W1211 Introduction to Statistics
Lecture 21

Wei Wang

Nov 19, 2012

The Invariance Principle

- One of the nice features of MLE's is that, the MLE of a function of parameters, is the function of the MLE's of the parameters.

- More specifically, we have

> Let $\hat{\theta}_1, \ldots, \hat{\theta}_m$ be the MLE's of the parameters $\theta_1, \ldots, \theta_m$. Then the MLE of any function $h(\theta_1, \ldots, \theta_m)$ of these parameters is $h(\hat{\theta}_1, \ldots, \hat{\theta}_m)$.

<u>Ex.</u> In the normal example, what is the MLE of σ?

Large Sample Behavior

- The following proposition says, for large samples, it is "optimal" to use MLE's, because it is asymptotically unbiased and has the minimal variance among all unbiased estimators.

- Proposition:

  > Under very general conditions on the joint distribution of the sample,
  > When the sample size n is large, the maximum likelihood estimator is
  > Approximately the MVUE of the parameter.

Confidence Intervals

- A point estimate, because it is a single number, by itself provides no information about the precision and reliability of estimation (the reason why we need standard error).

- An alternative to reporting a single sensible value for the parameter being estimated is to calculate and report an entire interval of plausible values – an *interval estimate* or *confidence interval* (*CI*).

- A confidence interval is always calculated by first selecting a *confidence level*, which is a measure of the degree of reliability of the interval.

- Construct a confidence interval for a standard normal random variable.

Illustration

- Let's first consider a simple, somewhat unrealistic problem situation.
  1. We are interested in the population mean parameter μ.
  2. The population distribution is normal.
  3. The value of the population standard deviation σ is known. (unlikely!)

- Suppose we have a random sample $X_1, X_2, \ldots, X_n$ from a normal distribution with mean value μ and standard deviation σ. As we know, $\bar{X}$ also follows a normal distribution with mean value μ and standard deviation $\sigma/\sqrt{n}$. Thus, we could get a standard normal distribution by normalizing $\bar{X}$.

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

Construction

- The smallest interval that contains 95% of the possible outcomes of Z is (-1.96, 1.96).

$$-1.96 < \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} < 1.96$$

$$\Downarrow$$

$$-1.96 \cdot \frac{\sigma}{\sqrt{n}} < \bar{X} - \mu < 1.96 \cdot \frac{\sigma}{\sqrt{n}}$$

$$\Downarrow$$

$$\bar{X} - 1.96 \cdot \frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + 1.96 \cdot \frac{\sigma}{\sqrt{n}}$$

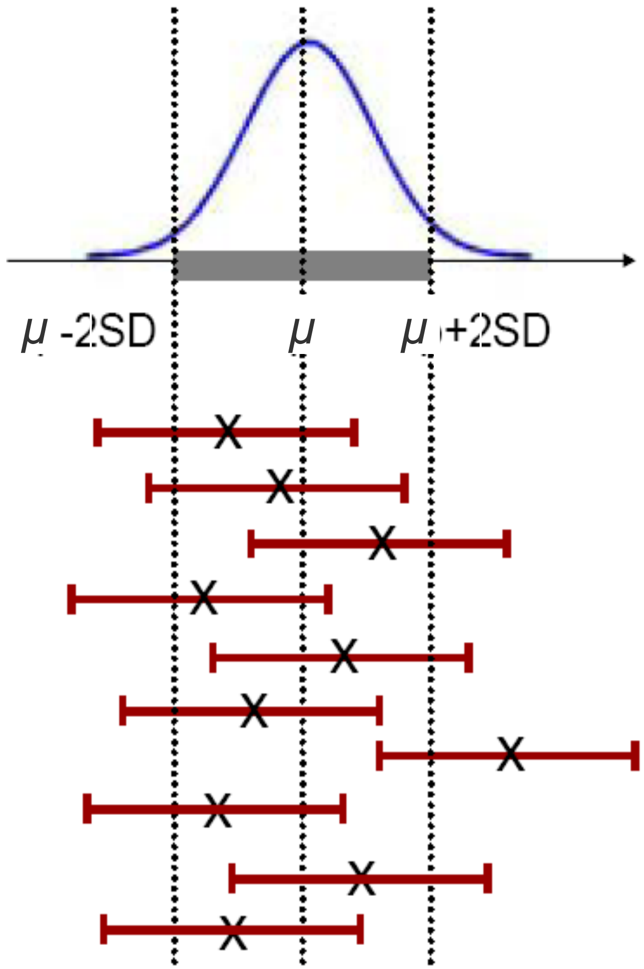
Interpretation

- Thus we have $\mathrm{P}\left(\bar{\mathrm{X}} - 1.96 \cdot \frac{\sigma}{\sqrt{n}} < \mu < \bar{\mathrm{X}} + 1.96 \cdot \frac{\sigma}{\sqrt{n}}\right) = 0.95$.
- Some people interpreted this as: the true parameter *μ* has 95% chance of falling in the interval of $(\bar{\mathrm{X}} - 1.96 \cdot \sigma/\sqrt{n}, \bar{\mathrm{X}} + 1.96 \cdot \sigma/\sqrt{n})$. Is it right?
- In fact, the two boundaries of the interval given above are random! Thus every time we sample n observations from the same population, we will get a different confidence interval!

Random Interval

- By constructing a confidence interval like this, we never be sure whether μ actually lies in our confidence interval. However, we know that about 95 out of 100 times intervals constructed using this method will capture the true parameter.

- Interpreted as: "*the probability is .95 that the random interval includes or covers the true value of μ.*"



Confidence Interval for the Mean of a Normal Population when Variance is assumed known

- A $100(1-\alpha)\%$ confidence interval for the mean μ of a normal population when the value of σ is known is given by
$$\left(\bar{x} - z_{\alpha/2}\cdot\frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\alpha/2}\cdot\frac{\sigma}{\sqrt{n}}\right)$$
- $z_{\alpha/2}$ is the upper $(100\cdot\alpha/2)\%$ percentile of a standard normal distribution, i.e., $P(Z > z_{\alpha/2}) = \alpha/2$.
- z_{α}'s are usually refered to as z critical values.

Remarks

- When constructing a confidence interval, *confidence level*, *precision*, and *sample size* are closely related. Is there a finite 100% confidence interval?

- The precision, or the width of the confidence interval when σ is known is, $2z_{\alpha/2}\sigma/\sqrt{n}$. Thus we can see, the confidence level of the interval is inversely related to its precision.

- The precision is also inversely related to the sample size.

- An appealing strategy is to specify both the desired confidence level and interval width and then determine the necessary sample size.

Sample Size Calculation

- The general formula for the sample size n necessary to ensure an interval width w is obtained from $w = 2 \cdot z_{\alpha/2} \cdot \sigma/\sqrt{n}$.

$$n = \left(2 \cdot z_{\alpha/2} \cdot \frac{\sigma}{w}\right)^2$$

- Ex. A new operating system has been installed, and we wish to estimate the true average response time μ to a particular editing command. Assuming that response times are normally distributed with $\sigma = 25$ millisec. How many tests should we do to ensure that the resulting 95% CI has a width of at most 10?

Sample Size Calculation

- The general formula for the sample size n necessary to ensure an interval width w is obtained from $w = 2 \cdot z_{\alpha/2} \cdot \sigma/\sqrt{n}$.

$$n = \left(2 \cdot z_{\alpha/2} \cdot \frac{\sigma}{w}\right)^2$$

- Ex. A new operating system has been installed, and we wish to estimate the true average response time μ to a particular editing command. Assuming that response times are normally distributed with $\sigma = 25$ millisec. How many tests should we do to ensure that the resulting 95% CI has a width of at most 10?
- Plug in into the formula

$$n = (2 \cdot 1.96 \cdot \frac{25}{10})^2 = 96.04$$

So we need at least 97 tests.

Constructing a CI

- The previous examples show the general procedure of constructing confidence intervals. Suppose X_1, X_2, …, X_n are the sample on which the CI for a parameter θ is to be based. Then we construct a so-called "pivotal" quantity whose distribution does not depend on parameters.

- In other words, the pivotal quantity is a function of both samples and parameters, i.e., $h(X_1, X_2, \ldots, X_n, \theta)$, and the distribution of $h(\cdot)$ does not depend on θ or any other unknowns.

- Then one can find *a* and *b* to satisfy $\mathrm{P}(a < h(\mathrm{X}_1, \mathrm{X}_2, \ldots, \mathrm{X}_n; \theta) < b) = 1 - \alpha$, by the pivotal property, *a* and *b* do not depend on θ. Then the inequality can be manipulated to isolate θ, giving the equivalent probability statement
$$\mathrm{P}(l(\mathrm{X}_1, \mathrm{X}_2, \ldots, \mathrm{X}_n) < \theta < u(\mathrm{X}_1, \mathrm{X}_2, \ldots, \mathrm{X}_n)) = 1 - \alpha$$

Large-Sample Confidence Intervals for a Population Mean and Proportion

- However, in most cases, it is impossible to locate a pivotal quantity. In the previous setting, we can do this because the unlikely assumption of knowing σ.
- We often need to resort to large-sample theory, namely Central Limit Theorem to construct CIs.
- The most common application is to construct CIs for a Population Mean and Proportion.

Key Results

- If $X_1, X_2, \ldots, X_n$ IID from a general distribution with mean μ and variance σ^2, then CLT tells us

$$\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$$

or

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$$

Key Results

- If $X_1, X_2, \ldots, X_n$ IID from a general distribution with mean μ and variance σ^2, then CLT tells us

$$\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$$

  or

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$$

- Further, if we substitute σ with its estimator $\hat{\sigma}$, this still holds

$$\frac{\bar{X} - \mu}{\hat{\sigma}/\sqrt{n}} \sim N(0, 1)$$

General Results

- Proposition:

  A 100(1-α)% confidence interval for the mean μ of any population when the value of σ is unknown and sample size n is sufficiently large is given by

  $$\left(\bar{x} - z_{\alpha/2} \cdot \frac{\hat{\sigma}}{\sqrt{n}}, \bar{x} + z_{\alpha/2} \cdot \frac{\hat{\sigma}}{\sqrt{n}}\right)$$

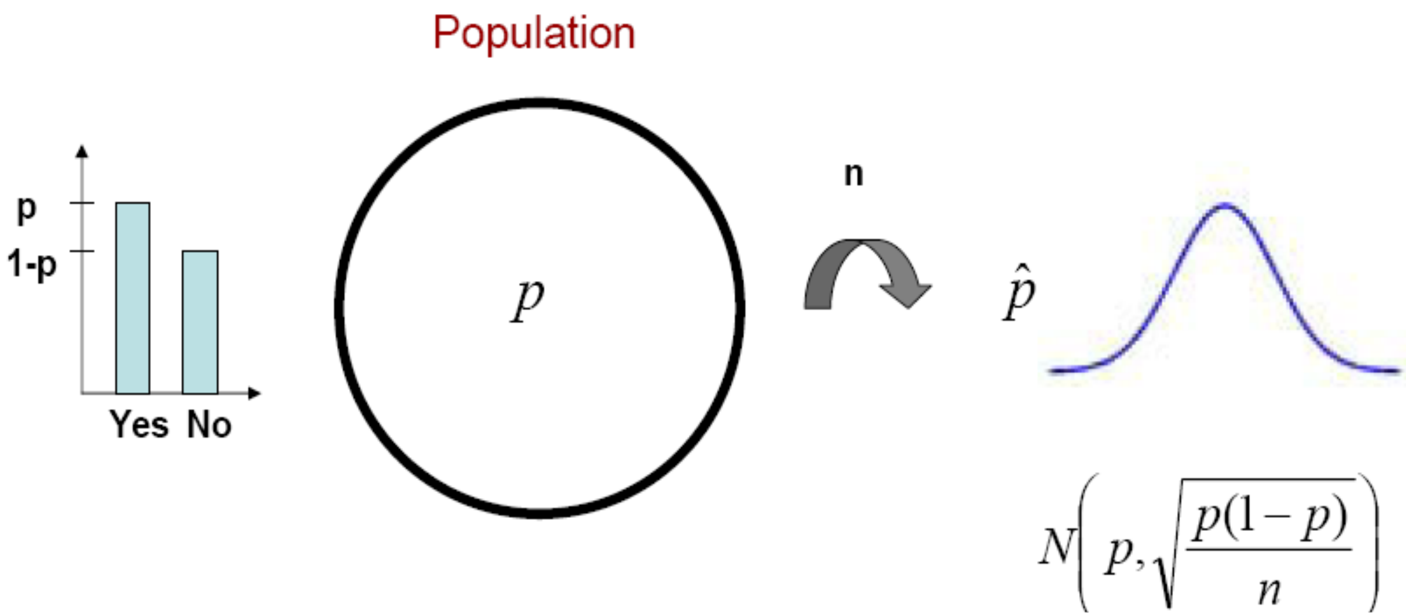
- Rule of Thumb: generally speaking, n>40 will be sufficient to justify the use of this interval. This is somewhat more conservative than the rule of thumb for the CLT, because of the additional randomness coming from $\hat{\sigma}$.

- One can also derive a similar sample size calculation formula in this case

  $$n = \left(2 \cdot z_{\alpha/2} \cdot \frac{\hat{\sigma}}{w}\right)^2$$

Proportions

- A special case of non-normal population is Bernoulli population. And the parameter of interest is the population proportion *p*.



Large Sample CI

- One can directly apply the proposition from the large sample case to construct the CI for the population proportion *p*.

$$\left(\bar{x} - z_{\alpha/2} \cdot \frac{\hat{\sigma}}{\sqrt{n}}, \bar{x} + z_{\alpha/2} \cdot \frac{\hat{\sigma}}{\sqrt{n}}\right)$$

- In this case $\bar{x} = \hat{p}$, $\hat{\sigma}^2 = \hat{p}(1 - \hat{p})$.

- If we set *q*=1-*p*, then the large sample confidence interval for *p* should be

$$\left(\hat{p} - z_{\alpha/2}\sqrt{\hat{p}\hat{q}/n}, \hat{p} + z_{\alpha/2}\sqrt{\hat{p}\hat{q}/n}\right)$$

- To calculate sample size: $n = \left(2 \cdot z_{\alpha/2} \cdot \frac{\sqrt{\hat{p}\hat{q}}}{w}\right)^2$

Another way

- The large sample confidence interval works fine if we have enough data. But for finite samples we can construct a better CI.

- Since in this case, we only have 1 parameter *p*, by CLT, we have

$$\mathrm{P}\left(-z_{\alpha/2} < \frac{\hat{p}-p}{\sqrt{p(1-p)/n}} < z_{\alpha/2}\right) \approx 1-\alpha$$

- If we solve the resulting quadratic function, we'll have a new confidence interval for *p*.

$$\left(\frac{\hat{p}+\frac{z_{\alpha/2}^2}{2n}-z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}+\frac{z_{\alpha/2}^2}{4n^2}}}{1+z_{\alpha/2}^2/n}, \frac{\hat{p}+\frac{z_{\alpha/2}^2}{2n}+z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}+\frac{z_{\alpha/2}^2}{4n^2}}}{1+z_{\alpha/2}^2/n}\right)$$

Remarks

- The latter confidence interval looks complicated, but it "can be recommended for use with nearly all sample sizes and parameter values". Therefore we don't have to check for large sample conditions.

- In the latter case, we can also derive a new sample size calculation formula

$$n = \frac{2z_{\alpha/2}^2 \hat{p}\hat{q} - z_{\alpha/2}^2 w^2 \pm \sqrt{4z_{\alpha/2}^4 \hat{p}\hat{q}(\hat{p}\hat{q} - w^2) + w^2 z_{\alpha/2}^4}}{w^2}$$

  "+" sign is used!

- When sample size is large, the confidence interval we just constructed and the sample size calculation formula will be equivalent to

$$\left(\hat{p} - z_{\alpha/2}\sqrt{\hat{p}\hat{q}/n}, \hat{p} + z_{\alpha/2}\sqrt{\hat{p}\hat{q}/n}\right) \quad \text{and} \quad n = \left(2 \cdot z_{\alpha/2} \cdot \frac{\sqrt{\hat{p}\hat{q}}}{w}\right)^2$$

One-sided CI

- In some situations, an investigator will want only one upper bound or one lower bound for the parameter.

- Follow a similar argument as in the two-sided case, we have the following result

A large sample 100(1-α)% confidence upper bound for the mean μ is

$$\mu < \bar{x} + z_\alpha \cdot \frac{\hat{\sigma}}{\sqrt{n}}$$

and a lower bound is

$$\mu > \bar{x} - z_\alpha \cdot \frac{\hat{\sigma}}{\sqrt{n}}$$

A one-sided confidence bound for p results from replacing $z_{\alpha/2}$ by z_α.