Bruce M. Boghosian

Properties of Estimators

Factorization theorems for sufficiency

Bruce M. Boghosian

Department of Mathematics

Tufts University

Outline

Definition of sufficiency

- Recall the definition: Let $X_j = k_j$ for $j = 1, \ldots, n$ be a random sample of size n from $p_X(k;\theta)$. The statistic $\hat{\theta} = h(X_1, \ldots, X_n)$ is *sufficient* for θ if the likelihood function $L(\theta)$ factors into the product of the probability distribution for $\hat{\theta}$ and constant that does not involve θ, i.e.,

$$L(\theta) = \prod_{j=1}^{n} p_X(k_j;\theta) = p_{\hat{\theta}}(\theta_e;\theta)\; b(k_1, \ldots, k_n).$$

- For continuous random variables $Y_j = y_j$ for $j = 1, \ldots, n$,

$$L(\theta) = \prod_{j=1}^{n} f_Y(y_j;\theta) = f_{\hat{\theta}}(\theta_e;\theta)\; b(y_1, \ldots, y_n).$$

- We used this definition for Bernoulli and binomial random variables, e.g.
- Problem: Sometimes it is difficult to figure out $p_{\hat{\theta}}$.

The second factorization criterion

Bruce M. Boghosian

- **Thm:** Let $X_j = k_j$ for $j = 1, \ldots, n$ be a random sample of size n from the discrete pdf $p_X(k;\theta)$. The statistic $\hat{\theta}$ is *sufficient* for θ iff there are functions $g[h(k_1,\ldots,k_n);\theta]$ and $b(k_1,\ldots,k_n)$ such that

$$L(\theta) = \prod_{j=1}^{n} p_X(k_j;\theta) = g[h(k_1,\ldots,k_n);\theta]\; b(k_1,\ldots,k_n)$$

- Statement for continuous r.v.s $Y_j = y_j$ for $j = 1, \ldots, n$,

$$L(\theta) = \prod_{j=1}^{n} f_Y(y_j;\theta) = g[h(y_1,\ldots,y_n);\theta]\; b(y_1,\ldots,y_n)$$

The second factorization criterion (continued)

- **Pf (for discrete case):** First suppose that $\hat{\theta}$ is sufficient. Then, by definition, we can write

$$L(\theta) = p_{\hat{\theta}}(\theta_e; \theta)\ b(k_1, \ldots, k_n),$$

which is of the desired form if we identify

$$g[h(k_1, \ldots, k_n); \theta] = p_{\hat{\theta}}(h(k_1, \ldots, k_n); \theta).$$

- Next suppose that

$$L(\theta) = \prod_{j=1}^{n} p_X(k_j; \theta) = g[h(k_1, \ldots, k_n); \theta]\ b(k_1, \ldots, k_n)$$

- We need to show that $g[h(k_1, \ldots, k_n); \theta]$ can always be rewritten in terms of the pdf of $\hat{\theta}$.

The second factorization criterion (continued)

- Can we write $g[h(k_1,\ldots,k_n);\theta]$ in terms of pdf of $\hat{\theta}$?
- Let c be some possible value of $h(k_1,\ldots,k_n)$, and let $A=\left\{\vec{k}\mid h(\vec{k})=c\right\}$ be the inverse image of c, so we write $A=h^{-1}(c)$. Then

$$\therefore\quad p_{\hat{\theta}}(c;\theta)=\sum_{\vec{k}\in A}p_{\vec{X}}(\vec{k})=\sum_{\vec{k}\in A}\prod_{j=1}^{n}p_{X_j}(k_j)=\sum_{\vec{k}\in A}g(c;\theta)b(\vec{k})=g(c;\theta)\left[\sum_{\vec{k}\in A}b(\vec{k})\right]$$

- It follows that

$$g(c;\theta)=\frac{p_{\hat{\theta}}(c;\theta)}{\sum_{\vec{k}\in A}b(\vec{k})}$$

$$\therefore\quad L(\theta)=p_{\hat{\theta}}(h(\vec{k});\theta)\frac{b(\vec{k})}{\sum_{\vec{k}\in A}b(\vec{k})}$$

and hence $\hat{\theta}$ is sufficient by definition. $\square$

Sufficient estimators – Example 5

- Show exponential distribution $f_Y(y) = \theta e^{-\theta y}$ is sufficient
- Likelihood function is

$$L(\theta) = \prod_{j=1}^{n} \left(\theta e^{-\theta y_j}\right) = \theta^n \exp\left(-\theta n \overline{y}\right)$$

- This is of the form

$$L(\theta) = g[h(y_1, \ldots, y_n); \theta]\, b(y_1, \ldots, y_n)$$

- where we have defined

$$\begin{aligned} h(y_1, \ldots, y_n) &:= \overline{y} \\ g(h; \theta) &:= \theta^n \exp\left(-\theta n h\right) \\ b(y_1, \ldots, y_n) &:= 1. \end{aligned}$$

- Hence the exponential distribution is sufficient.

Exponential form of pdf

- **Thm.:** Let $Y_1 = y_1, Y_2 = y_2, \ldots, Y_n = y_n$ be n independent random samples from a pdf of the general form

$$f_Y(y;\theta) = \exp\left[K(y)p(\theta) + S(y) + q(\theta)\right],$$

where θ is a parameter, and where the support of f does not depend on θ. Then the estimator

$$\hat{\theta}(\vec{y}) := \sum_{j=1}^{n} K(y_j),$$

which may or may not be unbiased for θ, is sufficient.

Proof of sufficiency

Bruce M. Boghosian

- For pdf $f_Y(y;\theta) = \exp\left[K(y)p(\theta) + S(y) + q(\theta)\right]$,

$$L(\theta) = \prod_{j=1}^{n} f_Y(y_j;\theta) = \prod_{j=1}^{n} \exp\left[K(y_j)p(\theta) + S(y_j) + q(\theta)\right]$$
$$= \exp\left[\sum_{j=1}^{n} K(y_j)p(\theta) + \sum_{j=1}^{n} S(y_j) + nq(\theta)\right]$$

- This is of the form $L(\theta) = g[h(y_1,\dots,y_n);\theta]\ b(y_1,\dots,y_n)$
- where we have defined

$$h(y_1,\dots,y_n) := \sum_{j=1}^{n} K(y_j)$$
$$g(h;\theta) := hp(\theta) + nq(\theta)$$
$$b(y_1,\dots,y_n) := \exp\left[\sum_{j=1}^{n} S(y_j)\right]$$

- Hence any pdf of this form is sufficient. $\square$

Sufficient estimators – Example 6

Bruce M. Boghosian

- Show exponential distribution $f_Y(y) = \theta e^{-\theta y}$ is sufficient
- We can write $f_Y(y) = e^{-\theta y + \ln\theta}$
- In exponential form, $\exp\left[K(y)p(\theta) + S(y) + q(\theta)\right]$, where

$$\begin{aligned} K(y) &:= y \\ p(\theta) &:= -\theta \\ S(y) &:= 0 \\ q(\theta) &:= \ln\theta. \end{aligned}$$

- Hence, once again, exponential distribution is sufficient.

Summary

- We have continued our study of sufficiency of estimators.
- We found three factorization theorems for proving sufficiency
  - Directly from definition (requires knowledge of $f_{\hat{\theta}}$).
  - Second factorization theorem
  - Exponential form