Quick Announcement:

- Write your submission ID for the class on the **top-left side** of all future homeworks (including homework 4)
- You can check your submission ID in the **HW 3 Grade - comment section** on BCourses.

(This will help us significantly in sorting future homeworks and save us a lot of time)

STAT 135 Lab 4
Recap, Efficiency, Sufficiency, Bias-Variance trade-off (Rao-Blackwell), Bayesian Inference

Rebecca Barter

February 23, 2015

Recap (what have we done so far)

Recap

Suppose we have a sample $X_1, ..., X_n$ of IID observations drawn from some population which has an underlying distribution, $F(\theta)$.

We know that $X_i \sim F(\theta)$, where we might have, for example:

- $F(\theta) = N(\mu, \sigma^2)$, in which case $\theta = (\mu, \sigma^2)$, or $\theta = \mu$
- $F(\theta) = Binomial(n, p)$, in which case $\theta = p$
- etc

Suppose that we know what the distribution family is (we know F) but we don't know the value of the parameter (**we don't know θ**).

Recap

Statistical inference is the process of deducing properties of the underlying distribution, $F(\theta)$, of a population by analysis of data $(X_1, ..., X_n)$ taken from the population.

So far, we have inferred properties about a population by **deriving estimates** (e.g. MLE, MOM) for the parameters (θ) of the underlying distribution $(F(\theta))$ of the population.

We found that if we use the maximum likelihood procedure to estimate the true parameter, θ_0, then when we have a large sample size this estimator has the following distribution

$$\hat{\theta}_{MLE} \sim N\left(\theta_0, \frac{1}{nI(\theta_0)}\right)$$

Recap

We found that if we use the maximum likelihood procedure to estimate the true parameter, θ_0, then when we have a large sample size this estimator has the following distribution

$$\hat{\theta}_{MLE} \sim N\left(\theta_0, \frac{1}{nI(\theta_0)}\right)$$

- θ_0 is a fixed (non-random), unknown property of the population
- $\hat{\theta}_{MLE}$ is a **random variable** since it is calculated from the random sample $X_1, ..., X_n$

Efficiency

Efficiency

Recall that

$$Var(\hat{\theta}_{MLE}) = \frac{1}{nI(\theta)}$$

How do other estimators compare with the MLE? It turns out that, if $\hat{\theta}$ is any other **unbiased** estimator for θ (if our density satisfies smoothness conditions), then

$$\boxed{Var(\hat{\theta}) \geq \frac{1}{nI(\theta)}}$$

This is the **Cramer-Rao lower bound**, and it says that there is no unbiased estimator that achieves smaller variance than the asymptotic MLE.

- Does this mean that the MLE is the best (for large samples)? Is it only the variance that we want to minimize?

Efficiency

One way to compare two estimators (say $\hat{\theta}$ and $\tilde{\theta}$) of θ, is to compare their variance.

- If $\hat{\theta}$ has lower variance than $\tilde{\theta}$, then we say it is **more efficient**.

Thus we can define the **efficiency of $\hat{\theta}$ relative to $\tilde{\theta}$** to be

$$eff(\hat{\theta}, \tilde{\theta}) = \frac{Var(\tilde{\theta})}{Var(\hat{\theta})}$$

What is the variance of the most efficient unbiased estimator?

Efficiency

The MLE asymptotically achieves the lowest variance of all unbiased estimators.

We thus say that an unbiased estimator, $\hat{\theta}$ is **efficient** if its variance is equal to

$$Var(\hat{\theta}) = \frac{1}{nI(\theta)}$$

i.e. if it achieves the Cramer-Rao lower bound.

- so the **MLE is asymptotically efficient**

Exercise

Exercise: Efficiency

Suppose we are interested in modeling the distance between mutations on a DNA strand. There are so many mutations on any given DNA strand that it is impossible to look at them all, so we take a sample of distances, $X_1, ..., X_n$. It is known that the distances are IID and follow an $Exponential(1/\lambda)$ distribution, but we don't know λ. We want to use our sample $X_1, ..., X_n$ to estimate the value of λ.
The density for the exponential distribution with parameter λ^{-1} is given by

$$f(x) = \frac{1}{\lambda} e^{-\frac{x}{\lambda}} \quad , \quad x > 0$$

1. Show that $\hat{\lambda} = nX_{(1)} = n \min(X_1, ..., X_n)$ is an unbiased estimator of λ
2. Consider the estimator $\tilde{\lambda} = \bar{X}$, and find the efficiency of $\hat{\lambda}$ relative to $\tilde{\lambda}$

Sufficiency

Sufficiency

The technical definition of sufficiency is:

> *A statistic $T(X_1, ..., X_n)$ is said to be* **sufficient** *for θ if the conditional distribution of $X_1, ..., X_n$, given $T = t$, does not depend on θ for any value of t*

where a *statistic* is defined to be any function of our sample $X_1, ..., X_n$. Examples (of which there are many!) include:

- The sample mean: $T = \bar{X}_n = \frac{\sum_{i=1}^{n} X_i}{n}$
- The sample variance: $T = Var(X) = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X}_n)^2$
- The sample maximum: $T = \max\{X_1, ..., X_n\}$
- $T = 5$ (technically, this is still a function of $X_1, ..., X_n$)

Sufficiency

We say a statistic T is an estimator of a population parameter θ, if T is usually close to θ. For example,

- The sample mean is an estimator for population mean
- The sample variance is an estimator for the population variance

Sufficiency

A more heuristic definition of a sufficient statistic:

Suppose that $X_1, ..., X_n \sim F(\theta)$. *Then* $T(X_1, ..., X_n)$ *is a* **sufficient statistic** *for* θ *if the statistician who knows the value of* T *can do just a good a job of estimating the unknown parameter* θ *as the statistician who knows the entire random sample.*

Sufficiency

Suppose that we just need to show that a given $T(X_1, ..., X_n)$ is sufficient. One way to do this is to calculate

$$P(X_1 = x_1, ..., X_n = x_n | T(X_1, ..., X_n))$$

and show that it is independent of θ.

But what if we don't already know $T(X_1, ..., X_n)$? There is an extremely useful theorem to 1) find sufficient statistics, and 2) to show that a given statistic is sufficient

Sufficiency

The factorization theorem

A necessary and sufficient condition for T to be sufficient for θ is that

$$f_\theta(x_1, ..., x_n) = g_\theta(T)h(x_1, ..., x_n)$$

i.e. that the density can be factored into a product such that one factor, h, does not depend on θ, and the other factor, which does depend on θ, depends on $(x_1, ..., x_n)$ only through T

- note that T is **not unique**.

Sufficiency

So, in summary, we can show that a statistic T is sufficient for θ by either:

1. Calculating

$$P(X_1 = x_1, ..., X_n = x_n | T(X_1, ..., X_n))$$

   and show that it is independent of θ, or
2. Show that the density can be factorized as

$$f_\theta(x_1, ..., x_n) = g_\theta(T)h(x_1, ..., x_n)$$

If we don't already know T, then the second approach can be used to find sufficient statistics since any function, $T(X_1, ..., X_n)$, satisfying the factorization property must be a sufficient statistic.

Exercise

Sufficiency: Exercise (Problem 21 of Section 8.10 of Rice)

Suppose that $X_1, ..., X_n$ are i.i.d with density function

$$f(x|\theta) = e^{-(x-\theta)}, \quad x \geq \theta$$

and $f(x|\theta) = 0$ otherwise.

1. Find the method of moments estimate of θ
2. Find the MLE of θ (Hint: be careful, and don't differentiate before thinking. For what values of θ is the likelihood positive?)
3. Find a sufficient statistic for θ

The Bias-Variance tradeoff

The Bias-Variance Tradeoff

- Note that previously we assessed the quality of an estimator in terms of efficiency (where we said that an estimator with lower variance is more efficient than an estimator with larger variance)

What other properties of an estimator might we want to consider?

The Bias-Variance Tradeoff

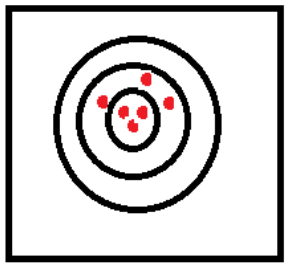
The two things we typically want to minimize are:

- **Bias**: we want our estimator, $\hat{\theta}$, on average, to be close to the true value, θ.
- **Variance**: we don't want our estimator, $\hat{\theta}$ to take values too far from the expected value, $E(\hat{\theta})$.

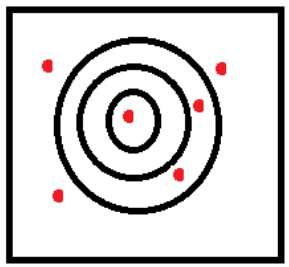
The Bias-Variance Tradeoff

However, there exists a relationship between bias and variance, such that in general

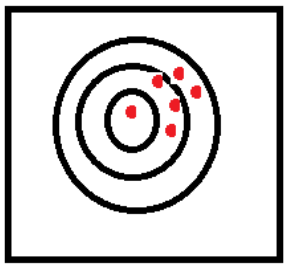
- Small variance (good) typically corresponds to high bias (bad)
- Small bias (good) typically corresponds to large variance (bad)



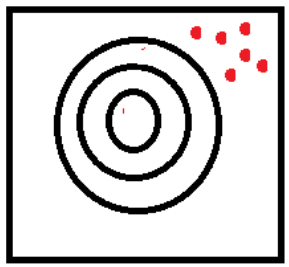
Small Variance - Small Bias



Large Variance - Small Bias



Small Variance - Large Bias



Small Variance - Huge Bias

The Bias-Variance Tradeoff

This bias-variance trade-off is captured in the Mean-Squared error

$$
\begin{aligned}
MSE(\hat{\theta}) &= E((\hat{\theta} - \theta)^2) \\
&= E\left(\hat{\theta}^2 - 2\theta\hat{\theta} + \theta^2\right) \\
&= E(\hat{\theta}^2) - 2\theta E(\hat{\theta}) + \theta^2 \\
&= Var(\hat{\theta}) + (E(\hat{\theta}))^2 - 2\theta E(\hat{\theta}) + \theta^2 \\
&= Var(\hat{\theta}) + \left(E(\hat{\theta}) - \theta\right)^2
\end{aligned}
$$

$$
\boxed{MSE(\hat{\theta}) = Var(\hat{\theta}) + \text{bias}(\hat{\theta})^2}
$$

So since we want both the variance and the bias to be small, the quantity that we really want to minimize is the MSE!

The Bias-Variance Tradeoff

We can now tie together the concept of sufficiency and MSE using the **Rao-Blackwell theorem**:
Suppose that

- $\hat{\theta}$ is an estimator for θ such that $E\left(\hat{\theta}^2\right) < \infty$.
- T is a sufficient statistic for θ.

If we define a new estimator to be

$$\tilde{\theta} = E\left(\hat{\theta} \middle| T\right)$$

Then

$$MSE(\tilde{\theta}) \leq MSE(\hat{\theta})$$

i.e. if we know a sufficient statistic T, and we have an estimator $\hat{\theta}$, then we can define a better estimator, $\tilde{\theta}$, for θ which has smaller MSE.

Exercise

Exercise: Rao-Blackwell

Suppose that $X_1, ..., X_n \sim Poisson(\lambda)$, then suppose that we are interested in the parameter $\theta = e^{-\lambda}$. Recall that the Poisson pmf is given by

$$P(X = x) = \frac{e^{-\lambda}\lambda^x}{x!}$$

Note that $\theta = P(X_1 = 0)$. Then

1. Show that $\sum_{i=1}^{n} X_i$ is a sufficient statistic for θ.
2. Show that $\hat{\theta} = \mathbb{1}_{(X_1=0)}$ is an unbiased estimator for θ.
3. Use the Rao-Blackwell theorem to find an estimator for θ with smaller MSE than $\hat{\theta}$.

Bayesian Inference

Bayes Theorem: $$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Bayesian Inference

- So far we have focused on **frequentist statistics**
- The two key differences between the Bayesian approach and the Frequentist approach are
  - in the interpretation of what a "probability" means.
  - in the treatment of unknown parameters as random (Bayesian) or fixed (Frequentist)

Bayesian Inference

Frequentist	**Bayesian**
Unknown parameters θ are treated as having fixed (non-random), but unknown values.	Unknown parameters θ are treated as random variables that follow some prior distribution.
The "probability" of an event is the frequency with which it would occur if the circumstances were repeated infinitely many times.	The "probability" of an event represents the degree of belief that the event occurs.
Parameters are fixed.	The data is fixed.

Bayesian Inference

- In Frequentest inference, we make inferences about the unknown parameter θ using methods such as MLE and MOM
- In Bayesian inference, we assume the parameter θ is random and has some prior distribution: $\theta \sim F_\Theta$.
  - This prior distribution encapsulates our assumptions about θ before seeing the data.
  - We use the observed data to "update our knowledge" about θ.
  - This updated knowledge is captured in the **posterior distribution** (the distribution of Θ given the data, X):

$$\text{Posterior probability} \propto \text{Likelihood} \times \text{Prior probability}$$

Bayesian Inference

Posterior distribution: the distribution of Θ given the data, X:

$$\text{Posterior probability} = \frac{\text{Likelihood} \times \text{Prior probability}}{\text{some normalizing constant}}$$

We know how to calculate:

- the likelihood, $f_{X|\Theta=\theta}(x|\theta)$, is the density of the data X given $\Theta = \theta$
- the prior distribution has density $f_\Theta(\theta)$

And so we can calculate the posterior distribution using Bayes Theorem:

$$f_{\Theta|X=x}(\theta|x) = \frac{f_{X|\Theta}(x|\theta)f_\Theta(\theta)}{f_X(x)} = \frac{f_{X|\Theta=\theta}(x|\theta)f_\Theta(\theta)}{\int f_{X|\Theta=\theta'}(x|\theta')f_\Theta(\theta')d\theta'}$$

Bayesian Inference

Why are we interested in the posterior distribution?

- The posterior distribution is the distribution of our unknown parameter, treated as a random variable, after taking into account the evidence (data) obtained from an experiment
- **We can use the posterior distribution to make inferences about θ**
- If we are interested in the expected value of our parameter, θ, then we can calculate the expected value of the posterior distribution (the posterior mean).

For example, we can estimate the mean value of θ using the posterior mean:

$$E(\theta) = \int \theta f_{\Theta|X=x}(\theta|x)d\theta$$

Exercise

Bayesian Inference: exercise (Rice chapter 8 exercise 66)

Suppose that the unknown probability that a basketball player makes a shot successfully is θ. Suppose that your prior on θ is uniform on $[0, 1]$ and that she then makes two shots in a row. Assume that the outcome of the two shots are independent.

1. What is the posterior density of θ?
2. What would you estimate the probability that she makes a third shot to be?