Expectation for Multiviariate Distributions

Definition. Suppose X and Y are jointly distributed random variables with joint probability function $f(x, y)$. Then for a function $g : \mathbb{R}^2 \to \mathbb{R}$,

$$E(g(X, Y)) = \sum_{(x,y)} g(x, y)f(x, y).$$

More generally, if $g : \mathbb{R}^n \to \mathbb{R}$, and $X_1, ..., X_n$ have joint probability function $f(x_1, ..., x_n)$, then

$$E(g(X_1, ..., X_n)) = \sum_{(x_1,...,x_n)} g(x_1, ..., x_n)f(x_1, ..., x_n).$$

Example. Suppose X and Y have joint probability function given by the following table:

		x		
	f(x,y)	0	1	2
y	0	.2	.3	.1
	2	.25	.13	.02

Compute $E(XY)$.

Properties of Expectation

(i) For random variables X and Y:

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

(ii) For functions g_1 and g_2:

$$E[a \cdot g_1(X,Y) + b \cdot g_2(X,Y)] = a \cdot E[g_1(X,Y)] + b \cdot E[g_2(X,Y)].$$

Proofs:

(i)

$$
\begin{aligned}
E(X+Y) &= \sum_{x,y}(x+y)f(x,y) = \sum_{x,y} xf(x,y) + \sum_{x,y} yf(x,y) \\
&= \sum_x \sum_y xf(x,y) + \sum_y \sum_x yf(x,y) \\
&= \sum_x [x \sum_y f(x,y)] + \sum_y [y \sum_x f(x,y)] \\
&= \sum_x xf_X(x) + \sum_y yf_Y(y) \\
&= E(X) + E(Y).
\end{aligned}
$$

(ii) Follows from (i) and the fact that for any constant a and a r.v. Z we have $E[aZ] = aE[Z]$.

Clicker Question(s).

Covariance and Correlation

Definition.[1] The **covariance** between X and Y, denoted $Cov(X, Y)$ or σ_{XY}, is

$$Cov(X, Y) := E[(X - E(X))(Y - E(Y))].$$

Shortcut formula:

$$Cov(X, Y) = E(XY) - E(X)E(Y).$$

Note

$$Cov(X, X) = Var(X)$$

and

$$Cov(X, Y) = Cov(Y, X).$$

[1] The definitions and properties of covariance and correlations are the same for discrete and continuous r.v.'s!

Example. Suppose X and Y have the following joint probability function:

	f(x,y)	x = 0	1	2
y	0	0.2	0.3	0.1
	2	0.25	0.13	0.02

Compute $Cov(X, Y)$.

Exercise. Compute $Cov(X^2, X)$ in the above example.

- The **sign**[2] of $Cov(X, Y)$ can be used to uncover certain forms of relationship between X and Y.
    - Suppose
        - **large** values of X tend to occur with **large** values of Y
        - **small** values of X tend to occur with **small** values of Y.

    Then
    $$(X - E(X)) \quad \text{and} \quad (Y - E(Y))$$
    will tend to be of the same sign.

    Therefore,
    $$(X - E(X))(Y - E(Y))$$
    will be on average **positive**, and hence
    $$Cov(X, Y) > 0.$$

[2] The numerical value of $Cov(X, Y)$ has no interpretation.

- Suppose now that
  - **large** values of X tend to occur with **small** values of Y
  - **small** values of X tend to occur with **large** values of Y .

  Then
  $$(X - E(X)) \quad \text{and} \quad (Y - E(Y))$$
  will tend to be of opposite signs.

  Therefore,
  $$(X - E(X))(Y - E(Y))$$
  will be on average **negative**, and hence
  $$Cov(X, Y) < 0.$$

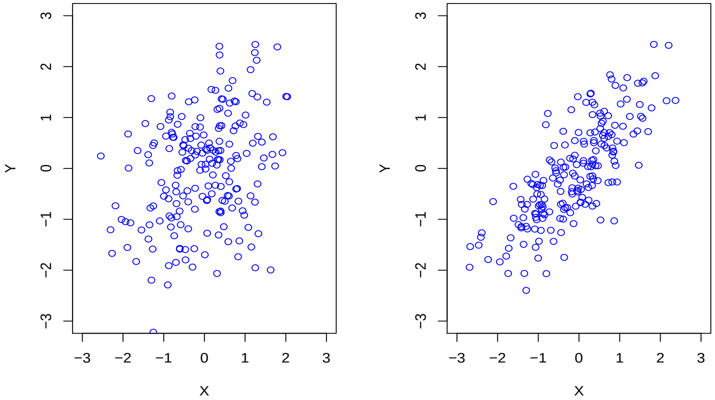


Figure: 200 observations from the joint distribution of (X, Y) with $Cov(X, Y) > 0$ (both panels).

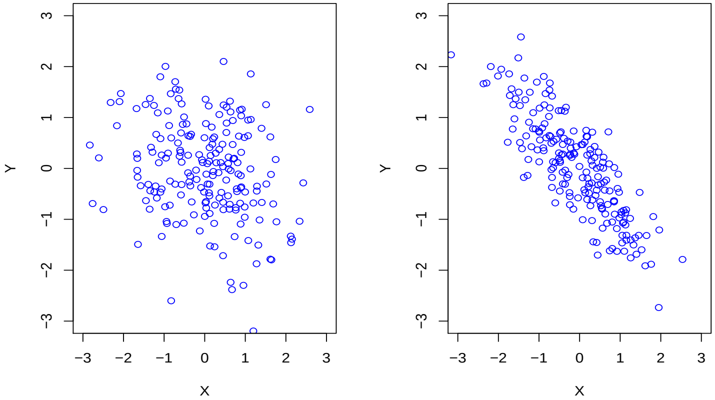


Figure: 200 observations from the joint distribution of (X, Y) with $Cov(X, Y) < 0$ (both panels).

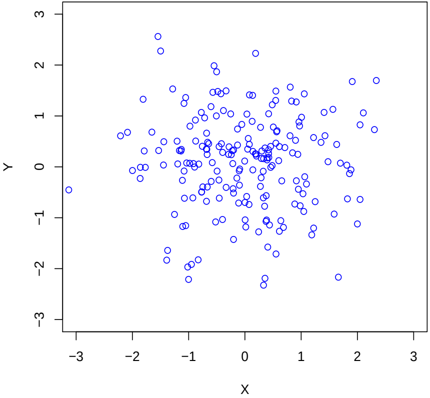


Figure: 200 observations from the joint distribution of (X, Y) where X and Y are independent.

Theorem. If X and Y are independent, then

$$Cov(X, Y) = 0.$$

- The converse statement is FALSE, namely if $Cov(X, Y) = 0$ then X and Y are not necessarily independent.

  Counter example: consider $X \sim N(0, 1)$ and $Y = X^2 - 1$.

- The above Theorem follows from the following more general result.

Theorem. If X and Y are independent, then for any two functions g_1 and g_2 we have

$$E[g_1(X)g_2(Y)] = E[g_1(X)]E[g_2(Y)].$$

Proof. Since $f(x,y) = f_X(x)f_Y(y)$, we have

$$\begin{aligned}
E[g_1(X)g_2(Y)] &= \sum_{x,y} g_1(x)g_2(y)f(x,y) \\
&= \sum_{x,y} g_1(x)g_2(y)f_X(x)f_Y(y) \\
&= \sum_{x}\sum_{y} g_1(x)g_2(y)f_X(x)f_Y(y) \\
&= [\sum_{x} g_1(x)f_X(x)][\sum_{y} g_2(y)f_Y(y)] \\
&= E[g_1(X)]E[g_2(Y)].
\end{aligned}$$

Clicker Question(s).

Correlation

Definition. The **correlation coefficient** of X and Y, denoted $corr(X, Y)$ or ρ, is defined by

$$corr(X, Y) \equiv \rho := \frac{Cov(X, Y)}{SD(X)SD(Y)}.$$

Properties of the correlation coefficient:

- For any r.v.'s X and Y, we always have

$$-1 \leq corr(X, Y) \leq 1.$$

  This results follows from the Cauchy-Schwarz inequality: for any (square integrable) random variables X and Y we have

$$[E(X \cdot Y)]^2 \leq E(X^2)E(Y^2).$$

- When $\rho = 0$ (or $Cov(X, Y) = 0$) we say that X and Y are uncorrelated.

  This is always the case when X and Y are independent!

- The **magnitude** of ρ **matters**: as $\rho \to \pm 1$ the relation between X and Y becomes one-to-one and linear.

  In the limit, when $|corr(X, Y)| = 1$, we have

  $$X = aY + b,$$

  for some constants a and b.

- Note that

  $$\text{sign}(\rho) = \text{sign}(Cov(X, Y)).$$

  Hence the interpretation of the sign of ρ is the same as for $Cov(X, Y)$!

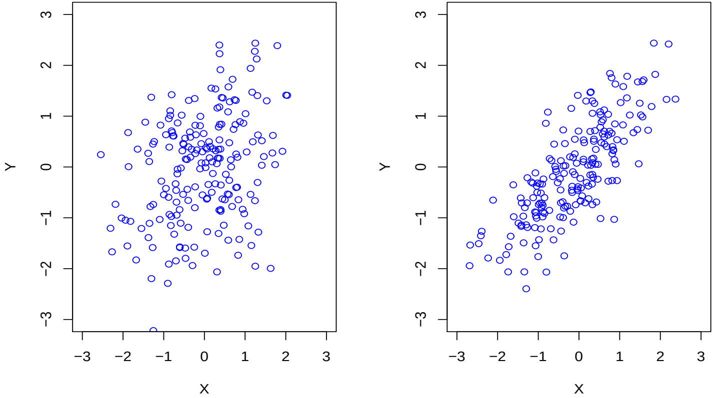


Figure: 200 observations from the joint distribution of (X, Y) with $\rho = 0.25$ (left panel) and $\rho = 0.8$ (right panel).

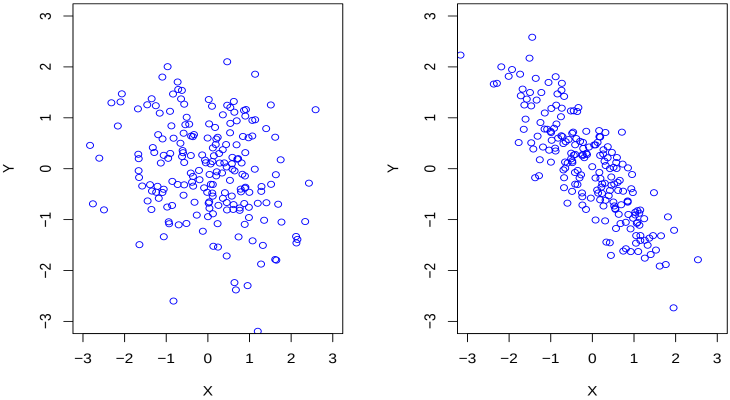


Figure: 200 observations from the joint distribution of (X, Y) with $\rho = -0.25$ (left panel) and $\rho = -0.8$ (right panel).

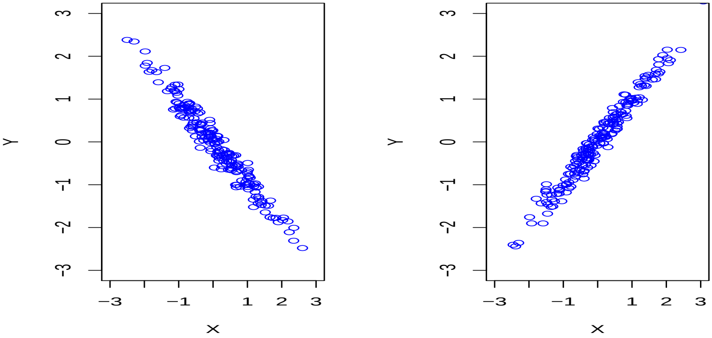


Figure: 200 observations from the joint distribution of (X, Y) with $\rho = -0.98$ (left panel) and $\rho = 0.98$ (right panel).

Clicker Question(s).

Mean and Variance of a Linear Combination

A **linear combination** of r.v's $X_1, \ldots, X_n$ is any random variable of the form

$$\sum_{i=1}^{n} a_i X_i,$$

where $a_1, \ldots, a_n \in \mathbb{R}$ are constants.

Examples:

- for $a_1 = a_2 = \cdots = a_n = 1$, we get the total

$$T = \sum_{i=1}^{n} X_i$$

- for $a_1 = a_2 = \cdots = a_n = 1/n$, we get the sample mean

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

Results for the Mean

Expected Value of a Linear Combination:

$$E\left(\sum_{i=1}^{n} a_i X_i\right) = \sum_{i=1}^{n} a_i E(X_i)$$

Examples:

- For $X_i \sim N(\mu_i, \sigma_i^2),\ i = 1, \cdots, n$, we get
$$E(\frac{1}{n}\sum_{i=1}^{n} X_i) = \frac{1}{n}\sum_{i=1}^{n} \mu_i.$$
- For $X_1, X_2, \ldots, X_n$ that are identically distributed with the common mean μ we have
$$E(\sum_{i=1}^{n} X_i) = n\mu.$$

Clicker Question(s).

Variance of a Linear Combination

Consider first just two variables:

$$
\begin{aligned}
Var(aX+bY) &= E[(aX+bY-a\mu_X-b\mu_Y)^2] \\
&= E\{[a(X-\mu_X)+b(Y-\mu_Y)]^2\} \\
&= E\{a^2(X-\mu_X)^2+2ab(X-\mu_X)(Y-\mu_Y)+b^2(Y-\mu_Y)^2\} \\
&= a^2E[(X-\mu_X)^2]+2abE[(X-\mu_X)(Y-\mu_Y)]+b^2E[(Y-\mu_Y)^2] \\
&= a^2Var(X)+2ab\cdot Cov(X,Y)+b^2Var(Y).
\end{aligned}
$$

Thus,

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

- Important special case:

  if X and Y are uncorrelated then

$$Var(aX + bY) = a^2 Var(X) + b^2 Var(Y).$$

This holds, in particular, when X and Y are independent!

For example, for independent X and Y we have

$$Var(X + 2Y - 3) = Var(X) + 4Var(Y)$$

$$Var(X - Y) = Var(X) + Var(Y).$$

Clicker Question(s).

- For n random variables we have

$$\begin{aligned} Var(\sum_{i=1}^{n} a_i X_i) &= \sum_{i=1}^{n} a_i^2 Var(X_i) + \sum_{1 \le i \neq j \le n} a_i a_j Cov(X_i, X_j). \\ &= \sum_{i=1}^{n} a_i^2 Var(X_i) + 2 \sum_{1 \le i < j \le n} a_i a_j Cov(X_i, X_j) \end{aligned}$$

Special case: if $X_1, X_2, ..., X_n$ are mutually uncorrelated ($Cov(X_i, X_j) = 0$ when $i \neq j$), then

$$Var(\sum_{i=1}^{n} a_i X_i) = \sum_{i=1}^{n} a_i^2 Var(X_i).$$

An important case where this holds is when $X_1, X_2, ..., X_n$ are independent.

- **Law of Large Numbers.** Suppose that $X_1, X_2, \ldots, X_n$ are independent and identically distributed (i.i.d) r.v's with the common mean μ and variance σ^2. Then

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

Note that as $n \to \infty$, we have

$$Var(\frac{1}{n}\sum_{i-1}^{n} X_i) \to 0,$$

and hence

$$\underbrace{\frac{1}{n}\sum_{i=1}^{n} X_i}_{\text{sample mean}} \to \underbrace{\mu}_{\text{model mean}} \quad \text{as } n \to \infty.$$

A formal statement of this fact is called the "Law of Large Numbers".

Covariance of a Linear Combination

For r.v's X, Y, U and V we have

$$Cov(aX + bY, cU + dV) = E[(aX + bY - a\mu_X - b\mu_Y)(cU + dV - c\mu_U - d\mu_V)]$$

$$= E[(a(X - \mu_X) + b(Y - \mu_Y)][c(U - \mu_U) + d(V - \mu_V)]$$

$$= ac \cdot E[(X - \mu_X)(U - \mu_U)] + ad \cdot E[(X - \mu_X)(V - \mu_V)]$$

$$+ bc \cdot E[(Y - \mu_Y)((U - \mu_U]) + bd \cdot E[(Y - \mu_Y)(V - \mu_V)]$$

$$= ac \cdot Cov(X, U) + ad \cdot Cov(X, V) + bc \cdot Cov(Y, U) + bd \cdot Cov(Y, V).$$

or (easier to remember):

$$\boxed{Cov(X + Y, U + V) = Cov(X, U) + Cov(X, V) + Cov(Y, U) + Cov(Y, V).}$$

Examples:

- Let $E(X) = 1$, $E(Y) = 2$, $Var(X) = Var(Y) = 1$, and $Cov(X, Y) = -1$. Then,

$$\begin{aligned} Var(3X - Y) &= 9\,Var(X) + 2 \cdot 3 \cdot (-1) \cdot Cov(X, Y) + Var(Y) \\ &= 9 \cdot 1 + 2 \cdot 3 \cdot (-1) \cdot (-1) + 1 = 16. \end{aligned}$$

- Let Z_1 and Z_2 be independent r.v's such that

$$Z_i \sim N(0, 1) \quad i = 1, 2.$$

Then, for $\rho \in [0, 1]$, we have

$$\begin{aligned} &Cov(\rho Z_1 + \sqrt{1-\rho^2} Z_2, Z_1) = \\ &\quad = \rho \cdot Cov(Z_1, Z_1) + \sqrt{1-\rho^2} \cdot Cov(Z_2, Z_1) \\ &\quad = \rho \cdot Cov(Z_1, Z_1) + 0 = \rho \cdot Var(Z_1) = \rho. \end{aligned}$$

and $\quad corr(\rho Z_1 + \sqrt{1-\rho^2} Z_2, Z_1) = \rho.$

Clicker Question(s).

Linear Combinations of Independent Normals

Theorem. Suppose that $X_1, ..., X_n$ are independent random variables such that

$$X_i \sim N(\mu_i, \sigma_i^2), \quad i = 1, \ldots, n.$$

Then[3]

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

- Note that the "only" new fact here is that the variable $\sum_{i=1}^{n} a_i X_i$ has a **normal distribution**, as the formulae for the mean and variance follow from the general rules that we have just discussed!

[3] This can be proven by using moment generating functions.

Examples. Suppose that $X_1, ..., X_n$ are independent and

$$X_i \sim N(\mu, \sigma^2), \quad i = 1, \ldots, n.$$

Then

$$X_1 - X_2 \sim N(\mu - \mu = 0, 2\sigma^2)$$

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

$$\sum_{i=1}^{n} X_i \sim N(n\mu, n\sigma^2).$$

Exercise. For $X \sim N(\mu, \sigma^2)$ show that

$$aX + b \sim N(a\mu + b, a^2\sigma^2).$$

Clicker Question(s).

Example. Suppose that the height of adult males in Canada is normally distributed with a mean of 70 inches and variance of 4^2 inches. Let $X_1, \ldots, X_n$ denote the heights of a random sample of n adult males and $\bar{X}_n$ denote the sample mean of these heights.

(i) Compute the probability that X_3 exceeds 75.

(ii) Compute the probability that for $n = 10$ the sample mean $\bar{X}_{10}$ exceeds 75.

(iii) What is the smallest value of n such that $SD(\bar{X}_n) \leq 2$?

(iv) What is the smallest value of n such that

$$P(|\bar{X}_n - 70| \leq 1) \geq 0.9?$$

Exercise. Suppose that the height of adult males in Canada is normally distributed with a mean of 70 inches and variance of 4^2 inches, while the height of adult females is normally distributed with a mean of 65 inches and variance of 3^2 inches.

Let $X_1, ..., X_{20}$ denote the heights of a random sample of 20 adult males, and $Y_1, ..., Y_{20}$ denote the heights of a random sample of 20 adult females.

Assuming that $\bar{X}_n$ and $\bar{Y}_n$ are independent, calculate the probability that $\bar{X}_n$ differs from $\bar{Y}_n$ by more than 5 inches.

Indicator Random Variables

Let $A \subset S$ be an event. We say that $\mathbb{1}_A$ is the **indicator** random variable of the event A if

$$\mathbb{1}_A(s) = \begin{cases} 1 & s \in A, \\ 0 & s \in \bar{A}. \end{cases}$$

Thus,

$$\mathbb{1}_A(s) = \begin{cases} 1 & \text{with probability } P(A), \\ 0 & \text{with probability } 1 - P(A). \end{cases}$$

- Why do we introduce indicator variables?

  To make many calculations, like computing the mean and variance, vastly easier.

Exercise. Show

(i)
$$E(\mathbb{1}_A) = P(A). \tag{1}$$

(ii) Since $(\mathbb{1}_A)^2 = \mathbb{1}_A$,
$$E[(\mathbb{1}_A)^2] = E(\mathbb{1}_A) = P(A).$$

(iii)
$$Var(\mathbb{1}_A) = P(A)(1 - P(A)). \tag{2}$$

(iv) For two events A and B:
$$Cov(\mathbb{1}_A, \mathbb{1}_B) = P(A \cap B) - P(A)P(B). \tag{3}$$

Thus, if A and B are independent, then
$$Cov(\mathbb{1}_A, \mathbb{1}_B) = 0.$$

Example. Suppose $X \sim Binomial(n, p)$. Show that

$$E(X) = np \quad \text{and} \quad Var(X) = np(1-p)$$

using indicator random variables.

- For each of the n experiments introduce an indicator r.v. as follows:

$$X_i = \begin{cases} 1 & \text{if the } i\text{'th trial is a success} \\ 0 & \text{if the } i\text{'th trial is a failure.} \end{cases}$$

  Observe that the total number of successes X can be now represented as

$$X = \sum_{i=1}^{n} X_i.$$

- By (1) – (2), we have for $i = 1, \dots, n$:

$$E(X_i) = p \quad \text{and} \quad Var(X_i) = p(1-p).$$

Therefore,

$$E(X) = E(\sum_{i=1}^{n} X_i) = n \cdot p.$$

$$\begin{aligned} Var(X) &= Var(\sum_{i=1}^{n} X_i) \\ &= \sum_{i=1}^{n} Var(X_i) + \sum_{i \neq j} Cov(X_i, X_j) \\ &= n \cdot p(1-p) + 0 = n \cdot p(1-p), \end{aligned}$$

where in the last line we used independence of $X_1, X_2, \dots, X_n$.

Example. Suppose $X \sim hyp(N, r, n)$. Show

$$E(X) = n\frac{r}{N} \quad \text{and} \quad Var(X) = n(\frac{r}{N})(1 - \frac{r}{N})\frac{N-n}{N-1}$$

using indicator random variables.

- For each of the n experiments introduce an indicator r.v. as follows:

$$X_i = \begin{cases} 1 & \text{if the } i\text{'th trial is a success} \\ 0 & \text{if the } i\text{'th trial is a failure.} \end{cases}$$

  Then, as for the Binomial distribution, we can represent the total number of successes X as

$$X = \sum_{i=1}^{n} X_i.$$

  Note, however, that now the variables $X_1, X_2, \ldots, X_n$ are not independent!

- Since

$$P(\text{success on the } i\text{'th draw}) = \frac{r}{N}, \quad i = 1, \ldots, n, \tag{4}$$

we have

$$E(X_i) = \frac{r}{N} \tag{5}$$

$$Var(X_i) = \frac{r}{N}(1 - \frac{r}{N}). \tag{6}$$

At first (4) may seem to be counterintuitive, as the probability of success on the i'th trial depends on the outcomes of the other trials.
However here we are calculating the expectation (and the probability of success) without knowing the other outcomes.

From (5) we get

$$E(X) = E(\sum_{i=1}^{n} X_i) = n\frac{r}{N}.$$

- To find $Var(\sum_{i=1}^{n} X_i)$ we need to compute

$$Cov(X_i, X_j) \text{ for each pair } i\ , j,\ i \neq j,$$

or, by (3),

$$P(X_i = 1, X_j = 1).$$

For the latter, we have

$$\begin{aligned} P(X_i = 1, X_j = 1) &= P(X_i = 1 | X_j = 1) P(X_j = 1) \\ &= (\frac{r-1}{N-1})(\frac{r}{N}), \end{aligned}$$

where we used a similar argument as the one for (4). Hence

$$\begin{aligned} Cov(X_i, X_j) &= (\frac{r-1}{N-1})(\frac{r}{N}) - (\frac{r}{N})^2 \\ &= -\frac{r(N-r)}{N^2(N-1)}. \end{aligned}$$

Thus,

$$\begin{aligned} Var(\sum_{i=1}^{n} X_i) &= \sum_{i=1}^{n} Var(X_i) + \sum_{i \neq j} Cov(X_i, X_j) \\ &= n\frac{r}{N}(1 - \frac{r}{N}) + [-\frac{r(N-r)}{N^2(N-1)}] \sum_{i \neq j} 1. \end{aligned}$$

For the last term

$$\sum_{1 \leq i \neq j \leq n} 1 = 2 \sum_{1 \leq i < j \leq n} 1 = 2\binom{n}{2},$$

since we can select and order two different integers out of n in $\binom{n}{2}$ ways. Thus,

$$Var(\sum_{i=1}^{n} X_i) = n\frac{r}{N}(1 - \frac{r}{N}) - 2\binom{n}{2}\frac{r(N-r)}{N^2(N-1)},$$

which, with simple algebra, gives the result.

Similar problems in the Notes:

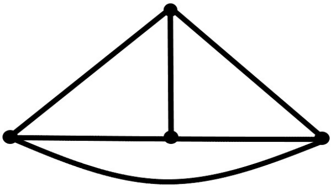
- Example on page 230.

  We have N letters to N different people, and N envelopes addressed to those N people. One letter is put in each envelope at random. Find the mean and variance of the number of letters placed in the right envelope.

- 9.71, 9.72, 33, and 34.
- Question 4 in the Sample Final Exam:

  Ten friends go to an all-you-can-eat sushi restaurant and sit at one large round table. Each person likes spicy food with probability 0.6, independently of each other. We say a "match" occurs when two people sitting next to each other BOTH like spicy food or BOTH do not like spicy food. (...) Find the expected total number of "matches" at the table.

Exercise. Four vertices are connected with edges as shown:



Each edge has probability p of being destroyed. Find the mean and variance of the number of isolated vertices in the final graph. (A vertex is isolated if all edges connecting it to other vertices have been destroyed).