Week 2-3
Sampling distribution of estimators

Bodhisattva Sen

November 8, 2017

1 The sampling distribution of a statistic

A **statistic** is a function of the data, and hence is itself a random variable with a distribution.

This distribution is called its **sampling distribution**. It tells us what values the statistic is likely to assume and how likely is it to take these values.

Formally, suppose that $X_1, \ldots, X_n$ are i.i.d from distribution P_θ, where $\theta \in \Omega \subset \mathbb{R}^k$.

Let T be a statistic, i.e., suppose that $T = \varphi(X_1, \ldots, X_n)$. Assume that $T \sim F_\theta$, where F_θ is the c.d.f of T (possibly dependent on θ).

The distribution of T (with θ fixed) is called the **sampling distribution** of T. Thus, the sampling distribution has c.d.f F_θ.

Example: Suppose that $X_1, \ldots, X_n$ are i.i.d $N(\mu, \sigma^2)$. Then we know that

$$\bar{X}_n \sim N\left(\mu, \frac{\sigma^2}{n}\right).$$

2 The gamma and the χ^2 distributions

2.1 The gamma distribution

The gamma function is a real-valued non-negative function defined on $(0, \infty)$ in the following manner

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x}\, dx \;\; , \;\; \alpha > 0\,.$$

The Gamma function enjoys some nice properties. Two of these are listed below:

$$(a)\ \Gamma(\alpha+1) = \alpha\,\Gamma(\alpha)\ \ ,\ \ (b)\ \Gamma(n) = (n-1)!\ \ (n\ \text{ integer})\,.$$

Property (b) is an easy consequence of Property (a). Start off with $\Gamma(n)$ and use Property (a) recursively along with the fact that $\Gamma(1) = 1$ (why?). Another important fact is that $\Gamma(1/2) = \sqrt{\pi}$ (Prove this at home!).

The gamma distribution with parameters $\alpha > 0, \lambda > 0$ (denoted by Gamma(α, λ)) is defined through the following density function:

$$f(x|\alpha,\lambda) = \frac{\lambda^\alpha}{\Gamma(\alpha)}\, e^{-\lambda\, x}\, x^{\alpha-1} I_{(0,\infty)}(x).$$

The first parameter α is called the *shape* parameter and the second parameter λ is called the *scale* parameter.

For fixed λ the shape parameter regulates the shape of the gamma density.

Here is a simple exercise that justifies the term "scale parameter" for λ.

Exercise: Let X be a random variable following Gamma(α, λ). Then show that $Y = \lambda\, X$ (thus X is Y scaled by λ) follows the Gamma$(\alpha, 1)$ distribution. What is the distribution of $c\, X$ for some arbitrary positive constant c? You can use the change of variable theorem in one-dimension to work this out.

Reproductive Property of the gamma distribution:

Let $X_1, X_2, \ldots, X_n$ be independent random variables with $X_i \sim$ Gamma(α_i, λ). Then $S_n := X_1 + X_2 + \ldots + X_n$ is distributed as Gamma$(\sum_{i=1}^n \alpha_i, \lambda)$.

If X follows the Gamma(α, λ) distribution, the mean and variance of X can be explicitly expressed in terms of the parameters:

$$\mathbb{E}(X) = \frac{\alpha}{\lambda}\ \text{ and }\ \text{Var}(X) = \frac{\alpha}{\lambda^2}\,.$$

We outline the computation of a general moment $\mathbb{E}(X^k)$, where k is a positive integer.

We have,

$$\begin{aligned}
\mathbb{E}(X^k) &= \int_0^\infty x^k \, \frac{\lambda^\alpha}{\Gamma(\alpha)} \, e^{-\lambda x} \, x^{\alpha-1} \, dx \\
&= \frac{\lambda^\alpha}{\Gamma(\alpha)} \int_0^\infty e^{-\lambda x} \, x^{k+\alpha-1} \, dx \\
&= \frac{\lambda^\alpha}{\Gamma(\alpha)} \frac{\Gamma(\alpha+k)}{\lambda^{\alpha+k}} \\
&= \frac{(\alpha+k-1)\cdots(\alpha)\,\Gamma(\alpha)}{\lambda^k \, \Gamma(\alpha)} \\
&= \frac{\Pi_{i=1}^k \, (\alpha+i-1)}{\lambda^k} \, .
\end{aligned}$$

The formulae for the mean and the variance should follow directly from the above computation. Note that in the above derivation, we have used the fact that

$$\int_0^\infty e^{-\lambda x} \, x^{k+\alpha-1} \, dx = \frac{\Gamma(\alpha+k)}{\lambda^{\alpha+k}} \, .$$

This is an immediate consequence of the fact that the gamma density with parameters $(\alpha+k, \lambda)$ integrates to 1.

Exercise: Here is an exercise that should follow from the discussion above. Let $S_n \sim$ Gamma(n, λ), where $\lambda > 0$. Show that for large n, the distribution of S_n is well approximated by a normal distribution (with parameters that you need to identify).

2.2 The Chi-squared distribution

We now introduce an important family of distributions, called the chi-squared family. To do so, we first define the **chi-squared distribution** with 1 degree of freedom (for brevity, we call it "chi-squared one" and write it as χ_1^2).

The χ_1^2 distribution: Let $Z \sim N(0,1)$. Then the distribution of $W := Z^2$ is called the χ_1^2 distribution, and W itself is called a χ_1^2 random variable.

Exercise: Show that W follows a Gamma$(1/2, 1/2)$ distribution. (You can do this by working out the density function of W from that of Z).

For any integer $d > 0$ we can now define the χ_d^2 distribution (chi-squared d distribution, or equivalently, the chi-squared distribution with d degrees of freedom).

The χ_d^2 distribution: Let $Z_1, Z_2, \ldots, Z_d$ be i.i.d $N(0,1)$ random variables. Then the distribution of

$$W_d := Z_1^2 + Z_2^2 + \ldots + Z_d^2$$

is called the χ^2_d distribution and W_d itself is called a χ^2_d random variable.

Exercise: Using the reproductive property of the Gamma distribution, show that $W_d \sim \text{Gamma}(d/2, 1/2)$.

Thus, it follows that the sum of k i.i.d χ^2_1 random variables is a χ^2_k random variable.

Exercise: Let Z_1, Z_2, Z_3 be i.i.d $N(0,1)$ random variables. Consider the vector (Z_1, Z_2, Z_3) as a random point in 3-dimensional space. Let R be the length of the radius vector connecting this point to the origin. Find the density functions of (a) R and (b) R^2.

Theorem 2.1. *If $X \sim \chi^2_m$ then $\mathbb{E}(X) = m$ and $\text{Var}(X) = 2m$.*

Theorem 2.2. *Suppose that $X_1, \ldots, X_k$ are independent and $X_i \sim \chi^2_{m_i}$ then the sum*

$$X_1 + \cdots + X_k \sim \chi^2_{\sum_{i=1}^k m_i}.$$

3 Sampling from a normal population

Let $X_1, X_2, \ldots, X_n$ be i.i.d $N(\mu, \sigma^2)$, where $\mu \in \mathbb{R}$, $\sigma > 0$ are unknown.

You could think of the X_i's for example as a set of randomly sampled SAT scores from the entire population of SAT scores. Then μ is the average SAT score of the entire population and σ^2 is the variance of SAT scores in the entire population. We are interested in estimating μ and σ^2 based on the data. Note that SAT scores are actually discrete in nature — $N(\mu, \sigma^2)$ provides a good approximation to the actual population distribution. In other words, $N(\mu, \sigma^2)$ is the **model** that we use for the SAT scores.

In statistics as in any other science, models are meant to provide insightful approximations to the true underlying nature of reality.

Natural estimates of the mean and the variance are given by:

$$\hat{\mu} = \overline{X}_n = \frac{X_1 + X_2 + \ldots + X_n}{n} \text{ and } \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \overline{X}_n)^2 .$$

These are the *sample mean* and *sample variance* (biased version). In what follows, we will use a slightly different estimator of σ^2 than the one proposed above. We will use

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \overline{X}_n)^2 .$$

One reason for using s^2 is that it has a natural interpretation as the multiple of a χ^2 random variable; further s^2 is an *unbiased estimator* of σ^2 whereas $\hat{\sigma}^2$ is not, i.e.,

$$\mathbb{E}(s^2) = \sigma^2 \text{ but } \mathbb{E}(\hat{\sigma}^2) \neq \sigma^2 .$$

For the sake of notational simplicity we will let S^2 denote the *residual sum of squares about the mean*, i.e., $S^2 := \sum_{i=1}^n (X_i - \overline{X}_n)^2$.

Here is an interesting (and fairly profound) proposition.

Proposition 3.1. *Let $X_1, X_2, \ldots, X_n$ be an i.i.d sample from some distribution F with mean μ and variance σ^2. Then F is the $N(\mu, \sigma^2)$ distribution if and only if for all n, $\overline{X}_n$ and s^2 are independent random variables.*

The "if" part is the profound part. It says that the independence of the natural estimates of the mean and the variance for any sample size forces the underlying distribution to be normal.

We will sketch a proof of the only if part, i.e., we will assume that F is $N(\mu, \sigma^2)$ and show that $\overline{X}_n$ and s^2 are independent.

Proof. To this end, define new random variables $Y_1, Y_2, \ldots, Y_n$ where for each i,

$$Y_i = (X_i - \mu)/\sigma.$$

These are the *standardized versions* of the X_i's and are i.i.d. $N(0, 1)$ random variables. Now, note that:

$$\overline{X} = \overline{Y}\,\sigma + \mu \text{ and } s^2 = \frac{\sigma^2 \sum_{i=1}^n (Y_i - \overline{Y})^2}{n-1} .$$

From the above display, we see that it suffices to show the independence of $\overline{Y}$ and $\sum_{i=1}^n (Y_i - \overline{Y})^2$.

The way this proceeds is outlined below: Let $\boldsymbol{Y}$ denote the $n \times 1$ column vector $(Y_1, Y_2, \ldots, Y_n)^\top$ and let P be an $n \times n$ orthogonal matrix with the first row of P (which has length n) being $(1/\sqrt{n}, 1/\sqrt{n}, \ldots, 1/\sqrt{n})$.

Recall that an orthogonal matrix satisfies

$$P^\top P = P P^\top = I$$

where I is the identity matrix.

Using standard linear algebra techniques it can be shown that such a P can always be constructed. Now define a new random vector

$$W = P\boldsymbol{Y}.$$

Then it can be established that the random vector $\boldsymbol{W} = (W_1, W_2, \ldots, W_n)^\top$ has the same distribution as $(Y_1, Y_2, \ldots, Y_n)^\top$; in other words, $W_1, W_2, \ldots, W_n$ are i.i.d $N(0,1)$ random variables.

Theorem 3.2. *Suppose that $Z_1, \ldots, Z_n$ are i.i.d $N(0,1)$. Suppose that A is an orthogonal matrix and*

$$\boldsymbol{V} = A\boldsymbol{Z}.$$

Then the random variables $V_1, \ldots, V_n$ are i.i.d $N(0,1)$. Also, $\sum_{i=1}^n V_i^2 = \sum_{i=1}^n Z_i^2$.

Note that

$$\boldsymbol{W}^\top \boldsymbol{W} = (P\boldsymbol{Y})^\top P\boldsymbol{Y} = \boldsymbol{Y}^\top P^\top P\boldsymbol{Y} = \boldsymbol{Y}^\top \boldsymbol{Y}$$

by the orthogonality of P; in other words, $\sum_{i=1}^n W_i^2 = \sum_{i=1}^n Y_i^2$. Also,

$$W_1 = Y_1/\sqrt{n} + Y_2/\sqrt{n} + \ldots + Y_n/\sqrt{n} = \sqrt{n}\,\overline{Y}\,.$$

Note that W_1 is independent of $W_2^2 + W_3^2 + \ldots + W_n^2$. But

$$\sum_{i=2}^n W_i^2 = \sum_{i=1}^n W_i^2 - W_1^2 = \sum_{i=1}^n Y_i^2 - n\,\overline{Y}^2 = \sum_{i=1}^n (Y_i - \overline{Y})^2\,.$$

It therefore follows that $\sqrt{n}\,\overline{Y}$ and $\sum_{i=1}^n (Y_i - \overline{Y})^2$ are independent, which implies that $\overline{Y}$ and $\sum_{i=1}^n (Y_i - \overline{Y})^2$ are independent. □

Note that $\overline{Y} \sim N(0, 1/n)$. Deduce that $\overline{X}$ follows $N(\mu, \sigma^2/n)$. Since $\sum_{i=1}^n (Y_i - \overline{Y})^2 = W_2^2 + W_3^2 + \ldots + W_n^2$, it follows that

$$\frac{S^2}{\sigma^2} = \sum_{i=1}^n (Y_i - \overline{Y})^2 \sim \chi^2_{n-1}\,.$$

Thus,

$$s^2 = \frac{S^2}{n-1} =_d \frac{\sigma^2}{n-1}\chi^2_{n-1}. \tag{1}$$

In the above display the symbol $=_d$ means "is equal in distribution to".

In the case $n = 2$, it is easy to check the details of the transformation leading from $\boldsymbol{Y}$ to $\boldsymbol{W}$. Set $\boldsymbol{W} = P\boldsymbol{Y}$ with

$$P = \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{pmatrix}.$$

Thus $W_1 = (Y_1 + Y_2)/\sqrt{2}$ and $W_2 = (Y_1 - Y_2)/\sqrt{2}$.

Exercise: Use the change of variable theorem to deduce that W_1 and W_2 are i.i.d $N(0,1)$.

Proof of Theorem 3.2: The joint p.d.f of $\boldsymbol{Z} = (Z_1, \dots Z_n)$ is

$$f_n(\boldsymbol{z}) = \frac{1}{(2\pi)^{n/2}} \exp\left(-\frac{1}{2}\sum_{i=1}^{n} z_i^2\right), \qquad \text{for } \boldsymbol{z} \in \mathbb{R}^n.$$

Note that as $\boldsymbol{Z} \mapsto A\boldsymbol{Z}$ is a linear transformation. The joint p.d.f of $\boldsymbol{V} = A\boldsymbol{Z}$ is

$$g_n(\boldsymbol{v}) = \frac{1}{|\det A|} f_n(A^{-1}\boldsymbol{v}), \qquad \text{for } \boldsymbol{v} \in \mathbb{R}^n.$$

Let $\boldsymbol{z} = A^{-1}\boldsymbol{v}$. Since A is orthogonal, $|\det A| = 1$ and $\boldsymbol{v}^\top \boldsymbol{v} = \sum_{i=1}^n v_i^2 = \boldsymbol{z}^\top \boldsymbol{z} = \sum_{i=1}^n z_i^2$. So,

$$g_n(\boldsymbol{v}) = \frac{1}{(2\pi)^{n/2}} \exp\left(-\frac{1}{2}\sum_{i=1}^{n} v_i^2\right), \qquad \text{for } \boldsymbol{v} \in \mathbb{R}^n.$$

Thus, $\boldsymbol{V}$ has the same joint p.d.f as $\boldsymbol{Z}$.

---

4 The t-distribution

Definition: Let $Z \sim N(0,1)$ and let $V \sim \chi^2_n$, independent of each other. Then,

$$T = \frac{Z}{\sqrt{V/n}}$$

is said to follow the *t-distribution* on n degrees of freedom. We write $T \sim t_n$.

The density of the t-distribution is derived in the text book (see Chapter 8.4). With a little bit of patience, you can also work it out, using the change of variable theorem appropriately (I won't go into the computational details here).

Exercise: Let X be a random variable that is distributed symmetrically about 0, i.e., X and $-X$ have the same distribution function (and hence the same density function). If f denotes the density, show that it is an even function, i.e. $f(x) = f(-x)$ for all x.

Conversely, if the random variable X has a density function f that is even, then it is symmetrically distributed about 0, i.e $X =_d -X$.

Here are some important facts about the t-distribution. Let $T \sim t_n$.

(a) T and $-T$ have the same distribution. Thus, the distribution of T is symmetric about 0 and it has an even density function.

From definition,

$$-T = \frac{-Z}{\sqrt{V/n}} = \frac{\tilde{Z}}{\sqrt{V/n}},$$

where $\tilde{Z} \equiv -Z$ follows $N(0,1)$, and is independent of V where V follows χ^2_n. Thus, by definition, $-T$ also follows the t-distribution on n degrees of freedom.

(b) As $n \to \infty$, the t_n distribution converges to the $N(0,1)$ distribution; hence the quantiles of the t-distribution are well approximated by the quantiles of the normal distribution.

This follows from the law of large numbers. Consider the term V/n in the denominator of T for large n. As V follows χ^2_n it has the same distribution as $K_1 + K_2 + \ldots + K_n$ where K_i's are i.i.d χ^2_1 random variables. But by the WLLN we know that

$$\frac{K_1 + K_2 + \ldots + K_n}{n} \xrightarrow{\mathbb{P}} \mathbb{E}(K_1) = 1 \ \text{(check!)}.$$

Thus V/n converges in probability to 1; hence the denominator in T converges in probability to 1 and T consequently, converges in distribution to Z, where Z is $N(0,1)$.

Theorem 4.1. *Suppose that $X_1, \ldots, X_n$ form a random sample from the normal distribution with mean μ and variance σ^2. Let $\overline{X}_n$ denote the sample mean, and define $s^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X}_n)^2$. Then*

$$\frac{\sqrt{n}(\overline{X}_n - \mu)}{s} \sim t_{n-1}.$$

5 Confidence intervals

Confidence intervals (CIs) provide a method of quantifying uncertainty to an estimator $\hat{\theta}$ when we wish to estimate an unknown parameter θ.

We want to find an interval (A, B) that we think has high probability of containing θ.

Definition: Suppose that $\boldsymbol{X}_n = (X_1, \ldots, X_n)$ is a random sample from a distribution P_θ, $\theta \in \Omega \subset \mathbb{R}^k$ (that depends on a parameter θ).

Suppose that we want to estimate $g(\theta)$, a real-valued function of θ.

Let $A \leq B$ be two statistics that have the property that for all values of θ,

$$\mathbb{P}_\theta(A \leq g(\theta) \leq B) \geq 1 - \alpha,$$

where $\alpha \in (0, 1)$.

Then the random interval (A, B) is called a *confidence interval* for $g(\theta)$ with level (coefficient) $(1 - \alpha)$.

If the inequality "$\geq 1 - \alpha$" is an equality for all θ, the the CI is called *exact*.

Example 1: Find a level $(1 - \alpha)$ CI for μ from data $X_1, X_2, \ldots, X_n$ which are i.i.d. $N(\mu, \sigma^2)$ where σ is **known**. Here $\theta = \mu$ and $g(\theta) = \mu$.

Step 1: We want to construct $\Psi(X_1, X_2, \ldots, X_n, \mu)$ such that the distribution of this object is known to us.

How do we proceed here?

The usual way is to find some decent estimator of μ and combine it along with μ in some way to get a "pivot", i.e., a random variable whose distribution does not depend on θ.

The most intuitive estimator of μ here is the sample mean $\overline{X}_n$. We know that

$$\overline{X}_n \sim N(\mu, \sigma^2/n).$$

The standardized version of the sample mean follows $N(0, 1)$ and can therefore act as a pivot. In other words, construct,

$$\Psi(\boldsymbol{X}_n, \mu) = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} = \frac{\sqrt{n}\,(\overline{X} - \mu)}{\sigma} \sim N(0, 1)$$

for every value of θ.

With z_β denoting the upper β-th quantile of $N(0,1)$ (i.e., $\mathbb{P}(Z > z_\beta) = \beta$ where Z follows $N(0,1)$) we can write:

$$\mathbb{P}_\mu\left(-z_{\alpha/2} \leq \frac{\sqrt{n}(\overline{X}-\mu)}{\sigma} \leq z_{\alpha/2}\right) = 1-\alpha.$$

From the above display we can find limits for μ such that the above inequalities are simultaneously satisfied. On doing the algebra, we get:

$$\mathbb{P}_\mu\left(\overline{X} - \frac{\sigma}{\sqrt{n}} z_{\alpha/2} \leq \mu \leq \overline{X} + \frac{\sigma}{\sqrt{n}} z_{\alpha/2}\right) = 1-\alpha.$$

Thus our level $(1-\alpha)$ CI for μ is given by

$$\left[\overline{X} - \frac{\sigma}{\sqrt{n}} z_{\alpha/2}, \overline{X} + \frac{\sigma}{\sqrt{n}} z_{\alpha/2}\right].$$

Often a standard method of constructing CIs is the following *method of pivots* which we describe below.

(1) Construct a function Ψ using the data $\boldsymbol{X}_n$ and $g(\theta)$, say $\Psi(\boldsymbol{X}_n, g(\theta))$, such that the distribution of this random variable under parameter value θ *does not depend on* θ and is known.

Such a Ψ is called a *pivot*.

(2) Let G denote the distribution function of the pivot. The idea now is to get a range of plausible values of the pivot. The level of confidence $1-\alpha$ is to be used to get the appropriate range.

This can be done in a variety of ways but the following is standard. Denote by $q(G;\beta)$ the β'th quantile of G. Thus,

$$\mathbb{P}_\theta[\Psi(\boldsymbol{X}_n, g(\theta)) \leq q(G;\beta)] = \beta.$$

(3) Choose $0 \leq \beta_1, \beta_2 \leq \alpha$ such that $\beta_1 + \beta_2 = \alpha$. Then,

$$\mathbb{P}_\theta[q(G;\beta_1) \leq \Psi(\boldsymbol{X}_n, g(\theta)) \leq q(G;1-\beta_2)] = 1-\beta_2-\beta_1 = 1-\alpha\,.$$

(4) Vary θ across its domain and choose your level $1-\alpha$ confidence interval (set) as the set of all $g(\theta)$ such that the two inequalities in the above display are simultaneously satisfied.

Example 2: The data are the same as in Example 1 but now σ^2 is no longer known. Thus, the parameter of unknowns $\theta = (\mu, \sigma^2)$ and we are interested in finding a CI for $g(\theta) = \mu$.

Clearly, setting

$$\Psi(\boldsymbol{X}_n, \mu) = \frac{\sqrt{n}(\overline{X} - \mu)}{\sigma}$$

will not work smoothly here. This certainly has a known ($N(0,1)$) distribution but involves the *nuisance parameter* σ making it difficult get a CI for μ directly.

However, one can replace σ by s, where s^2 is the natural estimate of σ^2 introduced before. So, set:

$$\Psi(\boldsymbol{X}_n, \mu) = \frac{\sqrt{n}(\overline{X} - \mu)}{s}\,.$$

This only depends on the data and $g(\theta) = \mu$. We claim that this is indeed a pivot.

To see this write

$$\frac{\sqrt{n}(\overline{X} - \mu)}{s} = \frac{\frac{\sqrt{n}(\overline{X} - \mu)}{\sigma}}{\sqrt{s^2/\sigma^2}}\,.$$

The numerator on the extreme right of the above display follows $N(0,1)$ and the denominator is independent of the numerator and is the square root of a χ^2_{n-1} random variable over its degrees of freedom (from display (1)).

It follows from definition that $\Psi(\boldsymbol{X}_n, \mu) \sim t_{n-1}$ distribution.

Thus, G here is the t_{n-1} distribution and we can choose the quantiles to be $q(t_{n-1}; \alpha/2)$ and $q(t_{n-1}; 1 - \alpha/2)$. By symmetry of the t_{n-1} distribution about 0, we have, $q(t_{n-1}; \alpha/2) = -q(t_{n-1}; 1 - \alpha/2)$. It follows that,

$$\mathbb{P}_{\mu,\sigma^2}\left[-q(t_{n-1}; 1 - \alpha/2) \leq \frac{\sqrt{n}(\overline{X} - \mu)}{s} \leq q(t_{n-1}; 1 - \alpha/2)\right] = 1 - \alpha\,.$$

As with Example 1, direct algebraic manipulations show that this is the same as the statement:

$$\mathbb{P}_{\mu,\sigma^2}\left[\overline{X} - \frac{s}{\sqrt{n}}\, q(t_{n-1}; 1 - \alpha/2) \leq \mu \leq \overline{X} + \frac{s}{\sqrt{n}}\, q(t_{n-1}; 1 - \alpha/2)\right] = 1 - \alpha\,.$$

This gives a level $1 - \alpha$ confidence set for μ.

Food for thought: In each of the above examples there are innumerable ways of decomposing α as $\beta_1 + \beta_2$. It turns out that when α is split equally the level $1 - \alpha$ CIs obtained in Examples 1 and 2 are the shortest.

What are desirable properties of confidence sets? On one hand, we require high levels of confidence; in other words, we would like α to be as small as possible.

On the other hand we would like our CIs to be shortest possible.

Unfortunately, we cannot simultaneously make the confidence levels of our CIs go up and the lengths of our CIs go down.

In Example 1, the length of the level $(1-\alpha)$ CI is

$$2\sigma\frac{z_{\alpha/2}}{\sqrt{n}}.$$

As we reduce α (for higher confidence), $z_{\alpha/2}$ increases, making the CI wider.

However, we can reduce the length of our CI for a fixed α by increasing the sample size.

If my sample size is 4 times yours, I will end up with a CI which has the same level as yours but has half the length of your CI.

Can we hope to get absolute confidence, i.e. $\alpha = 0$? That is too much of an ask. When $\alpha = 0$, $z_{\alpha/2} = \infty$ and the CIs for μ are infinitely large. The same can be verified for Example 2.

Asymptotic pivots using the central limit theorem: The CLT allows us to construct an *approximate pivot* for large sample sizes for estimating the population mean μ for any underlying distribution F.

Let $X_1, X_2, \ldots, X_n$ be i.i.d observations from some common distribution F and let

$$\mathbb{E}(X_1) = \mu \qquad \text{and} \qquad \text{Var}(X_1) = \sigma^2.$$

We are interested in constructing an approximate level $(1-\alpha)$ CI for μ.

By the CLT we have $\overline{X} \sim_{appx} N(\mu, \sigma^2/n)$ for large n; in other words,

$$\frac{\sqrt{n}(\overline{X}-\mu)}{\sigma} \sim_{appx} N(0,1)\,.$$

If σ is known the above quantity is an approximate pivot and following Example 1, we can therefore write,

$$\mathbb{P}_\mu\left(-z_{\alpha/2} \le \frac{\sqrt{n}\,(\overline{X}-\mu)}{\sigma} \le z_{\alpha/2}\right) \approx 1-\alpha\,.$$

As before, this translates to

$$\mathbb{P}_\mu\left(\overline{X} - \frac{\sigma}{\sqrt{n}}\,z_{\alpha/2} \le \mu \le \overline{X} + \frac{\sigma}{\sqrt{n}}\,z_{\alpha/2}\right) \approx 1-\alpha.$$

This gives an approximate level $(1-\alpha)$ CI for μ when σ is known.

The approximation will improve as the sample size n increases.

Note that the true coverage of the above CI will be different from $1-\alpha$ and can depend heavily on the nature of F and the sample size n.

Realistically however σ is unknown and is replaced by s. Since we are dealing with large sample sizes, s is with very high probability close to σ and the interval

$$\left(\overline{X} - \frac{s}{\sqrt{n}}\, z_{\alpha/2}\ ,\ \overline{X} + \frac{s}{\sqrt{n}}\, z_{\alpha/2}\right),$$

still remains an approximate level $(1-\alpha)$ CI.

Exercise: Suppose $X_1, X_2, \ldots, X_n$ are i.i.d Ber(θ). The sample size n is large.

Thus

$$\mathbb{E}(X_1) = \theta \qquad \text{and} \qquad \mathrm{Var}(X_1) = \theta(1-\theta).$$

We want to find a level $(1-\alpha)$ CI (approximate) for θ.

Note that both mean and variance are unknown.

Show that if $\hat{\theta}$ is natural estimate of θ obtained by computing the sample proportion of 1's, then

$$\left[\hat{\theta} - \sqrt{\frac{\hat{\theta}\,(1-\hat{\theta})}{n-1}}\, z_{\alpha/2}\ ,\ \hat{\theta} + \sqrt{\frac{\hat{\theta}\,(1-\hat{\theta})}{n-1}}\, z_{\alpha/2}\right]$$

is an approximate level $(1-\alpha)$ CI for θ.

See http://www.rossmanchance.com/applets/ConfSim.html and
http://www.ruf.rice.edu/~lane/stat_sim/conf_interval/ for illustrations of confidence intervals.

Interpretation of confidence intervals: Let (A, B) be a coefficient γ confidence interval for a parameter θ. Let (a, b) be the observed value of the interval.

It is NOT correct to say that "θ lies in the interval (a, b) with *probability* γ".

It is true that "θ will lie in the random intervals having endpoints $A(X_1, \ldots, X_n)$ and $B(X_1, \ldots, X_n)$ with probability γ".

After observing the specific values $A(X_1, \ldots, X_n) = a$ and $B(X_1, \ldots, X_n) = b$, it is not possible to assign a probability to the event that θ lies in the specific interval (a, b) without regarding θ as a random variable.

We usually say that there is *confidence* γ that θ lies in the interval (a, b).

6 The (Cramer-Rao) Information Inequality

We saw in the last lecture that for a variety of different models one could differentiate the log-likelihood function with respect to the parameter θ and set this equal to 0 to obtain the MLE of θ.

In these examples, the log-likelihood as a function of θ is strictly concave (looks like an inverted bowl) and hence solving for the stationary point gives us the unique maximizer of the log-likelihood.

We start this section by introducing some notation. Let X be a random variable with p.d.f $f(\cdot, \theta)$, where $\theta \in \Omega$, and

$$\ell(x, \theta) = \log f(x, \theta) \qquad \text{and} \qquad \dot{\ell}(x, \theta) = \frac{\partial}{\partial \theta} \ell(x, \theta).$$

As before, $\boldsymbol{X}_n$ denotes the vector $(X_1, X_2, \ldots, X_n)$ and $\boldsymbol{x}$ denotes a particular value $(x_1, x_2, \ldots, x_n)$ assumed by the random vector $\boldsymbol{X}_n$.

We denote by $f_n(\boldsymbol{x}, \theta)$ the value of the density of $\boldsymbol{X}_n$ at the point $\boldsymbol{x}$. Then,

$$f_n(\boldsymbol{x}, \theta) = \prod_{i=1}^{n} f(x_i, \theta).$$

Thus,

$$L_n(\theta, \boldsymbol{X}_n) = \prod_{i=1}^{n} f(X_i, \theta) = f_n(\boldsymbol{X}_n, \theta)$$

and

$$\ell_n(\boldsymbol{X}_n, \theta) = \log L_n(\theta, \boldsymbol{X}_n) = \sum_{i=1}^{n} \ell(X_i, \theta).$$

Differentiating with respect to θ yields

$$\dot{\ell}_n(\boldsymbol{X}_n, \theta) = \frac{\partial}{\partial \theta} \log f_n(\boldsymbol{X}_n, \theta) = \sum_{i=1}^{n} \dot{\ell}(X_i, \theta).$$

We call $\dot{\ell}(x, \theta)$ the **score function** and

$$\dot{\ell}_n(\boldsymbol{X}_n, \theta) = 0$$

the **score equation**. If differentiation is permissible for the purpose of obtaining the MLE, then $\hat{\theta}_n$, the MLE, solves the equation

$$\dot{\ell}_n(\boldsymbol{X}_n, \theta) \equiv \sum_{i=1}^{n} \dot{\ell}(X_i, \theta) = 0.$$

In this section, our first goal is to find a (nontrivial) **lower bound** on the **variance of unbiased estimators** of $g(\theta)$ where $g : \Omega \to \mathbb{R}$ is some differentiable function.

If we can indeed find such a bound (albeit under some regularity conditions) and there is an unbiased estimator of $g(\theta)$ that attains this lower bound, we can conclude that it is the MVUE of $g(\theta)$.

We now impose the following restrictions (regularity conditions) on the model.

(A.1) The set $A_\theta = \{x : f(x, \theta) > 0\}$ actually does NOT depend on θ and is subsequently denoted by A.

(A.2) If $W(\boldsymbol{X}_n)$ is a statistic such that $\mathbb{E}_\theta(|W(\boldsymbol{X}_n)|) < \infty$ for all θ, then,

$$\frac{\partial}{\partial\theta}\mathbb{E}_\theta[W(\boldsymbol{X}_n)] = \frac{\partial}{\partial\theta}\int_{A^n} W(\boldsymbol{x})\, f_n(\boldsymbol{x}, \theta)\, d\boldsymbol{x} = \int_{A^n} W(\boldsymbol{x})\, \frac{\partial}{\partial\theta} f_n(\boldsymbol{x}, \theta) d\boldsymbol{x}.$$

(A.3) The quantity $\frac{\partial}{\partial\theta}\log f(x, \theta)$ exists for all $x \in A$ and all $\theta \in \Omega$ as a well defined finite quantity.

The first condition says that the set of possible values of the data vector on which the distribution of $\boldsymbol{X}_n$ is supported does not vary with θ; this therefore rules out families of distribution like the uniform.

The second assumption is a "smoothness assumption" on the family of densities and is generally happily satisfied for most parametric models we encounter in statistics.

There are various types of simple sufficient conditions that one can impose on $f(x, \theta)$ to make the interchange of integration and differentiation possible — we shall however not bother about these for the moment.

We define the **information** about the parameter θ in the model, namely $I(\theta)$, by

$$I(\theta) := \mathbb{E}_\theta[\dot{\ell}^2(X, \theta)],$$

provided it exists as a finite quantity for every $\theta \in \Omega$.

We then have the following theorem.

Theorem 6.1 (Cramer-Rao inequality)**.** *All notation being as above, if $T(\boldsymbol{X}_n)$ is an unbiased estimator of $g(\theta)$, then*

$$\mathrm{Var}_\theta(T(\boldsymbol{X}_n)) \geq \frac{[g'(\theta)]^2}{nI(\theta)},$$

provided assumptions A.1, A.2 and A.3 hold, and $I(\theta)$ exists and is finite for all θ.

The above inequality is the celebrated **Cramer-Rao inequality** (or the information inequality) and is one of the most well-known inequalities in statistics and has important ramifications in even more advanced forms of inference.

Notice that if we take $g(\theta) = \theta$ then $n^{-1}I(\theta)^{-1}$ gives us a lower bound on the variance of unbiased estimators of θ in the model.

If $I(\theta)$ is small, the lower bound is large, so unbiased estimators are doing a poor job in general — in other words, the data is not that informative about θ (within the context of unbiased estimation).

On the other hand, if $I(\theta)$ is big, the lower bound is small, and so if we have a best unbiased estimator of θ that actually attains this lower bound, we are doing a good job. That is why $I(\theta)$ is referred to as the information about θ.

Proof of Theorem 6.1: Let ρ_θ denote the correlation between $T(\boldsymbol{X}_n)$ and $\dot{\ell}_n(\boldsymbol{X}_n, \theta)$. Then $\rho_\theta^2 \leq 1$ which implies that

$$\mathrm{Cov}_\theta^2 \left(T(\boldsymbol{X}_n), \dot{\ell}_n(\boldsymbol{X}_n, \theta) \right) \leq \mathrm{Var}_\theta(T(\boldsymbol{X}_n)) \cdot \mathrm{Var}_\theta(\dot{\ell}_n(\boldsymbol{X}_n, \theta)). \tag{2}$$

As,

$$1 = \int f_n(\boldsymbol{x}, \theta)\, d\boldsymbol{x}, \quad \text{for all } \theta \in \Omega,$$

on differentiating both sides of the above identity with respect to θ and using $A.2$ with $W(\boldsymbol{x}) \equiv 1$ we obtain,

$$\begin{aligned} 0 &= \int \frac{\partial}{\partial \theta} f_n(\boldsymbol{x}, \theta) d\boldsymbol{x} = \int \left(\frac{\partial}{\partial \theta} f_n(\boldsymbol{x}, \theta) \right) \frac{1}{f_n(\boldsymbol{x}, \theta)} f_n(\boldsymbol{x}, \theta) d\boldsymbol{x} \\ &= \int \left(\frac{\partial}{\partial \theta} \log f_n(\boldsymbol{x}, \theta) \right) f_n(\boldsymbol{x}, \theta) d\boldsymbol{x}. \end{aligned}$$

The last expression in the above display is precisely $\mathbb{E}_\theta[\dot{\ell}_n(\boldsymbol{X}_n, \theta)]$ which therefore is equal to 0. Note that,

$$\mathbb{E}_\theta[\dot{\ell}_n(\boldsymbol{X}_n, \theta)] = \mathbb{E}_\theta \left(\sum_{i=1}^n \dot{\ell}(X_i, \theta) \right) = n\mathbb{E}_\theta \left[\dot{\ell}(X, \theta) \right],$$

since the $\dot{\ell}(X_i, \theta)$'s are i.i.d. Thus, we have $\mathbb{E}_\theta \left(\dot{\ell}(X_1, \theta) \right) = 0$. This implies that

$$I(\theta) = \mathrm{Var}_\theta(\dot{\ell}(X, \theta)).$$

Further, let $I_n(\theta) := \mathbb{E}_\theta[\dot{\ell}_n^2(\boldsymbol{X}_n, \theta)]$. Then

$$\begin{aligned} I_n(\theta) &= \mathrm{Var}_\theta(\dot{\ell}_n(\boldsymbol{X}_n, \theta)) = \mathrm{Var}_\theta \left(\sum_{i=1}^n \dot{\ell}(X_i, \theta) \right) \\ &= \sum_{i=1}^n \mathrm{Var}_\theta(\dot{\ell}(X_i, \theta)) = nI(\theta). \end{aligned}$$

We will refer to $I_n(\theta)$ as the *information* based on n observations. Since $\mathbb{E}_\theta[\dot{\ell}_n(\boldsymbol{X}_n, \theta)] = 0$, it follows that

$$\begin{aligned}
\text{Cov}_\theta\left(T(\boldsymbol{X}_n), \dot{\ell}_n(\boldsymbol{X}_n, \theta)\right) &= \int T(\boldsymbol{x})\dot{\ell}_n(\boldsymbol{x}, \theta) f_n(\boldsymbol{x}, \theta) d\boldsymbol{x} \\
&= \int T(\boldsymbol{x}) \left(\frac{\partial}{\partial\theta} f_n(\boldsymbol{x}, \theta)\right) d\boldsymbol{x} \\
&= \frac{\partial}{\partial\theta} \int T(\boldsymbol{x}) f_n(\boldsymbol{x}, \theta) d\boldsymbol{x} \qquad \text{(by } A.2\text{)} \\
&= \frac{\partial}{\partial\theta} g(\theta) = g'(\theta).
\end{aligned}$$

Using the above in conjunction in (2) we get,

$$[g'(\theta)]^2 \leq \text{Var}_\theta(T(\boldsymbol{X}_n))\ I_n(\theta)$$

which is equivalent to what we set out to prove. □

There is an alternative expression for the information $I(\theta)$ in terms of the second derivative of the log-likelihood with respect to θ. If

$$\ddot{\ell}(x, \theta) := \frac{\partial^2}{\partial\theta^2} \log f(x, \theta)$$

exists for all $x \in A$ and for all $\theta \in \Theta$ then, we have the following identity:

$$I(\theta) = \mathbb{E}_\theta\left(\dot{\ell}(X, \theta)^2\right) = -\mathbb{E}_\theta\left(\ddot{\ell}(X, \theta)\right),$$

provided we can differentiate twice under the integral sign; more concretely, if

$$\int \frac{\partial^2}{\partial\theta^2} f(x, \theta) dx = \frac{\partial^2}{\partial\theta^2} \int f(x, \theta) dx = 0 \quad (\star).$$

To prove the above identity, first note that,

$$\dot{\ell}(x, \theta) = \frac{1}{f(x, \theta)} \left[\frac{\partial}{\partial\theta} f(x, \theta)\right].$$

Now,

$$\begin{aligned}
\ddot{\ell}(x, \theta) &= \frac{\partial}{\partial\theta}\left(\dot{\ell}(x, \theta)\right) = \frac{\partial}{\partial\theta}\left(\frac{1}{f(x, \theta)} \frac{\partial}{\partial\theta} f(x, \theta)\right) \\
&= \frac{\partial^2}{\partial\theta^2} f(x, \theta) \frac{1}{f(x, \theta)} - \frac{1}{f^2(x, \theta)} \left(\frac{\partial}{\partial\theta} f(x, \theta)\right)^2 \\
&= \frac{\partial^2}{\partial\theta^2} f(x, \theta) \frac{1}{f(x, \theta)} - \dot{\ell}(x, \theta)^2.
\end{aligned}$$

Thus,

$$\begin{aligned}\mathbb{E}_\theta[\ddot{\ell}(X,\theta)] &= \int \ddot{\ell}(x,\theta) f(x,\theta) dx \\ &= \int \frac{\partial^2}{\partial\theta^2} f(x,\theta) dx - \mathbb{E}_\theta[\dot{\ell}^2(X,\theta)] \\ &= 0 - \mathbb{E}_\theta[\dot{\ell}^2(X,\theta)],\end{aligned}$$

where the first term on the right side vanishes by virtue of $(\star)$. This establishes the desired equality. It follows that,

$$I_n(\theta) = \mathbb{E}_\theta[-\ddot{\ell}_n(\boldsymbol{X}_n,\theta)],$$

where $\ddot{\ell}_n(\boldsymbol{X}_n,\theta)$ is the second partial derivative of $\ell_n(\boldsymbol{X}_n,\theta)$ with respect to θ. To see this, note that,

$$\ddot{\ell}_n(\boldsymbol{X}_n,\theta) = \frac{\partial^2}{\partial\theta^2}\left(\sum_{i=1}^n \ell(X_i,\theta)\right) = \sum_{i=1}^n \ddot{\ell}(X_i,\theta),$$

so that

$$\mathbb{E}_\theta[\ddot{\ell}_n(\boldsymbol{X}_n,\theta)] = \sum_{i=1}^n \mathbb{E}_\theta[\ddot{\ell}(X_i,\theta)] = n\,\mathbb{E}_\theta[\ddot{\ell}(X,\theta)] = -n\,I(\theta).$$

We now look at some applications of the Cramer-Rao inequality.

Example 1: Let $X_1, X_2, \ldots, X_n$ be i.i.d Pois(θ), $\theta > 0$. Then

$$\mathbb{E}_\theta(X_1) = \theta \qquad \text{and} \qquad \mathrm{Var}_\theta(X_1) = \theta.$$

Let us first write down the likelihood of the data. We have,

$$f_n(\boldsymbol{x},\theta) = \prod_{i=1}^n \frac{e^{-\theta}\,\theta^{x_i}}{x_i!} = e^{-n\,\theta}\,\theta^{\sum_{i=1}^n x_i}\left(\prod_{i=1}^n x_i!\right)^{-1}.$$

Thus,

$$\begin{aligned}\ell_n(\boldsymbol{x},\theta) &= -n\theta + \log\theta\left(\sum_{i=1}^n x_i\right) - \log\prod_{i=1}^n x_i! \\ \dot{\ell}_n(\boldsymbol{x},\theta) &= -n + \frac{1}{\theta}\sum_{i=1}^n x_i.\end{aligned}$$

Thus the information about θ based on n observations is given by,

$$I_n(\theta) = \mathrm{Var}_\theta\left(-n + \frac{1}{\theta}\sum_{i=1}^n X_i\right) = \frac{1}{\theta^2}\mathrm{Var}_\theta\left(\sum_{i=1}^n X_i\right) = \frac{n\theta}{\theta^2} = \frac{n}{\theta}.$$

The assumptions needed for the Cramer-Rao inequality to hold are all satisfied for this model, and it follows that for any unbiased estimator $T(\boldsymbol{X}_n)$ of $g(\theta) = \theta$ we have,

$$\mathrm{Var}_\theta(T(\boldsymbol{X}_n)) \geq \frac{1}{I_n(\theta)} = \frac{\theta}{n}.$$

Since $\overline{X}_n$ is unbiased for θ and has variance θ/n we conclude that $\overline{X}_n$ is the best unbiased estimator (MVUE) of θ.

Example 2: Let $X_1, X_2, \ldots, X_n$ be i.i.d $N(0, V)$. Consider once again, the joint density of the n observations:

$$f_n(\boldsymbol{x}, V) = \frac{1}{(2\pi V)^{n/2}} \exp\left(-\frac{1}{2V}\sum_{i=1}^n x_i^2\right).$$

Now,

$$\begin{aligned}\dot{\ell}_n(\boldsymbol{x}, V) &= \frac{\partial}{\partial V}\left(-\frac{n}{2}\log 2\pi - \frac{n}{2}\log V - \frac{1}{2V}\sum_{i=1}^n x_i^2\right) \\ &= -\frac{n}{2V} + \frac{1}{2V^2}\sum_{i=1}^n x_i^2.\end{aligned}$$

Differentiating yet again we obtain,

$$\ddot{\ell}_n(\boldsymbol{x}, V) = \frac{n}{2V^2} - \frac{1}{V^3}\sum_{i=1}^n x_i^2.$$

Then, the information for V based on n observations is,

$$I_n(V) = -\mathbb{E}_V\left(\frac{n}{2V^2} - \frac{1}{V^3}\sum_{i=1}^n X_i^2\right) = \frac{n}{2\,V^2} + \frac{1}{V^3}\,nV = \frac{n}{2\,V^2}\,.$$

Now consider the problem of estimating $g(V) = V$. For any unbiased estimator $S(\boldsymbol{X}_n)$ of V, the Cramer-Rao inequality tells us that

$$\mathrm{Var}_V\left(S(\boldsymbol{X}_n)\right) \geq I_n(V)^{-1} = \frac{2V^2}{n}.$$

Consider, $\sum_{i=1}^n X_i^2/n$ as an estimator of V. This is clearly unbiased for V and the variance is given by,

$$\mathrm{Var}_V\left(\frac{1}{n}\sum_{i=1}^n X_i^2\right) = \frac{1}{n}\mathrm{Var}_V(X_1^2) = \frac{V^2}{n}\mathrm{Var}_V\left(\frac{X_1^2}{V}\right) = \frac{2V^2}{n},$$

since $X_1^2/V \sim \chi_1^2$ which has variance 2. It follows that $\sum X_i^2/n$ is the best unbiased estimator of V in this model.

7 Large Sample Properties of the MLE

In this section we study some of the large sample properties of the MLE in standard parametric models and how these can be used to construct confidence sets for θ or a function of θ. We will see in this section that in the long run MLEs are the best possible estimators in a variety of different models.

We will stick to models satisfying the restrictions (A1, A2 and A3) imposed in the last section. Hence our results will not apply to the uniform distribution (or ones similar to the uniform).

Let us throw our minds back to the Cramer-Rao inequality. When does an unbiased estimator $T(\boldsymbol{X}_n)$ of $g(\theta)$ attain the bound given by this inequality? This requires:

$$\mathrm{Var}_\theta(T(\boldsymbol{X}_n)) = \frac{(g'(\theta))^2}{n\, I(\theta)}\,.$$

But this is equivalent to the assertion that the correlation between $T(\boldsymbol{X}_n)$ and $\dot{\ell}_n(\boldsymbol{X}_n, \theta)$ is equal to 1 or -1.

This means that $\dot{\ell}_n(\boldsymbol{X}_n, \theta)$ can be expressed as a *linear function* of $T(\boldsymbol{X}_n)$.

In fact, this is a necessary and sufficient condition for the information bound to be attained by the variance of $T(\boldsymbol{X}_n)$.

It turns out that this is generally difficult to achieve. Thus, there will be many different functions of θ, for which best unbiased estimators will exist but whose variance will not hit the information bound. The example below will illustrate this point.

Example: Let $X_1, X_2, \ldots, X_n$ be i.i.d Ber(θ). We have,

$$f(x, \theta) = \theta^x (1-\theta)^{1-x} \qquad \text{for } x = 0, 1.$$

Thus,

$$\ell(x, \theta) = x \log \theta + (1-x) \log(1-\theta)\,,$$

$$\dot{\ell}(x, \theta) = \frac{x}{\theta} - \frac{1-x}{1-\theta}$$

and

$$\ddot{\ell}(x, \theta) = -\frac{x}{\theta^2} - \frac{1-x}{(1-\theta)^2}\,.$$

Thus,

$$\dot{\ell}_n(\boldsymbol{X}_n, \theta) = \sum_{i=1}^n \dot{\ell}(X_i, \theta) = \frac{\sum_{i=1}^n X_i}{\theta} - \frac{n - \sum_{i=1}^n X_i}{1-\theta}\,.$$

Recall that the MLE solves $\dot{\ell}_n(\boldsymbol{X}_n, \theta) = 0$.

Check that in this situation, this gives you precisely $\overline{X}_n$ as your MLE.

Let us compute the information $I(\theta)$. We have,

$$I(\theta) = -\mathbb{E}_\theta[\ddot{\ell}(X_1, \theta)] = \mathbb{E}_\theta\left(\frac{X_1}{\theta^2} + \frac{1 - X_1}{(1-\theta)^2}\right) = \frac{1}{\theta} + \frac{1}{1-\theta} = \frac{1}{\theta\,(1-\theta)}.$$

Thus,

$$I_n(\theta) = nI(\theta) = \frac{n}{\theta(1-\theta)}.$$

Consider unbiased estimation of $\Psi(\theta) = \theta$ based on $\boldsymbol{X}_n$. Let $T(\boldsymbol{X}_n)$ be an unbiased estimator of θ. Then, by the information inequality,

$$\mathrm{Var}_\theta(T(\boldsymbol{X}_n)) \geq \frac{\theta(1-\theta)}{n}.$$

Note that the variance of $\overline{X}$ is precisely $\theta(1-\theta)/n$, so that it is the MVUE of θ. Note that,

$$\dot{\ell}_n(\boldsymbol{X}_n, \theta) = \frac{n\overline{X}}{\theta} - \frac{n(1-\overline{X})}{1-\theta} = \left(\frac{n}{\theta} + \frac{n}{1-\theta}\right)\overline{X} - \frac{n}{1-\theta}.$$

Thus, $\overline{X}_n$ is indeed linear in $\dot{\ell}_n(\boldsymbol{X}_n, \theta)$.

Consider now estimating a different function of θ, say $g(\theta) = \theta^2$.

This is the probability of getting two consecutive heads. Suppose we try to find an unbiased estimator of this parameter.

Then $S(\boldsymbol{X}_n) = X_1X_2$ is an unbiased estimator ($\mathbb{E}_\theta(X_1X_2) = \mathbb{E}_\theta(X_1)\mathbb{E}_\theta(X_2) = \theta^2$), but then so is X_iX_j for any $i \neq j$.

We can find the best unbiased estimator of θ^2 in this model by using techniques beyond the scope of this course — it can be shown that any estimator $T(\boldsymbol{X}_n)$ that can be written as a function of $\overline{X}$ and is unbiased for θ^2 is an MVUE (and indeed there is one such).

Verify that,

$$T^*(\boldsymbol{X}_n) = \frac{n\,\overline{X}^2 - \overline{X}}{n-1}$$

is unbiased for θ^2 and is therefore an (in fact *the*) MVUE.

However, the variance of $T^*(\boldsymbol{X}_n)$ does not attain the information bound for estimating $g(\theta)$ which is $4\theta^3(1-\theta)/n$ (verify).

This can be checked by direct (somewhat tedious) computation or by noting that $T^*(\boldsymbol{X}_n)$ is not a linear function of $\dot{\ell}_n(\boldsymbol{X}_n, \theta)$.

The question then is whether we can propose an estimator of θ^2 that does achieve the bound, at least approximately, in the long run.

It turns out that this is actually possible. Since the MLE of θ is $\overline{X}$, the MLE of $g(\theta)$ is proposed as the plug-in value $g(\overline{X}) = \overline{X}^2$.

This is *not an unbiased estimator of* $g(\theta)$ in finite samples, but has excellent behavior in the long run. In fact,

$$\sqrt{n}\,(g(\overline{X}_n) - g(\theta)) \to_d N(0, 4\theta^3(1-\theta)).$$

Thus for large values of n, $g(\overline{X})$ behaves approximately like a normal random variable with mean $g(\theta)$ and variance $4\theta^3(1-\theta)/n$.

In this sense, $g(\overline{X}_n)$ is *asympotitcally (in the long run) unbiased* and *asympotically efficient* (in the sense that it has minimum variance).

This is not an isolated phenomenon but happens repeatedly. To this end, here is a general proposition — consequence of the delta method.

Proposition 7.1. *Suppose T_n is an estimator of $g(\theta)$ (based on i.i.d observations, $X_1, X_2, \ldots, X_n$ from P_θ) that satisfies:*

$$\sqrt{n}(T_n - g(\theta)) \to_d N(0, \sigma^2(\theta)).$$

Here $\sigma^2(\theta)$ is the limiting variance and depends on the underlying parameter θ. Then, for a continuously differentiable function h such that $h'(g(\theta)) \neq 0$, we have:

$$\sqrt{n}\,(h(T_n) - h(g(\theta))) \to_d N(0, h'(g(\theta))^2\,\sigma^2(\theta))\,.$$

Here is an important proposition that establishes the limiting behavior of the MLE.

Proposition 7.2. *If $\hat{\theta}_n$ is the MLE of θ obtained by solving*

$$\sum_{i=1}^{n} \dot{\ell}(X_i, \theta) = 0,$$

then the following representation for the MLE is valid:

$$\sqrt{n}(\hat{\theta}_n - \theta) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} I(\theta)^{-1} \dot{\ell}(X_i, \theta) + r_n,$$

where r_n converges to 0 in probability. It follows by a direct application of the CLT that,

$$\sqrt{n}(\hat{\theta}_n - \theta) \to_d N(0, I(\theta)^{-1}).$$

The above result shows MLE $\hat{\theta}$ is (asymptotically) the best possible estimator: Not only does its long term distribution center around θ, the quantity of interest, its distribution is also less spread out than that of any "reasonable" estimator of θ. If S_n is a "reasonable" estimator of θ, with

$$\sqrt{n}\,(S_n - \theta) \to_d N(0, \xi^2(\theta))\,,$$

then $\xi^2(\theta) \geq I(\theta)^{-1}$.

We can now deduce the limiting behavior of the MLE of $g(\theta)$ given by $g(\hat{\theta}_n)$ for any smooth function g such that $g'(\theta) \neq 0$.

Combining Proposition 7.2 with Proposition 7.1 yields (take $g(\theta) = \theta, T_n = \hat{\theta}_n$ and $h = g$)

$$\sqrt{n}(g(\hat{\theta}_n) - g(\theta)) \to_d N(0, g'(\theta)^2\, I(\theta)^{-1}).$$

Thus, for large n,

$$g(\hat{\theta}_n) \sim_{appx} N(g(\theta), g'(\theta)^2\,(n\, I(\theta))^{-1})\,.$$

Thus $g(\hat{\theta}_n)$ is asympotically unbiased for $g(\theta)$ (unbiased in the long run) and its variance is approximately the information bound for unbiased estimators of $g(\theta)$.

Constructing confidence sets for θ: Suppose that, for simplicity, θ takes values in a subset of $\mathbb{R}$. Since,

$$\sqrt{n}(\hat{\theta} - \theta) \to_d N(0, I(\theta)^{-1}),$$

it follows that

$$\sqrt{n\, I(\theta)}(\hat{\theta} - \theta) \to_d N(0, 1).$$

Thus, the left side acts as an *approximate pivot* for θ. We have,

$$\mathbb{P}_\theta \left(-z_{\alpha/2} \leq \sqrt{nI(\theta)}(\hat{\theta} - \theta) \leq z_{\alpha/2} \right) \approx 1 - \alpha.$$

An approximate level $1 - \alpha$ confidence set for θ is obtained as

$$\left\{ \theta : -z_{\alpha/2} \leq \sqrt{nI(\theta)}(\hat{\theta} - \theta) \leq z_{\alpha/2} \right\}.$$

To find the above confidence set, one needs to solve for all values of θ satisfying the inequalities in the above display; this can however be a potentially complicated exercise depending on the functional form for $I(\theta)$.

However, if the sample size n is large $I(\hat{\theta})$ can be expected to be close to $I(\theta)$ with high probability and hence the following is also valid:

$$P_\theta \left[-z_{\alpha/2} \leq \sqrt{n\, I(\hat{\theta})}(\hat{\theta} - \theta) \leq z_{\alpha/2} \right] \approx 1 - \alpha. \quad (\star\star)$$

This immediately gives an approximate level $1-\alpha$ CI for θ as:

$$\left[\hat{\theta} - \frac{1}{\sqrt{nI(\hat{\theta})}}\, z_{\alpha/2}, \hat{\theta} + \frac{1}{\sqrt{nI(\hat{\theta})}} z_{\alpha/2}\right].$$

Let's see what this implies for the Bernoulli example discussed above. Recall that $I(\theta) = (\theta\,(1-\theta))^{-1}$ and $\hat{\theta} = \overline{X}$. The approximate level $1-\alpha$ CI is then given by,

$$\left[\overline{X} - \sqrt{\frac{\overline{X}\,(1-\overline{X})}{n}}\, z_{\alpha/2}, \overline{X} + \sqrt{\frac{\overline{X}\,(1-\overline{X})}{n}}\, z_{\alpha/2}\right].$$

Exercise: Find explicitly

$$\left\{\theta : -z_{\alpha/2} \leq \sqrt{n\,I(\theta)}(\hat{\theta} - \theta) \leq z_{\alpha/2}\right\}$$

in the following cases (a) $X_1, X_2, \ldots, X_n$ are i.i.d Bernoulli(θ). (b) $X_1, X_2, \ldots, X_n$ are i.i.d Pois(θ).

You will see that this involves solving for the roots of a quadratic equation. As in the Bernoulli example, one can also get an approximate CI for θ in the Poisson setting on using $(\star\star)$. Verify that this yields the following level $1-\alpha$ CI for θ:

$$\left[\overline{X} - \sqrt{\frac{\overline{X}}{n}}\, z_{\alpha/2}, \overline{X} + \sqrt{\frac{\overline{X}}{n}}\, z_{\alpha/2}\right].$$

The recipe $(\star\star)$ is somewhat unsatisfactory because it involves one more level of approximation in that $I(\theta)$ is replaced by $I(\hat{\theta})$ (note that there is already one level of approximation in that the pivots being considered are only approximately $N(0,1)$ by the CLT).