

Lecture 6: Statistical Decision Theory and Sufficiency

Lecturer: Yiqiao Zhong

Last update: 19:34:31, 2024-09-24

Disclaimer: materials in this pdf are used for STAT 709 “Mathematical Statistics” at the University of Wisconsin–Madison, but they are not guaranteed to be accurate. Consult textbooks whenever possible.

Our goal is to introduce basic statistical notions with an emphasis on statistical thinking.

1 Data science in Scientific Experiments

Data science is an ever-evolving concept. With abundant data collected by the internet and digital device, data science has revolutionizing effects on our society.

Historically, statistics have been widely used by scientists and social workers to design experiments/surveys and analyze data.

- Procedure: scientific hypothesis $\rightarrow$ **data collection** $\rightarrow$ **experimental design** $\rightarrow$ **data analysis** $\rightarrow$ conclusions & scientific results.
- Additional steps: Descriptive data analysis, exploratory data analysis including data visualization.
- Focus of mathematical statistics: statistical inference and decision theory.

From the statistical view, there are two themes underlying data: **statistical modeling** and **uncertainty qualification**. The first provides interpretability and is strongly connected to the scientific hypothesis under investigation. The second provides the strength of confidence for the conclusion (e.g., whether the sample size is sufficiently large, whether there is systematic bias, etc.)

There is a different view from the engineering world: data are structured and **deterministic** resources for improving user’s experience. For example, generative AI uses data to create a **black-box model** to interact with humans. A key source of randomness comes from algorithms.

2 Basic Terminology in Statistical Models

Suppose that

$$\text{data}_{(\text{sample})} \sim (\Omega, \mathcal{F}, P).$$

Let P be the population distribution. Usually, $X_1, X_2, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} P$. Let $X = (x_1, x_2, \dots, x_n)^\top \in \mathbb{R}^{n \times p}$ be the data matrix with sample size n . For each x_i , it can be a unit, observation, measurement, or data point. Our goal is to deduce P based on data.

Sample vs. population. From the statistical perspective, our observed data are samples drawn from a large “universe” of population, so a first-step analysis is to understand how samples relate to population structure.

Example 1. (*Graph data*) In the Erdős–Rényi model $\mathcal{G}(n, p)$, we observe a graph G of n vertices where there is an edge between vertex i and j with probability p . In other words, we observe i.i.d. Bernoulli variable $A_{ij} \sim \text{Bern}(p)$ for $1 \leq i < j \leq n$. We view our observation as the symmetric matrix A , and the “population” matrix as its expectation

$$\mathbb{E} A = p \mathbf{1}_n \mathbf{1}_n^\top$$

which is a rank-one matrix. Under certain conditions, $A \approx \mathbb{E} A$, so for a statistical procedure $T(A)$ we should reasonably expect $T(A) \approx T(\mathbb{E} A)$. A further question is to investigate such approximation and the fluctuation.

2.1 Classifying statistical models

Parametric family. If we assume

$$P \in \{P_\theta : \theta \in \Theta \subset \mathbb{R}^p\},$$

where $\Theta \subset \mathbb{R}^p$ is the parameter space, and p is the dimension of $\theta \in \Theta$. Sometimes we denote by $\mathcal{P}$ the distribution family $\{P_\theta : \theta \in \Theta \subset \mathbb{R}^p\}$. Our goal is to infer θ or a function of θ based on the data.

Family of densities: If $P_\theta \ll \nu, \forall \theta \in \Theta$, then we only need to consider $\{\frac{dP_\theta}{d\nu} : \theta \in \Theta\}$, particularly pdf/pmf $P_\theta(x)$.

Example 2. $P \in \mathcal{P} = \{\mathcal{N}(\mu, \sigma^2) : \mu \in \mathbb{R}, \sigma^2 > 0\}$. The parameters are $\theta = (\mu, \sigma^2)$.

Example 3. (*Linear model*) $\mathcal{P} = \{\mathcal{N}(\beta^T X + \beta_0, \sigma^2) : \beta \in \mathbb{R}^p, \beta_0 \in \mathbb{R}, \sigma^2 > 0\}$, where $X \in \mathbb{R}^p$ is observed and fixed, usually known as covariates, features or attributes, and β_0 is called intercept or bias. An equivalent way of writing the linear model is

$$Y = \beta^T X + \beta_0 + \sigma \varepsilon, \quad \text{where } \varepsilon \sim N(0, 1) \quad (1)$$

Nonparametric family. Different from the parametric family, now the parameter space Θ is an infinite-dimensional space. Despite the infinite-dimensional parameter space, a statistical method often uses a finite number of estimates to approximate the unknown quantity.

Example 4. (*Density estimation*) Each distribution in the family is absolutely continuous w.r.t. the Lebesgue measure, and the set of pdf is $\{f(x) : |f'(x)| \leq L\}$ for a certain fixed $L > 0$. This family consists of density functions with a bounded Lipschitz constant.

Semiparametric family. For a semiparametric family, Θ contains finite-dimensional parameters of interest, as well as (infinite-dimensional) nuisance parameters.

Example 5. (*Linear model 2*) Consider the same linear model as in Eqn. 1, but we only assume that ε has zero mean and a bounded fourth moment. Our interest is the finite-dimensional parameters β, β_0 but there is a nuisance infinite-dimensional parameter, the distribution of ε .

2.2 Identifiability

$\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ is identifiable if and only if $P_{\theta_1} \neq P_{\theta_2}$ for $\theta_1 \neq \theta_2$. Identifiability is important in disciplines (e.g., economics) where interpretability is highly valued.

Example 6. (*Factor model*) Consider that each observation $Y \in \mathbb{R}^p$ has the structure

$$Y = Bf + u, \quad \text{where } f \sim N(0, I_r), \quad u \sim N(0, \Sigma) \quad (2)$$

and where Σ is known, and the unknown parameters are the matrix $B \in \mathbb{R}^{p \times r}$. Equivalently, we assume that $Y \sim N(0, BB^\top + \Sigma)$. Since $BR(BR)^\top = BB^\top$ for any orthogonal matrix $R \in O(r)$, this model is not identifiable unless we put further constraints on B .

In classical statistics/ML literature, latent variable models are a broad class of models where non-identifiability is an issue.

Example 7. (*Multinomial regression*) Suppose that each observation takes a value in a finite set $\{1, 2, \dots, K\}$ (categorical data) with probabilities $p_1, p_2, \dots, p_K$, and the probabilities are determined by

$$p_k = \frac{\exp(\beta_k^\top X)}{\sum_{j=1}^K \exp(\beta_j^\top X)}$$

where X is fixed and observed, and unknown parameters are $(\beta_1, \dots, \beta_K)$. For any $c \in \mathbb{R}^p$, both $(\beta_1, \dots, \beta_K)$ and $(\beta_1 + c, \dots, \beta_K + c)$ stipulate the same statistical model, thus creating an identifiability issue. Usually, we choose a reference class, say, the first class, to avoid non-identifiability:

$$p_k = \frac{\exp(\beta_k^\top X)}{\sum_{j=1}^{K-1} \exp(\beta_j^\top X) + 1}, \quad \text{where } k = 1, 2, \dots, K-1$$

Equivalently we set $\beta_k \leftarrow \beta_k - \beta_K$.

In machine learning, this is also known as **softmax regression**, but since interpretability is not a top priority, the identifiability issue is not a concern.

2.3 Statistics

Let X be the data drawn from P . For example, $X = (x_1, x_2, \dots, x_n)^\top \in \mathbb{R}^{n \times p}$. For any measurable function T , we call $T(X)$ a *statistic*.

By definition, $\sigma(T(X)) \subset \sigma(X)$. Intuitively $T(X)$ distills/summarizes information from X .

Example 8. *Common Statistics: sample mean $\bar{X}$ and sample variance S^2 , where*

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i, \quad S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2.$$

Note: empirical variance uses $\frac{1}{n}$ instead of $\frac{1}{n-1}$, though the difference is minor when n is large.

Order Statistics. Let $X_1, X_2, \dots, X_n$ be i.i.d. random variables. Let $X_{(i)}$ be the i th smallest in $X_1, \dots, X_n$. Then $X_{(1)}, \dots, X_{(n)}$ are called order statistics. Assume X_i has Lebesgue p.d.f. f . Then, the Lebesgue p.d.f. of $X_{(i)}$ exists and is given by

$$g_i(x) = \frac{n!}{(i-1)!(n-i)!} [F(x)]^{i-1} [1-F(x)]^{n-i} f(x).$$

To prove this claim, we only need to realize, for any constant $\varepsilon > 0$,

$$\begin{aligned} P(x \leq X_{(i)} \leq x + \varepsilon) \\ &= P(\exists \text{ non-identical } j_1, j_2, \dots, j_n \text{ such that } X_{j_1}, \dots, X_{j_{i-1}} < x \text{ and } x \leq X_{j_i} \leq x + \varepsilon \text{ and } X_{j_{i+1}}, \dots, X_{j_n} > x + \varepsilon) \\ &= \sum_{j_1, j_2, \dots, j_n} P(X_{j_1}, \dots, X_{j_{i-1}} < x) \cdot P(x \leq X_{j_i} \leq x + \varepsilon) \cdot P(X_{j_{i+1}}, \dots, X_{j_n} > x + \varepsilon) \end{aligned}$$

where $j_1, \dots, j_n$ is a permutation of $1, 2, \dots, n$. We then set $\varepsilon \rightarrow 0$ and use cdf/pdf to simplify the probabilities.

Remark 1. In a previous assignment, we used probability tail bound to study $\max_{i \leq n} X_i$ for i.i.d. normal variables X_i :

$$\frac{1}{\sqrt{2 \log n}} \max_{1 \leq i \leq n} X_i \xrightarrow{P} 1.$$

An alternative way is to use the p.d.f. formula for order statistics ([exercise](#)).

2.4 Exponential Family

Definition 1. (*Exponential Family*) A general form of exponential family is defined by the pdf

$$f_{\theta}(\omega) = \exp\{\eta(\theta)^T T(\omega) - \xi(\theta)\} h(\omega), \quad \forall \omega \in \Omega.$$

Its canonical form is

$$f_{\eta}(\omega) = \exp\{\eta^T T(\omega) - \xi(\eta)\} h(\omega), \quad \forall \omega \in \Omega.$$

Example 9. Binomial, normal family, and many distribution families are subsumed in the exponential family. (Read Example 2.5-2.7 in Chapter 2.1.3. of [3]).

Properties of exponential family. The exponential family is very useful and enjoys many useful properties. Below are the basic properties of differentiability.

Theorem 1. (Theorem 2.1 in Chapter 2.1.3 of [3])

(i) Marginal distributions / conditional distributions remain exponential family.

(ii) If f is a Borel function satisfying $\int |f| dP_{\mu_0} < \infty$, then

$$\int f(\omega) \exp(\eta^T T(\omega)) h(\omega) d\nu(\omega) \in C^{\infty},$$

where C^{∞} is the set of all smooth functions.

3 Basic Terminology in Statistical Decision Theory

- Decision rule: statistics $T : (\mathbb{R}^k, \mathcal{B}^k) \rightarrow (A, \mathcal{F}_A)$.
- Loss function: $L(P, T(X))$.
- Risk function: expected loss $R_T(P) = \mathbb{E}[L(P, T(X))]$. For parametric family, it is sometimes written as $R(\theta)$. For example, mean squared error (MSE),

$$\text{MSE}_{\hat{\theta}}(\theta) = \mathbb{E}_{\theta}[(\hat{\theta} - \theta)^2] = \int (\hat{\theta}(x) - \theta)^2 dP_{\theta}(x).$$

3.1 Optimal Rule

The decision rule T_* is called the optimal rule if

$$R_{T_*}(P) \leq R_T(P), \quad \forall P \in \mathcal{P}, \quad \forall T \in \{\text{Class of allowable decisions}\}.$$

The definition of optimal rules is stringent, and a natural rule may not be optimal. (In fact, an optimal rule may not exist.) For example, to estimate the mean μ based on an observation $X \sim N(\mu, 1)$, the ‘natural’ rule $T_1(X) = X$ and the ‘trivial’ rule $T_2(X) \equiv 0$ have MSE 1 and μ^2 respectively. So the ‘trivial’ rule actually wins if $|\mu| < 1$.

Sometimes, it is easy to calculate only the order of $R_T(P)$ (up to a constant) involving sample size, dimension, and other parameters. We call a decision rule T_* (especially an estimator) rate-optimal if $R_{T_*}(P) \leq C R_T(P)$ holds for a certain constant C .

Randomized decision rules. For every x , $\delta(x, \cdot)$ is a probability measure on $(A, \mathcal{F}_A)$. Often, $\delta = \sum_{j=1}^N \mathbb{I}_{\{Z=j\}} T_j$, where Z is a discrete random variable. The risk of a randomized rule is defined as $R_{\delta}(P) = \mathbb{E}[L(P, \delta, X)]$.

This definition includes randomized algorithms such as stochastic gradient descent.

Admissibility. Let $\mathcal{G}$ be a class of decision rules, $T \in \mathcal{G}$ is $\mathcal{G}$ -admissible if and only if there is no $S \in \mathcal{G}$, such that

$$R_S(P) \leq R_T(P), \forall P \in \mathcal{P},$$

and strict inequality holds for certain P .

Two rules. In the statistical literature, there are two main approaches to study the optimality of decision rules.

- Bayes rule: Given a probability distribution Π over P , T_* minimizes

$$r_T(\Pi) = \int R_T(P) d\Pi(P).$$

- Minimax rule: T_* minimizes

$$\sup_{P \in \mathcal{P}} R_T(P).$$

3.2 Estimation

Consider $T(X) = \bar{X}$, our goal is to estimate the mean of X_i , based on i.i.d. $X_1, \dots, X_n$. Use squared loss $L(P, a) = (\theta - a)^2$, then

$$R_{\bar{X}}(P) = \frac{\sigma^2}{n}, \text{ where } \sigma^2 = \text{Var}(X_i).$$

Bias. Given an estimator $T(X)$, and a parameter to estimate ν , define

$$b_T(P) = \mathbb{E}[T(X)] - \nu.$$

The estimator $T(X)$ is called unbiased if and only if $b_T(P) = 0$, $\forall P \in \mathcal{P}$. For example, for variance estimation, sample variance is unbiased, and empirical variance is biased.

3.3 Hypothesis Test

Hypothesis tests can be formulated as a statistical decision problem. Suppose that we are interested in the hypothesis test

$$H_0 : P \in \mathcal{P}_0 \text{ vs. } H_1 : P \in \mathcal{P}_1.$$

Let $T(X) \in \{0, 1\}$, $T(X)$ must have the form $I_C(X) = \{X : T(X) = 1\}$, where C is called rejection region. Consider the 0-1 loss,

$$L(P, j) = 0, \text{ if } P \in \mathcal{P}_j, \text{ and } 1 \text{ otherwise.}$$

Then,

$$R_T(P) = \begin{cases} P(T(X) = 1) = P(X \in C), & P \in \mathcal{P}_0 \\ P(T(X) = 0) = P(X \notin C), & P \in \mathcal{P}_1. \end{cases}$$

4 Sufficient Statistics

A statistic $T(X)$ is sufficient if and only if conditional distribution $X|T$ does not depend on P or θ (i.e., conditional distribution does not change with varying P).

Theorem 2. (Factorization Theorem) Suppose $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ is dominated by the Lebesgue measure ν (i.e., Lebesgue pdf exists for every θ). Then $T(X)$ is sufficient if and only if there exists nonnegative Borel functions h, g_θ , such that

$$\frac{dP_\theta}{d\nu}(x) = g_\theta(T(x))h(x).$$

Example 10. (*Exponential Family*) Recall that the exponential family is defined as

$$f_{\theta}(\omega) = \exp\{\eta(\theta)^T T(\omega) - \xi(\theta)\} h(\omega).$$

Take $g_{\theta}(t) = \exp(\eta(\theta)^T t - \xi(\theta))$. By the factorization theorem, $T(X)$ is a sufficient statistic for θ .

Example 11. (*Order Statistics*) Let $X_1, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} P \in \mathcal{P}$. Let $X_{(1)}, \dots, X_{(n)}$ be the corresponding order statistics. It is obvious that

$$f(X_1) \cdots f(X_n) = f(X_{(1)}) \cdots f(X_{(n)}).$$

Therefore, $T(X) = (X_{(1)}, \dots, X_{(n)})$ is sufficient for P .

In this example, the ranks of $X_1, \dots, X_n$ are not relevant to referring P .

Connection to information theory*. From the information-theoretic view, the unknown parameter θ is a random variable (Bayesian view), and $T(X)$ is a sufficient statistic if and only if

$$I(X; \theta) = I(T(X); \theta)$$

where I denotes the mutual information. Since the data processing inequality implies $I(T(X); \theta) \leq I(X; \theta)$, sufficiency means that $T(X)$ retains all relevant information for inferring θ based on observation X .

Example 12. (Read Chapter 2.2.2 Example 2.10 of [3]) Concrete calculation provided.

5 Further reading

Statistical decision theory is a classical research area and sufficiency is a classical notion for data reduction. Suggested reading chapters are Chap. 1.1–1.3 of [1] and Chap 2.1–2.3 of [3].

For information theory, a good reference is [4, Chapter 2.1-2.5, 2.8-2.9]. A modern treatment is [2].

Local-scale family. A typical example of local-scale family is the multivariate normal distribution $\mathcal{N}_p(\mu_p, \Sigma_{p \times p})$, for $\mu_p \in \mathbb{R}^p$ and $\Sigma_{p \times p} \in \mathbb{R}^{p \times p}$ being positive definite.

5.1 Data Processing Inequality

Suppose that X is independent of Z given Y . In other words, $X \rightarrow Y \rightarrow Z$ forms a Markov Chain. What are useful examples of conditional independence?

Example 13. Let $Z = \varphi(Y)$ for some Borel function φ . Note that

$$\mathbb{E}[f(X)g(Z)|Y] = \mathbb{E}[f(X)|Y]g(Z) = \mathbb{E}[f(X)|Y]\mathbb{E}[g(Z)|Y].$$

Example 14. Let $Z = \varphi(Y, \xi)$, where ξ is independent of (X, Y) .

The proofs for the two examples are left in the assignment. (HW 3)

Theorem 3 (Data processing inequality). If $X \rightarrow Y \rightarrow Z$ forms a Markov Chain, then

$$I(X; Z) \leq I(X; Y),$$

where the equality holds if and only if X is independent of Y given Z .

Interpretation: X is the quantity of interest, Y is the observation or data, and Z is the (random) transformation of data, which won't increase the information about X .

Remark 2. *Mutual information is a fundamental quantity in noisy communication channel.*

The proof of data processing inequality is based on the chain rule for mutual information.

$$I(X; Z) + I(X; Y|Z) = I(X; Y, Z) = I(X; Y) + I(X; Z|Y)$$

By the Markov property (conditional independence), we have $I(X; Z|Y) = 0$, so

$$I(X; Y) = I(X; Z) + I(X; Y|Z) \geq I(X; Z).$$

References

- [1] Peter J Bickel and Kjell A Doksum. *Mathematical statistics: basic ideas and selected topics, volumes I-II package*. Chapman and Hall/CRC, 2015.
- [2] Yury Polyanskiy and Yihong Wu. *Information Theory: From Coding to Learning*. Cambridge University Press, 2024.
- [3] Jun Shao. *Mathematical Statistics*. Springer Science & Business Media, 2nd edition, 2013.
- [4] MTCAJ Thomas and A Thomas Joy. *Elements of information theory*. Wiley-Interscience, 2006.