Undergraduate probability

These notes are based on *A First Course in Probability Theory*, 6th edition, by S. Ross.

1. Combinatorics.

The first basic principle is to multiply.

Suppose we have 4 shirts of 4 different colors and 3 pants of different colors. How many possibilities are there? For each shirt there are 3 possibilities, so altogether there are $4 \times 3 = 12$ possibilities.

Example. How many license plates of 3 letters followed by 3 numbers are possible?

Answer. $(26)^3(10)^3$, because there are 26 possibilities for the first place, 26 for the second, 26 for the third, 10 for the fourth, 10 for the fifth, and 10 for the sixth. We multiply.

How many ways can one arrange a, b, c? One can have

$$abc, \quad acb, \quad bac, \quad bca, \quad cab, \quad cba.$$

There are 3 possibilities for the first position. Once we have chosen the first position, there are 2 possibilities for the second position, and once we have chosen the first two possibilities, there is only 1 choice left for the third. So there are $3 \times 2 \times 1 = 3!$ arrangements. In general, if there are n letters, there are $n!$ possibilities.

Example. What is the number of possible batting orders with 9 players?

Answer. 9!

Example. How many ways can one arrange 4 math books, 3 chemistry books, 2 physics books, and 1 biology book on a bookshelf so that all the math books are together, all the chemistry books are together, and all the physics books are together.

Answer. 4!(4!3!2!1!). We can arrange the math books in 4! ways, the chemistry books in 3! ways, the physics books in 2! ways, and the biology book in $1! = 1$ way. But we also have to decide which set of books go on the left, which next, and so on. That is the same as the number of ways of arranging the letters M, C, P, B, and there are 4! ways of doing that.

How many ways can one arrange the letters a, a, b, c? Let us label them A, a, b, c. There are 4!, or 24, ways to arrange these letters. But we have repeats: we could have Aa or aA. So we have a repeat for each possibility, and so the answer should be $4!/2! = 12$. If there were 3 a's, 4 b's, and 2 c's, we would have

$$\frac{9!}{3!4!2!}.$$

Example. Suppose there are 4 Czech tennis players, 4 U.S. players, and 3 Russian players, in how many ways could they be arranged?

Answer. 11!/(4!4!3!).

What we just did was called the number of permutations.

Now let us look at what are known as combinations. How many ways can we choose 3 letters out of 5? If the letters are a, b, c, d, e and order matters, then there would be 5 for the first position, 4 for the second, and 3 for the third, for a total of $5 \times 4 \times 3$. But suppose the letters selected were a, b, c. If order

doesn't matter, we will have the letters a, b, c 6 times, because there are 3! ways of arranging 3 letters. The same is true for any choice of three letters. So we should have $5 \times 4 \times 3/3!$. We can rewrite this as

$$\frac{5 \cdot 4 \cdot 3}{3!} = \frac{5!}{3!2!}$$

This is often written $\binom{5}{3}$, read *"5 choose 3."* More generally,

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

Example. How many ways can one choose a committee of 3 out of 10 people?

Answer. $\binom{10}{3}$.

Example. Suppose there are 8 men and 8 women. How many ways can we choose a committee that has 2 men and 2 women?

Answer. We can choose 2 men in $\binom{8}{2}$ ways and 2 women in $\binom{8}{2}$ ways. The number of committees is then the product: $\binom{8}{2}\binom{8}{2}$.

Let us look at a few more complicated applications of combinations. One example is the binomial theorem:

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

To see this, the left hand side is

$$(x+y)(x+y)\cdots(x+y).$$

This will be the sum of 2^n terms, and each term will have n factors. How many terms have k x's and $n-k$ y's? This is the same as asking in a sequence of n positions, how many ways can one choose k of them in which to put x's? The answer is $\binom{n}{k}$, so the coefficient of $x^k y^{n-k}$ should be $\binom{n}{k}$.

One can derive some equalities involving these binomial coefficients by combinatorics. For example, let us argue that

$$\binom{10}{4} = \binom{9}{3} + \binom{9}{4}$$

without doing any algebra. Suppose we have 10 people, one of whom we decide is special, denoted A. $\binom{10}{4}$ represents the number of committees having 4 people out of the 10. Any such committee will either contain A or will not. $\binom{9}{3}$ is the number of committees that contain A and 3 out of the remaining 9 people, while $\binom{9}{4}$ is the number of committee that do not contain A and contain 4 out of the remaining 4 people.

Suppose one has 9 people and one wants to divide them into one committee of 3, one of 4, and a last of 2. There are $\binom{9}{3}$ ways of choosing the first committee. Once that is done, there are 6 people left and there are $\binom{6}{4}$ ways of choosing the second committee. Once that is done, the remainder must go in the third committee. So the answer is

$$\frac{9!}{3!6!}\frac{6!}{4!2!} = \frac{9!}{3!4!2!}.$$

In general, to divide n objects into one group of n_1, one group of n_2, ..., and a kth group of n_k, where $n = n_1 + \cdots + n_k$, the answer is

$$\frac{n!}{n_1! n_2! \cdots n_k!}.$$

These are known as multinomial coefficients.

Suppose one has 8 indistinguishable balls. How many ways can one put them in 3 boxes? Let us make sequences of o's and |'s; any such sequence that has | at each side, 2 other |'s, and 8 o's represents a way of arranging balls into boxes. For example, if one has

$$| \quad o \quad o \quad | \quad o \quad o \quad o \quad | \quad o \quad o \quad o \quad |,$$

this would represent 2 balls in the first box, 3 in the second, and 3 in the third. Altogether there are $8 + 4$ symbols, the first is a | as is the last. so there are 10 symbols that can be either | or o. Also, 8 of them must be o. How many ways out of 10 spaces can one pick 8 of them into which to put a o? The answer is $\binom{10}{8}$.

Example. How many quintuples $(x_1, x_2, x_3, x_4, x_5)$ of nonnegative integers whose sum is 20 are there?

Answer. This is the same as asking: how many nonnegative integer solutions are there to the equation $x_1 + x_2 + x_3 + x_4 + x_5 = 20$. View this as putting 20 balls in 5 boxes, with x_1 in the first box, x_2 in the second, and so on. So there are 20 o's, a | for the first and last spot in a sequence, and 4 other |'s. We can choose 20 spaces for the o's out of the 24 total in $\binom{24}{20}$ ways.

Example. Consider the same question as the example above, but where each x_i is at least 1.

Answer. First put one ball in each box. This leaves 15 balls to put in 5 boxes, and as above this can be done $\binom{19}{15}$ ways.

2. The probability set-up.

We will have a sample space, denoted S (sometimes Ω) that consists of all possible outcomes. For example, if we roll two dice, the sample space would be all possible pairs made up of the numbers one through six. An event is a subset of S.

Another example is to toss a coin 2 times, and let $S = \{HH, HT, TH, TT\}$; or to let S be the possible orders in which 5 horses finish in a horse race; or S the possible prices of some stock at closing time today; or $S = [0, \infty)$; the age at which someone dies; or S the points in a circle, the possible places a dart can hit.

We use the following usual notation: $A \cup B$ is the union of A and B and denotes the points of S that are in A or B or both. $A \cap B$ is the intersection of A and B and is the set of points that are in both A and B. $\emptyset$ denotes the empty set. $A \subset B$ means that A is contained in B and A^c is the complement of A, that is, the points in S that are not in A. We extend the definition to have $\cup_{i=1}^n A_i$ is the union of $A_1, \cdots, A_n$, and similarly $\cap_{i=1}^n A_i$.

An exercise is to show that $(\cup_{i=1}^n A_i)^c = \cap_{i=1}^n A_i^c$ and $(\cap_{i=1}^n A_i)^c = \cup_{i=1}^n A_i^c$. These are called DeMorgan's laws.

There are no restrictions on S. The collection of events, $\mathcal{F}$, must be a σ-field, which means that it satisfies the following:

(i) $\emptyset, S$ *is in* $\mathcal{F}$*;*
(ii) if A *is in* $\mathcal{F}$*, then* A^c *is in* $\mathcal{F}$*;*
(iii) if $A_1, A_2, \ldots$ *are in* $\mathcal{F}$*, then* $\cup_{i=1}^\infty A_i$ *and* $\cap_{i=1}^\infty A_i$ *are in* $\mathcal{F}$*.*

Typically we will take $\mathcal{F}$ to be all subsets of S, and so (i)-(iii) are automatically satisfied. The only times we won't have $\mathcal{F}$ be all subsets is for technical reasons or when we talk about conditional expectation.

So now we have a space S, a σ-field $\mathcal{F}$, and we need to talk about what a probability is. There are three axioms:

(1) $0 \leq \mathbb{P}(E) \leq 1$ *for all events* E.
(2) $\mathbb{P}(S) = 1$.
(3) If the E_i *are pairwise disjoint,* $\mathbb{P}(\cup_{i=1}^{\infty} E_i) = \sum_{i=1}^{\infty} \mathbb{P}(E_i)$.

Pairwise disjoint means that $E_i \cap E_j = \emptyset$ unless $i = j$.

Note that probabilities are probabilities of subsets of S, not of points of S. However it is common to write $\mathbb{P}(x)$ for $\mathbb{P}(\{x\})$.

Intuitively, the probability of E should be the number of times E occurs in n times, taking a limit as n tends to infinity. This is hard to use. It is better to start with these axioms, and then to prove that the probability of E is the limit as we hoped.

There are some easy consequences of the axioms.

Proposition 2.1. *(1)* $\mathbb{P}(\emptyset) = 0$.
(2) If $A_1, \ldots, A_n$ *are pairwise disjoint,* $\mathbb{P}(\cup_{i=1}^{n} A_i) = \sum_{i=1}^{n} \mathbb{P}(A_i)$.
(3) $\mathbb{P}(E^c) = 1 - \mathbb{P}(E)$.
(4) If $E \subset F$, *then* $\mathbb{P}(E) \leq \mathbb{P}(F)$.
(5) $\mathbb{P}(E \cup F) = \mathbb{P}(E) + \mathbb{P}(F) - \mathbb{P}(E \cap F)$.

Proof. For (1), let $A_i = \emptyset$ for each i. These are clearly disjoint, so $\mathbb{P}(\emptyset) = \mathbb{P}(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \mathbb{P}(A_i) = \sum_{i=1}^{\infty} \mathbb{P}(\emptyset)$. If $\mathbb{P}(\emptyset)$ were positive, then the last term would be infinity, contradicting the fact that probabilities are between 0 and 1. So the probability must be zero.

The second follows if we let $A_{n+1} = A_{n+2} = \cdots = \emptyset$. We still have pairwise disjointness and $\cup_{i=1}^{\infty} A_i = \cup_{i=1}^{n} A_i$, and $\sum_{i=1}^{\infty} \mathbb{P}(A_i) = \sum_{i=1}^{n} \mathbb{P}(A_i)$, using (1).

To prove (3), use $S = E \cup E^c$. By (2), $\mathbb{P}(S) = \mathbb{P}(E) + \mathbb{P}(E^c)$. By axiom (2), $\mathbb{P}(S) = 1$, so (1) follows.

To prove (4), write $F = E \cup (F \cap E^c)$, so $\mathbb{P}(F) = \mathbb{P}(E) + \mathbb{P}(F \cap E^c) \geq \mathbb{P}(E)$ by (2) and axiom (1).

Similarly, to prove (5), we have $\mathbb{P}(E \cup F) = \mathbb{P}(E) + \mathbb{P}(E^c \cap F)$ and $\mathbb{P}(F) = \mathbb{P}(E \cap F) + \mathbb{P}(E^c \cap F)$. Solving the second equation for $\mathbb{P}(E^c \cap F)$ and substituting in the first gives the desired result. □

It is very common for a probability space to consist of finitely many points, all with equally likely probabilities. For example, in tossing a fair coin, we have $S = \{H, T\}$, with $\mathbb{P}(H) = \mathbb{P}(T) = \frac{1}{2}$. Similarly, in rolling a fair die, the probability space consists of $\{1, 2, 3, 4, 5, 6\}$, each point having probability $\frac{1}{6}$.

Example. What is the probability that if we roll 2 dice, the sum is 7?

Answer. There are 36 possibilities, of which 6 have a sum of 7: $(1, 6), (2, 5), (3, 4), (4, 3), (5, 2), (6, 1)$. Since they are all equally likely, the probability is $\frac{6}{36} = \frac{1}{6}$.

Example. What is the probability that in a poker hand (5 cards out of 52) we get exactly 4 of a kind?

Answer. The probability of 4 aces and 1 king is $\binom{4}{4}\binom{4}{1} \Big/ \binom{52}{5}$. The probability of 4 jacks and one 3 is the same. There are 13 ways to pick the rank that we have 4 of and then 12 ways to pick the rank we have

one of, so the answer is

$$13 \cdot 12 \frac{\binom{4}{4}\binom{4}{1}}{\binom{52}{5}}.$$

Example. What is the probability that in a poker hand we get exactly 3 of a kind (and he other two cards are of different ranks)?

Answer. The probability of 3 aces, 1 king and 1 queen is $\binom{4}{3}\binom{4}{1}\binom{4}{1} \Big/ \binom{52}{5}$. We have 13 choices for the rank we have 3 of and $\binom{12}{2}$ choices for the other two ranks, so the answer is

$$13\binom{12}{2} \frac{\binom{4}{3}\binom{4}{1}\binom{4}{1}}{\binom{52}{5}}.$$

Example. In a class of 30 people, what is the probability everyone has a different birthday? (We assume each day is equally likely.)

Answer. Let the first person have a birthday on some day. The probability that the second person has a different birthday will be $\frac{364}{365}$. The probability that the third person has a different birthday from the first two people is $\frac{363}{365}$. So the answer is

$$1 - \frac{364}{365}\frac{363}{365} \cdots \frac{336}{365}.$$

Example. Suppose 10 people put a key into a hat and then withdraw one randomly. What is the probability at least one person gets his/her own key?

Answer. If E_i is the event that the ith person gets his/her own key, we want $\mathbb{P}(\cup_{i=1}^{10} E_i)$. One can show, either from a picture or an induction proof, that

$$\mathbb{P}(\cup_{i=1}^{10} E_i) = \sum_{i_1} \mathbb{P}(E_{i_1}) - \sum_{i_1<i_2} \mathbb{P}(E_{i_1} \cap E_{i_2}) + \sum_{i_1<i_2<i_3} \mathbb{P}(E_{i_1} \cap E_{i_2} \cap E_{i_3}) - \cdots.$$

Now the probability that at least the 1st, 3rd, 5th, and 7th person gets his or her own key is the number of ways the 2nd, 4th, 6th, 8th, 9th, and 10th person can choose a key out of 6, namely 6!, divided by the number of ways 10 people can each choose a key, namely 10!. so $\mathbb{P}(E_1 \cap E_3 \cap E_5 \cap E_7) = 6!/10!$. There are $\binom{10}{4}$ ways of selecting 4 people to have their own key out of 10, so

$$\sum_{i_1,i_2,i_3,i_4} \mathbb{P}(E_{i_1} \cap E_{i_2} \cap E_{i_3} \cap E_{i_4}) = \binom{10}{4} \frac{6!}{10!} = \frac{1}{4!}$$

The other terms are similar, and the answer is

$$\frac{1}{1!} - \frac{1}{2!} + \frac{1}{3!} - \cdots - \frac{1}{10!} \approx 1 - e^{-1}.$$

3. Conditional probability.

Suppose there are 200 men, of which 100 are smokers, and 100 women, of which 20 are smokers. What is the probability that a person chosen at random will be a smoker? The answer is 120/300. Now, let us ask, what is the probability that a person chosen at random is a smoker given that the person is a women? One would expect the answer to be 20/100 and it is.

What we have computed is

$$\frac{\text{number of women smokers}}{\text{number of women}} = \frac{\text{number of women smokers}/300}{\text{number of women}/300},$$

which is the same as the probability that a person chosen at random is a woman and a smoker divided by the probability that a person chosen at random is a woman.

With this in mind, we make the following definition. If $\mathbb{P}(F) > 0$, we define

$$\mathbb{P}(E \mid F) = \frac{\mathbb{P}(E \cap F)}{\mathbb{P}(F)}.$$

$\mathbb{P}(E \mid F)$ is read "the probability of E given F."

Note $\mathbb{P}(E \cap F) = \mathbb{P}(E \mid F)\mathbb{P}(F)$.

Suppose you roll two dice. What is the probability the sum is 8? There are five ways this can happen: $(2,6), (3,5), (4,4), (5,3), (6,2)$, so the probability is 5/36. Let us call this event A. What is the probability that the sum is 8 given that the first die shows a 3? Let B be the event that the first die shows a 3. Then $\mathbb{P}(A \cap B)$ is the probability that the first die shows a 3 and the sum is 8, or 1/36. $\mathbb{P}(B) = 1/6$, so $\mathbb{P}(A \mid B) = \frac{1/36}{1/6} = 1/6$.

Example. Suppose a box has 3 red marbles and 2 black ones. We select 2 marbles. What is the probability that second marble is red given that the first one is red?

Answer. Let A be the event the second marble is red, and B the event that the first one is red. $\mathbb{P}(B) = 3/5$, while $\mathbb{P}(A \cap B)$ is the probability both are red, or is the probability that we chose 2 red out of 3 and 0 black out of 2. The $\mathbb{P}(A \cap B) = \binom{3}{2}\binom{2}{0} \Big/ \binom{5}{2}$. Then $\mathbb{P}(A \mid B) = \frac{3/10}{3/5} = 1/2$.

Example. A family has 2 children. Given that one of the children is a boy, what is the probability that the other child is also a boy?

Answer. Let B be the event that one child is a boy, and A the event that both children are boys. The possibilities are bb, bg, gb, gg, each with probability 1/4. $\mathbb{P}(A \cap B) = \mathbb{P}(bb) = 1/4$ and $\mathbb{P}(B) = \mathbb{P}(bb, bg, gb) = 3/4$. So the answer is $\frac{1/4}{3/4} = 1/3$.

Example. Suppose the test for HIV is 98% accurate in both directions and 0.5% of the population is HIV positive. If someone tests positive, what is the probability they actually are HIV positive?

Let D mean HIV positive, and T mean tests positive.

$$\mathbb{P}(D \mid T) = \frac{\mathbb{P}(D \cap T)}{\mathbb{P}(T)} = \frac{(.98)(.005)}{(.98)(.005) + (.02)(.995)} = 19.8\%.$$

Suppose you know $\mathbb{P}(E \mid F)$ and you want $\mathbb{P}(F \mid E)$.

Example. Suppose 36% of families own a dog, 30% of families own a cat, and 22% of the families that have a dog also have a cat. A family is chosen at random and found to have a cat. What is the probability they also own a dog?

Answer. Let D be the families that own a dog, and C the families that own a cat. We are given $\mathbb{P}(D) = .36, \mathbb{P}(C) = .30, \mathbb{P}(C \mid D) = .22$ We want to know $\mathbb{P}(D \mid C)$. We know $\mathbb{P}(D \mid C) = \mathbb{P}(D \cap C)/\mathbb{P}(C)$. To find the numerator, we use $\mathbb{P}(D \cap C) = \mathbb{P}(C \mid D)\mathbb{P}(D) = (.22)(.36) = .0792$. So $\mathbb{P}(D \mid C) = .0792/.3 = .264 = 26.4\%$.

Example. Suppose 30% of the women in a class received an A on the test and 25% of the men received an A. The class is 60% women. Given that a person chosen at random received an A, what is the probability this person is a women?

Answer. Let A be the event of receiving an A, W be the event of being a woman, and M the event of being a man. We are given $\mathbb{P}(A \mid W) = .30, \mathbb{P}(A \mid M) = .25, \mathbb{P}(W) = .60$ and we want $\mathbb{P}(W \mid A)$. From the definition

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

As in the previous example,

$$\mathbb{P}(W \cap A) = \mathbb{P}(A \mid W)\mathbb{P}(W) = (.30)(.60) = .18.$$

To find $\mathbb{P}(A)$, we write

$$\mathbb{P}(A) = \mathbb{P}(W \cap A) + \mathbb{P}(M \cap A).$$

Since the class is 40% men,

$$\mathbb{P}(M \cap A) = \mathbb{P}(A \mid M)\mathbb{P}(M) = (.25)(.40) = .10.$$

So

$$\mathbb{P}(A) = \mathbb{P}(W \cap A) + \mathbb{P}(M \cap A) = .18 + .10 = .28.$$

Finally,

$$\mathbb{P}(W \mid A) = \frac{\mathbb{P}(W \cap A)}{\mathbb{P}(A)} = \frac{.18}{.28}.$$

To get a general formula, we can write

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

This formula is known as Bayes' rule.

Suppose $\mathbb{P}(E \mid F) = \mathbb{P}(E)$, i.e., knowing F doesn't help in predicting E. Then E and F are independent. What we have said is that in this case

$$\mathbb{P}(E \mid F) = \frac{\mathbb{P}(E \cap F)}{\mathbb{P}(F)} = \mathbb{P}(E),$$

or $\mathbb{P}(E \cap F) = \mathbb{P}(E)\mathbb{P}(F)$. We use the latter equation as a definition:

We say E and F are independent if

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

Example. Suppose you flip two coins. The outcome of heads on the second is independent of the outcome of tails on the first. To be more precise, if A is tails for the first coin and B is heads for the second, and we assume we have fair coins (although this is not necessary), we have $\mathbb{P}(A \cap B) = \frac{1}{4} = \frac{1}{2} \cdot \frac{1}{2} = \mathbb{P}(A)\mathbb{P}(B)$.

Example. Suppose you draw a card from an ordinary deck. Let E be you drew an ace, F be that you drew a spade. Here $\frac{1}{52} = \mathbb{P}(E \cap F) = \frac{1}{13} \cdot \frac{1}{4} = \mathbb{P}(E) \cap \mathbb{P}(F)$.

Proposition 3.1. *If E and F are independent, then E and F^c are independent.*

Proof.

$$\mathbb{P}(E \cap F^c) = \mathbb{P}(E) - \mathbb{P}(E \cap F) = \mathbb{P}(E) - \mathbb{P}(E)\mathbb{P}(F) = \mathbb{P}(E)[1 - \mathbb{P}(F)] = \mathbb{P}(E)\mathbb{P}(F^c).$$

□

We say E, F, and G are independent if E and F are independent, E and G are independent, F and G are independent, and $\mathbb{P}(E \cap F \cap G) = \mathbb{P}(E)\mathbb{P}(F)\mathbb{P}(G)$.

Example. Suppose you roll two dice, E is that the sum is 7, F that the first is a 4, and G that the second is a 3. E and F are independent, as are E and G and F and G, but E, F and G are not.

Example. What is the probability that exactly 3 threes will show if you roll 10 dice?

Answer. The probability that the 1st, 2nd, and 4th dice will show a three and the other 7 will not is $\frac{1}{6}^3\frac{5}{6}^7$. Independence is used here: the probability is $\frac{1}{6}\frac{1}{6}\frac{5}{6}\frac{1}{6}\frac{5}{6}\cdots\frac{5}{6}$. The probability that the 4th, 5th, and 6th dice will show a three and the other 7 will not is the same thing. So to answer our original question, we take $\frac{1}{6}^3\frac{5}{6}^7$ and multiply it by the number of ways of choosing 3 dice out of 10 to be the ones showing threes. There are $\binom{10}{3}$ ways of doing that.

This is a particular example of what are known as Bernoulli trials or the binomial distribution.

Suppose you have n independent trials, where the probability of a success is p. The the probability there are k successes is the number of ways of putting k objects in n slots (which is $\binom{n}{k}$) times the probability that there will be k successes and $n-k$ failures in exactly a given order. So the probability is $\binom{n}{k}p^k(1-p)^{n-k}$.

A problem that comes up in actuarial science frequently is gambler's ruin.

Example. Suppose you toss a fair coin repeatedly and independently. If it comes up heads, you win a dollar, and if it comes up tails, you lose a dollar. Suppose you start with \$50. What's the probability you will get to \$200 before you go broke?

Answer. It is easier to solve a slightly harder problem. Let $y(x)$ be the probability you get to 200 before 0 if you start with x dollars. Clearly $y(0) = 0$ and $y(200) = 1$. If you start with x dollars, then with probability $\frac{1}{2}$ you get a heads and will then have $x+1$ dollars. With probability $\frac{1}{2}$ you get a tails and will then have $x-1$ dollars. So we have

$$y(x) = \tfrac{1}{2}y(x+1) + \tfrac{1}{2}y(x-1).$$

Multiplying by 2, and subtracting $y(x) + y(x-1)$ from each side, we have

$$y(x+1) - y(x) = y(x) - y(x-1).$$

This says succeeding slopes of the graph of $y(x)$ are constant (remember that x must be an integer). In other words, $y(x)$ must be a line. Since $y(0) = 0$ and $y(200) = 1$, we have $y(x) = x/200$, and therefore $y(50) = 1/4$.

Example. Suppose we are in the same situation, but you are allowed to go arbitrarily far in debt. Let $y(x)$ be the probability you ever get to \$200. What is a formula for $y(x)$?

Answer. Just as above, we have the equation $y(x) = \frac{1}{2}y(x+1) + \frac{1}{2}y(x-1)$. This implies $y(x)$ is linear, and as above $y(200) = 1$. Now the slope of y cannot be negative, or else we would have $y > 1$ for some x and

that is not possible. Neither can the slope be positive, or else we would have $y < 0$, and again this is not possible, because probabilities must be between 0 and 1. Therefore the slope must be 0, or $y(x)$ is constant, or $y(x) = 1$ for all x. In other words, one is certain to get to \$200 eventually (provided, of course, that one is allowed to go into debt). There is nothing special about the figure 300. Another way of seeing this is to compute as above the probability of getting to 200 before $-M$ and then letting $M \to \infty$.

4. Random variables.

A random variable is a real-valued function on S. Random variables are usually denoted by $X, Y, Z, \ldots$

Example. If one rolls a die, let X denote the outcome (i.e., either 1,2,3,4,5,6).

Example. If one rolls a die, let Y be 1 if an odd number is showing and 0 if an even number is showing.

Example. If one tosses 10 coins, let X be the number of heads showing.

Example. In n trials, let X be the number of successes.

A discrete random variable is one that can only take countably many values. For a discrete random variable, we define the probability mass function or the density by $p(x) = \mathbb{P}(X = x)$. Here $\mathbb{P}(X = x)$ is an abbreviation for $\mathbb{P}(\{\omega \in S : X(\omega) = x\})$. This type of abbreviation is standard. Note $\sum_i p(x_i) = 1$ since X must equal something.

Let X be the number showing if we roll a die. The expected number to show up on a roll of a die should be $1 \cdot \mathbb{P}(X = 1) + 2 \cdot \mathbb{P}(X = 2) + \cdots + 6 \cdot \mathbb{P}(X = 6) = 3.5$. More generally, we define

$$\mathbb{E}\, X = \sum_{\{x : p(x) > 0\}} x p(x)$$

to be the expected value or expectation or mean of X.

Example. If we toss a coin and X is 1 if we have heads and 0 if we have tails, what is the expectation of X?

Answer.

$$p_X(x) = \begin{cases} \frac{1}{2} & x = 1 \\ \frac{1}{2} & x = 0 \\ 0 & \text{all other values of } x. \end{cases}$$

Hence $\mathbb{E}\, X = (1)(\frac{1}{2}) + (0)(\frac{1}{2}) = \frac{1}{2}$.

Example. Suppose $X = 0$ with probability $\frac{1}{2}$, 1 with probability $\frac{1}{4}$, 2 with probability $\frac{1}{8}$, and more generally n with probability $1/2^n$. This is an example where X can take infinitely many values (although still countably many values). What is the expectation of X?

Answer. Here $p_X(n) = 1/2^n$ if n is a nonnegative integer and 0 otherwise. So

$$\mathbb{E}\, X = (0)\tfrac{1}{2} + (1)\tfrac{1}{4} + (2)\tfrac{1}{8} + (3)\tfrac{1}{16} + \cdots.$$

This turns out to sum to 1. To see this, recall the formula for a geometric series:

$$1 + x + x^2 + x^3 + \cdots = \frac{1}{1 - x}.$$

If we differentiate this, we get

$$1 + 2x + 3x^2 + \cdots = \frac{1}{(1 - x)^2}.$$

We have

$$\begin{aligned}\mathbb{E}\,X &= 1(\tfrac{1}{4}) + 2(\tfrac{1}{8}) + 3(\tfrac{1}{16} + \cdots \\ &= \tfrac{1}{4}\Big[1 + 2(\tfrac{1}{2}) + 3(\tfrac{1}{4}) + \cdots\Big] \\ &= \tfrac{1}{4}\frac{1}{(1-\frac{1}{2})^2} = 1.\end{aligned}$$

Example. Suppose we roll a fair die. If 1 or 2 is showing, let $X = 3$; if a 3 or 4 is showing, let $X = 4$, and if a 5 or 6 is showing, let $X = 10$. What is $\mathbb{E}\,X$?

Answer. We have $\mathbb{P}(X = 3) = \mathbb{P}(X = 4) = \mathbb{P}(X = 10) = \frac{1}{3}$, so

$$\mathbb{E}\,X = \sum x\mathbb{P}(X = x) = (3)(\tfrac{1}{3}) + (4)(\tfrac{1}{3}) + (10)(\tfrac{1}{3}) = \tfrac{17}{3}.$$

We want to give a second definition of $\mathbb{E}\,X$. We set

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

Remember we are only working with discrete random variables here.

In the example we just gave, we have $S = \{1, 2, 3, 4, 5, 6\}$ and $X(1) = 3, X(2) = 3, X(3) = 4, X(4) = 4, X(5) = 10, X(6) = 10$, and each ω has probability $\frac{1}{6}$. So using the second definition,

$$\mathbb{E}\,X = 3(\tfrac{1}{6}) + 3(\tfrac{1}{6}) + 4(\tfrac{1}{6}) + 4(\tfrac{1}{6}) + 10(\tfrac{1}{6}) + 10(\tfrac{1}{6}) = \tfrac{34}{6} = \tfrac{17}{3}.$$

We see that the difference between the two definitions is that we write, for example, $3\mathbb{P}(X = 3)$ as one of the summands in the first definition, while in the second we write this as $3\mathbb{P}(X = 1) + 3\mathbb{P}(X = 2)$.

Let us give a proof that the two definitions are equivalent.

Proposition 4.1. *If X is a discrete random variable and S is countable, then the two definitions are equivalent.*

Proof. Starting with the first definition, we have

$$\sum_x x\mathbb{P}(X = x) = \sum_x x \sum_{\{\omega : X(\omega) = x\}} \mathbb{P}(\omega).$$

This is because the set $(X = x)$ is the union of the disjoint sets $\{\omega\}$, where the union is over those ω for which $X(\omega) = x$. Bringing the x inside the sum, and using the fact that $X(\omega) = x$, our expression is equal to

$$\sum_x \sum_{\{\omega : X(\omega) = x\}} x\mathbb{P}(\omega) = \sum_x \sum_{\{\omega : X(\omega) = x\}} X(\omega)\mathbb{P}(\omega).$$

Since the union of $\{\omega : X(\omega) = x\}$ over all possible values of x is just all ω, we then have

$$\sum_\omega X(\omega)\mathbb{P}(\omega),$$

which is the second definition. □

One advantage of the second definition is that linearity is easy. We write

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

Similarly we have $\mathbb{E}(cX) = c\mathbb{E}X$ if c is a constant. These linearity results are quite hard using the first definition.

It turns out there is a formula for the expectation of random variables like X^2 and e^X. To see how this works, let us first look at an example.

Suppose we roll a die and let X be the value that is showing. We want the expectation $\mathbb{E}X^2$. Let $Y = X^2$, so that $\mathbb{P}(Y = 1) = \frac{1}{6}$, $\mathbb{P}(Y = 4) = \frac{1}{6}$, etc. and

$$\mathbb{E}X^2 = \mathbb{E}Y = (1)\tfrac{1}{6} + (4)\tfrac{1}{6} + \cdots + (36)\tfrac{1}{6}.$$

We can also write this as

$$\mathbb{E}X^2 = (1^2)\tfrac{1}{6} + (2^2)\tfrac{1}{6} + \cdots + (6^2)\tfrac{1}{6},$$

which suggests that a formula for $\mathbb{E}X^2$ is $\sum_x x^2\mathbb{P}(X = x)$. This turns out to be correct.

The only possibility where things could go wrong is if more than one value of X leads to the same value of X^2. For example, suppose $\mathbb{P}(X = -2) = \frac{1}{8}, \mathbb{P}(X = -1) = \frac{1}{4}, \mathbb{P}(X = 1) = \frac{3}{8}, \mathbb{P}(X = 2) = \frac{1}{4}$. Then if $Y = X^2$, $\mathbb{P}(Y = 1) = \frac{5}{8}$ and $\mathbb{P}(Y = 4) = \frac{3}{8}$. Then

$$\mathbb{E}X^2 = (1)\tfrac{5}{8} + (4)\tfrac{3}{8} = (-1)^2\tfrac{1}{4} + (1)^2\tfrac{3}{8} + (-2)^2\tfrac{1}{8} + (2)^2\tfrac{1}{4}.$$

So even in this case $\mathbb{E}X^2 = \sum_x x^2\mathbb{P}(X = x)$.

Theorem 4.2. $\mathbb{E}g(X) = \sum g(x)p(x)$.

Proof. Let $Y = g(X)$. Then

$$\mathbb{E}Y = \sum_y y\mathbb{P}(Y = y) = \sum_y y \sum_{\{x:g(x)=y\}} \mathbb{P}(X = x) = \sum_x g(x)\mathbb{P}(X = x).$$

□

Example. $\mathbb{E}X^2 = \sum x^2 p(x)$.

$\mathbb{E}X^n$ is called the nth moment of X.

If $M = \mathbb{E}X$, then

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

is called the variance of X. The square root of $\mathrm{Var}(X)$ is the standard deviation of X.

The variance measures how much spread there is about the expected value.

Example. We toss a fair coin and let $X = 1$ if we get heads, $X = -1$ if we get tails. Then $\mathbb{E}X = 0$, so $X - \mathbb{E}X = X$, and then $\mathrm{Var}\,X = \mathbb{E}X^2 = (1)^2\frac{1}{2} + (-1)^2\frac{1}{2} = 1$.

Example. We roll a die and let X be the value that shows. We have previously calculated $\mathbb{E}X = \frac{7}{2}$. So $X - \mathbb{E}X$ equals $-\frac{5}{2}, -\frac{3}{2}, -\frac{1}{2}, \frac{1}{2}, \frac{3}{2}, \frac{5}{2}$, each with probability $\frac{1}{6}$. So

$$\mathrm{Var}\,X = (-\tfrac{5}{2})^2\tfrac{1}{6} + (-\tfrac{3}{2})^2\tfrac{1}{6} + (-\tfrac{1}{2})^2\tfrac{1}{6} + (\tfrac{1}{2})^2\tfrac{1}{6} + (\tfrac{3}{2})^2\tfrac{1}{6} + (\tfrac{5}{2})^2\tfrac{1}{6} = \tfrac{35}{12}.$$

Note that the expectation of a constant is just the constant. An alternate expression for the variance is

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

5. Some discrete distributions.

Bernoulli. A r.v. X such that $\mathbb{P}(X=1)=p$ and $\mathbb{P}(X=0)=1-p$ is said to be a Bernoulli r.v. with parameter p. Note $\mathbb{E}\,X=p$ and $\mathbb{E}\,X^2=p$, so $\operatorname{Var}X=p-p^2=p(1-p)$.

Binomial. A r.v. X has a binomial distribution with parameters n and p if $\mathbb{P}(X=k)=\binom{n}{k}p^k(1-p)^{n-k}$. The number of successes in n trials is a binomial. After some cumbersome calculations one can derive $\mathbb{E}\,X=np$. An easier way is to realize that if X is binomial, then $X=Y_1+\cdots+Y_n$, where the Y_i are independent Bernoulli's, so $\mathbb{E}\,X=\mathbb{E}\,Y_1+\cdots+\mathbb{E}\,Y_n=np$. We haven't defined what it means for r.v.'s to be independent, but here we mean that the events $(Y_k=1)$ are independent. The cumbersome way is as follows.

$$\begin{aligned}\mathbb{E}\,X&=\sum_{k=0}^{n}k\binom{n}{k}p^k(1-p)^{n-k}=\sum_{k=1}^{n}k\binom{n}{k}p^k(1-p)^{n-k}\\&=\sum_{k=1}^{n}k\frac{n!}{k!(n-k)!}p^k(1-p)^{n-k}=np\sum_{k=1}^{n}\frac{(n-1)!}{(k-1)!((n-1)-(k-1))!}p^{k-1}(1-p)^{(n-1)-(k-1)}\\&=np\sum_{k=0}^{n-1}\frac{(n-1)!}{k!((n-1)-k)!}p^k(1-p)^{(n-1)-k}=np\sum_{k=0}^{n-1}\binom{n-1}{k}p^k(1-p)^{(n-1)-k}=np.\end{aligned}$$

To get the variance of X, we have

$$\mathbb{E}\,X^2=\sum_{k=1}^{n}\mathbb{E}\,Y_k^2+\sum_{i\neq j}\mathbb{E}\,Y_iY_j.$$

Now

$$\mathbb{E}\,Y_iY_j=1\cdot\mathbb{P}(Y_iY_j=1)+0\cdot\mathbb{P}(Y_iY_j=0)=\mathbb{P}(Y_i=1,Y_j=1)=\mathbb{P}(Y_i=1)\mathbb{P}(Y_j=1)=p^2$$

using independence. The square of $Y_1+\cdots+Y_n$ yields n^2 terms, of which n are of the form Y_k^2. So we have n^2-n terms of the form Y_iY_j with $i\neq j$. Hence

$$\operatorname{Var}X=\mathbb{E}\,X^2-(\mathbb{E}\,X)^2=np+(n^2-n)p^2-(np)^2=np(1-p).$$

Later we will see that the variance of the sum of independent r.v.'s is the sum of the variances, so we could quickly get $\operatorname{Var}X=np(1-p)$. Alternatively, one can compute $\mathbb{E}\,(X^2)-\mathbb{E}\,X=\mathbb{E}\,(X(X-1))$ using binomial coefficients and derive the variance of X from that.

Poisson. X is Poisson with parameter λ if

$$\mathbb{P}(X=i)=e^{-\lambda}\frac{\lambda^i}{i!}.$$

Note $\sum_{i=0}^{\infty}\lambda^i/i!=e^{\lambda}$, so the probabilities add up to one.

To compute expectations,

$$\mathbb{E}\,X=\sum_{i=0}^{\infty}ie^{-\lambda}\frac{\lambda^i}{i!}=e^{-\lambda}\lambda\sum_{i=1}^{\infty}\frac{\lambda^{i-1}}{(i-1)!}=\lambda.$$

Similarly one can show that

$$\mathbb{E}\,(X^2)-\mathbb{E}\,X=\mathbb{E}\,X(X-1)=\sum_{i=0}^{\infty}i(i-1)e^{-\lambda}\frac{\lambda^i}{i!}=\lambda^2e^{-\lambda}\sum_{i=2}^{\infty}\frac{\lambda^{i-2}}{(i-2)!}=\lambda^2,$$

so $\mathbb{E}\,X^2 = \mathbb{E}\,(X^2 - X) + |EX = \lambda^2 + \lambda$, and hence $\operatorname{Var} X = \lambda$.

Example. Suppose on average there are 5 homicides per month in a given city. What is the probability there will be at most 1 in a certain month?

Answer. If X is the number of homicides, we are given that $\mathbb{E}\,X = 5$. Since the expectation for a Poisson is λ, then $\lambda = 5$. Therefore $\mathbb{P}(X = 0) + \mathbb{P}(X = 1) = e^{-5} + 5e^{-5}$.

Example. Suppose on average there is one large earthquake per year in California. What's the probability that next year there will be exactly 2 large earthquakes?

Answer. $\lambda = \mathbb{E}\,X = 1$, so $\mathbb{P}(X = 2) = e^{-1}(\frac{1}{2})$.

We have the following proposition.

Proposition 5.1. *If X_n is binomial with parameters n and p_n and $np_n \to \lambda$, then $\mathbb{P}(X_n = i) \to \mathbb{P}(Y = i)$, where Y is Poisson with parameter λ.*

The above proposition shows that the Poisson distribution models binomials when the probability of a success is small. The number of misprints on a page, the number of automobile accidents, the number of people entering a store, etc. can all be modeled by Poissons.

Proof. For simplicity, let us suppose $\lambda = np_n$. In the general case we use $\lambda_n = np_n$. We write

$$\begin{aligned}\mathbb{P}(X_n = i) &= \frac{n!}{i!(n-i)!} p_n^i (1-p_n)^{n-i} \\ &= \frac{n(n-1)\cdots(n-i+1)}{i!} \Big(\frac{\lambda}{n}\Big)^i \Big(1 - \frac{\lambda}{n}\Big)^{n-i} \\ &= \frac{n(n-1)\cdots(n-i+1)}{n^i} \frac{\lambda^i}{i!} \frac{(1-\lambda/n)^n}{(1-\lambda/n)^i}.\end{aligned}$$

The first factor tends to 1 as $n \to \infty$. $(1 - \lambda/n)^i \to 1$ as $n \to \infty$ and $(1 - \lambda/n)^n \to e^{-\lambda}$ as $n \to \infty$. □

Uniform. Let $\mathbb{P}(X = k) = \frac{1}{n}$ for $k = 1, 2, \ldots, n$. This is the distribution of the number showing on a die (with $n = 6$), for example.

Geometric. Here $\mathbb{P}(X = i) = (1-p)^{i-1}p$ for $i = 1, 2, \ldots$. In Bernoulii trials, if we let X be the first time we have a success, then X will be geometric. For example, if we toss a coin over and over and X is the first time we get a heads, then X will have a geometric distribution. To see this, to have the first success occur on the k^{th} trial, we have to have $k-1$ failures in the first $k-1$ trials and then a success. The probability of that is $(1-p)^{k-1}p$. Since $\sum_{n=0}^{\infty} nr^n = 1/(1-r)^2$ (differentiate the formula $\sum r^n = 1/(1-r)$), we see that $\mathbb{E}\,X = 1/p$. Similarly we have $\operatorname{Var} X = (1-p)/p^2$.

Negative binomial. Let r and p be parameters and set

$$\mathbb{P}(X = n) = \binom{n-1}{r-1} p^r (1-p)^{n-r}, \qquad n = r, r+1, \ldots.$$

A negative binomial represents the number of trials until r successes. To get the above formula, to have the r^{th} success in the n^{th} trial, we must exactly have $r-1$ successes in the first $n-1$ trials and then a success in the n^{th} trial.

Hypergeometric. Set

$$\mathbb{P}(X = i) = \frac{\binom{m}{i}\binom{N-m}{n-i}}{\binom{N}{n}}.$$

This comes up in sampling without replacement: if there are N balls, of which m are one color and the other $N-m$ are another, and we choose n balls at random without replacement, then X represents the probability of having i balls of the first color.

6. Continuous distributions.

A r.v. X is said to have a continuous distribution if there exists a nonnegative function f such that

$$\mathbb{P}(a \le X \le b) = \int_a^b f(x)dx$$

for every a and b. (More precisely, such an X is said to have an absolutely continuous distribution.) f is called the density function for X. Note $\int_{-\infty}^{\infty} f(x)dx = \mathbb{P}(-\infty < X < \infty) = 1$. In particular, $\mathbb{P}(X = a) = \int_a^a f(x)dx = 0$ for every a.

Example. Suppose we are given $f(x) = c/x^3$ for $x \ge 1$. Since $\int_{-\infty}^{\infty} f(x)dx = 1$ and

$$c\int_{-\infty}^{\infty} f(x)dx = c\int_1^{\infty} \frac{1}{x^3}dx = \frac{c}{2},$$

we have $c = 2$.

Define $F(y) = \mathbb{P}(-\infty < X \le y) = \int_{-\infty}^{y} f(x)dx$. F is called the distribution function of X. We can define F for any random variable, not just continuous ones, by setting $F(y) = \mathbb{P}(X \le y)$. In the case of discrete random variables, this is not particularly useful, although it does serve to unify discrete and continuous random variables. In the continuous case, the fundamental theorem of calculus tells us, provided f satisfies some conditions, that

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

By analogy with the discrete case, we define the expectation by

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

In the example above,

$$\mathbb{E}\,X = \int_1^{\infty} x\frac{2}{x^3}dx = 2\int_1^{\infty} x^{-2}dx = 2.$$

We give another definition of the expectation in the continuous case. First suppose X is nonnegative and bounded above by a constant M. Define $X_n(\omega)$ to be $k/2^n$ if $k/2^n \le X(\omega) < (k+1)/2^n$. We are approximating X from below by the largest multiple of 2^{-n}. Each X_n is discrete and the X_n increase to X. We define $\mathbb{E}\,X = \lim_{n\to\infty} \mathbb{E}\,X_n$.

Let us argue that this agrees with the first definition in this case. We have

$$\begin{aligned}\mathbb{E}\,X_n &= \sum_{k/2^n} \frac{k}{2^n}\mathbb{P}(X_n = k/2^n) = \sum_{k/2^n} \frac{k}{2^n}\mathbb{P}(k/2^n \le X < (k+1)/2^n) \\ &= \sum \frac{k}{2^n}\int_{k/2^n}^{(k+1)/2^n} f(x)dx = \sum \int_{k/2^n}^{(k+1)/2^n} \frac{k}{2^n}f(x)dx.\end{aligned}$$

If $x \in [k/2^n, (k+1)/2^n)$, then x differs from $k/2^n$ by at most $1/2^n$. So the last integral differs from

$$\sum \int_{k/2^n}^{(k+1)/2^n} xf(x)dx$$

by at most $\sum (1/2^n)\mathbb{P}(k/2^n \le X < (k+1)/2^n) \le 1/2^n$, which goes to 0 as $n \to \infty$. On the other hand,

$$\sum \int_{k/2^n}^{(k+1)/2^n} xf(x)dx = \int_0^M xf(x)dx,$$

which is how we defined the expectation of X.

We will not prove the following, but it is an interesting exercise: if X_m is any sequence of discrete random variables that increase up to X, then $\lim_{m\to\infty} \mathbb{E}\, X_m$ will have the same value $\mathbb{E}\, X$.

To show linearity, if X and Y are bounded positive random variables, then take X_m discrete increasing up to X and Y_m discrete increasing up to Y. Then $X_m + Y_m$ is discrete and increases up to $X + Y$, so we have

$$\mathbb{E}\,(X+Y) = \lim_{m\to\infty} \mathbb{E}\,(X_m + Y_m) = \lim_{m\to\infty} \mathbb{E}\, X_m + \lim_{m\to\infty} \mathbb{E}\, Y_m = \mathbb{E}\, X + \mathbb{E}\, Y.$$

If X is not bounded or not necessarily positive, we have a similar definition; we will not do the details. This second definition of expectation is mostly useful for theoretical purposes and much less so for calculations.

Similarly to the discrete case, we have

Proposition 6.2. $\mathbb{E}\, g(X) = \int g(x)f(x)dx$.

As in the discrete case,

$$\operatorname{Var} X = \mathbb{E}\,[X - \mathbb{E}\, X]^2.$$

As an example of these calculations, let us look at the uniform distribution. We say that a random variable X has a uniform distribution on $[a, b]$ if $f_X(x) = \frac{1}{b-a}$ if $a \le x \le b$ and 0 otherwise.

To calculate the expectation of X,

$$\mathbb{E}\, X = \int_{-\infty}^{\infty} xf_X(x)dx = \int_a^b x\frac{1}{b-a}dx = \frac{1}{b-a}\int_a^b x\,dx = \frac{1}{b-a}\Big(\frac{b^2}{2} - \frac{a^2}{2}\Big) = \frac{a+b}{2}.$$

This is what one would expect. To calculate the variance, we first calculate

$$\mathbb{E}\, X^2 = \int_{-\infty}^{\infty} x^2 f_X(x)dx = \int_a^b x^2 \frac{1}{b-a}dx = \frac{a^2+ab+b^2}{3}.$$

We then do some algebra to obtain

$$\operatorname{Var} X = \mathbb{E}\, X^2 - (\mathbb{E}\, X)^2 = \frac{(b-a)^2}{12}.$$

7. Normal distribution.

A r.v. is a standard normal (written $\mathcal{N}(0,1)$) if it has density

$$\frac{1}{\sqrt{2\pi}} e^{-x^2/2}.$$

A synonym for normal is Gaussian. The first thing to do is show that this is a density. Let $I = \int_0^\infty e^{-x^2/2}dx$. Then

$$I^2 = \int_0^\infty \int_{-\infty}^\infty e^{-x^2/2}e^{-y^2/2}dx\,dy.$$

Changing to polar coordinates,

$$I^2 = \int_0^{\pi/2} \int_0^\infty re^{-r^2/2}dr = \pi/2.$$

So $I = \sqrt{\pi/2}$, hence $\int_{-\infty}^\infty e^{-x^2/2}dx = \sqrt{2\pi}$ as it should.

Note

$$\int xe^{-x^2/2}dx = 0$$

by symmetry, so $\mathbb{E}\,Z = 0$. For the variance of Z, we use integration by parts:

$$\mathbb{E}\,Z^2 = \frac{1}{\sqrt{2\pi}}\int x^2e^{-x^2/2}dx = \frac{1}{\sqrt{2\pi}}\int x\cdot xe^{-x^2/2}dx.$$

The integral is equal to

$$-xe^{-x^2/2}\Big]_{-\infty}^\infty + \int e^{-x^2/2}dx = \sqrt{2\pi}.$$

Therefore $\mathrm{Var}\,Z = \mathbb{E}\,Z^2 = 1$.

We say X is a $\mathcal{N}(\mu,\sigma^2)$ if $X = \sigma Z + \mu$, where Z is a $\mathcal{N}(0,1)$. We see that

$$F_X(x) = \mathbb{P}(X \le x) = \mathbb{P}(\mu + \sigma Z \le x) = \mathbb{P}(Z \le (x-\mu)/\sigma) = F_Z((x-\mu)/\sigma)$$

if $\sigma > 0$. (A similar calculation holds if $\sigma < 0$.) Then by the chain rule X has density

$$f_X(x) = F_X'(x) = F_Z'((x-\mu)/\sigma) = \frac{1}{\sigma}f_Z((x-\mu)/\sigma).$$

This is equal to

$$\frac{1}{\sqrt{2\pi}\sigma}e^{-(x-\mu)^2/2\sigma^2}.$$

$\mathbb{E}\,X = \mu + \mathbb{E}\,Z$ and $\mathrm{Var}\,X = \sigma^2\mathrm{Var}\,Z$, so

$$\mathbb{E}\,X = \mu, \qquad \mathrm{Var}\,X = \sigma^2.$$

If X is $\mathcal{N}(\mu,\sigma^2)$ and $Y = aX+b$, then $Y = a(\mu+\sigma Z)+b = (a\mu+b)+(a\sigma)Z$, or Y is $\mathcal{N}(a\mu+b, a^2\sigma^2)$. In particular, if X is $\mathcal{N}(\mu,\sigma^2)$ and $Z = (X-\mu)/\sigma$, then Z is $\mathcal{N}(0,1)$.

The distribution function of a standard $\mathcal{N}(0,1)$ is often denoted $\Phi(x)$, so that

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

Tables of $\Phi(x)$ are often given only for $x > 0$. One can use the symmetry of the density function to see that

$$\Phi(-x) = 1 - \Phi(x);$$

this follows from

$$\begin{aligned}\Phi(-x) = \mathbb{P}(Z \le -x) &= \int_{-\infty}^{-x}\frac{1}{\sqrt{2\pi}}e^{-y^2/2}dy \\ &= \int_x^\infty \frac{1}{\sqrt{2\pi}}e^{-y^2/2}dy = \mathbb{P}(Z \ge x) = 1 - \mathbb{P}(Z < x) = 1 - \Phi(x).\end{aligned}$$

Example. Find $\mathbb{P}(1 \le X \le 4)$ if X is $\mathcal{N}(2, 25)$.

Answer. Write $X = 2 + 5Z$. So

$$\begin{aligned}\mathbb{P}(1 \le X \le 4) &= \mathbb{P}(1 \le 2 + 5Z \le 4) = \mathbb{P}(-1 \le 5Z \le 2) = \mathbb{P}(-0.2 \le Z \le .4)\\ &= \mathbb{P}(Z \le .4) - \mathbb{P}(Z \le -0.2) = \Phi(0.4) - \Phi(-0.2) = .6554 - [1 - \Phi(0.2)]\\ &= .6554 - [1 - .5793].\end{aligned}$$

Example. Find c such that $\mathbb{P}(|Z| \ge c) = .05$.

Answer. By symmetry we want c such that $\mathbb{P}(Z \ge c) = .025$ or $\Phi(c) = \mathbb{P}(Z \le c) = .975$. From the table we see $c = 1.96 \approx 2$. This is the origin of the idea that the 95% significance level is ± 2 standard deviations from the mean.

Proposition 7.1. *We have the following bounds. For $x > 0$*

$$\frac{1}{\sqrt{2\pi}}\Big(\frac{1}{x} - \frac{1}{x^3}\Big)e^{-x^2/2} \le 1 - \Phi(x) \le \frac{1}{\sqrt{2\pi}}\frac{1}{x}e^{-x^2/2}.$$

Proof. Take the inequalities

$$(1 - 3y^{-4})e^{-y^2/2} \le e^{-y^2/2} \le (1 + y^{-2})e^{-y^2/2}$$

and integrate. □

In particular, for x, large,

$$\mathbb{P}(Z \ge x) = 1 - \Phi(x) \sim \frac{1}{\sqrt{2\pi}}\frac{1}{x}e^{-x^2/2} \le e^{-x^2/2}.$$

8. Normal approximation to the binomial.

A special case of the central limit theorem is

Theorem 8.1. *If S_n is a binomial with parameters n and p, then*

$$\mathbb{P}\Big(a \le \frac{S_n - np}{\sqrt{np(1-p)}} \le b\Big) \to \mathbb{P}(a \le Z \le b),$$

as $n \to \infty$, where Z is a $\mathcal{N}(0, 1)$.

This approximation is good if $np(1-p) \ge 10$ and gets better the larger this quantity gets. Note np is the same as $\mathbb{E}\, S_n$ and $np(1-p)$ is the same as $\operatorname{Var} S_n$. So the ratio is also equal to $(S_n - \mathbb{E}\, S_n)/\sqrt{\operatorname{Var} S_n}$, and this ratio has mean 0 and variance 1, the same as a standard $\mathcal{N}(0, 1)$.

Note that here p stays fixed as $n \to \infty$, unlike the case of the Poisson approximation.

Example. Suppose a fair coin is tossed 100 times. What is the probability there will be more than 60 heads?

Answer. $np = 50$ and $\sqrt{np(1-p)} = 5$. We have

$$\mathbb{P}(S_n \ge 60) = \mathbb{P}((S_n - 50)/5 \ge 2) \approx \mathbb{P}(Z \ge 2) \approx .0228.$$

Example. Suppose a die is rolled 180 times. What is the probability a 3 will be showing more than 50 times?

Answer. Here $p = \frac{1}{6}$, so $np = 30$ and $\sqrt{np(1-p)} = 5$. Then $\mathbb{P}(S_n > 50) \approx \mathbb{P}(Z > 4)$, which is very small.

Example. Suppose a drug is supposed to be 75% effective. It is tested on 100 people. What is the probability more than 70 people will be helped?

Answer. Here S_n is the number of successes, $n = 100$, and $p = .75$. We have

$$\mathbb{P}(S_n \geq 70) = \mathbb{P}((S_n - 75)/\sqrt{300/16} \geq -1.154) \approx \mathbb{P}(Z \geq -1.154) \approx .87.$$

(The last figure came from a table.)

When $b-a$ is small, there is a correction that makes things more accurate, namely replace a by $a - \frac{1}{2}$ and b by $b + \frac{1}{2}$. This correction never hurts and is sometime necessary. For example, in tossing a coin 100 times, there ispositive probability that there are exactly 50 heads, while without the correction, the answer given by the normal approximation would be 0.

Example. We toss a coin 100 times. What is the probability of getting 49, 50, or 51 heads?

Answer. We write $\mathbb{P}(49 \leq S_n \leq 51) = \mathbb{P}(48.5 \leq S_n \leq 51.5)$ and then continue as above.

9. Some continuous distributions.

We look at some other continuous random variables besides normals.

Uniform. Here $f(x) = 1/(b-a)$ if $a \leq x \leq b$ and 0 otherwise. To compute expectations, $\mathbb{E}\, X = \frac{1}{b-a}\int_a^b x\, dx = (a+b)/2$.

Exponential. An exponential with parameter λ has density $f(x) = \lambda e^{-\lambda x}$ if $x \geq 0$ and 0 otherwise. We have $\mathbb{P}(X > a) = \int_a^\infty \lambda e^{-\lambda x} dx = e^{-\lambda a}$ and we readily compute $\mathbb{E}\, X = 1/\lambda$, $\operatorname{Var} X = 1/\lambda^2$. Examples where an exponential r.v. is a good model is the length of a telephone call, the length of time before someone arrives at a bank, the length of time before a light bulb burns out.

Exponentials are memory-less. This means that $\mathbb{P}(X > s + t \mid X > t) = \mathbb{P}(X > s)$, or given that the light bulb has burned 5 hours, the probability it will burn 2 more hours is the same as the probability a new light bulb will burn 2 hours. To prove this,

$$\mathbb{P}(X > s + t \mid X > t) = \frac{\mathbb{P}(X > s + t)}{\mathbb{P}(X > t)} = \frac{e^{-\lambda(s+t)}}{e^{-\lambda t}} = e^{-\lambda s} = \mathbb{P}(X > s).$$

Gamma. A gamma distribution with parameters λ and t has density

$$f(x) = \frac{\lambda e^{-\lambda t}(\lambda x)^{t-1}}{\Gamma(t)}$$

if $x \geq 0$ and 0 otherwise. Here $\Gamma(t) = \int_0^\infty e^{-y} y^{t-1} dt$ is the Gamma function, which interpolates the factorial function.

An exponential is the time for something to occur. A gamma is the time for t events to occur. A gamma with parameters $\frac{1}{2}$ and $\frac{n}{2}$ is known as a χ^2_n, a chi-squared r.v. with n degrees of freedom. Gammas and chi-squared's come up frequently in statistics. Another distribution that arises in statistics is the beta:

$$f(x) = \frac{1}{B(a,b)} x^{a-1}(1-x)^{b-1}, \qquad 0 < x < 1,$$

where $B(a,b) = \int_0^1 x^{a-1}(1-x)^{b-1}$.

Cauchy. Here

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

What is interesting about the Cauchy is that it does not have finite mean, that is, $\mathbb{E}\,|X| = \infty$.

Often it is important to be able to compute the density of $Y = g(X)$. Let us give a couple of examples. If X is uniform on $(0,1]$ and $Y = -\log X$, then $Y > 0$. If $x > 0$,

$$F_Y(x) = \mathbb{P}(Y \le x) = \mathbb{P}(-\log X \le x) = \mathbb{P}(\log X \ge -x) = \mathbb{P}(X \ge e^{-x}) = 1 - \mathbb{P}(X \le e^{-x}) = 1 - F_X(e^{-x}).$$

Taking the derivative,

$$f_Y(x) = \frac{d}{dx}F_Y(x) = -f_X(e^{-x})(-e^{-x}),$$

using the chain rule. Since $f_X = 1$, this gives $f_Y(x) = e^{-x}$, or Y is exponential with parameter 1.

For another example, suppose X is $\mathcal{N}(0,1)$ and $Y = X^2$. Then

$$\begin{aligned} F_Y(x) &= \mathbb{P}(Y \le x) = \mathbb{P}(X^2 \le x) = \mathbb{P}(-\sqrt{x} \le X \le \sqrt{x}) \\ &= \mathbb{P}(X \le \sqrt{x}) - \mathbb{P}(X \le -\sqrt{x}) = F_X(\sqrt{x}) - F_X(-\sqrt{x}). \end{aligned}$$

Taking the derivative and using the chain rule,

$$f_Y(x) = \frac{d}{dx}F_Y(x) = f_X(\sqrt{x})\Big(\frac{1}{2\sqrt{x}}\Big) - f_X(-\sqrt{x})\Big(-\frac{1}{2\sqrt{x}}\Big).$$

Remembering that $f_X(t) = \frac{1}{\sqrt{2\pi}}e^{-t^2/2}$ and doing some algebra, we end up with

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

which is a Gamma with parameters $\frac{1}{2}$ and $\frac{1}{2}$. (This is also a χ^2 with one degree of freedom.)

10. Multivariate distributions.

We want to discuss collections of random variables $(X_1, X_2, \ldots, X_n)$, which are known as random vectors. In the discrete case, we can define the density $p(x,y) = \mathbb{P}(X = x, Y = y)$. Remember that here the comma means "and."" In the continuous case a density is a function such that

$$\mathbb{P}(a \le X \le b, c \le Y \le d) = \int_a^b \int_c^d f(x,y)dy\,dx.$$

Example. If $f_{X,Y}(x,y) = ce^{-x}e^{-2y}$ for $0 < x < \infty$ and $x < y < \infty$, what is c?

Answer. We use the fact that a density must integrate to 1. So

$$\int_0^\infty \int_x^\infty ce^{-x}e^{-2y}dy\,dx = 1.$$

Recalling multivariable calculus, this is

$$\int_0^\infty ce^{-x}\tfrac{1}{2}e^{-2x}dx = \frac{c}{6},$$

so $c = 6$.

The multivariate distribution function of (X, Y) is defined by $F_{X,Y}(x, y) = \mathbb{P}(X \leq x, Y \leq y)$. In the continuous case, this is

$$\int_{-\infty}^{x} \int_{-\infty}^{y} f_{X,Y}(x, y) dy\, dx,$$

and so we have

$$f(x, y) = \frac{\partial^2 F}{\partial x \partial y}(x, y).$$

The extension to n random variables is exactly similar.

We have

$$\mathbb{P}(a \leq X \leq b, c \leq Y \leq d) = \int_a^b \int_c^d f_{X,Y}(x, y) dy\, dx,$$

or

$$\mathbb{P}((X, Y) \in D) = \int \int_D f_{X,Y} dy\, dx$$

when D is the set $\{(x, y) : a \leq x \leq b, c \leq y \leq d\}$. One can show this holds when D is any set. For example,

$$\mathbb{P}(X < Y) = \int \int_{\{x<y\}} f_{X,Y}(x, y) dy\, dx.$$

If one has the joint density of X and Y, one can recover the densities of X and of Y:

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x, y) dy, \qquad f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x, y) dx.$$

A multivariate random vector is $(X_1, \ldots, X_r)$ with

$$\mathbb{P}(X_1 = n_1, \ldots, X_r = n_r) = \frac{n!}{n_1! \cdots n_r!} p_1^{n_1} \cdots p_r^{n_r},$$

where $n_1 + \cdots + n_r = n$ and $p_1 + \cdots p_r = 1$.

In the discrete case we say X and Y are independent if $\mathbb{P}(X = x, Y = y) = \mathbb{P}(X = x, Y = y)$ for all x and y. In the continuous case, X and Y are independent if

$$\mathbb{P}(X \in A, Y \in B) = \mathbb{P}(X \in A)\mathbb{P}(Y \in B)$$

for all pairs of subsets A, B of the reals. The left hand side is an abbreviation for

$$\mathbb{P}(\{\omega : X(\omega) \text{ is in } A \text{ and } Y(\omega) \text{ is in } B\})$$

and similarly for the right hand side.

In the discrete case, if we have independence,

$$p_{X,Y}(x, y) = \mathbb{P}(X = x, Y = y) = \mathbb{P}(X = x)\mathbb{P}(Y = y) = p_X(x) p_Y(y).$$

In other words, the joint density $p_{X,Y}$ factors. In the continuous case,

$$\int_a^b \int_c^d f_{X,Y}(x, y) dy\, dx = \mathbb{P}(a \leq X \leq b, c \leq Y \leq d) = \mathbb{P}(a \leq X \leq b)\mathbb{P}(c \leq Y \leq d)$$
$$= \int_a^b f_X(x) dx \int_c^d f_Y(y) dy.$$

One can conclude from this that

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

or again the joint density factors. Going the other way, one can also see that if the joint density factors, then one has independence.

Example. Suppose one has a floor made out of wood planks and one drops a needle onto it. What is the probability the needle crosses one of the cracks? Suppose the needle is of length L and the wood planks are D across.

Answer. Let X be the distance from the midpoint of the needle to the nearest crack and let Θ be the angle the needle makes with the vertical. Then X and Θ will be independent. X is uniform on $[0, D/2]$ and Θ is uniform on $[0, \pi/2]$. A little geometry shows that the needle will cross a crack if $L/2 > X/\cos\Theta$. We have $f_{X,\Theta} = \frac{4}{\pi D}$ and so we have to integrate this constant over the set where $X < L\cos\Theta/2$ and $0 \le \Theta \le \pi/2$ and $0 \le X \le D/2$. The integral is

$$\int_0^{\pi/2} \int_0^{L\cos\theta/2} \frac{4}{\pi D} dx\, d\theta = \frac{2L}{\pi D}.$$

If X and Y are independent, then

$$\begin{aligned}
\mathbb{P}(X+Y \le a) &= \int\int_{\{x+y\le a\}} f_{X,Y}(x,y)dx\,dy = \int\int_{\{x+y\le a\}} f_X(x)f_Y(y)dx\,dy \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{a-y} f_X(x)f_Y(y)dx\,dy \\
&= \int F_X(a-y)f_Y(y)dy.
\end{aligned}$$

Differentiating with respect to a, we have

$$f_{X+Y}(a) = \int f_X(a-y)f_Y(y)dy.$$

There are a number of cases where this is interesting.

(1) If X is a gamma with parameters s and λ and Y is a gamma with parameters t and λ, then straightforward integration shows that $X+Y$ is a gamma with parameters $s+t$ and λ. In particular, the sum of n independent exponentials with parameter λ is a gamma with parameters n and λ.

(2) If Z is a $\mathcal{N}(0,1)$, then $F_{Z^2}(y) = \mathbb{P}(Z^2 \le y) = \mathbb{P}(-\sqrt{y} \le Z \le \sqrt{y}) = F_Z(\sqrt{y}) - F_Z(-\sqrt{y})$. Differentiating shows that $f_{Z^2}(y) = ce^{-y/2}(y/2)^{(1/2)-1}$, or Z^2 is a gamma with parameters $\frac{1}{2}$ and $\frac{1}{2}$. So using (1) above, if Z_i are independent $\mathcal{N}(0,1)$'s, then $\sum_{i=1}^n Z_i^2$ is a gamma with parameters $n/2$ and $\frac{1}{2}$, i.e., a χ^2_n.

(3) If X_i is a $\mathcal{N}(\mu_i, \sigma_i^2)$ and the X_i are independent, then some lengthy calculations show that $\sum_{i=1}^n X_i$ is a $\mathcal{N}(\sum \mu_i, \sum \sigma_i^2)$.

(4) The analogue for discrete random variables is easier. If X and Y takes only nonnegative integer values, we have

$$\mathbb{P}(X+Y=r) = \sum_{k=0}^{r} \mathbb{P}(X=k, Y=r-k) = \sum_{k=0}^{r} \mathbb{P}(X=k)\mathbb{P}(Y=r-k).$$

In the case where X is a Poisson with parameter λ and Y is a Poisson with parameter μ, we see that $X+Y$ is a Poisson with parameter $\lambda+\mu$.

Note that it is not always the case that the sum of two independent random variables will be a random variable of the same type.

If X and Y are independent normals, then $-Y$ is also a normal (with $\mathbb{E}(-Y) = -\mathbb{E}Y$ and $\operatorname{Var}(-Y) = (-1)^2 \operatorname{Var} Y = \operatorname{Var} Y$), and so $X - Y$ is also normal.

To define a conditional density in the discrete case, we write

$$p_{X|Y=y}(x \mid y) = \mathbb{P}(X = x \mid Y = y).$$

This is equal to

$$\frac{\mathbb{P}(X = x, Y = y)}{\mathbb{P}(Y = y)} = \frac{p(x, y)}{p_Y(y)}.$$

Analogously, we define in the continuous case

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

Just as in the one-dimensional case, there is a change of variables formula. Let us recall how the formula goes in one dimension. If X has a density f_X and $y = g(X)$, then

$$F_Y(y) = \mathbb{P}(Y \leq y) = \mathbb{P}(g(X) \leq y) = \mathbb{P}(X \leq g^{-1}(y)) = F_X(g^{-1}(y)).$$

Taking the derivative, using the chain rule, and recalling that the derivative of $g^{-1}(y)$ is $1/g'(y)$, we have

$$f_Y(y) = f_X(g^{-1}(y)) \frac{1}{g(y)}.$$

The higher dimensional case is very analogous. Suppose $Y_1 = g_1(X_1, X_2)$ and $Y_2 = g_2(X_1, X_2)$. Let h_1 and h_2 be such that $X_1 = h_1(Y_1, Y_2)$ and $X_2 = h_2(Y_1, Y_2)$. (This plays the role of g^{-1}.) Let J be the Jacobian of the mapping $(x_1, x_2) \to (g_1(x_1, x_2), g_2(x_1, x_2)$, so that $J = \frac{\partial g_1}{\partial x_1} \frac{\partial g_2}{\partial x_2} - \frac{\partial g_1}{\partial x_2} \frac{\partial g_2}{\partial x_1}$. (This is the analogue of $g'(y)$.) Using the change of variables theorem from multivariable calculus, we have

$$f_{Y_1, Y_2}(y_1, y_2) = f_{X_1, X_2}(x_1, x_2) |J|^{-1}.$$

Example. Suppose X_1 is $\mathcal{N}(0, 1)$, X_2 is $\mathcal{N}(0, 4)$, and X_1 and X_2 are independent. Let $Y_1 = 2X_1 + X_2, Y_2 = X_1 - 3X_2$. Then $y_1 = g_1(x_1, x_2) = 2x_1 + x_2, y_2 = g_2(x_1, x_2) = x_1 - x_3$, so

$$J = \begin{pmatrix} 2 & 1 \\ 1 & -3 \end{pmatrix} = -7.$$

(In general, J might depend on x, and hence on y.) Some algebra leads to $x_1 = \frac{3}{7} y_1 + \frac{1}{7} y_2$, $x_2 = \frac{1}{7} y_1 - \frac{2}{7} y_2$. Since X_1 and X_2 are independent,

$$f_{X_1, X_2}(x_1, x_2) = f_{X_1}(x_1) f_{X_2}(x_2) = \frac{1}{\sqrt{2\pi}} e^{-x_1^2/2} \frac{1}{\sqrt{8\pi}} e^{-x_2^2/8}.$$

Therefore

$$f_{Y_1, Y_2}(y_1, y_2) = \frac{1}{\sqrt{2\pi}} e^{-(\frac{3}{7} y_1 + \frac{1}{7} y_2)^2/2} \frac{1}{\sqrt{8\pi}} e^{-(\frac{1}{7} y_1 - \frac{2}{7} y_2)^2/8} \frac{1}{7}.$$

11. Expectations.

As in the one variable case, we have

$$\mathbb{E}\, g(X,Y) = \sum\sum g(x,y)p(x,y)$$

in the discrete case and

$$\mathbb{E}\, g(X,Y) = \int\int g(x,y)f(x,y)dx\,dy$$

in the continuous case.

If we set $g(x,y) = x+y$, then

$$\mathbb{E}\,(X+Y) = \int\int (x+y)f(x,y)dx\,dy = \int\int xf(x,y)dx\,dy + \int\int yf(x,y)dx\,dy.$$

If we now set $g(x,y) = x$, we see the first integral on the right is $\mathbb{E}\, X$, and similarly the second is $\mathbb{E}\, Y$. Therefore

$$\mathbb{E}\,(X+Y) = \mathbb{E}\, X + \mathbb{E}\, Y.$$

Proposition 11.1. *If X and Y are independent, then*

$$\mathbb{E}\,[h(X)k(Y)] = \mathbb{E}\, h(X)\mathbb{E}\, k(Y).$$

In particular, $\mathbb{E}\,(XY) = (\mathbb{E}\, X)(\mathbb{E}\, Y)$.

Proof. By the above with $g(x,y) = h(x)k(y)$,

$$\begin{aligned}
\mathbb{E}\,[h(X)k(Y)] &= \int\int h(x)k(y)f(x,y)dx\,dy \\
&= \int\int h(x)k(y)f_X(x)f_Y(y)dx\,dy \\
&= \int h(x)f_X(x)\int k(y)f_Y(y)dy\,dx = \int h(x)f_X(x)(\mathbb{E}\, k(Y))dx \\
&= \mathbb{E}\, h(X)\mathbb{E}\, k(Y).
\end{aligned}$$

□

The covariance of two random variables X and Y is defined by

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

As with the variance, $\mathrm{Cov}\,(X,Y) = \mathbb{E}\,(XY) - (\mathbb{E}\, X)(\mathbb{E}\, Y)$. It follows that if X and Y are independent, then $\mathbb{E}\,(XY) = (\mathbb{E}\, X)(\mathbb{E}\, Y)$, and then $\mathrm{Cov}\,(X,Y) = 0$.

Note

$$\begin{aligned}
\mathrm{Var}\,(X+Y) &= \mathbb{E}\,[((X+Y) - \mathbb{E}\,(X+Y))^2] = \mathbb{E}\,[((X-\mathbb{E}\, X) + (Y-\mathbb{E}\, Y))^2] \\
&= \mathbb{E}\,[(X-\mathbb{E}\, X)^2 + 2(X-\mathbb{E}\, X)(Y-\mathbb{E}\, Y) + (Y-\mathbb{E}\, Y)^2] \\
&= \mathrm{Var}\, X + 2\mathrm{Cov}\,(X,Y) + \mathrm{Var}\, Y.
\end{aligned}$$

We have the following corollary.

Proposition 11.2. *If X and Y are independent, then*

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

Proof. We have

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

□

Since a binomial is the sum of n independent Bernoulli's, its variance is $np(1-p)$. If we write $\overline{X} = \sum_{i=1}^n X_i/n$ and the X_i are independent and have the same distribution ($\overline{X}$ is called the sample mean), then $\mathbb{E}\,\overline{X} = \mathbb{E}\,X_1$ and $\operatorname{Var}\overline{X} = \operatorname{Var} X_1/n$.

We define the conditional expectation of X given Y by

$$\mathbb{E}[X \mid Y = y] = \int x f_{X|Y=y}(x)dx.$$

12. Moment generating functions.

We define the moment generating function m_X by

$$m_X(t) = \mathbb{E}\,e^{tX},$$

provided this is finite. In the discrete case this is equal to $\sum e^{tx}p(x)$ and in the continuous case $\int e^{tx}f(x)dx$.

Let us compute the moment generating function for some of the distributions we have been working with.

1. Bernoulli: $pe^t + (1-p)$.
2. Binomial: using independence,

   $$\mathbb{E}\,e^{t\sum X_i} = \mathbb{E}\prod e^{tX_i} = \prod \mathbb{E}\,e^{tX_i} = (pe^t + (1-p))^n,$$

   where the X_i are independent Bernoulli's.
3. Poisson:

   $$\mathbb{E}\,e^{tX} = \sum \frac{e^{tk}e^{-\lambda}\lambda^k}{k!} = e^{-\lambda}\sum \frac{(\lambda e^t)^k}{k!} = e^{-\lambda}e^{\lambda e^t} = e^{\lambda(e^t-1)}.$$

4. Exponential:

   $$\mathbb{E}\,e^{tX} = \int_0^\infty e^{tx}\lambda e^{-\lambda x}dx = \frac{\lambda}{\lambda - t}$$

   if $t < \lambda$ and ∞ if $t \geq \lambda$.
5. $\mathcal{N}(0,1)$:

   $$\frac{1}{\sqrt{2\pi}}\int e^{tx}e^{-x^2/2}dx = e^{t^2/2}\frac{1}{\sqrt{2\pi}}\int e^{-(x-t)^2/2}dx = e^{t^2/2}.$$

6. $\mathcal{N}(\mu,\sigma^2)$: Write $X = \mu + \sigma Z$. Then

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

Proposition 12.1. *If X and Y are independent, then $m_{X+Y}(t) = m_X(t)m_Y(t)$.*

Proof. By independence and Proposition 11.1,

$$m_{X+Y}(t) = \mathbb{E}\,e^{tX}e^{tY} = \mathbb{E}\,e^{tX}\mathbb{E}\,e^{tY} = m_X(t)m_Y(t).$$

□

Proposition 12.2. *If $m_X(t) = m_Y(t) < \infty$ for all t in an interval, then X and Y have the same distribution.*

We will not prove this, but this is essentially the uniqueness of the Laplace transform. Note $\mathbb{E}\, e^{tX} = \int e^{tx} f_X(x)dx$. If $f_X(x) = 0$ for $x < 0$, this is $\int_0^\infty e^{tx} f_X(x)dx = \mathcal{L}f_X(-t)$, where $\mathcal{L}f_X$ is the Laplace transform of f_X.

We can use this to verify some of the properties of sums we proved before. For example, if X is a $\mathcal{N}(a, b^2)$ and Y is a $\mathcal{N}(c, d^2)$ and X and Y are independent, then

$$m_{X+Y}(t) = e^{at+b^2t^2/2}e^{ct+d^2t^2/2} = e^{(a+c)t+(b^2+d^2)t^2/2}.$$

Proposition 12.2 then implies that $X + Y$ is a $\mathcal{N}(a + c, b^2 + d^2)$.

Similarly, if X and Y are independent Poisson random variables with parameters a and b, resp., then

$$m_{X+Y}(t) = m_X(t)m_Y(t) = e^{a(e^t-1)}e^{b(e^t-1)} = e^{(a+b)(e^t-1)},$$

which is the moment generating function of a Poisson with parameter $a + b$.

One problem with the moment generating function is that it might be infinite. One way to get around this, at the cost of considerable work, is to use the characteristic function $\varphi_X(t) = \mathbb{E}\, e^{itX}$, where $i = \sqrt{-1}$. This is always finite, and is the analogue of the Fourier transform.

The joint moment generating function of X and Y is $m_{X,Y}(s, t) = \mathbb{E}\, e^{sX+tY}$. If X and Y are independent, then $m_{X,Y}(s, t) = m_X(s)m_Y(t)$ by Proposition 12.2. We will not prove this, but the converse is also true: if $m_{X,Y}(s, t) = m_X(s)m_Y(t)$ for all s and t, then X and Y are independent.

13. Limit laws.

Suppose X_i are independent and have the same distribution. In the case of continuous or discrete random variables, this means they all have the same density. We say the X_i are i.i.d., which stands for "independent and identically distributed." Let $S_n = \sum_{i=1}^n X_i$. S_n is called the partial sum process.

Theorem 13.1. *Suppose $\mathbb{E}\,|X_i| < \infty$ and let $\mu = \mathbb{E}\, X_i$. Then*

$$\frac{S_n}{n} \to \mu.$$

This is known as the strong law of large numbers (SLLN). The convergence here means that $S_n(\omega)/n \to \mu$ for every $\omega \in S$, where S is the probability space, except possibly for a set of ω of probability 0.

The proof of Theorem 13.1 is quite hard, and we prove a weaker version, the weak law of large numbers (WLLN). The WLLN states that for every $a > 0$,

$$\mathbb{P}\Big(\Big|\frac{S_n}{n} - \mathbb{E}\, X_1\Big| > a\Big) \to 0$$

as $n \to \infty$. It is not even that easy to give an example of random variables that satisfy the WLLN but not the SLLN.

Before proving the WLLN, we need an inequality called Chebyshev's inequality.

Proposition 13.2. *If $Y \geq 0$, then for any A,*

$$\mathbb{P}(Y > A) \leq \frac{\mathbb{E}\, Y}{A}.$$

Proof. Let $B = \{Y > A\}$. Recall 1_B is the random variable that is 1 if $\omega \in B$ and 0 otherwise. Note $1_B \leq Y/A$. This is obvious if $\omega \notin B$, while if $\omega \in B$, then $Y(\omega)/A > 1 = 1_B(\omega)$. We then have

$$\mathbb{P}(Y > A) = \mathbb{P}(B) = \mathbb{E}\, 1_B \leq \mathbb{E}\,(Y/A) = \frac{\mathbb{E}\, Y}{A}.$$

□

We now prove the WLLN.

Theorem 13.3. *Suppose the X_i are i.i.d. and $\mathbb{E}\,|X_1|$ and $\operatorname{Var} X_1$ are finite. Then for every $a > 0$,*

$$\mathbb{P}\Big(\Big|\frac{S_n}{n} - \mathbb{E}\, X_1\Big| > a\Big) \to 0$$

as $n \to \infty$.

Proof. Recall $\mathbb{E}\, S_n = n\mathbb{E}\, X_1$ and by the independence, $\operatorname{Var} S_n = n\operatorname{Var} X_1$, so $\operatorname{Var}(S_n/n) = \operatorname{Var} X_1/n$. We have

$$\begin{aligned}
\mathbb{P}\Big(\Big|\frac{S_n}{n} - \mathbb{E}\, X_1\Big| > a\Big) &= \mathbb{P}\Big(\Big|\frac{S_n}{n} - \mathbb{E}\left(\frac{S_n}{n}\right)\Big| > a\Big) \\
&= \mathbb{P}\Big(\Big|\frac{S_n}{n} - \mathbb{E}\left(\frac{S_n}{n}\right)\Big|^2 > a^2\Big) \\
&\leq \frac{\mathbb{E}\,|\frac{S_n}{n} - \mathbb{E}\,(\frac{S_n}{n})|^2}{a^2} \\
&= \frac{\operatorname{Var}(\frac{S_n}{n})}{a^2} \\
&= \frac{\frac{\operatorname{Var} X_1}{n}}{a^2} \to 0.
\end{aligned}$$

The inequality step follows from Proposition 13.2 with $A = a^2$ and $Y = |\frac{S_n}{n} - \mathbb{E}\,(\frac{S_n}{n})|^2$. □

We now turn to the central limit theorem (CLT).

Theorem 13.4. *Suppose the X_i are i.i.d. Suppose $\mathbb{E}\, X_i^2 < \infty$. Let $\mu = \mathbb{E}\, X_i$ and $\sigma^2 = \operatorname{Var} X_i$. Then*

$$\mathbb{P}\Big(a \leq \frac{S_n - n\mu}{\sigma\sqrt{n}} \leq b\Big) \to \mathbb{P}(a \leq Z \leq b)$$

for every a and b, where Z is a $\mathcal{N}(0,1)$.

The ratio on the left is $(S_n - \mathbb{E}\, S_n)/\sqrt{\operatorname{Var} S_n}$. We do not claim that this ratio converges for any ω (in fact, it doesn't), but that the probabilities converge.

Example. If the X_i are i.i.d. Bernoulli random variables, so that S_n is a binomial, this is just the normal approximation to the binomial.

Example. Suppose we roll a die 3600 times. Let X_i be the number showing on the i^{th} roll. We know S_n/n will be close to 3.5. What's the probability it differs from 3.5 by more than 0.05?

Answer. We want

$$\mathbb{P}\Big(\Big|\frac{S_n}{n} - 3.5\Big| > .05\Big).$$

We rewrite this as

$$\begin{aligned}\mathbb{P}(|S_n - n\mathbb{E}\,X_1| > (.05)(3600)) &= \mathbb{P}\Big(\Big|\frac{S_n - n\mathbb{E}\,X_1}{\sqrt{n}\sqrt{\operatorname{Var} X_1}}\Big| > \frac{180}{(60)\sqrt{\frac{35}{12}}}\Big)\\ &\approx \mathbb{P}(|Z| > 1.756) \approx .08.\end{aligned}$$

Example. Suppose the lifetime of a human has expectation 72 and variance 36. What is the probability that the average of the lifetimes of 100 people exceeds 73?

Answer. We want

$$\begin{aligned}\mathbb{P}\Big(\frac{S_n}{n} > 73\Big) &= \mathbb{P}(S_n > 7300)\\ &= \mathbb{P}\Big(\frac{S_n - n\mathbb{E}\,X_1}{\sqrt{n}\sqrt{\operatorname{Var} X_1}} > \frac{7300 - (100)(72)}{\sqrt{100}\sqrt{36}}\Big)\\ &\approx \mathbb{P}(Z > 1.667) \approx .047.\end{aligned}$$

The idea behind proving the central limit theorem is the following. It turns out that if $m_{Y_n}(t) \to m_Z(t)$ for every t, then $\mathbb{P}(a \le Y_n \le b) \to \mathbb{P}(a \le Z \le b)$. (We won't prove this.) We are going to let $Y_n = (S_n - n\mu)/\sigma\sqrt{n}$. Let $W_i = (X_i - \mu)/\sigma$. Then $\mathbb{E}\,W_i = 0$, $\operatorname{Var} W_i = \frac{\operatorname{Var} X_i}{\sigma^2} = 1$, the W_i are independent, and

$$\frac{S_n - n\mu}{\sigma\sqrt{n}} = \frac{\sum_{i=1}^n W_i}{\sqrt{n}}.$$

So there is no loss of generality in assuming that $\mu = 0$ and $\sigma = 1$. Then

$$m_{Y_n}(t) = \mathbb{E}\,e^{tY_n} = \mathbb{E}\,e^{(t/\sqrt{n})(S_n)} = m_{S_n}(t/\sqrt{n}).$$

Since the X_i are i.i.d., all the X_i have the same moment generating function. Since $S_n = X_1 + \cdot + X_n$, then

$$m_{S_n}(t) = m_{X_1}(t)\cdots m_{X_n}(t) = [m_{X_1}(t)]^n.$$

If we expand e^{tX_1} as a power series,

$$m_{X_1}(t) = \mathbb{E}\,e^{tX_1} = 1 + t\mathbb{E}\,X_1 + \frac{t^2}{2!}\mathbb{E}\,(X_1)^2 + \frac{t^3}{3!}\mathbb{E}\,(X_1)^3 + \cdots.$$

We put the above together and obtain

$$\begin{aligned}m_{Y_n}(t) &= m_{S_n}(t/\sqrt{n})\\ &= [m_{X_1}(t/\sqrt{n})]^n\\ &= \Big[1 + t\cdot 0 + \frac{(t/\sqrt{n})^2}{2!} + R_n\Big]^n\\ &= \Big[1 + \frac{t^2}{2n} + R_n]^n,\end{aligned}$$

where $|R_n|/n \to 0$ as $n \to \infty$. This converges to $e^{t^2/2} = m_Z(t)$ as $n \to \infty$.