Comparing two samples (Chapter 11)

• Comparing two independent samples

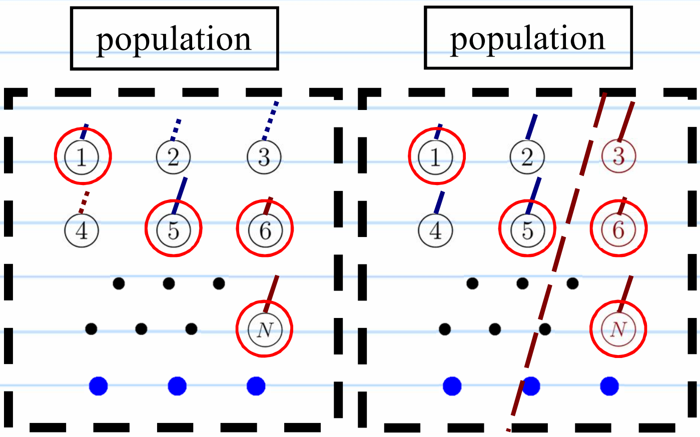
• Problem formulation and statistical modeling

Data

U	V
1	X_1
⋮	⋮
1	X_n
2	Y_1
⋮	⋮
2	Y_m

observed data from s.r.s (random variables)
$\{X_1, \ldots, X_n\}$
$\{Y_1, \ldots, Y_m\}$

- X_i's, Y_j's are continuous quantities of same characteristic
- $X-Y$ is meaningful



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

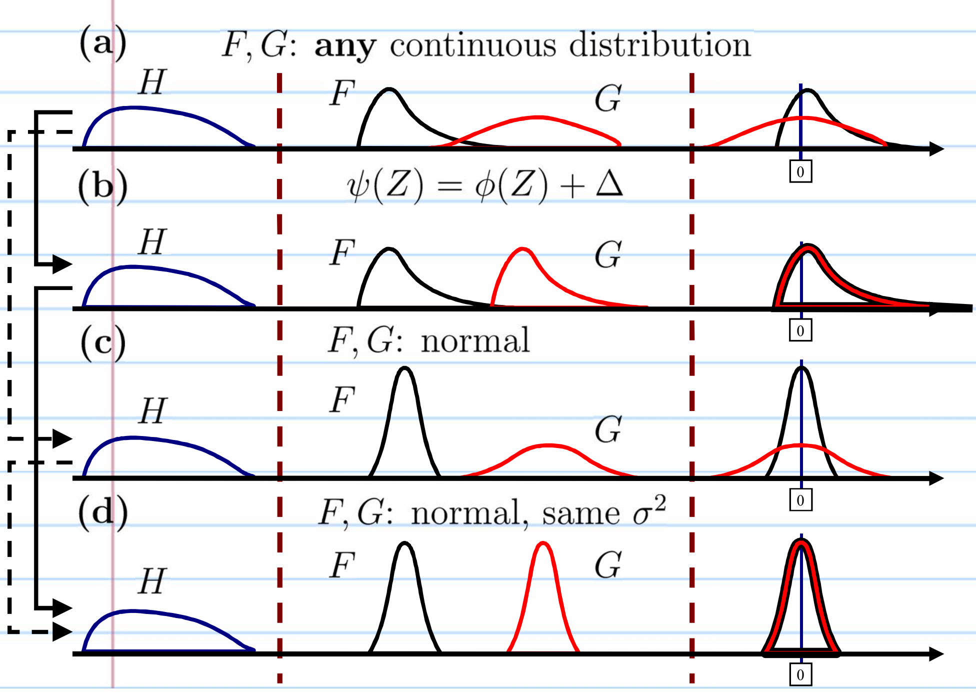
For example, in medical study,
- X_i's: treatment
- Y_j's: control

For example, in human population,
- X_i's: heights of males
- Y_j's: heights of females

Why?

- $X_1, \ldots, X_n \sim$ i.i.d. with a common continuous distribution F
- $Y_1, \ldots, Y_m \sim$ i.i.d. from a common continuous distribution G
- $\{X_1, \ldots, X_n\}$ and $\{Y_1, \ldots, Y_m\}$ are independent

- Let random variables $Z_1, \ldots, Z_n$ and $Z_{n+1}, \ldots, Z_{n+m}$ represent the variablility of the $n+m$ members sampled from the population.
- Assume $Z_1, \ldots, Z_{n+m}$ are i.i.d. from a population distribution H.
- Let F and G be the distributions of $X = \phi(Z)$ and $Y = \psi(Z)$, respectively.
- The transformations ϕ and ψ might contain random components, e.g., $\phi(Z) = \phi^*(Z) + \delta$, where ϕ^*: a fixed function and δ: a random variable.
- Let μ_X and μ_Y be the means of F and G, respectively.



- Let $X_i = \phi(Z_i)$, $i = 1, \ldots, n$, $\Rightarrow X_i \sim F$
- Let $Y_j = \psi(Z_{n+j})$, $j = 1, \ldots, m$, $\Rightarrow Y_j \sim G$
- $X_1, \ldots, X_n, Y_1, \ldots, Y_m$ are independent

- $X_i = \mu_X + \epsilon_{1,i}$, $i = 1, \ldots, n$,
- $Y_j = \mu_Y + \epsilon_{2,j}$, $j = 1, \ldots, m$,
- $E(\epsilon) = 0$ and ϵ's are independent

Note 1 (Some notes about comparing several samples)

- The samples are drawn under different conditions, and inferences must be made about possible effects of these conditions, e.g., treatment and control groups.
- Two-sample comparison (and ANOVA, multiple comparison):
  - methods for comparing samples from distributions that may be different
  - methods for making inference about how the distributions differ
- This chapter (and next chapter) will be concerned with analyzing measurements that are continuous in nature, e.g., temperature.

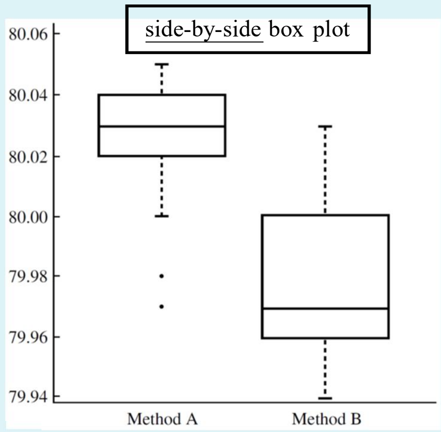
Example 1 (heat of fusion of ice, Natrella, 1963)

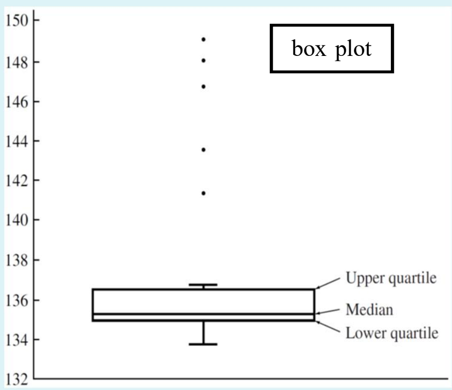
- Two methods, A and B, were used in a determination of the latent heat of fusion of ice.
- The investigators wished to find out how the methods differed.
- The following table gives the change in total heat from ice at -0.72°C to water 0°C in calories per gram of mass:

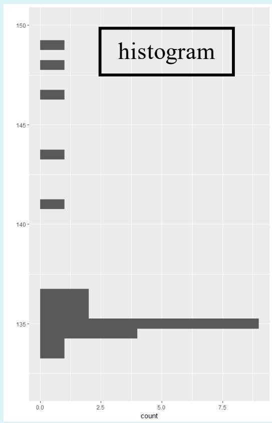
Method A	79.98	80.04	80.02	80.04	80.03	80.03	80.04	79.97
	80.05	80.03	80.02	80.00	80.02			
Method B	80.02	79.94	79.98	79.97	79.97	80.03	79.95	79.97

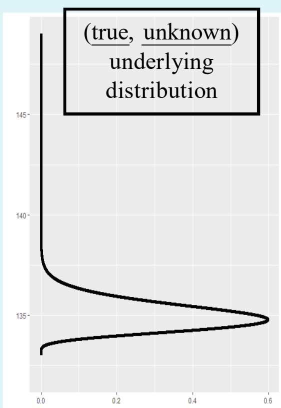
Definition 1 (box plot)

1. horizontal lines
   - at median, upper and lower quartiles (50%, 75%, 25% quantiles)
   - IQR = upper quartile − lower quartile
2. vertical lines: from upper (or lower) quartile to the most extreme data point that is within a distance of $1.5 \times IQR$ of the upper (or lower) quartile
3. each data point beyond the ends of the vertical lines is marked with an asterisk or dot (might be regarded as possible outliers)









Question 1.

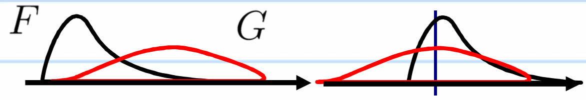
Q: How to define "two samples are identical" or "two samples are different"?

- Notice the distinction between "two identical random variables" ($X = Y$) and "two random variables with an identical distribution" ($X \sim Y$)
- In the statistical modeling of two-sample problem, $X \sim F$ and $Y \sim G$

Consider the different modelings in LNp.2,

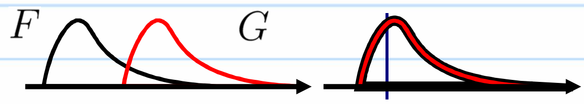
- For (a), $F = G$ vs. $F \neq G$
- For (a), $\mu_X = \mu_Y$ vs. $\mu_X \neq \mu_Y$
- For (a), $\tilde{\mu}_X = \tilde{\mu}_Y$ vs. $\tilde{\mu}_X \neq \tilde{\mu}_Y$
  ($\tilde{\mu}$: median)

(a)



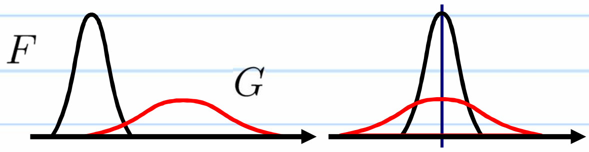
- For (b), $\Delta = 0$ vs. $\Delta \neq 0$

(b)



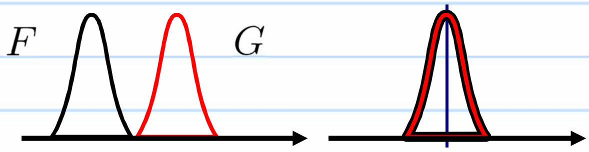
- For (c), $\mu_X = \mu_Y$ and $\sigma_X^2 = \sigma_Y^2$ vs.
  $\mu_X \neq \mu_Y$ or $\sigma_X^2 \neq \sigma_Y^2$
- For (c), $\mu_X = \mu_Y$ (i.e., different variances are allowed) vs. $\mu_X \neq \mu_Y$

(c)



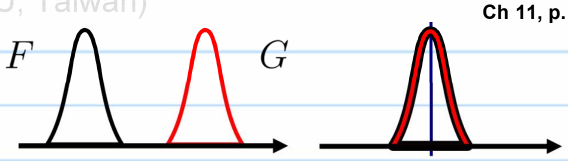
- For (c), $\sigma_X^2 = \sigma_Y^2$ (i.e., different means are allowed) vs. $\sigma_X^2 \neq \sigma_Y^2$

(d)



- For (d), $\mu_X = \mu_Y$ vs. $\mu_X \neq \mu_Y$

- Methods based on normality assumptions



- Assume that (1) F and G are normal, and (2) F and G have same variance.
- Thus, the statistical model is:

$$\left.\begin{array}{lll} \text{1st sample:} & X_1, \ldots, X_n & \sim \text{ i.i.d. } N(\mu_X, \sigma^2) \\ \text{2nd sample:} & Y_1, \ldots, Y_m & \sim \text{ i.i.d. } N(\mu_Y, \sigma^2) \end{array}\right\} \Leftarrow \text{independent} \quad (*)$$

  - This model contains three parameters: μ_X ($\in \mathbb{R}$), μ_Y ($\in \mathbb{R}$), σ^2 (> 0).
  - Under this model, the "difference" between F and G is simplified to be the difference between μ_X and μ_Y, i.e., $\Delta \equiv \mu_X - \mu_Y$ ($\Leftarrow$ called "effect"), and $\quad \mu_X - \mu_Y = 0 \;\Leftrightarrow\;$ no difference or no effect

Review 1 (estimation of the parameters in one-sample normal model)

Consider $X_1, \ldots, X_n \sim$ i.i.d. $N(\mu, \sigma^2)$, and the statistics

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{e} \mu \quad \text{and} \quad s_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2 \xrightarrow{e} \sigma^2$$

- distribution (exercise)
  - $\overline{X}$ and s_X^2 are independent
  - $\overline{X} \sim N(\mu, \sigma^2/n) \Rightarrow \sqrt{n}(\overline{X} - \mu)/\sigma \sim N(0,1)$
  - $(n-1)s_X^2 \sim \sigma^2\chi_{n-1}^2 \Rightarrow (n-1)s_X^2/\sigma^2 \sim \chi_{n-1}^2$; $n-1$: degrees of freedom

- $(T_1 = \sum_{i=1}^n X_i,\ \ T_2 = \sum_{i=1}^n X_i^2)$ is a sufficient and complete statistic (exercise, **Hint**. 2-parameter exponential family)
- Optimality
  - $\overline{X}$ is the uniformly minimum variance unbiased estimator (UMVUE) of μ (exercise, **Hint**. Lehmann-Scheffe Thm)
  - $\overline{X}$ is the maximum likelihood estimator (MLE) of μ (exercise, **Hint**.

$$\text{log-likelihood} \propto -\frac{n}{2}\log(\sigma^2) - \frac{\sum_{i=1}^n (X_i-\mu)^2}{2\sigma^2}\ \Bigg)$$

  - s_X^2 is the UMVUE of σ^2 (exercise, **Hint**. Lehmann-Scheffe Thm)
  - The MLE of σ^2 is $\dfrac{n-1}{n}s_X^2 = \dfrac{1}{n}\sum_{i=1}^n (X_i - \overline{X})^2$ (exercise)

Definition 1 (estimators of the parameters in the 2-sample normal model)

Under the two-sample normal model (∗) in LNp.6,

- an intuitive estimator of μ_X is $\overline{X} = \frac{1}{n}\sum_{i=1}^n X_i$,
- an intuitive estimator of μ_Y is $\overline{Y} = \frac{1}{m}\sum_{j=1}^m Y_j$,
- since $s_X^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i-\overline{X})^2$ and $s_Y^2 = \frac{1}{m-1}\sum_{j=1}^m (Y_j - \overline{Y})^2$ estimate the same parameter σ^2, we can pool them to get a better estimator:

$$s_p^2 = \frac{(n-1)}{(n-1)+(m-1)}s_X^2 + \frac{(m-1)}{(n-1)+(m-1)}s_Y^2 = \frac{(n-1)s_X^2 + (m-1)s_Y^2}{m+n-2}.$$

Note 1 (Some notes about the estimator of σ^2)

- s_p^2 is called the **pooled sample variance**
- s_p^2 is a weighted average of the sample variances of the X_i's and Y_j's, where
  - the weights are proportional to the degrees of freedom, it is appropriate since if one sample is of much larger size than the other, the estimate of σ^2 from that sample is more reliable $\Rightarrow$ it receives greater weight
  - since $E(s_X^2) = \sigma^2$ and $E(s_Y^2) = \sigma^2 \Rightarrow s_p^2$: an unbiased estimator of σ^2

Theorem 1 (distributions of the parameter estimators, 2-sample normal model)

- Since $(X_1, \ldots, X_n), (Y_1, \ldots, Y_m)$ are independent random variables

$$\Rightarrow (\ \overline{X},\ \ s_X^2,\ \ \overline{Y},\ \ s_Y^2\ ) \text{ are independent random variables}$$

$$\Rightarrow (\ \overline{X},\ \ \overline{Y},\ \ s_p^2\ ) \text{ are independent random variables}$$

- $\overline{X} \sim N(\mu_X, \sigma^2/n) \Rightarrow \sqrt{n}(\overline{X} - \mu_X)/\sigma \sim N(0,1)$
- $\overline{Y} \sim N(\mu_Y, \sigma^2/m) \Rightarrow \sqrt{m}(\overline{Y} - \mu_Y)/\sigma \sim N(0,1)$
- $\overline{X} - \overline{Y} \sim N\left(\mu_X - \mu_Y, \frac{\sigma^2}{n} + \frac{\sigma^2}{m}\right) \Rightarrow \frac{(\overline{X}-\overline{Y})-(\mu_X-\mu_Y)}{\sigma\sqrt{\frac{1}{n}+\frac{1}{m}}} \sim N(0,1)$
- Since **(i)** $(n-1)s_X^2/\sigma^2 \sim \chi^2_{n-1}$, **(ii)** $(m-1)s_Y^2/\sigma^2 \sim \chi^2_{m-1}$, and **(iii)** s_X^2 and s_Y^2 are independent,

$$\frac{(n-1)s_X^2 + (m-1)s_Y^2}{\sigma^2} = \frac{(m+n-2)s_p^2}{\sigma^2} \sim \chi^2_{m+n-2}$$

Theorem 2 (log-likelihood, 2-sample normal model)

Under the two-sample normal model (∗) in LNp.6, the log-likelihood is proportional to (exercise)

$$l(\mu_X, \mu_Y, \sigma^2) \;\propto\; -\frac{m+n}{2}\log(\sigma^2) - \frac{\sum_{i=1}^n (X_i-\mu_X)^2}{2\sigma^2} - \frac{\sum_{j=1}^m (Y_j-\mu_Y)^2}{2\sigma^2}$$

$$= -\frac{1}{2\sigma^2}\left(\sum_{i=1}^n X_i^2 + \sum_{j=1}^m Y_j^2\right) + \frac{\mu_X}{\sigma^2}\left(\sum_{i=1}^n X_i\right) + \frac{\mu_Y}{\sigma^2}\left(\sum_{j=1}^m Y_j\right)$$

$$-[(m+n)/2]\log(\sigma^2) - (n\,\mu_X^2)/(2\sigma^2) - (m\,\mu_Y^2)/(2\sigma^2)$$

$\in$ 3-parameter exponential family

From the log-likelihood, we have

- $\frac{\partial l}{\partial \mu_X} = \frac{1}{\sigma^2}[(\sum_{i=1}^n X_i) - n \times \mu_X]$
- $\frac{\partial l}{\partial \mu_Y} = \frac{1}{\sigma^2}[(\sum_{j=1}^m Y_j) - m \times \mu_Y]$
- $\frac{\partial l}{\partial \sigma^2} = -\frac{m+n}{2\sigma^2} + \frac{\sum_{i=1}^n (X_i-\mu_X)^2}{2\sigma^4} + \frac{\sum_{j=1}^m (Y_i-\mu_Y)^2}{2\sigma^4}$

Theorem 3 (UMVUE and MLE of the parameters in the 2-sample normal model)

- $\left(R_1 = \sum_{i=1}^n X_i, R_2 = \sum_{j=1}^m Y_j, R_3 = \sum_{i=1}^n X_i^2 + \sum_{j=1}^m Y_j^2\right)$ is a sufficient and complete statistic (**Hint**. 3-parameter exponential family)
- $\overline{X}$ $(= R_1/n)$ is the UMVUE (by Lehmann-Scheffe Thm) and MLE of μ_X
- $\overline{Y}$ $(= R_2/m)$ is the UMVUE (by Lehmann-Scheffe Thm) and MLE of μ_Y

- (by Lehmann-Scheffe Thm) The pooled sample variance s_p^2 is the UMVUE of σ^2, since **(i)** s_p^2 is unbiased, and **(ii)**

$$(m+n-2)s_p^2 = (n-1)s_X^2 + (m-1)s_Y^2$$
$$= \sum_{i=1}^n (X_i-\overline{X})^2 + \sum_{j=1}^m (Y_j-\overline{Y})^2$$
$$= \left(\sum_{i=1}^n X_i^2\right) - n\overline{X}^2 + \left(\sum_{j=1}^m Y_j^2\right) - m\overline{Y}^2 = R_3 - (R_1^2/n) - (R_2^2/m)$$

- The MLE of σ^2 is $\frac{m+n-2}{m+n}\, s_p^2 = \frac{(n-1)s_X^2 + (m-1)s_Y^2}{m+n}$

Question 2 (how to claim Δ=0 or Δ≠0?)

Under the two-sample normal model (∗) in LNp.6, consider the parameter

$$\Delta = \mu_X - \mu_Y.$$

Notice that

$$\Delta = 0 \quad \Leftrightarrow \quad \text{no difference in the two samples}$$

- The UMVUE (by Lehmann-Scheffe Thm and $\hat{\Delta} = R_1/n - R_2/m$) and MLE of Δ is $\hat{\Delta} = \overline{X} - \overline{Y}$.
- But, $\hat{\Delta} \neq 0$ is not a strong enough evidence to reject $\Delta = 0$ (**Note**. $P(\hat{\Delta} \neq 0) = 1$). A better way is to examine if a C.I. of Δ contains 0.
- Q: how to construct an interval estimator for Δ?

Review 2 (pivotal quantity of θ)

A **pivotal quantity** for θ is a function of data $X_1, \ldots, X_n$ and the parameter θ, denoted by

$$Q(\mathbf{X}, \theta) = Q(X_1, \ldots, X_n, \theta),$$

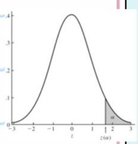
if the distribution of $Q(\mathbf{X}, \theta)$ is irrelevant to *all* parameters.

Theorem 4 (confidence interval of Δ, 2-sample normal model)

Under the two-sample normal model $(*)$ in LNp.6,

- σ^2 known (σ^2 is not a parameter)
  - a pivotal quantity of Δ is

$$Q_{Z,\Delta} = \frac{(\overline{X}-\overline{Y})-(\mu_X-\mu_Y)}{\sigma\sqrt{\frac{1}{n}+\frac{1}{m}}} = \frac{(\overline{X}-\overline{Y})-\Delta}{\sigma\sqrt{\frac{1}{n}+\frac{1}{m}}} \sim N(0,1)$$



  - a $100(1-\alpha)\%$ C.I. for Δ is $(\overline{X}-\overline{Y}) \pm z(\alpha/2) \times \left(\sigma\sqrt{\frac{1}{n}+\frac{1}{m}}\right)$ since

$$1-\alpha = P(\,|Q_{Z,\Delta}| < z(\alpha/2)\,)$$
$$= P\left((\overline{X}-\overline{Y}) - z(\alpha/2)\sigma\sqrt{\tfrac{1}{n}+\tfrac{1}{m}} < \Delta < (\overline{X}-\overline{Y}) + z(\alpha/2)\sigma\sqrt{\tfrac{1}{n}+\tfrac{1}{m}}\right)$$

- σ^2 unknown (σ^2 is a parameter)
  - a pivotal quantity of Δ is

$$Q_{T,\Delta} = \frac{(\overline{X}-\overline{Y})-(\mu_X-\mu_Y)}{s_p\sqrt{\frac{1}{n}+\frac{1}{m}}} = \frac{\left[(\overline{X}-\overline{Y})-\Delta\right]/\left(\sigma\sqrt{\frac{1}{n}+\frac{1}{m}}\right)}{\sqrt{\left[\frac{(m+n-2)s_p^2}{\sigma^2}\right]\frac{1}{m+n-2}}} \sim t_{m+n-2}$$

  - a $100(1-\alpha)\%$ C.I. for Δ is $(\overline{X}-\overline{Y}) \pm t_{m+n-2}(\alpha/2) \times \left(s_p\sqrt{\frac{1}{n}+\frac{1}{m}}\right)$

Note 2 (A note about the confidence intervals of Δ)

These confidence intervals are of the form

$$\text{(estimate)} \pm \text{(critical value)} \times [\text{(estimted) standard error}],$$

where the (estimated) standard error of $\overline{X}-\overline{Y}$ is $\sigma_{\overline{X}-\overline{Y}} = \sigma\sqrt{\frac{1}{n}+\frac{1}{m}}$ when σ^2 is known, and is $s_{\overline{X}-\overline{Y}} = s_p\sqrt{\frac{1}{n}+\frac{1}{m}}$ when σ^2 is unknown.

Example 2 (confidence interval of Δ, heat of fusion of ice, cont. Ex.1 in LNp.3)

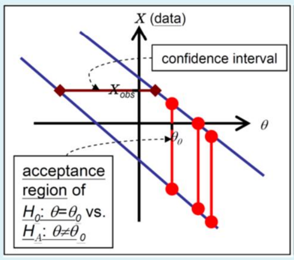
- $n = 13, \overline{X}_A = 80.02,\ s_A = 0.024; \quad m = 8, \overline{X}_B = 79.98,\ s_B = 0.031$
- $s_p = \sqrt{\frac{12}{19}\,s_A^2 + \frac{7}{19}\,s_B^2} = 0.027, \qquad s_{\overline{X}_A-\overline{X}_B} = s_p\sqrt{\frac{1}{13}+\frac{1}{8}} = 0.012$
- A 95% confidence interval for $\Delta = \mu_A - \mu_B$ is
  $(\overline{X}_A-\overline{X}_B) \pm t_{19}(0.025) \times s_{\overline{X}_A-\overline{X}_B} = (0.04) \pm (2.093) \times (0.012) = (0.015, 0.065)$.

Question 3 (how to perform testing of Δ=0?)

- **Recall.** duality between confidence interval and hypothesis testing
- **Q**: What are the hypothesis testings corresponding to these confidence intervals of Δ?

Theorem 5 (<u>z-test</u> and <u>t-test</u> for <u>Δ=Δ_0</u>, <u>2-sample</u> <u>normal</u> model)

Under the two-sample normal model ($*$) in LNp.6, consider the null and alternative hypotheses:
$$H_0 : \mu_X - \mu_Y = \Delta = \Delta_0 \quad \text{vs.} \quad H_A : \mu_X - \mu_Y = \Delta \neq \Delta_0$$
where Δ_0 is a known constant (**Note**. if $\Delta_0 = 0$, $H_0 : \mu_X = \mu_Y$ vs. $H_A : \mu_X \neq \mu_Y$), and H_A is a **two-sided alternative**. From the duality between C.I. and testing,
$$|Q_{Z_0,\Delta}| < z(\alpha/2) \quad \overset{\text{cf.}}{\longleftrightarrow} \quad |Q_{Z,\Delta_0}| < z(\alpha/2)$$
$$|Q_{T_0,\Delta}| < t_{m+n-2}(\alpha/2) \quad \overset{\text{cf.}}{\longleftrightarrow} \quad |Q_{T,\Delta_0}| < t_{m+n-2}(\alpha/2)$$



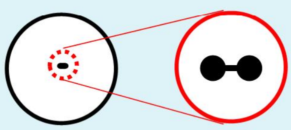
the corresponding test of these confidence intervals are:

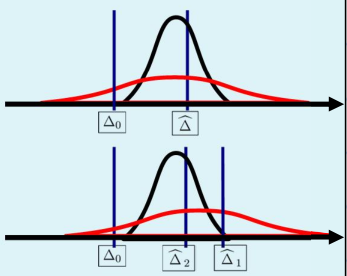
- test statistic
  - σ^2 known: $Z = \dfrac{(\overline{X} - \overline{Y}) - \Delta_0}{\sigma\sqrt{\frac{1}{n} + \frac{1}{m}}}$ $\left(\overset{\text{cf.}}{\longleftrightarrow} Q_{Z,\Delta} \text{ in LNp.11}\right)$
  - σ^2 unknown: $T = \dfrac{(\overline{X} - \overline{Y}) - \Delta_0}{s_p\sqrt{\frac{1}{n} + \frac{1}{m}}}$ $\left(\overset{\text{cf.}}{\longleftrightarrow} Q_{T,\Delta} \text{ in LNp.11}\right)$
- null distribution
  - σ^2 known: under H_0, $Z \sim N(0,1)$
  - σ^2 unknown: under H_0, $T \sim t_{m+n-2}$
- level-α rejection region
  - σ^2 known: $|Z| > z(\alpha/2)$, called z-test (reasonable?)
  - σ^2 unknown: $|T| > t_{m+n-2}(\alpha/2)$, called t-test (reasonable?)

Note. The t-test (or z-test) rejects H_0 if and only if its corresponding C.I. does not include Δ_0.

Note 3 (Some <u>notes</u> about <u>z-</u> and <u>t-tests</u>)

- For the null and alternative hypotheses:
$$H_0 : \Delta = \Delta_0 \text{ (or } \Delta \leqslant \Delta_0) \text{ vs. } H_A^* : \Delta > \Delta_0 \text{ (need domain knowledge)}$$
or
$$H_0 : \Delta = \Delta_0 \text{ (or } \Delta \geqslant \Delta_0) \text{ vs. } H_A^{**} : \Delta < \Delta_0 \text{ (need domain knowledge)}$$
where H_A^* and H_A^{**} are **one-sided alternatives**, the z- and t-tests are
  - σ^2 known: $Z > z(\alpha)$ for H_A^*, and $Z < -z(\alpha)$ for H_A^{**}
  - σ^2 unknown: $T > t_{m+n-2}(\alpha)$ for H_A^*, and $T < -t_{m+n-2}(\alpha)$ for H_A^{**}
- **FYI**. All the tests presented in LNp.13-14 are uniformly most powerful unbiased (UMPU) tests. (**Note**. Its proof follows a theorem of UMPU tests for exponential family with *nuisance parameters*)
- The test statistics are of the form:
$$\frac{(\overline{X} - \overline{Y}) - \Delta_0}{s_{\overline{X}-\overline{Y}}}.$$





  - In the numerator, $(\overline{X} - \overline{Y}) - \Delta_0$ estimates $\Delta - \Delta_0$.
  - **Q**: why is this estimate divided by $s_{\overline{X}-\overline{Y}}$ ($s_{\overline{X}-\overline{Y}} \downarrow$ when $m \uparrow$ and/or $n \uparrow$)?
- **Q**: if H_0 not rejected, do we really accept $\Delta = \Delta_0$, say $\mu_X = \mu_Y$? (better to claim "sample size is not large enough to reject H_0.")
- statistically significant difference vs. physically significant difference (example?)

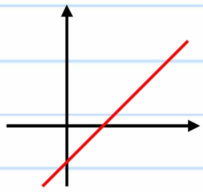
$$\text{statistical standard} \quad \overset{\text{different}}{\longleftrightarrow} \quad \text{physical standard}$$

Theorem 6 (likelihood ratio tests for $\Delta=\Delta_0$, 2-sample normal model)

All the tests presented in LNp.13-14 are likelihood ratio tests.

Proof: We only prove the case of two-sided hypothesis. For the case of one-sided hypothesis, its proof is similar (exercise).

- Recall that
  - the log-likelihood is
$$l = \log(\mathcal{L}) \propto -\frac{m+n}{2}\log(\sigma^2) - \frac{\sum_{i=1}^n (X_i-\mu_X)^2}{2\sigma^2} - \frac{\sum_{j=1}^m (Y_j-\mu_Y)^2}{2\sigma^2},$$
  - the test statistic of likelihood ratio test is
$$\Lambda = \frac{\sup_\omega \mathcal{L}}{\sup_\Omega \mathcal{L}} \quad \text{or} \quad \log(\Lambda) = \sup_\omega \log(\mathcal{L}) - \sup_\Omega \log(\mathcal{L}) = \sup_\omega l - \sup_\Omega l,$$
  where $\Omega = H_0 \bigcup H_A$ and $\omega = H_0$,
  - a likelihood ratio test rejects H_0 for small values of Λ (or $\log \Lambda$).
- σ^2 known
  - The parameter spaces Ω and ω are
$$\Omega = \{(\mu_X, \mu_Y) \mid \mu_X \in \mathbb{R},\ \mu_Y \in \mathbb{R}\}$$
$$\omega = \{(\mu_X, \mu_Y) \mid \mu_X \in \mathbb{R},\ \mu_Y = \mu_X - \Delta_0\}$$



  - Under Ω, the MLE's of (μ_X, μ_Y) are
$$\hat{\mu}_{X,\Omega} = \overline{X}, \qquad \hat{\mu}_{Y,\Omega} = \overline{Y},$$

$$\Rightarrow \sup_\Omega l = l(\hat{\mu}_{X,\Omega}, \hat{\mu}_{Y,\Omega}) \propto -\frac{\sum_{i=1}^n (X_i-\overline{X})^2 + \sum_{j=1}^m (Y_j-\overline{Y})^2}{2\sigma^2} - \frac{m+n}{2}\log(\sigma^2)$$

  - Under ω, the log-likelihood is proportional to
$$-\frac{m+n}{2}\log(\sigma^2) - \frac{\sum_{i=1}^n (X_i-\mu_X)^2}{2\sigma^2} - \frac{\sum_{j=1}^m [Y_j-(\mu_X-\Delta_0)]^2}{2\sigma^2},$$
  and the MLE's of (μ_X, μ_Y) are
$$\hat{\mu}_{X,\omega} = \frac{n}{m+n}\overline{X} + \frac{m}{m+n}(\overline{Y}+\Delta_0),$$
$$\hat{\mu}_{Y,\omega} = \hat{\mu}_{X,\omega} - \Delta_0 = \frac{n}{m+n}(\overline{X}-\Delta_0) + \frac{m}{m+n}\overline{Y},$$
$$\Rightarrow \sup_\omega l = l(\hat{\mu}_{X,\omega}, \hat{\mu}_{Y,\omega}) \propto -\frac{\sum_{i=1}^n (X_i-\hat{\mu}_{X,\omega})^2 + \sum_{j=1}^m (Y_j-\hat{\mu}_{Y,\omega})^2}{2\sigma^2} - \frac{m+n}{2}\log(\sigma^2)$$

  - Therefore, the log-likelihood-ratio is
$$\log(\Lambda) = l(\hat{\mu}_{X,\omega}, \hat{\mu}_{Y,\omega}) - l(\hat{\mu}_{X,\Omega}, \hat{\mu}_{Y,\Omega})$$
$$= -\frac{1}{2\sigma^2}\Big[\Big(\sum_{i=1}^n X_i^2 - 2n\overline{X}\hat{\mu}_{X,\omega} + n\hat{\mu}_{X,\omega}^2 + \sum_{j=1}^m Y_j^2 - 2m\overline{Y}\hat{\mu}_{Y,\omega} + m\hat{\mu}_{Y,\omega}^2\Big) - \Big(\sum_{i=1}^n X_i^2 - n\overline{X}^2 + \sum_{j=1}^m Y_j^2 - m\overline{Y}^2\Big)\Big]$$
$$= -\frac{1}{2\sigma^2}\left[n(\overline{X}-\hat{\mu}_{X,\omega})^2 + m(\overline{Y}-\hat{\mu}_{Y,\omega})^2\right]$$
$$= -\frac{1}{2\sigma^2}\left(\frac{mn}{m+n}\right)(\overline{X}-\overline{Y}-\Delta_0)^2$$

$$\overline{X} - \hat{\mu}_{X,\omega} = \frac{m}{m+n}(\overline{X}-\overline{Y}-\Delta_0)$$
$$\overline{Y} - \hat{\mu}_{Y,\omega} = \frac{-n}{m+n}(\overline{X}-\overline{Y}-\Delta_0)$$

- The likelihood ratio test rejects H_0 for

$$\text{small values of } \log(\Lambda) \quad \Leftrightarrow \quad \text{large values of } \left|(\overline{X}-\overline{Y})-\Delta_0\right|,$$

which is the z-test apart from constants that do not depend on the data.

- σ^2 unknown
  - The parameter spaces Ω and ω are

$$\begin{aligned}\Omega &= \{(\mu_X,\mu_Y,\sigma^2)\,|\,\mu_X\in\mathbb{R},\ \mu_Y\in\mathbb{R},\ \sigma^2>0\}\\ \omega &= \{(\mu_X,\mu_Y,\sigma^2)\,|\,\mu_X\in\mathbb{R},\ \mu_Y=\mu_X-\Delta_0,\ \sigma^2>0\}\end{aligned}$$

  - Under Ω, the MLE's of (μ_X,μ_Y,σ^2) are

$$\hat{\mu}_{X,\Omega}=\overline{X},\qquad \hat{\mu}_{Y,\Omega}=\overline{Y},$$

$$\hat{\sigma}^2_\Omega=\frac{1}{m+n}\left[\sum_{i=1}^n(X_i-\hat{\mu}_{X,\Omega})^2+\sum_{j=1}^m(Y_j-\hat{\mu}_{Y,\Omega})^2\right]=\frac{(n-1)S_X^2+(m-1)S_Y^2}{m+n}$$

$$\begin{aligned}\Rightarrow l(\hat{\mu}_{X,\Omega},\hat{\mu}_{Y,\Omega},\hat{\sigma}^2_\Omega) &\propto -\frac{m+n}{2}\log(\hat{\sigma}^2_\Omega)-\frac{\sum_{i=1}^n(X_i-\overline{X})^2+\sum_{j=1}^m(Y_j-\overline{Y})^2}{2\hat{\sigma}^2_\Omega}\\ &= -\frac{m+n}{2}\log(\hat{\sigma}^2_\Omega)-\frac{m+n}{2}\end{aligned}$$

  - Under ω, the log-likelihood is proportional to

$$-\frac{m+n}{2}\log(\sigma^2)-\frac{\sum_{i=1}^n(X_i-\mu_X)^2}{2\sigma^2}-\frac{\sum_{j=1}^m[Y_j-(\mu_X-\Delta_0)]^2}{2\sigma^2},$$

and the MLE's of (μ_X,μ_Y,σ^2) are

$$\hat{\mu}_{X,\omega} = \frac{n}{m+n}\overline{X}+\frac{m}{m+n}(\overline{Y}+\Delta_0),$$

$$\begin{aligned}\hat{\mu}_{Y,\omega} &= \hat{\mu}_{X,\omega}-\Delta_0=\frac{n}{m+n}(\overline{X}-\Delta_0)+\frac{m}{m+n}\overline{Y},\\ \hat{\sigma}^2_\omega &= \frac{1}{m+n}\left[\sum_{i=1}^n(X_i-\hat{\mu}_{X,\omega})^2+\sum_{j=1}^m(Y_j-\hat{\mu}_{Y,\omega})^2\right].\end{aligned}$$

$$\begin{aligned}\Rightarrow l(\hat{\mu}_{X,\omega},\hat{\mu}_{Y,\omega},\hat{\sigma}^2_\omega) &\propto -\frac{m+n}{2}\log(\hat{\sigma}^2_\omega)-\frac{\sum_{i=1}^n(X_i-\hat{\mu}_{X,\omega})^2+\sum_{j=1}^m(Y_j-\hat{\mu}_{Y,\omega})^2}{2\hat{\sigma}^2_\omega}\\ &= -\frac{m+n}{2}\log(\hat{\sigma}^2_\omega)-\frac{m+n}{2}\end{aligned}$$

  - Therefore, the log-likelihood-ratio is

$$\log(\Lambda)=l(\hat{\mu}_{X,\omega},\hat{\mu}_{Y,\omega},\hat{\sigma}^2_\omega)-l(\hat{\mu}_{X,\Omega},\hat{\mu}_{Y,\Omega},\hat{\sigma}^2_\Omega)=-\frac{m+n}{2}\log\left(\frac{\hat{\sigma}^2_\omega}{\hat{\sigma}^2_\Omega}\right)\quad\text{and}$$

$$\begin{aligned}\frac{\hat{\sigma}^2_\omega}{\hat{\sigma}^2_\Omega} &= \frac{\sum_{i=1}^n(X_i-\hat{\mu}_{X,\omega})^2+\sum_{j=1}^m(Y_j-\hat{\mu}_{Y,\omega})^2}{\sum_{i=1}^n(X_i-\overline{X})^2+\sum_{j=1}^m(Y_j-\overline{Y})^2}\\ &= \frac{\sum_{i=1}^n(X_i-\overline{X})^2+\sum_{j=1}^m(Y_j-\overline{Y})^2+\frac{mn}{m+n}(\overline{X}-\overline{Y}-\Delta_0)^2}{\sum_{i=1}^n(X_i-\overline{X})^2+\sum_{j=1}^m(Y_j-\overline{Y})^2}\\ &= 1+\frac{mn}{m+n}\times\frac{(\overline{X}-\overline{Y}-\Delta_0)^2}{\sum_{i=1}^n(X_i-\overline{X})^2+\sum_{j=1}^m(Y_j-\overline{Y})^2}\propto\frac{(\overline{X}-\overline{Y}-\Delta_0)^2}{(m+n-2)s_p^2}\end{aligned}$$

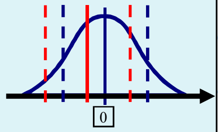
  - The likelihood ratio test rejects H_0 for

$$\text{small values of } \log(\Lambda) \quad \Leftrightarrow \quad \text{large values of } \frac{\left|(\overline{X}-\overline{Y})-\Delta_0\right|}{s_p},$$

which is the t-test apart from constants that do not depend on the data.

Theorem 7 (power of z-test, 2-sample normal model)

When σ^2 is known, the power function of a level-α z-test for $H_0: \Delta = \mu_X - \mu_Y = \Delta_0$ vs. $H_A: \Delta = \mu_X - \mu_Y \neq \Delta_0$ is



$$\beta_\Delta = 1 - \Phi\left(z(\alpha/2) - \frac{\Delta - \Delta_0}{\sigma\sqrt{\frac{1}{n}+\frac{1}{m}}}\right) + \Phi\left(-z(\alpha/2) - \frac{\Delta - \Delta_0}{\sigma\sqrt{\frac{1}{n}+\frac{1}{m}}}\right),$$

where $\Delta \in H_A$, Φ is the cdf of $N(0,1)$, and $z(\alpha/2)$ is the $(1-\alpha/2)$-quantile of $N(0,1)$.

Proof. power $\beta_\Delta = P\,(\text{rejection region} \mid \mu_X - \mu_Y = \Delta)$

$$= P\left(\left|\frac{(\overline{X}-\overline{Y})-\Delta_0}{\sigma\sqrt{\frac{1}{n}+\frac{1}{m}}}\right| > z\,(\alpha/2) \;\middle|\; \mu_X - \mu_Y = \Delta\right)$$

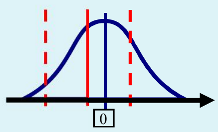
$$= P\left(\frac{(\overline{X}-\overline{Y})-\Delta}{\sigma\sqrt{\frac{1}{n}+\frac{1}{m}}} > z\,(\alpha/2) - \frac{\Delta-\Delta_0}{\sigma\sqrt{\frac{1}{n}+\frac{1}{m}}} \;\middle|\; \mu_X - \mu_Y = \Delta\right)$$

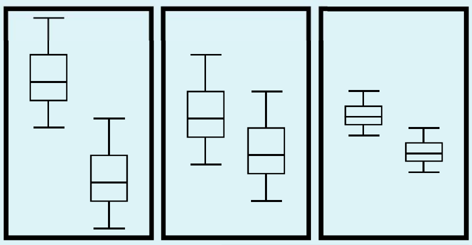
$$+\, P\left(\frac{(\overline{X}-\overline{Y})-\Delta}{\sigma\sqrt{\frac{1}{n}+\frac{1}{m}}} < -z\,(\alpha/2) - \frac{\Delta-\Delta_0}{\sigma\sqrt{\frac{1}{n}+\frac{1}{m}}} \;\middle|\; \mu_X - \mu_Y = \Delta\right),$$

where $[(\overline{X}-\overline{Y})-\Delta]\Big/\left(\sigma\sqrt{\frac{1}{n}+\frac{1}{m}}\right) \sim N(0,1)$ when $\mu_X - \mu_Y = \Delta$.

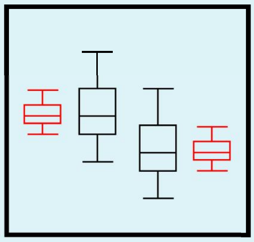
Note 4 (Some notes about the power function of z- and t-tests)

- The power $\beta_\Delta \uparrow 1$ as $\alpha \uparrow 1$ (reasonable?) or $\dfrac{|\Delta - \Delta_0|}{\sigma\sqrt{\frac{1}{n}+\frac{1}{m}}} \uparrow \infty$, i.e.,
  - as $|\Delta - \Delta_0|$ increases (reasonable?), or
  - as σ decreases (reasonable?), or
  - as n, m increase (reasonable?).





- When σ^2 is unknown, the exact power of the t-test can be similiarly calculated. But, this calculation requires the use of *noncentral t* distribution.



- **Sample size determination** using power.
  - The necessary sample sizes can be determined from α, σ, Δ, and β_Δ.
  - For example, when σ^2 is known and $n = m$,

$$\beta_\Delta = 1 - \underbrace{\Phi\left(z\left(\frac{\alpha}{2}\right) - \frac{\Delta-\Delta_0}{\sigma}\sqrt{\frac{n}{2}}\right)}_{\text{first term}} + \underbrace{\Phi\left(-z\left(\frac{\alpha}{2}\right) - \frac{\Delta-\Delta_0}{\sigma}\sqrt{\frac{n}{2}}\right)}_{\text{second term}}$$

  * Usually, one of these terms is negligible with respect to the other.

* If $\Delta - \Delta_0 > 0$, the first term will be dominant and

$$\beta_\Delta \approx 1 - \Phi\left(z\left(\frac{\alpha}{2}\right) - \frac{\Delta - \Delta_0}{\sigma}\sqrt{\frac{n}{2}}\right)$$

$$\Rightarrow \quad z(\beta_\Delta) \approx z\left(\frac{\alpha}{2}\right) - \frac{\Delta - \Delta_0}{\sigma}\sqrt{\frac{n}{2}}$$

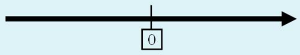
$$\Rightarrow \quad n \approx [z(\alpha/2) - z(\beta_\Delta)]^2 \frac{2\sigma^2}{(\Delta - \Delta_0)^2}$$

Q: What if $\Delta - \Delta_0 < 0$? (exercise)

- Determining sample sizes using power is equivalent to using length of C.I. For example, consider the case of σ^2 known.
  - Suppose that m, n are such that the half-length of the C.I. for Δ is $L_{n,m}$.
  - From Thm 4 (LNp.11), $L_{n,m} = z(\alpha/2)\,\sigma\sqrt{\frac{1}{n} + \frac{1}{m}}$.
  - From Thm 7 (LNp.19), the corresponding power function of $L_{n,m}$ is

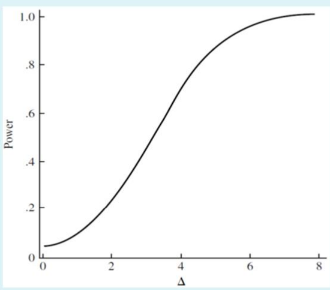
$$\beta_\Delta = 1 - \Phi\left(z(\alpha/2) - \frac{\Delta - \Delta_0}{L_{n,m}} z(\alpha/2)\right) + \Phi\left(-z(\alpha/2) - \frac{\Delta - \Delta_0}{L_{n,m}} z(\alpha/2)\right)$$

  - This property could be used to suggest sample sizes m, n under which the statistical standard (LNp.14) is more consistent with the physical standard.



Example 3 (power function and sample size determination)

- Figure 11.6 (textbook, p.435) gives the power function β_Δ when $n = m = 18$, $\sigma = 5$, $\Delta_0 = 0$, $\alpha = 0.05$ ($\Rightarrow z(\alpha/2) = 1.96$).



- Suppose we want to detect a difference of $\Delta = 1$ with probability (power β_Δ) 0.9. The sample size should be such that

$$0.1 = 1 - \beta_\Delta \approx \Phi\left(1.96 - (\Delta/\sigma)\sqrt{n/2}\right)$$

  - Solving for n, we find that the necessary sample size would be 525!
  - This is a consequence of a large $\sigma = 5$ relative to the difference $\Delta = 1$.
  - If the experimenters want to detect such a difference with a smaller sample size, some modification of the experimental technique to reduce σ would be necessary.

Note 5 (Some notes about z- and t-tests when 2-sample normal model does not hold)

- **Q**: Can we use z- or t-tests (or their corresponding C.I.) when the underlying distributions F, G of X, Y are not normal?

  Ans: Yes, if the sample sizes m, n are large. But, **why**? Consider the model:

$$\left.\begin{array}{llll} \text{1st sample:} & X_1, \ldots, X_n & \sim & \text{i.i.d. from } F \\ \text{2nd sample:} & Y_1, \ldots, Y_m & \sim & \text{i.i.d. from } G \end{array}\right\} \Leftarrow \text{independent}$$

  where F and G can be *any* continuous distributions with *same finite* variance.

- Denote the means of F, G by μ_X, μ_Y, respectively, and their (identical) variance by σ^2 $(<\infty)$. Consider testing $H_0: \mu_X = \mu_Y$ vs. $H_A: \mu_X \neq \mu_Y$.
- By CLT and LLN, when $m \to \infty$ and $n \to \infty$,

$$\overline{X} \overset{D}{\approx} N\left(\mu_X, \frac{\sigma^2}{n}\right), \ \overline{Y} \overset{D}{\approx} N\left(\mu_Y, \frac{\sigma^2}{m}\right) \Rightarrow \overline{X} - \overline{Y} \overset{D}{\approx} N\left(\mu_X - \mu_Y, \sigma^2\left(\frac{1}{n} + \frac{1}{m}\right)\right)$$

$$s_X^2 \xrightarrow{P} \sigma^2, \ s_Y^2 \xrightarrow{P} \sigma^2 \ \Rightarrow \ s_p^2 = \frac{n-1}{m+n-2} s_X^2 + \frac{m-1}{m+n-2} s_Y^2 \xrightarrow{P} \sigma^2$$

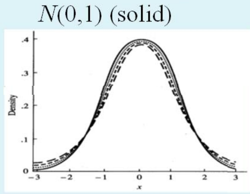
- Thus,
  - when σ^2 is known, $\quad Q_{Z,\Delta} = \dfrac{(\overline{X} - \overline{Y}) - (\mu_X - \mu_Y)}{\sigma\sqrt{\left(\frac{1}{n} + \frac{1}{m}\right)}} \overset{D}{\approx} N(0,1)$
  - when σ^2 is unknown,

$$Q_{T,\Delta} = \frac{(\overline{X} - \overline{Y}) - (\mu_X - \mu_Y)}{s_p\sqrt{\left(\frac{1}{n} + \frac{1}{m}\right)}} = \frac{\left[(\overline{X} - \overline{Y}) - (\mu_X - \mu_Y)\right] \Big/ \sqrt{\sigma^2\left(\frac{1}{n} + \frac{1}{m}\right)}}{\sqrt{s_p^2/\sigma^2}}$$

$$\overset{D}{\approx} \ N(0,1) \quad \text{(by Slutsky's Thm)}$$

and t_{m+n-2} tends to $N(0,1)$ as $m, n \to \infty$.

t_5 (long dash), t_{10} (short dash), t_{30} (dot), $N(0,1)$ (solid)



- **Q**: How to modify the z- and t-tests (or their corresponding C.I.) when the equal variance assumption in the 2-sample normal model does not hold? Consider the model:

$$\left.\begin{array}{llll} \text{1st sample:} & X_1, \ldots, X_n & \sim & \text{i.i.d. } N(\mu_X, \sigma_X^2) \\ \text{2nd sample:} & Y_1, \ldots, Y_m & \sim & \text{i.i.d. } N(\mu_Y, \sigma_Y^2) \end{array}\right\} \Leftarrow \text{independent}$$

where σ_X^2 and σ_Y^2 can be different. Consider $H_0: \mu_X = \mu_Y$ vs. $H_A: \mu_X \neq \mu_Y$.

- This model has 4 parameters: $\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2$, and their intuitive estimators are $\quad \overline{X} \xrightarrow{e} \mu_X, \quad \overline{Y} \xrightarrow{e} \mu_Y, \quad s_X^2 \xrightarrow{e} \sigma_X^2, \quad s_Y^2 \xrightarrow{e} \sigma_Y^2.$
- Under this model,
  - $\overline{X} - \overline{Y} \xrightarrow{e} \Delta = \mu_X - \mu_Y$ and $\overline{X} - \overline{Y} \sim N\left(\mu_X - \mu_Y, \frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}\right)$
  - when σ_X^2, σ_Y^2 are known, $\quad Q^*_{Z,\Delta} = \dfrac{(\overline{X} - \overline{Y}) - (\mu_X - \mu_Y)}{\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}} \sim N(0,1),$
  - when σ_X^2, σ_Y^2 are unknown,
    - $Var(\overline{X} - \overline{Y}) = \frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}$ can be estimated by $s^2_{\overline{X}-\overline{Y}} = \frac{s_X^2}{n} + \frac{s_Y^2}{m}$
    - it has been shown (Welch, 1938) the distribution of $\dfrac{(s_X^2/n) + (s_Y^2/m)}{(\sigma_X^2/n) + (\sigma_Y^2/m)}$ can be approximated by χ^2_ν/ν where

$$\nu = \frac{\left[(\sigma_X^2/n) + (\sigma_Y^2/m)\right]^2}{\frac{(\sigma_X^2/n)^2}{n-1} + \frac{(\sigma_Y^2/m)^2}{m-1}},$$

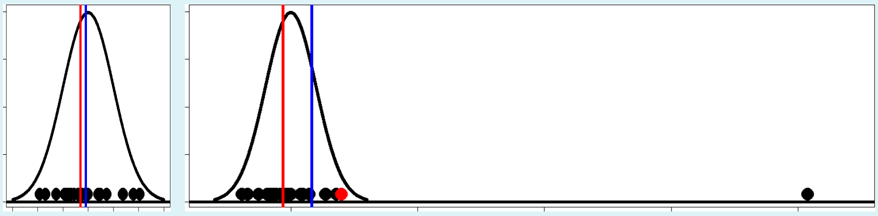
    - the degrees of freedom ν can be estimated by $\hat{\nu} = \dfrac{\left[(s_X^2/n) + (s_Y^2/m)\right]^2}{\frac{(s_X^2/n)^2}{n-1} + \frac{(s_Y^2/m)^2}{m-1}}$ and then rounded to the nearest integer,
    - thus,

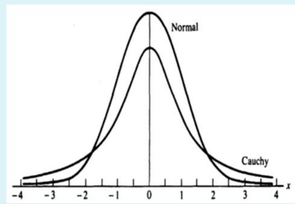
$$Q^*_{T,\Delta} = \frac{(\overline{X} - \overline{Y}) - (\mu_X - \mu_Y)}{\sqrt{\frac{s_X^2}{n} + \frac{s_Y^2}{m}}} = \frac{\left[(\overline{X} - \overline{Y}) - (\mu_X - \mu_Y)\right] \Big/ \sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}}{\sqrt{\left(\frac{s_X^2}{n} + \frac{s_Y^2}{m}\right) \Big/ \left(\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}\right)}} \overset{D}{\approx} t_\nu$$

and ν can be substituted by its integer estimate.

Note 6 (The circumstances under which z- and t-tests may be invalid)

- The distributions F, G of X, Y are not normal, and sample sizes are small ($\Rightarrow$ the null distribution of the statistic T (or Z) might not be close to t (or $N(0,1)$) distribution.)
- Data contains *outliers* (extreme values)
  - Problem of averaging data, such as $\overline{X}$, $\overline{Y}$, $\hat{\sigma}^2$: sensitive to extreme values



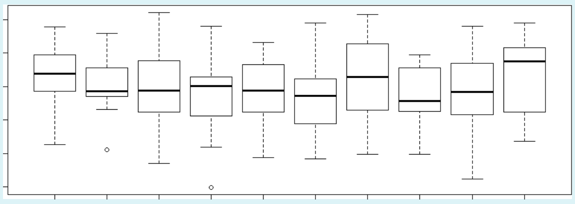


In contrast, median of data is insensitive to extreme values.

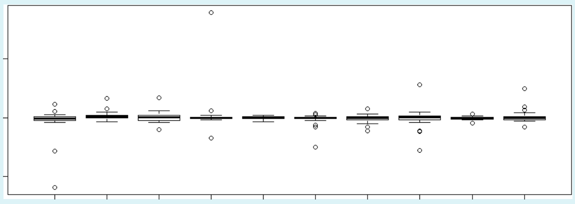
  - **Q**: In what distributions, observing extreme values is often expected?

Ans. heavy-tail distributions, such as Cauchy $C(\mu, \sigma)$.

$X_1, \ldots, X_{20} \sim$ i.i.d. $N(0,1)$ (repeat 10 times)



$X_1, \ldots, X_{20} \sim$ i.i.d. $C(0,1)$ (repeat 10 times)



  - Some properties of Cauchy distribution
    - it does not have finite moments of any order (a consequence of heavy tail)

- if $X_1, \ldots, X_n$ are i.i.d $\sim C(\mu, \sigma)$, then $\overline{X} \sim C(\mu, \sigma)$.

  ($\overset{\text{cf.}}{\longleftrightarrow}$ if $X_1, \ldots, X_n$ are i.i.d. from a distribution with *finite* variance σ^2, then $Var(\overline{X}) = \sigma^2/n \to 0$ when $n \to \infty$.)

  $\Rightarrow$ For 2-sample data, when F, G are Cauchy, even though the sample sizes m, n are large, the property "$\overline{X}, \overline{Y} \overset{D}{\approx}$ normal" does not hold.

  (**Q**: Why do LLN and CLT not apply to Cauchy?)

Question 4.

How to develop statistical methods for the circumstances under which z- and t-tests are inappropriate?

Q: What limits the validity of the tests?

Ans. Statistical models (i.e., joint distributions of data) covered in the 2-sample normal model (a *model space* of dimension *three*) are still not flexible enough to reflect the pattern of data

$\Rightarrow$ should include more joint distributions into the model space, say enlarge the model space to allow for F, G being *any* distributions

$\Rightarrow$ develop statistical methods under this enlarged model space

$\Rightarrow$ such statistical methods should be suitable for data of any patterns

❖ **Reading**: textbook, 11.1, 11.2.1, 11.2.2

- A nonparametric method for 2-sample problem --- Mann-Whitney Test

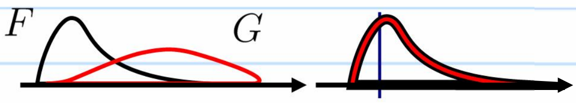
Question 5.

- In the materials taught before, we usually assume, in the statistical modeling, that the data follows a *particular* joint distribution which contains some unknown parameters of *finite* dimension.
- The statistical inferences, estimation and testing, are then based on a formulation of these parameters.

Q: What if we do not have any knowledge about the *particular* form of the joint distribution of data?

Consider the problem of 2-sample comparison.

- Let Ω be the collection of *all* continuous distributions
- Only assume that $F, G \in \Omega$
- Thus, the statistical model is:



$$\left.\begin{array}{llcl} \text{1st sample:} & X_1, \ldots, X_n & \sim & \text{i.i.d. from } F \\ \text{2nd sample:} & Y_1, \ldots, Y_m & \sim & \text{i.i.d. from } G \end{array}\right\} \Leftarrow \text{independent} \qquad (\square)$$

- This model contains parameters of *infinitely many* dimension because

$$\dim(\Omega) = \infty \quad (\textbf{why}?)$$

- Under this model, a 2-sample comparison examines the null and alternative hypotheses: $H_0 : F = G$ vs. $H_A : F \neq G$.

Definition 2 (nonparametric models and nonparametric methods)

- Nonparametric models do not assume any *particular* distributional form. Nonparametric models can be viewed as having *infinitely many* parameters. ($\stackrel{\text{cf.}}{\longleftrightarrow}$ parametric models: parameters are of *finite* dimension)
- Statistical methods developed under nonparametric models are called *nonparametric methods*.

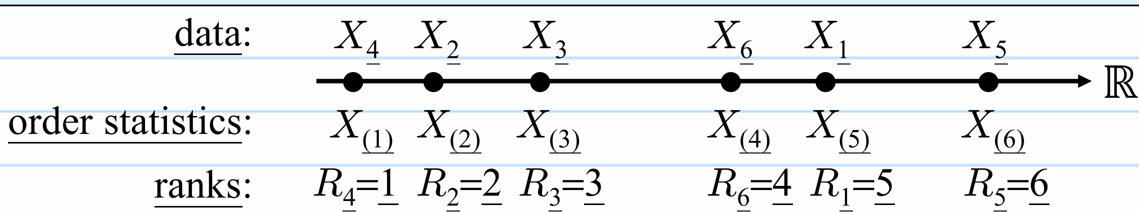
Review 3 (order statistics and ranks)

- Let $X_1, X_2, \ldots, X_n$ be random variables. We sort the X_i's and denote by $X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}$ the *order statistics*. Using the notation,

$$\begin{aligned} X_{(1)} &= \min(X_1, X_2, \ldots, X_n) \quad \text{is the } \textit{minimum}, \\ X_{(n)} &= \max(X_1, X_2, \ldots, X_n) \quad \text{is the } \textit{maximum}. \end{aligned}$$

- Let $R(X_1, X_2, \ldots, X_n) = (R_1, R_2, \ldots, R_n)$ such that $X_i = X_{(R_i)}$, $i = 1, \ldots, n$. Then, $(R_1, R_2, \ldots, R_n)$ is called the *ranks* of $X_1, X_2, \ldots, X_n$. Notice that

$$R_i = \sum_{j=1}^{n} \delta(X_i - X_j), \quad \text{where } \delta(t) = \begin{cases} 1, & \text{if } t \geq 0, \\ 0, & \text{if } t < 0. \end{cases}$$



Theorem 8 (sufficient and complete statistics for nonparametric models)

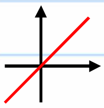
Let $X_1, \ldots, X_n$ be i.i.d. from F, where $F \in \Omega$.
Then, $(X_{(1)}, X_{(2)}, \ldots, X_{(n)})$ is sufficient and complete.
$\Big(\overset{\text{cf.}}{\longleftrightarrow}\ X_1, \ldots, X_n \sim$ i.i.d. $N(\mu, \sigma^2) \quad \Rightarrow \quad (\overline{X}, s_X^2)$ is sufficient and complete.$\Big)$

Proof. Denote the pdf of F by f. The joint pdf of $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$ is

$$f_{X_{(1)},X_{(2)},\ldots,X_{(n)}}(x_1, x_2, \ldots, x_n) = n! \times f(x_1)f(x_2)\cdots f(x_n),$$

for $x_1 < x_2 < \cdots < x_n$ and zero, otherwise.
The proof of sufficiency follows from the fact that the conditional probability of $X_1, \ldots, X_n$ given $X_{(1)}, \ldots, X_{(n)}$ is $\frac{1}{n!}$, which is irrelevant to F.

The proof of completeness is omitted (out of the scope of this course).



Note 7 (Some notes about order statistics and ranks)

- Order statistics and ranks are defined precisely, i.e., **no ties**, under the condition $P(X_i = X_j) = 0$, $i \neq j$ (**Note**. this condition holds when $X_1, \ldots, X_n \sim$ i.i.d. from F and F is a continuous distribution).
- Under Ω, the dimension of data (i.e., n) cannot be reduced without losing the information about F ($\in \Omega$).
- Under 1-sample model, ranks + order statistics = complete data
- Order statistics are intuitive estimator of quantiles, e.g., median.

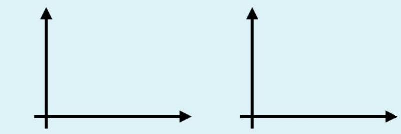
- Ranks are invariant under any monotonic transformation of data, i.e.,

$$R(X_1, \ldots, X_n) = R(H(X_1), \ldots, H(X_n)),$$

if H is a monotone increasing function and

$$R(X_1, \ldots, X_n) = (n+1) - R(H(X_1), \ldots, H(X_n)),$$

if H is a monotone decreasing function. $\big(\overset{\text{cf.}}{\longleftrightarrow}$ z- or t-tests may change significantly under monotonic transformations of data$\big)$.



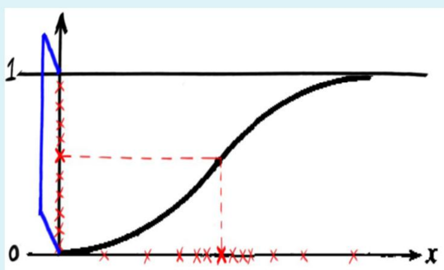
- Replacing the data by their ranks also has the effect of moderating the influence of outliers.
- Many nonparametric methods are based on order statistics and/or ranks.
- **Q**: Why are many nonparametric methods based on replacement of the data by ranks? What information of data are contained in their ranks?

(exercise) – **Recall**. Let $X_1, \ldots, X_n$ be i.i.d. from a continuous cdf F, and let $U_i = F(X_i)$, $i = 1, \ldots, n$. Then, $U_1, \ldots, U_n$ are i.i.d. from $U(0,1)$.

(exercise) – **Recall**. If $U_1, \ldots, U_n \sim$ i.i.d. $U(0,1)$, the pdf of the ith-order statistic $U_{(i)}$ is

$$f_{U_{(i)}}(u) = \frac{n!}{(i-1)!(n-i)!} u^{i-1}(1-u)^{n-i},$$

for $0 < u < 1$ and zero, otherwise.
Note that $E(U_{(i)}) = i/(n+1)$.



– $U_i = F(X_i)$ is not a statistic because F is an unknown function.

– But,

$$X_i = X_{(R_i)} \to U_{(R_i)} = F(X_{(R_i)}) \to R_i = (n+1)\frac{R_i}{n+1} \leftrightarrow (n+1)E\big[\, U_{(R_i)} \big|\, R_i \big].$$

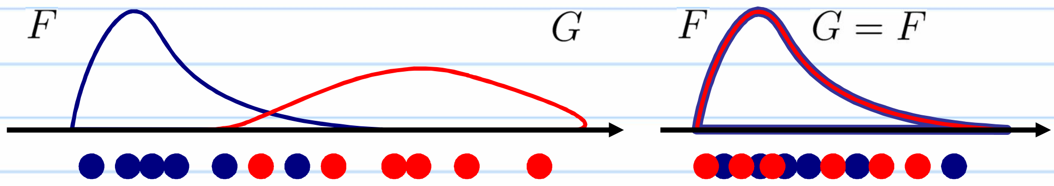
Question 6.

How to use ranks to compare two samples? Under the nonparametric model (□) in LNp.27, for the null and alternative hypotheses:

$$H_0 : F = G \quad \text{vs.} \quad H_A : F \neq G$$

what data are "more extreme," i.e., cast more doubts on H_0?

Intuition.

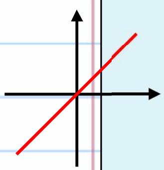


Theorem 9 (Mann-Whitney test or Wilcoxon rank sum test)

Consider the nonparametric model (□) in LNp.27.

- Pool all $m+n$ observations (i.e., $X_1, \ldots, X_n, Y_1, \ldots, Y_m$) together and rank them in order of increasing size, i.e.,

$$R(X_1, \ldots, X_n, Y_1, \ldots, Y_m) = (R_1, \ldots, R_n, R_{n+1}, \ldots, R_{m+n}).$$



- Test statistic W_X (or W_Y)
  – Let $W_X = \sum_{i=1}^{n} R_i$ and $W_Y = \sum_{j=1}^{m} R_{n+j}$. They are respectively the sums of the ranks of X_i's and Y_j's in the pooled data. Notice that

$$W_X + W_Y = 1 + 2 + \cdots + (m+n) = \frac{(m+n)(m+n+1)}{2}$$
$$\Rightarrow W_Y = \frac{(m+n)(m+n+1)}{2} - W_X.$$

  – Data with larger or smaller W_X are more extreme $\Rightarrow$ tend to reject H_0
- Null distribution of W_X
  – Under H_0 $(F = G)$,

$$\begin{array}{ccccccc} X_1, & \ldots, & X_n, & Y_1, & \ldots, & Y_m & \sim \text{ i.i.d. } F \\ \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \\ R_1, & \ldots, & R_n, & R_{n+1} & \ldots, & R_{m+n} & \sim \ ? \end{array}$$

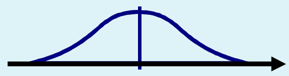
  – Any assignments of the ranks $\{1, \ldots, m+n\}$ to the pooled $m+n$ data are equally likely, and the total number of different assignments is $(m+n)!$.
  – Joint distribution of $R_1, \ldots, R_n$:
    * Consider an urn containing $m+n$ balls, labelled by $1, 2, \ldots, m+n$, respectively.
    * Sequentially draw n balls without replacement from the urn $\Rightarrow$ there are $\binom{m+n}{n} \times n!$ different outcomes, each with equal probability

* $*$ Let $r_1, r_2, \ldots, r_n$ be the numbers on the 1st, 2nd, …, nth balls drawn, respectively. Then,

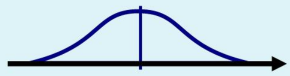
$$P(R_1 = r_1, \ldots, R_n = r_n) = \frac{1}{\binom{m+n}{n} \times n!} = \frac{m!}{(m+n)!}.$$

– The null distribution of $W_X = R_1 + \cdots + R_n$ (W_X is the sum of the numbers on the n balls) can be obtained from the joint distribution of $R_1, \ldots, R_n$.

- Rejection region
  – Let $n_1 = \min(n, m)$ be the smaller sample size, and W be the rank sum from that sample (i.e., $W = W_X$ if $n \le m$ and $W = W_Y$ if $n > m$).
  - $*$ Note that under H_0,
    - $\cdot$ $$E(W) = \left\{ \begin{array}{ll} E(R_1) + \cdots + E(R_n), & \text{if } n \le m \\ E(R_{n+1}) + \cdots + E(R_{n+m}), & \text{if } n > m \end{array} \right\} = \frac{n_1(m+n+1)}{2}.$$
    - $\cdot$ the null distribution of W is symmetric around $E(W)$ (exercise).
  - $*$ Let $W' = n_1(m+n+1) - W$.
  - $*$ Let $W^* = \min(W, W')$.



  – Reject H_0 when W^* is small, i.e., $W^* \le w$.



  – Table 8 of Appendix B in the textbook gives critical values w for W^*.

- We have assumed here that there are no ties among the observations. If there are only a small number of ties; tied observations are assigned average ranks.

Example 4 (Mann-Whitney test, heat of fusion of ice, cont. Ex.1 in LNp.3)

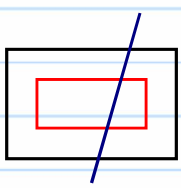
- The ranks are (ties ⇒ average rank)

Method A	7.5	19.0	11.5	19.0	15.5	15.5	19.0	4.5
	21.0	15.5	11.5	9.0	11.5			
Method B	11.5	1.0	7.5	4.5	4.5	15.5	2.0	4.5

- $n_1 = 8$, $W = W_B = 51$, $W' = 8(8+13+1) - W = 125$, $W^* = \min(W, W') = 51$
- two-sided test at level $\alpha = 0.01$, critical value $= 53$
  two-sided test at level $\alpha = 0.05$, critical value $= 60$
- Therefore, the Mann-Whitney test rejects the null hypothesis at $\alpha = 0.01$.

a comparison of parametric and nonparametric models

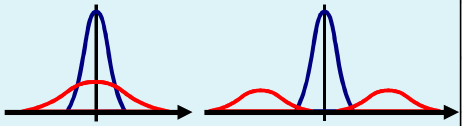
	model space	data reduction	robustness	power on H_A^p	power on $H_A^{np} \setminus H_A^p$
parametric models	small	low-dim	worse	higher	(usually) lower
nonparametric models	large	high-dim	better	lower	(usually) higher



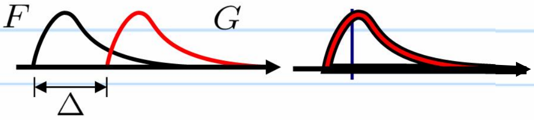
Question 7.

Does Mann-Whitney test have reasonably good powers over the whole $H_A: F \neq G$? Note that

$$H_0 \cup H_A = \{(F,G)\,|\,F,G \in \Omega\},$$
$$H_0 = \{(F,G)\,|\,F \in \Omega, G = F\}.$$



- Assume that the distributions (cdfs) $F, G \in \Omega$ and F, G have same shape.



  – If $X \sim F$ and $Y = X + \Delta$, where Δ is an unknown constant, then for the cdf $G(y)$ of Y, we have

$$G(y) \equiv P(Y \leq y) = P(X + \Delta \leq y) = P(X \leq y - \Delta) = F(y - \Delta),$$

and for the pdfs $f(x)$ of X and $g(y)$ of Y, we have

$$g(y) = \frac{d}{dy}G(y) = \frac{d}{dy}F(y - \Delta) = f(y - \Delta).$$

- Thus, the statistical model is:

$$\left.\begin{array}{lll} \text{1st sample:} & X_1, \ldots, X_n & \sim \text{ i.i.d. from } F \\ \text{2nd sample:} & Y_1, \ldots, Y_m & \sim \text{ i.i.d. from } G \end{array}\right\} \Leftarrow \text{independent} \qquad (\Diamond)$$

where $F \in \Omega$ and $G(x) = F(x - \Delta)$.

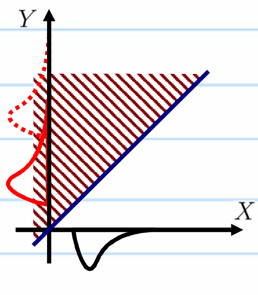
  – This model contains *infinitely many* parameters because $\dim(\Omega) = \infty$.

  – Under this model, the null $H_0: F = G$ become $H_0: \Delta = 0$, and the alternative $H_A: F \neq G$ becomes $H_A: \Delta \neq 0$, i.e.,

  * $H_0 \cup H_A = \{(F,G)\,|\,F \in \Omega,\ G(y) = F(y - \Delta),\ \Delta \in \mathbb{R}$ (or $\pi_\Delta \in [0,1]$)$\}$
  * $H_0 = \{(F,G)\,|\,F \in \Omega,\ G(y) = F(y - \Delta),\ \Delta = 0$ (or $\pi_\Delta = 1/2$)$\}$

Theorem 10 (An alternative formulation of H_0 and H_A)

- Suppose that (1) $X \sim F \in \Omega$, (2) $Y \sim G$, where $G(x) = F(x - \Delta)$, and (3) X, Y are independent. The joint pdf of (X, Y) is $f(x)g(y) = f(x)f(y - \Delta)$.
- Define $\pi_\Delta = P_\Delta(X < Y)$. Clearly, $0 \leq \pi_\Delta \leq 1$.
- Then, $\pi_\Delta = 1/2$ if and only if $\Delta = 0$.

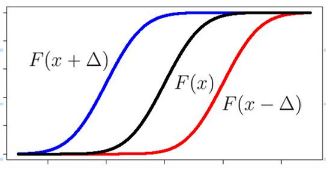


Proof.

- $P_\Delta(X < Y) = \int_{-\infty}^{\infty}\int_{x}^{\infty} f(x)\, f(y - \Delta)\, dy\, dx$

$$= \int_{-\infty}^{\infty} f(x)\, \left[F(y - \Delta)\big|_x^{\infty}\right] dx$$

$$= \int_{-\infty}^{\infty} [1 - F(x - \Delta)]\, f(x)\, dx = 1 - \int_{-\infty}^{\infty} F(x - \Delta)\, f(x)\, dx$$



- If $\Delta > 0$, then $F(x - \Delta) \leqslant F(x) \leqslant F(x + \Delta)$, $\forall\, x$, and there must exist a region A of x in which the inequalities are strict and $\int_A f(x)\, dx > 0$.

- Thus, for $\Delta > 0$,

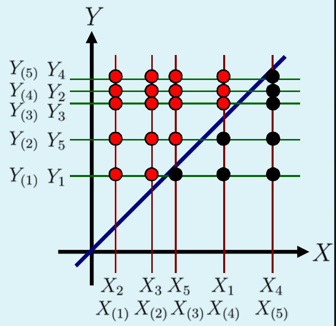
$$\underbrace{\int_{-\infty}^{\infty} F(x-\Delta)\, f(x)\, dx}_{1-P_\Delta(X<Y)} < \underbrace{\int_{-\infty}^{\infty} F(x)\, f(x)\, dx}_{1-P_{\Delta=0}(X<Y)} < \underbrace{\int_{-\infty}^{\infty} F(x+\Delta)\, f(x)\, dx}_{1-P_{-\Delta}(X<Y)}$$

- Then, the results follow from: $\displaystyle\int_{-\infty}^{\infty} F(x)\, f(x)\, dx = \int_0^1 z\, dz = \frac{1}{2} z^2 \Big|_0^1 = \frac{1}{2}.$

Theorem 11 (An alternative view of Mann-Whitney test)

Consider the nonparametric model (◊) in LNp.35.

- Estimation of π_Δ: the parameter $\pi_\Delta = P_\Delta(X < Y)$ can be estimated by the proportion of the comparisons for which X was less than Y, i.e.,
  - consider any pairs (X_i, Y_j), $1 \le i \le n$, $1 \le j \le m$,
  - let $Z_{ij} = \begin{cases} 1, & \text{if } X_i < Y_j, \\ 0, & \text{otherwise}, \end{cases} \Rightarrow \hat{\pi}_\Delta = \dfrac{1}{mn} \displaystyle\sum_{i=1}^{n} \sum_{j=1}^{m} Z_{ij}$

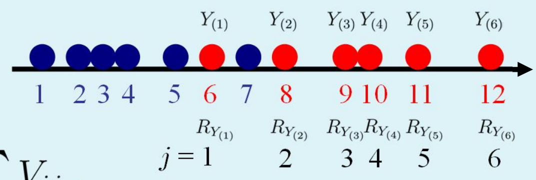


  - an alternative expression: consider the mn pairs $\left(X_{(i)}, Y_{(j)}\right)$, and let

$$V_{ij} = \begin{cases} 1, & \text{if } X_{(i)} < Y_{(j)}, \\ 0, & \text{otherwise}, \end{cases} \Rightarrow \hat{\pi}_\Delta = \frac{1}{mn} \sum_{i=1}^{n} \sum_{j=1}^{m} V_{ij}$$

$$\begin{aligned} \hat{\pi}_\Delta &= \tfrac{1}{mn}(\#\{X_i < Y_j\}) \\ &= \tfrac{1}{mn}(\#\{X_{(i)} < Y_{(j)}\}) \end{aligned}$$

- Test $H_0 : \pi_\Delta = 1/2$ ($\Leftrightarrow \Delta = 0$) vs. $H_A : \pi_\Delta \neq 1/2$ ($\Leftrightarrow \Delta \neq 0$)
  - intuitively, should reject H_0 if $\hat{\pi}_\Delta$ is too small (closer to 0) or too large (closer to 1)
  - test statistic U_Y (or U_X)



    * Define

$$U_Y \equiv (mn)\, \hat{\pi}_\Delta = \sum_{i=1}^{n} \sum_{j=1}^{m} Z_{ij} = \sum_{i=1}^{n} \sum_{j=1}^{m} V_{ij}\,.$$

      · Reject H_0 if U_Y is too small or too large (closer to 0 or mn).

    * Let $R_{Y_{(j)}}$ be the rank of $Y_{(j)}$ in the *pooled* sample. Then,

$$\sum_{j=1}^{m} R_{Y_{(j)}} = \text{rank sum of } Y_j\text{'s (or } Y_{(j)}\text{'s)} = R_{n+1} + \cdots + R_{n+m} = W_Y.$$

    * Notice that

$$U_Y = \sum_{j=1}^{m} \left(\sum_{i=1}^{n} V_{ij} \right) = \sum_{j=1}^{m} \underbrace{(R_{Y_{(j)}} - j)}_{\#\{X_{(i)} < Y_{(j)}\}} = \left(\sum_{j=1}^{m} R_{Y_{(j)}} \right) - \frac{m(m+1)}{2} = W_Y - [m(m+1)]/2.$$

    * Similarly, U_X can be defined by changing "$X_{(i)} < Y_{(j)}$" in V_{ij} to "$X_{(i)} > Y_{(j)}$", and

      · $\frac{1}{mn} U_X \xrightarrow{e} 1 - \pi_\Delta = P_\Delta(X > Y)$

- $U_X = mn - U_Y$
- $U_X = W_X - \frac{1}{2}n(n+1)$
- reject H_0 if U_X is too small or too large

– null distribution of U_Y: the pmf of U_Y under H_0 can be obtained from the null distribution of W_Y by

$$P(U_Y = u) = P\Big(W_Y - \frac{m(m+1)}{2} = u\Big) = P\Big(W_Y = u + \frac{m(m+1)}{2}\Big).$$

– The tests based on U_Y and W_Y (or U_X and W_X) are actually equivalent.

Note 8 (A comparison of t-test and Mann-Whitney (M-W) test)

- Unlike t-test, the M-W test does not depend on normality assumption.
- The M-W test is insensitive to outliers, where as the t-test is sensitive.
- When the normality assumption holds, the t-test is more powerful.
- However, under normality assumption, the M-W test is nearly as powerful as the t-test. It has been shown that to attain the same power
  – the total sample size required for the t-test is approximately 0.95 times the total sample size required for the M-W test.
- The M-W test is generally preferable, especially for small sample sizes.

Theorem 12 (means and variances of U_Y and W_Y under H_0)

Consider the nonparametric model ($\Diamond$) in LNp.35. If $\Delta = 0$ ($\Leftrightarrow \pi_\Delta = 1/2$),

- $E(W_Y) = [m(m+n+1)]/2$ and $Var(W_Y) = [mn(m+n+1)]/12$
  ($\Leftrightarrow E(W_X) = [n(m+n+1)]/2$ and $Var(W_X) = [mn(m+n+1)]/12$
  since $W_X = [(m+n)(m+n+1)]/2 - W_Y$)
- $E(U_Y) = mn/2$ and $Var(U_Y) = [mn(m+n+1)]/12$
  ($\Leftrightarrow E(U_X) = mn/2$ and $Var(U_X) = [mn(m+n+1)]/12$
  since $U_X = mn - U_Y$)

Proof. It is enough to prove the case of W_Y.

- Note that
$$W_Y = R_{n+1} + \cdots + R_{m+n}.$$
  Under $H_0: \Delta = 0$, $(R_{n+1}, \cdots, R_{m+n})$ can be viewed as a without-replacement simple random sample from the population
$$\{1, \ldots, n, n+1, \ldots, m+n\}.$$
- Let $N = m + n$. Since
$$\sum_{k=1}^N k = \frac{N(N+1)}{2} \quad \text{and} \quad \sum_{k=1}^N k^2 = \frac{N(N+1)(2N+1))}{12},$$
  the population mean μ and variance σ^2 of this population distribution are
$$\mu = \frac{1}{N}\Big(\sum_{k=1}^N k\Big) = \frac{N+1}{2} \quad \text{and} \quad \sigma^2 = \frac{1}{N}\Big(\sum_{k=1}^N k^2\Big) - \mu^2 = \frac{N^2-1}{12}.$$

- Let $\overline{R}$ ($= W_Y/m$) be the average of this without-replacement sample $(R_{n+1}, \ldots, R_{m+n})$. Then (by Thms 1 & 3 in LN, Ch7, p.16-18),

$$E(\overline{R}) = \mu \quad \text{and} \quad Var(\overline{R}) = (\sigma^2/m)[(N-m)/(N-1)].$$

- The results follows from $E(W_Y) = m\,E(\overline{R})$ and $Var(W_Y) = m^2\,Var(\overline{R})$.

Theorem 13 (Asymptotic null distribution of U_Y)

Consider the nonparametric model ($\Diamond$) in LNp.35 and the null $H_0 : \Delta = 0$ ($\Leftrightarrow \pi_\Delta = 1/2$). For m, n both greater than 10, the null distribution of U_Y (or U_X) is well approximated by a normal distribution, i.e.,

$$\frac{U_Y - E(U_Y)}{\sqrt{Var(U_Y)}} \overset{D}{\approx} N(0,1) \quad \left(\text{or} \quad \frac{U_X - E(U_X)}{\sqrt{Var(U_X)}} \overset{D}{\approx} N(0,1)\right).$$

The proof is omitted, but some **notes** are given below.

- This Thm does not follow immediatedly from the ordinary CLT although

$$U_Y = \textstyle\sum_i \sum_j Z_{ij} \quad \text{and} \quad Z_{ij} \sim \text{binomial}(1, \pi_\Delta).$$

  But, Z_{ij}'s are not independent.
- Similarly, the null distribution of W_Y (or W_X) can be approximated by normal, i.e.,

$$\frac{W_Y - E(W_Y)}{\sqrt{Var(W_Y)}} \overset{D}{\approx} N(0,1) \quad \left(\text{or} \quad \frac{W_X - E(W_X)}{\sqrt{Var(W_X)}} \overset{D}{\approx} N(0,1)\right).$$

Example 5 (Asymptotic null dist. of W_Y, heat of fusion of ice, cont. Ex.4 in LNp.34)

- $n = 13$ (method A), $m = 8$ (method B), $W_B = 51$.
- Under the null,

$$\mu_{W_B} = E(W_B) = [8(8+13+1)]/2 = 88,$$
$$\sigma_{W_B} = \sqrt{Var(W_B)} = \sqrt{[(8\times 13)(8+13+1)]/12} = 13.8.$$

- Because

$$\frac{W_B - \mu_{W_B}}{\sigma_{W_B}} = \frac{51-88}{13.8} = -2.68,$$

  the approximate p-value is $P(|N(0,1)| > 2.68) = 2 \times [1 - \Phi(2.68)] = 0.0074$ ($\Rightarrow$ reject H_0 at $\alpha = 0.01 \Rightarrow$ consistent with the testing result using exact null distribution in Ex.4)

Theorem 14 (Nonparametric confidence interval for Δ)

Consider the nonparametric model ($\Diamond$) in LNp.35.

- Q: How can we test

$$H_0^* : \Delta = \Delta_0 \quad \text{vs.} \quad H_A^* : \Delta \neq \Delta_0,$$

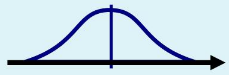
  where Δ_0 is a known constant?
  - Under H_0^*, we have (1) $X_i \sim F$, (2) $Y_j \sim G$, and (3) $G(x) = F(x - \Delta_0)$. Then,

$$X_1, \ldots, X_n, Y_1 - \Delta_0, \ldots, Y_m - \Delta_0 \sim \text{i.i.d. } F$$

  - The test of H_0^* vs. H_A^* using the data X_i's and Y_j's is equivalent to testing $H_0 : \Delta = 0$ vs. $H_A : \Delta \neq 0$ using the data X_i's and $(Y_j - \Delta_0)$'s.
  - To test $H_0^* : \Delta = \Delta_0$, can use

- $*$ the test statistic: $U_Y(\Delta_0) = \#\{X_i < Y_j - \Delta_0\} = \#\{Y_j - X_i > \Delta_0\}$,
- $*$ the acceptance region: $k(\alpha) \le U_Y(\Delta_0) \le mn - k(\alpha)$,

  where $k(\alpha)$ is the critical value determined by the significance level α (**Note**. The null distribution of $U_Y(\Delta_0)$ is symmetric about $mn/2$.)

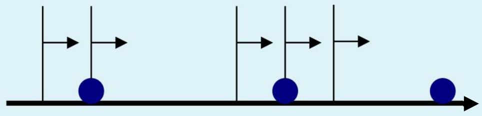


- By the duality of test and C.I., a $100(1-\alpha)\%$ confidence interval for Δ is

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

- 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 $U_Y(\Delta_0) = \#\{Y_j - X_i > \Delta_0\} = mn - k(\alpha)$,

    if $\Delta_0 < D_{(k(\alpha))}$, then $U_Y(\Delta_0) = \#\{Y_j - X_i > \Delta_0\} \ge mn - k(\alpha) + 1$,

    thus, $D_{(k(\alpha))}$ is the leftmost point of the confidence interval C,
  - $*$ if $\Delta_0 < D_{(mn-k(\alpha)+1)}$, then $U_Y(\Delta_0) = \#\{Y_j - X_i > \Delta_0\} \ge k(\alpha)$,

    if $\Delta_0 \ge D_{(mn-k(\alpha)+1)}$, then $U_Y(\Delta_0) = \#\{Y_j - X_i > \Delta_0\} \le k(\alpha) - 1$,

    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: $61 \le W_B \le 88 + (88 - 61) = 115$

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

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

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

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

  [$\overset{\text{cf.}}{\longleftrightarrow}$ the C.I. (0.015, 0.065) given in Ex.2 (LNp.12)
  - Note that the Δ here is the $-\Delta$ in Ex.2.
  - 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 ($\Diamond$) in LNp.35 or the nonparametric model ($\Box$) in LNp.27. (**Note**. (1) ($\Box$) has more models than ($\Diamond$) (2) $\pi_\Delta = P(X < Y)$ is well-defined in ($\Diamond$) and ($\Box$) (3) Δ is well-defined only in ($\Diamond$))

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

- In bootstrap, we

$$\left.\begin{array}{lll} X_1,\ldots,X_n & \sim & \text{i.i.d. from } F \\ Y_1,\ldots,Y_m & \sim & \text{i.i.d. from } G \end{array}\right\} \Leftarrow \text{independent}$$

  - 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:

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

  - $X'_1,\ldots,X'_n$ 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 $\{y_1,\ldots,y_m\}$.
- 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: $(\hat{\pi}'_{\Delta,1},\ldots,\hat{\pi}'_{\Delta,B})$
- 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 $[\hat{\pi}'_{\Delta,(B(\alpha/2))},\hat{\pi}'_{\Delta,(B(1-\alpha/2))}]$),
- the standard deviation of $(\hat{\pi}'_{\Delta,1},\ldots,\hat{\pi}'_{\Delta,B})\overset{e}{\rightarrow}$ the standard error of $\hat{\pi}_\Delta$.

❖ **Reading**: textbook, 11.2.3

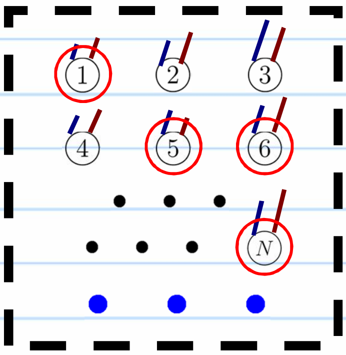
• Comparing paired samples ←→ Independent samples

• Problem formulation and statistical modeling

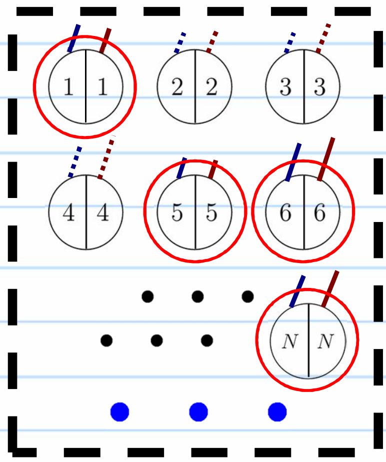
observed data (random variables)
$\{X_1, \ldots, X_n\}$
↕ ↕ ↕
$\{Y_1, \ldots, Y_n\}$

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

population



population



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

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

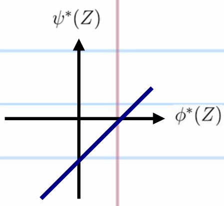
Independent samples

- Let random variables $Z_1, \ldots, Z_n$ represent the variablility of the n members sampled from the population.
- Assume $Z_1, \ldots, Z_n$ are i.i.d. from a population distribution $H(z)$.
- 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),
  - μ_X and μ_Y: the means of X and Y, respectively,
  - $\Delta = \mu_X - \mu_Y$.

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

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

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. $F(x, y)$



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

- Further assume that
  (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
  (2) Z, δ_1, δ_2 are independent
  (3) $\psi^*(Z) = \phi^*(Z) - \Delta \Rightarrow \Delta = \phi^*(Z) - \psi^*(Z)$

$$\mu_X = E[\phi^*(Z)] \qquad \mu_Y = E[\psi^*(Z)]$$

Then,

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

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

$$\begin{cases} X_i = \mu_X + \epsilon_{1i} = \phi^*(Z_i) + \delta_{1i} = \mu_X + (\phi^*(Z_i) - \mu_X) + \delta_{1i} \\ Y_j = \mu_Y + \epsilon_{2j} = \psi^*(Z_{n+j}) + \delta_{2j} = \mu_Y + (\phi^*(Z_{n+j}) - \mu_X) + \delta_{2j} \end{cases}$$

- A comparison
  - Increase sample sizes: increase information about μ_X and μ_Y (signal)
  - 2 independent $\rightarrow$ paired: suppress the variation of error (noise)

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

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

- $\epsilon = [\phi^*(Z) - \mu_X] + \delta \quad \Rightarrow \quad \underbrace{Var(\epsilon)}_{\equiv \sigma_\epsilon^2} = Var[\phi^*(Z)] + Var(\delta) \ge \underbrace{Var(\delta)}_{\equiv \sigma_\delta^2}$
- 2 independent samples $(n = m)$
  - $X_i - Y_j = (\mu_X - \mu_Y) + (\epsilon_{1i} - \epsilon_{2j})$
    $= (\mu_X - \mu_Y) + [\phi^*(Z_i) - \phi^*(Z_{n+j})] + (\delta_{1i} - \delta_{2j})$
  - $\overline{X} - \overline{Y} = (\mu_X - \mu_Y) + (\overline{\epsilon_1} - \overline{\epsilon_2})$
    $\Rightarrow \overline{X} - \overline{Y} \overset{e}{\rightarrow} \Delta$ and $Var(\overline{X} - \overline{Y}) = (\sigma_{\epsilon 1}^2 / n) + (\sigma_{\epsilon 2}^2 / n)$

- paired samples
  - $D_i \equiv X_i - Y_i = (\mu_X - \mu_Y) + (\delta_{1i} - \delta_{2i})$
  - $\overline{D} = \overline{X} - \overline{Y} = (\mu_X - \mu_Y) + (\overline{\delta_1} - \overline{\delta_2})$

$$\Rightarrow \quad \overline{X} - \overline{Y} \xrightarrow{e} \Delta = \mu_X - \mu_Y \quad \text{and}$$

$$Var(\overline{X} - \overline{Y}) = (\sigma_{\delta_1}^2 / n) + (\sigma_{\delta_2}^2 / n) \longleftrightarrow$$

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

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

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

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

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

  $E(Y) = E[\psi^*(Z) + \delta_2] = E[\psi^*(Z)] = \mu_Y$
- $Var(X) = Var[\phi^*(Z) + \delta_1] = Var[\phi^*(Z)] + Var(\delta_1) \equiv \sigma_X^2 \; (= \sigma_{\epsilon_1}^2)$

  $Var(Y) = Var[\psi^*(Z) + \delta_2] = Var[\psi^*(Z)] + Var(\delta_2) \equiv \sigma_Y^2 \; (= \sigma_{\epsilon_2}^2)$
- 2 independent samples $(n = m)$
  - $Cov(X_i, Y_j) = Cov[\phi^*(Z_i) + \delta_{1i},\, \psi^*(Z_{n+j}) + \delta_{2j}] = 0$
  - $E(\overline{X} - \overline{Y}) = \mu_X - \mu_Y = \Delta$
  - $Var(\overline{X} - \overline{Y}) = (\sigma_X^2 + \sigma_Y^2)/n$

- paired samples
  - $Cov(X_i, Y_i) = Cov[\phi^*(Z_i) + \delta_{1i}, \psi^*(Z_i) + \delta_{2i}]$

    $= Cov[\phi^*(Z_i), \psi^*(Z_i)] \equiv \sigma_{XY}$
    * **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 $\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$
    * Notice that $\rho_{XY} \neq$

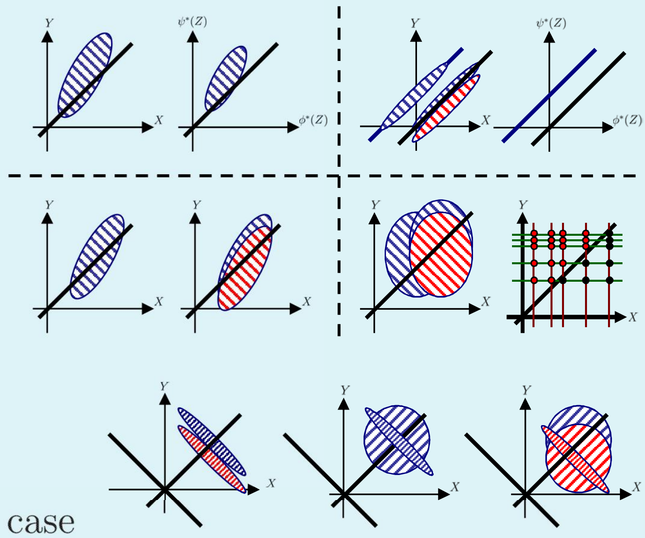
      $Cor[\phi^*(Z), \psi^*(Z)] = \frac{\sigma_{XY}}{\sigma_{\phi^*(Z)} \sigma_{\psi^*(Z)}}$
  - Let $D_i = X_i - Y_i$, $i = 1, \ldots, n$. Then,
    * $D_1, \ldots, D_n$ are i.i.d.
    * $E(D_i) = \mu_X - \mu_Y$
    * $Var(D_i) = Var(X_i) + Var(Y_i) - 2\,Cov(X_i, Y_i) = \sigma_X^2 + \sigma_Y^2 - 2\,\sigma_{XY}$
  - Since $\overline{D} = \overline{X} - \overline{Y} \quad (\xrightarrow{e} \Delta)$
    * $E(\overline{D}) = \mu_X - \mu_Y = \Delta$
    * $Var(\overline{D}) = Var(\overline{X} - \overline{Y}) = (\sigma_X^2 + \sigma_Y^2 - 2\,\sigma_{XY})/n$

      $= (\sigma_X^2 + \sigma_Y^2 - 2\,\rho_{XY}\,\sigma_X\,\sigma_Y)/n$

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

$$Cov\big[\phi^*(Z), \psi^*(Z)\big] = Cov\big[\phi^*(Z), \phi^*(Z) - \Delta\big] = Var\big[\phi^*(Z)\big] > 0.$$

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



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

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

in the unpaired case. The relative efficiency is $\sigma_{\overline{D}}^2/\sigma_{\overline{X}-\overline{Y}}^2 = 1 - \rho_{XY}$.
  - 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.
- From now on, the analyses of paired data are based on

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

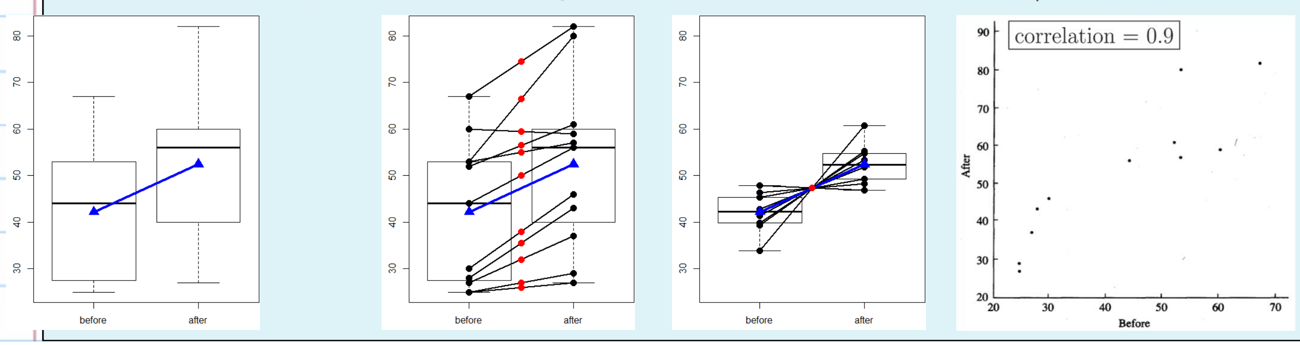
- Statistical modeling for D_i's: $D_1, \ldots, D_n \sim$ i.i.d. $F \Leftarrow$ one-sample model

Example 7 (Effect of cigarette smoking on platelet aggregation, Levine, 1973)

- Blood samples were drawn from 11 individuals before and after they smoked a cigarette to measure the extent to which the blood platlets aggregated.
- data: maximum percentage of all platelets that aggregated after being exposed to a stimulus.

	1	2	3	4	5	6	7	8	9	10	11
before (Y)	25	25	27	44	30	67	53	53	52	60	28
after (X)	27	29	37	56	46	82	57	80	61	59	43
difference (D)	2	4	10	12	16	15	4	27	9	−1	15

- Q: Do the differences D_i's indicate a clear pattern of $\Delta = \mu_X - \mu_Y \neq 0$?
- The two-sample (unpaired) t-test for the before and after data gives a p-value $=$ 0.1721 $\Rightarrow$ the null $H_0 : \mu_X = \mu_Y$ is not rejected under $\alpha = 0.1$ ($s_p^2 = 289.34$).
  - Q: Why did the 2-sample t-test not reject H_0 when the differences showed such a clear pattern of $\mu_X > \mu_Y$?
  - Note that in 2-sample t-test, the test statistic is $|T| = \dfrac{|\overline{X} - \overline{Y}|}{s_p\sqrt{\frac{1}{n} + \frac{1}{m}}}$, where s_p^2 estimates σ_ϵ^2, rather than σ_δ^2.



- Figure 11.7 (textbook, p.447) plots after-values vs. before-values. They are positively correlated with a sample correlation coefficient 0.9. Pairing was a natural and effective experimental design in this case: relative efficiency = 0.1.

- Methods based on normality assumptions
- **Recall**. $D_i = X_i - Y_i$, $i = 1, \ldots, n$, and $D_1, \ldots D_n \sim$ i.i.d. F.
- Assume that F is Normal.
- Thus, the statistical model for D_i's is

$$D_1, \ldots D_n \sim \text{i.i.d. } N(\mu_D, \sigma_D^2), \qquad (\triangle)$$

where $\mu_D = \mu_X - \mu_Y$.
  - This model contains two parameters: μ_D ($\in \mathbb{R}$) and σ_D^2 (> 0).
  - Under this model, we can only examine whether there exists "difference" between the means of the two paired samples, i.e.,

$$\mu_D = \mu_X - \mu_Y = 0 \Rightarrow \text{no difference or no effect}$$

Theorem 18 (test and confidence interval for μ_D, 1-sample normal model, paired data)

Consider the model ($\triangle$).
- **Recall** (Review 1 in LNp.6-7).
  - $\overline{D} = \frac{1}{n}\sum_{i=1}^n D_i \xrightarrow{e} \mu_D$, and $\overline{D} \sim N(\mu_D, \sigma_D^2/n) \Rightarrow \sqrt{n}(\overline{D} - \mu_D)/\sigma_D \sim N(0,1)$
  - $s_D^2 = \frac{\sum_{i=1}^n (D_i - \overline{D})^2}{n-1} \xrightarrow{e} \sigma_D^2$, and $(n-1)s_D^2 \sim \sigma_D^2 \chi_{n-1}^2 \Rightarrow (n-1)s_D^2/\sigma_D^2 \sim \chi_{n-1}^2$
  - $\overline{D}$ and s_D^2 are independent

- Pivotal quantity for μ_D
  - σ_D known: $$Q_{Z,\mu_D} \equiv \frac{\overline{D} - \mu_D}{\sigma_{\overline{D}}} = \frac{\overline{D} - \mu_D}{\sigma_D/\sqrt{n}} = \frac{\sqrt{n}(\overline{D} - \mu_D)}{\sigma_D} \sim N(0,1)$$
  - σ_D unknown: $$Q_{T,\mu_D} \equiv \frac{\overline{D} - \mu_D}{s_{\overline{D}}} = \frac{\overline{D} - \mu_D}{s_D/\sqrt{n}} = \frac{\sqrt{n}(\overline{D} - \mu_D)/\sigma_D}{\sqrt{\frac{(n-1)s_D^2/\sigma_D^2}{n-1}}} \sim t_{n-1}$$
- Test the null and alternative hypotheses at significance level α:

$$H_0 : \mu_D = \mu_{D,0} \quad \text{vs.} \quad H_A : \mu_D \neq \mu_{D,0}$$

where $\mu_{D,0}$ is a known constant.

(**Note**. if $\mu_{D,0} = 0$, this is equivalent to $H_0 : \mu_X = \mu_Y$ vs. $H_A : \mu_X \neq \mu_Y$.)
  - σ_D known: reject H_0 if $$\left| Z \equiv \frac{\overline{D} - \mu_{D,0}}{\sigma_{\overline{D}}} \right| > z(\alpha/2) \quad \Leftrightarrow \quad \left|\overline{D} - \mu_{D,0}\right| > z(\alpha/2)\,\sigma_{\overline{D}} = z(\alpha/2)\frac{\sigma_D}{\sqrt{n}}$$
  - σ_D unknown: reject H_0 if $$\left| T \equiv \frac{\overline{D} - \mu_{D,0}}{s_{\overline{D}}} \right| > t_{n-1}(\alpha/2) \Leftrightarrow \left|\overline{D} - \mu_{D,0}\right| > t_{n-1}(\alpha/2)\, s_{\overline{D}} = t_{n-1}(\alpha/2)\frac{s_D}{\sqrt{n}}$$
- A $100(1-\alpha)\%$ confidence interval for μ_D is
  - σ_D known: $\overline{D} \pm z(\alpha/2) \times \sigma_{\overline{D}} \Leftrightarrow \overline{D} \pm z(\alpha/2) \times (\sigma_D/\sqrt{n})$
  - σ_D unknown: $\overline{D} \pm t_{n-1}(\alpha/2) \times s_{\overline{D}} \Leftrightarrow \overline{D} \pm t_{n-1}(\alpha/2) \times (s_D/\sqrt{n})$

Example 8 (Effect of smoking, t-test for paired data, cont. Ex.7 In LNp.52)

- $n = 11$, $D_i = \text{after}_i - \text{before}_i$
- $\overline{D} = 10.27$, $s_{\overline{D}} = 2.405$ $\left(\Rightarrow s_D^2 = 11 \times 2.405^2 = 63.62\right.$

  $\left.\Rightarrow 63.62/2 = 31.81 \stackrel{e}{\to} \sigma_\delta^2 \stackrel{\text{cf.}}{\longleftrightarrow} s_p^2 = 289.34 \stackrel{e}{\to} \sigma_\epsilon^2 \text{ in Ex.7}\right)$

  If $\sigma_{\delta_1}^2 = \sigma_{\delta_2}^2 = \sigma_\delta^2$, then $Var(D_i) = 2\,\sigma_\delta^2$

- A 90% confidence interval for μ_D is

  $$\overline{D} \pm t_{10}(0.05)\, s_{\overline{D}} = 10.27 \pm 1.812 \times 2.40 = (5.9, 14.6),$$

  which does not contain zero $\left(\stackrel{\text{cf.}}{\longleftrightarrow} H_0 \text{ not rejected in Ex.7 using 2-sample } t\text{-test}\right)$
- The (one-sample) t-statistic is $T = (10.27 - 0)/2.40 = 4.28 > t_{10}(0.005) = 3.169$. The p-value of a two-sided test is less than 0.01. There is little doubt that smoking increases platelet aggregation.

Note 9 (Some notes about one-sample t-test when normality assumption does not hold)

- Consider the model: $D_1, \ldots D_n \sim$ i.i.d. F,

  where F can be *any* continuous distributions with *finite* variance.
  - By CLT and LLN, when $n \to \infty$ (sample size is large),

    $\overline{D} \stackrel{D}{\approx} N(\mu_D, \sigma_D^2/n)$ and $s_D^2 \stackrel{P}{\longrightarrow} \sigma_D^2$.
  - Thus, by Slutsky's Thm, $Q_{T,\mu_D} = \dfrac{(\overline{D} - \mu_D)/(\sigma_D/\sqrt{n})}{\sqrt{s_D^2/\sigma_D^2}} \stackrel{D}{\approx} N(0,1)$

    and t_{n-1} tends to $N(0,1)$ as $n \to \infty$.
- **Q**: What if the sample size n is small or population variance $= \infty$?

• A nonparametric method --- the signed rank test

- Let Ω be the collection of *all* continuous distributions $\Rightarrow \dim(\Omega)=\infty$
- Consider the nonparametric statistical model: $D_1, \ldots D_n \sim$ i.i.d. F, $\quad(\nabla)$

  where $F \in \Omega$.
- Let $\Omega_0 = \{F \mid F \in \Omega \text{ and } F \text{ is symmetric about } 0\}$
  - $\Omega_0 \subset \Omega$ and $\dim(\Omega_0)=\infty$
  - If $F \in \Omega_0$, then the median of F is 0. But, F with median zero is not necessary a distribution being symmetric about 0.
- Under the model (∇), we want to test the null and alternative hypotheses:

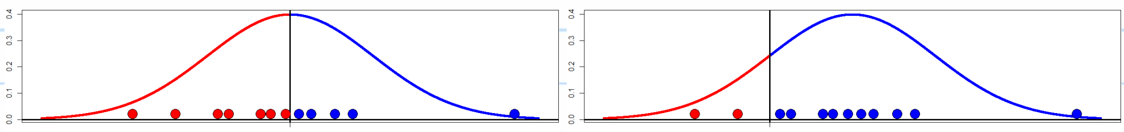
  $$H_0 : F \in \Omega_0 \quad \text{vs.} \quad H_A : F \in \Omega \setminus \Omega_0$$

  Q: Why add the "symmetric" condition in the null?

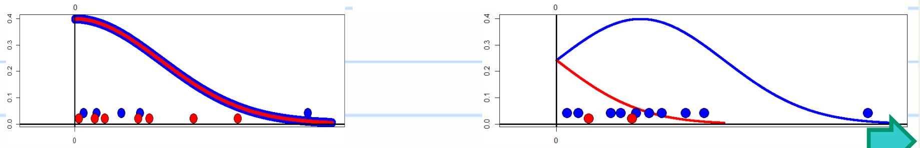
Question 8.

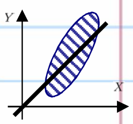
How to use ranks to examine "symmetric about 0"? What data are "more extreme," i.e., cast more doubts on H_0?

Intuition.



ranks:





- A brief comparison to (2-sample) rank sum test (assume no $D_i = 0$)
  - similarity: if in the paired case,
    - the data $\{-D_i \,|\, D_i < 0\}$ is treated as the 1st sample
    - the data $\{D_i \,|\, D_i > 0\}$ is treated as the 2nd sample
    - then, the calculation for the paired case is equivalent to the rank-sum statistic in the unpaired case
  - difference
    - In 2-sample unpaired cases, the sample sizes m, n are fixed numbers.
    - In the paired case, the sizes

$$N_- = \#\{D_i < 0\} \quad \text{and} \quad N_+ = \#\{D_i > 0\}$$

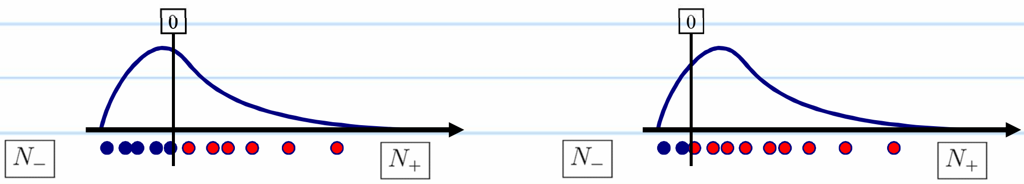
(**Note**. $N_- + N_+ = n$) are random variables.

    - Under H_0,
      - $I_{[D_1>0]}, \ldots, I_{[D_n>0]} \sim$ i.i.d. Bernoulli(1/2),
      - $N_+ = \sum_{i=1}^{n} I_{[D_i>0]} \sim \text{bin}(n, 1/2)$ and $N_- = n - N_+ \sim \text{bin}(n, 1/2)$
    - When conditioned on N_+ (or N_-), the null distribution of the test statistic in the paired case is identical to the null distribution of rank-sum statistic in the unpaired case.
- Alternative test: sign test (TBp.461, problem 12)
  - Consider the model (∇) in LNp.56.

  - Let $\Omega_0^* = \{F \,|\, F \in \Omega \text{ and } F \text{ has median } 0\}$
    - $\Omega_0 \subset \Omega_0^* \subset \Omega$ and $\dim(\Omega_0^*) = \infty$
  - Under the model (∇), test the null and alternative hypotheses:

$$H_0^* : F \in \Omega_0^* \quad \text{vs.} \quad H_A^* : F \in \Omega \backslash \Omega_0^*$$

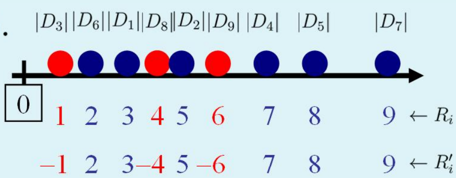
  - **Intuition**.



  - Reject H_0^* if N_+ (or N_-) is small (close to 0) or large (close to n)
  - Null distribution of N_+ (or N_-): bin(n, 1/2)

Theorem 19 (Wilcoxon signed rank test)

Consider the nonparametric model (∇) in LNp.56.



- test ststistic W_+ (or $W_- = \frac{n(n+1)}{2} - W_+$)

(1) Let R_i = rank of $|D_i|$, $i = 1, \ldots, n$.

(2) Restore the signs of D_i's to the ranks R_i''s, i.e., let $R_i' = \text{sign}(D_i) \times R_i$.

(3) $W_+ = \sum_{i=1}^{n} I_{[D_i>0]} R_i'$, i.e., sum of the ranks R_i''s that have positive signs.

- Q: What values of W_+ are more extreme? If there is no difference between the two paired conditions, we expect

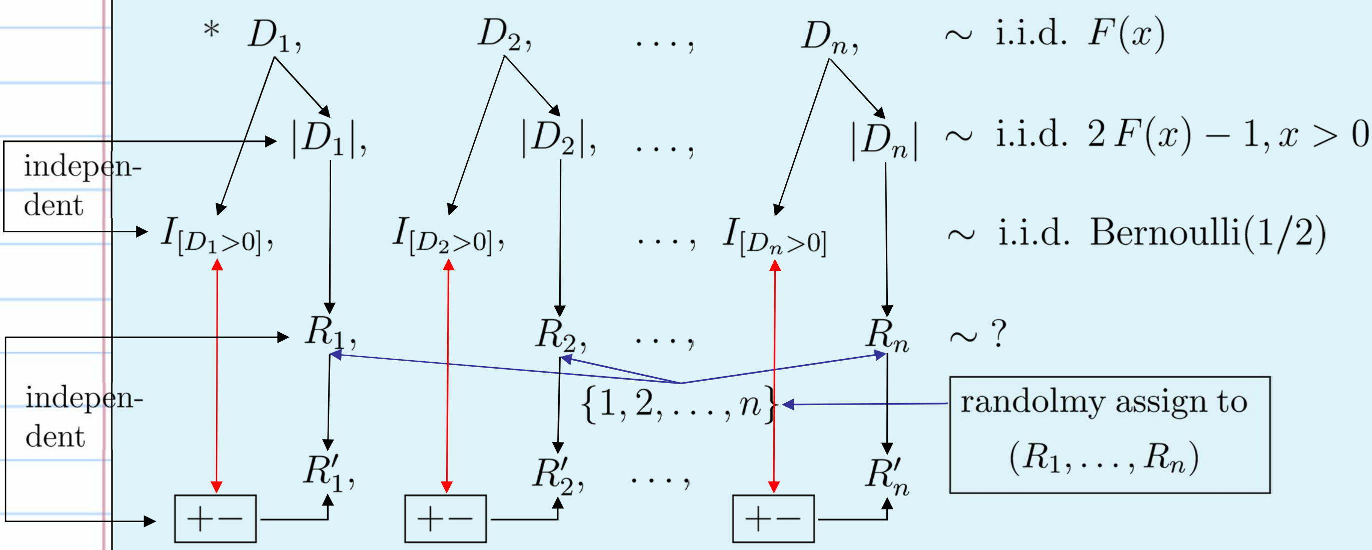
– about *half* the D_i's to be positive and *half* negative (median=0?)

– positive R_i''s and negative R_i''s *similarly* distributed (symmetric?)

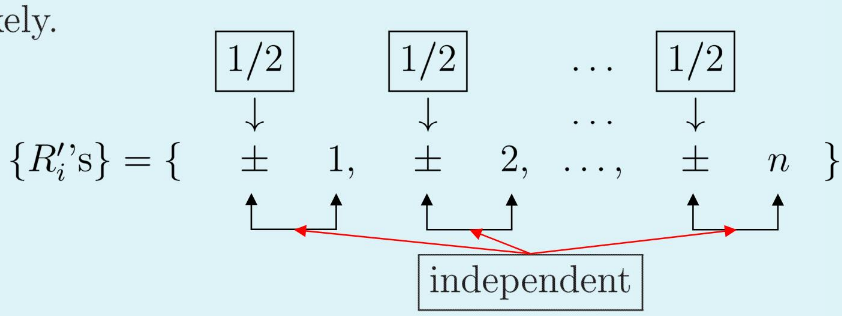
and W_+ will not be too small or too lagre

$\Rightarrow$ data with larger or smaller W_+ are more extreme $\Rightarrow$ tend to reject H_0

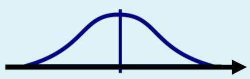
- Null distribution of W_+
  – $W_+ \in \{0, 1, 2, \ldots, \frac{n(n+1)}{2}\}$
  – Under the null H_0 (F is symmetric about 0)
    * $D_i \Leftrightarrow (I_{[D_i>0]}, |D_i|)$ and D_i has the same distribution as $-D_i$
    *



* Any particular assignment of $\{-, +\}$ signs to the integers $1, \ldots, n$ (the ranks) is equally likely.



* There are 2^n such assignments and for each we can calculate $W_+ \Rightarrow$ obtain 2^n values (not all distinct) of W_+, each with probability $1/2^n$.
* The probability of each distinct value of W_+ may thus be calculated, giving the desired null distribution.



- (Two-sided) rejection region

(exercise) – The null distribution of W_+ is symmetric around $E(W_+)$

– Reject H_0 when $\min(W_+, W_-)$ is small, i.e., $\min(W_+, W_-) \le w$

– Table 9 of Appendix B in textbook (TBp.A24) gives critical values w

- Ties
  – Tie between (X_i, Y_i): If some of the differences D_i's are zero, the most common technique is to discard those observations.

– Tie between $|D_i|$'s: If there are ties, each $|D_i|$ is assigned the average value of the ranks for which it is tied.
– If there are a large number of ties, modifications must be made. See Hollander and Wolfe (1973) or Lehmann (1975).

Example 9 (Smoking effect, signed-rank test for paired data, cont. Ex.7 In LNp.52)

- $n = 11$, $W_- = 1$ and $W_+ = [11(11+1)]/2 - W_- = 65 \Rightarrow \min(W_-, W_+) = 1$
- From Table 9 of Appendix B (TBp.A24), the critical value for two-sided test with significant level $\alpha = 0.01$ is 5.
- Since $\min(W_-, W_+) < 5$, reject H_0 at $\alpha = 0.01$ (consistent with the test result in Ex.8, LNp.55).

Note 10 (A comparison of one-sample t-test and signed rank test for paired data)

- Unlike (one-sample) t-test, the signed-rank test does not depend on normality assumption.
- The signed-rank test is insensitive to outliers, whereas the t-test is sensitive.
- When the normality assumption holds, the t-test is more powerful.
- However, it has been shown that even when normality assumption holds, the signed-rank test is nearly as powerful as the t-test (relatively efficiency of signed-rank test statistic to (one-sample) t-test statistic ≈ 0.95).
- The signed-rank test is generally preferable, especially for *small* sample sizes.

Theorem 20 (means and variances of W_+ under H_0)

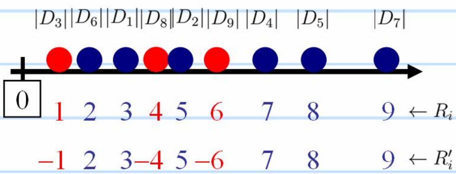
- Consider the nonparametric model (∇) in LNp.56.
- Under the null hypothesis H_0: F is symmetric about 0,

$$E(W_+) = \frac{n(n+1)}{4} \quad \text{and} \quad Var(W_+) = \frac{n(n+1)(2n+1)}{24}.$$

$$\Big(\Leftrightarrow E(W_-) = [n(n+1)]/4 \quad \text{and} \quad Var(W_-) = [n(n+1)(2n+1)]/24$$
$$\text{since } W_- = [n(n+1)]/2 - W_+ \Big)$$

Proof.

- For $k = 1, \ldots, n$, let $I_k = \begin{cases} 1, & \text{if the } k\text{th largest } |D_i| \text{ has } D_i > 0, \\ 0, & \text{otherwise.} \end{cases}$
- Under H_0,
  – $I_1, \ldots, I_n \sim$ i.i.d. Bernoulli(1/2),
  – $E(I_k) = 1/2$ and $Var(I_k) = 1/4$.



- Write
$$W_+ = \sum_{i=1}^{n} I_{[D_i>0]}\, R_i' = \sum_{k=1}^{n} k\, I_k.$$
- Thus,
$$E(W_+) = \sum_{k=1}^{n} k\, E(I_k) = \frac{1}{2}\Big(\sum_{k=1}^{n} k\Big) = \frac{n(n+1)}{4}$$
$$Var(W_+) = \sum_{k=1}^{n} k^2\, Var(I_k) = \frac{1}{4}\Big(\sum_{k=1}^{n} k^2\Big) = \frac{n(n+1)(2n+1)}{24}$$

Theorem 21 (Asymptotic null distribution of W_+)

- Consider the nonparametric model (∇) in LNp.56.
- Under the null $H_0 : F$ is symmetric about 0,
  if the sample size n is greater than 20, the null distribution of W_+ is well approximated by a normal distribution, i.e.,

$$\frac{W_+ - E(W_+)}{\sqrt{Var(W_+)}} \overset{D}{\approx} N(0,1) \quad \left(\text{or} \quad \frac{W_- - E(W_-)}{\sqrt{Var(W_-)}} \overset{D}{\approx} N(0,1) \right).$$

Hint for Proof. Use the expression $W_+ = \sum_{k=1}^{n} k\, I_k$ to find the moment generating function of W_+, and show it converges (after standardization) to the moment generating function of $N(0,1)$, which is $e^{t^2/2}$.

❖ **Reading**: textbook, 11.3