Modern Statistics

Xiangyu Chang

March 17, 2026

Abstract

To be undated.

1 Lecture 5: Random Vector and Expectation

In Lecture 3, we introduced random vectors (X, Y) and their joint distributions. A central question in statistics is: how does knowing one variable affect our beliefs about another? This lecture answers that question through **conditional distributions**. We then extend the framework to d-dimensional random vectors, introduce the multivariate normal and multinomial distributions, and finally define **expectation**—the first of many summary quantities that will play a fundamental role in estimation and inference in later lectures.

1.1 Recall: Joint and Marginal Distributions

We briefly recall the key concepts from Lecture 3. For a random vector (X, Y):

- **Joint CDF**: $F_{X,Y}(x, y) = P(X \leqslant x, Y \leqslant y)$.
- **Joint PMF** (discrete): $f_{X,Y}(x, y) = P(X = x, Y = y)$.
- **Joint PDF** (continuous): $f_{X,Y}(x, y) \geqslant 0$, $\int_{\mathbb{R}^2} f_{X,Y}(x, y)\ \mathrm{d}x\ \mathrm{d}y = 1$, and $P((X, Y) \in A) = \iint_A f_{X,Y}(x, y)\ \mathrm{d}x\ \mathrm{d}y$.
- **Marginal distribution**: $f_X(x) = \sum_y f_{X,Y}(x, y)$ (discrete) or $f_X(x) = \int_{\mathbb{R}} f_{X,Y}(x, y)\ \mathrm{d}y$ (continuous).
- **Independence**: X and Y are independent if $P(X \in A, Y \in B) = P(X \in A) \cdot P(Y \in B)$ for all Borel sets A, B, equivalently $f_{X,Y}(x, y) = f_X(x) f_Y(y)$.

With this foundation, we now study how to update our beliefs about Y when we observe X.

1.2 Conditional PMF and PDF

The **conditional distribution** of Y given $X = x$ describes how the distribution of Y changes when we know the value of X. It is the probabilistic analogue of "Y given X."

Definition 1.1 (Conditional PMF and PDF)**.** For random variables X and Y with joint PMF or PDF $f_{X,Y}$ and marginal $f_X(x) > 0$:

- **Discrete**: $P(Y = y \mid X = x) = \frac{P(X=x, Y=y)}{P(X=x)} = \frac{f_{X,Y}(x,y)}{f_X(x)}$.
- **Continuous**: $f_{Y|X}(y \mid x) = \frac{f_{X,Y}(x,y)}{f_X(x)}$.

In both cases, the **chain rule** holds:

$$f_{X,Y}(x, y) = f_{Y|X}(y \mid x) \cdot f_X(x).$$

When X and Y are independent, we have $f_{X,Y}(x, y) = f_X(x) f_Y(y)$. Substituting into the conditional formula yields

$$f_{Y|X}(y \mid x) = \frac{f_{X,Y}(x, y)}{f_X(x)} = \frac{f_X(x) f_Y(y)}{f_X(x)} = f_Y(y).$$

Thus, knowing X does not change the distribution of Y.

Example 1.2 (Conditional Uniform)**.** Let $X \sim U[0, 1]$ and suppose that given $X = x$, we have $Y \mid X = x \sim U[x, 1]$. Find the marginal PDF of Y.

Step 1: The marginal of X and the conditional of Y given X are

$$f_X(x) = \begin{cases} 1, & 0 \leqslant x \leqslant 1 \\ 0, & \text{otherwise} \end{cases} \quad \text{and} \quad f_{Y|X}(y \mid x) = \begin{cases} \frac{1}{1-x}, & x \leqslant y < 1 \\ 0, & \text{otherwise} \end{cases}.$$

Step 2: By the chain rule, the joint PDF is

$$f_{X,Y}(x, y) = f_{Y|X}(y \mid x) \cdot f_X(x) = \begin{cases} \frac{1}{1-x}, & 0 \leqslant x \leqslant y < 1 \\ 0, & \text{otherwise} \end{cases}.$$

Step 3: Marginalizing over x, for $y \in [0, 1]$:

$$f_Y(y) = \int_{\mathbb{R}} f_{X,Y}(x, y)\, \mathrm{d}x = \int_0^y \frac{1}{1-x}\, \mathrm{d}x = -\ln(1 - y).$$

Remark 1.3 (Linear Model)**.** *A simple regression model posits $Y_i = \beta^\top x_i + \varepsilon_i$ for observations $\{(x_i, y_i)\}_{i=1}^n$, where ε_i are random errors. The conditional distribution $Y \mid X = x$ is central to this framework; we will return to it in the regression lectures.*

1.3 Multivariate Random Vectors

We now extend the bivariate framework to d-dimensional random vectors $\boldsymbol{X} = (X_1, \ldots, X_d)^\top \in \mathbb{R}^d$.

Definition 1.4 (Multivariate CDF and PDF)**.** For $\boldsymbol{x} = (x_1, \ldots, x_d)^\top \in \mathbb{R}^d$:

- **Joint CDF**: $F_{\boldsymbol{X}}(\boldsymbol{x}) = P(X_1 \leqslant x_1, \ldots, X_d \leqslant x_d)$.
- **Joint PDF**: $f_{\boldsymbol{X}}(\boldsymbol{x}) \geqslant 0$, $\int_{\mathbb{R}^d} f_{\boldsymbol{X}}(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x} = 1$, and $P(\boldsymbol{X} \in A) = \int_A f_{\boldsymbol{X}}(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x}$ for Borel sets $A \subseteq \mathbb{R}^d$.

Definition 1.5 (Marginal Distribution)**.** The marginal PDF of $(X_1, \ldots, X_k)$ is obtained by integrating out the remaining variables:

$$f_{X_1,\ldots,X_k}(x_1, \ldots, x_k) = \int_{\mathbb{R}} \cdots \int_{\mathbb{R}} f_{\boldsymbol{X}}(\boldsymbol{x})\ \mathrm{d}x_{k+1} \cdots \mathrm{d}x_d.$$

Definition 1.6 (Independence)**.** Random variables $X_1, \ldots, X_d$ are **independent** if

$$f_{\boldsymbol{X}}(\boldsymbol{x}) = \prod_{i=1}^{d} f_{X_i}(x_i).$$

1.3.1 Multivariate Normal Distribution

The multivariate normal distribution generalizes the univariate normal to $\mathbb{R}^d$.

Definition 1.7 (Standard Multivariate Normal)**.** Let $Z = (Z_1, \ldots, Z_k)^\top$ where $Z_1, \ldots, Z_k \sim N(0,1)$ are independent. The density of Z is

$$f(\boldsymbol{z}) = \prod_{i=1}^{k} f_{Z_i}(z_i) = \frac{1}{(2\pi)^{k/2}} \exp\left\{-\frac{1}{2}\sum_{j=1}^{k} z_j^2\right\} = \frac{1}{(2\pi)^{k/2}} \exp\left\{-\frac{1}{2}\boldsymbol{z}^\top \boldsymbol{z}\right\}.$$

We write $Z \sim N(\boldsymbol{0}, I)$, where I is the $k \times k$ identity matrix.

Definition 1.8 (General Multivariate Normal)**.** A random vector $\boldsymbol{X}$ has a **multivariate normal distribution** $\boldsymbol{X} \sim N(\boldsymbol{\mu}, \Sigma)$ if its density is

$$f(\boldsymbol{x}; \boldsymbol{\mu}, \Sigma) = \frac{1}{(2\pi)^{k/2}|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})^\top \Sigma^{-1} (\boldsymbol{x} - \boldsymbol{\mu})\right\},$$

where $\boldsymbol{\mu} \in \mathbb{R}^k$ is the mean vector, Σ is a $k \times k$ symmetric positive definite covariance matrix, and $|\Sigma|$ denotes its determinant.

Lemma 1.9. *If $\boldsymbol{X} \sim N(\boldsymbol{0}, I)$, then $\boldsymbol{Z} = \boldsymbol{\mu} + \Sigma^{1/2}\boldsymbol{X} \sim N(\boldsymbol{\mu}, \Sigma)$.*

Proof. The transformation $\boldsymbol{g}^{-1}(\boldsymbol{z}) = \Sigma^{-1/2}(\boldsymbol{z} - \boldsymbol{\mu})$ has Jacobian $\nabla \boldsymbol{g}^{-1}(\boldsymbol{z}) = \Sigma^{-1/2}$. By the multivariate change-of-variable formula,

$$f_{\boldsymbol{Z}}(\boldsymbol{z}) = f_{\boldsymbol{X}}(\boldsymbol{g}^{-1}(\boldsymbol{z}))\left|\det \Sigma^{-1/2}\right| = \frac{1}{(2\pi)^{k/2}|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(\boldsymbol{z} - \boldsymbol{\mu})^\top \Sigma^{-1} (\boldsymbol{z} - \boldsymbol{\mu})\right\}.$$

■

Definition 1.10 (Matrix Square Root)**.** The **square root** $\Sigma^{1/2}$ of a symmetric positive definite matrix Σ satisfies:

- $\Sigma^{1/2}$ is symmetric.
- $\Sigma = \Sigma^{1/2}\Sigma^{1/2}$.
- $\Sigma^{1/2}\Sigma^{-1/2} = \Sigma^{-1/2}\Sigma^{1/2} = I$, where $\Sigma^{-1/2} = (\Sigma^{1/2})^{-1}$.
- If $\Sigma = UDU^\top$ is an eigendecomposition, then $\Sigma^{1/2} = UD^{1/2}U^\top$ and $\Sigma^{-1/2} = UD^{-1/2}U^\top$, where $D^{1/2} = \operatorname{diag}(\sqrt{D_{11}}, \ldots, \sqrt{D_{kk}})$.

1.3.2 Multinomial Distribution

Consider tossing a die with k faces n times. Let $p = (p_1, \ldots, p_k)$ with $p_j \geqslant 0$ and $\sum_{j=1}^k p_j = 1$, and let $X = (X_1, \ldots, X_k)^\top$ where X_j counts how many times face j appears, so $\sum_{j=1}^k X_j = n$.

Definition 1.11 (Multinomial Distribution)**.** We say $X \sim \text{Multinomial}(n, p)$ if

$$P(X_1 = n_1, X_2 = n_2, \ldots, X_k = n_k) = \frac{n!}{n_1! n_2! \cdots n_k!} p_1^{n_1} p_2^{n_2} \cdots p_k^{n_k},$$

where $n_1 + \cdots + n_k = n$.

Lemma 1.12. *If $X \sim \text{Multinomial}(n, p)$ with $X = (X_1, \ldots, X_k)^\top$, then the marginal distribution of X_j is $X_j \sim \text{Bin}(n, p_j)$.*

1.4 Expectation

We now introduce the **expected value** (or **mean**) of a random variable—a single number that summarizes the "center" of its distribution. Expectation will be the cornerstone of estimation, variance, and many limit theorems in later lectures.

Definition 1.13 (Expected Value)**.** The **expected value** (or **mean**, or **first moment**) of a random variable X is

$$\mathbb{E}[X] = \int x \, \mathrm{d}F_X(x) = \begin{cases} \sum_x x \, f_X(x) & \text{if } X \text{ is discrete} \\ \int_{\mathbb{R}} x \, f_X(x) \, \mathrm{d}x & \text{if } X \text{ is continuous} \end{cases},$$

provided the sum or integral exists (i.e., $\mathbb{E}[|X|] < \infty$).

Example 1.14 (Coin Toss)**.** Flip a fair coin twice. Let X be the number of heads. Then $X \in \{0, 1, 2\}$ with $f_X(0) = 1/4$, $f_X(1) = 1/2$, $f_X(2) = 1/4$. Thus

$$\mathbb{E}[X] = \sum_x x \, f_X(x) = 0 \cdot \frac{1}{4} + 1 \cdot \frac{1}{2} + 2 \cdot \frac{1}{4} = 1.$$

Definition 1.15 (Rule of the Lazy Statistician)**.** Let $Y = r(X)$ for a (measurable) function r. Then

$$\mathbb{E}[Y] = \mathbb{E}[r(X)] = \int r(x) \, \mathrm{d}F_X(x) = \int r(x) \, f_X(x) \, \mathrm{d}x \quad \text{(continuous case)}.$$

We do *not* need to find the distribution of Y; we integrate with respect to the distribution of X.

Example 1.16 (Indicator and Probability)**.** Let I_A be the indicator of event A: $I_A(x) = 1$ if $x \in A$ and $I_A(x) = 0$ otherwise. Then

$$\mathbb{E}[I_A(X)] = 0 \cdot P(X \notin A) + 1 \cdot P(X \in A) = P(X \in A).$$

Thus expectation of an indicator recovers the probability of the corresponding event.

1.5 The k-th Moment

We now extend the definition of expectation to higher powers of X. The k-th moment captures progressively finer shape information: the first moment is the mean, the second underlies variance, the third governs skewness, and so on.

Definition 1.17 (k-th Moment)**.** The **k-th moment** of X is $\mathbb{E}[X^k]$, provided $\mathbb{E}[|X|^k] < \infty$.

A fundamental hierarchy holds: the finiteness of a higher moment implies that of all lower moments.

Theorem 1.18 (Moment Hierarchy)**.** *If $\mathbb{E}[|X|^k] < \infty$ for some $k \geqslant 1$, then $\mathbb{E}[|X|^i] < \infty$ for all $1 \leqslant i \leqslant k$.*

Proof. Split the integral at $|x| = 1$:

$$\begin{aligned}
\mathbb{E}[|X|^i] &= \int_{\mathbb{R}} |x|^i \, \mathrm{d}F_X(x) \\
&= \int_{|x|>1} |x|^i \, \mathrm{d}F_X(x) + \int_{|x|\leqslant 1} |x|^i \, \mathrm{d}F_X(x) \\
&\leqslant \int_{|x|>1} |x|^i \, \mathrm{d}F_X(x) + 1 \\
&\leqslant \int_{|x|>1} |x|^k \, \mathrm{d}F_X(x) + 1 \quad (\text{since } |x| > 1 \text{ and } i \leqslant k \Rightarrow |x|^i \leqslant |x|^k) \\
&\leqslant \int_{\mathbb{R}} |x|^k \, \mathrm{d}F_X(x) + 1 \\
&= \mathbb{E}[|X|^k] + 1 < \infty.
\end{aligned}$$

■

1.6 Properties of Expectation

The following three properties make expectation tractable in practice.

1. **Linearity.** For constants $a_1, \ldots, a_n \in \mathbb{R}$ and integrable random variables $X_1, \ldots, X_n$:

$$\mathbb{E}\left[\sum_{i=1}^{n} a_i X_i\right] = \sum_{i=1}^{n} a_i \, \mathbb{E}[X_i].$$

2. **Multiplicativity under independence.** If $X_1, \ldots, X_k$ are mutually independent:

$$\mathbb{E}\left[\prod_{i=1}^{k} X_i\right] = \prod_{i=1}^{k} \mathbb{E}[X_i].$$

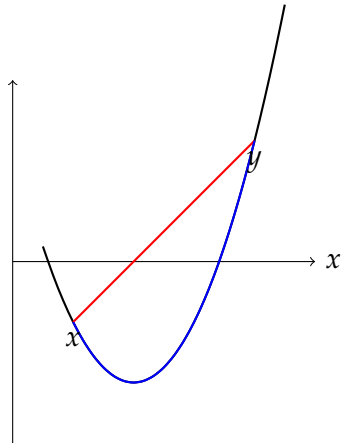
3. **Jensen's Inequality.** A function $g\colon \mathbb{R} \to \mathbb{R}$ is **convex** if for all $0 \leqslant \lambda \leqslant 1$:

$$g(\lambda x + (1-\lambda)y) \leqslant \lambda g(x) + (1-\lambda)g(y).$$

For any convex g:

$$\mathbb{E}[g(X)] \geqslant g(\mathbb{E}[X]).$$

Geometrically, the chord between two points on the graph of g lies above the arc, as illustrated below.



The red chord $\lambda g(x) + (1-\lambda)g(y)$ lies above the blue arc $g(\lambda x + (1-\lambda)y)$.

Example 1.19 (Expectation of Binomial(n,p))**.** Let $X \sim \text{Binomial}(n,p)$, so $P(X=k) = \binom{n}{k}p^k(1-p)^{n-k}$. We compute $\mathbb{E}[X]$ in two ways.

Method 1 (Direct calculation):

$$\begin{aligned}
\mathbb{E}[X] &= \sum_{k=0}^{n} k \cdot P(X=k) \\
&= \sum_{k=0}^{n} k \cdot \binom{n}{k} p^k (1-p)^{n-k} \\
&= \sum_{k=1}^{n} n \cdot \binom{n-1}{k-1} p^k (1-p)^{n-k} \\
&= np \sum_{k=1}^{n} \binom{n-1}{k-1} p^{k-1} (1-p)^{n-k} \\
&= np \sum_{j=0}^{n-1} \binom{n-1}{j} p^{j} (1-p)^{n-1-j} \\
&= np.
\end{aligned}$$

Method 2 (Decomposition): Write $X = \sum_{i=1}^{n} X_i$ where $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \text{Bernoulli}(p)$. Since $\mathbb{E}[X_i] = p$, linearity of expectation gives

$$\mathbb{E}[X] = \mathbb{E}\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} \mathbb{E}[X_i] = np.$$

References