Overview of Statistical Estimation Theory

Roni Rosenfeld

January 16, 1997

Based loosely on:

- "Theory of Point Estimation" by Lehmann.
- "Statistical Theory and Mehtodology" by Brownlee.
- "Nonparametric Probability Density Estimation", by Tapia & Thompson.

1 The Problem

A random sample: A set of values of statistically independent, identically distributed random variables.

Given: A random sample $\{x_1, \ldots, x_n\}$ (a.k.a. the *data*).

Assumption: the data come from some probability distribution P, which belongs to some known class $\mathcal{P}$ parameterized by $\Theta \stackrel{\text{def}}{=} \{\theta_1, \ldots, \theta_k\}$:

$$\mathcal{P} \stackrel{\text{def}}{=} \{P(x; \Theta)\} \stackrel{\text{def}}{=} \{P(x; \theta_1, \ldots, \theta_k\}$$

Examples:

1. A coin has probability θ_1 of falling on "heads":

$$\mathcal{P} = \{P(x;\theta_1)|0 \leq \theta_1 \leq 1\}$$

2. A die has a certain probability of landing on each of its 6 faces:

$$\mathcal{P} = \{P(x;\theta_1 \ldots,\theta_6)|\sum_{i=1}^{6}\theta_i = 1, \theta_i \geq 0\}$$

3. Cosmic ray strength has normal distribution:

$$\mathcal{P} = \{P(x;\mu,\sigma^2)|\mu,\sigma \in \mathcal{R}\}$$

We now ask: Given the data (results of coin flips, dice throwing, or cosmic ray recording), can we judiciously guess the value of (some of the) θ_i's ?

- find "the most suitable value" $\Longrightarrow$ *point estimation*.
- find a reasonable range $\Longrightarrow$ *confidence interval*.
- answer "is θ in this range?" $\Longrightarrow$ *hypothesis testing*.

We may want to estimate some subset of the θ_i's, or some function of them. The other θ_i's may or may not be known.

Assume for simplicity we want to estimate a single θ. This is the *estimand*.

2 Estimators and Estimates

We need to find a real-valued function $\hat{\theta}(X)$, which tends to be close to θ.

Note: θ is a fixed but unknown number. $\hat{\theta}(X)$ is a function of a random variable, and thus has an associated distribution.

Example: $\hat{\theta}(X)$ = # heads / sample size

The function $\hat{\theta}(X)$ is the *estimator*.

Given a sample $\mathbf{x}$, the value $\hat{\theta}(\mathbf{x})$ is an *estimate*.

What do we mean when we ask that the estimator "tend to be closed to θ"?

We could mean "close on average", as in:

$$Pr_{\Theta}(|\hat{\theta}(X) - \theta| > C) < \epsilon$$

or

$$E_{\Theta}[(\hat{\theta}(X) - \theta)^2] < \epsilon$$

More generally, define a *loss function* $L(\theta, y) \geq 0$ to reflect the negative impact of the estimation inaccuracy. $L(\theta, y) = 0$ iff $\theta = y$. Now we can rate any given estimator $\hat{\theta}(X)$ using a *risk function*:

$$R(\theta, \hat{\theta}) \stackrel{\text{def}}{=} \mathbf{E}_{\Theta}\{L(\theta, \hat{\theta})\}$$

Goal: Find an estimator $\hat{\theta}(X)$ which minimizes the risk R for all values of θ.

$\Longrightarrow$ Impossible! (why?)

Well, we should at least require the estimator to have some desired properties. Since an estimator is a (function of a) random variable, we can talk about its *expectation* $E[\hat{\theta}]$ and its *variance* $V[\hat{\theta}]$.

The *bias* of an estimator is the difference between its expectation and the estimand: $E[\hat{\theta}] - \theta$.

An *unbiased* estimator has zero bias: $E[\hat{\theta}] = \theta$.

Unbiased estimators do not always exist!

The *variance* of an estimator is simply $V[\hat{\theta}] = E[(\hat{\theta} - E[\hat{\theta}])^2]$.

We would like an estimator with minimal bias and minimal variance. Sometimes there will be a tradeoff between the two (see example).

One way to optimize bias and variance together is by choosing as the Loss function the square of the estimation error. Then the Risk function becomes the *mean square error*:

$$E[(\hat{\theta} - \theta)^2] = E[(\hat{\theta} - E[\hat{\theta}] + E[\hat{\theta}] - \theta)^2] = \text{variance} + \text{bias}^2$$

Among all unbiased estimators, the one with the minimum variance is called the *efficient* estimator. The *efficiency* of any other estimator is the ratio of its variance to that of the efficient estimator.

Another desired property is *consistency*: as the sample size grows, the error in the estimation should tend to zero in probability. More formally:

$$\forall \epsilon > 0, \Pr\{|\hat{\theta} - \theta| > \epsilon\} \longrightarrow 0 \quad \text{as} n \to \infty$$

We already mentioned that it is impossible to minimize the Risk function uniformly across all possible values of θ. Instead, we may choose to *minimize the maximal risk*:

$$\sup_{\theta} R(\theta, \hat{\theta})$$

Estimators designed with this criterion are called *minimax estimators*.

Alternatively, we may assume some prior probability distribution $g(\theta)$ over θ, and try to minimize the *weighted risk*:

$$\int R(\theta, \hat{\theta}) \cdot g(\theta) d\theta$$

$g(\theta)$ has an interesting interpretation: it captures our *prior* knowledge about the distribution we are trying to estimate. Thus θ itself is assumed to be a random variable. More on this later.

3 Maximum Likelihood Estimators

The idea: choose the θ which maximizes the likelihood of the model having generated the data

The *likelihood function*:

$$L(\mathbf{x}|\Theta) \stackrel{\text{def}}{=} \Pr(\mathbf{x}|\Theta) = \prod_{1}^{n} \Pr(x_i|\Theta)$$

The Maximum Likelihood Estimator (MLE):

$$MLE(\mathbf{x}) \stackrel{\text{def}}{=} \arg\max_{\Theta} L(\mathbf{x}|\Theta)$$

- The ML *estimator* is derived by holding $\mathbf{x}$ fixed and maximizing L over all possible values of θ (i.e. differentiating L wrt θ).
- The ML *estimate* is derived by plugging the value of $\mathbf{x}$ into the ML estimator.

Example: $\mathcal{P} = B(p, n)$ (the family of binomial distributions.) Suppose we observe k successes in n trials. Then the likelihood function is:

$$L(k|p) = \Pr(X = k; p, n) = \binom{n}{k} \theta^k (1 - \theta)^{n-k}$$

Maximizing L is the same as maximizing $\log L$ (why?), so:

$$\begin{aligned} \frac{\partial}{\partial \theta} \log L &= \frac{\partial}{\partial \theta} [\log \binom{n}{k} + k \log \theta + (n - k) \log(1 - \theta)] \\ &= \frac{k}{\theta} - \frac{n - k}{1 - \theta} = 0 \end{aligned} \tag{1}$$

with the solution $\hat{\theta} = k/n$.

If there are multiple unknown parameters, we solve a system of equations based on the partial derivatives wrt the various parameters.

MLEs are:

- consistent
- asymptotically efficient
- asymptotically normal
- invariant (if $\hat{\theta}$ is the MLE of θ, then $g(\hat{\theta})$ is the MLE of $g(\theta)$.)
- often biased! (but usually canbe corrected)

MLEs are used very commonly, and usually (but not always) wisely so.

In comparing ML and MVU (minimum-variance unbiased) estimators, usually neither one is uniformly better, and usually there is little practical difference between them for large samples.

4 Bayesian Analysis

Until now we assumed nothing about the possible values of θ, except that they belong to a given set. We tried to come up with estimators that will "do justice" no matter what the true value of θ is. We discovered this is difficult.

Suppose now that we have some prior knowledge about the likely values of θ, encoded in a pdf $g_0(\theta)$ called the *prior*. We could then answer questions about θ *without even looking at the data!* For example, if we are asked to estimate the value of θ, we may choose to provide:

- $\arg\max_\theta g_0(\theta)$ (the mode of g_0), or
- the median of g_0, or
- $\arg\min_\eta \int R(\theta, \eta) g_0(\theta) d\theta$ (that θ which minimizes the Risk), or
- $E[g_0(\theta)]$ (the expectation of g_0)

Once we see some data, we want to update our belief about the likely values of θ, by computing the *posterior* distribution g_1 of θ given the data and the prior (i.e. $g_1(\theta|\mathbf{x}, g_0)$). We do this using Bayes Formula:

$$g_1(\theta|\mathbf{x}) \;=\; \frac{g_0(\theta)\cdot L(\mathbf{x}|\theta)}{\Pr(\mathbf{x})} \;=\; \frac{g_0(\theta)\cdot L(\mathbf{x}|\theta)}{\int_\eta g_0(\eta)\cdot L(\mathbf{x}|\eta)\, d\eta}$$

Or, conceptually:

$$\begin{aligned} \text{Posterior(model}|\text{data)} &= \frac{\text{Prior(model)} \cdot \text{Likelihood(data}|\text{model)}}{\text{Pr(data)}} \\ &= \frac{\text{Prior(model)} \cdot \text{Likelihood(data}|\text{model)}}{\int_{\text{all models}} \text{Prior(model)} \cdot \text{Likelihood(data}|\text{model)}} \end{aligned} \tag{2}$$

After consulting the data and generating the posterior, we are in a similar situation to the one we had with the prior: all our knowledge (or belief) about the likely values of θ is now encoded in the posterior $g_1(\theta)$. So when asked to estimate θ, we can again respond in much the same way, only using $g_1(\theta)$ instead of $g_0(\theta)$:

- $\arg\max_\theta g_1(\theta)$ (the mode of g_1) $\Longrightarrow$ "MAP estimation"
- the median of g_1
- $\arg\min_\eta \int R(\theta, \eta) g(\theta) d\theta$ (that θ which minimizes the Risk) $\Longrightarrow$ "Bayesian Estimator".
- $E[g_1(\theta)]$ (the expectation of g_1)
  (special case when the risk is the squared error)

The mode (maximum) of a stochastic function is not nearly as stable as its mean or median. For small samples or questionable priors, MAP could lead to bad estimates.

Suppose now that we are given more data. We can treat $g_1(\theta)$ as the prior, and update to $g_2(\theta)$, etc.

Hopefully, if we do things right, updating in one batch should give the same result as updating in chunks.

Consider the sequence of posteriors $\{g_0, g_1, g_2, \ldots\}$.

- When there's no data: "posterior" = prior.
- When there's infinite data: the posterior sequence converges to the real distribution, independent of the prior.

MLE can now be seen as a special case of Bayesian MAP estimation, where the prior is the uniform distribution.