11/8

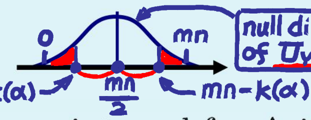
* the test statistic: $U_Y(\Delta_0) = \#\{X_i < Y_j - \Delta_0\} = \#\{Y_j - X_i > \Delta_0\}$, fixed, X_i^*, Y_j^*, changed
* the acceptance region: $k(\alpha) \le U_Y(\Delta_0) \le mn - k(\alpha)$, What if $\Delta_0 = 0$?

Assume it is an integer

where $k(\alpha)$ is the critical value determined by the significance level α

irrelevant to F

(**Note.** The null distribution of $U_Y(\Delta_0)$ is symmetric about $mn/2$.) check Thm10 (LNp.40)



null dist. of $U_Y(\Delta_0)$

- By the duality of test and C.I., a $100(1-\alpha)\%$ confidence interval for Δ is

$$C = \{\Delta \mid k(\alpha) \le U_Y(\Delta) \le mn - k(\alpha)\}.$$

changed, fixed

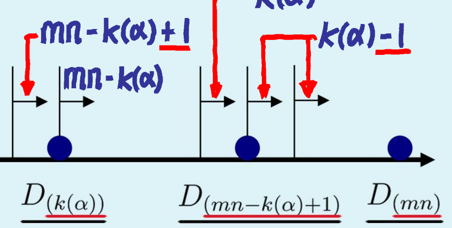
k(α) can be obtained from the critical value of W_Y (check Thm11, LNp.39) in Table8, TBp.A21

pivotal quantity

What is its dist.?

– Let $D_{(1)}, D_{(2)}, \ldots, D_{(mn)}$ denote the ordered mn differences $(Y_j - X_i)$'s. Then,

$$C = [D_{(k(\alpha))},\ D_{(mn-k(\alpha)+1)}].$$



– To see this,
  * if $\Delta_0 = D_{(k(\alpha))}$, then $U_Y(\Delta_0) = \#\{Y_j - X_i > \Delta_0\} = mn - k(\alpha)$, ← accept
    if $\Delta_0 < D_{(k(\alpha))}$, then $U_Y(\Delta_0) = \#\{Y_j - X_i > \Delta_0\} \ge mn - k(\alpha) + 1$, ← reject
    thus, $D_{(k(\alpha))}$ is the leftmost point of the confidence interval C,
  * if $\Delta_0 \le D_{(mn-k(\alpha)+1)}$, then $U_Y(\Delta_0) = \#\{Y_j - X_i > \Delta_0\} \ge k(\alpha)$, ← accept
    if $\Delta_0 > D_{(mn-k(\alpha)+1)}$, then $U_Y(\Delta_0) = \#\{Y_j - X_i > \Delta_0\} \le k(\alpha) - 1$, ← reject
    thus, $D_{(mn-k(\alpha)+1)}$ is the rightmost point of the confidence interval C.

Example 6 (C.I. for Δ, heat of fusion of ice, cont. Ex.4 in LNp.34 & Ex.5 in LNp.42)

- $n = 13$ (method A), $m = 8$ (method B), $W_B = 51$. Under null, $E(W_B) = 88$.
- Under the significant level $\alpha = 0.05$, the critical value for W_B^* is 60 (Ex.4, LNp.34) $\Rightarrow$ acceptance region: $61 \le W_B \le 88 + (88 - 61) = 115$

  $k(\alpha)$ → $\Leftrightarrow 25 \le U_B = W_B - [8(8+1)]/2 = W_B - 36 \le 79.$ ← $mn-k(\alpha)$

consistent with the test result in Ex.4

- After sorting the $mn = 8 \times 13 = 104$ differences $(Y_j - X_i)$'s, we get

$$D_{(k(\alpha)=25)} = -0.07 \quad \text{and} \quad D_{(mn-k(\alpha)+1=80)} = -0.01.$$

normality seems valid in this case (check box plot in LNp.4)

A 95% confidence interval for Δ is $(-0.07, -0.01)$, which does not contain 0. $\times(-1) = (0.01, 0.07)$

[$\overset{\text{cf.}}{\longleftrightarrow}$ the C.I. (0.015, 0.065) given in Ex.2 (LNp.12) under normality

– Note that the Δ here is the $-\Delta$ in Ex.2.

Δ can be defined as "difference of medians" under (□)

Check Note8 (LNp.39)

– In this case, the C.I. based on the nonparametric model is slightly wider than the one based on the normal model.

– But, the latter C.I. relies on the validity of normality assumption.]

Theorem 15 (Bootstrap confidence interval for π_Δ ($\leftrightarrow \Delta$))

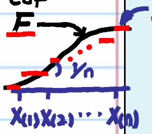
Consider the nonparametric model (◇) in LNp.35 or the nonparametric model (□) in LNp.27. (**Note.** (1) (□) has more models than (◇) (2) $\pi_\Delta = P(X < Y)$ is well-defined in (◇) and (□) (3) Δ is well-defined only in (◇))

TBp.284–285

- Bootstrapping is a numerical method that can be used to gain information about the sampling distribution of $\hat{\pi}_\Delta = \frac{1}{mn}(\#\{X_i < Y_j\}) \xrightarrow{e} \pi_\Delta$, and the estimated standard error of $\hat{\pi}_\Delta$. a r.v. ← only has one obs. of this r.v.

Under H_0 / Under H_A

true cdf



- In bootstrap, we $\hat{\pi}_\Delta \leftarrow \left\{ \begin{array}{ll} X_1, \ldots, X_n \sim \text{i.i.d. from } F \\ Y_1, \ldots, Y_m \sim \text{i.i.d. from } G \end{array} \right\} \Leftarrow$ independent ← cf.
  - replace the ture cdf F (unknown) by the empirical cdf $\hat{F}_n$ (known) of $(X_1, \ldots, X_n) = (x_1, \ldots, x_n)$ [$\hat{F}_n$: assigns x_i's equal probabilities $1/n$]
  - replace the true cdf G (unknown) by the empirical cdf $\hat{G}_m$ (known) of $(Y_1, \ldots, Y_m) = (y_1, \ldots, y_m)$ [$\hat{G}_m$: assigns y_j's equal probabilities $1/m$]
- Re-sample (generate data $X'_1, \ldots, X'_n, Y'_1, \ldots, Y'_m$ using simulation) from this model:

joint distribution of these r.v.'s: completely known

$$\left. \begin{array}{ll} X'_1, \ldots, X'_n \sim \text{i.i.d. from } \hat{F}_n \\ Y'_1, \ldots, Y'_m \sim \text{i.i.d. from } \hat{G}_m \end{array} \right\} \Leftarrow \text{independent}$$

Are they similar?

  - $X'_1, \ldots, X'_n$ is a with-replacement sample from the population $\{x_1, \ldots, x_n\}$
  - $Y'_1, \ldots, Y'_m$ is a with-replacement sample from the population $\{y_1, \ldots, y_m\}$

we can control this

known data

- Repeat the re-sampling procedure many times, say B times, and
  - at each time, compute $\hat{\pi}'_\Delta = \frac{1}{mn} \#\{X'_i < Y'_j\}$ from $(X'_1, \ldots, X'_n, Y'_1, \ldots, Y'_m)$
  - this produces a bootstrap sample: $(\hat{\pi}'_{\Delta,1}, \ldots, \hat{\pi}'_{\Delta,B})$ ← can be regarded as a sample of $\hat{\pi}_\Delta$
- A histogram of $(\hat{\pi}'_{\Delta,1}, \ldots, \hat{\pi}'_{\Delta,B})$ offers an indication of the sampling distribution of $\hat{\pi}_\Delta$ ($\Rightarrow$ a $100(1-\alpha)\%$ C.I. of π_Δ is $[\hat{\pi}'_{\Delta,(B(\alpha/2))}, \hat{\pi}'_{\Delta,(B(1-\alpha/2))}]$),

a r.v.

- the standard deviation of $(\hat{\pi}'_{\Delta,1}, \ldots, \hat{\pi}'_{\Delta,B}) \stackrel{e}{\rightarrow}$ the standard error of $\hat{\pi}_\Delta$.

a r.v.

❖ **Reading**: textbook, 11.2.3

• Comparing paired samples ↔ Independent samples

Recall. Strata in survey sampling. (LN, CH7, p.59-63)

• Problem formulation and statistical modeling

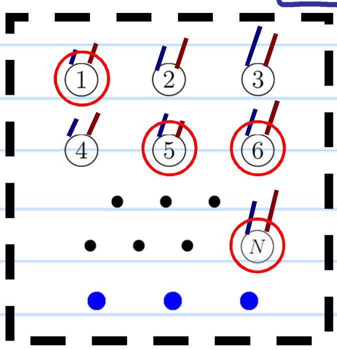
Such connection does not exist in 2-independent samples.

observed data (random variables)

$\{X_1, \ldots, X_n\}$
$\updownarrow \quad \updownarrow \quad \updownarrow$
$\{Y_1, \ldots, Y_n\}$

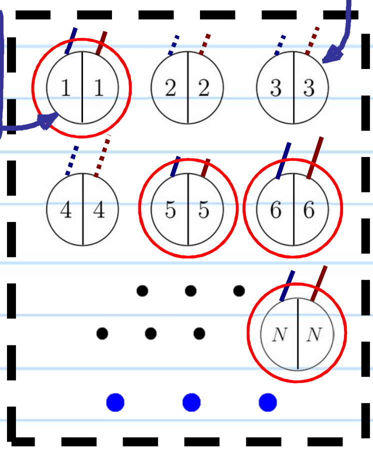
population

homogeneous within stratum



a stratum

population



- X_i's, Y_i's are continuous quantities
- X–Y is meaningful

the comparison of their means is meaningful.

2-indep samples

cf.

Data

U	block	V
1	1	X_1
1	2	X_2
⋮	⋮	⋮
1	n	X_n
2	1	Y_1
2	2	Y_2
⋮	⋮	⋮
2	n	Y_n

s.r.s., $N \rightarrow \infty$:
without replacement
≈ with replacement
($\Rightarrow$ i.i.d.)

For example, in human population,
- X_i's: left eye vision
- Y_i's: right eye vision

of the ith person

For example, in medical study,
- X_i's: treatment
- Y_i's: control

applied on the ith twins

2-indep samples cf

- $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n) \sim$ i.i.d. with a common continuous joint distribution $F(x, y)$ ← population distribution ← cf
- (X_i, Y_i), i.e., F, might not be independent

Estimation of a ratio (LN, CH7, p.38~39)

∵S.r.S., $N \to \infty$

- Let random variables $Z_1, \ldots, Z_n$ represent the variablility of the n members sampled from the population. ← cf. → Independent samples
- Assume $Z_1, \ldots, Z_n$ are i.i.d. from a population distribution $H(z)$. ← cf.

check LNp.2

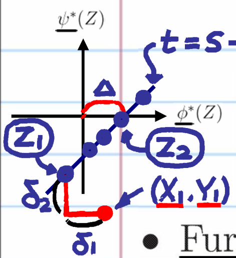
use e.g.'s in LNp.46 to understand

- Let $X = \phi(Z)$ and $Y = \psi(Z)$, where ϕ, ψ contain random components, and denote
  - $F(x, y)$: the joint distributions of (X, Y), ← not necessarily independent
  - μ_X and μ_Y: the means of X and Y, respectively,
  - $\Delta = \mu_X - \mu_Y$.

$$(\phi, \psi): \quad \underset{(H(z))}{Z} \to \underset{(F(x,y))}{(X, Y)}$$

Because $Z_1, \ldots, Z_n \sim$ i.i.d. $H(z)$, $(X_1, Y_1), \ldots, (X_n, Y_n) \sim$ i.i.d. $F(x, y)$

- Then, for $1 \le i \le n$, same ball
$$\begin{cases} X_i = \phi(\boxed{Z_i}) \\ Y_i = \psi(\boxed{Z_i}) \end{cases}$$



different balls cf.

$$\begin{cases} X_i = \phi(\boxed{Z_i}) &= \mu_X + \boxed{\epsilon_{1i}}, \\ Y_j = \psi(\boxed{Z_{n+j}}) &= \mu_Y + \boxed{\epsilon_{2j}}, \end{cases}$$
← mean zero in two independent samples case. check LNp.2

- Further assume that cf.

Δ: shift in r.v.'s, not only in their dist.'s

(1) $\phi(Z) = \phi^*(Z) + \delta_1$ and $\psi(Z) = \psi^*(Z) + \delta_2$, where ϕ^*, ψ^* are fixed functions and δ_1, δ_2 are independent random variables with mean 0 (X, Y) ← can have different dist

(2) Z, δ_1, δ_2 are independent

same r.v.

(3) $\psi^*(Z) = \phi^*(Z) - \Delta \Rightarrow \Delta = \phi^*(Z) - \psi^*(Z)$ ← cf. →
$$\mu_X = E[\phi^*(Z)] \qquad \mu_Y = E[\psi^*(Z)]$$

Then,
$$\begin{cases} X_i = \phi^*(Z_i) + \delta_{1i} = \phi^*(Z_i) + \delta_{1i} = \mu_X + [\phi^*(Z_i) - \mu_X] + \delta_{1i} \\ Y_i = \psi^*(Z_i) + \delta_{2i} = \phi^*(Z_i) - \Delta + \delta_{2i} = \mu_Y + [\phi^*(Z_i) - \mu_X] + \delta_{2i} \end{cases}$$
cf. ∵ (3)

same marginal distribution

- Q: What are the sources of variation in ϵ's and δ's? If we apply the above formulation to the case of two independent samples, then add (3)
$$\begin{cases} X_i = \mu_X + \epsilon_{1i} = \phi^*(Z_i) + \delta_{1i} = \mu_X + (\phi^*(Z_i) - \mu_X) + \delta_{1i} \\ Y_j = \mu_Y + \epsilon_{2j} = \psi^*(Z_{n+j}) + \delta_{2j} = \mu_Y + (\phi^*(Z_{n+j}) - \mu_X) + \delta_{2j} \end{cases}$$
cf.

Change data collection procedure

- A comparison variation in ϵ ← variation in Z & variation in δ $Var(\epsilon) \downarrow Var(\delta)$
  - Increase sample sizes: increase information about μ_X and μ_Y (signal)
  - 2 independent → paired: suppress the variation of error (noise)

cf.

Recall comparison of S.R.S & stratified random sampling in LN, CH7, P.73-74

Theorem 16 (A brief variance comparison of paired and independent samples) Under (3)

Consider the models in the *dashed* frames. Under the two models,

- $\epsilon = [\phi^*(Z) - \mu_X] + \delta \quad \Rightarrow \quad \underbrace{Var(\epsilon)}_{\equiv \sigma_\epsilon^2} = Var[\phi^*(Z)] + Var(\delta) \ge \underbrace{Var(\delta)}_{\equiv \sigma_\delta^2}$
  due to the variation of members in the population
- 2 independent samples $(n = m)$
  - $X_i - Y_j = (\mu_X - \mu_Y) + (\epsilon_{1i} - \epsilon_{2j})$
  $= (\mu_X - \mu_Y) + [\phi^*(Z_i) - \phi^*(Z_{n+j})] + (\delta_{1i} - \delta_{2j})$
  - $\overline{X} - \overline{Y} = (\mu_X - \mu_Y) + (\overline{\epsilon_1} - \overline{\epsilon_2})$ cf. → paired samples ← cf
  $\Rightarrow \overline{X} - \overline{Y} \xrightarrow{e} \Delta$ and $Var(\overline{X} - \overline{Y}) = (\sigma_{\epsilon_1}^2 / n) + (\sigma_{\epsilon_2}^2 / n)$

- paired samples $\phi^*(Z_i)-\mu_X$ cancelled out
  $D_i \equiv X_i - Y_i = (\mu_X - \mu_Y) + (\delta_{1i} - \delta_{2i})$
  $Var(D_i) = \sigma^2_{\delta_1} + \sigma^2_{\delta_2}$
  - $\overline{D} = \overline{X} - \overline{Y} = (\mu_X - \mu_Y) + (\overline{\delta}_1 - \overline{\delta}_2)$ 2-indep. samples
  same estimator of Δ as in 2 indep. samples $\Rightarrow$ $\overline{X} - \overline{Y} \xrightarrow{e} \Delta = \mu_X - \mu_Y$ and larger smaller
  $Var(\overline{X} - \overline{Y}) = (\sigma^2_{\delta_1}/n) + (\sigma^2_{\delta_2}/n)$ cf.

$Var(\overline{X} - \overline{Y})$ under the 2-independent-sample model with the sample size
$$n' = \frac{\sigma^2_\epsilon}{\sigma^2_\delta}\, n \quad (\geq n).$$

- Paired sample is more effective than independent samples in this case.

Theorem 17 (Conditions under which paired sample is more effective) remove assumption (3)

Consider the models in the *dotted* frames of LNp.48. Under the two models,

- $E(X) = E[\phi^*(Z) + \delta_1] = E[\phi^*(Z)] = \mu_X$
  $E(Y) = E[\psi^*(Z) + \delta_2] = E[\psi^*(Z)] = \mu_Y$

  Note. X (or Y) have same marginal dist. in 2-indep. & paired cases.
- $Var(X) = Var[\phi^*(Z) + \delta_1] = Var[\phi^*(Z)] + Var(\delta_1) \equiv \sigma^2_X \; (= \sigma^2_{\epsilon_1})$
  $Var(Y) = Var[\psi^*(Z) + \delta_2] = Var[\psi^*(Z)] + Var(\delta_2) \equiv \sigma^2_Y \; (= \sigma^2_{\epsilon_2})$
- 2 independent samples ($n = m$)
  - $Cov(X_i, Y_j) = Cov[\phi^*(Z_i) + \delta_{1i}, \psi^*(Z_{n+j}) + \delta_{2j}] = 0 \xleftrightarrow{cf.}$ paired samples
  - $E(\overline{X} - \overline{Y}) = \mu_X - \mu_Y = \Delta$
  - $Var(\overline{X} - \overline{Y}) = (\sigma^2_X + \sigma^2_Y)/n \xleftrightarrow{cf.}$ paired samples

- paired samples
  - $Cov(X_i, Y_i) = Cov[\phi^*(Z_i) + \delta_{1i}, \psi^*(Z_i) + \delta_{2i}]$ same r.v. → cov: $Z_i \leftrightarrow Z_i$; $Z_i \leftrightarrow \delta_{1i}$, $Z_i \leftrightarrow \delta_{2i}$, $\delta_{1i} \leftrightarrow \delta_{2i}$ indep
    $= Cov[\phi^*(Z_i), \psi^*(Z_i)] \equiv \sigma_{XY} \xleftrightarrow{cf.}$ 2-indep. samples
    * **Note**. We do not observe $(\phi^*(Z_i), \psi^*(Z_i))$'s. But, σ_{XY} can be estimated using (X_i, Y_i)'s data.
    * Denote the correlation of (X_i, Y_i) by $\rho_{XY} = \dfrac{\sigma_{XY}}{\sigma_X\, \sigma_Y}$
    * Notice that $\rho_{XY} \neq$ ($\leq$ in absolute value)
      $Cor[\phi^*(Z), \psi^*(Z)] = \dfrac{\sigma_{XY}}{\sigma_{\phi^*(Z)}\, \sigma_{\psi^*(Z)}}$ ($\sigma_{\phi^*(Z)} \leq \sigma_X$, $\sigma_{\psi^*(Z)} \leq \sigma_Y$)
  - Let $D_i = X_i - Y_i$, $i = 1, \ldots, n$. Then,
    * $D_1, \ldots, D_n$ are i.i.d. ← ∵ $(X_1, Y_1), \cdots, (X_n, Y_n)$ are independent
    * $E(D_i) = \mu_X - \mu_Y$
    * $Var(D_i) = Var(X_i) + Var(Y_i) - 2\,Cov(X_i, Y_i) = \sigma^2_X + \sigma^2_Y - 2\,\sigma_{XY}$
  - Since $\overline{D} = \overline{X} - \overline{Y}$ ($\xrightarrow{e} \Delta$)
    * $E(\overline{D}) = \mu_X - \mu_Y = \Delta$
    * $Var(\overline{D}) = Var(\overline{X} - \overline{Y}) = (\sigma^2_X + \sigma^2_Y - 2\,\sigma_{XY})/n$
      $= (\sigma^2_X + \sigma^2_Y - 2\,\rho_{XY}\,\sigma_X\,\sigma_Y)/n \xleftrightarrow{cf.}$ 2-indep. samples

Why?

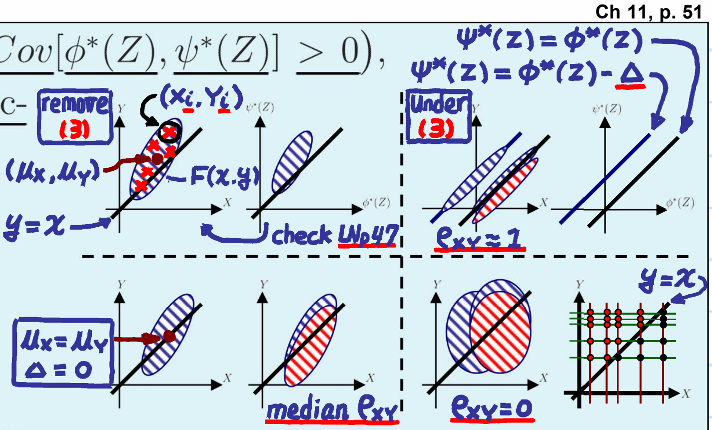
- If $\rho_{XY} > 0$ ($\Leftrightarrow \sigma_{XY} > 0 \Leftrightarrow Cov[\phi^*(Z), \psi^*(Z)] > 0$), then paired sample is more effective than independent samples.

Under (3)

- When $\psi^*(Z) = \phi^*(Z) - \Delta$,

$$\begin{aligned} &Cov[\phi^*(Z), \psi^*(Z)] \\ &= Cov[\phi^*(Z), \phi^*(Z) - \Delta] \\ &= Var[\phi^*(Z)] > 0. \end{aligned}$$

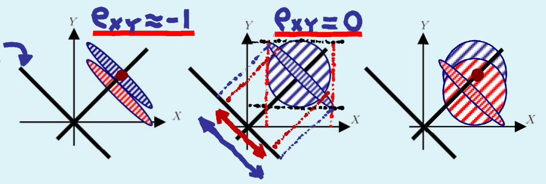
11/13



- Q: Why are independent samples more effective than paired samples when $\sigma_{XY} < 0$?

$y = -x$

$X - Y = (1, -1) * \binom{X}{Y}$



- If $\sigma_X^2 = \sigma_Y^2 = \sigma^2$, then in the paired case

$$\sigma_D^2 = Var(D) = [2\,\sigma^2(1 - \rho_{XY})]/n \quad \text{and} \quad \sigma_{\bar{X}-\bar{Y}}^2 = Var(\bar{X} - \bar{Y}) = 2\,\sigma^2/n$$

$\hat{\Delta}$ $\hat{\Delta}$

in the unpaired case. The relative efficiency is $\sigma_D^2/\sigma_{\bar{X}-\bar{Y}}^2 = 1 - \rho_{XY}$.

•paired
n X_i's
n Y_i's
•2-indep.
$2n$ X_i's
$2n$ Y_j's

- If $\rho_{XY} = 0.5$, a paired design with n pairs of subjects yields the same precision as an unpaired design with $2n$ subjects per treatment.

$\dfrac{\sigma_\delta^2}{\sigma_\varepsilon^2}$ under (3)

- From now on, the analyses of paired data are based on

$$D_i = X_i - Y_i, \quad i = 1, \ldots, n.$$

- Statistical modeling for D_i's: $D_1, \ldots, D_n \sim$ i.i.d. $F \Leftarrow$ one-sample model