n choose m" (for $m \leqslant n$) and calculates by $\binom{n}{m} = \frac{n!(n-m)!}{m!}$ 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 $A_1, A_2, \ldots$ 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.

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 $C_1, C_2, \ldots$ 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 $\cup_{i=1}^{\infty} A_i^* = \cup_{i=1}^{\infty} A_i$, 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 $\binom{52}{4} = \frac{52!(52-4)!}{4!} = 270,725$ distinct groups. Therefore, the probability of event A is $\Pr(A) = \frac{1}{270,725}$.
Note, $\binom{n}{m}$ reads "from n choose m" (for $m \leqslant n$) and calculates by $\binom{n}{m} = \frac{n!}{m!(n-m)!}$ 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, $\Pr(A, B) = \Pr(A \cap B)$.

Lecture 5: Aug 31

Last time

- Conditional Probability (1.3)
- Independence (1.3)

Today

- HW1 due 09/02
- Random variables (1.4)
- Distribution Functions (1.5)

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

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

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).

Lecture 6: Sept 2

Last time

- Random variables

Today

- HW1 due today
- no class next Monday (Labor day)
- Presentation: Approximate Bayesian Computation
- Presentation: Karl Pearson

Lecture 7: Sept 7

Last time

- Presentations

Today

- HW2 posted (due: Sept 15th)
- No lecture, but reviews on Fridays with two presentations
- Random variables
- Distribution Functions
- Types of Random Variables

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
$X(s)$	3	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.

Types of Random Variables

Definition A random variable X can be

- *discrete*:
  - X takes on a finite or countably infinite number of values
  - $F_X(x)$ is step-wise constant
- *continuous*:
  - the range of X consists of subsets of the real line
  - $F_X(x)$ is continuous.
- *mixed:* $F_X(x)$ is piecewise continuous.

Example A random variable has cdf

$$F(x) = \begin{cases} 0 & x < 0 \\ x/2 & 0 \leqslant x < 1 \\ 2/3 & 1 \leqslant x < 2 \\ 11/12 & 2 \leqslant x < 3 \\ 1 & 3 \leqslant x \end{cases}$$

Is this a valid cdf? Is it a discrete random variable or continuous random variable or mixed?
solution:
$F(x)$ satisfies the three properties of a cdf that

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 $\lim_{x \to b^+} F(x) = F(b)$.

Therefore, $F(x)$ is a valid cdf. The random variable X is a mixed type.

Lecture 8: Sept 9

Last time

- Random variables
- Distribution Functions
- Types of Random Variables

Today

- Presentation: Andrey Markov by Ryan Mortonson
- Presentation: Spatial Statistics by Camille Kreisel
- Review part 1

Review part 1

We briefly review what we have covered so far. We complement this review process with examples/questions taken from the book "Introduction to Probability Theory and Statistical Inference" 3rd ed. by Harold J. Larson.

We started with Set Theory.

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

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

An event occurs if any one of its elements is the outcome observed.

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.

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

Then we moved to define a probability function. To establish the domain for the probability function, we start with *sigma algebra*.

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).

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

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

which means that if we have property (1), (2) and (3) then we have property (1), (2), (3') and vise-versa (if we have property (1), (2) and (3') then we have property (1), (2), (3)).

This is because:

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

So that if we have property (3) that $A_1, A_2, \cdots \in \mathcal{B}$ and $\cup_{i=1}^{\infty} A_i \in \mathcal{B}$. Then by property (2), we know that $A_i^c \in \mathcal{B}$ for $i = 1, 2, \ldots$. And we can apply property (3) again such that if $A_1^c, A_2^c, \cdots \in \mathcal{B}$, then $(\cup_{i=1}^{\infty} A_i^c) \in \mathcal{B}$. Therefore, now we know $(\cup_{i=1}^{\infty} A_i^c) \in \mathcal{B}$ and we can apply property (2) again to get its complement which is also in the Borel field. Therefore, $(\cup_{i=1}^{\infty} A_i^c)^c \in \mathcal{B}$ which is $\cap_{i=1}^{\infty} A_i$.

For the other direction, we start from property (1), (2) and (3'). With property (3'), we have if $A_1, A_2, \cdots \in \mathcal{B}$, then $\cap_{i=1}^{\infty} A_i \in \mathcal{B}$. We again, first apply property (2) such that if $A_1, A_2, \cdots \in \mathcal{B}$, then $A_1^c, A_2^c, \cdots \in \mathcal{B}$. Now, by property (3'), we have $\cap_{i=1}^{\infty} A_i^c \in \mathcal{B}$. By applying property (2), we have $(\cap_{i=1}^{\infty} A_i^c)^c \in \mathcal{B}$. By substituting A_i with A_i^{*c} and taking complement at both side of equation $(\cup_{i=1}^{\infty} A_i^c)^c = \cap_{i=1}^{\infty} A_i$, we have $(\cup_{i=1}^{\infty} A_i^*) = (\cap_{i=1}^{\infty} A_i^{*c})^c$. Therefore, $\cup_{i=1}^{\infty} A_i = (\cap_{i=1}^{\infty} A_i^c)^c \in \mathcal{B}$ which is property (3).

Lecture 9: Sept 12

Last time

- Presentations

Today

- HW2 deadline extended (due: Sept 22nd)
- Random variables
- Distribution Functions
- Types of Random Variables

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.

Types of Random Variables

Definition A random variable X can be

- *discrete*:
  - X takes on a finite or countably infinite number of values
  - $F_X(x)$ is step-wise constant
- *continuous*:
  - the range of X consists of subsets of the real line
  - $F_X(x)$ is continuous.
- *mixed:* $F_X(x)$ is piecewise continuous.

Example A random variable has cdf

$$F(x) = \begin{cases} 0 & x < 0 \\ x/2 & 0 \leqslant x < 1 \\ 2/3 & 1 \leqslant x < 2 \\ 11/12 & 2 \leqslant x < 3 \\ 1 & 3 \leqslant x \end{cases}$$

Is this a valid cdf? Is it a discrete random variable or continuous random variable or mixed?
solution:
$F(x)$ satisfies the three properties of a cdf that

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 $\lim_{x \to b^+} F(x) = F(b)$.

Therefore, $F(x)$ is a valid cdf. The random variable X is a mixed type.

Discrete Random Variables

Suppose a random variable X takes only a finite or countable number of values. Let the sample space of X be $S = \{x_1, x_2, \dots\}$. Then the cdf can be expressed as:

$$F(x) = \sum_{x_i \leqslant x} \Pr(X = x_i).$$

Definition The *probability mass function* (pmf) of a discrete random variable X is given by

$$f_X(x) = \Pr(X = x) \text{ for all } x.$$

If the sample space of X is $X = \{x_1, x_2, \dots\}$, then

$$f(x_i) = \Pr(X = x_i) = \Pr(x_{i-1} < X \leqslant x_i) = F(x_i) - F(x_{i-1}).$$

Example (Geometric probabilities) Suppose we do an experiment that consists of tossing a coin until a head appears. Let p = probability of a head on any given toss, and define a random variable X = number of tosses required to get a head. Then for any $x = 1, 2, \dots,$

$$\Pr(X = x) = (1 - p)^{x-1} p,$$

since we must get $x - 1$ tails followed by a head for the event to occur and all trials are independent. What is the pmf of the above Geometric distribution? What is the cdf?

solution:
We have the pmf

$$f(x) = \Pr(X = x) = \begin{cases} (1-p)^{x-1}p & \text{for } x = 1, 2, \dots \\ 0 & \text{otherwise.} \end{cases}$$

For cdf, we have

$$\begin{aligned} F(x) = \Pr(X \leqslant x) &= \sum_{i=1}^{\lfloor x \rfloor} f(i) \\ &= \begin{cases} f(1) + f(2) + \cdots + f(\lfloor x \rfloor) & \text{for } x \geqslant 1 \\ 0 & \text{otherwise} \end{cases} \\ &= \begin{cases} 1 - (1-p)^{\lfloor x \rfloor} & \text{for } x \geqslant 1 \\ 0 & \text{for } x < 1 \end{cases} \end{aligned}$$

where $\lfloor x \rfloor$ denote the floor function that returns the largest integer smaller or equal to x and we used the summation of a geometric sequence.

Definition The *domain* of a random variable X is the set of all values of x for which $f(x) > 0$. This is also called *range* or *sample space*.

Properties of the pmf:

1. $f(x) > 0$ for at most a countable number of values x. For all other values x, $f(x) = 0$.
2. Let $\{x_1, x_2, \dots\}$ denote the domain of X. Then

$$\sum_{i=1}^{\infty} f(x_i) = 1.$$

An obvious consequence is that $f(x) \leqslant 1$ over the domain.

Example What is the pmf of a deterministic random variable (a constant)?
solution:

$$f(x) = \Pr(X = x) = \begin{cases} 1 & \text{for } x = c \\ 0 & \text{otherwise.} \end{cases}$$

This is equivalent as a constant of value c.

Example In many applications, a formula can be used to represent the pmf of a random variable. Suppose X can take values $1, 2, \dots$ with pmf

$$f(x) = \begin{cases} \frac{1}{x(x+1)} & \text{for } x = 1, 2, \dots \\ 0 & \text{otherwise.} \end{cases}$$

How would we determine if this is an allowable pmf?
solution:
We show that $f(x)$ satisfies the properties of pmf.

1. $f(x) > 0$ for a countable number of values x. For all other values x, $f(x) = 0$.
2. Let $\{x_1, x_2, \dots\}$ denote the domain of X. Then

$$\sum_{i=1}^{\infty} f(x_i) = \sum_{i=1}^{\infty} f(i) = \sum_{i=1}^{\infty} \left(\frac{1}{x} - \frac{1}{x+1} \right) = 1.$$

Lecture 10: Sept 14

Last time

- Random variables
- Distribution Functions
- Types of Random Variables

Today

- Continuous Random Variables
- Counting Techniques

Continuous Random Variables

Definition A random variable X is *continuous* if $F_X(x)$ is a continuous function of x.

Definition A random variable X is *absolutely continuous* if $F_X(x)$ is an absolutely continuous function of x.

Definition A function $F(x)$ is *absolutely continuous* if it can be written

$$F(x) = \int_{-\infty}^{x} f(x)dx.$$

Absolute continuity is stronger than continuity but weaker than differentiability. An example of an absolutely continuous function is one that is:

- continuous everywhere
- differentiable everywhere, except possibly for a countable number of points.

Definition The *probability density function* or pdf, $f_X(x)$, of a continuous random variable X is the function that satisfies

$$F_X(x) = \int_{-\infty}^{x} f_X(t)dt \quad \text{for all } x.$$

Notation: We write $X \sim F_X(x)$ for the expression "X has a distribution given by $F_X(x)$" where we read the symbol "$\sim$" as "is distributed as". Similarly, we can write $X \sim f_X(x)$ or , if X and Y have the same distribution, $X \sim Y$.

Theorem A function $f_X(x)$ is a pdf (or pmf) of a random variable X if and only if

1. $f_X(x) \geqslant 0$ for all x.
2. $\int_{-\infty}^{\infty} f_X(x)dx = 1$ (pdf) or $\sum_x f_X(x) = 1$ (pmf).

Example Suppose $F(x) = 1 - e^{-\lambda x}$ for $x > 0$ and $F(x) = 0$ otherwise. Is $F(x)$ a cdf? What is the associated pdf?
solution:
$F(x)$ satisfies the three properties of cdf

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 $\lim_{x \to b^+} F(x) = F(b)$.

$F(x)$ is a cdf. Actually, $F(x)$ is the cdf of exponential distribution.

To get the pdf, we only need to differentiate the cdf.

$$f(x) = \frac{dF(x)}{dx} = \begin{cases} \lambda e^{-\lambda x} & \text{for } x > 0 \\ 0 & \text{otherwise.} \end{cases}$$

Notes

- If X is a continuous random variable, then $f(x)$ is not the probability that $X = x$. In fact, if X is an absolutely continuous random variable with density function $f(x)$, then $\Pr(X = x) = 0$. (Why?)
  proof

$$\begin{aligned} \Pr(X = x) &= \lim_{h \to 0} \int_{x-h}^{x+h} f(u)du \\ &= \lim_{h \to 0} F(x + h) - F(x - h) \\ &= F(x+) - F(x-) \\ &= 0 \end{aligned}$$

- Because $\Pr(X = a) = 0$, all the following are equivalent:

$$\Pr(a \leqslant X \leqslant b), \quad \Pr(a \leqslant X < b) \quad , \quad \Pr(a < X \leqslant b) \quad \text{and} \quad \Pr(a < X < b)$$

- $f(x)$ can exceed one!

Counting Techniques

These sections are from 2.4 of "Introduction to Probability Theory and Statistical Inference" by Harold J. Larson. We employ them to discuss combinatorics.

When the equally likely assumption is made for a finite sample space, the probability of occurrence of any event A is given by the ratio of the number of elements belonging to A to the number of elements belonging to S. For such cases it is useful to be able to count the number of elements belonging to given sets.

A very simple technique that is frequently useful in counting problems is called the *multiplication principle*.

Definition If a first operation can be performed in any of n_1 ways and a second operation can then be performed in any of n_2 ways, both operations can be performed (the second immediately following the first) in $n_1 \cdot n_2$ ways.

Example If we can travel from town A to town B in 3 ways and from town B to town C in 4 ways, then we can travel from A to C via B in a total of $3 \cdot 4 = 12$ ways.

Example If the operation of tossing a die gives rise to 1 of 6 possible outcomes and the operation of tossing a second die gives rise to 1 of 6 possible outcomes, then the operation of tossing a pair of dice gives rise to $6 \cdot 6 = 36$ possible outcomes.

Definition An arrangement of n symbols in a definite order is called a *permutation* of n symbols.

Example Let's consider all possible n-tuples made by n different symbols. In listing all the possible n-tuples, we would perform n natural operations. First we must fill the leftmost position of n-tuples, we have all n symbols to choose from. Then we must fill the second leftmost position, where we have $n-1$ symbols to choose from. Then, the third position with $n-2$ symbols to choose from, and so on. Finally, when we reach the right most position, we have 1 symbol left.

Using the multiplication rule, the total number of ways we can perform all n operations will be

$$n! = n(n-1)(n-2)\cdots 2 \cdot 1,$$

where we write $n!$ (read n-factorial) and we define $0! = 1$.

Example Suppose the same 5 people park their cars on the same side of the street in the same block every night. How many different ordering of the 5 cars parked on the street are possible?
Solution:

$$5! = 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1 = 120$$

Example Suppose the same 5 people park their cars on the two sides of the street in the same block every night where one side has 3 slots and the other side has 2. How many different ordering of 3 cars out of 5 can be parked on the 3-slot side?
Solution:
For the first slot, we have 5 possible cars to choose from. For the second slot, we have 4 cars to choose from (one is taken for the first slot). For the third slot, we have 3 cars to choose from (two cars are taken for the other two slots). In total, there are

$$5 \cdot 4 \cdot 3 = 60$$

ways.

Definition The number of r-tuples we can make $r \leqslant n$, using n different symbols (each only once), is called the *number of permutations of n things r at a time* and is denoted by ${}^nP_r$ which is calculated as

$$ {}^nP_r = n(n-1)\cdots(n-r+1). $$

Example Fifteen cars enter a race. In how many different ways could trophies for first, second, and third place be awarded?
Solutions:

$$ {}^{15}P_3 = 15 \cdot 14 \cdot 13 = 2730. $$

Example How many of the 3-tuples just counted have car number 15 in the first position?
Solutions:

$$ {}^{14}P_2 = 14 \cdot 13 = 182. $$

Definition The number of distinct subsets, each of size r, that can be constructed from a set with n elements is called the number of *combinations* of n things r at a time: this number is represented by $\binom{n}{r}$ which reads n choose r.

$$ \binom{n}{r} = \frac{n!}{r!(n-r)!}. $$

Example How many distinct 5-card hands can be dealt from a standard 52-card deck?

$$ \binom{52}{5} = \frac{52!}{5!47!} = 2,598,960. $$

Theorem If x and y are any two real numbers and n is a positive integer, then

$$ (x+y)^n = \sum_{i=1}^{n} \binom{n}{i} x^i y^{n-i}, \quad \text{where } \binom{n}{i} = \frac{n!}{(n-i)!i!}. $$

Lecture 11: Sept 16

Last time

- Continuous Random Variables

Today

- Presentations
- Review part 2

Review part 2

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.

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. $\Pr(A^c) = 1 - \Pr(A)$.

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 $\Pr(A) \leqslant \Pr(B)$.

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.

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: $\Pr(B|B) = 1$.

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 $i = 1, 2, \ldots,$

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

Lecture 12: Sept 19

Last time

- Random variables
- Distribution Functions
- Types of Random Variables

Today

- Counting Techniques
- Transformations of Random Variables

Definition The number of r-tuples we can make $r \leqslant n$, using n different symbols (each only once), is called the *number of permutations of n things r at a time* and is denoted by ${}^nP_r$ which is calculated as

$$ {}^nP_r = n(n-1)\cdots(n-r+1). $$

Example Fifteen cars enter a race. In how many different ways could trophies for first, second, and third place be awarded?
Solutions:

$$ {}^{15}P_3 = 15 \cdot 14 \cdot 13 = 2730. $$

Example How many of the 3-tuples just counted have car number 15 in the first position?
Solutions:

$$ {}^{14}P_2 = 14 \cdot 13 = 182. $$

Definition The number of distinct subsets, each of size r, that can be constructed from a set with n elements is called the number of *combinations* of n things r at a time: this number is represented by $\binom{n}{r}$ which reads n choose r.

$$ \binom{n}{r} = \frac{n!}{r!(n-r)!}. $$

Example How many distinct 5-card hands can be dealt from a standard 52-card deck?

$$ \binom{52}{5} = \frac{52!}{5!47!} = 2,598,960. $$

Theorem If x and y are any two real numbers and n is a positive integer, then

$$ (x+y)^n = \sum_{i=0}^{n} \binom{n}{i} x^i y^{n-i}, \qquad \text{where } \binom{n}{i} = \frac{n!}{(n-i)!i!}. $$

Transformations of Random Variables

Theorem If X is a r.v. with sample space $\mathcal{X} \subset \mathbb{R}$ and cdf $F_X(x)$, then any function of X, say $Y = g(X)$ is also a random variable. The new random variable Y has a new sample space $\mathcal{Y} = g(X) \subset \mathbb{R}$. The objective is to find the cdf $F_Y(y)$ of Y.

Probability mapping: For any set $A \subset \mathcal{Y}$:

$$\begin{aligned}\Pr(Y \in A) &= \Pr(g(X) \in A) \\ &= \Pr(\{x \in \mathcal{X} : g(x) \in A\}) \\ &= \Pr(X \in g^{-1}(A)),\end{aligned}$$

where we have defined

$$g^{-1}(A) = \{x \in \mathcal{X} : g(x) \in A\}.$$

Notice that $g^{-1}(A)$ is well defined even if $g(\cdot)$ is not necessarily bijective.

Example (Binomial transformation) A discrete random variable X has a *binomial distribution* if its pmf is of the form

$$f_X(x) = \Pr(X = x) = \binom{n}{x} p^x (1-p)^{n-x}, \quad x = 0, 1, \ldots, n,$$

where n is a positive integer and $0 \leqslant p \leqslant 1$. Values such as n and p that can be set to different values, producing different probability distributions, are called *parameters*. Consider a random variable $Y = g(X)$, where $g(x) = n - x$; that is, $Y = n - X$. Here $\mathcal{X} = \{0, 1, \ldots, n\}$ and $\mathcal{Y} = \{y : y = g(x), x \in \mathcal{X}\} = \{0, 1, \ldots, n\}$. For any $y \in \mathcal{Y}$, $n - x = g(x) = y$ if and only if $x = n - y$. Therefore, $g^{-1}(y) = n - y$ and

$$\begin{aligned}f_Y(y) &= \sum_{x \in g^{-1}(y)} f_X(x) \\ &= f_X(n-y) \\ &= \binom{n}{n-y} p^{n-y} (1-p)^{n-(n-y)} \\ &= \binom{n}{y} (1-p)^y p^{n-y}.\end{aligned}$$

Therefore, Y also has a binomial distribution, but with parameters n and $1 - p$.

Example (exercise 2.3) Suppose X has the geometric pmf $f_X(x) = \frac{1}{3}(\frac{2}{3})^x, x = 0, 1, 2, \ldots$. Determine the probability distribution of $Y = X/(X+1)$. Note that here both X and Y are discrete random variables. To specify the probability distribution of Y, specify its pmf. *Solution:*

$$\Pr(Y = y) = \Pr(\frac{X}{X+1} = y) = \Pr(X = \frac{y}{1-y}) = \frac{1}{3}(\frac{2}{3})^{y/(1-y)}, y = 0, \frac{1}{2}, \frac{2}{3}, \ldots, \frac{x}{x+1}, \ldots$$

Lecture 13: Sept 21

Last time

- Counting Techniques
- Transformations of Random Variables

Today

- Transformations of continuous random variables

Theorem Suppose a continuous random variable X has cdf $F_X(x)$, let $Y = g(X)$, and let $\mathcal{X}$ and $\mathcal{Y}$ be defined as

$$\mathcal{X} = \{x : f(x) > 0\} \quad \text{and} \quad \mathcal{Y} = \{y : y = g(x) \text{ for some } x \in \mathcal{X}\}.$$

Then,

1. If g is an increasing function on $\mathcal{X}$, $F_Y(y) = F_X(g^{-1}(y))$ for $y \in \mathcal{Y}$.
2. If g is a decreasing function on $\mathcal{X}$, $F_Y(y) = 1 - F_X(g^{-1}(y))$ for $y \in \mathcal{Y}$.

Proof: We start with

$$\begin{aligned} F_Y(y) &= \Pr(Y \leqslant y) \\ &= \Pr(g(X) \leqslant y) \end{aligned}$$

1. If g is an increasing function, then $g(X) \leqslant y$ if and only if $X \leqslant g^{-1}(y)$. Therefore, $F_Y(y) = \Pr(g(X) \leqslant y) = \Pr(X \leqslant g^{-1}(y)) = F_X(g^{-1}(y))$.
2. Similarly, if g is a decreasing function, then $g(X) \leqslant y$ if and only if $X \geqslant g^{-1}(y)$. And $F_Y(y) = \Pr(g(X) \leqslant y) = \Pr(X \geqslant g^{-1}(y)) = 1 - F_X(g^{-1}(y))$.

Theorem Let X have pdf $f_X(x)$ and let $Y = g(X)$, where g is a monotone function. Let $\mathcal{X}$ and $\mathcal{Y}$ be defined as

$$\mathcal{X} = \{x : f(x) > 0\} \quad \text{and} \quad \mathcal{Y} = \{y : y = g(x) \text{ for some } x \in \mathcal{X}\}.$$

Suppose that $f_X(x)$ is continuous on $\mathcal{X}$ and that $g^{-1}(y)$ has a continuous derivative on $\mathcal{Y}$. Then the pdf of Y is given by

$$f_Y(y) = \begin{cases} f_X(g^{-1}(y))|\frac{d}{dy}g^{-1}(y)| & y \in \mathcal{Y} \\ 0 & otherwise. \end{cases}$$

Proof:
From last theorem, we have the cdf forms $F_Y(y)$. Then $f_Y(y) = \frac{d}{dy}F_Y(y)$. (finish the proof)
From last theorem, we have

$$F_Y(y) = \begin{cases} F_X(g^{-1}(y)) & \text{if } g \text{ is increasing} \\ 1 - F_X(g^{-1}(y)) & \text{if } g \text{ is decreasing.} \end{cases}$$

We have, by the chain rule,

$$f_Y(y) = \frac{d}{dy}F_Y(y) = \begin{cases} f_X(g^{-1}(y))\frac{d}{dy}g^{-1}(y) & \text{if } g \text{ is increasing} \\ -f_X(g^{-1}(y))\frac{d}{dy}g^{-1}(y) & \text{if } g \text{ is decreasing,} \end{cases}$$

where $\frac{d}{dy}g^{-1}(y) < 0$ when g is decreasing such that $-\frac{d}{dy}g^{-1}(y) = |\frac{d}{dy}g^{-1}(y)|$.

Example (Square transformation) Suppose X is a continuous random variable. For $y > 0$, the cdf of $Y = X^2$ is

$$F_Y(y) = \Pr(Y \leqslant y) = \Pr(X^2 \leqslant y) = \Pr(-\sqrt{y} \leqslant X \leqslant \sqrt{y}).$$

Because x is continuous, we can drop the equality from the left endpoint and obtain

$$\begin{aligned} F_Y(y) &= \Pr(-\sqrt{y} < X \leqslant \sqrt{y}) \\ &= \Pr(X \leqslant \sqrt{y}) - \Pr(X \leqslant -\sqrt{y}) = F_X(\sqrt{y}) - F_X(-\sqrt{y}). \end{aligned}$$

The pdf of Y can now be obtained from the cdf by differentiation:

$$\begin{aligned} f_Y(y) &= \frac{d}{dy}F_Y(y) \\ &= \frac{d}{dy}\left[F_X(\sqrt{y}) - F_X(-\sqrt{y})\right] \\ &= \frac{1}{2\sqrt{y}}f_X(\sqrt{y}) + \frac{1}{2\sqrt{y}}f_X(-\sqrt{y}), \end{aligned}$$

where we use the chain rule to differentiate $F_X(\sqrt{y})$ and $F_X(-\sqrt{y})$.

Example (Linear transformation) Suppose X is a continuous random variable with pdf $f_X(x)$. Let

$$Y = a + bX, \quad \frac{dy}{dx} = b.$$

Then

$$f_Y(y) = f_X\left[g^{-1}(y)\right]|\frac{dx}{dy}| = f_X(\frac{y-a}{b})\frac{1}{|b|}.$$

This transformation is often used when X has mean 0 and standard deviation 1. The linear transformation above creates a random variable Y with a distribution that has the same shape as that of X but has mean a and variance b^2.

Conversely, if Y has mean a and standard deviation b, then $X = (Y - a)/b$ has mean 0 and standard deviation 1. This is called sometimes the "Studentized" transformation.

Example (Normal distribution) Let $X \sim N(0,1)$:

$$f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}, \quad -\infty < x < \infty.$$

The transformation

$$Y = \mu + \sigma X, \quad X = \frac{Y - \mu}{\sigma}$$

yields

$$f_Y(y) = f_X(\frac{y-\mu}{\sigma})\frac{1}{\sigma} = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y-\mu)^2}{2\sigma^2}}.$$

More generally, a distribution is a member of the class of *location-scale* distributions if the distribution of a linear transformation of a random variable with that distribution has the same distribution, but with different parameters.

Example (Square root of an exponential RV) Suppose $X \sim exp(\lambda)$, so that

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

and consider the distribution of $Y = \sqrt{X}$. The transformation

$$y = g(x) = \sqrt{x}, \quad x \geqslant 0$$

is one-to-one and has an inverse $x = y^2$ with $dx/dy = 2y$. Thus

$$f_Y(y) = f_X(y^2)2y = 2\lambda y e^{-\lambda y^2}, \quad y \geqslant 0.$$

This distribution is a particular form of the Rayleigh distribution and is a special case of the Weibull distribution.

Lecture 14: Sept 23

Last time

- Transformations of continuous random variables

Today

- Practice examples

Example A random variable X has a discrete uniform $(1, N)$ distribution, $X \sim Unif\{1, N\}$, if

$$\Pr(X = x|N) = \frac{1}{N}, \quad x = 1, 2, \ldots, N,$$

where N is a specified integer. This distribution puts equal mass on each of the outcomes $1, 2, \ldots, N$. Question: what is the cdf of this r.v.?
Solutions:

$$F(x) = \Pr(X \leqslant x|N) = \begin{cases} 0, & x < 1 \\ \frac{1}{N}, & 1 \leqslant x < 2 \\ \frac{2}{N}, & 2 \leqslant x < 3 \\ \frac{3}{N}, & 3 \leqslant x < 4 \\ \vdots & \\ \frac{N-1}{N}, & N - 1 \leqslant x < N \\ 1, & N \leqslant x \end{cases}$$

Example The continuous uniform distribution is defined by spreading mass uniformly over an interval $[a, b]$. A random variable X has a continuous uniform $[a, b]$ distribution, $X \sim Unif(a, b)$, if its pdf is given by

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

Question: what is the cdf?
Solutions:

$$F(x) = \Pr(X \leqslant x) = \begin{cases} 0, & x < a \\ \frac{x-a}{b-a}, & a \leqslant x < b \\ 1, & b \leqslant x. \end{cases}$$

Lecture 15: Sept 26

Last time

- Presentation

Today

- HW3 posted
- Midterm Exam 1 10/10, will have a practice exam
- Probability integral transformation
- Expectations (2.2)

Theorem (Probability integral transformation) Let X have continuous cdf $F_X(x)$ and define the random variable Y as $Y = F_X(X)$. Then Y is uniformly distributed on $(0, 1)$, that is, $\Pr(Y \leqslant y) = y, 0 < y < 1$.

Before we prove this theorem, we will digress for a moment and look at F_X^{-1}, the inverse of the cdf F_X, in some detail. If F_X is strictly increasing, then F_X^{-1} is well defined by

$$F_X^{-1}(y) = x \iff F_X(x) = y.$$

However, if F_X is constant on some interval, then F_X^{-1} is not well defined as Figure 13.1 illustrates. Any $x_1 \leqslant x \leqslant x_2$ satisfies $F_X(x) = y$

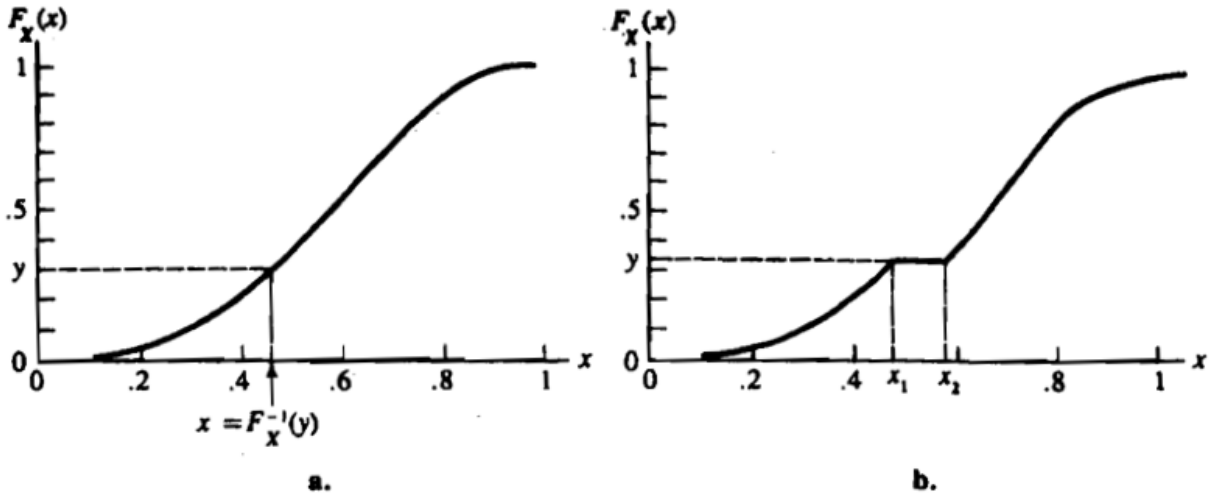

Figure 13.1: Figure 2.1.2. (a) $F_X(x)$ strictly increasing; (b) $F_X(x)$ nondecreasing

This problem is avoided by defining F_X^{-1} for $0 < y < 1$ by

$$F_X^{-1}(y) = \inf\{x : F_X(x) \geqslant y\}.$$

With this definition, for Figure 13.1(b), we have $F_X^{-1}(y) = x_1$.
Proof:
For $Y = F_X(X)$, we have, for $0 < y < 1$,

$$\begin{aligned}
\Pr(Y \leqslant y) &= \Pr(F_X(X) \leqslant y) \\
&= \Pr(F_X^{-1}[F_X(X)] \leqslant F_X^{-1}(y)) \quad (F_X^{-1} \text{ is increasing}) \\
&= \Pr(X \leqslant F_X^{-1}(y)) \\
&= F_X(F_X^{-1}(y)) \qquad (\text{definition of } F_X) \\
&= y.
\end{aligned}$$

One application of the probability integral transformation is in the generation of random samples from a particular distribution. If it is required to generate an observation X from a population with cdf F_X, we need only generate a uniform random number U, between 0 and 1, and solve for x in the equation $F_X(x) = u$.

Expected Values

Definition The *expected value* or *mean* of a random variable $g(X)$, denoted by $Eg(X)$, is

$$Eg(X) = \begin{cases} \int_{-\infty}^{\infty} g(x)f(x)dx & \text{if } X \text{ is continuous} \\ \sum_{x \in \mathcal{X}} g(x) \Pr(X = x) & \text{if } X \text{ is discrete} \end{cases}$$

Provided the integral or summation exists.

If we let $g(X) = X$, then we get

$$EX = \begin{cases} \int_{-\infty}^{\infty} xf(x)dx & \text{if } X \text{ is continuous} \\ \sum_{x \in \mathcal{X}} x \Pr(X = x) & \text{if } X \text{ is discrete} \end{cases}$$

Example (Exponential mean) Suppose X has an *exponential (λ) distribution*, $X \sim Exp(\lambda)$, that is, it has pdf given by

$$f_X(x) = \frac{1}{\lambda} e^{-x/\lambda}, \quad 0 \leqslant x < \infty, \lambda > 0.$$

Find out EX.
Solution:

$$
\begin{aligned}
EX &= \int_0^\infty \frac{1}{\lambda} x e^{-x/\lambda} dx \\
&= -x e^{-x/\lambda}\Big|_0^\infty + \int_0^\infty e^{-x/\lambda} dx \\
&= \int_0^\infty e^{-x/\lambda} dx \\
&= \lambda
\end{aligned}
$$

Example (Binomial mean) if X has a *binomial distribution*, $X \sim Binomial(n, p)$, its pmf is given by

$$\Pr(X = x) = \binom{n}{x} p^x (1-p)^{n-x}, \quad x = 0, 1, \ldots, n,$$

where n is a positive integer, $0 \leqslant p \leqslant 1$, and for every fixed pair n and p the pmf sums to 1. Find out EX.
Solution:

$$EX = \sum_{x=0}^{n} x \binom{n}{x} p^x (1-p)^{n-x}, = \sum_{x=1}^{n} x \binom{n}{x} p^x (1-p)^{n-x}.$$

Using the identity $x\binom{n}{x} = n\binom{n-1}{x-1}$, we have

$$
\begin{aligned}
EX &= \sum_{x=1}^{n} n \binom{n-1}{x-1} p^x (1-p)^{n-x} \\
&= \sum_{y=0}^{n-1} n \binom{n-1}{y} p^{y+1} (1-p)^{n-(y+1)} \\
&= np \sum_{y=0}^{n-1} \binom{n-1}{y} p^y (1-p)^{n-1-y} \\
&= np,
\end{aligned}
$$

since the last summation must be 1, being the sum over all possible values of a binomial$(n-1, p)$ pmf.

Lecture 16: Sept 28

Last time

- HW3 posted
- Midterm Exam 1 10/10, will have a practice exam
- Probability integral transformation
- Expectations (2.2)

Today

- Exam 1 covers up to next Monday's lecture
- Expectations (2.2)
- Moments and moment generating function

Expectation

The process of taking expectations is a linear operation, which means that the expectation of a linear function of X can be easily evaluated by noting that for any constants a and b, such that

$$E(aX + b) = aEX + b$$

Theorem Let X be a random variable and let a, b, and c be constants. Then for any functions $g_1(x)$ and $g_2(x)$ whose expectations exist,

1. $E(ag_1(X) + bg_2(X) + c) = aEg_1(X) + bEg_2(X) + c$.
2. If $g_1(x) \geqslant 0$ for all x, then $Eg_1(X) \geqslant 0$.
3. If $g_1(x) \geqslant g_2(x)$ for all x, then $Eg_1(X) \geqslant Eg_2(X)$.
4. If $a \leqslant g_1(x) \leqslant b$ for all x, then $a \leqslant Eg_1(X) \leqslant b$.

Proof:
We will give details for only the continuous case, the discrete case being similar. By definition

$$\begin{aligned} E(ag_1(X) + bg_2(X) + c) &= \int_{-\infty}^{\infty} [ag_1(x) + bg_2(x) + c] f_X(x)dx \\ &= \int_{-\infty}^{\infty} ag_1(x)f_X(x)dx + \int_{-\infty}^{\infty} bg_2(x)f_X(x)dx + \int_{-\infty}^{\infty} cf_X(x)dx \\ &= aEg_1(X) + bEg_2(X) + c \end{aligned}$$

The other three properties are proved in a similar manner (shown in class).

Example (Method of indicators) An example of how the above properties are useful. Let $X \sim Binomial(n, p)$ for n positive integer and $0 \leqslant p \leqslant 1$ (n is the number of independent identical binary trials and p is the probability of success). We can write

$$X = \sum_{i=1}^{n} I_i$$

where I_i is the indicator that i^{th} trial is a success (i.e. $I_i \overset{\text{i.i.d.}}{\sim} Bernoulli(p)$). We have

$$EI_i = 1 \cdot p + 0 \cdot (1 - p) = p.$$

Therefore,

$$EX = \sum_{i=1}^{n} EI_i = \sum_{i=1}^{n} p = np.$$

Theorem For a non-negative random variable X (i.e. $f(x) = 0$ for $x < 0$).

$$EX = \begin{cases} \int_0^\infty (1 - F(x))dx, & X \text{ continuous} \\ \sum_{x=0}^\infty (1 - F(x)), & X \text{ discrete} \end{cases}$$

Proof:
We prove the continuous case first,

$$\begin{aligned} \int_0^\infty [1 - F(x)]\, dx &= \int_0^\infty [1 - \Pr(X \leqslant x)]\, dx \\ &= \int_0^\infty \Pr(X > x) dx \\ &= \int_0^\infty \int_{y=x}^\infty f_X(y) dy dx \\ &= \int_0^\infty \int_{x=0}^y f_X(y) dx dy \\ &= \int_0^\infty y f_X(y) dy \\ &= EX. \end{aligned}$$

Then, for discrete case, we have

$$\begin{aligned}\sum_{x=0}^{\infty}(1-F(x)) &= \sum_{x=0}^{\infty}\Pr(X>x)\\ &= \sum_{x=0}^{\infty}\sum_{y=x+1}^{\infty}\Pr(X=y)\\ &= \sum_{y=1}^{\infty}\sum_{x=0}^{y-1}\Pr(X=y)\\ &= \sum_{y=1}^{\infty} y\Pr(X=y)\\ &= EX\end{aligned}$$

Moments

Example (Minimizing distance) The expected value of a random variable has another property, one that we can think of as relating to the interpretation of EX as a good guess at a value of X.

Suppose we measure the distance between a random variable X and a constant b by $(X-b)^2$. The closer b is to X, the smaller this quantity is. We can now determine the value of b that minimizes $E\left[(X-b)^2\right]$ and, hence, will provide us with a good predictor of X. (Note that it does no good to look for a value of b that minimizes $(X-b)^2$, since the answer would depend on X, making it a useless predictor of X.)

We could proceed with the minimization of $E(X-b)^2$ by using calculus, but there is a simpler method:

$$\begin{aligned}E(X-b)^2 &= E(X-EX+EX-b)^2\\ &= E\left[(X-EX)+(EX-b)\right]^2\\ &= E(X-EX)^2+(EX-b)^2+2E\left[(X-EX)(EX-b)\right],\end{aligned}$$

where we have expanded the square. Note that $E\left[(X-EX)(EX-b)\right] = (EX-b)E(X-EX) = 0$, since $EX-b$ is constant and comes out of the expectation, $E(X-EX) = EX-EX = 0$. This means

$$E(X-b)^2 = E(X-EX)^2+(EX-b)^2.$$

Such that $E(X-b)^2$ is minimized at $b=EX$. And $E(X-EX)^2$ is actually the variance of X ($VarX = E(X-EX)^2$).

The various moments of a distribution are an important class of expectations.

Definition For each integer n, the nth *moment* of X (or $F_X(x)$), μ_n', is

$$\mu_n' = EX^n.$$

The nth *central moment* of X, μ_n, is

$$\mu_n = E(X - \mu)^n,$$

where $\mu = \mu_1' = EX$.

Notes:

- $\mu_0' = EX^0 = 1$
- μ_1' is the *mean*, usually denoted by μ.
- $\mu_0 = E(X - \mu)^0 = 1$
- $\mu_1 = 0$
- $\mu_2 = E(X - EX)^2$ is the *variance*
- $\mu_3 = E(X - EX)^3$ is related to the *skewness*.
- $\mu_4 = E(X - EX)^4$ is related to the *kurtosis*.

Definition The *variance* of a random variable X is its second central moment, $\text{Var}\,(X) = E\left[(X - EX)^2\right]$. The positive square root of $\text{Var}\,(X)$ is the *standard deviation* of X.

The variance gives a measure of the degree of spread of a distribution around its mean. Figure 29.5 shows a plot of two samples, one sample draws 100 numbers from a normal distribution with mean 0 and variance 1, $N(0, 1)$. The other sample draws 100 numbers from a normal distribution with mean 0 and variance 100, $N(0, 100)$.

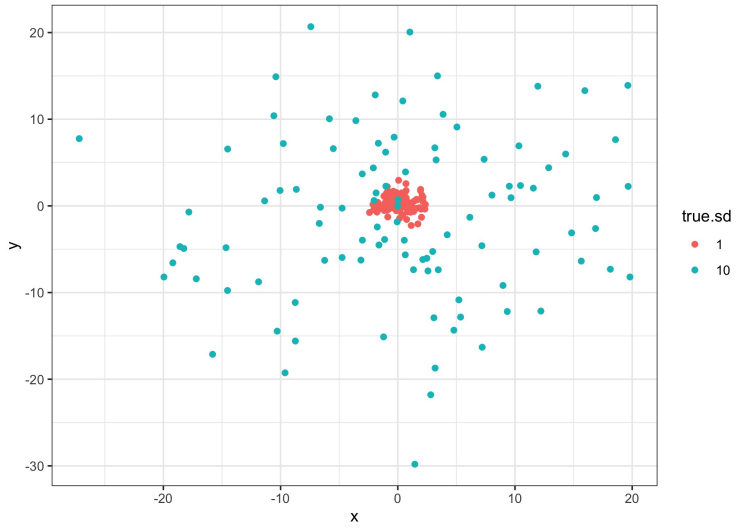

Figure 16.2: Figure 2.1.2. Two samples of 100 numbers drawn from $N(0, 1)$ and $N(0, 100)$.

Example (Exponential variance) Let X have the exponential(λ) distribution. We can calculate the variance of X now.

Solution:

$$\begin{aligned}\mathrm{Var}\,(X) &= E(X-\lambda)^2 \\ &= \int_0^\infty (x-\lambda)^2 \frac{1}{\lambda} e^{-x/\lambda} dx\end{aligned}$$

$$\begin{aligned}\mathrm{Var}\,(X) &= \int_0^\infty (x^2 - 2x\lambda + \lambda^2)\frac{1}{\lambda} e^{-x/\lambda} dx \\ &= \int_0^\infty x^2 \frac{1}{\lambda} e^{-x/\lambda} dx - 2\int_0^\infty x\lambda \frac{1}{\lambda} e^{-x/\lambda} dx + \lambda^2 \\ &= EX^2 - \lambda^2 \\ &= \lambda^2\end{aligned}$$

Theorem If X is a random variable with finite variance, then for any constants a and b,

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

Proof:
From the definition, we have

$$\begin{aligned}\mathrm{Var}\,(aX+b) &= E\left[(aX+b) - E(aX+b)\right]^2 \\ &= E(aX - aEX)^2 \\ &= a^2 E(X-EX)^2 \\ &= a^2 \mathrm{Var}\,(X)\,.\end{aligned}$$

It is sometimes to use an alternative formula for the variance, given by

$$\mathrm{Var}\,(X) = E(X^2) - (EX)^2,$$

which is easily established by

$$\begin{aligned}\mathrm{Var}\,(X) &= E(X-EX)^2 = E\left[X^2 - 2XEX + (EX)^2\right] \\ &= EX^2 - 2(EX)^2 + (EX)^2 \\ &= EX^2 - (EX)^2.\end{aligned}$$

Example (Binomial variance) Let $X \sim Binomial(n,p)$, that is ,

$$\Pr(X=x) = \binom{n}{x} p^x (1-p)^{n-x}.$$

What is the variance of X?
Solutions:

Method #1:
We want to find EX^2 first. We use the

$$EX^2 = \sum_{x=0}^{n} x^2 \binom{n}{x} p^x (1-p)^{n-x}.$$

we use the same property $x^2\binom{n}{x} = xn\binom{n-1}{x-1}$. We then have

$$\begin{aligned}
EX^2 &= n \sum_{x=1}^{n} x \binom{n-1}{x-1} p^x (1-p)^{n-x} \\
&= n \sum_{y=0}^{n-1} (y+1) \binom{n-1}{y} p^{y+1} (1-p)^{n-1-y} \\
&= np \sum_{y=0}^{n-1} y \binom{n-1}{y} p^y (1-p)^{n-1-y} + np \sum_{y=0}^{n-1} \binom{n-1}{y} p^y (1-p)^{n-1-y} \\
&= np \cdot (n-1)p + np \\
&= n(n-1)p^2 + np.
\end{aligned}$$

And now

$$\begin{aligned}
\mathrm{Var}\,(X) &= EX^2 - (EX)^2 \\
&= n(n-1)p^2 + np - (np)^2 \\
&= np - np^2 \\
&= np(1-p).
\end{aligned}$$

Method #2:
Recall that we could write $X = \sum_{i=1}^{n} I_i$, where $I_i \overset{\text{i.i.d.}}{\sim} Bernoulli(p)$. Then

$$\begin{aligned}
\mathrm{Var}\,(X) &= \mathrm{Var}\left(\sum_{i=1}^{n} I_i\right) \\
&= \sum_{i=0}^{n} \mathrm{Var}\,(I_i) \qquad (I_i\text{'s are independent}) \\
&= n\mathrm{Var}\,(I_i) \qquad (I_i\text{'s are identically distributed}) \\
&= n\left[E(I_i^2) - (EI_i)^2\right] \\
&= n\left[p - p^2\right] \\
&= np(1-p).
\end{aligned}$$

Lecture 17: Sept 30

Last time

- Exam 1 covers up to next Monday's lecture
- Expectations (2.2)
- Moments and moment generation function

Today

- Practice examples

Example A random variable X has a discrete uniform $(1, N)$ distribution, $X \sim Unif\{1, N\}$, if

$$\Pr(X = x|N) = \frac{1}{N}, \quad x = 1, 2, \ldots, N,$$

where N is a specified integer. This distribution puts equal mass on each of the outcomes $1, 2, \ldots, N$. Question: what is the cdf of this r.v.?
Solutions:

$$F(x) = \Pr(X \leqslant x|N) = \begin{cases} 0, & x < 1 \\ \frac{1}{N}, & 1 \leqslant x < 2 \\ \frac{2}{N}, & 2 \leqslant x < 3 \\ \frac{3}{N}, & 3 \leqslant x < 4 \\ \vdots & \\ \frac{N-1}{N}, & N - 1 \leqslant x < N \\ 1, & N \leqslant x \end{cases}$$

Example The continuous uniform distribution is defined by spreading mass uniformly over an interval $[a, b]$. A random variable X has a continuous uniform $[a, b]$ distribution, $X \sim Unif(a, b)$, if its pdf is given by

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

Question: what is the cdf?
Solutions:

$$F(x) = \Pr(X \leqslant x) = \begin{cases} 0, & x < a \\ \frac{x-a}{b-a}, & a \leqslant x < b \\ 1, & b \leqslant x. \end{cases}$$

Lecture 18: Oct 3

Last time

- Practice examples

Today

- Midterm 1 practice exam posted on canvas
- Moments and moment generating function

Moments

Example (Minimizing distance) The expected value of a random variable has another property, one that we can think of as relating to the interpretation of EX as a good guess at a value of X.

Suppose we measure the distance between a random variable X and a constant b by $(X-b)^2$. The closer b is to X, the smaller this quantity is. We can now determine the value of b that minimizes $E\left[(X-b)^2\right]$ and, hence, will provide us with a good predictor of X. (Note that it does no good to look for a value of b that minimizes $(X-b)^2$, since the answer would depend on X, making it a useless predictor of X.)

We could proceed with the minimization of $E(X-b)^2$ by using calculus, but there is a simpler method:

$$\begin{aligned} E(X-b)^2 &= E(X-EX+EX-b)^2 \\ &= E\left[(X-EX)+(EX-b)\right]^2 \\ &= E(X-EX)^2+(EX-b)^2+2E\left[(X-EX)(EX-b)\right], \end{aligned}$$

where we have expanded the square. Note that $E\left[(X-EX)(EX-b)\right] = (EX-b)E(X-EX) = 0$, since $EX-b$ is constant and comes out of the expectation, $E(X-EX) = EX-EX = 0$. This means

$$E(X-b)^2 = E(X-EX)^2+(EX-b)^2.$$

Such that $E(X-b)^2$ is minimized at $b = EX$. And $E(X-EX)^2$ is actually the variance of X ($VarX = E(X-EX)^2$).

The various moments of a distribution are an important class of expectations.

Definition For each integer n, the nth *moment* of X (or $F_X(x)$), μ_n', is

$$\mu_n' = EX^n.$$

The nth *central moment* of X, μ_n, is

$$\mu_n = E(X-\mu)^n,$$

where $\mu = \mu_1' = EX$.

Notes:

- $\mu_0' = EX^0 = 1$
- μ_1' is the *mean*, usually denoted by μ.
- $\mu_0 = E(X - \mu)^0 = 1$
- $\mu_1 = 0$
- $\mu_2 = E(X - EX)^2$ is the *variance*
- $\mu_3 = E(X - EX)^3$ is related to the *skewness*.
- $\mu_4 = E(X - EX)^4$ is related to the *kurtosis*.

Definition The *variance* of a random variable X is its second central moment, $\text{Var}(X) = E[(X - EX)^2]$. The positive square root of $\text{Var}(X)$ is the *standard deviation* of X.

The variance gives a measure of the degree of spread of a distribution around its mean. Figure 29.5 shows a plot of two samples, one sample draws 100 numbers from a normal distribution with mean 0 and variance 1, $N(0, 1)$. The other sample draws 100 numbers from a normal distribution with mean 0 and variance 100, $N(0, 100)$.

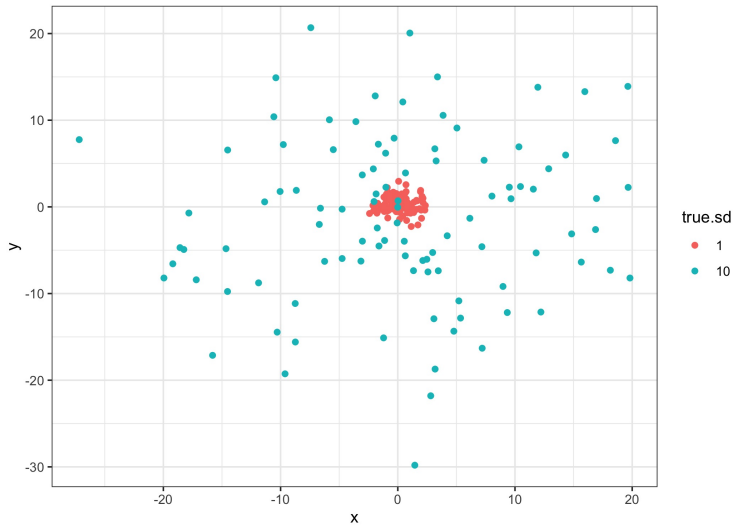

Figure 18.3: Figure 2.1.2. Two samples of 100 numbers drawn from $N(0, 1)$ and $N(0, 100)$.

Example (Exponential variance) Let X have the exponential(λ) distribution. We can calculate the variance of X now.
Solution:

$$\begin{aligned}\mathrm{Var}\,(X) &= E(X-\lambda)^2 \\ &= \int_0^\infty (x-\lambda)^2 \frac{1}{\lambda} e^{-x/\lambda} dx\end{aligned}$$

$$\begin{aligned}\mathrm{Var}\,(X) &= \int_0^\infty (x^2 - 2x\lambda + \lambda^2)\frac{1}{\lambda} e^{-x/\lambda} dx \\ &= \int_0^\infty x^2 \frac{1}{\lambda} e^{-x/\lambda} dx - 2\int_0^\infty x\lambda \frac{1}{\lambda} e^{-x/\lambda} dx + \lambda^2 \\ &= EX^2 - \lambda^2 \\ &= \lambda^2\end{aligned}$$

Theorem If X is a random variable with finite variance, then for any constants a and b,

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

Proof:
From the definition, we have

$$\begin{aligned}\mathrm{Var}\,(aX+b) &= E\left[(aX+b) - E(aX+b)\right]^2 \\ &= E(aX - aEX)^2 \\ &= a^2 E(X-EX)^2 \\ &= a^2 \mathrm{Var}\,(X)\,.\end{aligned}$$

It is sometimes to use an alternative formula for the variance, given by

$$\mathrm{Var}\,(X) = E(X^2) - (EX)^2,$$

which is easily established by

$$\begin{aligned}\mathrm{Var}\,(X) &= E(X-EX)^2 = E\left[X^2 - 2XEX + (EX)^2\right] \\ &= EX^2 - 2(EX)^2 + (EX)^2 \\ &= EX^2 - (EX)^2.\end{aligned}$$

Example (Binomial variance) Let $X \sim Binomial(n,p)$, that is ,

$$\Pr(X=x) = \binom{n}{x} p^x (1-p)^{n-x}.$$

What is the variance of X?
Solutions:
Method #1:
We want to find EX^2 first. We use the

$$EX^2 = \sum_{x=0}^{n} x^2 \binom{n}{x} p^x (1-p)^{n-x}.$$

we use the same property $x^2\binom{n}{x} = xn\binom{n-1}{x-1}$. We then have

$$\begin{aligned}
EX^2 &= n\sum_{x=1}^{n} x\binom{n-1}{x-1}p^x(1-p)^{n-x} \\
&= n\sum_{y=0}^{n-1}(y+1)\binom{n-1}{y}p^{y+1}(1-p)^{n-1-y} \\
&= np\sum_{y=0}^{n-1} y\binom{n-1}{y}p^y(1-p)^{n-1-y} + np\sum_{y=0}^{n-1}\binom{n-1}{y}p^y(1-p)^{n-1-y} \\
&= np\cdot(n-1)p + np \\
&= n(n-1)p^2 + np.
\end{aligned}$$

And now

$$\begin{aligned}
\mathrm{Var}\,(X) &= EX^2 - (EX)^2 \\
&= n(n-1)p^2 + np - (np)^2 \\
&= np - np^2 \\
&= np(1-p).
\end{aligned}$$

Method #2:
Recall that we could write $X = \sum_{i=1}^{n} I_i$, where $I_i \overset{\text{i.i.d.}}{\sim} Bernoulli(p)$. Then

$$\begin{aligned}
\mathrm{Var}\,(X) &= \mathrm{Var}\left(\sum_{i=1}^{n} I_i\right) \\
&= \sum_{i=0}^{n}\mathrm{Var}\,(I_i) \qquad (I_i\text{'s are independent}) \\
&= n\mathrm{Var}\,(I_i) \qquad (I_i\text{'s are identically distributed}) \\
&= n\left[E(I_i^2) - (EI_i)^2\right] \\
&= n\left[p - p^2\right] \\
&= np(1-p).
\end{aligned}$$

Definition Let X be a random variable with cdf F_X. The *moment generating function (mgf)* of X (or F_X), denoted by $M_X(t)$, is

$$M_X(t) = Ee^{tX},$$

provided that the expectation exists for t in some neighborhood of 0. That is, there is an $h > 0$ such that, for all t in $-h < t < h$, Ee^{tX} exists. If the expectation does not exist in a neighborhood of 0, we say that the moment generating function does not exist.

More explicitly, we can write the mgf of X as

$$M_X(t) = \int_{-\infty}^{\infty} e^{tx} f_X(x)dx, \quad \text{if } X \text{ is continuous,}$$

or

$$M_X(t) = \sum_x e^{tx} \Pr(X = x), \quad \text{if } X \text{ is discrete.}$$

It is easy to see how the mgf generates moments as in the following theorem.

Theorem If X has mgf $M_X(t)$, then

$$EX^n = M_X^{(n)}(0),$$

where we define

$$M_X^{(0)} = \frac{d^n}{dt^n} M_X(t)\bigg|_{t=0}.$$

That is, the n^{th} moment is equal to the n^{th} derivative of $M_X(t)$ evaluated at $t = 0$.

Proof:

$$\begin{aligned}
\frac{d}{dt} M_X(t) &= \frac{d}{dt} \int_{-\infty}^{\infty} e^{tx} f_X(x)dx \\
&= \int_{-\infty}^{\infty} \left(\frac{d}{dt} e^{tx} \right) f_X(x)dx \\
&= \int_{-\infty}^{\infty} \left(xe^{tx} \right) f_X(x)dx \\
&= E(Xe^{tX}).
\end{aligned}$$

Therefore,

$$\frac{d}{dt} M_X(t)\bigg|_{t=0} = E(Xe^{tX})\bigg|_{t=0} = EX.$$

Proceeding in an analogous manner, we can establish that

$$\frac{d^n}{dt^n} M_X(t)\bigg|_{t=0} = E(X^n e^{tX})\bigg|_{t=0} = EX^n.$$

Lecture 19: Oct 14

Last time

- Midterm exam 1 review

Today

- Internal midterm evaluation open
- Presentations
- Moment generating function

Definition Let X be a random variable with cdf F_X. The *moment generating function (mgf)* of X (or F_X), denoted by $M_X(t)$, is

$$M_X(t) = Ee^{tX},$$

provided that the expectation exists for t in some neighborhood of 0. That is, there is an $h > 0$ such that, for all t in $-h < t < h$, Ee^{tX} exists. If the expectation does not exist in a neighborhood of 0, we say that the moment generating function does not exist.

More explicitly, we can write the mgf of X as

$$M_X(t) = \int_{-\infty}^{\infty} e^{tx} f_X(x)dx, \quad \text{if } X \text{ is continuous,}$$

or

$$M_X(t) = \sum_x e^{tx} \Pr(X = x), \quad \text{if } X \text{ is discrete.}$$

It is easy to see how the mgf generates moments as in the following theorem.

Theorem If X has mgf $M_X(t)$, then

$$EX^n = M_X^{(n)}(0),$$

where we define

$$M_X^{(n)}(0) = \frac{d^n}{dt^n} M_X(t)\bigg|_{t=0}.$$

That is, the n^{th} moment is equal to the n^{th} derivative of $M_X(t)$ evaluated at $t = 0$.

Proof:

$$\begin{aligned}\frac{d}{dt}M_X(t) &= \frac{d}{dt}\int_{-\infty}^{\infty} e^{tx} f_X(x)dx \\ &= \int_{-\infty}^{\infty} \left(\frac{d}{dt}e^{tx}\right) f_X(x)dx \\ &= \int_{-\infty}^{\infty} \left(xe^{tx}\right) f_X(x)dx \\ &= E(Xe^{tX}).\end{aligned}$$

Therefore,

$$\frac{d}{dt}M_X(t)\bigg|_{t=0} = E(Xe^{tX})\bigg|_{t=0} = EX.$$

Proceeding in an analogous manner, we can establish that

$$\frac{d^n}{dt^n}M_X(t)\bigg|_{t=0} = E(X^n e^{tX})\bigg|_{t=0} = EX^n.$$

Example (Binomial mgf) Let $X \sim Binomial(n, p)$, then its mgf is

$$\begin{aligned}M_X(t) &= \sum_{x=0}^{n} e^{tx}\binom{n}{x}p^x(1-p)^{n-x} \\ &= \sum_{x=0}^{n}\binom{n}{x}(pe^t)^x(1-p)^{n-x} \\ &= [pe^t + (1-p)]^n.\end{aligned}$$

Theorem Let $F_X(x)$ and $F_Y(y)$ be two cdfs all of whose moments exist.

1. If X and Y have **bounded support**, then $F_X(u) = F_Y(u)$ for all u if and only if $EX^r = EY^r$ for all integers $r = 0, 1, 2, \ldots$.
2. If the moment generating functions exist and $M_X(t) = M_Y(t)$ for all t in some neighborhood of 0, then $F_X(u) = F_Y(u)$ for all u.

Theorem (Convergence of mgfs) Suppose $\{X_i, i = 1, 2, \ldots\}$ is a sequence of random variables, each with mgf $M_{X_i}(t)$. Furthermore, suppose that

$$\lim_{i\to\infty} M_{X_i}(t) = M_X(t), \quad \text{for all } t \text{ in a neighborhood of } 0,$$

and $M_X(t)$ is an mgf. Then there is a unique cdf F_X whose moments are determined by $M_X(t)$ and, for all x where $F_X(x)$ is continuous, we have

$$\lim_{i\to\infty} F_{X_i}(x) = F_X(x).$$

That is, *convergence*, for $|t| < h$, of mgfs to an mgf implies *convergence* of cdfs.

Poisson approximation One approximation that is usually taught in elementary statistics courses is that binomial probabilities can be approximated by Poisson probabilities. It is taught that the Poisson approximation is valid "when n is large and np is small", and rules of thumb are sometimes given.

The $Poisson(\lambda)$ pmf is given by

$$\Pr(X = x) = \frac{e^{-\lambda}\lambda^x}{x!}, \quad x = 0, 1, 2, \ldots,$$

where λ is a positive constant. The approximation states that if $X \sim Binomial(n, p)$ and $Y \sim Poisson(\lambda)$, with $\lambda = np$, then

$$\Pr(X = x) \approx \Pr(Y = x)$$

for large n and small np. We now show that the mgf converge, lending credence to this approximation. Recall that

$$M_X(t) = [pe^t + (1 - p)]^n.$$

For the $Poisson(\lambda)$ distribution, we can calculate (HW4, exercise 2.33)

$$M_Y(t) = e^{\lambda(e^t - 1)},$$

and if we define $p = \lambda/n$, then $M_X(t) = [1 + (e^t - 1)\lambda/n]^n$ such that $M_X(t) \to M_Y(t)$ as $n \to \infty$.

Theorem For any constant a and b, the mgf of the random variable $aX + b$ is given by

$$M_{aX+b} = e^{bt} M_X(at).$$

Proof:
By definition,

$$\begin{aligned} M_{aX+b} &= E\left(e^{(aX+b)t}\right) \\ &= E\left(e^{(aX)t} e^{bt}\right) \\ &= e^{bt} E\left(e^{(aX)t}\right) \\ &= e^{bt} M_X(at). \end{aligned}$$

Lecture 20: Oct 17

Last time

- Presentations
- Moment generating function

Today

- Internal midterm evaluation open
- Moment generating function
- Common Discrete Distributions (Chapter 3)

Definition Let X be a random variable with cdf F_X. The *moment generating function (mgf)* of X (or F_X), denoted by $M_X(t)$, is

$$M_X(t) = Ee^{tX},$$

provided that the expectation exists for t in some neighborhood of 0. That is, there is an $h > 0$ such that, for all t in $-h < t < h$, Ee^{tX} exists. If the expectation does not exist in a neighborhood of 0, we say that the moment generating function does not exist.

More explicitly, we can write the mgf of X as

$$M_X(t) = \int_{-\infty}^{\infty} e^{tx} f_X(x)dx, \quad \text{if } X \text{ is continuous,}$$

or

$$M_X(t) = \sum_x e^{tx} \Pr(X = x), \quad \text{if } X \text{ is discrete.}$$

It is easy to see how the mgf generates moments as in the following theorem.

Theorem If X has mgf $M_X(t)$, then

$$EX^n = M_X^{(n)}(0),$$

where we define

$$M_X^{(n)}(0) = \frac{d^n}{dt^n} M_X(t)\bigg|_{t=0}.$$

That is, the n^{th} moment is equal to the n^{th} derivative of $M_X(t)$ evaluated at $t = 0$.

Proof:

$$\begin{aligned}\frac{d}{dt}M_X(t) &= \frac{d}{dt}\int_{-\infty}^{\infty} e^{tx} f_X(x)dx \\ &= \int_{-\infty}^{\infty} \left(\frac{d}{dt}e^{tx}\right) f_X(x)dx \\ &= \int_{-\infty}^{\infty} \left(xe^{tx}\right) f_X(x)dx \\ &= E(Xe^{tX}).\end{aligned}$$

Therefore,

$$\frac{d}{dt}M_X(t)\bigg|_{t=0} = E(Xe^{tX})\bigg|_{t=0} = EX.$$

Proceeding in an analogous manner, we can establish that

$$\frac{d^n}{dt^n}M_X(t)\bigg|_{t=0} = E(X^n e^{tX})\bigg|_{t=0} = EX^n.$$

Example (Binomial mgf) Let $X \sim Binomial(n, p)$, then its mgf is

$$\begin{aligned}M_X(t) &= \sum_{x=0}^{n} e^{tx}\binom{n}{x}p^x(1-p)^{n-x} \\ &= \sum_{x=0}^{n}\binom{n}{x}(pe^t)^x(1-p)^{n-x} \\ &= [pe^t + (1-p)]^n.\end{aligned}$$

Theorem Let $F_X(x)$ and $F_Y(y)$ be two cdfs all of whose moments exist.

1. If X and Y have **bounded support**, then $F_X(u) = F_Y(u)$ for all u if and only if $EX^r = EY^r$ for all integers $r = 0, 1, 2, \ldots$.
2. If the moment generating functions exist and $M_X(t) = M_Y(t)$ for all t in some neighborhood of 0, then $F_X(u) = F_Y(u)$ for all u.

Theorem (Convergence of mgfs) Suppose $\{X_i, i = 1, 2, \ldots\}$ is a sequence of random variables, each with mgf $M_{X_i}(t)$. Furthermore, suppose that

$$\lim_{i\to\infty} M_{X_i}(t) = M_X(t), \quad \text{for all } t \text{ in a neighborhood of } 0,$$

and $M_X(t)$ is an mgf. Then there is a unique cdf F_X whose moments are determined by $M_X(t)$ and, for all x where $F_X(x)$ is continuous, we have

$$\lim_{i\to\infty} F_{X_i}(x) = F_X(x).$$

That is, *convergence*, for $|t| < h$, of mgfs to an mgf implies *convergence* of cdfs.

Poisson approximation One approximation that is usually taught in elementary statistics courses is that binomial probabilities can be approximated by Poisson probabilities. It is taught that the Poisson approximation is valid "when n is large and np is small", and rules of thumb are sometimes given.

The $Poisson(\lambda)$ pmf is given by

$$\Pr(X = x) = \frac{e^{-\lambda}\lambda^x}{x!}, \quad x = 0, 1, 2, \ldots,$$

where λ is a positive constant. The approximation states that if $X \sim Binomial(n, p)$ and $Y \sim Poisson(\lambda)$, with $\lambda = np$, then

$$\Pr(X = x) \approx \Pr(Y = x)$$

for large n and small np. We now show that the mgf converge, lending credence to this approximation. Recall that

$$M_X(t) = [pe^t + (1 - p)]^n.$$

For the $Poisson(\lambda)$ distribution, we can calculate (HW4, exercise 2.33)

$$M_Y(t) = e^{\lambda(e^t - 1)},$$

and if we define $p = \lambda/n$, then $M_X(t) = [1 + (e^t - 1)\lambda/n]^n$ such that $M_X(t) \to M_Y(t)$ as $n \to \infty$.

Theorem For any constant a and b, the mgf of the random variable $aX + b$ is given by

$$M_{aX+b} = e^{bt} M_X(at).$$

Proof:
By definition,

$$\begin{aligned} M_{aX+b} &= E\left(e^{(aX+b)t}\right) \\ &= E\left(e^{(aX)t} e^{bt}\right) \\ &= e^{bt} E\left(e^{(aX)t}\right) \\ &= e^{bt} M_X(at). \end{aligned}$$

Common Discrete Distribution

Why parametric models?

- *Parametric models* or *distribution families* have a specific form but can change according to a fixed number of parameters.
- The objective is to model a population. Parametric models are often appropriate in common situations with similar mechanisms.
- Parametric models have many known and useful properties and are easy to work with. When fitting a population, only a few parameters need to be estimated: *parametric inference.*

- Sometimes one does not want to make parametric assumptions and would rather work with non-parametric models. But non-parametric models can be infinite dimensional.
- In this course, we emphasize parametric models.

Discrete uniform X has the discrete unifrom$(1, N)$ distribution if X is equally likely to be one of $\{1, 2, \ldots, N\}$.

- Sample space: $\{1, 2, \ldots, N\}$
- pmf:
$$f_X(x) = \frac{1}{N}, \quad x = 1, 2, \ldots, N$$
- cdf:
$$F_X(x) = \Pr(X \leqslant x) = \begin{cases} 0 & x < 0 \\ \lfloor x \rfloor / N & 0 \leqslant x < N \\ 1 & N \leqslant x \end{cases}$$
- moments:
$$EX = \frac{N+1}{2}$$

Bernoulli Distribution Consider an experiment where outcomes are binary (say, Success or Failure) and the probability of success is p. Define the following random variable

$$Y = \begin{cases} 1 & \text{outcome is success} \\ 0 & \text{outcome is failure} \end{cases}$$

Then, Y has a Bernoulli Distribution.

- Sample space: $\{0, 1\}$.
- pmf: $\Pr(Y = 1) = p$ and $\Pr(Y = 0) = 1 - p$. We can write this as:
$$f(y) = \Pr(Y = y) = \begin{cases} p^y (1-p)^{1-y} & y = 0, 1 \\ 0 & othersie \end{cases}$$
- what are the cdf, mean and variance?

Binomial Distribution A $Binomial(n, p)$ random variable X is defined as the number of successes in n i.i.d. (independent, identically distributed) Bernoulli trials, each with probability p of success:

$$X = \sum_{i=1}^{n} Y_i, \quad Y_1, \ldots, Y_n \overset{\text{i.i.d.}}{\sim} Bernoulli(p)$$

- Sample space: $\{0, 1, \ldots, n\}$

- pmf:

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

- cdf:

$$F_X(x) = \sum_{s=0}^{x} \binom{n}{s} p^s (1-p)^{n-s} \quad \text{(no closed form)}$$

Poisson Distribution The Poisson distribution was derived by the French mathematician Poisson in 1837 as a limiting version of the binomial distribution. The Poisson distribution is often used to model the number of occurrences in a given time interval. One of the basic assumptions on which the Poisson distribution is built is that, for small time intervals, the probability of an arrival is proportional to the length of waiting time. This makes it a reasonable model for situations such as waiting for a bus, waiting for customers to arrive in a bank.

The Poisson distribution has a single parameter λ, sometimes called the intensity parameter. A Poisson random variable X, takes values in the nonnegative integers with pmf

$$\Pr(X = x|\lambda) = \frac{e^{-\lambda}\lambda^x}{x!}, \quad x = 0, 1, \ldots$$

To see that $\sum_{x=0}^{\infty} P(X = x|\lambda) = 1$, recall the Taylor series expansion of $e^{\lambda} = \sum_{i=0}^{\infty} \frac{\lambda^i}{i!}$. Thus

$$\sum_{x=0}^{\infty} \Pr(X = x|\lambda) = e^{-\lambda} \sum_{x=0}^{\infty} \frac{\lambda^x}{x!} = e^{-\lambda} e^{\lambda} = 1$$

What is the mean and variance of X?

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

Similarly

$$\begin{aligned} EX^2 &= \sum_{x=0}^{\infty} x^2 \frac{e^{-\lambda}\lambda^x}{x!} \\ &= \sum_{x=1}^{\infty} x \frac{e^{-\lambda}\lambda^x}{(x-1)!} \\ &= \sum_{x=1}^{\infty} \frac{e^{-\lambda}\lambda^x}{(x-1)!} + \sum_{x=2}^{\infty} \frac{e^{-\lambda}\lambda^x}{(x-2)!} \\ &= \lambda + \lambda^2 \end{aligned}$$

So that

$$Var(X) = EX^2 - (EX)^2 = \lambda$$

- Sample space: $\{0, 1, \dots\}$
- pmf: $\Pr(X = x) = \frac{e^{-\lambda}\lambda^x}{x!}$
- cdf: $F_X(x) = \sum_{s=0}^{x} \frac{e^{-\lambda}\lambda^s}{s!}$

Lecture 21: Oct 19

Last time

- Internal midterm evaluation open
- Moment generating function

Today

- Common Discrete Distributions (Chapter 3)

Binomial Distribution A $Binomial(n, p)$ random variable X is defined as the number of successes in n i.i.d. (independent, identically distributed) Bernoulli trials, each with probability p of success:

$$X = \sum_{i=1}^{n} Y_i, \quad Y_1, \ldots, Y_n \overset{\text{i.i.d.}}{\sim} Bernoulli(p)$$

- Sample space: $\{0, 1, \ldots, n\}$
- pmf:

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

- cdf:

$$F_X(x) = \sum_{s=0}^{x} \binom{n}{s} p^s (1-p)^{n-s} \quad \text{(no closed form)}$$

Poisson Distribution The Poisson distribution was derived by the French mathematician Poisson in 1837 as a limiting version of the binomial distribution. The Poisson distribution is often used to model the number of occurrences in a given time interval. One of the basic assumptions on which the Poisson distribution is built is that, for small time intervals, the probability of an arrival is proportional to the length of waiting time. This makes it a reasonable model for situations such as waiting for a bus, waiting for customers to arrive in a bank.

The Poisson distribution has a single parameter λ, sometimes called the intensity parameter. A Poisson random variable X, takes values in the nonnegative integers with pmf

$$\Pr(X = x|\lambda) = \frac{e^{-\lambda}\lambda^x}{x!}, \quad x = 0, 1, \ldots$$

To see that $\sum_{x=0}^{\infty} P(X = x|\lambda) = 1$, recall the Taylor series expansion of $e^{\lambda} = \sum_{i=0}^{\infty} \frac{\lambda^i}{i!}$. Thus

$$\sum_{x=0}^{\infty} \Pr(X = x|\lambda) = e^{-\lambda} \sum_{x=0}^{\infty} \frac{\lambda^x}{x!} = e^{-\lambda} e^{\lambda} = 1$$

What is the mean and variance of X?

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

Similarly

$$\begin{aligned} EX^2 &= \sum_{x=0}^{\infty} x^2 \frac{e^{-\lambda}\lambda^x}{x!} \\ &= \sum_{x=1}^{\infty} x \frac{e^{-\lambda}\lambda^x}{(x-1)!} \\ &= \sum_{x=1}^{\infty} \frac{e^{-\lambda}\lambda^x}{(x-1)!} + \sum_{x=2}^{\infty} \frac{e^{-\lambda}\lambda^x}{(x-2)!} \\ &= \lambda + \lambda^2 \end{aligned}$$

So that

$$Var(X) = EX^2 - (EX)^2 = \lambda$$

- Sample space: $\{0, 1, \dots\}$
- pmf: $\Pr(X = x) = \frac{e^{-\lambda}\lambda^x}{x!}$
- cdf: $F_X(x) = \sum_{s=0}^{x} \frac{e^{-\lambda}\lambda^s}{s!}$

Hypergeometric Distribution Suppose a population of N entities is made up of two types: M of the first type and $N - M$ of the second type. Suppose we take a sample of size K. We wish to know X, the number in the sample of the first type. The probability mass function of X is given by:

$$f_X(x) = \frac{\binom{M}{x}\binom{N-M}{k-x}}{\binom{N}{k}}$$

for $x = \max(0, M - N + K), \dots, \min(M, K)$.

The sample space is defined so that all binomial coefficients are valid. We must have:

$$0 \leqslant x \leqslant K, \quad 0 \leqslant x \leqslant M, \quad 0 \leqslant K - x \leqslant N - M$$

Often $K < M$ and $K < N - M$ so the range becomes $0 \leqslant x \leqslant K$.

Hypergeometric vs Binomial We can show that the limiting form of the hypergeometric pmf is the binomial pmf

$$\begin{aligned}\Pr(s) &= \frac{\binom{M}{s}\binom{N-M}{n-s}}{\binom{N}{n}} \\ &= \frac{\frac{M!}{s!(M-s)!}\frac{(N-M)!}{(n-s)!(N-M-n+s)!}}{\frac{N!}{n!(N-n)!}} \\ &= \frac{\frac{n!}{s!(n-s)!}\frac{M!}{(M-s)!}\frac{(N-M)!}{(N-M-n+s)!}}{\frac{N!}{(N-n)!}}\end{aligned}$$

Note

$$\begin{aligned}\frac{M!}{(M-s)!} &= \frac{M(M-1)(M-2)\dots(M-s)!}{(M-s)!} \\ &= M^s\left[1(1-\frac{1}{M})\dots(1-\frac{s-1}{M})\right] \\ \frac{N!}{(N-n)!} &= N^n\left[1(1-\frac{1}{N})\dots(1-\frac{n-1}{N})\right] \\ \frac{(N-M)!}{[(N-M)-(n-s)]!} &= (N-M)^{n-s}\left[1(1-\frac{1}{N-M})\dots(1-\frac{n-s-1}{N-M})\right]\end{aligned}$$

Letting $N \to \infty, M \to \infty, \frac{M}{N} \to p$, we have

$$\begin{aligned}\Pr(s) &= \frac{\binom{M}{s}\binom{N-M}{n-s}}{\binom{N}{s}} \\ &\approx \binom{n}{s}\frac{M^s(N-M)^{n-s}}{N^n} \\ &= \binom{n}{s}\left(\frac{M}{N}\right)^s\left(1-\frac{M}{N}\right)^{n-s} \\ &\to \binom{n}{s}p^s(1-p)^{n-s}\end{aligned}$$

In summary, we have

Hypergeometric	$\to$	Binomial	$\to$	Poisson
$N \to \infty$		$n \to \infty$		$\lambda = np$
$M \to \infty$		$p \to 0$		
$\frac{M}{N} \to p$		$np \to \lambda$		

Geometric Distribution Consider a series of iid Bernoulli Trials with p = probability of success in each trial. Define a random variable X representing the number of trials until first success. Note X includes the trial at which the success occurs (one parameterization). Then, X has a geometric distribution.

- Sample space: $\{1, 2, \dots\}$
- pmf:
$$f(x) = \Pr(X = x) = \begin{cases} p(1-p)^{x-1} & x = 1, 2, \dots \\ 0 & otherwise \end{cases}$$
- cdf:
$$F(x) = \Pr(X \leqslant x) = 1 - (1-p)^x$$
- Moments:
$$E(X) = 1/p$$
$$Var(X) = (1-p)/p^2$$

Memoryless property. Suppose $k > i$, then

$$\Pr(X > k | X > i) = \Pr(X > k - i)$$

Proof:

$$\begin{aligned} \Pr(X > k | X > i) &= \frac{\Pr(X > k)}{\Pr(X > i)} = \frac{(1-p)^k}{(1-p)^i} \\ &= (1-p)^{k-i} = \Pr(X > k - i) \end{aligned}$$

Example Suppose X is number of years you live, and X follows a geometric distribution, then

$$\begin{aligned} \Pr(\text{survive two more years}) &= \Pr(X > \text{current age} + 2 | X > \text{current age}) \\ &= \Pr(X > 2) \end{aligned}$$

This model is clearly too simple for human populations (since we do age).

Negative Binomial Distribution Still in the context of iid Bernoulli trials, define a random variable corresponding to the number of trials required to have s successes. We say $X \sim Negbin(s, p)$.

- Sample space: $\{s, (s+1), \dots\}$
- pmf: for $x = s, s+1, s+2, \dots$
$$\begin{aligned} f(x) &= \binom{x-1}{s-1} p^{s-1} q^{x-s} \cdot p \\ &= \binom{x-1}{s-1} p^s q^{x-s} \end{aligned}$$
- cdf: no closed form
- Expectation: $EX = s/p$.
- Variance: $Var(X) = s(1-p)/p^2$

Notes

- Why the name? See Casella & Berger p.95.
- $X \sim Negbin(1, p)$ is the same as $X \sim Geometric(p)$
- $Negbin(n, p)$ is the same as the sum of n $Geometric(p)$ random variables

Other parameterizations The negative binomial distribution is sometimes defined in terms of the random variable Y = number of failures before the rth success. Then

- Sample space: $\{0, 1, 2, \dots\}$
- pmf
$$f(y) = \binom{r+y-1}{y} p^r q^y, \quad y = 0, 1, 2, \dots$$
- cdf: no closed form
- Expectation: $EY = r(1-p)/p$
- Variance: $Var(Y) = r(1-p)/p^2$

Negarive binomial vs. Poisson The negative binomial distribution is often good for modeling count data as an alternative to the Poisson. In the previous parameterization, define

$$\lambda = \frac{r(1-p)}{p} \iff p = \frac{r}{r+\lambda}$$

Then we have

$$EX = \lambda$$
$$Var(X) = \frac{\lambda}{p} = \lambda(1 + \frac{\lambda}{r}) = \lambda + \frac{\lambda^2}{r}$$

For the Poisson we had that the variance equals the mean.
For the negative binomial, the variance is equal to the mean plus a quadratic term. Thus the negative binomial can capture overdispersion in count data.
In the previous parameterization, the pmf becomes

$$\begin{aligned} f(y) = \binom{r+y-1}{y} p^r q^y &= \frac{(r+y-1)!}{y!(r-1)!} \left(\frac{r}{r+\lambda}\right)^s \left(\frac{\lambda}{r+\lambda}\right)^y \\ &= \frac{\lambda^x}{x!} \frac{s(s+1)\dots(s+x-1)}{(s+\lambda)^x} \left(1 + \frac{\lambda}{s}\right)^{-s} \end{aligned}$$

Letting $s \to \infty$, we get

$$f(x) \to \frac{\lambda^x}{x!} e^{-\lambda}$$

So for large s, the negative binomial can be approximated by a Poisson with parameter $\lambda = r(1-p)/p$.

Lecture 22: Oct 21

Last time

- Common Discrete Distributions (Chapter 3)

Today

- Presentations
- Common Discrete Distributions (Chapter 3)

Geometric Distribution Consider a series of iid Bernoulli Trials with p = probability of success in each trial. Define a random variable X representing the number of trials until first success. Note X includes the trial at which the success occurs (one parameterization). Then, X has a geometric distribution.

- Sample space: $\{1, 2, \dots\}$
- pmf:
$$f(x) = \Pr(X = x) = \begin{cases} p(1-p)^{x-1} & x = 1, 2, \dots \\ 0 & otherwise \end{cases}$$
- cdf:
$$F(x) = \Pr(X \leqslant x) = 1 - (1-p)^x$$
- Moments:
$$E(X) = 1/p$$
$$Var(X) = (1-p)/p^2$$

Memoryless property. Suppose $k > i$, then

$$\Pr(X > k | X > i) = \Pr(X > k - i)$$

Proof:

$$\begin{aligned} \Pr(X > k | X > i) &= \frac{\Pr(X > k)}{\Pr(X > i)} = \frac{(1-p)^k}{(1-p)^i} \\ &= (1-p)^{k-i} = \Pr(X > k - i) \end{aligned}$$

Example Suppose X is number of years you live, and X follows a geometric distribution, then

$$\begin{aligned} \Pr(\text{survive two more years}) &= \Pr(X > \text{current age} + 2 | X > \text{current age}) \\ &= \Pr(X > 2) \end{aligned}$$

This model is clearly too simple for human populations (since we do age).

Negative Binomial Distribution Still in the context of iid Bernoulli trials, define a random variable corresponding to the number of trials required to have s successes. We say $X \sim Negbin(s, p)$.

- Sample space: $\{s, (s+1), \dots\}$
- pmf: for $x = s, s+1, s+2, \dots$

$$\begin{aligned} f(x) &= \binom{x-1}{s-1} p^{s-1} q^{x-s} \cdot p \\ &= \binom{x-1}{s-1} p^s q^{x-s} \end{aligned}$$

- cdf: no closed form
- Expectation: $EX = s/p$.
- Variance: $Var(X) = s(1-p)/p^2$

Notes

- Why the name? See Casella & Berger p.95.
- $X \sim Negbin(1, p)$ is the same as $X \sim Geometric(p)$
- $Negbin(n, p)$ is the same as the sum of n $Geometric(p)$ random variables

Other parameterizations The negative binomial distribution is sometimes defined in terms of the random variable Y = number of failures before the rth success. Then

- Sample space: $\{0, 1, 2, \dots\}$
- pmf

$$f(y) = \binom{r+y-1}{y} p^r q^y, \quad y = 0, 1, 2, \dots$$

- cdf: no closed form
- Expectation: $EY = r(1-p)/p$
- Variance: $Var(Y) = r(1-p)/p^2$

Negarive binomial vs. Poisson The negative binomial distribution is often good for modeling count data as an alternative to the Poisson. In the previous parameterization, define

$$\lambda = \frac{r(1-p)}{p} \iff p = \frac{r}{r+\lambda}$$

Then we have

$$\begin{aligned} EX &= \lambda \\ Var(X) &= \frac{\lambda}{p} = \lambda(1 + \frac{\lambda}{r}) = \lambda + \frac{\lambda^2}{r} \end{aligned}$$

For the Poisson we had that the variance equals the mean.
For the negative binomial, the variance is equal to the mean plus a quadratic term. Thus the negative binomial can capture overdispersion in count data.
In the previous parameterization, the pmf becomes

$$\begin{aligned} f(y) &= \binom{r+y-1}{y} p^r q^y = \frac{(r+y-1)!}{y!(r-1)!} \left(\frac{r}{r+\lambda}\right)^s \left(\frac{\lambda}{r+\lambda}\right)^y \\ &= \frac{\lambda^x}{x!} \frac{s(s+1)\dots(s+x-1)}{(s+\lambda)^x} \left(1+\frac{\lambda}{s}\right)^{-s} \end{aligned}$$

Letting $s \to \infty$, we get

$$f(x) \to \frac{\lambda^x}{x!} e^{-\lambda}$$

So for large s, the negative binomial can be approximated by a Poisson with parameter $\lambda = r(1-p)/p$.

Lecture 23: Oct 24

Last time

- Common Discrete Distributions (Chapter 3)

Today

- Will start taking attendance (no punishment)
- Example application of what you learn
- Negative binomial distribution
- Common Continuous Distributions

Negative Binomial Distribution Still in the context of iid Bernoulli trials, define a random variable corresponding to the number of trials required to have s successes. We say $X \sim Negbin(s, p)$.

- Sample space: $\{s, (s+1), \dots\}$
- pmf: for $x = s, s+1, s+2, \dots$

$$\begin{aligned} f(x) &= \binom{x-1}{s-1} p^{s-1} q^{x-s} \cdot p \\ &= \binom{x-1}{s-1} p^s q^{x-s} \end{aligned}$$

- cdf: no closed form
- Expectation: $EX = s/p$.
- Variance: $Var(X) = s(1-p)/p^2$

Notes

- Why the name? See Casella & Berger p.95.
- $X \sim Negbin(1, p)$ is the same as $X \sim Geometric(p)$
- $Negbin(n, p)$ is the same as the sum of n $Geometric(p)$ random variables

Other parameterizations The negative binomial distribution is sometimes defined in terms of the random variable Y = number of failures before the rth success. Then

- Sample space: $\{0, 1, 2, \dots\}$
- pmf

$$f(y) = \binom{r+y-1}{y} p^r q^y, \quad y = 0, 1, 2, \dots$$

- cdf: no closed form
- Expectation: $EY = r(1-p)/p$
- Variance: $Var(Y) = r(1-p)/p^2$

Negarive binomial vs. Poisson The negative binomial distribution is often good for modeling count data as an alternative to the Poisson. In the previous parameterization, define

$$\lambda = \frac{r(1-p)}{p} \iff p = \frac{r}{r+\lambda}$$

Then we have

$$EX = \lambda$$
$$Var(X) = \frac{\lambda}{p} = \lambda(1 + \frac{\lambda}{r}) = \lambda + \frac{\lambda^2}{r}$$

For the Poisson we had that the variance equals the mean.
For the negative binomial, the variance is equal to the mean plus a quadratic term. Thus the negative binomial can capture overdispersion in count data.
In the previous parameterization, the pmf becomes

$$f(y) = \binom{r+y-1}{y} p^r q^y = \frac{(r+y-1)!}{y!(r-1)!} \left(\frac{r}{r+\lambda}\right)^r \left(\frac{\lambda}{r+\lambda}\right)^y$$
$$= \frac{\lambda^y}{y!} \frac{r(r+1)\dots(r+y-1)}{(r+\lambda)^y} \left(1 + \frac{\lambda}{r}\right)^{-r}$$

Letting $r \to \infty$, we get

$$f(x) \to \frac{\lambda^x}{x!} e^{-\lambda}$$

So for large r, the negative binomial can be approximated by a Poisson with parameter $\lambda = r(1-p)/p$.

Common continuous distributions

Uniform Distribution A random variable X having a pdf

$$f(x) = \begin{cases} 1 & \text{for } 0 < x \leqslant 1 \\ 0 & otherwise \end{cases}$$

is said to have a *uniform distribution* over the interval $(0, 1)$.

The cdf is:

$$F(y) = \int_{-\infty}^{y} f(x)dx = \begin{cases} 0 & \text{for } y \leqslant 0 \\ y & \text{for } 0 \leqslant y \leqslant 1 \\ 1 & \text{for } y > 1 \end{cases}$$

- Unifrom; $Y \sim U[a, b]$

- sample space: $[a, b]$
- pdf:

$$f(y) = \begin{cases} \frac{1}{b-a} & \text{for } a < y \leqslant b \\ 0 & otherwise \end{cases}$$

- cdf:

$$F(y) = \int_{-\infty}^{y} f(x)dx = \begin{cases} 0 & \text{for } y \leqslant a \\ \frac{y-a}{b-a} & \text{for } a \leqslant y \leqslant b \\ 1 & \text{for } y > b \end{cases}$$

- moments:

$$E(Y) = (a+b)/2$$
$$Var(Y) = \frac{(b-a)^2}{12}$$

Notes

- The uniform extends to the continuous case the idea of equally likely outcomes.
- If $Y \sim U[0, 1]$, then $a + (b - a)Y \sim U[a, b]$

Exponential Distribution Denoted $X \sim Exp(\lambda)$:

- sample space: $x \geqslant 0$
- pdf:

$$f(x) = \begin{cases} \lambda e^{-\lambda y} & \text{for } y \geqslant 0 \\ 0 & otherwise \end{cases}$$

- cdf:

$$F(x) = \int_{-\infty}^{x} f(y)dy = \begin{cases} 1 - e^{-\lambda x} & \text{for } x \geqslant 0 \\ 0 & \text{for } x < 0 \end{cases}$$

- moments:

$$E(X) = 1/\lambda$$
$$Var(X) = 1/\lambda^2$$
$$M_X(t) = \lambda/(\lambda - t), \quad t < \lambda$$

Interpretation The exponential can be derived as the waiting time between Poisson events. Suppose that the number of events in a unit interval of time follows a Poisson(λ) distribution. Then, let Y be the time until the first event.

$$\Pr(Y > t) = \Pr(0 \text{ events in } [0, t])$$

and the number of events in $[0, t]$ follows a Poisson distribution with parameter λt. Therefore,

$$\Pr(Y > t) = e^{-\lambda t}.$$

The cdf of Y is

$$F(t) = 1 - \Pr(Y > t) = 1 - e^{-\lambda t}$$

and hence the density is $f(t) = \lambda e^{-\lambda t}$.

Alternative parameterization Many books write the density as

$$f(y) = \begin{cases} \frac{1}{\theta} e^{-y/\theta} & \text{for } y \geqslant 0 \\ 0 & otherwise \end{cases}$$

so that $E(Y) = \theta$ and $Var(Y) = \theta^2$. In this case $\theta = 1/\lambda$ is called the *mean parameter*, while $\lambda = 1/\theta$ is called the *rate parameter*.

Memoryless property The exponential has a memoryless property, just like the geometric.

$$\Pr(Y > s + t | Y > t) = \Pr(Y > s)$$

Same interpretation as the geometric for continuous time:

- The probability of an event in a time interval depends only on the length of the interval, not the absolute time of the interval.
- The underlying Poisson process is stationary: the rate λ is constant. (In the geometric case, the probability, p of getting an event in every discrete time unit is constant).

Shifted exponential Let $X \sim Exp(\lambda)$ and $Y = X + v, v \in \mathbb{R}$. Then, Y has the *shifted exponential distribution* with pdf:

$$f(y) = \begin{cases} \lambda e^{-(y-v)\lambda} & \text{for } y \geqslant v \\ 0 & otherwise \end{cases}$$

Interpretation:

- $v > 0$: Event is delayed
- $v < 0$: The news of the event is delayed

Does the shifted expoenential maintain the memoryless property?

Double exponential The *double exponential distribution* is formed by reflecting an exponential distribution around zero. It has pdf:

$$f(x) = \frac{1}{2} \lambda e^{-\lambda |x|}, \quad x \in \mathbb{R}$$

Suppose X has the above distribution with $\lambda = 1$. Now let $Y = \sigma X + \mu, \mu \in \mathbb{R}$ (shifting) and $\sigma > 0$ (scaling). Then Y has the *Laplace distribution* with pdf:

$$f_Y(y) = \frac{1}{2\sigma} \exp\left(-\frac{|y - \mu|}{\sigma} \right)$$

with moments

$$EY = \mu, \quad Var(Y) = 2\sigma^2$$

The Laplace distribution provides an alternative to the normal for centered data with fatter tails but all finite moments.

Lecture 24: Oct 26

Last time

- Common Discrete Distributions (Chapter 3)

Today

- Negative binomial distribution
- Common Continuous Distributions

Negarive binomial vs. Poisson The negative binomial distribution is often good for modeling count data as an alternative to the Poisson. In the previous parameterization, define

$$\lambda = \frac{r(1-p)}{p} \iff p = \frac{r}{r+\lambda}$$

Then we have

$$EX = \lambda$$
$$Var(X) = \frac{\lambda}{p} = \lambda(1 + \frac{\lambda}{r}) = \lambda + \frac{\lambda^2}{r}$$

For the Poisson we had that the variance equals the mean.
For the negative binomial, the variance is equal to the mean plus a quadratic term. Thus the negative binomial can capture overdispersion in count data.
In the previous parameterization, the pmf becomes

$$\begin{aligned} f(y) &= \binom{r+y-1}{y} p^r q^y = \frac{(r+y-1)!}{y!(r-1)!} \left(\frac{r}{r+\lambda}\right)^r \left(\frac{\lambda}{r+\lambda}\right)^y \\ &= \frac{\lambda^y}{y!} \frac{r(r+1)\dots(r+y-1)}{(r+\lambda)^y} \left(1+\frac{\lambda}{r}\right)^{-r} \end{aligned}$$

Letting $r \to \infty$, we get

$$f(x) \to \frac{\lambda^x}{x!} e^{-\lambda}$$

So for large r, the negative binomial can be approximated by a Poisson with parameter $\lambda = r(1-p)/p$.

Common continuous distributions

Uniform Distribution A random variable X having a pdf

$$f(x) = \begin{cases} 1 & \text{for } 0 < x \leqslant 1 \\ 0 & otherwise \end{cases}$$

is said to have a *uniform distribution* over the interval $(0, 1)$.

The cdf is:

$$F(y) = \int_{-\infty}^{y} f(x)dx = \begin{cases} 0 & \text{for } y \leqslant 0 \\ y & \text{for } 0 \leqslant y \leqslant 1 \\ 1 & \text{for } y > 1 \end{cases}$$

- Unifrom; $Y \sim U[a, b]$
- sample space: $[a, b]$
- pdf:

$$f(y) = \begin{cases} \frac{1}{b-a} & \text{for } a < y \leqslant b \\ 0 & otherwise \end{cases}$$

- cdf:

$$F(y) = \int_{-\infty}^{y} f(x)dx = \begin{cases} 0 & \text{for } y \leqslant a \\ \frac{y-a}{b-a} & \text{for } a \leqslant y \leqslant b \\ 1 & \text{for } y > b \end{cases}$$

- moments:

$$E(Y) = (a + b)/2$$
$$Var(Y) = \frac{(b-a)^2}{12}$$

Notes

- The uniform extends to the continuous case the idea of equally likely outcomes.
- If $Y \sim U[0, 1]$, then $a + (b - a)Y \sim U[a, b]$

Exponential Distribution Denoted $X \sim Exp(\lambda)$:

- sample space: $x \geqslant 0$
- pdf:

$$f(x) = \begin{cases} \lambda e^{-\lambda y} & \text{for } y \geqslant 0 \\ 0 & otherwise \end{cases}$$

- cdf:

$$F(x) = \int_{-\infty}^{x} f(y)dy = \begin{cases} 1 - e^{-\lambda x} & \text{for } x \geqslant 0 \\ 0 & \text{for } x < 0 \end{cases}$$

- moments:

$$E(X) = 1/\lambda$$
$$Var(X) = 1/\lambda^2$$
$$M_X(t) = \lambda/(\lambda - t), \quad t < \lambda$$

Interpretation The exponential can be derived as the waiting time between Poisson events. Suppose that the number of events in a unit interval of time follows a Poisson(λ) distribution. Then, let Y be the time until the first event.

$$\Pr(Y > t) = \Pr(0 \text{ events in } [0, t])$$

and the number of events in $[0, t]$ follows a Poisson distribution with parameter λt. Therefore,

$$\Pr(Y > t) = e^{-\lambda t}.$$

The cdf of Y is

$$F(t) = 1 - \Pr(Y > t) = 1 - e^{-\lambda t}$$

and hence the density is $f(t) = \lambda e^{-\lambda t}$.

Alternative parameterization Many books write the density as

$$f(y) = \begin{cases} \frac{1}{\theta} e^{-y/\theta} & \text{for } y \geqslant 0 \\ 0 & otherwise \end{cases}$$

so that $E(Y) = \theta$ and $Var(Y) = \theta^2$. In this case $\theta = 1/\lambda$ is called the *mean parameter*, while $\lambda = 1/\theta$ is called the *rate parameter*.

Memoryless property The exponential has a memoryless property, just like the geometric.

$$\Pr(Y > s + t | Y > t) = \Pr(Y > s)$$

Same interpretation as the geometric for continuous time:

- The probability of an event in a time interval depends only on the length of the interval, not the absolute time of the interval.
- The underlying Poisson process is stationary: the rate λ is constant. (In the geometric case, the probability, p of getting an event in every discrete time unit is constant).

Shifted exponential Let $X \sim Exp(\lambda)$ and $Y = X + v, v \in \mathbb{R}$. Then, Y has the *shifted exponential distribution* with pdf:

$$f(y) = \begin{cases} \lambda e^{-(y-v)\lambda} & \text{for } y \geqslant v \\ 0 & otherwise \end{cases}$$

Interpretation:

- $v > 0$: Event is delayed
- $v < 0$: The news of the event is delayed

Does the shifted expoenential maintain the memoryless property?

Double exponential The *double exponential distribution* is formed by reflecting an exponential distribution around zero. It has pdf:

$$f(x) = \frac{1}{2}\lambda e^{-\lambda|x|}, \quad x \in \mathbb{R}$$

Suppose X has the above distribution with $\lambda = 1$. Now let $Y = \sigma X + \mu, \mu \in \mathbb{R}$ (shifting) and $\sigma > 0$ (scaling). Then Y has the *Laplace distribution* with pdf:

$$f_Y(y) = \frac{1}{2\sigma} \exp\left(-\frac{|y-\mu|}{\sigma}\right)$$

with moments

$$EY = \mu, \quad Var(Y) = 2\sigma^2$$

The Laplace distribution provides an alternative to the normal for centered data with fatter tails but all finite moments.

Lecture 25: Oct 28

Last time

- Common Continuous Distributions

Today

- Common Continuous Distributions

Normal Distribution Introduced by De Moivre (1667 - 1754) in 1733 as an approximation to the binomial. Later studied by Laplace and others as part of the Central Limit Theorem. Gauss derived the normal as a suitable distribution for outcomes that could be thought of as sums of many small deviations.

- Sample space: $\mathbb{R} = (-\infty, \infty)$
- pdf: For $Y \sim N(\mu, \sigma^2)$,

$$f(y) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y-\mu)^2}{2\sigma^2}} \quad -\infty < y < \infty$$

- cdf: There is no closed form.
- When $\mu = 0$ and $\sigma = 1$, the distribution is called *standard normal*:

$$\Phi(y) = \Pr(Y \leqslant y), \quad \Phi(-y) = 1 - \Phi(y)$$

- Mean:

$$EY = \mu$$

- Variance:

$$Var(Y) = E(Y - \mu)^2 = \sigma^2$$

- Higher central moments:

$$E(Y-\mu)^m = \begin{cases} \frac{m!}{2^{m/2}(m/2)!}\sigma^m & m \text{ is even} \\ 0 & m \text{ is odd} \end{cases}$$

- In particular:

$$\mu_3 = E(Y-\mu)^3 = 0 (\text{Skewness})$$
$$\mu_4 = E(Y-\mu)^4 = 3\sigma^4$$

- Moment generating function:

$$M_Y(t) = \exp(\mu t + \sigma^2 t^2/2)$$

Standardization

$$Y \sim N(\mu, \sigma^2) \iff Z = \frac{Y - \mu}{\sigma} \sim N(0, 1)$$

Shifting and scaling:

$$Z \sim N(0, 1) \iff Y = \sigma Z + \mu \sim N(\mu, \sigma^2)$$

Notes

- Normal distribution is useful in many practical settings. E.g. measurement error.
- Plays an important role in *sampling distributions* in *large samples*, since the Central Limit Theorem syas that the sums of independent identically distributed random variables are approximately normal
- There are many important distributions that can be derived from functions of normal random variables (e.g. χ^2, t, F). We will briefly present the pdf's and sample spaces of these distributions.

Lecture 26: Oct 31

Last time

- Common Continuous Distributions

Today

- Common Continuous Distributions

Normal Distribution Introduced by De Moivre (1667 - 1754) in 1733 as an approximation to the binomial. Later studied by Laplace and others as part of the Central Limit Theorem. Gauss derived the normal as a suitable distribution for outcomes that could be thought of as sums of many small deviations.

- Sample space: $\mathbb{R} = (-\infty, \infty)$
- pdf: For $Y \sim N(\mu, \sigma^2)$,

$$f(y) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y-\mu)^2}{2\sigma^2}} \quad -\infty < y < \infty$$

- cdf: There is no closed form.
- When $\mu = 0$ and $\sigma = 1$, the distribution is called *standard normal*:

$$\Phi(y) = \Pr(Y \leqslant y), \quad \Phi(-y) = 1 - \Phi(y)$$

- Mean:

$$EY = \mu$$

- Variance:

$$Var(Y) = E(Y - \mu)^2 = \sigma^2$$

- Higher central moments:

$$E(Y-\mu)^m = \begin{cases} \frac{m!}{2^{m/2}(m/2)!}\sigma^m & m \text{ is even} \\ 0 & m \text{ is odd} \end{cases}$$

- In particular:

$$\mu_3 = E(Y-\mu)^3 = 0 (\text{Skewness})$$
$$\mu_4 = E(Y-\mu)^4 = 3\sigma^4$$

- Moment generating function:

$$M_Y(t) = \exp(\mu t + \sigma^2 t^2/2)$$

Standardization

$$Y \sim N(\mu, \sigma^2) \iff Z = \frac{Y - \mu}{\sigma} \sim N(0, 1)$$

Shifting and scaling:

$$Z \sim N(0, 1) \iff Y = \sigma Z + \mu \sim N(\mu, \sigma^2)$$

Notes

- Normal distribution is useful in many practical settings. E.g. measurement error.
- Plays an important role in *sampling distributions* in *large samples*, since the Central Limit Theorem syas that the sums of independent identically distributed random variables are approximately normal
- There are many important distributions that can be derived from functions of normal random variables (e.g. χ^2, t, F). We will briefly present the pdf's and sample spaces of these distributions.

Lecture 27: Nov. 2

Last time

- Normal Distributions

Today

- Common Continuous Distributions
- Families of Distributions

χ^2 distribution If $Z \sim N(0,1)$, then $X = Z^2$ has the χ^2 distribution with 1 degree of freedom. More generally, we have the χ^2 distribution with v degrees of freedom with pdf:

$$f(x) = \frac{(x/2)^{\frac{v}{2}-1}e^{-x/2}}{2\Gamma(v/2)}, \quad x > 0$$

where $\Gamma(a)$ is the complete gamma function,

$$\Gamma(a) = \int_0^\infty x^{a-1}e^{-x}dx$$

The $\chi^2(v)$ distribution is a special case of the gamma distribution, so it is easier to derive its properties from the gamma.

Facts about the Gamma function

- $\Gamma(a+1) = a\Gamma(a), a > 0$
- $\Gamma(1) = 1$
- $\Gamma(n) = (n-1)!$
- $\Gamma(1/2) = \sqrt{\pi}$

Student's t and F distributions Y has a t_k distribution (t with k degrees of freedom) if its pdf can be written as:

$$f(y) = \frac{\Gamma\left[(v+1)/2\right]}{\sqrt{v\pi}\Gamma(v/2)} \frac{1}{(1+y^2/v)^{(v+1)/2}}, \quad -\infty < y < \infty$$

Y has an $F(v_1, v_2)$ distribution if its pdf can be written as:

$$f(y) = \frac{(v_1/v_2)\Gamma\left[(v_1+v_2)/2\right](v_1 y/v_2)^{v_1/2-1}}{\Gamma(v_1/2)\Gamma(v_2/2)(1+v_1 y/v_2)^{(v_1+v_2)/2}}, \quad 0 \leqslant y < \infty$$

There are many important properties and relationships between these three distributions (e.g. χ^2_k is the distribution of the sum of the squares of k independent standard normals). We'll come back to these in a few weeks when we do *sampling distributions and transformations of the normal distribution* (if time permits).

Gamma distribution Notation: $Y \sim Gamma(a, \lambda)$.

- pdf:
$$f(y) = \frac{\lambda e^{-\lambda y}(\lambda y)^{a-1}}{\Gamma(a)}, \quad y \geqslant 0$$
where $\Gamma(a)$ is the gamma function,
$$\Gamma(a) = \int_0^\infty x^{a-1}e^{-x}dx$$
- cdf: In general, there is no closed form, unless a is an integer.
- moments:
$$E(Y) = a/\lambda$$
$$Var(Y) = a/\lambda^2$$
- MGF:
$$M_Y(t) = \left(\frac{1}{1 - t/\lambda}\right)^a, \quad t < \theta$$

Another parameterization Same as the exponential distribution, we can let $\beta = \frac{1}{\lambda}$, then we have

- pdf:
$$f(y) = \frac{y^{a-1}e^{-y/\beta}}{\Gamma(a)\beta^a}, \quad y \geqslant 0$$
- moments:
$$EX = \alpha\beta$$
$$Var(X) = \alpha\beta^2$$
- MGF:
$$M_Y(t) = \left(\frac{1}{1 - t\beta}\right)^a, \quad t < \frac{1}{\beta}$$

Notes:

- The special case $a = 1$ corresponds to an $exponential(\lambda)$
- The parameter a is known as the *shape parameter*, since it most influences the peakedness of the distribution.
- The parameter β is called the *scale parameter* since most of its influence is on the spread of the distribution.
- The special case $Gamma(a = n/2, \lambda = 1/2)$, for integer n, corresponds to the χ^2_n distribution with n degrees of freedom.
- The gamma distribution can be derived as the sum of a independent $exponential(\lambda)$ distributions.

Lecture 28: Nov. 4

Last time

- Common continuous distributions

Today

- Presentations
- Families of Distributions

Beta distribution Notation: $Y \sim Beta(a, b)$.

- Sample space: $[0, 1]$
- pdf:
$$f(y) = \frac{y^{a-1}(1-y)^{b-1}}{B(a,b)}, \quad 0 \leqslant y \leqslant 1$$
where $B(a, b)$ is the Beta function,
$$B(a,b) = \int_0^1 x^{a-1}(1-x)^{b-1}dx = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)},$$
and $\Gamma(a)$ is the gamma function. Note that if a and b are integers, then $B(a, b)$ can be calculated in closed form.
- cdf: In general, there is no closed form, except if a and b are integers.
- moments:
$$EY = \frac{a}{a+b}$$
$$Var(Y) = \frac{ab}{(a+b)^2(a+b+1)}$$
The beta distribution is very flexible, and can take a wide variety of shapes by varying its parameters.
- Special case: $Beta(1, 1) = U(0, 1)$.

Omitted distributions: Weibull distribution, and Cauchy distribution.

Exponential Families A family of pdfs or pmfs with vector parameter $\boldsymbol{\theta}$ is called an *exponential family* if it can be expressed as

$$f(x|\boldsymbol{\theta}) = h(x)c(\boldsymbol{\theta})exp\left(\sum_{j=1}^{k} w_j(\boldsymbol{\theta})t_j(x)\right), \quad x \in S \subset \mathbb{R} \tag{1}$$

where S is not defined in terms of $\boldsymbol{\theta}$, $h(x)$, $c(\boldsymbol{\theta}) \geqslant 0$ and the functions are just functions of the parameters specified; i.e. h is free of $\boldsymbol{\theta}$, $c(\boldsymbol{\theta})$ is free of x, etc...

Examples:

- One-dimensional: Exponential, Poisson
- Two-dimensional: Gaussian

Exponential family parameterizations are unique except for multiplying constant factors.

Example: Gaussian Let $f(x|\mu, \sigma^2)$ be the $n(\mu, \sigma^2)$ family of pdfs, where $\boldsymbol{\theta} = (\mu,$ $sigma)$. Then

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

Thus

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

The parameter space is $(\mu, \sigma^2) \in \mathbb{R} \times (0, \infty)$.

Example: Binomial Let $f(x|p)$ be the $binomial(n, p)$, $0 < p < 1$ family of pmfs.

$$\begin{aligned} f(x|p) &= \binom{n}{x} p^x (1-p)^{n-x} = \binom{n}{x} (1-p)^n \left[\frac{p}{1-p}\right]^x \\ &= \binom{n}{x} (1-p)^n \exp\left[\log\left(\frac{p}{1-p}\right) x\right] \end{aligned}$$

Thus,

$$\begin{aligned} h(x) = \tbinom{n}{x}, \quad x = 0, \ldots, n \quad & w_1(p) = \log\left(\tfrac{p}{1-p}\right) \\ c(p) = (1-p)^n, 0 < p < 1 \quad & t_1(x) = x \end{aligned}$$

Note that this works when p is considered the parameter, while n is fixed. Also, p cannot be 0 or 1. Otherwise, the range changes.

More examples The following distributions belong to Exponential families:

- Continuous: exponential, Gaussian, gamma, beta, χ^2
- Discrete: Poisson, geometric, binomial (fixed # trials), negative binomial (fixed # successes)

The following distributions not exponential families:

- Continuous: t, F, unifrom E.g.:$X \sim U(0, \theta)$

$$f_X(x) = \theta^{-1} 1(0 < x < \theta)$$

- Discrete: uniform, hypergeometric

Theorem If X is a random variable with pdf or pmf of the form 3, then

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

Although these equations may look formidable, when applied to specific cases they can work out quite nicely. Their advantage is that we can replace integration or summation by differentiation, which is often more straightforward.

Example (Normal exponential family) Let $f(x|\mu, \sigma^2)$ be the $N(\mu, \sigma^2)$ family of pdfs, where $\boldsymbol{\theta} = (\mu, \sigma), -\infty < \mu < \infty, \sigma > 0$. Then

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

Define

$$\theta_1 = \tfrac{1}{\sigma^2} > 0, \quad \theta_2 = \tfrac{\mu}{\sigma^2} \in \mathbb{R}$$

Then

$$f_X(x) = \frac{\sqrt{\theta_1}}{\sqrt{2\pi}} \exp\left(-\frac{\theta_2^2}{2\theta_1}\right) \exp\left(-\theta_1 \frac{x^2}{2} + \theta_2 x\right)$$

and

$$\begin{aligned} &h(x) = 1 \text{ for all } x; \\ &c(\boldsymbol{\theta}) = c(\theta_1, \theta_2) = \exp\left(-\tfrac{\theta_2^2}{2\theta_1}\right), \quad (\theta_1, \theta_2) \in (0, \infty) \times \mathbb{R} \\ &w_1(\boldsymbol{\theta}) = \theta_1 \qquad\qquad t_1(x) = -x^2/2 \\ &w_2(\boldsymbol{\theta}) = \theta_2 \qquad\qquad t_2(x) = x \end{aligned}$$

Therefore, by the above theorem

$$\begin{aligned} E(X) &= -\frac{\partial}{\partial \theta_2} \log c(\boldsymbol{\theta}) = \frac{\theta_2}{\theta_1} = \mu \\ Var(X) &= -\frac{\partial^2}{\partial \theta_2^2} \log c(\boldsymbol{\theta}) = -\frac{1}{\theta_1} = \sigma^2 \end{aligned} \tag{2}$$

Lecture 29: Nov. 7

Last time

- Presentations
- Exponential families

Today

- Exponential families
- Location and Scale families
- Chebychev's Inequality

Exponential Families A family of pdfs or pmfs with vector parameter $\boldsymbol{\theta}$ is called an *exponential family* if it can be expressed as

$$f(x|\boldsymbol{\theta}) = h(x)c(\boldsymbol{\theta})exp\left(\sum_{j=1}^{k} w_j(\boldsymbol{\theta})t_j(x)\right), \quad x \in S \subset \mathbb{R} \tag{3}$$

where S is not defined in terms of $\boldsymbol{\theta}$, $h(x)$, $c(\boldsymbol{\theta}) \geqslant 0$ and the functions are just functions of the parameters specified; i.e. h is free of $\boldsymbol{\theta}$, $c(\boldsymbol{\theta})$ is free of x, etc...

Theorem If X is a random variable with pdf or pmf of the form 3, then

$$E\left(\sum_{i=1}^{k} \frac{\partial w_i(\boldsymbol{\theta})}{\partial \theta_j} t_i(X)\right) = -\frac{\partial}{\partial \theta_j} \log c(\boldsymbol{\theta})$$

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

Although these equations may look formidable, when applied to specific cases they can work out quite nicely. Their advantage is that we can replace integration or summation by differentiation, which is often more straightforward.

Example (Normal exponential family) Let $f(x|\mu, \sigma^2)$ be the $n(\mu, \sigma^2)$ family of pdfs, where $-\infty < \mu < \infty, \sigma > 0$. Then

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

Define

$$\theta_1 = \tfrac{1}{\sigma^2} > 0, \quad \theta_2 = \tfrac{\mu}{\sigma^2} \in \mathbb{R}$$

Then

$$f_X(x) = \frac{\sqrt{\theta_1}}{\sqrt{2\pi}} \exp\left(-\frac{\theta_2^2}{2\theta_1}\right) \exp\left(-\theta_1 \frac{x^2}{2} + \theta_2 x\right)$$

and

$$\begin{aligned} &h(x) = 1 \text{ for all } x; \\ &c(\boldsymbol{\theta}) = c(\theta_1, \theta_2) = \tfrac{\sqrt{\theta_1}}{\sqrt{2\pi}} \exp\left(-\tfrac{\theta_2^2}{2\theta_1}\right), \quad (\theta_1, \theta_2) \in (0, \infty) \times \mathbb{R} \\ &w_1(\boldsymbol{\theta}) = \theta_1 \qquad\qquad t_1(x) = -x^2/2 \\ &w_2(\boldsymbol{\theta}) = \theta_2 \qquad\qquad t_2(x) = x \end{aligned}$$

Therefore, by the above theorem

$$\begin{aligned} E(X) &= -\frac{\partial}{\partial \theta_2} \log c(\boldsymbol{\theta}) = \frac{\theta_2}{\theta_1} = \mu \\ Var(X) &= -\frac{\partial^2}{\partial \theta_2^2} \log c(\boldsymbol{\theta}) = -\frac{1}{\theta_1} = \sigma^2 \end{aligned} \tag{4}$$

Location and Scale families

Let Z be a continuous random variable with pdf $f(z)$. Define the class of rvs

$$X_{\mu,\sigma} = \sigma Z + \mu, \quad \mu \in \mathbb{R}, \sigma > 0$$

Then

1. $X_{\mu,\sigma}$ has pdf
$$f_{\mu,\sigma}(x) = \frac{1}{\sigma} f\left(\frac{x-\mu}{\sigma}\right)$$
2. $$E(X) = \sigma E(Z) + \mu, \quad Var(X) = \sigma^2 Var(Z)$$
3. The variable $Z = X_{0,1}$ is called the *generator* and is a member of the class.

Location families and scale families

- The family of pdfs $f_{\mu,\sigma}(x)$ is called a *location-scale* family where μ is called the *location parameter*, and σ is called the *scale parameter*.
- The family of pdfs
$$f_{\mu,1}(x) = f(x-\mu)$$
with $\sigma = 1$ is called a *location* family.
- The family of pdfs
$$f_{0,\sigma}(x) = \frac{1}{\sigma} f\left(\frac{x}{\sigma}\right)$$
with $\mu = 0$ is called a *scale* family.

Example (Exponential location family) Let $f(x) = e^{-x}, x \geqslant 0$, and $f(x) = 0, x < 0$. To form a location family we replace x with $x - \mu$ to obtain

$$f(x|\mu) = \begin{cases} e^{-(x-\mu)} & x - \mu \geqslant 0 \\ 0 & x - \mu < 0 \end{cases}$$
$$= \begin{cases} e^{-(x-\mu)} & x \geqslant \mu \\ 0 & x < \mu \end{cases}$$

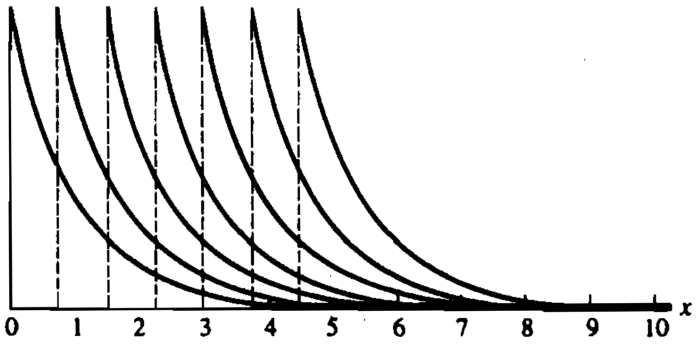

Figure 3.5.2. ***Exponential location densities***

Figure 29.4: Figure 3.5.2. Exponential location densities.

As shown in the above graph, the densities are shifted. Now the positive part of the density starts at μ rather than at 0. If X measures time, then μ might be restricted to be nonnegative so that X will be positive with probability 1 for every value of μ. In this type of model, where μ denotes a bound on the range of X, μ is sometimes called a *threshold parameter*.

The effect of introducing the scale parameter σ is either to stretch ($\sigma > 1$) or to contract ($\sigma < 1$) the graph of $f(x)$ while still maintaining the same basic shape of the graph. This is illustrated in the Figure below.

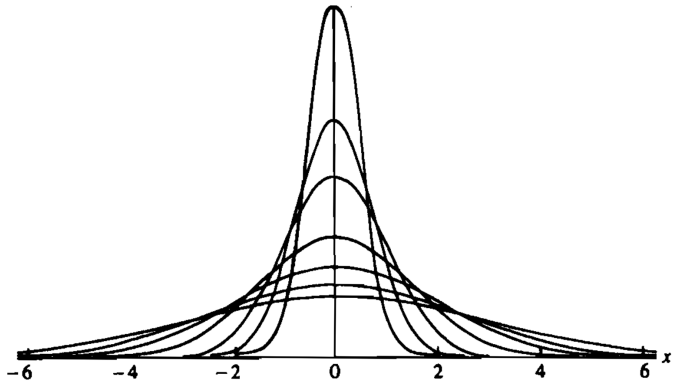

Figure 29.5: Figure 3.5.3. Members of the same scale family

Probability Inequalities

The most famous, and perhaps most useful, probability inequality is Chebychev's Inequality.

Theorem (Chebychev's Inequality) Let X be a random variable and let $g(x)$ be a nonnegative function. Then, for any $r > 0$,

$$\Pr\left(g(X) \geqslant r\right) \leqslant \frac{Eg(X)}{r}.$$

Proof:

$$\begin{aligned} Eg(X) &= \int_{-\infty}^{\infty} g(x) f_X(x) dx \\ &\geqslant \int_{\{x:g(x)\geqslant r\}} g(x) f_X(x) dx \qquad (g \text{ is nonnegative}) \\ &\geqslant r \int_{\{x:g(x)\geqslant r\}} f_X(x) dx \\ &= r \Pr(g(X) \geqslant r) \end{aligned}$$

Example The most widespread use of Chebychev's Inequality involves means and variances. Let $g(x) = (x-\mu)^2/\sigma^2$, where $\mu = EX$ and $\sigma^2 = Var(X)$. For convenience write $r = t^2$.Then

$$\Pr\left(\frac{(X-\mu)^2}{\sigma^2} \geqslant t^2\right) \leqslant \frac{1}{t^2} E\left[\frac{(X-\mu)^2}{\sigma^2}\right] = \frac{1}{t^2}.$$

This means

$$\Pr(|X-\mu| \geqslant t\sigma) \leqslant \frac{1}{t^2}$$

and its companion

$$\Pr(|X-\mu| < t\sigma) \geqslant 1 - \frac{1}{t^2},$$

which give a universal bound on the deviation $|X-\mu|$ in terms of σ. For example, taking $t = 2$, we get

$$\Pr(|X-\mu| \geqslant 2\sigma) \leqslant \frac{1}{2^2} = 0.25,$$

so there is at least a 75% chance that a random variable will be within 2σ of its mean. Have you heard of Six Sigma?

Lecture 30: Nov. 9

Last time

- Exponential families
- Location and Scale families
- Chebychev's Inequality

Today

- Multiple Random Variables (Chapter 4)

Joint and Marginal Distributions

In previous lectures, we have discussed probability models and computation of probability for events involving only one random variable. These are called *univariate models*.

In an experimental situation, it would be very unusual to observe only the value of one random variable. For example, in an experiment designed to gain information about some health characteristics of a population of people, the body weights of several people in the population might be measured. These different weights would be observations on difference random variables, one for each person measured. Multiple observations could also arise because several physical characteristics were measured on each person. Thus, we need to know how to describe and use probability models that deal with more than one random variable at a time.

Definition: An n-dimensional random vector $\boldsymbol{X} = (X_1, \ldots, X_n)$ is a function from a sample space S into $\mathbb{R}^n$.

- Each coordinate X_i is a random variable.
- The random vector is associated with a probability space $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n), F)$.
- For each Borel set B,
$$\Pr\{\boldsymbol{X} \in B\} = \Pr\{\boldsymbol{X}^{-1}(B)\} \tag{5}$$
where
$$\boldsymbol{X}^{-1}(B) = \{w : \boldsymbol{X}(w) \in B\}$$

Example (Bivariate random variable) A fair coin is flipped 3 times. Define the random vector (X, Y) where X represents the number of heads on the last toss and Y the total number of heads. Then, the probabilities of various outcomes are given in the following table:

Outcome	(x, y)	Pr(*outcome*)
(H, H, H)	(1, 3)	1/8
(H, T, H), (T, H, H)	(1, 2)	2/8
(H, H, T)	(0, 2)	1/8
(T, T, H)	(1, 1)	1/8
(T, H, T), (H, T, T)	(0, 1)	2/8
(T, T, T)	(0, 0)	1/8

Definition Two random variables X and Y are said to be jointly *discrete* if there is an associated *joint probability mass function*,

$$f_{X,Y}(x, y) = \Pr\{X = x, Y = y\}$$

which sums to 1 over a finite or possibly countable combinations of x and y for which $f_{X,Y}(x, y) > 0$, i.e.,

$$\sum_{x,y} f_{X,Y}(x, y) = 1$$

From this, one can also obtain the marginal pmfs of X and Y as follows:

$$f_X(x) = \Pr(X = x) = \sum_y f_{X,Y}(x, y)$$
$$f_Y(y) = \Pr(Y = y) = \sum_x f_{X,Y}(x, y)$$

Example Back to the fair coin example again. From the definition, we can construct the joint pmf of X and Y:

		Y			
		0	1	2	3
X	0	1/8	1/4	1/8	0
	1	0	1/8	1/4	1/8

The marginal distributions of X and Y are also easy to find. Note: Marginals do not determine joint pmf.

Bivariate cdfs Whether they are discrete or continuous or some combination of the two, we can always define the *joint cdf*. For $n = 2$, the *bivariate cumulative distribution function* is

$$F_{X,Y}(x, y) = \Pr\{X \leqslant x, Y \leqslant y\}$$

Properties:

- $F_{X,Y}(x, y) \geqslant 0$
- $F_{X,Y}(\infty, \infty) = 1$

- $F_{X,Y}(-\infty, y) = F_{X,Y}(x, -\infty) = 0$
- $F_{X,Y}(-\infty, -\infty) = 0$
- F is non-decreasing and right-continuous in each variable separately.

Joint probabilities All joint probability statements about X and Y can be answered in terms of their joint cdf:

$$\Pr(x_1 < X \leqslant x_2, y_1 < Y \leqslant y_2) = \\ F_{X,Y}(x_2, y_2) + F_{X,Y}(x_1, y_1) - F_{X,Y}(x_1, y_2) - F_{X,Y}(x_2, y_1)$$

Example

$$\Pr(X > x, Y > y) = 1 - F_X(x) - F_Y(y) + F_{X,Y}(x, y)$$

Note: To ensure that a bivariate function $F(x, y)$ is a proper cdf, it must satisfy all the properties mentioned above and the rectangular property above.

Marginal distributions From $F_{X,Y}$, we can derive the univariate distribution functions for X and Y. These are generally called *marginal distributions*.

$$F_X(x) = \Pr\{X \leqslant x\} = \Pr\{X \leqslant x, Y \leqslant \infty\} = F_{X,Y}(x, \infty)$$
$$F_Y(y) = \Pr\{Y \leqslant y\} = \Pr\{X < \infty, Y \leqslant y\} = F_{X,Y}(\infty, y)$$

Note: Although we can obtain $F_X(x)$ and $F_Y(y)$ from the joint cdf, we cannot do the reverse.

Continuous Bivariate RVs The random variables X and Y are said to be *jointly continuous* if there exists a function $f_{X,Y}(x, y)$, such that for any Borel set B of 2-tuples in $\mathbb{R}^2$,

$$\Pr\{(X, Y) \in B\} = \int\int_{(x,y)\in B} f_{X,Y}(x, y)dxdy.$$

The function $f_{X,Y}(x, y)$ is called the *joint probability density function* for X and Y. It follows in this case that

$$F_{X,Y}(x, y) = \int_{-\infty}^{x}\int_{-\infty}^{y} f_{X,Y}(s, t)dsdt,$$
$$f_{X,Y}(x, y) = \frac{\partial^2 F(x, y)}{\partial x \partial y}$$

Properties of the bivariate pdf

- $f_{X,Y}(x, y) \geqslant 0$
- $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{X,Y}(x, y)dxdy = 1$
- $f_{X,Y}(x, y)$ is not a probability, but can be thought of as a relative probability of (X, Y) falling into a small rectangle located at (x, y):

$$\Pr\{x < X \leqslant x + dx, y < Y \leqslant y + dy\} \approx f(x, y)dxdy$$

- The *marginal probability density functions* for X and Y can be obtained as

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)dy$$
$$f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)dx$$

Example 1

$$F_{X,Y}(x,y) = xy \qquad 0 < x \leqslant 1, 0 < y \leqslant 1$$
$$f_{X,Y}(x,y) = \frac{\partial^2 F_{X,Y}(x,y)}{\partial x \partial y} =$$
$$f_X(x) =$$
$$f_Y(y) =$$

Example 2

$$F_{X,Y}(x,y) = x - x\log\frac{x}{y} \qquad 0 < x \leqslant y \leqslant 1$$
$$f_{X,Y}(x,y) = \frac{\partial^2 F_{X,Y}(x,y)}{\partial x \partial y} =$$
$$f_X(x) =$$
$$f_Y(y) =$$

Note: Once we have $f_X(x)$ and $f_Y(y)$, we can obtain $F_X(x)$ and $F_Y(y)$ directly. Double check: $F_X(x) = F_{X,Y}(x,\infty)$.