Introduction to the Theory of Statistics Part 2

PM522b

Meredith Franklin

Division of Biostatistics, University of Southern California

Slides 2, 2015

Topics covered

- Data Reduction
  1. Statistics
  2. The Sufficiency Principle
  3. The Likelihood Principle
  4. The Equivariance Principle

Statistic

- The random samples we generated previously are vectors of observations that can be interpreted in statistically meaningful ways
- We want to use the information contained in our random sample to arrive at conclusions regarding our population
- A statistic:
    - is a form of data reduction
    - can be thought of as a partition of the sample space
    - is a summary quantity of our random sample
    - is a function of the sample

A statistic is a form of data reduction

- Data reduction means that we use a statistic $T(\mathbf{x})$ instead of the entire sample $\mathbf{x} = (x_1, ..., x_n)$ to make inferences about an unknown parameter θ.

Partitioning the sample space

The sample space $\mathcal{X}$ can be partitioned and subsequently the observations $\mathbf{x}$ can be reduced.
Let $\mathcal{T} = \{t : t = T(\mathbf{x}) \text{ for some } \mathbf{x} \in \mathcal{X}\}$ be the image of $\mathcal{X}$ under $T(\mathbf{x})$.
Then, the statistic $T(\mathbf{x})$ partitions the sample space $\mathcal{X}$ into sets $A_t, t \in \mathcal{T}$ where $A_t = \{\mathbf{x} : T(\mathbf{x}) = t\}$ for $t \in \mathcal{T}$

- So, rather than reporting the whole sample $\mathbf{x}$ we use $T(\mathbf{x}) = t$
- Reporting $T(\mathbf{x}) = t$ is equivalent to reporting $\mathbf{x} \in A_t$

A statistic is a function of the sample

- A statistic is formally defined as a function of the observable random variables in a sample and known constants.
- Functions of observed samples (i.e. data) are used to generate statistics.

Definition

For an iid sequence of random variables $X_1, X_2, ..., X_n$ sampled from our population with distribution function $f(\mathbf{X}|\theta)$, the function $T(\mathbf{X}) = T(X_1, X_2, ..., X_n)$ which does not contain the unknown parameter θ is called a *statistic*.

A statistic is a function of the sample

- The statistic $T(\mathbf{X})$, a function of random variables, is itself a random variable.
- When $T(\mathbf{X})$ is used for inference, two different random samples $\mathbf{x}$ and $\mathbf{y}$ that satisfy $T(\mathbf{x}) = T(\mathbf{y})$ lead to the same inference.
- The most frequently used statistics are measures of central tendency and measures of concentration or variation of the random sample.

Simple Examples

Where $T(X) = T(X_1, X_2, ..., X_n)$

$$T(X) = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \text{ (sample mean)}$$

$$T(X) = S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 \text{ (sample variance)}$$

$$T(X) = M_n \text{ (sample median)}$$

Statistic

- A statistic can be basically anything, but the choice of what we use as a statistic depends on the problem at hand.
- Some important things to note:
  - T need not be a continuous function, but it does need to be measurable, i.e. the mapping $T : \mathcal{X} \to \mathcal{T}$ is measurable.
  - By saying it cannot depend on parameter θ, that means that the parameter θ cannot appear in the formula for T. However, it is ok if the distribution of T depends on θ.

Principles of Data Reduction

- We need to evaluate how good a statistic really is, and to do this we rely on the three principles of data reduction:
    - Sufficiency
    - Likelihood
    - Equivariance

Sufficiency

We assess our statistic for certain properties:

- Does the statistic retain all of the information about the true population parameters?
- Has some information about our parameters been lost or obscured through the process of reducing our data?
- A sufficient statistic for θ is one that captures all of the information about θ contained in our sample.
- This leads to the sufficiency principle:

Sufficiency Principle, CB 6.2

If $T(\mathbf{X})$ is a sufficient statistic for θ, then inference about θ should depend on the sample $\mathbf{X}$ only through the value of the statistic $T(\mathbf{X})$. If $\mathbf{x}$ and $\mathbf{y}$ are two sample points such that $T(\mathbf{x}) = T(\mathbf{y})$, the inference about θ should be the same whether $\mathbf{X} = \mathbf{x}$ or $\mathbf{Y} = \mathbf{y}$ is observed.

Basically, if we know the value of the sufficient statistic T we can do just as good of a job estimating θ as someone who knows the entire sample.

Sufficiency

Definition: Sufficient statistics

For a random sample $X_1, X_2, ..., X_n$ with pdf $f(\mathbf{x}|\theta)$, the statistic $T(\mathbf{X})$ is said to be sufficient if the conditional distribution of $X_1, X_2, ..., X_n$ given $T(\mathbf{X})$ does not depend on θ.

- A statistic $T(\mathbf{X})$ is sufficient for θ if inferences about θ depend on $\mathbf{X}$ only through $T(\mathbf{X})$. (Informal definition)
- A statistic $T(\mathbf{X})$ is sufficient for θ if the conditional distribution of $\mathbf{X}$ given $T(\mathbf{X})$ does not depend on θ. (Formal definition)

Sufficiency

Example

To illustrate sufficiency, we devise a scenario where we have two 522b students A and B. Student A knows the entire random sample $X_1, ..., X_n = \mathbf{x}$ and can compute the statistic $T(\mathbf{X}) = t(\mathbf{x})$. This student can make inference about the parameter θ using this information. On the other hand, student B only knows the value of the statistic $T(\mathbf{X}) = t(\mathbf{x})$. Since the conditional distribution of $X_1, ..., X_n$ given $T(\mathbf{X})$ does not depend on θ, student B knows $P(\mathbf{X} = \mathbf{y}|T(\mathbf{X}) = t(\mathbf{x}))$, which is a probability distribution on $A_{T(x)} = \{\mathbf{y} : T(\mathbf{y} = T(\mathbf{x})\}$ that can be calculated without knowledge of the true value of θ. So, student B can use this distribution to generate a radom sample $\mathbf{y}$ satisfying $P(\mathbf{Y} = \mathbf{y}|T(\mathbf{X}) = t(\mathbf{x})) = P(\mathbf{X} = \mathbf{y}|T(\mathbf{X}) = t(\mathbf{x}))$. This means that for each θ, $\mathbf{X}$ and $\mathbf{Y}$ have he same unconditional pdf (shown on next slide).

Student B knows just as much about θ via $T(\mathbf{X}) = t(\mathbf{x})$ as student A who knows the entire sample $\mathbf{X} = \mathbf{x}$.

Sufficiency

Example, con't

For this example to work, $\mathbf{X}$ and $\mathbf{Y}$ must have he same unconditional distribution, namely $P_\theta(\mathbf{X} = \mathbf{x}) = P_\theta(\mathbf{Y} = \mathbf{x})$ $\forall \mathbf{x}$ and θ.

$$
\begin{aligned}
P_\theta(\mathbf{X} = \mathbf{x}) &= P_\theta(\mathbf{X} = \mathbf{x} \text{ and } T(\mathbf{X}) = T(\mathbf{x})) \\
&= P(\mathbf{X} = \mathbf{x} | T(\mathbf{X}) = T(\mathbf{x})) P_\theta(T(\mathbf{X}) = T(\mathbf{x})) \\
&= P(\mathbf{Y} = \mathbf{x} | T(\mathbf{X}) = T(\mathbf{x})) P_\theta(T(\mathbf{X}) = T(\mathbf{x})) \\
&= P_\theta(\mathbf{Y} = \mathbf{x} \text{ and } T(\mathbf{X}) = T(\mathbf{x})) \\
&= P_\theta(\mathbf{Y} = \mathbf{x})
\end{aligned}
$$

Sufficiency

To verify that a statistic $T(\mathbf{X})$ is indeed sufficient for parameter θ, we must verify that for any fixed values of $\mathbf{x}$ and t, $P_\theta(\mathbf{X} = \mathbf{x} | T(\mathbf{X}) = T(\mathbf{x}))$.

Theorem, CB 6.2.2

$T(\mathbf{X})$ is sufficient for θ iff the ratio $p(\mathbf{x}|\theta)/q(T(\mathbf{x}|\theta))$ is independent of θ where $p(\mathbf{x}|\theta)$ and $q(T(\mathbf{x}|\theta))$ are the joint pmfs or pdfs of $\mathbf{X}$ and $T(\mathbf{X})$, respectively.

$$\begin{aligned} P_\theta(\mathbf{X} = \mathbf{x} | T(\mathbf{X}) = T(\mathbf{x})) &= \frac{P_\theta(\mathbf{X} = \mathbf{x} \text{ and } T(\mathbf{X}) = T(\mathbf{x}))}{P_\theta(T(\mathbf{X}) = T(\mathbf{x}))} \\ &= \frac{P_\theta(\mathbf{X} = \mathbf{x})}{P_\theta(T(\mathbf{X}) = T(\mathbf{x}))} \\ &= \frac{p(\mathbf{x}|\theta)}{q(T(\mathbf{x}|\theta))} \end{aligned}$$

Sufficiency

A couple of examples:

Sufficiency of sample mean for the normal distribution

Given $X_1, ..., X_n$ iid $N(\mu, \sigma^2)$ with σ^2 known, is the sample mean, $\bar{X} = (X_1, ..., X_n)/n$ a sufficient statistic for μ?
The joint pdf for the sample $\mathbf{X}$ is

$$\begin{aligned}
f_X(\mathbf{x}|\mu) &= \prod_{i=1}^{n} (2\pi\sigma^2)^{-n/2} \exp(-(x_i - \mu)^2/(2\sigma^2)) \\
&= (2\pi\sigma^2)^{-n/2} \exp(-\sum_{i=1}^{n} (x_i - \mu)^2/(2\sigma^2)) \\
&= (2\pi\sigma^2)^{-n/2} \exp(-\sum_{i=1}^{n} (x_i - \bar{x} + \bar{x} - \mu)^2/(2\sigma^2)) \text{ (add and subtract } \bar{x}) \\
&= (2\pi\sigma^2)^{-n/2} \exp(-\sum_{i=1}^{n} (x_i - \bar{x})^2 + n(\bar{x} - \mu)^2/(2\sigma^2))
\end{aligned}$$

Sufficiency

Sufficiency of sample mean for the normal distribution, con't

The joint pdf for the sample mean $\bar{X}$ which is iid $N(\mu, \sigma^2/n)$ is:

$$f_{\bar{X}}(\bar{\mathbf{x}}|\mu) = (2\pi\sigma^2)^{-1/2}\exp(-n(\bar{x}-\mu)^2/(2\sigma^2))$$

So the ratio $p(\mathbf{x}|\theta)/q(T(\mathbf{x}|\theta))$ is $f_X(\mathbf{x}|\mu)/f_{\bar{X}}(\bar{\mathbf{x}}|\mu)$ which expands to:

$$\frac{f_X(\mathbf{x}|\mu)}{f_{\bar{X}}(\bar{\mathbf{x}}|\mu)} = \frac{(2\pi\sigma^2)^{-n/2}\exp(-\sum_{i=1}^n (x_i-\bar{x})^2 + n(\bar{x}-\mu)^2/(2\sigma^2))}{(2\pi\sigma^2)^{-1/2}\exp(-n(\bar{x}-\mu)^2/(2\sigma^2))}$$
$$= n^{-1/2}(2\pi\sigma^2)^{-(n-1)/2}\exp(-\sum_{i=1}^n (x_i-\bar{x})^2/(2\sigma^2))$$

and does not depend on μ. Thus the sample mean is a sufficient statistic for the parameter μ.

Sufficiency with order statistics

Sometimes we can't reduce the sample and have to resort to other means for determining sufficiency.

Sufficiency when density is unknown

Let $X_1, ..., X_n$ be iid with pdf f which is unknown. The best we can do in this case is show that the order statistics $X_{(1)}, ..., X_{(n)}$ are sufficient for f.
(Example in class)

Factorization Theorem

Theorem

A statistic $T(\mathbf{X})$ is sufficient for θ iff there exists functions $g(t|\theta)$ and $h(\mathbf{x})$ such that the joint pdf or pmf, $f(\mathbf{x}|\theta)$ can be written as:

$$f(\mathbf{x}|\theta) = g(T(\mathbf{x})|\theta)h(\mathbf{x})$$

Proof (discrete case):

$$\begin{aligned} f(\mathbf{x}|\theta) &= P_\theta(\mathbf{X} = \mathbf{x}) = P_\theta(\mathbf{X} = \mathbf{x} \text{ and } T(\mathbf{X}) = T(\mathbf{x})) \\ &= P_\theta(T(\mathbf{X}) = T(\mathbf{x}))P(\mathbf{X} = \mathbf{x}|T(\mathbf{X}) = T(\mathbf{x})) \\ &= g(T(\mathbf{x})|\theta)h(\mathbf{x}) \end{aligned}$$

Factorization Theorem and Exponential Families

- It is easy to find sufficient statistics for exponential family distributions using the Factorization Theorem.
- Exponential families are described in CB 3.4. They include many of the most common distributions (both discrete and continuous): normal, exponential, gamma, chi-squared, beta, Dirichlet, binomial, Bernoulli, negative binomial, Poisson, Wishart, Inverse Wishart.

Exponential Families

Distributions belonging to the exponential family can be expressed as:

$$f(x|\theta) = h(x)c(\theta)exp(\sum_{i=1}^{k} w_i(\theta)t_i(x))$$

Where $h(x) \geq 0$ and $t_1(x), ..., t_k(x)$ are real valued funtions of the observations x (they cannot depend on θ), $c(\theta) \geq 0$, and $w_1(\theta), ..., w_k(\theta)$ are real valued functions of the paramter(s) θ (they cannot depend on x).

Factorization Theorem and Exponential Families

- The important thing to notice is what characterizes the exponential family distributions–the parameter(s) and observation variable(s) must factorize.
- This means the distribution can be separated into products that each involve either the parameters or the observations.
- To verify that a pdf or pmf belongs to the exponential family, the functions $h(x)$, $c(\theta)$, $w_i(\theta)$ and $t_i(\theta)$ must be identified and shown to have the form shown above.
- Example in class of exponential family $N(\mu, \sigma^2)$

Factorization Theorem and Exponential Families

Sufficiency, Factorization Theorem and Exponential Families

Let $X_1, ..., X_n$ be iid observations from a pdf or pmf $f(x|\theta)$ that belongs to an exponential family:

$$f(x|\theta) = h(x)c(\theta)exp(\sum_{i=1}^{k} w_i(\theta)t_i(x))$$

where $\theta = (\theta_1, ..., \theta_d)$, $d \leq k$.
Then

$$T(\mathbf{X}) = (\sum_{j=1}^{n} t_1(X_j), ..., \sum_{j=1}^{n} t_k(X_j))$$

is sufficient for θ.
Examples in class of Poisson and normal exponential family factorization for finding sufficient statistics.

Minimal Sufficient Statistics

- As we shave seen, there are cases where there are many sufficient statistics for a particular model.
- Sufficient statistics are not unique. If $T(X)$ is sufficient and $T^*(X)$ is another statistic such that $T(X) = g_1(T^*(X))$ for some function g_1 then $T^*(X)$ is also sufficient.

$$
\begin{aligned}
f(x|\theta) &= g(T(x)|\theta)h(x) \\
&= g(g_1(T^*(x))|\theta)h(x) \\
&= g^*(T^*(x)|\theta)h(x)
\end{aligned}
$$

- So if T(X) is sufficient, so is $T^*(X) = (T(X), T_1(X))$ where $T_1(X)$ is any other statistic.
- If $T(X) = g_1(T^*(X))$ then the partition of $\mathcal{X}$ defined by $T(x)$ is coarser than that defined by $T^*(x)$.

Minimal Sufficient Statistics

- Given many possible sufficient statistics, are some better than others?
- Recall we want a statistic that provides data reduction without loss of information about the parameter θ. Thus, a statistic that achieves the most data reduction while retaining all the information about θ is preferable. Such a statistic is called a minimal sufficient statistic.

Minimal Sufficient Statistic

$T(X)$ is a minimal sufficient statistic if it is sufficient and for any other sufficient statistic $T^*(X)$, $T(X)$ is a function of $T^*(X)$.

- However, this definition of minimal sufficient statistics does not often help identify which of a group of sufficient statistics is actually minimal (normal model example).

Minimal Sufficient Statistics

Theorem for Minimal Sufficient Statistics

If $T(X)$ has the property that the ratio $f(x|\theta)/f(y|\theta)$ does not depend on θ iff $T(x) = T(y)$ then $T(X)$ is a minimal sufficient statistic for θ.
Proof:
Let $T(X)$ satisfy the condition of the theorem. We show that $T(X)$ is sufficient and that it is minimally sufficient.
Let x_t denote an element of A_t. Recall $A_t = \{x : T(x) = t\}$. So, $T(x_t) = t$ and $T(x_{T(x)}) = T(x)$
from the theorem, we have:

$$\frac{f(x|\theta)}{f(x_{T(x)}|\theta)} = h(x)$$

where $h(x)$ is some function that does not depend on θ.

$$f(x|\theta) = f(x_{T(x)}|\theta)h(x)$$

So by the factorization theorem, $T(X)$ is sufficient.

Minimal Sufficient Statistics

Theorem for Minimal Sufficient Statistics, con't

Let $T^*(X)$ be another sufficient statistic. By the factorization theorem we have

$$\frac{f(x|\theta)}{f(y|\theta)} = \frac{h^*(x)g^*(T^*(x)|\theta)}{h^*(y)g^*(T^*(y)|\theta)} = \frac{h^*(x)}{h^*(y)}$$

Thus $T^*(x) = T^*(y)$ implies that $f(x|\theta)/f(y|\theta)$ does not depend on θ. From the assumption that $T(x) = T(y)$ it follows that the partition of $\mathcal{X}$ induced by $T^*(x)$ is finer than that induced by $T(X)$. This implies that $T(X)$ is a minimal sufficient statistic.

Minimal Sufficient Statistics

Some general notes about sufficiency and minimal sufficiency:

- In terms of partitioning the sample space, any sufficient statistic introduces a partition of the sample space.
- The partition of the minimal sufficient statistic is the **coarsest** so that it achieves the greatest possible data reduction for a sufficient statistic.
- A minimal sufficient statistic eliminates all of the extra information in the sample and leaves only that which contains information about θ.

Ancillary Statistics

An ancillary statistic:

- contains no information about parameter θ; however, it provides a complimentary purpose to a sufficient statistic.
- An ancillary statistic by itstelf does not provide any information about a parameter, but in conjunction with another statistic it can (R.A. Fisher).
- is an observation on a random variable whose distribution is fixed and known, but unrelated to θ.
- is denoted as $S(X)$

The range statistic $R = X_{(n)} - X_{(1)}$ is a common example of an ancillary statistic because it does not depend on the distribution of the sample $\mathbf{x}$ but rather on the parameter of the distribution that relates to *location*.
Other examples include ancillary statistics belonging to the scale family, or a mixture of scale and location.

Ancillary Statistics

Location family θ is the location parameter:

$$\{F(x - \theta) : -\infty < \theta < \infty\}$$

Scale family θ is the scale parameter:

$$\{(1/\theta)F(x/\theta) : \theta > 0\}$$

Scale-Location family θ_1 is the scale parameter and θ_2 is the location parameter:

$$\{(1/\theta_1)F((x - \theta_2)/\theta_1) : \theta_1 > 0, -\infty < \theta_2 < \infty\}$$

Ancillary Statistics

Location Ancillary Statistics

We use the CDF to show how location ancillary statistics do not depend on the parameter θ. Let $X_1, ..., X_n$ be iid observation from a location parameter family with cdf $F(x-\theta)$. Let $Z_1, ..., Z_n$ be iid observations with cdf $F(x)$ (i.e. θ=0) with $X_1 = Z_1 + \theta, ..., X_n = Z_n + \theta$. Show the range $R = X_{(n)} - X_{(1)}$ is an ancillary statistic.

The cdf of R is

$$\begin{aligned} F(r|\theta) &= P(R \leq r) \\ &= P(\max X_i - \min X_i \leq r) \\ &= P(\max(Z_i+\theta) - \min(Z_i+\theta) \leq r) \\ &= P(\max Z_i - \min Z_i + \theta - \theta \leq r) \\ &= P(\max Z_i - \min Z_i \leq r) \end{aligned}$$

Which does not depend on θ because the distribution of $Z_1, ..., Z_n$ does not depend on θ.

In class example showing the range for Uniform$(\theta, \theta+1)$ is an ancillary statistic.

Ancillary Statistics

Scale Ancillary Statistics

Again we use the CDF to show how scale ancillary statistics do not depend on the parameter θ. Let $X_1, ..., X_n$ be iid observation from a scale parameter family with cdf $F(x/\theta)$. Any statistic that depends on the sample through its n-1 values $X_1/X_n, ...X_{n-1}/X_n$ is an ancillary statistic.
Let $Z_1, ..., Z_n$ be iid observations with cdf $F(x)$ (i.e. θ=1) with $X_i = \theta Z_i$. The joint CDF of $X_1/X_n, ...X_{n-1}/X_n$ is:

$$\begin{aligned} F(y_1, ..., y_{n-1}|\theta) &= P(X_1/X_n \le y_1, ..., X_{n-1}/X_n \le y_{n-1}) \\ &= P(\theta Z_1/(\theta Z_n) \le y_1, ..., \theta Z_{n-1}/(\theta Z_n) \le y_{n-1}) \\ &= P(Z_1/Z_n \le y_1, ..., Z_{n-1}/Z_n \le y_{n-1}) \end{aligned}$$

Which does not depend on θ because the distribution of $Z_1, ..., Z_n$ does not depend on θ.

Complete Statistics

Ancillary statistics in conjunction with sufficient statistics provides us with a definition for complete statistics.
Basically, if we have a sufficient statistic that optimally summarizes the observations, then there should be an ancillary statistic that is a function of that statistic.

Basu's Theorem

If $T(X)$ is complete and a minimal sufficient statistic then $T(X)$ is independent of every ancillary statistic.
(i.e. A complete sufficient statistic is independent of every ancillary statistic.)

(Proof in class)

Complete Statistics

Complete statistics apply to families of distributions, most importantly the exponential family.

Complete Statistics in the Exponential Family

Let $X_1, .X_2, ..., X_n$ be observations from an exponential family with pdf (or pmf) that has the form i

$$f(x|\theta) = h(x)c(\theta)\exp(\sum_{j=1}^{k} w(\theta_j)t_j(x))$$

where $\theta = (\theta_1, \theta_2, ..., \theta_k)$. Then the statistic

$$T(X) = (\sum_{i=1}^{n} t_1(X_i), \sum_{i=1}^{n} t_2(X_i), ... \sum_{i=1}^{n} t_k(X_i))$$

is complete as long as the parameter space Θ contains an open set in $\Re^k$

Likelihood

First we need to go over the likelihood principle:

- Likelihoods relate data to a population
- They arise from a probability distribution function $f(x|\theta)$ connecting data x to a population
- Used as a data reduction technique
- We assume data come from a family of distributions with unknown parameters
- We use the data to estimate these unknown parameters

Likelihood

Definition:The Likelihood Principle

The likelihood principle states that given a pdf $f(\mathbf{x}|\theta)$ and observed data $\mathbf{x}$, all of the relevant information regarding the unknown parameter(s) θ is contained in the likelihood function for the observed $\mathbf{x}$

Two likelihood functions contain the same information about θ if they are proportional to each other.

Likelihood

- In the context of random variables, $X_1, X_2, ..., X_n$ are an iid sample from a population with pdf $f(x|\theta_1, \theta_2, ..., \theta_k)$ where $\theta_i, i = 1, ..., k$ are unknown parameters
- The likelihood is f viewed as a function of of θ_i for fixed observed values of x
- The joint density of the data evaluated as a function of the parameters with the data fixed

Equivariance

- Given some data, statistical decisions should not be affected by simple transformations or reordering of the data.
- For example, the value of a point estimate will be affected by a transformation, but it should be *equivariant* in the sense that it reflects the transformation in a meaningful way.
- This is formalized by the equivariance principle through which appropriate classes of transformations are defined and rules that statistical decisions must satisfy are specified.