Statistics 110—Intro to Probability

Lectures by Joe Blitzstein
Notes by Max Wang

Harvard University, Fall 2011

Introduction

Statistics 110 is an introductory statistics course offered at Harvard University. It covers all the basics of probability—counting principles, probabilistic events, random variables, distributions, conditional probability, expectation, and Bayesian inference. The last few lectures of the course are spent on Markov chains.

These notes were partially live-TEXed—the rest were TEXed from course videos—then edited for correctness and clarity. I am responsible for all errata in this document, mathematical or otherwise; any merits of the material here should be credited to the lecturer, not to me.

Feel free to email me at mxawng@gmail.com with any comments.

Acknowledgments

In addition to the course staff, acknowledgment goes to Zev Chonoles, whose online lecture notes (`http://math.uchicago.edu/~chonoles/expository-notes/`) inspired me to post my own. I have also borrowed his format for this introduction page.

The page layout for these notes is based on the layout I used back when I took notes by hand. The LATEX styles can be found here: `https://github.com/mxw/latex-custom`.

Copyright

Copyright © 2011 Max Wang.

This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License. This means you are free to edit, adapt, transform, or redistribute this work as long as you

- include an attribution of Joe Blitzstein as the instructor of the course these notes are based on, and an attribution of Max Wang as the note-taker;
- do so in a way that does not suggest that either of us endorses you or your use of this work;
- use this work for noncommercial purposes only; and
- if you adapt or build upon this work, apply this same license to your contributions.

See `http://creativecommons.org/licenses/by-nc-sa/4.0/` for the license details.

Lecture 2 — 9/2/11

Definition 2.1. A sample space S is the set of all possible outcomes of an experiment.

Definition 2.2. An event $A \subseteq S$ is a subset of a sample space.

Definition 2.3. Assuming that all outcomes are equally likely and that the sample space is finite,

$$P(A) = \frac{\#\text{ favorable outcomes}}{\#\text{ possible outcomes}}$$

is the probability that A occurs.

Proposition 2.4 (Multiplication Rule)**.** *If there are r experiments and each experiment has n_i possible outcomes, then the overall sample space has size*

$$n_1 n_2 \cdots n_r$$

Example. The probability of a full house in a five card poker hand (without replacement, and without other players) is

$$P(\text{full house}) = \frac{13\binom{4}{3} \cdot 12\binom{4}{2}}{\binom{52}{5}}$$

Definition 2.5. The binomial coefficient is given by

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

or 0 if $k > n$.

Theorem 2.6 (Sampling Table)**.** *The number of subsets of size k chosen from a set of n distinct elements is given by the following table:*

	ordered	unordered	
replacement	n^k	$\binom{n+k-1}{k}$	
no replacement	$\frac{n!}{(n-k)!}$	$\binom{n}{k}$	

Lecture 3 — 9/7/11

Proposition 3.1. *The number of ways to choose k elements from a set of order n, with replacement and where order doesn't matter, is*

$$\binom{n+k-1}{k}$$

Proof. This count is equivalent to the number of ways to put k indistinguishable particles in n distinguishable boxes. Suppose we order the particles; then this count is simply the number of ways to place "dividers" between the particles, e.g.,

$$\bullet\bullet\bullet|\,\bullet\,|\,\bullet\bullet||\,\bullet\,|\bullet$$

There are

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

ways to place the particles, which determines the placement of the divisiors (or vice versa); this is our result. ■

Example.

1. $\binom{n}{k} = \binom{n}{n-k}$
2. $n\binom{n-1}{k-1} = k\binom{n}{k}$

   Pick k people out of n, then designate one as special. The RHS represents how many ways we can do this by first picking the k individuals and then making our designation. On the LHS, we see the number of ways to pick a special individual and then pick the remaining $k-1$ individuals from the remaining pool of $n-1$.
3. (Vandermonde)
   $\binom{n+m}{k} = \sum_{i=0}^{k} \binom{n}{i}\binom{m}{k-i}$

   On the LHS, we choose k people out of $n+m$. On the RHS, we sum up, for every i, how to choose i from the n people and $k-i$ from the m people.

Definition 3.2. A probability space consists of a sample space S along with a function $P : \mathcal{P}(S) \to [0,1]$ taking events to real numbers, where

1. $P(\emptyset) = 0$, $P(S) = 1$
2. $P(\bigcup_{n=1}^{\infty} A_n) = \sum_{n=1}^{\infty} P(A_n)$ if the A_n are disjoint

Lecture 4 — 9/9/11

Example (Birthday Problem)**.** The probability that at least two people among a group of k share the same birthday, assuming that birthdays are evenly distributed across the 365 standard days, is given by

$$\begin{aligned} P(\text{match}) &= 1 - P(\text{no match}) \\ &= 1 - \frac{365 \cdot 364 \cdots (365-k+1)}{365^k} \end{aligned}$$

Proposition 4.1.

1. *$P(A^C) = 1 - P(A)$.*
2. *If $A \subseteq B$, then $P(A) \leq P(B)$.*

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

Proof. All immediate. ■

Corollary 4.2 (Inclusion-Exclusion)**.** *Generalizing 3 above,*

$$P\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} P(A_i) - \sum_{i<j} P(A_i \cap A_j) + \sum_{i<j<k} P(A_i \cap A_j \cap A_k) - \cdots + (-1)^{n+1} P\left(\bigcap_{i=1}^{n} A_i\right)$$

Example (deMontmort's Problem)**.** Suppose we have n cards labeled $1, \ldots, n$. We want to determine the probability that for some card in a shuffled deck of such cards, the ith card has value i. Since the number of orderings of the deck for which a given set of matches occurs is simply the permutations on the remaining cards, we have

$$P(A_i) = \frac{(n-1)!}{n!} = \frac{1}{n}$$
$$P(A_1 \cap A_2) = \frac{(n-2)!}{n!} = \frac{1}{n(n-1)}$$
$$P(A_1 \cap \cdots \cap A_k) = \frac{(n-k)!}{n!}$$

So using the above corollary,

$$\begin{aligned} P(A_1 \cup \cdots \cup A_n) &= n \cdot \frac{1}{n} - \frac{n(n-1)}{2!} \cdot \frac{1}{n(n-1)} \\ &\quad + \frac{n(n-1)(n-2)}{3!} \cdot \frac{1}{n(n-1)(n-2)} \\ &= 1 - \frac{1}{2!} + \frac{1}{3!} - \cdots + (-1)^n \frac{1}{n!} \\ &\approx 1 - \frac{1}{e} \end{aligned}$$

Lecture 5 — 9/12/11

Note. Translation from English to inclusion-exclusion:

- Probability that at least one of the A_i occurs:
$$P(A_1 \cup \cdots \cup A_n)$$
- Probability that none of the A_i occurs:
$$1 - P(A_1 \cup \cdots \cup A_n)$$
- Probability that all of the A_i occur:
$$P(A_1 \cap \cdots \cap A_n) = 1 - P(A_1^C \cup \cdots \cup A_n^C)$$

Definition 5.1. The probability of two events A and B are independent if

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

In general, for n events $A_1, \ldots, A_n$, independence requires i-wise independence for every $i = 2, \ldots, n$; that is, say, pairwise independence alone does not imply independence.

Note. We will write $P(A \cap B)$ as $P(A, B)$.

Example (Newton-Pepys Problem)**.** Suppose we have some fair dice; we want to determine which of the following is most likely to occur:

1. At least one 6 given 6 dice.
2. At least two 6's with 12 dice.
3. At least three 6's with 18 dice.

For the first case, we have

$$P(A) = 1 - \left(\frac{-1}{5}\right)^6$$

For the second,

$$P(B) = 1 - \left(\frac{-1}{5}\right)^{12} - 12\left(\frac{-1}{1}\right)\left(\frac{-1}{5}\right)^{11}$$

and for the third,

$$P(C) = 1 - \sum_{k=0}^{2} \binom{18}{k} \left(\frac{-1}{1}\right)^k \left(\frac{-1}{5}\right)^{18-k}$$

(The summand on the RHS is called a binomial probability.)

Thus far, all the probabilities with which we have concerned ourselves have been unconditional. We now turn to *conditional probability*, which concerns how to update our beliefs (and computed probabilities) based on new evidence?

Definition 5.2. The probability of an event A given B is

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

if $P(B) > 0$.

Corollary 5.3.

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

or, more generally,

$$P(A_1 \cap \cdots \cap A_n) = P(A_1) \cdot P(A_2|A_1) \cdot P(A_3|A_1, A_2) \cdots P(A_n|A_1, \ldots, A_{n-1})$$

Theorem 5.4 (Bayes' Theorem)**.**

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Lecture 6 — 9/14/11

Theorem 6.1 (Law of Total Probability)**.** *Let S be a sample space and $A_1, \ldots, A_n$ a partition of S. Then*

$$\begin{aligned} P(B) &= P(B \cap A_1) + \cdots + P(B \cap A_n) \\ &= P(B|A_1)P(A_1) + \cdots + P(B|A_n)P(A_n) \end{aligned}$$

Example. Suppose we are given a random two-card hand from a standard deck.

1. What is the probability that both cards are aces given that we have an ace?

$$\begin{aligned} P(\text{both aces} \mid \text{have ace}) &= \frac{P(\text{both aces, have ace})}{P(\text{have ace})} \\ &= \frac{\binom{4}{2}/\binom{52}{2}}{1 - \binom{48}{2}/\binom{52}{2}} \\ &= \frac{1}{33} \end{aligned}$$

2. What is the probability that both cards are aces given that we have the ace of spades?

$$P(\text{both aces} \mid \text{ace of spades}) = \frac{3}{51} = \frac{1}{17}$$

Example. Suppose that a patient is being tested for a disease and it is known that 1% of similar patients have the disease. Suppose also that the patient tests positive and that the test is 95% accurate. Let D be the event that the patient has the disease and T the event that he tests positive. Then we know $P(T|D) = 0.95 = P(T^C|D^C)$. Using Bayes' theorem and the Law of Total Probability, we can compute

$$\begin{aligned} P(D|T) &= \frac{P(T|D)P(D)}{P(T)} \\ &= \frac{P(T|D)P(D)}{P(T|D)P(D) + P(T|D^C)P(D^C)} \\ &\approx 0.16 \end{aligned}$$

Definition 6.2. Two events A and B are conditionally independent of an event C if

$$P((A \cap B) \mid C) = P(A|C)P(B|C)$$

Example. Two conditionally independent events are not necessarily unconditionally independent. For instance, suppose we have a chess opponent of unknown strength. We might say that conditional on the opponent's strength, all games outcomes would be independent. However, *without* knowing the opponent's strength, earlier games would give us useful information about the opponent's strength; hence, without the conditioning, the game outcomes are not independent.

Example. Two independent events are not necessarily conditionally independent. Suppose we know that a fire alarm goes off (event A). Suppose there are only two possible causes, that a fire happened, F, or that someone was making popcorn, C, and suppose moreover that these events are independent. Given, however, that the alarm went off, we have

$$P(F|(A \cap C^C)) = 1$$

and hence we do not have conditional independence.

Lecture 7 — 9/16/11

Example (Monty Hall Problem)**.** Suppose there are three doors, behind two of which are goats and behind one of which is a car. Monty Hall, the game show host, knows the contents of each door, but we, the player, do not, and have one chance to choose the car. After choosing a door, Monty then opens one of the two remaining doors to reveal a goat (if both remaining doors have goats, he chooses with equal probability). We are then given the option to change our choice—should we do so?

In fact, we should; the chance that switching will give us the car is the same as the chance that we did not originally pick the car, which is $\frac{2}{3}$. However, we can also solve the problem by conditioning. Suppose we have chosen a door (WLOG, say the first). Let S be the event of finding the car by switching, and let D_i be the event that the car is in door i. Then by the Law of Total Probability,

$$\begin{aligned} P(S) &= P(S|D_1)\frac{1}{3} + P(S|D_2)\frac{1}{3} + P(S|D_3)\frac{1}{3} \\ &= 0 + 1 \cdot \frac{1}{3} + 1 \cdot \frac{1}{3} \\ &= \frac{2}{3} \end{aligned}$$

By symmetry, the probability that we succeed conditioned on the door Monty opens is the same.

Example (Simpson's Paradox)**.** Suppose we have the two following tables:

Hibbert

	heart	band-aid
success	70	10
failure	20	0

Nick

	heart	band-aid
success	2	81
failure	8	9

for the success of two doctors for two different operations.

Note that although Hibbert has a higher success rate conditional on each operation, Nick's success rate is higher overall. Let us denote A to be the event of a successful operation, B the event of being treated by Nick, and C the event of having heart surgery. In other words, then, we have

$$P(A|B,C) < P(A|B^C,C)$$

and

$$P(A|B,C^C) < P(A|B^C,C^C)$$

but

$$P(A|B) > P(A|B^C)$$

In this example, C is the confounder.

Lecture 8 — 9/19/11

Definition 8.1. A one-dimensional random walk models a (possibly infinite) sequence of successive steps along the number line, where, starting from some position i, we have a probability p of moving $+1$ and a probability $q = 1 - p$ of moving -1.

Example. An example of a one-dimensional random walk is the gambler's ruin problem, which asks: Given two individuals A and B playing a sequence of successive rounds of a game in which they bet \$1, with A winning B's dollar with probability p and A losing a dollar to B with probability $q = 1 - p$, what is the probability that A wins the game (supposing A has i dollars and B has $n-i$ dollars)? This problem can be modeled by a random walk with absorbing states at 0 and n, starting at i.

To solve this problem, we perform first-step analysis; that is, we condition on the first step. Let $p_i = P(A \text{ wins game} \mid A \text{ start at } i)$. Then by the Law of Total Probability, for $1 \le i \le n-1$.

$$p_i = pp_{i+1} + qp_{i-1}$$

and of course we have $p_0 = 0$ and $p_n = 1$. This equation is a difference equation.

To solve this equation, we start by guessing

$$p_i = x^i$$

Then we have

$$\begin{aligned} x^i &= px^{i+1} + qx^{i-1} \\ px^2 - x^i + q &= 0 \\ x &= \frac{1 \pm \sqrt{1-4pq}}{2p} \\ &= \frac{1 \pm \sqrt{1-4p(1-p)}}{2p} \\ &= \frac{1 \pm \sqrt{4p^2-4p+1}}{2p} \\ &= \frac{1 \mp (2p-1)}{2p} \\ &= 1, \frac{q}{p} \end{aligned}$$

As with differential equations, this gives a general solution of the form

$$p_i = A1^i + B\left(\frac{-1}{q}\right)^i$$

for $p \neq q$ (to avoid a repeated root). Our boundary conditions for p_0 and p_n give

$$B = -A$$

and

$$1 = A\left(1 - \frac{q}{p}\right)^n$$

To solve for the case where $p = q$, we can guess $x = \frac{q}{p}$ and take

$$\lim_{x\to 1} \frac{1-x^i}{1-x^n} = \lim_{x\to 1} \frac{ix^{i-1}}{nx^{n-1}} = \frac{i}{n}$$

So we have

$$p_i = \begin{cases} \dfrac{1-\left(\frac{-1}{q}\right)^i}{1-\left(\frac{-1}{q}\right)^n} & p \neq q \\ \dfrac{i}{n} & p = q \end{cases}$$

Now suppose that $p = 0.49$ and $i = n - i$. Then we have the following surprising table

N	$P(A \text{ wins})$
20	0.40
100	0.12
200	0.02

Note that this table is true when the odds are only slightly against A and when A and B start off with equal funding; it is easy to see that in a typical gambler's situation, the chance of winning is extremely small.

Definition 8.2. A random variable is a function

$$X : S \to \mathbb{R}$$

from some sample space S to the real line. A random variable acts as a "summary" of some aspect of an experiment.

Definition 8.3. A random variable X is said to have the Bernoulli distribution if X has only two possible values, 0 and 1, and there is some p such that

$$P(X = 1) = p \qquad P(X = 0) = 1 - p$$

We say that

$$X \sim \text{Bern}(p)$$

Note. We write $X = 1$ to denote the event

$$\{s \in S : X(s) = 1\} = X^{-1}\{1\}$$

Definition 8.4. The distribution of successes in n independent Bern(p) trials is called the binomial distribution and is given by

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

where $0 \leq k \leq n$. We write

$$X \sim \text{Bin}(n, p)$$

Definition 8.5. The probability mass function (PMF) of a discrete random variable (a random variable with enumerable values) is a function that gives the probability that the random variable takes some value. That is, given a discrete random variable X, its PMF is

$$f_X(x) = P(X = x)$$

Lecture 9 — 9/21/11

In addition to our definition of the binomial distribution by its PMF, we can also express a random variable $X \sim \text{Bin}(n, p)$ as a sum of indicator random variables,

$$X = X_1 + \cdots + X_n$$

where

$$X_i = \begin{cases} 1 & i\text{th trial succeeds} \\ 0 & \text{otherwise} \end{cases}$$

In other words, the X_i are i.i.d. (independent, identically distributed) Bern(p).

Definition 9.1. The cumulative distribution function (CDF) of a random variable X is

$$F_X(x) = P(X \leq x)$$

Note. The requirements for a PMF with values p_i is that each $p_i \geq 0$ and $\sum_i p_i = 1$. For Bin(n, p), we can easily verify this with the binomial theorem, which yields

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

Proposition 9.2. *If X, Y are independent random variables and $X \sim \text{Bin}(n, p)$, $Y \sim \text{Bin}(m, p)$, then*

$$X + Y \sim \text{Bin}(n + m, p)$$

Proof. This is clear from our "story" definition of the binomial distribution, as well as from our indicator r.v.'s. Let us also check this using PMFs.

$$\begin{aligned}
P(X+Y=k) &= \sum_{j=0}^{k} P(X+Y=k \mid X=j) P(X=j) \\
&= \sum_{j=0}^{k} P(Y=k-j \mid X-j) \binom{n}{j} p^j q^{n-j} \\
\text{independence} \quad &= \sum_{j=0}^{k} P(Y=k-j) \binom{n}{j} p^j q^{n-j} \\
&= \sum_{j=0}^{k} \binom{m}{k-j} p^{k-j} q^{m-(k-j)} \binom{n}{j} p^j q^{n-j} \\
&= p^k q^{n+m-k} \sum_{j=0}^{k} \binom{m}{k-j} \binom{n}{j} \\
\text{Vandermonde} \quad &= \binom{n+m}{k} p^k q^{n+m-k}
\end{aligned}$$

■

Example. Suppose we draw a random 5-card hand from a standard 52-card deck. We want to find the distribution of the number of aces in the hand. Let $X = \#\text{aces}$. We want to determine the PMF of X (or the CDF—but the PMF is easier). We know that $P(X = k) = 0$ except if $k = 0, 1, 2, 3, 4$. This is clearly not binomial since the trials (of drawing cards) are not independent. For $k = 0, 1, 2, 3, 4$, we have

$$P(X = k) = \frac{\binom{4}{k}\binom{48}{5-k}}{\binom{52}{5}}$$

which is just the probability of choosing k out of the 4 aces and $5-k$ of the non-aces. This is reminiscient of the elk problem in the homework.

Definition 9.3. Suppose we have w white and b black marbles, out of which we choose a simple random sample of n. The distribution of # of white marbles in the sample, which we will call X, is given by

$$P(X = k) = \frac{\binom{w}{k}\binom{b}{n-k}}{\binom{w+b}{n}}$$

where $0 \leq k \leq w$ and $0 \leq n - k \leq b$. This is called the hypergeometric distribution, denoted HGeom(w, b, n).

Proof. We should show that the above is a valid PMF. It is clearly nonnegative. We also have, by Vandermonde's identity,

$$\sum_{k=0}^{w} \frac{\binom{w}{k}\binom{b}{n-k}}{\binom{w+b}{n}} = \frac{\binom{w+b}{n}}{\binom{w+b}{n}} = 1$$

■

Note. The difference between the hypergometric and binomial distributions is whether or not we sample with replacement. We would expect that in the limiting case of $n \to \infty$, they would behave similarly.

Lecture 10 — 9/23/11

Proposition 10.1 (Properties of CDFs)**.** *A function F_X is a valid CDF iff the following hold about F_X:*

1. *monotonically nondecreasing*
2. *right-continuous*
3. $\lim_{x\to-\infty} F_X(x) = 0$ *and* $\lim_{x\to\infty} F_X(x) = 1$.

Definition 10.2. Two random variables X and Y are independent if $\forall x, y$,

$$P(X \le x, Y \le y) = P(X \le x)P(Y \le y)$$

In the discrete case, we can say equivalently that

$$P(X = x, Y = y) = P(X = x)P(Y = y)$$

Note. As an aside before we move on to discuss averages and expected values, recall that

$$\frac{1}{n}\sum_{i=1}^{n} i = \frac{n+1}{2}$$

Example. Suppose we want to find the average of $1, 1, 1, 1, 1, 3, 3, 5$. We could just add these up and divide by 8, or we could formulate the average as a weighted average,

$$\frac{5}{8}\cdot 1 + \frac{2}{8}\cdot 3 + \frac{1}{8}\cdot 5$$

Definition 10.3. The expected value or average of a discrete random variable X is

$$E(X) = \sum_{x\in \mathrm{Im}(X)} xP(X = x)$$

Observation 10.4. Let $X \sim \mathrm{Bern}(p)$. Then

$$E(X) = 1 \cdot P(X = 1) + 0 \cdot P(X = 0) = p$$

Definition 10.5. If A is some event, then an indicator random variable for A is

$$X = \begin{cases} 1 & A \text{ occurs} \\ 0 & \text{otherwise} \end{cases}$$

By definition, $X \sim \mathrm{Bern}(P(A))$, and by the above,

$$E(X) = P(A)$$

The above shows that to get the probability of an event, we can simply compute the expected value of an indicator.

Observation 10.6. Let $X \sim \mathrm{Bin}(n, p)$. Then (using the binomial theorem),

$$\begin{aligned} E(X) &= \sum_{k=0}^{n} k\binom{n}{k}p^k q^{n-k} \\ &= \sum_{k=1}^{n} k\binom{n}{k}p^k q^{n-k} \\ &= \sum_{k=1}^{n} n\binom{n-1}{k-1}p^k q^{n-k} \\ &= np\sum_{k=1}^{n}\binom{n-1}{k-1}p^{k-1} q^{n-k} \\ &= np\sum_{j=0}^{n-1}\binom{n-1}{j}p^{j} q^{n-1-j} \\ &= np \end{aligned}$$

Proposition 10.7. *Expected value is linear; that is, for random variables X and Y and some constant c,*

$$E(X + Y) = E(X) + E(Y)$$

and

$$E(cX) = cE(X)$$

Observation 10.8. Using linearity, given $X \sim \mathrm{Bin}(n, p)$, since we know

$$X = X_1 + \cdots + X_n$$

where the X_i are i.i.d. $\mathrm{Bern}(p)$, we have

$$X = p + \cdots + p = np$$

Example. Suppose that, once again, we are choosing a five card hand out of a standard deck, with $X = \#\text{aces}$. If X_i is an indicator of the ith card being an ace, we have

$$\begin{aligned} E(X) &= E(X_1 + \cdots + X_5) \\ &= E(X_1) + \cdots + E(X_5) \\ \text{by symmetry} \quad &= 5E(X_1) \\ &= 5P(\text{first card is ace}) \\ &= \frac{5}{13} \end{aligned}$$

Note that this holds even though the X_i are dependent.

Definition 10.9. The geometric distribution, $\mathrm{Geom}(p)$, is the number of failures of independent $\mathrm{Bern}(p)$ trials before the first success. Its PMF is given by (for $X \sim \mathrm{Geom}(p)$)

$$P(X = k) = q^k p$$

for $k \in \mathbb{N}$. Note that this PMF is valid since

$$\sum_{k=0}^{\infty} pq^k = p \cdot \frac{1}{1-q} = 1$$

Observation 10.10. Let $X \sim \text{Geom}(p)$. We have our formula for infinite geometric series,

$$\sum_{k=0}^{\infty} q^k = \frac{1}{1-q}$$

Taking the derivative of both sides gives

$$\sum_{k=1}^{\infty} kq^{k-1} = \frac{1}{(1-q)^2}$$

Then

$$E(X) = \sum_{k=0}^{\infty} kpq^k = p\sum_{k=0}^{\infty} kq^k = \frac{pq}{(1-q)^2} = \frac{q}{p}$$

Alternatively, we can use first step analysis and write a recursive formula for $E(X)$. If we condition on what happens in the first Bernoulli trial, we have

$$\begin{aligned} E(X) &= 0 \cdot p + (1 + E(X))q \\ E(X) - qE(X) &= q \\ E(X) &= \frac{q}{1-q} \\ E(X) &= \frac{q}{p} \end{aligned}$$

Lecture 11 — 9/26/11

Recall our assertion that E, the expected value function, is linear. We now prove this statement.

Proof. Let X and Y be discrete random variables. We want to show that $E(X+Y) = E(X) + E(Y)$.

$$\begin{aligned} E(X+Y) &= \sum_t tP(X+Y=t) \\ &= \sum_s (X+Y)(s)P(\{s\}) \\ &= \sum_s (X(s)+Y(s))P(\{s\}) \\ &= \sum_s X(s)P(\{s\}) + \sum_s Y(s)P(\{s\}) \\ &= \sum_x xP(X=x) + \sum_y yP(Y=y) \\ &= E(X) + E(Y) \end{aligned}$$

The proof that $E(cX) = cE(X)$ is similar. ■

Definition 11.1. The negative binomial distribution, $\text{NB}(r,p)$, is given by the number of failures of independent $\text{Bern}(p)$ trials before the rth success. The PMF for $X \sim \text{NB}(r,p)$ is given by

$$P(X=n) = \binom{n+r-1}{r-1} p^r (1-p)^n$$

for $n \in \mathbb{N}$.

Observation 11.2. Let $X \sim \text{NB}(r,p)$. We can write $X = X_1 + \cdots + X_r$ where each X_i is the number of failures between the $(i-1)$th and ith success. Then $X_i \sim \text{Geom}(p)$. Thus,

$$E(X) = E(X_1) + \cdots + E(X_r) = \frac{rq}{p}$$

Observation 11.3. Let $X \sim \text{FS}(p)$, where $\text{FS}(p)$ is the time until the first success of independent $\text{Bern}(p)$ trials, counting the success. Then if we take $Y = X - 1$, we have $Y \sim \text{Geom}(p)$. So,

$$E(X) = E(Y) + 1 = \frac{q}{p} + 1 = \frac{1}{p}$$

Example. Suppose we have a random permutation of $\{1, \ldots, n\}$ with $n \geq 2$. What is the expected number of local maxima—that is, numbers greater than both its neighbors?

Let I_j be the indicator random variable for position j being a local maximum ($1 \leq j \leq n$). We are interested in

$$E(I_1 + \cdots + I_n) = E(I_1) + \cdots + E(I_n)$$

For the non-endpoint positions, in each local neighborhood of three numbers, the probability that the largest number is in the center position is $\frac{1}{3}$.

$$5, 2, \cdots, \underbrace{28, 3, 8}, \cdots, 14$$

Moreover, these positions are all symmetrical. Analogously, the probability that an endpoint position is a local maximum is $\frac{1}{2}$. Then we have

$$E(I_1) + \cdots + E(I_n) = \frac{n-2}{3} + \frac{2}{2} = \frac{n+1}{3}$$

Example (St. Petersburg Paradox)**.** Suppose you are given the offer to play a game where a coin is flipped until a heads is landed. Then, for the number of flips i made up to and including the heads, you receive $\$2^i$. How much should you be willing to pay to play this game? That is, what price would make the game fair, or the expected value zero?

Let X be the number of flips of the fair coin up to and including the first heads. Clearly, $X \sim FS(\frac{1}{2})$. If we let $Y = 2^X$, we want to find $E(Y)$. We have

$$E(Y) = \sum_{k=1}^{\infty} 2^k \cdot \frac{1}{2^k} = \sum_{k=1}^{\infty} 1$$

This assumes, however, that our cash source is boundless. If we bound it at 2^K for some specific K, we should only bet K dollars for a fair game—this is a sizeable difference.

Lecture 12 — 9/28/11

Definition 12.1. The Poisson distribution, $\text{Pois}(\lambda)$, is given by the PMF

$$P(X = k) = \frac{e^{-\lambda}\lambda^k}{k!}$$

for $k \in \mathbb{N}$, $X \sim \text{Pois}(\lambda)$. We call λ the rate parameter.

Observation 12.2. Checking that this PMF is indeed valid, we have

$$\sum_{k=0}^{\infty} e^{-\lambda}\frac{\lambda^k}{k!} = e^{-\lambda}e^{\lambda} = 1$$

Its mean is given by

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

The Poisson distribution is often used for applications where we count the successes of a large number of trials where the per-trial success rate is small. For example, the Poisson distribution is a good starting point for counting the number of people who email you over the course of an hour. The number of chocolate chips in a chocolate chip cookie is another good candidate for a Poisson distribution, or the number of earthquakes in a year in some particular region.

Since the Poisson distribution is not bounded, these examples will not be precisely Poisson. However, in general, with a large number of events A_i with small $P(A_i)$, and where the A_i are all independent or "weakly dependent," then the number of the A_i that occur is approximately $\text{Pois}(\lambda)$, with $\lambda \approx \sum_{i=1}^{n} P(A_i)$. We call this a Poisson approximation.

Proposition 12.3. *Let $X \sim \text{Bin}(n, p)$. Then as $n \to \infty$, $p \to 0$, and where $\lambda = np$ is held constant, we have $X \sim \text{Pois}(\lambda)$.*

Proof. Fix k. Then as $n \to \infty$ and $p \to 0$,

$$\begin{aligned} \lim_{\substack{n\to\infty \\ p\to 0}} P(X = k) &= \lim_{\substack{n\to\infty \\ p\to 0}} \binom{n}{k} p^k (1-p)^{n-k} \\ &= \lim_{\substack{n\to\infty \\ p\to 0}} \frac{n(n-1)\cdots(n-k+1)}{k!}\left(\frac{-1}{\lambda}\right)^k \cdot \\ &\quad \left(1-\frac{\lambda}{n}\right)^n \left(1-\frac{\lambda}{n}\right)^{-k} \\ &= \frac{\lambda^k}{k!}\cdot e^{-\lambda} \end{aligned}$$

■

Example. Suppose we have n people and we want to know the *approximate* probability that at least three individuals have the same birthday. There are $\binom{n}{3}$ triplets of people; for each triplet, let I_{ijk} be the indicator r.v. that persons i, j, and k have the same birthday. Let $X = \#$ triple matches. Then we know that

$$E(X) = \binom{n}{3}\frac{1}{365^2}$$

To approximate $P(X \geq 1)$, we approximate $X \sim \text{Pois}(\lambda)$ with $\lambda = E(X)$. Then we have

$$P(X \geq 1) = 1 - P(X = 0) = 1 - e^{-\lambda}\frac{\lambda^0}{0!} = 1 - e^{-\lambda}$$

Lecture 13 — 9/30/11

Definition 13.1. Let X be a random variable. Then X has a probability density function (PDF) $f_X(x)$ if

$$P(a \leq X \leq b) = \int_a^b f_X(x)\,\mathrm{d}x$$

A valid PDF must satisfy

1. $\forall x, f_X(x) \geq 0$
2. $\displaystyle\int_{-\infty}^{\infty} f_X(x)\,\mathrm{d}x = 1$

Note. For $\epsilon > 0$ very small, we have

$$f_X(x_0)\cdot\epsilon \approx P\left(X \in (x_0 - \frac{\epsilon}{2}, x_0 + \frac{\epsilon}{2})\right)$$

Theorem 13.2. *If X has PDF f_X, then its CDF is*

$$F_X(x) = P(X \leq x) = \int_{-\infty}^{x} f_X(t)\,\mathrm{d}t$$

If X is continuous and has CDF F_X, then its PDF is

$$f_X(x) = F_X'(x)$$

Moreover,

$$P(a < X < b) = \int_a^b f_X(x)\,\mathrm{d}x = F_X(b) - F_X(a)$$

Proof. By the Fundamental Theorem of Calculus. ■

Definition 13.3. The expected value of a continuous random variable X is given by

$$E(X) = \int_{-\infty}^{\infty} x f_X(x)\,\mathrm{d}x$$

Giving the expected value is like giving a one number summary of the average, but it provides no information about the spread of a distribution.

Definition 13.4. The variance of a random variable X is given by

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

which is the expected value of the distance from X to its mean; that is, it is, on average, how far X is from its mean.

We can't use $E(X - EX)$ because, by linearity, we have

$$E(X - EX) = EX - E(EX) = EX - EX = 0$$

We would like to use $E|X - EX|$, but absolute value is hard to work with; instead, we have

Definition 13.5. The standard deviation of a random variable X is

$$\mathrm{SD}(X) = \sqrt{\mathrm{Var}(X)}$$

Note. Another way we can write variance is

$$\begin{aligned}\mathrm{Var}(X) &= E((X - EX)^2)\\ &= E(X^2 - 2X(EX) + (EX)^2)\\ &= E(X^2) - 2E(X)E(X) + (EX)^2\\ &= E(X^2) - (EX)^2\end{aligned}$$

Definition 13.6. The uniform distribution, $\mathrm{Unif}(a, b)$, is given by a completely random point chosen in the interval $[a, b]$. Note that the probability of picking a given point x_0 is exactly 0; the uniform distribution is continuous. The PDF for $U \sim \mathrm{Unif}(a, b)$ is given by

$$f_U(x) = \begin{cases} c & a \le x \le b \\ 0 & \text{otherwise} \end{cases}$$

for some constant c. To find c, we note that, by the definition of PDF, we have

$$\begin{aligned}\int_a^b c\,\mathrm{d}x &= 1\\ c(b - a) &= 1\\ c &= \frac{1}{b - a}\end{aligned}$$

The CDF, then, is given by

$$\begin{aligned}F_U(x) = \int_{-\infty}^{x} f_U(t)\,\mathrm{d}t &= \int_a^x f_U(t)\,\mathrm{d}t\\ &= \int_a^x \frac{1}{b - a}\,\mathrm{d}t\\ &= \frac{x - a}{b - a}\end{aligned}$$

Observation 13.7. The expected value of an r.v. $U \sim \mathrm{Unif}(a, b)$ is

$$\begin{aligned}E(U) &= \int_a^b \frac{x}{b - a}\,\mathrm{d}x\\ &= \frac{x^2}{2(b - a)}\Bigg|_a^b\\ &= \frac{b^2 - a^2}{2(b - a)}\\ &= \frac{(b + a)(b - a)}{2(b - a)}\\ &= \frac{b + a}{2}\end{aligned}$$

This is the midpoint of the interval $[a, b]$.

Finding the variance of $U \sim \mathrm{Unif}(a, b)$, however, is a bit more trouble. We need to determine $E(U^2)$, but it is too much of a hassle to figure out the PDF of U^2. Ideally, things would be as simple as

$$E(U^2) = \int_{-\infty}^{\infty} x^2 f_U(x)\,\mathrm{d}x$$

Fortunately, this is true:

Theorem 13.8 (Law of the Unconscious Statistician (LOTUS))**.** *Let X be a continuous random variable, $g : \mathbb{R} \to \mathbb{R}$ continuous. Then*

$$E(g(X)) = \int_{-\infty}^{\infty} g(x) f_X(x)\,\mathrm{d}x$$

where f_X is the PDF of X. This allows us to determine the expected value of $g(X)$ without knowing its distribution.

Observation 13.9. The variance of $U \sim \mathrm{Unif}(a, b)$ is given by

$$\begin{aligned}\mathrm{Var}(U) &= E(U^2) - (EU)^2\\ &= \int_a^b x^2 f_U(x)\,\mathrm{d}x - \left(\frac{b + a}{2}\right)^2\\ &= \frac{1}{b - a}\int_a^b x^2\,\mathrm{d}x - \left(\frac{b + a}{2}\right)^2\end{aligned}$$

$$= \frac{1}{b-a} \cdot \frac{x^3}{3}\bigg|_a^b - \left(\frac{b+a}{2}\right)^2$$
$$= \frac{b^3}{3(b-a)} - \frac{a^3}{3(b-a)} - \frac{(b+a)^2}{4}$$
$$= \frac{(b-a)^2}{12}$$

The following table is useful for comparing discrete and continuous random variables:

	discrete	continuous
P?F	PMF $P(X=x)$	PDF $f_X(x)$
CDF	$F_X(x) = P(X \le x)$	$F_X(x) = P(X \le x)$
$E(X)$	$\sum_x xP(X=x)$	$\int_{-\infty}^{\infty} x f_X(x)\,\mathrm{d}x$
$\mathrm{Var}(X)$	$EX^2 - (EX)^2$	$EX^2 - (EX)^2$
$E(g(X))$ [LOTUS]	$\sum_x g(x)P(X=x)$	$\int_{-\infty}^{\infty} g(x) f_X(x)\,\mathrm{d}x$

Lecture 14 — 10/3/11

Theorem 14.1 (Universality of the Uniform)**.** *Let us take $U \sim \mathrm{Unif}(0,1)$, F a strictly increasing CDF. Then for $X = F^{-1}(U)$, we have $X \sim F$. Moreover, for any random variable X, if $X \sim F$, then $F(X) \sim \mathrm{Unif}(0,1)$.*

Proof. We have

$$P(X \le x) = P(F^{-1}(U) \le x) = P(U \le F(x)) = F(x)$$

since $P(U \le F(x))$ is the length of the interval $[0, F(x)]$, which is $F(x)$. For the second part,

$$P(F(X) \le x) = P(X \le F^{-1}(x)) = F(F^{-1}(x)) = x$$

since F is X's CDF. But this shows that $F(X) \sim \mathrm{Unif}(0,1)$. ■

Example. Let $F(x) = 1 - e^{-x}$ with $x > 0$ be the CDF of an r.v. X. Then $F(X) = 1 - e^{-X}$ by an application of the second part of Universality of the Uniform.

Example. Let $F(x) = 1 - e^{-x}$ with $x > 0$, and also let $U \sim \mathrm{Unif}(0,1)$. Suppose we want to simulate F with a random variable X; that is, $X \sim F$. Then computing the inverse

$$F^{-1}(u) = -\ln(1-u)$$

yields $F^{-1}(U) = -\ln(1-U) \sim F$.

Proposition 14.2. *The standard uniform distribution is symmetric; that is, if $U \sim \mathrm{Unif}(0,1)$, then also $1 - U \sim \mathrm{Unif}(0,1)$.*

Intuitively, this is true because there is no difference between measuring U from the right vs. from the left of $[0,1]$.

The general uniform distribution is also linear; that is, $a + bU$ is uniform on some interval $[a,b]$. If a distribution is nonlinear, it is hence nonuniform.

Definition 14.3. We say that random variables $X_1, \ldots, X_n$ are independent if

- for continuous, $P(X_1 \le x_1, \ldots, X_n \le x_n) = P(X_1 \le x_1) \cdots P(X_n \le x_n)$
- for discrete, $P(X_1 = x_1, \ldots, X_n = x_n) = P(X_1 = x_1) \cdots P(X_n = x_n)$

The expressions on the LHS are called joint CDFs and joint PMFs respectively.

Note that pairwise independence does not imply independence.

Example. Consider the penny matching game, where $X_1, X_2 \sim \mathrm{Bern}(\frac{1}{2})$, i.i.d., and let X_3 be the indicator r.v. for the event $X_1 = X_2$ (the r.v. for winning the game). All of these are pairwise independent, but the X_3 is clearly dependent on the combined outcomes of X_1 and X_2.

Definition 14.4. The normal distribution, given by $\mathcal{N}(0,1)$, is defined by PDF

$$f(z) = ce^{-z^2/2}$$

where c is the normalizing constant required to have f integrate to 1.

Proof. We want to prove that our PDF is valid; to do so, we will simply determine the value of the normalizing constant that makes it so. We will integrate the square of the PDF sans constant because it is easier than integrating naïvely

$$\int_{-\infty}^{\infty} e^{-z^2/2}\,\mathrm{d}z \int_{-\infty}^{\infty} e^{-z^2/2}\,\mathrm{d}z$$
$$= \int_{-\infty}^{\infty} e^{-x^2/2}\,\mathrm{d}x \int_{-\infty}^{\infty} e^{-y^2/2}\,\mathrm{d}y$$
$$= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-(x^2+y^2)/2}\,\mathrm{d}x\,\mathrm{d}y$$
$$= \int_0^{2\pi}\int_0^{\infty} e^{-r^2/2} r\,\mathrm{d}r\,\mathrm{d}\theta$$

Substituting $u = \frac{r^2}{2}, \mathrm{d}u = r\,\mathrm{d}r$

$$= \int_0^{2\pi}\left(\int_0^{\infty} e^{-u}\,\mathrm{d}u\right)\mathrm{d}\theta$$
$$= 2\pi$$

So our normalizing constant is $c = \frac{1}{\sqrt{2\pi}}$. ■

Observation 14.5. Let us compute the mean and variance of $Z \sim \mathcal{N}(0,1)$. We have

$$EZ = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} z e^{-z^2/2} \, \mathrm{d}z = 0$$

by symmetry (the integrand is odd). The variance reduces to

$$\mathrm{Var}(Z) = E(Z^2) - (EZ)^2 = E(Z^2)$$

By LOTUS,

$$\begin{aligned} E(Z^2) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} z^2 e^{-z^2/2} \, \mathrm{d}z \\ \text{evenness} \quad &= \frac{2}{\sqrt{2\pi}} \int_{0}^{\infty} z^2 e^{-z^2/2} \, \mathrm{d}z \\ \text{by parts} \quad &= \frac{2}{\sqrt{2\pi}} \int_{0}^{\infty} \underbrace{z}_{u} \underbrace{z e^{-z^2/2} \, \mathrm{d}z}_{\mathrm{d}v} \\ &= \frac{2}{\sqrt{2\pi}} \left(uv \Big|_0^{\infty} + \int_0^{\infty} e^{-z^2/2} \, \mathrm{d}z \right) \\ &= \frac{2}{\sqrt{2\pi}} \left(0 + \frac{\sqrt{2\pi}}{2} \right) \\ &= 1 \end{aligned}$$

We use Φ to denote the standard normal CDF; so

$$\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-t^2/2} \, \mathrm{d}t$$

By symmetry, we also have

$$\Phi(-z) = 1 - \Phi(z)$$

Lecture 15 — 10/5/11

Recall the standard normal distribution. Let Z be an r.v., $Z \sim \mathcal{N}(0,1)$. Then Z has CDF Φ; it has $E(Z) = 0$, $\mathrm{Var}(Z) = E(Z^2) = 1$, and $E(Z^3) = 0$.[1] By symmetry, also $-Z \sim \mathcal{N}(0,1)$.

Definition 15.1. Let $X = \mu + \sigma Z$, with $\mu \in \mathbb{R}$ (the mean or center), $\sigma > 0$ (the SD or scale). Then we say $X \sim \mathcal{N}(\mu, \sigma^2)$. This is the general normal distribution.

If $X \sim \mathcal{N}(\mu, \sigma^2)$, we have $E(X) = \mu$ and $\mathrm{Var}(\mu + \sigma Z) = \sigma^2 \mathrm{Var}(Z) = \sigma^2$. We call $Z = \frac{X-\mu}{\sigma}$ the standardization of X. X has CDF

$$P(X \le x) = P\left(\frac{X-\mu}{\sigma} \le \frac{x-\mu}{\sigma} \right) = \Phi\left(\frac{x-\mu}{\sigma} \right)$$

which yields a PDF of

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\left(\frac{x-\mu}{\sigma}\right)^2/2}$$

We also have $-X = -\mu + \sigma(-Z) \sim \mathcal{N}(-\mu, \sigma^2)$.

Later, we will show that if $X_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$ are independent, then

$$X_i + X_j \sim \mathcal{N}(\mu_i + \mu_j, \sigma_i^2 + \sigma_j^2)$$

and

$$X_i - X_j \sim \mathcal{N}(\mu_i - \mu_j, \sigma_i^2 + \sigma_j^2)$$

Observation 15.2. If $X \sim \mathcal{N}(\mu, \sigma^2)$, we have

$$\begin{aligned} P(|X-\mu| \le \sigma) &\approx 68\% \\ P(|X-\mu| \le 2\sigma) &\approx 95\% \\ P(|X-\mu| \le 3\sigma) &\approx 99.7\% \end{aligned}$$

Observation 15.3. We observe some properties of the variance.

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

For any constant c,

$$\begin{aligned} \mathrm{Var}(X + c) &= \mathrm{Var}(X) \\ \mathrm{Var}(cX) &= c^2 \mathrm{Var}(X) \end{aligned}$$

Since variance is not linear, in general, $\mathrm{Var}(X+Y) \neq \mathrm{Var}(X) + \mathrm{Var}(Y)$. However, if X and Y are independent, we *do* have equality. On the other extreme,

$$\mathrm{Var}(X + X) = \mathrm{Var}(2X) = 4 \mathrm{Var}(X)$$

Also, in general,

$$\begin{aligned} &\mathrm{Var}(X) \ge 0 \\ &\mathrm{Var}(X) = 0 \iff \exists a : P(X = a) = 1 \end{aligned}$$

Observation 15.4. Let us compute the variance of the Poisson distribution. Let $X \sim \mathrm{Pois}(\lambda)$. We have

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

To reduce this sum, we can do the following:

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

Taking the derivative w.r.t. λ,

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

[1] These are called the first, second, and third moments.

$$\sum_{k=1}^{\infty} \frac{k\lambda^k}{k!} = \lambda e^{\lambda}$$

Repeating,

$$\sum_{k=1}^{\infty} \frac{k^2\lambda^{k-1}}{k!} = \lambda e^{\lambda} + e^{\lambda}$$

$$\sum_{k=1}^{\infty} \frac{k^2\lambda^k}{k!} = \lambda e^{\lambda}(\lambda + 1)$$

So,

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

So for our variance, we have

$$\mathrm{Var}(X) = (\lambda^2 + \lambda) - \lambda^2 = \lambda$$

Observation 15.5. Let us compute the variance of the binomial distribution. Let $X \sim \mathrm{Bin}(n, p)$. We can write

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

where the I_j are i.i.d. $\mathrm{Bern}(p)$. Then,

$$X^2 = I_1^2 + \cdots + I_n^2 + 2I_1I_2 + 2I_1I_3 + \cdots + 2I_{n-1}I_n$$

where I_iI_j is the indicator of success on both i and j.

$$\begin{aligned} E(X^2) &= nE(I_1^2) + 2\binom{n}{2}E(I_1I_2) \\ &= np + n(n-1)p^2 \\ &= np + n^2p^2 - np^2 \end{aligned}$$

So,

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

Proof. (of Discrete LOTUS)
We want to show that $E(g(X) = \sum_x g(x)P(X = x)$. To do so, once again we can "ungroup" our expected value expression:

$$\sum_x g(x)P(X = x) = \sum_{s \in S} g(X(s))P(\{s\})$$

We can rewrite this as

$$\begin{aligned} \sum_x \sum_{s:X(s)=x} g(X(s))P(\{s\}) &= \sum_x g(x) \sum_{s:X(s)=x} P(\{s\}) \\ &= \sum_x g(x)P(X = x) \end{aligned}$$

■

Lecture 17 — 10/14/11

Definition 17.1. The exponential distribution, $\mathrm{Expo}(\lambda)$, is defined by PDF

$$f(x) = \lambda e^{-\lambda x}$$

for $x > 0$ and 0 elsewhere. We call λ the rate parameter.

Integrating clearly yields 1, which demonstrates validity. Our CDF is given by

$$\begin{aligned} F(x) &= \int_0^x \lambda e^{-\lambda t}\,\mathrm{d}t \\ &= \begin{cases} 1 - e^{-\lambda x} & x > 0 \\ 0 & \text{otherwise} \end{cases} \end{aligned}$$

Observation 17.2. We can normalize any $X \sim \mathrm{Expo}(\lambda)$ by multiplying by λ, which gives $Y = \lambda X \sim \mathrm{Expo}(1)$. We have

$$P(Y \le y) = P(X \le \frac{y}{\lambda}) = 1 - e^{-\lambda y/\lambda} = 1 - e^{-y}$$

Let us now compute the mean and variance of $Y \sim \mathrm{Expo}(1)$. We have

$$\begin{aligned} E(Y) &= \int_0^{\infty} ye^{-y}\,\mathrm{d}y \\ &= (-ye^{-y})\Big|_0^{\infty} + \int_0^{\infty} e^{-y}\,\mathrm{d}y \\ &= 1 \end{aligned}$$

for the mean. For the variance,

$$\begin{aligned} \mathrm{Var}(Y) &= EY^2 - (EY)^2 \\ &= \int_0^{\infty} y^2e^{-y}\,\mathrm{d}y - 1 \\ &= 1 \end{aligned}$$

Then for $X = \frac{Y}{\lambda}$, we have $E(X) = \frac{1}{\lambda}$ and $\mathrm{Var}(X) = \frac{1}{\lambda^2}$.

Definition 17.3. A random variable X has a memoryless distribution if

$$P(X \ge s + t \mid X \ge s) = P(X \ge t)$$

Intuitively, if we have a random variable that we interpret as a waiting time, memorylessness means that no matter how long we have already waited, the probability of having to wait a given time more is invariant.

Proposition 17.4. *The exponential distribution is memoryless.*

Proof. Let $X \sim \text{Expo}(\lambda)$. We know that

$$P(X \geq t) = 1 - P(X \leq t) = e^{-\lambda t}$$

Meanwhile,

$$\begin{aligned} P(X \geq s+t \mid X \geq s) &= \frac{P(X \geq s+t, X \geq s)}{P(X \geq s)} \\ &= \frac{P(X \geq s+t)}{P(X \geq s)} \\ &= \frac{e^{-\lambda(s+t)}}{e^{-\lambda s}} \\ &= e^{-\lambda t} \\ &= P(X \geq t) \end{aligned}$$

which is our desired result. ■

Example. Let $X \sim \text{Expo}(\lambda)$. Then by linearity and by the memorylessness,

$$\begin{aligned} E(X \mid X > a) &= a + E(X - a \mid X > a) \\ &= a + \frac{1}{\lambda} \end{aligned}$$

Lecture 18 — 10/17/11

Theorem 18.1. *If X is a positive, continuous random variable that is memoryless (i.e., its distribution is memoryless), then there exists $\lambda \in \mathbb{R}$ such that $X \sim \text{Expo}(\lambda)$.*

Proof. Let F be the CDF of X and $G = 1 - F$. By memorylessness,

$$G(s+t) = G(s)G(t)$$

We can easily derive from this identity that $\forall k \in \mathbb{Q}$,

$$G(kt) = G(t)^k$$

This can be extended to all $k \in \mathbb{R}$. If we take $t = 1$, then we have

$$G(x) = G(1)^x = e^{x \ln G(1)}$$

But since $G(1) < 1$, we can define $\lambda = -\ln G(1)$, and we will have $\lambda > 0$. Then this gives us

$$F(x) = 1 - G(x) = 1 - e^{-\lambda x}$$

as desired. ■

Definition 18.2. A random variable X has moment-generating function (MGF)

$$M(t) = E(e^{tX})$$

if $M(t)$ is bounded on some interval $(-\epsilon, \epsilon)$ about zero.

Observation 18.3. We might ask why we call M "moment-generating." Consider the Taylor expansion of M:

$$E(e^{tX}) = E\left(\sum_{n=0}^{\infty} \frac{X^n t^n}{n!}\right) = \sum_{n=0}^{\infty} \frac{E(X^n) t^n}{n!}$$

Note that we cannot simply make use of linearity since our sum is infinite; however, this equation does hold for reasons beyond the scope of the course.

This observation also shows us that

$$E(X^n) = M^{(n)}(0)$$

Claim 18.4. *If X and Y have the same MGF, then they have the same CDF.*

We will not prove this claim.

Observation 18.5. If X has MGF M_X and Y has MGF M_Y, then

$$M_{X+Y} = E(e^{t(X+Y)}) = E(e^{tX})E(e^{tY}) = M_X M_Y$$

The second inequality comes from the claim (which we will prove later) that if for X, Y independent, $E(XY) = E(X)E(Y)$.

Example. Let $X \sim \text{Bern}(p)$. Then

$$M(t) = E(e^{tX}) = pe^t + q$$

Suppose now that $X \sim \text{Bin}(n, p)$. Again, we write $X = I_1 + \cdots + I_n$ where the I_j are i.i.d Bern(p). Then we see that

$$M(t) = (pe^t + q)^n$$

Example. Let $Z \sim \mathcal{N}(0, 1)$. We have

$$\begin{aligned} M(t) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{tz - z^2/2} \, \mathrm{d}z \\ \text{completing the square} \quad &= \frac{e^{t^2/2}}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-(1/2)(z-t)^2} \, \mathrm{d}z \\ &= e^{t^2/2} \end{aligned}$$

Example. Suppose $X_1, X_2, \ldots$ are conditionally independent (given p) random variables that are Bern(p). Suppose also that p is unknown. In the Bayesian approach, let us treat p as a random variable. Let $p \sim \text{Unif}(0, 1)$; we call this the prior distribution.

Let $S_n = X_1 + \cdots + X_n$. Then $S_n \mid p \sim \text{Bin}(n, p)$. We want to find the posterior distribution, $p \mid S_n$, which will give us $P(X_{n+1} = 1 \mid S_n = k)$. Using "Bayes' Theorem,"

$$f(p \mid S_n = k) = \frac{P(S_n = k \mid p) f(p)}{P(S_n = k)}$$

$$= \frac{P(S_n = k \mid p)}{P(S_n = k)}$$
$$\propto p^k (1-p)^{n-k}$$

In the specific case of $S_n = n$, normalizing is easier:

$$f(p \mid S_n = n) = (n+1)p^n$$

Computing $P(X_{n+1} = 1 \mid S_n = k)$ simply requires finding the expected value of an indicator with the above probability $p \mid S_n = n$.

$$P(X_{n+1} = 1 \mid S_n = n) = \int_0^1 p(n+1)p^n \, \mathrm{d}p = \frac{n+1}{n+2}$$

Lecture 19 — 10/19/11

Observation 19.1. Let $X \sim \text{Expo}(1)$. We want to determine the MGF M of X. By LOTUS,

$$\begin{aligned} M(t) &= E(e^{tX}) \\ &= \int_0^\infty e^{tx} e^{-x} \, \mathrm{d}x \\ &= \int_0^\infty e^{-x(1-t)} \, \mathrm{d}x \\ &= \frac{1}{1-t}, \ t < 1 \end{aligned}$$

If we write

$$\frac{1}{1-t} = \sum_{n=0}^\infty t^n = \sum_{n=0}^\infty n! \frac{t^n}{n!}$$

we get immediately that

$$E(X^n) = n!$$

Now take $Y \sim \text{Expo}(1)$ and let $X = \lambda Y \sim \text{Expo}(1)$. So $Y^n = \frac{X^n}{\lambda^n}$, and hence

$$E(Y^n) = \frac{n!}{\lambda^n}$$

Observation 19.2. Let $Z \sim \mathcal{N}(0,1)$, and let us determine all its moments. We know that for n odd, by symmetry,

$$E(Z^n) = 0$$

We previously showed that

$$M(t) = e^{t^2/2}$$

But we can write

$$\sum_{n=0}^\infty \frac{(t^2/2)^n}{n!} = \sum_{n=0}^\infty \frac{t^{2n}}{2^n n!} = \sum_{n=0}^\infty \frac{(2n)!}{2^n n!} \frac{t^{2n}}{(2n)!}$$

So

$$E(Z^{2n}) = \frac{(2n)!}{2^n n!}$$

Observation 19.3. Let $X \sim \text{Pois}(\lambda)$. By LOTUS, its MGF is given by

$$\begin{aligned} E(e^{tX}) &= \sum_{k=0}^\infty e^{tk} e^{-\lambda} \frac{\lambda^k}{k!} \\ &= e^{-\lambda} e^{\lambda e^t} \\ &= e^{\lambda(e^t - 1)} \end{aligned}$$

Observation 19.4. Now let $X \sim \text{Pois}(\lambda)$ and $Y \sim \text{Pois}(\mu)$ independent. We want to find the distribution of $X + Y$. We can simply multiply their MGFs, yielding

$$\begin{aligned} M_X(t) M_Y(t) &= e^{\lambda(e^t - 1)} e^{\mu(e^t - 1)} \\ &= e^{(\lambda + \mu)(e^t - 1)} \end{aligned}$$

Thus, $X + Y \sim \text{Pois}(\lambda + \mu)$.

Example. Suppose X, Y above are dependent; specifically, take $X = Y$. Then $X + Y = 2X$. But this cannot be Poisson since it only takes on even values. We could also compute the mean and variance

$$E(2X) = 2\lambda \qquad \text{Var}(2X) = 4\lambda$$

but they should be equal for the Poisson.

We now turn to the study of joint distributions. Recall that joint distributions for independent random variables can be given simply by multiplying their CDFs; we want also to study cases where random variables are not independent.

Definition 19.5. Let X, Y be random variables. Their joint CDF is given by

$$F(x,y) = P(X \le x, Y \le y)$$

In the discrete case, X and Y have a joint PMF given by

$$P(X = x, Y = y)$$

and in the continuous case, X and Y have a joint PDF given by

$$f(x,y) = \frac{\partial}{\partial x \partial y} F(x,y)$$

and we can compute

$$P((X,Y) \in B) = \iint_B f(x,y) \, \mathrm{d}x \, \mathrm{d}y$$

Their separate CDFs and PMFs (e.g., $P(X \le x)$) are referred to as marginal CDFs, PMFs, or PDFs. X and Y are independent precisely when the the joint CDF is equal to the product of the marginal CDFs:

$$F(x,y) = F_X(x) F_Y(y)$$

We can show that, equivalently, we can say

$$P(X = x, Y = y) = P(X = x)P(Y = y)$$

or

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

for all $x, y \in \mathbb{R}$.

Definition 19.6. To get the marginal PMF or PDF of a random variable X from its joint PMF or PDF with another random variable Y, we can marginalize over Y by computing

$$P(X = x) = \sum_y P(X = x, Y = y)$$

or

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

Example. Let $X \sim \text{Bern}(p)$, $X \sim \text{Bern}(q)$. Suppose they have joint PMF given by

	$Y = 0$	$Y = 1$	
$X = 0$	2/6	1/6	3/6
$X = 1$	2/6	1/6	3/6
	4/6	2/6	

Here we have computed the marginal probabilities (in the margin), and they demonstrate that X and Y are independent.

Example. Let us define the uniform distribution on the unit square, $\{(x, y) : x, y \in [0, 1]\}$. We want the joint PDF to be constant everywhere in the square and 0 otherwise; that is,

$$f(x, y) = \begin{cases} c & 0 \le x \le 1, 0 \le y \le 1 \\ 0 & \text{otherwise} \end{cases}$$

Normalizing, we simply need $c = \frac{1}{\text{area}} = 1$. It is apparent that the marginal PDFs are both uniform.

Example. Let us define the uniform distribution on the unit disc, $\{(x, y) : x^2 + y^2 \le 1\}$. Their joint PDF can be given by

$$f(x, y) = \begin{cases} \frac{1}{\pi} & x^2 + y^2 \le 1 \\ 0 & \text{otherwise} \end{cases}$$

Given $X = x$, we have $-\sqrt{1 - x^2} \le y \le \sqrt{1 - x^2}$. We might guess that Y is uniform, but clearly X and Y are dependent in this case, and it turns out that this is not the case.

Lecture 20 — 10/21/11

Definition 20.1. Let X and Y be random variables. Then the conditional PDF of $Y|X$ is

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

Note that $Y|X$ is shorthand for $Y \mid X = x$.

Example. Recall the PDF for our uniform distribution on the disk,

$$f(x, y) = \begin{cases} \frac{1}{\pi} & y^2 = 1 - x^2 \\ 0 & \text{otherwise} \end{cases}$$

and marginalizing over Y, we have

$$f_X(x) = \int_{-\sqrt{1-x^2}}^{\sqrt{1-x^2}} \frac{1}{\pi}\,\mathrm{d}y = \frac{2}{\pi}\sqrt{1 - x^2}$$

for $-1 \le x \le 1$. As a check, we could integrate this again with respect to $\mathrm{d}x$ to ensure that it is 1. From this, it is easy to find the conditional PDF,

$$f_{Y|X}(y|x) = \frac{1/\pi}{\frac{2}{\pi}\sqrt{1 - x^2}} = \frac{1}{2\sqrt{1 - x^2}}$$

for $-\sqrt{1 - x^2} \le y \le \sqrt{1 - x^2}$. Since we are holding x constant, we see that $Y|X \sim \text{Unif}(-\sqrt{1 - x^2}, \sqrt{1 - x^2})$. From these computations, it is clear, in many ways, that X and Y are not independent. It is not true that $f_{X,Y} = f_X f_Y$, nor that $f_{Y|X} = f_Y$

Proposition 20.2. *Let X, Y have joint PDF f, and let $g : \mathbb{R}^2 \to \mathbb{R}$. Then*

$$E(g(X, Y)) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x, y)f(x, y)\,\mathrm{d}x\,\mathrm{d}y$$

This is LOTUS in two dimensions.

Theorem 20.3. *If X, Y are independent random variables, then $E(XY) = E(X)E(Y)$.*

Proof. We will prove this in the continuous case. Using LOTUS, we have

$$\begin{aligned} E(XY) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xyf_{X,Y}(x, y)\,\mathrm{d}x\,\mathrm{d}y \\ \text{by independence} \quad &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xyf_X(x)f_Y(y)\,\mathrm{d}x\,\mathrm{d}y \\ &= \int_{-\infty}^{\infty} E(X)yf_Y(y)\,\mathrm{d}y \\ &= E(X)E(Y) \end{aligned}$$

as desired. ■

Example. Let $X, Y \sim \text{Unif}(0,1)$ i.i.d.; we want to find $E|X-Y|$. By LOTUS (and since the joint PDF is 1), we want to integrate

$$\begin{aligned} E|X-Y| &= \int_0^1 \int_0^1 |x-y| \,\mathrm{d}x \,\mathrm{d}y \\ &= \iint_{x>y} (x-y) \,\mathrm{d}x \,\mathrm{d}y + \iint_{x\le y} (y-x) \,\mathrm{d}x \,\mathrm{d}y \\ \text{by symmetry} &= 2\iint_{x>y} (x-y) \,\mathrm{d}x \,\mathrm{d}y \\ &= \int_0^1 \int_y^1 (x-y) \,\mathrm{d}x \,\mathrm{d}y \\ &= 2\int_0^1 \left(\frac{x^2}{2} - yx\right)\Big|_y^1 \mathrm{d}y \\ &= \frac{1}{3} \end{aligned}$$

If we let $M = \max\{X,Y\}$ and $L = \min\{X,Y\}$, then we would have $|X-Y| = M - L$, and hence also

$$E(M-L) = E(M) - E(L) = \frac{1}{3}$$

We also have

$$E(X+Y) = E(M+L) = E(M) + E(L) = 1$$

This gives

$$E(M) = \frac{2}{3} \quad E(L) = \frac{1}{3}$$

Example (Chicken-Egg Problem)**.** Suppose there are $N \sim \text{Pois}(\lambda)$ eggs, each hatching with probability p, independent (these are Bernoulli). Let X be the number of eggs that hatch. Thus, $X|N \sim \text{Bin}(N,p)$. Let Y be the number that don't hatch. Then $X + Y = N$.

Let us find the joint PMF of X and Y.

$$\begin{aligned} P(X=i,Y=j) &= \sum_{n=0}^{\infty} P(X=i,Y=j \mid N=n)P(N=n) \\ &= P(X=i,Y=j \mid N=i+j)P(N=i+j) \\ &= P(X=i \mid N=i+j)P(N=i+j) \\ &= \frac{(i+j)!}{i!j!} p^i q^j e^{-\lambda} \frac{\lambda^{i+j}}{(i+j)!} \\ &= \left(e^{-\lambda p}\frac{(\lambda p)^i}{i!}\right)\left(e^{-\lambda q}\frac{(\lambda q)^j}{j!}\right) \end{aligned}$$

In other words, the randomness of the number of eggs offsets the dependence of Y on X given a fixed number of X. This is a special property of the Poisson distribution.

Lecture 21 — 10/24/11

Theorem 21.1. *Let $X \sim \mathcal{N}(\mu_1, \sigma_1^2)$ and $Y \sim \mathcal{N}(\mu_2, \sigma_2^2)$ be independent random variables. Then $X+Y \sim \mathcal{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$.*

Proof. Since X and Y are independent, we can simply multiply their MGFs. This is given by

$$\begin{aligned} M_{X+Y}(t) &= M_X(t)M_Y(t) \\ &= \exp(\mu_1 t + \sigma_1 \frac{t^2}{2}) \exp(\mu_2 t + \sigma_2 \frac{t^2}{2}) \\ &= \exp((\mu_1+\mu_2)t + (\sigma_1^2+\sigma_2^2)\frac{t^2}{2}) \end{aligned}$$

which yields our desired result. ■

Example. Let $Z_1, Z_2 \sim \mathcal{N}(0,1)$, i.i.d.; let us find $E|Z_1 - Z_2|$. By the above, $Z_1 - Z_2 \sim \mathcal{N}(0,2)$. Let $Z \sim \mathcal{N}(0,1)$. Then

$$\begin{aligned} E|Z_1 - Z_2| &= E|\sqrt{2}Z| \\ &= \sqrt{2}E|Z| \\ &= \int_{-\infty}^{\infty} |z| \frac{1}{\sqrt{2\pi}} e^{-z^2/2} \,\mathrm{d}z \\ \text{evenness} \quad &= 2\int_0^{\infty} |z| \frac{1}{\sqrt{2\pi}} e^{-z^2/2} \,\mathrm{d}z \\ &= \sqrt{\frac{2}{\pi}} \end{aligned}$$

Definition 21.2. Let $\mathbf{X} = (X_1, \ldots, X_k)$ be a multivariate random variable, $\mathbf{p} = (p_1, \ldots, p_k)$ a probability vector with $p_j \ge 0$ and $\sum_j p_j = 1$. The multinomial distribution is given by assorting n objects into k categories, each object having probability p_j of being in category j, and taking the number of objects in each category, X_j. If X has the multinomial distribution, we write $X \sim \text{Mult}_k(n, \mathbf{p})$.

The PMF of X is given by

$$P(X_1 = n_1, \ldots, X_k = n_k) = \frac{n!}{n_1! \cdots n_k!} p_1^{n_1} \cdots p_k^{n_k}$$

if $\sum_k n_k = n$, and 0 otherwise.

Observation 21.3. Let $X \sim \text{Mult}_k(n,p)$. Then the marginal distribution of X_j is simply $X_j \sim \text{Bin}(n, p_j)$, since each object is either in j or not, and we have

$$E(X_j) = np_j \quad \text{Var}(X_j) = np_j(1-p_j)$$

Observation 21.4. If we "lump" some of our categories together for $X \sim \text{Mult}_k(n,p)$, then the result is still multinomial. That is, taking

$$Y = (X_1, \ldots, X_{l-1}, X_l + \cdots + X_k)$$

and

$$p' = (p_1, \ldots, p_{l-1}, p_l + \cdots + p_k)$$

we have $Y \sim \text{Mult}_l(n, p')$, and this is true for any combinations of lumpings.

Observation 21.5. Let $X \sim \text{Mult}_k(n, p)$. Then given $X_1 = n_1$,

$$(X_2, \ldots, X_k) \sim \text{Mult}_{k-1}(n - n_1, (p_2', \ldots, p_k'))$$

where

$$p_j' = \frac{p_j}{1 - p_1} = \frac{p_j}{p_2 + \cdots + p_k}$$

This is symmetric for all j.

Definition 21.6. The Cauchy distribution is a distribution of $T = \frac{X}{Y}$ with $X, Y \sim \mathcal{N}(0, 1)$ i.i.d.

Note. The Cauchy distribution has no mean, but has the property that an average of many Cauchy distributions is still Cauchy.

Observation 21.7. Let us compute the PDF of X with the Cauchy distribution. The CDF is given by

$$\begin{aligned}
P(\frac{X}{Y} \le t) &= P(\frac{X}{|Y|} \le t) \\
&= P(X \le t|Y|) \\
&= \int_{-\infty}^{\infty} \int_{-\infty}^{t|y|} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \frac{1}{\sqrt{2\pi}} e^{-y^2/2} \, \mathrm{d}x \, \mathrm{d}y \\
&= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-y^2/2} \int_{-\infty}^{t|y|} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \, \mathrm{d}x \, \mathrm{d}y \\
&= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-y^2/2} \Phi(t|y|) \, \mathrm{d}y \\
&= \sqrt{\frac{2}{\pi}} \int_{0}^{\infty} e^{-y^2/2} \Phi(ty) \, \mathrm{d}y
\end{aligned}$$

There is little we can do to compute this integral. Instead, let us compute the PDF, calling the CDF above $F(t)$. Then we have

$$\begin{aligned}
F'(t) &= \sqrt{\frac{2}{\pi}} \int_0^{\infty} y e^{-y^2/2} \frac{1}{\sqrt{2\pi}} e^{-t^2 y^2/2} \, \mathrm{d}y \\
&= \frac{1}{\pi} \int_0^{\infty} y e^{-(1+t^2)y^2/2} \, \mathrm{d}y \\
\text{Substituting } u = \frac{(1+t^2)y^2}{2} &\implies \mathrm{d}u = (1+t^2) y \, \mathrm{d}y, \\
&= \frac{1}{\pi(1+t)^2}
\end{aligned}$$

We could also have performed this computation using the Law of Total Probability. Let ϕ be the standard normal PDF. We have

$$\begin{aligned}
P(X \le t|Y|) &= \int_{-\infty}^{\infty} P(X \le t|Y| \mid Y = y) \phi(y) \, \mathrm{d}y \\
\text{by independence} \quad &= \int_{-\infty}^{\infty} P(X \le ty) \phi(y) \, \mathrm{d}y \\
&= \int_{-\infty}^{\infty} \Phi(ty) \phi(y) \, \mathrm{d}y
\end{aligned}$$

and then we proceed as before.

Lecture 22 — 10/26/11

Definition 22.1. The covariance of random variables X and Y is

$$\text{Cov}(X, Y) = E((X - EX)(Y - EY))$$

Note. The following properties are immediately true of covariance:

1. $\text{Cov}(X, X) = \text{Var}(X)$
2. $\text{Cov}(X, Y) = \text{Cov}(Y, X)$
3. $\text{Cov}(X, Y) = E(XY) - E(X)E(Y)$
4. $\forall c \in \mathbb{R}$, $\text{Cov}(X, c) = 0$
5. $\forall c \in \mathbb{R}$, $\text{Cov}(cX, Y) = c \, \text{Cov}(X, Y)$
6. $\text{Cov}(X, Y + Z) = \text{Cov}(X, Y) + \text{Cov}(X, Z)$

The last two properties demonstrate that covariance is bilinear. In general,

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

Observation 22.2. We can use covariance to compute the variance of sums:

$$\begin{aligned}
\text{Var}(X + Y) &= \text{Cov}(X, X) + \text{Cov}(X, Y) \\
&\quad + \text{Cov}(Y, X) + \text{Cov}(Y, Y) \\
&= \text{Var}(X) + 2 \, \text{Cov}(X, Y) + \text{Var}(Y)
\end{aligned}$$

and more generally,

$$\text{Var}(\sum X) = \sum \text{Var}(X) + 2 \sum_{i<j} \text{Cov}(X_i, X_j)$$

Theorem 22.3. *If X, Y are independent, then $\text{Cov}(X, Y) = 0$ (we say that they are uncorrelated).*

Example. The converse of the above is false. Let $Z \sim \mathcal{N}(0, 1)$, $X = Z$, $Y = Z^2$, and let us compute the covariance.

$$\begin{aligned}
\text{Cov}(X, Y) &= E(XY) - (EX)(EY) \\
&= E(Z^3) - (EZ)(EZ^2) \\
&= 0
\end{aligned}$$

E(Z3) -> odd power of normal is zero]

But X and Y are very dependent, since Y is a function of X.

E(Z) and E(Z2) are zero(given)

Definition 22.4. The correlation of two random variables X and Y is

$$\begin{aligned}\operatorname{Cor}(X,Y) &= \frac{\operatorname{Cov}(X,Y)}{\operatorname{SD}(X)\operatorname{SD}(Y)} \\ &= \operatorname{Cov}\left(\frac{X-EX}{\operatorname{SD}(X)}, \frac{Y-EY}{\operatorname{SD}(Y)}\right)\end{aligned}$$

The operation of

$$\frac{X-EX}{\operatorname{SD}(X)}$$

is called standardization; it gives the result a mean of 0 and a variance of 1.

Theorem 22.5. $|\operatorname{Cor}(X,Y)| \leq 1$.

Proof. We could apply Cauchy-Schwartz to get this result immediately, but we shall also provide a direct proof. WLOG, assume X and Y are standardized. Let $\rho = \operatorname{Cor}(X,Y)$. We have

$$\operatorname{Var}(X+Y) = \operatorname{Var}(X) + \operatorname{Var}(Y) + 2\rho = 2 + 2\rho$$

and we also have

$$\operatorname{Var}(X-Y) = \operatorname{Var}(X) + \operatorname{Var}(Y) - 2\rho = 2 - 2\rho$$

But since $\operatorname{Var} \geq 0$, this yields our result. ■

Example. Let $(X_1, \ldots, X_k) \sim \operatorname{Mult}_k(n,p)$. We shall compute $\operatorname{Cov}(X_i, X_j)$ for all i, j. If $i = j$, then

$$\operatorname{Cov}(X_i, X_i) = \operatorname{Var}(X_i) = np_i(1-p_i)$$

Suppose $i \neq j$. We can expect that the covariance will be negative, since more objects in category i means less in category j. We have

$$\operatorname{Var}(X_i + X_j) = np_i(1-p_i) + np_j(1-p_j) + 2\operatorname{Cov}(X_i, X_j)$$

But by "lumping" i and j together, we also have

$$\operatorname{Var}(X_i + X_j) = n(p_i+p_j)(1-(p_i+p_j))$$

Then solving for c, we have

$$\operatorname{Cov}(X_i, X_j) = -np_ip_j$$

Note. Let A be an event and I_A its indicator random variable. It is clear that

$$I_A^n = I_A$$

for any $n \in \mathbb{N}$. It is also clear that

$$I_A I_B = I_{A \cap B}$$

Example. Let $X \sim \operatorname{Bin}(n,p)$. Write $X = X_1 + \cdots + X_n$ where the X_j are i.i.d. $\operatorname{Bern}(p)$. Then

$$\begin{aligned}\operatorname{Var}(X_j) &= EX_j^2 - (EX_j)^2 \\ &= p - p^2 \\ &= p(1-p)\end{aligned}$$

It follows that

$$\operatorname{Var}(X) = np(1-p)$$

since $\operatorname{Cor}(X_i, X_j) = 0$ for $i \neq j$ by independence.

Lecture 23 — 10/28/11

Example. Let $X \sim \operatorname{HGeom}(w,b,n)$. Let us write $p = \frac{w}{w+b}$ and $N = w+b$. Then we can write $X = X_1 + \cdots + X_n$ where the X_j are $\operatorname{Bern}(p)$. (Note, however, that unlike with the binomial, the X_j are not independent.) Then

$$\begin{aligned}\operatorname{Var}(X) &= n\operatorname{Var}(X_1) + 2\binom{n}{2}\operatorname{Cov}(X_1, X_2) \\ &= np(1-p) + 2\binom{n}{2}\operatorname{Cov}(X_1, X_2)\end{aligned}$$

Computing the covariance, we have

$$\begin{aligned}\operatorname{Cov}(X_1, X_2) &= E(X_1X_2) - (EX_1)(EX_2) \\ &= \frac{w}{w+b}\left(\frac{w-1}{w+b-1}\right) - \left(\frac{w}{w+b}\right)^2 \\ &= \frac{w}{w+b}\left(\frac{w-1}{w+b-1}\right) - p^2\end{aligned}$$

and simplifying,

$$\operatorname{Var}(X) = \frac{N-n}{N-1}np(1-p)$$

The term $\frac{N-n}{N-1}$ is called the finite population correction; it represents the "offset" from the binomial due to lack of replacement.

Theorem 23.1. *Let X be a continuous random variable with PDF f_X, and let $Y = g(X)$ where g is differentiable and strictly increasing. Then the PDF of Y is given by*

$$f_Y(y) = f_X(x)\frac{\mathrm{d}x}{\mathrm{d}y}$$

where $y = g(x)$ and $x = g^{-1}(y)$. (Also recall from calculus that $\frac{\mathrm{d}x}{\mathrm{d}y} = \left(\frac{\mathrm{d}y}{\mathrm{d}x}\right)^{-1}$.)

Proof. From the CDF of Y, we get

$$P(Y \leq y) = P(g(X) \leq y)$$

$$\begin{aligned} &= P(X \le g^{-1}(y)) \\ &= F_X(g^{-1}(y)) \\ &= F_X(x) \end{aligned}$$

Then, differentiating, we get by the Chain Rule that

$$f_Y(y) = f_X(x)\frac{\mathrm{d}x}{\mathrm{d}y}$$ ■

Example. Consider the log normal distribution, which is given by $Y = e^Z$ for $Z \sim \mathcal{N}(0,1)$. We have

$$f_Y(y) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}$$

To put this in terms of y, we substitute $z = \ln y$. Moreover, we know that

$$\frac{\mathrm{d}y}{\mathrm{d}z} = e^z = y$$

and so,

$$f_Y(y) = \frac{1}{y}\frac{1}{\sqrt{2\pi}} e^{-\ln y/2}$$

Theorem 23.2. *Suppose that X is a continuous random variable in n dimensions, $Y = g(X)$ where $g : \mathbb{R}^n \to \mathbb{R}^n$ is continuously differentiable and invertible. Then*

$$f_Y(y) = f_X(x)\left|\det \frac{\mathrm{d}x}{\mathrm{d}y}\right|$$

where

$$\frac{\mathrm{d}x}{\mathrm{d}y} = \begin{pmatrix} \frac{\partial x_1}{\partial y_1} & \cdots & \frac{\partial x_n}{\partial y_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial x_n}{\partial y_1} & \cdots & \frac{\partial x_n}{\partial y_n} \end{pmatrix}$$

is the Jacobian matrix.

Observation 23.3. Let $T = X + Y$, where X and Y are independent. In the discrete case, we have

$$P(T = t) = \sum_x P(X = x)P(Y = t - x)$$

For the continuous case, we have

$$\begin{aligned} f_T(t) &= (f_X * f_Y)(t) \\ &= \int_{-\infty}^{\infty} f_X(x) f_Y(t - x)\,\mathrm{d}x \end{aligned}$$

This is true because we have

$$\begin{aligned} F_T(t) &= P(T \le t) \\ &= \int_{-\infty}^{\infty} P(X + Y \le t \mid X = x) f_X(x)\,\mathrm{d}x \\ &= \int_{-\infty}^{\infty} F_Y(t - x) f_X(x)\,\mathrm{d}x \end{aligned}$$

Then taking the derivative of both sides,

$$f_T(t) = \int_{-\infty}^{\infty} f_X(x) f_Y(t - x)\,\mathrm{d}x$$

We now briefly turn our attention to proving the existence of objects with some desired property A using probability. We want to show that $P(A) > 0$ for some random object, which implies that some such object must exist.

Reframing this question, suppose each object in our universe of objects has some kind of "score" associated with this property; then we want to show that there is some object with a "good" score. But we know that there is an object with score at least equal to the average score, i.e., the score of a random object. Showing that this average is "high enough" will prove the existence of an object without specifying one.

Example. Suppose there are 100 people in 15 committees of 20 people each, and that each person is on exactly 3 committees. We want to show that there exist 2 committees with overlap ≥ 3. Let us find the average of two random committees. Using indicator random variables for the probability that a given person is on both of those two committees, we get

$$E(\text{overlap}) = 100 \cdot \frac{\binom{3}{2}}{\binom{15}{2}} = \frac{300}{105} = \frac{20}{7}$$

Then there exists a pair of committees with overlap of at least $\frac{20}{7}$. But since all overlaps must be integral, there is a pair of committees with overlap ≥ 3.

Lecture 24 — 10/31/11

Definition 24.1. The beta distribution, $\text{Beta}(a, b)$ for $a, b > 0$, is defined by PDF

$$f(x) = \begin{cases} cx^{a-1}(1-x)^{b-1} & 0 < x < 1 \\ 0 & \text{otherwise} \end{cases}$$

where c is a normalizing constant (defined by the beta function).

The beta distribution is a flexible family of continuous distributions on $(0, 1)$. By flexible, we mean that the appearance of the distribution varies significantly depending on the values of its parameters. If $a = b = 1$, the beta reduces to the uniform. If $a = 2$ and $b = 1$, the beta appears as a line with positive slope. If $a = b = \frac{1}{2}$, the beta appears to be concave-up and parabolic; if $a = b = 2$, it is concave down.

The beta distribution is often used as a *prior* distribution for some parameter on $(0, 1)$. In particular, it is the *conjugate prior* to the binomial distribution.

Observation 24.2. Suppose that, based on some data, we have $X \mid p \sim \text{Bin}(n, p)$, and that our prior distribution for p is $p \sim \text{Beta}(a, b)$. We want to determine the posterior distribution of p, $p \mid X$. We have

$$\begin{aligned} f(p \mid X = k) &= \frac{P(X = k \mid p) f(p)}{P(X = k)} \\ &= \frac{\binom{n}{k} p^k (1-p)^{n-k} c p^{a-1} (1-p)^{b-1}}{P(X = k)} \\ &\propto c p^{a+k-1} (1-p)^{b+n-k-1} \end{aligned}$$

So, we have $p \mid X \sim \text{Beta}(a + X, b + n - X)$. We call the Beta the conjugate prior to the binomial because both its prior and posterior distribution are Beta.

Observation 24.3. Let us find a specific case of the normalizing constant

$$c^{-1} = \int_0^1 x^k (1-x)^{n-k} \, \mathrm{d}x$$

To do this, consider the story of "Bayes' billiards." Suppose we have $n+1$ billiard balls, all white; then we paint one pink and throw them along $(0, 1)$ all independently. Let X be the number of balls to the left of the pink ball. Then conditioning on where the pink ball ends up, we have

$$\begin{aligned} P(X = k) &= \int_0^1 P(X = k \mid p) \underbrace{f(p)}_{1} \, \mathrm{d}p \\ &= \int_0^1 \binom{n}{k} p^k (1-p)^{n-k} \, \mathrm{d}p \end{aligned}$$

where, given the pink ball's location, X is simply binomial (each white ball has an independent chance p of landing to the left). Note, however, that painting a ball pink and then throwing the balls along $(0, 1)$ is the same as throwing the balls along the real line and then painting one pink. But then it is clear that there is an equal chance for any given number from 0 to n of white balls to be to the pink ball's left. So we have

$$\int_0^1 \binom{n}{k} p^k (1-p)^{n-k} \, \mathrm{d}p = \frac{1}{n+1}$$

Lecture 25 — 11/2/11

Definition 25.1. The gamma function is given by

$$\begin{aligned} \Gamma(a) &= \int_0^\infty x^{a-1} e^{-x} \, \mathrm{d}x \\ &= \int_0^\infty x^a e^{-x} \frac{1}{x} \, \mathrm{d}x \end{aligned}$$

for any $a > 0$. The gamma function is a continuous extension of the factorial operator on natural numbers. For n a positive integer,

$$\Gamma(n) = (n-1)!$$

More generally,

$$\Gamma(x+1) = x\Gamma(x)$$

Definition 25.2. The standard gamma distribution, $\text{Gamma}(a, 1)$, is defined by PDF

$$\frac{1}{\Gamma(a)} x^{a-1} e^{-x}$$

for $x > 0$, which is simply the integrand of the normalized Gamma function. More generally, let $X \sim \text{Gamma}(a, 1)$ and $Y = \frac{X}{\lambda}$. We say that $Y \sim \text{Gamma}(a, \lambda)$. To get the PDF of Y, we simply change variables; we have $x = \lambda y$, so

$$\begin{aligned} f_Y(y) &= f_X(x) \frac{\mathrm{d}x}{\mathrm{d}y} \\ &= \frac{1}{\Gamma(a)} (\lambda y)^a e^{-\lambda y} \frac{1}{\lambda y} \lambda \\ &= \frac{1}{\Gamma(a)} (\lambda y)^a e^{-\lambda y} \frac{1}{y} \end{aligned}$$

Definition 25.3. We define a Poisson process as a process in which events occur continuously and independently such that in any time interval t, the number of events which occur is $N_t \sim \text{Pois}(\lambda t)$ for some fixed rate parameter λ.

Observation 25.4. The time T_1 until the first event occurs is $\text{Expo}(\lambda)$:

$$P(T_1 > t) = P(N_t = 0) = e^{-\lambda t}$$

which means that

$$P(T_1 \le t) = 1 - e^{-\lambda t}$$

as desired. More generally, the time until the next event is always $\text{Expo}(\lambda)$; this is clear from the memoryless property.

Proposition 25.5. *Let T_n be the time of the nth event in a Poisson process with rate parameter λ. Then, for X_j i.i.d.* $\text{Expo}(\lambda)$*, we have*

$$T_n = \sum_{j=1}^n X_j \sim \text{Gamma}(n, \lambda)$$

The exponential distribution is the continuous analogue of the geometric distribution; in this sense, the gamma distribution is the continuous analogue of the negative binomial distribution.

Proof. One method of proof, which we will not use, would be to repeatedly convolve the PDFs of the i.i.d. X_j. Instead, we will use MGFs. Suppose that the X_j are i.i.d Expo(1); we will show that their sum is Gamma$(n, 1)$.

The MGF of X_j is given by

$$M_{X_j}(t) = \frac{1}{1-t}$$

for $t < 1$. Then the MGF of T_n is

$$M_{T_n}(t) = \left(\frac{-1}{1}\right)^n$$ (1/(1-t))n <- given expression is wrong

also for $t < 1$. We will show that the gamma distribution has the same MGF.

Let $Y \sim \text{Gamma}(n, 1)$. Then by LOTUS,

$$\begin{aligned} E(e^{tY}) &= \frac{1}{\Gamma(n)} \int_0^\infty e^{ty} y^n e^{-y} \frac{1}{y} \, \mathrm{d}y \\ &= \frac{1}{\Gamma(n)} \int_0^\infty y^n e^{(1-t)y} \frac{1}{y} \, \mathrm{d}y \end{aligned}$$

Changing variables, with $x = (1-t)y$, then

$$\begin{aligned} E(e^{tY}) &= \frac{(1-t)^{-n}}{\Gamma(n)} \int_0^\infty x^n e^{-x} \frac{1}{x} \, \mathrm{d}x \\ &= \left(\frac{-1}{1}\right)^n \frac{\Gamma(n)}{\Gamma(n)} \\ &= \left(\frac{-1}{1}\right)^n \end{aligned}$$

■

Note that this is the MGF for any $n > 0$, although the sum of exponentials expression requires integral n.

Observation 25.6. Let us compute the moments of $X \sim \text{Gamma}(a, 1)$. We want to compute $E(X^c)$. We have

$$\begin{aligned} E(X^c) &= \frac{1}{\Gamma(a)} \int_0^\infty x^c x^a e^{-x} \frac{1}{x} \, \mathrm{d}x \\ &= \frac{1}{\Gamma(a)} \int_0^\infty x^{a+c} e^{-x} \frac{1}{x} \, \mathrm{d}x \\ &= \frac{\Gamma(a+c)}{\Gamma(a)} \\ &= \frac{a(a+1)\cdots(a+c)\Gamma(a)}{\Gamma(a)} \\ &= a(a+1)\cdots(a+c) \end{aligned}$$

If instead, we take $X \sim \text{Gamma}(a, \lambda)$, then we will have

$$E(X^c) = \frac{a(a+1)\cdots(a+c)}{\lambda^c}$$

Lecture 26 — 11/4/11

Observation 26.1 (Gamma-Beta)**.** Let us take $X \sim \text{Gamma}(a, \lambda)$ to be your waiting time in line at the bank, and $Y \sim \text{Gamma}(b, \lambda)$ your waiting time in line at the post office. Suppose that X and Y are independent. Let $T = X + Y$; we know that this has distribution Gamma$(a + b, \lambda)$.

Let us compute the joint distribution of T and of $W = \frac{X}{X+Y}$, the fraction of time spent waiting at the bank. For simplicity of notation, we will take $\lambda = 1$. The joint PDF is given by

$$\begin{aligned} f_{T,W}(t, w) &= f_{X,Y}(x, y) \left| \det \frac{\partial(x, y)}{\partial(t, w)} \right| \\ &= \frac{1}{\Gamma(a)\Gamma(b)} x^a e^{-x} y^b e^{-y} \frac{1}{xy} \left| \det \frac{\partial(x, y)}{\partial(t, w)} \right| \end{aligned}$$

We must find the determinant of the Jacobian (here expressed in silly-looking notation). We know that

$$x + y = t \qquad \frac{x}{x+y} = w$$

Solving for x and y, we easily find that

$$x = tw \qquad y = t(1-w)$$

Then the determinant of our Jacobian is given y

$$\begin{vmatrix} w & t \\ 1-w & -t \end{vmatrix} = -tw - t(1-w) = -t$$

Taking the absolute value, we then get

$$\begin{aligned} f_{T,W}(t, w) &= \frac{1}{\Gamma(a)\Gamma(b)} x^a e^{-x} y^b e^{-y} \frac{1}{xy} t \\ &= \frac{1}{\Gamma(a)\Gamma(b)} w^{a-1} (1-w)^{b-1} t^{a+b} e^{-t} \frac{1}{t} \\ &= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} w^{a-1} (1-w)^{b-1} \frac{1}{\Gamma(a+b)} t^{a+b} e^{-t} \frac{1}{t} \end{aligned}$$

This is a product of some function of w with the PDF of T, so we see that T and W are independent. To find the marginal distribution of W, we note that the PDF of T integrates to 1 just like any PDF, so we have

$$\begin{aligned} f_W(w) &= \int_{-\infty}^\infty f_{T,W}(t, w) \, \mathrm{d}t \\ &= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} w^{a-1} (1-w)^{b-1} \end{aligned}$$

This yields $W \sim \text{Beta}(a, b)$ and also gives the normalizing constant of the beta distribution.

It turns out that if X were distributed according to any other distribution, we would not have independence, but proving so is out of the scope of the course.

Observation 26.2. Let us find $E(W)$ for $W \sim \text{Beta}(a,b)$. Let us write $W = \frac{X}{X+Y}$ with X and Y defined as above. We have

wrong expression -> $E(W)=E(X/X+Y) = E\left(\frac{-1}{X}\right) = \frac{E(X)}{E(X+Y)} = \frac{a}{a+b}$

Note that in general, the first equality is false! However, because $X+Y$ and $\frac{X}{X+Y}$, they are uncorrelated and hence linear. So

$$E\left(\frac{-1}{X}\right)E(X+Y) = E(X)$$

Definition 26.3. Let $X_1, \dots, X_n$ be i.i.d. The order statistics of this sequence is

$$X_{(1)} \le X_{(2)} \le \cdots \le X_{(n)}$$

where

$$X_{(1)} = \min\{X_1, \dots, X_n\}$$
$$X_{(n)} = \max\{X_1, \dots, X_n\}$$

and the remaining $X_{(j)}$ fill out the order. If n is odd, we have the median $X_{(\frac{-1}{n+1})}$. The order statistics lets us find arbitrary quantiles for the sequence.

The order statistics are hard to work with because they are dependent (and positively correlated), even though we started with i.i.d. random variables. They are particularly tricky in the discrete case because of ties, so we will assume that the X_j are continuous.

Observation 26.4. Let $X_1, \dots, X_n$ be i.i.d. continuous with PDF f_j and CDF F_j. We want to find the CDF and PDF of $X_{(j)}$. For the CDF, we have

$$\begin{aligned} P(X_{(j)} \le x) &= P(\text{at least } j \text{ of the } X_j\text{'s are} \le x) \\ &= \sum_{k=j}^{n} \binom{n}{k} P(X_1 \le x)^k (1 - P(X_1 \le x))^{n-k} \\ &= \sum_{k=j}^{n} \binom{n}{k} F(x)^k (1-F(x))^{n-k} \end{aligned}$$

Turning now to the PDF, recall that a PDF gives a density rather than a probability. We can multiply the PDF of $X_{(j)}$ at a point x by a tiny interval $\mathrm{d}x$ about x in order to obtain the probability that $X_{(j)}$ is in that interval. Then we can simply count the number of ways to have one of the X_i be in that interval and precisely $j-1$ of the X_i below the interval. So,

$$f_{X_{(j)}}(x)\,\mathrm{d}x = n(f(x)\,\mathrm{d}x)\binom{n-1}{j-1}F(x)^{j-1}(1-F(x))^{n-j}$$
$$f_{X_{(j)}}(x) = n\binom{n-1}{j-1}F(x)^{j-1}(1-F(x))^{n-j}f(x)$$

Example. Let $U_1, \dots, U_n$ be i.i.d. Unif$(0,1)$, and let us determine the distribution of $U_{(j)}$. Applying the above result, we have

$$f_{U_{(j)}}(x) = n\binom{n-1}{j-1}x^{j-1}(1-x)^{n-j}$$

for $0 < x < 1$. Thus, we have $U_{(j)} \sim \text{Beta}(j, n-j+1)$. This confirms our earlier result that, for U_1 and U_2 i.i.d. Unif$(0,1)$, we have

$$E|U_1 - U_2| = E(U_{\max}) - E(U_{\min}) = \frac{1}{3}$$

because $U_{\max} \sim \text{Beta}(2,1)$ and $U_{\min} \sim \text{Beta}(1,2)$, which have means $\frac{2}{3}$ and $\frac{1}{3}$ respectively.

Lecture 27 — 11/7/11

Example (Two Envelopes Paradox)**.** Suppose we have two envelopes containing sums of money X and Y, and suppose we are told that one envelope has twice as much money as the next. We choose one envelope; by symmetry, take X WLOG. Then it appears that Y has equal probabilities of containing $\frac{X}{2}$ and of $2X$, and thus averages $1.25X$. So it seems that we ought to switch to envelope Y. But then, by the same reasoning, it would seem we ought to switch back to X.

We can argue about this paradox in two ways. First, we can say, by symmetry, that

$$E(X) = E(Y)$$

which is simple and straightforward. We might also, however, try to condition on the value of Y with respect to X using the Law of Total Probability

$$\begin{aligned} E(Y) &= E(Y \mid Y = 2X)P(Y = 2X) \\ &\quad + E(Y \mid Y = \frac{X}{2})P(Y = \frac{X}{2}) \\ &= E(2X)\frac{1}{2} + E(\frac{X}{2})\frac{1}{2} \\ &= \frac{5}{4}E(X) \end{aligned}$$

Assuming that X and Y are not 0 or infinite, these cannot both be correct, and the argument from symmetry is immediately correct.

The flaw in our second argument is that, in general,

$$E(Y \mid Y = Z) \neq E(Z)$$

because we cannot drop the condition that $Y = Z$; we must write

$$E(Y \mid Y = Z) = E(Z \mid Y = Z)$$

In other words, if we let I be the indicator for $Y = 2X$, we are saying that X and I are dependent.

Example (Patterns in coin flips)**.** Suppose we repeatedly flip a fair coin. We want to determine how many flips it takes until HT is observed (including the H and T); similarly, we can ask how many flips it takes to get HH. Let us call these random variables W_{HT} and W_{HH} respectively. Note that, by symmetry,

$$E(W_{HH}) = E(W_{TT})$$

and

$$E(W_{HT}) = E(W_{TH})$$

Let us first consider W_{HT}. This is the time to the first H, which we will call W_1, plus the time W_2 to the next T. Then we have

$$E(W_{HT}) = E(W_1) + E(W_2) = 2 + 2 = 4$$

because $W_i - 1 \sim \text{Geom}(\frac{1}{2})$.

Now let us consider W_{HH}. The distinction here is that no "progress" can be easily made; once we get a heads, we are not decidedly halfway to the goal, because if the next flip is tails, we lose all our work. Instead, we make use of conditional expectation. Let H_i be the event that the ith toss is heads, $T_i = H_i^C$ the event that it is tails. Then

$$\begin{aligned} E(W_{HH}) &= E(W_{HH} \mid H_1)\frac{1}{2} + E(W_{HH} \mid T_1)\frac{1}{2} \\ &= \left(E(W_{HH} \mid H_1, H_2)\frac{1}{2} + E(W_{HH} \mid H_1, T_2)\frac{1}{2}\right)\frac{1}{2} \\ &\quad + (1 + E(W_{HH}))\frac{1}{2} \\ &= \left(1 + (2 + E(W_{HH}))\frac{1}{2}\right)\frac{1}{2} + (1 + E(W_{HH}))\frac{1}{2} \end{aligned}$$

Solving for $E(W_{HH})$ gives

$$E(W_{HH}) = 6$$

So far, we have been conditioning expectations on events. Let X and Y be random variables; then this kind of conditioning includes computing $E(Y \mid X = x)$. If Y is discrete, then

$$E(Y \mid X = x) = \sum_y yP(Y = y \mid X = x)$$

and if Y is continuous,

$$\begin{aligned} E(Y \mid X = x) &= \int_{-\infty}^{\infty} y f_{Y|X=x}(y|x)\,\mathrm{d}y \\ \text{if } X \text{ continuous} \quad &= \int_{-\infty}^{\infty} y\frac{f_{X,Y}(x,y)}{f_X(x)}\,\mathrm{d}y \end{aligned}$$

Definition 27.1. Now let us write

$$g(x) = E(Y \mid X = x)$$

Then

$$E(Y|X) = g(X)$$

So, suppose for instance that $g(x) = x^2$; then $g(X) = X^2$. We can see that $E(Y|X)$ is a random variable and a function of X. This is a conditional expectation.

Example. Let X and Y be i.i.d. Pois(λ). Then

$$\begin{aligned} E(X + Y \mid X) &= E(X|X) + E(Y|X) \\ X \text{ is a function of itself} \quad &= X + E(Y|X) \\ X \text{ and } Y \text{ independent}) \quad &= X + E(Y) \\ &= X + \lambda \end{aligned}$$

Note that, in general,

$$E(h(X) \mid X) = h(X)$$

Now let us determine $E(X \mid X + Y)$. We can do this in two different ways. First, let $T = X + Y$ and let us find the conditional PMF.

$$\begin{aligned} P(X = k \mid T = n) &= \frac{P(T = n \mid X = k)P(X = k)}{P(T = n)} \\ &= \frac{P(Y = n - k)P(X = k)}{P(T = n)} \\ &= \frac{\frac{e^{-\lambda}\lambda^{n-k}}{(n-k)!}e^{-\lambda}\frac{\lambda^k}{k!}}{e^{-2\lambda}\frac{(2\lambda)^n}{n!}} \\ &= \binom{n}{k}\left(\frac{-1}{1}\right)^n \end{aligned}$$

That is, $X \mid T = n \sim \text{Bin}(n, \frac{1}{2})$. Thus, we have

$$E(X \mid T = n) = \frac{n}{2}$$

which means that

$$E(X|T) = \frac{T}{2}$$

In our second method, first we note that

$$E(X \mid X + Y) = E(Y \mid X + Y)$$

by symmetry (since they are i.i.d.). We have

$$\begin{aligned} E(X \mid X + Y) + E(Y \mid X + Y) &= E(X + Y \mid X + Y) \\ &= X + Y \\ &= T \end{aligned}$$

So, without even using the Poisson, $E(X|T) = \frac{T}{2}$.

Proposition 27.2 (Adam's Law)**.** *Let X and Y be random variables. Then*

$$E(E(Y|X)) = E(Y)$$

Lecture 28 — 11/9/11

Example. Let $X \sim \mathcal{N}(0,1)$, $Y = X^2$. Then

$$E(Y|X) = E(X^2|X) = X^2 = Y$$

On the other hand,

$$E(X|Y) = E(X|X^2) = 0$$

since, after observing $X^2 = a$, then $X = \pm\sqrt{a}$ with equal likelihood of being positive or negative (since the standard normal is symmetric about 0). Note that this doesn't mean X and X^2 are independent.

Example. Suppose we have a stick, break off a random piece, and then break off another random piece. We can model this as $X \sim \text{Unif}(0,1)$, $Y|X \sim \text{Unif}(0,X)$. We know that

$$E(Y \mid X = x) = \frac{x}{2}$$

and hence

$$E(Y|X) = \frac{X}{2}$$

Note that

$$E(E(Y|X)) = \frac{1}{4} = E(Y)$$

That is, on average, we take half the stick and then take half of that stick, which matches our intuition.

Proposition 28.1. *Let X and Y be random variables.*

1. *$E(h(X)Y \mid X) = h(X)E(Y|X)$.*
2. *$E(Y|X) = E(Y)$ if X and Y are independent (the converse, however, is not true in general).*
3. *$E(E(Y|X)) = E(Y)$. This is called iterated expectation or Adam's Law; it is usually more useful to think of this as computing $E(Y)$ by choosing a simple X to work with.*
4. *$E((Y - E(Y|X))h(X)) = 0$. In words, the residual (i.e., $Y - E(Y|X)$) is uncorrelated with $h(X)$:*

$$\begin{aligned}&\text{Cov}(Y - E(Y|X), h(X))\\ &= \underbrace{E((Y - E(Y|X))h(X))}_{0} - \underbrace{E(Y - E(Y|X))}_{0}E(h(X))\end{aligned}$$

To better understand (4), we can think of the functions X and Y as vectors (the vector space has inner product $\langle X, Y\rangle = E(XY)$). We can think of $E(Y|X)$ as the projection of Y onto the plane consisting of all functions of X. In this picture, the residual vector $Y - E(Y|X)$ is orthogonal to the plane of all functions of X, and thus $\langle Y - E(Y|X), h(X)\rangle = 0$.

Proof. We will prove all the properties above.

1. Since we know X, we know $h(X)$, and this is equivalent to factoring out at constant (by linearity).
2. Immediate.
3. We will prove the discrete case. Let $E(Y|X) = g(X)$. Then by discrete LOTUS, we have

$$\begin{aligned}Eg(X) &= \sum_x g(x)P(X = x)\\ &= \sum_x E(Y \mid X = x)P(X = x)\\ &= \sum_x \left(\sum_y yP(Y = y \mid X = x)\right)P(X = x)\\ &= \sum_x \left(\sum_y yP(Y = y \mid X = x)P(X = x)\right)\end{aligned}$$

conditional PMF times marginal PMF = joint PMF

$$\begin{aligned}&= \sum_x \sum_y yP(Y = y, X = x)\\ &= \sum_y \sum_x yP(Y = y, X = x)\\ &= \sum_y yP(Y = y)\\ &= E(Y)\end{aligned}$$

4. We have

$$\begin{aligned}&E((Y - E(Y|X))h(X))\\ &\quad= E(Yh(X)) - E(E(Y|X)h(X))\\ &\quad= E(Yh(X)) - E(E(h(X)Y|X))\\ &\quad= E(Yh(X)) - E(Yh(X))\\ &\quad= 0\end{aligned}$$

■

Definition 28.2. We can define the conditional variance much as we did conditional expectation. Let X and Y be random variables. Then

$$\begin{aligned}\text{Var}(Y|X) &= E(Y^2|X) - (E(Y|X))^2\\ &= E((Y - E(Y|X))^2 \mid X)\end{aligned}$$

Proposition 28.3 (Eve's Law)**.**

$$\text{Var}(Y) = E(\text{Var}(Y|X)) + \text{Var}(E(Y|X))$$

Example. Suppose we have three populations, where $X = 1$ is the first, $X = 2$ the second, and $X = 3$ the third, and suppose we know the mean and variance of the height Y of individuals in each of the separate populations. Then Eve's law says we can take the variance of all three means, and add it to the mean of all three variances, to get the total variance.

Example. Suppose we choose a random city and then choose a random sample of n people in that city. Let X be the number of people with a particular disease, and Q the proportion of people in the chosen city with the disease. Let us determine $E(X)$ and $\text{Var}(X)$, assuming $Q \sim \text{Beta}(a, b)$ (a mathematically convenient, flexible distribution).

Assume that $X|Q \sim \text{Bin}(n, Q)$. Then

$$\begin{aligned} E(X) &= E(E(X|Q)) \\ &= E(nQ) \\ &= n\frac{a}{a+b} \end{aligned}$$

and

$$\begin{aligned} \text{Var}(X) &= E(\text{Var}(X|Q)) + \text{Var}(E(X|Q)) \\ &= E(nQ(1-Q)) + n^2 \text{Var}(Q) \end{aligned}$$

We have

$$\begin{aligned} E(Q(1-Q)) &= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \int_0^1 q^a (1-q)^b \, \mathrm{d}q \\ &= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \frac{\Gamma(a+1)\Gamma(b+1)}{\Gamma(a+b+2)} \\ &= \frac{ab\Gamma(a+b)}{(a+b+1)(a+b)\Gamma(a+b)} \\ &= \frac{ab}{(a+b)(a+b+1)} \end{aligned}$$

and

$$\text{Var}(Q) = \frac{\mu(1-\mu)}{a+b+1}$$

where $\mu = \frac{a}{a+b}$. This gives us all the information we need to easily compute $\text{Var}(X)$.

Lecture 29 — 11/14/11

Example. Consider a store with a random number N of customers. Let X_j be the amount the jth customer spends, with $E(X_j) = \mu$ and $\text{Var}(X_j) = \sigma^2$. Assume that $N, X_1, X_2, \ldots$ are independent. We want to determine the mean and variance of

$$X = \sum_{j=1}^{N} X_j$$

We might, at first, mistakenly invoke linearity to claim that $E(X) = N\mu$. But this is incoherent; the LHS is a real number whereas the RHS is a random variable. However, this error highlights something useful: we want to make N a constant, so let us condition on N. Then using the Law of Total Probability, we have

$$\begin{aligned} E(X) &= \sum_{n=0}^{\infty} E(X \mid N = n) P(N = n) \\ &= \sum_{n=0}^{\infty} \mu n P(N = n) \\ &= \mu E(N) \end{aligned}$$

Note that we can drop the conditional because N and the X_j are independent; otherwise, this would not be true.

We could also apply Adam's Law to get

$$E(X) = E(E(X|N)) = E(\mu N) = \mu E(N)$$

To get the variance, we apply Eve's Law to get

$$\begin{aligned} \text{Var}(X) &= E(\text{Var}(X|N)) + \text{Var}(E(X|N)) \\ &= E(N\sigma^2) + \text{Var}(\mu N) \\ &= \sigma^2 E(N) + \mu^2 \text{Var}(N) \end{aligned}$$

We now turn our attention to statistical inequalities.

Theorem 29.1 (Cauchy-Schwartz Inequality)**.**

$$|E(XY)| \leq \sqrt{E(X^2)E(Y^2)}$$

If X and Y are uncorrelated, $E(XY) = (EX)(EY)$, so we don't need inequality.

We will not prove this inequality in general. However, if X and Y have mean 0, then

$$|\,\text{Corr}(X, Y)| = \left| \frac{E(XY)}{\sqrt{E(X^2)E(Y^2)}} \right| \leq 1$$

Theorem 29.2 (Jensen's Inequality)**.** *If $g : \mathbb{R} \to \mathbb{R}$ is convex (i.e., $g'' > 0$), then*

$$Eg(X) \geq g(EX)$$

If g ic concave (i.e., $g'' < 0$), then

$$Eg(X) \leq g(EX)$$

Example. If X is positive, then

$$E(\frac{1}{X}) \geq \frac{1}{EX}$$

and

$$E(\ln X) \leq \ln(EX)$$

Proof. It is true of any convex function g that

$$g(x) \geq a + bx$$

if $a + bx$ is the line tangent to any point $(x_0, g(x_0))$ on the graph of g. Take $x_0 = E(X)$. Then we have

$$\begin{aligned} g(x) &\geq a + bx \\ g(X) &\geq a + bX \\ Eg(X) &\geq E(a + bX) \\ &= a + bE(X) \\ &= g(EX) \end{aligned}$$

■

Theorem 29.3 (Markov Inequality)**.**

$$P(|X| \geq a) \leq \frac{E|X|}{a}$$

for any $a > 0$.

Proof. Let $I_{|X|\geq a}$ be the indicator random variable for the event $|X| \geq a$. It is always true that

$$aI_{|X|\geq a} \leq |X|$$

because if $I_{|X|\geq a} = 1$, then $|X| \geq a$ and then inequality holds, and if $I_{|X|\geq a} = 0$, the inequality is trivial since $|X| \geq 0$. Then, taking expected values, we have

$$aEI_{|X|\geq a} \leq E|X|$$

as desired. ■

Example. Suppose we have 100 people. It is easily possible that at least 95% of the people are younger than average in the group. However, it is *not* possible that at least 50% are older than twice the average age.

Theorem 29.4 (Chebyshev Inequality)**.**

$$P(|X - \mu| \geq a) \leq \frac{\text{Var}(X)}{a^2}$$

for $\mu = EX$ and $a > 0$. Alternatively, we can write

$$P(|X - \mu| \geq c\,\text{SD}(X)) \leq \frac{1}{c^2}$$

for $c > 0$.

Proof.

$$\begin{aligned} P(|X - \mu| \leq a) &= P((X - \mu)^2 \leq a^2) \\ \text{by Markov} \quad &\leq \frac{E((X - \mu)^2)}{a^2} \\ &= \frac{\text{Var}(X)}{a^2} \end{aligned}$$

■

Lecture 30 — 11/16/11

Definition 30.1. Let $X_1, X_2, \ldots$ be i.i.d. random variables with mean μ and variance σ^2. The sample mean of the first n random variables is

$$\bar{X}_n = \frac{1}{n}\sum_{j=1}^{n} X_j$$

We want to answer the question: What happens to the sample mean when n gets large?

Theorem 30.2 (Law of Large Numbers)**.** *With probability 1, as $n \to \infty$,*

$$\bar{X}_n \to \mu$$

pointwise. That is, the sample mean of a collection of i.i.d. random variables converges to the true mean.

Example. Suppose that $X_j \sim \text{Bern}(p)$. The Law of Large Numbers says that $\frac{1}{n}(X_1 + \cdots + X_n) \to p$.

Note that the Law of Large Numbers says nothing about the value of any individual X_j. For instance, in the above example with simple success and failures (which we may model as a series of coin flips), flipping heads many times does not mean that a tails is on its way. Rather, it means that the large but finite string of heads is "swamped" by the infinite flips yet to come.

Theorem 30.3 (Weak Law of Large Numbers)**.** *For any $c > 0$, as $n \to 0$,*

$$P(|\bar{X}_n - \mu| > c) \to 0$$

Proof. (of Weak LoLN) By Chebyshev's inequality,

$$\begin{aligned} P(|\bar{X}_n - \mu| > c) &\leq \frac{\text{Var}(\bar{X}_n)}{c^2} \\ &= \frac{\frac{1}{n^2}n\sigma^2}{c^2} \\ &= \frac{\sigma^2}{nc^2} \\ &\to 0 \end{aligned}$$

■

Note that the Law of Large Numbers does not tell us anything about the distribution of $\bar{X}_n$. To study this distribution, and in particular the rate at which $\bar{X}_n \to 0$, we might consider

$$n^i(\bar{X}_n - \mu)$$

for various values of i.

Theorem 30.4 (Central Limit Theorem)**.** *As $n \to \infty$,*

$$n^{1/2}\frac{(\bar{X}_n - \mu)}{\sigma} \to \mathcal{N}(0, 1)$$

in distribution; that is, the CDFs converge. Equivalently,

$$\frac{\sum_{j=1}^{n} X_j - n\mu}{\sqrt{n}\sigma} \to \mathcal{N}(0, 1)$$

Proof. We will prove the CLT assuming that the MGF $M(t)$ of the X_j exists (note that we have been assuming all along that the first two moments exist). We will show that the MGFs converge, which will imply that the CDFs converge (however, we will not show this fact).

Let us assume WLOG that $\mu = 0$ and $\sigma = 1$. Let

$$S_n = \sum_{j=1}^{n} X_j$$

We will show that the MGF of $\frac{S_n}{\sqrt{n}}$ converges to the MGF of $\mathcal{N}(0,1)$. We have

$$\begin{aligned} & E(e^{tS_n/\sqrt{n}}) \\ \text{uncorrelated} \quad & \text{since independent} \\ & = E(e^{tX_1/\sqrt{n}}) \cdots E(e^{tX_n/\sqrt{n}}) \\ & = E(e^{tX_j/\sqrt{n}})^n \\ & = M\left(\frac{t}{\sqrt{n}}\right)^n \end{aligned}$$

Taking the limit results in the indeterminate form 1^∞, which is hard to work with. Instead, we take the log of both sides and then take the limit, to get

$$\begin{aligned} \lim_{n\to\infty} n \ln M\left(\frac{t}{\sqrt{n}}\right) &= \lim_{n\to\infty} \frac{\ln M(\frac{t}{\sqrt{n}})}{\frac{1}{n}} \\ \text{substitute } y = \frac{1}{\sqrt{n}} & \\ &= \lim_{y\to 0} \frac{\ln M(ty)}{y^2} \\ \text{L'Hopital's} \quad &= \lim_{y\to 0} \frac{tM'(ty)}{2yM(ty)} \\ [M(0)=1, M'(0)=0] \quad &= \frac{t}{2} \lim_{y\to 0} \frac{M'(ty)}{y} \\ \text{L'Hopital's} \quad &= \frac{t^2}{2} \lim_{y\to 0} \frac{M''(ty)}{1} \\ &= \frac{t^2}{2} \\ &= \ln e^{t^2/2} \end{aligned}$$

and $e^{t^2/2}$ is the $\mathcal{N}(0,1)$ MGF. ■

Corollary 30.5. *Let $X \sim \text{Bin}(n,p)$ with $X = \sum_{j=1}^n X_j$, $X_j \sim \text{Bern}(p)$ i.i.d.*

$$\begin{aligned} P(a \le X \le b) &= P\left(\frac{a-np}{\sqrt{npq}} \le \frac{X-np}{\sqrt{npq}} \le \frac{b-np}{\sqrt{npq}}\right) \\ &\approx \Phi\left(\frac{-1}{b-np}\right) - \Phi\left(\frac{-1}{a-np}\right) \end{aligned}$$

The Poisson approximation works well when n is large, p is small, and $\lambda = np$ is moderate. In contrast, the Normal approximation works well when n is large and p is near $\frac{1}{2}$ (to match the symmetry of the normal).

It seems a little strange that we are approximating a discrete distribution with a continuous distribution. In general, to correct for this, we can write

$$P(X = a) = P(a - \epsilon < X < a + \epsilon)$$

where $(a - \epsilon, a + \epsilon)$ contains only a

Lecture 31 — 11/18/11

Definition 31.1. Let $V = Z_1^2 + \cdots + Z_n^2$ where the $Z_j \sim \mathcal{N}(0,1)$ i.i.d. Then V has the chi-squared distribution with n degrees of freedom, $V \sim \chi_n^2$.

Observation 31.2. It is true, but we will not prove, that

$$\chi_1^2 = \text{Gamma}(\frac{1}{2}, \frac{1}{2})$$

Since $\chi_n^2 = \sum \chi_1^2$, we have

$$\chi_n^2 = \text{Gamma}(\frac{n}{2}, \frac{1}{2})$$

Definition 31.3. Let $Z \sim \mathcal{N}(0,1)$ and $V \sim \chi_n^2$ be independent. Let

$$T = \frac{Z}{\sqrt{V/n}}$$

Then T has the Student-t distribution with n degrees of freedom, $T \sim t_n$.

Observation 31.4. The Student-t is symmetric; that is $-T \sim t_n$. Note that if $n = 1$, then T is the ratio of two i.i.d. standard normals, so T becomes the Cauchy distribution (and hence has no mean).

If $n \ge 2$, then

$$E(T) = E(Z)E(\frac{1}{\sqrt{V/n}}) = 0$$

Note that in general, $T \sim t_n$ will only have moments up to (but not including) the nth.

Observation 31.5. We proved that

$$E(Z^2) = 1, \quad E(Z^4) = 1 \cdot 3, \quad E(Z^6) = 1 \cdot 3 \cdot 5$$

using MGFs. We can also prove this by noting that

$$E(Z^{2n}) = E((Z^2)^n)$$

and that $Z^2 \sim \chi_1^2 = \text{Gamma}(\frac{1}{2}, \frac{1}{2})$. Then we can simply use LOTUS to get our desired mean.

Observation 31.6. The Student-t distribution looks much like the normal distribution but is heavier-tailed, especially if n is small. As $n \to \infty$, we claim that the Student-t converges to the standard normal.

Let

$$T_n = \frac{Z}{\sqrt{V_n/n}}$$

where $Z_1, Z_2, \ldots \sim \mathcal{N}(0,1)$ i.i.d., $V_n = Z_1^2 + \cdots + Z_n^2$, and $Z \sim \mathcal{N}(0,1)$ independent of the Z_j. By the Law of Large numbers, with probability 1,

$$\lim_{n\to\infty} \frac{V_n}{n} = 1$$

So $T_n \to Z$, which is standard normal as desired.

Definition 31.7. Let $X = (X_1, \ldots, X_k)$ be a random vector. We say that X has the multivariate normal distribution (MVN) if every linear combination

$$t_1X_1 + \cdots t_kX_k$$

of the X_j is normal.

Example. Let Z, W be i.i.d. $\mathcal{N}(0,1)$. Then $(Z + 2W, 3Z + 5W)$ is MVN, since

$$s(Z + 2W) + t(3Z + 5W) = (s + 3t)Z + (2s + 5t)W$$

is a sum of independent normals and hence normal.

Example. Let $Z \sim \mathcal{N}(0,1)$. Let S be a random sign (± 1 with equal probabilities) independent of Z. Then Z and SZ are marginally standard normal. However, (Z, SZ) is *not* multivariate normal, since $Z + SZ$ is 0 with probability $\frac{1}{2}$.

Observation 31.8. Recall that the MGF for $X \sim \mathcal{N}(\mu, \sigma^2)$ is given by

$$E(e^{tX}) = e^{t\mu + t^2\sigma^2/2}$$

Suppose that $X = (X_1, \ldots, X_k)$ is MVN. Let $\mu_j = EX_j$. Then the MGF of X is given by

$$E(e^{t_1X_1 + \cdots + t_kX_k})$$
$$= \exp(t_1\mu_1 + \cdots + t_k\mu_k + \frac{1}{2}\operatorname{Var}(t_1X_1 + \cdots + t_kX_k))$$

Theorem 31.9. *Let $X = (X_1, \ldots, X_k)$ be MVN. Then within X, uncorrelated implies independence. For instance, if we write $X = (\mathbf{X_1}, \mathbf{X_2})$, if every component of $\mathbf{X_1}$ is uncorrelated with every component of $\mathbf{X_2}$, then $\mathbf{X_1}$ is independent of $\mathbf{X_2}$.*

Example. Let X, Y be i.i.d. $\mathcal{N}(0,1)$. Then $(X+Y, X-Y)$ is MVN. We also have that

$$\begin{aligned}\operatorname{Cov}(X+Y, X-Y) &= \operatorname{Var}(X) + \operatorname{Cov}(X,Y) - \operatorname{Cov}(X,Y) - \operatorname{Var}(Y) \\ &= 0\end{aligned}$$

So by our above theorem, $X+Y$ and $X-Y$ are independent.

Lecture 32 — 11/21/11

Definition 32.1. Let $X_0, X_1, X_2, \ldots$ be sequence of random variables. We think of X_n as the state of a finite system at a discrete time n (that is, the X_n have discrete indices and each has finite range). The sequence has the Markov property if

$$\begin{aligned}P(X_{n+1} = j \mid X_n = i, X_{n-1} = i_{n-1}, \ldots, X_0 = i_0) \\ = P(X_{n+1} = j \mid X_n = i)\end{aligned}$$

In casual terms, in a system with the Markov property, the past and future are *conditionally independent* given the present. Such a system is called a Markov chain.

If $P(X_{n+1} = j \mid X_n = i)$ does not depend on time n, then we denote

$$q_{ij} := P(X_{n+1} = j \mid X_n = i)$$

called the transition probability, and we call the sequence a homogenous Markov chain.

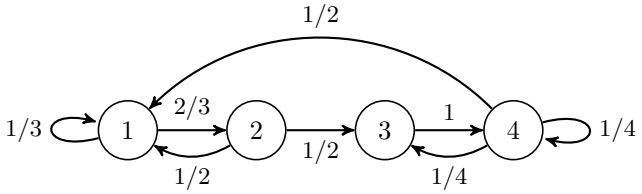
To describe a homogenous Markov chain we simply need to show the states of the process and the transition probabilities. We could, instead, array the q_{ij}'s as a matrix,

$$Q = \left(q_{ij}\right)$$

called the transition matrix.

Note. More generally, we could consider continuous systems (i.e., spaces) at continous times and more broadly study *stochastic processes*. However, in this course, we will restrict our study to homogenous Markov chains.

Example. The following diagram describes a (homogenous) Markov chain:



We could alternatively describe the same Markov chain by specifying its transition matrix

$$Q = \begin{pmatrix} \frac{1}{3} & \frac{2}{3} & 0 & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 \\ 0 & 0 & 0 & 1 \\ \frac{1}{2} & 0 & \frac{1}{4} & \frac{1}{4} \end{pmatrix}$$

Observation 32.2. Suppose that at time n, X_n has distribution s (a row vector in the transition matrix, which represents the PMF). Then

$$\begin{aligned} P(X_{n+1} = j) &= \sum_i P(X_{n+1} = j \mid X_n = i) P(X_n = i) \\ &= \sum_i q_{ij} s_i \\ &= sQ \end{aligned}$$

So sQ is the distribution of X_{n+1}. More generally, we have that sQ^j is the distribution of X_{n+j}.

We can also compute the two-step transition probability:

$$\begin{aligned} &P(X_{n+2} = j \mid X_n = i) \\ &= \sum_k P(X_{n+2} = j \mid X_{n+1} = k, X_n = i) \\ &\qquad P(X_{n+1} = k \mid X_n = i) \\ &= \sum_k P(X_{n+2} = j \mid X_{n+1} = k) P(X_{n+1} = k \mid X_n = i) \\ &= \sum_k q_{kj} q_{ik} \\ &= \sum_k q_{ik} q_{kj} \\ &= (Q^2)_{ij} \end{aligned}$$

More generally, we have

$$P(X_{n+m} = j \mid X_n = i) = (Q^m)_{ij}$$

Definition 32.3. Let s be some probability vector for a Markov chain with transition matrix Q. We say that s is stationary for the chain if

$$sQ = s$$

We also call s a stationary distribution. Note that this is the transpose of an eigenvector equation.

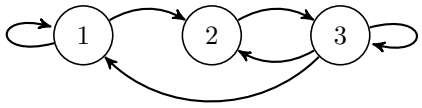
This definition raises the following questions:

1. Does a stationary distribution exist for every Markov chain?
2. Is the stationary distribution unique?
3. Does the chain (in some sense) converge to the stationary distribution?
4. How can we compute it (efficiently)?

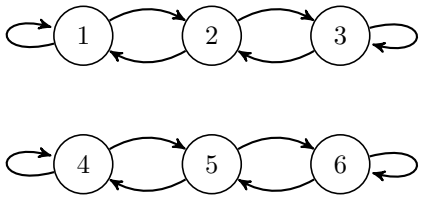
Lecture 33 — 11/28/11

Example. The following are some pathological examples of Markov chains (sans transition probabilities), in state-diagram form:

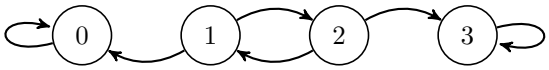
1. Unpathological Markov chain



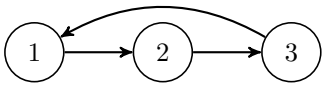
2. Disconnected Markov chain



3. Markov chain with absorbing states



4. Periodic Markov chain



Definition 33.1. A state is recurrent if, starting from that state, there is probability 1 of transitioning back to that state after a finite number of transitions. If a state is not recurrent, it is transient.

Definition 33.2. A Markov chain is irreducible if it is possible (with positive probability) to transition from any state to any other state in a finite number of transitions. Note that in an irreducible chain, all states are recurrent; over an infinite number of transitions, any nonzero probability of returning to a state means that the event of return will occur with probability 1.

Observation 33.3. In our example above, Markov chains 1 and 4 are irreducible; chains 2 and 3 are not. All the states of chain 2 are recurrent; even though the chain itself has two connected components, we will always (i.e., with probability 1), return to the state which we started from.

However, in chain 3, states 1 and 2 are transient. With probability 1, from states 1 and 2, we will at some point transition to state 0 or 3; after that point, we will never return to state 1 or 2. On the other hand, if we start in 0 or 3, we stay there forever; they are clearly recurrent.

Theorem 33.4. *For any irreducible Markov chain,*

1. *A stationary distribution s exists.*

2. *s is unique.*
3. *$s_i = \dfrac{1}{r_i}$, where r_i is the average time to return to state i starting from state i.*
4. *If Q^m is strictly positive for some m, then*

$$\lim_{n\to\infty} P(X_n = i) = s_i$$

Alternatively, if t is any (starting-state) probability vector, then

$$\lim_{n\to\infty} tQ = s$$

Definition 33.5. A Markov chain with transition matrix Q is reversible if there is a probability vector s such that

$$s_i q_{ij} = s_j q_{ji}$$

for all pairs of states i and j.

Reversibility is also known as time-reversibility. Intuitively, the progression of a reversible Markov chain could be played back backwards, and the probabilities would be consistent with the original Markov chain.

Theorem 33.6. *If a Markov chain is reversible with respect to s, then s is stationary.*

Proof. We know that $s_i q_{ij} = s_j q_{ji}$ for some s. Summing over all states,

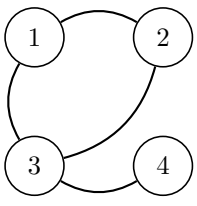
$$\begin{aligned}\sum_i s_i q_{ij} &= \sum_j s_j q_{ji} \\ &= s_j \sum_j q_{ji} \\ &= s_j\end{aligned}$$

But since this is true for every j, this is exactly the statement of

$$sQ = s$$

as desired. ■

Example (Random walk on an undirected network)**.** Consider the following example *undirected* Markov chain



Let d_i be the degree of i (so, $d_1 = 2$, $d_2 = 2$, $d_3 = 3$, $d_4 = 1$). Then we claim that (in general)

$$d_i q_{ij} = d_j q_{ji}$$

for all i, j.

Assume $i \neq j$. Since the Markov chain is undirected, q_{ij} and q_{ji} are either both zero or both nonzero. If (i, j) is an edge, then

$$q_{ij} = \frac{1}{d_i}$$

since our Markov chain represents a random walk. But this suffices to prove our claim.

Let us now normalize d_i to a stationary vector s_i. This is easy; we can simply take

$$s_i = \frac{d_i}{\sum_j d_j}$$

and we have thus found our desired stationary distribution.