Parameter Estimation

The RV which estimates a parameter of a distribution (such as $\bar{X}$ used to estimate the mean μ of a Normal distribution) is called an ESTIMATOR of the parameter; being a RV, it has a distribution of its own (the so-called *sampling distribution,* often too difficult to find). Once a RIS of size n is taken and the resulting *value* of any such estimator is referred to as the parameter's ESTIMATE.

We start by assuming (to simplify things) that there is only **one parameter**, say θ, of the *sampled* distribution to estimate (if there are other parameters, their exact value must be known). Also note that some parameters can have any value for a certain interval (such as p of the binomial distribution) while others need to be integers (such as n of the binomial distribution) - here, we will consider the former type *only.*

First issue is: what are *desired properties* of an estimator, say $\hat{\theta}(X_1, X_2, ..., X_n)$, of a parameter θ ?

The most important property is to be UNBIASED, meaning

$$\mathbb{E}(\hat{\theta}) = \theta$$

or, as a second best, *asymptotically unbiased* , i.e.

$$\mathbb{E}(\hat{\theta}) \underset{n\to\infty}{\longrightarrow} \theta$$

The $\mathbb{E}(\hat{\theta}) - \theta$ difference is called the BIAS of an estimator, and is identically equal to 0 in the former case, and tends to zero as $n \to \infty$ in the latter case.

The second essential property of an estimator is to have its variance as small as possible.

Some definitions:

CONSISTENT ESTIMATOR must meet two properties: be asymptotically unbiased, and have a variance which tends, as $n \to \infty$, to zero. This means that, with enough sampling, we can always pinpoint the value of θ to any required accuracy. Yet, such an estimator may still be very inefficient!

MINIMUM VARIANCE UNBIASED ESTIMATOR (MVUE) is an (fully) *unbiased estimator* whose *variance* is smaller or equal to the variance of any other *unbiased* estimator, for all potential values of θ. This thus defines the 'best' estimator, but things are not so easy: given an unbiased estimator, it is clearly impossible to compare its variance with every other such estimator (infinitely many of them), to see whether it fits the bill. Luckily, there is a *theoretical lower bound* (a function of θ) on the variance of all unbiased estimators of the REGULAR type; when an estimator achieves this bound, it must be automatically MVUE (REGULAR means that the parameter appears in the pdf of the sampled distribution, but *not* in the corresponding support). Note that most distributions we have ever seen were regular (uniform distribution being the only exception).

Rao-Cramér inequality (sometimes they reverse the names)
Preliminaries: Define a new RV by

$$U = \frac{\partial \ln f(X \mid \theta)}{\partial \theta}$$

Note that this notation does not imply conditional probability; the bar only separates the variable(s) from the parameter(s).

EXAMPLE: For $\mathcal{E}(\beta)$, this yields

$$U = \frac{\partial \left(-\frac{X}{\beta} - \ln \beta\right)}{\partial \beta} = \frac{X}{\beta^2} - \frac{1}{\beta}$$

It follows that (in general)

$$\mathbb{E}(U) = \int (\ln \overset{\bullet}{f}) \cdot f \; dx = \int \frac{\overset{\bullet}{f}}{f} \cdot f \; dx = \int \overset{\bullet}{f} \; dx = \frac{\partial}{\partial \theta} \int f \; dx = 0$$

where f is a shorthand for the original pdf, $\overset{\bullet}{\ldots}$ denotes differentiation with respect to θ, and the dx integration is over all x values. Differentiating one more time results in

$$\begin{aligned}
&\int (\ln \overset{\bullet\bullet}{f}) \cdot f \; dx + \int (\ln \overset{\bullet}{f}) \cdot \overset{\bullet}{f} \; dx = \int (\ln \overset{\bullet\bullet}{f}) \cdot f \; dx \\
&+ \int (\ln \overset{\bullet}{f}) \cdot \frac{\overset{\bullet}{f}}{f} \cdot f \; dx = \int (\ln \overset{\bullet\bullet}{f}) \cdot f \; dx + \int (\ln \overset{\bullet}{f})^2 \cdot f \; dx \\
&= \int (\ln \overset{\bullet\bullet}{f}) \cdot f \; dx + \int U^2 \cdot f \; dx = 0
\end{aligned}$$

This implies that the variance of U can be alternately computed by

$$\mathrm{Var}(U) = -\mathbb{E}\left(\frac{\partial^2 \ln f(X \mid \theta)}{\partial \theta^2}\right)$$

Actual derivation: Now, let Θ be an *unbiased* estimator of θ, i.e.

$$\mathbb{E}(\Theta) = \int \cdots \int \Theta \cdot \prod_{i=1}^{n} f_i \; dx_1 ... dx_n = \theta$$

where f_i is a shorthand for $f(x_i \mid \theta)$. Differentiating with respect to θ yields

$$\int \cdots \int \Theta \cdot \sum_{j=1}^{n} \frac{\overset{\bullet}{f_j}}{f_j} \prod_{i=1}^{n} f_i \; dx_1 ... dx_n = 1$$

or, equivalently

$$\begin{aligned}
1 &= \mathbb{E}\left(\Theta \cdot \sum_{j=1}^{n} U_j\right) = \mathrm{Cov}\left(\Theta, \sum_{j=1}^{n} U_j\right) \\
&\leq \sqrt{\mathrm{Var}(\Theta) \cdot n \; \mathrm{Var}(U)}
\end{aligned}$$

implying

$$\mathrm{Var}\,(\Theta) \geq \frac{1}{n\ \mathrm{Var}\,(U)} = \frac{-1}{n\ \mathbb{E}\left(\frac{\partial^2 \ln f(X\ \mid\ \theta)}{\partial \theta^2}\right)}$$

The RHS expression is the Rao-Cramer bound on the variance of *any* unbiased estimator (RCV for short). ■

For the $\mathcal{E}\,(\beta)$ example this means that the variance of an unbiased estimator of β can never be smaller than $\frac{\beta^2}{n}$ (whatever the β value is). Since $\bar{X}$ has exactly this variance, it is the MVUE (stop searching for anything better).

When the distribution is discrete, integration becomes summation and $f(X \mid \theta)$ becomes pmf instead of pdf but the rest is the same. Thus, for a geometric distribution with $f(X) = p\ (1-p)^{X-1}$, RCV is equal to

$$\frac{-1}{n\ \mathbb{E}\left(\frac{\partial^2 (\ln p + (X-1)\ln(1-p))}{\partial \theta^2}\right)} = \frac{p^2(1-p)}{n}$$

For an unbiased estimator, divide RCV by the estimator's variance to find its EFFICIENCY. Assuming 'large' sample size (which is our main emphasis here) and a *regular* case, there is always a way of finding *asymptotically* unbiased and asymptotically efficient estimator by Maximum Likelihood technique (described later).

But there are other ways of finding an estimator, which we need to discuss first.

The simplest way is of course by sheer guessing; we did it trying to find the center of Cauchy distribution: the $\bar{X}$ guess was not even a consistent estimator, but $\tilde{X}$ proved to be quite decent ('almost' the best, as we will see later).

But now, we will try to be a bit more systematic.

An old technique (practically obsolete, but we will still go over it) is the

Method of moments (MM)

With one parameter (the current assumption), it works as follows: make the expected value of X equal to $\bar{X}$ and solve for the parameter (say θ) - the solution (always a function of $\bar{X}$) is your estimator.

Note that

- this will not work when $\mathbb{E}\,(X)$ is indefinite (or infinite), or when it is *not* a function of θ,
- but it does not require the case to be regular (even though the corresponding MM estimator is then rather inferior).
- The resulting $\hat{\theta} = g(\bar{X})$ is then always asymptotically (i.e. when n is 'large') Normal, with the asymptotic mean of $g(\mu) = \theta$ and the asymptotic variance of

$$\frac{g'(\mu)^2 \sigma^2}{n}$$

since

$$g(\bar{X}) \simeq g(\mu) + g'(\mu) \cdot (\bar{X} - \mu) + ...$$

Examples:

1. Estimating p of Geometric distribution can be done by $\hat{p} = \frac{1}{\bar{X}}$, whose *sampling* distribution is approximately (as n get bigger) Normal with the asymptotic mean of p (that's always the case - that's how it has been arranged) and asymptotic variance of $\frac{p^2(1-p)}{n}$ (RCV, when lucky, we can get the 'best' estimator even by this technique). Note that $\sqrt{\frac{\hat{p}^2(1-\hat{p})}{n}}$ - substituting our *estimate* into the variance formula - called the STANDARD ERROR gives a good idea about the estimate's accuracy (this should be done routinely with all numerical estimates).

2. Estimating b of $\mathcal{U}(0, \theta)$ distribution leads to $\hat{\theta} = 2\bar{X}$, whose sampling distribution is approximately Normal with the (exact) mean of θ and variance of $\frac{\theta^2}{3n}$. Later on, we will see that this is a very inefficient way of estimating θ.

The next technique is based on the so-called **sufficient statistic** (SS). To find it, we simplify the sample's joint pdf, namely

$$\prod_{i=1}^{n} f(X_i \mid \theta)$$

as much as possible and delete factors free of θ (if any) and factors free of X_i (if any). If the remaining expression contains only a single combination of the X_is (let us denote it Φ) this combination is the corresponding SUFFICIENT STATISTIC for estimating θ (note that it may not always exist; when it does, it is normally either a sum or a product of n terms - when it is a product, we should immediately convert it into a sum by taking its ln). One can prove that a sufficient statistic contains all the sample's information about the value of θ, implying that we can always match or exceed the quality of all other unbiased estimators by a properly designed function of Φ. To find such **sufficient estimator** , we do something similar to the MM technique (see the examples). The big advantage of this technique is that it often works even in non-regular cases (in which case the Maximum Likelihood technique below will find it somehow more directly, so the concept of sufficiency is more of a theoretical interest).

Examples:

1. The same geometric distribution as before (regular case). The joint pdf is

$$\prod_{i=1}^{n} p(1-p)^{X_i - 1} = \frac{p^n}{(1-p)^n} \cdot (1-p)^{\sum_{i=1}^{n} X_i}$$

implying that $\sum_{i=1}^{n} X_i$ is sufficient statistic. Since its expected value is $\frac{n}{p}$, making these equal and solving for p yields the same estimator as MM (now we have a second reason to claim that this is the best we can do).

2. The same uniform distribution (non-regular case). The joint pdf is

$$\frac{1}{\theta^n}\prod_{i=1}^{n} G_{0,b}[X_i] = \frac{1}{\theta^n}\cdot G_{0,\theta}[X_{(n)}]\cdot G_{0,X_{(n)}}[X_{(1)}]$$

where $G_{a,b}[X]$ equals to 1 when $a \le X \le b$ and 0 otherwise. This implies that $X_{(n)}$ is a sufficient statistic. Since $\frac{X_{(n)}}{\theta} \in \mathsf{beta}(n,1)$, the expected value of $X_{(n)}$ is $\frac{n}{n+1}\cdot\theta$. Making them equal and solving for θ yields $\hat{\theta} = \frac{n+1}{n}\cdot X_{(n)}$. This is now a (fully) unbiased estimator whose variance is $\left(\frac{n+1}{n}\right)^2\theta^2\frac{n}{(n+1)^2(n+2)} = \frac{\theta^2}{n(n+2)}$. The RELATIVE EFFICIENCY of the MM estimator to this SS estimator is $\frac{3}{n+2}$ (going to 0 as n increases)!

3. Sampling a (special case of beta) distribution with $f(x) = c\cdot x^{c-1}$ where $0 < x < 1$ and $c > 0$ (regular case). The joint pdf is

$$c^n\left(\prod_{i=1}^{n} X_i\right)^{c-1}$$

making $\prod_{i=1}^{n} X_i$ or, equivalently, $\sum_{i=1}^{n}\ln X_i$ the corresponding SS for estimating c. The expected value of the latter is $-\frac{n}{c}$; solving for c yields

$$\hat{c} = -\frac{n}{\sum_{i=1}^{n}\ln X_i} = -\frac{1}{\overline{\ln X}} \simeq c + c^2\left(\overline{\ln X} + \frac{1}{c}\right) + ...$$

The last expansion makes it obvious that its asymptotic variance is $\frac{(c^2)^2}{c^2 n} = \frac{c^2}{n}$; this is guaranteed to be the same as RCV (as it does, check it out).

Finally, the technique which always works and is guaranteed to provide the best asymptotic estimator (whenever it exists - and that covers all regular cases and all cases having a sufficient statistic, which is all we need) is the **maximum likelihood** (ML) estimation. It works as follows: considering the joint sample pdf (or its ln, if it's more convenient) a function of θ (keeping the X_is fixed - like they have already been observed), find θ (a function of the X_is) which maximizes this joint pdf (or its ln) - that yields the corresponding ML estimator. One can show that in a *regular* case, its asymptotic variance is RCV (its asymptotic mean is always θ) and its sampling distribution is approximately Normal. The only (small - we live in the computer age) problem one may encounter (rarely) is that the solution may not always be a simple 'analytic' function of the X_is - sometimes, we may be able to get the corresponding *numerical* solution only (given the observed, numerical values of the X_is). But that's ultimately what we always want in the end!

Examples:

1. The same geometric distribution as before. The ln of the joint pdf is

$$n \ln p + \ln(1-p) \sum_{i=1}^{n} (X_i - 1)$$

The corresponding first derivative with respect to p leads to the following so-called NORMAL EQUATION

$$\frac{n}{p} + \frac{n - \sum_{i=1}^{n} X_i}{1-p} = 0$$

Solving for p yields the same old $\frac{1}{\bar{X}}$.

2. The same uniform distribution. Maximizing

$$\frac{1}{\theta^n} \cdot G_{0,\theta}[X_{(n)}] \cdot G_{0,X_{(n)}}[X_{(1)}]$$

is achieved by making θ as small as possible while keeping it no less than $X_{(n)}$. That makes $X_{(n)}$ itself the MLE, having the expected value of $\frac{n}{n+1}\theta$ (the small bias tends to 0 as $n \to \infty$) and the variance of $\frac{n}{(n+1)^2(n+2)}\theta^2$.

3. The same $\mathsf{beta}(c, 1)$ distribution. To maximize

$$n \ln c + (c-1) \sum_{i=1}^{n} \ln X_i$$

solve

$$\frac{n}{c} + \sum_{i=1}^{n} \ln X_i = 0$$

which yields the same $\hat{c} = -\frac{1}{\bar{X}}$ as the SS technique (perhaps a bit faster). To get the corresponding RCV is now also so much easier, since the second derivative (with respect to c) of $\ln c + (c-1)\ln x$ is $-\frac{1}{c^2}$.

4. Sampling from $\mathcal{C}(\tilde{\mu}, 1)$. To maximize

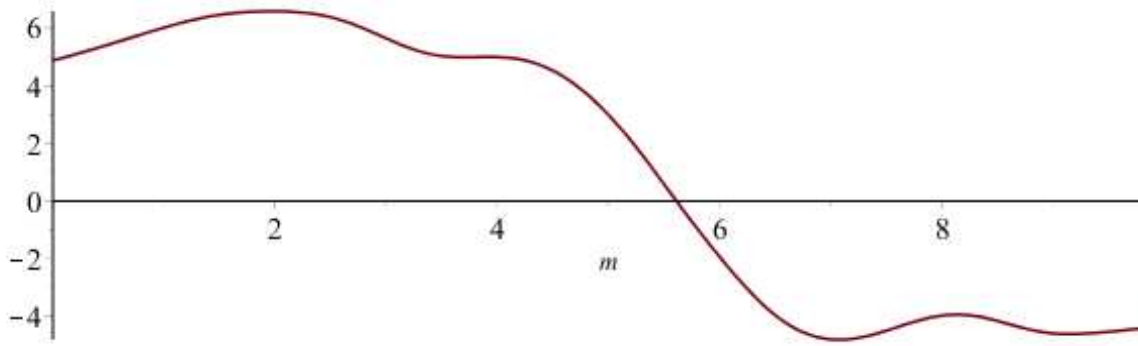
$$-n \ln \pi - \sum_{i=1}^{n} \ln\left(1 + (X_i - \tilde{\mu})^2\right)$$

we need to solve

$$2 \sum_{i=1}^{n} \frac{X_i - \tilde{\mu}}{1 + (X_i - \tilde{\mu})^2} = 0$$

This can be done only numerically. Given the following sample of 30 observations from $\mathcal{C}(\tilde{\mu}, 1)$: 6.5,8.3,2.6,8.6,5.4,1.8,4.2, 5.1,12.6,14.3,3.1,16.2, 12.6,5.0,6.4,2.9,4.7,8.7,6.3,13.0,5.5,-4.2,7.0,9.8, 5.6,6.1,6.0,4.6,5.5,3.3 it is

quite easy to ask Maple for a numerical solution to the previous equation (use 'fsolve'), plotting its LHS cannot hurt either, getting



The estimate turns out to be 5.610, its standard error is the square root of the corresponding RCV, namely $\sqrt{\frac{2}{n}} = 0.2582$.