Statistical Estimation

We often observe data that can be thought of as independent draws from a distribution, which is unknown.

We want to be able to say something about the distribution.

Examples

1. Pollsters ask a sample of individuals how they plan to vote in an upcoming election. They want to know how the population will vote.

2. I need to decide whether to buy LED or compact fluorescent lightbulbs. LEDs are more expensive, but last longer. I have data on the lifetimes, in thousands of hours, of a sample of 100 lightbulbs of each type.

3. A manufacturer of washing machines wants to set its warranty period such that ~~more than~~ at least 90% of its products won't fail during the warranty period. It has data on the time to first failure of 1,000 of its machines.

In examples 2 & 3 above, and in a sense also in 1, we were interested only in estimating a summary statistic of the unknown distribution (the mean lifetime in example 2, and the 90th percentile of the time to the first failure in example 3), rather than the distribution itself.

Even though most applications only require such summary statistics, we usually pose the problem of estimating the whole distribution. However, we then "cheat" and restrict ourselves to distributions that are parametrised by a small number of summary statistics. There are usually good arguments or evidence to justify the restriction to those distributions.

So, the abstract problem formulation is as follows:

Problem: Let $X_1, X_2, X_3, \ldots, X_n$ be iid random samples from a distribution F.

Construct an estimator $\hat{F}$ of F from the data $(X_1, \ldots, X_n)$.

Q: What is a good estimator?

Example : Lifetimes of objects (lightbulbs, refrigerators, washing machines) are commonly modelled using Exponential distributions. It is a good model for things that fail due to one sudden, random event rather than an accumulation of wear and tear.

Q. Assume that lightbulbs from a given manufacturer have an $\text{Exp}(\lambda)$ lifetime, where λ is unknown. Given data $x_1, x_2, \ldots, x_n$ on the lifetimes of n lightbulbs, estimate λ.

The question has deliberately been posed rather loosely. In practice, we don't always know precise criteria by which to judge the quality of an estimator. As such, the question may not have a single correct answer. Instead, we are going to study a couple of natural choices of estimator, and study their properties.

Method of Moments Estimators

We continue with the lightbulbs example.

Assume $X_1, X_2, X_3, \ldots$ are iid random variables with an $\text{Exp}(\lambda)$ distribution, and that we have observed $X_1, \ldots, X_n$. How can we estimate λ from these observations?

Clearly, any estimator $\hat{\lambda}$ has to be a function of the observed data, i.e.,

$$\hat{\lambda} = g(X_1, X_2, \ldots, X_n)$$

How should we choose the function g?

One natural approach is to pick some summary statistic of the data, and compare it to the same statistic from the $\text{Exp}(\lambda)$ distribution. Then pick the value of λ that matches the statistic.

Let us illustrate with two choices of summary statistic.

1. Mean, $\mathbb{E}X_1$: Recall that an $\text{Exp}(\lambda)$ distribution has mean equal to $1/\lambda$. Matching this to the sample mean, we get

$$\hat{\lambda}_1 = \frac{1}{\text{sample mean}} = \frac{n}{X_1 + X_2 + \ldots + X_n}$$

$$= g_1(X_1, X_2, \ldots, X_n)$$

2. Median :

Recall that if X_1 has an $\text{Exp}(\lambda)$ distribution, then $\mathbb{P}(X_1 > x) = e^{-\lambda x} \quad \forall\ x \geq 0$.

The median is that value of x for which this probability is $1/2$, i.e.,

$$\text{median} = \{ x : e^{-\lambda x} = \tfrac{1}{2} \}$$

$$= -\frac{1}{\lambda} \ln\left(\tfrac{1}{2}\right) \approx \frac{0.7}{\lambda}$$

This suggests the estimator

$$\hat{\lambda}_2 = \frac{0.7}{\text{sample median}} = g_2(X_1, X_2, \ldots, X_n)$$

There is a slight problem with this estimator : while the sample mean is always well-defined, the sample median is unique only if the number of data points is odd. If it is even, then any number between the two mid-ranking data points is a valid ~~per~~ median.

This is not an insuperable problem. We just need to agree on a definition of the median when there is an even number of data points. For instance, we could agree that it should be the average of the two mid-ranking data points.

The two examples above illustrate how we can estimate the parameters of a distribution by matching the moments of the distribution (which can be expressed as a function of the unknown parameters) with corresponding moments of the sample.

Definition : Let X be a random variable with distribution or cdf F and density f. The k^{th} moment of X is defined as

$$\mathbb{E}(X^k) = \int_{-\infty}^{\infty} x^k f(x)\, dx.$$

In particular, the first moment is the mean of X.

Example : Suppose X is an $Exp(\lambda)$ random variable with density

$$f_X(t) = \begin{cases} \lambda e^{-\lambda t}, & t \geq 0 \\ 0, & t < 0 \end{cases}$$

$$\mathbb{E}[X^k] = \int_0^{\infty} x^k \lambda e^{-\lambda x}\, dx$$

$$= \lambda^{-k} \int_0^{\infty} (\lambda x)^k e^{-\lambda x}\, d(\lambda x)$$

$$= \lambda^{-k} \int_0^{\infty} z^k e^{-z}\, dz$$

$$= \lambda^{-k} \cdot k! \quad \text{(Exercise !)}$$

In particular, if $X \sim \text{Exp}(\lambda)$, then $\mathbb{E}X = \frac{1}{\lambda}$, and so a method of moments estimator of λ is

$$\hat{\lambda} = \frac{1}{\text{sample mean}} = \frac{n}{x_1 + x_2 + \dots + x_n}$$

where $x_1, x_2, \dots, x_n$ are the observations.

While this is the standard MoM estimator of λ, in principle, any moment could have been used. Thus,

$$\hat{\lambda}_k = \sqrt[k]{\frac{n\,k!}{x_1^k + x_2^k + \dots + x_n^k}}$$

is also a MoM estimator of λ.

Typically, as k gets larger, the estimator becomes more sensitive to outliers, so smaller values of k are preferred.

Later, we will look at some measures of the quality of an estimator.

Example : Suppose $X_1, X_2, \ldots, X_n$ are iid with an $N(\mu, \sigma^2)$ distribution, i.e., a normal or Gaussian distribution with mean μ and variance σ^2. Recall that this is the distribution with density

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad x \in \mathbb{R}$$

We now have two parameters, μ & σ, to estimate, so we will need at least two different moments of the distribution. We know that if $X \sim N(\mu, \sigma^2)$, then

$\mathbb{E}[X] = \mu$ and $\mathrm{Var}(X) = \sigma^2$, so

$\mathbb{E}[X^2] = \mu^2 + \sigma^2$.

By matching them with the sample moments, we get two equations :

$$\hat{\mu} = \frac{1}{n}(x_1 + x_2 + \ldots + x_n),$$

$$\hat{\mu}^2 + \hat{\sigma}^2 = \frac{1}{n}(x_1^2 + x_2^2 + \ldots + x_n^2),$$

which we can easily solve for $\hat{\sigma}^2$. We get

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

Exercises

1. The Uniform distribution on an interval $[0, \theta]$ is the distribution with density

$$f(x) = \begin{cases} 1/\theta, & 0 \leq x \leq \theta \\ 0, & \text{otherwise} \end{cases}$$

It has mean $\theta/2$ and variance $\theta^2/12$

Find a MoM estimator of θ given observations $x_1, x_2, \ldots, x_n$.

2. Repeat the above for a Uniform distribution on $[-\theta, \theta]$, which has density

$$f(x) = \begin{cases} 1/2\theta, & 0 \leq x \leq \theta \\ 0, & \text{otherwise} \end{cases}$$

3. Suppose you are told that the data $x_1, x_2, \ldots, x_n$ come from a normal distribution with zero mean, but unknown variance σ^2. Find a MoM estimator for σ^2.

Maximum Likelihood Estimators (MLEs)

Another natural way to estimate an unknown parameter is to find the parameter value that makes the data "most likely", or in other words that maximises the likelihood of the data.

It is easier to interpret "likelihood" with discrete data, so let us start with such an example.

Example: Consider a coin with unknown probability p of coming up heads. Define the random variable X_i to take the value 1 if the i^{th} coin toss comes up heads, and 0 if it comes up tails. Then

$\mathbb{P}(X_i = 1) = p$, $\mathbb{P}(X_i = 0) = 1 - p$.

X_i is called a Bernoulli(p) random variable.

Given the outcomes $x_1, x_2, \ldots, x_n$ of n independent coin tosses, how should we estimate p?

(Exercise: Find an MoM estimate of p.)

Example (continued)

We can work out the probability of observing the particular sequence $x_1, x_2, \ldots, x_n$:

$$\mathbb{P}(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n)$$

$$= \prod_{k=1}^{n} p^{x_k} (1-p)^{1-x_k}$$

$$= p^{(\text{no. of 1s})} (1-p)^{(\text{no. of 0s})}$$

The probability, or likelihood, of observing a particular sequence is a function of the unknown parameter, p. We denote the likelihood by $L(\text{data}; p)$. Thus, in this example, we have

$$L(x_1, x_2, \ldots, x_n; p) = p^{(\text{no. of 1s})} (1-p)^{(\text{no. of 0s})}$$

We want to find the value of p that makes the data most likely, i.e., we want to solve the optimisation problem

$$\max L(\text{data}; p) \quad \text{over } p \in [0, 1].$$

Example (continued)

It is easy to maximise the function $L(p)$ (for convenience, let's write $L(p)$ when our main interest is in how L depends on p rather than on the data):

- differentiate $L(p)$ with respect to p, set the derivative to 0, solve for p, and check if the solution is a maximum (sometimes we also have to check the end-points of the range).

It's even easier if we take logarithms first, & maximise $\ell(p) = \ln L(p)$ instead:

$$\max. \quad \ell(p) = (\text{no. of 1s}) \cdot \ln p + (\text{no. of 0s}) \ln(1-p)$$
$$\text{over} \quad p \in [0, 1].$$

We get $\ell'(p) = \dfrac{\text{no. of 1s}}{p} - \dfrac{\text{no. of 0s}}{1-p}$

$$\therefore \ \ell'(p) = 0 \iff \frac{p}{1-p} = \frac{\text{no. of 1s}}{\text{no. of 0s}},$$

$$\text{i.e., } \hat{p} = \frac{1}{n}(\text{no. of 1s})$$

$$= \text{observed fraction of 1s.}$$

You might have guessed the answer even if you knew no statistics!

Exercise : 1) Check that the above value of p does indeed maximise $l(p)$ over $[0,1]$.

2) Check that $\hat{p}$ coincides with the MoM estimator.

Remarks

Why take logarithms ?

Firstly, note that it is OK to do so. Maximising $L(p)$ is equivalent to maximising any monotone increasing function of $L(p)$, & $\ln(x)$ is an increasing function of x.

But why $\ln(x)$ rather than $\sqrt{x}$ or e^x ?

The answer is that, when we have independent observations, as is often the case, then the likelihood of a data set is the product of the likelihoods of the individual observations.

Taking logarithms turns this product into a sum, which is much more convenient for differentiation.

Example : Suppose $X_1, X_2, \ldots, X_n$ are iid with a Poisson(λ) distribution, i.e.,

$$\mathbb{P}(X_1 = k) = \frac{\lambda^k}{k!} e^{-\lambda}$$

Hence, we have the likelihood function

$$L(x_1, \ldots, x_n ; \lambda) = \prod_{i=1}^{n} \left(\frac{\lambda^{x_i}}{x_i!} e^{-\lambda} \right)$$

Taking logarithms, we get

$$\ell(\lambda) = \sum_{i=1}^{n} \left(x_i \log \lambda - \lambda - \log x_i! \right)$$

$$= \sum_{i=1}^{n} \log(x_i!) + (\log \lambda) \sum_{i=1}^{n} x_i - n\lambda$$

(All logs are natural, unless otherwise specified.)

The term $\sum_{i=1}^{n} \log(x_i!)$ doesn't depend on λ, and can be ignored. We get

$$\ell'(\lambda) = \frac{1}{\lambda} \sum_{i=1}^{n} x_i - n$$

$$= 0 \iff \lambda = \frac{1}{n} \sum_{i=1}^{n} x_i$$

Thus, the MLE estimator is

$$\hat{\lambda} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

Check that this coincides with the MoM estimator - Exercise!

Does the MLE approach extend to continuous random variables?

It does, but we can no longer equate likelihood & probability because, with continuous random variables, the probability of any set of observations is zero!

How then should we interpret likelihood?

The density function seems a natural & justifiable choice.

Example: Suppose $X_1, \ldots, X_n$ are iid $\text{Exp}(\lambda)$ random variables. Estimate λ from the data $x_1, x_2, \ldots, x_n$.

Writing the likelihood in terms of the density function, we get

$$L(x_1, \ldots, x_n; \lambda) = \prod_{i=1}^{n} \lambda e^{-\lambda x_i}, \text{ and}$$

$$\ell(x_1, \ldots, x_n; \lambda) = n\lambda - \lambda \sum_{i=1}^{n} x_i$$

Optimising ℓ over $\lambda \geq 0$, we get

$$\hat{\lambda} = \frac{1}{n} \sum_{i=1}^{n} x_i,$$

which again coincides with the MoM estimator!

Remark : It is somewhat misleading that the MLE & MoM estimators have coincided in all our examples so far. It is by no means necessary that they should do so!

That they have done so in all our examples is an artifact, a consequence of the fact that many popular parametric families of distributions are so-called ~~"linear"~~ "exponential families."

Example : Let $X_1, X_2, \dots, X_n$ be iid $N(\mu, \sigma^2)$ random variables. We want to estimate μ and σ^2 from the data. We have the log-likelihood function

$$\ell(x_1, \dots, x_n; \mu, \sigma^2)$$

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

$$= \text{const.} - n \ln \sigma - \sum_{i=1}^{n} \frac{(x_i - \mu)^2}{2\sigma^2}$$

$$\therefore \frac{\partial \ell}{\partial \mu} = \sum_{i=1}^{n} \frac{x_i - \mu}{\sigma^2} = 0 \Leftrightarrow \mu = \frac{1}{n} \sum_{i=1}^{n} x_i \quad - (1)$$

Example (cont.)

$$\frac{\partial \ell}{\partial \sigma} = \frac{-n}{\sigma} + \sum_{i=1}^{n} \frac{(x_i - \mu)^2}{\sigma^3} = 0 \quad - \quad (2)$$

We want to solve eqns. (1) & (2) for μ & σ. It is clear that the (unique) solution is

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

There is an example where the MLE doesn't coincide with the MoME.

Exercise : Suppose $X_1, \ldots, X_n$ are iid $\mathcal{U}([0, \theta])$ where θ is unknown. Find the MLE for θ given $x_1, \ldots, x_n$

Hint : The likelihood function $L(x_1, \ldots, x_n ; \theta)$ is non-zero only if $\theta \geq \max\{x_1, \ldots, x_n\}$

Sufficient Statistics

Let us revisit the example of estimating the probability that a coin comes up Heads when it is tossed. We saw that the likelihood depended only on the number of Heads in a coin toss sequence:

$$L(x_1, \ldots, x_n; p) = p^H (1-p)^T$$

where H & T denote the no. of heads & tails respectively. In particular, the likelihood does not depend on where in the sequence the heads & tails appear.

The no. of heads is called a sufficient statistic for this model - it is a summary statistic that is sufficient for inference, in the sense that it contains all the relevant information in the data

In the same way, in the Exp(λ) example,

$$l(x_1, \ldots, x_n; \lambda) = n\lambda \left(1 - \frac{1}{n} \sum_{i=1}^{n} x_i\right),$$

so the sample mean is a sufficient statistic

Definition: We say that a summary statistic, i.e., a function $S(x_1, \ldots, x_n)$ of the data, is a summary statistic, if the likelihood factorises as

$$L(x_1, \ldots, x_n; \theta) = f(S; \theta)\, g(x_1, \ldots, x_n; S)$$

In words, the data $x_1, \ldots, x_n$ are conditionally independent of the parameter θ given the summary statistic, S. As they are conditionally independent, they have no further information about the parameter.

Example: If $X_1, \ldots, X_n$ are iid $N(\mu, \sigma^2)$, then we can rewrite log-likelihood as

$$\ell(x_1, \ldots, x_n; \mu, \sigma^2) =$$

$$\text{const.} - n \ln \sigma - \frac{1}{2\sigma^2} \sum_{i=1}^{n} x_i^2 + \frac{\mu}{\sigma^2} \sum_{i=1}^{n} x_i - \frac{n\mu^2}{2\sigma^2}$$

Hence, in this case, $(\Sigma x_i^2, \Sigma x_i)$ is a sufficient statistic.

This example illustrates that a sufficient statistic need not be a single real number - it could be a vector.

Figures of Merit

How do we judge the quality of a statistical estimator?

Intuitively, a good estimator is one that is close to the true value.

Given a statistical model, namely a probability distribution F_θ from some family, parametrised by a parameter θ (scalar or vector-valued), the observations are modelled as iid random variables $X_1, X_2, \ldots$ with common distribution F_θ.

Any estimator $\hat{\theta}$ of θ is some function of the random variables $X_1, X_2, \ldots, X_n$, i.e., $\hat{\theta}$ is also a random variable.

The probability distribution of $\hat{\theta}$ obviously depends on θ, or equivalently, F_θ.

There are two commonly used measures of the quality of an estimator:

1. Bias $(\hat{\theta}) = \mathbb{E}(\hat{\theta}) - \theta$: ideally should be 0

2. Mean-square error : $\mathrm{MSE}(\hat{\theta}) = \mathbb{E}[(\hat{\theta}-\theta)^2] \geq 0$

   Want this to be as small as possible

Remarks

Both bias & mse are expectations, i.e., they are numbers, not random variables. So they don't depend on the data, only on the choice of estimator (e.g. MLE vs. MoME, in case these are different).

Both bias & mse do depend on the value of the unknown parameter θ, and on n, the amount of data available.

For any "good" estimator, both bias and mse should ~~decrease~~ go to 0 as $n \to \infty$.

Def.: An estimator $\hat{\theta}$ of θ is called consistent if $\hat{\theta}$ converges to θ in probability as $n \to \infty$, i.e., if

$\forall\ \varepsilon > 0,\ \mathbb{P}(|\hat{\theta} - \theta| > \varepsilon) \to 0$ as $n \to \infty$

A sufficient condition for consistency is that the bias & mse both ~~decrease~~ tend to 0 as $n \to \infty$