生物统计学

第十三章 非参数检验

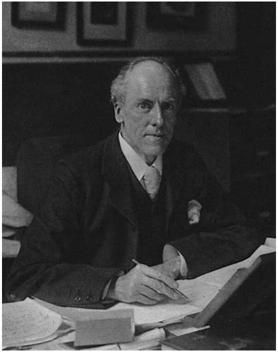
云南大学 生命科学学院

會澤百家 至公天下

关于样本平均数、样本方差的假设检验，包括方差分析、回归分析和相关分析（仅限 Pearson 相关分析），都是关于总体参数的检验。因此，统计学上将它们统称为参数检验(parametric test)。

对于非正态的、未知的总体，则须用非参数检验(non-parametric test)。它是一类与总体分布无关的检验方法，对总体分布的具体形式不作任何限制性的假定，不以总体参数的具体数值估计或检验为目的。

1. 卡方检验
2. 符号检验
3. 秩和检验



Karl Pearson (1857-1936)

- 1892 年，35 岁的 Pearson 抛硬币 2400 次，后来他的学生 C.L.T. Griffith 又抛了 8178 次作为补充。
- 1895 年，Pearson 提出了频率多边形的“平均百分比误差”作为衡量拟合度的标准。
- 1900 年，Pearson 发表拟合优度检验(goodness-of-fit test)。

Pearson 从 n 个服从正态分布的随机变量的联合概率密度函数出发，再次导出了 χ^2 分布的密度函数，随后将理论频率 m_i 和观测频率 m_i'，通过三角转换引入 χ^2 的表达式中，最终得到

$$\sum_{i=1}^{n+1} \frac{(m_i - m_i')^2}{m_i} \sim \chi_n^2 \tag{13.1}$$

Pearson 从 n 个服从正态分布的随机变量的联合概率密度函数出发，再次导出了 χ^2 分布的密度函数，随后将理论频率 m_i 和观测频率 m_i'，通过三角转换引入 χ^2 的表达式中，最终得到

$$\sum_{i=1}^{n+1} \frac{(m_i - m_i')^2}{m_i} \sim \chi_n^2 \tag{13.1}$$

检验了“每次投 12 颗骰子，投 26306 次的结果中 5 或 6 点出现的次数”数据

$$\chi^2 = 43.87241, \quad P = 0.000016$$

非参数方法的卡方检验主要有两个功能：

- 判断理论值与观测值之间的适合程度（适合性检验）
- 判断两个试验因素或变量是否独立（独立性检验）

1. 卡方检验
   - 适合性检验
   - 独立性检验
   - 卡方检验的分解
2. 符号检验
3. 秩和检验

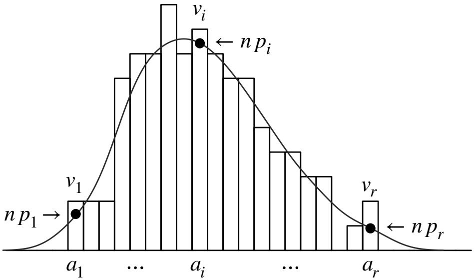


图 **13.1** 理论分布的拟合

$$\chi^2 = \sum_{i=1}^{r} \frac{(v_i - np_i)^2}{np_i} \quad (13.2)$$

v_i：观测频数（实际发生的频数值）；

np_i：理论频数（频数的理论期望值）。

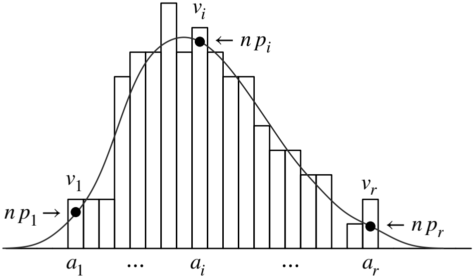


图 **13.1** 理论分布的拟合

$$\chi^2 = \sum_{i=1}^{r} \frac{(v_i - np_i)^2}{np_i} \quad (13.2)$$

$$\chi^2 = \sum_{i=1}^{r} \frac{(O_i - E_i)^2}{E_i} \quad (13.3)$$

v_i：观测频数（实际发生的频数值）；

np_i：理论频数（频数的理论期望值）。

对一组具体的样本算出 χ_c^2 后，如果零假设 H_0 成立，即理论值与观测值相等，或称拟合，出现像 χ_c^2 这么大的差异或更大的差异的概率 $P(\chi^2 \geq \chi_c^2 | H_0)$，就是拟合优度。

对一组具体的样本算出 χ_c^2 后，如果零假设 H_0 成立，即理论值与观测值相等，或称拟合，出现像 χ_c^2 这么大的差异或更大的差异的概率 $P(\chi^2 \geq \chi_c^2|H_0)$，就是拟合优度。

相伴概率 P 值的定义：检验统计量取到当前值，以及比当前值更极端值的概率。

计算 χ^2 值需要注意以下两点：

1. 任一理论频数值 E_i 都必须大于 5。如果 $E_i = np_i \leq 5$，统计量会明显偏离 χ^2 分布。
2. 当 $r = 2$ 时，相应的 χ^2 分布有自由度 $df = r - 1 = 1$。需要对 χ^2 统计量进行连续性矫正，矫正公式为

$$\chi^2 = \sum_{i=1}^{r} \frac{(|O_i - E_i| - 0.5)^2}{E_i} \tag{13.4}$$

例 (13.1)

荷包红鲤（红色，纯隐性性状）与湘江野鲤（青灰色，纯显性性状）杂交，得到青灰色子代 1503 条，红色子代 99 条。试问这一组试验资料是否符合孟德尔一对等位基因的遗传规律，即鲤鱼体色青：红 $= 3:1$。

解

1. 设定零假设 H_0：子代体色性状分离符合 $3:1$ 的比例，备择假设 H_1：子代体色性状分离不符合 $3:1$ 的比例。
2. 选取显著性水平 $\alpha = 0.01$。
3. 计算检验统计量和 P 值。
   - 检验统计量 $\chi_c^2 = \sum_{i=1}^{2} \frac{(|O_i - E_i| - 0.5)^2}{E_i} = \frac{(|1503-1201.5|-0.5)^2}{1201.5} + \frac{(|99-400.5|-0.5)^2}{400.5} \approx 301.636$。
   - 单尾检验的 P 值 $P(\chi^2 \geq \chi_c^2 \approx 301.626|H_0) = 1.456996e - 67$。
4. 作出统计推断。

   子代体色性状分离不符合 $3:1$ 的比例。

例 (13.2)

孟德尔用豌豆的两对相对性状进行杂交试验，黄色（显性）圆滑（显性）种子与绿色（隐性）褶皱（隐性）种子杂交后得子代 556 粒种子。表型数据分布情况为：黄圆 315 粒、黄皱 101 粒、绿圆 108 粒、绿皱 32 粒。试问此结果是否符合自由组合规律，即黄圆：黄皱：绿圆：绿皱 $= 9:3:3:1$。

独立性检验是研究两个及以上因素彼此之间是否有关联的统计方法。

表 **13.1** $r \times c$ 列联表的一般形式

	B_1	B_2	$\cdots$	B_c	合计
A_1	O_{11}	O_{12}	$\cdots$	O_{1c}	$\sum_{j=1}^{c} O_{1j}$
A_2	O_{21}	O_{22}	$\cdots$	O_{2c}	$\sum_{j=1}^{c} O_{2j}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$	
A_r	O_{r1}	O_{r2}	$\cdots$	O_{rc}	$\sum_{j=1}^{c} O_{rj}$
合计	$\sum_{i=1}^{r} O_{i1}$	$\sum_{i=1}^{r} O_{i2}$	$\cdots$	$\sum_{i=1}^{r} O_{ic}$	$\sum_{i=1}^{r} \sum_{j=i}^{c} O_{ij}$

列联表(contingency table) 是观测数据按两个或更多属性（定性变量）分类时所列出的频数表。

例 (13.5)

为研究吸烟与患气管炎之间的关系，对人群进行随机抽样调查（共 500 例），数据如表 13.2 所示。试检验吸烟与患气管炎两种因素之间的关联性。

表 **13.2** 吸烟与气管炎患病的抽样调查资料

不同人群	气管炎		合计
	患病	不患病	
吸烟	50	250	300
不吸烟	5	195	200
合计	55	445	500

例 (13.5)

为研究吸烟与患气管炎之间的关系，对人群进行随机抽样调查（共 500 例），数据如表 13.2 所示。试检验吸烟与患气管炎两种因素之间的关联性。

表 **13.2** 吸烟与气管炎患病的抽样调查资料

不同人群	气管炎		合计
	患病	不患病	
吸烟	50 (**33**)	250 (**267**)	300
不吸烟	5 (**22**)	195 (**178**)	200
合计	55	445	500

配对试验的结果如果组织成列联表，称为配对列联表。

配对列联表的独立性检验又称为 McNemar-Bowker 检验。检验统计量为

$$\chi_c^2 = \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} \frac{(O_{ij} - O_{ji})^2}{O_{ij} + O_{ji}} \tag{13.5}$$

服从 χ^2 分布，有自由度 $df = \frac{k(k-1)}{2}$。

13.5 A 和 B 两种检测方法的调查资料

A 法	B 法		合计
	阳性	阴性	
阳性	56 (O_{11})	23 (O_{12})	79
阴性	16 (O_{21})	5 (O_{22})	21
合计	72	28	100

在 2×2 列联表的独立性检验，以及二分类的适合性检验中，当理论频数小于 5 时，可用**Fisher 的精确检验法**(Fisher's exact test) 来解决。

在 2×2 列联表的独立性检验，以及二分类的适合性检验中，当理论频数小于 5 时，可用**Fisher 的精确检验法**(Fisher's exact test) 来解决。

表 **13.9** 品茶试验的数据资料

制茶方式	女士判断结果		合计
	茶倒入奶	奶倒入茶	
茶倒入奶	3	1	4
奶倒入茶	1	3	4
合计	4	4	8

1. 卡方检验
2. 符号检验
   - 单样本的符号检验
   - 成对数据的符号检验
3. 秩和检验

假设有一个未知分布的总体，其中位数为 ξ。现在从该总体中随机抽取 n 个观测值，记作 $x_1, x_2, \cdots, x_n$。令 $\delta_i = x_i - \xi$。

假设有一个未知分布的总体，其中位数为 ξ。现在从该总体中随机抽取 n 个观测值，记作 $x_1, x_2, \cdots, x_n$。令 $\delta_i = x_i - \xi$。

$$\underbrace{+,+,\cdots,+,+}_{n}\Big|\underbrace{+,+,\cdots,+}_{n-1},-\Big|\cdots\Big|+\underbrace{-,-,\cdots,-,-}_{n-1}\Big|\underbrace{-,-,\cdots,-,-}_{n}$$

假设有一个未知分布的总体，其中位数为 ξ。现在从该总体中随机抽取 n 个观测值，记作 $x_1, x_2, \cdots, x_n$。令 $\delta_i = x_i - \xi$。

$$\underbrace{+,+,\cdots,+,+}_{n} \Big| \underbrace{+,+,\cdots,+,}_{n-1} - \Big| \cdots \Big| + \underbrace{-,-,\cdots,-,-}_{n-1} \Big| \underbrace{-,-,\cdots,-,-}_{n}$$

记 n 个符号中“+”的个数为 n_+，$n_+ = k$ 的概率为

$$P(n_+ = k) = \mathrm{C}_n^k \frac{1}{2}^k (1 - \frac{1}{2})^{n-k} = \mathrm{C}_n^k (\frac{1}{2})^n, \quad k \in (0, 1, 2, \cdots, n) \tag{1}$$

例 (13.11)

假设某项研究要求一批玉米种子发芽率要达到 90%。现随机抽取 10 袋玉米种子，每袋取 200 粒种子做发芽试验，得发芽数分别为 $168, 170, 174, 181, 175, 178, 183, 169, 179, 181$，试检验该批玉米种子的发芽率是否为 90%。

解

1. 设定零假设 $H_0: \xi = 200 \times 0.9 = 180$ 粒，备择假设 $H_1: \xi \neq 180$ 粒。
2. 选取显著性水平 $\alpha = 0.05$。
3. 计算差值并记录符号。

   得 $n_+ = 3$，$n_- = 7$。
4. 计算相伴概率 P 值。

   相伴概率 $P = 2 \times \sum_{i=0}^{3} \mathrm{C}_{10}^{i} (\frac{1}{2})^i \approx 0.344$。
5. 作出统计推断。

   该批种子的发芽率达到了 90% 的要求。

对于成对数据，也可使用符号检验来完成比较。将数据按照配对关系做减法，记录差值的符号。

问题就转化为了单样本的符号检验，

成对 t 检验判断的是差值的总体平均数是否等于 0，
符号检验判断的是差值的总体中位数是否等于 0。

例 (13.12)

猪场随机挑选 15 头猪，记录它们运动前后的心率数据（次/分钟，见 `pigHR` 数据集）。试问运动对猪的心率是否有影响。

解

1. 设定零假设 H_0：运动前后心率差值 d 的总体中位数 $=0$，备择假设 H_1：运动前后心率差值 d 的总体中位数 $\neq 0$。
2. 选取显著性水平 $\alpha = 0.05$。
3. 计算差值并记录符号。

   得 $n_+ = 2, n_- = 12$。
4. 计算相伴概率 P 值。$P = 2 \times \sum_{i=0}^{2} \mathrm{C}_{14}^{i} (\frac{1}{2})^i \approx 0.013$。
5. 作出统计推断。

   运动对猪的心率有显著影响。

1945 年统计学家 Frank Wilcoxon 提出了一种改进方法，称为秩和检验(rank sum test)。

- 将观测值由小到大排列，每一个观测值按照次序排列中的位置编号，也就是秩(rank)，重新编码。
- 然后计算出秩和(rank sum) 进行检验。

秩和检验效率高于符号检验，因为它除了比较差值的符号外，还比较差值的秩大小。

随机选取 4 块地不允许有杂草，其它 4 块地每行间正好有 3 株羊角芹。实验得玉米产量 (kg/亩):

无杂草地块 $700.14, 723.24, 693.00, 742.98$（记作 A 组）

少量杂草地块 $666.12, 740.88, 643.02, 655.20$（记作 B 组）

643.02 655.20 666.12 **693.00** **700.14** **723.24** 740.88 **742.98**

643.02 655.20 666.12 **693.00** **700.14** **723.24** 740.88 **742.98**

将以上排序队列中的数字转换成位置编号，也就是秩，得

1 2 3 **4** **5** **6** 7 **8**

643.02 655.20 666.12 **693.00** **700.14** **723.24** 740.88 **742.98**

将以上排序队列中的数字转换成位置编号，也就是秩，得

1 2 3 **4** **5** **6** 7 **8**

再将 A 组和 B 组的秩分别求和，得秩和

13 **23**

假设少量杂草对产量没有影响（零假设 H_0），我们应当看到 $W_1 = 18$。但是，当前的实际情况是 $W_1 = 13$，偏离了零假设成立时的理论值 18。

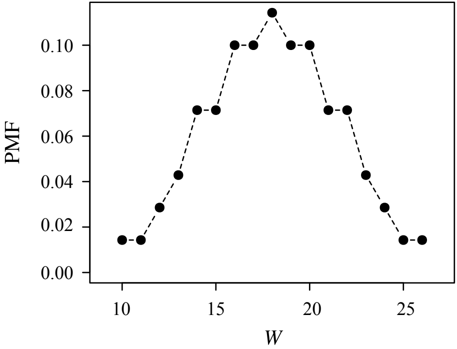


图 **13.3** 秩和统计量的概率分布

为克服符号检验未充分用数据信息的缺点，Wilcoxon 对其进行了改进，提出了符号秩检验(signed rank test)，也称 Wilcoxon 配对检验。

例题 13.12 中，五头猪在运动前心率 (BHR) 和运动后心率 (AHR) 数据，以及它们的差值 (DIF) 如下

$$\begin{aligned} \mathrm{BHR}: &60 \quad 65 \quad 68 \quad 64 \quad 74 \\ \mathrm{AHR}: &76 \quad 80 \quad 85 \quad 77 \quad 72 \\ \mathrm{DIF}: &16 \quad 15 \quad 17 \quad 13 \quad -2 \end{aligned}$$

例题 13.12 中，五头猪在运动前心率 (BHR) 和运动后心率 (AHR) 数据，以及它们的差值 (DIF) 如下

$$\begin{aligned}
\text{BHR}: 60 \quad 65 \quad 68 \quad 64 \quad 74 \\
\text{AHR}: 76 \quad 80 \quad 85 \quad 77 \quad 72 \\
\text{DIF}: 16 \quad 15 \quad 17 \quad 13 \quad -2
\end{aligned}$$

对差值的绝对值从小到大排序，记录它们的秩，并对原差值是正值的作加粗处理，将得到

$$\begin{array}{ccccc}
2 & \mathbf{13} & \mathbf{15} & \mathbf{16} & \mathbf{17} \\
1 & \mathbf{2} & \mathbf{3} & \mathbf{4} & \mathbf{5}
\end{array}$$

将所有负差值的秩求和得 $W_{-} = 1$，将所有正差值的秩求和得 $W_{+} = 14$。两个秩和中的较小者，即为**Wilcoxon** 符号秩统计量。

本章小结

1. 卡方检验
   - 适合性检验
   - 独立性检验
   - 卡方检验的分解
2. 符号检验
   - 单样本的符号检验
   - 成对数据的符号检验
3. 秩和检验
   - 成组数据的秩和检验
   - 成对数据的符号秩检验
   - 多组数据的秩和检验