II. Preliminary Knowledge: Statistics

Young-geun Kim

Department of Statistics and Probability

STT 997 (SS 2025)

Motivating Example

- *Juryeonggu* (酒令具; AD668 $\sim$ AD935) is an old dice used for drinking games during a historical dynasty in what is now South Korea.
- It has 14 faces, six square and eight hexagonal, with each face assigning a penalty, e.g., 三盞一去: Drink three cups at once.

Left: A restored copy at Gyeongju National Museum. Right: Lettering on each side at the time of excavation.

Motivating Example

- **Q:** This dice has two different types of faces. What outcomes can we expect when we roll it?

Left: A restored copy at Gyeongju National Museum. Right: Lettering on each side at the time of excavation.

Outline

1. Basic Terminology
2. Distribution of Multivariate Random Variable
3. Density Model
4. Likelihood Maximization Principle

Probability

- **Outcome:** Possible results of experiments or trials, e.g., H when we toss a coin.
- **Event:** A set of outcomes, e.g., $\{$H, T$\}$ representing the set of all possible outcomes.
- **Probability ($\mathbb{P}$):** A function that maps events to real numbers, e.g., $\mathbb{P}(\{\text{H}\}) = 1/2$ and $\mathbb{P}(\{\text{H, T}\}) = 1$.

Probability

- Let Ω, $\mathcal{F}$, and $\mathbb{P}$, respectively, be the set of all outcomes, the set of all events, and the probability function. The triple $(\Omega, \mathcal{F}, \mathbb{P})$ is called *probability space*. For example, when we consider tossing 2 coins:
  $\Omega = \{\text{HH, HT, TH, TT}\}$
  $\mathcal{F} = \{\phi, \{\text{HH}\}, \ldots, \{\text{HH, HT, TH, TT}\}\}$
  $\mathbb{P}(\phi) = 0$, $\mathbb{P}(\{\text{HH}\}) = 1/2$, $\ldots$, $\mathbb{P}(\Omega = \{\text{HH, HT, TH, TT}\}) = 1$.
- **Probability Axioms:** $\mathbb{P}$ should hold the following three axioms:
  1. For any event $E \in \mathcal{F}$, $\mathbb{P}(E) \geq 0$
  2. $\mathbb{P}(\Omega) = 1$
  3. For any countable sequence of disjoint events $(E_i)_{i=1}^{\infty}$, $\mathbb{P}(\cup_{i=1}^{\infty} E_i) = \sum_{i=1}^{\infty} \mathbb{P}(E_i)$.
- There are diverse approaches to describing what probabilities mean. For example, from a frequentist perspective, the probability of an event is described as the limit[1] of the relative frequencies of the event.

[1] The limit point will be formally defined in a later slide titled ‘Convergence of Random Variable’.

Conditional Probability

- For any events A and B such that $\mathbb{P}(A) > 0$, we denote the conditional probability of event B given A by $\mathbb{P}(B|A) := \mathbb{P}(A \cap B)/\mathbb{P}(A)$. That is, $\mathbb{P}(B|A)$ represents the possibility of B occuring when A is treated as the set of all outcomes.
- **Example:** Consider a population of $10,000$ people, consisting of 40 with the flu and $9,960$ without the flu. A flu diagnostic kit is used, which has a 90% probability of correctly diagnosing the flu in those who have it, and a 95% of correctly identifying those who are healthy. Given this, determine:
  1. The probability that a person diagnosed with the flu actually has the flu.
  2. The probability that a person diagnosed without the flu actually does not have the flu.

Bayes' Theorem

- Bayes' theorem (or Bayes' law or Bayes' rule) is a useful formula for inverting conditional probabilities, expressed as ($\mathbb{P}(A) > 0$ and $\mathbb{P}(B) > 0$):

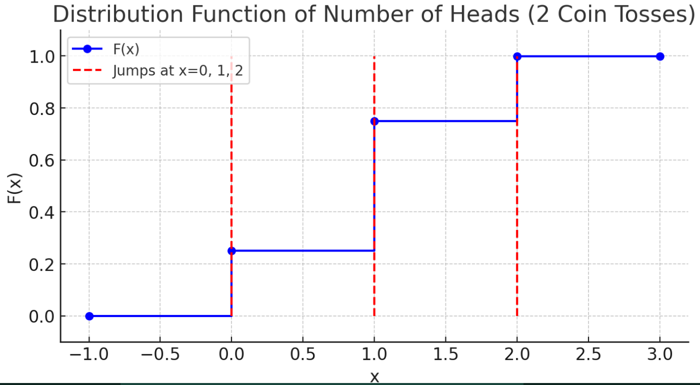
$$\mathbb{P}(A|B) = \mathbb{P}(B|A)\mathbb{P}(A)/\mathbb{P}(B). \tag{1}$$

- In the flu example, by Bayes' theorem, $\mathbb{P}(\text{Actually has the flu}|\text{Diagnosed with the flu})$ can be expressed as:

$$\begin{aligned} &\frac{\mathbb{P}(\text{Diagnosed with the flu}|\text{Actually has the flu})\mathbb{P}(\text{Actually has the flu})}{\mathbb{P}(\text{Diagnosed with the flu})} \\ &= \frac{90\% \times 0.4\%}{90\% \times 0.4\% + 5\% \times 99.6\%} \approx 6.7\%. \end{aligned} \tag{2}$$

Random Variable, Distribution Function, and Moment

- In this course, X is called a *random variable* when it maps outcomes to a real numbers, and $\vec{X} := (X_1, \ldots, X_p)^T$ is called a *random vector* if it maps outcomes to real vectors.
- The *distribution function* (or *cumulative distribution function*) of X is defined as $F(x) := \mathbb{P}(X \leq x)$, and for $\vec{X}$, it is defined as $F(\vec{x}) := \mathbb{P}(X_1 \leq x_1, \ldots, X_p \leq x_p)$.



Random Variable, Distribution Function, and Moment

- The probability of X being in the interval $(a, b]$ can be expressed as $\mathbb{P}(a < X \leq b) = F(x = b) - F(x = a)$, and being at a specific value a can be expressed as $\mathbb{P}(X = a) = F(x = a) - \lim_{x \to a-} F(x)$.
- When X is a continuous random variable, the *probability density function* is defined as $p(x) := dF(x)/dx$. When it is discrete, the *probability mass function* is defined as $p(x) := \mathbb{P}(X = x)$.[2]
- For example, X is said to follow the standard Gaussian distribution when $F(x) = \int_{-\infty}^{x} p(x = x')dx'$ where $p(x) = (\sqrt{2\pi})^{-1} \exp(-x^2/2)$.

[2]Both probability density and mass functions are Radon–Nikodym derivatives with respect to (w.r.t.) the Lebesgue measure and counting measure, respectively.

Random Variable, Distribution Function, and Moment

- The expectation over the distribution of X is denoted by $\mathbb{E}_X$, and the s-th moment of $g(X)$ for any function g is denoted by $\mathbb{E}_X g(X)^s$.
- The $\mathbb{E}_X g(X)^s = \int g(x)^s p(x)\, dx$ for continuous variables, and $\mathbb{E}_X g(X)^s = \sum_x g(x)^s p(x)$ for discrete variables. For example, the population mean is the first moment of X.
- The population variance is denoted by $\mathrm{Var}_X(g(X)) := \mathbb{E}_X\Big(g(X) - \mathbb{E}_X(g(X))\Big)^2 = \mathbb{E}_X\Big(g(X)^2\Big) - \Big(\mathbb{E}_X(g(X))\Big)^2$.

Some Useful Inequalities about Moments

- **Jensen's Inequality:** For any random variable X having a finite first moment, and any real-valued convex function ϕ defined on the support of X,

$$\phi(\mathbb{E}_X(X)) \leq \mathbb{E}_X(\phi(X)). \tag{3}$$

  Hint: Since ϕ is convex, it has a subderivative at $\mathbb{E}(X)$, i.e., there exists a real number a such that $\phi(X) \geq a(X - \mathbb{E}_X(X)) + \phi(\mathbb{E}_X(X))$.

- **Liapounov's Inequality:** For any random variable X having a finite s-th moment and any positive real number $r < s$,

$$\left(\mathbb{E}_X(|X|^r)\right)^{1/r} \leq \left(\mathbb{E}_X(|X|^s)\right)^{1/s} \tag{4}$$

  Hint: Jensen's inequality with $\phi(u) = u^{s/r}$ for $u \geq 0$.

Some Useful Inequalities about Moments

- **Markov's Inequality:** For any random variable X having a finite r-th moment and any positive real number k,

$$\mathbb{P}(|X| \geq k) \leq \mathbb{E}_X|X|^r/k^r. \tag{5}$$

Hint: $I(|X| \geq k) \leq (|X|/k)^r I(|X| \geq k) \leq (|X|/k)^r$.

- **Chebyshev's Inequality:** For any random variable X having a finite second moment and any positive real number k,

$$\mathbb{P}(|X - \mathbb{E}_X(X)| \geq k) \leq \mathrm{Var}_X(X)/k^2 \tag{6}$$

Hint: Markov's inequality with $X - \mathbb{E}_X(X)$ and $r = 2$.

Outline

Joint Distribution

- Generative models aim to learn the population of data having multivariate features. In this subsection, we formulate the target of generative models, distributions of multivariate random variables (or random vectors).
- let $\{w|\vec{X}(w) \leq \vec{x}\} := \{w|X_1(w) \leq x_1, \ldots, X_p(w) \leq x_p\}$. For any two random vectors $\vec{X} \in \mathbb{R}^p$ and $\vec{Y} \in \mathbb{R}^q$,
  Joint Distribution Function: $F(\vec{x}, \vec{y}) := \mathbb{P}(\vec{X} \leq \vec{x}, \vec{Y} \leq \vec{y})$
- When $\vec{X}$ and $\vec{Y}$ are continuous random vectors, their *joint density function* is defined as $p(\vec{x}, \vec{y}) := \partial^{p+q} F(\vec{x}, \vec{y}) / \partial x_1 \cdots \partial x_p \partial y_1 \cdots \partial y_q$. When they are discrete, their *joint probability mass function* is defined as $p(\vec{x}, \vec{y}) := \mathbb{P}(\vec{X} = \vec{x}, \vec{Y} = \vec{y})$.[3]

[3] The $p(\vec{x}, \vec{y})$ is defined as the Radon-Nikodym derivative in general.

Joint Distribution

- The mean vector of $\vec{X}$ is defined as $\mathbb{E}_{\vec{X}}(\vec{X})$.
- The covariance matrix between $\vec{X}$ and $\vec{Y}$ is defined as $\text{Cov}_{(\vec{X},\vec{Y})}(\vec{X},\vec{Y}) := \mathbb{E}_{(\vec{X},\vec{Y})}\left[(\vec{X}-\mathbb{E}_{\vec{X}}(\vec{X}))(\vec{Y}-\mathbb{E}_{\vec{Y}}(\vec{Y}))^T\right]$.
- For any two random vectors, $\vec{X}$ and $\vec{Y}$, they are called *uncorrelated* if their covariance matrix is the zero matrix.
- With these definitions, we can formulate a popular distribution of random vectors, the multivariate Gaussian distribution, which has a mean vector $\vec{\mu}$ and covariance matrix Σ. The density function is given by:

$$p(\vec{x};\vec{\mu},\Sigma) := (2\pi)^{-p/2}|\Sigma|^{-1/2}\exp\left(-\frac{1}{2}(\vec{x}-\vec{\mu})^T\Sigma^{-1}(\vec{x}-\vec{\mu})\right). \tag{7}$$

Marginal Distribution

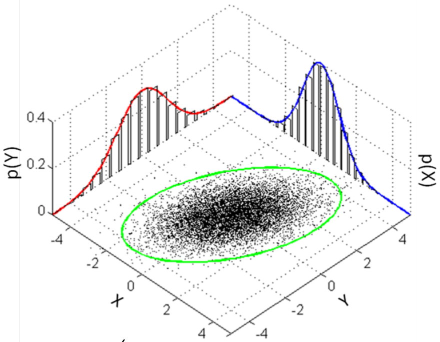
- We refer to $F(\vec{x})$ and $F(\vec{y})$ as the *marginal distributions* of $F(\vec{x}, \vec{y})$ on $\vec{X}$ and $\vec{Y}$, respectively.

Q: Why should we separately define the joint distribution? Can we express $F(\vec{x}, \vec{y})$ as a function of marginal ones, $F(\vec{x})$ and $F(\vec{y})$?

Joint Distribution $\rightarrow$ Marginal Distribution

- For a given joint distribution function $F(\vec{x}, \vec{y})$, we can obtain the marginal distributions using $F(\vec{x}) = F(\vec{x}, \infty)$ and $F(\vec{y}) = F(\infty, \vec{y})$.
- That is, when we view $(x, y, F(x, y))$ as a surface in $\mathbb{R}^3$, the marginal distribution functions represent the limits of slices of the joint distribution onto the planes defined by $(x, y = \infty, F(x, y = \infty))$ and $(x = \infty, y, F(x = \infty, y))$.
- This indicates that the complete information of the joint structure naturally implies the complete information of the marginal structures.

Joint Distribution → Marginal Distribution



- For example, when $(X, Y)^T \sim p\left(x, y; \begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix}, \begin{pmatrix} \sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2 \end{pmatrix}\right)$, $X \sim p(x; \mu_X, \sigma_X^2)$ and $Y \sim p(y; \mu_Y, \sigma_Y^2)$.

Image source: `https://en.wikipedia.org/wiki/Multivariate_normal_distribution`.

Joint Distribution $\leftarrow$ Marginal Distribution

- Marginal distributions generally do not identify their joint. Since marginal structures provide information from a finite number of slices of the joint, they are insufficient to completely recover the joint structure without additional assumptions (or unless some variables are deterministic constants).
- For instance, if X and Y follow standard Gaussian distributions, we cannot guarantee that their joint distribution is Gaussian.

  Hint: Consider $(X, Y)^T \sim \frac{1}{2} p\left(x, y; \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right) + \frac{1}{2} p\left(x, y; \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & -\rho \\ -\rho & 1 \end{pmatrix}\right)$.
- In contrast, under the same conditions, when we further assume that $(X, Y)^T$ follows a bivariate Gaussian distribution with $\mathrm{Cov}(X, Y) = \rho$, their joint distribution is uniquely specified by $p\left(x, y; \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right)$.

Conditional Distribution

- When $\vec{X} \in \mathbb{R}^p$ and $\vec{Y} \in \mathbb{R}^q$ are continuous random vectors, the *conditional distribution function* of $\vec{Y}$ given $\vec{X} = \vec{x}$ is defined as

$$\begin{aligned} F(\vec{y}|\vec{x}) &:= \mathbb{P}(\vec{Y} \leq \vec{y}|\vec{X} = \vec{x}) \\ &:= \frac{\partial^p \mathbb{P}(\vec{X} \leq \vec{x}, \vec{Y} \leq \vec{y})}{\partial x_1 \cdots \partial x_p} \Bigg/ \frac{\partial^p \mathbb{P}(\vec{X} \leq \vec{x})}{\partial x_1 \cdots \partial x_p}. \end{aligned} \tag{8}$$

- The *conditional probability density function* of $\vec{Y}$ given $\vec{X} = \vec{x}$ is defined as

$$p(\vec{y}|\vec{x}) := \frac{\partial^q F(\vec{y}|\vec{x})}{\partial y_1 \cdots \partial y_q} = \frac{p(\vec{x}, \vec{y})}{p(\vec{x})}.$$

  In this context, $\mathbb{P}(\vec{X} = \vec{x})$ is sometimes used to refer to $p(\vec{x})$.

- For example, when $(X, Y)^T \sim p\left(x, y; \begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix}, \begin{pmatrix} \sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2 \end{pmatrix}\right)$, the conditional distribution of Y given $X = x$ is $N(y; \mu_Y + \rho\sigma_Y(x - \mu_X)/\sigma_X, (1 - \rho^2)\sigma_Y^2)$.

Conditional Distribution

- Conditional expectations and variances are defined as follows:
  $\mathbb{E}_{\vec{Y}|\vec{X}}(g(\vec{Y})|\vec{X}) := \int g(\vec{y})p(\vec{y}|\vec{x})d\vec{y}$
  $\mathrm{Var}_{\vec{Y}|\vec{X}}(g(\vec{Y})|\vec{X}) := \mathbb{E}_{\vec{Y}|\vec{X}}(g(\vec{Y}) - \mathbb{E}_{\vec{Y}|\vec{X}}(g(\vec{Y})|\vec{X}))(g(\vec{Y}) - \mathbb{E}_{\vec{Y}|\vec{X}}(g(\vec{Y})|\vec{X}))^T$
- The marginal variance can be linearly decomposed as follows:

$$\mathrm{Var}_{\vec{Y}}(\vec{Y}) = \mathbb{E}_{\vec{X}}\left(\mathrm{Var}_{\vec{Y}|\vec{X}}(\vec{Y}|\vec{X})\right) + \mathrm{Var}_{\vec{X}}\left(\mathbb{E}_{\vec{Y}|\vec{X}}(\vec{Y}|\vec{X})\right) \tag{9}$$

Hint: $\mathbb{E}_{\vec{Y}}g(\vec{Y}) = \mathbb{E}_{\vec{X}}\left(\mathbb{E}_{\vec{Y}|\vec{X}}\left(g(\vec{Y})|\vec{X}\right)\right)$ and
$\vec{Y} - \mathbb{E}_{\vec{Y}}(\vec{Y}) = (\vec{Y} - \mathbb{E}_{\vec{Y}|\vec{X}}(\vec{Y}|\vec{X})) + (\mathbb{E}_{\vec{Y}|\vec{X}}(\vec{Y}|\vec{X}) - \mathbb{E}_{\vec{Y}}(\vec{Y}))$

Change of Variables

- Let $\vec{Y} = g(\vec{X})$ where g is a bijective and differentiable function. Then, the probability density function of $\vec{Y}$ can be expressed as follows: $p(\vec{y}) = p(\vec{x} = g^{-1}(\vec{y})) \left| \frac{dg^{-1}(\vec{z})}{d\vec{z}} \right|_{\vec{z}=\vec{y}}$.
- The rigorous proof requires several technical details. Roughly speaking,

$$\begin{aligned} p(\vec{y}) &\approx \frac{\mathbb{P}(\vec{y} \leq \vec{Y} \leq \vec{y} + \Delta\vec{y})}{\Delta\vec{y}} \\ &\approx \frac{\mathbb{P}(\vec{X} \text{ is in between } g^{-1}(\vec{y}) \text{ and } g^{-1}(\vec{y} + \Delta\vec{y}))}{\Delta\vec{y}} \\ &\approx p(\vec{x} = g^{-1}(\vec{y})) \left| \frac{\Delta g^{-1}(\vec{y})}{\Delta\vec{y}} \right|. \end{aligned} \tag{10}$$

- See Section A.8 of Bickel and Doksum (2015) for a detailed proof.

Change of Variables

- Let $(X, Y)^T \sim p\left(x, y; \begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix}, \begin{pmatrix} \sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2 \end{pmatrix}\right)$ and define $(Z, W)^T$ as follows:

$$\begin{aligned} Z &= \frac{X - \mu_X}{\sigma_X}, \\ W &= \frac{1}{\sqrt{1-\rho^2}}\left(\frac{Y - \mu_Y}{\sigma_Y} - \rho\frac{X - \mu_X}{\sigma_X}\right) \end{aligned} \tag{11}$$

  Then, $(Z, W)^T$ follows the bivariate standard Gaussian distribution, with the mean vector being zero and the covariance matrix being the identity matrix.
- With this representation, Y can be expressed as $Y = \mu_Y + \rho\sigma_Y(X - \mu_X)/\sigma_X + \sqrt{1-\rho^2}W$. The covariance between X and W is zero. Since $p(y)$ is a marginal of the bivariate Gaussian, it is a univariate Gaussian. Thus, the conditional distribution of Y given $X = x$ is $N\left(y; \mu_Y + \rho\sigma_Y(x - \mu_X)/\sigma_X, (1-\rho^2)\sigma_Y^2\right)$.

Independency

- For any two events A and B, they are called *independent* when $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$.
- For any two random variables X and Y, they are called *independent* when events $\{X \leq x\} := \{w \in \Omega | X(w) \leq x\}$ and $\{Y \leq y\} := \{w \in \Omega | Y(w) \leq y\}$ are independent for any $x, y \in \mathbb{R}$. That is, $F(x, y) = F(x)F(y)$, which is equivalent to $p(x, y) = p(x)p(y)$ or $p(x|y) = p(x)$.
- The independency of two random variables implies that observing one does not alter (conditional) distributions of the other.

Independency

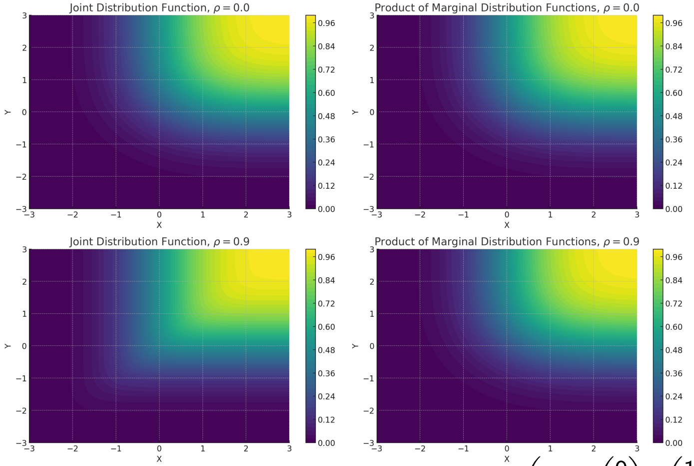


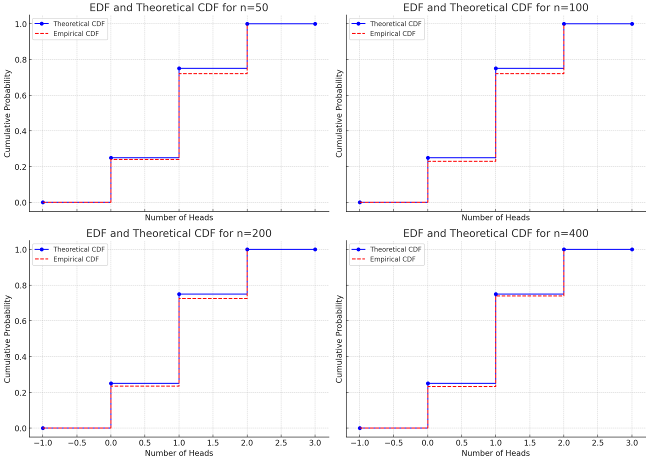
Figure: $F(x, y)$ vs. $F(x)F(y)$ when $(X, Y)^T \sim N\left(x, y; \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right)$

Random Sample and Empirical Estimation

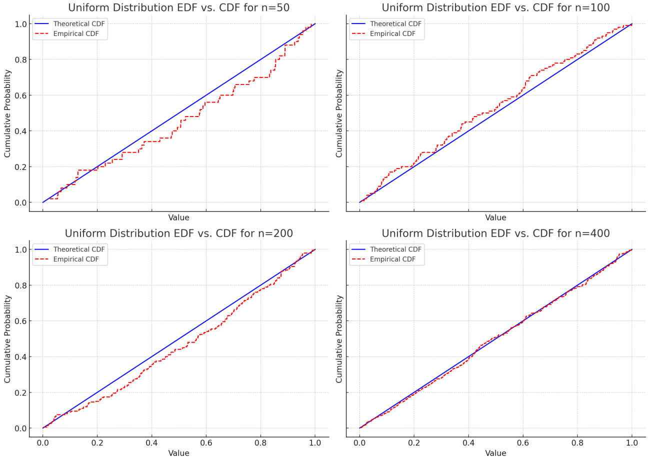
- Random vectors $\vec{x}_1, \ldots, \vec{x}_n$ are called independent and identically distributed (i.i.d.) samples from $p(\vec{x})$ when they are independent and each follows the distribution $p(\vec{x})$.
- *Empirical Distribution Function*: $\widehat{F}_n(x) := n^{-1} \sum_{i=1}^{n} \mathbf{1}(x_i \leq x)$. Note that $\widehat{F}_n$ is discontinuous even when X is a continuous random variable.
- *Empirical Measure*: $\mathbb{P}_n(E) := n^{-1} \sum_{i=1}^{n} \mathbf{1}(x_i \in E) = n^{-1} \sum_{i=1}^{n} \delta_{x_i}(E)$ for any event E.[4]
- *Empirical Estimate of $\mathbb{E}_X g(X)$*: $n^{-1} \sum_{i=1}^{n} g(x_i)$, where $\vec{x}_1, \ldots, \vec{x}_n$ are i.i.d. samples from $p(x)$.

[4] We sometimes use p_n to refer to the empirical measure.

Empirical vs. True Distribution Function (Discrete Random Variable)



Empirical vs. True Distribution Function (Continuous Random Variable)



Outline

Density Model

- We have formulated distributions for random vectors and the dependencies between their elements.
- Generative models introduce density models that learn the distribution of high-dimensional and complex data.
- In this subsection, we motivate the problem of density estimation and introduce several traditional statistical models.

Recapping Motivating Example

- **Q:** This dice has two different types of faces. How can we estimate the distribution of outcomes when the dice is rolled?

Left: A restored copy at Gyeongju National Museum. Right: Lettering on each side at the time of excavation.

Motivating Example: A Non-parametric Approach

- Histograms directly estimate the probabilities of each event (without parametrization).
- Wang et al. (2005) rolled the dice $n = 7,000$ times using two different materials, Wooden and Fiberglass Reinforced Plastic (FRP). The histogram is as follows (Wooden/FRP):

Outcome	1st □	2nd □	3rd □	4th □	5th □	6th □	1st ⬡
Frequency	523/531	533/522	477/519	525/504	518/552	492/504	456/481

Outcome	2nd ⬡	3rd ⬡	4th ⬡	5th ⬡	6th ⬡	7th ⬡	8th ⬡
Frequency	512/505	506/487	488/490	505/479	487/484	482/470	496/472

Motivating Example: A Parametric Approach

- When we assume that faces of the same type have equal probabilities of appearing, we can express all the probabilities of events, i.e., the distribution, using a single parameter p as follows:

Outcome	1st □	2nd □	3rd □	4th □	5th □	6th □	1st ⬡
Probability	p	p	p	p	p	p	$\frac{1-6p}{8}$

Outcome	2nd ⬡	3rd ⬡	4th ⬡	5th ⬡	6th ⬡	7th ⬡	8th ⬡
Probability	$\frac{1-6p}{8}$	$\frac{1-6p}{8}$	$\frac{1-6p}{8}$	$\frac{1-6p}{8}$	$\frac{1-6p}{8}$	$\frac{1-6p}{8}$	$\frac{1-6p}{8}$

- In the experiment with $n = 7,000$, wooden dice, empirical estimates based on frequencies are $\hat{p} = (3,068/7,000)/6 \approx 7.3\%$ and $(1 - 6\hat{p})/8 \approx 7.0\%$.

Parametric Model

- Parametric models utilize (low-dimensional) parameters, denoted by $\theta \in \Theta$, to describe the distribution of random variables.
- For discrete random variables, examples include:
  1. Bernoulli distribution
  2. Binomial distribution
  3. Categorical distribution (or Multinoulli distribution)
  4. Multinomial distribution
- For continuous cases,
  1. Gaussian distribution (or Normal distribution)

Parametric Model: Discrete Examples

- *Bernoulli Distribution:* Let $p \in [0,1]$. We say $X \sim \text{Ber}(p)$ when X equals 1 with probability (w.p.) p and 0 w.p. $1-p$.

x	0	1
$\mathbb{P}_p(X=x)$	$1-p$	p

- *Binomial Distribution:* Let $n \in \mathbb{N}$ and $p \in [0,1]$. We say $X \sim \text{B}(n,p)$ when X can be expressed as the summation of n i.i.d. samples from $\text{Ber}(p)$.

x	0	$\cdots$	x	$\cdots$	n
$\mathbb{P}_{(n,p)}(X=x)$	p^n	$\cdots$	$\frac{n!}{(x)!(n-x)!}p^x(1-p)^{n-x}$	$\cdots$	$(1-p)^n$

Parametric Model: Discrete Examples

- *Categorical Distribution (or Multinoulli Distribution):* Let $\vec{p} \in \mathbb{R}^p$ satisfy $\sum_{j=1}^p p_j = 1$ and $p_j \geq 0$, and $\vec{e}_j$ be the j-th standard basis vector. We say $\vec{X} \sim \mathrm{Cate}(\vec{p})$ where $\vec{X} \in \mathbb{R}^p$ when $p(\vec{X} = \vec{e}_j) = p_j$.

$\vec{x}$	$\vec{e}_1$	$\cdots$	$\vec{e}_p$
$\mathbb{P}_{\vec{p}}(\vec{X} = \vec{x})$	p_1	$\cdots$	p_p

- *Multinomial Distribution:* Let $n \in \mathbb{N}$ and $\vec{p} \in \mathbb{R}^p$ satisfy $\sum_{j=1}^p p_j = 1$ and $p_j \geq 0$. We say $\vec{X} \sim \mathrm{Multi}(n, \vec{p})$ when $\vec{X}$ can be expressed as the summation of n i.i.d. samples from $\mathrm{Cate}(\vec{p})$.

$\vec{x}$	$n\vec{e}_1$	$\cdots$	$\sum_{j=1}^p x_j \vec{e}_j$	$\cdots$	$n\vec{e}_p$
$\mathbb{P}_{(n,\vec{p})}(\vec{X} = \vec{x})$	p_1^n	$\cdots$	$\frac{n!}{(x_1)!\cdots(x_p)!} p_1^{x_1} \cdots p_p^{x_p}$	$\cdots$	p_p^n

Parametric Model: Continuous Examples

- *Gaussian distribution (or Normal distribution)*: $p_\theta(\vec{x}) := p(\vec{x}; \vec{\mu}, \Sigma)$, where $\theta := (\vec{\mu}, \Sigma)$. In this case, $\mathbb{P}_\theta(\vec{X} \leq \vec{x}) = \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_p} p_\theta(\vec{x}) d\vec{x}$. Thus, we can express the probabilities of all possible events using θ.
- When we have a p-dimensional $\vec{X}$, the number of parameters is p (for $\vec{\mu}$) plus $p(p+1)/2$ (for Σ), which is $O(p^2)$. Assuming a sparse covariance structure, e.g., $\Sigma_{i,j} = 0$ when $|i - j| > 1$, reduces this to $O(p)$.

Parametric vs. Non-parametric

- Parametric methods assume that the distribution can be characterized by a finite number of parameters (sometimes just a few). In contrast, non-parametric methods, such as kernel density estimation, impose minimal assumptions on the distribution.
- Most deep generative models are parametric, utilizing parameters from neural networks.
- Roughly speaking, parametric methods are more effective when the true underlying distribution conforms to their assumptions, while non-parametric methods are preferable in situations where this is not the case.

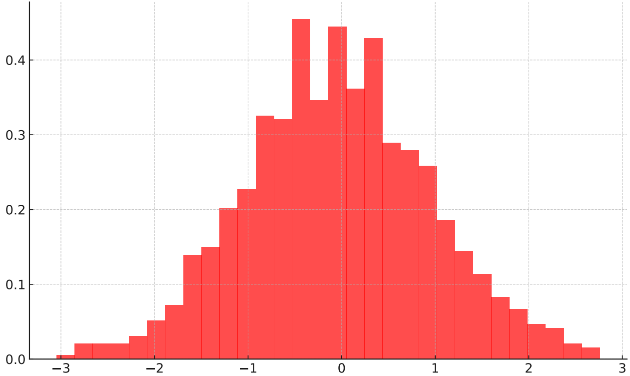
Outline

Likelihood Maximization Principle

- There are diverse approaches to estimating the distribution of multivariate data, which aim to identify the best model distributions among specified density models.
- In this subsection, we review the principle of likelihood maximization, focusing on maximum likelihood estimators and their relationship with Kullback-Leibler divergence, a measure of statistical distance.

Likelihood Maximization: Gaussian Example

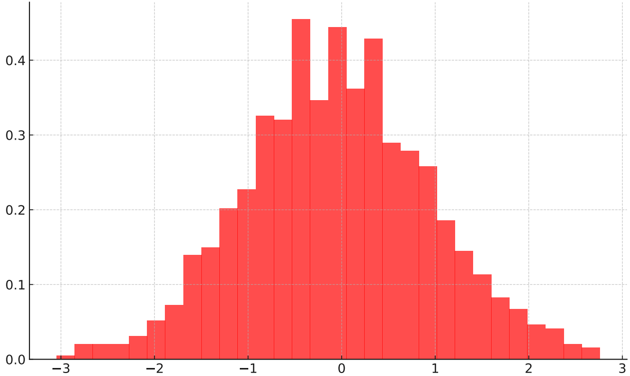
- Let's begin with a basic example. Assume we observed n univariate samples $x_1, \ldots, x_n$ having the following histogram:



Likelihood Maximization: Gaussian Example

- Given its bell-shaped curve, Gaussian distributions may be considered:

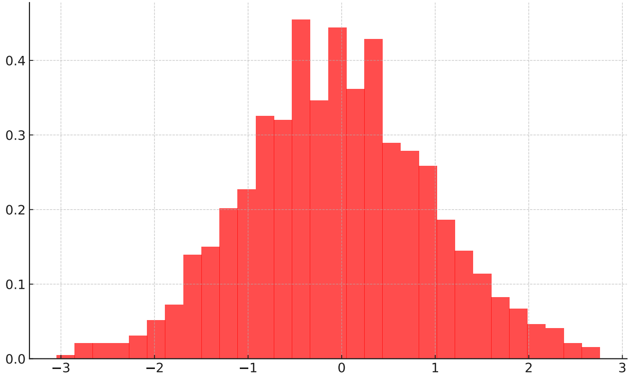
$$p_{\theta:=(\mu,\sigma^2)}(x) = (\sqrt{2\pi\sigma^2})^{-1} \exp(-(x-\mu)^2/2\sigma^2).$$



Likelihood Maximization: Gaussian Example

- Which θ would be better to describe these observations: $\theta_1 = (0,5)^T$ or $\theta_2 = (3,1)^T$?

$$p_{\theta:=(\mu,\sigma^2)}(x) = (\sqrt{2\pi\sigma^2})^{-1} \exp(-(x-\mu)^2/2\sigma^2).$$



Likelihood Maximization: Gaussian Example

- *Likelihoods*, defined as $L_n(\theta) := \prod_{i=1}^n p_\theta(x_i)$ are one of the popular criteria.
- Note that $p_\theta(x_i) := d\mathbb{P}_\theta(X \le x)/dx|_{x=x_i}$ indicates the possibility of X being (nearby) x_i, if the true density were p_θ.
- Therefore, the likelihood quantifies the possibility of obtaining the observations under the assumption that the model distribution is exactly the same as the true distribution.
- In this example, the log-likelihoods can be expressed as $l_n(\theta) := \sum_{i=1}^n \log p_\theta(x_i) = -(n/2)\log\sigma^2 - \sum_{i=1}^n (x_i - \mu)^2/(2\sigma^2) - (n/2)\log 2\pi$.
- Maximizers of (log-)likelihoods are called *maximum likelihood estimators* (MLEs). In this example, the MLE can be expressed as:

$$\arg\max_\theta l_n(\theta) = (\hat{\mu}_n, \hat{\sigma_n^2})^T = \left(\bar{x}, n^{-1}\sum_{i=1}^n (x_i - \bar{x})^2\right)^T. \tag{12}$$

Density Estimation with Maximum Likelihood Estimator

- A better θ is expected to yield better density estimation with the corresponding p_θ.
- Consistency and asymptotic normality of MLEs are important to characterize and justify their efficacy.
- Note that MLEs are functions of random samples. To discuss the properties of MLEs, we first review the convergence of random variables.

Convergence of Random Variable

- Let $(X_n)_{n=1}^N$ be a sequence of relative frequencies of heads obtained from tossing a coin n times, where $X_n = n^{-1} \sum_{i=1}^n I(\text{the } i\text{-th coin is head})$.
- What can we expect for X_N as $N \to \infty$? Since X_n is a random quantity, the sequence $(X_n)_{n=1}^N$ is also random. For example, $\mathbb{P}(X_1 = \cdots = X_N = 0) = 1/2^{N(N+1)/2}$ (all tails) and $\mathbb{P}(X_1 = \cdots = X_N = 1) = 1/2^{N(N+1)/2}$ (all heads).
- We need a notion of convergence tailored to address such randomness. Popular ones include:
  1. Convergence in probability
  2. Convergence in distribution

Convergence of Random Variable

1. **Convergence in Probability:** In the previous example, for any $\epsilon > 0$, the sequence $(\mathbb{P}(|X_n - 1/2| \geq \epsilon))_{n=1}^{N}$, consisting of deterministic real numbers, converges to zero. By Chebyshev's inequality,

$$\mathbb{P}(|X_n - 1/2| \geq \epsilon) \leq \epsilon^{-2} \mathrm{Var}_{X_n}(X_n) = \epsilon^{-2} n^{-1}/4. \tag{13}$$

- We say that $(X_n)_n$ converges to $c \in \mathbb{R}$ in probability, denoted by $X_n \xrightarrow{p} c$, when for any $\epsilon > 0$, $\mathbb{P}(|X_n - c| \geq \epsilon) \to 0$.
- Let $\bar{X}$ be the mean of n i.i.d. samples from a distribution with a finite first moment μ and second moment.[5] Then, $\bar{X} \xrightarrow{p} \mu$.

[5] When we remove this condition on the finiteness of the second moment, it becomes a special case of the (weak) law of large numbers.

Convergence of Random Variable

2. **Convergence in Distribution:** In the previous example, $\mathbb{P}(X_n \leq x) \to \mathbb{P}(1/2 \leq x)$ for any $x \in \mathbb{R}$:
   When $x < 1/2$, $\mathbb{P}(X_n \leq x) \leq \mathbb{P}(|X_n - 1/2| \geq 1/2 - x) \to 0$.
   When $x \geq 1/2$,
   $\mathbb{P}(X_n \leq x) = 1 - \mathbb{P}(X_n - 1/2 > x - 1/2) \leq 1 - \mathbb{P}(|X_n - 1/2| \leq x - 1/2) \to 1$.

- We say that $(X_n)_n$ converges to X in distribution, denoted by $X_n \xrightarrow{d} X$, when for any $x \in \mathbb{R}$, $\mathbb{P}(X_n \leq x) \to \mathbb{P}(X \leq x)$. Note that convergence in probability implies convergence in distribution.
- *Central Limit Theorem:* Let $\bar{X}$ be the mean of n i.i.d. samples from a distribution with a finite population mean μ and standard deviation σ. Then, $\sqrt{n}(\bar{X} - \mu)/\sigma \xrightarrow{d} Z \sim N(0, 1)$.
- See Durrett (2019) for another notion of convergence of random variables, **almost sure convergence**, as well as details on the central limit theorem and the strong law of large numbers.

Consistency

- Let $x_1, \ldots, x_n$ be i.i.d. samples from p_{θ_0}, a model among specified parametric density models $\{p_\theta | \theta \in \Theta\}$.
- An estimator based on the samples, $\hat{\theta}_n$, is called *consistent* when $\hat{\theta}_n \xrightarrow{p} \theta_0$.
- Under some conditions[6], MLEs are consistent. For example, when the density model class is Gaussian distributions and $\theta_0 := (\mu_0, \sigma_0^2)^T$, the MLE is

$$(\hat{\mu}_n, \hat{\sigma_n^2})^T = \left(\bar{x}, n^{-1} \sum_{i=1}^{n} (x_i - \bar{x})^2\right)^T. \tag{14}$$

and $(\hat{\mu}_n, \hat{\sigma_n^2})^T \xrightarrow{p} (\mu_0, \sigma_0^2)^T$ by the law of large numbers.

[6] See Section 5 of Bickel and Doksum (2015) for details on the conditions and proof.

Asymptotic Normality

- Under some conditions[7], the $\sqrt{n}\left(\hat{\theta}_n - \theta_0\right)$ converges to a normal distribution:

$$\sqrt{n}\left(\hat{\theta}_n - \theta_0\right) \xrightarrow{d} N\left(0, (I(\theta))^{-1}\right) \tag{15}$$

where $I(\theta) := -\mathbb{E}_{X\sim p_\theta(x)}\left(\partial^2 \log p_\theta(X)/\partial\theta^2\right)$ represents the *Fisher information*.

Sketch of Proof: Under some conditions, $\hat{\theta}_n$ is a solution to $\dot{l}_n(\theta) = 0$. A Taylor expansion around $\theta = \theta_0$ yields $0 = \dot{l}_n(\hat{\theta}_n) \approx \dot{l}_n(\theta_0) + \ddot{l}_n(\hat{\theta}_n)(\hat{\theta}_n - \theta_0)$, implying that $\sqrt{n}(\hat{\theta}_n - \theta_0) \approx -\sqrt{n}\left(\ddot{l}_n(\theta_0)\right)^{-1}\dot{l}_n(\theta_0)$. Now, applying law of large numbers to $\ddot{l}_n(\theta_0)$ and central limit theorem to $\dot{l}_n(\theta_0)$ imply that the RHS converges to the Gaussian distribution.

[7] Again, see Section 5 of Bickel and Doksum (2015) for details on the conditions and proof.

Likelihood Maximization and Statistical Distance Minimization

- We can explain the maximum likelihood principle using Kullback-Leibler (KL) divergence.
- Note that $n^{-1}l_n(\theta) = \int \Big(\log p_\theta(x) \Big) \mathbb{P}_n(dx)$. It implies

$$\begin{aligned} n^{-1}l_n(\theta) &= \int \Big(\log \mathbb{P}_\theta(dx)/\mathbb{P}_n(dx) \Big) \mathbb{P}_n(dx) + \int \Big(\log d\mathbb{P}_n(x)/dx \Big) \mathbb{P}_n(dx) \\ &= -\mathrm{KL}(\mathbb{P}_n || \mathbb{P}_\theta) + \mathrm{C} \end{aligned} \tag{16}$$

  where C is a constant w.r.t. θ.[8]
- Thus, MLEs are minimizers of the KL divergence between the empirical measure and the model density.

[8] Formally, $\mathrm{KL}(\mathbb{P}||\mathbb{Q}) := \int \log \Big(\mathbb{P}(dx)/\mathbb{Q}(dx) \Big) \mathbb{P}(dx)$ where $\mathbb{P} \ll \mathbb{Q}$.

References I

Bickel, P. J. and Doksum, K. A. (2015). *Mathematical statistics: basic ideas and selected topics, volumes I-II package*. Chapman and Hall/CRC.

Durrett, R. (2019). *Probability: theory and examples*, volume 49. Cambridge university press.

Wang, M.-O., Seo, J.-C., and Lim, I.-K. (2005). Probability research of wooden die for drinking game as ethnic custom mathematics. *Journal for History of Mathematics*, 18(4):67–84.