STA355F: Theory of Statistical Practice **Fall 2026**

Lecture 7: November 6, 2026

Lecturer: Wenlong Mou

1 Sufficient statistics

Idea: in some statistical problems, we receive a large amount of data, but we do not need to store all of them, and we can compress the data into a smaller representation without losing information about the unknown parameter.

There are certain cases where compression does not lose information about the unknown parameter. This is captured by the notion of sufficient statistics.

Note: it is always relative to some parameter of interest (in a statistical model).

Definition 7.1 (Sufficient statistic)**.** *Let $(\mathbb{P}_\theta : \theta \in \Theta)$ be a statistical model, and X be sampled from $\mathbb{P}_\theta$. We call a statistic $T(X)$ sufficient for θ if for any $\theta \in \Theta$ and t in the range of T, the conditional distribution of X given $T(X) = t$ does not depend on θ.*

Intuition: if $T(X)$ is sufficient for θ, then we can discard the original data X and only keep $T(X)$. When you need to compute things using the original X, you can just sample $\widetilde{X}$ independently from the conditional distribution of X given $T(X)$, which does not depend on θ, and then use $\widetilde{X}$ to compute things. By the definition, we have $X \overset{d}{=} \widetilde{X}$ under $\mathbb{P}_\theta$.

Example: Let $X_1, X_2, \cdots, X_n \overset{\text{i.i.d.}}{\sim} \text{Ber}(p)$, where $p \in (0, 1)$ is unknown. We can let

$$T(X_1, X_2, \cdots, X_n) = \sum_{i=1}^{n} X_i \sim \text{Binom}(n, p).$$

Conditionally on $T(X_1, X_2, \cdots, X_n) = t$, the conditional distribution of $(X_1, X_2, \cdots, X_n)$ is uniform over all binary vectors with exactly t ones. This conditional distribution does not depend on p, so $T(X)$ is sufficient for p.

Factorization theorem: Checking by definition is not very convenient for sufficient statistics. But there is a more convenient characterization.

Theorem 7.1 (Factorization theorem)**.** *Suppose that the statistical model $(\mathbb{P}_\theta : \theta \in \Theta)$ has a density $p_\theta(x)$ with respect to some base measure (either pdf or pmf). Then $T(X)$ is sufficient for θ if and only if there exist functions $g_\theta(t)$ and $h(x)$ such that*

$$p_\theta(x) = g_\theta(T(x)) \cdot h(x), \quad \textit{for all } x \in \mathcal{X}, \theta \in \Theta.$$

Intuitively, $g_\theta(T(x))$ captures the interaction between θ and X (but only through $T(X)$), and $h(x)$ captures the part that does not depend on θ.

Proof idea:

"⇐": Suppose the conditional density of X given $T(X) = t$ exists, and we can write it as

$$p_\theta(x \mid T(x) = t) = \frac{g_\theta(t) \cdot h(x)\mathbf{1}_{T(x)=t}}{\int g_\theta(t) \cdot h(z)\mathbf{1}_{T(z)=t} dz} = \frac{h(x)\mathbf{1}_{T(x)=t}}{\int h(z)\mathbf{1}_{T(z)=t} dz},$$

which does not depend on θ.

"⇒": Suppose $T(X)$ is sufficient for θ. Assuming that the density of $T(X)$ exists, we can construct

$$g_\theta(t) = p_\theta(T(X) = t), \qquad h(x) = p_{\theta_0}(X = x \mid T(X) = T(x)),$$

for some fixed $\theta_0 \in \Theta$. Since the conditional density does not depend on θ, we have

$$g_\theta(t) \cdot h(x) = p_\theta(T(X) = t) \cdot p_{\theta_0}(X = x \mid T(X) = T(x))$$

Examples. Let $X_1, X_2, \cdots, X_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, 1)$, where μ is unknown. We can write the density as

$$\begin{aligned} p_\mu(X_1, \cdots, X_n) &= \frac{1}{(2\pi)^{n/2}} \exp\Big(-\frac{1}{2}\sum_{i=1}^n (X_i - \mu)^2 \Big) \\ &= \underbrace{\frac{1}{(2\pi)^{n/2}} \exp\Big(\mu \cdot \sum_{i=1}^n X_i - n\mu^2/2\Big)}_{g_\mu(T(X))} \cdot \underbrace{\exp\Big(-\frac{1}{2}\sum_{i=1}^n X_i^2\Big)}_{h(X)}, \end{aligned}$$

where the sufficient statistic is $T(X) = \sum_{i=1}^n X_i$.

On the other hand, we can also consider $X_1, X_2, \cdots, X_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2)$, where both μ and σ^2 are unknown. We can write the density as

$$\begin{aligned} p_{\mu,\sigma^2}(X_1, \cdots, X_n) &= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\Big(-\frac{1}{2\sigma^2}\sum_{i=1}^n (X_i - \mu)^2 \Big) \\ &= \underbrace{\frac{1}{(2\pi\sigma^2)^{n/2}} \exp\Big(-\frac{1}{2\sigma^2}\sum_{i=1}^n X_i^2 + \frac{\mu}{\sigma^2}\sum_{i=1}^n X_i - \frac{n\mu^2}{2\sigma^2}\Big)}_{g_{\mu,\sigma^2}(T(X))} \cdot 1, \end{aligned}$$

where the sufficient statistic is $T(X) = (\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2)$.

When is sufficient statistics useful? From the motivating example, we have seen a use case: we can store $T(X)$ only, and simulate $\widetilde{X}$ from the conditional distribution of X given $T(X)$, which does not depend on θ. This allows us to mimic any statistical procedure that uses the original data X. But sometimes, we can do even better: Rao–Blackwellization.

Rao–Blackwellization procedure: Given an estimator $\widehat{g}$ of $g(\theta)$ based on X, and let $T(X)$ be a sufficient statistic for θ. We can construct a new estimator

$$\widehat{g}_{\text{RB}} := \mathbb{E}\big[\widehat{g} \mid T(X)\big].$$

Since the conditional distribution of X given $T(X)$ does not depend on θ, this conditional expectation is computable without knowing θ. So $\widehat{g}_{\text{RB}}$ is a valid estimator of $g(\theta)$.

Key properties:

- $\mathbb{E}_\theta[\widehat{g}_{\mathrm{RB}}] = \mathbb{E}_\theta[\widehat{g}]$ for all $\theta \in \Theta$, so $\widehat{g}_{\mathrm{RB}}$ is unbiased if $\widehat{g}$ is unbiased.
- $\mathrm{var}_\theta(\widehat{g}_{\mathrm{RB}}) \le \mathrm{var}_\theta(\widehat{g})$ for all $\theta \in \Theta$, so $\widehat{g}_{\mathrm{RB}}$ is at least as good as $\widehat{g}$ in terms of variance.

Exponential family and sufficient statistics.

Definition 7.2 (Exponential family)**.** *A parametric family of distributions* $(\mathbb{P}_\theta : \theta \in \Theta)$ *is called an exponential family if the density can be written as*

$$p_\theta(x) = h(x) \exp\Big(\eta(\theta)^\top T(x) - B(\theta)\Big),$$

where $T(x)$ *is called the sufficient statistic,* $\eta(\theta)$ *is called the natural parameter, and* $B(\theta)$ *is called the log-partition function.*

From the definition, it is clear that $T(X)$ is a sufficient statistic for θ in an exponential family.

Let $X_1, X_2, \cdots, X_n \overset{\text{i.i.d.}}{\sim} \mathbb{P}_\theta$, where $\mathbb{P}_\theta$ is an exponential family. Then the joint density can be written as

$$p_\theta^{(n)}(x_1, x_2, \ldots, x_n) = \prod_{i=1}^n p_\theta(x_i) = \prod_{i=1}^n h(x_i) \cdot \exp\Big(\eta(\theta)^\top \sum_{i=1}^n T(x_i) - nB(\theta)\Big),$$

which tells us that the sufficient statistic for θ based on n i.i.d. samples is

$$T(X_1, X_2, \ldots, X_n) = \sum_{i=1}^n T(X_i).$$

Example of exponential family: Gaussian distribution, Bernoulli distribution, Poisson distribution.

Application of the Rao–Blackwellization procedure: let $X_1, X_2, \cdots, X_n \overset{\text{i.i.d.}}{\sim} \mathrm{Ber}(p)$, and we want to estimate $g(p) = p^2$. We can of course do plug-in estimator $\widehat{g} = \big(\frac{1}{n}\sum_{i=1}^n X_i\big)^2$, which is biased. Alternatively, we can also try to construct an unbiased estimator of $g(p)$. One approach is by Rao–Blackwellization. We consider

$$\widehat{g} = X_1 X_2.$$

Starting from this estimator, we can construct a new estimator

$$\widehat{g}_{\mathrm{RB}} = \mathbb{E}[X_1 X_2 \mid T(X)] = \mathbb{E}[X_1 X_2 \mid \sum_{i=1}^n X_i] = \frac{T(X)(T(X)-1)}{n(n-1)}.$$

In general, $\widehat{g}_{\mathrm{RB}}$ and $(T(X)/n)^2$ may not be comparable. However, it is known that for exponential families, under certain regularity conditions, the Rao–Blackwellized estimator is the unique minimum variance unbiased estimator (UMVUE).

2 Hypothesis testing

Binary decision problem: decide whether $\theta \in \Theta_0$ or $\theta \in \Theta_1$, based on the observed data X, where Θ_0 and Θ_1 are disjoint subsets of Θ.

Key concepts: Type I error, Type II error, power function, size/level of a test.

Goal of testing: constrained multi-objective optimization problem.

$$\text{maximize}_\phi \Big\{ \beta_\phi(\theta) \text{ for } \theta \in \Theta_1 \Big\}, \quad \text{subject to } \beta_\phi(\theta) \le \alpha, \text{ for } \theta \in \Theta_0,$$

where ϕ is the test function, and $\beta_\phi(\theta) = \mathbb{E}_\theta[\phi(X)]$ is the probability of rejection.

Sometimes, it is not easy to construct an exact valid test, and we may need to use asymptotic theory to construct an approximate test. We say a test is asymptotically valid if

$$\limsup_{n\to\infty} \sup_{\theta\in\Theta_0} \beta_\phi(\theta) \le \alpha.$$

2.1 One-dimensional testing

We can consider one-sided and two-sided testing.

- Two-sided testing: $H_0 : \theta = \theta_0$ vs. $H_1 : \theta \neq \theta_0$.
- One-sided testing: $H_0 : \theta \le \theta_0$ vs. $H_1 : \theta > \theta_0$, or $H_0 : \theta \ge \theta_0$ vs. $H_1 : \theta < \theta_0$.

For one-sided testing, let us consider the case $H_0 : \theta \le \theta_0$ vs. $H_1 : \theta > \theta_0$. Usually we compute some test statistics $T(X)$, and reject H_0 when $T(X)$ is large.

Examples.

- Let $X_1, X_2, \cdots, X_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\theta, 1)$, and we want to test $H_0 : \theta \le 0$ vs. $H_1 : \theta > 0$. We can use the test statistic $T(X) = \sqrt{n}\bar{X}$, and reject H_0 when $T(X) > z_{1-\alpha}$, where $z_{1-\alpha}$ is the $(1-\alpha)$-quantile of standard normal distribution.
- Let $X_1, X_2, \cdots, X_n \overset{\text{i.i.d.}}{\sim} \text{Ber}(p)$, and we want to test $H_0 : p \le 1/2$ vs. $H_1 : p > 1/2$. We can use the test statistic $T(X) = \sum_{i=1}^n X_i$. But here the rejection threshold needs a bit more careful treatment, since the distribution of $T(X)$ is discrete. We may not be able to find a threshold c such that $\mathbb{P}_{p=1/2}(T(X) > c) = \alpha$. In such a case, we can use randomized test to achieve exact level α. In particular, we find c such that $\mathbb{P}_{p=1/2}(T(X) > c) \le \alpha$ and $\mathbb{P}_{p=1/2}(T(X) \ge c) \ge \alpha$. We reject H_0 when $T(X) > c$, and reject H_0 with probability $\gamma \in [0, 1]$ when $T(X) = c$, such that

$$\mathbb{P}_{p=1/2}(T(X) > c) + \gamma \cdot \mathbb{P}_{p=1/2}(T(X) = c) = \alpha.$$

Similarly, for two-sided testing, we use a test statistic $T(X)$, and reject H_0 when $T(X)$ is too large or too small. For continuous $T(X)$, we can take

$$\phi(X) = \begin{cases} 1, & \text{if } T(X) > c_1 \text{ or } T(X) < c_2, \\ 0, & \text{otherwise,} \end{cases}$$

For discrete $T(X)$, we also use randomized test to achieve exact level α.

$$\phi(X) = \begin{cases} 1, & \text{if } T(X) > c_1 \text{ or } T(X) < c_2, \\ \gamma_1, & \text{if } T(X) = c_1, \\ \gamma_2, & \text{if } T(X) = c_2, \\ 0, & \text{otherwise.} \end{cases}$$

In general, we have one degree of freedom to choose the threshold c_1 and c_2 (and γ_1, γ_2 for discrete case) to achieve exact level α. You can make the test more powerful on one side by choosing c_1 and c_2 asymmetrically. If you balance the power on both sides, this corresponds to the "unbiased" two-sided test, which is the most common choice.

2.2 Popular choices of methodologies

Wald test: Consider the two-sided testing problem $H_0 : \theta = \theta_0$ vs. $H_1 : \theta \neq \theta_0$. Suppose we have an estimator $\widehat{\theta}_n$ of θ, and we know that under $\mathbb{P}_{\theta_0}$, we have

$$\sqrt{n}(\widehat{\theta}_n - \theta_0) \xrightarrow{d} \mathcal{N}(0, \sigma^2), \quad \text{as } n \to \infty.$$

Let $\widehat{\text{se}}$ be a consistent estimator of σ. Then we can use the test statistic

$$T(X) = \frac{\sqrt{n}(\widehat{\theta}_n - \theta_0)}{\widehat{\text{se}}},$$

and we reject H_0 when $|T(X)| > z_{1-\alpha/2}$, where $z_{1-\alpha/2}$ is the $(1-\alpha/2)$-quantile of standard normal distribution. From previous lectures, we know that $T(X) \xrightarrow{d} \mathcal{N}(0,1)$ under $\mathbb{P}_{\theta_0}$, so this test is asymptotically valid.

We can also do similar construction for one-sided testing.

χ^2 tests: Suppose that $\theta \in \mathbb{R}^d$, and we want to test

$$H_0 : \theta = \theta_0 \quad \text{vs.} \quad H_1 : \theta \neq \theta_0.$$

Similar to Wald test, we start with an estimator $\widehat{\theta}_n$ of θ, which is known to satisfy

$$\sqrt{n}(\widehat{\Sigma}_n)^{-1/2}(\widehat{\theta}_n - \theta_0) \xrightarrow{d} \mathcal{N}(0, I_d), \quad \text{as } n \to \infty,$$

under $\mathbb{P}_{\theta_0}$, where $\widehat{\Sigma}_n$ is a consistent estimator of the asymptotic covariance matrix of $\widehat{\theta}_n$. Then we can use the test statistic

$$T(X) = n(\widehat{\theta}_n - \theta_0)^\top \widehat{\Sigma}_n^{-1}(\widehat{\theta}_n - \theta_0) = \Big(\Big\| \sqrt{n}(\widehat{\Sigma}_n)^{-1/2}(\widehat{\theta}_n - \theta_0) \Big\|_2 \Big)^2$$

Under $\mathbb{P}_{\theta_0}$, we have

$$T(X) \xrightarrow{d} \chi^2_d, \quad \text{as } n \to \infty,$$

where χ^2_d is the chi-squared distribution with d degrees of freedom, i.e., the distribution of the sum of squares of d independent standard normal random variables. And the χ^2 test rejects H_0 when $T(X)$ is larger than the $(1-\alpha)$-quantile of χ^2_d distribution. The χ^2 test is asymptotically valid.

Remark: you can also use the variance computed exactly under $\mathbb{P}_{\theta_0}$ instead of the estimated variance. The resulting test is still asymptotically valid.

Special case of Wald's test: Let $X_1, X_2, \cdots, X_n \overset{\text{i.i.d.}}{\sim} \text{Cat}(p_1, p_2, \cdots, p_d)$, where $p_i \geq 0$ and $\sum_{i=1}^d p_i = 1$. We want to test

$$H_0 : p = p^0 \quad \text{vs.} \quad H_1 : p \neq p^0.$$

We take the estimator

$$\widehat{p}_n(j) = \frac{1}{n} \sum_{i=1}^n \mathbf{1}_{X_i = e_j}, \quad j = 1, 2, \cdots, d.$$

It is known that under $\mathbb{P}_{p^0}$, $\sqrt{n}\big(\widehat{p}_n - p^0\big) \xrightarrow{d} \mathcal{N}(0, \Sigma_0)$, where Σ_0 is a degenerate covariance matrix with rank $d-1$. We can use a change-of-variable trick to reduce the dimension by 1 (e.g., replacing p_d by $1 - \sum_{j=1}^{d-1} p_j$), and then we can apply the χ^2 test to the reduced-dimension problem. After some algebra, the test statistic can be simplified to

$$T(X) = n \sum_{j=1}^{d} \frac{(\widehat{p}_n(j) - p^0(j))^2}{p^0(j)} \overset{d}{\Rightarrow} \chi^2_{d-1}, \quad \text{as } n \to \infty.$$

This is called Pearson's χ^2 test, which is a classical test for categorical data.

Student t-test: Here we deal with exactly Gaussian data, but we do not want to take the asymptotic approximation. Let $X_1, X_2, \cdots, X_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2)$, where both μ and σ^2 are unknown. We want to test

$$H_0 : \mu = \mu_0 \quad \text{vs.} \quad H_1 : \mu \neq \mu_0.$$

A natural test statistic is

$$T = \frac{\sqrt{n}(\overline{X} - \mu_0)}{\widehat{se}_n},$$

where $\widehat{se}_n = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2}$ is the sample standard deviation. We know that as $n \to \infty$, $T \xrightarrow{d} \mathcal{N}(0,1)$ under H_0. Student computed the exact distribution of T under H_0, which is known as Student's t-distribution with $n-1$ degrees of freedom. The t-test rejects H_0 when $|T| > t_{1-\alpha/2, n-1}$, where $t_{1-\alpha/2,n-1}$ is the $(1-\alpha/2)$-quantile of Student's t-distribution with $n-1$ degrees of freedom. The t-test is an exact test, and it is also asymptotically valid.

Here, the definition of t-distribution is as follows: let $Z \sim \mathcal{N}(0,1)$ and $Y \sim \chi^2_\nu$ be independent, then we say that $T = Z/\sqrt{Y/\nu}$ follows Student's t-distribution with ν degrees of freedom.

2.3 p-values

Definition 7.3. *Suppose that $T(X)$ is the statistic, and we reject H_0 when $T(X) \in R_\alpha$, where R_α is the rejection region for level α. We define the p-value as*

$$\inf\{\alpha : T(X) \in R_\alpha\}.$$

Note that p-value is a random variable, and it is a function of the observed data X. Intuitively, the p-value is the smallest level α at which we can reject H_0 based on the observed data.

If we can compute the p-value, then the test can be equivalently described as: reject H_0 if p-value $\leq \alpha$.

Example: consider a one-dimensional two-sided Wald test, where we use

$$T(X) = \frac{\sqrt{n}(\widehat{\theta}_n - \theta_0)}{\widehat{\text{se}}}.$$

Rejection region is

$$R_\alpha = \{x : |T(x)| > z_{1-\alpha/2}\}.$$

Then the p-value is

$$\inf\{\alpha : |T(X)| > z_{1-\alpha/2}\} = 2(1 - \Phi(|T(X)|)),$$

where Φ is the CDF of standard normal distribution.