# Nonparametric Methods

# Outline

- Types of data
- Nonparametric tests
- Sign test
- Sign rank test
- Sign sum test
- Sample size calculation

# Types of data

- The type of data collected in a study determine the type of statistical analysis used.
- Data can be broadly classified into three main types:
  - Cardinal data
    - Interval scale
    - Ratio scale
  - Ordinal data
  - Nominal data — **Categorical** Data

# Cardinal Data

- on a scale where it is meaningful to measure the distance between possible data values.

- Examples
  - Height
  - Age
  - Exam marks
  - Size of bicycle frame
  - Time to complete a statistics test
  - Number of cigarettes smoked

# Ordinal Data

- data can be ordered but do not have specific numeric values.

- Examples
  - Degree of illness
    - none, mild, moderate, acute, chronic.
  - Opinion of students about stats classes
    - Very unhappy, unhappy, neutral, happy, ecstatic!
  - Attitudes towards the death penalty
    - Strongly disagree, disagree, neutral, agree, strongly agree.

# Nominal Data

- different data values can be classified into categories but the categories have no specific ordering.

- Examples
  - Type of Bicycle
    - Mountain bike, road bike, chopper, folding, BMX.
  - Newspapers:
    - The Sun, The Mail, The Times, The Guardian, the Telegraph.
  - Smoking status
    - smoker, non-smoker

# Nonparametric tests

- “Distribution free" methods require fewer assumptions than parametric methods
- Focus on testing rather than estimation
- Not sensitive to outlying observations
- Especially useful for cruder data (like ranks)
- “Throws away” some of the information in the data
- May be less powerful than parametric counterparts, when the parametric assumptions are true
- For large samples, are equally efficient to parametric counterparts

# Example
# **Dermatology**

- Suppose we want to compare the effectiveness of two ointments (A, B) in reducing excessive redness in people who cannot otherwise be exposed to sunlight.

- Ointment A is randomly applied to either the left or right arm, and ointment B is applied to the corresponding area on the other arm.

- The person is then exposed to 1 hour of sunlight, and the two arms are compared for degrees of redness.

# Which test can you use?

- Denote by A/B the redness score of arm being given ointment A/B,

  Sample 1, (A1,B1)

  Sample 2, (A2, B2)

  …
- What if
  - The exact value of A/B is not available, but we only know which one is better than the other
  - The exact value of A/B is available
  - The experiment is done in two independent populations

| pair | A>B |
|---|---|
| 1 | + |
| 2 | + |
| 3 | - |
| 4 | / |
| 5 | + |
| 6 | - |
| 7 | - |
| 8 | - |
| 9 | + |
| 10 | + |
| … | … |

- Case 1:
  - Does the number of + equal to the number of -?
  - P(+|not /)=0.5

# Sign Test

- Let $d_i$ = difference (A-B)
- Let C$= \sum_i I(d_i > 0)$ be the number of whom $d_i$ > 0 out of the total of $n$ people with nonzero $d_i$
  - C is then $\text{binomial}(n, p)$
- The sign test tests whether $H_0: p = .5$ using C and n

- Let Δ be the population median of the $d_i$
- Notice that $\Delta = 0$ iff $p = P(d > 0 | d \neq 0) = .5$
- $H_0: \Delta = 0$ versus $H_1: \Delta \neq 0$ $(\text{or} > \text{or} <)$

# Normal-Theory Method

- Since $E(C) = np = n/2$ and $Var(C) = npq = n/4$

$$\frac{C - n/2}{\sqrt{n/4}} \xrightarrow{D} N(0,1)$$

- Using a continuity correction,

$$\frac{C - \frac{n}{2} - \frac{1}{2}}{\sqrt{n/4}} \xrightarrow{D} N(0,1)$$

- we have (for $n \geq 20$):

- To test $H_0 : \Delta = 0$ versus $H_1 : \Delta \neq 0$, if

$$C > c_2 = \frac{n}{2} + \frac{1}{2} + z_{1-\alpha/2}\sqrt{\frac{n}{4}} \text{ or } C < c_2 = \frac{n}{2} - \frac{1}{2} - z_{1-\alpha/2}\sqrt{\frac{n}{4}}$$

then $H_0$ is rejected. Otherwise, $H_0$ is accepted.

# Exact Method

- If $n < 20$, use the exact binomial probabilities instead.

**Computation of the $p$-Value for the Sign Test (Exact Test)**

If $C > n/2$, $\quad p = 2 \times \sum_{k=C}^{n} \binom{n}{k} \left(\frac{1}{2}\right)^n$

If $C < n/2$, $\quad p = 2 \times \sum_{k=0}^{C} \binom{n}{k} \left(\frac{1}{2}\right)^n$

If $C = n/2$, $\quad p = 1.0$

| pair | A>B | A-B |
|---|---|---|
| 1 | + | 5 |
| 2 | + | 10 |
| 3 | - | -1 |
| 4 | / | 0 |
| 5 | + | 7 |
| 6 | - | -1 |
| 7 | - | -2 |
| 8 | - | -1 |
| 9 | + | 8 |
| 10 | + | 6 |
| … | … | |

- Case 2
  - Do you agree that A has the same effect as B?
- Wilcoxon signed-rank test

# Signed rank procedure

① Take the paired differences

② Take the absolute values of the differences

③ Rank these absolute values, throwing out the 0s

④ Multiply the ranks by the sign of the difference (+1 for a positive difference and -1 for a negative difference)

⑤ Calculate the rank sum $W_+$ of the positive ranks

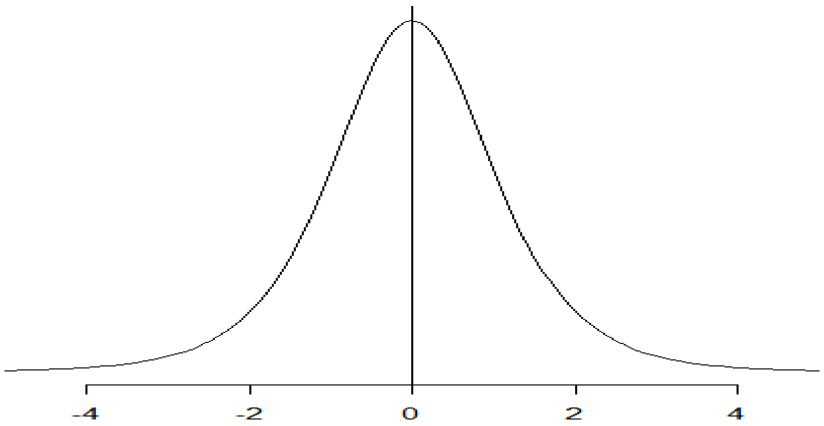
# Signed-rank test

- The statistic:

$$W_+ = \sum_{i=1}^{n} R_i \varphi_i$$

where $R_i$ is the rank of $|d_i|$, and $\varphi_i$ is the indicator variable of $d_i$, $\varphi_i = \mathrm{I}(d_i > 0)$

- $H_0$ ?



# distribution of $W_+$

- Under $H_0$ and if there are no ties

$$E(W_+) = n(n+1)/4$$

$$Var(W_+) = \frac{\mathrm{n(n+1)(2n+1)}}{24}$$

  - There is a correction term necessary for ties

$$Var(W_+) = \frac{\mathrm{n(n+1)(2n+1)}}{24} - \sum_{i=1}^{g} \frac{t_i^3 - t_i}{48}$$

  where $t_i$ refers to the number of differences with the same absolute value in the $i$th tied group and $g$ is the number of tied groups

- For large sample size n, $W_+$ follows a normal distribution

# The Wilcoxon Rank-Sum Test

Also known as the Mann-Whitney test

# Example
## Health Services Administration

- Suppose we want to compare the length of hospital stay for patients with the same diagnosis at two different hospitals.
- The results are shown in Table 9.8.

| Table 9.8 Comparison of length of stay in 2 hospitals | |
|---|---|
| First hospital | 21, 10, 32, 60, 8, 44, 29, 5, 13, 26, 33 |
| Second hospital | 86, 27, 10, 68, 87, 76, 125, 60, 35, 73, 96, 44, 238 |

# The Wilcoxon Rank-Sum Procedure

① Discard the treatment labels

② Rank the observations

③ Calculate the sum of the ranks in the first treatment, denoted by $W$

④ Judgment

- calculate the asymptotic normal distribution of this statistic
- compare with the exact distribution under the null hypothesis( $\min(n_1, n_2) < 10$ )

# Wilcoxon Rank-sum test

- $X_1, X_2, \ldots, X_{n_1} \sim F, Y_1, Y_2, \ldots, Y_{n_2} \sim G$
- The statistic

$$W = \sum_{i=1}^{n_1} rank(X_i)$$

$$= \sum_{i=1}^{n_1} \left( \sum_{j=1}^{n_1} I(X_j \le X_i) + \sum_{j=1}^{n_2} I(Y_j \le X_i) \right)$$

$$= \sum_{i=1}^{n_1} \sum_{j=1}^{n_1} I(X_j \le X_i) + \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} I(Y_j \le X_i)$$

$$= \frac{n_1(n_1 + 1)}{2} + \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} I(Y_j \le X_i)$$

Mann-Whitney Statistic U

# Distribution of $W$

- $H_0: F = G$
- Under $H_0, W$ has an exact distribution. If there are no ties
  - $E(W) = (n_1\ (n_1 + n_2 + 1))/2$
  - $Var(W) = \mathrm{n}_1\mathrm{n}_2(\mathrm{n}_1 + \mathrm{n}_2 + 1)/12$
- There is a correction term necessary for ties
  - $Var(W_+) = \frac{\mathrm{n}_1\mathrm{n}_2(\mathrm{n}_1+\mathrm{n}_2+1)}{12} - \sum_{i=1}^{g} \frac{n_1 n_2 \left(t_i^3 - t_i\right)}{12(n_1+n_2)(n_1+n_2+1)}$

  where $t_i$ refers to the number of differences with the same absolute value in the $i$th tied group and $g$ is the number of tied groups

  > ?dwilcox
- For large sample size n, $W$ follows a normal distribution

# Hypotheses

– Wilcoxon Signed rank

$f(\cdot)$ is symmetric around Δ

$H_0$: Δ=0 vs. $H_1$: Δ ≠0

– Wilcoxon Rank sum

Location shift model: $G(x) = F(x - \Delta)$ or $Y = \mathrm{X} + \Delta$

$H_0$: Δ=0 vs. $H_1$: Δ ≠0

# Sample-size Calculation

# Sign test

- One-sample case: $Z_1, Z_2, \ldots, Z_n \sim f$
  - Let $\theta$ be the population median of the $f(x)$
  - $H_0: \theta = 0$ versus $H_1: \theta \neq 0$
    $\Leftrightarrow H_0:\ p = P(Z > 0 | Z \neq 0) = 0.5 \text{ vs. } p \neq 0.5$
  - $C = \sum_{i=1}^{n} I(Z_i > 0)$, the test statistic $T = \frac{C - n/2}{\sqrt{n/4}}$
  - Under the alternative, $C \sim B(n, p_1)$

$$1 - \beta = P\left(T > z_{1-\alpha/2} \middle| p = p_1\right)$$

# Wilcoxon signed rank test

- One-sample case: $Z_1, Z_2, \ldots, Z_n \sim f, f(\cdot)$ is symmetric around $\theta$
  - $H_0$: $\theta = 0$ vs. $H_1$: $\theta \neq 0$
  - The working model:

$$Z_{\mathrm{i}} = \theta + e_{\mathrm{i}}$$

  - Test statistic: $T_+ = \sum_{i=1}^{n} R_i \varphi_i$

  wilcox sample size.pdf
- Two-sample Wilcoxon rank-sum test: also refer to the linked paper

# AUC & M-W U

- For two independent samples

$$X_1, X_2, \ldots, X_{n_1}; \text{ and } Y_1, Y_2, \ldots, Y_{n_2}$$

Giving labels $Z = (\underbrace{0, \ldots 0}_{n_1}, \underbrace{1, \ldots 1}_{n_2})$, the AUC of $(X_1, \ldots, X_{n_1}, Y_1, \ldots, Y_{n_2})$ for predicting $Z$

$$AUC = \frac{U}{n_1 n_2}$$

Where $U$ is the Mann-Whitney Statistic.

https://en.wikipedia.org/wiki/Mann%E2%80%93Whitney_U_test#Relation_to_other_tests

# C-index

- For random variables $\xi = (X_1, \ldots, X_{n_1}, Y_1, \ldots, Y_{n_2})$ and Z,

$$\text{c-index}=P(\xi_1 < \xi_2 | Z_1 < Z_2)$$
$$\Leftrightarrow P(\xi_1 < \xi_2 | Z_1 = 0, Z_2 = 1)$$
$$\Leftrightarrow P(X_1 < Y_1)$$

- More on AUC, M-W U & C-index
  - They are equivalent for classification problems
  - How about regression? i.e., Z is continuous?
    - Only C-index works
    - One still can make a new variable Z'=I(Z>C), and take Z' as Z, but the equivalence between AUC (M-W U) and C-index disappear
    - Extension to survival analysis