Appendix B: Probability and Vectors

Joshua French

Overview of probability, random variables, and random vectors

Probability Basics

Probability attempts to quantify how likely certain outcomes are, where the outcomes are produced by a random experiment (defined below).

Required Background: basic set theory.

The table below summarizes basic probability-related terminology.

term	notation	definition
experiment	N/A	A mechanism that produces outcomes that cannot be predicted with absolute certainty.
outcome	ω	The simplest kind of result produced by an experiment.
sample space	Ω	The set of all possible outcomes an experiment can produce.
event	A, A_i, B, etc.	Any subset of Ω.
empty set	$\emptyset$	The event that includes no outcomes.

Some comments about the terms,

- **Outcomes**: also called **points**, **realizations**, or **elements**.
- **Event:** a subset of outcomes.
- The **empty set** is a subset of Ω, but not an outcome of Ω.
- The **empty set** is a subset of every event $A \subseteq \Omega$.

Basic Set Operations

Let A and B be two events contained in Ω.

- The **intersection** of A and B is the set of outcomes that are common to both A and B,
  - Denoted $A \cap B$
  - Set definition: $A \cap B = \{\omega \in \Omega : \omega \in A \text{ and } \omega \in B\}$.
- Events A and B are **disjoint** if $A \cap B = \emptyset$, or A and B have no common outcomes.
- The **union** of A and B is the set of outcomes that are in A or B or both.
  - Denoted $A \cup B$
  - Set definition: $A \cup B = \{\omega \in \Omega : \omega \in A \text{ or } \omega \in B\}$.
- The **complement** of A is the set of outcomes that are in Ω but are not in A.
  - Denoted A^c, $\overline{A}$, or A'.
  - Set definition: $A^c = \{\omega \in \Omega : \omega \notin A\}$.
- The set **difference** between A and B is the elements of A that are not in B.
  - Denoted $A \quad B$
  - Set definition: $A \quad B = \{\omega \in A : \omega \notin B\}$.
  - The set difference between A and B may also be denoted by $A - B$.
  - The set difference is order specific, i.e., $(A \quad B) \neq (B \quad A)$ in general.

Probability Function

A function P that assigns a real number $P(A)$ to every event A is a probability distribution if it satisfies three properties:

1. $P(A) \geq 0$ for all $A \in \Omega$.
2. $P(\Omega) = P(\omega \in \Omega) = 1$. Alternatively, $P(A \subseteq \Omega) = 1$.
3. If $A_1, A_2, ...$ are disjoint, then $P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i)$.

A set of events $\{A_i : i \in I\}$ are **independent** if

$$P\left(\cap_{i \in J} A_i\right) = \prod_{i \in J} P(A_i)$$

for every finite subset $J \subseteq I$.

The **conditional probability** of A given B, denoted as $P(A \mid B)$, is the probability that A occurs given that B has occurred, and is defined as

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}, \quad P(B) > 0.$$

Some additional facts about probabilities:

- **Complement rule**: $P(A^c) = 1 - P(A)$.
- **Addition rule**: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.
- **Bayes' rule**: Assuming $P(A) > 0$ and $P(B) > 0$, then

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

- **Law of Total Probability**: Let $B_1, B_2, ...$ be a countably infinite partition of Ω. Then

$$P(A) = \sum_{i=1}^{\infty} P(A \cap B_i) = \sum_{i=1}^{\infty} P(A \mid B_i)P(B_i).$$

Random Variables

A **random variable** Y is a mapping/function

$$Y : \Omega \to \mathbb{R}$$

that assigns a real number $Y(\omega)$ to each outcome ω. (We typically drop the (ω) notation for simplicity.)

The **cumulative distribution function (CDF)** of Y, F_Y, is a function $F_Y : \mathbb{R} \to [0, 1]$ defined by

$$F_Y(y) = P(Y \leq y).$$

The subscript of F indicates the random variable the CDF describes. E.g., F_X denotes the CDF of the random variable X and F_Y denotes the CDF of the random variable Y. The subscript can be dropped when the context makes it clear what random variable the CDF describes. An F-distributed random variable is one that has the F distribution.

The **support** of Y, $\mathcal{S}$, is the smallest set such that $P(Y \in \mathcal{S}) = 1$.

Discrete random variables

Y is a **discrete** random variable if it takes countably many values $\{y_1, y_2, ...\} = \mathcal{S}$.

The **probability mass function (pmf)** for Y is $f_Y(y) = P(Y = y)$, where $y \in \mathbb{R}$, and must have the following properties:

1. $0 \leq f_Y(y) \leq 1$.
2. $\sum_{y \in \mathcal{S}} f_Y(y) = 1$.

Additionally, the following statements are true:

- $F_Y(c) = P(Y \leq c) = \sum_{y \in \mathcal{S} : y \leq c} f_Y(y)$.

- $P(Y \in A) = \sum_{y \in A} f_Y(y)$ for some event A.
- $P(a \leq Y \leq b) = \sum_{y \in \mathcal{S}: a \leq y \leq b} f_Y(y)$.

The **expected value**, **mean**, or first moment of Y is defined as

$$E(Y) = \sum_{y \in \mathcal{S}} y f_Y(y),$$

assuming the sum is well-defined.

The **variance** of Y is defined as

$$\begin{aligned} \text{var}(Y) &= E(Y - E(Y))^2 \\ &= \sum_{y \in \mathcal{S}} (y - E(Y))^2 f_Y(y). \end{aligned}$$

Note that $\text{var}(Y) = E(Y - E(Y))^2 = E(Y^2) - [E(Y)]^2$. The last expression is often easier to compute.

The **standard deviation** of Y is

$$SD(Y) = \sqrt{\text{var}(Y)}.$$

Example (Bernoulli)

A random variable Y is said to have a Bernoulli distribution with probability θ, denoted $Y \sim \mathsf{Bernoulli}(\theta)$, if:

- $\mathcal{S} = \{0, 1\}$
- $P(Y = 1) = \theta$, where $\theta \in (0, 1)$.

Bernoulli PMF:

$$f_Y(y) = \theta^y (1 - \theta)^{(1-y)}.$$

Determine the mean and variance of Y.

Mean:

$$E(Y) = 0(1 - \theta) + 1(\theta) = \theta.$$

Variance

$$\begin{aligned} \text{var}(Y) &= (0 - \theta)^2 (1 - \theta) + (1 - \theta)^2 \theta \\ &= \theta(1 - \theta). \end{aligned}$$

Continuous random variables

Y is a **continuous** random variable if there exists a function $f_Y(y)$ such that:

1. $f_Y(y) \geq 0$ for all y,
2. $\int_{-\infty}^{\infty} f_Y(y)dy = 1$,
3. $a \leq b$, $P(a < Y < b) = \int_a^b f_Y(y)dy$.

The function f_Y is called the **probability density function (pdf)**.

Additionally, $F_Y(y) = \int_{-\infty}^{y} f_Y(y)dy$ and $f_Y(y) = F'_Y(y)$ for any point y at which F_Y is differentiable.

The **mean** of a continuous random variables Y is defined as

$$E(Y) = \int_{-\infty}^{\infty} y f_Y(y)dy = \int_{y \in \mathcal{S}} y f_Y(y).$$

assuming the integral is well-defined.

The **variance** of a continuous random variable Y is defined by

$$\mathrm{var}(Y) = E(Y - E(Y))^2 = \int_{-\infty}^{\infty} (y - E(Y))^2 f_Y(y)dy = \int_{y \in \mathcal{S}} (y - E(Y))^2 f_Y(y)dy$$

Example (Exponential distribution)

A random variable Y is said to have an exponential distribution rate parameter λ, denoted with $Y \sim \mathsf{Exp}(\lambda)$ if $\mathcal{S} = \{y \in \mathbb{R} : y \geq 0\}$ and has distribution,

Exponential PDF:

$$f_Y(y) = \lambda \exp(-\lambda y)$$

Determine the mean and variance of Y.

Mean:

$$\begin{aligned} E(Y) &= \int_0^{\infty} y\lambda \exp(-\lambda y)\ dy \\ &= -\exp(-\lambda y)(\lambda^{-1} + y)\Big]_0^{\infty} \\ &= \frac{1}{\lambda}. \end{aligned}$$

Note that this process involves integration by parts, which is not shown. Similarly, $E(Y^2) = \frac{2}{\lambda^2}$. Thus,

Variance:

$$\begin{aligned}
\mathrm{var}(Y) &= E(Y^2) - [E(Y)]^2 \\
&= 2\lambda^{-2} - [\lambda^{-1}]^2 \\
&= \frac{2}{\lambda^2}.
\end{aligned}$$

Useful facts for transformations of random variables

Let Y be a random variable and $a \in \mathbb{R}$ be a constant. Then:

- $E(a) = a$.
- $E(aY) = aE(Y)$.
- $E(a + Y) = a + E(Y)$.
- $\mathrm{var}(a) = 0$.
- $\mathrm{var}(aY) = a^2\mathrm{var}(Y)$.
- $\mathrm{var}(a + Y) = \mathrm{var}(Y)$.
- For a discrete random variable and a function g,

  $$E(g(Y)) = \sum_{y\in\mathcal{S}} g(y) f_Y(y),$$

  assuming the sum is well-defined.
- For a continuous random variable and a function g,

  $$E(g(Y)) = \int_{y\in\mathcal{S}} g(y) f_Y(y)\ dy,$$

  assuming the integral is well-defined.

Multivariate distributions

Basic properties

Let $Y_1, Y_2, \dots, Y_n$ denote n random variables with supports $\mathcal{S}_1, \mathcal{S}_2, \dots, \mathcal{S}_n$, respectively.

If the random variables are **jointly discrete** (i.e., all discrete), then the joint pmf $f(y_1, \dots, y_n) = P(Y_1 = y_1, \dots, Y_n = y_n)$ satisfies the following properties:

1. $0 \leq f(y_1, \dots, y_n) \leq 1$,
2. $\sum_{y_1\in\mathcal{S}_1} \cdots \sum_{y_n\in\mathcal{S}_n} f(y_1, \dots, y_n) = 1$,
3. $P((Y_1, \dots, Y_n) \in A) = \sum_{(y_1,\dots,y_n)\in A} f(y_1, \dots, y_n)$.

In this context,

$$E(Y_1 \cdots Y_n) = \sum_{y_1 \in \mathcal{S}_1} \cdots \sum_{y_n \in \mathcal{S}_n} y_1 \cdots y_n f(y_1, \dots, y_n).$$

In general,

$$E(g(Y_1, \dots, Y_n)) = \sum_{y_1 \in \mathcal{S}_1} \cdots \sum_{y_n \in \mathcal{S}_n} g(y_1, \dots, y_n) f(y_1, \dots, y_n),$$

where g is a function of the random variables.

If the random variables are **jointly continuous**, then $f(y_1, \dots, y_n)$ is the joint pdf if it satisfies the following properties:

1. $f(y_1, \dots, y_n) \geq 0$,
2. $\int_{y_1 \in \mathcal{S}_1} \cdots \int_{y_n \in \mathcal{S}_n} f(y_1, \dots, y_n) dy_n \cdots dy_1 = 1$,
3. $P((Y_1, \dots, Y_n) \in A) = \int \cdots \int_{(y_1, \dots, y_n) \in A} f(y_1, \dots, y_n) dy_n \dots dy_1$.

In this context,

$$E(Y_1 \cdots Y_n) = \int_{y_1 \in \mathcal{S}_1} \cdots \int_{y_n \in \mathcal{S}_n} y_1 \cdots y_n f(y_1, \dots, y_n) dy_n \dots dy_1.$$

In general,

$$E(g(Y_1, \dots, Y_n)) = \int_{y_1 \in \mathcal{S}_1} \cdots \int_{y_n \in \mathcal{S}_n} g(y_1, \dots, y_n) f(y_1, \dots, y_n) dy_n \cdots dy_1,$$

where g is a function of the random variables.

Marginal distributions

If the random variables are jointly discrete, then the marginal pmf of Y_1 is obtained by summing over the other variables $Y_2, \dots, Y_n$:

$$f_{Y_1}(y_1) = \sum_{y_2 \in \mathcal{S}_2} \cdots \sum_{y_n \in \mathcal{S}_n} f(y_1, \dots, y_n).$$

Similarly, if the random variables are jointly continuous, then the marginal pdf of Y_1 is obtained by integrating over the other variables $Y_2, \dots, Y_n$:

$$f_{Y_1}(y_1) = \int_{y_2 \in \mathcal{S}_2} \cdots \int_{y_n \in \mathcal{S}_n} f(y_1, \dots, y_n) dy_n \cdots dy_2.$$

Independence of random variables

Random variables X and Y are independent if

$$F(x, y) = F_X(x)F_Y(y).$$

Alternatively, X and Y are independent if

$$f(x, y) = f_X(x)f_Y(y).$$

Conditional distributions

Let X and Y be random variables. Then assuming $f_Y(y) > 0$, the conditional distribution of X given $Y = y$, denoted $X|Y = y$ comes from Bayes' formula:

$$f(x|y) = \frac{f(x, y)}{f_Y(y)}, \quad f_Y(y) > 0.$$

Covariance

The covariance between random variables X and Y is

$$\text{cov}(X, Y) = E[(X - E(X))(Y - E(Y))] = E(XY) - E(X)E(Y).$$

Useful facts for transformations of multiple random variables

Let a and b be scalar constants. Let Y and Z be random variables. Then:

- $E(aY + bZ) = aE(Y) + bE(Z)$.
- $\text{var}(Y + Z) = \text{var}(Y) + \text{var}(Z) + 2\text{cov}(Y, Z)$.
- $\text{cov}(a, Y) = 0$.
- $\text{cov}(Y, Y) = \text{var}(Y)$.
- $\text{cov}(aY, bZ) = ab\text{cov}(Y, Z)$.
- $\text{cov}(a + Y, b + Z) = \text{cov}(Y, Z)$.

If Y and Z are also independent, then:

- $E(YZ) = E(Y)E(Z)$.
- $\text{cov}(Y, Z) = 0$.

In general, if $Y_1, Y_2, \ldots, Y_n$ are a set of random variables, then:

- $E(\sum_{i=1}^{n} Y_i) = \sum_{i=1}^{n} E(Y_i)$, i.e., the expectation of the sum of random variables is the sum of the expectation of the random variables.
- $\text{var}(\sum_{i=1}^{n} Y_i) = \sum_{i=1}^{n} \text{var}(Y_i) + \sum_{j=1}^{n} \sum_{1 \le i < j \le n} 2\text{cov}(Y_i, Y_j)$, i.e., the variance of the sum of random variables is the sum fo the variables' variances plus the sum of twice all possible pairwise covariances.

If in addition, $Y_1, Y_2, \dots, Y_n$ are all independent of each other, then:

- $\text{var}(\sum_{i=1}^{n} Y_i) = \sum_{i=1}^{n} \text{var}(Y_i)$ since all pairwise covariances are 0.

Example (Binomial distribution)

A random variable Y is said to have a Binomial distribution with n trials and probability of success θ, denoted $Y \sim \mathsf{Bin}(n, \theta)$ when $\mathcal{S} = \{0, 1, 2, \dots, n\}$ and the pmf is:

Binomial PMF:

$$f(y \mid \theta) = \binom{n}{y} \theta^y (1-\theta)^{(n-y)}.$$

An alternative explanation of a Binomial random variable is that it is the sum of n independent and identically-distributed Bernoulli random variables. Alternatively, let $Y_1, Y_2, \dots, Y_n \overset{i.i.d.}{\sim} \mathsf{Bernoulli}(\theta)$, where i.i.d. stands for independent and identically distributed, i.e., $Y_1, Y_2, \dots, Y_n$ are independent random variables with identical distributions. Then $Y = \sum_{i=1}^{n} Y_i \sim \mathsf{Bin}(n, \theta)$.

A Binomial random variable with $\theta = 0.5$ models the question: what is the probability of flipping y heads in n flips?

Determine the mean and variance of Y.

Mean:

$E(Y_i) = \theta$ for $i = 1, 2, \dots, n$.

Variance:

$\text{var}(Y_i) = \theta(1-\theta)$ for $i = 1, 2, \dots, n$.

We determine that:

$$
\begin{aligned}
E(Y) &= E\left(\sum_{i=1}^{n} Y_i\right) \\
&= \sum_{i=1}^{n} E(Y_i) \\
&= \sum_{i=1}^{n} \theta \\
&= n\theta.
\end{aligned}
$$

Similarly, since $Y_1, Y_2, \ldots, Y_n$ are i.i.d., we see that

$$
\begin{aligned}
\mathrm{var}(Y) &= \mathrm{var}(\sum_{i=1}^{n} Y_i) \\
&= \sum_{i=1}^{n} \mathrm{var}(Y_i) \\
&= \sum_{i=1}^{n} \theta(1-\theta) \\
&= n\theta(1-\theta)
\end{aligned}
$$

Example (Continuous bivariate distribution)

Hydration is important for health. Like many people, the author has a water bottle he uses to say hydrated through the day and drinks several liters of water per day. Let's say the author refills his water bottle every 3 hours.

- Let Y denote the proportion of the water bottle filled with water at the beginning of the 3-hour window.
- Let X denote the amount of water the author consumes in the 3-hour window (measured in the the proportion of total water bottle capacity).

We know that $0 \le X \le Y \le 1$. The joint density of the random variables is

$$
f(x, y) = 4y^2, \quad 0 \le x \le y \le 1,
$$

and 0 otherwise.

We answer a series of questions about this distribution.

Q1: Determine $P(0.5 \le X \le 1, 0.75 \le Y)$.

$$
\begin{aligned}
&\int_{3/4}^{1}\int_{0.5}^{y} 4y^2 \, dx \, dy \\
&= \int_{3/4}^{1} 4y^2x\Big]_{1/2}^{y} \, dy \\
&= \int_{3/4}^{1} 4y^4 - 2y^2 \, dy \\
&= y^4 - \frac{2}{3}y^3\Big]_{3/4}^{1} \\
= \; &= \left(1-\frac{2}{3}\right) - \left(\frac{81}{256} - \frac{2(27)}{3(64)}\right) \\
&= 229/768 \approx 0.30.
\end{aligned}
$$

Q2: Determine the marginal distributions of X and Y.

$$
\begin{aligned}
f_X(x) &= \int_{x}^{1} 4y^2 \, dy \\
&= \frac{4}{3}y^3\Big]_{x}^{1} \\
&= \frac{4}{3}(1-x^3), \quad 0 \le x \le 1.
\end{aligned}
$$

$$
\begin{aligned}
f_Y(y) &= \int_{0}^{y} 4y^2 \, dx \\
&= 4y^2x\Big]_{0}^{y} \\
&= 4y^3, \quad 0 \le y \le 1.
\end{aligned}
$$

Q3: Determine the means of X and Y.

The mean of X is the integral of $xf_X(x)$ over the support of X, i.e.,

$$
\begin{aligned}
E(X) &= \int_0^1 x\left(\frac{4}{3}(1-x^3)\right)\,dx \\
&= \frac{2}{3}x^2 - \frac{4}{15}x^4\Big]_0^1 \\
&= \frac{2}{3} - \frac{4}{15} \\
&= \frac{10}{15} - \frac{4}{15} \\
&= \frac{2}{5}.
\end{aligned}
$$

Similarly,

$$
\begin{aligned}
E(Y) &= \int_0^1 y(4y^3)\,dy \\
&= \frac{4}{5}y^5\Big]_0^1 \\
&= \frac{4}{5}.
\end{aligned}
$$

Q4: Determine the variances of X and Y.

We use the formula $\mathrm{var}(X) = E(X^2) - [E(X)^2]$ to compute the variances. First,

$$
\begin{aligned}
E(X^2) &= \int_0^1 x^2\left(\frac{4}{3}(1-x^3)\right)\,dx \\
&= \int_0^1 \frac{4}{3}x^2 - \frac{4}{3}x^5\,dx \\
&= \frac{4}{9}x^3 - \frac{4}{18}x^6\Big]_0^1 \\
&= \frac{4}{9} - \frac{4}{18} \\
&= \frac{8}{18} - \frac{4}{18} \\
&= \frac{4}{18} \\
&= \frac{2}{9}.
\end{aligned}
$$

Second,

$$E(Y^2) = \int_0^1 y^2(4y^3)\ dy = \frac{2}{3}$$

Thus,

$$\begin{aligned}
E(Y^2) &= \int_0^1 y^2(4y^3)\ dy \\
&= \frac{4}{6}y^6\Big]_0^1 \\
&= \frac{4}{6} \\
&= \frac{2}{3}.
\end{aligned}$$

$$\text{var}(X) = 2/9 - (2/5)^2 = \frac{14}{225}.$$

$$\text{var}(Y) = 2/3 - (4/5)^2 = \frac{2}{75}.$$

Q5: Determine the mean of XY.

$$\begin{aligned}
E(XY) &= \int_0^1 \int_0^y xy(4y^2)\ dx\ dy \\
&= \int_0^1 2x^2y^3\Big]_0^y\ dy \\
&= \int_0^1 2y^5\ dy \\
&= \frac{2}{6}y^6\Big]_0^1 \\
&= \frac{2}{6} \\
&= \frac{1}{3}.
\end{aligned}$$

Q6: Determine the covariance of X and Y.

Using our previous work, we see that,

$$
\begin{aligned}
\operatorname{cov}(X, Y) &= E(XY) - E(X)E(Y) \\
&= 1/3 - (2/5)(4/5) \\
&= \frac{1}{3} - \frac{8}{25} \\
&= \frac{25}{75} - \frac{24}{75} \\
&= \frac{1}{75}.
\end{aligned}
$$

Q7: Determine the mean and variance of $Y - X$, i.e., the average amount of water remaining after a 3-hour window and the variability of that amount.

$$
\begin{aligned}
E(Y - X) &= E(Y) - E(X) \\
&= 4/5 - 2/5 \\
&= \frac{2}{5}
\end{aligned}
$$

$$
\begin{aligned}
\operatorname{var}(Y - X) &= \operatorname{var}(Y) + \operatorname{var}(X) - 2\operatorname{cov}(Y, X) \\
&= 2/75 + 14/225 - 2(1/75) \\
&= 14/225.
\end{aligned}
$$

Random vectors

Definition

A **random vector** is a vector of random variables. A random vector is assumed to be a column vector unless otherwise specified.

Additionally, a **random matrix** is a matrix of random variables.

Mean, variance, and covariance

Let $\mathbf{y} = [Y_1, Y_2, \dots, Y_n]$ be an $n \times 1$ random vector.

The mean of a random vector is the vector containing the means of the random variables in the vector. More specifically, the mean of $\mathbf{y}$ is defined as

$$
E(\mathbf{y}) = \begin{bmatrix} E(Y_1) \\ E(Y_2) \\ \vdots \\ E(Y_n) \end{bmatrix}.
$$

The variance of a random vector isn't a number. Instead, it is the matrix of covariances of all pairs of random variables in the random vector. The variance of $\mathbf{y}$ is

$$\begin{aligned}
\operatorname{var}(\mathbf{y}) &= E(\mathbf{y}\mathbf{y}^T) - E(\mathbf{y})E(\mathbf{y})^T \\
&= \begin{bmatrix} \operatorname{var}(Y_1) & \operatorname{cov}(Y_1, Y_2) & \dots & \operatorname{cov}(Y_1, Y_n) \\ \operatorname{cov}(Y_2, Y_1) & \operatorname{var}(Y_2) & \dots & \operatorname{cov}(Y_2, Y_n) \\ \vdots & \vdots & \vdots & \vdots \\ \operatorname{cov}(Y_n, Y_1) & \operatorname{cov}(Y_n, Y_2) & \dots & \operatorname{var}(Y_n) \end{bmatrix}.
\end{aligned}$$

Alternatively, the variance of $\mathbf{y}$ is called the **covariance matrix** of $\mathbf{y}$ or the **variance-covariance matrix** of $\mathbf{y}$.

Note: $\operatorname{var}(\mathbf{y}) = \operatorname{cov}(\mathbf{y}, \mathbf{y})$.

Let $\mathbf{x} = [X_1, X_2, \dots, X_n]$ be an $n \times 1$ random vector.

The covariance matrix between $\mathbf{x}$ and $\mathbf{y}$ is defined as

$$\operatorname{cov}(\mathbf{x}, \mathbf{y}) = E(\mathbf{x}\mathbf{y}^T) - E(\mathbf{x})E(\mathbf{y})^T.$$

Properties of transformations of random vectors

Define:

- $\mathbf{a}$ to be an $n \times 1$ vector of constants (not necessarily the same constant).
- $\mathbf{A}$ to be an $m \times n$ matrix of constants (not necessarily the same constant).
- $\mathbf{x} = [X_1, X_2, \dots, X_n]$ to be an $n \times 1$ random vector.
- $\mathbf{y} = [Y_1, Y_2, \dots, Y_n]$ to be an $n \times 1$ random vector.
- $\mathbf{z} = [Z_1, Z_2, \dots, Z_n]$ to be an $n \times 1$ random vector.
- $0_{n\times n}$ to be an $n \times n$ matrix of zeros.

Then:

- $E(\mathbf{A}\mathbf{y}) = \mathbf{A}E(\mathbf{y})$.
- $E(\mathbf{y}\mathbf{A}^T) = E(\mathbf{y})\mathbf{A}^T$.
- $E(\mathbf{x} + \mathbf{y}) = E(\mathbf{x}) + E(\mathbf{y})$.
- $\operatorname{var}(\mathbf{A}\mathbf{y}) = \mathbf{A}\operatorname{var}(\mathbf{y})\mathbf{A}^T$.
- $\operatorname{cov}(\mathbf{x} + \mathbf{y}, \mathbf{z}) = \operatorname{cov}(\mathbf{x}, \mathbf{z}) + \operatorname{cov}(\mathbf{y}, \mathbf{z})$.
- $\operatorname{cov}(\mathbf{x}, \mathbf{y} + \mathbf{z}) = \operatorname{cov}(\mathbf{x}, \mathbf{y}) + \operatorname{cov}(\mathbf{x}, \mathbf{z})$.
- $\operatorname{cov}(\mathbf{A}\mathbf{x}, \mathbf{y}) = \mathbf{A}\ \operatorname{cov}(\mathbf{x}, \mathbf{y})$.
- $\operatorname{cov}(\mathbf{x}, \mathbf{A}\mathbf{y}) = \operatorname{cov}(\mathbf{x}, \mathbf{y})\mathbf{A}^T$.
- $\operatorname{var}(\mathbf{a}) = 0_{n\times n}$.
- $\operatorname{cov}(\mathbf{a}, \mathbf{y}) = 0_{n\times n}$.
- $\operatorname{var}(\mathbf{a} + \mathbf{y}) = \operatorname{var}(\mathbf{y})$.

Example (Continuous bivariate distribution continued)

Using the definitions we introduced, we want to answer **Q7** of the hydration example. Summarizing only the essential details, we have a random vector $\mathbf{z} = [X, Y]$ with mean $E(\mathbf{z}) = [2/5, 4/5]$ and covariance matrix

$$\text{var}(\mathbf{z}) = \begin{bmatrix} 14/225 & 1/75 \\ 1/75 & 2/75 \end{bmatrix}.$$

Determine $E(Y - X)$ and $\text{var}(Y - X)$.

Define $\mathbf{A} = [-1, 1]^T$ (the ROW vector with 1 and -1). Then,

$$\mathbf{Az} = \begin{bmatrix} -1 & 1 \end{bmatrix} \begin{bmatrix} X \\ Y \end{bmatrix} = Y - X$$

and,

$$\begin{aligned} E(Y - X) &= E(\mathbf{Az}) \\ &= \begin{bmatrix} -1 & 1 \end{bmatrix} \begin{bmatrix} 2/5 \\ 4/5 \end{bmatrix} \\ &= -2/5 + 4/5 \\ &= 2/5. \end{aligned}$$

Additionally,

$$\begin{aligned} &\text{var}(Y - X) \\ &= \text{var}(\mathbf{Az}) \\ &= \mathbf{A}\textbf{var}(\mathbf{z})\mathbf{A}^T \\ &= \begin{bmatrix} -1 & 1 \end{bmatrix} \begin{bmatrix} 14/225 & 1/75 \\ 1/75 & 2/75 \end{bmatrix} \begin{bmatrix} -1 \\ 1 \end{bmatrix} \\ &= \begin{bmatrix} -14/225 + 1/75 & -1/75 + 2/75 \end{bmatrix} \begin{bmatrix} -1 \\ 1 \end{bmatrix} \\ &= 14/225 + 2/75 - 2(1/75) \\ &= 14/225. \end{aligned}$$

Multivariate normal (Gaussian) distribution

Definition

The random vector $\mathbf{y} = [Y_1, \dots, Y_n]$ has a multivariate normal distribution with mean $E(\mathbf{y}) = \mu$ (an $n \times 1$ vector) and covariance matrix $\text{var}(\mathbf{y}) = \Sigma$ (an $n \times n$ matrix) if its joint pdf is,

$$f(\mathbf{y}) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{y}-\mu)^T\Sigma^{-1}(\mathbf{y}-\mu)\right),$$

where $|\Sigma|$ is the determinant of Σ. Note that Σ must be symmetric and positive definite.

In this case, we would denote the distribution of $\mathbf{y}$ as

$$\mathbf{y} \sim \mathsf{N}(\mu, \Sigma).$$

Linear functions of a multivariate normal random vector

A linear function of a multivariate normal random vector (i.e., $\mathbf{a} + \mathbf{A}\mathbf{y}$, where $\mathbf{a}$ is an $m \times 1$ vector of constant values and $\mathbf{A}$ is an $m \times n$ matrix of constant values) is also multivariate normal (though it could collapse to a single random variable if $\mathbf{A}$ is a $1 \times n$ vector).

Application: Suppose that $\mathbf{y} \sim \mathsf{N}(\mu, \Sigma)$. For an $m \times n$ matrix of constants $\mathbf{A}$, $\mathbf{A}\mathbf{y} \sim \mathsf{N}(\mathbf{A}\mu, \mathbf{A}\Sigma\mathbf{A}^T)$.

More generally, the most common estimators used in linear regression are linear combinations of a (typically) multivariate normal random vector, meaning that many of the estimators also have a (multivariate) normal distribution.

Example (OLS matrix form)

Ordinary least squares regression is a method for fitting a linear regression model to data. Suppose that we have observed variables $X_1, X_2, X_3, \dots, X_{p-1}, Y$ for each of n subjects from some population, with $X_{i,j}$ denoting the value of X_j for observation i and Y_i denoting the value of Y for observation i. In general, we want to use $X_1, \dots, X_{p-1}$ to predict the value of Y. Let,

$$\mathbf{X} = \begin{bmatrix} 1 & X_{1,1} & X_{1,2} & \cdots & X_{1,n} \\ 1 & X_{2,1} & X_{2,2} & \cdots & X_{2,n} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & X_{n,1} & X_{n,2} & \cdots & X_{n,n} \end{bmatrix}$$

be a full-rank matrix of size $n \times p$ and

$$\mathbf{y} = (Y_1, Y_2, \dots, Y_n)^T,$$

be an $n \times 1$ vector of responses. It is common to assume that,

$$\mathbf{y} \sim \mathsf{N}(\mathbf{X}\beta, \sigma^2 \mathbf{I}_{n\times n}).$$

where $\beta = (\beta_0, \beta_1, \ldots, \beta_{p-1})$ is a p-dimensional vector of constants.

The matrix $\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ projects $\mathbf{y}$ into the space spanned by the vectors in $\mathbf{X}$.

Determine the distribution of $\mathbf{Hy}$.