통계분석
Statistical Analysis

Statistical Inferences

Statistics Inferences

- **Point Estimation: single value estimate from sample data**

$$\text{mean } \mu, \text{variance } \sigma^2$$

- **Interval Estimation: [a,b] interval estimate**

Confidence interval for population mean

$$a < \mu < b$$

Interval Estimation

Point Estimation?

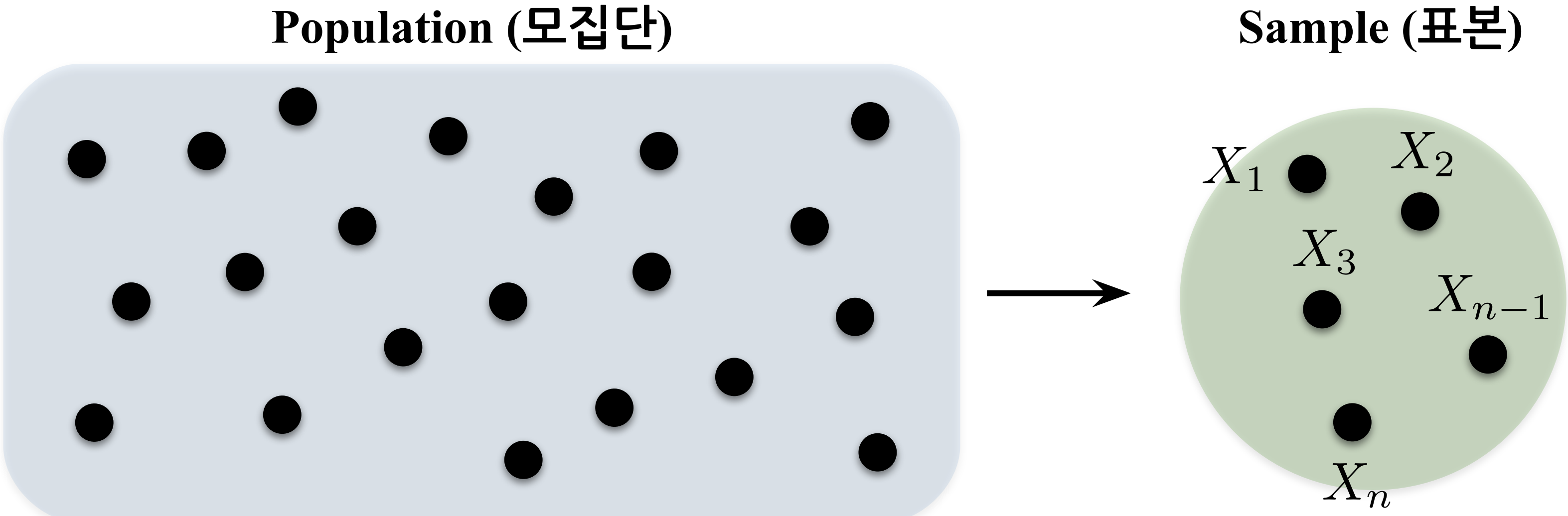
$$(\hat{\theta} - \theta) = \text{Error of estimation}$$

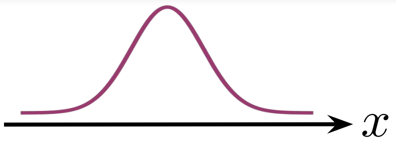
- Point estimation gives just a single number, but it cannot say how close to the parameter such an estimation is.
- Instead of estimating a single number, let us try to say that the parameter is inside some interval of plausible values with a certain probability.

$$P\left(a < \theta < b\right) = 1 - \alpha$$

→ This inferential approach is called "***interval estimation***" or "***confidence interval***."

Sampling to Estimate Population Properties



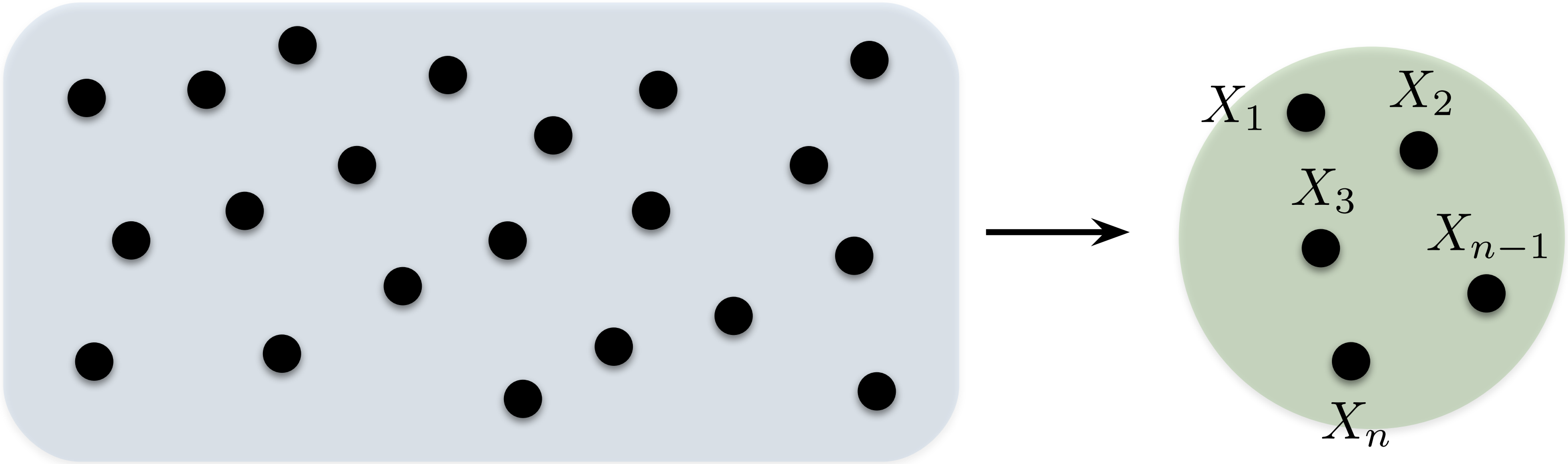


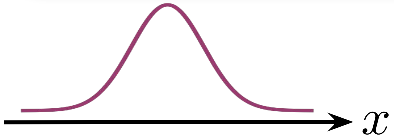
Probability distribution (Population)

- We want to know population properties (parameters), but we cannot investigate the population.
- Instead we take a sample from the population and estimate population properties from sample statistics
- Sample statistics : 표본 통계량

Distribution of Sample Statistics

Population (모집단) **Sample (표본)**





Probability distribution (Population)

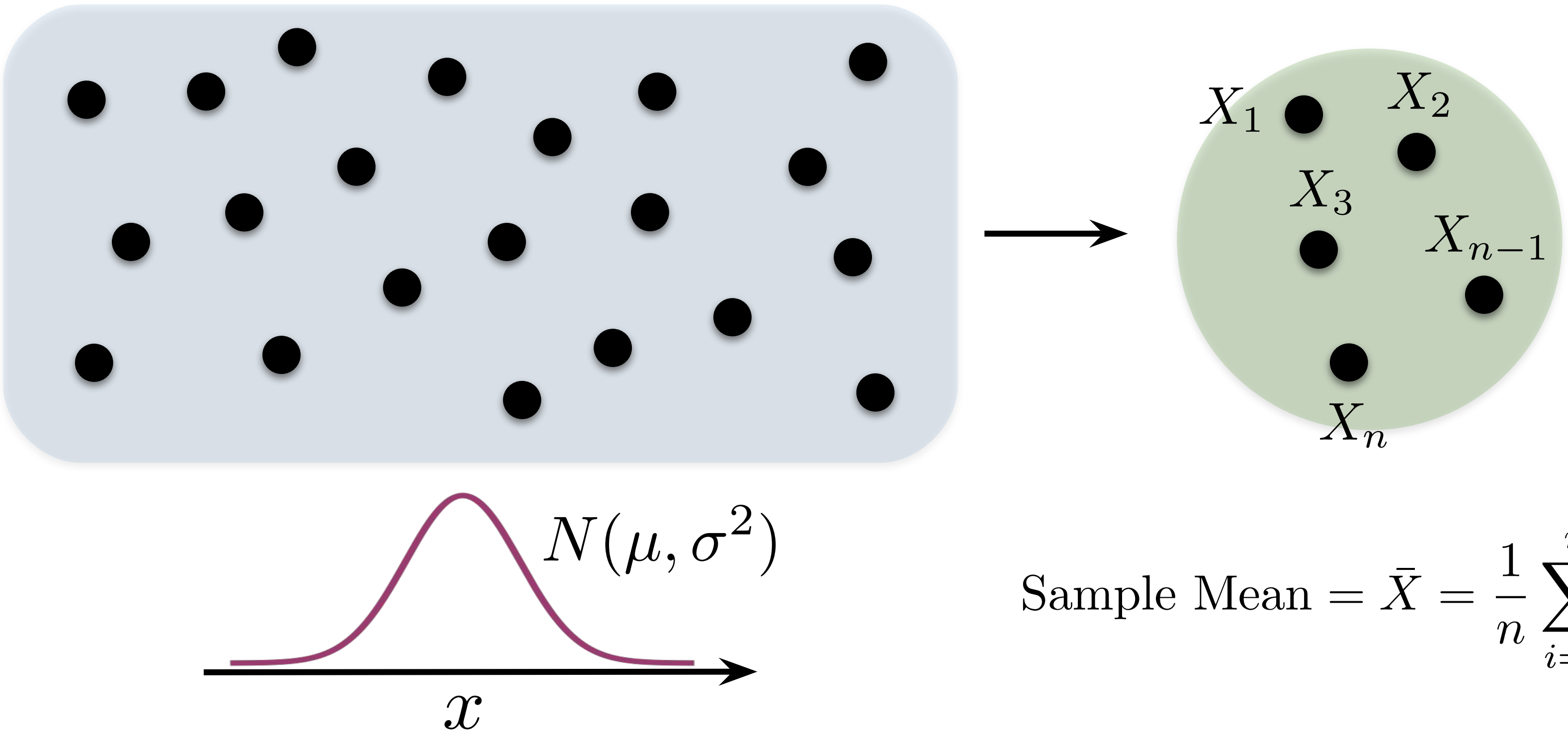
$$\text{Sample statistic} = h(X_1, X_2, \cdots, X_n)$$

- We need to know the probability distribution of the sample statistic.
- The distribution of the sample statistic is derived from the distribution of the population.

Example: Distribution of Sample Mean

Population (모집단)

Sample (표본)



$$\text{Sample Mean} = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Distribution of Sample Mean : Case I

$X_1, X_2, \cdots, X_n$: Random sample from $N(\mu, \sigma^2)$

Here we consider

- $X_i \sim N(\mu, \sigma^2)$
- σ is known. (In general, population std is also unknown.)
- μ is an unknown parameter to be estimated here.

Sample Mean $\bar{X} \sim N(\mu, \sigma^2/n) \longrightarrow Z = \dfrac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$

Distribution of Sample Mean : Case I

$$X_1, X_2, \cdots, X_n : \text{Random sample from } N(\mu, \sigma^2)$$

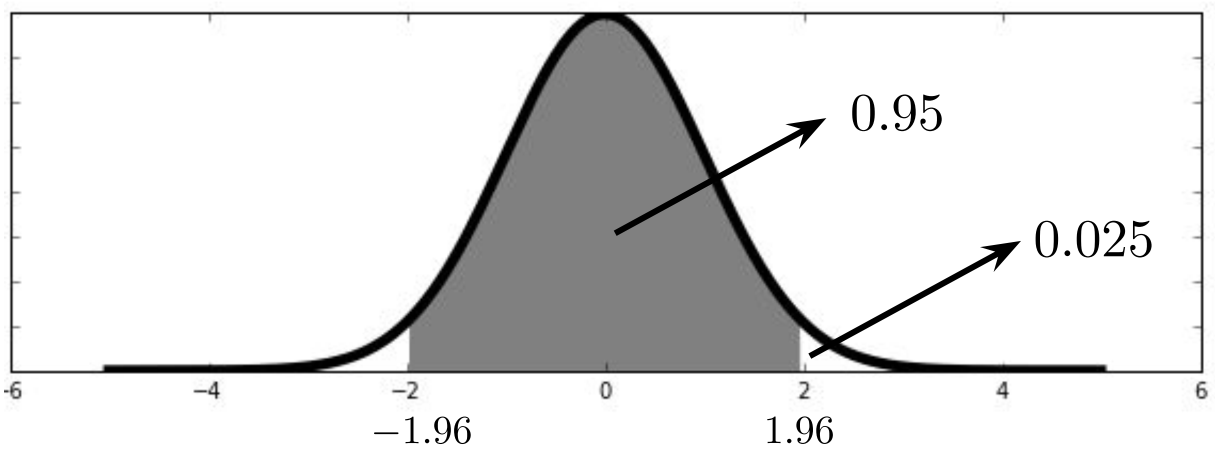
Sample Mean $\bar{X} \sim N(\mu, \sigma^2/n) \longrightarrow Z = \dfrac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$

$$P\left(-a < \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < a\right) = 1 - \alpha : \text{ The probability that } -a < Z < a \text{ is } 1 - \alpha$$

- **How to find the confidence interval**

1. First specify the probability $1 - \alpha$.
   The probability $1 - \alpha$ is defined as the confidence level.
2. Find out the value a that leads to $1 - \alpha$.
3. Reorganizing the above interval, the interval for μ can be determined.

Distribution of Sample Mean : Case I



- Confidence level $1-\alpha = 0.95 \longrightarrow a = 1.96$
- $$P\left(-1.96 < \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} < 1.96\right) = 0.95$$
- $$-1.96 < \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} < 1.96 \quad \longrightarrow \bar{X} - 1.96\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}$$

Distribution of Sample Mean : Case I

- $P\left(-1.96 < \dfrac{\bar{X}-\mu}{\sigma/\sqrt{n}} < 1.96\right) = 0.95$

- $-1.96 < \dfrac{\bar{X}-\mu}{\sigma/\sqrt{n}} < 1.96 \quad \longrightarrow \bar{X} - 1.96\dfrac{\sigma}{\sqrt{n}} < \mu < \bar{X} + 1.96\dfrac{\sigma}{\sqrt{n}}$

- Probability that $\mu \in \left(\bar{X} - 1.96\dfrac{\sigma}{\sqrt{n}}, \bar{X} + 1.96\dfrac{\sigma}{\sqrt{n}}\right)$ is 0.95
- The random variable is $\bar{X}$. μ is the parameter value.
- $\left(\bar{X} - 1.96\dfrac{\sigma}{\sqrt{n}}, \bar{X} + 1.96\dfrac{\sigma}{\sqrt{n}}\right)$ is a random interval depending on $\bar{X}$

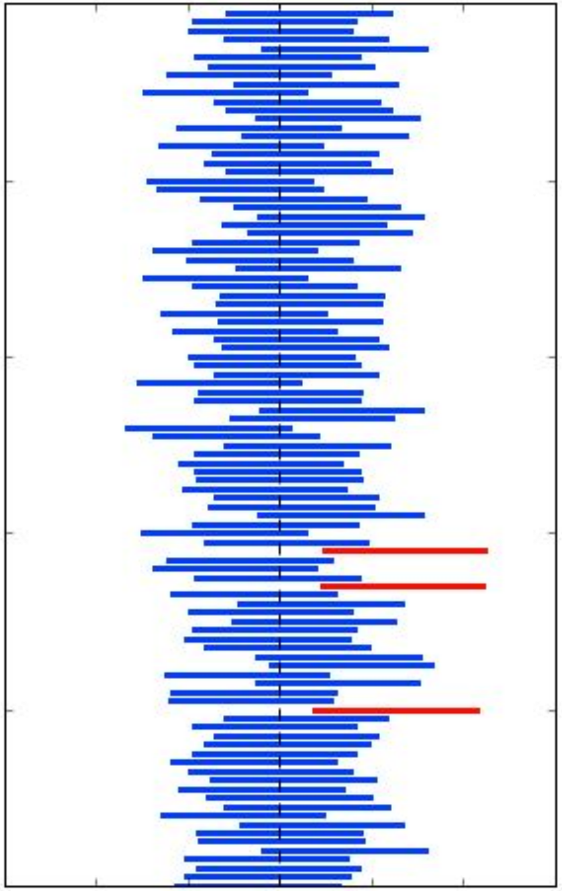
Confidence Interval: Interpretation

$$P\left(\bar{X} - 1.96\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}\right) = 0.95$$

0.95 confidence interval for the parameter μ.

- Interpretation: The probability is 0.95 with which the ***random interval*** includes the true value of the parameter.
- The ***random interval*** means that the interval changes sample by sample.

Confidence Interval: Interpretation



100 random intervals,

each of which has $n = 20$ elements

The true value of μ is fixed.

Among 100, three intervals do not include μ.

Do not be confused with that the interval is fixed.
The true value of the parameter is included in a random interval with probability 0.95.

- $\mu =$ a fixed (unknown) constant, which is not random
- $\left(\bar{X} - 1.96\frac{\sigma}{\sqrt{n}}, \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}\right)$ is a random interval.

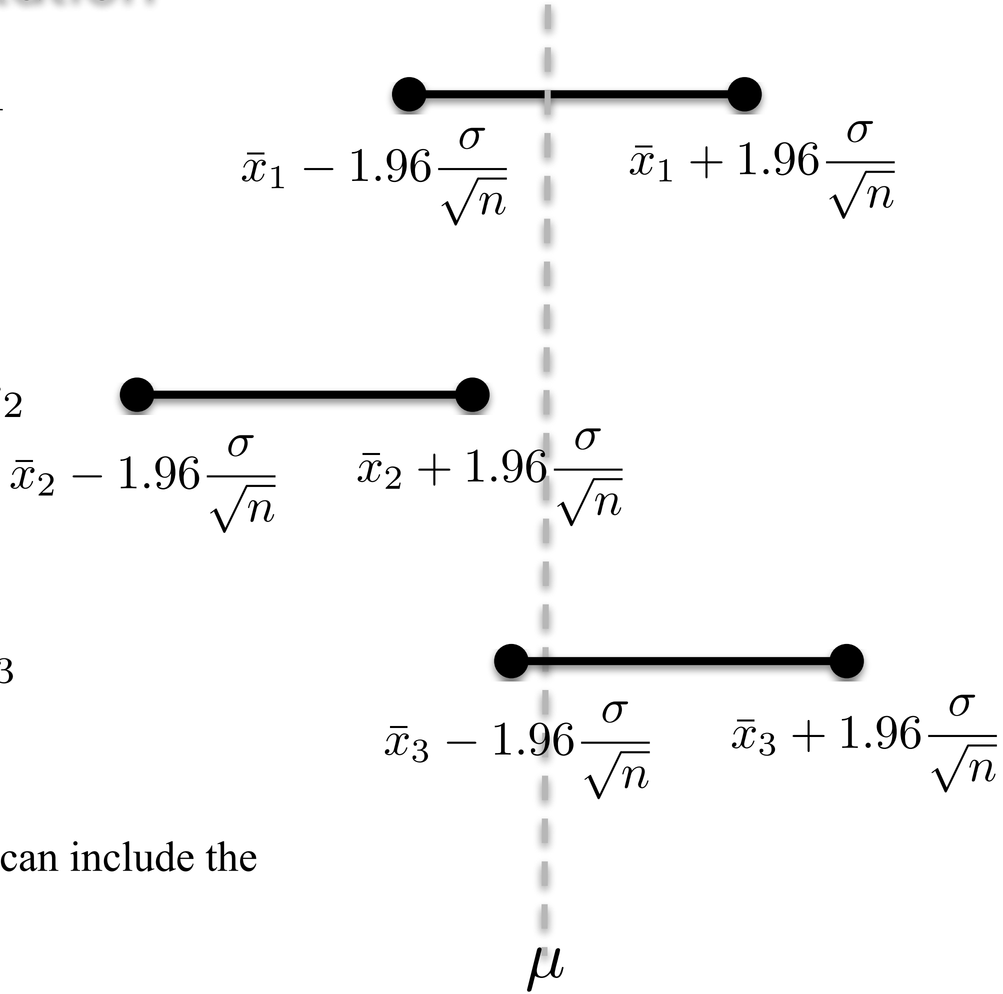
This interval is not fixed, but different sample by sample.

Confidence Interval: Interpretation

1st sample: $x_{11}, x_{12}, \cdots x_{1n} \rightarrow \bar{x}_1$

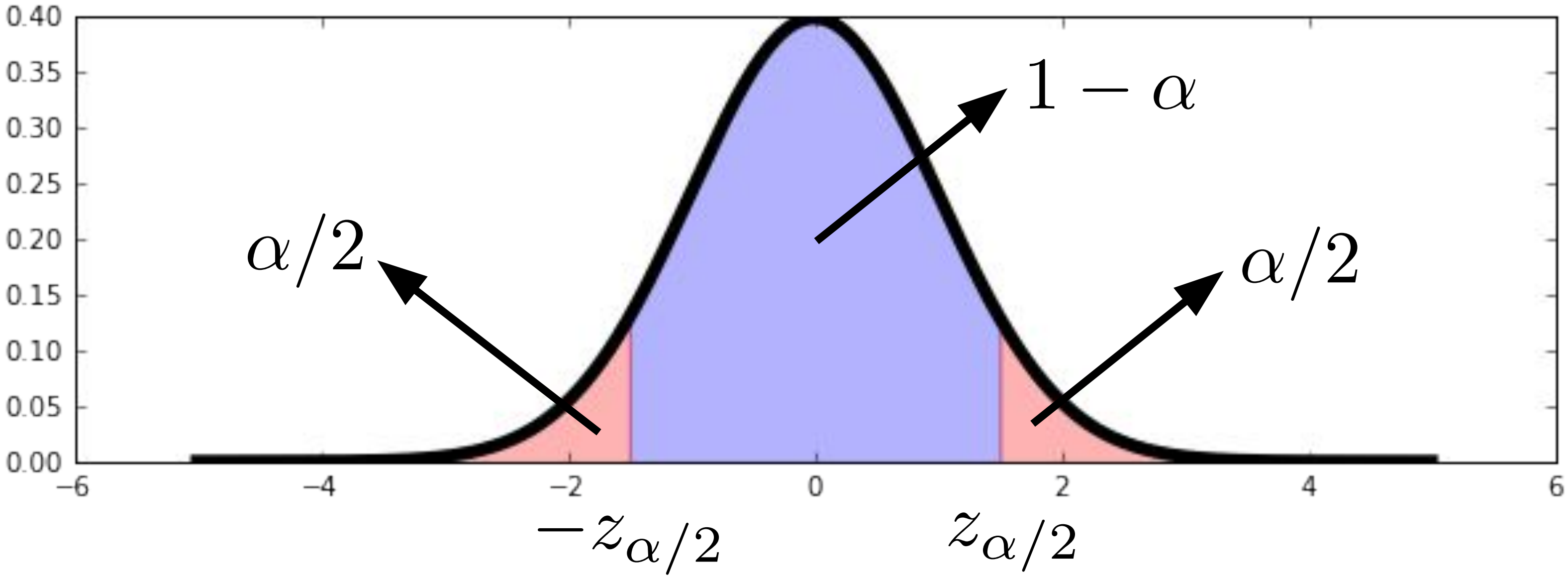
2nd sample: $x_{21}, x_{22}, \cdots x_{2n} \rightarrow \bar{x}_2$

3rd sample: $x_{31}, x_{32}, \cdots x_{3n} \rightarrow \bar{x}_3$



The interval differs sample by sample.
With probability 95%, this random interval can include the true value of the parameter.

Confidence Interval for Normal Distribution

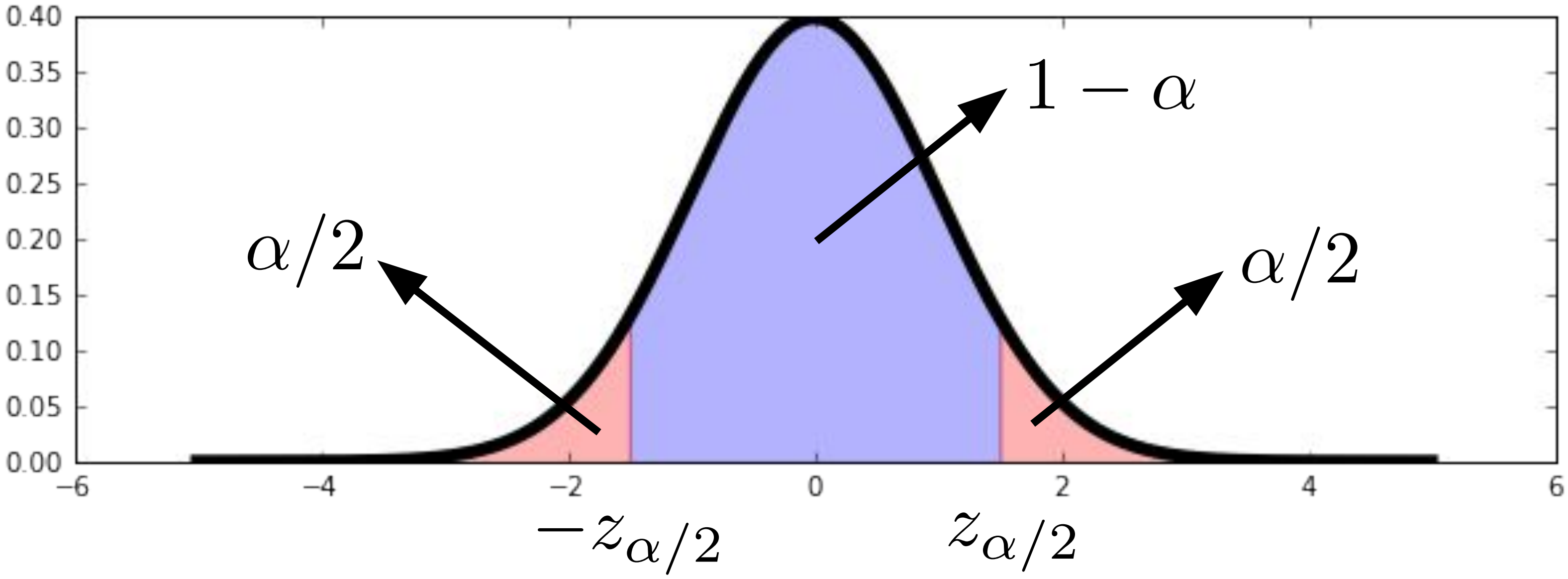


Recall z-values : $P\left(Z > z_{\alpha/2}\right) = \alpha/2$

$$P\left(Z < -z_{\alpha/2}\right) = \alpha/2$$

$100(1-\alpha)\%$ confidence level: $P\left(-z_{\alpha/2} < Z < z_{\alpha/2}\right) = 1-\alpha$

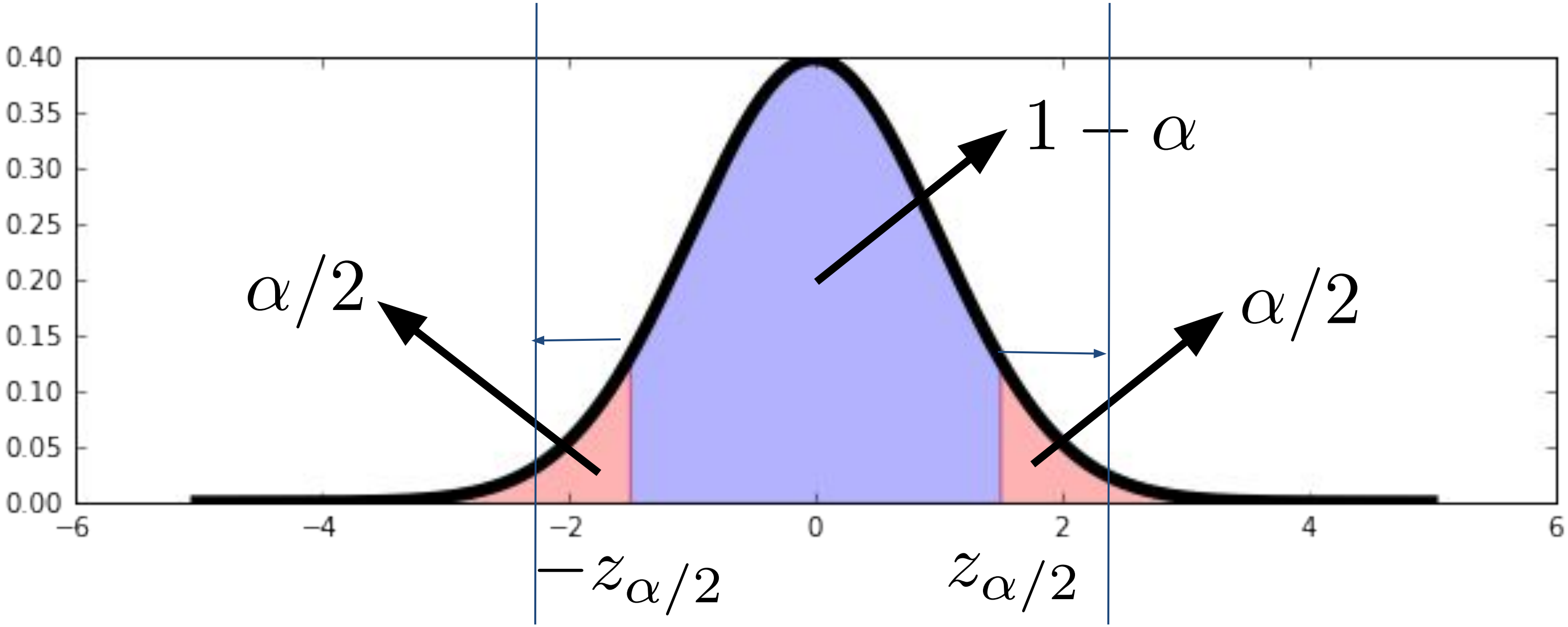
Confidence Interval for Normal Distribution



$$100(1-\alpha)\%\ \text{Confidence Interval with } \sigma \text{ known :}$$

$$-z_{\alpha/2} < Z < z_{\alpha/2} \longrightarrow \bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

Trade-off between Confidence Level and Interval Size



$$100(1-\alpha)\% \text{ Confidence Interval with } \sigma \text{ known :}$$

$$-z_{\alpha/2} < Z < z_{\alpha/2} \longrightarrow \bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

Trade-off between Confidence Level and Interval Size

1. When you increase the confidence level, alpha decreases.
2. Magnitude of corresponding critical values also increases.
3. We will get a larger estimated interval; It is more difficult to specify your population parameter.

90% confidence interval -> [15, 20]
99% confidence interval -> [5,   40]

Confidence Intervals

This case was previously discussed.

- Confidence Interval of normal distribution with unknown mean, but with known variance

- Confidence Interval of normal distribution with both population mean and population variance unknown

- Confidence Interval of large sample

Confidence Intervals

- Confidence Interval of normal distribution with unknown mean, but with known variance
- Confidence Interval of normal distribution with both mean and variance unknown
- Confidence Interval of large sample

Case II : Average and Variance are Unknown

Recall

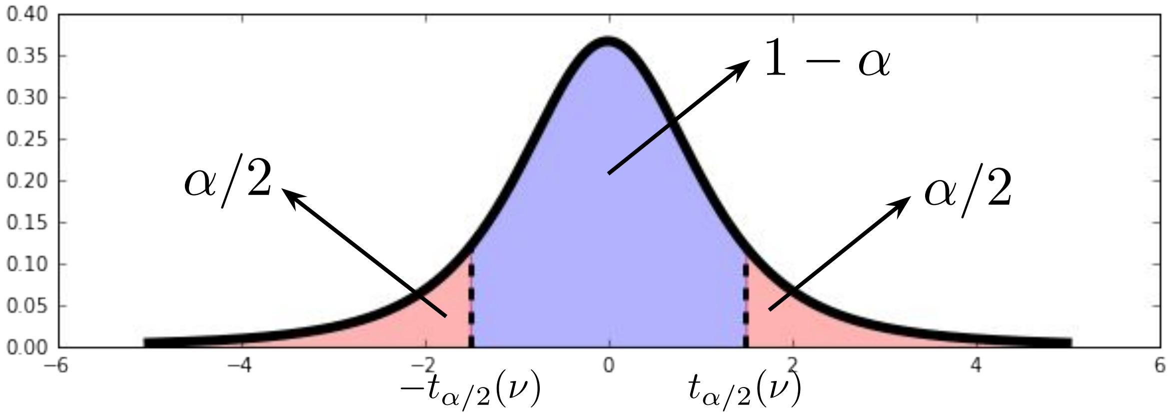
$$X_1, X_2, \cdots, X_n : \text{Random sample from } N(\mu, \sigma^2)$$

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t(n-1)$$

Student's t-distribution with (n-1) degrees of freedom

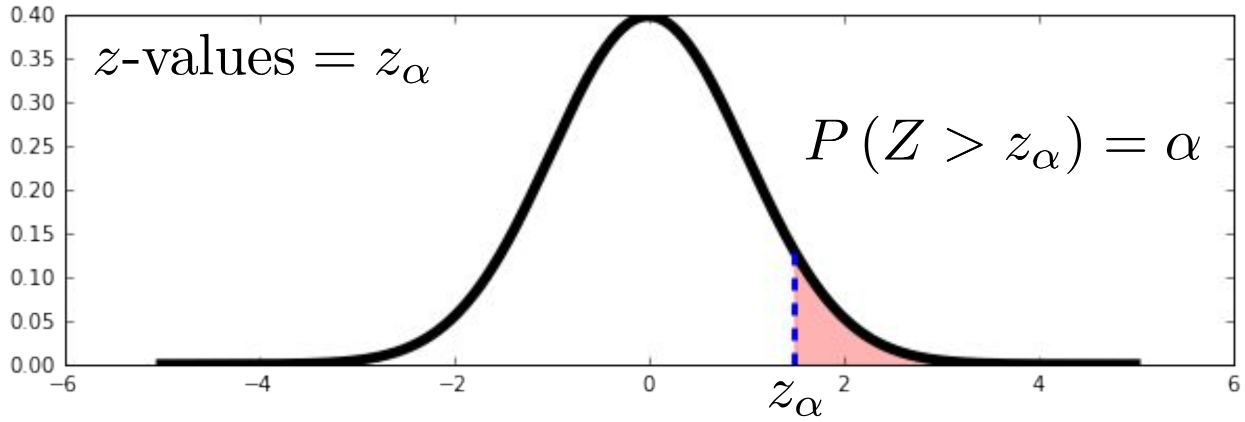
- By calculating the sample average and the sample variance, one can obtain Confidence Interval for the population average from Student's *t*-distribution.

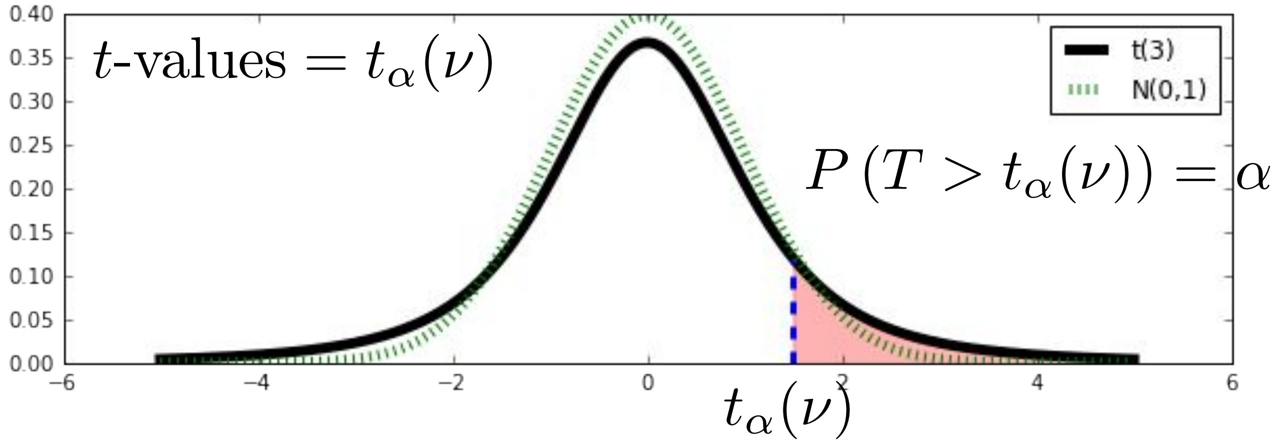
Case II : Average and Variance are Unknown



- Sample Mean satisfies $T = \dfrac{\bar{X}-\mu}{S/\sqrt{n}} \sim t(n-1)$
- $100(1-\alpha)\%$ confidence level: $P\left(-t_{\alpha/2}(\nu) < T < t_{\alpha/2}(\nu)\right) = 1-\alpha$
- $-t_{\alpha/2}(\nu) < T < t_{\alpha/2}(\nu) \longrightarrow \bar{X} - t_{\alpha/2}(n-1)\dfrac{S}{\sqrt{n}} < \mu < \bar{X} + t_{\alpha/2}(n-1)\dfrac{S}{\sqrt{n}}$

Recall : t-Values





Confidence Intervals

- Confidence Interval of normal distribution with unknown mean, but with known variance
- Confidence Interval of normal distribution with both mean and variance unknown
- Confidence Interval of large sample

Confidence Intervals for Large Sample

- Here we do NOT assume that the population distribution is normal.

- Without knowing the type of the population distribution, we can estimate the confidence interval, ***especially when the sample is large***.

How large is the sample?

Central Limit Theorem

For a sufficiently large n,

$\bar{X}$ is approximately a normal distribution with $\mu_{\bar{X}} = \mu, \sigma^2_{\bar{X}} = \sigma^2/n$

When n is large, $\bar{X} \sim N(\mu, \sigma^2/n)$

Confidence Intervals for Large Sample

Central Limit Theorem

For a sufficiently large n,

$\bar{X}$ is approximately a normal distribution with $\mu_{\bar{X}} = \mu, \sigma^2_{\bar{X}} = \sigma^2/n$

When n is large, $\bar{X} \sim N(\mu, \sigma^2/n)$ $\longrightarrow$ $Z = \dfrac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$

We would like to infer ***<u>the population average</u>***,

1. when the variance is known: $$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

2. when the variance is unknown: $$Z = \frac{\bar{X} - \mu}{s/\sqrt{n}} \sim N(0,1)$$

$$\sigma \approx s \text{ (point estimate)}$$

- If n is very large, this substitution is quite good.
- Recall the minimum variance unbiased estimator of population variance.

Confidence Intervals for Large Sample

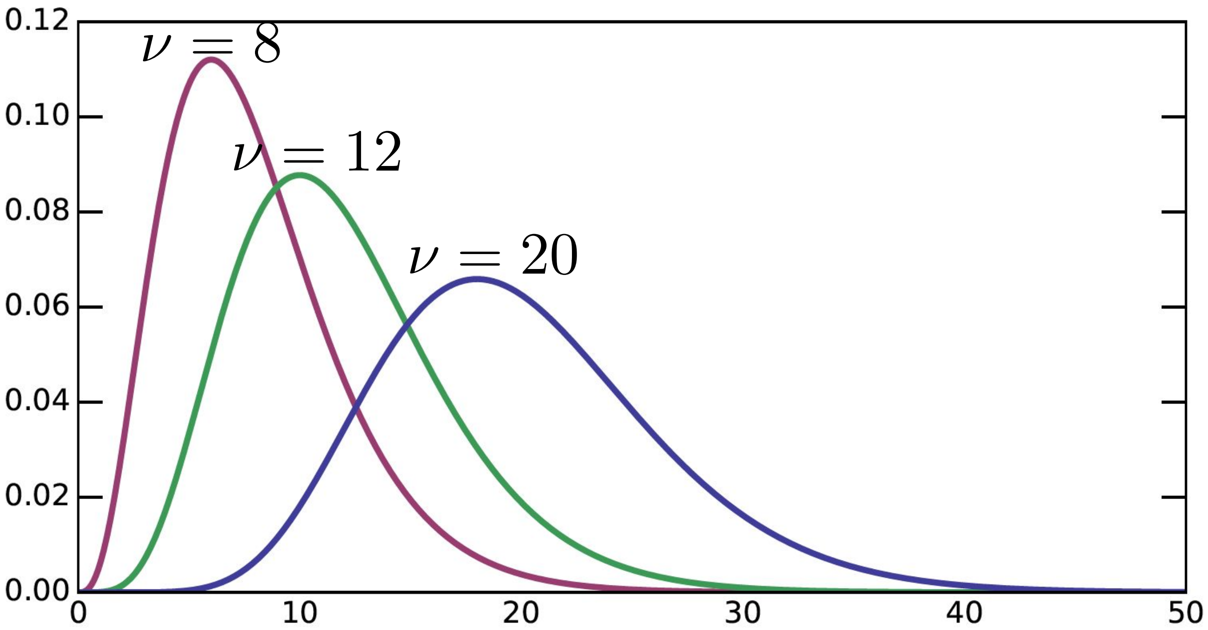
1. when the variance is known: $Z = \dfrac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$

2. when the variance is unknown: $Z = \dfrac{\bar{X} - \mu}{s/\sqrt{n}} \sim N(0, 1)$

100(1-α)% confidence interval of the population mean for large sample

$$P\left(-z_{\alpha/2} < Z < z_{\alpha/2}\right) \approx 1 - \alpha$$

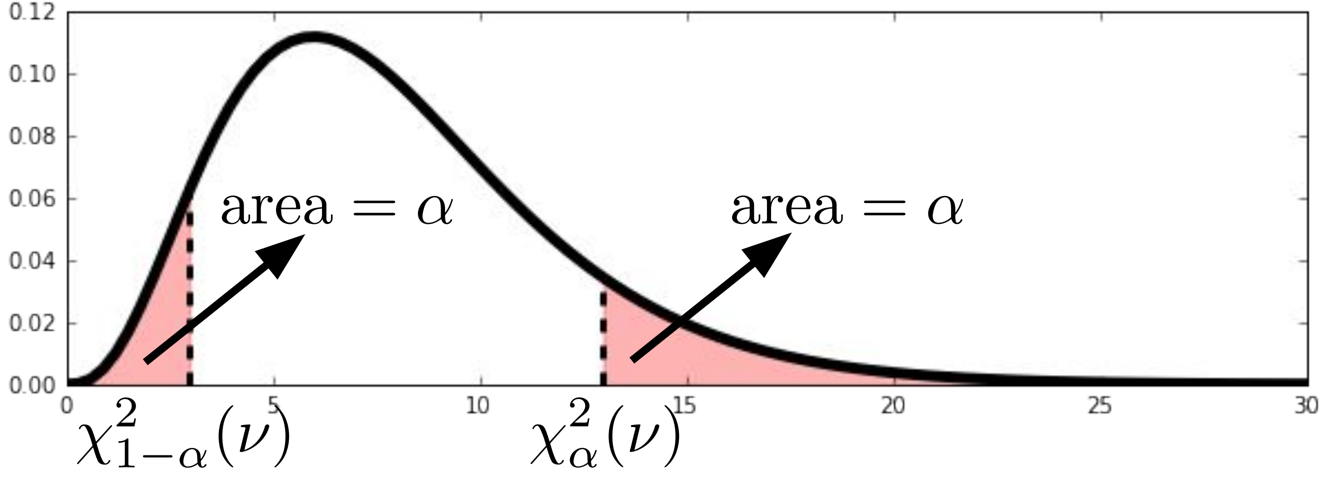
Confidence Intervals for Variances of Normal Distributions



$$Y = \sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2 / \sigma^2 = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$$

Chi-Squared distribution with (n-1) degrees of freedom

Chi-Squared Critical Values

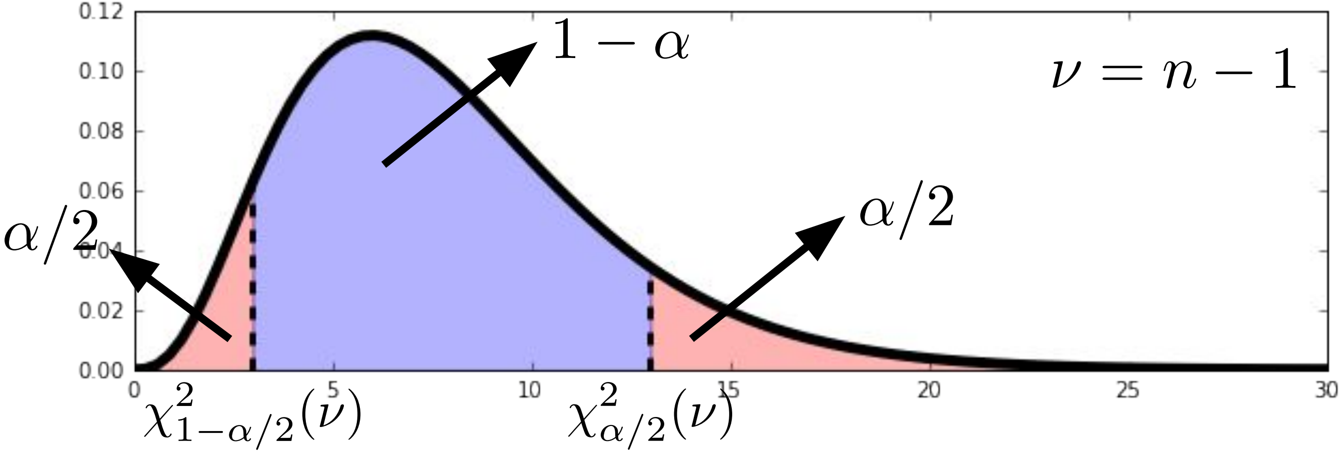


$$P\left(Y > \chi^2_\alpha(\nu)\right) = \alpha \qquad P\left(Y > \chi^2_{1-\alpha}(\nu)\right) = 1 - \alpha$$

Unlike normal or t distributions which are symmetric with respect to 0,

$$\chi^2_{1-\alpha}(\nu) \neq -\chi^2_\alpha(\nu)$$

Confidence Intervals with Chi²



$$P\left(\chi^2_{1-\alpha/2}(\nu) < Y = \frac{(n-1)S^2}{\sigma^2} < \chi^2_{\alpha/2}(\nu)\right) = 1-\alpha$$

$$\longrightarrow \frac{(n-1)S^2}{\chi^2_{\alpha/2}(\nu)} < \sigma^2 < \frac{(n-1)S^2}{\chi^2_{1-\alpha/2}(\nu)}$$

100(1-α)% confidence interval for the population variance of normal distribution.

Summary: Steps for Confidence Intervals

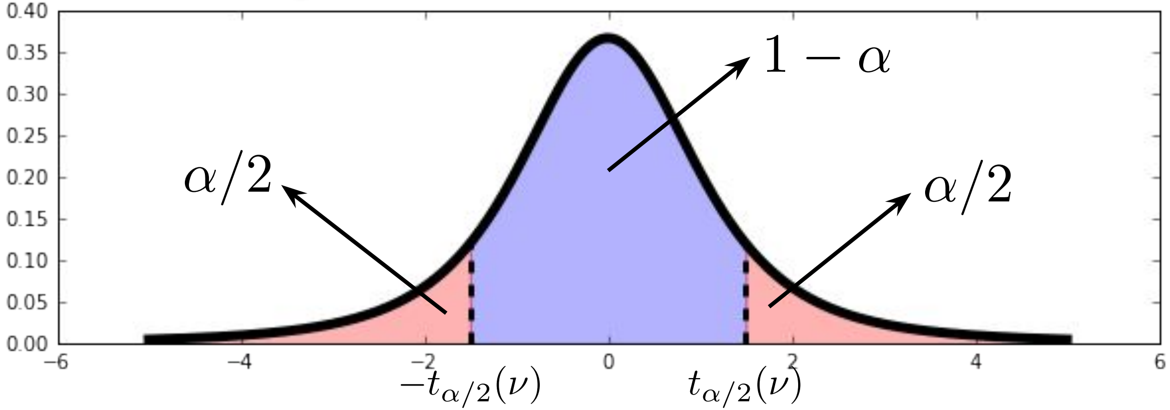
1. What is the population parameter to know?
2. What is the sample statistic to use for the population parameter estimation?
3. What is the distribution of the sample statistic? What variable $\boldsymbol{X}$ follows the distribution? Here the variable $\boldsymbol{X}$ involves the population parameter and the sample statistic.
4. Specify the confidence level.
5. Find out the interval (critical values) of the variable $\boldsymbol{X}$ satisfying the confidence level.
6. Reformulate the above interval to obtain the interval of the population parameters.

Prediction Intervals for a Single Future Value

Here the objective is to predict a single value of a variable to be observed at some future time, rather than to estimate the mean value of that variable. We have available a random sample $X_1, X_2, \cdots, X_n$ from a normal population distribution, and wish to predict the value of X_{n+1}, a single future observation.

- A point predictor $= \bar{X}$
- The prediction error $= \bar{X} - X_{n+1}$
- $E\left(\bar{X} - X_{n+1}\right) = E\left(\bar{X}\right) - E\left(X_{n+1}\right) = \mu - \mu = 0$
- $V\left(\bar{X} - X_{n+1}\right) = V\left(\bar{X}\right) + V\left(X_{n+1}\right) = \dfrac{\sigma^2}{n} + \sigma^2 = \sigma^2\left(1 + \dfrac{1}{n}\right)$
- $Z = \dfrac{\left(\bar{X} - X_{n+1}\right) - 0}{\sqrt{\sigma^2\left(1 + \frac{1}{n}\right)}} = \dfrac{\bar{X} - X_{n+1}}{\sqrt{\sigma^2\left(1 + \frac{1}{n}\right)}} \sim N(0, 1)$
- if σ is unknown, σ is replaced by S. $\rightarrow T = \dfrac{\bar{X} - X_{n+1}}{\sqrt{S^2\left(1 + \frac{1}{n}\right)}} \sim t(n-1)$

Prediction Intervals for a Single Future Value



- The prediction error satisfies $T = \dfrac{\bar{X} - X_{n+1}}{\sqrt{S^2\left(1+\frac{1}{n}\right)}} \sim t(n-1)$

- $100(1-\alpha)\%$ confidence level: $P\left(-t_{\alpha/2}(\nu) < T < t_{\alpha/2}(\nu)\right) = 1-\alpha$

- $-t_{\alpha/2}(\nu) < T < t_{\alpha/2}(\nu) \longrightarrow$

$$\bar{X} - t_{\alpha/2}(n-1)\cdot S\sqrt{1+\frac{1}{n}} < X_{n+1} < \bar{X} + t_{\alpha/2}(n-1)\cdot S\sqrt{1+\frac{1}{n}}$$