FALL 2013
Stat 8003: Statistical Methods I
Lecture 6

Jichun Xie

1 Parameter Estimation

1.1 Bias, Variance and MSE

Among all the possible choices, which estimate should we use? What is a good estimate?

- What kind of estimators get the answer close to the truth on average? (Bias)
- What kind of estimators get the answer close to the truth most of the time? (Variance)
- When sample size is large, can we get an estimator approaches to the truth? (Convergence)

1.1.1 Bias

Let T be out estimator. On average (in repeated experiments), we want T to be equal to the parameter of interest θ.

Definition of Bias. Let $T = g(Y_1, \ldots, Y_n)$. Suppose the parameter of interest is θ. Define

$$\texttt{Bias}(T) = \texttt{E}(T) - \theta.$$

If T is unbiased, $\texttt{Bias}(T) = 0$.

Example.

1. Let $Y_1, \ldots, Y_n \sim \text{Bernoulli}(\theta)$, *i.i.d.* Suppose $W = \frac{1}{n}\sum_{i=1}^{n} Y_i$. Then

$$\mathtt{E}(W) = \frac{1}{n}\sum_{i=1}^{n} \mathtt{E}(Y_i) = \theta.$$

W is unbiased.

2. Let $Y_1, \ldots, Y_n \sim N(\mu, \sigma^2)$, *i.i.d.* Assure yourself that $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} Y_i$ is unbiased estimator of μ. Suppose μ is known. Assure yourself that $\tilde{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \mu)^2$ is unbiased estimator σ^2.

3. Following Example 2, if μ is unknown, then we cannot use it in the statistic for estimating σ^2. Then

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

is biased. Assure yourself that it is true. (Key: in the bracket, add $+\mu - \mu$)

From Example 3, we know that

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

Let $s^2 = \frac{n}{n-1}\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$. Then s^2 is unbiased. And when the sample size n is large, the bias of $\hat{\sigma}^2$ approaches to zero. We call $\hat{\sigma}^2$ asymptotically unbiased.

1.1.2 Efficiency

Let T_1 and T_2 be unbiased estimators of θ. T_1 is more efficient than T_2 if

$$\texttt{Var}(T_1) < \texttt{Var}(T_2).$$

The relative efficiency is $R(T_1, T_2) = \texttt{Var}(T_2)/\texttt{Var}(T_1)$. When $R(T_1, T_2) < 1$, T_2 is more efficient than T_1.

Example.

Suppose $Y_1, \ldots, Y_n \sim \text{N}(\mu, \sigma^2)$, *i.i.d.*. Let $W_1 = \bar{Y}$ and $W_2 = Y_1$. Note that $\texttt{E}(W_1) = \texttt{E}(W_2) = \mu$. They are both unbiased. However $\texttt{Var}(W_1) = \sigma^2/n$ and $\texttt{Var}(W_2) = \sigma^2$. It leads to $R(W_1, W_2) = n \geq 1$, so that W_1 is more efficient than W_2 unless $n = 1$.

1.1.3 Mean Square Error

Suppose that W_1 and W_2 are two statistics of the parameter of interest θ. At least of them is biased. How to compare their performance?

Example. Suppose $Y_1, \ldots, Y_n \sim N(\mu, \sigma^2)$, *i.i.d.*, with both μ and σ^2 unknown. We just discussed that there are two estimators of σ^2: $\hat{\sigma}^2$ and s^2. We know that

$$\texttt{Bias}(\hat{\sigma}) = -\frac{1}{n}\sigma^2, \ \texttt{Bias}(s^2) = 0$$
$$\texttt{Var}(\hat{\sigma}) = \frac{(n-1)^2}{n^2}\texttt{Var}(s^2)$$

Which one should we use?

Definition of Mean Square Error. Suppose $T = g(Y_1, \ldots, Y_n)$ is a statistic for the parameter of interest θ. Then

$$\texttt{MSE}(T) = \texttt{E}(T - \theta)^2.$$

Note that

$$\texttt{MSE}(T) = \texttt{E}\{(T - \texttt{E}T) + (\texttt{E}T - \theta)\}^2 = \texttt{E}(T - \texttt{E}T)^2 + (\texttt{E}T - \theta)^2 = \texttt{Var}(T) + \texttt{Bias}(T)^2.$$

MSE is a combined measure of Bias and MSE.

Example.

In the above example, what are the MSE of $\hat{\sigma}^2$ and s^2? We know the biases for both. We only need to get the variances. To get the variance, we need the following result:

If $Y_1, \ldots, Y_n \sim \mathrm{N}(\mu, \sigma^2)$, *i.i.d.*, then

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}(Y_i - \bar{Y})^2 \sim \chi^2(n-1).$$

For $Z \sim \chi^2(k)$, $\mathtt{E}(Z) = k$ and $\mathtt{Var}(Z) = 2k$.

Then it is easy to see that

$$\mathtt{Var}(s^2) = \frac{2}{n-1}\sigma^4$$
$$\mathtt{Var}(\hat{\sigma}^2) = \frac{2(n-1)}{n^2}\sigma^4$$

Then

$$\mathtt{MSE}(s^2) = \{\mathtt{Bias}(s^2)\}^2 + \mathtt{Var}(s^2) = \frac{2}{n-1}\sigma^4$$
$$\mathtt{MSE}(\hat{\sigma}^2) = \{\mathtt{Bias}(\hat{\sigma}^2)\}^2 + \mathtt{Var}(\hat{\sigma}^2) = \frac{2n-1}{n^2}\sigma^4$$

To compare their MSE,

$$\frac{\mathtt{MSE}(\hat{\sigma}^2)}{\mathtt{MSE}(s^2)} = \frac{(2n-1)(n-1)}{2n^2} = \frac{(n-1/2)(n-1)}{n^2} < 1.$$

Therefore, in terms of MSE, $\hat{\sigma}^2$ is more efficient than s^2. When the sample size is large, they are asymptotically equivalently efficient.

$$\lim_{n\to\infty}\frac{\mathtt{MSE}(\hat{\sigma}^2)}{\mathtt{MSE}(s^2)} = 1.$$

Summary of Criterion of Comparison:

1. Unbiasedness
2. Relative Efficiency
3. MSE

2 Confidence Interval

2.1 Spread of the Data

Consider a sample of *i.i.d.*random variables, $Y_1, \ldots, Y_n$. The most common estimate of the spread is the sample variance, *i.e.*

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

- $\mathtt{E}(s^2) = \sigma^2 = \mathtt{E}(Y_i - \mu)^2$, unbiased
- Not robust, sensitive to outliers. Why?

Other estimate of spread:

- Range: $Y_{(n)} - Y_{(1)}$.
- Interquartile: $Q_{0.75} - Q_{0.25}$.
- Median absolute deviation (MAD): $\mathtt{MAD} = \mathtt{median}_i|Y_i - \mathtt{median}_i Y_i|$

2.2 Example

CNN/ORC Poll. Sept. 6-8, 2013. $N = 1,022$ adults nationwide. Margin of error ± 3. "Which of the following is the most important issue facing the country today? The economy. Health care. The situation in Syria. The federal budget deficit. The environment. Gun policy. Immigration."

	%
The economy	41
Health care	16
The situation in Syria	15
The federal budget deficit	13
The environment	5
Gun policy	5
Immigration	3
Other (vol.)	1

What does margin of error (MOE) mean?

Let

$$Y_i = \begin{cases} 1 & \text{if the } i\text{-th person chooses "The economy";} \\ 0 & \text{otherwise.} \end{cases}$$

It is reasonable to assume $Y_i \sim \text{Bernoulli}(p)$, $i = 1, \ldots, N$.

First, how to estimate p? The poll results showed that

$$\hat{p} = \sum_{i=1}^{N} Y_i/N = 0.41.$$

Assure yourself that $\hat{p}$ is both MLE and MOM estimator of p. And it is unbiased. But how to describe the precision of this estimator?

2.3 Definition of Confidence Interval

A $(1-\alpha)\%$ confidence interval (CI) is an interval (L, U) that traps the parameter of interest, say θ, with $(1-\alpha)\%$ "confidence". Thus, for all $\theta \in \Theta$,

$$\mathtt{P}(L < \theta < U) \geq 1 - \alpha.$$

Note that L and U are functions of the data. They are also random variables.

From a frequencist's point of view, θ is fixed. We repeat the experiment 100 times, and calculate how many of those times do we expect the interval to contain θ. Suppose with $(1-\alpha)\%$ of the times, the interval contains θ. Then the CI is called a $(1-\alpha)\%$ CI.

2.4 Methods to Calculate CI

2.4.1 Pivot Method.

1. This approach finds a pivot (*e.g.*, often based on differences for a location parameter or ratios of a scale parameter) first. The pivot is a function of the data and the parameter of interest, whose distribution is known and independent of the parameter of interest.
2. We then make a probability statement about the pivot.
3. Invert the results from Step 2 to get a confidence interval.

Example. Let $Y_1, \ldots, Y_n i.i.d. \text{N}(\mu, \sigma^2)$, with σ^2 known. We would like to derive a $(1-\alpha)\%$ confidence interval of μ.

$$\bar{Y} \sim \text{N}(\mu, \sigma^2/n)$$
$$\bar{Y} - \mu \sim \text{N}(0, \sigma^2/n)$$

Note that $\bar{Y} - \mu$ is a function of data and the parameter of interest. Its distribution is independent of μ. Also note that $Z = \sqrt{n}(\bar{Y} - \mu)/\sigma$ is also a pivot.

Suppose we want a 95% confidence interval.

Use R instead of normal table, we have

```
> qnorm(.25)
[1] -1.959964
```

Let the $(1 - \alpha/2)\%$-th quantile of standard normal is $z_{\alpha/2}$. Then

$$\mathtt{P}(-1.96 \leq Z \leq 1.96) = 95\%.$$

Invert the above result.

$$\begin{aligned} 95\% &= \mathtt{P}\left(-1.96 \leq \frac{\sqrt{n}(\bar{Y} - \mu)}{\sigma} \leq 1.96\right) \\ &= \mathtt{P}(\bar{Y} - 1.96\sqrt{\sigma^2/n} \leq \mu \leq \bar{Y} + 1.96\sqrt{\sigma^2/n}) \end{aligned}$$

And therefore,

$$L = \bar{Y} - 1.96\sqrt{\sigma^2/n}, \quad U = \bar{Y} + 1.96\sqrt{\sigma^2/n}.$$

More generally, how to get a $(1 - \alpha)\%$ confidence interval?

Just replace 1.96 with $z_{1-\alpha/2}$, which is the $(1-\alpha/2)$-th quantile of the standard normal. The $(1 - \alpha)\%$ confidence interval is

$$L = \bar{Y} - z_{\alpha/2}\sqrt{\sigma^2/n}, \quad U = \bar{Y} + z_{\alpha/2}\sqrt{\sigma^2/n}.$$

Remark: If σ is unknown, we can use similar ideas. We can replace σ^2 in the pivot with

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

Then

$$T = \frac{\bar{Y} - \mu}{s^2/n}.$$

T has a student-T distribution with degree of freedom $df = n - 1$. What would be the confidence interval then?

Actually, T distribution is very close to normal distribution when it's degree of freedom is large. Therefore, if n is large, in the above confidence interval, we can use normal too.

Also, when the sample size is large, we can make confidence interval of this form even when the data are not normally distributed.

Rationale. Central Limit Theorem. Suppose $Y_1, \ldots, Y_n$ are *i.i.d.* observations following a distribution with

$$\mathtt{E}(Y_i) = \mu, \quad \mathtt{Var}(Y_i) = \sigma^2.$$

Then as $n \to \infty$,

$$\frac{\sqrt{n}(\bar{Y} - \mu)}{\sigma} \xrightarrow{D.} \mathrm{N}(0, 1).$$

Remark: The notation $\xrightarrow{D.}$ indicates convergence in distribution, sometimes referred to as convergence in law. It means that the cdf of $\sqrt{n}(\bar{Y} - \mu)/\sigma$ converges to the standard normal cdf $\Phi(x)$ for all continuity points x of $\Phi(x)$. It can also be denoted as

$$\bar{Y} \dot{\sim} \mathrm{N}(\mu, \sigma^2/n).$$

Example. Now let's get back to the polling example. We would like to estimate $p =$ the true proportion of Americans who think the economy is the most important issue facing the country today. And also, we would like to derive a 95% confidence interval for p.

We derived $\hat{p} = \sum_{i=1}^{N} Y_i/N$ as the estimator of p.

$$\mathtt{Var}(\hat{p}) = \frac{1}{N} \sum_{i=1}^{N} \mathtt{Var}(Y_i) = \frac{p(1-p)}{N}.$$

Based on the central limit theorem,

$$\frac{\sqrt{n}(\hat{p} - p)}{p(1-p)} \xrightarrow{D.} \mathrm{N}(0, 1).$$

Following the same procedure, we have

$$L = \hat{p} - z_{1-\alpha/2}\sqrt{\frac{p(1-p)}{n}}, \quad U = \hat{p} + z_{1-\alpha/2}\sqrt{\frac{p(1-p)}{n}}.$$

Unfortunately, we don't know p here, and therefore, we cannot use p in the variance expression. However, since p can be estimated with $\hat{p}$, we can substitute them in the above expression:

$$L = \hat{p} - z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \quad U = \hat{p} + z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}.$$

Rationale. Slutsky's Theorem. Consider a sequence of random variable $\mathbf{Y} = (Y_1, \ldots, Y_n)$. As n goes to infinity, if

$$g(\mathbf{Y}, \mu) \xrightarrow{D.} \mathrm{N}(0,1) \text{ and } h(\mathbf{Y}) \xrightarrow{P.} \mu,$$

then

$$g(\mathbf{Y}, h(\mathbf{Y})) \xrightarrow{D.} \mathrm{N}(0,1).$$

Remark: The notation $\xrightarrow{P.}$ denotes convergence in probability. A sequence of random variables $(X_1, \ldots, X_n)$ converges in probability to a constant c, if for every $\epsilon > 0$,

$$\lim_{n\to\infty} \mathtt{P}(|X_n - c| \geq \epsilon) = 0.$$

Now we continue the pooling example. Suppose we are interested in getting a 95% CI.

$$\mathtt{MOE} = 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{N}} = 0.03.$$

The 95% CI is $(\hat{p} - \mathtt{MOE}, \hat{p} + \mathtt{MOE}) = (0.38, 0.44)$.

Now think about definition of confidence intervals. Does it mean with probability 95%, the true rate for choosing economy is within the range $(0.38, 0.44)$?

2.4.2 Variance Stabilization

Now let's revisit the polling example. The variance of the pivot $\hat{p}$ actually involves p. As the estimator changes, the variance will also change. It is easy to see that the variance of the pivot is related to the width of the confidence interval. Therefore, using the pivot method in

the polling example might cause the confidence interval not stable enough (sometimes wide, sometimes narrow). How to solve this problem. One intuitive method is to do transformation, so that we construct a new pivot, $g(\hat{p})$, whose variance is irrelevant of p. But to get the confidence interval, we need the distribution of $g(\hat{p})$. How to get its distribution?

Delta Method. Suppose

$$a_n(W_n - b) \xrightarrow{D.} \mathrm{N}(0, 1),$$

where W_n is a random variable depending on the sequence of random variable $\{Y_n\}$, a_n is a constant dependent on n, and b is a constant. It is easy to see that $\mathtt{E}(W_n) = b$ or $\mathtt{E}(W_n) \xrightarrow{P.} b$. Then for a real valued function $g(\cdot)$ with the first derivative $g'(b) \neq 0$, we have

$$a_n\{g(W_n) - g(b)\} \xrightarrow{D.} \mathrm{N}(0, \{g'(b)\}^2).$$

It can also be reformulated as

$$\frac{a_n\{g(W_n) - g(b)\}}{|g'(b)|} \xrightarrow{D.} \mathrm{N}(0, 1).$$

Now lets get back to the polling example. Let $a_n = \sqrt{n}/\sqrt{p(1-p)}$. Then

$$a_n(\hat{p} - p) \xrightarrow{D.} \mathrm{N}(0, 1).$$

Since $\mathtt{E}(\hat{p}) = p$. Now for a real function g, we have

$$\frac{a_n\{g(\hat{p}) - g(p)\}}{|g'(p)|} = \frac{\sqrt{n}\{g(\hat{p}) - g(p)\}}{\sqrt{p(1-p)}|g'(p)|} \xrightarrow{D.} \mathrm{N}(0, 1).$$

How to choose g? Remember our goal is to make the variance term of $g(\hat{p})$ irrelevant of p. Therefore, we should make

$$|g'(p)| = \{p(1-p)\}^{-1/2}.$$

By integration, $g(p) = 2\arcsin(\sqrt{p})$. Thus

$$2\sqrt{n}\{\arcsin(\sqrt{\hat{p}}) - \arcsin(\sqrt{p})\} \xrightarrow{D.} \mathrm{N}(0, 1).$$

Using this method to construct confidence interval, we have

$$L = \left[\sin\left\{\arcsin(\sqrt{\hat{p}}) - \frac{1.96}{2\sqrt{n}}\right\}\right]^2, \quad U = \left[\sin\left\{\arcsin(\sqrt{\hat{p}}) + \frac{1.96}{2\sqrt{n}}\right\}\right]^2.$$

Based on the polling example, $L = 0.38$, $U = 0.44$, same as the one given by the pivot method. When the sample size is very large, these two methods yield similar results.

But when they are different, how should we choose which one to use? What properties do we want a confidence interval to have?

- Good coverage
- Width (precision)
- Small sample versus large sample properties