1 Multivariate Distributions

We now develop a theory of probability to describe the simultaneous behavior of multiple (possibly dependent) random variables. To simplify the notation, we focus on bivariate random variables, but the ideas extend in the obvious way to several random variables. This is the analogue of multi-variable functions from calculus.

1.1 Bivariate Distributions

We first consider the bivariate (2 variable) case where X and Y are random variables defined on the same sample space taking values $(x, y) \in \mathbb{R}^2$. The case with n random variables is similar and will be described in Section 1.3. We will see that all definitions are straightforward generalization of the univariate (single variable) case.

Remark 1. We will define everything for discrete and continuous random variables to be precise, but the ideas of the joint and marginal distributions are very similar between the two. We just replace the PMF with the PDF and replace sums with integrals just like in the univariate case.

1.1.1 Joint Distributions

The probabilities of objects involving both X and Y are encoded by the joint CDF.

Definition 1 (Joint Cumulative Distribution Function)**.** The *joint CDF* of random variables X and Y is the function $F_{X,Y} : \mathbb{R}^2 \to [0, 1]$ defined by

$$F_{X,Y}(x, y) = \mathbb{P}(X \le x, Y \le y) \qquad x, y \in \mathbb{R}.$$

Just like in the univariate case, the CDF allows us to compute the probability of any random variable taking values in any subset of $\mathbb{R}^2$. Just like the PMF and PDF, in the case when the random variables are discrete or continuous, there is an equivalent notion of probability functions.

Definition 2 (Joint Probability Mass Function)**.** The *joint PMF* of X and Y is

$$p_{X,Y}(x, y) = \mathbb{P}(\{\omega \in \Omega : X(\omega) = x\} \cap \{\omega \in \Omega : Y(\omega) = y\}) = \mathbb{P}(X = x, Y = y)$$

for $x \in \mathcal{X}, y \in \mathcal{Y}$ and 0 otherwise.

The joint PMF is still a probability function in the sense that

1. $0 \le p_{X,Y}(x, y) \le 1$
2. $\sum_{x,y} p_{X,Y}(x, y) = 1$.

Definition 3 (Joint Probability Density Function)**.** The *joint probability density function* of X and Y is

$$f_{X,Y}(x, y) = \frac{\partial^2}{\partial x \partial y} F_{X,Y}(x, y).$$

The joint PDF is not a probability (much like in the univariate case), but it satisfies the normalization property

1. $f_{X,Y} \ge 0$
2. $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{X,Y}(x, y)\, dxdy = 1$.

Using the joint distributions, we can compute the probability of any set $A \subset \mathbb{R}^2$,

$$\mathbb{P}((X, Y) \in A) = \sum_{(x,y)\in A} p_{X,Y}(x, y) \qquad \text{or} \qquad \mathbb{P}((X, Y) \in A) = \iint_A f_{X,Y}(x, y)\, dxdy.$$

1.1.2 Marginal Distributions

The probabilities of only one random variable are encoded by the marginal distributions. These notions are the same as in the univariate case, and can be recovered by "integrating out" the other random variables we are not interested in.

Definition 4 (Marginal Cumulative Distribution Function)**.** Suppose that X and Y are random variables with joint CDF $F_{X,Y}(x, y)$. The *marginal CDF* of X is

$$F_X(x) = \mathbb{P}(X \le x) = \mathbb{P}(X \le x, Y \le \infty) = \lim_{y\to\infty} F_{X,Y}(x, y).$$

Similarly, the marginal CDF of Y is

$$F_Y(y) = \mathbb{P}(Y \le y) = \mathbb{P}(X \le \infty, Y \le y) = \lim_{x\to\infty} F_{X,Y}(x, y).$$

When X and Y are either discrete or continuous, we have the following notions of the probability functions, which behave like the PMF and PDF we covered earlier.

Definition 5 (Marginal Probability Mass Function)**.** Suppose that X and Y are *discrete* random variables with joint PMF $p_{X,Y}(x, y)$. The *marginal PMF* of X is

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

Similarly, the marginal distribution of Y is

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

Definition 6 (Marginal Probability Density Function)**.** Suppose that X and Y are *continuous* random variables with joint probability function $f_{X,Y}(x, y)$. The *marginal PDF* of X is

$$f_X(x) = \int f_{X,Y}(x, y)\, dy.$$

Similarly, the marginal distribution of Y is

$$f_Y(y) = \int f_{X,Y}(x, y)\, dx.$$

Remark 2. We can go from the joint distributions to the marginal distributions, but we cannot go the other way around. By only looking at the marginal distributions, we do not know how the random variables behave together without further assumptions.

1.1.3 Independence

Recall we say that events A and B are independent, if $\mathbb{P}(A\cap B) = \mathbb{P}(A)\cdot\mathbb{P}(B)$. Since the CDFs encode the behaviors of random variables we get the following definition.

Definition 7 (Independence)**.** X and Y are *independent* random variables if

$$F_{X,Y}(x, y) = \mathbb{P}(X \le x, Y \le y) = \mathbb{P}(X \le x)\,\mathbb{P}(Y \le y) = F_X(x)F_Y(y)$$

for all values of (x, y).

There is a converse of this result as well, that says that if the probability functions factorizes, then it must correspond to independent random variables.

Proposition 1 (*Factorization Characterization*)

If the joint PDF of X and Y factorizes as

$$f_{X,Y}(x,y) = g(x)h(y) \qquad \text{for all } x, y \in \mathbb{R}$$

for some non-negative functions g and h, then X and Y are independent. Furthermore, if either g or h is a valid PDF, then the other is one too, and they correspond to the marginal PDFs of X and Y respectively. An analogous statement holds for the PMF.

The factorization property is what makes it easy to add dimensions. Going from the joint distribution of n variables to $n+1$ can be achieved by simply multiplying by another marginal. This is convenient when we are working with several independent observations.

Remark 3. If X and Y are independent, then it is possible to recover the joint distribution from the marginals. In this case, the joint distribution is simply the product of the marginals.

1.2 Joint Summary Statistics

We now introduce the summary statistics that generalizes the expected value and variances to the multivariate setting.

1.2.1 Expected Value

Definition 8 (Expected Value)**.** Suppose X and Y are discrete/continuous random variables with joint probability functions $p_{X,Y}/f_{X,Y}$. Then for any function $g : \mathbb{R}^2 \to \mathbb{R}$,

$$\mathbb{E}\left[g(X,Y)\right] = \sum_{(x,y)} g(x,y)p_{X,Y}(x,y) \qquad \text{or} \qquad \mathbb{E}\left[g(X,Y)\right] = \iint_{\mathbb{R}^2} g(x,y)f_{X,Y}(x,y)\,dxdy$$

depending on if X, Y are jointly discrete or continuous.

Properties:

1. *Linearity of Expectation:* If X and Y are any random variables, then

$$\mathbb{E}[ag_1(X,Y) + bg_2(X,Y)] = a \cdot \mathbb{E}[g_1(X,Y)] + b \cdot \mathbb{E}[g_2(X,Y)].$$

   In particular, if X and Y are any random variables (not necessarily independent), then

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

2. *Product of two Independent Random Variables:* If X and Y are independent, then

$$\mathbb{E}[g_1(X)g_2(Y)] = \mathbb{E}[g_1(X)]\,\mathbb{E}[g_2(Y)].$$

1.2.2 Covariance

The covariance measures the joint variability of two random variables.

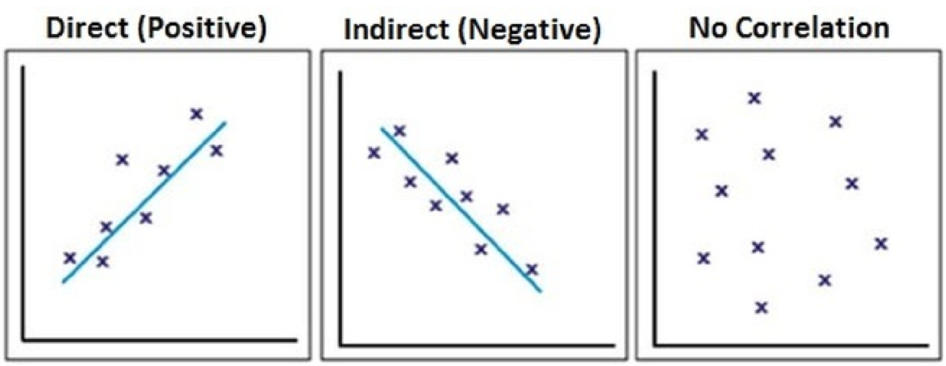
Definition 9 (Covariance)**.** For two random variables X and Y, the covariance between X and Y is

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

provided the expression exists.

1.2.3 Properties

1. *Relationship with Variance:* $\mathrm{Cov}(X, X) = \mathrm{Var}(X)$.
2. *Equivalent formula:*
$$\mathrm{Cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\,\mathbb{E}[Y].$$
3. *Relationship with independence I:* If X and Y are independent,
$$\mathrm{Cov}(X, Y) = 0.$$
The converse of this statement is false!. There are pairs of random variables that have zero covariance, but are dependent (see Problem 1.14).
4. *Relationship with Independence II:* If X and Y have zero covariance, then
$$\mathbb{E}[XY] = \mathbb{E}[X]\,\mathbb{E}[Y].$$
5. *Cauchy–Schwarz Inequality:* For any random variables X and Y,
$$|\,\mathbb{E}[XY]| \le \sqrt{\mathbb{E}(X^2)}\sqrt{\mathbb{E}(Y^2)} \qquad \text{or} \qquad |\mathrm{Cov}(X, Y)| \le \sqrt{\mathrm{Var}(X)}\sqrt{\mathrm{Var}(Y)}.$$
6. *The Sign of the Covariance*: Suppose X, Y are positively related (when X large, Y likely large; when X small, Y likely small), then
$$\mathrm{Cov}(X, Y) > 0$$
Conversely, suppose X, Y are negatively related (when X large, Y likely small; when X small, Y likely large), then
$$\mathrm{Cov}(X, Y) < 0.$$



1.2.4 Correlation

The correlation measures how linearly related two random variables are.

Definition 10. The *correlation* of X and Y, denoted $\mathrm{Corr}(X, Y)$, is defined by
$$\rho = \mathrm{Corr}(X, Y) = \frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{Var}(X)}\sqrt{\mathrm{Var}(Y)}} = \frac{\mathrm{Cov}(X, Y)}{\mathrm{SD}(X)\mathrm{SD}(Y)}.$$
We say that X and Y are *uncorrelated* if $\mathrm{Cov}(X, Y) = 0$ (or equivalently $\mathrm{Corr}(X, Y) = 0$). We have implicitly assumed that X and Y have non-zero variance in this definition

The correlation satisfies the following properties

1. $-1 \le \rho \le 1$
2. $|\rho| = 1 \;\Leftrightarrow\; X = aY + b$. If $a > 0$, then $\rho = 1$, and if $a < 0$, then $\rho = -1$.
3. $\mathrm{Corr}(X, Y) = 0 \not\Rightarrow X, Y$ independent in general (see Problem 1.14).

1.3 Multivariate Random Variables

We considered the case of bivariate random variables above, but all the terminology above can be extended to a collection $X_1, X_2, \dots, X_n$ of random variables in the obvious way. We will state the results for discrete random variables, but the obvious modifications also defines the continuous analogue of every definition.

Definition 11 (Multivariate Joint PMF)**.** For a collection of n discrete random variables, $X_1, ..., X_n$, the joint probability function is defined as

$$p_{X_1,...,X_n}(x_1, x_2, ..., x_n) = \mathbb{P}(X_1 = x_1, X_2 = x_2, ..., X_n = x_n).$$

and we call the vector $(X_1, \dots, X_n)$ a random vector.

The notion of independence generalized naturally.

Definition 12 (Multivariate Independence)**.** $X_1, X_2, \dots, X_n$ are *independent* if

$$p_{X_1,...,X_n}(x_1, x_2, \dots, x_n) = p_{X_1}(x_1)p_{X_2}(x_2)\cdots p_{X_n}(x_n)$$

for all values of $(x_1, \dots, x_n)$.

The definition of the expected value generalizes naturally as well.

Definition 13 (Multivariate Expected Value)**.** If $g : \mathbb{R}^n \to \mathbb{R}$, and $X_1, ..., X_n$ are discrete random variables with joint PMF $p_{X_1,...,X_n}(x_1, ..., x_n)$, then

$$\mathbb{E}\left[g(X_1, ..., X_n)\right] = \sum_{(x_1,...,x_n)} g(x_1, ..., x_n)p_{X_1,...,X_n}(x_1, ..., x_n).$$

1.3.1 Covariance Matrix

The notion of a variance is a bit trickier since there are many possible covariances to deal with. The covariances between all entries is summarized by a single matrix called the covariance matrix.

Definition 14 (Covariance Matrix)**.** The covariance matrix of $\boldsymbol{X} = (X_1, X_2, \dots, X_n) \in \mathbb{R}^n$ is a $n \times n$ matrix with entries given by $\mathrm{Cov}(X_i, X_j)$, i.e.

$$\mathrm{Cov}(\boldsymbol{X}) = [\mathrm{Cov}(X_i, X_j)]_{i,j\le n} \in \mathbb{R}^{n\times n}.$$

To compute the covariance matrix of $\boldsymbol{X} \in \mathbb{R}^n$, we have the following familiar formula

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

where the expectation is applied to each entry of the vector or matrix.

Remark 4. This is a generalization of the univariate formula

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

In fact, one can simply recognize that the vector formula is simply the definition of the covariance applied entrywise

$$\mathrm{Cov}(\boldsymbol{X}) = \mathbb{E}[\boldsymbol{X}\boldsymbol{X}^\intercal] - \mathbb{E}[\boldsymbol{X}]\,\mathbb{E}[\boldsymbol{X}]^\intercal = [\mathbb{E}[X_iX_j] - \mathbb{E}[X_i]\,\mathbb{E}[X_j]]_{i,j\le n}$$

Proposition 2

The covariance matrix is symmetric and positive semidefinite.

1.3.2 Linear Combinations of Random Variables

We are often interested in the linear combinations of random variables.

Definition 15. A *linear combination* of the random variables $X_1, ..., X_n$ is any random variable $Y \in \mathbb{R}$ of the form

$$Y = \sum_{i=1}^{n} a_i X_i \qquad \text{where } a_1, ..., a_n \in \mathbb{R}.$$

Let $a_1, \ldots, a_n, b_1, \ldots, b_n \in \mathbb{R}$. We have the following properties about linear combinations of random variables.

1. **Linearity of Expectation:** For any random variables $X_1, \ldots, X_n$,

$$\mathbb{E}\left[\sum_{i=1}^{n} a_i X_i\right] = \sum_{i=1}^{n} a_i \,\mathbb{E}[X_i].$$

2. **Bi-Linearity of Covariance:** For any random variables $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$,

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

   In particular, for random variables X, Y, U, V be random variables, and $a, b, c, d \in \mathbb{R}$. Then,

$$\begin{aligned}&\mathrm{Cov}(aX + bY, cU + dV)\\ &= ac\mathrm{Cov}(X, U) + ad\mathrm{Cov}(X, V) + bc\mathrm{Cov}(Y, U) + bd\mathrm{Cov}(Y, V)\end{aligned}$$

3. **Variance of Linear Combinations:** The following result shows how the variance of a linear combination is "broken down" into pieces:

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

   In particular, for random variables X, Y, and $a, b \in \mathbb{R}$,

$$\mathrm{Var}(aX + bY) = a^2 \,\mathrm{Var}(X) + b^2 \,\mathrm{Var}(Y) + 2ab\mathrm{Cov}(X, Y).$$

   If the $X_1, \ldots, X_n$ are independent, then they are uncorrelated, so in this case

$$\mathrm{Var}\left(\sum_{i=1}^{n} a_i X_i\right) = \sum_{i=1}^{n} a_i^2 \,\mathrm{Var}(X_i)$$

1.3.3 Sums of Independent Random Variables

We now list the distributions of some linear combinations of random vectors. Many of these properties we have already encountered through examples.

1. **Sum of Independent Binomials is Binomial:** If $X \sim \mathrm{Bin}(n, p)$ and $Y \sim \mathrm{Bin}(m, p)$ are independent, then

$$T = X + Y \sim \mathrm{Bin}(n + m, p).$$

2. **Sum of Independent Normal is Normal:** If $X \sim \mathrm{N}(\mu_1, \sigma_1^2)$ and $Y \sim \mathrm{N}(\mu_2, \sigma_2^2)$ are independent, then

$$T = X + Y \sim \mathrm{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2).$$

3. **Sum of Independent Poisson is Poisson:** If $X \sim \text{Poi}(\lambda_1)$ and $Y \sim \text{Poi}(\lambda_2)$ are independent, then
$$T = X + Y \sim \text{Poi}(\lambda_1 + \lambda_2).$$

4. **Sum of Independent Bernouilli is Binomial:** Let $X_1, X_2, \ldots, X_n$ be independent $\text{Bern}(p)$ random variables. Then,
$$T = X_1 + X_2 + \ldots + X_n \sim \text{Bin}(n, p).$$

5. **Sum of Independent Geometric is Negative Binomial:** Let $X_1, X_2, \ldots, X_k$ be independent $\text{Geo}(p)$ random variables. Then,
$$T = X_1 + X_2 + \ldots + X_k \sim \text{NegBin}(k, p).$$

Remark 5. Properties 1, 4, and 5 follow directly from the construction of these random variables.

1.4 Example Problems

Problem 1.1. Let $X \in \{1, 2, 3\}$ and $Y \in \{1, 2\}$, and suppose that every outcome of (X, Y) is equally likely. What is the joint PMF for the vector (X, Y)?

Solution 1.1. We can compute all the probabilities one by one and encode the joint PMF of X and Y in the table

$p_{X,Y}(x,y)$		x: 1	2	3	$p_Y(y)$
y	1	1/6	1/6	1/6	3/6
	2	1/6	1/6	1/6	3/6
$p_X(x)$		2/6	2/6	2/6	1

Problem 1.2. Suppose a fair coin is tossed 3 times. Define the random variables X = "number of Heads", and
$$Y = \begin{cases} 1 & \text{Head occurs on the first toss,} \\ 0 & \text{Tail occurs on the first toss.} \end{cases}$$

1. Find the joint PMF for (X, Y).
2. Are X and Y independent?
3. What is the probability that $X + Y = 2$?

Solution 1.2.

Part 1: We can compute all the probabilities one by one and encode the joint PMF of X and Y in the table

$p_{X,Y}(x,y)$		x: 0	1	2	3	$p_Y(y)$
y	0	1/8	2/8	1/8	0	1/2
	1	0	1/8	2/8	1/8	1/2
$p_X(x)$		1/8	3/8	3/8	1/8	1

Part 2: We can see

$$p_{X,Y}(0,1) = 0 \neq \frac{1}{8} \cdot \frac{1}{2} = p_X(0)p_Y(1)$$

which implies that X and Y are not independent (which makes perfect sense, as the number of heads we have should depend on whether we had heads in the first toss). **Part 3:** We have $X + Y = 2$ if and only if $X = 2, Y = 0$ or $X = 1, Y = 1$. We can sum these terms up in the joint PMF

$$\mathbb{P}(X + Y = 2) = p_{X,Y}(2,0) + p_{X,Y}(1,1) = \frac{1}{8} + \frac{1}{8} = \frac{1}{4}.$$

Problem 1.3. Let X and Y be any discrete random variables. Show that

1. $0 \le p_{X,Y}(x,y) \le 1$
2. $p_{X,Y}(x,y) \le p_X(x)$
3. $p_{X,Y}(x,y) \le p_Y(y)$

Solution 1.3.

1. We have $p_{X,Y}(x,y) = \mathbb{P}(X = x, Y = y)$ and all probabilities must be between 0 and 1.
2. We have $p_{X,Y}(x,y) = \mathbb{P}(X = x, Y = y) \le \mathbb{P}(X = x) = p_X(x)$ since $\{X = x, Y = y\} \subseteq \{X = x\}$.
3. We have $p_{X,Y}(x,y) = \mathbb{P}(X = x, Y = y) \le \mathbb{P}(Y = y) = p_Y(y)$ since $\{X = x, Y = y\} \subseteq \{Y = y\}$.

Problem 1.4. Suppose X and Y have joint PMF

$$p_{X,Y}(x,y) = \frac{1}{6}\left(\frac{1}{2}\right)^x \left(\frac{2}{3}\right)^y, \quad x, y = 0, 1, 2...$$

Find the marginal PMFs p_X and p_Y of X and Y.

Solution 1.4. Recall the identity

$$\sum_{k=0}^{\infty} q^k = \frac{1}{1-q}, \quad 0 < q < 1.$$

Part 1: The X marginal is

$$\begin{aligned} p_X(x) &= \sum_{y=0}^{\infty} \frac{1}{6}\left(\frac{1}{2}\right)^x \left(\frac{2}{3}\right)^y \\ &= \frac{1}{6}\left(\frac{1}{2}\right)^x \sum_{y=0}^{\infty} \left(\frac{2}{3}\right)^y = \frac{1}{6}\left(\frac{1}{2}\right)^x \frac{1}{1-\frac{2}{3}} \\ &= \frac{1}{2}\left(\frac{1}{2}\right)^x, \quad x = 0, 1, \ldots \end{aligned}$$

from which we conclude that $X \sim \text{Geo}(1/2)$.

Part 2: The Y marginal is

$$\begin{aligned} p_Y(y) &= \sum_{x=0}^{\infty} \frac{1}{6}\left(\frac{1}{2}\right)^x \left(\frac{2}{3}\right)^y \\ &= \frac{1}{6}\left(\frac{2}{3}\right)^y \sum_{x=0}^{\infty}\left(\frac{1}{2}\right)^x = \frac{1}{6}\left(\frac{2}{3}\right)^y \frac{1}{1-\frac{1}{2}} \\ &= \frac{1}{3}\left(\frac{2}{3}\right)^y, \quad y = 0, 1, \dots \end{aligned}$$

from which we conclude that $Y \sim \text{Geo}(1/3)$.

Problem 1.5. Suppose $X \sim \text{Poi}(2)$, $Y \sim \text{Poi}(3)$, and that X and Y are independent. What is the joint probability mass function of X and Y?

Solution 1.5. By independence, we that for all integer valued $x, y \geq 0$,

$$p_{X,Y}(x,y) = p_X(x)p_Y(y) = e^{-2}\frac{2^x}{x!}e^{-3}\frac{3^y}{y!} = e^{-5}\frac{2^x}{x!}\frac{3^y}{y!}.$$

Problem 1.6. If we roll a die n times, let's denote by $X_1, \dots, X_6$ the number of times we rolled a 1, 2,. . . , 6.

1. What is the distribution (or marginal probability function) of X_j for $j = 1, \dots, 6$?
2. Are $X_1, X_2, \dots, X_6$ independent?
3. What is the joint probability function of $(X_1, \dots, X_6)$?
4. Let's denote by $T = X_1 + X_2$ the number of times we had a 1 or two. What's the distribution of $T = X_1 + X_2$?

Solution 1.6.

Part 1: By definition, if X_j denotes the number of times we roll a j in n rolls, then

$$X_j \sim \text{Bin}\left(n, \frac{1}{6}\right).$$

Part 2: Intuitively, these are not independent because we must have $X_1 + \cdots + X_6 = n$ so X_6 is totally determined by X_1 to X_5. For example, if we consider the case

$$\mathbb{P}(X_1 = n, X_2 = n, \dots, X_6 = n) = 0$$

but

$$\mathbb{P}(X_1 = n) \cdots \mathbb{P}(X_6 = n) = \left(\frac{1}{6}^n\right)^6 > 0$$

so they are not independent.

Part 3: Let $x_1, \dots, x_6 \in \{0, 1, \dots, n\}$. As noted earlier, if $x_1 + x_2 + \cdots + x_6 \neq n$, then $\mathbb{P}(X_1 =$

$x_1, \dots, X_6 = x_6) = 0$. Thus, let $x_1 + x_2 + \cdots + x_6 = n$. We can arrange the x_1 rolls of 1, x_2 rolls of 2,..., x_6 of rolls of 6, among the n trials in

$$\frac{n!}{x_1!x_2!\dots x_6!}$$

many ways, using the formula for the arrangements with repeated objects: the 1 is repeated x_1 times, the 2 is repeated x_2 times, etc. Each of these arrangements has probability

$$\left(\frac{1}{6}\right)^{x_1} \cdot \left(\frac{1}{6}\right)^{x_2} \cdot \dots \cdot \left(\frac{1}{6}\right)^{x_6} = \left(\frac{1}{6}\right)^{x_1+\cdots+x_6} = \left(\frac{1}{6}\right)^{n}$$

Hence, the joint PMF of $(X_1, \dots, X_6)$ is

$$p_{X_1,\dots,X_6}(x_1, \dots, x_6) = \begin{cases} \frac{n!}{x_1!x_2!\dots x_6!}\left(\frac{1}{6}\right)^n, & \text{if } x_1 + x_2 + \cdots + x_6 = n, \\ 0 & \text{otherwise.} \end{cases}$$

Part 4: T counts the number of 1's and 2's after n rolls. The probability of rolling a 1 or 2 is $\frac{1}{3}$, so

$$T \sim \text{Bin}\Big(n, \frac{1}{3}\Big).$$

Remark 6. We will see in the appendix that we could have used the fact that $(X_1, \dots, X_6) \sim \text{Mult}(n, \frac{1}{6}, \dots, \frac{1}{6})$ and used the properties of the multinomial to derive all of the above parts.

Problem 1.7. Let X_1 and X_2 be independent exponential random variables with mean $\theta_1 = \frac{1}{\lambda_1}$ and $\theta_2 = \frac{1}{\lambda_2}$ respectively. Let Y denote the index of the smaller of X_1 and X_2, that is

$$Y = \begin{cases} 1 & X_1 < X_2 \\ 2 & X_2 < X_1. \end{cases}$$

In the case that $X_1 = X_2$ we set $Y = 1$. Furthermore, let $X_Y = \min(X_1, X_2)$ to denote the value of the smaller of X_1 and X_2.

1. Find the distribution of X_Y.
2. Find the distribution of Y.
3. Are X_Y and Y independent?

Solution 1.7. Recall that if $X \sim \text{Exp}(\theta)$ then $\mathbb{P}(X > t) = e^{-\lambda t}$ where $\lambda = \frac{1}{\theta}$.

Part 1: We have for $t \geq 0$,

$$F_{X_Y}(t) = \mathbb{P}(\min(X_1, X_2) \leq t) = 1 - \mathbb{P}(\min(X_1, X_2) > t).$$

Notice that by independence,

$$\mathbb{P}(\min(X_1, X_2) > t) = \mathbb{P}(X_1 > t)\,\mathbb{P}(X_2 > t) = e^{-\lambda_1 t}e^{-\lambda_2 t} = e^{-(\lambda_1+\lambda_2)t}.$$

To get the PDF, we differentiate with respect to t,

$$f_{X_Y}(t) = \frac{d}{dt}F_{X_Y}(t) = (\lambda_1 + \lambda_2)e^{-(\lambda_1+\lambda_2)t}$$

so X_Y is exponential with parameter $\theta = (\lambda_1 + \lambda_2)^{-1}$.

Part 2: There are only two cases, so

$$\begin{aligned}\mathbb{P}(Y = 2) = \mathbb{P}(X_1 > X_2) &= \int_0^\infty \int_{x_2}^\infty \lambda_1 \lambda_2 e^{-\lambda_1 x_1} e^{-\lambda_2 x_2} \, dx_1 dx_2 \\ &= \int_0^\infty \lambda_2 e^{-\lambda_1 x_2} e^{-\lambda_2 x_2} \, dx_2 \\ &= \frac{\lambda_2}{\lambda_1 + \lambda_2}.\end{aligned}$$

Therefore,

$$\mathbb{P}(Y = 1) = 1 - \mathbb{P}(Y = 2) = \frac{\lambda_1}{\lambda_1 + \lambda_2}.$$

Remark 7. An alternative way to compute this is using independence. We have

$$\mathbb{P}(X_1 > X_2) = \mathbb{E}[\mathbb{1}(X_1 > X_2)] = \mathbb{E}_{X_2} \mathbb{E}_{X_1}[\mathbb{1}(X_1 > X_2)] = \mathbb{E}_{X_2}[e^{-\lambda_1 X_2}] = \frac{1}{1 + \frac{\lambda_1}{\lambda_2}} = \frac{\lambda_2}{\lambda_1 + \lambda_2}$$

using the MGF formula for $X \sim \mathrm{Exp}(\theta)$, $M_X(t) = \mathbb{E}[e^{tX}] = \frac{1}{1-\theta t}$ for $t < \frac{1}{\theta}$.

Part 3: We have

$$\begin{aligned}\mathbb{P}(X_Y > t, Y = 1) = \mathbb{E}[\mathbb{1}(X_1 > t, Y = 1)] &= \int_t^\infty \int_{x_1}^\infty \lambda_1 \lambda_2 e^{-\lambda_1 x_1} e^{-\lambda_2 x_2} \, dx_2 dx_1 \\ &= \int_t^\infty \lambda_1 e^{-(\lambda_1 + \lambda_2) x_1} \, dx_1 \\ &= \frac{\lambda_1}{\lambda_1 + \lambda_2} e^{-(\lambda_1 + \lambda_2)t} \\ &= \mathbb{P}(X_Y > t) \, \mathbb{P}(Y = 1)\end{aligned}$$

since the indicator on the set $\{X_1 > t, Y = 1\}$ means that we integrate over the region $X_2 > X_1 > t$. An identical computation show that

$$\mathbb{P}(X_Y > t, Y = 2) = \mathbb{P}(X_Y > t) \, \mathbb{P}(Y = 2),$$

so X_Y and Y are independent. This is a somewhat surprising result since X_Y looks like it depends on the minimal index Y.

Remark 8. Notice that in this problem X_1, X_2 and X_Y are continuous random variables while Y is discrete. However, the notations and definitions for the joint distributions naturally extend to this case as well.

Problem 1.8. Let (X, Y) have the following joint probability mass function

$X \backslash Y$	0	1	2
1	0.1	0.2	0.1
2	0.1	0.3	0.2

1. Compute $\mathbb{E}[XY]$.
2. Compute $\mathbb{E}[X^2 Y]$.

Solution 1.8.

Part 1: From the definition of the expected value,

$$\mathbb{E}[XY] = \sum_{x=1}^{2}\sum_{y=0}^{2} xyp_{X,Y}(x,y) = 0 + 0.2 + 0.2 + 0 + 0.6 + 0.8 = 1.8$$

Part 2: From the definition of the expected value,

$$\mathbb{E}[X^2Y] = \sum_{x=1}^{2}\sum_{y=0}^{2} x^2yp_{X,Y}(x,y) = 0 + 0.2 + 0.2 + 0 + 1.2 + 1.6 = 3.2.$$

Problem 1.9. Let (X, Y) have the joint probability density function

$$f_{X,Y}(x,y) = \begin{cases} 24xy, & 0 \le x \le 1,\ 0 \le y \le 1-x, \\ 0, & \text{otherwise.} \end{cases}$$

1. Compute $\mathbb{E}[XY]$.
2. Compute $\mathbb{E}[X^2Y]$.

Solution 1.9.

Part 1: From the definition of the expected value,

$$\begin{aligned}
\mathbb{E}[XY] = \int_0^1 \int_0^{1-x} xy\, f_{X,Y}(x,y)\, dy\, dx &= \int_0^1 \int_0^{1-x} 24x^2y^2\, dy\, dx \\
&= \int_0^1 24x^2 \cdot \frac{(1-x)^3}{3}\, dx \\
&= 8\int_0^1 (x^2 - 3x^3 + 3x^4 - x^5)\, dx \\
&= 8\left(\frac{1}{3} - \frac{3}{4} + \frac{3}{5} - \frac{1}{6}\right) = \frac{2}{15}
\end{aligned}$$

Part 2: From the definition of the expected value,

$$\begin{aligned}
\mathbb{E}[X^2Y] = \int_0^1 \int_0^{1-x} x^2y\, f_{X,Y}(x,y)\, dy\, dx &= \int_0^1 \int_0^{1-x} 24x^3y^2\, dy\, dx \\
&= 8\int_0^1 (x^3 - 3x^4 + 3x^5 - x^6)\, dx \\
&= 8\left(\frac{1}{4} - \frac{3}{5} + \frac{1}{2} - \frac{1}{7}\right) = \frac{2}{35}.
\end{aligned}$$

Problem 1.10. Let X, Y be independent random variables with $\mathrm{Var}(X) = \mathrm{Var}(Y) = 1$. What is $\mathrm{Var}(X - Y)$?

Solution 1.10. By independence, $\mathrm{Cov}(X,Y) = 0$, so the variance for linear combinations formula implies

$$\mathrm{Var}(X - Y) = \mathrm{Var}(X) + (-1)^2\,\mathrm{Var}(Y) - 2\mathrm{Cov}(X,Y) = 1 + 1 = 2.$$

1.5 Proofs of Key Results

Problem 1.11. If X and Y are discrete random variables, show that

$$\mathbb{E}[ag_1(X,Y) + bg_2(X,Y)] = a \cdot \mathbb{E}[g_1(X,Y)] + b \cdot \mathbb{E}[g_2(X,Y)].$$

In particular, if $g_1 = x$ and $g_2 = y$ then

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

Solution 1.11. We have by the definition,

$$\begin{aligned}
\mathbb{E}[ag_1(X,Y) + bg_2(X,Y)] &= \sum_{(x,y)} [ag_1(x,y) + bg_2(x,y)] p_{X,Y}(x,y) \\
&= a \sum_{(x,y)} g_1(x,y) p_{X,Y}(x,y) + b \sum_{(x,y)} g_2(x,y) p_{X,Y}(x,y) \\
&= a \cdot \mathbb{E}[g_1(X,Y)] + b \cdot \mathbb{E}[g_2(X,Y)].
\end{aligned}$$

By taking $g_1(x,y) = x$ and $g_2(x,y) = y$ we immediately arrive at the fact that

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

Remark 9. We have by the definition of the marginal PMF

$$\mathbb{E}[X] = \sum_{(x,y)} x f_{X,Y}(x,y) = \sum_x \sum_y x f_{X,Y}(x,y) = \sum_x x \sum_y p_{X,Y}(x,y) = \sum_x x p_X(x)$$

so $\mathbb{E}[X]$ coincides with the expected value for single random variables we saw before.

Problem 1.12. If X and Y are independent discrete random variables, show that

$$\mathbb{E}[g_1(X) g_2(Y)] = \mathbb{E}[g_1(X)]\, \mathbb{E}[g_2(Y)].$$

In particular, if $g_1 = x$ and $g_2 = y$ then

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

Solution 1.12. Since $p_{X,Y}(x,y) = p_X(x) p_Y(y)$ by independence, we have by the definition of the expected value,

$$\begin{aligned}
\mathbb{E}[g_1(X) g_2(Y)] &= \sum_{(x,y)} (g_1(x) g_2(y)) f_{X,Y}(x,y) \\
\text{independence} &= \sum_{(x,y)} g_1(x) g_2(y) f_X(x) f_Y(y) \\
&= \left(\sum_x g_1(x) f_X(x) \right) \left(\sum_y g_2(y) f_Y(y) \right) = \mathbb{E}[g_1(X)]\, \mathbb{E}[g_2(Y)].
\end{aligned}$$

By taking $g_1(x) = x$ and $g_2(y) = y$ we immediately arrive at the fact that

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

A similar proof holds for continuous random variables. The integration is used instead of summation and we apply Fubini's theorem to split the sum into two parts.

Problem 1.13. Prove Proposition 1.

Solution 1.13. Independence follows immediately if $g(x)$ and $h(y)$ integrated to 1, since in that case we can take $g(x) = f_X(x)$ and $h(y) = f_Y(y)$. However, if we only know that

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

then we don't know that g and h integrate to 1. However, we can normalize the functions so that they always integrate to 1 without loss of generality. We define $c = \int_{-\infty}^{\infty} h(y)\,dy$ and consider the functions $\tilde{g}(x) = cg(x)$ and $\tilde{h}(y) = \frac{h(y)}{c}$

$$f_{X,Y}(x,y) = g(x)h(y) = cg(x)\frac{h(y)}{c} = \tilde{g}(x)\tilde{h}(y).$$

Notice that by definition of c,

$$\int_{-\infty}^{\infty} \tilde{h}(y)\,dy = \frac{1}{c}\int_{-\infty}^{\infty} h(y)\,dy = 1.$$

Likewise, by Fubini's theorem and the fact that $f_{X,Y}$ is a joint PDF,

$$\int_{-\infty}^{\infty} \tilde{g}(x)\,dx = \underbrace{\left(\int_{-\infty}^{\infty} \tilde{h}(y)\,dy\right)}_{=1}\int_{-\infty}^{\infty} \tilde{g}(x)\,dx = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \underbrace{\tilde{g}(x)\tilde{h}(y)}_{=f_{X,Y}(x,y)}\,dxdy = 1.$$

This means that X and Y are independent with marginal PDFs $f_X = \tilde{g}$ and $f_Y = \tilde{h}$. The proof for discrete random variables is identical.

Problem 1.14. Suppose that X and Y are independent. Show that

$$\mathrm{Cov}(X,Y) = 0.$$

Show that the converse is false by providing a counterexample.

Solution 1.14. Suppose that X and Y are independent. We know that $\mathbb{E}[XY] = \mathbb{E}[X]\,\mathbb{E}[Y]$. Therefore, using the equivalent formula,

$$\mathrm{Cov}(X,Y) = \mathbb{E}[XY] - \mathbb{E}[X]\,\mathbb{E}[Y] = \mathbb{E}[X]\,\mathbb{E}[Y] - \mathbb{E}[X]\,\mathbb{E}[Y] = 0.$$

Counterexample: Let $X \sim \mathrm{Unif}(-1,1)$, and let $Y = X^2$. X and Y are not independent because

$$0 = \mathbb{P}\left(X > \frac{1}{2}, Y < \frac{1}{4}\right) \neq \mathbb{P}\left(X > \frac{1}{2}\right)\mathbb{P}\left(Y < \frac{1}{4}\right) > 0$$

since $X > \frac{1}{2} \implies X^2 > \frac{1}{4}$ so it is impossible that $Y = X^2 < \frac{1}{4}$ as well. However, we can compute the covariance,

$$\mathrm{Cov}(X,Y) = \mathbb{E}[XY] - \mathbb{E}[X]\,\mathbb{E}[Y] = \mathbb{E}[X^3] - \mathbb{E}[X]\,\mathbb{E}[X^2] = 0$$

since the PDF of X is symmetric, so $\mathbb{E}[X^3] = 0$ and $\mathbb{E}[X] = 0$.

Problem 1.15. Prove the Cauchy–Schwarz inequality,

$$|\,\mathbb{E}[XY]| \leq \sqrt{\mathbb{E}(X^2)}\sqrt{\mathbb{E}(Y^2)}.$$

Equality holds if and only if $Y = aX$ for some constant a.

Solution 1.15. Notice that the statement is trivial if either $X = 0$ or $Y = 0$, so we consider the non-trivial cases.

For any $t \in \mathbb{R}$, we have

$$0 \leq \mathbb{E}[(tX - Y)^2] = at^2 - 2bt + c$$

where $a = \mathbb{E}[X^2]$, $b = \mathbb{E}[XY]$ and $c = \mathbb{E}[Y^2]$. A quadratic polynomial $at^2 - 2bt + c$ is non-negative if and only if it has at most one root, which happens if the discriminant satisfies

$$D = 4b^2 - 4ac \leq 0 \implies b^2 \leq ac \implies |b| \leq \sqrt{ac}$$

so $|\mathbb{E}[XY]| \leq \sqrt{\mathbb{E}[X^2]}\sqrt{\mathbb{E}[Y^2]}$. This proves the first part of the statement.

We now consider the equality case. Suppose now that we have equality $|\mathbb{E}[XY]| = \sqrt{\mathbb{E}(X^2)}\sqrt{\mathbb{E}(Y^2)}$, so $|b| = \sqrt{ac}$. This implies that $D = 0$, so the quadratic polynomial has exactly one real root. Let $\lambda = \frac{b}{a}$ denote the value of this root, so

$$\mathbb{E}[(\lambda X - Y)^2] = a\lambda^2 - 2b\lambda + c = 0.$$

We have that $\mathbb{E}[(\lambda X - Y)^2] = 0$ if and only if $\lambda X - Y = 0$ with probability one, so $Y = \lambda X = \frac{\mathbb{E}[XY]}{\mathbb{E}[X^2]}X$ with probability 1. Therefore, if $X \neq aY$ for any a, then $Y \neq \frac{\mathbb{E}[XY]}{\mathbb{E}[X^2]}X$ so $|\mathbb{E}[XY]| \neq \sqrt{\mathbb{E}(X^2)}\sqrt{\mathbb{E}(Y^2)}$.

To prove the converse, suppose that $X = aY$. We have

$$|\mathbb{E}[XY]| = |a||\mathbb{E}[Y^2]| = \sqrt{\mathbb{E}((aY)^2)}\sqrt{\mathbb{E}(Y^2)} = |a|\,\mathbb{E}[Y^2]$$

so equality holds.

Problem 1.16. Show that $\rho = \mathrm{Corr}(X,Y)$ satisfies $|\rho| \leq 1$ and $|\rho| = 1$ if and only $Y = aX + b$ for some constants a and b.

Solution 1.16. By the Cauchy–Schwarz inequality, applied to $X - \mathbb{E}[X]$ and $Y - \mathbb{E}[Y]$ we have

$$|\mathrm{Cov}(X,Y)| = |\mathbb{E}[(X-\mathbb{E}[X])(Y-\mathbb{E}[Y])]| \leq \sqrt{\mathbb{E}[(X-\mathbb{E}[X])^2]}\sqrt{\mathbb{E}[(Y-\mathbb{E}[Y])^2]} = \sqrt{\mathrm{Var}(X)}\sqrt{\mathrm{Var}(Y)}.$$

Rearranging terms implies that

$$|\mathrm{Corr}(X,Y)| = \frac{|\mathrm{Cov}(X,Y)|}{\sqrt{\mathrm{Var}(X)}\sqrt{\mathrm{Var}(Y)}} \leq 1.$$

Next, we have that equality happens if and only if $Y - \mathbb{E}[Y] = a(X - \mathbb{E}[X])$ for some constant a. This means that there must be a linear relation between Y and X if equality were to hold. To see that any linear relation achieves equality, suppose that $Y = aX + b$ for some constants a and b, so by bilinearity

$$|\mathrm{Cov}(X,Y)| = |\mathrm{Cov}(X, aX + b)| = |a\mathrm{Cov}(X,X) + \mathrm{Cov}(X,b))| = |a|\,\mathrm{Var}(X)$$

and

$$\sqrt{\mathrm{Var}(Y)} = \sqrt{\mathrm{Var}(aX+b)} = |a|\sqrt{\mathrm{Var}(X)},$$

so

$$|\mathrm{Corr}(X,Y)| = 1.$$

Remark 10. We can repeat the second computation without the absolute values to conclude that $\mathrm{Corr}(X,Y) = 1$ implies that $Y = aX + b$ for some constant $a > 0$ and $\mathrm{Corr}(X,Y) = -1$ implies that $Y = aX + b$ for some constant $a < 0$

Problem 1.17. Prove the bilinearity property of covariances

$$\mathrm{Cov}\left[\sum_{i=1}^n a_i X_i, \sum_{i=1}^n b_i Y_i\right] = \sum_{i=1}^n \sum_{j=1}^m a_i b_j \mathrm{Cov}(X_i, Y_j).$$

Solution 1.17. This is a direct consequence of linearity of expectation and the distributive property of numbers

$$\sum_{i=1}^n a_i \times \sum_{j=1}^m b_j = \sum_{i=1}^n \sum_{j=1}^m a_i b_j.$$

By the definition of the covariance,

$$\begin{aligned}
\mathrm{Cov}\left[\sum_{i=1}^n a_i X_i, \sum_{j=1}^n b_j Y_j\right] &= \mathbb{E}\left[\left(\sum_{i=1}^n a_i X_i - \mathbb{E}\left[\sum_{i=1}^n a_i X_i\right]\right)\left(\sum_{i=1}^m b_i Y_i - \mathbb{E}\left[\sum_{j=1}^m b_j Y_j\right]\right)\right] \\
\text{linearity of expectation} &= \mathbb{E}\left[\left(\sum_{i=1}^n a_i (X_i - \mathbb{E}[X_i])\right)\left(\sum_{j=1}^m b_j (Y_j - \mathbb{E}[Y_j])\right)\right] \\
\text{distributive property} &= \mathbb{E}\left[\sum_{i=1}^n \sum_{j=1}^m a_i b_J (X_i - \mathbb{E}[X_i])(Y_j - \mathbb{E}[Y_j])\right] \\
\text{linearity of expectation} &= \sum_{i=1}^n \sum_{j=1}^m a_i b_j \, \mathbb{E}[(X_i - \mathbb{E}[X_i])(Y_j - \mathbb{E}[Y_j])] = \sum_{i=1}^n \sum_{j=1}^m a_i b_j \mathrm{Cov}(X_i, Y_j).
\end{aligned}$$

Problem 1.18. Prove the formula for the variance of linear combinations of random variables,

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

Solution 1.18. Since $\mathrm{Var}(X) = \mathrm{Cov}(X, X)$, the proof follows directly from the bilinearity of covariance. We have

$$\begin{aligned}
\mathrm{Var}\left(\sum_{i=1}^n a_i X_i\right) &= \mathrm{Cov}\left[\sum_{i=1}^n a_i X_i, \sum_{i=1}^n a_i X_i\right] \\
\text{bilinearity} &= \sum_{i,j=1}^n a_i a_j \mathrm{Cov}(X_i, X_j) \\
\text{split into diagonal and offdiagonal} &= \sum_{i}^n a_i^2 \mathrm{Cov}(X_i, X_i) + \sum_{i \neq j} a_i a_j \mathrm{Cov}(X_i, X_j) \\
\mathrm{Cov}(X,Y) = \mathrm{Cov}(Y,X), \mathrm{Var}(X) = \mathrm{Cov}(X,X) &= \sum_{i}^n a_i^2 \, \mathrm{Var}(X_i) + 2 \sum_{i<j} a_i a_j \mathrm{Cov}(X_i, X_j).
\end{aligned}$$

Problem 1.19. If $X \sim \mathrm{Poi}(\lambda_1)$ and $Y \sim \mathrm{Poi}(\lambda_2)$ are independent, show that

$$T = X + Y \sim \mathrm{Poi}(\lambda_1 + \lambda_2).$$

Solution 1.19. We have $X + Y = n$ if and only if $X = m$ and $Y = n - m$ for $m = 0, 1, \ldots, n$. Therefore,

$$\begin{aligned}
p_T(n) = \mathbb{P}(X+Y=n) &= \sum_{(x,y):x+y=n} \mathbb{P}(X=x, Y=y) \\
&= \sum_{m=0}^{n} \mathbb{P}(X=m, Y=n-m) \\
\text{independence} \quad &= \sum_{m=0}^{n} \mathbb{P}(X=m)\,\mathbb{P}(Y=n-m) \\
&= \sum_{m=0}^{n} e^{-\lambda_1}\frac{\lambda_1^m}{m!} e^{-\lambda_2}\frac{\lambda_2^{n-m}}{(n-m)!} \\
&= \frac{e^{-(\lambda_1+\lambda_2)}}{n!}\sum_{m=0}^{n}\frac{n!}{m!(n-m)!}\lambda_1^m\lambda_2^{n-m} \\
\text{Binomial thm} \quad &= \frac{e^{-(\lambda_1+\lambda_2)}}{n!}(\lambda_1+\lambda_2)^n.
\end{aligned}$$

Problem 1.20. Let $X \sim \mathrm{Poi}(\lambda_1)$ and $Y \sim \mathrm{Poi}(\lambda_2)$ be independent. Show that

$$X \mid X+Y=n \sim \mathrm{Bin}\left(n, \frac{\lambda_1}{\lambda_1+\lambda_2}\right).$$

Similarly, for Y, we have

$$Y \mid X+Y=n \sim \mathrm{Bin}\left(n, \frac{\lambda_2}{\lambda_1+\lambda_2}\right)$$

Solution 1.20. Since $X + Y \sim \mathrm{Poi}(\lambda_1 + \lambda_2)$, we have

$$\begin{aligned}
p_{X \mid X+Y} = \frac{p_{X,X+Y}(x,n)}{p_{X+Y}(n)} &= \frac{\mathbb{P}(X=x, X+Y=n)}{\mathbb{P}(X+Y=n)} \\
\text{independence} \quad &= \frac{\mathbb{P}(X=x)\,\mathbb{P}(Y=n-x)}{\mathbb{P}(X+Y=n)} \\
&= \frac{e^{-\lambda_1}\frac{\lambda_1^x}{x!}e^{-\lambda_2}\frac{\lambda_2^{n-x}}{(n-x)!}}{e^{-(\lambda_1+\lambda_2)}\frac{(\lambda_1+\lambda_2)^n}{n!}} \\
&= \frac{n!}{x!(n-x)!}\left(\frac{\lambda_1}{\lambda_1+\lambda_2}\right)^x\left(\frac{\lambda_2}{\lambda_1+\lambda_2}\right)^{n-x}
\end{aligned}$$

which we recognize as the PMF of a $\mathrm{Bin}\left(n, \frac{\lambda_1}{\lambda_1+\lambda_2}\right)$ random variable. The proof for the Y given $X + Y = n$ is identical.

Problem 1.21. Let $X_i \sim \mathrm{N}(\mu_i, \sigma_i^2),\ i = 1, 2, \ldots, n$ be independent then,

$$\sum_{i=1}^{n}(a_iX_i + b_i) \sim \mathrm{N}\left(\sum_{i=1}^{n} a_i\mu_i + b_i, \sum_{i=1}^{n} a_i^2\sigma_i^2\right).$$

Solution 1.21. The easiest proof of this fact uses moment generating functions, which will be explained a later week. In the meantime, we will present a geometric proof of this result using a multidimensional change of variables. For simplicity, we consider the case that $n = 2$ and the means of the random variables are 0 and $a_i = 1$ and $b_i = 0$. The general case can be recovered by an induction argument and a standardization argument.

Let $X_1 \sim \mathrm{N}(0, \sigma_1^2)$ and $X_2 \sim \mathrm{N}(0, \sigma_2^2)$ be independent. We need to show that

$$X_1 + X_2 \sim \mathrm{N}(0, \sigma_1^2 + \sigma_2^2).$$

By standardization, $X_1 = \sigma_1 Z_1$ and $X_2 = \sigma_2 Z_2$. We want to find the PDF of

$$F_{X_1+X_2}(t) = \mathbb{P}(\sigma_1 Z_1 + \sigma_2 Z_2 \leq t) = \frac{1}{2\pi} \iint_{\sigma_1 z_1 + \sigma_2 z_2 \leq t} e^{-\frac{z_1^2+z_2^2}{2}} dz_1 dz_2.$$

Using the change of variables corresponding to the rotation of the half plane to one perpendicular to the w_1 axis,

$$w_1 = \frac{\sigma_1}{\sqrt{\sigma_1^2+\sigma_2^2}} z_1 + \frac{\sigma_2}{\sqrt{\sigma_1^2+\sigma_2^2}} z_2 \qquad w_2 = \frac{-\sigma_2}{\sqrt{\sigma_1^2+\sigma_2^2}} z_1 + \frac{\sigma_1}{\sqrt{\sigma_1^2+\sigma_2^2}} z_2$$

we have

$$\{\sigma_1 z_1 + \sigma_2 z_2 \leq t\} = \left\{ w_1 \leq \frac{t}{\sqrt{\sigma_1^2+\sigma_2^2}} \right\}$$

and

$$dw_1 dw_2 = \begin{vmatrix} \frac{\sigma_1}{\sqrt{\sigma_1^2+\sigma_2^2}} & \frac{\sigma_2}{\sqrt{\sigma_1^2+\sigma_2^2}} \\ -\frac{\sigma_2}{\sqrt{\sigma_1^2+\sigma_2^2}} & \frac{\sigma_1}{\sqrt{\sigma_1^2+\sigma_2^2}} \end{vmatrix} dz_1 dz_2 = dz_1 dz_2$$

so

$$\frac{1}{2\pi} \iint_{\sigma_1 z_1 + \sigma_2 z_2 \leq t} e^{-\frac{z_1^2+z_2^2}{2}} dz_1 dz_2 = \frac{1}{2\pi} \iint_{w_1 \leq \frac{t}{\sqrt{\sigma_1^2+\sigma_2^2}}} e^{-\frac{w_1^2+w_2^2}{2}} dw_1 dw_2 = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\frac{t}{\sqrt{\sigma_1^2+\sigma_2^2}}} e^{-\frac{w_1^2}{2}} dw_1.$$

We conclude that

$$F_{X_1+X_2}(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\frac{t}{\sqrt{\sigma_1^2+\sigma_2^2}}} e^{-\frac{w_1^2}{2}} dw_1 = \mathbb{P}\left(Z \leq \frac{t}{\sqrt{\sigma_1^2+\sigma_2^2}} \right) = \mathbb{P}(\sqrt{\sigma_1^2+\sigma_2^2} Z \leq t)$$

which is the CDF of a $\mathrm{N}(0, \sigma_1^2 + \sigma_2^2)$ random variable.

Remark 11. Essentially we have exploited the rotational invariance of the standard normal. Notice that the PDF is rotational symmetric since

$$f_{Z_1,Z_2}(z_1, z_2) = \frac{1}{2\pi} e^{-\frac{z_1^2+z_2^2}{2}} = \frac{1}{2\pi} e^{-\frac{r^2}{2}} = f_{Z_1,Z_2}(r)$$

is a function of $r = \sqrt{z_1^2 + z_2^2}$, which means that the density only depends on the distance from points to the origin. The special change of variables rotated the half plane $\{\sigma_1 z_1 + \sigma_2 z_2 \leq t\}$ to be perpendicular to the w_1 axis to reduce the problem to computing the probability $\sqrt{\sigma_1^2 + \sigma_2^2} Z \leq t$.

Problem 1.22. Prove Proposition 2, that is the covariance matrix is a positive semidefinite matrix.

Solution 1.22. Since $\mathrm{Cov}(X^i, X^j) = \mathrm{Cov}(X^j, X^i)$, the matrix $\boldsymbol{C} = \mathrm{Cov}(\mathbf{X})$ must be **symmetric**. Moreover, if $\mathbf{y} = (y_1, \dots, y_n)^\top \in \mathbb{R}^n$, then

$$\mathbf{y}^\top \boldsymbol{C} \mathbf{y} = \mathbb{E}(\mathbf{y}^\top (\mathbf{X} - \mathbb{E}[\mathbf{X}])(\mathbf{X} - \mathbb{E}[\mathbf{X}])^\top \mathbf{y}) = \mathbb{E}(\mathbf{y}^\top (\mathbf{X} - \mathbb{E}[\mathbf{X}]))^2 = \mathrm{Var}(\mathbf{y}^\top \mathbf{X}) \geq 0,$$

and so $\boldsymbol{C}$ has to be **positive semidefinite**.

2 Appendix: Important Multivariate Distributions

2.1 Multinomial Distribution

The multinomial distribution models the number of each outcome in multiple independent experiments with k possible outcomes. The multinomial distribution is a generalization of the binomial distribution.

Definition 16 (Multinomial Distribution)**.** Consider an experiment in which:

1. Individual trials have k possible outcomes, and the probabilities of each individual outcome are denoted $p_i,\ 1 \leq i \leq k$, so that $p_1 + p_2 + \cdots + p_k = 1$.

2. Trials are independently repeated n times, with X_i denoting the number of times outcome i occurred, so that $X_1 + X_2 + \cdots + X_k = n$.

We say that $X_1, ..., X_k$ has a *multinomial distribution* with parameters n and $p_1, ..., p_k$, if $\boldsymbol{X}$ has joint PMF

$$p_{X_1,...,X_k}(x_1, ..., x_k) = \frac{n!}{x_1!x_2!\cdots x_k!} p_1^{x_1} \cdots p_k^{x_k} = \binom{n}{x_1, \ldots, x_k} p_1^{x_1} \cdots p_k^{x_k},$$

and is denoted by

$$\mathbf{X} = (X_1, ..., X_k) \sim \mathrm{Mult}(n, p_1, ..., p_k).$$

The terms $\frac{n!}{x_1!x_2!\cdots x_k!} = \binom{n}{x_1,\ldots,x_k}$ are called multinomial coefficients.

Remark 12. Since we must have $p_1 + p_2 + \cdots + p_k = 1$ and $X_1 + X_2 + \cdots + X_k = n$, the kth variable is uniquely determined by the first $k-1$ variables,

$$p_k = 1 - p_1 - p_2 - \ldots - p_{k-1} \qquad \text{and} \qquad x_k = n - x_1 - x_2 - \ldots - x_{k-1}$$

so the joint PMF is sometimes written as

$$p_{X_1,...,X_{k-1}}(x_1, ..., x_{k-1}) = \frac{n!}{x_1!x_2!\cdots x_{k-1}!(n - \sum_{i=1}^{k-1} x_i)!} p_1^{x_1} \cdots p_{k-1}^{x_{k-1}} \left(1 - \sum_{i=1}^{k-1} p_i\right)^{n - \sum_{i=1}^{k-1} x_i}$$

Remark 13. Notice that when $k = 2$, then we have the PMF of the binomial distribution.

Example 1. The following experiments can be modeled by a multinomial distribution

Experiment	X	Distribution
Draw 10 cards from a deck with replacement	# of each suit	$\mathrm{Mult}(10, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$
Roll a dice n times	# of each roll	$\mathrm{Mult}(n, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6})$
20 customers order from a menu of 3 items	# of each item	$\mathrm{Mult}(20, p_1, p_2, p_3)$

2.1.1 Properties

1. **Marginal PMF:** The number of times the outcome i occurred is

$$X_j \sim \mathrm{Bin}(n, p_j), \quad \text{for } j = 1, 2, \ldots, k\ .$$

2. **Sum of Marginals:** The number of times the outcomes i or j occurred is

$$X_i + X_j \sim \mathrm{Bin}(n, p_i + p_j), \quad \text{for } i \neq j.$$

3. **Expected Values:** The expected value of the outcomes are given by

$$\mathbb{E}[X_i X_j] = n(n-1)p_i p_j \text{ for } i \neq j \qquad \text{and} \qquad \mathbb{E}[X_i] = np_i \text{ for } i = 1, \ldots, k$$

2.2 Multivariate Normal

The multivariate normal is the most commonly seen multivariate distribution in statistics, data science and many other applied fields. It models the behavior of the sums of i.i.d. random vectors. It is the multivariate generalization of the normal distribution.

Definition 17 (Multivariate Normal Distribution)**.** We say that the vector $\boldsymbol{X} = (X_1, X_2, \dots, X_n) \in \mathbb{R}^n$ has a *multivariate normal distribution* with mean $\boldsymbol{\mu} = (\mu_1, \dots, \mu_n)$ and positive definite covariance matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{n\times n}$ if X has joint PDF

$$f_{\boldsymbol{X}}(\boldsymbol{x}) = \frac{1}{(2\pi)^{\frac{n}{2}}} \frac{1}{\sqrt{\det(\Sigma)}} e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^{\intercal}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})}, \quad \boldsymbol{x} \in \mathbb{R}^n.$$

This is denoted by

$$\boldsymbol{X} \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}).$$

Remark 14. In the special case when $n = 1$, the above density is equivalent to the univariate normal we saw before.

A special case of the multivariate normal is when all of its entries are i.i.d. standard normal.

Definition 18 (Standard Normal Vector)**.** We say that the vector $\boldsymbol{X} = (X_1, X_2, \dots, X_n) \in \mathbb{R}^n$ has a *standard normal distribution* if it has a multivariate normal distribution with mean $\boldsymbol{\mu} = (0, 0, \dots, 0)$ and covariance matrix $\boldsymbol{I}$. In this case, $\boldsymbol{X}$ has density

$$f_{\boldsymbol{X}}(\boldsymbol{x}) = \frac{1}{(2\pi)^{\frac{n}{2}}} e^{-\frac{1}{2}\|\boldsymbol{x}\|_2^2} = \frac{1}{(2\pi)^{\frac{n}{2}}} e^{-\frac{1}{2}\sum_{i=1}^n x_i^2}, \quad \boldsymbol{x} \in \mathbb{R}^n.$$

and is denoted by

$$\boldsymbol{X} \sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{I}).$$

Remark 15. It follows immediately that if $\boldsymbol{X} = (X_1, X_2, \dots, X_n) \sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{I})$, then its joint PDF is simply the product of n standard normal PDFs,

$$f_{\boldsymbol{X}}(\boldsymbol{x}) = \prod_{i=1}^n f_Z(x_i) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x_i^2}$$

so the formal definition is consistent with the interpretation of X having i.i.d. standard normal entries.

2.2.1 Constructing the Multivariate Normal

Recall that in the univariate case, we can construct the normal distribution $X \sim \mathrm{N}(\mu, \sigma^2)$ by taking linear transformation of a standard normal distribution $Z \sim \mathrm{N}(0, 1)$ through a method called (de)-standardizing,

$$X \overset{d}{=} \sigma Z + \mu.$$

The mean and variance uniquely determine the normal distribution, and one can easily check that

$$\mathbb{E}[\sigma Z + \mu] = \mu \qquad \text{and} \qquad \mathrm{Var}[\sigma Z + \mu] = \sigma^2.$$

The multivariate normal can be constructed in the same way. We can construct the normal distribution $\mathbf{X} \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \in \mathbb{R}^n$ by taking a linear transformation of a standard normal distribution $\mathbf{Z} \sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{I})$

$$\mathbf{X} \overset{d}{=} \boldsymbol{\Sigma}^{\frac{1}{2}}\mathbf{Z} + \boldsymbol{\mu}.$$

The mean and variance uniquely determine the normal distribution, and one can easily check (see Problem 2.11) that

$$\mathbb{E}[\boldsymbol{\Sigma}^{\frac{1}{2}}\mathbf{Z} + \boldsymbol{\mu}] = \boldsymbol{\mu} \qquad \text{and} \qquad \mathrm{Cov}[\boldsymbol{\Sigma}^{\frac{1}{2}}\mathbf{Z} + \boldsymbol{\mu}] = \boldsymbol{\Sigma}.$$

Remark 16. From this construction, it is possible to define a multivariate normal even if $\mathbf{\Sigma}$ is not positive definite since $\mathbf{\Sigma}^{\frac{1}{2}}$ is well defined whenever $\mathbf{\Sigma}$ is positive semidefinite. However, in this case the density of $\boldsymbol{X}$ will not be explicit.

2.2.2 Properties

1. **Marginal PDF:** If $\mathbf{X} \sim \mathrm{N}(\boldsymbol{\mu}, \mathbf{\Sigma})$, then distribution of the marginal X_i is normally distributed with mean μ_i and variance $\mathrm{Var}(X_i) = \mathrm{Cov}(X_i, X_i) = \Sigma_{ii}$,

$$X_i \sim \mathrm{N}(\mu_i, \Sigma_{ii}).$$

Remark 17. It is not true that if a vector $\mathbf{X} = (X_1, \dots, X_n)$ has marginals $X_i \sim \mathrm{N}(\mu_i, \sigma_i^2)$, then $\mathbf{X}$ is a multivariate normal in general. This is because knowing the marginal distribution isn't enough to recover the joint distribution.

2. **Stability:** The linear combination of independent normally distributed random variables are normally distributed. Let $X_i \sim \mathrm{N}(\mu_i, \sigma_i^2),\ i = 1, 2, \dots, n$ are independent then,

$$\sum_{i=1}^n (a_i X_i + b_i) \sim \mathrm{N}\left(\sum_{i=1}^n a_i \mu_i + b_i, \sum_{i=1}^n a_i^2 \sigma_i^2\right).$$

Remark 18. Notice that by linearity, we have

$$\mathbb{E}\left[\sum_{i=1}^n (a_i X_i + b_i)\right] = \sum_{i=1}^n (a_i\, \mathbb{E}[X_i] + b_i) = \sum_{i=1}^n a_i \mu_i + b_i$$

and by independence

$$\mathrm{Var}\left(\sum_{i=1}^n (a_i X_i + b_i)\right) = \mathrm{Var}\left(\sum_{i=1}^n a_i X_i\right) = \sum_{i=1}^n a_i^2\, \mathrm{Var}(X_i) = \sum_{i=1}^n a_i^2 \sigma_i^2$$

which precisely matches the mean and variance of the linear combination.

Remark 19. If $X \sim \mathrm{N}(\mu, \sigma^2)$ and $Y = aX + b$, where $a, b \in \mathbb{R}$, then we have the following standardization result

$$Y \sim \mathrm{N}(a\mu + b, a^2\sigma^2).$$

3. **Uncorrelated implies independent:** If (X, Y) is a bivariate normal and $\mathrm{Corr}(X, Y) = 0$, then X and Y are independent. This is generally not true if (X, Y) are any random variables.

4. **Covariance Characterization of PSD matrices:** The covariance matrix of any random vector is a positive semidefinite matrix.

   Conversely, if $\boldsymbol{C}$ is any given symmetric and positive semidefinite $n \times n$ matrix, then there exists an $n \times n$ matrix $\boldsymbol{A}$ such that $\boldsymbol{C} = \boldsymbol{A}\boldsymbol{A}^\top$. So $\boldsymbol{C}$ is the covariance matrix of some normally distributed random vector $\boldsymbol{X} = \boldsymbol{A}\boldsymbol{Z} + \boldsymbol{\mu}$ where $\boldsymbol{Z}$ is a standard Gaussian vector.

2.3 Example Problems

Problem 2.1. Let

$$(X_1, X_2, X_3) \sim \text{Mult}(10, 0.5, 0.3, 0.2).$$

Compute $\text{Cov}(X_1, X_2)$.

Solution 2.1. From the properties of the multinomial distribution, we know that if $(X_1, \dots, X_k) \sim \text{Mult}(n, p_1, \dots, p_k)$ then

$$\mathbb{E}[X_i X_j] = n(n-1)p_i p_j, \qquad \mathbb{E}[X_i] = np_i.$$

Applied to this problem using the equivalent formula for the covariance,

$$\text{Cov}(X_1, X_2) = \mathbb{E}[X_1 X_2] - \mathbb{E}[X_1]\,\mathbb{E}[X_2] = n(n-1)p_1p_2 - np_1np_2 = -np_1p_2 = -10 \cdot 0.5 \cdot 0.3 = -1.5.$$

Problem 2.2. Consider drawing 5 cards from a standard 52 card deck of playing cards (4 suits, 13 kinds) **with replacement**. What is the probability that 2 of the drawn cards are hearts, 2 are spades, and 1 is a diamond?

Solution 2.2. Denote by H, S, D, C the number of Hearts, Spades, Diamonds, and Clubs. Then

$$(H, S, D, C) \sim \text{Mult}(5, 0.25, 0.25, 0.25, 0.25)$$

and

$$\mathbb{P}(H = 2, S = 2, D = 1, C = 0) = \frac{5!}{2!2!1!0!}\left(\frac{1}{4}\right)^5$$

Problem 2.3. In a manufacturing process, two pieces of metal are combined to form a new piece of metal. Due to variations in the production process, we assume that the lengths of the two pieces, say L_1 and L_2, follow continuous uniform distributions as $L_1 \sim \text{Unif}(0.9, 1.1)$ and $L_2 \sim \text{Unif}(1.5, 1.7)$. Furthermore, due to variations joining process of the two pieces, the length of the new piece is not exactly $L_1 + L_2$, but instead $L = L_1 + L_2 + \varepsilon$ where $\varepsilon \sim N(0, 0.1^2)$. Compute the expected total length, $\mathbb{E}(L)$.

Solution 2.3. From the formula sheet, we see $\mathbb{E}(L_1) = \frac{0.9+1.1}{2} = 1$, $\mathbb{E}(L_2) = \frac{1.5+1.7}{2} = 1.6$ and $\mathbb{E}(\varepsilon) = 0$. By linearity,

$$\mathbb{E}(L) = \mathbb{E}(L_1 + L_2 + \varepsilon) = \mathbb{E}(L_1) + \mathbb{E}(L_2) + \mathbb{E}(\varepsilon) = 1 + 1.6 + 0 = 2.6.$$

Problem 2.4. In a certain cooking process, the target temperature, say C, follows a normal distribution (in celsius) with mean 57 and standard deviation 2. Your American friend asks you: What is the distribution of the target temperature in Fahrenheit?

Aside: The relationship between the temperature in Celsius c and Fahrenheit f is $f = c \cdot 9/5 + 32$.

Solution 2.4. By the stability property, $C \cdot 9/5 + 32 \sim \text{N}(57 \cdot 9/5 + 32, (9/5)^2 \cdot 2^2) = \text{N}(134.6, 12.96)$.

Problem 2.5. Let $X \sim \text{N}(\mu_1, \sigma^2)$ be independent of $Y \sim \text{N}(\mu_2, \sigma^2)$. What is the distribution of $X - Y$?

Solution 2.5. By the stability property, we have $\mathbb{E}[X-Y]=\mu_1-\mu_2$ and $\mathrm{Var}(X-Y)=\mathrm{Var}(X)+\mathrm{Var}(Y)=2\sigma^2$, so

$$X-Y\sim \mathrm{N}(\mu_1-\mu_2,2\sigma^2).$$

Problem 2.6. Three cylindrical parts are joined end to end to make up a shaft in a machine: 2 type-A parts and 1 type-B part. The lengths of the parts vary a little, and have the following distributions:

$$A\sim N(6,0.4),\quad B\sim N(35.2,0.6).$$

The overall length of the assembled shaft must lie between 46.8 and 47.5 or else the shaft has to be scrapped. Assume the lengths of different parts are independent. What percentage of assembled shafts has to be scrapped?

Solution 2.6. Let A_1,A_2 and B denote the three independent parts. The total length is $L=A_1+A_2+B$ satisfies

$$L\sim N(6+6+35.2,0.4+0.4+0.6)\quad\Rightarrow\quad L\sim N(47.2,1.4)$$

The part is scrapped if $L<46.8$ or $L>47.5$, so

$$\begin{aligned}\mathbb{P}(\text{“scrapped”})&=\mathbb{P}(L<46.8)+\mathbb{P}(L>47.5)\\&=\mathbb{P}\left(Z<\frac{46.8-47.2}{\sqrt{1.4}}\right)+\mathbb{P}\left(Z>\frac{47.5-47.2}{\sqrt{1.4}}\right)\\&=F_Z(-0.37)+(1-F_Z(0.27))\\&=(1-F_Z(0.37))+(1-F_Z(0.27))\\&=0.749.\end{aligned}$$

Remark 20. A common mistake is to say that A_1+A_2+B is the same as $L=2A_1+B$ (A_1 and A_2 have the same distribution, after all), and conclude

$$L=2A_1+B\sim N(2\cdot 6+35.2,2^2\cdot 0.4+0.6)\Rightarrow L\sim N(47.2,2.2).$$

The linearity of expectation (which holds even if the random variables are dependent) is not affected by this mistake; but the variance is affected by this mistake. This is because A_1+A_2 and $2A_1$ are very different objects since the first is a sum of two independent random variables and the latter is the sum of two very dependent random variables.

Problem 2.7. Let $X_1,\dots,X_n$ be independent and $X_i\sim N(\mu,\sigma^2)$ for all $i=1,\dots,n$. Show that

$$\bar{X}_n=\frac{1}{n}\sum_{i=1}^n X_i\sim N\left(\mu,\frac{\sigma^2}{n}\right).$$

Solution 2.7. By the Gaussian stability, we have

$$\bar{X}_n=\frac{1}{n}\sum_{i=1}^n X_i$$

is normally distributed. We just have to compute the mean and variance. By linearity,

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

and the variance of a linear combination (the covariance is 0 by independence) gives us

$$\mathrm{Var}(\bar{X}_n) = \sum_{i=1}^{n} \frac{1}{n^2} \mathrm{Var}(X_i) = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}.$$

Remark 21. As n increases, the variance σ^2/n decreases, so the distribution of $\bar{X}_n$ becomes more concentrated around μ. This is intuitive because if for example, you want to estimate the average of the midterm (or any other event that is normally distributed), then asking 5 people gives us a less reliable result than asking 50 people.

Problem 2.8. Suppose that the height of adult males in Canada is normally distributed with a mean of 70 inches and variance of 4^2 inches, and let $X_1, ..., X_{10}$ denote the heights of a random sample of adult males. Suppose $\bar{X}_{10}$ denotes the sample mean of these heights.

Let

$$p_1 = \mathbb{P}(68 \le X_1 \le 72)$$

and

$$p_{10} = \mathbb{P}(68 \le \bar{X}_{10} \le 72).$$

Which of the following is true?

1. $p_1 > p_{10}$
2. $p_1 = p_{10}$
3. $p_1 < p_{10}$

Solution 2.8. The interval contains the mean, so this result should be intuitive because a larger sample means less variance, so p_{10} should be bigger. To reinforce this, we can compute this explicitly.

We find

$$\begin{aligned}
p_1 &= \mathbb{P}(68 \le X_1 \le 72) \\
&= \mathbb{P}\left(\frac{68-70}{4} \le Z \le \frac{72-70}{4}\right), \quad Z \sim N(0,1) \\
&= F_Z(0.5) - F_Z(-0.5) = 2F_Z(0.5) - 1 \\
&= 2 \cdot 0.69146 - 1 = 0.38292
\end{aligned}$$

Next,

$$\bar{X}_{10} = \frac{1}{10}\sum_{i=1}^{10} X_i \sim N\left(\frac{1}{10}\sum_{i=1}^{10} 70, \frac{1}{10^2}\sum_{i=1}^{10} 4^2\right) \Rightarrow \bar{X}_{10} \sim N(70, 1.6)$$

so

$$\begin{aligned}
p_{10} &= \mathbb{P}(68 \le \bar{X}_{10} \le 72) \\
&= \mathbb{P}\left(\frac{68-70}{\sqrt{1.6}} \le Z \le \frac{72-70}{\sqrt{1.6}}\right), \quad Z \sim \mathrm{N}(0,1) \\
&= F_Z(1.58) - F_Z(-1.58) = 2F_Z(1.58) - 1 \\
&= 2 \cdot 0.94295 - 1 = 0.8859
\end{aligned}$$

2.4 Proofs of Key Results

Problem 2.9. If $(X_1, \dots, X_n) \sim \text{Mult}(n, p_1, \dots, p_k)$, show that

1. $$X_j \sim \text{Bin}(n, p_j), \quad \text{for } j = 1, 2, \dots, k\ .$$
2. $$X_i + X_j \sim \text{Bin}(n, p_i + p_j), \quad \text{for } i \neq j.$$
3. $$\mathbb{E}[X_i X_j] = n(n-1)p_i p_j \quad \text{for } i \neq j.$$

Solution 2.9.

Part 1: By definition, X_j denotes the number of occurrences of outcome j in n trials and each occurence has probability p_j of happening so

$$X_j \sim \text{Bin}(n, p_j), \quad \text{for } j = 1, 2, \dots, k\ .$$

Part 2: By definition, $X_i + X_j$ denotes the number of occurrences of outcome i or j in n trials and the probability of either i or j happening is $p_i + p_j$ so

$$X_i + X_j \sim \text{Bin}(n, p_i + p_j), \quad \text{for } i \neq j.$$

Part 3: We need to compute (noting that $x_i + x_j \leq n$ needs to hold):

$$\begin{aligned}
\mathbb{E}[X_i X_j] &= \sum_{\substack{x_i \geq 0, x_j \geq 0 \\ x_i + x_j \leq n}} x_i \cdot x_j \cdot \frac{n!}{x_i! x_j! (n - x_i - x_j)!} p_i^{x_i} p_j^{x_j} (1 - p_i - p_j)^{n - x_i - x_j} \\
&= \sum_{\substack{x_i \geq 1, x_j \geq 1 \\ x_i + x_j \leq n}} x_i \cdot x_j \cdot \frac{n!}{x_i! x_j! (n - x_i - x_j)!} p_i^{x_i} p_j^{x_j} (1 - p_i - p_j)^{n - x_i - x_j} \\
&= \sum_{\substack{x_i \geq 1, x_j \geq 1 \\ x_i + x_j \leq n}} \frac{n!}{(x_i - 1)! (x_j - 1)! (n - x_i - x_j)!} p_i^{x_i} p_j^{x_j} (1 - p_i - p_j)^{n - x_i - x_j}
\end{aligned}$$

Like in the computation of the expected value of a binomial, we factor out terms to make the summation look like the sum of a PMF,

$$\begin{aligned}
&= n(n-1)p_i p_j \sum_{\substack{x_i - 1 \geq 0, x_j - 1 \geq 0 \\ x_i - 1 + x_j - 1 \leq n - 2}} \frac{(n-2)! \times p_i^{x_i - 1} p_j^{x_j - 1} (1 - p_i - p_j)^{n - 2 - (x_i - 1) - (x_j - 1)}}{(x_i - 1)! (x_j - 1)! (n - 2 - (x_i - 1) - (x_j - 1))!} \\
&= n(n-1)p_i p_j \underbrace{\sum_{\substack{y_i \geq 0, y_j \geq 0 \\ y_i + y_j \leq n - 2}} \frac{(n-2)!}{(y_i)! (y_j)! (n - 2 - y_i - y_j)!} p_i^{y_i} p_j^{y_j} (1 - p_i - p_j)^{n - 2 - y_i - y_j}}_{=1 \text{ Sum of PMF of Mult}(n-2, p_i, p_j, 1 - p_i - p_j)} \\
&= n(n-1)p_i p_j
\end{aligned}$$

where we used the change of variables $y_i = x_i - 1$, $y_j = x_j - 1$.

Alternative Proof: We can compute the expected value using linearity of expectation. We can write $X_i = \sum_{k=1}^n \mathbb{1}_{A_k}$ where A_k is the event that outcome i occured on the kth trial, and

$$\mathbb{1}_{A_k} = \begin{cases} 1 & A_k \text{ happens} \\ 0 & A_k \text{ does not happens.} \end{cases}$$

Similarly, $X_j = \sum_{\ell=1}^n \mathbb{1}_{B_\ell}$ where B_ℓ is the event that outcome j occured on the ℓth trial, and

$$\mathbb{1}_{B_\ell} = \begin{cases} 1 & B_\ell \text{ happens} \\ 0 & B_\ell \text{ does not happens.} \end{cases}$$

Therefore,

$$\mathbb{E}[X_i X_j] = \mathbb{E}\left[\sum_{k=1}^n \mathbb{1}_{A_k} \sum_{\ell=1}^n \mathbb{1}_{B_\ell}\right] = \sum_{k,\ell=1}^n \mathbb{E}\left[\mathbb{1}_{A_k}\mathbb{1}_{B_\ell}\right].$$

We have two cases

1. $k = \ell$: Suppose that $k = \ell$. Since $\mathbb{1}(A_k)\mathbb{1}(B_k) = 1$ if and only if A_k and B_k happen, we have
$$\mathbb{E}[\mathbb{1}_{A_k}\mathbb{1}_{B_\ell}] = \mathbb{E}[\mathbb{1}_{A_k}\mathbb{1}_{B_k}] = \mathbb{P}(A_k \cap B_k) = 0$$
since both outcome i and j can't happen at the same time.

2. $k \neq \ell$: Suppose that $k \neq \ell$. Since $\mathbb{1}_{A_k}\mathbb{1}_{B_\ell} = 1$ if and only if A_k and B_ℓ happens
$$\mathbb{E}[\mathbb{1}_{A_k}\mathbb{1}_{B_\ell}] = \mathbb{P}(A_k \cap B_\ell) = \mathbb{P}(A_k)\,\mathbb{P}(B_\ell) = p_i p_j$$
since the trials are independent, so the outcomes A_k and B_ℓ are independent (they refer to different trials).

Since there are $n(n-1)$ ways to pick indices $k \neq \ell$, we have

$$\mathbb{E}[X_i X_j] = \sum_{k,\ell=1}^n \mathbb{E}\left[\mathbb{1}_{A_k}\mathbb{1}_{B_\ell}\right] = n(n-1)\,\mathbb{E}\left[\mathbb{1}_{A_1}\mathbb{1}_{B_2}\right] = n(n-1)p_i p_j.$$

Problem 2.10. Suppose that $\mathbf{X} \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \in \mathbb{R}^n$ has a multivariate normal distribution. If $\boldsymbol{\Sigma}$ is such that $\mathrm{Corr}(X_i, X_j) = 0$, show that X_i and X_j are independent.

Solution 2.10. The joint distribution of (X_i, X_j) is normally distributed with mean $\mathbb{E}[X_i] = \mu_i$, $\mathbb{E}[X_j] = \mu_j$ and $\mathrm{Var}(X_i) = \mathrm{Cov}(X_i, X_i) = \Sigma_{ii}$, $\mathrm{Var}(X_j) = \mathrm{Cov}(X_i, X_i) = \Sigma_{jj}$, $\mathrm{Cov}(X_i, X_j) = 0$ since $\mathrm{Corr}(X_i, X_j) = 0$. That is, $(X_i, X_j) \sim \mathrm{N}(\tilde{\bar{\mu}}, \tilde{\boldsymbol{\Sigma}})$ where

$$\tilde{\bar{\mu}} = (\mu_i, \mu_j) \qquad \text{and} \qquad \tilde{\boldsymbol{\Sigma}} = \begin{pmatrix} \Sigma_{ii} & 0 \\ 0 & \Sigma_{jj} \end{pmatrix}.$$

Therefore, the joint PDF of (X_i, X_j) is

$$\begin{aligned} p_{X_i,X_j}(x_i, x_j) &= \frac{1}{(2\pi)} \frac{1}{\sqrt{\det(\tilde{\Sigma})}} e^{-\frac{1}{2}\boldsymbol{x}^\intercal \tilde{\boldsymbol{\Sigma}}^{-1} \boldsymbol{x}} \\ &= \frac{1}{2\pi} \frac{1}{\sqrt{\Sigma_{ii}}\sqrt{\Sigma_{jj}}} e^{-\frac{1}{2}((x_i-\mu_i)^2\Sigma_{ii}^{-1} + (x_j-\mu_j)^2\Sigma_{jj}^{-1})} \\ &= \frac{1}{\sqrt{2\pi\Sigma_{ii}}} e^{-\frac{1}{2}\frac{(x_i-\mu_i)^2}{\Sigma_{ii}}} \cdot \frac{1}{\sqrt{2\pi\Sigma_{jj}}} e^{-\frac{1}{2}\frac{(x_j-\mu_j)^2}{\Sigma_{jj}}} = p_{X_i}(x_i) p_{X_j}(x_j) \end{aligned}$$

so X_i and X_j are independent.

Problem 2.11. Show that any positive semidefinite matrix $\boldsymbol{C}$ corresponds to the covariance of some normally distributed random variable.

Solution 2.11. If $\boldsymbol{C}$ is a positive semidefinite matrix, then there exists a matrix $\boldsymbol{A}$ (which may not necessarily be unique) such that $\boldsymbol{C} = \boldsymbol{A}\boldsymbol{A}^\intercal$. Then if we define $\mathbf{X} = \boldsymbol{A}\mathbf{Z} + \boldsymbol{\mu}$ where $\mathbf{Z}$ is a standard Gaussian vector, then it follows that, for $i, j = 1, ..., n$,

$$\mathbb{E}[X_i] = \mu_i \text{ and } \mathrm{Cov}(X_i, X_j) = \mathrm{Cov}\Big(\sum_{k=1}^n A_{ik} Z_k, \sum_{\ell=1}^n A_{j\ell} Z_\ell\Big) = \sum_{k=1}^n A_{ik} A_{jk} = C_{ij}.$$

Therefore, the covariance matrix $\boldsymbol{C}$ of $\mathbf{X}$ is given by

$$\boldsymbol{C} = \boldsymbol{A}\boldsymbol{A}^\top.$$