EE 381 Probability & Statistic with Applications to Computing (Fall 2020)

Lecture 11
Estimation Theory

Duc H. Tran, PhD

Estimation Theory

- Unbiased Estimates, Efficient Estimates
- Confidence Intervals

Introduction about Estimation Theory

Estimation theory is a branch of statistics that deals with estimating the values of parameters based on measured empirical data that has a random component.

Example:

- Alternatively, it is desired to estimate the probability of a voter voting for a candidate, based on some demographic features (i.e., age).
- We use radar to find the range of objects (airplanes, boats, etc.) by analyzing the two-way transit timing of received echoes of transmitted pulses. Since the reflected pulses are unavoidably embedded in electrical noise, their measured values are randomly distributed, so that the transit time must be estimated.

Introduction about Estimation Theory

Three main estimation methods:

- **Maximum likelihood estimation**
- Minimum mean squared error (MMSE) estimation
- Bayesian estimation.

These methods are widely used, all of them have pros and cons.

Unbiased Estimates and Efficient Estimates

- A statistic is called an *unbiased estimator* of a population parameter if the mean or expectation of the statistics is equal to the parameter.
- The corresponding value of the statistic is called an *unbiased estimate* of the parameter.

Example: Theorem 5-1

The mean of the sampling distribution of means equal to the mean of the population.

$$E(\bar{X}) = \mu_{\bar{X}} = \mu$$

Unbiased Estimates and Efficient Estimates

- If the sampling distributions of two statistics have the same mean, the statistic with the smaller variance is called a *more efficient estimator* of the mean.
- The corresponding value of the efficient statistic is then called an *efficient estimate*.

Example: For a normal population, the sampling distribution of the *mean* and *median* both have the same mean, namely, the population mean. However, the variance of the sampling distribution of means is smaller than that of the sampling distribution of medians.

→ The *mean* provides a more *efficient estimate* than the *median*.

Unbiased Estimates and Efficient Estimates

Example: A sample of five measurements of the diameter of a sphere were recorded by a scientist as 6.33, 6.37, 6.36, 6.32, and 6.37 cm. Determine unbiased and efficient estimates of the true mean and the true variance. (Assumed that the measured diameter is normally distributed).

Solution:

An unbiased and efficient estimate of the true mean (i.e., population mean) is:

$$\bar{x} = \frac{\sum x}{n} = \frac{6.33 + 6.37 + 6.36 + 6.32 + 6.37}{5} = 6.35\ cm$$

An unbiased and efficient estimate of the true variance (i.e., population variance) is:

$$\hat{s}^2 = \frac{n}{n-1} S^2 = \frac{\sum (x - \bar{x})^2}{n-1} = \frac{(6.33 - 6.35)^2 + \cdots + (6.37 - 6.35)^2}{5-1} = 0.00055$$

However, $\hat{s}$ is a biased estimator of σ, since $E(\hat{S}) \neq \sigma$. Meaning that $\hat{s} = \sqrt{0.00055} = 0.023$ is an estimate of the true standard deviation, but this estimate is not unbiased.

Point Estimates and Interval Estimates

- An estimate of a population parameter given by a single number is called a *point estimate* of the parameter.
- An estimate of a population parameter given by two numbers between which the parameter may be considered to lie is called an *interval estimate* of the parameter.

Example: If we say that a distance is 5.28 feet, we are giving a point estimate.

If we say that the distance is 5.28 ± 0.03 feet, we are giving an interval estimate.

- A statement of the error or precision of an estimate is often called its *reliability*.

Confidence Interval

- Confidence interval: the measure of the quality of an estimate. It is a range of values that are close to the estimate.
- A small confidence interval means the estimate is likely to be close to the correct value, a larger one indicates the estimate may be inaccurate.
- As expected, confidence intervals shrink as more data are obtained.

Confidence Interval Estimates of Population Parameters

Let μ_S and σ_S be the mean and standard deviation of the sampling distribution of a statistic S.

- If the sampling distribution of S is approximately normal, we can expect to find S lying in the intervals $(\mu_S - \sigma_S,\ \mu_S + \sigma_S)$, $(\mu_S - 2\sigma_S,\ \mu_S + 2\sigma_S)$, or $(\mu_S - 3\sigma_S,\ \mu_S + 3\sigma_S)$ about 68.27%, 95.45%, and 99.73% of the time, respectively.
- Equivalently, we can be confident of finding μ_S in the intervals $(S - \sigma_S,\ S + \sigma_S)$, $(S - 2\sigma_S,\ S + 2\sigma_S)$, or $(S - 3\sigma_S,\ S + 3\sigma_S)$ about 68.27%, 95.45%, and 99.73% of the time, respectively.
- We calls these intervals the 68.27%, 95.45%, and 99.73% *confidence intervals* for estimating μ_S.
- The end numbers of these intervals $(S \pm \sigma_S, S \pm 2\sigma_S, S \pm 3\sigma_S)$ are then called the 68.27%, 95.45%, and 99.73% *confidence limits*.

Confidence Interval Estimates of Population Parameters

Example: $S \pm 1.96\sigma_S$ and $S \pm 2.58\sigma_S$ are 95% and 99% (or 0.95 and 0.99) confidence limits for μ_S.

- The percentage confidence is often called the *confidence level*.
- The number 1.96, 2.58 in the confidence limits are called *critical values*, and are denoted by z_c. From *confidence levels*, we can find *critical values* and conversely.

Confidence Level	99.73%	99%	98%	96%	95.45%	95%	90%	80%	68.27%	50%
z_c	3.00	2.58	2.33	2.05	2.00	1.96	1.645	1.28	1.00	0.6745

Some values of corresponding to various confidence levels used in practice. (Appendix C)

Confidence Interval for Means

Large Samples ($\boldsymbol{n \geq 30}$): If the statistic S is the sample mean $\bar{X}$, the confidence limits for the population mean are given by:

$$\bar{X} \pm z_c \frac{\sigma}{\sqrt{n}}$$

in case sampling from an infinite population or sampling with replacement from a finite population.

Or given by:

$$\bar{X} \pm z_c \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}$$

in case sampling is without replacement from a population of finite size N.

Example: If the statistic S is the sample mean $\bar{X}$, then 95% and 99% confidence limits for estimation of the population mean μ are given by $\bar{X} \pm 1.96\sigma_{\bar{X}}$ and $\bar{X} \pm 2.58\sigma_{\bar{X}}$, respectively.

The confidence limits are given by $\bar{X} \pm z_c\sigma_{\bar{X}}$.

Confidence Interval for Means

Small Samples ($\boldsymbol{n < 30}$) and Population Normal: In this case, we use the Student's t distribution to obtain confidence levels. The confidence limits for population means are given by:

$$\bar{X} \pm t_c \frac{\hat{S}}{\sqrt{n}}$$

Example: if $-t_{0.975}$ and $t_{0.975}$ are the values of T for which 2.5% of the area lies in each tail of the Student's t distribution, then a 95% confidence interval for T is given by:

$$-t_{0.975} < \frac{(\bar{X} - \mu)\sqrt{n}}{\hat{S}} < t_{0.975}$$

from which we see that μ can be estimated to lie in the interval:

$$\bar{X} - t_{0.975} \frac{\hat{S}}{\sqrt{n}} < \mu < \bar{X} + t_{0.975} \frac{\hat{S}}{\sqrt{n}}$$

with 95% confidence.

Confidence Interval for Means

Example: Find the 95% confidence interval for estimating the mean height of the XYZ University students. Suppose that the heights of 100 male students represent a random sample of the heights of all 1546 male students at the university.

Solution: The 95% confidence limits are $\bar{X} \pm 1.96 \frac{\sigma}{\sqrt{n}}$.

Previously calculated: $\bar{x} = 67.45$ inches and $\hat{s} = 2.93$ inches.

Use them as an estimate of σ, the confidence limits are: $67.45 \pm 1.96 \frac{2.93}{\sqrt{100}}$, or 67.45 ± 0.57 inches.

Then the 95% confidence interval for the population mean μ is: $66.88 < \mu < 68.02$ inches.

Note: we assumed that the population was infinite or so large that we could consider conditions to be the same as sampling with replacement.

Confidence Interval for Proportions

Suppose that the statistic S is the proportion of successes in a sample of size $n \geq 30$ drawn from a binomial population in which p is the proportion of successes (the probability of success).

The confidence limits for p are given by

$$P \pm z_c\sqrt{\frac{pq}{n}} = P \pm z_c\sqrt{\frac{p(1-p)}{n}}$$

in case sampling is from an infinite population or if sampling is with replacement from a finite population.

Similarly, if sampling is without replacement from a population of finite size N, the confidence limits are:

$$P \pm z_c\sqrt{\frac{pq}{n}}\sqrt{\frac{N-n}{N-1}}$$

Confidence Interval for Proportions

Example: A sample poll of 100 voters chosen at random from all voters in a given district indicated that 55% of them were in favor of a particular candidate. Find 95% and 99% confidence limits for the proportion of all the voters in favor of this candidate.

Solution:

a) The 95% confidence limits for the population *p* are:

$$P \pm 1.96\sigma_P = P \pm 1.96\sqrt{\frac{p(1-p)}{n}} = 0.55 \pm 1.96\sqrt{\frac{0.55 * 0.45}{100}} = 0.55 \pm 0.1$$

Where we have used the sample proportion 0.55 to estimate *p*.

b) The 99% confidence limits for the population *p* are:

$$0.55 \pm 2.58\sqrt{\frac{0.55 * 0.45}{100}} = 0.55 \pm 0.13$$

Confidence Interval for Differences and Sums

- If S_1 and S_2 are two sample statistics with approximately normal sampling distributions, confidence limits for the differences of the population parameters corresponding to S_1 and S_2 are given by:

$$S_1 - S_2 \pm z_c \sigma_{S_1 - S_2} = S_1 - S_2 \pm z_c \sqrt{\sigma_{S_1}^2 + \sigma_{S_2}^2}$$

- The confidence limits for the sum of the population parameters are:

$$S_1 + S_2 \pm z_c \sigma_{S_1 + S_2} = S_1 + S_2 \pm z_c \sqrt{\sigma_{S_1}^2 + \sigma_{S_2}^2}$$

Example: The confidence limits for the difference of two population means, in the case where the populations are infinite and have known standard deviations σ_1, σ_2 are given by:

$$\bar{X}_1 - \bar{X}_2 \pm z_c \sigma_{\bar{X}_1 - \bar{X}_2} = \bar{X}_1 - \bar{X}_2 \pm z_c \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

where $\bar{X}_1$, n_1 and $\bar{X}_2$, n_2 are the respective means and sizes of the two samples drawn from the populations.

Confidence Interval for Differences and Sums

Example: A sample of 150 brand A light bulbs showed a mean lifetime of 1400 hours and a standard deviation of 120 hours. A sample of 200 brand B light bulbs showed a mean lifetime of 1200 hours and a standard deviation of 80 hours. Find the 95% confidence limits for the difference of the mean lifetimes of the populations of brands A and B.

Solution:

Confidence limits for the difference in means of brands A and B are:

$$\bar{X}_A - \bar{X}_B \pm z_c \sqrt{\frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B}}$$

The 95% confidence limits are: $1400 - 1200 \pm 1.96\sqrt{\frac{120^2}{150} + \frac{80^2}{100}} = 200 \pm 24.8$

Therefore, we can be 95% confident that the difference of population means lies between 175 and 225.

Maximum Likelihood Estimates

- Although confidence limits are valuable for estimating a population parameter, it is still convenient to have a single or point estimate → we employ *maximum likelihood method.*
- Assume that the population has a density function that contains a population parameter, θ, which is to be estimated by a certain statistic. The density function can be denoted by $f(x, \theta)$.
- Assuming that there are *n* independent observations, $X_1, \ldots, X_n$, the joint density function for these observations is:

$$L = f(x_1, \theta) f(x_2, \theta) \ldots f(x_n, \theta)$$

which is called the *likelihood*.

Maximum Likelihood Estimates

The *maximum likelihood* can then be obtained by taking the derivative of L with respect to θ and setting it equal to 0.

In some cases, it is easier to work with the log likelihood function:

$$\ln L(x_1, x_2, \ldots, x_n, \theta)$$

Meaning that it is convenient to first take logarithms, and then take the derivative.

The solution of this equation, for θ in terms of the x_k, is known as the *maximum likelihood estimator of* θ.

In case there are several parameters, take the partial derivatives with respect to each parameter, set them equal to 0, and solve the resulting equations simultaneously.

Maximum Likelihood Estimates

Example: Suppose that n observations $X_1, \dots, X_n$ are made from a normally distributed population of which the mean is unknown, and the variance is known. Find the maximum likelihood estimate of the mean.

Solution:

Since $f(x_k, \mu) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_k-\mu)^2}{2\sigma^2}}$

We have: $L = f(x_1, \mu) f(x_2, \mu) \dots f(x_n, \mu) = 2\pi\sigma^{2^{-\left(\frac{n}{2}\right)}} e^{-\sum\frac{(x_k-\mu)^2}{2\sigma^2}}$

Therefore,

$$\ln L = -\left(\frac{n}{2}\right)\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum(x_k - \mu)^2$$

Taking the partial derivative with respect to μ yields:

$$\frac{1}{L}\frac{\partial L}{\partial \mu} = \frac{1}{\sigma^2}\sum(x_k - \mu)$$

Maximum Likelihood Estimates

Example (ctn.):

Setting $\frac{\partial L}{\partial \mu} = 0$ gives:

$$\sum (x_k - \mu) = 0 \qquad i.e. \sum x_k - n\mu = 0$$

Or
$$\mu = \frac{\sum x_k}{n} = \mu_{\bar{X}}$$

Therefore, not surprisingly, the maximum likelihood estimate of the mean is the sample mean $\mu_{\bar{X}}$. (theorem 5-1).

Maximum Likelihood Estimates

Example: Assume we have a bag that contains 3 balls. Each ball is either red or blue, but we have no information in addition to this. Let call the number of blue balls θ, it might be 0, 1, 2, 3. We are allowed to choose 4 balls at random from the bag with replacement. Let random variables X_1, X_2, X_3, X_4 represent the 4 chosen balls, and they are defined as follows:

$$X_i = \begin{cases} 1 & if\ the\ ith\ chosen\ ball\ is\ blue \\ 0 & if\ the\ ith\ chosen\ ball\ is\ red \end{cases}$$

Estimate the number of blue balls, θ, in the bag if the observed sample are: $(x_1, x_2, x_3, x_4) = (1, 0, 1, 1)$.

Note:

❖ X_i's are independent and identically distributed (i.i.d.) and $X_i \sim Bernoulli(\frac{\theta}{3})$

❖ The observed sample are: $(x_1, x_2, x_3, x_4) = (1, 0, 1, 1)$, meaning that we observe 3 blue balls and 1 red ball.

Maximum Likelihood Estimates

Solution: Since $X_i \sim Bernoulli\left(\frac{\theta}{3}\right)$, we have

$$P_{X_i}(x) = \begin{cases} \dfrac{\theta}{3} & for\ x = 1 \\ 1 - \dfrac{\theta}{3} & for\ x = 0 \end{cases}$$

Since X_i's are independent, the joint PMF of X_i's will be:

$$P_{X_1X_2X_3X_4}(x_1, x_{2,}\, x_3, x_4) = P_{X_1}(x_1)P_{X_2}(x_2)P_{X_3}(x_3)P_{X_4}(x_4)$$

$$P_{X_1X_2X_3X_4}(1,0,1,1) = \frac{\theta}{3}.\left(1 - \frac{\theta}{3}\right).\frac{\theta}{3}.\frac{\theta}{3} = (\frac{\theta}{3})^3\left(1 - \frac{\theta}{3}\right)$$

The probability of the observed sample is maximized for $\theta = 2$. This mean that the observed data is most likely to occur for $\theta = 2$.

Therefore, we estimate $\theta = 2$.

We called this the MLE of θ.

θ	$P_{X_1X_2X_3X_4}(1,0,1,1;\theta)$
0	0
1	0.0247
2	0.0988
3	0

Maximum Likelihood Estimates

Solution (ctn.): let confirm the result by using the general formula of maximum likelihood method.

We have: $L = P_{X_1 X_2 X_3 X_4}\left(x_1, x_{2,} x_3, x_4\right) = (\frac{\theta}{3})^3\left(1 - \frac{\theta}{3}\right) = (\frac{\theta}{3})^3 - (\frac{\theta}{3})^4$

Take the derivative of L with respect to θ:

$$\frac{\partial L}{\partial \theta} = 3(\frac{\theta}{3})^2 - 4(\frac{\theta}{3})^3$$

Set it equal 0, we get: $\theta = 2.25$.

Since number of blue balls should be an integer, we estimate $\theta = 2$.

Reference

Notes, equations, and figures in the lecture are based on or taken from materials in the course textbook:

"Probability and Statistics", by Spiegel, Schiller and Srinivasan, ISBN 987-007-179557-9 (McGraw-Hill/Schaun's)