Estimation, Confidence Interval and Hypothesis Testing

Tonglin Zhang
Data

- Many kinds of data: numeric, text, image, audio, and et, but we focus on numeric data.
- In statistics, we only have data.
- The simplest case: repeat the same thing independently.
- We denote them by \( \{X_1, X_2, \ldots, X_n\} \)
  
  if we treat data as random variables (we are going to collect data), or
  
  \( \{x_1, x_2, \ldots, x_n\} \)
  
  if we treat data as real values (we have obtained the data).
- To understand the methodology, we adopt the first case.
Some Important Concepts

- **Parameters**: A parameter is an unknown constant which affects the distribution of random variables. Parameters can only be estimated (learned, CS term) from data.

- **Statistic**: A function of data, which becomes a real value or a real vector if data are available.

- **Estimation**: The entire method. You need to provide the details for the derivation.

- **Estimator**: A mathematical formula. It is usually provided by estimation.

- **Estimate**: A real value or a real vector. Based on data, an estimator becomes an estimate.
Data are treated as random before they are available, or as values after they are available. Only data are real.

The goal of theoretical statistics is to provide methods for data.

The goal of applied statistics is to implement methods to data.

Models or assumptions: artificial or subjective, may be changed, but useful.

Notations (textbook): upper cases mean random variables (e.g. $X$, $Y$, and etc); Lower cases mean observations (e.g. $x$, $y$ and etc.)
**Example 1:** Let $X \sim N(\mu, \sigma^2)$. Suppose we repeat observe $X$ many times such that the data can be expressed as $X_1, X_2, \ldots, X_n$. Then, the statistical model can be expressed as

$$X_1, X_2, \ldots, X_n \sim iid \ N(\mu, \sigma^2).$$

Because $\mu$ and $\sigma^2$ are unknown, we can only estimate (learn) them by data. The method is to use the sample mean

$$\hat{\mu} = \bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

for $\mu$ and the sample variance

$$\hat{\sigma}^2 = S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$$

for $\sigma^2$. 
Then,

- The method for the derivation of the above two formulae is called estimation.
- The two formulae are called estimators.
- After you have the data and put them inside, you will have two numeric values. They are called the estimate of $\mu$ and the estimate of $\sigma^2$, respectively.
- Consider the data:
  Then, we have $\bar{x} = 5.384$ and $s^2 = 0.663^2$. Thus, the estimate of $\mu$ is 5.384 and the estimate of $\sigma^2$ is $0.663^2 = 0.4395$. 
Example 2: Suppose that one wants to study the quality of a nail machine. The person examines 1000 nails and find that 47 of those were bad. What is the probability for bad nails.

Solution. Let $p = P(\text{the nail is bad})$ and $X$ be the number of nails in $n$ nails. Then, $X \sim Bin(n, p)$. The estimator of $p$ is

$$\hat{p} = \frac{X}{n}.$$

The estimate of $p$ is

$$\frac{47}{1000} = 0.047.$$

Thus, the (estimated) probability for bad nails is 0.047.
**Example 3:** It is known that the number of radioactive particles within a given period follows Poisson distribution. Suppose one studies this problem for a particular radioactive isotopes with a certain mass. The person observes the number of particles as 

$$22, 14, 27, 36, 32.$$  

Find the intensity of the radiations of the isotopes with the same time period and same mass.  

**Solution:** Let $\lambda$ be the intensity. Based on statistical theory (you will learn it in the future), the estimate of $\lambda$

$$\hat{\lambda} = \frac{1}{5}(22 + 14 + 27 + 36 + 32) = 26.2.$$  

Thus, the estimate of the intensity is 26.2.
Evaluation of Estimation

- The standard measure for estimation is MSE (mean square error). Let $\hat{\theta}$ be the estimator of $\theta$. The MSE is defined as

$$E[(\hat{\theta} - \theta)^2].$$

- In the comparison of two estimators, we say one is better than the other if it has the lower MSE values.

- The simulation method is an important tool in comparison of estimators.
Confidence Interval

We consider two models here.

- Normal model: Suppose data are observed from $N(\mu, \sigma^2)$ such that they can be expressed as

$$X_1, X_2, \cdots, X_n \sim N(\mu, \sigma^2).$$

We want to provide the confidence interval for $\mu$.

- Binomial model: Suppose we repeat F/S experiments $n$ times such that we have many F/S results. We use $X$ to represent the number of successes. Then, we have

$$X \sim Bin(n, \theta).$$

We want to provide the confidence interval for $\theta$. 
Example 4: (normal confidence interval). Let $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$. Then,

- If $\sigma$ is known, the $1 - \alpha$ level confidence interval for $\mu$ is
  \[ \bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}. \]

- If $\sigma$ is unknown and $n$ is large (e.g., $n \geq 40$), the $1 - \alpha$ level confidence interval for $\mu$ is
  \[ \bar{x} \pm z_{\alpha/2} \frac{s}{\sqrt{n}}. \]

- If $\sigma$ is unknown and $n$ is small (e.g., $n < 40$), the $1 - \alpha$ level confidence interval for $\mu$ is
  \[ \bar{x} \pm t_{\alpha/2, n-1} \frac{s}{\sqrt{n}}. \]
A measure for evaluate confidence interval is the coverage probability, which is defined as

\[ P\{\text{lower} \leq \theta \leq \text{upper}\}, \]

which should be a function of \( \theta \). Thus, it is often written as

\[ P_\theta\{\text{lower} \leq \theta \leq \text{upper}\}. \]

We use simulations to evaluate coverage probabilities.
**Example 5:** Let $X \sim Bin(n, \theta)$. The $1 - \alpha$ level confidence interval for $\theta$ is

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

where $\hat{\theta} = x/n$. We use simulation to evaluate coverage probabilities.
Example 6: Let $\mu_0$ be a target value in Example 4. Let $\mu_0$ be the target value. There are three kinds of testing problems. Let $\alpha$ be the significance level.

(1). Testing whether the true value of $\mu$ is not over the target value $\mu_0$. It is expressed as

\[ H_0 : \mu \leq \mu_0 \leftrightarrow H_1 : \mu > \mu_0. \]
When $\sigma$ is known, we conclude $H_0$ is False if
\[
Z = \frac{\bar{X} - \mu_0}{\sigma / \sqrt{n}} > z_\alpha.
\]

When $\sigma$ is unknown and $n$ is large (e.g. $> 40$), we conclude $H_0$ is False if
\[
Z = \frac{\bar{X} - \mu_0}{S / \sqrt{n}} > z_\alpha.
\]

When $\sigma$ is unknown and $n$ is small (e.g. $< 40$), we conclude $H_0$ is False if
\[
T = \frac{\bar{X} - \mu_0}{S / \sqrt{n}} > t_{\alpha,n-1}.
\]
(2). Test whether the true value of $\mu$ is not less than the target value $\mu_0$. It is expressed as

$$H_0 : \mu \geq \mu_0 \leftrightarrow H_1 : \mu < \mu_0.$$ 

- When $\sigma$ is known, we conclude $H_0$ is False if

$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} < -z_\alpha.$$ 

- When $\sigma$ is unknown and $n$ is large, we conclude $H_0$ is False if

$$Z = \frac{\bar{X} - \mu_0}{S/\sqrt{n}} < -z_\alpha.$$ 

- When $\sigma$ is unknown and $n$ is small, we conclude $H_0$ is False if

$$T = \frac{\bar{X} - \mu_0}{S/\sqrt{n}} < -t_{\alpha,n-1}.$$
(3). Test whether the true value of $\mu$ is equal to the target value $\mu_0$. It is expressed as

$$H_0 : \mu = \mu_0 \leftrightarrow H_1 : \mu \neq \mu_0.$$ 

- When $\sigma$ is known, we conclude $H_0$ is False if

$$|Z| = \left| \frac{\bar{X} - \mu_0}{\sigma / \sqrt{n}} \right| > z_{\alpha/2}.$$ 

- When $\sigma$ is unknown and $n$ is large, we conclude $H_0$ is False if

$$|Z| = \left| \frac{\bar{X} - \mu_0}{S / \sqrt{n}} \right| > z_{\alpha/2}.$$ 

- When $\sigma$ is unknown and $n$ is small, we conclude $H_0$ is False if

$$|T| = \left| \frac{\bar{X} - \mu_0}{S / \sqrt{n}} \right| > t_{\alpha/2,n-1}.$$
Example 7. Let $\theta_0$ be the target value in Example 5. We still have three testing problems. Let $\alpha$ be the significance level and

$$Z = \frac{\hat{\theta} - \theta_0}{\sqrt{\hat{\theta}(1 - \hat{\theta})/n}}.$$
Hypothesis Testing

- (1) Test whether the true $\theta$ is not over $\theta_0$ or not. Here, we have

$$H_0 : \theta \leq \theta_0 \iff H_1 : \theta > \theta_0.$$ 

We conclude $H_0$ is False if $Z > z_{\alpha}$.

- (2) Test whether the true $\theta$ is at least $\theta_0$ or not. Here, we have

$$H_0 : \theta \geq \theta_0 \iff H_1 : \theta < \theta_0.$$ 

We conclude $H_0$ is False if $Z =< -z_{\alpha}$.

- (3) Test whether the true $\theta$ is equal $\theta_0$ or not. Here, we have

$$H_0 : \theta = \theta_0 \iff H_1 : \theta \neq \theta_0.$$ 

We conclude $H_0$ is False if $|Z| > z_{\alpha/2}$. 

Tonglin Zhang, Department of Statistics, Purdue University

Estimation, Confidence Interval, and Hypothesis Testing
Evaluations

Since the decision can only be made based on data, one cannot guarantee that the decision is always consistent with the truth. Therefore, there are two types of errors based on the following table. I am going to interpret those by a few examples.

<table>
<thead>
<tr>
<th>Truth</th>
<th>Conclusion</th>
<th>Truth</th>
</tr>
</thead>
<tbody>
<tr>
<td>True</td>
<td>Correct</td>
<td>False</td>
</tr>
<tr>
<td>False</td>
<td>Type I error</td>
<td>Correct</td>
</tr>
</tbody>
</table>
**Example.** Let $X_1, \cdots, X_{10} \sim iid N(\mu, 1)$. Then, $\sigma_0^2 = 1$. Suppose we want to test

$$H_0 : \mu \leq 2 \leftrightarrow H_1 : \mu > 2.$$ 

Then, we choose $\mu_0 = 2$, which can be understood as the threshold value. Choose significance level $\alpha = 0.05$. By the formula, we reject $H_0$ if

$$\bar{X} > \mu_0 + \frac{1.645\sigma_0}{\sqrt{10}} = 2.5202.$$ 

The type I error probability is a function of $\mu$ given by

$$P\{\bar{X} > 2.5201|\mu \leq \mu_0\} = P(\text{Conclude } H_1|H_0)$$

under $X_1, \cdots, X_{10} \sim iid N(\mu, 1)$, which is equivalent to

$$\bar{X} \sim N(\mu, 1/10).$$
The type II error probability is also a function of $\mu$ given by

$$P\{\bar{X} \leq 2.5201|\mu > \mu_0\} = P(\text{Conclude } H_0|H_1).$$

The power function is

$$P\{\bar{X} > 2.5201\} = \begin{cases} 
P(\text{Type I}) & \text{under } H_0 \\
1 - P(\text{Type II}) & \text{under } H_1
\end{cases}$$

The significance level is

$$\alpha = \max P(\text{type I}).$$
In practice, there is an important concept called $p$-value, which describes the probability of seeing the data given that the statement is true.

The mathematical formula for $p$-value has been developed.

Please read my R code for more interpretations.
Example. In failure/success experiment, we can assume the total number of successes $X \sim Bin(n, p)$. Suppose we want to test

$$H_0 : p \leq p_0 \leftrightarrow H_1 : p > p_0$$

for a certain $p_0$. For example, we can choose $p_0 = 0.5$ in gambling problems. Choose significance level $\alpha = 0.05$. By the formula, we reject $H_0$ if

$$X/n > p_0 + 1.645 \sqrt{\frac{p_0(1-p_0)}{n}}.$$ 

The type I error probabilities is

$$P\{X/n > p_0 + 1.645 \sqrt{\frac{p_0(1-p_0)}{n}} \mid p \leq p_0\}$$

under $X \sim Bin(n, p)$. 
The type II error probability is also a function of $p$ given by

$$P\{X/n \leq p_0 + 1.645 \sqrt{\frac{p_0(1-p_0)}{n}} | p > p_0\} = P(\text{Conclude } H_0 | H_1).$$

The power function is

$$P\{X/n > p_0 + 1.645 \sqrt{\frac{p_0(1-p_0)}{n}}\} = \begin{cases} P(\text{Type I}) & \text{under } H_0 \\ 1 - P(\text{Type II}) & \text{under } H_1 \end{cases}$$

The significance level is

$$\alpha = \max P(\text{type I}).$$

I am going to explain more based on my R code.