Math 5010–1
Introduction to Probability

Based on D. Stirzaker's book

Cambridge University Press

Davar Khoshnevisan
University of Utah

Lecture 1

1. The sample space, events, and outcomes

- Need a math model for describing "random" events that result from performing an "experiment."
- Ω denotes a sample space. We think of the elements of Ω as "outcomes" of the experiment.
- $\mathscr{F}$ is a collection of subsets of Ω; elements of $\mathscr{F}$ are called "events." We wish to assign a "probability" $P(A)$ to every $A \in \mathscr{F}$. When Ω is finite, $\mathscr{F}$ is always taken to be the collection of all subsets of Ω.

Example 1.1. Roll a six-sided die; what is the probability of rolling a six? First, write a sample space. Here is a natural one:

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

In this case, Ω is finite and we want $\mathscr{F}$ to be the collection of all subsets of Ω. That is,

$$\mathscr{F} = \Big\{\varnothing, \Omega, \{1\}, \ldots, \{6\}, \{1,2\}, \ldots, \{1,6\}, \ldots, \{1,2,\ldots,6\}\Big\}.$$

Example 1.2. Toss two coins; what is the probability that we get two heads? A natural sample space is

$$\Omega = \Big\{(H_1, H_2), (H_1, T_2), (T_1, H_2), (T_1, T_2)\Big\}.$$

Once we have readied a sample space Ω and an event-space $\mathscr{F}$, we need to assign a probability to every event. This assignment cannot be made at whim; it has to satisfy some properties.

2. Rules of probability

Rule 1. $0 \leqslant P(A) \leqslant 1$ for every event A.

Rule 2. $P(\Omega) = 1$. "Something will happen with probability one."

Rule 3 (Addition rule)**.** If A and B are disjoint events [i.e., $A \cap B = \varnothing$], then the probability that at least one of the two occurs is the sum of the individual probabilities. More precisely put,

$$P(A \cup B) = P(A) + P(B).$$

Lemma 1.3. *Choose and fix an integer $n \geqslant 1$. If $A_1, A_2, \ldots, A_n$ are disjoint events, then*

$$P\left(\bigcup_{i=1}^{n} A_i\right) = P(A_1) + \cdots + P(A_n).$$

Proof. The proof uses *mathematical induction*.

Claim. If the assertion is true for $n-1$, then it is true for n.

The assertion is clearly true for $n = 1$, and it is true for $n = 2$ by Rule 3. Because it is true for $n = 2$, the Claim shows that the assertion holds for $n = 3$. Because it holds for $n = 3$, the Claim implies that it holds for $n = 4$, etc.

Proof of Claim. We can write $A_1 \cup \cdots \cup A_n$ as $A_1 \cup B$, where $B = A_2 \cup \cdots \cup A_n$. Evidently, A_1 and B are disjoint. Therefore, Rule 3 implies that $P(A) = P(A_1 \cup B) = P(A_1) + P(B)$. But B itself is a disjoint union of $n-1$ events. Therefore $P(B) = P(A_2) + \cdots + P(A_n)$, thanks to the assumption of the Claim ["the induction hypothesis"]. This ends the proof. □

Lecture 2

1. Properties of probability

Rules 1–3 suffice if we want to study only finite sample spaces. But infinite samples spaces are also interesting. This happens, for example, if we want to write a model that answers, "what is the probability that we toss a coin 12 times before we toss heads?" This leads us to the next, and final, rule of probability.

Rule 4 (Extended addition rule)**.** If $A_1, A_2, \ldots$ are [countably-many] disjoint events, then

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

This rule will be extremely important to us soon. It looks as if we might be able to derive this as a consequence of Lemma 1.3, but that is not the case ... it needs to be assumed as part of our model of probability theory.

Rules 1–4 have other consequences as well.

Example 2.1. Recall that A^c, the complement of A, is the collection of all points in Ω that are not in A. Thus, A and A^c are disjoint. Because $\Omega = A \cup A^c$ is a disjoint union, Rules 2 and 3 together imply then that

$$\begin{aligned} 1 &= \mathrm{P}(\Omega) \\ &= \mathrm{P}(A \cup A^c) \\ &= \mathrm{P}(A) + \mathrm{P}(A^c). \end{aligned}$$

Thus, we obtain the physically–appealing statement that

$$\mathrm{P}(A) = 1 - \mathrm{P}(A^c).$$

For instance, this yields $P(\varnothing) = 1 - P(\Omega) = 0$. "Chances are zero that nothing happens."

Example 2.2. If $A \subseteq B$, then we can write B as a disjoint union: $B = A \cup (B \cap A^c)$. Therefore, $P(B) = P(A) + P(B \cap A^c)$. The latter probability is $\geqslant 0$ by Rule 1. Therefore, we reach another physically-appealing property:

$$\text{If } A \cup B, \text{ then } P(A) \leqslant P(B).$$

Example 2.3. Suppose $\Omega = \{\omega_1, \ldots, \omega_N\}$ has N distinct elements ("N distinct outcomes of the experiment"). One way of assigning probabilities to every subset of Ω is to just let

$$P(A) = \frac{|A|}{|\Omega|} = \frac{|A|}{N},$$

where $|E|$ denotes the number of elements of E. Let us check that this probability assignment satisfies Rules 1–4. Rules 1 and 2 are easy to verify, and Rule 4 holds vacuously because Ω does not have infinitely-many disjoint subsets. It remains to verify Rule 3. If A and B are disjoint subsets of Ω, then $|A \cup B| = |A| + |B|$. Rule 3 follows from this. In this example, each outcome ω_i has probability $1/N$. Thus, these are "equally likely outcomes."

Example 2.4. Let

$$\Omega = \Big\{(H_1, H_2), (H_1, T_2), (T_1, H_2), (T_1, T_2)\Big\}.$$

There are four possible outcomes. Suppose that they are equally likely. Then, by Rule 3,

$$\begin{aligned} P(\{H_1\}) &= P\Big(\{H_1, H_2\} \cup \{H_1, T_2\}\Big) \\ &= P(\{H_1, H_2\}) + P(\{H_1, T_2\}) \\ &= \frac{1}{4} + \frac{1}{4} \\ &= \frac{1}{2}. \end{aligned}$$

In fact, in this model for equally-likely outcomes, $P(\{H_1\}) = P(\{H_2\}) = P(\{T_1\}) = P(\{T_2\}) = 1/2$. Thus, we are modeling two fair tosses of two fair coins.

Example 2.5. Let us continue with the sample space of the previous example, but assign probabilities differently. Here, we define $P(\{H_1, H_2\}) = P(\{T_1, T_2\}) = 1/2$ and $P(\{H_1, T_2\}) = P(\{T_1, H_2\}) = 1/2$. We compute, as we

did before, to find that $P(\{H_1\}) = P(\{H_2\}) = P(\{H_3\}) = P(\{H_4\}) = 1/2$. But now the coins are not tossed fairly. In fact, the results of the two coin tosses are the same in this model.

The following generalizes Rule 3, because $P(A \cap B) = 0$ when A and B are disjoint.

Lemma 2.6 (Another addition rule)**.** *If A and B are events (not necessarily disjoint), then*

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

Proof. We can write $A \cup B$ as a disjoint union of three events:

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

By Rule 3,

$$P(A \cup B) = P(A \cap B^c) + P(A^c \cap B) + P(A \cap B). \tag{1}$$

Similarly, write $A = (A \cap B^c) \cup (A \cap B)$, as a disjoint union, to find that

$$P(A) = P(A \cap B^c) + P(A \cap B). \tag{2}$$

There is a third identity that is proved the same way. Namely,

$$P(B) = P(A^c \cap B) + P(A \cap B). \tag{3}$$

Add (2) and (3) and solve to find that

$$P(A \cap B^c) + P(A^c \cap B) = P(A) + P(B) - 2P(A \cap B).$$

Plug this in to the right-hand side of (1) to finish the proof. □

2. An example

Roll two fair dice fairly; all possible outcomes are equally likely.

2.1. A good sample space is

$$\Omega = \begin{Bmatrix} (1,1) & (1,2) & \cdots & (1.6) \\ \vdots & \vdots & \ddots & \vdots \\ (6,1) & (6,2) & \cdots & (6.6) \end{Bmatrix}$$

We have seen already that $P(A) = |A|/|\Omega|$ for any event A. Therefore, the first question we address is, "how many items are in Ω?" We can think of Ω as a 6-by-6 table; so $|\Omega| = 6 \times 6 = 36$, by second-grade arithmetic.

Before we proceed with our example, let us document this observation more abstractly.

Proposition 2.7 (The first principle of counting)**.** *If we have* $\mathfrak{m}$ *distinct forks and* $\mathfrak{n}$ *distinct knives, then* $\mathfrak{mn}$ *distinct knife–fork combinations are possible.*

. . . not to be mistaken with . . .

Proposition 2.8 (The second principle of counting)**.** *If we have* $\mathfrak{m}$ *distinct forks and* $\mathfrak{n}$ *distinct knives, then there are* $\mathfrak{m}+\mathfrak{n}$ *utensils.*

. . . back to our problem . . .

2.2. What is the probability that we roll doubles? Let

$$A = \{(1,1), (2,2), \ldots, (6,6)\}.$$

We are asking to find $P(A) = |A|/36$. But there are 6 items in A; hence, $P(A) = 6/36 = 1/6$.

2.3. What are the chances that we roll a total of five dots? Let

$$A = \{(1,4), (2,3), (3,2), (4,1)\}.$$

We need to find $P(A) = |A|/36 = 4/36 = 1/9$.

2.4. What is the probability that we roll somewhere between two and five dots (inclusive)? Let

$$A = \left\{ \overbrace{(1,1)}^{\text{sum = 2}}, \underbrace{(1,2), (2,1)}_{\text{sum =3}}, \overbrace{(1,3), (2,2), (3,1)}^{\text{sum =4}}, \underbrace{(1,4), (4,1), (2,3), (3,2)}_{\text{sum=5}} \right\}.$$

We are asking to find $P(A) = 10/36$.

2.5. What are the odds that the product of the number of dots thus rolls is an odd number? The event in question is

$$A := \begin{Bmatrix} (1,1), & (1,3), & (1,5) \\ (3,1), & (3,3), & (3,5) \\ (5,1), & (5,3), & (5,5) \end{Bmatrix}.$$

And $P(A) = 9/36 = 1/4$.

Lecture 3

1. Easy cards

There are 52 cards in a deck. You deal two cards, all pairs equally likely.

Math model: Ω is the collection of all pairs [drawn without replacement from an ordinary deck]. What is $|\Omega|$? To answer this note that $2|\Omega|$ is the number of all possible ways to give a pair out; i.e., $2|\Omega| = 52 \times 51$, by the principle of counting. Therefore,

$$|\Omega| = \frac{52 \times 51}{2} = 1326.$$

- The probability that the second card is an ace is $(4 \times 51)/2 = 102$ divided by 1326. This probability is $\simeq 0.0769$
- The probability that both cards are aces is $(4 \times 3)/2 = 6$ divided by 1326, which is $\simeq 0.0045$.
- The probability that both cards are the same is P{ace and ace} + $\cdots$ + P{King and King} $= 13 \times 0.0769 \simeq 0.0588$.

2. The birthday problem

n people in a room; all birthdays are equally likely, and assigned at random. What are the chances that no two people in the room are born on the same day? You may assume that there are 365 days a years, and that there are no leap years.

Let $p(n)$ denote the probability in question.

To understand this consider finding $p(2)$ first. There are two people in the room.

The sample space is the collection of all pairs of the form (D_1, D_2), where D_1 and D_2 are birthdays. Note that $|\Omega| = 365^2$ [principle of counting].

In general, Ω is the collection of all "n-tuples" of the form $(D_1, \ldots, D_n)$ where the D_i's are birthdays; $|\Omega| = 365^n$. Let A denote the collection of all elements $(D_1, \ldots, D_n)$ of Ω such that all the D_i's are distinct. We need to find $|A|$.

To understand what is going on, we start with $n = 2$. In order to list all the elements of A, we observe that we have to assign two separate birthdays. [Forks = first birthday; knives = second birthday]. There are therefore 365×364 outcomes in A when $n = 2$. Similarly, when $n = 3$, there are $365 \times 364 \times 363$, and in general, $|A| = 365 \times \cdots \times (365 - n + 1)$. **Check this with induction!**

Thus,

$$p(n) = \frac{|A|}{|\Omega|} = \frac{365 \times \cdots \times (365 - n + 1)}{365^n}.$$

For example, check that $p(10) \simeq 0.88$, which may seem very high at first.

3. An urn problem

n purple and n orange balls are in an urn. You select balls at random [without replacement]. What are the chances that they have different colors?

Here, Ω denotes the collection of all pairs of colors. Note that $|\Omega| = 2n(2n-1)$ [principle of counting].

$$P\{\text{two different colors}\} = 1 - P\{\text{the same color}\}.$$

Also,

$$P\{\text{the same color}\} = P(P_1 \cap P_2) + P(O_1 \cap O_2),$$

where O_j denotes the event that the jth ball is orange, and P_k the event that the kth ball is purple. The number of elements of $P_1 \cap P_2$ is $n(n-1)$; the same holds for $O_1 \cap O_2$. Therefore,

$$\begin{aligned} P\{\text{different colors}\} &= 1 - \left[\frac{n(n-1)}{2n(2n-1)} + \frac{n(n-1)}{2n(2n-1)}\right] \\ &= \frac{n}{2n-1}. \end{aligned}$$

In particular, regardless of the value of n, we always have

$$P\{\text{different colors}\} > \frac{1}{2}.$$

Lecture 4

1. Conditional Probabilities

Example 4.1. There are 5 women and 10 men in a room. Three of the women and 9 of the men are employed. You select a person at random from the room, all people being equally likely to be chosen. Clearly, Ω is the collection of all 15 people, and

$$\mathrm{P}\{\text{male}\} = \frac{2}{3}, \quad \mathrm{P}\{\text{female}\} = \frac{1}{3}, \quad \mathrm{P}\{\text{employed}\} = \frac{4}{5}.$$

Also,

$$\mathrm{P}\{\text{male and employed}\} = \frac{8}{15}, \quad \mathrm{P}\{\text{female and employed}\} = \frac{4}{15}.$$

Someone has looked at the result of the sample, and tells us that the person sampled is employed. Let $\mathrm{P}(\text{female} \,|\, \text{employed})$ denote the conditional probability of "female" given this piece of information. Then,

$$\mathrm{P}(\text{female} \,|\, \text{employed}) = \frac{|\text{female among employed}|}{|\text{employed}|} = \frac{3}{12} = \frac{1}{4}.$$

Definition 4.2. If A and B are events and $\mathrm{P}(B) > 0$, then the *conditional probability of* A *given* B is

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

For the previous example, this amounts to writing

$$\mathrm{P}(\text{Female} \,|\, \text{employed}) = \frac{|\text{female and employed}|/|\Omega|}{|\text{employed}|/|\Omega|} = \frac{1}{4}.$$

Example 4.3. If we deal two cards fairly from a standard deck, the probability of $K_1 \cap K_2$ [$K_j = \{\text{King on the } j \text{ draw}\}$] is

$$P(K_1 \cap K_2) = P(K_1)P(K_2 \,|\, K_1) = \frac{4}{52} \times \frac{3}{51}.$$

This agrees with direct counting: $|K_1 \cap K_2| = 4 \times 3$, whereas $|\Omega| = 52 \times 51$.

Similarly,

$$\begin{aligned} P(K_1 \cap K_2 \cap K_3) &= P(K_1)P(K_2 \cap K_3 \,|\, K_1) \\ &= \frac{4}{52} \times \frac{P(K_1 \cap K_2 \cap K_3)}{P(K_1)} \\ &= \frac{4}{52} \times \frac{P(K_2 \,|\, K_1)}{P(K_2 \cap K_1)} \times \frac{P(K_1 \cap K_2 \cap K_3)}{P(K_2 \cap K_1)} \\ &= \frac{4}{52} \times \frac{3}{51} \times P(K_3 \,|\, K_2 \cap K_2) \\ &= \frac{4}{52} \times \frac{3}{51} \times \frac{2}{50}. \end{aligned}$$

Or for that matter,

$$P(K_1 \cap K_2 \cap K_3 \cap K_4) = \frac{4}{52} \times \frac{3}{51} \times \frac{2}{50} \times \frac{1}{49}.$$

(Check!)

Theorem 4.4 (Law of total probability)**.** *For all events* A *and* B,

$$P(A) = P(A \cap B) + P(A \cap B^c).$$

If, in addition, $0 < P(B) < 1$, *then*

$$P(A) = P(A \,|\, B)P(B) + P(A \,|\, B^c)P(B^c).$$

Proof. For the first statement, note that $A = (A \cap B) \cup (A \cap B^c)$ is a disjoint union. For the second, write $P(A \cap B) = P(A \,|\, B)P(B)$ and $P(A \cap B^c) = P(A \,|\, B^c)P(B^c)$. □

Example 4.5 (The Monte Hall problem)**.** Three doors: behind one is a nice prize; behind the other two lie goats. You choose a door at random. The host (Monte Hall) opens another door, and gives you the option of changing your choice to the remaining unopened door. Should you take his offer?

The answer is "yes." Indeed, if W denotes the event that you win, then under the "not switch" model, we have

$$P(W) = \frac{1}{3}. \tag{4}$$

Under the "switch model,"

$$P(W) = P(W \mid R)P(R) + P(W \mid R^c)P(R^c),$$

where R denotes the event that you had it right in your first guess. Now $P(R) = 1/3$, but because you are going to switch, $P(W \mid R) = 0$ and $P(W \mid R^c) = 1$. Therefore,

$$P(W) = \frac{2}{3}.$$

Compare this with (4) to see that you should always switch. What are my assumptions on Ω? This is an important issue, as can be seen from reading the nice discussion (p. 75) of the text.

Example 4.6. There are three types of people: poor (π), middle-income (μ), and rich (ρ). 40% of all π, 45% of μ, and 60% of ρ are over 25 years old (Θ). Find $P(\Theta)$. The result of Theorem 4.4 gets replaced with

$$\begin{aligned} P(\Theta) &= P(\Theta \cap \pi) + P(\Theta \cap \mu) + P(\Theta \cap \rho) \\ &= P(\Theta \mid \pi)P(\pi) + P(\Theta \mid \mu)P(\mu) + P(\Theta \mid \rho)P(\rho) \\ &= 0.4P(\pi) + 0.45P(\mu) + 0.6P(\rho). \end{aligned}$$

If we knew $P(\pi)$ and $P(\mu)$, then we could solve. For example, suppose $P(\pi) = 0.1$ and $P(\mu) = 0.3$. Then $P(\rho) = 0.6$ (why?), and

$$P(\Theta) = (0.4 \times 0.1) + (0.45 \times 0.3) + (0.6 \times 0.6) = 0.535.$$

2. Bayes's Theorem

The following question arises from time to time: Suppose A and B are two events of positive probability. If we know $P(B \mid A)$ but want $P(A \mid B)$, then we can proceed as follows:

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

If we know only the conditional probabilities, then we can write $P(B)$, in turn, using Theorem 4.4, and obtain

Theorem 4.7 (Bayes's Formula)**.** *If A, A^c and B are events of positive probability, then*

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B \mid A)P(A) + P(B \mid A^c)P(A^c)}.$$

Example 4.8. There are two coins on a table. The first tosses heads with probability $1/2$, whereas the second tosses heads with probability $1/3$. You select one at random and toss it. What are the chances that you toss heads?

Question: What is Ω?

Question: Someone tells you that the end result of this game was heads. What are the odds that it was the first coin that was chosen?

Let C denote the event that you selected the first coin. Let H denote the event that you tossed heads. We know: $P(C) = 1/2$, $P(H \mid C) = 1/2$, and $P(H \mid C^c) = 1/3$. By Bayes's formula,

$$\begin{aligned} P(C \mid H) &= \frac{P(H \mid C)P(C)}{P(H \mid C)P(C) + P(H \mid C^c)P(C^c)} \\ &= \frac{\frac{1}{2} \times \frac{1}{2}}{\left(\frac{1}{2} \times \frac{1}{2}\right) + \left(\frac{1}{3} \times \frac{1}{2}\right)} \\ &= \frac{3}{5}. \end{aligned}$$

Lecture 5

1. Independence

- Events A and B are said to be *independent* if

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

Divide both sides by $P(B)$, if it is positive, to find that A and B are independent if and only if

$$P(A \mid B) = P(A).$$

"Knowledge of B tells us nothing new about A."

Two experiments are *independent* if A_1 and A_2 are independent for all outcomes A_j of experiment j.

Example 5.1. Toss two fair coins; all possible outcomes are equally likely. Let H_j denote the event that the jth coin landed on heads, and $T_j = H_j^c$. Then,

$$P(H_1 \cap H_2) = \frac{1}{4} = P(H_1)P(H_2).$$

In fact, the two coins are independent because $P(T_1 \cap T_2) = P(T_1 \cap H_2) = P(H_1 \cap H_2) = 1/4$ also. Conversely, if two fair coins are tossed independently, then all possible outcomes are equally likely to occur. What if the coins are not fair, say $P(H_1) = P(H_2) = 1/4$?

- Three events A_1, A_2, A_3 are *independent* if any two of them. Events A_1, A_2, A_3, A_4 are independent if any three of are. And in general, once we have defined the independence of $n-1$ events, we define n events $A_1, \ldots, A_n$ to be *independent* if any $n-1$ of them are independent.
- One says that n experiments are *independent*, for all $n \geqslant 2$, if any $n-1$ of them are independent.

You should check that this last one is a well-defined (albeit inductive) definition.

2. Gambler's ruin formula

You, the "Gambler," are playing independent repetitions of a fair game against the "House." When you win, you gain a dollar; when you lose, you lose a dollar. You start with k dollars, and the House starts with K dollars. What is the probability that the House is ruined before you?

Define P_j to be the conditional probability that when the game ends you have $K+j$ dollars, given that you start with j dollars initially. We want to find P_k.

Two easy cases are: $P_0 = 0$ and $P_{k+K} = 1$.

By Theorem 4.4 and independence,

$$P_j = \frac{1}{2}P_{j+1} + \frac{1}{2}P_{j-1} \qquad \text{for } 0 < j < k+K.$$

In order to solve this, write $P_j = \frac{1}{2}P_j + \frac{1}{2}P_j$, so that

$$\frac{1}{2}P_j + \frac{1}{2}P_j = \frac{1}{2}P_{j+1} + \frac{1}{2}P_{j-1} \qquad \text{for } 0 < j < k+K.$$

Multiply both side by two and solve:

$$P_{j+1} - P_j = P_j - P_{j-1} \qquad \text{for } 0 < j < k+K.$$

In other words,

$$P_{j+1} - P_j = P_1 \qquad \text{for } 0 < j < k+K.$$

This is the simplest of all possible "difference equations." In order to solve it you note that, since $P_0 = 0$,

$$\begin{aligned} P_{j+1} &= (P_{j+1} - P_j) + (P_j - P_{j-1}) + \cdots + (P_1 - P_0) \qquad \text{for } 0 < j < k+K \\ &= (j+1)P_1 \qquad \text{for } 0 < j < k+K. \end{aligned}$$

Apply this with $j = k+K-1$ to find that

$$1 = P_{k+K} = (k+K)P_1, \quad \text{and hence} \quad P_1 = \frac{1}{k+K}.$$

Therefore,

$$P_{j+1} = \frac{j+1}{k+K} \qquad \text{for } 0 < j < k+K.$$

Set $j = k-1$ to find the following:

Theorem 5.2 (Gambler's ruin formula)**.** *If you start with* k *dollars, then the probability that you end with* $k+K$ *dollars before losing all of your initial fortune is* $k/(k+K)$ *for all* $1 \leqslant k \leqslant K$.

3. Conditional probabilities as probabilities

Suppose B is an event of positive probability. Consider the conditional probability distribution, $Q(\cdots) = P(\cdots \mid B)$.

Theorem 5.3. *Q is a probability on the new sample space B. [It is also a probability on the larger sample space Ω, why?]*

Proof. Rule 1 is easy to verify: For all events A,

$$0 \leqslant Q(A) = \frac{P(A \cap B)}{P(B)} \leqslant \frac{P(B)}{P(B)} = 1,$$

because $A \cap B \subseteq B$ and hence $P(A \cap B) \leqslant P(B)$.

For Rule 2 we check that

$$Q(B) = P(B \mid B) = \frac{P(B \cap B)}{P(B)} = 1.$$

Next suppose $A_1, A_2, \ldots$ are disjoint events. Then,

$$Q\left(\bigcup_{n=1}^{\infty} A_n\right) = \frac{1}{P(B)} P\left(\bigcup_{n=1}^{\infty} A_n \cap B\right).$$

Note that $\cup_{n=1}^{\infty} A_n \cap B = \cup_{n=1}^{\infty} (A_n \cap B)$, and $(A_1 \cap B), (A_2 \cap B), \ldots$ are disjoint events. Therefore,

$$Q\left(\bigcup_{n=1}^{\infty} A_n\right) = \frac{1}{P(B)} \sum_{N=1}^{\infty} P(A_n \cap B) = \sum_{n=1}^{\infty} Q(A_n).$$

This verifies Rule 4, and hence Rule 3. □

Lecture 6

1. Combinatorics

Recall the two basic principles of counting [combinatorics]:

First principle: m distinct garden forks plus n distinct fish forks equals $m + n$ distinct forks.

Second principle: m distinct knives and n distinct forks equals mn distinct ways of taking a knife and a fork.

2. Unordered Selection

Example 6.1. 6 dice are rolled. What is the probability that they all show different faces?

$\Omega =?$

$|\Omega| = 6^6$.

If A is the event in question, then $|A| = 6 \times 5 \times 4 \times 3 \times 2 \times 1$.

Definition 6.2. If k is an integer $\geqslant 1$, then we define "k factorial" as the following integer:

$$k! = k \cdot (k-1) \cdot (k-2) \cdots 2 \cdot 1.$$

For consistency of future formulas, we define also

$$0! = 1.$$

Example 6.3. Five rolls of a fair die. What is $P(A)$, where A is the event that all five show different faces? Note that $|A|$ is equal to 6 [which face is left out] times 6^5. Thus,

$$P(A) = \frac{6 \cdot 5!}{6^5} = \frac{6!}{6^5}.$$

3. Ordered Selection

Example 6.4. Two-card poker.

$$P(\text{doubles}) = \frac{13 \times \binom{4\times 3}{2}}{\binom{52\times 51}{2}}.$$

Theorem 6.5. *n objects are divided into r types. n_1 are of type 1; n_2 of type 2; $\ldots$; n_r are of type r. Thus, $n = n_1 + \cdots + n_r$. Objects of the same type are indistinguishable. The number of permutations is*

$$\binom{n}{n_1, \ldots, n_r} = \frac{n!}{n_1! \cdots n_r!}.$$

Proof. Let N denote the number of permutations; we seek to find N. For every permtation in N there are $n_1! \cdots n_r!$ permutations wherein all n objects are treated differently. Therefore, $n_1! \cdots n_r! N = n!$. Solve to finish. □

Example 6.6. n people; choose r of them to form a "team." The number of different teams is then

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

You have to choose r of type 1 ("put this one in the team"), and $n - r$ of type 2 ("leave this one out of the team").

Definition 6.7. We write the preceding count statistic as "n choose r," and write it as

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

Example 6.8. Roll 4 dice; let A denote the event that all faces are different. Then,

$$|A| = \binom{6}{4} 4! = \frac{6!}{2!} = \frac{6!}{2}.$$

The 6-choose-4 is there because that is how many ways we can choose the different faces. Thus,

$$\mathrm{P}(\mathrm{A}) = \frac{6!}{2 \times 4^6}.$$

Example 6.9. There are

$$\binom{52}{5} = 2,598,960$$

different standard poker hands possible.

Lecture 7

1. Unordered Selection, continued

Let us recall the following:

Theorem 7.1. *The number of ways to create a team of r things among n is "n choose r." Its numerical value is*

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

Example 7.2. If there are n people in a room, then they can shake hands in $\binom{n}{2}$ many different ways. Indeed, the number of possible hand shakes is the same as the number of ways we can list all pairs of people, which is clearly $\binom{n}{2}$. Here is another, equivalent, interpretation. If there are n vertices in a "graph," then there are $\binom{n}{2}$ many different possible "edges" that can be formed between distinct vertices. The reasoning is the same.

Example 7.3 (Recap)**.** There are $\binom{52}{5}$ many distinct poker hands.

Example 7.4 (Poker)**.** The number of different "pairs" $[a, a, b, c, d]$ is

$$\underbrace{13}_{\text{choose the } a} \times \underbrace{\binom{4}{2}}_{\text{deal the two } a\text{'s}} \times \underbrace{\binom{12}{3}}_{\text{choose the } b,\ c,\ \text{and } d} \times \underbrace{4^3}_{\text{deal } b, c, d}.$$

Therefore,

$$\mathrm{P}(\text{pairs}) = \frac{13 \times \binom{4}{2} \times \binom{12}{3} \times 4^3}{\binom{52}{5}} \approx 0.42.$$

Example 7.5 (Poker)**.** Let A denote the event that we get two pairs $[a, a, b, b, c]$. Then,

$$|A| = \underbrace{\binom{13}{2}}_{\text{choose } a,\, b} \times \underbrace{\binom{4}{2}^2}_{\text{deal the } a,\, b} \times \underbrace{13}_{\text{choose } c} + \underbrace{4}_{\text{deal } c}.$$

Therefore,

$$P(\text{two pairs}) = \frac{\binom{13}{2} \times \binom{4}{2}^2 \times 13 \times 4}{\binom{52}{5}} \approx 0.06.$$

Example 7.6. How many subsets does $\{1, \ldots, n\}$ have? Assign to each element of $\{1, \ldots, n\}$ a zero ["not in the subset"] or a one ["in the subset"]. Thus, the number of subsets of a set with n distinct elements is 2^n.

Example 7.7. Choose and fix an integer $r \in \{0, \ldots, n\}$. The number of subsets of $\{1, \ldots, n\}$ that have size r is $\binom{n}{r}$. This, and the preceding proves the following amusing combinatorial identity:

$$\sum_{j=0}^{n} \binom{n}{j} = 2^n.$$

You may need to also recall the first principle of counting.

The preceding example has a powerful generalization.

Theorem 7.8 (The binomial theorem)**.** *For all integers $n \geqslant 0$ and all real numbers x and y,*

$$(x+y)^n = \sum_{j=0}^{n} \binom{n}{j} x^j y^{n-j}.$$

Remark 7.9. When $n = 2$, this yields the familiar algebraic identity

$$(x+y)^2 = x^2 + 2xy + y^2.$$

For $n = 3$ we obtain

$$\begin{aligned}(x+y)^3 &= \binom{3}{0} x^0 y^3 + \binom{3}{1} x^1 y^2 + \binom{3}{2} x^2 y^1 + \binom{3}{3} x^3 y^0 \\ &= y^3 + 3xy^2 + 3x^2 y + x^3.\end{aligned}$$

Proof. This is obviously correct for $n = 0, 1, 2$. We use induction. Induction hypothesis: True for $n - 1$.

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

Change variables [$k = j + 1$ for the first sum, and $k = j$ for the second] to deduce that

$$\begin{aligned}(x+y)^n &= \sum_{k=1}^{n}\binom{n-1}{k-1}x^k y^{n-k} + \sum_{k=0}^{n-1}\binom{n-1}{k}x^k y^{n-k}\\ &= \sum_{k=1}^{n-1}\left\{\binom{n-1}{k-1} + \binom{n-1}{k}\right\}x^k y^{n-k} + x^n + y^n.\end{aligned}$$

But

$$\begin{aligned}\binom{n-1}{k-1} + \binom{n-1}{k} &= \frac{(n-1)!}{(k-1)!(n-k)!} + \frac{(n-1)!}{k!(n-k-1)!}\\ &= \frac{(n-1)!}{(k-1)!(n-k-1)!}\left\{\frac{1}{n-k} + \frac{1}{k}\right\}\\ &= \frac{(n-1)!}{(k-1)!(n-k-1)!}\times\frac{n}{(n-k)k}\\ &= \frac{n!}{k!(n-k)!}\\ &= \binom{n}{k}.\end{aligned}$$

The binomial theorem follows. □

Lecture 8

1. Random Variables

We would like to say that a random variable X is a "numerical outcome of a complicated experiment." This is not sufficient. For example, suppose you sample 1,500 people at random and find that their average age is 25. Is $X = 25$ a "random variable"? Surely there is nothing random about the number 25!

What is random? The procedure that led to 25. This procedure, for a second sample, is likely to lead to a different number. Procedures are functions, and thence

Definition 8.1. A random variable is a function X from Ω to some set D which is usually [for us] a subset of the real line $\mathbf{R}$, or d-dimensional space $\mathbf{R}^d$.

In order to understand this, let us construct a random variable that models the number of dots in a roll of a fair six-sided die.

Define the sample space,

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

We assume that all outcome are equally likely [fair die].

Define $X(\omega) = \omega$ for all $\omega \in \Omega$, and note that for all $k = 1, \ldots, 6$,

$$P(\{\omega \in \Omega : X(\omega) = k\}) = P(\{k\}) = \frac{1}{6}. \tag{5}$$

This probability is zero for other values of k. Usually, we write $\{X \in A\}$ in place of the set $\{\omega \in \Omega : X(\omega) \in A\}$. In this notation, we have

$$P\{X = k\} = \begin{cases} \frac{1}{6} & \text{if } k = 1, \ldots, 6, \\ 0 & \text{otherwise.} \end{cases} \tag{6}$$

This is a math model for the result of a coin toss.

2. General notation

Suppose X is a random variable, defined on some probability space Ω. By the *distribution* of X we mean the collection of probabilities $P\{X \in A\}$, as A ranges over all sets in $\mathscr{F}$.

If X takes values in a finite, or countably-infinite set, then we say that X is a *discrete random variable*. Its distribution is called a *discrete distribution*. The function

$$f(x) = P\{X = x\}$$

is then called the *mass function* of X. Note that $f(x) = 0$ for all but a countable number of values of x. The values x for which $f(x) > 0$ are called the *possible values* of X.

Some important properties of mass functions:

- $0 \leqslant f(x) \leqslant 1$ for all x. [Easy]
- $\sum_x f(x) = 1$. Proof: $\sum_x f(x) = \sum_x P\{X = x\}$, and this is equal to $P(\cup_x\{X = x\}) = P(\Omega)$, since the union is a countable disjoint union.

3. The binomial distribution

Suppose we perform n independent trials; each trial leads to a "success" or a "failure"; and the probability of success per trial is the same number $p \in (0,1)$.

Let X denote the total number of successes in this experiment. This is a discrete random variable with possible values $0,\dots,n$. We say then that X is a binomial random variable ["$X = \text{Bin}(n,p)$"].

Math modelling questions:

- Construct an Ω.
- Construct X on this Ω.

Let us find the mass function of X. We seek to find $f(x)$, where $x = 0,\dots,n$. For all other values of x, $f(x) = 0$.

Now suppose x is an integer between zero and n. Note that $f(x) = P\{X = x\}$ is the probability of getting exactly x successes and $n-x$ failures. Let S_i denote the event that the ith trial leads to a success. Then,

$$f(x) = P\left(S_1 \cap \cdots \cap S_x \cap S_{x+1}^c \cap \cdots S_n^c\right) + \cdots$$

where we are summing over all possible ways of distributing x successes and $n-x$ failures in n spots. By independence, each of these probabilities

is $p^x(1-p)^{n-x}$. The number of probabilities summed is the number of ways we can distributed x successes and $n-x$ failures into n slots. That is, $\binom{n}{x}$. Therefore,

$$f(x) = P\{X = x\} = \begin{cases} \binom{n}{x} p^x (1-p)^{n-x} & \text{if } x = 0, \ldots, n, \\ 0 & \text{otherwise.} \end{cases}$$

Note that $\sum_x f(x) = 1$ by the binomial theorem. So we have not missed anything.

3.1. An example. Consider the following sampling question: *Ten percent of a certain population smoke. If we take a random sample [without replacement] of* 5 *people from this population, what are the chances that at least 2 people smoke in the sample?*

Let X denote the number of smokers in the sample. Then $X = \text{Bin}(n, p)$ ["success" = "smoker"]. Therefore,

$$\begin{aligned} P\{X \geqslant 2\} &= 1 - P\{X \leqslant 1\} \\ &= 1 - P(\{X = 0\} \cup \{X = 1\}) \\ &= 1 - [p(0) + p(1)] \\ &= 1 - \left[\binom{n}{0} p^0 (1-p)^{n-0} + \binom{n}{1} p^1 (1-p)^{n-1} \right] \\ &= 1 - (1-p)^n - np(1-p)^{n-1}. \end{aligned}$$

Alternatively, we can write

$$P\{X \geqslant 2\} = P(\{X = 2\} \cup \cdots \{X = n\}) = \sum_{j=2}^{n} f(j),$$

and then plug in $f(j) = \binom{n}{j} p^j (1-p)^{n-j}$.

4. The geometric distribution

A p-*coin* is a coin that tosses heads with probability p and tails with probability $1-p$. Suppose we toss a p-coin until the first time heads appears. Let X denote the number of tosses made. Then X is a so-called geometric random variable ["$X = \text{Geom}(p)$"].

Evidently, if n is an integer greater than or equal to one, then $P\{X = n\} = (1-p)^{n-1}p$. Therefore, the mass function of X is given by

$$f(x) = \begin{cases} p(1-p)^{x-1} & \text{if } x = 1, 2, \ldots, \\ 0 & \text{otherwise.} \end{cases}$$

Lecture 9

1. The geometric distribution, continued

1.1. An example. A couple has children until their first son is born. Suppose the sexes of their children are independent from one another [unrealistic], and the probability of girl is 0.6 every time [not too bad]. Let X denote the number of their children to find then that $X = \text{Geom}(0.4)$. In particular,

$$\begin{aligned} P\{X \leqslant 3\} &= f(1) + f(2) + f(3) \\ &= p + p(1-p) + p(1-p)^2 \\ &= p\left[1 + 1 - p + (1-p)^2\right] \\ &= p\left[3 - 3p + p^2\right] \\ &= 0.784. \end{aligned}$$

1.2. The tail of the distribution. Now you may be wondering why these random variables are called "geometric." In order to answer this, consider the tail of the distribution of X (probability of large values). Namely, for all $n \geqslant 1$,

$$\begin{aligned} P\{X \geqslant n\} &= \sum_{j=n}^{\infty} p(1-p)^{j-1} \\ &= p \sum_{k=n-1}^{\infty} (1-p)^k. \end{aligned}$$

Let us recall an elementary fact from calculus.

Lemma 9.1 (Geometric series). *If* $r \in (0,1)$, *then for all* $n \geqslant 0$,

$$\sum_{j=n}^{\infty} r^j = \frac{r^n}{1-r}.$$

Proof. Let $s_n = r^n + r^{n+1} + \cdots = \sum_{j=n}^{\infty} r^j$. Then, we have two relations between s_n and s_{n+1}:

(1) $rs_n = \sum_{j=n+1}^{\infty} r^j = s_{n+1}$; and

(2) $s_{n+1} = s_n - r^n$.

Plug (2) into (1) to find that $rs_n = s_n - r^n$. Solve to obtain the lemma. □

Return to our geometric random variable X to find that

$$P\{X \geqslant n\} = p\frac{(1-p)^{n-1}}{1-(1-p)} = (1-p)^{n-1}.$$

That is, $P\{X \geqslant n\}$ vanishes geometrically fast as $n \to \infty$.

In the couples example (§1.1),

$$P\{X \geqslant n\} = 0.6^{n-1} \qquad \text{for all } n \geqslant 1.$$

2. The negative binomial distribution

Suppose we are tossing a p-coin, where $p \in (0,1)$ is fixed, until we obtain r heads. Let X denote the number of tosses needed. Then, X is a discrete random variable with possible values $r, r+1, r+2, \ldots$. When $r = 1$, then X is Geom(p). In general,

$$f(x) = \begin{cases} \binom{x-1}{r-1} p^r (1-p)^{x-r} & \text{if } x = r, r+1, r+2, \ldots, \\ 0 & \text{otherwise.} \end{cases}$$

This X is said to have a *negative binomial distribution with parameters* r *and* p. Note that our definition differs slightly from that of your text (p. 117).

3. The Poisson distribution

Choose and fix a number $\lambda > 0$. A random variable X is said to have the *Poisson distribution with parameter* λ (written Poiss(λ)) if its mass function is

$$f(x) = \begin{cases} \dfrac{e^{-\lambda}\lambda^x}{x!} & \text{if } x = 0, 1, \ldots, \\ 0 & \text{otherwise.} \end{cases} \tag{7}$$

In order to make sure that this makes sense, it suffices to prove that $\sum_x f(x) = 1$, but this is an immediate consequence of the Taylor expansion of e^λ, viz.,

$$e^\lambda = \sum_{k=0}^{\infty} \frac{\lambda^k}{k!}.$$

3.1. Law of rare events. Is there a physical manner in which Poiss(λ) arises naturally? The answer is "yes." Let $X = \text{Bin}(n, \lambda/n)$. For instance, X could denote the total number of sampled people who have a rare disease (population percentage $= \lambda/n$) in a large sample of size n. Then, for all fixed integers $k = 0, \ldots, n$,

$$f_X(k) = \binom{n}{k} \left(\frac{\lambda}{n}\right)^k \left(1 - \frac{\lambda}{n}\right)^{n-k}. \tag{8}$$

Poisson's "law of rare events" states that if n is large, then the distribution of X is approximately Poiss(λ). In order to deduce this we need two computational lemmas.

Lemma 9.2. *For all* $z \in \mathbf{R}$,

$$\lim_{n\to\infty} \left(1 + \frac{z}{n}\right)^n = e^z.$$

Proof. Because the natural logarithm is continuous on $(0, \infty)$, it suffices to prove that

$$\lim_{n\to\infty} n \ln\left(1 + \frac{z}{n}\right) = z. \tag{9}$$

By Taylor's expansion,

$$\ln\left(1 + \frac{z}{n}\right) = \frac{z}{n} + \frac{\theta^2}{2},$$

where θ lies between 0 and z/n. Equivalently,

$$\frac{z}{n} \leqslant \ln\left(1 + \frac{z}{n}\right) \leqslant \frac{z}{n} + \frac{z^2}{2n^2}.$$

Multiply all sides by n and take limits to find (9), and thence the lemma. □

Lemma 9.3. *If* $k \geqslant 0$ *is a fixed integer, then*

$$\binom{n}{k} \sim \frac{n^k}{k!} \qquad \text{as } n \to \infty.$$

where $a_n \sim b_n$ *means that* $\lim_{n\to\infty}(a_n/b_n) = 1$.

Proof. If $n \geqslant k$, then

$$\begin{aligned}\frac{n!}{n^k(n-k)!} &= \frac{n(n-1)\cdots(n-k+1)}{n^k} \\ &= \frac{n}{n} \times \frac{n-1}{n} \times \cdots \times \frac{n-k+1}{n} \\ &\to 1 \qquad \text{as } n \to \infty.\end{aligned}$$

The lemma follows upon writing out $\binom{n}{k}$ and applying the preceding to that expression. □

Thanks to Lemmas 9.2 and 9.3, and to (8),

$$f_X(k) \sim \frac{n^k}{k!}\frac{\lambda^k}{n^k}e^{-\lambda} = \frac{e^{-\lambda}\lambda^k}{k!}.$$

That is, when n is large, X behaves like a Poiss(λ), and this proves our assertion.

Lecture 10

1. (Cumulative) distribution functions

Let X be a discrete random variable with mass function f. The (cumulative) *distribution function* F of X is defined by

$$F(x) = P\{X \leqslant x\}.$$

Here are some of the properties of distribution functions:

(1) $F(x) \leqslant F(y)$ whenever $x \leqslant y$; therefore, F is non-decreasing.
(2) $1 - F(x) = P\{X > x\}$.
(3) $F(b) - F(a) = P\{a < X \leqslant b\}$ for $a < b$.
(4) $F(x) = \sum_{y:\, y \leqslant x} f(y)$.
(5) $F(\infty) = 1$ and $F(-\infty) = 0$. [Some care is needed]
(6) F is right-continuous. That is, $F(x+) = F(x)$ for all x.
(7) $f(x) = F(x) - F(x-)$ is the size of the jump [if any] at x.

Example 10.1. Suppose X has the mass function

$$f_X(x) = \begin{cases} \frac{1}{2} & \text{if } x = 0, \\ \frac{1}{2} & \text{if } x = 1, \\ 0 & \text{otherwise.} \end{cases}$$

Thus, X has equal chances of being zero and one. Define a new random variable $Y = 2X - 1$. Then, the mass function of Y is

$$f_Y(x) = f_X\left(\frac{x+1}{2}\right) = \begin{cases} \frac{1}{2} & \text{if } x = -1, \\ \frac{1}{2} & \text{if } x = 1, \\ 0 & \text{otherwise.} \end{cases}$$

The procedure of this example actually produces a theorem.

Theorem 10.2. *If* $Y = g(X)$ *for a function* g*, then*

$$f_Y(x) = \sum_{z:\, g(z)=x} f_X(z).$$

2. Expectation

The *expectation* EX of a random variable X is defined formally as

$$EX = \sum_x x f(x).$$

If X has infinitely-many possible values, then the preceding sum must be defined. This happens, for example, if $\sum_x |x| f(x) < \infty$. Also, EX is always defined [but could be $\pm\infty$] if $P\{X \geqslant 0\} = 1$, or if $P\{X \leqslant 0\} = 1$. The *mean* of X is another term for EX.

Example 10.3. If X takes the values ± 1 with respective probabilities $1/2$ each, then $EX = 0$.

Example 10.4. If $X = \text{Bin}(n, p)$, then I claim that $EX = np$. Here is why:

$$\begin{aligned} EX &= \sum_{k=0}^{n} k \overbrace{\binom{n}{k} p^k q^{n-k}}^{f(k)} \\ &= \sum_{k=1}^{n} \frac{n!}{(k-1)!(n-k)!} p^k q^{n-k} \\ &= np \sum_{k=1}^{n} \binom{n-1}{k-1} p^{k-1} q^{(n-1)-(k-1)} \\ &= np \sum_{j=0}^{n-1} \binom{n-1}{j} p^j q^{(n-1)-j} \\ &= np, \end{aligned}$$

thanks to the binomial theorem.

Example 10.5. Suppose $X = \text{Poiss}(\lambda)$. Then, I claim that $EX = \lambda$. Indeed,

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

because $e^{\lambda} = \sum_{j=0}^{\infty} \lambda^j/j!$, thanks to Taylor's expansion.

Example 10.6. Suppose X is negative binomial with parameters r and p. Then, $EX = r/p$ because

$$\begin{aligned} EX &= \sum_{k=r}^{\infty} k \binom{k-1}{r-1} p^r q^{k-r} \\ &= \sum_{k=r}^{\infty} \frac{k!}{(r-1)!(k-r)!} p^r q^{k-r} \\ &= r \sum_{k=r}^{\infty} \binom{k}{r} p^r q^{k-r} \\ &= \frac{r}{p} \sum_{k=r}^{\infty} \binom{k}{r} p^{r+1} q^{(k+1)-(r+1)} \\ &= \frac{r}{p} \sum_{j=r+1}^{\infty} \underbrace{\binom{j-1}{(r+1)-1} p^{r+1} q^{j-(r+1)}}_{P\{\text{Negative binomial } (r+1,p)=j\}} \\ &= \frac{r}{p}. \end{aligned}$$

Thus, for example, $E[\text{Geom}(p)] = 1/p$.

Finally, two examples to test the boundary of the theory so far.

Example 10.7 (A random variable with infinite mean)**.** Let X be a random variable with mass function,

$$f(x) = \begin{cases} \dfrac{1}{Cx^2} & \text{if } x = 1, 2, \ldots, \\ 0 & \text{otherwise,} \end{cases}$$

where $C = \sum_{j=1}^{\infty}(1/j^2)$. Then,

$$EX = \sum_{j=1}^{\infty} j \cdot \frac{1}{Cj^2} = \infty.$$

But $P\{X < \infty\} = \sum_{j=1}^{\infty} 1/(Cj^2) = 1$.

Example 10.8 (A random variable with an undefined mean)**.** Let X be a random with mass function,

$$f(x) = \begin{cases} \dfrac{1}{Dx^2} & \text{if } x = \pm 1, \pm 2, \ldots, \\ 0 & \text{otherwise,} \end{cases}$$

where $D = \sum_{j \in \mathbf{Z} \setminus \{0\}}(1/j^2)$. Then, EX is undefined. If it were defined, then it would be

$$\lim_{n,m\to\infty} \left(\sum_{j=-m}^{-1} \frac{j}{Dj^2} + \sum_{j=1}^{n} \frac{j}{Dj^2} \right) = \frac{1}{D} \lim_{n,m\to\infty} \left(\sum_{j=-m}^{-1} \frac{1}{j} + \sum_{j=1}^{n} \frac{1}{j} \right).$$

But the limit does not exist. The rough reason is that if N is large, then $\sum_{j=1}^{N}(1/j)$ is very nearly $\ln N$ plus a constant (Euler's constant). "Therefore," if n, m are large, then

$$\left(\sum_{j=-m}^{-1} \frac{1}{j} + \sum_{j=1}^{n} \frac{1}{j} \right) \approx -\ln m + \ln n = \ln\left(\frac{n}{m}\right).$$

If $n = m \to \infty$, then this is zero; if $m \gg n \to \infty$, then this goes to $-\infty$; if $n \gg m \to \infty$, then it goes to $+\infty$.

Lecture 11

1. Some properties of expectations

Suppose X is a random variable with mass function f. If $Y = g(X)$ for some function g, then what is the expectation of Y? One way to address this is to first find the mass function f_Y of Y, and then to compute EY as $\sum_a a f_Y(a)$ [provided that the sum makes sense, of course]. But there is a more efficient method.

Theorem 11.1. *If X has mass function f and g is some function, then*

$$E[g(X)] = \sum_x g(x)f(x),$$

provided that either $g(x) \geqslant 0$ for all x, or $\sum_x |g(x)|f(x) < \infty$.

Proof. Let $y_1, y_2, \ldots$ denote the possible values of $g(X)$. Consider the set $A_j = \{x : \ g(x) = y_j\}$ for all $j \geqslant 1$. Because the y_j's are distinct, it follows that the A_j's are disjoint. Moreover,

$$\begin{aligned} E[g(X)] &= \sum_{j=1}^{\infty} y_j P\{g(X) = y_j\} = \sum_{j=1}^{\infty} y_j P\{X \in A_j\} \\ &= \sum_{j=1}^{\infty} y_j \sum_{x \in A_j} f(x) = \sum_{j=1}^{\infty} \sum_{x \in A_j} g(x)f(x). \end{aligned}$$

Because the A_j's are disjoint,

$$\sum_{j=1}^{\infty} \sum_{x \in A_j} g(x)f(x) = \sum_{x \in \cup_{j=1}^{\infty} A_j} g(x)f(x).$$

The theorem follows from making one final observation: $\cup_{j=1}^{\infty} A_j$ is the collection of all possible values of X. □

One can derive other properties of expectations by applying similar arguments. Here are some useful properties. For proof see the text.

Theorem 11.2. *Let* X *be a discrete random variable with mass function* f *and finite expectation* EX*. Then:*

(1) $E(aX + b) = aE(X) + b$ *for all constants* a, b*;*

(2) $Ea = a$ *for all nonrandom (constant) variables* a*;*

(3) *If* $P\{a \leqslant X \leqslant b\} = 1$*, then* $a \leqslant EX \leqslant b$*;*

(4) *If* $g(X)$ *and* $h(X)$ *have finite expectations, then*

$$E[g(X) + h(X)] = E[g(X)] + E[h(X)].$$

This is called linearity.

2. A first example

Suppose X has mass function

$$f(x) = \begin{cases} 1/4 & \text{if } x = 0, \\ 3/4 & \text{if } x = 1, \\ 0 & \text{otherwise.} \end{cases}$$

Recall: $EX = (\frac{1}{4} \times 0) + (\frac{3}{4} \times 1) = \frac{3}{4}$. Now let us compute $E(X^2)$ using Theorem 11.1:

$$E(X^2) = \left(\frac{1}{4} \times 0^2\right) + \left(\frac{3}{4} \times 1^2\right) = \frac{3}{4}.$$

Two observations:

(1) This is obvious because $X = X^2$ *in this particular example*; and

(2) $E(X^2) \neq (EX)^2$. In fact, the difference between $E(X^2)$ and $(EX)^2$ is an important quantity, called the *variance of* X. We will return to this topic later.

3. A second example

If $X = \text{Bin}(n, p)$, then what is $E(X^2)$? It may help to recall that $EX = np$. By Theorem 11.1,

$$E(X^2) = \sum_{k=0}^{n} k^2 \binom{n}{k} p^k q^{n-k} = \sum_{k=1}^{n} k \, \frac{n!}{(k-1)!(n-k)!} \, p^k q^{n-k}.$$

The question is, "how do we reduce the factor k further"? If we had $k-1$ instead of k, then this would be easy to answer. So let us first solve a related problem.

$$\begin{aligned}
\mathrm{E}\big[X(X-1)\big] &= \sum_{k=0}^{n} k(k-1)\binom{n}{k}p^k q^{n-k} = \sum_{k=2}^{n} k(k-1)\,\frac{n!}{k!(n-k)!}\,p^k q^{n-k}\\
&= n(n-1)\sum_{k=2}^{n}\frac{(n-2)!}{(k-2)!\big([n-2]-[k-2]\big)!}\,p^k q^{n-k}\\
&= n(n-1)\sum_{k=2}^{n}\binom{n-2}{k-2}p^k q^{n-k}\\
&= n(n-1)p^2\sum_{k=2}^{n}\binom{n-2}{k-2}p^{k-2}q^{[n-2]-[k-2]}\\
&= n(n-1)p^2\sum_{\ell=0}^{n-2}\binom{n-2}{\ell}p^{\ell}q^{[n-2]-\ell}.
\end{aligned}$$

The summand is the probability that $\mathrm{Bin}(n-2,p)$ is equal to ℓ. Since that probability is added over all of its possible values, the sum is one. Thus, we obtain $\mathrm{E}[X(X-1)] = n(n-1)p^2$. But $X(X-1) = X^2 - X$. Therefore, we can apply Theorem 11.2 to find that

$$\begin{aligned}
\mathrm{E}(X^2) &= \mathrm{E}[X(X-1)] + \mathrm{E}X = n(n-1)p^2 + np\\
&= (np)^2 + npq.
\end{aligned}$$

4. Expectation inequalities

Theorem 11.3 (The triangle inequality)**.** *If X has a finite expectation, then*

$$\big|\mathrm{E}X\big| \leqslant \mathrm{E}(|X|).$$

Proof. Let $g(x) = |x| - x$. This is a positive function, and $\mathrm{E}[g(X)] = \mathrm{E}(|X|) - \mathrm{E}X$. But $\mathrm{P}\{g(X) \geqslant 0\} = 1$. Therefore, $\mathrm{E}[g(X)] \geqslant 0$ by Theorem 11.2. This proves that $\mathrm{E}X \leqslant \mathrm{E}(|X|)$. Apply the same argument to $-X$ to find that $-\mathrm{E}X = \mathrm{E}(-X) \leqslant \mathrm{E}(|-X|) = \mathrm{E}(|X|)$. This proves that $\mathrm{E}X$ and $-\mathrm{E}X$ are both bounded above by $\mathrm{E}(|X|)$, which is the desired result. □

Theorem 11.4 (The Cauchy–Schwarz inequality)**.** *If $\mathrm{E}(|X|) < \infty$, then*

$$\mathrm{E}(|X|) \leqslant \sqrt{\mathrm{E}(X^2)}.$$

Proof. Expand the trivial bound $(|X| - E(|X|))^2 \geqslant 0$ to obtain:

$$X^2 - 2|X|E(|X|) + \left|E(|X|)\right|^2 \geqslant 0.$$

Take expectations, and note that $b = E(|X|)$ is nonrandom. This proves that

$$E(X^2) - 2E(|X|)E(|X|) + \left|E(|X|)\right|^2 \geqslant 0.$$

The left-hand side is manifestly equal to $E(X^2) - |E(|X|)|^2$, whence follows the theorem. □

One can use more advanced methods to prove the following:

$$E(|X|) \leqslant \sqrt{E(X^2)} \quad \text{for all random variables } X.$$

Note that $|X|$ and X^2 are nonnegative. So the expectations are defined, though possibly infinite. The preceding form of the Cauchy–Schwarz inequality implies that if $E(X^2)$ is finite, then so is $E(|X|)$.

Lecture 12

By the Cauchy–Schwarz inequality, if $E(X^2) < \infty$, then EX is well defined and finite as well. In that case, the *variance* of X is defined as

$$\operatorname{Var}(X) = E(X^2) - |EX|^2 .$$

In order to understand why this means anything, note that

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

Thus:

(1) We predict the as-yet-unseen value of X by the nonrandom number EX;

(2) $\operatorname{Var}(X)$ is the expected squared-error in this prediction. Note that $\operatorname{Var}(X)$ is also a nonrandom number.

1. Example 1

If $X = \text{Bin}(n, p)$, then we have seen that $EX = np$ and $E(X^2) = (np)^2 + npq$. Therefore, $\operatorname{Var}(X) = npq$.

2. Example 2

Suppose X has mass function

$$f(x) = \begin{cases} 1/4 & \text{if } x = 0, \\ 3/4 & \text{if } x = 1, \\ 0 & \text{otherwise.} \end{cases}$$

We saw in Lecture 11 that $EX = 3/4$. Now we compute the variance by first calculating

$$E(X^2) = \left(0^2 \times \frac{1}{4}\right) + \left(1^2 \times \frac{3}{4}\right) = \frac{3}{4}.$$

Thus,

$$\text{Var}(X) = \frac{3}{4} - \left(\frac{3}{4}\right)^2 = \frac{3}{4}\left(1 - \frac{3}{4}\right) = \frac{3}{16}.$$

3. Example 3

Let n be a fixed positive integer, and X takes any of the values $1, \ldots, n$ with equal probability. Then, $f(x) = 1/n$ if $x = 1, \ldots, n$; $f(x) = 0$, otherwise. Let us calculate the first two "moments" of X.[1] In this way, we obtain the mean and the variance of X.

The first moment is the expectation, or the mean, and is

$$EX = \sum_{k=1}^{n} \frac{k}{n} = \frac{1}{n} \times \frac{(n+1)n}{2} = \frac{n+1}{2}.$$

In order to compute $E(X^2)$ we need to know the algebraic identity:

$$\sum_{k=1}^{n} k^2 = \frac{(2n+1)(n+1)n}{6}. \tag{10}$$

This is proved by induction: For $n = 1$ it is elementary. Suppose it is true for $n-1$. Then write

$$\sum_{k=1}^{n} k^2 = \sum_{k=1}^{n-1} k^2 + n^2 = \frac{(2(n-1)+1)(n-1+1)(n-1)}{6} + n^2,$$

thanks to the induction hypothesis. Simplify to obtain

$$\begin{aligned}\sum_{k=1}^{n} k^2 &= \frac{(2n-1)n(n-1)}{6} + n^2 = \frac{(2n-1)(n^2-n)}{6} + n^2 \\ &= \frac{2n^3 - 3n^2 + n}{6} + \frac{6n^2}{6} = \frac{2n^3 + 3n^2 + n}{6} = \frac{n(2n^2+3n+1)}{6},\end{aligned}$$

which easily yields (10).

Thus,

$$E(X^2) = \sum_{k=1}^{n} \frac{k^2}{n} = \frac{1}{n} \times \frac{(2n+1)(n+1)n}{6} = \frac{(2n+1)(n+1)}{6}.$$

[1] It may help to recall that the pth moment of X is $E(X^p)$.

Therefore,

$$\begin{aligned}\mathrm{Var}(X) &= \frac{(2n+1)(n+1)}{6} - \left(\frac{n+1}{2}\right)^2 = \frac{2n^2+3n+1}{6} - \frac{n^2+2n+1}{4} \\ &= \frac{4n^2+6n+2}{12} - \frac{3n^2+6n+3}{12} \\ &= \frac{n^2-1}{12}.\end{aligned}$$

4. Example 4

Suppose $X = \mathrm{Poisson}(\lambda)$. We saw in Lecture 10 that $EX = \lambda$. In order to compute $E(X^2)$, we first compute $E[X(X-1)]$ and find that

$$\begin{aligned}E[X(X-1)] &= \sum_{k=0}^{\infty} k(k-1)\frac{e^{-\lambda}\lambda^k}{k!} = \sum_{k=2}^{\infty} \frac{e^{-\lambda}\lambda^k}{(k-2)!} \\ &= \lambda^2 \sum_{k=2}^{\infty} \frac{e^{-\lambda}\lambda^{k-2}}{(k-2)!}.\end{aligned}$$

The sum is equal to one; change variables ($j = k-2$) and recognize the jth term as the probability that $\mathrm{Poisson}(\lambda) = j$. Therefore,

$$E[X(X-1)] = \lambda^2.$$

Because $X(X-1) = X^2 - X$, the left-hand side is $E(X^2) - EX = E(X^2) - \lambda$. Therefore,

$$E(X^2) = \lambda^2 + \lambda.$$

It follows that

$$\mathrm{Var}(X) = \lambda.$$

5. Example 5

Suppose $f(x) = pq^{x-1}$ if $x = 1, 2, \ldots$; and $f(x) = 0$ otherwise. This is the Geometric(p) distribution. [The mass function for the first time to heads for a p-coin; see Lecture 8.] We have seen already that $EX = 1/p$ (Lecture 10). Let us find a new computation for this fact, and then go on and find also the variance.

$$\begin{aligned}EX &= \sum_{k=1}^{\infty} kpq^{k-1} = p\sum_{k=1}^{\infty} kq^{k-1} \\ &= p\,\frac{d}{dq}\left(\sum_{k=0}^{\infty} q^k\right) = p\,\frac{d}{dq}\left(\frac{1}{1-q}\right) = \frac{p}{(1-q)^2} = \frac{1}{p}.\end{aligned}$$

Next we compute $E(X^2)$ by first finding

$$\begin{aligned} E[X(X-1)] &= \sum_{k=1}^{\infty} k(k-1)pq^{k-1} = \frac{p}{q}\sum_{k=1}^{\infty} k(k-1)q^{k-2} \\ &= pq\,\frac{d^2}{dq^2}\left(\sum_{k=0}^{\infty} q^k\right) = \frac{p}{q}\,\frac{d^2}{dq^2}\left(\frac{1}{1-q}\right) \\ &= pq\,\frac{d}{dq}\left(\frac{1}{(1-q)^2}\right) = pq\,\frac{2}{(1-q)^3} = \frac{2q}{p^2}. \end{aligned}$$

Because $E[X(X-1)] = E(X^2) - EX = E(X^2) - (1/p)$, this proves that

$$E(X^2) = \frac{2q}{p^2} + \frac{1}{p} = \frac{2q+p}{p^2} = \frac{2-p}{p^2}.$$

Consequently,

$$\mathrm{Var}(X) = \frac{2-p}{p^2} - \frac{1}{p^2} = \frac{1-p}{p^2} = \frac{q}{p^2}.$$

For a wholly different solution, see Example (13) on page 124 of your text.

Lecture 13

1. Inequalities

Let us start with an inequality.

Lemma 13.1. *If h is a nonnegative function, then for all $\lambda > 0$,*

$$P\{h(X) \geqslant \lambda\} \leqslant \frac{E[h(X)]}{\lambda}.$$

Proof. We know already that

$$E[h(X)] = \sum_x h(x)f(x) \geqslant \sum_{x:\ h(x)\geqslant\lambda} h(x)f(x).$$

If x is such that $h(x) \geqslant \lambda$, then $h(x)f(x) \geqslant \lambda f(x)$, obviously. Therefore,

$$E[h(X)] \geqslant \lambda \sum_{x:\ h(x)\geqslant\lambda} f(x) = \lambda P\{h(X) \geqslant \lambda\}.$$

Divide by λ to finish. □

Thus, for example,

$$P\{|X| \geqslant \lambda\} \leqslant \frac{E(|X|)}{\lambda} \qquad \text{"Markov's inequality."}$$

$$P\{|X - EX| \geqslant \lambda\} \leqslant \frac{\mathrm{Var}(X)}{\lambda^2} \qquad \text{"Chebyshev's inequality."}$$

To get Markov's inequality, apply Lemma 13.1 with $h(x) = |x|$. To get Chebyshev's inequality, first note that $|X-EX| \geqslant \lambda$ if and only if $|X-EX|^2 \geqslant \lambda^2$. Then, apply Lemma 13.1 to find that

$$P\{|X - EX| \geqslant \lambda\} \leqslant \frac{E\left(|X - EX|^2\right)}{\lambda^2}.$$

Then, recall that the numerator is $\mathrm{Var}(X)$.

In words:

- If $E(|X|) < \infty$, then the probability that $|X|$ is large is small.
- If $\mathrm{Var}(X)$ is small, then with high probability $X \approx EX$.

2. Conditional distributions

If X is a random variable with mass function f, then $\{X = x\}$ is an event. Therefore, if B is also an event, and if $P(B) > 0$, then

$$P(X = x \mid B) = \frac{P(\{X = x\} \cap B)}{P(B)}.$$

As we vary the variable x, we note that $\{X = x\} \cap B$ are disjoint. Therefore,

$$\sum_x P(X = x \mid B) = \frac{\sum P(\{X = x\} \cap B)}{P(B)} = \frac{P\left(\cup_x \{X = x\} \cap B\right)}{P(B)} = 1.$$

Thus,

$$f(x \mid B) = P(X = x \mid B)$$

defines a mass function also. This is called the *conditional mass function of* X *given* B.

Example 13.2. Let X be distributed uniformly on $\{1, \ldots, n\}$, where n is a fixed positive integer. Recall that this means that

$$f(x) = \begin{cases} \dfrac{1}{n} & \text{if } x = 1, \ldots, n, \\ 0 & \text{otherwise.} \end{cases}$$

Choose and fix two integers a and b such that $1 \leqslant a \leqslant b \leqslant n$. Then,

$$P\{a \leqslant X \leqslant b\} = \sum_{x=a}^{b} \frac{1}{n} = \frac{b - a + 1}{n}.$$

Therefore,

$$f(x \mid a \leqslant X \leqslant b) = \begin{cases} \dfrac{1}{b - a + 1} & \text{if } x = a, \ldots, b, \\ 0 & \text{otherwise.} \end{cases}$$

3. Conditional expectations

Once we have a (conditional) mass function, we have also a conditional expectation at no cost. Thus,

$$E(X \mid B) = \sum_x x f(x \mid B).$$

Example 13.3 (Example 13.2, continued)**.** In Example 13.2,

$$E(X \mid a \leqslant X \leqslant b) = \sum_{k=a}^{b} \frac{k}{b-a+1}.$$

Now,

$$\begin{aligned}\sum_{k=a}^{b} k &= \sum_{k=1}^{b} k - \sum_{k=1}^{a-1} k \\ &= \frac{b(b+1)}{2} - \frac{(a-1)a}{2} \\ &= \frac{b^2+b-a^2+a}{2}.\end{aligned}$$

Write $b^2 - a^2 = (b-a)(b+a)$ and factor $b+a$ to get

$$\sum_{k=a}^{b} k = \frac{b+a}{2}(b-a+1).$$

Therefore,

$$E(X \mid a \leqslant X \leqslant b) = \frac{b+a}{2}.$$

This should not come as a surprise. Example 13.2 actually shows that given $B = \{a \leqslant X \leqslant b\}$, the conditional distribution of X given B is uniform on $\{a, \ldots, b\}$. Therefore, the conditional expectation is the expectation of a uniform random variable on $\{a, \ldots, b\}$.

Theorem 13.4 (Bayes's formula for conditional expectations)**.** *If* $P(B) > 0$, *then*

$$EX = E(X \mid B)P(B) + E(X \mid B^c)P(B^c).$$

Proof. We know from the ordinary Bayes's formula that

$$f(x) = f(x \mid B)P(B) + f(x \mid B^c)P(B^c).$$

Multiply both sides by x and add over all x to finish. □

Remark 13.5. The more general version of Bayes's formula works too here: Suppose $B_1, B_2, \ldots$ are disjoint and $\cup_{i=1}^{\infty} B_i = \Omega$; i.e., "one of the B_i's happens." Then,

$$EX = \sum_{i=1}^{\infty} E(X \mid B_i) P(B_i).$$

Example 13.6. Suppose you play a fair game repeatedly. At time 0, before you start playing the game, your fortune is zero. In each play, you win or lose with probability 1/2. Let T_1 be the first time your fortune becomes $+1$. Compute $E(T_1)$.

More generally, let T_x denote the first time to win x dollars, where $T_0 = 0$.

Let W denote the event that you win the first round. Then, $P(W) = P(W^c) = 1/2$, and so

$$E(T_x) = \frac{1}{2} E(T_x \mid W) + \frac{1}{2} E(T_x \mid W^c). \tag{11}$$

Suppose $x \neq 0$. Given W, T_x is one plus the first time to make $x-1$ more dollars. Given W^c, T_x is one plus the first time to make $x+1$ more dollars. Therefore,

$$\begin{aligned} E(T_x) &= \frac{1}{2}\Big[1 + E(T_{x-1})\Big] + \frac{1}{2}\Big[1 + E(T_{x+1})\Big] \\ &= 1 + \frac{E(T_{x-1}) + E(T_{x+1})}{2}. \end{aligned}$$

Also $E(T_0) = 0$.

Let $g(x) = E(T_x)$. This shows that $g(0) = 0$ and

$$g(x) = 1 + \frac{g(x+1) + g(x-1)}{2} \qquad \text{for } x = \pm 1, \pm 2, \ldots .$$

Because $g(x) = (g(x) + g(x))/2$,

$$g(x) + g(x) = 2 + g(x+1) + g(x-1) \qquad \text{for } x = \pm 1, \pm 2, \ldots .$$

Solve to find that for all integers $x \geqslant 1$,

$$g(x+1) - g(x) = -2 + g(x) - g(x-1).$$

Lecture 14

Example 14.1 (St.-Petersbourg paradox, continued)**.** We continued with our discussion of the St.-Petersbourg paradox, and note that for all integers $N \geqslant 1$,

$$\begin{aligned} g(N) &= g(1) + \sum_{k=2}^{N} \Big(g(k) - g(k-1) \Big) \\ &= g(1) + \sum_{k=1}^{N-1} \Big(g(k+1) - g(k) \Big) \\ &= g(1) + \sum_{k=1}^{N-1} \Big(-2 + g(k) - g(k-1) \Big) \\ &= g(1) - 2(N-1) + \sum_{k=1}^{N} \Big(g(k) - g(k-1) \Big) \\ &= g(1) - 2(N-1) + g(N). \end{aligned}$$

If $g(1) < \infty$, then $g(1) = 2(N-1)$. But N is arbitrary. Therefore, $g(1)$ cannot be finite; i.e.,

$$E(T_1) = \infty.$$

This shows also that $E(T_x) = \infty$ for all $x \geqslant 1$, because for example $T_2 \geqslant 1 + T_1$! By symmetry, $E(T_x) = \infty$ if x is a negative integer as well.

1. Joint distributions

If X and Y are two discrete random variables, then their *joint mass function* is

$$f(x, y) = P\{X = x, Y = y\}.$$

We might write $f_{X,Y}$ in place of f in order to emphasize the dependence on the two random variables X and Y.

Here are some properties of $f_{X,Y}$:

- $f(x,y) \geqslant 0$ for all x, y;
- $\sum_x \sum_y f(x,y) = 1$;
- $\sum_{(x,y)\in C} f(x,y) = P\{(X,Y) \in C\}$.

Example 14.2. You roll two fair dice. Let X be the number of 2s shown, and Y the number of 4s. Then X and Y are discrete random variables, and

$$f(x,y) = P\{X = x, Y = y\}$$
$$= \begin{cases} \frac{1}{36} & \text{if } x = 2 \text{ and } y = 0, \\ \frac{1}{36} & \text{if } x = 0 \text{ and } y = 2, \\ \frac{2}{36} & \text{if } x = y = 1, \\ \frac{8}{36} & \text{if } x = 0 \text{ and } y = 1, \\ \frac{8}{36} & \text{if } x = 1 \text{ and } y = 0, \\ \frac{16}{36} & \text{if } x = y = 0, \\ 0 & \text{otherwise.} \end{cases}$$

Some times it helps to draw up a table of "joint probabilities":

$x \setminus y$	**0**	**1**	**2**
0	16/36	8/36	1/36
1	8/36	2/36	0
2	1/36	0	0

From this we can also calculate f_X and f_Y. For instance,

$$f_X(1) = P\{X = 1\} = f(1,0) + f(1,1) = \frac{10}{36}.$$

In general, you compute the row sums (f_X) and put them in the margin; you do the same with the column sums (f_Y) and put them in the bottom row. In this way, you obtain:

$x \setminus y$	**0**	**1**	**2**	f_X
0	16/36	8/36	1/36	25/36
1	8/36	2/36	0	10/36
2	1/36	0	0	1/36
f_Y	25/36	10/36	1/36	**1**

The "**1**" designates the right-most column sum (which should be one), and/or the bottom-row sum (which should also be one). This is also the sum of the elements of the table (which should also be one).

En route we have discovered the next result, as well.

Theorem 14.3. *For all* x, y*:*

(1) $f_X(x) = \sum_b f(x, b)$.
(2) $f_Y(y) = \sum_a f(a, y)$.

2. Independence

Definition 14.4. Let X and Y be discrete with joint mass function f. We say that X and Y are *independent* if for all x, y,

$$f(x, y) = f_X(x) f_Y(y).$$

- Suppose A and B are two sets, and X and Y are independent. Then,

$$\begin{aligned} P\{X \in A, Y \in B\} &= \sum_{x \in A} \sum_{y \in B} f(x, y) \\ &= \sum_{x \in A} f_X(x) \sum_{y \in B} f_Y(y) \\ &= P\{X \in A\} P\{Y \in B\}. \end{aligned}$$

- Similarly, if h and g are functions, then $h(X)$ and $g(Y)$ are independent as well.
- All of this makes sense for more than 2 random variables as well.

Example 14.5 (Example 14.2, continued)**.** Note that in this example, X and Y are not independent. For instance,

$$f(1, 2) = 0 \neq f_X(1) f_Y(2) = \frac{10}{36} \times \frac{1}{36}.$$

Now, let us find the distribution of $Z = X + Y$. The possible values are 0, 1, and 2. The probabilities are

$$\begin{aligned} f_Z(0) &= f_{X,Y}(0, 0) = \frac{16}{36} \\ f_Z(1) &= f_{X,Y}(1, 0) + f_{X,Y}(0, 1) = \frac{8}{36} + \frac{8}{36} = \frac{16}{36} \\ f_Z(2) &= f_{X,Y}(0, 2) + f_{X,Y}(2, 0) + f_{X,Y}(1, 1) = \frac{1}{36} + \frac{1}{36} + \frac{2}{36} = \frac{4}{36}. \end{aligned}$$

That is,

$$f_Z(x) = \begin{cases} \frac{16}{36} & \text{if } x = 0 \text{ or } 1, \\ \frac{4}{36} & \text{if } x = 2, \\ 0 & \text{otherwise.} \end{cases}$$

Example 14.6. *Let* $X = \text{geometric}(p_1)$ *and* $Y = \text{geometric}(p_2)$ *be independent. What is the mass function of* $Z = \min(X, Y)$*?*

Recall from Lecture 9 that $P\{X \geqslant n\} = q_1^{n-1}$ and $P\{Y \geqslant n\} = q_2^{n-1}$ for all integers $n \geqslant 1$. Therefore,

$$\begin{aligned} P\{Z \geqslant n\} &= P\{X \geqslant n, Y \geqslant n\} = P\{X \geqslant n\}P\{Y \geqslant n\} \\ &= (q_1 q_2)^{n-1}, \end{aligned}$$

as long as $n \geqslant 1$ is an integer. Because $P\{Z \geqslant n\} = P\{Z = n\} + P\{Z \geqslant n+1\}$, for all integers $n \geqslant 1$,

$$\begin{aligned} P\{Z = n\} &= P\{Z \geqslant n\} - P\{Z \geqslant n+1\} = (q_1 q_2)^{n-1} - (q_1 q_2)^n \\ &= (q_1 q_2)^{n-1}(1 - q_1 q_2). \end{aligned}$$

Else, $P\{Z = n\} = 0$. Thus, $Z = \text{geometric}(p)$, where $p = 1 - q_1 q_2$.

Lecture 15

1. Expectations

Theorem 15.1. *Let* g *be a real-valued function of two variables, and* (X, Y) *have joint mass function* f*. If the sum converges then*

$$E[g(X,Y)] = \sum_x \sum_y g(x,y) f(x,y).$$

Corollary 15.2. *For all* a, b *real,*

$$E(aX + bY) = aEX + bEY.$$

Proof. Setting $g(x,y) = ax + by$ yields

$$\begin{aligned} E(aX + bY) &= \sum_x \sum_y (ax + by) f(x,y) \\ &= \sum_x ax \sum_y f(x,y) + \sum_x \sum_y by f(x,y) \\ &= a \sum_x x f_X(x) + b \sum_y y \sum_x f(x,y) \\ &= aEX + b \sum_y f_Y(y), \end{aligned}$$

which is $aEX + bEY$. □

2. Covariance and correlation

Theorem 15.3 (Cauchy–Schwarz inequality)**.** *If* $\mathrm{E}(X^2)$ *and* $\mathrm{E}(Y^2)$ *are finite, then*

$$|\mathrm{E}(XY)| \leqslant \sqrt{\mathrm{E}(X^2)\,\mathrm{E}(Y^2)}.$$

Proof. Note that

$$\begin{aligned}&\left(X\mathrm{E}(Y^2) - Y\mathrm{E}(XY)\right)^2\\ &\qquad = X^2\left(\mathrm{E}(Y^2)\right)^2 + Y^2\left(\mathrm{E}(XY)\right)^2 - 2XY\mathrm{E}(Y^2)\mathrm{E}(XY).\end{aligned}$$

Therefore, we can take expectations of both side to find that

$$\begin{aligned}&\mathrm{E}\left[\left(X\mathrm{E}(Y^2) - Y\mathrm{E}(XY)\right)^2\right]\\ &\qquad = \mathrm{E}(X^2)\left(\mathrm{E}(Y^2)\right)^2 + \mathrm{E}(Y^2)\left(\mathrm{E}(XY)\right)^2 - 2\mathrm{E}(Y^2)\left(\mathrm{E}(XY)\right)^2\\ &\qquad = \mathrm{E}(X^2)\left(\mathrm{E}(Y^2)\right)^2 - \mathrm{E}(Y^2)\left(\mathrm{E}(XY)\right)^2.\end{aligned}$$

The left-hand side is $\geqslant 0$. Therefore, so is the right-hand side. Solve to find that

$$\mathrm{E}(X^2)\mathrm{E}(Y^2) \geqslant \left(\mathrm{E}(XY)\right)^2.$$

[If $\mathrm{E}(Y^2) > 0$, then this is OK. Else, $\mathrm{E}(Y^2) = 0$, which means that $\mathrm{P}\{Y = 0\} = 1$. In that case the result is true, but tautologically.] □

Thus, if $\mathrm{E}(X^2)$ and $\mathrm{E}(Y^2)$ are finite, then $\mathrm{E}(XY)$ is finite as well. In that case we can define the *covariance* between X and Y to be

$$\mathrm{Cov}(X, Y) = \mathrm{E}\left[(X - \mathrm{E}X)(Y - \mathrm{E}Y)\right]. \tag{12}$$

Because $(X - \mathrm{E}X)(Y - \mathrm{E}Y) = XY - X\mathrm{E}Y - Y\mathrm{E}X + \mathrm{E}X\mathrm{E}Y$, we obtain the following, which is the computationally useful formula for covariance:

$$\mathrm{Cov}(X, Y) = \mathrm{E}(XY) - \mathrm{E}(X)\mathrm{E}(Y). \tag{13}$$

Note, in particular, that $\mathrm{Cov}(X, X) = \mathrm{Var}(X)$.

Theorem 15.4. *Suppose* $\mathrm{E}(X^2)$ *and* $\mathrm{E}(Y^2)$ *are finite. Then, for all nonrandom* a, b, c, d*:*

(1) $\mathrm{Cov}(aX + b\,, cY + d) = ac\mathrm{Cov}(X, Y)$*;*

(2) $\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\mathrm{Cov}(X, Y)$.

Proof. Let $\mu = \mathrm{E}X$ and $\nu = \mathrm{E}Y$ for brevity. We then have

$$\begin{aligned}\mathrm{Cov}(aX + b\,, cY + d) &= \mathrm{E}\left[(aX + b - (a\mu + b))(cY + d - (c\nu + d))\right]\\ &= \mathrm{E}\left[(a(X - \mu))(c(Y - \nu))\right]\\ &= ac\mathrm{Cov}(X, Y).\end{aligned}$$

Similarly,

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

Now identify the terms. □

Lecture 16

1. Some examples

Example 16.1 (Example 14.2, continued)**.** We find that

$$\mathrm{E}(XY) = \left(1 \times 1 \times \frac{2}{36}\right) = \frac{2}{36}.$$

Also,

$$\mathrm{E}X = \mathrm{E}Y = \left(1 \times \frac{10}{36}\right) + \left(2 \times \frac{1}{36}\right) = \frac{12}{36}.$$

Therefore,

$$\mathrm{Cov}(X, Y) = \frac{2}{36} - \left(\frac{12}{36} \times \frac{12}{36}\right) = -\frac{72}{1296} = -\frac{1}{18}.$$

The *correlation* between X and Y is the quantity,

$$\rho(X, Y) = \frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{Var}(X)\ \mathrm{Var}(Y)}}. \tag{14}$$

Example 16.2 (Example 14.2, continued)**.** Note that

$$\mathrm{E}(X^2) = \mathrm{E}(Y^2) = \left(1^2 \times \frac{10}{36}\right) + \left(2^2 \times \frac{1}{36}\right) = \frac{14}{36}.$$

Therefore,

$$\mathrm{Var}(X) = \mathrm{Var}(Y) = \frac{14}{36} - \left(\frac{12}{36}\right)^2 = \frac{360}{1296} = \frac{5}{13}.$$

Therefore, the correlation between X and Y is

$$\rho(X, Y) = -\frac{1/18}{\sqrt{\left(\frac{5}{13}\right)\left(\frac{5}{13}\right)}} = -\frac{13}{90}.$$

2. Correlation and independence

The following is a variant of the Cauchy–Schwarz inequality. I will not prove it, but it would be nice to know the following.

Theorem 16.3. *If* $E(X^2)$ *and* $E(Y^2)$ *are finite, then* $-1 \leqslant \rho(X,Y) \leqslant 1$.

We say that X and Y are *uncorrelated* if $\rho(X,Y) = 0$; equivalently, if $\mathrm{Cov}(X,Y) = 0$. A significant property of uncorrelated random variables is that $\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$; see Theorem 15.4(2).

Theorem 16.4. *If* X *and* Y *are independent [with joint mass function* f*], then they are uncorrelated.*

Proof. It suffices to prove that $E(XY) = E(X)E(Y)$. But

$$\begin{aligned} E(XY) &= \sum_x \sum_y xy f(x,y) = \sum_x \sum_y xy f_X(x) f_Y(y) \\ &= \sum_x x f_X(x) \sum_y y f_Y(y) = E(X)E(Y), \end{aligned}$$

as planned. □

Example 16.5 (A counter example)**.** Sadly, it is only too common that people some times think that the converse to Theorem 16.4 is also true. So let us dispel this with a counterexample: Let Y and Z be two independent random variables such that $Z = \pm 1$ with probability $1/2$ each; and $Y = 1$ or 2 with probability $1/2$ each. Define $X = YZ$. Then, I claim that X and Y are uncorrelated but not independent.

First, note that $X = \pm 1$ and ± 2, with probability $1/4$ each. Therefore, $E(X) = 0$. Also, $XY = Y^2 Z = \pm 1$ and ± 4 with probability $1/4$ each. Therefore, again, $E(XY) = 0$. It follows that

$$\mathrm{Cov}(X,Y) = \underbrace{E(XY)}_{0} - \underbrace{E(X)}_{0} E(Y) = 0.$$

Thus, X and Y are uncorrelated. But they are not independent. Intuitively speaking, this is clear because $|X| = Y$. Here is one way to logically justify our claim:

$$P\{X = 1, Y = 2\} = 0 \neq \frac{1}{8} = P\{X = 1\}P\{Y = 2\}.$$

Example 16.6 (Binomials)**.** Let $X = \mathrm{Bin}(n,p)$ denote the total number of successes in n independent success/failure trials, where $P\{\text{success per trial}\} = p$. Define I_j to be one if the jth trial leads to a success; else $I_j = 0$. The key observation is that

$$X = I_1 + \cdots + I_n.$$

Note that $E(I_j) = 1 \times p = p$ and $E(I_j^2) = E(I_j) = p$, whence $\text{Var}(I_j) = p - p^2 = pq$. Therefore,

$$E(X) = \sum_{j=1}^{n} E(I_j) = np \quad \text{and} \quad \text{Var}(X) = \sum_{j=1}^{n} \text{Var}(I_j) = npq.$$

3. The law of large numbers

Theorem 16.7. *Suppose $X_1, X_2, \ldots, X_n$ are independent, all with the same mean μ and variance $\sigma^2 < \infty$. Then for all $\epsilon > 0$, however small,*

$$\lim_{n\to\infty} P\left\{\left|\frac{X_1 + \cdots + X_n}{n} - \mu\right| \geqslant \epsilon\right\} = 0. \tag{15}$$

Lemma 16.8. *Suppose $X_1, X_2, \ldots, X_n$ are independent, all with the same mean μ and variance $\sigma^2 < \infty$. Then:*

$$E\left(\frac{X_1 + \cdots + X_n}{n}\right) = \mu$$
$$\text{Var}\left(\frac{X_1 + \cdots + X_n}{n}\right) = \frac{\sigma^2}{n}.$$

Proof of Theorem 16.7. Recall *Chebyshev's inequality*: For all random variables Z with $E(Z^2) < \infty$, and all $\epsilon > 0$,

$$P\{|Z - EZ| \geqslant \epsilon\} \leqslant \frac{\text{Var}(Z)}{\epsilon^2}.$$

We apply this with $Z = (X_1 + \cdots + X_n)/n$, and then use use Lemma 16.8 to find that for all $\epsilon > 0$,

$$P\left\{\left|\frac{X_1 + \cdots + X_n}{n} - \mu\right| \geqslant \epsilon\right\} \leqslant \frac{\sigma^2}{n\epsilon^2}.$$

Let $n \nearrow \infty$ to finish. □

Proof of Lemma 16.8. It suffices to prove that

$$E(X_1 + \cdots + X_n) = n\mu$$
$$\text{Var}(X_1 + \cdots + X_n) = n\sigma^2.$$

We prove this by induction. Indeed, this is obviously true when $n = 1$. Suppose it is OK for all integers $\leqslant n - 1$. We prove it for n.

$$\begin{aligned} E(X_1 + \cdots + X_n) &= E(X_1 + \cdots + X_{n-1}) + EX_n \\ &= (n-1)\mu + EX_n, \end{aligned}$$

by the induction hypothesis. Because $EX_n = \mu$, the preceding is equal to $n\mu$, as planned. Now we verify the more interesting variance computation.

Once again, we assume the assertion holds for all integers $\leqslant n-1$, and strive to check it for n.

Define

$$Y = X_1 + \cdots + X_{n-1}.$$

Because Y is independent of X_n, $\mathrm{Cov}(Y, X_n) = 0$. Therefore, by Lecture 15,

$$\begin{aligned}\mathrm{Var}\,(X_1 + \cdots + X_n) &= \mathrm{Var}(Y + X_n)\\ &= \mathrm{Var}(Y) + \mathrm{Var}(X_n) + \mathrm{Cov}(Y, X_n)\\ &= \mathrm{Var}(Y) + \mathrm{Var}(X_n).\end{aligned}$$

We know that $\mathrm{Var}(X_n) = \sigma^2$, and by the induction hypothesis, $\mathrm{Var}(Y) = (n-1)\sigma^2$. The result follows. □

Lecture 17

1. Wrap-up of Lecture 16

Proof of Lemma 16.8. It suffices to prove that

$$\begin{aligned}\mathrm{E}(X_1 + \cdots + X_n) &= n\mu\\ \mathrm{Var}(X_1 + \cdots + X_n) &= n\sigma^2.\end{aligned}$$

We prove this by induction. Indeed, this is obviously true when $n = 1$. Suppose it is OK for all integers $\leqslant n-1$. We prove it for n.

$$\begin{aligned}\mathrm{E}(X_1 + \cdots + X_n) &= \mathrm{E}(X_1 + \cdots + X_{n-1}) + EX_n\\ &= (n-1)\mu + EX_n,\end{aligned}$$

by the induction hypothesis. Because $EX_n = \mu$, the preceding is equal to $n\mu$, as planned. Now we verify the more interesting variance computation.

Once again, we assume the assertion holds for all integers $\leqslant n-1$, and strive to check it for n.

Define

$$Y = X_1 + \cdots + X_{n-1}.$$

Because Y is independent of X_n, $\mathrm{Cov}(Y, X_n) = 0$. Therefore, by Lecture 15,

$$\begin{aligned}\mathrm{Var}(X_1 + \cdots + X_n) &= \mathrm{Var}(Y + X_n)\\ &= \mathrm{Var}(Y) + \mathrm{Var}(X_n) + \mathrm{Cov}(Y, X_n)\\ &= \mathrm{Var}(Y) + \mathrm{Var}(X_n).\end{aligned}$$

We know that $\mathrm{Var}(X_n) = \sigma^2$, and by the induction hypothesis, $\mathrm{Var}(Y) = (n-1)\sigma^2$. The result follows. □

2. Conditioning

2.1. Conditional mass functions. For all y, define the conditional mass function of X given that $Y = y$ as

$$\begin{aligned} f_{X|Y}(x\,|\,y) = P\left(X = x \,\middle|\, Y = y\right) &= \frac{P\{X = x\,, Y = y\}}{P\{Y = y\}} \\ &= \frac{f(x\,,y)}{f_Y(y)}, \end{aligned} \tag{16}$$

provided that $f_Y(y) > 0$.

As a function in x, $f_{X|Y}(x\,|\,y)$ is a probability mass function. That is:

(1) $0 \leqslant f_{X|Y}(x\,|\,y) \leqslant 1$;

(2) $\sum_x f_{X|Y}(x\,|\,y) = 1$.

Example 17.1 (Example 14.2, Lecture 14, continued)**.** In this example, the joint mass function of (X, Y), and the resulting marginal mass functions, were given by the following:

$x \setminus y$	**0**	**1**	**2**	f_X
0	16/36	8/36	1/36	25/36
1	8/36	2/36	0	10/36
2	1/36	0	0	1/36
f_Y	25/36	10/36	1/36	**1**

Let us calculate the conditional mass function of X, given that $Y = 1$:

$$\begin{aligned} f_{X|Y}(0\,|\,1) &= \frac{f(0\,,1)}{f_Y(1)} = \frac{8}{10} \\ f_{X|Y}(1\,|\,1) &= \frac{f(1\,,1)}{f_Y(1)} = \frac{2}{10} \\ f_{X|Y}(x\,|\,1) &= 0 \text{ for other values of } x. \end{aligned}$$

Similarly,

$$\begin{aligned} f_{X|Y}(0\,|\,0) &= \frac{16}{25} \\ f_{X|Y}(1\,|\,0) &= \frac{8}{25} \\ f_{X|Y}(2\,|\,0) &= \frac{1}{25} \\ f_{X|Y}(x\,|\,0) &= 0 \text{ for other values of } x, \end{aligned}$$

and

$$f_{X|Y}(0\,|\,2) = 1$$
$$f_{X|Y}(x\,|\,2) = 0 \text{ for other values of } x.$$

2.2. Conditional expectations. Define conditional expectations, as we did ordinary expectations. But use conditional probabilities in place of ordinary probabilities, viz.,

$$E(X\,|\,Y = y) = \sum_x x f_{X|Y}(x\,|\,y). \tag{17}$$

Example 17.2 (Example 17.1, continued)**.** Here,

$$E(X\,|\,Y = 1) = \left(0 \times \frac{8}{10}\right) + \left(1 \times \frac{2}{10}\right) = \frac{2}{10} = \frac{1}{5}.$$

Similarly,

$$E(X\,|\,Y = 0) = \left(0 \times \frac{16}{25}\right) + \left(1 \times \frac{8}{25}\right) + \left(2 \times \frac{1}{25}\right) = \frac{10}{25} = \frac{2}{5},$$

and

$$E(X\,|\,Y = 2) = 0.$$

Note that $E(X) = 12/36 = 1/3$, which is none of the preceding. If you know that $Y = 0$, then your best bet for X is $2/5$. But if you have no extra knowledge, then your best bet for X is $1/3$.

However, let us note the Bayes's formula in action:

$$\begin{aligned}
&E(X)\\
&= E(X\,|\,Y = 0)P\{Y = 0\} + E(X\,|\,Y = 1)P\{Y = 1\} + E(X\,|\,Y = 2)P\{Y = 2\}\\
&= \left(\frac{2}{5} \times \frac{25}{36}\right) + \left(\frac{1}{5} \times \frac{10}{36}\right) + \left(0 \times \frac{1}{36}\right)\\
&= \frac{12}{36},
\end{aligned}$$

as it should be.

3. Sums of independent random variables

Theorem 17.3. *If X and Y are independent, then*

$$f_{X+Y}(z) = \sum_x f_X(x) f_Y(z - x).$$

Proof. We note that $X + Y = z$ if $X = x$ for some x and $Y = z - x$ for that x. For example, suppose X is integer-valued and $\geqslant 1$. Then $\{X + Y = z\} = \cup_{x=1}^{\infty} P\{X = x, Y = z - x\}$. In general,

$$f_{X+Y}(z) = \sum_x P\{X = x, Y = z - x\} = \sum_x P\{X = x\}P\{Y = z - x\}.$$

This is the desired result. □

Example 17.4. Suppose $X = \text{Poisson}(\lambda)$ and $Y = \text{Poisson}(\gamma)$ are independent. Then, I claim that $X + Y = \text{Poisson}(\lambda + \gamma)$. We verify this by directly computing as follows: The possible values of $X + Y$ are $0, 1, \ldots$. Let $z = 0, 1, \ldots$ be a possible value, and then check that

$$\begin{aligned} f_{X+Y}(z) &= \sum_{x=0}^{\infty} f_X(x) f_Y(z - x) \\ &= \sum_{x=0}^{\infty} \frac{e^{-\lambda}\lambda^x}{x!} f_Y(z - x) \\ &= \sum_{x=0}^{z} \frac{e^{-\lambda}\lambda^x}{x!} \frac{e^{-\gamma}\gamma^{z-x}}{(z-x)!} \\ &= \frac{e^{-(\lambda+\gamma)}}{z!} \sum_{x=0}^{z} \binom{z}{x} \lambda^x \gamma^{z-x} \\ &= \frac{e^{-(\lambda+\gamma)}}{z!} (\lambda + \gamma)^z, \end{aligned}$$

thanks to the binomial theorem. For other values of z, it is easy to see that $f_{X+Y}(z) = 0$.

Lecture 18

1. The distribution of the sum of two independent random variables, continued

Recall that if X and Y are independent, then

$$f_{X+Y}(z) = \sum_x f_X(x) f_Y(z-x).$$

Now we work out three examples of this. [We have seen another already at the end of Lecture 17.]

Example 18.1. Suppose $X = \pm 1$ with probability $1/2$ each; and $Y = \pm 2$ with probability $1/2$ each. Then,

$$f_{X+Y}(z) = \begin{cases} 1/4 & \text{if } z = 3, -3, 1, -1, \\ 0 & \text{otherwise.} \end{cases}$$

Example 18.2. Let X and Y denote two independent geometric(p) random variables with the same parameter $p \in (0,1)$. What is the mass function of $X+Y$? If $z = 2, 3, \ldots$, then

$$\begin{aligned} f_{X+Y}(z) &= \sum_x f_X(x) f_Y(z-x) = \sum_{x=1}^{\infty} p q^{x-1} f_Y(z-x) \\ &= \sum_{x=1}^{z+1} p q^{x-1} p q^{z-x-1} = p^2 \sum_{x=1}^{z+1} q^{z-2} = (z+1) p^2 q^{z-2}. \end{aligned}$$

Else, $f_{X+Y}(z) = 0$. This shows that $X+Y$ is a negative binomial. Can you deduce this directly, and by other means?

Example 18.3. If $X = \text{bin}(n, p)$ and $Y = \text{bin}(m, p)$ for the same parameter $p \in (0, 1)$, then what is the distribution of $X + Y$? If $z = 0, 1, \ldots, n + m$, then

$$\begin{aligned} f_{X+Y}(z) &= \sum_x f_X(x) f_Y(z-x) = \sum_{x=0}^{n} \binom{n}{x} p^x q^{n-x} \\ &= \sum_{\substack{0 \leqslant x \leqslant n \\ 0 \leqslant z-x \leqslant m}} \binom{n}{x} p^x q^{n-x} \binom{m}{z-x} p^{z-x} q^{m-(z-x)} \\ &= p^z q^{m+n-z} \sum_{\substack{0 \leqslant x \leqslant n \\ z-m \leqslant x \leqslant z}} \binom{n}{x} \binom{m}{z-x}. \end{aligned}$$

[The sum is over all integers x such that x is between 0 and n, *and* x is also betweem $z - m$ and m.] For other values of z, $f_{X+Y}(z) = 0$.

Equivalently, we can write for all $z = 0, \ldots, n + m$,

$$f_{X+Y}(z) = \binom{n+m}{z} p^z q^{m+n-z} \sum_{\substack{0 \leqslant x \leqslant n \\ z-m \leqslant x \leqslant z}} \frac{\binom{n}{x}\binom{m}{z-x}}{\binom{n+m}{z}}.$$

Thus, if we showed that the sum is one, then $X + Y = \text{bin}(n + m, p)$. In order to show that the sum is one consider an urn that has n white balls and m black balls. We choose z balls at random, without replacement. The probability that we obtain exactly x white and $z - x$ black is precisely,

$$\frac{\binom{n}{x}\binom{m}{z-x}}{\binom{n+m}{z}}.$$

Therefore, if we add this probability over all possible values of x we should get one. This does the job.

Can you find a direct way to prove that $X + Y = \text{bin}(n + m, p)$?

2. Transformations of a mass function

Let f denote the mass function of a random variable. For technical reasons, one often "transforms" f into a new function which is easier to analyze some times. The transformation can be fairly arbitrary, but it should be possible, in principle, to compute f from that transformation as well. In

this way, the computations for the transform will often yield useful computations for the original mass function. [We do this only when it is very hard to work with the mass function directly.]

In this course we will study only two transformations: The generating function, and the moment generating function.

2.1. The generating function. If X is integer valued, then its *generating function* [also known as the "probability generating function," or p.g.f., for short] G is the function

$$G(s) = \sum_k s^k f(s) \qquad \text{for all } s \in (-1,1).$$

That is, we start with some mass function f, and transform it into another function—the generating function—G. Note that

$$G(s) = E[s^X].$$

This is indeed a useful transformation. Indeed,

Theorem 18.4 (Uniqueness). *If* $G_X(s) = G_Y(s)$ *for all* $s \in (-1,1)$, *then* $f_X = f_Y$.

In order to go from G to f we need a lot of examples. In this course, we will work out a few. Many more are known.

Example 18.5. Suppose X is uniformly distributed on $\{-n,\dots,m\}$, where n and m are positive integers. This means that $f(x) = 1/(m+n+1)$ if $x = -n,\dots,m$ and $f(x) = 0$ otherwise. Consequently, for all $s \in (-1,1)$,

$$\begin{aligned} G(s) &= \sum_{x=-n}^{m} \frac{s^x}{n+m+1} = \frac{1}{n+m+1} \sum_{x=-n}^{m} s^x \\ &= \frac{s^{-n} - s^{m+1}}{(n+m+1)(1-s)}, \end{aligned}$$

using facts about geometric series.

Example 18.6. Suppose

$$G(s) = \frac{(\alpha-1)s}{\alpha - s} \qquad \text{for all } s \in (-1,1),$$

where $\alpha > 1 > 0$. I claim that G is a p.g.f. The standard way to do this is to expand G into a Taylor expansion. Define

$$h(s) = \frac{1}{\alpha - s} = (\alpha - s)^{-1}.$$

Then, $h'(s) = (\alpha - s)^{-2}$, $h''(s) = 2(\alpha - s)^{-3}$, etc., and in general,

$$h^{(n)}(s) = n!(\alpha - s)^{-(n+1)}.$$

According to the Taylor-MacLaurin expansion of h,

$$h(s) = \sum_{n=0}^{\infty} \frac{1}{n!} s^n h^{(n)}(0).$$

Note that $h^{(n)}(0) = \alpha^{-1} n! \alpha^{-n}$. Therefore, as long as $0 < s/\alpha < 1$,

$$\frac{1}{\alpha - s} = \frac{1}{\alpha} \sum_{n=0}^{\infty} \left(\frac{s}{\alpha}\right)^n.$$

In particular,

$$G(s) = \frac{(\alpha - 1)s}{\alpha} \sum_{n=0}^{\infty} s^n (1/\alpha)^n = \frac{\alpha - 1}{\alpha} \sum_{k=1}^{\infty} s^k (1/\alpha)^{k-1}.$$

By the uniqueness theorem,

$$f(k) = \begin{cases} \dfrac{\alpha - 1}{\alpha} \left(\dfrac{1}{\alpha}\right)^{k-1} & \text{if } k = 1, 2, \ldots, \\ 0 & \text{otherwise.} \end{cases}$$

Thus, in fact, $X = \text{geometric}(1/\alpha)$.

Lecture 19

1. Transformations of a mass function

1.1. The generating function. Recall that if X is an integer-valued random variable, then its [probability] *generating function*(p.g.f.) is

$$G(s) = E[s^X] = \sum_{k=-\infty}^{\infty} s^k f(k) \qquad \text{for all } -1 < s < 1.$$

1.2. The moment generating function. The *moment generating function* (m.g.f.) of a random variable X is

$$M(s) = E[e^{sX}] = \sum_x e^{sx} f(x),$$

provided that the sum exists.

This is indeed a useful transformation, viz.,

Theorem 19.1 (Uniqueness)**.** *If there exists* $s_0 > 0$ *such that* $M_X(s)$ *and* $M_Y(s)$ *are finite and equal for all* $s \in (-s_0, s_0)$*, then* $f_X = f_Y$.

Example 19.2. If

$$M(s) = \frac{1}{2}e^s + \frac{1}{4}e^{-\pi s} + \frac{1}{4}e^{es},$$

then M is an m.g.f. with

$$f(x) = \begin{cases} 1/2 & \text{if } x = 1, \\ 1/4 & \text{if } x = -\pi \text{ or } x = e, \\ 0 & \text{otherwise.} \end{cases}$$

2. Sums of independent random variables

Theorem 19.3. *If $X_1, \ldots, X_n$ are independent, with respective generating functions $G_{X_1}, \ldots, G_{X_n}$, then $\sum_{i=1}^n X_i$ has the p.g.f.,*

$$G(s) = G_{X_1}(s) \times \cdots \times G_{X_n}(s).$$

Proof. By induction, it suffices to do this for $n = 2$ (why?). But then

$$G_{X_1+X_2}(s) = E\left[s^{X_1+X_2}\right] = E\left[s^{X_1} \times s^{X_2}\right].$$

By independence, this is equal to the product of $E[s^{X_1}]$ and $E[s^{X_2}]$, which is the desired result. □

Example 19.4. Suppose $X = \text{bin}(n, p)$. Then we can write $X = I_1 + \cdots + I_n$, where $I_1, \ldots, I_n$ are independent, each taking the values zero (with probability $q = 1 - p$) and one (with probability p). Let us first compute

$$G_{I_j}(s) = E[s^{I_j}] = qs^0 + ps^1 = q + ps.$$

We can apply Theorem 19.3 then to find that

$$G_X(s) = (q + ps)^n.$$

Example 19.5. If $X = \text{bin}(n\,p)$ and $Y = \text{bin}(m, p)$ are independent, then by the previous example and Theorem 19.3,

$$G_{X+Y}(s) = (q + ps)^n (q + ps)^m = (q + ps)^{n+m}.$$

By the uniqueness theorem, $X + Y = \text{bin}(n + m, p)$. We found this out earlier by applying much harder methods. See Example 18.3.

Example 19.6. If $X = \text{Poisson}(\lambda)$, then

$$\begin{aligned} G(s) = E\left[s^X\right] &= \sum_{k=0}^{\infty} s^k e^{-\lambda} \frac{\lambda^k}{k!} \\ &= e^{-\lambda} \sum_{k=0}^{\infty} \frac{(s\lambda)^k}{k!}. \end{aligned}$$

The sum gives the Taylor expansion of $\exp(s\lambda)$. Therefore,

$$G(s) = \exp\left\{\lambda(s-1)\right\}.$$

Example 19.7. Now suppose $X = \text{Poisson}(\lambda)$ and $Y = \text{Poisson}(\gamma)$ are independent. We apply the previous example and Theorem 19.3, in conjunction, to find that

$$\begin{aligned} G_{X+Y}(s) &= \exp\{\lambda(s-1)\}\exp\{\gamma(s-1)\} \\ &= \exp\{(\lambda+\gamma)(s-1)\}. \end{aligned}$$

Thus, $X + Y = \text{Poisson}(\gamma + \lambda)$, thanks to the uniqueness theorem and Example 19.6. For a harder derivation of the same fact see Example 17.4.

Next is another property of generating function, applied to random sums.

Theorem 19.8. *Suppose $X_0, X_1, X_2, \ldots$ and N are all independent, and $N \geqslant 0$. Suppose also that all X_is have the same distribution, with common p.g.f. G. Then, the p.g.f. of $S = \sum_{i=0}^{N} X_i$ is*

$$G_Z(s) = G_N(G(s)).$$

Proof. We know that

$$\begin{aligned} G_Z(s) = E\left[s^Z\right] &= \sum_{n=0}^{\infty} E\left(s^Z \mid N = n\right) P\{N = n\} \\ &= P\{N = 0\} + \sum_{n=1}^{\infty} E\left(s^{X_1+\cdots+X_n}\right) P\{N = n\}, \end{aligned}$$

by the independence of $X_1, X_2, \ldots$ and N. Therefore,

$$\begin{aligned} G_Z(s) &= \sum_{n=0}^{\infty} (G(s))^n P\{N = n\} \\ &= E\left[(G(s))^N\right], \end{aligned}$$

which is the desired result. □

3. Example: Branching processes

Branching processes are mathematical models for population genetics. The simplest branching process models asexual reproduction of genes, for example. It goes as follows: At time $n = 0$ there is one gene of a given (fixed) type. At time $n = 1$, this gene splits into a random number of "offspring genes." All subsequent genes split in the same way in time. We assume that all genes behave independently from all other genes, but the offspring distribution is the same for all genes as well. So here is the math model: Let $X_{i,j}$ be independent random variables, all with the same distribution (mass function). Let $Z_0 = 1$ be the population size at time 0,

and define $Z_1 = X_{1,1}$. This is the population size at time $n = 1$. Then, $Z_2 = \sum_{j=1}^{Z_1} X_{2,j}$ be the population size in generation 2, and more generally,

$$Z_n = \sum_{j=1}^{Z_{n-1}} X_{n,j}.$$

The big question of branching processes, and one of the big questions in population genetics, is "what happens to Z_n as $n \to \infty$"?

Let G denote the common generating function of the $X_{i,j}$'s, and let G_n denote the generating function of Z_n. Because $Z_0 = 1$, $G_0(s) = s$. Furthermore,

$$G_1(s) = E\left[s^{X_{1,1}}\right] = G(s) = G_0(G(s)).$$

In general,

$$G_{n+1}(s) = E\left[s^{\sum_{j=1}^{Z_n} X_{n+1,j}}\right] = G_n\left(G(s)\right),$$

thanks to Theorem 19.8. Because this is true for all $n \geqslant 0$, we have $G_1(s) = G(s)$,, $G_2(s) = G(G(s))$, and more generally,

$$G_k(s) = \overbrace{G(G(\cdots G(s)\cdots))}^{k \text{ times}} \qquad \text{for all } k \geqslant 0.$$

Note that $\{Z_n = 0\}$ is the event that the population has gone extinct by the nth generation. These events are increasing, therefore rule 4 of probabilities tells us that

$$P\{\text{ultimate extinction}\} = \lim_{n\to\infty} P\{Z_n = 0\}.$$

Theorem 19.9 (A. N. Kolmogorov)**.** *The extinction probability above is equal to the smallest nonnegative solution* s *to the equation*

$$G(s) = s.$$

Lecture 20

Example 20.1. Suppose that the offspring mass function is given by

$$f(k) = \begin{cases} 1/4 & \text{if } k = 0, \\ 1/4 & \text{if } k = 1, \\ 1/2 & \text{if } k = 2. \end{cases}$$

Then, $G(s) = \frac{1}{4} + \frac{1}{4}s + \frac{1}{2}s^2$, and hence $G(s) = s$ is the same equation as

$$2s^2 - 3s + 1 = 0.$$

The solutions are

$$s = \frac{3 \pm \sqrt{9-8}}{4} = \frac{1}{2} \text{ and } 1.$$

Thus, the probability of ultimate extinction is $1/2$.

...examples of mgf's

Lecture 21

1. Continuous Random Variables

Definition 21.1. We say that X is a *continuous* random variable with *density function* f if f is a piecewise continuous nonnegative function, and for all real numbers x,

$$P\{X \leqslant x\} = \int_{-\infty}^{x} f(y)\,dy.$$

In this case,

$$F(x) = P\{X \leqslant x\} = \int_{-\infty}^{x} f(y)\,dy$$

defines the *distribution function* of X.

Some basic properties:

(1) We have $F(\infty) - F(-\infty) = \int_{-\infty}^{\infty} f(y)\,dy = 1$.

(2) Because f is integrable and nonnegative, for all real numbers x we have

$$F(x+h) - F(x) = \int_{x}^{x+h} f(y)\,dy \to 0 \qquad \text{as } h \searrow 0.$$

But the left-most term is $P\{x < X \leqslant x+h\}$. Therefore, by Rule 4 of probabilities,

$$P\{X = x\} = F(x) - F(x-) = 0 \qquad \text{for all } x.$$

(3) If f is continuous at x, then by the fundamental theorem of calculus,

$$F'(x) = f(x).$$

This shows that $F' = f$ at all but at most countably-many points.

For examples, we merely need to construct *any* f such that $f(x) \geqslant 0$ and $\int_{-\infty}^{x} f(y)\,dy = 1$, together with the property that f is continuous piecewise. Here are some standard examples.

Example 21.2 (Uniform density)**.** If $a < b$ are fixed, then the uniform density on (a, b) is the function

$$f(x) = \begin{cases} \dfrac{1}{b-a} & \text{if } a \leqslant x \leqslant b, \\ 0 & \text{otherwise.} \end{cases}$$

In this case, we can compute the distribution function as follows:

$$F(x) = \begin{cases} 0 & \text{if } x < a, \\ \dfrac{x}{b-a} & \text{if } a \leqslant x \leqslant b, \\ 1 & \text{if } x > b. \end{cases}$$

Example 21.3 (Exponential densities)**.** Let $\lambda > 0$ be fixed. Then

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x \geqslant 0, \\ 0 & \text{if } x < 0 \end{cases}$$

is a density, and is called the *exponential density with parameter* λ. It is not hard to see that

$$F(x) = \begin{cases} 1 - e^{-\lambda x} & \text{if } x \geqslant 0, \\ 0 & \text{if } x < 0. \end{cases}$$

Example 21.4 (The Cauchy density)**.** Define for all real numbers x,

$$f(x) = \frac{1}{\pi}\,\frac{1}{1+x^2}.$$

Because

$$\frac{d}{dx}\arctan x = \frac{1}{1+x^2},$$

we have

$$\int_{-\infty}^{\infty} f(y)\,dy = \frac{1}{\pi}\int_{-\infty}^{\infty} \frac{1}{1+y^2}\,dy = \frac{1}{\pi}\left[\arctan(\infty) - \arctan(-\infty)\right] = 1.$$

Also,

$$\begin{aligned} F(x) &= \frac{1}{\pi}\int_{-\infty}^{x} f(y)\,dy = \frac{1}{\pi}\left[\arctan(x) - \arctan(-\infty)\right] \\ &= \frac{1}{\pi}\arctan(x) + \frac{1}{2} \qquad \text{for all real } x. \end{aligned}$$

Lecture 22

1. Examples of continuous random variables

Example 22.1 (Standard normal density)**.** I claim that

$$\phi(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right)$$

defines a density function. Clearly, $\phi(x) \geqslant 0$ and is continuous at all points x. So it suffices to show that the area under ϕ is one. Define

$$A = \int_{-\infty}^{\infty} \phi(x)\, dx.$$

Then,

$$\begin{aligned} A^2 &= \frac{1}{2\pi} \int_{-\infty}^{\infty} \exp\left(-\frac{x^2+y^2}{2}\right) dx\, dy \\ &= \frac{1}{2\pi} \int_0^{2\pi} \int_0^{\infty} \exp\left(-\frac{r^2}{2}\right) r\, dr\, d\theta. \end{aligned}$$

Let $s = r^2/2$ to find that the inner integral is $\int_0^{\infty} \exp(-s)\, ds = 1$. Therefore, $A^2 = 1$ and hence $A = 1$, as desired. [Why is A not -1?]

The distribution function of ϕ is

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-z^2/2}\, dz.$$

One can prove that there is "no nice formula" that "describes" $\Phi(x)$ for all x (theorem of Liouville). Usually, people use tables of integrals to evaluate $\Phi(x)$ for concrete values of x.

Example 22.2 (Gamma densities)**.** Choose and fix two numbers (parameters) $\alpha, \lambda > 0$. The *gamma density* with parameters α and λ is the probability

density function that is proportional to

$$\begin{cases} x^{\alpha-1}e^{-\lambda x} & \text{if } x \geqslant 0, \\ 0 & \text{if } x < 0. \end{cases}$$

Now,

$$\int_0^\infty x^{\alpha-1}e^{-\lambda x}\,dx = \frac{1}{\lambda^\alpha}\int_0^\infty y^{\alpha-1}e^{-y}\,dy.$$

Define the *gamma function* as

$$\Gamma(\alpha) = \int_0^\infty y^{\alpha-1}e^{-y}\,dy \qquad \text{for all } \alpha > 0.$$

One can prove that there is "no nice formula" that "describes" $\Gamma(\alpha)$ for all α (theorem of Liouville). Thus, the best we can do is to say that the following is a Gamma density with parameters $\alpha, \lambda > 0$:

$$f(x) = \begin{cases} \dfrac{\lambda^\alpha}{\Gamma(\alpha)}x^{\alpha-1}e^{-\lambda x} & \text{if } x \geqslant 0, \\ 0 & \text{if } x < 0. \end{cases}$$

You can probably guess by now (and correctly!) that $F(x) = \int_{-\infty}^x f(y)\,dy$ cannot be described by nice functions either. Nonetheless, let us finish by making the observation that $\Gamma(\alpha)$ is computable for some reasonable values of $\alpha > 0$. The key to unraveling this remark is the following "reproducing property":

$$\Gamma(\alpha+1) = \alpha\Gamma(\alpha) \qquad \text{for all } \alpha > 0. \tag{18}$$

The proof uses integration by parts:

$$\begin{aligned} \Gamma(\alpha+1) &= \int_0^\infty x^\alpha e^{-x}\,dx \\ &= \int_0^\infty u(x)v'(x)\,dx, \end{aligned}$$

where $u(x) = x^\alpha$ and $v'(x) = e^{-x}$. Integration by parts states that[1]

$$\int uv' = uv - \int v'u \qquad \text{for indefinite integrals.}$$

[1]This follows immediately from integrating the product rule: $(uv)' = u'v + uv'$.

Evidently, $u'(x) = \alpha x^{\alpha-1}$ and $v(x) = -e^{-x}$. Hence,

$$\begin{aligned}\Gamma(\alpha+1) &= \int_0^\infty x^\alpha e^{-x}\,dx \\ &= uv\Big|_0^\infty - \int_0^\infty v'u \\ &= \left(-\alpha x^{\alpha-1}e^{-x}\right)\Big|_0^\infty + \alpha\int_0^\infty x^{\alpha-1}e^{-x}\,dx.\end{aligned}$$

The first term is zero, and the second (the integral) is $\alpha\Gamma(\alpha)$, as claimed. Now, it easy to see that $\Gamma(1) = \int_0^\infty e^{-x}\,dx = 1$. Therefore, $\Gamma(2) = 1\times\Gamma(1) = 1$, $\Gamma(3) = 2\times\Gamma(2) = 2$, ..., and in general,

$$\Gamma(n+1) = n! \qquad \text{for all integers } n \geqslant 0.$$

2. Functions of a continuous random variable

The basic problem: If $Y = g(X)$, then how can we compute f_Y in terms of f_X?

Example 22.3. Suppose X is uniform on $(0,1)$, and $Y = -\ln X$. Then, we compute f_Y by first computing F_Y, and then using $f_Y = F_Y'$. Here are the details:

$$F_Y(a) = P\{Y \leqslant a\} = P\{-\ln X \leqslant a\}.$$

Now, $-\ln(x)$ is a decreasing function. Therefore, $-\ln(x) \leqslant a$ if and only if $x \geqslant e^{-a}$, and hence,

$$F_Y(a) = P\left\{X \geqslant e^{-a}\right\} = 1 - F_X(e^{-a}).$$

Consequently,

$$f_Y(a) = -f_X(e^{-a})\frac{d}{da}(e^{-a}) = e^{-a}f_X(e^{-a}).$$

Now recall that $f_X(u) = 1$ if $0 \leqslant u \leqslant 1$ and $f_X(u) = 0$ otherwise. Now e^{-a} is between zero and one if and only if $a \geqslant 0$. Therefore,

$$f_X(e^{-a}) = \begin{cases} 1 & \text{if } a \geqslant 0,\\ 0 & \text{if } a < 0.\end{cases}$$

It follows then that

$$f_Y(a) = \begin{cases} e^{-a} & \text{if } a \geqslant 0,\\ 0 & \text{otherwise.}\end{cases}$$

Thus, $-\ln X$ has an exponential density with parameter $\lambda = 1$. More generally, if $\lambda > 0$ is fixed, then $-(1/\lambda)\ln X$ has an exponential density with parameter λ.

Lecture 23

1. Functions of a continuous random variable, continued

The problem: $Y = g(X)$; find f_Y in terms of f_X.

The solution: First compute F_Y, by hand, in terms of F_X, and then use the fact that $F'_Y = f_Y$ and $F'_X = f_X$.

Example 23.1. Suppose X has density f_X. Then let us find the density function of $Y = X^2$. Again, we seek to first compute F_Y. Now, for all $a > 0$,

$$F_Y(a) = P\{X^2 \leqslant a\} = P\left\{-\sqrt{a} \leqslant X \leqslant \sqrt{a}\right\} = F_X\left(\sqrt{a}\right) - F_X\left(-\sqrt{a}\right).$$

Differentiate $[d/da]$ to find that

$$f_Y(a) = \frac{f_X\left(\sqrt{a}\right) + f_X\left(-\sqrt{a}\right)}{2\sqrt{a}}$$

On the other hand, $f_Y(a) = 0$ if $a \leqslant 0$. For example, consider the case that X is standard normal. Then,

$$f_{X^2}(a) = \begin{cases} \dfrac{e^{-a}}{\sqrt{2\pi a}} & \text{if } a > 0, \\ 0 & \text{if } a \leqslant 0. \end{cases}$$

Or if X is Cauchy, then

$$f_{X^2}(a) = \begin{cases} \dfrac{1}{\pi\sqrt{a}(1+a)} & \text{if } a > 0, \\ 0 & \text{if } a \leqslant 0. \end{cases}$$

Or if X is uniform $(0,1)$, then

$$f_{X^2}(a) = \begin{cases} \frac{1}{2\sqrt{a}} & \text{if } 0 < a < 1, \\ 0 & \text{otherwise.} \end{cases}$$

Example 23.2. Suppose $\mu \in \mathbf{R}$ and $\sigma > 0$ are fixed constants, and define $Y = \mu + \sigma X$. Find the density of Y in terms of that of X. Once again,

$$F_Y(a) = P\{\mu + \sigma X \leqslant a\} = P\left\{X \leqslant \frac{a-\mu}{\sigma}\right\} = F_X\left(\frac{a-\mu}{\sigma}\right).$$

Therefore,

$$f_Y(a) = \frac{1}{\sigma} f_X\left(\frac{a-\mu}{\sigma}\right).$$

For example, if X is standard normal, then

$$f_{\mu+\sigma X}(a) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right).$$

This is the socalled $N(\mu, \sigma^2)$ density.

Example 23.3. Suppose X is uniformly distributed on $(0,1)$, and define

$$Y = \begin{cases} 0 & \text{if } 0 \leqslant X < \frac{1}{3}, \\ 1 & \text{if } \frac{1}{3} \leqslant X < \frac{2}{3}, \\ 2 & \text{if } \frac{2}{3} \leqslant X < 1. \end{cases}$$

Then, Y is a discrete random variable with mass function,

$$f_Y(x) = \begin{cases} \frac{1}{3} & \text{if } x = 0, 1, \text{ or } 2, \\ 0 & \text{otherwise.} \end{cases}$$

For instance, in order to compute $f_Y(1)$ we note that

$$f_Y(1) = P\left\{\frac{1}{3} \leqslant X < \frac{2}{3}\right\} = \int_{1/3}^{2/3} \underbrace{f_X(y)}_{\equiv 1}\, dy = \frac{1}{3}.$$

Example 23.4. Another common transformation is $g(x) = |x|$. In this case, let $Y = |X|$ and note that if $a > 0$, then

$$F_Y(a) = P\{-a < X < a\} = F_X(a) - F_X(-a).$$

Else, $F_Y(a) = 0$. Therefore,

$$f_Y(a) = \begin{cases} f_X(a) + f_X(-a) & \text{if } a > 0, \\ 0 & \text{if } a \leqslant 0. \end{cases}$$

For instance, if X is standard normal, then

$$f_{|X|}(a) = \begin{cases} \frac{2}{\pi}\, e^{-a^2/2} & \text{if } a > 0, \\ 0 & \text{if } a \leqslant 0. \end{cases}$$

Or if X is Cauchy, then

$$f_{|X|}(a) = \begin{cases} \frac{2}{\pi}\, \frac{1}{1+a^2} & \text{if } a > 0, \\ 0 & \text{otherwise.} \end{cases}$$

Can you guess $f_{|X|}$ when X is uniform $(-1\,,1)$?

Lecture 24

1. Functions of a random variable, continued

Example 24.1. It is best to try to work on these problems on a case-by-case basis. Here is an example where you need to do that. Consider X to be a uniform $(0,1)$ random variable, and define $Y = \sin(\pi X/2)$. Because $X \in (0,1)$, it follows that $Y \in (0,1)$ as well. Therefore, $F_Y(a) = 0$ if $a < 0$, and $F_Y(1) = 1$ if $a > 1$. If $0 \leqslant a \leqslant 1$, then

$$F_Y(a) = P\left\{\sin\left(\frac{\pi X}{2}\right) \leqslant a\right\} = P\left\{X \leqslant \frac{2}{\pi}\arcsin a\right\} = \frac{2}{\pi}\arcsin a.$$

You need to carefully plot the arcsin curve to deduce this. Therefore,

$$f_Y(a) = \begin{cases} \dfrac{2}{\pi\sqrt{1-a^2}} & \text{if } 0 < a < 1, \\ 0 & \text{otherwise.} \end{cases}$$

Finally, a transformation of a continuous random variable into a discrete one ....

Example 24.2. Suppose X is uniform $(0,1)$ and define $Y = \lfloor 2X \rfloor$ to be the largest integer $\leqslant 2X$. Find f_Y.

First of all, we note that Y is discrete. Its possible values are 0 (this is when $0 < X < 1/2$) and 1 (this is when $1/2 < X < 1$). Therefore,

$$f_Y(0) = P\left\{0 < X < \frac{1}{2}\right\} = \int_0^{1/2} dy = \frac{1}{2} = 1 - f_Y(1) = \frac{1}{2}.$$

This is thrown in just so we remember that it is entirely possible to start out with a continuous random variable, and then transform it into a discrete one.

2. Expectation

If X is a continuous random variable with density f, then its *expectation* is defined to be

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

provided that either $X \geqslant 0$, or $\int_{-\infty}^{\infty} |x| f(x)\,dx < \infty$.

Example 24.3 (Uniform)**.** Suppose X is uniform $(a\,,b)$. Then,

$$E(X) = \int_a^b x \frac{1}{b-a}\,dx = \frac{1}{2}\frac{b^2-a^2}{b-a}.$$

It is easy to check that $b^2 - a^2 = (b-a)(b+a)$, whence

$$E(X) = \frac{b+a}{2}.$$

N.B.: The formula of the first example on page 303 of your text is wrong.

Example 24.4 (Gamma)**.** If X is Gamma$(\alpha\,,\lambda)$, then for all positive values of x we have $f(x) = \lambda^\alpha/\Gamma(\alpha)x^{\alpha-1}e^{-\lambda x}$, and $f(x) = 0$ for $x < 0$. Therefore,

$$\begin{aligned} E(X) &= \frac{\lambda^\alpha}{\Gamma(\alpha)} \int_0^\infty x^\alpha e^{-\lambda x}\,dx \\ &= \frac{1}{\lambda\Gamma(\alpha)} \int_0^\infty z^\alpha e^{-z}\,dz \qquad (z = \lambda x) \\ &= \frac{\Gamma(\alpha+1)}{\lambda\Gamma(\alpha)} \\ &= \frac{\alpha}{\lambda}. \end{aligned}$$

In the special case that $\alpha = 1$, this is the expectation of an exponential random variable with parameter λ.

Example 24.5 (Normal)**.** Suppose $X = N(\mu\,,\sigma^2)$. That is,

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

Then,

$$\begin{aligned}
\mathrm{E}(X) &= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} x \exp\left(-\frac{(x-\mu)^2}{2}\right) \, dx \\
&= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} (\mu + \sigma z) e^{-z^2/2} \, dz \qquad (z = (x-\mu)/\sigma) \\
&= \mu \underbrace{\int_{-\infty}^{\infty} \frac{e^{-z^2/2}}{\sqrt{2\pi}} \, dz}_{1} + \frac{\sigma}{\sqrt{2\pi}} \underbrace{\int_{-\infty}^{\infty} z e^{-z^2/2} \, dz}_{0, \text{ by symmetry}} \\
&= \mu.
\end{aligned}$$

Example 24.6 (Cauchy)**.** In this example, $f(x) = \pi^{-1}(1+x^2)^{-1}$. Note that the expectation is defined only if the following limit exists regardless of how we let n and m tend to ∞:

$$\int_{-m}^{n} \frac{y}{1+y^2} \, dy.$$

Now I argue that the limit does not exist; I do so by showing two different choices of (n, m) which give rise to different limiting "integrals."

First suppose $m = n$, so that by symmetry,

$$\int_{-n}^{n} \frac{y}{1+y^2} \, dy = 0.$$

Let $n \to \infty$ to obtain zero as the limit of the left-hand side.

Next, suppose $m = 2n$. Again by symmetry,

$$\begin{aligned}
\int_{-2n}^{n} \frac{y}{1+y^2} \, dy &= \int_{-2n}^{-n} \frac{y}{1+y^2} \, dy \\
&= -\int_{n}^{2n} \frac{y}{1+y^2} \, dy \\
&= -\frac{1}{2} \int_{1+n^2}^{1+4n^2} \frac{dz}{z} \qquad (z = 1+y^2) \\
&= -\frac{1}{2} \ln\left(\frac{1+4n^2}{1+n^2}\right) \\
&\to -\frac{1}{2} \ln 4 \quad \text{as } n \to \infty.
\end{aligned}$$

Therefore, the Cauchy density does not have a well-defined expectation. [That is not to say that the expectation is well defined, but infinite.]

Lecture 25

1. Expectations, continued

Theorem 25.1. *If X is a positive random variable with density f, then*

$$E(X) = \int_0^\infty P\{X > x\}\,dx = \int_0^\infty (1 - F(x))\,dx.$$

Proof. The second identity is a consequence of the fact that $1 - F(x) = P\{X > x\}$. In order to prove the first identity note that $P\{X > x\} = \int_x^\infty f(y)\,dy$. Therefore,

$$\begin{aligned}\int_0^\infty P\{X > x\}\,dx &= \int_0^\infty \int_x^\infty f(y)\,dy\,dx \\ &= \int_0^\infty f(y) \int_0^y dx\,dy \\ &= \int_0^\infty y f(y)\,dy,\end{aligned}$$

and this is $E(X)$. □

Question: Why do we need X to be positive? [To find the answer you need to think hard about the change of variables formula of calculus.]

Theorem 25.2. *If $\int_{-\infty}^\infty |g(a)|f(a)|\,da < \infty$, then*

$$E[g(X)] = \int_{-\infty}^\infty g(a)f(a)\,da.$$

Proof. I will prove the result in the special case that $g(x) \geqslant 0$, but will not assume that $\int_{-\infty}^\infty g(a)f(a)\,da < \infty$.

The preceding theorem implies that

$$E[g(X)] = \int_0^\infty P\{g(X) > x\}\,dx.$$

But $P\{g(X) > x\} = P\{X \in A\}$ where $A = \{y :\ g(y) > x\}$. Therefore,

$$\begin{aligned} E[g(X)] &= \int_0^\infty \int_{\{y:\ g(y)>x\}} f(y)\,dy\,dx \\ &= \int_0^\infty \int_0^{g(y)} f(y)\,dx\,dy \\ &= \int_0^\infty g(y)f(y)\,dy, \end{aligned}$$

as needed. □

Properties of expectations:

(1) If $g(X)$ and $h(X)$ have finite expectations, then

$$E\left[g(X) + h(X)\right] = E[g(X)] + E[h(X)].$$

(2) If $P\{a \leqslant X \leqslant b\} = 1$ then $a \leqslant EX \leqslant b$.

(3) Markov's inequality: If $h(x) \geqslant 0$, then

$$P\left\{h(X) \geqslant a\right\} \leqslant \frac{E[h(X)]}{a} \qquad \text{for all } a > 0.$$

(4) Cauchy–Schwarz inequality:

$$E[X^2] \geqslant \{E(|X|)\}^2.$$

In particular, if $E[X^2] < \infty$, then $E(|X|)$ and EX are both finite.

Definition 25.3. The *variance* of X is defined as

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

Alternative formula:

$$\mathrm{Var}(X) = E\left[(X - EX)^2\right].$$

Example 25.4 (Moments of Uniform$(0,1)$)**.** If X is uniform$(0,1)$, then for all integers $n \geqslant 1$,

$$E(X^n) = \int_0^1 x^n\,dx = \frac{1}{n+1}.$$

Example 25.5 (Moments of $N(0,1)$)**.** Compute $E(X^n)$, where $X = N(0,1)$ and $n \geqslant 1$ is an integer:

$$\begin{aligned} E(X^n) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} a^n e^{-a^2/2} \, da \\ &= 0 \quad \text{if } n \text{ is odd, by symmetry.} \end{aligned}$$

If n is even, then

$$\begin{aligned} E(X^n) &= \frac{2}{\sqrt{2\pi}} \int_0^{\infty} a^n e^{-a^2/2} \, da = \sqrt{\frac{2}{\pi}} \int_0^{\infty} a^n e^{-a^2/2} \, da \\ &= \sqrt{\frac{2}{\pi}} \int_0^{\infty} (2z)^{n/2} e^{-z} \underbrace{\left((2z)^{-1/2} \, dz \right)}_{da} \qquad \left(z = a^2/2 \;\Leftrightarrow\; a = \sqrt{2z} \right) \\ &= \frac{2^{n/2}}{\sqrt{\pi}} \int_0^{\infty} z^{(n-1)/2} e^{-z} \, dz \\ &= \frac{2^{n/2}}{\sqrt{\pi}} \Gamma\left(\frac{n+1}{2} \right). \end{aligned}$$

Lecture 26

1. Moment generating functions

Let X be a continuous random variable with density f. Its *moment generating function* is defined as

$$M(t) = E[e^{tX}] = \int_{-\infty}^{\infty} e^{tx} f(x)\,dx, \tag{19}$$

provided that the integral exists.

Example 26.1 (Uniform$(0,1)$)**.** If $X = \text{Uniform}(0,1)$, then

$$M(t) = E[e^{tX}] = \int_0^1 e^{tx}\,dx = \frac{e^t - 1}{t}.$$

Example 26.2 (Gamma)**.** If $X = \text{Gamma}(\alpha,\lambda)$, then

$$\begin{aligned} M(t) &= \int_0^\infty e^{tx} \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x}\,dx \\ &= \frac{\lambda^\alpha}{\Gamma(\alpha)} \int_0^\infty x^{\alpha-1} e^{-(\lambda-t)x}\,dx. \end{aligned}$$

If $t \geqslant \lambda$, then the integral is infinite. On the other hand, if $t < \lambda$, then

$$\begin{aligned} M(t) &= \frac{\lambda^\alpha}{\Gamma(\alpha)} \int_0^\infty \frac{z^{\alpha-1}}{(\lambda-t)^{\alpha-1}} e^{-z} \frac{dz}{\lambda - t} \qquad (z = (\lambda - t)x) \\ &= \frac{\lambda^\alpha}{\Gamma(\alpha) \times (\lambda-t)^\alpha} \underbrace{\int_0^\infty z^{\alpha-1} e^{-z}\,dz}_{\Gamma(\alpha)} \\ &= \frac{\lambda^\alpha}{(\lambda-t)^\alpha}. \end{aligned}$$

Thus,

$$M(t) = \begin{cases} \left(\dfrac{\lambda}{\lambda - t}\right)^{\alpha} & \text{if } t < \lambda, \\ \infty & \text{otherwise.} \end{cases}$$

Example 26.3 ($N(0,1)$)**.** If $X = N(0,1)$, then

$$\begin{aligned} M(t) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{tx} e^{-x^2/2}\, dx \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\left(-\frac{x^2 - 2tx}{2}\right) dx \\ &= \frac{e^{t^2/2}}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\left(-\frac{x^2 - 2tx + t^2}{2}\right) dx \\ &= \frac{e^{t^2/2}}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\left(-\frac{(x-t)^2}{2}\right) dx \\ &= \frac{e^{t^2/2}}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-u^2/2}\, du \qquad (u = x - t) \\ &= e^{t^2/2}. \end{aligned}$$

2. Relation to moments

Suppose we know the function $M(t) = E \exp(tX)$. Then, we can compute the moments of X from the function M by successive differentiation. For instance, suppose X is a continuous random variable with moment generating function M and density function f, and note that

$$M'(t) = \frac{d}{dt}\left(E[e^{tX}]\right) = \frac{d}{dt} \int_{-\infty}^{\infty} e^{tx} f(x)\, dx.$$

Now, if the integral converges absolutely, then a general fact states that we can take the derivative under the integral sign. That is,

$$M'(t) = \int_{-\infty}^{\infty} x e^{tx} f(x)\, dx = E\left[Xe^{tX}\right].$$

The same end-result holds if X is discrete with mass function f, but this time,

$$M'(t) = \sum_{x} x e^{tx} f(x) = E\left[Xe^{tX}\right].$$

Therefore, in any event:

$$M'(0) = E[X].$$

In general, this procedure yields,

$$M^{(n)}(t) = E\left[X^n e^{tX}\right].$$

Therefore,

$$M^{(n)}(0) = E\left[X^n\right].$$

Example 26.4 (Uniform)**.** We saw earlier that if X is distributed uniformly on $(0,1)$, then for all real numbers t,

$$M(t) = \frac{e^t - 1}{t}.$$

Therefore,

$$M'(t) = \frac{te^t - e^t + 1}{t^2}, \quad M''(t) = \frac{t^2e^t - 2te^t + 2e^t - 2}{t^3},$$

whence

$$EX = M'(0) = \lim_{t \searrow 0} \frac{te^t - e^t + 1}{t^2} = \lim_{t \searrow 0} \frac{te^t}{2t} = \frac{1}{2},$$

by l'Hopital's rule. Similarly,

$$E\left[X^2\right] = \lim_{t \searrow 0} \frac{t^2e^t - 2te^t + 2e^t - 2}{t^3} = \lim_{t \searrow 0} \frac{t^2e^t}{3t^2} = \frac{1}{3}.$$

Alternatively, these can be checked by direct computation, using the fact that $E[X^n] = \int_0^1 x^n \, dx = 1/(n+1)$.

A discussion of physical CLT machines . . . .

Lecture 27

1. Two important properties

Theorem 27.1 (Uniqueness)**.** *If* X *and* Y *are two random variables—discrete* or *continuous—with moment generating functions* M_X *and* M_Y*, and if there exists* $\delta > 0$ *such that* $M_X(t) = M_Y(t)$ *for all* $t \in (-\delta, \delta)$*, then* $M_X = M_Y$ *and* X *and* Y *have the same distribution. More precisely:*

(1) X *is discrete if and only if* Y *is, in which case their mass functions are the same;*

(2) X *is continuous if and only if* Y *is, in which case their density functions are the same.*

Theorem 27.2 (Lévy's continuity theorem)**.** *Let* X_n *be a random variables—discrete* or *continuous—with moment generating functions* M_n*. Also, let* X *be a random variable with moment generating function* M*. Suppose there exists* $\delta > 0$ *such that:*

(1) *If* $-\delta < t < \delta$*, then* $M_n(t), M(t) < \infty$ *for all* $n \geqslant 1$*; and*

(2) $\lim_{n\to\infty} M_n(t) = M(t)$ *for all* $t \in (-\delta, \delta)$*, then*

$$\lim_{n\to\infty} F_{X_n}(a) = \lim_{n\to\infty} P\{X_n \leqslant a\} = P\{X \leqslant a\} = F_X(a),$$

for all numbers a *at which* F_X *is continuous.*

Example 27.3 (Law of rare events)**.** Suppose $X_n = \text{binomal}(n, \lambda/n)$, where $\lambda > 0$ is fixed, and $n \geqslant \lambda$. Then, recall that

$$M_{X_n}(t) = \left(q + pe^{-t}\right)^n = \left(1 - \frac{\lambda}{n} + \frac{\lambda e^{-t}}{n}\right)^n \to \exp\left(-\lambda + \lambda e^{-t}\right).$$

Note that the right-most term is $M_X(t)$, where $X = \text{Poisson}(\lambda)$. Therefore, by Lévy's continuity theorem,

$$\lim_{n\to\infty} P\{X_n \leqslant a\} = P\{X \leqslant a\}, \tag{20}$$

at all a where F_X is continuous. But X is discrete and integer-valued. Therefore, F_X is continuous at a if and only if a is not a nonnegative integer. If a *is* a nonnegative integer, then we can choose a non-integer $b \in (a, a+1)$ to find that

$$\lim_{n\to\infty} P\{X_n \leqslant b\} = P\{X \leqslant b\}.$$

Because X_n and X are both non-negative integers, $X_n \leqslant b$ if and only if $X_n \leqslant a$, and $X \leqslant b$ if and only if $X \leqslant a$. Therefore, (20) holds for all a.

Example 27.4 (The de Moivre–Laplace central limit theorem)**.** Suppose $X_n = \text{binomial}(n, p)$, where $p \in (0,1)$ is fixed, and define Y_n to be its *standardization*. That is, $Y_n = (X_n - EX_n)/\sqrt{\text{Var}X_n}$. Alternatively,

$$Y_n = \frac{X_n - np}{\sqrt{npq}}.$$

We know that for all real numbers t,

$$M_{X_n}(t) = \left(q + pe^{-t}\right)^n.$$

We can use this to compute M_{Y_n} as follows:

$$M_{Y_n}(t) = E\left[\exp\left(t \cdot \frac{X_n - np}{\sqrt{npq}}\right)\right].$$

Recall that $X_n = I_1 + \cdots + I_n$, where I_j is one if the jth trial succeeds; else, $I_j = 0$. Then, $I_1, \ldots, I_n$ are independent binomial$(1, p)$'s, and $X_n - np = \sum_{j=1}^n (I_j - p)$. Therefore,

$$\begin{aligned}
E\left[\exp\left(t \cdot \frac{X_n - np}{\sqrt{npq}}\right)\right] &= E\left[\frac{t}{\sqrt{npq}} \sum_{j=1}^n (I_j - p)\right] \\
&= \left(E\left[\exp\left(\frac{t}{\sqrt{npq}}(I_1 - p)\right)\right]\right)^n \\
&= \left(p \exp\left\{\frac{t}{\sqrt{npq}}(1-p)\right\} + q \exp\left\{\frac{t}{\sqrt{npq}}(0-p)\right\}\right)^n \\
&= \left(p \exp\left\{t\sqrt{\frac{q}{np}}\right\} + q \exp\left\{-t\sqrt{\frac{p}{nq}}\right\}\right)^n.
\end{aligned}$$

According to the Taylor–MacLaurin expansion,

$$\exp\left\{t\sqrt{\frac{q}{np}}\right\} = 1 + t\sqrt{\frac{q}{np}} + \frac{t^2 q}{2np} + \text{smaller terms},$$
$$\exp\left\{-t\sqrt{\frac{p}{nq}}\right\} = 1 - t\sqrt{\frac{p}{nq}} + \frac{t^2 p}{2nq} + \text{smaller terms}.$$

Therefore,

$$\begin{aligned}
&p\exp\left\{t\sqrt{\frac{q}{np}}\right\} + q\exp\left\{-t\sqrt{\frac{p}{nq}}\right\} \\
&= p\left(1 + t\sqrt{\frac{q}{np}} + \frac{t^2 q}{2np} + \cdots\right) + q\left(1 - t\sqrt{\frac{p}{nq}} + \frac{t^2 p}{2nq} + \cdots\right) \\
&= p + t\sqrt{\frac{pq}{n}} + \frac{t^2 q}{2n} + \cdots + q - t\sqrt{\frac{pq}{n}} + \frac{t^2 p}{2n} + \cdots \\
&= 1 + \frac{t^2}{2n} + \text{smaller terms}.
\end{aligned}$$

Consequently,

$$M_{Y_n}(t) = \left(1 + \frac{t^2}{2n} + \text{smaller terms}\right)^n \to \exp\left(-\frac{t^2}{2}\right).$$

We recognize the right-hand side as $M_Y(t)$, where $Y = N(0,1)$. Because F_Y is continuous, this prove the *central limit theorem* of de Moivre: For all real numbers a,

$$\lim_{n\to\infty} P\{Y_n \leqslant a\} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{a} e^{-x^2/2}\,dx.$$

2. Jointly distributed continuous random variables

Definition 27.5. We say that (X, Y) is jointly distributed with *joint density function* f if f is piecewise continuous, and for all "nice" two-dimensional sets A,

$$P\{(X, Y) \in A\} = \iint_A f(x, y)\,dx\,dy.$$

If (X, Y) has a joint density function f, then:

(1) $f(x, y) \geqslant 0$ for all x and y;
(2) $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x, y)\,dx\,dy = 1$.

Any function f of two variables that satisfies these properties will do.

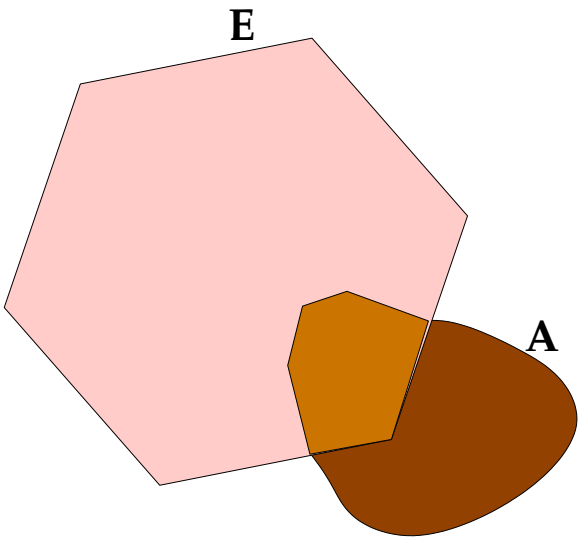

Figure 1. Region of integration in Example 27.6

Example 27.6 (Uniform joint density)**.** Suppose A is a subset of the plane that has a well-defined finite area $|A| > 0$. Define

$$f(x,y) = \begin{cases} \dfrac{1}{|A|} & \text{if } (x,y) \in A, \\ 0 & \text{otherwise.} \end{cases}$$

Then, f is a joint density function, and the corresponding random vector (X,Y) is said to be distributed *uniformly* on A. Moreover, for all planar sets E with well-defined areas,

$$P\{(X,Y) \in E\} = \iint\limits_{E\cap A} \frac{1}{|A|}\,dx\,dy = \frac{|E \cap A|}{|A|}.$$

See Figure 1.

Example 27.7. Suppose (X,Y) has joint density

$$f(x,y) = \begin{cases} Cxy & \text{if } 0 < y < x < 1, \\ 0 & \text{otherwise.} \end{cases}$$

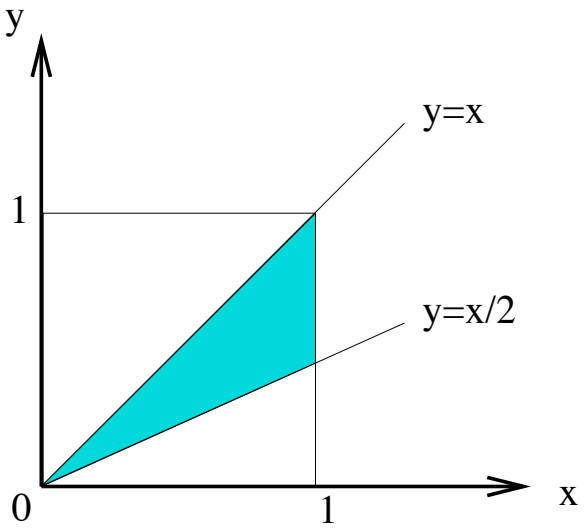

Figure 2. Region of integration in Example 27.7

Let us first find C, and then $P\{X \leqslant 2Y\}$. To find C:

$$1 = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)\,dx\,dy = \int_0^1\int_0^x Cxy\,dy\,dx$$
$$= C\int_0^1 x\underbrace{\left(\int_0^x y\,dy\right)}_{\frac{1}{2}x^2}dx = \frac{C}{2}\int_0^1 x^3\,dx = \frac{C}{8}.$$

Therefore, $C = 8$, and hence

$$f(x,y) = \begin{cases} 8xy & \text{if } 0 < y < x < 1, \\ 0 & \text{otherwise.} \end{cases}$$

Now

$$P\{X \leqslant 2Y\} = P\{(X,Y) \in A\} = \iint_A f(x,y)\,dx\,dy,$$

where A denotes the collection of all points (a,b) in the plane such that $a \leqslant 2b$. Therefore,

$$P\{X \leqslant 2Y\} = \int_0^1\int_{x/2}^x 8xy\,dy\,dx = \frac{3}{32}.$$

See Figure 2.

Lecture 28

Lecture 29

1. Marginals, distribution functions, etc.

If (X, Y) has joint density f, then

$$F_X(a) = P\{X \leqslant a\} = P\{(X, Y) \in A\},$$

where $A = \{(x\,y) : x \leqslant a\}$. Thus,

$$F_X(a) = \int_{-\infty}^{a} \left(\int_{-\infty}^{\infty} f(x\,,y)\,dy \right) dx.$$

Differentiate, and apply the fundamental theorem of calculus, to find that

$$f_X(a) = \int_{-\infty}^{\infty} f(a\,,y)\,dy.$$

Similarly,

$$f_Y(b) = \int_{-\infty}^{\infty} f(x\,,b)\,dx.$$

Example 29.1. Let

$$f(x\,,y) = \begin{cases} 8xy & \text{if } 0 < y < x < 1, \\ 0 & \text{otherwise.} \end{cases}$$

Then,

$$\begin{aligned} f_X(a) &= \begin{cases} \int_0^a 8ay\,dy & \text{if } 0 < a < 1, \\ 0 & \text{otherwise.} \end{cases} \\ &= \begin{cases} 4a^3 & \text{if } 0 < a < 1, \\ 0 & \text{otherwise.} \end{cases} \end{aligned}$$

[Note the typo in the text, page 341.] Similarly,

$$\begin{aligned} f_Y(b) &= \begin{cases} \int_b^1 8xb\,dx & \text{if } 0 < b < 1, \\ 0 & \text{otherwise.} \end{cases} \\ &= \begin{cases} 4b(1-b^2) & \text{if } 0 < b < 1, \\ 0 & \text{otherwise.} \end{cases} \end{aligned}$$

Example 29.2. Suppose (X, Y) is distributed uniformly on the square that joins the origin to the points $(1\,,0)$, $(1\,,1)$, and $(0\,,1)$. Then,

$$f(x\,,y) = \begin{cases} 1 & \text{if } 0 < x < 1 \text{ and } 0 < y < 1, \\ 0 & \text{otherwise.} \end{cases}$$

It follows that X and Y are both distributed uniformly on $(0\,,1)$.

Example 29.3. Suppose (X, Y) is distributed uniformly in the circle of radius one about $(0\,,0)$. That is,

$$f(x\,,y) = \begin{cases} \dfrac{1}{\pi} & \text{if } x^2 + y^2 \leqslant 1, \\ 0 & \text{otherwise.} \end{cases}$$

Then,

$$\begin{aligned} f_X(a) &= \begin{cases} \displaystyle\int_{-\sqrt{1-a^2}}^{\sqrt{1-a^2}} \frac{1}{\pi} dy & \text{if } -1 < a < 1, \\ 0 & \text{otherwise.} \end{cases} \\ &= \begin{cases} \dfrac{2}{\pi}\sqrt{1-a^2} & \text{if } -1 < a < 1, \\ 0 & \text{otherwise.} \end{cases} \end{aligned}$$

N.B.: f_Y is the same function. Therefore, in particular,

$$\begin{aligned} EX &= EY \\ &= \frac{2}{\pi}\int_{-1}^{1} a\sqrt{1-a^2}\,da \\ &= 0, \qquad \text{by symmetry.} \end{aligned}$$

2. Functions of a random vector

Basic problem: If (X, Y) has joint density f, then what, if any, is the joint density of (U, V), where $U = u(X, Y)$ and $V = v(X, Y)$? Or equivalently, $(U, V) = T(X, Y)$, where

$$T(x, y) = \begin{pmatrix} u(x, y) \\ v(x, y) \end{pmatrix}.$$

Example 29.4. Let (X, Y) be distributed uniformly in the circle of radius $R > 0$ about the origin in the plane. Thus,

$$f_{X,Y}(x, y) = \begin{cases} \dfrac{1}{\pi R^2} & \text{if } x^2 + y^2 \leqslant R^2, \\ 0 & \text{otherwise.} \end{cases}$$

We wish to write (X, Y), in polar coordinates, as (R, Θ), where

$$R = \sqrt{X^2 + Y^2} \quad \text{and} \quad \Theta = \arctan(Y/X).$$

Then, we compute first the *joint distribution function* $F_{R,\Theta}$ of (R, Θ) as follows:

$$\begin{aligned} F_{R,\Theta}(a, b) &= P\{R \leqslant a, \Theta \leqslant b\} \\ &= P\{(X, Y) \in A\}, \end{aligned}$$

where A is the "partial cone" $\{(x, y) : x^2 + y^2 \leqslant a^2, \arctan(y/x) \leqslant b\}$. If a is not between 0 and R, or $b \notin (-\pi, \pi)$, then $F_{R,\Theta}(a, b) = 0$. Else,

$$\begin{aligned} F_{R,\Theta}(a, b) &= \iint_A f_{X,Y}(x, y)\, dx\, dy \\ &= \int_0^b \int_0^a \frac{1}{\pi R^2} r\, dr\, d\theta, \end{aligned}$$

after the change of variables $r = \sqrt{x^2 + y^2}$ and $\theta = \arctan(y/x)$. Therefore, for all $a \in (0, R)$ and $b \in (-\pi, \pi)$,

$$F_{R,\Theta}(a, b) = \begin{cases} \dfrac{a^2 b}{2\pi R^2} & \text{if } 0 < a < R \text{ and } -\pi < b < \pi, \\ 0 & \text{otherwise.} \end{cases}$$

It is easy to see that

$$f_{R,\Theta}(a, b) = \frac{\partial^2 F_{R,\Theta}}{\partial a \partial b}(a, b).$$

Therefore,

$$f_{R,\Theta}(a, b) = \begin{cases} \dfrac{a}{\pi R^2} & \text{if } 0 < a < R \text{ and } -\pi < b < \pi, \\ 0 & \text{otherwise.} \end{cases}$$

The previous example can be generalized.

Suppose T is invertible with inverse function

$$T^{-1}(u,v) = \begin{pmatrix} x(u,v) \\ y(u,v) \end{pmatrix}.$$

The *Jacobian* of this transformation is

$$J(u,v) = \frac{\partial x}{\partial u}\frac{\partial y}{\partial v} - \frac{\partial x}{\partial v}\frac{\partial y}{\partial u}.$$

Theorem 29.5. *If T is "nice," then*

$$f_{U,V}(u,v) = f_{X,Y}(x(u,v), y(u,v))|J(u,v)|.$$

Example 29.6. In the polar coordinates example($r = u$, $\theta = v$),

$$\begin{aligned} r(x,y) &= \sqrt{x^2+y^2}, \\ \theta(x,y) &= \arctan(y/x) = \theta, \\ x(r,\theta) &= r\cos\theta, \\ y(r,\theta) &= r\sin\theta. \end{aligned}$$

Therefore, for all $r > 0$ and $\theta \in (-\pi, \pi)$,

$$\begin{aligned} J(r,\theta) &= (\cos(\theta) \times r\cos(\theta)) - (-r\sin(\theta) \times \sin(\theta)) \\ &= r\cos^2(\theta) + u\sin^2(\theta) = r. \end{aligned}$$

Hence,

$$f_{R,\Theta}(r,\theta) = \begin{cases} r f_{X,Y}(r\cos\theta, r\sin\theta) & \text{if } r > 0 \text{ and } \pi < \theta < \pi, \\ 0 & \text{otherwise.} \end{cases}$$

You should check that this yields Example 29.4, for instance.

Example 29.7. Let us compute the joint density of $U = X$ and $V = X + Y$. Here,

$$\begin{aligned} u(x,y) &= x \\ v(x,y) &= x + y \\ x(u,v) &= u \\ y(u,v) &= v - u. \end{aligned}$$

Therefore,

$$J(u,v) = (1 \times 1) - (0 \times -1) = 1.$$

Consequently,

$$f_{U,V}(u,v) = f_{X,Y}(u, v-u).$$

This has an interesting by-product: The density function of $V = X + Y$ is

$$\begin{aligned} f_V(v) &= \int_{-\infty}^{\infty} f_{U,V}(u, v)\, du \\ &= \int_{-\infty}^{\infty} f_{X,Y}(u, v-u)\, du. \end{aligned}$$