Chapter 7 Point Estimation

Point estimator definition:

Any function of a sample is a point estimate; Any staticstic is a point estimator.

Method of Moments

Equating the first k sample moments to the corresponding k population moments and solving the resulting system of simultaneous equations.

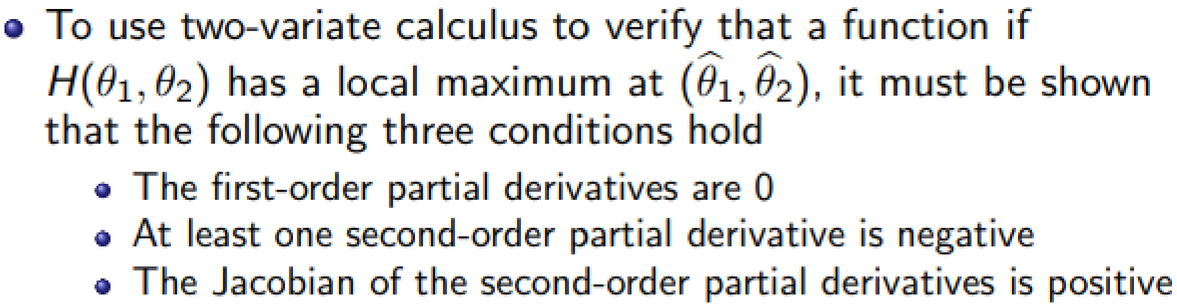
- Pros: Always provide an estimator;Somestimes helpful.
- Cons: Might occur range of estimator and range of parameter does not coincide thus draw unreasonable estimation.

Maximum Likelihood Estimators

$$\max_{\theta} L(\theta|x) = \Pi_{i=1}^{n} f(x_i|\theta)$$

Pros: Range of MLE always coincides with the range of parameters. Also satisfy some optimal condition.
Cons: Sensitivity problem; Maximum problem might be hard to solve.

- Invariance Property: If $\hat{\theta}$ is the MLE of θ, then for any function $\tau(\theta)$, the MLE of $\tau(\theta)$ is $\tau(\hat{\theta})$
- Two variate calculas:

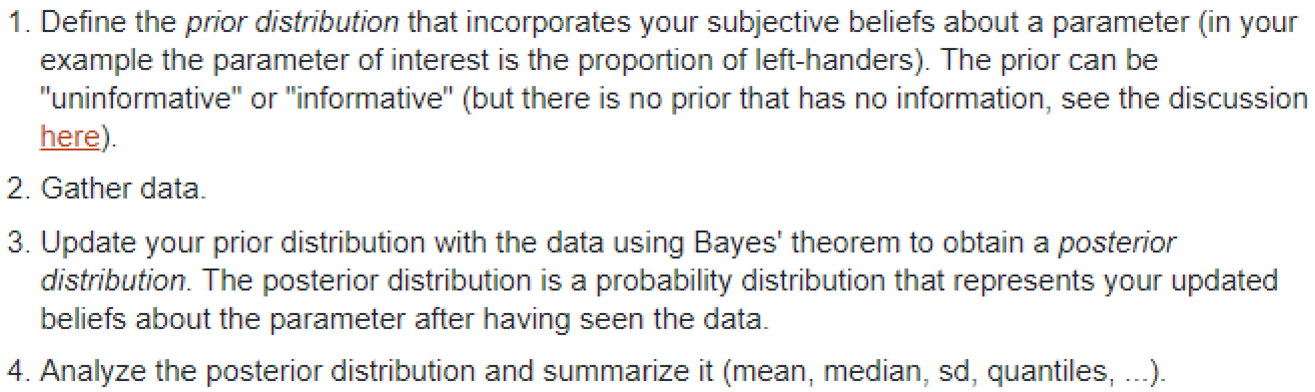


- To use two-variate calculus to verify that a function if $H(\theta_1, \theta_2)$ has a local maximum at $(\widehat{\theta_1}, \widehat{\theta_2})$, it must be shown that the following three conditions hold
  - The first-order partial derivatives are 0
  - At least one second-order partial derivative is negative
  - The Jacobian of the second-order partial derivatives is positive

Notice that the second and third condition guarantees that the Hessian matrix is negatice-definite matrix, thus the function is a concave one, thus the maximum is obtained.

Bayesian Estimator

Some steps in Bayesian estimation:



1. Define the *prior distribution* that incorporates your subjective beliefs about a parameter (in your example the parameter of interest is the proportion of left-handers). The prior can be "uninformative" or "informative" (but there is no prior that has no information, see the discussion here).
2. Gather data.
3. Update your prior distribution with the data using Bayes' theorem to obtain a *posterior distribution*. The posterior distribution is a probability distribution that represents your updated beliefs about the parameter after having seen the data.
4. Analyze the posterior distribution and summarize it (mean, median, sd, quantiles, ...).

- Conjugate family: A class Π of prior distributions is a conjugate family of F if the posterior distribution is in the class Π for all $f \in F, x \in X$ and all priors in Π.
  The definition of conjugate family make sure that the prior and posterior is in the same distribution family. Some examples: Binomial and beta family; Multinomial distribution and Dirichlet priors; Poisson distribution and gamma priors.
  What's interesting is that the posterior estimator is usually a combination of the prior estimation and the data's information. For Normal distribution, It's of the form:

$$E[\theta|x] = \frac{\tau^2}{\tau^2+\sigma^2}x + \frac{\sigma^2}{\tau^2+\sigma^2}\mu, Var[\theta|x] = \frac{\sigma^2\tau^2}{\tau^2+\sigma^2}$$

  Where $X - N(\theta, \sigma^2), \theta - N(\mu, \tau^2)$.

Method of Evaluating Estimates

MSE

Definition: An estimator W of parameter θ is the function of θ defined by $E_\theta(W-\theta)^2$.

$$E_\theta(W-\theta)^2 = Var_\theta W + (E_\theta W - \theta)^2 = Var_\theta W + (Bias_\theta W)^2$$

We call an estimator is unbiased if $E_\theta W = \theta$. In this case, the MSE is just the variance of W.

Notice

- MSE is just a measurement we use to judge whether an estimator is better than the other, it has its flaws, for example, it tends to be forgiving of underestimation in the scale case(suppose you're estimating a variance, σ^2. Then, if you consciously underestimate the quantity your MSE can be at most σ^4, while overestimation can produce an MSE that far exceeds σ^4, perhaps even by an unbounded amount). Another flaw is that when the data has many outliers, then the estimator can be very sensitive to these outliers. Even though, this partial information can provide with us some guidelines for choosing between estimators.

- A scenario MSE can be useful is when facing a class of equivariante estimators.

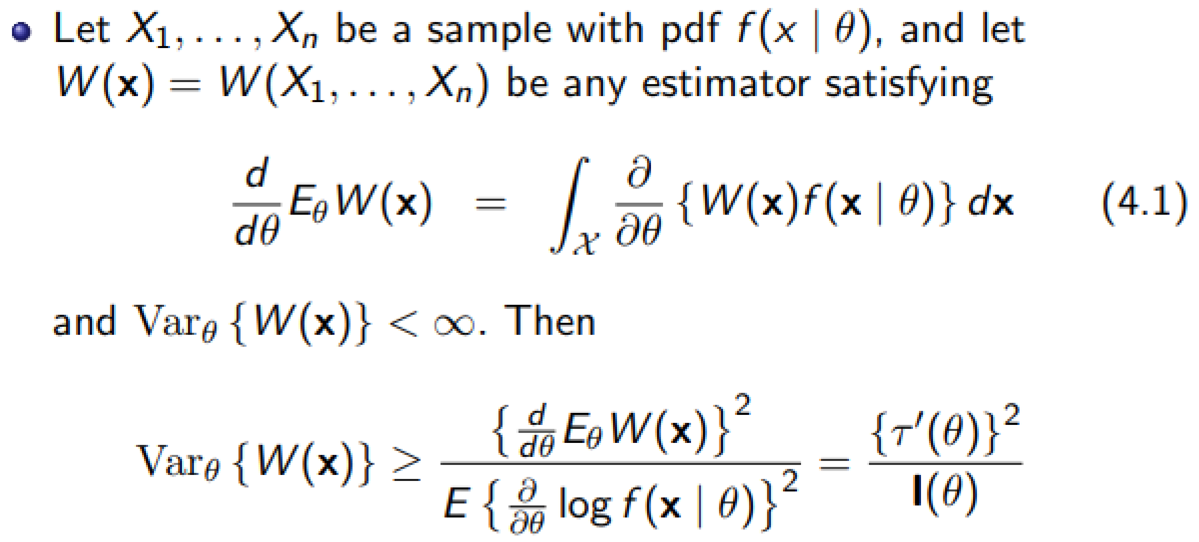
Best Unbiased Estimators

The motivation is that there is no best MSE estimator since the space is too large. In this part, we focus on those who are unbiased. Try to find the one with the smallest variance.

- "Best" is defined in the following way: An estimator W^* is an unbiased best estimator of $\tau(\theta)$ if it satisfies $E_\theta W^* = \tau(\theta)$, also, it's the one with the smallest variance.

Since finding the best unbiased estimator is hard, two ways are explored to achieve this goal:
(1) The first one is give the variance a lower bound and if we manage to find an estimator whose variance is just equal to the lower bound, then the best is found;
(2) Use the ss and cs we defined in chapter 6 and build a relationship of best estimator and cs.

Cramer-Rao Inequality



- Let $X_1, \ldots, X_n$ be a sample with pdf $f(x \mid \theta)$, and let $W(\mathbf{x}) = W(X_1, \ldots, X_n)$ be any estimator satisfying

$$\frac{d}{d\theta} E_\theta W(\mathbf{x}) = \int_{\mathcal{X}} \frac{\partial}{\partial\theta} \{W(\mathbf{x}) f(\mathbf{x} \mid \theta)\}\, d\mathbf{x} \qquad (4.1)$$

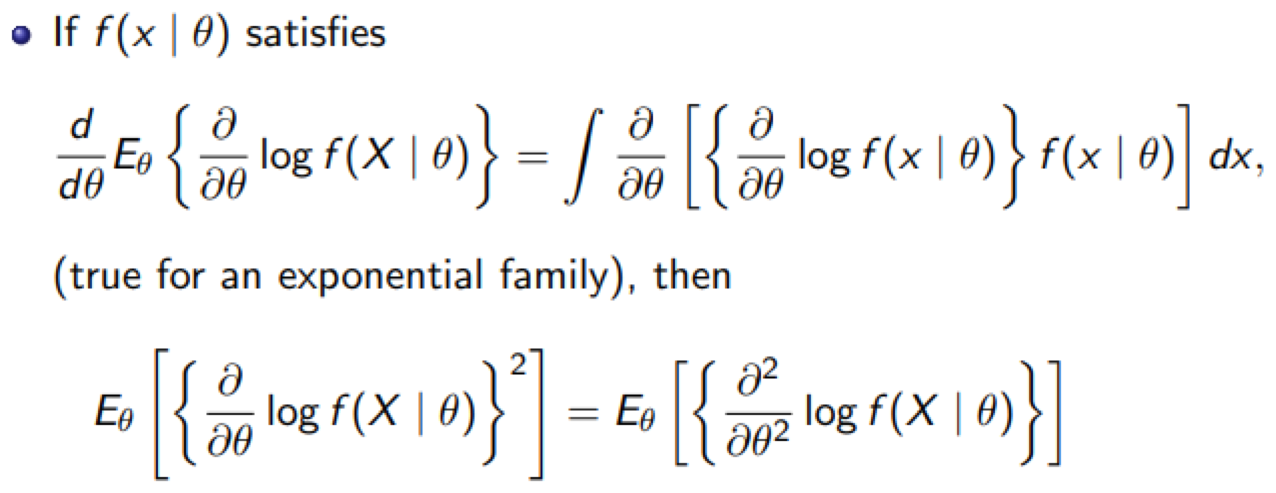
and $\mathrm{Var}_\theta\{W(\mathbf{x})\} < \infty$. Then

$$\mathrm{Var}_\theta\{W(\mathbf{x})\} \geq \frac{\left\{\frac{d}{d\theta} E_\theta W(\mathbf{x})\right\}^2}{E\left\{\frac{\partial}{\partial\theta} \log f(\mathbf{x} \mid \theta)\right\}^2} = \frac{\{\tau'(\theta)\}^2}{\mathbf{I}(\theta)}$$

The first condition just says that integration and derivative operation can be interchanged. The proof is by Cauchy-Schwarz ineq:

$$[Cov(X,Y)]^2 \leq (VarX)(VarY)$$

where X to be the estimator $W(X)$.
An useful lemma to simplify computation:



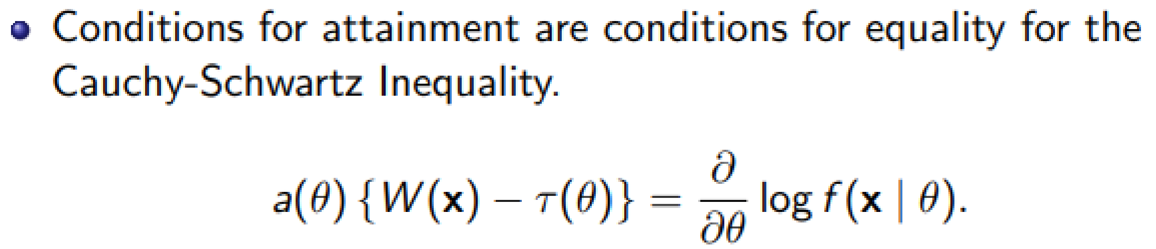
- If $f(x \mid \theta)$ satisfies

$$\frac{d}{d\theta} E_\theta \left\{ \frac{\partial}{\partial \theta} \log f(X \mid \theta) \right\} = \int \frac{\partial}{\partial \theta} \left[\left\{ \frac{\partial}{\partial \theta} \log f(x \mid \theta) \right\} f(x \mid \theta) \right] dx,$$

(true for an exponential family), then

$$E_\theta \left[\left\{ \frac{\partial}{\partial \theta} \log f(X \mid \theta) \right\}^2 \right] = E_\theta \left[\left\{ \frac{\partial^2}{\partial \theta^2} \log f(X \mid \theta) \right\} \right]$$

Then there is this corallay that helps us determine when this lower bound can be achieved.



- Conditions for attainment are conditions for equality for the Cauchy-Schwartz Inequality.

$$a(\theta) \left\{ W(\mathbf{x}) - \tau(\theta) \right\} = \frac{\partial}{\partial \theta} \log f(\mathbf{x} \mid \theta).$$

Using ss and cs to characterize bue

Main Theorem : Let T be a css for a parameter θ, and let $\phi(T)$ be any estimator based only on T. Then $\phi(T)$ is the unique best ubbiased estimator of its expected value.
How does this theorem conclude:

1. (Rao-Blackwell) Let W be any unbiased estimator of $\tau(\theta)$ and let T be a sufficient statistic for θ. Define $\phi(T) = E(W|T)$. Then $E_\theta \phi(T) = \tau(theta)$ and $Var_\theta \phi(T) \leq Var_\theta W$ for all θ; That is $\phi(T)$ is a uniformly better unbiased estimator of $\tau(\theta)$.

- Comment: R-B theorem implies that to find a ble, we just need to focus on the set of estimators which conditions on a ss, since conditioning on a ss will improve an estimator's performance on variance without alternating its mean.

2. If W is a bue of $\tau(\theta)$. Then W is unique.

- This result partly answers the question that if we find an estimator who achieves the CR lower bound, do we find anything problem.

3. If $E_\theta W = \tau(\theta)$, W is the best unbiased estimator of $\tau(\theta)$ iff W is uncorrelated with all unbiased estimator of 0.
-Comment: An estimator with mean 0 is purely a white noise. An estimator correlated to it implies

> the estimator itself constains noise in it. Intuitively thinking, extracting this noise out of the estimator will decrease the variance without hurting the useful information and an ideal bue should have this property.

Now we look at the main theorem, it simply says that given a css for a param, then any unbiased estimator conditioning on it is a best unbiased estimator of some kind. So this theorem transform the difficulty of finding bue to finding a css. Notice that the complete condition guarantees that there is no noise in the stats, sufficient makes sure that all the information about the param is provided. Thus, this theorem is an elegant bridge of complete-sufficient theory and bue.