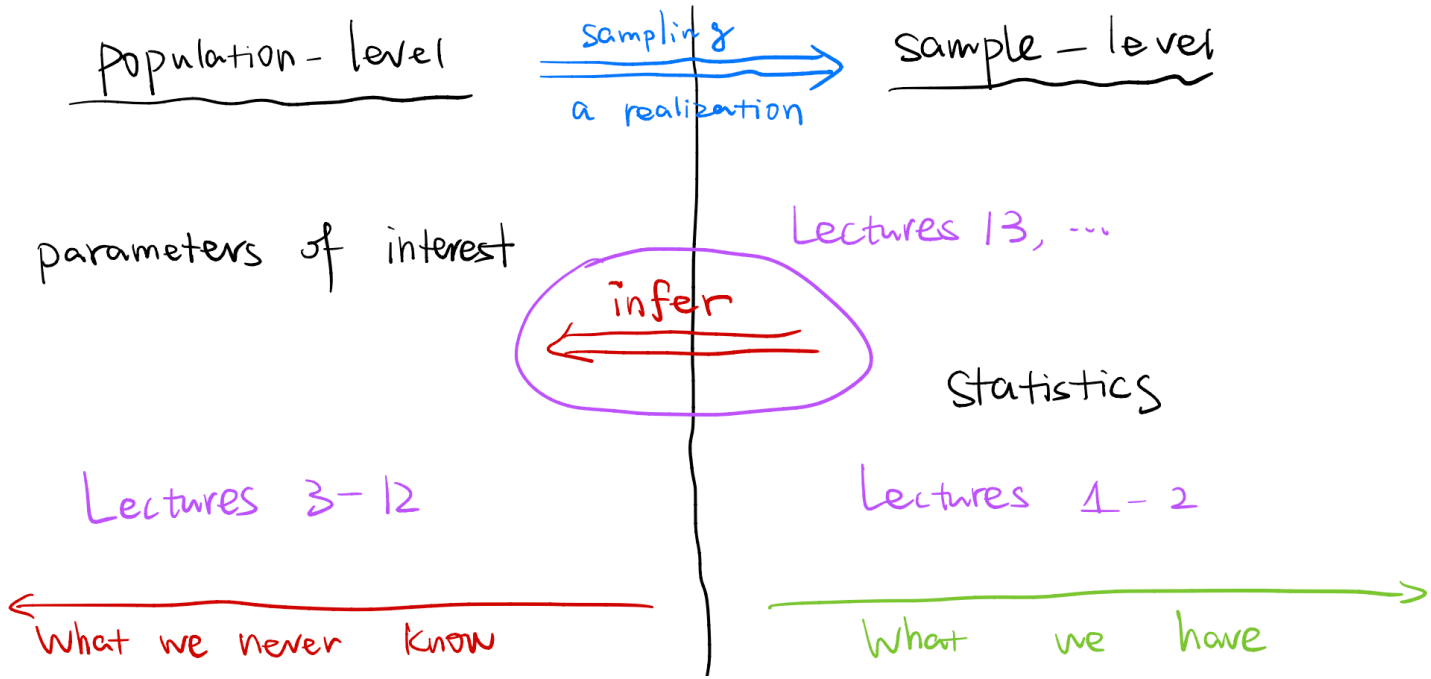
Lecture 12: Confidence Intervals (I)

Ye Tian

Department of Statistics, Columbia University
Calculus-based Introduction to Statistics (S1201)

July 28, 2022

Recap: Some concepts



- **Goal of statistics:** To make inferences on **parameters** of a **population**, which are assumed to be **fixed but unknown**.
  Examples: population proportion, mean, standard deviation, median etc.
- **Statistic:** A function of samples $X_1, \ldots, X_n$, which is calculable given the samples. It does NOT depend on unknown parameters.
- **Estimator:** A statistic used to estimate the parameter.
- **Estimate:** A numerical value of the estimator.

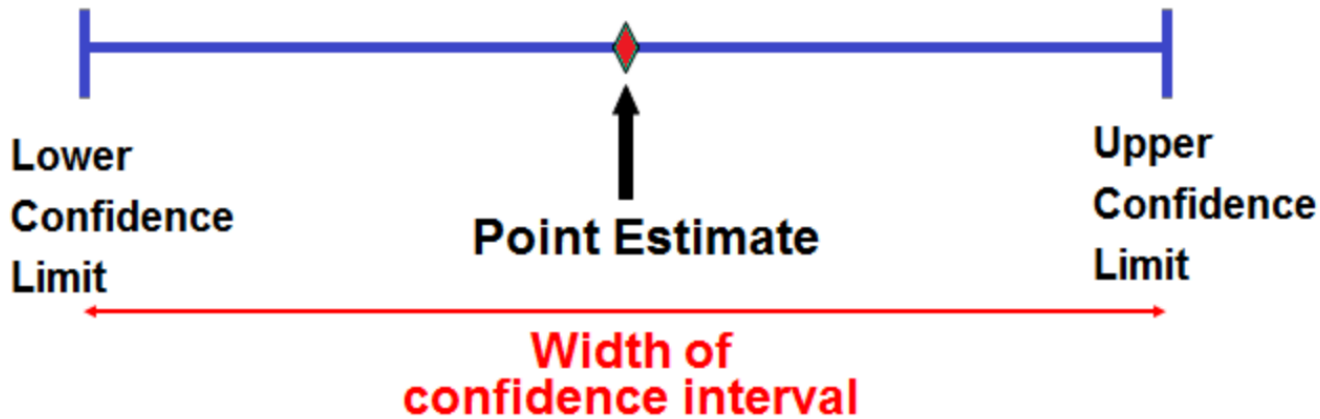
Two types of estimation

Two types of estimation:

- **Point estimation:** use a **single value** as the estimate of the parameter
- **Confidence interval:** point out a **range (interval)** which is very likely to cover the parameter

An example: If we want to estimate the acceptance rate of Columbia next year...

- Mike: It might be 7%. $\longrightarrow$ a point estimation
- Lee: It could be between 6% and 8%. $\longrightarrow$ a confidence interval



Today's goal

- Understand confidence intervals and know how to construct them

Confidence Interval Intuitions

A previous example

10 individuals with Columbia email accounts are selected, and the first, third, and tenth individuals are using a strong password, whereas the others do not. If for each randomly sampled person, the probability that they use a strong password is p. Estimate p.

$X_i = \mathbb{1}(i\text{-th individual uses a strong password})$, and $\hat{p}_{\text{MLE}} = \hat{p}_{\text{Moment}} = \bar{X}$. It's good...but for different samples, we have different $\bar{X}$ since it's a random variable. Is it possible to have $\bar{X} = p$ every time?

Not at all. Actually the probability would be ZERO if $p \neq 0, \frac{1}{n}, \frac{2}{n}, \ldots, \frac{n-1}{n}, 1$ (why?)

Point estimation is good, but...

Point estimation:

- It's a procedure that only needs the input of samples, and it will output a single value as the estimate of the parameter.
- Provides no information about the precision and reliability of estimation, although sometimes we know it will be close to the parameter value when n is very large (e.g.: $\bar{X}$ is close to $\theta = \mathbb{E}X$ by Law of Large Number!)

Now we want to find **another procedure**, which

- also only needs the input of samples
- outputs an interval as a range estimate of the parameter, which **in a long run** can cover the true parameter value

The first example: estimating μ in $N(\mu, 1)$

Suppose we have samples $X_1, \ldots, X_n \overset{i.i.d.}{\sim} N(\mu, 1)$, where the mean parameter μ is **unknown**. We want to estimate μ.

Consider the following procedure: We know $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i \sim N(\mu, 1/n)$. Then by standardization, $\frac{\bar{X}-\mu}{\sqrt{1/n}} \sim N(0,1)$. Thus,

$$\mathbb{P}\left(z_{0.975} \leq \frac{\bar{X}-\mu}{\sqrt{1/n}} \leq z_{0.025} \right) = 95\%,$$

where z_α is the $(100(1-\alpha))\%$ quantile of $N(0,1)$, i.e. $\Phi(z_\alpha) = 1-\alpha$. Then

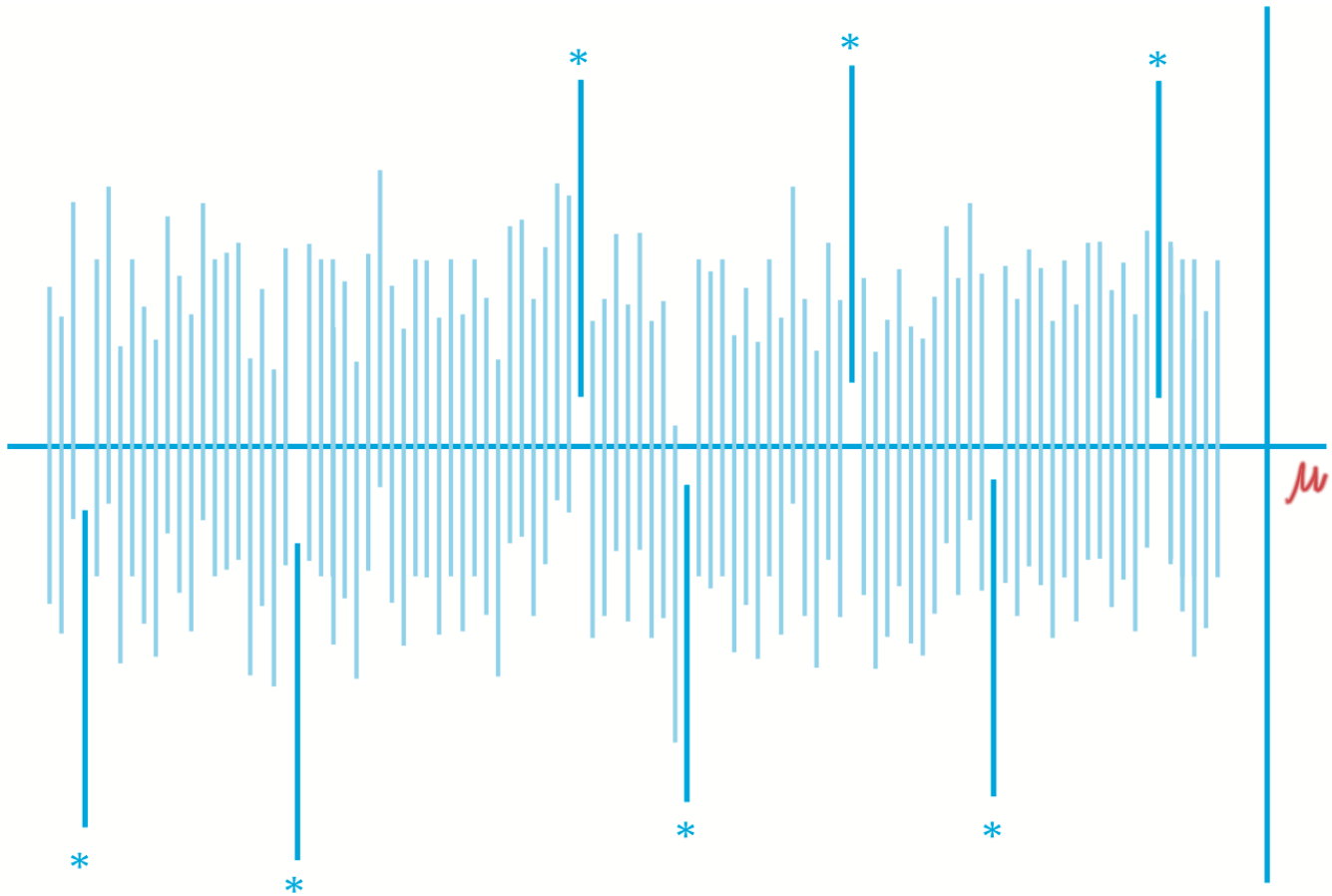
$$\mathbb{P}\left(\bar{X} - \frac{z_{0.025}}{\sqrt{n}} \leq \mu \leq \bar{X} - \frac{z_{0.975}}{\sqrt{n}} \right) = 95\%.$$

We get a **random** interval $\left[\bar{X} - \frac{z_{0.025}}{\sqrt{n}}, \bar{X} + \frac{z_{0.025}}{\sqrt{n}}\right]$ which can cover μ with 95% probability!

How to understand this 95% coverage probability

$$\mathbb{P}\left(\bar{X} - \frac{z_{0.025}}{\sqrt{n}} \leq \mu \leq \bar{X} + \frac{z_{0.025}}{\sqrt{n}}\right) = 95\%.$$

We get a **random** interval $\left[\bar{X} - \frac{z_{0.025}}{\sqrt{n}}, \bar{X} + \frac{z_{0.025}}{\sqrt{n}}\right]$ which can cover μ with 95% probability!



*: The "realization" of the random interval that does NOT cover μ

More intuitions

$$\mathbb{P}\left(\bar{X} - \frac{z_{0.025}}{\sqrt{n}} \leq \mu \leq \bar{X} + \frac{z_{0.025}}{\sqrt{n}}\right) = 95\%.$$

Recall that we want to find **a new procedure**, which

- also only needs the input of samples ✓
- outputs an interval as a range estimate of the parameter, which **in a long run** can cover the true parameter value
  ✓: The **random** interval $\left[\bar{X} - \frac{z_{0.025}}{\sqrt{n}}, \bar{X} + \frac{z_{0.025}}{\sqrt{n}}\right]$ covers μ with 95% probability!

Question: Is such random interval unique?

- What about $\left[\bar{X} - \frac{z_{0.05}}{\sqrt{n}}, \bar{X}\right]$?
- What about $\left[\bar{X} - \frac{z_{0.015}}{\sqrt{n}}, \bar{X} + \frac{z_{0.035}}{\sqrt{n}}\right]$?
- What about $\left[\bar{X} - \frac{z_{\alpha}}{\sqrt{n}}, \bar{X} + \frac{z_{0.05-\alpha}}{\sqrt{n}}\right]$, where $0 \leq \alpha \leq 0.025$?

Confidence Interval Basics

Confidence intervals

Definition: Suppose we have samples $X_1, \ldots, X_n$. If we can construct an random interval $[l(X_1, \ldots, X_n), u(X_1, \ldots, X_n)]$ which satisfies

$$\mathbb{P}(l(X_1, \ldots, X_n) \leq \theta \leq u(X_1, \ldots, X_n)) = 1 - \alpha,$$

then it can be called as a $(100(1-\alpha))\%$ **confidence interval (CI)**. $(100(1-\alpha))\%$ is the **confidence level**.

Example: Suppose we have samples $X_1, \ldots, X_n \overset{i.i.d.}{\sim} N(\mu, 1)$, where the mean parameter μ is **unknown**. All of the followings are 95% CI of μ.

- $\left[\bar{X} - \frac{z_{0.05}}{\sqrt{n}}, \bar{X}\right]$
- $\left[\bar{X} - \frac{z_{0.015}}{\sqrt{n}}, \bar{X} - \frac{z_{0.035}}{\sqrt{n}}\right]$
- $\left[\bar{X} - \frac{z_{0.025}}{\sqrt{n}}, \bar{X} + \frac{z_{0.025}}{\sqrt{n}}\right]$

Which one do you prefer?

People usually prefer "splitting the confidence level". It can also be proved that under the setting of normal distribution, the width of the third CI above is the shortest among all 95% CIs.

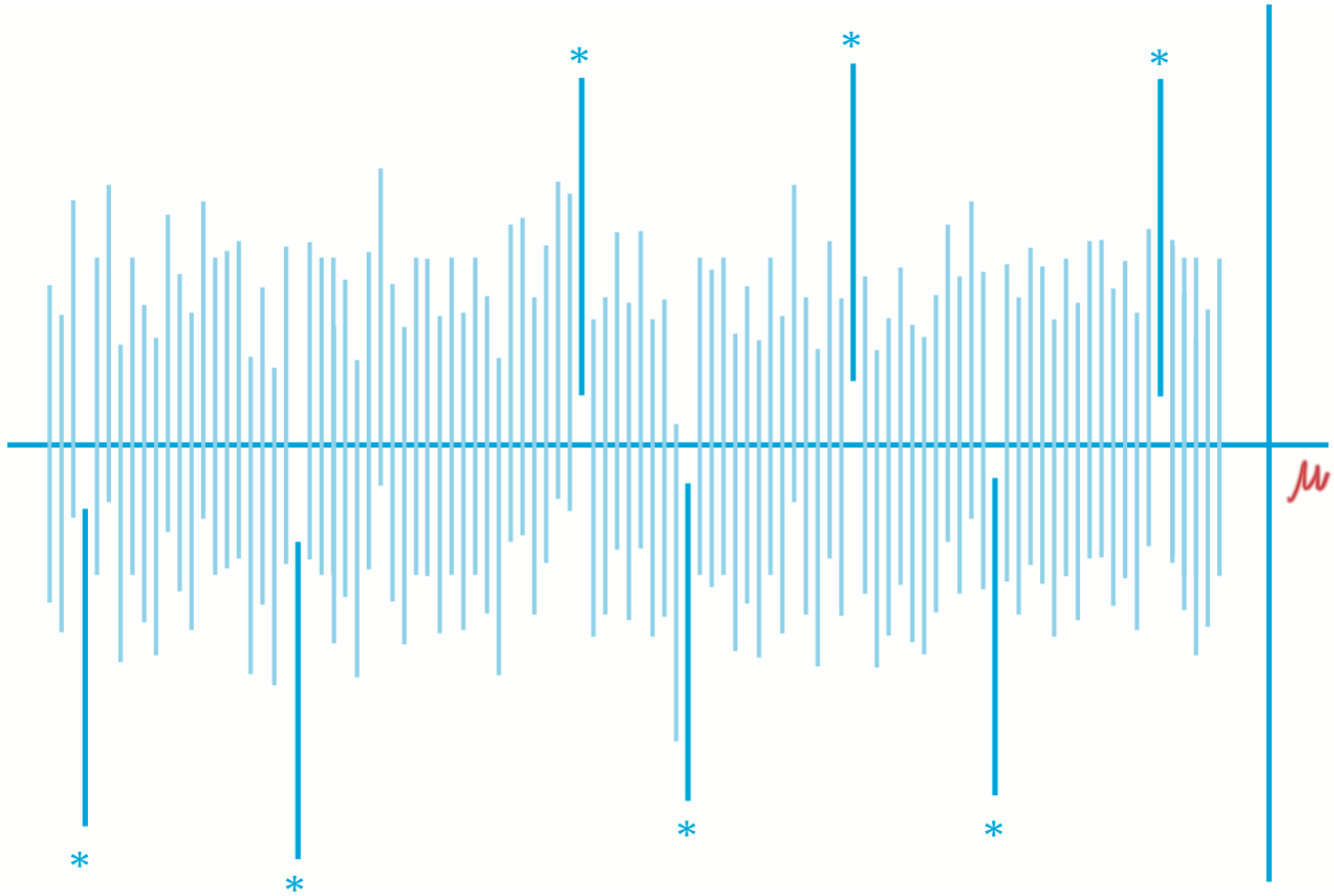
Interpretations

Suppose we have samples $X_1, \ldots, X_n \overset{i.i.d.}{\sim} N(\mu, 1)$, where the mean parameter μ is **unknown**. In our observations, $\bar{x} = 5$ and $n = 10$. We want to estimate μ.

$$\mathbb{P}\left(\bar{X} - \frac{z_{0.025}}{\sqrt{n}} \leq \mu \leq \bar{X} + \frac{z_{0.025}}{\sqrt{n}}\right) = 95\%.$$

- Do NOT write it as $\mathbb{P}\left(5 - \frac{1.96}{\sqrt{10}} \leq \mu \leq 5 + \frac{1.96}{\sqrt{10}}\right) = 95\%$!
- Do NOT say "μ lies in $\left[5 - \frac{1.96}{\sqrt{10}}, 5 + \frac{1.96}{\sqrt{10}}\right]$ with probability 95%"
- μ is a **fixed** and **unknown** number
- The correct statement: "If the experiment is performed over and over again, in the long run the random interval $\left[\bar{X} - \frac{1.96}{\sqrt{10}}, \bar{X} + \frac{1.96}{\sqrt{10}}\right]$ will cover μ with probability 95%"

Interpretations



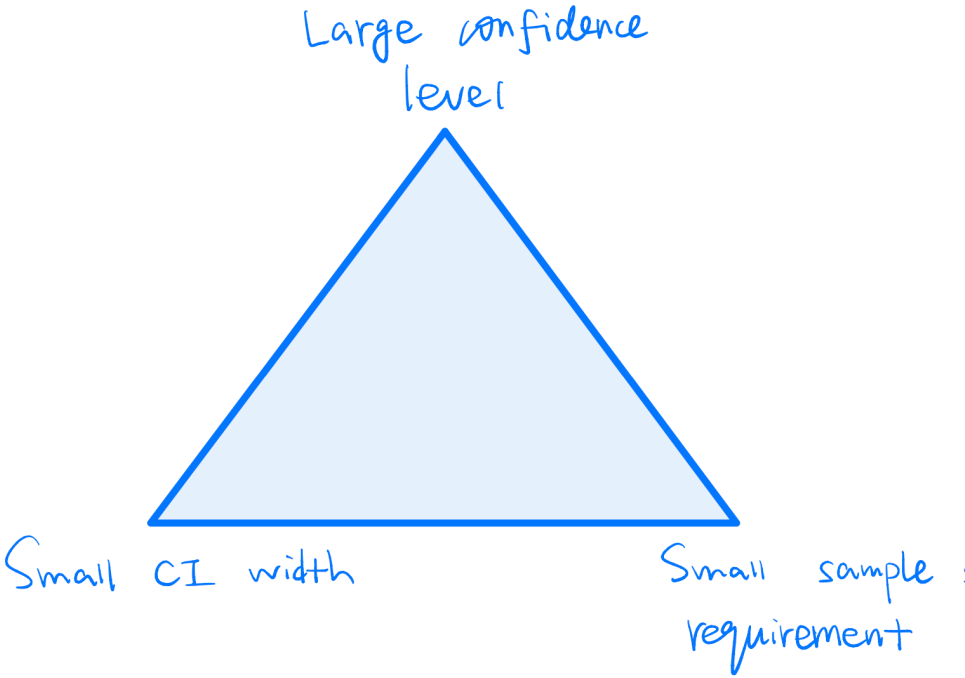
The correct statement: "If the experiment is performed over and over again, in the long run the random interval $\left[\bar{X} - \frac{1.96}{\sqrt{10}}, \bar{X} + \frac{1.96}{\sqrt{10}}\right]$ will cover μ with probability 95%"

Confidence level, precision (width), and sample size

The ideal scenario:

- **Large** confidence level
- **Small** CI width
- **Small** sample size requirement

But this is an **impossible trinity**!



$$\mathbb{P}\left(\bar{X} - \frac{z_{\alpha/2}}{\sqrt{n}} \leq \mu \leq \bar{X} + \frac{z_{\alpha/2}}{\sqrt{n}}\right) = 1 - \alpha.$$

Confidence level $= 100(1-\alpha)\%$, CI width $= \frac{2z_{\alpha/2}}{\sqrt{n}}$, sample size n.

- Fixed confidence level: CI width $\searrow$, n needs to be $\nearrow$
- Fixed confidence level: n $\searrow$, CI width $\nearrow$
- Fixed sample size n: CI width $\nearrow$, corresponding confidence level $\nearrow$
- Fixed sample size CI width: n $\nearrow$, corresponding confidence level $\nearrow$

Confidence level, precision (width), and sample size

Example: Extensive monitoring of a computer time-sharing system has suggested that response time to a particular editing command is normally distributed with standard deviation 1 millisec. We wish to estimate the true average response time μ. What sample size is necessary to ensure that the resulting 95% CI has a width of (at most) 0.5?

Recall that the 95% CI is $\left[\bar{X} - \frac{z_{0.025}}{\sqrt{n}}, \bar{X} + \frac{z_{0.025}}{\sqrt{n}}\right] = \left[\bar{X} - \frac{1.96}{\sqrt{n}}, \bar{X} + \frac{1.96}{\sqrt{n}}\right]$. The width $= 2 \times \frac{1.96}{\sqrt{n}}$. Thus we need

$$2 \times \frac{1.96}{\sqrt{n}} \leq 0.5 \Rightarrow n \geq \left(\frac{2 \times 1.96}{0.5}\right)^2 = 61.4656.$$

Therefore, we need sample size $n \geq 62$.

Confidence intervals

Definition: Suppose we have samples $X_1, \ldots, X_n$. If we can construct an random interval $[l(X_1, \ldots, X_n), u(X_1, \ldots, X_n)]$ which satisfies

$$\mathbb{P}(l(X_1, \ldots, X_n) \leq \theta \leq u(X_1, \ldots, X_n)) = 1 - \alpha, \tag{1}$$

then it can be called as a $(100(1-\alpha))\%$ **confidence interval (CI)**. $(100(1-\alpha))\%$ is the **confidence level**.

Steps to construct a CI: Suppose we have samples $X_1, \ldots, X_n$, and θ is the unknown parameter.

- **Step 1:** Find a r.v. V which is a function of both $X_1, \ldots, X_n$ and θ
- **Step 2:** Verify that the distribution of V does NOT depend on θ or any other unknown parameters
- **Step 3:** Derive equation (1) from the fact that $\mathbb{P}(v_{1-\alpha/2} \leq V(X_1, \ldots, X_n, \theta) \leq v_{\alpha/2}) = 1 - \alpha$, where $\mathbb{P}(V \geq v_\beta) = \beta$ for any $\beta \in [0, 1]$.

Revisit the derivation of CI of μ in $N(\mu, 1)$

Steps to construct a CI: Suppose we have samples $X_1, \ldots, X_n$, and θ is the unknown parameter.

- **Step 1:** Find a r.v. V which is a function of both $X_1, \ldots, X_n$ and θ
- **Step 2:** Verify that the distribution of V does NOT depend on θ or any other unknown parameters
- **Step 3:** Derive equation (1) from the fact that $\mathbb{P}(v_{1-\alpha/2} \leq V(X_1, \ldots, X_n, \theta) \leq v_{\alpha/2}) = 1 - \alpha$, where $\mathbb{P}(V \geq v_\beta) = \beta$ for any $\beta \in [0, 1]$.

Suppose $X_1, \ldots, X_n \overset{i.i.d.}{\sim} N(\mu, 1)$. $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i \sim N(\mu, 1/n)$. By standardization, $\frac{\bar{X}-\mu}{\sqrt{1/n}} \sim N(0, 1)$. Thus,

$$\mathbb{P}\left(z_{0.975} \leq \frac{\bar{X} - \mu}{\sqrt{1/n}} \leq z_{0.025}\right) = 95\%,$$

Then

$$\mathbb{P}\left(\bar{X} - \frac{z_{0.025}}{\sqrt{n}} \leq \mu \leq \bar{X} - \frac{z_{0.975}}{\sqrt{n}}\right) = 95\%.$$

More Examples

χ^2-distribution (Chi-squared distribution)

Suppose $Z_1, \ldots, Z_p \overset{i.i.d.}{\sim} N(0,1)$. Then we say variable $V = \sum_{i=1}^{p} Z_i^2$ follows the **χ^2-distribution** with **degree of freedom** p, denoted as $V \sim \chi_p^2$.

Property: $\mathbb{E}V = p$, $\mathrm{Var}(V) = 2p$.

Applications:

- Suppose $X_1, \ldots, X_n \overset{i.i.d.}{\sim} N(\mu, \sigma^2)$. The sample variance $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$. Then $\frac{(n-1)s^2}{\sigma^2} \sim \chi_{n-1}^2$.
- Suppose $X_1, \ldots, X_n \overset{i.i.d.}{\sim} \mathrm{Exp}(\lambda)$. Then $2\lambda \sum_{i=1}^{n} X_i \sim \chi_{2n}^2$. (You will use this result in HW4)

You may find Table A.7 in our textbook useful, which gives the table of quantile values of χ_p^2 distribution with different p.

CI of σ^2 in $N(\mu, \sigma^2)$

Suppose $X_1, \ldots, X_n \overset{i.i.d.}{\sim} N(\mu, \sigma^2)$. The sample variance $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$.Then $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$.
Then

$$\mathbb{P}\left(\chi^2_{n-1,1-\alpha/2} \leq \frac{(n-1)S^2}{\sigma^2} \leq \chi^2_{n-1,\alpha/2}\right) = 1-\alpha,$$

which implies

$$\mathbb{P}\left(\frac{(n-1)S^2}{\chi^2_{n-1,\alpha/2}} \leq \sigma^2 \leq \frac{(n-1)S^2}{\chi^2_{n-1,1-\alpha/2}}\right) = 1-\alpha.$$

So a $100(1-\alpha)\%$ CI of σ^2 is $\left[\frac{(n-1)S^2}{\chi^2_{n-1,\alpha/2}}, \frac{(n-1)S^2}{\chi^2_{n-1,1-\alpha/2}}\right]$.

Example: In one observation, we have $n = 10$ and $s^2 = 9.06$, then the 95% CI of σ^2 is $\left[\frac{9\times 9.06}{19.022}, \frac{9\times 9.06}{2.7}\right] = [4.29, 30.20]$.

t-distribution (Student's t-distribution)

Suppose $Z \sim N(0,1)$, $Q \sim \chi^2_p$, and Z is independent with Q. Then we say the variable $\frac{Z}{\sqrt{Q/p}}$ follows **t-distribution with degree of freedom** p, denoted as $\frac{Z}{\sqrt{Q/p}} \sim t_p$.

Applications: Suppose $X_1, \ldots, X_n \overset{i.i.d.}{\sim} N(\mu, \sigma^2)$. The sample variance $S^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X})^2$. Then $\frac{\sqrt{n}(\bar{X}-\mu)}{S} \sim t_{n-1}$.

Proof: We need to use the fact (not trivial to prove!) that $\bar{X} \perp\!\!\!\perp S$.

- $\bar{X} \sim N(\mu, \sigma^2/n) \Rightarrow \frac{\sqrt{n}(\bar{X}-\mu)}{\sigma} \sim N(0,1)$.
- On the other hand, $\frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{n-1}$.

Finally, since $\frac{\sqrt{n}(\bar{X}-\mu)}{\sigma} \perp\!\!\!\perp \frac{(n-1)s^2}{\sigma^2}$, we have

$$\frac{\sqrt{n}(\bar{X}-\mu)}{\sigma} \Big/ \frac{(n-1)s^2}{\sigma^2} = \frac{\sqrt{n}(\bar{X}-\mu)}{S} \sim t_{n-1}.$$

Many thanks to

- Joyce Robbins
- Yang Feng
- Chengliang Tang
- Owen Ward
- Wenda Zhou
- And all my teachers in the past 25 years