Confidence Interval

Some concepts: Interval estimate, coverage probability, confidence coefficient, confidence interval (CI)

Definition: an **interval estimate** for a real-valued parameter θ based on a sample $\underline{X} \equiv (X_1, \ldots, X_n)$ is a *pair of functions* $L(\underline{X})$ and $U(\underline{X})$ so that $L(\underline{X}) \leq U(\underline{X})$ for all $\underline{X}$, that is $[L(\underline{X}), U(\underline{X})]$.

Note:

• *The above is a two-sided confidence interval,* one can also define one-sided intervals: $(-\infty, U(\underline{X})]$ or $[L(\underline{X}), \infty)$.

Definition: the **coverage probability** of an interval estimator is

$$P_\theta(\theta \in [L(\underline{X}), U(\underline{X})]) = P_\theta(L(\underline{X}) \leq \theta, U(\underline{X}) \geq \theta)$$

Note:

• This is the probability that the random interval $[L(\underline{X}), U(\underline{X})]$ covers the true θ.

• One problem about the coverage probability is that it can vary depend on what θ is.

Definition: For an interval estimator $[L(\underline{X}), U(\underline{X})]$ of a parameter θ, the **confidence coefficient** $\equiv inf_\theta P_\theta(\theta \in [L(\underline{X}), U(\underline{X})])$.

Note:

• The term **confidence interval** refers to the interval estimate along with its confidence coefficient.

There are two general approaches to derive the confidence interval: (1) the privotal quantitymethod, and (2) invert the test, a introduced next.

1. General approach for deriving CI's :

The Pivotal Quantity Method

Definition: A pivotal quantity is a function of the sample and the parameter of interest. Furthermore, its distribution is entirely known.

Example. Point estimator and confidence interval for μ when the population is normal and the population variance is known.

- Let $X_1, X_2, \ldots, X_n$ be a random sample for a normal population with mean μ and variance σ^2. That is, $X_i \overset{iid.}{\sim} N(\mu, \sigma^2), i = 1, \ldots, n$.
- For now, we assume that σ^2 is known.

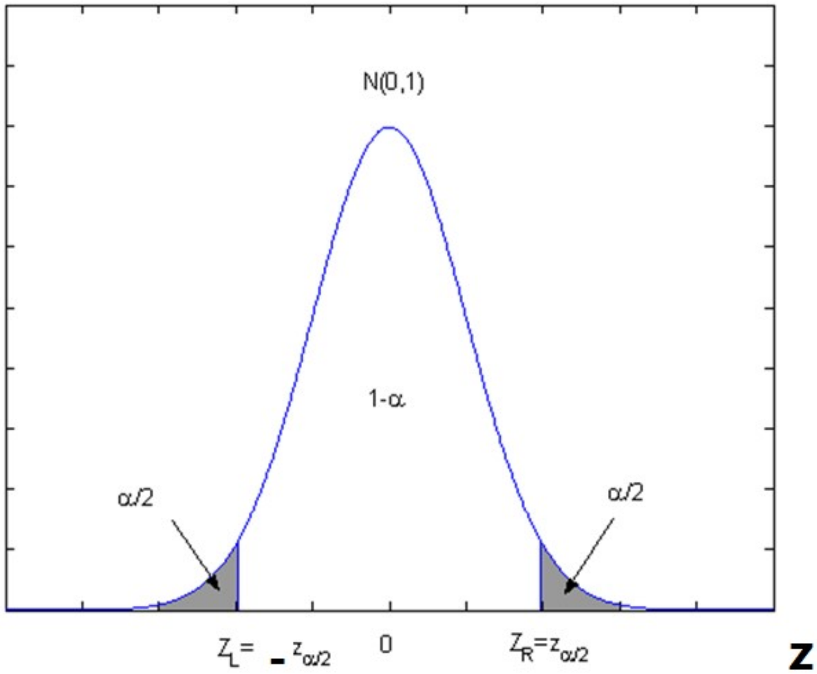
(1). We start by looking at the point estimator of μ :

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

(2). Then we found the pivotal quantity Z:

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

Now we shall start the derivation for the symmetrical CI's for μ from the PDF of the pivotal quantity Z



100(1-α)% CI for μ, 0<α<1

(e.g. α=0.05 ⇒ 95% C.I.)

$$P(-Z_{\alpha/2} \le Z \le Z_{\alpha/2}) = 1-\alpha$$

$$P(-Z_{\alpha/2} \le \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \le Z_{\alpha/2}) = 1-\alpha$$

$$P(-Z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} \le \overline{X}-\mu \le Z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}) = 1-\alpha$$

$$P(-\overline{X} - Z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} \le -\mu \le -\overline{X} + Z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}) = 1-\alpha$$

$$P(\overline{X} + Z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} \ge \mu \ge \overline{X} - Z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}) = 1-\alpha$$

$$P(\overline{X} - Z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} \le \mu \le \overline{X} + Z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}) = 1-\alpha$$

(3) ∴ the 100(1-α)% C.I. for μ is $[\overline{X} - Z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}, \overline{X} + Z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}]$

***Note, some special values for α and the corresponding $Z_{\alpha/2}$ values are:**

1. The 95% CI, where $\boldsymbol{\alpha} = \mathbf{0.05}$ and the corresponding $\mathbf{Z}_{\frac{\alpha}{2}} = \mathbf{Z}_{0.025} = \mathbf{1.96}$

2. The 90% CI, where $\boldsymbol{\alpha} = \mathbf{0.1}$ and the corresponding $\mathbf{Z}_{\frac{\alpha}{2}} = \mathbf{Z}_{0.05} = \mathbf{1.645}$

3. The 99% CI, where $\boldsymbol{\alpha} = \mathbf{0.01}$ and the corresponding $\mathbf{Z}_{\frac{\alpha}{2}} = \mathbf{Z}_{0.005} = \mathbf{2.575}$

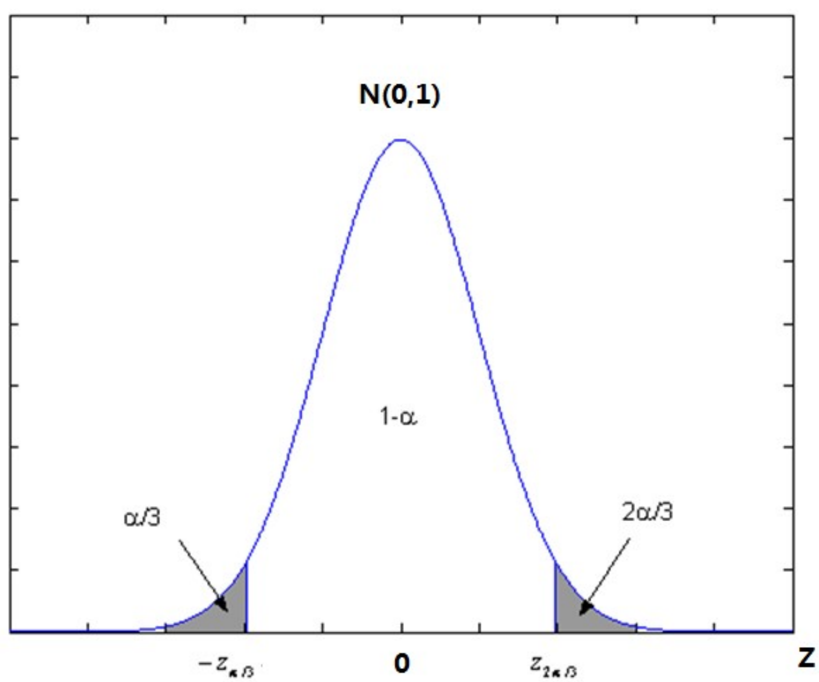
∴Recall the 100(1-α)% symmetric C.I. for μ is

$$[\overline{X} - Z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}, \overline{X} + Z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}]$$

*Please note that this CI is symmetric around $\overline{\boldsymbol{X}}$

The length of this CI is: $L_{sy} = 2 \cdot Z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$

(4) Now we derive a non-symmetrical CI:



$P(-Z_{\alpha/3} \le Z \le Z_{2/3\alpha}) = 1-\alpha$

100(1-α)% C.I. for μ

$\Rightarrow [\overline{X} - Z_{2/3\alpha} \cdot \frac{\sigma}{\sqrt{n}}, \overline{X} + Z_{1/3\alpha} \cdot \frac{\sigma}{\sqrt{n}}]$

Compare the lengths of the C.I.'s, one can prove theoretically

that: $L = (Z_{\alpha/3} + Z_{2/3\alpha}) \cdot \frac{\sigma}{\sqrt{n}} > L_{sy} = 2 \cdot Z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$

You can try a few numerical values for α, and see for yourself. For example,

$\boldsymbol{\alpha = 0.05}$

Theorem: Let $f(y)$ be a unimodal pdf. If the interval satisfies

(i) $\int_a^b f(y)\,dy = 1-\alpha$

(ii) $f(a) = f(b) > 0$

(iii) $a \le y^* \le b$, where y^* is a mode of $f(y)$, then $[a, b]$ is the shortest lengthed interval satisfying (i).

Note:

- y in the above theorem denotes the pivotal statistic upon which the CI is based
- $f(y)$ need not be symmetric: (graph)
- However, when $f(y)$ is symmetric, and $y^* = 0$, then $a = -b$. This is the case for the $N(0, 1)$ density and the t density.

Example. Large Sample Confidence interval for a population mean (*any population) and a population proportion p

<Theorem> Central Limit Theorem

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \xrightarrow{n \to \infty} N(0,1)$$

When n is large enough, we have

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

That means Z follows approximately the normal (0,1) distribution.

Application #1. Inference on μ when the population distribution is unknown but the sample size is large

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

By Slutsky's Theorem We can also obtain another pivotal quantity when σ is unknown by plugging the sample standard deviation S as follows:

$$Z = \frac{\bar{X} - \mu}{S/\sqrt{n}} \dot{\sim} N(0,1)$$

We subsequently obtain the $100(1-\alpha)\%$ C.I. using the second P.Q. for μ: $\bar{X} \pm Z_{\alpha/2}\frac{S}{\sqrt{n}}$

Application #2. Inference on one population proportion p when the population is Bernoulli(p) ***

Let $X_i \overset{i.i.d.}{\sim} Bernoulli(p),\ i = 1, \cdots, n$, please find the 100(1-α)% CI for p.

Point estimator : $\hat{p} = \bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$ (ex. $n = 1000$, $\hat{p} = 0.6$)

Our goal: derive a 100(1-α)% C.I. for p

Thus for the Bernoulli population, we have:

$$\mu = E(X) = p$$

$$\sigma^2 = Var(X) = p(1-p)$$

Thus by the CLT we have:

$$Z = \frac{\bar{X} - p}{\sqrt{\frac{p(1-p)}{n}}} \dot{\sim} N(0,1)$$

Furthermore, we have for this situation: $\bar{X} = \hat{p}$

Therefore we obtain the following pivotal quantity Z for p:

$$Z = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \dot{\sim} N(0,1)$$

By Slustky's theorem, we can replace the population proportion in the denominator with the sample proportion and obtain another pivotal quantity for p:

$$Z^* = \frac{\hat{p} - p}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} \dot{\sim} N(0,1)$$

 # Thus the $100(1-\alpha)\%$ (approximate, or large sample) C.I. for p based on the second pivotal quantity Z^* is:

$$P(-z_{\alpha/2} \le Z^* \le z_{\alpha/2}) = 1-\alpha$$

$$P(-z_{\alpha/2} \le \frac{\hat{p} - p}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} \le z_{\alpha/2}) = 1-\alpha$$

$$P(-\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \le -p \le -\hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}) = 1-\alpha$$

$$P(\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \le p \le \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}) = 1-\alpha$$

=> The $100(1-\alpha)\%$ large sample C.I. for p is

$$[\hat{p} - Z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + Z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}].$$

CLT => n large usually means $n \ge 30$

special case for the inference on p based on a Bernoulli population. The sample size n is large means

Let $X = \sum_{i=1}^{n} X_i$, large sample means:

$n\hat{p} = X \ge 5$ (*Here X= total # of 'S'), and

$n(1-\hat{p}) = n - X \ge 5$ (*Here n-X= total # of 'F')

Example: normal population, σ^2 unknown

1. Point estimation : $\overline{X} \sim N(\mu, \frac{\sigma^2}{n})$
2. $Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$
3. **Theorem.** Sampling from normal population
   a. $Z \sim N(0,1)$
   b. $W = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$
   c. Z and W are independent.

Definition. $T = \frac{Z}{\sqrt{W/(n-1)}} = \frac{\overline{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}$

------ Derivation of CI, normal population, σ^2 is unknown ------

$\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$ is not a pivotal quantity.

$\bar{X} - \mu \sim N(0, \frac{\sigma^2}{n})$ is not a pivotal quantity.

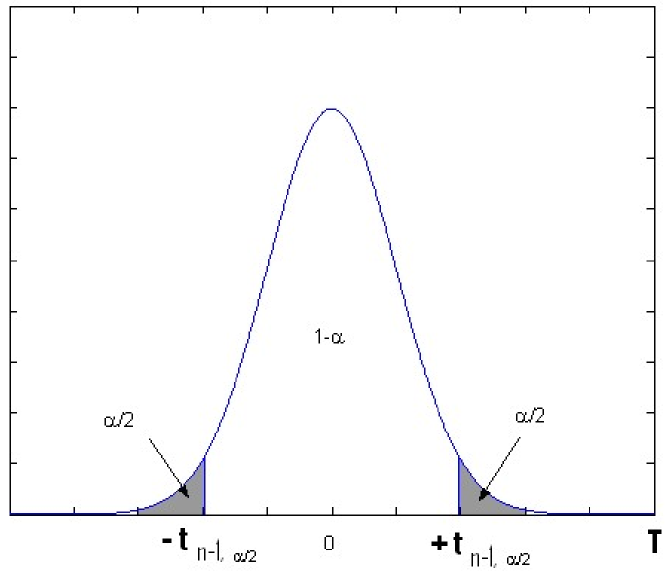
$Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \sim N(0,1)$ is not a pivotal quantity.

Remove σ !!!

Therefore $\boxed{T = \frac{\bar{X} - \mu}{S / \sqrt{n}} \sim t_{n-1}}$ is a pivotal quantity.

Now we will use this pivotal quantity to derive the 100(1-α)% confidence interval for μ.

We start by plotting the pdf of the t-distribution with n-1 degrees of freedom as follows:



The above pdf plot corresponds to the following probability statement:

$$P(-t_{n-1,\alpha/2} \le T \le t_{n-1,\alpha/2}) = 1-\alpha$$

$$=> \ P(-t_{n-1,\alpha/2} \le \frac{\bar{X}-\mu}{S/\sqrt{n}} \le t_{n-1,\alpha/2}) = 1-\alpha$$

$$=> \ P(-t_{n-1,\alpha/2}\frac{S}{\sqrt{n}} \le \bar{X}-\mu \le t_{n-1,\alpha/2}\frac{S}{\sqrt{n}}) = 1-\alpha$$

$$=> \ P(-\bar{X}-t_{n-1,\alpha/2}\frac{S}{\sqrt{n}} \le -\mu \le -\bar{X}+t_{n-1,\alpha/2}\frac{S}{\sqrt{n}}) = 1-\alpha$$

$$=> \ P(\bar{X}+t_{n-1,\alpha/2}\frac{S}{\sqrt{n}} \ge \mu \ge \bar{X}-t_{n-1,\alpha/2}\frac{S}{\sqrt{n}}) = 1-\alpha$$

$$=> \ P(\bar{X}-t_{n-1,\alpha/2}\frac{S}{\sqrt{n}} \le \mu \le \bar{X}+t_{n-1,\alpha/2}\frac{S}{\sqrt{n}}) = 1-\alpha$$

=> Thus the $100(1-\alpha)\%$ C.I. for μ when σ^2 is unknown is

$[\bar{X}-t_{n-1,\alpha/2}\frac{S}{\sqrt{n}}, \bar{X}+t_{n-1,\alpha/2}\frac{S}{\sqrt{n}}]$. (*Please note that $t_{n-1,\alpha/2} \ge Z_{\alpha/2}$)

Example. Inference on 2 population means, when both populations are normal. We have 2 independent samples, the population variances are unknown but equal ($\sigma_1^2 = \sigma_2^2 = \sigma^2$)$\Rightarrow$ pooled-variance t-test.

Data: $X_1, \ldots, X_{n_1} \overset{iid}{\sim} N(\mu_1, \sigma_1^2)$

$Y_1, \ldots, Y_{n_2} \overset{iid}{\sim} N(\mu_2, \sigma_2^2)$

Goal: Compare μ_1 and μ_2

1) **Point estimator:**

$$\widehat{\mu_1 - \mu_2} = \bar{X} - \bar{Y} \sim N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right) = N\left(\mu_1 - \mu_2, \left(\frac{1}{n_1} + \frac{1}{n_2}\right)\sigma^2\right)$$

2) **Pivotal quantity:**

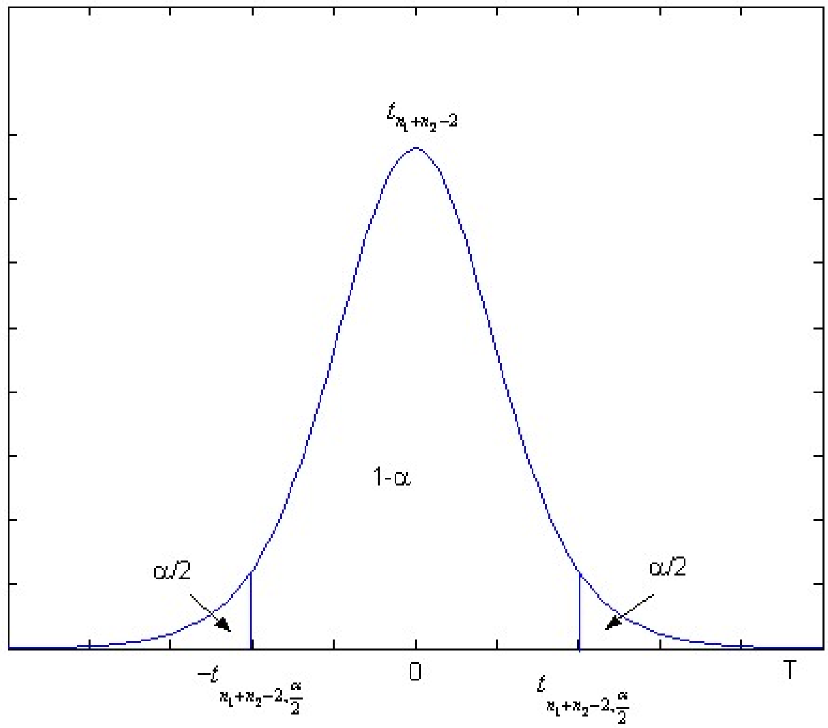
$$T = \frac{Z}{\sqrt{\dfrac{W}{n_1 + n_2 - 2}}} = \frac{\dfrac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sigma \cdot \sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}}}{\sqrt{\dfrac{\dfrac{(n_1 - 1)S_1^2}{\sigma^2} + \dfrac{(n_2 - 1)S_2^2}{\sigma^2}}{n_1 + n_2 - 2}}} = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{S_p \sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}} \sim t_{n_1 + n_2 - 2}$$

.

where $S_p^2 = \dfrac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$ is the pooled variance.

This is the PQ of the inference on the parameter of interest $(\mu_1 - \mu_2)$

3) Confidence Interval for $(\mu_1 - \mu_2)$



$$1-\alpha = P(-t_{n_1+n_2-2,\frac{\alpha}{2}} \le T \le t_{n_1+n_2-2,\frac{\alpha}{2}})$$

$$1-\alpha = P\left(-t_{n_1+n_2-2,\frac{\alpha}{2}} \le \frac{(\bar{X}-\bar{Y})-(\mu_1-\mu_2)}{S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \le t_{n_1+n_2-2,\frac{\alpha}{2}}\right)$$

$$1-\alpha = P(-t_{n_1+n_2-2,\frac{\alpha}{2}} \cdot S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}} \le (\bar{X}-\bar{Y})-(\mu_1-\mu_2) \le t_{n_1+n_2-2,\frac{\alpha}{2}} \cdot S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}})$$

$$1-\alpha = P(\bar{X}-\bar{Y}-t_{n_1+n_2-2,\frac{\alpha}{2}} \cdot S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}} \le \mu_1-\mu_2 \le \bar{X}-\bar{Y}+t_{n_1+n_2-2,\frac{\alpha}{2}} \cdot S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}})$$

$\therefore$ This is the $100(1-\alpha)\%$ C.I for $(\mu_1 - \mu_2)$

2. General approach for deriving CI's :

Inverting a Test

Hypothesis test: under a given $H_0 : \theta = \theta_0$,

$$P_{\theta_0}(T_n \in R) = \alpha \Leftrightarrow P_{\theta_0}(T_n \notin R) = 1 - \alpha$$

where T_n is a test statistic.

We can use this to construct a (1 - α) confidence interval:

• Define acceptance region $A = \mathbb{R} \backslash R$.

• If you *fix* α, but vary the null hypothesis θ_0, then you obtain $R(\theta_0)$, a rejection region for each θ_0 such that, by construction:

$$\forall \theta_0 \in \Theta: P_{\theta_0}\left(T_n \notin R(\theta_0)\right) = P_{\theta_0}\left(T_n \in A(\theta_0)\right) = 1 - \alpha$$

• Now, for a given sample $X \sim \equiv X_1, \ldots, X_n$, consider the set

$$C(\underline{X}) \equiv \{\theta: T_n(\underline{X}) \in A(\theta)\}$$

By construction:

$$P_\theta\left(\theta \in C(\underline{X})\right) = P_\theta\left(T_n(\underline{X}) \in A(\theta)\right), \forall \theta \in \Theta$$

Therefore, $C(\underline{X})$ is a (1 - α) confidence interval for θ.

• The confidence interval $C(\underline{X})$ is the set of θ's such that, for the given data $\underline{X}$ and for each $\theta_0 \in C(\underline{X})$, you would not be able to reject the null hypothesis
$H_0 : \theta = \theta_0$.

• In hypothesis testing, the acceptance region is the set of $\underline{X}$ which are very likely for a fixed θ_0.

In interval estimation, the confidence interval is the set of θ's which make $\underline{X}$ very likely, for a fixed $\underline{X}$.

Example: $X_1, \ldots, X_n \sim i.i.d.\ N(\mu, 1)$.
We want to construct a 95% CI for μ by inverting the Z-test.

• We know that, under each null hypothesis $H_0: \mu = \mu_0$,

$$\sqrt{n}(\overline{X}_n - \mu_0) \sim N(0,1)$$

• Hence, for each μ_0, a 95% acceptance region is

$$\{-1.96 \le \sqrt{n}(\overline{X}_n - \mu_0) \le 1.96\}$$

$$\Leftrightarrow \left\{\overline{X}_n - 1.96\frac{1}{\sqrt{n}} \le \mu_0 \le \overline{X}_n + 1.96\frac{1}{\sqrt{n}}\right\}$$

• Now consider what happens when we invert one-sided test.
Consider the hypotheses $H_0: \mu \le \mu_0$vs. $H_a: \mu > \mu_0$. Then a 95% acceptance region is

$$\{\sqrt{n}(\overline{X}_n - \mu_0) \le 1.645\}$$

$$\Leftrightarrow \left\{\mu_0 \ge \overline{X}_n - 1.645\frac{1}{\sqrt{n}}\right\}$$

Quiz:

Let the random sample $X_1, X_2, \ldots X_n \sim N(\mu, \sigma^2)$, where both μ and σ^2 are unknown

(1) Derive the 100(1-α)% CI for σ^2 using the pivotal quantity method;
(2) Derive the 100(1-α)% CI for σ^2 by inverting the two sided test $H_0: \sigma^2 = \sigma_0^2 \quad vs \quad H_a: \sigma^2 \ne \sigma_0^2$
(3) Are your CIs in (1) and (2) the same?
(4) Are your CI(s) optimal? If not, please derive the optimal CI.