4 Concepts from last lecture

- Likelihood Principle

- Sufficiency

- Principles of Bayesian Inference

$$\pi(\theta \mid \theta) \propto \pi(\theta) p(x \mid \theta)$$

4 Frequentist point estimation

Chapter 4 focuses on Frequentist point estimation and describes the principles behind it. We discuss some desirable properties of the estimators.

4.1 Review of criteria for good point estimators

For the moment we assume

- there is a single (scalar) unknown parameter θ of interest,
- we estimate θ using an estimator $T = T(X)$.
- Note that T is a function of X and therefore it is a random variable.

The quality of the estimator (a random variable) can be assessed by looking at the properties of its distribution: the sampling distribution of T.

Properties of particular interest are the mean, variance and mean squared error.

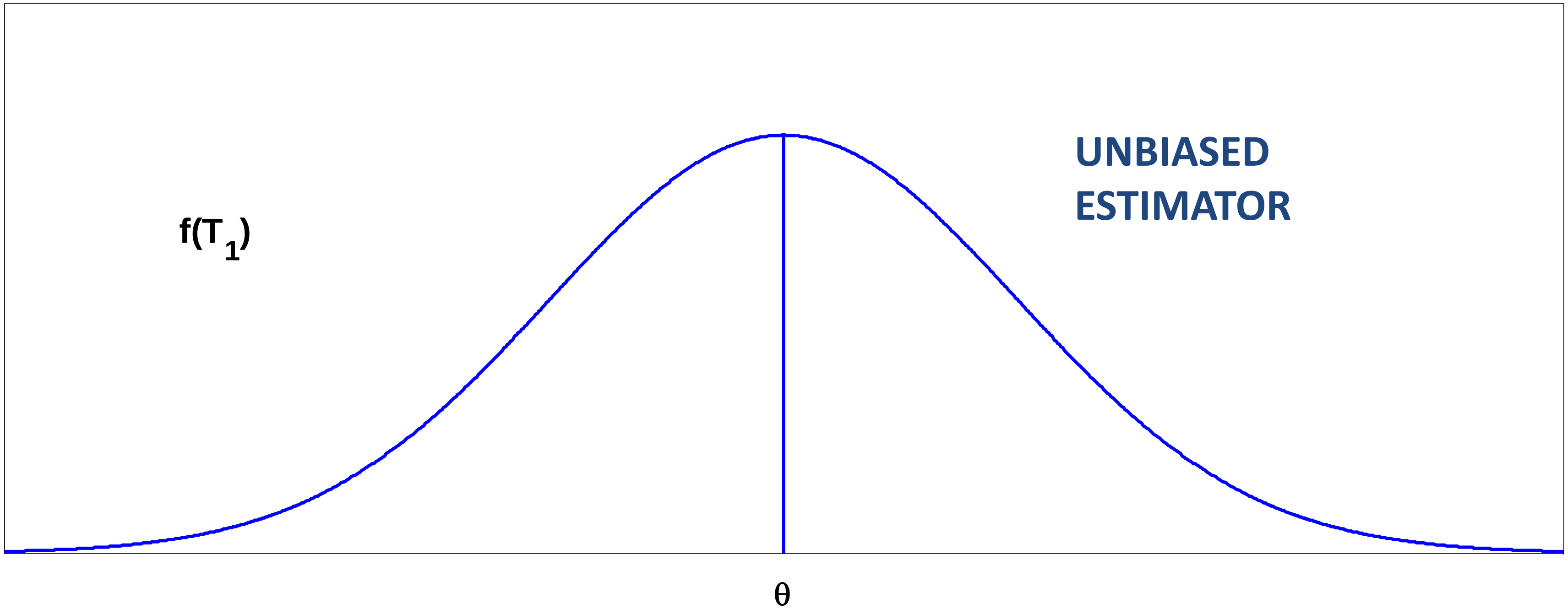
Unbiasedness

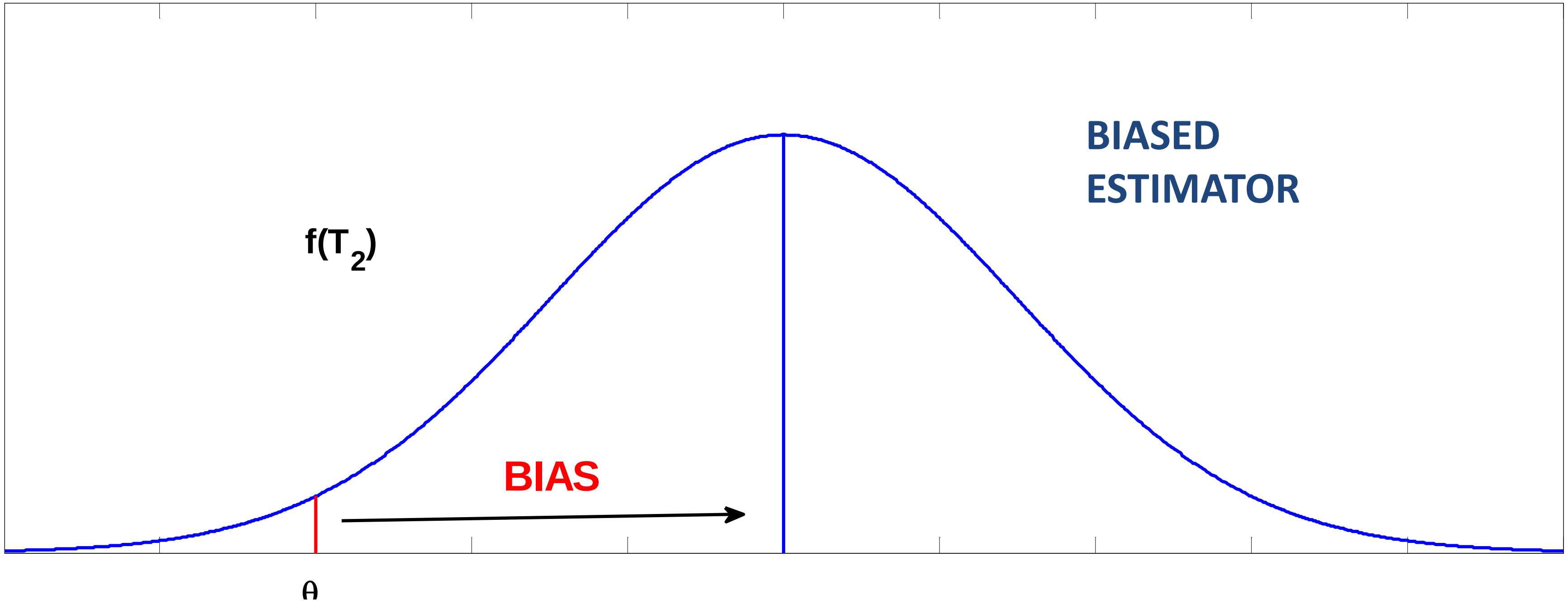
Let $T(X)$ be an estimator of θ, with $X = (X_1, \ldots, X_n)$

The difference E$(T) - \theta = b_T(\theta)$ is the bias of the estimator.

If $b_T(\theta) \equiv 0$ (i.e. $b_T(\theta)$ is zero for all values of θ) then T is an unbiased estimator of θ.

The expectation above is taken with respect the sampling distribution of $T(X)$.


f(T_1)
UNBIASED
ESTIMATOR
θ


f(T_2)
BIASED
ESTIMATOR
BIAS
θ

- **Unbiased estimators are 'fair' in the sense that they do not consistently either over- or under-estimate the unknown parameter.**

- **Unbiasedness is not an invariant property in that if T is an unbiased estimator of θ then, in general, $g(T)$ is a biased estimator of $g(\theta)$. For example, $\mathrm{E}(T^2) = \mathrm{var}(T) + \theta^2 > \theta^2$.**

- **Unbiased estimators for parametric functions can often be found by adjusting an 'obvious' estimator.**

Example 4.1

Suppose we have a random sample $(X_1, \ldots, X_n)$ from a population with mean μ and variance σ^2, and we wish to find an unbiased estimator of μ^2.

We know that $\overline{X}$ is an unbiased estimator of μ so we might guess that $\overline{X}^2$ would be suitable

Then

$$\begin{aligned}
E(X^2) &= \mu^2 + \sigma^2 \\
E(\overline{X}^2) &= E\left(\frac{1}{n^2}\sum_i \sum_j X_i X_j\right) = \frac{1}{n^2}\left\{nE(X^2) + n(n-1)(E(X))^2\right\} \\
&= \frac{1}{n}\left\{\sigma^2 + \mu^2 + (n-1)\mu^2\right\} = \frac{\sigma^2}{n} + \mu^2
\end{aligned}$$

Therefore $\overline{X}^2$ is a biased estimator of μ^2.

Based on the previous results we look for an unbiased estimator.

Consider

$$\begin{aligned} S^2 &= \frac{\sum_i (X_i - \overline{X})^2}{n-1} = \frac{1}{n-1}\left\{\sum_i X_i^2 - n\overline{X}^2\right\} \\ E(S^2) &= \frac{1}{n-1}\left\{n(\sigma^2+\mu^2) - n(\mu^2+\sigma^2/n)\right\} = \sigma^2 \end{aligned}$$

Combining this with the previous results we get

$$E(\overline{X}^2 - \frac{S^2}{n}) = \mu^2$$

Thus $\overline{X}^2 - \frac{S^2}{n}$ is unbiased for μ^2.

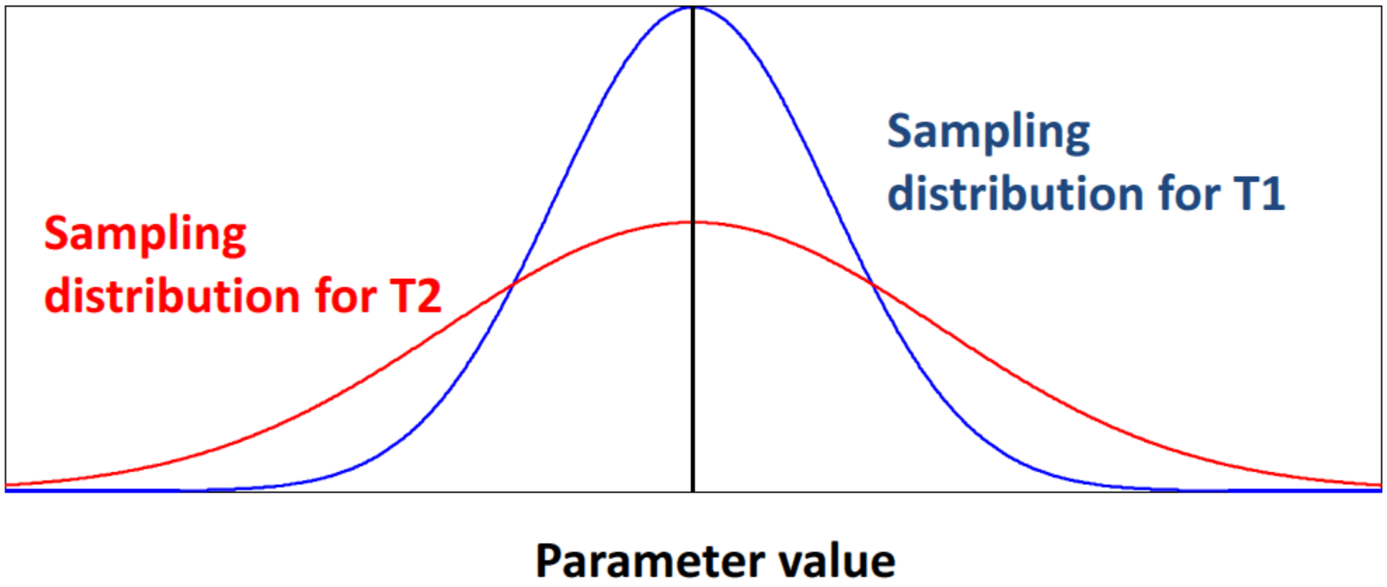
End of Example 4.1 ■

Example 4.1 shows a somewhat rough and ready, but effective, method of finding unbiased estimators;

- **guess at a sensible estimator,**
- **derive its expectation,**
- **then adjust the estimator accordingly to remove the bias.**

Mean squared error

Although unbiasedness is often regarded as desirable, there are cases where the values of a biased estimator, T', with a small variance will usually be closer to θ than those of the unbiased estimator, T, with a larger variance.



- Even if an estimator T is unbiased for θ, its distribution might be closely concentrated around θ or widely spread out.

- For example, if an unbiased estimator is equally likely to underestimate or overestimate θ by 1000000 units , it would never yield an estimate close to θ.

- The fact that an estimator is unbiased does not necessarily imply it is a good estimator.

- If an unbiased estimator of θ has also small variance, then its distribution will be concentrated around its mean θ and there will be high probability that the estimator will be close to θ.

- Therefore, the study of unbiased estimators is largely devoted to the search for unbiased estimator with small variance.

In order to formally assess the performance of a point estimator a quadratic loss function is often used (see Chapter 1, where the loss is taken to be the square of the estimation error.)

$$l(\theta, T) = (T - \theta)^2$$

The risk in using an estimator T of θ is then the loss averaged over the possible values of the estimator:

$$R(\theta, T) = E(l(\theta, T)) = E(T - \theta)^2 = \int_{\mathcal{X}} (T(x) - \theta)^2 p(x; \theta) \mathrm{d}x$$

Thus we might try to choose an estimator with a small mean square error (mse)

$$\text{mse}(T;\theta) = \mathrm{E}\{(T-\theta)^2\} = \text{var}(T) + \{b_T(\theta)\}^2$$

- For unbiased estimators the mse is the same as the variance.
- More generally, the mse criterion allows a trade-off between variance and bias.
- Note that $\text{mse}(T;\theta)$ is often a function of the unknown θ, so that in comparing T and T', it may not be the case that $\text{mse}(T;\theta) \leq \text{mse}(T';\theta)$ *for all* θ.

If $\mathrm{mse}(T;\theta) \leq \ \mathrm{mse}(T';\theta)$ does not hold *for all* θ then our choice of estimator would depend on the *unknown* θ, so that use of mean squared error as a criterion for comparison of estimators does not provide a complete answer to the problem.

We are not able to select a best estimator on the basis of mse or even always to select between estimators T and T'.

The mse concept does, however, enable us to eliminate some estimators. Suppose, for example, that the estimator T' is such that $\text{mse}(T';\theta) \geq \text{mse}(T;\theta)$ for all values of θ. Then we would never want to use T', since T is uniformly (in θ) preferable to T'.

In this case T' is said to be inadmissible.

An estimator that is not inadmissible is called admissible. An admissible estimator is one for which there is no uniformly better (in the sense of mse) estimator.

Most of the standard estimators are admissible.

Consistency

Exact unbiasedness arises largely for reasons of mathematical simplicity, and some common estimators are biased.

For example, if we wish to estimate the variance of the distribution of a random variable, given a random sample $X_1, \dots, X_n$, we usually use

$$S^2 = \frac{1}{n-1} \sum_{1=1}^{n} (X_i - \overline{X})^2$$

as it is unbiased. The more natural one $\frac{1}{n} \sum_{1=1}^{n} (X_i - \overline{X})^2$ is biased. It is obvious that as n increases the two estimators become closer.

Let us try to formalize this concept

Suppose that $\{T_n\}$ is a sequence of estimators of θ, where T_n is based on a simple random sample X_n of size n.

We would like $\{T_n\}$ to have the property that, as n increases (so that we obtain estimates based on larger and larger samples of data), the corresponding estimates t_n get increasingly close to θ.

An estimator is said to be (weakly) consistent for θ if, for all $\epsilon > 0$ and θ,

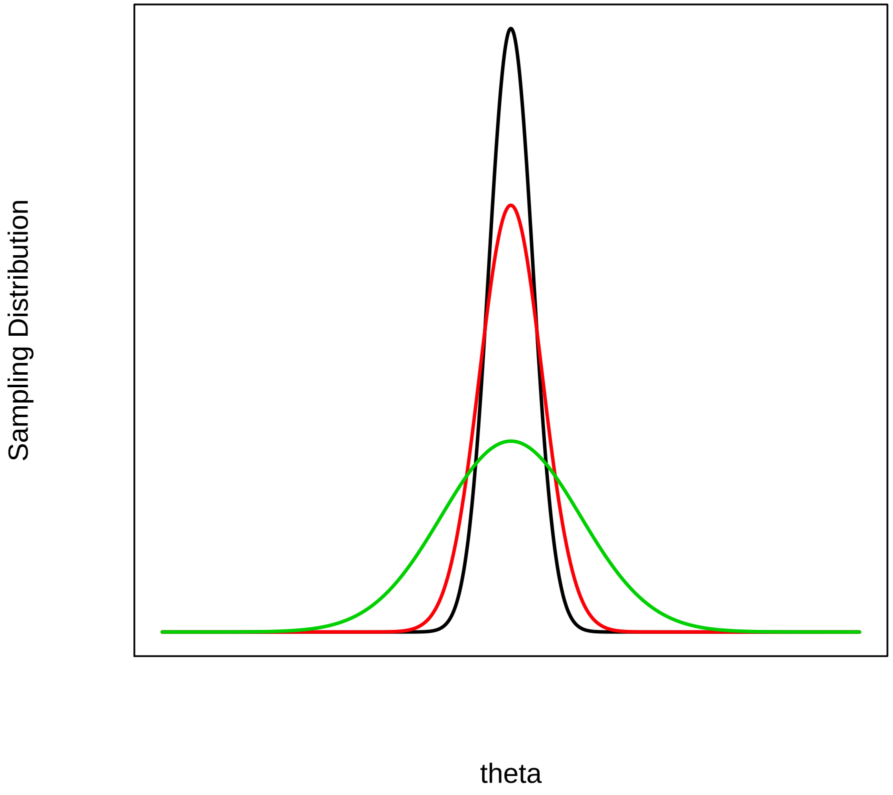
$$\mathbf{P}(\mid T_n(X_n) - \theta \mid < \epsilon) \to 1$$

as $n \to \infty$ (T_n converges in probability to θ).

T_n is said to be strongly consistent for θ if $T_n \to \theta$ with probability 1 (T_n converges almost surely to θ):

$$\Pr(\lim_{n\to\infty} T_n \to \theta) = 1$$

Consistency conveys the idea that the location of the sampling distribution of the estimator $T(X)$ is approximately correct for large n. It states that as $n \to \infty$, the sampling distribution of $T(X)$ becomes more and more concentrated around the true value θ.



It can be shown that if the mse($T_n;\theta$) $\to$ 0 as $n \to \infty$, which requires that both b_{T_n} and var(T_n) tend to zero, then T_n will be (weakly) consistent for θ. (This is a sufficient condition but it is not necessary.)

Example 4.2

Suppose $X_i \overset{iid}{\sim}$ Exp(λ) with $i = 1, \ldots, n$

Let $\mu = 1/\lambda$. We want to estimate μ.

We have

$$\begin{aligned} \mathcal{L}(\lambda) &= p(x;\theta) = \prod_i \lambda \exp\{-\lambda x_i\} = \lambda^n \exp\left\{-\lambda \sum_i x_i\right\} \\ l(\lambda) &= \log \mathcal{L}(\lambda) = n \log \lambda - \lambda \sum_i x_i \end{aligned}$$

$\Longrightarrow \overline{X}$ is the mle of μ and it is unbiased, with

$$\text{Var}(\overline{X}) = \frac{1}{n\lambda^2}$$

Var$(\overline{X}) \to 0$ as $n \to \infty$ so $\overline{X}$ is consistent for μ.

Consider now $X_{(1)} = \min X_i$. Then

$$\Pr(X_{(1)} > x) = \Pr(X_i > x, i = 1, \ldots, n) = \prod_i \Pr(X_i > x) = \exp\{-n\lambda x\}$$

$\Longrightarrow X_{(1)} \sim$ Exp$(n\lambda)$ and $E(X_{(1)}) = \frac{1}{n\lambda} = \frac{\mu}{n}$

therefore $nX_{(1)}$ is also unbiased for μ but

$$\text{Var}(nX_{(1)}) = n^2 \frac{1}{n^2\lambda^2} = \mu^2$$

so $nX_{(1)}$ is not consistent for μ.

End of Example 4.2 ■

Minimum variance

Among unbiased estimators, the mean squared error is equivalent to the variance. If T and T' are distinct unbiased estimators for a parameter θ then we would prefer T if var$(T) <$ var(T').

Definition: If var$(T) <$ var(T'), T is said to be *more efficient* than T'.

The efficiency of T' relative to T is simply the variance of T expressed as a percentage of that of T', that is 100var(T)/var(T')%.

Example: if var$(T) = 1.2$ and var$(T') = 1.8$, then the efficiency of T' relative to T is 67%. We may also consider the efficiency of an estimator relative to the Cramér-Rao lower bound, to be discussed in the next section.

Example 4.3

Suppose that $X \sim \mathrm{N}(\theta, 4)$ and $Y \sim \mathrm{N}(\theta, 9)$, with X and Y independent.

Consider 2 estimators of θ: $U = (X+Y)/2$ and $V = (X+2Y)/3$. We have

- $E(U) = E(V) = \theta$, both unbiased
- $\text{var}(U) = \frac{1}{4}(4+9) = 13/4 = 3.25$
- $\text{var}(V) = \frac{1}{9}(4+4\times 9) = \frac{40}{9} = 4.4$
- U is preferred to V having uniformly smaller variance.
- The efficiency of V relative U is $\frac{13}{4}\frac{9}{40} \times 100 \approx 73\%$

Consider the class of all linear functions of X and Y that are unbiased estimators of θ. Then

$$\begin{aligned} W &= aX + (1-a)Y \qquad \text{for} 0 \leq a \leq 1 \\ \text{var}(W) &= a^2 4 + (1-a)^2 9 \end{aligned}$$

What a gives the smallest variance?

$$\frac{\text{d}}{\text{d}a}\text{var}(W) = 8a - 18(1-a)$$

$\Longrightarrow$ var(W) is minimised when $a = 18/26 = 9/13 \approx 0.692$ and therefore the best linear unbiased estimator is $\frac{9}{13}X + \frac{4}{13}Y$ which has variance 2.77.

The efficiency of U and V relative to W are 85% and 63% respectively.

End of Example 4.3 ■

Note that minimising the variance of an estimator is, by itself, not a useful criterion.

If $T(x) = c$ for some constant c, regardless of the data x, then var$(T) = 0$ for all θ, but T will usually be a very poor estimator of θ!

4.2 Information and Variance Bounds

The results and proofs in this section depend on certain regularity conditions which justify changing the order of differentiation and summation/integration.

In particular, this requires that the range of the data should not be a function of the unknown parameter θ. For convenience, we give the proofs below in the continuous case, i.e. when the random variable X has a joint density.

The Score Function

We have denoted the log-likelihood function by $\ell(\theta; x) = \ln \mathcal{L}(\theta; x)$. In this section, we study its derivative V evaluated at X, defined by

$$V(X) = \frac{\partial}{\partial\theta}\,\ell(\theta; X) = \frac{\partial}{\partial\theta}\,\ln \mathcal{L}(\theta; X) = \frac{1}{p(X;\theta)}\frac{\partial}{\partial\theta} p(X;\theta)\,,$$

which is therefore a random variable. This variable is often called the score function.

Mean of The Score Function

Start from the fact that

$$1 = \int p(x;\theta)dx.$$

Differentiating with respect to θ :

$$0 = \int \frac{\partial}{\partial\theta} p(x;\theta)dx = \int p(x;\theta)\frac{\partial}{\partial\theta}\ln p(x;\theta)dx = \mathrm{E}(V). \qquad (\star\star)$$

Thus V has mean zero.

It follows that

$$\text{var}(V) = \text{E}(V^2) = \mathrm{E}\left[\left\{\frac{\partial}{\partial\theta}\ln p(X;\theta)\right\}^2\right],$$

where var(V) is called the Fisher information in the sample, and is denoted by $I(\theta)$.

Differentiating $(\star\star)$ again with respect to θ (the right hand integral), we have

$$\begin{aligned} 0 &= \int \frac{\partial}{\partial\theta} p(x;\theta) \frac{\partial}{\partial\theta} \ln p(x;\theta) dx + \int p(x;\theta) \frac{\partial^2}{\partial\theta^2} \ln p(x;\theta) dx \\ &= \int p(x;\theta) \left\{ \frac{\partial}{\partial\theta} \ln p(x;\theta) \right\}^2 dx + \int p(x;\theta) \frac{\partial^2}{\partial\theta^2} \ln p(x;\theta) dx \\ &= \mathrm{E}(V^2) + \mathrm{E} \left\{ \frac{\partial^2}{\partial\theta^2} \ln p(x;\theta) dx \right\}. \end{aligned}$$

Thus

$$\mathrm{var}(V) = \mathrm{E}(V^2) = \mathrm{E} \left\{ -\frac{\partial^2}{\partial\theta^2} \ln p(x;\theta) \right\},$$

which is usually simpler to compute.

- **The Fisher information $I(\theta)$ is a measure of the information in the data x about the unknown parameter θ.**

- **It is additive, in the sense that if x_1 and x_2 are two independent samples whose distributions, $p^{(1)}$ and $p^{(2)}$ say, both depend on the same unknown parameter θ, then $I(\theta; x_1, x_2) = I^{(1)}(\theta; x_1) + I^{(2)}(\theta; x_2)$.**

- **In particular, if $(x_1, \ldots, x_n)$ is a simple random sample of size n, then the information $I_n(\theta)$ based on a sample of size n satisfies $I_n(\theta) = ni(\theta)$, where $i(\theta)$ is Fisher's information in the observation X_i.**

Suppose that we are interested in the parameter $\phi = g(\theta)$.

Then, given data x, the likelihood $\mathcal{L}^*$ of ϕ satisfies $\mathcal{L}^*(\phi; X) = \mathcal{L}(\theta; X)$.

Differentiating with respect to ϕ, we obtain

$$\frac{\partial}{\partial\phi} \ln \mathcal{L}^*(\phi; X) = \frac{\partial}{\partial\theta} \ln \mathcal{L}(\theta; X)\frac{\partial\theta}{\partial\phi}$$

where $\frac{\partial\theta}{\partial\phi} = \{\frac{\partial\phi}{\partial\theta}\}^{-1} = \{g'(\theta)\}^{-1}$.

Thus we see that the information $I^*(\phi)=$ E$[\{\frac{\partial}{\partial\phi} \ln \mathcal{L}^*(\phi; X)\}^2]$ about ϕ in the sample is given by

$$I^*(\phi) = I(\theta)/\{g'(\theta)\}^2$$

The Cramér-Rao lower bound

Suppose that T is an unbiased estimator of $m(\theta)$, i.e. E$(T) = m(\theta)$. Then

$$m(\theta) = \mathrm{E}(T) = \int T(x)p(x;\theta)dx.$$

Differentiating with respect to θ:

$$\begin{aligned} m'(\theta) &= \int T(x)\frac{\partial}{\partial\theta}p(x;\theta)dx \\ &= \int T(x)\frac{\partial}{\partial\theta}\ln p(x;\theta)p(x;\theta)dx \\ &= \mathrm{E}(TV) \\ &= \mathrm{cov}(T,V) \text{ since } \mathrm{E}(V) = 0 \text{ and } \mathrm{E}(V)\mathrm{E}(T) = 0. \end{aligned}$$

NB: T and V are, in general, dependent random variables since both are functions of X.

But, corr$^2(T,V) \leq 1$ or, equivalently, $\text{cov}^2(V,T) \leq \text{var}(V)\,\text{var}(T)$, and so

$$\text{var}(T) \geq \frac{[m'(\theta)]^2}{\text{var}(V)} = \frac{[m'(\theta)]^2}{I(\theta)}.$$

This inequality is known as the Cramér–Rao lower bound.

If T is unbiased for θ, then cov$(V,T) = 1$ and the lower bound is simply $1/I(\theta)$ (since $m'(\theta) = 1$ and formula in previous slide).