Sufficiency and Exponential Families

February 14, 2025

Outline

Intuition

- Consider an asymmetric coin with Heads turning out with probability $\theta \in (0, 1)$.

Intuition

- Consider an asymmetric coin with Heads turning out with probability $\theta \in (0, 1)$.
- Ann tossed it 1000 times and reported that the total number of times that Heads turned out is 768.
- Based on this information, John wants to say something about the probability of Heads turning out.
- John sends a request to Ann to report the whole sequence of outcomes not just a single number.
- He argues that the more information he has the more accurate conclusions he can draw.
- Do you think that he has the point to request the information at which specific tosses Heads turned out? or ...

Intuition

- Consider an asymmetric coin with Heads turning out with probability $\theta \in (0, 1)$.
- Ann tossed it 1000 times and reported that the total number of times that Heads turned out is 768.
- Based on this information, John wants to say something about the probability of Heads turning out.
- John sends a request to Ann to report the whole sequence of outcomes not just a single number.
- He argues that the more information he has the more accurate conclusions he can draw.
- Do you think that he has the point to request the information at which specific tosses Heads turned out? or ... the information John got from Ann is *sufficient*.

Example – Bernoulli trials

Let $X = (X_1, ..., X_n)$ be a vector of i.i.d Bernoulli(θ) random variables. The pmf function of X is:

$$\begin{aligned} p(X|\theta) &= P(X_1 = x_1|\theta) \cdot P(X_2 = x_2|\theta) \cdot \cdots \cdot P(X_n = x_n|\theta) \\ &= \theta^{x_1}(1-\theta)^{1-x_1} \cdot \theta^{x_2}(1-\theta)^{1-x_2} \ldots \theta^{x_n}(1-\theta)^{1-x_n} \\ &= \theta^{\sum x_i}(1-\theta)^{n-\sum x_i} \end{aligned}$$

Consider $T(X) = \sum_{i=1}^{n} X_i$ whose distribution has the binomial pmf:

$$\binom{n}{t}\theta^t(1-\theta)^{n-t}, 0 \leq t \leq n$$

1. Find the distribution of X given $T(X)$.

Example – Bernoulli trials

Let $X = (X_1, ..., X_n)$ be a vector of i.i.d Bernoulli(θ) random variables. The pmf function of X is:

$$\begin{aligned} p(X|\theta) &= P(X_1 = x_1|\theta) \cdot P(X_2 = x_2|\theta) \cdot \dots \cdot P(X_n = x_n|\theta) \\ &= \theta^{x_1}(1-\theta)^{1-x_1} \cdot \theta^{x_2}(1-\theta)^{1-x_2} \dots \theta^{x_n}(1-\theta)^{1-x_n} \\ &= \theta^{\sum x_i}(1-\theta)^{n-\sum x_i} \end{aligned}$$

Consider $T(X) = \sum_{i=1}^{n} X_i$ whose distribution has the binomial pmf:

$$\binom{n}{t}\theta^t(1-\theta)^{n-t}, 0 \leq t \leq n$$

1. Find the distribution of X given $T(X)$.
2. The distribution is uniform over the n-tuples X such that $T(X) = t$.

Conclusion

1. Once you know $T(X) = \sum_{i=1}^{n} X_i$ the distribution of X given this information does not longer depend on the parameter θ.
2. $T(X) = \sum_{i=1}^{n} X_i$ 'took away' all the information about θ.
3. To make decision concerning θ, only the information of $T(X) = t$ is needed, since the value of X given t reflects only the order information in X which is independent of θ.

Homework *Variables X_i, $i = 1, \ldots, d$ are called exchangeable if the distribution of a vector $\mathbf{X} = (X_1, \ldots, X_n)$ is invariant on permutations of coordinates. Argue that iid samples are exchangeable. Give an example of exchangeable variables that are not iid. Show that the vector of order statistics for exchangeable random variables is a sufficient statistics.*

Sufficiency – the formal definition

Definition

Let $X \sim P_\theta$, $\theta \in \Theta$ and $T(X)$ is a statistic of X. The statistic T is sufficient for θ if the conditional distribution of X given $T = t$ is independent of θ.

Example 1.5.2 Customers arrive at a service counter according to a Poisson process with arrival rate parameter θ. Let X_1 and X_2 be the inter-arrival times of first two customers. (From time 0, customer 1 arrives at time X_1 and customer 2 at time $X_1 + X_2$.) Prove that $T(X_1, X_2) = X_1 + X_2$ is sufficient for θ.

Solution (Sketch) What is the distribution of $X_1 + X_2$? If X_1 and X_2 are independent random variables with $\Gamma(p, \theta)$ and $\Gamma(q, \theta))$ distributions, then $Y_1 = X_1 + X_2$ and $Y_2 = X_1/(X_1 + X_2)$ are independent and $Y_1 \sim \Gamma(p + q, \theta)$ and $Y_2 \sim B(p, q)$.

Thus, with $p = q = 1$, $T \sim \Gamma(2, \theta)$ and $Y_2 \sim U(0, 1)$, and independent.

$$[(X_1, X_2)|T = t] \sim (X, Y)$$

with $X \sim U(0, t)$; $Y = t - X$.

Factorization theorem

In general, checking sufficiency directly is difficult because we need to compute the conditional distribution which often leads to a singular distribution. Fortunately, a simple necessary and sufficient criterion for a statistic to be sufficient is available. This result was proved in various forms by Fisher, Neyman, also by Halmos and Savage. It is often referred to as the factorization theorem for sufficient statistics.

Theorem 1.5.1.

In a **regular model**, a statistic $T(X)$ is sufficient for θ if, and only if, there exists a function $g(t, \theta)$ and a function h defined on the space of values of X such that

$$p(x|\theta) = g(T(x), \theta)h(x)$$

The proof in the book given for the discrete case is recommended to follow.

Applications of the theorem

Example 1.5.2 Let $X_1, X_2, \ldots, X_n$ be inter-arrival times for n customers which are iid Exponential(θ) r.v.'s

$$p(x_1, \ldots, x_n|\theta) = \theta^n e^{-\theta \sum_{i=1}^n x_i},$$

where $0 < x_i, i = 1, \ldots, n$. Thus $T(X_1, ..., X_n) = \sum_{i=1}^n x_i$ is sufficient.

Example – sample from Uniform Distribution Let $X_1, \ldots, X_n$ be a sample from the U(α, β) distribution:

$$p(x_1, \ldots, x_n|\alpha, \beta) = \frac{1}{(\beta - \alpha)^n}; \mathbf{x} \in [\alpha, \beta]^n.$$

We note that this density can be written as

$$\begin{aligned} p(x_1, \ldots, x_n|\alpha, \beta) &= \frac{1}{(\beta - \alpha)^n} \mathbb{I}_{[\alpha,\beta]}(\min x_i) \mathbb{I}_{[\alpha,\beta]}(\max x_i) \\ &= \frac{1}{(\beta - \alpha)^n} \mathbb{I}_{[\alpha,\infty)}(\min x_i) \mathbb{I}_{(-\infty,\beta]}(\max x_i), \end{aligned}$$

and thus the statistic $T(x_1, \ldots, x_n) = (\min x_i, \max x_i)$ is sufficient for $\theta = (\alpha, \beta)$. If α (β) is known, then $\max x_i$ ($\min x_i$) is sufficient for β (α).

Example 1.5.4– Normal Sample

Let $X_1, ..., X_n$ be iid N(μ, σ^2). The joint density is

$$\begin{aligned} p(x_1, \ldots, x_n|\theta) &= \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-\sum_{i=1}^{n}(x_i-\mu)^2/(2\sigma^2)} \\ &= \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-n(\overline{x^2}-2\mu\bar{x})/(2\sigma^2)} e^{-n\mu^2/(2\sigma^2)} \end{aligned}$$

Thus $T(X_1, \ldots, X_n) = (\overline{X}, \overline{X^2})$ is sufficient. But also $\tilde{T}(X_1, \ldots, X_n) = (\overline{X}, S^2)$ is sufficient since any one-to-one mapping of a sufficient statistics is also sufficient.

Homework *Argue formally that any one-to-one mapping of a sufficient statistics is also a sufficient statistics. How the 'one-to-one' assumption can be weakened.*

Minimal Sufficiency

Issue: Probability models often admit many sufficient statistics. Suppose $X = (X_1, \ldots, X_n)$ where X_i are iid from P_θ, $\theta \in \Theta$. $T(X) = (X_1, ..., X_n)$ is (trivially) sufficient.
If X_i' are iid $N(\theta, 1)$, then $T' = \bar{X}$ is sufficient and provides a greater reduction of the data.

Homework *Provide formal argument for the above statements.*

Definition

A statistic $T(X)$ is minimally sufficient if it is sufficient and provides a greater reduction of the data than any other sufficient statistic, i.e. if $S(X)$ is any sufficient statistic, then there exists a mapping r such that

$$T(X) = r(S(X)).$$

Example 1.5.1 (continued)

$X_1, \ldots, X_n$ are iid Bernoulli(θ) and $T(X) = \sum_{i=1}^n X_i$ is sufficient. We will show that it is also minimal.
Let $S(X)$ be any other sufficient statistic. By the factorization theorem:

$$p(x|\theta) = g(S(x), \theta)h(x),$$

Using the pmf of X we have

$$\theta^{T(x)}(1-\theta)^{(n-T(x))} = g(S(x), \theta)h(x)$$

Fix any two values of θ, say θ_1 and θ_2 and take the ratio of the pmfs:

$$(\theta_1/\theta_2)^{T(x)}[(1-\theta_1)/(1-\theta_2)]^{n-T(x)} = g(S(x), \theta_1)/g(S(x), \theta_2)$$

Take logarithm of both sides and solve for $T(x)$. E.g., $\theta_1 = 2/3$ and $\theta_2 = 1/3$

$$T(x) = r(S(x)) = \log[2^n g(S(x), 2/3)/g(S(x), 1/3)]/2\log 2.$$

The likelihood ratio

Definition – the Likelihood Function

For $X \sim P_\theta$, $\theta \in \Theta$. Let $p(x|\theta)$ be the pmf or density function. The likelihood function L for a given observed data value $X = x$ is a function of the parameter

$$\theta \mapsto L_x(\theta) = p(x|\theta)$$

Theorem (Dynkin, Lehmann, and Scheffe)

Suppose there exists θ_0 such that the support of $p(x|\theta_0)$ contains all the supports of $p(x|\theta)$ $\theta \in \Theta$ (support is the set on which the density (pmf) is positive). Let

$$\Lambda_x(\cdot) = \frac{L_x(\cdot)}{L_x(\theta_0)} : \Theta \mapsto \mathbb{R}.$$

Then $x \mapsto \Lambda_x$ is a function valued minimal statistics.

Outline

Exponential family

Definition

$\{P_\theta\}$, $\theta \in \Theta$ is a k-parameter exponential family if

$$p(x|\theta) = h(x) \exp\left(\sum_{j=1}^{k} \eta_j(\theta) T_j(x) - B(\theta)\right), \quad x \in \mathbb{R}^q$$

- $\eta_1, \ldots, \eta_k$ and B are real-valued functions mapping $\Theta \mapsto \mathbb{R}$,
- $T_1, \ldots, T_k$ and h are real-valued functions mapping $\mathbb{R}^q \mapsto \mathbb{R}$.

Note: By the Factorization Theorem (Theorem 1.5.1)

- $\mathbf{T}(X) = (T_1(X), \ldots, T_k(X))$ is sufficient.
- For an iid sample $X_1, \ldots, X_n$ from P_θ, its distribution is a k-parameter exponential family with natural sufficient statistic

$$T^{(n)} = \sum_{i=1}^{n} (T_1(X_i), \ldots, T_k(X_i))$$

Importance and convenience of exponential families

There are many benefits of using exponential families

- The minimal sufficient statistic can be conveniently established.
- Moment generating functions for such a statistic can be obtained in a general form.
- Mean and variance have an explicit form.
- Questions of existence and uniqueness of maximum likelihood estimates can be answered completely and elegantly. This is largely a consequence of the strict concavity of the log likelihood in the natural parameter $\eta = (\eta_1, \ldots, \eta_k)$.

Canonical form of exponential families

Canonical parameter

$$q(x,\eta) = h(x)\exp(\mathbf{T}^\top(x)\eta - A(\eta))$$
$$A(\eta) = \log \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} h(x)\exp(\mathbf{T}^\top(x)\eta)\, dx.$$

In the discrete case, $A(\eta)$ is defined in the same way except integrals are replaced by sums. The natural parameter space is

$$\mathcal{E} = \{\eta \in \mathbf{R}^k : -\infty < A(\eta) < \infty\}.$$

An exponential family is of rank k iff the generating statistic $\mathbf{T}$ is k-dimensional and $1, T_1(X), \ldots, T_k(X)$ are linearly independent with positive probability.

Homework *Show that every k-parameter exponential family is also k'-dimensional with $k' > k$ but the rank of an exponential is a minimal dimension.*

Moment generating function and moments

MGF

Both $\mathcal{E}$ and A are convex. If η_0 is in the interior of $\mathcal{E}$ (so A is convex around η_0), then

$$M(\mathbf{s}) = \exp(A(\eta_0 + \mathbf{s}) - A(\eta_0)),$$

for all $\mathbf{s}$ such that $\eta_0 + \mathbf{s} \in \mathcal{E}$.

We obtain

$$E_{\eta_0}(\mathbf{T}) = \dot{A}(\eta_0), \quad Var_{\eta_0}(T(X)) = \ddot{A}(\eta_0),$$

where $\dot{A}(\eta_0)$ is the gradient (vector of the derivatives) and $\ddot{A}(\eta_0)$ is the Hessian (matrix of the second derivatives).

Homework *Recall what is the moment generating function for the multivariate variable. Using this definition, derive the above form of MGF for the exponential family in the canonical form*

Examples

- Poisson distribution

$$p(x|\theta) = e^{-\theta}\theta^x/x!.$$

thus $h(x) = 1/x!$, $B(\theta) = \theta$, $\eta(\theta) = \log\theta$, $T(x) = x$.

- Gamma distribution

$$p(x|\lambda, p) = \frac{\lambda^p x^{p-1}}{\Gamma(p)} e^{-\lambda x}$$
$$= e^{(p-1)\log x - \lambda x - \log\Gamma(p) + p\log\lambda},$$

$h(x) = 1$, $B(\lambda, p) = p\log\lambda - \log\Gamma(p)$, $\eta_1(p, \lambda) = (p-1)$, $\eta_2(p, \lambda) = -\lambda$, $T_1(x) = \log x$, $T_2(x) = x$.

Homework *Derive the canonical form for the gamma distribution and compute from it, its mean and variance.*