STA 522, Spring 2022
Introduction to Theoretical Statistics II

Lecture 4

Department of Biostatistics
University at Buffalo

AGENDA

- Comments on Exam 1
- Ancillary Statistics
- Complete Statistics
- Likelihood Principle and Equivariance

Review: Factorization Theorem, Minimal Sufficiency

- **Factorization Theorem:** Let $f(\underline{x} \mid \theta)$ denote the joint pdf or pmf of a sample $\underline{X}$. A statistic $T(\underline{x})$ is a sufficient statistic for θ *if and only if* there exist functions $g(t \mid \theta)$ and $h(\underline{x})$ such that, for all sample points $\underline{x}$ and all parameter points θ,

$$f(\underline{x} \mid \theta) = g(T(\underline{x}) \mid \theta) \cdot h(\underline{x}).$$

- **Finding minimal sufficient statistics:** Let $f(\underline{x} \mid \theta)$ be the pdf/pmf of a sample $\underline{X}$. Suppose there exists a function $T(\underline{X})$ such that, for every $\underline{x}$ and $\underline{y}$, the ratio

$$\frac{f(\underline{x} \mid \theta)}{f(\underline{y} \mid \theta)}$$

  is constant as a function of θ if and only if $T(\underline{x}) = T(\underline{y})$. Then $T(\underline{X})$ is a minimal sufficient statistic for θ.

Ancillary Statistics

Definition: A statistic $S(\underline{X})$ whose distribution does not depend on the parameter θ is called an **ancillary statistic**.

- Alone, an ancillary statistic contains no information on the parameter θ.
- However, when used in conjunction with other statistics, ancillary statistics sometimes do provide valuable information for inferences about θ.

Example (Uniform Ancillary Statistic): Let $X_1, X_2, \ldots, X_n$ be iid Uniform $\left(\theta - \frac{1}{2}, \theta + \frac{1}{2}\right)$. We saw that $\left(X_{(1)}, X_{(n)}\right)$ is minimal sufficient for θ. We will show that $R = X_{(n)} - X_{(1)}$ is ancillary for θ by showing that the pdf of R doesn't depend on θ.

To this end we first find the joint pdf of $(X_{(1)}, X_{(n)})$. Note that the common pdf and cdf for each X_i is:

$$f(x \mid \theta) = I\left(\theta - \frac{1}{2} < x < \theta + \frac{1}{2}\right), \text{ and}$$

$$F(x \mid \theta) = \begin{cases} 0 & x \leq \theta - \frac{1}{2} \\ x - \theta + \frac{1}{2} & \theta - \frac{1}{2} < x < \theta + \frac{1}{2} \\ 1 & x \geq \theta + \frac{1}{2} \end{cases}$$

From a result discussed in class, we have:

$$\begin{aligned} f_{X_{(1)}, X_{(n)}}(x_1, x_n) = \frac{n!}{(1-1)!(n-1-1)!(n-n)!}\ & f(x_1 \mid \theta)\ f(x_n \mid \theta) \\ & \times \left[F(x_1 \mid \theta)\right]^{1-1}\ \left[F(x_n \mid \theta) - F(x_1 \mid \theta)\right]^{n-1-1} \\ & \times \left[1 - F(x_n \mid \theta)\right]^{n-n} \end{aligned}$$

So, here

$$f_{X_{(1)},X_{(n)}}(x_1, x_n \mid \theta) = n(n-1)\ (x_n - x_1)^{n-2}\ I\left(\theta - \frac{1}{2} < x_1 < x_n < \theta + \frac{1}{2}\right)$$

To find the distribution of $R = X_{(n)} - X_{(1)}$, consider $M = (X_{(1)} + X_{(n)})/2$. The inverse transformation is $X_{(1)} = (2M - R)/2$ and $X_{(n)} = (2M + R)/2$ with Jacobian of transformation being 1.

Joint support for (R, M):
$\left\{(r, s) : 0 < r < 1, \theta - \frac{1}{2} + \frac{r}{2} < m < \theta + \frac{1}{2} - \frac{r}{2}\right\}$. Therefore

$$\begin{aligned} f_R(r \mid \theta) &= \int_{\theta - \frac{1}{2} + \frac{r}{2}}^{\theta + \frac{1}{2} - \frac{r}{2}} n(n-1)\ r^{n-2}\ dm \\ &= n(n-1)\ r^{n-2}\ (1-r); \qquad 0 < r < 1 \end{aligned}$$

This is the pdf of Beta($\alpha = n - 1$, $\beta = 2$) distribution, and $f_R(r \mid \theta)$ is the same for all θ.

Thus the distribution of R does not depend on $\theta \implies R$ is ancillary for θ.

Example (Location Family Ancillary Statistic): Let $X_1, X_2, \ldots, X_n$ be iid random variables from a location parameter family with cdf $F(x-\theta)$ for some $\theta \in \mathbb{R}$. Then $R = X_{(n)} - X_{(1)}$ is an ancillary statistic.

To see this, define $Z_i = X_i - \theta$. Then Z_i are iid with common cdf $F(x)$. The cdf of the range statistic R is:

$$\begin{aligned}
F_R(r \mid \theta) &= P_\theta(R \le r) \\
&= P_\theta\left(\max_i X_i - \min_i X_i \le r\right) \\
&= P_\theta\left(\max_i (Z_i + \theta) - \min_i (Z_i + \theta) \le r\right) \\
&= P_\theta\left(\max_i Z_i - \min_i Z_i \le r\right)
\end{aligned}$$

The last quantity does not depend on θ as the common cdf of $Z_1, \ldots, Z_n$ does not depend on θ.

Hence, R is ancillary for θ.

Example (Scale family ancillary statistic) Let $X_1, X_2, \dots, X_n$ be iid random variables from a scale parameter family with cdf $F(x/\sigma)$ for some $\sigma > 0$. Then any statistic that depends on the sample only through the $n-1$ values $X_1/X_n, \dots, X_{n-1}/X_n$ is an ancillary statistic.

For example $\dfrac{X_1 + \cdots + X_n}{X_n} = \dfrac{X_1}{X_n} + \cdots + \dfrac{X_{n-1}}{X_n} + 1$ is an ancillary statistic.

To see this, define $Z_i = X_i/\sigma$ so that Z_i are iid from $F(x)$ (free of σ).

$$\begin{aligned}
F_{X_1/X_n,\dots,X_{n-1}/X_n}(y_1, \dots, y_{n-1} \mid \sigma) &= P_\sigma\left(X_1/X_n \le y_1, \dots, X_{n-1}/X_n \le y_{n-1}\right) \\
&= P_\sigma\left(Z_1/Z_n \le y_1, \dots, Z_{n-1}/Z_n \le y_{n-1}\right)
\end{aligned}$$

which does not depend on σ.

Ancillary and Minimal Sufficient Statistics

- Ancillary statistic by itself does not contain any information on θ while a minimal sufficient statistic contains all information. So it may seem that ancillary and minimal sufficient statistics should be unrelated, or statistically independent. This is not necessarily the case.

- Consider the Uniform$(\theta - \frac{1}{2}, \theta + \frac{1}{2})$ example. Here $\big(R = X_{(n)} - X_{(1)}, M = (X_{(1)} + X_{(n)})/2\big)$ is minimal sufficient for θ, but R is ancillary.

Completeness

Definition: Let $f(t \mid \theta)$ be a family of pdfs or pmfs for a statistic $T(\underline{X})$. The family of probability distributions is called **complete** if

$$\mathrm{E}_\theta(g(T)) = 0 \text{ for all } \theta \implies P_\theta(g(T) = 0) = 1 \text{ for all } \theta$$

In this case, $T(\underline{X})$ is called a **complete statistic**.

Notes

- completeness is a property of a family of distributions, not of a particular distribution. For example, if $X \sim \mathrm{N}(0, 1)$ then defining $g(x) = x$, we have that $E[g(X)] = E(X) = 0$, but $g(x)$ satisfies $P(g(X) = 0) = P(X = 0) = 0$ and not 1. However, this a particular distribution and not a family of distributions.
- Instead, if we consider $X \sim \mathrm{N}(\theta, 1)$, $-\infty < \theta < \infty$, then we will see that no function of X, except one that is 0 with probability 1 for all θ, satisfies $E_\theta(g(X)) = 0$ for all θ.

Example (Binomial Complete Sufficient Statistic): Suppose that $T \sim \text{Binomial}(n, p)$, where $0 < p < 1$ is an unknown parameter and n is a fixed integer. Then T is complete.

To see this, let g be a function such that $E_p(g(T)) = 0$. Then

$$
\begin{aligned}
0 = E_p(g(T)) &= \sum_{t=0}^{n} g(t) \binom{n}{t} p^t (1-p)^{n-t} \\
&= (1-p)^n \sum_{t=0}^{n} g(t) \binom{n}{t} \left(\frac{p}{1-p}\right)^t \text{ for all } p \in (0,1) \\
\implies 0 &= \sum_{t=0}^{n} g(t) \binom{n}{t} \left(\frac{p}{1-p}\right)^t \text{ for all } p \in (0,1) \\
\implies 0 &= \sum_{t=0}^{n} g(t) \binom{n}{t} r^t \text{ for all } r \in (0,\infty)
\end{aligned}
$$

The last expression is a polynomial of degree n in $r \in (0, \infty)$. For the polynomial to be 0 for all r, each coefficient has to be 0, implying $g(t) = 0$ for $t = 0, \ldots, n \implies P_p(g(T) = 0) = 1$ for all p.

Example (Uniform Complete Sufficient Statistics): Suppose $X_1, X_2, \ldots, X_n$ iid Uniform$(0, \theta)$. It follows that $T(\underline{X}) = \max_i X_i$ is a sufficient statistic. We shall show that T is also complete.

Using results on order statistics, the pdf of T is obtained as

$$f(t \mid \theta) = \frac{n}{\theta^n}\ t^{n-1}\ I(0 < t < \theta)$$

Suppose $g(t)$ is a function satisfying $E_\theta(g(T)) = 0$, for all $0 < \theta < \infty$. Since $E_\theta(g(T))$ is constant in θ,

$$\begin{aligned} 0 = \frac{d}{d\theta} E_\theta(g(T)) &= \frac{d}{d\theta} \int_0^\theta g(t)\ \frac{n}{\theta^n}\ t^{n-1}\ dt \\ &= \frac{n}{\theta^n}\ \frac{d}{d\theta}\left(\int_0^\theta g(t)\ t^{n-1}\ dt\right) + \left(\frac{d}{d\theta}\frac{n}{\theta^n}\right)\int_0^\theta g(t)\ t^{n-1}\ dt \\ &= \frac{n}{\theta^n}\ g(\theta)\ \theta^{n-1} + 0\ = \frac{n}{\theta}\ g(\theta), \quad \text{for all } \theta \end{aligned}$$

Since $\frac{n}{\theta} \neq 0$, we must have $g(\theta) = 0$ for all $\theta > 0$. Hence T is complete.

Complete Statistics in the Exponential Family

Theorem 6.2.25

Let $X_1, X_2, \ldots, X_n$ be iid observations from an exponential family with pdf or pmf of the form $f(x \mid \boldsymbol{\theta}) = h(x)c(\boldsymbol{\theta}) \exp\left(\sum_{j=1}^{k} w_j(\boldsymbol{\theta}) t_j(x)\right)$, where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$. Then $T(\underline{X}) = \left(\sum_{i=1}^{n} t_1(X_i), \ldots, \sum_{i=1}^{n} t_k(X_i)\right)$ is a complete statistic for $\boldsymbol{\theta}$ as long as the parameter space Θ contains an open set in $\mathbb{R}^k$.

NOTES

(a) The dimensions of $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$ and $\underline{w} = (w_1(\boldsymbol{\theta}), w_2(\boldsymbol{\theta}), \ldots, w_k(\boldsymbol{\theta}))$ must be the same.

(b) The parameter space Θ doesn't need to be an open set, it just needs to contain an open set in $\mathbb{R}^k$. Note that this is not possible if the entries in θ are functionally related (i.e., lies on a lower hyper-plane). Example includes the $\mathrm{N}(\theta, \theta^2)$ family.

Example (Exercise 6.15) Suppose $X_1, X_2, \ldots, X_n \sim$ iid $\mathrm{N}(\theta, \theta^2)$. We argued that the previous result on exponential family cannot be used here. Is this family complete?

Consider the sufficient statistic $T(\underline{X}) = (\overline{X}, S^2)$. Note that

$$\mathrm{E}_\theta \left(\overline{X}^2\right) = \mathrm{Var}_\theta \left(\overline{X}\right) + \left(\mathrm{E}_\theta(\overline{X})\right)^2 = \theta^2/n + \theta^2 = \frac{n+1}{n}\theta^2$$

and

$$E_\theta \left(S^2\right) = \theta^2$$

.

Therefore if we define

$$g(T(\underline{X})) = \frac{n}{n+1}\overline{X}^2 - S^2$$

Then $\mathrm{E}_\theta \left(g(T(\underline{X})\right) = 0$ for all θ but $P_\theta \left(g(T(\underline{X}) = 0\right) = 0$.

Hence, the family is not complete.

Basu's Theorem

Theorem 6.2.24

If $T(\underline{X})$ is a complete and minimal sufficient statistic, then $T(\underline{X})$ is independent of every ancillary statistic.

Proof: (Only for discrete distributions.) Let $S(\underline{X})$ be any ancillary statistic for the parameter θ. Then $P(S(\underline{X}) = s)$ does not depend on θ.

Again, since $T(\underline{X})$ is sufficient,
$P(S(\underline{X}) = s \mid T(\underline{X}) = t) = P(\underline{X} \in \{\underline{x} : S(\underline{x}) = s\} \mid T(\underline{X}) = t)$ does not depend on θ.

So enough to show that

$$P(S(\underline{X}) = s \mid T(\underline{X}) = t) = P(S(\underline{X}) = s) \text{ for all } t \in \mathcal{T}.$$

We have

$$P(S(\underline{X}) = s) = \sum_{t \in \mathcal{T}} P(S(\underline{X}) = s \mid T(\underline{X}) = t)\ P_\theta(T(\underline{X}) = t)$$

and from $\sum_{t \in \mathcal{T}} P_\theta(T(\underline{X}) = t) = 1$,

$$P(S(\underline{X}) = s) = \sum_{t \in \mathcal{T}} P(S(\underline{X}) = s)\ P_\theta(T(\underline{X}) = t)$$

Define $g(t) = P(S(\underline{X}) = s \mid T(\underline{X}) = t) - P(S(\underline{X}) = s)$. Then from the above two equations we have

$$\begin{aligned} \mathrm{E}_\theta(g(T)) &= \sum_{t \in \mathcal{T}} g(t)\ P_\theta(T(\underline{X}) = t) \text{ for all } \theta \\ \implies g(t) &= 0 \text{ for all } t \in \mathcal{T} \qquad\qquad (T \text{ is complete}) \\ \implies P(S(\underline{X}) = s \mid T(\underline{X}) = t) &= P(S(\underline{X}) = s) \text{ for all } t \in \mathcal{T} \end{aligned}$$

Using Basu's Theorem

Example: Let $X_1, X_2, \ldots, X_n$ be iid exponential observations with parameter θ. Find the expected value of

$$g(\underline{X}) = \frac{X_n}{X_1 + \cdots + X_n}.$$

Note on the outset that exponential is a scale family, and so from a previous example it follows that $g(\underline{X})$ is ancillary for θ.

From the results on exponential family, it follows that $T(\underline{X}) = \sum_{i=1}^n X_i$ is complete and sufficient (verify).

Hence, by Basu's Theorem, $T(\underline{X})$ and $g(\underline{X})$ are independent, meaning $E_\theta[g(\underline{X})T(\underline{X})] = E[g(\underline{X})]E_\theta[T(\underline{X})]$. Here $E_\theta[g(\underline{X})T(\underline{X})] = E_\theta[X_n] = \theta$, and $E_\theta[T(\underline{X})] = E_\theta[\sum_{i=1}^n X_i] = n\theta$.

This implies $E[g(X)] = 1/n$.

Example: Let $X_1, X_2, \ldots, X_n$ be iid observations from $\mathrm{N}(\mu, \sigma^2)$ distribution. We can establish the independence between $\overline{X}$ and S^2 using Basu's theorem.

First fix σ^2 to some arbitrary value say σ_0^2.

Then $\overline{X}$ is a complete sufficient statistic for μ.

For any fixed σ_0^2, the family $\mathrm{N}(\mu, \sigma_0^2)$ is a location family with location parameter μ. It can be shown that (homework) S^2, a statistic based on $X_1 - \overline{X}, \ldots, X_n - \overline{X}$, is ancillary for μ.

So, for any fixed σ_0^2, $\overline{X}$ and S^2 are independent (Basu's theorem).

Since σ_0^2 is arbitrary, therefore, $\overline{X}$ and S^2 are independent for any (μ, σ^2)

The Likelihood Function

Definition: Let $f(\underline{x} \mid \theta)$ denote the joint pdf or pmf of the sample $\underline{X} = (X_1, X_2, \ldots, X_n)$. Then, given that $\underline{X} = \underline{x}$ is observed, the function of θ defined by

$$L(\theta \mid \underline{x}) = f(\underline{x} \mid \theta)$$

is called the **likelihood function**.

If $\underline{X}$ is a discrete random vector, then $L(\theta \mid \underline{x}) = P_\theta(\underline{X} = \underline{x})$. If for two parameter points θ_1 and θ_2,
$P_{\theta_1}(\underline{X} = \underline{x}) = L(\theta_1 \mid \underline{x}) > L(\theta_2 \mid \underline{x}) = P_{\theta_2}(\underline{X} = \underline{x})$, then $\underline{x}$ is more likely to have occurred if $\theta = \theta_1$ than if $\theta = \theta_2$.

Likelihood Principle: If $\underline{x}$ and $\underline{y}$ are two sample points such that $L(\theta \mid \underline{x})$ is proportional to $L(\theta \mid \underline{y})$, that is, there exists a constant $C(\underline{x}, \underline{y})$ such that $L(\theta \mid \underline{x}) = C(\underline{x}, \underline{y})\, L(\theta \mid \underline{y})$ for all θ, then the conclusions drawn from $\underline{x}$ and $\underline{y}$ should be identical.

Example (Negative Binomial Likelihood): Let X have a negative binomial distribution with $r = 3$ and success probability p. Find the likelihood function if $x = 2$ is observed, and also for general $X = x$.

If $x = 2$, the likelihood function for $0 \leq p \leq 1$ is

$$L(p \mid 2) = P_p(X = 2) = \binom{4}{2} \, p^3 \, (1 - p)^2$$

For general $X = x$ the likelihood function is:

$$L(p \mid x) = P_p(X = x) = \binom{3 + x - 1}{x} \, p^3 \, (1 - p)^x$$

Example (Poisson Likelihood): Let $\underline{X} = (X_1, X_2, \ldots, X_n)$ denote a random sample from a Poisson distribution with mean λ. The likelihood function for $0 < \lambda < \infty$ is given by:

$$L(\lambda \mid \underline{x}) = P_\lambda(\underline{X} = \underline{x}) = \exp(-n\lambda)\ \frac{\lambda^{\sum_{i=1}^n x_i}}{\prod_{i=1}^n x_i!}$$

Example (Normal Likelihood): Let $\underline{X} = (X_1, X_2, \ldots, X_n)$ denote a random sample from a $\mathrm{N}(\mu, \sigma^2)$ distribution. The likelihood function for $-\infty < \mu < \infty$ and $\sigma > 0$ is given by

$$L(\mu, \sigma \mid \underline{x}) = f(\underline{x} \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}}\ (\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2\right)$$

Equivariance

Equivariance Principle: If $\underline{Y} = g(\underline{X})$ is a change of measurement scale such that the model for $\underline{Y}$ has the same formal structure as the model for $\underline{X}$, then an inference procedure should be both measurement equivariant and formally equivariant.

Example (Binomial equivariance): Suppose $X \sim \text{Binomial}(n, p)$ and we want to "estimate" p using x, say using the statistic $T(x)$.

Now $X \sim \text{Binomial}(n, p) \implies Y = n - X \sim \text{Binomial}(n, q = 1 - p)$, so $T(y)$ should be an estimator for $q = 1 - p$.

Since $p + q = 1$, it is reasonable to ensure that their estimator also satisfies this relationship, i.e.,

$$T(x) + T(y) = 1 \implies T(x) = 1 - T(y) = 1 - T(n - x)$$

If we only consider estimator satisfying this relationship, then we get a greatly reduced and simplified set of estimators.

Homework

- Read p. 282 – 291.
- Exercises: TBA.