Lecture 6: Nonparametric methods

BTBI30081

統計應用方法Applied Methods in Statistics

2025/3/26

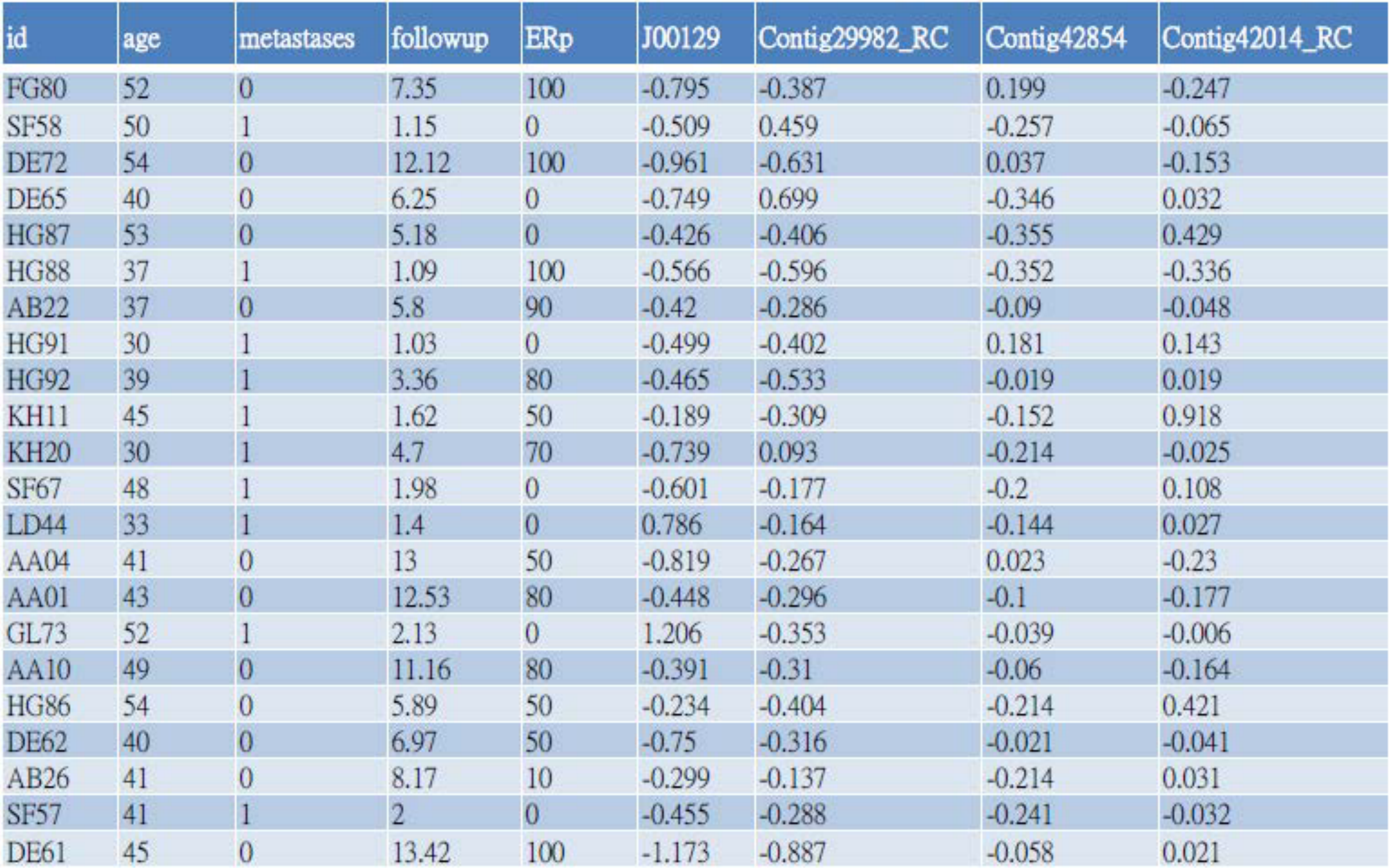
Example: Gene expression microarray data

- Data from a study using gene expression profiling to predict breast cancer outcomes (http://www.nature.com/nature/journal/v415/n6871/full/415530a.html)
- 78 breast cancer: 44 remained disease-free for an interval of at least five years after their initial diagnosis (good prognosis group), while 34 patients had developed distant metastases within five years (poor prognosis group)

samplexprs.csv

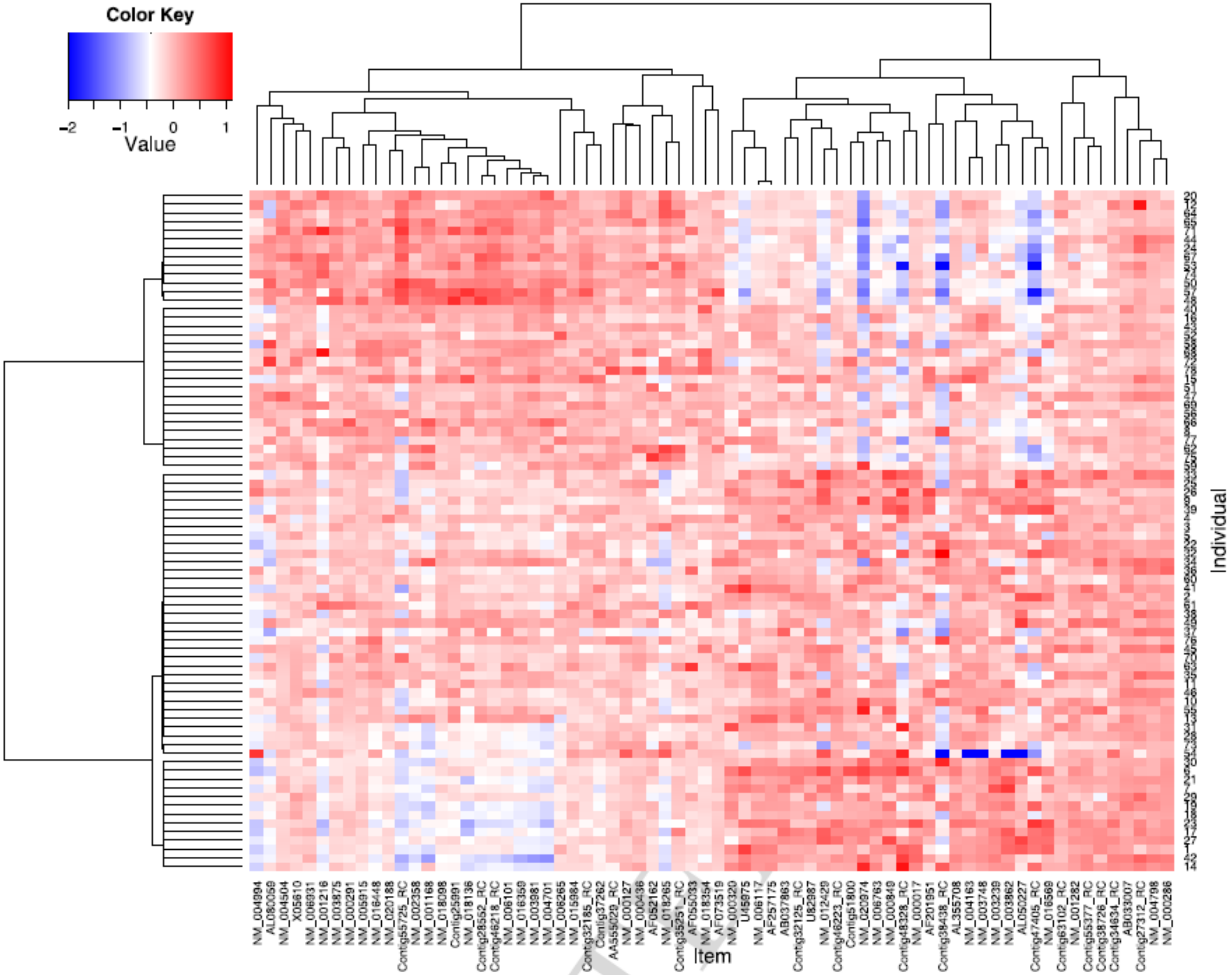
Variable	Description
id	An unique identification number
age	Age at diagnosis of breast cancer (year)
metastases	Developing distant metastases: 0=no (good prognosis group), 1=yes (poor prognosis group)
followup	Follow-up time (year)
ERp	ER-α expression level
J00129	log$_{10}$ gene expression intensity ratios
Contig29982_RC	log$_{10}$ gene expression intensity ratios

- RMD_example 06.1



id	age	metastases	followup	ERp	J00129	Contig29982_RC	Contig42854	Contig42014_RC
FG80	52	0	7.35	100	-0.795	-0.387	0.199	-0.247
SF58	50	1	1.15	0	-0.509	0.459	-0.257	-0.065
DE72	54	0	12.12	100	-0.961	-0.631	0.037	-0.153
DE65	40	0	6.25	0	-0.749	0.699	-0.346	0.032
HG87	53	0	5.18	0	-0.426	-0.406	-0.355	0.429
HG88	37	1	1.09	100	-0.566	-0.596	-0.352	-0.336
AB22	37	0	5.8	90	-0.42	-0.286	-0.09	-0.048
HG91	30	1	1.03	0	-0.499	-0.402	0.181	0.143
HG92	39	1	3.36	80	-0.465	-0.533	-0.019	0.019
KH11	45	1	1.62	50	-0.189	-0.309	-0.152	0.918
KH20	30	1	4.7	70	-0.739	0.093	-0.214	-0.025
SF67	48	1	1.98	0	-0.601	-0.177	-0.2	0.108
LD44	33	1	1.4	0	0.786	-0.164	-0.144	0.027
AA04	41	0	13	50	-0.819	-0.267	0.023	-0.23
AA01	43	0	12.53	80	-0.448	-0.296	-0.1	-0.177
GL73	52	1	2.13	0	1.206	-0.353	-0.039	-0.006
AA10	49	0	11.16	80	-0.391	-0.31	-0.06	-0.164
HG86	54	0	5.89	50	-0.234	-0.404	-0.214	0.421
DE62	40	0	6.97	50	-0.75	-0.316	-0.021	-0.041
AB26	41	0	8.17	10	-0.299	-0.137	-0.214	0.031
SF57	41	1	2	0	-0.455	-0.288	-0.241	-0.032
DE61	45	0	13.42	100	-1.173	-0.887	-0.058	0.021

Example: Gene expression microarray data (samplexprs.csv)



Heatmap for gene expression microarray data (samplexprs.csv)

Nonparametric statistical methods

- A family of probability distributions that can be described by a few parameters is a parametric family.

ex、常態分布：平均數、變異數

- Parametric statistical procedures can be used when the sampling distribution is from a parametric family (e.g., normal or approximately normal).
- A family of probability distributions is nonparametric if it cannot be easily described by a few parameters.

Use of nonparametric methods

- Nonparametric statistical procedures can be used when:
  - sample size(s) are small 小樣本下常態分佈假設較難驗證
  - assumptions of parametric testing procedures cannot be met 常態性、變異數同質性、連續性假設
- Pro and con of nonparametric methods:
  - pro: insensitive to weird observations (outliers)
  - con: only looking at sign and rank → lose information, less powerful. ①符號 ②排名

Nonparametric versus Parametric

Type of test	Nonparametric	Parametric
One sample test	Sign test	One-sample t-test
Paired data	Wilcoxon signed-rank test	Paired t-test
Two sample test	Wilcoxon rank-sum test (Mann-Whitney test)	Two-sample t-test
More than 2 samples	Kruskal-Wallis test	ANOVA
Correlation	Spearman rank correlation	Pearson correlation

One sample test

- E.g., New ointment for reducing sun burn, measure degree of protection

Subject	1	2	3	4	5	6	7	8	9	10
Redless (X)	37	39	31	39	38	47	35	30	25	40

- $X \sim N(\mu, \sigma^2)$

  The mean protection of old ointment $\mu_0 = 34$

  $$\begin{cases} H_0\colon\ \mu = \mu_0 \\ H_a\colon\ \mu > \mu_0 \end{cases}$$

- Parametric test: one-sample t-test

  $t = \frac{\bar{X} - \mu_0}{(s_X/\sqrt{N})} \xrightarrow[H_0]{} t(N-1)$, where $s_X^2 = \frac{1}{N-1}\sum_{i=1}^{N}(X_i - \bar{X})^2$

  E.g., Ointment: $t = \frac{36.1-34}{6.17/\sqrt{10}} = 1.08$, p-value = 0.16

  (RMD_example 06.2)

Sign test

- Nonparametric test: sign test

  Test statistic $S = \#\text{ of } X > 34$

  If S is large ⇒ evidence that new ointment is better than old one

假設檢定

H_0 新舊藥膏無顯著差異，中位數=34

H_1 更好 中位數>34

- Model:

  X's independent, from some continuous distribution

  Under H_0, $\Pr(X > 34) = p_+$, $\Pr(X \leq 34) = p_-$, $p_+ = p_- = 1/2$

  $S = 7$ (# of $X > 34$) 因為每個觀測值大於或小於34都是0.5

  Under H_0, $S \sim \text{Binomial}(N, 1/2)$, $N = 10$

  P-value = $\Pr(S \geq 7 | H_0) = 0.17$

  (RMD_example 06.3)

① 比較每筆資料與34的大小

② 檢定統計量 $S = \#(X > 34) = 7$

③ 在虛無假設下 H_0 $S \sim B(10, 0.5)$

④ 計算 p-value

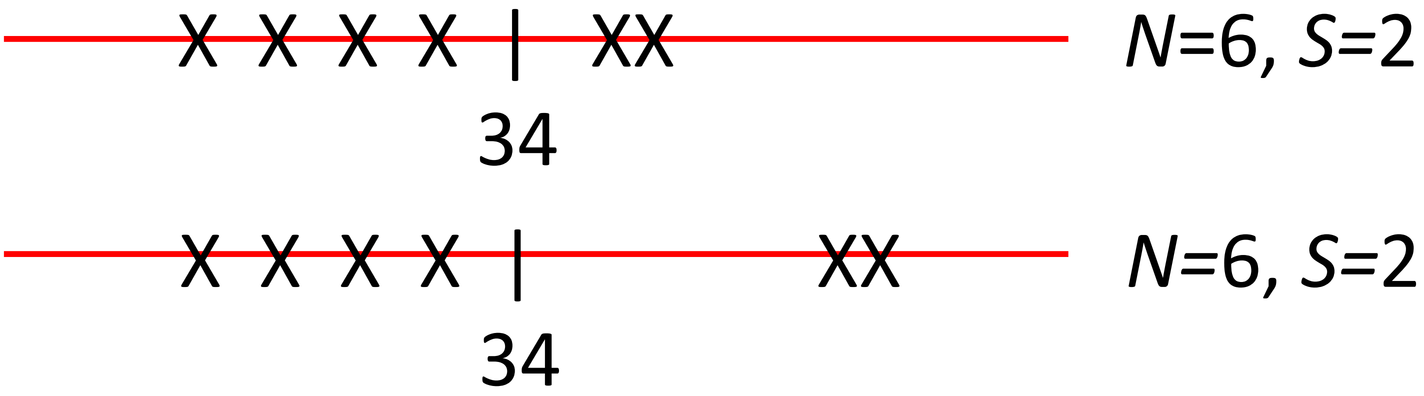
Notes:

1. Only need +, - data
2. Did not assume normality of X
3. With large N, we can use Normal approximation to Binomial

   $N \gg,\ S\sim\mathrm{Normal}(\mu = N/2\,, \sigma^2 = N/4)$ under H_0
4. For small N, get exact p-value

5. Not sensitive to wild observations

E.g.,



We get the same p-value by sign test, but not by t-test t-test 對極端值非常敏感

6. Limitation of sign test: only use signs, ignore magnitude of X

7. Sign test is in fact for H_0: median of X distribution $= m_0$, however usually median $\approx$ mean.

sign test 實質上是中位數的檢定

Sign test—for gene expression data

- Use for one-sample test
- Test whether or not the population mean of $\log_{10}$ gene expression intensity ratios on gene J00129 (μ_{J00129}) is equal to -0.5.
- $H_0\colon \mu_{J00129} = -0.5 \quad H_a\colon \mu_{J00129} \neq -0.5$
- It depends on the sign of the differences between observations and given number; not on their actual magnitude.

  (RMD_example 06.4)

Paired data

- E.g., Twins, 1st born got Tx 1, 2nd born got Tx 2

1st born	659	984	397	574	447	479	676	761	647	402
2nd born	452	507	460	787	351	277	234	516	577	338
D (diff.)	207	397	-63	-213	96	202	442	245	70	64

① 幾組資料

② 是否來自相同受試單位

③ 差值是否符合常態分布

- $D \sim N(\mu_D, \sigma_D^2)$

$$\begin{cases} H_0: & \begin{array}{c} \text{No treatment effect} \\ \text{Distribution of } D'\text{s symmetric about 0} \\ \mu_D = 0 \end{array} \\ H_a: & \begin{array}{c} \text{Distribution of } D'\text{s is skewed to left} \\ \mu_D > 0 \end{array} \end{cases}$$

無差異

Tx1效果較好

- Parametric test: paired t-test

$t = \dfrac{\overline{D}}{(s_D/\sqrt{N})} \xrightarrow[H_0]{} t(N-1)$, where $s_D^2 = \frac{1}{N-1}\sum_{i=1}^{N}(D_i - \overline{D})^2$ 可計算出來的

E.g., Twins: $t = 2.29$, p-value = 0.024

(RMD_example 06.5)

Wilcoxon signed-rank test

- Nonparametric test 1: sign test for D: $\begin{cases} H_0\colon\ \mu_D = 0 \\ H_a\colon\ \mu_D > 0 \end{cases}$

  E.g., Twins: $S = 8$, p-value = 0.055 (one-sided) (RMD_example 06.6)

  Limitation: only use signs, ignore magnitude of D

- Nonparametric test 2: Wilcoxon signed-rank test → account for the magnitude of D

  Test statistic

  1. Find absolute value of D: $|D|$
  2. Rank in increasing order of $|D|$'s
  3. Test statistic T^+ = sum of the ranks of $|D|$ for which original D was positive

E.g., Twins

D	207	397	-63	-213	96	202	442	245	70	64
Rank of $\lvert D\rvert$	6	9	1	7	4	5	10	8	3	2
	+	+	-	-	+	+	+	+	+	+

$T^+ = 6 + 9 + 4 + 5 + 10 + 8 + 3 + 2 = 47$

$T^- = 1 + 7 = 8$

- Reject H_0 if T^+ is large

- Derive the exact distribution of T^+ under H_0
  - E.g., for $N = 3$

    Under H_0, $\Pr(D+) = \Pr(D-) = 1/2$

Rank			T^+	Prob.
1	2	3		
+	+	+	6	$(1/2)^3$
+	-	+	4	$(1/2)^3$
-	+	+	5	$(1/2)^3$
-	-	+	3	$(1/2)^3$
-	+	-	2	$(1/2)^3$
+	-	-	1	$(1/2)^3$
-	-	-	0	$(1/2)^3$
+	+	-	3	$(1/2)^3$

Distribution of T^+

t	0	1	2	3	4	5	6
$\Pr(T^+ = t)$	$1/8$	$1/8$	$1/8$	$2/8$	$1/8$	$1/8$	$1/8$
$\Pr(T^+ \geq t)$	$8/8$	$7/8$	$6/8$	$5/8$	$3/8$	$2/8$	$1/8$

If $T^+ = 5, \Pr_{H_0}(T^+ \geq 5) = 2/8 = 0.25$ (one-sided p-value)

- E.g., Twins: $N = 10, T^+ = 47,$ p-value = $\Pr_{H_0}(T^+ \geq 47) = 0.024$ (one-sided)

(RMD_example 06.7)

- Large sample approximation ($N \geq 15$)

  Large sample approximation of T^+ under H_0:
  distribution of D's symmetric about 0

  $$\frac{T^+ - \mathrm{E}_{H_0}(T^+)}{\sqrt{\mathrm{Var}_{H_0}(T^+)}} \sim N(0, 1)$$

  for $N \geq 15$

Wilcoxon signed-rank test—for gene expression data

- Use for paired data
- Test whether or not the difference of the $\log_{10}$ expression intensity ratio on gene J00129 and the one on Contig29982_RC *from the same individual* is equal to 0.
- H_0: $\mu_{J00129} = \mu_{Contig29982\_RC}$
  H_a: $\mu_{J00129} \neq \mu_{Contig29982\_RC}$
- Replaces the observed paired differences with ranks (RMD_example 06.8)

Two sample test

- E.g., Aspirin concentration (mg%)

Aspirin X $m = 5$	15	26	13	28	17	
Aspirin Y $n = 6$	12	20	10	21	18	22

- $X \sim N(\mu_X, \sigma_X^2), Y \sim N(\mu_Y, \sigma_Y^2)$

$$\begin{cases} H_0\colon \mu_X = \mu_Y \\ H_a\colon \mu_X > \mu_Y \end{cases}$$

- Parametric test: two-sample t-test

  Assume normality on X's and Y's; $\sigma_X^2 = \sigma_Y^2$

  $$t = \frac{\bar{X} - \bar{Y}}{s_P\sqrt{\frac{1}{m} + \frac{1}{n}}} \xrightarrow[H_0]{} t(m+n-2)$$

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

  E.g., Aspirin: $t = 0.74$, p-value = 0.24

  (RMD_example 06.9)

Wilcoxon rank-sum test

- Nonparametric test: Wilcoxon rank-sum test (or Mann-Whitney test)

  不需要正態性或變異數假設
  相等

  No normality, variance assumptions

  Test statistic

  1. Pool X's and Y's together 把兩組資料X和Y合併
  2. Rank in increasing order
  3. Test statistic W_X = sum of ranks from X's

- E.g., Aspirin concentration

	10	12	13	15	17	18	20	21	22	26	28
Rank	1	2	3	4	5	6	7	8	9	10	11
Aspirin	Y	Y	X	X	X	Y	Y	Y	Y	X	X

$$W_X = 3 + 4 + 5 + 10 + 11 = \underline{33}$$ 統計量

- Reject H_0 if W_X is large (extreme a lot more than what we would expected if H_0 is true)

- Exact distribution of W_X under H_0
  - $N = m + n$
  - Under H_0, $\binom{N}{m}$ possible assignments of ranks are all equally likely with probability $1/\binom{N}{m}$
  - E.g.,$m = 2, n = 3$, $N = 5, \binom{5}{2} = 10$

m : X組sample數
n : Y組sample數

X ranks	Prob.	W_X
1, 2	$1/10$	3
1, 3	$1/10$	4
1, 4	$1/10$	5
1, 5	$1/10$	6
2, 3	$1/10$	5
2, 4	$1/10$	6
2, 5	$1/10$	7
3, 4	$1/10$	7
3, 5	$1/10$	8
4, 5	$1/10$	9

H_0: X組與Y組來自同一個分布

H_1: X與Y組分布不同

Wx很大or很小 ⇒ 排序明顯偏向某一邊 ⇒ X,Y分布不同

双尾檢定 (左右極端拒絕 H_0)

Distribution of W_X under H_0

c	3	4	5	6	7	8	9
$\Pr(W_X = c)$	0.1	0.1	0.2	0.2	0.2	0.1	0.1
$\Pr(W_X \geq c)$	1.0	0.9	0.8	0.6	0.4	0.2	0.1

If $W_X = 9, \Pr_{H_0}(W_X \geq 9) = 0.1$ (one-sided p-value)

- E.g., Aspirin concentration

  $m = 5, n = 6, N = 11, \binom{11}{5} = 462, W_X = 33,$ p-value = 0.33 ① 在全462種可能rank分配中，有多少組rank sum = 33

  (RMD_example 06.10) $p(w_x = 33) = \frac{k}{462}$ k組符合

$p(w_x \geq 33) = \frac{\text{rank sum of} > 33}{462}$

- Large sample approximation ($m > 10,\ n > 10$)

Large sample approximation of W_X under H_0:

$$\frac{W_X - \mathrm{E}_{H_0}(W_X)}{\sqrt{\mathrm{Var}_{H_0}(W_X)}} \sim N(0,1)$$

for $m > 10,\ n > 10$

$$E_{H_0}(W_X) = \frac{m(N+1)}{2}$$

$$Var_{H_0}(W_X) = \frac{mn(N+1)}{12}$$

如果樣本數夠大 ($m>10, n>10$)

可以使用常態分布近似

不用列所有可能組合

Wilcoxon rank-sum test—for gene expression data

- Related to the Mann-Whitney test
- Used with two independent samples
- Test whether or not the difference of population mean $\log_{10}$ expression intensity ratios on gene J00129 between good and poor prognosis groups is equal to 0.
- $H_0\colon \mu_G = \mu_P \quad H_a\colon \mu_G \neq \mu_P$
- The two samples are combined and observations are ranked.

  (RMD_example 06.11)

More than 2 samples

- There are more than two independent groups for comparison
  - Parametric test: **ANOVA** (RMD_example 06.12)
  - Nonparametric test: **Kruskal-Wallis test** – generalization of Wilcoxon rank-sum test to compare more than 2 groups (RMD_example 06.13)

Kruskal-Wallis test—for gene expression data

- More than 2 samples
- Test the equality of population mean $\log_{10}$ expression intensity ratios on gene J00129 among 11 ERp groups (0, 5, 10, 30, 40, 50, 60, 70, 80, 90, 100)
- H_0: $\mu_0 = \mu_5 = \mu_{10} = \mu_{30} = \mu_{40} = \mu_{50} = \mu_{60} = \mu_{70} = \mu_{80} = \mu_{90} = \mu_{100}$

  H_a: not H_0

  (RMD_example 06.13)

Pearson correlation

衡量2個連續變數之間線性關係強度

- Measure the **linear** relationship between two **continuous** random variables X and Y
- Population correlation coefficient, ρ, is defined as

$$\rho = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}}$$

- Given a data set of N points of observations $(x_1, y_1), \cdots, (x_N, y_N)$ from (X, Y), we can use sample correlation coefficient, r, to estimate ρ

$$r = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N}(x_i - \bar{x})^2 \sum_{i=1}^{N}(y_i - \bar{y})^2}}$$

- Hypothesis testing of ρ

  For the test H_0: $\rho = 0$

$$t = r\sqrt{\frac{N-2}{1-r^2}} \xrightarrow[H_0]{} t(N-2)$$

(RMD_example 06.14)

測量2個變數之間單調關係

Rank correlation

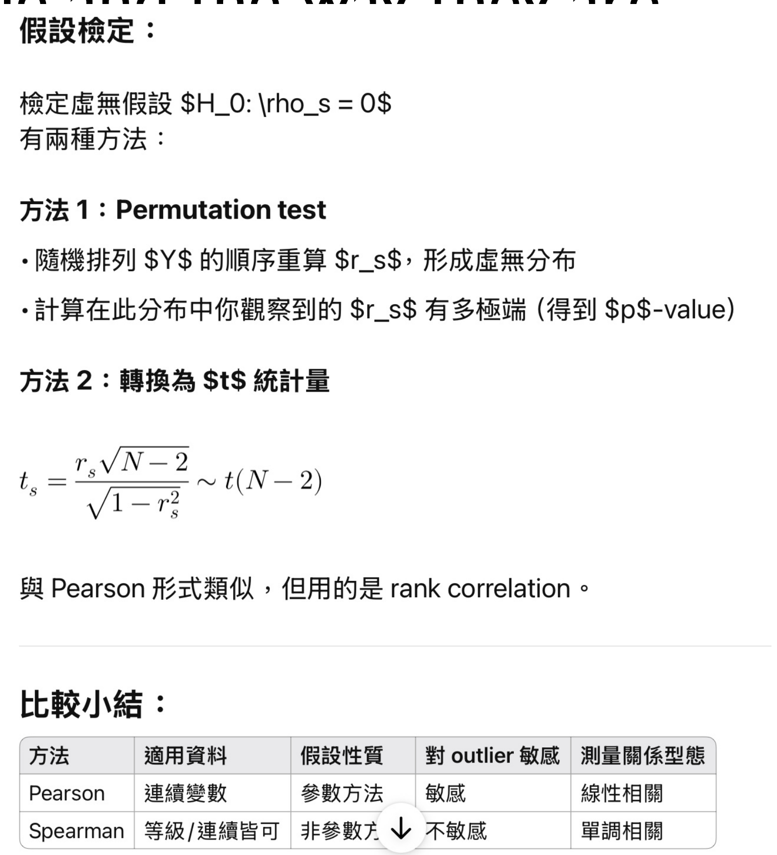
- Used to describe the **monotonic** relationship (whether linear or not) between two ordinal variables (e.g., very difficult, a little difficult, no difficulty), or between one ordinal variable and one continuous variable ① 順序型 var ② outliers
- A non-parametric approach is to use the ranks of the variables to calculate the correlation: **Spearman rank correlation coefficient** ρ_S (population), r_S (sample)

1. Rank the values of X from 1 to N
2. Rank the values of Y from 1 to N
3. Compute the correlation coefficient ρ or r based on ranks

- The value of ρ_s $(\text{or } r_s)$ is less sensitive to outliers
- The significance of the Spearman rank correlation $(H_0\text{: } \rho_s = 0)$ is tested by
  1. Using a permutation test. An advantage of this approach is that it automatically takes into account the number of tied data values in the sample and the way they are treated in computing the ra
  2. Using

$$t_s = r_s\sqrt{\frac{N-2}{1-r_s^2}} \xrightarrow[H_0]{} t$$

(RMD_example 06.15)



假設檢定：

檢定虛無假設 $H_0: \rho_s = 0$
有兩種方法：

方法 1：Permutation test

- 隨機排列 Y 的順序重算 r_s，形成虛無分布
- 計算在此分布中你觀察到的 r_s 有多極端（得到 p-value）

方法 2：轉換為 t 統計量

$$t_s = \frac{r_s\sqrt{N-2}}{\sqrt{1-r_s^2}} \sim t(N-2)$$

與 Pearson 形式類似，但用的是 rank correlation。

比較小結：

方法	適用資料	假設性質	對 outlier 敏感	測量關係型態
Pearson	連續變數	參數方法	敏感	線性相關
Spearman	等級/連續皆可	非參數方	不敏感	單調相關

Summary

- Nonparametric statistical methods provide an alternative to parametric methods when the parametric assumptions cannot be met.
- The use of a nonparametric method does not require knowledge of the underlying population distribution(s) or the Central Limit Theorem.
- A nonparametric test result is usually more conservative than a parametric test result.
- Nonparametric methods are often used with laboratory studies and small sample sizes.