Econometrics: Lecture 6

Hang Miao

Rutgers University

January 21, 2021

Overview

1. Review
   - Types of Convergence
   - Law of Large Number
   - Central Limit Theorem
2. Statistical Inference on Population Mean
   - Hypothesis Testing
   - P-Value
   - Confidence intervals
3. Statistical Inference on Means of Two Different Population
   - Hypothesis Testing
   - P-Value for Mean Difference
   - Confidindce Interval for Mean Difference

Review

Types of Convergence

Convergence of Random Variable

Throughout the following, we assume that $\{X_n\}$ is a sequence of random variables, and X is a random variable, and all of them are defined on the same probability space $(\Omega, \mathcal{F}, \Pr)$

Definition (Converge Almost Surely)

$\{X_n\}$ converge almost surely to X if $\Pr(\lim_{n\to\infty} X_n = X) = 1$,
or $\Pr\left(\omega \in \Omega : \lim_{n\to\infty} X_n(\omega) \neq X(\omega)\right) = 0$
Denoted as $X_n \xrightarrow{\text{a.s.}} X$

Definition (Converge in Probability)

$\{X_n\}$ converge in probability to X if $\lim_{n\to\infty} \Pr\left(|X_n - X| \geq \varepsilon\right) = 0$.
Denoted as $X_n \xrightarrow{P} X$, or $\operatorname*{plim}_{n\to\infty} X_n = X$.

Types of Convergence

Definition (Converge in Distribution)

$\{X_n\}$ converges in distribution, or converge weakly, or converge in law to X if

$$\lim_{n\to\infty} F_n(x) = F(x), \quad \forall x \in \mathbb{R}$$

Denoted as $X_n \xrightarrow{d} X$. where $F_n(\cdot)$ and $F(\cdot)$ are the cumulative distribution functions (CDF) of random variables $\{X_n\}$ and X, respectively.

Definition (Converge in Mean)

Given a real number $r \geq 1$, $\{X_n\}$ converges in the **r-th mean (or in the L^r-norm)** to X, if the r-th absolute moments $\mathrm{E}(| X_n |^r)$ and $\mathrm{E}(| X |^r)$ of $\{X_n\}$ and X exist, and

$$\lim_{n\to\infty} \mathrm{E}\left(|X_n - X|^r\right) = 0,$$

Denoted as $X_n \xrightarrow{L^r} X$.

Relationship Between Different Types of Convergence

- Almost sure convergence implies convergence in probability:

$$X_n \xrightarrow{\text{a.s.}} X \quad \Rightarrow \quad X_n \xrightarrow{p} X$$

- Convergence in probability implies there exists a sub-sequence k_n s.t X_{k_n} converges almost surely to X:

$$X_n \xrightarrow{p} X \quad \Rightarrow \quad X_{k_n} \xrightarrow{\text{a.s.}} X$$

- Convergence in probability implies convergence in distribution:

$$X_n \xrightarrow{p} X \quad \Rightarrow \quad X_n \xrightarrow{d} X$$

- Convergence in r-th order mean implies convergence in probability:

$$X_n \xrightarrow{L^r} X \quad \Rightarrow \quad X_n \xrightarrow{p} X$$

Law of Large Number

Assumption

Suppose $\{X_n\}$ is a sequence of random variables. The sample mean of $\{X_n\}$ is defined as follow

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Definition (Weak Law of Large Numbers (Chebychev's LLN))

$$\lim_{n\to\infty} \mathrm{E}\left(\overline{X}_n\right) = \mu, \ \lim_{n\to\infty} \mathrm{Var}\left(\overline{X}_n\right) = 0 \implies \overline{X}_n \xrightarrow{p} \mu$$

Definition (Strong Law of Large Numbers (Kolmogorov's LLN))

$$\lim_{n\to\infty} \mathrm{E}\left(\overline{X}_n\right) = \mu, \ \{X_n\}\text{i.i.d.} \implies \overline{X}_n \xrightarrow{a.s} \mu$$

Central Limit Theorem

Assumption

The central limit theorem (CLT) establishes that, in some situations, when independent random variables are added, their properly normalized sum tends toward a normal distribution (informally a "bell curve") even if the original variables themselves are not normally distributed.

Definition (Lindeberg–Levy CLT)

Suppose $\{X_n\}$ is a sequence of i.i.d. random variables with $\mathrm{E}[X_i] = \mu$ and $\mathrm{Var}[X_i] = \sigma^2 < \infty$. Then as $n \to \infty$,

$$\sqrt{n}(\overline{X}_n - \mu) \xrightarrow{d} N(0, \sigma^2)$$

Or equivalently,

$$\frac{\overline{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}} \xrightarrow{d} N(0, 1)$$

Statistical Inference on Population Mean

Hypothesis Testing for Population Mean with Known Variance σ^2

Motivation

Equipped with Central Limit Theorem we could make the statistical inference about the population mean from sample mean by Hypothesis Testing introduced before.

Prerequisite Setting

Suppose there is a sequence of i.i.d. observed sample $\{X_n\}$ from population. The variance of the underlying population is known as σ^2. The population mean μ is unknown.

Hypothesis Testing

We could propose certain value μ_0 for μ as null hypothesis $H_0 : \mu = \mu_0$ then the alternative hypothesis should be $H_1 : \mu \neq \mu_0$.
The object is to see how credible H_0 is and how likely we make a type I mistake if we reject the H_0 while in favor of the alternative H_1

Hypothesis Testing for Population Mean with Known Variance σ^2

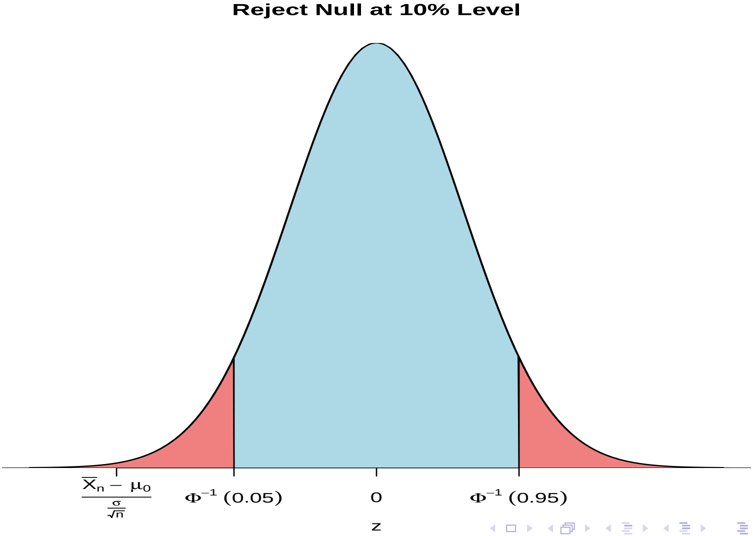
Statistics

According to the central limit theorem (CLT), the Z-score statistic $Z_n = \frac{\overline{X}_n - \mu_0}{\frac{\sigma}{\sqrt{n}}}$ converge in distribution to standard normal $N(0, 1)$. If the sample size is large enough, Z_n approximately follows standard normal distribution. As a result, we would expect to see Z_n near around the mean 0 rather than deviated far away from mean 0. This is because of the bell-shape of the Normal Probability Density Function put a high weight on center while less weight on tails.

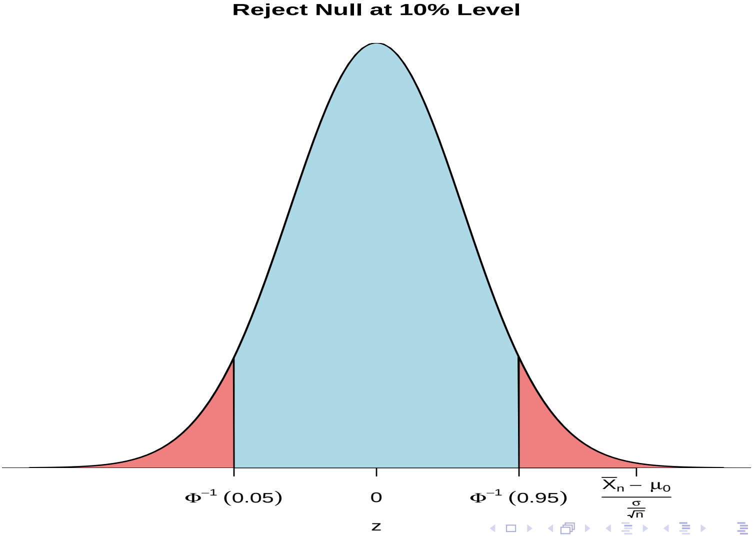
Reject the Null

Suppose we do spot the Z_n lies on the tails under the Null Hypothesis $H_0 : \mu = \mu_0$. Since such kind of event is less likely to happen, the null H_0 is less credible. Depending on the incredibility of H_0, we could reject it at three regularly used different significant level: $10\%, 5\%, 1\%$

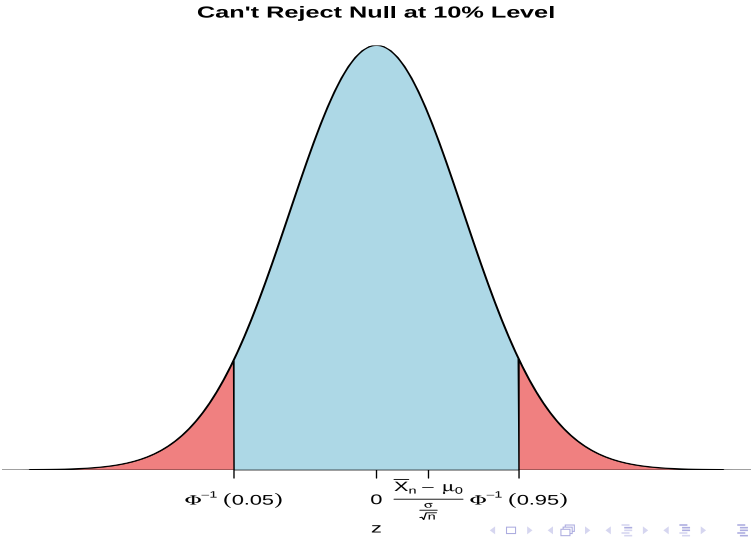
Reject the Null at 10% Significant Level: Case I



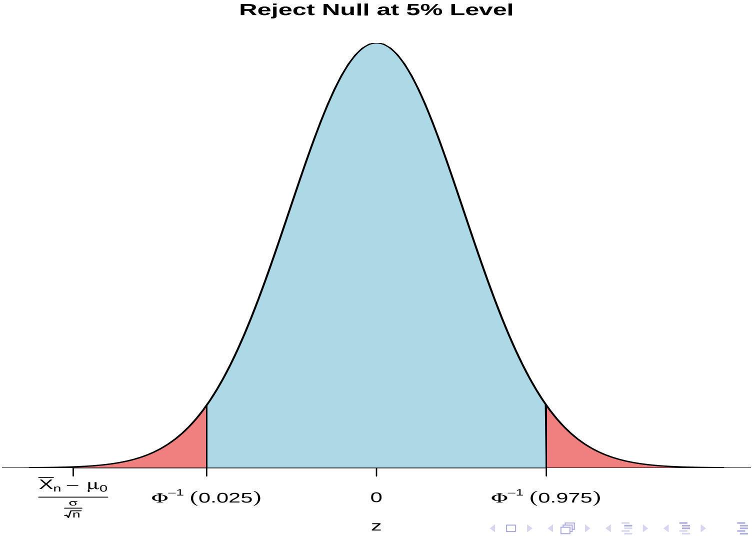
Reject the Null at 10% Significant Level: Case II



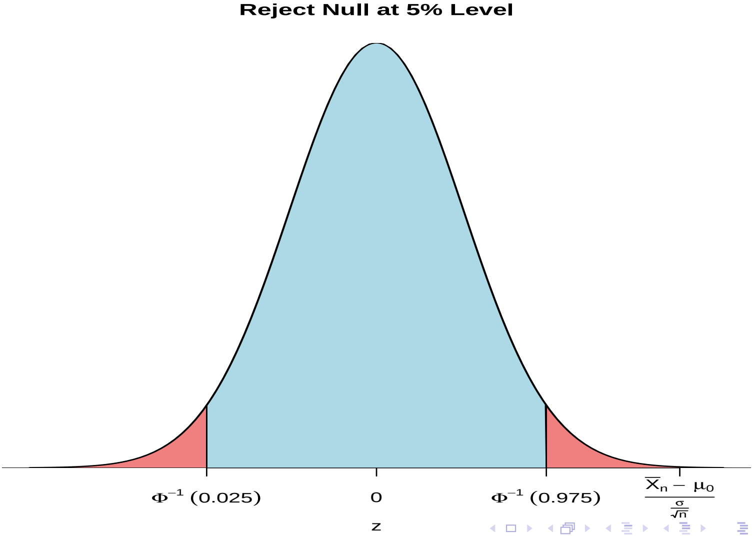
Can't Reject the Null at 10% Significant Level



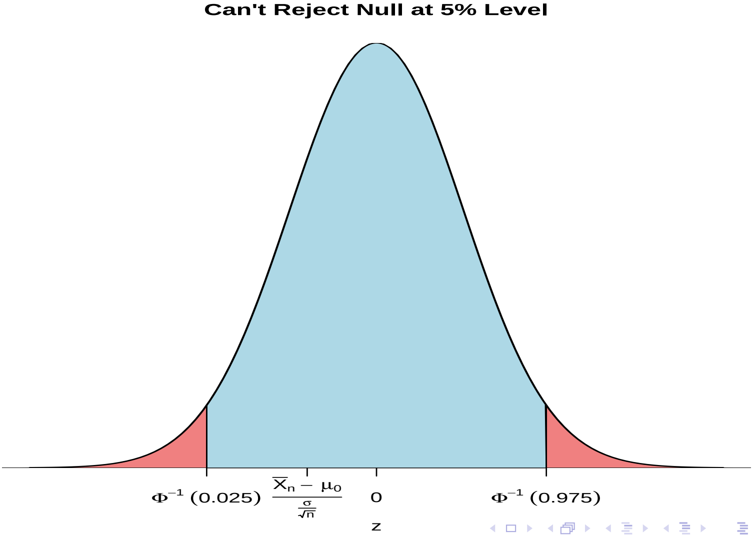
Reject the Null at 5% Significant Level: Case I



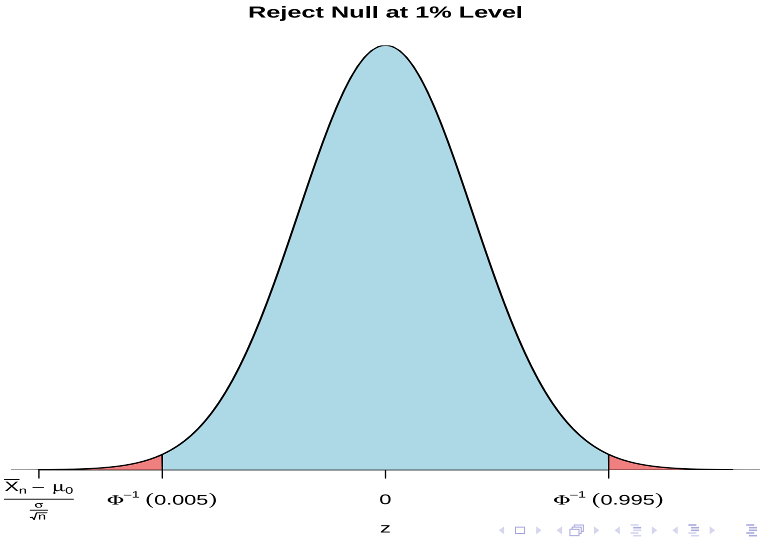
Reject the Null at 5% Significant Level: Case II



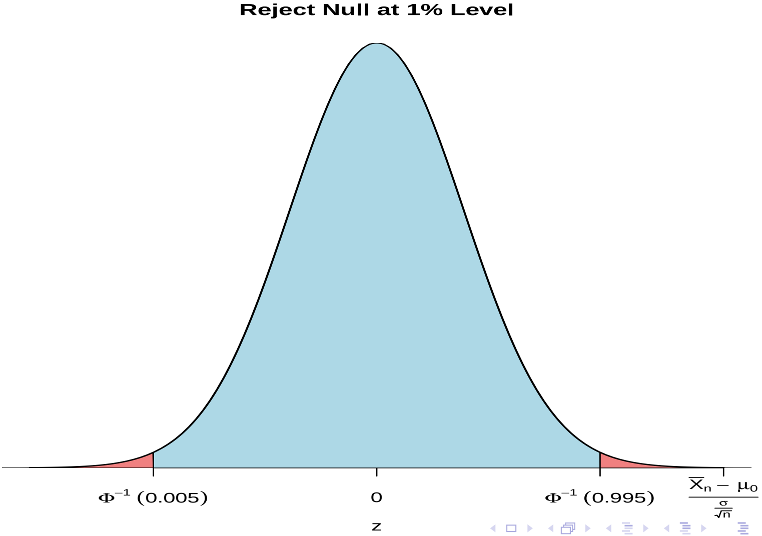
Can't Reject the Null at 5% Significant Level



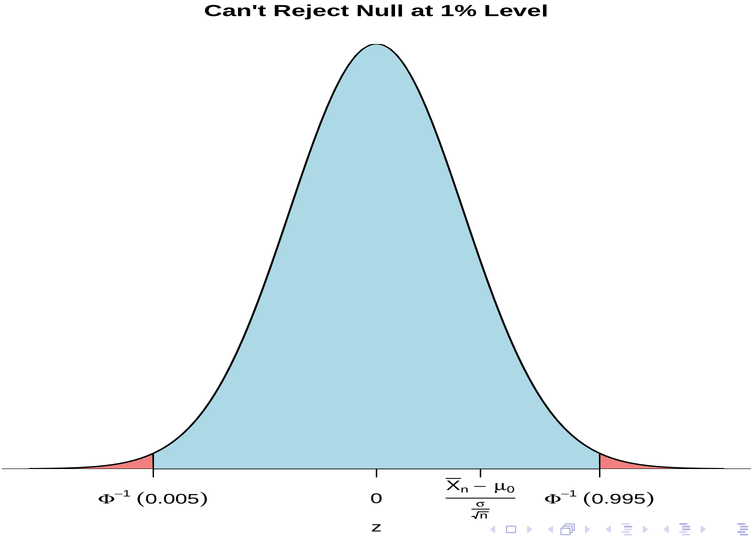
Reject the Null at 1% Significant Level: Case I



Reject the Null at 1% Significant Level: Case II



Can't Reject the Null at 1% Significant Level



P-Value

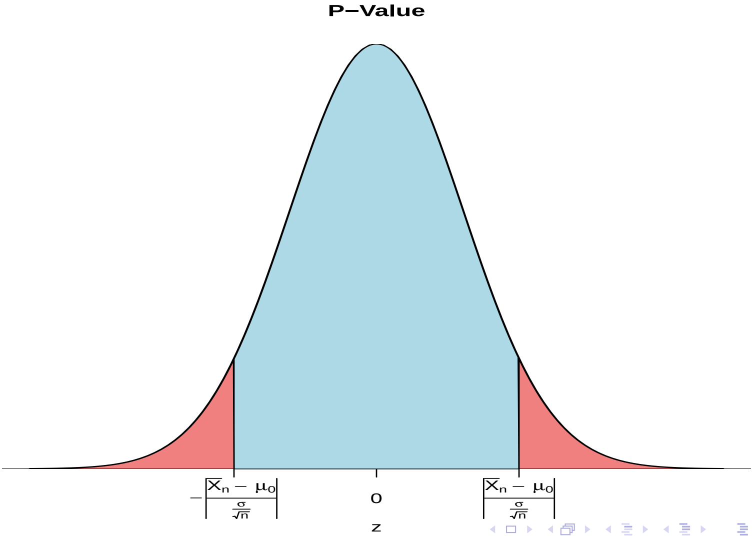
Motivation

From the above figures, we see that the Null H_0 could be reject at a lower significant level if it could be rejected at a higher significant level. It is inconvenient to compare the value of Z_n with each critical value $\Phi^{-1}(\cdot)$ at different significant level.

Definition (P-Value)

In order to resolve the inconvenience mentioned above. The p-value is introduced. It is the probability that we could spot a statistic as extreme as the current spotted one Z_n. Under the null hypothesis $H_0 : \mu = \mu_0$. The p-value is the total area of two tails under PDF of Normal(0,1) started $\pm Z_n$. i.e it is $2\Phi(Z_n)$ if $Z_n \leq 0$, or $2\big(1 - \Phi(Z_n)\big)$ if $Z_n \geq 0$. In R, the p-value is calculated by the CDF of normal function
$2 * pnorm(Z_n,\ lower.tail = T)$ if if $Z_n \leq 0$, or
$2 * pnorm(Z_n,\ lower.tail = F)$ if $Z_n \geq 0$. The null hypothesis is rejected at significant level α, if the p-value less than α.

P-Value


P−Value
$-\left|\frac{\overline{X}_n - \mu_0}{\frac{\sigma}{\sqrt{n}}}\right|$
0
$\left|\frac{\overline{X}_n - \mu_0}{\frac{\sigma}{\sqrt{n}}}\right|$
z

Confidence Intervals

Definition (Confidence intervals)

Confidence Interval (CI) is a type of interval estimate, computed from the statistics of the observed data, that might contain the true value of an unknown population parameter.

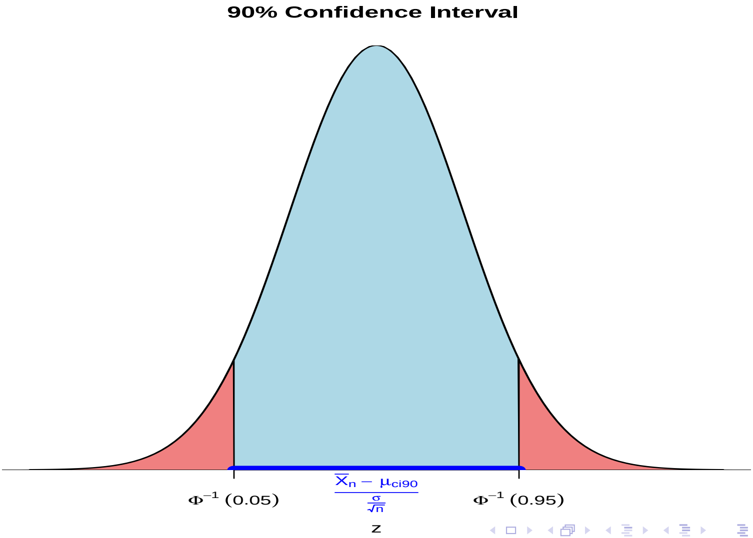
Confidence Intervals for the Population Mean

$$99\% \text{ confidence interval for } \mu_X = \left[\overline{X}_n \pm 2.58 \times \frac{\sigma}{\sqrt{n}}\right]$$

$$95\% \text{ confidence interval for } \mu_X = \left[\overline{X}_n \pm 1.96 \times \frac{\sigma}{\sqrt{n}}\right]$$

$$90\% \text{ confidence interval for } \mu_X = \left[\overline{X}_n \pm 1.64 \times \frac{\sigma}{\sqrt{n}}\right]$$

90% Confidence Interval



90% Confidence Interval

We want the population mean μ_{ci90} takes the value such that statistic Z_n lie in the region where the area (probability) under the PDF is 90%.

$$\Phi^{-1}(0.05) \leq \frac{\bar{X}_n - \mu_{ci90}}{\frac{\sigma}{\sqrt{n}}} \leq \Phi^{-1}(0.95)$$

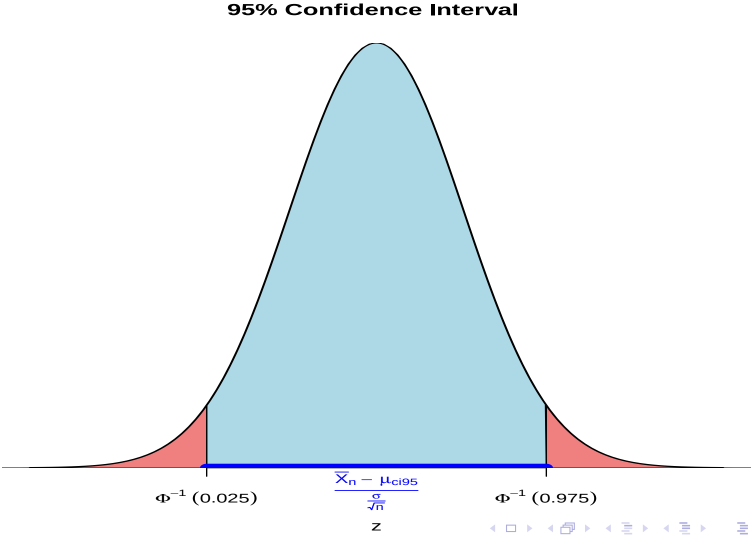
$$\frac{\sigma}{\sqrt{n}}\Phi^{-1}(0.05) \leq \bar{X}_n - \mu_{ci90} \leq \frac{\sigma}{\sqrt{n}}\Phi^{-1}(0.95)$$

$$\bar{X}_n - \frac{\sigma}{\sqrt{n}}\Phi^{-1}(0.95) \leq \mu_{ci90} \leq \bar{X}_n - \frac{\sigma}{\sqrt{n}}\Phi^{-1}(0.05)$$

Since normal distribution is symmetrical, $\Phi^{-1}(0.05) = -\Phi^{-1}(0.95) = 1.64$. Therefore,

$$\mu_{ci90} \in \left[\bar{X}_n - 1.64\frac{\sigma}{\sqrt{n}},\ \bar{X}_n + 1.64\frac{\sigma}{\sqrt{n}}\right]$$

95% Confidence Interval



95% Confidence Interval

We want the population mean μ_{ci95} takes the value such that statistic Z_n lie in the region where the area (probability) under the PDF is 95%.

$$\Phi^{-1}(0.025) \leq \frac{\bar{X}_n - \mu_{ci95}}{\frac{\sigma}{\sqrt{n}}} \leq \Phi^{-1}(0.975)$$

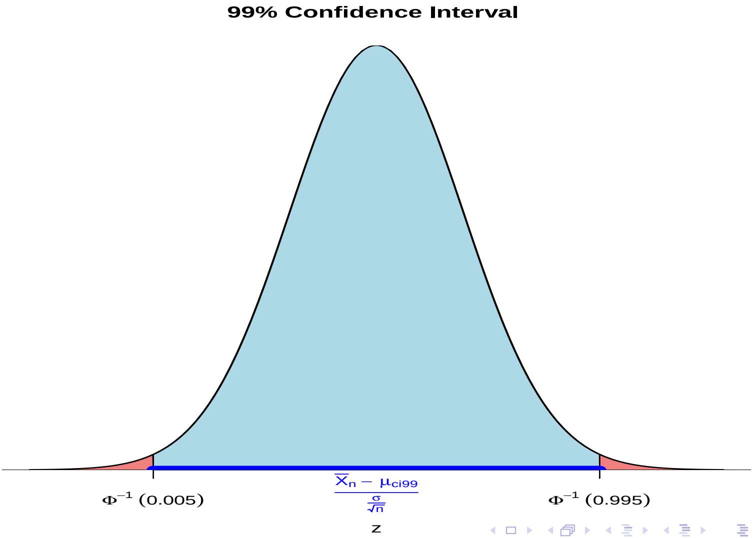
$$\frac{\sigma}{\sqrt{n}}\Phi^{-1}(0.025) \leq \bar{X}_n - \mu_{ci95} \leq \frac{\sigma}{\sqrt{n}}\Phi^{-1}(0.975)$$

$$\bar{X}_n - \frac{\sigma}{\sqrt{n}}\Phi^{-1}(0.975) \leq \mu_{ci95} \leq \bar{X}_n - \frac{\sigma}{\sqrt{n}}\Phi^{-1}(0.025)$$

Since normal distribution is symmetrical, $\Phi^{-1}(0.025) = -\Phi^{-1}(0.975) = 1.96$. Therefore,

$$\mu_{ci95} \in \left[\bar{X}_n - 1.96\frac{\sigma}{\sqrt{n}},\ \bar{X}_n + 1.96\frac{\sigma}{\sqrt{n}}\right]$$

99% Confidence Interval



99% Confidence Interval

We want the population mean μ_{ci99} takes the value such that statistic Z_n lie in the region where the area (probability) under the PDF is 99%.

$$\Phi^{-1}(0.005) \leq \frac{\bar{X}_n - \mu_{ci99}}{\frac{\sigma}{\sqrt{n}}} \leq \Phi^{-1}(0.995)$$

$$\frac{\sigma}{\sqrt{n}}\Phi^{-1}(0.005) \leq \bar{X}_n - \mu_{ci99} \leq \frac{\sigma}{\sqrt{n}}\Phi^{-1}(0.995)$$

$$\bar{X}_n - \frac{\sigma}{\sqrt{n}}\Phi^{-1}(0.995) \leq \mu_{ci99} \leq \bar{X}_n - \frac{\sigma}{\sqrt{n}}\Phi^{-1}(0.005)$$

Since normal distribution is symmetrical,
$\Phi^{-1}(0.005) = -\Phi^{-1}(0.995) = 2.58$. Therefore,

$$\mu_{ci99} \in \left[\bar{X}_n - 2.58\frac{\sigma}{\sqrt{n}},\ \bar{X}_n + 2.58\frac{\sigma}{\sqrt{n}}\right]$$

Statistical Inference on Means of Two Different Population

Hypothesis Testing for Mean Difference of Two Different Population with Known Variance σ_1^2 and σ_2^2

Motivation

In many cases, a researcher is interesting in gathering information about two populations in order to compare them. As in statistical inference for one population parameter: Hypothesis tests of significance, P-Value and confidence intervals are useful statistical tools for the difference between two population parameters.

Prerequisite Setting

Suppose there are two sequence of i.i.d. observed sample $\{X_n\}$ and $\{Y_m\}$ from population. The variance of the two underlying population is known as σ_1^2 and σ_2^2. The population mean μ_1 and μ_2 are unknown. We are interested to know whether there is there is a significant difference between the means of the two populations.

Hypothesis Testing for Mean Difference of Two Different Population with Known Variance σ_1^2 and σ_2^2

Hypothesis Testing

Consider μ_1 and μ_2 as their true population mean, we are interest to test the null hypothesis $H_0 : \mu_1 = \mu_2$ or $H_0 : \mu_1 - \mu_2 = 0$ against the alternative hypothesis should be $H_1 : \mu_1 \neq \mu_2$ or $H_1 : \mu_1 - \mu_2 \neq 0$
The object is to see how credible H_0 is and how likely we make a type I mistake if we reject the H_0 while in favor of the alternative H_1

Statistics

Let the Z-score statistics constructed in the following way:

$$Z = \frac{\left(\overline{X}_n - \overline{Y}_m\right) - \left(\mu_1 - \mu_2\right)}{\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}}$$

We claim that sample size is large enough, Z approximately follows standard normal distribution by the Central Limit Theorem.

Hypothesis Testing for Population Mean with Known Variance σ^2

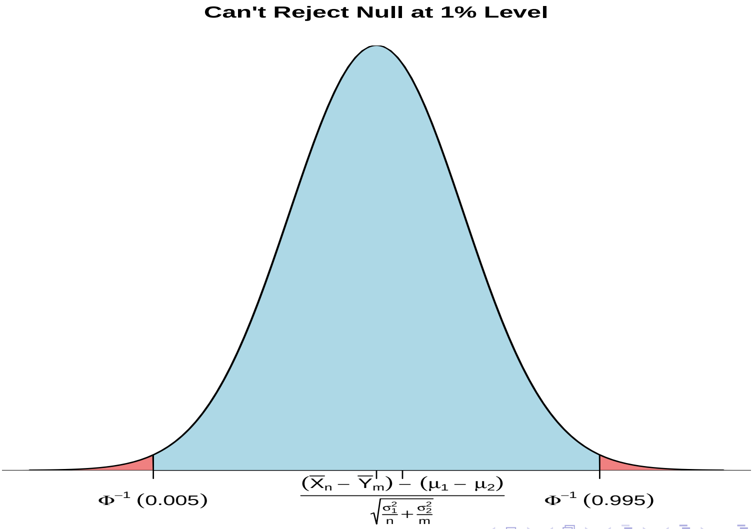
To apply the central limit theorem, we need to check the statistics Z satisfies the condition as follow:

- $\mathrm{E}\left[\overline{X}_n - \overline{Y}_m\right] = \mu_1 - \mu_2$
- $\mathrm{Var}\left[\overline{X}_n - \overline{Y}_m\right] = \frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}$

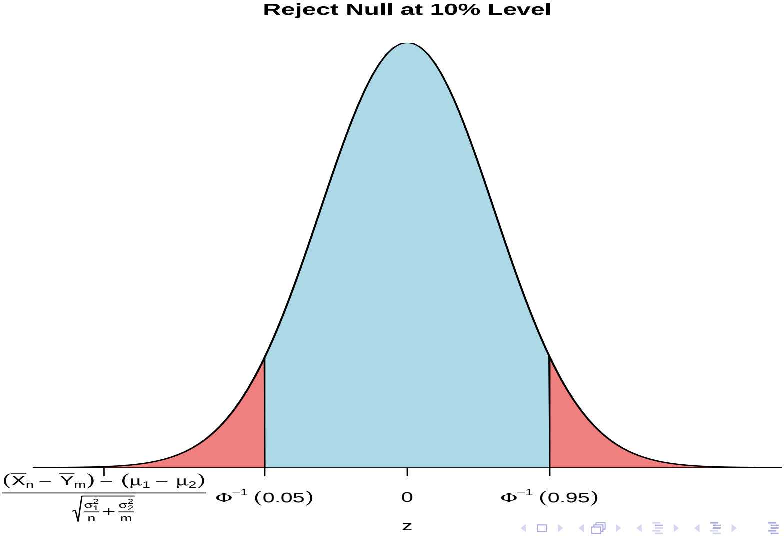
Proof:

$$
\begin{aligned}
\mathrm{E}\left[\overline{X}_n - \overline{Y}_m\right] &= \mathrm{E}\left[\overline{X}_n\right] - \mathrm{E}\left[\overline{Y}_m\right] = \mu_1 - \mu_2 \\
\mathrm{Var}\left[\overline{X}_n - \overline{Y}_m\right] &= \mathrm{Cov}\left[\overline{X}_n - \overline{Y}_m, \overline{X}_n - \overline{Y}_m\right] \\
&= \mathrm{Cov}\left[\overline{X}_n, \overline{X}_n\right] + \mathrm{Cov}\left[\overline{X}_n, -\overline{Y}_m\right] \\
&\quad + \mathrm{Cov}\left[-\overline{Y}_m, \overline{X}_n\right] + \mathrm{Cov}\left[-\overline{Y}_m, -\overline{Y}_m\right] \\
&= \mathrm{Var}[\overline{X}_n] + \mathrm{Var}[\overline{Y}_m] = \frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}
\end{aligned}
$$

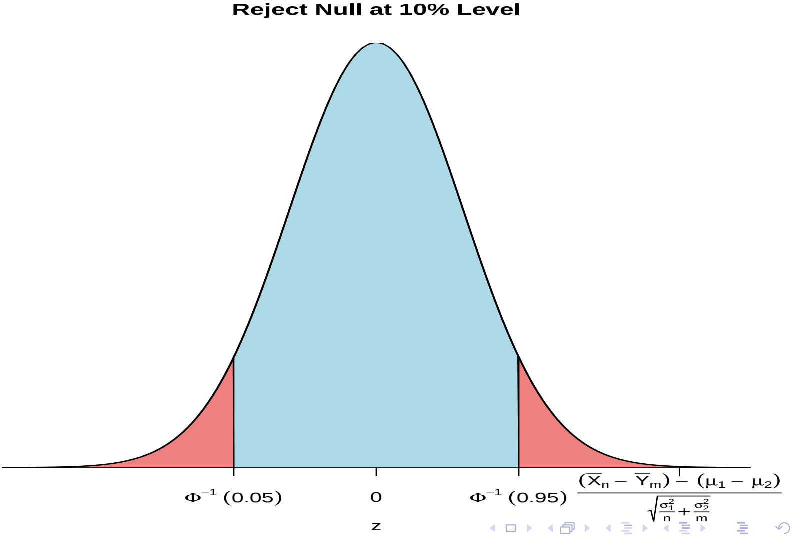
Can't Reject the Null at 1% Significant Level



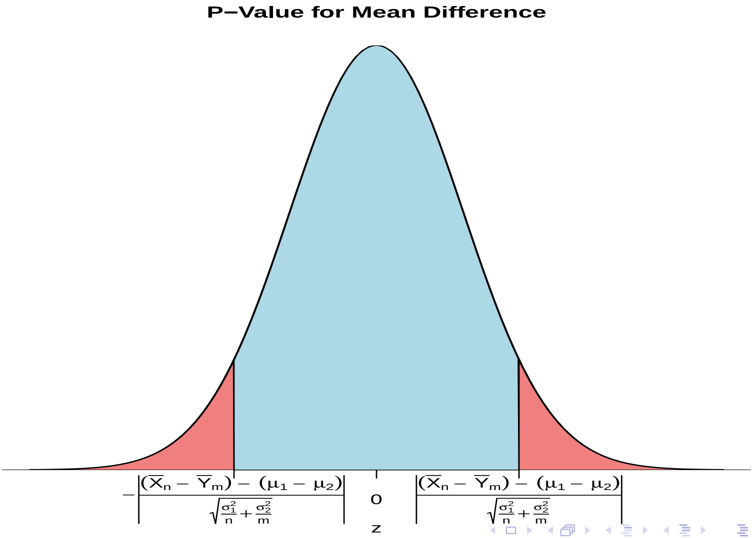
Reject the Null at 10% Significant Level: Case I



Reject the Null at 10% Significant Level: Case II

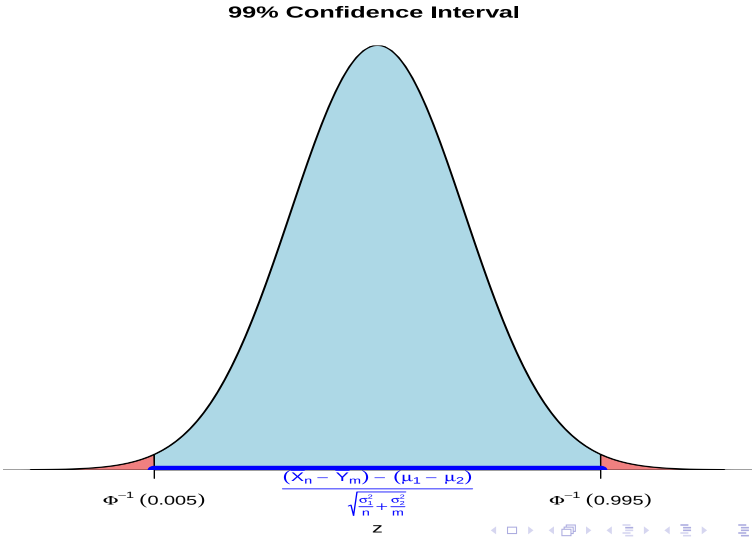


P-Value



P−Value for Mean Difference
$-\left|\frac{(\overline{X}_n - \overline{Y}_m) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n}+\frac{\sigma_2^2}{m}}}\right|$
0
$\left|\frac{(\overline{X}_n - \overline{Y}_m) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n}+\frac{\sigma_2^2}{m}}}\right|$
z

Confidince Interval for Mean Difference



99% Confidence Interval

We want the population mean μ_{ci99} takes the value such that statistic Z lie in the region where the area (probability) under the PDF is 99%.

$$\Phi^{-1}(0.005) \leq \frac{(\overline{X}_n - \overline{Y}_m) - \mu_{ci99}}{\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}} \leq \Phi^{-1}(0.995)$$

$$\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}\Phi^{-1}(0.005) \leq (\overline{X}_n - \overline{Y}_m) - \mu_{ci99} \leq \sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}\Phi^{-1}(0.995)$$

$$(\overline{X}_n - \overline{Y}_m) - \sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}\Phi^{-1}(0.995) \leq \mu_{ci99} \leq (\overline{X}_n - \overline{Y}_m) - \sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}\Phi^{-1}(0.005)$$

99% Confidence Interval

Since normal distribution is symmetrical,
$\Phi^{-1}(0.005) = -\Phi^{-1}(0.995) = 2.58$. Therefore,

$$(\overline{X}_n - \overline{Y}_m) - \sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}\Phi^{-1}(0.995) \leq \mu_{ci99} \leq (\overline{X}_n - \overline{Y}_m) + \sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}\Phi^{-1}(0.995)$$

$$\mu_{ci99} \in \left[(\overline{X}_n - \overline{Y}_m) - 2.58\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}},\ (\overline{X}_n - \overline{Y}_m) + 2.58\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}\right]$$