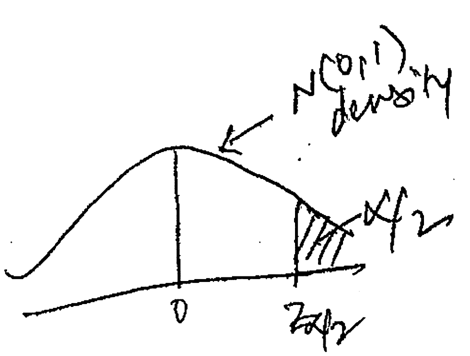
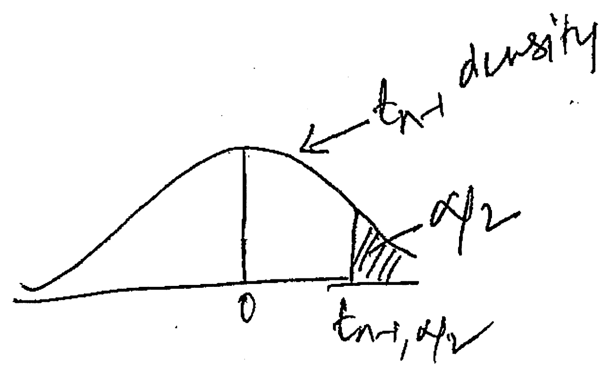
Recap:

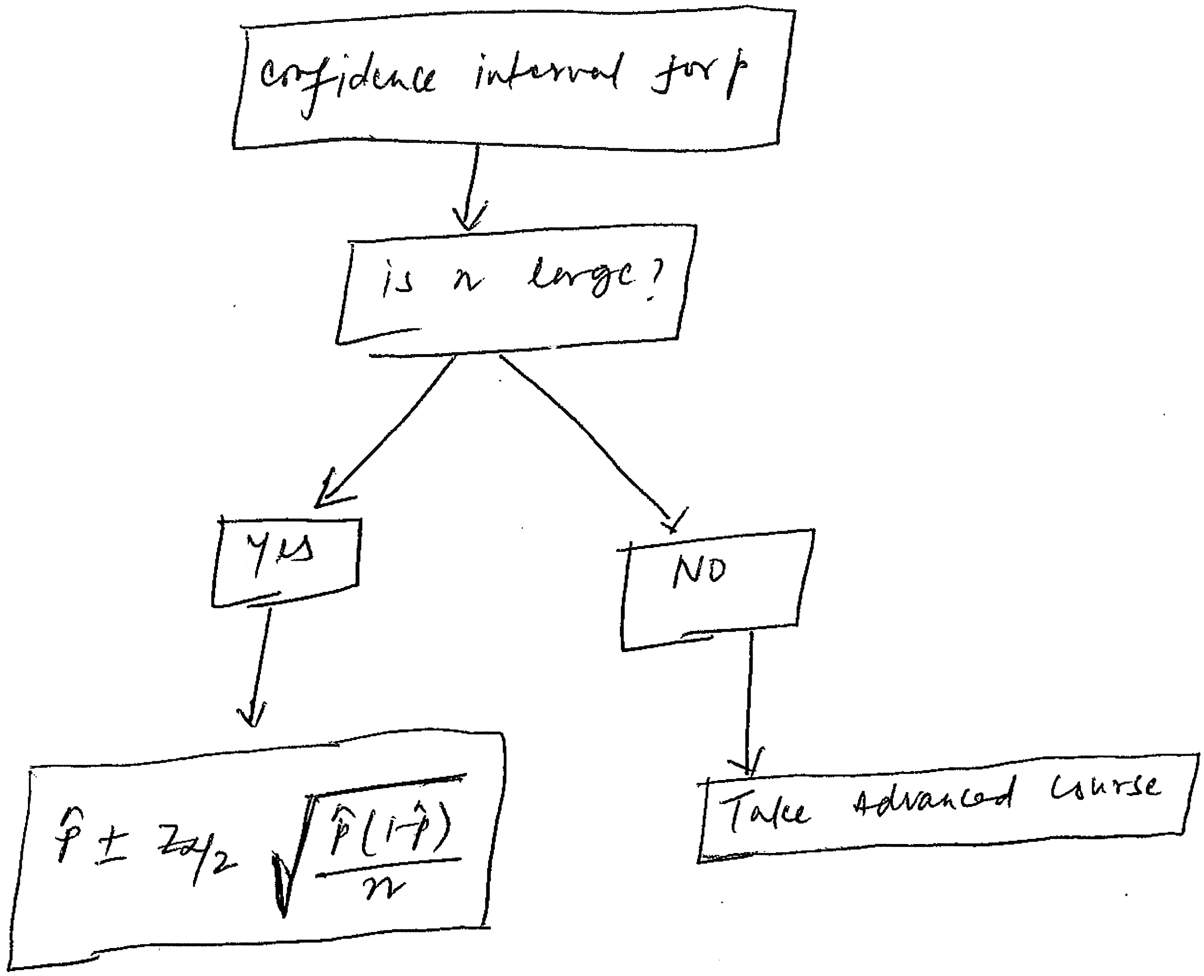
Confidence interval for population mean μ

Population normal?

- Yes
  - σ^2 known?
    - Yes → $\bar{X} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$
    - NO → $\bar{X} \pm t_{n-1,\alpha/2} \frac{s}{\sqrt{n}}$
- NO
  - n large?
    - Yes → $\bar{X} \pm z_{\alpha/2} \frac{s}{\sqrt{n}}$ or $t_{n-1,\alpha/2}$
    - NO → Take advanced course






confidence interval for p
is n large?
Yes
NO
$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$
Take Advanced Course

Recap.

Paired data:

Subject	X	Y
1	X_1	Y_1
2	X_2	Y_2
$\vdots$		
n	X_n	Y_n

- Each subject gives both X and Y
- The X and Y ~~xxxx~~ samples are not independent.

Two-independent samples:

$$\left.\begin{array}{l} n_1 \text{ subjects in group 1}: \quad X_1, X_2, \ldots, X_{n_1} \\ n_2 \text{ " " " 2}: \quad Y_1, Y_2, \ldots, Y_{n_2} \end{array}\right\}$$

- The subjects in the two groups are different.

Inference on difference in population means.

Case 1: Paired data :

$$\boxed{\begin{array}{c} X \sim f_{\theta_1}(x) \\ E(X) = \mu_X \end{array}} \qquad \boxed{\begin{array}{c} Y \sim f_{\theta_2}(x) \\ E(X) = \mu_Y \end{array}}$$

$$\downarrow \qquad\qquad\qquad \downarrow$$

$$X_1, \ldots, X_n \text{ (RS)} \qquad (Y_1, Y_2, \ldots, Y_n)$$

- Inference on $\mu_X - \mu_Y$
- Data are paired, i.e, they are collected as $(X_1, Y_1), \ldots, (X_n, Y_n)$
- Define: $D_i = X_i - Y_i$, $i = 1, 2, \ldots, n$

$$\Rightarrow E[D_i] = \mu_X - \mu_Y = \mu_D \text{ (say)}.$$

- Apply one-sample procedure to the differences.
- $100(1-\alpha)\%$ CI for μ_D assuming $D_1, \ldots, D_n \overset{RS}{\sim} N[\mu_D^2, \sigma_D^2]$:

$$\begin{cases} \bar{D} \pm z_{\alpha/2} \cdot \dfrac{\sigma_D}{\sqrt{n}} & \text{if } \sigma_D^2 \text{ is known} \\ \bar{D} \pm t_{n-1, \alpha/2} \dfrac{S_D}{\sqrt{n}} & \text{if } \sigma_D^2 \text{ is unknown.} \end{cases}$$

Here : $S_D = SD$ of the differences

- Approx $100(1-\alpha)\%$ CI for μ_D if n is large:

$$\bar{D} \pm z_{\alpha/2} \frac{S_D}{\sqrt{n}}$$ (Normality of $D_1, \ldots, D_n$ is not needed)

Q. What is the pivot here?

$$Z = \frac{\bar{D} - \mu_D}{\sigma_D/\sqrt{n}} \quad \mu_D \leftarrow E[\bar{D}], \quad \sigma_D/\sqrt{n} \leftarrow SE(\bar{D})$$

$$\sim \quad T = \frac{\bar{D} - \mu_D}{S_D/\sqrt{n}} \quad \mu_D \leftarrow E[\bar{D}], \quad S_D/\sqrt{n} \leftarrow \hat{SE}(\bar{D})$$

Case 2: Two independent samples

Assume: $X_1, X_2, \ldots, X_{n_1} \overset{iid}{\sim} N[\mu_X, \sigma_X^2]$
$Y_1, Y_2, \ldots, Y_{n_2} \overset{iid}{\sim} N[\mu_Y, \sigma_Y^2]$ $\Big\}$ σ_X^2, σ_Y^2 unknown.

Scenario 1: No assumption regarding σ_X^2 and σ_Y^2 — they may be equal or unequal

$$E[\bar{X} - \bar{Y}] = \mu_X - \mu_Y$$

- Estimate
  - μ_X by $\bar{X}$
  - μ_Y by $\bar{Y}$
  - σ_X^2 by $S_X^2 = \frac{1}{n_1 - 1} \sum_{i=1}^{n_1} (X_i - \bar{X})^2$
  - σ_Y^2 by $S_Y^2 = \frac{1}{n_2 - 1} \sum_{i=1}^{n_2} (Y_i - \bar{Y})^2$

$$SE[\bar{X} - \bar{Y}] = \sqrt{var(\bar{X} - \bar{Y})} \underset{\text{indep.}}{=} \sqrt{\frac{\sigma_X^2}{n_1} + \frac{\sigma_Y^2}{n_2} - 0}$$

- Pivot: $T = \dfrac{(\bar{X}-\bar{Y}) - (\mu_X - \mu_Y)}{\sqrt{\dfrac{s_X^2}{n_1} + \dfrac{s_Y^2}{n_2}}}$ $\sqrt{\dfrac{s_X^2}{n_1} + \dfrac{s_Y^2}{n_2}} \longrightarrow \widehat{SE}(\bar{X}-\bar{Y})$.

- The distribution of the pivot can be approximated by a t_ν distribution where

$$\nu = \frac{\left(\dfrac{s_X^2}{n_1} + \dfrac{s_Y^2}{n_2}\right)^2}{\dfrac{s_X^4}{n_1^2(n_1-1)} + \dfrac{s_Y^4}{n_2^2(n_2-1)}}$$

- Known as Satterthwaite's approximation

- Approx $100(1-\alpha)\%$ CI for $\mu_X - \mu_Y$: $(\bar{X}-\bar{Y}) \pm t_{\nu,\alpha/2}\, \widehat{SE}(\bar{X}-\bar{Y})$

- When n_1 and n_2 are large: $(\bar{X}-\bar{Y}) \pm z_{\alpha/2}\, \widehat{SE}(\bar{X}-\bar{Y})$

  (no need to make normality assumption).

<u>Scenario 2:</u> Assume that $\sigma_X^2 = \sigma_Y^2 = \sigma^2$ (common variance)

- Estimate μ_X by $\bar{X}$, μ_Y by $\bar{Y}$
- Estimate σ^2 by pooling the two samples to get

$$S_p^2 = \frac{\sum_{i=1}^{n_1}(X_i-\bar{X})^2 + \sum_{i=1}^{n_2}(Y_i-\bar{Y})^2}{n_1+n_2-2} = \frac{(n_1-1)S_X^2 + (n_2-1)S_Y^2}{n_1+n_2-2}$$

$\uparrow$ pooled-sample variance

$$E[S_p^2] = \frac{1}{(n_1+n_2-2)}\left[E\left[(n_1-1)S_X^2\right] + (n_2-1)S_Y^2\right]$$

$$= \frac{1}{n_1+n_2-2}\left[(n_1-1)\sigma^2 + (n_2-1)\sigma^2\right]$$

$$= \sigma^2 \Rightarrow S_p^2 \text{ is unbiased for } \sigma^2$$

- ~~[illegible]~~
- $SE(\bar{X}-\bar{Y}) = \sqrt{var(\bar{X}-\bar{Y})}$

$$= \sqrt{\frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}} = \sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$$

- $100(1-\alpha)\%$ CI for $\mu_X - \mu_Y$:
- Approx CI when n_1 and n_2 are large:

5

Pivot:

$$T = \frac{(\bar{X}-\bar{Y}) - (\mu_X - \mu_Y)}{S_p \sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \sim t_{n_1+n_2-2}$$

$$\widehat{SE}(\bar{X}-\bar{Y})$$

$100(1-\alpha)\%$ CI for $\mu_X - \mu_Y$:

$$(\bar{X}-\bar{Y}) \pm t_{n_1+n_2-2,\,\alpha/2} \, \widehat{SE}(\bar{X}-\bar{Y})$$

If n_1 and n_2 are large: $(\bar{X}-\bar{Y}) + z_{\alpha/2} \, \widehat{SE}(\bar{X}-\bar{Y})$

(no need of the normality assumption).

• Two-indep sample situation $100(1-\alpha)\%$ CI for $(p_1 - p_2)$:

difference in means of two Bernoulli populations

Estimator: $\hat{p}_1 - \hat{p}_2$

$$SE[\hat{p}_1 - \hat{p}_2] = \sqrt{var(\hat{p}_1 - \hat{p}_2)} \underset{\text{indep.}}{=} \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

← Just replace (p_1, p_2) by $(\hat{p}_1, \hat{p}_2)$ to get $\hat{SE}(\hat{p}_1 - \hat{p}_2)$.

Pivot: $$Z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\hat{SE}(\hat{p}_1 - \hat{p}_2)} \dot\sim N(0,1)$$ when n_1 and n_2 are large

$\Rightarrow$ $100(1-\alpha)\%$ CI for $p_1 - p_2$ is:

$$(\hat{p}_1 - \hat{p}_2) \pm z_{\alpha/2} \cdot \hat{SE}(\hat{p}_1 - \hat{p}_2)$$ [when n_1 and n_2 are large].