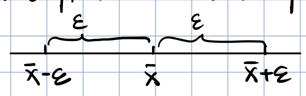
Sample size

Question: If we want to get confidence interval (CI) for population mean μ of a certain confidence level $1-\alpha$ and length 2ε, how large should the sample size n be?

Recall CI $[\bar{X}-\varepsilon, \bar{X}+\varepsilon]$ confidence level $1-\alpha$ if $\mathbb{P}(\mu \in \underbrace{[\bar{X}-\varepsilon, \bar{X}+\varepsilon]}_{\text{random interval}}) = 1-\alpha$



Basic idea:

Higher confidence level $1-\alpha$ ⟵ Make the length 2ε bigger

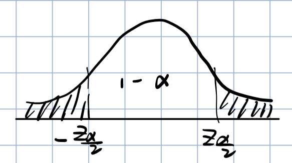
Higher confidence level $1-\alpha$ ⟵ Make the sample size n larger

Example $X \sim \mathcal{N}(\mu, \sigma^2)$ σ known what's a $100(1-\alpha)\%$ CI for μ?

$$\varepsilon \geq z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$$

standard derivation of $\bar{X}$ $\quad \frac{\sigma^2}{n} = \mathrm{Var}(\bar{X})$

$z_{\frac{\alpha}{2}}$: # standard derivations from μ.



$$\frac{\varepsilon}{z_{\frac{\alpha}{2}} \cdot \sigma} \geq \frac{1}{\sqrt{n}}$$

$$\frac{z_{\frac{\alpha}{2}} \cdot \sigma}{\varepsilon} \leq \sqrt{n}, \quad \left(\frac{z_{\frac{\alpha}{2}} \cdot \sigma}{\varepsilon}\right)^2 \leq n$$

To get CI for μ with confidence level $\geq 1-\alpha$

X normal	σ known	$z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} \leq \varepsilon$, or $n \geq \left(\frac{z_{\frac{\alpha}{2}} \cdot \sigma}{\varepsilon}\right)^2$
	σ unknown	$t_{\frac{\alpha}{2}}(n-1) \frac{S}{\sqrt{n}} \leq \varepsilon$
X not normal	σ known	approximately $n \geq \left(\frac{z_{\frac{\alpha}{2}} \cdot \sigma}{\varepsilon}\right)^2$

Problem (1) $X \sim \mathcal{N}(\mu, \sigma)$ want 95% CI for μ with length ≤ 0.1. What's the smallest sample size to guarantee this?

$\sigma^2 = 25 \quad \sigma = 5 \qquad 1-\alpha = 95\%$

$2\varepsilon = 0.1, \ \varepsilon = 0.05 \qquad z_{\frac{\alpha}{2}} \approx 1.96$

$$n \geq \left(\frac{z_{\frac{\alpha}{2}} \cdot \sigma}{\varepsilon}\right)^2$$

$$= \left(\frac{1.96 \times 5}{0.05}\right)^2 = (1.96 \times 100)^2 = 196^2 = 38416.$$

(2) What if length ≤ 0.01?

$n = \left(\frac{z_{\frac{\alpha}{2}} \cdot \sigma}{\varepsilon}\right)^2$ is proportional to $\frac{1}{\varepsilon^2}$ if α, σ fixed.

So if ε becomes $\frac{1}{10}$ of original, then n becomes 100 thus because $\frac{1}{(\frac{1}{10})^2} = 100$.

So $n = 3841600$.

(3) What if $\sigma = 0.5$ instead of 5? n proportional to σ^2 if α, ε fixed.

So, if σ becomes $\frac{1}{10}$ of original, then n because $\frac{1}{100}$ times

So $n = 384.16$ round up $n \geq 385$

Example: $X \sim N(\mu_X, \sigma_X^2)$, $Y \sim N(\mu_Y, \sigma_Y^2)$ want CI for $\mu_X - \mu_Y$ with confidence level $1-\alpha$ and length 2ε what should sample sizes n, m be?

$X_1, \dots, X_n$ $Y_1, \dots, Y_m$

Assume σ_X, σ_Y known.

$[\bar{X} - \bar{Y} - \varepsilon, \bar{X} - \bar{Y} + \varepsilon]$ $\varepsilon \geq Z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}$ $Var(\bar{X} - \bar{Y})$

To get CI for $\mu_X - \mu_Y$ with confidence level $\geq 1-\alpha$

X, Y normal	σ_X, σ_Y known	$Z_{\frac{\alpha}{2}} \sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}} \leq \varepsilon$
	$\sigma_X = \sigma_Y = \sigma$ unknown	$t_{\frac{\alpha}{2}}(n+m-2)\, S_p \sqrt{\frac{1}{n} + \frac{1}{m}} \leq \varepsilon$
	$\sigma_X \neq \sigma_Y$ unknown	$t_{\frac{\alpha}{2}}(n-1) \frac{S_D}{\sqrt{n}} \leq \varepsilon$, $D = X - Y$
X, Y not normal	σ_X, σ_Y known	approximately $Z_{\frac{\alpha}{2}} \sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}} \leq \varepsilon$

Problem 1. (X, Y) are a pair of RV, $D = X - Y$ $(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)$

$D \sim N(\mu_D, \sigma_D^2)$ $n = 25$, $S_D^2 = 4$

What should length of CI be to guarantee 95% level of confidence?

- Answer: exactly asking you to constance of 95% CI (problem done before)

$n - 1 = 24$ $1 - \alpha = 95\%$, $t_{\frac{\alpha}{2}}(24) = 2.064$ $S_D = 2$

$2.064 \times \frac{2}{5} \leq \varepsilon$, $\varepsilon \geq 0.8256$ length $2\varepsilon \geq 2 \times 0.8256$.

2. If length $= 0.5$, what's the maximum confidence level guarantee?

$2\varepsilon = 1$, $\varepsilon = 0.5$

$P(\mu \in [\bar{X} - \varepsilon, \bar{X} + \varepsilon])$

$t_{\frac{\alpha}{2}}(24) \cdot \frac{2}{5} \leq 0.5$

$t_{\frac{\alpha}{2}}(24) \leq 1.25$

$T \sim t(24)$

$P(T > t_{\frac{\alpha}{2}}(24)) = \frac{\alpha}{2}$

$P(T > 1.25) = \frac{\alpha}{2}$

$\quad \| $

0.11168

Problem $X \sim N(\mu_X, 5^2)$, $Y \sim N(\mu_Y, 3^2)$ want 95% CI for $\mu_X - \mu_Y$, length = 0.1

What's the sample size n, m we should choose to minimize $n+m$? $N = n+m$

$z_{\frac{\alpha}{2}} = 1.96$, $\varepsilon = 0.05$

$m = 3k$

$n = 5k$

$m+n = 8k = 8^2 \times 1600$

$n = \sigma_X \cdot k$

$m = \sigma_Y \cdot k$

$$1.96\sqrt{\frac{5^2}{n} + \frac{3^2}{m}} \le 0.05$$

$$\frac{5^2}{n} + \frac{3^2}{m} \le \left(\frac{0.05}{1.96}\right)^2 \approx \left(\frac{0.05}{2}\right)^2 = \left(\frac{0.1}{4}\right)^2 = \left(\frac{1}{40}\right)^2 = \frac{1}{1600}$$

To minimize

$$\frac{5^2}{n} + \frac{3^2}{m} = \frac{1}{1600}$$

$$\Rightarrow \frac{5^2}{5k} + \frac{3^2}{3k} = \frac{1}{1600},$$

$$\frac{5}{k} + \frac{3}{k} = \frac{1}{1600}$$

$$\frac{8}{k} = \frac{1}{1600}$$

$$k = 8 \times 1600$$

$N = n+m$

$$\frac{5^2}{n} + \frac{3^2}{N-n} = \frac{1}{1600}$$

$$\frac{d}{dn} : -\frac{5^2}{n^2} + \frac{3^2}{(N-n)^2} \cdot (+1) = 0$$

$$\frac{3^2}{(N-n)^2} = \frac{5^2}{n^2}, \quad \frac{3}{N-n} = \frac{5}{n}, \quad \frac{N-n}{3} = \frac{n}{5}$$

$$\frac{m}{3} = \frac{n}{5} = k.$$