Chapter 5: Estimation

By Joan Llull*

Probability and Statistics.
QEM Erasmus Mundus Master. Fall 2016

Main references:

— Goldberger: 11.1-11.13, 12.1, 12.2

— Lindgren: 8.1, 8.2, 7.7, 7.10, 2.8, 2.3, 8.7

I. Analogy Principle

The ***estimation problem*** consists of obtaining an approximation to a population characteristic θ_0 combining the information provided in a random sample. An ***estimator*** is a rule for calculating an ***estimate*** of a given quantity based on observed data. Hence, an estimator is a function from the sample $(X_1, ..., X_N)$, which we denote by $\hat{\theta}(X_1, ..., X_N)$, and an estimate is the result of implementing the estimator to the given sample, denoted by $\hat{\theta}(x_1, ..., x_N)$. In general, when the context is clear, we typically abuse of notation and simply use $\hat{\theta}$ to denote both the estimator and the estimate. The estimator is a statistic, and the estimate is a particular realization of this statistic.

A general rule to decide which estimator to implement is to define in the sample a statistic that satisfies similar properties to those satisfied by the ***true parameter*** in the population. This is called the ***analogy principle***. For example, to estimate the population mean, we often compute the sample mean; to estimate the variance, we often compute the sample variance; to estimate the median, we compute the median in the sample.

II. Desirable Properties of an Estimator

We define now certain criteria to determine the "quality" of an estimator. An estimator is good if it is a good approximation to the true parameter no matter which is the true value of the parameter. For example, consider a population with mean θ. If we are interested in estimating θ, we could propose as an estimator $\hat{\theta} = 3$ (regardless of what information I have in my sample). This estimator will be very good only if $\theta = 3$, but, in general, it is going to be bad. Instead, if we propose $\bar{X}_N$, the estimator will be generally good, as we expected $\bar{X}_N$ to be centered around θ, no matter what is the real value of θ.

* Departament d'Economia i Història Econòmica. Universitat Autònoma de Barcelona. Facultat d'Economia, Edifici B, Campus de Bellaterra, 08193, Cerdanyola del Vallès, Barcelona (Spain). E-mail: joan.llull[at]movebarcelona[dot]eu. URL: http://pareto.uab.cat/jllull.

A measure of how good is an estimator is the ***mean squared error*** (MSE):

$$MSE(\hat{\theta}) \equiv \mathbb{E}[(\hat{\theta} - \theta)^2]. \tag{1}$$

The MSE can be decomposed as follows:

$$\begin{aligned} MSE(\hat{\theta}) &= \mathbb{E}[(\hat{\theta} - \theta)^2] = \mathbb{E}[(\hat{\theta} - \mathbb{E}[\hat{\theta}] + \mathbb{E}[\hat{\theta}] - \theta)^2] \\ &= \mathbb{E}[(\hat{\theta} - \mathbb{E}[\hat{\theta}])^2] + \mathbb{E}[(\mathbb{E}[\hat{\theta}] - \theta)^2] + 2\,\mathbb{E}[\hat{\theta} - \mathbb{E}[\hat{\theta}]](\mathbb{E}[\hat{\theta}] - \theta) \\ &= \mathrm{Var}(\hat{\theta}) + (\mathbb{E}[\hat{\theta}] - \theta)^2. \end{aligned} \tag{2}$$

The difference $\mathbb{E}[\hat{\theta}] - \theta$ is called ***bias***. Thus, a good estimator, which is one that has small MSE, is an estimator that have small variance (precision), and small (or no) bias. An ***unbiased*** estimator is an estimator that has zero bias, that is, an estimator that satisfies $\mathbb{E}[\hat{\theta}] = \theta$.

For example, we saw in Chapter 4 that $\mathrm{Var}(\bar{X}_N) = \frac{\mathrm{Var}(X)}{N}$, and $\mathbb{E}[\bar{X}_N] = \mathbb{E}[X]$. Thus, $MSE(\bar{X}_N) = \frac{\mathrm{Var}(X)}{N}$. We could do the same for s^2_N or for $\hat{\sigma}^2_N$. Indeed, in an exercise in this chapter, you will compare the MSE of these two estimators of $\mathrm{Var}(X)$. There, you will see that $\hat{\sigma}^2_N$, even though it is a biased estimator of $\mathrm{Var}(X)$, has a lower MSE than the unbiased estimator s^2, no matter what the value of σ^2 or the sample size are. There are other examples where this is not true in general, but only for some values of the true parameter or for certain sample sizes.

Then the question is: what is more important, absence of a bias or lower MSE? There is obviously a trade-off, and no clear answer. What is clear, though, is that among the estimators that are unbiased, we prefer those that have less variance. We say that an estimator is more ***efficient*** than another if, being both unbiased, it has lower variance. Among all the estimators that are unbiased, the on that has the minimum possible variance is called ***best unbiased estimator*** (BUE), the minimum variance unbiased estimator, or simply the most efficient estimator. A more restrictive criterion is to search for the ***best linear unbiased estimator*** (BLUE), which is the best unbiased estimator of the class of estimators that are linear combinations of the data.

III. Moments and Likelihood Problems

A ***moments problem*** is defined by two equivalent conditions on the parameter of interest:

- It optimizes an expectation function. E.g.:

$$\mu = \arg\min_c \mathbb{E}[(Y - c)^2] \quad \text{or} \quad (\alpha, \beta) = \arg\min_{(a,b)} \mathbb{E}[(Y - a - bX)^2]. \tag{3}$$

- It solves a moment condition. E.g.:

$$\mathbb{E}[(Y-\mu)]=0 \quad \text{or} \quad \mathbb{E}\left[(Y-\alpha-\beta X)\begin{pmatrix}1\\X\end{pmatrix}\right]=\begin{pmatrix}0\\0\end{pmatrix}. \tag{4}$$

Notice that any of these two requirements makes any assumption on the population distribution beyond the relevant moments. A method of moments estimator would use sample analogs to these conditions, and would obtain $\hat{\mu}$ or $(\hat{\alpha},\hat{\beta})$ that satisfy them.

The ***likelihood problem*** is defined by a more completely specified environment. We assume that the population distribution is a known function, except for the parameters of interest, which are unknown. For instance, we know that the population pdf is $\{f(X;\theta):\theta\in\Theta\}$, where Θ is the space of parameters for which the function is defined, but we do not know the value $\theta=\theta_0$, which is the true parameter value.

The fact that we know more information about the population of interest allows us to obtain better estimators from a given sample, provided that is extra information (the functional form of the population distribution) is correct.

As we prove below, the true parameter satisfies:

$$\theta_0=\arg\max_{\theta\in\Theta}\mathbb{E}[\ln f(X;\theta)]. \tag{5}$$

To prove it, consider the first order condition of the above optimization problem:

$$\begin{aligned}
&\mathbb{E}\left[\frac{\partial \ln f(X;\theta)}{\partial\theta'}\right]\equiv\mathbb{E}[z(X;\theta)]=0\\
&\Leftrightarrow\int_{-\infty}^{\infty}z(X;\theta)f(X;\theta_0)dX=0\\
&\Leftrightarrow\int_{-\infty}^{\infty}\frac{\partial \ln f(X;\theta)}{\partial\theta'}f(X;\theta_0)dX=0\\
&\Leftrightarrow\int_{-\infty}^{\infty}\frac{1}{f(X;\theta)}\frac{\partial f(X;\theta)}{\partial\theta'}f(X;\theta_0)dX=0
\end{aligned} \tag{6}$$

Now, note that, because $f(X;\theta_0)$ is a pdf, it must integrate to 1:

$$\begin{aligned}
&\int_{-\infty}^{\infty}f(X;\theta_0)dX=1\\
&\Leftrightarrow\frac{\partial}{\partial\theta'}\int_{-\infty}^{\infty}f(X;\theta_0)dX=0\\
&\Leftrightarrow\int_{-\infty}^{\infty}\frac{\partial f(X;\theta_0)}{\partial\theta'}dX=0.
\end{aligned} \tag{7}$$

(note that we are assuming that the range of integration does not depend on θ_0). Hence, replacing $\theta = \theta_0$ in Equation (6), we obtain the same expression as in Equation (7), and thus we conclude that θ_0 is a solution of the problem in Equation (5), because it satisfies the first order condition. We call $z(X;\theta_0)$ the ***score***, and the fact that $\mathbb{E}[z(X;\theta_0)] = 0$ is called ***zero expected score*** condition. Finally, note that the likelihood problem can also be seen as a moments problem.

IV. Maximum Likelihood Estimation

Consider a sample of size X, $(X_1, ..., X_N)$. In a likelihood problem, the pdf of this sample is known (up to parameter values), as we assume that $\{f(X;\theta) : \theta \in \Theta\}$ is known. The pdf of the sample written as a function of θ is known as the ***likelihood function***, and is equal to:

$$L_N(\theta) = \prod_{i=1}^{N} f(X_i;\theta). \tag{8}$$

$L_N(\theta)$ can be seen both as a function of the data for a given parameter θ (the pdf of the sample if $\theta = \theta_0$), or a function $\{L_N(\theta) : \theta \in \Theta\}$ of the parameter for a fixed sample (the likelihood function). The log-likelihood function is defined as:

$$\mathcal{L}_{\mathrm{N}}(\theta) \equiv \ln L_N(\theta) = \sum_{i=1}^{N} \ln f(X_i;\theta). \tag{9}$$

Given a sample with log-likelihood $\mathcal{L}_{\mathrm{N}}(\theta)$, we define the ***Maximum Likelihood Estimator*** (MLE) as:

$$\hat{\theta}_{MLE} \equiv \arg\max_{\theta\in\Theta} \mathcal{L}_{\mathrm{N}}(\theta) = \arg\max_{\theta\in\Theta} \frac{1}{N} \sum_{i=1}^{N} \ln f(X_i;\theta). \tag{10}$$

Note that this estimator is the sample analog of the condition that the true parameter θ_0 satisfies in the population, as described by Equation (5). The idea of the MLE is to approximate θ_0 by the value of θ that maximizes the likelihood (probability in the discrete case, density in the continuous case) of obtaining the sample that we observe. This is called the ***likelihood principle***.

The MLE satisfies the first order conditions of the optimization problem in Equation (10):

$$\frac{1}{N} \sum_{\mathrm{i}=1}^{\mathrm{N}} \frac{\partial \ln f(X_i;\hat{\theta}_{MLE})}{\partial \theta'} = 0. \tag{11}$$

These conditions are sample analogs of the zero expected score rule.

V. The Cramer-Rao Lower Bound

The variance of the score is called the ***Fisher information*** (or simply information), $I(\theta_0) \equiv \text{Var}(z(X;\theta_0))$. The name information is motivated by the fact that $I(\theta_0)$ is a way of measuring the amount of information that a random variable X contains about an unknown parameter θ_0 for the distribution that models X (variation in the slope of the likelihood function). The ***information equality*** is an interesting property satisfied by the variance of the score:

$$\text{Var}(z(X;\theta_0)) = \mathbb{E}[z(X;\theta_0)z(X;\theta_0)'] = -\,\mathbb{E}\left[\frac{\partial^2 \mathcal{L}_{\text{N}}(\theta_0)}{\partial\theta\partial\theta'}\right]. \tag{12}$$

To prove it, note that, if the expected score is equal to zero, the derivative of the expected score should also be equal to zero:

$$\begin{aligned}
&\frac{\partial}{\partial\theta}\int_{-\infty}^{\infty}\frac{\partial\,\mathcal{L}_{\text{N}}(\theta_0)}{\partial\theta'}f(X;\theta_0)dX = 0\\
&\Leftrightarrow \int_{-\infty}^{\infty}\frac{\partial\,\mathcal{L}_{\text{N}}(\theta_0)}{\partial\theta\partial\theta'}f(X;\theta_0)dX + \int_{-\infty}^{\infty}\frac{\partial\,\mathcal{L}_{\text{N}}(\theta_0)}{\partial\theta}\frac{\partial f(X;\theta_0)}{\partial\theta'}dX = 0\\
&\Leftrightarrow \mathbb{E}\left[\frac{\partial^2\,\mathcal{L}_{\text{N}}(\theta_0)}{\partial\theta\partial\theta'}\right] + \int_{-\infty}^{\infty}\frac{\partial\,\mathcal{L}_{\text{N}}(\theta_0)}{\partial\theta}\frac{\partial f(X;\theta_0)}{\partial\theta'}\frac{1}{f(X;\theta_0)}f(X;\theta_0)dX = 0\\
&\Leftrightarrow \mathbb{E}\left[\frac{\partial^2\,\mathcal{L}_{\text{N}}(\theta_0)}{\partial\theta\partial\theta'}\right] + \mathbb{E}\left[\frac{\partial\,\mathcal{L}_{\text{N}}(\theta_0)}{\partial\theta}\frac{\partial\,\mathcal{L}_{\text{N}}(\theta_0)}{\partial\theta'}\right] = 0,
\end{aligned} \tag{13}$$

which, after rearranging the terms, delivers the result.

The ***Cramer-Rao inequality*** states that any unbiased estimator $\tilde{\theta}$ satisfies:

$$\text{Var}(\tilde{\theta}) \geq I(\theta_0)^{-1}. \tag{14}$$

Thus, this inequality indicates that the inverse of the information matrix is the lower bound for the variance for any unbiased estimator. Therefore, an unbiased estimator that has variance equal to the Cramer-Rao lower bound is the BUE. As we will be able to discuss after we introduce some concepts in Chapter 8, when the sample size tends to infinity, the variance of the MLE tends to the Cramer-Rao lower bound, and, hence, in that case, it is the BUE. Moreover, if a BUE exists, it is the MLE (we are not going to prove it).

We are going to prove the Cramer-Rao inequality for the simple case in which θ_0 is a scalar (and so is $I(\theta_0)$), but the results are directly generalizable to the case in which θ_0 is a vector (and thus $I(\theta_0)$ is a matrix). To do so, we are going to use the Schwartz inequality that we proved in Chapter 3. The Schwartz inequality states that $\text{Cov}(X,Y)^2/(\text{Var}(X)\,\text{Var}(Y)) \leq 1$. We are going to define two random

variables: $\tilde{\theta}$ and $z(X;\theta_0)$. To compute $\text{Cov}(\tilde{\theta}, z(X;\theta_0))$ we start from the fact that $\tilde{\theta}$ is an unbiased estimator, and hence satisfies $\mathbb{E}[\tilde{\theta}] = \theta_0$:

$$\begin{aligned}
&\mathbb{E}[\tilde{\theta}] = \theta_0 \\
\Leftrightarrow &\frac{\partial}{\partial \theta_0'}\mathbb{E}[\tilde{\theta}] = \frac{\partial}{\partial \theta_0'}\int_{-\infty}^{\infty}\tilde{\theta}f(X;\theta_0)dX = \int_{-\infty}^{\infty}\tilde{\theta}\frac{\partial f(X;\theta_0)}{\partial \theta_0'}dX = 1 \\
\Leftrightarrow &\int_{-\infty}^{\infty}\tilde{\theta}\frac{\partial \ln f(X;\theta_0)}{\partial \theta_0'}f(X;\theta_0)dX = \mathbb{E}[\tilde{\theta}z(X;\theta_0)] = 1. \qquad (15)
\end{aligned}$$

Because $\mathbb{E}[x(X;\theta_0)] = 0$, this implies that $\text{Cov}(\tilde{\theta}, z(X;\theta_0)) = 1$. Thus, the Schwartz inequality reads as:

$$\frac{1}{\text{Var}(\tilde{\theta})\,\text{Var}(z(X;\theta_0))} \leq 1 \quad \Leftrightarrow \quad \text{Var}(\tilde{\theta}) \geq \frac{1}{\text{Var}(z(X;\theta_0))} = \frac{1}{I(\theta_0))}, \qquad (16)$$

proving the result. The same logic with a bit more tedious algebra applies to the multivariate case.

Let us illustrate all this with the normal distribution. Consider a random variable $X \sim \mathcal{N}(\mu_0, \sigma_0^2)$, for which we have an i.i.d. sample $(X_1, ..., X_N)$. We are interested in estimating the parameters $\theta = (\mu, \sigma^2)'$. The likelihood function for the sample is:

$$L_N(\mu, \sigma^2) = \prod_{i=1}^{N} f(x_i) = (2\pi)^{-\frac{N}{2}}(\sigma^2)^{-\frac{N}{2}}\exp\left(-\frac{1}{2\sigma^2}\sum_{\text{i=1}}^{\text{N}}(x_i - \mu)^2\right), \qquad (17)$$

and the log-likelihood is:

$$\mathcal{L}_{\text{N}}(\mu, \sigma^2) = -\frac{N}{2}\ln(2\pi) - \frac{N}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\sum_{\text{i=1}}^{\text{N}}(x_i - \mu)^2. \qquad (18)$$

The score is:

$$z(X;\theta) = \frac{\partial\,\mathcal{L}_{\text{N}}(\mu, \sigma^2)}{\partial(\mu, \sigma^2)} = \begin{pmatrix} \frac{1}{\sigma^2}\sum_{\text{i=1}}^{\text{N}}(x_i - \mu) \\ -\frac{N}{2}\frac{1}{\sigma^2} + \frac{1}{2\sigma^4}\sum_{\text{i=1}}^{\text{N}}(x_i - \mu)^2 \end{pmatrix}. \qquad (19)$$

Evaluated at $\theta = (\mu_0, \sigma_0^2)$, it is easy to check that the expected score is equal to zero. The MLE picks $\theta = (\hat{\mu}, \hat{\sigma}^2)'$ such that $z(X;\hat{\theta}) = 0$ (i.e., the first order condition is satisfied), which, with simple algebra, delivers:

$$\hat{\mu} = \frac{1}{N}\sum_{\text{i=1}}^{\text{N}} X_i = \bar{X}_N \quad \text{and} \quad \hat{\sigma}^2 = \frac{1}{N}\sum_{\text{i=1}}^{\text{N}}(X_i - \bar{X}_N)^2 = \hat{\sigma}_N^2. \qquad (20)$$

To compute the Cramer-Rao lower bound, we can use the information matrix equality and compute $I(\theta_0) = -\,\mathbb{E}\left[\frac{\partial^2 z(X;\theta_0)}{\partial\theta\partial\theta'}\right]$:

$$\mathbb{E}\left[\begin{pmatrix} N\frac{1}{\sigma_0^2} & \frac{1}{\sigma_0^4}\sum_{\mathrm{i=1}}^{\mathrm{N}}(x_i-\mu_0) \\ \frac{1}{\sigma_0^4}\sum_{\mathrm{i=1}}^{\mathrm{N}}(x_i-\mu_0) & -\frac{N}{2}\frac{1}{\sigma_0^4}+\frac{1}{\sigma_0^6}\sum_{\mathrm{i=1}}^{\mathrm{N}}(x_i-\mu_0)^2 \end{pmatrix}\right] = \begin{pmatrix} \frac{N}{\sigma_0^2} & 0 \\ 0 & \frac{N}{2\sigma_0^4} \end{pmatrix}. \tag{21}$$

Therefore, the Cramer-Rao lower bound is:

$$I(\theta_0)^{-1} = \begin{pmatrix} \frac{\sigma_0^2}{N} & 0 \\ 0 & \frac{2\sigma_0^4}{N} \end{pmatrix}, \tag{22}$$

which allows us to extract three conclusions:

1) $\hat{\mu} = \bar{X}_N$ is the best unbiased estimator (BUE) of μ_0: it is an unbiased estimator (we saw in Chapter 4 that $\mathbb{E}[\bar{X}_N] = \mu_0$), and the variance of the sample mean (derived in Chapter 4) is equal to the Cramer-Rao lower bound.

2) $\hat{\sigma}^2 = \hat{\sigma}_N^2$ is a biased estimator of σ^2, as we noted in Chapter 4, and, hence, the Cramer-Rao result is not applicable. On the other hand, s_N^2, which is an unbiased estimator of σ_0^2, is not the BUE, because its variance is $\mathrm{Var}(s_N^2) = \frac{2\sigma^4}{N-1} + \frac{\mu_4 - 3\sigma^4}{N}$ (as noted in Chapter 4), larger than the Cramer-Rao bound. The latter does not surprise us, because we knew that, if there exist a BUE, it is the MLE.

3) If we knew μ_0, and we were to estimate only σ_0^2, the estimator that we would obtain would be $\hat{\sigma} = \tilde{\sigma}_N^2$, which is unbiased, and whose variance is $\mathrm{Var}(\tilde{\sigma}_N^2) = \frac{\mu_4 - \sigma_0^4}{N}$, where μ_4 is the fourth central moment. It turns out that, for the normal distribution, $\mu_4 = 3\sigma^4$, which would imply that the "ideal" estimator of the variance is indeed a BUE, because its variance is equal to the Cramer-Rao bound.

As a final note, it is left as an exercise to check that the equality of the information, which we have used in Equation (21), holds.

VI. Bayesian Inference

Recall the Bayes theorem in Chapter 3:

$$P(\mathcal{A}\,|\,\mathcal{B}) = \frac{P(\mathcal{B}\,|\,\mathcal{A})P(\mathcal{A})}{P(\mathcal{B})} = \frac{P(\mathcal{B}\,|\,\mathcal{A})P(\mathcal{A})}{P(\mathcal{B}\,|\,\mathcal{A})P(\mathcal{A}) + P(\mathcal{B}\,|\,\mathcal{A}^c)P(\mathcal{A}^c)}. \tag{23}$$

The second equality is new: we have used the fact that $P(\mathcal{B}) = P(\mathcal{B} \cap \mathcal{A}) + P(\mathcal{B} \cap \mathcal{A}^c)$. If instead of $\mathcal{A}$ and $\mathcal{A}^c$ we partition the sample space in N mutually exclusive sets that cover the entire sample space, $\mathcal{A}_1, ..., \mathcal{A}_N$, we can write:

$$P(\mathcal{A}_i \mid \mathcal{B}) = \frac{P(\mathcal{B} \mid \mathcal{A}_i)P(\mathcal{A}_i)}{P(\mathcal{B} \mid \mathcal{A}_1)P(\mathcal{A}\,1) + ... + P(\mathcal{B} \mid \mathcal{A}_N)P(\mathcal{A}_N)}. \tag{24}$$

We often have *a priori* beliefs about the probability that an event occurs. Imagine we want to estimate whether a coin is fair or not. It is natural to start from the belief that the probabilities assigned to heads and tails are 0.5 each. If we are only able to toss the coin once, it makes sense to give a lot of weight to our belief, and little to the sample, it looks a better strategy than assigning an estimate of $\hat{P}(\text{heads}) \equiv \hat{p} = 1$ for the outcome we obtained. If we can keep tossing the coin, we will update our beliefs with each toss. If after 10 times we have obtained 8 heads and only 2 tails, we will start to think that maybe the coin is not fair. If, after tossing it 1,000 times, we obtained 800 heads and 200 tails, then we will probably conclude that the coin is not fair, and our estimate will be $\hat{p} = \frac{8}{10}$. This is the intuition of the type of estimators that we are introducing in this section.

More formally, define ***subjective probability*** as the probability function that describes our beliefs about the true probabilities of the different outcomes. We can define subjective probabilities even when we do not have any *a priori* information or belief. We assume that we know the likelihood of the sample $f_N(X|\theta)$ up to the parameter θ, the ***a priori distribution*** $g(\theta)$ that describes our beliefs about the true parameter before observing the sample. The ***Bayesian inference*** is based on the ***a posteriori distribution*** of the parameter given the information in the sample, obtained from the application of the Bayes theorem:

$$h(\theta|X) = \frac{f(X|\theta)g(\theta)}{\int_{-\infty}^{\infty} f(X|c)g(c)dc} \propto f(X|\theta)g(\theta). \tag{25}$$

Importantly, note that we are treating θ (the "parameter") as a random variable now, not as a given (but unknown) value as we have been doing so far. The approaches used so far form the basis of the ***frequentist inference***, which is opposed to the Bayesian inference. In a frequentist approach, the unknown parameters of the model are not capable of being treated as random variables in any way. In contrast, a Bayesian approach to inference does allow probabilities to be associated with unknown parameters. Normally, these associated probabilities are probabilities in its Bayesian interpretation, which is a quantity that we assign for the purpose of representing a state of knowledge or belief.

Therefore, in Bayesian estimation, we are primarily interested in obtaining a

posterior distribution $h(\theta|X)$. However, we can also obtain point estimates using the posterior distribution $h(\theta|X)$. For example, the mean of the posterior distribution minimizes the expected quadratic loss:

$$\mathbb{E}_h[\theta|X] = \arg\min_c \int_{-\infty}^{\infty} (c-\theta)^2 h(\theta|X) d\theta. \tag{26}$$

Likewise, the median of the posterior distribution minimizes the expected absolute loss, and the mode maximizes the posterior density.

Let us illustrate all this by retaking the example of tossing a coin. Let X be a random variable that takes the value of 1 if the outcome is a head, and let p denote the probability that $X = 1$ ($1-p$ is the probability of $X = 0$, and 0 is the probability for any other value of X). The likelihood of a sample of size N for X is:

$$f(X|p) = p^r (1-p)^{N-r}, \tag{27}$$

where $r \equiv \sum_{\mathrm{i}=1}^{\mathrm{N}} X_i$. Now we need a prior. Consider the beta distribution with given values for parameters α and β (such that $\alpha, \beta > 0$) as our prior:

$$g(p) = \begin{cases} p^{\alpha-1}(1-p)^{\beta-1} \frac{1}{B(\alpha,\beta)} & \text{if } p \in [0,1] \\ 0 & \text{otherwise,} \end{cases} \tag{28}$$

where the last term $B(\alpha, \beta) \equiv \int_0^1 z^{\alpha-1}(1-z)^{\beta-1}$ is a constant that guarantees the distribution to integrate to 1. In this case, the posterior distribution of p given the data is:

$$h(p|X) \propto p^{r+\alpha-1}(1-p)^{N-r+\beta-1}. \tag{29}$$

A point estimate for p is the mean of the posterior distribution:

$$\tilde{p} = \mathbb{E}_h[p|X] = \frac{\alpha + r}{\alpha + r + \beta + N - r} = \frac{\alpha + r}{\alpha + \beta + N}. \tag{30}$$

Note that we can rewrite $\tilde{p}$ as follows:

$$\tilde{p} = \frac{N}{\alpha+\beta+N}\frac{r}{N} + \frac{\alpha+\beta}{\alpha+\beta+N}\frac{\alpha}{\alpha+\beta} \equiv w(N)\hat{p} + (1-w(N))\,\mathbb{E}_g[p], \tag{31}$$

with $w(N) \in (0,1)$, and $\partial w(N)/\partial N = (\alpha+\beta)/(\alpha+\beta+N)^2 > 0$. Therefore, our estimate is a convex combination of the sample mean and the mean of our prior, with weights that are such that the weight on the former increases (and that on the latter decreases as a result) when the sample size N increases.