

S1211Q Introduction to Statistics

Lecture 14

Wei Wang

July 24, 2012

Estimator, Its Standard Error and Estimated Standard Error

- ▶ Now we are trying to estimate the probability of getting heads of a biased coin, so each flip X_i is a Bernoulli RV with parameter p , the estimator of parameter p is the sample mean/proportion

$$\hat{p} = \frac{\sum_{i=1}^n X_i}{n}$$

Estimator, Its Standard Error and Estimated Standard Error

- ▶ Now we are trying to estimate the probability of getting heads of a biased coin, so each flip X_i is a Bernoulli RV with parameter p , the estimator of parameter p is the sample mean/proportion

$$\hat{p} = \frac{\sum_{i=1}^n X_i}{n}$$

- ▶ If we flipped 100 times and observed 75 heads, then our estimate of p is

$$\hat{p} = \frac{75}{100} = 0.75$$

Estimator, Its Standard Error and Estimated Standard Error

- ▶ Now we are trying to estimate the probability of getting heads of a biased coin, so each flip X_i is a Bernoulli RV with parameter p , the estimator of parameter p is the sample mean/proportion

$$\hat{p} = \frac{\sum_{i=1}^n X_i}{n}$$

- ▶ If we flipped 100 times and observed 75 heads, then our estimate of p is

$$\hat{p} = \frac{75}{100} = 0.75$$

- ▶ Also, we need to report how good our estimator is through its Standard Error. This is also related to the Interval Estimation.

Estimator, Its Standard Error and Estimated Standard Error

- ▶ The standard error is $Var(\hat{p}) = \frac{p(1-p)}{n}$, but we cannot report it since we don't know what p is.
- ▶ So we can only report the estimated standard error of the estimator \hat{p}

$$\widehat{Var}(\hat{p}) = \frac{\hat{p}(1 - \hat{p})}{n}$$

Another Example

- ▶ Now we have X_1, X_2, \dots, X_N IID with mean μ and variance σ^2 , what's the estimator of μ ?

Another Example

- ▶ Now we have X_1, X_2, \dots, X_N IID with mean μ and variance σ^2 , what's the estimator of μ ?

▶

$$\hat{\mu} = \bar{X}$$

Another Example

- ▶ Now we have X_1, X_2, \dots, X_N IID with mean μ and variance σ^2 , what's the estimator of μ ?

▶

$$\hat{\mu} = \bar{X}$$

- ▶ The standard error of $\hat{\mu}$ is $\sqrt{\text{Var}(\hat{\mu})} = \frac{\sigma}{\sqrt{n}}$. Can we report $\frac{\sigma}{\sqrt{n}}$?

Another Example

- ▶ Now we have X_1, X_2, \dots, X_N IID with mean μ and variance σ^2 , what's the estimator of μ ?

▶

$$\hat{\mu} = \bar{X}$$

- ▶ The standard error of $\hat{\mu}$ is $\sqrt{\text{Var}(\hat{\mu})} = \frac{\sigma}{\sqrt{n}}$. Can we report $\frac{\sigma}{\sqrt{n}}$?
- ▶ It really depends on whether or not we know σ . If we know it, then we can report $\frac{\sigma}{\sqrt{n}}$; otherwise, we can only report $\frac{\hat{\sigma}}{\sqrt{n}}$.

Methods of Point Estimation

- The definition of unbiasedness does not in general indicate how unbiased estimators can be derived.
- There are two commonly used “constructive” methods for obtaining point estimators: the [method of moments](#) and the [method of maximum likelihood](#).
- Although maximum likelihood estimators are generally preferable to moment estimators because of certain efficiency properties, they often require significantly more computation than do moment estimators.
- It is **NOT** guaranteed that these two methods would yield unbiased estimators.

Population Moment and Sample Moment

- ▶ Let X_1, \dots, X_n be a random sample from a pmf or pdf $f(x)$. For $k = 1, 2, \dots$, the k th population moment is $E(X^k)$. The k th sample moment is $(1/n) \sum_{i=1}^n X_i^k$.
- ▶ The essence of the Methods of Moment is to equate population moments with sample moments and solve the resulting equations.

Moment Estimators

- Definition:

Let X_1, X_2, \dots, X_n be an i.i.d. sample from a pmf or pdf $f(x)$. For $k = 1, 2, 3, \dots$, the moment estimator for the k th population moment, is the k th sample moment, i.e.,

$$\widehat{E(X^k)} = \frac{\sum_{i=1}^n X_i^k}{n}$$

Remarks

- Notice that, the sample k th moment will **converge** to the population k th moment, as sample size $n \rightarrow \infty$, thanks to the **LLN**.
- For any finite sample size n , the sample k th moment in general will **not** be **equal** to the population k th moment. But it is a good candidate for estimation.
- The larger the n , the better the estimation is!
- In practice, if the underlying pmf or pdf has m parameters, namely, we have $f(x; \theta_1, \dots, \theta_m)$, where $\theta_1, \dots, \theta_m$ are parameters whose values are unknown. Then the moment estimators are obtained by **equating the first m sample moments to the corresponding first m population moments and solving for $\theta_1, \dots, \theta_m$** .

Example

Ex. Show that the sample proportion is the moment estimator of the population probability.

Example

Ex. Let X_1, X_2, \dots, X_n be an i.i.d. normal sample, and assume that the underlying normal distribution is $N(\mu, \sigma^2)$ where μ, σ^2 are unknown. How can we construct moment estimators to estimate the two unknown parameters?

As we already know if $X \sim N(\mu, \sigma^2)$, then $E(X) = \mu$, and $E(X^2) = \mu^2 + \sigma^2$.

Therefore, we have two equations:

$$\begin{cases} \hat{\mu} = \sum_{i=1}^n X_i / n \\ \hat{\mu}^2 + \hat{\sigma}^2 = \sum_{i=1}^n X_i^2 / n \end{cases} \longrightarrow \begin{cases} \hat{\mu} = \sum_{i=1}^n X_i / n \\ \hat{\sigma}^2 = \sum_{i=1}^n X_i^2 / n - \bar{X}^2 \end{cases}$$

Is the variance estimator unbiased?

Example

Ex. Let X_1, X_2, \dots, X_n be an i.i.d. sample from exponential distribution with parameter λ which is unknown. How do we estimate λ using moment estimator?

As we already know if $X \sim \text{Exp}(\lambda)$, then $E(X) = 1/\lambda$.

Thus, we have equation $1/\hat{\lambda} = \bar{X} \rightarrow \hat{\lambda} = 1/\bar{X}$.

Is this estimator unbiased?

Maximum Likelihood Est.

- The method of **maximum likelihood** was first introduced by **R.A. Fisher**, a geneticist and statistician, in the 1920s. It is by far the most commonly used method to obtain estimators.
- **Likelihood function** is just another way of looking at the *joint pmf or the pdf*. In particular, let X_1, X_2, \dots, X_n (not necessarily i.i.d.) have joint pmf or pdf

$$f(x_1, x_2, \dots, x_n; \theta_1, \dots, \theta_m)$$

where $\theta_1, \dots, \theta_m$ are parameters whose values are unknown. When x_1, x_2, \dots, x_n are the observed sample values and $f(\cdot)$ is then regarded as a function of $\theta_1, \dots, \theta_m$, it is called the **likelihood function**.

Example

Ex. A biased coin has been flipped for 10 times. Let X_1, X_2, \dots, X_{10} denote the outcomes of the coin flips. Assume the probability of having a head is p (parameter of interest), and the sample we observed is $\{0, 1, 1, 0, 0, 0, 1, 0, 0, 0\}$. Write down the likelihood function for p .

$$f(x_1, x_2, \dots, x_n; p) = f(x_1; p) f(x_2; p) \dots f(x_n; p) = (1-p) p p (1-p) \dots (1-p) = p^3(1-p)^7$$

Idea of **Maximum Likelihood**: can we find a p that can **maximize** the above function?

MLE

- The **maximum likelihood estimates** (mle's) $\hat{\theta}_1, \dots, \hat{\theta}_m$ are those values of θ_i 's that maximize the likelihood function, so that

$$f(x_1, \dots, x_n; \hat{\theta}_1, \dots, \hat{\theta}_m) \geq f(x_1, \dots, x_n; \theta_1, \dots, \theta_m) \quad \text{for all } \theta_1, \dots, \theta_m$$

when the X_i 's are substituted in place of the x_i 's.

- **Remark:** the likelihood function tells us how likely the observed sample is as a function of the possible parameter values. Maximizing the likelihood gives the parameter values for which **the observed sample is most likely to have been generated** – that is, the parameter values that “**agree most closely**” with the observed data.
- In practice, in stead of maximizing the likelihood itself, people usually choose to maximize the **log-likelihood function**.

Example

Ex. Let X_1, X_2, \dots, X_n be an i.i.d. sample from exponential distribution with parameter λ which is unknown. Write down the likelihood function for λ . What is the MLE of λ ? Is the MLE unbiased?

Since we have an i.i.d. sample, it is easy to see that the likelihood function is a product of the individual pdf's:

$$f(x_1, \dots, x_n; \lambda) = (\lambda e^{-\lambda x_1}) \dots (\lambda e^{-\lambda x_n}) = \lambda^n e^{-\lambda \sum x_i}$$



$$\log[f(x_1, \dots, x_n; \lambda)] = n \log(\lambda) - \lambda \sum x_i$$



$$\hat{\lambda} = n / \sum X_i$$

Example

Ex. Let X_1, X_2, \dots, X_n be an i.i.d. sample from normal distribution with parameter μ and σ^2 which is unknown. Write down the likelihood function. What are the MLE's of the two parameters? Are they unbiased?


Some complications

- The following is an example that MLE's can't be calculated analytically.

Ex. Let X_1, X_2, \dots, X_n be an i.i.d. sample from Weibull distribution with parameters α and β and pdf

$$f(x; \alpha, \beta) = \begin{cases} \frac{\alpha}{\beta^\alpha} \cdot x^{\alpha-1} \cdot e^{-(x/\beta)^\alpha} & x \geq 0, \\ 0 & \text{otherwise.} \end{cases}$$

by solving equations $\frac{\partial \log(f)}{\partial \alpha} = 0 \quad \frac{\partial \log(f)}{\partial \beta} = 0$


$$\hat{\alpha} = \left[\frac{\sum x_i^{\hat{\alpha}} \cdot \log(x_i)}{\sum x_i^{\hat{\alpha}}} - \frac{\sum \log(x_i)}{n} \right]^{-1} \quad \hat{\beta} = \left(\frac{\sum x_i^{\hat{\alpha}}}{n} \right)^{1/\hat{\alpha}}$$

Some Complications

- ▶ Also, sometimes we cannot use calculus to get the MLE, such as when the density is not differentiable.
- ▶ Read Example 6.22 on textbook P.262.

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

Ex. 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, \dots, 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 σ/\sqrt{n} . Thus, we could get a standard normal distribution by normalizing \bar{X} .

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

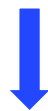
Construction

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

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



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



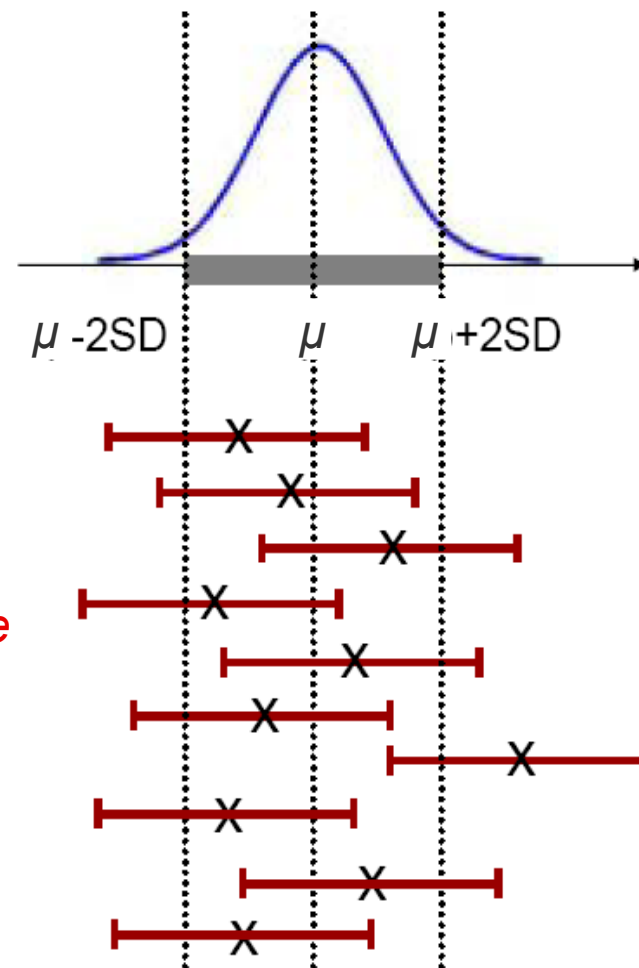
$$\bar{X} - 1.96 \cdot \frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + 1.96 \cdot \frac{\sigma}{\sqrt{n}}$$

Interpretation

- Thus we have $P\left(\bar{X} - 1.96 \cdot \frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + 1.96 \cdot \frac{\sigma}{\sqrt{n}}\right) = 0.95$.
- Some people interpreted this as: the true parameter μ has 95% chance of falling in the interval of $(\bar{X} - 1.96 \cdot \sigma/\sqrt{n}, \bar{X} + 1.96 \cdot \sigma/\sqrt{n})$. 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

- Definition:

A 100(1- α)% 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 $\alpha/2$ quantile of a standard normal distribution, i.e., $P(Z > z_{\alpha/2}) = \alpha/2$.

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?

Non-normal and Unknown Variance

- Previously we constructed a confidence interval for normal population mean with known variance. The next question would then be, what if we don't have normality and what if we don't know the underlying variance?
- If we have large enough sample size, the celebrated **CLT** can help us construct a confidence interval for the mean parameter of a population with unknown distribution and unknown variance. Consider the following quantity

$$\frac{\bar{X} - \mu}{\hat{\sigma}/\sqrt{n}} = \underbrace{\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}}_{\text{CLT} \rightarrow N(0,1)} \cdot \underbrace{\frac{\sigma}{\hat{\sigma}}}_{\text{LLN} \rightarrow 1}$$

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

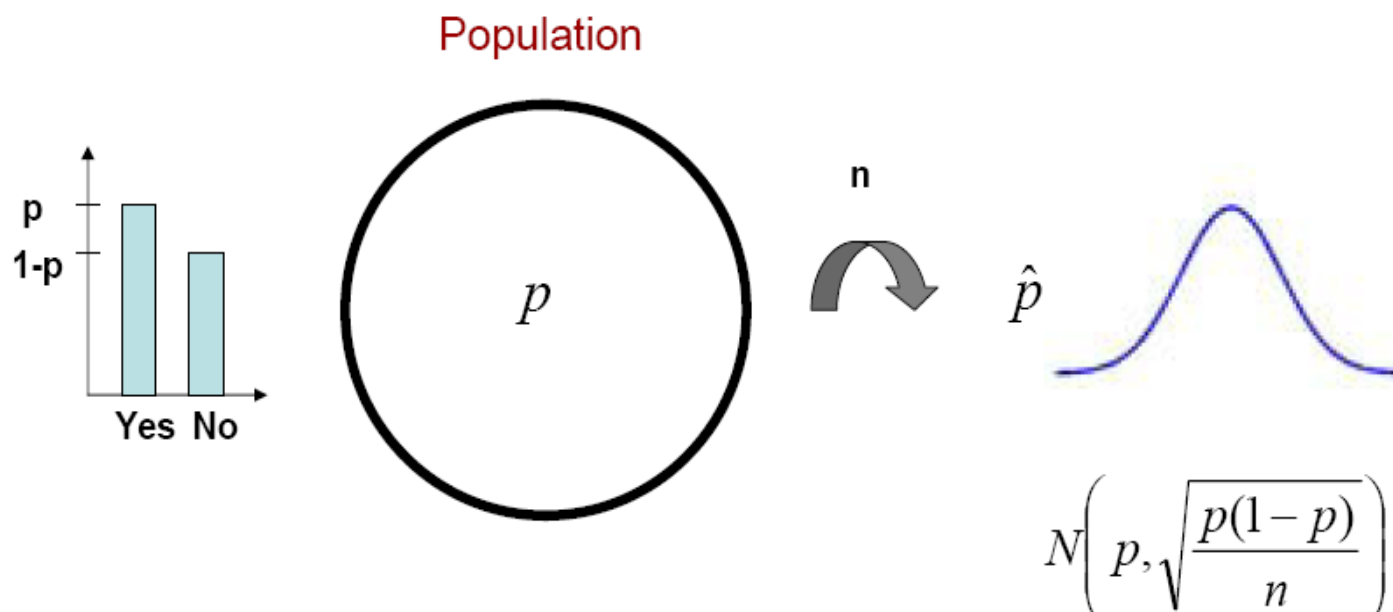
$$\left(\bar{x} - z_{\alpha/2} \cdot \frac{\hat{\sigma}}{\sqrt{n}}, \bar{x} + z_{\alpha/2} \cdot \frac{\hat{\sigma}}{\sqrt{n}} \right)$$

- **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

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

- 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 $100(1-\alpha)\%$ 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} .

Constructing a CI

- The previous examples show the general procedure of constructing confidence intervals. Suppose X_1, X_2, \dots, 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., $h(X_1, X_2, \dots, X_n, \theta)$, and the distribution of $h(\cdot)$ does not depend on θ or any other unknowns.
- Then one can find a and b to satisfy $P(a < h(X_1, X_2, \dots, X_n; \theta) < b) = 1 - \alpha$, 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

$$P(l(X_1, X_2, \dots, X_n) < \theta < u(X_1, X_2, \dots, X_n)) = 1 - \alpha$$