Topic 8: Estimation

Rohini Somanathan

Course 003, 2018

Random Samples

We cannot usually look at the population as a whole because it would take too long, be too expensive or impractical for other reasons (we crash cars to see know how sturdy they are)

We would like to choose a sample which is representative of the population or process that interests us. A random sample is one in which all objects in the population have an equal chance of being selected.

Definition (random sample): *Let $f(x)$ be the density function of a continuous random variable X. Consider a sample of size n from this distribution. We can think of the first value drawn as a realization of the random variable X_1, similarly for $X_2 \dots X_n$. $(x_1, \dots, x_n)$ is a random sample if $f(x_1, \dots, x_n) = f(x_1)f(x_2)\dots f(x_n)$.*

This is often hard to implement in practice unless we think through the possible pitfalls.

Example: We have a bag of sweets and chocolates of different types (eclairs, five-stars, gems...) and want to estimate the average weight of a items in the bag. If we pass the bag around, each student puts their hand in and picks 5 items and replaces these, how do you think these sample averages would compare with the true average?

Now think about caveats when collecting a household sample to estimate consumption.

Statistical Models

Definition (statistical model): *A statistical model for a random sample consists of a parametric functional form, $f(x;\Theta)$ together with a parameter space Ω which defines the potential candidates for Θ.*

Examples: **We may specify that our sample comes from**

- **a Bernoulli distribution and $\Omega = \{p : p \in [0,1]\}$**
- **a Normal distribution where $\Omega = \{(\mu,\sigma^2) : \mu \in (-\infty,\infty), \sigma > 0\}$**

Note that Ω could be much more restrictive. For example, we could have $p \in (\frac{1}{2},1)$ in the first case and $\mu \in (0,100)$ in the second case.

Estimators and Estimates

Definition (estimator): *An estimator of the parameter θ, based on the random variables $X_1, \ldots, X_n$, is a real-valued function $\delta(X_1, \ldots, X_n)$ which specifies the estimated value of θ for each possible set of values of $X_1, \ldots, X_n$.*

Since an estimator $\delta(X_1, \ldots, X_n)$ is a function of random variables, $X_1, \ldots, X_n$, the estimator is itself a random variable and its probability distribution can be derived from the joint distribution of $X_1, \ldots, X_n$.

A point estimate is a specific value of the estimator $\delta(x_1, \ldots, x_n)$ that is determined by using the observed values $x_1, \ldots, x_n$.

There are lots of potential functions of the random sample, δ, what criteria should we use to choose among these?

Desirable Properties of Estimators

1. **Unbiasedness :** $E(\hat{\theta}_n) = \theta \ \forall \ \theta \in \Omega$.
2. **Consistency:** $\lim_{n\to\infty} P(|\hat{\theta}_n - \theta| > \epsilon) = 0$ **for every** $\epsilon > 0$.
3. **Minimum MSE:** $E(\hat{\theta}_n - \theta)^2 \leq E(\tilde{\theta}_n - \theta)^2$ **for any** $\tilde{\theta}_n$.

Using the MSE criterion could lead us to choose biased estimators because

$$MSE(\hat{\theta}) = Var(\hat{\theta}) + Bias(\hat{\theta}, \theta)^2$$

$$MSE(\hat{\theta}) = E(\hat{\theta}-\theta)^2 = E[\hat{\theta}-E(\hat{\theta})+E(\hat{\theta})-\theta]^2 = E[(\hat{\theta}-E(\hat{\theta}))^2+(E(\hat{\theta})-\theta)^2+2(\hat{\theta}-E(\hat{\theta}))(E(\hat{\theta})-\theta)] = Var(\hat{\theta})+Bias(\hat{\theta},\theta)^2+0$$

A Minimum Variance Unbiased Estimator (MVUE) is an estimator which has the smallest variance among the class of unbiased estimators.

A Best Linear Unbiased Estimator (BLUE) is an estimator which has the smallest variance among the class of linear unbiased estimators (the estimates must be linear functions of sample values).

Maximum Likelihood Estimators

Definition (M.L.E.): *Suppose that the random variables* $X_1, \ldots, X_n$ *form a random sample from a discrete or continuous distribution for which the p.f. or p.d.f is* $f(x|\theta)$, *where* θ *belongs to some parameter space* Ω. *For any observed vector* $\mathbf{x} = (x_1, \ldots, x_n)$, *let the value of the joint p.f. or p.d.f. be denoted by* $f_n(\mathbf{x}|\theta)$. *When* $f_n(\mathbf{x}|\theta)$ *is regarded a function of* θ *for a given value of* $\mathbf{x}$, *it is called the likelihood function.*

For each possible observed vector $\mathbf{x}$, *let* $\delta(\mathbf{x}) \in \Omega$ *denote a value of* $\theta \in \Omega$ *for which the likelihood function* $f_n(\mathbf{x}|\theta)$ *is a maximum, and let* $\hat{\theta} = \delta(\mathbf{X})$ *be the estimator of* θ *defined in this way. The estimator* $\hat{\theta}$ *is called the maximum likelihood estimator of* θ *(M.L.E.).*

For a given sample, $\delta(\mathbf{x})$ *is the maximum likelihood estimate of* θ *(also called M.L.E.)*

M.L.E..of a Bernoulli parameter

- The Bernoulli density can be written as $f(x;\theta) = \theta^x(1-\theta)^{1-x}, \quad x \in \{0,1\}$.
- For any observed values $x_1, \dots, x_n$, where each x_i is either 0 or 1, the likelihood function is given by: $f_n(x|\theta) = \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i} = \theta^{\sum x_i}(1-\theta)^{n-\sum x_i}$
- The value of θ that will maximize this will be the same as that which maximizes the log of the likelihood function, $L(\theta)$ which is given by:

$$L(\theta) = \Big(\sum_{i=1}^{n} x_i\Big)\ln\theta + \Big(n - \sum_{i=1}^{n} x_i\Big)\ln(1-\theta)$$

The first order condition for an extreme point is given by: $\frac{\left(\sum_{i=1}^{n} x_i\right)}{\hat{\theta}} = \frac{n-\left(\sum_{i=1}^{n} x_i\right)}{1-\hat{\theta}}$ and solving this, we get $\hat{\theta}_{MLE} = \frac{\sum_{i=1}^{n} x_i}{n}$.

Confirm that the second derivate of $L(\theta)$ is in fact negative, so we do have a maximum.

Sampling from a normal distribution

$$f_n(x|\mu,\sigma^2) = \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} e^{-\frac{1}{2\sigma^2}(\sum\limits_{i=1}^{n}(x_i-\mu)^2)}$$

$$L(\mu,\sigma^2) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2$$

If the variance is known, we have to maximize this function w.r.t. μ, and our first-order condition is: $\frac{1}{\sigma^2}(\sum\limits_{i=1}^{n} x_i - n\mu) = 0$, so $\hat{\mu} = \bar{x}_n$.

If both the mean and variance are unknown, the likelihood function has to be maximized w.r.t. both μ and σ^2 and we have two first-order conditions:

$$\frac{\partial L}{\partial \mu} = \frac{1}{\sigma^2}(\sum_{i=1}^{n} x_i - n\mu) \quad (1)$$

$$\frac{\partial L}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(x_i-\mu)^2 \quad (2)$$

We obtain $\hat{\mu} = \bar{x}_n$ from setting $\frac{\partial L}{\partial \mu} = 0$ and substitute this into the second condition to obtain $\hat{\sigma}^2 = \frac{1}{n}\sum\limits_{i=1}^{n}(x_i - \bar{x}_n)^2$

The maximum likelihood estimators are therefore $\hat{\mu} = \bar{X}_n$ and $\hat{\sigma^2} = \frac{1}{n}\sum\limits_{i=1}^{n}(X_i - \bar{X}_n)^2$. M.L.E's in general are not unbiased as seen here.

Sampling from a uniform distribution

Maximum likelihood estimators need not exist and when they do, they may not be unique as the following examples illustrate:

If $X_1, \dots, X_n$ is a random sample from a uniform distribution on $[0, \theta]$, the likelihood function is

$$f_n(x; \theta) = \frac{1}{\theta^n}$$

This is decreasing in θ and is therefore maximized at the smallest admissible value of θ which is given by $\hat{\theta} = max(X_1 \dots X_n)$.

If instead, the support is $(0, \theta)$ instead of $[0, \theta]$, then no M.L.E. exists since the maximum sample value is no longer an admissible candidate for θ.

If the random sample is from a uniform distribution on $[\theta, \theta + 1]$. Now θ could lie anywhere in the interval $[\max(x_1, \dots, x_n) - 1, \min(x_1, \dots, x_n)]$ and the method of maximum likelihood does not provide us with a unique estimate.

Likelihood functions are often complicated and we use numerical optimization methods to compute the M.L.E. (Gamma, Cauchy distributions)

Properties of Maximum Likelihood Estimators

Invariance: If $\hat{\theta}$ is the maximum likelihood estimator of θ, and $g(\theta)$ is a one-to-one function of θ, then $g(\hat{\theta})$ is a maximum likelihood estimator of $g(\theta)$

Example: The sample mean and sample variance are the M.L.E.s of the mean and variance of a random sample from a normal distribution so

- the M.L.E. of the standard deviation is the square root of the sample variance
- the M.L.E of $E(X^2)$ is equal to the sample variance plus the square of the sample mean, i.e. since $E(X^2) = \sigma^2 + \mu^2$, the M.L.E of $E(X^2) = \hat{\sigma}^2 + \hat{\mu}^2$

Consistency: If there exists a unique M.L.E. $\hat{\theta}_n$ of a parameter θ for a sample of size n, then plim $\hat{\theta}_n = \theta$.

Note: MLEs are not, in general, unbiased.

Example: The MLE of the variance of a normally distributed variable is given by $\hat{\sigma}^2_n = \frac{\sum\limits_{i=1}^{n}(X_i - \bar{X}_n)^2}{n}$. Let's rewrite this and take its expectation:

$$\hat{\sigma}^2_n = \frac{1}{n}\sum_{i=1}^{n}(X_i^2 - 2X_i\bar{X} + \bar{X}^2) = \frac{1}{n}(\sum X_i^2 - 2\bar{X}\sum X_i + \sum \bar{X}^2) = \frac{1}{n}(\sum X_i^2 - 2n\bar{X}^2 + n\bar{X}^2)] = \frac{1}{n}(\sum X_i^2 - n\bar{X}^2)$$

$$E[\hat{\sigma}^2_n] = E[\frac{1}{n}(\sum X_i^2 - n\bar{X}^2)] = \frac{1}{n}[\sum E(X_i^2) - nE(\bar{X}^2)] = \frac{1}{n}[n(\sigma^2 + \mu^2) - n(\frac{\sigma^2}{n} + \mu^2)] = \sigma^2\frac{n-1}{n}$$

Notice that $\frac{n}{n-1}E[\hat{\sigma}^2_n] = \sigma^2$ so an unbiased estimate is $\frac{\sum(X_i - \bar{X}_n)^2}{n-1}$

Sufficient Statistics

- We have seen that M.L.E's may not exist, or may not be unique. Where should our search for other estimators start? A natural starting point is the set of sufficient statistics for the sample.
- Suppose that in a specific estimation problem, two statisticians A and B would like to estimate θ; A observes the realized values of $X_1, \dots X_n$, while B only knows the value of a certain statistic $T = r(X_1, \dots, X_n)$.
- A can now choose any function of the observations $(X_1, \dots, X_n)$ whereas B can choose only functions of T. If B does just as well as A because the single function T has all the relevant information in the sample for choosing a suitable θ, then T is a sufficient statistic.

In this case, given $T = t$, we can *generate* an alternative sample $X_1' \dots X_n'$ in accordance with this conditional joint distribution (auxiliary randomization). Suppose A uses $\delta(X_1 \dots X_n)$ as an estimator. Well B could just use $\delta(X_1' \dots X_n')$, which has the same probability distribution as $A's$ estimator.

Think about what such an auxiliary randomization would be for a Bernoulli sample.

Neyman factorization and the Rao-Blackwell Theorem

Result (The Factorization Criterion (Fisher (1922) ; Neyman (1935)): *Let* $(X_1, \dots, X_n)$ *form a random sample from either a continuous or discrete distribution for which the p.d.f. or the p.f. is* $f(x|\theta)$*, where the value of* θ *is unknown and belongs to a given parameter space* Ω*. A statistic* $T = r(X_1, \dots, X_n)$ *is a sufficient statistic for* θ *if and only if, for all values of* $x = (x_1, \dots, x_n) \in \mathbb{R}^n$ *and all values of* $\theta \in \Omega$*,* $f_n(x|\theta)$ *of* $(X_1, \dots, X_n)$ *can be factored as follows:*

$$f_n(x|\theta) = u(x)v[r(x), \theta]$$

The functions u *and* v *are nonnegative; the function* u *may depend on* x *but does not depend on* θ*; and the function* v *will depend on* θ *but depends on the observed value* x *only through the value of the statistic* $r(x)$.

Result: Rao-Blackwell Theorem: *An estimator that is not a function of a sufficient statistic is dominated by one that is (in terms of having a lower MSE)*

Sufficient Statistics: examples

Let $(X_1,\dots,X_n)$ form a random sample from the distributions given below:

Poisson Distribution with unknown mean θ:

$$f_n(x|\theta) = \prod_{i=1}^{n} \frac{e^{-\theta}\theta^{x_i}}{x_i!} = \Big(\prod_{i=1}^{n}\frac{1}{x_i!}\Big)e^{-n\theta}\theta^{y}$$

where $y = \sum_{i=1}^{n} x_i$. We observe that $T = \sum_{i=1}^{n} X_i$ is a sufficient statistic for θ.

Normal distribution with known variance and unknown mean: **The joint p.f. $f_n(x|\theta)$ of $X_1,\dots X_n$ has already been derived as:**

$$f_n(x|\mu) = \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}}\exp\Big(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}x_i^2\Big)\exp\Big(\frac{\mu}{\sigma^2}\sum_{i=1}^{n}x_i - \frac{n\mu^2}{2\sigma^2}\Big)$$

The last term is $v(r(x),\theta)$, so once again, $T = \sum_{i=1}^{n} X_i$ is a sufficient statistic for μ.

Jointly Sufficient Statistics

If our parameter space is multi-dimensional, and often even when it is not, there may not exist a single sufficient statistic T, but we may be able to find a set of statistics, $T_1 \dots T_k$ which are jointly sufficient statistics for estimating our parameter.

The corresponding factorization criterion is now

$$f_n(x|\theta) = u(x)v[r_1(x), \dots r_k(x), \theta]$$

Example: If both the mean and the variance of a normal distribution is unknown, the joint p.d.f.

$$f_n(x|\mu) = \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} \exp\Big(-\frac{1}{2\sigma^2}\sum_{i=1}^{n} x_i^2\Big) \exp\Big(\frac{\mu}{\sigma^2}\sum_{i=1}^{n} x_i - \frac{n\mu^2}{2\sigma^2}\Big)$$

can be seen to depend on x only through the statistics $T_1 = \sum X_i$ and $T_2 = \sum X_i^2$. These are therefore jointly sufficient statistics for μ and σ^2.

If $T_1 \dots, T_k$ are jointly sufficient for some parameter vector θ and the statistics $T_1', \dots T_k'$ are obtained from these by a one-to-one transformation, then $T_1', \dots T_k'$ are also jointly sufficient. So the sample mean and sample variance are also jointly sufficient in the above example, since $T_1' = \frac{1}{n}T_1$ and $T_2' = \frac{1}{n}T_2 - \frac{1}{n^2}T_1^2$

Minimal Sufficient Statistics and Order Statistics

Definition (minimal sufficient statistic): *A statistic T is a minimal sufficient statistic if T is a sufficient statistic and every function of T which is a sufficient statistic is a one-to-one function of T.*

Minimally sufficient statistics cannot be reduced further without destroying the property of sufficiency. Minimal jointly sufficient statistics are defined in an analogous manner.

Definition (order statistics): *Let Y_1 denote the smallest value in the sample, Y_2 the next smallest, and so on, with Y_n the largest value in the sample. We call $Y_1, \ldots Y_n$ the order statistics of a sample.*

Order statistics are always jointly sufficient. To see this, note that the likelihood function is given by

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

Since the order of the terms in this product are irrelevant, we could as well write this expression as

$$f_n(x|\theta) = \prod_{i=1}^{n} f(y_i|\theta).$$

For some distributions (such as the Cauchy) these (or a one-to-one function of them) are the only jointly sufficient statistics and are therefore minimally jointly sufficient.

If a sufficient statistic $r(x)$ exists, the MLE must be a function of this statistic (this follows from the factorization criterion). It turns out that if MLE is a sufficient statistic, it is minimally sufficient.

Remarks

- Suppose we are picking a sample from a normal distribution, we may be tempted to use $Y_{(n+1)/2}$ as an estimate of the median m and $Y_n - Y_1$ as an estimate of the variance. Yet we know that we would do better using the sample mean for m and the sample variance must be a function of $\sum X_i$ and $\sum X_i^2$.
- A statistic is always sufficient with respect of a particular probability distribution, $f(x|\theta)$ and may not be sufficient w.r.t. , say, $g(x|\theta)$. Instead of choosing functions of the sufficient statistic we obtain in one case, we may want to find a robust estimator that does well for many possible distributions.
- In non-parametric inference, we do not know the likelihood function, and so our estimators are based on functions of the order statistics.