Estimation

Dr Wan Nor Arifin
Unit of Biostatistics and Research Methodology, Universiti Sains Malaysia.
wnarifin@usm.my

Last update: 1 October, 2018

Outlines

Introduction....2
  Statistics....2
  Statistical inference....2
Estimation....2
  Estimation....2
  Estimator and estimate....2
    Estimator....2
    Estimate....3
    Notations....3
  Properties of good estimators*....3
    Unbiased estimator....3
    Small variability....3
    Small error....3
    Efficient....4
    Consistent....4
  Interval estimate: Confidence interval....4
  Interval estimator and estimate....5
    Estimator....5
    Estimate....5
  Coverage probability....5
  Confidence interval: Simulation....6
One Population Mean....6
  Standard normal distribution, *z*....6
  Student's t distribution....7
One Population Proportion....8
One Population Variance....10
Maximum Likelihood Estimation*....10
  Likelihood....10
  Maximum likelihood estimation (MLE)....11
Topics for Self-study....12
References....12

Introduction

Statistics

- It "is a field of study concerned with the collection, organization, summarization and analysis of data, and the drawing of inferences about a body of data when only part of the data is observed" (Daniel, 1995).

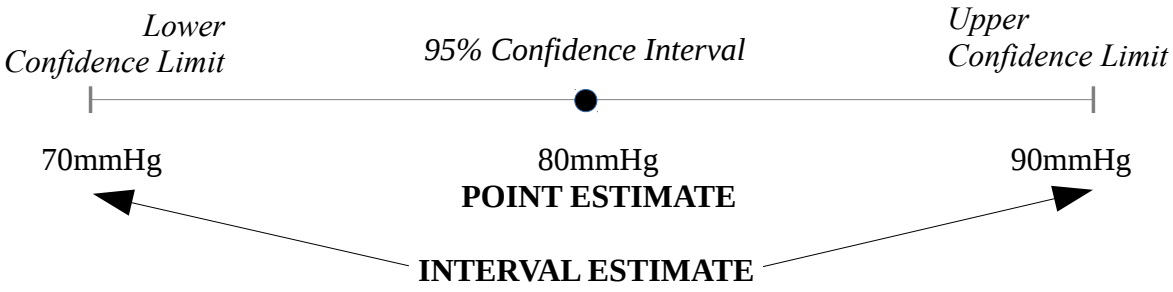
Statistical inference

- It "is the procedure by which we reach a conclusion about a population on the basis of information contained in a sample drawn from that population" (Daniel, 1995).
- It reflects the definition of statistics i.e. making inference about a body of data (population) from part of the data (sample).
- Numerical values calculated from:
  - Population → Parameter (denoted as theta, θ).
  - Sample → Statistic.
- Two approaches of statistical inference: *Estimation* and *Hypothesis testing*.

Estimation

Estimation

- It is the process of calculating a *statistic* from a sample data as an approximation of a *parameter* of the population from which the sample was drawn.
- An *estimate* is an approximation of a *parameter*.
- For each parameter, two types of estimate are possible:
  - *Point estimate* – a single numerical value used as an estimation of a parameter value.
  - *Interval estimate* – consists of two numerical values presented in form of range/interval within which we believe the true parameter value is included, given with specified *confidence level*. This is known as *confidence interval*.
- For example, in a study it was reported that "the mean value of diastolic blood pressure is 80mmHg (95% CI: 70, 90)". Thus, visually



Estimator and estimate

Estimator

- A *point estimator* is a function of a sample, say $W(X_{1,}\ldots,X_n)$.
- This function is referred as a statistic.

- For example, mean is a function

$$mean(X_{1,}\dots,X_n)=\frac{1}{n}\sum_{i=1}^{n}X_i=\bar{X}$$

Estimate

- A *point estimate* is the realized value of a statistic/estimator, say $W(x_{1,}\dots,x_n)$.
- For example, the mean of a sample is

$$mean(x_{1,}\dots,x_n)=\frac{1}{n}\sum_{i=1}^{n}x_i=\bar{x}$$

Notations

$$\begin{aligned}\text{parameter} &= \theta\\ \text{estimate} &= \hat{\theta}\end{aligned}$$

for example, of mean

$$\begin{aligned}\text{parameter} &= \mu\\ \text{estimate} &= \bar{x}\end{aligned}$$

Properties of good estimators*

Unbiased estimator

- Bias is the difference between the expected value of W and θ,

$$Bias_\theta=E_\theta(W)-\theta$$

- When the expected value of W gets close to θ, it gets less biased.

Small variability

- The variance of an estimator,

$$Var_\theta(W)$$

or in form of its standard error,

$$SE_\theta(W)=\sqrt{Var_\theta(W)}$$

- Smaller variance/standard error is desirable.

Small error

- Simultaneous measure of both bias and variance of an estimator is given by mean squared

error (MSE),

$$MSE_\theta(W) = E_\theta(W-\theta)^2 = Var_\theta(W) + (E_\theta(W)-\theta)^2 = Var_\theta(W) + Bias_\theta(W)^2$$

for an unbiased estimator (bias = 0) we get,

$$MSE_\theta(W) = E_\theta(W-\theta)^2 = Var_\theta(W)$$

- To put this in the context of the estimator of mean,

$$E(\bar{X}-\mu)^2 = Var(\bar{X}) = \frac{\sigma^2}{n}$$

Efficient

- When we have two estimators W and V, we can compare the relative efficiency by the ratio of their variances or MSEs,

$$Eff_\theta(W, V) = \frac{Var_\theta(V)}{Var_\theta(W)}$$

$$Eff_\theta(W, V) = \frac{MSE_\theta(V)}{MSE_\theta(W)}$$

Consistent

- An estimator W is a consistent estimator of a parameter θ if the estimate gets closer to the true parameter value as sample size *n* increases to infinity.
- W converges in probability to θ, with the probability converges to 1,

$$P(|W-\theta| < \epsilon) \rightarrow 1 \quad \text{as } n \rightarrow \infty \text{ for any } \epsilon > 0.$$

Interval estimate: Confidence interval

- Presenting **point estimate** (a single value) is not enough, it should be accompanied by its **interval estimate** in form of a **confidence interval**.
- When we estimate, we make an educated guess (estimate) of the true population parameter.
- It is like saying "I think that based on my sample, the mean DBP is 80 mmHg and I am 95% sure that the real mean DBP in population is between 70 to 90mmHg."
- We present in form of a point estimate followed by its interval estimates for a given confidence level:

point estimate (% confidence level: lower confidence limit, upper confidence limit)

- Usually 95% confidence level for the confidence interval.
- Generally to obtain confidence interval:

$$\text{point estimate} \pm (\text{reliability coefficient}) \times (\text{standard error})$$

or in short

$$\text{point estimate} \pm \text{precision}$$

- **Reliability coefficient** depends on the statistical distribution and confidence level.
- **Standard error** is the standard deviation of the sampling distribution. It shows how "erratic" are the samples of a population.
- Smaller standard error gives us narrower confidence interval, thus smaller precision for a given confidence level.
- Think of an analog watch which can give you the time precise to the closest one second as compared to a digital atomic clock used in advanced research center which can give you reading as precise as a fraction of a second.
- A more formal definition of a confidence interval is "a range of values that has been calculated in such a way that if the calculation is repeated on a large number of samples, the percentage of confidence intervals that cover the true parameter value will be equal to the desired confidence interval" (Marschner, 2015). We will see this in form of simulations later.

Interval estimator and estimate

Estimator

- An *interval estimator* is a pair of functions of a sample, $[L(X_1, \ldots, X_n), U(X_1, \ldots, X_n)]$.

Estimate

- A *interval estimate* is a pair of the realized values the interval estimator, $[L(x_1, \ldots, x_n), U(x_1, \ldots, x_n)]$.
- These satisfy $L(x_{1,\ldots}, x_n) \leq U(x_1, \ldots, x_n)$, so that $L(x_1, \ldots, x_n) \leq \theta \leq U(x_1, \ldots, x_n)$.
- For example, the interval estimate of mean of $X \sim N(\mu, \sigma^2)$

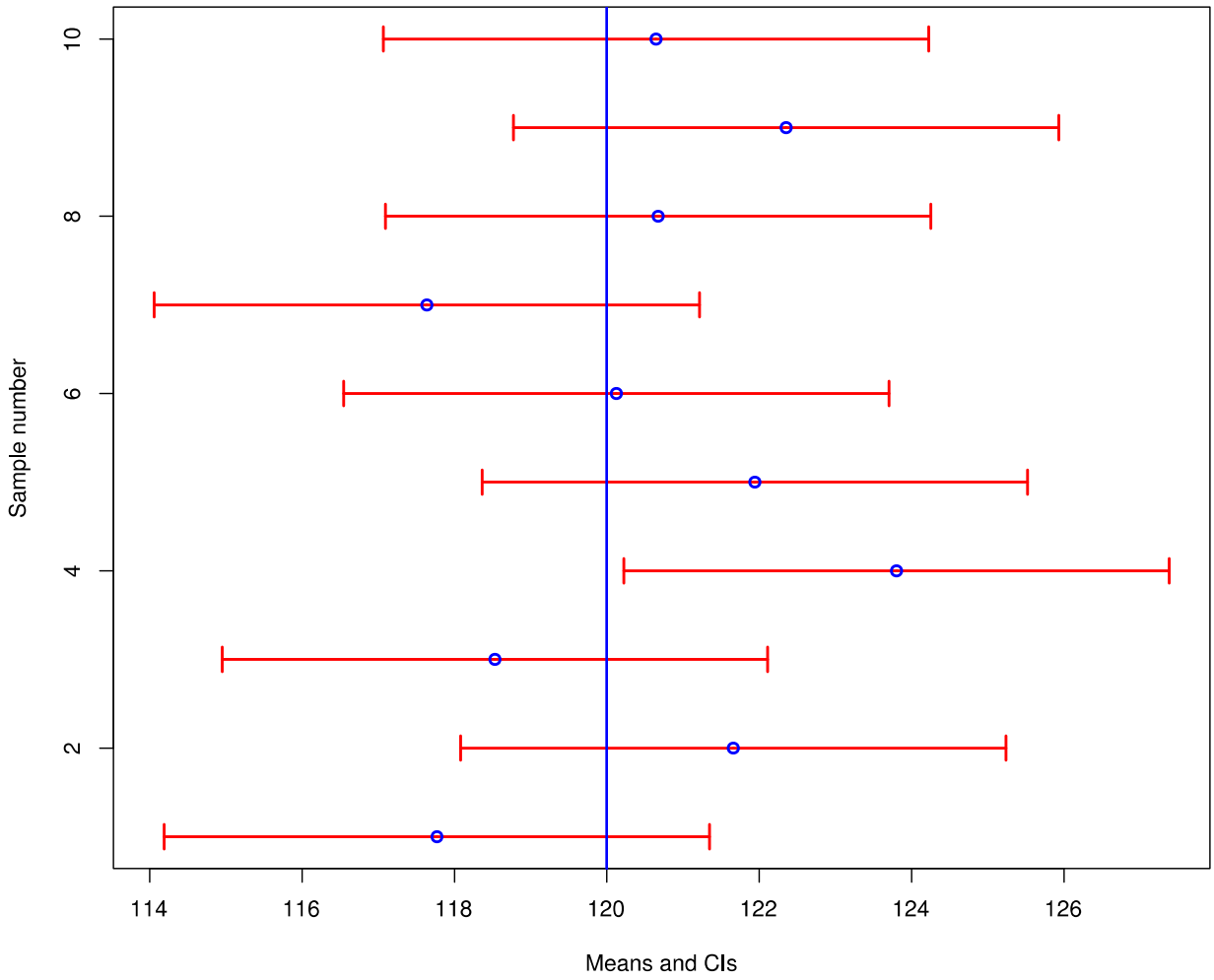
$$\left[\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$$

for $(1 - \alpha)$ confidence level

$$P_{\mu}(\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}) = 1 - \alpha$$

Coverage probability

- Coverage probability is the percentage of repeated samples of a population that have confidence intervals including the true parameter value.
- The coverage probability should be equal to the confidence level to say that a confidence interval is valid.
- For a simple example, below is a plot means and 95% CIs of samples $n = 10$ from $X \sim N(120, 10^2)$. The coverage probability = 9/10 samples include true mean of 120 = 90% (which is less than the confidence level of 95%).



Confidence interval: Simulation

- Using R: Simulate confidence intervals of calculated from samples of $X \sim N(120, 10^2)$. Open `estimation.R` script.

One Population Mean

- Recall that to find average/mean $\bar{x}$ of numerical data:

$$\bar{x} = \frac{\sum x_i}{n}$$

- To find confidence interval for one sample mean, there are two ways: Using **standard normal distribution, *z*** or **Student's t distribution, *t***.

Standard normal distribution, *z*

- Data are normally distributed.
- Population standard deviation σ, is known.
- If σ is not known (which is the case most of the time), for a large sample size (usually 30 or

more), the sample standard deviation *s*, *could* be used in place of σ.

- The confidence interval is given by,

$$\text{point estimate} \pm (\text{reliability coefficient}) \times (\text{standard error})$$

$$\bar{x} \pm z_{(1-\alpha/2)} \times \sigma_{\bar{x}}$$

$$\bar{x} \pm z_{(1-\alpha/2)} \times \frac{\sigma}{\sqrt{(n)}}$$

- Commonly used reliability coefficient using *z* distribution $z_{(1-\alpha/2)}$ by $(1-\alpha)\times 100\,\%$ confidence level are:

$$\alpha = 0.10,\ (1-\alpha)100\% = 90\% \rightarrow 1.65$$
$$\alpha = 0.05,\ (1-\alpha)100\% = 95\% \rightarrow 1.96$$
$$\alpha = 0.01,\ (1-\alpha)100\% = 99\% \rightarrow 2.58$$

Example 1:

From his data on systolic blood pressure (SBP) collected from 30 patients, a researcher found that the mean SBP was 120mmHg with SD of 15mmHg. Estimate with 95% confidence the population mean of SBP.

$$\bar{x} = 120$$
$$s = 15 \approx \sigma$$
$$\bar{x} \pm z_{(1-\alpha/2)} \times \sigma / \sqrt{(n)}$$
$$120 \pm 1.96 \times 15 / \sqrt{(30)}$$
$$120 \pm 1.96 \times 2.739$$
$$95\ CI:\ 114.6, 125.4$$

Stated in sentence:

"We are 95% confident that the population mean is between 114.6 and 125.4."

or in journal presentation form:

mean = 120mmHg (95% CI: 114.6, 125.4)

Using R: `estimation.R` script.

Student's t distribution

- When we cannot use *z* distribution (i.e when sample size is too small, *s* is not a reliable approximation of σ in case σ is not known).
- Includes degree of freedom *df*, given by *df = n-1* to take into account the sample size.
- Different values of $t_{(1-\alpha/2)}$ for different *df*s, thus we will have different reliability coefficients.
- For larger *n*, $t_{(1-\alpha/2)}$ values are very close to $z_{(1-\alpha/2)}$ values.
- Confidence interval is given by,

$$\text{point estimate} \pm (\text{reliability coefficient}) \times (\text{standard error})$$

$$\bar{x} \pm t_{(1-\alpha/2)} \times \sigma_{\bar{x}}$$

$$\bar{x} \pm t_{(1-\alpha/2)} \times \frac{s}{\sqrt{(n)}}$$

- Find reliability coefficient using *t* distribution $t_{(1-\alpha/2)}$ by $(1-\alpha)\times 100\%$ confidence level using R:

  ```
  qt(p = 1 – alpha/2, df = n - 1)
  ```

  Let say, for n = 15, $(1-\alpha)\times 100\%$ confidence level (i.e. alpha = 0.1, 0.05, 0.01), $t_{(1-\alpha/2)}$

  90% → ?
  95% → ?
  99% → ?

Example 2:

From his data on systolic blood pressure (SBP) collected from 20 patients, a researcher found that the mean SBP was 120mmHg with SD of 15mmHg. Estimate with 95% confidence the population mean of SBP.

$$\bar{x} = 120$$
$$s = 15$$
$$n = 20$$
$$df = n - 1 = 19$$
$$\bar{x} \pm t_{(1-\alpha/2,\, df=19)} \times s/\sqrt{(n)}$$
$$120 \pm 2.0930 \times 15/\sqrt{(20)}$$
$$120 \pm 2.0930 \times 3.354$$
$$95\ CI: 113.0, 127.0$$

Stated in sentence:

"We are 95% confident that the population mean is between 113.0 and 127.0."

or in journal presentation form:

mean = 120mmHg (95% CI: 113.0, 127.0)

Using R: `estimation.R` script.

One Population Proportion

- Common in medical and health sciences; percentage of HIV positive among drug addicts, proportion of smokers died of lung cancer etc.
- Sample proportion $\hat{p}$. Sample size *n*.

- Sampling distribution of $\hat{p}$ is quite close to normal distribution when both *np* and *n(1-p)* greater than 5 (i.e when sample size *n* is large and the proportion in population *p* not too small).
- Examples,

$n = 1000$, $p = 0.05$, $1 - p = 0.95$; $np = 50$, $n(1 - p) = 950$
→ use z distribution.

$n = 50$, $p = 0.05$, $1 - p = 0.95$; $np = 2.5$, $n(1 - p) = 47.5$
→ cannot calculate confidence interval based on *z* distribution.

- Confidence interval is given by,

$$\text{point estimate} \pm (\text{reliability coefficient}) \times (\text{standard error})$$

$$\hat{p} \pm z_{(1-\alpha/2)} \times \sigma_{\hat{p}}$$

$$\hat{p} \pm z_{(1-\alpha/2)} \times \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

- Similarly, commonly used reliability coefficient using *z* distribution $z_{(1-\alpha/2)}$ by confidence level are:

90% → 1.65
95% → 1.96
99% → 2.58

Example 3:

It was found that in a study among drug addicts in Kelantan, 130 out of 200 are HIV positive. Construct 99% confidence interval for the proportion of HIV positive among the addicts.

$$\hat{p} = 130/200 = .65$$
$$n = 200$$
$$\hat{p} \pm z_{(1-\alpha/2)} \times \sqrt{\hat{p}(1-\hat{p})/n}$$
$$.65 \pm 2.58 \times \sqrt{(.65(.35)/200)}$$
$$99\ CI: 0.5623, 0.7377$$

Stated in sentence:

"We are 99% confident that the population proportion *p* is between 56.23% and 73.77%."

or in journal presentation form:

proportion = 65.0% (95% CI: 56.23%, 73.77%)

Using R: `estimation.R` script.

One Population Variance

- Concerns the confidence interval of sample variance and standard deviation.
- Confidence interval of variance is given by,

$$\frac{(n-1)s^2}{\chi^2_{1-\alpha/2}} < \sigma^2 < \frac{(n-1)s^2}{\chi^2_{\alpha/2}}$$

- Confidence of interval of standard deviation is given by,

$$\sqrt{\frac{(n-1)s^2}{\chi^2_{1-\alpha/2}}} < \sigma < \sqrt{\frac{(n-1)s^2}{\chi^2_{\alpha/2}}}$$

Example 4:

In a study measuring systolic blood pressure involving 30 subjects, it was found the standard deviation was 15. Construct 95% confidence interval for the variance.

$$s = 15$$
$$s^2 = 15^2 = 225$$
$$n = 30$$
$$\frac{(n-1)s^2}{\chi^2_{1-\alpha/2}} < \sigma^2 < \frac{(n-1)s^2}{\chi^2_{\alpha/2}}$$
$$\frac{(29)225}{45.72} < \sigma^2 < \frac{(29)225}{16.05}$$
$$142.71 < \sigma^2 < 406.62$$
$$95\ CI: 142.71, 406.62$$

Stated in sentence:

> "We are 95% confident that the population variance σ^2 is between 142.71 and 406.62."

or in journal presentation form:

s^2 = 225.00 (95% CI: 142.71, 406.62)

Using R: `estimation.R` script.

Maximum Likelihood Estimation*

Likelihood

- Likelihood of a parameter value given data or sample value, written as $L(\theta \mid x)$. This is likelihood function.
- Likelihood function equals the pdf/pmf that,

$$L(\theta|x) = f_X(x|\theta)$$

- Simple example (example 2 from *Probability Distribution* lecture), using binomial distribution, say proportion of diabetics $p = 0.3$, for a sample $n = 10$, what is the likelihood of x = 4?

$$L(\theta|x)=P(X=x|\theta)=f_X(x|\theta)$$
$$L(0.3|4)=P(X=4|0.3)=f_X(4|0.3)=0.21$$

- However, while pdf/pmf $f_X(x|\theta)$ considers varying values of x given θ, likelihood function $L(\theta|x)$ considers varying values of θ given x. In other words, we are looking for the most plausible parameter value whenever we have the sample value at hand.
- Since we are dealing with a sample of size n, with $\mathbf{x}=(x_1,\ldots,x_n)$ observations, the likelihood function is a product of pdfs,

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

for *k* plausible parameters $\theta=(\theta_1,\ldots,\theta_k)$. Or in form of log likelihood (to reduce computational burden),

$$l(\theta|\mathbf{x})=\sum_{i=1}^{n} \log f(x_i|\theta)$$

Maximum likelihood estimation (MLE)

- It is a general method of estimation whenever there is no obvious method to estimate parameters in complicated situations. We learn about it here to learn its principles.
- MLE is used to find the most supported parameter value given data, that is the one with the highest likelihood based the likelihood function. We want to find the parameter value that **maximizes** the likelihood function.
- This is done by differentiating the likelihood function with respect to θ to find its derivative and setting it to zero, and solve

$$\frac{d}{d\theta}L(\theta|\mathbf{x})=0$$

to find the maximum stationary point (θ)of the likelihood function.
- This is followed by finding the second derivative,

$$\frac{d^2}{d^2\theta}L(\theta|\mathbf{x})<0$$

solving for θ found above. If the value is negative, then θ is the maximum point.
- For complex likelihood functions, the maximum is found by numerical method through a number of iterations by computer software. Hence our purpose here is to know the principles behind the MLE.

Topics for Self-study

Based on your knowledge of the sampling distributions from lecture Sampling Distributions, construct confidence interval for the

- difference between two population means.
- difference between two population proportions.
- ratio between two population variances.

Estimation methods:

- Least squares estimation.
- Method of moments estimation.

References

Casella, G., & Berger, R. L. (2002). Statistical inference. Delhi, India: Cengage Learning.

Daniel, W. W. (1995). *Biostatistics: A foundation for analysis in the health sciences* (6th ed.). USA: John Wiley & Sons.

Marschner, I. C. (2015). Inference principles for biostatisticians. USA: CRC Press.

Rice, J. A. (1995). *Mathematical statistics and data analysis* (2nd ed.). USA: Duxbury Press.

Tijms, H. (2007). *Understanding probability: Chances rules in everyday life* (2nd ed.). New York, USA: Cambridge University Press.