补充

1. χ^2 分布 (卡方分布)

核心直觉：它是“误差的平方和”。用于描述方差的波动。

定义与构造

设 $X_1, X_2, \ldots, X_n$ 相互独立，且都服从标准正态分布 $N(0,1)$，那么它们的平方和：$Y = \sum_{i=1}^{n} X_i^2$

服从自由度为 n 的 χ^2 分布，记为 $Y \sim \chi^2(n)$。

- **自由度 (n)**：参与求和的独立标准正态变量的个数。

关键性质

- **取值范围**：$(0, +\infty)$（因为是平方和，肯定是正的）。
- **期望与方差**：
  - $E(Y) = n$
  - $Var(Y) = 2n$
- **可加性**：如果 $Y_1 \sim \chi^2(n_1), Y_2 \sim \chi^2(n_2)$ 且独立，那么 $Y_1 + Y_2 \sim \chi^2(n_1 + n_2)$。
- **大样本**：当 n 很大时，$\chi^2(n)$ 趋近于正态分布。

主要用途

- **检验方差**：推断正态总体的方差（如 σ^2 的置信区间）。
- **拟合优度检验**：判断数据是否符合某种分布。
- **列联表分析**：独立性检验（卡方检验）。

---

2. t 分布 (学生氏 t 分布)

核心直觉：它是“小样本下的正态分布”。当我们不知道总体标准差 σ，只能用样本标准差 S 代替时，就不再是正态分布，而是 t 分布。

定义与构造

设 $X \sim N(0,1)$，$Y \sim \chi^2(n)$，且 X 与 Y 相互独立，则变量：$T = \frac{X}{\sqrt{Y/n}}$

服从自由度为 n 的 t 分布，记为 $T \sim t(n)$。

- **口诀**：“上边是标准正态，下边是根号下的卡方除以自由度”。

关键性质

- **形态**：钟形曲线，关于 $x = 0$ 对称。
- **与正态分布的区别**：**“厚尾”**。t 分布的尾部比标准正态分布略高，中间略低。这意味着小样本情况下，极端值出现的概率比正态分布预测的要大。
- **极限**：当自由度 $n \to \infty$ 时，$t(n)$ 收敛于标准正态分布 $N(0,1)$。

主要用途

- **均值检验**：在方差未知的情况下，检验正态总体的均值（t-test）。
- **回归系数检验**：线性回归中检验系数是否显著不为0。

3. F 分布

核心直觉：它是"两个方差的比值"。用于比较两个群体的波动程度是否一致。

定义与构造

设 $U \sim \chi^2(n_1)$，$V \sim \chi^2(n_2)$，且 U 与 V 相互独立，则变量：$F = \frac{U/n_1}{V/n_2}$

服从自由度为 (n_1, n_2) 的 F 分布，记为 $F \sim F(n_1, n_2)$。

- **口诀**："两个卡方分布，分别除以自己的自由度，然后再相除"。

关键性质

- **取值范围**：$(0, +\infty)$。
- **倒数性质**：如果 $X \sim F(n_1, n_2)$，那么 $\frac{1}{X} \sim F(n_2, n_1)$。这在查表和计算置信区间时非常有用。
- **非对称**：偏态分布，向右拖尾。

主要用途

- **方差分析 (ANOVA)**：比较三个或更多组的均值是否存在显著差异（本质是在比组间方差和组内方差）。
- **等方差检验**：检验两个正态总体的方差是否相等（F-test）。

4.泊松分布

若随机变量 X 服从参数为 λ 的泊松分布，记为 $X \sim P(\lambda)$ 或 $X \sim \mathrm{Poi}(\lambda)$。

其概率质量函数为：$P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}, \quad k = 0, 1, 2, \ldots$

期望:$E[X] = \lambda$

方差 :$\mathrm{Var}(X) = \lambda$

5. 指数分布

若随机变量 X 服从参数为 λ ($\lambda > 0$) 的指数分布，记为 $X \sim Exp(\lambda)$。

- 概率密度函数 (PDF):

  描述了在某一点取值的相对可能性。可以看到随着 x 增加，概率密度呈指数衰减。

  $f(x; \lambda) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0 \\ 0 & x < 0 \end{cases}$

  这里 λ 被称为**率参数 (Rate Parameter)**，表示单位时间内事件发生的平均次数。

- 累积分布函数 (CDF):

  描述了事件在时间 x 之前发生的概率。$F(x; \lambda) = \begin{cases} 1 - e^{-\lambda x} & x \geq 0 \\ 0 & x < 0 \end{cases}$

期望与方差 (Mean & Variance)

指数分布的均值和方差都仅依赖于参数 λ。

- **期望 (Mean):** $E[X] = \frac{1}{\lambda}$
  - *注意：如果 λ 是故障率（每小时坏几次），那么 $1/\lambda$ 就是平均故障间隔时间。*
- **方差 (Variance):** $Var(X) = \frac{1}{\lambda^2}$

6. 几何分布

概率质量函数 (PMF)

对于定义 A（第 k 次才成功）：$P(X=k)=(1-p)^{k-1}p$

累积分布函数 (CDF)

$P(X\leq k)=1-(1-p)^k$

无记忆性 —— **最独特的性质**

期望值：$E[X]=\frac{1}{p}$

方差：$Var(X)=\frac{1-p}{p^2}$

7. 二项分布

概率质量函数 (PMF)：

$P(X=k)=\binom{n}{k}p^k(1-p)^{n-k},\quad k=0,1,\ldots,n$

性质：

- **期望**：$E(X)=np$
- **方差**：$Var(X)=np(1-p)$

8. 均匀分布

概率密度函数 (PDF)：

$$f(x)=\begin{cases}\frac{1}{b-a}, & a\leq x\leq b\\ 0, & \text{其他}\end{cases}$$

性质：

- **期望**：$E(X)=\frac{a+b}{2}$ （区间的中点）
- **方差**：$Var(X)=\frac{(b-a)^2}{12}$

9. 分位数

(1) 对标准正态分布 N(0,1)，用 zp 表示其分布的 p 分位数，即$P\{X\leqslant z_p\}=p$

由于标准正态分布的概率密度函数图形关于y 轴对称，因此有$-z_p=z_{1-p}$。

(2) 对自由度为 n 的 χ^2 分布 $\chi^2(n)$，用 $\chi_p^2(n)$ 表示其分布的 p 分位数，即$P\{\chi^2\leqslant\chi_p^2(n)\}=p$

(3)对自由度为 n 的 t 分布 $t(n)$，用 $t_p(n)$ 表示其分布的 p 分位数，即$P\{T\leqslant t_p(n)\}=p$

由于 t 分布的概率密度函数图形关于 y 轴对称，因此有 $-t_p(n)=t_{1-p}(n)$。

(4) 对自由度为 n_1,n_2 的 F 分布 $F(n_1,n_2)$，用 $F_p(n_1,n_2)$ 表示其分布的 p 分位数，即 $P\{F\leqslant F_p(n_1,n_2)\}=p$，且有$F_p(n_2,n_1)=\frac{1}{F_{1-p}(n_1,n_2)}$

第一章

频率估计概率

这是概率论中最基础的思想之一，也是“频率学派”统计学的核心基石。

- **定义**：在大量重复进行的独立试验中，如果某个事件 A 发生的次数为 m，试验总次数为 n，那么频率 $f_n(A) = \frac{m}{n}$。当 n 趋于无穷大时，频率会稳定在某个常数附近，我们用这个频率值作为该事件概率 $P(A)$ 的估计值。
- **公式**：$\hat{p} = \frac{m}{n}$

矩估计法

矩估计法的基本方程组就是令：

$$总体原点矩 = 样本原点矩$$

$$E[X^k] = \frac{1}{n}\sum_{i=1}^{n} X_i^k$$

一阶矩估计

一阶矩即 $k = 1$ 时的矩，也就是我们最熟悉的**期望（均值）**。

- **总体一阶矩：** $E[X] = \mu$ (理论上的均值)
- **样本一阶矩：** $A_1 = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ (样本的平均值)

二阶矩估计

二阶矩即 $k = 2$ 时的矩，描述的是数据的平方的期望。它通常与**方差**密切相关，因为 $\mathrm{Var}(X) = E[X^2] - (E[X])^2$。

- **总体二阶矩：** $E[X^2]$
- **样本二阶矩：** $A_2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2$ (样本平方的平均值)

充分统计量

定理： 设总体分布族为 $\{P_\theta, \theta \in \Theta\}$, $T(X)$ 是充分的，当且仅当存在一个定义在 $I \times \Theta$ 上的函数 $g(t, \theta)$ 及定义在 R^n 上的函数 $h(x)$ 使得

$$p(x, \theta) = g(T(x), \theta)h(x)$$

对所有的 $x \in R^n$ 都成立，其中 I 是 $T(x)$ 的值域， $p(x, \theta)$ 是样本的联合概率密度函数或分布率。$T(x)$ 为充分统计量。

$\bar{y}$和$\sum(y_i - \bar{y})^2$

样本均值 $\bar{y}$ 与 样本离差平方和 $\sum(y_i - \bar{y})^2$ 相互独立。

极大似然估计的不变性原理

如果 $\hat{\theta}$ 是参数 θ 的极大似然估计，而 $g(\theta)$ 是 θ 的一个函数，那么 $g(\hat{\theta})$ 就是 $g(\theta)$ 的极大似然估计。

均方误差 (MSE)

$$MSE(\hat{\theta}) = E[(\hat{\theta} - \theta)^2] = Var(\hat{\theta}) + [Bias(\hat{\theta})]^2$$

充分完备统计量

设 $x_1, x_2, \cdots, x_n$ 是来自总体 $\{P_\theta : \theta \in \Theta\}$ 的简单样本，总体的密度函数（或分布列）为 $p(x; \theta)$，且样本 $x_1, x_2, \cdots, x_n$ 的联合密度函数（或联合分布列）可分解为

$$p(x_1, x_2, \cdots, x_n; \theta) = c(\theta)h(x_1, x_2, \cdots, x_n)\exp\{\sum_{k=1}^{m} w_k(\theta)T_k(x_1, x_2, \cdots, x_n)\}$$

其中 $h(x_1,x_2,\cdots,x_n)$ 仅是 $x_1,x_2,\cdots,x_n$ 的函数，$w=w(\theta)=(w_1(\theta),\cdots,w_m(\theta))$ 是定义在 m 维参数空间 Θ 上取值于 $\Lambda\subset\mathbf{R}^m$ 的向量函数，$c(\theta)$ 仅是 θ 的函数。如果 $w(\theta)$ 值域 Λ 包含内点，则 m 维统计量$T(x_1,x_2,\cdots,x_n)=(T_1(x_1,x_2,\cdots,x_n),T_2(x_1,x_2,\cdots,x_n),\cdots,T_m(x_1,x_2,\cdots,x_n))$ 是完全充分的。

UMVUE

设 $S(x)$ 是完全充分统计量，$\varphi(x)$ 是 $q(\theta)$ 的方差有限的无偏估计，即 $\varphi(x)\in U_q$，则

$$T(x)=E_\theta(\varphi(x)\mid S(x))$$

是 $q(\theta)$ **唯一的**一致最小方差无偏估计（UMVUE）。

- 其一，若能获得 $q(\theta)$ 的无偏估计 $\varphi(x)$，则 $\varphi(x)$ 关于 $S(x)$ 的条件数学期望 $T(x)=E_\theta(\varphi(x)\mid S(x))$ 就是 $q(\theta)$ 的一致最小方差无偏估计；
- 其二，由于 $q(\theta)$ 的一致最小方差无偏估计 $T(x)$ 一定是完全充分统计量 $S(x)$ 的函数，所以若能获得完全充分统计量 $S(x)$ 的函数 $h(S(x))$，并将其无偏化，就可获得 $q(\theta)$ 的一致最小方差无偏估计。

Fisher 信息量

$$I(\theta)=E\left[\left(\frac{\partial}{\partial\theta}\ln f(X;\theta)\right)^2\right]=-E\left[\frac{\partial^2}{\partial\theta^2}\ln f(X;\theta)\right]$$

有效估计

定义 2.3.2 设分布族 $\{P_\theta:\theta\in\Theta\}$ 是 Cramer-Rao 正则族，$q(\theta)$ 是可估参数，若存在某个无偏估计 $\hat{q}\in U_q$，对所有的 $\theta\in\Theta$，有$\mathrm{Var}_\theta(\hat{q})=\frac{[q'(\theta)]^2}{nI(\theta)}$，则称 $\hat{q}$ 为参数 $q(\theta)$ 的有效估计。

定义 2.3.3 对可估参数 $q(\theta)$ 的任一无偏估计 $T\in U_q$，令$e(T,q(\theta))=\frac{[q'(\theta)]^2}{nI(\theta)}\Big/\mathrm{Var}_\theta(T)$，则称 $e(T,q(\theta))$ 为使用 T 估计 $q(\theta)$ 的有效率。

渐近无偏估计

设 $\hat{\theta}_n$ 是基于 n 个样本对参数 θ 的估计量。如果满足以下极限条件：$\lim_{n\to\infty}E[\hat{\theta}_n]=\theta$

或者等价地写成偏差（Bias）趋于0：$\lim_{n\to\infty}Bias(\hat{\theta}_n)=\lim_{n\to\infty}(E[\hat{\theta}_n]-\theta)=0$，则称 $\hat{\theta}_n$ 是 θ 的渐近无偏估计。

Cramér-Rao 不等式

对于单参数 λ 的函数 $g(\lambda)$，如果 T 是它的无偏估计量，那么 T 的方差一定满足以下不等式：

$Var(T)\geq\frac{[g'(\lambda)]^2}{I_n(\lambda)}$

其中：

- **分子** $[g'(\lambda)]^2$ 是你要估计的函数对参数求导后的平方。
- **分母** $I_n(\lambda)$ 是样本的 **Fisher 信息量 (Fisher Information)**。

相合估计

如果 $\hat{q}_n(X)$ 是参数 $q(\theta)$ **相合**估计，且函数 $h(y)$ 在 $y=q(\theta)$ 处连续，则 $h(\hat{q}_n)$ 是 $h(q(\theta))$ 的**相合**估计。

$\lim_{n\to\infty}E(\hat{\theta}_n)=\theta$ 且 $\lim_{n\to\infty}D(\hat{\theta}_n)=0$

*(或者写成：$\hat{\theta}_n$ **的均方误差趋于0**，即$\lim_{n\to\infty} E[(\hat{\theta}_n - \theta)^2] = 0$)*

第二章

势和势函数

☑ 势 (Power)：正确拒绝 H_0 的能力

> 当 H_0 不成立时，我们 **正确拒绝** H_0 的概率，叫做检验的势或功效。

- ◇ 定义：$\gamma(\theta) = P_\theta\{x \in W\} = 1 - \beta(\theta), \quad \theta \in \Theta_1$
- ☑ 所以：势 = 1 - 第二类错误概率

☑ 势函数 (功效函数)

> 把检验的"势"看成关于参数 θ 的函数，就是 **势函数** (也叫功效函数)：

$$g(\theta) = P_\theta\{x \in W\} = E_\theta[\varphi(x)], \quad \theta \in \Theta$$

其中：

- $g(\theta)$ 表示当真实参数为 θ 时，拒绝 H_0 的概率。
- 当 $\theta \in \Theta_0$ 时，$g(\theta) = \alpha(\theta)$ （第一类错误概率）
- 当 $\theta \in \Theta_1$ 时，$g(\theta) = \gamma(\theta)$ （势）

两类错误

☑ 1. 第一类错误（弃真错误）

> 当 H_0 实际上成立，但我们却错误地拒绝了它。

- ◇ 定义：$\alpha(\theta) = P_\theta\{x \in W\}, \quad \theta \in \Theta_0$
- $\alpha(\theta)$ **是当参数取** $\theta \in \Theta_0$ **（即** H_0 **成立）时，样本落入拒绝域的概率。**
- **也就是犯第一类错误的概率。**

☑ 2. 第二类错误（取伪错误）

> 当 H_0 实际上不成立（即 H_1 为真），但我们却错误地接受了 H_0。

- ◇ 定义：$\beta(\theta) = P_\theta\{x \notin W\} = 1 - P_\theta\{x \in W\}, \quad \theta \in \Theta_1$
- 即当 H_1 成立时，样本落在接受域的概率。
- 也就是犯第二类错误的概率。

Neyman-Pearson 假设检验原理

为了处理这个矛盾（指无法同时降低两类错误），统计学家 Neyman 和 Pearson 提出了一个普遍做法：

> 📌 **固定第一类错误的概率不超过某个小值** α**，然后在所有满足这一条件的检验中，寻找势最大（即第二类错误最小）的那个检验。**

一致最优检验

设总体分布族为 $\{p(x;\theta) : \theta \in \Theta\}$**，考虑假设检验问题，** $H_0 : \theta \in \Theta_0, \quad H_1 : \theta \in \Theta_1$**，将水平为** α **的所有检验函数的集合记为** $\Phi_\alpha = \{\varphi(x) : \sup_{\theta\in\Theta_0} E_\theta\varphi(x) \le \alpha\}$**，定义最优势检验如下。**

对假设检验问题，若存在水平为 α 的检验函数 $\varphi^* \in \Phi_\alpha$，使得对任一水平为 α 的检验函数 $\varphi \in \Phi_\alpha$ 有不等式$E_\theta(\varphi^*(x)) \geq E_\theta(\varphi(x))$，对所有的 $\theta \in \Theta_1$ 都成立，则称 $\varphi^*(x)$ 是水平为 α 的一致最优势检验(Uniformly Most Powerful Test)，简记为 UMPT。

一致最优无偏检验

设 $\varphi(x)$ 是假设检验问题 $H_0: \theta \in \Theta_0$，　$H_1: \theta \in \Theta_1$ 的检验函数，若其势函数 $g_\varphi(\theta) = E_\theta(\varphi(x))$ 满足

$$\begin{cases} g_\varphi(\theta) \leq \alpha, & \theta \in \Theta_0 \\ g_\varphi(\theta) \geq \alpha, & \theta \in \Theta_1 \end{cases}$$

则称 $\varphi(x)$ 是水平为 α 的无偏检验，显然，水平为 α 的一致最优势检验一定是无偏检验。

Karlin-Rubin 定理的方向判定口诀

单参数指数族分布，定理要求概率密度函数 $p(x,\theta)$ 必须能写成这种形式：
$p(x,\theta) = d(\theta)h(x)\exp\{c(\theta)T(x)\}$

- $T(x)$：这是我们提取出来的**统计量**（充分统计量）。
- $c(\theta)$：这是关于参数 θ 的函数。

总结出一个简单的"正负得负"规律，用来迅速判断拒绝域是 $T > C$ 还是 $T < C$。

如果这是离散分布题目：你必须写出 γ，因为边界概率 $P(T = C) \neq 0$。

因为这是连续分布题目：边界概率 $P(T = C) = 0$，等号是否取、γ 是多少都不影响积分结果。为了简洁，直接写 $\leq C$ 是标准做法。

1. **函数 $c(\theta)$ 的单调性**：
   - 单调递增 (↗)：记为 **正 (+)**
   - 单调递减 (↘)：记为 **负 (-)**
2. **备择假设 H_1 的方向**：
   - $H_1: \theta > \theta_0$ (右边)：记为 **正 (+)**
   - $H_1: \theta < \theta_0$ (左边)：记为 **负 (-)**

判定法则：将这两个符号相乘：

- **正正得正 (+)** → 拒绝域为大号 $T(x) > C$
- **负负得正 (+)** → 拒绝域为大号 $T(x) > C$
- **正负得负 (-)** → 拒绝域为小号 $T(x) < C$
- **负正得负 (-)** → 拒绝域为小号 $T(x) < C$

第三章

贝叶斯统计

记作 $p(x|\theta)$。意味着 θ 是一个**随机变量**。这里的 p 代表的是在 θ 取某个特定值时的**条件概率密度**。

贝叶斯公式

连续形式：$\pi(\theta|x) = \frac{p(x|\theta)\pi(\theta)}{\int_\Theta p(x|\theta)\pi(\theta)d\theta}$

离散形式：$P(\theta_i|X) = \frac{P(X|\theta_i)P(\theta_i)}{\sum P(X|\theta_j)P(\theta_j)}$

风险最小的那个"最佳估计值" θ^*，正好就是**后验分布的期望值（均值）** $\theta^* = E(\theta|x) = \int \theta p(\theta|x)d\theta$

第四章

性质

$\boldsymbol{\mu}$ 是一个 p 维向量，称为均值向量。当 $\boldsymbol{A}$ 和 $\boldsymbol{B}$ 为常数矩阵时，由定义可立即推出如下性质

(1) $E(\boldsymbol{AX}) = \boldsymbol{A}E(\boldsymbol{X})$

(2) $E(\boldsymbol{AXB}) = \boldsymbol{A}E(\boldsymbol{X})\boldsymbol{B}$

(3) $D(\boldsymbol{AX}) = \boldsymbol{A}D(\boldsymbol{X})\boldsymbol{A}^T = \boldsymbol{A\Sigma A}^T$

(4) $cov(\boldsymbol{AX}, \boldsymbol{BY}) = \boldsymbol{A}cov(\boldsymbol{X}, \boldsymbol{Y})\boldsymbol{B}^T$

(5) 设 $\boldsymbol{X}$ 为 p 维随机向量，期望和协方差存在，记 $\boldsymbol{\mu} = E(\boldsymbol{X}), \boldsymbol{\Sigma} = D(\boldsymbol{X})$，$\boldsymbol{A}$ 为 $p \times p$ 常数阵，则

$E(\boldsymbol{X}^T\boldsymbol{AX}) = tr(\boldsymbol{A\Sigma}) + \boldsymbol{\mu}^T\boldsymbol{A\mu}$

随机向量 $\boldsymbol{X} = (X_1, X_2, \cdots, X_p)^T$，其协方差阵 $\boldsymbol{\Sigma}$ 都是对称阵，同时总是半正定的。大多数情形下是正定的。

多元正态分布

定义4.2.3 若对任何非零向量 $\boldsymbol{a} \in R^p$， $\boldsymbol{X}$ 的线性组合$\boldsymbol{a}^T\boldsymbol{X}$ 服从一元正态分布 $N(\boldsymbol{a}^T\boldsymbol{\mu}, \boldsymbol{a}^T\boldsymbol{V}\boldsymbol{a})$，则称 $\boldsymbol{X}$ 服从 p 元正态分布 $N_p(\boldsymbol{\mu}, \boldsymbol{V})$。

多元正态条件分布

设 X~Np(μ,V),V>0 (正定)。我们将 X 分块为 X1 和 X2。

1. 给定 X2 时 X1 的条件分布： 服从$N_k(\boldsymbol{\mu}_{1\cdot2}, \boldsymbol{V}_{11\cdot2})$，其中：

- **条件均值：** $\boldsymbol{\mu}_{1\cdot2} = \boldsymbol{\mu}_1 + \boldsymbol{V}_{12}\boldsymbol{V}_{22}^{-1}(\boldsymbol{X}_2 - \boldsymbol{\mu}_2)$
- **条件协方差 (舒尔补)：** $\boldsymbol{V}_{11\cdot2} = \boldsymbol{V}_{11} - \boldsymbol{V}_{12}\boldsymbol{V}_{22}^{-1}\boldsymbol{V}_{21}$

2. 给定 X1 时 X2 的条件分布： 服从$N_{p-k}(\boldsymbol{\mu}_{2\cdot1}, \boldsymbol{V}_{22\cdot1})$，其中：

- **条件均值：** $\boldsymbol{\mu}_{2\cdot1} = \boldsymbol{\mu}_2 + \boldsymbol{V}_{21}\boldsymbol{V}_{11}^{-1}(\boldsymbol{X}_1 - \boldsymbol{\mu}_1)$
- **条件协方差：** $\boldsymbol{V}_{22\cdot1} = \boldsymbol{V}_{22} - \boldsymbol{V}_{21}\boldsymbol{V}_{11}^{-1}\boldsymbol{V}_{12}$

多元正态分布的通用密度函数公式

$f(\mathbf{x}) = \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}}\exp\left\{-\frac{1}{2}(\mathbf{x}-\mu)^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\mu)\right\}$

第五章

符号秩和检验

设有两个总体 $F(x)$ 与 $G(y)$，检验问题为 $H_0: F(x) = G(x), H_1: F(x) \neq G(x)$。如今获得成对数据 $(x_i, y_i), i = 1, 2, \cdots, n$，令 $z_i = x_i - y_i$，记 R_i 为 $|z_i|$ 在 $|z_1|, |z_2|, \cdots, |z_n|$ 中的秩，V_i 定义为

$$V_i = \begin{cases} 1, & z_i > 0 \\ 0, & z_i \leq 0 \end{cases} \quad i = 1, 2, \cdots, n$$

符号秩和检验用的统计量是 $W^+ = \sum_{i=1}^n V_i R_i$，这一统计量实质上是 $x_i > y_i$ 的观测值的差的绝对值 $|z_i|$ 的秩和。

W^+ 是所有满足 $x_i > y_i$ 的那些差值的绝对值所对应的秩之和。

◇ **构造差值与秩**

令：$z_i = x_i - y_i \quad (i = 1, 2, \ldots, n)$

这是每一对中的差值。

然后定义：

- $|z_i|$ 是差值的绝对值
- 将所有 $|z_i|$ 从小到大排序，得到它们的秩 (rank)

记：

- R_i：$|z_i|$ 在所有 $|z_1|, |z_2|, \ldots, |z_n|$ 中的秩（按升序排列）
- 注意：如果出现相等的 $|z_i|$，则使用平均秩处理（如并列第3和第4，则都取秩为3.5）

◇ **定义符号变量 V_i**

为了判断方向，引入符号变量：

$$V_i = \begin{cases} 1, & \text{若 } z_i > 0 \Rightarrow x_i > y_i \\ 0, & \text{若 } z_i \leq 0 \Rightarrow x_i \leq y_i \end{cases}$$

☑ 这个 V_i 表示第 i 对中 x_i 是否大于 y_i

◇ **核心统计量：W^+**

定义符号秩和检验的统计量为：$W^+ = \sum_{i=1}^{n} V_i R_i$

即W^+ 是所有满足 $x_i > y_i$ 的那些差值的绝对值所对应的秩之和。

在 H_0 为真时，$x_i > y_i$ 与 $x_i < y_i$ 出现的可能性应该是相同的，因而在 H_0 为真时 W^+ 不应过大，也不应过小，从而拒绝域的合理形式为：$\{W^+ \leq d \text{或} W^+ \geq c\}$

在小样本场合 $(n \leq 20)$，后面附表给出了 $P(W^+ \geq c) \leq \alpha$ 的临界值 c，$d = \frac{n(n+1)}{2} - c$；

在大样本场合，可以证明 $(W^+)^* = \frac{W^+ - E(W^+)}{\sqrt{Var(W^+)}}$ 近似 $N(0,1)$，其中 $E(W^+) = \frac{n(n+1)}{4}$，

$Var(W^+) = \frac{n(n+1)(2n+1)}{24}$，从而水平为 α 的拒绝域为 $\{|(W^+)^*| \geq u_{1-\frac{\alpha}{2}}\}$