1 Important Multivariable Distributions

1.1 Mulitinomial 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 1 (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, $1 \le i \le k$, 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

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

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

Remark 1. 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 2. 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	$\text{Mult}(10, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$
Roll a dice n times	# of each roll	$\text{Mult}(n, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6})$
20 customers order from a menu of 3 items	# of each item	$\text{Mult}(n, p_1, p_2, p_3)$

1.1.1 Properties

1. **Marginal PMF:** The number of times the outcome i occurred is
$$X_j \sim \text{Bin}(n, p_j), \quad \text{for } j = 1, 2, \ldots, k\ .$$
2. **Sum of Marginals:** The number of times the outcomes i or j occurred is
$$X_i + X_j \sim \text{Bin}(n, p_i + p_j), \quad \text{for } i \neq j.$$
3. **Conditional PMF:** The number of times i occured given that i and j occurred t times is
$$X_i \mid X_i + X_j = t \sim \text{Bin}\left(t, \frac{p_i}{p_i + p_j}\right), \quad \text{for } i \neq j.$$
4. **Expected Values:** The expected value of the outcomes are given by
$$\mathbb{E}[X_i X_j] = n(n-1)p_i p_j \text{ for } i \neq j \qquad \text{and} \qquad \mathbb{E}[X_i] = np_i \text{ for } i = 1, \ldots, k$$

1.2 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 2 (Multivariate Normal Distribution)**.** We say that the vector $\boldsymbol{X} = (X_1, X_2, \ldots, X_n) \in \mathbb{R}^n$ has a *multivarwiate normal distribution* with mean $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)$ and positive definite covariance matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{n \times n}$ if X has joint PDF

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

This is denoted by

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

Remark 3. In the special case when $n = 1$, the above density is equivalent to the multivariate normal we saw before.

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

Definition 3 (Standard Normal Vector)**.** We say that the vector $\boldsymbol{X} = (X_1, X_2, \ldots, X_n) \in \mathbb{R}^n$ has a *standard normal distribution* if it has a multivariate normal distribution with mean $\boldsymbol{\mu} = (0, 0, \ldots, 0)$ 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 4. It follows immediately that if $\boldsymbol{X} = (X_1, X_2, \ldots, X_n) \in \mathbb{R}^n$, 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.

1.2.1 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 1.13) 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 5. From this construction, it is possible to define a multivariate normal even if $\boldsymbol{\Sigma}$ is not positive definite since $\boldsymbol{\Sigma}^{\frac{1}{2}}$ is well defined whenever $\boldsymbol{\Sigma}$ is positive semidefinite. However, in this case the density of $\boldsymbol{X}$ will not be explicit.

1.2.2 Properties

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

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

In particular, 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

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

Remark 6. 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.

2. **Correlation 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.

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

   Conversely, if $\boldsymbol{C}$ is any given symmetric and positive semidefinite $n \times n$ matrix, then there exists a $n \times n$ 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.

1.3 Example Problems

Problem 1.1. Let

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

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

Solution 1.1. From the properties of the multinomial distribution (see Week 10), we know that if $(X_1, \ldots, X_k) \sim \mathrm{Mult}(n, p_1, \ldots, p_k)$ 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,

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

Problem 1.2. 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 1.2. 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

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

Problem 1.3. In the game of Roulette, a small ball is spun around a wheel in such a way so that the probability it lands in a black or red box is 18/38 each, and the probability it lands in a green box is 2/38. Suppose 10 games are played, and let B, R and G denote the number of times the ball landed on black, red, and green, respectively.

- Write down the probability function of (B, R, G) along with all its constraints.
- Given that $B = 5$, calculate the probability that $R = 5$.

Solution 1.3.

Part 1: We know $(B, R, G) \sim \text{Mult}(10, 18/38, 18/38, 2/38)$ so

$$\mathbb{P}(B = b, R = r, G = g) = \frac{10!}{b!r!g!}\left(\frac{18}{38}\right)^{b+r}\left(\frac{2}{38}\right)^{g},$$

when $b, r, g \in \{0, 1, \ldots, 10\}$ with $b + r + g = 10$ and 0 otherwise.

Part 2: By definition of conditional probability, and using that marginally $B \sim Bin(10, 18/38)$, we find

$$\begin{aligned}\mathbb{P}(R = 5 \mid B = 5) &= \frac{\mathbb{P}(R = 5, B = 5)}{\mathbb{P}(B = 5)} = \frac{\mathbb{P}(R = 5, B = 5, G = 0)}{\mathbb{P}(B = 5)} \\ &= \frac{\frac{10!}{5!5!}\left(\frac{18}{38}\right)^{10}}{\frac{10!}{5!5!}\left(\frac{18}{38}\right)^{5}\left(\frac{20}{38}\right)^{5}} = \left(\frac{18}{20}\right)^{5} \approx 0.59049\end{aligned}$$

Problem 1.4. We can model n rounds of fair, independent rock-paper-scissors game using multinomial distribution:

$$(R, P, C) \sim \text{Mult}(n, 1/3, 1/3, 1/3).$$

Suppose that I play 5 games of R-P-S. Given that the sum of Rocks and Papers is 4, what would be the distribution of the number of Rocks I played?

Solution 1.4. Using the conditional probability formula for the multinomial with with $n = 5$, $p_j = 1/3$ for $j = 1, 2, 3$ and $t = 4$, we find

$$R \mid R + P = 4 \sim \text{Bin}\left(4, \frac{1/3}{1/3 + 1/3}\right) = \text{Bin}\left(4, \frac{1}{2}\right)$$

Problem 1.5. 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 1.5. 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,

$$\mathbb{E}(L) = \mathbb{E}(L_1 + L_2 + \varepsilon) = \mathbb{E}(L_1) + \mathbb{E}(L_2) + \mathbb{E}(L_3) = 1 + 1.6 + 0 = 2.6.$$

Problem 1.6. 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 1.6. 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 1.7. 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 1.7. By the stability property, we have $\mathbb{E}[X - Y] = \mu_1 - \mu_2$ and $\text{Var}(X - Y) = \text{Var}(X) + \text{Var}(Y) = 2\sigma^2$, so

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

Problem 1.8. 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 1.8. 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 7. 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 1.9. Let $X_1, \ldots, X_n$ be independent and $X_i \sim N(\mu, \sigma^2)$ for all $i = 1, \ldots, n$. Show that

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

Solution 1.9. 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 8. 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 1.10. 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

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

and

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

Which of the following is true?

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

Solution 1.10. 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

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

1.4 Proofs of Key Results

Problem 1.11. If $(X_1, \ldots, X_n) \sim \mathrm{Mult}(n, p_1, \ldots, p_n)$, show that

1. $$X_j \sim \mathrm{Bin}(n, p_j), \quad \text{for } j = 1, 2, \ldots, k\ .$$
2. $$X_i + X_j \sim \mathrm{Bin}(n, p_i + p_j), \quad \text{for } i \ne j.$$
3. $$X_i \mid X_i + X_j = t \sim \mathrm{Bin}\left(t, \frac{p_i}{p_i + p_j}\right), \quad \text{for } i \ne j.$$
4. $$\mathbb{E}[X_i X_j] = n(n-1)p_i p_j \quad \text{for } i \ne j.$$

Solution 1.11.

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

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

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

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

Part 3: Notice that if $X_i+X_j=t$, then X_i takes values in $\{0,1,\dots,t\}$. Therefore, for $x\in\{0,1,\dots,t\}$ we have

$$f_{X_i|X_i+X_j}=\frac{\mathbb{P}(X_i=x)}{\mathbb{P}(X_i+X_j=t)}=\frac{\mathbb{P}(X_i=x,X_j=t-x)}{\mathbb{P}(X_i+X_j=t)}=\frac{\mathbb{P}(X_i=x,X_j=t-x,\sum_{k\neq i,j}X_k=n-t)}{\mathbb{P}(X_i+X_j=t)}$$

since the total of all outcomes must be n. From the second part, we know that $X_i+X_j\sim\text{Bin}(n,p_i+p_j)$

$$f_{X_i|X_i+X_j}=\frac{\frac{n!}{x!(t-x)!(n-t)!}p_i^x p_j^{t-x}(1-p_i-p_j)^{n-t}}{\frac{n!}{t!(n-t)!}(p_i+p_j)^t(1-p_i-p_j)^{n-t}}=\frac{t!}{x!(t-x)!}\left(\frac{p_i}{p_i+p_j}\right)^x\left(\frac{p_j}{p_i+p_j}\right)^{t-x}$$

which we recognize as the PMF of a $\text{Bin}\left(t,\frac{p_i}{p_i+p_j}\right)$ random variable.

Remark 9. This result is intuitive. Since we are given that $X_j+X_j=t$ we know that we have t total occurrences of X_i and X_j. We have that

$$\mathbb{P}(i\text{ happens}\mid i\text{ or }j\text{ happens})=\frac{\mathbb{P}(i\text{ happens})}{\mathbb{P}(i\text{ or }j\text{ happens})}=\frac{p_i}{p_i+p_j}.$$

Therefore, the number of times i happens given that i or j happens at total of t times is $\text{Bin}\left(t,\frac{p_i}{p_i+p_j}\right)$

Part 4: We need to compute (noting that $x_i+x_j\leq n$ needs to hold):

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

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,

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

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 an 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 an 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 1.12. Show that any covariance matrix must be positive semidefinite.

Solution 1.12. Since $\mathrm{Cov}(X^i, X^j) = \mathrm{Cov}(X^j, X^i)$, the matrix C must be **symmetric**. Moreover, if $\mathbf{y} = (y^1, \ldots, y^n)^\top \in \mathbb{R}^n$, 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**.

Problem 1.13. Show that any positive semidefinite matrix $\boldsymbol{C}$ corresponds to the covariance of some normally distributed random variable.

Solution 1.13. If $\boldsymbol{C}$ is a positive semidefinite matrix, then there exists a matrix $\boldsymbol{A}$ (which may not necessarily be unique) such that $\boldsymbol{C} = \boldsymbol{A}\boldsymbol{A}^\intercal$. A very convenient choice for A is the **Cholesky decomposition** of $\boldsymbol{C}$ into the form

$$\boldsymbol{C} = \boldsymbol{L}\boldsymbol{L}^\top,$$

where the matrix $\boldsymbol{L}$ is a lower triangular matrix,

$$\boldsymbol{L} = \begin{pmatrix} \ell_{11} & 0 & 0 & 0 & \cdots & 0 \\ \ell_{21} & \ell_{22} & 0 & 0 & \cdots & 0 \\ \ell_{31} & \ell_{32} & \ell_{33} & 0 & \cdots & 0 \\ \vdots & & \vdots & \ddots & \ddots & \vdots \\ \ell_{d1} & \ell_{d2} & \ell_{d3} & \cdots & \cdots & \ell_{dd} \end{pmatrix}$$

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 $i, j = 1, ..., n$,

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

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

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