# Lecture 11. Two-sample Comparison (II):

## Nonparametric method

兩個母群體中位數之比較：無母數之方法

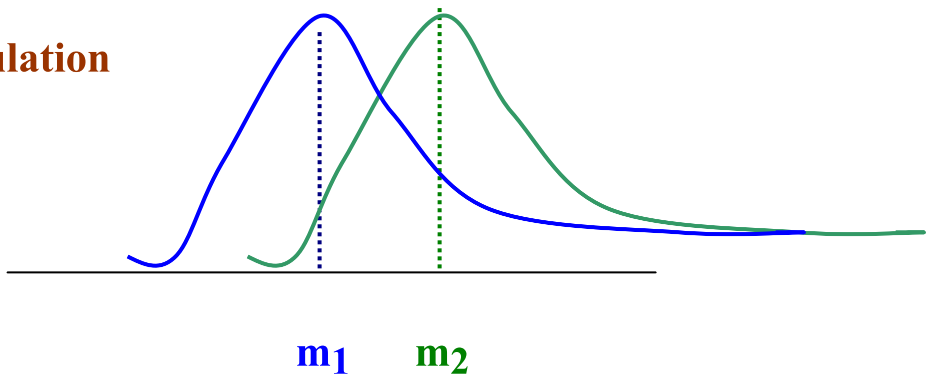
### Nonparametric method

非參數化之方法；無母數之方法

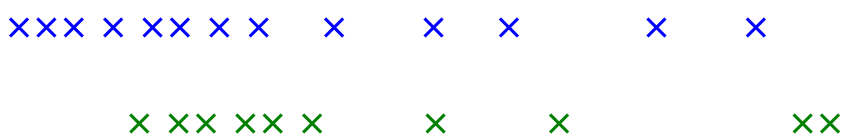
對應前述之參數化方法：t-test, paired t-test (assuming normal distribution(s))

**When the underlying distributions are very similar, but both are far from normal (Gaussian) [e.g., right-skewed], nonparametric methods can be applied.**

**Population**




**Data (sample)**



# One-tailed and Two-tailed test

- **One-tailed**

  $H_0: m_1 \leqq m_2$ (null) vs. $Ha: m_1 > m_2$ (alternative)

- **Two-tailed**

  $H_0: m_1 = m_2$ vs. $Ha: m_1 \neq m_2$

**Note:** comparison between parametric and nonparametric methods for **two-sample** problem.

|  FIG. 4. *Frank Wilcoxon.* | Independent samples | Paired samples |
|---|---|---|
| parametric | T test | Paired t-test |
| nonparametric | Wilcoxon rank sum test | Wilcoxon signed rank test |

**Procedure：**

- To construct the test statistic:
- Sampling distribution of the test statistic under Ho
- Significance level (1-α)
- Type I error (α) and the p-value

# Independent samples

**Data:**

表 13.3　兩組患有苯酮尿症樣本孩童的標準智商年齡分數（nMA）

| 低暴露濃度組 (<10.0 mg/dl) | | 高暴露濃度組 (<10.0 mg/dl) | |
|---|---|---|---|
| nMA（個月） | 等級 | nMA（個月） | 等級 |
| 34.5 | 2.0 | 28.0 | 1.0 |
| 37.5 | 6.0 | 35.0 | 3.0 |
| 39.5 | 7.0 | 37.0 | 4.5 |
| 40.0 | 8.0 | 37.0 | 4.5 |
| 45.5 | 11.5 | 43.5 | 9.0 |
| 47.0 | 14.5 | 44.0 | 10.0 |
| 47.0 | 14.5 | 45.5 | 11.5 |
| 47.5 | 16.0 | 46.0 | 13.0 |
| 48.7 | 19.5 | 48.0 | 17.0 |
| 49.0 | 21.0 | 48.3 | 18.0 |
| 51.0 | 23.0 | 48.7 | 19.5 |
| 51.0 | 23.0 | 51.0 | 23.0 |
| 52.0 | 25.5 | 52.0 | 25.5 |
| 53.0 | 28.0 | 53.0 | 28.0 |
| 54.0 | 31.5 | 53.0 | 28.0 |
| 54.0 | 31.5 | 54.0 | 31.5 |
| 55.0 | 34.5 | 54.0 | 31.5 |
| 56.5 | 36.0 | 55.0 | 34.5 |
| 57.0 | 37.0 | | 313.0 |
| 58.5 | 38.5 | | |
| 58.5 | 38.5 | | |
| | 467.0 | | |

- $X_1, X_2, \ldots, X_n$ ; $Y_1, Y_2, \ldots, Y_m$
- Pooled sample: $Z_1 < Z_2 < Z_3 < Z_4 < \ldots < Z_{n+m}$
- Rank: $1, 2, 3, 4, \ldots, (n+m)$

## (Wilcoxon's) rank sum test:

**If the smaller rank sum (of X-group or Y-group)=W, and the sample size of the corresponding group is $n_S$ (the other is $n_L$). If $n_S$=n, then $n_L$=m; if $n_S$=m, then $n_L$=n.**

**Under null ($H_0$)** $\Longrightarrow$

$$E(W)=[(n_S+n_L)\times(n_S+n_L+1)/2]\times[n_S/(n_S+n_L)]$$

$$=n_S\times(n_S+n_L+1)/2\,;$$

$Var(W)=n_S n_L(n_S+n_L+1)/12$ **(Proof: see Appendix11.1)**

$SE(W)=\sqrt{[n_S n_L(n_S+n_L+1)/12]}$ **(SE=standard error)**

**Approximate Z-score:**

$[W-E(W)]\div SE(W)\sim N(0,1)$, **approximately, when n and m are both large.**

**Example:** **(Ref. to Table 13.3)**

W=313

$E(W)=18\times(18+21+1)\div 2=360$

$Var(W)=18\times 21\times(18+21+1)\div 12=1260$

$SE(W)=\sqrt{(1260)}=35.5$

$Z_W=(313-360)\div 35.5=-1.32$;

P value=$2\times 0.093=0.186$ ……….

# Paired samples

## Data: Table 13.2

表 13.2　囊腫纖維病變患者樣本的強迫性肺活量（FVC）減少的情形

| 病患 | FVC 減少量 (ml) | | 差距 | 等級 | 符號化等級 | |
|---|---|---|---|---|---|---|
| | 安慰劑 | 利尿劑 | | | | |
| 1 | 224 | 213 | 11 | 1 | 1 | |
| 2 | 80 | 95 | −15 | 2 | | −2 |
| 3 | 75 | 33 | 42 | 3 | 3 | |
| 4 | 541 | 440 | 101 | 4 | 4 | |
| 5 | 74 | −32 | 106 | 5 | 5 | |
| 6 | 85 | −28 | 113 | 6 | 6 | |
| 7 | 293 | 445 | −152 | 7 | | −7 |
| 8 | −23 | −178 | 155 | 8 | 8 | |
| 9 | 525 | 367 | 158 | 9 | 9 | |
| 10 | −38 | 140 | −178 | 10 | | −10 |
| 11 | 508 | 323 | 185 | 11 | 11 | |
| 12 | 255 | 10 | 245 | 12 | 12 | |
| 13 | 525 | 65 | 460 | 13 | 13 | |
| 14 | 1023 | 343 | 680 | 14 | 14 | |
| | | | | | 86 | −19 |

$X_1, X_2, \ldots, X_n$ ;

$Y_1, Y_2, \ldots, Y_n$

**Difference:** $d_1, d_2, \ldots, d_n$ ; $d_i = X_i - Y_i$

**Rank** $r_1, r_2, \ldots, r_n$ (for $|d_i|$)

**Sign** + + - -…+ (+: for $d_i$ positive;

- : for $d_i$ negative)

# (Wilcoxon's) signed rank test

**If the smaller |rank sum|=T, (absolute value of rank sum)**

**Under null ($H_0$) $\Longrightarrow$**

**$E(T)=[n(n+1)/2]\div 2=n(n+1)/4$;**

**$Var(T)=n(n+1)(2n+1)/24$ (Homework) (Proof: see below*)**

**$SE(T)=\sqrt{[n(n+1)(2n+1)/24]}$ (SE=standard error)**

**Approximate Z-score:**

**$[T-E(T)]\div SE(T)\sim N(0,1)$, approximately, when n is large.**

**Example: (Table 13.2)**

**$T=19$**

**$E(T)=14\times 15\div 4=52.5$**

**$Var(T)=14\times 15\times 29\div 24=253.75$**

**$SE(T)=15.93$**

**$Z_T=(19-52.5)\div 15.93=-2.10$;**

**P value=$2\times 0.018=0.036$ ……….**

**Proof(★):** Let $W=\Sigma U_i$, where $U_i=0$ (with probability=1/2), and$=r_i$ (with probability=1/2). It is then interesting to note that: the W defined in this way have the same distribution with the statistic T. Also, {Ui} are independent of each other because the outcome of Ui is independent of that of Uj for $i\neq j$. So,

$E(W)=\Sigma(EU_i)=\Sigma(0+i/2)=(1/2)\Sigma_i i=(1/2)(n(n+1)/2)= n(n+1)/4$;

$Var(W)= \Sigma(VarUi)= \Sigma(EU_i^2-( EU_i)^2)= \Sigma(i^2/2-( i/2)^2)= \Sigma(i^2/4)$

$=(1/4)( n(n+1)(2n+1)/6)$ QED

# SAS code

```
data twospl;
input dlco group $ @@;
cards;
7.51    emp   10.81   emp    11.75   emp    12.59    emp
13.47   emp   14.18   emp    15.25   emp    17.40    emp
17.75   emp   19.13   emp    20.93   emp    25.73    emp
26.16   emp
6.19    no_emp    12.11    no_emp    14.12    no_emp
15.50   no_emp    15.52    no_emp    16.56    no_emp
17.06   no_emp    19.59    no_emp    20.21    no_emp
20.35   no_emp    21.05    no_emp    21.41    no_emp
23.39   no_emp    23.60    no_emp    24.05    no_emp
25.59   no_emp    25.79    no_emp    26.29    no_emp
29.60   no_emp    30.88    no_emp    31.42    no_emp
32.66   no_emp    36.16    no_emp
;
proc univariate plot normal data=twospl;
var dlco;
by group;
run;

proc npar1way Wilcoxon;
class group;
var dlco;
/* exact Wilcoxon; */ /* Time consuming for non-sparse data*/
run;
```

# output

Wilcoxon Scores (Rank Sums) for Variable dlco
Classified by Variable group

| group | N | Sum of Scores | Expected Under H0 | Std Dev Under H0 | Mean Score |
|---|---|---|---|---|---|
| emp | 13 | 168.0 | 240.50 | 30.363081 | 12.923077 |
| no_emp | 23 | 498.0 | 425.50 | 30.363081 | 21.652174 |

Wilcoxon Two-Sample Test

| | |
|---|---|
| Statistic (S) | 168.0000 |
| Normal Approximation | |
| Z | -2.3713 |
| One-Sided Pr < Z | 0.0089 |
| Two-Sided Pr > \|Z\| | 0.0177 |
| t Approximation | |
| One-Sided Pr < Z | 0.0117 |
| Two-Sided Pr > \|Z\| | 0.0234 |
| Exact Test | |
| One-Sided Pr <= S | 0.0081 |
| Two-Sided Pr >= \|S - Mean\| | 0.0162 |

Z includes a continuity correction of 0.5.

Kruskal-Wallis Test

| | |
|---|---|
| Chi-Square | 5.7014 |
| DF | 1 |
| Pr > Chi-Square | 0.0170 |

# Appendix 11.1 (★)

## Variance of the Wilcoxon rank sum statistic

Let $\mathbf{g} = (\mathbf{r}_1, \dots, \mathbf{r}_{n_S}, \mathbf{q}_1, \dots, \mathbf{q}_{n_L})^T$, where $\mathbf{a}^T$ denotes the *transpose* of matrix $\mathbf{a}$; $N = n_S + n_L$, and

$$W = \mathbf{r}_1 + \dots + \mathbf{r}_{n_S}.$$

The main quantity we want to calculate is

$$\mathrm{VAR}(W) = \mathrm{E}(W^2) - (\mathrm{E}W)^2.$$

It is easy to deduce that $\mathrm{E}(W) = n_S(N+1)/2$ according to the 'uniformly distributed' principle (**UDP**) [explained in the class, not a generally used terminology in Statistics!]. The term remained to be calculated is $\mathrm{E}(W^2) = \mathrm{E}(\sum_{i=1}^{n_S} \mathbf{r}_i^2 + 2\sum_{i<j} \mathbf{r}_i\mathbf{r}_j)$. To this end, consider the cross-product matrix $\mathbf{g}\mathbf{g}^T$

$$\mathbf{g}\mathbf{g}^T \equiv (g_{ij}) = \begin{pmatrix} \mathcal{G}_1 & \mathcal{G}_2 \\ \mathcal{G}_3 & \mathcal{G}_4 \end{pmatrix},$$

where $\mathcal{G}_1 = (\mathbf{r}_i\mathbf{r}_j)$, $\mathcal{G}_2 = (\mathbf{r}_i\mathbf{q}_j)$ $\mathcal{G}_3 = \mathcal{G}_2^T$, and $\mathcal{G}_4 = (\mathbf{q}_i\mathbf{q}_j)$. Note that the sum of all elements in $\mathbf{g}\mathbf{g}^T$ is

$$\sum_{i,j} g_{ij} = (1 + \dots + N)^2 = \frac{N^2(N+1)^2}{4} = \sum_{i,j} \mathcal{G}_{1,ij} + \sum_{i,j} \mathcal{G}_{2,ij} + \sum_{i,j} \mathcal{G}_{3,ij} + \sum_{i,j} \mathcal{G}_{4,ij},$$

because the element of $\mathbf{g}$ is only a *re-alignment* of $(1, 2 \dots, N)^T$. Further, $\sum_{i,j} \mathcal{G}_{1,ij} = \mathrm{E}(W^2)$.

The diagonal part of $\mathbf{g}\mathbf{g}^T$ has the sum $1^2 + \dots + N^2$; so under $H_0$ and according to the **UDP**, the sum of diagonal part of $\mathcal{G}_1$ has the expectation:

$$\mathrm{E}(\sum_{i=1}^{n_S} \mathbf{r}_i^2) = \{\frac{1}{6}N(N+1)(2N+1)\} \times \frac{n_S}{N}. \tag{1}$$

There remains $N^2 - N$ and $n_S^2 - n_S$ off-diagonal terms in $\mathbf{g}\mathbf{g}^T$ and $\mathcal{G}_1$, respectively. By excluding the squared terms, the $\mathbf{g}\mathbf{g}^T$-matrix has a sum of the cross-product terms as

$$\sum_{i \neq j} g_{ij} = (1 + \dots + N)^2 - (1^2 + \dots + N^2) = \frac{N^2(N+1)^2}{4} - \frac{1}{6}N(N+1)(2N+1).$$

So by a simialr argument with **UDP**, the expectation (under $H_0$) of the sum of off-diagonal terms in $\mathcal{G}_1$ is

$$2\mathrm{E}(\sum_{i<j} \mathbf{r}_i\mathbf{r}_j) = \{\frac{N^2(N+1)^2}{4} - \frac{1}{6}N(N+1)(2N+1)\} \times \frac{n_S(n_S-1)}{N(N-1)}. \tag{2}$$

The variance of $W$ is then easily calculated as

$$\mathrm{VAR}(W) = (1) + (2) - (\frac{n_S(N+1)}{2})^2 = \dots = \frac{n_S n_L(N+1)}{12}$$