UC Irvine Math 130B Spring 2024

Probability II

Professor: Luke Smith
Teaching Assistant: Yiyun He
Notes: Timothy Cho

June 2024
Lecture Note Series #11

Introduction

These notes come from both the lecture and the discussion, and are roughly sorted by content. Sections are numbered chronologically using the following scheme by taking the section number modulo 10. Note that we have occasionally merged two sections for continuity reasons.

Day	Lecture	Discussion
Monday	0	1
Tuesday	2	3
Wednesday	4	5
Thursday	6	7
Friday	8	9

Additionally, the first digit (first two if the section number is three digits long) denotes the week that the lecture/discussion occurred in. It should be noted that not every lecture is recorded in these notes: some lectures were skipped, but despite this the notes should be comprehensible.

The text used was *A First Course in Probability*, 8e, by Sheldon Ross. Numbers in [brackets] refer to sections in the text.

11 Review of Math 130A

Recall the following definition.

Definition 11.1. A *random variable* (abbreviated RV) is a function $X : S \to \mathbb{R}$. The set S is called the *sample space* of X.

We can discuss probability that a random variable X takes on a particular subset $A \subseteq \mathbb{R}$, denoted $\Pr(X \in A)$. Random variables, on their own, are not particularly useful — they are merely representations of the sample space in $\mathbb{R}$. Hence, we consider their *distributions*, *expectation*, and *variance*.

Definition 11.2. Let X be a random variable. The *variance* of X is given by $\operatorname{Var} X := \mathbb{E}[(X - \mathbb{E}(X))^2]$.

Intuitively, the variance measures the mean squared error, but we will almost never compute it as such, due to the proposition below.

Proposition 11.3. *Let X be a random variable. Then $\operatorname{Var} X = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$, where $X^2 = X \cdot X$ is a multiplication of real-valued functions.*

Proof. Recall that the expectation $\mathbb{E}[X]$ is linear. Thus

$$\begin{aligned}\operatorname{Var} X = \mathbb{E}[(X - \mathbb{E}(X))^2] &= \mathbb{E}[(X - \mathbb{E}(X))(X - \mathbb{E}(X))] \\ &= \mathbb{E}[X^2 - 2[\mathbb{E}(X)]X + (\mathbb{E}[X])^2] \\ &= \mathbb{E}[X^2] - 2\mathbb{E}[X] \cdot \mathbb{E}[X] + \mathbb{E}[x]^2 \\ &= \mathbb{E}[X^2] - \mathbb{E}[X]^2,\end{aligned}$$

noting that $\mathbb{E}[X]$ is a constant which can be manipulated as such. □

We recall the following common types of random variables, with their expectations and variance.

Definition 11.4. Let X denote a random variable, and $p \in [0,1]$.

1. We say $X : S \to \{0,1\}$ is a *Bernoulli RV* and write $X \sim \mathsf{Ber}(p)$ is we have $\Pr(X = 0) = p$ and $\Pr(X = 1) = 1 - p$.
2. We say $X : S \to \{0,1,\dots,n\}$ is a *binomial RV* and write $X \sim \mathsf{Bin}(n,p)$ if $\Pr(X = i) = \binom{n}{i} p^i (1-p)^{n-i}$ for all $i \leq n$.
3. We say $X : S \to \mathbb{Z}_{\geq 0}$ is a *Poisson RV* and write $X \sim \mathsf{Poi}(\lambda)$ if
$$\Pr(X = i) = e^{-\lambda} \cdot \frac{\lambda^i}{i!}$$
for all $i \in \mathbb{Z}^+$.
4. We say $X : S \to \mathbb{Z}^+$ is a *geometric RV* and write $X \sim \mathsf{Geo}(p)$ if $\Pr(X = i) = (1-p)^{i-1}p$ for all $i \in \mathbb{Z}^+$.

Theorem 11.5. *Let X denote a random variable.*

1. *If $X \sim \mathsf{Ber}(p)$, then $\mathbb{E}[X] = p$ and $\operatorname{Var} X = p(1-p)$.*
2. *If $X \sim \mathsf{Bin}(n,p)$, then $\mathbb{E}[X] = np$ and $\operatorname{Var} X = np(1-p)$.*
3. *If $X \sim \mathsf{Poi}(\lambda)$, then $\mathbb{E}[X] = \operatorname{Var} X = \lambda$.*
4. *If $X \sim \mathsf{Geo}(n,p)$, then $\mathbb{E}[X] = 1/p$ and $\operatorname{Var} X = (1-p)/p^2$.*

14 Joint Random Variables (I)

In Math 130A, we only considered problems with a single random variable, so now we will generalize to problems with multiple random variables, with varying degrees in relation between each other.

Definition 14.1. Let X and Y be two discrete random variables. We define the *joint probability mass function* by $p_{X,Y}(x,y) := \Pr(X = x \text{ and } Y = y)$.

Example 14.2. Let an urn consist of 3 green, 7 yellow, and 2 purple balls. If X is the number of green balls chosen, and Y is the number of yellow balls, both in four draws without replacement, then
$$p_{X,Y}(1,3) = \Pr(X = 1 \text{ and } Y = 3) = \frac{\binom{3}{1}\binom{7}{3}\binom{2}{0}}{\binom{12}{4}} = \boxed{\frac{7}{33}}.$$
Similarly, we may compute
$$p_{X,Y}(0,2) = \Pr(X = 0, Y = 2) = \frac{\binom{3}{0}\binom{7}{2}\binom{2}{2}}{\binom{12}{4}} = \boxed{\frac{7}{165}}.$$

In the above example, we note that by the law of total probability, we have for all $i \leq 4$
$$\sum_{j=0}^{4} p_{X,Y}(X = i, Y = j) = \Pr(X = i, Y \leq 4) = \Pr(X = i) = p_X(i).$$

A similar identity holds for $p_Y(i)$, so that we are able to pull apart "single variable" information from a joint mass function.

Definition 14.3. Let X and Y be discrete random variables, with the joint probability mass function $p(i,j)$. Then the *marginal probability mass functions* for X resp. Y, denoted p_X resp. p_Y, are given by

$$\begin{aligned} p_X(i) &= \Pr(X=i) = \sum_j p(i,j) \text{ and} \\ p_Y(i) &= \Pr(Y=j) = \sum_i p(i,j). \end{aligned}$$

Example 14.4. Consider the number of children per familiy in the data given below:

Number of Children	0	1	2	3
Percentage of Families	15%	20%	35%	30%

Suppose further that boys and girls are equally distributed. Picking a family at random, let G denote the number of girls in the family, and B the number of boys. Letting $p(i,j) := p_{B,G}(i,j)$, we may calculate

$$\begin{aligned} p(0,0) &= \Pr(\text{no children}) = 0.15 = \boxed{\frac{3}{20}}, \\ p(0,1) &= \Pr(1 \text{ child, } 1 \text{ girl}) = \frac{1}{2}\cdot 0.2 = \boxed{\frac{1}{10}}, \\ p(0,2) &= \Pr(2 \text{ children, } 2 \text{ girls}) \\ &= \Pr(2 \text{ girls} \,|\, 2 \text{ children}) \cdot \Pr(2 \text{ children}) = \frac{1}{4}\cdot 0.35 = \boxed{\frac{7}{80}}, \\ p(0,3) &= \Pr(3 \text{ children, } 3 \text{ girls}) = \frac{1}{8}\cdot 0.3 = \boxed{\frac{3}{80}}. \end{aligned}$$

From here, we compute the marginal

$$p_B(0) = \Pr(B=0) = \frac{3}{20} + \frac{1}{10} + \frac{7}{80} + \frac{3}{80} = \boxed{\frac{3}{8}}.$$

Now, we define distribution functions.

Definition 14.5. Let X and Y be (any) random variables. We define the *joint cumulative distribution function* by $F_{X,Y}(a,b) := \Pr(X \le a, Y \le b)$.

Similarly, we define the *marginal cumulative distribution functions* by

$$\begin{aligned} F_X(a) &:= \Pr(X \le a) = \Pr(X \le a, Y \in \mathbb{R}) = \lim_{b\to\infty} F(a,b); \\ F_Y(b) &:= \Pr(Y \le b) = \Pr(X \in \mathbb{R}, Y \le b) = \lim_{a\to\infty} F(a,b). \end{aligned}$$

However, the limiting behavior is not always needed to solve problems.

Example 14.6. Suppose we roll 10 dice. Let X be the number of threes, and Y be the number of odd rolls. Then

$$F_X(5) = \Pr(X \leq 5) = \sum_{n=0}^{5} \binom{10}{n} \left(\frac{1}{6}\right)^n \left(\frac{5}{6}\right)^{10-n} = \boxed{\frac{10053125}{10077696}}.$$

We observe that we did not need $F_{X,Y}$ to compute F_X.

18 Joint Random Variables (II)

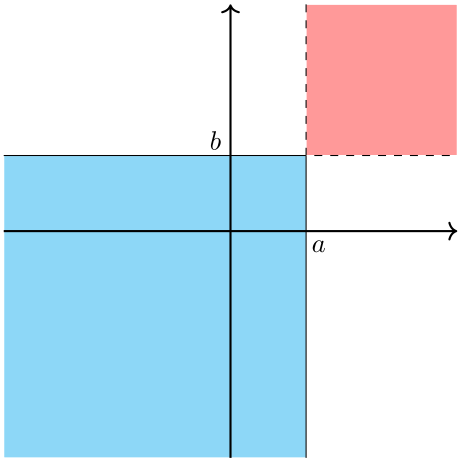
Example 18.1. Let X, Y be as defined in Example 14.6 (above). We can compute

$$F_{X,Y}(1,4) = \Pr(X \leq 1, Y \leq 4) = \sum_{x=0}^{1} \sum_{y=0}^{4} \binom{10}{x, y-x, 10-y} \left(\frac{1}{6}\right)^x \left(\frac{1}{3}\right)^{y-x} \left(\frac{1}{2}\right)^{10-y}.$$

Now, suppose we wanted to find $\Pr(X > 1, Y > 4)$. We know that

$$\Pr(X > a) = 1 - \Pr(X \leq a) = 1 - F_X(a),$$

but it is not the case that $\Pr(X > a, Y > b) \overset{?}{=} 1 - \Pr(X \leq a, Y \leq b) = 1 - F_{X,Y}(a,b)$. Geometrically, we are stating that the complement of $(-\infty, a] \times (-\infty, b] \subseteq \mathbb{R}^2$ is not $(a, \infty) \times (b, \infty)$, as seen in the figure below:



Instead, the correct identity is

$$\Pr(X > a, Y > b) = 1 - \Pr(X \leq a, Y \leq b) - \Pr(X \leq a, Y < b) - \Pr(X > a, Y \leq b). \quad (1)$$

Continuous Random Variables

Recall that if X is a continuous RV with a probability density function $f_X(x)$, then

$$\Pr(a \leq X \leq b) = \int_a^b f_X(x)\, dx.$$

Now, we generalize the above:

Definition 18.2. Let X, Y be continuous random variables. Then X and Y are *jointly continuous* if

$$\Pr\big((X,Y) \in D\big) = \iint_D f(x,y)\, dA,$$

where $D \subseteq \mathbb{R}^2$ is any region. The function $f(x,y) =: f_{X,Y}(x,y)$ is the *joint probability density function* of X and Y.

Example 18.3. Let $f(x,y) = \begin{cases} 2e^{-x}e^{-2y} & x, y > 0 \\ 0 & \text{otherwise.} \end{cases}$ It is not too hard to check that $f(x,y)$ defines a joint probability density function for two random variables X and Y, so we compute

(a) $\Pr(X > 1, Y < 1)$:

$$\begin{aligned} \Pr(X > 1, Y < 1) &= \int_1^\infty \int_{-\infty}^1 2e^{-x}e^{-2y}\, dy\, dx \\ &= \int_1^\infty \int_0^1 2e^{-x}e^{-2y}\, dy\, dx \\ &= \int_1^\infty -e^{-x}e^{-2y}\Big|_0^1\, dx = -\int_1^\infty e^{-x}(e^{-2} - 1)\, dx \\ &= (1 - e^{-2}) \int_1^\infty e^{-x}\, dx = \boxed{e^{-1} + e^{-3}}; \end{aligned}$$

(b) $\Pr(X < Y) = \displaystyle\int_0^\infty \int_0^y 2e^{-x}e^{-2y}\, dx\, dy = \int_0^\infty 2(e^{-2y} - e^{-3y})\, dy = \boxed{\frac{1}{3}};$

(c) $F_X(a)$:

$$\begin{aligned} F_X(a) = \Pr(X \le a) &= \Pr(X \le a, Y < \infty) \\ &= \int_0^\infty \int_0^a 2e^{-x}e^{-2y}\, dx\, dy \\ &= \int_0^\infty 2e^{-2y}\, dy \cdot \int_0^a e^{-x}\, dx = \boxed{1 - e^{-a}}, \end{aligned}$$

where the splitting of the integral in (c) is legal by Fubini's Theorem.

More generally, we should be able to get the marginal probability density function from the joint probability density function by writing

$$f_X(x) = \int_D f_{X,Y}(x,y)\, dy \text{ and } f_Y(y) = \int_D f_{X,Y}(x,y)\, dx.$$

Here, the bound D is the *support* of the function $f_{X,Y}$ (i.e., where the function is nonzero).

20 Jointly Continuous Random Variables

Recall that for a single continuous random variable X, we have

$$F_X(a) = \int_{-\infty}^{a} f_X(x)\, dx \implies f_X(a) = \frac{d}{da} F_X(a). \tag{2}$$

Similarly, for jointly continuous random variables X and Y, we have

$$F_{X,Y}(a,b) = \int_{-\infty}^{b} \int_{-\infty}^{a} f_{X,Y}(x,y)\, dx\, dy \implies f_{X,Y}(a,b) = \frac{\partial^2}{\partial a \partial b} F_{X,Y}(a,b). \tag{3}$$

Example 20.1. Pick a point (X,Y) randomly inside the circle $x^2 + y^2 = R^2$, picked with uniform density. Then we may verify that $f_{X,Y}(x,y) = 1/(\pi R^2)$ whenever $x^2 + y^2 \le R^2$, and $f_{X<Y}(x,y) =$ otherwise. From this, we may find $f_X(x)$:

$$f_X(x) = \int\limits_{x^2+y^2 \le R^2} f_{X,Y}(x,y)\, dy = \int\limits_{-\sqrt{R^2-x^2}}^{\sqrt{R^2-x^2}} \frac{dy}{\pi R^2} = \boxed{\frac{2\sqrt{R^2 - x^2}}{\pi R^2}}$$

whenever $x^2 + y^2 \le R^2$, and $\boxed{0}$ otherwise.

Now, define $D := \sqrt{X^2 + Y^2}$, we can find the cumulative distribution function $F_D(t)$: notice that $F_D(t) = 0$ whenever $t < 0$, and $F_D(t) = 1$ whenever $t > R$. Letting $0 \le t \le R$, we know

$$\begin{aligned} F_D(t) := \Pr(D \le t) &= \Pr(X^2 + Y^2 \le t^2) \\ &= \iint\limits_{x^2+y^2 \le t^2} f_{X,Y}(x,y)\, dA \\ &= \frac{1}{\pi R^2} \iint\limits_{x^2+y^2 \le t^2} dA = \frac{\pi t^2}{\pi R^2} = \frac{d^2}{R^2}. \end{aligned}$$

Hence $F_D(t) = (d/R)^2$ when $0 \le t \le R$. We can then compute, for example, $\Pr(D \le \frac{1}{4}R) = F_D(\frac{1}{4}R) = \frac{1}{16}$.

Finally, we compute $\mathbb{E}[D]$: notice $f_D(t) = \left[\frac{t^2}{D^2}\right]' = \frac{2}{R^2}t$, so

$$\mathbb{E}[D] = \int_0^R t f_D(t)\, dt = \int_0^R \frac{2t^2}{R^2}\, dt = \frac{2}{3R^2} t^3 \Big|_0^R = \boxed{\frac{2}{3}R}.$$

Once we are comfortable with joint distributions for two variables, we can describe joint distributions for multiple random variables.

Definition 20.2. Let $X_1, X_2, \ldots, X_n$ be discrete random variables. Then the *joint probability mass function* of the X_i is

$$p(x_1, \ldots, x_n) := \Pr(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n).$$

Definition 20.3. Let $X_1, X_2, \ldots, X_n$ be continuous random variables. Then they are *jointly continuous* if there exists a *joint probability density function* f such that for all $C \subseteq \mathbb{R}^n$, we have

$$\Pr\big((X_1, X_2, \ldots, X_n) \in C\big) = \overbrace{\int \cdots \int_C}^{n \text{ times}} f(x_1, x_2, \ldots, x_n)\, d(x_1, x_2, \ldots, x_n).$$

We note that the marginal mass functions, in the multivariable cases, are not just made of one variable: if $A \subsetneq \{X_1, \ldots, X_n\}$ is nonempty, then f_A is called a marginal probability mass (density) function regardless of the number of variables in A.

Example 20.4. Consider nine dice rolls, and let X_i be the number of i'ts, where $i \in \{1, 2, 3, 4, 5, 6\}$. Then we have $X_i \sim \mathsf{Bin}(9, 1/6)$, but the X_i are jointly distributed as a *multinomial distribution*. For example,

$$\Pr(X_1 = 3, X_2 = X_3 = 2, X_4 = X_5 = 1, X_6 = 0) = \binom{9}{3, 2, 2, 1, 1, 0}\left(\frac{1}{6}\right)^9 = \boxed{\frac{9!}{3!2!2!6^9}}.$$

In contrast, the sums of these X_i are still distributed binomially. For example,

$$\Pr(\text{exactly 5 rolls greater than 2}) = \Pr(X_3 + X_4 + X_5 + X_6 = 5)$$
$$= \binom{9}{5}\left(\frac{4}{6}\right)^5\left(\frac{2}{6}\right)^{9-5} = \boxed{\binom{9}{5}\left(\frac{2}{3}\right)^5\left(\frac{1}{3}\right)^4},$$

so that $X_3 + \cdots + X_6 \sim \mathsf{Bin}(9, 2/3)$.

21 The Gaussian and Exponential Distributions

Recall the following from Math 130A.

Definition 21.1. A continuous random variable X has a *Gaussian* (or *normal*) distribution if it has the probability density function

$$f(t) := \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(t-\mu)^2}{2\sigma^2}\right),$$

for parameters μ and σ. We write $X \sim \mathcal{N}(\mu, \sigma^2)$.

Since $\mathbb{E}(X - \mu) = \mathbb{E}(x) - \mu$ and $\operatorname{Var}(X/\sigma) = \operatorname{Var}(X)/\sigma$, we perform a change of variables from $X \sim \mathcal{N}(\mu, \sigma^2)$ by setting $Z := (X - \mu)/\sigma$, so that $\mathbb{Z} \sim \mathcal{N}(0, 1)$. Note that Z is still normal, so it suffices to study the distribution $\mathcal{N}(0, 1)$ instead, with the simpler density function

$$f(t) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}t^2}.$$

We verify that this is indeed a probability density function. We compute

$$\Pr(X \in \mathbb{R}) = \int_R f(x)\, dx = \frac{1}{2\pi}\int_{\mathbb{R}} e^{-\frac{1}{2}x^2}\, dx,$$

so we must show that $I := \displaystyle\int_{\mathbb{R}} e^{-\frac{1}{2}x^2}\, dx = \sqrt{2\pi}$. This is the famous Gaussian integral, which is best computed by writing

$$I^2 = \left(\int_{\mathbb{R}} e^{-\frac{1}{2}x^2}\,dx\right)^2 = \int_{\mathbb{R}} e^{-\frac{1}{2}x^2}\,dx \cdot \int_{\mathbb{R}} e^{-\frac{1}{2}y^2}\,dy = \iint_{\mathbb{R}^2} e^{-\frac{1}{2}(x^2+y^2)}\,dA,$$

so we make a change into polar coordinates and write

$$I^2 = \iint_{\mathbb{R}^2} e^{-\frac{1}{2}(x^2+y^2)}\,dA = \int_0^\infty \int_0^{2\pi} e^{-\frac{1}{2}r^2} r\,d\theta\,dr,$$

which is straightforward:

$$I^2 = \int_0^\infty \int_0^{2\pi} re^{-\frac{1}{2}r^2}\,d\theta\,dr = \int_0^\infty 2\pi re^{-\frac{1}{2}r^2}\,dr = 2\pi \int_0^\infty e^{-u}\,du = 2\pi,$$

from which it follows that $\boxed{I = \sqrt{2\pi}}$, as the integrand is positive. Now

$$\mathbb{E}(Z) = \int_{\mathbb{R}} xf(x)\,dx = \int_{\mathbb{R}} \frac{x}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}\,dx = 0$$

by symmetry. From here, we observe that

$$\operatorname{Var}(Z) = \mathbb{E}(Z^2) - \cancelto{0}{\mathbb{E}(Z)^2} = \int_{\mathbb{R}} x^2 f(x)\,dx = \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} x^2 e^{-\frac{1}{2}x^2}\,dx =: \frac{J}{\sqrt{2\pi}},$$

so we evaluate J by integration by parts:

$$J = \int_{\mathbb{R}} x \cdot xe^{-\frac{1}{2}x^2}\,dx = \cancelto{0}{x\left(-e^{-\frac{1}{2}x^2}\right)\Big|_{\mathbb{R}}} - \int_{\mathbb{R}} \left(-e^{-\frac{1}{2}x^2}\right)\,dx = I = \sqrt{2\pi},$$

so that $\operatorname{Var}(Z) = J/\sqrt{2\pi} = 1$. This proves the following, after reversing our change of variables:

Proposition 21.2. *Suppose $X \sim \mathcal{N}(\mu, \sigma^2)$. Then $\mathbb{E}[X] = \mu$ and $\operatorname{Var} X = \sigma^2$.*

Definition 21.3. We say that a continuous random variable X is *exponentially distributed* if it has the probability density function $f(t) := \lambda e^{-\lambda t}$ for some parameter λ, whenever $t \geq 0$, and 0 otherwise. We write $X \sim \mathsf{Exp}(\lambda)$.

We verify that f above is indeed a probability density function:

$$\int_0^\infty \lambda e^{-\lambda t} = e^{-\lambda t}\Big|_0^\infty = 0 + 1 = 1.$$

We now compute the expectation and the variance of X:

$$\mathbb{E}[X] = \int_0^\infty x\lambda e^{-\lambda x}\,dx = \cancelto{0}{-xe^{-\lambda x}\Big|_0^\infty} - \int_0^\infty (-e^{-\lambda x})\,dx = \int_0^\infty e^{-\lambda x}\,dx = \frac{1}{\lambda};$$

by integration by parts twice, we may verify $\operatorname{Var} X = 1/\lambda^2$. This proves the following:

Proposition 21.4. *If $X \sim \mathsf{Exp}(\lambda)$, then $\mathbb{E}[X] = 1/\lambda$ and $\operatorname{Var} X = 1/\lambda^2$.*

24 Independence of Random Variables

Recall that events E, F are independent if $\Pr(E \cap F) = \Pr(E) \cdot \Pr(F)$, which holds if and only if $\Pr(E|F) = \Pr(E)$, as $\Pr(E \cap F) = \Pr(E|F) \cdot \Pr(F)$. We now define a similar notion for random variables.

Definition 24.1. Random variables X and Y are *independent* if for all $A, B \subseteq \mathbb{R}$, we have

$$\Pr(X \in A, Y \in B) = \Pr(X \in A) \cdot \Pr(Y \in B).$$

Equivalently, we observe $F_{X,Y}(a, b) = F_X(a)F_Y(b)$.

We have the following tests of independence.

Theorem 24.2. *Let X and Y be discrete random variables. Then X and Y are independent if and only if $p_{X,Y}(x, y) = p_X(x)p_Y(y)$ for all $x, y \in \mathbb{R}$.*

Proof. Let X and Y denote discrete random variables.

($\Longrightarrow$): Suppose X and Y are independent. Take $A = \{x\}$ and $B = \{y\}$, so that

$$p_{X,Y}(x, y) = \Pr(X = x, Y = y) = \Pr(X \in A, Y \in B) = \Pr(X \in A) \Pr(Y \in B)$$
$$= \Pr(X = x) \Pr(Y = y) = p_X(x)p_Y(y).$$

($\Longleftarrow$): Suppose X and Y are discrete and $p_{X,Y}(x, y) = p_X(x)p_Y(y)$ holds identically. Then

$$\Pr(X \in A, Y \in B) = \sum_{x \in A} \sum_{y \in B} p(x, y) = \sum_{x \in A} \sum_{y \in B} p_X(x)p_Y(y)$$
$$= \sum_{x \in a} p_X(x) \cdot \sum_{y \in B} p_Y(y) = \Pr(X \in A) \cdot \Pr(Y \in B).$$

Hence X and y are independent. □

Theorem 24.3. *Let X and Y are continuous random variables. Then X and Y are independent if and only if $f_{X,Y}(x, y) = f_X(x)f_Y(y)$ holds identically.*

Proof. Let X and Y be continuous random variables.

($\Longrightarrow$): Suppose X and Y are independent. Then we know $F(x, y) = F_X(x)F_Y(y)$, so that

$$\frac{\partial^2}{\partial x \partial y} F(x, y) = \frac{\partial^2}{\partial x \partial y}[F_X(x)F_Y(y)] = \frac{\partial}{\partial x} F_X(x) \cdot \frac{\partial}{\partial y} F_Y(y),$$

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

($\Longleftarrow$): Suppose $f_{X,Y}(x, y) = f_X(x)f_Y(y)$ holds identically. Then by Fubini's Theorem, for all $A, B \subseteq \mathbb{R}$,

$$\Pr(X \in A, Y \in B) = \iint\limits_{A \times B} f(x, y)\, dA = \int_A f_X(x)\, dx \cdot \int_B f_Y(y)\, dy = \Pr(X \in A) \Pr(Y \in B),$$

which completes the proof. □

Example 24.4. Consider a sequence of $n+m$ independent trials, where the probability of success is p. Let X be the number of successes in the first n trials, and Y be the number of successes in the last m trials. Then we see $X \sim \mathsf{Bin}(n,p)$ and $Y \sim \mathsf{Bin}(m,p)$, so that

$$p_{X,Y}(x,y) = \Pr(X=x, Y=y) = \binom{n}{x} p^x (1-p)^{n-x} \binom{m}{y} p^y (1-p)^{m-y} = p_X(x) p_Y(y),$$

so X and Y are indeed independent, as we intuitively expect.

In contrast, if Z is the *total* number of successes (i.e., $X = X+Y$), then X and Z are *dependent*: note $\Pr(X=1, Z=0) = 0$, but $\Pr(X=1) \cdot \Pr(Z=0) > 0 \cdot 0 = 0$, so $p_{X,Z} \neq p_X p_Z$.

Example 24.5. Say that the number of people entering the post office per day is distributed $\mathsf{Poi}(\lambda)$. For each person entering, say we have $\Pr(\text{male}) = p$ and $\Pr(\text{female}) = 1-p$. Let X be the number of males in the post office, and Y be the number of females in the post office, so that $X \sim \mathsf{Poi}(p\lambda)$, $Y \sim \mathsf{Poi}((1-p)\lambda)$, and $X+Y \sim \mathsf{Poi}(\lambda)$. We show that X and Y are independent.

Proof. Define the event E by "$X=i$ and $Y=j$" and the event F by "$X+Y=i+j$," so the law of total probability gives

$$\Pr(E) = \Pr(E|F)\Pr(F) + \Pr(E|F^c)\Pr(F^c).$$

Clearly, $\Pr(E|F^c) = 0$, so we see $\Pr(E) = \Pr(E|F)\Pr(F)$. Now, we compute

$$\begin{aligned}
\Pr(X=i, Y=j) &= \left[\binom{i+j}{i} p^i (1-p)^j\right]\left(e^{-\lambda}\frac{\lambda^{i+j}}{(i+j)!}\right) \\
&= \frac{\cancel{(i+j)!}}{i!j!} p^i (1-p)^j e^{-(p\lambda + (1-p)\lambda)} \frac{\lambda^i \lambda^j}{\cancel{(i+j)!}} \\
&= \frac{p^i (1-p)^j e^{-p\lambda} e^{-(1-p)\lambda} \lambda^i \lambda^j}{i!j!} \\
&= \frac{(p\lambda)^i e^{-p\lambda}}{i!} \cdot \frac{((1-p)\lambda)^j e^{-(1-p)\lambda}}{j!} = \Pr(X=i)\Pr(Y=j),
\end{aligned}$$

which establishes independence. □

30 Properties of Continuous Distributions

In this section, we view more examples utilizing continuous distributions. First, we begin with a definition.

Definition 30.1. A random variable X is *memoryless* if

$$\Pr(X > s+t \,|\, X > t) = \Pr(X > s)$$

holds identically. Equivalently, $\Pr(X > s+t) = \Pr(X > s)\Pr(X > t)$ holds identically.

Intuitively, the time spent "waiting" for something to happen (t) does not "cause" it to be more or less liekly to happen in the next s units of times. Note that memorylessness is not independence; for example, if X is memoryless, it may be the case that

$$\Pr(X > 12 \,|\, X > 2) \neq \Pr(X > 12).$$

Example 30.2. Let $U \sim \mathcal{U}(0, 20)$. Then $\Pr(U > 10) = \frac{1}{2}$, and we check

$$\Pr(U > 15 \,|\, U > 5) = \frac{\Pr(U > 15, U > 5)}{\Pr(U > 5)} = \frac{\Pr(U > 15)}{\Pr(5)} = \frac{1/4}{3/4} = \frac{1}{3},$$

which is not $\Pr(U > 15 - 5) = \frac{1}{2}$. Hence, U is not memoryless — knowing "where" a variable is distributed (i.e., in the range $U > 5$) makes it more likely for us to locate it.

It can be proven that if X is memoryless, then in fact $X \sim \mathsf{Exp}(\lambda)$ for some $\lambda > 0$.

Example 30.3. Let $X \sim \mathsf{Exp}(1)$, so that $f_X(x) = e^{-x}$ when $x \geq 0$ and 0 otherwise. Now

$$\Pr(X > 10) = \int_{10}^{\infty} e^{-x}\, dx = -e^{-x}\Big|_{10}^{\infty} = 0 + e^{-10} = e^{-10},$$

and $\Pr(X > 15 \,|\, X > 5) = \Pr(X > 15)/\Pr(X > 5) = e^{-15}/e^{-5} = e^{-10} = \Pr(X > 15 - 5)$. This holds because X is memoryless.

Example 30.4. Say that the average phone call is 10 minutes long, and a person is in line ahead of you on the phone (as with the phone booths of old). We expect that the time in minutes that the person spends on the call, X, is exponentially distributed — waiting a while does not mean that the call is closer to being finished. Granted, people do not generally stay on phone calls for 10 hours, but that is so unlikely that we might as well model this situation as memoryless (and hence exponential). We are given $\mathbb{E}[X] = 10$, so that $X \sim \mathsf{Exp}(\frac{1}{10})$. Then we may compute

$$\Pr(X > 10) = \int_{10}^{\infty} \frac{1}{10} e^{-x/10}\, dx = \frac{1}{e} \approx 0.368, \text{ and}$$

$$\Pr(10 < X < 20) = e^{-x/10}\Big|_{10}^{20} = e^{-1} - e^{-2} \approx 0.233.$$

Example 30.5. A target is randomly shot at. Let X be the horizontal miss distance and Y be the vertical miss distance. We will assume that X, Y are independent random variables and their joint probability density function is differentiable, and that the joined density depends only on the coordinate (X, Y) via $X^2 + Y^2$, the square of the distanced missed; i.e., $f_{X,Y}(x, y) = f_X(x) f_Y(y) = g(x^2 + y^2)$ for some function g. We claim that in fact $X, Y \sim \mathcal{N}(0, \sigma^2)$, for some variance σ^2.

Proof. We know $f_X(x) f_Y(y) = g(x^2 + y^2)$, so taking the partial derivative with respect to x gives

$$f_X'(x) f_Y(y) = 2x g'(x^2 + y^2).$$

Dividing over by what we are originally give, we obtain

$$\frac{f_X'(x)}{f_X(x)} = \frac{2x \cdot g'(x^2 + y^2)}{g(x^2 + y^2)} \implies \frac{f_X'(x)}{x f_X(x)} = \frac{2g'(x^2 + y^2)}{g(x^2 + y^2)}.$$

Now, the left is dependent only on x, and the right only depends on $x^2 + y^2$, so we claim that both expressions are constant. To see this, let $x_1, x_2 \in \mathbb{R}$, with $|x_1| \leq x_2$. Consider the circle $x^2 + y^2 = x_2^2$, so there must exist some $y_1 > 0$ such that (x_1, y_1) sits on the circle. But in that case, $x_1^2 + y_1^2 = x_2^2 + 0^2 = x_2^2$, so that $g(x_1^2 + y_1^2) = g(x_2^2)$, so that

$$\frac{f_X'(x_1)}{xf_X(x_1)} = \frac{2g'(x_1^2 + y_1^2)}{g(x_1^2 + y_1^2)} = \frac{2g'(x_2^2 + 0^2)}{g(x_2^2 + 0^2)} = \frac{f_X'(x_2)}{x_2 f_X(x_2)},$$

so that the function $f_X'(x)/(xf_X(x)) =: c$ is constant. This, after rearrangement, yields the differential equation $f_X'/f_X = cx$, which is separable, so we obtain $f_X(x) = ke^{-cx^2/2}$. Clearly, $\mathbb{E}[X] = 0$, so now the substitution $c =: 1/\sigma^2$ shows $X \sim \mathcal{N}(0, \sigma^2)$. A symmetric argument shows that $Y \sim \mathcal{N}(0, \sigma^2)$ as well. □

34 Independence Revisited

Recall that two continuous random variables X, Y are independent if and only if $f_{X,Y}(x, y) = f_X(x)f_Y(y)$, with a similar result holding for discrete random variables. We prove a statement that is slightly stronger than this:

Proposition 34.1. *Continuous random variables X, Y are independent if and only if their joint probability density function has the form*

$$f_{X,Y}(x, y) = h(x) \cdot g(y)$$

for some functions h, g, for all $x, y \in \mathbb{R}$. A similar result holds for discrete random variables.

Proof. ($\implies$): This direction is immediate from Theorem 24.3.

($\impliedby$): Suppose $f_{X,Y}(x, y) = h(x) \cdot g(y)$ holds identically. We know that

$$1 = \iint_{\mathbb{R}^2} f_{X,Y}(x, y)\, dA = \int_{\mathbb{R}} g(y)\, dy \cdot \int_{\mathbb{R}} h(x)\, dx = c_1 c_2$$

for constants $c_1 := \int_{\mathbb{R}} g(y)\, dy$ and $c_2 := \int_{\mathbb{R}} h(x)\, dx$. Certainly, $c_1, c_2 \in \mathbb{R}^+$ (i.e., the integrals on the right actually converge), but

$$f_Y(y) = \int_{\mathbb{R}} f_{X,Y}(x, y)\, dx = g(y) \int_{\mathbb{R}} h(x)\, dx = g(y)c_2 \text{ and}$$

$$f_X(x) = \int_{\mathbb{R}} f_{X,Y}(x, y)\, dy = h(x) \int_{\mathbb{R}} g(y)\, dy = h(x)c_1.$$

Hence

$$f_{X,Y}(x, y) = h(x)g(y) = \frac{f_X(x)f_Y(y)}{c_1 c_2} = f_X(x)f_Y(y),$$

so X and Y are indeed independent. □

Example 34.2. Let $f(x, y) = 6e^{-2x}e^{-2y}$ whenever $x, y \in \mathbb{R}^+$, and 0 elsewhere. Certainly, we may write $f(x, y) = 6e^{-2x} \cdot e^{-3y}$, so it seems like X and Y are independent — however, this function is defined piecewise, so we must check the piecewise parts. But

$$f(x, y) = \begin{cases} 6e^{-2x} & x > 0 \\ 0 & x \leq 0 \end{cases} \cdot \begin{cases} e^{-3y} & y > 0 \\ 0 & y \leq 0 \end{cases},$$

which properly shows that X and Y are independent.

Example 34.3. Let $f(x, y) = 24xy$ whenever $x, y, x + y \in (0, 1)$, and 0 everywhere else. This function *looks* independent, but we write

$$f(x, y) = 24x \cdot y \cdot \begin{cases} 1 & x, y, x + y \in (0, 1) \\ 0 & \text{elsewhere.} \end{cases}$$

The condition $x + y \in (0, 1)$ seems to be an issue — it involves *both* variables subject to a relation, so we suspect that this function does not represent a joint probability density function of two independent random variables X, Y. Indeed, we can check that this is the case by computing $f_X(x)$ and $f_Y(y)$ directly, taking care of the piecewise function.

Definition 34.4. Let $X_1, X_2, \ldots$ be a sequence of random variables, independent and identically distributed (IID). We say X_n is a *record value* if $X_n > X_i$ for all $1 \le i < n$.

Example 34.5. Let $X_1, X_2, \ldots$ be IID, and let A_i be the event that X_i is a record value. Is A_{n+1} independent of A_n? This is not too obvious, but by symmetry we may ask the equivalent question, "is A_n equivalent to A_{n+1}?" But now X_n being a record is not dependent on $X_{n+1}, X_{n+2}, \ldots$, i.e., X_n being a record is independent of future records. Thus, A_{n+1} **is** independent of A_n, as unintuitive as it may or may not seem.

We remark that independence is symmetric, which may be occasionally useful.

38 Sums of Independent Random Variables

Suppose X and Y are independent. If X and Y are discrete, then

$$p_{X+Y}(a) := \Pr(X + Y = a) = \sum_k \Pr(X = k)\Pr(Y = a - k) = \sum_k p_X(k)p_Y(a - k). \tag{4}$$

Similarly, if X and Y are continuous and independent, we find the cumulative

$$F_{X+Y}(t) := \Pr(X + Y \le t) = \Pr(X \le t - Y) = \int_{\mathbb{R}} \int_{-\infty}^{t-y} f_X(x)f_Y(y)\,dx\,dy,$$

so that

$$f_{X+Y}(t) = \frac{d}{dt}F_{X+Y}(t) = \int_{\mathbb{R}} \frac{d}{dt}\left[\int_{-\infty}^{t-y} f_X(x)f_Y(y)\,dx\right] dy = \int_{\mathbb{R}} f_X(t - y)f_Y(y)\,dy. \tag{5}$$

The last expression should be familiar from differential equations (Math 3D):

Definition 38.1. Let f, g be integrable functions on $\mathbb{R}$. We define the (continuous) *convolution* of f and g by

$$(f * g)(t) := \int_{\mathbb{R}} f(t - y)g(y)\,dy,$$

provided that the integral converges.

Proposition 38.2. *We have that $(f * g)(t) = (g * f)(t)$ whenever the integral for the convolution is defined.*

Proof. Check that

$$(f*g)(t) = \int_{-\infty}^{\infty} f(t-y)g(y)\,dy = -\int_{\infty}^{-\infty} f(u)g(t-u)\,du = \int_{-\infty}^{\infty} f(u)g(t-u)\,du = (g*f)(t),$$

which completes the proof. □

We remark that similar results hold for the *discrete convolution*

$$(f * g)(n) := \sum_{k=-\infty}^{\infty} f(k)g(n-k).$$

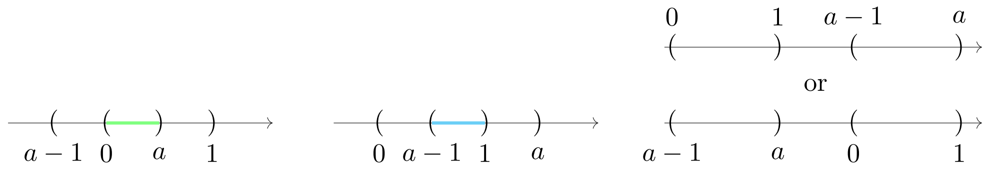
Equations (4) and (5) show that the distributions of sums of random variables are found by convoluting the marginal distributions for each variable.

Example 38.3. Let $X, Y \sim \mathcal{U}(0,1)$ be independent. Then $f_X(t) = f_Y(t) = 1$ whenever $t \in [0,1]$, and 0 elsewhere. Then

$$f_{X+Y}(a) = \int_{\mathbb{R}} f_X(a-y)f_Y(y)\,dy = \int_{-\infty}^{\infty} 1^*\,dy,$$

where the $\cdot^*$ indicates that this function is sometimes zero off its natural support. Hence, we need to simplify the bounds, so we consider three cases:

Notice that $f_Y(a-y) = 1$ when $a-1 < y < a$, while $f_X(y) = 1$ when $0 < y < 1$, so the function 1^* is nonzero whenever we lie in the intersection of intervals $(0,1) \cap (a-1,a)$. We sketch our three cases below:



(a) Case I (b) Case II (c) Case III

Case I: $0 < a < 1$. In this case, $(0,1) \cap (a-1,a) = (0,a)$, so

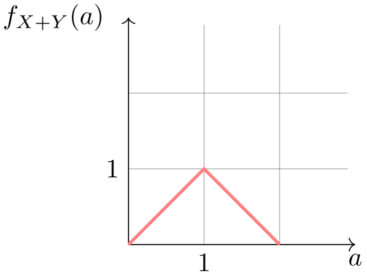
$$f_{X+Y}(a) = \int_0^a 1\,dy = y\Big|_0^a = a.$$

Case II: $1 < a < 2$. In this case, $(0,1) \cap (a-1,a) = (a-1,1)$, so

$$f_{X+Y}(a) = \int_{a-1}^1 1\,dy = y\Big|_{a-1}^1 = 1-(a-1) = 2-a.$$

Case III: $a \in (-\infty,0] \cup [2,\infty)$. in this case, $(0,1) \cap (a-1,a) = \varnothing$, so the integral is 0.

Hence, the density of the sum is given by $f_{X+Y}(a) = \begin{cases} a & a \in (0,1] \\ 2-a & a \in (1,2) : \\ 0 & \text{elsewhere} \end{cases}$



Notably, the graph on the previous page reveals that $X+Y$ is not uniform.

However, for other types of distributions, the sum behaves "nicely" assuming certain conditions are met.

Proposition 38.4. *Let X, Y be independent with $X \sim \mathsf{Poi}(\lambda_1)$ and $Y \sim \mathsf{Poi}(\lambda_2)$. Then $X+Y \sim \mathsf{Poi}(\lambda_1+\lambda_2)$.*

Proof. We compute via the discrete convolution

$$\begin{aligned} p_{X+Y}(n) = \sum_{k=0}^{n} p_X(k) p_Y(n-k) &= \sum_{k=0}^{n} \left[\frac{e^{-\lambda_1}\lambda_1^k}{k!} \cdot \frac{e^{-\lambda_2}\lambda_2^{n-k}}{(n-k)!} \right] \\ &= e^{-\lambda_1-\lambda_2} \sum_{k=0}^{n} \frac{\lambda_1^k \lambda_2^{n-k}}{k!(n-k)!} \cdot \frac{n!}{n!} \\ &= e^{-(\lambda_1+\lambda_2)} \sum_{k=0}^{n} \binom{n}{k} \lambda_1^k \lambda_2^{n-k} \frac{1}{n!} \\ &= \frac{e^{-(\lambda_1+\lambda_2)}}{n!} (\lambda_1+\lambda_2)^n, \end{aligned}$$

where the last step follows from the binomial theorem. This shows $X+Y \sim \mathsf{Poi}(\lambda_1+\lambda_2)$. □

Proposition 38.5. *Let X, Y be independent with $X \sim \mathsf{Bin}(n,p)$ and $Y \sim \mathsf{Bin}(m,p)$. Then $X+Y \sim \mathsf{Bin}(n+m,p)$.*

Proof. We first state the combinatorial identity

$$\binom{n+m}{k} = \sum_{i=0}^{k} \binom{n}{i} \binom{m}{k-i},$$

which can be verified by induction. Now using this identity, we compute

$$\begin{aligned} p_{X+Y}(k) &= \sum_{i=0}^{k} p_X(i) p_Y(k-i) \\ &= \sum_{i=0}^{k} \binom{n}{i} p^i (1-p)^{n-i} \binom{m}{k-i} p^{k-i} (1-p)^{m-(k-i)} \\ &= \sum_{i=0}^{k} \binom{n}{i} \binom{m}{k-i} p^k (1-p)^{(n+m)-k} \\ &= p^k (1-p)^{(n+m)-k} \sum_{i=0}^{k} \binom{n}{i} \binom{m}{k-i} = \binom{n+m}{k} p^k (1-p)^{(n+m)-k}, \end{aligned}$$

so that it follows $X+Y \sim \mathsf{Bin}(n+m,p)$. □

We also state the following result about sums of normal random variables. We omit the proof, for it is quite messy and unenlightening.

Proposition 38.6. *If $X_1, \dots, X_n$ are independent with $X_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$, then*

$$X_1 + X_2 + \cdots + X_n \sim \mathcal{N}(\mu_1 + \cdots + \mu_n, \sigma_1^2 + \cdots + \sigma_n^2).$$

Proof. See here[1] for the proof where $n = 2$, from which the result follows by induction. □

Example 38.7. Suppose a basketball team plays 44 games, 26 against division A and 10 against division B teams, and suppose Pr(win against A) = 0.4 and Pr(win against B) = 0.7. Let X_A resp. X_b be the number of wins against division A resp. B teams, so that $X_A \sim \mathsf{Bin}(26, 0.4)$ and $X_B \sim \mathsf{Bin}(18, 0.7)$. Now, if we wanted to find the total number of games won, we need to find $X_A + X_B$, which is cumbersome. But we may compute $\mathbb{E}[X_A] = 10.4, \mathrm{Var}(X_A) = 6.24, \mathbb{E}[X_B] = 12.6$, and $\mathrm{Var}(X_B) = 3.78$, so that $X_A \approx \mathcal{N}(10.4, 6.24)$ and $X_B \approx \mathcal{N}(12.6, 3.78)$, so by Proposition 38.6, $X_A + X_B \approx \mathcal{N}(23, 10.02)$. Now

$$\Pr(\geq 25 \text{ games won}) = \Pr(X_A + X_B \geq 25) \approx \int_{24.5}^{\infty} \frac{1}{10.02\sqrt{2\pi}} \exp\left(-\frac{(x-23)^2}{2 \cdot 10.02}\right) dx,$$

where the lower bound of 24.5 comes from *continuity correction.* This integral may be computed by calculator, or by reducing to the standard normal $\mathcal{N}(0,1)$ via z-scores and using a lookup table, but we obtain $\boxed{32\%}$ as our approximate probability.

Notice that we disregarded the upper bound of 44 total games in the computation above, but neglecting this bound only gives a negligible error with on an order of magnitude comparable to 10^{-10}.

40 Gamma Distributions (I)

Recall the following.

Definition 40.1. Let $f : U \to \mathbb{R}$ be a function. Then the *support* of f is the largest set on which f is nonzero:

$$\mathsf{supp}(f) := \{x \in \mathsf{dom}(f) : f(x) \neq 0\}.$$

The following proposition will be useful.

Proposition 40.2. *Let $f, g : \mathbb{R} \to \mathbb{R}$. If $\mathsf{supp}(f), \mathsf{supp}(g) \subseteq [0, \infty)$, then*

$$(f * g)(t) = \int_0^t f(t-y)g(y)\, dy.$$

Proof. If $g(y) \neq 0$, then $y \in [0, \infty)$. Similarly, if $f(t-y) \neq 0$, we have $0 < t - y < \infty$, so that $y \in (-\infty, t]$. The intersection of these intervals is $[0, t]$ (allowing this to be degenerate if $t \leq 0$), from which the proposition follows immediately. □

We remark that in terms of probability, the above proposition allows us to simplify convolution calculations if necessary, especially if we know that our probability density function only takes non-negative inputs.

[1] `https://en.wikipedia.org/wiki/Sum_of_normally_distributed_random_variables`

We now examine gamma distributions. Recall that the time that one event randomly occurs is usually modeled by an exponential random variable: $Y \sim \mathsf{Exp}(\lambda)$. We generalize this: the gamma distribution models the time when t events occur.

Definition 40.3. Let $\lambda, t \in \mathbb{R}^+$. Then we define the *gamma distribution*, $Y \sim \mathsf{Gam}(t, \lambda)$, by the probability density function

$$f_Y(y) := \frac{\lambda e^{-\lambda y}(\lambda y)^{t-1}}{\Gamma(t)} \text{ when } y > 0,$$

and $f_Y(y) = 0$ otherwise, where $\Gamma(t)$ denotes the *gamma function* $\Gamma(t) := \displaystyle\int_0^\infty x^{t-1}e^{-x}\,dx$.

We check that f_Y above is actually a well-defined probability density function: we have

$$\int_{\mathbb{R}} f_Y(y)\,dy = \int_0^\infty \frac{\lambda e^{-\lambda y}(\lambda y)^{t-1}}{\Gamma(t)}\,dy = \frac{\lambda^t}{\Gamma(t)}\int_0^\infty e^{-\lambda y}y^{t-1}\,dy.$$

To evaluate this last integral, make the change of variables $u = \lambda y$, so $du = \lambda\,dy$ and

$$\int_0^\infty e^{-\lambda y}y^{t-1}\,dy = \int_0^\infty e^{-u}\left(\frac{u}{\lambda}\right)^{t-1}\cdot\frac{du}{\lambda} = \lambda^{-t}\int_0^\infty e^{-u}u^{t-1}\,du = \frac{\Gamma(t)}{\lambda^t}.$$

From here, it follows that $\displaystyle\int_{\mathbb{R}} f_Y(y)\,dy = \frac{\lambda^t}{\Gamma(t)}\cdot\frac{\Gamma(t)}{\lambda^t} = 1$, so this is a well-defined probability density function.

We also note the relationship $\mathsf{Exp}(\lambda) = \Gamma(1, \lambda)$, which follows from the following properties of the gamma function:

Proposition 40.4. *If $t \in (1, \infty)$, then $\Gamma(t) = (t-1)\Gamma(t-1)$.*

Proof. We have, by integration by parts,

$$\begin{aligned}\Gamma(t) = \int_0^\infty y^{t-1}e^{-y}\,dy &= -y^{t-1}e^{-y}\Big|_0^\infty + \int_0^\infty (t-1)y^{t-2}e^{-y}\,dy \\ &= 0 + (t-1)\Gamma(t-1) = (t-1)\Gamma(t-1),\end{aligned}$$

which completes the proof. □

Corollary 40.5. *For all $n \in \mathbb{Z}^+$, we have $\Gamma(n) = (n-1)!$.*

Proof. Proceed by induction. □

44 Gamma Distributions (II)

Our work in the previous section allows us to derive the following important property of the gamma distribution.

Theorem 44.1. *Let X, Y be independent with $X \sim \mathsf{Gam}(s, \lambda)$ and $Y \sim \mathsf{Gam}(t, \lambda)$. Then*

$$X + Y \sim \mathsf{Gam}(s+t, \lambda).$$

Proof. Let $a > 0$. Notice that $\mathsf{supp}(f_X) = \mathsf{supp}(f_Y) = [0, \infty)$, so Proposition 40.2 applies and we compute

$$\begin{aligned}
f_{X+Y}(a) &= \int_0^a f_X(a-y) f_Y(y)\, dy \\
&= \frac{1}{\Gamma(s)\Gamma(t)} \int_0^a \lambda e^{-\lambda(a-y)} \lambda^{s-1} (a-y)^{s-1} \cdot \lambda e^{-\lambda y} \lambda^{t-1} y^{t-1}\, dy \\
&= \frac{\lambda e^{-\lambda a} \lambda^{s+t-1}}{\Gamma(s)\Gamma(t)} \int_0^a (a-y)^{s-1} y^{t-1}\, dy \\
&= \frac{\lambda e^{-\lambda a} \lambda^{s+t-1}}{\Gamma(s)\Gamma(t)} \int_0^1 (a-au)^{s-1} (au)^{t-1} \cdot a\, du \quad (y =: au, dy = a\, du) \\
&= \frac{\lambda e^{-\lambda a} (\lambda a)^{s+t-1}}{\Gamma(s)\Gamma(t)} \int_0^1 (1-u)^{s-1} u^{t-1}\, du.
\end{aligned}$$

Now, the value

$$c := \frac{1}{\Gamma(s)\Gamma(t)} \int_0^1 (1-u)^{s-1} u^{t-1}\, du$$

is constant with respect to a, but the total area under the probability density function f_{X+Y} must equal 1, i.e.,

$$\int_0^\infty f_{X+Y}(a)\, da = c \int_0^\infty \lambda e^{-\lambda a} (\lambda a)^{s+t-1}\, da = 1.$$

But we know the distribution of $\mathsf{Gam}(s+t, \lambda)$, which is $\dfrac{\lambda e^{-\lambda a} (\lambda a)^{s+t-1}}{\Gamma(s+t)}$, so this implies that $f_{X+Y} \propto \mathsf{Gam}(s+t, \lambda)$. But both of these are probability density functions, so equality must hold, so this completes the proof. □

One of the integrals above is important in its own right.

Definition 44.2. Let $s, t \in \mathbb{R}$. The *beta function* $B(s,t)$ is the integral

$$B(s,t) := \int_0^1 (1-x)^{s-1} x^{t-1}\, dx,$$

provided that the integral converges.

Then, fixing notation as per our proof above, we observe

$$c = \frac{B(s,t)}{\Gamma(s)\Gamma(t)} = \frac{1}{\Gamma(s+t)}.$$

This proves the following:

Lemma 44.3. *Let $s, t \in \mathbb{R}$. Then $B(s,t) = \dfrac{\Gamma(s)\Gamma(t)}{\Gamma(s+t)}$, provided that everything is well-defined.*

From here, of course we may use induction on Theorem 44.1 to show that if $X_1, \ldots, X_n$ are independent with $X_i \sim \mathsf{Gam}(t_i, \lambda)$, then $X_1 + \cdots + X_n \sim \mathsf{Gam}(t_1 + \cdots + t_n, \lambda)$. This gives us the following corollary:

Corollary 44.4. *If $X_1, \ldots, X_n$ are independent with $X_i \sim \mathsf{Exp}(\lambda)$, then $X_1 + \cdots + X_n \sim \mathsf{Gam}(n, \lambda)$.*

Chi-Square Distributions

Let us examine a distribution that, at first, seems completely unrelated to the gamma distribution. However, we will uncover a very enlightening connection, which will prove a fairly substantial result about the gamma function.

Definition 44.5. Let $Z_1, Z_2, \ldots, Z_n$ be independent standard normal variables: $Z_i \sim \mathcal{N}(0,1)$. We define the *chi-square distribution with n degrees of freedom* by $Y := Z_1^2 + Z_2^2 + \cdots + Z_n^2$, and we write $Y \sim \chi^2(n)$.

Intuitively, $\chi^2(n)$ tells us the distribution of the square of the distance from the origin in $\mathbb{R}^n$, where points are chosen so that each coordinate is randomly distributed normally. Our main goal will be to find the probability density function of $\chi^2(n)$. We start with $\chi^2(1)$: let $Y \sim \chi^2(1)$, so that we write $Y = Z^2$, where $Z \sim \mathcal{N}(0,1)$. Its *cumulative distribution* function $F_Y(y)$ is 0 when $y \le 0$, and when $y > 0$,

$$\begin{aligned} F_Y(y) := \Pr(Z^2 \le y) = \Pr(|Z| \le \sqrt{y}) \\ = \int_{-\sqrt{y}}^{\sqrt{y}} \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\,dx = \frac{2}{\sqrt{2\pi}} \int_0^{\sqrt{y}} e^{-x^2/2}\,dx. \end{aligned}$$

This is usually a very tricky integral, but we do not need to compute it: the fundamental theorem of calculus gives

$$f_Y(y) = \frac{d}{dy}\left[\frac{2}{\sqrt{2\pi}} \int_0^{\sqrt{y}} e^{-x^2/2}\,dx\right] = \frac{2}{\sqrt{2\pi}} e^{-y/2} \cdot \frac{1}{2\sqrt{y}} = \frac{1}{\sqrt{2\pi y}} e^{-y/2} \text{ where } y > 0.$$

But now, we compare the following probability density function for a *gamma* distribution:

$$\mathsf{Gam}\left(\frac{1}{2}, \frac{1}{2}\right) = \frac{\frac{1}{2} e^{-y/2} \left(\frac{1}{2} y\right)^{-1/2}}{\Gamma(\frac{1}{2})} = \frac{1}{2\Gamma\left(\frac{1}{2}\right)} e^{-y/2} \sqrt{\frac{2}{y}} = \frac{1}{\Gamma\left(\frac{1}{2}\right)\sqrt{2y}} e^{-y/2}.$$

This density function is *proportional* to $f_Y(y)$; since both are probability density functions, we must have equality: $\chi^2(1) = \mathsf{Gam}\left(\frac{1}{2}, \frac{1}{2}\right)$. Furthermore, from this we know that if $Y \sim \chi^2(n)$, then $Y = Z_1^2 + \cdots + Z_n^2$ with $Z_i \sim \mathcal{N}(0,1)$, we observe $Z_i^2 \sim \chi^2(1) = \Gamma\left(\frac{1}{2}, \frac{1}{2}\right)$, so Theorem 44.1 gives the following result:

Theorem 44.6. *Let $n \in \mathbb{Z}^+$. Then* $\chi^2(n) = \mathsf{Gam}\left(\frac{n}{2}, \frac{1}{2}\right)$.

Incidentally, we have proven the following by comparing coefficients:

Proposition 44.7. $\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$.

Now, Proposition 40.4 tells us that we immediately know what $\Gamma\left(\frac{2k+1}{2}\right)$ is for all $k \in \mathbb{Z}$:

$$\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi} \implies \Gamma\left(\frac{3}{2}\right) = \frac{1}{2}\Gamma\left(\frac{1}{2}\right) = \frac{\sqrt{\pi}}{2} \implies \Gamma\left(\frac{5}{2}\right) = \frac{3}{2}\Gamma\left(\frac{3}{2}\right) = \frac{3\sqrt{\pi}}{4} \implies \cdots.$$

This is a fairly interesting result, and in fact the values of the gamma function at integers and half-integers are the only ones we have closed-form expressions for.

48 Log-Normal Distributions and Conditionals

Definition 48.1. A continuous random variable Y is *log-normal* if $\ln(Y) \sim \mathcal{N}(\mu, \sigma^2)$. We write $Y \sim \mathsf{Log}(\mu, \sigma)$.

We remark that it follows $Y = e^X$, so $X \sim \mathcal{N}(\mu, \sigma^2)$ if Y is log-normal. The log-normal distribution is used to model growth with many small percentage changes, or simply just for convenience, noting the property $\ln(ab) = \ln a + \ln b$.

Example 48.2. Let S_n denote the price of a security after n weeks (after some fixed reference time), for all $n \in \mathbb{Z}^+$. A common model for this is to set $X_n := S_n/S_{n-1} \sim \mathsf{Log}(\mu, \sigma)$. Say, in this case, we have $\mu = 0.0165$ and $\sigma = 0.0730$. Then we may write

$$\begin{aligned} \Pr(\text{price higher next week}) &= \Pr(X_n > 1) \\ &= \Pr(\ln X_n > 0) \approx \boxed{59\%}, \end{aligned}$$

noting that $\ln X_n$ is normally distributed with mean μ and variance σ^2. Similarly, we compute, assuming the X_i are independent:

$$\begin{aligned} \Pr(\text{price grows 2 weeks in a row}) &= \Pr(X_n > 1 \text{ and } X_{n+1} > 1) \\ &= \Pr(X_n > 1)^2 \approx \boxed{35\%}, \end{aligned}$$

where the squaring is legal because the variables X_i are identical, and we made the independence assumption. Finally,

$$\begin{aligned} \Pr(\text{price is higher after 2 weeks}) &= \Pr\left(\frac{S_{n+2}}{S_n} > 1\right) \\ &= \Pr\left(\frac{S_{n+2}}{S_{n+1}} \cdot \frac{S_{n+1}}{S_n} > 1\right) = \Pr(X_{n+2}X_{n+1} > 1) \\ &= \Pr(\ln X_{n+2} + \ln X_{n+1} > 0). \end{aligned}$$

Now, $\ln X_i \sim \mathcal{N}(\mu, \sigma^2)$, so that $\ln X_{n+2} + \ln X_{n+1} \sim \mathcal{N}(2\mu, 2\sigma^2)$, so calculating this probability directly gives $\boxed{63\%}$.

[6.4]/[6.5] Conditional Distributions

Recall that if E and F are events, then $\Pr(E|F) = \Pr(E \cap F)/\Pr(F)$. Similarly, we define the following for random variables:

Definition 48.3. Let X and Y be discrete random variables. We define the *conditional probability mass function* by

$$p_{X|Y}(x|y) := \Pr(X = x \,|\, Y = y) = \frac{\Pr(X = x, Y = y)}{\Pr(Y = y)} = \frac{p_{X,Y}(x, y)}{p_Y(y)}.$$

[The function $p_{Y|X}$ is defined analogously.] Similarly, we define the *conditional cumulative distribution function* by

$$F_{X|Y}(a|y) : \Pr(X \le a \,|\, Y = y) = \sum_{x \le a} p_{X|Y}(x|y).$$

We immediately get the following result.

Proposition 48.4. *Let X, Y be discrete random variables. Then X and Y are independent if and only if $p_{X|Y}(x|y) = p_X(x)$ and similarly $p_{Y|X}(y|x) = p_Y(y)$.*

Proof. We know X and Y are independent if and only if $p_{X,Y}(x, y) = p_X(x)p_Y(y)$. Dividing over by p_X or p_Y completes the proof. □

The continuous case requires a bit more justification, but it is basically the the same as in the discrete case:

Definition 48.5. Let X, Y be continuous random variables. We define the continuous *conditional probability density function* by

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

and the *cumulative distribution function* by

$$F_{X|Y}(a|y) := \int_{-\infty}^{a} f_{X|Y}(x|y)\, dx.$$

We explain why the definition for the conditional probability density above makes sense. Suppose we wanted to determine the probability $\Pr(X \in (a, b) \,|\, Y = y)$, for $a < b$. For a small Δx, we can consider the probabilities

$$\Pr(X \in (x_i, x_i + \Delta x) \,|\, Y = y),$$

taking $a = x_0 < x_1 < \cdots < x_n = b$. Now for sufficiently small Δy, we have

$$\begin{aligned}\Pr(X \in (x_i, x_i + \Delta x) \,|\, Y = y) &\approx \Pr(X \in (x_i, x_i + \Delta x) \,|\, Y \in (y, y + \Delta y)) \\ &= \frac{\Pr(X \in (x_i, x_i + \Delta x) \text{ and } Y \in (y, y + \Delta y))}{\Pr(Y \in (y, y + \Delta y))} \\ &\approx \frac{f_{X,Y}(x_i, y)\Delta x \,\cancel{\Delta y}}{f_Y(y)\cancel{\Delta y}},\end{aligned}$$

so summing and taking the limit as $n \to \infty$ (alternatively, $\Delta x \to 0$) gives

$$\Pr(X \in (a, b) \,|\, Y = y) \approx \sum_{i=0}^{n} \frac{f_{X,Y}(x_i, y)}{f_Y(y)} \Delta x \longrightarrow \int_a^b \frac{f_{X,Y}(x, y)}{f_Y(y)}\, dx,$$

so our definition is reasonable.

Example 48.6. Let $X \sim \mathsf{Poi}(\lambda_1)$ and $Y \sim \mathsf{Poi}(\lambda_2)$ be independent random variables. We compute the conditional probability mass function of X, given that $X + Y = n$. By independence, we have $X + Y \sim \mathsf{Poi}(\lambda_1 + \lambda_2)$, so we have

$$\begin{aligned}\Pr(X = k \,|\, X + Y = n) = \frac{\Pr(X = k, X + Y = n)}{\Pr(X + Y = n)} &= \frac{\Pr(X = k, Y = n - k)}{\Pr(X + Y = n)} \\ &= \frac{\Pr(X = k) \Pr(Y = n - k)}{\Pr(X + Y = n)} \\ &= \frac{\cancel{e^{-\lambda_1}} \frac{\lambda_1^k}{k!} \cdot \cancel{e^{-\lambda_2}} \frac{\lambda_2^{n-k}}{(n-k)!}}{\cancel{e^{-(\lambda_1+\lambda_2)}} \frac{(\lambda_1+\lambda_2)^n}{n!}} \\ &= \binom{n}{k} \frac{\lambda_1^k \lambda_2^{n-k}}{(\lambda_1 + \lambda_2)^n}.\end{aligned}$$

Now, some rearrangement yields

$$\binom{n}{k}\frac{\lambda_1^k\lambda_2^{n-k}}{(\lambda_1+\lambda_2)^n} = \binom{n}{k}\frac{\lambda_1^k}{(\lambda_1+\lambda_2)^k}\cdot\frac{\lambda_2^{n-k}}{(\lambda_1+\lambda_2)^{n-k}},$$

so that $X \,|\, X+Y=n \sim \mathsf{Bin}\left(n, \frac{\lambda_1}{\lambda_1+\lambda_2}\right)$.

Example 48.7. Let $f(x,y)=e^{-x/y}e^{-y}/y$ when $x,y>0$, and $f(x,y)=0$ otherwise. We compute the probability $\Pr(X>1\,|\,Y=y)$ whenever $y>0$. First, we find

$$f_Y(y)=\int_0^\infty \frac{e^{-x/y}e^{-y}}{y}\,dx \overset{u=x/y}{=} e^{-y}\int_0^\infty e^{-u}\,du = -e^{-y}e^{-u}\Big|_{u=0}^{\infty} = e^{-y}.$$

Hence

$$f_{X|Y}(x|y)=\frac{f(x,y)}{f_Y(y)}=\frac{\frac{1}{y}e^{-x/y}\cancel{e^{-y}}}{\cancel{e^{-y}}}=\frac{e^{-x/y}}{y} \text{ whenever } x,y>0,$$

and 0 otherwise. Now we simply need to compute

$$\Pr(X>1\,|\,Y=y)=\int_1^\infty f_{X|Y}(x|y)\,dx=\int_1^\infty\frac{e^{-x/y}}{y}\,dx=-e^{-x/y}\Big|_1^\infty=\boxed{e^{-1/y}}.$$

50 Combining Continuous and Discrete Variables

Let X be a *continuous* random variable and N be a *discrete* random variable, and suppose we want to find X conditioned on N, or vice versa. Certainly, we **cannot** write "$f_{X|N}(x|n)=f(x,n)/p_N(n)$" or "$p_{N|X}(n|x)=p(n,x)/f_X(x)$," because the "joint distributions" $f(x,n)$ and $p(n,x)$ are not even well-defined. Instead, we can get around this issue by approximating

$$f_X(x)\approx\frac{\Pr(X\in(x,x+\Delta x))}{\Delta x} \implies f_{X|N}(x|n)\approx\frac{\Pr(X\in(x,x+\Delta X)\,|\,N=n)}{\Delta x},$$

so by Bayes' Theorem,

$$\frac{\Pr(X\in(x,x+\Delta x),N=n)}{\Pr(N=n)\Delta x}\overset{\text{Bayes}}{=}\frac{\Pr(N=n\,|\,X\in(x,x+\Delta x))\cdot\Pr(X\in(x,x+\Delta x))}{\Pr(N=n)\Delta x}.$$

Taking $\Delta x\to 0$, we get $f_{X|N}(x|n)=\dfrac{\Pr(N=n\,|\,X=x)}{\Pr(N=n)}\cdot f_X(x)$. Hence, we make the following definition.

Definition 50.1. Let X be a continuous random variable and N be a discrete random variable. Then we define the *conditional distributions* $p_{N|X}$ and $f_{X|N}$ by

$$\begin{aligned} &p_{N|X}(n|x):=\Pr(N=n\,|\,X=x) \text{ and}\\ &f_{X|N}(x|n):=\frac{\Pr(N=n\,|\,X=x)}{\Pr(N=n)}\cdot f_X(x)=\frac{p_{N|X}(n|x)f_X(x)}{p_N(n)}.\end{aligned}$$

We note that unlike the derivation for $f_{X|N}$, the definition of $p_{N|X}$ is straightforward.

Example 50.2. Let $X \sim \mathcal{U}(0,1)$. When $X = x$, take $n+m$ independent trials with probability of success x. If N is the number of successes, then certainly $N \,|\, X = x \sim \mathsf{Bin}(n+m, x)$, so

$$p_{N|X}(n|x) = \binom{n+m}{n} x^n (1-x)^m.$$

For $x \in (0,1)$, we find $f_{X|N}(x|n)$:

$$f_{X|N}(x|n) = \frac{p_{N|X}(n|x) f_X(x)}{p_N(n)} = \frac{\binom{n+m}{n} x^n (1-x)^m}{p_N(n)} \cdot 1.$$

Now, define $c := \binom{n+m}{n} / p_N(n)$, which is constant with respect to x. Treating $f_{X|N}(x|n)$ as a function of x, write $f_{X|N}(x|n) = cx^n(1-x)^m$, but we know that

$$1 = \int_0^1 f_{X|N}(x|n)\,dx = c\int_0^1 x^n(1-x)^m\,dx = c \cdot B(n+1, m+1),$$

where B denotes the beta function. Hence, $c = 1/B(n+1, m+1)$. This distribution has a special name:

Definition 50.3. Let $a, b > 0$. We say that a random variable X has a *beta distribution*, denoted $X \sim \mathsf{Beta}(a,b)$, if its probability density function is

$$f_X(x) = \frac{x^{a-1}(1-x)^{b-1}}{B(a,b)} \text{ for } x \in (0,1),$$

and $f_X(x) = 0$ otherwise, where B is the beta function.

Thus, the distribution in the previous example is $\mathsf{Beta}(n+1, m+1)$.

60 Functions of Random Variables

Let X be a continuous random variable, and let $W := g(X)$, where g is a real-valued function. Then to find the density of W, we find its cumulative distribution first: $F_W(t) = \Pr(g(X) \le t)$. If g is invertible, then we have

$$F_W(t) = \Pr(g(X) \le t) = \Pr(X \le g^{-1}(t)) = F_X(g^{-1}(t)).$$

[Of course, if g were monotonically decreasing, we will flip the sign $\Pr(X \ge g^{-1}(t))$.] Hence

$$f_W(t) = \frac{d}{dt} F_W(t) = f_X(g^{-1}(t)) \cdot (g^{-1}(t))'.$$

We now generalize this procedure to joint distributions.

Theorem 60.1. *Let X_1, X_2 be jointly continuous, and let $Y_1 := g_1(X_1, X_2)$ and $Y_2 := g_2(X_1, X_2)$, where the g_i have continuous first partial derivatives. Suppose the system $y_1 = g_1(x_1, x_2), y_2 = g_2(x_1, x_2)$ has a unique solution $x_1 = h_1(y_1, y_2), x_2 = h_2(y_1, y_2)$, and for all pairs (x_1, x_2), the Jacobian*

$$J(x_1, x_2) = \frac{\partial(g_1, g_2)}{\partial(x_1, x_2)} = \begin{vmatrix} (g_1)_{x_1} & (g_1)_{x_2} \\ (g_2)_{x_1} & (g_2)_{x_2} \end{vmatrix}$$

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

Proof Sketch. We do what we did for a single variable. First, we compute the joint cumulative distribution

$$F_{Y_1,Y_2}(y_1,y_2) = \Pr(Y_1 \le y_1, Y_2 \le y_2) = \Pr(g(X_1,X_2) \le y_1, g(X_1,X_2) \le y_2).$$

Now, let $R \subseteq \mathbb{R}^2$ be the set of all pairs (x_1,x_2) such that $g_1(x_1,x_2) \le y_1$ and $g_2(x_1,x_2) \le y_2$. But this implies

$$F_{Y_1,Y_2}(y_1,y_2) = \iint_R f_{X_1,X_2}(x_1,x_2)\,dA \implies \frac{\partial^2}{\partial y_1 \partial y_2} F_{Y_1,Y_2} = \frac{\partial^2}{\partial y_1 \partial y_2} \iint_R f_{X_1,X_2}\,dA,$$

so that $f_{Y_1,Y_2}(y_1,y_2) = f_{X_1,X_2}(x_1,x_2) \cdot |J(x_1,x_2)|^{-1}$. Here, this last step requires justification using multivariable analysis, but the intuitive explanation of the Jacobian (i.e., it shows up in a change of variables) should explain its presence here. □

Example 60.2. Let X_1, X_2 be jointly continuous, and let $Y_1 := X_1 + X_2$, $Y_2 := X_1 - X_2$. Then we can find f_{Y_1,Y_2} via Theorem 60.1: we have the transformation $y_1 = x_1 + x_2$, $y_2 = x_1 - x_2$, which has the Jacobian

$$J(x_1,x_2) = \begin{vmatrix} \partial y_1/\partial x_1 & \partial y_1/\partial x_2 \\ \partial y_2/\partial x_1 & \partial y_2/\partial x_2 \end{vmatrix} = \begin{vmatrix} 1 & 1 \\ 1 & -1 \end{vmatrix} = -2.$$

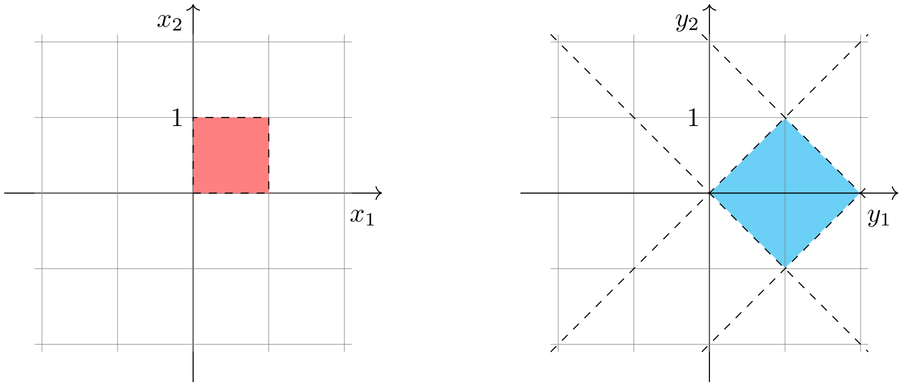
Now, solving for x_1, x_2 in terms of y_1, y_2 gives $x_1 = \frac{1}{2}(y_1 + y_2)$ and $x_2 = \frac{1}{2}(y_1 - y_2)$. Hence

$$f_{Y_1,Y_2}(y_1,y_2) = f_{X_1,X_2}(x_1,x_2) \cdot |J(x_1,x_2)|^{-1} = \boxed{\frac{1}{2} f_{X_1,X_2}\left(\frac{y_1+y_2}{2}, \frac{y_1-y_2}{2}\right)}.$$

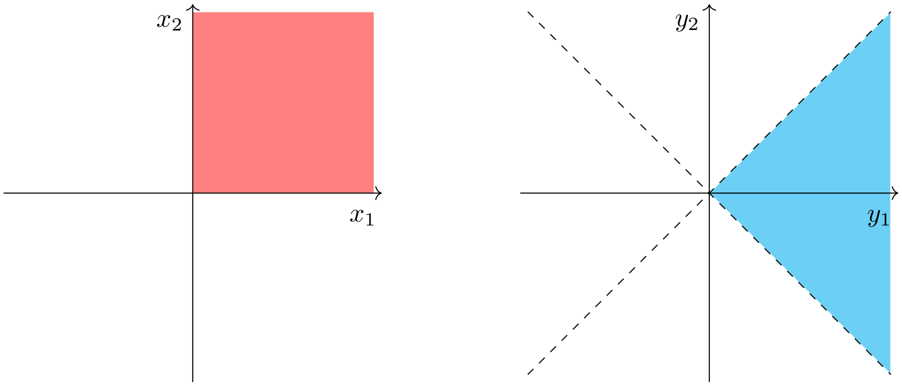
Example 60.3. Now, if $X_1, X_2 \sim \mathcal{U}(0,1)$ and are independent, and we take Y_1, Y_2 as in the preceding example. Then $f_{X_1,X_2}(x_1,x_2) = 1$ when $(x_1,x_2) \in (0,1)^2$ and 0 otherwise. Hence $f_{Y_1,Y_2}(y_1,y_2) = \frac{1}{2}$ whenever $(x_1,x_2) \in (0,1)^2$, i.e., $(y_1+y_2, y_1-y_2) \in (0,2)^2$, and 0 otherwise:

$$f_{Y_1,Y_2}(y_1,y_2) = \begin{cases} \frac{1}{2} & 0 < y_1 \pm y_2 < 2 \\ 0 & \text{otherwise.} \end{cases}$$

Geometrically, we have the following transformation which takes us from x_i-coordinates to y_i-coordinates:



Example 60.4. Similarly, take $X_1 \sim \mathsf{Exp}(\lambda_1)$ and $X_2 \sim \mathsf{Exp}(\lambda_2)$ to be independent, and take the same Y_1, Y_2. Then $f_{X_1,X_2}(x_1,x_2) = \lambda_1\lambda_2 e^{-\lambda_1 x_1}e^{-\lambda_2 x_2}$ whenever $(x_1,x_2) \in (0,\infty)^2$, so that $f_{Y_1,Y_2}(y_1,y_2) = \lambda_1\lambda_2 e^{-\lambda_1\left(\frac{y_1+y_2}{2}\right)}e^{-\lambda_2\left(\frac{y_1-y_2}{2}\right)}$ whenever we have $\frac{1}{2}(y_1 \pm y_2) \geq 0 \iff y_1 \pm y_2 = 0$. The transformation in the plane looks like



Example 60.5. Let $X \sim \mathsf{Gam}(a,\lambda)$ and $Y \sim \mathsf{Gam}(b,\lambda)$ be independent. It can be shown that $U := X+Y \sim \mathsf{Gam}(a+b,\lambda)$ and $V = \frac{X}{X+Y} \sim \mathsf{Beta}(a,b)$, and suppose we want to find the joint distribution $f_{U,V}(u,v)$. First, by independence,

$$f_{X,Y}(x,y) = f_X(x)f_Y(y) = \frac{\lambda e^{-ax}(\lambda x)^{a-1}}{\Gamma(a)} \cdot \frac{\lambda e^{-by}(xy)^{b-1}}{\Gamma(b)} = \frac{\lambda^{a+b}}{\Gamma(a)\Gamma(b)}e^{-\lambda(x+y)}x^{a-1}y^{b-1},$$

where $x,y > 0$. Now

$$J(x,y) = \frac{\partial(u,v)}{\partial(x,y)} = \begin{vmatrix} u_x & u_y \\ v_x & v_y \end{vmatrix} = \begin{vmatrix} 1 & 1 \\ \frac{y}{(x+y)^2} & \frac{-x}{(x+y)^2} \end{vmatrix} = \frac{-\cancel{(x+y)}}{(x+y)^{\cancel{2}}} = -\frac{1}{x+y} = -\frac{1}{u}.$$

We notice that $V = X/U$, so $X = UV$. From here, it follows that $Y = U - X = U - UV$. Note that $x,y > 0$,and $0 < x \leq x+y$, so $0 < v < 1$. Certainly, $u = x+y > 0+0 = 0$, so our (u,v)-domain is $(u,v) \in (0,\infty) \times (0,1)$. Hence

$$f_{U,V}(u,v) = f_{X,Y}(uv, u(1-v))\left|-\frac{1}{u}\right|^{-1} = \frac{\lambda e^{-\lambda u}(\lambda u)^{a+b-1}}{\Gamma(a)\Gamma(b)} \cdot v^{a-1}(1-v)^{b-1}.$$

Incidentally, we have proven that U and V are also independent.

68 Properties of Expectation

Recall that if X is a random variable, then $\mathbb{E}[x] = \sum_x xp(x)$ or $\int_{\mathbb{R}} xf(x)\,dx$, depending on whether X is discrete or continuous. Similarly, $\mathbb{E}[X^2] = \sum_x x^2p(x)$ or $\int_{\mathbb{R}} x^2f(x)\,dx$. We now generalize this to multiple jointly continuous random variables.

Proposition 68.1. *Let X, Y be random variables, and g be a real-valued function of two variables. If X, Y are both discrete, then $\mathbb{E}[g(X,Y)] = \sum_x\sum_y g(x,y)p(x,y)$. Similarly, if X, Y are both continuous, then $\mathbb{E}[g(X,Y)] = \iint_{\mathbb{R}^2} g(x,y)f(x,y)\,dA$.*

Proof. Exercise. □

Example 68.2. Consider a road of length $L = 1$, and say an accident occurs on the road at milepost X, with an ambulance at milepost Y, where X, Y are independent with $X, Y \sim \mathcal{U}(0,1)$. Find the expected distance $|X - Y|$.

Solution. We compute $f_{X,Y}(x,y) = 1$ when $(x,y) \in (0,1)^2$, and 0 otherwise. Then, we find

$$\mathbb{E}[|X-Y|] = \int_0^1 \int_0^1 |x-y| \cdot 1 \, dy \, dx = \int_0^1 \int_0^x (x-y) \, dy \, dx + \int_0^1 \int_x^1 (y-x) \, dy \, dx.$$

Notice that these integrals are actually equal in value, by symmetry. Hence, we compute

$$\frac{1}{2}\mathbb{E}[|X-Y|] = \int_0^1 \int_0^x (x-y) \, dy \, dx = \int_0^1 xy - \frac{y^2}{2}\bigg|_0^1 dx = \int_0^1 \frac{x^2}{2} \, dx = \frac{1}{6},$$

so the expected value is $\boxed{\frac{1}{3}}$. •

Corollary 68.3. *If* $\mathbb{E}[X]$ *and* $\mathbb{E}[Y]$ *are both finite, then* $\mathbb{E}[X+Y] = \mathbb{E}[X] + \mathbb{E}[Y]$.

We remark that this works even if X, Y are *dependent.*

Proof. We prove the version of this theorem where both X, Y are continuous; the discrete case is similar. We compute

$$\begin{aligned}
\mathbb{E}[X+Y] &= \iint_{\mathbb{R}^2} (x+y) f(x,y) \, dy \, dx \\
&= \iint_{\mathbb{R}^2} x f(x,y) + y f(x,y) \, dy \, dx \\
&= \iint_{\mathbb{R}^2} x f(x,y) \, dy \, dx + \iint_{\mathbb{R}^2} y f(x,y) \, dx \, dy \\
&= \iint_{\mathbb{R}} x \left(\int_{\mathbb{R}} f(x,y) \, dy \right) dx + \int_{\mathbb{R}} y \left(\int_{\mathbb{R}} f(x,y) \, dx \right) dy \\
&= \int_{\mathbb{R}} x f_X(x) \, dx + \int_{\mathbb{R}} y f_Y(y) \, dy = \mathbb{E}[X] + \mathbb{E}[Y],
\end{aligned}$$

which completes the proof. □

Corollary 68.4. *If* $X \geq Y$ *for all outcomes, then* $\mathbb{E}[X] \geq \mathbb{E}[Y]$.

Proof. We have $X - Y \geq 0$, so $\mathbb{E}[X - Y] \geq 0$, but now $\mathbb{E}[X] - \mathbb{E}[Y] \geq 0$. □

Let us view some applications of this idea.

Definition 68.5. A *sample* from a distribution function F is a set of random variables $X_1, X_2, \ldots, X_n$, independent and identically distributed with cumulative distribution $F_{X_i} = F$. Given a sample, we also define the *sample mean* $\overline{X} := \frac{1}{n}\sum_{i=1}^n X_i$.

We now demonstrate that the sample mean is appropriately named. If $\mathbb{E}[X_i] = \mu$ for all X_i, we see

$$\mathbb{E}[\bar{X}] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n X_i\right] = \frac{1}{n}\mathbb{E}\left[\sum_{i=1}^n X_i\right] = \frac{1}{n}\sum_{i=1}^n \mathbb{E}[X_{]} = \frac{1}{\cancel{n}}\mu\cancel{n} = \mu.$$

Here is another important application.

Definition 68.6. The *indicator* of an event A is the random variable $\mathbb{1}_A$, which equals 1 when A occurs, and 0 otherwise.

Certainly, $\mathbb{E}[\mathbb{1}_A] = 1 \cdot \Pr(\mathbb{1}_A = 1) + \cancel{0 \cdot \Pr(\mathbb{1}_A = 0)} = \Pr(A)$.

Example 68.7. Let $A_1, A_2, \dots, A_n$ be a list of events, and define $X := \sum_{i=1}^n \mathbb{1}_{A_i}$ and $Y := \mathbb{1}_{X \geq 1}$. Certainly, $X \geq Y$ for any outcome, so $\mathbb{E}[X] \geq \mathbb{E}[Y]$.Now

$$\mathbb{E}[X] = \mathbb{E}\left[\sum_{i=1}^n \mathbb{1}_{A_i}\right] = \sum_{i=1}^n \mathbb{E}\left[\mathbb{1}_{A_i}\right] = \sum_{i=1}^n \Pr(A_i),$$

but now $\mathbb{E}[Y] = \Pr\left(\bigcup_{i=1}^n A_i\right)$. This proves *finite subadditivity*: given any finite list of events A_i, we have

$$\Pr\left(\bigcup_{i=1}^n A_i\right) \leq \sum_{i=1}^n \Pr(A_i).$$

Furthermore, summing indicators gives easy proofs of expected values of well-known distributions:

Example 68.8. Suppose $X \sim \mathsf{Bin}(n, p)$. Take the events A_i to be where the ith trial is a success, so that $X = \sum_{i=1}^n \mathbb{1}_{A_i}$. Now

$$\mathbb{E}[X] = \mathbb{E}\left[\sum_{i=1}^n \mathbb{1}_{A_i}\right] = \sum_{i=1}^n \mathbb{E}[\mathbb{1}_{A_i}] = \sum_{i=1}^n p = np.$$

Example 68.9. X has a *negative binomial distribution*, written $X \sim \mathsf{NB}(r, p)$, where X represents the number of trials to get r successes, with probability p. Without knowing $p_X(n)$, we can still derive the expectation of X. Let X_i, $1 \leq i \leq r$, to be the number of trials between the $(i-1)$st success and the ith success. Then $X = X_1 + \cdots + X_r$, and we have $X_i \sim \mathsf{Geo}(p)$. Hence $\mathbb{E}[X] = \sum_{i=1}^r \mathbb{E}[X_i] = \frac{1}{p} \cdot r = \frac{r}{p}$.

Example 68.10. X has a *hypergeometric distribution*, written $X \sim \mathsf{Hyp}(n, N, m)$, where X represents the number of green balls obtained when selecting, without replacement, $n < N$ balls from an urn of N balls, and $m < N$ is the total number of green balls. To find $\mathbb{E}[X]$, we enumerate the green balls $1, 2, \dots, m$, and we let X_i be the indicator for choosing the ith green ball. Then $X = X_1 + \cdots + X_m$, and

$$\mathbb{E}[X_i] = \Pr(\text{green ball } i \text{ is chosen}) = \frac{\binom{1}{1}\binom{N-1}{n-1}}{\binom{N}{n}} = \frac{n}{N}.$$

Hence $\mathbb{E}[X] = \displaystyle\sum_{i=1}^m \mathbb{E}[X_i] = \frac{mn}{N}$.

70 Indicator Sums: Examples

Example 70.1. Say that N people throw their hats in the middle of the room, and then each person takes a hat at random. Let X be the number of people who select their own hat. What is $\mathbb{E}[X]$?

Solution. Define the variables X_i to be the indicator that person i chooses their own hat, for $1 \le i \le N$. Then we see that $X = \sum_{i=1}^{N} X_i$, and $\mathbb{E}[X_i] = \Pr(i\text{th person picks own hat}) = \frac{1}{N}$. Hence $\mathbb{E}[X] = \sum_{i=1}^{n} \mathbb{E}[X_i] = \frac{N}{N} = \boxed{1}$. •

Example 70.2. There are N types of coupons, with an unlimited amount of each. Each time a coupon is collected, it is equally likely to be any one of the N types. Let X be the number of coupons collected to collect all types. What is $\mathbb{E}[X]$?

Solution. Define the variables X_i to be the numbers of coupons collected between having $(i-1)$ types of coupons, until collecting the ith type, for $1 \le i \le N$. Then $X = X_1 + \cdots + X_N$, and we may verify $X_i \sim \mathsf{Geo}\left(\frac{N-i+1}{N}\right)$, for all $2 \le i \le N$. Then

$$\begin{aligned}\mathbb{E}[X] = \sum_{i=1}^{N} \mathbb{E}[X_i] &= \sum_{i=1}^{N} \frac{N}{N-i+1} \\ &= N \sum_{i=1}^{N} \frac{1}{N-i+1} = \boxed{N\left(1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{N-1}\right)}.\end{aligned}$$

•

Example 70.3. Consider a random walk on $\mathbb{R}^2$ starting at $(0,0)$, and for each step i, we pick an angle $\Theta_i \sim \mathcal{U}(0, 2\pi)$, chosen all independently, then walk a unit length in the direction Θ_i. What is $\mathbb{E}[D_n^2]$, where D_n is the distance from the origin after n steps?

Solution. Define (X_i, Y_i) to be the displacement during the ith step, so that $X_i = \cos\Theta_i$ and $Y_i = \sin\Theta_i$. Hence

$$D_n^2 = \left(\sum_{i=1}^{n} X_i\right)^2 + \left(\sum_{i=1}^{n} Y_i\right)^2 = \sum_{i=1}^{n} (X_i^2 + Y_i^2) + \sum_{i \ne j} (X_i X_j + Y_i Y_j).$$

Now $X_i^2 + Y_i^2 = \cos^2\Theta_i + \sin^2\Theta_i = 1$, and we may compute $\mathbb{E}[X_i] = \mathbb{E}[\cos\Theta_i] = \int_0^{2\pi} \cos\theta \cdot \frac{1}{2\pi}\, d\theta = 0$, and similarly $\mathbb{E}[Y_i] = 0$. By independence, when $i \ne j$ we have[2] $\mathbb{E}[X_i X_j] = \mathbb{E}[X_i]\mathbb{E}[X_j] = 0$, and similarly $\mathbb{E}[Y_i Y_j] = 0$. Hence $\mathbb{E}[D_n^2] = \boxed{n}$. •

74 Infinite Sums and Moments

In general, $\mathbb{E}\left[\sum_{i=1}^{\infty} X_i\right] \ne \sum_{i=1}^{\infty} \mathbb{E}[X_i]$, as doing this requires an interchange of limits:

$$\mathbb{E}\left[\lim_{n\to\infty} \sum_{i=1}^{n} X_i\right] \overset{?}{=} \lim_{n\to\infty} \mathbb{E}\left[\sum_{i=1}^{n} X_i\right].$$

To do so, the sum $\sum_{i=1}^{n} \mathbb{E}[X_i]$ must be absolutely convergent, which we can guarentee in two special cases:

[2] We will prove this later.

Proposition 74.1. *Let $X_1, X_2, \dots$ be a sequence of random variables, and suppose $\mathbb{E}[X_i]$ is finite for all $i \in \mathbb{Z}^+$. Then*

1. *If $X_i \geq 0$ for all outcomes, then $\sum_{i=1}^{\infty} \mathbb{E}[X_i]$ is absolutely convergent.*
2. *If $\sum_{i=1}^{\infty} \mathbb{E}[|X_i|]$ is finite, then $\sum_{i=1}^{\infty} \mathbb{E}[X_i]$ is absolutely convergent.*

Proposition 74.2. *Let X be a nonnegative, integer-valued random variable. Then $\mathbb{E}[X] = \sum_{i=1}^{\infty} \Pr(X \geq i)$.*

Proof. Notice that $\Pr(X \geq i) = \mathbb{E}[\mathbb{1}_{X \geq i}]$. Now, write

$$\sum_{i=1}^{\infty} \mathbb{1}_{X \geq i} = \sum_{i=1}^{X} \mathbb{1}_{X \geq i} + \sum_{i=X+1}^{\infty} \mathbb{1}_{X \geq i} = \sum_{i=1}^{X} 1 + \sum_{i=X+1}^{\infty} 0 = X.$$

Since $\mathbb{E}[\mathbb{1}_{X \geq i}] \geq 0$, Proposition 74.1 tells us that $\sum_{i=1}^{\infty} \mathbb{E}[\mathbb{1}_{X \geq i}]$ is absolutely convergent, so

$$\mathbb{E}[X] = \mathbb{E}\left[\sum_{i=1}^{\infty} \mathbb{1}_{X \geq i}\right] = \sum_{i=1}^{\infty} \mathbb{E}[1_{X \geq i}] = \sum_{i=1}^{\infty} \Pr(X \geq i),$$

as claimed. □

Moments on the Number of Events that Occur

Recall that if $A_1, A_2, \dots, A_n$ are events, and if X is the number of events among the A_i that occur, then $\mathbb{E}[X] = \sum_{i=1}^{n} \Pr(A_i)$. We also recall the following.

Definition 74.3. Let X be a random variable, and $n \in \mathbb{Z}^+$. The *nth moment* of X is $\mathbb{E}[X^n]$.

Now, we are concerned about pairs of events. It is not too hard to see $\mathbb{1}_{A_i}\mathbb{1}_{A_j} = \mathbb{1}_{A_i \cap A_j}$, so that

$$\mathbb{E}[\mathbb{1}_{A_i}\mathbb{1}_{A_j}] = \Pr(A_i \cap A_j).$$

If X is the number of events occurring, then $\binom{X}{2}$ is the number of pairs that occur. Now, $\binom{X}{2} = \frac{1}{2}(X^2 - X)$, so that

$$\mathbb{E}\left[\binom{X}{2}\right] = \frac{1}{2}\mathbb{E}[X^2] - \frac{1}{2}\mathbb{E}[X].$$

Alternatively, $\binom{X}{2} = \sum_{i<j} \mathbb{1}_{A_i}\mathbb{1}_{A_j}$, so that $\mathbb{E}\left[\binom{X}{2}\right] = \mathbb{E}[\mathbb{1}_{A_i}\mathbb{1}_{A_j}] = \sum_{i<j} \Pr(A_i \cap A_j)$. This has many useful applications.

Example 74.4. Let $X \sim \mathsf{Bin}(n,p)$. As before, let A_i be the event where the ith trial is a success (for all $1 \leq i \leq n$), so that $X_i := \mathbb{1}_{A_i} \sim \mathsf{Ber}(p)$. Now, when $i < j$, we have (by independence)

$$\mathbb{E}[X_iX_j] = \Pr(A_i \cap A_j) = \Pr(A_i)\Pr(A_j) = p \cdot p = p^2.$$

We also compute $\mathbb{E}\left[\binom{X}{2}\right] = \mathbb{E}\left[\frac{1}{2}(X^2 - X)\right] = \frac{1}{2}\mathbb{E}[X^2] - \frac{np}{2}$, but we also know

$$\mathbb{E}\left[\binom{X}{2}\right] = \sum_{i<j} \Pr(A_i \cap A_j) = \sum_{i<j} p^2 = \binom{n}{2}p^2 = \frac{n(n-1)p}{2}.$$

Solving for $\mathbb{E}[X^2]$ yields $\mathbb{E}[X^2] = n^2p^2 - np^2 + np$, so $\mathrm{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = np(1-p)$, thus giving us another way to derive the variance for a binomial random variable.

78 Higher Moments

We now generalize the example in the preceding section. If we have events $A_1, A_2, \dots, A_n$, and we have X be the number of events that occur, then

$$\mathbb{E}\left[\binom{X}{k}\right] = \sum\sum_{i_1<i_2<\cdots<i_k}\cdots\sum \Pr(A_{i_1} \cap A_{i_2} \cap \cdots \cap A_{i_k}),$$

where $1 \le i \le n$. Using this process, we may recursively determine the kth moment from prior moments.

Example 78.1. Let $X \sim \mathsf{Bin}(n,p)$, and let X_i be the indicator for which the ith trial is a success. Then by independence, $\mathbb{E}(X_{i_1}X_{i_2}\cdots X_{i_k}) = \prod_{j=1}^{k}\mathbb{E}[X_{i_j}] = p^k$, so

$$\mathbb{E}\left[\binom{X}{k}\right] = \sum\sum_{i_1<i_2<\cdots<i_k}\cdots\sum p^k = \binom{n}{k}p^k.$$

Taking $k = 3$, we have $\binom{X}{3} = \frac{1}{6}X(X-1)(X-2)$, so

$$\begin{aligned}\mathbb{E}\left[\binom{X}{3}\right] &= \frac{X(X-1)(X-2)}{6} = \frac{n(n-1)(n-2)}{6}p^3 \\ \implies \mathbb{E}[X^3] - 3\mathbb{E}[X^2] + 2\mathbb{E}[X] &= n(n-1)(n-2)p^3.\end{aligned}$$

Since $\mathbb{E}[X] = np$ and we know $\mathbb{E}[X^2] = n^2p^2 - np^2 + np$, the reader can solve for $\mathbb{E}[X^3]$ at this point.

Example 78.2. Let $X \sim \mathsf{Hyp}(n, N, m)$, and let A_i be the event that the ith ball chosen is green. Then we know that $\mathbb{E}[\mathbb{1}_{A_i}] = \Pr(A_i) = m/N$, and $\mathbb{E}[X] = \mathbb{E}[\mathbb{1}_{A_1} + \cdots + \mathbb{1}_{A_n}] = mn/N$. Now, if $i < j$, we have

$$\mathbb{E}[\mathbb{1}_{A_i}\mathbb{1}_{A_j}] = \Pr(A_i \cap A_j) = \frac{\binom{m}{2}}{\binom{N}{2}} = \frac{m}{n}\cdot\frac{m-1}{n-1} = \Pr(A_i)\Pr(A_j|A_i).$$

Hence

$$\mathbb{E}\left[\binom{X}{2}\right] = \sum_{i<j}\mathbb{E}[\mathbb{1}_{A_i}\mathbb{1}_{A_j}] = \sum_{i<j}\frac{m(m-1)}{N(N-1)} = \binom{n}{2}\frac{m(m-1)}{N(N-1)} = \frac{mn(m-1)(n-1)}{2N(N-1)},$$

so that

$$\begin{aligned}\frac{1}{2}\mathbb{E}[X(X-1)] = \frac{mn(m-1)(n-1)}{2N(N-1)} &\implies \mathbb{E}[X^2] - \mathbb{E}[X] = \frac{mn(m-1)(n-1)}{N(N-1)} \\ &\implies \mathbb{E}[X^2] = \frac{mn}{N} + \frac{mn(m-1)(n-1)}{N(N-1)}.\end{aligned}$$

From here, a computation shows that $\operatorname{Var}(X) = \dfrac{mn}{N}\left(1 - \dfrac{mn}{N} + \dfrac{(m-1)(n-1)}{N-1}\right)$.

80 Covariance

First, we state a preliminary proposition.

Proposition 80.1. *If X and Y are independent and g, h are real-valued functions of one variable, then* $\mathbb{E}[g(X)h(Y)] = \mathbb{E}[g(X)\mathbb{E}[h(y)]$.

Proof. We prove the continuous case; the discrete one is similar. By independence, write

$$\begin{aligned}\mathbb{E}[g(X)h(Y)] &= \iint_{\mathbb{R}^2} g(x)h(y)f(x,y)\,dA \\ &= \int_{\mathbb{R}}\int_{\mathbb{R}} g(x)h(y)f_X(x)f_Y(y)\,dy\,dx \\ &= \int_{\mathbb{R}} g(x)f_X(x)\,dx \cdot \int_{\mathbb{R}} h(y)f_Y(y)\,dy \\ &= \mathbb{E}[g(X)]\mathbb{E}[h(Y)],\end{aligned}$$

which completes the proof. □

This gives us the following very important corollary.

Corollary 80.2. *If X and Y are independent, then* $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$.

Now, we generalize the idea of variance, to multiple variables. Recall that $\mathrm{Var}(X) := \mathbb{E}[(X - \mathbb{E}[X])^2]$, which measured the "spread" of a random variable. Given two variables, we want to see how they, roughly, "vary with each other."

Definition 80.3. Let X, Y be random variables. We define the *covariance* of X and Y by

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

This is a clear generalization of variance, as $\mathrm{Cov}(X,X) = \mathbb{E}[(X - \mathbb{E}[X])^2] = \mathrm{Var}(X)$. We also have a following, simpler formula to calculate covariance:

Proposition 80.4. *For random variables X, Y, we have* $\mathrm{Cov}(X,Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$.

Proof. The proof proceeds by definition:

$$\begin{aligned}\mathrm{Cov}(X,Y) &= \mathbb{E}\big[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])\big] \\ &= \mathbb{E}\big[XY - X\mathbb{E}[Y] - Y\mathbb{E}[X] + \mathbb{E}[X]\mathbb{E}[Y]\big] \\ &= \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] - \cancel{\mathbb{E}[Y]\mathbb{E}[X]} + \cancel{\mathbb{E}[X]\mathbb{E}[Y]},\end{aligned}$$

from which the result follows. □

By Corollary 80.2, it follows that if X and Y are independent, then $\mathrm{Cov}(X,Y) = 0$. Hence, the covariance is only "interesting" when X and Y are dependent.

Example 80.5. We give an example to show that when $\mathrm{Cov}(X,Y) = 0$, it is not necessarily true that X and Y are independent. Let $X \sim \mathcal{U}(\{-1,0,1\})$ and $Y = \mathbb{1}_{X=0}$. Then

$$XY = \left\{\begin{matrix} X \cdot 1 & X = 0 \\ X \cdot 0 & X \neq 0 \end{matrix}\right\} = 0,$$

so $\mathbb{E}[XY] = 0$. Also, $\mathbb{E}[X] = \frac{1}{3}(-1+0+1) = 0$ so that $\mathrm{Cov}(X,Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] = 0 - 0\mathbb{E}[Y] = 0$. But clearly X and Y are *dependent*:

$$\frac{1}{2} = \Pr(X = 1 \mid Y = 0) \neq \Pr(X = 1) = \frac{1}{3}.$$

Proposition 80.6. *The covariance satisfies the following properties:*

1. *Symmetry:* $\mathrm{Cov}(X,Y)=\mathrm{Cov}(Y,X)$.
2. *Bilinearity:* $\mathrm{Cov}(X,aY)=\mathrm{Cov}(aX,Y)=a\,\mathrm{Cov}(X,Y)$ *for all* $a\in\mathbb{R}$*, and for sums of random variables, we have*

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

Proof. Part (1) is clear, so we prove the second part of (2). We compute

$$\begin{aligned}\mathrm{Cov}\left(\sum_{i=1}^n X_i,\sum_{j=1}^m Y_j\right)&=\mathbb{E}\left[\sum_{i=1}^n X_i\sum_{j=1}^m Y_j\right]-\mathbb{E}\left[\sum_{i=1}^n X_i\right]\mathbb{E}\left[\sum_{j=1}^m Y_j\right]\\&=\sum_{i=1}^n\sum_{j=1}^m\big[\mathbb{E}[X_iY_j]-\mathbb{E}[X_i]\mathbb{E}[Y_j]\big]\\&=\sum_{i=1}^n\sum_{j=1}^m\mathrm{Cov}(X_i,Y_j),\end{aligned}$$

which completes the proof. □

From bilinearity, we get several important results.

Corollary 80.7. *Let X be a random variable and $a\in\mathbb{R}$. Then* $\mathrm{Var}(aX)=a^2\,\mathrm{Var}(X)$.

Proof. Write $\mathrm{Var}(aX)=\mathrm{Cov}(aX,aX)=a\,\mathrm{Cov}(X,aX)=a^2\,\mathrm{Cov}(X,X)=a^2\,\mathrm{Var}(X)$. □

Corollary 80.8. *Let $X_1,X_2,\dots,X_n$ be random variables. Then*

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

Proof. By bilinearity, write

$$\begin{aligned}\mathrm{Var}\left(\sum_{i=1}^n X_i\right)&=\mathrm{Cov}\left(\sum_{i=1}^n X_i,\sum_{j=1}^n X_j\right)\\&=\sum_{i=1}^n\sum_{j=1}^n\mathrm{Cov}(X_i,X_j)\\&=\sum_{i=1}^n\mathrm{Cov}(X_i,X_i)+\sum_{i<j}\mathrm{Cov}(X_i,X_j)+\sum_{i>j}\mathrm{Cov}(X_i,X_j)\\&=\sum_{i=1}^n\mathrm{Var}(X_i)+2\sum_{i<j}\mathrm{Cov}(X_i,X_j),\end{aligned}$$

where the last step follows from symmetry of the covariance. □

In view of a past discussion, we may also establish the following.

Corollary 80.9. *If $X_1,X_2,\dots,X_n$ are pairwise independent, then* $\mathrm{Var}(X_1+\cdots+X_n)=\mathrm{Var}(X_1)+\cdots+\mathrm{Var}(X_n)$.

Some Statistical Applications

Let $X_1, X_2, \dots, X_n$ be independent and identically distributed random variables, with identical $\mathbb{E}[X_i] = \mu$ and $\mathrm{Var}(X) = \sigma^2$. Recall that the sample mean is the random variable $\overline{X} = \frac{1}{n}(X_1 + \cdots + X_n)$, with $\mathbb{E}[\overline{X}] = \mu$. We also define the following:

Definition 80.10. Given $X_1, \dots, X_n$ as above, we define the *sample deviation* of X_i by $X_i - \overline{X}$. From here, we also define the *sample variance*

$$S^2 := \sum_{i=1}^{n} \frac{(X_i - \overline{X})^2}{n-1}.$$

The term $(n-1)$ in the denominator of S^2 above is *Bessel's correction*, which often goes mysteriously unexplained in statistics texts. We now explain why it is there — and it is because of this very nice property.

Theorem 80.11. *We have* $\mathbb{E}[S^2] = \sigma^2$.

Proof. Because we need it in the computation later, we compute the variance of the sample mean $\overline{X}$ first:

$$\mathrm{Var}(\overline{X}) = \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\,\mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n},$$

where we have applied the independence of the X_i.

Now, we do some manipulations to the random variable S^2 to make it easier to work with. First, notice that $(n-1)S^2 = \sum_{i=1}^{n}(X_i - \overline{X})^2$, and we may write

$$\begin{aligned}
(n-1)S^2 = \sum_{i=1}^{n}(X_i - \overline{X})^2 &= \sum_{i=1}^{n} \left[(X_i - \mu) + (\mu - \overline{X})\right]^2 \\
&= \sum_{i=1}^{n} \left[(X_i - \mu)^2 + 2(X_i - \mu)(\mu - \overline{X}) + (\mu - \overline{X})^2\right] \\
&= \sum_{i=1}^{n}(X_i - \mu)^2 + 2(\mu - \overline{X})\sum_{i=1}^{n}(X_i - \mu) + (\mu - \overline{X})\sum_{i=1}^{n} 1.
\end{aligned}$$

Now, $\sum_{i=1}^{n}(X_i - \mu) = \sum_{i=1}^{n} X_i - \sum_{i=1}^{n} \mu = n\overline{X} - n\mu = -n(\mu - \overline{X})$, so that

$$\begin{aligned}
(n-1)S^2 &= \sum_{i=1}^{n}(X_i - \mu)^2 + 2(\mu - \overline{X})\sum_{i=1}^{n}(X_i - \mu) + (\mu - \overline{X})\sum_{i=1}^{n} 1 \\
&= \sum_{i=1}^{n}(X_i - \mu)^2 - 2n(\mu - \overline{X})^2 + n(\mu - \overline{X})^2 \\
&= \sum_{i=1}^{n} \left[(X_i - \mu)^2\right] - n(\overline{X} - \mu)^2.
\end{aligned}$$

Now, we are in a position to take the expectation:

$$\begin{aligned}(n-1)\mathbb{E}[S^2] &= \mathbb{E}\left[\sum_{i=1}^{n}(X_i-\mu)^2\right] - n\mathbb{E}[(\overline{X}-\mu)^2] \\ &= \sum_{i=1}^{n}\mathbb{E}[(X_i-\mu)^2] - n\mathbb{E}[(\overline{X}-\mu)^2] \\ &= \sum_{i=1}^{n}\mathrm{Var}(X_i) - n\,\mathrm{Var}(\overline{X}) \\ &= n\sigma^2 - \cancel{n}\cdot\frac{\sigma^2}{\cancel{n}} = (n-1)\sigma^2.\end{aligned}$$

Cancelling the $n-1$ shows $\mathbb{E}[S^2] = \sigma^2$, as claimed. □

84 Correlation

In the preceding section, we saw many properties of covariance. Now, we see one main application of covariance, which is finding *correlation* between two random variables. First, we recall the following definition.

Definition 84.1. Let X be a random variable. We define the *standard deviation* of X, denoted σ_X, by $\sigma_X := \sqrt{\mathrm{Var}(X)}$.

Definition 84.2. Let X, Y be random variables. We define the *correlation coefficient* $\rho(X,Y)$ by

$$\rho(X,Y)) = \frac{\mathrm{Cov}(X,Y)}{\sigma_X\sigma_Y}.$$

Hence, by the definition above, covariance is essentially "un-normalized" correlation.

Proposition 84.3. *Let X, Y be random variables. Then $|\rho(X,Y)| \le 1$.*

Proof. We know $\mathrm{Var}(X) = \sigma_X^2$ and $\mathrm{Var}(Y) = \sigma_Y^2$. Since the variance is non-negative,

$$\begin{aligned}&0 \le \mathrm{Var}\left(\frac{X}{\sigma_X} + \frac{Y}{\sigma_Y}\right) = \cancelto{1}{\mathrm{Var}\left(\frac{X}{\sigma_X}\right)} + 2\,\mathrm{Cov}\left(\frac{X}{\sigma_X}, \frac{Y}{\sigma_Y}\right) + \cancelto{1}{\mathrm{Var}\left(\frac{Y}{\sigma_Y}\right)} \\ \implies\ &0 \le 2 + 2\,\mathrm{Cov}\left(\frac{X}{\sigma_X}, \frac{Y}{\sigma_Y}\right) = 2\left[1 + \frac{\mathrm{Cov}(X,Y)}{\sigma_X\sigma_Y}\right] \\ \implies\ &0 \le 1 + \rho(X,Y),\end{aligned}$$

which demonstrates $-1 \le \rho(X,Y)$. Noting that $0 \le \mathrm{Var}\left(\frac{X}{\sigma_X} - \frac{Y}{\sigma_Y}\right)$ and expanding shows the other inequality $\rho(X,Y) \le 1$, which completes the proof. □

Proposition 84.4. *Let X, Y be random variables. Then $\rho(X,Y) = \pm 1$ if and only if $Y = a + bX$, where $b = \pm\sigma_Y/\sigma_X$, taking the same signs as for ρ and b.*

Proof. ($\Longleftarrow$): This is a direct computation.

($\Longrightarrow$): Suppose $\rho(X,Y) = 1$. Then $\mathrm{Var}\left(\frac{X}{\sigma_X} - \frac{Y}{\sigma_Y}\right) = 2 - 2\rho(X,Y) = 2 - 2 = 0$, which forces $\frac{X}{\sigma_X} - \frac{Y}{\sigma_Y} =: c$ to be constant, so rearranging yields the formula. This is similar when $\rho(X,Y) = -1$. □

We see that the correlation coefficient does what it is supposed to, at least statistically. In general, statisticians say that X, Y are *correlated* if $|\rho(X,Y)| \geq 0.95$, and *uncorrelated* if $|\rho(X,Y)| \leq 0.05$. We will say that X and Y are uncorrelated if $\rho(X,Y)$ is identically 0, however.

Example 84.5. Let A, B be events with indicators $\mathbb{1}_A$ and $\mathbb{1}_B$. Then $\operatorname{Cov}(\mathbb{1}_A, \mathbb{1}_B) = \mathbb{E}[\mathbb{1}_A \mathbb{1}_B] - \mathbb{E}[\mathbb{1}_A]\mathbb{E}[\mathbb{1}_B] = \Pr(A \cap B) - \Pr(A)\Pr(B) = [\Pr(A|B) - \Pr(A)]\Pr(B)$. Since $\Pr(B) \geq 0$, we observe that if $\operatorname{Cov}(\mathbb{1}_A, \mathbb{1}_B) > 0$, then $\Pr(A|B) > \Pr(A)$; similarly, if $\operatorname{Cov}(\mathbb{1}_A, \mathbb{1}_B) < 0$, then $\Pr(A|B) < \Pr(A)$. Finally, if $\operatorname{Cov}(\mathbb{1}_A, \mathbb{1}_B) = 0$, then A and B are independent. Hence, for *indicators*, $\operatorname{Cov}(\mathbb{1}_A, \mathbb{1}_B) = 0$ if and only if they are independent.

Example 84.6. Consider m independent trials, each with r possible outcomes. Take $\Pr(\text{outcome } i \text{ per trial}) = p_i$, so that $p_1 + p_2 + \cdots + p_r = 1$. Take N_i to be the number of times i occurs. Then $N_i \sim \mathsf{Bin}(m, p_i)$, and $N_1, \ldots, N_r$ are distributed multinomially. We prove that $\operatorname{Cov}(N_i, N_j) < 0$ whenever $i \neq j$.

Proof. Define $\mathbb{1}_i(k) = 1$ if trial k results in outcome i, and 0 otherwise. Then $N_i = \sum_{k=1}^m \mathbb{1}_i(k)$ and $N_j = \sum_{\ell=1}^m \mathbb{1}_j(\ell)$. By bilinearity of the covariance, observe

$$\operatorname{Cov}(N_i, N_j) = \sum_{\ell=1}^m \sum_{k=1}^m \operatorname{Cov}\left(\mathbb{1}_i(k), \mathbb{1}_j(\ell)\right) = \sum_{i=1}^m \operatorname{Cov}(\mathbb{1}_1(k), \mathbb{1}_i(\ell)) + \sum_{k \neq \ell} \operatorname{Cov}(\mathbb{1}_i(k), \mathbb{1}_j(\ell)).$$

If $k \neq \ell$, then $\mathbb{1}_i(k)$ and $\mathbb{1}_j(\ell)$ are independent, so the second sum vanishes. When $k = \ell$, observe that $\operatorname{Cov}(\mathbb{1}_i(\ell), \mathbb{1}_j(\ell)) = \cancel{\Pr(i, j \text{ both occur})} - \Pr(i)\Pr(j) < 0$, which proves the claim. □

94 Conditional Expectation

If X and Y are discrete random variables, we can find conditional expected value just as we do regular expected value:

$$\mathbb{E}[X \mid Y = y] = \sum_x x \Pr(X = x \mid Y = y) = \sum_x x p_{X|Y}(x|y).$$

Example 94.1. Let X, Y be independent, both with distribution $\mathsf{Bin}(n,p)$. We find $\mathbb{E}[X \mid X + Y = m]$. First, we calculate

$$p_{X|X+Y}(x|m) = \frac{\Pr(X = x, X + Y = m)}{\Pr(X + Y = m)}.$$

By independence, $X + Y \sim \mathsf{Bin}(2n, p)$, so

$$\begin{aligned} p_{X|X+Y}(x|m) &= \frac{\Pr(X = x, Y = m - x)}{\Pr(X + Y = m)} = \frac{\Pr(X = x)\Pr(Y = m - x)}{\Pr(X + Y = m)} \\ &= \frac{\binom{n}{x}\cancel{p^x(1-p)^{n-x}}\binom{n}{m-x}\cancel{p^{m-x}(1-p)^{n-(m-x)}}}{\binom{2n}{m}\cancel{p^m(1-p)^{2n-m}}} \\ &= \frac{\binom{n}{x}\binom{n}{m-x}}{\binom{2n}{m}}, \end{aligned}$$

hence $(X|X+Y) \sim \mathsf{Hyp}(n, 2n, m)$. Using past work about the hypergeometric random variable, we see $\mathbb{E}[X\,X + Y = m] = \boxed{m/2}$.

Likewise, if X, Y are continuous, then

$$\mathbb{E}[X \mid Y = y] = \int_{\mathbb{R}} x f_{X|Y}(x|y)\, dx = \int_{\mathbb{R}} \frac{x f(x,y)}{f_Y(y)}\, dx,$$

provided that $f_Y(y) \neq 0$.

Example 94.2. Let $f(x,y) = \frac{1}{y} e^{-x/y} e^{-y}$ for $x, y > 0$. Find $\mathbb{E}[X \mid Y = y]$.

Solution. We have

$$f_Y(y) = \int_{\mathbb{R}} f(x,y)\, dx = \frac{e^{-y}}{y} \int_0^\infty e^{-x/y}\, dx = e^{-y} e^{-x/y}\Big|_{x=0}^{\infty} = e^{-y}$$

for $y > 0$, so that $f_{X|Y}(x|y) = \frac{1}{y} e^{-x/y}$, for $x, y > 0$. Hence

$$\mathbb{E}[X \mid Y = y] = \int_0^\infty \frac{x e^{-x/y}}{y}\, dx = -(x+y)e^{-x/y}\Big|_0^\infty = \boxed{y},$$

where we have done the integral by parts. •

We remark that since conditional probability has the same properties as regular probability, conditional expectation has the same theorems as regular expected value. For example, the following things hold:

Proposition 94.3. *In the following, let g be a function of one variable. We have*

1. $\mathbb{E}[g(X) \mid Y = y] = \int_{\mathbb{R}} g(x) f_{X|Y}(x|y)\, dx$ *or* $= \sum_x g(x) p_{X|Y}(X|Y)$,

2. $\mathbb{E}\left[\sum_{i=1}^n X_i \,\middle|\, Y = y\right] = \sum_{i=1}^n \mathbb{E}[X_i \mid Y = y]$.

Also, note that $\mathbb{E}[X|Y]$ is a random variable and a function of Y. We apply this to our previous examples:

1. In Example 94.1, we observe that since $\mathbb{E}[X \mid M = m] = \frac{m}{2}$, we have $\mathbb{E}[X|M] = \frac{M}{2}$, where $M = X + Y$.

2. In Example 94.2, we have $\mathbb{E}[X \mid Y = y] = y$, so $\mathbb{E}[X|Y] = Y$.

(Notice that the expression $X|Y$ is not a random variable, but $\mathbb{E}[X|Y]$ is.)

Now, letting $g(Y) = \mathbb{E}[X|Y]$ above, observe that

$$\mathbb{E}\Big[\mathbb{E}[X|Y]\Big] = \int_{\mathbb{R}} \mathbb{E}[X \mid Y = y] f_Y(y)\, dy.$$

This is very useful, because of the following result:

Theorem 94.4 (Law of Total Expectation)**.** *If X, Y are random variables, then*

$$\mathbb{E}[X] = \mathbb{E}\Big[\mathbb{E}[X|Y]\Big].$$

We refer to [7.5.2] in the text for the proof of the discrete case.

We now see how the law of total expectation can be used to solve problems.

Example 94.5. Let X, Y be independent with $X, Y \sim \mathsf{Bin}(3, \frac{1}{4})$. Then $X+Y \sim \mathsf{Bin}(6, \frac{1}{4})$. We know that $\mathbb{E}[X] = \frac{3}{4}$, but we can also check that $\mathbb{E}[X \mid X+Y = x+y] = \frac{1}{2}(x+y)$, so

$$\mathbb{E}\Big[\mathbb{E}[X|X+Y]\Big] = \mathbb{E}\left[\frac{1}{2}(X+Y)\right] = \frac{1}{2}\mathbb{E}[X+Y] = \frac{1}{2}\cdot\frac{6}{4} = \frac{3}{4} = \mathbb{E}[X].$$

Example 94.6. A miner is trapped in a mine, in a room with 3 doors. Each time he faces these doors, he picks one of them with equal probability:

Door	Result
1	Safety in 3 hours
2	Return to the same room in 5 hours
3	Return to the same room in 7 hours.

Let X be the time for the miner to get to safety. Find $\mathbb{E}[X]$.

Solution. In the normal case, $\mathbb{E}[X]$ is very difficult to calculate directly, so we change strategies. Define D to be the door that the miner chooses at random, so $D \sim \mathcal{U}(\{1,2,3\})$. Then $\mathbb{E}[X \mid D = 1] = 3$, but we can also find

$$\mathbb{E}[X \mid D=2] = 5 + \mathbb{E}[X] \text{ and } \mathbb{E}[X \mid D=3] = 7 + \mathbb{E}[X].$$

But by the law of total expectation, we see

$$\mathbb{E}[X] = \mathbb{E}\Big[\mathbb{E}[X|D]\Big] = \frac{1}{3}\left(\mathbb{E}[X|D=1] + \mathbb{E}[X|D=2] + \mathbb{E}[X|D=3]\right)$$
$$\implies 3\mathbb{E}[X] = 3 + 5 + \mathbb{E}[X] + 7 + \mathbb{E}[X] \implies \mathbb{E}[X] = \boxed{15}.$$

Hence, the miner will get to safety in 15 hours, on average. •

Example 94.7. Let N be the number of people entering a store per day, and $X_1, X_2, \ldots, X_N$ be the amount that each customer spends. Suppose $\mathbb{E}[N] = 50$ and $\mathbb{E}[X_i] = 8$, and suppose that all the random variables we defined are independent. Find the expected value of the total money spent per day.

Solution. Intuitively, we should get $50 \cdot 8 = 400$ dollars spent per day. We prove that this is correct. The desired expectation we want is $\mathbb{E}\left[\sum_{i=1}^N X_i\right]$, which would be annoying to do, except that we can condition on N:

$$\mathbb{E}\left[\sum_{i=1}^N X_i \,\middle|\, N = n\right] = \mathbb{E}\left[\sum_{i=1}^n X_i\right] = 8n.$$

This implies $\mathbb{E}\left[\sum_{i=1}^N X_i \,\middle|\, N\right] = 8N$, so now taking expectations on both sides and using the law of total expectation, we get

$$\mathbb{E}\left[\sum_{i=1}^N X_i\right] = \mathbb{E}\left[\mathbb{E}\left[\sum_{i=1}^N X_i \middle| N\right]\right] = \mathbb{E}[8N] = 8\mathbb{E}[N] = 8 \cdot 50 = \boxed{400},$$

which matches with our intuitive answer. •

Example 94.8. Let $U_1, U_2, U_3, \ldots \sim \mathcal{U}(0,1)$, and for $x \in \mathbb{R}^+$, define

$$N(x) := \min\left\{ n : \sum_{i=1}^{n} U_i > x \right\}.$$

That is, $N(x)$ is the minimum number of U_i's needed in order to surpass x. Let $m(x) = \mathbb{E}[N(x)]$. Find a formula for $m(x)$.

Solution. This is a nightmare to do without conditioning, but our strategy here is to condition on U_1 and then try to take advantage of recursion. It is easy to see

$$\mathbb{E}[N(x) \,|\, U_1 = y] = \begin{cases} 1 & y > x \\ 1 + \mathbb{E}[N(x-y)] & y \geq x. \end{cases}$$

Now, applying the law of total expectation,

$$\begin{aligned} m(x) &= \mathbb{E}[N(x)] \\ &= \mathbb{E}\Big[\mathbb{E}[(N(x) \,|\, U_1]\Big] = \int_0^1 \mathbb{E}[N(x) \,|\, U_1 = y] \cdot 1 \, dy \\ &= \int_0^x \big(1 + \mathbb{E}[N(x-y)]\big) \, dy + \int_x^1 1 \, dy \\ &= \int_0^1 dy + \int_0^x \mathbb{E}[N(x-y)] \, dy \\ \implies m(x) &= 1 + \int_0^x m(x-y) \, dy. \end{aligned}$$

This is a hidden differential equation, but first we make the substitution $u := x - y$ and $du = -dy$ to obtain

$$m(x) = 1 + \int_0^x m(u) \, du \implies m'(x) = m(x).$$

Now, we know what the solution is: it is $m(x) = Ce^x$ for some constant C. Now, $m(0) = 1$, so we just have $\boxed{m(x) = e^x}$. •

98 Conditional Variance

Just as we can define conditional expectation, we can also define conditional variance. Similarly, just as $\mathbb{E}[X|Y]$ is a random variable, so is $\mathrm{Var}(X|Y)$, as we define below.

Definition 98.1. Let X, Y be random variables. We define the *conditional variance* of X, given $Y = y$, by

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

The formula above is quite messy, so we do the following. Expanding out as we do with regular variance, we see that

$$\mathrm{Var}(X|Y) = \mathbb{E}[X^2|Y] - \big(\mathbb{E}[X|Y]\big)^2. \tag{6}$$

We also have the following crude lemma:

Lemma 98.2. *For random variables X, Y, we have* $\mathbb{E}[\operatorname{Var}(X|Y)] = \mathbb{E}[X^2] - \mathbb{E}[(\mathbb{E}[X|Y])^2]$.

Proof. Take expectations on both sides of equation (6), and simplify using the law of total expectation. □

To clean up this lemma, notice that for a function of a single variable g, we have $\operatorname{Var}(g(Y)) = \mathbb{E}[g^2(Y)] - \mathbb{E}[g(Y)]^2$, so taking $g(y) = \mathbb{E}[X|Y]$, we derive the following.

Proposition 98.3 (Law of Total Variance)**.** *Let X, Y be random variables. Then*

$$\operatorname{Var}(X) = \mathbb{E}[\operatorname{Var}(X|Y)] + \operatorname{Var}(\mathbb{E}[X|Y]).$$

Proof. From the above discussion, notice $\operatorname{Var}(\mathbb{E}[X|Y]) = \mathbb{E}[(\mathbb{E}[X|Y])^2] - \mathbb{E}[\mathbb{E}[X|Y]]^2 = \mathbb{E}[(\mathbb{E}[X|Y])^2] - \mathbb{E}[X]^2$, so

$$\begin{aligned}\mathbb{E}[\operatorname{Var}(X|Y)] + \operatorname{Var}(\mathbb{E}[X|Y]) &= \mathbb{E}[X^2] - \cancel{\mathbb{E}[(\mathbb{E}[X|Y])^2]} + \cancel{\mathbb{E}[(\mathbb{E}[X|Y])^2]} - \mathbb{E}[X]^2 \\ &= \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \operatorname{Var}(X),\end{aligned}$$

which proves the claim. □

We demonstrate that the above proposition is actually useful.

Example 98.4. Let $N(t)$ be the number of people who arrived at a train station by time t, so we may model $N(t) \sim \mathsf{Poi}(\lambda t)$ for some parameter $\lambda > 0$. Let $Y \sim \mathcal{U}(0, \tau)$ be the arrival time of the train, with $N(t)$ and Y independent. When the train arrives, everyone at the station boards. Find the mean and the variance of the number of people entering the train.

Solution. The number of people entering the train is $N(Y)$. Notice that $(N(Y)|Y = y) = N(y) \sim \mathsf{Poi}(\lambda y)$, so taking expectations, $\mathbb{E}[N(Y)|Y = y] = \lambda y = \operatorname{Var}(N(Y)|Y = y)$, only using the fact that this random variable is distributed Poisson. Now,

$$\mathbb{E}[N(Y)] = \mathbb{E}[\lambda Y] = \lambda \mathbb{E}[Y] = \boxed{\frac{\lambda\tau}{2}}$$

because Y is distributed uniformly. For the expectation, we have not used anything new, but for the variance, we apply Proposition 98.3:

$$\begin{aligned}\operatorname{Var}(N(Y)) &= \mathbb{E}[\operatorname{Var}(N(Y)|Y)] + \operatorname{Var}(\mathbb{E}[N(Y)|Y]) \\ &= \mathbb{E}[\lambda Y] + \operatorname{Var}(\lambda Y) \\ &= \lambda\mathbb{E}[Y] + \lambda^2 \operatorname{Var}(Y) \\ &= \boxed{\frac{\lambda\tau}{2} + \frac{(\lambda\tau)^2}{12}}.\end{aligned}$$

The fact that these were known distributions was very helpful here. •

Example 98.5. Fix the notation from Example 94.7. We compute $\operatorname{Var}\left(\sum_{i=1}^N X_i\right)$. First, we can show (as an exercise) $\operatorname{Var}\left(\sum_{i=1}^N X_i \middle| N\right) = N \operatorname{Var}(X)$. Then we write, using the law of total variance,

$$\begin{aligned}\operatorname{Var}\left(\sum_{i=1}^{N} X_i\right) &= \mathbb{E}\left[\operatorname{Var}\left(\sum_{i=1}^{N} X_i \middle| N\right)\right] + \operatorname{Var}\left(\mathbb{E}\left[\sum_{i=1}^{n} X_i \middle| N\right]\right)\\ &= \mathbb{E}[N \operatorname{Var}(X)] + \operatorname{Var}(N\mathbb{E}[X])\\ &= \boxed{\operatorname{Var}(X)\mathbb{E}[N] + (\mathbb{E}[X])^2 \operatorname{Var}(N)}.\end{aligned}$$

100 Moment Generating Functions

Recall that

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

which is absolutely convergent for every real number t. Now, let us try to plug in a random variable into the exponent:

$$\begin{aligned} e^{tX} &= \sum_{n=0}^{\infty} \frac{(tX)^n}{n!} = \sum_{n=0}^{\infty} \frac{t^n X^n}{n!}\\ \implies \mathbb{E}[e^{tX}] &= \mathbb{E}\left[\sum_{n=0}^{\infty} \frac{t^n X^n}{n!}\right] = \sum_{n=0}^{\infty} \frac{t^n \mathbb{E}[X^n]}{n!}.\end{aligned}$$

Notice that this power series contains all of the moments of X, so we define the following:

Definition 100.1. Let X be a random variable. We define the *moment generating function* of X by $M_X(t) := \mathbb{E}[e^{tX}]$.

Now, taking derivatives, we can do the following:

$$M_X'(t) = \frac{d}{dt}\mathbb{E}[e^{tX}] = \mathbb{E}\left[\frac{d}{dt}e^{tX}\right] = \mathbb{E}[Xe^{tX}].$$

Plugging in $t = 0$, we see $M_X'(0) = \mathbb{E}[X]$, and it follows by induction that $M_X^{(n)}(t) = \mathbb{E}[X^n e^{tX}]$ so that $M_X^{(n)}(0) = \mathbb{E}[X^n]$ for all $n \in \mathbb{Z}^+$. We record this result.

Proposition 100.2. *Let X be a random variable. Then $\mathbb{E}[X^n] = M_X^{(n)}(0)$.*

Example 100.3. Let $X \sim \mathsf{Exp}(\lambda)$, with $\lambda > 0$ and thus $X > 0$. Let $t < \lambda$, so $\lambda - t > 0 \implies -(\lambda - t) < 0$. We find $M_X(t)$ directly:

$$M_X(t) = \mathbb{E}[e^{tX}] = \int_0^{\infty} e^{tx}\lambda e^{-\lambda x}\,dx = \lambda\int_0^{\infty} e^{(t-\lambda)x}\,dx = \frac{-\lambda}{\lambda - t}e^{-(\lambda-t)x}\bigg|_0^{\infty}.$$

Since $-(\lambda - t) < 0$, we just have[3] $M(t) = \dfrac{\lambda}{\lambda - t} = \lambda(\lambda - t)^{-1}$ for $t < \lambda$. Now, we compute

[3] We drop the subscript X when it is clear from context.

$$\begin{aligned} M'(t) &= \frac{\lambda}{(\lambda-t)^2} \implies \mathbb{E}[X] = M'(0) = \frac{1}{\lambda}, \\ M''(t) &= \frac{2\lambda}{(\lambda-t)^3} \implies \mathbb{E}[X^2] = M''(0) = \frac{2}{\lambda^2}, \\ M'''(t) &= \frac{6\lambda}{(\lambda-t)^4} \implies \mathbb{E}[X^3] = M'''(0) = \frac{6}{\lambda^3}. \end{aligned}$$

Inducting, it is not too hard to see $M^{(n)}(t) = \dfrac{n! \cdot \lambda}{(\lambda-t)^{n+1}}$, so $\mathbb{E}[X^n] = \boxed{\dfrac{n!}{\lambda^n}}$.

Example 100.4. Let $X \sim \mathsf{Poi}(\lambda)$; we compute $M_X(t)$. By definition,

$$M(t) = \mathbb{E}[e^{tX}] = \sum_{n=0}^{\infty} e^{tn} \frac{e^{-\lambda}\lambda^n}{n!} = e^{-\lambda} \sum_{n=0}^{\infty} \frac{e^{tn}\lambda^n}{n!} = e^{-\lambda} \sum_{n=0}^{\infty} \frac{(e^t\lambda)^n}{n!} = e^{-\lambda} e^{\lambda e^t} = \boxed{e^{\lambda(e^t-1)}}.$$

Example 100.5. Let $X \sim \mathsf{Bin}(n,p)$; we compute $M_X(t)$. By definition,

$$M_X(t) = \mathbb{E}[e^{tX}] = \sum_{k=0}^{n} e^{tk} \binom{n}{k} p^k (1-p)^{n-k} = \sum_{k=1}^{n} \binom{n}{k} (e^t p)^k (1-p)^{n-k} = \boxed{(e^t p + (1-p))^n}$$

by the binomial theorem.

The following result is very useful.

Proposition 100.6. *Random variables X, Y are independent if and only if $M_{X+Y}(t) = M_X(t) M_y(t)$ holds identically.*

Proof. Suppose X and Y are independent. Then certainly $\mathbb{E}[g(X)h(Y)] = \mathbb{E}[g(X)]\mathbb{E}[h(Y)]$ for functions g, h of one variable, so it follows that

$$\mathbb{E}[e^{t(X+Y)}] = \mathbb{E}[e^{tX}e^{tY}] = \mathbb{E}[e^{tX}]\mathbb{E}[e^{tY}].$$

The converse is much more difficult, and requires inverse Laplace transforms (as in Math 3D), so we skip it here. □

This gives two quick proofs of two results that we have seen before.

Example 100.7. Say X, Y are independent with $X \sim \mathsf{Poi}(\lambda_1)$ and $Y \sim \mathsf{Poi}(\lambda_2)$. Then

$$M_{X+Y}(t) = M_X(t) M_Y(t) = e^{\lambda_1(e^t-1)} e^{\lambda_2(e^t-1)} = e^{(\lambda_1+\lambda_2)(e^t-1)},$$

so we see $X + Y \sim \mathsf{Poi}(\lambda_1 + \lambda_2)$.

Example 100.8. Say X, Y are independent with $X \sim \mathsf{Bin}(n,p)$ and $Y \sim \mathsf{Bin}(m,p)$. Then

$$M_{X+Y}(t) = M_X(t) M_Y(t) = (e^t p + (1-p))^n (e^t p + (1-p))^m = (e^t p + (1-p))^{m+n},$$

so we see $X + Y \sim \mathsf{Bin}(m+n, p)$.

104 Markov's and Chebyshev's Inequalities

We state two preliminary results, which are important in their own right but lead to several larger theorems.

Proposition 104.1 (Markov's Inequality). *If X is a nonnegative random variable, then if $a > 0$, then $\Pr(X \geq a) \leq \frac{\mathbb{E}[X]}{a}$.*

Proof. Consider the indicator variable $\mathbb{1}_{X \geq a}$. If $0 < X < a$, then $\mathbb{1}_{X \geq a} = 0$. In particular, $\mathbb{1}_{X \geq a} < X$ as X is nonnegative, so we have $\mathbb{1}_{X \geq a} < \frac{X}{a}$. When $X \geq a$, we have $\mathbb{1}_{X \geq a} = 1 \leq \frac{X}{a}$. In either case, we have $\mathbb{1}_{X \geq a} \leq \frac{X}{a}$, and taking expectations proves the result. □

Notice that when $a \to \infty$, then $\Pr(X \geq a)$ is nicely bounded, assuming $\mathbb{E}[X]$ is finite.

Proposition 104.2 (Chebyshev's Inequality). *If X is a random variable with $\mathbb{E}[X] =: \mu$ and $\mathrm{Var}(X) =: \sigma^2$, then for all $k > 0$, we have $\Pr(|X - \mu| \geq k) \leq \frac{\sigma^2}{k^2}$.*

Proof. Notice $\Pr(|X - \mu| \geq k) = \Pr((X - \mu)^2 \geq k^2)$, which is justified as $k > 0$. Now, the random variable $(X - \mu)^2$ is non-negative, so Markov's Inequality applies:

$$\Pr((X - \mu)^2 \geq k^2) \leq \frac{\mathbb{E}[(X - \mu)^2]}{k^2} = \frac{\mathrm{Var}(X)}{k^2} = \frac{\sigma^2}{k^2},$$

so we are done. □

Example 104.3. Let X be a random variable with $\mathbb{E}[X] = \mu = 30$. Then Markov's Inequality states that $\Pr(X \geq 45) \leq \frac{30}{45} = \frac{2}{3}$, then, in addition, if we know $\mathrm{Var}(X) = \sigma^2 = 300$, then by Chebyshev's Inequality,

$$\Pr(10 \leq X \leq 50) = \Pr(|X - 30| \leq 20) = 1 - \Pr(|X - \mu| \geq 20) \geq 1 - \frac{300}{20^2} = \frac{1}{4}.$$

Of course, these are merely approximations. If $X \sim \mathcal{U}(0, 60)$, we have $\mathbb{E}[X] = 30$ and $\mathrm{Var}(X) = 300$; we see $\Pr(X \geq 45) = \int_{45}^{60} \frac{1}{60}\, dx = \frac{1}{4} < \frac{2}{3}$; similarly, $\Pr(10 \leq X \leq 50) = \frac{2}{3}$.

Example 104.4. Let $X_1, X_2, \ldots, X_n$ be independent and identically distributed random variables with $X_i \sim \mathcal{N}(50, 400)$. Let $\bar{X}$ be the sample mean of these variables. Find the smallest value of n such that $\Pr(45 < \bar{X} < 55) \geq 0.99$.

Solution. Recall that $\mathbb{E}[\bar{X}] = \mathbb{E}[X_i] = 50$, and $\mathrm{Var}(\bar{X}) = \frac{\mathrm{Var}(X_i)}{n} = \frac{400}{n}$. Now, by Chebyshev,

$$\Pr(45 < \bar{X} < 55) = 1 - \Pr(|\bar{X} - \mathbb{E}[\bar{X}]| \geq 5) \geq 1 - \frac{\mathrm{Var}(\bar{X})}{25} = 1 - \frac{16}{n}.$$

Now we want $\Pr(45 < \bar{X} < 55) \geq \frac{99}{100}$, so we set $1 - \frac{16}{n} \geq \frac{99}{100}$:

$$1 - \frac{16}{n} \geq \frac{99}{100} \implies \frac{16}{n} \leq \frac{1}{100} \implies n \geq 1600.$$

Hence, we must have at least $\boxed{1600}$ samples in order for $\bar{X}$ to be within 5 of the sample mean, with probability 99%. •

108 Limit Theorems

The preceding example is a specific case of the following idea.

Theorem 108.1 (Weak Law of Large Numbers)**.** *Let $X_1, X_2, X_3, \ldots$ be independent and identically distributed random variables, with $\mathbb{E}[X_i] = \mu$. Then for all $\varepsilon > 0$, we have*

$$\Pr\left(\left|\frac{X_1 + X_2 + \cdots + X_n}{n} - \mu\right| \geq \varepsilon\right) \to 0 \text{ as } n \to \infty.$$

Proof. The proof is essentially similar to the working of Example 104.4 — apply Chebyshev's Inequality. □

Similarly, we have seen allusions to the following theorem:

Theorem 108.2 (Central Limit Theorem)**.** *Let $X_1, X_2, X_3, \ldots$ be independent and identically distributed random variables, with $\mathbb{E}[X_i] = \mu$ and $\operatorname{Var}(X_i) = \sigma^2$. Then*

$$\frac{X_1 + X_2 + \cdots + X_n - n\mu}{\sigma\sqrt{n}} \to N(0,1) \text{ as } n \to \infty.$$

More specifically, if $Z \sim N(0,1)$, then for all $a \in \mathbb{R}$,

$$\lim_{n\to\infty} \Pr\left(\frac{X_1 + X_2 + \cdots + X_n - n\mu}{\sigma\sqrt{n}} \leq a\right) = \Pr(Z \leq a).$$

That is, as we increase the sample size n, the normalized sample mean $(\sigma X_i - n\mu)/(\sigma\sqrt{n})$ approaches the standard normal. If we do not standardize, this is equivalent to saying that the sample mean itself approaches a normal variable (which may or may not be standardized).

Example 108.3. Say that X is the number of students enrolled in a class, distributed $X \sim \mathsf{Poi}(100)$, so that $\mathbb{E}[X] = \operatorname{Var}(X) = 100$. If $X \geq 120$, then the department split the class into two sections. Approximate $\Pr(X \geq 120)$.

Solution. Take $Z \sim \mathcal{N}(0,1)$, so that $\Pr(X \geq 120) \approx \Pr(Z \geq 2) = 1 - \Pr(Z \leq 2) \approx 0.025$. This seems to work fine, but the central limit theorem seems to only apply to sums of random variables? To explain this, write

$$X = \sum_{i=1}^{100} X_i,$$

where X_i is the number of students enrolled in one slot of the class. Note that $X_i \sim \mathsf{Poi}(1)$: on average, 1 person enrolls in 1 slot of the class, but classes could be under-enrolled or waitlisted. Certainly, now X is a sum, so the central limit theorem applies. •