1 Introduction to Probability

Probability is the area of mathematics concerned with describing uncertain or random events.

1.1 Basic Concepts

We first begin by recalling some crucial definitions that will appear throughout this course.

1.1.1 Probability Space

A probability space $(\Omega, \mathcal{F}, \mathbb{P})$ consists of three parts: Ω the set of all possible outcomes, $\mathcal{F}$ the events we can assign probability to, and $\mathbb{P}$ the probability of each event. This forms the foundation of probability theory.

Definition 1. A nonempty set Ω is called a **sample space** denotes the set of all possible outcomes.

Definition 2. A collection of events $\mathscr{F}$ is called a **σ-algebra** on Ω if

- $\Omega \in \mathscr{F}$
- if $A \in \mathscr{F}$, then $A^c \in \mathscr{F}$
- if $A_1, A_2, \ldots \in \mathscr{F}$, then $\bigcup_{i=1}^{\infty} A_i \in \mathscr{F}$.

The elements of $\mathcal{F}$ are often called **events**.

Remark 1. When $\Omega = \mathbb{R}^n$, it is typically not possible to take $\mathcal{F}$ to be the power set consisting of all subsets of $\mathbb{R}^n$. Some sets in $\mathbb{R}^n$ are very weird, and need to be excluded. The classical choice of σ-algebra on $\mathbb{R}^n$ is the **Borel σ-algebra**

$$\mathscr{B}(\mathbb{R}^n),$$

which is defined as the smallest σ-algebra that contains all open cubes

$$(a_1, b_1) \times \cdots \times (a_n, b_n) \qquad \text{with } a_i < b_i$$

It is quite difficult to construct sets that are not Borel sets., so $\mathscr{B}(\mathbb{R}^n)$ contains all "reasonable" sets.

Definition 3. A function $\mathbb{P}$ is a **probability measure** on the σ-algebra $\mathcal{F}$ if

1. $\mathbb{P}(\Omega) = 1$;
2. $\mathbb{P} \geq 0$ for all $A \in \mathcal{F}$;
3. $\mathbb{P}$ is countably additive, i.e. if $A_1, A_2, \ldots \in \mathscr{F}$ are disjoint ($A_i \cap A_j = \emptyset$ for all $i \neq j$), then

$$\mathbb{P}\left(\bigsqcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mathbb{P}(A_i).$$

1.1.2 Independence

A natural property of many events and random variables is independence. We will see later that independence allows us to control the randomness and make very precise statements about multiple random events.

Definition 4 (Independence of Events)**.** Two events A, B are called **pairwise independent** (under the probability measure $\mathbb{P}$) if

$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\,\mathbb{P}(B).$$

More generally, n events $A_1, \ldots, A_n$ are **independent** if for all $k = 2, \ldots, n$ and distinct numbers $i_1, \ldots, i_k \in \{1, \ldots, n\}$

$$\mathbb{P}(A_{i_1} \cap \cdots \cap A_{i_k}) = \mathbb{P}(A_{i_1}) \cdots \mathbb{P}(A_{i_k}).$$

Remark 2. Independence is stronger than requiring only pairwise independence (see Problem 1.9)

1.1.3 Conditional Probabilities

Our probabilities change when given more information. Conditional probabilities are the way to describe how probabilities change as we get more information.

Definition 5. Let any probability space $(\Omega, \mathscr{F}, \mathbb{P})$ be given. For two events $A, B \in \mathscr{F}$ with $\mathbb{P}(B) \neq 0$, the **conditional probability** $\mathbb{P}(A|B)$ is defined as

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}.$$

This leads to the alternative intuitive definition of independence.

Definition 6. If A and B are independent, then

$$\mathbb{P}(A|B) = \mathbb{P}(A).$$

The following rule allows us to compute the probability an event by breaking it down into cases.

Theorem 1 (*Law of Total Probability*)

Consider a ***partition*** $B_1, B_2, \ldots$ *of* Ω*, i.e.,* $B_i \cap B_j = \emptyset$ *and* $\bigcup_{i=1}^{\infty} B_i = \Omega$*. Assume that* $A, B_1, B_2, \ldots \in \mathscr{F}$*. It follows that*

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

Remark 3. The conditions of a partition essentially mean that the cases are distinct $B_i \cap B_j = \emptyset$ to avoid over counting, and that no cases are left out $\bigcup_{i=1}^{\infty} B_i = \Omega$ to avoid under counting.

1.2 Random Variables

Instead of working with abstract probability spaces $(\Omega, \mathcal{F}, \mathbb{P})$ it is often convenient to work with $\Omega = \mathbb{R}$ or $\Omega = \mathbb{R}^n$. Random variables are functions defined on the underlying probability space that encode numerical outcomes of a random experiment.

Example 1. Suppose we are interested in the number of heads obtained in an experiment of tossing two coins. Let X be the number of heads. X is a function of the outcome $\omega \in \Omega = \{HH, HT, TH, TT\}$ of our coin toss:

$$X(HH) = 2, \quad X(HT) = 1, \quad X(TH) = 1, \quad X(TT) = 0.$$

Definition 7. Consider a probability space $(\Omega, \mathscr{F}, \mathbb{P})$. A function $X : (\Omega, \mathcal{F}) \to (\mathbb{R}, \mathscr{B})$ is called an **$\mathscr{F}$-measurable random variable** if

$$\{X \leq x\} := \{\omega \in \Omega : X(\omega) \leq x\} \in \mathscr{F}, \quad \text{for all } x \in \mathbb{R}.$$

The set $\mathscr{B}$ is called the **Borel σ-algebra**, the σ-algebra generated by the open sets.

The distribution of the random variable encodes the information of a probability measure on $\mathbb{R}$.

Definition 8. The function defined as

$$F_X(x) = \mathbb{P}(X \leq x) := \mathbb{P}(\{X \leq x\}), \quad x \in \mathbb{R}$$

is called the **cumulative distribution function (CDF)** of X.

Theorem 2 (*Characterization of a CDF*)

The cdf F satisfies

(i) right-continuous,

(ii) non-decreasing,

(iii) satisfies $\lim_{x\to-\infty} F(x) = 0$ and $\lim_{x\to\infty} F(x) = 1$.

Conversely, any function F with these properties (i), (ii) and (iii) is the cdf of some random variable.

Definition 9. For a random variable X, the k**th moment** is defined as

$$\mathbb{E}[X^k] = \int_{-\infty}^{+\infty} x^k \, \mathrm{d}F_X(x).$$

The k**th central moment** is defined as

$$\mathbb{E}[(X-\mu)^k] = \int_{-\infty}^{+\infty} (x-\mu)^k \, \mathrm{d}F_X(x).$$

The first moment is called the **mean** or **expectation** of X, and it is commonly denoted by μ_X or μ. The second central moment is called the **variance** of X and it is denoted by σ_X^2 or simply σ^2. The square root of the variance, i.e., σ_X or σ, is called the **standard deviation**.

Remark 4. The mean of the random variable can be interpreted as the average value of many samples of a random variable. The variance of the random variable can be interpreted as a measure of the spread of the random variable.

More generally, the expected value of a function of a random variable is computed in the obvious way.

Proposition 1 (*Law of the Unconscious Statistician*)

For any real-valued function g, we have

$$\mathbb{E}[g(X)] = \int_{-\infty}^{+\infty} g(x) \, \mathrm{d}F_X(x),$$

and the integration has different expressions depending on the type (discrete, continuous, and mixed types) of X.

In this course, we will commonly work with two classes of random variables.

Definition 10. We say that X is a **discrete random variable** if X only takes at most countable (including finitely countable) number of real values $\{x_1, x_2, ...\}$. We denote by p_X its **probability mass function (pmf)**, i.e.,

$$p_X(x_i) = \mathbb{P}(X = x_i), \quad i \in \mathbb{N}.$$

For a discrete random variable X, we have that

- cdf:

$$F_X(x) = \mathbb{P}(X \le x) = \sum_{i=1}^{\infty} p_X(x_i) \, \mathbb{1}[x_i, \infty)(x)$$

with $\mathbb{1}[x_i, \infty)$ denoting the **indicator function** of the interval $[x_i, \infty)$.

- for a general function g,
$$\mathbb{E}\left[g(X)\right] = \sum_{i\in\mathbb{N}} g(x_i)p_X\left(x_i\right)$$

Example 2. Examples of discrete distributions include the Poisson, binomial, and negative binomial distributions.

Definition 11. We say that X is a **continuous random variable** if the distribution function of X is continuous everywhere and differentiable almost everywhere. We denote by f_X or f its **probability density function (pdf)**, i.e.,

$$f_X(x) = \frac{\mathrm{d}}{\mathrm{d}x}F_X(x) = F_X'(x), \qquad x \in \mathbb{R}.$$

For a continuous random variable X, we have that

- cdf:
$$F_X\left(x\right) = \int_{-\infty}^{x} f_X(t)\mathrm{d}t, \qquad x \in \mathbb{R}.$$
- for a general function g,
$$\mathbb{E}\left[g(X)\right] = \int_{-\infty}^{\infty} g(t)f_X(t)\mathrm{d}t$$

Example 3. Examples of continuous distributions include the exponential, Gamma, Weibull, and normal distributions.

Definition 12. We call X a **mixed random variable** if it has both discrete and continuous components.

For a mixed random variable X, let $\{x_1, x_2, ...\}$ be those real numbers x for which $p(x) = \mathbb{P}(X = x) > 0$, and let f be a density function for the continuous component. We have that

- cdf:
$$F_X(x) = \mathbb{P}(X \le x) = \sum_{i=1}^{\infty} p(x_i)\mathbb{1}[x_i\,\infty)(x) + \int_{-\infty}^{x} f(y)\,\mathrm{d}y.$$
- for a general function g,
$$\mathbb{E}\left[g(X)\right] = \sum_{i=1}^{\infty} g(x_i)p(x_i) + \int_{-\infty}^{\infty} g(x)f(x)\,dx$$

1.2.1 Random Vectors

We now define a high dimensional analogue of a random variables, which maps the probability space to $\mathbb{R}^n$ instead of $\mathbb{R}$.

Definition 13. For a random vector $\boldsymbol{X} = (X_1, X_2, ..., X_n)$ defined on a probability space $(\Omega, \mathscr{F}, \mathbb{P})$, we define its **joint distribution function** $F_{\boldsymbol{X}} : \mathbb{R}^n \to [0, 1]$ by

$$F_{\boldsymbol{X}}\left(\boldsymbol{x}\right) = \mathbb{P}\left(\boldsymbol{X} \le \boldsymbol{x}\right) = \mathbb{P}\left(X_1 \le x_1, ..., X_n \le x_n\right)$$

where $\boldsymbol{x} = (x_1, x_2, ..., x_n) \in \mathbb{R}^n$.

If $\boldsymbol{X}$ is a vector of discrete or continuous random variables, then the joint distribution function is encoded by its mass or density function.

Definition 14. If $\boldsymbol{X}$ is a vector of discrete random variables, then the **joint mass function** $p_{\boldsymbol{X}} : \mathbb{R}^n \to [0,1]$ given by

$$p_{\boldsymbol{X}}(x_1, \ldots, x_n) = \mathbb{P}\left(X_1 = x_1, ..., X_n = x_n\right)$$

and if $\boldsymbol{X}$ is a vector of continuous random variables, then the **joint density function** $f_{\boldsymbol{X}} : \mathbb{R}^n \to [0, \infty)$ is given by

$$F_{\boldsymbol{X}}\left(\boldsymbol{x}\right) = \int_{-\infty}^{x_1} \int_{-\infty}^{x_2} \cdots \int_{-\infty}^{x_n} f_{\boldsymbol{X}}(y_1, \ldots, y_n)\, \mathrm{d}y_n \cdots \mathrm{d}y_1.$$

Often, it makes sense to look at the law of just one component X_i of the random vector $\boldsymbol{X}$.

Definition 15. The cdf of X_i is called **marginal cdf** of X_i and denoted F_{X_i}.

Definition 16. The random variables $X_1, ..., X_n$ are **independent** if

$$F_{\boldsymbol{X}}\left(\boldsymbol{x}\right) = \prod_{i=1}^{n} F_{X_i}\left(x_i\right) \quad \text{for all } \boldsymbol{x} = (x_1, \ldots, x_n) \in \mathbb{R}^n \tag{1}$$

where F_{X_i} denotes the cdf of X_i.

Remark 5. Note that (1) can be re-written as

$$\mathbb{P}\left(\{X_1 \leq x_1\} \cap \cdots \cap \{X_n \leq x_n\}\right) = \mathbb{P}(X_1 \leq x_1) \cdots \mathbb{P}(X_n \leq x_n), \quad \text{for all } x_1, \ldots x_n \in \mathbb{R},$$

which coincides with the notion of independence for sets.

Proposition 1. *The random variables $X_1, ..., X_n$ are independent if and only if for all functions $g_1, \ldots, g_n$,*

$$\mathbb{E}\left[\prod_{i=1}^{n} g_i(X_i)\right] = \prod_{i=1}^{n} \mathbb{E}[g_i(X_i)].$$

Proposition 2. *If $X_1, \ldots X_n$ are independent random variables and $g_1, \ldots, g_n$ are functions, then $g_1(X_1), \ldots, g_n(X_n)$ are also independent.*

Definition 17.

- The **covariance** for two random variables X_1, X_2 is defined as

$$\mathrm{Cov}\left(X_1, X_2\right) = \mathbb{E}\left[\left(X_1 - \mathbb{E}\left[X_1\right]\right)\left(X_2 - \mathbb{E}\left[X_2\right]\right)\right]$$

- The **variance** of a random variable X is defined as average squared deviation from the mean:

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

Note that the covariance is a measure for the possible dependence for two random variables, and the variance is a common measure for the variability of a random variable.

Proposition 3. *We have*

1. *For any two random variables*

$$\mathrm{Cov}\left(X_1, X_2\right) = \mathbb{E}\left[X_1 X_2\right] - \mathbb{E}\left[X_1\right]\mathbb{E}\left[X_2\right]$$

2. *If X_1 and X_2 are independent, then* $\mathrm{Cov}\left(X_1, X_2\right) = 0$.

Remark 6. However, the converse of Proposition 3 (b) is not true! Two random variables with zero covariance are called as **uncorrelated random variables.** Note that uncorrelated random variables are NOT necessarily independent (see Problem 1.15).

Proposition 4. *The variance has the following properties.*

1. *For any random variable X,*
$$\mathrm{Var}(X) = \mathbb{E}[X^2] - \big(\mathbb{E}[X]\big)^2$$
2. *If $a, b \in \mathbb{R}$, then*
$$\mathrm{Var}(aX + b) = a^2\mathrm{Var}(X)$$
3. *For any two random variables X and Y,*
$$\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\mathrm{Cov}(X, Y)$$
4. *If X and Y are independent random variables, then*
$$\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$$

1.3 Transformations of random variables

Various transformations can be applied to random variables. A substantial number of well-known distributions can be viewed as transformations of other distributions.

1.3.1 Change of Variables in One-Dimension

For a random variable X and a known function g, we define the random variable $Y := g(X)$. One may be interested in the distribution of $Y = g(X)$, namely

$$F_Y(y) = F_{g(X)}(y) = \mathbb{P}(g(X) \leq y).$$

subsectionChange of Variables Formula (Discrete Random Variables)

Suppose that we know the PDF $f_X(x)$ of X. Our goal is to recover the the PMF $f_Y(y)$ of the random variable $Y = g(X)$. This can be done directly using the following steps

1. Using the range of X, find the range of $Y = g(X)$ denoted by computing the image of the range of X:
$$Y(S) = g(X(S)).$$
2. Compute the PMF of Y by expressing it in terms of the PMF of X:
$$f_Y(y) = \mathbb{P}(Y = y) = \mathbb{P}(g(X) = y) = \sum_{x:g(x)=y} \mathbb{P}(X = x) = \sum_{x \in g^{-1}(\{y\})=y} f_X(x).$$

This procedure is summarized by the change of variables formula.

Theorem 3 (*Change of Variables Formula (Discrete)*)

If X is discrete and $g : \mathbb{R} \to \mathbb{R}$, then

$$f_Y(y) = \sum_{x \in g^{-1}(\{y\})} f_X(x), \quad y \in Y(S) = g(X(S)).$$

subsectionChange of Variables Formula (Continuous Random Variables)

Suppose that we know the PDF $f_X(x)$ of X. Our goal is to recover the the PDF $f_Y(y)$ of the random variable $Y = g(X)$. This can be done directly using the following steps

1. Use the support of X to find the support of $Y = g(X)$:
$$\text{supp}(Y) = \text{cl}(\{y \in \mathbb{R} : f_Y(y) > 0\}) = g(\text{supp}(X)).$$
2. Compute the CDF of Y for $y \in \text{supp}(Y)$ by expressing it in terms of the CDF of X:
$$F_Y(y) = \mathbb{P}(g(X) \leq y) = \dots$$
When the function g is not strictly increasing (or decreasing) over the support of X, then we must be careful when rewriting the inequality $\mathbb{P}(g(X) \leq y)$.
3. Compute the PDF of Y by differentiating the CDF of Y,
$$f_Y(y) = F_Y'(y) \qquad y \in \text{supp}(Y).$$

Remark 7. Technically, the function F_Y is only differentiable on the interior of $\text{supp}(Y)$. This is not an issue since we can simply define $f_Y(y)$ on the boundaries by continuity.

When g is invertible, the above procedure gives us the change of density formula.

Theorem 4 (*Change of Variables Formula (Continuous)*)

Let X_1 and X_2 be a pair of continuous random variables with joint density f_{X_1,X_2}. Let X be a continuous random variable and g be invertible and differentiable with inverse g^{-1} on the support of Y, then

$$f_Y(y) = |(g^{-1})'(y)| f_X(g^{-1}(y)) = \frac{1}{|g'(g^{-1}(y))|} f_X(g^{-1}(y)), \quad y \in \text{supp}(Y).$$

1.3.2 Change of Variables in Higher Dimensions

We now consider bivariate cases (general n dimensional random variables are handled similarly). Let X_1 and X_2 be two random variables with joint density function f_{X_1,X_2}. The following rule from calculus provides us a recipe to compute the joint density of transformations of X_1 and X_2.

Theorem 5

Let $Y_1 = g(X_1, X_2)$ and $Y_2 = h(X_1, X_2)$ and suppose the map

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \longmapsto \begin{pmatrix} g(x_1, x_2) \\ h(x_1, x_2) \end{pmatrix} = \begin{pmatrix} y_1 \\ y_2 \end{pmatrix}$$

is invertible. Denote its inverse by

$$\begin{pmatrix} x_1(y_1, y_2) \\ x_2(y_1, y_2) \end{pmatrix}.$$

Then the joint density of (Y_1, Y_2) is given by

$$f_{Y_1,Y_2}(y_1, y_2) = f_{X_1,X_2}(x_1(y_1, y_2), x_2(y_1, y_2)) |J(y_1, y_2)|,$$

on the possible region for (y_1, y_2), where the Jacobian of the transformation is given by

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

We will apply this formula to compute the distribution of the sums of random variables.

Proposition 2

Let X_1 and X_2 be two random variables with joint density function f_{X_1,X_2}. Define $S = X_1 + X_2$. The density function of S is given by

$$f_S(s) = \int_{-\infty}^{\infty} f_{X_1,X_2}(x, s-x)\, dx. \tag{2}$$

Proof. To see why this is so, consider the bivariate transformation

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \longmapsto \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} := \begin{pmatrix} x_1 \\ x_1 + x_2 \end{pmatrix}$$

which maps (X_1, X_2) into (X_1, S). The inverse is given by

$$x_1(y_1, y_2) = y_1, \quad x_2(y_1, y_2) = y_2 - y_1$$

Hence, the Jacobian is

$$J(y_1, y_2) = \begin{vmatrix} \frac{\partial x_1}{\partial y_1} & \frac{\partial x_2}{\partial y_1} \\ \frac{\partial x_1}{\partial y_2} & \frac{\partial x_2}{\partial y_2} \end{vmatrix} = \begin{vmatrix} 1 & -1 \\ 0 & 1 \end{vmatrix} = 1.$$

Therefore, by the result from Section 1.3.1, the joint density of X_1 and S is

$$f_{X_1,S}(x, s) = f_{X_1,X_2}(x, s-x)$$

and so the marginal density of S is

$$f_S(s) = \int f_{X_1,S}(x, s)\, \mathrm{d}x = \int f_{X_1,X_2}(x, s-x)\, \mathrm{d}x.$$

□

A special case of this formula arises when X_1 and X_2 are independent.

Corollary 1

When X_1 and X_2 are two independent random variables, (2) becomes

$$f_S(s) = \int_{-\infty}^{\infty} f_{X_1}(x)\, f_{X_2}(s-x)\, \mathrm{d}x \tag{3}$$

*and f_S is called **convolution** of the densities f_{X_1} and f_{X_2}.*

Definition 18. A sequence $X_1, X_2, \ldots$ of random variables on $(\Omega, \mathscr{F}, \mathbb{P})$ is called **independent and identically distributed (i.i.d.)** if the random variables are independent and they have the same cdf.

The next result states that the sample mean converges to its expected value. This reconciles the frequentist and the axiomatic approaches to probability.

Theorem 6 (*Strong Law of Large Numbers*)

Let $X_1, X_2, \ldots$ be an i.i.d. sequence of random variables. Then there exists an event $\Omega_0 \in \mathscr{F}$ such

that $\mathbb{P}(\Omega_0) = 1$ *and*

$$\frac{1}{n}\sum_{k=1}^{n} X_k(\omega) \longrightarrow \mathbb{E}[X_1] \quad \text{as } n \to \infty, \qquad \text{for all } \omega \in \Omega_0.$$

In other words, the sample mean converges to the theoretical mean ***almost surely****.*

The central limit theorem is, next to the law of large numbers, one of the main pillars of modern probability and statistics. It states that the distribution of the standardized sample mean has a universal limit.

Theorem 7 (*Central Limit Theorem*)

Let $X_1, X_2, \ldots$ *an i.i.d. sequence with finite variance* $\sigma^2 := \mathrm{Var}(X_i)$ *and common mean* $\mu := \mathbb{E}[X_i]$*. Define the standardized sample mean*

$$\widehat{S}_n := \frac{1}{\sigma\sqrt{n}}\sum_{k=1}^{n}(X_k - \mu).$$

Then, for any $x \in \mathbb{R}$*,*

$$\mathbb{P}(\widehat{S}_n \leq x) \longrightarrow \Phi(x), \qquad \text{as } n \to \infty$$

where Φ *is the cdf of* $N(0,1)$*. In other words, the standardized sample mean converges to a Gaussian random variable* ***in distribution****.*

1.3.3 Laplace transform

We now define a generating function for a random variable which encodes the distribution as a real valued function.

Definition 19. The **Laplace transform** of a positive random variable X:

$$L_X(t) = \mathbb{E}\left[e^{-tX}\right], \quad t \geq 0.$$

The Laplace transform of a positive random variable is sometimes much simpler than its cdf. Moreover, there is a one-to-one correspondence between a Laplace transform and a cdf. The Laplace transform technique is also very useful to study the sum of a sequence of positive and independent random variables.

Proposition 3

Let $X_1, ..., X_n$ *be independent random variables with Laplace transforms*

$$L_{X_i}(t) = \mathbb{E}\left[e^{-tX_i}\right], \quad i = 1, ..., n.$$

Define the sum $S = X_1 + \cdots + X_n$*. Then*

$$L_S(t) = \prod_{i=1}^{n} L_{X_i}(t).$$

Proof. By independence, the Laplace transform of S is given by

$$L_S(t) = \mathbb{E}\left[e^{-tS}\right] = \mathbb{E}\left[e^{-t(X_1+\cdots+X_n)}\right] = \mathbb{E}\left[\prod_{i=1}^{n} e^{-tX_i}\right] = \prod_{i=1}^{n} L_{X_i}(t).$$

□

1.4 Generating random variables with given distribution

1.4.1 The quantile method

There is a nice relationship between general random variables and uniform random variables.

Proposition 4

Let X be a random variable with continuous CDF F_X. Then $U := F_X(X)$ has a uniform distribution on $(0,1)$. That is,

$$F_U(t) = \mathbb{P}(U \le t) = \begin{cases} 0 & \text{if } t \le 0, \\ t & \text{if } 0 < t < 1, \\ 1 & \text{if } t \ge 1. \end{cases}$$

This suggests a way to generate random variables if one can "invert" the CDF.

Definition 20. Let F_X be an arbitrary CDF and define the **quantile function** of X as

$$F_X^{-1}(t) = \min\{x \in \mathbb{R} : F_X(x) \ge t\}, \qquad 0 < t < 1.$$

Then $F_X^{-1}(t)$ will be a $t \times 100\%$-quantile for the distribution of X. In particular $F_X^{\leftarrow}(\frac{1}{2})$ is called the **median**.

The quantile function can be interpreted as a generalized inverse of F_X. If F_X is strictly increasing, then the quantile function is simply the inverse function of F_X. The next result allows us to generate random variables using uniform random variables.

Proposition 5 (*Inverse Transform Sampling*)

Suppose that U is a random variable with a uniform distribution on $(0,1)$ and F is an arbitrary cdf. Then

$$X := F^{-1}(U)$$

has cdf F.

The effectiveness of this method depends on the efficiency with which F^{-1} can be computed. If this is relatively easy, as for exponential or Gumbel distributions, then the method is highly effective, as only one uniform random variable needs to be generated for each random variable with cdf F. In many cases. however, the computation of F^{-1} is difficult or computationally very costly.

1.4.2 The Box–Muller algorithm

The following is method to generate independent normally distributed random variables. It uses the rotational invariance of normally distributed random variables.

Corollary 2

Let U_1 and U_2 be two independent random variables with a uniform distribution on $(0,1)$. Then

$$X_1 := \sqrt{-2\log U_1}\cos(2\pi U_2)$$
$$X_2 := \sqrt{-2\log U_1}\sin(2\pi U_2)$$

are independent $N(0,1)$-distributed random variables.

Proof. Suppose X_1 and X_2 are two independent standard normal random variables. Writing (X_1, X_2) in polar coordinates gives rise to two random variables R and Ψ such that

$$R := \sqrt{X_1^2 + X_2^2}$$

and Ψ is such that

$$X_1 = R\cos\Psi, \qquad X_2 = R\sin\Psi.$$

Then

$$x_1(r,\psi) = r\cos\psi \qquad x_2(r,\psi) = r\sin\psi,$$

and so

$$J(r,\psi) = \begin{vmatrix} \frac{\partial x_1}{\partial r} & \frac{\partial x_2}{\partial r} \\ \frac{\partial x_1}{\partial \psi} & \frac{\partial x_2}{\partial \psi} \end{vmatrix} = \begin{vmatrix} \cos\psi & \sin\psi \\ -r\sin\psi & r\cos\psi \end{vmatrix} = r(\cos\psi)^2 + r(\sin\psi)^2 = r.$$

Therefore, the joint density of R and Ψ is

$$f_{R,\Psi}(r,\psi) = \frac{1}{2\pi} e^{-((r\cos\psi)^2+(r\sin\psi)^2)/2} \cdot r \cdot \mathbb{1}[0,2\pi)(\psi) = \frac{r}{2\pi} e^{-r^2/2} \cdot \mathbb{1}[0,2\pi)(\psi)$$

It follows that the joint CDF of R and Ψ is for $a > 0$ and $b \in [0, 2\pi)$ given by

$$\begin{aligned} \mathbb{P}(R \le a,\, \Psi \le b) &= \int_0^a \int_0^b \frac{r}{2\pi} e^{-r^2/2}\, \mathrm{d}\psi\, \mathrm{d}r = \frac{b}{2\pi} \cdot \int_0^a r e^{-r^2/2}\, \mathrm{d}r \\ &\overset{s=r^2}{=} \frac{b}{2\pi} \cdot \frac{1}{2} \int_0^{a^2} e^{-s/2}\, \mathrm{d}s \end{aligned}$$

Thus, Ψ has a uniform distribution on $[0, 2\pi)$, R has the same distribution as $\sqrt{Z}$, where Z has an exponential distribution with parameter $1/2$, and Ψ and R are independent. □

The previous result provides an algorithm for the exact simulation of $N(0,1)$-random variables, provided that one is able to generate independent uniform random variables (which most computer systems can).

1.4.3 Generating random vectors with a multivariate normal distribution

Let $n \in \mathbb{N}$ be a fixed dimension.

Definition 21. A random vector

$$X = \begin{pmatrix} X^1 \\ \vdots \\ X^n \end{pmatrix} = (X^1, \ldots, X^n)^\top$$

has a **multivariate normal distribution** if there exist a random vector $Z = (Z^1, \ldots, Z^n)^\top$ whose components $Z^1, \ldots, Z^n$ are independent standard normally distributed random variables, a vector $m \in \mathbb{R}^n$ and a $n \times n$-matrix A such that

$$X = AZ + m.$$

It follows that, for $i, j = 1, ..., n$,

$$\mathbb{E}[X^i] = m^i \text{ and } \mathrm{Cov}(X^i, X^j) = \mathrm{Cov}\Big(\sum_{k=1}^d A_{ik} Z^k, \sum_{\ell=1}^d A_{j\ell} Z^\ell\Big) = \sum_{k=1}^d A_{ik} A_{jk} = C_{ij}$$

where the **covariance matrix** C is given by

$$C = AA^\top.$$

We say that X has distribution $N(m, C)$ and write $X \sim N(m, C)$. In particular, $Z \sim N(0, \boldsymbol{I})$, where $\boldsymbol{I}$ is the identity matrix.

Proposition 6

The covariance matrix is symmetric and positive semidefinite.

Proof. Since $\text{Cov}(X^i, X^j) = \text{Cov}(X^j, X^i)$, the matrix C must be **symmetric**. Moreover, if $y = (y^1, \ldots, y^d)^\top \in \mathbb{R}^d$ and $\langle \cdot, \cdot \rangle$ denotes the **Euclidian inner product** on $\mathbb{R}^n$, then

$$\langle y, Cy \rangle = y^\top Cy = \text{Cov}(y^\top X, y^\top X) = \text{Var}(y^\top X) \geq 0,$$

and so C has to be **positive semidefinite**. □

Remark 8. Recall that a matrix is positive semidefinite matrix if and only if it has non-negative eigenvalues. Since real symmetric matrices $\boldsymbol{A}$ are also diagonalizable by orthogonal matrices, if $\boldsymbol{A}$ is positive semi-definite then it can be written as

$$\boldsymbol{A} = \boldsymbol{Q}\boldsymbol{D}\boldsymbol{Q}^\intercal$$

where $\boldsymbol{Q}$ is an orthgonal matrix and $\boldsymbol{D}$ is diagonal with non-negative entries.

Conversely, if C is any given symmetric and nonnegative definite $n \times n$ matrix, then there exists a $n \times n$ matrix A such that $C = AA^\top$. So C is the covariance matrix of some normally distributed random vector X.

However, A need not be unique. But a very convenient choice for A is the **Cholesky decomposition** of C into the form

$$C = LL^\top,$$

where the matrix L is a lower triangular matrix,

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

The Cholesky decomposition is unique and available through standard computational software packages.

1.5 Example Problems

1.5.1 Basic Definitions

Problem 1.1. Suppose two six sided dice are rolled, and the number of dots facing up on each die is recorded.

1. Write down the sample space S.
2. Write down, as a set, the event $A =$ "The sum of the dots is 7".
3. Write down, as a set, the event B^c, where $B =$ "The sum of the numbers is at least 4".
4. Write down, as a set, the events $A \cap B^c$ and $A \cup B^c$.

Solution 1.1.

1. The sample space for a pair of dice is the a pair of the outcomes of each die roll

$$S = \{1, \ldots, 6\} \times \{1, \ldots, 6\} = \{(x, y) : x, y \in \{1, 2, \ldots, 6\}\}$$

2. We can simply write down all the combinations

$$A = \{(1,6),(2,5),(3,4),(4,3),(5,2),(6,1)\}$$

3. If $B = \{\text{sum is at least } 4\}$ then $B^c = \{\text{sum is at most } 3\}$, so

$$B^c = \{(1,1),(1,2),(2,1)\}$$

4. It follows that $A \cap B^c = \emptyset$ and

$$A \cup B^c = \{1,6),(2,5),(3,4),(4,3),(5,2),(6,1),(1,1),(1,2),(2,1)\}$$

Problem 1.2. Show the *monotonicity* property of probability,

$$\text{if } A \subseteq B \text{ then } \mathbb{P}(A) \leq \mathbb{P}(B).$$

Solution 1.2. This follows directly from the axioms. If $A \subseteq B$, then $B = A \cup A \setminus B$ and the sets A and $A \setminus B$ are disjoint. Therefore, by countable additivity,

$$\mathbb{P}(B) = \mathbb{P}(A) + \mathbb{P}(A \setminus B) \geq \mathbb{P}(A)$$

since $\mathbb{P}(A \setminus B) \geq 0$ by the non-negativity property.

Problem 1.3. Show that the axiomatic definition of a probability implies that

$$0 \leq \mathbb{P}(A) \leq 1$$

for any event A.

Solution 1.3. Suppose for the sake of contradiction that $\mathbb{P}(A) > 1$ for some event A. By the monontonicity property, since $A \subseteq S$,

$$\mathbb{P}(S) \geq \mathbb{P}(A) > 1$$

which contradicts the fact that $\mathbb{P}(S) = 1$. Therefore, $\mathbb{P}(A) \leq 1$.

Problem 1.4. Show that the axiomatic defintiion of a probability implies that

$$\mathbb{P}(A) = 1 - \mathbb{P}(A^c)$$

for any event A.

Solution 1.4. Notice that $A \cup A^c = S$ and A and A^c are disjoint. From finite additivity, we conclude that

$$\mathbb{P}(A) + \mathbb{P}(A^c) = \mathbb{P}(S) = 1 \implies \mathbb{P}(A) = 1 - \mathbb{P}(A^c).$$

Problem 1.5. Suppose that $A \subseteq B$. Show that $\mathbb{P}(A) \leq \mathbb{P}(B)$.

Solution 1.5. This property is called *monotonicity*. By the law of total probability,

$$\mathbb{P}(B) = \mathbb{P}(A \cap B) + \mathbb{P}(A^c \cap B).$$

Since $A \subseteq B$, we have $\mathbb{P}(A \cap B) = \mathbb{P}(A)$ and $\mathbb{P}(A^c \cap B) \geq 0$, so

$$\mathbb{P}(A) \leq \mathbb{P}(B).$$

Problem 1.6. Show that for arbitrary events $A_1, A_2, \ldots, A_n$,

$$\mathbb{P}(\cup_{k=1}^n A_k) \leq \sum_{k=1}^n \mathbb{P}(A_k).$$

Solution 1.6. This is called the *union bound*. By the inclusion exclusion principle, we see that

$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$$

since $\mathbb{P}(A \cap B) \geq 0$. The general statement then follows by (strong) induction. We have

$$\begin{aligned}
\mathbb{P}(A_1 \cup A_2 \cup A_{n-1} \cup A_n) &= \mathbb{P}((A_1 \cup A_2 \cup A_{n-1}) \cup A_n) \\
&\leq \mathbb{P}(A_1 \cup A_2 \cup A_{n-1}) + \mathbb{P}(A_n) \\
&\leq \sum_{k=1}^{n-1} \mathbb{P}(A_k) + \mathbb{P}(A_n) = \sum_{k=1}^{n} \mathbb{P}(A_k).
\end{aligned}$$

Remark 9. It is true that equality is attained if $A_1, \ldots, A_n$ are mutually exclusive. However, we can have equality even if $A_1, \ldots, A_n$ are not mutually exclusive. We can consider the distribution on the two points $\{a, b\}$ such that $\mathbb{P}(\{a\}) = 0$ and $\mathbb{P}(\{b\}) = 1$. Then $A = \{a, b\}$ and $B = \{a\}$ satisfy the union bound, but they are not mutually exclusive.

Problem 1.7. Suppose that A and B are independent.

1. Show that A^c and B^c are independent.
2. Show that A and B^c are independent.

Solution 1.7.

1. We check the definition,

$$\begin{aligned}
\mathbb{P}(A^c \cap B^c) = 1 - \mathbb{P}((A^c \cap B^c)^c) &= 1 - \mathbb{P}(A \cup B) && \text{De Morgan's law} \\
&= 1 - \mathbb{P}(A) - \mathbb{P}(B) + \mathbb{P}(A \cap B) && \text{inclusion exclusion} \\
&= 1 - \mathbb{P}(A) - \mathbb{P}(B) + \mathbb{P}(A)\,\mathbb{P}(B) && \text{independence} \\
&= (1 - \mathbb{P}(A))(1 - \mathbb{P}(B)) \\
&= \mathbb{P}(A^c)\,\mathbb{P}(B^c).
\end{aligned}$$

2. We check the definition,

$$\begin{aligned}
\mathbb{P}(A^c \cap B) = \mathbb{P}(B) - \mathbb{P}(A \cap B) &= \mathbb{P}(B) - \mathbb{P}(A)\,\mathbb{P}(B) && \text{independence} \\
&= (1 - \mathbb{P}(A))\,\mathbb{P}(B) \\
&= \mathbb{P}(A^c)\,\mathbb{P}(B).
\end{aligned}$$

1.5.2 Random Variables

Problem 1.8. 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 conditional distribution of X given Y?
4. What is the probability that $X + Y = 2$?

Solution 1.8.

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

		x				
$f_{X,Y}(x,y)$		0	1	2	3	$f_Y(y)$
y	0	1/8	2/8	1/8	0	1/2
	1	0	1/8	2/8	1/8	1/2
$f_X(x)$		1/8	3/8	3/8	1/8	1

Part 2: We can see

$$f_{X,Y}(0,1) = 0 \neq \frac{1}{8} \cdot \frac{1}{2} = f_X(0) f_Y(1)$$

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: Using the formula $f_{X \mid Y}(x \mid y) = f_{X,Y}(x,y)/f_Y(y)$ we find

	x			
	0	1	2	3
$f_{X \mid Y}(x \mid y = 0)$	2/8	4/8	2/8	0
$f_{X \mid Y}(x \mid y = 1)$	0	2/8	4/8	2/8

Part 4: 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

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

Problem 1.9. Let X, Y, Z be i.i.d. Rademacher random variables, i.e.

$$\mathbb{P}(X = \pm 1) = \frac{1}{2}.$$

Show that XY, YZ, ZX are pairwise independent but not mutually independent.

Solution 1.9. To check pairwise independence, we see that

$$\mathbb{P}(XY = \pm 1, YZ = \pm 1) = \frac{1}{4} = \mathbb{P}(XY = \pm 1)\,\mathbb{P}(YZ = \pm 1)$$

The other cases are similar because any two elements from XY, YZ, ZX share exactly one common random variable. However, they are not mutually independent because

$$\mathbb{P}(XY = 1, YZ = 1, ZX = 1) = \frac{1}{4} \neq \frac{1}{8} = \mathbb{P}(XY = \pm 1)\,\mathbb{P}(YZ = \pm 1)\,\mathbb{P}(ZX = \pm 1)$$

since $\{XY = 1, YZ = 1, ZX = 1\}$ happens only when all of X, Y and Z have the same sign.

Problem 1.10. Prove the following properties for the quantile function

1. For all $x \in \mathbb{R}$, $F_X^{-1}(F_X(x)) \leq x$
2. For all $p \in [0, 1]$, $F_X(F_X^{-1}(p)) \geq p$
3. $F_X^{-1}(p) \leq x \Leftrightarrow p \leq F_X(x)$
4. $F_X^{-1}(p)$ is non-decreasing and left-continuous (except for the endpoints $p = 0$ or $p = 1$)

Solution 1.10.

1. We have
$$F_X^{-1}(F_X(x)) = \inf_{t \in \mathbb{R}} \{F_X(t) \geq F_X(x)\} \leq x$$
since $x \in \{t \in \mathbb{R} : F_X(t) \geq F_X(x)\}$.

2. Since F_X is right continuous and increasing we have $\{F_X(x) \geq p\}$ is a closed set, so it attains its infimum. Therefore, $c_p \in \{F_X(x) \geq p\}$ so
$$F_X(F_X^{-1}(p)) = F_X(c_p) \geq p.$$

3. On one hand, $F_X^{-1}(p) \leq x$ implies that $x \in \{t : F_X(t) \geq p\}$ so $p \leq F_X(x)$. On the other hand, if $p \leq F_X(x)$ then $x \in \{t : F_X(t) \geq p\}$ so $F_X^{-1}(p) \leq x$ since $F_X^{-1}(p)$ is the infimum of all $\{t : F_X(t) \geq p\}$.

4. Suppose that $p_1 \leq p_2$. Then
$$F_X^{-1}(p_1) = \inf_{x \in \mathbb{R}} \{F_X(x) \geq p_1\} \leq \inf_{x \in \mathbb{R}} \{F_X(x) \geq p_2\} = F_X^{-1}(p_2)$$
since $\{F_X(x) \geq p_1\} \subseteq \{F_X(x) \geq p_2\}$, so F_X^{-1} is non-decreasing.

   To see left continuity, notice that monotone functions can only have jump discontinuities, so it suffices to show that $\sup_{q<p} F_X^{-1}(q) = F_X^{-1}(p)$. For each $q < p$ and $\epsilon > 0$, we have by definition of the supremum
$$\sup_{q<p} F_X^{-1}(q) + \epsilon \geq F_X^{-1}(q) \overset{(3)}{\Longrightarrow} F_X(\sup_{q<p} F_X^{-1}(q) + \epsilon) \geq q.$$
   So taking $\epsilon \to 0$ by right continuity of F_X implies that $F_X(\sup_{q<p} F_X^{-1}(q)) \geq q$ for all $q < p$ so $F_X(\sup_{q<p} F_X^{-1}(q)) \geq p$. Property 3 above implies that
$$\sup_{q<p} F_X^{-1}(q) \geq F_X^{-1}(p).$$
   This combined with monotonicity $\sup_{q<p} F_X^{-1}(q) \leq F_X^{-1}(p)$ implies that $\sup_{q<p} F_X^{-1}(q) = F_X^{-1}(p)$ as required.

Problem 1.11. Let X be an exponential random variable with mean θ, and define $Y = X^{\frac{1}{s}}$ for $s > 0$. What is the distribution of Y?

Solution 1.11. It follows that for $y > 0$

$$\begin{aligned} F_Y(y) &= \mathbb{P}\left(X^{\frac{1}{s}} \leq y\right) \\ &= \mathbb{P}\left(X \leq y^s\right) \\ &= F_X\left(y^s\right) \\ &= 1 - e^{-\frac{y^s}{\theta}} \\ &= 1 - e^{-\left(\frac{y}{\alpha}\right)^s}, \quad y \geq 0, \end{aligned}$$

where $\alpha = \theta^{\frac{1}{s}}$. Hence, Y follows a Weibull distribution with parameters α and s.

Problem 1.12. Suppose $F(x) = 0$ if $x \leq 0$ and $F(x) = 1 - e^{-\lambda x}$ for $x > 0$, where $\lambda > 0$ (that is, F is the cdf of an exponential distribution with parameter λ and mean $1/\lambda$). What is the quantile function of X? How can you generate a realization of X using a uniform random variable?

Solution 1.12. We have

$$F^{-1}(t) = -\frac{1}{\lambda}\log(1-t).$$

So if U has a uniform distribution on $(0, 1)$, then $-\frac{1}{\lambda}\log(1-U)$ will have an exponential distribution with parameter λ.

Problem 1.13. Compute the Laplace transform of a Poisson distributed random variable with parameter $\lambda > 0$.

Solution 1.13. Let X have a Poisson distribution with parameter $\lambda > 0$. Then

$$\begin{aligned} L_X(t) &= \mathbb{E}\left[e^{-tX}\right] = \sum_{k=0}^{\infty} e^{-tk} \cdot \mathbb{P}[X = k] \\ &= \sum_{k=0}^{\infty} e^{-tk} e^{-\lambda} \frac{\lambda^k}{k!} = e^{-\lambda} \sum_{k=0}^{\infty} \frac{(e^{-t}\lambda)^k}{k!} \\ &= e^{-\lambda} e^{e^{-t}\lambda} = e^{-\lambda(1-e^{-t})} \end{aligned}$$

Problem 1.14. If $X \sim \text{Poi}(\lambda)$ and $Y \sim \text{Poi}(\mu)$ are independent, what is the distribution of $X + Y$?

Solution 1.14. We know that $L_X(t) = e^{-\lambda(e^t-1)}$ and $L_Y(t) = e^{-\mu(e^t-1)}$. Since X and Y are independent, the Laplace transform of $X + Y$ is

$$L_{X+Y}(t) = L_X(t)L_Y(t) = e^{-\lambda(e^t-1}e^{-\mu(e^t-1)} = e^{-(\lambda+\mu)(e^t-1)}$$

which we recognize as the Laplace transform of a $\text{Poi}(\lambda + \mu)$ random variables. By uniqueness of the Laplace transform, we conclude $X + Y \sim \text{Poi}(\lambda + \mu)$.

Problem 1.15. Let X be a random variable takes values 0 and 1 with equal probability,

$$\mathbb{P}(X=0)=\mathbb{P}(X=1)=\frac{1}{2}.$$

Let Y be a random variable that is independent of X and takes values -1 and 1 with equal probability,

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

We let $Z=XY$. Show that Z and X are uncorrelated, but not independent.

Solution 1.15. First, note that

$$\mathbb{E}[Z]=\mathbb{E}[XY]=\mathbb{E}[X]\cdot\mathbb{E}[Y]=\frac{1}{2}\cdot 0=0.$$

Therefore,

$$\operatorname{Cov}(X,Z)=\mathbb{E}[XZ]-\mathbb{E}[X]\,\mathbb{E}[Z]=\mathbb{E}[X^2]\mathbb{E}[Y]=0.$$

On the other hand, we have $Z^2=X$ and hence, for $g_1(t)=t$ and $g_2(t)=t^2$,

$$\mathbb{E}[g_1(X)\cdot g_2(Z)]=\mathbb{E}[X^2]=\frac{1}{2}\neq\frac{1}{4}=\mathbb{E}[g_1(X)]\cdot\mathbb{E}[g_2(Z)].$$

Therefore, by Proposition 1, X, Z are not independent.
(Equivalently, since $g_1(X)=X$ and $g_2(Z)=Z^2=X$ are not independent, we conclude that X and Z cannot be independent by Proposition 2.)

Problem 1.16. A course is evaluated with a midterm exam and a final exam. The midterm has weight α and the final has weight $1-\alpha$, where $\alpha\in[0,1]$. Assuming the percentage scores of a given student in both exams are independent and identically distributed $[0,100]$-valued random variables, is it preferable for the student to write both exams, or should they instead get VIFs and transfer the weight of the midterm to the final?

Solution 1.16. Let M denote the result of the midterm and F the result of the final. Both are independent random variables with identical distribution on $[0,100]$. The final grade is either $F_\alpha:=\alpha M+(1-\alpha)F$ or $F_0:=F$, depending on whether or not the midterm exam is taken. Clearly,

$$\mathbb{E}[F_\alpha]=\alpha\mathbb{E}[M]+(1-\alpha)\mathbb{E}[F]=\mathbb{E}[F]=\mathbb{E}[F_0].$$

So the expectations in both cases are the same. Next,

$$\begin{aligned}\operatorname{Var}(F_\alpha)&=\operatorname{Var}(\alpha M+(1-\alpha)F)=\operatorname{Var}(\alpha M)+\operatorname{Var}((1-\alpha)F)\\&=\big(\alpha^2+(1-\alpha)^2\big)\operatorname{Var}(F_0)=\big(1+2\alpha^2-2\alpha\big)\operatorname{Var}(F_0)\end{aligned}$$

If $0<\alpha<1$, the right-hand side is strictly smaller than $\operatorname{Var}(F_0)$. So from that point of view it is better to write the midterm exam.