18.600: Lecture 1

Permutations and combinations, Pascal's triangle, learning to count

Scott Sheffield

MIT

My office hours: Wednesdays 3 to 5 in 2-249

Take a selfie with Norbert Wiener's desk.

Outline

Outline

Politics

- Suppose that, in some election, betting markets place the probability that your favorite candidate will be elected at 58 percent. Price of a contact that pays 100 dollars if your candidate wins is 58 dollars.
- Market seems to say that your candidate will probably win, if "probably" means with probability greater than .5.
- The price of such a contract may fluctuate in time.
- Let $X(t)$ denote the price at time t.
- Suppose $X(t)$ is known to vary continuously in time. What is probability p it reaches 59 before 57?
- If $p > .5$, we can make money in expecation by buying at 58 and selling when price hits 57 or 59.
- If $p < .5$, we can sell at 58 and buy when price hits 57 or 59.
- Efficient market hypothesis (a.k.a. "no free money just lying around" hypothesis) suggests $p = .5$ (with some caveats...)
- Natural model for prices: repeatedly toss coin, adding 1 for heads and -1 for tails, until price hits 0 or 100.

Which of these statements is "probably" true?

- 1. $X(t)$ will go below 50 at some future point.
- 2. $X(t)$ will get all the way below 20 at some point
- 3. $X(t)$ will reach both 70 and 30, at different future times.
- 4. $X(t)$ will reach both 65 and 35 at different future times.
- 5. $X(t)$ will hit 65, then 50, then 60, then 55.
- Answers: 1, 2, 4.
- Full explanations coming toward the end of the course.
- Problem sets in this course explore applications of probability to politics, medicine, finance, economics, science, engineering, philosophy, dating, etc. Stories motivate the math and make it easier to remember.
- Provocative question: what simple advice, that would greatly benefit humanity, are we unaware of? Foods to avoid? Exercises to do? Books to read? How would we know?
- Let's start with easier questions.

Outline

Outline

Remark, just for fun

Permutations

Counting tricks

Binomial coefficients

Problems

Permutations

- How many ways to order 52 cards?
- Answer: $52 \cdot 51 \cdot 50 \cdot \ldots \cdot 1 = 52! = 80658175170943878571660636856403766975289505600883277824 \times 10^{12}$
- n hats, n people, how many ways to assign each person a hat?
- Answer: $n!$
- n hats, $k < n$ people, how many ways to assign each person a hat?
- $n \cdot (n-1) \cdot (n-2) \ldots (n-k+1) = n!/(n-k)!$

Permutation notation

- A **permutation** is a function from $\{1, 2, \ldots, n\}$ to $\{1, 2, \ldots, n\}$ whose range is the whole set $\{1, 2, \ldots, n\}$. If σ is a permutation then for each j between 1 and n, the the value $\sigma(j)$ is the number that j gets mapped to.
- For example, if $n = 3$, then σ could be a function such that $\sigma(1) = 3$, $\sigma(2) = 2$, and $\sigma(3) = 1$.
- If you have n cards with labels 1 through n and you shuffle them, then you can let $\sigma(j)$ denote the label of the card in the jth position. Thus orderings of n cards are in one-to-one correspondence with permutations of n elements.
- One way to represent σ is to list the values $\sigma(1), \sigma(2), \ldots, \sigma(n)$ in order. The σ above is represented as $\{3, 2, 1\}$.
- If σ and ρ are both permutations, write $\sigma \circ \rho$ for their composition. That is, $\sigma \circ \rho(j) = \sigma(\rho(j))$.

Cycle decomposition

- Another way to write a permutation is to describe its cycles:
- For example, taking $n = 7$, we write $(2, 3, 5), (1, 7), (4, 6)$ for the permutation σ such that $\sigma(2) = 3, \sigma(3) = 5, \sigma(5) = 2$ and $\sigma(1) = 7, \sigma(7) = 1$, and $\sigma(4) = 6, \sigma(6) = 4$.
- If you pick some j and repeatedly apply σ to it, it will "cycle through" the numbers in its cycle.
- Visualize this by writing down numbers 1 to n and drawing arrow from each k to $\sigma(k)$. Trace through a cycle by following arrows.
- Generally, a function f is called an **involution** if $f(f(x)) = x$ for all x.
- A permutation is an involution if all cycles have length one or two.
- A permutation is "fixed point free" if there are no cycles of length one.

Outline

Remark, just for fun

Permutations

Counting tricks

Binomial coefficients

Problems

Outline

Remark, just for fun

Permutations

Counting tricks

Binomial coefficients

Problems

Fundamental counting trick

- n ways to assign hat for the first person. No matter what choice I make, there will remain $n-1$ ways to assign hat to the second person. No matter what choice I make there, there will remain $n-2$ ways to assign a hat to the third person, etc.
- This is a useful trick: break counting problem into a sequence of stages so that one always has the same number of choices to make at each stage. Then the total count becomes a product of number of choices available at each stage.
- Easy to make mistakes. For example, maybe in your problem, the number of choices at one stage actually *does* depend on choices made during earlier stages.

Another trick: overcount by a fixed factor

- If you have 5 indistinguishable black cards, 2 indistinguishable red cards, and three indistinguishable green cards, how many distinct shuffle patterns of the ten cards are there?
- Answer: if the cards were distinguishable, we'd have 10!. But we're overcounting by a factor of 5!2!3!, so the answer is 10!/(5!2!3!).

Outline

Remark, just for fun

Permutations

Counting tricks

Binomial coefficients

Problems

Outline

Remark, just for fun

Permutations

Counting tricks

Binomial coefficients

Problems

$\binom{n}{k}$ notation

- How many ways to choose an ordered sequence of k elements from a list of n elements, with repeats allowed?
- Answer: n^k
- How many ways to choose an ordered sequence of k elements from a list of n elements, with repeats forbidden?
- Answer: $n!/(n-k)!$
- How many way to choose (unordered) k elements from a list of n without repeats?
- Answer: $\binom{n}{k} := \frac{n!}{k!(n-k)!}$
- What is the coefficient in front of x^k in the expansion of $(x+1)^n$?
- Answer: $\binom{n}{k}$.

Pascal's triangle

- Arnold principle.
- A simple recursion: $\binom{n}{k} = \binom{n-1}{k-1} + \binom{n-1}{k}$.
- What is the coefficient in front of x^k in the expansion of $(x+1)^n$?
- Answer: $\binom{n}{k}$.
- $(x+1)^n = \binom{n}{0} \cdot 1 + \binom{n}{1}x^1 + \binom{n}{2}x^2 + \ldots + \binom{n}{n-1}x^{n-1} + \binom{n}{n}x^n$.
- Question: what is $\sum_{k=0}^n \binom{n}{k}$?
- Answer: $(1+1)^n = 2^n$.

Outline

Remark, just for fun

Permutations

Counting tricks

Binomial coefficients

Problems

Outline

Remark, just for fun

Permutations

Counting tricks

Binomial coefficients

Problems

More problems

- How many hands that have four cards of the same suit, one card of another suit?
- $4\binom{13}{4} \cdot 3\binom{13}{1}$
- How many 10 digit numbers with no consecutive digits that agree?
- If initial digit can be zero, have $10 \cdot 9^9$ ten-digit sequences. If initial digit required to be non-zero, have 9^{10}.
- How many ways to assign a birthday to each of 23 distinct people? What if no birthday can be repeated?
- 366^{23} if repeats allowed. $366!/343!$ if repeats not allowed.

More problems

- How many full house hands in poker?
- $13\binom{4}{3} \cdot 12\binom{4}{2}$
- How many "2 pair" hands?
- $13\binom{4}{2} \cdot 12\binom{4}{2} \cdot 11\binom{4}{1}/2$
- How many royal flush hands?
- 4

18.600: Lecture 2

Multinomial coefficients and more counting problems

Scott Sheffield

MIT

Outline

Outline

Multinomial coefficients

Integer partitions

More problems

Partition problems

- You have eight distinct pieces of food. You want to choose three for breakfast, two for lunch, and three for dinner. How many ways to do that?
- Answer: $8!/(3!2!3!)$
- One way to think of this: given any permutation of eight elements (e.g., 12435876 or 87625431) declare first three as breakfast, second two as lunch, last three as dinner. This maps set of $8!$ permutations on to the set of food-meal divisions in a many-to-one way: each food-meal division comes from $3!2!3!$ permutations.
- How many 8-letter sequences with 3 A's, 2 B's, and 3 C's?
- Answer: $8!/(3!2!3!)$. Same as other problem. Imagine 8 "slots" for the letters. Choose 3 to be A's, 2 to be B's, and 3 to be C's.

Partition problems

- In general, if you have n elements you wish to divide into r distinct piles of sizes $n_1, n_2 \ldots n_r$, how many ways to do that?
- Answer $\binom{n}{n_1, n_2, \ldots, n_r} := \frac{n!}{n_1! n_2! \ldots n_r!}$.

One way to understand the binomial theorem

- Expand the product $(A_1 + B_1)(A_2 + B_2)(A_3 + B_3)(A_4 + B_4)$.
- 16 terms correspond to 16 length-4 sequences of A's and B's.

$$A_1A_2A_3A_4 + A_1A_2A_3B_4 + A_1A_2B_3A_4 + A_1A_2B_3B_4+$$

$$A_1B_2A_3A_4 + A_1B_2A_3B_4 + A_1B_2B_3A_4 + A_1B_2B_3B_4+$$

$$B_1A_2A_3A_4 + B_1A_2A_3B_4 + B_1A_2B_3A_4 + B_1A_2B_3B_4+$$

$$B_1B_2A_3A_4 + B_1B_2A_3B_4 + B_1B_2B_3A_4 + B_1B_2B_3B_4$$

- What happens to this sum if we erase subscripts?
- $(A + B)^4 = B^4 + 4AB^3 + 6A^2B^2 + 4A^3B + A^4$. Coefficient of A^2B^2 is 6 because 6 length-4 sequences have 2 A's and 2 B's.
- Generally, $(A + B)^n = \sum_{k=0}^n \binom{n}{k} A^k B^{n-k}$, because there are $\binom{n}{k}$ sequences with k A's and $(n-k)$ B's.

How about trinomials?

- Expand $(A_1 + B_1 + C_1)(A_2 + B_2 + C_2)(A_3 + B_3 + C_3)(A_4 + B_4 + C_4)$. How many terms?
- Answer: 81, one for each length-4 sequence of A's and B's and C's.
- We can also compute $(A + B + C)^4 = A^4+4A^3B+6A^2B^2+4AB^3+B^4+4A^3C+12A^2BC+12AB^2C+4B^3C + 6A^2C^2 + 12ABC^2 + 6B^2C^2 + 4AC^3 + 4BC^3 + C^4$
- What is the sum of the coefficients in this expansion? What is the combinatorial interpretation of coefficient of, say, ABC^2?
- Answer $81 = (1 + 1 + 1)^4$. ABC^2 has coefficient 12 because there are 12 length-4 words have one A, one B, two C's.

Multinomial coefficients

- Is there a higher dimensional analog of binomial theorem?
- Answer: yes.
- Then what is it?
- $$(x_1+x_2+\ldots+x_r)^n = \sum_{n_1,\ldots,n_r : n_1+\ldots+n_r=n} \binom{n}{n_1, \ldots, n_r} x_1^{n_1} x_2^{n_2} \ldots x_r^{n_r}$$
- The sum on the right is taken over all collections $(n_1, n_2, \ldots, n_r)$ of r non-negative integers that add up to n.
- Pascal's triangle gives coefficients in binomial expansions. Is there something like a "Pascal's pyramid" for trinomial expansions?
- Yes (look it up) but it is a bit tricker to draw and visualize than Pascal's triangle.

By the way...

- If $n!$ is the product of all integers in the interval with endpoints 1 and n, then $0! = 0$.
- Actually, we say $0! = 1$. What are the reasons for that?
- **Because** there is one map from the empty set to itself.
- **Because** we want the formula $\binom{n}{k} = \frac{n!}{k!(n-k)!}$ to still make sense when $k = 0$ and $k = n$. There is clearly 1 way to choose n elements from a group of n elements. And 1 way to choose 0 elements from a group of n elements so $\frac{n!}{n!0!} = \frac{n!}{0!n!} = 1$.
- **Because** we want the recursion $n(n-1)! = n!$ to hold for $n = 1$. (We won't define factorials of negative integers.)
- **Because** we want $n! = \int_0^\infty t^n e^{-t} dt$ to hold for all non-negative integers. (Check for positive integers by integration by parts.) This is one of those formulas you should just know. Can use it to define $n!$ for non-integer n.
- Another common notation: write $\Gamma(z) := \int_0^\infty t^{z-1} e^{-t} dt$ and define $n! := \Gamma(n+1) = \int_0^\infty t^n e^{-t} dt$, so that $\Gamma(n) = (n-1)!$.

Outline

Outline

Integer partitions

- How many sequences $a_1, \ldots, a_k$ of non-negative integers satisfy $a_1 + a_2 + \ldots + a_k = n$?
- Answer: $\binom{n+k-1}{n}$. Represent partition by $k-1$ bars and n stars, e.g., as $**|**||****|*$.

Outline

Multinomial coefficients

Integer partitions

More problems

Outline

Multinomial coefficients

Integer partitions

More problems

More counting problems

- In 18.821, a class of 27 students needs to be divided into 9 teams of three students each? How many ways are there to do that?
- $\frac{27!}{(3!)^9 9!}$
- You teach a class with 90 students. In a rather severe effort to combat grade inflation, your department chair insists that you assign the students exactly 10 A's, 20 B's, 30 C's, 20 D's, and 10 F's. How many ways to do this?
- $\binom{90}{10,20,30,20,10} = \frac{90!}{10!20!30!20!10!}$
- You have 90 (indistinguishable) pieces of pizza to divide among the 90 (distinguishable) students. How many ways to do that (giving each student a non-negative integer number of slices)?
- $\binom{179}{90} = \binom{179}{89}$

More counting problems

- How many 13-card bridge hands have 4 of one suit, 3 of one suit, 5 of one suit, 1 of one suit?
- $4!\binom{13}{4}\binom{13}{3}\binom{13}{5}\binom{13}{1}$
- How many bridge hands have at most two suits represented?
- $\binom{4}{2}\binom{26}{13} - 8$
- How many hands have either 3 or 4 cards in each suit?
- Need three 3-card suits, one 4-card suit, to make 13 cards total. Answer is $4\binom{13}{3}^3\binom{13}{4}$

18.600: Lecture 3
What is probability?

Scott Sheffield

MIT

Outline

Outline

What does "I'd say there's a thirty percent chance it will rain tomorrow" mean?

- **Neurological:** When I think "it will rain tomorrow" the "truth-sensing" part of my brain exhibits 30 percent of its maximum electrical activity.
- **Frequentist:** Of the last 1000 days that meteorological measurements looked this way, rain occurred on the subsequent day 300 times.
- **Market preference ("risk neutral probability"):** The market price of a contract that pays 100 if it rains tomorrow agrees with the price of a contract that pays 30 tomorrow no matter what.
- **Personal belief:** If you offered *me* a choice of these contracts, I'd be indifferent. (If need for money is different in two scenarios, I can replace dollars with "units of utility.")

Outline

Formalizing probability

Sample space

DeMorgan's laws

Axioms of probability

Outline

Formalizing probability

Sample space

DeMorgan's laws

Axioms of probability

Even more fundamental question: defining a set of possible outcomes

- Roll a die n times. Define a **sample space** to be $\{1,2,3,4,5,6\}^n$, i.e., the set of $a_1, \ldots, a_n$ with each $a_j \in \{1,2,3,4,5,6\}$.
- Shuffle a standard deck of cards. Sample space is the set of 52! permutations.
- Will it rain tomorrow? Sample space is $\{R, N\}$, which stand for "rain" and "no rain."
- Randomly throw a dart at a board. Sample space is the set of points on the board.

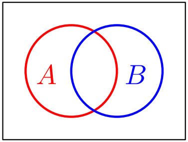
Event: subset of the sample space

- If a set A is comprised of some of the elements of B, say A is a **subset** of B and write $A \subset B$.
- Similarly, $B \supset A$ means A is a subset of B (or B is a superset of A).
- If S is a finite sample space with n elements, then there are 2^n subsets of S.
- Denote by $\emptyset$ the set with no elements.

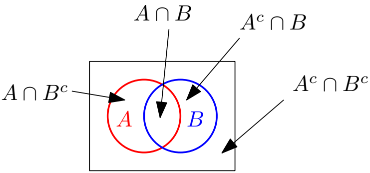
Intersections, unions, complements

- $A \cup B$ means the union of A and B, the set of elements contained in at least one of A and B.
- $A \cap B$ means the intersection of A and B, the set of elements contained on both A and B.
- A^c means complement of A, set of points in whole sample space S but not in A.
- $A \setminus B$ means "A minus B" which means the set of points in A but not in B. In symbols, $A \setminus B = A \cap (B^c)$.
- $\cup$ is associative. So $(A \cup B) \cup C = A \cup (B \cup C)$ and can be written $A \cup B \cup C$.
- $\cap$ is also associative. So $(A \cap B) \cap C = A \cap (B \cap C)$ and can be written $A \cap B \cap C$.

Venn diagrams



Venn diagrams



Outline

Outline

Formalizing probability

Sample space

DeMorgan's laws

Axioms of probability

DeMorgan's laws

- "It will not snow or rain" means "It will not snow and it will not rain."
- If S is event that it snows, R is event that it rains, then $(S \cup R)^c = S^c \cap R^c$
- More generally: $(\cup_{i=1}^n E_i)^c = \cap_{i=1}^n (E_i)^c$
- "It will not both snow and rain" means "Either it will not snow or it will not rain."
- $(S \cap R)^c = S^c \cup R^c$
- $(\cap_{i=1}^n E_i)^c = \cup_{i=1}^n (E_i)^c$

Outline

Formalizing probability

Sample space

DeMorgan's laws

Axioms of probability

Outline

Formalizing probability

Sample space

DeMorgan's laws

Axioms of probability

Axioms of probability

- $P(A) \in [0,1]$ for all $A \subset S$.
- $P(S) = 1$.
- Finite additivity: $P(A \cup B) = P(A) + P(B)$ if $A \cap B = \emptyset$.
- Countable additivity: $P(\cup_{i=1}^{\infty} E_i) = \sum_{i=1}^{\infty} P(E_i)$ if $E_i \cap E_j = \emptyset$ for each pair i and j.

- **Neurological:** When I think "it will rain tomorrow" the "truth-sensing" part of my brain exhibits 30 percent of its maximum electrical activity. Should have $P(A) \in [0,1]$ and presumably $P(S) = 1$ but not necessarily $P(A \cup B) = P(A) + P(B)$ when $A \cap B = \emptyset$.
- **Frequentist:** $P(A)$ is the fraction of times A occurred during the previous (large number of) times we ran the experiment. Seems to satisfy axioms...
- **Market preference ("risk neutral probability"):** $P(A)$ is price of contract paying dollar if A occurs divided by price of contract paying dollar regardless. Seems to satisfy axioms, assuming no arbitrage, no bid-ask spread, complete market...
- **Personal belief:** $P(A)$ is amount such that I'd be indifferent between contract paying 1 if A occurs and contract paying $P(A)$ no matter what. Seems to satisfy axioms with some notion of utility units, strong assumption of "rationality"...

18.600: Lecture 4

Axioms of probability and inclusion-exclusion

Scott Sheffield

MIT

Outline

Outline

Axioms of probability

- $P(A) \in [0,1]$ for all $A \subset S$.
- $P(S) = 1$.
- Finite additivity: $P(A \cup B) = P(A) + P(B)$ if $A \cap B = \emptyset$.
- Countable additivity: $P(\cup_{i=1}^{\infty} E_i) = \sum_{i=1}^{\infty} P(E_i)$ if $E_i \cap E_j = \emptyset$ for each pair i and j.

- **Neurological:** When I think "it will rain tomorrow" the "truth-sensing" part of my brain exhibits 30 percent of its maximum electrical activity.
- **Frequentist:** $P(A)$ is the fraction of times A occurred during the previous (large number of) times we ran the experiment.
- **Market preference ("risk neutral probability"):** $P(A)$ is price of contract paying dollar if A occurs divided by price of contract paying dollar regardless.
- **Personal belief:** $P(A)$ is amount such that I'd be indifferent between contract paying 1 if A occurs and contract paying $P(A)$ no matter what.

- **Neurological:** When I think "it will rain tomorrow" the "truth-sensing" part of my brain exhibits 30 percent of its maximum electrical activity. Should have $P(A) \in [0, 1]$, maybe $P(S) = 1$, not necessarily $P(A \cup B) = P(A) + P(B)$ when $A \cap B = \emptyset$.
- **Frequentist:** $P(A)$ is the fraction of times A occurred during the previous (large number of) times we ran the experiment. Seems to satisfy axioms...
- **Market preference ("risk neutral probability"):** $P(A)$ is price of contract paying dollar if A occurs divided by price of contract paying dollar regardless. Seems to satisfy axioms, assuming no arbitrage, no bid-ask spread, complete market...
- **Personal belief:** $P(A)$ is amount such that I'd be indifferent between contract paying 1 if A occurs and contract paying $P(A)$ no matter what. Seems to satisfy axioms with some notion of utility units, strong assumption of "rationality"...

Axiom breakdown

- What if personal belief function doesn't satisfy axioms?
- Consider an A-contract (pays 10 if candidate A wins election) a B-contract (pays 10 dollars if candidate B wins) and an A-or-B contract (pays 10 if either A or B wins).
- Friend: "I'd say A-contract is worth 1 dollar, B-contract is worth 1 dollar, A-or-B contract is worth 7 dollars."
- **Amateur response:** "Dude, that is, like, so messed up. Haven't you heard of the axioms of probability?"
- **Cynical professional response:** "I fully understand and respect your opinions. In fact, let's do some business. You sell me an A contract and a B contract for 1.50 each, and I sell you an A-or-B contract for 6.50."
- Friend: "Wow... you've beat by suggested price by 50 cents on each deal. Yes, sure! You're a great friend!"
- Axioms breakdowns are money-making opportunities.

Outline

Outline

Intersection notation

- We will sometimes write AB to denote the event $A \cap B$.

Consequences of axioms

- Can we show from the axioms that $P(A^c) = 1 - P(A)$?
- Can we show from the axioms that if $A \subset B$ then $P(A) \leq P(B)$?
- Can we show from the axioms that $P(A \cup B) = P(A) + P(B) - P(AB)$?
- Can we show from the axioms that $P(AB) \leq P(A)$?
- Can we show from the axioms that if S contains finitely many elements $x_1, \ldots, x_k$, then the values $\big(P(\{x_1\}), P(\{x_2\}), \ldots, P(\{x_k\})\big)$ determine the value of $P(A)$ for any $A \subset S$?
- What k-tuples of values are consistent with the axioms?

Famous 1982 Tversky-Kahneman study (see wikipedia)

- People are told "Linda is 31 years old, single, outspoken, and very bright. She majored in philosophy. As a student, she was deeply concerned with issues of discrimination and social justice, and also participated in anti-nuclear demonstrations."
- They are asked: Which is more probable?
  - Linda is a bank teller.
  - Linda is a bank teller and is active in the feminist movement.
- 85 percent chose the second option.
- Could be correct using neurological/emotional definition. Or a "which story would you believe" interpretation (if witnesses offering more details are considered more credible).
- But axioms of probability imply that second option cannot be more likely than first.

Outline

Axions of probability

Consequences of axioms

Inclusion exclusion

Outline

Axioms of probability

Consequences of axioms

Inclusion exclusion

Inclusion-exclusion identity

- Imagine we have n events, $E_1, E_2, \ldots, E_n$.
- How do we go about computing something like $P(E_1 \cup E_2 \cup \ldots \cup E_n)$?
- It may be quite difficult, depending on the application.
- There are some situations in which computing $P(E_1 \cup E_2 \cup \ldots \cup E_n)$ is a priori difficult, but it is relatively easy to compute probabilities of *intersections* of any collection of E_i. That is, we can easily compute quantities like $P(E_1 E_3 E_7)$ or $P(E_2 E_3 E_6 E_7 E_8)$.
- In these situations, the inclusion-exclusion rule helps us compute unions. It gives us a way to express $P(E_1 \cup E_2 \cup \ldots \cup E_n)$ in terms of these intersection probabilities.

Inclusion-exclusion identity

- Can we show from the axioms that $P(A \cup B) = P(A) + P(B) - P(AB)$?
- How about $P(E \cup F \cup G) = P(E) + P(F) + P(G) - P(EF) - P(EG) - P(FG) + P(EFG)$?
- More generally,

$$P(\cup_{i=1}^n E_i) = \sum_{i=1}^n P(E_i) - \sum_{i_1 < i_2} P(E_{i_1} E_{i_2}) + \ldots$$
$$+ (-1)^{(r+1)} \sum_{i_1 < i_2 < \ldots < i_r} P(E_{i_1} E_{i_2} \ldots E_{i_r})$$
$$+ \ldots + (-1)^{n+1} P(E_1 E_2 \ldots E_n).$$

- The notation $\sum_{i_1 < i_2 < \ldots < i_r}$ means a sum over all of the $\binom{n}{r}$ subsets of size r of the set $\{1, 2, \ldots, n\}$.

Inclusion-exclusion proof idea

- Consider a region of the Venn diagram contained in exactly $m > 0$ subsets. For example, if $m = 3$ and $n = 8$ we could consider the region $E_1 E_2 E_3^c E_4^c E_5 E_6^c E_7^c E_8^c$.
- This region is contained in three single intersections (E_1, E_2, and E_5). It's contained in 3 double-intersections ($E_1 E_2$, $E_1 E_5$, and $E_2 E_5$). It's contained in only 1 triple-intersection ($E_1 E_2 E_5$).
- It is counted $\binom{m}{1} - \binom{m}{2} + \binom{m}{3} + \ldots \pm \binom{m}{m}$ times in the inclusion exclusion sum.
- How many is that?
- Answer: 1. (Follows from binomial expansion of $(1-1)^m$.)
- Thus each region in $E_1 \cup \ldots \cup E_n$ is counted exactly once in the inclusion exclusion sum, which implies the identity.

18.600: Lecture 5

Problems with all outcomes equally likely, including a famous hat problem

Scott Sheffield

MIT

Outline

Outline

Equal likelihood

- If a sample space S has n elements, and all of them are equally likely, then each one has to have probability $1/n$
- What is $P(A)$ for a general set $A \subset S$?
- Answer: $|A|/|S|$, where $|A|$ is the number of elements in A.

Outline

Equal likelihood

A few problems

Hat problem

A few more problems

Outline

Equal likelihood

A few problems

Hat problem

A few more problems

Problems

- Roll two dice. What is the probability that their sum is three?
- $2/36 = 1/18$
- Toss eight coins. What is the probability that exactly five of them are heads?
- $\binom{8}{5}/2^8$
- In a class of 100 people with cell phone numbers, what is the probability that nobody has a number ending in 37?
- $(99/100)^{100} \approx 1/e$
- Roll ten dice. What is the probability that a 6 appears on exactly five of the dice?
- $\binom{10}{5}5^5/6^{10}$
- In a room of 23 people, what is the probability that two of them have a birthday in common?
- $1 - \prod_{i=0}^{22} \frac{365-i}{365}$

Outline

Equal likelihood

A few problems

Hat problem

A few more problems

Outline

Equal likelihood

A few problems

Hat problem

A few more problems

Recall the inclusion-exclusion identity

- $$P(\cup_{i=1}^n E_i) = \sum_{i=1}^n P(E_i) - \sum_{i_1<i_2} P(E_{i_1}E_{i_2}) + \ldots$$
  $$+ (-1)^{(r+1)} \sum_{i_1<i_2<\ldots<i_r} P(E_{i_1}E_{i_2}\ldots E_{i_r})$$
  $$= + \ldots + (-1)^{n+1} P(E_1E_2\ldots E_n).$$
- The notation $\sum_{i_1<i_2<i_r}$ means a sum over all of the $\binom{n}{r}$ subsets of size r of the set $\{1,2,\ldots,n\}$.

Famous hat problem

- n people toss hats into a bin, randomly shuffle, return one hat to each person. Find probability nobody gets own hat.
- Inclusion-exclusion. Let E_i be the event that ith person gets own hat.
- What is $P(E_{i_1}E_{i_2}\ldots E_{i_r})$?
- Answer: $\frac{(n-r)!}{n!}$.
- There are $\binom{n}{r}$ terms like that in the inclusion exclusion sum. What is $\binom{n}{r}\frac{(n-r)!}{n!}$?
- Answer: $\frac{1}{r!}$.
- $P(\cup_{i=1}^n E_i) = 1 - \frac{1}{2!} + \frac{1}{3!} - \frac{1}{4!} + \ldots \pm \frac{1}{n!}$
- $1 - P(\cup_{i=1}^n E_i) = 1 - 1 + \frac{1}{2!} - \frac{1}{3!} + \frac{1}{4!} - \ldots \pm \frac{1}{n!} \approx 1/e \approx .36788$

Outline

Equal likelihood

A few problems

Hat problem

A few more problems

Outline

Equal likelihood

A few problems

Hat problem

A few more problems

Problems

- What's the probability of a full house in poker (i.e., in a five card hand, 2 have one value and three have another)?
- Answer 1:
$$\frac{\#\text{ ordered distinct-five-card sequences giving full house}}{\#\text{ ordered distinct-five-card sequences}}$$
- That's $\binom{5}{2} * 13 * 12 * (4 * 3 * 2) * (4 * 3) / (52 * 51 * 50 * 49 * 48) = 6/4165$.
- Answer 2:
$$\frac{\#\text{ unordered distinct-five-card sets giving full house}}{\#\text{ unordered distinct-five-card sets}}$$
- That's $13 * 12 * \binom{4}{3} * \binom{4}{2} / \binom{52}{5} = 6/4165$.
- What is the probability of a two-pair hand in poker?
- Fix suit breakdown, then face values: $\binom{4}{2} \cdot 2 \cdot \binom{13}{2}\binom{13}{2} \cdot 13 / \binom{52}{5}$
- How about bridge hand with 3 of one suit, 3 of one suit, 2 of one suit, 5 of another suit?
- $\binom{4}{2} \cdot 2 \cdot \binom{13}{3}\binom{13}{3}\binom{13}{2}\binom{13}{5} / \binom{52}{13}$

18.600: Lecture 6

Conditional probability

Scott Sheffield

MIT

Outline

Definition: probability of A given B

Examples

Multiplication rule

Outline

Definition: probability of A given B

Examples

Multiplication rule

Conditional probability

- Suppose I have a sample space S with n equally likely elements, representing possible outcomes of an experiment.
- Experiment is performed, but I don't know outcome. For some $F \subset S$, I ask, "Was the outcome in F?" and receive answer yes.
- I think of F as a "new sample space" with all elements equally likely.
- Definition: $P(E|F) = P(EF)/P(F)$.
- Call $P(E|F)$ the "conditional probability of E given F" or "probability of E conditioned on F".
- Definition makes sense even without "equally likely" assumption.

Outline

Definition: probability of A given B

Examples

Multiplication rule

Outline

Definition: probability of A given B

Examples

Multiplication rule

More examples

- Probability have rare disease given positive result to test with 90 percent accuracy.
- Say probability to have disease is p.
- $S = \{\text{disease, no disease}\} \times \{\text{positive, negative}\}$.
- $P(\text{positive}) = .9p + .1(1-p)$ and $P(\text{disease, positive}) = .9p$.
- $P(\text{disease}|\text{positive}) = \frac{.9p}{.9p+.1(1-p)}$. If p is tiny, this is about $9p$.
- Probability suspect guilty of murder given a particular suspicious behavior.
- Probability plane will come eventually, given plane not here yet.

Another famous Tversky/Kahneman study (Wikipedia)

- Imagine you are a member of a jury judging a hit-and-run driving case. A taxi hit a pedestrian one night and fled the scene. The entire case against the taxi company rests on the evidence of one witness, an elderly man who saw the accident from his window some distance away. He says that he saw the pedestrian struck by a blue taxi. In trying to establish her case, the lawyer for the injured pedestrian establishes the following facts:
  - There are only two taxi companies in town, "Blue Cabs" and "Green Cabs." On the night in question, 85 percent of all taxis on the road were green and 15 percent were blue.
  - The witness has undergone an extensive vision test under conditions similar to those on the night in question, and has demonstrated that he can successfully distinguish a blue taxi from a green taxi 80 percent of the time.
- Study participants believe blue taxi at fault, say witness correct with 80 percent probability.

Outline

Definition: probability of A given B

Examples

Multiplication rule

Outline

Definition: probability of A given B

Examples

Multiplication rule

Multiplication rule

- $P(E_1E_2E_3 \ldots E_n) = P(E_1)P(E_2|E_1)P(E_3|E_1E_2) \ldots P(E_n|E_1 \ldots E_{n-1})$
- Useful when we think about multi-step experiments.
- For example, let E_i be event ith person gets own hat in the n-hat shuffle problem.
- Another example: roll die and let E_i be event that the roll *does not lie* in $\{1, 2, \ldots, i\}$. Then $P(E_i) = (6 - i)/6$ for $i \in \{1, 2, \ldots, 6\}$.
- What is $P(E_4|E_1E_2E_3)$ in this case?

Monty Hall problem

- Prize behind one of three doors, all equally likely.
- You point to door one. Host opens either door two or three and shows you that it doesn't have a prize. (If neither door two nor door three has a prize, host tosses coin to decide which to open.)
- You then get to open a door and claim what's behind it. Should you stick with door one or choose other door?
- Sample space is $\{1, 2, 3\} \times \{2, 3\}$ (door containing prize, door host points to).
- We have $P\big((1, 2)\big) = P\big((1, 3)\big) = 1/6$ and $P\big((2, 3)\big) = P\big((3, 2)\big) = 1/3$. Given host points to door 2, probability prize behind 3 is $2/3$.

Another popular puzzle (see Tanya Khovanova's blog)

- Given that your friend has exactly two children, one of whom is a son born on a Tuesday, what is the probability the second child is a son.
- Make the obvious (though not quite correct) assumptions. Every child is either boy or girl, and equally likely to be either one, and all days of week for birth equally likely, etc.
- Make state space matrix of $196 = 14 \times 14$ elements
- Easy to see answer is $13/27$.

18.600: Lecture 7

Bayes' formula and independence

Scott Sheffield

MIT

Outline

Bayes' formula

Independence

Outline

Bayes' formula

Independence

Recall definition: conditional probability

- Definition: $P(E|F) = P(EF)/P(F)$.
- Equivalent statement: $P(EF) = P(F)P(E|F)$.
- Call $P(E|F)$ the "conditional probability of E given F" or "probability of E conditioned on F".

Dividing probability into two cases

- $$\begin{aligned} P(E) &= P(EF) + P(EF^c) \\ &= P(E|F)P(F) + P(E|F^c)P(F^c) \end{aligned}$$
- In words: want to know the probability of E. There are two scenarios F and F^c. If I know the probabilities of the two scenarios and the probability of E conditioned on each scenario, I can work out the probability of E.
- Example: $D =$ "have disease", $T =$ "positive test."
- If $P(D) = p$, $P(T|D) = .9$, and $P(T|D^c) = .1$, then $P(T) = .9p + .1(1-p)$.
- What is $P(D|T)$?

Bayes' theorem

- Bayes' theorem/law/rule states the following: $P(A|B) = \frac{P(B|A)P(A)}{P(B)}$.
- Follows from definition of conditional probability: $P(AB) = P(B)P(A|B) = P(A)P(B|A)$.
- Tells how to update estimate of probability of A when new evidence restricts your sample space to B.
- So $P(A|B)$ is $\frac{P(B|A)}{P(B)}$ times $P(A)$.
- Ratio $\frac{P(B|A)}{P(B)}$ determines "how compelling new evidence is".
- What does it mean if ratio is zero?
- What if ratio is $1/P(A)$?

Bayes' theorem

- Bayes' formula $P(A|B) = \frac{P(B|A)P(A)}{P(B)}$ is often invoked as tool to guide intuition.
- Example: A is event that suspect stole the $\$10,000$ under my mattress, B is event that suspect deposited several thousand dollars in cash in bank last week.
- Begin with subjective estimates of $P(A)$, $P(B|A)$, and $P(B|A^c)$. Compute $P(B)$. Check whether B occurred. Update estimate.
- Repeat procedure as new evidence emerges.
- Caution required. My idea to check whether B occurred, or is a lawyer selecting the provable events $B_1, B_2, B_3, \ldots$ that maximize $P(A|B_1B_2B_3\ldots)$? Where did my probability estimates come from? What is my state space? What assumptions am I making?

"Bayesian" sometimes used to describe philosophical view

- Philosophical idea: we assign subjective probabilities to questions we can't answer. Will candidate win election? Will Red Sox win world series? Will stock prices go up this year?
- Bayes essentially described probability of event as
$$\frac{\text{value of right to get some thing if event occurs}}{\text{value of thing}}.$$
- Philosophical questions: do we have subjective probabilities/hunches for questions we can't base enforceable contracts on? Do there exist other universes? Are there other intelligent beings? Are there beings smart enough to simulate universes like ours? Are we part of such a simulation?...
- Do we use Bayes subconsciously to update hunches?
- Should we think of Bayesian priors and updates as part of the epistemological foundation of science and statistics?

Updated "odds"

- Define "odds" of A to be $P(A)/P(A^c)$.
- Define "conditional odds" of A given B to be $P(A|B)/P(A^c|B)$.
- Is there nice way to describe ratio between odds and conditional odds?
- $\frac{P(A|B)/P(A^c|B)}{P(A)/P(A^c)} = ?$
- By Bayes $P(A|B)/P(A) = P(B|A)/P(B)$.
- After some algebra, $\frac{P(A|B)/P(A^c|B)}{P(A)/P(A^c)} = P(B|A)/P(B|A^c)$
- Say I think A is 5 times as likely as A^c, and $P(B|A) = 3P(B|A^c)$. Given B, I think A is 15 times as likely as A^c.
- Gambling sites (look at oddschecker.com) often list $P(A^c)/P(A)$, which is basically amount house puts up for bet on A^c when you put up one dollar for bet on A.

$P(\cdot|F)$ is a probability measure

- We can check the probability axioms: $0 \leq P(E|F) \leq 1$, $P(S|F) = 1$, and $P(\cup E_i|F) = \sum P(E_i|F)$, if i ranges over a countable set and the E_i are disjoint.
- The probability measure $P(\cdot|F)$ is related to $P(\cdot)$.
- To get former from latter, we set probabilities of elements outside of F to zero and multiply probabilities of events inside of F by $1/P(F)$.
- It $P(\cdot)$ is the *prior* probability measure and $P(\cdot|F)$ is the *posterior* measure (revised after discovering that F occurs).

Outline

Outline

Independence

- Say E and F are **independent** if $P(EF) = P(E)P(F)$.
- Equivalent statement: $P(E|F) = P(E)$. Also equivalent: $P(F|E) = P(F)$.
- Example: toss two coins. Sample space contains four equally likely elements $(H,H),(H,T),(T,H),(T,T)$.
- Is event that first coin is heads independent of event that second coin heads.
- Yes: probability of each event is $1/2$ and probability of both is $1/4$.
- Is event that first coin is heads independent of event that number of heads is odd?
- Yes: probability of each event is $1/2$ and probability of both is $1/4$...
- despite fact that (in everyday English usage of the word) oddness of the number of heads "depends" on the first coin.

Independence: another example

- Shuffle 4 cards with labels 1 through 4. Let $E_{j,k}$ be event that card j comes before card k. Is $E_{1,2}$ independent of $E_{3,4}$?
- Is $E_{1,2}$ independent of $E_{1,3}$?
- No. In fact, what is $P(E_{1,2}|E_{1,3})$?
- $2/3$
- Generalize to $n > 7$ cards. What is $P(E_{1,7}|E_{1,2}E_{1,3}E_{1,4}E_{1,5}E_{1,6})$?
- $6/7$

Independence of multiple events

- Say $E_1 \dots E_n$ are independent if for each $\{i_1, i_2, \dots, i_k\} \subset \{1, 2, \dots n\}$ we have $P(E_{i_1}E_{i_2} \dots E_{i_k}) = P(E_{i_1})P(E_{i_2}) \dots P(E_{i_k})$.
- In other words, the product rule works.
- Independence implies $P(E_1E_2E_3|E_4E_5E_6) = \frac{P(E_1)P(E_2)P(E_3)P(E_4)P(E_5)P(E_6)}{P(E_4)P(E_5)P(E_6)} = P(E_1E_2E_3)$, and other similar statements.
- Does pairwise independence imply independence?
- No. Consider these three events: first coin heads, second coin heads, odd number heads. Pairwise independent, not independent.

18.600: Lecture 8

Discrete random variables

Scott Sheffield

MIT

Outline

Outline

Random variables

- A random variable X is a function from the state space to the real numbers.
- Can interpret X as a quantity whose value depends on the outcome of an experiment.
- Example: toss n coins (so state space consists of the set of all 2^n possible coin sequences) and let X be number of heads.
- Question: What is $P\{X = k\}$ in this case?
- Answer: $\binom{n}{k}/2^n$, if $k \in \{0, 1, 2, \ldots, n\}$.

Independence of multiple events

- In n coin toss example, knowing the values of some coin tosses tells us nothing about the others.
- Say $E_1 \ldots E_n$ are independent if for each $\{i_1, i_2, \ldots, i_k\} \subset \{1, 2, \ldots n\}$ we have $P(E_{i_1} E_{i_2} \ldots E_{i_k}) = P(E_{i_1}) P(E_{i_2}) \ldots P(E_{i_k})$.
- In other words, the product rule works.
- Independence implies $P(E_1 E_2 E_3 | E_4 E_5 E_6) = \frac{P(E_1)P(E_2)P(E_3)P(E_4)P(E_5)P(E_6)}{P(E_4)P(E_5)P(E_6)} = P(E_1 E_2 E_3)$, and other similar statements.
- Does pairwise independence imply independence?
- No. Consider these three events: first coin heads, second coin heads, odd number heads. Pairwise independent, not independent.

Examples

- Shuffle n cards, and let X be the position of the jth card. State space consists of all $n!$ possible orderings. X takes values in $\{1, 2, \ldots, n\}$ depending on the ordering.
- Question: What is $P\{X = k\}$ in this case?
- Answer: $1/n$, if $k \in \{1, 2, \ldots, n\}$.
- Now say we roll three dice and let Y be sum of the values on the dice. What is $P\{Y = 5\}$?
- $6/216$

Indicators

- Given any event E, can define an **indicator** random variable, i.e., let X be random variable equal to 1 on the event E and 0 otherwise. Write this as $X = 1_E$.
- The value of 1_E (either 1 or 0) *indicates* whether the event has occurred.
- If $E_1, E_2, \ldots E_k$ are events then $X = \sum_{i=1}^k 1_{E_i}$ is the number of these events that occur.
- Example: in n-hat shuffle problem, let E_i be the event ith person gets own hat.
- Then $\sum_{i=1}^n 1_{E_i}$ is total number of people who get own hats.
- Writing random variable as sum of indicators: frequently useful, sometimes confusing.

Outline

Outline

Defining random variables

Probability mass function and distribution function

Recursions

Probability mass function

- Say X is a **discrete** random variable if (with probability one) it takes one of a countable set of values.
- For each a in this countable set, write $p(a) := P\{X = a\}$. Call p the **probability mass function**.
- For the cumulative distribution function, write $F(a) = P\{X \le a\} = \sum_{x \le a} p(x)$.
- Example: Let $T_1, T_2, T_3, \ldots$ be sequence of independent fair coin tosses (each taking values in $\{H, T\}$) and let X be the smallest j for which $T_j = H$.
- What is $p(k) = P\{X = k\}$ (for $k \in \mathbb{Z}$) in this case?
- $p(k) = (1/2)^k$
- What about $F_X(k)$?
- $1 - (1/2)^k$

Another example

- Another example: let X be non-negative integer such that $p(k) = P\{X = k\} = e^{-\lambda}\lambda^k/k!$.
- Recall Taylor expansion $\sum_{k=0}^{\infty} \lambda^k/k! = e^{\lambda}$.
- In this example, X is called a **Poisson** random variable with intensity λ.
- Question: what is the state space in this example?
- Answer: Didn't specify. One possibility would be to define state space as $S = \{0, 1, 2, \ldots\}$ and define X (as a function on S) by $X(j) = j$. The probability function would be determined by $P(S) = \sum_{k \in S} e^{-\lambda}\lambda^k/k!$.
- Are there other choices of S and P — and other functions X from S to P — for which the values of $P\{X = k\}$ are the same?
- Yes. "X is a Poisson random variable with intensity λ" is statement only about the *probability mass function* of X.

Outline

Defining random variables

Probability mass function and distribution function

Recursions

Outline

Defining random variables

Probability mass function and distribution function

Recursions

Using Bayes' rule to set up recursions

- Gambler one has positive integer m dollars, gambler two has positive integer n dollars. Take turns making one dollar bets until one runs out of money. What is probability first gambler runs out of money first?
- $n/(m+n)$
- **Gambler's ruin:** what if gambler one has an unlimited amount of money?
- Wins eventually with probability one.
- **Problem of points:** in sequence of independent fair coin tosses, what is probability $P_{n,m}$ to see n heads before seeing m tails?
- Observe: $P_{n,m}$ is equivalent to the probability of having n or more heads in first $m+n-1$ trials.
- Probability of exactly n heads in $m+n-1$ trials is $\binom{m+n-1}{n}$.
- Famous correspondence by Fermat and Pascal. Led Pascal to write *Le Triangle Arithmétique*.

18.600: Lecture 9

Expectations of discrete random variables

Scott Sheffield

MIT

Outline

Outline

Expectation of a discrete random variable

- Recall: a random variable X is a function from the state space to the real numbers.
- Can interpret X as a quantity whose value depends on the outcome of an experiment.
- Say X is a **discrete** random variable if (with probability one) it takes one of a countable set of values.
- For each a in this countable set, write $p(a) := P\{X = a\}$. Call p the **probability mass function**.
- The **expectation** of X, written $E[X]$, is defined by

$$E[X] = \sum_{x:p(x)>0} xp(x).$$

- Represents weighted average of possible values X can take, each value being weighted by its probability.

Simple examples

- Suppose that a random variable X satisfies $P\{X=1\}=.5$, $P\{X=2\}=.25$ and $P\{X=3\}=.25$.
- What is $E[X]$?
- Answer: $.5\times 1+.25\times 2+.25\times 3=1.75$.
- Suppose $P\{X=1\}=p$ and $P\{X=0\}=1-p$. Then what is $E[X]$?
- Answer: p.
- Roll a standard six-sided die. What is the expectation of number that comes up?
- Answer: $\frac{1}{6}1+\frac{1}{6}2+\frac{1}{6}3+\frac{1}{6}4+\frac{1}{6}5+\frac{1}{6}6=\frac{21}{6}=3.5$.

Expectation when state space is countable

- If the state space S is countable, we can give **SUM OVER STATE SPACE** definition of expectation:

$$E[X]=\sum_{s\in S}P\{s\}X(s).$$

- Compare this to the **SUM OVER POSSIBLE X VALUES** definition we gave earlier:

$$E[X]=\sum_{x:p(x)>0}xp(x).$$

- Example: toss two coins. If X is the number of heads, what is $E[X]$?
- State space is $\{(H,H),(H,T),(T,H),(T,T)\}$ and summing over state space gives $E[X]=\frac{1}{4}2+\frac{1}{4}1+\frac{1}{4}1+\frac{1}{4}0=1$.

A technical point

- If the state space S is countable, is it possible that the sum $E[X]=\sum_{s\in S}P(\{s\})X(s)$ somehow depends on the order in which $s\in S$ are enumerated?
- In principle, yes... We only say expectation is defined when $\sum_{s\in S}P(\{x\})|X(s)|<\infty$, in which case it turns out that the sum does not depend on the order.

Outline

Outline

Expectation of a function of a random variable

- If X is a random variable and g is a function from the real numbers to the real numbers then $g(X)$ is also a random variable.
- How can we compute $E[g(X)]$?
- **SUM OVER STATE SPACE:**
$$E[g(X)] = \sum_{s \in S} P(\{s\})g(X(s)).$$
- **SUM OVER X VALUES:**
$$E[g(X)] = \sum_{x:p(x)>0} g(x)p(x).$$
- Suppose that constants a, b, μ are given and that $E[X] = \mu$.
- What is $E[X + b]$?
- How about $E[aX]$?
- Generally, $E[aX + b] = aE[X] + b = a\mu + b$.

More examples

- Let X be the number that comes up when you roll a standard six-sided die. What is $E[X^2]$?
- $\frac{1}{6}(1 + 4 + 9 + 16 + 25 + 36) = 91/12$
- Let X_j be 1 if the jth coin toss is heads and 0 otherwise. What is the expectation of $X = \sum_{i=1}^n X_j$?
- Can compute this directly as $\sum_{k=0}^n P\{X = k\}k$.
- Alternatively, use symmetry. Expected number of heads should be same as expected number of tails.
- This implies $E[X] = E[n - X]$. Applying $E[aX + b] = aE[X] + b$ formula (with $a = -1$ and $b = n$), we obtain $E[X] = n - E[X]$ and conclude that $E[X] = n/2$.

Additivity of expectation

- If X and Y are distinct random variables, then can one say that $E[X + Y] = E[X] + E[Y]$?
- Yes. In fact, for real constants a and b, we have $E[aX + bY] = aE[X] + bE[Y]$.
- This is called the **linearity of expectation**.
- Another way to state this fact: given sample space S and probability measure P, the expectation $E[\cdot]$ is a **linear** real-valued function on the space of random variables.
- Can extend to more variables $E[X_1 + X_2 + \ldots + X_n] = E[X_1] + E[X_2] + \ldots + E[X_n]$.

More examples

- Now can we compute expected number of people who get own hats in n hat shuffle problem?
- Let X_i be 1 if ith person gets own hat and zero otherwise.
- What is $E[X_i]$, for $i \in \{1, 2, \dots, n\}$?
- Answer: $1/n$.
- Can write total number with own hat as $X = X_1 + X_2 + \dots + X_n$.
- Linearity of expectation gives $E[X] = E[X_1] + E[X_2] + \dots + E[X_n] = n \times 1/n = 1$.

Outline

Outline

Why should we care about expectation?

- **Laws of large numbers:** choose lots of independent random variables with same probability distribution as X — their average tends to be close to $E[X]$.
- Example: roll $N = 10^6$ dice, let Y be the sum of the numbers that come up. Then Y/N is probably close to 3.5.
- **Economic theory of decision making:** Under "rationality" assumptions, each of us has utility function and tries to optimize its expectation.
- **Financial contract pricing:** under "no arbitrage/interest" assumption, price of derivative equals its expected value in so-called **risk neutral probability**.
- **Comes up everywhere** probability is applied.

Expected utility when outcome only depends on wealth

- Contract one: I'll toss 10 coins, and if they all come up heads (probability about one in a thousand), I'll give you 20 billion dollars.
- Contract two: I'll just give you ten million dollars.
- What are expectations of the two contracts? Which would you prefer?
- Can you find a function $u(x)$ such that given two random wealth variables W_1 and W_2, you prefer W_1 whenever $E[u(W_1)] < E[u(W_2)]$?
- Let's assume $u(0) = 0$ and $u(1) = 1$. Then $u(x) = y$ means that you are indifferent between getting 1 dollar no matter what and getting x dollars with probability $1/y$.

18.600: Lecture 10

Variance and standard deviation

Scott Sheffield

MIT

Outline

Outline

Defining variance

Examples

Properties

Decomposition trick

Recall definitions for expectation

- Recall: a random variable X is a function from the state space to the real numbers.
- Can interpret X as a quantity whose value depends on the outcome of an experiment.
- Say X is a **discrete** random variable if (with probability one) it takes one of a countable set of values.
- For each a in this countable set, write $p(a) := P\{X = a\}$. Call p the **probability mass function**.
- The **expectation** of X, written $E[X]$, is defined by

$$E[X] = \sum_{x:p(x)>0} xp(x).$$

- Also,

$$E[g(X)] = \sum_{x:p(x)>0} g(x)p(x).$$

Defining variance

- Let X be a random variable with mean μ.
- The variance of X, denoted $\mathrm{Var}(X)$, is defined by $\mathrm{Var}(X) = E[(X-\mu)^2]$.
- Taking $g(x) = (x-\mu)^2$, and recalling that $E[g(X)] = \sum_{x:p(x)>0} g(x)p(x)$, we find that

$$\mathrm{Var}[X] = \sum_{x:p(x)>0} (x-\mu)^2 p(x).$$

- Variance is one way to measure the amount a random variable "varies" from its mean over successive trials.

Very important alternatative formula

- Let X be a random variable with mean μ.
- We introduced above the formula $\mathrm{Var}(X) = E[(X-\mu)^2]$.
- This can be written $\mathrm{Var}[X] = E[X^2 - 2X\mu + \mu^2]$.
- By additivity of expectation, this is the same as $E[X^2] - 2\mu E[X] + \mu^2 = E[X^2] - \mu^2$.
- This gives us our very important alternative formula: $\mathrm{Var}[X] = E[X^2] - (E[X])^2$.
- Seven words to remember: "expectation of square minus square of expectation."
- Original formula gives intuitive idea of what variance is (expected square of difference from mean). But we will often use this alternative formula when we have to actually compute the variance.

Outline

Defining variance

Examples

Properties

Decomposition trick

Outline

Defining variance

Examples

Properties

Decomposition trick

Variance examples

- If X is number on a standard die roll, what is $\mathrm{Var}[X]$?
- $\mathrm{Var}[X] = E[X^2] - E[X]^2 =$
$\frac{1}{6}1^2 + \frac{1}{6}2^2 + \frac{1}{6}3^2 + \frac{1}{6}4^2 + \frac{1}{6}5^2 + \frac{1}{6}6^2 - (7/2)^2 = \frac{91}{6} - \frac{49}{4} = \frac{35}{12}$.
- Let Y be number of heads in two fair coin tosses. What is $\mathrm{Var}[Y]$?
- Recall $P\{Y = 0\} = 1/4$ and $P\{Y = 1\} = 1/2$ and $P\{Y = 2\} = 1/4$.
- Then $\mathrm{Var}[Y] = E[Y^2] - E[Y]^2 = \frac{1}{4}0^2 + \frac{1}{2}1^2 + \frac{1}{4}2^2 - 1^2 = \frac{1}{2}$.

More variance examples

- You buy a lottery ticket that gives you a one in a million chance to win a million dollars.
- Let X be the amount you win. What's the expectation of X?
- How about the variance?
- Variance is more sensitive than expectation to rare "outlier" events.
- At a particular party, there are four five-foot-tall people, five six-foot-tall people, and one seven-foot-tall person. You pick one of these people uniformly at random. What is the expected height of the person you pick?
- $E[X] = .4 \cdot 5 + .5 \cdot 6 + .1 \cdot 7 = 5.7$
- Variance?
- $.4 \cdot 25 + .5 \cdot 36 + .1 \cdot 49 - (5.7)^2 = 32.9 - 32.49 = .41$,

Outline

Outline

Identity

- If $Y = X + b$, where b is constant, then does it follow that $\mathrm{Var}[Y] = \mathrm{Var}[X]$?
- Yes.
- We showed earlier that $E[aX] = aE[X]$. We claim that $\mathrm{Var}[aX] = a^2\mathrm{Var}[X]$.
- Proof: $\mathrm{Var}[aX] = E[a^2X^2] - E[aX]^2 = a^2E[X^2] - a^2E[X]^2 = a^2\mathrm{Var}[X]$.

Standard deviation

- Write $\mathrm{SD}[X] = \sqrt{\mathrm{Var}[X]}$.
- Satisfies identity $\mathrm{SD}[aX] = a\mathrm{SD}[X]$.
- Uses the same units as X itself.
- If we switch from feet to inches in our "height of randomly chosen person" example, then X, $E[X]$, and $\mathrm{SD}[X]$ each get multiplied by 12, but $\mathrm{Var}[X]$ gets multiplied by 144.

Outline

Outline

Number of aces

- Choose five cards from a standard deck of 52 cards. Let A be the number of aces you see.
- Let's compute $E[A]$ and $\mathrm{Var}[A]$.
- To start with, how many five card hands total?
- Answer: $\binom{52}{5}$.
- How many such hands have k aces?
- Answer: $\binom{4}{k}\binom{48}{5-k}$.
- So $P\{A=k\} = \frac{\binom{4}{k}\binom{48}{5-k}}{\binom{52}{5}}$.
- So $E[A] = \sum_{k=0}^{4} kP\{A=k\}$,
- and $\mathrm{Var}[A] = \sum_{k=0}^{4} k^2 P\{A=k\} - E[A]^2$.

Number of aces revisited

- Choose five cards from a standard deck of 52 cards. Let A be the number of aces you see.
- Choose five cards in order, and let A_i be 1 if the ith card chosen is an ace and zero otherwise.
- Then $A = \sum_{i=1}^{5} A_i$. And $E[A] = \sum_{i=1}^{5} E[A_i] = 5/13$.
- Now $A^2 = (A_1 + A_2 + \ldots + A_5)^2$ can be expanded into 25 terms: $A^2 = \sum_{i=1}^{5}\sum_{j=1}^{5} A_i A_j$.
- So $E[A^2] = \sum_{i=1}^{5}\sum_{j=1}^{5} E[A_i A_j]$.
- Five terms of form $E[A_i A_j]$ with $i=j$ five with $i \neq j$. First five contribute $1/13$ each. How about other twenty?
- $E[A_i A_j] = (1/13)(3/51) = (1/13)(1/17)$. So $E[A^2] = \frac{5}{13} + \frac{20}{13\times 17} = \frac{105}{13\times 17}$.
- $\mathrm{Var}[A] = E[A^2] - E[A]^2 = \frac{105}{13\times 17} - \frac{25}{13\times 13}$.

Hat problem variance

- In the n-hat shuffle problem, let X be the number of people who get their own hat. What is $\mathrm{Var}[X]$?
- We showed earlier that $E[X]=1$. So $\mathrm{Var}[X] = E[X^2] - 1$.
- But how do we compute $E[X^2]$?
- Decomposition trick: write variable as sum of simple variables.
- Let X_i be one if ith person gets own hat and zero otherwise. Then $X = X_1 + X_2 + \ldots + X_n = \sum_{i=1}^{n} X_i$.
- We want to compute $E[(X_1 + X_2 + \ldots + X_n)^2]$.
- Expand this out and using linearity of expectation:

$$E[\sum_{i=1}^{n} X_i \sum_{j=1}^{n} X_j] = \sum_{i=1}^{n}\sum_{j=1}^{n} E[X_i X_j] = n\cdot\frac{1}{n} + n(n-1)\frac{1}{n(n-1)} = 2.$$

- So $\mathrm{Var}[X] = E[X^2] - (E[X])^2 = 2 - 1 = 1$.

18.600: Lecture 11

Binomial random variables and repeated trials

Scott Sheffield

MIT

Outline

Outline

Bernoulli random variables

- Toss fair coin n times. (Tosses are independent.) What is the probability of k heads?
- Answer: $\binom{n}{k}/2^n$.
- What if coin has p probability to be heads?
- Answer: $\binom{n}{k}p^k(1-p)^{n-k}$.
- Writing $q = 1 - p$, we can write this as $\binom{n}{k}p^k q^{n-k}$
- Can use binomial theorem to show probabilities sum to one:
- $1 = 1^n = (p+q)^n = \sum_{k=0}^n \binom{n}{k}p^k q^{n-k}$.
- Number of heads is **binomial random variable with parameters** (n, p).

Examples

- Toss 6 fair coins. Let X be number of heads you see. Then X is binomial with parameters (n, p) given by $(6, 1/2)$.
- Probability mass function for X can be computed using the 6th row of Pascal's triangle.
- If coin is biased (comes up heads with probability $p \neq 1/2$), we can still use the 6th row of Pascal's triangle, but the probability that $X = i$ gets multiplied by $p^i(1-p)^{n-i}$.

Other examples

- Room contains n people. What is the probability that exactly i of them were born on a Tuesday?
- Answer: use binomial formula $\binom{n}{i}p^i q^{n-i}$ with $p = 1/7$ and $q = 1 - p = 6/7$.
- Let $n = 100$. Compute the probability that nobody was born on a Tuesday.
- What is the probability that exactly 15 people were born on a Tuesday?

Outline

Outline

Expectation

- Let X be a binomial random variable with parameters (n, p).
- What is $E[X]$?
- Direct approach: by definition of expectation, $E[X] = \sum_{i=0}^{n} P\{X = i\} i$.
- What happens if we modify the nth row of Pascal's triangle by multiplying the i term by i?
- For example, replace the 5th row $(1, 5, 10, 10, 5, 1)$ by $(0, 5, 20, 30, 20, 5)$. Does this remind us of an earlier row in the triangle?
- Perhaps the prior row $(1, 4, 6, 4, 1)$?

Useful Pascal's triangle identity

- Recall that $\binom{n}{i} = \frac{n \times (n-1) \times \ldots \times (n-i+1)}{i \times (i-1) \times \ldots \times (1)}$. This implies a simple but important identity: $i\binom{n}{i} = n\binom{n-1}{i-1}$.
- Using this identity (and $q = 1 - p$), we can write

$$E[X] = \sum_{i=0}^{n} i \binom{n}{i} p^i q^{n-i} = \sum_{i=1}^{n} n \binom{n-1}{i-1} p^i q^{n-i}.$$

- Rewrite this as $E[X] = np \sum_{i=1}^{n} \binom{n-1}{i-1} p^{(i-1)} q^{(n-1)-(i-1)}$.
- Substitute $j = i - 1$ to get

$$E[X] = np \sum_{j=0}^{n-1} \binom{n-1}{j} p^j q^{(n-1)-j} = np(p+q)^{n-1} = np.$$

Decomposition approach to computing expectation

- Let X be a binomial random variable with parameters (n, p). Here is another way to compute $E[X]$.
- Think of X as representing number of heads in n tosses of coin that is heads with probability p.
- Write $X = \sum_{j=1}^{n} X_j$, where X_j is 1 if the jth coin is heads, 0 otherwise.
- In other words, X_j is the number of heads (zero or one) on the jth toss.
- Note that $E[X_j] = p \cdot 1 + (1-p) \cdot 0 = p$ for each j.
- Conclude by additivity of expectation that

$$E[X] = \sum_{j=1}^{n} E[X_j] = \sum_{j=1}^{n} p = np.$$

Interesting moment computation

- Let X be binomial (n, p) and fix $k \geq 1$. What is $E[X^k]$?
- Recall identity: $i\binom{n}{i} = n\binom{n-1}{i-1}$.
- Generally, $E[X^k]$ can be written as

$$\sum_{i=0}^{n} i \binom{n}{i} p^i (1-p)^{n-i} i^{k-1}.$$

- Identity gives

$$E[X^k] = np \sum_{i=1}^{n} \binom{n-1}{i-1} p^{i-1} (1-p)^{n-i} i^{k-1} =$$

$$np \sum_{j=0}^{n-1} \binom{n-1}{j} p^j (1-p)^{n-1-j} (j+1)^{k-1}.$$

- Thus $E[X^k] = npE[(Y+1)^{k-1}]$ where Y is binomial with parameters $(n-1, p)$.

Computing the variance

- Let X be binomial (n,p). What is $E[X]$?
- We know $E[X] = np$.
- We computed identity $E[X^k] = npE[(Y+1)^{k-1}]$ where Y is binomial with parameters $(n-1,p)$.
- In particular $E[X^2] = npE[Y+1] = np[(n-1)p+1]$.
- So $\mathrm{Var}[X] = E[X^2] - E[X]^2 = np(n-1)p + np - (np)^2 = np(1-p) = npq$, where $q = 1-p$.
- Commit to memory: variance of binomial (n,p) random variable is npq.
- This is n times the variance you'd get with a single coin. Coincidence?

Compute variance with decomposition trick

- $X = \sum_{j=1}^n X_j$, so $E[X^2] = E[\sum_{i=1}^n X_i \sum_{j=1}^n X_j] = \sum_{i=1}^n \sum_{j=1}^n E[X_i X_j]$
- $E[X_i X_j]$ is p if $i = j$, p^2 otherwise.
- $\sum_{i=1}^n \sum_{j=1}^n E[X_i X_j]$ has n terms equal to p and $(n-1)n$ terms equal to p^2.
- So $E[X^2] = np + (n-1)np^2 = np + (np)^2 - np^2$.
- Thus $\mathrm{Var}[X] = E[X^2] - E[X]^2 = np - np^2 = np(1-p) = npq$.

Outline

Outline

More examples

- An airplane seats 200, but the airline has sold 205 tickets. Each person, independently, has a .05 chance of not showing up for the flight. What is the probability that more than 200 people will show up for the flight?
- $\sum_{j=201}^{205} \binom{205}{j} .95^j .05^{205-j}$
- In a 100 person senate, forty people always vote for the Republicans' position, forty people always for the Democrats' position and 20 people just toss a coin to decide which way to vote. What is the probability that a given vote is tied?
- $\binom{20}{10}/2^{20}$
- You invite 50 friends to a party. Each one, independently, has a 1/3 chance of showing up. What is the probability that more than 25 people will show up?
- $\sum_{j=26}^{50} \binom{50}{j} (1/3)^j (2/3)^{50-j}$

18.600: Lecture 12

Poisson random variables

Scott Sheffield

MIT

Outline

Poisson random variable definition

Poisson random variable properties

Poisson random variable problems

Outline

Poisson random variable definition

Poisson random variable properties

Poisson random variable problems

Poisson random variables: motivating questions

- How many raindrops hit a given square inch of sidewalk during a ten minute period?
- How many people fall down the stairs in a major city on a given day?
- How many plane crashes in a given year?
- How many radioactive particles emitted during a time period in which the expected number emitted is 5?
- How many calls to call center during a given minute?
- How many goals scored during a 90 minute soccer game?
- How many notable gaffes during 90 minute debate?
- **Key idea for all these examples:** Divide time into large number of small increments. Assume that during each increment, there is some small probability of thing happening (independently of other increments).

Remember what e is?

- The number e is defined by $e = \lim_{n\to\infty}(1+1/n)^n$.
- It's the amount of money that one dollar grows to over a year when you have an interest rate of 100 percent, continuously compounded.
- Similarly, $e^\lambda = \lim_{n\to\infty}(1+\lambda/n)^n$.
- It's the amount of money that one dollar grows to over a year when you have an interest rate of 100λ percent, continuously compounded.
- It's also the amount of money that one dollar grows to over λ years when you have an interest rate of 100 percent, continuously compounded.
- Can also change sign: $e^{-\lambda} = \lim_{n\to\infty}(1-\lambda/n)^n$.

Bernoulli random variable with n large and $np = \lambda$

- Let λ be some moderate-sized number. Say $\lambda = 2$ or $\lambda = 3$. Let n be a huge number, say $n = 10^6$.
- Suppose I have a coin that comes up heads with probability λ/n and I toss it n times.
- How many heads do I expect to see?
- Answer: $np = \lambda$.
- Let k be some moderate sized number (say $k = 4$). What is the probability that I see exactly k heads?
- Binomial formula: $\binom{n}{k}p^k(1-p)^{n-k} = \frac{n(n-1)(n-2)\dots(n-k+1)}{k!}p^k(1-p)^{n-k}$.
- This is approximately $\frac{\lambda^k}{k!}(1-p)^{n-k} \approx \frac{\lambda^k}{k!}e^{-\lambda}$.
- A **Poisson random variable** X with parameter λ satisfies $P\{X=k\} = \frac{\lambda^k}{k!}e^{-\lambda}$ for integer $k \geq 0$.

Outline

Outline

Probabilities sum to one

- A **Poisson random variable** X with parameter λ satisfies $p(k) = P\{X = k\} = \frac{\lambda^k}{k!}e^{-\lambda}$ for integer $k \geq 0$.
- How can we show that $\sum_{k=0}^{\infty} p(k) = 1$?
- Use Taylor expansion $e^{\lambda} = \sum_{k=0}^{\infty} \frac{\lambda^k}{k!}$.

Expectation

- A **Poisson random variable** X with parameter λ satisfies $P\{X = k\} = \frac{\lambda^k}{k!}e^{-\lambda}$ for integer $k \geq 0$.
- What is $E[X]$?
- We think of a Poisson random variable as being (roughly) a Bernoulli (n, p) random variable with n very large and $p = \lambda/n$.
- This would suggest $E[X] = \lambda$. Can we show this directly from the formula for $P\{X = k\}$?
- By definition of expectation

$$E[X] = \sum_{k=0}^{\infty} P\{X = k\}k = \sum_{k=0}^{\infty} k\frac{\lambda^k}{k!}e^{-\lambda} = \sum_{k=1}^{\infty} \frac{\lambda^k}{(k-1)!}e^{-\lambda}.$$

- Setting $j = k - 1$, this is $\lambda \sum_{j=0}^{\infty} \frac{\lambda^j}{j!}e^{-\lambda} = \lambda$.

Variance

- Given $P\{X = k\} = \frac{\lambda^k}{k!}e^{-\lambda}$ for integer $k \geq 0$, what is $\mathrm{Var}[X]$?
- Think of X as (roughly) a Bernoulli (n, p) random variable with n very large and $p = \lambda/n$.
- This suggests $\mathrm{Var}[X] \approx npq \approx \lambda$ (since $np \approx \lambda$ and $q = 1 - p \approx 1$). Can we show directly that $\mathrm{Var}[X] = \lambda$?
- Compute

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

- Setting $j = k - 1$, this is

$$\lambda \left(\sum_{j=0}^{\infty} (j+1)\frac{\lambda^j}{j!}e^{-\lambda} \right) = \lambda E[X + 1] = \lambda(\lambda + 1).$$

- Then $\mathrm{Var}[X] = E[X^2] - E[X]^2 = \lambda(\lambda + 1) - \lambda^2 = \lambda$.

Outline

Outline

Poisson random variable problems

- A country has an average of 2 plane crashes per year.
- How reasonable is it to assume the number of crashes is Poisson with parameter 2?
- Assuming this, what is the probability of exactly 2 crashes? Of zero crashes? Of four crashes?
- $e^{-\lambda}\lambda^k/k!$ with $\lambda = 2$ and k set to 2 or 0 or 4
- A city has an average of five major earthquakes a century. What is the probability that there is at least one major earthquake in a given decade (assuming the number of earthquakes per decade is Poisson)?
- $1 - e^{-\lambda}\lambda^k/k!$ with $\lambda = .5$ and $k = 0$
- A casino deals one million five-card poker hands per year. Approximate the probability that there are exactly 2 royal flush hands during a given year.
- Expected number of royal flushes is $\lambda = 10^6 \cdot 4/\binom{52}{5} \approx 1.54$. Answer is $e^{-\lambda}\lambda^k/k!$ with $k = 2$.

18.600: Lecture 13

Lectures 1-12 Review

Scott Sheffield

MIT

Outline

Counting tricks and basic principles of probability

Discrete random variables

Outline

Counting tricks and basic principles of probability

Discrete random variables

Selected counting tricks

- Break "choosing one of the items to be counted" into a sequence of stages so that one always has the same number of choices to make at each stage. Then the total count becomes a product of number of choices available at each stage.
- Overcount by a fixed factor.
- If you have n elements you wish to divide into r distinct piles of sizes $n_1, n_2 \ldots n_r$, how many ways to do that?
- Answer $\binom{n}{n_1, n_2, \ldots, n_r} := \frac{n!}{n_1! n_2! \ldots n_r!}$.
- How many sequences $a_1, \ldots, a_k$ of non-negative integers satisfy $a_1 + a_2 + \ldots + a_k = n$?
- Answer: $\binom{n+k-1}{n}$. Represent partition by $k-1$ bars and n stars, e.g., as $**|**||****|*$.

Axioms of probability

- Have a set S called *sample space*.
- $P(A) \in [0,1]$ for all (measurable) $A \subset S$.
- $P(S) = 1$.
- Finite additivity: $P(A \cup B) = P(A) + P(B)$ if $A \cap B = \emptyset$.
- Countable additivity: $P(\cup_{i=1}^{\infty} E_i) = \sum_{i=1}^{\infty} P(E_i)$ if $E_i \cap E_j = \emptyset$ for each pair i and j.

Consequences of axioms

- $P(A^c) = 1 - P(A)$
- $A \subset B$ implies $P(A) \leq P(B)$
- $P(A \cup B) = P(A) + P(B) - P(AB)$
- $P(AB) \leq P(A)$

Inclusion-exclusion identity

- Observe $P(A \cup B) = P(A) + P(B) - P(AB)$.
- Also, $P(E \cup F \cup G) =$
  $P(E) + P(F) + P(G) - P(EF) - P(EG) - P(FG) + P(EFG)$.
- More generally,

$$\begin{aligned} P(\cup_{i=1}^n E_i) &= \sum_{i=1}^n P(E_i) - \sum_{i_1 < i_2} P(E_{i_1} E_{i_2}) + \ldots \\ &+ (-1)^{(r+1)} \sum_{i_1 < i_2 < \ldots < i_r} P(E_{i_1} E_{i_2} \ldots E_{i_r}) \\ &= + \ldots + (-1)^{n+1} P(E_1 E_2 \ldots E_n). \end{aligned}$$

- The notation $\sum_{i_1 < i_2 < \ldots < i_r}$ means a sum over all of the $\binom{n}{r}$ subsets of size r of the set $\{1, 2, \ldots, n\}$.

Famous hat problem

- n people toss hats into a bin, randomly shuffle, return one hat to each person. Find probability nobody gets own hat.
- Inclusion-exclusion. Let E_i be the event that ith person gets own hat.
- What is $P(E_{i_1} E_{i_2} \ldots E_{i_r})$?
- Answer: $\frac{(n-r)!}{n!}$.
- There are $\binom{n}{r}$ terms like that in the inclusion exclusion sum. What is $\binom{n}{r} \frac{(n-r)!}{n!}$?
- Answer: $\frac{1}{r!}$.
- $P(\cup_{i=1}^n E_i) = 1 - \frac{1}{2!} + \frac{1}{3!} - \frac{1}{4!} + \ldots \pm \frac{1}{n!}$
- $1 - P(\cup_{i=1}^n E_i) = 1 - 1 + \frac{1}{2!} - \frac{1}{3!} + \frac{1}{4!} - \ldots \pm \frac{1}{n!} \approx 1/e \approx .36788$

Conditional probability

- Definition: $P(E|F) = P(EF)/P(F)$.
- Call $P(E|F)$ the "conditional probability of E given F" or "probability of E conditioned on F".
- Nice fact: $P(E_1E_2E_3 \dots E_n) = P(E_1)P(E_2|E_1)P(E_3|E_1E_2)\dots P(E_n|E_1 \dots E_{n-1})$
- Useful when we think about multi-step experiments.
- For example, let E_i be event ith person gets own hat in the n-hat shuffle problem.

Dividing probability into two cases

- $$\begin{aligned} P(E) &= P(EF) + P(EF^c) \\ &= P(E|F)P(F) + P(E|F^c)P(F^c) \end{aligned}$$
- In words: want to know the probability of E. There are two scenarios F and F^c. If I know the probabilities of the two scenarios and the probability of E conditioned on each scenario, I can work out the probability of E.

Bayes' theorem

- Bayes' theorem/law/rule states the following: $P(A|B) = \frac{P(B|A)P(A)}{P(B)}$.
- Follows from definition of conditional probability: $P(AB) = P(B)P(A|B) = P(A)P(B|A)$.
- Tells how to update estimate of probability of A when new evidence restricts your sample space to B.
- So $P(A|B)$ is $\frac{P(B|A)}{P(B)}$ times $P(A)$.
- Ratio $\frac{P(B|A)}{P(B)}$ determines "how compelling new evidence is".

$P(\cdot|F)$ is a probability measure

- We can check the probability axioms: $0 \leq P(E|F) \leq 1$, $P(S|F) = 1$, and $P(\cup E_i) = \sum P(E_i|F)$, if i ranges over a countable set and the E_i are disjoint.
- The probability measure $P(\cdot|F)$ is related to $P(\cdot)$.
- To get former from latter, we set probabilities of elements outside of F to zero and multiply probabilities of events inside of F by $1/P(F)$.
- $P(\cdot)$ is the *prior* probability measure and $P(\cdot|F)$ is the *posterior* measure (revised after discovering that F occurs).

Independence

- Say E and F are **independent** if $P(EF) = P(E)P(F)$.
- Equivalent statement: $P(E|F) = P(E)$. Also equivalent: $P(F|E) = P(F)$.

Independence of multiple events

- Say $E_1 \ldots E_n$ are independent if for each $\{i_1, i_2, \ldots, i_k\} \subset \{1, 2, \ldots n\}$ we have $P(E_{i_1} E_{i_2} \ldots E_{i_k}) = P(E_{i_1}) P(E_{i_2}) \ldots P(E_{i_k})$.
- In other words, the product rule works.
- Independence implies $P(E_1 E_2 E_3 | E_4 E_5 E_6) = \frac{P(E_1)P(E_2)P(E_3)P(E_4)P(E_5)P(E_6)}{P(E_4)P(E_5)P(E_6)} = P(E_1 E_2 E_3)$, and other similar statements.
- Does pairwise independence imply independence?
- No. Consider these three events: first coin heads, second coin heads, odd number heads. Pairwise independent, not independent.

Outline

Outline

Random variables

- A random variable X is a function from the state space to the real numbers.
- Can interpret X as a quantity whose value depends on the outcome of an experiment.
- Say X is a **discrete** random variable if (with probability one) if it takes one of a countable set of values.
- For each a in this countable set, write $p(a) := P\{X = a\}$. Call p the **probability mass function**.
- Write $F(a) = P\{X \leq a\} = \sum_{x \leq a} p(x)$. Call F the **cumulative distribution function**.

Indicators

- Given any event E, can define an **indicator** random variable, i.e., let X be random variable equal to 1 on the event E and 0 otherwise. Write this as $X = 1_E$.
- The value of 1_E (either 1 or 0) *indicates* whether the event has occurred.
- If $E_1, E_2, \ldots, E_k$ are events then $X = \sum_{i=1}^k 1_{E_i}$ is the number of these events that occur.
- Example: in n-hat shuffle problem, let E_i be the event ith person gets own hat.
- Then $\sum_{i=1}^n 1_{E_i}$ is total number of people who get own hats.

Expectation of a discrete random variable

- Say X is a **discrete** random variable if (with probability one) it takes one of a countable set of values.
- For each a in this countable set, write $p(a) := P\{X = a\}$. Call p the **probability mass function**.
- The **expectation** of X, written $E[X]$, is defined by

$$E[X] = \sum_{x:p(x)>0} xp(x).$$

- Represents weighted average of possible values X can take, each value being weighted by its probability.

Expectation when state space is countable

- If the state space S is countable, we can give **SUM OVER STATE SPACE** definition of expectation:

$$E[X] = \sum_{s \in S} P\{s\}X(s).$$

- Agrees with the **SUM OVER POSSIBLE X VALUES** definition:

$$E[X] = \sum_{x:p(x)>0} xp(x).$$

Expectation of a function of a random variable

- If X is a random variable and g is a function from the real numbers to the real numbers then $g(X)$ is also a random variable.
- How can we compute $E[g(X)]$?
- Answer:
$$E[g(X)] = \sum_{x:p(x)>0} g(x)p(x).$$

Additivity of expectation

- If X and Y are distinct random variables, then $E[X+Y] = E[X] + E[Y]$.
- In fact, for real constants a and b, we have $E[aX+bY] = aE[X] + bE[Y]$.
- This is called the **linearity of expectation**.
- Can extend to more variables $E[X_1 + X_2 + \ldots + X_n] = E[X_1] + E[X_2] + \ldots + E[X_n]$.

Defining variance in discrete case

- Let X be a random variable with mean μ.
- The variance of X, denoted $\mathrm{Var}(X)$, is defined by $\mathrm{Var}(X) = E[(X-\mu)^2]$.
- Taking $g(x) = (x-\mu)^2$, and recalling that $E[g(X)] = \sum_{x:p(x)>0} g(x)p(x)$, we find that
$$\mathrm{Var}[X] = \sum_{x:p(x)>0} (x-\mu)^2 p(x).$$
- Variance is one way to measure the amount a random variable "varies" from its mean over successive trials.
- Very important alternate formula: $\mathrm{Var}[X] = E[X^2] - (E[X])^2$.

Identity

- If $Y = X + b$, where b is constant, then $\mathrm{Var}[Y] = \mathrm{Var}[X]$.
- Also, $\mathrm{Var}[aX] = a^2\mathrm{Var}[X]$.
- Proof: $\mathrm{Var}[aX] = E[a^2X^2] - E[aX]^2 = a^2E[X^2] - a^2E[X]^2 = a^2\mathrm{Var}[X]$.

Standard deviation

- Write $\mathrm{SD}[X] = \sqrt{\mathrm{Var}[X]}$.
- Satisfies identity $\mathrm{SD}[aX] = a\mathrm{SD}[X]$.
- Uses the same units as X itself.
- If we switch from feet to inches in our "height of randomly chosen person" example, then X, $E[X]$, and $\mathrm{SD}[X]$ each get multiplied by 12, but $\mathrm{Var}[X]$ gets multiplied by 144.

Bernoulli random variables

- Toss fair coin n times. (Tosses are independent.) What is the probability of k heads?
- Answer: $\binom{n}{k}/2^n$.
- What if coin has p probability to be heads?
- Answer: $\binom{n}{k}p^k(1-p)^{n-k}$.
- Writing $q = 1-p$, we can write this as $\binom{n}{k}p^k q^{n-k}$
- Can use binomial theorem to show probabilities sum to one:
- $1 = 1^n = (p+q)^n = \sum_{k=0}^n \binom{n}{k}p^k q^{n-k}$.
- Number of heads is **binomial random variable with parameters** (n, p).

Decomposition approach to computing expectation

- Let X be a binomial random variable with parameters (n, p). Here is one way to compute $E[X]$.
- Think of X as representing number of heads in n tosses of coin that is heads with probability p.
- Write $X = \sum_{j=1}^n X_j$, where X_j is 1 if the jth coin is heads, 0 otherwise.
- In other words, X_j is the number of heads (zero or one) on the jth toss.
- Note that $E[X_j] = p \cdot 1 + (1-p) \cdot 0 = p$ for each j.
- Conclude by additivity of expectation that

$$E[X] = \sum_{j=1}^n E[X_j] = \sum_{j=1}^n p = np.$$

Compute variance with decomposition trick

- $X = \sum_{j=1}^n X_j$, so $E[X^2] = E[\sum_{i=1}^n X_i \sum_{j=1}^n X_j] = \sum_{i=1}^n \sum_{j=1}^n E[X_i X_j]$
- $E[X_i X_j]$ is p if $i = j$, p^2 otherwise.
- $\sum_{i=1}^n \sum_{j=1}^n E[X_i X_j]$ has n terms equal to p and $(n-1)n$ terms equal to p^2.
- So $E[X^2] = np + (n-1)np^2 = np + (np)^2 - np^2$.
- Thus $\mathrm{Var}[X] = E[X^2] - E[X]^2 = np - np^2 = np(1-p) = npq$.
- Can show generally that if $X_1, \ldots, X_n$ independent then $\mathrm{Var}[\sum_{j=1}^n X_j] = \sum_{j=1}^n \mathrm{Var}[X_j]$

Bernoulli random variable with n large and $np = \lambda$

- Let λ be some moderate-sized number. Say $\lambda = 2$ or $\lambda = 3$. Let n be a huge number, say $n = 10^6$.
- Suppose I have a coin that comes on heads with probability λ/n and I toss it n times.
- How many heads do I expect to see?
- Answer: $np = \lambda$.
- Let k be some moderate sized number (say $k = 4$). What is the probability that I see exactly k heads?
- Binomial formula: $\binom{n}{k} p^k (1-p)^{n-k} = \frac{n(n-1)(n-2)\ldots(n-k+1)}{k!} p^k (1-p)^{n-k}$.
- This is approximately $\frac{\lambda^k}{k!}(1-p)^{n-k} \approx \frac{\lambda^k}{k!} e^{-\lambda}$.
- A **Poisson random variable** X with parameter λ satisfies $P\{X = k\} = \frac{\lambda^k}{k!} e^{-\lambda}$ for integer $k \geq 0$.

Expectation and variance

- A **Poisson random variable** X with parameter λ satisfies $P\{X = k\} = \frac{\lambda^k}{k!} e^{-\lambda}$ for integer $k \geq 0$.
- Clever computation tricks yield $E[X] = \lambda$ and $\mathrm{Var}[X] = \lambda$.
- We think of a Poisson random variable as being (roughly) a Bernoulli (n, p) random variable with n very large and $p = \lambda/n$.
- This also suggests $E[X] = np = \lambda$ and $\mathrm{Var}[X] = npq \approx \lambda$.

Poisson point process

- A Poisson point process is a random function $N(t)$ called a Poisson process of rate λ.
- For each $t > s \geq 0$, the value $N(t) - N(s)$ describes the number of events occurring in the time interval (s, t) and is Poisson with rate $(t - s)\lambda$.
- The numbers of events occurring in disjoint intervals are independent random variables.
- Probability to see zero events in first t time units is $e^{-\lambda t}$.
- Let T_k be time elapsed, since the previous event, until the kth event occurs. Then the T_k are independent random variables, each of which is exponential with parameter λ.

Geometric random variables

- Consider an infinite sequence of independent tosses of a coin that comes up heads with probability p.
- Let X be such that the first heads is on the Xth toss.
- Answer: $P\{X = k\} = (1-p)^{k-1} p = q^{k-1} p$, where $q = 1 - p$ is tails probability.
- Say X is a **geometric random variable with parameter** p.
- Some cool calculation tricks show that $E[X] = 1/p$.
- And $\mathrm{Var}[X] = q/p^2$.

Negative binomial random variables

- Consider an infinite sequence of independent tosses of a coin that comes up heads with probability p.
- Let X be such that the rth heads is on the Xth toss.
- Then $P\{X = k\} = \binom{k-1}{r-1} p^{r-1}(1-p)^{k-r} p$.
- Call X **negative binomial random variable with parameters** (r, p).
- So $E[X] = r/p$.
- And $\mathrm{Var}[X] = rq/p^2$.

18.600: Lecture 15

Poisson processes

Scott Sheffield

MIT

Outline

Outline

Poisson random variables

What should a Poisson point process be?

Poisson point process axioms

Consequences of axioms

Properties from last time...

- A **Poisson random variable** X with parameter λ satisfies $P\{X = k\} = \frac{\lambda^k}{k!}e^{-\lambda}$ for integer $k \geq 0$.
- The probabilities are approximately those of a binomial with parameters $(n, \lambda/n)$ when n is very large.
- Indeed,

$$\binom{n}{k} p^k (1-p)^{n-k} = \frac{n(n-1)(n-2)\dots(n-k+1)}{k!} p^k (1-p)^{n-k} \approx$$

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

- General idea: if you have a large number of unlikely events that are (mostly) independent of each other, and the expected number that occur is λ, then the total number that occur should be (approximately) a Poisson random variable with parameter λ.

Properties from last time...

- Many phenomena (number of phone calls or customers arriving in a given period, number of radioactive emissions in a given time period, number of major hurricanes in a given time period, etc.) can be modeled this way.
- A **Poisson random variable** X with parameter λ has expectation λ and variance λ.
- Special case: if $\lambda = 1$, then $P\{X = k\} = \frac{1}{k!e}$.
- Note how quickly this goes to zero, as a function of k.
- Example: number of royal flushes in a million five-card poker hands is approximately Poisson with parameter $10^6/649739 \approx 1.54$.
- Example: if a country expects 2 plane crashes in a year, then the total number might be approximately Poisson with parameter $\lambda = 2$.

A cautionary tail

- Example: Joe works for a bank and notices that his town sees an average of one mortgage foreclosure per month.
- Moreover, looking over five years of data, it seems that the number of foreclosures per month follows a rate 1 Poisson distribution.
- That is, roughly a $1/e$ fraction of months has 0 foreclosures, a $1/e$ fraction has 1, a $1/(2e)$ fraction has 2, a $1/(6e)$ fraction has 3, and a $1/(24e)$ fraction has 4.
- Joe concludes that the probability of seeing 10 foreclosures during a given month is only $1/(10!e)$. Probability to see 10 or more (an extreme *tail event* that would destroy the bank) is $\sum_{k=10}^{\infty} 1/(k!e)$, less than one in million.
- Investors are impressed. Joe receives large bonus.
- But probably shouldn't....

Outline

Outline

How should we define the *Poisson process*?

- Whatever his faults, Joe was a good record keeper. He kept track of the precise *times* at which the foreclosures occurred over the whole five years (not just the total numbers of foreclosures). We could try this for other problems as well.
- Let's encode this information with a function. We'd like a random function $N(t)$ that describe the number of events that occur during the first t units of time. (This could be a model for the number of plane crashes in first t years, or the number of royal flushes in first $10^6 t$ poker hands.)
- So $N(t)$ is a **random non-decreasing integer-valued function** of t with $N(0) = 0$.
- For each t, $N(t)$ is a random variable, and the $N(t)$ are functions on the same sample space.

Outline

Outline

Poisson process axioms

- Let's back up and give a precise and minimal list of properties we want the random function $N(t)$ to satisfy.
- 1. $N(0) = 0$.
- 2. **Independence:** Number of events (jumps of N) in disjoint time intervals are independent.
- 3. **Homogeneity:** Prob. distribution of # events in interval depends only on length. (Deduce: $E[N(h)] = \lambda h$ for some λ.)
- 4. **Non-concurrence:** $P\{N(h) \geq 2\} << P\{N(h) = 1\}$ when h is small. Precisely:
  - $P\{N(h) = 1\} = \lambda h + o(h)$. (Here $f(h) = o(h)$ means $\lim_{h \to 0} f(h)/h = 0$.)
  - $P\{N(h) \geq 2\} = o(h)$.
- A random function $N(t)$ with these properties is a **Poisson process with rate** λ.

Outline

Poisson random variables

What should a Poisson point process be?

Poisson point process axioms

Consequences of axioms

Outline

Poisson random variables

What should a Poisson point process be?

Poisson point process axioms

Consequences of axioms

Consequences of axioms: time till first event

- Can we work out the probability of no events before time t?
- We assumed $P\{N(h) = 1\} = \lambda h + o(h)$ and $P\{N(h) \geq 2\} = o(h)$. Taken together, these imply that $P\{N(h) = 0\} = 1 - \lambda h + o(h)$.
- Fix λ and t. Probability of no events in interval of length t/n is $(1 - \lambda t/n) + o(1/n)$.
- Probability of no events in first n such intervals is about $\big(1 - \lambda t/n + o(1/n)\big)^n \approx e^{-\lambda t}$.
- Taking limit as $n \to \infty$, can show that probability of no event in interval of length t is $e^{-\lambda t}$.
- $P\{N(t) = 0\} = e^{-\lambda t}$.
- Let T_1 be the time of the first event. Then $P\{T_1 \geq t\} = e^{-\lambda t}$. We say that T_1 is an **exponential random variable with rate** λ.

Consequences of axioms: time till second, third events

- Let T_2 be time between first and second event. Generally, T_k is time between $(k-1)$th and kth event.
- Then the $T_1, T_2, \ldots$ are independent of each other (informally this means that observing some of the random variables T_k gives you no information about the others). Each is an exponential random variable with rate λ.
- This finally gives us a way to construct $N(t)$. It is determined by the sequence T_j of independent exponential random variables.
- Axioms can be readily verified from this description.

Back to Poisson distribution

- Axioms should imply that $P\{N(t) = k\} = e^{-\lambda t}(\lambda t)^k/k!$.
- One way to prove this: divide time into n intervals of length t/n. In each, probability to see an event is $p = \lambda t/n + o(1/n)$.
- Use binomial theorem to describe probability to see event in exactly k intervals.
- Binomial formula: $\binom{n}{k} p^k (1-p)^{n-k} = \frac{n(n-1)(n-2)\dots(n-k+1)}{k!} p^k (1-p)^{n-k}$.
- This is approximately $\frac{(\lambda t)^k}{k!}(1-p)^{n-k} \approx \frac{(\lambda t)^k}{k!} e^{-\lambda t}$.
- Take n to infinity, and use fact that expected number of intervals with two or more points tends to zero (thus probability to see any intervals with two more points tends to zero).

Summary

- We constructed a random function $N(t)$ called a Poisson process of rate λ.
- For each $t > s \geq 0$, the value $N(t) - N(s)$ describes the number of events occurring in the time interval (s, t) and is Poisson with rate $(t-s)\lambda$.
- The numbers of events occurring in disjoint intervals are independent random variables.
- Let T_k be time elapsed, since the previous event, until the kth event occurs. Then the T_k are independent random variables, each of which is exponential with parameter λ.

18.600: Lecture 16

More discrete random variables

Scott Sheffield

MIT

Outline

Outline

Geometric random variables

- Consider an infinite sequence of independent tosses of a coin that comes up heads with probability p.
- Let X be such that the first heads is on the Xth toss.
- For example, if the coin sequence is $T, T, H, T, H, T, \ldots$ then $X = 3$.
- Then X is a random variable. What is $P\{X = k\}$?
- Answer: $P\{X = k\} = (1-p)^{k-1}p = q^{k-1}p$, where $q = 1 - p$ is tails probability.
- Can you prove directly that these probabilities sum to one?
- Say X is a **geometric random variable with parameter** p.

Geometric random variable expectation

- Let X be a geometric with parameter p, i.e., $P\{X = k\} = (1-p)^{k-1}p = q^{k-1}p$ for $k \geq 1$.
- What is $E[X]$?
- By definition $E[X] = \sum_{k=1}^{\infty} q^{k-1}pk$.
- There's a trick to computing sums like this.
- Note $E[X-1] = \sum_{k=1}^{\infty} q^{k-1}p(k-1)$. Setting $j = k-1$, we have $E[X-1] = q\sum_{j=0}^{\infty} q^{j-1}pj = qE[X]$.
- Kind of makes sense. $X-1$ is "number of extra tosses after first." Given first coin heads (probability p), $X-1$ is 0. Given first coin tails (probability q), conditional law of $X-1$ is geometric with parameter p. In latter case, conditional expectation of $X-1$ is same as a priori expectation of X.
- Thus $E[X] - 1 = E[X-1] = p \cdot 0 + qE[X] = qE[X]$ and solving for $E[X]$ gives $E[X] = 1/(1-q) = 1/p$.

Geometric random variable variance

- Let X be a geometric random variable with parameter p. Then $P\{X = k\} = q^{k-1}p$.
- What is $E[X^2]$?
- By definition $E[X^2] = \sum_{k=1}^{\infty} q^{k-1}pk^2$.
- Let's try to come up with a similar trick.
- Note $E[(X-1)^2] = \sum_{k=1}^{\infty} q^{k-1}p(k-1)^2$. Setting $j = k-1$, we have $E[(X-1)^2] = q\sum_{j=0}^{\infty} q^{j-1}pj^2 = qE[X^2]$.
- Thus $E[(X-1)^2] = E[X^2 - 2X + 1] = E[X^2] - 2E[X] + 1 = E[X^2] - 2/p + 1 = qE[X^2]$.
- Solving for $E[X^2]$ gives $(1-q)E[X^2] = pE[X^2] = 2/p - 1$, so $E[X^2] = (2-p)/p^2$.
- $\mathrm{Var}[X] = (2-p)/p^2 - 1/p^2 = (1-p)/p^2 = 1/p^2 - 1/p = q/p^2$.

Example

- Toss die repeatedly. Say we get 6 for first time on Xth toss.
- What is $P\{X = k\}$?
- Answer: $(5/6)^{k-1}(1/6)$.
- What is $E[X]$?
- Answer: 6.
- What is $\mathrm{Var}[X]$?
- Answer: $1/p^2 - 1/p = 36 - 6 = 30$.
- Takes $1/p$ coin tosses on average to see a heads.

Outline

Outline

Negative binomial random variables

- Consider an infinite sequence of independent tosses of a coin that comes up heads with probability p.
- Let X be such that the rth heads is on the Xth toss.
- For example, if $r = 3$ and the coin sequence is $T, T, H, H, T, T, H, T, T, \ldots$ then $X = 7$.
- Then X is a random variable. What is $P\{X = k\}$?
- Answer: need exactly $r - 1$ heads among first $k - 1$ tosses and a heads on the kth toss.
- So $P\{X = k\} = \binom{k-1}{r-1} p^{r-1}(1-p)^{k-r}p$. Can you prove these sum to 1?
- Call X **negative binomial random variable with parameters** (r, p).

Expectation of binomial random variable

- Consider an infinite sequence of independent tosses of a coin that comes up heads with probability p.
- Let X be such that the rth heads is on the Xth toss.
- Then X is a **negative binomial random variable with parameters** (r, p).
- What is $E[X]$?
- Write $X = X_1 + X_2 + \ldots + X_r$ where X_k is number of tosses (following $(k-1)$th head) required to get kth head. Each X_k is geometric with parameter p.
- So $E[X] = E[X_1 + X_2 + \ldots + X_r] = E[X_1] + E[X_2] + \ldots + E[X_r] = r/p$.
- How about $\mathrm{Var}[X]$?
- Turns out that $\mathrm{Var}[X] = \mathrm{Var}[X_1] + \mathrm{Var}[X_2] + \ldots + \mathrm{Var}[X_r]$. So $\mathrm{Var}[X] = rq/p^2$.

Outline

Outline

Problems

- Nate and Natasha have beautiful new baby. Each minute with .01 probability (independent of all else) baby cries.
- **Additivity of expectation:** How many times do they expect the baby to cry between 9 p.m. and 6 a.m.?
- **Geometric random variables:** What's the probability baby is quiet from midnight to three, then cries at exactly three?
- **Geometric random variables:** What's the probability baby is quiet from midnight to three?
- **Negative binomial:** Probability fifth cry is at midnight?
- **Negative binomial expectation:** How many minutes do I expect to wait until the fifth cry?
- **Poisson approximation:** Approximate the probability there are exactly five cries during the night.
- **Exponential random variable approximation:** Approximate probability baby quiet all night.

More fun problems

- Suppose two soccer teams play each other. One team's number of points is Poisson with parameter λ_1 and other's is independently Poisson with parameter λ_2. (You can google "soccer" and "Poisson" to see the academic literature on the use of Poisson random variables to model soccer scores.) Using Mathematica (or similar software) compute the probability that the first team wins if $\lambda_1 = 2$ and $\lambda_2 = 1$. What if $\lambda_1 = 2$ and $\lambda_2 = .5$?
- Imagine you start with the number 60. Then you toss a fair coin to decide whether to add 5 to your number or subtract 5 from it. Repeat this process with independent coin tosses until the number reaches 100 or 0. What is the *expected* number of tosses needed until this occurs?

18.600: Lecture 17

Continuous random variables

Scott Sheffield

MIT

Outline

Outline

Continuous random variables

Expectation and variance of continuous random variables

Uniform random variable on $[0, 1]$

Uniform random variable on $[\alpha, \beta]$

Measurable sets and a famous paradox

Continuous random variables

- Say X is a **continuous random variable** if there exists a **probability density function** $f = f_X$ on $\mathbb{R}$ such that $P\{X \in B\} = \int_B f(x)dx := \int 1_B(x)f(x)dx$.
- We may assume $\int_{\mathbb{R}} f(x)dx = \int_{-\infty}^{\infty} f(x)dx = 1$ and f is non-negative.
- Probability of interval $[a, b]$ is given by $\int_a^b f(x)dx$, the area under f between a and b.
- Probability of any single point is zero.
- Define **cumulative distribution function** $F(a) = F_X(a) := P\{X < a\} = P\{X \leq a\} = \int_{-\infty}^a f(x)dx$.

Simple example

- Suppose $f(x) = \begin{cases} 1/2 & x \in [0,2] \\ 0 & x \notin [0,2]. \end{cases}$
- What is $P\{X < 3/2\}$?
- What is $P\{X = 3/2\}$?
- What is $P\{1/2 < X < 3/2\}$?
- What is $P\{X \in (0,1) \cup (3/2,5)\}$?
- What is F?
- We say that X is **uniformly distributed on the interval** $[0,2]$.

Another example

- Suppose $f(x) = \begin{cases} x/2 & x \in [0,2] \\ 0 & 0 \notin [0,2]. \end{cases}$
- What is $P\{X < 3/2\}$?
- What is $P\{X = 3/2\}$?
- What is $P\{1/2 < X < 3/2\}$?
- What is F?

Outline

Outline

Expectations of continuous random variables

- Recall that when X was a discrete random variable, with $p(x) = P\{X = x\}$, we wrote
$$E[X] = \sum_{x:p(x)>0} p(x)x.$$
- How should we define $E[X]$ when X is a continuous random variable?
- Answer: $E[X] = \int_{-\infty}^{\infty} f(x)xdx$.
- Recall that when X was a discrete random variable, with $p(x) = P\{X = x\}$, we wrote
$$E[g(X)] = \sum_{x:p(x)>0} p(x)g(x).$$
- What is the analog when X is a continuous random variable?
- Answer: we will write $E[g(X)] = \int_{-\infty}^{\infty} f(x)g(x)dx$.

Variance of continuous random variables

- Suppose X is a continuous random variable with mean μ.
- We can write $\mathrm{Var}[X] = E[(X-\mu)^2]$, same as in the discrete case.
- Next, if $g = g_1 + g_2$ then $E[g(X)] = \int g_1(x)f(x)dx + \int g_2(x)f(x)dx = \int \big(g_1(x) + g_2(x)\big)f(x)dx = E[g_1(X)] + E[g_2(X)]$.
- Furthermore, $E[ag(X)] = aE[g(X)]$ when a is a constant.
- Just as in the discrete case, we can expand the variance expression as $\mathrm{Var}[X] = E[X^2 - 2\mu X + \mu^2]$ and use additivity of expectation to say that $\mathrm{Var}[X] = E[X^2] - 2\mu E[X] + E[\mu^2] = E[X^2] - 2\mu^2 + \mu^2 = E[X^2] - E[X]^2$.
- This formula is often useful for calculations.

Outline

Outline

Continuous random variables

Expectation and variance of continuous random variables

Uniform random variable on $[0,1]$

Uniform random variable on $[\alpha,\beta]$

Measurable sets and a famous paradox

Recall continuous random variable definitions

- Say X is a **continuous random variable** if there exists a **probability density function** $f = f_X$ on $\mathbb{R}$ such that $P\{X \in B\} = \int_B f(x)dx := \int 1_B(x)f(x)dx$.
- We may assume $\int_{\mathbb{R}} f(x)dx = \int_{-\infty}^{\infty} f(x)dx = 1$ and f is non-negative.
- Probability of interval $[a,b]$ is given by $\int_a^b f(x)dx$, the area under f between a and b.
- Probability of any single point is zero.
- Define **cumulative distribution function** $F(a) = F_X(a) := P\{X < a\} = P\{X \le a\} = \int_{-\infty}^a f(x)dx$.

Uniform random variables on $[0,1]$

- Suppose X is a random variable with probability density function $f(x) = \begin{cases} 1 & x \in [0,1] \\ 0 & x \notin [0,1]. \end{cases}$
- Then for any $0 \le a \le b \le 1$ we have $P\{X \in [a,b]\} = b - a$.
- Intuition: all locations along the interval $[0,1]$ equally likely.
- Say that X is a **uniform random variable on** $[0,1]$ or that X is **sampled uniformly from** $[0,1]$.

Properties of uniform random variable on $[0,1]$

- Suppose X is a random variable with probability density function $f(x) = \begin{cases} 1 & x \in [0,1] \\ 0 & x \notin [0,1], \end{cases}$ which implies

$$F_X(a) = \begin{cases} 0 & a < 0 \\ a & a \in [0,1] \\ 1 & a > 1 \end{cases}.$$

- What is $E[X]$?
- Guess $1/2$ (since $1/2$ is, you know, in the middle).
- Indeed, $\int_{-\infty}^{\infty} f(x)xdx = \int_0^1 xdx = \frac{x^2}{2}\Big|_0^1 = 1/2$.
- What is the general moment $E[X^k]$ for $k \ge 0$?
- Answer: $1/(k+1)$.
- What would you guess the variance is? Expected square of distance from $1/2$?
- It's obviously less than $1/4$, but how much less?
- $\mathrm{Var} E[X^2] - E[X]^2 = 1/3 - 1/4 = 1/12$.

Outline

Outline

Continuous random variables

Expectation and variance of continuous random variables

Uniform random variable on $[0,1]$

Uniform random variable on $[\alpha,\beta]$

Measurable sets and a famous paradox

Uniform random variables on $[\alpha,\beta]$

- Fix $\alpha<\beta$ and suppose X is a random variable with probability density function $f(x)=\begin{cases}\frac{1}{\beta-\alpha} & x\in[\alpha,\beta]\\ 0 & x\notin[\alpha,\beta].\end{cases}$
- Then for any $\alpha\le a\le b\le\beta$ we have $P\{X\in[a,b]\}=\frac{b-a}{\beta-\alpha}$.
- Intuition: all locations along the interval $[\alpha,\beta]$ are equally likely.
- Say that X is a **uniform random variable on** $[\alpha,\beta]$ or that X is **sampled uniformly from** $[\alpha,\beta]$.

Uniform random variables on $[\alpha,\beta]$

- Suppose X is a random variable with probability density function $f(x)=\begin{cases}\frac{1}{\beta-\alpha} & x\in[\alpha,\beta]\\ 0 & x\notin[\alpha,\beta].\end{cases}$
- What is $E[X]$?
- Intuitively, we'd guess the midpoint $\frac{\alpha+\beta}{2}$.
- What's the cleanest way to prove this?
- One approach: let Y be uniform on $[0,1]$ and try to show that $X=(\beta-\alpha)Y+\alpha$ is uniform on $[\alpha,\beta]$.
- Then expectation linearity gives $E[X]=(\beta-\alpha)E[Y]+\alpha=(1/2)(\beta-\alpha)+\alpha=\frac{\alpha+\beta}{2}$.
- Using similar logic, what is the variance $\mathrm{Var}[X]$?
- Answer: $\mathrm{Var}[X]=\mathrm{Var}[(\beta-\alpha)Y+\alpha]=\mathrm{Var}[(\beta-\alpha)Y]=(\beta-\alpha)^2\mathrm{Var}[Y]=(\beta-\alpha)^2/12$.

Outline

Continuous random variables

Expectation and variance of continuous random variables

Uniform random variable on $[0,1]$

Uniform random variable on $[\alpha,\beta]$

Measurable sets and a famous paradox

Outline

Uniform measure on $[0, 1]$

- One of the very simplest probability density functions is $f(x) = \begin{cases} 1 & x \in [0,1] \\ 0 & 0 \notin [0,1]. \end{cases}$.
- If $B \subset [0, 1]$ is an interval, then $P\{X \in B\}$ is the length of that interval.
- Generally, if $B \subset [0, 1]$ then $P\{X \in B\} = \int_B 1dx = \int 1_B(x)dx$ is the "total volume" or "total length" of the set B.
- What if B is the set of all rational numbers?
- How do we mathematically define the volume of an arbitrary set B?

Idea behind parodox

- What if we could partition $[0, 1]$ into a countably infinite collection of disjoint sets that all looked the same (up to a translation, say) and thus all had to have the same probability?
- Well, if that probability was zero, then (by countable additivity) probability of whole interval would be zero, a contradiction.
- But if that probability were a number greater than zero the probability of whole interval would be infinite, also a contradiction...
- Related problem: if you can cut a cake into countably infinitely many pieces all of the same weight, how much does each piece weigh?

Formulating the paradox precisely

- Uniform probability measure on $[0, 1)$ should satisfy **translation invariance:** If B and a horizontal translation of B are both subsets $[0, 1)$, their probabilities should be equal.
- Consider **wrap-around translations** $\tau_r(x) = (x + r) \mod 1$.
- By translation invariance, $\tau_r(B)$ has same probability as B.
- Call x, y "equivalent modulo rationals" if $x - y$ is rational (e.g., $x = \pi - 3$ and $y = \pi - 9/4$). An **equivalence class** is the set of points in $[0, 1)$ equivalent to some given point.
- There are uncountably many of these classes.
- Let $A \subset [0, 1)$ contain **one** point from each class. For each $x \in [0, 1)$, there is **one** $a \in A$ such that $r = x - a$ is rational.
- Then each x in $[0, 1)$ lies in $\tau_r(A)$ for **one** rational $r \in [0, 1)$.
- Thus $[0, 1) = \cup \tau_r(A)$ as r ranges over rationals in $[0, 1)$.
- If $P(A) = 0$, then $P(S) = \sum_r P(\tau_r(A)) = 0$. If $P(A) > 0$ then $P(S) = \sum_r P(\tau_r(A)) = \infty$. Contradicts $P(S) = 1$ axiom.

Three ways to get around this

- 1. **Re-examine axioms of mathematics:** the very *existence* of a set A with one element from each equivalence class is consequence of so-called **axiom of choice**. Removing that axiom makes paradox goes away, since one can just suppose (pretend?) these kinds of sets don't exist.
- 2. **Re-examine axioms of probability:** Replace *countable additivity* with *finite additivity*? (Doesn't fully solve problem: look up Banach-Tarski.)
- 3. **Keep the axiom of choice and countable additivity but don't define probabilities of all sets:** Instead of defining $P(B)$ for *every* subset B of sample space, restrict attention to a family of so-called **"measurable"** sets.
- Most mainstream probability and analysis takes the third approach.
- In practice, sets we care about (e.g., countable unions of points and intervals) tend to be measurable.

Perspective

- More advanced courses in probability and analysis (such as 18.125 and 18.175) spend a significant amount of time rigorously constructing a class of so-called **measurable sets** and the so-called **Lebesgue measure**, which assigns a real number (a measure) to each of these sets.
- These courses also replace the **Riemann integral** with the so-called **Lebesgue integral**.
- We will not treat these topics any further in this course.
- We usually limit our attention to probability density functions f and sets B for which the ordinary Riemann integral $\int 1_B(x)f(x)dx$ is well defined.
- Riemann integration is a mathematically rigorous theory. It's just not as robust as Lebesgue integration.

18.600: Lecture 18

Normal random variables

Scott Sheffield

MIT

Outline

Outline

Tossing coins

- Suppose we toss a million fair coins. How many heads will we get?
- About half a million, yes, but how close to that? Will we be off by 10 or 1000 or 100,000?
- How can we describe the error?
- Let's try this out.

Tossing coins

- Toss n coins. What is probability to see k heads?
- Answer: $2^{-k}\binom{n}{k}$.
- Let's plot this for a few values of n.
- Seems to look like it's converging to a curve.
- If we replace fair coin with p coin, what's probability to see k heads.
- Answer: $p^k(1-p)^{n-k}\binom{n}{k}$.
- Let's plot this for $p = 2/3$ and some values of n.
- What does limit shape seem to be?

Outline

Outline

Standard normal random variable

- Say X is a (standard) **normal random variable** if $f_X(x) = f(x) = \frac{1}{\sqrt{2\pi}}e^{-x^2/2}$.
- Clearly f is always non-negative for real values of x, but how do we show that $\int_{-\infty}^{\infty} f(x)dx = 1$?
- Looks kind of tricky.
- Happens to be a nice trick. Write $I = \int_{-\infty}^{\infty} e^{-x^2/2}dx$. Then try to compute I^2 as a two dimensional integral.
- That is, write

$$I^2 = \int_{-\infty}^{\infty} e^{-x^2/2}dx \int_{-\infty}^{\infty} e^{-y^2/2}dy = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-x^2/2}dxe^{-y^2/2}dy.$$

- Then switch to polar coordinates.

$$I^2 = \int_0^{\infty}\int_0^{2\pi} e^{-r^2/2}rd\theta dr = 2\pi\int_0^{\infty} re^{-r^2/2}dr = -2\pi e^{-r^2/2}\Big|_0^{\infty},$$

so $I = \sqrt{2\pi}$.

Standard normal random variable mean and variance

- Say X is a (standard) **normal random variable** if $f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$.
- Question: what are mean and variance of X?
- $E[X] = \int_{-\infty}^{\infty} x f(x) dx$. Can see by symmetry that this zero.
- Or can compute directly:

$$E[X] = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} x dx = \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \Big|_{-\infty}^{\infty} = 0.$$

- How would we compute $\mathrm{Var}[X] = \int f(x) x^2 dx = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} x^2 dx$?
- Try integration by parts with $u = x$ and $dv = xe^{-x^2/2}dx$. Find that $\mathrm{Var}[X] = \frac{1}{\sqrt{2\pi}} (-xe^{-x^2/2}\Big|_{-\infty}^{\infty} + \int_{-\infty}^{\infty} e^{-x^2/2} dx) = 1$.

General normal random variables

- Again, X is a (standard) **normal random variable** if $f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$.
- What about $Y = \sigma X + \mu$? Can we "stretch out" and "translate" the normal distribution (as we did last lecture for the uniform distribution)?
- Say Y is normal with parameters μ and σ^2 if $f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2}$.
- What are the mean and variance of Y?
- $E[Y] = E[X] + \mu = \mu$ and $\mathrm{Var}[Y] = \sigma^2 \mathrm{Var}[X] = \sigma^2$.

Cumulative distribution function

- Again, X is a standard normal random variable if $f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$.
- What is the cumulative distribution function?
- Write this as $F_X(a) = P\{X \le a\} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{a} e^{-x^2/2} dx$.
- How can we compute this integral explicitly?
- Can't. Let's just give it a name. Write $\Phi(a) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{a} e^{-x^2/2} dx$.
- Values: $\Phi(-3) \approx .0013$, $\Phi(-2) \approx .023$ and $\Phi(-1) \approx .159$.
- Rough rule of thumb: "two thirds of time within one SD of mean, 95 percent of time within 2 SDs of mean."

Outline

Outline

DeMoivre-Laplace Limit Theorem

- Let S_n be number of heads in n tosses of a p coin.
- What's the standard deviation of S_n?
- Answer: $\sqrt{npq}$ (where $q = 1 - p$).
- The special quantity $\frac{S_n - np}{\sqrt{npq}}$ describes the number of standard deviations that S_n is above or below its mean.
- What's the mean and variance of this special quantity? Is it roughly normal?
- **DeMoivre-Laplace limit theorem (special case of central limit theorem):**

$$\lim_{n\to\infty} P\{a \le \frac{S_n - np}{\sqrt{npq}} \le b\} \to \Phi(b) - \Phi(a).$$

- This is $\Phi(b) - \Phi(a) = P\{a \le X \le b\}$ when X is a standard normal random variable.

Problems

- Toss a million fair coins. Approximate the probability that I get more than $501,000$ heads.
- Answer: well, $\sqrt{npq} = \sqrt{10^6 \times .5 \times .5} = 500$. So we're asking for probability to be over two SDs above mean. This is approximately $1 - \Phi(2) = \Phi(-2) \approx .159$.
- Roll 60000 dice. Expect to see 10000 sixes. What's the probability to see more than 9800?
- Here $\sqrt{npq} = \sqrt{60000 \times \frac{1}{6} \times \frac{5}{6}} \approx 91.28$.
- And $200/91.28 \approx 2.19$. Answer is about $1 - \Phi(-2.19)$.

18.600: Lecture 19

Exponential random variables

Scott Sheffield

MIT

Outline

Outline

Exponential random variables

Minimum of independent exponentials

Memoryless property

Relationship to Poisson random variables

Exponential random variables

- Say X is an **exponential random variable of parameter λ** when its probability distribution function is

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0 \\ 0 & x < 0 \end{cases}.$$

- For $a > 0$ have

$$F_X(a) = \int_0^a f(x)dx = \int_0^a \lambda e^{-\lambda x} dx = -e^{-\lambda x}\Big|_0^a = 1 - e^{-\lambda a}.$$

- Thus $P\{X < a\} = 1 - e^{-\lambda a}$ and $P\{X > a\} = e^{-\lambda a}$.
- Formula $P\{X > a\} = e^{-\lambda a}$ is very important in practice.

Moment formula

- Suppose X is exponential with parameter λ, so $f_X(x) = \lambda e^{-\lambda x}$ when $x \geq 0$.
- What is $E[X^n]$? (Say $n \geq 1$.)
- Write $E[X^n] = \int_0^\infty x^n \lambda e^{-\lambda x} dx$.
- Integration by parts gives $E[X^n] = -\int_0^\infty n x^{n-1} \lambda \frac{e^{-\lambda x}}{-\lambda} dx + x^n \lambda \frac{e^{-\lambda x}}{-\lambda} \Big|_0^\infty$.
- We get $E[X^n] = \frac{n}{\lambda} E[X^{n-1}]$.
- $E[X^0] = E[1] = 1$, $E[X] = 1/\lambda$, $E[X^2] = 2/\lambda^2$, $E[X^n] = n!/\lambda^n$.
- If $\lambda = 1$, then $E[X^n] = n!$. Could take this as definition of $n!$. It makes sense for $n = 0$ and for non-integer n.
- Variance: $\mathrm{Var}[X] = E[X^2] - (E[X])^2 = 1/\lambda^2$.

Outline

Outline

Minimum of independent exponentials is exponential

- **CLAIM:** If X_1 and X_2 are independent and exponential with parameters λ_1 and λ_2 then $X = \min\{X_1, X_2\}$ is exponential with parameter $\lambda = \lambda_1 + \lambda_2$.
- How could we prove this?
- Have various ways to describe random variable Y: via density function $f_Y(x)$, or cumulative distribution function $F_Y(a) = P\{Y \leq a\}$, or function $P\{Y > a\} = 1 - F_Y(a)$.
- Last one has simple form for exponential random variables. We have $P\{Y > a\} = e^{-\lambda a}$ for $a \in [0, \infty)$.
- Note: $X > a$ if and only if $X_1 > a$ *and* $X_2 > a$.
- X_1 and X_2 are independent, so $P\{X > a\} = P\{X_1 > a\}P\{X_2 > a\} = e^{-\lambda_1 a} e^{-\lambda_2 a} = e^{-\lambda a}$.
- If $X_1, \ldots, X_n$ are independent exponential with $\lambda_1, \ldots \lambda_n$, then $\min\{X_1, \ldots X_n\}$ is exponential with $\lambda = \lambda_1 + \ldots + \lambda_n$.

Outline

Exponential random variables

Minimum of independent exponentials

Memoryless property

Relationship to Poisson random variables

Outline

Exponential random variables

Minimum of independent exponentials

Memoryless property

Relationship to Poisson random variables

Memoryless property

- Suppose X is exponential with parameter λ.
- **Memoryless property:** If X represents the time until an event occurs, then *given* that we have seen no event up to time b, the conditional distribution of the remaining time till the event is the same as it originally was.
- To make this precise, we ask what is the probability distribution of $Y = X - b$ *conditioned on* $X > b$?
- We can characterize the conditional law of Y, given $X > b$, by computing $P(Y > a|X > b)$ for each a.
- That is, we compute $P(X - b > a|X > b) = P(X > b + a|X > b)$.
- By definition of conditional probability, this is just $P\{X > b + a\}/P\{X > b\} = e^{-\lambda(b+a)}/e^{-\lambda b} = e^{-\lambda a}$.
- Thus, conditional law of $X - b$ *given* that $X > b$ is same as the original law of X.

Memoryless property for geometric random variables

- Similar property holds for geometric random variables.
- If we plan to toss a coin until the first heads comes up, then we have a .5 chance to get a heads in one step, a .25 chance in two steps, etc.
- Given that the first 5 tosses are all tails, there is conditionally a .5 chance we get our first heads on the 6th toss, a .25 chance on the 7th toss, etc.
- Despite our having had five tails in a row, our expectation of the amount of time remaining until we see a heads is the same as it originally was.

Exchange overheard on Logan airport shuttle

- **Bob:** There's this really interesting problem in statistics I just learned about. If a coin comes up heads 10 times in a row, how likely is the next toss to be heads?
- **Alice:** Still fifty fifty.
- **Bob:** That's a common mistake, but you're wrong because the 10 heads in a row increase the conditional probability that there's something funny going on with the coin.
- **Alice:** You never said it might be a funny coin.
- **Bob:** That's the point. You should always suspect that there might be something funny with the coin.
- **Alice:** It's a math puzzle. You always assume a normal coin.
- **Bob:** No, that's your mistake. You should never assume that, because maybe somebody tampered with the coin.

Exchange overheard on a Logan airport shuttle

- **Alice:** Yeah, yeah, I get it. I can't win here.
- **Bob:** No, I don't think you get it yet. It's a subtle point in statistics. It's very important.
- Exchange continued for duration of shuttle ride (Alice increasingly irritated, Bob increasingly patronizing).
- Raises interesting question about memoryless property.
- Suppose the duration of a couple's relationship is exponential with λ^{-1} equal to two weeks.
- Given that it has lasted for 10 weeks so far, what is the conditional probability that it will last an additional week?
- How about an additional four weeks? Ten weeks?

Remark on Alice and Bob

- Alice assumes Bob means "independent tosses of a fair coin." Under this assumption, all 2^{11} outcomes of eleven-coin-toss sequence are equally likely. Bob considers HHHHHHHHHHH more likely than HHHHHHHHHHT, since former could result from a faulty coin.
- Alice sees Bob's point but considers it annoying and churlish to ask about coin toss sequence and criticize listener for assuming this means "independent tosses of fair coin".
- Without that assumption, Alice has no idea what context Bob has in mind. (An environment where two-headed novelty coins are common? Among coin-tossing cheaters with particular agendas?...)
- Alice: you need assumptions to convert stories into math.
- Bob: good to question assumptions.

Radioactive decay: maximum of independent exponentials

- Suppose you start at time zero with n radioactive particles. Suppose that each one (independently of the others) will decay at a random time, which is an exponential random variable with parameter λ.
- Let T be amount of time until no particles are left. What are $E[T]$ and $\mathrm{Var}[T]$?
- Let T_1 be the amount of time you wait until the first particle decays, T_2 the amount of *additional* time until the second particle decays, etc., so that $T = T_1 + T_2 + \dots T_n$.
- Claim: T_1 is exponential with parameter $n\lambda$.
- Claim: T_2 is exponential with parameter $(n-1)\lambda$.
- And so forth. $E[T] = \sum_{i=1}^n E[T_i] = \lambda^{-1}\sum_{j=1}^n \frac{1}{j}$ and (by independence) $\mathrm{Var}[T] = \sum_{i=1}^n \mathrm{Var}[T_i] = \lambda^{-2}\sum_{j=1}^n \frac{1}{j^2}$.

Outline

Exponential random variables

Minimum of independent exponentials

Memoryless property

Relationship to Poisson random variables

Outline

Exponential random variables

Minimum of independent exponentials

Memoryless property

Relationship to Poisson random variables

Relationship to Poisson random variables

- Let $T_1, T_2, \ldots$ be independent exponential random variables with parameter λ.
- We can view them as waiting times between "events".
- How do you show that the number of events in the first t units of time is Poisson with parameter λt?
- We actually did this already in the lecture on Poisson point processes. You can break the interval $[0, t]$ into n equal pieces (for very large n), let X_k be number of events in kth piece, use memoryless property to argue that the X_k are independent.
- When n is large enough, it becomes unlikely that any interval has more than one event. Roughly speaking: each interval has one event with probability $\lambda t/n$, zero otherwise.
- Take $n \to \infty$ limit. Number of events is Poisson λt.

18.600: Lecture 20

More continuous random variables

Scott Sheffield

MIT

Outline

Outline

Defining gamma function Γ

- Last time we found that if X is geometric with rate 1 and $n \geq 0$ then $E[X^n] = \int_0^\infty x^n e^{-x} dx = n!$.
- This expectation $E[X^n]$ is actually well defined whenever $n > -1$. Set $\alpha = n + 1$. The following quantity is well defined for any $\alpha > 0$:
  $\Gamma(\alpha) := E[X^{\alpha-1}] = \int_0^\infty x^{\alpha-1} e^{-x} dx = (\alpha - 1)!$.
- So $\Gamma(\alpha)$ extends the function $(\alpha - 1)!$ (as defined for *strictly positive* integers α) to the positive reals.
- Vexing notational issue: why define Γ so that $\Gamma(\alpha) = (\alpha - 1)!$ instead of $\Gamma(\alpha) = \alpha!$?
- At least it's kind of convenient that Γ is defined on $(0, \infty)$ instead of $(-1, \infty)$.

Recall: geometric and negative binomials

- The sum X of n independent geometric random variables of parameter p is negative binomial with parameter (n, p).
- Waiting for the nth heads. What is $P\{X = k\}$?
- Answer: $\binom{k-1}{n-1} p^{n-1}(1-p)^{k-n}p$.
- What's the continuous (Poisson point process) version of "waiting for the nth event"?

Poisson point process limit

- Recall that we can approximate a Poisson process of rate λ by tossing N coins per time unit and taking $p = \lambda/N$.
- Let's fix a rational number x and try to figure out the probability that that the nth coin toss happens at time x (i.e., on exactly xNth trials, assuming xN is an integer).
- Write $p = \lambda/N$ and $k = xN$. (Note $p = \lambda x/k$.)
- For large N, $\binom{k-1}{n-1} p^{n-1}(1-p)^{k-n}p$ is

$$\frac{(k-1)(k-2)\dots(k-n+1)}{(n-1)!} p^{n-1}(1-p)^{k-n}p$$

$$\approx \frac{k^{n-1}}{(n-1)!} p^{n-1} e^{-x\lambda} p = \frac{1}{N}\Big(\frac{(\lambda x)^{(n-1)} e^{-\lambda x}\lambda}{(n-1)!}\Big).$$

Defining Γ distribution

- The probability from previous side, $\frac{1}{N}\Big(\frac{(\lambda x)^{(n-1)} e^{-\lambda x}\lambda}{(n-1)!}\Big)$ suggests the form for a continuum random variable.
- Replace n (generally integer valued) with α (which we will eventually allow be to be any real number).
- Say that random variable X has gamma distribution with parameters (α, λ) if $f_X(x) = \begin{cases} \frac{(\lambda x)^{\alpha-1} e^{-\lambda x}\lambda}{\Gamma(\alpha)} & x \ge 0 \\ 0 & x < 0 \end{cases}$.
- Waiting time interpretation makes sense only for integer α, but distribution is defined for general positive α.

Outline

Outline

Cauchy distribution

- A standard **Cauchy random variable** is a random real number with probability density $f(x) = \frac{1}{\pi}\frac{1}{1+x^2}$.
- There is a "spinning flashlight" interpretation. Put a flashlight at $(0,1)$, spin it to a uniformly random angle in $[-\pi/2, \pi/2]$, and consider point X where light beam hits the x-axis.
- $F_X(x) = P\{X \le x\} = P\{\tan\theta \le x\} = P\{\theta \le tan^{-1}x\} = \frac{1}{2} + \frac{1}{\pi}\tan^{-1}x$.
- Find $f_X(x) = \frac{d}{dx}F(x) = \frac{1}{\pi}\frac{1}{1+x^2}$.

Cauchy distribution: Brownian motion interpretation

- The light beam travels in (randomly directed) straight line. There's a windier random path called Brownian motion.
- If you do a simple random walk on a grid and take the grid size to zero, then you get Brownian motion as a limit.
- We will not give a complete mathematical description of Brownian motion here, just one nice fact.
- FACT: start Brownian motion at point (x,y) in the upper half plane. Probability it hits negative x-axis before positive x-axis is $\frac{1}{2} + \frac{1}{\pi}\tan^{-1}\frac{y}{x}$. Linear function of angle between positive x-axis and line through $(0,0)$ and (x,y).
- Start Brownian motion at $(0,1)$ and let X be the location of the first point on the x-axis it hits. What's $P\{X < a\}$?
- Applying FACT, translation invariance, reflection symmetry: $P\{X < x\} = P\{X > -x\} = \frac{1}{2} + \frac{1}{\pi}\tan^{-1}\frac{1}{x}$.
- So X is a standard Cauchy random variable.

Question: what if we start at $(0,2)$?

- Start at $(0,2)$. Let Y be first point on x-axis hit by Brownian motion. Again, same probability distribution as point hit by flashlight trajectory.
- Flashlight point of view: Y has the same law as $2X$ where X is standard Cauchy.
- Brownian point of view: Y has same law as $X_1 + X_2$ where X_1 and X_2 are standard Cauchy.
- But wait a minute. $\mathrm{Var}(Y) = 4\mathrm{Var}(X)$ and by independence $\mathrm{Var}(X_1 + X_2) = \mathrm{Var}(X_1) + \mathrm{Var}(X_2) = 2\mathrm{Var}(X_2)$. Can this be right?
- Cauchy distribution doesn't have finite variance or mean.
- Some standard facts we'll learn later in the course (central limit theorem, law of large numbers) don't apply to it.

Outline

Gamma distribution

Cauchy distribution

Beta distribution

Outline

Gamma distribution

Cauchy distribution

Beta distribution

Beta distribution: Alice and Bob revisited

- Suppose I have a coin with a heads probability p that I don't know much about.
- What do I mean by not knowing anything? Let's say that I think p is equally likely to be any of the numbers $\{0, .1, .2, .3, .4, \ldots, .9, 1\}$.
- Now imagine a multi-stage experiment where I first choose p and then I toss n coins.
- Given that number h of heads is $a - 1$, and $b - 1$ tails, what's *conditional* probability p was a certain value x?
- $P\Big(p = x | h = (a-1)\Big) = \frac{\frac{1}{11}\binom{n}{a-1}x^{a-1}(1-x)^{b-1}}{P\{h=(a-1)\}}$ which is $x^{a-1}(1-x)^{b-1}$ times a constant that doesn't depend on x.

Beta distribution

- Suppose I have a coin with a heads probability p that I *really* don't know anything about. Let's say p is uniform on $[0, 1]$.
- Now imagine a multi-stage experiment where I first choose p uniformly from $[0, 1]$ and then I toss n coins.
- If I get, say, $a - 1$ heads and $b - 1$ tails, then what is the *conditional* probability density for p?
- Turns out to be a constant (that doesn't depend on x) times $x^{a-1}(1-x)^{b-1}$.
- $\frac{1}{B(a,b)}x^{a-1}(1-x)^{b-1}$ on $[0, 1]$, where $B(a, b)$ is constant chosen to make integral one. Can be shown that $B(a, b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$.
- What is $E[X]$?
- Answer: $\frac{a}{a+b}$.

18.600: Lecture 21

Joint distributions functions

Scott Sheffield

MIT

Outline

Outline

Distribution of function of random variable

- Suppose $P\{X \leq a\} = F_X(a)$ is known for all a. Write $Y = X^3$. What is $P\{Y \leq 27\}$?
- Answer: note that $Y \leq 27$ if and only if $X \leq 3$. Hence $P\{Y \leq 27\} = P\{X \leq 3\} = F_X(3)$.
- Generally $F_Y(a) = P\{Y \leq a\} = P\{X \leq a^{1/3}\} = F_X(a^{1/3})$
- This is a general principle. If X is a continuous random variable and g is a strictly increasing function of x and $Y = g(X)$, then $F_Y(a) = F_X(g^{-1}(a))$.
- How can we use this to compute the probability density function f_Y from f_X?
- If $Z = X^2$, then what is $P\{Z \leq 16\}$?

Outline

Distributions of functions of random variables

Joint distributions

Independent random variables

Examples

Outline

Distributions of functions of random variables

Joint distributions

Independent random variables

Examples

Joint probability mass functions: discrete random variables

- If X and Y assume values in $\{1, 2, \ldots, n\}$ then we can view $A_{i,j} = P\{X = i, Y = j\}$ as the entries of an $n \times n$ matrix.
- Let's say I don't care about Y. I just want to know $P\{X = i\}$. How do I figure that out from the matrix?
- Answer: $P\{X = i\} = \sum_{j=1}^{n} A_{i,j}$.
- Similarly, $P\{Y = j\} = \sum_{i=1}^{n} A_{i,j}$.
- In other words, the probability mass functions for X and Y are the row and columns sums of $A_{i,j}$.
- Given the joint distribution of X and Y, we sometimes call distribution of X (ignoring Y) and distribution of Y (ignoring X) the **marginal** distributions.
- In general, when X and Y are jointly defined discrete random variables, we write $p(x, y) = p_{X,Y}(x, y) = P\{X = x, Y = y\}$.

Joint distribution functions: continuous random variables

- Given random variables X and Y, define $F(a, b) = P\{X \leq a, Y \leq b\}$.
- The region $\{(x, y) : x \leq a, y \leq b\}$ is the lower left "quadrant" centered at (a, b).
- Refer to $F_X(a) = P\{X \leq a\}$ and $F_Y(b) = P\{Y \leq b\}$ as **marginal** cumulative distribution functions.
- Question: if I tell you the two parameter function F, can you use it to determine the marginals F_X and F_Y?
- Answer: Yes. $F_X(a) = \lim_{b \to \infty} F(a, b)$ and $F_Y(b) = \lim_{a \to \infty} F(a, b)$.

Joint density functions: continuous random variables

- Suppose we are given the joint distribution function $F(a,b) = P\{X \le a, Y \le b\}$.
- Can we use F to construct a "two-dimensional probability density function"? Precisely, is there a function f such that $P\{(X,Y) \in A\} = \int_A f(x,y)dxdy$ for each (measurable) $A \subset \mathbb{R}^2$?
- Let's try defining $f(x,y) = \frac{\partial}{\partial x}\frac{\partial}{\partial y}F(x,y)$. Does that work?
- Suppose first that $A = \{(x,y) : x \le a, \le b\}$. By definition of F, fundamental theorem of calculus, fact that $F(a,b)$ vanishes as either a or b tends to $-\infty$, we indeed find $\int_{-\infty}^{b}\int_{-\infty}^{a}\frac{\partial}{\partial x}\frac{\partial}{\partial y}F(x,y)dxdy = \int_{-\infty}^{b}\frac{\partial}{\partial y}F(a,y)dy = F(a,b)$.
- From this, we can show that it works for strips, rectangles, general open sets, etc.

Outline

Outline

Independent random variables

- We say X and Y are independent if for any two (measurable) sets A and B of real numbers we have

$$P\{X \in A, Y \in B\} = P\{X \in A\}P\{Y \in B\}.$$

- Intuition: knowing something about X gives me no information about Y, and vice versa.
- When X and Y are discrete random variables, they are independent if $P\{X = x, Y = y\} = P\{X = x\}P\{Y = y\}$ for all x and y for which $P\{X = x\}$ and $P\{Y = y\}$ are non-zero.
- What is the analog of this statement when X and Y are continuous?
- When X and Y are continuous, they are independent if $f(x,y) = f_X(x)f_Y(y)$.

Sample problem: independent normal random variables

- Suppose that X and Y are independent normal random variables with mean zero and variance one.
- What is the probability that (X, Y) lies in the unit circle? That is, what is $P\{X^2 + Y^2 \leq 1\}$?
- First, any guesses?
- Probability X is within one standard deviation of its mean is about .68. So $(.68)^2$ is an upper bound.
- $f(x, y) = f_X(x)f_Y(y) = \frac{1}{\sqrt{2\pi}}e^{-x^2/2}\frac{1}{\sqrt{2\pi}}e^{-y^2/2} = \frac{1}{2\pi}e^{-r^2/2}$
- Using polar coordinates, we want $\int_0^1 (2\pi r)\frac{1}{2\pi}e^{-r^2/2}dr = -e^{-r^2/2}\big|_0^1 = 1 - e^{-1/2} \approx .39$.

Outline

Outline

Repeated die roll

- Roll a die repeatedly and let X be such that the first even number (the first 2, 4, or 6) appears on the Xth roll.
- Let Y be the the number that appears on the Xth roll.
- Are X and Y independent? What is their joint law?
- If $j \geq 1$, then

$$P\{X = j, Y = 2\} = P\{X = j, Y = 4\}$$

$$= P\{X = j, Y = 6\} = (1/2)^{j-1}(1/6) = (1/2)^j(1/3).$$

- Can we get the marginals from that?

Continuous time variant of repeated die roll

- On a certain hiking trail, it is well known that the lion, tiger, and bear attacks are independent Poisson processes with respective λ values of $.1/\mathrm{hour}$, $.2/\mathrm{hour}$, and $.3/\mathrm{hour}$.
- Let $T \in \mathbb{R}$ be the amount of time until the first animal attacks. Let $A \in \{\mathrm{lion}, \mathrm{tiger}, \mathrm{bear}\}$ be the species of the first attacking animal.
- What is the probability density function for T? How about $E[T]$?
- Are T and A independent?
- Let T_1 be the time until the first attack, T_2 the subsequent time until the second attack, etc., and let $A_1, A_2, \ldots$ be the corresponding species.
- Are all of the T_i and A_i independent of each other? What are their probability distributions?

Buffon's needle problem

- Drop a needle of length one on a large sheet of paper (with evenly spaced horizontal lines spaced at all integer heights).
- What's the probability the needle crosses a line?
- Need some assumptions. Let's say vertical position X of lowermost endpoint of needle modulo one is uniform in $[0, 1]$ and independent of angle θ, which is uniform in $[0, \pi]$. Crosses line if and only there is an integer between the numbers X and $X + \sin\theta$, i.e., $X \leq 1 \leq X + \sin\theta$.
- Draw the box $[0, 1] \times [0, \pi]$ on which (X, θ) is uniform. What's the area of the subset where $X \geq 1 - \sin\theta$?

More lions, tigers, bears

- Lion, tiger, and bear attacks are independent Poisson processes with λ values $.1/\mathrm{hour}$, $.2/\mathrm{hour}$, and $.3/\mathrm{hour}$.
- Distribution of time T_{tiger} till first tiger attack?
- Exponential $\lambda_{\mathrm{tiger}} = .2/\mathrm{hour}$. So $P\{T_{\mathrm{tiger}} > a\} = e^{-.2a}$.
- How about $E[T_{\mathrm{tiger}}]$ and $\mathrm{Var}[T_{\mathrm{tiger}}]$?
- $E[T_{\mathrm{tiger}}] = 1/\lambda_{\mathrm{tiger}} = 5$ hours, $\mathrm{Var}[T_{\mathrm{tiger}}] = 1/\lambda_{\mathrm{tiger}}^2 = 25$ hours squared.
- Time until 5th attack by any animal?
- Γ distribution with $\alpha = 5$ and $\lambda = .6$.
- X, where Xth attack is 5th bear attack?
- Negative binomial with parameters $p = 1/2$ and $n = 5$.
- Can hiker breathe sigh of relief after 5 attack-free hours?

18.600: Lecture 22

Sums of independent random variables

Scott Sheffield

MIT

Summing two random variables

- Say we have independent random variables X and Y and we know their density functions f_X and f_Y.
- Now let's try to find $F_{X+Y}(a) = P\{X + Y \leq a\}$.
- This is the integral over $\{(x, y) : x + y \leq a\}$ of $f(x, y) = f_X(x)f_Y(y)$. Thus,
- $$P\{X + Y \leq a\} = \int_{-\infty}^{\infty} \int_{-\infty}^{a-y} f_X(x)f_Y(y)dxdy$$
  $$= \int_{-\infty}^{\infty} F_X(a - y)f_Y(y)dy.$$
- Differentiating both sides gives $f_{X+Y}(a) = \frac{d}{da} \int_{-\infty}^{\infty} F_X(a-y)f_Y(y)dy = \int_{-\infty}^{\infty} f_X(a-y)f_Y(y)dy$.
- Latter formula makes some intuitive sense. We're integrating over the set of x, y pairs that add up to a.

Independent identically distributed (i.i.d.)

- The abbreviation i.i.d. means independent identically distributed.
- It is actually one of the most important abbreviations in probability theory.
- Worth memorizing.

Summing i.i.d. uniform random variables

- Suppose that X and Y are i.i.d. and uniform on $[0, 1]$. So $f_X = f_Y = 1$ on $[0, 1]$.
- What is the probability density function of $X + Y$?
- $f_{X+Y}(a) = \int_{-\infty}^{\infty} f_X(a - y)f_Y(y)dy = \int_0^1 f_X(a - y)$ which is the length of $[0, 1] \cap [a - 1, a]$.
- That's a when $a \in [0, 1]$ and $2 - a$ when $a \in [1, 2]$ and 0 otherwise.

Review: summing i.i.d. geometric random variables

- A geometric random variable X with parameter p has $P\{X = k\} = (1-p)^{k-1}p$ for $k \geq 1$.
- Sum Z of n independent copies of X?
- We can interpret Z as time slot where nth head occurs in i.i.d. sequence of p-coin tosses.
- So Z is negative binomial (n, p). So $P\{Z = k\} = \binom{k-1}{n-1}p^{n-1}(1-p)^{k-n}p$.

Summing i.i.d. exponential random variables

- Suppose $X_1, \ldots X_n$ are i.i.d. exponential random variables with parameter λ. So $f_{X_i}(x) = \lambda e^{-\lambda x}$ on $[0, \infty)$ for all $1 \leq i \leq n$.
- What is the law of $Z = \sum_{i=1}^n X_i$?
- We claimed in an earlier lecture that this was a gamma distribution with parameters (λ, n).
- So $f_Z(y) = \frac{\lambda e^{-\lambda y}(\lambda y)^{n-1}}{\Gamma(n)}$.
- We argued this point by taking limits of negative binomial distributions. Can we check it directly?
- By induction, would suffice to show that a gamma $(\lambda, 1)$ plus an independent gamma (λ, n) is a gamma $(\lambda, n+1)$.

Summing independent gamma random variables

- Say X is gamma (λ, s), Y is gamma (λ, t), and X and Y are independent.
- Intuitively, X is amount of time till we see s events, and Y is amount of subsequent time till we see t more events.
- So $f_X(x) = \frac{\lambda e^{-\lambda x}(\lambda x)^{s-1}}{\Gamma(s)}$ and $f_Y(y) = \frac{\lambda e^{-\lambda y}(\lambda y)^{t-1}}{\Gamma(t)}$.
- Now $f_{X+Y}(a) = \int_{-\infty}^{\infty} f_X(a-y)f_Y(y)dy$.
- Up to an a-independent multiplicative constant, this is

$$\int_0^a e^{-\lambda(a-y)}(a-y)^{s-1}e^{-\lambda y}y^{t-1}dy = e^{-\lambda a}\int_0^a (a-y)^{s-1}y^{t-1}dy.$$

- Letting $x = y/a$, this becomes $e^{-\lambda a}a^{s+t-1}\int_0^1 (1-x)^{s-1}x^{t-1}dx$.
- This is (up to multiplicative constant) $e^{-\lambda a}a^{s+t-1}$. Constant must be such that integral from $-\infty$ to ∞ is 1. Conclude that $X + Y$ is gamma $(\lambda, s+t)$.

Summing two normal variables

- X is normal with mean zero, variance σ_1^2, Y is normal with mean zero, variance σ_2^2.
- $f_X(x) = \frac{1}{\sqrt{2\pi}\sigma_1}e^{\frac{-x^2}{2\sigma_1^2}}$ and $f_Y(y) = \frac{1}{\sqrt{2\pi}\sigma_2}e^{\frac{-y^2}{2\sigma_2^2}}$.
- We just need to compute $f_{X+Y}(a) = \int_{-\infty}^{\infty} f_X(a-y)f_Y(y)dy$.
- We could compute this directly.
- Or we could argue with a multi-dimensional bell curve picture that if X and Y have variance 1 then $f_{\sigma_1 X + \sigma_2 Y}$ is the density of a normal random variable (and note that variances and expectations are additive).
- Or use fact that if $A_i \in \{-1, 1\}$ are i.i.d. coin tosses then $\frac{1}{\sqrt{N}}\sum_{i=1}^{\sigma^2 N} A_i$ is approximately normal with variance σ^2 when N is large.
- Generally: if independent random variables X_j are normal (μ_j, σ_j^2) then $\sum_{j=1}^n X_j$ is normal $(\sum_{j=1}^n \mu_j, \sum_{j=1}^n \sigma_j^2)$.

Other sums

- Sum of an independent binomial (m, p) and binomial (n, p)?
- Yes, binomial $(m + n, p)$. Can be seen from coin toss interpretation.
- Sum of independent Poisson λ_1 and Poisson λ_2?
- Yes, Poisson $\lambda_1 + \lambda_2$. Can be seen from Poisson point process interpretation.

18.600: Lecture 23

Conditional probability, order statistics, expectations of sums

Scott Sheffield

MIT

Outline

Conditional probability densities

Order statistics

Expectations of sums

Outline

Conditional probability densities

Order statistics

Expectations of sums

Conditional distributions

- Let's say X and Y have joint probability density function $f(x,y)$.
- We can *define* the conditional probability density of X given that $Y=y$ by $f_{X|Y=y}(x) = \frac{f(x,y)}{f_Y(y)}$.
- This amounts to restricting $f(x,y)$ to the line corresponding to the given y value (and dividing by the constant that makes the integral along that line equal to 1).
- This definition assumes that $f_Y(y) = \int_{-\infty}^{\infty} f(x,y)dx < \infty$ and $f_Y(y) \neq 0$. Is that safe to assume?
- Usually...

Remarks: conditioning on a probability zero event

- Our standard definition of conditional probability is $P(A|B) = P(AB)/P(B)$.
- Doesn't make sense if $P(B) = 0$. But previous slide defines "probability conditioned on $Y = y$" and $P\{Y = y\} = 0$.
- When can we (somehow) make sense of conditioning on probability zero event?
- Tough question in general.
- Consider conditional law of X given that $Y \in (y - \epsilon, y + \epsilon)$. If this has a limit as $\epsilon \to 0$, we can call *that* the law conditioned on $Y = y$.
- Precisely, define $F_{X|Y=y}(a) := \lim_{\epsilon \to 0} P\{X \leq a | Y \in (y - \epsilon, y + \epsilon)\}$.
- Then set $f_{X|Y=y}(a) = F'_{X|Y=y}(a)$. Consistent with definition from previous slide.

A word of caution

- Suppose X and Y are chosen uniformly on the semicircle $\{(x, y) : x^2 + y^2 \leq 1, x \geq 0\}$. What is $f_{X|Y=0}(x)$?
- Answer: $f_{X|Y=0}(x) = 1$ if $x \in [0, 1]$ (zero otherwise).
- Let (θ, R) be (X, Y) in polar coordinates. What is $f_{X|\theta=0}(x)$?
- Answer: $f_{X|\theta=0}(x) = 2x$ if $x \in [0, 1]$ (zero otherwise).
- Both $\{\theta = 0\}$ and $\{Y = 0\}$ describe the same probability zero event. But our interpretation of what it means to condition on this event is different in these two cases.
- Conditioning on (X, Y) belonging to a $\theta \in (-\epsilon, \epsilon)$ wedge is very different from conditioning on (X, Y) belonging to a $Y \in (-\epsilon, \epsilon)$ strip.

Outline

Outline

Maxima: pick five job candidates at random, choose best

- Suppose I choose n random variables $X_1, X_2, \dots, X_n$ uniformly at random on $[0,1]$, independently of each other.
- The n-tuple $(X_1, X_2, \dots, X_n)$ has a constant density function on the n-dimensional cube $[0,1]^n$.
- What is the probability that the *largest* of the X_i is less than a?
- ANSWER: a^n.
- So if $X = \max\{X_1, \dots, X_n\}$, then what is the probability density function of X?
- Answer: $F_X(a) = \begin{cases} 0 & a < 0 \\ a^n & a \in [0,1] \\ 1 & a > 1 \end{cases}$. And $f_x(a) = F_X'(a) = na^{n-1}$.

General order statistics

- Consider i.i.d random variables $X_1, X_2, \dots, X_n$ with continuous probability density f.
- Let $Y_1 < Y_2 < Y_3 \dots < Y_n$ be list obtained by *sorting* the X_j.
- In particular, $Y_1 = \min\{X_1, \dots, X_n\}$ and $Y_n = \max\{X_1, \dots, X_n\}$ is the maximum.
- What is the joint probability density of the Y_i?
- Answer: $f(x_1, x_2, \dots, x_n) = n! \prod_{i=1}^n f(x_i)$ if $x_1 < x_2 \dots < x_n$, zero otherwise.
- Let $\sigma : \{1, 2, \dots, n\} \to \{1, 2, \dots, n\}$ be the permutation such that $X_j = Y_{\sigma(j)}$
- Are σ and the vector $(Y_1, \dots, Y_n)$ independent of each other?
- Yes.

Example

- Let $X_1, \dots, X_n$ be i.i.d. uniform random variables on $[0,1]$.
- Example: say $n = 10$ and condition on X_1 being the third largest of the X_j.
- Given this, what is the conditional probability density function for X_1?
- Write $p = X_1$. This kind of like choosing a random p and then conditioning on 7 heads and 2 tails.
- Answer is beta distribution with parameters $(a,b) = (8,3)$.
- Up to a constant, $f(x) = x^7(1-x)^2$.
- General beta (a,b) expectation is $a/(a+b) = 8/11$. Mode is $\frac{(a-1)}{(a-1)+(b-1)} = 2/9$.

Outline

Outline

Conditional probability densities

Order statistics

Expectations of sums

Properties of expectation

- Several properties we derived for discrete expectations continue to hold in the continuum.
- If X is discrete with mass function $p(x)$ then $E[X] = \sum_x p(x)x$.
- Similarly, if X is continuous with density function $f(x)$ then $E[X] = \int f(x)xdx$.
- If X is discrete with mass function $p(x)$ then $E[g(x)] = \sum_x p(x)g(x)$.
- Similarly, X if is continuous with density function $f(x)$ then $E[g(X)] = \int f(x)g(x)dx$.
- If X and Y have joint mass function $p(x,y)$ then $E[g(X,Y)] = \sum_y \sum_x g(x,y)p(x,y)$.
- If X and Y have joint probability density function $f(x,y)$ then $E[g(X,Y)] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x,y)f(x,y)dxdy$.

Properties of expectation

- For both discrete and continuous random variables X and Y we have $E[X+Y] = E[X]+E[Y]$.
- In both discrete and continuous settings, $E[aX] = aE[X]$ when a is a constant. And $E[\sum a_i X_i] = \sum a_i E[X_i]$.
- But what about that delightful "area under $1-F_X$" formula for the expectation?
- When X is non-negative with probability one, do we always have $E[X] = \int_0^\infty P\{X > x\}$, in both discrete and continuous settings?
- Define $g(y)$ so that $1 - F_X(g(y)) = y$. (Draw horizontal line at height y and look where it hits graph of $1-F_X$.)
- Choose Y uniformly on $[0,1]$ and note that $g(Y)$ has the same probability distribution as X.
- So $E[X] = E[g(Y)] = \int_0^1 g(y)dy$, which is indeed the area under the graph of $1-F_X$.

18.600: Lecture 24

Covariance and some conditional expectation exercises

Scott Sheffield

MIT

Outline

Covariance and correlation

Paradoxes: getting ready to think about conditional expectation

Outline

Covariance and correlation

Paradoxes: getting ready to think about conditional expectation

A property of independence

- If X and Y are independent then $E[g(X)h(Y)] = E[g(X)]E[h(Y)]$.
- Just write $E[g(X)h(Y)] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x)h(y)f(x,y)dxdy$.
- Since $f(x,y) = f_X(x)f_Y(y)$ this factors as $\int_{-\infty}^{\infty} h(y)f_Y(y)dy \int_{-\infty}^{\infty} g(x)f_X(x)dx = E[h(Y)]E[g(X)]$.

Defining covariance and correlation

- Now define covariance of X and Y by $\mathrm{Cov}(X,Y) = E[(X - E[X])(Y - E[Y])]$.
- Note: by definition $\mathrm{Var}(X) = \mathrm{Cov}(X,X)$.
- Covariance (like variance) can also written a different way. Write $\mu_x = E[X]$ and $\mu_Y = E[Y]$. If laws of X and Y are known, then μ_X and μ_Y are just constants.
- Then

$$\mathrm{Cov}(X,Y) = E[(X-\mu_X)(Y-\mu_Y)] = E[XY-\mu_X Y-\mu_Y X+\mu_X\mu_Y] =$$

$$E[XY] - \mu_X E[Y] - \mu_Y E[X] + \mu_X\mu_Y = E[XY] - E[X]E[Y].$$

- Covariance formula $E[XY] - E[X]E[Y]$, or "expectation of product minus product of expectations" is frequently useful.
- Note: if X and Y are independent then $\mathrm{Cov}(X,Y) = 0$.

Basic covariance facts

- Using $\mathrm{Cov}(X,Y) = E[XY] - E[X]E[Y]$ as a definition, certain facts are immediate.
- $\mathrm{Cov}(X,Y) = \mathrm{Cov}(Y,X)$
- $\mathrm{Cov}(X,X) = \mathrm{Var}(X)$
- $\mathrm{Cov}(aX,Y) = a\mathrm{Cov}(X,Y)$.
- $\mathrm{Cov}(X_1 + X_2, Y) = \mathrm{Cov}(X_1,Y) + \mathrm{Cov}(X_2,Y)$.
- **General statement of bilinearity of covariance:**

$$\mathrm{Cov}(\sum_{i=1}^m a_i X_i, \sum_{j=1}^n b_j Y_j) = \sum_{i=1}^m \sum_{j=1}^n a_i b_j \mathrm{Cov}(X_i, Y_j).$$

- Special case:

$$\mathrm{Var}(\sum_{i=1}^n X_i) = \sum_{i=1}^n \mathrm{Var}(X_i) + 2 \sum_{(i,j):i<j} \mathrm{Cov}(X_i, X_j).$$

Defining correlation

- Again, by definition $\mathrm{Cov}(X,Y) = E[XY] - E[X]E[Y]$.
- **Correlation** of X and Y defined by

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

- Correlation doesn't care what units you use for X and Y. If $a > 0$ and $c > 0$ then $\rho(aX + b, cY + d) = \rho(X,Y)$.
- Satisfies $-1 \leq \rho(X,Y) \leq 1$.
- Why is that? Something to do with $E[(X+Y)^2] \geq 0$ and $E[(X-Y)^2] \geq 0$?
- If a and b are constants and $a > 0$ then $\rho(aX + b, X) = 1$.
- If a and b are constants and $a < 0$ then $\rho(aX + b, X) = -1$.

Important point

- Say X and Y are uncorrelated when $\rho(X,Y) = 0$.
- Are independent random variables X and Y always uncorrelated?
- Yes, assuming variances are finite (so that correlation is defined).
- Are uncorrelated random variables always independent?
- No. Uncorrelated just means $E[(X - E[X])(Y - E[Y])] = 0$, i.e., the outcomes where $(X - E[X])(Y - E[Y])$ is positive (the upper right and lower left quadrants, if axes are drawn centered at $(E[X], E[Y])$) balance out the outcomes where this quantity is negative (upper left and lower right quadrants). This is a much weaker statement than independence.

Examples

- Suppose that $X_1, \ldots, X_n$ are i.i.d. random variables with variance 1. For example, maybe each X_j takes values ± 1 according to a fair coin toss.
- Compute $\mathrm{Cov}(X_1 + X_2 + X_3, X_2 + X_3 + X_4)$.
- Compute the correlation coefficient $\rho(X_1 + X_2 + X_3, X_2 + X_3 + X_4)$.
- Can we generalize this example?
- What is variance of number of people who get their own hat in the hat problem?
- Define X_i to be 1 if ith person gets own hat, zero otherwise.
- Recall formula $\mathrm{Var}(\sum_{i=1}^n X_i) = \sum_{i=1}^n \mathrm{Var}(X_i) + 2\sum_{(i,j):i<j} \mathrm{Cov}(X_i, X_j)$.
- Reduces problem to computing $\mathrm{Cov}(X_i, X_j)$ (for $i \neq j$) and $\mathrm{Var}(X_i)$.

Outline

Covariance and correlation

Paradoxes: getting ready to think about conditional expectation

Outline

Covariance and correlation

Paradoxes: getting ready to think about conditional expectation

Famous paradox

- Certain corrupt and amoral banker dies, instructed to spend some number n (of banker's choosing) days in hell.
- At the end of this period, a (biased) coin will be tossed. Banker will be assigned to hell forever with probability $1/n$ and heaven forever with probability $1 - 1/n$.
- After 10 days, banker reasons, "If I wait another day I reduce my odds of being here forever from $1/10$ to $1/11$. That's a reduction of $1/110$. A $1/110$ chance at infinity has infinite value. Worth waiting one more day."
- Repeats this reasoning every day, stays in hell forever.
- Standard punch line: this is actually what banker deserved.
- Fairly dark as math humor goes (and no offense intended to anyone...) but dilemma is interesting.

- **Paradox:** decisions seem sound individually but together yield worst possible outcome. Why? Can we demystify this?
- **Variant without probability:** Instead of tossing $(1/n)$-coin, person deterministically spends $1/n$ fraction of future days (every nth day, say) in hell.
- **Even simpler variant:** infinitely many identical money sacks have labels $1, 2, 3, \dots$ I have sack 1. You have all others.
- You offer me a deal. I give you sack 1, you give me sacks 2 and 3. I give you sack 2 and you give me sacks 4 and 5. On the nth stage, I give you sack n and you give me sacks $2n$ and $2n+1$. Continue until I say stop.
- Lets me get arbitrarily rich. But if I go on forever, I return every sack given to me. If nth sack confers right to spend nth day in heaven, leads to hell-forever paradox.
- I make infinitely many good trades and end up with less than I started with. "Paradox" is really just existence of 2-to-1 map from (smaller set) $\{2, 3, \dots\}$ to (bigger set) $\{1, 2, \dots\}$.

Money pile paradox

- You have an infinite collection of money piles with labeled $0, 1, 2, \dots$ from left to right.
- Precise details not important, but let's say you have $1/4$ in the 0th pile and $\frac{3}{8}5^j$ in the jth pile for each $j > 0$. Important thing is that pile size is increasing exponentially in j.
- Banker proposes to transfer a fraction (say $2/3$) of each pile to the pile on its left and remainder to the pile on its right. Do this simultaneously for all piles.
- Every pile is bigger after transfer (and this can be true even if banker takes a portion of each pile as a fee).
- Banker seemed to make you richer (every pile got bigger) but really just reshuffled your infinite wealth.

Two envelope paradox

- X is geometric with parameter $1/2$. One envelope has 10^X dollars, one has 10^{X-1} dollars. Envelopes shuffled.
- You choose an envelope and, after seeing contents, are allowed to choose whether to keep it or switch. (Maybe you have to pay a dollar to switch.)
- Maximizing conditional expectation, it seems it's always better to switch. But if you always switch, why not just choose second-choice envelope first and avoid switching fee?
- Kind of a disguised version of money pile paradox. But more subtle. One has to replace "jth pile of money" with "restriction of expectation sum to scenario that first chosen envelop has 10^j". Switching indeed makes each pile bigger.
- However, "Higher expectation given amount in first envelope" may not be right notion of "better." If S is payout with switching, T is payout without switching, then S has same law as $T-1$. In that sense S is worse.

Moral

- Beware infinite expectations.
- Beware unbounded utility functions.
- They can lead to strange conclusions, sometimes related to "reshuffling infinite (actual or expected) wealth to create more" paradoxes.
- Paradoxes can arise even when total transaction is finite with probability one (as in envelope problem).

18.600: Lecture 25

Conditional expectation

Scott Sheffield

MIT

Outline

Outline

Recall: conditional probability distributions

- It all starts with the definition of conditional probability: $P(A|B) = P(AB)/P(B)$.
- If X and Y are jointly discrete random variables, we can use this to define a probability mass function for X *given* $Y = y$.
- That is, we write $p_{X|Y}(x|y) = P\{X = x|Y = y\} = \frac{p(x,y)}{p_Y(y)}$.
- In words: first restrict sample space to pairs (x, y) with given y value. Then divide the original mass function by $p_Y(y)$ to obtain a probability mass function on the restricted space.
- We do something similar when X and Y are continuous random variables. In that case we write $f_{X|Y}(x|y) = \frac{f(x,y)}{f_Y(y)}$.
- Often useful to think of sampling (X, Y) as a two-stage process. First sample Y from its marginal distribution, obtain $Y = y$ for some particular y. Then sample X from its probability distribution *given* $Y = y$.
- Marginal law of X is weighted average of conditional laws.

Example

- Let X be value on one die roll, Y value on second die roll, and write $Z = X + Y$.
- What is the probability distribution for X given that $Y = 5$?
- Answer: uniform on $\{1, 2, 3, 4, 5, 6\}$.
- What is the probability distribution for Z given that $Y = 5$?
- Answer: uniform on $\{6, 7, 8, 9, 10, 11\}$.
- What is the probability distribution for Y given that $Z = 5$?
- Answer: uniform on $\{1, 2, 3, 4\}$.

Outline

Outline

Conditional probability distributions

Conditional expectation

Interpretation and examples

Conditional expectation

- Now, what do we mean by $E[X|Y = y]$? This should just be the expectation of X in the conditional probability measure for X given that $Y = y$.
- Can write this as $E[X|Y = y] = \sum_x xP\{X = x|Y = y\} = \sum_x xp_{X|Y}(x|y)$.
- Can make sense of this in the continuum setting as well.
- In continuum setting we had $f_{X|Y}(x|y) = \frac{f(x,y)}{f_Y(y)}$. So $E[X|Y = y] = \int_{-\infty}^{\infty} x \frac{f(x,y)}{f_Y(y)} dx$

Example

- Let X be value on one die roll, Y value on second die roll, and write $Z = X + Y$.
- What is $E[X|Y = 5]$?
- What is $E[Z|Y = 5]$?
- What is $E[Y|Z = 5]$?

Conditional expectation as a random variable

- Can think of $E[X|Y]$ as a function of the random variable Y. When $Y = y$ it takes the value $E[X|Y = y]$.
- So $E[X|Y]$ is itself a random variable. It happens to depend only on the value of Y.
- Thinking of $E[X|Y]$ as a random variable, we can ask what *its* expectation is. What is $E[E[X|Y]]$?
- **Very useful fact:** $E[E[X|Y]] = E[X]$.
- In words: what you expect to expect X to be *after learning* Y is same as what you *now* expect X to be.
- Proof in discrete case: $E[X|Y = y] = \sum_x xP\{X = x|Y = y\} = \sum_x x\frac{p(x,y)}{p_Y(y)}$.
- Recall that, in general, $E[g(Y)] = \sum_y p_Y(y)g(y)$.
- $E[E[X|Y = y]] = \sum_y p_Y(y) \sum_x x\frac{p(x,y)}{p_Y(y)} = \sum_x \sum_y p(x,y)x = E[X]$.

Conditional variance

- Definition: $\mathrm{Var}(X|Y) = E\left[(X - E[X|Y])^2|Y\right] = E\left[X^2 - E[X|Y]^2|Y\right]$.
- $\mathrm{Var}(X|Y)$ is a random variable that depends on Y. It is the variance of X in the conditional distribution for X given Y.
- Note $E[\mathrm{Var}(X|Y)] = E[E[X^2|Y]] - E[E[X|Y]^2|Y] = E[X^2] - E[E[X|Y]^2]$.
- If we subtract $E[X]^2$ from first term and add equivalent value $E[E[X|Y]]^2$ to the second, RHS becomes $\mathrm{Var}[X] - \mathrm{Var}[E[X|Y]]$, which implies following:
- **Useful fact:** $\mathrm{Var}(X) = \mathrm{Var}(E[X|Y]) + E[\mathrm{Var}(X|Y)]$.
- One can discover X in two stages: first sample Y from marginal and compute $E[X|Y]$, then sample X from distribution given Y value.
- Above fact breaks variance into two parts, corresponding to these two stages.

Example

- Let X be a random variable of variance σ_X^2 and Y an independent random variable of variance σ_Y^2 and write $Z = X + Y$. Assume $E[X] = E[Y] = 0$.
- What are the covariances $\mathrm{Cov}(X, Y)$ and $\mathrm{Cov}(X, Z)$?
- How about the correlation coefficients $\rho(X, Y)$ and $\rho(X, Z)$?
- What is $E[Z|X]$? And how about $\mathrm{Var}(Z|X)$?
- Both of these values are functions of X. Former is just X. Latter happens to be a constant-valued function of X, i.e., happens not to actually depend on X. We have $\mathrm{Var}(Z|X) = \sigma_Y^2$.
- Can we check the formula $\mathrm{Var}(Z) = \mathrm{Var}(E[Z|X]) + E[\mathrm{Var}(Z|X)]$ in this case?

Outline

Conditional probability distributions

Conditional expectation

Interpretation and examples

Outline

Conditional probability distributions

Conditional expectation

Interpretation and examples

Interpretation

- Sometimes think of the expectation $E[Y]$ as a "best guess" or "best predictor" of the value of Y.
- It is best in the sense that at among all constants m, the expectation $E[(Y-m)^2]$ is minimized when $m = E[Y]$.
- But what if we allow non-constant predictors? What if the predictor is allowed to depend on the value of a random variable X that we can observe directly?
- Let $g(x)$ be such a function. Then $E[(y-g(X))^2]$ is minimized when $g(X) = E[Y|X]$.

Examples

- Toss 100 coins. What's the conditional expectation of the number of heads given that there are k heads among the first fifty tosses?
- $k + 25$
- What's the conditional expectation of the number of aces in a five-card poker hand given that the first two cards in the hand are aces?
- $2 + 3 \cdot 2/50$

18.600: Lecture 26

Moment generating functions and characteristic functions

Scott Sheffield

MIT

Outline

Outline

Moment generating functions

- Let X be a random variable.
- The **moment generating function** of X is defined by $M(t) = M_X(t) := E[e^{tX}]$.
- When X is discrete, can write $M(t) = \sum_x e^{tx} p_X(x)$. So $M(t)$ is a weighted average of countably many exponential functions.
- When X is continuous, can write $M(t) = \int_{-\infty}^{\infty} e^{tx} f(x) dx$. So $M(t)$ is a weighted average of a continuum of exponential functions.
- We always have $M(0) = 1$.
- If $b > 0$ and $t > 0$ then $E[e^{tX}] \geq E[e^{t \min\{X,b\}}] \geq P\{X \geq b\} e^{tb}$.
- If X takes both positive and negative values with positive probability then $M(t)$ grows at least exponentially fast in $|t|$ as $|t| \to \infty$.

Moment generating functions actually generate moments

- Let X be a random variable and $M(t) = E[e^{tX}]$.
- Then $M'(t) = \frac{d}{dt}E[e^{tX}] = E\left[\frac{d}{dt}(e^{tX})\right] = E[Xe^{tX}]$.
- in particular, $M'(0) = E[X]$.
- Also $M''(t) = \frac{d}{dt}M'(t) = \frac{d}{dt}E[Xe^{tX}] = E[X^2e^{tX}]$.
- So $M''(0) = E[X^2]$. Same argument gives that nth derivative of M at zero is $E[X^n]$.
- Interesting: knowing all of the derivatives of M at a single point tells you the moments $E[X^k]$ for all integer $k \geq 0$.
- Another way to think of this: write $e^{tX} = 1 + tX + \frac{t^2X^2}{2!} + \frac{t^3X^3}{3!} + \dots$.
- Taking expectations gives $E[e^{tX}] = 1 + tm_1 + \frac{t^2m_2}{2!} + \frac{t^3m_3}{3!} + \dots$, where m_k is the kth moment. The kth derivative at zero is m_k.

Moment generating functions for independent sums

- Let X and Y be independent random variables and $Z = X + Y$.
- Write the moment generating functions as $M_X(t) = E[e^{tX}]$ and $M_Y(t) = E[e^{tY}]$ and $M_Z(t) = E[e^{tZ}]$.
- If you knew M_X and M_Y, could you compute M_Z?
- By independence, $M_Z(t) = E[e^{t(X+Y)}] = E[e^{tX}e^{tY}] = E[e^{tX}]E[e^{tY}] = M_X(t)M_Y(t)$ for all t.
- In other words, adding independent random variables corresponds to multiplying moment generating functions.

Moment generating functions for sums of i.i.d. random variables

- We showed that if $Z = X + Y$ and X and Y are independent, then $M_Z(t) = M_X(t)M_Y(t)$
- If $X_1 \dots X_n$ are i.i.d. copies of X and $Z = X_1 + \dots + X_n$ then what is M_Z?
- Answer: M_X^n. Follows by repeatedly applying formula above.
- This a big reason for studying moment generating functions. It helps us understand what happens when we sum up a lot of independent copies of the same random variable.

Other observations

- If $Z = aX$ then can I use M_X to determine M_Z?
- Answer: Yes. $M_Z(t) = E[e^{tZ}] = E[e^{taX}] = M_X(at)$.
- If $Z = X + b$ then can I use M_X to determine M_Z?
- Answer: Yes. $M_Z(t) = E[e^{tZ}] = E[e^{tX+bt}] = e^{bt}M_X(t)$.
- Latter answer is the special case of $M_Z(t) = M_X(t)M_Y(t)$ where Y is the constant random variable b.

Examples

- Let's try some examples. What is $M_X(t) = E[e^{tX}]$ when X is binomial with parameters (p, n)? Hint: try the $n = 1$ case first.
- Answer: if $n = 1$ then $M_X(t) = E[e^{tX}] = pe^t + (1-p)e^0$. In general $M_X(t) = (pe^t + 1 - p)^n$.
- What if X is Poisson with parameter $\lambda > 0$?
- Answer: $M_X(t) = E[e^{tx}] = \sum_{n=0}^{\infty} \frac{e^{tn}e^{-\lambda}\lambda^n}{n!} = e^{-\lambda}\sum_{n=0}^{\infty} \frac{(\lambda e^t)^n}{n!} = e^{-\lambda}e^{\lambda e^t} = \exp[\lambda(e^t - 1)]$.
- We know that if you add independent Poisson random variables with parameters λ_1 and λ_2 you get a Poisson random variable of parameter $\lambda_1 + \lambda_2$. How is this fact manifested in the moment generating function?

More examples: normal random variables

- What if X is normal with mean zero, variance one?
- $M_X(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{tx} e^{-x^2/2} dx = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\{-\frac{(x-t)^2}{2} + \frac{t^2}{2}\} dx = e^{t^2/2}$.
- What does that tell us about sums of i.i.d. copies of X?
- If Z is sum of n i.i.d. copies of X then $M_Z(t) = e^{nt^2/2}$.
- What is M_Z if Z is normal with mean μ and variance σ^2?
- Answer: Z has same law as $\sigma X + \mu$, so $M_Z(t) = M(\sigma t)e^{\mu t} = \exp\{\frac{\sigma^2 t^2}{2} + \mu t\}$.

More examples: exponential random variables

- What if X is exponential with parameter $\lambda > 0$?
- $M_X(t) = \int_0^{\infty} e^{tx} \lambda e^{-\lambda x} dx = \lambda \int_0^{\infty} e^{-(\lambda - t)x} dx = \frac{\lambda}{\lambda - t}$.
- What if Z is a Γ distribution with parameters $\lambda > 0$ and $n > 0$?
- Then Z has the law of a sum of n independent copies of X. So $M_Z(t) = M_X(t)^n = \left(\frac{\lambda}{\lambda - t}\right)^n$.
- Exponential calculation above works for $t < \lambda$. What happens when $t > \lambda$? Or as t approaches λ from below?
- $M_X(t) = \int_0^{\infty} e^{tx} \lambda e^{-\lambda x} dx = \lambda \int_0^{\infty} e^{-(\lambda - t)x} dx = \infty$ if $t \geq \lambda$.

More examples: existence issues

- Seems that unless $f_X(x)$ decays superexponentially as x tends to infinity, we won't have $M_X(t)$ defined for all t.
- What is M_X if X is standard Cauchy, so that $f_X(x) = \frac{1}{\pi(1+x^2)}$.
- Answer: $M_X(0) = 1$ (as is true for any X) but otherwise $M_X(t)$ is infinite for all $t \neq 0$.
- Informal statement: moment generating functions are not defined for distributions with fat tails.

Outline

Moment generating functions

Characteristic functions

Continuity theorems and perspective

Outline

Moment generating functions

Characteristic functions

Continuity theorems and perspective

Characteristic functions

- Let X be a random variable.
- The **characteristic function** of X is defined by $\phi(t) = \phi_X(t) := E[e^{itX}]$. Like $M(t)$ except with i thrown in.
- Recall that by definition $e^{it} = \cos(t) + i\sin(t)$.
- Characteristic functions are similar to moment generating functions in some ways.
- For example, $\phi_{X+Y} = \phi_X \phi_Y$, just as $M_{X+Y} = M_X M_Y$.
- And $\phi_{aX}(t) = \phi_X(at)$ just as $M_{aX}(t) = M_X(at)$.
- And if X has an mth moment then $E[X^m] = i^m \phi_X^{(m)}(0)$.
- But characteristic functions have a distinct advantage: they are always well defined for all t even if f_X decays slowly.

Outline

Moment generating functions

Characteristic functions

Continuity theorems and perspective

Outline

Moment generating functions

Characteristic functions

Continuity theorems and perspective

Continuity theorems

- Let X be a random variable and X_n a sequence of random variables.
- We say that X_n **converge in distribution** or **converge in law** to X if $\lim_{n\to\infty} F_{X_n}(x) = F_X(x)$ at all $x \in \mathbb{R}$ at which F_X is continuous.
- **Lévy's continuity theorem (see Wikipedia):** if $\lim_{n\to\infty} \phi_{X_n}(t) = \phi_X(t)$ for all t, then X_n converge in law to X.
- **Moment generating analog:** if moment generating functions $M_{X_n}(t)$ are defined for all t and n and $\lim_{n\to\infty} M_{X_n}(t) = M_X(t)$ for all t, then X_n converge in law to X.

Perspective

- In later lectures, we will see that one can use moment generating functions and/or characteristic functions to prove the so-called *weak law of large numbers* and *central limit theorem*.
- Proofs using characteristic functions apply in more generality, but they require you to remember how to exponentiate imaginary numbers.
- Moment generating functions are central to so-called *large deviation theory* and play a fundamental role in statistical physics, among other things.
- Characteristic functions are *Fourier transforms* of the corresponding distribution density functions and encode "periodicity" patterns. For example, if X is integer valued, $\phi_X(t) = E[e^{itX}]$ will be 1 whenever t is a multiple of 2π.

18.600: Lecture 27

Lectures 15-27 Review

Scott Sheffield

MIT

Outline

Continuous random variables

Problems motivated by coin tossing

Random variable properties

Outline

Continuous random variables

Problems motivated by coin tossing

Random variable properties

Continuous random variables

- Say X is a **continuous random variable** if there exists a **probability density function** $f = f_X$ on $\mathbb{R}$ such that $P\{X \in B\} = \int_B f(x)dx := \int 1_B(x)f(x)dx$.
- We may assume $\int_{\mathbb{R}} f(x)dx = \int_{-\infty}^{\infty} f(x)dx = 1$ and f is non-negative.
- Probability of interval $[a, b]$ is given by $\int_a^b f(x)dx$, the area under f between a and b.
- Probability of any single point is zero.
- Define **cumulative distribution function** $F(a) = F_X(a) := P\{X < a\} = P\{X \le a\} = \int_{-\infty}^{a} f(x)dx$.

Expectations of continuous random variables

- Recall that when X was a discrete random variable, with $p(x) = P\{X = x\}$, we wrote
$$E[X] = \sum_{x:p(x)>0} p(x)x.$$
- How should we define $E[X]$ when X is a continuous random variable?
- Answer: $E[X] = \int_{-\infty}^{\infty} f(x)xdx$.
- Recall that when X was a discrete random variable, with $p(x) = P\{X = x\}$, we wrote
$$E[g(X)] = \sum_{x:p(x)>0} p(x)g(x).$$
- What is the analog when X is a continuous random variable?
- Answer: we will write $E[g(X)] = \int_{-\infty}^{\infty} f(x)g(x)dx$.

Variance of continuous random variables

- Suppose X is a continuous random variable with mean μ.
- We can write $\mathrm{Var}[X] = E[(X-\mu)^2]$, same as in the discrete case.
- Next, if $g = g_1 + g_2$ then $E[g(X)] = \int g_1(x)f(x)dx + \int g_2(x)f(x)dx = \int \big(g_1(x) + g_2(x)\big)f(x)dx = E[g_1(X)] + E[g_2(X)]$.
- Furthermore, $E[ag(X)] = aE[g(X)]$ when a is a constant.
- Just as in the discrete case, we can expand the variance expression as $\mathrm{Var}[X] = E[X^2 - 2\mu X + \mu^2]$ and use additivity of expectation to say that $\mathrm{Var}[X] = E[X^2] - 2\mu E[X] + E[\mu^2] = E[X^2] - 2\mu^2 + \mu^2 = E[X^2] - E[X]^2$.
- This formula is often useful for calculations.

Outline

Continuous random variables

Problems motivated by coin tossing

Random variable properties

Outline

Continuous random variables

Problems motivated by coin tossing

Random variable properties

It's the coins, stupid

- Much of what we have done in this course can be motivated by the i.i.d. sequence X_i where each X_i is 1 with probability p and 0 otherwise. Write $S_n = \sum_{i=1}^n X_n$.
- **Binomial** (S_n — number of heads in n tosses), **geometric** (steps required to obtain one heads), **negative binomial** (steps required to obtain n heads).
- **Standard normal** approximates law of $\frac{S_n - E[S_n]}{\mathrm{SD}(S_n)}$. Here $E[S_n] = np$ and $\mathrm{SD}(S_n) = \sqrt{\mathrm{Var}(S_n)} = \sqrt{npq}$ where $q = 1 - p$.
- **Poisson** is limit of binomial as $n \to \infty$ when $p = \lambda/n$.
- **Poisson point process**: toss one λ/n coin during each length $1/n$ time increment, take $n \to \infty$ limit.
- **Exponential**: time till first event in λ Poisson point process.
- **Gamma distribution**: time till nth event in λ Poisson point process.

Discrete random variable properties derivable from coin toss intuition

- **Sum of two independent binomial random variables** with parameters (n_1, p) and (n_2, p) is itself binomial $(n_1 + n_2, p)$.
- **Sum of n independent geometric random variables** with parameter p is negative binomial with parameter (n, p).
- **Expectation of geometric random variable** with parameter p is $1/p$.
- **Expectation of binomial random variable** with parameters (n, p) is np.
- **Variance of binomial random variable** with parameters (n, p) is $np(1-p) = npq$.

Continuous random variable properties derivable from coin toss intuition

- **Sum of n independent exponential random variables** each with parameter λ is gamma with parameters (n, λ).
- **Memoryless properties:** given that exponential random variable X is greater than $T > 0$, the conditional law of $X - T$ is the same as the original law of X.
- Write $p = \lambda/n$. **Poisson random variable expectation** is $\lim_{n\to\infty} np = \lim_{n\to\infty} n\frac{\lambda}{n} = \lambda$. **Variance** is $\lim_{n\to\infty} np(1-p) = \lim_{n\to\infty} n(1-\lambda/n)\lambda/n = \lambda$.
- **Sum of λ_1 Poisson and independent λ_2 Poisson** is a $\lambda_1 + \lambda_2$ Poisson.
- **Times between successive events** in λ Poisson process are independent exponentials with parameter λ.
- **Minimum of independent exponentials** with parameters λ_1 and λ_2 is itself exponential with parameter $\lambda_1 + \lambda_2$.

DeMoivre-Laplace Limit Theorem

- **DeMoivre-Laplace limit theorem (special case of central limit theorem):**

$$\lim_{n\to\infty} P\{a \le \frac{S_n - np}{\sqrt{npq}} \le b\} \to \Phi(b) - \Phi(a).$$

- This is $\Phi(b) - \Phi(a) = P\{a \le X \le b\}$ when X is a standard normal random variable.

Problems

- Toss a million fair coins. Approximate the probability that I get more than $501,000$ heads.
- Answer: well, $\sqrt{npq} = \sqrt{10^6 \times .5 \times .5} = 500$. So we're asking for probability to be over two SDs above mean. This is approximately $1 - \Phi(2) = \Phi(-2)$.
- Roll 60000 dice. Expect to see 10000 sixes. What's the probability to see more than 9800?
- Here $\sqrt{npq} = \sqrt{60000 \times \frac{1}{6} \times \frac{5}{6}} \approx 91.28$.
- And $200/91.28 \approx 2.19$. Answer is about $1 - \Phi(-2.19)$.

Properties of normal random variables

- Say X is a (standard) **normal random variable** if $f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$.
- Mean zero and variance one.
- The random variable $Y = \sigma X + \mu$ has variance σ^2 and expectation μ.
- Y is said to be normal with parameters μ and σ^2. Its density function is $f_Y(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2}$.
- Function $\Phi(a) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{a} e^{-x^2/2} dx$ can't be computed explicitly.
- Values: $\Phi(-3) \approx .0013$, $\Phi(-2) \approx .023$ and $\Phi(-1) \approx .159$.
- Rule of thumb: "two thirds of time within one SD of mean, 95 percent of time within 2 SDs of mean."

Properties of exponential random variables

- Say X is an **exponential random variable of parameter** λ when its probability distribution function is $f(x) = \lambda e^{-\lambda x}$ for $x \geq 0$ (and $f(x) = 0$ if $x < 0$).
- For $a > 0$ have
$$F_X(a) = \int_0^a f(x)dx = \int_0^a \lambda e^{-\lambda x} dx = -e^{-\lambda x}\Big|_0^a = 1 - e^{-\lambda a}.$$
- Thus $P\{X < a\} = 1 - e^{-\lambda a}$ and $P\{X > a\} = e^{-\lambda a}$.
- Formula $P\{X > a\} = e^{-\lambda a}$ is very important in practice.
- Repeated integration by parts gives $E[X^n] = n!/\lambda^n$.
- If $\lambda = 1$, then $E[X^n] = n!$. Value $\Gamma(n) := E[X^{n-1}]$ defined for real $n > 0$ and $\Gamma(n) = (n-1)!$.

Defining Γ distribution

- Say that random variable X has gamma distribution with parameters (α, λ) if $f_X(x) = \begin{cases} \frac{(\lambda x)^{\alpha-1} e^{-\lambda x} \lambda}{\Gamma(\alpha)} & x \geq 0 \\ 0 & x < 0 \end{cases}$.
- Same as exponential distribution when $\alpha = 1$. Otherwise, multiply by $x^{\alpha-1}$ and divide by $\Gamma(\alpha)$. The fact that $\Gamma(\alpha)$ is what you need to divide by to make the total integral one just follows from the definition of Γ.
- Waiting time interpretation makes sense only for integer α, but distribution is defined for general positive α.

Outline

Continuous random variables

Problems motivated by coin tossing

Random variable properties

Outline

Continuous random variables

Problems motivated by coin tossing

Random variable properties

Properties of uniform random variables

- Suppose X is a random variable with probability density function $f(x) = \begin{cases} \frac{1}{\beta-\alpha} & x \in [\alpha, \beta] \\ 0 & x \notin [\alpha, \beta]. \end{cases}$
- Then $E[X] = \frac{\alpha+\beta}{2}$.
- And $\mathrm{Var}[X] = \mathrm{Var}[(\beta-\alpha)Y + \alpha] = \mathrm{Var}[(\beta-\alpha)Y] = (\beta-\alpha)^2\mathrm{Var}[Y] = (\beta-\alpha)^2/12$.

Distribution of function of random variable

- Suppose $P\{X \le a\} = F_X(a)$ is known for all a. Write $Y = X^3$. What is $P\{Y \le 27\}$?
- Answer: note that $Y \le 27$ if and only if $X \le 3$. Hence $P\{Y \le 27\} = P\{X \le 3\} = F_X(3)$.
- Generally $F_Y(a) = P\{Y \le a\} = P\{X \le a^{1/3}\} = F_X(a^{1/3})$
- This is a general principle. If X is a continuous random variable and g is a strictly increasing function of x and $Y = g(X)$, then $F_Y(a) = F_X(g^{-1}(a))$.

Joint probability mass functions: discrete random variables

- If X and Y assume values in $\{1, 2, \dots, n\}$ then we can view $A_{i,j} = P\{X = i, Y = j\}$ as the entries of an $n \times n$ matrix.
- Let's say I don't care about Y. I just want to know $P\{X = i\}$. How do I figure that out from the matrix?
- Answer: $P\{X = i\} = \sum_{j=1}^{n} A_{i,j}$.
- Similarly, $P\{Y = j\} = \sum_{i=1}^{n} A_{i,j}$.
- In other words, the probability mass functions for X and Y are the row and columns sums of $A_{i,j}$.
- Given the joint distribution of X and Y, we sometimes call distribution of X (ignoring Y) and distribution of Y (ignoring X) the **marginal** distributions.
- In general, when X and Y are jointly defined discrete random variables, we write $p(x, y) = p_{X,Y}(x, y) = P\{X = x, Y = y\}$.

Joint distribution functions: continuous random variables

- Given random variables X and Y, define $F(a, b) = P\{X \le a, Y \le b\}$.
- The region $\{(x, y) : x \le a, y \le b\}$ is the lower left "quadrant" centered at (a, b).
- Refer to $F_X(a) = P\{X \le a\}$ and $F_Y(b) = P\{Y \le b\}$ as **marginal** cumulative distribution functions.
- Question: if I tell you the two parameter function F, can you use it to determine the marginals F_X and F_Y?
- Answer: Yes. $F_X(a) = \lim_{b \to \infty} F(a, b)$ and $F_Y(b) = \lim_{a \to \infty} F(a, b)$.
- Density: $f(x, y) = \frac{\partial}{\partial x} \frac{\partial}{\partial y} F(x, y)$.

Independent random variables

- We say X and Y are independent if for any two (measurable) sets A and B of real numbers we have

$$P\{X \in A, Y \in B\} = P\{X \in A\}P\{Y \in B\}.$$

- When X and Y are discrete random variables, they are independent if $P\{X = x, Y = y\} = P\{X = x\}P\{Y = y\}$ for all x and y for which $P\{X = x\}$ and $P\{Y = y\}$ are non-zero.
- When X and Y are continuous, they are independent if $f(x, y) = f_X(x)f_Y(y)$.

Summing two random variables

- Say we have independent random variables X and Y and we know their density functions f_X and f_Y.
- Now let's try to find $F_{X+Y}(a) = P\{X + Y \le a\}$.
- This is the integral over $\{(x, y) : x + y \le a\}$ of $f(x, y) = f_X(x)f_Y(y)$. Thus,
- $$P\{X + Y \le a\} = \int_{-\infty}^{\infty} \int_{-\infty}^{a-y} f_X(x)f_Y(y)dxdy$$
  $$= \int_{-\infty}^{\infty} F_X(a - y)f_Y(y)dy.$$
- Differentiating both sides gives $f_{X+Y}(a) = \frac{d}{da} \int_{-\infty}^{\infty} F_X(a-y)f_Y(y)dy = \int_{-\infty}^{\infty} f_X(a-y)f_Y(y)dy$.
- Latter formula makes some intuitive sense. We're integrating over the set of x, y pairs that add up to a.

Conditional distributions

- Let's say X and Y have joint probability density function $f(x,y)$.
- We can *define* the conditional probability density of X given that $Y=y$ by $f_{X|Y=y}(x) = \frac{f(x,y)}{f_Y(y)}$.
- This amounts to restricting $f(x,y)$ to the line corresponding to the given y value (and dividing by the constant that makes the integral along that line equal to 1).

Maxima: pick five job candidates at random, choose best

- Suppose I choose n random variables $X_1, X_2, \dots, X_n$ uniformly at random on $[0,1]$, independently of each other.
- The n-tuple $(X_1, X_2, \dots, X_n)$ has a constant density function on the n-dimensional cube $[0,1]^n$.
- What is the probability that the *largest* of the X_i is less than a?
- ANSWER: a^n.
- So if $X = \max\{X_1, \dots, X_n\}$, then what is the probability density function of X?
- Answer: $F_X(a) = \begin{cases} 0 & a < 0 \\ a^n & a \in [0,1] \\ 1 & a > 1 \end{cases}$. And $f_x(a) = F_X'(a) = na^{n-1}$.

General order statistics

- Consider i.i.d random variables $X_1, X_2, \dots, X_n$ with continuous probability density f.
- Let $Y_1 < Y_2 < Y_3 \dots < Y_n$ be list obtained by *sorting* the X_j.
- In particular, $Y_1 = \min\{X_1, \dots, X_n\}$ and $Y_n = \max\{X_1, \dots, X_n\}$ is the maximum.
- What is the joint probability density of the Y_i?
- Answer: $f(x_1, x_2, \dots, x_n) = n! \prod_{i=1}^n f(x_i)$ if $x_1 < x_2 \dots < x_n$, zero otherwise.
- Let $\sigma : \{1, 2, \dots, n\} \to \{1, 2, \dots, n\}$ be the permutation such that $X_j = Y_{\sigma(j)}$
- Are σ and the vector $(Y_1, \dots, Y_n)$ independent of each other?
- Yes.

Properties of expectation

- Several properties we derived for discrete expectations continue to hold in the continuum.
- If X is discrete with mass function $p(x)$ then $E[X] = \sum_x p(x)x$.
- Similarly, if X is continuous with density function $f(x)$ then $E[X] = \int f(x)xdx$.
- If X is discrete with mass function $p(x)$ then $E[g(x)] = \sum_x p(x)g(x)$.
- Similarly, X if is continuous with density function $f(x)$ then $E[g(X)] = \int f(x)g(x)dx$.
- If X and Y have joint mass function $p(x,y)$ then $E[g(X,Y)] = \sum_y \sum_x g(x,y)p(x,y)$.
- If X and Y have joint probability density function $f(x,y)$ then $E[g(X,Y)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x,y)f(x,y)dxdy$.

Properties of expectation

- For both discrete and continuous random variables X and Y we have $E[X+Y] = E[X] + E[Y]$.
- In both discrete and continuous settings, $E[aX] = aE[X]$ when a is a constant. And $E[\sum a_i X_i] = \sum a_i E[X_i]$.
- But what about that delightful "area under $1 - F_X$" formula for the expectation?
- When X is non-negative with probability one, do we always have $E[X] = \int_0^\infty P\{X > x\}$, in both discrete and continuous settings?
- Define $g(y)$ so that $1 - F_X(g(y)) = y$. (Draw horizontal line at height y and look where it hits graph of $1 - F_X$.)
- Choose Y uniformly on $[0,1]$ and note that $g(Y)$ has the same probability distribution as X.
- So $E[X] = E[g(Y)] = \int_0^1 g(y)dy$, which is indeed the area under the graph of $1 - F_X$.

A property of independence

- If X and Y are independent then $E[g(X)h(Y)] = E[g(X)]E[h(Y)]$.
- Just write $E[g(X)h(Y)] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x)h(y)f(x,y)dxdy$.
- Since $f(x,y) = f_X(x)f_Y(y)$ this factors as $\int_{-\infty}^{\infty} h(y)f_Y(y)dy \int_{-\infty}^{\infty} g(x)f_X(x)dx = E[h(Y)]E[g(X)]$.

Defining covariance and correlation

- Now define covariance of X and Y by $\mathrm{Cov}(X,Y) = E[(X - E[X])(Y - E[Y])$.
- Note: by definition $\mathrm{Var}(X) = \mathrm{Cov}(X,X)$.
- Covariance formula $E[XY] - E[X]E[Y]$, or "expectation of product minus product of expectations" is frequently useful.
- If X and Y are independent then $\mathrm{Cov}(X,Y) = 0$.
- Converse is not true.

Basic covariance facts

- $\mathrm{Cov}(X,Y) = \mathrm{Cov}(Y,X)$
- $\mathrm{Cov}(X,X) = \mathrm{Var}(X)$
- $\mathrm{Cov}(aX,Y) = a\mathrm{Cov}(X,Y)$.
- $\mathrm{Cov}(X_1 + X_2, Y) = \mathrm{Cov}(X_1,Y) + \mathrm{Cov}(X_2,Y)$.
- **General statement of bilinearity of covariance:**

$$\mathrm{Cov}(\sum_{i=1}^m a_i X_i, \sum_{j=1}^n b_j Y_j) = \sum_{i=1}^m \sum_{j=1}^n a_i b_j \mathrm{Cov}(X_i, Y_j).$$

- Special case:

$$\mathrm{Var}(\sum_{i=1}^n X_i) = \sum_{i=1}^n \mathrm{Var}(X_i) + 2\sum_{(i,j):i<j} \mathrm{Cov}(X_i, X_j).$$

Defining correlation

- Again, by definition $\mathrm{Cov}(X,Y) = E[XY] - E[X]E[Y]$.
- **Correlation** of X and Y defined by

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

- Correlation doesn't care what units you use for X and Y. If $a > 0$ and $c > 0$ then $\rho(aX+b, cY+d) = \rho(X,Y)$.
- Satisfies $-1 \leq \rho(X,Y) \leq 1$.
- If a and b are positive constants and $a > 0$ then $\rho(aX+b, X) = 1$.
- If a and b are positive constants and $a < 0$ then $\rho(aX+b, X) = -1$.

Conditional probability distributions

- It all starts with the definition of conditional probability: $P(A|B) = P(AB)/P(B)$.
- If X and Y are jointly discrete random variables, we can use this to define a probability mass function for X *given* $Y = y$.
- That is, we write $p_{X|Y}(x|y) = P\{X = x|Y = y\} = \frac{p(x,y)}{p_Y(y)}$.
- In words: first restrict sample space to pairs (x,y) with given y value. Then divide the original mass function by $p_Y(y)$ to obtain a probability mass function on the restricted space.
- We do something similar when X and Y are continuous random variables. In that case we write $f_{X|Y}(x|y) = \frac{f(x,y)}{f_Y(y)}$.
- Often useful to think of sampling (X,Y) as a two-stage process. First sample Y from its marginal distribution, obtain $Y = y$ for some particular y. Then sample X from its probability distribution *given* $Y = y$.

Conditional expectation

- Now, what do we mean by $E[X|Y=y]$? This should just be the expectation of X in the conditional probability measure for X given that $Y = y$.
- Can write this as $E[X|Y=y] = \sum_x xP\{X = x|Y = y\} = \sum_x xp_{X|Y}(x|y)$.
- Can make sense of this in the continuum setting as well.
- In continuum setting we had $f_{X|Y}(x|y) = \frac{f(x,y)}{f_Y(y)}$. So $E[X|Y=y] = \int_{-\infty}^{\infty} x\frac{f(x,y)}{f_Y(y)}dx$

Conditional expectation as a random variable

- Can think of $E[X|Y]$ as a function of the random variable Y. When $Y = y$ it takes the value $E[X|Y=y]$.
- So $E[X|Y]$ is itself a random variable. It happens to depend only on the value of Y.
- Thinking of $E[X|Y]$ as a random variable, we can ask what *its* expectation is. What is $E[E[X|Y]]$?
- **Very useful fact:** $E[E[X|Y]] = E[X]$.
- In words: what you expect to expect X to be *after learning* Y is same as what you *now* expect X to be.
- Proof in discrete case: $E[X|Y=y] = \sum_x xP\{X = x|Y = y\} = \sum_x x\frac{p(x,y)}{p_Y(y)}$.
- Recall that, in general, $E[g(Y)] = \sum_y p_Y(y)g(y)$.
- $E[E[X|Y=y]] = \sum_y p_Y(y)\sum_x x\frac{p(x,y)}{p_Y(y)} = \sum_x \sum_y p(x,y)x = E[X]$.

Conditional variance

- Definition:
  $\mathrm{Var}(X|Y) = E\big[(X - E[X|Y])^2|Y\big] = E\big[X^2 - E[X|Y]^2|Y\big]$.
- $\mathrm{Var}(X|Y)$ is a random variable that depends on Y. It is the variance of X in the conditional distribution for X given Y.
- Note $E[\mathrm{Var}(X|Y)] = E[E[X^2|Y]] - E[E[X|Y]^2|Y] = E[X^2] - E[E[X|Y]^2]$.
- If we subtract $E[X]^2$ from first term and add equivalent value $E[E[X|Y]]^2$ to the second, RHS becomes $\mathrm{Var}[X] - \mathrm{Var}[E[X|Y]]$, which implies following:
- **Useful fact:** $\mathrm{Var}(X) = \mathrm{Var}(E[X|Y]) + E[\mathrm{Var}(X|Y)]$.
- One can discover X in two stages: first sample Y from marginal and compute $E[X|Y]$, then sample X from distribution given Y value.
- Above fact breaks variance into two parts, corresponding to these two stages.

Example

- Let X be a random variable of variance σ_X^2 and Y an independent random variable of variance σ_Y^2 and write $Z = X + Y$. Assume $E[X] = E[Y] = 0$.
- What are the covariances $\mathrm{Cov}(X, Y)$ and $\mathrm{Cov}(X, Z)$?
- How about the correlation coefficients $\rho(X, Y)$ and $\rho(X, Z)$?
- What is $E[Z|X]$? And how about $\mathrm{Var}(Z|X)$?
- Both of these values are functions of X. Former is just X. Latter happens to be a constant-valued function of X, i.e., happens not to actually depend on X. We have $\mathrm{Var}(Z|X) = \sigma_Y^2$.
- Can we check the formula $\mathrm{Var}(Z) = \mathrm{Var}(E[Z|X]) + E[\mathrm{Var}(Z|X)]$ in this case?

Moment generating functions

- Let X be a random variable and $M(t) = E[e^{tX}]$.
- Then $M'(0) = E[X]$ and $M''(0) = E[X^2]$. Generally, nth derivative of M at zero is $E[X^n]$.
- Let X and Y be independent random variables and $Z = X + Y$.
- Write the moment generating functions as $M_X(t) = E[e^{tX}]$ and $M_Y(t) = E[e^{tY}]$ and $M_Z(t) = E[e^{tZ}]$.
- If you knew M_X and M_Y, could you compute M_Z?
- By independence, $M_Z(t) = E[e^{t(X+Y)}] = E[e^{tX}e^{tY}] = E[e^{tX}]E[e^{tY}] = M_X(t)M_Y(t)$ for all t.
- In other words, adding independent random variables corresponds to multiplying moment generating functions.

Moment generating functions for sums of i.i.d. random variables

- We showed that if $Z = X + Y$ and X and Y are independent, then $M_Z(t) = M_X(t)M_Y(t)$
- If $X_1 \ldots X_n$ are i.i.d. copies of X and $Z = X_1 + \ldots + X_n$ then what is M_Z?
- Answer: M_X^n. Follows by repeatedly applying formula above.
- This a big reason for studying moment generating functions. It helps us understand what happens when we sum up a lot of independent copies of the same random variable.
- If $Z = aX$ then $M_Z(t) = E[e^{tZ}] = E[e^{taX}] = M_X(at)$.
- If $Z = X + b$ then $M_Z(t) = E[e^{tZ}] = E[e^{tX+bt}] = e^{bt}M_X(t)$.

Examples

- If X is binomial with parameters (p, n) then $M_X(t) = (pe^t + 1 - p)^n$.
- If X is Poisson with parameter $\lambda > 0$ then $M_X(t) = \exp[\lambda(e^t - 1)]$.
- If X is normal with mean 0, variance 1, then $M_X(t) = e^{t^2/2}$.
- If X is normal with mean μ, variance σ^2, then $M_X(t) = e^{\sigma^2 t^2/2+\mu t}$.
- If X is exponential with parameter $\lambda > 0$ then $M_X(t) = \frac{\lambda}{\lambda - t}$.

Cauchy distribution

- A standard **Cauchy random variable** is a random real number with probability density $f(x) = \frac{1}{\pi}\frac{1}{1+x^2}$.
- There is a "spinning flashlight" interpretation. Put a flashlight at $(0, 1)$, spin it to a uniformly random angle in $[-\pi/2, \pi/2]$, and consider point X where light beam hits the x-axis.
- $F_X(x) = P\{X \le x\} = P\{\tan\theta \le x\} = P\{\theta \le tan^{-1}x\} = \frac{1}{2} + \frac{1}{\pi}\tan^{-1} x$.
- Find $f_X(x) = \frac{d}{dx}F(x) = \frac{1}{\pi}\frac{1}{1+x^2}$.

Beta distribution

- Two part experiment: first let p be uniform random variable $[0, 1]$, then let X be binomial (n, p) (number of heads when we toss n p-coins).
- **Given** that $X = a - 1$ and $n - X = b - 1$ the conditional law of p is called the β distribution.
- The density function is a constant (that doesn't depend on x) times $x^{a-1}(1-x)^{b-1}$.
- That is $f(x) = \frac{1}{B(a,b)}x^{a-1}(1-x)^{b-1}$ on $[0, 1]$, where $B(a, b)$ is constant chosen to make integral one. Can show $B(a, b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$.
- Turns out that $E[X] = \frac{a}{a+b}$ and the mode of X is $\frac{(a-1)}{(a-1)+(b-1)}$.

18.600: Lecture 29

Weak law of large numbers

Scott Sheffield

MIT

Outline

Outline

Markov's and Chebyshev's inequalities

- **Markov's inequality:** Let X be a random variable taking only non-negative values. Fix a constant $a > 0$. Then $P\{X \geq a\} \leq \frac{E[X]}{a}$.
- **Proof:** Consider a random variable Y defined by $Y = \begin{cases} a & X \geq a \\ 0 & X < a \end{cases}$. Since $X \geq Y$ with probability one, it follows that $E[X] \geq E[Y] = aP\{X \geq a\}$. Divide both sides by a to get Markov's inequality.
- **Chebyshev's inequality:** If X has finite mean μ, variance σ^2, and $k > 0$ then
$$P\{|X - \mu| \geq k\} \leq \frac{\sigma^2}{k^2}.$$
- **Proof:** Note that $(X - \mu)^2$ is a non-negative random variable and $P\{|X - \mu| \geq k\} = P\{(X - \mu)^2 \geq k^2\}$. Now apply Markov's inequality with $a = k^2$.

Markov and Chebyshev: rough idea

- **Markov's inequality:** Let X be a random variable taking only non-negative values with finite mean. Fix a constant $a > 0$. Then $P\{X \geq a\} \leq \frac{E[X]}{a}$.
- **Chebyshev's inequality:** If X has finite mean μ, variance σ^2, and $k > 0$ then

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

- Inequalities allow us to deduce limited information about a distribution when we know only the mean (Markov) or the mean and variance (Chebyshev).
- **Markov:** if $E[X]$ is small, then it is not too likely that X is large.
- **Chebyshev:** if $\sigma^2 = \mathrm{Var}[X]$ is small, then it is not too likely that X is far from its mean.

Statement of weak law of large numbers

- Suppose X_i are i.i.d. random variables with mean μ.
- Then the value $A_n := \frac{X_1 + X_2 + \ldots + X_n}{n}$ is called the *empirical average* of the first n trials.
- We'd guess that when n is large, A_n is typically close to μ.
- Indeed, **weak law of large numbers** states that for all $\epsilon > 0$ we have $\lim_{n \to \infty} P\{|A_n - \mu| > \epsilon\} = 0$.
- Example: as n tends to infinity, the probability of seeing more than $.50001n$ heads in n fair coin tosses tends to zero.

Proof of weak law of large numbers in finite variance case

- As above, let X_i be i.i.d. random variables with mean μ and write $A_n := \frac{X_1 + X_2 + \ldots + X_n}{n}$.
- By additivity of expectation, $\mathbb{E}[A_n] = \mu$.
- Similarly, $\mathrm{Var}[A_n] = \frac{n\sigma^2}{n^2} = \sigma^2/n$.
- By Chebyshev $P\{|A_n - \mu| \geq \epsilon\} \leq \frac{\mathrm{Var}[A_n]}{\epsilon^2} = \frac{\sigma^2}{n\epsilon^2}$.
- No matter how small ϵ is, RHS will tend to zero as n gets large.

Outline

Outline

Weak law of large numbers: Markov/Chebyshev approach

Weak law of large numbers: characteristic function approach

Extent of weak law

- Question: does the weak law of large numbers apply no matter what the probability distribution for X is?
- Is it always the case that if we define $A_n := \frac{X_1+X_2+\ldots+X_n}{n}$ then A_n is typically close to some fixed value when n is large?
- What if X is Cauchy?
- Recall that in this strange case A_n actually has the same probability distribution as X.
- In particular, the A_n are not tightly concentrated around any particular value even when n is very large.
- But in this case $E[|X|]$ was infinite. Does the weak law hold as long as $E[|X|]$ is finite, so that μ is well defined?
- Yes. Can prove this using characteristic functions.

Characteristic functions

- Let X be a random variable.
- The **characteristic function** of X is defined by $\phi(t) = \phi_X(t) := E[e^{itX}]$. Like $M(t)$ except with i thrown in.
- Recall that by definition $e^{it} = \cos(t) + i\sin(t)$.
- Characteristic functions are similar to moment generating functions in some ways.
- For example, $\phi_{X+Y} = \phi_X\phi_Y$, just as $M_{X+Y} = M_X M_Y$, if X and Y are independent.
- And $\phi_{aX}(t) = \phi_X(at)$ just as $M_{aX}(t) = M_X(at)$.
- And if X has an mth moment then $E[X^m] = i^m \phi_X^{(m)}(0)$.
- But characteristic functions have an advantage: they are well defined at all t for all random variables X.

Continuity theorems

- Let X be a random variable and X_n a sequence of random variables.
- Say X_n **converge in distribution** or **converge in law** to X if $\lim_{n\to\infty} F_{X_n}(x) = F_X(x)$ at all $x \in \mathbb{R}$ at which F_X is continuous.
- The weak law of large numbers can be rephrased as the statement that A_n converges in law to μ (i.e., to the random variable that is equal to μ with probability one).
- **Lévy's continuity theorem (see Wikipedia):** if
$$\lim_{n\to\infty} \phi_{X_n}(t) = \phi_X(t)$$
for all t, then X_n converge in law to X.
- By this theorem, we can prove the weak law of large numbers by showing $\lim_{n\to\infty} \phi_{A_n}(t) = \phi_\mu(t) = e^{it\mu}$ for all t. In the special case that $\mu = 0$, this amounts to showing $\lim_{n\to\infty} \phi_{A_n}(t) = 1$ for all t.

Proof of weak law of large numbers in finite mean case

- As above, let X_i be i.i.d. instances of random variable X with mean zero. Write $A_n := \frac{X_1+X_2+\ldots+X_n}{n}$. Weak law of large numbers holds for i.i.d. instances of X if and only if it holds for i.i.d. instances of $X-\mu$. Thus it suffices to prove the weak law in the mean zero case.
- Consider the characteristic function $\phi_X(t) = E[e^{itX}]$.
- Since $E[X]=0$, we have $\phi_X'(0) = E[\frac{\partial}{\partial t} e^{itX}]_{t=0} = iE[X] = 0$.
- Write $g(t) = \log \phi_X(t)$ so $\phi_X(t) = e^{g(t)}$. Then $g(0)=0$ and (by chain rule) $g'(0) = \lim_{\epsilon\to 0} \frac{g(\epsilon)-g(0)}{\epsilon} = \lim_{\epsilon \to 0} \frac{g(\epsilon)}{\epsilon} = 0$.
- Now $\phi_{A_n}(t) = \phi_X(t/n)^n = e^{ng(t/n)}$. Since $g(0)=g'(0)=0$ we have $\lim_{n\to\infty} ng(t/n) = \lim_{n\to\infty} t\frac{g(\frac{t}{n})}{\frac{t}{n}} = 0$ if t is fixed. Thus $\lim_{n\to\infty} e^{ng(t/n)} = 1$ for all t.
- By Lévy's continuity theorem, the A_n converge in law to 0 (i.e., to the random variable that is 0 with probability one).

18.600: Lecture 30

Central limit theorem

Scott Sheffield

MIT

Outline

Central limit theorem

Proving the central limit theorem

Outline

Central limit theorem

Proving the central limit theorem

Recall: DeMoivre-Laplace limit theorem

- Let X_i be an i.i.d. sequence of random variables. Write $S_n = \sum_{i=1}^n X_n$.
- Suppose each X_i is 1 with probability p and 0 with probability $q = 1 - p$.
- **DeMoivre-Laplace limit theorem:**

$$\lim_{n\to\infty} P\{a \le \frac{S_n - np}{\sqrt{npq}} \le b\} \to \Phi(b) - \Phi(a).$$

- Here $\Phi(b) - \Phi(a) = P\{a \le Z \le b\}$ when Z is a standard normal random variable.
- $\frac{S_n - np}{\sqrt{npq}}$ describes "number of standard deviations that S_n is above or below its mean".
- Question: Does a similar statement hold if the X_i are i.i.d. but have some other probability distribution?
- **Central limit theorem:** Yes, if they have finite variance.

Example

- Say we roll 10^6 ordinary dice independently of each other.
- Let X_i be the number on the ith die. Let $X = \sum_{i=1}^{10^6} X_i$ be the total of the numbers rolled.
- What is $E[X]$?
- $10^6/6$
- What is $\mathrm{Var}[X]$?
- $10^6 \cdot (35/12)$
- How about $\mathrm{SD}[X]$?
- $1000\sqrt{35/12}$
- What is the probability that X is less than a standard deviations above its mean?
- Central limit theorem: should be about $\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{a} e^{-x^2/2}dx$.

Example

- Suppose earthquakes in some region are a Poisson point process with rate λ equal to 1 per year.
- Let X be the number of earthquakes that occur over a ten-thousand year period. Should be a Poisson random variable with rate 10000.
- What is $E[X]$?
- 10000
- What is $\mathrm{Var}[X]$?
- 10000
- How about $\mathrm{SD}[X]$?
- 100
- What is the probability that X is less than a standard deviations above its mean?
- Central limit theorem: should be about $\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{a} e^{-x^2/2}dx$.

General statement

- Let X_i be an i.i.d. sequence of random variables with finite mean μ and variance σ^2.
- Write $S_n = \sum_{i=1}^{n} X_i$. So $E[S_n] = n\mu$ and $\mathrm{Var}[S_n] = n\sigma^2$ and $\mathrm{SD}[S_n] = \sigma\sqrt{n}$.
- Write $B_n = \frac{X_1+X_2+\ldots+X_n-n\mu}{\sigma\sqrt{n}}$. Then B_n is the difference between S_n and its expectation, measured in standard deviation units.
- **Central limit theorem:**

$$\lim_{n\to\infty} P\{a \le B_n \le b\} \to \Phi(b) - \Phi(a).$$

Outline

Outline

Central limit theorem

Proving the central limit theorem

Recall: characteristic functions

- Let X be a random variable.
- The **characteristic function** of X is defined by $\phi(t) = \phi_X(t) := E[e^{itX}]$. Like $M(t)$ except with i thrown in.
- Recall that by definition $e^{it} = \cos(t) + i\sin(t)$.
- Characteristic functions are similar to moment generating functions in some ways.
- For example, $\phi_{X+Y} = \phi_X \phi_Y$, just as $M_{X+Y} = M_X M_Y$, if X and Y are independent.
- And $\phi_{aX}(t) = \phi_X(at)$ just as $M_{aX}(t) = M_X(at)$.
- And if X has an mth moment then $E[X^m] = i^m \phi_X^{(m)}(0)$.
- Characteristic functions are well defined at all t for all random variables X.

Rephrasing the theorem

- Let X be a random variable and X_n a sequence of random variables.
- Say X_n **converge in distribution** or **converge in law** to X if $\lim_{n\to\infty} F_{X_n}(x) = F_X(x)$ at all $x \in \mathbb{R}$ at which F_X is continuous.
- Recall: the weak law of large numbers can be rephrased as the statement that $A_n = \frac{X_1+X_2+\ldots+X_n}{n}$ converges in law to μ (i.e., to the random variable that is equal to μ with probability one) as $n \to \infty$.
- The central limit theorem can be rephrased as the statement that $B_n = \frac{X_1+X_2+\ldots+X_n-n\mu}{\sigma\sqrt{n}}$ converges in law to a standard normal random variable as $n \to \infty$.

Continuity theorems

- **Lévy's continuity theorem (see Wikipedia):** if
$$\lim_{n\to\infty} \phi_{X_n}(t) = \phi_X(t)$$
for all t, then X_n converge in law to X.
- By this theorem, we can prove the central limit theorem by showing $\lim_{n\to\infty} \phi_{B_n}(t) = e^{-t^2/2}$ for all t.
- **Moment generating function continuity theorem:** if moment generating functions $M_{X_n}(t)$ are defined for all t and n and $\lim_{n\to\infty} M_{X_n}(t) = M_X(t)$ for all t, then X_n converge in law to X.
- By this theorem, we can prove the central limit theorem by showing $\lim_{n\to\infty} M_{B_n}(t) = e^{t^2/2}$ for all t.

Proof of central limit theorem with moment generating functions

- Write $Y = \frac{X-\mu}{\sigma}$. Then Y has mean zero and variance 1.
- Write $M_Y(t) = E[e^{tY}]$ and $g(t) = \log M_Y(t)$. So $M_Y(t) = e^{g(t)}$.
- We know $g(0) = 0$. Also $M'_Y(0) = E[Y] = 0$ and $M''_Y(0) = E[Y^2] = \mathrm{Var}[Y] = 1$.
- Chain rule: $M'_Y(0) = g'(0)e^{g(0)} = g'(0) = 0$ and $M''_Y(0) = g''(0)e^{g(0)} + g'(0)^2 e^{g(0)} = g''(0) = 1$.
- So g is a nice function with $g(0) = g'(0) = 0$ and $g''(0) = 1$. Taylor expansion: $g(t) = t^2/2 + o(t^2)$ for t near zero.
- Now B_n is $\frac{1}{\sqrt{n}}$ times the sum of n independent copies of Y.
- So $M_{B_n}(t) = \left(M_Y(t/\sqrt{n})\right)^n = e^{ng(\frac{t}{\sqrt{n}})}$.
- But $e^{ng(\frac{t}{\sqrt{n}})} \approx e^{n(\frac{t}{\sqrt{n}})^2/2} = e^{t^2/2}$, in sense that LHS tends to $e^{t^2/2}$ as n tends to infinity.

Proof of central limit theorem with characteristic functions

- Moment generating function proof only applies if the moment generating function of X exists.
- But the proof can be repeated almost verbatim using characteristic functions instead of moment generating functions.
- Then it applies for any X with finite variance.

Almost verbatim: replace $M_Y(t)$ with $\phi_Y(t)$

- Write $\phi_Y(t) = E[e^{itY}]$ and $g(t) = \log \phi_Y(t)$. So $\phi_Y(t) = e^{g(t)}$.
- We know $g(0) = 0$. Also $\phi'_Y(0) = iE[Y] = 0$ and $\phi''_Y(0) = i^2 E[Y^2] = -\mathrm{Var}[Y] = -1$.
- Chain rule: $\phi'_Y(0) = g'(0)e^{g(0)} = g'(0) = 0$ and $\phi''_Y(0) = g''(0)e^{g(0)} + g'(0)^2 e^{g(0)} = g''(0) = -1$.
- So g is a nice function with $g(0) = g'(0) = 0$ and $g''(0) = -1$. Taylor expansion: $g(t) = -t^2/2 + o(t^2)$ for t near zero.
- Now B_n is $\frac{1}{\sqrt{n}}$ times the sum of n independent copies of Y.
- So $\phi_{B_n}(t) = \left(\phi_Y(t/\sqrt{n})\right)^n = e^{ng(\frac{t}{\sqrt{n}})}$.
- But $e^{ng(\frac{t}{\sqrt{n}})} \approx e^{-n(\frac{t}{\sqrt{n}})^2/2} = e^{-t^2/2}$, in sense that LHS tends to $e^{-t^2/2}$ as n tends to infinity.

Perspective

- The central limit theorem is actually fairly robust. Variants of the theorem still apply if you allow the X_i not to be identically distributed, or not to be completely independent.
- We won't formulate these variants precisely in this course.
- But, roughly speaking, if you have a lot of little random terms that are "mostly independent" — and no single term contributes more than a "small fraction" of the total sum — then the total sum should be "approximately" normal.
- Example: if height is determined by lots of little mostly independent factors, then people's heights should be normally distributed.
- Not quite true... certain factors by themselves can cause a person to be a whole lot shorter or taller. Also, individual factors not really independent of each other.
- *Kind of* true for homogenous population, ignoring outliers.

18.600: Lecture 31

Strong law of large numbers and Jensen's inequality

Scott Sheffield

MIT

Outline

Outline

Pedro's hopes and dreams

- Pedro is considering two ways to invest his life savings.
- One possibility: put the entire sum in government insured interest-bearing savings account. He considers this completely risk free. The (post-tax) interest rate equals the inflation rate, so the real value of his savings is guaranteed not to change.
- Riskier possibility: put sum in investment where every month *real* value goes up 15 percent with probability .53 and down 15 percent with probability .47 (independently of everything else).
- How much does Pedro make in expectation over 10 years with risky approach? 100 years?

Pedro's hopes and dreams

- How much does Pedro make in expectation over 10 years with risky approach? 100 years?
- Answer: let R_i be i.i.d. random variables each equal to 1.15 with probability .53 and .85 with probability .47. Total value after n steps is initial investment times $T_n := R_1 \times R_2 \times \ldots \times R_n$.
- Compute $E[R_1] = .53 \times 1.15 + .47 \times .85 = 1.009$.
- Then $E[T_{120}] = 1.009^{120} \approx 2.93$. And $E[T_{1200}] = 1.009^{1200} \approx 46808.9$

Pedro's financial planning

- How would you advise Pedro to invest over the next 10 years if Pedro wants to be completely sure that he doesn't lose money?
- What if Pedro is willing to accept substantial risk if it means there is a good chance it will enable his grandchildren to retire in comfort 100 years from now?
- What if Pedro wants the money for himself in ten years?
- Let's do some simulations.

Logarithmic point of view

- We wrote $T_n = R_1 \times \ldots \times R_n$. Taking logs, we can write $X_i = \log R_i$ and $S_n = \log T_n = \sum_{i=1}^n X_i$.
- Now S_n is a sum of i.i.d. random variables.
- $E[X_1] = E[\log R_1] = .53(\log 1.15) + .47(\log .85) \approx -.0023$.
- By the law of large numbers, if we take n extremely large, then $S_n/n \approx -.0023$ with high probability.
- This means that, when n is large, S_n is *usually* a very negative value, which means T_n is *usually* very close to zero (even though its expectation is very large).
- Bad news for Pedro's grandchildren. After 100 years, the portfolio is probably in bad shape. But what if Pedro takes an even longer view? Will T_n converge to zero with probability one as n gets large? Or will T_n perhaps always *eventually* rebound?

Outline

Outline

A story about Pedro

Strong law of large numbers

Jensen's inequality

Strong law of large numbers

- Suppose X_i are i.i.d. random variables with mean μ.
- Then the value $A_n := \frac{X_1+X_2+\ldots+X_n}{n}$ is called the *empirical average* of the first n trials.
- Intuition: when n is large, A_n is typically close to μ.
- Recall: **weak law of large numbers** states that for all $\epsilon > 0$ we have $\lim_{n\to\infty} P\{|A_n - \mu| > \epsilon\} = 0$.
- The **strong law of large numbers** states that with probability one $\lim_{n\to\infty} A_n = \mu$.
- It is called "strong" because it implies the weak law of large numbers. But it takes a bit of thought to see why this is the case.

Strong law implies weak law

- Suppose we know that the strong law holds, i.e., with probability 1 we have $\lim_{n\to\infty} A_n = \mu$.
- Strong law implies that for every ϵ the random variable $Y_\epsilon = \max\{n : |A_n - \mu| > \epsilon\}$ is finite with probability one. It has some probability mass function (though we don't know what it is).
- Note that if $|A_n - \mu| > \epsilon$ for some n value then $Y_\epsilon \geq n$.
- Thus for each n we have $P\{|A_n - \mu| > \epsilon\} \leq P\{Y_\epsilon \geq n\}$.
- So $\lim_{n\to\infty} P\{|A_n - \mu| > \epsilon\} \leq \lim_{n\to\infty} P\{Y_\epsilon \geq n\} = 0$.
- If the right limit is zero for each ϵ (strong law) then the left limit is zero for each ϵ (weak law).

Proof of strong law assuming $E[X^4] < \infty$

- Assume $K := E[X^4] < \infty$. Not necessary, but simplifies proof.
- Note: $\mathrm{Var}[X^2] = E[X^4] - E[X^2]^2 > 0$, so $E[X^2]^2 \leq K$.
- The strong law holds for i.i.d. copies of X if and only if it holds for i.i.d. copies of $X - \mu$ where μ is a constant.
- So we may as well assume $E[X] = 0$.
- Key to proof is to bound fourth moments of A_n.
- $E[A_n^4] = n^{-4}E[S_n^4] = n^{-4}E[(X_1 + X_2 + \ldots + X_n)^4]$.
- Expand $(X_1 + \ldots + X_n)^4$. Five kinds of terms: $X_iX_jX_kX_l$ and $X_iX_jX_k^2$ and $X_iX_j^3$ and $X_i^2X_j^2$ and X_i^4.
- The first three terms all have expectation zero. There are $\binom{n}{2}$ of the fourth type and n of the last type, each equal to at most K. So $E[A_n^4] \leq n^{-4}\left(6\binom{n}{2} + n\right)K$.
- Thus $E[\sum_{n=1}^\infty A_n^4] = \sum_{n=1}^\infty E[A_n^4] < \infty$. So $\sum_{n=1}^\infty A_n^4 < \infty$ (and hence $A_n \to 0$) with probability 1.

Outline

A story about Pedro

Strong law of large numbers

Jensen's inequality

Outline

A story about Pedro

Strong law of large numbers

Jensen's inequality

Jensen's inequality statement

- Let X be random variable with finite mean $E[X] = \mu$.
- Let g be a **convex** function. This means that if you draw a straight line connecting two points on the graph of g, then the graph of g lies below that line. If g is twice differentiable, then convexity is equivalent to the statement that $g''(x) \geq 0$ for all x. For a concrete example, take $g(x) = x^2$.
- **Jensen's inequality:** $E[g(X)] \geq g(E[X])$.
- **Proof:** Let $L(x) = ax + b$ be tangent to graph of g at point $(E[X], g(E[X]))$. Then L lies below g. Observe
$$E[g(X)] \geq E[L(X)] = L(E[X]) = g(E[X])$$
- **Note:** if g is **concave** (which means $-g$ is convex), then $E[g(X)] \leq g(E[X])$.
- If your utility function is concave, then you always prefer a safe investment over a risky investment with the same expected return.

More about Pedro

- Disappointed by the strong law of large numbers, Pedro seeks a better way to make money.
- Signs up for job as "hedge fund manager". Allows him to manage $C \approx 10^9$ dollars of somebody else's money. At end of each year, he and his staff get two percent of principle plus twenty percent of profit.
- Precisely: if X is end-of-year portfolio value, Pedro gets
$$g(X) = .02C + .2 \max\{X - C, 0\}.$$
- Pedro notices that g is a convex function. He can therefore increase his expected return by adopting risky strategies.
- Pedro has strategy that increases portfolio value 10 percent with probability .9, loses everything with probability .1.
- He repeats this yearly until fund collapses.
- With high probability Pedro is rich by then.

Perspective

- The "two percent of principle plus twenty percent of profit" is common in the hedge fund industry.
- The idea is that fund managers have both guaranteed revenue for expenses (two percent of principle) and incentive to make money (twenty percent of profit).
- Because of Jensen's inequality, the convexity of the payoff function is a genuine concern for hedge fund investors. People worry that it encourages fund managers (like Pedro) to take risks that are bad for the client.
- This is a special case of the "principal-agent" problem of economics. How do you ensure that the people you hire genuinely share your interests?

18.600: Lecture 32

Markov Chains

Scott Sheffield

MIT

Outline

Outline

Markov chains

- Consider a sequence of random variables $X_0, X_1, X_2, \ldots$ each taking values in the same state space, which for now we take to be a finite set that we label by $\{0, 1, \ldots, M\}$.
- Interpret X_n as state of the system at time n.
- Sequence is called a **Markov chain** if we have a fixed collection of numbers P_{ij} (one for each pair $i, j \in \{0, 1, \ldots, M\}$) such that whenever the system is in state i, there is probability P_{ij} that system will next be in state j.
- Precisely, $P\{X_{n+1} = j | X_n = i, X_{n-1} = i_{n-1}, \ldots, X_1 = i_1, X_0 = i_0\} = P_{ij}$.
- Kind of an "almost memoryless" property. Probability distribution for next state depends only on the current state (and not on the rest of the state history).

Simple example

- For example, imagine a simple weather model with two states: rainy and sunny.
- If it's rainy one day, there's a .5 chance it will be rainy the next day, a .5 chance it will be sunny.
- If it's sunny one day, there's a .8 chance it will be sunny the next day, a .2 chance it will be rainy.
- In this climate, sun tends to last longer than rain.
- Given that it is rainy today, how many days to I expect to have to wait to see a sunny day?
- Given that it is sunny today, how many days to I expect to have to wait to see a rainy day?
- Over the long haul, what fraction of days are sunny?

Matrix representation

- To describe a Markov chain, we need to define P_{ij} for any $i, j \in \{0, 1, \ldots, M\}$.
- It is convenient to represent the collection of transition probabilities P_{ij} as a matrix:

$$A = \begin{pmatrix} P_{00} & P_{01} & \ldots & P_{0M} \\ P_{10} & P_{11} & \ldots & P_{1M} \\ \cdot & & & \\ \cdot & & & \\ \cdot & & & \\ P_{M0} & P_{M1} & \ldots & P_{MM} \end{pmatrix}$$

- For this to make sense, we require $P_{ij} \geq 0$ for all i, j and $\sum_{j=0}^{M} P_{ij} = 1$ for each i. That is, the rows sum to one.

Transitions via matrices

- Suppose that p_i is the probability that system is in state i at time zero.
- What does the following product represent?

$$\begin{pmatrix} p_0 & p_1 & \ldots & p_M \end{pmatrix} \begin{pmatrix} P_{00} & P_{01} & \ldots & P_{0M} \\ P_{10} & P_{11} & \ldots & P_{1M} \\ \cdot & & & \\ \cdot & & & \\ \cdot & & & \\ P_{M0} & P_{M1} & \ldots & P_{MM} \end{pmatrix}$$

- Answer: the probability distribution at time one.
- How about the following product?

$$\begin{pmatrix} p_0 & p_1 & \ldots & p_M \end{pmatrix} A^n$$

- Answer: the probability distribution at time n.

Powers of transition matrix

- We write $P_{ij}^{(n)}$ for the probability to go from state i to state j over n steps.
- From the matrix point of view

$$\begin{pmatrix} P_{00}^{(n)} & P_{01}^{(n)} & \ldots & P_{0M}^{(n)} \\ P_{10}^{(n)} & P_{11}^{(n)} & \ldots & P_{1M}^{(n)} \\ \cdot & & & \\ \cdot & & & \\ \cdot & & & \\ P_{M0}^{(n)} & P_{M1}^{(n)} & \ldots & P_{MM}^{(n)} \end{pmatrix} = \begin{pmatrix} P_{00} & P_{01} & \ldots & P_{0M} \\ P_{10} & P_{11} & \ldots & P_{1M} \\ \cdot & & & \\ \cdot & & & \\ \cdot & & & \\ P_{M0} & P_{M1} & \ldots & P_{MM} \end{pmatrix}^n$$

- If A is the one-step transition matrix, then A^n is the n-step transition matrix.

Questions

- What does it mean if all of the rows are identical?
- Answer: state sequence X_i consists of i.i.d. random variables.
- What if matrix is the identity?
- Answer: states never change.
- What if each P_{ij} is either one or zero?
- Answer: state evolution is deterministic.

Outline

Outline

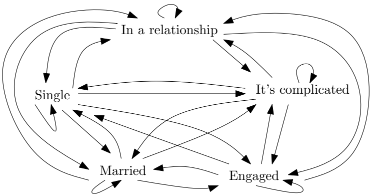
Simple example

- Consider the simple weather example: If it's rainy one day, there's a .5 chance it will be rainy the next day, a .5 chance it will be sunny. If it's sunny one day, there's a .8 chance it will be sunny the next day, a .2 chance it will be rainy.
- Let rainy be state zero, sunny state one, and write the transition matrix by

$$A = \begin{pmatrix} .5 & .5 \\ .2 & .8 \end{pmatrix}$$

- Note that
$$A^2 = \begin{pmatrix} .64 & .35 \\ .26 & .74 \end{pmatrix}$$
- Can compute $A^{10} = \begin{pmatrix} .285719 & .714281 \\ .285713 & .714287 \end{pmatrix}$

Does relationship status have the Markov property?



- Can we assign a probability to each arrow?
- Markov model implies time spent in any state (e.g., a marriage) before leaving is a geometric random variable.
- Not true... Can we make a better model with more states?

Outline

Markov chains

Examples

Ergodicity and stationarity

Outline

Markov chains

Examples

Ergodicity and stationarity

Ergodic Markov chains

- Say Markov chain is **ergodic** if some power of the transition matrix has all non-zero entries.
- Turns out that if chain has this property, then $\pi_j := \lim_{n\to\infty} P_{ij}^{(n)}$ exists and the π_j are the unique non-negative solutions of $\pi_j = \sum_{k=0}^{M} \pi_k P_{kj}$ that sum to one.
- This means that the row vector

$$\pi = \begin{pmatrix} \pi_0 & \pi_1 & \dots & \pi_M \end{pmatrix}$$

is a left eigenvector of A with eigenvalue 1, i.e., $\pi A = \pi$.
- We call π the *stationary distribution* of the Markov chain.
- One can solve the system of linear equations $\pi_j = \sum_{k=0}^{M} \pi_k P_{kj}$ to compute the values π_j. Equivalent to considering A fixed and solving $\pi A = \pi$. Or solving $(A - I)\pi = 0$. This determines π up to a multiplicative constant, and fact that $\sum \pi_j = 1$ determines the constant.

Simple example

- If $A = \begin{pmatrix} .5 & .5 \\ .2 & .8 \end{pmatrix}$, then we know

$$\pi A = \begin{pmatrix} \pi_0 & \pi_1 \end{pmatrix} \begin{pmatrix} .5 & .5 \\ .2 & .8 \end{pmatrix} = \begin{pmatrix} \pi_0 & \pi_1 \end{pmatrix} = \pi.$$

- This means that $.5\pi_0 + .2\pi_1 = \pi_0$ and $.5\pi_0 + .8\pi_1 = \pi_1$ and we also know that $\pi_0 + \pi_1 = 1$. Solving these equations gives $\pi_0 = 2/7$ and $\pi_1 = 5/7$, so $\pi = \begin{pmatrix} 2/7 & 5/7 \end{pmatrix}$.
- Indeed,

$$\pi A = \begin{pmatrix} 2/7 & 5/7 \end{pmatrix} \begin{pmatrix} .5 & .5 \\ .2 & .8 \end{pmatrix} = \begin{pmatrix} 2/7 & 5/7 \end{pmatrix} = \pi.$$

- Recall that $A^{10} = \begin{pmatrix} .285719 & .714281 \\ .285713 & .714287 \end{pmatrix} \approx \begin{pmatrix} 2/7 & 5/7 \\ 2/7 & 5/7 \end{pmatrix} = \begin{pmatrix} \pi \\ \pi \end{pmatrix}$

18.600: Lecture 33

Entropy

Scott Sheffield

MIT

Outline

Outline

What is entropy?

- Entropy is an important notion in thermodynamics, information theory, data compression, cryptography, etc.
- Familiar on some level to everyone who has studied chemistry or statistical physics.
- Kind of means amount of randomness or disorder.
- But can we give a mathematical definition? In particular, how do we define the entropy of a random variable?

Information

- Suppose we toss a fair coin k times.
- Then the state space S is the set of 2^k possible heads-tails sequences.
- If X is the random sequence (so X is a random variable), then for each $x \in S$ we have $P\{X = x\} = 2^{-k}$.
- In information theory it's quite common to use log to mean $\log_2$ instead of $\log_e$. We follow that convention in this lecture. In particular, this means that

$$\log P\{X = x\} = -k$$

for each $x \in S$.
- Since there are 2^k values in S, it takes k "bits" to describe an element $x \in S$.
- Intuitively, could say that when we learn that $X = x$, we have learned $k = -\log P\{X = x\}$ "bits of information".

Shannon entropy

- Shannon: famous MIT student/faculty member, wrote *The Mathematical Theory of Communication* in 1948.
- Goal is to define a notion of how much we "expect to learn" from a random variable or "how many bits of information a random variable contains" that makes sense for general experiments (which may not have anything to do with coins).
- If a random variable X takes values $x_1, x_2, \dots, x_n$ with positive probabilities $p_1, p_2, \dots, p_n$ then we define the **entropy** of X by

$$H(X) = \sum_{i=1}^{n} p_i(-\log p_i) = -\sum_{i=1}^{n} p_i \log p_i.$$

- This can be interpreted as the expectation of $(-\log p_i)$. The value $(-\log p_i)$ is the "amount of surprise" when we see x_i.

Twenty questions with Harry

- Harry always thinks of one of the following animals:

x	$P\{X = x\}$	$-\log P\{X = x\}$
Dog	1/4	2
Cat	1/4	2
Cow	1/8	3
Pig	1/16	4
Squirrel	1/16	4
Mouse	1/16	4
Owl	1/16	4
Sloth	1/32	5
Hippo	1/32	5
Yak	1/32	5
Zebra	1/64	6
Rhino	1/64	6

- Can learn animal with $H(X) = \frac{47}{16}$ questions on average.

Other examples

- Again, if a random variable X takes the values $x_1, x_2, \dots, x_n$ with positive probabilities $p_1, p_2, \dots, p_n$ then we define the **entropy** of X by

$$H(X) = \sum_{i=1}^{n} p_i(-\log p_i) = -\sum_{i=1}^{n} p_i \log p_i.$$

- If X takes one value with probability 1, what is $H(X)$?
- If X takes k values with equal probability, what is $H(X)$?
- What is $H(X)$ if X is a geometric random variable with parameter $p = 1/2$?

Outline

Entropy

Noiseless coding theory

Conditional entropy

Outline

Entropy

Noiseless coding theory

Conditional entropy

Coding values by bit sequences

- David Huffman (as MIT student) published in "A Method for the Construction of Minimum-Redundancy Code" in 1952.
- If X takes four values A, B, C, D we can code them by:

$$A \leftrightarrow 00$$
$$B \leftrightarrow 01$$
$$C \leftrightarrow 10$$
$$D \leftrightarrow 11$$

- Or by

$$A \leftrightarrow 0$$
$$B \leftrightarrow 10$$
$$C \leftrightarrow 110$$
$$D \leftrightarrow 111$$

- No sequence in code is an extension of another.
- What does 100111110010 spell?
- A coding scheme is equivalent to a twenty questions strategy.

Twenty questions theorem

- **Noiseless coding theorem:** Expected number of questions you need is always at least the entropy.
- **Note:** The expected number of questions *is* the entropy if each question divides the space of possibilities exactly in half (measured by probability).
- In this case, let X take values $x_1, \ldots, x_N$ with probabilities $p(x_1), \ldots, p(x_N)$. Then if a valid coding of X assigns n_i bits to x_i, we have

$$\sum_{i=1}^{N} n_i p(x_i) \geq H(X) = -\sum_{i=1}^{N} p(x_i) \log p(x_i).$$

- **Data compression:** $X_1, X_2, \ldots, X_n$ be i.i.d. instances of X. Do there exist encoding schemes such that the expected number of bits required to encode the entire sequence is about $H(X)n$ (assuming n is sufficiently large)?
- Yes. We can cut space of N^n possibilities *close* to exactly in half at each stage (up till near end maybe).

Outline

Entropy

Noiseless coding theory

Conditional entropy

Outline

Entropy

Noiseless coding theory

Conditional entropy

Entropy for a pair of random variables

- Consider random variables X, Y with joint mass function $p(x_i, y_j) = P\{X = x_i, Y = y_j\}$.
- Then we write

$$H(X, Y) = -\sum_i \sum_j p(x_i, y_j) \log p(x_i, y_i).$$

- $H(X, Y)$ is just the entropy of the pair (X, Y) (viewed as a random variable itself).
- Claim: if X and Y are independent, then

$$H(X, Y) = H(X) + H(Y).$$

  Why is that?

Conditional entropy

- Let's again consider random variables X, Y with joint mass function $p(x_i, y_j) = P\{X = x_i, Y = y_j\}$ and write

$$H(X, Y) = -\sum_i \sum_j p(x_i, y_j) \log p(x_i, y_i).$$

- But now let's not assume they are independent.
- We can define a **conditional entropy** of X *given* $Y = y_j$ by

$$H_{Y=y_j}(X) = -\sum_i p(x_i|y_j) \log p(x_i|y_j).$$

- This is just the entropy of the conditional distribution. Recall that $p(x_i|y_j) = P\{X = x_i | Y = y_j\}$.
- We similarly define $H_Y(X) = \sum_j H_{Y=y_j}(X) p_Y(y_j)$. This is the *expected* amount of conditional entropy that there will be in Y after we have observed X.

Properties of conditional entropy

- Definitions: $H_{Y=y_j}(X) = -\sum_i p(x_i|y_j)\log p(x_i|y_j)$ and $H_Y(X) = \sum_j H_{Y=y_j}(X)p_Y(y_j)$.
- **Important property one:** $H(X,Y) = H(Y) + H_Y(X)$.
- In words, the expected amount of information we learn when discovering (X,Y) is equal to expected amount we learn when discovering Y *plus* expected amount when we subsequently discover X (given our knowledge of Y).
- To prove this property, recall that $p(x_i,y_j) = p_Y(y_j)p(x_i|y_j)$.
- Thus, $H(X,Y) = -\sum_i\sum_j p(x_i,y_j)\log p(x_i,y_j) =$ $-\sum_i\sum_j p_Y(y_j)p(x_i|y_j)[\log p_Y(y_j) + \log p(x_i|y_j)] =$ $-\sum_j p_Y(y_j)\log p_Y(y_j)\sum_i p(x_i|y_j) -$ $\sum_j p_Y(y_j)\sum_i p(x_i|y_j)\log p(x_i|y_j) = H(Y) + H_Y(X)$.

Properties of conditional entropy

- Definitions: $H_{Y=y_j}(X) = -\sum_i p(x_i|y_j)\log p(x_i|y_j)$ and $H_Y(X) = \sum_j H_{Y=y_j}(X)p_Y(y_j)$.
- **Important property two:** $H_Y(X) \le H(X)$ with equality if and only if X and Y are independent.
- In words, the expected amount of information we learn when discovering X *after* having discovered Y can't be more than the expected amount of information we would learn when discovering X *before* knowing anything about Y.
- Proof: note that $\mathcal{E}(p_1,p_2,\ldots,p_n) := -\sum p_i\log p_i$ is concave.
- The vector $v = \{p_X(x_1), p_X(x_2),\ldots,p_X(x_n)\}$ is a weighted average of vectors $v_j := \{p_X(x_1|y_j), p_X(x_2|y_j),\ldots,p_X(x_n|y_j)\}$ as j ranges over possible values. By (vector version of) Jensen's inequality, $H(X) = \mathcal{E}(v) = \mathcal{E}(\sum p_Y(y_j)v_j) \ge \sum p_Y(y_j)\mathcal{E}(v_j) = H_Y(X)$.

18.600: Lecture 34

Martingales and the optional stopping theorem

Scott Sheffield

MIT

Outline

Outline

Martingale definition

- Let S be a probability space.
- Let $X_0, X_1, X_2, \ldots$ be a sequence of random variables. Informally, we will imagine that we acquiring information about S in a sequence of stages, and each X_j represents a quantity that is known to us at the jth stage.
- If Z is any random variable, we let $E[Z|\mathcal{F}_n]$ denote the conditional expectation of X given all the information that is available to us on the nth stage. If we don't specify otherwise, we assume that this information consists precisely of the values $X_0, X_1, \ldots, X_n$, so that $E[Z|\mathcal{F}_n] = E[Z|X_0, X_1, \ldots, X_n]$. (In some applications, one could imagine there are other things known as well at stage n.)
- We say X_n sequence is a **martingale** if $E[|X_n|] < \infty$ for all n and $E[X_{n+1}|\mathcal{F}_n] = X_n$ for all n.
- "Taking into account all the information I have at stage n, the expected value at stage $n+1$ is the value at stage n."

Martingale definition

- Example: Imagine that X_n is the price of a stock on day n.
- Martingale condition: "Expected value of stock tomorrow, given all I know today, is value of the stock today."
- Question: If you are given a mathematical description of a process $X_0, X_1, X_2, \dots$ then how can you check whether it is a martingale?
- Consider all of the information that you know after having seen $X_0, X_1, \dots, X_n$. Then try to figure out what additional (not yet known) randomness is involved in determining X_{n+1}. Use this to figure out the conditional expectation of X_{n+1}, and check to see whether this is necessarily equal to the known X_n value.

Martingale examples

- Suppose that $A_1, A_2, \dots$ are i.i.d. random variables each equal to -1 with probability .5 and 1 with probability .5.
- Let $X_0 = 0$ and $X_n = \sum_{i=1}^n A_i$ for $n > 0$. Is the X_n sequence a martingale?
- Answer: yes. To see this, note that $E[X_{n+1}|\mathcal{F}_n] = E[X_n + A_{n+1}|\mathcal{F}_n] = E[X_n|\mathcal{F}_n] + E[A_{n+1}|\mathcal{F}_n]$, by additivity of conditional expectation (given $\mathcal{F}_n$).
- Since X_n is known at stage n, we have $E[X_n|\mathcal{F}_n] = X_n$. Since we know nothing more about A_{n+1} at stage n than we originally knew, we have $E[A_{n+1}|\mathcal{F}_n] = 0$. Thus $E[X_{n+1}|\mathcal{F}_n] = X_n$.
- Informally, I'm just tossing a new fair coin at each stage to see if X_n goes up or down one step. If I know the information available up to stage n, and I know $X_n = 10$, then I see $X_{n+1} = 11$ and $X_{n+1} = 9$ as equally likely, so $E[X_{n+1}|\mathcal{F}_n] = 10 = X_n$.

Another martingale example

- What if each A_i is 1.01 with probability .5 and .99 with probability .5 and we write $X_0 = 1$ and $X_n = \prod_{i=1}^n A_i$ for $n > 0$? Then is X_n a martingale?
- Answer: yes. Note that $E[X_{n+1}|\mathcal{F}_n] = E[A_{n+1}X_n|\mathcal{F}_n]$. At stage n, the value X_n is known, and hence can be treated as a known constant, which can be factored out of the expectation, i.e., $E[A_{n+1}X_n|\mathcal{F}_n] = X_n E[A_{n+1}|\mathcal{F}_n]$.
- Since I know nothing new about A_{n+1} at stage n, we have $E[A_{n+1}|\mathcal{F}_n] = E[A_{n+1}] = 1$. Hence $E[A_{n+1}X_n|\mathcal{F}_n] = X_n$.
- Informally, I'm just tossing a new fair coin at each stage to see if X_n goes up or down by a percentage point of its current value. If I know all the information available up to stage n, and I know $X_n = 5$, then I see $X_{n+1} = 5.05$ and $X_{n+1} = 4.95$ as equally likely, so $E[X_{n+1}|\mathcal{F}_n] = 5$.
- **Two classic martingale examples:** sums of independent random variables (each with mean zero) and products of independent random variables (each with mean one).

Another example

- Suppose A is 1 with probability .5 and -1 with probability .5. Let $X_0 = 0$ and write $X_n = (-1)^n A$ for all $n > 0$.
- What is $E[X_n]$, as a function of n?
- $E[X_n] = 0$ for all n.
- Does this mean that X_n is a martingale?
- No. If $n \geq 1$, then given the information available up to stage n, I can figure out what A must be, and can hence deduce exactly what X_{n+1} will be — and it is not the same as X_n. In particular, $E[X_{n+1}|\mathcal{F}_n] = -X_n \neq X_n$.
- Informally, X_n alternates between 1 and -1. Each time it goes up and hits 1, I know it will go back down to -1 on the next step.

Stopping time definition

- Let T be a non-negative integer valued random variable.
- Think of T as giving the time the asset will be sold if the price sequence is $X_0, X_1, X_2, \ldots$.
- Say that T is a **stopping time** if the event that $T = n$ depends only on the values X_i for $i \leq n$. In other words, the decision to sell at time n depends only on prices up to time n, not on (as yet unknown) future prices.

Stopping time examples

- Let $A_1, \ldots$ be i.i.d. random variables equal to -1 with probability .5 and 1 with probability .5 and let $X_0 = 0$ and $X_n = \sum_{i=1}^n A_i$ for $n \geq 0$.
- Which of the following is a stopping time?
  1. The smallest T for which $|X_T| = 50$
  2. The smallest T for which $X_T \in \{-10, 100\}$
  3. The smallest T for which $X_T = 0$.
  4. The T at which the X_n sequence achieves the value 17 for the 9th time.
  5. The value of $T \in \{0, 1, 2, \ldots, 100\}$ for which X_T is largest.
  6. The largest $T \in \{0, 1, 2, \ldots, 100\}$ for which $X_T = 0$.
- Answer: first four, not last two.

Outline

Outline

Optional stopping overview

- **Doob's optional stopping time theorem** is contained in many basic texts on probability and Martingales. (See, for example, Theorem 10.10 of *Probability with Martingales*, by David Williams, 1991.)
- Essentially says that you can't make money (in expectation) by buying and selling an asset whose price is a martingale.
- Precisely, if you buy the asset at some time and adopt any strategy at all for deciding when to sell it, then the expected price at the time you sell is the price you originally paid.
- If market price is a martingale, you cannot make money in expectation by "timing the market."

Doob's Optional Stopping Theorem: statement

- **Doob's Optional Stopping Theorem:** If the sequence $X_0, X_1, X_2, \ldots$ is a **bounded** martingale, and T is a stopping time, then the expected value of X_T is X_0.
- When we say martingale is bounded, we mean that for some C, we have that with probability one $|X_i| < C$ for all i.
- Why is this assumption necessary?
- Can we give a counterexample if boundedness is not assumed?
- Theorem can be proved by induction if *stopping time* T is bounded. Unbounded T requires a limit argument. (This is where boundedness of martingale is used.)

Martingales applied to finance

- Many asset prices are believed to behave approximately like martingales, at least in the short term.
- **Efficient market hypothesis**: new information is instantly absorbed into the stock value, so expected value of the stock tomorrow should be the value today. (If it were higher, statistical arbitrageurs would bid up today's price until this was not the case.)
- But what about interest, risk premium, etc.?
- According to the **fundamental theorem of asset pricing**, the discounted price $\frac{X(n)}{A(n)}$, where A is a risk-free asset, is a martingale with respected to **risk neutral probability**. More on this next lecture.

Martingales as successively revised best guesses

- The two-element sequence $E[X], X$ is a martingale.
- In previous lectures, we interpreted the conditional expectation $E[X|Y]$ as a random variable.
- Depends only on Y. Describes expectation of X given observed Y value.
- We showed $E[E[X|Y]] = E[X]$.
- This means that the three-element sequence $E[X], E[X|Y], X$ is a martingale.
- More generally if Y_i are any random variables, the sequence $E[X], E[X|Y_1], E[X|Y_1, Y_2], E[X|Y_1, Y_2, Y_3], \ldots$ is a martingale.

Martingales as real-time subjective probability updates

- Ivan sees email from girlfriend with subject "some possibly serious news", thinks there's a 20 percent chance she'll break up with him by email's end. Revises number after each line:
- Oh Ivan, I've missed you so much! 12
- I have something crazy to tell you, 24
- and so sorry to do this by email. (Where's your phone!?) 38
- I've been spending lots of time with a guy named Robert, 52
- a visiting database consultant on my project 34
- who seems very impressed by my work. 23
- Robert wants me to join his startup in Palo Alto. 38
- Exciting!!! Of course I said I'd have to talk to you first, 24
- because you are absolutely my top priority in my life, 8
- and you're stuck at MIT for at least three more years... 11
- but honestly, I'm just so confused on so many levels. 15
- Call me!!! I love you! Alice 0

More conditional probability martingale examples

- Example: let C be the amount of oil available for drilling under a particular piece of land. Suppose that ten geological tests are done that will ultimately determine the value of C. Let C_n be the **conditional expectation** of C *given* the outcome of the first n of these tests. Then the sequence $C_0, C_1, C_2, \ldots, C_{10} = C$ is a martingale.
- Let A_i be my best guess at the probability that a basketball team will win the game, given the outcome of the first i minutes of the game. Then (assuming some "rationality" of my personal probabilities) A_i is a martingale.

18.600: Lecture 35

Martingales and risk neutral probability

Scott Sheffield

MIT

Outline

Martingales and stopping times

Risk neutral probability and martingales

Outline

Martingales and stopping times

Risk neutral probability and martingales

Recall martingale definition

- Let S be the probability space. Let $X_0, X_1, X_2, \dots$ be a sequence of real random variables. Interpret X_i as price of asset at ith time step.
- Say X_n sequence is a **martingale** if $E[|X_n|] < \infty$ for all n and $E[X_n|\mathcal{F}_n] := E[X_{n+1}|X_0, X_1, X_2, \dots, X_n] = X_n$ for all n.
- "Given all I know today, expected price tomorrow is the price today."
- If you are given a mathematical description of a process $X_0, X_1, X_2, \dots$ then how can you check whether it is a martingale?
- Consider all of the information that you know after having seen $X_0, X_1, \dots, X_n$. Then try to figure out what additional (not yet known) randomness is involved in determining X_{n+1}. Use this to figure out the conditional expectation of X_{n+1}, and check to see whether this is always equal to the known X_n value.

Recall stopping time definition

- Let T be a non-negative integer valued random variable.
- Think of T as giving the time the asset will be sold if the price sequence is $X_0, X_1, X_2, \ldots$.
- Say that T is a **stopping time** if the event that $T = n$ depends only on the values X_i for $i \leq n$. In other words, the decision to sell at time n depends only on prices up to time n, not on (as yet unknown) future prices.

Examples

- Suppose that an asset price is a martingale that starts at 50 and changes by increments of ± 1 at each time step. What is the probability that the price goes down to 40 before it goes up to 70?
- What is the probability that it goes down to 45 then up to 55 then down to 45 then up to 55 again — all before reaching either 0 or 100?

Outline

Outline

Martingales applied to finance

- Many asset prices are believed to behave approximately like martingales, at least in the short term.
- **Efficient market hypothesis**: new information is instantly absorbed into the stock value, so expected value of the stock tomorrow should be the value today. (If it were higher, statistical arbitrageurs would bid up today's price until this was not the case.)
- But there are some caveats: interest, risk premium, etc.
- According to the **fundamental theorem of asset pricing**, the discounted price $\frac{X(n)}{A(n)}$, where A is a risk-free asset, is a martingale with respected to **risk neutral probability**.

Risk neutral probability

- "Risk neutral probability" is a fancy term for "market probability". (The term "market probability" is arguably more descriptive.)
- That is, it is a probability measure that you can deduce by looking at prices on market.
- For example, suppose somebody is about to shoot a free throw in basketball. What is the price in the sports betting world of a contract that pays one dollar if the shot is made?
- If the answer is .75 dollars, then we say that the risk neutral probability that the shot will be made is .75.
- Risk neutral probability is the probability determined by the market betting odds.

Risk neutral probability of outcomes known at fixed time T

- **Risk neutral probability of event A:** $P_{RN}(A)$ denotes

$$\frac{\text{Price}\{\text{Contract paying 1 dollar at time } T \text{ if } A \text{ occurs }\}}{\text{Price}\{\text{Contract paying 1 dollar at time } T \text{ no matter what }\}}.$$

- If risk-free interest rate is constant and equal to r (compounded continuously), then denominator is e^{-rT}.
- Assuming no **arbitrage** (i.e., no risk free profit with zero upfront investment), P_{RN} satisfies axioms of probability. That is, $0 \leq P_{RN}(A) \leq 1$, and $P_{RN}(S) = 1$, and if events A_j are disjoint then $P_{RN}(A_1 \cup A_2 \cup \ldots) = P_{RN}(A_1) + P_{RN}(A_2) + \ldots$
- **Arbitrage example:** if A and B are disjoint and $P_{RN}(A \cup B) < P(A) + P(B)$ then we sell contracts paying 1 if A occurs and 1 if B occurs, buy contract paying 1 if $A \cup B$ occurs, pocket difference.

Risk neutral probability differ vs. "ordinary probability"

- At first sight, one might think that $P_{RN}(A)$ describes the market's best guess at the probability that A will occur.
- But suppose A is the event that the government is dissolved and all dollars become worthless. What is $P_{RN}(A)$?
- Should be 0. Even if people think A is *likely*, a contract paying a dollar when A occurs is worthless.
- Now, suppose there are only 2 outcomes: A is event that economy booms and everyone prospers and B is event that economy sags and everyone is needy. Suppose purchasing power of dollar is the same in both scenarios. If people think A has a .5 chance to occur, do we expect $P_{RN}(A) > .5$ or $P_{RN}(A) < .5$?
- Answer: $P_{RN}(A) < .5$. People are risk averse. In second scenario they need the money more.f

Non-systemic event

- Suppose that A is the event that the Boston Red Sox win the World Series. Would we expect $P_{RN}(A)$ to represent (the market's best assessment of) the probability that the Red Sox will win?
- Arguably yes. The amount that *people in general* need or value dollars does not depend much on whether A occurs (even though the financial needs of specific individuals may depend on heavily on A).
- Even if some people bet based on loyalty, emotion, insurance against personal financial exposure to team's prospects, etc., there will arguably be enough in-it-for-the-money statistical arbitrageurs to keep price near a reasonable guess of what well-informed informed experts would consider the true probability.

Extensions of risk neutral probability

- Definition of risk neutral probability depends on choice of currency (the so-called *numéraire*).
- In 2016 presidential election, investors predicted value of Mexican peso (in US dollars) would be lower
- Risk neutral probability can be defined for variable times and variable interest rates — e.g., one can take the numéraire to be amount one dollar in a variable-interest-rate money market account has grown to when outcome is known. Can define $P_{RN}(A)$ to be price of contract paying this amount if and when A occurs.
- For simplicity, we focus on fixed time T, fixed interest rate r in this lecture.

Risk neutral probability is objective

- Check out binary prediction contracts at predictwise.com, oddschecker.com, predictit.com, etc.
- Many financial derivatives are essentially bets of this form.
- Unlike "true probability" (what does that mean?) the "risk neutral probability" is an objectively measurable price.
- Pundit: The market predictions are ridiculous. I can estimate probabilities much better than they can.
- Listener: Then why not make some bets and get rich? If your estimates are so much better, law of large numbers says you'll surely come out way ahead eventually.
- Pundit: Well, you know... been busy... scruples about gambling... more to life than money...
- Listener: Yeah, that's what I thought.

Prices as expectations

- By assumption, the price of a contract that pays one dollar at time T if A occurs is $P_{RN}(A)e^{-rT}$.
- If A and B are disjoint, what is the price of a contract that pays 2 dollars if A occurs, 3 if B occurs, 0 otherwise?
- Answer: $(2P_{RN}(A) + 3P_{RN}(B))e^{-rT}$.
- Generally, in absence of arbitrage, price of contract that pays X at time T should be $E_{RN}(X)e^{-rT}$ where E_{RN} denotes expectation with respect to the risk neutral probability.
- Example: if a non-divided paying stock will be worth X at time T, then its price today should be $E_{RN}(X)e^{-rT}$.
- So-called **fundamental theorem of asset pricing** states that (assuming no arbitrage) interest-discounted asset prices are martingales with respect to risk neutral probability. Current price of stock being $E_{RN}(X)e^{-rT}$ follows from this.

18.600: Lecture 36

Risk Neutral Probability and Black-Scholes

Scott Sheffield

MIT

Outline

Black-Scholes

Call quotes and risk neutral probability

Outline

Black-Scholes

Call quotes and risk neutral probability

Overview

- The mathematics of today's lecture will not go far beyond things we know.
- Main mathematical tasks will be to compute expectations of functions of log-normal random variables (to get the Black-Scholes formula) and differentiate under an integral (to compute risk neutral density functions from option prices).
- Will spend time giving financial *interpretations* of the math.
- Can interpret this lecture as a sophisticated story problem, illustrating an important application of the probability we have learned in this course (involving probability axioms, expectations, cumulative distribution functions, etc.)
- Brownian motion (as mathematically constructed by MIT professor Norbert Wiener) is a *continuous time martingale*.
- Black-Scholes theory assumes that the log of an asset price is a process called *Brownian motion with drift* with respect to *risk neutral probability*. Implies option price formula.

Black-Scholes: main assumption and conclusion

- More famous MIT professors: Black, Scholes, Merton.
- 1997 Nobel Prize.
- **Assumption:** the log of an asset price X at fixed future time T is a normal random variable (call it N) with some known variance (call it $T\sigma^2$) and some mean (call it μ) with respect to risk neutral probability.
- **Observation:** N normal $(\mu, T\sigma^2)$ implies $E[e^N] = e^{\mu + T\sigma^2/2}$.
- **Observation:** If X_0 is the current price then $X_0 = E_{RN}[X]e^{-rT} = E_{RN}[e^N]e^{-rT} = e^{\mu+(\sigma^2/2-r)T}$.
- **Observation:** This implies $\mu = \log X_0 + (r - \sigma^2/2)T$.
- **Conclusion:** If g is any function then the price of a contract that pays $g(X)$ at time T is

$$E_{RN}[g(X)]e^{-rT} = E_{RN}[g(e^N)]e^{-rT}$$

where N is normal with mean μ and variance $T\sigma^2$.

Black-Scholes example: European call option

- A **European call option** on a stock at **maturity date** T, **strike price** K, gives the holder the right (but not obligation) to purchase a share of stock for K dollars at time T.

> The document gives the bearer the right to purchase one share of MSFT from me on May 31 for 35 dollars. SS

- If X is the value of the stock at T, then the value of the option at time T is given by $g(X) = \max\{0, X - K\}$.
- **Black-Scholes:** price of contract paying $g(X)$ at time T is $E_{RN}[g(X)]e^{-rT} = E_{RN}[g(e^N)]e^{-rT}$ where N is normal with variance $T\sigma^2$, mean $\mu = \log X_0 + (r - \sigma^2/2)T$.
- Write this as

$$\begin{aligned} e^{-rT}E_{RN}[\max\{0, e^N - K\}] &= e^{-rT}E_{RN}[(e^N - K)1_{N \geq \log K}] \\ &= \frac{e^{-rT}}{\sigma\sqrt{2\pi T}} \int_{\log K}^{\infty} e^{-\frac{(x-\mu)^2}{2T\sigma^2}} (e^x - K)dx. \end{aligned}$$

The famous formula

- Let T be time to maturity, X_0 current price of underlying asset, K strike price, r risk free interest rate, σ the volatility.
- We need to compute $e^{-rT}\int_{\log K}^{\infty} e^{-\frac{(x-\mu)^2}{2T\sigma^2}}(e^x - K)dx$ where $\mu = rT + \log X_0 - T\sigma^2/2$.
- Can use complete-the-square tricks to compute the two terms explicitly in terms of standard normal cumulative distribution function Φ.
- Price of European call is $\Phi(d_1)X_0 - \Phi(d_2)Ke^{-rT}$ where $d_1 = \frac{\ln(\frac{X_0}{K}) + (r + \frac{\sigma^2}{2})(T)}{\sigma\sqrt{T}}$ and $d_2 = \frac{\ln(\frac{X_0}{K}) + (r - \frac{\sigma^2}{2})(T)}{\sigma\sqrt{T}}$.

Outline

Outline

Black-Scholes

Call quotes and risk neutral probability

Determining risk neutral probability from call quotes

- If $C(K)$ is price of European call with strike price K and $f = f_X$ is risk neutral probability density function for X at time T, then $C(K) = e^{-rT} \int_{-\infty}^{\infty} f(x) \max\{0, x - K\} dx$.
- Differentiating under the integral, we find that

$$e^{rT} C'(K) = \int f(x)(-1_{x>K})dx = -P_{RN}\{X > K\} = F_X(K) - 1,$$

$$e^{rT} C''(K) = f(K).$$

- We can look up $C(K)$ for a given stock symbol (say GOOG) and expiration time T at cboe.com and work out approximately what F_X and hence f_X must be.

Perspective: implied volatility

- Risk neutral probability densities derived from call quotes are not quite lognormal in practice. Tails are too fat. Main Black-Scholes assumption is only approximately correct.
- "Implied volatility" is the value of σ that (when plugged into Black-Scholes formula along with known parameters) predicts the current market price.
- If Black-Scholes were completely correct, then given a stock and an expiration date, the implied volatility would be the same for all strike prices. In practice, when the implied volatility is viewed as a function of strike price (sometimes called the "volatility smile"), it is not constant.

Perspective: why is Black-Scholes not exactly right?

- **Main Black-Scholes assumption:** risk neutral probability densities are lognormal.
- **Heuristic support for this assumption:** If price goes up 1 percent or down 1 percent each day (with no interest) then the risk neutral probability must be .5 for each (independently of previous days). Central limit theorem gives log normality for large T.
- **Replicating portfolio point of view:** in the simple binary tree models (or continuum Brownian models), we can transfer money back and forth between the stock and the risk free asset to ensure our wealth at time T equals the option payout. Option price is required initial investment, which is risk neutral expectation of payout. "True probabilities" are irrelevant.
- **Where arguments for assumption break down:** Fluctuation sizes vary from day to day. Prices can have big jumps.
- **Fixes:** variable volatility, random interest rates, Lévy jumps....

18.600: Lecture 37

Review: practice problems

Scott Sheffield

MIT

Expectation and variance

- Eight athletic teams are ranked 1 through 8 after season one, and ranked 1 through 8 again after season two. Assume that each set of rankings is chosen uniformly from the set of 8! possible rankings and that the two rankings are independent. Let N be the number of teams whose rank does not change from season one to season two. Let N_+ the number of teams whose rank improves by exactly two spots. Let N_- be the number whose rank declines by exactly two spots. Compute the following:
  - $E[N]$, $E[N_+]$, and $E[N_-]$
  - $\mathrm{Var}[N]$
  - $\mathrm{Var}[N_+]$

Expectation and variance — answers

- Let N_i be 1 if team ranked ith first season remains ith second seasons. Then $E[N] = E[\sum_{i=1}^{8} N_i] = 8 \cdot \frac{1}{8} = 1$. Similarly, $E[N_+] = E[N_-] = 6 \cdot \frac{1}{8} = 3/4$
- $\mathrm{Var}[N] = E[N^2] - E[N]^2$ and $E[N^2] = E[\sum_{i=1}^{8} \sum_{j=1}^{8} N_i N_j] = 8 \cdot \frac{1}{8} + 56 \cdot \frac{1}{56} = 2$.
- N_+^i be 1 if team ranked ith has rank improve to $(i-2)$th for second seasons. Then $E[(N_+)^2] = E[\sum_{3=1}^{8} \sum_{3=1}^{8} N_+^i N_+^j] = 6 \cdot \frac{1}{8} + 30 \cdot \frac{1}{56} = 9/7$, so $\mathrm{Var}[N_+] = 9/7 - (3/4)^2$.

Conditional distributions

- Roll ten dice. Find the conditional probability that there are exactly 4 ones, given that there are exactly 4 sixes.

Conditional distributions — answers

- Straightforward approach: $P(A|B) = P(AB)/P(B)$.
- Numerator: is $\frac{\binom{10}{4}\binom{6}{4}4^2}{6^{10}}$. Denominator is $\frac{\binom{10}{4}5^6}{6^{10}}$.
- Ratio is $\binom{6}{4}4^2/5^6 = \binom{6}{4}(\frac{1}{5})^4(\frac{4}{5})^2$.
- Alternate solution: first condition on location of the 6's and then use binomial theorem.

Poisson point processes

- Suppose that in a certain town earthquakes are a Poisson point process, with an average of one per decade, and volcano eruptions are an independent Poisson point process, with an average of two per decade. The V be length of time (in decades) until the first volcano eruption and E the length of time (in decades) until the first earthquake. Compute the following:
  - $\mathbb{E}[E^2]$ and $\mathrm{Cov}[E, V]$.
  - The expected number of calendar years, in the next decade (ten calendar years), that have no earthquakes and no volcano eruptions.
  - The probability density function of $\min\{E, V\}$.

Poisson point processes — answers

- $E[E^2] = 2$ and $\mathrm{Cov}[E, V] = 0$.
- Probability of no earthquake or eruption in first year is $e^{-(2+1)\frac{1}{10}} = e^{-.3}$ (see next part). Same for any year by memoryless property. Expected number of quake/eruption-free years is $10e^{-.3} \approx 7.4$.
- Probability density function of $\min\{E, V\}$ is $3e^{-(2+1)x}$ for $x \geq 0$, and 0 for $x < 0$.

18.600: Lecture 38

Review: practice problems

Scott Sheffield

MIT

Order statistics

- Let X be a uniformly distributed random variable on $[-1,1]$.
  - Compute the variance of X^2.
  - If $X_1,\ldots,X_n$ are independent copies of X, what is the probability density function for the smallest of the X_i

Order statistics — answers

- $$\mathrm{Var}[X^2] = E[X^4] - (E[X^2])^2$$
  $$= \int_{-1}^{1} \frac{1}{2}x^4 dx - (\int_{-1}^{1} \frac{1}{2}x^2 dx)^2 = \frac{1}{5} - \frac{1}{9} = \frac{4}{45}.$$
- Note that for $x \in [-1,1]$ we have
  $$P\{X > x\} = \int_x^1 \frac{1}{2} dx = \frac{1-x}{2}.$$
  If $x \in [-1,1]$, then
  $$P\{\min\{X_1,\ldots,X_n\} > x\}$$
  $$= P\{X_1 > x, X_2 > x, \ldots, X_n > x\} = (\frac{1-x}{2})^n.$$
  So the density function is
  $$-\frac{\partial}{\partial x}(\frac{1-x}{2})^n = \frac{n}{2}(\frac{1-x}{2})^{n-1}.$$

Moment generating functions

- Suppose that X_i are independent copies of a random variable X. Let $M_X(t)$ be the moment generating function for X. Compute the moment generating function for the average $\sum_{i=1}^n X_i/n$ in terms of $M_X(t)$ and n.

Moment generating functions — answers

- Write $Y = \sum_{i=1}^{n} X_i/n$. Then

$$M_Y(t) = E[e^{tY}] = E[e^{t\sum_{i=1}^{n} X_i/n}] = (M_X(t/n))^n.$$

Entropy

- Suppose X and Y are independent random variables, each equal to 1 with probability $1/3$ and equal to 2 with probability $2/3$.
  - Compute the entropy $H(X)$.
  - Compute $H(X+Y)$.
  - Which is larger, $H(X+Y)$ or $H(X,Y)$? Would the answer to this question be the same for any discrete random variables X and Y? Explain.

Entropy — answers

- $H(X) = \frac{1}{3}(-\log\frac{1}{3}) + \frac{2}{3}(-\log\frac{2}{3})$.
- $H(X+Y) = \frac{1}{9}(-\log\frac{1}{9}) + \frac{4}{9}(-\log\frac{4}{9}) + \frac{4}{9}(-\log\frac{4}{9})$
- $H(X,Y)$ is larger, and we have $H(X,Y) \geq H(X+Y)$ for any X and Y. To see why, write $a(x,y) = P\{X=x, Y=y\}$ and $b(x,y) = P\{X+Y=x+y\}$. Then $a(x,y) \leq b(x,y)$ for any x and y, so
  $H(X,Y) = E[-\log a(x,y)] \geq E[-\log b(x,y)] = H(X+Y)$.

18.600: Lecture 39

Review: practice problems

Scott Sheffield

MIT

Markov chains

- Alice and Bob share a home with a bathroom, a walk-in closet, and 2 towels.
- Each morning a fair coin decide which of the two showers first.
- After Bob showers, if there is at least one towel in the bathroom, Bob uses the towel and leaves it draped over a chair in the walk-in closet. If there is no towel in the bathroom, Bob grumpily goes to the walk-in closet, dries off there, and leaves the towel in the walk-in closet
- When Alice showers, she first checks to see if at least one towel is present. If a towel is present, she dries off with that towel and returns it to the bathroom towel rack. Otherwise, she cheerfully retrieves both towels from the walk-in closet, then showers, dries off and leaves both towels on the rack.
- **Problem:** describe towel-distribution evolution as a Markov chain and determine (over the long term) on what fraction of days Bob emerges from the shower to find no towel.

Markov chains — answers

- Let state $0, 1, 2$ denote bathroom towel number.
- Shower state change Bob: $2 \to 1$, $1 \to 0$, $0 \to 0$.
- Shower state change Alice: $2 \to 2$, $1 \to 1$, $0 \to 2$.
- Morning state change AB: $2 \to 1$, $1 \to 0$, $0 \to 1$.
- Morning state change BA: $2 \to 1$, $1 \to 2$, $0 \to 2$.
- Markov chain matrix:

$$M = \begin{pmatrix} 0 & .5 & .5 \\ .5 & 0 & .5 \\ 0 & 1 & 0 \end{pmatrix}$$

- Row vector π such that $\pi M = \pi$ (with components of π summing to one) is $\begin{pmatrix} \frac{2}{9} & \frac{4}{9} & \frac{1}{3} \end{pmatrix}$.
- Bob finds no towel only if morning starts in state zero and Bob goes first. Over long term Bob finds no towel $\frac{2}{9} \times \frac{1}{2} = \frac{1}{9}$ fraction of the time.

Optional stopping, martingales, central limit theorem

Suppose that $X_1, X_2, X_3, \ldots$ is an infinite sequence of independent random variables which are each equal to 1 with probability $1/2$ and -1 with probability $1/2$. Let $Y_n = \sum_{i=1}^n X_i$. Answer the following:

- What is the the probability that Y_n reaches -25 before the first time that it reaches 5?
- Use the central limit theorem to approximate the probability that $Y_{9000000}$ is greater than 6000.

Optional stopping, martingales, central limit theorem — answers

- $p_{-25}25 + p_5 5 = 0$ and $p_{-25} + p_5 = 1$. Solving, we obtain $p_{-25} = 1/6$ and $p_5 = 5/6$.
- One standard deviation is $\sqrt{9000000} = 3000$. We want probability to be 2 standard deviations above mean. Should be about $\int_2^\infty \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx$.

Martingales

- Let X_i be independent random variables with mean zero. In which of the cases below is the sequence Y_i necessarily a martingale?
  - $Y_n = \sum_{i=1}^n iX_i$
  - $Y_n = \sum_{i=1}^n X_i^2 - n$
  - $Y_n = \prod_{i=1}^n (1 + X_i)$
  - $Y_n = \prod_{i=1}^n (X_i - 1)$

Martingales

- Yes, no, yes, no.

Calculations like those needed for Black-Scholes derivation

- Let X be a normal random variable with mean 0 and variance 1. Compute the following (you may use the function $\Phi(a) := \int_{-\infty}^a \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx$ in your answers):
  - $E[e^{3X-3}]$.
  - $E[e^X 1_{X \in (a,b)}]$ for fixed constants $a < b$.

Calculations like those needed for Black-Scholes derivation – answers

$$\begin{aligned} E[e^{3X-3}] &= \int_{-\infty}^{\infty} e^{3x-3}\frac{1}{\sqrt{2\pi}}e^{-x^2/2}dx \\ &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}}e^{-\frac{x^2-6x+6}{2}}dx \\ &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}}e^{-\frac{x^2-6x+9}{2}}e^{3/2}dx \\ &= e^{3/2}\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}}e^{-\frac{(x-3)^2}{2}}dx \\ &= e^{3/2} \end{aligned}$$

Calculations like those needed for Black-Scholes derivation – answers

$$\begin{aligned} E[e^X 1_{X\in(a,b)}] &= \int_a^b e^x\frac{1}{\sqrt{2\pi}}e^{-x^2/2}dx \\ &= \int_a^b e^x\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}dx \\ &= \int_a^b \frac{1}{\sqrt{2\pi}}e^{-\frac{x^2-2x+1-1}{2}}dx \\ &= e^{1/2}\int_a^b \frac{1}{\sqrt{2\pi}}e^{-\frac{(x-1)^2}{2}}dx \\ &= e^{1/2}\int_{a-1}^{b-1} \frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}dx \\ &= e^{1/2}(\Phi(b-1)-\Phi(a-1)) \end{aligned}$$

If you want *more* probability and statistics...

- **UNDERGRADUATE:**
  (a) 18.615 Introduction to Stochastic Processes
  (b) 18.642 Topics in Math with Applications in Finance
  (c) 18.650 Statistics for Applications
- **GRADUATE LEVEL PROBABILITY**
  (a) 18.175 Theory of Probability
  (b) 18.176 Stochastic calculus
  (c) 18.177 Topics in stochastic processes (topics vary — repeatable, offered twice next year)
- **GRADUATE LEVEL STATISTICS**
  (a) 18.655 Mathematical statistics
  (b) 18.657 Topics in statistics (topics vary — topic this year was machine learning; repeatable)
- **OUTSIDE OF MATH DEPARTMENT**
  (a) Look up new MIT minor in statistics and data sciences.
  (b) Look up *long* list of probability/statistics courses (about 78 total) at `https://stat.mit.edu/academics/subjects/`
  (c) Ask other MIT faculty how they use probability and statistics in their research.

Thanks for taking the course!

- Considering previous generations of mathematically inclined MIT students, and adopting a frequentist point of view...
- You will probably do some important things with your lives.
- I hope your probabilistic shrewdness serves you well.
- Thinking more short term...
- Happy exam day!
- And may the odds be ever in your favor.