Lecture 2 — Probability review

Recall from Lecture 1 our "fundamental principle": Data is a realization of a random process. Throughout this course, we will model the data using **random variables**. The goal of this lecture is to review, with a statistical focus, relevant concepts concerning random variables and their distributions.

2.1 Random variables

A **discrete** random variable X can take a finite or countably infinite number of possible values. We use discrete random variables to model categorical data (for example, which presidential candidate a voter supports) and count data (for example, how many cups of coffee a graduate student drinks in a day). The distribution of X is specified by its **probability mass function (PMF)**:

$$f_X(x) = \mathbb{P}[X = x].$$

Then for any set A of values that X can take,

$$\mathbb{P}[X \in A] = \sum_{x \in A} f_X(x).$$

A **continuous** random variable X takes values in $\mathbb{R}$ and models continuous real-valued data (for example, the height of a person). For any single value $x \in \mathbb{R}$, $\mathbb{P}[X = x] = 0$. Instead, the distribution of X is specified by its **probability density function (PDF)** $f_X(x)$, which satisfies for any set $A \subseteq \mathbb{R}$

$$\mathbb{P}[X \in A] = \int_A f_X(x) dx.$$

In both cases, when it is clear which random variable is being referred to, we will simply write $f(x)$ for $f_X(x)$.

Example 2.1. A **Bernoulli** random variable $X \sim \text{Bernoulli}(p)$ (for $p \in [0, 1]$) is discrete and takes two possible values $\{0, 1\}$. Its PMF is given by

$$f(x) = \begin{cases} p & x = 1 \\ 1 - p & x = 0. \end{cases}$$

Example 2.2. A **Binomial** random variable $X \sim \text{Binomial}(n, p)$ (for a positive integer n and $p \in [0, 1]$) is discrete and takes values in $\{0, 1, 2, \ldots, n\}$. Its PMF is given by (for $x = 0, 1, 2, \ldots, n$)

$$f(x) = \binom{n}{x} p^x (1 - p)^{n-x}.$$

Example 2.3. A **Gamma** random variable $X \sim \text{Gamma}(\alpha, \beta)$ (for $\alpha, \beta > 0$) is continuous and takes positive values. Its PDF is given by

$$f(x) = \begin{cases} \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} & x > 0 \\ 0 & x \le 0. \end{cases}$$

In the above, $\Gamma(\alpha)$ is called the Gamma function, defined by

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x} dx.$$

You can think of the Gamma function as extending the factorial function to all positive real numbers. (For positive integers n, $\Gamma(n) = (n-1)!$.)

Example 2.4. A **Normal** random variable $X \sim \mathcal{N}(\mu, \sigma^2)$ (for $\mu \in \mathbb{R}$ and $\sigma^2 > 0$) is continuous and can take any real value. Its PDF is given by

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

For any random variable X and real-valued function g, the **expectation** or mean of $g(X)$ is its "average value". If X is discrete with PMF $f_X(x)$, then

$$\mathbb{E}[g(X)] = \sum_x g(x) f_X(x)$$

where the sum is over all possible values of X. If X is continuous with PDF $f_X(x)$, then

$$\mathbb{E}[g(X)] = \int_{\mathbb{R}} g(x) f_X(x) dx.$$

The expectation is *linear*: For any random variables $X_1, \ldots, X_n$ (not necessarily independent) and any $c \in \mathbb{R}$,

$$\mathbb{E}[X_1 + \ldots + X_n] = \mathbb{E}[X_1] + \ldots + \mathbb{E}[X_n], \quad \mathbb{E}[cX] = c\,\mathbb{E}[X].$$

If $X_1, \ldots, X_n$ are independent, then

$$\mathbb{E}[X_1 \ldots X_n] = \mathbb{E}[X_1] \ldots \mathbb{E}[X_n].$$

The **variance** of X is defined by the two equivalent expressions

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

For any $c \in \mathbb{R}$, $\text{Var}[cX] = c^2 \,\text{Var}[X]$. If $X_1, \ldots, X_n$ are independent, then

$$\text{Var}[X_1 + \ldots + X_n] = \text{Var}[X_1] + \ldots + \text{Var}[X_n].$$

If $X_1, \ldots, X_n$ are not independent, then this is not true—$\text{Var}[X_1 + \ldots + X_n]$ will depend on the covariance between each pair of variables. The **standard deviation** of X is $\sqrt{\text{Var}[X]}$.

The distribution of X can also be specified by its **cumulative distribution function (CDF)** $F_X(x) = \mathbb{P}[X \leq x]$. In the discrete and continuous cases, respectively, this is given by

$$F_X(x) = \sum_{y:\, y \leq x} f_X(y), \qquad F_X(x) = \int_{-\infty}^{x} f_X(y)dy.$$

In the continuous case, the fundamental theorem of calculus implies

$$f_X(x) = \frac{d}{dx} F_X(x).$$

By definition, F_X is monotonically increasing: $F_X(x) \leq F_X(y)$ if $x < y$. If F_X is continuous and strictly increasing, meaning $F_X(x) < F_X(y)$ for all $x < y$, then F_X has an inverse function $F_X^{-1} : (0,1) \to \mathbb{R}$ called the **quantile function**: For any $t \in (0,1)$, $F_X^{-1}(t)$ is the t^{th} quantile of the distribution of X. I.e. the probability that X is less than this value is exactly t.

2.2 Moment generating functions

A tool that will be particularly useful for us is the **moment generating function (MGF)** of a random variable X. This is a function of a single argument $t \in \mathbb{R}$, defined as

$$M_X(t) = \mathbb{E}[e^{tX}].$$

Depending on the random variable X, $M_X(t)$ might be infinite for some values of t. Here are two examples:

Example 2.5 (Normal MGF). Suppose $X \sim \mathcal{N}(0,1)$. Then

$$M_X(t) = \mathbb{E}[e^{tX}] = \int e^{tx} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx = \int \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2+2tx}{2}} dx.$$

To compute this integral, we complete the square:

$$\int \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2+2tx}{2}} dx = \int \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2+2tx-t^2}{2} + \frac{t^2}{2}} dx = e^{\frac{t^2}{2}} \int \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-t)^2}{2}} dx.$$

The quantity inside the last integral above is the PDF of the $\mathcal{N}(t,1)$ distribution—hence it must integrate to 1. Then $M_X(t) = e^{t^2/2}$.

Now suppose $X \sim \mathcal{N}(\mu, \sigma^2)$. Then $X = \mu + \sigma Z$, where $Z \sim \mathcal{N}(0,1)$. So

$$M_X(t) = \mathbb{E}[e^{tX}] = \mathbb{E}[e^{\mu t + \sigma t Z}] = e^{\mu t}\mathbb{E}[e^{\sigma t Z}] = e^{\mu t} M_Z(\sigma t) = e^{\mu t + \frac{\sigma^2 t^2}{2}}.$$

For a normal random variable X, $M_X(t)$ is finite for all $t \in \mathbb{R}$.

Example 2.6 (Gamma MGF). Suppose $X \sim \text{Gamma}(\alpha, \beta)$, for $\alpha, \beta > 0$. Then

$$M_X(t) = \mathbb{E}[e^{tX}] = \int_0^\infty e^{tx} \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} dx = \int_0^\infty \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{(t-\beta)x} dx.$$

If $t > \beta$, then $\lim_{x\to\infty} x^{\alpha-1}e^{(t-\beta)x} = \infty$, so certainly the integral above is infinite. If $t = \beta$, note that $\int_0^\infty x^{\alpha-1}dx = \frac{1}{\alpha}x^\alpha\Big|_0^\infty = \infty$, since $\alpha > 0$. Hence $M_X(t) = \infty$ for any $t \geq \beta$. For $t < \beta$, let us rewrite the above to isolate the PDF of the Gamma$(\alpha, \beta - t)$ distribution:

$$M_X(t) = \frac{\beta^\alpha}{(\beta-t)^\alpha}\int_0^\infty \frac{(\beta-t)^\alpha}{\Gamma(\alpha)}x^{\alpha-1}e^{-(\beta-t)x}dx.$$

As the PDF of the Gamma$(\alpha, \beta - t)$ distribution integrates to 1, we obtain finally

$$\begin{aligned} M_X(t) &= \begin{cases} \infty & t \geq \beta \\ \frac{\beta^\alpha}{(\beta-t)^\alpha} & t < \beta \end{cases} \\ &= \begin{cases} \infty & t \geq \beta \\ (1-\beta^{-1}t)^{-\alpha} & t < \beta. \end{cases} \end{aligned}$$

If the MGF of a random variable X is finite in any interval that contains 0 as an interior point, as in the above two examples, then (like the PDF or CDF) it also completely specifies the distribution of X. This is the content of the following theorem (which we will not prove in this class):

Theorem 2.7. *Let X and Y be two random variables such that, for some $h > 0$ and every $t \in (-h, h)$, both $M_X(t)$ and $M_Y(t)$ are finite and $M_X(t) = M_Y(t)$. Then X and Y have the same distribution.*

The reason why the MGF will be useful for us is because if $X_1, \ldots, X_n$ are independent, then the MGF of their sum satisfies

$$M_{X_1+\ldots+X_n}(t) = \mathbb{E}[e^{t(X_1+\ldots+X_n)}] = \mathbb{E}[e^{tX_1}] \times \ldots \times \mathbb{E}[e^{tX_n}] = M_{X_1}(t)\ldots M_{X_n}(t).$$

This gives us a very simple tool to understand the distributions of sums of independent random variables.