1. LECTURE 1: SEPTEMBER 24, 2010

1.1. **Birthday Problem.** If there are 2 people, the chance that they do *not* have the same birthday is

$$\frac{364}{365}.$$

So the chance that they *do* have the same birthday is

$$1 - \frac{364}{365} = \frac{1}{365} \approx 0.28\%.$$

If there are 3 people, you and 2 others, the chance that *neither of the other two* shares *your specific birthday* is

$$\frac{364}{365} \cdot \frac{364}{365},$$

and so the chance that no one else shares *your* birthday is

$$1 - \frac{364}{365} \cdot \frac{364}{365} \approx 0.55\%.$$

However, the other two might have the same birthday, not equal to yours. The chance that *all 3 people have different birthdays* is

$$\frac{365}{365} \cdot \frac{364}{365} \cdot \frac{363}{365};$$

hence, the probability that *not all three birthdays are distinct* (i.e. *at least two share the same birthday*) is

$$1 - \frac{365}{365} \cdot \frac{364}{365} \cdot \frac{363}{365} \approx 0.82\%.$$

Continuing this way, we see that in a group of $n \leq 365$ people, the chance that at least two share the same birthday is

$$1 - \frac{365 \cdot 364 \cdots (365 - n + 1)}{365^n}.$$

For large n, this is very computationally intensive to calculate exactly. For $n = 91$, advanced computational tools (like Maple) can calculate it exactly; to 10 decimal places,

$$1 - \frac{365 \cdot 364 \cdots (275)}{365^{91}} \approx 0.9999953652.$$

But we can make a useful approximation. Calculus shows us that, for $0 \leq p \leq 1$,

$$1 - p < e^{-p}.$$

Thus, if we let P_n be the probability that there are $n \leq 365$ distinct birthdays,

$$P_n = \prod_{k=1}^{n-1} \left(1 - \frac{k}{365}\right) < \prod_{k=1}^{n-1} e^{-k/365} = \exp\left\{-\sum_{k=1}^{n-1} \tfrac{k}{365}\right\} = e^{-\frac{n(n-1)}{2 \cdot 365}}.$$

Plugging in $n = 91$ on a scientific calculate yields the approximation for $1 - P_n$ (the probability that at least two share a birthday)

$$1 - P_n > 1 - e^{-\frac{n(n-1)}{2 \cdot 365}}; 1 - P_{91} > e^{-\frac{91 \cdot 90}{2 \cdot 365}} \approx 0.9999865856.$$

By comparison, consider the *different* question (as above) of the likelihood that someone else has *your* birthday. The probability that no one else has your birthday, in a crowd of size n, is

$$Q_n = \left(\frac{364}{365}\right)^{n-1}.$$

For example, with $n = 91$,

$$1 - Q_{91} \approx 21.8\%.$$

In order for the probability of at least one other person to share your birthday to exceed 50%, we need n large enough that

$$1 - Q_n \geq 0.5; \quad \Longrightarrow \quad n > 253.$$

Many people find this surprising (and would expect the answer to be something like $365/2 \approx 183$). One partial explanation for the counter-intuitive high answer is that, among the others, there are likely to be many pairs that share the same birthday; in 253 people, the number of distinct birthdays represented may be many fewer than 253.

1.2. **Harder Birthday Problems.** It is easy to quickly find very hard problems.

- **Question**: In a crowd of n people, what are the chances that 3 people share the same birthday? 4 people? $m \leq n$ people?
- **Question**: In a crowd of n people, what are the chances that 2 pairs share birthdays? 3 pairs? m pairs? m triples? m sets of size k?

None of these have "nice" answers. (For a set of 3 in a crowd of n, there is an exact formula, but it involves hypergeometric functions.) No formulas are known in greater generality. One can estimate without a formula (this is typically how probability theory is done), but this is quite hard in this setting. In fact, good estimates for the probability that m people share the same birthday were not discovered until 1995.

2. LECTURE 2: SEPTEMBER 27, 2010

2.1. **Experiments and Outcomes.** Probability theory is rooted in experimental science. Say we conduct an experiment, and look for a particular outcome. There are many random / uncontrollable factors that may influence the outcome; if we only conduct the experiment once, we do not get an accurate picture. So we repeat the experiment many times, and average the results.

If the experimental setup is the same each trial (so that the results of previous trials cannot influence future ones), then we expect (with a large number of trials) to see different outcomes occurring with well-defined frequencies.

Example 2.1. Our experiment is flipping a (two-faced) coin. Random effects (the velocity we give it, friction from the air, etc.) determine the outcome, either Heads H or Tails T. In practice, if we perform many trials, we find that

$$\frac{\#\{\text{the outcome } H \text{ occurs}\}}{\text{number of flips}}$$

(i.e. the frequency of Heads) becomes stable as the number of flips grows large. In other words, if $N_n(H)$ is the number of times H occurs in n flips, we expect that

$$\lim_{n\to\infty} \frac{N_n(H)}{n} \quad \text{exists.}$$

This limit is a real number in the interval $[0, 1]$. A coin is called *fair* if this limiting frequency is exactly $\frac{1}{2}$. You might be surprised to learn that a U.S. quarter is *not fair*; experiments have shown that, when tossed in the standard manner, the frequency of heads is a tiny bit higher than tails (probably due to weight distribution).

We abstract this setup and talk about an experiment with a set Ω of possible **outcomes**. This set is called the **sample space**. There may be more than two outcomes. In general, we are interested in the frequencies of occurrence of sets of outcomes, called **events**. (That is, events are subsets of Ω.)

Example 2.2. Consider the experiment of rolling a die. There are 6 outcomes, so the sample space $\Omega = \{1, 2, 3, 4, 5, 6\}$. The set of possible events, the subsets of $\{1, 2, 3, 4, 5, 6\}$, number $2^6 = 64$ in total. For example, one event is $E = \{2, 4, 6\}$ (which we can describe as the event that we roll an *even* number on the die). Let $N_n(E)$ denote the number of times we observe E in n rolls. As usual, we expect that

$$\lim_{n\to\infty} \frac{N_n(E)}{n} \quad \text{exists.}$$

The die is called *fair* if, for $k \in \{1, 2, 3, 4, 5, 6\}$, $\lim_{n\to\infty} N_n(k)/n = \frac{1}{6}$.

Example 2.3. A real number is chosen randomly from the interval $[0, 1]$. In this case, the sample space of outcomes is $[0, 1]$. So events are subsets of $[0, 1]$. This example is trickier. When we talk about choosing a random number, we expect (for example) the first three decimal digits to each be chosen randomly from $\{1, 2, \ldots, 10\}$ with equal frequencies. But that means that, if E_{291} is the event that the number chosen is $0.291\ldots$ (for any choice of continuing digits), we must have

$$\lim_{n\to\infty} \frac{N_n(E_{291})}{n} = \frac{1}{1000}$$

since the 1000 possible 3-digit numbers (ranging from 000 up to 999) are each seen with equal frequencies as the first three digits. Similarly, any particular k-digit string will only show up with frequency 10^{-k}, which is very small for large k. Thus, we have the difficult situation that, for any particular outcome $x \in [0,1]$,

$$\lim_{n\to\infty} \frac{N_n(x)}{n} = 0.$$

In this example, we think in terms of intervals instead: how likely is it the number will be between 0.1 and 0.2, for example. A random number is called **uniform** if, for each interval $[a,b] \subseteq [0,1]$, the frequency it is in the interval $[a,b]$ is

$$\lim_{n\to\infty} \frac{N_n([a,b])}{n} = b - a.$$

Let Ω be a sample space. Here are a few properties that limiting frequencies of events in Ω possess.

- One possible even is Ω itself, the set of all possible outcomes. Since these are the only possible outcomes of the experiment, one of them is bound to happen every time. So the frequency of Ω is 1.
- Another event is $\varnothing$, the event that contains no outcomes. Since *some* outcome occurs each trial, the frequency of $\varnothing$ is 0.
- Suppose that E and F are two events in Ω, and they are **disjoint**: no outcome in E is also in F, and vice versa. Say $E = \{e_1, \ldots, e_k\}$ and $F = \{f_1, \ldots, f_m\}$. Hence, counting up, we find that

$$\begin{aligned} N_n(E \cup F) &= N_n(\{e_1, \ldots, e_k, f_1, \ldots, f_m\}) \\ &= N_n(e_1) + \cdots + N_n(e_k) + N_n(f_1) + \cdots + N_n(f_m) \\ &= N_n(E) + N_n(F). \end{aligned}$$

  Thus, the frequencies also satisfy

$$\lim_{n\to\infty} \frac{N_n(E \cup F)}{n} = \lim_{n\to\infty} \frac{N_n(E)}{n} + \lim_{n\to\infty} \frac{N_n(F)}{n}.$$

Remark 2.4. Given the first and third points above, the second follows. This is because the events Ω and $\varnothing$ are disjoint, and so the frequency of $\Omega \cup \varnothing$ is equal to the sum of the frequencies of Ω and $\varnothing$, or $1 + 0 = 1$. On the other hand, $\Omega \cup \varnothing = \Omega$ which has frequency 1. Hence, the frequency of $\varnothing$ must be $1 - 1 = 0$.

These properties lead us to the abstract **Axioms of Probability Theory**.

2.2. **Axioms of Probability Theory.** Let Ω be a set (the sample space). Let $\mathscr{F}$ be a collection of subsets of Ω (the events). A **probability** or **probability measure** is a function

$$\mathbb{P}\colon \mathscr{F} \to [0,1]$$

which has the following properties.

(1) $\mathbb{P}(\Omega) = 1$.
(2) If $E, F \in \mathscr{F}$ are disjoint (i.e. $E \cap F = \varnothing$) then $\mathbb{P}(E \cup F) = \mathbb{P}(E) + \mathbb{P}(F)$.

(3) If $E_1, E_2, \ldots$ is an infinite sequence of *pairwise disjoint* events (i.e. $E_i \in \mathscr{F}$ for each i and for all $i \neq j$ $E_i \cap E_j = \varnothing$), then

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

You may wonder about property (3). It is important to realize that is *does not follow* from property (2). By induction on property (2), it *does* follow that, for any finite n,

$$\mathbb{P}\left(\bigcup_{i=1}^{n} E_i\right) = \sum_{i=1}^{n} \mathbb{P}(E_i)$$

when $E_1, E_2, \ldots, E_n$ are pairwise disjoint events. If we naively take limits, we conclude that

$$\lim_{n\to\infty} \mathbb{P}\left(\bigcup_{i=1}^{n} E_i\right) = \sum_{i=1}^{\infty} \mathbb{P}(E_i).$$

This is not, a priori, the same as property (3) above; declaring these equal requires a kind of *continuity* for $\mathbb{P}$.

Definition 2.5. *Let $F_1, F_2, \ldots$ be a sequence of subsets of a set Ω. If $F_1 \subseteq F_2 \subseteq \cdots$, and if $F = \bigcup_{n=1}^{\infty} F_n$, say that*

$$F_n \uparrow F \text{ as } n \to \infty.$$

Proposition 2.6. *Let Ω be a set and $\mathscr{F}$ a collection of subsets of Ω. Suppose $\mathbb{Q}\colon \mathscr{F} \to [0,1]$ is a function which satisfies properties (1) and (2) above (with $\mathbb{Q}$ in place of $\mathbb{P}$). Then $\mathbb{Q}$ is a probability if and only if it also satisfies the following condition.*

(3′) *If $F_n \uparrow F$ as $n \to \infty$, then $\lim_{n\to\infty} \mathbb{Q}(F_n) = \mathbb{Q}(F)$.*

Proof. First suppose that (3′) holds. Let $E_1, E_2, \ldots$ be a sequence of disjoint events. Then set

$$F_n = E_1 \cup E_2 \cup \cdots \cup E_n$$

for each n. Then $F_{n+1} = F_n \cup E_{n+1}$, and so $F_n \subseteq F_{n+1}$. Also, setting $F = \bigcup_{i=1}^{\infty} E_i$, we have $F_n \uparrow F$ as $n \to \infty$. Thus, by property (3′), it follows that

$$\mathbb{Q}\left(\bigcup_{i=1}^{\infty} E_i\right) = \mathbb{Q}(F) = \lim_{n\to\infty} \mathbb{Q}(F_n).$$

From property (2),

$$\mathbb{Q}(F_n) = \mathbb{Q}(E_1 \cup \cdots \cup E_n) = \sum_{i=1}^{n} \mathbb{Q}(E_i)$$

because the E_i are disjoint. Thus,

$$\mathbb{Q}\left(\bigcup_{i=1}^{\infty} E_i\right) = \lim_{n\to\infty} \sum_{i=1}^{n} \mathbb{Q}(E_i) = \sum_{i=1}^{\infty} \mathbb{Q}(E_i).$$

So property (3) holds, and $\mathbb{Q}$ is a probability.

Conversely, suppose that $\mathbb{Q}$ is a probability. Let $F_1 \subseteq F_2 \subseteq \cdots$ be any increasing sequence of events, with $F_n \uparrow F$ as $n \to \infty$. Define $E_1 = F_1$, and for $i > 1$ set $E_i = F_i - F_{i-1} = F_i \cap F_{i-1}^c$. Then if $i \neq j$, say $i < j$, we have

$$E_i \cap E_j = F_i \cap F_{i-1}^c \cap F_j \cap F_{j-1}^c \subseteq F_i \cap F_{j-1}^c.$$

But $i < j$ so $i \leq j - 1$. Since $F_i \subseteq F_{i+1} \subseteq \cdots \subseteq F_{j-1}$, it follows that $F_i \cap F_{j-1}^c = \varnothing$. In other words, the events E_i are disjoint. Hence, since $\mathbb{Q}$ is a probability, by property (3) we have

$$\mathbb{Q}\left(\bigcup_{i=1}^{\infty} E_i\right) = \sum_{i=1}^{\infty} \mathbb{Q}(E_i).$$

Note that

$$F_n = (F_n - F_{n-1}) \cup (F_{n-1} - F_{n-2}) \cup \cdots \cup (F_2 - F_1) \cup F_1 = E_n \cup E_{n-1} \cup \cdots \cup E_2 \cup E_1.$$

Since the E_i are pairwise disjoint, property (2) yields $\mathbb{Q}(F_n) = \mathbb{Q}(E_1) + \cdots + \mathbb{Q}(E_n)$. In other words, we have

$$\mathbb{Q}\left(\bigcup_{i=1}^{\infty} E_i\right) = \lim_{n\to\infty} \mathbb{Q}(F_n).$$

Now, what is the set $\bigcup_{i=1}^{\infty} E_i$? To say an outcome x is in this union is to say that it is in at least one E_i. That is, x is in some F_i and not in F_{i-1}. Since it is in some F_i, this means $x \in \bigcup_{i=1}^{\infty} F_i = F$. On the other hand, if y is any element of $F = \bigcup_{i=1}^{\infty} F_i$, then there is some unique i so that $x \in F_i$. If $i > 1$, then $x \notin F_{i-1}$, and hence $x \in E_i$, so $x \in \bigcup E_i$. This shows that $\bigcup_{i=1}^{\infty} E_i = F$, and so we have proven that

$$\mathbb{Q}(F) = \lim_{n\to\infty} \mathbb{Q}(F_n),$$

thus verifying condition (3′).

□

Thus, a probability measure is an abstraction of the "long-time frequency of occurrence" discussion in Section 2.1 above, with one additional condition: *continuity*. There is no very good intuitive way to see why we should require this condition (because it is hard to reason about infinite collections of events). In fact, there are some mathematicians and statisticians who believe condition (3) should not be used. The theory that follows without it, however, is less useful and less powerful, so we will *always* assume property (3) (known as "countable additivity"). Note: in many examples we will consider, the sample space Ω is finite, which means there are only finitely many possible events, and condition (3) is vacuous anyhow.

Remark 2.7. The astute reader may be bothered by the proof of Proposition 2.6 (and also by the statements of the Axioms of Probability) for the following reason. We didn't insist that the collection $\mathscr{F}$ of events consist of *all* subsets of the sample space Ω. (Indeed, there are good technical reasons to allow it to be smaller – see Remark 2.8.) In order for the machinations of the proof to work, then, $\mathscr{F}$ has to allow certain operations. The careful reader should note that if we assume $\mathscr{F}$ is closed under countable unions – $\bigcup_{i=1}^{\infty} E_i$ is in $\mathscr{F}$ whenever $E_i \in \mathscr{F}$ for all i – and complements – $E^c \in \mathscr{F}$ whenever $E \in \mathscr{F}$ – along with the required $\Omega \in \mathscr{F}$ to validate property (1), then the proof of Proposition 2.6 makes good sense. Such a class $\mathscr{F}$ of events is called a σ-**field**. To do probability theory properly, the class of events must form a σ-field. We will gloss over this detail in Math 180A; much of

the time, our sample space Ω will be finite, in which case we always take $\mathscr{F}$ to consist of all possible subsets of Ω (and this *is* a σ-field).

Remark 2.8. Why would we ever want to have the class of events $\mathscr{F}$ not include all subsets of the sample space Ω? One reason might be a (bad) model of the experiment where some of the possible outcomes do not really have probabilities. For example, in the coin-tossing experiment, you might include extra information like "heads, and I'm wearing jeans" versus "heads, and I'm wearing a skirt". These outcomes will not have probabilities: over time, your outfit will change as you continue the experiment, and the inconsistency will mean the limiting frequency does not exist. Of course, in this case, one should make a better model that leaves out data you're not interested in measuring anyhow.

There is a better, technical reason why $\mathscr{F}$ must sometimes exclude subsets of the sample space. Consider, again, Example 2.3. Here the sample space is the interval $[0, 1]$. If one takes $\mathscr{F}$ to equal the (absolutely, unfathomably enormous) collection of all possible subsets of $[0, 1]$, then it is a theorem (of measure theory, beyond the scope of this course) that **there does not exist a probability $\mathbb{P}$ defined on all of $\mathscr{F}$ which satisfies $\mathbb{P}([a, b]) = b - a$ for all intervals $[a, b] \subseteq [0, 1]$.** In other words, if we want to include all possible subsets as events, there is no such thing as a uniform random number. To get around this problem, we restrict the considered events to a subclass called the *Borel sets* in this example; these are the subsets that are generated by intervals (which must, of course, be included) under the operations of countable union and complement.

2.3. **Examples of Probabilities.** Suppose $\mathbb{P}$ is a probability on a finite sample space

$$\Omega = \{\omega_1, \ldots, \omega_n\}.$$

Let $E = \{\omega_{i_1} \ \ldots, \omega_{i_r}\}$ be an event. By Axiom (2), we have

$$\mathbb{P}(E) = \mathbb{P}(\{\omega_{i_1}\}) + \cdots + \mathbb{P}(\{\omega_{i_r}\}).$$

In other words, $\mathbb{P}$ is completely determined by the probabilities it assigns to the singleton events. So, in this finite setting, a probability is just an assignment of numbers $\mathbb{P}(\omega_i) = p_i$ to elements $\omega_i \in \Omega$, such that $p_i \geq 0$ and $p_1 + \cdots + p_n = 1$.

In this case, a common choice is $p_i = 1/n$ for each i; this is the **uniform** probability measure. Uniform measures describe *fair* or *completely random* events.

Example 2.9. Suppose two dice are rolled. The sample space here is the set of all pairs of numbers, each in $\{1, 2, 3, 4, 5, 6\}$; i.e. $\Omega = \{1, 2, 3, 4, 5, 6\}^2$. There are 36 such elements, and so the uniform probability measure assigns $\mathbb{P} = 1/36$ to each such pair. If the dice are fair, this is how we expect them to behave.

For example, consider the event S_7 that the sum of the two dice is 7. This event is

$$S_7 = \{(i, j) \ : \ 1 \leq i, j \leq 6, i + j = 7\}.$$

The event can be enumerated, $S_7 = \{(1, 6), (2, 5), (3, 4), (4, 3), (5, 2), (6, 1)\}$, having 6 elements. Thus, under the uniform probability, $\mathbb{P}(S_7) = 6/36 = 1/6$. Similarly $\mathbb{P}(S_6) = \mathbb{P}(S_8) = 5/36$, and $\mathbb{P}(S_2) = \mathbb{P}(S_{12}) = 1/36$.

3. Lecture 3: September 29, 2010

Recall: when $\Omega = \{\omega_1, \ldots, \omega_n\}$ is finite, a probability is just a choice of numbers $\mathbb{P}(\{\omega_i\}) = p_i$ where $p_i \geq 0$ and $p_1 + \cdots + p_n = 1$. The same goes for an infinite but discrete sample space, like $\Omega = \mathbb{N}$; in this case a probability is just a sequence of non-negative numbers $(p_n)_{n\in\mathbb{N}}$ with $\sum_{n=1}^{\infty} p_n = 1$. For the first 12 lectures in this course, we deal exclusively with this discrete setting. This means we don't need to worry about technical problems with nasty subsets of the sample space, so we **always assume that all subsets of Ω are events**. Also, if any outcome $\omega \in \Omega$ has $\mathbb{P}(\{\omega\}) = 0$, we can simply throw it away in this setting, so we are free to assume that **the only event with probability 0 is $\varnothing$**.

3.1. Elementary Properties of Probability.

The following properties hold for probabilities in general, not just the discrete ones we're considering for the time-being.

Proposition 3.1 (Monotonicity)**.** *Let $\mathbb{P}$ be a probability. If E and F are events and $E \subseteq F$, then $\mathbb{P}(E) \leq \mathbb{P}(F)$.*

Proof. Since $E \cup E^c = \Omega$, we have $F = (F \cap E) \cup (F \cap E^c)$, and these two pieces are disjoint. Now, since $E \subseteq F$, the first term is $F \cap E = E$. Thus $F = E \cup (F \cap E^c)$ where the two are disjoint. Hence

$$\mathbb{P}(F) = \mathbb{P}(E) + \mathbb{P}(F \cap E^c) \geq \mathbb{P}(E) + 0 = \mathbb{P}(E).$$

□

Proposition 3.2. *Let $\mathbb{P}$ be a probability. For any event E,*

$$\mathbb{P}(E^c) = 1 - \mathbb{P}(E).$$

Proof. Since E and E^c are disjoint, condition (2) in the Axioms of Probability says that $\mathbb{P}(E \cup E^c) = \mathbb{P}(E) + \mathbb{P}(E^c)$. But any outcome ω in the sample space Ω is either in E or its complement E^c; hence $E \cup E^c = \Omega$. Condition (1) of the Axioms of Probability says $\mathbb{P}(\Omega) = 1$, so $1 = \mathbb{P}(\Omega) = \mathbb{P}(E \cup E^c) = \mathbb{P}(E) + \mathbb{P}(E^c)$; subtracting now yields the result. □

Proposition 3.3. *Let $\mathbb{P}$ be a probability, and let E, F be two events. Then*

$$\mathbb{P}(E \cup F) = \mathbb{P}(E) + \mathbb{P}(F) - \mathbb{P}(E \cap F).$$

Proof. One can see this by drawing a picture, and noticing that $\mathbb{P}(E) + \mathbb{P}(F)$ "overcounts" the probability of $\mathbb{P}(E \cup F)$ by counting the region $E \cap F$ twice. To be rigorous, note that $E = (E \cap F) \cup (E \cap F^c)$ where $E \cap F$ and $E \cap F^c$ are disjoint; similarly, $F = (E \cap F) \cup (E^c \cap F)$ where the two are disjoint. So

$$\mathbb{P}(E) = \mathbb{P}(E \cap F) + \mathbb{P}(E \cap F^c)$$
$$\mathbb{P}(F) = \mathbb{P}(E \cap F) + \mathbb{P}(E^c \cap F).$$

Adding these two yields

$$\mathbb{P}(E) + \mathbb{P}(F) = 2\mathbb{P}(E \cap F) + \mathbb{P}(E \cap F^c) + \mathbb{P}(E^c \cap F).$$

Subtracting $\mathbb{P}(E \cap F)$ from both sides,

$$\mathbb{P}(E) + \mathbb{P}(F) - \mathbb{P}(E \cap F) = \mathbb{P}(E \cap F) + \mathbb{P}(E \cap F^c) + \mathbb{P}(E^c \cap F).$$

Now, as above, the first two terms on the right sum to $\mathbb{P}(E \cap F) + \mathbb{P}(E \cap F^c) = \mathbb{P}(E)$. Note that E and $E^c \cap F$ are disjoint, and so $\mathbb{P}(E) + \mathbb{P}(E^c \cap F) = \mathbb{P}(E \cup (E^c \cap F))$. The reader can quickly check that $E \cup (E^c \cap F) = E \cup F$, and this concludes the proof. □

Example 3.4. *Out of 100 students surveyed, 90 own laptops and 65 own smart phones. Only 3 students own neither. How many own both a laptop and a smart phone?*

Take the sample space to be the collection of students, and let the probability be uniform: $\mathbb{P}(\omega) = \frac{1}{100}$ for each student ω. So for any event E, the probability of E is just the fraction $\mathbb{P}(E) = \#E/100$. Let L be the event "owns a laptop" and S the event "owns a smart phone". Thus $\mathbb{P}(L) = 0.9$ and $\mathbb{P}(S) = 0.65$. Since 3 students own neither, $\mathbb{P}(L^c \cap S^c) = 0.03$. Then

$$\mathbb{P}(L \cup S) = 1 - \mathbb{P}((L \cup S)^c) = 1 - \mathbb{P}(L^c \cap S^c) = 1 - 0.03 = 0.97.$$

Hence,

$$\mathbb{P}(L \cap S) = \mathbb{P}(L) + \mathbb{P}(S) - \mathbb{P}(L \cup S) = 0.9 + 0.65 - 0.97 = 0.58.$$

In other words, the number of students who own both a laptop and a smart phone is 58.

3.2. **Independence.** "Independence" in science is a metaphysical concept. Two events are independent if they can have no influence on one-another under any circumstances. Formalizing this in abstract probability theory requires the (counterpoint) notion of *conditional probability*.

Example 3.5. Roll two fair dice ("fair" meaning all 36 outcomes are equally likely). Then the probability of the sum being 8 is 5/36, as computed earlier.

But suppose you have some extra information. The dice are rolled one at a time; you don't look the first time, but your friend tells you the first die is not a 5 or a 6. Given this information, how likely is it that the sum is 8?

This is a new experiment, so the old numbers don't apply. Now the set of possible outcomes is

$$(1,1),(1,2),(1,3),(1,4),(1,5),(1,6)$$
$$(2,1),(2,2),(2,3),(2,4),(2,5),(\mathbf{2},\mathbf{6})$$
$$(3,1),(3,2),(3,3),(3,4),(\mathbf{3},\mathbf{5}),(3,6)$$
$$(4,1),(4,2),(4,3),(\mathbf{4},\mathbf{4}),(4,5),(4,6)$$

24 outcomes instead of 36. Of these outcomes, the ones whose sum is 8 are in bold above; there are 3. Now, each of these outcomes was equally likely before, and so (barring those that we know do not occur now) each is equally likely. Thus, the new probability of the sum being 8, **conditioned** on the knowledge that the first die is in $\{1,2,3,4\}$, is $3/24 = 1/8 < 5/36$.

In Example 3.5, we didn't really need to count up; we could have used the following scheme instead. Let N be the event that the first die is not 5 or 6. Let S be the event that the sum is 8. We are interested in the event $N \cap S$; but the **sample space has been reduced** to N, so we need to renormalize. The ratio of all N-allowed outcomes in the universe in which N occurs is

$$\frac{\mathbb{P}(N \cap S)}{\mathbb{P}(N)}.$$

(Note: if N is an event with probability 0, this is problematic; in finite sample spaces this never comes up, but when we get to continuous probability, we'll have to fiddle with this.) This is what we call **conditional probability**.

Definition 3.6. *Let B be an event in a probability space, with $\mathbb{P}(B) > 0$. For any event A, the* **conditional probability** *of A given B is*

$$\mathbb{P}(A|B) \equiv \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}.$$

We will spend a lot of time on conditional probability in Lectures 11 and beyond. For now, we use this notion just to formalize independence.

Definition 3.7. *Two events A, B are called* **independent** *if the probability of A is not changed by conditioning on B (and vice versa). In other words, they are independent if $\mathbb{P}(A|B) = \mathbb{P}(A)$ and $\mathbb{P}(B|A) = \mathbb{P}(B)$. Multiplying out, both of these say the same thing:*

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

Example 3.8. A fair coin is tossed 3 times. The events H_1 that the first is a head and T_3 that the third is a tail are independent. This is because, by definition of fairness, each of the 8 outcomes

$$HHH, HHT, HTH, HTT, THH, THT, TTH, TTT$$

is equally likely. The event H_1 is equal to $\{HHH, HHT, HTH, HTT\}$ with probability $\frac{1}{2}$; the event T_3 is equal to $\{HHT, HTT, THT, TTT\}$ with probability $\frac{1}{2}$; the event $H_1 \cap T_3$ is equal to $\{HHT, HTT\}$ with probability $\frac{1}{4} = \frac{1}{2} \cdot \frac{1}{2}$.

Example 3.9. *Choosing two balls from an urn.* Suppose an urn is filled with 2 red balls and 2 blue balls. You take two out and record their colors. Consider the events E that the first ball is blue, and F that the second ball is blue. Denoting the balls B_1, B_2, R_1, R_2, the set of possible outcomes in the experiment is

$$B_1B_2, B_1R_1, B_1R_2$$
$$B_2B_1, B_2R_1, B_2R_2,$$
$$R_1B_1, R_2B_2, R_1R_2,$$
$$R_2B_1, R_2B_2, R_2R_1.$$

Counting up, we see that E and F each have 6 elements out of the total 12, and so $\mathbb{P}(E) = \mathbb{P}(F) = \frac{1}{2}$. However, we can count in this example that $E \cap F = \{B_1B_2, B_2B_1\}$ so $\mathbb{P}(E \cap F) = 2/12 = 1/6 < \frac{1}{2} \cdot \frac{1}{2}$. So E, F are *not* independent. (Naturally so, since the first ball being blue leaves fewer blue balls, decreasing the odds that the second one is blue.)

We could have reached the same conclusion without enumerating the sample space as follows. Taking two balls out is the same as randomly labeling the balls 1,2,3,4 and noting the colors of balls 1 and 2. Since $\frac{1}{2}$ the balls are blue, for any one of these labels (1, 2, 3, or 4), chances are $\frac{1}{2}$ it will be blue. (Keep in mind, for example, the event that the 2nd labeled ball is blue includes labelings where the 1st labeled is red.) Thus $\mathbb{P}(E) = \mathbb{P}(F) = \frac{1}{2}$. On the other hand, we can calculate $\mathbb{P}(E \cap F)$ from the definition of conditional probability as

$$\mathbb{P}(E \cap F) = \mathbb{P}(E) \cdot \mathbb{P}(F|E).$$

Given that E occurs (so a blue ball has been removed), there are 3 balls remaining in the urn, and only 1 is blue. Hence, $\mathbb{P}(F|E) = \frac{1}{3}$, and so $\mathbb{P}(E \cap F) = \frac{1}{2} \cdot \frac{1}{3} = \frac{1}{6}$.

This way of using conditional probability is the usual way it's used. (I.e. it is often easier to calculate $\mathbb{P}(A|B)$ than $\mathbb{P}(A \cap B)$.)

What about independence of more than two events? Heuristically, to say A, B, C are independent means that there is no interaction between them. There is a tricky point here: it is possible for no one event to influence another, but at the same time *two of them* can influence the third.

Example 3.10. Toss a fair coin twice. Let A be the event that the first toss came up heads, B be the event that the second toss came up heads, and let C be the event that *exactly one* of the two coins came up heads. Fairness means A and B are independent: indeed, $A = \{HH, HT\}$ has $\mathbb{P}(A) = \frac{1}{2}$, $B = \{HH, TH\}$ has $\mathbb{P}(B) = \frac{1}{2}$, and $A \cap B = \{HH\}$ has $\mathbb{P}(A \cap B) = \frac{1}{4}$. Now, $C = \{HT, TH\}$ has $\mathbb{P}(C) = \frac{1}{2}$, and $A \cap C = \{HT\}$ and $B \cap C = \{TH\}$ each has $\mathbb{P}(A \cap C) = \mathbb{P}(B \cap C) = \frac{1}{4}$, so we have

$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B), \quad \mathbb{P}(A \cap C) = \mathbb{P}(A)\mathbb{P}(C), \quad \mathbb{P}(B \cap C) = \mathbb{P}(B)\mathbb{P}(C).$$

That is, each pair of events (A, B), (A, C), and (B, C) is independent. But notice that if A and B both occur, then C does *not* occur: so A and B together influence C!

Example 3.10 highlights the fact that there are multiple ways we might generalize independence to many events.

Definition 3.11. *Let $A_1, \ldots, A_n$ be a collection of events in a probability space. Say they are* **pairwise independent** *if, for all $i \neq j$, $\mathbb{P}(A_i \cap A_j) = \mathbb{P}(A_i)\mathbb{P}(A_j)$. Say they are* **independent** *if, for any choice of indices $1 \leq i_1 < i_2 < \cdots < i_k \leq n$ for any $k = 1, 2, \ldots, n$,*

$$\mathbb{P}(A_{i_1} \cap \cdots \cap A_{i_k}) = \mathbb{P}(A_{i_1}) \cdots \mathbb{P}(A_{i_k}).$$

For example, with three events, A, B, C are independent means *four* conditions:

$$\begin{aligned} &\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B) \\ &\mathbb{P}(A \cap C) = \mathbb{P}(A)\mathbb{P}(C) \qquad \textbf{AND} \qquad \mathbb{P}(A \cap B \cap C) = \mathbb{P}(A)\mathbb{P}(B)\mathbb{P}(C). \\ &\mathbb{P}(B \cap C) = \mathbb{P}(B)\mathbb{P}(C) \end{aligned}$$

This last triple-intersection condition is what we did not verify in Example 3.10. Indeed, it fails in that example: $A \cap B \cap C = \varnothing$ so $\mathbb{P}(A \cap B \cap C) = 0$ in that example, while $\mathbb{P}(A)\mathbb{P}(B)\mathbb{P}(C) = \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{8}$.

Another word of caution. Independence for multiple events can fail in the opposite way from Example 3.10, as the following example shows.

Example 3.12. Let $\Omega = \{1, 2, 3, 4, 5, 6, 7, 8\}$ be a sample space with $\mathbb{P}(k) = \frac{1}{8}$ for each k. Consider the events

$$A = \{1, 2, 3, 4\}, \quad B = \{1, 2, 5, 6\}, \quad C = \{1, 3, 7, 8\}.$$

Then $\mathbb{P}(A) = \mathbb{P}(B) = \mathbb{P}(C) = \frac{1}{2}$. Now, $A \cap B \cap C = \{1\}$ so $\mathbb{P}(A \cap B \cap C) = \frac{1}{8} = \mathbb{P}(A)\mathbb{P}(B)\mathbb{P}(C)$. *But* $B \cap C = \{1\}$ also has probability $\frac{1}{8}$ while $\mathbb{P}(B)\mathbb{P}(C) = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}$. So these events are not pairwise independent, though they are "3-wise independent".

4. Lecture 4: October 1, 2010

4.1. **Independence.** To say n events are independent, we must consider all collections of 2, 3, ..., $n-1$, and n events chosen from the list, and verify that the probabilities of all such intersections are products of the individual probabilities.

In practice, it is usually just the top-level intersection that is (most) important: i.e. the most-used independence property is

$$\mathbb{P}(A_1 \cap A_2 \cap \cdots \cap A_n) = \mathbb{P}(A_1)\mathbb{P}(A_2)\cdots\mathbb{P}(A_n).$$

Example 4.1. According to the 2009 census, 1 in 6 Americans is Latino/Hispanic. If 6 Americans are chosen randomly *and independently*, what is the probability at least one of them is Latino/Hispanic?

Let L be the event that at least 1 of the 6 is Latino/Hispanic. Then L^c is the event that *none* are. In other words,

$$L^c = N_1 \cap N_2 \cap \cdots \cap N_6,$$

where N_k is the event that the kth person is not Latino/Hispanic. The census data says that each person in America has a $1 - \frac{1}{6} = \frac{5}{6}$ chance of being non-Latino/Hispanic. Thus, by independence,

$$\mathbb{P}(L^c) = \mathbb{P}(N_1 \cap N_2 \cap \cdots \cap N_6) = \mathbb{P}(N_1)\mathbb{P}(N_2)\cdots\mathbb{P}(N_6) = \frac{5}{6}\cdot\frac{5}{6}\cdots\frac{5}{6} = \left(\frac{5}{6}\right)^6 \approx 0.33490.$$

Thus, $\mathbb{P}(L) = 1 - \mathbb{P}(L^c) \approx 0.66510$.

Remark 4.2. Actually, since the number of Latino/Hispanics in America at a given time is fixed and finite, if the first random person is Latino/Hispanic, this slighty reduces the chances the second one will be. In other words, it is impossible for the selected people to be *truly* independent. But since the sample size is over 3×10^8, for calculation purposes it is fine to ignore this and pretend the sample size is infinite, so that independence is possible.

In general, if the fraction of people having a certain characteristic L is p, then the chance that a group of n independently-chosen random people contains no people with characteristic L is

$$(1-p)^n.$$

So we can ask questions like: how many people do we need before the chance of finding a person with characteristic L is $> \frac{1}{2}$? Bigger than 0.9? If we want the probability to be $> q$, this means we must have

$$1 - (1-p)^n > q, \quad \text{or} \quad (1-p)^n < 1 - q.$$

Taking logarithms, and noting that both $1-p$ and $1-q$ are in $(0,1)$ so have *negative logarithms*, this means we need

$$n > \frac{\ln(1-q)}{\ln(1-p)}.$$

For example, when $p = \frac{1}{6}$, to get $q > \frac{1}{2}$ requires $n > \ln(1-\frac{1}{2})/\ln(1-\frac{1}{6}) \approx 3.8$, so 4 people are required to have better than even odds. To get $q > 0.9$ requires $n > \ln(1-0.9)/\ln(1-\frac{1}{6}) \approx 12.6$, so 13 people are required.

Example 4.3. Most major sports franchises have a "best of 7" championship. This means team A and team B face off at least 4 times (and up to 7); the first to win 4 games is the champion. Let's assume the teams are evenly matched, so that each game is won by team A or team B with probability $\frac{1}{2}$. Let's also ignore factors like home-team advantage, and "momentum" (whereby a team that has won a few games is more likely, for psychological reasons, to win future ones). Thus, we model the successive games as *independent*.

There are two ways the series can be settled in only 4 games: $AAAA$ and $BBBB$. The probability of each of these events, by independence, is $\mathbb{P}(AAAA) = \mathbb{P}(BBBB) = (\frac{1}{2})^4 = \frac{1}{16}$; hence, the probability that the series ends in 4 games is $2 \cdot \frac{1}{16} = \frac{1}{8}$.

In 5 games, there are 8 possibilities: four each for A and B as champion. Here are the A-champion possibilities.

$$BAAAA, ABAAA, AABAA, AAABA.$$

Note: $AAAAB$ is not valid since the fifth game would not have happened in this case. By independence, each of these outcomes has probability $\mathbb{P}(BAAAA) = \cdots = \mathbb{P}(AAABA) = (\frac{1}{2})^4 = \frac{1}{32}$. So with 8 possible outcomes, the probability that the series ends in 5 games is $8 \cdot \frac{1}{32} = \frac{1}{4}$.

In 6 games there are 20 possibilities; the 10 where A wins are enumerated as follows:

$$BBAAAA, BABAAA, BAABAA, BAAABA,$$
$$ABBAAA, ABABAA, ABAABA,$$
$$AABBAA, AABAB,$$
$$AAABBA.$$

Thus, the probability of the series ending in 6 games is $20 \cdot (\frac{1}{2})^6 = \frac{20}{64} = \frac{5}{16}$.

It is quite messy to enumerate all possible outcomes of the series that go to 7 games. But we don't need to, because the series *must end* in 4, 5, 6, or 7 games. In other words,

$$\mathbb{P}(\text{ends in 4}) + \mathbb{P}(\text{ends in 5}) + \mathbb{P}(\text{ends in 6}) + \mathbb{P}(\text{ends in 7}) = 1,$$

and so

$$\mathbb{P}(\text{ends in 7}) = 1 - \mathbb{P}(\text{ends in 4}) - \mathbb{P}(\text{ends in 5}) - \mathbb{P}(\text{ends in 6}) = 1 - \frac{1}{8} - \frac{1}{4} - \frac{5}{16} = \frac{5}{16}.$$

By the way: we can use this to count the number of outcomes that go to 7 games. Each such outcome (e.g. $BBBAAAA$) has probability $(\frac{1}{2})^7 = \frac{1}{128}$ by independence. So if n is the number of 7-game championship outcomes, we must have

$$\frac{n}{128} = \frac{5}{16}, \qquad \text{therefore} \qquad n = 40.$$

Sport	**4 game series**	**5 game series**	**6 game series**	**7 game series**
Random (∞)	0.125	0.250	0.313	0.313
Basketball (57)	0.122	0.228	0.386	0.263
Baseball (94)	0.181	0.224	0.224	0.372
Hockey (74)	0.270	0.216	0.230	0.284

The chart above shows the ideal probabilities in this model (rounded to 3 decimal places), compared to actual statistics from Basketball, Baseball, and Hockey over the last half-century or more. Basketball appears to be completely random. (You may as well watch two people tossing a coin 7 times!) The only sport that is statistically significantly different from random is Hockey. (Oh Canada...)

4.2. **Random Variables.** In most of the examples we've considered, the events are determined by measuring certain numerical values. These numerical values are called **random variables**. They are the bread and butter of all of science - they're what we *measure*. In terms of our formulation of probability, let's be precise.

Definition 4.4. *Let Ω be a sample space. A function $X : \Omega \to \mathbb{R}$ is called a* **random variable**. *In probability theory (and science in general), typical events are of the form $\{X = x\}$ or $\{a \leq X \leq b\}$ for some constants $a, b, x \in \mathbb{R}$.*

To be clear,

$$\{a \leq X \leq b\} \equiv \{\omega \in \Omega \,:\, a \leq X(\omega) \leq b\}.$$

We might similarly write $\{X \in [a, b]\}$, or related expressions like $\{X \leq x\}$, $\{X > x\}$, defining events in Ω.

Example 4.5. Consider rolling two dice. The sample space Ω is the set of pairs of integers between 1 and 6, $\Omega = \{(i, j) \,:\, i, j \in \{1, 2, 3, 4, 5, 6\}\}$. a common random variable we've been considering is

$$X(i, j) = i + j,$$

the sum of the dice. This sum can take integer values $x = 2$ through $x = 12$; if the dice are fair, we can easily count that the probabilities of the events $\{X = x\}$ are given in the following chart.

x	2	3	4	5	6	7	8	9	10	11	12
$\mathbb{P}(X = x)$	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

Example 4.6. A very drunk man walks along a straight sidewalk. Each step he takes, he chooses randomly (with probability $\frac{1}{2}$ either way) whether to go forward or backward. This so-called **random walk** is the pre-eminent example of a **Markov chain**, which will be discussed in 180B and 180C. Each possible walk, with n steps, can be represented as a string of n $+$'s and $-$'s, each independent with $\mathbb{P}(+) = \mathbb{P}(-) = \frac{1}{2}$. (In fact, this is the same as keeping track of tosses of a fair coin). Here are two interesting random variables:

- $X_n =$ his position (in steps) relative to his starting position, after n steps.

The sample space on which X_n is defined is the set of all walks with n steps, a finite sample space. We can quickly write down the values of X_3 for each of the 8 walks of length 3.

walk	$+++$	$++-$	$+-+$	$+--$	$-++$	$-+-$	$--+$	$---$
X_3	3	1	1	-1	1	-1	-1	-3

Since the steps are independent, each configuration (e.g. $+-+$) has probability $\mathbb{P}(+-+) = \mathbb{P}(+)\mathbb{P}(-)\mathbb{P}(+) = \left(\frac{1}{2}\right)^3 = \frac{1}{8}$. So, we can just count up that

$$\mathbb{P}(X_3 = 3) = \mathbb{P}(X_3 = -3) = \frac{1}{8}, \qquad \mathbb{P}(X_3 = 1) = \mathbb{P}(X_3 = -1) = \frac{3}{8},$$

while X_3 cannot assume any value other than $\pm 1, \pm 3$.

- $T_k =$ the length of time (i.e. number of steps n he takes) until $X_n = k$.

The sample space on which T_k is defined is the set of walks of *infinite* length (infinite sequences of $+$'s and $-$'s), an infinite sample space. It is much harder to calculate the probabilities of the events $T_k = n$ for different n. While $-n \leq X_n \leq n$ surely, all we can say for sure is that $T_k \geq k$. In fact, T_k can take *any* integer value $\geq k$: the drunkard might keep moving back and forth and take an arbitrarily long time to get more than k steps from where he started. There are many walks ω for which $T_k(\omega) = \infty$ (if the drunkard *never* manages to get k steps forward; for example $++-------\cdots$, $+-+------\cdots$, $+--+---$, etc. all have $T_3 = \infty$). In fact $\mathbb{P}(T_k = \infty) = 0$ (again, wait until 180B to discuss these details).

4.3. **Distributions.** Given a random variable on a probability space, we can calculate the probabilities of different numerical outcomes. For example, the table in Example 4.5 lists all the probabilities of all possible outcomes of measuring the variable X, the sum of two dice. From there we can quickly answer questions like: what is the probability that the sum is more than 8? This is

$$\mathbb{P}(X > 8) = \mathbb{P}(X = 9) + \mathbb{P}(X = 10) + \mathbb{P}(X = 11) + \mathbb{P}(X = 12) = \frac{4}{36} + \frac{3}{36} + \frac{2}{36} + \frac{1}{36} = \frac{5}{18}.$$

If U is any subset of the possible values for X, we can calculate the probability $\mathbb{P}(X \in U)$ accordingly. In fact, $\mathbb{P}(X \in \cdot\,)$ is a **new probability**, defined on the set of values of X.

Definition 4.7. *Let $(\Omega, \mathbb{P})$ be a probability space, and let $X : \Omega \to S$ be a random variable, where $S \subseteq \mathbb{R}$ is the* **state space** *of X. (We might have $S = \mathbb{R}$, or $S = [0, 5]$, or $S = \{2, 3, 5, 8, 13\}$, or $S = \mathbb{N}$, for example.) The function defined on subsets $U \subseteq S$ given by*

$$U \mapsto \mathbb{P}(X \in U)$$

is a probability on S. It is called the **distribution** *or* **law** *of X.*

Remark 4.8. The range S of the random variable X is called the **state space** because the value of X reports on the *state of the system* in our experiment. In physical experiments, we never see the sample space – it is inaccessible to us, abstract. What we measure is the sample space.

Example 4.9 (Geometric Distribution)**.** Suppose a given experiment has probability p of success. We repeat *independent trials* of the experiment until it succeeds. Let N be the random variable $N =$ number of trials required before success.

The sample space for N can be described by sequences of S (success) and F (failure). Since we repeat until we succeed (and then stop), the sample space is $\{S, FS, FFS, FFFS, \ldots\}$ (together with the one infinite string $FFFFF\cdots$). Thus, the event $N = n$ is just one

outcome: $\{N = n\} = \{FFF \cdots FS\}$ where there are $n - 1$ F's. By independence of the trials,

$$\mathbb{P}(N = n) = \mathbb{P}(FFF \cdots FS) = \mathbb{P}(F)\mathbb{P}(F)\mathbb{P}(F) \cdots \mathbb{P}(F)\mathbb{P}(S) = (1-p)^{n-1}p.$$

When $n = 1$, $\mathbb{P}(N = 1) = p$. Thus, the distribution of N is the probability on $\mathbb{N}$ which assigns to the number n the probability $(1-p)^n p$. Note,

$$\sum_{n=1}^{\infty}(1-p)^{n-1}p = p\sum_{n=1}^{\infty}(1-p)^{n-1} = p\sum_{k=0}^{\infty}(1-p)^{k} = p\frac{1}{1-(1-p)} = 1,$$

so indeed this distribution is a probability. It is called the **geometric distribution** on $\mathbb{N}$, with parameter p.

For example, consider the drunkard's random walk again. If our experiment is observing his steps, and "success" is considered "stepping forward", then the length of time it takes before his first step forward has a geometric distribution with parameter $\frac{1}{2}$. For instance: the probability that his first step forward is his 3rd step is $(1-\frac{1}{2})^2\frac{1}{2} = \frac{1}{8}$.

5. LECTURE 5: OCTOBER 4, 2010

Example 5.1. Let's look again at Example 4.1. Suppose we randomly, independently call people up and ask if they are Latino/Hispanic. If N is the number of people we call until we find someone who answers affirmatively, then N has a geometric distribution with parameter $\frac{1}{6}$. Hence, $\mathbb{P}(N = 6) = (1 - \frac{1}{6})^5 \frac{1}{6} \approx 0.067$, less than 7%. But in Example 4.1, we calculated that the probability of finding a Latino/Hispanic in a random group of 6 people is more than 66%. Is this a contradiction?

No. In Example 4.1, we calculated the probability that *at least 1* person in a group of 6 is Latino/Hispanic. The event $N = 6$ is much more restrictive: it means *only the 6th person answered yes*. In fact, the event that at least 1 answers yes (out of 6 randomly chosen) is the same as $\{1 \leq N \leq 6\}$. In general, when N has a geometric distribution with parameter p, we can calculate

$$\mathbb{P}(1 \leq N \leq n) = \sum_{k=1}^{n} \mathbb{P}(N = k) = \sum_{k=1}^{n} (1-p)^{k-1} p = p \sum_{\ell=0}^{n-1} (1-p)^{\ell} = p \frac{1 - (1-p)^n}{1 - (1-p)} = 1 - (1-p)^n.$$

This is just as we calculated in Example 4.1; with $p = \frac{1}{6}$ and $n = 6$, we get $\mathbb{P}(1 \leq N \leq 6) = 1 - (1 - \frac{1}{6})^6 \approx 0.66510$.

5.1. Expected Value.

Suppose we play a gambling game in which there are n outcomes $\omega_1, \ldots, \omega_n$, with probabilities $p_1, \ldots, p_n$. Let X be the random variable representing our winnings in each outcome (so $X(\omega)$ is the number of dollars we win if outcome ω occurs). Note: $X(\omega)$ is negative for some outcomes ω. We want to decide whether we should play the game.

Returning to the frequency interpretation of probability, when we say $\mathbb{P}(\omega_k) = p_k$, we mean that if we conduct many independent trials of the experiment/game, the fraction of the time we get outcome ω_k is p_k. Let's look at an example.

Example 5.2 (Roulette)**.** In a round of Roulette, there is a wheel with 38 numbered pockets around the outside edge; 18 are black, 18 are red, and 2 are green. In a basic wager on red, the wheel is spun, and a ball rotates around eventually falling randomly into one of the pockets. If the pocket is red, you win \$1; otherwise you lose \$1.

Let X be the number of dollars you win in one round of Roulette. Then either $X = 1$ or $X = -1$, so the state space is $\{\pm 1\}$. Assuming all pockets are equally likely outcomes (which is physically realistic), $\mathbb{P}(X = 1) = \frac{18}{38} = \frac{9}{19}$ while $\mathbb{P}(X = -1) = \frac{20}{38} = \frac{10}{19}$; this is the distribution (or law) of winnings in Roulette.

What happens if you play many rounds of Roulette? Say you play N rounds (for N large) Then we expect that in about $\frac{9}{19}N$ rounds, you win \$1, while in $\frac{10}{19}N$ rounds, you lose \$1. So, in total, the net money you expect to have "won" after N rounds is about

$$\tfrac{9}{19}N(\$1) + \tfrac{10}{19}N(-\$1) = -\$\tfrac{1}{19}N.$$

One way to say this is that *over time, you lose* $\$\frac{1}{19} \approx 5.26¢$ *on average per round*.

In Example 5.2, what we calculated (the average winnings per round) is called the **expected value** or **expectation** of the winning random variable.

Definition 5.3. *Let $X\colon \Omega \to S$ be a discrete random variable. The* **expectation** *or* **expected value** *of X (if it exists), is*

$$\mathbb{E}(X) = \sum_{\omega\in\Omega} X(\omega)\mathbb{P}(\omega).$$

That is: the expectation is the weighted-average of the possible values of X, weighted according to their likelihoods.

There is an alternative way to write $\mathbb{E}(X)$. If we group outcomes according to common X-value, we get

$$\mathbb{E}(X) = \sum_{x\in S} \sum_{\substack{\omega\in\Omega \\ X(\omega)=x}} X(\omega)\mathbb{P}(\omega) = \sum_{x\in S} x \sum_{\substack{\omega\in\Omega \\ X(\omega)=x}} \mathbb{P}(\omega).$$

For the inside sum, we are adding up the probabilities of all outcomes in the event $\{\omega \in \Omega : X(\omega) = x\}$. By additivity of probabilities, we can therefore write

$$\mathbb{E}(X) = \sum_{x\in S} x\,\mathbb{P}(X = x). \tag{5.1}$$

Example 5.4. Let X be the sum of two fair dice. What is the expected value of X? Referring to the chart on page 14, we have 11 different possible values for X, and so

$$\begin{aligned}\mathbb{E}(X) = 2\cdot\frac{1}{36} + 3\cdot\frac{2}{36} + 4\cdot\frac{3}{36} + 5\cdot\frac{4}{36} + 6\cdot\frac{5}{36} + 7\cdot\frac{6}{36} \\ + 8\cdot\frac{5}{36} + 9\cdot\frac{4}{36} + 10\cdot\frac{3}{36} + 11\cdot\frac{2}{36} + 12\cdot\frac{1}{36} = 7.\end{aligned}$$

(In fact, we could have guessed this without adding: the probabilities are symmetric about the mid-point 7, and in this case the mid-point will always be the expected value.)

In Example 5.4, we used only the distribution of X to calculate $\mathbb{E}(X)$. Equation 5.1 shows that this is true in general: $\mathbb{E}(X)$ depends only on the distribution of X. In other words, if X and Y have the same distribution (but possibly take different values), they will have the same expectation.

Example 5.5. Let $\Omega = \{1, 2, 3\}$ and $S = \{-1, 1\}$. Let $\mathbb{P}$ be uniform on Ω. Then the two random variables

$$X(1) = X(2) = -1, X(3) = 1; \qquad Y(1) = -1, Y(2) = 1, Y(3) = -1$$

Have the same distribution:

$$\mathbb{P}(X = 1) = \mathbb{P}(Y = 1) = 1 = \tfrac{1}{3}, \mathbb{P}(X = -1) = \mathbb{P}(Y = -1) = \tfrac{2}{3}.$$

Hence, they have the same expectation:

$$\mathbb{E}(X) = \mathbb{E}(Y) = \frac{1}{3}\cdot(1) + \frac{2}{3}\cdot(-1) = -\frac{1}{3}.$$

Example 5.6. It is a fact (first calculated by Euler in the 1700s) that

$$\sum_{n=1}^{\infty} \frac{1}{n^2} = \frac{\pi^2}{6}.$$

Suppose X is an $\mathbb{N}$-valued random variable, with

$$\mathbb{P}(X = n) = \frac{6}{\pi^2} \cdot \frac{1}{n^2}.$$

(This is possible because the sum of all the $\mathbb{P}(X = n)$ is 1.) Then

$$\mathbb{E}(X) = \sum_{n=1}^{\infty} n\,\mathbb{P}(X = n) = \sum_{n=1}^{\infty} n \cdot \frac{6}{\pi^2} \cdot \frac{1}{n^2} = \frac{6}{\pi^2} \sum_{n=1}^{\infty} \frac{1}{n} = \infty.$$

So this random variable X has infinite expected value. Worse yet, if Y is a random variable with $\mathbb{P}(Y = 0) = 0$ and $\mathbb{P}(Y = n) = \mathbb{P}(Y = -n) = \frac{3}{\pi^2} \cdot \frac{1}{n^2}$, then

$$\mathbb{E}(Y) = \frac{3}{\pi^2} \sum_{n=-\infty}^{-1} \frac{1}{n^2} + \frac{3}{\pi^2} \sum_{n=1}^{\infty} \frac{1}{n^2},$$

and this sum doesn't even make sense. So Y *does not have an expectation*.

Example 5.7. Suppose N is a geometric random variable with parameter p. This means its distribution is given by

$$\mathbb{P}(N = n) = p(1-p)^{n-1}, \quad n \in \mathbb{N}.$$

Then the expectation of N is

$$\mathbb{E}(N) = \sum_{n \geq 0} n\mathbb{P}(N = n) = \sum_{n=0}^{\infty} np(1-p)^{n-1}.$$

We use calculus to evaluate this sum. Let $q = 1 - p$. Then we think of $\mathbb{E}(N)$ as a function of q:

$$\mathbb{E}(N) = f(q) = \sum_{n=0}^{\infty} n(1-q)q^{n-1} = (1-q) \sum_{n=0}^{\infty} nq^{n-1}.$$

Noting that $nq^{n-1} = \frac{d}{dq}(q^n)$, we can write this as

$$f(q) = (1-q) \sum_{n=0}^{\infty} \frac{d}{dq}(q^n) = (1-q) \frac{d}{dq} \sum_{n=0}^{\infty} q^n.$$

The sum is the geometric series, which sums to $\frac{1}{1-q}$. Its derivative is $\frac{1}{(1-q)^2}$. Hence

$$\mathbb{E}(N) = f(q) = (1-q) \frac{d}{dq} \frac{1}{1-q} = (1-q) \frac{1}{(1-q)^2} = \frac{1}{1-q} = \frac{1}{p}.$$

In other words: if we perform an experiment repeatedly, where each trial has probability p of success, the average time we can expect to wait for the first success is $1/p$ trials.

Here is a very important property that $\mathbb{E}$ has.

Proposition 5.8. *If $X, Y \colon \Omega \to S$ are random variables and $a \in \mathbb{R}$, then $\mathbb{E}(X + Y) = \mathbb{E}(X) + \mathbb{E}(Y)$ and $\mathbb{E}(aX) = a\mathbb{E}(X)$.*

Proof. From the definition of expectation, this is a simple calculation:

$$\begin{aligned}\mathbb{E}(X+Y) = \sum_{\omega\in\Omega}(X+Y)(\omega)\mathbb{P}(\omega) &= \sum_{\omega\in\Omega}(X(\omega)+Y(\omega))\mathbb{P}(\omega)\\ &= \sum_{\omega\in\Omega}X(\omega)\mathbb{P}(\omega) + \sum_{\omega\in\Omega}Y(\omega)\mathbb{P}(\omega) = \mathbb{E}(X)+\mathbb{E}(Y).\end{aligned}$$

Similarly,

$$\mathbb{E}(aX) = \sum_{\omega\in\Omega}(aX)(\omega)\mathbb{P}(\omega) = a\sum_{\omega\in\Omega}X(\omega)\mathbb{P}(\omega) = a\mathbb{E}(X).$$

□

Remark 5.9. Many sources, like Durrett's book, take Equation 5.1 as the *definition* of $\mathbb{E}$. If you do this, then Proposition 5.8 is not at all obvious (and fairly tricky to prove). Our definition is more natural, and is easily seen to be equivalent to Equation 5.1 as we showed above.

Example 5.10 (China's one-child policy)**.** One plan China thought to enact to counter their overpopulation problem was a "one-son" policy (instead of the "one-child" policy they did adopt). Under this plan, a family could have as many female children as they like, but only one son. (I.e. once they have a son, they must stop having children.) This plan was criticized by those who felt it would create a huge imbalance in the genders.

Assume that the probability of either sex is equal, $\frac{1}{2}$, and the genders of siblings are independent. Then if N is the number of children any family has, N has a geometric distribution with parameter $\frac{1}{2}$. From Example 5.7, it follows that $\mathbb{E}(N) = 2$. Now, let N_m and N_f be the number of male vs. female children each family has. Then $N = N_m + N_f$. By Proposition 5.8, this means $\mathbb{E}(N_m) + \mathbb{E}(N_f) = \mathbb{E}(N) = 2$, and since (by the stopping condition of the experiment) $N_m = 1$, we have $\mathbb{E}(N_f) = 2 - \mathbb{E}(N_m) = 2 - 1 = 1$. Hence, under the one-son policy, the expected numbers of male vs. female children will remain equal.

One may object that, in reality, a family may stop trying for a boy at a certain point, and so there *will* be more girls. So let's modify the setup. Suppose that each family has children until a boy arrives, or until they have 4 children. The possible outcomes are

$$M, FM, FFM, FFFM, FFFF.$$

By independence, these outcomes have probabilities

$$\mathbb{P}(M) = \tfrac{1}{2}, \quad \mathbb{P}(FM) = \tfrac{1}{4}, \quad \mathbb{P}(FFM) = \tfrac{1}{8}, \quad \mathbb{P}(FFFM) = \tfrac{1}{16}, \quad \mathbb{P}(FFFF) = \tfrac{1}{16}.$$

Now, the random variables N_m and N_f can be quickly computed:

$$\begin{aligned}N_m(M) = N_m(FM) = N_m(FFM) = N_m(FFFM) = 1, N_m(FFFF) = 0;\\ N_f(M) = 0, N_f(FM) = 1, N_f(FFM) = 2, N_f(FFFM) = 3, N_f(FFFF) = 4.\end{aligned}$$

Thus, we calculate

$$\begin{aligned}\mathbb{E}(N_m) &= \tfrac{1}{2}\cdot 1 + \tfrac{1}{4}\cdot 1 + \tfrac{1}{8}\cdot 1 + \tfrac{1}{16}\cdot 1 + \tfrac{1}{16}\cdot 0 = \tfrac{15}{16}\\ \mathbb{E}(N_f) &= \tfrac{1}{2}\cdot 0 + \tfrac{1}{4}\cdot 1 + \tfrac{1}{8}\cdot 2 + \tfrac{1}{16}\cdot 3 + \tfrac{1}{16}\cdot 4 = \tfrac{15}{16}.\end{aligned}$$

This is not a miracle: no matter what stopping condition you impose, the expectations of N_m and N_f will always be the same!

6. LECTURE 6: OCTOBER 6, 2010

6.1. **Functions of Random Variables.** Let $X\colon \Omega \to S$ be a random variable. Suppose $f\colon S \to \mathbb{R}$ is a function. Then we can compose the two to get a new random variable:

$$f \circ X = f(X).$$

$f(X)$ is defined on the same sample space, but its state space (determined by the range of f) may be a different subset of $\mathbb{R}$ than S.

Example 6.1. Let X be a random variable. Then X^2 and $X - c$ are both random variables, as is $(X-c)^2$, for any constant c.

Example 6.2. Let $X\colon \Omega \to S$ be a random variable. For any number $x \in S$, consider the *indicator function* $\mathbb{1}_x\colon S \to \mathbb{R}$ defined by

$$\mathbb{1}_x(t) = \begin{cases} 1, & t = x \\ 0, & t \neq x \end{cases}$$

Then $\mathbb{1}_x(X)$ is a random variables that takes only two values: 0 and 1. That is, $\mathbb{1}_x(X)(\omega) = 1$ if $X(\omega) = x$ and $\mathbb{1}_x(X)(\omega) = 0$ otherwise. Note, then, that

$$\{\mathbb{1}_x(X) = 1\} = \{X = x\}.$$

This will come in handy in Section 7.3.

There is a nice relationship between $\mathbb{E}(f(X))$ and the distribution of X.

Proposition 6.3. *Let $X\colon \Omega \to S$ be a random variable and $f\colon S \to \mathbb{R}$ a function. Then*

$$\mathbb{E}(f(X)) = \sum_{x \in S} f(x)\, \mathbb{P}(X = x).$$

Proof. The random variable $Y = f(X)$ has sample space $f(S)$. Equation 5.1 therefore asserts

$$\mathbb{E}(Y) = \sum_{y \in f(S)} y\, \mathbb{P}(Y = y).$$

For each $y \in f(S)$, consider the set of all $x \in S$ with $f(x) = y$. We then have

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

Hence, $\mathbb{P}(Y = y) = \sum_{x\,:\, f(x)=y} \mathbb{P}(X = x)$, so

$$\mathbb{E}(Y) = \sum_{y \in f(X)} y \sum_{x\,:\, f(x)=y} \mathbb{P}(X = x) = \sum_{y \in f(X)} \sum_{x\,:\, f(x)=y} f(x)\, \mathbb{P}(X = x).$$

Finally, the double sum is just a way of summing over all $x \in S$, by grouping S into the blocks of x that all have common f-value. In other words,

$$\mathbb{E}(Y) = \sum_{x \in S} f(x)\, \mathbb{P}(X = x).$$

□

Notice: the proposition shows that, for *any* f, the expectation of $f(X)$ is determined by the distribution of X (not any particular values). (Later on, we'll see that the whole distribution of $f(X)$ is determined by only the distribution of X.)

Example 6.4. Let X be the sum of two fair dice. Let's calculate $\mathbb{E}(X^2)$. Proposition 6.3 tells us that

$$\mathbb{E}(X^2) = \sum_{n=2}^{12} n^2 \, \mathbb{P}(X = n).$$

Referring to the chart on page 14, we can calculate this:

$$\mathbb{E}(X^2) = 2^2 \cdot \tfrac{1}{36} + 3^2 \cdot \tfrac{2}{36} + 4^2 \cdot \tfrac{3}{36} + \cdots + 12^2 \cdot \tfrac{1}{36} = 54\tfrac{5}{6} \approx 54.83.$$

Example 6.4 demonstrates that Proposition 6.3 is quite natural. To paraphrase it: if X takes values $x_1, \ldots, x_n$ with probabilities $p_1, \ldots, p_n$ then $f(X)$ takes values $f(x_1), \ldots, f(x_n)$ with probabilities $p_1, \ldots, p_n$. Thus, the average (expected) value of $f(X)$ is the weighted average $f(x_1)p_1 + \cdots + f(x_n)p_n$.

Example 6.5. If $X : \Omega \to S$ is any random variable and $x \in S$, we can use Proposition 6.3 to compute that

$$\mathbb{E}(\mathbb{1}_x(X)) = \sum_{t \in S} \mathbb{1}_x(t) \, \mathbb{P}(X = t) = 1 \cdot \mathbb{P}(X = x) + \sum_{t \neq x} 0 \cdot \mathbb{P}(X = t) = \mathbb{P}(X = x).$$

Since, as we saw, $\{\mathbb{1}_x(X) = 1\} = \{X = x\}$ and $\{1_x(X) = 0\} = \{X \neq x\}$, this makes sense.

6.2. **Variance.** For a given random variable X, the numbers $\mathbb{E}(X^n)$ for $n = 1, 2, 3, \ldots$ are called the **moments** of X. (They may more may not exist for different n.) The first moment is just $\mathbb{E}(X)$, the expectation. The second moment has special meaning as well – but only when we *remove the first moment contribution*.

Definition 6.6. *If X is a random variable, the* **Variance** *of X, denoted $\mathrm{Var}X$, is the number*

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

Note that, by the linearity of $\mathbb{E}$ (5.8) and the fact that $\mathbb{E}(\mathbb{E}(X)) = \mathbb{E}(X)$ since $\mathbb{E}(X)$ is a constant,

$$\mathrm{Var}X = \mathbb{E}[X^2 - 2\mathbb{E}(X)X + \mathbb{E}(X)^2] = \mathbb{E}(X^2) - 2\mathbb{E}(X)\,\mathbb{E}(X) + \mathbb{E}(X)^2 = \mathbb{E}(X^2) - \mathbb{E}(X)^2.$$

Example 6.7. Let X be the sum of two fair dice. In Example 6.4, we calculated that $\mathbb{E}(X^2) = 54f\tfrac{5}{6}$. Note also that

$$\mathbb{E}(X) = 2 \cdot \tfrac{1}{36} + 3 \cdot \tfrac{2}{36} + 4 \cdot \tfrac{3}{36} + \cdots + 12 \cdot \tfrac{1}{36} = 7.$$

Thus,

$$\mathrm{Var}X = \mathbb{E}(X^2) - \mathbb{E}(X^2) = 54\tfrac{5}{6} - 49 = 5\tfrac{5}{6}.$$

Example 6.8. Let N be a geometric random variable with parameter $p \in (0,1)$. Then

$$\mathbb{E}(N^2) = \sum_{n\geq 1} n^2 \mathbb{P}(N=n) = \sum_{n=1}^{\infty} n^2 p(1-p)^{n-1}.$$

To evaluate this sum, we use calculus once again. Setting $q = 1-p$,

$$\mathbb{E}(N^2) = (1-q)\sum_{n=1}^{\infty} n^2 q^{n-1}.$$

We expect derivatives to be involved: this time, two of them. Note that, if $n \geq 2$,

$$\frac{d^2}{dq^2}(q^n) = n(n-1)q^{n-2}.$$

Thus, we should express n^2 in terms of $n^2 - n$.

$$\mathbb{E}(N^2) = (1-q)\sum_{n=1}^{\infty}(n^2-n+n)q^{n-1} = (1-q)\left(\sum_{n=1}^{\infty} nq^{n-1} + \sum_{n=1}^{\infty}(n^2-n)q^{n-1}\right).$$

The first term in this sum, $(1-q)\sum_{n=1}^{\infty} nq^{n-1}$ is equal to $\mathbb{E}(N) = \frac{1}{p}$ as we calculated in Example 5.7. In the second sum, note that the first term is 0, since $1^2 - 1 = 0$, and so the second sum is

$$(1-q)\sum_{n=2}^{\infty} n(n-1)q^{n-1} = (1-q)q\sum_{n=2}^{\infty} n(n-1)q^{n-2} = (1-q)q\sum_{n=2}^{\infty}\frac{d^2}{dq^2}q^n.$$

Now, $\frac{d^2}{dq^2}(1+q) = 0$, and so we can rewrite this as

$$\begin{aligned}\mathbb{E}(N^2) = \frac{1}{p} + (1-q)q\frac{d^2}{dq^2}\sum_{n=0}^{\infty} q^n &= \frac{1}{p} + (1-q)q\frac{d^2}{dq^2}\frac{1}{1-q} \\ &= \frac{1}{p} + (1-q)q\cdot\frac{2}{(1-q)^3} = \frac{1}{p} + \frac{2q}{(1-q)^2}.\end{aligned}$$

Subbing in $q = 1-p$ yields

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

Thus,

$$\mathrm{Var}N = \mathbb{E}(N^2) - \mathbb{E}(N^2) = \frac{2-p}{p^2} - \left(\frac{1}{p}\right)^2 = \frac{1-p}{p^2}.$$

Notice that in Examples 6.7 and 6.8, the variance was positive. Actually, this is always true by definition: since $(X - \mathbb{E}(X))^2 \geq 0$, it follows that $\mathrm{Var}X = \mathbb{E}[(X-\mathbb{E}(X))^2] \geq 0$.

It is possible, however, for $\mathrm{Var}X$ to equal 0.

Example 6.9. Let $X = c$ be a constant random variable. Then $\mathbb{E}(X) = c$, so $X - \mathbb{E}(X) = c - c = 0$. Thus, $(X - \mathbb{E}(X))^2 = 0$, and so $\mathrm{Var}X = 0$.

Actually, this is the *only* way $\mathrm{Var}X$ can possibly equal 0. For let $\mathbb{E}(X) = c$. If X is not constant, then there is some number $x \neq c$ in the sample space such that $\mathbb{P}(X = x) > 0$. But then the random variable $Y = (X - c)^2$ is > 0 on the set $\{X = x\}$, and so

$$\mathrm{Var}X = \mathbb{E}(Y) = \sum_t (t-c)^2 \mathbb{P}(X = t) \geq (x-c)^2 \mathbb{P}(X = x) > 0.$$

Thus: a random variable X is a constant if and only if $\mathrm{Var}X = 0$.

Definition 6.10. *The* **standard deviation** *of a random variable X, denoted $\sigma(X)$, is*

$$\sigma(X) = \sqrt{\mathrm{Var}X}.$$

Standard deviation is supposed to be a measure of how "spread-out" the distribution of X is. Nearer the end of the course, we'll discuss statements like that. For now, **a word of caution**. You may have heard statistics like "68% of the distribution is within one standard deviation of the mean, and 95% is within two standard deviations". These statements are **generally false**. (The sense in which they are, in some cases, approximately true, has to do with the Central Limit Theorem, our final topic in the course.) The statements are trying to quantify the following:

$$\mathbb{P}(|X - \mathbb{E}(X)| \leq z \cdot \sigma(X)),$$

where z is a positive number (i.e. the number of standard deviations). The claims above are that with $z = 1$ this probability is about 0.68, and with $z = 2$ it's about 0.95. But we can see from many of the examples we've done that these statements are quite false.

Example 6.11. Let N be a geometric random variable with parameter p. In Example 5.7, we calculated that $\mathbb{E}(N) = \frac{1}{p}$, and in Example 6.8 we saw that $\mathrm{Var}N = \frac{1-p}{p^2}$, so $\sigma(N) = \frac{\sqrt{1-p}}{p}$. To fix some numbers, let's take $p = \frac{1}{2}$, so $\mathbb{E}(N) = 2$ and $\sigma(N) = \sqrt{2}$. Then the quantities we're interested in are

$$\mathbb{P}(|N - 2| \leq z \cdot \sqrt{2}).$$

For example, with $z = 1$, $|N - 2| \leq \sqrt{2} < 1.5$ means $1 \leq N \leq 3$ since N is positive-integer valued, and this probability is

$$\begin{aligned}\mathbb{P}(|N-2| \leq 1) &= \mathbb{P}(1 \leq N \leq 3) \\ &= \mathbb{P}(N=1) + \mathbb{P}(N=2) + \mathbb{P}(N=3) = \tfrac{1}{2} + \left(\tfrac{1}{2}\right)^2 + \left(\tfrac{1}{2}\right)^3 = \tfrac{7}{8} = 0.875.\end{aligned}$$

Similarly, with $z = 2$, $|N - 2| \leq 2\sqrt{2} < 2.9$ is the same as $1 \leq N \leq 4$ (since $N \geq 1$ always), and this set has probability $\frac{15}{16} = 0.9375$. Neither matches the commonly quoted statistic.

Example 6.12. Consider a random variable $X \colon \Omega \to \{-1, 0, 1\}$ with distribution $\mathbb{P}(X = \pm 1) = 0.405$ and $\mathbb{P}(X = 0) = 0.19$. Then

$$\mathbb{E}(X) = 0.405(-1) + 0.1(0) + 0.405(1) = 0, \quad \mathbb{E}(X^2) = 0.405(-1)^2 + 0.1(0)^2 + 0.405(1)^2 = 0.81.$$

Hence $\mathrm{Var}X = 0.81 - 0$ and so $\sigma(X) = \sqrt{0.81} = 0.9$. Thus, $\mathbb{P}(|X - 0| \leq 0.9) = \mathbb{P}(X = 0) = 0.19$ is quite small, while $\mathbb{P}(|X - 0| \leq 2 \cdot 0.9) = 1$ is certain.

7. LECTURE 7: OCTOBER 8, 2010

7.1. **Independence of Random Variables.** In scientific experiments, the basic assumption is that different trials are independent. In practice, we measure some random variable in each trial, so what is important is that *the random variables are independent*. What does this mean? Remember: typical events are of the form $X = x$ for some value x.

Definition 7.1. *Let $X_1, \dots, X_n$ be random variables defined on the same sample space Ω, with discrete state space S. They are said to be* **independent** *if, for any choice of numbers $x_1, \dots, x_n$ in S, the events $\{X_1 = x_1\}, \dots, \{X_n = x_n\}$ are independent. In other words,*

$$\mathbb{P}(X_1 = x_1, \dots, X_n = x_n) = \mathbb{P}(X_1 = x_1) \cdots \mathbb{P}(X_n = x_n).$$

The notation $\mathbb{P}(X_1 = x_1, \dots, X_n = x_n)$ is shorthand for $\mathbb{P}(\{X_1 = x_1\} \cap \cdots \cap \{X_n = x_n\})$.

Example 7.2. Roll three fair dice. Let X_1, X_2, X_3 be the numbers on each of the three. Then for any x_1, the event $X_1 = x_1$ describes 36 outcomes (all possible rolls for the other two dice), so $\mathbb{P}(X_1 = x_1) = \frac{36}{6^3} = \frac{1}{6}$. The same is true for the events $X_2 = x_2$ and $X_3 = x_3$. Now, the event $\{X_1 = x_1\} \cap \{X_2 = x_2\} \cap \{X_3 = x_3\}$ describes exactly one outcome (the roll $x_1x_2x_3$) out of the possible 6^3. Hence,

$$\mathbb{P}(X_1 = x_1, X_2 = x_2, X_3 = x_3) = \frac{1}{6^3} = \mathbb{P}(X_1 = x_1)\mathbb{P}(X_2 = x_2)\mathbb{P}(X_3 = x_3),$$

and so the random variables X_1, X_2, X_3 are independent.

Similar calculations show that the random variables giving the values of n fair coin tosses are independent. These are the canonical kinds of examples of independent random variables.

Example 7.3. Roll three fair dice. Let S_{ij} be the sum of die i and die j; that is, $S_{ij} = X_i + X_j$ from Example 7.2. Then S_{12}, S_{23}, S_{31} are *not* independent random variables. For example, consider the events $\{S_{12} = 2\}, \{S_{23} = 2\}, \{S_{31} = 2\}$. The event $\{X_1 + X_2 = 2\}$ can only occur if $X_1 = X_2 = 1$, and this event has probability $\frac{1}{6^2}$. Similarly $\mathbb{P}(S_{23} = 2) = \mathbb{P}(S_{31} = 2) = \frac{1}{6^2}$. But the event $\{S_{12} = S_{23} = S_{31} = 2\}$ can only occur if $X_1 = X_2 = X_3 = 1$, and this even has probability $\frac{1}{6^3}$. Thus

$$\mathbb{P}(S_{12} = 2, S_{23} = 2, S_{31} = 2) = \frac{1}{6^3} \neq \left(\frac{1}{6^2}\right)^3 = \mathbb{P}(S_{12} = 2)\mathbb{P}(S_{23} = 2)\mathbb{P}(S_{31} = 2).$$

These events are also not pairwise independent. The event $\{S_{12} = S_{23} = 2\}$ again contains only the single outcome $X_1 = X_2 = X_3 = 1$ and so has probability $\mathbb{P}(S_{12} = 2, S_{23} = 2) = \frac{1}{6^3}$, whereas $\mathbb{P}(S_{12} = 2)\mathbb{P}(S_{23} = 2) = \frac{1}{6^2} \cdot \frac{1}{6^2} = \frac{1}{6^4}$.

Example 7.4. If a fair coin is tossed twice, let Y_1, Y_2 record the values of the tosses (where heads is 1 and tails is 0). Then Y_1 and Y_2 are independent (same type of argument as in Example 7.2). Now, let $Y_3 = Y_1 + Y_2$. Since Y_3 is determined by Y_1 and Y_2, we should expect that Y_1, Y_2, Y_3 are *not* independent. Indeed, the events $A = \{Y_1 = 1\}$, $B = \{Y_2 = 1\}$, and $C = \{Y_3 = 1\}$ are the events described in Example 3.10, which we noted are not independent. On the other hand, as we showed in that example, the events A, B, C are

pairwise-independent. In fact, if we take the sum modulo 2 (that is, $Y_3 = Y_1 + Y_2$ except in the case $Y_1 = Y_2 = 1$ in which case we define $Y_3 = 0$) this is true for any values of the three variables, and so Y_1, Y_2, Y_3 are pairwise independent. This makes sense, too: consider Y_1 and $Y_1 + Y_2$. Since Y_1, Y_2 are independent, knowing the value of Y_1 gives us no information about the value of $Y_1 + Y_2 \pmod 2$.

7.2. **Sums of Independent Random Variables.** If X, Y are random variables, and you know their distributions, what can you say about the distribution of $X + Y$?

Example 7.5. Suppose X has sample space $\{\pm 1\}$ with distribution $\mathbb{P}(X = \pm 1) = \frac{1}{2}$. Then $\{-X = \pm 1\} = \{X = \mp 1\}$, so $\mathbb{P}(-X = \pm 1) = \frac{1}{2}$ as well – i.e. X and $-X$ has the same distribution. Let $Y_1 = X$ and $Y_2 = -X$. Let Y_3 be a different random variable with this same distribution $\mathbb{P}(Y_3 = \pm 1) = \frac{1}{2}$, such that X, Y_3 are independent. Then we can calculate:

$$X + Y_1 = X + X = 2X, \qquad \mathbb{P}(X + Y_1 = \pm 2) = \frac{1}{2}$$
$$X + Y_2 = X - X = 0, \qquad \mathbb{P}(X + Y_2 = 0) = 1.$$

So, even though Y_1, Y_2 have the same distribution, $X + Y_1$ and $X + Y_2$ have very different distributions. Things are different again with Y_3. Since $X = \pm 1$ and $Y_3 = \pm 1$, there are four possibilities for the values of the two. Independence yields

$$\mathbb{P}(X = 1, Y_3 = 1) = \mathbb{P}(X = 1)\mathbb{P}(Y_3 = 1) = \tfrac{1}{2} \cdot \tfrac{1}{2} = \tfrac{1}{4}$$
$$\mathbb{P}(X = 1, Y_3 = -1) = \mathbb{P}(X = 1)\mathbb{P}(Y_3 = -1) = \tfrac{1}{2} \cdot \tfrac{1}{2} = \tfrac{1}{4}$$
$$\mathbb{P}(X = -1, Y_3 = 1) = \mathbb{P}(X = -1)\mathbb{P}(Y_3 = 1) = \tfrac{1}{2} \cdot \tfrac{1}{2} = \tfrac{1}{4}$$
$$\mathbb{P}(X = -1, Y_3 = -1) = \mathbb{P}(X = -1)\mathbb{P}(Y_3 = -1) = \tfrac{1}{2} \cdot \tfrac{1}{2} = \tfrac{1}{4}.$$

So the possible values for $X + Y_3$ are $1 + 1 = 2$, $1 + (-1) = (-1) + 1 = 0$, and $(-1) + (-1) = -2$. They have probabilities

$$\mathbb{P}(X + Y_3 = 2) = \mathbb{P}(X = 1, Y_3 = 1) = \tfrac{1}{4}$$
$$\mathbb{P}(X + Y_3 = 0) = \mathbb{P}(X = 1, Y_3 = -1) + \mathbb{P}(X = -1, Y_3 = 1) = \tfrac{1}{2}$$
$$\mathbb{P}(X + Y_3 = -2) = \mathbb{P}(X = -1, Y_3 = -1) = \tfrac{1}{4}.$$

All in all, we have three random variables Y_1, Y_2, Y_3 all with the same distribution, but the distributions of $X + Y_1, X + Y_2, X + Y_3$ are all distinct. They even have different ranges: $\{\pm 2\}$ for the first, $\{0\}$ for the second, and $\{0, \pm 2\}$ for the third.

Example 7.5 demonstrates that adding random variables doesn't just add their distributions. In general the distribution of $X + Y$ depends not only on the distributions of X, Y but also on the values of the variables. We will discuss this at greater length later in the quarter, when we talk about *joint distributions*.

There is one case where the distribution of a sum $X + Y$ is completely determined by the distributions of X and Y; we saw a case of it in Example 7.5 with X, Y_3. This is the case when X, Y are *independent* random variables.

Theorem 7.6. *Let $X, Y : \Omega \to S$ be independent random variables with discrete sample spaces S. Then*

$$\mathbb{P}(X + Y = t) = \sum_{x \in S} \mathbb{P}(X = x)\mathbb{P}(Y = t - x).$$

Proof. Since X, Y take values in S, the event $\{X + Y = t\}$ can be broken up as a union

$$\{X + Y = t\} = \bigcup_{x \in S} \{X = x\} \cap \{Y = t - x\}.$$

This is a disjoint union. Thus

$$\mathbb{P}(X + Y = t) = \sum_{x \in S} \mathbb{P}(X = x, Y = t - x).$$

By the independence of X, Y, $\mathbb{P}(X = x, Y = t - x) = \mathbb{P}(X = x)\mathbb{P}(Y = t - x)$, proving the theorem. □

The importance of Theorem 7.6 is that, in order to calculate the probability that the sum $X + Y = t$ for some t, you only need to know the quantities $\mathbb{P}(X = x)$ and $\mathbb{P}(Y = y)$ for x, y ranging through the sample space; i.e. you only need to know the distributions of X and Y.

Example 7.7. Let's look *again* at the sum X of two fair dice. Let X_1, X_2 be the numbers on the dice; then X_1, X_2 are independent (following Example 7.2). We can use Theorem 7.6 to quickly calculate the distribution of X. For example,

$$\begin{aligned}\mathbb{P}(X = 8) = \mathbb{P}(X_1 + X_2 = 0) &= \sum_{x=1}^{6} \mathbb{P}(X_1 = x)\mathbb{P}(X_2 = 8 - x) \\ &= \mathbb{P}(X_1 = 1)\mathbb{P}(X_2 = 7) + \mathbb{P}(X_1 = 2)\mathbb{P}(X_2 = 6) + \mathbb{P}(X_1 = 3)\mathbb{P}(X_2 = 5) \\ &\quad + \mathbb{P}(X_1 = 4)\mathbb{P}(X_2 = 4) + \mathbb{P}(X_1 = 5)\mathbb{P}(X_2 = 3) + \mathbb{P}(X_1 = 6)\mathbb{P}(X_2 = 2).\end{aligned}$$

The first term $\mathbb{P}(X_1 = 1)\mathbb{P}(X_2 = 7) = 0$ of course, since $X_2 \neq 7$. Each of the other terms is the product $\frac{1}{6} \cdot \frac{1}{6} = \frac{1}{36}$, and there are 5 terms, so $\mathbb{P}(X = 8) = \frac{5}{36}$, as we have calculated before.

7.3. **Independence and Expectation.** If X is a random variable and $x \in \mathbb{R}$, we can write the event $\{X = x\}$ as $\{\mathbb{1}_x(X) = 1\}$. This means, as we saw in Example 6.5, that

$$\mathbb{E}(\mathbb{1}_x(X)) = \mathbb{P}(X = x).$$

Now, let X, Y be random variables, and x, y numbers. Consider the function $\mathbb{1}_x(X)\mathbb{1}_y(Y)$. Since $\mathbb{1}_x(X) = 0$ unless $X = x$, and $\mathbb{1}_y(Y) = 0$ unless $Y = y$, this product $= 0$ unless $X = x$ and $Y = y$, in which case it is 1. Thus,

$$\mathbb{E}(\mathbb{1}_x(X)\mathbb{1}_y(Y)) = 1 \cdot \mathbb{P}(X = x, Y = y) + 0 \cdot \mathbb{P}(\cdots) = \mathbb{P}(X = x, Y = y).$$

Putting these together, we have the following rewriting of the definition of independence of random variables.

Lemma 7.8. *Let $X_1, \ldots, X_n$ be (discrete) random variables. They are independent if and only if, for any numbers $x_1, \ldots, x_n$,*

$$\mathbb{E}[\mathbb{1}_{x_1}(X_1) \cdots \mathbb{1}_{x_n}(X_n)] = \mathbb{E}[\mathbb{1}_{x_1}(X_1)] \cdots \mathbb{E}[\mathbb{1}_{x_n}(X_n)].$$

This is a nice kind of algebraic way of expressing things. But we can do better than this. Lets look at a more complicated function than $\mathbb{1}_x$; consider the function $f = \mathbb{1}_{x_1} - 3\mathbb{1}_{x_2}$. (This is just a funny way to write the function f defined by $f(x_1) = 1$, $f(x_2) = -3$, and $f(x) = 0$ for all $x \neq x_1, x_2$.) If X and Y are independent, then for any y, Proposition 5.8 allows us to expand

$$\mathbb{E}[f(X)\mathbb{1}_y(Y)] = \mathbb{E}[(\mathbb{1}_{x_1}(X) - 3\mathbb{1}_{x_2}(X))\mathbb{1}_y(Y)] = \mathbb{E}[\mathbb{1}_{x_1}(X)\mathbb{1}_y(Y)] - 3\mathbb{E}[\mathbb{1}_{x_2}(X)\mathbb{1}_y(Y)].$$

By independence, these terms factor as

$$\mathbb{E}[\mathbb{1}_{x_1}(X)]\mathbb{E}[\mathbb{1}_y(Y)] - 3\mathbb{E}[\mathbb{1}_{x_2}(X)]\mathbb{E}[\mathbb{1}_y(Y)].$$

Now we can collect terms to get

$$(\mathbb{E}[\mathbb{1}_{x_1}(X)] - 3\mathbb{E}[\mathbb{1}_{x_2}(X)]) \cdot \mathbb{1}_y(Y).$$

Applying Proposition 5.8 again, we have

$$\mathbb{E}[\mathbb{1}_{x_1}(X)] - 3\mathbb{E}[\mathbb{1}_{x_2}(X)] = \mathbb{E}[\mathbb{1}_{x_1}(X) - 3\mathbb{1}_{x_2}(X)] = \mathbb{E}[f(X)].$$

Combining everything, what we have is

$$\mathbb{E}[f(X)\mathbb{1}_y(Y)] = \mathbb{E}[f(X)]\mathbb{E}[\mathbb{1}_y(Y)].$$

So this factorization works for linear combinations of indicator functions too; and it would work just as well in the Y variable.

But this gets us *all* functions (in the discrete case). For the sample space is at most countable, $S = \{x_1, x_2, x_3, \ldots\}$, and so for any function $f\colon S \to \mathbb{R}$, if we let $y_n = f(x_n)$, we can rewrite the formula for f as

$$f = \sum_{n=1}^{\infty} y_n \mathbb{1}_{x_n}.$$

So doing calculations like the ones above, and using Lemma 7.8, gives us the following beautiful restatement of independence of random variables.

Theorem 7.9. *Let $X_1, \ldots, X_n\colon \Omega \to S$ be discrete random variables. They are independent if and only if, for any functions $f_1, \ldots, f_n\colon S \to \mathbb{R}$,*

$$\mathbb{E}[f_1(X_1) \cdots f_n(X_n)] = \mathbb{E}[f_1(X_1)] \cdots \mathbb{E}[f_n(X_n)].$$

8. LECTURE 8: OCTOBER 11, 2010

Recall the wonderful restatement of independence of random variables from last time: random variables $X, Y : \Omega \to S$ are independent if, for all functions $f, g \colon S \to \mathbb{R}$

$$\mathbb{E}[f(X)g(Y)] = \mathbb{E}[f(X)]\mathbb{E}[g(Y)].$$

Example 8.1. Let's take another look at Example 6.4. There we calculated the expectation of X^2 where X is the sum of two fair dice. Another approach is as follows. Let X_1, X_2 be the values of the two dice; then X_1, X_2 are independent (following Example 7.2), and $X = X_1 + X_2$. Then $X^2 = X_1^2 + X_1X_2 + X_2^2$ and so

$$\mathbb{E}(X^2) = \mathbb{E}[X_1^2 + 2X_1X_2 + X_2^2] = \mathbb{E}[X_1^2] + 2\mathbb{E}[X_1][X_2] + \mathbb{E}[X_2^2],$$

where the second equality follows from Proposition 5.8 and Theorem 7.9. Note that X_1 and X_2 each have the uniform distribution on $\{1, 2, 3, 4, 5, 6\}$, so

$$\mathbb{E}(X_1) = \mathbb{E}(X_2) = 1 \cdot \tfrac{1}{6} + 2 \cdot \tfrac{1}{6} + \cdots + 6 \cdot \tfrac{1}{6} = \tfrac{21}{6} = 3\tfrac{1}{2}.$$

Similarly

$$\mathbb{E}(X_1^2) = \mathbb{E}(X_2^2) = 1^2 \cdot \tfrac{1}{6} + 2^2 \cdot \tfrac{1}{6} + \cdots + 6^2 \cdot \tfrac{1}{6} = \tfrac{91}{6} = 15\tfrac{1}{6}.$$

Thus,

$$\mathbb{E}(X^2) = 15\tfrac{1}{6} + 2(3\tfrac{1}{2})(3\tfrac{1}{2}) + 15\tfrac{1}{6} = 54\tfrac{5}{6},$$

confirming our earlier answer.

Theorem 7.9 gives us many very important computational tools; here is one of the most useful.

Proposition 8.2. *Let $X_1, \ldots, X_n$ be independent random variables. Then*

$$\mathrm{Var}(X_1 + \cdots + X_n) = \mathrm{Var}X_1 + \cdots + \mathrm{Var}X_n.$$

Proof. We'll just verify this for $n = 2$; in general, the calculation is the same but more tedious. We have

$$(X_1 + X_2)^2 = X_1^2 + X_2^2 + 2X_1X_2.$$

Hence, by independence,

$$\mathbb{E}[(X_1 + X_2)^2] = \mathbb{E}[X_1^2] + \mathbb{E}[X_2^2] + 2\mathbb{E}[X_1]\mathbb{E}[X_2]. \tag{8.1}$$

On the other hand,

$$(\mathbb{E}[X_1 + X_2])^2 = (\mathbb{E}[X_1] + \mathbb{E}[X_2])^2 = \mathbb{E}[X_1]^2 + 2\mathbb{E}[X_1]\mathbb{E}[X_2] + \mathbb{E}[X_2]^2. \tag{8.2}$$

Subtracting (8.2) from (8.1), we get

$$\begin{aligned}\mathrm{Var}(X_1 + X_2) &= \big(\mathbb{E}[X_1^2] + \mathbb{E}[X_2^2] - 2\mathbb{E}[X_1]\mathbb{E}[X_2]\big) - \big(\mathbb{E}[X_1]^2 + 2\mathbb{E}[X_1]\mathbb{E}[X_2] + \mathbb{E}[X_2]^2\big) \\ &= \big(\mathbb{E}(X_1^2) - \mathbb{E}(X_1)^2\big) + \big(\mathbb{E}(X_2^2) - \mathbb{E}(X_2)^2\big) = \mathrm{Var}X_1 + \mathrm{Var}X_2.\end{aligned}$$

□

Remark 8.3. If we hadn't assumed independence, that calculation shows that

$$\mathrm{Var}(X_1 + X_2) = \mathrm{Var}X_1 + \mathrm{Var}X_2 + \mathbb{E}(X_1X_2) - \mathbb{E}(X_1)\mathbb{E}(X_2).$$

This correction term is called the **covariance** of X_1 and X_2; that is,

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

If X, Y are independent then $\mathrm{Cov}(X, Y) = 0$. But the converse is false. In general, the result of Proposition 8.2 holds under the weaker assumption that the covariances of all pairs of variables $X_1, \ldots, X_n$ are 0. In this case the random variables are said to be **uncorrelated**, which is weaker than independent.

Example 8.4. Suppose that $X_1, X_2, X_2, \ldots, X_n$ are independent, and all have the same Variance σ^2. Then if $S_n = X_1 + \cdots + X_n$, $\mathrm{Var}S_n = \mathrm{Var}X_1 + \cdots + \mathrm{Var}X_n = n\sigma^2$. Thus the standard deviation $\sigma(S_n) = \sqrt{n}\,\sigma$. I.e. **the standard deviation of an independent sum of n variables grows at the rate $\sqrt{n}$.**

Example 8.5. Let $X_1, X_2, \ldots, X_n$ be independent random variables, all with the same distribution. (We call them *i.i.d.*: **independent and identically distributed**.) Let $A_n = \frac{1}{n}(X_1 + \cdots + X_n)$, the (empirical) average. Then we can calculate

$$\mathrm{Var}A_n = \frac{1}{n^2}\mathrm{Var}(X_1 + \cdots + X_n) = \frac{1}{n^2}(\mathrm{Var}X_1 + \cdots + \mathrm{Var}X_n) = \frac{1}{n^2} \cdot n \cdot \mathrm{Var}X_1.$$

So $\sigma(A_n) = \frac{1}{\sqrt{n}}\sigma(X_1)$. Now, we can ask the question: how likely is it that A_n is at most z standard deviations from its expected value?

$$\mathbb{P}[|A_n - \mathbb{E}(A_n)| \leq z \cdot \sigma(A_n)]$$

Since $\mathbb{E}(A_n) = \frac{1}{n}(\mathbb{E}(X_1) + \cdots + \mathbb{E}(X_n))$, we can rewrite this as

$$\mathbb{P}\left(\left|\tfrac{[X_1 - \mathbb{E}(X_1)] + \cdots + [X_n - \mathbb{E}(X_n)]}{n}\right| \leq z \cdot \tfrac{1}{\sqrt{n}}\sigma(X_1)\right).$$

Multiplying through by n gives us

$$\mathbb{P}\left(|[X_1 - \mathbb{E}(X_1)] + \cdots + [X_n - \mathbb{E}(X_n)]| \leq z \cdot \sqrt{n} \cdot \sigma(X_1)\right).$$

Since $X_1 - \mathbb{E}(X_1), \ldots, X_n - \mathbb{E}(X_n)$ are all independent, we can actually use Theorem 7.6 to evaluate this quantity. This will be the end-goal of the course. Remarkably, it turns out that as n grows, this number doesn't depend much on the distribution of X_1, and also doesn't depend much on n. In fact, we will eventually prove the **Central Limit Theorem** which says that this number is approximately equal to

$$\mathbb{P}[|A_n - \mathbb{E}(A_n)| \leq z \cdot \sigma(A_n)] \approx \int_{-z}^{z} \tfrac{1}{\sqrt{2\pi}}\, e^{-x^2/2}\, dx, \quad n \text{ large}.$$

With $z = 1$, this gives about 0.68, and with $z = 2$, this gives about 0.95. *This* is where those statistics come from. They *do not hold* for any given random variable, but they do hold (approximately) when a bunch of *independent copies* of a given random variable are averaged.

8.1. **Combinatorial Tools in Probability.** In finite probability spaces (or at least with random variables with finite state spaces), lots of counting problems arise. There are some basic counting tools we'll use often. *Combinatorics* is a term for a branch of mathematics that has a lot to do with counting the number of elements in sets with some nice structures. Let's proceed with some examples.

Example 8.6. *In the television program* Dancing with the Stars*, 12 celebrities compete over the course of a season; at the end, through various competitions, all 12 are ranked 1 through 12. In how many possible ways can the 12 be ranked?*

The question is the same as asking "how many ways are there to order 12 objects"? We begin by given the 12 objects names $a_1, \ldots, a_{12}$. To order them, we must pick one of them a_i to be first; there are 12 such choices. Now, we must pick one to be second, out of the remaining objects $a_1, \ldots, a_{i-1}, a_{i+1}, \ldots, a_{12}$. Hence, there are 11 choices here. We must pick a third one from the remaining 10. Continuing this way, we see that the number of such orderings is

$$12 \cdot 11 \cdot 10 \cdots 3 \cdot 2 \cdot 1 = 479,001,600.$$

In general, the numbers of orders of n objects is denoted $n!$, n **factorial**.

$$n! = n \cdot (n-1) \cdot (n-2) \cdots 3 \cdot 2 \cdot 1.$$

Thus number grows very quickly with n; a famous approximation is *Stirling's formula*,

$$n! \approx \sqrt{2\pi n} \left(\frac{n}{e}\right)^n,$$

which holds in the sense that the ratio of $n!$ to Stirling's approximation tends to 1 as $n \to \infty$.

Example 8.7. *In the NHL, there are 30 teams: 15 Eastern Conference and 15 Western Conference. Only 8 teams from each conference will make the playoffs; they will be ranked 1 through 8 based on their win-loss-tie record during the 82 regular-season games. This ranking will determine who is matched with whom in the play-offs; it is called the playoff lineup. How many possible playoff lineups are there for the Western Conference?*

The analysis is much the same as in Example 8.7. There are 15 teams, and we must choose 8 of them and order them 1 through 8. So we have 15 choices for #1; once this one is selected, there are 14 choices for #2, and so forth. But we are not choosing all 15 teams; once we have chosen 7 teams to rank #1 through #7, there are 8 teams left, and we choose one of them to be the #8 ranked team, and then we're done. Thus, the number of Western Conference playoff lineups is

$$15 \cdot 14 \cdot 13 \cdot 12 \cdot 11 \cdot 10 \cdot 9 \cdot 8 = 259,459,200.$$

Note: the fact that we went down to 8 is not because there are 8 teams; it is because $15 - 8 = 7$, so we must select all but 7 teams; it just so happens that this means the last factor is 8. If we were selecting 8 teams out of 16, we would have multiplied $16 \cdot 15 \cdots 10 \cdot 9$.

In general, the number of ways of selecting and ordering k objects from a list of $n \geq k$ objects is

$$n \cdot (n-1) \cdot (n-2) \cdots (n-k+2) \cdot (n-k+1).$$

We call this n *permute* k, and denote it ${}_nP_k$ or $P_{n,k}$. Note, we can write this in terms of factorials, since

$$\begin{aligned} n! &= n \cdot (n-1) \cdot (n-2) \cdots (n-k+2) \cdot (n-k+1) \cdot (n-k) \cdots 3 \cdot 2 \cdot 1 \\ &= n \cdot (n-1) \cdot (n-2) \cdots (n-k+2) \cdot (n-k+1) \cdot (n-k)! = {}_nP_k \cdot (n-k)!, \end{aligned}$$

and so

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

Example 8.8. (Refer to Example 8.7) *Anything can happen once the playoffs start. Suppose we forget about the ranking of the* 8 *teams, and only want to know how many different combinations of* 8 *teams are possible.*

To work this out, we combine the last two examples. If we do count different rankings of the 8 teams as different, we get ${}_{15}P_8$ possibilities. Now, if we forget about different orderings, then many of these configurations are really the same as far as we're concerned? How much have we overcounted? Let's select 8 particular teams we care about: the Flames, the Oilers, the Canucks, the Red Wings, the Blackhawks, the Ducks, the Kings, and the Sharks. Among the ${}_{15}P_8$ possible playoff lineups, how many have exactly these teams? The number of ways of ordering them is $8!$, and so all $8!$ orderings appear, and these are the only ways these exact 8 teams can appear.

This is true for any collection of 8 teams. Hence, when counting with order mattering, ${}_{15}P_8$ overcounts the number of unordered lineups by a factor of $8! = 40320$. Thus, the number of possible combinations of teams in the playoffs is

$$\frac{{}_{15}P_8}{8!} = \frac{259,459,200}{40320} = 6435.$$

In general, the number of *combinations* of k objects out of $n \geq k$ (i.e. the number of ways of selecting k objects out of n, not caring about their order) is

$$\frac{{}_nP_k}{k!} = \frac{n!}{k!(n-k)!} = \frac{n \cdot (n-1) \cdots (n-k+1)}{k!}.$$

It is denoted ${}_nC_k$, or $C_{n,k}$, or (most commonly) $\binom{n}{k}$.

Let's revisit an old example.

Example 8.9. In Example 4.3, we counted the number of ways that a team can win a best-of-7 championship tournament (such as in the NHL) in exactly 4 games, or 5 games, or 6 games, or 7 games. There, we did it "by hand" (by writing out all possibilities). But with these new combinatorial tools, it's quite a bit easier.

In order for Team A to win in 7 games, they must win the 7th game, and then 3 other games selected from the first 6. Since the 7th game is fixed as an A-win, this means the number of tournaments in which they win can be counted by the number of ways they can win exactly 3 out of the first 6; i.e.

$$\binom{6}{3} = \frac{6 \cdot 5 \cdot 4}{3 \cdot 2 \cdot 1} = 20.$$

For team A to win in 6 games, on the other hand, means they win the 6th game and 3 of the first 5. Hence, the number of such tournaments is

$$\binom{5}{3} = \frac{5 \cdot 4 \cdot 3}{3 \cdot 2 \cdot 1} = 10.$$

Both of these confirm our earlier results. Moreover, we can count for much larger tournaments now: in a best-of-11, the winner is the first to win 6 games. The number of ways team A wins in 9 games is (since they must win the 9th game and some combination of 5 of the first 8)

$$\binom{8}{5} = \frac{8 \cdot 7 \cdot 6}{5!} = 56.$$

9. LECTURE 9: OCTOBER 13, 2010

9.1. **Poker.** The permutation and combination formulas make it easy to analyze probabilities in well-known card games, like Poker.

Example 9.1. *A standard deck of 52 cards is well-shuffled. Seven cards are dealt to each player (this is called Seven Card Stud). What are the odds that your hand contains a pair? Three-of-a-kind? A full house?*

The number of possible combinations of 7 cards is $\binom{52}{7} = 133,784,560$. Saying that the deck is well-shuffled is exactly to say that each of these possible hands is equally likely to have been dealt to any given player. So we need only count the number of hands that contain a pair, or three of a kind, or a full house.

- **A pair**. This means exactly two of the cards have the same value. There are 13 values for the card (A, 1, 2, ..., J, Q, K). The two will have two different suits chosen from among the 4 suits; this can be done in $\binom{4}{2} = 6$ ways. Since we are looking for a pair and nothing else, the only other constraint is that the 5 remaining cards not have matching values, and none should have the same value as the pair already found. So we choose 5 different values from the 12 remaining ones, in $\binom{12}{5}$ ways, and assign each of them one of the four suits, in 4^5 ways. Thus, the number of hands containing exactly (and only) a pair is

$$13 \cdot \binom{4}{2} \cdot \binom{12}{5} \cdot 4^5 = 63,258,624.$$

  So, the odds of getting a pair (and nothing else) are

$$\frac{63,258,624}{133,784,560} \approx 47.3\%.$$

Three of a kind. The analysis is similar to "a pair". There are 13 values, with 3 suits, giving us $13 \cdot \binom{4}{3}$. The other 4 cards have distinct values, different from the one already chosen, with 4 suits; this gives $\binom{12}{4} \cdot 4^4$. Thus, the odds of 3 of a kind are

$$\frac{13 \cdot \binom{4}{3} \cdot \binom{12}{4} \cdot 4^4}{\binom{52}{7}} \approx 4.9\%.$$

A full house. This means a pair and three of a kind. We choose one value (out of 13) for the pair, and two suits $\binom{4}{2}$; we choose a second value (out of 12) for the three of a kind, and three suits $\binom{4}{3}$; the two remaining cards have distinct values chosen from among the remaining 11, giving $\binom{11}{2}$, with two suits, 4^2. Thus, the odds of a full house are

$$\frac{13 \cdot \binom{4}{2} \cdot 12 \cdot \binom{4}{3} \cdot \binom{11}{2} \cdot 4^2}{\binom{52}{7}} \approx 2.5\%.$$

Example 9.2. *In Texas Hold'em Poker, each player is dealt 2 cards. 3 "community cards" (called "the flop") are dealt face up in the middle, and betting begins. In the course of betting, 2 more cards ("the turn" and "the river") are dealt face up.*

Suppose your two cards are both hearts, and two of the flop cards are also a hearts. What are the odds you will get a flush: five cards out of the 7 (your two plus the community cards) having the same suit? What if only one of the flop cards was a heart?

- **2 hearts in the flop**. The values of 5 cards are known to you – your two, and the flop. There are 13 hears in total, and 4 of them are among the 5 face-up. The remaining 47 cards include $13-4=9$ hearts, and so contain $47-9=38$ non-hearts. There are $\binom{47}{2}=1081$ ways the final two cards can be chosen; of those, $\binom{38}{2}$ have both non-hearts. Hence, the odds that at least one heart will come up as the turn or the river (thus giving you a flush) are
$$1-\frac{\binom{38}{2}}{\binom{47}{2}}\approx 35.0\%.$$
- **1 heart in the flop**. This time the remaining 47 cards include $13-3=10$ harts, and 37 non-hearts. We need both the turn and the river to be hearts; this can only happen in $\binom{10}{2}$ ways. So the odds of a flush of hearts in this scenario are quite bleak:
$$\frac{\binom{10}{2}}{\binom{47}{2}}=\frac{45}{1081}\approx 4.2\%.$$

9.2. **More than two Categories.** We can also use the coefficients $\binom{n}{k}$ to count the number of ways of dividing a group into more than two pieces.

Example 9.3. *In a house with 12 rooms, we want to paint 3 of them red, 4 of them white, and 5 of them blue. How many ways can we do this?*

First pick 3 of the rooms to be red; this can be done in $\binom{12}{3}$ ways. From the remaining $12-3=9$ rooms, select 4 to be painted white; this can be done in $\binom{9}{4}$ ways. The remaining 5 rooms will get painted blue, as required. Thus, the number of painting configurations is
$$\binom{12}{3}\cdot\binom{9}{4}=\frac{12!}{3!9!}\cdot\frac{9!}{4!5!}=\frac{12!}{3!4!5!}=27720.$$
Note, we could, instead, have painted first the white rooms, then the blue rooms, and left the red for last; had we counted this way, we would have gotten
$$\binom{12}{4}\cdot\binom{8}{5}=\frac{12!}{4!8!}\cdot\frac{8!}{5!3!}=\frac{12!}{3!4!5!}.$$
Naturally, we had to get the same answer. This also shows us a more general combination-counting tool.

In general, the number of ways of dividing a set of n objects into $m\le n$ groups of sizes $n_1,n_2,\dots,n_m$ (meaning $n_1+\cdots+n_m=n$) is denoted $\binom{n}{n_1\ n_2\ \cdots\ n_m}$. It is equal to
$$\binom{n}{n_1\ n_2\ \cdots\ n_m}=\frac{n!}{n_1!\cdots n_m!}.$$
It is easy to check that
$$\binom{n}{n_1\ n_2\ \cdots\ n_m}=\binom{n}{n_1}\cdot\binom{n-n_1}{n_2}\cdot\binom{n-n_1-n_2}{n_3}\cdots\binom{n-n_1-n_2-\cdots-n_{m-1}}{n_m},$$

which (following Example 9.3) is how we come up with the general formula.

The coefficients $\binom{n}{k}$ are often called **binomial coefficients**; this is due to their appearance in the **binomial theorem** (Theorem 10.4, next Lecture). The coefficients $\binom{n}{n_1\ n_2\ \cdots\ n_m}$ care called **multinomial coefficients**, as they appear in the **multinomial theorem**, Theorem 10.6 (a generalization of the binomial theorem, also covered next lecture).

9.3. **Binomial and Poisson Distributions.** Recall the geometric distribution: if an experiment with success probability p is repeated, the probability that it takes n trials for a success to occur is $p(1-p)^{n-1}$. Suppose instead we ask a more scientifically relevant question. Suppose we perform n trials (continuing whether or not some succeed or fail). What is the probability that a specific number k of the trials succeed?

The sample space should be modeled as all sequences $SSFSFFSFFFFSFSSSFSFS$ of successes and failures, of length n. For any such outcome, if there are k successes and $n-k$ failures, since the trials are independent, the probability of such an outcome is $p^k(1-p)^{n-k}$. Now, the event "there are exactly k successes" means that we must choose k of the trials to be S, and the remaining $n-k$ to be F. This can be done in precisely $\binom{n}{k}$ ways. Hence: if $N_{n,p}$ counts the number of successful trials among n, then

$$\mathbb{P}(N_{n,p}=k)=\binom{n}{k}p^k(1-p)^{n-k}.$$

This is called the **binomial distribution**, $\mathrm{binomial}(n,p)$. Since there are only n trials, $N_{n,p}$ is a random variable with state space $\{0,1,\ldots,n\}$. We can quickly check from the binomial theorem that

$$\sum_{k=0}^{n}\mathbb{P}(N_{n,p}=k)=\sum_{k=0}^{n}\binom{n}{k}p^k(1-p)^{n-k}=(p+(1-p))^n=1^n=1.$$

Hence, $\mathrm{binomial}(n,p)$ really is a probability distribution. We can compute its expected value, either brute force, or more cleverly as follows. For $i=1,\ldots,n$, let X_i be a random variable, $X_i=1$ if the ith trial is a success, $X_i=0$ if the ith trial is a failure. Then $N_{n,p}=X_1+\cdots+X_n$. Note that $\mathbb{P}(X_i=1)=p$ and $\mathbb{P}(X_i=0)=(1-p)$, so $\mathbb{E}(X_i)=p$. Hence, $\mathbb{E}(N_{p,n})=\mathbb{E}(X_1+\cdots+X_n)=\mathbb{E}(X_1)+\cdots+\mathbb{E}(X_n)=np$.

We can also calculate the variance, by noting that these random variables X_i are independent (since the trials are). Hence

$$\mathrm{Var}(N_{n,p})=\mathrm{Var}(X_1+\cdots+X_n)=\mathrm{Var}(X_n)+\cdots+\mathrm{Var}(X_n).$$

Each of the variables X_i has

$$\mathbb{E}(X_i^2)=p(1)^2+(1-p)(0)^2=p,$$

and so $\mathrm{Var}(X_i)=\mathbb{E}(X_i^2)-\mathbb{E}(X_i)^2=p-p^2$. Thus, $\mathrm{Var}(N_{n,p})=np(1-p)$. Let's record this for posterity:

If $N_{n,p}$ is $\mathrm{binomial}(n,p)$, then $\quad \mathbb{E}(N_{n,p})=np, \quad \mathrm{Var}(N_{n,p})=np(1-p).$

Example 9.4. *A student guesses randomly on a multiple choice test with* 5 *questions (each with* 4 *choices). What are the odds he will pass (i.e. get at least* 3 *correct)?*

The number N of questions he answers correctly is a $\mathrm{binomial}(5, \frac{1}{4})$ random variable; so the probability that he answers exactly k correct is

$$\mathbb{P}(N = k) = \binom{5}{k} \left(\frac{1}{4}\right)^k \left(\frac{3}{4}\right)^{5-k}.$$

We want to calculate $\mathbb{P}(N \geq 3) = \mathbb{P}(N = 3) + \mathbb{P}(N = 4) + \mathbb{P}(N = 5)$. This is

$$\mathbb{P}(N \geq 3) = \binom{5}{3} \left(\frac{1}{4}\right)^3 \left(\frac{3}{4}\right)^2 + \binom{5}{4} \left(\frac{1}{4}\right)^4 \left(\frac{3}{4}\right) + \binom{5}{5} \left(\frac{1}{4}\right)^5 = \frac{53}{512} \approx 10.4\%.$$

Next time, he should probably study!

In example 9.4, we had to do a fair amount of arithmetic. It might have been a lot worse: if the quiz had had 15 questions, calculating the odds he gets at least 8 correct means adding up 8 terms, each of which involves calculating a binomial coefficient $\binom{15}{k}$. This can get hard quickly. To help with such problems, there is a different distribution which is a very good approximation of $\mathrm{binomial}(n, p)$ for large n and small p.

Definition 9.5. *Let* $\lambda > 0$. *The* **Poisson distribution** $\mathrm{Poisson}(\lambda)$, *is a probability distribution on the natural numbers* $\mathbb{N} = \{0, 1, 2, \ldots\}$. *The distribution is as follows: a random variable* X *is* $\mathrm{Poisson}(\lambda)$ *if, for each* $k \in \mathbb{N}$,

$$\mathbb{P}(X = k) = e^{-\lambda} \frac{\lambda^k}{k!}, \quad \textit{for } k = 0, 1, 2, \ldots$$

Note that, if X is $\mathrm{Poisson}(\lambda)$, then

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

So, indeed, $\mathrm{Poisson}(\lambda)$ is a probability distribution. It is also very easy to compute that

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

Theorem 9.6 (Poisson Approximation)**.** *Suppose* N_n *is a* $\mathrm{binomial}(n, p_n)$ *random variable for some* p_n. *If* $n \cdot p_n \to \lambda$ *as* $n \to \infty$, *then for* $k \in \mathbb{N}$,

$$\lim_{n \to \infty} \mathbb{P}(N_n = k) = e^{-\lambda} \frac{\lambda^k}{k!}.$$

In other words: if p is small and n is large, then a $\mathrm{binomial}(n, p)$ random variable is approximately $\mathrm{Poisson}(np)$. Since np is the expected value of $\mathrm{binomial}(n, p)$ and λ is the expected value of $\mathrm{Poisson}(\lambda)$, this is the way it had to be.
Before we prove Theorem 9.6, let's apply it to Example 9.4 to see how good an approximation we get.

Example 9.7. If N_5 is $\mathrm{binomial}(5, \frac{1}{4})$, the Poisson approximation indicates the distribution is close to $\mathrm{Poisson}(\frac{5}{4})$. Hence

$$\mathbb{P}(N_5 = 3) + \mathbb{P}(N_5 = 4) + \mathbb{P}(N_5 = 5) \approx e^{-\frac{5}{4}} \frac{(\frac{5}{4})^3}{3!} + e^{-\frac{5}{4}} \frac{(\frac{5}{4})^4}{4!} + e^{-\frac{5}{4}} \frac{(\frac{5}{4})^5}{5!} \approx 13.0\%.$$

This isn't a bad approximation, considering that 5 is not a very good approximation of ∞, nor is $\frac{1}{4}$ a good approximation of 0. And the amount of arithmetic here is a lot less, especially as n grows. For example, suppose the test had had 15 questions. In this case $n = 15$ so $pn = \frac{15}{4}$. Hence, the Poisson approximation gives

$$\mathbb{P}(8 \leq N_{15} \leq 15) \approx \sum_{k=8}^{15} e^{-\frac{15}{4}} \frac{(\frac{15}{4})^k}{k!} \approx 3.8\%.$$

On the other hand, if we calculate exactly from the distribution $\mathrm{binomial}(15, \frac{1}{4})$, we get

$$\mathbb{P}(8 \leq N_{15} \leq 15) = \sum_{k=8}^{15} \binom{15}{k} \left(\frac{1}{4}\right)^k \left(\frac{3}{4}\right)^{15-k} \approx 1.7\%.$$

Ratio-wise, this is still not a great approximation (although the absolute difference has improved a little). The reason is that p has not shrunk; in fact, shrinking p has a much stronger effect than increasing n.

Remark 9.8. Example 9.7 demonstrates an important feature of the Poisson approximation, which is hard to prove but true: if $n \cdot p_n$ *increases* to λ as $n \to \infty$, then

$$\mathbb{P}(N_n = k) \leq e^{-\lambda} \frac{\lambda^k}{k!}.$$

That is: the Poisson approximation gives a (good) upper-bound. This was first proved in 1968 in the Annals of Statistics. It is not easy to see from the following proof of the Poisson approximation.

10. LECTURE 10: OCTOBER 15, 2010

10.1. **Proof of the Poisson Approximation.** The Poisson approximation says that a $\mathrm{binomial}(n,p)$ random variable can be approximated by a $\mathrm{Poisson}(np)$ random variable, provided n is large and p is small. To make this precise, we restate the theorem here.

Theorem 10.1 (Poisson Approximation)**.** *For each $n \in \mathbb{N}$, let p_n be a number in $(0,1)$, such that $p_n \to 0$ at the rate $1/n$; i.e. suppose there is some constant $\lambda > 0$ such that $n \cdot p_n \to \lambda$. If N_n is a $\mathrm{binomial}(n,p_n)$ random variable, then for $k \in \mathbb{N}$,*

$$\lim_{n\to\infty} \mathbb{P}(N_n = k) = e^{-\lambda}\frac{\lambda^k}{k!}.$$

Proof. Define $\lambda_n = np_n$. Then we can write

$$\begin{aligned}\mathbb{P}(N_n = k) = \binom{n}{k} p_n^k (1-p_n)^{n-k} &= \frac{n(n-1)\cdots(n-k+1)}{k!}\left(\frac{\lambda_n}{n}\right)^k \left(1-\frac{\lambda_n}{n}\right)^{n-k} \\ &= \frac{n(n-1)\cdots(n-k+1)}{n^k}\cdot\frac{\lambda_n^k}{k!}\cdot\left(1-\frac{\lambda_n}{n}\right)^{n-k} \\ &= \frac{n}{n}\frac{n-1}{n}\cdots\frac{n-k+1}{n}\cdot\frac{\lambda_n^k}{k!}\cdot\left(1-\frac{\lambda_n}{n}\right)^{n}\left(1-\frac{\lambda_n}{n}\right)^{-k}.\end{aligned}$$

For fixed k, each of the k factors at the front tends to 1. By assumption, $\frac{\lambda_n^k}{k!} \to \frac{\lambda^k}{k!}$. The last term also tends to 1 because $\lambda_n \to \lambda$ so $\lambda_n/n \to 0$. Hence,

$$\lim_{n\to\infty} \mathbb{P}(N_n = k) = \frac{\lambda^k}{k!}\cdot\lim_{n\to\infty}\left(1-\frac{\lambda_n}{n}\right)^n.$$

We can use Calculus to simplify this limit. For example, taking the logarithm,

$$\ln\left(1-\frac{\lambda_n}{n}\right)^n = n\ln(1-\lambda_n/n) = n(-\lambda_n/n + o(1/n)) = -\lambda_n + o(1),$$

using Taylor's theorem. Hence, the logarithm tends to $-\lim_{n\to\infty}\lambda_n = -\lambda$, and so the final term tends to $e^{-\lambda}$, completing the proof. □

Example 10.2. *In the summer of 2001, there were 6 shark attacks in Florida, while the yearly average is 2. Is this unusual?*

Suppose that, on any given day, there is a probability p of a shark attack. The summer is 100 days long, and so the number N of shark attacks during the summer is $\mathrm{binomial}(100,p)$. Then $\mathbb{E}(N) = 100p$, and since we know (empirically) that $\mathbb{E}(N) = 2$, we get $p = \frac{1}{50}$ is quite small. Hence, we use the Poisson approximation. The odds that $N \le 5$ are approximately

$$\mathbb{P}(N \le 5) = \sum_{k=0}^{5}\mathbb{P}(N=k) \approx \sum_{k=0}^{5} e^{-2}\frac{2^k}{k!} \approx 98.3\%.$$

(We can do the more laborious exact calculation instead; the result is $\approx 98.5\%$.) In other words, the chances that there should be 6 shark attacks, knowing the long-term behaviour of the sharks, is less than 2%. *This is very unusual*, and points to a change in the sharks' behaviour.

In an article in the September 7, 2001, *National Post*, Prof. David Kelton (then at U. Penn., now the Director of the Master of Science in Quantitative Analysis Program at the University of Cincinnati) stated (about the high number of shark attacks) "Just because you see events happening in a rash this does not imply that there is some physical driver causing them to happen. It is characteristic of random processes that they have bursty behaviour." **This is very wrong**. We are not talking about a single random event with low probability, we are talking about an aggregate of events; as the above analysis shows, without a change in the system, this should not happen with more than 98% certainty. This only goes to show that you should not trust everything you read, even if it is quoted from (so-called) experts. Experts make mistakes, and (more likely) are often misquoted and misunderstood by the media.

10.2. **Binomial Coefficients.** $\binom{n}{k}$ counts the number of ways of choosing one group of size k out of a group of size n; we can equivalently think of this as dividing a group of n into two parts: one of size k and the other of size $n-k$. Thinking in those terms, it becomes obvious that

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

which is also easy to check looking at the formula. Another neat property these coefficients have is the following *Pascal relation*.

Proposition 10.3. *If $1 \leq k < n$, then*

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

Proof. This can be checked with some easy calculations from the formulas; but it is much easier, and more informative, to prove this by understanding *what it says*. Take our group of n, and select one of them (Bob). When we choose k people from the group, we have a choice: either we include Bob or we don't. If we decide to include Bob, we must choose an additional $k-1$ people from the remaining $n-1$: $\binom{n-1}{k-1}$. On the other hand, if we exclude Bob (poor Bob), we must choose all k folks out of the other $n-1$ people: $\binom{n-1}{k}$. Since we must either include or exclude Bob, these must add up to all the ways to choose k out of n: $\binom{n}{k}$. □

If we line up the numbers $\binom{n}{0}, \binom{n}{1}, \binom{n}{2}, \ldots, \binom{n}{n}$ in rows and stack these rows in a pyramid as n increases, we have a triangular array of numbers called *Pascal's triangle*. The relation of Proposition 10.3 says that, in the triangle, the value of any number is equal to the sum of the two numbers above it.

A famous theorem, originally proved by Newton, relates the combination numbers $\binom{n}{k}$ with polynomials.

Theorem 10.4 (The Binomial Theorem)**.** *For any real x, y and natural n,*

$$\begin{aligned}(x+y)^n &= \binom{n}{0}x^n + \binom{n}{1}x^{n-1}y + \binom{n}{2}x^{n-2}y^2 + \cdots + \binom{n}{n-1}xy^{n-1} + \binom{n}{n}y^n \\ &= \sum_{k=0}^{n}\binom{n}{k}x^k y^{n-k}.\end{aligned}$$

Proof. It's convenient to rename $x = x_1$ and $y = x_0$. If we expand out

$$(x_0 + x_1)^n = (x_0 + x_1)(x_0 + x_1) \cdots (x_0 + x_1),$$

forgetting for the moment that x_0, x_1 commute, we get 2^n different terms, each of the form $x_{i_1} x_{i_2} \cdots x_{i_n}$ where $i_1, \ldots, i_n$ are in $\{0, 1\}$. For example,

$$\begin{aligned}(x_0 + x_1)(x_0 + x_1)(x_0 + x_1) = {} & x_0x_0x_0 + x_0x_0x_1 + x_0x_1x_0 + x_0x_1x_1 \\ & + x_1x_0x_0 + x_1x_0x_1 + x_1x_1x_0 + x_1x_1x_1.\end{aligned}$$

Of course, x_0, x_1 *do* commute, and two of these terms $x_{i_1} x_{i_2} \cdots x_{i_n}$ and $x_{j_1} x_{j_2} \cdots x_{j_n}$ are equal if and only if there are the same number of 1s and 0s in the i's as in the j's: i.e. if $i_1 + \cdots + i_n = j_1 + \cdots + j_n$. this sum can be anything from 0 to n; if $i_1 + \cdots + i_n = k$, then each such term is equal to $x_1^k x_0^{n-k} = x^k y^{n-k}$. The number of such terms can be counted: we must choose k 1s out of the n slots, so there are $\binom{n}{k}$ such terms. Hence

$$(x + y)^n = \sum_{k=0}^{n} \sum_{i_1 + \cdots + i_n = k} x_{i_1} \cdots x_{i_n} = \sum_{k=0}^{n} x^k y^{n-k} \sum_{i_1 + \cdots + i_n = k} 1 \;=\; \sum_{k=0}^{n} x^k y^{n-k} \binom{n}{k}.$$

□

Each of the terms $x_{i_1} \cdots x_{i_n} = x^k y^{n-k}$ is called a *monomial* (a polynomial that is a product of powers of the variables). A sum of two monomials is called a *binomial*. Hence this is the *binomial theorem*. For this reason, the numbers $\binom{n}{k}$ are usually called the **binomial coefficients**.

An easy consequence of the binomial theorem is that we can add up all the binomial coefficients of a given degree.

Corollary 10.5. $\sum_{k=0}^{n} \binom{n}{k} = 2^n$.

Proof. Just sub in $x = y = 1$ in the binomial theorem:

$$2^n = (1 + 1)^n = \sum_{k=0}^{n} \binom{n}{k} 1^k 1^{n-k} = \sum_{k=0}^{n} \binom{n}{k}.$$

□

Note: $\binom{n}{k}$ counts the number of subsets of size k from a set of size n. Any subset must have size k for some $k \in \{0, 1, 2, \ldots, n\}$, and so the sum of the binomial coefficients *counts the number of subset of a set* of size n. As claimed earlier, this is equal to 2^n.

The more general numbers $\binom{n}{n_1\ n_2 \cdots\ n_m}$ count the number of ways of dividing a group of n objects into m groups of sizes $n_1, n_2, \ldots, n_m$. These numbers are called **multinomial coefficients**. They arise just like the binomial coefficients, when expanding powers of a multinomial (a sum of, in this case, m monomials).

Theorem 10.6 (Multinomial Theorem)**.** *Let $x_1, x_2, \ldots, x_m$ be real numbers, and let n be a positive integer. Then*

$$(x_1 + x_2 + \cdots + x_m)^n = \sum_{n_1 + \cdots + n_m = n} \binom{n}{n_1\ n_2\ \cdots\ n_m} x_1^{n_1} x_2^{n_2} \cdots x_m^{n_m}.$$

The proof of Theorem 10.6 is very similar to the proof of the binomial theorem, just more notation-intensive.

As with the binomial distribution, multinomial coefficients come into play with computing the probabilities of the number of outcomes of a group of (more than two) events with known probabilities in multiple trials.

Example 10.7. Suppose we perform an experiment in which there are m possible outcomes $\omega_1, \ldots, \omega_m$, with probabilities $\mathbb{P}(\omega_i) = p_i$. If we perform n independent trials of the experiment, then the probability that outcome ω_1 occurs exactly n_1 times, ω_2 occurs n_2 times, and so on through ω_m occurring n_m times (so $n = n_1 + \cdots + n_m$) is

$$\binom{n}{n_1\ n_2\ \cdots\ n_m} p_1^{n_1} p_2^{n_2} \cdots p_m^{n_m}.$$

For example: suppose a die has A on 3 faces, B on 2 faces, and C on 1 face. Then in each roll, $\mathbb{P}(A) = \frac{3}{6} = \frac{1}{2}$, $\mathbb{P}(B) = \frac{2}{6} = \frac{1}{3}$, and $\mathbb{P}(C) = \frac{1}{6}$. If we roll the die 10 times, then

$$\mathbb{P}(5\ A, 3\ B, 2\ C) = \binom{10}{5\ 3\ 2} \left(\frac{1}{2}\right)^5 \left(\frac{1}{3}\right)^3 \left(\frac{1}{6}\right)^2 = \frac{10!}{5!3!2!} \cdot \frac{1}{2^5 3^3 6^2} = \frac{35}{432} \approx 8.1\%.$$

10.3. **Ball and Urn Problems.** Many problems in probability have the following flavour.

Example 10.8. *An urn contains 15 white balls and 10 red balls. If 6 are pulled out randomly, what are the odds that 4 are white and 2 are red?*

There are 25 balls in total, and so the number of ways of pulling 6 out is $\binom{25}{6} = 177100$. We assume that the balls are selected independently, and each is chosen with probability $\frac{1}{25}$; the result is all 177100 combinations are equally likely. So, we must count the number of them in which 4 are white and 2 are red. There are 15 red balls, so there are $\binom{15}{4} = 1365$ was of selecting four white balls; there are 10 red balls, so there are $\binom{10}{2} = 45$ ways of selecting 2 red balls. Hence, there are $1365 \cdot 45 = 61425$ such combinations, and the odds are

$$\frac{61425}{177100} \approx 34.7\%.$$

This might seem surprising: out of the 7 possible color configurations, this lone one happens more than $\frac{1}{3}$ of the time. In fact, we can use similar computations to calculate the probabilities of all possible configurations.

white/red	6/0	5/1	4/2	3/3	2/4	1/5	0/6
$\mathbb{P}$	2.8%	17.0%	34.7%	30.8%	12.5%	2.1%	0.1%

As Example 10.8 demonstrates, if an urn contains n balls painted two colors, n_1 of color A and $n_2 = n - n_1$ of color B, the the probability of randomly selecting k balls with k_1 of color A and $k_2 = k - k_1$ of color B is

$$\mathbb{P}(k_1\ A, k_2\ B) = \frac{\binom{n_1}{k_1}\binom{n_2}{k_2}}{\binom{n}{k}} = \frac{\binom{n_1}{k_1}\binom{n-n_1}{k-k_1}}{\binom{n}{k}}.$$

Since the experiment involves pulling k balls out of the urn and they must be divided in *some* manner between the two colors, this proves the following identity for binomial coefficients:

$$\sum_{k_1=0}^{k} \binom{n_1}{k_1}\binom{n-n_1}{k-k_1} = \binom{n}{k}.$$

Example 10.9. Most lotteries are ball and urn games. For example, Lotto 6/49 (popular in Western Canada) is played as follows: 49 balls (labeled 1 through 49) are randomized, and 6 of them are selected. People buy tickets with 6 distinct numbers on them; if all 6 numbers on the ticket match the numbers chosen from the urn, the ticket holder wins the grand prize (valued in the millions of dollars). There are also smaller prizes for matching at least 3 numbers.

There are $\binom{49}{6} = 13,983,816$ possible lotto draws. So, the probability of matching all 6 is, of course, 1 out of $13,983,816$ (about 7 out of 100 million). But matching 3 numbers is much more likely. Here, we have divided the balls into two groups: the 6 that were chosen, and the remaining 43. So the probability of 6 randomly chosen numbers (those on your ticket) having 3 matching (from the 6 winning) and 3 not matching (from the 43 others) is

$$\frac{\binom{6}{3}\binom{43}{3}}{\binom{49}{6}} \approx 1.8\%.$$

Similarly, we can calculate the probabilities of any number of balls matching: the probability of $k \in \{0,1,2,3,4,5,6\}$ matching is

$$\frac{\binom{6}{k}\binom{43}{6-k}}{\binom{49}{6}}.$$

These numbers are approximated in the following chart.

k	0	1	2	3	4	5	6
$\mathbb{P}$	43.6%	41.3%	13.2%	1.8%	0.1%	0.002%	0.000007%

Example 10.10. In a district election for local government in Queens, NY in 1968, Andrew V. Ippolito received 1405 votes, while his opponent received 1422 (a margin of only 17 votes). After the election, it was noticed that 101 more votes were cast than the number of registered voters in the district. Thus, 101 votes should be disqualified – though there is no way to know which candidates those 101 voted for. A district court judge ordered a new election, on the grounds that "it does not strain the probabilities to assume a likelihood that the questioned votes produced or could produce a change in the result."

We can view this as a ball and urn question. There were $1405+1422 = 2827$ votes cast, 1405 for Ippolito, and 1422 for his opponent. Suppose we choose 101 of these votes randomly. The probability that k are for Ippolito and $101-k$ are for his opponent is

$$\frac{\binom{1405}{k}\binom{1422}{101-k}}{\binom{2827}{101}}.$$

If we throw these randomly selected 101 votes away, there are 2726 votes remaining, with $1405-k$ for Ippolito, and $1422-(101-k)=1321+k$ for his opponent. In order to reverse the result of the election, it must be that $1405-k>1321+k$, which means $k<42$. Hence, the probability that a randomly-selected 101 votes removed would reverse the election results is

$$\sum_{k=0}^{41}\frac{\binom{1405}{k}\binom{1422}{101-k}}{\binom{2827}{101}}.$$

It would be no fun to sum this up by hand! But a computer program like Maple can handle it with little trouble: it is $\approx 3.87\%$. It is, of course, a matter of opinion, but many would say that this *does* "strain the probabilities" (since the disqualification would still result in a win for the opponent with over 96% certainty).

11. LECTURE 11: OCTOBER 20, 2010

11.1. **The Principle of Inclusion-Exclusion.** We have already seen (and multiply-used) the identity

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

What if we are interested in the probability of a triple union, $\mathbb{P}(A \cup B \cup C)$? We can handle this iteratively. Let $D = A \cup B$; then

$$\mathbb{P}(A \cup B \cup C) = \mathbb{P}(D \cup C) = \mathbb{P}(D) + \mathbb{P}(C) - \mathbb{P}(D \cap C).$$

Well, for $\mathbb{P}(D)$ we have

$$\mathbb{P}(D) = \mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B).$$

Also, $D \cap C = (A \cup B) \cap C = (A \cap C) \cup (B \cap C)$, and so

$$\mathbb{P}(D \cap C) = \mathbb{P}\left((A \cap C) \cup (B \cap C)\right) = \mathbb{P}(A \cap C) + \mathbb{P}(B \cap C) - \mathbb{P}\left((A \cap C) \cap (B \cap C)\right).$$

Finally, since $(A \cap C) \cap (B \cap C) = A \cap B \cap C$, putting all the pieces together we get

$$\mathbb{P}(A \cup B \cup C) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B) + \mathbb{P}(C) - [\mathbb{P}(A \cap C) + \mathbb{P}(B \cap C) - \mathbb{P}(A \cap B \cap C)].$$

Simplifying and reordering, this is neatly expressed as

$$\mathbb{P}(A \cup B \cup C) = \mathbb{P}(A) + \mathbb{P}(B) + \mathbb{P}(C) - \mathbb{P}(A \cap B) - \mathbb{P}(A \cap C) - \mathbb{P}(B \cap C) + \mathbb{P}(A \cap B \cap C).$$

This is the triple union version of the *principle of inclusion-exclusion*.

Example 11.1. *Three fair dice are rolled. What is the probability that at least one 6 comes up?*

We could calculate this by looking at the complement: there are $5^3 = 125$ outcomes with no 6s out of the possible 216, hence the probability we are after is $1 - \frac{125}{216} = \frac{91}{216}$. Now, let's calculated it instead using inclusion-exclusion. Let A_i be the event that 6 comes up on the ith roll. Then we are interested in $\mathbb{P}(A_1 \cup A_2 \cup A_3)$, which we can calculate as

$$\mathbb{P}(A_1 \cup A_2 \cup A_3) = \mathbb{P}(A_1) + \mathbb{P}(A_2) + \mathbb{P}(A_3) - \mathbb{P}(A_1 \cap A_2) - \mathbb{P}(A_1 \cap A_3) - \mathbb{P}(A_2 \cap A_3) + \mathbb{P}(A_1 \cap A_2 \cap A_3).$$

Because the dice are fair, $\mathbb{P}(A_1) = \mathbb{P}(A_2) = \mathbb{P}(A_3) = \frac{1}{6}$. Because the rolls are independent, $\mathbb{P}(A_1 \cap A_2) = \mathbb{P}(A_1 \cap A_3) = \mathbb{P}(A_2 \cap A_3) = \frac{1}{36}$, and $\mathbb{P}(A_1 \cap A_2 \cap A_3) = \frac{1}{216}$. Hence,

$$\mathbb{P}(A_1 \cup A_2 \cup A_3) = 3 \cdot \frac{1}{6} - 3 \cdot \frac{1}{36} + \frac{1}{216} = \frac{91}{216},$$

as expected.

One way to view the inclusion-exclusion formula is by successive approximations. If A, B, C are all disjoint, then we know $\mathbb{P}(A \cup B \cup C) = \mathbb{P}(A) + \mathbb{P}(B) + \mathbb{P}(C)$; in general, this sum counts intersecting pieces more than once, and so

$$\mathbb{P}(A \cup B \cup C) \leq \mathbb{P}(A) + \mathbb{P}(B) + \mathbb{P}(C).$$

We can try to account for the intersecting pieces, which are $A \cap B$, $A \cap C$, and $B \cap C$, each of which is counted twice when adding up $\mathbb{P}(A) + \mathbb{P}(B) + \mathbb{P}(C)$; so we subtract them off. But this overcompensates, since the common intersection $A \cap B \cap C$ then gets subtracted out 3 times (when we only wanted to subtract it out 2 times). So we have

$$\mathbb{P}(A \cup B \cup C) \geq \mathbb{P}(A) + \mathbb{P}(B) + \mathbb{P}(C) - \mathbb{P}(A \cap B) - \mathbb{P}(A \cap C) - \mathbb{P}(B \cap C).$$

Finally, to recompensate again, we add back in $\mathbb{P}(A \cup B \cup C)$ and get equality. The two inequalities we proved en route can be useful in themselves; they are called the *Bonferroni inequalities*.

This compensating counting procedure allows us to get the general inclusion-exclusion formula for any finite number of sets in a union.

Theorem 11.2 (The Inclusion-Exclusion Principle)**.** *Let $A_1, A_2, \ldots, A_n$ be events in a probability space $(\Omega, \mathbb{P})$. Then*

$$\begin{aligned}\mathbb{P}(A_1 \cup \cdots \cup A_n) = &\sum_{1 \le i \le n} \mathbb{P}(A_i) - \sum_{1 \le i < j \le n} \mathbb{P}(A_i \cap A_j) + \sum_{1 \le i < j < j \le n} \mathbb{P}(A_i \cap A_j \cap A_k) \\ &- \cdots + (-1)^{n+1} \mathbb{P}(A_1 \cap \cdots \cap A_n).\end{aligned}$$

Proof. We could prove this by induction, following the method we used to prove the $n = 3$ case above. Instead, let's think about it in terms of compensating and overcounting. (The argument here only really works in the discrete case.) For $0 \le k \le \Omega$, define $B_k \subseteq \Omega$ to be the set of all outcomes ω that are in exactly k of the sets $A_1, A_2, \ldots, A_k$. Then $A_1 \cup \cdots \cup A_n = B_1 \cup \cdots \cup B_n$. But the B_k are disjoint. Thus

$$\mathbb{P}(A_1 \cup \cdots \cup A_n) = \sum_{k=1}^{n} \mathbb{P}(B_k) = \sum_{k=1}^{n} \sum_{\omega \in B_k} \mathbb{P}(\omega).$$

Now, suppose $\omega \in B_k$.

- The sum $\sum_i \mathbb{P}(A_i)$ counts $\mathbb{P}(\omega)$ exactly $\binom{k}{1}$ times.
- The sum $\sum_{i<j} \mathbb{P}(A_i \cap A_j)$ counts $\mathbb{P}(\omega)$ exactly $\binom{k}{2}$ times.

  $\vdots$
- The sum $\sum_{i_1 < i_2 < \cdots < i_k} \mathbb{P}(A_{i_1} \cap \cdots \cap A_{i_k})$ counts $\mathbb{P}(\omega)$ exactly $\binom{k}{k}$ times.

For $m < k$, none of the terms in the sum $\sum_{i_1 < \cdots < i_m} \mathbb{P}(A_{i_1} \cap \cdots A_{i_m})$ does not count $\mathbb{P}(\omega)$ as all. Thus, the inclusion-exclusion formula above counts $\mathbb{P}(\omega)$

$$\binom{k}{1} - \binom{k}{2} + \binom{k}{3} - \cdots + (-1)^{k+1} \binom{k}{k}$$

times. Well, cleverly invoking the binomial theorem, we have

$$0 = (-1+1)^k = \sum_{j=0}^{k} \binom{k}{j} (-1)^j (1)^{k-j} = \binom{k}{0} - \binom{k}{1} + \binom{k}{2} - \cdots + (-1)^k \binom{k}{k}.$$

Subtracting the last k terms from both sides gives us

$$\binom{k}{1} - \binom{k}{2} + \binom{k}{3} - \cdots + (-1)^{k+1} \binom{k}{k} = \binom{k}{0} = 1.$$

Whence, the inclusion-exclusion formula counts the probability of each $\omega \in B_k$ exactly once. Since the disjoint union of the B_k's is the union $A_1 \cup \cdots \cup A_n$, this proves the theorem. □

Example 11.3. *A fair die is rolled 10 times. How likely is it that we don't see each of the numbers at least once?*

Let A_i be the event that we never see the number i. Since the probability of i in any given roll is $\frac{1}{6}$, each of the events A_i have probability $(\frac{5}{6})^{10}$. Now, if $i \neq j$, the event $A_i \cap A_j$ is the event that the other four numbers (rather than i, j) appear in the 10 independent trials, so $\mathbb{P}(A_i \cap A_j) = (\frac{4}{6})^{10}$. In general, if $i_1 < i_2 < \cdots < i_k$ with $k \leq 6$, we have $\mathbb{P}(A_{i_1} \cap \cdots \cap A_{i_k}) = (\frac{6-k}{6})^{10}$. (In particular, when $k = 6$ we have $\mathbb{P}(A_1 \cap \cdots \cap A_6) = 0$, of course, since this event is that no number ever comes up, which is impossible.)

The event that not all numbers show up is the event that at least one number never shows up, i.e. $A_1 \cup A_2 \cup \cdots \cup A_6$. Now, in the inclusion-exclusion formula, all the terms in the sum

$$\sum_{1 \leq i_1 < \cdots < i_k \leq 6} \mathbb{P}(A_{i_1} \cap A_{i_k})$$

are equal to $(\frac{6-k}{6})^{10}$, and so this sum is equal to

$$\binom{6}{k} \left(\frac{6-k}{6}\right)^{10}.$$

Hence, the inclusion-exclusion formula tells us that

$$\begin{aligned}\mathbb{P}(A_1 \cup \cdots \cup A_6) &= \binom{6}{1}\left(\frac{5}{6}\right)^{10} - \binom{6}{2}\left(\frac{4}{6}\right)^{10} + \binom{6}{3}\left(\frac{3}{6}\right)^{10} - \binom{6}{4}\left(\frac{2}{6}\right)^{10} + \binom{6}{5}\left(\frac{1}{6}\right)^{10}. \\ &= \frac{101923}{139968} \approx 72.8\%.\end{aligned}$$

What if we'd rolled the dice more times? Changing 10 to 15 in the formula gives an answer of about 35.6%. Much smaller. If we roll 20 times, the probability drops to about 15.2%. In general, it falls off exponentially as the number of rolls increases.

Example 11.4. *A deck of n cards are shuffled well. You are asked to guess at what order they are in. What is the probability you get at least one card right?*

Let A_i be the event that you get the ith card right. (A_i includes outcomes where more than the ith card is correct.) We want to calculate $\mathbb{P}(A_1 \cup \cdots \cup A_n)$. To use inclusion-exclusion, we need to calculate all probabilities of intersections $\mathbb{P}(A_{i_1} \cap \cdots \cap A_{i_k})$ for $1 \leq i_1 < \cdots < i_k \leq n$, for all $k \leq n$. Fortunately, these probabilities don't depend in the particular indices but only on the number of them (as in Example 11.3). The event $A_{i_1} \cap \cdots \cap A_{i_k}$ is the event that cards $i_1, \ldots, i_k$ are correct, and the others may or may not be. There are $n - k$ remaining cards, and there are therefore $(n-k)!$ possible orderings of them; since the cards were well shuffled, all of these orderings are equally likely. There are $n!$ possible orderings in total, and hence

$$\mathbb{P}(A_{i_1} \cap \cdots A_{i_k}) = \frac{(n-k)!}{n!}.$$

Thus, the inclusion-exclusion formula gives us

$$\mathbb{P}(A_1 \cup \cdots \cup A_n) = \binom{n}{1}\frac{(n-1)!}{n!} - \binom{n}{2}\frac{(n-2)!}{n!} + \binom{n}{3}\frac{(n-3)!}{n!} - \cdots + (-1)^{n+1}\binom{n}{n}\frac{0!}{n!}.$$

In this sum, we have terms of the form

$$(-1)^{k+1}\binom{n}{k}\frac{(n-k)!}{n!} = (-1)^{k+1}\frac{n!}{k!(n-k)!}\frac{(n-k)!}{n!} = \frac{(-1)^{k+1}}{k!}.$$

That is

$$\mathbb{P}(A_1 \cup \cdots \cup A_n) = \sum_{k=1}^{n} \frac{(-1)^{k+1}}{k!}.$$

We could calculate this exactly for different n; for example, when $n = 52$ we get approximately 0.63212. It is much easier, however, to notice that

$$1 - \mathbb{P}(A_1 \cup \cdots \cup A_n) = 1 + \sum_{k=1}^{n} \frac{(-1)^k}{k!} = \sum_{k=0}^{n} \frac{(-1)^k}{k!}.$$

This is the nth Taylor-series approximation for the series $e^{-1} = \sum_{k=0}^{\infty} \frac{(-1)^k}{k!}$. Hence, for large n,

$$\mathbb{P}(A_1 \cup \cdots \cup A_n) \approx 1 - e^{-1}.$$

In fact, with $n = 52$, this agrees with the exact answer to 69 decimal places.

Of particular interest is the following consequence: no matter *how* many cards there are, there is still a roughly 36.8% chance you will get *none* right!

12. LECTURE 12: OCTOBER 22, 2010

12.1. **Estimates Using Inclusion-Exclusion.** Thinking back to the $n = 3$ case in the proof of the inclusion-exclusion formula, if we cut off the sum at some point, we get either an upper or a lower bound for the actual value.

Theorem 12.1 (Bonferroni's Inequalities)**.** *Let $1 \leq m \leq n$, and define*

$$P_m \equiv \sum_{j=1}^{m} (-1)^{j+1} \sum_{1 \leq i_1 < \cdots < i_j \leq n} \mathbb{P}(A_{i_1} \cap \cdots \cap A_{i_j}).$$

(The principle of inclusion-exclusion, Theorem 11.2, states that $P_n = \mathbb{P}(A_1 \cup \cdots \cup A_n)$.) Then for $1 \leq \ell < n/2$,

$$P_{2\ell} \leq \mathbb{P}(A_1 \cup \cdots \cup A_n) \leq P_{2\ell-1}.$$

In other words, if you cut off the sum in inclusion-exclusion, you get an upper-bound if the sum ends in a $+$, or a lower bound if the sum ends in a $-$. In particular,

$$\sum_{i=1}^{n} \mathbb{P}(A_i) - \sum_{1 \leq i < j \leq n} \mathbb{P}(A_i \cap A_j) \ \leq \ \mathbb{P}(A_1 \cup \cdots \cup A_n) \ \leq \ \sum_{i=1}^{n} \mathbb{P}(A_j).$$

Proof. We approach this just as in the proof of Theorem 11.2. Let B_k be the set of outcomes ω that are in exactly k of the events $A_1, \ldots, A_n$ for $1 \leq k \leq n$. Then, following exactly the proof of Theorem 11.2, the sum P_m counts the element $\mathbb{P}(\omega)$

$$\binom{k}{1} - \binom{k}{2} + \binom{k}{3} - \cdots + (-1)^{m+1} \binom{k}{m} = \sum_{j=1}^{m} (-1)^{j+1} \binom{k}{j}$$

times, provided $m \leq k$. (If $k < m$ then P_m counts $\mathbb{P}(\omega)$ as many times as does P_n since ω is not in any of the higher intersections; as we showed in the proof of Theorem 11.2, $\mathbb{P}(\omega)$ is counted exactly once in this case. We can dispense with this discussion by defining $\binom{k}{m} = 0$ when $m > k$.) We want to see how different this is from the 1 time we want it to be counted, so we look at $P_m - 1$, which counts $\mathbb{P}(\omega)$

$$-1 + \sum_{j=1}^{m} (-1)^{j+1} \binom{k}{j} = \sum_{j=0}^{m} (-1)^{j+1} \binom{k}{j}$$

times. We want to compare the sign of this with the sign of the final term in the sum P_m, which is $(-1)^{m+1}$. So it makes sense to divide by this sign: $(-1)^{-m-1}(P_m - 1)$ counts $\mathbb{P}(\omega)$ exactly

$$(-1)^{-m-1} \sum_{k=0}^{m} (-1)^{j+1} \binom{k}{j} = \sum_{j=0}^{m} (-1)^{m-j} \binom{k}{j}$$

times. Remarkably, this sum can be explicitly calculated. Recall the relation from Pascal's triangle: $\binom{k}{j} = \binom{k-1}{j} + \binom{k-1}{j-1}$. Thus

$$\sum_{j=0}^{m} (-1)^{m-j} \binom{k}{j} = \sum_{j=0}^{m} (-1)^{m-j} \left[\binom{k-1}{j} + \binom{k-1}{j-1} \right].$$

This is actually a telescoping sum. Let's write out a few terms to see this, starting at the top $j = m$:

$$\left[\binom{k-1}{m} + \binom{k-1}{m-1}\right] - \left[\binom{k-1}{m-1} + \binom{k-1}{m-2}\right] + \left[\binom{k-1}{m-2} + \binom{k-1}{m-3}\right] - \cdots$$
$$\cdots + (-1)^{m-1}\left[\binom{k-1}{1} + \binom{k-1}{0}\right] + (-1)^m\left[\binom{k-1}{0}\right].$$

(The last term is just $\binom{k}{0} = \binom{k-1}{0}$; Pascal's triangle cuts off at the edges. Again we could dispense with this by making the convention $\binom{k}{j} = 0$ when $j < 0$.) All of the terms cancel in pairs except for the very first, and so we have the remarkable formula

$$\sum_{j=0}^{m}(-1)^{m-j}\binom{k}{j} = \binom{k-1}{m}. \tag{12.1}$$

Thus, $(-1)^{-m-1}(P_m - 1)$ counts $\mathbb{P}(\omega)$ $\binom{k-1}{m}$ times, which always non-negative. So $P_m - 1$ counts $\mathbb{P}(\omega)$ $(-1)^m\binom{k-1}{m}$ times, which is positive when P_m ends in a $+$ and negative when P_m ends in a $-$,regardless of the value of k (between 1 and n). This therefore holds true for any $\omega \in A_1 \cup \cdots \cup A_n$: $\mathbb{P}(\omega)$ is counted at least 1 time of P_m ends in a $+$, and is counted at most 1 time if P_m ends in a $-$. This proves the theorem. □

Remark 12.2. One might hope that taking more terms in inclusion-exclusion will decrease the error in Bonferroni's inequalities; i.e. one might hope that $|P_m - \mathbb{P}(A_1 \cup \cdots \cup A_n)|$ decreases as m increases. Unfortunately, this is typically not true. For example, consider the case $A_1 = A_2 = \cdots = A_n$. Then the union is equal to A_1, and so every point in the union is in the intersection of n of the sets; hence, the above analysis shows that for any point $\omega \in A_1$, $\mathbb{P}(\omega)$ is counted $\binom{n-1}{m}$ times by $(-1)^{-m-1}(P_m - 1)$. In other words, each $\mathbb{P}(\omega)$ is counted $1 + (-1)^{m+1}\binom{n-1}{m}$ times by P_m, and so $P_m = (1 + (-1)^{m+1}\binom{n-1}{m})\mathbb{P}(A_1)$. Since $\mathbb{P}(A_1) = \mathbb{P}(A_1 \cup \cdots \cup A_n)$, we see that in this case

$$|P_m - \mathbb{P}(A_1 \cup \cdots \cup A_n)| = \binom{n-1}{m}\mathbb{P}(A_1).$$

As m grows, this grows for a while, maxing out near $\frac{n-1}{2}$, then decreasing down to 0 when $m = n$.

Example 12.3. Let's use Bonferroni's inequalities to estimate the probability of three people sharing the same birthday in a crowd of n (random, independent) people. There are $\binom{n}{3}$ possible groups of 3 in the crowd; for each such group g let A_g be the event that the three have the same birthday. Then $\mathbb{P}(A_g) = \frac{1}{365^2}$ for any group g (pick one of the three; the other two must each have the same birthday is the one chosen). Now, we want to calculate

$$\mathbb{P}\left(\bigcup_g A_g\right).$$

To get an upper-bound, Bonferroni's first inequality yields

$$\mathbb{P}\left(\bigcup_g A_g\right) \leq \sum_g \mathbb{P}(A_g) = \binom{n}{3}\frac{1}{365^2}.$$

For example, when $n = 60$, this gives 25.7%; when $n = 80$, we get 61.7%; and when $n = 100$, we get 121.4% – this last one is a pretty bad estimate! Now for the lower bound,

$$\mathbb{P}\left(\bigcup_g A_g\right) \geq \sum_g \mathbb{P}(A_g) - \sum_{g<g'} \mathbb{P}(A_g \cap A_{g'}).$$

(Here $g < g'$ means we are counting over pairs (g, g') where $g \neq g'$, but we don't count both (g, g') and (g', g), only one of them.) For distinct groups g, g', the probability of the event $A_g \cap A_{g'}$ depends on the intersection of g, g'. Since $g \neq g'$, the two groups of 3 can overlap by 0, 1, or 2 people. So we have (using the upper-bound calculation above)

$$\mathbb{P}(\bigcup_g A_g) \geq \binom{n}{3}\frac{1}{365^2} - \sum_{|g\cap g'|=0} \mathbb{P}(A_g \cap A_{g'}) - \sum_{|g\cap g'|=1} \mathbb{P}(A_g \cap A_{g'}) - \sum_{|g\cap g'|=2} \mathbb{P}(A_g \cap A_{g'}).$$

If the two groups g, g' do not intersect ($|g \cap g'| = 0$), then the event $A_g \cap A'_g$ simply means the first three share a birthday, and the second three share a birthday; the probability of this is $(\frac{1}{365^2})^2$. (In other words: if g and g' do not intersect, then the events A_g and $A_{g'}$ are independent.) The number of such disjoint pairs of 3-person groups is counted by a multinomial coefficient: we must divide the n people into two groups of 3 along with the remaining group of $n - 6$. Thus

$$\sum_{|g\cap g'|=0} \mathbb{P}(A_g \cap A_{g'}) = \binom{n}{3\ 3\ n-6}\frac{1}{365^4}.$$

Now, consider those pairs of groups g, g' that have one common member, $|g \cap g'| = 1$. In this case, the event $A_g \cap A_{g'}$, that everyone in the first group has the same birthday and everyone in the second group has the same birthday, forces the 5 people to share the same birthday, so $\mathbb{P}(A_g \cap A_{g'}) = \frac{1}{365^4}$. Choosing such groups is equivalent to dividing the n people into groups of size $1, 2, 2, n - 5$ (the intersection, the two others in each of the groups, and everyone else). Thus

$$\sum_{|g\cap g'|=1} \mathbb{P}(A_g \cap A_{g'}) = \binom{n}{1\ 2\ 2\ n-5}\frac{1}{365^4}.$$

For the final terms, consider pairs of groups g, g' with an overlap of two people, $|g\cap g'| = 2$. Again, the event $A_g \cap A_{g'}$ in this case forces all people in the two groups to share one birthday; there are 4 people in the two groups, so $\mathbb{P}(A_g \cap A_{g'}) = \frac{1}{365^3}$. Choosing two such groups is equivalent to dividing the n people into groups of sizes $2, 1, 1, n - 4$ (the intersection, the one other for each group, and everyone else). Thus,

$$\sum_{|g\cap g'|=2} \mathbb{P}(A_g \cap A_{g'}) = \binom{n}{2\ 1\ 1\ n-4}\frac{1}{365^3}.$$

So, finally, the second Bonferroni inequality gives us the lower bound

$$\binom{n}{3}\frac{1}{365^2} - \binom{n}{3\ 3\ n-6}\frac{1}{365^4} - \binom{n}{1\ 2\ 2\ n-5}\frac{1}{365^4} - \binom{n}{2\ 1\ 1\ n-4}\frac{1}{365^3}$$

for the probability that at least one group of three people in n share a common birthday. Evaluating at $n = 60$ yields 7.1%; with $n = 80$ we get -15.3%; with $n = 100$, we get -122.4%.

So, the Bonferroni inequalities give us no information quickly as n grows. We do, however, have the information that

$$7.1\% \leq \mathbb{P}(3 \text{ people among } 60 \text{ share a birthday}) \leq 25.7\%.$$

As you can see, the birthday problem (with more than two people) is quite hard!

12.2. **Conditional Probability.** Recall the definition of conditional probability: if A, B are two events in a probability space $(\Omega, \mathbb{P})$ and $\mathbb{P}(B) > 0$, then the probability of A *given that* B *occurs* is

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

As motivation, think again in terms of the frequency-interpretation of probability: if we perform many many trials,

$$\mathbb{P}(A) \approx \frac{\#\text{ trials where } A \text{ occurs}}{\#\text{ trials}}.$$

Now, if we want to observe the frequency of A given that B occurs, we look through all of our data, single out those trials where B occurs, and look at how many among *them* where A also occurs; thus

$$\mathbb{P}(A|B) \approx \frac{\#\text{ trials where } B \text{ occurs, and } A \text{ also occurs}}{\#\text{ trials where } B \text{ occurs}}.$$

Dividing top and bottom by the number of trials yields

$$\begin{aligned}\mathbb{P}(A|B) &\approx \frac{\#\text{ trials where } B \text{ occurs, and } A \text{ also occurs}}{\#\text{ trials where } B \text{ occurs}} \\ &= \frac{(\#\text{ trials where } B \text{ occurs, and } A \text{ also occurs})/(\#\text{ trials})}{(\#\text{ trials where } B \text{ occurs})/(\#\text{ trials})} \\ &\approx \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}.\end{aligned}$$

Returning to mathematics, notice the following properties that $\mathbb{P}(A|B)$ has.

- $\mathbb{P}(\Omega|B) = \frac{\mathbb{P}(\Omega \cap B)}{\mathbb{P}(B)} = \frac{\mathbb{P}(B)}{\mathbb{P}(B)} = 1$. Thus, $\mathbb{P}(\,\cdot\,|B)$ assigns mass 1 to the whole sample space.
- If A_1, A_2 are disjoint, then

  $$\mathbb{P}(A_1 \cup A_2|B) = \frac{\mathbb{P}((A_1 \cup A_2) \cap B)}{\mathbb{P}(B)} = \frac{\mathbb{P}([A_1 \cap B] \cup [A_2 \cap B])}{\mathbb{P}(B)}.$$

  Then $A_1 \cap B$ and $A_2 \cap B$ are also disjoint, and so since $\mathbb{P}$ is additive over disjoint unions,

$$\mathbb{P}(A_1 \cup A_2|B) = \frac{\mathbb{P}(A_1 \cap B) + \mathbb{P}(A_2 \cap B)}{\mathbb{P}(B)} = \frac{\mathbb{P}(A_1 \cap B)}{\mathbb{P}(B)} + \frac{\mathbb{P}(A_2 \cap B)}{\mathbb{P}(B)} = \mathbb{P}(A_1|B) + \mathbb{P}(A_2|B).$$

  In other words, $\mathbb{P}(\,\cdot\,|B)$ is additive over disjoint unions.

- Suppose $A_n \uparrow A$ (see Lecture 1: this means $A_1 \subseteq A_2 \subseteq \cdots$ and $A = \bigcup_{n=1}^{\infty} A_n$). Then $A_1 \cap B \subseteq A_2 \cap B \subseteq \cdots$ and $\bigcup_{n=1}^{\infty} A_n \cap B = A \cap B$, so $A_n \cap B \uparrow A \cap B$, and by the continuity of $\mathbb{P}$ we get

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

  as $n \to \infty$. In other words, $\mathbb{P}(\,\cdot\,|B)$ is continuous.

What these items show is that *the conditional probability $\mathbb{P}(\,\cdot\,|B)$ is, itself, a probability measure on Ω.* And actually, it is often easier to *start* with this probability (i.e. we often have more intuition about it in a given experiment), to work out its constituents $\mathbb{P}(A)$ and $\mathbb{P}(A \cap B)$ indirectly.

Example 12.4. *Two cards are taken from a well-shuffled deck of 52 cards. What is the probability they are both clubs?*

Let C_i be the event that the ith card is a club, $i = 1, 2$. We are interested in calculating $\mathbb{P}(C_1 \cap C_2)$. Well, $\mathbb{P}(C_2|C_1) = \mathbb{P}(C_1 \cap C_2)/\mathbb{P}(C_1)$. The denominator is easy: picking one cards out of 52, the chance that it is a club is $13/52$, so $\mathbb{P}(C_1) = \frac{1}{4}$. Now, *in the event that we chose a club* (i.e. conditioning on C_1), there are 51 cards remaining, and 12 of them are clubs. Hence, $\mathbb{P}(C_2|C_1) = \frac{12}{51}$. So, we have the equation

$$\frac{12}{51} = \frac{\mathbb{P}(C_1 \cap C_2)}{\frac{13}{52}},$$

and so $\mathbb{P}(C_1 \cap C_2) = \frac{13}{52} \cdot \frac{12}{51}$.

The method demonstrated in Example 12.4 is so ubiquitous, we give it a name: the **multiplication rule**. It is a trivial consequence of the definition, but it is very useful in this form:

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

In fact, we can continue this way: if we have 3 events A_1, A_2, A_3, then

$$\mathbb{P}(A_1 \cap A_2 \cap A_3) = \mathbb{P}(A_1) \cdot \frac{\mathbb{P}(A_1 \cap A_2)}{\mathbb{P}(A_1)} \cdot \frac{\mathbb{P}(A_1 \cap A_2 \cap A_3)}{\mathbb{P}(A_1 \cap A_2)}.$$

The second term is $\mathbb{P}(A_2|A_1)$. If we write the numerator of the third term as $\mathbb{P}((A_1 \cap A_2) \cap A_3)$, we recognize this as $\mathbb{P}(A_3|A_1 \cap A_2)$, and so the multiplication rule says

$$\mathbb{P}(A_1 \cap A_2 \cap A_3) = \mathbb{P}(A_1)\mathbb{P}(A_2|A_1)\mathbb{P}(A_3|A_1 \cap A_2).$$

In general, we have

$$\mathbb{P}(A_1 \cap \cdots \cap A_n) = \mathbb{P}(A_1)\mathbb{P}(A_2|A_1)\mathbb{P}(A_3|A_1 \cap A_2) \cdots \mathbb{P}(A_n|A_1 \cap \cdots \cap A_{n-1}).$$

Example 12.5. *Five cards are dealt from a well-shuffled deck of 52 cards. What is the probability they form a flush (i.e. all cards of the same suit)?*

For $n = 1, 2, 3, 4, 5$ let F_n be the event that the nth card is the same suit as the $n-1$st card (so F_1 is just the event that we get a card, i.e. $F_1 = \Omega$). A flush is the event $F_1 \cap F_2 \cap F_3 \cap F_4 \cap F_5$. We calculate this as

$$\mathbb{P}(F_1)\mathbb{P}(F_2|F_1)\mathbb{P}(F_3|F_1 \cap F_2)\mathbb{P}(F_4|F_1 \cap F_2 \cap F_3)\mathbb{P}(F_5|F_1 \cap F_2 \cap F_3 \cap F_4).$$

By definition, $\mathbb{P}(F_1) = 1$. Now, once F_1 occurs, there are only 12 cards left in its suit, out of 51 total, so $\mathbb{P}(F_2|F_1) = \frac{12}{51}$. Once F_1 and F_2 occur, there are only 11 cards left of that suit, among 50 total, so $\mathbb{P}(F_3|F_1 \cap F_2) = \frac{11}{50}$. Continuing this way, we have

$$\mathbb{P}(F_1 \cap \cdots \cap F_5) = 1 \cdot \frac{12}{51} \cdot \frac{11}{50} \cdot \frac{10}{49} \cdot \frac{9}{48} \approx 0.198\%.$$

13. LECTURE 13: OCTOBER 25, 2010

13.1. **Law of Total Probability.** There are many cases in which conditional probabilities are easier to determine than probabilities of intersections directly. This can help us calculate probabilities using the following general setup. Let A be an event, and let $B_1, B_2, \ldots, B_n$ be a collection of disjoint events whose union is all of Ω. (In other words, in any outcome, exactly one of the B_i occurs.) Such a collection is called a **partition**. Then

$$A = A \cap \Omega = A \cap (B_1 \cup \cdots \cup B_n) = (A \cap B_1) \cup \cdots \cup (A \cap B_n),$$

and the union is disjoint. Hence

$$\mathbb{P}(A) = \sum_{k=1}^{n} \mathbb{P}(A \cap B_k).$$

Now, we use the multiplication rule, $\mathbb{P}(A \cap B_k) = \mathbb{P}(A|B_k)\mathbb{P}(B_k)$, and thus we have the **law of total probability**:

$$\mathbb{P}(A) = \sum_{k=1}^{n} \mathbb{P}(B_k)\mathbb{P}(A|B_k).$$

Example 13.1. *An urn contains 5 red and 10 black balls. 2 balls are drawn from the urn, without replacement. What is the probability that the second ball drawn is red?*

Let R_i be the event that the ith draw is red, and B_i the event that the ith draw is black. We want to calculate $\mathbb{P}(R_2)$. To do so, we notice that $\{B_1, R_1\}$ is a partition, so the law of total probability gives us

$$\mathbb{P}(R_2) = \mathbb{P}(R_1)\mathbb{P}(R_2|R_1) + \mathbb{P}(B_1)\mathbb{P}(R_2|B_1).$$

There are 15 balls in total, so the first ball is red with probability $\frac{5}{15} = \frac{1}{3}$ and black with probability $\frac{10}{15} = \frac{2}{3}$; so

$$\mathbb{P}(R_2) = \frac{1}{3}\mathbb{P}(R_2|R_1) + \frac{2}{3}\mathbb{P}(R_2|B_1).$$

Now, if the first ball is red, there are only 4 red balls left out of the 14, so $\mathbb{P}(R_2|R_1) = \frac{4}{14} = \frac{2}{7}$. On the other hand, if a blue ball is chosen first, then all 5 red balls remain out of the 14, so $\mathbb{P}(R_2|B_1) = \frac{5}{14}$. In total,

$$\mathbb{P}(R_2) = \frac{1}{3} \cdot \frac{2}{7} + \frac{2}{3} \cdot \frac{5}{14} = \frac{2}{21} + \frac{10}{42} = \frac{1}{3}.$$

Many examples in conditional probability can be phrased in the form of a **two-stage experiment**: one trial is carried out, and the results of that trial determine how to proceed with the second trial. (Example 13.1 can be thought of in those terms, except that there the second trial is conducted the same way no matter what happened in the first trial.)

Example 13.2. *A fair die is rolled; then a fair coin is tossed the number of times that came up on the die. What is the probability of exactly 3 heads?*

Let D_k be the event that the die comes up $k \in \{1, 2, 3, 4, 5, 6\}$. Then the D_k form a partition. In this case, $\mathbb{P}(D_k) = \frac{1}{6}$ for all k. Setting A to be the event that we get exactly 3 heads, the

law of total probability says that

$$\mathbb{P}(A) = \sum_{k=1}^{6} \mathbb{P}(D_k)\mathbb{P}(A|D_k) = \frac{1}{6}\sum_{k=1}^{6} \mathbb{P}(A|D_k).$$

Now, if D_1 or D_2 occur, then A is impossible, so $\mathbb{P}(A|D_1) = \mathbb{P}(A|D_2) = 0$. For $k \geq 3$, the occurrence of D_k means that we can flip the coin $k \geq 3$ times, and so the number of heads is a $\text{binomial}(k, \frac{1}{2})$ random variable, so $\mathbb{P}(A|D_k) = \binom{k}{3}(\frac{1}{2})^k$. Thus

$$\begin{aligned}\mathbb{P}(A) &= \frac{1}{6}\sum_{k=3}^{6} \binom{k}{3}\left(\frac{1}{2}\right)^k = \frac{1}{6}\left(\binom{3}{3}2^{-3} + \binom{4}{3}2^{-4} + \binom{5}{3}2^{-5} + \binom{6}{3}2^{-6}\right) \\ &= \frac{1}{6}\left(\frac{1}{8} + 4\cdot\frac{1}{16} + 10\cdot\frac{1}{32} + 20\cdot\frac{1}{64}\right) = \frac{1}{6}\cdot\frac{8+16+20+20}{64} = \frac{1}{6}.\end{aligned}$$

Example 13.3 (The Monty Hall Problem)**.** *On the television program* Let's Make a Deal *(1963–1976), the host Monty Hall would often put contestants in the following situation. Three doors are presented to you; you are told that two of them hide goats, while the third hides a valuable prize. You must choose a door at random, to be opened revealing your haul. After you choose, Monty opens one of the other two doors, always revealing a goat. You are then given a choice: stay with the door you selected, or switch to the remaining door. Should you switch?*

Our näive intuition would tell us that it doesn't matter whether we switch: there are two doors, one contains a prize, so the probability is $\frac{1}{2}$ that we get the prize, and $\frac{1}{2}$ that we get the goat. But this is wrong. Let's see why.

Number the doors $1, 2, 3$ with $\#1$ being the door you initially chose. Let D_i be the event that the prize is behind door $i \in \{1, 2, 3\}$. Let M_j be the event that Monty opens door $j \in \{1, 2, 3\}$. Then the nine events $\{D_i \cap M_j\}_{1\leq i,j\leq 3}$ form a partition. But five of these sets are actually empty, as the following table shows.

	Door #1	Door #2	Door #3	Monty's action
D_1	prize	goat	goat	opens door #2 or #3
D_2	goat	prize	goat	opens door #3
D_3	goat	goat	prize	opens door #2

Hence, only the events $D_1 \cap M_2$, $D_1 \cap M_3$, $D_2 \cap M_3$, and $D_3 \cap M_2$ are nonempty, and these form a partition. Now, let W be the event that you win the prize without switching. Then

$$\begin{aligned}\mathbb{P}(W) = \mathbb{P}(D_1 \cap M_2)\mathbb{P}(W|D_1 \cap M_2) + \mathbb{P}(D_1 \cap M_3)\mathbb{P}(W|D_1 \cap M_3) \\ + \mathbb{P}(D_2 \cap M_3)\mathbb{P}(W|D_2 \cap M_3) + \mathbb{P}(D_3 \cap M_2)\mathbb{P}(W|D_2 \cap M_2).\end{aligned}$$

But W only occurs (without switching) if D_1 occurs, so $\mathbb{P}(W|D_1 \cap M_2) = \mathbb{P}(W|D_1 \cap M_3) = 1$, while $\mathbb{P}(W|D_2 \cap M_3) = \mathbb{P}(W|D_3 \cap M_2) = 0$. Thus

$$\mathbb{P}(W) = \mathbb{P}(D_1 \cap M_2) + \mathbb{P}(D_1 \cap M_3).$$

We can now calculate these probabilities from the multiplication rule, since Monty selects either door 2 or door 3 randomly (with probability $\frac{1}{2}$ either way).

$$\mathbb{P}(D_1 \cap M_2) = \mathbb{P}(D_1)\mathbb{P}(M_2|D_1) = \frac{1}{3} \cdot \frac{1}{2} = \frac{1}{6}$$
$$\mathbb{P}(D_1 \cap M_3) = \mathbb{P}(D_1)\mathbb{P}(M_3|D_1) = \frac{1}{3} \cdot \frac{1}{2} = \frac{1}{6}$$

Hence, without switching, we have $\mathbb{P}(W) = \frac{1}{6} + \frac{1}{6} = \frac{1}{3}$. Thus, we should switch: we will win with probability $\frac{2}{3}$ in that case!

The faulty reasoning that might lead to the answer $\frac{1}{2}$ for the probability of winning with or without switching arises from the (false) believe that the four events forming the above partition are equally likely. This is definitely not true: in fair *two-stage* experiments (where later outcomes are determined by earlier ones), total probabilities are rarely uniform.

Remark 13.4. This problem was popularized by Marilyn vos Savant (who is in the Guiness Book of World Records for having the highest IQ on record: 228), in her column in *Parade* magazine in 1990. (This is a column where she solves challenging puzzles sent in by readers.) It was first posed as a mathematical problem in a letter to the *American Statistician* in 1975. In 1991, Monty Hall was interviewed in the New York Times, where he explained that the analysis didn't apply to the actual game, since he relied on his perceptions of the psychology of the contestant to subtly influence their decisions.

13.2. **Bayes' Theorem.** Consider the following illuminating example.

Example 13.5. *Suppose that an HIV test has 99% accuracy: if a patient is $HIV+$ then the test gives a positive result (denoted $T+$) with probability 0.99, and gives a negative result ($T-$) with probability 0.01. Similarly, if a patient is $HIV-$, the event $T+$ has probability 0.01, while the event $T-$ has probability 0.99. In other words,*

$$\mathbb{P}(T+|HIV+) = \mathbb{P}(T-|HIV-) = 0.99 \quad \mathbb{P}(T+|HIV-) = \mathbb{P}(T-|HIV+) = 0.01.$$

If a random patient tests positive, what is the probability s/he is $HIV+$?

The näive answer is 99%, but this is incorrect. What we want to know is $\mathbb{P}(HIV+|T+)$. Thinking this must also be 0.99 is known as the *prosecutor's fallacy*: the incorrect assumption that $\mathbb{P}(A|B) = \mathbb{P}(B|A)$ for different events A, B.

To see how false this assumption may be, consider the San Diego Metro area, with close to 3 million residents; the $HIV+$ population is estimated at around 500. Now, if every single San Diego resident got tested, here are the results we would expect:

- Of the 500 $HIV+$ people, $0.99 \cdot 500 = 495$ get $T+$, while $0.01 \cdot 500 = 5$ get $T-$.
- Of the $2,999,500$ $HIV-$ people, $0.01 \cdot 2,999,500 = 29,995$ get $T+$, while $0.99 \cdot 2,999,500 = 2,969,505$ get $T-$.

So, looking at those people who test $T+$, we see that 495 of them are $HIV+$ while $29,995$ are $HIV-$. There are a total of $495 + 29,995 = 30,490$ positive tests, and so the fraction of those that are actually positive is $\frac{495}{30490} \approx 1.6\%$. In other words,

$$\mathbb{P}(HIV+|T+) \approx 1.62\%, \text{ even though } \mathbb{P}(T+|HIV+) = 99\%.$$

Remark 13.6. There is nothing medical or statistical that forces the symmetry in Example 13.5. It is more typical to find a situation like $\mathbb{P}(T+|HIV+) = 0.95$ while $\mathbb{P}(T-|HIV-) = 0.85$. This will be clear in Example 13.8.

As Example 13.5 demonstrates, $P(A|B)$ and $P(B|A)$ may bear little resemblance to each other. We can see this just from their definitions:

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

So their ratio is

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

i.e.

$$\mathbb{P}(B|A) = \frac{\mathbb{P}(B) \cdot \mathbb{P}(A|B)}{\mathbb{P}(A)}. \tag{13.1}$$

We saw this in Example 13.5:

$$\mathbb{P}(HIV+|T+) = \frac{\mathbb{P}(HIV+)}{\mathbb{P}(T+)}\mathbb{P}(T+|HIV+),$$

and since $\mathbb{P}(HIV+)/\mathbb{P}(T+)$ is very small (in the example it's $\frac{500}{30490}$), the two quantities $\mathbb{P}(HIV+|T+)$ and $\mathbb{P}(T+|HIV+)$ are dramatically different.

Equation 13.1 is only useful if we have some way of knowing (or estimating) the ratio $\mathbb{P}(B)/\mathbb{P}(A)$. We can use the law of total probability to help with this: we can express $\mathbb{P}(A) = \mathbb{P}(A \cap B) + \mathbb{P}(A \cap B^c) = \mathbb{P}(B)\mathbb{P}(A|B) + \mathbb{P}(B^c)\mathbb{P}(A|B^c)$. Thus, Equation 13.1 becomes

$$\mathbb{P}(B|A) = \frac{\mathbb{P}(B)\mathbb{P}(A|B)}{\mathbb{P}(B)\mathbb{P}(A|B) + \mathbb{P}(B^c)\mathbb{P}(A|B^c)}. \tag{13.2}$$

This is sometimes called *Bayes' Formula* or *Bayes' Theorem*. **You should not memorize it.** Rather, you should remember the reasoning process that got us to it, and follow your nose with given examples.

Example 13.7. *Approximately* 1% *of American women over the age of* 50 *have breast cancer. A woman with breast cancer is* 90% *likely to test positive, while a woman who does not have breast cancer is* 10% *likely to get a false-positive. What is the probability that a woman has breast cancer, given that she gets a positive test?*

Let B be the event "she has breast cancer" and let T be the event "she tests positive". The information we have is that $\mathbb{P}(B) = 0.01$, while $\mathbb{P}(T|B) = 0.9$ and $\mathbb{P}(T|B^c) = 0.1$. We want to calculate $\mathbb{P}(B|T)$ (the probability of breast cancer, given a positive test). We could use Bayes' formula, but instead let's basically derive it from the definitions.

$$\mathbb{P}(B|T) = \frac{\mathbb{P}(B \cap T)}{\mathbb{P}(T)} = \frac{\mathbb{P}(B \cap T)}{\mathbb{P}(B \cap T) + \mathbb{P}(B^c \cap T)}.$$

Using the multiplication rule, the numerator is $\mathbb{P}(B \cap T) = \mathbb{P}(T|B)\mathbb{P}(B) = (0.9)(0.01) = 0.009$. In the denominator, the first term we just calculated, and the second is $\mathbb{P}(B^c \cap T) = \mathbb{P}(T|B^c)\mathbb{P}(B^c) = (0.1)(1 - 0.01) = 0.099$. Thus

$$\mathbb{P}(B|T) = \frac{0.009}{0.009 + 0.099} \approx 8.33\%.$$

Example 13.8. *In an election between two candidates Anderson and Bradley, exit polls indicated that candidate Anderson was winning 60% to 40%. But when all the votes were counted, candidate Bradley won by 55% to 45%. How could this have happened?*

Suppose that, among those people who voted for Anderson, only the fraction p stopped to answer the exit poll, while for those who voted for Bradley, the fraction q stopped to answer the poll. There is no reason to suppose $p = q$. So let's let A denote the event "voted for Anderson"; assuming no other candidates (and no spoiled ballots) then A^c is the event "voted for Bradley". Let P be the event "stopped to answer the exit pollster"; so $\mathbb{P}(P|A) = p$ while $\mathbb{P}(P|A^c) = q$. Then tallying up the exit polls means calculating $\mathbb{P}(A|P)$ and $\mathbb{P}(A^c|P)$ (out of the sample of those who answered the pollster, what fraction voted for Anderson vs. Bradley). Of coursr $\mathbb{P}(A|P) = 1 - \mathbb{P}(A^c|P)$ since $\mathbb{P}(\,\cdot\,|P)$ is a probability. Approaching this using Baye's theorem, we have

$$\mathbb{P}(A|P) = \frac{\mathbb{P}(A \cap P)}{\mathbb{P}(P)} = \frac{\mathbb{P}(A \cap P)}{\mathbb{P}(A \cap P) + \mathbb{P}(A^c \cap P)}.$$

Using the multiplication formula, the numerator is $\mathbb{P}(A \cap P) = \mathbb{P}(P|A)\mathbb{P}(A)$ and the denominator is $\mathbb{P}(A \cap P) + \mathbb{P}(A^c \cap P) = \mathbb{P}(P|A)\mathbb{P}(A) + \mathbb{P}(P|A^c)\mathbb{P}(A^c)$; so

$$\mathbb{P}(A|P) = \frac{\mathbb{P}(P|A)\mathbb{P}(A)}{\mathbb{P}(P|A)\mathbb{P}(A) + \mathbb{P}(P|A^c)\mathbb{P}(A^c)}.$$

The final election results say that $\mathbb{P}(A) = 0.45$ while $\mathbb{P}(A^c) = 0.55$. On the other hand, the exit polls indicated that $\mathbb{P}(A|P) = 0.6$. Thus, with $\mathbb{P}(P|A) = p$ and $\mathbb{P}(P|A^c) = q$, we want to find p, q such that

$$0.6 = \frac{0.45p}{0.45p + 0.55q}.$$

There are many solutions to this; we can only solve for a ratio of p to q; indeed, it says

$$(0.6)(0.45)p + (0.6)(0.55)q = (0.45)p$$

which implies

$$\frac{p}{q} = \frac{(0.6)(0.55)}{(0.4)(0.45)} = \frac{11}{6}.$$

One solution is $q = 0.3$ and $p = 0.55$. In other words, if only 30% of those who voted for Bradley answered the exit poll, while 55% of those who voted for Anderson answered the poll, the surprising election results would have ensued.

Another reason not to memorize Bayes' theorem is that the technique applies more widely than the formula (as stated).

Example 13.9. *A tech company purchases chip-sets from three factories, F_1, F_2, F_3; they get 20% from F_1, 30% from F_2, and 50% from F_3. The three factories have defect rates of 4%, 3%, and 2% respectively. If a defective chip is found, what is the probability it came from factory F_2?*

For any chip, the probability it came from factory F_i is $\mathbb{P}(F_1) = 0.2$, $\mathbb{P}(F_2) = 0.3$, $\mathbb{P}(F_3) = 0.5$. Let D denote the event that a chip is defective. The probability a chip is defective, given it comes from factory F_i, is $\mathbb{P}(D|F_1) = 0.04$, $\mathbb{P}(D|F_2) = 0.03$, $\mathbb{P}(D|F_3) = 0.02$. We are interested in the probability that the chip came from factory F_2 given that it's defective: $\mathbb{P}(F_2|D)$. We proceed with Bayes' approach, using the fact that F_1, F_2, F_3 form a partition (each chip-sets comes from one of the three factories).

$$\mathbb{P}(F_2|D) = \frac{\mathbb{P}(F_2 \cap D)}{\mathbb{P}(D)} = \frac{\mathbb{P}(F_2 \cap D)}{\mathbb{P}(F_1 \cap D) + \mathbb{P}(F_2 \cap D) + \mathbb{P}(F_3 \cap D)}.$$

By the multiplication rule, $\mathbb{P}(F_i \cap D) = \mathbb{P}(D|F_i)\mathbb{P}(F_i)$. We can calculated these from the data given:

$$\mathbb{P}(F_1 \cap D) = \mathbb{P}(D|F_1)\mathbb{P}(F_1) = (0.04)(0.2) = 0.008$$
$$\mathbb{P}(F_2 \cap D) = \mathbb{P}(D|F_2)\mathbb{P}(F_2) = (0.03)(0.3) = 0.009$$
$$\mathbb{P}(F_3 \cap D) = \mathbb{P}(D|F_3)\mathbb{P}(F_3) = (0.02)(0.5) = 0.010$$

The sum of these three (the denominator in the calculation of $\mathbb{P}(F_i|D)$) is $0.008 + 0.009 + 0.010 = 0.027$, so

$$\mathbb{P}(F_2|D) = \frac{0.009}{0.027} = \frac{1}{3}.$$

We can similarly calculate that $\mathbb{P}(F_1|D) = \frac{8}{27}$ while $\mathbb{P}(F_3|D) = \frac{10}{27}$. Notice that the *conditional* defect rates are in the reverse order from the *unconditioned* defect rates of the three factories (due to the unequal proportions of chip-sets purchased from each).

Example 13.9 demonstrates a slightly more general form of Bayes' formula which we can state as follows. Let $A, B_1, B_2, \ldots, B_n$ be events, where $B_1, \ldots, B_n$ form a partition of the sample space. Then for $j \in \{1, \ldots, n\}$,

$$\mathbb{P}(B_j|A) = \frac{\mathbb{P}(B_j)\mathbb{P}(A|B_j)}{\sum_{i=1}^{n} \mathbb{P}(B_i)\mathbb{P}(A|B_i)}.$$

Again, I stress that you should not memorize this (since it will be difficult to remember which symbol matches which event in a given example); rather, you should work from the basics as we did in Examples 13.7–13.9 using the definition of conditional probability, partitioning the sample space, and using the multiplication formula, to relate $\mathbb{P}(B|A)$ with $\mathbb{P}(A|B)$.

14. LECTURE 14: OCTOBER 27, 2010

14.1. **Joint Distributions.** Recall that, if $X\colon \Omega \to S$ is a random variable with state space S, the distribution μ_X of X is a probability measure *on* S, given by

$$\mu_X(x) = \mathbb{P}(X = x) = \mathbb{P}\left(\{\omega \in \Omega \,:\, X(\omega) = x\}\right), \quad x \in S.$$

Frequently we will measure more than one random variable simultaneously; they may not be independent. In general, if $X\colon \Omega \to S$ and $Y\colon \Omega \to T$ are random variables (on the same sample space) with state spaces S, T, their **joint distribution** is the probability measure $\mu_{X,Y}$ *on* $S \times T$ given by

$$\mu_{X,Y}(x,y) = \mathbb{P}(X = x, Y = y) = \mathbb{P}\left(\{\omega \in \Omega \,:\, X(\omega) = x\} \cap \{\omega \in \Omega \,:\, Y(\omega) = y\}\right),$$

for $x \in S$ and $y \in T$.

Example 14.1. In a (mythical) community, 15% of families have no children, 20% have one child, 35% have two children, and 30% have three children. Let B be the number of boys, and G the number of girls each family has. (We suppose that each child born has probability $\frac{1}{2}$ of being a boy, and that genders of successive births are independent.) Then we can calculate the distribution $\mu_{B,G}$ as follows. Here are a few examples.

- $\mathbb{P}(B = 0, G = 0) = \mathbb{P}(\text{no children}) = 0.15$.
- $\mathbb{P}(B = 0, G = 1) = \mathbb{P}(1 \text{ chid, and } G = 1) = \mathbb{P}(1 \text{ chid})\mathbb{P}(G = 1 | 1 \text{ chid}) = (0.20)\frac{1}{2} = 0.10$.
- $\mathbb{P}(B = 1, G = 2) = \mathbb{P}(3 \text{ children, and } B = 1) = \mathbb{P}(3 \text{ children})\mathbb{P}(B = 1 | 3 \text{ children}) = (0.30)\binom{3}{1}\frac{1}{2}(\frac{1}{2})^2 = 0.1125$.

Proceeding in this way, we can build up a table of all the values $\mu_{B,G}(x,y)$ where x, y range through $\{0,1,2,3\}$.

$\mu_{B,G}(i,j)$	$i=0$	$i=1$	$i=2$	$i=3$	row sum
$j=0$	0.15	0.10	0.0875	0.0375	0.3750
$j=1$	0.10	0.175	0.1125	0	0.3875
$j=2$	0.0875	0.1125	0	0	0.2000
$j=3$	0.0375	0	0	0	0.0375
col sum	0.3750	0.3875	0.2000	0.0375	1.0000

The lower-right portion of the table has all 0s, since the sum $B + G$ is always ≤ 3.

Example 14.2. Suppose X, Y are independent (discrete) random variables. Then, by definition,

$$\mu_{X,Y}(x,y) = \mathbb{P}(X = x, Y = y) = \mathbb{P}(X = x)\mathbb{P}(Y = y) = \mu_X(x)\mu_Y(y). \tag{14.1}$$

This is another way to state independence: *X, Y are independent if their joint distribution $\mu_{X,Y}$ is the product of their individual distributions μ_X, μ_Y*, as in Equation 14.1.

In Example 14.1, we listed the column and row sums. Why? These numbers allow us to recover the distributions of the two random variables separately.

Proposition 14.3. *If X, Y are two discrete random variables, then for any x_0 in the state space of X and any y_0 in the state space of Y,*

$$\mu_X(x_0) = \mathbb{P}(X = x_0) = \sum_y \mathbb{P}(X = x_0, Y = y) = \sum_y \mu_{X,Y}(x_0, y) \quad \textit{(col sum)}$$

$$\mu_Y(y_0) = \mathbb{P}(Y = y_0) = \sum_x \mathbb{P}(X = x, Y = y_0) = \sum_x \mu_{X,Y}(x, y_0) \quad \textit{(row sum)}$$

Proof. This is just the law of total probability. The events $\{Y = y\}$, as y ranges through all possible values y in the state space of Y, form a partition of Ω. Hence

$$\mathbb{P}(X = x_0) = \sum_y \mathbb{P}(\{X = x_0\} \cap \{Y = y\}).$$

The second equation is proved similarly. □

Thus, the sum of the column sums is equal to $\sum_x \mathbb{P}(X = x)$ which must equal 1.

In general, if we are presented with a table of number $\mu(x, y)$ and are told it is the joint distribution of a pair of random variables X, Y, we can recover the distributions μ_X and μ_Y by taking the row and column sums:

$$\mathbb{P}(X = x) = \sum_y \mu(x, y), \quad \mathbb{P}(Y = y) = \sum_x \mu(x, y).$$

In the table in Example 14.1, we wrote the row and column sums in the margins of the table. This is common, and for this reason, the distributions you get by summing the rows and columns of a two variable distribution are called the **marginal distributions**. So Proposition 14.3 can be restated as follows: *the marginals of $\mu_{X,Y}$ are the distributions μ_X and μ_Y.*

We can do this for any number of random variables; the generalization is straightforward.

Example 14.4. Let X_1, X_2, X_3 be the outcomes of tossing three fair coins. The joint distribution μ_{X_1,X_2,X_3} is the probability measure on $\{0, 1\}^3$ given (due to fairness) by

$$\mu_{X_1,X_2,X_3}(x_1, x_2, x_3) = \mathbb{P}(X_1 = x_1, X_2 = x_2, X_3 = x_3) = \frac{1}{8}$$

for each triple $(x_1, x_2, x_3) \in \{0, 1\}^3$. In this case, the marginals require us to sum over the *other two* variables:

$$\sum_{x_2,x_3} \mu_{X_1,X_2,X_3}(x, x_2, x_3) = \frac{1}{8} + \frac{1}{8} + \frac{1}{8} + \frac{1}{8} = \frac{1}{2},$$

and indeed this is equal to $\mathbb{P}(X_1 = x)$ for either choice $x \in \{0, 1\}$.

Example 14.5. Let B_1, B_2 be a partition of the sample space Ω. Let N be a $\mathrm{Poisson}(\lambda)$ random variable. Suppose we perform an experiment N times (i.e. a *random* number of times). Let X_i be the number of times that B_i occurs among the random number of trials. (So there is randomness in the trials, and in the number we perform.)

First, we can calculate the probability that the number of trials we perform is exactly n: $\mathbb{P}(N = n) = e^{-\lambda}\frac{\lambda^n}{n!}$. Now, suppose we want to know the probability that B_1 occurs n_1

times, while B_2 occurs n_2 times, for fixed n_1, n_2. In this case, we know that the number of trials performed was $n = n_1 + n_2$. We can then calculate as follows:

$$\begin{aligned}\mathbb{P}(X_1 = n_1, X_2 = n_2) &= \mathbb{P}(N = n, X_1 = n_1, X_2 = n_2) \\ &= \mathbb{P}(N = n)\mathbb{P}(X_1 = n_1, X_2 = n_2 | N = n) \\ &= e^{-\lambda}\frac{\lambda^n}{n!}\mathbb{P}(X_1 = n_1, X_2 = n_2 | N = n).\end{aligned}$$

Now, given that the number of trials is the fixed number n, the number of times B_1 occurs is binomial: we repeat an experiment n times, and look for the number of "success" outcomes B_1 (which has probability $p = \mathbb{P}(B_1)$). Thus

$$\begin{aligned}\mathbb{P}(X_1 = n_1, X_2 = n_2 | N = n) &= \binom{n}{n_1} p^{n_1}(1-p)^{n_2} \\ &= \frac{n!}{n_1! n_2!}\mathbb{P}(B_1)^{n_1}\mathbb{P}(B_2)^{n_2}.\end{aligned}$$

Thus,

$$\mathbb{P}(X_1 = n_1, X_2 = n_2) = e^{-\lambda}\frac{\lambda^n}{n!} \cdot \frac{n!}{n_1! n_2!}\mathbb{P}(B_1)^{n_1}\mathbb{P}(B_2)^{n_2}.$$

We cancel the $n!$s, and sneakily rewrite the initial terms as

$$e^{-\lambda}\lambda^n = e^{-\lambda\mathbb{P}(B_1)}e^{-\lambda\mathbb{P}(B_2)}\lambda^{n_1}\lambda^{n_2}.$$

Finally, this gives

$$\mu_{X_1,X_2}(n_1, n_2) = \mathbb{P}(X_1 = n_1, X_2 = n_2) = e^{-\lambda\mathbb{P}(B_1)}\frac{\lambda^{n_1}}{n_1!}\mathbb{P}(B_1)^{n_1} \cdot e^{-\lambda\mathbb{P}(B_2)}\frac{\lambda^{n_2}}{n_2!}\mathbb{P}(B_2)^{n_2}.$$

So we have calculated the joint distribution of X_1, X_2. This allows us to calculate the distributions of X_1, X_2 by taking the marginals.

$$\begin{aligned}\mu_{X_1}(n_1) &= \sum_{n_2=0}^{\infty} e^{-\lambda\mathbb{P}(B_1)}\frac{(\lambda\mathbb{P}(B_1))^{n_1}}{n_1!} \cdot e^{-\lambda\mathbb{P}(B_2)}\frac{(\lambda\mathbb{P}(B_2))^{n_2}}{n_2!} \\ &= e^{-\lambda\mathbb{P}(B_1)}\frac{(\lambda\mathbb{P}(B_1))^{n_1}}{n_1!}\sum_{n_2=0}^{\infty} e^{-\lambda\mathbb{P}(B_2)}\frac{(\lambda\mathbb{P}(B_2))^{n_2}}{n_2!} \\ &= e^{-\lambda\mathbb{P}(B_1)}\frac{(\lambda\mathbb{P}(B_1))^{n_1}}{n_1!}.\end{aligned}$$

That is, X_1 is $\mathrm{Poisson}(\lambda\mathbb{P}(B_1))$. Similarly, we can calculate that X_2 is $\mathrm{Poisson}(\lambda\mathbb{P}(B_2))$. And, better yet: the joint distribution μ_{X_1,X_2} is the product of the marginals. In other words, X_1, X_2 **are independent**.

This last point is remarkable. If n is fixed, then $X_1 + X_2 = n$, which means that they are maximally dependent: one determines the other completely! But if we perform the experiment a random (Poisson) number of times, they become statistically independent! To see how this can be useful, consider this concrete example. Suppose the number of cars that come through a fast-food drive-through in an hour is $\mathrm{Poisson}$. Then the numbers of cars with male drivers, and the number with female drivers, are each independent $\mathrm{Poisson}$s as well.

In Example 14.4, we saw that μ_{X_1,X_2} was a product of two distributions. We then calculated that these two distributions were, in fact, its marginals, and so concluded that X_1, X_2 were independent. Actually, we didn't need to go through this last step.

Proposition 14.6. *Let $\mu\colon S \times T \to \mathbb{R}$ be a probability distribution, and let μ_1, μ_2 denote its marginals. Suppose there are functions $f\colon S \to \mathbb{R}$ and $g\colon T \to \mathbb{R}$ such that $\mu(x,y) = f(x)g(y)$. Then there is a constant c so that $f(x) = c\mu_1(x)$ and $g(y) = \frac{1}{c}\mu_2(y)$.*

Proof. We simply take the row and column sums.

$$\sum_y \mu(x,y) = \sum_y f(x)g(y) = f(x)\sum_y g(y), \quad \sum_x \mu(x,y) = \sum_x f(x)g(y) = g(y)\sum_x f(x).$$

Let $c_1 = \sum_y g(y)$ and $c_2 = \sum_x f(x)$. Then, using the definitions of the marginals,

$$\mu_1(x) = \sum_y \mu(x,y) = c_1 f(x), \quad \mu_2(y) = \sum_x \mu(x,y) = c_2 g(y).$$

Finally, since μ_1 and μ_2 are probability distributions, we have

$$1 = \sum_x \mu_1(x) = \sum_x c_1 f(x) = c_1 c_2, \quad 1 = \sum_y \mu_2(y) = \sum_y c_2 g(y) = c_2 c_1.$$

Thus $c_1 = 1/c_2$. So taking $c = c_2$ we have $c_1 = \frac{1}{c}$, and so $f(x) = c\mu_1(x)$ and $g(x) = \frac{1}{c}\mu_2(x)$ as claimed. Also, note

$$\mu(x,y) = f(x)g(y) = \frac{f(x)g(y)}{c_1 c_2} = \mu_1(x)\mu_2(y),$$

and as Example 14.2 shows, this means that the two random variables are independent. □

The upshot of Proposition 14.6 is that, to prove to random variable are independent, we only need to see that their joint distribution factors as some kind of product; from that alone, we get independence (and can pick out the marginal distributions without calculation).

15. LECTURE 15: OCTOBER 29, 2010

15.1. **Computing from Joint Distributions.** The joint distribution is the trump-card when it comes to computing anything about a collection of random variables.

Theorem 15.1. *Let $X_1, \dots, X_n$ be a collection of discrete random variables with state spaces $S_1, \dots, S_n$. Let $\mu_{X_1,\dots,X_n}$, the joint distribution, be known. If $f\colon S_1 \times \cdots \times S_n \to \mathbb{R}$ is* any *function, then the distribution of $f(X_1, \dots, X_n)$ can be calculated as*

$$\begin{aligned}\mathbb{P}(f(X_1,\dots,X_n)=k) &= \sum_{\substack{(x_1,\dots,x_n)\\ f(x_1,\dots,x_n)=k}} \mathbb{P}(X_1=x_1,\dots,X_n=x_n)\\ &= \sum_{\substack{(x_1,\dots,x_n)\\ f(x_1,\dots,x_n)=k}} \mu_{X_1,\dots,X_n}(x_1,\dots,x_n).\end{aligned}$$

Example 15.2. Let $f(x,y)=x+y$. Theorem 15.1, in this case, says

$$\mathbb{P}(X+Y=k)=\sum_{\substack{(x,y)\\ x+y=k}}\mathbb{P}(X=x,Y=y)=\sum_x \mathbb{P}(X=x,Y=k-x)=\sum_x \mu_{X,Y}(x,k-x).$$

For example, consider the random variables B, G from Example 14.1. The table in that example gives the joint distribution $\mu_{B,G}$. So we can compute, for example,

$$\begin{aligned}\mathbb{P}(B+G=2)=\sum_{i=0}^{3}\mu_{B,G}(i,2-i) &= \mu_{B,G}(0,2)+\mu_{B,G}(1,1)+\mu_{B,G}(2,0)+\mu_{B,G}(3,-1)\\ &= 0.0875+0.175+0.0875=0.35.\end{aligned}$$

Of course, $B+G$ is the total number of children, and we started that example with the knowledge that a couple has 2 children 35% of the time; so this is consistent.

Example 15.3. Suppose X, Y are independent. Then we know $\mu_{X,Y}(x,y)=\mu_X(x)\mu_Y(y)$. As calculated in Example 15.2, $\mathbb{P}(X+Y=k)=\sum_x \mu_{X,Y}(x,k-x)$; in this case of independence, that says $\mathbb{P}(X+Y=k)=\sum_x \mu_X(x)\mu_Y(k-x)$ – a fact we proved already in Theorem 7.6. But we can write down rules for other functions of the two variables. For example,

$$\mathbb{P}(XY=k)=\sum_{\substack{(x,y)\\ xy=k}}\mu_{X,Y}(x,y)=\sum_x \mu_{X,Y}(x,k/x)=\sum_x \mu_X(x)\mu_Y(k/x).$$

Let's consider a specific example. Roll two fair dice, and let X, Y be the values that come up. Then

$$\mathbb{P}(XY=6)=\sum_{i=1}^{6}\mu_X(i)\mu_Y(6/i).$$

Now, $6/1=6$, $6/2=3$, $6/3=2$, and $6/6=1$ are all in the state space of Y; on the other hand, $6/4$ and $6/5$ are not, and so in the sum those terms give probability 0. Hence

$$\mathbb{P}(XY=6)=\mu_X(1)\mu_Y(6)+\mu_X(2)\mu_Y(3)+\mu_X(3)\mu_Y(2)+\mu_X(6)\mu_Y(1)=4\cdot\frac{1}{6}\cdot\frac{1}{6}=\frac{1}{9}.$$

On the other hand, the only way to make 25 as a product of number in $\{1,2,3,4,5,6\}$ is $25=5^2$, so $\mathbb{P}(XY=25)=\mu_X(5)\mu_Y(5)=(\frac{1}{6})^2=\frac{1}{36}$.

15.2. **Conditional Distributions.** In Examples 14.1 and 14.4, we calculated joint distributions using conditioning. This leads us to the notion of a **conditional distribution**:

$$\mathbb{P}(X = x|Y = y) = \frac{\mathbb{P}(X = x, Y = y)}{\mathbb{P}(Y = y)} = \frac{\mu_{X,Y}(x, y)}{\mu_Y(y)}.$$

For fixed y, this is a probability on the state space S of X: $\mathbb{P}(X = \cdot|Y = y) = \frac{\mu_{X,Y}(\cdot,y)}{\mu_Y(y)}$. It is important not to confuse it with the joint distribution $\mathbb{P}(X = x, Y = y)$, which does not sum to 1 over x (only over both x, y).

Example 15.4. Let n be fixed, and suppose we perform and experiment n times. There are three outcomes we're looking for, B_1, B_2, B_3, which partition the sample space. Set $p_i = \mathbb{P}(B_i)$. If $X_i =$ the number of times B_i occurs, then the joint distribution of X_1, X_2, X_3 is multinomial:

$$\mathbb{P}(X_1 = n_1, X_2 = n_2, X_3 = n_3) = \binom{n}{n_1\ n_2\ n_3} p_1^{n_1} p_2^{n_2} p_3^{n_3}.$$

Let's compute the conditional distribution of X_1, X_2 given that $X_3 = k$ (for some fixed $k \leq n$).

$$\mathbb{P}(X_1 = n_1, X_2 = n_2|X_3 = k) = \frac{\mathbb{P}(X_1 = n_1, X_2 = n_2, X_3 = k)}{\mathbb{P}(X_3 = k)}.$$

The denominator is the distribution of X_3 evaluated at k. Viewing $B_1 \cup B_2$ as the complement of B_3, we see X_3 has a binomial distribution:

$$\mathbb{P}(X_3 = k) = \binom{n}{k} p_3^k (1 - p_3)^{n-k}.$$

Thus,

$$\mathbb{P}(X_1 = n_1, X_2 = n_2|X_3 = k) = \frac{\binom{n}{n_1\ n_2\ k} p_1^{n_1} p_2^{n_2} p_3^k}{\binom{n}{k} p_3^k (1 - p_3)^{n-k}}.$$

We may simplify

$$\frac{\binom{n}{n_1\ n_2\ n_3}}{\binom{n}{k}} = \frac{\frac{n!}{n_1!n_2!k!}}{\frac{n!}{k!(n-k)!}} = \frac{(n-k)!}{n_1!n_2!}$$

Hence, canceling the p_3^k terms, and noting that $1 - p_3 = p_1 + p_2$, we have

$$\mathbb{P}(X_1 = n_1, X_2 = n_2|X_3 = k) = \frac{(n-k)!}{n_1!n_2!} \frac{p_1^{n_1} p_2^{n_2}}{(p_1 + p_2)^{n-k}}.$$

Note that $n_1 + n_2 = n - k$, and so we can write this finally as

$$\mathbb{P}(X_1 = n_1, X_2 = n_2|X_3 = k) = \binom{n_1 + n_2}{n_1} \left(\frac{p_1}{p_1 + p_2}\right)^{n_1} \left(\frac{p_2}{p_1 + p_2}\right)^{n_1}.$$

Observe that $\frac{p_2}{p_1+p_2} = 1 - \frac{p_1}{p_1+p_2}$. Hence, we see that the conditional distribution is binomial $\text{binomial}(n_1 + n_2, \frac{p_1}{p_1+p_2})$ with outcomes n_1. This is what we should have expected: once we know that k of the outcomes are B_3, there are $n - k = n_1 + n_2$ remaining trails. Since $\mathbb{P}(B_1) = p_1$ and $\mathbb{P}(B_2) = p_2$, the fraction of them we expect to be B_1 is $\frac{p_1}{p_1+p_2}$, and so we expect a binomial with $n_1 + n_2$ trials, with probability of success $\frac{p_1}{p_1+p_2}$, given n_1 trials – i.e. the answer we calculated.

15.3. **Cumulative Distribution Functions.** So far we have considered random variables $X\colon \Omega \to S$ with state spaces $S \subseteq \mathbb{N}$ (i.e. *discrete* random variables). For such random variables, $\mathbb{P}(X = x)$ is non-zero for any $x \in S$. We want to move to more general kinds of random variables, where it might even happen that $\mathbb{P}(X = x) = 0$ for any $x \in S$ (like a uniformly random number in the interval $[0, 1]$). To avoid this problem, we can always view $X\colon \Omega \to \mathbb{R}$, and then think about the numbers

$$F_X(x) = \mathbb{P}(X \leq x) \tag{15.1}$$

instead of the numbers $\mu_X(x) = \mathbb{P}(X = x)$. The function in Equation 15.1 is called the **cumulative distribution function** of X. Let's see what it looks like in the discrete setting we've been discussing.

Example 15.5. Let $X\colon \Omega \to S \subset \mathbb{R}$ be a discrete random variable; for the sake of argument, let's take $S = \mathbb{N}$. Then for any x, the event $\{X \leq x\}$ can be decomposed as

$$\{X \leq x\} = \bigcup_{\substack{n \in \mathbb{N} \\ n \leq x}} \{X = n\},$$

and the union is disjoint. Hence,

$$F_X(x) = \mathbb{P}(X \leq x) = \sum_{\substack{n \in \mathbb{N} \\ n \leq x}} \mathbb{P}(X = n) = \sum_{\substack{n \in \mathbb{N} \\ n \leq x}} \mu_X(n).$$

So, if x is not a positive integer, then there is an integer m with $m < x < m + 1$, and so $\{n \in \mathbb{N} : n \leq x\} = \{n \in \mathbb{N} : n \leq m\}$. In other words, for ALL $x \in (n, n + 1)$ we have $F_X(x) = F_X(m)$. So F_X is a **step-function**.

Now, for any positive integer m,

$$F_X(m) - F_X(m - 1) = \sum_{n=0}^{m} \mu_X(n) - \sum_{n=0}^{m} \mu_X(n) = \mu_X(m).$$

So the size of the mth jump is exactly $\mu_X(m)$. Thus, the function F_X encodes exactly the same information as μ_X for discrete random variable.

For a concrete example, let X be the sum of 2 fair dice (where 0 means tails and 1 means heads). Then we know that

$$\mu_X(0) = \frac{1}{4}, \mu_X(1) = \frac{1}{2}, \mu_X(2) = \frac{1}{4}.$$

Then, we can calculate the cumulative distribution function of X:

$$F_X(x) = \begin{cases} 0, & x < 0 \\ \frac{1}{4}, & 0 \leq x < 1 \\ \frac{3}{4}, & 1 \leq x < 2 \\ 1, & x \geq 2 \end{cases}.$$

16. LECTURE 16: NOVEMBER 1, 2010

16.1. **Cumulative Distribution Functions.** Example 15.5 shows that if X is a discrete random variable, then its cumulative distribution function F_X is a step function, with steps at the states of X of heights equal to the probabilities of the states. It follows then that $F_X(x)$ is 0 for x sufficiently small, and $F_X(x) = 1$ for x sufficiently large. Also, F_X is a *monotone increasing* function. Of course, F_X is not continuous. But the value of F_X at the jumps is equal to the new, higher value. One way of saying this is that F_X is *right-continuous*:

$$\lim_{x \to x_0+} F_X(x) = F_X(x_0).$$

These properties actually follow directly from the axioms of probability. As the following theorem shows, it is here (finally) that Axiom 3 (the countable additivity of $\mathbb{P}$, a.k.a. the continuity of $\mathbb{P}$) comes into play.

Theorem 16.1. *Let $(\Omega, \mathbb{P})$ be a probability space, and let $X : \Omega \to \mathbb{R}$ be any random variable (not necessarily discrete). Then the function $F_X(x) = \mathbb{P}(X \leq x)$ has the following properties:*

(a) *F_X is monotone increasing: if $x < y$ then $F_X(x) \leq F_X(y)$.*

(b) $\lim_{x \to -\infty} F_X(x) = 0$, $\lim_{x \to +\infty} F_X(x) = 1$.

(c) *F_X is right-continuous: for any x_0,* $\lim_{x \to x_0+} F_X(x) = F_X(x_0)$.

Proof. (a) By definition, $F_X(x) = \mathbb{P}(X \leq x) = \mathbb{P}(\{\omega \in \Omega : X(\omega) \leq x\})$. If $x < y$, then $X(\omega) \leq x$ implies that $X(\omega) \leq y$; in other words, $\{X \leq x\} \subseteq \{X \leq y\}$. Hence, $\mathbb{P}(X \leq x) \leq P(X \leq y)$.

(b) We'll just look at $x \in \mathbb{N}$ here. $F_X(n) = \mathbb{P}(X \leq n)$. The events $A_n = \{X \leq n\}$ satisfy $A_n \leq A_{n+1}$. Note that, since X takes only finite values, $\bigcup_n A_n = \Omega$, meaning $A_n \uparrow \Omega$ as $n \to \infty$ and so by the continuity axiom of probability, $\lim_{n\to\infty} F_X(n) = \lim_{n\to\infty} \mathbb{P}(A_n) = \mathbb{P}(\Omega) = 1$. On the other hand, $A_n \downarrow \varnothing$ as $n \to -\infty$ (since any value of X is bigger than *some* small negative number). Hence, by the continuity axiom of probability, $\lim_{n\to-\infty} F_X(n) = \lim_{n\to-\infty} \mathbb{P}(A_n) = \mathbb{P}(\varnothing) = 0$.

(c) Take $A_n = \{X \leq x_0 + \frac{1}{n}\}$. Then $A_n \downarrow \{X \leq x_0\}$ as $n \to \infty$, and so

$$\begin{aligned} \lim_{x \to x_0+} F_X(x) = \lim_{x \to x_0+} \mathbb{P}(X \leq x) &= \lim_{n\to\infty} \mathbb{P}(X \leq x_0 + \tfrac{1}{n}) \\ &= \lim_{n\to\infty} \mathbb{P}(A_n) = \mathbb{P}(X \leq x_0) = F_X(x_0). \end{aligned}$$

□

If F is *any* function $\mathbb{R} \to [0, 1]$ satisfying properties (a),(b),(c) in Theorem 16.1, we call F a **cumulative distribution function**. In general, any such function is the cumulative distribution function of some random variable. Example 15.5 shows that, when X is discrete, F_X is a step function. Here's an example that is not.

Example 16.2. Consider the function F given by

$$F(x) = \begin{cases} 0, & x < 0 \\ x, & 0 \leq x < 1 \\ 1, & x \geq 1 \end{cases}.$$

A random variable X with this cumulative distribution function $F_X = F$ is called **uniform** on $[0, 1]$. We encountered this distribution in Example 2.3, where our intuition told us that we should find $\mathbb{P}(X = x_0) = 0$ for any x_0. We can see this right from the form of F. The distribution of X is given by $\mathbb{P}(X \leq x) = F_X(x)$. Let's look at the probability that X lies in a certain interval $(a, b]$. Note that

$$(a, b] = (-\infty, b] \cap (a, \infty) = (-\infty, b] \cap (-\infty, a]^c.$$

Thus,

$$\{X \in (a, b]\} = \{X \leq b\} \cap \{X \leq a\}^c,$$

and so

$$\mathbb{P}(X \in (a, b]) = \mathbb{P}(X \leq b) - \mathbb{P}(X \leq a) = F_X(b) - F_X(a).$$

So, for the distribution in this example, if $a, b \in [0, 1]$ then we have $F_X(a) = a$ and $F_X(b) = b$, and so $\mathbb{P}(X \in (a, b]) = b - a$.

What about $\mathbb{P}(X \in [a, b])$? This is $\mathbb{P}(X = a) + \mathbb{P}(X \in (a, b]) = \mathbb{P}(X = a) + b - a$. So how do we evaluate $\mathbb{P}(X = a)$? Well, fix some small number $\epsilon > 0$, and look at

$$\mathbb{P}(X \in (a - \epsilon, a + \epsilon]) = (a + \epsilon) - (a - \epsilon) = 2\epsilon.$$

Notice that $\{a\} \subset (a - \epsilon, a + \epsilon]$ for any small $\epsilon > 0$, so $\{X = a\} \subseteq \{X \in (a - \epsilon, a + \epsilon]\}$. Thus, $\mathbb{P}(X = a) \leq \mathbb{P}(X \in (a - \epsilon, a + \epsilon]) = 2\epsilon$. This is true for **any** $\epsilon > 0$. For example, taking $\epsilon = \frac{1}{2} \times 10^{-10}$, this means $\mathbb{P}(X = a) < 0.0000000001$. Taking $\epsilon > 0$ smaller and smaller, we have $\mathbb{P}(X = a) = 0$ as we thought.

In Example 16.2, we saw two very important general facts. First, as we calculated (in general),

$$\mathbb{P}(X \in (a, b]) = F_X(b) - F_X(a). \tag{16.1}$$

The second fact, which we calculated implicitly, is

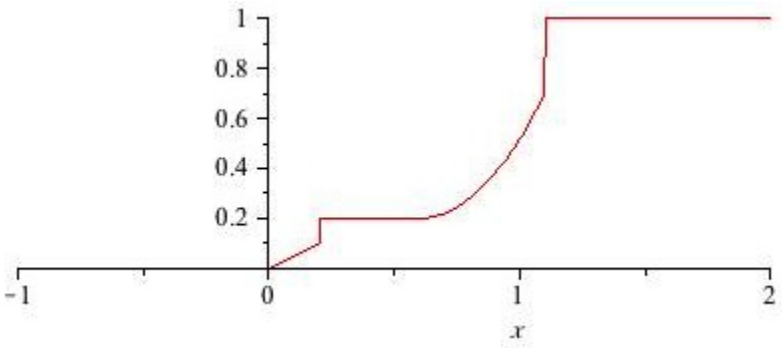
$$\mathbb{P}(X = a) = \lim_{\epsilon \downarrow 0} \left[F_X(a + \epsilon) - F_X(a - \epsilon) \right]. \tag{16.2}$$

Actually, since F is right-continuous at all points, we can simplify this to

$$\mathbb{P}(X = a) = F_X(a) - \lim_{\epsilon \downarrow 0} F_X(a - \epsilon).$$

From Equation 16.2, we see that $\mathbb{P}(X = a) = 0$ if F_X is *continuous at* a. In general, the only points a where $\mathbb{P}(X = a) > 0$ are the *jumps* in F_X, where $\mathbb{P}(X = a)$ is the height of the jump.

Example 16.3. Consider the following graph of the function



$$F(x) = \begin{cases} 0, & x < 0 \\ \frac{1}{2}x, & 0 \le x < 0.2 \\ 0.2, & 0.2 \le x < 0.6 \\ 0.2 + 2(x-0.6)^2, & 0.6 < x < 1.1 \\ 1, & x \ge 1.1 \end{cases}.$$

This function is monotone increasing, 0 when $x \to -\infty$, 1 when $x \to \infty$, and continuous except at the points 0.2 and 1.1 where it is right continuous. Hence, F is a cumulative distribution function. If X is a random variable with this cumulative distribution function $F_X = F$, then we have $\mathbb{P}(X = x) = 0$ unless $x \in \{0.2, 1.1\}$, where

$$\mathbb{P}(X = 0.2) = F(0.2) - \lim_{\epsilon \downarrow 0} F(0.2 - \epsilon) = 0.2 - \lim_{\epsilon \downarrow 0} \frac{1}{2}(0.2 - \epsilon) = 0.2 - 0.1 = 0.1$$

and

$$\begin{aligned} \mathbb{P}(X = 0.6) = F(1.1) - \lim_{\epsilon \downarrow 0} F(1.1 - \epsilon) &= 1 - \lim_{\epsilon \downarrow 0}(0.2 + 2(1.1 - \epsilon - 0.6)^2) \\ &= 1 - (0.2 + 2(0.5)^2) = 0.3. \end{aligned}$$

Although all other points have probability 0 of occurring, we can compute the probability that $X \in [a, b]$ for any interval. For example,

$$\begin{aligned} \mathbb{P}(X \in [0.1, 0.8]) = \mathbb{P}(X = 0.1) + \mathbb{P}(X \in (0.1, 0.8]) &= 0 + F(0.8) - F(0.1) \\ &= (0.2 + 2(0.8 - 0.6)^2) - \tfrac{1}{2}(0.2) \\ &= 0.22 - 0.1 = 0.12. \end{aligned}$$

16.2. **Probability Densities.** Now let's leave discreteness behind us, and forget about jumps altogether. If F is a cumulative distribution function without jumps, then it is a continuous, monotone increasing function (increasing from 0 to 1). To get the intuition right, let's assume F is even nicer than this: **we temporarily assume that F is C^1, continuously-differentiable**. In this case, F has a continuous derivative $f = F'$. Since F is monotone-increasing, $f \ge 0$. And we can recover F from f with the Fundamental Theorem of Calculus:

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

More generally, if $a < b$ then

$$F(b) - F(a) = \int_{a}^{b} f(x)\, dx. \tag{16.3}$$

The derivative f therefore allows us to recover the cumulative distribution function F. The only constraint on f (other than $f \ge 0$) relates to the fact that $\lim_{x \to \infty} F(x) = 1$. This means we must have

$$\int_{-\infty}^{\infty} f(x) = 1.$$

Such a function is called a **probability density**.

The Fundamental Theorem of Calculus goes two ways, actually: *if* there exists a continuous f that makes 16.3 true, then F is differentiable, and $F' = f$. But Equation 16.3 holds true for some discontinuous functions as well; we only need f to be *Riemann integrable*.

Since we are assuming F is continuous, if X is a random variable with $F_X = F$, then $\mathbb{P}(X = x) = 0$ for all x. Thus $\mathbb{P}(X \in [a,b]) = \mathbb{P}(X \in (a,b]) = \mathbb{P}(X \in [a,b)) = \mathbb{P}(X \in (a,b)) = F(b) - F(a)$.

Definition 16.4. *Let X be a real-valued random variable. If there exists a Riemann-integrable function $f_X \geq 0$ with total integral $\int_{-\infty}^{\infty} f_X(x)\,dx = 1$ such that*

$$F_X(b) = \int_{-\infty}^{b} f_X(x)\,dx, \quad b \in \mathbb{R}$$

then we call f_X the **probability density** *of X. If X has a density, we call it a* **continuous** *random variable. In this case, for any $a \leq b$ we have*

$$\mathbb{P}(X \in [a,b]) = \int_a^b f_X(x)\,dx. \tag{16.4}$$

Remark 16.5. We interpret "Riemann integrable" loosely in Definition 16.4. Technically, a Riemann integrable function is bounded; we will allow unbounded densities, provided the improper Riemann integral exists. Recall that if f is Riemann integrable on $[a,x$ for all $x < b$, the *improper integral* of f on $[a,b]$ is defined to be the limit

$$\int_a^b f(t)\,dt = \lim_{x \uparrow b} \int_a^x f(t)\,dt.$$

Similarly,

$$\int_{-\infty}^{\infty} f(t)\,dt = \lim_{r \to \infty} \int_{-r}^{r} f(t)\,dt.$$

Note that, if X is a continuous random variable (according to Definition 16.4), then it has a density f_X which is Riemann integrable. This means that the cumulative distribution function is continuous.

Proposition 16.6. *If f is (improperly) Riemann integrable on $(-\infty,\infty)$ and $F(x) = \int_{-\infty}^{x} f(t)\,dx$, then F is continuous.*

Proof. Let $x \in \mathbb{R}$. Then by definition

$$F(x) = \int_{-\infty}^{x} f(t)\,dx = \lim_{r \to \infty} \int_{-r}^{x} f(t)\,dt.$$

If x is a vertical asymptote of f, then the value of this integral is *defined* to be the limit

$$F(x) = \lim_{y \to x} \int_{-r}^{y} f(t)\,dy = \lim_{y \to x} F(y),$$

and so F is automatically continuous at x. If x is not a vertical asymptote, then (for $-r$ sufficiently close to x) we have f is bounded on $[-r,x]$. That is, there is a constant M such that $f(t) \leq M$ for $t \in [-r,x]$. Hence, for $y \in [-r,x]$,

$$F(x) - F(y) = \int_{-r}^{x} f(t)\,dt - \int_{-r}^{y} f(t)\,dt = \int_{y}^{x} f(t)\,dt \leq \int_{y}^{x} M\,dt = M(y-x),$$

and this tends to 0 as $y \to x$; hence, F is continuous at x. $\square$

This is the reason we call X a continuous random variable: if X is a continuous r.v. (i.e. if it has a density f_X), then F_X is a continuous function. But if the density f_X is not continuous, then the cumulative distribution F_X is not differentiable (though it is continuous).

So, if X is a discrete random variable, or more generally if F_X is not continuous at some point, then X cannot possibly have a density f_X. (If it did have one, then at the jump points in F_X the "function" f_X would have to have infinite spikes.) The density is a substitute for the distribution value at a point: you can think of $f_X(x)$ as containing the same sort of information that $\mu_X(x) = \mathbb{P}(X = x)$ does in the discrete case. To be a little more precise: if f_X is close to constant on a tiny interval $[x, x + \Delta x]$ containing x (for example if f_X is continuous), then

$$\mathbb{P}(x \leq X \leq x + \Delta x) \approx f_X(x)\Delta x.$$

As $\Delta x \to 0$, both sides tend to 0, which is why f_X is not a probability, it is a probability *density*.

Example 16.7. For $a < b$ in $\mathbb{R}$, consider the function

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

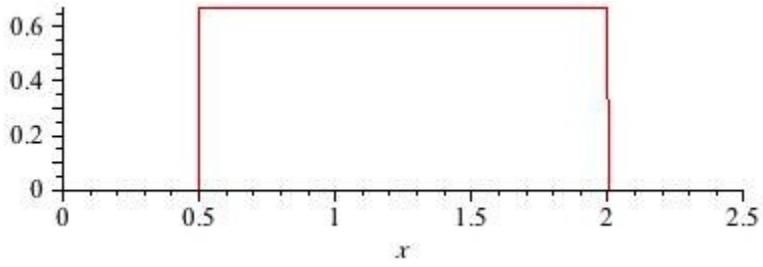


FIGURE 1. The probability density f.

This function is Riemann integrable, and non-negative, so it is a probability density. The antiderivative F of f is also a piecewise function,

$$F(x) = \int_{-\infty}^{x} f(t)\,dt = \begin{cases} 0, & x < a \\ \frac{x-a}{b-a}, & a \leq x \leq b \\ 1, & x > b \end{cases}.$$

This cumulative distribution function is continuous, but non-differentiable at the points a and b. If $a \leq c < d \leq b$, then the random variable X with density f satisfies

$$\mathbb{P}(X \in [c, d]) = F(d) - F(c) = \frac{d-c}{b-a}.$$

I.e. the probability that X is in $[c, d]$ is the ratio of the length of $[c, d]$ to the whole interval $[a, b]$. We call such a random variable **uniform** on $[a, b]$.

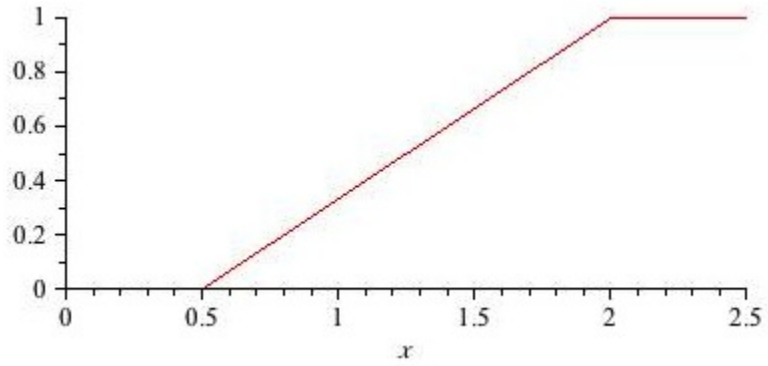


FIGURE 2. The cumulative distribution function F.

Remark 16.8. A word of caution. We showed above that *if* X is a continuous random variable (namely it has a probability density f_X) then its cumulative distribution function F_X is continuous. *The converse of this statement is not always true*. That is: there are continuous cumulative distribution functions F for which no density exists. These are pathological examples that are beyond the scope of this course. In analysis, a function F which has a density is called *absolutely continuous*. It would be more accurate to use the term *absolutely continuous random variable* for an X which possesses a density f_X; but the terminology is very standard, so we will stick with it.

17. LECTURE 17: NOVEMBER 3, 2010

Let's review our new view-point, and the new (and old) objects we're considering.

- *Random variable*: a function $X : \Omega \to \mathbb{R}$ where $(\Omega, \mathbb{P})$ is a probability space.
- *Distribution* μ_X: the probability *on* $\mathbb{R}$ defined as follows: for a subset U of $\mathbb{R}$,
$$\mathbb{P}(X \in U) = \mathbb{P}(\{\omega \in \Omega \,:\, X(\omega) \in U\}) \equiv \mu_X(U).$$
- *Discrete Distribution*: **If** the set of values S that X can assume (i.e. S is the state space) is finite (or infinite but discrete, like the natural numbers $\mathbb{N}$), then for any subset $\{s_1, s_2, \ldots, s_n\}$ of S,
$$\mu_X(\{s_1, \ldots, s_n\}) = \mu_X(s_1) + \cdots + \mu_X(s_n).$$
So to know μ_X, we only need to know the **probability mass function** (also denoted μ_X):
$$\mu_X : S \to [0,1], \quad \mu_X(s) = \mathbb{P}(X = s).$$
But there are lots of examples where $\mathbb{P}(X = s) = 0$ for all s in the state space; we need another tool to understand μ_X in this case.
- *Cumulative Distribution Function*: For any random variable X, we define a function $F_X : \mathbb{R} \to [0,1]$ by
$$F_X(s) = \mathbb{P}(X \leq s).$$
This function is non-decreasing, right continuous, and at $\pm\infty$ has the following limits: $\lim_{s\to-\infty} F_X(s) = 0$, and $\lim_{s\to+\infty} F_X(s) = 1$. If there is a value s where $\mathbb{P}(X = s) = h > 0$, then F_X has a jump-discontinuity at s, with height h. If $\mathbb{P}(X = s) = 0$ for all s, then F_X is continuous.
- *Probability Density*: **Sometimes**, the cumulative distribution function F_X of a random variable is continuous enough that it is an *antiderivative*. If there exists a function $f_X : \mathbb{R} \to \mathbb{R}$ such that
$$F_X(s) = \int_{-\infty}^{s} f_X(x)\, dx$$
that function f_X is called the **probability density** of X. If it exists, F_X *must* be continuous; thus, if there is any value s where $\mathbb{P}(X = s) > 0$ (e.g. the discrete case), then f_X does not exist. If a density exists, then the properties of its antiderivative F_X tell us that $f_X(x) \geq 0$ for all x, and $\int_{-\infty}^{\infty} f(x)\, dx = 1$.
- *Fundamental Theorem of Calculus*: **If** F_X is even nicer, namely C^1 (continuously differentiable, meaning that F_X' is a continuous function), then by the Fundamental Theorem of Calculus
$$\frac{d}{ds} F_X(s) = \frac{d}{ds} \int_{-\infty}^{s} f_X(x)\, dx = f_X(s).$$
I.e. the density is the derivative of the cumulative distribution function, **in the case that it is differentiable**. But there are lots of important examples where f_X exists (but is not continuous); then F_X has "sharp-corners" at those points.

Example 17.1 (Exponential Density)**.** Let $\lambda > 0$. The function $f(x) = \lambda e^{-\lambda x}$ for $x \geq 0$ and $f(x) = 0$ for $x < 0$ is a probability density: it is ≥ 0, Riemann integrable, and

$$\int_0^\infty \lambda e^{-\lambda x} = -\, e^{-\lambda x}\big|_{x=0}^{x=\infty} = -(0-1) = 1.$$

If X is a random variable with $f_X(x) = \lambda e^{-\lambda x}$, we call X an $\mathbf{exponential}(\lambda)$ random variable.

We can calculate F_X from f_X, by integrating.

$$F_X(x) = \int_{-\infty}^x f_X(t)\,dx = \int_0^x \lambda e^{-\lambda t}\,dt = -e^{-\lambda t}\big|_{t=-\infty}^{t=x} = -(e^{-\lambda x} - 1) = 1 - e^{-\lambda x}.$$

To demonstrate how useful these tools (F_X and f_X) can be, let's look at a "memory" property of these random variables. Since an $\mathrm{exponential}(\lambda)$ random variable X is continuous, $\mathbb{P}(X = x) = 0$ for all x. But we can look at the event that $X > x$. From the calculation we just did, we have

$$\mathbb{P}(X > x) = 1 - \mathbb{P}(X \leq x) = 1 - F_X(x) = 1 - (1 - e^{-\lambda x}) = e^{-\lambda x}.$$

So, let's look at conditional probabilities. Suppose we know already that $X > x$. How likely is it that $X > x + y$? Well,

$$\mathbb{P}(X > x + y | X > x) = \frac{\mathbb{P}(X > x + y, X > x)}{\mathbb{P}(X > x)}.$$

Assuming that $y > 0$, the event $\{X > x + y\}$ is a subset of the event $\{X > x\}$. Thus $\mathbb{P}(X > x + y, X > x) = \mathbb{P}(X > x + y)$, and so

$$\mathbb{P}(X > x + y | X > x) = \frac{\mathbb{P}(X > x + y)}{\mathbb{P}(X > x)} = \frac{e^{-\lambda(x+y)}}{e^{-\lambda x}} = e^{-\lambda y} = \mathbb{P}(X > y).$$

This is often described by saying that X is "memoryless". If X describes the amount of time you must wait for an event, then if you have been waiting x, the probability you must wait an additional y is the same as if you hadn't been waiting at all!

Exponential random variables are used to model radioactive decay: if X denotes the length of time before any particular particle decays, then X is assumed to have an $\mathrm{exponential}(\lambda)$ distribution for some λ. This parameter has physical meaning: $\ln 2/\lambda$ is called the *half-life*. We will come back to this we discuss medians.

Example 17.2 (Power Laws)**.** Let $\rho > 1$, and define

$$f(x) = \begin{cases} (\rho - 1)x^{-\rho}, & x \geq 1 \\ 0, & x < 1 \end{cases}.$$

Then $f \geq 0$, and is Riemann integrable with

$$\int_{-\infty}^\infty f(x) = \int_1^\infty (\rho - 1)x^{-\rho}\,dx = (\rho - 1)\,\frac{1}{-\rho + 1}x^{-\rho+1}\bigg|_{x=1}^{x=\infty} = (-1)(0 - 1) = 1.$$

A random variable with this function as its density is said to have a *power law* with exponent ρ. If X has a power law, then

$$\begin{aligned} F_X(x) = \int_{-\infty}^{x} f_X(t)\,dx = \int_1^x (\rho-1)t^{-\rho}\,dt &= (\rho-1)\,\frac{1}{-\rho+1}t^{-\rho+1}\Big|_{t=1}^{t=x} \\ &= -(x^{-\rho+1}-1) \\ &= 1-x^{-(\rho-1)}. \end{aligned}$$

Hence, we can calculate: if X has a power law with parameter 2, say, then

$$\mathbb{P}(3 \le X \le 4) = \mathbb{P}(X=3) + \mathbb{P}(3 < X \le 4) = 0 + F_X(4) - F_X(3) = (1-\tfrac{1}{4}) - (1-\tfrac{1}{3}) = \tfrac{1}{12}.$$

Example 17.3 (Normal Law)**.** Let $t > 0$, and define

$$f(x) = \tfrac{1}{\sqrt{2\pi t}} e^{-x^2/2t}.$$

This is called the **normal denisty** or **Gaussian density** with variance t. (We will soon see that it is appropriate to use the word variance here.) This distribution is so important that we give it a short symbol: we call it

$$N(0,t).$$

(The 0 refers to the expected value, which we'll get to next lecture.) It is strictly positive for all x. To calculate its integral, we use a trick with polar coordinates:

$$\left(\int_{\mathbb{R}} f(x)\,dx\right)^2 = \int_{\mathbb{R}^2} f(x)f(y)\,dxdy = \frac{1}{2\pi t}\int_{\mathbb{R}^2} e^{-(x^2+y^2)/2t}\,dxdy.$$

Substituting polar coordinates this becomes

$$\frac{1}{2\pi t}\int_0^{2\pi}\int_0^{\infty} e^{-r^2/2t} r\,dr\,d\theta = \frac{1}{2\pi t}\left(\int_0^{2\pi} d\theta\right)\left(\int_0^{\infty} re^{-r^2/2t}\,dr\right).$$

The first integral is equal to 2π. The second integrand is an antiderivative:

$$\frac{d}{dr}e^{-r^2/2t} = -\frac{2r}{2t}e^{-r^2/2t} = -\frac{1}{t}re^{-r^2/2t},$$

and so the second integral is

$$\int_0^{\infty} re^{-r^2/2t}\,dr = -t\int_0^{\infty}\frac{d}{dr}e^{-r^2/2t}\,dr = -t\,e^{-r^2/2t}\Big|_{r=0}^{r=\infty} = -t(0-1) = t.$$

Thus, the factor $\frac{1}{2\pi t}$ gets canceled out, and so $\left(\int_{\mathbb{R}} f(x)\,dx\right)^2 = 1$. Since $f \ge 0$, the integral is positive, and so the total mass is 1, making f a probability density.

As we will see in the next few weeks, this is the most important probability density in the world. It can be a little challenging to work with, however. If we wanted to compute the cumulative distribution function of a normal random variable X (with variance t), we would have to evaluate the integral

$$F_X(x) = \int_{-\infty}^{x}\frac{1}{\sqrt{2\pi t}}e^{-y^2/2t}\,dy.$$

It is a long-known fact that this function (while very smooth and having a nice power-series expansion) cannot be written down in "closed-form" (i.e. cannot be expressed as a

composition of polynomials, radicals, exponential/logarithmic, and trigonometric functions). F_X is sometimes called the **error function**. Using numerical integration, some of its properties can be approximated as closely as we like. For example, if X has the normal density with variance t, then it can be computed that

$$\mathbb{P}(|X| \leq \sqrt{t}) = F_X(\sqrt{t}) - F_X(-\sqrt{t}) \approx 0.68$$
$$\mathbb{P}(|X| \leq 2\sqrt{t}) = F_X(2\sqrt{t}) - F_X(-2\sqrt{t}) \approx 0.95$$

17.1. **Transforming random variables.** One very helpful feature of continuous random variables is that we have tools like the fundamental theorem of calculus to help us figure out how distributions change when we compose with new functions.

Example 17.4. *Suppose X is* $\text{exponential}(\lambda)$*. What is the density of aX for some $a > 0$?*

To answer this, we will first calculate the cumulative distribution function.

$$F_{aX}(x) = \mathbb{P}(aX \leq x) = \mathbb{P}(X \leq x/a) = F_X(x/a).$$

Actually, this calculation works for *any* distribution. Now, as calculated in Example 17.1, we can compute

$$F_X(x/a) = 1 - e^{-\lambda x/a}.$$

So this is the cumulative distribution function of aX. To find the density of aX, we simply have to differentiate (which is legal here since F_X is differentiable).

$$f_{aX}(x) = \frac{d}{dx}(1 - e^{-\lambda x/a}) = \frac{\lambda}{a} e^{-\frac{\lambda}{a}x}.$$

In other words, aX is a an $\text{exponential}(\lambda/a)$ random variable.

Actually, using calculus, we can see how this kind of transformation works for any (smooth enough) distribution.

Theorem 17.5. *Let X be a random variable with a continuous density f_X. Let $r\colon \mathbb{R} \to \mathbb{R}$ be a differentiable function that is strictly-increasing. Then the density of the random variable $r(X)$ is*

$$f_{r(X)}(x) = f_X(r^{-1}(x)) \cdot (r^{-1})'(x).$$

Proof. We work with cumulative distribution functions.

$$F_{r(X)}(x) = \mathbb{P}(r(X) \leq x) = \mathbb{P}(X \leq r^{-1}(x)).$$

The last equality is valid since r is strictly-increasing and differentiable, so it has an inverse, and $r(y) \leq x$ if and only if $y \leq r^{-1}(x)$. Well, $\mathbb{P}(X \leq r^{-1}(x)) = F_X(r^{-1}(x))$. Since the density of X is continuous, we can use the fundamental theorem of calculus to get the new density.

$$f_{r(X)}(x) = \frac{d}{dx} F_{r(X)}(x) = \frac{d}{dx} F_X(r^{-1}(x)) = F_X'(r^{-1}(x)) \cdot \frac{d}{dx} r^{-1}(x) = f_X(r^{-1}(x))(r^{-1})'(x),$$

as claimed. □

Example 17.6. Suppose X has an $\text{exponential}(\lambda)$ distribution. Then we can calculate the density of the new random variable e^X. The function $r(x) = e^x$ is strictly-increasing and differentiable, so we apply Theorem 17.5 to find

$$f_{e^X}(x) = f_X(r^{-1}(x)) \cdot (r^{-1})'(x).$$

We can explicitly calculate that $r^{-1}(x) = \ln x$ and $(r^{-1})'(x) = \frac{1}{x}$. So

$$f_{e^X}(x) = f_X(\ln x) \cdot \frac{1}{x}.$$

When $\ln x < 0$, this is 0 since $f_X(y) = 0$ for $y < 0$. For $\ln x \geq 0$ (i.e. $x \geq 1$), the density of X is $f_Y(y) = \lambda e^{-\lambda y}$, so in this regime

$$f_{e^X}(x) = \lambda e^{-\lambda \ln x} \cdot \frac{1}{x}.$$

We can simplify $e^{-\lambda x} = e^{\ln(x^{-\lambda})} = x^{-\lambda}$, so for $x \geq 1$,

$$f_{e^X}(x) = \lambda x^{-\lambda} \cdot \frac{1}{x} = \lambda x^{-(\lambda+1)}.$$

In other words, e^X has a power law distribution with parameter $\rho = \lambda + 1$.

18. LECTURE 18: NOVEMBER 5, 2010

Example 18.1. *Point a flashlight at a(n infinitely tall) wall, with an angle chosen uniformly from* $[-\frac{\pi}{2}, \frac{\pi}{2}]$. *Let* X *be the height on the wall where the light hits. What is the distribution of* X*?*

Let Θ be the random angle chosen. Θ is uniform on $[-\frac{\pi}{2}, \frac{\pi}{2}]$, which means that its density is

$$f_\Theta(\theta) = \begin{cases} \frac{1}{\pi}, & -\frac{\pi}{2} \leq \theta \leq \frac{\pi}{2} \\ 0, & \text{otherwise} \end{cases}$$

Now, from trigonometry, we see that the height X is the opposite side of a right triangle with angle Θ. If you stand distance d away from the wall, this means $\frac{X}{d} = \tan\Theta$. So $X = r(\Theta) = d\tan\Theta$. The function r is strictly-increasing and differentiable, so by Theorem 17.5,

$$f_X(x) = f_{r(\Theta)}(x) = f_\Theta(r^{-1}(x))(r^{-1})'(x).$$

The inverse function is $r^{-1}(x) = \tan^{-1}(x/d)$. This function is *always* valued in $(-\frac{\pi}{2}, \frac{\pi}{2})$, and so for any x the variable $\theta = \tan^{-1}(x/d)$ has $f_\Theta(\theta) = \frac{1}{\pi}$. Now,

$$\frac{d}{dx} r^{-1}(x) = \frac{d}{dx}\tan^{-1}(x/d) = \frac{1}{1+(x/d)^2} \cdot \frac{1}{d} = \frac{d}{d^2+x^2}.$$

Hence,

$$f_X(x) = \frac{d}{\pi} \cdot \frac{1}{d^2+x^2}.$$

Normalizing to $d = 1$, this is known as the **Cauchy density**, the density of the **Cauchy distribution**.

Example 18.2. *Suppose that* X *is* $\text{exponential}(\lambda)$. *Let* $Y = 1 - e^{-\lambda X}$. *What is the density of* Y*?*

Let $F(x) = 1 - e^{-\lambda x}$. This function is strictly-increasing and differentiable, so we may change variables a la Theorem 17.5. We will do it from scratch here. First note that $e^{-\lambda x} \leq 1$ when $x \geq 0$, and $e^{-\lambda x} > 0$ for all x. Since X is $\text{exponential}(\lambda)$, $\mathbb{P}(X < 0) = 0$, and so we have $\mathbb{P}(F(X) \in (0,1)) = 1$. Thus we can immediately see that $F_Y(y) = 1$ for $y \geq 1$ and $F_Y(y) = 0$ for $y \leq 0$. For $y \in (0,1)$, we calculate

$$\begin{aligned} F_Y(y) = \mathbb{P}(Y \leq y) &= \mathbb{P}(1 - e^{-\lambda X} \leq y) \\ &= \mathbb{P}(1 - y \leq e^{-\lambda X}) \\ &= \mathbb{P}(\ln(1-y) \leq -\lambda X) \\ &= \mathbb{P}(X \leq -\tfrac{1}{\lambda}\ln(1-y)) = F_X(-\tfrac{1}{\lambda}\ln(1-y)). \end{aligned}$$

When $y \in (0,1)$, $\ln(1-y) \in (-\infty, 0)$, and so $x = -\frac{1}{\lambda}\ln(1-y) > 0$. Thus

$$F_X(x) = 1 - e^{-\lambda x} = 1 - e^{-\lambda\cdot(-\frac{1}{\lambda})\ln(1-y))} = y.$$

That is, the cumulative distribution function of Y is

$$F_Y(y) = \begin{cases} 0, & y \leq 0 \\ y, & 0 < y < 1 \\ 1, & y \geq 1 \end{cases}.$$

In other words, Y has a **uniform** distribution on $[0,1]$.

The result of Example 18.2 is actually quite general; we can state it as the following.

Proposition 18.3. *Let X be a continuous random variable with cumulative distribution function F_X. Then the random variable $Y = F_X(X)$ is uniform on $[0, 1]$.*

Proof. Since the range of F_X is contained in $[0, 1]$, the same is true for Y, and so we know already that $F_Y(y) = 0$ for $y \leq 0$ and $F_Y(y) = 1$ for $y \geq 1$. In between, let us first make the simplifying assumption that the function F_X is strictly increasing and continuously differentiable. (This is the same as assuming that the density f_X is continuous and strictly positive.) Then we can apply Theorem 17.5 directly to get for $y \in (0, 1)$

$$f_Y(y) = f_X(F_X^{-1}(y)) \cdot (F_X^{-1})'(y) = F_X'(F_X^{-1}(y)) \cdot (F_X^{-1})'(y).$$

The chain rule says that

$$(F_X \circ F_X^{-1})'(y) = F_X'(F_X^{-1}(y)) \cdot (F_X^{-1})'(y).$$

But $F_X \circ F_X^{-1}(y) = y$, so the derivative of this function is just 1. Thus, for $0 < y < 1$, $f_Y(y) = 1$, showing that Y is uniform on $[0, 1]$.

Actually, we don't need this fancy an approach, or the differentiability assumption. If F_X is strictly increasing, then so is F_X^{-1}, and we can compute directly with the cumulative distribution functions: for $y \in (0, 1)$

$$F_Y(y) = \mathbb{P}(F_X(X) \leq y) = \mathbb{P}(X \leq F_X^{-1}(y)) = F_X(F_X^{-1}(y)) = y. \tag{18.1}$$

This is problematic if F_X isn't strictly-increasing; but we can fix this by thinking about the graph. If F_X is flat, $F_X(x) = y_0$ for all $x \in [x_0, x_1]$, then the reflection of the graph of F_X across the line $y = x$ gives a function which has a jump at y_0 from x_0 up to x_1. To formalize this, we can generally define

$$F_X^{-1}(y) \equiv \min\{x \,:\, F(x) \geq y\}. \tag{18.2}$$

This is equal to the inverse in the case that F_X is strictly-increasing; in general, it gives a function with jumps and puts the value at the *bottom* of the jump. (I.e. F_X^{-1} is *left*-continuous, not right-continuous.) With this definition, the argument of Equation 18.1 works in general. In fact, it is not even necessary for X to be continuous! (F_X^{-1} will have jumps where X is flat, and will be flat where X has jumps.) □

We can turn this around to give a very useful construction of a random variable with any given distribution.

Theorem 18.4. *Let U be a uniform random variable on $[0, 1]$. Let F be* any *cumulative distribution function. Then the random variable $X = F^{-1}(U)$ has cumulative distribution function F.*

Proof. The function F^{-1} used in the statement is the one in Equation 18.2. The idea of the proof is captured in the special case that F is strictly-increasing, so we'll stick to that case here. We have

$$F_X(x) = \mathbb{P}(X \leq x) = \mathbb{P}(F^{-1}(U) \leq x) = \mathbb{P}(U \leq F(x)) = F_U(F(x)).$$

Now, the function $F_U(y) = y$ when $y \in [0, 1]$, and since F is a cumulative distribution function, $F(x) \in [0, 1]$ for sure. Thus, $F_X(x) = F_U(F(x)) = F(x)$ – i.e. $F_X = F$, as desired.

To make this work in general, the only step that needs to be verified is the condition

$$F^{-1}(y) \leq x \qquad \Longleftrightarrow \qquad y \leq F(x).$$

The function of Equation 18.2 is designed exactly to make this true for all x, y. □

Theorem 18.4 is a vital theoretical and practical tool. Theoretically, it allows us to conclude that any distribution function really is the distribution function of *some* random variable: all we need to know is that a uniform random variable exists! This is also of great practical interest.

Example 18.5. Suppose that I want to simulate an $\mathrm{exponential}(3)$ random variable; how do I do it? First I generate a list of uniformly random data points $u_1, u_2, \ldots, u_n$ in $[0, 1]$. Then I take the cumulative distribution function of the $\mathrm{exponential}(3)$ distribution, $F(x) = 1 - e^{-3x}$, I find its inverse $F^{-1}(y) = -\frac{1}{3}\ln(y-1)$, and then I compose: since $u_1, u_2, \ldots, u_n$ are samples of a uniform random variable U, the data set

$$-\frac{1}{3}\ln(u_1 - 1), \ldots, -\frac{1}{3}\ln(u_n - 1)$$

are samples from the random variable $F^{-1}(U)$, which is $\mathrm{exponential}(3)$ according to Theorem 18.4. This is actually how computer software (like Matlab, Maple, Mathematica, and statistics packages) generate random data with some given distribution: all the program really knows how to do is generate random uniform data; it then transforms it using Theorem 18.4 to any desired distribution.

18.1. **Expected Value.** For discrete random variables X, we defined the expected value $\mathbb{E}(X)$ (when it exists) to be

$$\mathbb{E}(X) = \sum_x x\mathbb{P}(X = x).$$

This won't do as a good definition for a continuous random variable, since $\mathbb{P}(X = x) = 0$ for all x, and the sum would have a continuum of terms. To figure out what $\mathbb{E}(X)$ should mean in the continuous setting, we approach the problem like Riemann did when formally defining the integral as a limit of sums.

Suppose X is a continuous random variable with real values. To simplify matters, let's assume $\mathbb{P}(X \in [a, b]) = 1$ for some finite interval $[a, b]$. (Such a random variable is called **bounded**; equivalently, we are assuming that f_X is 0 outside $[a, b]$.) So the cumulative distribution function $F = F_X$ is a continuous function with $F(x) = 0$ for $x \leq a$ and $F(x) = 1$ for $x \geq b$. Let's further suppose that F is actually C^1 on the interval $[a, b]$. Now, let's approximate this function by a step function: let $a = x_0 < x_1 < x_2 < \cdots < x_n = b$ be a partition of this interval. Let Y be a discrete random variable with

$$F_Y(x) = F(x_j), \quad x_j \leq x < x_{j+1}.$$

(I.e. we do left-end-point approximation.) This means that

$$\mathbb{P}(Y = x_j) = F(x_j) - F(x_{j-1}).$$

So

$$\mathbb{E}(Y) = \sum_{j=1}^{n} x_j \mathbb{P}(Y = x_j) = \sum_{j=1}^{n} x_j [F(x_j) - F(x_{j-1})].$$

Now, since F is C^1 on each interval $[x_{j-1}, x_j]$, the mean value theorem shows us that there is a point $x_j^* \in [x_{j-1}, x_j]$ where

$$F(x_j) - F(x_{j-1}) = F'(x_j^*)(x_j - x_{j-1}).$$

Since F is differentiable, $F'(x) = f_X(x)$ is the density of X. So, putting the pieces together, we have

$$\mathbb{E}(Y) = \sum_{j=1}^{n} x_j f_X(x_j^*)(x_j - x_{j-1}).$$

This is a Riemann sum! If we take the partition points x_j closer and closer together, we get

$$\sum_{j=1}^{n} x_j f_X(x_j^*)(x_j - x_{j-1}) \to \int_a^b x f_X(x)\, dx.$$

On the other hand, when the partition points are very close together, Y is a good approximation to X. All this motivates our definition.

Definition 18.6. *Let X be a continuous random variable, with density f_X. The* **expectation** *or* **expected value** *of X, denoted $\mathbb{E}(X)$ is defined to be*

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

when this integral exists and is finite.

Example 18.7. Let $X \sim \text{exponential}(\lambda)$. Then $f_X(x) = \lambda e^{-\lambda x}$ for $x > 0$, and so

$$\mathbb{E}(X) = \int_0^{\infty} x \cdot \lambda e^{-\lambda x}\, dx.$$

This is an integration by parts problem: setting $U = x$ and $dV = \lambda e^{-\lambda x}\, dx$, we have

$$\mathbb{E}(X) = \int_0^{\infty} U\, dV = UV|_0^{\infty} - \int_0^{\infty} V\, dU.$$

Note that $V = \int dV = \int \lambda e^{-\lambda x}\, dx = -e^{-\lambda x} + C$. (When interpreting this integral as the cumulative distribution function of X, we need to take $C = 1$, but we can take any C we like since it will cancel between the integral and boundary terms. So it is most convenient here to take $C = 0$.) Hence

$$\mathbb{E}(X) = -xe^{-\lambda x}\Big|_{x=0}^{x=\infty} + \int_0^{\infty} e^{-\lambda x}\, dx.$$

When $x = 0$, $xe^{-\lambda x} = 0 \cdot 1 = 0$; as $x \to \infty$, $xe^{-\lambda x} \to 0$ as well, so the boundary terms are 0. For the integral, we have

$$\mathbb{E}(X) = \int_0^{\infty} e^{-\lambda x}\, dx = -\frac{1}{\lambda} e^{-\lambda x}\Big|_{x=0}^{x=\infty} = -\frac{1}{\lambda}(0 - 1) = \frac{1}{\lambda}.$$

So an $\text{exponential}(\lambda)$ random variable has expectation $\frac{1}{\lambda}$.

19. LECTURE 19: NOVEMBER 8, 2010

Example 19.1. Let X have a power law, $f_X(x) = (\rho - 1)x^{-\rho}$, $x \geq 1$, for some $\rho > 1$. Then

$$\mathbb{E}(X) = \int_1^\infty x \cdot (\rho - 1)x^{-\rho}\, dx = (\rho - 1)\int_1^\infty x^{-\rho+1}\, dx.$$

Now we run into a little trouble. If $1 < \rho \leq 2$, then $-\rho + 1 \in [-1, 0)$, and so $x^{-\rho+1}$ is not integrable. Thus, the power law does not have a finite expectation for $1 < \rho \leq 2$. When $\rho > 2$, we have

$$\mathbb{E}(X) = (\rho - 1)\left.\frac{1}{-\rho+2}x^{-\rho+2}\right|_{x=1}^{x=\infty} = \frac{\rho-1}{2-\rho}(0-1) = \frac{1-\rho}{2-\rho}.$$

Asking when $\mathbb{E}(X)$ exists is a little dicey. The trouble is that the integrand $xf_X(x)$ is ≥ 0 when $x > 0$ and is ≤ 0 when $x < 0$. So strange cancelations can occur depending how you let the limits of integration tend to $\pm\infty$.

Example 19.2. Let X have the Cauchy distribution from Example 18.1, $f_X(x) = \frac{1}{\pi} \cdot \frac{1}{1+x^2}$. Then

$$\int_{-\infty}^\infty x f_X(x)\, dx = \frac{1}{\pi}\int_{-\infty}^\infty \frac{x}{1+x^2}\, dx.$$

Since $\frac{x}{1+x^2} \sim \frac{1}{x}$ as $x \to \infty$, a function that is not integrable, this integral is ill-defined. To see why, note that the definition is actually a limit:

$$\int_{-\infty}^\infty \frac{x}{1+x^2}\, dx = \lim_{a\to-\infty}\lim_{b\to\infty}\int_a^b \frac{x}{1+x^2}\, dx.$$

We can evaluate the integral over $[a, b]$:

$$\int_a^b \frac{x}{1+x^2}\, dx = \frac{1}{2}\ln(1+x^2)\Big|_{x=a}^{x=b} = \frac{1}{2}\ln\left(\frac{1+b^2}{1+a^2}\right).$$

Now, we could decide to take $-a = b \to \infty$, in which case we would get

$$\lim_{a\to-\infty}\frac{1}{2}\ln\left(\frac{1+a^2}{1+a^2}\right) = 0,$$

suggesting that the expected value should be 0. On the other hand, in this double limit, we could have decided instead to let $b = -2a \to \infty$, in which case we would get

$$\lim_{a\to-\infty}\frac{1}{2}\ln\left(\frac{1+4a^2}{1+a^2}\right) = \ln 2.$$

Since the answer depends on how we perform the limit, this improper integral is *not well-defined*. The problem is that

$$\begin{aligned}\int_{-\infty}^\infty \left|\frac{x}{1+x^2}\right| dx = \int_{-\infty}^0 \frac{-x}{1+x^2}\, dx + \int_0^\infty \frac{x}{1+x^2}\, dx &= 2\int_0^\infty \frac{x}{1+x^2}\, dx \\ &= 2\lim_{b\to\infty}\int_0^b \frac{x}{1+x^2}\, dx\end{aligned}$$

and this limit is

$$\lim_{b\to\infty}\int_0^b \frac{x}{1+x^2}\,dx = \lim_{b\to\infty}\frac{1}{2}\ln(1+b^2) = \infty.$$

So, even though the Cauchy distribution is symmetric, its expected value is of the form $\infty - \infty$ which is not-defined.

This suggests we should look at the absolute value.

Proposition 19.3. *Let X be a continuous random variable. The expected value $\mathbb{E}(X)$ exists if and only if the positive integral*

$$\int_{-\infty}^{\infty} |xf_X(x)|\;dx = \int_{-\infty}^{\infty} |x| f_X(x)\,dx < \infty.$$

Proof. For any $a < 0 < b$ we have

$$\int_a^b |x| f_X(x)\,dx = \int_0^b xf_X(x)\,dx + \int_a^0 |x| f_X(x)\,dx.$$

Hence

$$\int_{-\infty}^{\infty} |x| f_X(x)\,dx = \lim_{b\to\infty}\int_0^b xf_X(x)\,dx + \lim_{a\to-\infty}\int_a^0 |x| f_X(x)\,dx.$$

Both of these integrals are increasing positive limits, so they either both exists or the sum is infinite. Thus, if the integral of $|x|f_X(x)$ is finite therefore implies that both limits exist, and so

$$\mathbb{E}(X) = \int_{-\infty}^{\infty} xf_X(x)\,dx = \lim_{b\to\infty}\int_0^b xf_X(x)\,dx - \lim_{a\to-\infty}\int_a^0 |x| f_X(x)\,dx$$

is a difference of two finite limits that exist, so it exists. On the other hand, if the full integral of $|x|f_X(x)$ is infinite, then at least one of the two one-sides limits is infinite. If the $\lim_{b\to\infty}\int_0^b xf_X(x)\,dx = \infty$ but $\lim_{a\to-\infty}\int_a^0 |x|f_X(x)\,dx$ exists, then their difference tends to ∞ so $\mathbb{E}(X) = \infty$. If the $\lim_{b\to\infty}\int_0^b xf_X(x)\,dx$ exists but $\lim_{a\to-\infty}\int_a^0 |x|f_X(x)\,dx = -\infty$, then their difference tends to $-\infty$ so $\mathbb{E}(X) = -\infty$. If both limits tend to ∞, then $\mathbb{E}(X) = \infty - \infty$ is an indeterminate form, and will take on different values depending how the two limits are performed, as in Example 19.2. □

Example 19.4. Let $X \sim N(0,t)$, a normal distribution, with density $f_X(x) = \frac{1}{\sqrt{2\pi t}}e^{-x^2/2t}$. Following Proposition 19.3, we need to check the positive integral

$$\int_{-\infty}^{\infty} |x| f_X(x)\,dx = \frac{1}{\sqrt{2\pi t}}\int_{-\infty}^{\infty} |x| e^{-x^2/2t}\,dx.$$

Breaking this integral up into two pieces,

$$\begin{aligned}\int_{-\infty}^{\infty} |x| e^{-x^2/2t}\,dx &= \int_{-\infty}^{0} |x| e^{-x^2/2t}\,dx + \int_0^{\infty} |x| e^{-x^2/2t}\,dx \\ &= \int_{-\infty}^{0} (-x) e^{-x^2/2t}\,dx + \int_0^{\infty} xe^{-x^2/2t}\,dx.\end{aligned}$$

In the first integral, we make the change of variables $y = -x$. Then $dx = -dy$ and so this integral becomes

$$\int_{-\infty}^{0} (-x)e^{-x^2/2t}\,dx = -\int_{\infty}^{0} ye^{-y^2/2t}\,dy = \int_{0}^{\infty} ye^{-y^2/2t}\,dy,$$

the same as the first integral. Hence

$$\int_{-\infty}^{\infty} |x| f_X(x)\,dx = \frac{2}{\sqrt{2\pi t}}\int_{0}^{\infty} xe^{-x^2/2t}\,dt\frac{2}{\sqrt{2\pi t}}\cdot -te^{-x^2/2t}\Big|_{x=0}^{x=\infty} = \sqrt{\frac{2t}{\pi}} < \infty.$$

Hence, $\mathbb{E}(X)$ exists. We can calculate it using the same calculations above:

$$\begin{aligned}\int_{-\infty}^{\infty} xf_X(x)\,dx &= \frac{1}{\sqrt{2\pi t}}\int_{-\infty}^{0} xe^{-x^2/2t}\,dx + \frac{1}{\sqrt{2\pi t}}\int_{0}^{\infty} xe^{-x^2/2t}\,dx \\ &= -\frac{1}{\sqrt{2\pi t}}\int_{0}^{\infty} ye^{-y^2/2t}\,dy + \frac{1}{\sqrt{2\pi t}}\int_{0}^{\infty} xe^{-x^2/2t}\,dx = 0.\end{aligned}$$

So, if X has a normal $N(0,t)$ distribution, $\mathbb{E}(X) = 0$. (This is the reason for the 0 in $N(0,t)$.)

Remark 19.5. The phenomenon that we saw in Example 19.4, with the two integrals over the positive and negative half-lines canceling, is because the normal distribution has the property that $f_X(x) = f_X(-x)$. Such distributions are called *symmetric*. This is the same condition as the one you saw on Homework 6, problem 7(a): integrating both sides we find

$$\begin{aligned}F_X(x) = \int_{-\infty}^{x} f_X(t)\,dt = \int_{-\infty}^{x} f_X(-t)\,dt &= \int_{\infty}^{-x} f_X(s)\,(-ds) \\ &= \int_{-x}^{\infty} f_X(s)\,ds \\ &= 1 - \int_{-\infty}^{-x} f_X(s)\,ds = 1 - F_X(-x).\end{aligned}$$

That is, $F_X(x)+F_X(-x) = 1$. (As you showed, this is equivalent to $\mathbb{P}(a \le X \le b) = \mathbb{P}(-b \le X \le -a)$ for all $a < b$.) In such a case, X distributes half of its mass in $(-\infty, 0)$ and the other half in $(0,\infty)$, so it is natural to expect that $\mathbb{E}(X) = 0$. This is true whenever X is symmetric, *provided* $\mathbb{E}(X)$ *exists*. As Example 19.2 showed, there are symmetric random variables that do not have an expectation.

19.1. **Properties of Expectation.** The expectation on continuous random variables has the same properties as it does on discrete random variables. The main property (linearity of $\mathbb{E}$) we will leave until a future lecture, when we discuss joint densities and joint distributions of continuous random variables. But we can prove a few useful properties along these lines now.

Example 19.6. Let X be a continuous random variable and $a \in \mathbb{R}$. Then, as we calculated in Example 17.4, $F_{aX}(x) = F_X(x/a)$. Differentiating, this shows that

$$f_{aX}(x) = \frac{d}{dx}F_X(x/a) = \frac{1}{a}f_X(x/a).$$

Therefore

$$\mathbb{E}(aX) = \int_{-\infty}^{\infty} x f_{aX}(x)\, dx = \int_{-\infty}^{\infty} \frac{x}{a} f_X(x/a)\, dx.$$

Making the substitution $y = x/a$ (which takes $(-\infty, \infty)$ to itself) gives $dx = a dy$ and so

$$\mathbb{E}(aX) = \int_{-\infty}^{\infty} y f_X(y)\, a\, dy = a\mathbb{E}(X).$$

Similarly, we can calculate that

$$F_{X+a}(x) = \mathbb{P}(X + a \leq x) = \mathbb{P}(X \leq x - a) = F_X(x - a),$$

and so differentiating yields $f_{X+a}(x) = f_X(x - a)$. Hence

$$\mathbb{E}(X + a) = \int_{-\infty}^{\infty} x f_{X+a}(x)\, dx = \int_{-\infty}^{\infty} x f_X(x - a)\, dx.$$

Making the substitution $y = x - a$ (which takes $(-\infty, \infty)$ to itself) gives $dx = dy$ and so

$$\mathbb{E}(X + a) = \int_{-\infty}^{\infty} (y + a) f_X(y)\, dy = \int_{-\infty}^{\infty} y f_X(y)\, dy + a \int_{-\infty}^{\infty} f_X(y)\, dy = \mathbb{E}(X) + a \cdot 1.$$

Example 19.7. Let $X \sim N(0, t)$. Then the density of $X + a$ is

$$f_{X+a}(x) = f_X(x - a) = \frac{1}{\sqrt{2\pi t}} e^{-(x-a)^2/2t}.$$

Using the calculations in Examples 19.4 and 19.6, we have $\mathbb{E}(X + a) = 0 + a$. We call this density $N(a, t)$, a **normal law with mean** a **and variance** t. (We'll explain the variance part in the next section.)

19.2. **Expectations of functions of continuous random variables.** We can also compute expectations of *functions* of a random variable, just as in the discrete case (provided the new composed variable has an expectation). Let X have density f_X. Let $r \colon \mathbb{R} \to \mathbb{R}$ be a differentiable function. Then

$$\mathbb{E}(r(X)) = \int_{-\infty}^{\infty} x f_{r(X)}\, dx.$$

Let's consider the case that r is strictly-increasing, and maps $\mathbb{R}$ onto $\mathbb{R}$. Then we have a formula, Theorem 17.5, for $f_{r(X)}$:

$$f_{r(X)}(x) = f_X(r^{-1}(x)) \cdot (r^{-1})'(x).$$

So we have

$$\mathbb{E}(r(X)) = \int_{-\infty}^{\infty} x f_X(r^{-1}(x)) \cdot (r^{-1})'(x)\, dx.$$

This is setup perfectly for the substitution (i.e. reverse chain) rule. Set $y = r^{-1}(x)$. Then $dy = (r^{-1})'(x)\, dx$, and $x = r(y)$, so this last integral therefore becomes

$$\mathbb{E}(r(X)) = \int_{-\infty}^{\infty} r(y) f_X(y)\, dy.$$

Actually, this last equality holds in full generality (even when r is not onto, or strictly-increasing).

Theorem 19.8. *Let X be a random variable with density f_X, and let $r\colon \mathbb{R} \to \mathbb{R}$. If the integral $\int_{-\infty}^{\infty} |r(x)| f_X(x)\,dx$ is finite, then $\mathbb{E}(r(X))$ exists and*

$$\mathbb{E}(r(X)) = \int_{-\infty}^{\infty} r(x) f_X(x)\,dx.$$

One important application of this is calculating variances.

Example 19.9. Let $X \sim N(0,t)$. Then

$$\mathbb{E}(X^2) = \int_{-\infty}^{\infty} x^2 f_X(x)\,dx = \frac{1}{\sqrt{2\pi t}} \int_{-\infty}^{\infty} x^2 e^{-x^2/2t}\,dx.$$

We can do this integration by parts. Set $U(x) = x$ and $dV(x) = xe^{-x^2/2t}\,dx$. Then $V(x) = -te^{-x^2/2t}$, and so

$$\begin{aligned}\int_{-\infty}^{\infty} x^2 e^{-x^2/2t}\,dx = \int_{-\infty}^{\infty} U\,dV &= UV|_{-\infty}^{\infty} - \int_{-\infty}^{\infty} V\,dU \\ &= -txe^{-x^2/2t}\Big|_{x=-\infty}^{x=\infty} + \int_{-\infty}^{\infty} te^{-x^2/2t}\,dx.\end{aligned}$$

The boundary terms vanish as $x \to \infty$. So, we have

$$\mathbb{E}(X^2) = \frac{1}{\sqrt{2\pi t}} \int_{-\infty}^{\infty} te^{-x^2/2t}\,dx = t\int_{-\infty}^{\infty} f_X(x)\,dx = t \cdot 1.$$

So $\mathbb{E}(X^2) = t$. Since $\mathbb{E}(X) = 0$, we also have

$$\mathrm{Var}(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2 = t.$$

This is the reason we call the distribution $N(0,t)$ – a normal random variable with expectation 0 and variance t.

Example 19.10. Let X be uniform on $[a,b]$. Then

$$\mathbb{E}(X^2) = \int_{-\infty}^{\infty} x^2 f_X(x)\,dx = \int_a^b x^2 \frac{1}{b-a}\,dx = \frac{1}{b-a}\,\frac{1}{3}x^3\Big|_{x=a}^{x=b} = \frac{b^3-a^3}{3(b-a)} = \frac{1}{3}(b^2+ab+a^2).$$

On the other hand,

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

Hence,

$$\mathrm{Var}(X) = \frac{1}{3}(b^2+ab+a^2) - \left(\frac{1}{2}(a+b)\right)^2 = \frac{(b-a)^2}{12}.$$

Example 19.11. Let $X \sim \text{exponential}(\lambda)$. Then

$$\mathbb{E}(X^2) = \int_0^{\infty} x^2 \cdot \lambda e^{-\lambda x}\,dx$$

The integral can, once again, be done by parts. Letting $U = x^2$ and $dV = \lambda e^{-\lambda x} dx$, we have $V(x) = -e^{-\lambda x}$ and so

$$\mathbb{E}(X^2) = \int_0^\infty U dV = UV|_0^\infty - \int_0^\infty V dU = -x^2 e^{-\lambda x}\big|_{x=0}^{x=\infty} + \int_0^\infty e^{-\lambda x} \cdot 2x \, dx.$$

The boundary terms vanish, and so we have

$$\mathbb{E}(X^2) = 2\int_0^\infty x e^{-\lambda x} \, dx = \frac{2}{\lambda} \int_0^\infty x \cdot \lambda e^{-\lambda x} \, dx = \frac{2}{\lambda}\mathbb{E}(X).$$

We already calculated that $\mathbb{E}(X) = \frac{1}{\lambda}$, so $\mathbb{E}(X^2) = \frac{2}{\lambda^2}$. Hence

$$\mathrm{Var}(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2 = \frac{2}{\lambda^2} - \left(\frac{1}{\lambda}\right)^2 = \frac{1}{\lambda^2}.$$

20. LECTURE 20: NOVEMBER 10, 2010

20.1. **Joint Distributions.** As we saw in the discrete setting, if we have two random variables X, Y we could completely describe their distribution (and relationship with each other) by knowing the numbers $\mathbb{P}(X = x, Y = y)$ for all states x, y that they can take on. In the continuous case, these numbers are typically 0. Instead, we use the same trick, looking at an inequality instead of equality.

Definition 20.1. *Let X, Y be any two $\mathbb{R}$-valued random variables. The function*

$$F_{X,Y}(x, y) = \mathbb{P}(X \leq x, Y \leq y)$$

is called the **joint cumulative distribution function** *of the pair.*

Notice that

$$\begin{aligned} F_{X,Y}(x, y) = \mathbb{P}(X \leq x, Y \leq y) &\leq \mathbb{P}(X \leq x) = F_X(x) \\ &\leq \mathbb{P}(Y \leq y) = F_Y(y) \end{aligned}$$

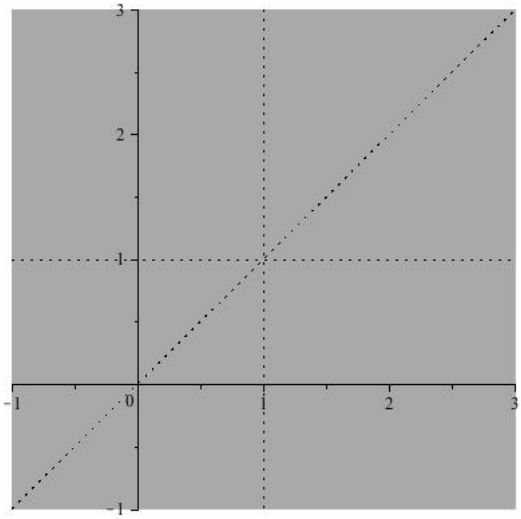
so $F_{X,Y}(x, y) \leq \min\{F_X(x), F_Y(y)\}$. In particular, this means that $\lim_{x\to-\infty} F_{X,Y}(x, y) = 0$ for any y, and $\lim_{y\to-\infty} F_{X,Y}(x, y) = 0$ for any x.

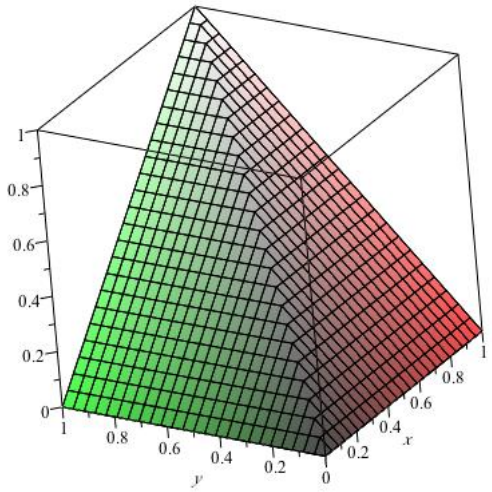
Example 20.2. *Suppose X, Y are uniform random variables on $[0, 1]$, and in fact $X = Y$. What does $F_{X,Y}$ look like?*

First, suppose $x < y$. Then $\mathbb{P}(X \leq x, Y \leq y) = \mathbb{P}(X \leq x, X \leq y)$. The event $\{X \leq x\} \cap \{X \leq y\}$ is the same as the even $\{X \leq \min\{x, y\}\}$. Thus,

$$F_{X,Y}(x, y) = \mathbb{P}(X \leq \min\{x, y\}) = \begin{cases} 0, & \min\{x, y\} \leq 0 \\ \min\{x, y\}, & 0 \leq \min\{x, y\} \leq 1 \\ 1, & \min\{x, y\} \geq 1 \end{cases}.$$

Even in this simple case, the function $F_{X,Y}$ has quite a complicated graph. The function





is 0 except in the first quadrant. It is equal to 1 in the region where $x, y \geq 1$. In the intervening "L"-shaped region, it equals x in the top part (where $0 \leq x \leq 1$ and $y \geq x$) and equals y in the bottom part (where $0 \leq y \leq 1$ and $x \geq y$.

In general, if $X = Y$, the joint distribution function $F_{X,Y} = F_{X,X}$ is actually an undesirable object. In the discrete case, where we look instead directly at the distribution function $\mu_{X,Y}(x, y) = \mathbb{P}(X = x, Y = y)$, when $X = Y$ we would have $\mu_{X,X}(x, y) = 0$ unless $x = y$: i.e. the only non-zero values are on the diagonal. In the continuous case, the analogue of the distribution function μ_X is the *probability density function* f_X; so to understand things better, we should figure out what a joint density looks like.

20.2. **Joint Density.** Following what we did in the 1-variable case, let's consider the special case that the joint distribution function $F_{X,Y}$ is actually a smooth function. Since there are two variables now, we're going to want to differentiate in each variable, which means two derivatives – so we consider the case that $F_{X,Y}$ is $C^2(\mathbb{R}^2)$. Consider the function

$$f_{X,Y}(x, y) = \frac{\partial}{\partial x}\frac{\partial}{\partial y}F_{X,Y}(x, y).$$

This is called the **joint density** of X and Y. We can recover the joint cumulative distribution from it by integrating. Since we know $\lim_{x\to-\infty} F_{X,Y}(x, y) = \lim_{y\to-\infty} F_{X,Y}(x, y) = 0$, the fundamental theorem of calculus tells us that

$$\begin{aligned}\int_{-\infty}^{a} f_{X,Y}(x, y)\,dx = \int_{-\infty}^{a} \frac{\partial}{\partial x}\frac{\partial}{\partial y}F_{X,Y}(x, y)\,dx &= \frac{\partial}{\partial y}F_{X,Y}(a, y) - \frac{\partial}{\partial y}F_{X,Y}(-\infty, y)\\ &= \frac{\partial}{\partial y}F_{X,Y}(a, y).\end{aligned}$$

So, integrating again,

$$\begin{aligned}\int_{-\infty}^{b}\int_{-\infty}^{a} f_{X,Y}(x, y)\,dx\,dy = \int_{-\infty}^{b} \frac{\partial}{\partial y}F_{X,Y}(a, y)\,dy &= F_{X,Y}(a, b) - F_{X,Y}(a, -\infty)\\ &= F_{X,Y}(a, b).\end{aligned}$$

So the joint density $f_{X,Y}$ contains the same information as $F_{X,Y}$. But it allows us to calculate many things much more easily.

Example 20.3. *Let X and Y be continuous random variables with joint cumulative distribution function $F_{X,Y}(x, y) = \mathbb{P}(X \le x, Y \le y)$. Consider the rectangle $B = [a, b] \times [c, d]$; what is $\mathbb{P}((X, Y) \in B)$?*

We want to calculate $\mathbb{P}(a \le X \le b, c \le Y \le d)$. Well,

$$\{X \in [a, b], Y \in [c, d]\} = \{X \le b, Y \in [c, d]\} \cap \{X < a, Y \in [c, d]\}^c.$$

Therefore

$$\mathbb{P}((X, Y) \in B) = \mathbb{P}(X \le b, Y \in [c, d]) - \mathbb{P}(X < a, Y \in [c, d]).$$

We can then further break things up:

$$\begin{aligned}\{X \le b, Y \in [c, d]\} &= \{X \le b, Y \le d\} \cap \{X \le b, Y < c\}^c,\\ \{X < a, Y \in [c, d]\} &= \{X < a, Y \le d\} \cap \{X < a, Y < c\}^c.\end{aligned}$$

So

$$\begin{aligned}\mathbb{P}(X \le b, Y \in [c, d]) &= \mathbb{P}(X \le b, Y \le d) - \mathbb{P}(X \le b, Y < c)\\ \mathbb{P}(X < a, Y \in [c, d]) &= \mathbb{P}(X < a, Y \le d) - \mathbb{P}(X < a, Y < c).\end{aligned}$$

Now, since X, Y are continuous, all the inequalities $<$ and $\leq$ are interchangeable. So putting everything together, we find that

$$\mathbb{P}((X,Y) \in B) = [F_{X,Y}(b,d) - F_{X,Y}(b,c)] - [F_{X,Y}(a,d) - F_{X,Y}(a,c)]\,.$$

This looks very complicated. But lets look at this through the filter of the fundamental theorem of calculus:

$$F_{X,Y}(b,d) - F_{X,Y}(b,c) = \int_c^d \frac{\partial}{\partial y} F_{X,Y}(b,y)\,dy$$
$$F_{X,Y}(a,d) - F_{X,Y}(a,c) = \int_c^d \frac{\partial}{\partial y} F_{X,Y}(a,y)\,dy.$$

So we have

$$\begin{aligned}\mathbb{P}((X,Y) \in B) &= \int_c^d \frac{\partial}{\partial y} F_{X,Y}(b,y)\,dy - \int_c^d \frac{\partial}{\partial y} F_{X,Y}(a,y)\,dy \\ &= \int_c^d \frac{\partial}{\partial y}[F_{X,Y}(b,y) - F_{X,Y}(a,y)]\,dy.\end{aligned}$$

Now, we do it again:

$$F_{X,Y}(b,y) - F_{X,Y}(a,y) = \int_a^b \frac{\partial}{\partial x} F_{X,Y}(x,y)\,dx.$$

So finally, we come to the equation

$$\begin{aligned}\mathbb{P}((X,Y) \in B) &= \int_c^d \frac{\partial}{\partial y}\left[\int_a^b \frac{\partial}{\partial x} F_{X,Y}\,dx\right] dy \\ &= \int_c^d \int_a^b \frac{\partial}{\partial y}\frac{\partial}{\partial x} F_{X,Y}(x,y)\,dy\,dx \\ &= \int_B f_{X,Y}(x,y)\,dx\,dy.\end{aligned}$$

Example 20.3 shows us that we can calculate the probability that the *random vector* (X,Y) is in a rectangle by *integrating the joint density* $f_{X,Y}$ over that rectangle. The thing is, any (reasonable) two-dimensional solid can be approximated by a collection of very small rectangles (this is the whole point of double integrals). So we actually have the following theorem.

Theorem 20.4. *Let X, Y be random variables with a joint density $f_{X,Y}$, and let A be a region in $\mathbb{R}^2$. Then*

$$\mathbb{P}((X,Y) \in A) = \iint_A f_{X,Y}(x,y)\,dxdy.$$

From Theorem 20.4, we see that the joint density must be ≥ 0 (since probabilities are always ≥ 0), and it must be that

$$\iint_{\mathbb{R}^2} f_{X,Y}(x,y)\,dxdy = 1.$$

As with single random variable, the joint density is a very efficient way to record the information contained in the joint distribution.

Example 20.5. Suppose that X, Y have joint density

$$f_{X,Y}(x,y) = \begin{cases} \frac{1}{2}, & 1 \le x \le 2 \ \& \ -1 \le y \le 1 \\ 0, & \text{otherwise} \end{cases}.$$

This is a joint density function since $f_{X,Y} \ge 0$ and

$$\int_{\mathbb{R}^2} f_{X,Y}(x,y)\,dxdy = \int_{-1}^{1}\int_{1}^{2} \frac{1}{2}\,dxdy = \frac{1}{2}\cdot 2 \cdot 1 = 1.$$

Let D be the disk of radius $2/3$ centered at $(2,0)$. On the left-half of D, $f_{X,Y} = \frac{1}{2}$; on the right half, $f_{X,Y} = 0$. Thus

$$\mathbb{P}((X,Y) \in D) = \int_D f_{X,Y}(x,y)\,dxdy = \frac{1}{2}\cdot(\text{half the area of } D) = \frac{1}{2}\cdot\frac{1}{2}\pi\left(\frac{2}{3}\right)^2 = \frac{\pi}{9} \approx 34.9\%.$$

Example 20.6. The joint distribution of Example 20.2 is not a differentiable function – it has sharp ridges along (some of) the lines in the figure in that example. But corners are no problem for the existence of densities in one-variable, so let's differentiate anyhow. At all points not along the lines $x = y$, $x = 0$, $x = 1$, $y = 0$, or $y = 1$, the joint distribution function $F_{X,Y}$ is a *linear* function: its values are 0, 1, x, or y, depending on what region. But $\frac{\partial^2}{\partial x \partial y} 0 = \frac{\partial^2}{\partial x \partial y} 1 = \frac{\partial^2}{\partial x \partial y} x = \frac{\partial^2}{\partial x \partial y} y = 0$. Hence, if the pair $(X,Y) = (X,X)$ had a density $f_{X,X}$, that density would have to be 0 except on those lines. Since the area of those lines is 0, however, $\iint_{\mathbb{R}^2} f_{X,Y}(x,y)\,dxdy = 0$, not 1. In other words, **the pair (X,X) does not have a joint density**.

Remark 20.7. Example 20.6 shows that the existence of a *joint* density is harder to come by than just a one-variable density. For a single random variable X, the main obstacle to the existence of a density f_X is any discrete values for X: i.e. if $\mathbb{P}(X = x) > 0$ for any one value x. But even if both X and Y have densities (no discrete part), the pair (X,Y) can fail to have a density. The obstacle here is, as in Example 20.6, the set of possible values for the *pair* (X,Y) has to have positive area on the plane. For any random variable X, the pair (X,X) has values lying in the line $y = x$ which has 0 area; hence, there can be no joint density function.

20.3. **Independence.** Remember, from the beginning of the quarter, that we call two events A, B independent if

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

We translated this into a condition for independence of discrete *random variables* by insisting that, for any values x, y in the state spaces, the events $\{X = x\}$ and $\{Y = y\}$ are independent; i.e.

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

Now, if X, Y are *continuous* random variables, then this equation is always satisfies, since both sides are always 0. This is not very informative. So to properly define independence for continuous random variables, we need to go back to basics and look at some non-trivial events.

Definition 20.8. *Let X, Y be continuous random variables. Say that X, Y are* **independent** *if the events $\{X \leq x\}$ and $\{Y \leq y\}$ are independent for any values x, y in the state spaces. In other words, they are independent if*

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

In terms of cumulative distribution functions,

$$F_{X,Y}(x, y) = F_X(x)F_Y(y).$$

Differentiating this $\frac{\partial^2}{\partial x \partial y}$, the third and most useful way to state independence is

$$f_{X,Y}(x, y) = f_X(x)f_Y(y).$$

Example 20.9. Suppose that X is uniform on $[1, 2]$ and Y is uniform on $[-1, 1]$. This means that

$$f_X(x) = \begin{cases} 1, & 1 \leq x \leq 2 \\ 0, & \text{otherwise} \end{cases} \qquad f_Y(y) = \begin{cases} \frac{1}{2}, & -1 \leq y \leq 1 \\ 0, & \text{otherwise} \end{cases}$$

If X, Y are independent, then

$$f_{X,Y}(x, y) = f_X(x)f_Y(y) = \begin{cases} \frac{1}{2}, & 1 \leq x \leq 2 \ \& \ -1 \leq y \leq 1 \\ 0, & \text{otherwise} \end{cases}.$$

This is exactly the joint density we saw in Example 20.5, so in fact that was an example of a joint density of two independent random variables.

20.4. **Sums of Independent Random Variables.** In the *discrete* case, we had a nice concrete formula for determining the distribution of the sum of two random variables:

$$\mathbb{P}(X + Y = k) = \sum_x \mathbb{P}(X = x)\mathbb{P}(Y = k - x).$$

Since both sides are 0 in the continuous case, this doesn't make much sense for continuous random variables. But if we write things in terms of probability densities, there is a direct analogue.

Theorem 20.10. *Let X, Y be continuous random variables, with densities f_X and f_Y. If X and Y are independent, then the density of $X + Y$ is given by*

$$f_{X+Y}(t) = \int_{-\infty}^{\infty} f_X(x)f_Y(t - x)\, dx.$$

Proof. By definition,

$$f_{X+Y}(t) = \frac{d}{dt}F_{X+Y}(t) = \frac{d}{dt}\mathbb{P}(X + Y \leq t).$$

Now, the event $\{X + Y \leq t\}$ can be written as an event in terms of the random vector (X, Y):

$$\{X + Y \leq t\} = \{(X, Y) \in H\} \quad \text{where} \quad H = \{(x, y) \,:\, x + y \leq t\}.$$

Hence, using Theorem 20.4,

$$\mathbb{P}(X + Y \leq t) = \mathbb{P}((X, Y) \in H) = \iint_H f_{X,Y}(x, y)\, dxdy.$$

The region H is a slanted half-plane. We can do the double integral as an iterated-integral: for fixed x, we integrate y from $-\infty$ up to $t-x$, so

$$\mathbb{P}(X+Y\leq t)=\int_{-\infty}^{\infty}\left(\int_{-\infty}^{t-x}f_{X,Y}(x,y)dy\right)dx.$$

If we now differentiate both sides with respect to t, we find

$$\begin{aligned}f_{X+Y}(t)=\frac{d}{dt}\mathbb{P}(X+Y\leq t)&=\frac{d}{dt}\int_{-\infty}^{\infty}\left(\int_{-\infty}^{t-x}f_{X,Y}(x,y)dy\right)dx\\&=\int_{-\infty}^{\infty}\left(\frac{d}{dt}\int_{-\infty}^{t-x}f_{X,Y}(x,y)dy\right)dx\\&=\int_{-\infty}^{\infty}f_{X,Y}(x,t-x)\,dx.\end{aligned}$$

(The final equality is just the Fundamental Theorem of Calculus.) So far, everything we have written holds true for any continuous random variables. But we know that X,Y are independent, which means that $f_{X,Y}(x,t-x)=f_X(x)f_Y(t-x)$. This proves the result. □

Example 20.11. *Suppose X and Y are uniform on $[0,1]$. What is the probability that $X+Y\leq 1$?*

We will calculate the density of $X+Y$ using Theorem 20.10.

$$f_{X+Y}(t)=\int_{-\infty}^{\infty}f_X(x)f_Y(t-x)\,dx=\int_0^1 f_Y(t-x)\,dx$$

since $f_X(x)=1$ when $0\leq x\leq 1$ and $f_X=0$ outside the unit interval. Now we do a change of variables: $u=t-x$, so $du=-dx$. When $x=0$, $u=t-x=t$ and when $x=1$, $u=t-x=t-1$. So

$$f_{X+Y}(t)=\int_t^{t-1}f_Y(u)\,(-du)=\int_{t-1}^{t}f_Y(u)\,du.$$

Now, $f_Y(u)=1$ if $u\in[0,1]$ and is 0 otherwise. So to calculate this integral, we have to divide into cases depending on the value of t.

- If $t\leq 0$, then $t-1\leq 0$, and so $f_Y(u)=0$ on $[t-1,t]$. So in this case, $f_{X+Y}(t)=0$.
- If $0\leq t\leq 1$, then $t-1\leq 0$, and so $f_Y(u)=0$ for $u\in[t-1,0]$ and $f_Y(u)=1$ for $u\in[0,t]$; hence
$$f_{X+Y}(t)=\int_0^t du=t$$
for t in this range.
- If $1\leq t\leq 2$, then $t-1\in[0,1]$ but $t\geq 1$ so $f_Y(u)=1$ for $u\in[t-1,1]$ and $f_Y(u)=0$ for $u\in[1,t]$; hence
$$f_{X+Y}(t)=\int_{t-1}^1 du=1-(t-1)=2-t$$
for t in this range.
- If $t\geq 2$ then $t-1\geq 1$ and $t\geq 1$, so $f_Y(u)=0$ for all $u\in[t-1,t]$, and so $f_{X+Y}(t)=0$.

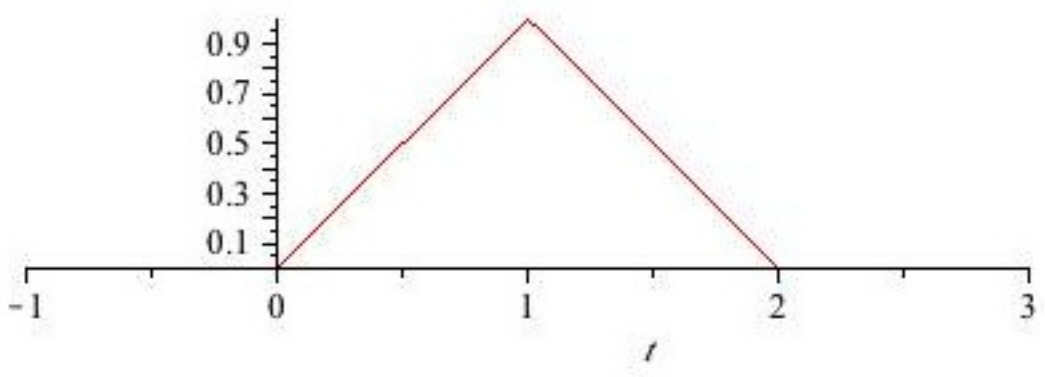


FIGURE 3. The graph of the density f_{X+Y} when X, Y are independent uniform $[0, 1]$ random variables.

Altogether, then, f_{X+Y} is a *tent-function*:

To answer the question at hand, we must evaluate

$$\mathbb{P}(X + Y \leq 1) = \int_{-\infty}^{1} f_{X+Y}(t)\, dt = \int_{0}^{1} t\, dt = \frac{1}{2},$$

as you might expect. On the other hand,

$$\mathbb{P}(X + Y \leq 1/2) = \int_{0}^{1/2} f_{X+Y}(t)\, dt = \int_{0}^{1/2} t\, dt = \frac{1}{8},$$

which might seem small; similarly,

$$\mathbb{P}(X + Y \leq 3/2) = \int_{0}^{3/2} f_{X+Y}(t)\, dt = \int_{0}^{1} t\, dt + \int_{1}^{3/2} (2 - t)\, dt = \frac{1}{2} + \left(2t - \frac{t^2}{2}\right|_{t=1}^{t=3/2} = \frac{7}{8}.$$

Example 20.12. We can use similar techniques to answer related questions. For example, if X and Y are independent, then how can we calculate the probability that $X - Y \leq a$? One approach would be to define $Z = -Y$, and note that

$$f_{X,Z}(x, z) = f_{X,-Y}(x, z) = f_{X,Y}(x, -z) = f_X(x) f_Y(-z) = f_X(x) f_{-Y}(z) = f_X(x) f_Z(z).$$

In other words, $X, -Z$ are also independent. So we can use Theorem 20.10:

$$\begin{aligned} f_{X-Y}(t) = f_{X+Z}(t) = \int_{-\infty}^{\infty} f_X(x) f_Z(t - x)\, dx &= \int_{-\infty}^{\infty} f_X(x) f_Y(-(t - x))\, dx \\ &= \int_{-\infty}^{\infty} f_X(x) f_Y(x - t)\, dx. \end{aligned}$$

If we know f_X and f_Y, we can (in principle) calculate this integral, and therefore we have the density of $X - Y$, which we can use to answer all probabilistic questions about $X - Y$. For example, if (as in Example 20.11) X and Y are both uniform on $[0, 1]$, an analogous

analysis here shows that we get another tent functions:

$$f_{X-Y}(t) = \begin{cases} 0, & t \leq -1 \text{ or } t \geq 1 \\ 1+t, & -1 \leq t \leq 0 \\ 1-t, & 0 \leq t \leq 1 \end{cases}$$

So, for example, $\mathbb{P}(X - Y \in [-0.1, 0.1])$ can be calculated by integrating this function:

$$\mathbb{P}(X - Y \in [-0.1, 0.1]) = \int_{-0.1}^{0.1} f_{X-Y}(t)\, dt = \int_{-0.1}^{0} (1+t)\, dt + \int_{0}^{0.1} (1-t)\, dt = 0.19.$$

21. LECTURE 21: NOVEMBER 12, 2010

21.1. **Adding independent normal random variables.**

Example 21.1. Suppose $X \sim N(0,1)$ and $Y \sim N(0,1)$. If X, Y are independent, we have

$$f_{X+Y}(u) = \int_{-\infty}^{\infty} f_X(x) f_Y(u-x)\, dx = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \frac{1}{\sqrt{2\pi}} e^{-(u-x)^2/2}\, dx$$
$$= \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-\frac{1}{2}(x^2+(u-x)^2)}\, dx.$$

We can simplify the quadratic polynomial in the exponent:

$$x^2 + (u-x)^2 = x^2 + x^2 - 2ux + u^2 = 2x^2 - 2ux + u^2.$$

There is a standard trick here: since we are integrating against x, we will *complete the square* with respect to x:

$$2x^2 - 2ux + u^2 = 2(x^2 - ux) + u^2 = 2(x^2 - ux + u^2/4 - u^2/4) + u^2 = 2(x - u/2)^2 - u^2/2.$$

Hence,

$$f_{X+Y}(u) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-\frac{1}{2}(2(x-u/2)^2 - u^2/2)}\, dx = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-(x-u/2)^2} e^{-u^2/4}\, dx.$$

Now, we can factor out the $e^{-u^2/4}$, since it is constant with respect to x:

$$f_{X+Y}(u) = e^{-u^2/4} \cdot \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-(x-u/2)^2}\, dx.$$

Now, in the integral, we make the change of variables $y = x - u/2$. Then $dy = dx$, and the interval of integration $(-\infty, \infty)$ *does not change*. Thus

$$f_{X+Y}(u) = e^{-u^2/4} \cdot \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-y^2}\, dy.$$

We can actually evaluate this integral, but there is no need: the integral $\frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-y^2}\, dy$ it is just some constant c, so we know that

$$f_{X+Y}(u) = ce^{-u^2/4}.$$

Since this *is* a probability density, we know that c is the unique constant which makes $\int_{-\infty}^{\infty} f_{X+Y}(u)\, du = 1$. We recognize that the function itself is Gaussian: it has the form $ce^{-u^2/2t}$ where $t = 2$; i.e. this is a normal $N(0,2)$ density, which means the constant must equal $\frac{1}{\sqrt{4\pi}}$.

This is a remarkable fact: the sum of two independent normal random variables is a normal random variable! In fact, more detailed calculations would have shown that if $X \sim N(0,t)$ and $Y \sim N(0,s)$, and if X, Y are independent, then $X + Y \sim N(0, t+s)$.

21.2. **Marginal Densities.** The joint cumulative distribution function $F_{X,Y}$ of a pair of random variables contains all the probabilistic information there is about X, Y. In particular, we must be able to recover the distributions of X and Y from $F_{X,Y}$. In fact, this is quite easy, since

$$F_X(x) = \mathbb{P}(X \leq x) = \mathbb{P}(X \leq x, Y < \infty) = \lim_{y\to\infty} F_{X,Y}(x, y).$$

Similarly,

$$F_Y(y) = \mathbb{P}(Y \leq y) = \mathbb{P}(X < \infty, Y \leq y) = \lim_{x\to\infty} F_{X,Y}(x, y).$$

Now, if X, Y have a joint density $f_{X,Y}$, then

$$F_{X,Y}(x, y) = \int_{-\infty}^{y}\int_{-\infty}^{x} f_{X,Y}(u, v)\, dudv.$$

Hence

$$F_X(a) = \lim_{y\to\infty}\int_{-\infty}^{y}\int_{-\infty}^{x} f_{X,Y}(u, v)\, dudv = \int_{-\infty}^{\infty}\int_{-\infty}^{x} f_{X,Y}(u, v)\, dudv.$$

If we differentiate with respect to x, we get an expression for the density of X:

$$f_X(x) = \frac{d}{dx}\int_{-\infty}^{\infty}\int_{-\infty}^{x} f_{X,Y}(u, v)\, dudv = \int_{-\infty}^{\infty} f_{X,Y}(x, v)\, dv.$$

We can similarly compute that

$$f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(u, y)\, du.$$

These are the **marginal densities**. They are akin to the marginal distribution functions in the discrete setting that we discussed in Lecture 14: they are the row and column "sums" (actually integrals here).

Example 21.2. Suppose that the random vector (X, Y) is uniformly distributed on the unit ball:

$$f_{X,Y}(x, y) = \begin{cases} \frac{1}{\pi}, & x^2 + y^2 \leq 1 \\ 0, & x^2 + y^2 > 1 \end{cases}$$

Let's find the densities of X and Y. We must compute the marginals.

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

For fixed x, the function $f_{X,Y}(x, y)$ is only non-zero when $y^2 \leq 1 - x^2$, which means $-\sqrt{1-x^2} \leq y \leq \sqrt{1-x^2}$. This set is empty unless $|x| \leq 1$, so $f_X(x) = 0$ if $|x| > 1$. For $|x| \leq 1$, the integral becomes

$$\int_{-\infty}^{\infty} f_{X,Y}(x, y)\, dy = \int_{-\sqrt{1-x^2}}^{\sqrt{1-x^2}} \frac{1}{\pi}\, dy = \frac{2}{\pi}\sqrt{1-x^2}.$$

Thus, we have found that

$$f_X(x) = \begin{cases} \frac{2}{\pi}\sqrt{1-x^2}, & |x| \leq 1 \\ 0, & |x| > 1 \end{cases}$$

This is called the **semicircle law**. (The graph is not strictly a semicircle, but rather a semi-ellipse, since the factor $\frac{2}{\pi}$ is requires to normalize the integral.) As for f_Y, we could calculate it similarly, but we can also just note that, in this example,

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

That is, the two functions $f_{X,Y}$ and $f_{Y,X}$ are equal. When this happens, it means that X and Y must have the same distribution, so $f_X = f_Y$.

Example 21.3. Suppose X, Y are independent random variables, with joint density $f_{X,Y}$. In this case, we know that $f_{X,Y}(x,y) = f_X(x)f_Y(y)$, so it is easier to pick out the separate densities of X, Y. Let's see what happens when we compute the marginals:

$$\int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dy = \int_{-\infty}^{\infty} f_X(x)f_Y(y)\,dy = f_X(x)\int_{-\infty}^{\infty} f_Y(y)\,dy = f_X(x)\cdot 1,$$

as required.

We can use marginals to give an easy way to check whether a given function $f(x,y)$ is the joint density of a pair of independent random variables. (This is the analogue of Proposition 14.6 for discrete distributions.)

Proposition 21.4. *Suppose $f_{X,Y}$ is a non-negative function of two variables with $\iint_{\mathbb{R}^2} f(x,y)\,dxdy = 1$. Suppose there are one-variable functions g, h such that $f(x,y) = g(x)h(y)$. Then $f_{X,Y}$ is the joint density of a pair of random variables that are* independent, *and there is a constant $c > 0$ so that $f_X(x) = cg(x)$ and $f_Y(y) = \frac{1}{c}h(y)$.*

Proof. We simply compute the marginals of f:

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dy = \int_{-\infty}^{\infty} g(x)h(y)\,dy = g(x)\cdot\int_{-\infty}^{\infty} h(y)\,dy = g(x)\cdot c$$

where $c = \int_{\mathbb{R}} h(y)\,dy$. Similarly,

$$g_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dx = \int_{-\infty}^{\infty} g(x)h(y)\,dx = h(y)\cdot\int_{-\infty}^{\infty} g(x)\,dx = h(y)\cdot c'$$

where $c' = \int_{\mathbb{R}} g(x)\,dx$. Finally, we just note that

$$1 = \int_{\mathbb{R}^2} f_{X,Y}(x,y)\,dxdy = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x)h(y)\,dxdy = \left(\int_{-\infty}^{\infty} g(x)\,dx\right)\left(\int_{-\infty}^{\infty} h(y)\,dy\right) = cc',$$

so $c' = \frac{1}{c}$ as claimed. Thus, $f_{X,Y}(x,y) = g(x)h(y) = \frac{1}{c}f_X(x)\cdot xf_Y(y) = f_X(x)f_Y(y)$, so X, Y are independent. □

Example 21.5. Suppose that $f_{X,Y}(x,y) = \frac{1}{\pi}e^{-(x^2+y^2)}$. Notice that

$$f_{X,Y}(x,y) = \frac{1}{\pi}e^{-x^2}\cdot e^{-y^2}$$

is a product of two one-variable functions. Thus, X, Y are independent. Moreover, their distributions are both (constants) times e^{-x^2}. We can even tell quickly that, since both distributions must be the same, the constant must equally distribute the $\frac{1}{\pi}$, so $f_X(x) = \frac{1}{\sqrt{\pi}}e^{-x^2} = f_Y(x)$.

Example 21.6. *Suppose that*

$$f_{X,Y}(x,y) = \begin{cases} e^{-x}, & x > 0 \ \ \& \ \ 0 < y < 1 \\ 0, & \textit{otherwise} \end{cases}$$

Are X, Y independent?

The domain where $f_{X,Y} \neq 0$ is a rectangle, so there is a change $f_{X,Y}$ is a product. Indeed, we see that $f_{X,Y}(x,y) = g(x)h(y)$, where

$$g(x) = \begin{cases} e^{-x}, & x > 0 \\ 0, & x \leq 0 \end{cases} \qquad h(y) = \begin{cases} 1, & 0 < y < 1 \\ 0, & \text{otherwise} \end{cases}$$

So $f_{X,Y}$ is the joint density of independent random variables, where X is $\text{exponential}(1)$ and Y is uniform on $[0,1]$.

21.3. **Conditional Densities.** If X, Y are discrete random variables then the event $\mathbb{P}(Y = y) > 0$ for each y in the state space of Y. So it makes sense to condition on this event:

$$\mathbb{P}(X \in [a,b] | Y = y) = \frac{\mathbb{P}(X \in [a,b], Y = y)}{\mathbb{P}(Y = y)}.$$

But when X, Y are continuous random variables, the denominator is 0. However, the numerator is *also* 0, so $\mathbb{P}(X \in [a,b] | Y = y)$ has a shot of making sense (since it is a $\frac{0}{0}$ indeterminate form). To see how we should make sense of it, let's look at a conditional probability that does make sense: suppose that f_Y is not 0 on the interval $[y, y + \Delta y]$ (so that $\mathbb{P}(Y \in [y, y + \Delta y]) > 0$). Then

$$\mathbb{P}(X \in [a,b] | Y \in [y, y + \Delta y]) = \frac{\mathbb{P}(X \in [a,b], Y \in [y, y + \Delta y])}{\mathbb{P}(Y \in [y, y + \Delta y])}.$$

The numerator is

$$\mathbb{P}(X \in [a,b], Y \in [y, y + \Delta y]) = \int_y^{y + \Delta y} \int_a^b f_{X,Y}(x,v)\, dx dv.$$

The function $v \mapsto \int_a^b f_{X,Y}(x,v)\, dx$ is continuous, and so for sufficiently small Δy, we have

$$\int_a^b f_{X,Y}(x,v)\, dx \approx \int_a^b f_{X,Y}(x,y)\, dx, \quad y \leq v \leq y + \Delta y.$$

So when we integrate, the numerator is close to

$$\mathbb{P}(X \in [a,b], Y \in [y, y + \Delta y]) \approx \int_y^{y + \Delta y} \int_a^b f_{X,Y}(x,y)\, dx dv = \int_a^b f_{X,Y}(x,y)\, dx \cdot \Delta y.$$

Similarly, the denominator is close to

$$\mathbb{P}(Y \in [y, y + \Delta y]) = \int_y^{y + \Delta y} f_Y(v)\, dv \approx f_Y(y) \cdot \Delta y.$$

So

$$\mathbb{P}(X \in [a,b] | Y \in [y, y + \Delta y]) \approx \frac{\int_a^b f_{X,Y}(x,y)\, dx \cdot \Delta y}{f_Y(y) \cdot \Delta y}.$$

Both the numerator and denominator tend to 0 as $\Delta y \to 0$, but they cancel and we see that

$$\lim_{\Delta y \to 0} \mathbb{P}(X \in [a,b] | Y \in [y, y+\Delta y]) = \frac{\int_a^b f_{X,Y}(x,y)\,dx}{f_Y(y)} = \int_a^b \frac{f_{X,Y}(x,y)}{f_Y(y)}\,dx.$$

So this motivates our definition of **conditional density**.

Definition 21.7. *Let X, Y be jointly continuous random variables, with joint density $f_{X,Y}$. The* **conditional density** *of X, Y given that $Y = y$ is the function*

$$f_{X|Y}(x,y) = \frac{f_{X,Y}(x,y)}{f_Y(y)} = \frac{f_{X,Y}(x,y)}{\int_{\mathbb{R}} f(x,y)\,dx}.$$

This function is defined for all (x,y) where $f_Y(y) > 0$. (If $f_Y(y) = 0$, we just set $f_{X|Y}(x,y) = 0$.)

The meaning of the conditional density is made clear from the derivation we gave for its expression: for $a < b$ and y,

$$\mathbb{P}(X \in [a,b] | Y = y) = \int_a^b f_{X|Y}(x,y)\,dx.$$

That is: even though the event $\{Y = y\}$ has probability 0, conditioning on this event makes sense provided the *density* of probabilities near this point is positive. Recall: if X, Y are discrete, the conditional distribution was

$$\mu_{X|Y}(x,y) = \frac{\mathbb{P}(X = x, Y = y)}{\mathbb{P}(Y = y)} = \frac{\mu_{X,Y}(x,y)}{\mu_Y(y)}.$$

The conditional density is the infinitesimal form of this for continuous random variables.

Example 21.8. As in Example 21.2, let (X, Y) be a uniform random vector in the unit disk. Let's calculate the conditional density $f_{X|Y}(x,y)$. As calculated above,

$$f_Y(y) = \begin{cases} \frac{2}{\pi}\sqrt{1-y^2}, & |y| \le 1 \\ 0, & |y| > 1 \end{cases}$$

So first, automatically, when $|y| > 1$ we must have $f_{X|Y}(x,y) = 0$. When $|y| \le 1$, we have

$$f_{X|Y}(x,y) = \frac{f_{X,Y}(x,y)}{f_Y(y)} = \begin{cases} \frac{1/\pi}{2/\pi\sqrt{1-y^2}}, & x^2 + y^2 \le 1 \\ 0, & x^2 + y^2 > 1 \end{cases}$$

To be clear: we are thinking of y as fixed here, so we should write this as:

$$f_{X|Y}(x,y) = \begin{cases} \frac{1}{2\sqrt{1-y^2}}, & |x| \le \sqrt{1-y^2} \\ 0, & |x| > \sqrt{1-y^2} \end{cases}$$

when $|y| \le 1$, and $f_{X|Y}(x,y) = 0$ when $|y| > 1$. Hence, the conditional distribution of X given that $Y = y$ is uniform on the interval $[-\sqrt{1-y^2}, \sqrt{1-y^2}]$.

Remark 21.9. Comparing Examples 21.2 and 21.8, we see the following properties of the uniform distribution on the unit disk.

- If we choose many samples of (X, Y) uniform on the disk, and then look at the distribution of their first coordinate (i.e. we project all the points down onto the x-axis), the resulting distribution of points will *not* be uniform: the histogram has a semicircular shape. This is because there are more points along each vertical line to project near the centre of the disk than near the outside.
- On the other hand, if we look at a narrow horizontal strip at height y and look at the x-coordinates of points along this line, stretching from $(-\sqrt{1-y^2}, y)$ to $(\sqrt{1-y^2}, y)$, these points *are* distributed uniformly.

Example 21.10. *Let X be uniform on $[0, 1]$. Let Y be uniform on $[0, X]$. (That is: once we pick a random number $X = x$ in $[0, 1]$, we pick a uniform random number in $[0, x]$.) What is the joint density of (X, Y)?*

First, as stated, we have the density of X and the conditional density of Y given X:

$$f_X(x) = \begin{cases} 1, & 0 < x < 1 \\ 0, & \text{otherwise} \end{cases}$$

$$f_{Y|X}(y, x) = \begin{cases} \frac{1}{x}, & 0 < y < x \\ 0, & \text{otherwise} \end{cases}$$

Now, by definition

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

and so we can recover the joint density by multiplying:

$$f_{Y,X}(y, x) = f_{Y|X}(y, x) f_X(x) = \begin{cases} \frac{1}{x}, & 0 < y < x < 1 \\ 0, & \text{otherwise} \end{cases}$$

This density is 0 except on the lower-triangular part of the unit square. Note that $f_{Y,X}(y, x) = f_{X,Y}(x, y)$, so we have the desired joint density. Just as a sanity check, let's verify it actually is a probability density. It is, of course, ≥ 0, and

$$\iint_{\mathbb{R}^2} f_{X,Y}(x, y)\, dxdy = \int_0^1 \left(\int_0^x \frac{1}{x} dy \right) dx = \int_0^1 \frac{1}{x}(x - 0)\, dx = 1.$$

Now, knowing the joint distribution of X, Y, we know everything. For example, we can calculate the unconditional distribution of Y. This is the marginal:

$$f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x, y)\, dx = \int_y^1 \frac{1}{x}\, dx = \ln x|_{x=y}^{x=1} = \ln 1 - \ln y = \ln \frac{1}{y}$$

when $0 < y \leq 1$, and it's 0 outside the unit interval. This density gets very large as $y \downarrow 0$. Again, as a sanity check, let's verify that it integrates to 1:

$$\int_{-\infty}^{\infty} f_Y(y)\, dy = \int_0^1 -\ln y\, dy = -\left(y \ln y - y|_{y=0}^{y=1} = 1,\right.$$

where we have used the fact that $\lim_{y \to 0+} y \ln y = 0$ (which can be verified, for example, with l'Hôpital's rule). So Y is most definitely *not* uniform; it is much more likely to be

close to 0 than close to 1. This makes sense: we choose X first, and then choose $Y \leq X$, so Y has fewer chances to be large than to be small.

21.4. **Back to Expectation.** We can use joint densities to study expectations of sums of jointly-continuous random variables.

Theorem 21.11. *Let X, Y be random variables with a joint density function $f_{X,Y}$. If $\mathbb{E}(|X|) < \infty$ and $\mathbb{E}(|Y|) < \infty$, then $X + Y$ has an expectation, and*

$$\mathbb{E}(X + Y) = \mathbb{E}(X) + \mathbb{E}(Y).$$

Remark 21.12. Actually, this theorem holds for *any* random variables, whether they have a joint density or not. We have already seen this in the discrete setting. In Example 20.6, we saw that the pair (X, X) never has a joint density; however, $\mathbb{E}(X + X) = \mathbb{E}(2X) = 2\mathbb{E}(X) = \mathbb{E}(X) + \mathbb{E}(X)$ from Example 19.6. We only have the tools to prove this theorem in the jointly-continuous case, however; to prove it in general would require the more sophisticated machinery of measure theory. (On the other hand, we could also prove it by approximating with discrete distributions and using the theorem from the discrete world; this would work, but would be quite messy to sort out all the details.)

Proof. By definition

$$\mathbb{E}(X + Y) = \int_{-\infty}^{\infty} t f_{X+Y}(t)\, dt.$$

We are *not* assuming X, Y are independent, so we cannot express $f_{X+Y}(t)$ as in Theorem 20.10. We can, however, follow some of the proof there. In general, we have

$$F_{X+Y}(t) = \mathbb{P}(X + Y \leq t) = \mathbb{P}((X, Y) \in H_t)$$

where $H_t = \{(x, y) : x + y \leq t\}$. So by the definition of the joint density $f_{X,Y}$,

$$F_{X+Y}(t) = \iint_{H_t} f_{X,Y}(x, y)\, dxdy.$$

We can evaluate this integral as an iterated integral. The half-plane H_t is the set of points below-and-to-the-left of the line $y = t - x$, so

$$F_{X+Y}(t) = \int_{-\infty}^{\infty} \left(\int_{-\infty}^{t-x} f_{X,Y}(x, y)\, dy \right) dx.$$

Differentiating with respect to t gives us

$$f_{X+Y}(t) = \int_{-\infty}^{\infty} \left(\frac{d}{dt} \int_{-\infty}^{t-x} f_{X,Y}(x, y)\, dy \right) dx = \int_{-\infty}^{\infty} f_{X,Y}(x, t - x)\, dx.$$

This is *almost* the formula from Theorem 20.10 for the density of a sum of independent random variables; but this formula actually holds in general. (Independence would give us the one final step that $f_{X,Y}(x, t - x) = f_X(x) f_Y(t - x)$, which is not true in general.)

Combining this with the definition of expectation gives us

$$\mathbb{E}(X + Y) = \int_{-\infty}^{\infty} t \left(\int_{-\infty}^{\infty} f_{X,Y}(x, t - x)\, dx \right) dt.$$

Now, we reverse the order of integration. (This is not always legal to do when evaluating improper integrals; but the assumptions that $\mathbb{E}(|X|) = \int_{-\infty}^{\infty} x f_X(x)\,dx < \infty$ and $\mathbb{E}(|Y|) = \int_{-\infty}^{\infty} y f_Y(y)\,dy < \infty$ actually make it legal. We don't work out the details of this application of *Fubini's theorem* here.) So

$$\mathbb{E}(X+Y) = \int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} t f_{X,Y}(x, t-x)\,dt \right) dx.$$

Now we do something funny: we write $t = t - x + x$. The inside integral then becomes

$$\int_{-\infty}^{\infty} (t-x+x) f_{X,Y}(x,t-x)\,dt = \int_{-\infty}^{\infty} (t-x) f_{X,Y}(x,t-x)\,dt + \int_{-\infty}^{\infty} x f_{X,Y}(x,t-x)\,dt.$$

We now make the change of variables $u = t - x$ in both integrals. This does not change the domain of integration, and $du = dt$, so we get

$$\int_{-\infty}^{\infty} u f_{X,Y}(x,u)\,du + x \int_{-\infty}^{\infty} f_{X,Y}(x,u)\,du.$$

Plugging this back into the double integral for $\mathbb{E}(X+Y)$ we get

$$\mathbb{E}(X+Y) = \int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} u f_{X,Y}(x,u)\,du \right) dx + \int_{-\infty}^{\infty} \left(x \int_{-\infty}^{\infty} f_{X,Y}(x,u)\,du \right) dx.$$

Once more, we reverse the order of integration, in the first integral only, to get

$$\mathbb{E}(X+Y) = \int_{-\infty}^{\infty} u \left(\int_{-\infty}^{\infty} f_{X,Y}(x,u)\,dx \right) du + \int_{-\infty}^{\infty} x \left(\int_{-\infty}^{\infty} f_{X,Y}(x,u)\,du \right) dx.$$

We recognize the two inside integrals as the *marginal densities*:

$$\int_{-\infty}^{\infty} f_{X,Y}(x,u)\,dx = f_Y(u)$$
$$\int_{-\infty}^{\infty} f_{X,Y}(x,u)\,du = f_X(x).$$

So finally we have

$$\mathbb{E}(X+Y) = \int_{-\infty}^{\infty} u f_Y(u)\,du + \int_{-\infty}^{\infty} x f_X(x)\,dx = \mathbb{E}(X) + \mathbb{E}(Y).$$

□

22. Lecture 22: November 17, 2010

22.1. **Probability vs. Expectation.** Let A be a subset of the real numbers $\mathbb{R}$. Remember the function

$$\mathbb{1}_A(y) = \begin{cases} 1, & y \in A \\ 0, & y \notin A \end{cases}$$

Let X be a random variable (say with a density f_X). Then consider the random variable $\mathbb{1}_A(X)$.

$$\mathbb{E}(\mathbb{1}_A(X)) = \int_{-\infty}^{\infty} \mathbb{1}_A(x) f_X(x)\, dx = \int_A f_X(x)\, dx = \mathbb{P}(X \in A).$$

We saw this relation in the discrete world, too. So we can use this to translate between expectations and probability. In particular,

$$\mathbb{E}(\mathbb{1}_{(-\infty,x]}(X)) = \mathbb{P}(X \in (-\infty, x]) = \mathbb{P}(X \leq x) = F_X(x).$$

So the cumulative distribution function F_X that we've been working with can be expressed as an (x-dependent) expectation. As we've seen, F_X (and its "derivative" f_X) contain all the information about the distribution of X.

We could consider lots of other x-dependent expectations of a random variable, and see what kind of information they contain.

Definition 22.1. *Let X be a random variable. The* **moment-generating function** *M_X of X is the function*

$$M_X(t) = \mathbb{E}(e^{tX}).$$

Note: this function may be infinite for some t, or may not exist at all for some t, depending on the distribution of X.

Example 22.2. Let X be uniform on $[0, 1]$. Then

$$M_X(t) = \mathbb{E}(e^{tX}) = \int_0^1 e^{tx}\, dx = \frac{1}{t} e^{tx}\Big|_{x=0}^{x=1} = \frac{e^t - 1}{t}.$$

This function is well-defined and finite everywhere, except possibly at $t = 0$ where the formula doesn't make sense. But actually,

$$\lim_{t \to 0} \frac{e^t - 1}{t} = \frac{d}{dt} e^t \Big|_{t=0} = 1$$

and so we can make sense of the formula even there. And actually, this matches up the value of the function:

$$M_X(0) = \mathbb{E}(e^{0X}) = \mathbb{E}(1) = 1.$$

In fact, we see from this that $M_X(0) = 1$ for *all* random variables X.

Example 22.3. Let X have an $\text{exponential}(\lambda)$ distribution. Then

$$M_X(t) = \int_{-\infty}^{\infty} e^{tx} f_X(x)\, dx = \int_0^{\infty} e^{tx} \lambda e^{-\lambda x}\, dx = \lambda \int_0^{\infty} e^{-(\lambda - t)x}\, dx.$$

Of course, this integral only converges if $\lambda - t > 0$. In this domain, we get

$$M_X(t) = \frac{\lambda}{-(\lambda - t)} e^{-(\lambda - t)} \Big|_{x=0}^{x=\infty} = \frac{\lambda}{-(\lambda - t)}(-1) = \frac{\lambda}{\lambda - t}.$$

In other words (simplifying a little), we have

$$M_X(t) = \begin{cases} \dfrac{1}{1-t/\lambda}, & -\infty < t < \lambda \\ \infty, & t \geq \lambda \end{cases}$$

Example 22.4. Let $X \sim N(0, \sigma^2)$. Then

$$M_X(t) = \int_{-\infty}^{\infty} e^{tx} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-x^2/2\sigma^2}\, dx = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} e^{-x^2/2\sigma^2 + tx} dx.$$

We can evaluate this integral by making a change of variables. The idea is to complete the square:

$$\begin{aligned} -\frac{x^2}{2\sigma^2} + tx = -\frac{1}{2\sigma^2}(x^2 - 2t\sigma^2 x) &= -\frac{1}{2\sigma^2}(x^2 - 2t\sigma^2 x + (t\sigma^2)^2 - (t\sigma^2)^2) \\ &= -\frac{1}{2\sigma^2}(x - t\sigma^2)^2 + \frac{1}{2}t^2\sigma^2. \end{aligned}$$

Hence

$$M_X(t) = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} e^{-\frac{1}{2\sigma^2}(x-t\sigma^2)^2 + \frac{1}{2}t^2\sigma^2}\, dx = e^{\frac{1}{2}\sigma^2 t^2} \cdot \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} e^{-\frac{1}{2\sigma^2}(x-t\sigma^2)^2}\, dx.$$

Now, for the integral, we can make the change of variables $u = x - t\sigma^2$. Since t is a constant with respect to x, this means that $du = dx$, and also the domain $(-\infty, \infty)$ does not change, so

$$\frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} e^{-\frac{1}{2\sigma^2}(x-t\sigma^2)^2}\, dx = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} e^{-u^2/2\sigma^2}\, du = 1$$

because this is exactly the $N(0, \sigma^2)$ density. Thus,

$$M_X(t) = e^{\frac{1}{2}\sigma^2 t^2}.$$

Example 22.5. If X has a power law, $f_X(x) = (\rho - 1)x^{-\rho}$ for $x \geq 1$, where $\rho > 1$, then

$$M_X(t) = (\rho - 1) \int_1^{\infty} e^{tx} x^{-\rho}\, dx.$$

The function $e^{tx}x^{-\rho}$ does not have an elementary anti-derivative, so we cannot write a formula here and take limits. But we can actually easily see that, for any $t > 0$ and any $\rho > 1$, the function $e^{tx}x^{-\rho} \to \infty$ as $x \to \infty$, so there is no hope this integral con converge. So we have

$$M_X(t) = \infty \quad \text{when} \quad t > 0.$$

For $t = 0$ we get $M_X(0) = 1$ as usual. When $t < 0$, the integral does converge, though it cannot be written in simple terms.

There is no need to restrict ourselves to the continuous world.

Example 22.6. Let X represent a fair coin toss (a Bernoulli random variable), with $\mathbb{P}(X = 1) = \frac{1}{2}$ and $\mathbb{P}(X = 0) = \frac{1}{2}$. Then

$$M_X(t) = \mathbb{E}(e^{tX}) = e^{t\cdot 1}\mathbb{P}(X = 1) + e^{t\cdot 0}\mathbb{P}(X = 0) = \frac{1}{2}(1 + e^t).$$

If we had instead decided to let -1 represent tails, so $\mathbb{P}(Y = 1) = \mathbb{P}(Y = -1) = \frac{1}{2}$, then

$$M_Y(t) = \mathbb{E}(e^{tY}) = \frac{1}{2}e^{t\cdot 1} + \frac{1}{2}e^{t\cdot(-1)} = \frac{e^t + e^{-t}}{2} = \cosh t.$$

Example 22.7. Suppose X is a $\mathrm{Poisson}(\lambda)$ random variable. Then

$$\begin{aligned} M_X(t) = \mathbb{E}(e^{tX}) &= \sum_{n=0}^{\infty} e^{tn}\mathbb{P}(X = n) \\ &= \sum_{n=0}^{\infty} e^{tn}e^{-\lambda}\frac{\lambda^n}{n!} = e^{-\lambda}\sum_{n=0}^{\infty}\frac{(e^t\lambda)^n}{n!} = e^{-\lambda}\cdot e^{e^t\lambda} = e^{\lambda(e^t-1)}. \end{aligned}$$

Example 22.8. Let X be $\mathrm{binomial}(n, p)$. Then

$$M_X(t) = \sum_{k=0}^{n} e^{tk}\mathbb{P}(X = k) = \sum_{k=0}^{n} e^{tk}\binom{n}{k}p^k(1-p)^{n-k}.$$

To evaluate this sum, the trick is to combine e^{tk} with p^k to give $(e^tp)^k$:

$$M_X(t) = \sum_{k=0}^{n}\binom{n}{k}(e^tp)^k(1-p)^{n-k}.$$

The binomial theorem then tells us that

$$M_X(t) = (e^tp + (1-p))^n.$$

22.2. **Why 'moment-generating'?** We've seen how useful the cumulative distribution function and density function can be in calculating probabilities associated to random variables. The moment-generating function isn't as good for such calculations, but there's one things it's *really* good at.

Proposition 22.9. *Let X be a random variable, and suppose that the function $M_X(t) = \mathbb{E}(e^{tX})$ is differentiable in a neighbourhood of $t = 0$. Then for positive integers n*

$$\left.\frac{d^n}{dt^n}M_X(t)\right|_{t=0} = \mathbb{E}(X^n).$$

The quantities $\mathbb{E}(X^n)$ are called the **moments** of X. Of particular interest are, of course, the $n = 1$ case (the expectation) and the $n = 2$ case (which we can use, with the expectation, to compute the variance).

Proof. The idea is very simple:

$$\frac{d^n}{dt^n}\mathbb{E}(e^{tX}) = \mathbb{E}\left(\frac{d^n}{dt^n}e^{tX}\right).$$

Once we pass the derivatives inside, we are differentiating the function $t \mapsto e^{Xt}$ where X is a constant; this nth derivative is $\frac{d^n}{dt^n}e^{Xt} = X^n e^{Xt}$. Hence

$$\frac{d^n}{dt^n}\mathbb{E}(e^{tX}) = \mathbb{E}(X^n e^{tX}).$$

Setting $t = 0$ gives the result. □

Remark 22.10. The only technical point here is the question of passing the derivative through the expectation. If X has a density, this is the question of whether

$$\frac{d}{dt}\int_{-\infty}^{\infty} e^{tx} f_X(x)\,dx = \int_{-\infty}^{\infty} \frac{d}{dx}e^{tx} f_X(x)\,dx.$$

Interchanging integrals and derivatives (with respect to different variables) is usually legal, but it is an interchange of limits and needs some justification. The assumption of the proposition, that M_X is differentiable in a neighbourhood of 0, is actually strong enough to imply this interchange is valid; if you are interested in the details, you should consider taking Math 140A/B/C.

Remark 22.11. Another way to prove the proposition is to expand e^{tX} in a power series:

$$e^{tX} = \sum_{n=0}^{\infty} \frac{(tX)^n}{n!}.$$

Then using the linearity of $\mathbb{E}$, we have

$$M_X(t) = \mathbb{E}(e^{tX}) = \sum_{n=0}^{\infty} \frac{\mathbb{E}((tX)^n}{n!} = \sum_{n=0}^{\infty} \frac{1}{n!}\mathbb{E}(X^n)t^n.$$

(Note: we have again interchanged limits, since we only really know $\mathbb{E}$ is linear over *finite* sums. So this step also needs technical justification.) For any convergent power series $g(t) = \sum_{n=0}^{\infty} a_n t^n$, it is well-known (from Taylor's theorem) that the nth derivative at 0 is $g^{(n)}(0) = n!a_n$. In this case, $a_n = \mathbb{E}(X^n)/n!$, so this gives the result.

Example 22.12. Let X be a standard normal random variable $N(0,1)$. As calculated in Example 22.4,

$$M_X(t) = e^{\frac{1}{2}t^2}.$$

So we can calculate all the moment of X by repeatedly differentiating. Actually, we can do this faster by expanding in a power series:

$$M_X(t) = \sum_{n=0}^{\infty} \frac{1}{n!}\left(\frac{1}{2}t^2\right)^n = \sum_{n=0}^{\infty} \frac{1}{2^n n!}t^{2n}.$$

This power series has only even powers of t in it, and so we have $\mathbb{E}(X^n) = 0$ when n is odd. When n is even, we can use the trick from Remark 22.11:

$$\left.\frac{d^{2n}}{dt^{2n}}M_X(t)\right|_{t=0} = (2n)! \cdot \frac{1}{2^n n!} = \frac{(2n)!}{2^n n!}.$$

We can simplify this a little:

$$\frac{(2n)!}{2^n n!} = \frac{2n(2n-1)(2n-2)(2n-3)\cdots 4\cdot 3\cdot 2\cdot 1}{2n\cdot 2(n-1)\cdot 2(n-2)\cdots 2(2)\cdot 2(1)} = (2n-1)(2n-3)\cdots 3\cdot 1.$$

This product is sometimes referred to as a *double-factorial*, $\mathbb{E}(X^{2n}) = (2n-1)!!$

In particular, we can quickly list off that $\mathbb{E}(X) = 0$ and $\mathbb{E}(X^2) = 1$, so $\mathrm{Var}X = 1$.

23. LECTURE 23: NOVEMBER 19, 2010

Example 23.1. Let X be $\text{exponential}(\lambda)$ with $\lambda > 0$. Then $M_X(t) = \frac{\lambda}{\lambda - t}$ is differentiable near 0, and so we can differentiate

$$M_X'(t) = -\frac{\lambda}{(\lambda - t)^2} \cdot (-1) = \frac{\lambda}{(\lambda - t)^2}.$$

So $M'(0) = \frac{\lambda}{\lambda^2} = \frac{1}{\lambda}$, as we've already calculated. Similarly,

$$M_X''(t) = \frac{d}{dt}\frac{\lambda}{(\lambda - t)^2} = \frac{-2\lambda}{(\lambda - t)^3} \cdot (-1) = \frac{2\lambda}{(\lambda - t)^3}.$$

So $M''(0) = \frac{2}{\lambda^2}$, and so $\text{Var}(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2 = M_X''(0) - M_X'(0)^2 = \frac{2}{\lambda^2} - \frac{1}{\lambda^2} = \frac{1}{\lambda^2}$.

We can calculate all the higher moments simultaneously with a power-series approach:

$$M_X(t) = \frac{1}{1 - t/\lambda} = \sum_{n=0}^{\infty} (t/\lambda)^n = \sum_{n=0}^{\infty} \frac{1}{\lambda^n} t^n.$$

Hence, the nth derivative is $n!$ times the coefficient of t^n, and so

$$\mathbb{E}(X^n) = \frac{n!}{\lambda^n}.$$

Example 23.2. Let X be $\text{binomial}(n, p)$. Calculating moments directly involves doing some very tricky binomial sums. But, as we calculated in Example 22.8,

$$M_X(t) = (e^t p + (1 - p))^n$$

so we can quickly calculate moments by differentiating.

$$M_X'(t) = n(e^t p + (1 - p))^{n-1} e^t p,$$

so $\mathbb{E}(X) = M_X'(0) = n(e^0 p + 1 - p)^{n-1} e^0 p = np$. Now, using the product rule,

$$M_X''(t) = n(n-1)(e^t p + (1 - p))^{n-1}(e^t p)^2 + n(e^t p + (1 - p))^{n-1} e^t p$$

so $\mathbb{E}(X^2) = M_X''(0) = n(n-1)p^2 + np$. So we can calculate

$$\text{Var}X = \mathbb{E}(X^2) - \mathbb{E}(X)^2 = n(n-1)p^2 + np - (np)^2 = -np^2 + np = np(1 - p).$$

23.1. Independence.

Suppose X and Y are independent random variables. Remember, this means that for any $x, y \in \mathbb{R}$,

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

Now, fix $t \in \mathbb{R}$, and consider the new random variables e^{tX} and e^{tY}. Except at $t = 0$, for any $a, b > 0$ we can express the event

$$\{e^{tX} \le a\} = \{tX \le \ln a\} = \{X \le \frac{1}{t} \ln a\}$$

$$\{e^{tY} \le b\} = \{tY \le \ln b\} = \{Y \le \frac{1}{t} \ln b\}.$$

Hence,

$$\mathbb{P}(e^{tX} \le a, e^{tY} \le b) = \mathbb{P}(X \le \frac{1}{t}\ln a, Y \le \frac{1}{t}\ln b) = \mathbb{P}(X \le \frac{1}{t}\ln a)\mathbb{P}(Y \le \frac{1}{t}\ln b)$$
$$= \mathbb{P}(e^{tX} \le a)\mathbb{P}(e^{tY} \le b).$$

In other words, for $t \neq 0$, the random variables e^{tX} and e^{tY} are also independent. When $t = 0$, on the other hand, $e^{tX} = e^{tY} = e^0 = 1$, and the constant random variable 1 *is* independent from itself. So it works for all t.

Why is this important? Recall that, whenever random variables A, B are independent, $\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$. (We proved this in the discrete case; in the continuous case, the proof is similar.) This is a very powerful tool for calculating the density of a sum of independent random variables.

Example 23.3. Let $X \sim N(0, \sigma_1^2)$ and $Y \sim N(0, \sigma_2^2)$ be normal random variables. As we calculated in Example 22.4,

$$M_X(t) = e^{\sigma_1^2 t^2/2}, \quad M_Y(t) = e^{\sigma_2^2 t^2/2}.$$

Thus, if X and Y are independent,

$$M_{X+Y}(t) = e^{\sigma_1^2 t^2/2} e^{\sigma_2^2 t^2/2} = e^{(\sigma_1^2+\sigma_2^2)t^2/2}.$$

But this is the moment-generating function of a $N(0, \sigma_1^2 + \sigma_2^2)$ random variable. So we have *very quickly* proved that $X + Y$ has a normal $N(0, \sigma_1^2 + \sigma_2^2)$ distribution.

So, in the presence of independence, the moment generating function transforms addition $X + Y$ into multiplication $M_X(t)M_Y(t)$. It might be more natural to have a transform that converts addition to addition. We can accomplish this by taking the logarithm:

$$\ln M_{X+Y}(t) = \ln\left(M_X(t)M_Y(t)\right) = \ln M_X(t) + \ln M_Y(t)$$

when X, Y are independent. So we define a new transform

$$C_X(t) = \ln M_X(t) = \ln \mathbb{E}(e^{tX}).$$

If we expand $M_X(t)$ in a power series, we get the moments:

$$M_X(t) = \sum_{n=0}^{\infty} \frac{\mathbb{E}(X^n)}{n!} t^n = \sum_{n=0}^{\infty} \frac{m_n}{n!} t^n.$$

What happens if we expand C_X in a power series?

$$C_X(t) = \sum_{n=0}^{\infty} \frac{c_n}{n!} t^n.$$

What is the relationship between the c_n and the moments m_n? To figure this out, we need to use Taylor's theorem (a lot). Let's just look at the first few c_ns.

$$c_0 + c_1 t + \frac{1}{2}c_2 t^2 + \cdots = C_X(t) = \ln M_X(t) = \ln(m_0 + m_1 t + \frac{1}{2}m_2 t^2 + \cdots)$$

First, remember that $m_0 = \mathbb{E}(X^0) = \mathbb{E}(1) = 1$. Now, Taylor's theorem says that

$$\ln(1 + x) = x - \frac{1}{2}x^2 + \frac{1}{3}x^3 - \cdots$$

If we plug in $x = m_1 t + \frac{1}{2} m_2 t^2 + \cdots$, we'll get a bunch of power series. The trick is to note that

$$(m_1 t + \frac{1}{2} m_2 t^2 + \cdots)^n = t^n (m_1 + \frac{1}{2} m_2 t + \cdot)$$

and so these terms won't show up until the nth power of t is considered. Thus, we only have to use the first two terms:

$$\begin{aligned} c_0 + c_1 t + \frac{1}{2} c_2 t^2 + \cdots &= (m_1 t + \frac{1}{2} m_2 t^2 + \cdots) - \frac{1}{2}(m_1 t + \frac{1}{2} m_2 t^2 + \cdots)^2 + \cdots \\ &= (m_1 t + \frac{1}{2} m_2 t^2 + \cdots) - \frac{1}{2}(m_1^2 t^2 + m_1 m_2 t^3 + \frac{1}{4} m_2 t^4 + \cdots) + \cdots \end{aligned}$$

Comparing coefficients,

$$c_0 = 0, \quad c_1 = m_1, \quad \frac{1}{2} c_2 = \frac{1}{2} m_2 - \frac{1}{2} m_1^2.$$

That is, $c_1 = \mathbb{E}(X)$, while $c_2 = \mathbb{E}(X^2) - \mathbb{E}(X)^2 = \mathrm{Var} X$.

The function C_X is called the **cumulant generating function**. The coefficients c_n

$$C_X(t) = \sum_{n=1}^{\infty} \frac{c_n}{n!} t^n = \mathbb{E}(X) t + \mathrm{Var} X t^2 + \cdots$$

are called the **cumulants** of X. They are polynomials in the moments of X. For example, if we continued the above Taylor series computations, we'd find that

$$c_3 = \mathbb{E}(X^3) - 3\mathbb{E}(X^2)\mathbb{E}(X) + 2\mathbb{E}(X)^3.$$

They have statistical meaning, just like the mean and variance. c_3 is called **skewness**; it is a measure of whether the distribution is skewed to the left or right. c_4 is called **curtosis**; it is measure of whether the distribution is tall and skinny, or short and wide.

Example 23.4. If X has a $N(0, \sigma^2)$ law, then $M_X(t) = e^{\sigma^2 t^2/2}$, so

$$C_X(t) = \ln M_X(t) = \frac{1}{2} \sigma^2 t^2.$$

In other words, $\mathbb{E}(X) = c_1 = 0$ and $\mathrm{Var} X = c_2 = \sigma^2$, as we knew, and $c_n = 0$ for all $n > 2$. This is yet another sign that Gaussians are very special. It also tells us that cumulants are, in general, a measure of how far we are from a normal distribution.

23.2. **Using sums of independent random variables.** The moment generating function is a great tool for calculating moments, and (as we just saw) for dealing with sums of independent random variables. This is a boon for us, since many of the problems we've already worked on involve sums of independent random variables, overtly or implicitly.

Example 23.5. Let $X_1, \ldots, X_n$ be indpendent random variables that all have the same (discrete) distribution:

$$\mathbb{P}(X_j = 1) = p, \quad \mathbb{P}(X_j = 0) = 1 - p, \quad 1 \le j \le n.$$

Think of $X_1, \ldots, X_n$ as n independent tosses of a biased coin (with probability p of heads). Now, let

$$S_n = X_1 + \cdots + X_n.$$

What is the distribution of S_n? Well, the state space of S_n is $\{0, 1, \ldots, n\}$. To say that $S_n = k$ means that exactly k of the X_j's equal 1 and the other $n - k$ equal 0. In other words, the event $\{S_n = k\}$ is the event of k successes out of n independent trials, each with success probability p. This is precisely the definition of the $\mathrm{Binomial}(n, p)$ law, so

$$\mathbb{P}(S_n = k) = \binom{n}{k} p^k (1-p)^{n-k}.$$

The point is: the binomial distribution arises as a sum of independent random variables.

24. LECTURE 24: NOVEMBER 22, 2010

Example 24.1 (Coupon Collector Problem)**.** *You collect coupons of some kind (e.g. baseball cards). There are n different coupons. Every time you receive one, it is indpendently equally likely to be any one of the n. How long does it take, on average, to collect all n coupons?*

We measure time in numbers of coupons collected. Define random variables $T_1, T_2, \ldots, T_n$ inductively as follows: once you have just collected $j-1$ distinct coupons, start counting; let T_j be the number of coupons you collect before you get a new coupon you haven't collected already. So $T_1 = 1$ (since the first coupon is always new). What we're interested in is the total time to collect all n distinct coupons, or $S_n = T_1 + T_2 + \cdots + T_n$.

First we can calculate expectation. We have $\mathbb{E}(S_n) = \mathbb{E}(T_1) + \mathbb{E}(T_2) + \cdots + \mathbb{E}(T_n)$. Now, we can actually calculate the distribution of T_j. Once we have collected $j-1$ distinct coupons, there are $n-(j-1)$ remaining ones. Each time we get a new coupon, it is equally likely to be any of the total n; so the probability that it is one of the desirable $n-(j-1)$ is $\frac{n-(j-1)}{n}$. Each new coupon collection is therefore an independent trial with success probability $p_j = \frac{n-(j-1)}{n}$, meaning that the time T_j of first success is a $\mathrm{geometric}(p_j)$ random variable.

In particular, $\mathbb{E}(T_j) = \frac{1}{p_j} = \frac{n}{n-(j-1)}$. Therefore

$$\mathbb{E}(S_n) = \sum_{j=1}^{n} \frac{n}{n-(j-1)}.$$

If we reindex this sum with $k = n-(j-1)$, then k ranges from 1 (when $j = n$) up to n (when $j = 1$), so

$$\mathbb{E}(S_n) = \sum_{k=1}^{n} \frac{n}{k} = n \cdot \left(1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{n}\right).$$

The number $H_n = 1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{n}$ is called the *harmonic number*. There is no simple formula for it, so the most precise answer we can give is that $\mathbb{E}(S_n) = nH_n$. But we can get a sense for how big the number is by approximating the sum by an integral:

$$\sum_{k=1}^{n} \frac{1}{k} \approx \int_1^n \frac{dt}{t} = \ln n.$$

So $\mathbb{E}(S_n) \approx n \ln n$. For example, if the coupons are actually playing cards, so $n = 52$, we get $\mathbb{E}(S_{52}) = 52H_{52} \approx 52 \ln 52 = 205.46467$ to 5 decimal places. In this case we can also explicitly calculate (using a computer) to find that $52H_{52} = 235.97829$ to 5 decimal places. So the approximation gives the right order of magnitude, but isn't that great actually.

We can do better by asking how close H_n is to $\ln n$ in general. Euler did this. In fact, it can be shown that H_n is always bigger than $\ln n$, but that the difference $H_n - \ln n$ stays bounded; in fact, it decreases, and it has a limit:

$$\lim_{n\to\infty} (H_n - \ln n) = \gamma = 0.5772156649\ldots$$

This number is called **Euler's constant**. Nobody knows a nice expression for this number. Nobody knows if its rational or irrational! But knowing how big it is to fairly good accuracy, we can make the better approximation

$$H_n \approx \ln n + \gamma,$$

so that

$$\mathbb{E}(S_n) = nH_n \approx n \ln n + \gamma n.$$

When $n = 52$, this gives

$$52 \ln 52 + \gamma \cdot 52 = 235.47989$$

to 5 decimal places – this is much better, accurate to within 0.5. (In fact, an even better approximation of nH_n is $n \ln n + \gamma n + \frac{1}{2}$; with $n = 52$ this is accurate to within 0.002.)

This is just the expectation of S_n. We would like to know more information about its distribution. To calculate more, we notice the following. Suppose we know that $T_1 = t_1, T_2 = t_2, \ldots, T_{j-1} = t_{j-1}$ for some particular values of $t_1, \ldots, t_{j-1}$. (In fact, we know that $t_1 = 1$.) Consider the *conditional distribution*

$$\mathbb{P}(T_j = t | T_1 = t_1, T_2 = t_2, \ldots, T_{j-1} = t_{j-1}).$$

The point here is that, past information is actually no help here. Once we have passed time $t_1 + \cdots + t_{j-1}$, there is some random collection of $j - 1$ coupons we have collected. But each now collection is independent of all past (and future) ones, and so the length of time it takes to find a coupon not in that group of $j - 1$ does not depend at all on the earlier times for success. In other words

$$\mathbb{P}(T_j = t | T_1 = t_1, T_2 = t_2, \ldots, T_{j-1} = t_{j-1}) = \mathbb{P}(T_j = t).$$

In other words, $T_1, \ldots, T_j$ are *independent*. This is true for all j, so $T_1, \ldots, T_n$ are independent. We could, therefore, in principle calculate the distribution of S_n, since $S_n = T_1 + \cdots + T_n$ is the sum of independent random variables, the distribution of each we know. In fact we could write down a formula for $\mathbb{P}(S_n = k)$, but instead let us just use this independence information to calculate variance. Recall that, when $T_1, \ldots, T_n$ are independent, $\mathrm{Var}(T_1 + \cdots + T_n) = \mathrm{Var}T_1 + \cdots + \mathrm{Var}T_n$. So

$$\mathrm{Var}S_n = \sum_{j=1}^{n} \mathrm{Var}T_n.$$

Since T_j is $\mathrm{geometric}(p_j)$ where $p_j = \frac{n-(j-1)}{n} = 1 - \frac{j-1}{n}$, we have (as calculated earlier in the quarter, and as you are recalculating using moment-generating functions on Homework 8)

$$\mathrm{Var}T_j = \frac{1 - p_j}{p_j^2} = \frac{\frac{j-1}{n}}{(1 - \frac{j-1}{n})^2} = n \frac{j-1}{(n-j+1)^2}.$$

Thus

$$\mathrm{Var}S_n = n \sum_{j=1}^{n} \frac{j-1}{(n-j+1)^2}.$$

This is another sum that cannot be explicitly simplified. Making the same change of variables $k = n - (j-1)$ we did before, the sum runs from $k = 1$ (when $j = n$) to $k = n$ when $j = 1$), giving

$$\mathrm{Var}S_n = n \sum_{j=1}^{n} \frac{n-k}{k^2} = n^2 \sum_{k=1}^{n} \frac{1}{k^2} - n \sum_{j=1}^{n} \frac{1}{k}.$$

The subtracted term is $nH_n = \mathbb{E}(S_n)$, which we already approximated as close to $n \ln n + \gamma n$. Now, the sum $\sum_{k=1}^{n} \frac{1}{k^2}$ grows with n, but has a finite limit as $n \to \infty$: $\sum_{k=1}^{n} \frac{1}{k^2} = \frac{\pi^2}{6}$ (also computed by Euler). Thus

$$\mathrm{Var} S_n \approx \frac{\pi^2}{6} n^2 - n \ln n - \gamma n.$$

When $n = 52$, this approximation gives $\mathrm{Var} S_{52} \approx 4212.42183$ to 5 decimal places; this is compared to the actual value which we can compute (with a computer) 4160.42023.

24.1. **Markov and Chebyshev's Inequalities.** There is an important reason we keep calculating variances. We have identified the standard deviation $\sigma(X) = \sqrt{\mathrm{Var} X}$ as a measure of how spread-out the distribution of X is. We will now rigorously prove a statement to this effect.

Lemma 24.2 (Markov's inequality)**.** *Let X be a random variable, with finite absolute expectation $\mathbb{E}(|X|)$. Then for any $t > 0$,*

$$\mathbb{P}(|X| > t) \leq \frac{\mathbb{E}(|X|)}{t}.$$

Proof. This is another exercise in the relationship between expectation and probability. We have a random variable $X : \Omega \to \mathbb{R}$ defined on some sample space Ω. We are interested in the probability of the event $\{|X| > t\}$. For any event A we have

$$\mathbb{P}(A) = \mathbb{E}(\mathbb{1}_A).$$

So, with $A = \{|X| > t\}$,

$$\mathbb{P}(|X| > t) = \mathbb{E}(\mathbb{1}_{|X|>t}).$$

Now, here's the key observation: when $|X| > t$, $\frac{1}{t}|X| > 1$. On the other hand, when $|X| \leq t$, $\frac{1}{t}|X| \geq 0$, which is the value of $\mathbb{1}_{|X|>t}$ when $|X| \leq t$. Altogether, this means

$$\frac{1}{t}|X| \geq \mathbb{1}_{|X|>t}.$$

Taking expectations gives

$$\frac{1}{t}\mathbb{E}(|X|) \geq \mathbb{E}(\mathbb{1}_{|X|>t}) = \mathbb{P}(|X| > t).$$

□

Corollary 24.3 (Chebyshev's inequality)**.** *Let X be a random variable with finite expectation and variance. Then for any $t > 0$,*

$$\mathbb{P}\left(|X - \mathbb{E}(X)| > s\sigma(X)\right) \leq \frac{1}{s^2}.$$

Proof. We apply Markov's inequality to the random variable $Y = (X - \mathbb{E}(X))^2$. Notice that $Y \geq 0$, so $|Y| = Y$, and by definition $\mathbb{E}(Y) = \mathbb{E}((X - \mathbb{E}(X))^2) = \mathrm{Var} X$. Therefore

$$\mathbb{P}(|Y| > t) \leq \frac{\mathbb{E}(|Y|)}{t} = \frac{\mathrm{Var} X}{t}.$$

Now, $|Y| > t$ means $(X - \mathbb{E}(X))^2 > t$, which means $|X - \mathbb{E}(X)| > \sqrt{t}$. Thus

$$\mathbb{P}(|X - \mathbb{E}(X)| > \sqrt{t}) \leq \frac{\mathrm{Var}X}{t}.$$

Finally, take $t = s^2\mathrm{Var}X$. Then $\sqrt{t} = s\sqrt{\mathrm{Var}X} = s\sigma(X)$, and so

$$\mathbb{P}\left(|X - \mathbb{E}(X)| > s\sigma(X)\right) \leq \frac{\mathrm{Var}X}{s^2\mathrm{Var}X} = \frac{1}{s^2}.$$

□

Chebyshev's inequality gives a quantitative bound on how likely a random variable is to be far from its mean, measured in number of standard deviations. The inequality gives *no information* when $s < 1$, since there it asserts that a certain probability is less than $1/s^2 > 1$ - no duh! But when $s = 2$ for example, it asserts that

$$\mathbb{P}(|X - \mathbb{E}(X)| > 2\sigma(X)) \leq \frac{1}{4}.$$

In other words, any random variable spends at least 75% of its time within two standard deviations of its mean.

Example 24.4. As we have calculated, if X is $N(0, \sigma^2)$, then

$$\mathbb{P}(|X| \leq 2\sigma^2) \approx 95\%,$$

much better than 75%. But this is *only true for Gaussians*. In general, Chebyshev's inequality cannot be improved: there are random variables (like ones we've seen with lots of mass at the edges and very little in the middle) that achieve the bound exactly.

Example 24.5. Let's look at the Coupon Collector Problem of Example 24.1 once more. We showed that, the time S_n to collect all n coupons has expectation $\mathbb{E}(S_n) = nH_n \approx n\ln n + \gamma n$ while $\mathrm{Var}S_n = n^2P_n - nH_n \approx \frac{\pi^2}{6}n^2 - n\ln n - \gamma n$ where $\gamma = 0.5772156649\ldots$ is Euler's constant, and $P_n = \sum_{k=1}^{n} \frac{1}{k^2}$. Thus

$$\sigma(S_n) = n\sqrt{P_n - H_n/n}$$

where $\sqrt{P_n - H_n/n}$ is close to $\pi/\sqrt{6}$. For example, when $n = 52$, exact calculation gives $\sigma(S_52) = 64.50132$ to 5 decimal places, while $52 \cdot \pi/\sqrt{6} = 66.69259$ to 5 decimal places. The upshot is that, since $\mathbb{E}(S_{52})$ is about 236 while $\sigma(S_{52})$ is about 65, we know that with probability at least 75%, $|S_{52} - 236| \leq 2 \cdot 65$, or

$$\mathbb{P}(S_{52} \in [106, 366]) \geq 75\%.$$

If we look at a $3\sigma = 195$ interval, we get a probability of $1 - \frac{1}{9} = 88.888\ldots\%$, so

$$\mathbb{P}(S_{52} \in [41, 431]) \geq 88.8\%.$$

Note, in this latter case, we must have enough time to collect all 52 coupons, so the lower bound of 41 is meaningless: we know $S_52 \geq 52$. But the upper bound is still meaningful: collecting 52 coupons will take at most 431 collections, with probability at least 88.8%.

24.2. **Covariance.** As we proved earlier in the quarter, if $X_1, X_2, \ldots, X_n$ are independent, then

$$\mathrm{Var}(X_1 + \cdots + X_n) = \mathrm{Var}X_1 + \cdots + \mathrm{Var}X_n.$$

Something interesting came out of the calculation that proved this formula. In general, *without assuming independence*, we had

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

The quantity added to the sum of the variances is called **covariance**:

$$\mathrm{Cov}(X,Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y) = \mathbb{E}[(X - \mathbb{E}(X))(Y - \mathbb{E}(Y))].$$

If X, Y are independent, then $\mathrm{Cov}(X,Y) = 0$. The converse is **not true**. In general, random variables that have covariance 0 are called **uncorrelated**.

Example 24.6. Let X be uniform on $[-1,1]$. Let $Y = X^2$. Then X and Y are certainly not independent: for example, $\mathbb{P}(Y \le \frac{1}{4}) = \mathbb{P}(-\frac{1}{2} \le X \le \frac{1}{2}) = \frac{1}{2}$, but

$$\mathbb{P}(X \le \frac{1}{2}, Y \le \frac{1}{4}) = \mathbb{P}(X \le \frac{1}{2}, -\frac{1}{2} \le X \le \frac{1}{2}) = \mathbb{P}(-\frac{1}{2} \le X \le \frac{1}{2}) = \frac{1}{2}$$

while

$$\mathbb{P}(X \le \frac{1}{2})\mathbb{P}(Y \le \frac{1}{4}) = \frac{3}{4} \cdot \frac{1}{2} = \frac{3}{8}.$$

However, we can quickly calculate

$$\mathbb{E}(X) = \int_{-1}^{1} x \cdot \frac{1}{2}\, dx = 0$$

$$\mathbb{E}(Y) = \mathbb{E}(X^2) = \int_{-1}^{1} x^2 \cdot \frac{1}{2}\, dx = \frac{1}{3}x^3\Big|_{x=-1}^{x=1} = \frac{2}{3}$$

$$\mathbb{E}(XY) = \mathbb{E}(X^3) = \int_{-1}^{1} x^3 \cdot \frac{1}{2}\, dx = 0.$$

Thus, $\mathbb{E}(XY) = 0 = \mathbb{E}(X)\mathbb{E}(Y)$, so $\mathrm{Cov}(X,Y) = 0$.

Uncorrelated random variables are "approximately independent"; precisely, they are independent *to second order*. By construction, if $X_1, X_2, \ldots, X_n$ are all uncorrelated $\mathrm{Cov}(X_i, X_j) = 0$ for $i \neq j$, then

$$\mathrm{Var}(X_1 + \cdots + X_n) = \mathrm{Var}X_1 + \cdots + \mathrm{Var}X_n$$

just as for independent random variables. In the real world, when we collect data, it is difficult (essentially impossible) in most cases to determine if variables are independent; but it is easy to compute covariance. As we will see in the next few lectures, many of the major theorems about large aggregates of data work just as well with uncorrelated random variables as with truly independent random variables.

25. Lecture 25: November 24, 2010

We now stand ready to finally justify that our model of probability theory matches up with our intuitive notion of probability being long-term frequency of success.

25.1. **The weak law of large numbers.** Let $X_1, X_2, X_3, \ldots$ be a sequence of random variables defined on a common probability space $(\Omega, \mathbb{P})$. We think of X_k as the number measured in the kth trial of an experiment. Hence, we assume the X_k are **independent**, and all have the **same distribution**. We then look at the average value

$$\overline{X}_n = \frac{S_n}{n} = \frac{X_1 + \cdots + X_n}{n}.$$

Our first theorem shows that this random variable $\overline{X}_n$ is very close to constant for large n.

Theorem 25.1 (WLLN)**.** *Suppose that the X_k are independent, all with the same distribution, and let $\mathbb{E}(X_k) = \mu$ and $\mathrm{Var}(X_k) = \sigma^2$ exist. Then for any $\epsilon > 0$,*

$$\lim_{n\to\infty} \mathbb{P}(|\overline{X}_n - \mu| > \epsilon) = 0.$$

In other words, no matter how small a window $(\mu - \epsilon, \mu + \epsilon)$ we want to consider, eventually $\overline{X}_n$ stays within this window with high probability.

Proof. Note that

$$\overline{X}_n - \mu = \frac{S_n}{n} - \mu = \frac{S_n - \mu n}{n}.$$

Now, $\mu = \mathbb{E}(X_1) = \mathbb{E}(X_2) = \cdots = \mathbb{E}(X_n)$, so by linearity of $\mathbb{E}$,

$$\mathbb{E}(S_n) = \mathbb{E}(X_1 + \cdots + X_n) = \mathbb{E}(X_1) + \cdots + \mathbb{E}(X_n) = n\mu.$$

Therefore

$$\mathbb{P}(|\overline{X}_n - \mu| > \epsilon) = \mathbb{P}\left(\frac{|S_n - \mathbb{E}(S_n)|}{n} > \epsilon\right) = \mathbb{P}(|S_n - \mathbb{E}(S_n)| > n\epsilon).$$

We want to use Chebyshev's inequality to estimate this; to do so, we need to have the standard deviation of S_n on the right hand side.

$$\mathbb{P}(|S_n - \mathbb{E}(S_n)| > n\epsilon) = \mathbb{P}(|S_n - \mathbb{E}(S_n)| > \frac{n\epsilon}{\sigma(S_n)}\sigma(S_n)) \leq \frac{1}{(\frac{n\epsilon}{\sigma(S_n)})^2} = \frac{\mathrm{Var}(S_n)}{n^2\epsilon^2}.$$

Now, since the X_k are independent,

$$\mathrm{Var}S_n = \mathrm{Var}(X_1 + \cdots + X_n) = \mathrm{Var}X_1 + \cdot + \mathrm{Var}X_n = n\sigma^2.$$

So, we find altogether that

$$\mathbb{P}(|\overline{X}_n - \mu| > \epsilon) = \mathbb{P}(|S_n - \mathbb{E}(S_n)| > n\epsilon) \leq \frac{\mathrm{Var}(S_n)}{n^2\epsilon^2} = \frac{n\sigma^2}{n^2\epsilon^2} = \frac{\sigma^2}{\epsilon^2 n}.$$

For any $\epsilon > 0$, this converges to 0 as $n \to \infty$, proving the theorem. □

Remark 25.2. Actually, this proof didn't require as many conditions as the theorem required. The only place independence was used was to justify $\mathrm{Var}S_n = n\sigma^2$, which holds more generally if the variables are *uncorrelated*. Also, the only thing about the distribution we needed to know was that $\mathbb{E}(X_k) = \mu$ and $\mathrm{Var}X_k = \sigma^2$ for all k.

25.2. **The strong law of large numbers.** The weak law of large numbers helps to justify our intuition about long-term frequencies: the (empirical) average of a set of independent data is very concentrated near its mean, and gets more and more concentrated at the rate of about $\frac{1}{n}$ as $n \to \infty$. However, in Lecture 2, we presented our intuition about what probability should mean in terms of the existence of certain limits: an event E has probability p if the limit

$$\lim_{n\to\infty} \frac{N_n(E)}{n} = p$$

where $N_n(E)$ is the number of times E occurs in n independent trials. The very existence of this limit was the big outstanding question. To see that we have produced a theory that can answer this question, we need the following **strong law of large numbers**.

Theorem 25.3 (SLLN)**.** *Let $X_1, X_2, X_3, \ldots$ be a sequence of independent random variables, each with the same distribution. Suppose that $\mathbb{E}(X_k^4) < \infty$. Let $\mathbb{E}(X_k) = \mu$. Setting $\overline{X}_n = \frac{S_n}{n}$ where $S_n = X_1 + \cdots + X_n$, then*

$$\mathbb{P}\left(\lim_{n\to\infty} \overline{X}_n = \mu\right) = 1.$$

This theorem says that the random variables $\overline{X}_n$ actually converge to a constant (with probability 1). This is the precise statement we were looking for, on the existence of limits to justify probability theory as a theory of limiting frequencies. This proof requires the 4th moment to exist; in fact, there are proofs that require only the expectation to exist, but they are much technically harder.

Proof. First, let's subtract off the μ: we will prove that $\lim_{n\to\infty}(\overline{X}_n - \mu) = 0$. Let's actually define centered random variables from the start: $Y_n = X_n - \mu$, so that $\mathbb{E}(Y_n) = 0$. Then with $\overline{Y}_n = \frac{1}{n}(Y_1 + \cdots + Y_n)$, we have $\overline{Y}_n = \frac{1}{n}(X_1 - \mu + X_2 - \mu + \cdots + X_n - \mu) = \frac{1}{n}(S_n - n\mu) = \overline{X}_n - \mu$; in other words, we want to prove that $\lim_{n\to\infty} \overline{Y}_n = 0$ with probability 1.

The random variables Y_n are also independent, and identically distributed. We can compute that

$$\mathbb{E}(Y_k^4) = \mathbb{E}((X_k - \mu)^4) = \mathbb{E}(X_k^4) - 4\mu\mathbb{E}(X_k^3) + 6\mu^2\mathbb{E}(X_k^2) - 4\mu^3\mathbb{E}(X_k) + \mu^4,$$

and since $|\mathbb{E}(X_k)| \le |\mathbb{E}(X_k^2)| \le |\mathbb{E}(X_k^3)| \le \mathbb{E}(X_k^4)$, this is also finite. Let's give it a name: $\mathbb{E}(Y_k^4) = \kappa^4$. Also, $\mathbb{E}(Y_k^2) = \mathbb{E}((X_k - \mu)^2) = \mathbb{E}(X_k - \mathbb{E}(X_k))^2 = \mathrm{Var} X_1$; call this σ^2 Why should we care about these quantities? Well, let $T_n = Y_1 + \cdots + Y_n$, so that $\overline{Y}_n = \frac{T_n}{n}$. Then

$$\mathbb{E}(\overline{Y}_n^4) = \frac{1}{n^4}\mathbb{E}(T_n^4) = \frac{1}{n^4}\mathbb{E}\left(\sum_{k=1}^n Y_k\right)^4.$$

Let's expand out this sum.

$$\begin{aligned}\left(\sum_{k=1}^n Y_k\right)^4 &= \left(\sum_{i=1}^n Y_i\right)\left(\sum_{j=1}^n Y_j\right)\left(\sum_{k=1}^n Y_k\right)\left(\sum_{\ell=1}^n Y_\ell\right) \\ &= \sum_{i,j,k,\ell=1}^n Y_i Y_j Y_k Y_\ell.\end{aligned}$$

When we take expectations, different things happen depending on how many of the indices are equal or not. For example, one of the terms that comes up is $Y_1Y_3Y_1Y_7$; using independence, we have

$$\mathbb{E}(Y_1Y_3Y_1Y_7) = \mathbb{E}(Y_1^2Y_3Y_7) = \mathbb{E}(Y_1^2)\mathbb{E}(Y_3)\mathbb{E}(Y_7)$$

and this equals 0 since $\mathbb{E}(Y_3) = 0$. This kind of argument shows that a lot of the terms in the big quadruple sum have expectation 0. In fact: whenever one of i, j, k, ℓ is distinct from the other 3 (lets say it's i), we have

$$\mathbb{E}(Y_iY_jY_kY_\ell) = \mathbb{E}(Y_i)\mathbb{E}(Y_jY_kY_\ell) = 0$$

since Y_i is independent from $Y_jY_kY_\ell$. So when we take expectations, the only terms that survive in the quadruple sum are those for which every index is repeated at least once. This can happen in two ways:

- All four indices are the same: $i = j = k = \ell$. There are n such terms, and in each case
$$\mathbb{E}(Y_iY_jY_kY_\ell) = \mathbb{E}(Y_i^4) = \kappa^4.$$
- The indices come in two pairs: that is, we divide the set of four indices into 2 groups of size 2 (which can be done in $\binom{4}{2} = 6$ ways), and then assign two distinct values to the two groups (which can be done in $\binom{n}{2} = \frac{1}{2}n(n-1)$ ways). For each of these $3n(n-1)$ terms, the expectation is
$$\mathbb{E}(Y_iY_jY_kY_\ell) = \mathbb{E}(Y_1^2)\mathbb{E}(Y_2^2) = \sigma^4.$$

So, we can actually calculate that

$$\mathbb{E}\left(\sum_{k=1}^{n} Y_k\right)^4 = n\kappa^4 + 3n(n-1)\sigma^4.$$

Thus,

$$\mathbb{E}(\overline{Y}_n^4) = \frac{n\kappa^4 + 3n(n-1)\sigma^4}{n^4} \leq \frac{(3\sigma^4 + \kappa^4)n^2}{n^4} = \frac{3\sigma^4 + \kappa^4}{n^2}.$$

The constant on the top doesn't matter; what matters is that it is finite. The sequence $\frac{1}{n^2}$ is summable; that is

$$\sum_{n=1}^{\infty} \mathbb{E}(\overline{Y}_n^4) \leq \sum_{n=1}^{\infty} \frac{3\sigma^4 + \kappa^4}{n^2} < \infty.$$

But now we interchange the sum and the expectation

$$\sum_{n=1}^{\infty} \mathbb{E}(\overline{Y}_n^4) = \mathbb{E}\left(\sum_{n=1}^{\infty} \overline{Y}_n^4\right)$$

so the random variable $\sum_{n=1}^{\infty} \overline{Y}_n^4$ has finite expectation. This cannot happen if this non-negative random variable is infinite on a set with positive probability! Hence, we have proved that

$$\mathbb{P}(\sum_{n=1}^{\infty} \overline{Y}_n^4 < \infty) = 1.$$

But for any sequence a_n, if $\sum_{n=1}^{\infty} a_n < \infty$ then $a_n \to 0$ as $n \to \infty$. Hence,

$$\left\{ \sum_{n=1}^{\infty} \overline{Y}_n^4 < \infty \right\} \subset \left\{ \lim_{n\to\infty} \overline{Y}_n^4 = 0 \right\} = \left\{ \lim_{n\to\infty} \overline{Y}_n = 0 \right\}.$$

So finally, we have

$$1 = \mathbb{P}\left(\sum_{n=1}^{\infty} \overline{Y}_n^4 < \infty \right) \leq \mathbb{P}\left(\lim_{n\to\infty} \overline{Y}_n = 0 \right).$$

This completes the proof. □

With this, we can now justify our formulation of probability theory.

Corollary 25.4 (Borel's law of large numbers)**.** *Let E be an event in some sample space Ω. Perform repeated independent trials, and record the number of times $N_n(E)$ that E occurs in the first n trials. Then*

$$\lim_{n\to\infty} \frac{N_n(E)}{n} \quad \textit{exists.}$$

Proof. This is just a question of interpreting what is meant by "perform repeated independent trials". The precise interpretation we give is: let $T_1, T_2, T_3, \ldots$ be a sequence if independent, identically distributed random variables taking values in Ω. We then interpret the statement "E occurs in the nth trial" as the event $T_n \in E$.

Now define new random variables $X_1, X_2, X_3, \ldots$ by

$$X_n = \mathbb{1}_E(T_n).$$

Since the T_n are independent, so are the X_n. We can also quickly work out the distribution of X_n: it takes only two values, 0 and 1, with probabilities

$$\mathbb{P}(X_n = 1) = \mathbb{P}(\mathbb{1}_E(T_n) = 1) = \mathbb{P}(T_n \in E) = \mathbb{P}(T_1 \in E) \equiv p,$$
$$\mathbb{P}(X_n = 0) = \mathbb{P}(\mathbb{1}_E(T_n) = 0) = \mathbb{P}(T_n \notin E) = \mathbb{P}(T_1 \notin E) = 1 - p.$$

Thus, the X_n are also independent and identically distributed. Finally, we compute $\mathbb{E}(X_n) = 1 \cdot p + 0 \cdot (1-p) = p$. The strong law of large numbers therefore asserts that

$$\lim_{n\to\infty} \overline{X}_n = p, \quad \text{with probability } 1.$$

But what does $\overline{X}_n$ represent? It is

$$\overline{X}_n = \frac{X_1 + \cdots + X_n}{n}.$$

Since $X_j = 1$ iff $T_j \in E$ and 0 otherwise, the sum $X_1 + \cdots + X_n$ counts exactly the number of times E occurs in the first n trials; that is, $X_1 + \cdots + X_n = N_n(E)$, and this concludes the proof. □

Example 25.5. *Suppose that the stock market behaves in the following ridiculously oversimplifies manner: each year, it either grows 20% or shrinks 20%, each possibility occurring independently with probability $\frac{1}{2}$. After n years (for large n), how do we expect the market has changed?*

Let M_n be the value of the market after n years. Then the model is that $M_n = M_{n-1}X_n$ where X_n is a random variable taking values 1.2 and 0.8 with probabilities $\frac{1}{2}$ each, and where the $X_1, X_2, X_3, \ldots$ are independent. Thus

$$M_n = M_0 X_1 X_2 \cdots X_n$$

where M_0 is the initial value of the market. We know how to handle sums of independent random variables, so to get this into a form we can understand, we take logarithms:

$$\ln M_n = \ln M_0 + \ln X_1 + \cdot + \ln X_n.$$

Now, the random variables $\ln X_1, \ln X_2, \ln X_3, \ldots$ are independent, and each has distribution

$$\mathbb{P}(\ln X_k = \ln 1.2) = \mathbb{P}(\ln X_k = \ln 0.8) = \frac{1}{2}.$$

So the expected value of $\ln X_k$ is

$$\mathbb{E}(\ln X_k) = \frac{1}{2}(\ln 1.2 + \ln 0.8) = \frac{1}{2}\ln(1.2 \cdot 0.8) = \frac{1}{2}\ln 0.96.$$

Not that this is *negative*. The strong law of large numbers now tells us that, with probability 1,

$$\lim_{n\to\infty} \frac{\ln X_1 + \cdots + \ln X_n}{n} = \frac{1}{2}\ln 0.96.$$

So, for large n,

$$\ln X_1 + \cdots + \ln X_n \approx \frac{1}{2}\ln 0.96 \cdot n.$$

Exponentiating, the product is

$$X_1 \cdots X_n = e^{\ln X_1 + \cdots + \ln X_n} \approx e^{\frac{1}{2} n \ln 0.96} = (0.96)^{n/2}.$$

So we have that the value of the market after n years is, for large n, approximately

$$M_n \approx M_0 \cdot (0.96)^{n/2}.$$

In other words, in this model, the market decays exponentially over time. This is therefore not a good model! It may be surprising that, with equal probabilities of positive and negative growth, we get overall negative growth. This is precisely because the growths are measured in percentages of the current total, not some absolute: a 20% reduction of 120% of the initial value corresponds to 24% of the initial value; but a 20% increase of 80% of the initial value is only a 16% increase from the initial value.

26. LECTURE 26: NOVEMBER 29, 2010

26.1. **Rate of Convergence.** The (weak and strong) laws of large numbers show that if $X_1, X_2, X_3, \ldots$ are independent and identically distributed, with expectation μ, then the empirical average $\overline{X}_n$ converges to μ:

$$\lim_{n\to\infty} \overline{X}_n = \lim_{n\to\infty} \frac{X_1 + \cdots + X_n}{n} = \mu.$$

How fast does it converge? To answer this question, its better to "standardize" by subtracting μ:

$$\lim_{n\to\infty} (\overline{X}_n - \mu) = 0.$$

The question is: how fast does this converge to 0? An exact answer is pretty tricky to find, but we can get a very close approximation by looking at the standard deviation. This is actually how we proved the weak law of large numbers.

$$\mathrm{Var}(\overline{X}_n - \mu) = \mathrm{Var}\overline{X}_n = \mathrm{Var}\left(\frac{X_1 + \cdots + X_n}{n}\right) = \frac{1}{n^2}\mathrm{Var}(X_1 + \cdots + X_n).$$

Since the X_k's are independent, $\mathrm{Var}(X_1 + \cdots + X_n) = \mathrm{Var}X_1 + \cdots + \mathrm{Var}X_n = n\sigma^2$ where $\sigma^2 = \mathrm{Var}X_k$ is the common variance. Thus

$$\mathrm{Var}(\overline{X}_n - \mu) = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}.$$

Taking square roots, the standard deviation is

$$\sigma(\overline{X}_n - \mu) = \frac{\sigma}{\sqrt{n}}.$$

Hence, the standard deviation (a measure of spread) of $\overline{X}_n - \mu$ decays like $\sigma/\sqrt{n}$. It is customary to "standardize" by dividing by σ; so we have that

$$\text{the standard deviation of } \frac{\overline{X}_n - \mu}{\sigma} \text{ is } = \frac{1}{\sqrt{n}}.$$

So a good guess is that the sequence $[\overline{X}_n(\omega) - \mu]/\sigma$ converges to 0 at the rate $\frac{1}{\sqrt{n}}$ for each specific outcome ω. Of course, this can't be exact since there is randomness. In fact, the randomness results in a very slightly larger rate of precise decay:

$$\frac{\overline{X}_n - \mu}{\sigma} \approx \sqrt{\frac{2\ln\ln n}{n}} \text{ as } n \to \infty.$$

This is called the **law of the iterated logarithm** (because it involves an iterated logarithm), and was first proved by Khinchin in 1924. But this is beyond the tools we've developed in this class; if you continue into 180B and 180C, you may say it.

Note, this means that if we multiply $\overline{X}_n - \mu$ by n^α for any $\alpha < \frac{1}{2}$, it *still* converges to 0. It is better to write this in terms of the empirical sum rather than the empirical average: $S_n = X_1 + \cdots + X_n$, so $S_n = n\overline{S}_n$. So

$$\text{the standard deviation of } \frac{S_n/n - \mu}{\sigma} = \frac{S_n - n\mu}{n\sigma} \text{ is } = \frac{1}{\sqrt{n}} = n^{-1/2}.$$

So, multiplying both sides by n^α

$$\text{the standard deviation of } \; n^\alpha \frac{S_n - n\mu}{n\sigma} = \frac{S_n - n\mu}{n^{1-\alpha}\sigma} \; \text{ is } \; = n^{\alpha - 1/2}.$$

Since $n^{\alpha-1/2} \to 0$ as $n \to \infty$ provided $\alpha < 1/2$, we can divide $S_n - n\mu$ by something *smaller* than n and still get 0. But this stops working when we get to $\alpha = 1/2$, where the statement is

$$\text{the standard deviation of } \; \frac{S_n - n\mu}{\sigma\sqrt{n}} \; \text{ is } \; = 1.$$

So there's no way this can converge to 0. So as $n \to \infty$, this rescaled random variable maintains its randomness. In fact, something miraculous happens. The limiting distribution of this random variable is *completely universal*.

26.2. **The Central Limit Theorem.**

Theorem 26.1 (Central Limit Theorem)**.** *Let $X_1, X_2, X_3, \ldots$ are independent, identically-distributed random variables, with common expectation μ and standard deviation σ. Let $S_n = X_1 + \cdots + X_n$. Then as $n \to \infty$, the distribution of $\frac{S_n - n\mu}{\sigma\sqrt{n}}$ converges to a standard normal $N(0,1)$. That is: for any $a \leq b$,*

$$\lim_{n\to\infty} \mathbb{P}\left(a \leq \frac{S_n - n\mu}{\sigma\sqrt{n}} \leq b \right) = \frac{1}{\sqrt{2\pi}} \int_a^b e^{-x^2/2} dx.$$

We will outline a proof of the Central Limit Theorem, but first it is important to fully understand what it is saying. It may be helpful to return to the empirical mean notation.

$$\left\{ a \leq \frac{S_n - n\mu}{\sigma\sqrt{n}} \leq b \right\} = \left\{ a \leq \frac{n\overline{X}_n - n\mu}{\sigma\sqrt{n}} \leq b \right\} = \left\{ \frac{a}{\sqrt{n}} \leq \frac{\overline{X}_n - \mu}{\sigma} \leq \frac{b}{\sqrt{n}} \right\}.$$

So, the strong law of large number together with the central limit theorem say the following.

> *Let $X_1, X_2, X_3, \ldots$ be a sequence of independent trials, each with the same distribution, having mean μ and standard deviation σ. Then the standardized average $\frac{\overline{X}_n - \mu}{\sigma}$ converges to 0 at the rate of about $\frac{1}{\sqrt{n}}$. Moreover*
>
> *the probability that $\dfrac{\overline{X}_n - \mu}{\sigma}$ is in the small interval $\left[-\dfrac{a}{\sqrt{n}}, \dfrac{a}{\sqrt{n}} \right]$ is close to $\dfrac{1}{\sqrt{2\pi}} \displaystyle\int_{-a}^{a} e^{-x^2/2} dx$.*

Example 26.2. *An experimental scientist performs independent trials, in the nth trial making a measurement X_n. The experimental setup is identical for each trial, but there is measurement error, and uncontrollable random effects that make the measurements all different. In the end, s/he will average the data to get an approximation of the* true *value μ.*

If s/he knows that the true *standard deviation of the data is 2, how many measurements should s/he make so that it is 95% likely that the measured empirical average is within 0.5 units of the true mean μ?*

This is a typical formulation of a problem in statistics. The reason that 95% is the common "sureness" probability is that

$$\frac{1}{\sqrt{2\pi}} \int_{-2}^{2} e^{-x^2/2} dx \approx 0.9545$$

Now, the central limit theorem says that, for large n,

$$\mathbb{P}\left(\overline{X}_n - \mu \in \left[-\frac{\sigma a}{\sqrt{n}}, \frac{\sigma a}{\sqrt{n}}\right]\right) = \mathbb{P}\left(\frac{\overline{X}_n - \mu}{\sigma} \in \left[-\frac{a}{\sqrt{n}}, \frac{a}{\sqrt{n}}\right]\right) \approx \frac{1}{\sqrt{2\pi}} \int_{-a}^{a} e^{-x^2/2} dx.$$

So, setting $a = 2$, this probability is over 95%. Thus, if we choose n large enough that $\frac{2\sigma}{\sqrt{n}} \leq 0.5$, we will get

$$\mathbb{P}(\overline{X}_n - \mu \in [-0.5, 0.5]) \geq \mathbb{P}\left(\overline{X}_n - \mu \in \left[-\frac{2\sigma}{\sqrt{n}}, \frac{2\sigma}{\sqrt{n}}\right]\right) \approx \frac{1}{\sqrt{2\pi}} \int_{-2}^{2} e^{-x^2/2} dx \geq 95\%.$$

Since we know $\sigma = 2$ in this experiment, we need n large enough so that $\frac{4}{\sqrt{n}} \leq 0.5$, which means $\sqrt{n} \geq 8$, so $n \geq 64$.

Now, there is a bit of uncertainty here. The central limit theorem only gives an *approximation* for finite n. So, when all is said and done, we have to ask whether the approximation in the above step is good enough when $n = 64$ to justify the calculation! This is actually a bit tricky to answer precisely, but in general the normal approximation is a *very* good one. Nevertheless, if the scientist wanted to be absolutely *certain*, s/he could use Chebyshev's inequality:

$$\mathbb{P}(|\overline{X}_n - \mu| > 0.5) \leq \frac{\mathrm{Var}\overline{X}_n}{0.5^2} = \frac{\sigma^2}{n \cdot 0.5^2} = \frac{16}{n}.$$

Thus

$$\mathbb{P}(|\overline{X}_n - \mu| \leq 0.5) \geq 1 - \frac{16}{n}.$$

To make sure this is at least 95%, we should choose n large enough that $1 - \frac{16}{n} \geq \frac{19}{20}$, which means $n \geq 16 \cdot 20 = 320$. This is a *lot* more trials than the central limit theorem approach suggested (a factor of 5). This is the trade-off: if we want to be sure, we can use Chebyshev's inequality, which is a very weak bound in general. The central limit theorem gives only an approximation for the probability (though one that improves with the number of trials), but the answer it gives is generally a lot more practical.

Example 26.3. *A fair coin is tossed* 100 *times. What is the probability that there are at least* 55 *heads?*

We model a coin toss as a random variable X taking values 1 (heads) and 0 (tails) each with probability $\frac{1}{2}$. The coin tosses are independent trials $X_1, \ldots, X_{100}$ each with mean 0.5 and variance $(0.5)(1-0.5) = 0.25$ so standard deviation 0.5. Letting $S_n = X_1 + \cdots + X_n$ as usual, we want to calculate $\mathbb{P}(S_{100} \geq 55)$. We can evaluate this exactly by adding the Binomial distribution, but instead we will use the central limit theorem to approximate it.

$$\mathbb{P}(S_{100} \geq 55) = \mathbb{P}(S_{100} - 100 \cdot 0.5 \geq 5) = \mathbb{P}\left(\frac{S_{100} - 50}{0.5\sqrt{100}} \geq 1\right)$$

and the central limit theorem says that this is close to

$$\frac{1}{\sqrt{2\pi}} \int_1^\infty e^{-x^2/2} dx \doteq 15.87\%.$$

The precise answer is

$$\sum_{n=55}^{100} \binom{100}{k} \left(\frac{1}{2}\right)^{100} \doteq 18.4101\%.$$

So we're not far off, but there is still a significant gap. It turns out this is not because n is too small; it is because we are using a continuous distribution (the Gaussian) to approximate a discrete one (the binomial). Indeed, note that

$$\mathbb{P}(S_{100} \geq 55) + \mathbb{P}(S_{100} \leq 54) = 1$$

but a continuous distribution will have some positive probability of lying in the interval $(54, 55)$. This is where the discrepancy comes from. There is an easy fix for this. Since $\mathbb{P}(S_{100} \leq 54) = \mathbb{P}(S_{100} \leq 54.5)$, we should also replace $\mathbb{P}(S_{100} \geq 55)$ with $\mathbb{P}(S_{100} \geq 54.5)$ when doing the normal approximation. This is called the **histogram correction**. To wit,

$$\mathbb{P}(S_{100} \geq 55) \approx \mathbb{P}(S_{100} \geq 54.5) = \mathbb{P}(S_{100} - 50 \geq 4.5) = \mathbb{P}\left(\frac{S_{100} - 50}{0.5\sqrt{100}} \geq 0.9\right) \approx 18.406\%.$$

27. LECTURE 27: DECEMBER 1, 2010

Example 27.1. Back in Example 10.10, we analyzed a disputed election. Candidate A received 1405 votes, while candidate B received 1422. After the election, it was noted that 101 more votes we cast than the number of registered voters, so 101 votes should be disqualified – but which ones? The question is, how likely is it that throwing out a random set of 101 votes would overturn the election results?

Instead of analyzing this is a ball and urn problem, we could just select 101 votes at random (independently) from the batch. This is equivalently to flipping a biased coin 101 times; we identify candidate A with 1 and candidate B with 0; then each coin toss $X_1, \ldots, X_{101}$ has $\mathbb{P}(X_k = 1) = \frac{1405}{2827}$ and $\mathbb{P}(X_k = 0) = \frac{1422}{2827}$. Hence

$$\mathbb{E}(X_k) = \frac{1405}{2827} \approx 0.496993 \qquad \mathrm{Var} X_k = \frac{1997910}{7991929} \approx 0.249991.$$

These are *really* close to $\frac{1}{2}$ and $\frac{1}{4}$ (the results for a fair coin), so since we are doing an approximation anyhow, we will use a fair coin; that is, the mean is 0.5 and the standard deviation is $\sqrt{0.25} = 0.5$. In order to reverse the election results, at most 41 of the contested (throw-away) votes can be for candidate A. Making the histogram correction,

$$\begin{aligned}\mathbb{P}(X_1 + \cdots + X_{101} \leq 41) &= \mathbb{P}(S_{101} \leq 41.5) \\ &= \mathbb{P}\left(S_{101} - 101 \cdot 0.5 \leq 41.5 - 101 \cdot 0.5 = -9\right) \\ &= \mathbb{P}\left(\frac{S_{101} - 101 \cdot 0.5}{0.5\sqrt{101}} \leq \frac{-9}{0.5\sqrt{101}} \doteq -1.7911\right).\end{aligned}$$

So, the central limit theorem tells us that this is

$$\approx \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{-1.7911} e^{-x^2/2} dx \doteq 3.66\%.$$

(The actual value we computed with ball and urn methods in Lecture 10 was 3.87%.)

The histogram correction doesn't always improve matters. Like with numerical integration schemes: the midpoint approximation is often better than the right-end-point, but not always (depending on convexity of the function being integrated).

Example 27.2. *The number X of students who pre-enroll in this course is* $\mathrm{Poisson}(80)$. *This room seats* 100. *What is the probability there will not be enough seats in the first lecture?*

We want to compute $\mathbb{P}(X > 100)$. Since we know $\mathbb{P}(X = n) = e^{-80}\frac{80^n}{n!}$, we can actually compute this exactly:

$$\mathbb{P}(X > 100) = 1 - \mathbb{P}(X \leq 100) = 1 - \sum_{n=0}^{100} \mathbb{P}(X = n) = 1 - e^{-80} \sum_{n=0}^{100} \frac{80^n}{n!}.$$

A computational software package like Maple can compute this quickly; it is $\doteq 1.317\%$. But we can do a normal approximation if we remember the result (from Exam 1) that if $X_1, X_2, \ldots, X_n$ are $\mathrm{Poisson}(1)$ and *independent* then $S_n = X_1 + \cdots + X_n$ is $\mathrm{Poisson}(n)$. So we can realize $X = S_{80}$ for independent random variables with

$$\mathbb{E}(X_k) = 1, \quad \mathrm{Var} X_k = 1.$$

Hence

$$\mathbb{P}(X > 100) = \mathbb{P}(S_{80} - 80 \cdot 1 > 20) = \mathbb{P}\left(\frac{S_{80} - 80 \cdot 1}{1 \cdot \sqrt{80}} > \frac{20}{\sqrt{80}}\right).$$

Using the central limit theorem, we can approximate this as

$$\frac{1}{\sqrt{2\pi}} \int_{20/\sqrt{80}}^{\infty} e^{-x^2/2} dx \doteq 1.267\%.$$

This is fairly close to the true value, but not fantastic. We might expect the histogram correction to improve matters. This would mean

$$\mathbb{P}(X > 100) \approx \mathbb{P}(X > 99.5) = \mathbb{P}(S_{80} - 80 > 19.5) = \mathbb{P}\left(\frac{S_{80} - 80}{\sqrt{80}} > \frac{19.5}{\sqrt{80}}\right)$$

and the normal approximation would give

$$\frac{1}{\sqrt{2\pi}} \int_{19.5/\sqrt{80}}^{\infty} e^{-x^2/2} dx \doteq 1.462\%$$

which is actually a worse overestimate than the original answer was an underestimate. Nevertheless, both are pretty close.

27.1. **A Proof of the Central Limit Theorem.** Here we will give a non-quantitative proof of the theorem, using moment generating functions. Remember (though we have not proved it) that the distribution of X can be recovered from its moment generating function $M_X(t) = \mathbb{E}(e^{tX})$. The standard normal distribution has moment generating function

$$M_{N(0,1)}(t) = e^{t^2/2}.$$

So, one way to prove the central limit theorem is the following.

Theorem 27.3 (Central Limit Theorem)**.** *Let $X_1, X_2, X_3, \ldots$ be a sequence of independent, identically-distributed random variables. Suppose that their common moment generating function $M = M_{X_1}$ is C^2 at 0. Let $S_n = X_1 + \cdots + X_n$, let $\mu = \mathbb{E}(X_1)$, and $\sigma^2 = \mathrm{Var} X_1$. Let*

$$M_n(t) = M_{\frac{S_n - n\mu}{\sigma\sqrt{n}}}(t) = \mathbb{E}\left(e^{t\frac{S_n - n\mu}{\sigma\sqrt{n}}}\right).$$

Then

$$\lim_{n\to\infty} M_n(t) = M_{N(0,1)}(t) = e^{t^2/2}.$$

Proof. First, let's let $s = t/\sqrt{n}$; then

$$\begin{aligned} M_n(t) = \mathbb{E}\left(e^{s(S_n - n\mu)/\sigma}\right) &= \mathbb{E}\left(e^{s[(X_1-\mu)/\sigma + \cdots + (X_n-\mu)/\sigma]}\right) \\ &= \mathbb{E}\left(e^{s(X_1-\mu)/\sigma} \cdots e^{s(X_n-\mu)/\sigma}\right) \\ &= \mathbb{E}\left(e^{s(X_1-\mu)/\sigma}\right) \cdots \mathbb{E}\left(e^{s(X_n-\mu)/\sigma}\right) \end{aligned}$$

because of independence. Now, since each of the random variables $(X_1 - \mu)/\sigma, \ldots, (X_n - \mu)/\sigma$ has the same distribution, they all have the same moment generating function; so

$$M_n(t) = [M_{(X_1-\mu)/\sigma}(s)]^n.$$

Now, moment generating functions are always strictly positive, so we can take logarithms:

$$\ln M_n(t) = n \ln M_{(X_1-\mu)/\sigma}(s).$$

Let's define $M(s) = M_{(X_1-\mu)/\sigma}(s)$, and $L(s) = \ln M(s) = \ln M_{(X_1-\mu)/\sigma}(s)$. So L is the cumulant generating function of $(X_1 - \mu)/\sigma$, and it is also C^2 at 0. As we showed two weeks ago, this means that $L'(0) = \mathbb{E}((X_1 - \mu)/\sigma) = 0$, and $L''(0) = \mathrm{Var}((X_1 - \mu)/\sigma) = 1$. So we have

$$\ln M_n(t) = nL(s) = nL(t/\sqrt{n}).$$

We want to find $\lim_{n\to\infty} M_n(t)$; it suffices to find $\lim_{n\to\infty} \ln M_n(t)$. So we use L'Hôpital's rule (differentiating with respect to n).

$$\begin{aligned}
\lim_{n\to\infty} \ln M_n(t) &= \lim_{n\to\infty} \frac{L(t/\sqrt{n})}{1/n} \\
&= \lim_{n\to\infty} \frac{L'(t/\sqrt{n}) \cdot -\frac{t}{2} n^{-3/2}}{-\frac{1}{n^2}} \\
&= \frac{t}{2} \lim_{n\to\infty} \frac{L'(t/\sqrt{n})}{n^{-1/2}}.
\end{aligned}$$

The equality from the first line to the second is L'Hôpital's rule, which we can apply here since the top is tending to $L(0) = \ln M(0) = \ln 1 = 0$, and the bottom is also tending to 0. Now, the new limit we've found is a ratio where the top is tending to $L'(0) = 0$ we showed above, and the bottom is also tending to 0. So we can apply L'Hôpital's rule again, differentiating with respect to n:

$$\begin{aligned}
\lim_{n\to\infty} \ln M_n(t) &= \frac{t}{2} \lim_{n\to\infty} \frac{L'(t/\sqrt{n})}{n^{-1/2}} \\
&= \frac{t}{2} \lim_{n\to\infty} \frac{L''(t/\sqrt{n}) \cdot -\frac{t}{2} n^{-3/2}}{-\frac{1}{2} n^{-3/2}} \\
&= \frac{t^2}{2} \lim_{n\to\infty} L''(t/\sqrt{n}).
\end{aligned}$$

Since $L''(0) = 1$, we have thus proved that

$$\lim_{n\to\infty} \ln M_n(t) = \frac{t^2}{2},$$

which means that

$$\lim_{n\to\infty} M_n(t) = e^{t^2/2}$$

as desired.

□

28. LECTURE 28: DECEMBER 3, 2010

28.1. **The rate of convergence in the Central Limit Theorem.** We developed the CLT as a way of quantifying how quickly the sample mean (of a collection of independent random variables) converges to the expected value. But once we rescale appropriately (by $\frac{1}{\sqrt{n}}$, the central limit theorem becomes an enormously powerful approximation tool, as we've seen in many examples above. The question then arises: how good an approximation is it?

That is to say: we know that if $X_1, X_2, X_3, \ldots$ is a sequence of i.i.d. random variables with common mean μ and common standard deviation σ, then

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{S_n - n\mu}{\sigma\sqrt{n}} \leq x\right) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-t^2/2}\, dt = \mathbb{P}(\chi \leq x)$$

where χ is a standard normal random variable. But how fast does this limit converge? If we want to use it to approximate probabilities for finite n, we should know how large n needs to be in general to get the difference small:

$$\left|\mathbb{P}\left(\frac{S_n - n\mu}{\sigma\sqrt{n}} \leq x\right) - \mathbb{P}(\chi \leq x)\right| \leq \ ??(n)$$

The way to approach this question is to again use the relationship between probability and expectation. First, let's introduce some notation:

$$\overline{S}_n = \frac{S_n - n\mu}{\sigma\sqrt{n}}.$$

I.e. $\overline{S}_n$ is the *standardized* sum of $X_1, \ldots, X_n$. So we want to find a bound on

$$|\mathbb{P}(\overline{S}_n \leq x) - \mathbb{P}(\chi \leq x)|.$$

Well, this is the same as

$$|\mathbb{E}(\mathbb{1}_{(-\infty,x]}(\overline{S}_n)) - \mathbb{1}_{(-\infty,x]}(\chi)|$$

where, as usual,

$$\mathbb{1}_{(-\infty,x]}(y) = \begin{cases} 1, & y \leq x \\ 0, & y > x \end{cases}$$

So, in general, we need a way to find an upper bound on quantities of the form

$$\left|\mathbb{E}(f(\overline{S}_n) - \mathbb{E}(f(\chi))\right|$$

for functions $f\colon \mathbb{R} \to \mathbb{R}$. Now, the function $f = \mathbb{1}_{(-\infty,x]}$ is not a very nice function – it is discontinuous at x. But we can approximate it by nicer functions – even C^∞ functions. The following theorem, the *Berry-Esseen Theorem*, will be proved for smooth functions; using a careful approximation scheme afterward, the same holds for the function $f = \mathbb{1}_{(-\infty,x]}$, giving a worst-case scenario rate of convergence in the CLT.

Theorem 28.1 (Berry, Esseen, 1941)**.** *Let $X_1, X_2, X_3, \ldots, X_n, \ldots$ be a sequence of i.i.d. random variables, with common mean μ and common standard deviation σ, and common (assumed finite) third moment $\rho = \mathbb{E}(|X_1 - \mu|^3)$. Let $\overline{S}_n$ be the standardized sum, and let χ be a standard normal random variable. Let $f\colon \mathbb{R} \to \mathbb{R}$ be a C^3 function with bounded third derivative. There is a constant $C > 0$ such that, for all n,*

$$\left|\mathbb{E}(f(\overline{S}_n) - \mathbb{E}(f(\chi))\right| \leq \frac{C\rho}{\sigma^3\sqrt{n}}.$$

The constant C will depend on f. So, after we do the approximation, we want to know what is the best possible C for $f = \mathbb{1}_{(-\infty,x]}$. This is an active area of research. The original 1941 theorem gave $C \leq 7.59$. Over time, it has been shrunk. In fact, the most recent improvement was in 2009: it is now known that $C \leq 0.4875$. In 1956, Esseen proved that $C \geq 0.40973$, so we're closing in on the precise value.

Proof. The key idea is to use the fact that, if $\chi_1, \chi_2, \ldots, \chi_n$ are independent normal random variables

$$\chi_j \sim N(0, \tfrac{1}{\sqrt{n}})$$

then (as we calculated some weeks ago)

$$\chi_1 + \cdots + \chi_n \sim N(0,1) \sim \chi.$$

Since we only care about the distribution, we can assume freely that these variables $\chi_1, \ldots, \chi_n$ are also independent of $X_1, \ldots, X_n$. Now, $\overline{S}_n$ is also a sum of independent random variables:

$$\overline{S}_n = \frac{S_n - n\mu}{\sigma\sqrt{n}} = \frac{X_1 - \mu}{\sigma\sqrt{n}} + \cdots + \frac{X_n - \mu}{\sigma\sqrt{n}} \equiv Y_1 + \cdots + Y_n.$$

So, we want to estimate

$$\mathbb{E}[f(Y_1 + \cdots + Y_n)] - \mathbb{E}[f(\chi_1 + \cdot + \chi_n)].$$

The next (clever) idea is to write this as a telescoping sum, as follows.

$$\begin{aligned}
& \mathbb{E}[f(Y_1 + \cdots + Y_n)] - \mathbb{E}[f(\chi_1 + \cdots + \chi_n)] \\
= \; & \mathbb{E}[f(Y_1 + \cdots + Y_n)] - \mathbb{E}[f(\chi_1 + Y_2 + \cdots + Y_n)] \\
+ \; & \mathbb{E}[f(\chi_1 + Y_2 + \cdots + Y_n)] - \mathbb{E}[(f(\chi_1 + \chi_2 + Y_3 + \cdots + Y_n)] \\
+ \; & \mathbb{E}[f(\chi_1 + \chi_2 + Y_3 + \cdots + Y_n)] - \cdots \\
& \cdots \\
+ \; & \mathbb{E}[f(\chi_1 + \cdots + \chi_{n-2} + \chi_{n-1} + Y_n)] - \mathbb{E}[f(\chi_1 + \cdots + \chi_{n-2} + \chi_{n-1} + \chi_n)].
\end{aligned}$$

The reason to do this is that now we just need to compare each of n terms of the form

$$\mathbb{E}[f(\chi_1 + \cdots + \chi_{k-1} + Y_k + \cdots + Y_n)] - \mathbb{E}[f(\chi_1 + \cdots + \chi_k + Y_{k+1} + \cdots + Y_n)]. \tag{28.1}$$

If we let

$$U_k = \chi_1 + \cdots + \chi_{k-1} + 0 + Y_{k+1} + \cdots + Y_n$$

then the terms in 28.1 are

$$\mathbb{E}[f(U_k + Y_k) - f(U_k + \chi_k)].$$

To estimate this, we use Taylor's theorem:

$$f(x+h) = f(x) + f'(x)h + \frac{1}{2!}f''(x)h^2 + \frac{1}{3!}f'''(x_0)h^3$$

where x_0 is some point between 0 and x. Thus, letting $U = U_k$, $Y = Y_k$, and $\chi = \chi_k$,

$$f(U+Y) - f(U+\chi) = f'(U)(Y-\chi) + \frac{1}{2!}f''(U)(Y^2-\chi^2) + \frac{1}{3!}f'''(U_0)(Y^3-\chi^3)$$

where U_0 is some (random) point between 0 and U_0. Now, taking expectations, the first two terms are

$$\mathbb{E}[f'(U)(Y-\chi)] + \frac{1}{2}\mathbb{E}[f''(U)(Y^2-\chi^2)].$$

Since U_k is independent from both Y_k and χ_k, these terms become

$$\mathbb{E}[f'(U)]\mathbb{E}(Y-\chi) + \mathbb{E}[f''(U)]\mathbb{E}(Y^2-\chi^2).$$

But Y and χ have mean 0, so the first term is 0; also, Y and χ are both standardized, so

$$\mathbb{E}(Y^2) = \mathbb{E}\left[\left(\frac{X_n-\mu}{\sigma\sqrt{n}}\right)^2\right] = \frac{1}{n}\frac{1}{\sigma^2}\mathrm{Var}X_k = \frac{1}{n} = \mathbb{E}(\chi_k^2).$$

So both of these terms vanish! And so we have

$$\mathbb{E}[f(U+Y) - f(U+\chi)] = \frac{1}{3!}\mathbb{E}\left[f'''(U_0)(Y^3-\chi^3)\right].$$

Now, U_0 is some random point, so we don't know if it's independent from Y or χ. The best we can say is

$$\left|\mathbb{E}\left[f'''(U_0)(Y^3-\chi^3)\right]\right| \le \mathbb{E}\left[|f'''(U_0)|(|Y|^3+|\chi|^3)\right] \le M\mathbb{E}(|Y|^3+|\chi|^3)$$

where $M = \max_x |f'''(x)|$. Now, $\chi \sim N(0, \frac{1}{\sqrt{n}}) \sim \frac{1}{\sqrt{n}}N(0,1)$, and so

$$\mathbb{E}(|\chi|^2) = \frac{1}{n^{3/2}}\int_{-\infty}^{\infty}|x|^3\frac{1}{\sqrt{2\pi}}e^{-x^2/2}dx = \sqrt{\frac{8}{\pi}}\cdot\frac{1}{n^{3/2}}.$$

Also,

$$\mathbb{E}(|Y|^3) = \mathbb{E}\left[\left(\frac{|X_k-\mu|}{\sigma\sqrt{n}}\right)^3\right] = \frac{1}{\sigma^3 n^{3/2}}\mathbb{E}[|X_1-\mu|^3] = \frac{\rho}{\sigma^3}\frac{1}{n^{3/2}}.$$

Putting everything together, we see that $\mathbb{E}(f(\overline{S}_n) - \mathbb{E}(f(\chi))$ is a sum of n terms, each of which is

$$\le \frac{1}{n^{3/2}}\frac{M}{3!}\left(\sqrt{\frac{8}{\pi}} + \frac{\rho}{\sigma^3}\right).$$

So, adding them all up, we get

$$|\mathbb{E}[f(\overline{S}_n) - f(\chi)]| \le \frac{1}{n^{3/2}}\frac{M}{3!}\left(\sqrt{\frac{8}{\pi}} + \frac{\rho}{\sigma^3}\right)\cdot n.$$

Now, $\rho^{1/3} = \mathbb{E}[|X-\mathbb{E}(X)|^3]^{1/3} \ge \mathbb{E}[|X-\mathbb{E}(X)|^2]^{1/2} = \sigma$, and so $\rho \ge \sigma^3$, so $\rho/\sigma^3 \ge 1$. So we can simplify

$$|\mathbb{E}[f(\overline{S}_n) - f(\chi)]| \le \frac{1}{n^{1/2}}\frac{M}{3!}\left(\sqrt{\frac{8}{\pi}} + 1\right)\frac{\rho}{\sigma^3}.$$

This proves the theorem with $C = \frac{1}{6}\left(1+\sqrt{\frac{8}{\pi}}\right)M$. □