Technical notes on statistical inference (estimation)

Yang Xu

Inferential statistics provides the mathematical means and procedures for deducing properties of a population, typically through analyses of a sample - a subset of a population. Broadly speaking, statistical inference includes estimation, i.e., inference of unknown parameters that characterize one or more populations, and testing, i.e., evaluation of hypotheses about one or more populations. This technical note focuses on some bare essentials of statistical estimation.

1. Statistic *vs.* parameter

   A statistic is a function of a sample, e.g., arithmetic mean of a sample. A parameter is a descriptor of a population, value of which is inferred from a sample of data, e.g., mean of a normally distributed population is typically inferred from a sample of that population.

2. Independent and identically distributed (iid) assumption

   A common assumption in statistical estimation is that samples are independent, such that the value of a particular sample is not determined by or dependent on values of any other samples, and simultaneously, that samples are identically distributed, such that there exists an underlying distribution from which all samples are generated.

3. Maximum likelihood estimator (MLE)

   Likelihood, or $L(\theta|\mathbf{x})$, is a function that models an observed sample $\mathbf{x}$ with unknown parameter(s) θ. Estimators of θ are derived from data in $\mathbf{x}$. A maximum likelihood estimator (Fisher, 1912, 1922) corresponds to parameter values of θ that maximize $L(\theta|\mathbf{x})$. It follows from the maximum likelihood principle that the "optimal" parameter values for a model are the ones that maximize the likelihood function with respect to observed data.

   Example: Obtaining MLEs for normal (or Gaussian) likelihood
   Suppose that $\mathbf{x} = \{x_1, ..., x_n\}$ is an iid sample distributed according to $\mathcal{N}(\mu, \sigma^2)$, the likelihood function based on the specified model and the observed data can then be formulated as:

   $$L(\theta|\mathbf{x}) = L(\mu, \sigma|\mathbf{x}) \overset{iid}{=} \prod_{i=1}^{n} f(x_i|\mu, \sigma) = \prod_{i=1}^{n} \mathcal{N}(x_i|\mu, \sigma^2). \tag{1}$$

   The MLEs correspond to a set of parameter values (μ,σ) that maximize $\prod_{i=1}^{n} \mathcal{N}(x_i|\mu, \sigma^2)$. Typically, MLEs are obtained by maximizing the logarithm of the likelihood function:

   $$\log L(\theta|\mathbf{x}) = \log \prod_{i=1}^{n} \mathcal{N}(x_i|\mu, \sigma^2) = -\frac{n}{2}\log(2\pi\sigma^2) - \sum_{i=1}^{n} \frac{(x_i-\mu)^2}{2\sigma^2}. \tag{2}$$

In this case, the MLEs can be derived analytically by setting the partial first derivatives to 0:

$$\frac{\partial \log L(\theta|\mathbf{x})}{\partial \mu} = 0 \Rightarrow \hat{\mu} = \frac{\sum_i x_i}{n}, \textit{ i.e. sample mean}; \tag{3}$$

$$\frac{\partial \log L(\theta|\mathbf{x})}{\partial \sigma^2} = 0 \Rightarrow \hat{\sigma^2} = \frac{\sum_i (x_i - \hat{\mu})^2}{n}, \textit{ i.e. sample variance.} \tag{4}$$

4. Sampling distribution and standard error

The sampling distribution is a probability distribution of a statistic, considering that a statistic is a random variable, or a function of a sample. The standard error of a statistic is the standard deviation of its sampling distribution - it measures the variability or dispersion of the statistic due to sampling from a population.

5. Central limit theorem (CLT)

The central limit theorem (in its coarse form) states that for a sequence of iid random variables that have finite means and variances, the average of that sequence would be normally distributed in the limit, i.e. when the sample size is sufficiently large.

Implications of CLT (proofs omitted)

5.1. If $x_1, ..., x_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$, then it follows:

$$\bar{x} \sim \mathcal{N}(\mu, \frac{\sigma^2}{n}), \textit{ when } n \rightarrow \infty, \tag{5}$$

i.e. standard error of the mean (SEM) of normal variables, or the standard deviation of the sampling distribution of the normal mean, would itself converge to a normal in the limit.

5.2. If $y_1, ..., y_n$ are iid variables of any arbitrary distribution with finite mean μ_y and variance σ_y^2, then $\bar{y} \sim \mathcal{N}(\mu_y, \frac{\sigma_y^2}{n})$, when $n \rightarrow \infty$, i.e. sampling distribution of the mean of any arbitrary variables would converge to a normal distribution in the limit.

6. Confidence interval

Confidence interval specifies a range of values that would likely include the true parameter value, where the range is estimated from a given sample. The level of a confidence interval corresponds to the probability that it would include the true parameter value.

Example: 95% confidence interval for the normal mean
Based on CLT (6.1.), it follows that the average of a set of normal varibles, $\bar{x}$, follows a standard normal distribution after transformation:

$$\frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} = z \sim \mathcal{N}(0, 1). \tag{6}$$

Since $p(-1.96 < z < 1.96) = 0.95$ covers 95% area under the probability density function of z, the 95% confidence interval for $\bar{x}$ is then $\bar{x} \pm 1.96 \times \frac{\sigma}{\sqrt{n}}$, or $[\bar{x} - 1.96\frac{\sigma}{\sqrt{n}}, \bar{x} + 1.96\frac{\sigma}{\sqrt{n}}]$.

7. Bootstrap

Not all sampling distributions have analytical forms as in the case of the mean. Bootstrap (Efron, 1979) is a simulation-based method for estimating sampling distribution of an arbitrary statistic based on the procedure of resampling, or more precisely, sampling with replacement. Efron (1979) postulates that the simulated sampling distribution (of a parameter), or the empirical bootstrap distribution, based on a given sample, should converge in probability to the true distribution in the limit, when the resampling repetitions approach infinity.

Example: Procedures for bootstrapping the mean
Given a sample $\mathbf{x} = \{x_1, ..., x_n\}$, the sampling distribution of $\bar{x}$ can be approximated by:
i. Repeat the following steps m times (m is typically large, e.g., $m \geq 1,000$):
- Randomly sample x's with replacement from $\mathbf{x}$ up to n data points, denoting it as $\mathbf{x}^b$;
- Compute the average of the bootstrapped sample $\mathbf{x}^b$, denoting it as $\bar{x^b}$;
ii. Obtain the empirical distribution for $\bar{x}$ based on the bootstrapped means $\{\bar{x^b_1}, ..., \bar{x^b_m}\}$.
SEM or confidence interval of $\bar{x}$ can then be estimated from this empirical distribution.

8. Pearson correlation

Pearson correlation (Pearson, 1895) measures linear dependence between a pair of variables, X and Y, by comparing the degree to which they co-vary against the degrees that they vary individually. Formally, the Pearson correlation coefficient is defined as follows:

$$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} = \frac{E[X - \mu_X][Y - \mu_Y]}{\sqrt{E[(X - \mu_X)^2]}\sqrt{E[(Y - \mu_Y)^2]}}. \tag{7}$$

It can be shown that this correlation coefficient has a maximal value of 1 and a minimal value of -1, indicating perfectly positive and negative dependence between the two variables respectively. When the cofficient is close to 0, there is minimal linear dependence between the two variables.

9. Linear regression model

A univariate linear regression model specifies a linear relationship between an observed independent variable X and an observed dependent variable Y:

$$Y = f(X) + \epsilon = \alpha + \beta X + \epsilon, \tag{8}$$

where it says that the value of Y is a linear transformation of that of X via the function $f(X) = \alpha + \beta X$, subject to some error ϵ due to random noise in the data. A commonly used distribution for the error term is a normal distribution with zero mean (i.e., white noise):

$$\epsilon \sim \mathcal{N}(0, \sigma^2) \tag{9}$$

Therefore Y is modelled as a normal distribution that centers at $f(X)$ with variance σ^2. It follows that for n iid samples $\mathbf{x} = \{x_1, ..., x_n\}$ and $\mathbf{y} = \{y_1, ..., y_n\}$, the likelihood function is:

$$L(\theta|\mathbf{x}, \mathbf{y}) = L(\alpha, \beta, \sigma|\mathbf{x}, \mathbf{y}) = \prod_i f(y_i|x_i, \alpha, \beta, \sigma) = \prod_i \mathcal{N}(y_i|\alpha + \beta x_i, \sigma^2). \tag{10}$$

The logarithm of the likelihood function is then:

$$\log L(\theta|\mathbf{x}, \mathbf{y}) = \log \prod_i \mathcal{N}(y_i|\alpha + \beta x_i, \sigma^2) = -\frac{n}{2}\log(2\pi\sigma^2) - \sum_{i=1}^{n} \frac{(y_i - f(x_i))^2}{2\sigma^2}. \tag{11}$$

Taking partial derivatives yields MLE solutions (a.k.a. ordinary least squares or OLE):

$$\begin{aligned}\frac{\partial \log L(\theta|\mathbf{x},\mathbf{y})}{\partial\beta} &= 0 \Rightarrow \hat{\beta} = \frac{\sigma_{\mathbf{xy}}}{\sigma_{\mathbf{x}}^2}; \quad (12)\\ \frac{\partial \log L(\theta|\mathbf{x},\mathbf{y})}{\partial\alpha} &= 0 \Rightarrow \hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}. \quad (13)\end{aligned}$$

Given the MLEs, the residuals r of the linear regression model describe the degree of deviation between values of the target dependent variable and the estimated values of those:

$$r_i = y_i - \hat{y}_i = y_i - \hat{f}(x_i) = y_i - (\hat{\alpha} + \hat{\beta}x_i),\ i = 1, ..., n. \quad (14)$$

10. Mean squared error (MSE).

Mean squared error measures the average squared deviation between a target and its estimate. In the case of linear regression, MSE can be used to measure the mean of squared errors between values of the dependent variable and those of its expected values:

$$MSE(\hat{Y}) = E[(\hat{Y} - Y)^2] = E[(\hat{f}(X) - Y)^2] = \frac{\sum_i(\hat{f}(x_i) - y_i)^2}{n}. \quad (15)$$

It turns out that this MSE can be understood as a tradeoff between variance and bias (of an estimator):

$$MSE(\hat{Y}) = Bias(\hat{Y})^2 + Var(\hat{Y}) + \sigma_Y^2, \quad (16)$$

where $Bias(\hat{Y}) = E[\hat{f}(X)] - f(X)$, $Var(\hat{Y}) = E[(\hat{f}(X) - E[\hat{f}(X)])^2]$, and σ_Y^2 represents irreducible error due to random fluctuations in the data.