Rank Tests

Applied Statistics

Fall 2025

目录

1 非参数检验简介 Introduction to Non-parametric Tests 4

1.1 动机 Motivation 4

1.2 处理非正态数据的方法 Measures for Non-normal Data 4

1.3 秩和检验概述 Overview of Rank Tests 5

1.4 常用秩和检验对照表 Common Rank Tests Comparison 6

2 Wilcoxon 秩和检验（两个独立样本）Wilcoxon Rank Sum Test 6

2.1 基本概念与动机 Basic Concept and Motivation 6

2.2 秩变换 The Rank Transformation 6

2.3 检验统计量 The Test Statistic 7

2.4 精确抽样分布与计算 The Exact Sampling Distribution and Calculation . 8

2.4.1 构建精确分布 8

2.4.2 示例计算 8

2.5 理论均值与标准差 Theoretical Mean and Standard Deviation 9

2.6 正态近似 The Normal Approximation 10

2.7 检验的原假设与备择假设 Hypotheses 10

2.8 实例与软件输出 Example and Software Output 11

2.8.1 练习题：顶级水疗中心的房间数 Practice: Numbers of rooms in top spas 11

2.8.2 Stata 软件输出 Stata Output 12

2.9 处理“结”（Ties） 13

2.10 历史注记 A Little Bit of History 14

3 Wilcoxon 符号秩检验（配对样本）Wilcoxon Signed Rank Sum Test 15

3.1 基本概念与动机 15

3.2 检验步骤与统计量 16

3.2.1 示例计算 17

3.3 原假设下的行为与精确分布 17

3.4 正态近似 17

3.5 处理“结”与零差值 Ties and Zero Differences 18

3.6 软件输出示例 Software Output Example 19

4 符号检验 Sign Test **20**

4.1 基本概念 20

4.2 何时使用 When to Use 21

4.3 示例 21

5 应用、比较与总结 Applications, Comparisons and Summary **22**

5.1 在顺序数据上的应用 Application on Ordinal Data 22

5.2 秩和检验与 t 检验的比较 Comparing Rank Tests with t-tests 24

5.3 选择检验的指导原则 Guidelines for Choosing a Test 25

大纲 Outline

1. 非参数检验简介 Introduction to Non-parametric Tests
2. Wilcoxon 秩和检验（两个独立样本）Wilcoxon Rank Sum Test (Two Independent Samples)
3. Wilcoxon 符号秩检验（配对样本）Wilcoxon Signed Rank Sum Test (Paired Samples)
4. 符号检验 Sign Test
5. 应用与比较 Applications and Comparisons

1 非参数检验简介 Introduction to Non-parametric Tests

1.1 动机 Motivation

- 我们之前学习的大多数方法（如 t 检验）都假设总体服从正态分布或样本量足够大（中心极限定理）。

  Most methods we learned (e.g., t tests) assume Normal populations or large samples (Central Limit Theorem).

- 在实践中，没有分布是严格正态的。

  In practice, no distribution is exactly normal.

- 单样本和两样本 t 检验是相当稳健的：结果对中等程度的非正态性不敏感，特别是当样本量合理时。

  The one-sample and two-sample t procedures are quite robust: the results are not very sensitive to moderate lack of Normality, especially when the samples are reasonably large.

- 如果图表显示数据不服从正态分布，特别是当我们只有少量观测值时，该怎么办？

  What if plots suggest that the data are not Normal, especially when we have only a few observations?

- 即使样本量不大，我们仍然希望检查研究结果的稳健性。

  Even if the sample size is modest, we still want to check the robustness of our findings.

1.2 处理非正态数据的方法 Measures for Non-normal Data

- **处理异常值 Outliers:** 如果非正态性是由异常值引起的，合理地剔除异常值可能是可行的。

  If the lack of normality is due to outliers, it may be legitimate to remove the outliers.

- **数据转换 Transformation:** 有时我们可以转换数据，使其分布更接近正态。例如，对数变换可以拉回右偏分布的长尾，这特别有用。

  Sometimes we can transform our data so that their distribution is more nearly Normal. Transformations such as the logarithm that pull in the long tail of right-skewed distributions are particularly helpful.

- **非参数检验 Non-parametric tests:** 当数据转换无效或不合适时，可以考虑使用不依赖特定分布假设的检验方法。

  When data transformation is ineffective or inappropriate, consider using tests that do not rely on specific distributional assumptions.

  - **秩和检验 Rank tests:** 将数据视为秩次而不是实际值进行分析。
    The data is treated as ranks instead of using the actual value.
  - **重抽样方法 Resampling:** 如自助法（Bootstrap）和置换检验（Permutation tests），通过重抽样数据直接生成用于推断的抽样分布。
    Bootstrap methods and permutation tests, resampling the data to directly generate the sampling distribution for inference.

1.3 秩和检验概述 Overview of Rank Tests

- 秩和检验是一种非参数方法，不要求总体分布具有特定形式。
  Rank tests are nonparametric methods that do not require any specific form of population distribution.
- 其核心思想是基于每个观测值的秩次（排序中的位置）。这些检验关注的是一个或多个总体的位置。
  The core idea is to focus on the rank (place in order) of each observation. These tests concern the position of a population or populations.
- **优点 Advantages:**
  - 通过排序，减少了极端值和分布形状的影响。
    By ranking the data, we reduce the influence of extreme values and distribution shape.
  - 对异常值不敏感。
    Less sensitive to outliers.
  - 适用于顺序数据（Ordinal Data），例如满意度评分。
    Applicable to ordinal data (e.g., satisfaction scores).
- **缺点 Disadvantages:**
  - 失去了关于实际数值大小的信息。
    We lose information about the actual magnitudes.
  - 当数据确实服从正态分布时，其统计功效（Power）通常低于对应的参数检验（如 t 检验）。
    When the data are indeed Normal, rank tests are generally less powerful than their parametric counterparts (e.g., t-test).

1.4 常用秩和检验对照表 Common Rank Tests Comparison

场景	参数检验 (正态假设)	秩和检验
单样本	单样本 t 检验	Wilcoxon 符号秩检验
配对样本	基于差异的单样本 t 检验	Wilcoxon 符号秩检验
两个独立样本	两样本 t 检验	Wilcoxon 秩和检验

表 1: 不同场景下的检验方法选择 Selection of Test Methods in Different Settings

2 Wilcoxon 秩和检验（两个独立样本）Wilcoxon Rank Sum Test

2.1 基本概念与动机 Basic Concept and Motivation

- 用于比较两个独立样本的位置。

  To compare the positions of two independent samples.

- **示例 Example:** 研究田间杂草是否影响玉米产量。

  Does having weeds in the field affect corn yield?

杂草数量 (每米) Weeds per meter	产量 (蒲式耳/英亩) Yield (bu/acre)			
0	166.7	172.2	165.0	176.9
3	158.6	176.4	153.1	156.0

- 样本量太小，无法充分评估正态性或依赖两样本 t 检验的稳健性。

  The samples are too small to assess Normality adequately or to rely on the robustness of the two-sample t test.

2.2 秩变换 The Rank Transformation

- 首先将所有观测值（来自两个样本）从小到大排序。

  First arrange all observations from the smallest to the largest.

- 每个观测值的秩就是它在有序列表中的位置，最小的观测值秩为 1。

  The rank of each observation is its position in this ordered list, starting with rank 1 for the smallest observation.

- **示例数据排序 Ranking Example Data:**

产量 Yield	153.1	156.0	158.6	165.0	166.7	172.2	176.4	176.9
秩 Rank	1	2	3	4	5	6	7	8
组别 Group	Weeds	Weeds	Weeds	**No Weeds**	**No Weeds**	**No Weeds**	Weeds	**No Weeds**

- **粗体**表示无杂草组（第一样本）的产量。

  The **boldface** entries are the yields with no weeds present (the first sample).

- 秩变换只保留了观测值的顺序信息，而丢弃了数值大小。

  The rank transformation retains only the ordering of the observations, not numerical values.

- 秩允许我们不假设分布的形状。

  Ranks allow us to make no assumptions about the shape of the distribution.

2.3 检验统计量 The Test Statistic

- 从 1 到 8 的秩和总是等于 36。如果杂草没有影响，我们期望每组秩和都为 18（36 的一半）。

  The sum of the ranks from 1 to 8 is always equal to 36. If weeds have no effect, we would expect the sum of the ranks in each group to be 18 (half of 36).

- 如果杂草的存在降低了玉米产量，我们预期杂草组的秩更小。如果杂草组的秩和太小，我们可以拒绝原假设。

  If the presence of weeds reduces corn yields, we expect the ranks of the weeds plot to be smaller. If the sum of ranks for the weeds plot is too small, we can conclude that the null hypothesis may be rejected.

- 因此，第一个样本的秩和可以作为我们的检验统计量 W。

  The rank sum of the first sample can thus be our test statistic W.

处理组 Treatment	**秩和 Sum of Ranks**
无杂草 No weeds	$4+5+6+8=23$
有杂草 Weeds	$1+2+3+7=13$

本例中，$W=23$（以“无杂草”组为第一样本）。

In this example, $W=23$ (taking the ”No weeds” group as the first sample).

2.4 精确抽样分布与计算 The Exact Sampling Distribution and Calculation

2.4.1 构建精确分布

- 在原假设下，我们可以通过列出所有可能的组合、计算秩和并统计频率来构建精确的抽样分布。

  We can build up the exact sampling distribution under the null hypothesis by listing every possible combination, calculating the rank sums, and counting the frequency.

- **步骤 Steps:**

  1. 枚举所有可能的秩分配组合。工作量增长很快：每组 4 个观测值，我们需要处理 $C_8^4 = 70$ 种组合。如果每组 12 个观测值呢？

     Enumerate all possible combinations of ranks. The workload explodes fast; with 4 obs in each group we are facing $C_8^4 = 70$ combinations to work with. How about 12 obs in each group?

  2. 计算每种组合下第一样本的秩和 W。

     Calculate the rank sum W for the first sample in each combination.

  3. 任一秩和 W 出现的概率等于其出现次数除以总组合数。

     The probability of any rank sum W is the number of occurrences divided by the total number of combinations.

- 这个分布取决于两个样本量 n_1 和 n_2。

  This distribution depends on the two sample sizes n_1 and n_2.

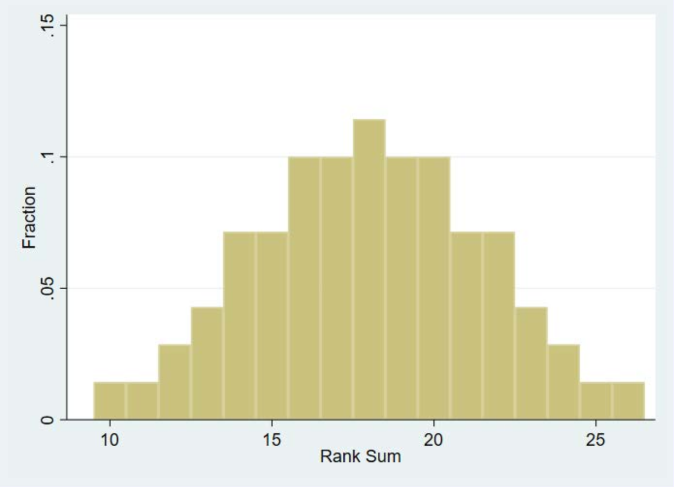
2.4.2 示例计算

精确分布计算示例

对于 $n_1 = 4, n_2 = 4$，所有可能的秩组合存储在“exact”工作表中。R 代码示例如下：

For $n_1 = 4, n_2 = 4$, all possible rank combinations are stored in the ”exact” sheet. Example R code:

```
n1 <- 4 # size group 1
n2 <- 4 # size group 2
N <- n1 + n2 # total number of subjects
rankMat <- combn(1:N, n1) # all possible ranks within group 1
LnPl <- colSums(rankMat) # all possible rank sums for group 1
write.csv(rankMat, "sampling.csv") # export combinations to excel
```



在之前的例子中，我们得到秩和 $W = 23$。根据精确分布，我们可以计算 $P(W \geq 23)$。

In the previous example, we observed $W = 23$. Based on the exact distribution, we can calculate $P(W \geq 23)$.

- 计算得到精确分布的均值 $\mu_W = 18$，标准差 $\sigma_W = 3.464$。
  The exact distribution has mean $\mu_W = 18$ and standard deviation $\sigma_W = 3.464$.

- 在 70 种组合中，秩和大于或等于 23 的组合有 7 种。
  There are 7 combinations with rank sum greater or equal to 23.

- 因此，单侧 P 值 $P(W \geq 23) = 7/70 = 0.1$。
  Therefore, the one-sided P-value is $P(W \geq 23) = 7/70 = 0.1$.

2.5 理论均值与标准差 Theoretical Mean and Standard Deviation

- 设两个样本的观测值数量分别为 n_1 和 n_2，总观测数 $N = n_1 + n_2$。
  Suppose two samples have n_1 and n_2 observations and $N = n_1 + n_2$.

- 第一样本的秩和 W 称为 Wilcoxon 秩和统计量。
  The sum W of the ranks for the first sample is the Wilcoxon rank sum statistic.

- 如果两个总体具有相同的连续分布（原假设），则 W 的均值和标准差为：
  If the two populations have the same continuous distribution (the null hypothesis),

then W has mean and standard deviation:

$$\mu_W = \frac{n_1(N+1)}{2}$$

$$\sigma_W = \sqrt{\frac{n_1 n_2(N+1)}{12}}$$

- **推导备注 Note on Derivation:** 公式推导涉及组合数学。详细推导和如何处理数据中的“结”（ties）可参考相关文献（https://www.stat.berkeley.edu/）。

  The algebra of how to derive the μ_W and σ_W can be found in the literature. The website also gives a more detailed discussion of how to deal with ties in the data.

2.6 正态近似 The Normal Approximation

- 随着两个样本量的增加，秩和统计量 W 的分布近似于正态分布。

  The rank sum statistic W becomes approximately Normal as the two sample sizes increase.

- 我们可以通过标准化 W 来构造 z 统计量：

  We can form a z statistic by standardizing W:

$$z = \frac{W - \mu_W}{\sigma_W} = \frac{W - n_1(N+1)/2}{\sqrt{n_1 n_2(N+1)/12}}$$

- 然后使用标准正态分布计算 P 值。

  Use standard Normal probability calculations to find P-values.

- **继续玉米产量示例 Continuing the Corn Yield Example:**

$$z = \frac{23-18}{3.464} \approx 1.44$$

$$\text{单侧}P\text{ 值} = P(z \geq 1.44) = 0.0749$$

- 使用正态近似的 Wilcoxon 秩和检验有时也被称为 **Mann-Whitney U 检验**。这两个检验本质上是等价的，只是统计量的表达形式不同。

  The Wilcoxon rank sum test with a normal approximation is sometimes referred to as the **Mann-Whitney test**. The two tests are essentially equivalent, differing only in the form of the test statistic.

2.7 检验的原假设与备择假设 Hypotheses

- 非参数检验检测的是分布上的差异，而不仅仅是均值或中位数。

  Nonparametric tests detect differences in distributions, not just means or medians.

- **典型表述 Typical Formulation:**
  - H_0：两个总体的分布相同。

    H_0: The two distributions are the same.
  - H_1：一个总体的值系统地大于另一个总体。

    H_1: One distribution has values that are systematically larger. (This is a location shift alternative.)
- **直观理解 Intuition:** 以玉米产量为例，设 X_1 为无杂草产量，X_2 为有杂草产量。“系统地更大”意味着，对于任意一个给定的阈值（例如 160），无杂草产量超过该阈值的概率大于有杂草产量超过该阈值的概率。

  Let X_1 be corn yield with no weeds and X_2 to be corn yield with weeds: ”systematically larger” would mean that yields higher than a given value (say 160) should be more likely in weed-free yields. $P(X_1 > 160) > P(X_2 > 160)$.
- **关于中位数的假设 Hypotheses about Medians:**
  - 原假设并不一定是关于总体中位数相等的。

    The hypothesis is not necessarily about population medians.
  - 只有在附加了一个很强的假设——**两个总体的分布形状完全相同**时，位置参数的检验才可以等价地表述为中位数的检验。

    Only with an additional assumption: both populations must have distributions of the **same shape**, the hypotheses can be stated in terms of medians:

    $$H_0 : \text{median}_1 = \text{median}_2 \quad \text{vs} \quad H_1 : \text{median}_1 > \text{median}_2$$
  - 在实践中，建议用文字描述关于分布的假设，以避免误解。

    In practice, it is recommended to express the hypotheses on the distributions in words.

2.8 实例与软件输出 Example and Software Output

2.8.1 练习题：顶级水疗中心的房间数 Practice: Numbers of rooms in top spas

练习题

数据 Data:

- A 组（高排名水疗中心随机选取 5 个）：552, 448, 68, 243, 30

  Group A (5 spas randomly selected from top-ranked spas): 552, 448, 68, 243, 30
- B 组（低排名水疗中心）：329, 780, 560, 540, 240

Group B (from the lower-ranked spas): 329, 780, 560, 540, 240

任务 Tasks:

1. 将所有观测值一起排序，列出 A 组和 B 组的秩。
   Rank all of the observations together and make a list of the ranks for Group A and Group B.
2. 陈述此场景下合适的原假设和备择假设，并计算检验统计量 W 的值。
   State appropriate null and alternative hypotheses for this setting and calculate the value of W, the test statistic.
3. 计算 μ_W, σ_W 和标准化的秩和统计量 z。然后使用正态近似给出近似的 P 值。
   Find μ_W, σ_W, and the standardized rank sum statistic z. Then give an approximate P-value using the Normal approximation.
4. 你得出了什么结论？
   What do you conclude?

2.8.2 Stata 软件输出 Stata Output

Stata 输出结果

```
Two-sample Wilcoxon rank-sum (Mann-Whitney) test

group |     obs    rank sum    expected
-------------+---------------------------------
A |       5         21         27.5
B |       5         34         27.5
-------------+---------------------------------
combined |      10          55          55

unadjusted variance       22.92
adjustment for ties        0.00
-----------------
adjusted variance         22.92

Ho: room(group==A) = room(group==B)
z =  -1.358
```

```
    Prob > |z| =   0.1745
```

解读 Interpretation:

- 以 A 组为第一样本，其秩和 $W = 21$，期望秩和 $\mu_W = 27.5$。
  Taking Group A as the first sample, its rank sum $W = 21$, expected rank sum $\mu_W = 27.5$.

- 调整后方差为 22.92，故 $\sigma_W = \sqrt{22.92} \approx 4.787$。
  The adjusted variance is 22.92, so $\sigma_W = \sqrt{22.92} \approx 4.787$.

- 标准化统计量 $z = (21 - 27.5)/4.787 \approx -1.358$。
  The standardized statistic $z = (21 - 27.5)/4.787 \approx -1.358$.

- 双侧 P 值 $= 0.1745$。
  Two-sided P-value $= 0.1745$.

- **结论：**在常规显著性水平（如 0.05）下，没有足够的证据拒绝原假设，即不能认为两组水疗中心的房间数分布存在系统性差异。
  Conclusion: At conventional significance levels (e.g., 0.05), there is not enough evidence to reject the null hypothesis. We cannot conclude that there is a systematic difference in the distribution of room numbers between the two groups of spas.

注意： Stata 16 及更高版本也会为小样本量报告精确 P 值。
Note: Stata 16 also reports the exact p-value in its output for small sample sizes.

2.9 处理“结”（Ties）

- 当观测值出现相同数值时，即产生“结”。
  Ties occur when observations have the same value.

- **标准做法 Standard Practice:** 将所有 tied 值赋予它们所占位置的平均秩。
  Assign all tied values the average of the ranks they occupy.

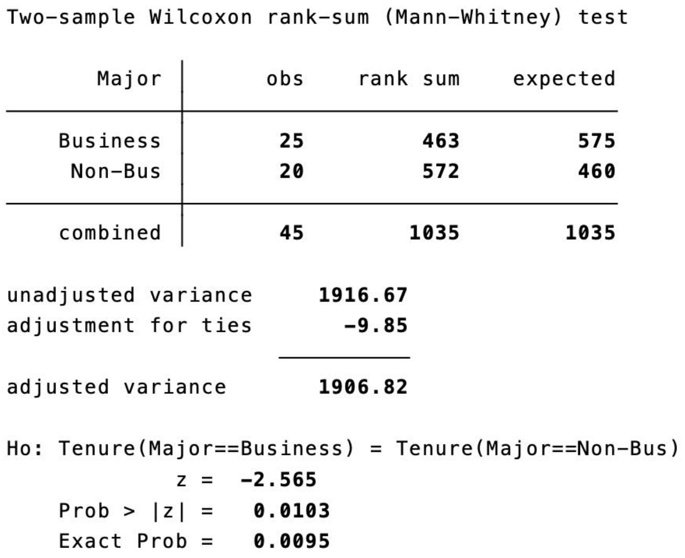
- **示例 Example:**

观测值 Observation	153	155	158	158	161	164
秩 Rank	1	2	3.5	3.5	5	6

- **对检验的影响 Impact on the Test:**
  - 如果数据中存在结，Wilcoxon 秩和统计量 W 的精确分布会发生变化，标准差 σ_W 必须进行调整。

    The exact distribution for the Wilcoxon rank sum W changes if the data contain ties and the standard deviation σ_W must be adjusted.
  - 在实践中，如果数据包含结，需要使用软件进行计算。统计软件会检测结，进行必要的调整，并使用正态近似。

    In practice, software is required if you want to use rank tests when the data contain tied values. Statistical software will detect ties, make the necessary adjustment, and switch to the Normal approximation.



```
Two-sample Wilcoxon rank-sum (Mann-Whitney) test

       Major |      obs    rank sum    expected
-------------+---------------------------------
    Business |       25         463         575
     Non-Bus |       20         572         460
-------------+---------------------------------
    combined |       45        1035        1035

unadjusted variance     1916.67
adjustment for ties       -9.85
                     ----------
adjusted variance       1906.82

Ho: Tenure(Major==Business) = Tenure(Major==Non-Bus)
             z =  -2.565
    Prob > |z| =   0.0103
    Exact Prob =   0.0095
```

2.10 历史注记 A Little Bit of History

历史背景

- **Frank Wilcoxon (1892-1965)** 出生在爱尔兰，父母是美国人。在从事过多种职业（包括商船船员、油井泵操作员、树木外科医生）后，他定居于化学领域，在罗格斯大学和康奈尔大学获得学位，并在多家公司工作。

  Frank Wilcoxon (1892-1965) was born in Ireland to American parents. After working in various occupations (including merchant seaman, oil-well pump attendant, and tree surgeon), he settled in chemistry, gaining degrees from Rutgers and Cornell and employment from various companies.

- 主要从事杀菌剂和杀虫剂的开发，Wilcoxon 于 1925 年开始对统计学产生兴趣，并对非参数方法做出了几项关键贡献。从工业界退休后，他在佛罗里达

州立大学教授统计学直至去世。

Working mainly on the development of fungicides and insecticides, Wilcoxon became interested in statistics in 1925 and made several key contributions to nonparametric methods. After retiring from industry, he taught statistics at Florida State until his death.

- 绝大多数文献以 Wilcoxon、Mann 和 Whitney 的名字命名这些检验。然而，这些检验在 20 世纪 40 年代末 50 年代初被其他几位研究者独立提出。除了 Wilcoxon、Mann 和 Whitney，功劳还应归于 Festinger (1946)、Whitfield (1947)、Haldane 和 Smith (1947) 以及 Van der Reyden (1952)。

  The great majority of the literature names these tests for Wilcoxon, Mann, and Whitney. However, they were independently developed by several other researchers in the late 1940s and early 1950s. In addition to Wilcoxon, Mann, and Whitney, credit is due to Festinger (1946), Whitfield (1947), Haldane and Smith (1947), and Van der Reyden (1952).

- Leon Festinger (1919-1989, 认知失调理论)、John Burdon Sanderson Haldane (1892-1964, 群体遗传学) 和 Cedric Austen Bardell Smith (1917-2002, 统计遗传学) 因其他工作而闻名，但关于 Whitfield 或 van der Reyden 的信息似乎知之甚少。

  Leon Festinger (1919-1989), John Burdon Sanderson Haldane (1892-1964), and Cedric Austen Bardell Smith (1917-2002) are well known for other work, but little seems to be known about Whitfield or van der Reyden.

- 详细研究请参阅 Berry, Mielke, and Johnston (2012)。

  For a detailed study, including information on these researchers, see Berry, Mielke, and Johnston (2012).

3 Wilcoxon 符号秩检验（配对样本）Wilcoxon Signed Rank Sum Test

3.1 基本概念与动机

- 适用于配对样本或单样本情形，对差值进行推断。

  The rank test for the matched pairs setting: inference on the differences.

- 示例 **Example:** 比较两种故事输入方法对儿童复述故事的影响：故事 1 仅被朗读，故事 2 在朗读的同时配有插图。

Compare storytelling with two input methods: story 1 was only read to them, and story 2 had been read but also illustrated with pictures.

儿童 Child	1	2	3	4	5
故事 2 得分 Story 2 Score	0.77	0.49	0.66	0.28	0.38
故事 1 得分 Story 1 Score	0.40	0.72	0.00	0.36	0.55
差值 Difference (Story 2 - Story 1)	0.37	-0.23	0.66	-0.08	-0.17

- 我们想知道插图是否改善了儿童复述故事。
  We wonder if illustrations improve how the children retell a story.
- **假设 Hypotheses:**
  - H_0：两种故事输入方法的得分分布相同。
    H_0: Scores have the same distribution for both stories.
  - H_1：故事 2 的得分系统地高于故事 1。
    H_1: Scores are systematically higher for Story 2.

3.2 检验步骤与统计量

1. 计算每对数据的差值 D_i。
   Compute the difference D_i for each pair.
2. 忽略差值的符号，取其绝对值 $|D_i|$。
   Ignore the signs and take the absolute values $|D_i|$.
3. 将所有绝对值从小到大排序，赋予秩次（1 到 n）。
   Rank all absolute differences from smallest to largest, assigning ranks from 1 to n.
4. 将对应原始差值为正的那些绝对值的秩次标出。
   Mark the ranks corresponding to positive differences.
5. 检验统计量 W^+ 为正差值的秩和。
   The test statistic W^+ is the sum of the ranks for the positive differences. (One could also use the sum of negative ranks, or the difference between the two.)

3.2.1 示例计算

故事示例计算

差值绝对值排序 Ranking Absolute Differences:

绝对值 Absolute value	0.08	0.17	0.23	**0.37**	**0.66**
秩 Rank	1	2	3	**4**	**5**
原始符号 Original Sign	-	-	-	+	+

正差值的秩和为： $W^+ = 4 + 5 = 9$。

The sum of ranks for positive differences is: $W^+ = 4 + 5 = 9$.

3.3 原假设下的行为与精确分布

- 如果配对内的不同处理不影响反应的分布（原假设），那么每个秩次为正或负的概率都是 1/2。

  If the distribution of the responses is not affected by the different treatments within pairs (the null hypothesis), every rank has a 50/50 chance of being positive.

- 通过枚举所有可能的正负号分配组合（共 2^n 种），计算每种组合下的 W^+，可以得出 W^+ 在原假设下的精确分布。

  The exact distribution of W^+ under H_0 can be derived by enumerating all 2^n possible assignments of positive/negative signs to the ranks and calculating W^+ for each.

- Wilcoxon 符号秩检验在秩和 W^+ 远离其均值时，拒绝“配对内无系统性差异”的假设。

  The Wilcoxon signed rank test rejects the hypothesis that there are no systematic differences within pairs when the rank sum W^+ is far from its mean.

3.4 正态近似

- 当样本量 n（配对数量）较大时，W^+ 的分布近似正态。

  The distribution of W^+ under H_0 becomes approximately Normal as the sample size n becomes large.

- W^+ 在原假设下的均值和标准差为：

  W^+ has mean and standard deviation under H_0:

$$\mu_{W^+} = \frac{n(n+1)}{4}$$

$$\sigma_{W^+} = \sqrt{\frac{n(n+1)(2n+1)}{24}}$$

其中 n 是差值非零的对子数（通常忽略差值为 0 的对子）。

where n is the number of pairs with non-zero differences (pairs with $D = 0$ are usually dropped).

- 标准化统计量：

  The standardized z statistic:

  $$z = \frac{W^+ - \mu_{W^+}}{\sigma_{W^+}}$$

- **故事示例计算：** $n = 5$, $\mu_{W^+} = 5 \times 6/4 = 7.5$, $\sigma_{W^+} = \sqrt{(5 \times 6 \times 11)/24} = \sqrt{330/24} = \sqrt{13.75} \approx 3.708$。

  Story example calculation: $n = 5$, $\mu_{W^+} = 5 \times 6/4 = 7.5$, $\sigma_{W^+} = \sqrt{(5 \times 6 \times 11)/24} = \sqrt{330/24} = \sqrt{13.75} \approx 3.708$.

  $$z = \frac{9 - 7.5}{3.708} \approx 0.404$$

  $$P(W^+ \geq 9) \approx P(z \geq 0.404) = 0.343$$

  这是一个单侧 P 值。该值较大，因此没有足够证据表明插图能系统性地提高得分。

  This is a one-sided P-value. The value is large, so there is not enough evidence that illustrations systematically improve scores.

3.5 处理“结”与零差值 Ties and Zero Differences

- **两种类型的“结” Two Types of Ties:**

  1. **绝对值之间的结 Ties among absolute differences:** 处理方法与秩和检验相同，赋予平均秩。

     Handled by assigning average ranks.

  2. **配对内的结（差值为零）Ties within a pair (difference of zero):** 差值为零既不取正也不取负。标准程序是简单地将这样的配对从样本中剔除。然而，过多的零差值会使结果偏向备择假设。为什么？因为如果有很多零差值被剔除，剩下的非零差值中正负号的分布可能更容易表现出系统性，从而增加拒绝 H_0 的机会。

     A tie within a pair creates a difference of zero. Because these are neither positive nor negative, the usual procedure simply drops such pairs from the sample. However, too many zeros would bias the result toward the alternative. Why? Because dropping many zero differences may leave a sample where the remaining signs are more easily imbalanced.

- 标准差 σ_{W^+} 必须针对结进行调整。软件会自动完成。
  The standard deviation σ_{W^+} must be adjusted for the ties. Software will do this.

3.6 软件输出示例 Software Output Example

汽车舒适度调查

背景：调查汽车舒适度，1= 最不舒适，5= 最舒适。比较欧洲车与美国车。
Background: Cars' comfort survey: 1=least comfortable, 5=most comfortable. Comparing European vs American cars.

```
Wilcoxon signed-rank test

sign |      obs   sum ranks    expected
-------------+---------------------------------
positive |       17          259         161
negative |        6           63         161
zero |        2           3           3
-------------+---------------------------------
all |       25         325         325

unadjusted variance     1381.25
adjustment for ties      -89.38
adjustment for zeros     -1.25
-------------
adjusted variance       1290.62

Ho: European = American
z =    2.728
Prob > |z| =    0.0064
Exact Prob = 0.0048
```

解读 Interpretation:

- 有效配对数为 $n = 25 - 2 = 23$（剔除了 2 个差值为零的配对）。
  The effective number of pairs is $n = 25 - 2 = 23$ (2 zero-difference pairs dropped).
- 正差值的秩和 $W^+ = 259$。
  The sum of ranks for positive differences $W^+ = 259$.

- 期望秩和 $\mu_{W^+} = 161$。
  Expected rank sum $\mu_{W^+} = 161$.

- 调整后方差为 1290.62，故 $\sigma_{W^+} \approx 35.93$。
  Adjusted variance is 1290.62, so $\sigma_{W^+} \approx 35.93$.

- $z = (259 - 161)/35.93 \approx 2.728$。
  $z = (259 - 161)/35.93 \approx 2.728$.

- 双侧 P 值 $= 0.0064$（精确 P 值 $= 0.0048$）。
  Two-sided P-value $= 0.0064$ (Exact P-value $= 0.0048$).

- **结论：**在 0.05 显著性水平下，拒绝原假设。有显著证据表明欧洲车与美国车的舒适度评分分布存在差异（从数据看，欧洲车评分可能更高）。
  Conclusion: At the 0.05 significance level, we reject the null hypothesis. There is significant evidence that the distributions of comfort ratings differ between European and American cars (the data suggest European cars may be rated higher).

注意： Stata 实际上使用 $W^+ - W^-$ 作为检验统计量，但这对检验结果没有影响。
Note: Stata actually uses $W^+ - W^-$ as the test statistic, but it is just numerically different and has no impact on the test outcome.

4 符号检验 Sign Test

4.1 基本概念

- 一种更简单的非参数检验，适用于单样本或配对数据场景。
  A simpler nonparametric test for one-sample or paired-data settings.

- 仅利用差异的**方向**（正号或负号），忽略其大小。
  Uses only the **direction** of differences (+ / -), ignoring magnitudes.

- **检验统计量 Test Statistic:** 正差异的个数（在剔除差值为 $D = 0$ 的结之后）。
  Number of positive differences, after discarding $D = 0$ ties.

- **假设 Hypotheses:**

$$H_0 : \Pr(D > 0) = \Pr(D < 0) = 0.5$$

$$H_1 : \Pr(D > 0) \neq 0.5 \quad (\text{双侧}) \quad \text{或} \quad \Pr(D > 0) > 0.5 \quad (\text{单侧})$$

- 在原假设下，统计量（正差异的个数 S^+）服从二项分布 Binomial$(n, 0.5)$，其中 n 是非零差异的对子数。

  Under H_0, the statistic (number of positive differences S^+) follows a Binomial$(n, 0.5)$ distribution, where n is the number of pairs with non-zero differences.

4.2 何时使用 When to Use

- **差异的大小不可靠或没有意义时**（例如，顺序数据）。

  When the magnitudes of differences are unreliable or not meaningful (e.g., ordinal data).

- **存在强烈异常值影响秩次时**。符号检验对异常值完全不敏感。

  In the presence of strong outliers affecting ranks. The sign test is completely insensitive to outliers.

- 作为 Wilcoxon 符号秩检验的一个更稳健但功效较低的替代方法。

  As a more robust but less powerful alternative to the Wilcoxon signed-rank test.

4.3 示例

符号检验示例

沿用前面的故事示例，有 5 个儿童，差值分别为：+0.37, -0.23, +0.66, -0.08, -0.17。

- 剔除差值为零的对子：本例中没有。
- 正差异的个数 $S^+ = 2$。
- $n = 5$。
- 原假设 $H_0 : p = 0.5$，其中 p 是出现正差异的概率。
- 备择假设 $H_1 : p > 0.5$（插图能提高得分）。
- P 值 $= P(S^+ \geq 2 \mid n = 5, p = 0.5) = 1 - P(S^+ \leq 1) = 1 - [P(S^+ = 0) + P(S^+ = 1)]$。
- 使用二项分布计算：$P(S^+ = 0) = C_5^0(0.5)^5 = 0.03125$，$P(S^+ = 1) = C_5^1(0.5)^5 = 0.15625$。
- 所以 P 值 $= 1 - (0.03125 + 0.15625) = 0.8125$。

这个 P 值非常大，远大于 0.05，因此没有证据支持备择假设。注意，相比 Wilcoxon 符号秩检验得到的 P 值（约 0.343），符号检验的 P 值更大，说明其功效更低，未

能捕捉到差值大小中包含的信息。

This P-value is very large, far greater than 0.05, so there is no evidence to support the alternative hypothesis. Note that compared to the P-value from the Wilcoxon signed-rank test (0.343), the sign test's P-value is larger, indicating its lower power, as it fails to capture information contained in the magnitudes of the differences.

5 应用、比较与总结 Applications, Comparisons and Summary

5.1 在顺序数据上的应用 Application on Ordinal Data

止痛药效果研究

背景：一家制药公司计划测试一种新的止痛药。随机选取 30 人，其中 15 人服用新止痛药，15 人服用阿司匹林。疗效评分如下：

Background: A pharmaceutical company is planning to test a new painkiller. 30 people were randomly selected, of whom 15 were given the new painkiller and 15 were given aspirin. Effectiveness rating:

- 5 = 药物极其有效。
  5 = The drug was extremely effective.
- 4 = 药物相当有效。
  4 = The drug was quite effective.
- 3 = 药物有些效果。
  3 = The drug was somewhat effective.
- 2 = 药物稍微有效。
  2 = The drug was slightly effective.
- 1 = 药物完全无效。
  1 = The drug was not at all effective.

Wilcoxon 秩和检验结果 Wilcoxon Rank Sum Test Results (Stata 输出):

```
Two-sample Wilcoxon rank-sum (Mann-Whitney) test

Drug |      obs    rank sum    expected
-------------+---------------------------------
```

```
Aspirin |      15       188.5       232.5
New |      15       276.5       232.5
-------------+---------------------------------
combined |      30         465         465

unadjusted variance      581.25
adjustment for ties      -28.45
-----------------
adjusted variance        552.80

Ho: PainLevel(Drug==Aspirin) = PainLevel(Drug==New)
z =  -1.871
Prob > |z| =   0.0613
Exact Prob = 0.0673
```

两样本 t 检验结果 Two-sample t Test Results (对比):

```
Two-sample t test with equal variances

Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
Aspirin |      15    2.933333    .3157254    1.222799     2.25617    3.610497
New |      15         3.8    .3265986    1.264911    3.099516    4.500484
---------+--------------------------------------------------------------------
combined |      30    3.366667    .2372415    1.299425    2.881453     3.85188
---------+--------------------------------------------------------------------
diff |           -.8666667    .4542568               -1.797169    .0638361

diff = mean(Aspirin) - mean(New)
t =  -1.9079
Ho: diff = 0                         degrees of freedom =       28

Ha: diff < 0                  Ha: diff != 0                Ha: diff > 0
Pr(T < t) = 0.0334        Pr(|T| > |t|) = 0.0667         Pr(T > t) = 0.9666
```

分析与讨论 Analysis and Discussion:

- 即使每组只有 15 个观测值，两样本 t 检验的 P 值（双侧 0.0667）与 Wilcoxon 检验的 P 值（0.0613 或精确 0.0673）非常相似。
  Even with only 15 observations in each group, the P-value for the two-sample

t (two-sided 0.0667) is very similar to that for the Wilcoxon test (0.0613 or exact 0.0673).

- t 检验将响应值 1 到 5 视为有意义的数字。然而，这是顺序数据，“极其有效”和“相当有效”之间的差异可能与其他等级之间的差异不相同。

  The t statistic treats the response values 1 through 5 as meaningful numbers. However, it is ordinal data and the difference between “extremely effective” and “quite effective” may not be the same as the difference between the other two levels.

- 一些从业者避免在没有完全有意义的测量尺度时使用 t 检验。

  Some practitioners avoid using t procedures when there is not a fully meaningful scale of measurement.

- 秩和检验只使用响应的顺序，而响应确实是从最无效到最有效排列的，因此秩检验在这里是合理的。

  The rank test uses only the order of the responses and the responses are arranged in order from least to most effective, so the rank test makes sense.

- **结论：**在这个例子中，两种检验得出的结论一致（在 0.05 水平上均不显著）。但考虑到数据的顺序性质，Wilcoxon 秩和检验可能是更合适的选择。

  Conclusion: In this example, both tests lead to the same conclusion (not significant at the 0.05 level). However, given the ordinal nature of the data, the Wilcoxon rank sum test might be the more appropriate choice.

5.2 秩和检验与 t 检验的比较 Comparing Rank Tests with t-tests

- **小样本与非正态分布 Small samples and non-normal distributions:**

  - 当样本量小且分布形状相同但非正态时，Wilcoxon 检验比两样本 t 检验更可靠。

    When our samples are small and show non-normal distributions of the same shape, the Wilcoxon test is more reliable than the two-sample t test.

- **适用范围 Scope of application:**

  - 基于秩的推断在很大程度上局限于简单的设置。正态性推断可以扩展到复杂的实验设计和多元回归方法，而非参数检验则不行。

    Inference based on ranks is largely restricted to simple settings. Normal inference extends to methods for use with complex experimental designs and multiple regression, but nonparametric tests do not.

- **顺序数据 Ordinal data:** 如上例所示，对于顺序数据，秩检验在概念上更合适。
  As shown in the previous example, for ordinal data, rank tests are conceptually more appropriate.

- **功效 Power:** 如果数据确实服从或近似服从正态分布，t 检验通常比对应的秩检验功效更高。如果数据严重偏离正态，秩检验可能功效更高。
  If the data are indeed Normal or approximately Normal, the t-test is generally more powerful than the corresponding rank test. If the data are severely non-Normal, the rank test may be more powerful.

5.3 选择检验的指导原则 Guidelines for Choosing a Test

考虑因素 Consideration	**优先考虑参数检验（如 t 检验）Prefer Parametric Test (e.g., t-test)**	**优先考虑非参数检验（如秩检验）Prefer Nonparametric Test (e.g., Rank Test)**
样本量 Sample Size	大样本（如每组 $n > 30$），可依赖中心极限定理。Large sample (e.g., each group $n > 30$), can rely on CLT.	小样本，且无法确认正态性。Small sample, and cannot confirm normality.
数据分布 Data Distribution	数据显示近似正态（可通过 Q-Q 图、直方图等检查）。Data appear approximately normal (checked by Q-Q plot, histogram, etc.).	数据明显非正态（严重偏态、多峰、存在极端异常值）。Data are clearly non-normal (severely skewed, multi-modal, extreme outliers).
数据类型 Data Type	连续数据，且测量尺度是等距或比率的。Continuous data with interval or ratio scale.	顺序数据（Ordinal Data），或连续数据但测量尺度不精确。Ordinal data, or continuous data with imprecise measurement scale.
研究设计 Research Design	需要扩展到更复杂的模型（如 ANOVA，回归）。Need to extend to more complex models (e.g., ANOVA, regression).	仅需进行简单的组间比较。Only simple group comparisons are needed.
异常值 Outliers	数据中没有或有少量可解释的异常值。No or few explainable outliers.	存在多个极端异常值，且无法合理解释或剔除。Presence of multiple extreme outliers that cannot be reasonably explained or removed.
方差齐性 Homogeneity of Variance	对于两样本 t 检验，如果方差异质，可以使用 Welch 校正。对于秩和检验，不要求方差齐性，但要求分布形状相同才能解释为位置检验.	

表 2: 检验方法选择指南 Guidelines for Test Selection

本章核心公式与概念总结 Summary of Key Formulas and Concepts

Wilcoxon 秩和检验 (Mann-Whitney U 检验)

- **适用场景：**两个独立样本的位置比较。

  Setting: Comparison of two independent samples.

- **原假设 H_0：**两个总体的分布相同。

  H_0: The two populations have identical distributions.

- **备择假设 H_1：**一个总体的值系统地大于另一个（或分布不同）。

  H_1: One population's values are systematically larger than the other's (or distributions differ).

- **检验统计量：** W = 第一样本的秩和。

  Test Statistic: W = Sum of ranks in the first sample.

- **均值和标准差 (无结，H_0 下)：**

  $$\mu_W = \frac{n_1(N+1)}{2}, \quad \sigma_W = \sqrt{\frac{n_1 n_2 (N+1)}{12}}, \quad N = n_1 + n_2$$

- **正态近似：**

  $$z = \frac{W - \mu_W}{\sigma_W} \sim N(0,1) \quad \text{(大样本时)}$$

Wilcoxon 符号秩检验

- **适用场景：**配对样本或单样本的位置比较。

  Setting: Paired samples or one sample.

- **原假设 H_0：**差值总体的中位数为 0，且分布对称。

  H_0: The median of the difference population is zero and the distribution is symmetric.

- **检验统计量：** W^+ = 正差值的秩和。

  Test Statistic: W^+ = Sum of ranks of positive differences.

- **均值和标准差 (无结，H_0 下，忽略零差值的对子数 n)：**

  $$\mu_{W^+} = \frac{n(n+1)}{4}, \quad \sigma_{W^+} = \sqrt{\frac{n(n+1)(2n+1)}{24}}$$

- **正态近似：**

  $$z = \frac{W^+ - \mu_{W^+}}{\sigma_{W^+}} \sim N(0,1) \quad \text{(大样本时)}$$

符号检验

- **适用场景：**配对样本或单样本，仅关心方向。
  Setting: Paired samples or one sample, only direction matters.

- **原假设 H_0：**正差值的概率 $p = 0.5$。
  H_0: Probability of a positive difference $p = 0.5$.

- **检验统计量：** S^+ = 正差值的个数（剔除 $D = 0$）。
  Test Statistic: S^+ = Number of positive differences (discarding $D = 0$).

- **精确分布：** $S^+ \sim \text{Binomial}(n, 0.5)$，其中 n 是非零差值的对子数。
  Exact Distribution: $S^+ \sim \text{Binomial}(n, 0.5)$, where n is the number of non-zero difference pairs.

关于“结”（Ties）的处理

- **排序时：**赋予 tied values 平均秩。
  When ranking: Assign average ranks to tied values.

- **方差调整：**对于 Wilcoxon 检验，标准差 σ_W 和 σ_{W^+} 需要针对结进行调整。公式较复杂，通常由软件完成。
  Variance adjustment: For Wilcoxon tests, the standard deviations σ_W and σ_{W^+} need to be adjusted for ties. Formulas are complex; typically done by software.

- **零差值：**在配对检验中，通常将差值为零的配对剔除。
  Zero differences: In paired tests, pairs with zero difference are usually dropped.

本章核心要点 Core Takeaways

- **非参数检验的核心优势在于其稳健性：**它们不依赖于总体分布的具体形式（如正态性），对异常值不敏感，适用于顺序数据和小样本。
  The core strength of nonparametric tests lies in their robustness: They do not rely on specific population distributional forms (e.g., normality), are insensitive to outliers, and are suitable for ordinal data and small samples.

- **正确选择检验：**
  - 两个独立样本 → Wilcoxon 秩和检验 (Mann-Whitney U 检验)。
    Two independent samples → Wilcoxon Rank Sum Test (Mann-Whitney U test).
  - 配对样本 → Wilcoxon 符号秩检验或符号检验。
    Paired samples → Wilcoxon Signed Rank Test or Sign Test.

  - 符号检验更简单稳健，但功效最低；Wilcoxon 符号秩检验利用了大小信息，功效更高。

    The Sign Test is simpler and more robust but least powerful; the Wilcoxon Signed Rank Test uses magnitude information and is more powerful.

- **理解假设：**
  - 秩和检验的原假设通常是“两个分布相同”，而不仅仅是中位数相等。只有在附加“分布形状相同”的假设下，才能解释为位置（中位数）检验。

    The null hypothesis of the rank-sum test is typically ”two distributions are identical”, not just equal medians. Only under the additional assumption of ”same distribution shape” can it be interpreted as a test of location (median).
  - 配对检验假设差值分布对称（对于 Wilcoxon 符号秩检验）。

    Paired tests assume the distribution of differences is symmetric (for the Wilcoxon signed-rank test).

- **软件实现：**现代统计软件（如 Stata, R, SPSS）可以方便地执行这些检验，并自动处理结的调整和提供精确 P 值。

  Software Implementation: Modern statistical software (e.g., Stata, R, SPSS) can easily perform these tests, automatically handle tie adjustments, and provide exact P-values.

- **权衡：**非参数检验牺牲了一些信息（数值大小）以换取稳健性，因此在数据满足参数检验假设时，其统计功效可能较低。在数据分析中，应根据数据特征和研究问题谨慎选择检验方法。

  Trade-off: Nonparametric tests sacrifice some information (magnitude) for robustness, thus they may have lower statistical power when the data meet the assumptions of parametric tests. In data analysis, the choice of test should be made carefully based on data characteristics and research questions.