Probability Review
Master of Science in Financial Mathematics
University of Chicago
Gregory F. Lawler
Last updated: July 27, 2023

This set of notes will combine a review of material generally covered in an undergraduate course in probability with some standard brain teasers in probability. When discussing "undergraduate" probability, we mean topics that do not use measure theory. We will first consider discrete probability spaces and then go to continuous spaces.

1 Discrete probability

1.1 Random variables and expectation

A *probability space* is a set Ω, sometimes denoted by S, called the sample space, as well as a collection $\mathcal{F}$ of subsets of Ω called the events and a probability (measure) $\mathbb{P}$ assigned to events.

- The elements of Ω are sometimes called outcomes.
- The probability space can be discrete, that is, consist of a finite or countable infinite set of points that can be enumerated $\{x_1, x_2, \ldots\}$ or can be uncountable. If the space is discrete we generally can let $\mathcal{F}$ denote the set of all subsets of Ω. This is not true of the space is continuous but this usually is not important for applied purposes since every subset one can think of is in $\mathcal{F}$.
- The probability is a function $\mathbb{P}$ that assigns to each event $A \in \mathcal{F}$, a number $\mathbb{P}(A) \in [0,1]$ which denotes the probability that A occurs. It satisfies $\mathbb{P}(\Omega) = 1$ and countable additivity: if $A_1, A_2, \ldots$ are disjoint events, then

$$\mathbb{P}\left(A_1 \cup A_2 \cup \cdots\right) = \sum_{k=1}^{\infty} \mathbb{P}(A_k).$$

Most of the time one does not worry about the probability space — the notion of probability is usually well defined on the possible events that one is looking at. However, it is important to remember the countable additivity rule.

We will use $\mathbb{P}$ for probability but P is also a standard notation.

Recall that $\emptyset$ denotes the empty set. We also write $\Omega \setminus A$ for the *complement* of A. As a subset, it is the set of points in Ω that are not in A. In probability language it is the

event "A does not happen". Events $A_1, A_2, \ldots$ are disjoint (sometimes called mutually exclusive) if for each $j \neq k$, $A_j \cap A_k = \emptyset$.

In measure theory, one learns that one must assume that the set of events $\mathcal{F}$ is a *σ-algebra* which means that it satisfies:

- $\emptyset \in \mathcal{F}$
- If $A \in \mathcal{F}$ then $\Omega \setminus A \in \mathcal{F}$.
- If $A_1, A_2, \ldots \in \mathcal{F}$, then so is $A_1 \cup A_2 \cup \cdots$.

For a discrete probability space $\Omega = \{a_1, a_2, \ldots\}$ a probability is the same thing as a function $p : \Omega \to [0, 1]$ with $p(a_1) + p(a_2) + \cdots = 1$. An event A is a subset of Ω and

$$\mathbb{P}(A) = \sum_{a \in A} p(a).$$

Example Suppose we want to model the flipping of a fair coin 10 times. For our probability space we could choose the set of all finite sequences of heads and tails: for example $HHTHHTTTHH$ would represent the outcome that tails came on tosses number 3,6,7,8 and heads came on the others. There are 2^{10} outcomes and since the coin is fair we assign probability 2^{-10} to each one. We often write events using words. For example we might write {exactly 9 heads} for the event

$$\{HHHHHHHHHT,\ HHHHHHHHTH,\ HHHHHHHTHH,\ HHHHHHTHHH,$$
$$HHHHHTHHHH,\ HHHHTHHHHH,\ HHHTHHHHH,\ HHTHHHHHH$$
$$HTHHHHHHH,\ THHHHHHHHH\}.$$

For practical purposes, one does not worry so much about the probability space but rather on random variables which are numerical outcomes from some random phenomenon. Formally we define a random variable on a discrete probability space to be a function from Ω to the real line. The standard notation for random variables is capital letters, X, Y, Z being the most typical.

A discrete random variable X takes on either a finite or countably infinite number of different values. The probability (mass) function for the random variable is the function

$$p(x) = p_X(x) = \mathbb{P}\{X = x\}.$$

This is nonzero only on a countable set, say, $\{x_1, x_2, \ldots\}$. It satisfies

$$1 = \mathbb{P}\{-\infty < X < \infty\} = \sum_x p(x) = \sum_{j=1}^{\infty} p(x_j).$$

To specify the *distribution* of a discrete random variable is to give its probability function. Equivalently, one can give the (cumulative) distribution function defined by

$$F(t) = F_X(t) = \mathbb{P}\{X \leq t\} = \sum_{x \leq t} p(x).$$

Example Uniform distribution: If V if a finite set with n elements, a random variable has a *uniform distribution* on V if each point is equally likely to be chosen.

$$p(x) = \frac{1}{n}, \quad x \in V.$$

The expectation (also called expected value, mean, average value) of a discrete random variable $\mathbb{E}[X]$ is the average value that one would expect in the long run from repeated trials from this distribution. It is defined by

$$\mathbb{E}[X] = \sum_x x\, p(x)$$

provided that

$$\mathbb{E}[|X|] = \sum_x |x|\, p(x) < \infty.$$

We say that X is an integrable random variable if $\mathbb{E}[|X|] < \infty$. Other standard notations are $E[X], \mu_X$.

Let us recall some facts about infinite sums. If $a_1, a_2, \ldots$ is a sequence of real numbers, we say that the sum (or series)

$$\sum_{j=1}^{\infty} a_j$$

is absolutely convergent if

$$\sum_{j=1}^{\infty} |a_j| < \infty.$$

In this case the sum above is well defined. There is a notion of conditional convergence in calculus but we will not use it here — in order for the expectation of a random variable to exist the sum must be absolutely convergent.

If a random variable takes on only nonnegative values and $\sum xp(x) = \infty$, then we will write $\mathbb{E}[X] = \infty$. However if

$$\sum_{x>0} x\, p(x) = \infty \qquad \text{and} \qquad \sum_{x<0} |x|\, p(x) = \infty,$$

we just say that the expectation is not defined.

Example Suppose X is the value of a roll of a standard 6-sided die. Then

$$p(1) = p(2) = \cdots = p(6) = \frac{1}{6}$$

and

$$\mathbb{E}[X] = 1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + \cdots + 6 \cdot \frac{1}{6} = 3.5.$$

Note that $\mathbb{E}[X]$ does not have to be a possible value for X.

Example Suppose we roll a die until we get a 6. Let X be the number of rolls needed until we get a 6. Note that $\mathbb{P}\{X = 1\} = 1/6$ and more generally if $k \geq 1$ is an integer, the probability that it takes k rolls to get a 6 is the probability that the first $k-1$ rolls are not a 6 and the last roll is a 6,

$$p(k) = \mathbb{P}\{X = k\} = \left(\frac{5}{6}\right)^{k-1} \frac{1}{6},$$

$$\mathbb{E}[X] = \sum_{k=1}^{\infty} k \left(\frac{5}{6}\right)^{k-1} \frac{1}{6}.$$

There are various ways to compute the last sum. One trick is to use a rule that the expectation must satisfy. We have $X \geq 1$ since we must always roll at least once. With probability $1/6$ we stop after the first roll. Otherwise, the expected number of rolls from then on is the same as if we started at the beginning. This gives the equation

$$\mathbb{E}[X] = 1 + \frac{5}{6}\,\mathbb{E}[X]$$

and hence $\mathbb{E}[X] = 6$.

There is a very important property about expectation.

- **Linearity**. If X and Y are random variables and a, b are real numbers then,

$$\mathbb{E}[aX + bY] = a\,\mathbb{E}[X] + b\,\mathbb{E}[Y].$$

This is easy to derive from the definition but it is a very powerful result. As the next example illustrates, this result holds regardless of how dependent the random variables are on each other.

Example Suppose we roll two independent fair dice one of which is red and the other is green. Let

$$X = \text{value of red die}, \qquad Y = \text{value of green die}.$$

As we have already seen $\mathbb{E}[X] = \mathbb{E}[Y] = 3.5$. Hence

$$\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y] = 7.$$

But it is also true that $\mathbb{E}[X + X] = \mathbb{E}[X] + \mathbb{E}[X] = 7$. This is not to say that $X + Y$ and $X + X$ have the same distribution; they just have the same expectation. The random variable $X + X$ gives probability $1/6$ to each element of $\{2, 4, 6, 8, 10, 12\}$. It is not hard to show (check it if you have never seen it!) that $X + Y$ has density

$$p(2) = p(12) = \frac{1}{36}, \quad p(3) = p(11) = \frac{2}{36}, \quad p(4) = p(10) = \frac{3}{36},$$

$$p(5) = p(9) = \frac{4}{36}, \quad p(6) = p(8) = \frac{5}{36}, \quad p(7) = \frac{6}{36}.$$

If you enjoy arithmetic you can check that

$$2 \cdot \frac{1}{6} + 4 \cdot \frac{1}{6} + 6 \cdot \frac{1}{6} + 8 \cdot \frac{1}{6} + 10\frac{1}{6} + 12 \cdot \frac{1}{6} = 7,$$

$$2 \cdot \frac{1}{36} + 3 \cdot \frac{2}{36} + 4 \cdot \frac{3}{36} + 5 \cdot \frac{4}{36} + 6 \cdot \frac{5}{36} + 7 \cdot \frac{6}{36} + 8 \cdot \frac{5}{36} + 9 \cdot \frac{4}{36} + 10 \cdot \frac{3}{36} + 11 \cdot \frac{2}{36} + 12 \cdot \frac{1}{36} = 7.$$

In optimization problems in applied mathematics, one often chooses to maximize the *expected value* of a quantity such as profit. Expected value is one possible thing to maximize but there are other choices which involve risk factors. What people often do not say is that one reason to try to maximize expected value is that it is one of the easiest things to work with. It is the linearity of expectation, even for dependent random variables, that makes it easier.

1.2 Independence and conditional probability

We have already used the notion of independence which is very intuitive. If A and B are two events and they are independent, that is, whether one occurs does not affect whether the other occurs, then the probabilities satisfy a product rule, $\mathbb{P}(A \cap B) = \mathbb{P}(A)\,\mathbb{P}(B)$. How do we make this notion mathematically precise? We go backwards and use this intuitive rule as the *definition* of independence.

Definition Two events A and B are independent if they satisfy the product rule

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

When more than two events are involved, the definition becomes a little more complicated. We want to say that a collection of events $\{A_j\}$ is (mutually) independent if for each j, no information about the other events is useful in determining if A_j occurs. It is not sufficient to assume that $\mathbb{P}(A_j \cap A_k) = \mathbb{P}(A_j)\,\mathbb{P}(A_k)$ for each $j \neq k$.

Definition A (possibly infinite) collection of event $\{A_k\}$ is (mutually) independent if for every distinct $j_1, \ldots, j_k$,

$$\mathbb{P}(A_{j_1} \cap \cdots \cap A_{j_k}) = \mathbb{P}(A_{j_1})\,\mathbb{P}(A_{j_2}) \cdots \mathbb{P}(A_{j_k}).$$

The parentheses indicate that when we say independent we will implicitly mean mutually independent. If the events satisfy $\mathbb{P}(A_j \cap A_k) = \mathbb{P}(A_j)\,\mathbb{P}(A_k)$ for each $j \neq k$, then the events are called *pairwise independent*. As the next example shows, pairwise independence does not imply independence.

Example Suppose we roll two dice, one red and one green, and let

$$X = \text{value on red die}\ , \quad Y = \text{value on green die}\ .$$

Consider the events

$$A = \{X = 2\}, \quad B = \{Y = 5\}, \quad C = \{X + Y = 7\}.$$

Then one can check that $\mathbb{P}(A) = \mathbb{P}(B) = \mathbb{P}(C) = \frac{1}{6}$. Also, the events $A \cap B, A \cap C, B \cap C$ all represent the event $\{X = 2, Y = 5\}$ and hence

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

This shows that the events are pairwise independent. However, the event $A \cap B \cap C$ is also the event $\{X = 2, Y = 5\}$ and hence

$$\frac{1}{36} = \mathbb{P}(A \cap B \cap C) \neq \mathbb{P}(A)\,\mathbb{P}(B)\,\mathbb{P}(C),$$

and hence the events are not (mutually) independent. We can think of this intuitively. If one is rolling a pair of dice and wants the sum of the values to equal 7, then having someone tell you that the red die is 2 does not change the probability of success. Similarly, having someone tell you that the green die is 5 does not change the probability. However, having someone tell you both that $X = 2$ and that $Y = 5$ changes the probability significantly!

The product rule for independence can be stated as: if B is independent of A, then the probability that B occurs given that A has occurred is $\mathbb{P}(B)$, exactly the same as without that information. We wish to generalize to the idea that $\mathbb{P}(B \mid A)$ is the "probability that B occurs given that A occurs". In this case the generalization of the product rule is

$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\,\mathbb{P}(B \mid A). \tag{1}$$

Similarly to independence, we use this intuition as the definition of conditional probability.

Definition If A, B are events with $\mathbb{P}(A) > 0$, the conditional probability of B given A denoted by $\mathbb{P}(B \mid A)$ is defined by

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

If $\mathbb{P}(A) = 0$, we do not define $\mathbb{P}(B \mid A)$. The definition immediately implies that (1) holds if $\mathbb{P}(A) > 0$. If $\mathbb{P}(A) = 0$, we can say that it still holds since we do not care what $\mathbb{P}(B \mid A)$ is. As can be seen in the next example, $\mathbb{P}(B \mid A)$ and $\mathbb{P}(A \mid B)$ are different quantities.

Example Let X, Y denote the values of the red and green die as above and consider the events

$$A = \{X = 2\}, \quad B = \{X + Y = 8\},$$

and recall that $\mathbb{P}(A) = 1/6, \mathbb{P}(B) = 5/36$. Note that $A \cap B = \{X = 2, Y = 6\}$ and hence $\mathbb{P}(A \cap B) = 1/36$. Therefore,

$$\mathbb{P}(B \mid A) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(A)} = \frac{1}{6}, \qquad \mathbb{P}(A \mid B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} = \frac{1}{5}.$$

One of the most important rules in probability goes under the name the law of total probability. A collection of events $A_1, A_2, \ldots, A_n$ is called a partition of the sample space if exactly one of the events will occur, in other words,

$$A_j \cap A_k = \emptyset \text{ if } j \neq k, \qquad \bigcup_{j=1}^{n} A_j = \Omega.$$

If B is another event, then $A_1 \cap B, A_2 \cap B, \ldots, A_n \cap B$ are disjoint events whose union is B and the rules for probability tell us that

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

If $\mathbb{P}(A_j) > 0$, we can write $\mathbb{P}(A_j \cap B) = \mathbb{P}(A_j)\,\mathbb{P}(B \mid A_j)$.

Fact If $A_1, \ldots, A_n$ is a partition of the sample space, and B is another event,

$$\mathbb{P}(B) = \sum_{j=1}^{n} \mathbb{P}(A_j)\,\mathbb{P}(B \mid A_j). \qquad \text{(Law of Total Probability)}$$

This also holds if we have a countable partition $A_1, A_2, \ldots$.

Example We will do a complicated example using the law of total probability to compute the probability of winning the game of craps. We start by giving the rules. The player starts by rolling two dice, and let X be the sum of the two rolls. Suppose that $X = k$.

- If $k = 2, 3, 12$, the player has lost the game.
- If $k = 7, 11$, the player has won the game.
- If $k = 4, 5, 6, 8, 9, 10$, then the number k becomes the player's "point" and the game continues.

If the game continues, the player rolls the dice until the player rolls either a 7 or the point k. In this case

- If the player rolls k first, the player wins.

- If the player rolls 7 first, the player loses.

We let W denote the event that the player wins the game, and we consider the partition $A_2, A_3, \ldots, A_{12}$ where A_k is the event that the player's first roll is k. The law of total probability gives

$$\mathbb{P}(W) = \sum_{j=2}^{12} \mathbb{P}(A_j)\,\mathbb{P}(W \mid A_j). \tag{2}$$

We know how to calculate $\mathbb{P}(A_j)$. Also, it is immediate from the rules for the game that

$$\mathbb{P}(W \mid A_k) = \begin{cases} 0, & k = 2, 3, 12 \\ 1, & k = 7, 11 \end{cases}.$$

What is not immediate is the value of $\mathbb{P}(W \mid A_k)$ when $k = 4, 5, 6, 8, 9, 10$. We will find that by doing a separate problem.

Suppose we repeat an experiment such that at each time one of three outcomes happens: I,II,III with the probability of I being p, II being q and III being $1 - p - q$. We ask the question, what is the probability of a I outcome before the first II outcome? Again we use the law of total probability using the partition

$$C_1 = \{\text{first outcome is I}\}, \qquad C_2 = \{\text{first outcome is II}\}, \qquad C_3 = \{\text{first outcome is III}\}.$$

If B is the event that we get a I before a II, then

$$\mathbb{P}(B) = \mathbb{P}(C_1)\,\mathbb{P}(B \mid C_1) + \mathbb{P}(C_2)\,\mathbb{P}(B \mid C_2) + \mathbb{P}(C_3)\,\mathbb{P}(B \mid C_3).$$

Clearly $\mathbb{P}(B \mid C_1) = 1, \mathbb{P}(B \mid C_2) = 0$. Finding $\mathbb{P}(B \mid C_3)$ is as hard as the original problem — indeed, if one thinks about it, it is exactly the same problem, $\mathbb{P}(B \mid C_3) = \mathbb{P}(B)$. Plugging in we get

$$\mathbb{P}(B) = p \cdot 1 + q \cdot 0 + (1 - p - q)\,\mathbb{P}(B)$$

giving

$$\mathbb{P}(B) = \frac{p}{p + q}.$$

Using this we get

$$\mathbb{P}(W \mid A_4) = \mathbb{P}(W \mid A_{10}) = \frac{3}{9}, \qquad \mathbb{P}(W \mid A_5) = \mathbb{P}(W \mid A_9) = \frac{4}{10},$$

$$\mathbb{P}(W \mid A_6) = \mathbb{P}(W \mid A_8) = \frac{5}{11}.$$

We can plug the appropriate numbers back into (2) to get

$$\mathbb{P}(W) = .4929292929\cdots.$$

This is the closest thing to a fair game as you will find in a casino.

There is an immediate corollary of the law of total tricks that is better known, Bayes theorem or Bayes rule. It follows from the fact that we can write

$$\mathbb{P}(A_j \cap B) = \mathbb{P}(B)\,\mathbb{P}(A_j \mid B) = \mathbb{P}(A_j)\,\mathbb{P}(A_j \mid B).$$

Fact If $A_1, \ldots, A_n$ is a partition of the sample space, and B is another event,

$$\mathbb{P}(A_j \mid B) = \frac{\mathbb{P}(A_j)\,\mathbb{P}(B \mid A_j)}{\mathbb{P}(A_1)\,\mathbb{P}(B \mid A_1) + \cdots + \mathbb{P}(A_n)\,\mathbb{P}(B \mid A_n)} \qquad \text{Bayes Theorem.}$$

Example One of the most common applications of Bayes theorem is to medical diagnosis. Here is a simple example. Suppose there is a disease that affects .01% of the population and there is a test that always comes out positive if the patient has the disease but for which there is a 5% chance of a "false positive" for patients without the disease. Suppose a patient tests positive — what is the probability that the patient has the disease? Letting A be the event that the test is positive, D the event that the patient has the disease, and O the event that there is no disease, we have

$$\mathbb{P}(A \mid D) = 1, \quad \mathbb{P}(A \mid O) = .05, \quad \mathbb{P}(D) = 1 - \mathbb{P}(O) = .0001$$

and

$$\mathbb{P}(D \mid A) = \frac{\mathbb{P}(D)\,\mathbb{P}(A \mid D)}{\mathbb{P}(D)\,\mathbb{P}(A \mid D) + \mathbb{P}(O)\,\mathbb{P}(A \mid O)} = \frac{(.0001)\cdot 1}{.0001 \cdot 1 + (.9999)\cdot .05} = .001996\cdots .$$

1.3 Independence of random variables and variance

The definition of independence to discrete random variables is essentially the same as for events. A collection of random variables $\{X_j\}$ is *(mutually) independent* if for each j, no information about the other random variables is useful in predicting the value of X_j.

Definition A collection of random variables $\{X_j\}$ is (mutually) independent if for every distinct $j_1, \ldots, j_k$ and subsets $V_1, \ldots, V_k$ of the real line

$$\mathbb{P}\{X_{j_1} \in V_1, \ldots, X_{j_k} \in V_k\} = \mathbb{P}\{X_{j_1} \in V_1\}\,\mathbb{P}\{X_{j_2} \in V_2\}\cdots\mathbb{P}\{X_{j_k} \in V_k\}.$$

We say that the collection is pairwise independent if for each $j \neq k$, X_j and X_k are independent random variables.

Fact If X and Y are independent discrete random variables, then

$$\mathbb{E}[XY] = \mathbb{E}[X]\,\mathbb{E}[Y] \tag{3}$$

This is not hard to show. We give the derivation to show where independence is used.

$$\begin{aligned}\mathbb{E}[XY] &= \sum_{a,b} ab\,\mathbb{P}\{X=a, Y=b\} \\ &= \sum_{a,b} ab\,\mathbb{P}\{X=a\}\,\mathbb{P}\{Y=b\} \qquad \text{(Independence)} \\ &= \left[\sum_a a\,\mathbb{P}\{X=a\}\right]\left[\sum_b b\,\mathbb{P}\{Y=b\}\right] = \mathbb{E}[X]\,\mathbb{E}[Y].\end{aligned}$$

It is possible for the product rule (3) to hold without the random variables being independent. As a simple example, suppose that $\mathbb{P}\{X=1\} = \mathbb{P}\{X=2\} = 1/2$ and Y is an independent random variable with $\mathbb{P}\{Y=1\} = \mathbb{P}\{Y=-1\} = 0$. Let $Z = XY$. Then we leave it to you to check that X and Z are not independent but $\mathbb{E}[Z] = 0, \mathbb{E}[XZ] = 0$ and hence X, Z satisfy (3).

Definition A collection of random variables $\{X_j\}$ is called an uncorrelated collection if for every $j \neq k$, $\mathbb{E}[X_j\,X_k] = \mathbb{E}[X_j]\,\mathbb{E}[X_k]$.

Independent random variables are uncorrelated but uncorrelated does not imply independent.

Definition The variance of a random variable X is defined by

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

The last equality follows from the simple calculation

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

Two simple but important properties of the variance are: if a, b are real numbers

$$\mathrm{Var}[aX] = a^2\,\mathrm{Var}[X], \qquad \mathrm{Var}[X+b] = \mathrm{Var}[X].$$

Definition The (theoretical) standard deviation of a random variable X, is given by

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

The first definition for variance shows that $\mathrm{Var}[X] \geq 0$ for all X with $\mathrm{Var}[X] = 0$ only for the trivial random variable that takes on only a single value. We can use the second definition to see that $\mathbb{E}[X^2] \geq \mathbb{E}[X]^2$. More generally if $q \geq 1$, $\mathbb{E}[|X|^q] \geq \mathbb{E}[|X|]^q$.

We use the term theoretical standard deviation to distinguish this from a statistical or sample standard deviation that we will describe later. When people talk about the standard deviation of a collection of data they are referring to the sample standard deviation.

There are many possible ways to measure the expected deviation from the mean for a random variable. The variance is chosen because of its nice properties, in particular, because of the next proposition.

Most of the times the following proposition is used for independent random variables but we state it more generally for uncorrelated random variables.

Fact If $X_1, \ldots, X_n$ are uncorrelated random variables, then

$$\text{Var}\left[X_1 + \cdots + X_n\right] = \text{Var}\left[X_1\right] + \text{Var}\left[X_2\right] + \cdots + \text{Var}\left[X_n\right].$$

In particular, this holds if $X_1, \ldots, X_n$ are independent.

This follows from the product rule for expectation as we now demonstrate. Suppose for ease that $\mathbb{E}[X_j] = 0$ for each j. If not, let $Y_j = X_j - \mathbb{E}[X_j]$. Then

$$\begin{aligned}
\text{Var}[X_1 + \cdots + X_n] &= \mathbb{E}[(X_1 + \cdots X_n)^2] \\
&= \sum_{j=1}^{n} \mathbb{E}[X_j^2] + \sum_{j \neq k} \mathbb{E}[X_j\, X_k] \\
&= \sum_{j=1}^{n} \mathbb{E}[X_j^2] + \sum_{j \neq k} \mathbb{E}[X_j]\, \mathbb{E}[X_k] \qquad \text{(Uncorrelated)} \\
&= \sum_{j=1}^{n} \text{Var}[X_j].
\end{aligned}$$

We will introduce a notion of conditional expectation — this idea will be expanded later. Suppose X is a random variable and $A_1, A_2, \ldots$ is a partition of the probability space. If $\mathbb{P}(A_j) > 0$, we define the conditional expectation of X given A_j by

$$E\left[X \mid A_j\right] = \frac{\mathbb{E}\left[X\, 1_{A_j}\right]}{\mathbb{P}(A_j)}.$$

We do not define this if $\mathbb{P}(A_j) = 0$. Here we have introduced a notation. If A is an event, then the indicator function 1_A is the random variable that equals 1 if A occurs and equals 0 if A does not occur. Hence $X\, 1_{A_j}$ equals X if A_j occurs and otherwise equals zero. We then have the law of total expectation.

Fact If $A_1, A_2, \ldots$ is a partition of the probability space and X is a random variable, then

$$\mathbb{E}\left[X\right] = \sum_{j=1}^{\infty} \mathbb{P}(A_j)\, E[X \mid A_j]. \qquad \text{Law of Total Expectation}$$

In this sum, if $\mathbb{P}(A_j) = 0$, we let $\mathbb{P}(A_j)\, E[X \mid A_j] = 0$ even if $E[X \mid A_j]$ is not defined.

Example Suppose we roll two dice and get a number k and then we flip a fair coin k times. What is the expected number of heads? If we let X be the number of heads and let A_k be the event that k is rolled, then

$$E(X \mid A_k) = \frac{k}{2},$$

and hence

$$\mathbb{E}[X] = \sum_{j=2}^{12} \mathbb{P}(A_j)\, E(X \mid A_j) = \frac{7}{2}.$$

Let us write this somewhat differently. Let $\mathcal{A}$ denote the partition and define the random variable $E[X \mid \mathcal{A}]$ to be the random variable that outputs $E[X|A_j]$ if A_j occurs. The random variable $E[X \mid \mathcal{A}]$ is $\mathcal{A}$-measurable. This means that if you know which of the events $A_1, A_2, \ldots$ occurred, then you know the value of $E[X \mid \mathcal{A}]$.

We will say that an event V is $\mathcal{A}$-measurable if we can write

$$V = \bigcup_{k=1}^{\infty} A_{j_k},$$

that is, if V is a union of sets in the partition. Then one can check from the definition that

$$\mathbb{E}\left[1_V\, X\right] = \mathbb{E}\left[1_V\, E[X \mid \mathcal{A}]\right].$$

We will see that these two properties characterize conditional expectation.

1.4 Some discrete distributions

1.4.1 Indicator random variables

Definition If A is an event then the indicator function or indicator random variable associated to A is the random variable 1_A which equals one if A occurs and 0 if A does not occur.

Note that

$$\mathbb{E}[1_A] = \mathbb{P}(A), \qquad \mathrm{Var}[1_A] = \mathbb{E}[1_A^2] - \mathbb{E}[1_A]^2 = \mathbb{P}(A) - \mathbb{P}(A)^2 = \mathbb{P}(A)\,[1 - \mathbb{P}(A)].$$

If A is an event written say as $A = \{X = 2, Y = 3\}$ we often write $1\{X = 2, Y = 3\}$ for 1_A.

1.4.2 Uniform distribution

We have already stated that a random variable has a uniform distribution on a finite set if it gives the same probability to each element. Let us consider a uniform random variable X on $\{1, \ldots, n\}$. The die roll we have considered is the case $n = 6$. Then

$$\mathbb{E}[X] = \sum_{j=1}^{n} j \cdot \frac{1}{n} = \frac{1}{n}\,\frac{n(n+1)}{2} = \frac{n+1}{2},$$

$$\mathbb{E}[X^2] = \sum_{j=1}^{n} j^2 \frac{1}{n} = \frac{1}{n}\sum_{j=1}^{n} j^2 = \frac{1}{n} \cdot \frac{n(n+1)(2n+1)}{6} = \frac{(n+1)(2n+1)}{6}.$$

$$\mathrm{Var}[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \frac{(n+1)(2n+1)}{6} - \left(\frac{n+1}{2}\right)^2 = \frac{n^2-1}{12}.$$

1.4.3 Binomial distribution

Bernoulli trials are repeated independent experiments each with the same probability $p \in (0,1)$ of success. There are several distributions coming from Bernoulli trials.

The binomial distribution gives the number of successes in n Bernoulli trials.

Definition A random variable X has a binomial distribution with parameters n and p if

$$\mathbb{P}\{X = k\} = \binom{n}{k} p^k (1-p)^{n-k}, \qquad k = 0, 1, 2, \ldots, n.$$

Recall that $\binom{n}{k}$ is the binomial coefficient

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

It represents the number of different ways that we can arrange k S's (S = success) and $(n-k)$ F's (F = failure). For example if $n = 10$ and $k = 6$ one possibility is SFFFSSSSFS. Each particular ordering of S's and F's has probability $p^k (1-p)^{n-k}$ of happening (this uses the independence assumption) so the total probability is the number of orderings times the probability for each one.

We will compute the expectation and variance. We could do this directly from the definition using

$$\mathbb{E}[X] = \sum_{k=0}^{n} k \binom{n}{k} p^k (1-p)^{n-k}, \qquad \mathbb{E}[X^2] = \sum_{k=0}^{n} k^2 \binom{n}{k} p^k (1-p)^{n-k}$$

and $\mathrm{Var}[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2$ but these sums are a little tricky. We use a simpler approach with indicator random variables. Let I_j denote the indicator function of the event that the jth trial is a success. Then the Bernoulli assumptions imply that $I_1, \ldots, I_n$ are independent, and by definition we see that

$$X = I_1 + I_2 + \cdots + I_n.$$

We have seen that $\mathbb{E}[I_j] = p, \mathrm{Var}[I_j] = p(1-p)$. Therefore,

$$\mathbb{E}[X] = \mathbb{E}[X_1] + \cdots + \mathbb{E}[X_n] = np,$$

$$\mathrm{Var}[X] = \mathrm{Var}[X_1] + \cdots + \mathrm{Var}[X_n] = np(1-p).$$

1.4.4 Geometric distribution

There are two very similar distributions that go under the name geometric distribution. They also use the assumption of Bernoulli trials and give the amount of time until a success.

$$X = \text{number of trials until a success}, \qquad Y = \text{number of failures before a success.}$$

Note that $X = Y + 1$; the only difference between the two is whether one includes the first success in the total time. The word geometric can be used for either one so one must be careful when reading which is being used by a particular author.

Definition A random variable X has a geometric distribution representing the number of trials until a success if it has distribution

$$\mathbb{P}\{X = k\} = (1-p)^{k-1}\, p, \qquad k = 1, 2, 3, \ldots.$$

The random variable $Y = X - 1$ has a geometric distribution representing the number of failures until a success and has distribution

$$\mathbb{P}\{Y = k\} = (1-p)^{k}\, p, \qquad k = 0, 1, 2, \ldots.$$

The rules for expectation and variance tell us that $\mathbb{E}[Y] = \mathbb{E}[X] - 1$ and $\mathrm{Var}[Y] = \mathrm{Var}[X]$ so it suffices to compute $\mathbb{E}[X], \mathrm{Var}[X]$. This can be done directly

$$\mathbb{E}\left[X\right] = \sum_{k=1}^{\infty} k\,(1-p)^{k-1}\, p = \frac{1}{p},$$

$$\mathbb{E}\left[X^2\right] = \sum_{k=1}^{\infty} k^2\,(1-p)^{k-1}\, p = \frac{2-p}{p^2},$$

$$\mathrm{Var}[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \frac{1-p}{p^2}.$$

$$\mathbb{E}[Y] = \mathbb{E}[X] - 1 = \frac{1-p}{p}, \qquad \mathrm{Var}[Y] = \mathrm{Var}[X] = \frac{1-p}{p^2}.$$

There is a slick way to get $\mathbb{E}[X]$ by noting that

$$\mathbb{E}[X] = 1 + \,(1-p)\,\mathbb{E}[X],$$

and then solving for $\mathbb{E}[X]$.

There is a trick for computing these infinite sums. If $0 < p < 1$ the geometric series tells us that

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

We can differentiate both sides with respect to p twice to get the formulas

$$\sum_{k=1}^{\infty} k\, p^{k-1} = \frac{1}{(1-p)^2}, \qquad \sum_{k=2}^{\infty} k\,(k-1)\, p^{k-2} = \frac{2}{(1-p)^3}.$$

These give

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

$$\sum_{k=0}^{\infty} k^2\; p^{k-1}\,(1-p) = \frac{2p}{(1-p)^2} + \sum_{k=0}^{\infty} k\, p^{k-1}\,(1-p) = \frac{2p}{(1-p)^2} + \frac{1}{1-p} = \frac{p+1}{(1-p)^2}.$$

1.4.5 Poisson distribution

This section will use two facts about the exponential function from calculus, the exponential series

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

and the limit

$$\lim_{n\to\infty} \left(1 + \frac{x}{n}\right)^n = e^x.$$

The Poisson distribution describes "rare events". It can be considered as a limit of the binomial distribution where the number of trials n goes to infinity while the probability p of success goes to zero in a way so that the expected number of successes np stays fixed at some constant $\lambda > 0$. If we set $p = \lambda/n$ then the probability of exactly k successes is

$$\binom{n}{k} p^k\,(1-p)^{n-k} = \binom{n}{k} \left(\frac{\lambda}{n}\right)^k \left(1 - \frac{\lambda}{n}\right)^{n-k}.$$

The Poisson distribution is obtained by taking the limit of this expression as n goes to infinity with k, λ fixed. After a little manipulation, the right-hand side can be written as

$$\frac{\lambda^k}{k!} \left[\frac{n}{n} \cdot \frac{n-1}{n} \cdots \frac{n-k+1}{n}\right] \left[1 - \frac{\lambda}{n}\right]^{-k} \left(1 - \frac{\lambda}{n}\right)^n.$$

If we fix λ, k and let n go to infinity, the first term does not depend on n, the middle two terms go to 1, and the third term becomes a familiar exponential limit from calculus.

Definition A random variable X has a Poisson distribution with parameter (mean) λ if

$$\mathbb{P}\{X = k\} = e^{-\lambda}\frac{\lambda^k}{k!}, \qquad k = 0, 1, 2, \ldots.$$

This is a well-defined distribution because

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

Also,

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

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

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

1.5 Some problems and brain teasers

1.5.1 St. Petersburg Paradox

One interpretation of expectation is the amount of money that one would bet in a game so that the payoff would be a fair game. However, one must be careful with this interpretation. Suppose we flip a fair coin until it comes up heads and let T denote the number of tails that were flipped before getting the heads. Let $X = 2^T$. Note that $T = k$ corresponds to the first k flips being tails and the $(k+1)$st flip being heads; therefore,

$$\mathbb{P}\{T = k\} = \left(\frac{1}{2}\right)^{k+1}, \qquad k = 0, 1, 2, 3, \ldots,$$

and hence

$$\mathbb{P}\{X = 2^k\} = \left(\frac{1}{2}\right)^{k+1}, \qquad k = 0, 1, 2, 3, \ldots.$$

and

$$\mathbb{E}[X] = 1\cdot\frac{1}{2} + 2\cdot\frac{1}{4} + 4\cdot\frac{1}{8} + 8\cdot\frac{1}{16} + \cdots = \infty \quad !$$

This seems to indicate that one would be willing to bet any amount of money in order to play a game with this payoff! Of course, this is not correct. Here are several reasons:

- The mathematical modeling of the problem has assumed that the "utility function of money is linear" which is a fancy way of saying that 2 billion dollars is twice as valuable to a person as 1 billion dollars.

- Our idealized problem has assumed that the casino you are playing against has an infinite amount of money. Suppose that the casino has assets of only a billion dollars. In this case, your payoff will be the smaller of X and one billion dollars. In this case, since $2^{29} < 1,000,000,000 < 2^{30}$, if $T < 30$, we get 2^T but if $T \geq 30$ we get "only" a billion dollars. The expected value is

$$1 \cdot \frac{1}{2} + 2 \cdot \frac{1}{4} + 4 \cdot \frac{1}{8} + \cdots + 29 \cdot \frac{1}{2^{30}} + (1,000,000,000) \cdot \frac{1}{2^{30}} < 15.$$

1.5.2 Monty Hall problem

The *Monty Hall* or *"Let's Make a Deal"* problem is a well known problem that comes from a simplification of a TV game show. Monty Hall was the host of Let's Make a Deal. Although this is not exactly what happened on the show, we will give the problem as usually stated. There are three curtains numbered $1, 2, 3$ and there is a prize behind one of the curtains. A contestant chooses a curtain and before Monty opens the curtain he opens one of the other curtains and shows that there is no prize behind that curtain. He then gives the contestant the chance to change their mind and choose the other unopened curtain. Should the contestant do this?

As stated the question is incomplete. As is often the case in the real world, there were some implicit assumptions being made and since we are being mathematical we will make them explicit. Assume that:

- The prize is equally likely to be behind any of the three curtains.
- Monty Hall knows which curtain the prize is behind.
- No matter which curtain the contestant chooses, Monty Hall always opens a different curtain that does not have the prize behind it. (He can do this since he knows where the prize is.)

These assumptions are not good enough to determine the conditional probability that the prize is behind the curtain the contestant originally chose. There is another implicit assumption that needs to be made esplicit.

- If the contestant has chosen the correct curtain at the beginning so that Monty has two choices for which curtain to open, he is equally likely to pick either one.

For ease, let us assume that the contestant chooses curtain 1 at the beginning. Let A_j be the event that the prize is behind curtain j and let O_k be the probability that Monty opened curtain k. Then our assumptions give

$$\mathbb{P}(A_1) = \mathbb{P}(A_2) = \mathbb{P}(A_3) = \frac{1}{3},$$

$$\mathbb{P}(O_1 \mid A_k) = 0, \quad k = 1, 2, 3,$$

$$\mathbb{P}(O_2 \mid A_1) = \frac{1}{2}, \quad \mathbb{P}(O_2 \mid A_2) = 0, \quad \mathbb{P}(O_2 \mid A_3) = 1,$$
$$\mathbb{P}(O_3 \mid A_1) = \frac{1}{2}, \quad \mathbb{P}(O_3 \mid A_2) = 1, \quad \mathbb{P}(O_3 \mid A_3) = 0.$$

Therefore, if Monty Hall opens curtain number 2,

$$\mathbb{P}(A_2 \mid O_2) = \frac{\mathbb{P}(A_2)\,\mathbb{P}(O_2 \mid A_2)}{\mathbb{P}(A_1)\,\mathbb{P}(O_2 \mid A_1) + \mathbb{P}(A_2)\,\mathbb{P}(O_2 \mid A_2) + \mathbb{P}(A_3)\,\mathbb{P}(O_2 \mid A_3)} = \frac{2}{3}.$$

So if all our assumptions are correct, one should choose the other curtain.

1.5.3 Hat check problem

Suppose there is an entering class of $N = 120$ masters students at a university and they all need to be handed their new ID card. A lazy administrator decided to take the ID cards and hand them out completely randomly. What is the probability that anyone gets their own card in this procedure?

This is a bit tricky, but it is much easier to compute the *expected* number of students who get their correct card. We will do it for general N. Let I_j be the indicator function that that the jth student receives their correct card. Clearly the probability that this happens is $1/N$ and hence $\mathbb{E}[I_j] = 1/N$. If X denotes the total number of students that get their own card, then

$$\mathbb{E}\,[X] = \mathbb{E}\left[\sum_{j=1}^{N} I_j\right] = \sum_{j=1}^{N} \mathbb{E}[I_j] = \sum_{j=1}^{N} \frac{1}{N} = 1.$$

Are the random variables $I_1, I_2, \ldots, I_N$ independent? The answer is "almost but not quite" if N is large. For the moment let us assume that they were and see what we can do. If A_j denotes the event that the jth student does not get the correct card, then

$$\mathbb{P}\{X = 0\} = \mathbb{P}(A_1 \cap \cdots \cap A_N) \approx \mathbb{P}(A_1)\,\mathbb{P}(A_2)\,\cdots\,\mathbb{P}(A_N) = \left(\frac{N-1}{N}\right)^N.$$

We wrote $\approx$ to indicate that this step is not exact — the product rule would require independence. But if it were true we would have

$$\lim_{N\to\infty} \mathbb{P}\{X = 0\} = \lim_{N\to\infty} \left(\frac{N-1}{N}\right)^N = \lim_{N\to\infty} \left(1 - \frac{1}{N}\right)^N = \frac{1}{e}.$$

We can give an exact answer using the inclusion-exclusion principle. We will not prove it here but we will give the idea. Let V_j be the event that the jth student gets the correct card. Then the probability that at least one person gets their card is $\mathbb{P}(V_1 \cup \cdots \cup V_N)$. As a first approximation, we might estimate this by

$$\sum_{j=1}^{N} \mathbb{P}(V_j).$$

But this will be an overestimate because there may be more than one student who gets their correct card. To compensate we subtract the pairwise intersections

$$-\sum_{j\neq k}\mathbb{P}(V_j\cap V_k).$$

This subtracts too much and we have to compensate by adding the probabilities of triple intersections:

$$+\sum_{j,k,l}\mathbb{P}(V_j\cap V_k\cap V_l)$$

where this is the sum over all distinct j,k,l;.

Fact Inclusion-Exclusion Principle

$$\mathbb{P}(V_1\cup\cdots\cup V_N)=\sum_{k=1}^{N}(-1)^{k+1}\sum_{j_1,\ldots,j_k}\mathbb{P}(V_{j_1}\cap V_{j_2}\cap\cdots\cap V_{j_k})$$

where the sum is over all subsets $\{j_1,\ldots,j_k\}$ of $\{1,2,\ldots,N\}$ with exactly k elements.

The inclusion-exclusion formula is generally too bulky to give precise answers for large N, but this problem is an exception. If we choose k students, then the probability that they all get the correct card is

$$\frac{1}{N}\,\frac{1}{N-1}\,\frac{1}{N-2}\cdots\frac{1}{N-k+1}$$

This can be seen by conditional probabilities. For example the probability that the first student gets the correct card is $1/N$ and the conditional probability that the second student gets the correct card given that the first student gets the right card is $1/(N-1)$. The number of k element subsets is $\binom{N}{k}$ and hence for our problem, the probability that no one gets their correct card is

$$\begin{aligned}1-\mathbb{P}(V_1\cup\cdots\cup V_N)&=1+\sum_{k=1}^{N}(-1)^k\sum_{j_1,\ldots,j_k}\mathbb{P}(V_{j_1}\cap V_{j_2}\cap\cdots\cap V_{j_k})\\&=1+\sum_{k=1}^{N}(-1)^k\binom{N}{k}\frac{1}{N\cdot(N-1)\cdots(N-k+1)}\\&=\sum_{k=0}^{N}\frac{(-1)^k}{k!}.\end{aligned}$$

The exact answer is the first $N+1$ terms of the exponential series for e^{-1}. While this is not exactly equal to e^{-1} it is very close even for relatively small N; indeed,

$$\left|\frac{1}{e}-\sum_{k=0}^{N}\frac{(-1)^k}{k!}\right|\leq\frac{1}{(N+1)!}.$$

This problem is often phrased in terms of individuals giving their hats at a hat-check and having them returned randomly. This is why it is often called the hat-check problem.

1.5.4 Time until a pattern appears

Suppose one flips a fair coin until one flips three heads in a row. What is the expected number of flips until one stops? What about for j heads in a row?

There are various ways of doing this problem. One approach is to let $e(j)$ be the expected number of steps until we have had a string of j heads. In order to have a string of j heads, one need to get a string of $j-1$ heads first. Then either we will get a heads and succeed, or we get a tails and we are back to square one. We can write this as

$$e(j) = e(j-1) + 1 + \frac{1}{2}\, e(j), \qquad e(j) = 2\, e(j-1) + 2.$$

This is a recursive equation. We have $e(0) = 0, e(1) = 2$ and we can keep on going. If we are only interested in $e(3)$ we can just plug in

$$e(2) = 2\, e(1) + 2 = 6, \qquad e(3) = 2\, e(2) + 2 = 14.$$

To try to get a closed formula, we see that $e(j)$ is approximately doubling at each stage, so let $g(j) = 2^{-j}\, e(j)$. then the recursion becomes

$$g(j) = g(j-1) + 2^{-(j-1)}$$

and hence

$$g(j) = \sum_{k=0}^{j-1} 2^{-k} = 2 - 2^{-(j-1)}, \qquad e(j) = 2^j\, g(j) = 2^{j+1} - 2.$$

1.5.5 Two envelopes problem

We will give a paradox here that we will not completely resolve (this makes it more fun!). Here is the problem:

- I choose a random positive number that I will call N.
- After doing so, I write the numbers N and $2N$ on pieces of identical paper, place them in identical envelopes and seal the envelopes.
- You have a simple game — you choose an envelope (without opening it) and after opening it you receive the amount of money in the envelope.

This is pretty dull but we add a slight twist similar to the Let's Make a Deal game. After choosing the envelope but before opening it, I give you the opportunity to switch envelopes. Should you switch?

Let J be the amount of money in the envelope that you have chosen and let K be the amount in the other envelope. Note that J is equally likely to be either N or $2N$. Therefore

K is equally to be $J/2$ or $2J$. If you keep your envelope, you will receive J while your expected amount if you switch envelopes is

$$\mathbb{E}[K] = \frac{1}{2} \cdot \frac{J}{2} + \frac{1}{2} \cdot (2J) = \frac{5}{4} J.$$

Therefore, it must be right to switch envelopes!

However, since you always switch, after you switch (and before you open) the number you hold is either N or $2N$. I ask you again if you want to switch back and the same argument tell you that switching back has an expectation of $(5/4)$ times keeping the envelope. By doing this twice we now have increased our expectation by a factor of $(5/4)^2$, but we have exactly the same number that we started with! What's wrong here?

Part of the problem comes with the idea "choose a number at random". There is no way to choose a number uniformly over all real numbers. I must be using some distribution. For example, suppose that I choose N uniformly on $\{1, \ldots, 100\}$. The number J can then be in $A_O := \{1, 3, 5, \ldots, 99\}$, $A_E := \{2, 4, 6, \ldots, 100\}$ or $B := \{102, 104, 106, \ldots, 200\}$. If $k \in A_O$, then the only way for $J = k$ is to choose $N = k$. Similarly if $k \in B$, then the only way that $J = k$ is that $N = k/2$ and we have chosen $2N$. For $k \in A_E$, there are two ways that J can equal k: either $N = k$ or $N = k/2$. From this we see that

$$\mathbb{P}\{J = k\} = \begin{cases} \frac{1}{200} & k \in A_O \cup B, \\ \frac{1}{100} & k \in A_E, \end{cases}$$

Using the rule for conditional expectation, we get that $\mathbb{E}[K]$ equals

$$\sum_k \mathbb{P}\{J = k\}\, \mathbb{E}[K \mid J = k\} = \sum_{k \in A_O} \mathbb{P}\{J = k\}\, 2k + \sum_{k \in A_E} \mathbb{P}\{J = k\}\, \frac{5k}{4} + \sum_{k \in B} \mathbb{P}\{J = k\}\, \frac{k}{2}.$$

which one can check is the same as $\mathbb{E}[J]$ which is the same as $(3/2)\, \mathbb{E}[N]$.

1.5.6 Secretary problem

Suppose someone will be interviewing N candidate for a job (traditionally called a secretarial position). The (rather implausible) assumptions are the following: the interviewer will interview the candidate one by one. The interviewer does not know the quality of the candidates ahead of time, but after each interview, the interviewer will know if the current candidate is better than all of the other candidates interviewed. The interviewer will have the option to offer the job to that candidate immediately but it must be done before seeing any of the other candidates. If the interviewer does not immediately offer the job, then the person will be unavailable later to accept it.

What is the optimal strategy for the interviewer? We will make the assumption that the candidates are in a random order. However, one cannot do optimization problems unless one decides what quantity are we trying to optimize? For a secretarial position, one would like to get the best candidate, but in most cases the second or third best candidate would also be

suitable for the position. However, in this problem our goal is to maximize the probability that the best candidate is chosen.

Suppose we decide that for large N the best strategy is to see the first bN candidates, accept none of them, and then choose the first candidate after that who is better than all the previous candidates. What is the probability that the best candidate is chosen? We partition into the events

$$A_0 = \{\text{best candidate occurs before time } bn\},$$

$$A_1 = \{\text{best occurs after } bn \text{ but second best occurs before}\}$$

$$A_2 = \{\text{best two occur after } bn \text{ but the third best occurs before}\}$$

$$\vdots$$

If we let B be the event that the best candidate is chosen we have

$$\mathbb{P}(B \mid A_0) = 0, \quad \mathbb{P}(B \mid A_1) = 1, \quad \mathbb{P}(B \mid A_2) = \frac{1}{2}, \quad \cdots, \quad \mathbb{P}(B \mid A_n) = \frac{1}{n}.$$

Also, in the limit as $N \to \infty$

$$\mathbb{P}(A_n) = b\,(1-b)^n.$$

Therefore (in the limit) using the law of total probability,

$$\mathbb{P}(B) = \sum_{n=0}^{\infty} \mathbb{P}(A_n)\,\mathbb{P}(B \mid A_n) = \sum_{n=1}^{\infty} b(1-b)^n\,\frac{1}{n}.$$

The Taylor series for $\log(1-x)$ about $x = 0$ is

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

and hence we see that (again, in the limit as $N \to \infty$),

$$\mathbb{P}(B) = -b\,\log b.$$

To find which value of b maximizes this we use calculus

$$f(b) = -b\,\log b, \quad f'(b) = -\log b - 1.$$

Setting $f'(b) = 0$ gives $b = 1/e$. Therefore for large N, the optimal strategy is to interview the first N/e candidates and then choose the first person that is better than all the candidates seen at that point. The probability that this succeeds is $f(1/e) = 1/e$.

2 Continuous random variables

2.1 Density and distribution function

Continuous random variables take values in $\mathbb{R}$; they have the property that for each real number the probability of getting exactly that number is 0. In order to describe the distribution of such a random variable, one uses the (cumulative) distribution function

$$F(t) = F_X(t) = \mathbb{P}\{X \le t\}.$$

The distribution function satisfies the following conditions:

$$\text{(Increasing)} \qquad F(s) \le F(t) \text{ if } s \le t,$$

$$F(-\infty) = 0, \qquad F(\infty) = 1,$$

$$\text{(Right continuous)} \qquad F(t) = F(t+) := \lim_{s \downarrow t} F(s).$$

A random variable is continuous if and only if the distribution function is continuous. We will restrict ourselves to a subclass of continuous functions, those with densities.

Definition The (probability) density (function) or pdf of a continuous random variable X is a function f such that for all $a < b$,

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

Since X is a continuous random variable, the probability of getting exactly a or b is zero and so we could also write the left-hand side as $\mathbb{P}\{a \le X \le b\}$. The conditions that a density satisfy are

$$f(x) \ge 0 \text{ for all } x, \qquad \int_{-\infty}^{\infty} f(x)\,dx = 1.$$

Definition A random variable has a uniform distribution on the interval $[a, b]$ if it has density

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

Here we use a standard practice to define f only at the places where it is nonzero. It is implicit that $f(x) = 0$ if $x < a$ or $x > b$. Also, since the integral of a function does not change if we change the value at a single point, it does matter what the values of $f(a), f(b)$ are. If we know the density, then we can get the distribution function immediately,

$$F(t) = \mathbb{P}\{X \le t\} = \int_{-\infty}^{t} f(x)\,dx. \tag{4}$$

For example, for the uniform distribution on $[a, b]$ we have

$$F(t) = \begin{cases} 0, & t \le a \\ \frac{t-a}{b-a}, & a \le t \le b \\ 1, & t \ge b \end{cases}.$$

Conversely, if we know the distribution function, we can get the density by differentiating (4) with respect to t. However, we have to be a little careful. The fundamental theorem of calculus tells us that the function $F(t)$ defined in (4) is differentiable in t only at those points where $f(t)$ is continuous.

Fact If F and f are the distribution function and the density, respectively, for a continuous random variable X, then

$$f(x) = F'(x)$$

at all x at which f is continuous.

One can check that this holds for the uniform distribution except at the points a and b which are the points at which f is not continuous. When trying to find the density of a random variable it is often easier to first find the distribution function and then to differentiate as this next example demonstrates.

Example Suppose X is a uniform random variable on $[0, 2]$. Let $Y = X^2$. Find the density of Y.

Note that

$$F_Y(y) = \mathbb{P}\{Y \le y\} = \mathbb{P}\{X \le \sqrt{y}\} = \begin{cases} 0, & y \le 0 \\ \frac{\sqrt{y}}{2}, & 0 \le y \le 4 \\ 1, & y \ge 4 \end{cases}.$$

Differentiating gives

$$f_Y(y) = F_Y'(y) = \frac{1}{4\sqrt{y}}, \quad 0 < y < 4.$$

The expectation and variance for continuous random variables with a density f are defined analogously to the discrete definitions, replacing the summation with an integral.

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x\, f(x)\, dx, \qquad \mathbb{E}[X^2] = \int_{-\infty}^{\infty} x^2\, f(x)\, dx,$$

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

and satisfy the same properties

$$\mathbb{E}[aX + bY] = a\,\mathbb{E}[X] + b\,\mathbb{E}[Y], \qquad \mathrm{Var}[aX + b] = \mathrm{Var}[aX] = a^2\,\mathrm{Var}[X].$$

For example, if X is uniform on $[0, 2]$,

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x\, f(x)\, dx = \int_0^2 x\, \frac{1}{2}\, dx = 1,$$

$$\mathbb{E}[X^2] = \int_{-\infty}^{\infty} x^2\, f(x)\, dx = \int_0^2 x^2\, \frac{1}{2}\, dx = \frac{4}{3}, \qquad \mathrm{Var}[X] = \frac{4}{3} - 1^2 = \frac{1}{3}.$$

2.1.1 Normal distribution

Probably the most important distribution in probability is the normal distribution which arises as the limit of the distribution of the average of many quantities. Here we will only define it, but later we will discuss the central limit theorem which discusses one way to take a limit.

Definition A random variable X has a normal (or Gaussian) distribution with mean μ and variance σ^2 if it has density

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-\mu)^2}{2\sigma^2}} = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{\frac{-(x-\mu)^2}{2\sigma^2}\right\}, \qquad -\infty < x < \infty.$$

If $\mu = 0, \sigma^2 = 1$, then X has a standard normal distribution. We will write $X \sim N(\mu, \sigma^2)$.

The notation $\exp(x)$ or $\exp\{x\}$ is the same as e^x. It is often used to aid reading when the expression in the exponent is complicated.

In order for this to be a probability density we need to show that it integrates to one. Here is a derivation of this for the standard normal using polar coordinates. Suppose

$$I = \int_{-\infty}^{\infty} e^{-\frac{x^2}{2}}\, dx.$$

Then

$$\begin{aligned} I^2 &= \left[\int_{-\infty}^{\infty} e^{-\frac{x^2}{2}}\, dx\right]\left[\int_{-\infty}^{\infty} e^{-\frac{y^2}{2}}\, dy\right] \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-\frac{x^2+y^2}{2}}\, dx\, dy \\ &= \int_{0}^{2\pi}\int_{0}^{\infty} e^{-\frac{r^2}{2}}\, r\, dr\, d\theta \\ &= 2\pi. \end{aligned}$$

Here we use the rule $dx\, dy = r\, dr\, d\theta$ and the extra factor of r makes the integral doable!

Let us consider the standard normal first. The density does not have an antiderivative that can be written in closed form. (It is actually a theorem in mathematics that one cannot write it in a nice form, so don't waste any time trying to find one!). We can still define the standard normal distribution function and we will use the letter Φ,

$$\Phi(t) = \mathbb{P}\{X \le t\} = \int_{-\infty}^{t} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}\, dx.$$

(Approximate) numerical values for Φ can be found either in tables or in computer packages. As would be expected from the definition,

$$\mathbb{E}[X] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} x\, e^{-\frac{x^2}{2}}\, dx = 0, \qquad \mathrm{Var}[X] = \mathbb{E}[X^2] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} x^2\, e^{-\frac{x^2}{2}}\, dx = 1.$$

The second integral can be done by integration by parts. For other values of μ, σ^2 we use the following simple fact.

Fact If X has a normal distribution with mean μ and variance σ^2 and $Y = (X - \mu)/\sigma$, then Y has a standard normal distribution. In particular, $\mathbb{E}[X] = \mu, \mathrm{Var}[X] = \sigma^2$ and X has distribution function

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

To see this we can just check the distribution function.

$$\begin{aligned}
\mathbb{P}\{Y \le y\} &= \mathbb{P}\left\{\frac{X - \mu}{\sigma} \le y\right\} \\
&= \mathbb{P}\{X \le \sigma y + \mu\} \\
&= \int_{-\infty}^{\sigma y + \mu} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} dx \\
&= \int_{-\infty}^{y} \frac{1}{\sqrt{2\pi}}\, e^{-\frac{t^2}{2}}\, dt.
\end{aligned}$$

The last step uses the change of variables $x = \mu + \sigma\, t$.

2.1.2 Exponential distribution

The exponential distribution comes from making the assumption of the memoryless property. Suppose T is the time at which some event will occur. The memoryless property states that if we have waited for t time units and the event has not occurred, then the probability that the event will occur in the next s times units is exactly the same as the probability it would have occurred in the first s time units. We write this assumption as

$$\mathbb{P}\{T > s + t \mid T > t\} = \mathbb{P}\{T > s\}.$$

Since,

$$\mathbb{P}\{T > s + t \mid T > t\} = \frac{\mathbb{P}\{T > s + t, T > t\}}{\mathbb{P}\{T > t\}} = \frac{\mathbb{P}\{T > s + t\}}{\mathbb{P}\{T > t\}},$$

we can write this as

$$G(s + t) = G(s)\, G(t) \qquad \text{where } G(t) = \mathbb{P}\{T > t\}.$$

The only continuous functions of t with this property satisfying $G(0) = 1, G(\infty) = 0$ are $G(t) = e^{-\lambda t}$.

Definition A random variable T has an exponential distribution with parameter (rate) $\lambda > 0$ if it has distribution function

$$F(t) = \mathbb{P}\{T \leq t\} = 1 - e^{-\lambda t}, \quad t \geq 0,$$

or equivalently, if it has density function

$$f(t) = \lambda\, e^{-\lambda t}, \quad t \geq 0.$$

By using integration by parts we can evaluate the integrals to show that

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} t\, \lambda\, e^{\lambda t}\, dt = \frac{1}{\lambda},$$

$$\mathbb{E}[X^2] = \int_{-\infty}^{\infty} t^2\, \lambda\, e^{\lambda t}\, dt = \frac{2}{\lambda^2}, \qquad \mathrm{Var}[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \frac{1}{\lambda^2}.$$

The expectation makes sense if we view T as the expected time until an event occurs when events occur at rate λ per time unit. In this case, the expected time for the next event is $1/\lambda$.

2.2 Joint densities and independence

Dealing with multiple random variables in the continuous case is very similar to the discrete case although the notation gets a little bulky. We will consider the case of a random vector (also called a random variable taking values in $\mathbb{R}^n$) $\boldsymbol{X} = (X_1, \ldots, X_n)$ where $X_1, \ldots, X_n$ are continuous random variables.

Definition

- The joint (cumulative) distribution function $F = F_{\boldsymbol{X}}$ is the function from $\mathbb{R}^n$ to $[0, 1]$ given by
$$F(t_1, \ldots, t_n) = \mathbb{P}\{X_1 \leq t_1, \ldots, X_n \leq t_n\}.$$
- The joint (probability) density (function) $f = f_{\boldsymbol{x}}$ is a function from $\mathbb{R}^n$ to $[0, \infty)$ such that for every subset $V \subset \mathbb{R}^n$,
$$\mathbb{P}\{\boldsymbol{X} \in V\} = \int_V f(\boldsymbol{x})\, d\boldsymbol{x}.$$
Here we write $\boldsymbol{x} = (x_1, \ldots, x_n)$ for a point in $\mathbb{R}^n$ and $d\boldsymbol{x}$ denotes usual integrals in $\mathbb{R}^n$.
- The distribution function and density of a particular component X_j are called marginal distribution function and marginal density.

One can get the joint distribution from the joint density and vice versa using

$$F(t_1, \ldots, t_n) = \int_{-\infty}^{t_n} \cdots \int_{-\infty}^{t_1} f(x_1, x_2, \ldots, x_n)\, dx_1 \cdots dx_n,$$

$$f(x_1, \ldots, x_n) = \partial_{x_1}\, \partial_{x_2} \cdots \partial_{x_n} F(x_1, \ldots, x_n).$$

The marginal density for a particular component can be found from the joint density by integrating out the other variables, e.g.,

$$f_{X_1}(x) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_{\boldsymbol{X}}(x, x_2, \ldots, x_n)\, dx_2\, dx_3\, \cdots\, dx_n.$$

There also is a notion of conditional densities. The definition is a natural analogue of conditional probability. Suppose Y is another random variable so that $(\boldsymbol{X}, Y)$ is a random vector in $\mathbb{R}^{n+1}$. Let us write its density as $f(\boldsymbol{x}, y)$ and

$$f_{\boldsymbol{X}}(\boldsymbol{x}) = \int_{-\infty}^{\infty} f(\boldsymbol{x}, y)\, dy$$

for the corresponding marginal.

Definition The conditional density for Y given $\boldsymbol{X} = \boldsymbol{x}$ is given by

$$f(y \mid \boldsymbol{x}) = \frac{f(\boldsymbol{x}, y)}{f_{\boldsymbol{X}}(\boldsymbol{x})},$$

provided that $f_{\boldsymbol{X}}(\boldsymbol{x}) \neq 0$. If $f_{\boldsymbol{X}}(\boldsymbol{x}) = 0$, then $f(y \mid \boldsymbol{x})$ is not defined. In other words,

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

Example Consider the two-dimensional random vector (X, Y) with a uniform distribution over the triangle

$$D = \{(x, y) : 0 \leq x \leq y \leq 1\}.$$

This triangle has area $1/2$ and hence the joint density is given by

$$f(x, y) = 2, \quad 0 \leq x \leq y \leq 1.$$

The marginal densities are

$$f_X(x) = \int_{-\infty}^{\infty} f(x, y)\, dy = 2\,(1 - x), \quad 0 \leq x \leq y,$$

$$f_Y(y) = \int_{-\infty}^{\infty} f(x)\, dx = 2\, y, \quad x \leq y \leq 1,$$

and the conditional densities are given by

$$f(x \mid y) = \frac{f(x,y)}{f_Y(y)} = \frac{1}{y}, \quad 0 \le x \le y,$$

$$f(y \mid x) = \frac{f(x,y)}{f_X(x)} = \frac{1}{1-x}, \quad x \le y \le 1.$$

In particular, the conditional densities have uniform distributions.

The notion of (mutual) independence for continuous random variables is essentially the same as for discrete but we can write the condition in terms of the joint distribution and joint density. A finite collection of continuous random variables $\boldsymbol{X} = (X_1, \ldots, X_n)$ is independent if and only if the joint distribution function and the joint density are given by the product of the marginal quantities,

$$F_{\boldsymbol{X}}(x_1, \ldots, x_n) = F_{X_1}(x_1)\, F_{X_2}(x_2) \cdots F_{X_n}(x_n),$$

$$f_{\boldsymbol{X}}(x_1, \ldots, x_n) = f_{X_1}(x_1)\, f_{X_2}(x_2) \cdots f_{X_n}(x_n).$$

Often we are given the densities of the individual random variables and the assumption that they are independent. Then we can write the joint density as the product of the individual densities.

Example Suppose X, Y are independent exponential random variables each with rate λ. Find the density of $Z = X + Y$.

The joint density of (X, Y) is given by the product of the two individual densities and hence is

$$f(x,y) = \lambda\, e^{-\lambda x}\, \lambda\, e^{-\lambda y} = \lambda^2\, e^{-\lambda(x+y)}, \qquad x, y > 0.$$

In order to find the density for Z we will find the distribution function first and then differentiate. If $z > 0$,

$$\begin{aligned} F_Z(z) &= \mathbb{P}\{Z \le z\} = \mathbb{P}\{X + Y \le z\} \\ &= \int\int_{x+y \le z} \lambda^2\, e^{-\lambda(x+y)} dy\, dx \\ &= \int_0^z \int_0^{z-x} \lambda^2\, e^{-\lambda(x+y)} dy\, dx \\ &= \int_0^z \lambda e^{-\lambda x} [1 - e^{-\lambda(z-x)}]\, dx \\ &= \int_0^z \lambda\, [e^{-\lambda x} - e^{-\lambda z}]\, dx \\ &= 1 - e^{-\lambda z} - \lambda\, z\, e^{-\lambda z}. \end{aligned}$$

$$f_Z(z) = F_Z'(z) = \lambda^2\, z\, e^{-\lambda z}, \qquad z > 0.$$

An infinite collection of random variables is (mutually) independent if each finite subset of them is independent. The basic properties of independent random variables extend to continuous random variables — they are important enough to repeat them here.

Fact

- If X, Y are independent random variables, then $\mathbb{E}[XY] = \mathbb{E}[X]\,\mathbb{E}[Y]$.
- If $X_1, \ldots, X_n$ are independent (or, more generally, uncorrelated), then

$$\mathrm{Var}[X_1 + \cdots + X_n] = \mathrm{Var}[X_1] + \cdots + \mathrm{Var}[X_n].$$

There are two other quantities associated to pairs of random variables.

Definition

- The covariance of X and Y is defined by

$$\mathrm{Cov}(X, Y) = \mathbb{E}\left[(X - \mathbb{E}[X])\,(Y - \mathbb{E}[Y])^2\right] = \mathbb{E}[XY] - \mathbb{E}[X]\,\mathbb{E}[Y].$$

- The correlation coefficient is defined by

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

  Note that $\mathrm{Cov}(X, Y) = 0$ if and only if X, Y are uncorrelated. The correlation coefficient satisfies $-1 \leq \rho(X, Y) \leq 1$. Roughly speaking, if $\mathrm{Cov}(X, Y) > 0$, or equivalently, $\rho(X, Y) > 0$, then when X is larger we expect Y to be larger. In this case we say that X, Y are positively correlated.

3 Moment generating function and sums of independent random variables

There are two closely related quantities associated to random variables that at first may not be motivated.

Definition Suppose X is a random variable.

- The moment generating function (mgf) of X is defined to be

$$M(s) = M_X(s) = \mathbb{E}[e^{sX}]$$

  provided that the expectation exists.

- The characteristic function (cf) of X is defined to be

$$\phi(s) = \phi_X(s) = \mathbb{E}[e^{isX}].$$

Here $i = \sqrt{-1}$. If x is a real number, then

$$e^{ix} = \cos x + i \sin x,$$

and

$$|e^{ix}|^2 = \cos^2 x + \sin^2 x = 1.$$

For this reason, we see that $|e^{isX}| \leq 1$ for all s and hence the expectation in the definition of the characteristic polynomial always exists. One advantage of characteristic function over moment generating function is that the expectation always exists. One disadvantage is that one has to deal with complex numbers. When M_X exists, then we have $\phi_X(s) = M_X(is)$.

We will give some examples later, but let us discuss some properties first for which we will not give complete proofs. Clearly,

$$M_X(0) = \phi_X(0) = 1.$$

Fact If two random variables have the same mgf or have the same cf, then they have the same distribution.

If we formally differentiate the formula with respect to s and interchange the derivative and the expectation we get

$$M_X^{(n)}(s) = \mathbb{E}\left[X^n e^{sX}\right], \qquad \phi_X^{(n)}(s) = \mathbb{E}\left[(iX)^n e^{isX}\right].$$

Here the (n) denotes the nth derivative. One needs to do some work to justify this, but it can be shown to be valid for the mgf if it exists. It is also valid for the cf provided that $\mathbb{E}[|X|^n] < \infty$. If we plug in $s = 0$ we get the following.

- If the moment generating function exists for s in a neighborhood of the origin, then for all positive integers n,

$$\mathbb{E}[X_n] = M^{(n)}(0).$$

- For all positive integers n, If $\mathbb{E}[|X|^n] < \infty$, then

$$i^n \, \mathbb{E}[X_n] = \phi^{(n)}(0).$$

The number $\mathbb{E}[X^n]$ is called the nth moment of the random variable and we see where the term mgf comes from.

If a, b are real numbers, we can write the mgf or cf of $aX + b$ in terms of the mgf or cf of X as follows:

$$M_{aX+b}(s) = \mathbb{E}[e^{s(aX+b)}] = e^{bs}\,\mathbb{E}[e^{(as)X}] = e^{bs}\,M_X(as),$$
$$\phi_{aX+b}(s) = \mathbb{E}[e^{is(aX+b)}] = e^{ibs}\,\mathbb{E}[e^{i(as)X}] = e^{ibs}\,\phi_X(as),$$

Finally, we get a nice formula for the mgf or cf of the sum of *independent* random variables.

Fact If $X_1, X_2, \ldots, X_n$ are independent random variables and $S = X_1 + \cdots + X_n$, then

$$M_S(s) = M_{X_1}(s)\,M_{X_2}(s)\cdots M_{X_n}(s), \qquad \phi_S(s) = \phi_{X_1}(s)\,\phi_{X_2}(s)\cdots\phi_{X_n}(s).$$

This follows from

$$M_S(s) = \mathbb{E}[e^{sX_1}e^{sX_2}\cdots e^{sX_n}] = \mathbb{E}[e^{sX_1}]\,\mathbb{E}[e^{sX_2}]\cdots\mathbb{E}[e^{sX_n}] = M_{X_1}(s)\,M_{X_2}(s)\,\cdots M_{X_n}(s),$$

and similarly for $\phi_S(s)$. The second equality uses the independence of the random variables.

Example

- If $X \sim N(0,1)$,

$$M(s) = \int_{-\infty}^{\infty} e^{sx}\,\frac{1}{\sqrt{2\pi}}\,e^{-x^2/2}\,dx = e^{s^2/2}, \quad \phi(s) = M(si) = e^{-s^2/2}.$$

- if $X \sim N(\mu, \sigma^2)$, then $X = \sigma\,Y + \mu$ where $Y \sim N(0,1)$ and

$$M(s) = e^{\mu s}\,M_Y(\sigma s) = e^{\mu s}\,e^{\sigma^2 s^2/2}, \quad \phi(s) = M(is) = e^{i\mu s}\,e^{-\sigma^2 s^2/2}.$$

- If X is Poisson with mean λ,

$$M(s) = \sum_{n=0}^{\infty} e^{sn}\,e^{-\lambda}\,\frac{\lambda^n}{n!} = e^{-\lambda}\sum_{n=0}^{\infty}\frac{(\lambda e^s)^n}{n!} = e^{-\lambda}\,e^{\lambda e^s} = \exp\left\{\lambda(e^s - 1)\right\},$$

$$\phi(s) = M(is) = \exp\left\{i\lambda(e^s - 1)\right\}.$$

- If X is exponential with rate λ,

$$M(s) = \int_0^\infty e^{sx}\,\lambda e^{-\lambda x}\,dx = \frac{\lambda}{\lambda - s}, \qquad \phi(s) = M(is) = \frac{\lambda}{\lambda - is}.$$

  $M(s)$ is valid only for $s < \lambda$ but $\phi(s)$ exists for all s.

We use this to establish some important results about sums of independent random variables.

Fact If $X_1, \ldots, X_n$ are independent random variables with $X_j \sim N(\mu_j, \sigma_j^2)$, then $X_1 + \cdots + X_n \sim N(\mu, \sigma^2)$, where $\mu = \mu_1 + \cdots + \mu_n, \sigma^2 = \sigma_1^2 + \cdots + \sigma_n^2$.

The fact that

$$\mathbb{E}[X_1 + \cdots + X_n] = \mu, \quad \mathrm{Var}[X_1 + \cdots + X_n] = \sigma^2,$$

follows from properties of expectations and variance of sums. The important fact is that the distribution of the sum is normal. To check this we just check that

$$\begin{aligned} M_{X_1+\cdots+X_n}(s) &= M_{X_1}(s)\,M_{X_2}(s)\cdots M_{X_n}(s) \\ &= \prod_{j=1}^n \exp\left\{\mu_j\,s + \frac{\sigma_j^2\,s^2}{2}\right\} \\ &= \exp\left\{\mu\,s + \frac{\sigma^2\,s^2}{2}\right\} \end{aligned}$$

But this is the mgf of a $N(\mu, \sigma^2)$ random variable, and the mgf determines the distribution.

Fact If $X_1, \ldots, X_n$ are independent random variables with X_j being Poisson with rate λ_j, then $X_1 + \cdots + X_n$ is Poisson with rate $\lambda_1 + \cdots + \lambda_n$.

To check this we just check that

$$\begin{aligned} M_{X_1+\cdots+X_n}(s) &= M_{X_1}(s)\,M_{X_2}(s)\cdots M_{X_n}(s) \\ &= \prod_{j=1}^n \exp\left\{\lambda_j\,(e^s - 1)\right\} \\ &= \exp\left\{(\lambda_1 + \cdots + \lambda_n)\,(e^s - 1)\right\} \end{aligned}$$

which is the mgf of a Poisson random variable with rate $\lambda_1 + \cdots + \lambda_n$.

4 Sums of independent random variables

In this section we will assume that $X_1, X_2, \ldots$ are independent, identically distributed (i.i.d.) random variables with $\mathbb{E}[X_j] = \mu, \mathrm{Var}[X_j] = \sigma^2 < \infty$ and let

$$S_n = X_1 + X_2 + \cdots + X_n.$$

We know that

$$\mathbb{E}[S_n] = \mu\, n, \qquad \mathrm{Var}[S_n] = \sigma^2\, n, \quad \mathrm{SD}[S_n] = \sigma\sqrt{n}.$$

This simple calculation underpins one of the most important facts about independent random trials.

Fact The "typical error" (standard deviation) of the sum of n i.i.d. quantities grows like $\sqrt{n}$.

We will start by giving a couple of simple inequalities that relate expectation and variance to probabilities of being far from the mean.

Fact Suppose $a > 0$.

- Markov inequality:

$$\mathbb{P}\{|X| \geq a\} \leq \frac{\mathbb{E}[|X|]}{a},$$

- Chebyshev inequality:

$$\mathbb{P}\{|X - \mathbb{E}(X)| \geq a\} \leq \frac{\mathrm{Var}[X]}{a^2}.$$

The first follows from

$$\mathbb{E}[|X|] \geq \mathbb{E}[|X|\, 1\{|X| \geq a\}] \geq a\, \mathbb{P}\{|X| \geq a\},$$

and the second follow from applying the first to the random variable $|X - \mathbb{E}(X)|^2$.

The law of large numbers (LLN) concerns the average of n trials,

$$\overline{X}_n = \frac{S_n}{n}.$$

Note that

$$\mathbb{E}\left[\overline{X}_n\right] = \frac{\mathbb{E}[S_n]}{n} = \mu, \qquad \mathrm{Var}\left[\overline{X}_n\right] = \frac{\mathrm{Var}[S_n]}{n^2} = \frac{\sigma^2}{n}, \qquad \mathrm{SD}\left[\overline{X}_n\right] = \sqrt{\mathrm{Var}[\overline{X}_n]} = \frac{\sigma}{\sqrt{n}}.$$

Note that for every $\epsilon > 0$, Chebyshev's inequality gives,

$$\mathbb{P}\{|\overline{X}_n - \mu| \geq \epsilon\} \leq \frac{\mathrm{Var}[\overline{X}_n]}{\epsilon^2} = \frac{\sigma^2}{\epsilon^2\, n}$$

which goes to zero as $n \to \infty$.

We have demonstrated what is often called the weak law of large numbers (WLLN). There is a stronger statement called the strong law of large numbers (SLLN) which states that with probability one $\overline{X}_n \to \mu$. To make this precise we need measure theory so we will not do it here. One can show that the SLLN and WLLN hold if the expectation exists even if $\sigma^2 = \infty$.

The standardized sum associated to S_n is the one shifted and normalized so that it have mean zero and variance 1,

$$Z_n = \frac{S_n - \mu\, n}{\sigma\, \sqrt{n}}, \qquad \mathbb{E}[Z_n] = 0, \qquad \mathrm{Var}[Z_n] = 1.$$

Probably the most important theorem in probability is the central limit theorem which states that as $n \to \infty$, the distribution of Z_n approaches that of a standard normal.

Fact Central Limit Theorem If $a < b$, then

$$\lim_{n\to\infty} \mathbb{P}\{a \le Z_n \le b\} = \Phi(b) - \Phi(a) = \frac{1}{\sqrt{2\pi}} \int_a^b e^{-\frac{x^2}{2}}\, dx.$$

We will not prove this, but we will discuss part of the argument so that the reader can see where the normal distribution arises. Without loss of generality we will assume that $\mu = 0$ and $\sigma^2 = 1$ for otherwise we can consider $Y_j = (X_j - \mu)/\sigma$. In this case $Z_n = S_n/\sqrt{n}$. Let ϕ be the characteristic function for X_j; we know that $\phi(0) = 1, \phi'(0) = i\,\mathbb{E}[X_j] = 0$ and $\phi''(0) = -\,\mathbb{E}[X_j^2] = -1$. Let us expand ϕ about the origin

$$\phi(s) = 1 - \frac{s^2}{2} + o(s^2), \qquad s \to 0.$$

The characteristic function for Z_n is

$$\phi_{Z_n}(s) = \phi_{S_n}\left(\frac{s}{\sqrt{n}}\right) = \left[\phi\left(\frac{s}{\sqrt{n}}\right)\right]^n = \left[1 - \frac{s^2}{2n} + o(n^{-1})\right]^n.$$

As $n \to \infty$, the right-hand side approaches $e^{-s^2/2}$ which is the characteristic function for a $N(0,1)$ random variable.

---

5 Some other Distributions

5.1 Gamma distribution

Definition A random variable X has a Gamma distribution with shape parameter r and rate λ if it has density function

$$f(x) = C_{\lambda,r}\, x^{r-1}\, e^{-\lambda x}, \qquad x > 0,$$

where $C_{\lambda,r}$ is the constant so that f integrates to one. In fact,

$$C_{\lambda,r} = \frac{\lambda^r}{\Gamma(r)},$$

where $\Gamma(r)$ denotes the Gamma function

$$\Gamma(r) = \int_0^\infty x^{r-1}\, e^{-x}\, dx.$$

If $r = 1$, the Gamma distribution is the same as the exponential with rate λ.

The Gamma function is the generalization of the factorial function to all positive numbers. Note that $r > 0$ is needed for the integral to converge near 0. Using integration by parts one can prove the rule

$$\Gamma(r+1) = r\,\Gamma(r).$$

If r is an integer, then $\Gamma(r) = (r-1)!$.

It has

$$\mathbb{E}[X] = \frac{r}{\lambda}, \qquad \mathrm{Var}[X] = \frac{r}{\lambda^2},$$

and moment generating function

$$M(s) = \left[\frac{\lambda}{\lambda - s}\right]^r, \qquad s < \lambda.$$

Using this we get the following property. Note that X and Y have the same λ in the statement.

Fact If X and Y are independent random variables and X is Gamma with rate λ and shape parameter r_1 and Y is Gamma with rate parameter λ and shape parameter r_2, then $X + Y$ is Gamma with rate λ and shape parameter $r_1 + r_2$.

The following can be checked.

Fact Iy X has a Gamma distribution with shape parameter r and rate λ and $a > 0$, then aX has a Gamma distribution with shape parameter r and rate λ/a.

If we consider a Gamma distribution as measuring times, then we can view multiplication by a constant as changing the time unit.

5.2 Chi-square distribution

Definition A random variable X has a chi-square distribution with n degrees of freedom, written χ^2_n, if it has the distribution of

$$Z_1^2 + Z_2^2 + \cdots + Z_n^2$$

where $Z_1, Z_2, \ldots, Z_n$ are independent standard normal random variables.

To find the density of Z_1^2 we compute the distribution function F first.

$$\begin{aligned}
\mathbb{P}\{Z_1^2 \le t\} &= \mathbb{P}\{-\sqrt{t} \le Z_1 \le \sqrt{t}\} \\
&= \int_{-\sqrt{t}}^{\sqrt{t}} \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\, dx \\
&= 2\int_{0}^{\sqrt{t}} \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\, dx
\end{aligned}$$

We cannot find a closed form for this integral. But we can still differentiate it using the fundamental theorem of calculus and the chain rule.

$$f(t) = F'(t) = 2 \cdot \frac{1}{\sqrt{2\pi}} e^{-(\sqrt{t})^2/2} \cdot \frac{1}{2\sqrt{t}} = \frac{1}{\sqrt{2\pi}} t^{-1/2} e^{-t/2}.$$

From this we see that that the χ^2_1 distribution is the same as the Gamma distribution with shape parameter $r = 1/2$ and rate parameter $\lambda = 1/2$. Using the rule for sums of independent Gamma random variables we get the following general rule.

Fact The χ^2_n distribution is the same as the Gamma distribution with shape parameter $n/2$ and rate $1/2$.

In particular. if $X \sim \chi^2_n$,

$$\mathbb{E}[X] = n, \qquad \mathrm{Var}[X] = 2n.$$

5.3 Beta distribution

The beta distribution arises in Bayesian statistics from the following problem. Suppose we have a coin with probability p of coming up heads but we do not know p. Since we know nothing we assume a *prior distribution* of a uniform distribution on $[0, 1]$. Suppose we flip the coin $m + n - 2$ times and observe $m - 1$ heads and $n - 1$ tails. What is our *posterior density* for p? This can be worked out using a form of Bayes theorem. Let A be the event of $m - 1$ heads and $n - 1$ tails and write the conditional density on p given A as

$$f(p \mid A) = \frac{f(p)\,\mathbb{P}(A \mid p)}{\mathbb{P}(A)}.$$

Note that the denominator does not depend on p, $f(p) = 1$, and given p the probability that A occurs is $\binom{m+n}{m} p^m (1-p)^n$. Hence

$$f(p \mid A) = \frac{1}{\beta(m,n)} p^{m-1} (1-p)^{n-1}$$

where $\beta(m,n)$ is the constant needed to make this a probability density. This is the basis for this definition.

Definition A random variable X has a beta distribution with parameters m and n if it has density

$$\frac{x^{m-1}(1-x)^{n-1}}{\beta(m,n)}, \qquad 0 < x < 1$$

where

$$\beta(m,n) = \int_0^1 x^{m-1}(1-x)^{n-1}\,dx = \frac{\Gamma(m)\,\Gamma(n)}{\Gamma(m+n)} = \frac{(m-1)!\,(n-1)!}{(m+n-1)!}.$$

Fact The beta distribution with parameters m and n is the posterior distribution for p when there are Bernoulli trials with probability of success p, the prior distribution on p is uniform on $[0,1]$, and $m+n-2$ trials have been done with exactly $m-1$ successes.

The computation of $\beta(m,n)$, often called the beta function, is straightforward. Integration by parts gives

$$\beta(m,n) = \int_0^1 x^{m-1}(1-x)^{n-1}\,dx = \frac{n-1}{m}\int_0^1 x^m (1-x)^{n-2},$$

which gives the rule

$$\beta(m,n) = \frac{n-1}{m}\beta(m+1,n-1).$$

Also $\beta(1,n) = 1/n$.

5.4 Multivariate normal distribution

Note: this section is cut and paste from *Introduction to Stochastic Calculus with Applications.*

Although the normal or Gaussian distribution is a little inconvenient in the sense that the distribution function cannot be computed exactly, there are many other aspects that make the distribution very convenient. In particular, when dealing with many variables, assuming a *joint or multivariate* normal distribution makes computations tractable. In this section we will give the basic definitions. Roughly speaking, the basic assumption is that

if $(X_1, \ldots, X_n, Y)$ have a joint normal distribution then not only does each variable have a normal distribution but also, the conditional distribution of Y given $X_1, \ldots, X_n$ is normal with mean $E[Y|X_1, \ldots, X_n]$ and a variance that depends on the joint distribution but not on the observed data points. There are a number of equivalent ways to define a joint normal distribution. We will use the following.

Definition A finite sequence of random variables $(X_1, \ldots, X_n)$ has a *joint (or multivariate) normal (or Gaussian)* distribution if they are linear combinations of independent standard normal random variables. In other words, if there exist independent random variables $(Z_1, \ldots, Z_m)$, each $N(0, 1)$, and constants m_j, a_{jk} such that for $j = 1, \ldots, n$,

$$X_j = m_j + a_{j1}\, Z_1 + a_{j2}\, Z_2 + \cdots + a_{jm}\, Z_m.$$

Clearly $\mathbb{E}[X_j] = m_j$. Let us consider the case of mean-zero (also called *centered*) joint normals, in which case the equation above can be written in matrix form

$$\boldsymbol{X} = A\, \mathbf{Z},$$

where

$$\boldsymbol{X} = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{pmatrix}, \quad \mathbf{Z} = \begin{pmatrix} Z_1 \\ Z_2 \\ \vdots \\ Z_m \end{pmatrix},$$

and A is the $n \times m$ matrix with entries a_{jk}. Each X_j is a normal random variable with mean zero and variance

$$\mathbb{E}[X_j^2] = a_{j1}^2 + \cdots + a_{jm}^2.$$

More generally, the covariance of X_j and X_k is given by

$$\mathrm{Cov}(X_j, X_k) = \mathbb{E}[X_j X_k] = \sum_{l=1}^{m} a_{jl} a_{kl}.$$

Let $\Gamma = AA^T$ be the $n \times n$ matrix whose entries are

$$\Gamma_{jk} = \mathbb{E}[X_j X_k].$$

Then Γ is called the *covariance matrix*.

We list some important properties. Assume $(X_1, \ldots, X_n)$ has a joint normal distribution with mean zero and covariance matrix Γ.

- Each X_j has a normal distribution. In fact, if $b_1, \ldots, b_n$ are constants, then

  $$b_1\, X_1 + \cdots + b_n\, X_n,$$

  has a normal distribution. We can see this easily since we can write the sum above as a linear combination of the independent normals $Z_1, \ldots, Z_m$.

- The matrix Γ is symmetric, $\Gamma_{jk} = \Gamma_{kj}$. Moreover, it is *positive semi-definite* which means that if $\mathbf{b} = (b_1, \ldots, b_n)$ is a vector in $\mathbb{R}^n$, then

$$\mathbf{b} \cdot \Gamma \mathbf{b} = \sum_{j=1}^{n} \sum_{k=1}^{n} \Gamma_{jk} b_j b_k \geq 0. \tag{5}$$

  (If the $\geq$ is replaced with > 0 for all $\mathbf{b} = (b_1, \ldots, b_n) \neq (0, \ldots, 0)$, then the matrix is called *positive definite*.) The inequality (5) can be derived by noting that the left-hand side is the same as

$$\mathbb{E}\left[(b_1 X_1 + \cdots + b_n X_n)^2\right],$$

  which is clearly nonnegative.

- If Γ is a positive semidefinite, symmetric matrix, then it is the covariance matrix for a joint normal distribution. The proof of this fact, which we omit, uses linear algebra to deduce that there exists an $n \times n$ matrix A with $AA^T = \Gamma$. (The A is not unique.)

- The distribution of a mean-zero joint normal is determined by the covariance matrix Γ.

In order to show that the covariance matrix Γ determines the distribution of a mean-zero joint normal, we compute the characteristic function. Suppose that $\Gamma = AA^T$ where A is $n \times n$. Using the independence of $Z_1, \ldots, Z_n$ and the characteristic function of the standard normal, $\mathbb{E}[e^{itZ_k}] = e^{-t^2/2}$, we see that the characteristic function of

$(X_1, \ldots, X_n)$ is

$$\begin{aligned}
\phi(\theta_1, \ldots, \theta_n) &= \mathbb{E}\left[\exp\left\{i(\theta_1 X_1 + \ldots + \theta_n X_n)\right\}\right] \\
&= \mathbb{E}\left[\exp\left\{i \sum_{j=1}^n \theta_j \sum_{k=1}^n a_{jk} Z_k\right\}\right] \\
&= \mathbb{E}\left[\exp\left\{i \sum_{k=1}^n Z_k (\sum_{j=1}^n \theta_j a_{jk})\right\}\right] \\
&= \prod_{k=1}^n \mathbb{E}\left[\exp\left\{i Z_k (\sum_{j=1}^n \theta_j a_{jk})\right\}\right] \\
&= \exp\left\{-\frac{1}{2} \sum_{k=1}^n \left(\sum_{j=1}^n \theta_j a_{jk}\right)^2\right\} \\
&= \exp\left\{-\frac{1}{2} \sum_{k=1}^n \sum_{j=1}^n \sum_{l=1}^n \theta_j \theta_i a_{jk} a_{lk}\right\} \\
&= \exp\left\{-\frac{1}{2} \theta A A^T \theta^T\right\} \\
&= \exp\left\{-\frac{1}{2} \theta \Gamma \theta^T\right\}
\end{aligned}$$

where we write $\theta = (\theta_1, \ldots, \theta_n)$. Even though we used A, which is not unique, in our computation, the answer only involves Γ. Since the characteristic function determines the distribution, the distribution depends only on the covariance matrix.

- If Γ is invertible, then $(X_1, \ldots, X_n)$ has a density. We write it in the case that the random variables have mean $\mathbf{m} = (m_1, \ldots, m_n)$,

$$f(x_1, \ldots, x_n) = f(\mathbf{x}) =$$
$$\frac{1}{(2\pi)^{n/2} \sqrt{\det \Gamma}} \exp\left\{-\frac{(\mathbf{x} - \mathbf{m}) \cdot \Gamma^{-1} (\mathbf{x} - \mathbf{m})^T}{2}\right\}.$$

  Sometimes this density is used as a definition of a joint normal. The formula for the density looks messy, but note that if $n = 1$, $\mathbf{m} = m, \Gamma = [\sigma^2]$, then the right-hand side is the density of a $N(m, \sigma^2)$ random variable.

- If (X_1, X_2) have a mean-zero joint normal density, and $\mathbb{E}(X_1 X_2) = 0$, then X_1, X_2 are independent random variables. To see this let $\sigma_j^2 = \mathbb{E}[X_j^2]$. Then the covariance matrix of (X_1, X_2) is the diagonal matrix with diagonal entries σ_j^2. If (Z_1, Z_2) are independent $N(0, 1)$ random variables and $Y_1 = \sigma_1 Z_1, Y_2 = \sigma_2 Z_2$, then by our definition (Y_1, Y_2) are joint normal with the same covariance matrix. Since the covariance matrix determines the distribution, X_1, X_2 must be independent,

It is a special property about joint normal random variables that uncorrelated implies independent. In our construction of Brownian motion, we will use a particular case, that we state as a lemma.

Proposition 5.1. *Suppose X, Y are independent $N(0,1)$ random variables and*

$$Z = \frac{X+Y}{\sqrt{2}}, \qquad W = \frac{X-Y}{\sqrt{2}}.$$

Then Z and W are independent $N(0,1)$ random variables.

Proof. By definition (Z, W) has a joint normal distribution and Z, W clearly have mean 0. Using $\mathbb{E}[X^2] = \mathbb{E}[Y^2] = 1$ and $\mathbb{E}[XY] = 0$, we get

$$\mathbb{E}[Z^2] = 1, \quad \mathbb{E}[W^2] = 1, \quad \mathbb{E}[ZW] = 0.$$

Hence the covariance matrix for (Z, W) is the identity matrix and this is the covariance matrix for independent $N(0,1)$ random variables. □

5.5 Distributions from statistics: F and (student's) T

We will discuss two distributions that arise in the statistical analysis of normal random variables. For this section, we will assume that $X_1, X_2, \ldots$ are independent random variables each with a $N(\mu, \sigma^2)$ distribution but where at least one of μ, σ^2 is unknown. We let

$$\overline{X}_n = \frac{X_1 + \cdots + X_n}{n}$$

which is called the sample mean. We have already seen that $\overline{X}_n \sim N(\mu, \sigma^2/n)$, or equivalently,

$$\frac{\sqrt{n}\,(\overline{X}_n - \mu)}{\sigma} \sim N(0,1). \tag{6}$$

The best statistical estimator of μ is $\overline{X}_n$ and if we know σ (which is a big assumption!), we can estimate the probability of a large error in this estimator by

$$\mathbb{P}\{\left|\overline{X}_n - \mu\right| \geq a\} = \mathbb{P}\left\{\frac{\sqrt{n}\,|\overline{X}_n - \mu|}{\sigma} \geq \frac{a\sqrt{n}}{\sigma}\right\} = 2\,\Phi\left(-\frac{a\sqrt{n}}{\sigma}\right).$$

Similarly, if we knew μ and wanted to estimate σ we could use the estimator

$$\overline{V}_n = \frac{1}{n}\sum_{j=1}^{n}(X_j - \mu)^2 = \frac{\sigma^2}{n}\sum_{j=1}^{n}\left(\frac{X_j - \mu}{\sigma}\right)^2.$$

The summation on the right is the sum of the squares of independent $N(0,1)$ random variables. Therefore, we get

$$\frac{n\,\overline{V}_n}{\sigma} \sim \chi^2_n.$$

The more realistic situation is when both μ and σ^2 are unknown. In this case, we still use $\overline{X}_n$ as the estimator for μ but we replace the estimator for σ^2 or σ by the sample variance or (sample) standard deviation defined by

$$S_n^2 = \frac{1}{n-1}\sum_{j=1}^{n}(X_j - \bar{X}_n)^2, \qquad S_n = \sqrt{\frac{1}{n-1}\sum_{j=1}^{n}(X_j - \bar{X}_n)^2}.$$

Note that $\overline{X}_n$ and S_n are statistics of the data, that is, they are functions of the data points. One can compute these numbers without regard to the distribution of the data. However, if we assume that the data is coming from i.i.d. normal random variables we can give the distribution on these statistics.

Note that we are dividing by $1/(n-1)$ rather then $1/n$ in the definition of S_n^2. Roughly speaking this is because if we estimate the mean by the sample mean, then the data which produced the sample mean will tend to be a bit closer to the sample mean than the actual mean. As an extreme, note that if we have only one piece of data ($n = 1$), then we have no estimate for variance since our one data point agrees with the sample mean. To show that $n-1$ is the correct denominator, let us show that $\mathbb{E}[S_n^2] = \sigma^2$ assuming $n \geq 2$. For ease assume that $\mu = 0$ in which case $\sigma^2 = \mathbb{E}[X_j^2] = n\,\mathbb{E}[\overline{X}_n^2]$. Then,

$$\begin{aligned}\mathbb{E}\left[\sum_{j=1}^{n}(X_j - \bar{X}_n)^2\right] &= \sum_{j=1}^{n}\left(\mathbb{E}[X_j^2] - 2\mathbb{E}[X_j\,\bar{X}_n] + \mathbb{E}[\bar{X}_n^2]\right)\\ &= \sum_{j=1}^{n}\frac{n-1}{n}\,\sigma^2 = (n-1)\,\sigma^2.\end{aligned}$$

Here we use

$$\mathbb{E}[X_j^2] = \sigma^2, \quad \mathbb{E}[\bar{X}_n^2] = \mathrm{Var}[X_n] = \frac{\sigma^2}{n},$$

$$\mathbb{E}[X_j\,\overline{X}_n] = \frac{1}{n}\left[\mathbb{E}[X_j^2] + \sum_{k\neq j}\mathbb{E}[X_jX_k]\right] = \frac{1}{n}\,\mathbb{E}[X_j^2] = \frac{\sigma^2}{n}.$$

The next fact is surprising — it uses our assumption of i.i.d. normal random variables.

Fact

- $\bar{X}_n$ and S_n^2 are independent random variables.
- The distribution of $(n-1)\,S_n^2$ is χ^2_{n-1}.

Here we show how to prove this fact. Assume that $\mu = 0, \sigma^2 = 1$ for otherwise we could consider $(X_j - \mu)/\sigma$.

Let $Y_j = X_j - \bar{X}_n$ and $\mathbf{Y} = (Y_1, \ldots, Y_{n-1})$. $\mathbf{Y}$ has a centered multivariate normal distribution and we denote its covariance matrix by Γ. Note also that $(\mathbf{Y}, \bar{X}_n)$ has a multivariate normal distribution with $\mathbb{E}[Y_j \bar{X}_n] = 0, \mathbb{E}[\bar{X}_n^2] = 1/n$, and hence the covariance matrix of $(\mathbf{Y}, \bar{X}_n)$ has a block form

$$\begin{bmatrix} \Gamma & 0 \\ 0 & 1/n \end{bmatrix}.$$

This is the same matrix that one obtains by choosing $\bar{X}_n$ to be a centered normal independent of $\mathbf{Y}$ and by uniqueness of distribution, we see that $\bar{X}_n$ is independent of $Y_1, \ldots, Y_{n-1}$. Note that

$$Y_n = -(Y_1 + \cdots + Y_{n-1})$$

and hence

$$S_n^2 = \frac{1}{n-1} \sum_{j=1}^{n} Y_j^2$$

is a function of $Y_1, \ldots, Y_{n-1}$ and is also independent of $\overline{X}_n$.

To see the distribution of $(n-1)\, S_n^2 = \sum (X_j - \overline{X}_n)^2$ we first check that

$$\sum_{j=1}^{n} X_j^2 = \sum_{j=1}^{n} (X_j - \overline{X}_n)^2 + n\, \overline{X}_n^2 = (n-1)\, S_n^2 + \left(\frac{\overline{X}_n}{1/\sqrt{n}} \right)^2.$$

The left-hand side has a χ_n^2 distribution and the second term on the right has a χ_1^2 distribution. Since the two terms on the right are independent it will follow that the first term has a χ_{n-1}^2 distribution.

Definition If m, n are positive integers, the F-distribution with m and n degrees of freedom is the distribution of

$$\frac{Y/m}{Z/n}$$

where Y, Z are independent, $Y \sim \chi_m^2, Z \sim \chi_n^2$. It has density

$$c_{m,n}\, x^{\frac{m}{2}-1} \left(1 + \frac{m}{n}\, x\right)^{-\frac{m+n}{2}}$$

where $c_{m,n}$ is the constant so that it integrates to one.

The constant $c_{m,n}$ can be given explicitly but we will not write it here. We think of the definition as the ratio of the chi-square random variables, and the computation of the density is an exercise in multivariate calculus. The density is not something to be memorized!

The student t-distribution is what one gets when one replaces the actual standard deviation in (6) with the standard deviation.

Definition The student t-distribution with $n-1$ degrees of freedom is the distribution of

$$\frac{\overline{X}_n - \mu}{S_n/\sqrt{n}}.$$

This can also be defined as the distribution of

$$\frac{Z/\sqrt{n}}{Y/\sqrt{n}}$$

where Y, Z are independent, $Z \sim N(0,1)$, and $Y \sim \chi^2_{n-1}$. The density is given by

$$c_n \left(1 + \frac{x^2}{n}\right)^{-\frac{n+1}{2}},$$

where c_n is the constant so that it integrates to one.

As $n \to \infty$, the t-density approaches the standard normal density. We can see this as

$$\lim_{n\to\infty} \left(1 + \frac{x^2}{n}\right)^{-\frac{n+1}{2}} = e^{-x^2/2}.$$

From a practical perspective, if one is summing a large amount of data (even thirty data points makes a good approximation), one can use the standard normal rather than the t-distribution.