MATH 525 Notes

Faye Jackson

September 1, 2020

1 Kolmogorov Axioms

We start with some experiment

Example. Here are some expirements

a) Flip a coin

b) Roll a die

c) Draw a ball from an urn

d) Pick a number between 0 and 1.

We will use the Kolmogorov axioms for a Probability Space $(\Omega, \mathcal{F}, P)$.

1.1 The Sample Space

Let's detail what Ω–the sample space–is in each example.

a) $\Omega = \{\mathrm{H}, \mathrm{T}\}$

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

c) $\Omega = \{\mathrm{red}, \mathrm{blue}, \mathrm{black}\}$

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

In the first three cases these are finite, and in the last case the sample space is infinite. Sometimes these are really hard to figure out!

Example. The experiment is the time I will wake up tomorrow? (say between 6am-10am). There are many possible sample spaces!

- $\Omega = [6, 10]$, measuring time with perfect precision.
- $\Omega = [6, 10] \cap \mathbb{Q}$, measuring time with arbitrary, but finite, precision
- $\Omega =$ finite, measuring time with precision up to the minute.

In the abstract set-up, Ω is any set.

1.2 The Events

This has to do with $\mathcal{F}$. An event is some set of outcomes:

Example. Going back to our first examples

a) $A = \{\mathrm{H}\} A = \{\mathrm{T}\}$

b) $A =$ any even number $= \{2, 4, 6\}$

c) $A =$ not black $= \{\text{red}, \text{blue}\}$

d) $A =$ a number less than $\frac{1}{2} = \left[0, \frac{1}{2}\right)$

An event then is a subset of Ω. For technical reasons, not all subsets of Ω will be allowed. Thus:

$$\mathcal{F} = \text{the set of subsets of } \Omega \text{ that are allowed to be events} \subseteq \mathcal{P}(\Omega)$$

We want to be able to perform the following operations provided that A and B are events:

- $A \cup B$ is an event (this represents inclusive or). We interpret this as "at least one of A or B happens."
- $A \cap B$ is an event (this represents and). We interpret this as "both A and B happen."
- $A \setminus B$ is an event. We interpret this as "A happens but B does not happen"

- $A \triangle B = (A \cup B) \setminus (A \cap B)$ is an event, called the symmetric difference in set theory. We interpret this as "exactly one of A and B happens."

More generally, we require:

- If $A_1, A_2, A_3, \ldots$ are events, then so is their union:

$$\bigcup_{n=1}^{\infty} A_n = \{x \in \Omega \mid \text{there exists some } n \text{ such that } x \in A_n\}$$

  This expresses inclusive or as above.

- If A is an event then so is $A^c = \Omega \setminus A$.

Definition. *A set $\mathcal{F}$ of subsets of Ω is called a σ-field (or σ-algebra) provided that it satisfies the above two axioms and $\Omega \in \mathcal{F}$ (we call Ω the "sure" event).*

Example. The simplest example is just let $\mathcal{F} = \mathcal{P}(\Omega)$, the set of all subsets of Ω (that is the powerset). When Σ is finite we almost always allow $\mathcal{F} = \mathcal{P}(\Omega)$. When the sample space is infinite, we run intro trouble with P.

Exercise. *If $A_1, A_2, \ldots$ are events then so is their intersection:*

$$\bigcap_{n=1}^{\infty} A_n = \{x \in \Omega \mid \textit{for all } n, x \in A_n\}$$

This is justified by DeMorgan's law:

$$\left(\bigcup_{n=1}^{\infty} A_n\right)^c = \bigcap_{n=1}^{\infty} A_n^c$$

1.3 Group Work

Justify the following:

1. Given two subsets A, B argue that $(A \cup B)^c = A^c \cap B^c$
2. Generalize this to DeMorgan's law as given above
3. Conclude that $A \cap B$, $A \setminus B$, and $\bigcap_{n=1}^{\infty} A_n$ are all events.

Proof of 1. We will do this by two-way inclusion:

($\subseteq$) Fix $x \in (A \cup B)^c$. Then $x \notin A \cup B$, and therefore x is not in A and x is not in B. Therefore $x \in A^c$ and $x \in B^c$, implying $x \in A^c \cap B^c$.

($\supseteq$) Fix $x \in A^c \cap B^c$. Then $x \in A^c$ and $x \in B^c$, implying that x is not in A and x is not in B. Therefore $x \notin A \cup B$, and so $x \in (A \cup B)^c$.

Proof of 2. Let's go!

($\subseteq$) Fix $x \in (\bigcup_{n=1}^{\infty} A_n)^c$. Then we know that there does not exist $n \in \mathbb{N}$ so that $x \in A_n$. Therefore for all $n \in \mathbb{N}$ we must have $x \notin A_n$, and thus $x \in A_n^c$ for all $n \in \mathbb{N}$, giving us that $x \in \bigcap_{n=1}^{\infty} A_n^c$.

($\supseteq$) Fix $x \in \bigcap_{n=1}^{\infty} A_n^c$. Then for all $n \in \mathbb{N}$ we must have $x \in A_n^c$, that is $x \notin A_n$. By negation then there deos not exists an $n \in \mathbb{N}$ such that $x \in A_n$, and so:

$$x \in \left(\bigcup_{n=1}^{\infty} A_n \right)^c$$

Just as desired ☺

Proof of 3. Suppose $A, B, A_1, A_2, \ldots$ are events. We now proceed

- We know that A^c and B^c are then events by defintion of a σ-field, and thus we also have $A^c \cup B^c$ is an event as well. This then gives:

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

  is an event, as desired

- Since B is an event we know that B^c is an event. Therefore by the earlier proof we know $A \cap B^c$ is an event. We now show that $A \setminus B = A \cap B^c$.

- ($\subseteq$) Fix $x \in A \setminus B$, then $x \in A \subseteq \Omega$ and $x \notin B$. Therefore $x \in B^c$ as well and so we have $x \in A \cap B^c$.
- ($\supseteq$) Fix $x \in A \cap B^c$, then $x \in A$ and $x \in B^c$, implying that $x \notin B$. Therefore $x \in A \setminus B$ as desired.

This means we must have that $A \setminus B$ is an event.

- Consider that by assumptions we know A_n^c is an event for all $n \in \mathbb{N}$. Therefore we must also have that $\bigcup_{n=1}^{\infty} A_n^c$ is an event. By taking complements we then know:

$$\left(\bigcup_{n=1}^{\infty} A_n^c\right)^c = \bigcap_{n=1}^{\infty} (A_n^c)^c = \bigcap_{n=1}^{\infty} A_n$$

And so this is an event as well, just as desired.

Great! We win ☺

1.4 The probability measure P

P is a function $\mathcal{F} \to [0,1]$, which assigns a probability to each event. P must satisfy the axioms:

- $P(\Omega) = 1$.
- If $A_1, A_2, \ldots$ are events such that $A_n \cap A_m$ is empty for all naturals $n \neq m$. Then:

$$P\left(\sum_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} P(A_n)$$

This is called σ-additivity.

Definition. *A function P as above is called a probability measure*

Definition. *A probability space is a triple $(\Sigma, \mathcal{F}, P)$ where Ω is any set, $\mathcal{F}$ is a σ-field of subsets of Ω, and $P : \mathcal{F} \to [0,1]$ is a probability measure.*

MATH 525 Notes

Faye Jackson

September 3, 2020

1 Last Time

We talked about what a probability space $(\Omega, \mathcal{F}, P)$ is:

- Ω is a sample space.
- $\mathcal{F}$ is a σ-field of events
- $P : \mathcal{F} \to [0, 1]$ is a probability measure.

2 Finite Sample Spaces

We will assume for today that Ω is finite and take $\mathcal{F} = P(\Omega)$.

Definition. *For any $x \in \Omega$ such that $\{x\} \in \mathcal{F}$, we call $\{x\}$ an atomic event*

Note that for Ω finite and $\mathcal{F} = P(\Omega)$ we must have all the possible atomic events, and moreover:

$$P(A) = P\left(\bigcup_{a \in A} \{a\}\right) = \sum_{a \in A} P(\{a\})$$

We can thus conclude that knowing the probability of the atomic events determines the whole probability measure.

2.1 Symmetric Spaces

Definition. *We call a probability space a symmetric probability space provided that $P(\{x\})$ does not depend on x for all $x \in \Omega$. That is all outcomes are equally likely. Thus when Ω is finite we have:*

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

Example. Flip a fair coin

Non-Example. Flip an unfair coin

Example. Let there be n people, including Ali and Bo, who are seated in a row. All arrangments are equally likely. What is $P(\text{Ali is left of Bo})$?

Consider the function $f : \Omega \to \Omega$, where we switch the places of Ali and Bo. f is an involution with no fixed points, and consider that:

$$A = \{\text{Ali is left of Bo}\}$$
$$A^c = \{\text{Ali is right of Bo}\}$$
$$f : A \leftrightarrow A^c$$
$$|A| = |A^c|$$

Because f is in particular a bijection. But now we win! Watch:

$$\begin{aligned} 1 &= P(\Omega) = P(A) + P(A^c) \\ &= P(A) + P(A) \\ P(A) &= \frac{1}{2} \end{aligned}$$

Example. Same set up. but now $P(\text{Ali is directly left of Bo})$. We split into two cases, in the first we consider that $P(\text{Bo is left-most}) = \frac{1}{n}$, and so $P(\text{Bo is not left-most}) = \frac{n-1}{n}$.

In the later case, $P(\text{Ali is directly left of Bo}) = \frac{1}{n-1}$ sinc there are $n-1$

people. Therefore:

$$P(\text{Ali is directly left of Bo}) = \frac{n-1}{n} \cdot \frac{1}{n-1} = \frac{1}{n}$$

2.2 Group Work

For any finite Ω, show that:

- $P(A^c) = 1 - P(A)$
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
- This one is nasty

  $$\begin{aligned} P(A \cup B \cup C) = &\, P(A) + P(B) + P(C) \\ &- P(A \cap B) - P(A \cap C) - P(B \cap C) + P(A \cap B \cap C) \end{aligned}$$

Proof. Let's go!

- Note that $A \cap A^c = \emptyset$. Thus:

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

  Rearranging we find $P(A^c) = 1 - P(A)$.

- Let $D = B \setminus (A \cap B)$. Then $B = D \cup (A \cap B)$, where this is a disjoint union so we can say $P(B) = P(D) + P(A \cap B)$. Now note that $A \cup B = A \cup D$, where the latter union is disjoint:

  $$\begin{aligned} P(A \cup B) &= P(A \cup D) = P(A) + P(D) \\ &= P(A) + P(B) - P(A \cap B) \end{aligned}$$

- Consider by the second bullet:

$$\begin{aligned}
P((A\cup B)\cup C) &= P(A\cup B) + P(C) - P((A\cup B)\cap C)\\
&= P(A) + P(B) - P(A\cap B) - P((A\cap C)\cup(B\cap C))\\
&= P(A) + P(B) - P(A\cap B)\\
&\quad - P(A\cap C) - P(B\cap C) + P((A\cap C)\cap(B\cap C))\\
&= P(A) + P(B) - P(A\cap B)\\
&\quad - P(A\cap C) - P(B\cap C) + P(A\cap B\cap C)
\end{aligned}$$

Example. Ali and Bo are taking a class together. They have been able to estimate that:

$$\begin{aligned}
P(\text{Ali gets a B}) &= 0.3\\
P(\text{Bo gets a B}) &= 0.4\\
P(\text{neither gets an A but at least one gets a B}) &= 0.1
\end{aligned}$$

They want to know the probability that neither gets a C but at least one gets a B. Note that the sample space is:

$$\begin{gathered}
\Omega = \{A, B, C\}^2\\
P(\{BA, BB, BC\}) = 0.3\\
P(\{AB, BB, CB\}) = 0.4\\
P(\{BB, BC, CB\}) = 0.1\\
P(\{AB, BA, BB\}) = ?
\end{gathered}$$

We then note that:

$$\begin{aligned}
P(\{AB, BA, BB\}) &= P(\{AB, BA, BB, BC, CB\}) - P(\{CB, BC\})\\
&= P(\{BA, BB, BC\}) + P(\{AB, BB, CB\}) - P(\{BB\}) - P(\{CB, BC\})\\
&= 0.3 + 0.4 - 0.1 = 0.6
\end{aligned}$$

3 Why bother with $\mathcal{F}$

Take $\Omega = [0,1)$. For any $q \in \mathbb{Q}$ and any subset $A \subseteq [0,1)$, define a new subset $A + q \subseteq \Omega$ by first translating the set and then taking the fractional part of each element of the translated set:

$$A + q = \{x \in \Omega \mid \exists z \in \mathbb{Z},\ x + z - q \in A\}$$

There exists a set $A \subseteq \Omega$ such that:

- $A \cap A + q = \emptyset$ for any $q \in [0,1) \cap \mathbb{Q}$. Note that $A + p \cap A + q = A \cap (A + (q - p)) + p$.
- Also:

$$\Omega = \bigcup_{q \in [0,1) \cap \mathbb{Q}} A + q$$

But then the axioms of a probability space tell us that:

$$1 = P(\Omega) = \sum_{q \in [0,1) \cap \mathbb{Q}} P(A + q) = \sum_{q \in [0,1) \cap \mathbb{Q}} P(A)$$

We get the last equality because we want our measure to be "translation invariant," and so $P(A+q) = P(A)$ for any rational q. If $P(A) = 0$ then we get $1 = 0$, and if $P(A) > 0$ then we get $1 = \infty$. This is bad. Thus A should not be allowed to be an event.

To get A we use the concept of an equivalence relation and the Axiom of Choice. Declare $x, y \in \Omega$ equivalent if $x - y \in \mathbb{Q}$. Then we must have that Ω is the disjoint union of the equivalence classes. The axiom of choice gives a set A which contains exactly one number from each equivalence class.

Example. Let $\Omega = [0,1]$. Let $\mathcal{F}$ be the smallest σ-field that contains all intervals. This is called the Borel σ-field. You can prove that this exists!

We let P be the unique probability measure such that for any $a \leq b$:

$$P((a,b)) = P([a,b]) = b - a$$

This P is called the Lebesgue measure. Note that for any $x \in [0,1]$ we have:

$$P(\{x\}) = 0$$

3.1 Properties of the Lebesgue measure

There are some nice properties of P.

Theorem. *Consider that if $A_1 \subseteq A_2 \subseteq \cdots$ is an ascending chain of events. Then:*

$$P\left(\bigcup_{n=1}^{\infty} A_n\right) = \lim_{n\to\infty} P(A_n)$$

Proof. Let $B_n = A_n \setminus A_{n-1}$ where $A_0 = \emptyset$. Then:

$$\bigcup_{n=1}^{\infty} A_n = \coprod_{n=1}^{\infty} B_n$$

Therefore:

$$\begin{aligned} P\left(\bigcup_{n=1}^{\infty} A_n\right) &= P\left(\coprod_{n=1}^{\infty} B_n\right) = \sum_{n=1}^{\infty} P(B_n) \\ &= \lim_{n\to\infty}\left(\sum_{k=1}^{n} P(B_k)\right) = \lim_{n\to\infty} P(A_n) \end{aligned}$$

Corrolary. *If $A_1 \supseteq A_2 \supseteq \cdots$ is a descending tower of events then:*

$$P\left(\bigcap_{n=1}^{\infty} A_n\right) = \lim_{n\to\infty} P(A_n)$$

Proof. But note that $A_1^c \subseteq A_2^c \subseteq \cdots$ is an ascending chain and so:

$$\begin{aligned} P\left(\bigcap_{n=1}^{\infty} A_n\right) &= 1 - P\left(\left(\bigcap_{n=1}^{\infty} A_n\right)^c\right) \\ &= 1 - P\left(\bigcup_{n=1}^{\infty} A_n^c\right) \\ &= 1 - \lim_{n\to\infty} A_n^c \\ &= \lim_{n\to\infty} (1 - A_n^c) \\ &= \lim_{n\to\infty} A_n \end{aligned}$$

MATH 525 Notes

Faye Jackson

September 8, 2020

1 Conditional Probability

Let AB be events and assume $P(B) > 0$. Then we define:

Definition. *The following probability:*

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

We call this "the probability of A assuming B"

Remark. Formally we replace Ω with $\Omega_B = B$ and $\mathcal{F}$ by $\mathcal{F}_B = \{A \cap B \mid A \in \mathcal{F}\}$. Then $P_B(A \cap B) = P(A \mid B)$.

Definition. *We say that A is <u>independent</u> of B provided that $P(A \mid B) = P(A)$. Furthermore we say that B <u>attracts</u> A provided that $P(A \mid B) > P(A)$, and that B <u>repels</u> A provided that $P(A \mid B) < P(A)$.*

Remark. The following holds for A and B:

$$\begin{aligned}
P(A \cap B) = P(A) \cdot P(B) &\iff A \text{ is independent from } B \\
P(A \cap B) < P(A) \cdot P(B) &\iff B \text{ repels } A \\
P(A \cap B) > P(A) \cdot P(B) &\iff B \text{ attracts } A
\end{aligned}$$

Also the relations of independence, attraction, and repelling are symmetric.

Example. Let A be the probability that a card is an ace and B be the probability that the card is spades. Note that:

$$\begin{aligned} P(A) &= \frac{4}{52} = \frac{1}{12} \\ P(B) &= \frac{1}{4} \\ P(A \cap B) &= \frac{1}{52} = \frac{1}{12} \cdot \frac{1}{4} = P(A) \cdot P(B) \end{aligned}$$

Now remove the Jack of Hearts. Same question:

$$\begin{aligned} P(A) &= \frac{4}{51} \\ P(B) &= \frac{13}{51} \\ P(A \cap B) &= \frac{1}{51} < P(A) \cdot P(B) = \frac{52}{51} \cdot \frac{1}{51} \end{aligned}$$

Observe 1. A, B are independent is not at all the same as $A \cap B = \emptyset$.

2 Group Work

1) Let A, B be events, $0 < P(B) < 1$. Show that:

$$P(A) = P(A \mid B) \cdot P(B) + P(A \mid B^c) \cdot P(B^c)$$

2) If $A \subseteq B$ then $P(A) \leq P(B)$. This inclusion means B implies A.

3) If $A \subseteq B$ and $P(A) = P(B)$. Does it follows that $A = B$?

Proof of 1. Note that $0 < P(B^c) = 1 - P(B) < 1$. And so:

$$\begin{aligned} (A \cap B) \sqcup (A \cap B^c) &= A \\ \frac{P(A \cap B)}{P(B)} \cdot P(B) + \frac{P(A \cap B^c)}{P(B^c)} \cdot P(B^c) &= P(A \cap B) + P(A \cap B^c) \\ &= P(A) \end{aligned}$$

Proof of 2. Note that:

$$P(B) = P(A) + P(B \setminus A) \geq P(A)$$

And so we win! ☺

Non-Example. Let A, B be events with $P(A) > 0$ and $P(B) = 0$ and $A, B \neq \emptyset$ and $A \cap B = \emptyset$. Then:

$$\begin{aligned} A &\subseteq A \cup B \\ P(A) &= P(A \cup B) \\ A &\neq A \cup B \end{aligned}$$

For another example take A to be the probability that a randomly chosen number from $[0, 1]$ is less than a half but not equal to one third. Then take B to be the probability that a randomly chosen number from $[0, 1]$ is less than one half.

Corrolary. *If $B_1, \ldots, B_n$ are events such that:*

- $\Omega = B_1 \cup \cdots \cup B_n$
- $B_i \cap B_j = \emptyset$ *for any* $1 \leq i, j \leq n$ *and* $i \neq j$
- $P(B_i) > 0$ *for* $1 \leq i \leq n$

Then for any event A we have:

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

Example. There are two kinds of covid tests. The Type I tests are defective 20% of the time. Then Type II tests are defective 5% of the time. There are twice as many Type II tests.

Question: Having chosen a random test what is the chance that it's not defective? This is A.

Let B be the event tha tthe test is of type I, then:

$$\begin{aligned} P(A) &= P(A \mid B) \cdot P(B) + P(A \mid B^c) \cdot P(B^c) \\ &= \frac{4}{5} \cdot \frac{2}{3} + \frac{19}{20} \cdot \frac{1}{3} = \frac{51}{60} \end{aligned}$$

You can flip things around and ask what is the probability that a defective test is of type I:

$$\begin{aligned} P(B \mid A^c) &= \frac{P(B \cap A^c)}{P(A^c)} = \frac{P(A^c \cap B)}{P(A^c)} \\ &= \frac{P(A^c \mid B) \cdot P(B)}{P(A^c)} \\ &= \frac{\frac{1}{5} \cdot \frac{2}{3}}{\frac{9}{60}} = \frac{8}{9} \end{aligned}$$

Example. Coin tosses. Consider all patterns that occur when tossing a coin three times. You and I choose a pattern. We begin tossing a fair coin. Whoever's pattern appears first wins.

Claim. *If you choose first. I can always choose a pattern such that my win probability is at least two thirds.*

Example. You choose HTH as your pattern. I choose the pattern HHT given this. Let X be the probability that I win. If we toss tails then the win probability doesn't change. If we toss heads twice then I always win, because eventually we will toss a tail. If we toss HTH then you win. If we go HTT then the probability is x again. Therefore:

$$\begin{aligned} x &= \frac{1}{2}\left(x + \frac{1}{2}\left(1 + \frac{1}{2}(0 + x)\right)\right) \\ x &= \frac{5}{8}x + \frac{1}{4} \implies x = \frac{2}{3} \end{aligned}$$

We can always do this kind of choosing, so we win at least two thirds of the time.

Definition. *Take a collection $\{A_i\}_{i \in I}$ be a collection of events indexed by a set I of arbitrary cardinality. The collection $\{A_i\}_{i \in I}$ is independent provided* *that for any finite subcollection $J \subseteq I$ we have:*

$$P\left(\bigcap_{j\in J} A_j\right) = \prod_{j\in J} P(A_j)$$

The collection is pairwise independent provided that any two events are independent.

Remark. Note that independence easily implies pairwise independence. However the converse is not true.

Example. Roll a die twice. Look at the following three events. A says the first die shows a three. B says second die shows four. C says the sum of the two rolls is seven:

$$P(A) = P(B) = P(C) = \frac{1}{6}$$
$$P(A\cap B) = P(A\cap C) = P(B\cap C) = \frac{1}{36}$$
$$P(A\cap B\cap C) = \frac{1}{36}$$

Thus this collection is pairwise independent but not independent

MATH 525 Notes

Faye Jackson

September 10, 2020

Basic Combinatorics

Let k, n be natural numbers:

Repeating Experiments

Say you have an experiment with k outcomes. Repeating n times gives k^n outcomes. If Ω is the sample space of the experiment. Then Ω^n is the sample space of the experiment repeated n times. Furthermore:

$$P(\{(x_1, \dots, x_n)\}) = P(\{x_1\}) \cdot P(\{x_2\}) \cdots P(\{x_n\})$$

The underlying assumption is that these events are independent.

Definition. *If* $(\Omega_1, \mathcal{F}_1, P_1)$ *and* $(\Omega_2, \mathcal{F}_2, P_2)$ *are probability spaces. Their product is* $(\Omega, \mathcal{F}, P)$ *where:*

- $\Omega = \Omega_1 \times \Omega_2 = \{(x_1, x_2) \mid x_1 \in \Omega_1, x_2 \in \Omega_2\}$
- $\mathcal{F}$ *is the smallest* σ*-field containing all* $A \times B$ *for* $A \in \mathcal{F}_1, B \in \mathcal{F}_2$
- P *is uniquely determined by* $P(A \times B) = P(A) \cdot P(B)$ *for all* $A \in \mathcal{F}_1$ *and* $B \in \mathcal{F}_2$.

Great! This should be proved as an exercise.

Proof. First lets prove that there does exist a smallest σ-field containing all $A \times B$ for $A \in \mathcal{F}_1$ and $B \in \mathcal{F}_2$. We do this by letting:

$$\mathcal{F}_* := \{A \times B \mid A \in \mathcal{F}_1, B \in \mathcal{F}_2\}$$
$$\mathcal{F} := \bigcap_{\substack{\mathcal{G} \in \Sigma(\Omega) \\ \mathcal{F}_* \subseteq \mathcal{G}}} \mathcal{G}$$

Where $\Sigma(\Omega)$ is the set of all σ-algebras over Ω. We must show this is a σ-algebra.

- Let $C_1, C_2, \ldots$ be events in $\mathcal{F}$. Now fix a $\mathcal{G} \in \Sigma(\Omega)$ such that $\mathcal{F}_* \subseteq \mathcal{G}$. Then by definition $C_1, C_2, \ldots$ are events in $\mathcal{G}$, and so $\bigcup_{n=1}^{\infty} C_n \in \mathcal{G}$. Therefore $\bigcup_{n=1}^{\infty} C_n \in \mathcal{F}$ just as desired.
- Let $X \in \mathcal{F}$. Then fix $\mathcal{G} \in \Sigma(\Omega)$ such that $\mathcal{F}_* \subseteq \mathcal{G}$. Then note that $X \in \mathcal{G}$ so $X^c \in \mathcal{G}$. Therefore we must have $X^c \in \mathcal{F}$ just as desired.

Note also that clearly $\mathcal{F}_* \subseteq \mathcal{F}$ as desired, and $\mathcal{F}$ is the smallest σ-field containing $\mathcal{F}_*$.

TODO: Probability measure

Example. Toss a coin 3 times, then $\Omega = \{H, T\}$ and $\Omega^3 = \{HHH, HHT, \ldots\}$.

Selecting a committee

We have n people and want a committee of k people, $n \geq k$. First select the most important person, then the next important, etc.

$$\text{Total \# of choices} = n(n-1)\cdots(n-k+1) = \frac{n!}{(n-k)!}$$

This number records importance. If one doesn't want to record order, must divide by # of orders, which is $k!$. Thus the number of committees without order is the integer $\frac{n!}{(n-k)!k!} =: \binom{n}{k}$. Note that:

$$\binom{n}{k} = \binom{n}{n-k}$$

Example. You are writing a letter to each of n people. Each letter is put into a random envelope. What is the probability that at least one letter is in the correct envelope:

$$A_i = i\text{-th letter is in the correct envelope}$$

$$\begin{aligned} P\left(\bigcup_{i=1}^{n} A_i\right) &= \sum_{i=1}^{n} P(A_i) - \sum_{1 \le i < j \le n} P(A_i \cap A_j) + \cdots \\ &= \sum_{k=1}^{\infty} (-1)^{k-1} \sum_{1 \le i_1 < i_2 < \cdots < i_k \le n} P\left(\bigcap_{j=1}^{k} A_{i_j}\right) \end{aligned}$$

Note first that $P(A_i) = \frac{1}{n}$ and $P(A_i \cap A_j) = \frac{1}{n} \cdot \frac{1}{n-1}$. More generally:

$$P\left(\bigcap_{j=1}^{k} A_{i_j}\right) = \frac{1}{n(n-1)\cdots(n-k+1)} = \frac{(n-k)!}{n!}$$

Note that this summand is independent of the indexing set. Look at an example:

- Note first that:

  $$|\{(i,j) \mid 1 \le i < j \le n\}| = \frac{1}{2} \cdot |\{(i,j), 1 < i \ne j \le n\}| = \frac{1}{2} \cdot n \cdot (n-1)$$

- In general:

  $$|\{(i_1, \ldots, i_k) \mid 1 \le i_1 < i_2 < \cdots < i_k \le n\}| = \binom{n}{k}$$

Therefore we have that:

$$
\begin{aligned}
P\left(\bigcup_{i=1}^{n} A_i\right) &= \sum_{k=1}^{\infty}(-1)^{k-1} \sum_{1\le i_1<i_2<\cdots<i_k\le n} \frac{(n-k)!}{n!} \\
&= \sum_{k=1}^{\infty}(-1)^{k-1} \cdot \binom{n}{k} \cdot \frac{(n-k)!}{n!} \\
&= -\sum_{k=1}^{\infty} \frac{(-1)^k}{k!} = 1 - e^{-1} \approx 0.63
\end{aligned}
$$

By recalling that:

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

Group Work

You deal 5 cards. What is the probability that exactly three are aces?

$$
\begin{aligned}
\#\text{ total deals} &= \binom{52}{5} \\
\#\text{ exactly three are aces} &= \binom{4}{3} \cdot \binom{48}{2} \\
P(A) &= \frac{\binom{4}{3} \cdot \binom{48}{2}}{\binom{52}{5}}
\end{aligned}
$$

There are two ways of thinking about this

a) Think of individual arrangements. Then:

$$\binom{5}{3} \cdot \frac{4}{52} \cdot \frac{3}{51} \cdot \frac{2}{50} \cdot \frac{48}{49} \cdot \frac{47}{48}$$

b) Another way to compute it is:

$$\frac{\binom{4}{3} \cdot \binom{48}{2}}{\binom{52}{5}}$$

Random Walks

Some setups

- A frog jumps each minutes, and with probability p goes left, and with probability $1-p$ moves right
- Flip a coin. Probability p it's heads, and $1-p$ that it's tails
- A gambler wins \$1 with probability p and loses \$1 with probability $1-p$

Formally we start with an integer k, we add 1 with probability p, and subtract 1 with probability $1-p$. A certain path is called a trajectory. You can ask:

a) Number of trajectories from $(0,k)$ to (t,n). Note that:

$$\begin{aligned} t &= \text{ups} + \text{downs} \\ n-k &= \text{ups} - \text{downs} \end{aligned}$$

But then:

$$\begin{aligned} \text{ups} &= \frac{1}{2}(\text{ups} + \text{downs}) + \frac{1}{2}(\text{ups} - \text{downs}) = \frac{1}{2}(t+n-k) \\ \text{downs} &= \frac{1}{2}(\text{ups} + \text{downs}) - \frac{1}{2}(\text{ups} - \text{downs}) = \frac{1}{2}(t-n+k) \end{aligned}$$

Then the number of trajectories is equal to:

$$\binom{t}{\frac{t+n-k}{2}} = \binom{t}{\frac{t-n+k}{2}}$$

Note that $t+n-k$ must be even or else this is not a well-defined idea. Because then the number of ups would not be a non-negative integer. Likewise if $|n-k| > t$ there are negative ups or downs. Thus we assume $|n-k| \leq t$.

b) The Reflective Principle. Let $n > k > 0$. The # of trajecctories from $(0,k)$ to (t,n) which touch the x-axis is equal to the number of all trajectories from $(0,-k)$ to (t,n). Just reflect the part before the first time the trajectory touches the x-axis

c) We can compute the number of trajectories from $(0,k)$ to (t,n) that do not touch the x axis. So we have this as:

$$\binom{t}{\frac{t+n-k}{2}} - \binom{t}{\frac{t+n+k}{2}}$$

Example. A gambler plays a fair game where he wins/loses \$1, 50 turns. What's his probability of starting at \$10and ending at \$20, without going broke.

$$2^{-50} \cdot \left(\binom{50}{\frac{50+20-10}{2}} - \binom{50}{\frac{50+20+10}{2}} \right) = 2^{-50} \cdot \left(\binom{50}{30} - \binom{50}{40} \right) = 0.04$$

In an unfair game with 0.6 change of going up, we get:

$$0.6^{30} \cdot (0.4)^{20} \cdot \left(\binom{50}{\frac{50+20-10}{2}} - \binom{50}{\frac{50+20+10}{2}} \right) = 0.6^{30} \cdot (0.4)^{20} \cdot \left(\binom{50}{30} - \binom{50}{40} \right) = 0.11$$

MATH 525 Notes

Faye Jackson

September 15, 2020

Last Time

Recall. Random Walk: The trajectories from $(0, k)$ to (t, n). This is:

$$\binom{t}{\frac{t+n-k}{2}} = \binom{t}{\frac{t-n+k}{2}}$$

Trajectories from $(0, k)$ to (t, n) that touches the x-axis

$$\binom{t}{\frac{t+n+k}{2}}$$

And so trajectories that do not touch the x-axis are:

$$\binom{t}{\frac{t+n-k}{2}} - \binom{t}{\frac{t+n+k}{2}}$$

Example. Flip a fair coin $2n$ times. Think of n being large. Which is more likely?
A: at the end there is exactly as many H as T. B: at any time, there are more H than T, or at any time, there are more T than H.

We model with a symmetric random walk. For A, we are interested in trajectories which go from $(0, 0)$ to $(2n, 0)$.

$$\#A = \binom{2n}{n} \qquad\qquad P(A) = 2^{-2n}\binom{2n}{n}$$

For B, note that the two events making it up are disjoint and symmetric:

$$P(B) = 2 \cdot P(\text{at any time, more } H \text{ than } T)$$

So we go $(0,0) \to (1,1) \to (2n, 2k)$ for $0 < k < n$ without touching the x-axis, we count these:

$$\binom{2n-1}{\frac{2n-1+2k-1}{2}} - \binom{2n-1}{\frac{2n-1+2k+1}{2}} = \binom{2n-1}{n+k-1} - \binom{2n-1}{n+k}$$

If $k = n$, there's only one way to go from $(1,1)$ to $(2n, 2n)$.

$$\text{Total \# of trajectories} = 1 + \sum_{k=1}^{n-1} \left(\binom{2n-1}{n+k-1} - \binom{2n-1}{n+k} \right)$$

But wait this is a telescoping sum!

$$\text{Total \# of these trajectories} = 1 + \binom{2n-1}{n} - \binom{2n-1}{2n-1} = \binom{2n-1}{n}$$

and so we then have:

$$\begin{aligned} P(B) &= 2^{-2n} \cdot 2 \cdot \binom{2n-1}{n} = 2^{-2n} \cdot 2 \cdot \frac{(2n-1)!}{n!(n-1)!} \\ &= 2^{-2n} \cdot \frac{2n(2n-1)!}{n!n!} = 2^{-2n} \cdot \binom{2n}{n} \end{aligned}$$

And so $P(A) = P(B)$!!!! Wow!

Stirling's Formula:

$$n! \sim n^n \cdot e^{-n} \cdot \sqrt{2\pi n}$$

With this formula:

$$P(A) \sim 2^{-2n} \cdot \frac{(2n)^{(2n)} e^{-2n} \cdot \sqrt{4\pi n}}{n^{2n} \cdot e^{-2n} \cdot 2\pi n} = \frac{1}{\sqrt{\pi \cdot n}}$$

Absorbing Barriers

Example. We consider again a gambler. They start with $\$k$ and they play until:

- They go broke
- They end up with $\$n$ and buy a Porsche.

Let s_k be the probability of success, starting at k. Let f_k be the probability of failure. We also let p be the probability of getting \$1 in each thurn

We'll apply the conditional probability formula:

$$P(A) = P(A \mid B) \cdot P(B) + P(A \mid B^c) \cdot P(B^c)$$

Then we can write that:

$$s_k = p \cdot s_{k+1} + (1-p)s_{k-1}$$

This is a difference equation, which is a discrete analog of a differential equation. Thus we're essentially solving a differential equation with boundary conditions. Namely we know $s_0 = 0$ and $s_n = 1$.

Case 1: $p = \frac{1}{2}$ Then we have:

$$s_k = \frac{1}{2}(s_{k+1} + s_{k-1})$$
$$2s_k = s_{k-1} - s - k - 1$$
$$s_{k+1} - s_k = s_k - s_{k-1}$$

This means that the difference in values never changes, thus call $x :=$

$s_1 - s_0$, so that:

$$\sum_{j=1}^{k}(s_j - s_{j-1}) = k \cdot x$$
$$\sum_{j=1}^{k}(s_j - s_{j-1}) = s_k - s_0 = s_k$$
$$s_n = n \cdot x = 1 \implies x = \frac{1}{n}$$

Thus $s_k = \frac{k}{n}$. Note that geometrically success and failure are exactly symmetric by a reflection, and we also know $p = \frac{1}{2}$ so this is truly symmetric. Thus $f_k = s_{n-k} = \frac{n-k}{n}$. Thus:

$$s_k + f_k = 1$$

This is weird, the event of bouncing around in the middle forever is non-empty.

Claim. *In fact if $n \geq 4$, the number of trajectories that are not absorbed is uncountable.*

Proof. It is enough to assume $n = 4$. Then we count trajectories between 1 and 3. At every second step we know the particle on this trajectory is at 2 and has 2 choices. Thus these trajectories are in bijection with:

$$\{f : 2\mathbb{N} \to \{+1, -1\}\} \cong P(\mathbb{N}) \cong \mathbb{R}$$

Thus there are uncountably many such trajectories.

Case 2: $p \neq \frac{1}{2}$ Then $s_k = p \cdot s_{k+1} + (1-p)s_{k-1}$. Then:

$$s_{k+1} - s_k = \frac{1-p}{p}(s_k - s_{k-1})$$

Where $\alpha := \frac{1-p}{p}$. Let $x := s_1 - s_0$. Then:

$$
\begin{aligned}
s_k &= (s_k - s_{k-1}) + (s_{k-1} - s_{k-2}) + \cdots + (s_1 - s_0) \\
&= x(\alpha^{k-1} + \alpha^{k-2} + \cdots + \alpha + 1) \\
&= x \cdot \frac{\alpha^k - 1}{\alpha - 1} \\
1 = s_n & \qquad\qquad = x \cdot \frac{\alpha^n - 1}{\alpha - 1} = x \cdot \frac{\alpha^n - 1}{\alpha - 1} \\
x &= \frac{\alpha - 1}{\alpha^n - 1} \\
s_k &= \frac{\alpha^k - 1}{\alpha^n - 1}
\end{aligned}
$$

Now we have symmetry again, but we also have to replace p with $1-p$, therefore:

$$f_k = \frac{\alpha^{k-n} - 1}{\alpha^{-n} - 1}$$

Therefore:

$$s_k + f_k = 1$$

What happens if we let $n \to \infty$. In case 1:

$$s_k = \frac{k}{n} \to 0$$

This is called the gambler's ruin. If you play against a bank with unlimited money. You lose with probability 1. For case 2:

$$s_k = \frac{\alpha^k - 1}{\alpha^n - 1} = \begin{cases} 0 & \text{if } \alpha > 1, p < \frac{1}{2} \\ 1 - \alpha^k & \text{if } \alpha < 1, p > \frac{1}{2} \end{cases}$$

Group Work

Same game, so 0 is ruin, n is success, k is start. p is the probability of $+$. Then the decision is to get \$1 or \$$\frac{1}{2}$. Which is better?

Consider that:

$$s^{\frac{1}{2}}_{k,n} = s^{1}_{2k,2n}$$

So in both cases

- When $p = \frac{1}{2}$:

$$s^{1}_{k,n} = \frac{k}{n}$$
$$s^{\frac{1}{2}}_{k,n} = \frac{2k}{2n} = \frac{k}{n}$$

- When $p \neq \frac{1}{2}$:

$$s^{1}_{k,n} = \frac{\alpha^k - 1}{\alpha^n - 1}$$
$$s^{\frac{1}{2}}_{k,n} = \frac{\alpha^{2k} - 1}{\alpha^{2n} - 1}$$

So lets divide them:

$$\begin{aligned}\frac{s^{\frac{1}{2}}_{k,n}}{s^{1}_{k,n}} &= \frac{\alpha^{2k} - 1}{\alpha^k - 1} \cdot \frac{\alpha^n - 1}{\alpha^{2n} - 1} \\ &= \frac{\alpha^k + 1}{\alpha^n + 1} = \begin{cases} > 1 & \text{if } \alpha < 1, p > \frac{1}{2} \\ < 1 & \text{if } \alpha > 1, p < \frac{1}{2} \end{cases}\end{aligned}$$

MATH 525 Notes

Faye Jackson

September 17, 2020

Random Variables

The definition

Definition. *Let $(\Omega, \mathcal{F}, P)$ be a probability space. A random variable is a function $x : \Omega \to \mathbb{R}$. such that for every real number y, $\{\omega \in \Omega \mid x(\omega) \leq y\}$ is an event. As shorthand, we write this event as $\{x \leq y\}$*

Example. We're going to toss a coin twice, counting the number of H. Then $\Omega = \{HH, TH, HT, TT\}$, and each point gets assigned values $2, 1, 1, 0$.

Remark. Let's think about how that condition works!

- If Ω is finite, $\mathcal{F} = P(\Omega)$, then the condition that $\{x \leq y\}$ is an event is tautologically true. We can ignore it
- In general there is a natural σ-field of subsets of $\mathbb{R}$, the Borel σ-field. It is generated by any of the following collections of subsets:
  - All open intervals
  - All closed intervals
  - $(-\infty, a]$ for all $a \in \mathbb{R}$
  - $(-\infty, a)$ for all $a \in \mathbb{R}$
  - (a, ∞) for all $a \in \mathbb{R}$
  - $[a, \infty)$ for all $a \in \mathbb{R}$.

The condition that $\{x \le y\}$ for all $y \in \mathbb{R}$ is equivalent to the condition that: For every $B \subseteq \mathbb{R}$ in the Borel σ-field, $\{\omega \in \Omega \mid x(\omega) \in B\}$ is an event. We write this in shorthand as $\{x \in B\}$. Such a function x is called measurable

Definition. *Suppose we have two spaces with σ-fields $(\Omega_1, \mathcal{F}_1)$ and $(\Omega_2, \mathcal{F}_2)$. We call a function $f : \Omega_1 \to \Omega_2$ measurable provided that for every $A \in \mathcal{F}_2$ we have $f^{-1}(A) \in \mathcal{F}_1$.*

Example. For any $a \in \mathbb{R}$, $\{a\}$ is in the Borel σ-field so the event $\{x = a\} = \{\omega \in \Omega \mid x(\omega) = a\} \in \mathcal{F}$.

For any $a, b \in \mathbb{R}$ all of (a, b), $(a, b]$, $[a, b)$, $[a, b]$ are in the Borel σ-field so:

$$\begin{array}{ll} \{a < x < b\} & \{a < x \le b\} \\ \{a \le x < b\} & \{a \le x \le b\} \end{array}$$

Remark. You can think of a random variable as a real number that depends on a certain experiment.

Definition. *If $A \subseteq \Omega$ is an event. Then the characteristic function:*

$$\begin{aligned} \mathbb{1}_A &: \Omega \to \mathbb{R} \\ \omega &\mapsto \begin{cases} 1 & \text{if } \omega \in A \\ 0 & \text{if } \omega \notin A \end{cases} \end{aligned}$$

is a random variable.

Proof. Note that for any y:

$$\{\mathbb{1}_A \le y\} = \begin{cases} \Omega & \text{if } y \ge 1 \\ A^c & \text{if } y < 1 \end{cases}$$

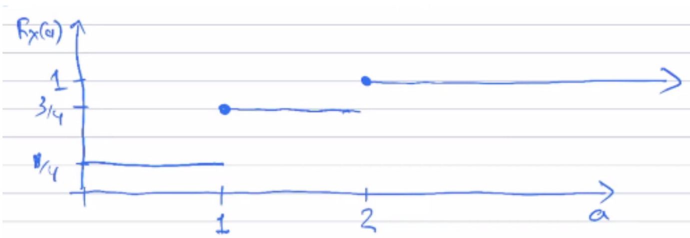
Therefore this is always an event

Lemma. *If X and Y are random variables and c is a constant, then $X+Y$, XY, and c are random variables*

Proof. **TODO**

Definition. *The cumulative distribution function of a random variable X is $F_X : \mathbb{R} \to [0,1]$ defined as $F_X(a) = P(x \leq a)$.*

Example. Toss a coin twice, $X = \#H$.



Group Work

1) We throw a dart at a circle of radius 1. We let X be the distance from the center. Compute F_X

2) Let X be any random variable whatsoever. Prove that

   i) $0 \leq F_X \leq 1$
   ii) $\lim_{a\to-\infty} F_X(a) = 0$
   iii) $\lim_{a\to\infty} F_X(a) = 1$.
   iv) $\lim_{b\downarrow a} F_X(b) = F(a)$ (The notation $b \downarrow a$ means b approaches a from above)
   v) If $a \leq b$ then $F_X(a) \leq F_X(b)$

Let's Go!

1) If a is non-negative then $F_X(a) = P(X \leq a) = a^2$ because the area of the inner circle we want to hit is πa^2 and the area of the whole dartboard is π. If a is negative then $F_X(a) = 0$.

2) Proof stuff

   i) $0 \leq F_X(a) = P(X \leq a) \leq 1$ because P is a probability measure.

ii) Pick a monotonic sequence of real numbers (a_k) converging to $-\infty$. We want to show that $\lim_{k\to\infty} F_X(a_k) = 0$. Define:

$$A_k = \{X \leq a_k\}$$

We know from (v) that $A_{k+1} \subseteq A_k$. This is a descending tower of events. Using lemma from class:

$$\lim_{k\to\infty} F_X(a_k) = \lim_{k\to\infty} P(A_k) = P\left(\bigcap_{k=1}^{\infty} A_k\right)$$

We claim that this intersection is empty. Fix some $\omega \in \Omega$, we can choose $a_j < X(\omega)$. Then $\omega \notin A_j$, and so $\omega \notin \bigcap_{k=1}^{\infty} A_k$. This is the end of the proof.

iii) Pick a monotonic sequence of real numbers (a_k) converging to ∞. We want to show that $\lim_{k\to\infty} F_X(a_k) = 0$. We can define $A_k = \{X \leq a_k\}$ as above. This is then an ascending tower of events so:

$$\lim_{k\to\infty} F_X(a_k) = \lim_{k\to\infty} P(A_k) = P\left(\bigcup_{k=1}^{\infty} A_k\right)$$

For any $\omega \in \Omega$, let $a_j > X(\omega)$, so $\omega \in A_j \subseteq \bigcup_{k=1}^{\infty}$. Thus this union is the whole sample space and we win.

iv) Consider a monotonically decreasing sequence (a_k) converging to a. Then A_k is a descending tower and we have:

$$\bigcap_{k=1}^{\infty} A_k = A = \{X \leq a\}$$

We omit the two-way containment proof. Therefore:

$$\lim_{k\to\infty} F_X(a_k) = \lim_{k\to\infty} P(A_k) = P\left(\bigcap_{k=1}^{\infty} A_k\right) = P(A)$$

v) Let $a \leq b$. Then note that:

$$\{X \leq a\} \subseteq \{X \leq b\}$$
$$P(X \leq a) \leq P(X \leq b)$$
$$F_X(a) \leq F_X(b)$$

Remark. It is not true that $\lim_{b \uparrow a} F_X(b) = F_X(a)$.

MATH 525 Notes

Faye Jackson

September 22, 2020

Recall. We have a random variable $X : \Omega \to \mathbb{R}$. We have a cumulative distribution function $F_X : \mathbb{R} \to [0, 1]$:

$$F_X(a) = P(X \leq a) = P(\{\omega \in \Omega \mid X(\omega) \leq a\})$$

We proved that:

i) $0 \leq F_X(a) \leq 1$

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

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

iv) $\lim_{a\downarrow b} F_X(a) = F_X(b)$

v) If $a \leq b$ then $F_X(a) \leq F_X(b)$.

Property (iv) says that F_X is right continuous, but it might not be left continuous. Note that F_X is left-continuous if and only if $P(X = a) = 0$ for all $a \in \mathbb{R}$. However, we do know that F_X has left limits.

Definition. *A function $F : \mathbb{R} \to [0, 1]$ is called a distribution function provided that it satisfies*

i) $0 \leq F(a) \leq 1$

ii) $\lim_{a\to\infty} F(a) = 0$

iii) $\lim_{a\to-\infty} F(a) = 0$

iv) $\lim_{a \downarrow b} F(a) = F(b)$

v) If $a \leq b$ *then* $F(a) \leq F(b)$.

Proposition. *If F is a distribution function, then there exists a probability space $(\Omega, \mathcal{F}, P)$ and a random variable $X : \Omega \to \mathbb{R}$ so that $F_X = F$.*

Proof of a Special Case. Suppose that F is a continuous and strictly increasing distribution function:

$$\Omega = (0,1) \qquad \mathcal{F} = \text{Borel } \sigma\text{-field} \qquad P = \text{Lebesgue Measure}$$

Basic knowledge from real analysis tells us that F has an inverse function $G : (0,1) \to \mathbb{R}$. Now set $X = G$, then:

$$\{X \leq a\} = G^{-1}([-\infty, a]) = F((-\infty, a]) = (0, F(a)]$$

Since $(0, F(a)]$ is an interval, it lies in the Borel σ-field, so $\{X \leq a\}$ is an event. Thus X is a random variable. Moreover:

$$F_X(a) = P(\{X \leq a\}) = P((0, F(a)]) = F(a)$$

Definition. *A random variable X can be given two names:*

1) *discrete provided that $X(\Omega)$ is countable (or finite). In that case we define its probability mass function as $f_x(a) = P(X = a)$. Then:*

$$F_X(a) = \sum_{b=-\infty}^{a} f_x(b)$$

2) *absolutely continuous provided that there exists an integrable function $f_x : \mathbb{R} \to [0, \infty)$ such that:*

$$F_X(a) = \int_{-\infty}^{a} f_x(b)\, \mathrm{d}b$$

We call f_x the probability density function

Caution: There are random variables that are neither discrete nor continuous. There is also some disagreement in the literature as to the definition of a continuous random variable (some say continuous [Tasho and book], I choose to say absolutely continuous)

Remark. Some quick things:

- If Ω is finite, we know X is discrete immediately.
- Also the indicator random variable $\mathbb{1}_A$ is a discrete random variable whatever $A \subseteq \Omega$ is.
- The cumulative distribution function of a discrete random variable that takes finitely many values is piecewise constant.
- The cumulative distribution function of an absolutely continuous random variable is in fact absolutely continuous (and thus continuous).
- The probability density function of an absolutely continuous random variable plays the same role as the probability mass function of a discrete random variable. However for an absolutely continuous random variable X we know $P(X = a) = 0$ for all $a \in \mathbb{R}$. You can think of the mass function as the density function $f_x(a) \cdot \mathrm{d}a$.
- The probability function of a discrete random variable X is uniquely determined by X. However the probability density function of an absolutely continuous random variable is not uniquely determined by X, but if f_x can be chosen to be continuous, then its continuous form is uniquely determined by X.

Interesting stuff!

Group Work

Let:

$$\Omega = [0, 1] \qquad \mathcal{F} = \text{Borel } \sigma\text{-field} \qquad P = \text{Lebesgue measure}$$

Let $X(\omega) = \frac{1}{2}(\omega + 1)$. Compute the cumulative distribution function and decide if X is discrete or absolutely continuous. Compute the probability mass function or the probability density function accordingly. Note that:

$$X(\Omega) = \frac{1}{2} \cdot (\Omega + 1) = \frac{1}{2} \cdot [1, 2] = \left[\frac{1}{2}, 1\right]$$

Therefore X is not discrete since this interval has uncountably many elements. Then note that for $\omega \in \Omega$:

$$\begin{aligned} X(\omega) &\leq a & \frac{1}{2}(\omega + 1) &\leq a \\ \omega + 1 &\leq 2a & \omega &\leq 2a - 1 \end{aligned}$$

So then we have for $\frac{1}{2} \leq a \leq 1$:

$$F_X(a) = P(X \leq a) = P([0, 2a - 1]) = 2a - 1$$

For any $a < \frac{1}{2}$ we know $F_X(a) = 0$, and if $a > 1$ we know $F_X(a) = 1$. Therefore we can see that:

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

We then compute:

$$f_x(a) = \begin{cases} 2 & \text{if } \frac{1}{2} \leq x \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

$$F_X(a) = \int_{-\infty}^{a} f_x(t)\, \mathrm{d}t$$

Example. Recall the dart example, $F_X(a) = a^2$ and $f_x(a) = 2a$.

Remark. Any non-negative integrable function $f : \mathbb{R} \to [0, \infty)$ which satisfies:

$$\int_{-\infty}^{\infty} f(t)\,\mathrm{d}t = 1$$

Is the probability density function of some random variable. Just define:

$$F(a) = \int_{-\infty}^{a} f(t)\,\mathrm{d}t$$

and check this is a cumulative distribution function, and then use the proposition from today.

Example. A very important example:

$$f(t) = \frac{1}{\sqrt{2\pi}} e^{\frac{-t^2}{2}}$$

This the normal probability density function. It is clearly non-negative. Why is its integral equal to one? This is a very beautiful computation.

- Let's go with a trick. We will compute the square, that is if:

  $$\left(\int_{-\infty}^{\infty} f(t)\,\mathrm{d}t\right)^2 = 1$$

  Then we are done:

  $$\begin{aligned} \left(\int_{-\infty}^{\infty} f(t)\,\mathrm{d}t\right)^2 &= \left(\int_{-\infty}^{\infty} f(x)\,\mathrm{d}x\right) \cdot \left(\int_{-\infty}^{\infty} f(y)\,\mathrm{d}y\right) \\ &= \frac{1}{2\pi} \int_{\mathbb{R}^2} e^{-\frac{x^2+y^2}{2}}\,\mathrm{d}x\,\mathrm{d}y \end{aligned}$$

  Great!

- There are two coordinate systems on the plane—cartesian and polar—and we know:

  $$\int_{\mathbb{R}^2} f(x,y)\,\mathrm{d}x\,\mathrm{d}y = \int_0^{2\pi} \int_0^{\infty} f(r\cos(\phi), r\sin(\phi)) \cdot r \cdot \mathrm{d}r\,\mathrm{d}\phi$$

Then we must have:

$$\begin{aligned}
\frac{1}{2\pi}\int_{\mathbb{R}^2} e^{-\frac{x^2+y^2}{2}}\,\mathrm{d}x\,\mathrm{d}y &= \frac{1}{2\pi}\int_0^{2\pi}\int_0^{\infty} e^{-\frac{(r\cos(\phi))^2+(r\sin(\phi))^2}{2}}\cdot r\cdot \mathrm{d}r\,\mathrm{d}\phi \\
&= \frac{1}{2\pi}\int_0^{2\pi}\int_0^{\infty} re^{-\frac{r^2}{2}}\cdot \mathrm{d}r\,\mathrm{d}\phi \\
&= \int_0^{\infty} re^{-\frac{r^2}{2}} = \left[-e^{-\frac{r^2}{2}}\right]_0^{\infty} \\
&= -0-(-1) = 1
\end{aligned}$$

Just as desired! Great!

Definition. *A function $f : I \to \mathbb{R}$ is called absolutely continuous provided that for all $\varepsilon > 0$ there exists a $\delta > 0$ such that for all finite collections of intervals $\{(x_i, y_i)\}$ such that $\sum(y_i - x_i) < \delta$, we have that $\sum |f(y_i) - f(x_i)| < \varepsilon$.*

MATH 525 Notes

Faye Jackson

September 24, 2020

Review: Sets and σ-fields

Definition (Cantor)**.** *A set is a collection of definite and distinct objects of our perception or our mind*

Example. $\{1, 2, 3, 4\}$, $\{1, 2, 3, 4, \ldots\} = \mathbb{N}$, $\{\heartsuit, H, T\}$

If A is a set we write $x \in A$ to say that x is an object in A.

Example. $2 \in \mathbb{N}$, $-3 \notin \mathbb{N}$.

- If A, B are sets. Then $A \subseteq B$ means $x \in A \implies x \in B$.
- If $P(x)$ is a true/false condition for each $x \in A$ we can form the subset $B = \{x \in A \mid P(x)\}$.
- One has $A \cap B$, $A \cup B$, ...
- If A_1, A_2 are subsets of B determined by the conditions P_1 and P_2 then:

$$A_1 \cap A_2 = \{x \in B \mid P_1(x) \text{ and } P_2(x)\}$$
$$A_1 \cup A_2 = \{x \in B \mid P_1(x) \text{ or } P_2(x)\}$$

- If Ω is a set we have the powerset:

$$P(\Omega) = \{B \mid B \subseteq \Omega\}$$

Then for $\Omega = \{H, T\}$ has $P(\Omega) = \{\emptyset, \Omega, \{H\}, \{T\}\}$, note that we have $|P(\Omega)| = 2^{|\Omega|}$

Definition. *Given a set Ω, a σ-field of subsets of Ω is a subset $\mathcal{F} \subseteq P(\Omega)$ such that*

- $\Omega \in \mathcal{F}$
- *If $A_1, A_2, \ldots \in \mathcal{F}$ then $\bigcup_{i \in \mathbb{N}} A_i \in \mathcal{F}$*
- *If $A \in \mathcal{F}$ then $A^c = \Omega \setminus A \in \mathcal{F}$.*

Example. $P(\Omega)$ is always a σ-field

Example. Let $\Omega = \{0, 1, 2\}$. Then the following are σ-fields:

- $\mathcal{F} = P(\Omega)$
- $\mathcal{F} = \{\emptyset, \Omega\}$
- $\mathcal{F} = \{\emptyset, \Omega, \{1\}, \{0, 2\}\}$. (and correspondingly for the other singletons)

Axioms of set theory

Russel's autinomy: Let A be the set $\{B \mid B \notin B\}$. Question, is $A \in A$? Either way you get a contradiction.

There are ten axioms in total for set theory. One of the most controversial is the axiom of choice: If $\{A_i\}$ is a collection of sets, then there exists a set A which contains a unique element from each A_i. This axiom leads to paradoxes like the Banach-Tarski paradox, but no contradictions!

Back to Random Variables

If $X : \Omega \to \mathbb{R}$ is a random variable. We introduced the following shorthands:

$$\{X \leq a\} = \{\omega \in \Omega \mid X(\omega) \leq a\}$$
$$\{X = a\} = \{\omega \in \Omega \mid X(\omega) = a\}$$

Example. Some nice examples

- $\Omega = [0,1]$, and $X : \Omega \to \mathbb{R}$ is $X(\omega) = \omega$. Then:

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

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

- $f_X(t) = \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}$. We can write:

$$F_X(a) = \int_{-\infty}^{a} f_X(t)\,\mathrm{d}t$$

  There is no closed formula for F_X in terms of elementary functions. We use the Greek letter Φ to represent this cumulative distribution function. Now take $\Omega = (0,1)$ and then $X = \Phi^{-1}$.

- Bernoulli variable: Let $\Omega = \{H, T\}$, $P(H) = p$, $P(T) = 1 - p$. And then let $X(H) = 1$ and $X(T) = 0$. Then:

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

Proposition 1. If $f : \mathbb{R} \to \mathbb{R}$ is continuous and $X : \mathbb{R} \to \mathbb{R}$ is a random variable. Then $f(X) = f \circ X$ is a random variable.

Proof. Since f is continuous, preimages of open sets under f are open. Since the Bore σ-field is generated by all open sets, the measurability of X implies the measurability of $f(X)$.

Example. If X is a random variable, then so are X^2, X^3, $\sqrt{X}$, $\ln X$, e^X, $\sin X$, $|X|$, $\frac{1}{X}$, ...

If $f : \mathbb{R} \to \mathbb{R}$ is continuous, then if $U \subseteq \mathbb{R}$ is open, then:

$$f^{-1}(U) = \{x \in \mathbb{R} \mid f(x) \in U\}$$

is open. This is in fact equivalent to continuity. We need $f \circ X$ measurable, which means that some $(f \circ X)^{-1}(U)$ is an event for all U in the Borel σ-field on $\mathbb{R}$. The Borel σ-field is generated by open sets, so lets just check opens U:

$$(f \circ X)^{-1}(U) = X^{-1}(f^{-1}(U))$$

And since $f^{-1}(U)$ is open, it is in the Borel σ-field, and so $X^{-1}(f^{-1}(U))$ is an event because X is a random variable.

A useful skill: Given the probability density function of X, compute the probability density function of $f(X)$. The general strategy for doing this is to work through the cumulative distribution function.

Example. Say X has density $\frac{1}{\sqrt{2\pi}}e^{-x^2/2}$. Compute the density of $Y = X^3$.

$$\begin{aligned} F_Y(a) &= P(Y \leq a) = P(X^3 \leq a) \\ &= P(X \leq \sqrt[3]{a}) = F_X(\sqrt[3]{a}) \\ f_Y(a) &= F_Y'(a) = f_x(\sqrt[3]{a}) \cdot \frac{1}{3}a^{-\frac{2}{3}} \\ &= \frac{1}{\sqrt{2\pi}a^{\frac{2}{3}}}e^{-\frac{a^{\frac{2}{3}}}{2}} \end{aligned}$$

Group Work

1) Start with $\Omega = [0, 1]$, X uniform, and $Y = X^2$. Compute the densities.

2) Start with $\Omega = [-1, 1]$, X uniform, $Y = |X|$. Compute the densities.

Let's go!

$$F_Y(a) = P(Y \le a) = P(X^2 \le a) = \begin{cases} 0 & \text{if } a \le 0 \\ P(X \le \sqrt{a}) & \text{if } 0 \le a \le 1 \\ 1 & \text{if } a \ge 1 \end{cases}$$

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

$$f_Y(a) = F_Y'(a) = \begin{cases} \frac{1}{2\sqrt{a}} & \text{if } 0 < a < 1 \\ 0 & \text{otherwise} \end{cases}$$

Now for the next one. Note first that:

$$f_X(a) = \begin{cases} \frac{1}{2} & \text{if } -1 \le a \le 1 \\ 0 & \text{otherwise} \end{cases}$$

So then

$$F_Y(a) = P(Y \le a) = P(|X| \le a) = \begin{cases} 0 & \text{if } a < 0 \\ P(-a \le X \le a) & \text{if } 0 \le a \le 1 \\ 1 & \text{if } a > 1 \end{cases}$$

Now we compute that:

$$\begin{aligned} P(-a \le X \le a) &= P(X \le a) - P(X \le -a) - P(X = a) \\ &= \frac{1}{2}(a+1) - \frac{1}{2}(-a+1) - 0 \\ &= a \end{aligned}$$

And so

$$F_Y(a) = \begin{cases} 0 & \text{if } a < 0 \\ a & \text{if } 0 \leq a \leq 1 \\ 1 & \text{if } a > 1 \end{cases} \qquad f_Y(a) = F'_Y(a) = \begin{cases} 1 & \text{if } 0 \leq a \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

Great!

MATH 525 Notes

Faye Jackson

September 29, 2020

Expectation

Example. Flip a coin 10,000 times. We expect roughly as many H as T. If X is a Bernoulli Random Variable with $p = \frac{1}{2}$. Then we should let $\mathrm{E}[X] = \frac{1}{2}$

Example. Consider the following game: Choose a number $1 \leq z \leq 6$. Roll three fair die. If z comes up three times, get \$3, if twice \$2, if once \$1, if none, pay \$1. Should you play? There are $6^3 = 216$ outcomes. If we play 216 times times and get each outcome then:

$$1 \cdot 3 + (5 + 5 + 5) \cdot 2 + (25 + 25 + 25) \cdot 1 + 125(-1) = -17$$

We would like $\mathrm{E}[X] = \frac{-17}{216} \approx -0.079$

Definition. *If $(\Omega, \mathcal{F}, P)$ is a probability space and $X : \Omega \to \mathbb{R}$ then we define the expectation of X as:*

$$\mathrm{E}[X] = \int_\Omega X \, \mathrm{d}P \in \mathbb{R}$$

Example. If $X = \mathbb{1}_A$ for some $A \in \mathcal{F}$. Then $\mathrm{E}[\mathbb{1}_A] = P(A)$.

A Special Case

Definition (Sort Of)**.** *If X is a discrete random variable then:*

$$\mathrm{E}[X] = \sum_{a \in \mathbb{R}} a \cdot P(X = a) = \sum_{a \in \mathbb{R}} a \cdot f_X(a)$$

If X is a continuous random variable then:

$$\mathrm{E}[X] = \int_{-\infty}^{\infty} a \cdot f_X(a)\,\mathrm{d}a$$

Properties of Expectation:

- $\mathrm{E}[X+Y] = \mathrm{E}[X] + \mathrm{E}[Y]$
- $\mathrm{E}[cX] = c\,\mathrm{E}[X]$, $c \in \mathbb{R}$.
- $\mathrm{E}[X] \geq 0$ if $X \geq 0$.

Example. If $X = \mathbb{1}_A$ and $Y = \mathbb{1}_B$ then:

$$(X+Y)(\omega) = \begin{cases} 2 & \text{if } \omega \in A \cap B \\ 1 & \text{if } \omega \in A \triangle B \\ 0 & \text{otherwise} \end{cases}$$

So then:

$$\begin{aligned} \mathrm{E}[X+Y] &= 2 \cdot P(A \cap B) + 1 \cdot P(A \triangle B) \\ &= P(A \cap B) + P(A \cap B) + P(A \cap B^c) + P(B \cap A^c) \\ &= P(A) + P(B) = \mathrm{E}[X] + \mathrm{E}[Y] \end{aligned}$$

Great!

Warning: $\mathrm{E}[X]$ may not exist!!!

Example. Let $\Omega = (0,1)$ and let $X(\omega) = \frac{1}{\omega}$. And so:

$$\mathrm{E}[X] = \int_0^1 \frac{1}{x}\,\mathrm{d}x = \infty$$

Which is not truly defined

In the General Setting

Definition. *X is called simple if it takes finitely many values $a_1, a_2, \ldots, a_n$. In that case we define $\mathrm{E}[X]$ to be:*

$$\mathrm{E}[X] = \sum_{i=1}^{n} a_i \cdot P(X = a_i)$$

Definition. *X is called integrable if there exists a sequence (X_n) of simple random variables such that:*

- *(X_n) is Cauchy, i.e. for any $\varepsilon > 0$ there exists an $N \in \mathbb{N}$ so that if $n, m > N$ then for all $\omega \in \Omega$ we have $\mathrm{E}[|X_n - X_m|] < \varepsilon$. Note that this is defined since $|X_n - X_m|$ is simple.*
- *$X_n \to X$ almost surely, that is:*

$$P\left(\left\{\omega \in \Omega \mid \lim_{n\to\infty} X_n(\omega) = X(\omega)\right\}\right) = 1$$

- *In this situaton we define the expectation of X as:*

$$\mathrm{E}[X] = \lim_{n\to\infty} \mathrm{E}[X_n]$$

Theorem. *This limit always exists for such a sequence X_n, and the quantity depends only on X, not on the sequence (X_n).*

Strategy: To compute expectations. Try to write X as a sum of indicators

Example. Recall the problem of letters. We write n personalized letters and send randomly. Let X be the number of people who got their letters. What is $\mathrm{E}[X]$. Write $X = X_1 + \cdots + X_n$ where X_i is 1 if person i got their letter and 0 if not. We know X_i is a Bernoulli variable and:

$$E(X_i) = P(i\text{-th person got letter}) = \frac{1}{n}$$

By linearity we have that:

$$\mathrm{E}[X] = \sum_{i=1}^{n} \mathrm{E}[X_i] = n \cdot \frac{1}{n} = 1$$

Group Work

Let $a_1, \dots, a_n$ be a permutation of $1, \dots, n$. We say that a_k is a record if $a_k > a_i$ for $1 \leq i < k$. By convention, a_1 is always a record. Let X be the number of records. Compute $\mathrm{E}[X]$.

<u>Tasho's Grace</u>: Let X_i be the indicator of the event where a_i is a record.

$$P(a_i \text{ is a record}) = P(a_i \text{ is largest among } a_1, \dots, a_i) = \frac{1}{i}$$

And so:

$$\mathrm{E}[X] = 1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{n} \geq \ln(n)$$

$$\lim_{n \to \infty} \left(\sum_{i=1}^{n} \frac{1}{i} - \ln(n) \right) = \gamma \approx 0.577$$

And so $\mathrm{E}[X] \approx \gamma + \ln(n)$

MATH 525 Notes

Faye Jackson

October 1, 2020

Recall. Given $(\Omega, \mathcal{F}, P)$ a probability space and a random variable $X : \Omega \to \mathbb{R}$ we defined:

$$\mathrm{E}[X] = \int_\Omega X \, \mathrm{d}P$$

And we have two special cases:

- If X is discrete $\mathrm{E}[X] = \sum_{a \in \mathbb{R}} a \cdot P(X = a)$.
- If X is continuous $\mathrm{E}[X] = \int_{-\infty}^{\infty} a \cdot f_X(x) \, \mathrm{d}a$

If X is a random variable and $g : \mathbb{R} \to \mathbb{R}$ is continuous then $g(X)$ is a Random variable

How do we compute $\mathrm{E}[g(X)]$. In principle:

$$\mathrm{E}[g(X)] = \int_{-\infty}^{\infty} a \cdot f_{g(X)}(a) \, \mathrm{d}a$$

Provided that $g(X)$ is continuous.

Theorem (Law of the unconscious statistician)**.** *Two parts:*

a) If X is discrete then the expectation of $g(X)$ is:

$$\mathrm{E}[g(X)] = \sum_{a \in \mathbb{R}} g(x) \cdot f_X(a)$$

b) If X and $g(X)$ are continuous then:

$$\mathrm{E}[g(X)] = \int_{-\infty}^{\infty} g(a) \cdot f_X(a) \, \mathrm{d}a$$

Wow!

Proof. We'll handle both cases:

a) Suppose X is discrete, then:

$$\begin{aligned}
\mathrm{E}[g(X)] &= \sum_{a \in \mathbb{R}} a \cdot f_{g(X)}(a) \\
&= \sum_{a \in \mathbb{R}} a \cdot P(g(X) = a) \\
&= \sum_{a \in \mathbb{R}} a \cdot \sum_{\substack{b \in \mathbb{R} \\ g(b) = a}} P(X = b) \\
&= \sum_{b \in \mathbb{R}} g(b) \cdot P(X = b)
\end{aligned}$$

Great!

b) We'll do a special case. Suppose X is continuous and g is strictly increasing and differentiable. So then:

$$F_{g(X)}(a) = P(g(X) \leq a) = P(X \leq g^{-1}(a)) = F_X(g^{-1}(a))$$

So then:

$$f_{g(X)}(a) = f_X(g^{-1}(a)) \cdot \left(g^{-1}\right)'(a)$$

So then:

$$\mathrm{E}[g(X)] = \int_{-\infty}^{\infty} a \cdot f_{g(X)}(a) \, \mathrm{d}a = \int_{-\infty}^{\infty} a \cdot f_X(g^{-1}(a)) \cdot (g^{-1})'(a) \, \mathrm{d}a$$

But by substitution this is just:

$$\mathrm{E}[g(X)] = \int_{-\infty}^{\infty} g(a) \cdot f_X(a) \, \mathrm{d}a$$

And so we're done!

This completes the proof ☺

Aside for Measure Theory

Consider a measure space $(\Omega_1, \mathcal{F}_1, P_1)$ and $(\Omega_2, \mathcal{F}_2)$. Let $g : \Omega_1 \to \Omega_2$ be a measurable function. Define the pushforward measure $P_2 = g_\star P_1$ by the formula $P_2(B) = P_1(g^{-1}(B))$. For any integrable $f : \Omega_2 \to \mathbb{R}$ we have:

$$\int_{\Omega_2} f(y) \, \mathrm{d}P_2(y) = \int_{\Omega_1} f(g(x)) \, \mathrm{d}P_1(x)$$

This is called the *change of variables theorem*. We get from this:

1) If $X : \Omega \to \mathbb{R}$ a discrete random variable we apply with $g = X$ and $f(x) = x$ and get:

$$\int_{\Omega} X \, \mathrm{d}P = \int_{\mathbb{R}} a \cdot \mathrm{d}X_\star P(a) = \sum_{a \in \mathbb{R}} a \cdot P(X = a)$$

2) For $X : \Omega \to \mathbb{R}$ a continuous random variable we get from absolute continuity of F_X that $\mathrm{d}X_\star P = f_X \, \mathrm{d}a$ So then:

$$\int_{\Omega} X \, \mathrm{d}P = \int_{\mathbb{R}} a \cdot \mathrm{d}X_\star P(a) = \int_{\mathbb{R}} a \cdot f_X(a) \, \mathrm{d}a \tag{1}$$

3) We get unconcious by using the same formula with $X = f$ and $g = g$.

Ok, end the Aside!

Example. Consider the uniform distribution on $[0, 1] = \Omega$ and $X(\omega) = \omega$.

Then:

$$f_X(a) = \begin{cases} 1 & \text{if } 0 \leq a \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

So then:

$$\begin{aligned} \mathrm{E}[X] &= \int_{\mathbb{R}} a \cdot f_X(a) \, \mathrm{d}a = \int_0^1 a \, \mathrm{d}a = \left[\frac{1}{2}a^2\right)_0^1 = \frac{1}{2} \\ \mathrm{E}\left[X^2\right] &= \int_{\mathbb{R}} a^2 \cdot f_X(a) \, \mathrm{d}a = \int_0^1 a^2 \, \mathrm{d}a = \left[\frac{1}{3}a^3\right)_0^1 = \frac{1}{3} \end{aligned}$$

Great!

Conditional Expectation

Definition. *Let $X : \Omega \to \mathbb{R}$ be a random variable and let $B \in \mathcal{F}$ be an event with $P(B) > 0$. We define $\mathrm{E}[X \mid E]$ as follows, both definitions are equivalent:*

1) If X is simple then $\mathrm{E}[X \mid B] = \sum_{a \in \mathbb{R}} a \cdot P(X = a \mid B)$. For general X, take a limit as we did last time.

2) Consider the probability space $(B, \mathcal{F}_B, P_B)$ and if we consider the random variable $X\big|_B : B \to \mathbb{R}$. Then $\mathrm{E}[X \mid B] = \mathrm{E}\left[X\big|_B\right]$.

Theorem. *If $\Omega = \coprod_{i=1}^n B_i$ a disjoint union then:*

$$\mathrm{E}[X] = \sum_{i=1}^n \mathrm{E}[X \mid B_i] \cdot P(B_i)$$

This follows directly from the usual conditioning formula

Example. Let X be the number of rolls of a fair die until the last 6. Condition on the outcome of the first roll, let B_i be the event that the outcome is i. Then:

$$\begin{aligned}
\mathrm{E}[X] &= \sum_{i=1}^{6} \mathrm{E}[X \mid B_i] \cdot P(B_i) \\
&= \mathrm{E}[X \mid B_6] \cdot P(B_6) + \sum_{i=1}^{5} \mathrm{E}[X \mid B_i] \cdot P(B_i) \\
&= \frac{1}{6} + \frac{5}{6} \cdot (\mathrm{E}[X] + 1)
\end{aligned}$$

$$\mathrm{E}[X] = 6$$

Example. Let X be the expected number of tosses of a fair coin until we get two H in a row. Condition on the first two tosses:

$$\begin{aligned}
\mathrm{E}[X] &= \frac{1}{4} \cdot \mathrm{E}[X \mid HH] \cdot \frac{1}{4} \cdot \mathrm{E}[X \mid HT] + \frac{1}{2} \cdot \mathrm{E}[X \mid T] \\
&= \frac{2}{4} + \frac{1}{4}(\mathrm{E}[X] + 2) + \frac{1}{2}(\mathrm{E}[X] + 1) \\
\mathrm{E}[X] &= 6
\end{aligned}$$

Group Work

Same setup but waiting for HT, call that event X. Let Y be the number of turns it takes to get a T. Condition on the first tosse:

$$\begin{aligned}
\mathrm{E}[Y] &= \frac{1}{2}\,\mathrm{E}[Y \mid H] + \frac{1}{2}\,\mathrm{E}[Y \mid T] \\
&= \frac{1}{2}(\mathrm{E}[Y] + 1) + \frac{1}{2} \\
\mathrm{E}[Y] &= 2 \\
\mathrm{E}[X] &= \frac{1}{2} \cdot \mathrm{E}[X \mid H] + \frac{1}{2} \cdot \mathrm{E}[X \mid T] \\
&= \frac{1}{2} \cdot (\mathrm{E}[Y] + 1) + \frac{1}{2}(\mathrm{E}[X] + 1) \\
2\,\mathrm{E}[X] &= 3 + \mathrm{E}[X] + 1 \\
\mathrm{E}[X] &= 4
\end{aligned}$$

Example. Buffon's Needle. Consider parallel lines of distance 1 apart.

Throw down a needle of length 1. What is the probability that the needle intersects a line?

Call ℓ the distance from the midpoint to the closest line. Also let ϕ be the angle between the needle and the lines. We know $0 \leq \ell \leq \frac{1}{2}$ and $0 \leq \phi \leq \pi$. The condition that they intersect is $\ell \leq \frac{1}{2}\sin(\phi)$. Assuming uniform distribution of (ℓ, ϕ) in $[0, 1/2] \times [0, \pi]$ we have:

$$P\left(\ell \leq \frac{1}{2}\sin(\phi)\right) = P\left(\left\{\ell, \phi\right\} \in [0, 1/2] \times [0, \pi] \mid \ell \leq \frac{1}{2}\sin(\phi)\right\}\right)$$

And so we compute the probability with area like this:

$$\frac{2}{\pi}\int_0^\pi \int_0^{\frac{1}{2}\sin(\phi)} 1 \,\mathrm{d}\ell \,\mathrm{d}\phi = \frac{2}{\pi}\int_0^\pi \frac{1}{2}\sin(\phi)\,\mathrm{d}\phi = \frac{1}{\pi}[-\cos]_0^\pi = \frac{2}{\pi}$$

MATH 525 Notes

Faye Jackson

October 6, 2020

Recall. We defined conditional expectation $\mathrm{E}[X \mid B]$ for a random variable $X : \Omega \to \mathbb{R}$ and $B \in \Omega$ where $P(B) > 0$. And we have the conditioning formula:

$$\mathrm{E}[X] = \sum_{i=1}^{n} \mathrm{E}[X \mid B_i] \cdot P(B_i)$$

where $\Omega = \coprod_{i=1}^{n} B_i$ is a disjoint union and $P(B_i) > 0$ for $1 \leq i \leq n$.

Example. A symmetric random walk with absorbing barriers at 0 and n. Let X_k be the duration of the walk starting at k. We want to know $e_k = \mathrm{E}[X_k]$.

We condition on the first step:

$$\begin{aligned} e_k = \mathrm{E}[X_k] &= \frac{1}{2}\mathrm{E}[X_k \mid L] + \frac{1}{2}\mathrm{E}[X_k \mid R] \\ &= \frac{1}{2}(e_{k-1} + 1) + \frac{1}{2}(e_{k+1} + 1) \\ 2e_k &= e_{k-1} + e_{k+1} + 2 \\ e_{n+1} - e_k &= e_k - e_{k-1} - 2 \end{aligned}$$

We have a difference equation with boundary conditions, namely:

$$e_0 = 0 \qquad\qquad e_n = 0$$

Let $x = e_1 - e_0$. Then $e_2 - e_1 = x - 2$, and in general $e_k - e_{k-1} = x - 2(k-1)$:

$$
\begin{aligned}
e_k &= (e_k - e_{k-1}) + (e_{k-1} - e_{k-2}) + \cdots + (e_1 - e_0) + e_0 \\
&= x - 2(k-1) + x - 2(k-2) + \cdots + x + 0 \\
&= kx - 2 \cdot \frac{k(k-1)}{2} = kx - k(k-1) \\
e_n &= nx - n(n-1) = 0 \\
x &= n - 1 \\
e_k &= k \cdot (n-1) - k \cdot (k-1) = k(n-k)
\end{aligned}
$$

With this we have our general formula! Great! Remove the barrier at n by taking the limit as $n \to \infty$, then $e_k \to \infty$. Recall that the probability of being absorbed at 0 goes to 1 as $n \to \infty$. Ponder this!

What does it mean to have $e_k = \infty = \mathrm{E}[X_k] = \sum_{t=1}^{\infty} t \cdot P(X_k = t)$. This tells us that $P(X_k = t)$ need to go to zero faster than $\frac{1}{t^2}$ in order to get something finite, but of course this is not the case.

Independence of Random Variables

Definition. *Consider random variables $X_1, \dots, X_n : \Omega \to \mathbb{R}$. We call them independent provided that for any $a_1, \dots, a_n \in \mathbb{R}$ the events $\{X_1 \leq a_1\}, \dots, \{X_n \leq a_n\}$ are independent.*

Remark. If $X_i = \mathbb{1}_{A_i}$ for some event A_i, then $X_1, \dots, X_n$ are independent if and only if $A_1, \dots, A_n$ are independent.

Remark. A few things to get us started:

- Just as in the case of events, we can generalize this to an arbitrary collection of random variables by requiring independence of every finite subcollection.

- Because the sets $(-\infty, a]$ generate the Borel σ-field. For any Borel sets $A_1, \dots, A_n \subseteq \mathbb{R}$ the events $\{X_1 \in A_1\}, \dots, \{X_n \in A_n\}$ are independent.

- In particular, given $a_1, \dots, a_n \in \mathbb{R}$ the events $\{X_1 = a_1\}, \dots, \{X_n = a_n\}$ are independent. In fact, if $X_1, \dots, X_n$ are discrete, then this is equivalent to their independence.

Group Work

You are selling your horse. You'll accept the first offer above \$K. Assume all offers are independent and have the same distribution. The question is then, how long do you expect to wait? Let p be the probability that any individual offer is above \$K. Let $X : \Omega \to \mathbb{R}$ tell you how long you wait. Let Y_i be the random variable that gives the i-th offer

$$\begin{aligned}
\mathrm{E}[X] &= p\,\mathrm{E}[X \mid Y_1 > \$K] + (1-p)\,\mathrm{E}[X \mid Y_1 \le \$K] \\
&= p + (1-p)(\mathrm{E}[X] + 1) = (1-p)\,\mathrm{E}[X] + 1 \\
p\,\mathrm{E}[X] = 1 &\implies \mathrm{E}[X] = \frac{1}{p}
\end{aligned}$$

Note that this works because, by independence we can see that $\mathrm{E}[X \mid Y_1 \le \$K] = \mathrm{E}[X] + 1$ since we can restart the game, and treat Y_2 as Y_1 and so on.

Making Independent Clones

Let $X : \Omega_1 \to \mathbb{R}$, $Y : \Omega_2 \to \mathbb{R}$ be random variables. We can construct a probability space Ω and random variables $X', Y' : \Omega \to \mathbb{R}$ which are independent and $F_{X'} = F_X$ and $F_{Y'} = F_Y$. Explicitly let $(\Omega, \mathcal{F}, P)$ be the product of $(\Omega_1, \mathcal{F}_1, P_1)$ and $(\Omega_2, \mathcal{F}_2, P_2)$. Define $X'(\omega_1, \omega_2) = X(\omega_1)$ and $Y'(\omega_1, \omega_2) = Y(\omega_2)$. Given $a, b \in \mathbb{R}$ we look at:

$$\begin{aligned}
P(\{X' \le a\} \cap \{Y' \le b\}) &= P(\{(\omega_1, \omega_2) \mid X(\omega_1) \le a\} \cap \{(\omega_1, \omega_2) \mid Y(\omega_2) \le b\}) \\
&= P(\{X \le a\} \times \Omega_2 \cap \Omega_1 \times \{Y \le b\}) \\
&= P(\{X \le a\} \times \{Y \le b\}) \\
&= P_1(X \le a) \cdot P_2(Y \le b) \\
&= P(\{X \le a\} \times \Omega_2) \cdot P(\Omega_1 \times \{Y \le b\}) \\
&= P(X' \le a) \cdot P(Y' \le b)
\end{aligned}$$

Intuition: X and Y are independent if each acts on their own coordinate.

Theorem. *If $X_1, \ldots, X_n : \Omega \to \mathbb{R}$ are independent Random Variables then:*

$$\mathrm{E}[X_1 \cdots X_n] = \mathrm{E}[X_1] \cdots \mathrm{E}[X_n]$$

Proof. There are three parts to the proof:

- Assume $X_i = \mathbb{1}_{A_i}$. Then:

  $$X_1 \cdots X_n = \mathbb{1}_{A_1} \cdots \mathbb{1}_{A_n} = \mathbb{1}_{A_1 \cap \cdots \cap A_n}$$

  So then:

  $$\mathrm{E}[X_1 \cdots X_n] = P(A_1 \cap \cdots A_n) = P(A_1) \cdots P(A_n) = \mathrm{E}[X_1] \cdots \mathrm{E}[X_n]$$

- Assume each X_i is simple. Note that the identity that we are trying to prove is linear in each X_i. This then reduces to the case that each X_i is an indicator.

- For X_i general take $X_i = \lim_{k \to \infty} X_{i,k}$ for $X_{i,k}$ simple and note that the equation we're trying to prove respects limits.

With this we're done.

Joint Distribution

Definition. *Consider a collection $X_1, \ldots, X_n : \Omega \to \mathbb{R}$ of random variables. Their joint distribution is:*

$$\begin{aligned} F_{X_1, \cdots, X_n}(a_1, \ldots, a_n) &= P(\{X_1 \leq a_1\} \cap \cdots \cap \{X_n \leq a_n\}) \\ &= P(X_1 \leq a, X_2 \leq a_2, \ldots, X_n \leq a_n) \end{aligned}$$

Great!

Remark. If $X_1, \ldots, X_n$ are independent then:

$$F_{X_1, \ldots, X_n}(a_1, \ldots, a_n) = F_{X_1}(a_1) \cdots F_{X_n}(a_n)$$

Just by chasing definitions.

Definition. $X_1, \dots, X_n$ *have* *joint density* *function if there exists a function* $f_{X_1,\dots,X_n} : \mathbb{R}^n \to \mathbb{R}$ *non-negative and integrable:*

$$F_{X_1,\dots,X_n}(a_1, \dots, a_n) = \int_{-\infty}^{a_1} \cdots \int_{-\infty}^{a_n} f_{X_1,\dots,X_n}(t_1, \dots, t_n) \, \mathrm{d}t_n \cdots \mathrm{d}t_1$$

Great!

Remark. If $X_1, \dots, X_n$ are idnependent and have densities $f_{X_1}, \dots, f_{X_n}$ then:

$$f_{X_1,\dots,X_n}(t_1, \dots, t_n) = \prod_{i=1}^{n} f_{X_i}(t_i)$$

Great!

MATH 525 Notes

Faye Jackson

October 8, 2020

Recall. Let $X_1, \dots, X_n : \Omega \to \mathbb{R}$ be random variables:

- The Joint Distribution $F_{X_1,\dots,X_n} : \mathbb{R}^n \to [0,1]$ is defined as:

$$F_{X_1,\dots,X_n}(a_1, \dots, a_n) = P(X_1 \leq a_1, \dots, X_n \leq a_n)$$

- A function $f_{X_1,\dots,X_n} : \mathbb{R}^n \to \mathbb{R}$ is called joint density if it is non-negative and integrable and:

$$F_{X_1,\dots,X_n}(a_1, \dots, a_n) = \int_{-\infty}^{a_1} \cdots \int_{-\infty}^{a_n} f_{X_1,\dots,X_n}(t_1, \dots, t_n) \, \mathrm{d}t_n \cdots \mathrm{d}t_1$$

  And it is called a joint mass if it is non-negative, summable, and:

$$F_{X_1,\dots,X_n}(a_1, \dots, a_n) = \sum_{t_1 \in \mathbb{R}} \cdots \sum_{t_n \in \mathbb{R}} f_{X_1,\dots,X_n}(t_1, \dots, t_n)$$

- $X_1, \dots, X_n$ are independent if the following equivalent conditions hold
  - For any Borel subsets $A_1, \dots, A_n \in \mathbb{R}$ the events $\{X_1 \in A_1\}, .., \{X_n \in A_n\}$ are independent
  - For any $a_1, \dots, a_n \in \mathbb{R}$ the events $\{X_1 \leq a_1\}, \dots, \{X_n \leq a_n\}$ are independent
  - The joint distribution function factors as:

$$F_{X_1,\dots,X_n}(a_1, \dots, a_n) = F_{X_1}(a_1) \cdots F_{X_n}(a_n)$$

- If $f_{X_1}, \ldots, f_{X_n}$ are densities or masses for $X_1, \ldots, X_n$ then the product is a joint density or joint mass:

$$f_{X_1,\ldots,X_n}(t_1, \ldots, t_n) = f_{X_1}(t_1) \cdots f_{X_n}(t_n)$$

Fact: If $f_{X_1,\ldots,X_n}$ is a joint density and $A \subseteq \mathbb{R}^n$ is a Borel set, then:

$$P((X_1, \ldots, X_n) \in A) = \int_A f_{X_1,\ldots,X_n}(t_1, \ldots, t_n)\, \mathrm{d}t_1 \cdots \mathrm{d}t_n$$

Proof Illustration for $n = 2$. Consider the following cases:

- $A = (-\infty, a] \times (-\infty, b]$, then by definition we're done.
- Let $A = [c, d] \times [d, b]$ Then:

$$\begin{aligned} A = (-\infty, a] \times (-\infty, b] &\setminus (-\infty, c] \times (-\infty, b] \\ &\setminus (-\infty, a] \times (-\infty, d] \cup (-\infty, c \times (-\infty, d] \end{aligned}$$

  And so this follows from the previous case since the integral will distribute as we wish

- In general, we rasterize and approximate by rectangles.

Remark. Even if each X_i has a density, they may not have a joint density. For example let $X : \Omega \to \mathbb{R}$ be a continuous random variable with density f. Take $Y = X$. Assume for contradiction that the pair (X, Y) has a joint density g. Let $\Delta = \{(x, x) \mid x \in \mathbb{R}\} \subseteq \mathbb{R}^2$ be the diagonal:

$$1 = P(X = Y) = P((X, Y) \in \Delta) = \int_\Delta g(x, y)\, \mathrm{d}x\, \mathrm{d}y = 0$$

Since the integral of any integrable function over the diagonal is zero (the diagonal has measure zero). This is our contradiction!

Remark. If $F_{X_1,\ldots,X_n}$ is sufficiently differentiable, then:

$$f(t_1,\ldots,t_n) = \frac{\partial^n}{\partial a_1,\ldots,\partial a_n} F_{X_1,\ldots,X_n}(t_1,\ldots,t_n)$$

Remark. If $X_1,\ldots,X_n$ have joint density and $g : \mathbb{R}^n \to \mathbb{R}$ is continuous then $g(X_1,\ldots,X_n)$ is a random variable, then:

$$\mathrm{E}[g(X_1,\ldots,X_n)] = \int_{\mathbb{R}^n} g(t_1,\ldots,t_n) f_{X_1,\ldots,X_n}(t_1,\ldots,t_n)\,\mathrm{d}t_1\cdots\mathrm{d}t_n$$

Example. We have:

$$\mathrm{E}[X_1\cdots X_n] = \int_{\mathbb{R}^n} t_1\cdots t_n f_{X_1,\ldots,X_n}(t_1,\ldots,t_n)\,\mathrm{d}t_1\cdots\mathrm{d}t_n$$

Marginal Distribution

Consider two random variables $X, Y : \Omega \to \mathbb{R}$ with the joint distribution $F_{X,Y}(a,b) = P(X \le a, Y \le b)$. Note that:

$$\begin{aligned}
\{X \le a\} &= \bigcup_{b\in\mathbb{R}} \{X \le a, Y \le b\} \\
F_X(a) = P(X \le a) &= \lim_{b\to\infty} P(\{X \le a, Y \le b\}) \\
&= \lim_{b\to\infty} F_{X,Y}(a,b)
\end{aligned}$$

This is what people call marginal distribution. If X and Y are discrete and we have a joint mass then:

$$f_X(a) = \sum_{b\in\mathbb{R}} f_{X,Y}(a,b)$$

And if X, Y have joint density then:

$$f_X(a) = \int_{b\in\mathbb{R}} f_{X,Y}(a,b)\,\mathrm{d}b$$

This is marginal mass and marginal density respectively.

Group Work

Question Given $X, Y : \Omega \to \mathbb{R}$ random variables with $\mathrm{E}[XY] = \mathrm{E}[X]\,\mathrm{E}[Y]$ are X, Y independent?
Consider $\Omega = [-1, 1]$ with uniform measure. Let $X : \Omega \to \mathbb{R}$ have $X(\omega) = \omega$ and let $Y = \Omega \to \mathbb{R}$ be given by $Y = |X|$. Check

i) $\mathrm{E}[XY] \stackrel{?}{=} \mathrm{E}[X]\,\mathrm{E}[Y]$.

ii) Are X and Y independent?

This is not too difficult

i) Consider that:

$$\begin{aligned}
\mathrm{E}[X] &= \int_\Omega X \,\mathrm{d}P = \int_{-1}^{1} \frac{1}{2} X(\omega)\,\mathrm{d}\omega = \int_{-1}^{1} \frac{1}{2}\omega\,\mathrm{d}\omega = 0 \\
\mathrm{E}[Y] &= \frac{1}{2}\int_{-1}^{1} |\omega|\,\mathrm{d}\omega = \frac{1}{2} \\
\mathrm{E}[XY] &= \frac{1}{2}\int_{-1}^{1} \omega\,|\omega|\,\mathrm{d}\omega = 0 = \mathrm{E}[X]\cdot\mathrm{E}[Y]
\end{aligned}$$

ii) Consider since $X(\omega) \le Y(\omega)$ for all $\omega \in \mathbb{R}$:

$$\begin{gathered}
P\left(X \le \frac{1}{2}, Y \le \frac{1}{2}\right) = P\left(Y \le \frac{1}{2}\right) = P\left(\left[-\frac{1}{2}, \frac{1}{2}\right]\right) = \frac{1}{2} \\
P\left(X \le \frac{1}{2}\right) = P\left(\left[-1, \frac{1}{2}\right]\right) = \frac{3}{4} \\
P\left(X \le \frac{1}{2}\right)\cdot P\left(Y \le \frac{1}{2}\right) = \frac{1}{2}\cdot\frac{3}{4} \ne \frac{1}{2} = P\left(X \le \frac{1}{2}, Y \le \frac{1}{2}\right)
\end{gathered}$$

Definition. *Let $X, Y : \Omega \to \mathbb{R}$ be random variables. Define:*

- *The covariance of X, Y, given as:*

$$\operatorname{cov}(X, Y) = \mathrm{E}[(X - \mathrm{E}[X])\cdot(Y - \mathrm{E}[Y])]$$

- *The variance of X, defined as:*

$$\operatorname{var}(X) = \operatorname{cov}(X, X) = \mathrm{E}\left[(X - \mathrm{E}[X])^2\right]$$

Remark. We can unfold these using linearity as:

$$\begin{aligned} \operatorname{cov}(X, Y) &= \mathrm{E}[XY] - \mathrm{E}[X]\,\mathrm{E}[Y] \\ \operatorname{var}(X) &= \mathrm{E}\left[X^2\right] - (\mathrm{E}[X])^2 \end{aligned}$$

Thus $\operatorname{cov}(X, Y) = 0$ is exactly when $\mathrm{E}[XY] = \mathrm{E}[X]\,\mathrm{E}[Y]$, these random variables are called uncorrelated

Remark. What happens when $Z \geq 0$ and $\mathrm{E}[Z] = 0$. Then Z is itself 0 almost surely.

Definition. *$X = Y$ almost surely provided that $P(X = Y) = 1$.*

Remark. Therefore $\operatorname{var}(X) = 0$ if and only if $(X - \mathrm{E}[X])^2 = 0$ almost surely, nthat is $X = \mathrm{E}[X]$ almost surely. So $\operatorname{var}(X)$ measures how “dispersed X is”

MATH 525 Notes

Faye Jackson

October 13, 2020

Recall. $X, Y : \Omega \to \mathbb{R}$ random variables. We write:

$$\begin{aligned}
\operatorname{cov}(X,Y) &= \mathrm{E}[(X - \mathrm{E}[X]) \cdot (Y - \mathrm{E}[Y])] = \mathrm{E}[XY] - \mathrm{E}[X] \cdot \mathrm{E}[Y] \\
\operatorname{var}(X) &= \mathrm{E}\big[(X - \mathrm{E}[X])^2\big] = \mathrm{E}\big[X^2\big] - \mathrm{E}[X]^2 \\
\operatorname{cov}(X,Y) = 0 &\iff \mathrm{E}[XY] = \mathrm{E}[X]\,\mathrm{E}[Y] \impliedby X, Y \text{ independent} \\
\operatorname{var}(X) = 0 &\iff X \text{ is constant almost surely}
\end{aligned}$$

Note that there is squaring in the variable, so the units are not preserved

Definition. *The standard deviation of X is $\sigma(X) = \sqrt{\operatorname{var}(X)}$. The correlation of X and Y is $\operatorname{cor}(X,Y) = \frac{\operatorname{cov}(X,Y)}{\sigma(X)\sigma(Y)}$*

Definition. *Covariance and variance may not exist. A necessary condition is that $\mathrm{E}[X], \mathrm{E}[Y]$ exist. But this is not enough.*

Theorem (Cauchy-Schwartz Inequality)**.** *Let $X, Y \to \mathbb{R}$ be random variables. Then $\mathrm{E}[XY]^2 \leq \mathrm{E}\big[X^2\big] \cdot \mathrm{E}\big[Y^2\big]$. Moreover, equality holds if and only if $aX = bY$ almost surely for some $a, b \in \mathbb{R}$ which are not both zero.*

Proof. If $\mathrm{E}\big[X^2\big] = 0$, then $X = 0$ almost surely, so $XY = 0$ almost surely, and $|\,\mathrm{E}[XY] = 0$. Thus we're done. Of course the same works if we have $\mathrm{E}\big[Y^2\big] = 0$.

Assume that $\mathrm{E}\big[X^2\big]$ and $\mathrm{E}\big[Y^2\big]$ are nonzero. Consider $Z = aX + bY$ for some real numbers $a, b \in \mathbb{R}$. Now:

$$0 \leq \mathrm{E}\big[Z^2\big] = \mathrm{E}\big[a^2X^2 + 2abXY + b^2Y^2\big] = a^2\,\mathrm{E}\big[X^2\big] + 2ab\,\mathrm{E}[XY] + b^2\,\mathrm{E}\big[Y^2\big]$$

Fix $b \neq 0$ and then we have a quadratic equation. We cannot have two roots, so we either have one root or no roots, so looking at the discriminant:

$$\begin{aligned}
4b^2 \operatorname{E}[XY]^2 - 4b^2 \operatorname{E}\left[X^2\right] \operatorname{E}\left[Y^2\right] &\leq 0 \\
4b^2(\operatorname{E}[XY]^2 - \operatorname{E}\left[X^2\right] \operatorname{E}\left[Y^2\right]) &\leq 0 \\
\operatorname{E}[XY]^2 - \operatorname{E}\left[X^2\right] \operatorname{E}\left[Y^2\right] &\leq 0 \\
\operatorname{E}[XY]^2 &\leq \operatorname{E}\left[X^2\right] \operatorname{E}\left[Y^2\right]
\end{aligned}$$

Great! This is what we want, but now we see that we get equality if and onl if there is an $a \in \mathbb{R}$ so that $0 = \operatorname{E}\left[Z^2\right] = \operatorname{E}\left[(aX + bY)^2\right]$. But of course this happens if and only if $aX + bY = 0$ almost surely. Thus $-aX = bY$, and b is nonzero. Perfect ☺

Corrolary. *$|\operatorname{cor}(X, Y)| \leq 1$ and $|\operatorname{cor}(X, Y)| = 1$ if and only if $aX = bY$ almost surely for some $a, b \in \mathbb{R}$ not both zero.*

Proof. Apply the above theorem to $X - \operatorname{E}[X]$ and $Y - \operatorname{E}[Y]$.

Group Work

Let $X, Y : \Omega \to \mathbb{R}$ be discrete random variables with joint mass function:

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

The joint mass is given by:

$y \backslash x$	-1	1
-1	$\frac{1}{3}$	$\frac{1}{6}$
1	$\frac{1}{6}$	$\frac{1}{3}$

$$f_X(-1) = \frac{1}{2} \qquad f_X(1) = \frac{1}{2}$$

$$f_Y(-1) = \frac{1}{2} \qquad f_Y(1) = \frac{1}{2}$$

Therefore:

$$\begin{aligned}
\mathrm{E}[X] = \mathrm{E}[Y] &= -1 \cdot \frac{1}{2} + 1 \cdot \frac{1}{2} = 0 \\
\mathrm{E}\big[X^2\big] = \mathrm{E}\big[Y^2\big] &= (-1)^2 \cdot \frac{1}{2} + 1^2 \cdot \frac{1}{2} = 1 \\
\mathrm{E}[XY] &= \frac{1}{3} - \frac{1}{6} - \frac{1}{6} + \frac{1}{3} = \frac{1}{3} \\
\mathrm{var}(X) &= \mathrm{var}(Y) = \sigma(X) = \sigma(Y) = 1 \\
\mathrm{cov}(X, Y) = \mathrm{cor}(X, Y) &= \frac{1}{3}
\end{aligned}$$

In particular X and Y are not independent, they are correlated.

Midterm Review

(1) Probability space $(\Omega, \mathcal{F}, P)$

- Ω is any set, finite or infinite
- $\mathcal{F}$ is a σ-field of subsets of Ω. Generally when Ω is finite take all subsets, and when Ω is infinite we avoid pathological subsets
- $P : \mathcal{F} \to [0, 1]$ is a probability measure. That is $P(\Omega) = 1$, $P(A^c) = 1 - P(A)$, and $P\left(\coprod_{k=1}^{\infty} A_k\right) = \sum_{k=1}^{\infty} P(A_k)$
- A finite probability space can be symmetric whenever $P(A) = \frac{\#A}{\#\Omega}$, for example a fair coin. It can also be asymmetric, for example a biased coin.
- Inclusion Exclusion Formulas:

$$\begin{aligned}
P(A \cup B) &= P(A) + P(B) - P(A \cap B) \\
P(A \cup B \cup C) &= P(A) + P(B) + P(C) - P(A \cap B) \\
&\quad - P(B \cap C) - P(A \cap C) + P(A \cap B \cap C)
\end{aligned}$$

- Towers. An Ascending tower $A_1 \subseteq A_2 \subseteq \cdots$, then:

$$P\left(\bigcup_{k=1}^{\infty} A_k\right) = \lim_{k\to\infty} P(A_k)$$

And then a descending tower $A_1 \supseteq A_2 \supseteq \cdots$, then:

$$P\left(\bigcap_{k=1}^{\infty} A_k\right) = \lim_{k\to\infty} P(A_k)$$

(2) Conditioning and Independence

- Given $B \in \mathcal{F}$ where $P(B) \neq 0$ we have:

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

- This gives us the conditioning formula if $\Omega = B_1 \sqcup \cdots \sqcup B_n$ where each $P(B_i) > 0$ then:

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

- Two events A and B are independent if and only if $P(A \mid B)$ equals $P(A)$ or in other words $P(A \cap B) = P(A)P(B)$.
- $A_1, \ldots, A_n$ is independent if and only if for any finite subset $P(A_{i_1} \cdots A_{i_k})$ is equal to $P(A_{i_1}) \cdots P(A_{i_k})$

(3) Combinatorics

- Repeated experiments are given by products of probability spaces defined by:

$$\begin{aligned}
(\Omega, \mathcal{F}, P) &= (\Omega_1, \mathcal{F}_1, P_1) \times (\Omega_2, \mathcal{F}_2, P_2) \\
\Omega &= \Omega_1 \times \Omega_2 \\
\mathcal{F} &= \text{smallest } \sigma\text{-field containing } A_1 \times A_2 \text{ for all } A_i \in \mathcal{F}_i \\
P(A_1 \times A_2) &= P_1(A_1) \cdot P_2(A_2)
\end{aligned}$$

- Choosing k out of n with order is $\frac{n!}{(n-k)!}$ and without order is $\frac{n!}{(n-k)!k!} = \binom{n}{k}$.

(4) Random Variables

- $X : \Omega \to \mathbb{R}$ is a random variable if and only if $X : \Omega \to \mathbb{R}$ is measurable, that is $\{X \in A\} \in \mathcal{F}$ for every Borel subset $A \subseteq \mathbb{R}$. This is true if and only if $\{X \leq a\} \in \mathcal{F}$ for every $a \in \mathbb{R}$.
- Indicators, of the form:

$$\mathbb{1}_A(x) = \begin{cases} 1 & \text{if } x \in A \\ 0 & \text{otherwise} \end{cases}$$

- Cumulative distribution functions, mass functions, and density functions:

$$F_X(a) = P(X \leq a) \tag{cdf}$$
$$f_X(a) = P(X = a) \tag{mass, X discrete}$$
$$F_X(a) = \int_{-\infty}^{a} f_X(t)\,\mathrm{d}t \tag{X continuous, f_X pdf}$$

These have characterizing properties

  - $0 \leq F_X \leq 1$, $\lim_{a\to\infty} F_X(a) = 1$, $\lim_{a\to-\infty} F_X(a) = 0$, $a \leq b$ implies $F_X(a) \leq F_X(b)$, $\lim_{b\downarrow a} F_X(b) = F_X(a)$.
  - $f_X \geq 0$, $\sum f_X(a) = 1$ for X discrete
  - $f_X \geq 0$, $\int f_X(a)\,\mathrm{d}a = 1$, X continous.

(5) Expectations

- For $X : \Omega \to \mathbb{R}$ a random variable, then:

$$\mathrm{E}[X] = \int_\Omega X\,\mathrm{d}P$$

This definition can be useful when Ω and P are simple and explicit or Ω is finite. Well we have:

$$\begin{gathered}
\mathrm{E}[\mathbb{1}_A] = P(A) \\
\mathrm{E} \text{ is linear} \\
\mathrm{E}[\lim X_k] = \lim \mathrm{E}[X_k] \text{ when } X_k \text{ is Cauchy}
\end{gathered}$$

We have the formulas:

$$\mathrm{E}[X] = \sum_{x \in \mathbb{R}} x \cdot f_X(x) \qquad (X \text{ discrete})$$

$$\mathrm{E}[X] = \int_{\mathbb{R}} x \cdot f_X(x) \, \mathrm{d}x \qquad (X \text{ continuous})$$

Moreover the law of unconscious statistician says:

$$\mathrm{E}[g(X)] = \sum_{x \in \mathbb{R}} g(x) \cdot f_X(x) \qquad (X \text{ discrete})$$

$$\mathrm{E}[g(X)] = \int_{\mathbb{R}} g(x) \cdot f_X(x) \, \mathrm{d}x \qquad (X \text{ continuous})$$

MATH 525 Notes

Faye Jackson

October 20, 2020

Sum of random variables

Consider two random variables $X, Y : \Omega \to \mathbb{R}$ we know that $Z = X + Y$ is a random variable

Question: If X, Y are continuous is Z also continuous? If so, what is the density of Z

Answer: In general, no

Theorem. *If $X, Y : \Omega \to \mathbb{R}$ have joint density $f_{X,Y}$, then $Z = X + Y$ is continuous and has density:*

$$f_Z(a) = \int_{-\infty}^{\infty} f_{X,Y}(x, a - x)\,\mathrm{d}x = \int_{-\infty}^{\infty} f_{X,Y}(a - y, y)\,\mathrm{d}y$$

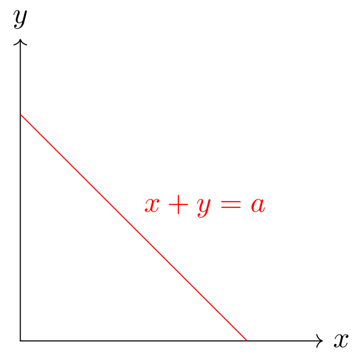
Proof. Recall if $A \subseteq \mathbb{R}^2$ is a borel set then we know:

$$P((X, Y) \in A) = \int_A f_{X,Y}(x, y)\,\mathrm{d}x\,\mathrm{d}y$$

Then if we consider the following A we know:

$$A := \{(x, y) \in \mathbb{R}^2 \mid x + y \leq a\}$$
$$F_Z(a) = P(Z \leq a) = P((X, Y) \in A)$$

There are two ways to integrate over this set, look at the picture



So we get with a u-substitution of $u = y + x$:

$$\begin{aligned} F_Z(a) &= \int_{-\infty}^{\infty} \int_{-\infty}^{a-x} f_{X,Y}(x,y)\,\mathrm{d}y\,\mathrm{d}x \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{a} f_{X,Y}(x,u-x)\,\mathrm{d}y\,\mathrm{d}x \\ &= \int_{-\infty}^{a} \int_{-\infty}^{\infty} f_{X,Y}(x,u-x)\,\mathrm{d}x\,\mathrm{d}u \end{aligned}$$

But then $f_Z(u) = \int_{-\infty}^{\infty} f_{X,Y}(x,u-x)\,\mathrm{d}x$ satisfies the conditions for a joint density.

Special Case

If X, Y are independent then $f_{X,Y}(x,y) = f_X(x) \cdot f_Y(y)$ is a joint density and so:

$$f_Z(a) = \int_{-\infty}^{\infty} f_X(x) \cdot f_Y(a-x)\,\mathrm{d}x = \int_{-\infty}^{\infty} f_X(a-y) \cdot f_Y(y)\,\mathrm{d}y$$

This integral is called the convolution of f_X and f_Y denoted as:

$$\boxed{f_{X+Y} = f_X * f_Y}$$

Discrete Version:

If X, Y are discrete random variables they automatically have joint mass

$f_{X,Y}(x,y) = P(X = x, Y = y)$. One sees as in the above proof that:

$$f_Z(a) = \sum_{x=-\infty}^{\infty} f_{X,Y}(x, a-x) = \sum_{y=-\infty}^{\infty} f_{X,Y}(a-y, y)$$

If X, Y are independent then $f_{X,Y} = f_X \cdot f_Y$ and so:

$$f_Z(a) = \sum_{x=-\infty}^{\infty} f_X(x) f_Y(a-x) = \sum_{y=-\infty}^{\infty} f_X(a-y) f_Y(y)$$

And so of course $f_{X+Y} = f_X * f_Y$.

Group Work

a) Let X, Y be independent continuous random variables with normal distributions, i.e. density given by $\frac{1}{\sqrt{2\pi}} e^{-x^2/2}$. Compute the density of $X + Y$.

We compute this by the theorem as a convolution:

$$\begin{aligned} f_Z(a) &= \int_{-\infty}^{\infty} f_X(x) \cdot f_Y(a-x)\, \mathrm{d}x \\ &= \int_{-\infty}^{\infty} \frac{1}{2\pi} e^{-x^2/2} e^{-(a-x)^2/2}\, \mathrm{d}x \\ &= \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-x^2/2} e^{-a^2/2+ax-x^2/2}\, \mathrm{d}x \\ &= \frac{e^{-a^2/2}}{2\pi} \int_{-\infty}^{\infty} e^{-x^2+ax}\, \mathrm{d}x \\ &= \frac{e^{-a^2/2}}{2\pi} \int_{-\infty}^{\infty} e^{-(x-a/2)^2+a^2/4}\, \mathrm{d}x \\ &= \frac{e^{-a^2/4}}{2\pi} \int_{-\infty}^{\infty} e^{-(x-a/2)^2}\, \mathrm{d}x \\ &= \frac{e^{-a^2/4}}{2\pi} \int_{-\infty}^{\infty} e^{-u^2}\, \mathrm{d}u = \frac{e^{-a^2/4}}{2\pi} \cdot \sqrt{\pi} = \frac{e^{-a^2/4}}{2\sqrt{\pi}} \end{aligned}$$

b) Let $X_1, \ldots, X_n$ be independent discrete random variables with bernoulli distribution $P(X_i = 1) = p, P(X_i = 0) = 1 - p$. Compute the mass of $X_1 + X_2 + \cdots + X_n$.

First we'll consider when $n = 2$, then:

$$\begin{aligned} f_{X_1+X_2}(a) &= \sum_{x=-\infty}^{\infty} P(X_1 = x)P(X_2 = a - x) \\ &= P(X_1 = 1)P(X_2 = a-1) \cdot P(X_1 = 0)P(X_2 = a) \\ &= \begin{cases} p^2 & \text{if } a = 2 \\ 2p(1-p) & \text{if } a = 1 \\ (1-p)^2 & \text{if } a = 0 \\ 0 & \text{otherwise} \end{cases} \end{aligned}$$

Guess for $Z_n = x_1 + \cdots + X_n$:

$$f_{Z_n}(k) = \binom{n}{k} p^k (1-p)^{n-k}$$

We can prove this by induction.

Definition. *A discrete random variable is called binomial with parameters n, p if $P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$.*

Example. If a coin is tossed n times and X is the number of heads, then X is binomial.

Example. Consider k fixed, $n \to \infty$, $p \to 0$ so that np remains fixed. Set $\lambda = np$. Then:

$$\begin{aligned} \binom{n}{k} p^k (1-p)^{n-k} &= \binom{n}{k} \left(\frac{\lambda}{n}\right)^k \left(1 - \frac{\lambda}{n}\right)^{n-k} \\ &= \frac{n(n-1)\cdots(n-k+1)}{k!} \cdot \frac{\lambda^k}{n^k} \cdot \left(1 - \frac{\lambda}{n}\right)^{n-k} \\ &= \frac{n(n-1)\cdots(n-k+1)}{n^k} \cdot \frac{\lambda^k}{k!} \cdot \left(1 - \frac{\lambda}{n}\right)^{n-k} \end{aligned}$$

But then:

$$\frac{n}{n} \cdot \frac{n-1}{n} \cdots \frac{n-k+1}{n} \to 1$$
$$\frac{\lambda^k}{k!} \to \frac{\lambda^k}{k!}$$
$$\left(1 - \frac{\lambda}{n}\right)^{n-k} \to e^{-\lambda}$$

And so:

$$\binom{n}{k} p^k (1-p)^{n-k} \to \frac{\lambda^k}{k!} e^{-\lambda}$$

In fact we can calculate that:

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

Definition. *A discrete random variable X is called Poisson for $\lambda > 0$ if $P(X = k) = \frac{\lambda^k}{k!} e^{-\lambda}$ for $k \geq 0$ and zero for $k < 0$.*

These random variables describe the number of successes in great number of trials with very small success probability

Example. The number of fender benders on Washtenaw avenue on a given day. A great number of cars go by, and the probability of accident for a given car is very small. Say the expectation is two, then:

- The probability of three accidents is $\frac{2^3}{3!} e^3 = 0.18$
- Probability of at least one accident is $1 - e^{-2} \approx 0.8$

MATH 525 Notes

Faye Jackson

October 22, 2020

Recall. We have the following:

- For X a Bernoulli Random Variable with parameter P: $P(X = 1) = p$, $P(X = 0) = 1 - p$
- X a Binomial Random variable with parameters n, p:

$$P(X = k) = \begin{cases} \binom{n}{k} p^k (1-p)^{n-k} & \text{if } 0 \leq k \leq n \\ 0 & \text{otherwise} \end{cases}$$

  So $X = X_1 + \cdots + X_n$ where X_i are independent Bernoulli with parameter p
- X is a Poisson Random Variable with parameter $\lambda > 0$:

$$P(X = k) = \begin{cases} \frac{\lambda^k}{k!} e^{-\lambda} & \text{if } k \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

  Obtained by taking the limit $n \to \infty$ of this binomial mass function with parameter, $(\lambda/n, n)$.

Poisson Process

Consider a call center. For a given time interval $[t_1, t_2]$ we have some probability of a call, or multiple calls, arriving. Set:

$$X_{[t_1,t_2]} := \#\text{ of calls received in the time interval } [t_1, t_2]$$

The key modeling assumption is that if the time interval is very small, i.e. $[t, t + \Delta t]$, then:

a) The probability of one call arriving is $\lambda \cdot \Delta t$

b) The probability of more than one call arriving is zero.

Reasonable: No simulataneous calls
Consequence: $X_{[t_1,t_2]}$ is the sum of $N = \frac{t_2 - t_1}{\Delta t}$ independent trials with success probability $\lambda \cdot \Delta t = p$. Thus, $X_{[t_1,t_2]}$ is Poisson with parameter $\lambda(t_2 - t_1)$. Also:

- $X_{[t_1,t_3]} = X_{[t_1,t_2]} + X_{[t_2,t_3]}$ for $t_1 < t_2 < t_3$
- $X_{[t_1,t_2]}$ and $X_{[t_3,t_4]}$ are independent for $t_1 < t_2 < t_3 < t_4$.

Definition. *A Poisson process with intensity λ is a family of Random Variables $X_{[s,t]} : \Omega \to \mathbb{R}$ indexed by $s < t \in \mathbb{R}$ sucht hat:*

a) $X_{[s,t]}$ is Poisson with parameter $\lambda(t - s)$.

b) $X_{[r,t]} = X_{[r,s]} + X_{[s,t]}$ for $r < s < t$.

c) If $[s_i, t_i]$ are finitely many pairwise disjoint intervals then the collection of random variables $\{X_{[s_i,t_i]}\}$ is independent.

Consider a Poisson process of intensity λ and ask: How long do we have to wait for the first arrival? Let Z be the time before first arrival:

$$F_Z(t) = P(Z \leq t) = 1 - P(Z > t)$$

$$= 1 - P(X_{[0,t]} = 0)$$
$$= 1 - e^{-\lambda t}$$

Note that F_Z is differentiable, so Z is a continuous Random Variable with density:

$$f_Z(t) = F'_Z(t) = \begin{cases} \lambda e^{-\lambda t} & \text{if } t > 0 \\ 0 & \text{otherwise} \end{cases}$$

Definition. *A continuous random variable Z is called exponential with parameter $\lambda > 0$ if its density is:*

$$f_Z(t) = \begin{cases} \lambda e^{-\lambda t} & \textit{if } t > 0 \\ 0 & \textit{otherwise} \end{cases}$$

Remark. And then:

- f_Z is indeed a density function:

$$\int_0^\infty \lambda e^{-\lambda t}\,\mathrm{d}t = \left(-e^{-\lambda t}\right]_0^\infty = 1$$

- We can calculate the expectation:

$$\begin{aligned} \mathrm{E}[Z] &= \int_0^\infty t \cdot \lambda e^{-\lambda t}\,\mathrm{d}t = \int_0^\infty u \cdot e^{-u} \frac{\mathrm{d}u}{\lambda} \\ &= \frac{1}{\lambda}\left([-ue^{-u}]_0^\infty + \int_0^\infty e^{-u}\,\mathrm{d}u \right) = \frac{1}{\lambda}\int_0^\infty e^{-u}\,\mathrm{d}u \\ &= \frac{1}{\lambda}\left(-e^{-u}\right]_0^\infty = \frac{1}{\lambda} \end{aligned}$$

- As we have seen, exponential Random Variables model waiting time between unpredictable events. We can tie this back to Bernoulli trials as follows. Say we perform independent trials with success probability p times $\Delta t, 2\Delta t, 3\Delta t$. Let Z be the time before first success:

$$P(Z > k \cdot \Delta t) = (1-p)^k$$

Now fix a time t, then in the interval of time $[0, t]$ there are roughly $k = \frac{t}{\Delta t}$ many trials. Let $\Delta t \to 0$ and $p = \lambda \cdot \Delta t$. Then:

$$\begin{aligned} P(Z > t) &= \lim_{\Delta t \to 0} (1 - \lambda \Delta t)^{\frac{t}{\Delta t}} \\ &= \lim_{n \to \infty} \left(1 - \frac{\lambda}{n}\right)^{\frac{t}{n}} = e^{-\lambda t} \end{aligned}$$

Group Work

We upgrade our modeling of fender benders on Washtenaw to use the Poisson process. We expect 2 accidents per 24 hours. Two Questions:

a) What is the probability that we must wait more than 12 hours for the first accident

We model with the Poisson process, so we have a random variable $X_{[0,24]}$ where λ is the intensity and $X_{[0,24]}$ is Poisson with parameter 24λ. Then:

$$\mathrm{E}\big[X_{[0,24]}\big] = 24\lambda = 2$$
$$\lambda = \frac{1}{12}$$

Let Z be the waiting random variable then:

$$P(Z > 12) = P(X_{[0,12]=0}) = e^{-\lambda 12} = e^{-1}$$

And so we're done with this part!

b) How long do we have to wait?

We know that the expectation of the waiting variable is $\frac{1}{\lambda} = 12$ from previous calculations.

Memoryless property

Lemma. *Ley Z be an exponential random variable with parameter λ. Then for any $a, b \geq 0$:*

$$P(Z > a + b \mid Z > a) = P(Z > b)$$

Proof. Well we have:

$$\begin{aligned} P(Z > a + b \mid Z > a) &= \frac{P(Z > a + b, Z > a)}{P(Z > a)} = \frac{P(Z > a + b)}{P(Z > a)} \\ &= \frac{e^{-\lambda(a+b)}}{e^{-\lambda a}} = e^{-\lambda b} = P(Z > b) \end{aligned}$$

The interpretation is that: How long you'll have to wait is not dependent on how long you have been waiting.

Lemma. *Let X, Y be independent exponential Random Variables with parameters λ, μ then $Z = \min(X, Y)$ is exponential with parameter $\lambda + \mu$.*

Proof. Well consider that:

$$\begin{aligned} F_Z(t) &= P(Z \geq t) = 1 - P(Z > t) \\ &= 1 - P(X > t, Y > t) = 1 - P(X > t)P(Y > t) \\ &= 1 - e^{-\lambda t}e^{-\mu t} = 1 - e^{-(\lambda+\mu)t} \end{aligned}$$

And so we're done!

Example. The expected number of crashes in 24 hours is 2 on Washtenaw and 4 on Main, and these are independent.

a) How long do we have to wait before an accident on either street?

> Of course our Wasthenaw parameter is $\lambda = \frac{1}{12}$ and our Main parameter is $\mu = \frac{1}{6}$. Thus $Z = \min(X, Y)$ is exponential with parameter $\lambda + \mu = \frac{1}{4}$. Thus $\mathrm{E}[Z] = 4$.

b) What is the probability that the first accident occurs on main

> Well we wish to find:
>
> $$P(X > Y) = P((X, Y) \in A)$$

where A is the graph below $y = x$. So then:

$$\begin{aligned}
P(X > Y) &= \int_A f_X(x) f_Y(y)\, \mathrm{d}x\, \mathrm{d}y \\
&= \int_0^\infty f_Y(y) \int_y^\infty f_X(x)\, \mathrm{d}x\, \mathrm{d}y \\
&= \int_0^\infty \frac{1}{6} e^{-\frac{1}{6}y} \int_y^\infty \frac{1}{12} e^{-\frac{1}{12}x}\, \mathrm{d}x\, \mathrm{d}y \\
&= \int_0^\infty \frac{1}{6} e^{-\frac{1}{6}y} e^{-\frac{1}{12}y}\, \mathrm{d}y \\
&= \frac{1}{6} \int_0^\infty e^{-\frac{1}{4}y}\, \mathrm{d}y = \frac{2}{3}
\end{aligned}$$

MATH 525 Notes

Faye Jackson

October 27, 2020

Conditioning on Random Variable

Recall. Given an event B with $P(B) > 0$ we have:

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

And likewise $\mathrm{E}[X \mid B]$. Now also if Ω is the disjoint of $B_1, \ldots, B_n$ with $P(B_i) > 0$ we have:

$$P(A) = \sum_{i=1}^{n} P(A \mid B_i)P(B_i) \qquad \mathrm{E}[X] = \sum_{i=1}^{n} \mathrm{E}[X \mid B_i]P(B_i)$$

Special Case: Discrete Y

Definition. *Consider two random variables $X, Y : \Omega \to \mathbb{R}$. First assume that Y is discrete*

- *Then define the conditional distribution $F_{X|Y}(x \mid y) = P(X \leq x \mid Y = y)$ defined for all $y \in \mathbb{R}$ such that $P(Y = y) > 0$.*
- *If X is also discrete we define the conditional mass function $f_{X|Y}(x \mid y) = P(X = x \mid Y = y)$ when $P(Y = y) > 0$.*
- *The conditional expectation $\mathrm{E}[X \mid Y]$ is the function:*

$$y \xmapsto{\mathrm{E}[X|Y]} \mathrm{E}[X \mid Y = y]$$

This function is defined for $y \in \mathbb{R}$ where $P(Y = y) > 0$. Extend this to a function on all of $\mathbb{R}$ by zero. Thus $\mathrm{E}[X \mid Y] : \mathbb{R} \to \mathbb{R}$.

Remark. If X is discrete, then:

$$\mathrm{E}[X \mid Y = y] = \sum_{x \in \mathbb{R}} x \cdot f_{X|Y}(x \mid y)$$

Which is nice

Remark. $\mathrm{E}[X \mid Y]$ is a random variable on the probability space $(\Omega_Y, \mathcal{F}_Y, P_Y)$. We take $\Omega_Y = \mathbb{R}$, $\mathcal{F}$ as the Borel σ-field, and $P_Y = Y_* P$. Since Y is discrete:

$$P_Y(A) = \sum_{a \in A} P(Y = a)$$

Proposition 1. $\mathrm{E}[\mathrm{E}[X \mid Y]] = \mathrm{E}[X]$.

Recall. We can calculate expectation in two ways:

$$\mathrm{E}[Z] = \int_\Omega Z \,\mathrm{d}P = \begin{cases} \int_{\mathbb{R}} z \cdot f_Z(z) \,\mathrm{d}x & \text{if } Z \text{ is continuous} \\ \sum_{x \in \mathbb{R}} f_Z(z) & \text{if } Z \text{ is discrete} \end{cases}$$

Proof. We calculate using the first method:

$$\begin{aligned} \mathrm{E}[\mathrm{E}[X \mid Y]] &= \int_\Omega \mathrm{E}[X \mid Y] \,\mathrm{d}P = \sum_{y \in \mathbb{R}} \mathrm{E}[X \mid Y = y] \,\mathrm{d}P(y) \\ &= \sum_{y \in \mathbb{R}} \mathrm{E}[X \mid Y = y] P(Y = y) = \mathrm{E}[X] \end{aligned}$$

Because Ω is the disjoint union of the events $Y = y$ for $y \in \mathbb{R}$.

Group Work

A hen lays N eggs, where N is Poisson with parameter λ. Each egg hatches with probability p independently of the other eggs. Let K be the number of chicks. Find $\mathrm{E}[K]$.

Consider for any $n \in \mathbb{N}$ the random variables $K_1, \dots, K_n$ where K_i tells us that the i-th chick hatched:

$$f_{K|N}(k \mid n) = P(K = k \mid N = n) = \binom{n}{k} p^k (1-p)^{n-k}$$

$$\mathrm{E}[K \mid N = n] = \sum_{k=0}^{\infty} k \cdot f_{K|N}(k \mid n) = np$$

But then:

$$\begin{aligned}
\mathrm{E}[K] = \mathrm{E}[\mathrm{E}[K \mid N]] &= \sum_{n=0}^{\infty} \mathrm{E}[K \mid N = n] \cdot P(N = n) \\
&= \sum_{n=0}^{\infty} np \cdot P(N = n) = p \sum_{n=0}^{\infty} n \cdot P(N = n) \\
&= p \, \mathrm{E}[N] = p\lambda
\end{aligned}$$

Special Case: Continuous Random Variables

Definition. *Suppose that X and Y are continuous with joint density. Then define:*

- *The conditional desnity:*

  $$f_{X|Y}(x \mid y) = \frac{f_{X,Y}(x,y)}{f_Y(y)}$$

  This is of course defined for all $y \in \mathbb{R}$ for which $f_Y(y) > 0$.

- *The conditional distribution:*

  $$F_{X|Y}(x \mid y) = \int_{-\infty}^{x} f_{X|Y}(z, y) \, \mathrm{d}z$$

  We want to have $F_{X|Y} =$ "$P(X \leq x \mid Y = y)$", but this is nonsense since $P(Y = y) = 0$ given that Y is continuous. So consider for very *small Δy:*

$$\begin{aligned}
P(X \leq x \mid y \leq Y \leq y + \Delta y) &= \frac{P(X \leq x, y \leq Y \leq y + \Delta y)}{P(y \leq Y \leq y + \Delta y)} \\
&= \frac{\int_{-\infty}^{x} \int_{y}^{y+\Delta y} f_{X,Y}(s,t)\,\mathrm{d}t\,\mathrm{d}s}{\int_{y}^{y+\Delta y} f_Y(t)\,\mathrm{d}t} \\
&= \frac{\int_{-\infty}^{x} \frac{1}{\Delta y} \int_{y}^{y+\Delta y} f_{X,Y}(s,t)\,\mathrm{d}t\,\mathrm{d}s}{\frac{1}{\Delta y} \int_{y}^{y+\Delta y} f_Y(t)\,\mathrm{d}t}
\end{aligned}$$

Now as $\Delta y \to 0$ we have:

$$\begin{aligned}
\frac{1}{\Delta y} \int_{y}^{y+\Delta y} f_{X,Y}(s,t)\,\mathrm{d}t &\to f_{X,Y}(s,y) \\
\frac{1}{\Delta y} \int_{y}^{y+\Delta y} f_Y(t)\,\mathrm{d}t &\to f_Y(y)
\end{aligned}$$

And so we have that:

$$\begin{aligned}
P(X \leq x \mid y \leq Y \leq y + \Delta y) &\to \frac{\int_{-\infty}^{x} f_{X,Y}(s,y)\,\mathrm{d}s}{f_Y(y)} \\
&= \int_{-\infty}^{x} f_{X|Y}(s \mid y)\,\mathrm{d}s
\end{aligned}$$

- *Now to define the conditional expectation we take* $\mathrm{E}[X \mid Y] : \mathbb{R} \to \mathbb{R}$ *that is defined as* 0 *when* $f_Y(y) = 0$ *and otherwise is defined as:*

  $$y \xmapsto{\mathrm{E}[X|Y]} \mathrm{E}[X \mid Y = y] = \int_{-\infty}^{\infty} x f_{X|Y}(x \mid y)\,\mathrm{d}x$$

  This is a random variable on $(\Omega_Y, \mathcal{F}_Y, P_Y)$*. And we also get that* $\mathrm{d}P_Y(y) = f_Y(y)\,\mathrm{d}y$

Proposition 2. $\mathrm{E}[\mathrm{E}[X \mid Y]] = \mathrm{E}[X]$

Proof. We have that:

$$\begin{aligned}
\mathrm{E}[\mathrm{E}[X \mid Y]] &= \int_{\Omega_Y} \mathrm{E}[X \mid Y = y] \, \mathrm{d}P_Y(y) \\
&= \int_{\Omega_Y} \left(\int_{-\infty}^{\infty} x \cdot f_{X|Y}(x \mid y) \, \mathrm{d}x \right) \mathrm{d}P_Y(y) \\
&= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x f_{X|Y}(x \mid y) f_Y(y) \, \mathrm{d}y \, \mathrm{d}x \\
&= \int_{-\infty}^{\infty} x \int_{-\infty}^{\infty} f_{X,Y}(x, y) \, \mathrm{d}y \, \mathrm{d}x \\
&= \int_{-\infty}^{\infty} x f_X(x) \, \mathrm{d}x = \mathrm{E}[X]
\end{aligned}$$

Just as desired

Group Work 2: Electric Boogaloo

Choose a point Y uniformly at random from $[0, 1]$. Choose a point X uniformly at random from $[0, Y]$. Compute $\mathrm{E}[X], f_X$.

Note that:

$$\mathrm{E}[X \mid Y = y] = \int_0^y \frac{x}{y} \, \mathrm{d}x = \frac{y}{2}$$

$$\begin{aligned}
\mathrm{E}[X] = \mathrm{E}[\mathrm{E}[X \mid Y]] &= \int_0^1 \mathrm{E}[X \mid Y = y] \, \mathrm{d}y \\
&= \int_0^1 \frac{y}{2} \, \mathrm{d}y = \frac{1}{4}
\end{aligned}$$

Looking for the density f_X we first see:

$$f_Y(y) = \begin{cases} 1 & \text{if } 0 \le y \le 1 \\ 0 & \text{otherwise} \end{cases} \qquad f_{X|Y}(x \mid y) = \begin{cases} \frac{1}{y} & \text{if } 0 \le x \le y \\ 0 & \text{otherwise} \end{cases}$$

Then for $0 < x \leq 1$

$$\begin{aligned} f_X(x) &= \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,\mathrm{d}y \\ &= \int_{-\infty}^{\infty} f_{X|Y}(x \mid y) \cdot f_Y(y) \\ &= \int_x^1 \frac{1}{y}\,\mathrm{d}y = \ln(1) - \ln(x) = -\ln(x) \end{aligned}$$

And otherwise, $f_X(x) = 0$.

MATH 525 Notes

Faye Jackson

October 29, 2020

Generating Functions

Definition. *Let X be a discrete Random Variable taking values in $\mathbb{Z}_{\geq 0}$. Define the probability generating function as the formal sum:*

$$G_X(s) = \sum_{n=0}^{\infty} P(X = n) \cdot s^n$$

Great! We can now do analysis to this!

Remark. Let's do some analysis!

- This power series converges for $|s| \leq 1$, and gives an analytic function. Clearly if $s = 1$ then the sum of $P(X = n)$ over all n must be 1 since X takes values in $\mathbb{Z}_{\geq 0}$. This gives convergence at $s = -1$ also. Then for $0 \leq s < 1$ we know $P(X = n) \cdot s^n \leq s^n$, and so we get convergence by the geometric series test. Great!!!

- We have an alternative expression:

  $$G_X(s) = \sum_{n=0}^{\infty} f_X(n) \cdot s^n = \mathrm{E}\big[s^X\big]$$

  By applying the law of the unconscious statistician.

- We know that G_X encodes a lot of information about X:

  - $G_X(0) = P(X = 0)$.

- $G_X(1) = 1$.
- $G'_X(1) = \sum_{n=1}^{\infty} n \cdot P(X = n) \cdot 1^{n-1} = \sum_{n=1}^{\infty} n \cdot P(X = n) = \mathrm{E}[X]$.

Example. If X is Poisson with parameter λ, then:

$$G_X(s) = \sum_{n=0}^{\infty} \frac{\lambda^n}{n!} e^{-\lambda} = e^{-\lambda} \sum_{n=0}^{\infty} \frac{(\lambda - s)^n}{n!} = e^{-\lambda} \cdot e^{\lambda - s} = e^{\lambda(s-1)}$$

Proposition 1. If $X_1, \ldots, X_n$ are independent and $X = X_1 \cdots X_n$ then $G_X(s) = G_{X_1}(s) \cdots G_{X_n}(s)$.

Proof. We use induction! For $n = 1$ it is trivial. So assume the result holds for n and fix $n+1$ independent variables $X_1, \ldots, X_{n+1}$.. Let $Y = x_2 \cdots X_{n+1}$. Then cosnider that:

$$\begin{aligned}
G_X(s) &= \sum_{n=0}^{\infty} P(X = n) s^n \\
&= \sum_{\ell=0}^{\infty} \sum_{m=0}^{\infty} P(X_1 = \ell, Y = m) s^{\ell m} \\
&= \sum_{\ell=0}^{\infty} \sum_{m=0}^{\infty} P(X_1 = \ell) P(Y = m) s^{\ell} s^m \\
&= \sum_{\ell=0}^{\infty} P(X_1 = \ell) s^{\ell} \cdot \sum_{m=0}^{\infty} P(Y = m) s^m \\
&= G_{X_1}(s) \cdot G_Y(s) = G_{X_1}(s) \cdot G_{X_2}(s) \cdots G_{X_{n+1}}(s)
\end{aligned}$$

Where we use the inductive hypothesis in the last line!

Example. Say that $X_1, \ldots, X_n$ are independent Bernoulli random variables with $X = X_1 + \cdots + X_n$. Then:

$$\begin{aligned}
G_{X_i}(s) &= (1 - p) + p \cdot s = 1 + p(s - 1) \\
G_X(s) &= ((1 - p) + ps)^n = \sum_{k=0}^{n} \underbrace{\binom{n}{k} p^k (1 - p)^{n-k}}_{P(X=k)} s^k
\end{aligned}$$

This is yet another method to prove the formula for a binomial random variable's density.

Let Y be the waiting time until the first success and condition on the first trial. We write S for success and F for failure:

$$\begin{aligned} G_Y(s) &= \mathrm{E}\big[s^Y\big] = \mathrm{E}\big[s^Y \mid S\big]p + \mathrm{E}\big[s^Y \mid F\big](1-p) \\ &= p \cdot s + (1-p)\,\mathrm{E}\big[s^{Y+1}\big] \\ &= p \cdot s + s(1-p)\,\mathrm{E}\big[s^Y\big] \\ &= p \cdot s + s(1-p)G_Y(s) \end{aligned}$$

We can solve for $G_Y(s)$ to get:

$$G_Y(s) = \frac{ps}{1 - s(1-p)}$$

To turn this into a power series we use that:

$$\begin{aligned} \sum_{k=0}^{\infty} x^k &= \frac{1}{1-x} \\ G_Y(s) &= ps \cdot \frac{1}{1-s(1-p)} \\ &= ps \sum_{k=0}^{\infty} (s(1-p))^k \\ &= \sum_{k=0}^{\infty} p(1-p)^k \cdot s^{k+1} \end{aligned}$$

Note that $P(Y = k)$ is the coefficient of s^k in $G_Y(s)$ and so $P(Y = k) = p(1-p)^{k-1}$ as expected

0.1 Back to the Random Walk

Consider a symmetric random walk with absorbing barries at $x = 0$ and $x = 5$. Let X_k be the duration of the walk starting at k, where k can be one of $0, 1, 2, 3, 4, 5$. We want to compute $P(X_2) = n$ for any n. We will compute $G_{X_k}(s) = G_k(s)$. Observe that $G_0(s) = G_5(s) = 1$; and by symmetry $G_1(s) = G_4(s)$ and $G_2(s) = G_3(s)$. To compute G_k for $0 < k < 5$ and $n > 0$, we condition on the first step:

$$\begin{aligned}
P(X_k = n) &= \frac{1}{2}P(X_k = n \mid L) + \frac{1}{2}P(X_k = n \mid R) \\
&= \frac{1}{2}P(X_{k-1} = n-1) + \frac{1}{2}P(X_{k+1} = n-1)
\end{aligned}$$

Then since $P(X_k = 0) = 0$ when $0 < k < 5$ we know:

$$\begin{aligned}
G_k(s) &= \sum_{n=1}^{\infty} P(X_k = n) \cdot s^n \\
&= \frac{1}{2}\sum_{n=1}^{\infty} \left(P(X_{k-1} = n-1 + P(X_{k+1} = n-1)\right) s^n \\
&= \frac{s}{2}\left(\sum_{n=0}^{\infty} P(X_{k-1} = n-1)s^{n-1} + \sum_{n=0}^{\infty} P(X_{k+1} = n-1)s^{n-1}\right) \\
&= \frac{s}{2}(G_{k-1}(s) + G_{k+1}(s))
\end{aligned}$$

We know that:

$$\begin{aligned}
G_1(s) &= \frac{s}{2}(G_0(s) + G_2(s) = \frac{s}{2} + \frac{s}{2}G_2(s) \\
G_2(s) &= \frac{s}{2}(G_1(s) + G_3(s)) = \frac{s}{2}(G_1(s) + G_2(s)) \\
&= \frac{s^2}{4} + \frac{s^2}{4}G_2(s) + \frac{s}{2}G_2(s) \\
G_2(s) &= \frac{s^2}{4 - 2s - s^2}
\end{aligned}$$

We want to expand this into a power series. We use partial fractions and geometric series:

$$\frac{s^2}{4 - 2s - s^2} = -1 + \frac{2s - 4}{s^2 + 2s - 4}$$

We factorize $s^2 + 2s - 4$ as $(s - \alpha)(s - \beta)$ where:

$$\alpha = -1 + \sqrt{s} \qquad\qquad \beta = -1 - \sqrt{s}$$

We now set up partial fractions as:

$$\frac{2s-4}{s^2+2s-4} = \frac{A}{s-\alpha} + \frac{B}{s-\beta}$$

We get the equations $A+B=2$ and and $A\beta + B\alpha = 4$. From which we get that:

$$\alpha = 1 - \frac{3}{\sqrt{5}} \qquad\qquad B = 1 + \frac{3}{\sqrt{5}}$$

We then apply the geometric series:

$$\begin{aligned}
\frac{A}{s-\alpha} &= -\frac{A}{\alpha} \cdot \frac{1}{1-\frac{s}{\alpha}} = -\frac{A}{\alpha} \cdot \sum_{n=0}^{\infty} \frac{s^n}{\alpha^n} \\
\frac{B}{s-\beta} &= -\frac{B}{\beta} \cdot \frac{1}{1-\frac{s}{\beta}} = -\frac{B}{\beta} \cdot \sum_{n=0}^{\infty} \frac{s^n}{\beta^n}
\end{aligned}$$

We put this all together to get:

$$\begin{aligned}
G_2(s) &= -1 + \frac{\alpha}{s-\alpha} + \frac{B}{s-\beta} \\
&= -1 - \frac{A}{\alpha}\sum_{n=0}^{\infty} \frac{s^n}{\alpha^n} - \frac{B}{\beta}\sum_{n=0}^{\infty} \frac{s^n}{\beta^n} \\
&= -1 - \sum_{n=0}^{\infty} \left(\frac{1-\frac{3}{\sqrt{5}}}{\sqrt{5}-1} \cdot \frac{1}{(-1+\sqrt{5})^n} + \frac{1+\frac{3}{\sqrt{5}}}{-1-\sqrt{5}} \cdot \frac{1}{(-1-\sqrt{5})^n} \right) s^n \\
&= -1 + \sum_{n=0}^{\infty} \left(\frac{3-\sqrt{5}}{5-\sqrt{5}} \cdot \frac{1}{(\sqrt{5}-1)^n} + (-1)^n \frac{3+\sqrt{5}}{5+\sqrt{5}} \cdot \frac{1}{(\sqrt{5}+1)^n} \right) \cdot s^n
\end{aligned}$$

And therefore for $n > 0$ we have:

$$P(X_2 = n) = \frac{3-\sqrt{5}}{5-\sqrt{5}} \cdot \frac{1}{(\sqrt{5}-1)^n} + (-1)^n \frac{3+\sqrt{5}}{5+\sqrt{5}} \cdot \frac{1}{(\sqrt{5}+1)^n} \qquad (n > 0)$$

A good sanity check is $P(X_2 = 0) = 0$, which must be given by:

$$\begin{aligned}
P(X_2 = 0) &= -1 + \frac{3-\sqrt{5}}{5-\sqrt{5}} + \frac{3+\sqrt{5}}{5+\sqrt{5}} \\
&= -1 + \frac{(3-\sqrt{5}(5+\sqrt{5}) + (3+\sqrt{5)(5-\sqrt{5})}}{25-5} \\
&= -1 + \frac{15-5+3\sqrt{5}+15-5-3\sqrt{5}+5\sqrt{5}}{20} \\
&= -1 + \frac{10+10}{20} = 0 && (\checkmark)
\end{aligned}$$

Group Work

Develop the rational function $\frac{1}{2-s^2}$ in a power series. To do this we can break it up as $\frac{1}{(\sqrt{2}-s)(\sqrt{2}+s)}$. Then we write:

$$\frac{1}{2-s^2} = \frac{A}{\sqrt{2}-s} + \frac{B}{\sqrt{2}+s}$$

This gives us a system of equations:

$$\begin{aligned}
A(\sqrt{2}+s) + B(\sqrt{2}-s) &= 1 \\
A\sqrt{2} + B\sqrt{2} &= 1 \\
As - Bs = (A-B)s &= 0
\end{aligned}$$

This implies that $A = B$ and $2A\sqrt{2} = 1$ and $A = B = \frac{1}{2\sqrt{2}}$. Then we may write:

$$\begin{aligned}
\frac{1}{2-s^2} &= \frac{1}{2\sqrt{2}(\sqrt{2}-s)} + \frac{1}{2\sqrt{2}(\sqrt{2}+s)} \\
&= \frac{1}{4} \cdot \frac{1}{1-\frac{s}{\sqrt{2}}} + \frac{1}{4} \cdot \frac{1}{1+\frac{s}{\sqrt{2}}}
\end{aligned}$$

We may then write the generating function for each of these:

$$\frac{1}{2-s^2} = \frac{1}{4}\left(\sum_{n=0}^{\infty}\frac{s^n}{(\sqrt{2})^n} + \sum_{n=0}^{\infty}\frac{(-1)^n s^n}{(\sqrt{2})^n}\right)$$
$$= \frac{1}{4}\cdot\sum_{n=0}^{\infty}(1+(-1)^n)\cdot\frac{1}{(\sqrt{2})^n}\cdot s^n$$

And therefore the coefficients are:

$$a_n = \frac{1}{4}\cdot(1+(-1)^n)\cdot\frac{1}{(\sqrt{2})^n}$$
$$a_{2n} = \frac{1}{2^{n+1}}$$
$$a_{2n+1} = 0$$

MATH 525 Notes

Faye Jackson

November 5, 2020

Stochastic Processes

Recall. Remember that we defined a generating function for a discrete random variable X taking values in the non-negative integers:

$$G_X(s) = \mathrm{E}\big[s^X\big] = \sum_{x=0}^{\infty} s^x \cdot P(X = x)$$

This function has very nice properties such as:

$$\begin{aligned} G_X(0) &= P(X = 0) \\ G_X(1) &= 1 \\ G'_X(1) &= \mathrm{E}[X] \end{aligned}$$

If $X_1, \ldots, X_n$ are independent and $X = X_1 + \cdots + X_n$, then:

$$G_X(s) = G_{X_1}(s) \cdots G_{X_n}(s)$$

Galton-Watson Process

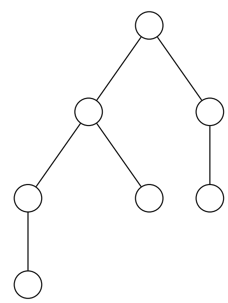
Definition. *Let $p : \mathbb{Z}_{\geq 0} \to \mathbb{Z}_{\geq 0}$ be a mass function, so $\sum_{k=0}^{\infty} p(k) = 1$. We consider a tree growing from a single root. At time zero there is only the root. At time n, each leaf at depth n grows k descendents with probability $p(k)$ independently of any other leaf. Let X_n be the total number of leaves at depth n. This is the Galton Watson Process*

Analytically, $X_0 = 1$ *and:*

$$X_n = \sum_{i=1}^{X_{n-1}} Y_{i,n-1}$$

where $Y_{1,n-1}, \dots, Y_{n-1,n-1}$ *are independent discrete random variables with mass function* p*. Note!* X_1 *has mass function* p*, i.e.* $Y_{i,n-1}$ *are independent copies of* X_1

Picture:



Then we have the following table of values:

n	0	1	2	3	4
X_n	1	2	3	1	0

Models: Extinction of family names and survival of genes in a gene pool

Question: The extinction probability at a particular time n, that is $P(X_n = 0)$, and eventually $\lim_{n\to\infty} P(X_n = 0)$. We also want to look at the expectation $\mathrm{E}[X_n]$

We will study this using generating functions, which we compute recursively:

$$G_n(s) = G_{X_n}(s) = \mathrm{E}\big[s^{X_n}\big]$$
$$g(s) := G_1(s) = \sum_{k=0}^{\infty} s^k \cdot P(X_1 = k) = \sum_{k=0}^{\infty} p(k)|cdots^k$$

We're going to condition on the value of X_1:

$$
\begin{aligned}
G_2(s) &= \mathrm{E}\big[s^{X_2}\big] \\
\mathrm{E}\big[s^{X_2} \bigm| X_1 = 0\big] &= 1 \\
\mathrm{E}\big[s^{X_2} \bigm| X_1 = 1\big] &= \mathrm{E}\big[s^{Y_{1,1}}\big] = g(s) \\
\mathrm{E}\big[s^{X_2} \bigm| X_1 = k\big] &= \mathrm{E}\big[s^{Y_{1,1}+\cdots+Y_{1,k}}\big] = (g(s))^k
\end{aligned}
$$

By conditioning, we then have that:

$$
\begin{aligned}
G_2(s) &= \sum_{k=0}^{\infty} \mathrm{E}\big[s^{X_1} \bigm| X_1 = k\big] \cdot P(X_1 = k) \\
&= \sum_{k=0}^{\infty} p(k) \cdot (g(s))^k \\
&= g(g(s))
\end{aligned}
$$

Iterating, we get that:

$$
\begin{aligned}
G_n(s) &= \underbrace{g(g(g(\cdots g(s)\cdots)}_{n \text{ times}} \\
G_n &= \underbrace{g \circ \cdots \circ g}_{n \text{ times}}
\end{aligned}
$$

Consider now the question about eventual extinction:

$$
\lim_{n\to\infty} P(X_n = 0) = \lim_{n\to\infty} G_n(0)
$$

This is the sequence $G_n(0)$ defined recursively as follows:

$$
\begin{aligned}
d_1 &= g(0) = p(0) \\
d_2 &= G_2(0) = g(g(0)) = g(d_1) \\
d_{n+1} &= g(d_n)
\end{aligned}
$$

Such a thing is called a dynamical system.

We have a simple special case where if $p(0) = 0$ then $d_n = 0$ for all n, so

$\lim_{n\to\infty} d_n = 0$. This would tell you that the process never dies, we could have anticipated this answer. This is intuitive since $p(0)$ is the probability that any given leaf has no descendents, thus saying $p(0) = 0$ means that every leaf must have at least one descendent

Now let's deal with the general case where $p(0) > 0$. We need to examine the properties that g has. Namely $g(0) = p(0) > 0$, $g(1) = 1$, g is non-decreasing between 0 and 1, and g is continuous.

Therefore we may consider $g : [0, 1] \to [0, 1]$ as a continuous and non-decreasing function. Remember that a non-decreasing bounded sequence has a limit, and note that d_n is a non-decreasing and bounded sequence, so we must have a limit called d. Now consider that:

$$g(d) = g(\lim_{n\to\infty} d_n) = \lim_{n\to\infty} g(d_n) = \lim_{n\to\infty} d_{n+1} = d$$

Therefore d is some fixed point of g. Moreover, if d' is any fixed point of g between 0 and 1 then, noting that:

- $d' > 0$ because $g(0) = p(0) > 0$.
- $g(0) \le g(d')$ since $0 < d'$, but then $d_1 = g(0) \le g(d') = d'$.
- But then $d_2 = g(d_1) \le g(d') = d'$.
- Therefore $d_n \le d'$ for all n, and so

$d = \lim_{n\to\infty} d_n \le d'$.

Conclusion: The eventual death probability d is the smallest fixed point of g between zero and one. Note that 1 is always a fixed point.

Example. Suppose that $p(0) = p(1) = p(2) = p(3) = \frac{1}{4}$. We are looking for fixed points:

$$\begin{aligned}
&s = g(s) = \frac{1}{4} + \frac{1}{4}s + \frac{1}{4}s^2 + \frac{1}{4}s^3 \\
&4s = 1 + s + s^2 + s^3 \\
&s^3 + s^2 - 3s + 1 = 0
\end{aligned}$$

We'll do long division of this polynomial by $s-1$, do this in the secrecy of your own homes.

$$\frac{s^3+s^2-3s+1}{s-1} = s^2+2s-1$$

And so we end up with three fixed points:

$$\begin{aligned} s &= 1 \\ s &= -1 \pm \sqrt{2} \end{aligned}$$

We need to look only between $[0,1]$, and so the death probability is $d = -1+\sqrt{2} \approx 0.4$.

Group Work

Same question, but $p(0) = \frac{1}{2}$, $p(1) = p(2) = p(3) = \frac{1}{6}$.

So we want to solve the equation:

$$\begin{aligned} s = g(s) &= \frac{1}{2} + \frac{1}{6}s + \frac{1}{6}s^2 + \frac{1}{6}s^3 \\ 6s &= 3 + s + s^2 + s^3 \\ s^3 + s^2 - 5s + 3 &= 0 \end{aligned}$$

We know 1 is a root of this equation, so we divide by $s-1$:

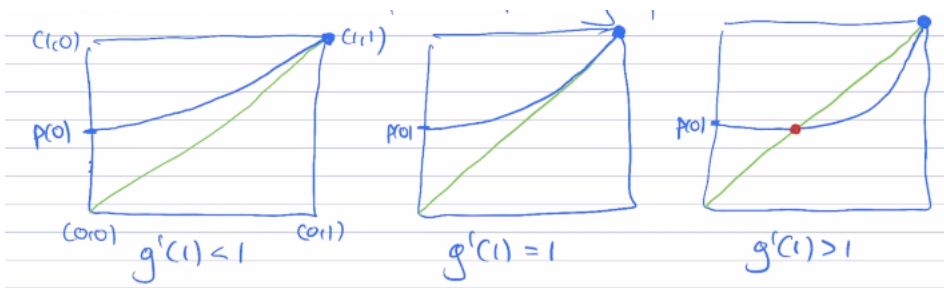
$$\frac{s^3+s^2-5s+3}{s-1} = s^2+2s-3$$

So then solving this equation we get three roots:

$$\begin{aligned} s &= 1 \\ s &= \frac{-2 \pm \sqrt{4+12}}{2} = -1 \pm 2 \end{aligned}$$

And so $s = 1, -3, 1$. Since $s = -3$ is outside of $[0,1]$ and thus $d = 1$.

Question: When is the survival probability positive? Equivalently, $d < 1$, that is g has a positive fixed point less than one. Note that g is convex (not strictly) so its graph can have one of the following forms:



Then $d < 1$ holds if and only if $\mathrm{E}[X_1] = g'(1) > 1$ which says exactly that the expected number of descendents of a given node is greater than one

What about extinction at time $n : P(X_n = 0)$. This is much harder, we must compute G_n. Sometimes possible.

Example. Start with $p(k) = 2^{-(k+1)}$ for $k = 0, 1, 2, \ldots$.

$$\begin{aligned} g(s) &= \sum_{k=0}^{\infty} p(k) s^k = \frac{1}{2} \cdot \sum_{k=0}^{\infty} \left(\frac{s}{2}\right)^k \\ &= \frac{1}{2} \cdot \frac{1}{1 - \frac{s}{2}} = \frac{1}{2 - s} \\ G_2(s) &= g(g(s)) = \frac{1}{2 - \frac{1}{2-s}} = \frac{2 - s}{3 - 2s} \\ G_3(s) &= g(G_2(s)) = \frac{1}{2 - \frac{2-s}{3-2s}} = \frac{3 - 2s}{4 - 3s} \\ G_n(s) &= \frac{n - (n-1)s}{(n+1) - ns} \end{aligned}$$

Recall the procedure to convert this into a power series:

$$\begin{aligned} \frac{n - (n-1)s}{(n+1) - ns} &= \frac{n-1}{2} + \frac{1}{n(n+1)} \cdot \frac{1}{1 - \frac{n}{n+1} s} \\ &= \frac{n-1}{2} + \frac{1}{n(n+1)} \sum * \left(\frac{n}{n+1} s\right)^k \end{aligned}$$

And so:

$$P(X_n = k) = \frac{1}{n(n+1)} \left(\frac{n}{n+1}\right)^k \qquad (k > 0)$$
$$P(X_n = 0) = \frac{n}{n+1}$$

Great!

We now look at the expectations:

$$\mathrm{E}[X_n] = G_n'(1) = g'(G_{n-1}(1))\cdot$$

MATH 525 Notes

Faye Jackson

November 10, 2020

Recall. X, Y independent $\implies \mathrm{E}[XY] = \mathrm{E}[X]\,\mathrm{E}[Y]$.

This motivated the definition of:

$$\begin{aligned}\operatorname{cov}(X,Y) &= \mathrm{E}[XY] - \mathrm{E}[X]\,\mathrm{E}[Y] = \mathrm{E}[(X-\mathrm{E}[X])(Y-\mathrm{E}[Y]] \\ \operatorname{var}(X) &= \operatorname{cov}(X,X) = \mathrm{E}\big[X^2\big] - \mathrm{E}[X]^2\end{aligned}$$

We introduced correlation and standard deviation as normalized versions of these concepts

We also talked about computation using density / mass functions using unconscious statistician:

$$\begin{aligned}\operatorname{var}(X) &= \sum_{k\in\mathbb{R}} k^2 \cdot f_X(k) - \left(\sum_{k\in\mathbb{R}} k \cdot f_X(k)\right)^2 \\ \operatorname{var}(X) &= \int_{\mathbb{R}} t^2 f_X(t)\,\mathrm{d}t - \left(\int_{\mathbb{R}} t \cdot f_X(t)\,\mathrm{d}t\right)^2\end{aligned}$$

Let's do some examples using random variables we've seen before

Example. Let X be Poisson with parameter λ, then:

$$\begin{aligned}
f_X(k) &= \frac{\lambda^k}{k!}e^{-\lambda} && (k = 0, 1, \ldots) \\
\mathrm{E}[X] &= \sum_{k=0}^{\infty} k \cdot \frac{\lambda^k}{k!}e^{-\lambda} \\
&= \lambda \cdot e^{-\lambda} \cdot \sum_{k=1}^{\infty} \frac{\lambda^{k-1}}{(k-1)!} \\
&= \lambda \cdot e^{-\lambda} \cdot \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} \\
&= \lambda \cdot e^{-\lambda} \cdot e^{\lambda} = \lambda
\end{aligned}$$

So then:

$$\mathrm{var}(X) = -\lambda^2 + \sum_{k=0}^{\infty} k^2 \frac{\lambda^k}{k!}e^{-\lambda}$$

We compute the sum:

$$\begin{aligned}
\sum_{k=0}^{\infty} k^2 \frac{\lambda^k}{k!}e^{-\lambda} &= \sum_{k=1}^{\infty} k \cdot \frac{\lambda^k}{(k-1)!}e^{-\lambda} \\
&= \lambda e^{-\lambda} \sum_{k=0}^{\infty} (k+1)\frac{\lambda^k}{k!} \\
&= \lambda e^{-\lambda} \left(\sum_{k=0}^{\infty} k\frac{\lambda^k}{k!} + \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} \right) \\
&= \lambda e^{-\lambda}(\lambda \cdot e^{\lambda} + e^{\lambda}) = \lambda^2 + \lambda
\end{aligned}$$

And so:

$$\mathrm{var}(X) = \lambda^2 + \lambda - \lambda^2 = \lambda$$

Great!

Example. Let X be exponential with parameter λ. Then:

$$\begin{aligned} f_X(t) &= \lambda e^{-t\lambda} && (t > 0) \\ \mathrm{E}[X] &= \int_0^\infty t\lambda e^{-t\lambda}\,\mathrm{d}t \\ &= \left[-te^{-t\lambda}\right)_0^\infty + \int_0^\infty e^{-t\lambda}\,\mathrm{d}t \\ &= 0 + \left[-\frac{1}{\lambda}e^{-t\lambda}\right)_0^\infty = \frac{1}{\lambda} \end{aligned}$$

Now to compute the variance we consider that:

$$\begin{aligned} \int_0^\infty t^2 \cdot \lambda e^{-t\lambda}\,\mathrm{d}t &= \left[-t^2 e^{-t\lambda}\right)_0^\infty + \int_0^\infty 2te^{-t\lambda}\,\mathrm{d}t \\ &= 0 + \frac{2}{\lambda} \cdot \int_0^\infty t\lambda e^{-t\lambda}\,\mathrm{d}t \\ &= \frac{2}{\lambda} \cdot \frac{1}{\lambda} = \frac{2}{\lambda^2} \end{aligned}$$

And therefore:

$$\mathrm{var}(X) = \frac{2}{\lambda^2} - \left(\frac{1}{\lambda}\right)^2 = \frac{1}{\lambda^2}$$

Great!

We now introduce a technique using variance and generating functions:

Lemma. *Let X be a discrete random variable taking values in $\mathbb{Z}_{\geq 0}$. Then:*

$$\mathrm{var}(X) = G_X''(1) + G_X'(1) - \left(G_X'(1)\right)^2$$

Proof. Recall that:

$$G_X(t) = \sum_{k=0}^{\infty} f_X(t) s^k$$
$$G'_X(t) = \sum_{k=1}^{\infty} k f_X(t) s^{k-1}$$
$$G''_X(t) = \sum_{k=2}^{\infty} k(k-1) f_X(t) s^{k-1}$$

Just from this we see that $\mathrm{E}[X] = G'_X(1)$ as we did last time. Now consider:

$$\begin{aligned} G''_X(1) &= \sum_{k=2}^{\infty} k(k-1) f_X(k) = \sum_{k=1}^{\infty} k(k-1) f_X(k) \\ &= \sum_{k=1}^{\infty} k^2 f_X(k) - \sum_{k=1}^{\infty} k f_X(k) \\ &= \mathrm{E}\left[X^2\right] - \mathrm{E}[X] \end{aligned}$$

But wait! This means that $\mathrm{E}\left[X^2\right] = G''_X(1) + G'_X(1)$. Therefore:

$$\mathrm{var}(X) = \mathrm{E}\left[X^2\right] - \mathrm{E}[X]^2 = G''_X(1) + G'_X(1) - \left(G'_X(1)\right)^2$$

Example. We will think about the branching processs that we considered in the last two lectures. Let $p : \mathbb{Z}_{\geq 0} \to [0,1]$ be a mass functon. We interpret $p(k)$ as the probability of a given leaf having k descendents. We have the generating function:

$$g(s) = \sum p(k) s^k$$

Letting X_n be the total number of leaves at depth n. We argued that:

$$G_n(s) = G_{X_n}(s) = \underbrace{g \circ \cdots \circ g}_{n \text{ times}} = g \circ G_{n-1}$$

Then to compute $\text{var}(X_n)$ we compute:

$$G_n'(s) = g'(G_{n-1}(s)) \cdot G_{n-1}'(s)$$
$$G_n''(s) = g''(G_{n-1}(s)) \cdot (G_{n-1}'(s))^2 + g'(G_{n-1}(s)) \cdot G_{n-1}''(s)$$

Then:

$$G_n'(s) = g'(1) \cdot G_{n-1}; (1) = (g'(1))^n$$
$$G_n''(1) = g''(1) \cdot (g'(1))^{2(n-1)} + g'(1) \cdot G_{n-1}''(1)$$

For simplicity, assume $g'(1) = 1$, then:

$$\begin{aligned} G_n'(1) &= 1 \\ G_n''(1) &= g''(1) + G_{n-1}''(1) \\ &= ng''(1) \end{aligned}$$

And therefore:

$$\text{var}(X_1) = g''(1) + 1 - (1)^2 = g''(1)$$
$$\text{var}(X_n) = ng''(1) + 1 - (1)^2 = ng''(1) = n\,\text{var}(X_1)$$

Great!

Proposition. *Let $X_1, \ldots, X_n$ be random variables. We are not assuming $X_1, \ldots, X_n$ are independent. Let $X = X_1 + \cdots + X_n$. Then:*

$$\begin{aligned} \text{var}(X) &= \sum_{i=1}^{n} \text{var}(X_i) + \sum_{1 \leq i \neq j \leq n} \text{cov}(X_i, X_j) \\ &= \sum_{i=1}^{n} \text{var}(X_i) + 2 \cdot \sum_{1 \leq i < j \leq n} \text{cov}(X_i, X_j) \end{aligned}$$

In particular, if $X_1, \ldots, X_n$ are independent, then:

$$\operatorname{var}(X) = \sum_{i=1}^{n} \operatorname{var}(X_i)$$

Proof. We just compute!

$$\begin{aligned}
\operatorname{var}(X) &= \mathrm{E}\big[(X - \mathrm{E}[X])^2\big] \\
&= \mathrm{E}\left[\left(\sum_{i=1}^{n} X_i - \mathrm{E}[X_i]\right)^2\right] \\
&= \mathrm{E}\left[\sum_{i=1}^{n} X_i - \mathrm{E}[X_i] + \sum_{1 \le i \neq j \le n} (X_i - \mathrm{E}[X_i])(X_j - \mathrm{E}[X_j])\right]
\end{aligned}$$

And then we know by linearity:

$$\operatorname{var}(X) = \sum_{i=1}^{n} \operatorname{var}(X_i) + \sum_{1 \le i \neq j \le n} \operatorname{cov}(X_i, X_j)$$

Example. Let X be Bernoulli with parameter p and $q := (1 - p)$. Then:

$$\begin{aligned}
\mathrm{E}[X] &= p \\
\mathrm{E}\big[X^2\big] &= p \\
\operatorname{var}(X) &= p - p^2 \\
&= p(1 - p) = pq
\end{aligned}$$

Example. X is binomial wit parameter n, p. Then $X = X_1 + \ldots + X_n$ for X_i Bernoulli with parameter p and independent. Then $\mathrm{E}[X] = np$ and $\operatorname{var}(X) = np(1 - p)$

Group Work

Recall the problem of letters. n personalized letters sent out randomly. Let X be the numbe of letters that reach their intended recipient. Compute $\text{var}(X)$.

Let X_i be the Bernoulli random variable that tells us if person i receives their letter and note that X_i has parameter $\frac{1}{n}$ since there is one letter for person i out of n letters. In a nice turn of events, $X = X_1 + \cdots + X_n$.

Now let $1 \leq i \neq j \leq n$. We compute $\text{cov}(X_i, X_j)$, to do this it's convenient to compute that:

$$\mathrm{E}[X_i X_j] = P(X_i = 1, X_j = 1)$$

Since $X_i X_j$ is either 0 or 1. The 0 does not contribute anything to the probability and $X_i X_j = 1$ only when both X_i and X_j are equal to 1. To compute this probability note that we must first send the correct letter to i, with probability $\frac{1}{n}$, and then send the correct letter to j with probability $\frac{1}{n-1}$:

$$\mathrm{E}[X_i X_j] = \frac{1}{n(n-1)}$$

Therefore:

$$\text{cov}(X_i, X_j) = \frac{1}{n(n-1)} - \frac{1}{n} \cdot \frac{1}{n} = \frac{1}{n^2(n-1)}$$

This allows us to write by the above that:

$$\begin{aligned}
\text{var}(X) &= \sum_{i=1}^{n} \text{var}(X_i) + \sum_{1 \leq i \neq j \leq n} \text{cov}(X_i, X_j) \\
&= \sum_{i=1}^{n} \frac{n-1}{n^2} + \sum_{1 \leq i \neq j \leq n} \frac{1}{n^2(n-1)} \\
&= n \cdot \frac{n-1}{n^2} + \frac{n(n-1)}{n^2(n-1)}
\end{aligned}$$

This is simple to compute:

$$\operatorname{var}(X) = \frac{n-1}{n} + \frac{1}{n} = 1$$

Note that this only works when $n > 1$, when $n = 1$ we have $\operatorname{var}(X) = 0$.

Example. We toss a coin $n > 5$ times and let X be the number of switches from H to T. We want to compute the variance of X. Write $X = X_1 + \cdots + X_{n-1}$ where X_i is the indicator that the i-th toss is H and $i+1$-th toss is T.

Each indicator is Bernoulli with parameter $\frac{1}{4}$ so $\mathrm{E}[X_i] = \frac{1}{4}$ and $\operatorname{var}(X_i) = \frac{3}{16}$. Note that if $i \neq j$ then X_i and X_j are independent unless i, j are consecutive.

If $j = i + 1$ then:

$$\begin{aligned}\operatorname{cov}(X_i, X_j) &= \mathrm{E}[X_i X_j] - \mathrm{E}[X_i]\,\mathrm{E}[X_j] \\ &= 0 - \frac{1}{16} = -\frac{1}{16}\end{aligned}$$

This occurs since X_i and X_j are Bernoulli for disjoint events, the $i+1$-th toss cannot be both heads and tails. Then:

$$\begin{aligned}\operatorname{var}(X) &= \sum_{i=1}^{n-1} \operatorname{var}(X_i) + 2 \sum_{1 \leq i < j \leq n} \operatorname{cov}(X_i, X_j) \\ &= \frac{3(n-1)}{16} + 2(n-2) \cdot -\frac{1}{16} = \frac{n+1}{16}\end{aligned}$$

MATH 525 Notes

Faye Jackson

November 12, 2020

Concentration Inequalities

Recall. If $X : \Omega \to \mathbb{R}$ is a Random Variable such that $\mathrm{E}[X]$ and $\mathrm{var}(X)$ both exist and $\mathrm{var}(X) = 0$. Then $X = \mathrm{E}[X]$ almost surely

Proposition (Markov)**.** *Let* $X : \Omega \to \mathbb{R}$ *be a Random Variable with non-negative values and assume* $\mathrm{E}[X]$ *exists. Then for any* $a > 0$ *we can estimate:*

$$P(X \geq a) \leq a^{-1}\,\mathrm{E}[X]$$

Proof. Let $Y : \Omega \to \mathbb{R}$ be defined as:

$$Y(\omega) = \begin{cases} a & \text{if } X(\omega) \geq a \\ 0 & \text{if } X(\omega) < a \end{cases}$$

Then Y is a Random Variable because $Y = a \cdot \mathbb{1}_{X \geq a}$. Then note that $X \geq Y$ and so $\mathrm{E}[X] \geq \mathrm{E}[Y]$. But wait:

$$\mathrm{E}[X] \geq \mathrm{E}[Y] = a \cdot P(X \geq a)$$

And so we're done!

Corrolary (Chebyshev's Inequality)**.** *Let* $X : \Omega \to \mathbb{R}$ *be a Random Variable.*

Assume that $\mathrm{E}[X]$ *and* $\mathrm{var}(X)$ *both exist. Then for* $a > 0$*:*

$$P(|X - \mathrm{E}[X]| \geq a) \leq \frac{\mathrm{var}(X)}{a^2}$$

Proof. Let $Y = (X - \mathrm{E}[X])^2$. Then Y is a Random Variable, moreover $Y \geq 0$ and $\mathrm{E}[Y] = \mathrm{var}(X)$ by definition. We apply Markov's Inequality to Y:

$$P(|X - \mathrm{E}[X]| \geq a) = P(Y \geq a^2) \leq \frac{\mathrm{var}(X)}{a^2}$$

And so we win!

Example. We roll a fair die 10,000 times. Let X be the total score, i.e. $X = X_1 + \cdots + X_{10,000}$ where X_i is the score at the i-th roll. Then $\mathrm{E}[X_i] = \frac{7}{2}$, so $\mathrm{E}[X] = 35,000$. Then consider that:

$$\mathrm{var}(X_i) = \mathrm{E}\left[X_i^2\right] - \mathrm{E}[X_i]^2 = \frac{1^2 + 2^2 + \cdots + 6^2}{6} - \frac{49}{4} = \frac{35}{12}$$

And since each X_i is independent:

$$\mathrm{var}(X) = \mathrm{var}(X_1) + \cdots + \mathrm{var}(X_{10,000}) = \frac{350,000}{12}$$

Applying Chebyshev's inequality we see that:

$$P(|X - 35,000| \geq 1000) \leq \frac{\frac{350,000}{12}}{10^6} = \frac{35}{1200} \approx 0.029$$

Cool! But we can get a much better bound

Proposition (Bernstein Inequality)**.** *Let* $X_1, \ldots, X_n$ *be independent Random Variables such that* $\mathrm{E}[X_i] = 0$ *and* $|X_i| \leq 1$*. Let* $X = X_1 + \cdots + X_n$*. For any* $a \geq 0$ *we have:*

$$P(X \geq a) \leq e^{-a^2/2n} \qquad\qquad P(X \leq -a) \leq e^{-a^2/2n}$$

Proof. The collection $-X_1, -X_2, \ldots, -X_n$ satsifies the assumption so $P(X \leq$

$-a) \leq e^{a^2/2n}$ follows from $P(X \geq a) \leq e^{-a^2/2n}$ applied to $-X_1, \ldots, -X_n$.

Thus it is enough to prove $P(X \geq a) \leq e^{-a^2/2n}$, and it is enough to assume that $a > 0$, because for $a = 0$ we get:

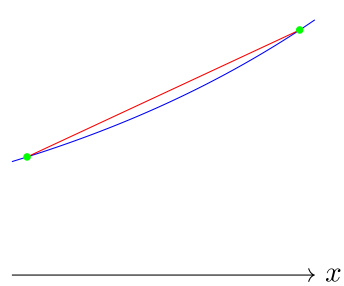
$$P(X \geq 0) \leq e^0 = 1$$

But the probability of any event is less than or equal to one!

Now let $a > 0$ and let $t > 0$ be a variable, to be fixed later. Then by independence and Markov's Inequality:

$$\begin{aligned} P(X \geq a) &= P(e^{tX} \geq e^{ta}) \\ &\leq e^{-ta} \cdot \mathrm{E}\big[e^{tX}\big] \\ &= e^{-ta} \cdot \mathrm{E}\big[e^{tX_1} \cdots e^{tX_n}\big] \\ &= e^{-ta} \cdot \prod_{i=1}^{n} \mathrm{E}\big[e^{tX_i}\big] \end{aligned}$$

Note that $\mathrm{E}\big[e^{tX_i}\big]$ exists because X_i is bounded. Now note that $x \mapsto e^{tx}$ is bounded:



Then the exponential is convex. What does this mean? well f is convex if and only if for every $z \leq x \leq w$. we have:

$$\begin{aligned} f(x) &\leq \frac{f(w) - f(z)}{w - z} \cdot w - \frac{f(w) - f(z)}{w - z} \cdot z + f(z) \\ &= \frac{f(w) - f(z)}{w - z} - \frac{f(w)z - f(z)w}{w - z} \end{aligned}$$

Thus for $f(x) = e^{tx}$, $w = 1$, $z = -1$ we get:

$$e^{tx} \leq \frac{e^t - e^{-t}}{2} \cdot x + \frac{e^t + e^{-t}}{2}$$

So then:

$$\mathrm{E}\big[e^{tX_i}\big] \leq \mathrm{E}\left[\frac{e^t - e^{-t}}{2} \cdot X_i + \frac{e^t + e^{-t}}{2}\right] = \frac{e^t + e^{-t}}{2}$$

Now consider that:

$$\frac{e^t + e^{-t}}{2} = \frac{1}{2}\sum_{k=0}^{\infty} \frac{t^k + (-t)^k}{k!} = \sum_{k=0}^{\infty} \frac{t^{2k}}{(2k)!}$$
$$e^{t^2/2} = \sum_{k=0}^{\infty} \frac{t^{2k}}{2^k \cdot k!}$$

And since $2^k \cdot k! \leq (2k)!$ we have:

$$\mathrm{E}\big[e^{tX_i}\big] \leq \frac{e^t + e^{-t}}{2} \leq e^{t^2/2}$$

Putting this together we have:

$$\begin{aligned} P(X \geq a) &\leq e^{-ta} \cdot \prod_{i=1}^{n} \mathrm{E}\big[e^{tX_i}\big] \\ &\leq e^{-ta} \cdot e^{nt^2/2} \end{aligned}$$

Then if we set $t = a/n$ we get:

$$P(X \geq a) \leq e^{-2a^2/2n} \cdot e^{a^2/2n} = e^{-a^2/2n}$$

Great! This is exactly what we want ☺

Example. We can renormalize each X_i in our previous example as:

$$Y_i = \frac{2}{5}\left(X_i - \frac{7}{2}\right)$$

to achieve that $\mathrm{E}[X] = 0$ and $|Y_i| \leq 1$. Then set $Y = Y_1 + \cdots + Y_{10,000}$. Then:

$$|X - 35,000| \geq 1000 \iff |Y| \geq 400$$

Therefore:

$$\begin{aligned} P(|X - 35,000| \geq 1000) &= P(Y \leq -400) + P(Y \geq 400) \\ &\leq 2 \cdot e^{-400^2/20,0000} = 2 \cdot e^{-8} \approx 0.00067 \end{aligned}$$

Group Work

Consider a symmetric Random Walk with no barriers starting at 0. The Question is how far do you get in n steps. There exists the trajectory that goes to position n and the trajectory that goes to n. More precisely, compute $P(|X| \geq k)$ and deduce from that how far you are likely to go.

Let $X = X_1 + \cdots + X_n$ and each X_i is "Bernoulli" with values in $\{\pm 1\}$. These are independent and so:

$$P(|X| \geq k) \leq 2e^{-k^2/2n}$$

Then write $k = a \cdot \sqrt{n}$. Then:

$$P(|X| \geq a\sqrt{n}) \leq ee^{-a^2/2}$$

Then:

a	$P(\lvert X\rvert \geq a\sqrt{n})$
1	1
2	0.27
3	0.022
4	0.00007

Thus, we will likely get as far as about $\sqrt{n}$.

Example. Lets go for the St. Petersburg paradox. We play the following game (one round). You flip a fair coin until you get tails. If you had to flip m times to get tails, I will pay you $\$2^m$ dollars

Question: What is the fair entrance fee?

Let X be the winning amount. We compute:

$$\mathrm{E}[X] = \frac{1}{2} \cdot 2 + \frac{1}{4} \cdot 4 + \frac{1}{8} \cdot 8 + \cdots = \infty$$

Let's play n rounds and a_n is the fair entrance fee for one round given the information that the game will be played n rounds. Let X_i be the win amount in round i. The let $Y_n = X_1 + \cdots + X_n$. Then:

$$P\left(\left|\frac{Y_n}{na_n} - 1\right| > \varepsilon\right) \to 0 \text{ as } n \to \infty \qquad (\varepsilon > 0)$$

We have $P(X_i = 2^m) = \frac{1}{2^m}$. Let m_n be an integer. Let $\overline{X}_i$ give X_i provided that we flip at most 2^{m_n} times and 0 otherwise. Then $\overline{Y}_n = \overline{X}_1 + \cdots + \overline{X}_n$.

Let $a_n = \mathrm{E}\left[\overline{X}_i\right] = m_n$. Then $\mathrm{E}\left[\overline{Y}_n\right] = n \cdot m_n$. Then:

$$p\left(\left|\frac{Y_n}{n \cdot a_n} - 1\right| > \varepsilon\right) \leq P\left(\left|\frac{\overline{Y}_n}{na_n} - 1\right|\right) + P(\overline{Y}_n \neq Y_n)$$

Now $\mathrm{var}(\overline{X}_i) \leq \mathrm{E}\left[\overline{X}_i^2\right] \leq 2^{m_n+1}$. Now Chebyshev implies that:

$$P\left(\left|\frac{\overline{Y}_n}{na_n} - 1\right|\right) \leq \frac{2^{m_n+1}}{\varepsilon^2 n m_n}$$

So then:

$$P(\overline{Y}_n \neq Y_n) \leq \sum_{i=1}^{n} P(\overline{X}_i \neq X_i) = n \cdot 2^{m_n}$$

We see that $m_n = a_n = \log_2 n + \log_2 \log_2 n$ makes both probabilities go to zero.

MATH 525 Notes

Faye Jackson

November 17, 2020

Recall. Chebyshev said that:

$$P(|X - \mathrm{E}[X]| \geq a) \leq \frac{\mathrm{var}(X)}{a^2}$$

Lets rescale a by $\sigma(X) = \sqrt{\mathrm{var}(X)}$. Then we get that:

$$P(|X - \mathrm{E}[X]| \geq a \cdot \sigma(X)) \leq \frac{1}{a^2}$$

For $0 < a < 1$ we get no information. But for $a > 1$ we do get information:

$$P(|X - \mathrm{E}[X]| < 2 \cdot \sigma(X)) \geq 1 - \frac{1}{4} = 75\% \qquad (a = 2)$$
$$P(|X - \mathrm{E}[X]| < 3 \cdot \sigma(X)) \geq 1 - \frac{1}{9} = 89\% \qquad (a = 3)$$

Of course Chebyshev doesn't always give the best estimate.

Example. If X is normally distributed, i.e. X is continuous and its density is $f_X(t) = \frac{1}{\sqrt{2\pi}} e^{-t^2/2}$, then $\mathrm{E}[X] = 0$ and $\sigma(X) = 1$. Then if we compute:

$$\begin{aligned} P(|X - \mathrm{E}[X]| < k\sigma(X)) &= P(|X| < k) \\ &= \int_{-k}^{k} f_X(t)\,\mathrm{d}t = \begin{cases} 95\% & \text{if } k = 2 \\ 99\% & \text{if } k = 3 \end{cases} \end{aligned}$$

Wow!

Example. But, the point is that Chebyshev cannot be improved for general

X. Consider a $X : \Omega \to \{-1, 0, 1\}$ such that:

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

Then we immediately see that $\mathrm{E}[X] = 0$, $\mathrm{var}(X) = \frac{1}{k^2}$, and $\sigma(X) = \frac{1}{k}$. So then:

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

The Law of Large Numbers

We expect that the theory of probability that we've developed conforms to our intuition. That is the probability of an event, or the expected value of a Random Variable, is reflected in the statistical outcomes. For example, $\mathrm{E}[X]$ should equal the average of all outcomesof "measuring X" many times independently.

In our framework, we take Random Variables $X_1, X_2, \ldots, X_n$ that are independent and have the same distribution as X (clones!). Define $S_n = \frac{1}{n}(X_1 + \cdots + X_n)$, the average. We expect:

$$S_n \to \mathrm{E}[X] \text{ as } n \to \infty$$

We just need to formalize this convergence!

Definition. *Let $Y_1, Y_2, \ldots$ be a sequence of Random Variables and let Y be a random variable. We say $\underline{Y_n \to Y}$:*

- *<u>almost surely</u> provided that:*

  $$P(\{\omega \in \Omega \mid Y_n(\omega) \to Y(\omega)\}) = 1$$

  In this case we write $\underline{Y_n \xrightarrow{a.s.} Y}$

- *<u>in probability</u> provided that for every $\varepsilon > 0$ we have that:*

  $$P(|Y_n - Y| > \varepsilon) \to 0 \text{ as } n \to \infty$$

In this case we write $Y_n \xrightarrow{p} Y$

Proposition 1. $Y_n \xrightarrow{a.s.} Y$ implies that $Y_n \xrightarrow{p} Y$.

Proof. Let $\varepsilon > 0$. Consider the event:

$$\begin{aligned} A_n(\varepsilon) &= \{\omega \in \Omega \mid \; |Y_n(\omega) - Y(\omega)| > \varepsilon\} \\ A(\varepsilon) &= \{\omega \in \Omega \mid \omega \in A_n(\varepsilon) \text{ for infinitely many } n\} \end{aligned}$$

First note that if $\varepsilon_1 < \varepsilon$ then $A_n(\varepsilon_1) \supseteq A_n(\varepsilon)$ and $A(\varepsilon_1) \supseteq A(\varepsilon)$. Note that:

$$A(\varepsilon) = \bigcap_{m=1}^{\infty} \bigcup_{n \geq m} A_n(\varepsilon)$$

Now we can consider that:

$$\{\omega \in \Omega \mid Y_n(\omega) \to Y(\omega)\} = \bigcap_{\varepsilon > 0} A(\varepsilon)^c = \left(\bigcup_{\varepsilon > 0} A(\varepsilon) \right)^c$$

Buw wait! We know that the probability of the left hand side is 1 by almost sure convergence, and so since this is an ascending tower:

$$P\left(\bigcup_{\varepsilon > 0} A(\varepsilon) \right) = 0 \implies P(A(\varepsilon)) = 0 \qquad (\forall \varepsilon > 0)$$

And then we get a descending tower:

$$\begin{aligned} 0 = P(A(\varepsilon)) = P\left(\bigcap_{m=1}^{\infty} \bigcup_{n \geq m} A_n(\varepsilon) \right) \\ = \lim_{m \to \infty} P\left(\bigcup_{n \geq m} A_n(\varepsilon) \right) \\ \geq \lim_{m \to \infty} P(A_m(\varepsilon)) \end{aligned}$$

But wait! This exactly means that Y_n converges to Y in probability!!!

Remark. The converse is not true, i.e. convergence in probability is strictly weaker than almost sure convergence.

Consider the sequence of events in $\Omega = [0, 1]$. First let:

$$A_1 = [0, 1/2], A_2 = [1/2, 1]$$
$$A_3 = [0, 1/3], A_4 = [1/3, 2/3], A_5 = [2/3, 1], \ldots$$

Define $Y_n = \mathbb{1}_{A_n}$. Then we want Y_n to converge probabilistically to 0. Well:

$$P(|Y_n| > \varepsilon) = P(A_n) \to 0 \text{ as } n \to \infty$$

But! Y_n does not converge to 0 almost surely. For any $\omega \in \Omega$ we can find sequences i_n, j_n such that $\omega \in A_{i_n}$ and $\omega \notin A_{j_n}$. But then $Y_{i_n}(\omega) = 1$ and $Y_{j_n}(\omega) = 0$, and so the limit of $Y_n(\omega)$ as $n \to \infty$ cannot exist. Therefore:

$$\{\omega \in \Omega \mid Y_n(\omega) \to 0\} = \emptyset$$

And so of course we cannot have Y_n converges to 0 almost surely.

Definition. *Let $X_1, X_2, \ldots$ be a sequence of independent identically distributed random variables (iid). We then define:*

$$S_n = \frac{1}{n}(X_1 + \cdots + X_n)$$

This sequence satisfies:

- *the strong law of large numbers provided that S_n converges to $\mathrm{E}[X_1]$ almost surely*
- *the weak law of large numbers provided that S_n converges to $\mathrm{E}[X_1]$ probabilistically.*

Theorem. *If $\mathrm{E}[|X_1|] < +\infty$, then the strong law holds. The proof is difficult and complicated.*

Proposition. *If $\mathrm{var}(X_1) < \infty$ then the weak law holds.*

Proof. Consider $S_n = \frac{1}{n}(X_1 + \cdots + X_n)$. Then consider that:

$$\mathrm{E}[S_n] = \frac{1}{n}\,\mathrm{E}[X_1 + \cdots + X_n] = \frac{n}{n}\,\mathrm{E}[X_1] = \mathrm{E}[X$$

$$\mathrm{var}(S_n) = \frac{1}{n^2}\,\mathrm{var}(X_1 + \cdots + X_n) = \frac{1}{n^2} \cdot n\,\mathrm{var}(X_1) = \frac{\mathrm{var}(X_1)}{n}$$

And then we have by Chebyshev's inequality that:

$$P(|S_n - \mathrm{E}[X_1]| > \varepsilon) \leq \frac{\mathrm{var}(S_n)}{\varepsilon^2} = \frac{\mathrm{var}(X_1)}{n \cdot \varepsilon^2} \to 0 \text{ as } n \to \infty$$

Remark. This proof even tells us that the rate of convergence is linear with proportionality $\frac{\mathrm{var}(X_1)}{\varepsilon^2}$.

Characteristic Functions

Recall. If X is a discrete Random Variable taking values in $0, 1, 2, \ldots$ then we have a generating function:

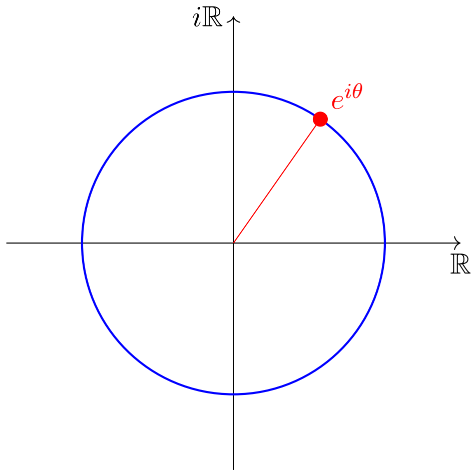
$$G_X(s) = \sum_{k=0}^{\infty} f_X(k) s^k = \mathrm{E}\left[s^X\right]$$

The question: can we do this for an random variable X. The answer is yes.

Definition. *For any Random Variable X, the characteristic function ϕ_X is defined as:*

$$\phi_X(t) := \mathrm{E}\left[e^{itX}\right] = \mathrm{E}[\cos(tX) + i\sin(tX)] = \mathrm{E}[\cos(tX)] + i\,\mathrm{E}[\sin(tX)]$$

We know this because if we recall from calculus that $\theta \mapsto e^{i\theta}$ traces out a circle:



Here are some nice properties:

- The integral $\mathrm{E}\left[e^{itX}\right]$ always converse
- If X is discrete with mass f_X then:

$$\phi_X(t) = \sum_{x=-\infty}^{\infty} f_X(x) \cdot e^{itx}$$

  Great! And furthermore if X takes values in $0, 1, \ldots$ then:

$$\phi_X(t) = G_X(e^{it})$$

- If X is continuous with density f_X then:

$$\phi_X(t) = \int_{-\infty}^{\infty} f_X(t) e^{itx} \, \mathrm{d}x$$

Example. There are of course many examples:

- If X is Bernoulli with parameter p then:

$$\phi_X(t) = \mathrm{E}\left[e^{itX}\right] = 1 \cdot (1-p) + e^{it} \cdot p$$

- If X is Binomial with parameter n, p we get that:

$$G_X(s) = ((1-p) + ps)^n \qquad \phi_X(t) = ((1-p) + pe^{it})^n$$

- If X is Poisson with parameter λ then:

$$G_X(s) = e^{\lambda(s-1)} \qquad \phi_X(t) = e^{\lambda\left(e^{it}-1\right)}$$

- Suppose X is exponential with parameter λ, then:

$$\begin{aligned}\phi_X(t) &= \int_0^\infty e^{itx} \cdot \lambda e^{-\lambda x} \, \mathrm{d}x = \lambda \int_0^\infty e^{x(it-\lambda)} \, \mathrm{d}x \\ &= \lambda \left[\frac{e^{x(it-\lambda)}}{it-\lambda} \right)_0^\infty = \frac{\lambda}{\lambda - it}\end{aligned}$$

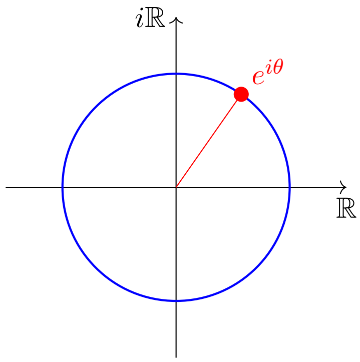
MATH 525 Notes

Faye Jackson

November 19, 2020

Recall. If X is an Random Variable its characteristic function $\phi_X(t) = \mathrm{E}\left[e^{itX}\right]$. This is based on:

$$e^{it} = \cos(t) + i\sin(t)$$



Example. Let $X : \Omega \to [-1, 1]$ where $P(X = -1) = P(X = 1) = \frac{1}{2}$. Then:

$$\phi_X(t) = \frac{e^{it} + e^{-it}}{2} = \cos(t)$$

If $X = \mu \in \mathbb{R}$ is constant then:

$$\phi_X(t) = e^{it\mu} = \cos(t\mu) + i\sin(t\mu)$$

Properties of ϕ_X:

- $\phi'(0) = i\,\mathrm{E}[X]$, $\phi''(0) = -\,\mathrm{E}\left[X^2\right]$, provided these expectations exist.

Then:

$$\mathrm{var}(X) = -\phi''(0) + \phi'(0)^2$$

If $\phi'(0)$ exists then $\mathrm{E}[X]$ may or may not exist. However if $\phi''(0)$ exists then both $\mathrm{E}[X]$ and $\mathrm{E}\left[X^2\right]$ exist.

- If ϕ_X is integrable, then X is continuous with density:

$$f_X(t) = \frac{1}{2\pi} \cdot \int_{-\infty}^{\infty} e^{-itx} \phi_X(x) \, \mathrm{d}x$$

- More generally, we can get information about F_X as follows:

$$\lim_{T\to\infty} \frac{1}{2\pi} \cdot \int_{-T}^{T} \frac{e^{-ita} - e^{-itb}}{it} \phi_X(t) \, \mathrm{d}t = P(a < x < b) + \frac{1}{2} P(X = a) + \frac{1}{2} P(X = b)$$

  This is called Levy's Theorem

- Levy's Continuity Theorem: If (X_n) is a sequence of Random Variable and X is a random Variable such that $\phi_{X_n} \to \phi_X$ then we have $F_{X_n}(y) \to F_X(y)$ for all y where F_X is continuous.

Proposition. *An sequence (X_n) of independent identical distributed Random Variables satisfies the weak law of large numbers without assuming $\mathrm{var}(X_i) < \infty$.*

Proof. Consider $\phi = \phi_{X_n}$. Let $S_n = \frac{1}{n}(X_1 + \cdots + X_n)$ and $\mu = \mathrm{E}[X_i]$. Then we know that ϕ is differentiable at zero by the properties. Then:

$$\begin{aligned} \phi(t) &= 1 + \phi'(0) \cdot t + \alpha(t) \cdot t \\ &= 1 + i\mu t + \alpha(t) \cdot t \end{aligned}$$

Such that $\alpha(t) \to 0$ as $t \to 0$. Then

$$
\begin{aligned}
\phi_{S_n}(t) &= \phi(t/n)^n \\
\phi_{S_n}(t) &= \lim_{n\to\infty} \phi(t/n)^n \\
&= \lim_{n\to\infty} \left(1 + i\mu\frac{t}{n} + \alpha\left(\frac{t}{n}\right)\cdot\frac{t}{n}\right)^n \\
&= e^{i\mu t}
\end{aligned}
$$

By Levy's Continuity Theorem $F_{S_n}(y) \to F_\mu(y)$ for all y where F_μ is continuous. But wait! F_μ is continuous except at μ:

$$
F_{S_n}(y) \to \begin{cases} 1 & \text{if } y > \mu \\ 0 & \text{if } y < \mu \end{cases}
$$

Let $\varepsilon > 0$. We need to show that:

$$
\lim_{n\to\infty} P(|S_n - \mu| \geq \varepsilon) = 0
$$

Well consider this as two separate events:

$$
\begin{aligned}
0 &\leq \lim_{n\to\infty} P(|S_n - \mu| \geq \varepsilon) \\
&= \lim_{n\to\infty} P(S_n \geq \mu + \varepsilon) + P(S_n \leq \mu - \varepsilon) \\
&\leq \lim_{n\to\infty} P\left(S_n > \mu + \frac{\varepsilon}{2}\right) + P(S_n \leq \mu - \varepsilon) \\
&= \lim_{n\to\infty} 1 - F_{S_n}\left(\mu + \frac{\varepsilon}{2}\right) + F_{S_n}(\mu - \varepsilon) \\
&= 1 - 1 + 0 = 0
\end{aligned}
$$

And so we get the squeeze theorem to get that S_n converges in probability to μ.

Example. Let X be normal, i.e. let X be continuous with the density

$f_X(t) = \frac{1}{\sqrt{2\pi}} e^{-t^2/2}$. We want to compute $\phi_X(t)$:

$$\begin{aligned}\phi_X(t) = \mathrm{E}\big[e^{itX}\big] &= \int_{-\infty}^{\infty} e^{itx} \cdot \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \,\mathrm{d}x \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{itx - x^2/2} \,\mathrm{d}x\end{aligned}$$

Instead of it, consider a real number $a \in \mathbb{R}$:

$$\begin{aligned}\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{ax - x^2/2} \,\mathrm{d}x &= \frac{1}{\sqrt{2\pi}} \cdot \int_{-\infty}^{\infty} e^{-(a-x)^2/2 + a^2/2} \,\mathrm{d}x \\ &= \frac{e^{a^2/2}}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-(a-x)^2/2} \,\mathrm{d}x \\ &= \frac{e^{a^2/2}}{\sqrt{2\pi}} \cdot \int_{-\infty}^{\infty} e^{-u^2/2} \,\mathrm{d}u \\ &= e^{a^2/2}\end{aligned}$$

Therefore the functions:

$$a \mapsto \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{ax} e^{-x^2/2} \,\mathrm{d}x \qquad\qquad a \mapsto e^{a^2/2}$$

Agree for all $a \in \mathbb{R}$. But these are defined for all $a \in \mathbb{C}$ and are analytic. By the principle of permanence, they agree for all $a \in \mathbb{C}$, in particular for $a = it$. Therefore we get the important formula:

$$\boxed{\phi_X(t) = \frac{1}{\sqrt{2\pi}} \cdot \int_{-\infty}^{\infty} e^{itX} e^{-x^2/2} \,\mathrm{d}x = e^{(it)^2/2} = e^{-t^2/2}}$$

Note! Any Random variable with $\phi_X(t) = e^{-t^2/2}$ is automaticall normal. In this case, ϕ_X is integrable and so X is continuous with density:

$$f_X(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-itx} e^{-x^2/2} \,\mathrm{d}x = \frac{1}{\sqrt{2\pi}} e^{-t^2/2}$$

Definition. *Let X be normal with $\mu \in \mathbb{R}$, $\sigma > 0$, then $Y := \sigma \cdot X + \mu$ is a*

Guassian Random Variable. We then have:

$$\mathrm{E}[Y] = \mu \qquad\qquad \mathrm{var}(Y) = \sigma^2$$

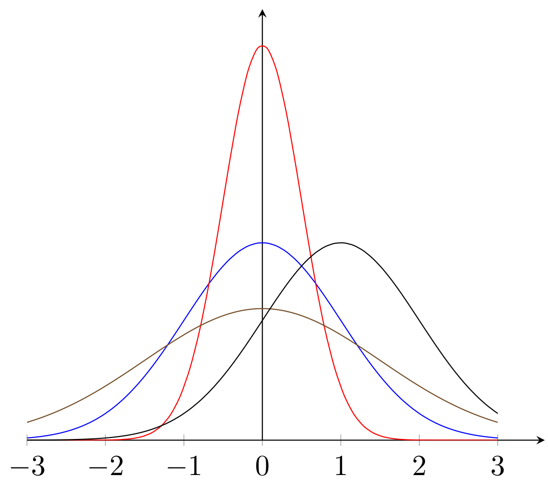
And also:

$$F_Y(t) = P(Y \le t) = P\left(X \le \frac{t-u}{\sigma}\right) = F_X\left(\frac{t-u}{\sigma}\right)$$

And tehrefore:

$$\begin{aligned} f_Y(t) = F_Y'(t) &= f_X\left(\frac{t-u}{\sigma}\right) \cdot \frac{1}{\sigma} \\ &= \frac{1}{\sigma \cdot \sqrt{2\pi}} e^{-\left(\frac{t-\mu}{\sigma\sqrt{2}}\right)^2} \end{aligned}$$

Here's the picture



And it has the characteristic function:

$$\begin{aligned} \phi_Y(t) = \mathrm{E}\left[e^{itY}\right] = \mathrm{E}\left[e^{it(\sigma X+\mu)}\right] &= \mathrm{E}\left[e^{it\mu} \cdot e^{it\sigma X}\right] \\ &= e^{it\mu} \cdot \phi_X(\sigma t) = e^{it\mu} \cdot e^{-(\sigma t)^2/2} \\ &= e^{it\mu - (\sigma t)^2/2} \end{aligned}$$

Corrolary. *If X, Y are Guassian and independent then $X+Y$ is also Gaus-*

sian and $\mathrm{E}[X+Y] = \mathrm{E}[X] + \mathrm{E}[Y]$ *and* $\mathrm{var}(X+Y) = \mathrm{var}(X) + \mathrm{var}(Y)$*.*

Proof. The claims about $\mathrm{E}[X+Y]$ and $\mathrm{var}(X+Y)$ are clear. To prove that $X+Y$ is Guassian, we compute the characteristic function:

$$\begin{aligned}\phi_{X+Y}(t) &= \phi_X(t) \cdot \phi_Y(t) \\ &= e^{it\mu_X + (\sigma_X t)^2/2} \cdot e^{it\mu_2 + (\sigma_2 t)^2/2} \\ &= e^{it(\mu_X+\mu_Y) + t^2/2 \cdot (\sigma_X^2 + \sigma_Y^2)}\end{aligned}$$

Proposition. *Here's an interesting property of Gaussians. Consider two independent identically distributed Random Variables* X, Y*. Assume that* $\mathrm{E}[X]$ *and* $\mathrm{var}(X)$ *exist and the distribution of* X *is symmetric. Assume that* $X+Y$ *and* $X-Y$ *are independent. Then* X *and* Y *are Gaussian:*

Proof. Let $\phi := \phi_X = \phi_Y$. Since the distribution of X is symmetric we have $\phi(t) = \phi(-t)$. The independence of X, Y gives:

$$\begin{aligned}\phi_{X+Y}(t) &= \phi_X(t) \cdot \phi_Y(t) = \phi(t)^2 \\ \phi_{X-Y}(t) &= \phi_X(t) \cdot \phi_Y(-t) = \phi(t)^2\end{aligned}$$

But then:

$$\begin{aligned}2X &= (X+Y) + (X-Y) \\ \phi_X(2t) = \phi_{2X}(t) &= \phi_{X+Y}(t) \cdot \phi_{X-Y}(t) = \phi(t)^4\end{aligned}$$

We may rewrite this as:

$$\phi_X(t) = \phi(t/2)^4$$

We then may iterate to get:

$$\phi_X(t) = \phi\left(\frac{t}{2^n}\right)^{4^n} = \lim_{n\to\infty} \phi\left(\frac{t}{2^n}\right)^{4^n}$$

The existence of $\mu := \mathrm{E}[X] = 0$ and $\sigma^2 := \mathrm{var}(X)$ gives that ϕ is twice differentiable at zero, that is:

$$\phi(t) = 1 + 0 \cdot t - \frac{\sigma^2 t^2}{2} + \alpha(t) \cdot t^2 \qquad\qquad \alpha(t) \to 0, t \to 0$$

Therefore we may write that:

$$\begin{aligned}\phi_X(t) &= \lim_{n\to\infty} \left(1 - \frac{\sigma^2 t^2}{2 \cdot 4^n} + \alpha\left(\frac{t}{2^n}\right) \cdot \frac{t^2}{4^n}\right)^{4^n} \\ &= e^{-\sigma^2 t^2/2}\end{aligned}$$

And so we are done! Wow!

Applications

Consider particles moving randomly on a plane. A 2D model of air particles in a room. Introduce a coordinate system and let X and Y be the velocities in the two directions of a randomly chosen particle. Physics tells us that X, Y are independent, identically distributed, and have symmetric distribution. If we rotate the coordinate system by 45°and let X', Y' be the new variables. Again X' and Y' are independent, but $X' = \frac{1}{\sqrt{2}}(X+Y)$ and $Y' = \frac{1}{\sqrt{2}}(X-Y)$. Therefore X and Y are Gaussian!!! Wow!

MATH 525 Notes

Faye Jackson

December 8, 2020

Final Review

Logistics

- Final is on Thursday December 17, 10:30-12:30pm (Ann Arbor Time) on Zoom, office hours zoom link (private meeting room)
- Modalities are as for the midterm
  - Video on
  - Stay muted
  - Cheat sheet of 5 pages that you can prepare
- Cumulative, 6 problems. 2-3 problems from the first half, and 3-4 problems from the second half.

Basic Concepts

- Probability Space $(\Omega, \mathcal{F}, P)$
  - Elements of Ω are samples, or atomic events. E.g., when you roll a die $\Omega = \{1, 2, 3, 4, 5, 6\}$.
  - General events are then just subsets of Ω which lie in the σ-alegbra $\mathcal{F} \subseteq \mathcal{P}(\Omega)$. When Ω is finite we can just take $\mathcal{F} = \mathcal{P}(\Omega)$. When Ω is uncountably infinite we often have to do something different, but for Ω a subset of $\mathbb{R}$ we usually take the Borel σ-algebra.

– P assigns probability to each event, i.e. $P : \mathcal{F} \to [0, 1]$. Furthermore it is σ-additive, that is:

$$P\left(\coprod_{k=1}^{\infty} A_k\right) \sum_{k=1}^{\infty} P(A_k)$$

In particular if Ω is countable then $P(\{\omega\})$, the atomic events, determine everything. For $\Omega = [0, 1]$, atomic events determine nothing.

- Random Variables. Formalized as a function $X : \Omega \to \mathbb{R}$

  – We intuitively think of X as a real number that is undecided. For $\omega \in \Omega$ then $X(\omega) \in \mathbb{R}$ is the value of X assuming "ω happened."

  – The expectation $\mathrm{E}[X]$ is the average of all possible values of X weighted by likelihood. If Ω is countable then:

$$\mathrm{E}[X] = \sum_{\omega \in \Omega} X(\omega) \cdot P(\{\omega\})$$

More generally, for any Ω, if X takes at most countably many values then:

$$\mathrm{E}[X] = \sum_{y \in \mathbb{R}} y \cdot P(X = y)$$

For a general X, we let $\mathrm{E}[X] - \int_{\Omega} X(\omega) \cdot \mathrm{d}P(\omega)$.

We say that X is integrable if there exists a sequence X_n of simple Random Variables (taking finitely many values) such that $X_n \to X$ almost surely, and for any $\varepsilon > 0$ there is an $N > 0$ so that for $n, m > N$ we have $\mathrm{E}[|X_n - X_m|] < \varepsilon$. We then define $\mathrm{E}[X] = \lim_{n\to\infty} \mathrm{E}[X_n]$.

Caution: Not every Random Variable is integrable.

  – We also have the distribution defined as $F_X(a) = P(X \leq a)$. These are defined for every random variable and has the properties that it's monotonically increasing from 0 to 1, right continuous, has left limits.

Need not be continuous, so it breaks the situation into two extremes:

1. When X takes countably many values, then F_X has a discontinuity at each of these values and is constant in between. We then call X discrete, and there is the function $f_X(a) = P(X = a)$ which determines everything, called the mass function. I.e. $\mathrm{E}[X] = \sum a \cdot f_X(a)$
2. F_X is absolutely continuous, this means that there exists an integrable density function f_X such that:

$$F_X(a) = \int_{-\infty}^{a} f_X(t)\,\mathrm{d}t$$

   This density determines everything, e.g. $\mathrm{E}[X] = \int_{-\infty}^{\infty} t \cdot f_X(t)\,\mathrm{d}t$. Note that the density is not uniquely determined, unless it can be chosen to be continuous.

What's important for these is to pay attention to the domains.

- Conditioning
  - For events A, B and $P(B) > 0$ we have:

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

    Intuitively we think of $P(A \mid B)$ as the probability that A happens given that B is known to happen
  - We also have $\mathrm{E}[X \mid B]$ which is the expectation of $X\big|_B$ with respect to $P_B = P(- \mid B)$.
  - The conditioning formulas for $\Omega = \coprod_{k=1}^{\infty} B_k$ where $P(B_k) > 0$

then we have:

$$P(A) = \sum_{k=1}^{\infty} P(A|B_k) \cdot P(B_k)$$
$$\mathrm{E}[X] = \sum_{k=1}^{\infty} \mathrm{E}[X \mid B_k] \cdot P(B_k)$$

Note Conditioning applies to any $Y = g(X)$ so we can use conditioning to compute $\mathrm{var}(X)$, generating function, and characteristic function.

– We may also condition on a random variable. Given X, Y random variables, we may define when Y is discrete that:

$$\begin{aligned} \mathrm{E}[X \mid Y] &: \mathbb{R} \to \mathbb{R} \\ y &\mapsto \begin{cases} \mathrm{E}[X \mid Y = y] & \text{if } P(Y = y) > 0 \\ 0 & \text{if } P(Y = y) = 0 \end{cases} \end{aligned}$$

This is a random variable on $(\mathbb{R}, \mathrm{Borel}, Y_* P)$, which isdefined as:

$$Y_* P(A) = \sum_{y \in A} P(Y = y)$$

We then have the useful formula that:

$$\mathrm{E}[\mathrm{E}[X \mid Y]] = \mathrm{E}[X]$$

We can also define it when X, Y have a joint density, by defining the conditional density $f_{X|Y}(x|y)$ as $f_{X,y}(x,y)/f_Y(y)$, defined when $f_Y(y) > 0$. Then:

$$\begin{aligned} \mathrm{E}[X \mid Y] &: \mathbb{R} \to \mathbb{R} \\ y &\mapsto \begin{cases} \int_{-\infty}^{\infty} x \cdot f_{X|Y}(x|y)\,\mathrm{d}x & \text{if } f_Y(y) > 0 \\ 0 & \text{otherwise} \end{cases} \end{aligned}$$

Again a random variable on $(\mathbb{R}, \mathrm{Borel}, Y_* P)$ where:

$$\mathrm{d}(Y_* P)(y) = f_Y(y) \cdot \mathrm{d}y$$

And again we have:

$$\mathrm{E}[\mathrm{E}[X \mid Y]] = \mathrm{E}[X]$$

- Independence
  - A, B are independent events provided that $P(A \cap B) = P(A) \cdot P(B)$, equivalently $P(A \mid B) = P(A)$
  - X, Y are independent random variables provided that $\{X \in A\}$ and $\{Y \in B\}$ are independent for any Borel sets $A, B \subseteq \mathbb{R}$. We can only check this on $\{X \leq a\}$ and $\{Y \leq b\}$ for $a, b \in \mathbb{R}$.

    When X, Y are discrete this is $f_X \cdot f_Y$ is a joint mass function, and when X, Y are continuous this is $f_X \cdot f_Y$ is a joint density:

$$P(X \leq a, Y \leq b) = \int_{-\infty}^{a} \int_{-\infty}^{b} f_{X,Y}(s,t)\, \mathrm{d}s\, \mathrm{d}t$$

Fun with Random Variables

- The discrete gang:
  - The Bernoulli (p): $f(1) = p$, $f(0) = 1 - p = q$, and $\mathrm{E}[X] = p$, $\mathrm{var}(X) = pq$
  - Binomial (n, p): $q := 1 - p$, $f(k) = \begin{cases} \binom{n}{k} p^k q^{n-k} & \text{if } 0 \leq k \leq n \\ 0 & \text{otherwise} \end{cases}$, and $\mathrm{E}[X] = np$ and $\mathrm{var}(X) = npq$
  - Geometric (p): $f(k) = q^{k-1} \cdot p$, $\mathrm{E}[X] = \frac{1}{p}$ and $\mathrm{var}(X) = \frac{q}{p^2}$
  - Poisson $(\lambda > 0)$: $f(k) = \begin{cases} \frac{\lambda^k}{k!} e^{-\lambda} & \text{if } k \geq 0 \\ 0 & \text{otherwise} \end{cases}$, $\mathrm{E}[X] = \lambda$ and $\mathrm{var}(X) = \lambda$.
- The continuous gang:

- Uniform on (a, b): $f(x) = \begin{cases} \frac{1}{b-a} & \text{if } a < x < b \\ 0 & \text{otherwise} \end{cases}$, and $\mathrm{E}[X] = \frac{a+b}{2}$ and $\mathrm{var}(X) = \frac{(a-b)^2}{12}$
- Exponential (λ): $f(x) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x > 0 \\ 0 & \text{otherwise} \end{cases}$., so $\mathrm{E}[X] = \frac{1}{\lambda}$, $\mathrm{var}\, 9X) = \frac{1}{\lambda^2}$
- Normal: $f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$, $\mathrm{E}[X] = 0$, $\mathrm{var}(X) = 1$
- Gaussian (μ, σ^2): $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/(2\sigma^2)}$, $\mathrm{E}[X] = \mu$, $\mathrm{var}(X) = \sigma^2$

- Generating Functions
  - Defined for $X : \Omega \to \mathbb{Z}_{\geq 0}$ as:

$$G_X(s) = \mathrm{E}\left[s^X\right] = \sum_{k=0}^{\infty} P(X = k) s^k$$

    G_X is analytic and converges at least for $|s| \leq 1$. We also know that
  - $G(0) = P(X = 0)$, $G(1) = 1$, $G'(1) = \mathrm{E}[X]$, $G''(1) + G'(1) - (G'(1))^2 = \mathrm{var}(X)$
  - If G_X is not presented as a power series $\sum a_k s^k$, then you can compute each a_k as:

$$a_k = \frac{G_X^{(k)}(0)}{k!}$$

  - If X, Y are independent then $G_{X+Y} = G_X \cdot G_Y$.
  - Bernoulli: $q + ps$, Binomial: $(q + ps)^n$, geometric: $\frac{p}{1-qs}$, Poisson: $e^{\lambda(s-1)}$

- Characteristic Functions

– For any X any random variable we define:

$$\begin{aligned}\phi_X : \mathbb{R} &\to \mathbb{C} \\ t &\mapsto \mathrm{E}\left[e^{itX}\right]\end{aligned}$$

But this has worse analytic properties

– $\phi(0) = 1$, $\phi'(0) = i\,\mathrm{E}[X]$, $\phi''(0) = -\,\mathrm{E}\left[X^2\right]$, $-\phi''(0) + (\phi'(0))^2 = \mathrm{var}(X)$.

– X, Y are independent then $\phi_{X+Y} = \phi_X \cdot \phi_Y$.

– If X has a generating function then $\phi_X(t) = G_X(e^{it})$

– If X is continuous with density f then:

$$\phi(t) = \int_{-\infty}^{\infty} e^{itx} f(x)\,\mathrm{d}x$$

– Conversely if ϕ is integrable then X is continuous and by Fourier inversion:

$$f_X(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-itx}\phi(t)\,\mathrm{d}t$$

– Levy's Continuity Theorem tells us that if $\phi_{X_n} \to \phi_X$ then X_n converges to X in distribution

– Some examples

  ∗ Uniform on (a, b): $\frac{2}{(b-a)t}\sin(t(b-a)/2)e^{i(a+b)t/2}$

  ∗ Exponential (λ): $\frac{\lambda}{\lambda - it}$

  ∗ Normal: $e^{-t^2/2}$

  ∗ Gaussian (μ, σ^2): $e^{it\mu-(t\sigma)^2/2}$

Convergence and Bounds

- Types of Convergence:

  1. $X_n \xrightarrow{a.s.} X$ when $P(\{\omega \in \Omega \mid X_n(\omega) \to X(\omega)\}) = 1$.

2. $X_n \xrightarrow{P} X$ when for every $\varepsilon > 0$ we have $\lim_{n\to\infty} P(|X_n - X| > \varepsilon) = 0$.
3. $X_n \xrightarrow{d} X$ when for every $y \in \mathbb{R}$ where F_X is continuous we have $F_{X_n}(y) \to F_X(y)$.

- We have $1 \implies 2 \implies 3$, but in general the converses fail.
- The Law of Large Numbers. If you have (X_n) independently identically distributed then:

$$\frac{1}{n}(X_1 + \cdots + X_n) \xrightarrow{a.s.} \mathrm{E}[X_1]$$

- The Central Limit Theorem. If you have (X_n) iid and $\mathrm{E}[X_1] = 0$ and $\mathrm{var}(X_1) = \sigma^2 < \infty$ then:

$$\frac{1}{\sqrt{n}}(X_1 + \cdots + X_n) \xrightarrow{d} N(0, \sigma^2)$$

- Variance and Covariance
  - We define:

$$\mathrm{cov}(X,Y) := \mathrm{E}[XY] - \mathrm{E}[X]\,\mathrm{E}[Y] \quad \mathrm{var}(X) := \mathrm{cov}(X,X)$$

  - If X, Y are independent implies $\mathrm{cov}(X,Y) = 0$, and if $\mathrm{var}(X) = 0$ then $X = \mathrm{E}[X]$ almost surely.
  - We also have $\sigma(X) = \sqrt{\mathrm{var}(X)}$ the standard deviation, which preserves units. We also can define:

$$\mathrm{cor}(X,Y) = \frac{\mathrm{cov}(X,Y)}{\sigma(X)\sigma(Y)}$$

    If we have $\mathrm{cor}(X,Y) = \pm 1$ then $aX + bY + c = 0$ for some $a, b, c \in \mathbb{R}$ not all zero almost surely.

- We then have:

$$\begin{aligned}\operatorname{var}\left(\sum_n X_n\right) &= \sum_n \operatorname{var}(X_n) + \sum_{n\neq m} \operatorname{cov}(X_n, X_m)\\ &= \sum_n \operatorname{var}(X_n) + 2\sum_{n<m} \operatorname{cov}(X_n, X_m)\end{aligned}$$

  In particular if these are independent then we have:

$$\operatorname{var}\left(\sum_n X_n\right) = \sum_n \operatorname{var}(X_n)$$

- Concentration Inequalities
  - Chebyshev's inequality says that:

$$P(|X - \mathrm{E}[X]| \geq a) \leq \frac{\operatorname{var}(X)}{a^2}$$

  - Bernstein's inequality: $X = X_1 + \cdots + X_n$ are independent and $\mathrm{E}[X_n] = 0$ and $|X_n| \leq 1$ then we have:

$$P(X \geq a) \leq e^{-a^2/2n} \qquad P(X \leq -a) \leq e^{-a^2/2n}$$