DECEMBER 14, 2015

LECTURE 15
INTRODUCTION TO STATISTICS: ESTIMATION

1 Statistics

Probability Theory is concerned with studying and describing the mathematical laws of uncertainty. Statistics is a closely related discipline, in which uncertainty plays the central role. The main focus of statistics is, using a finite number of measurements of variables of interest (data), to learn about the features of the probability model that governs the relationships between the variables.

Example 1. Suppose the investor has measured the monthly returns on a stock over a certain period of time (say, one year). Modeling those returns as random draws from some *unknown* distribution, the investor is interested in learning the properties of that distribution. For example, the mean of the distribution will described the expected rate of return, the variance of the distribution will describe the level of risk associated with investing in this stock. Note that both the mean and the variance are unknown, and the investor must *estimate* their values using the twelve observed monthly measurements.

Example 2. Suppose the investor has a finite number of measurements (data) of the returns on two stocks. The investor wishes to form the optimal investment portfolio of the two stocks that has a certain pre-determined expected level of return and the smallest variance (level of risk). Modeling the returns on the two stocks as random draws from some unknown bivariate distribution, the optimal portfolio (described the weights of the two stocks stocks) depends on the following properties of that distribution: the expected returns on the two stocks, their variances, and their covariance. Neither the expected returns nor the variances or the covariance are known. The investor must estimate estimate their values from the data.

Example 3. Suppose the economist has data on unemployment and inflation over the period of thirty years: for each $t = 1, \ldots, 30$, the economist has a pair of values U_t and π_t for unemployment and inflation respectively. He is interested in checking whether there is a trade-off between inflation and unemployment, i.e. if the Phillips Curve holds in practice. Assuming that for every year, U_t and π_t are draws from some unknown bivariate distribution, the economist can model the relationship between the two variables as the regression $U_t = \alpha + \beta\pi_t + \varepsilon_t$ or through the conditional expectation $E(U_t|\pi_t)$. In the regression case, the economist would be interested in estimating β to check whether the slope parameter is negative or zero. Similarly, in the conditional expectation case, the economist would be interested in estimating the conditional expectation function to check whether it is decreasing or constant (as a function of π_t).

Some key concepts of Statistics:

Data (or **sample**) is a finite collection of measurements of variables of interest. We have to distinguish between data viewed as a collection of random draws from some (multivariate) distribution, and data viewed as a collection of numbers. In the former case, we deal with a collection of *random variables* denoted $X_1, \ldots, X_n$ In the latter case - with *realizations* of the random variables denoted $x_1, \ldots, x_n$ (also called **observations**).

Population is a probability model that describes the joint distribution of random data $X_1, \ldots, X_n$. It is a mathematical law that describes the probabilistic behavior of random data or the relationships between the random variables in the data. Thus, population in statistics does not correspond to any physical population.

Estimator is any function of random data: $u(X_1, \ldots, X_n)$. It is used to learn or estimate certain properties of population (the underlying statistical model). The same function evaluated at the observed values $x_1, \ldots, x_n$, $u(x_1, \ldots, x_n)$, is called an *estimate*. An estimator is a random variable, an estimate is a realized value of the estimator.

One of the central questions in statistics is how to construct estimators with certain desirable properties. Methods of construction of estimators, as well as some important properties of estimators are discussed below for two specific models: the Bernoulli model and the Normal location-scale model.

2 Maximum Likelihood estimation and the Method of Moments

Suppose that one repeats n independent and identical Bernoulli trials that have the probability of success equal to $p \in [0, 1]$, where p is now viewed as the *unknown parameter* of interest. The random data (sample) in this example is given by $X_1, \ldots, X_n$, which are n *independent, identically distributed* (**iid**) random variables such that $X_i \sim Bernoulli(p)$. Our goal is to construct an estimator

$$\hat{p} = u(X_1, \ldots, X_n)$$

for the unknown probability of success p so that the computed estimate $u(x_1, \ldots x_n)$ is close/ informative about the unknown parameter (p).

Recall that for $x_i \in \{0, 1\}$,

$$P(X_i = x_i) = p^{x_i}(1-p)^{1-x_i}.$$

The probability of observing realizations $x_1, \ldots, x_n$ is given by

$$\begin{aligned} P(X_1 = x_1, \ldots, X_n = x_n) &= P(X_1 = x_1) \times \ldots \times P(X_n = x_n) \\ &= p^{x_1}(1-p)^{1-x_1} \times \ldots \times p^{x_n}(1-p)^{1-x_n} \\ &= p^{\sum_{i=1}^n x_i}(1-p)^{n-\sum_{i=1}^n x_i}. \end{aligned}$$

where the equality in the first line holds by the independence of $X_1, \ldots, X_n$. The Maximum Likelihood Estimator (MLE) of p is the value of p that maximizes the probability $P(X_1 = x_1, \ldots, X_n = x_n)$. In other words, we estimate p by the value that corresponds to the largest probability of observing the sample $x_1, \ldots, x_n$.

Definition 4. (Maximum Likelihood principle) Suppose that the joint distribution of data depends on a parameter θ. When viewed as a function of θ, the joint distribution distribution is called the *likelihood function*. According to the Maximum Likelihood (ML) principle, the estimator of θ is constructed by maximizing the likelihood function.

In the Bernoulli example, the likelihood function is given by

$$L(p) = p^{\sum_{i=1}^n x_i}(1-p)^{n-\sum_{i=1}^n x_i}.$$

Since $\ln(\cdot)$ (natural logarithm) is a monotone increasing one-to-one transformation, instead of maximizing $L(p)$ one can equivalently maximize $\ln L(p)$, which is called the *log-likelihood function*. As you will see below, using the log-likelihood function can substantially simplify the problem. Here, the log-likelihood function is

$$\ln L(p) = \sum_{i=1}^n x_i \times \ln p + \left(n - \sum_{i=1}^n x_i \right) \times \ln(1-p).$$

Taking the derivative of $\ln L(p)$ with respect to p, we obtain:

$$\frac{\sum_{i=1}^n x_i}{\hat{p}_{MLE}} - \frac{n - \sum_{i=1}^n x_i}{1 - \hat{p}_{MLE}} = 0.$$

Solving for $\hat{p}_{MLE}$, we obtain:

$$\hat{p}_{MLE} = \frac{1}{n} \sum_{i=1}^n x_i.$$

Hence, the MLE in the Bernoulli model is

$$\hat{p}_{MLE} = \frac{1}{n} \sum_{i=1}^n X_i = \bar{X}. \tag{1}$$

Comments:

1. One can clearly see from equation (1) that $\hat{p}_{MLE}$ is a random variable. Thus, in general it is different from the true parameter p.

2. Recall that in the Bernoulli model, p is the *population* probability of success. Similarly, $\hat{p}_{MLE} = \bar{x}$ is the sample frequency of success.

3. Recall also that in the Bernoulli model

$$p = EX_i.$$

   Thus, the parameter of interest can be computed as the population average of the variable of interest. The MLE is defined analogously as the sample average of the variable of interest:

$$\hat{p}_{MLE} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Definition 5. (Method of Moments) Suppose that some parameter of interest is defined through the expectation of some random variable X_i: $\theta = EX_i$. According to the Method of Moments (MM), one can construct an estimator for θ by replacing the expectation with the sample average: $\hat{\theta}_{MM} = n^{-1}\sum_{i=1}^{n} X_i$.

In the Bernoulli model,

$$\hat{p}_{MM} = \hat{p}_{MLE}.$$

While $\hat{p}_{MLE}$ is different from p in general, it is nevertheless informative about p. First, consider the expected value of the estimator:

$$\begin{aligned} E\hat{p}_{MLE} &= E\left(n^{-1}\sum_{i=1}^{n} X_i\right) \\ &= n^{-1}\sum_{i=1}^{n} EX_i \\ &= n^{-1}\sum_{i=1}^{n} p \\ &= p. \end{aligned}$$

Hence, the distribution of $\hat{p}_{MLE}$ is centered around the true value p. We call this property *unbiasedness.*

Definition 6. An estimator $\hat{\theta}$ of a parameter θ is called *unbiased* if $E\hat{\theta} = \theta$.

Next, consider the variance of $\hat{p}_{MLE}$.

$$\begin{aligned} Var\left(\hat{p}_{MLE}\right) &= Var\left(n^{-1}\sum_{i=1}^{n} X_i\right) \\ &= n^{-1}Var(X_1) \\ &= p(1-p)/n. \end{aligned}$$

Hence, the variance of $\hat{p}_{MLE}$ is inversely related to the sample size. As the sample size increases ($n \to \infty$), the distribution of $\hat{p}_{MLE}$ becomes more and more concentrated around the true value p. This property is known as the Law of Large Numbers: By increasing the sample size n, one can make the deviation of $\hat{p}_{MLE}$ from the true value p arbitrarily small.

3 Normal location-scale model

3.1 The model and derivation of the estimators

The Normal location-scale model is a prototypical example in Statistics: many more complicated models can be viewed as generalizations of the Normal location-scale model.

Suppose that $Y_1, \ldots, Y_n$ are iid $N(\mu, \sigma^2)$. We are interested in estimating two parameters: μ (the mean or location of the distribution) and $\sigma^2 > 0$ (the variance or scale of the distribution). To construct the estimators of μ and σ^2, we will consider the Methods of Moments approach first. Since

$$\mu = EY_i,$$

its Method of Moments estimator is

$$\hat{\mu}_{MM} = \frac{1}{n}\sum_{i=1}^{n} Y_i = \bar{Y}.$$

Since

$$\sigma^2 = Var(Y_i) = E\left(Y_i - \mu\right)^2,$$

its Method of Moments estimator is

$$\begin{aligned} \hat{\sigma}^2_{MM} &= \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{\mu}_{MM}\right)^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y})^2. \end{aligned}$$

Next, consider the MLEs of μ and σ^2. Since Y_i's are continuously distributed random

variables, we should use the PDF instead of the PMF. The normal PDF at y is given by

$$f(y;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(y-\mu)^2\right).$$

Suppose we observe $Y_1 = y_1, \ldots, Y_n = y_n$. Using the independence of Y's, the joint PDF evaluated at $y_1, \ldots, y_n$ (or the likelihood function) is given by

$$\begin{aligned}
\Pi_{i=1}^n f(y_i;\mu,\sigma^2) &= \Pi_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(y_i-\mu)^2\right) \\
&= \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^n (y_i-\mu)^2\right) \\
&\equiv L(\mu,\sigma^2).
\end{aligned}$$

Hence, the log-likelihood function is

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

Since only the last term on the right-hand side of (2) depends on μ,

$$\begin{aligned}
\hat{\mu}_{MLE} &= \arg\max_{\mu} \ln L(\mu,\sigma^2) \\
&= \arg\min_{\mu} \sum_{i=1}^n (y_i-\mu)^2. \qquad (3)
\end{aligned}$$

Hence, in the Normal location-scale model, the MLE for the location parameter can be obtained by minimizing the sum of the squared distances between the observations y_i's and the model. This estimator is also known as the Ordinary Least Squares (OLS) or the Least Squares (LS) estimator. The solution to the minimization problem in (3) is given by

$$\hat{\mu}_{MLE} = \bar{y} = \frac{1}{n}\sum_{i=1}^n y_i,$$

which can be obtained in the same manner as the result of Theorem 4 in Lecture 5.

Thus in the Normal location-scale model, all the three approaches for deriving estimators (Maximum Likelihood, Method of Moments, and Least Squares) produce the same estimator:

$$\hat{\mu}_{MLE} = \hat{\mu}_{MM} = \hat{\mu}_{LS} = \bar{Y} = \frac{1}{n}\sum_{i=1}^n Y_i.$$

Next, to obtain the MLE of σ^2, we evaluate the log-likelihood function at $\mu = \bar{y}$:

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

and note that the MLE of σ^2 is given by $\hat{\sigma}^2_{MLE} = \arg\max_{\sigma^2} \ln L(\bar{y}, \sigma^2)$. Differentiating $\ln L(\bar{y}, \sigma^2)$ with respect to σ^2, we obtain the following first-order condition:

$$-\frac{n}{2\hat{\sigma}^2_{MLE}} + \frac{1}{2\hat{\sigma}^4_{MLE}}\sum_{i=1}^{n}(y_i - \bar{y})^2 = 0,$$

and therefore,

$$\hat{\sigma}^2_{MLE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2.$$

Again, we find that the MLE and the Method of Moments estimators coincide:

$$\hat{\sigma}^2_{MLE} = \hat{\sigma}^2_{MM} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y})^2.$$

Note that μ, the mean of the distribution $N(\mu, \sigma^2)$, is estimated by the sample average $\bar{Y}$. Similarly, the variance of the distribution, σ^2, is estimated by the sample variance $n^{-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$.

3.2 Properties of the MLE of μ

Here we find that the MLE of μ is an unbiased estimator:

$$\begin{aligned} E\hat{\mu}_{MLE} &= E\bar{Y} \\ &= E\left(n^{-1}\sum_{i=1}^{n} Y_i\right) \\ &= \mu. \end{aligned}$$

For the variance of the MLE of μ, since the latter is given by the average of iid variables,

$$\begin{aligned} Var(\hat{\mu}_{MLE}) &= Var(Y_i)/n \\ &= \sigma^2/n. \end{aligned}$$

As in the case of the Bernoulli model, the distribution of the MLE of μ is centered at the true value of the parameter and becomes less dispersed as the sample size increases. In this case, it is also easy to characterize the entire distribution of $\hat{\mu}_{MLE}$.

Theorem 7. *Suppose that $Y_1, \ldots, Y_n$ are iid $N(\mu, \sigma^2)$. Then,*

$$\bar{Y} = n^{-1} \sum_{i=1}^{n} Y_i \sim N(\mu, \sigma^2/n).$$

Remark. The theorem shows that an average of iid normal random variables is also normally distributed.

Proof. To show the result, we will use the MGF of the normal distribution: it suffices to show that the MGF of $\bar{Y}$ at t is given by

$$M_{\bar{Y}}(t) = \exp\left(\mu t + \frac{(\sigma^2/n)t^2}{2}\right),$$

see the derivation of the MGF of the normal distribution in Theorem 2 of Lecture 11. For simplicity, suppose that $n = 2$, and recall that the MGF of Y_i is

$$M_{Y_i}(t) = \exp\left(\mu t + \frac{\sigma^2 t^2}{2}\right). \tag{4}$$

Consider the MGF of $Y_1 + Y_2$:

$$\begin{aligned}
M_{Y_1+Y_2}(t) &= E\exp\left(t(Y_1 + Y_2)\right) \\
&= E\left(\exp(tY_1)\exp(tY_2)\right) \\
&= E\exp(tY_1)E\exp(tY_2) \quad \text{(by independence of } Y_1 \text{ and } Y_2\text{)} \\
&= \exp\left(\mu t + \frac{\sigma^2 t^2}{2}\right) \times \exp\left(\mu t + \frac{\sigma^2 t^2}{2}\right) \quad \text{(by (4))} \\
&= \exp\left(2\mu t + \frac{2\sigma^2 t^2}{2}\right).
\end{aligned}$$

Hence,

$$Y_1 + Y_2 \sim N(2\mu, 2\sigma^2).$$

It follows now from Theorem 4 in Lecture 11 that

$$\frac{Y_1 + Y_2}{2} \sim N\left(\mu, \frac{\sigma^2}{2}\right).$$

Following the same steps in the general case, we obtain

$$\begin{aligned} \sum_{i=1}^{n} Y_i &\sim N(n\mu, n\sigma^2), \text{ and} \\ n^{-1}\sum_{i=1}^{n} Y_i &\sim N\left(\mu, \frac{\sigma^2}{n}\right). \end{aligned}$$

□

3.3 Unbiased estimation of σ^2

Unlike the MLE of μ, the MLE of σ^2 turns out to be biased.

Theorem 8. *Suppose that $Y_1, \ldots, Y_n$ are iid $N(\mu, \sigma^2)$. For $\hat{\sigma}^2_{MLE} = n^{-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$,*

$$E\hat{\sigma}^2_{MLE} = \frac{n-1}{n}\sigma^2.$$

Proof. Using the same arguments as in Theorem 2(d) in Lecture 6, we can write

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \frac{1}{n}\sum_{i=1}^{n} Y_i^2 - \left(\bar{Y}\right)^2.$$

Hence

$$E\left(\frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y})^2\right) = \frac{1}{n}\sum_{i=1}^{n} EY_i^2 - E\left(\bar{Y}\right)^2. \tag{5}$$

Also, since for any random variable W, $Var(W) = EW^2 - (EW)^2$, we can write $EW^2 = Var(W) + (EW)^2$. Therefore,

$$\begin{aligned} EY_i^2 &= Var(Y_i) + (EY_i)^2 \\ &= \sigma^2 + \mu^2, \qquad (6) \\ E\left(\bar{Y}\right)^2 &= Var(\bar{Y}) + (E\bar{Y})^2 \\ &= \frac{\sigma^2}{n} + \mu^2. \qquad (7) \end{aligned}$$

Substituting (6) and (7) into (5), we obtain:

$$\begin{aligned} E\hat{\sigma}^2_{MLE} &= \left(\sigma^2 + \mu^2\right) - \left(\frac{\sigma^2}{n} + \mu^2\right) \\ &= \frac{n-1}{n}\sigma^2. \end{aligned}$$

□

While the estimator $\hat{\sigma}^2_{MLE}$ is biased, its bias takes the scaling form, where the scaling factor depends only on the known characteristics of the data: the sample size n. It is therefore very easy to come up with bias correction and propose an alternative unbiased estimator.

Theorem 9. *Let $Y_1, \ldots, Y_n$ be iid $N(0, \sigma^2)$. Define the following estimator of σ^2:*

$$s^2 = \frac{n}{n-1}\hat{\sigma}^2_{MLE} = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2.$$

Then s^2 is an unbiased estimator, i.e.

$$Es^2 = \sigma^2.$$

Proof. Using the result of Theorem 8,

$$\begin{aligned} Es^2 &= \frac{n}{n-1}E\hat{\sigma}^2_{MLE} \\ &= \frac{n}{n-1} \times \frac{n-1}{n}\sigma^2 \\ &= \sigma^2. \end{aligned}$$

□

The following intuitive explanation is often provided for the division by $n-1$ instead of n in the definition of s^2. The definition of the variance

$$\sigma^2 = E(Y_i - \mu)^2$$

involves μ, the mean of the distribution. If μ were known, one could construct an unbiased estimator of σ^2 as

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

(it is very easy to check that this estimator would be unbiased). When $\mu$is unknown and must be estimated first by the average $\bar{Y}$, one must account for that by changing the denominator in the definition of s^2 from n to $n-1$.