Stat 200 : February 28th,2001

Summary of preceding lecture:
Efficiency and Cramer-Rao lower bound tell us how much variability we can expect from an unbiased estimator.

Properties of estimators ; after efficiency and a lower bound on an estimator's variance ($\frac{1}{nI(\theta)}$), I will introduce the notion of sufficiency of an estimator, if an estimator is sufficient for a parameter θ we can compute just that estimate and throw away all the other data.

Definition:
A statistic is that it is sufficient iff the conditional distribution (density or frequency) of the vector $\underline{X}$ given $T = t$, does not depend on θ for any value of $T = t$.
Neither in the fucntion, nor in the domain.

Forr iid samples, as is usually the case, this says:

$$\frac{f(x_1|t)f(x_2|t)\dots f(x_n|t)}{f_T(t)}$$

does not involve θ .

The binomial is the typical example:
$X_1, \dots, X_n$ a sequence of iid Bernouilli rv's, with $P(X = 1) = \theta$. Then $T = \sum_{i=1}^n X_i$ is sufficient for θ .

$$P(X_1 = x_1, X_2 = x_2, \dots, X_n = x_n|T = t) = \frac{P(X_1 = x_1, X_2 = x_2, \dots, X_n = x_n, T = t))}{P(T = t)}$$

$$P(X_1 = x_1, X_2 = x_2, \dots, X_n = x_n, T = t) = \begin{cases} 0 & \text{if } \sum_{i=1}^n x_i \neq t \\ \prod_{i=1}^n \theta^{x_i}(1-\theta)^{1-x_i} & \text{otherwise} \end{cases}$$

So

$$\frac{\theta^t(1-\theta)^{n-t}}{P(T=t)} = \frac{\theta^t(1-\theta)^{n-t}}{\binom{n}{k}\theta^t(1-\theta)^{n-t}} = \frac{1}{\binom{n}{k}} = \frac{t!(n-t)!}{n!}$$

This does not depend on θ. Here is a necessary and sufficient condition for sufficiency:

Theorem 8.1 *A necessary and sufficient condition for* $T(\underline{X}) \equiv T(X_1, X_2, \dots, X_n)$ *to be sufficient for a parameter is that the joint distribution (density or frequency) factors into two parts, one that depends on* $\hat{\theta}$ *and on* $\underline{x}$ *only through* $T(\underline{x})$ *the other that does not depend on* θ *:*

$$f(x_1, x_2, \dots, x_n|\theta) = g[T(x_1, x_2, \dots, x_n), \theta]h(x_1, x_2, \dots, x_n)$$

or

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

Proof: the condition is sufficient, i.e. if we have the condition we will have sufficiency.

First partition

$$
\begin{aligned}
P(T=t) &= \sum_{T(\underline{X})=t} P(\underline{X}=\underline{x}) \\
&= g(t,\theta) \sum_{T(\underline{X})=t} h(\underline{x}) = g(t,\theta) H(\underline{x}) \\
P(\underline{X}=\underline{x}|T=t) &= \frac{P(\underline{X}=\underline{x}, T=t)}{P(T=t)} \\
&= \frac{h(\underline{x}) g(t,\theta)}{H(\underline{x}) g(t,\theta)}
\end{aligned}
$$

Cancellation giving the result.

The other direction, i.e. sufficency implies the condition: T is sufficient for θ means we can write: $P(\underline{X}=\underline{x}|T=t)$ as a function of $\underline{x}$, call it h: $P(\underline{X}=\underline{x}|T=t)=h(\underline{x})$, we then have:

$$P(\underline{X}=\underline{x}|\theta) = P(\underline{X}=\underline{x}|T=t)P(t=t|\theta) = h(\underline{x})g(t,\theta)$$

8.7.1 Exponential Families

Probability distributions with sufficient statistics the same dimension as the parameter space, regardless of sample size. One paarameter families:

$$f(x|\theta) = exp[c(\theta)K(x) + d(\theta) + S(x)]$$

Joint density of an iid sample from this distribution will be :

$$f(\underline{x}|\theta) = \prod exp[c(\theta)K(x_i) + d(\theta) + S(x_i)]$$

$$= exp[c(\theta)\sum K(x_i) + nd(\theta)]exp[\sum S(x_i)]$$

So that $T(\underline{x}) = \sum K(x_i)$ is a sufficient statistic.

8.7.2 Bernouilli Example

$P(X=x) = \theta^x(1-\theta)^{1-x} = exp[x log(\frac{\theta}{1-\theta}) + log(1-\theta)]$ $K(x)=x, T=\sum X_i$ is the sufficient statistic.

The form of the density of an m-parameter exponential family:

$$f(x|\theta) = exp[\sum_{i=1}^{m} c_i(\theta)K_i(x) + d(\theta) + S(x)], \qquad x \in A$$

A must not depend on $\hat{\theta}$ either.

8.7.3 Normal Example

$$\begin{aligned} f(x|\mu,\sigma) &= \prod \frac{1}{\sigma\sqrt{2\pi}} exp[-\frac{1}{2\sigma^2}(x_i-\mu)^2] \\ &= \frac{1}{\sigma^n 2\pi^{\frac{n}{2}}} exp[-\frac{1}{2\sigma^2}(\sum_{i=1}^n x_i^2 - 2\mu\sum_{i=1}^n x_i + n\mu^2)] \end{aligned}$$

This is only a function of $\sum_{i=1}^n x_i$ and $\sum_{i=1}^n x_i^2$, thus they are sufficient statistics. Dimension of sufficient statistic= 2= dimension of parameter space : exponential family.

Corollary of the factorization theorem:
If T is sufficient for θ the mle is a function of T.

Proof:
The mle is built by maximising $f(\underline{x}\,|\theta\,)$ which can be factored as: $g(T,\theta\,)h(\underline{x}\,)$ the dependence on θ is only through T. To maximise this we only need to look at $g(T,\theta\,)$.

The following quantifies how much better it can be to use a sufficient statistic as a basis for an estimator, it always provides a method for improving an estimator.

Theorem 8.2 (Rao Blackwell) *Let $\hat{\theta}$ be any finite-varianced estimator of θ . Suppose that we have a sufficient statistic for θ we call T. Now taking as a new estimate $\tilde{\theta} = E(\hat{\theta}\,|T)$ we will have a better estimator because it has smaller MSE:*

$$E(\tilde{\theta} - \theta\,)^2 \leq E(\hat{\theta} - \theta\,)^2$$

The equality is strict unless $\hat{\theta} = \tilde{\theta}$.

Proof:
Uses the conditional expectation and variance formulas:

$$\begin{aligned} E(E(Y|X)) &= E(Y) \\ Var(Y) &= Var(E(Y|X)) + E(Var(Y|X)) \\ E(\tilde{\theta}\,) &= E(\hat{\theta}\,) \\ Var(\hat{\theta}\,) &= Var(E(\hat{\theta}\,|T)) + E(Var(\hat{\theta}\,|T)) \\ Var(\hat{\theta}\,) &= Var(\tilde{\theta}\,) + E(Var(\hat{\theta}\,|T)) \end{aligned}$$

Example of Rao-Blackwellisation:
$X_1, X_2, \ldots X_n \sim \mathcal{N}(\theta\,,\sigma^2)$ we want to estimate θ , using the silly estimate : $g(\underline{X}\,) = X_1$, and we know a sufficient statistic: $X_1 + X_2 + \cdots X_n$. Then the Rao-Blackwellisation would give us :

$$E[X_1|X_1 + X_2 + \cdots + X_n] = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

Because $E(X|X+Y) + E(Y|X+Y) = 2E(X|X+Y) = E(X+Y|X+Y) = X+Y$. So just the one step of conditionning on a sufficient statistic took us a long way.

Extension to other loss functions than the MSE, any convex $W(\tilde{\theta}\,,\theta\,)$ is such that Rao-Blackwellisation makes things better.

Example:

$$X_1, X_2, \ldots, X_n \sim \mathcal{N}(0,\sigma^2)$$

. and we ant to estimate θ . Estimator: first observation: X_1, why is this silly?

$$\hat{g}(\underline{X}) = X_1.$$

But we have a sufficient statistic: $X_1 + X_2 + \cdots X_n$.

$$E[X_1|X_1 + X_2 + X_3 + \cdots + X_n] = \frac{X_1 + X_2 \cdots X_n}{n}$$

In one step of conditionning we can make things much better.

Extension to other loss functions: $Jensens\ Inequality$

$$E(f(x)) \geq f(E(x))$$

Suppose we have a convex loss function $W(\tilde{\theta}, \theta)$.

$$E[W(\hat{\theta}, \theta)|T] \geq W(E(\hat{\theta}|T), \theta) = W(\tilde{\theta}, \theta)$$

$$E[W(\hat{\theta}, \theta)] \geq E[W(\tilde{\theta}, \theta)]$$

15 Decision Theory

Choose an action a from a set A, based on the observation of a random variable X which has a distribution depending on a parameter (state of nature) θ .

The decision $\mathbf{d}$ maps the sample space onto the the action space, $a = \mathbf{d}(X)$.

A loss $l(\theta, \mathbf{d}(X))$ depends on θ and $\mathbf{d}(X)$. Comparinf different decisions is based on the risk, or expected loss.

$$R(\theta, \mathbf{d}) = E[l(\theta, \mathbf{d}(X))]$$

We have just seen, a very detailed account of estimation as a decision, and mostly we used as our loss functionm the quadratic function, thus the risk is the MSE.

Finding the best $\mathbf{d}$ is not trivial, there might be two different states of nature, (parameter values) that give different orderings for the risks.

Two ways to address this:

- Minimax:
  The worst the risk could be is
  $$\max_{\hat{\theta} \in \Theta}[R(\theta, \mathbf{d})]$$
  Choose the decision function $\mathbf{d}^*$ that minimizes that worst case.
  $$\min_{\mathbf{d}} \left\{ \max_{\hat{\theta} \in \Theta}[R(\theta, \mathbf{d})] \right\}$$
- Bayes.