第二章　单变量位置推断问题

单一随机变量位置点估计、置信区间估计和假设检验是参数统计推断的基本内容, 其中t 统计量和t 检验作为正态分布总体期望均值的推断工具,是我们所熟知的. 如果数据不服从正态分布, 或有明显的偏态表现,在t 统计量和t 检验推断下的结论不一定可靠. 本章将关注三方面的推断问题：(1) 非参数位置检验基本检验; (2) 非参数置信区间构造问题; (3) 分布的检验. 主要内容包括符号检验和分位数推断及其扩展应用、对称分布的Wicoxon 秩和检验和推断、估计量的稳健性评价、正态记分检验和应用、单一总体拟合优度检验等. 最后一节给出单一总体中心位置各种不同检验的渐近相对效率的相关理论结果.

§2.1　符号检验和分位数推断

§2.1.1　基本概念

符号检验(sign test)是非参数统计中最古老的检验方法之一,最早可追溯到1701年苏格兰数理统计学家约翰 • 阿巴斯诺特(John Arbuthnot) 有关伦敦出生的男婴女婴比率是否超过1/2 的性别比平衡性研究.该检验被称为符号检验的一个原因是因为它所关心的信息只与两类观测值有关。如果用符号"+"和"−"区分,符号检验就是通过符号"+" 和"−"的个数来进行统计推断的,故称为符号检验.

首先看一个例子.

例2.1　假设某城市16座预出售的楼盘均价(单位：　百元/m^2)如表2.1所示

表2.1　16座预出售的楼盘均价

36	32	31	25	28	36	40	32
41	26	35	35	32	87	33	35

问：该地平均楼盘价格是否与媒体公布的37百元/m^2的说法相符?

解　这是一个实际的问题，可以将其转化成一个单一随机变量分布位置参数的假设检验问题.在参数假设检验中，我们所熟知的是正态分布未知参数的检验问题。假设在某一统计时点上楼盘价格服从正态分布$N(\mu,\sigma^2)$, 依照题意和参数统计的基本原理和步骤, 可以建立如下原假设和备择假设：

$$H_0:\mu=37 \leftrightarrow H_1:\mu\neq 37,$$

其中, μ是分布均值,原假设$\mu_0 = 37$, 根据样本数据计算样本均值和样本方差分别为$\bar{X} = 36.50, S^2 = 200.53$.

由于$n = 16 < 30$为小样本, 采用t统计量计算检验统计值：

$$t = \frac{\bar{X} - \mu_0}{S/\sqrt{n}} = -0.1412,$$

根据自由度为$n - 1 = 16 - 1 = 15$, 得t检验的p值为0.89, 在显著性水平$\alpha < 0.89$以下都不能拒绝原假设.

R中t检验参考程序和输出结果如下：

```
> attach(build.price)
  [1] 36 32 31 25 28 36 40 32 41 26 35 35 32 87 33 35
> mean(build.price)
  [1] 36.5
> var(build.price)
  [1] 200.5333
> length(build.price)
  16
> t.test(build.price-37)

  One-sample t-Test
  data:  build.price - 37
  t = -0.1412, df = 15, p-value = 0.8896
  alternative hypothesis: true mean is not equal to 0
  95 percent confidence interval:
   -8.045853  7.045853
  sample estimates:
   mean of x
        -0.5
```

t检验的结果是接受原假设。我们注意到在以上16个数据中, 3个楼盘的均价高于37百元/m^2, 而另外13 个楼盘的均价都低于37 百元/m^2. 由正态分布的对称性可知, 如果37 百元/m^{2}可以作为正态分布的平均水平,那么从该正态总体中取出的样本分布在37 百元左侧（小于37）与右侧（大于37）的数量应大致相等,不会出现大比例失衡. 然而观察数据发现,3:13 显然难以支持37 百元/m^2 作为正态分布的对称中心的说法,这与t检验选择接受原假设的结论并不一致,是数据量太少显著性证据不充分呢? 还是方法使用不当呢？

我们先来回答第一个有关样本量是否不足的问题。让我们换一个角度考虑位置检验推断问题. 不妨先试试将37 理解为总体的中位数,那么数据中应该差不多各有一半在37 的两侧.计算每一个数据与37的差, 大于37位于右侧的样本个数为3, 小于37 位于37左侧样本个数为13, 这是一个中位数为37 的分布应有的样本特征吗？在原假设和独立同分布的随机抽样的条件下,每一个样本理应等可能地出现在37的左与右，3:13 是一个在中位数两侧分布比例均衡的结果呢？还是一个明显的分布不均的结果？为此需要考虑出现在中位数左侧的样本量，它服从二项分布$\sim B(16,0.5)$. 在这个分布下很容易计算出出现3 个以下样本的可能性是小于0.05 的（请思考这是怎么计算出来的？）。这表明如果从中位数的角度来看这个问题的话，中位数37 受到拒绝，这表明对数据量不充分的猜测完全是错误的。这个分析思路实际上就是符号检验的基本原理.下面给出规范的符号检验推断过程.

假设$X_1, X_2, \cdots, X_n$是来自总体$\mathcal{F}(M_{\mathrm{e}})$的简单随机样本,$M_{\mathrm{e}}$ 是总体的中位数, 有位置模型(location model)

$$X_i = M_{\mathrm{e}} + \epsilon_i, i = 1, 2, \cdots, n. \tag{2.1}$$

我们感兴趣的是如下假设检验问题：

$$H_0 : M_{\mathrm{e}} = M_0 \leftrightarrow H_1 : M_{\mathrm{e}} \neq M_0. \tag{2.2}$$

式中, M_0是事先给定的待检验中位数值. 定义新变量：$Y_i = I\{X_i > M_0\}, Z_i = I\{X_i < M_0\}, i = 1, 2, \cdots, n$

$$S^+ = \sum_{i=1}^{n} Y_i, \quad S^- = \sum_{i=1}^{n} Z_i.$$

$S^+ + S^- = n'(n' \leqslant n)$, 令$K = \min\{S^+, S^-\}$.在原假设之下, 假设检验问题(2.2)等价于另一个随机变量Y的检验问题，如式(2.3)所示. 其中$Y \sim b(1, p)$,$p = P(X > M_0)$,

$$H_0 : p = 0.5 \leftrightarrow H_1 : p \neq 0.5. \tag{2.3}$$

此时, $K \leqslant k$可以按照抽样分布$B(n', 0.5)$求解得到, 在显著性水平α下，检验的拒绝域为

$$2 \times P_{\mathrm{binom}}(K \leqslant k | n', p = 0.5) \leqslant \alpha.$$

其中, k是满足上式最大的值. 也可以通过计算统计量K的p值做决策：如果统计量K的值是k, p值$= 2 \times P_{\mathrm{binom}}(K \leqslant k | n', p = 0.5)$,当$\alpha > p$时,拒绝原假设. 也就是说, 当大部分数据都在$M_0$的右边, 此时$S^+$较大, S^-较小, 则认为数据的中心位置大于M_0;反之, 当大部分数据都在M_0的左边时S^-较大, S^+ 较小, 则认为数据的中心位置小于M_0. 两种现象都是M_{e}不等于M_0 的直接证据。

例2.2(例2.1续解) 根据符号检验，假设检验问题表示为

$$H_0 : M_{\mathrm{e}} = 37 \leftrightarrow H_1 : M_{\mathrm{e}} \neq 37. \tag{2.4}$$

式中, M_{e}是总体的中位数. 如果原假设为真,即37 是总体的中位数. 用S^+ 表示位于37 右边点的个数, S^- 表示位于37左边点的个数, 数据中没有等于37的数, $S^+ + S^- = 16$.在原假设和独立同分布的随机抽样的条件下,每个样本等可能出现在37 的左与右。也就是说, $S^+ \sim B(n, 0.5)$. 从有利于接受备择假设的角度出发, S^+ 过大或过小, 都表示37不能作为总体的中心.

取$k = 3, 2 \times P(K \leqslant k|n = 16, p = 0.5) = 2 \times \sum_{i=0}^{3} \binom{16}{i} \left(\frac{1}{2}\right)^{16} \approx 0.0213$. 于是, 在显著性水平0.05之下, 拒绝原假设, 认为这些数据的中心位置与37百元/m^2存在显著性差异.

符号检验的R程序及输出结果如下：

```
> binom.test(sum(build.price<37),length(build.price),0.5)

   Exact binomial test

   data:  sum(build.price > 37) and length(build.price)
   number of successes = 3, number of trials = 16, p-value = 0.02127
   alternative hypothesis: true probability of success is not equal to 0.5
   95 percent confidence interval:
    0.04047373 0.45645655
   sample estimates:
   probability of success
                   0.1875
```

对结果的讨论

我们注意到, 在相同的显著性水平之下,t检验和符号检验看似得到了相反的结论,t 检验的结果平均值等于37，符号检验的结果平均值不等于37，我们应该采用哪一种分析结论呢?在回答这个问题之前, 首先应该明确的是,仅从两个检验过程的结论来评价两种检验并不恰当, 这是因为两种方法的分析目标是很不一样的, 一个将37 作为正态分布的均值考虑,而另一个则将37 作为中位数考虑, 各司其职，从不同的角度呈现数据中隐含的参数信息，对问题的理解角度不同，得到看似不同的结论在数据分析中是很常见的。

然而结合推断过程的决策细节来做一个选择还是有可能的.首先，在t检验中,结论是不能拒绝原假设, 它并不表示接受原假设, 而是表示要拒绝原假设还需要收集

更多的证据.我们知道要做出接受原假设的决策,还需要计算决策的势, 也就是不犯第Ⅱ类错误的概率，这样t 检验的做出接受原假设的决策的可靠性没有保证。由于符号检验在仅假定数据服从常规连续分布的情况下就得到了拒绝的结论,这一决定的风险至少有0.05 的显著性水平作为保证,表明已收集到的数据对于形成可靠性决定而言，其提供的数据是充分的.

另外，一个经典的假设检验过程通常由以下几个步骤构成：假定随机变量分布族→ 确定假设→ 检验统计量在原假设下的抽样分布→ 由抽样分布计算拒绝域或计算p值或与预设的显著性比较→ 做出决策。我们知道单一连续数据总体中心位置的参数有中位数和均值, 样本均值的点估计是总体均值, 样本中位数的点估计是总体中位数. 对来自正态分布的样本而言, 均值与中位数相等; 但对于非对称分布而言, 中位数较均值而言是对总体中心位置更稳健的估计.

t检验是在正态总体的前提假定下得到结果, 接受原假设也必须回到正态的假设分布中，这样就出现了结论不一致的问题，因为在正态分布中这两个位置是一个位置。既然数据是充分的，使用正态分布的假定又出现了自相矛盾的结果，而符号检验给出了拒绝原假设的可靠性结论。综合而言，不当的分布假定导致t 检验使用不当才是t 检验没有成功的原因. 也就是说,正是由于分布假设错误, 导致了本该充分的证据没有产生对参数作出可靠性推断的结论,也正因为如此才导致两种检验的结果看似不一致，实际上这种不一致仅仅是在正态假设中的不一致。而放松了对总体分布的假定，对要回答的问题的参数进行另一种选择，就会发掘出数据背后可靠的信息，这里符号检验的结果较t 检验的结果更可信.

类似地, 给出符号检验单边假设检验问题的检验方法, 如表2.2所示.

表2.2 符号检验单边假设检验问题检验方法

左边检验	$H_0: M_{\mathrm{e}} \leqslant M_0 \leftrightarrow H_1: M_{\mathrm{e}} > M_0$	$P_{\mathrm{binom}}(S^- \leqslant k \mid n', p=0.5) \leqslant \alpha$ 其中k是满足上式最大的K
右边检验	$H_0: M_{\mathrm{e}} \geqslant M_0 \leftrightarrow H_1: M_{\mathrm{e}} < M_0$	$P_{\mathrm{binom}}(S^+ \leqslant k \mid n', p=0.5) \leqslant \alpha$ 其中k是满足上式最大的K

说明P_{binom}表示二项分布的分布函数.

§2.1.2 大样本的检验方法

当样本量较大时, 可以使用二项分布的正态近似进行检验, 也就是说, 当$S^+ \sim$

$B\left(n', \frac{1}{2}\right)$ 时, $S^+ \dot{\sim} N\left(\frac{n'}{2}, \frac{n'}{4}\right)$, 定义

$$Z = \frac{S^+ - \frac{n'}{2}}{\sqrt{\frac{n'}{4}}} \xrightarrow{\mathcal{L}} N(0,1), \quad n \to +\infty. \tag{2.5}$$

当n'不够大时, 可以用Z的正态性修正, 其式如下：

$$Z = \frac{S^+ - \frac{n'}{2} + C}{\sqrt{\frac{n'}{4}}} \xrightarrow{\mathcal{L}} N(0,1). \tag{2.6}$$

一般, 当$S^+ < \frac{n'}{2}$时, $C = -\frac{1}{2}$; 当$S^+ > \frac{n'}{2}$时, $C = \frac{1}{2}$. 相应的p值为$2P_{N(0,1)}(Z < z)$. 同理, 可以得到单侧检验方法如表2.3所示.

表2.3 符号检验大样本假设检验方法

左侧检验	$H_0 : M_e \leqslant M_0 \leftrightarrow H_1 : M_e > M_0,$	p值为$P_{N(0,1)}(Z \geqslant z)$
右侧检验	$H_0 : M_e \geqslant M_0 \leftrightarrow H_1 : M_e < M_0,$	p值为$P_{N(0,1)}(Z \leqslant z)$

关于正态性修正的讨论

对离散分布应用正态性修正是非参数统计推断中较为普遍的做法. 我们知道, 很多检验统计量都可以表达为独立随机变量和形式的随机变量，它的抽样分布的近似分布都是正态分布。然而，不同抽样分布渐近性的收敛速度却可能很不同,有的分布在样本量较小的时候近似效果就不错,而有的分布则在样本量很大的时候,近似效果还不够理想.有的时候还会受到参数本身数值的影响。为克服利用连续分布对离散分布估计在样本量不大时可能出现的尾部概率估计偏差, 在对离散分布左右两侧点的概率分布值进行计算时,不直接采用正态分布值估计, 而是通过对分布中心位置作出一定的平移进而取得一定的修正效果.

正态性修正的具体定义为： 假设X服从离散分布,X所有的可能取值为$\{0, 1, 2, \cdots, n\}$, 如果X近似的正态分布为$N(\mu, \sigma^2)$,当待估计的点$X = k > n/2$时, k处的概率分布函数$P(X \leqslant k)$用正态分布$N(\mu - C, \sigma^2)$ 在k处的分布函数估计, $C = 1/2$, 这相当于用位置参数向右平移$1/2$单位的分布来估计k的概率分布; 同理, 当待估计的点$X = k < n/2$时, k处的概率分布函数$P(X \leqslant k)$用正态分布$N(\mu - C, \sigma^2)$在k处的分布函数估计, $C = -1/2$, 这相当于用位置参数向左平移$1/2$单位的分布来估计k的概率分布. 当$n = 30$时,二项分布、正态分布和正态性正负修正的左右两端代表点上的分布函数比较如表2.4所示.

表2.4 二项分布$B(30,0.5)$、正态分布和正态性正负修正之间的分布函数$P(X \leqslant k)$ 比较

k	0	1	2	21	22	23
$b(30,0.5)$	9.31e−10	2.89−08	4.34e−07	9.79e−01	9.92e−01	9.98e−01
$N(15,7.5)$	2.16e−08	1.59e−07	1.03e−06	9.66e−01	9.86e−01	9.95e−01
$N(15-1/2,7.5)$	—	—	—	9.78e−01	9.91e−01	9.97e−01
$N(15+1/2,7.5)$	7.58e−09	5.96e−08	4.12e−07	—	—	—

由表(2.4)可以看出, 对较大点处的分布函数做正态分布正修正结果$\left(C=\frac{1}{2}\right)$ 与二项分布精确分布比较接近, 对较小点处的分布函数做正态分布负修正结果$\left(C=-\frac{1}{2}\right)$与二项分布精确分布比较接近.

例2.3 设某大学有A和B两个食堂,为了解教职工对两个食堂的餐饮服务是否存在倾向性差异,每隔一个月随机安排共计50位教职工填写意向度调查表,每位回答者只能从两个食堂中选择一个作为自己倾向性更高的食堂，某月得到以下数据：

喜欢A食堂的人数： 29人

喜欢B食堂的人数： 18人

不能区分的人数： 3人

分析在显著性水平$\alpha = 0.10$下, 是否可以认为，在该大学校园两家食堂被喜爱的程度存在差异。

解 假设检验问题：

$H_0 : P(\mathrm{A}) = P(\mathrm{B})$, 喜欢 A 食堂的客户与喜欢 B 食堂的客户比例相等,
$H_1 : P(\mathrm{A}) \neq P(\mathrm{B})$, 喜欢 A 食堂的客户与喜欢 B 食堂的客户比例不等.

分析：这是定性数据的假设检验问题, 可以应用符号检验, 喜欢A食堂的人数设为S^+, $S^+ = 29$; 喜欢B食堂的人数设为S^-, $S^- = 18$, $S^+ + S^- = n' = 47$, $\frac{n'}{2} = 23.5$, 由于$S^+ > 23.5$,所以取正修正, 应用式(2.6)有

$$Z = \frac{29 - 23.5 + \frac{1}{2}}{\sqrt{\frac{47}{4}}} = 1.75 > Z_{0.05} = 1.64$$

其中$Z_{0.05}$是标准正态分布的0.05尾分位点。

结论：在显著性水平$\alpha = 0.10$下拒绝原假设, 证据显示A食堂和B食堂用餐者的倾向性存在显著差异.

§2.1.3 符号检验在配对样本比较中的应用

在对两总体进行比较的时候, 配对样本是经常遇到的情况, 比如生物的雌雄、人体疾病的有无、前后两次试验的结果, 意见的赞成或反对等. 这时, 设配对观测值

为$(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n)$. 在$n$对样本数据中, 若$x_i<y_i$, 则记为"+"; 若$x_i>y_i$, 则记为"−"; 若$x_i=y_i$, 则记为0. 于是数据可能被分成三类(+,−,0). 我们只比较"+"和"−" 的个数, 记"+"和"−"的个数和为$n',n'\leqslant n$.问题是比较两类数据的比例是否相等. 假设P_+为"+" 的比例, P_-为"−"的比例, 则可以有假设检验:

$$H_0:P_+=P_-,$$

$$H_1:P_+\neq P_-.$$

这类问题由于只涉及符号, 自然可以用符号检验来分析。看下面的例题.

例2.4 表2.5所示为某瑜伽教练的一个小班12位学员每周两次跟班训练一年前后的体重数据, 用符号检验分析参加该教练的瑜伽训练活动对体重的影响效果如何（$\alpha=0.05$）.

解 假设检验问题:

$$H_0:P(\text{瑜伽训练前体重})=P(\text{瑜伽训练一年后体重}),$$

$$H_1:P(\text{瑜伽训练前体重})\neq P(\text{瑜伽训练一年后体重}).$$

分析:这是定性数据的假设检验问题, 可以应用符号检验. 瑜伽训练前体重大于瑜伽训练后体重的样本个数记为S^+, $S^+=6$; 瑜伽训练前体重小于瑜伽训练后体重的样本个数记为S^-, $S^-=4$; $S^++S^-=n'=10$, $\dfrac{n'}{2}=5$. 应用式(2.6), 有

$$Z=\frac{6-5+\dfrac{1}{2}}{\sqrt{\dfrac{10}{4}}}=0.9487<Z_{0.025}=1.96$$

式中$Z_{0.05}$是标准正态分布的0.05尾分位点。
结论:证据不足不能拒绝原假设,没有充分证据显示瑜伽训练前与训练后学员体重有明显减低.

表2.5 瑜伽训练一年前后学员体重变化比较表(单位:kg)

学员号	瑜伽训练前体重	瑜伽训练后体重	符号
1	71	66	+
2	78.5	73	+
3	69	70	0
4	74.5	70	+
5	61.5	64	−
6	68	72	−
7	59	63	−
8	68	63	+
9	57	56.5	0
10	63	67	−
11	62	55	+
12	70	64	+

我们注意到，在面对体重的数据分析中，进行符号计数的时候，对学员前后体重差异未超过1kg的都未做符号计数，这是因为采取了个体体重测量误差在1kg以内属于正常体重偏差这个通用体重测量误差标准。

值得注意的是，在本例结论部分,我们并没有草率地选择接受原假设,而是较为谨慎地选择了没有充分证据拒绝原假设来表述结论。这样做的目的是提醒假设检验的使用者注意接受原假设可能犯第 II 类错误的潜在风险.

§2.1.4 分位数检验—— 符号检验的推广

以上我们主要介绍了中位数的符号检验,实际上以上方法完全可以扩展到单一随机变量分布的任意p 分位数的检验. 假设单一随机变量$\mathcal{F}(M_p)$, M_p是总体的p 分位数, 对于假设检验问题：

$$H_0 : M_p = M_{p_0} \leftrightarrow H_1 : M_p \neq M_{p_0},$$

M_{p_0}是待检验的p_0分位数. 上述检验问题等价于

$$H_0 : p = p_0 \leftrightarrow H_1 : p \neq p_0.$$

类似中位数检验, 定义$Y_i = I\{X_i > M_{p_0}\}, Z_i = I\{X_i < M_{p_0}\}$, 我们注意到在原假设之下, $Y_i \sim B(1, 1-p_0), Z_i \sim B(1, p_0)$,

$$S^+ = \sum_{i=1}^{n} Y_i, \quad S^- = \sum_{i=1}^{n} Z_i.$$

注意到S^+是数据落在M_{p_0}右边的数据量, S^-是数据落在M_{p_0}左边的数据量. 假设有效数据量$n' = S^+ + S^-$, 原假设下$S^- \sim B(n', p_0)$, $S^+ \sim B(n', 1-p_0)$, 容易注意到此时二项分布不再是对称分布，所以得到假设检验问题的结果如表2.6 所示.

表2.6 分位数符号检验问题结果

$H_0: M_p = M_{p_0} \leftrightarrow H_1: M_p \neq M_{p_0}$	$p_0 > 0.5$ 时,$P_{\text{binom}}\{S^- \leqslant k_1 \mid n', p = p_0\}$ $+P_{\text{binom}}(\{S^+ \leqslant k_2 \mid n', p = 1 - p_0\}) \leqslant \alpha$ $p_0 < 0.5$时, $P_{\text{binom}}\{S^+ \leqslant k_1 \mid n', p = 1 - p_0\}$ $+P_{\text{binom}}(\{S^- \leqslant k_2 \mid n', p = p_0\}) \leqslant \alpha$ 其中k_1, k_2 是满足上式最大的k_1, k_2
$H_0: M_p \leqslant M_{p_0} \leftrightarrow H_1: M_p > M_{p_0}$	$P_{\text{binom}}(S^- \leqslant k \mid n', p = p_0) \leqslant \alpha$ 其中k是满足上式最大的k
$H_0: M_p \geqslant M_{p_0} \leftrightarrow H_1: M_p < M_{p_0}$	$P_{\text{binom}}(S^+ \leqslant k \mid n', p = 1 - p_0) \leqslant \alpha$ 其中k是满足上式最大的k

例2.4续 根据医学知识，降低体重必须通过热量的高消耗来实现，而普通的瑜伽训练重在改善肌肉骨骼结构增强体质，个体差异比较大，并不都达到实现减重的目标。然而小部分学员可通过课内和课外训练结合的方式，来达到热量消耗从而降低体重。如此一来，考虑将降低体重的中位目标降低到3/4分位目标，学员中能否有1/4的学员通过瑜伽训练能够减重超过2kg，学员训练前后体重降低值如表2.7所示，显著性水平设为$\alpha = 0.05$.

表2.7 学员训练前后体重降低值

	5.0	5.5	4.5	-2.5	-4.0	-4.0	5.0	-4.0	7.0	6.0
$M_{0.75}$是否> 2	+	+	+	-	-	-	+	-	+	+

解 假设检验问题是：

$$H_0: M_{0.75} \leqslant 2 \leftrightarrow H_1: M_{0.75} > 2.$$

$S^+ = 6, S^- = 4$, 计算$\times P_{\text{binom}(4|10,0.75)}(S^- \leqslant 4) = 0.02 < \alpha = 0.05$, 因而拒绝原假设, 认为3/4分位数大于2，于是参加瑜伽训练班个人努力增强脂肪代谢，可以期待身体素质较高的学员能够达到降低体重的目标。

§2.2 Cox-Stuart趋势存在性检验

在客观世界中会遇到各种各样随时间变动的数据序列,人们通常关心这些数据随时间变化的规律,其中趋势分析是几乎都会分析的内容. 在趋势分析中,人们首先关心的是趋势是否存在, 比如, 收入是否下降了?农产品的产量或某地区历年的降雨量是否随着时间增加了? 如果趋势存在,则根据实际需要建立更精细的模型以刻画或度量趋势的程度或变化规律.一些分析习惯于将存在性问题和确定性问题以及影响程度等问题放在一起由一个模型统一来回答,比如，回归分析就是最常用的趋势

分析的工具. 通常的做法是用线性回归拟合直线, 然后再通过检验验证线性假设的合理性,如果检验通过, 则表示回归模型是合适的, 线性趋势存在.如果检验未通过, 那么趋势是否存在呢? 也就是说,当线性趋势没有得到肯定时, 是否也该否定其他可能的趋势的存在性呢? 显然, 答案是否定的. 因为存在性是一个一般性的问题, 而特定形式的模型是所有可能趋势中的某一种,用特殊的形式回答一般性的问题, 显然存在不可回答的风险. 也就是说, 当线性趋势被否定时, 也许有结构假定不恰当等多种原因,并不能一概否定其他趋势的存在性.我们显然也无法通过穷尽所有可能的结构来回答存在性问题. 而即便模型通过了检验, 也只能说在模型的假设之下, 数据的趋势是存在的.

考克斯(Cox)与斯托特(Stuart)在研究数列趋势问题的时候注意到了这一点.他们于1955 年提出了一种不依赖于趋势模型的快速判断趋势是否存在的方法,这一方法称为Cox-Stuart趋势存在性检验, 它的理论基础正是2.1 节的符号检验. 他们的想法是：如果数据有上升趋势,那么排在后面数的取值比排在前面的数显著地大; 反之,如果数据有明显的下降趋势,那么排在后面的数的取值比排在前面的数显著地小. 换句话讲, 我们可能生成一些数对,每一个数对是从前后两段数据中各选出一个数构成的,这些数对可以反映前后数据的变化.为保证数对同分布,前后两个数的间隔尽可能大, 这就意味着可以将数据一分为二,自然形成前后数对;在数据量充足的情况下，也可以将数据一分为三，将中间部分忽略不计，取前后两段数据。为保证数对不受局部干扰, 前后两个数的间隔应较大, 另外数对的数量也不能过少. 具体而言，考虑序列y_s的趋势问题表示如下：

$$y_s = \alpha + \Delta s + \epsilon_s, (s = 1, 2, \cdots, N) \tag{2.7}$$

$$H_0 : \Delta \leqslant 0 \quad \leftrightarrow \quad H_1 : \Delta > 0. \tag{2.8}$$

$\Delta > 0$表示正趋势参数，ϵ_s是随机误差项，定义$s < t$, 计算得分：

$$h_{st} = \begin{cases} 1, & y_s > y_t. \\ 0, & y_s < y_t. \end{cases}$$

Cox-Stuart统计量定义为

$$S = \sum_{s<t} w_{st} h_{st} \tag{2.9}$$

§2.2.1 最优权重Cox-Stuart统计量基本原理

先考虑正态的情况（假设N为偶数），$y_s \sim N(s\Delta, 1)$,(忽略α.) H_0 之下，有$h_{st} \sim$

$b(1,\frac{1}{2})$, $\mu(h_{st})=P(y_s-y_t>0)$, ϕ 是标准正态分布的分布函数，于是有

$$\mu(h_{st})=\phi\left(-\frac{(s-t)\Delta}{\sqrt{2}}\right); \tag{2.10}$$

$$\mu'(h_{st})=\left[\frac{\partial\mu(h_{st})}{\partial\Delta}\right]_{\Delta=0}=\frac{s-t}{2\sqrt{\pi}}. \tag{2.11}$$

令$r_{st}=s-t$,将(2.10)和(2.11)代入(2.9)有：

$$\mu'_S=\sum w_{st}\mu'(h_{st})=\frac{1}{2\sqrt{\pi}}\sum w_{st}r_{st}; \tag{2.12}$$

$$\sigma^2_{S|H_0}=\sum w_{st}^2\sigma^2(h_{st}|H_0)=\frac{1}{4}\sum w_{st}^2 \tag{2.13}$$

$$C_S^2=\sum\frac{\mu'_S}{\sigma^2_{S|H_0}}=\frac{1}{\pi}\frac{(\sum w_{st}r_{st})^2}{\sum w_{st}^2} \tag{2.14}$$

这里的C_S^2是第一章讨论的效率。使(2.14)式取比较大的值相当于一个两阶段的求解问题：先固定r_{st}, 令w_{st}变化使（2.14）取大，再令r_{st}变动使C_S^2 继续取大。于是(2.14)最大值的求解问题相当于求下面的最优化问题：

$$f(w,r)=\sum w_{st}r_{st}-\lambda\sum w_{st}^2 \tag{2.15}$$

$$\text{对}w\text{求导得到：}\ r_{st}+w_{st}\frac{\partial r_{st}}{\partial w_{st}}-2\lambda w_{st}=0 \tag{2.16}$$

$$i.e.\frac{r_{st}}{w_{st}}+\frac{\partial r_{st}}{\partial w_{st}}=2\lambda \tag{2.17}$$

$$w_{st}=\lambda r_{st}\text{满足(2.17)式} \tag{2.18}$$

这表示S表达式中的权重与序号间隔成正比时，可以使统计量的效率最大化，由此可以构造出第一个S 统计量,记作S_1：

$$S_1=\sum_{k=1}^{\lfloor N/2\rfloor}(N-2k+1)h_{k,N-k+1}$$

注意到这时$\{s,t\}=\{(1,N),(2,N-2),\cdots,(K,N-K+1)\},K=\lfloor N/2\rfloor$.

$$\mu_{S_1|H_0}=\frac{1}{2}\sum(N-2k+1)=\frac{1}{8}N^2\quad\sigma^2_{S_1|H_0}=\frac{1}{4}\sum(N-2k+1)^2=\frac{1}{24}N(N^2-1)$$

于是可以产生如下的第一种假设检验方法,当$y_s\sim N(\Delta s,1)$,

$$S_1^*=\frac{S_1-\frac{1}{8}N^2}{\sqrt{\frac{1}{24}N(N^2-1)}}\sim_{N\to+\infty}N(0,1)$$

$S_1^* > Z_\alpha$时有下降趋势; 反之, $S_1^* < Z_{-\alpha}$时, 有上升趋势,Z_α是标准正态分布α尾分位点。

例2.5 南美洲某国2015-2017连续三年统计月度失业率数据如表2.8所示，请根据失业率数据进行分析，判断失业率在2015年以后是否有逐年下降的趋势？

表2.8 某国36个月失业率数据表（月度数据，单位%）

年月	1501	1502	1503	1504	1505	1506	1507	1508	1509
失业率	8.5	7.1	8.2	11.5	7.0	8.2	9.5	7.8	9.2
年月	1510	1511	1512	1601	1602	1603	1604	1605	1606
失业率	10.2	9.0	9.4	9.2	8.9	10.5	8.9	7.3	8.8
年月	1607	1608	1609	1610	1611	1612	1701	1702	1703
失业率	8.4	6.9	8.0	7.8	6.3	7.5	8.7	7.0	8.4
年月	1704	1705	1706	1707	1708	1709	1710	1711	1712
失业率	9.4	8.2	8.6	8.0	7.6	11.1	7.3	5.5	7.0

解 假设检验问题：

H_0 : 该地区36个月失业率无变化 $\leftrightarrow$ H_1 : 该地区前36 个月失业率有下降的趋势.

分析：令$K = N/2 = 36/2 = 18$, 前后观测值为

表2.9 失业率数据的Cox-StuartS_1统计量数据对形成表

y	(y_1, y_{36})	(y_2, y_{35})	(y_3, y_{34})	$\cdots$	(y_{18}, y_{19})
数对	(8.5,7)	(7.1,5.5)	(8.2,7.3)	$\cdots$	(8.8,8.4)
w_{st}	35	33	31	$\cdots$	1
h_{st}	+	+	+	$\cdots$	+

本例中, 这18个数据对按权重相加$S_1 = 257$,$\mu(S_1) = 162$, $\sigma^2(S_1) = 1942.5$

$$S_1^* = \frac{S_1 - \frac{1}{8}N^2}{\sqrt{\frac{1}{24}N(N^2-1)}} = \frac{257 - 162}{44.07} = 2.155.$$

标准正态分布p值为$P(S_1^* > 2.155) = 0.0156 < \alpha = 0.05$, 表明该地失业率有下降趋势. R程序如下：

```
> UNE.rate
  [1]  8.5  7.1  8.2 11.5  7.0  8.2  9.5  7.8  9.2 10.2  9.0  9.4
 [13]  9.2  8.9 10.5  8.9  7.3  8.8  8.4  6.9  8.0  7.8  6.3  7.5
 [25]  8.7  7.0  8.4  9.4  8.2  8.6  8.0  7.6 11.1  7.3  5.5  7.0
> N=length(UNE.rate)
> N
```

```
  [1] 36
> k=N/2
> w=seq(N-2*1+1,N-2*k+1,-2)
> S=sum(w*(UNE.rate[1:(N/2)]-rev(UNE.rate[(N/2+1):N])>0))
> S
 [1] 257
> ES=N^2/8
> DS=1/24*N*(N^2-1)
> S.star=(S-ES)/sqrt(DS)
> 1-pnorm(S.star)
 [1] 0.01556232
```

§2.2.2 无权重Cox-Stuart统计量

除了最优权重Cox-Stuart统计量以外，根据Cox-Stuart(1955)文献显示，还有两种无权重的Cox-Stuart统计量，分别记为S_2 和S_3，定义如表2.10所示：

表2.10 Cox-Stuart趋势检验方法汇总表

统计量表达式	效率	与S_1的$A.R.E.$比较	对y的分布要求
$S_1=\sum_{k=1}^{\lfloor N/2\rfloor}(N-2k+1)h_{k,N-k+1}$	$R^2(S_1)=\frac{N^3}{6\pi}$	$A.R.E.(S_1,S_1)=1$	正态
$S_2=\sum_{k=1}^{\lfloor N/2\rfloor}h_{k,\lfloor N/2\rfloor+k}$	$R^2(S_2)=\frac{N^3}{8\pi}$	$A.R.E.(S_2,S_1)=0.91$	与分布无关
$S_3=\sum_{k=1}^{\lfloor N/3\rfloor}h_{k,\frac{2}{3}N+k}$	$R^2(S_3)=\frac{4N^3}{27\pi}$	$A.R.E.(S_3,S_1)=0.96$	与分布无关

从表2.10中可以看到S_2与S_1的不同有两点：（1）.每个$h_{s,t}$的权重都相同；（2）.数据对的构成方式不同，修改了S_1 首尾相接的组对方式，替换成了间隔相等的顺序组对方式，组对的两个数据点之间的时间间隔由不等长更新为等长。而且数据序列间隔足够长的组对方式保证组对数据点彼此的独立性，这可以看成是对S_1 的改进。S_2 和S_3 的等权重计算方式简化了计算，可直接用符号检验解决该问题。虽然在效率上有所损失，但从表的第三列来看效率损失并不大，几乎和带权重的S_1 效率相当。S_3的用法是将数据截成三段，只使用序列最早一段和最末一段数据，S_3 是对S_2的改进，改进后的效率有所提升，而且S_2和S_3 都不依赖于正态分布，有更好的适用性。

下面我们以双边检验为例具体介绍S_2的用法.假设检验问题：

$$H_0:\ \text{数据序列无趋势} \leftrightarrow H_1:\text{数据序列有增长或减少趋势.}$$

假设数据序列$y_1,y_2,\cdots,y_n$独立, 在原假设之下, 同分布为$F(y)$, 令

$$c=\begin{cases} N/2, & \text{如果}N\text{ 是偶数,} \\ (N+1)/2, & \text{如果}N\text{ 是奇数.} \end{cases}$$

取y_i和y_{i+c} 组成数对(y_i, y_{i+c}). 当N 为偶数时, 共有N 对,当N为奇数时, 共有$N-1$ 对.计算每一数对前后两值之差: $D_i = y_i - y_{i+c}$. 用D_i的符号度量增减. 令S^+为正D_i 的数目, 令S^- 为负D_i 的数目, $S^+ + S^- = N', N' \leqslant N$.令$K = \min\{S^+, S^-\}$, 显然当正号太多或负号太多, 即$K$过小的时候, 有趋势存在.

在没有趋势的原假设下, K服从二项分布$B(N', 0.5)$, 该检验在某种意义上是符号检验的应用的拓展.

对于单边检验问题:

$$H_0: \text{ 数据序列有下降趋势} \leftrightarrow H_1: \text{数据序列有上升趋势},$$

$$H_0: \text{ 数据序列有上升趋势} \leftrightarrow H_1: \text{数据序列有下降趋势}.$$

结果是类似的, S^+很大时(或S^-很小时), 有下降趋势; 反之, S^+ 很小时(或S^-很大时), 有上升趋势.

和符号检验几乎类似, Cox-StuartS_2趋势检验过程总结于表2.11.

表2.11 Cox-Stuart S_2趋势检验

原假设: H_0	备择假设: H_1	检验统计量(K)	p值
H_0 : 无上升趋势	H_1 : 有上升趋势	$S^+ = \sum \text{sign}(D_i)$	$P(S^+ \leqslant k)$
H_0 : 无下降趋势	H_1 : 有下降趋势	$S^- = \sum \text{sign}(-D_i)$	$P(S^- \leqslant k)$
H_0 : 无趋势	H_1 : 有上升或下降趋势	$K = \min\{S^-, S^+\}$	$2P(K \leqslant k)$
小样本时, 用近似正态统计量$Z = (K \pm 0.5 - N'/2)/\sqrt{N'/4}$ $K < N'/2$时取减号, $K > N'/2$ 时取加号 大样本时, 用近似正态统计量$Z = (K - N'/2)/\sqrt{N'/4}$			
对水平α, 如果p值$< \alpha$,拒绝H_0;否则不能拒绝			

例2.6 某地区32年来的降雨量如表2.12所示:

表2.12 某地区20年来降雨量数据表

年份	1971	1972	1973	1974	1975	1976	1977	1978
降雨量/mm	206	223	235	264	229	217	188	204
年份	1979	1980	1981	1982	1983	1984	1985	1986
降雨量/mm	182	230	223	227	242	238	207	208
年份	1987	1988	1989	1990	1991	1992	1993	1994
降雨量/mm	216	233	233	274	234	227	221	214
年份	1995	1996	1997	1998	1999	2000	2001	2002
降雨量/mm	226	228	235	237	243	240	231	210

问: (1)该地区前10年降雨量是否有变化?

(2) 该地区32年来降雨量是否有变化?

解 (1) 假设检验问题：

$$H_0: \text{该地区前10年降雨量无趋势},$$

$$H_1: \text{该地区前10年降雨量有上升或下降趋势}.$$

分析：令$C = N/2 = 10/2 = 5$, 前后观测值如表2.13所示

表2.13 降雨量数据前后观察值

(y_1, y_6)	(y_2, y_7)	(y_3, y_8)	(y_4, y_9)	(y_5, y_{10})
(206,217)	(223,188)	(235,204)	(264,182)	(229,230)
$-$	$+$	$+$	$+$	$-$

本例中, 这5个数据对的符号为2负3正,取$K = \min\{S^+, S^-\}$, p值为$P(K \leqslant 2) = \frac{1}{2^{N'}} \sum_{i=0}^{2} \binom{N'}{i} = \frac{1}{2^5}(1+5) = 0.5 > \alpha = 0.05$, 于是表明该地区前10年的降雨量没有趋势. 这里的数据量太少, 一般来说要拒绝原假设是很困难的, 没有拒绝原假设, 也很难说问题出在什么地方.

(2) 这里的数据对增加到16个,如表2.14所示.

表2.14 降雨量数据的Cox-StuatS_2分析

206	223	235	264	229	217	188	204
216	233	233	274	234	227	221	214
$-$	$-$	$+$	$-$	$-$	$-$	$-$	$-$
182	230	223	227	242	238	207	208
226	228	235	237	243	240	231	210
$-$	$+$	$-$	$-$	$-$	$-$	$-$	$-$

这16个数据对的符号为2正14负. 取$K = \min\{S^+, S^-\}$, p值为$2P(K \leqslant 2) = \frac{1}{2^{N'}} \sum_{i=0}^{k} \binom{N'}{i} = 0.0021 < \alpha$, 对$\alpha = 0.05$, 可以拒绝原假设, 这表明该地区前32年的降雨量有明显的趋势. 为比较结果, 我们直接做线性回归模型, R程序如下：

```
data(rain);
year=seq(1971,2002);
anova(lm(rain~(year)))    #anova
function
```

```
     Analysis of Variance Table

Response: rain

          Df  Sum Sq Mean Sq F value Pr(>F)

year       1   535.4   535.4  1.5792 0.2186

Residuals 30 10170.1

339.0
```

结果表明数据的线性趋势并不显著, 数据的趋势图如图2.1所示.综合分析，

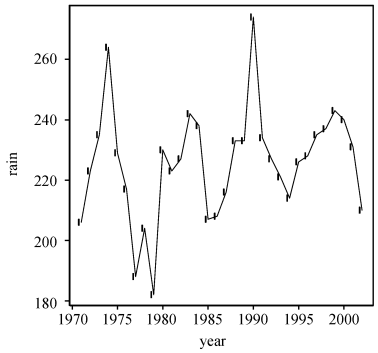

图 2.1 32年降雨量的变化趋势图

§2.3 随机游程检验

§2.3.1 两类随机游程检验

在实际中, 经常需要考虑一个序列中的数据出现是否与出现顺序无关,比如奖券的中奖是否随机出现，股票价格的变换是否随机，一个机械制程中产品故障的出现是否有规律，一个大型赛事中赢球是否有规律.若事件的发生并非随机的，而是有规律可循的，那么就可以做出相应的决策。在参数统计中, 研究这一问题相当困难, 要证明数据独立同分布则更难. 但是从非参数的角度来看,如果数据有上升或下降的趋势, 或有呈周期性变化的规律等特征,均可能表示数据与顺序是有关的, 或者说序列不是随机出现的.比如进出口的逆差和顺差是否随时间呈现某种规律, 一个机械流程中产品的次品出现是否存在一定的规律等.

这类问题一般伴随着一个二元0/1序列，感兴趣的是其中的1或0出现的顺序是否随机的问题.在一个二元序列中, 0 和1 交替出现.首先引入以下概念：在一个二元

序列中, 一个由0 或1 连续构成的串称为一个 **游程**, 一个游程中数据的个数称为 **游程的长度**. 一个序列中 **游程个数** 用R 表示, R表示0和1交替轮换的频繁程度. 容易看出, R是序列中0 和1交替轮换的总次数加1.

例2.7 在下面的0/1序列中, 总共有20个数, 0的总个数为$n_0 = 10$,1 的总个数为$n_1 = 10$. 共有4个0 游程, 4 个1 游程, 一共8 个游程($R = 8$).

$$1\ 0\ 0\ 0\ 0\ 1\ 1\ 1\ 0\ 1\ 1\ 0\ 0\ 0\ 0\ 1\ 1\ 1\ 1\ 0$$

如果0/1序列中0和1出现的顺序规律性不强, 随机性强, 则0和1的出现不会太集中, 也不会太分散. 换句话说, 可以通过0和1出现的集中程度度量序列随机性的强与弱. 我们注意到,如果不考虑序列的长度和序列中0/1的个数, 孤立地谈随机性意义不大.一个序列的顺序随机性是相对的, 只有固定了0和1的个数时才有意义.在固定序列长度n 时, n_1 表示序列中1 的个数, 如果游程个数过少,则说明0 和1 相对比较集中; 如果游程个数过多,则说明0 和1 交替周期特征明显, 这都不符合序列随机性要求.于是，这提供了一种通过游程个数过多或过少来判断序列非随性出现的可能性.

随机游程检验也称为Wald－Wolfowitz游程检验，是波兰的A·瓦尔德(Abraham Wald) 和J·沃夫维兹(Jacob Wolfowitz) 两位统计学家提出来的。关于这一问题的检验：设$X_1, X_2, \cdots, X_n$是一列由0或1 构成的序列,假设检验问题

$$H_0: \text{数据出现顺序随机} \leftrightarrow H_1: \text{数据出现顺序不随机}$$

设R为游程个数, $1 \leqslant R \leqslant n$. 在原假设成立的情况下, $X_i \sim B(1, p)$, p 是1 出现的概率, 由n_1/n 确定, R 的分布与p 有关. 假设有n_0个0和n_1个1, $n_0 + n_1 = n$,出现任何一种不同结构序列的可能性是$1\Big/\binom{n}{n_1} = 1\Big/\binom{n}{n_0}$, 注意到0游程和1游程之间最多差1, 于是得到$R$的条件分布为

$$\begin{aligned} P(R = 2k) &= \frac{2\begin{pmatrix} n_1 - 1 \\ k - 1 \end{pmatrix}\begin{pmatrix} n_0 - 1 \\ k - 1 \end{pmatrix}}{\begin{pmatrix} n \\ n_1 \end{pmatrix}}, \\ P(R = 2k + 1) &= \frac{\begin{pmatrix} n_1 - 1 \\ k - 1 \end{pmatrix}\begin{pmatrix} n_0 - 1 \\ k \end{pmatrix} + \begin{pmatrix} n_1 - 1 \\ k \end{pmatrix}\begin{pmatrix} n_0 - 1 \\ k - 1 \end{pmatrix}}{\begin{pmatrix} n \\ n_1 \end{pmatrix}}. \end{aligned}$$

建立了抽样分布, 根据分布公式就可以得出在H_0(即随机性)成立时$P(R \geqslant r)$或$P(R \leqslant r)$的值, 计算拒绝域进行检验. 这些值在n_0和n_1不大时可以计算或查表得出. 通常, 表中给出的水平$\alpha = 0.025,\ 0.05$ 及n_0, n_1 时临界值c_1 和c_2 的值, 满足$P(R \leqslant c_1) \leqslant \alpha$及$P(R \geqslant c_2) \leqslant \alpha$.

当数据序列的量很大时, 即当$n \to \infty$时,在原假设下,根据精确分布的性质可以得到:

$$E(R) = \frac{2n_1n_0}{n_1 + n_0} + 1,$$

$$\mathrm{var}(R) = \frac{2n_1n_0(2n_1n_0 - n_0 - n_1)}{(n_1 + n_0)^2(n_1 + n_0 - 1)}.$$

当$\dfrac{n_1}{n_0} \to \gamma$时,

$$E(R) = \frac{2n_1}{(1+\gamma)} + 1,$$

$$\mathrm{var}(R) \approx 4\gamma n_1/(1+\gamma)^3,$$

于是

$$Z = \frac{R - E(R)}{\sqrt{\mathrm{var}(R)}} = \frac{R - 2n_1/(1+\gamma)}{\sqrt{4\gamma n_1/(1+\gamma)^3}} \xrightarrow{\mathcal{L}} N(0,1).$$

因此可以用正态分布表得到p值和检验结果。这时,在给定水平α 后, 可以用近似公式得到拒绝域的临界值:

$$r_{\mathrm{l}} = \frac{2n_1n_0}{n_1 + n_0}\left[1 + \frac{Z_{\frac{\alpha}{2}}}{\sqrt{n_1 + n_0}}\right],$$

$$r_{\mathrm{u}} = 1 + \frac{2n_1n_0}{n_1 + n_0}\left[1 - \frac{Z_{\frac{\alpha}{2}}}{\sqrt{n_1 + n_0}}\right].$$

例2.8 超市一早开门营业，观察购物的男性和女性是否随机出现，记录下26位顾客到来的性别记录(用M表示男性,用F表示女性)依次如下：

M M F F F F F M M M M M F F F M M M M F F F F M M F

解 假设检验问题如下：

$$H_0:\ \text{男女出现顺序随机} \leftrightarrow H_1:\ \text{男女出现顺序不随机}.$$

统计分析：$n = 26, n_0 = 13, n_1 = 13, \alpha = 0.05$, 对于$n_0 = n_1$的情况，可以调用R里面的函数来直接分析，

```
library(tseries)#安装软件包
cusq=c(1,1,0,0,0,0,0,1,1,1,1,1,0,0,0,1,1,1,1,0,0,0,0,1,1,0)
#输入0/1数据
runs.test(cusq)
data: cusq
statistic = -2.4019, runs = 8, n1 = 13, n2 = 13, n = 26, p-value = 0.01631
alternative hypothesis: nonrandomness
```

结论：由于实际观测值为$R=8, p$值很小，拒绝原假设.北京退休的大爷大妈们常常三五成群晨练，晨练结束后三三两两有说有笑地顺便去超市赶着肉菜刚上架的生鲜时机把一家人一天的餐桌所需买回家，这样就出现了性别上三五成群结伴进超市的壮观景象，对超市卖场而言也形成了一个有规律的小小的早高峰。

例2.9 在试验设计中, 经常关心试验误差(experiment error)是否与序号无关. 假设有A,B,C 三个葡萄品种,采取完全试验设计，每个品种需要重复测量4次, 安排在12块试验田中栽种,共得到12 组数据, 每块试验田试验结果收成如表2.15所示.试问：误差分布是否按序号随机?

表2.15 12块试验田试验收成表(单位：kg)

(1) B	(2) C	(3) B	(4) B	(5) C	(6) A	(7) A	(8) C	(9) A	(10) B	(11) C	(12) A
23	24	18	23	19	11	6	22	14	22	27	15

解 假设检验：

H_0 : 试验误差在试验田序号上随机出现 $\leftrightarrow H_1$: 试验误差随试验田序号不随机.

如式(4.1)所示, 完全随机设计观测值

$$x_{ij}=\mu+\mu_i+\varepsilon_{ij}=\bar{x}_{..}+(\bar{x}_{i.}-\bar{x}_{..})+(x_{ij}-\bar{x}_{i.})$$
$$i=1,2,\cdots,k; j=1,2,\cdots,n.$$

试验误差为$\varepsilon_{ij}=x_{ij}-\bar{x}_{i.}$, 首先计算每个品种的均值$\bar{A}=11.5, \bar{B}=21.5, \bar{C}=23$, 各试验田实际收成与各自误差成分之间出现顺序为正和负的记录为

表2.16 误差成分正负记录表

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
+	+	−	+	−	−	−	−	+	+	+	+

统计分析：现在$n=12, n_1=7, n_0=5, \alpha=0.05$,查出$r_{\mathrm{l}}=3, r_{\mathrm{u}}=11$.

结论：由于实际观测值为$3<r=5<11$, 因此不能拒绝原假设.

对于连续型数据, 也关心数据是否随机出现,这时可以将连续的数据二元化, 将连续数据的随机性问题转化成二元数据的离散化问题, 这是Mood于1940年给出的中位数检验法. 看下面的例子.

例2.10 实习生在实习期迟到的情况被门镜系统记录下来, N表示正常, F表示迟到, 根据表2.17的记录判断这名学生迟到是否随机($\alpha=0.10$).

表2.17 实习生迟到情况统计表

1	2	3	4	5	6	7
NNN	FF	NNNNNNN	F	NN	FFF	NNNNNN
8	9	10	11	12		13
F	NNNN	FF	NNNNN	F		NNNNNNNNNNNN

解 假设检验问题：

$$H_0:\ \text{该实习生迟到是随机的} \leftrightarrow H_1:\ \text{该实习生迟到不随机}.$$

本例中$n_1=40, n_0=10, R=13$, 根据超几何分布, 计算R在大样本下的近似正态分布均值和方差如下：

$$\begin{aligned} E_R &= \frac{2n_1n_0}{n_1+n_0}+1=17, \\ \mathrm{Sd}_R &= \sqrt{\frac{2n_1n_0(2n_1n_0-n_0-n_1)}{(n_1+n_0)^2(n_1+n_0-1)}}=2.213, \\ Z &= \frac{R-E_R}{\mathrm{Sd}_R}=-1.81. \end{aligned}$$

取$\alpha=0.10, -1.64>Z=-1.81$, 于是可以认为这名学生迟到违反随机性,有一定的规律。鉴于游程数小于平均数，这表明该学生在实习前期迟到状况频繁出现，而在实习后期迟到习惯有明显的改善。

§2.3.2 三类及多类游程检验

有时候会碰上三类或多类的问题，比如足球比赛有赢球、输球和平局三种比赛结果状态，如果要看比赛结果随比赛进程是否有规律，可以用多值游程检验。假设一串游程有k 个不同的值，每类的数据量分别记为$n_1, n_2, \cdots, n_k, \sum_{i=1}^{k} n_i = n$. $p_i=\frac{n_i}{n}$,D.E. 巴顿(Barton D.E.) 和F.N. 戴维(David F.N.)(1957) 提出了可以用近似

正态的方法来解决三类或多类游程检验问题，可以证明游程数有期望和方差如下：

$$\mathrm{E}(R) = n\left(1 - \sum_{i=1}^{k} p_i^2\right) + 1 \tag{2.19}$$

$$\mathrm{var}(R) = n\left[\sum_{i=1}^{k}(p_i^2 - 2p_i^3) + (\sum_{i=1}^{k} p_i^2)^2\right] \tag{2.20}$$

于是可以通过

$$Z = \frac{R - \mathrm{E}(R)}{\sqrt{\mathrm{var}(R)}}$$

来检验，经验指出，当$n > 12$ 时，Z 值过大或过小拒绝原假设。

例2.11 15支足球队通过积分赛制争夺冠军，分析冠军队在14场比赛中的成绩，有人说其冠军相不明显，有人说其获得冠军偶然性较大，请问根据表2.8的比赛成绩，能否判断获胜是遵循一定规律的还是随机的？其中W 表示赢得比赛，D 表示平局，L 表示输掉比赛，$\alpha = 0.10$.

表2.18 比赛获胜队统计表

1	2	3	4	5	6	7	8	9	10	11	12	13	14
W	W	W	W	D	D	W	W	L	L	L	L	L	W

解 假设检验问题：

$$H_0 : \text{足球队赢球是随机的} \leftrightarrow H_1 : \text{足球队赢球不是随机的.}$$

本例中，不妨记n_W为赢球场数$n_W = 7$，n_L为输球场数$n_L = 5$,n_D为平局场数$n_D = 2$, $n = n_W + n_L + n_W = 14$ 根据式(2.7) 和式(2.8), 计算赢球概率p_W=7/14，输球概率p_L=5/14,平局概率p_D=2/14。根据式(2.7)和式(2.8)，R 在大样本下的近似正态分布均值和方差如下：

$$E_R = n \times (1 - p_W^2 - p_L^2 - p_D^2) = 14 \times \left[1 - \left(\frac{1}{2}\right)^2 - \left(\frac{5}{14}\right)^2 - \left(\frac{2}{14}\right)^2\right] + 1 = 9.43,$$

$$\mathrm{Sd}_R = \sqrt{n \times ((p_{\mathrm{W}}^2 + p_{\mathrm{L}}^2 + p_{\mathrm{D}}^2 - 2p_{\mathrm{W}}^3 - 2p_{\mathrm{L}}^3 - 2p_D^3) + (p_{\mathrm{W}}^2 + p_{\mathrm{L}}^2 + p_{\mathrm{D}}^2)^2)} = 1.71,$$

$$Z = \frac{R - E_R}{\mathrm{Sd}_R} = -2.59.$$

取$\alpha = 0.10, -1.64 > Z = -2.59$, 于是可以认为该冠军球队获胜是有一定规律。因为游程数小于平均数，表明该冠军队获胜的原因是前期赢球数较多，总的不输场次占到近7 成，保证了后程遭遇强劲对手时虽屡屡落败但依然顽强坚持下来并赢得最后两场比赛的胜利，巩固了前程的总积分，最终登上了冠军的领奖台，这是在积分赛制规则下摆脱弱队晋升强队所必备的战术“先打分散和孤立的弱队争取积分上的领先，再集中士气攻克实力强大的劲敌”。

§2.4 Wilcoxon符号秩检验

§2.4.1 基本概念

Wilcoxon秩和检验是由美国化学家统计学家F.威尔科克森（Frank • Wilcoxon）于1945年提出来的用于单变量分布位置的检验。前几节的统计推断都只依赖数据的符号,这样一类方法对连续分布的形态没有要求. 本节主要讨论对称分布,研究对称分布的位置具有普遍意义. 原因是,许多不对称的单峰数据分布可能通过变换化为对称分布.多峰分布通过混合分布整体表示后,每一个分布也可以用单峰对称分布表示. 就对称分布而言,对称中心只有一个, 中位数却可能有很多. 下面的定理指出, 对称分布的对称中心是总体分布的中位数之一. 毫无疑问,对称中心是比中位数更重要的位置. 因此, 作为总体的对称中心,有两点需要考虑：

(1) 由于对称中心是中位数,因此在对称中心的两侧应大致各有一半左右的数据量;

(2) 在对称中心的两侧,数据的分布疏密程度应类似.

这时, 只考虑数据的符号就不够了, 作为刻画数据中心位置的对称中心,要求数据在其两边分布的疏密情况是对称的. 不仅如此, 如果对称分布的中位数唯一, 则中位数就是对称中心,中位数与期望是一致的. 因此, 就对称分布而言, 可以比较不同统计量的检验效率, 继而从理论上比较参数方法和非参数方法的效率.

首先,给出对称分布的一些记号如下：称连续分布$F(x)$关于θ对称,如果对$\forall x \in \mathbb{R}$, $F(\theta - x) = P(X < \theta - x) = P(X > \theta + x) = 1 - F(x + \theta)$, 此时称$\theta$是分布的**对称中心**.

定理2.1 X服从分布函数为$F(\theta)$的分布, 且$F(\theta)$关于θ对称, 总体的对称中心是总体的中位数之一.

证明 对称分布X有对称中心θ, 那么$X - \theta$与$\theta - X$关于零点对称, 而且有相同的分布：

$$\forall x, P(X - \theta < x) = P(\theta - X < x).$$

特别地, 取$x = 0$, 则

$$P(X < \theta) = P(X > \theta) \Rightarrow P(X < \theta) \leqslant \frac{1}{2}.$$

以下证明$P(X \leqslant \theta) \geqslant \frac{1}{2}$. 应用反证法, 如果$P(X \leqslant \theta) < \frac{1}{2}$, 那么

$$P(X > \theta) = P(X < \theta) = 1 - P(X \leqslant \theta) > \frac{1}{2}.$$

这与上面结论矛盾, 综合两者, 有

$$P(X < \theta) \leqslant \frac{1}{2} \leqslant P(X \leqslant \theta).$$

即θ是X的一个中位数.

先看一个例子：对数据

$$-3.56 \quad -2.22 \quad -0.31 \quad -0.14 \quad 11.12 \quad 12.30 \quad 14.1 \quad 14.3$$

来说, 0是这组数据的中位数, 两侧有相等数量的正号和负号; 如果只看秩,而不看数据的取值, 直觉上认为这是一个以0为中心的样本. 但实际上,取负值的数据相对在0左侧聚集, 取正值的数据并非在0值右侧相等距离的位置聚集，而是在较远处的10 附近比较集中,而左侧距离0 间隔相当的位置上负值也不密集。这不满足对称性要求在对称中心两边的分布相同的特点. 为什么符号法失败了? 问题出在没有考虑数据绝对值的大小上. Wilcoxon符号秩统计量的思想是,首先把样本的绝对值$|X_1|,|X_2|,\cdots,|X_n|$ 排序. 其顺序统计量为$|X|_{(1)},|X|_{(2)},\cdots,|X|_{(n)}$. 如果数据关于零点对称, 对称中心两侧关于对称中心距离相等的位置上数据的疏密情况应该大致相同.这表现为, 当数据取绝对值以后,原来取正值的数据与原来取负值的数据交错出现,取正值数据在绝对值样本中的秩和与取负值数据在绝对值样本中的秩和应近似相等.

具体而言，用R_j^+表示$|X_j|$在绝对值样本中的秩, 即$|X_j|=|X|_{(R_j^+)}$. 如果用$S(x)$表示示性函数$I(x>0)$, 它在$x>0$ 时为 1 , 否则为 0 . 为方便起见, 我们引入 **反秩**(antirank)的概念.反秩D_j是由$|X_{D_j}|=|X|_{(j)}$ 定义的.我们还用W_j 表示与$|X|_{(j)}$ 相应的原样本点的符号函数,即$W_j=S(X_{D_j})$, 且称$R_j^+S(X_j)$为符号秩统计量. Wilcoxon **符号秩统计量** 定义为

$$W^+=\sum_{j=1}^{n}jW_j=\sum_{j=1}^{n}R_j^+S(X_j).$$

它是正的样本点按绝对值所得秩的和. 为说明这些概念, 看如下例子.

例2.12 如样本值为$9,13,-7,10,-18,4$, 则相应的统计量值如表2.16所示。

表2.19 符号秩统计量取值

X_1	X_2	X_3	X_4	X_5	X_6
9	13	-7	10	-18	4
$\|X\|_{(3)}$	$\|X\|_{(5)}$	$\|X\|_{(2)}$	$\|X\|_{(4)}$	$\|X\|_{(6)}$	$\|X\|_{(1)}$
$R_1^+=3$	$R_2^+=5$	$R_3^+=2$	$R_4^+=4$	$R_5^+=6$	$R_6^+=1$
$W_3=1$	$W_5=1$	$W_2=0$	$W_4=1$	$W_6=0$	$W_1=1$
$D_3=1$	$D_5=2$	$D_2=3$	$D_4=4$	$D_6=5$	$D_1=6$

显然，$W^+=3+5+4+1=13$.

设$F(x-\theta)$对称, 原假设为$H_0:\theta=0$, 有下面3 个定理.

定理2.2 如果原假设$H_0:\theta=0$成立, 则$S(X_1),S(X_2),\cdots,S(X_n)$独立于$(R_1^+,R_2^+,\cdots,R_n^+)$.

证明 事实上, 因为$(R_1^+,R_2^+,\cdots,R_n^+)$是$|X_1|,|X_2|,\cdots,|X_n|$的函数, 而出自随机样本的$(S(X_i),|X_j|),(i,j=1,2,\cdots,n,j\neq i)$是互相独立的数据对,因此我们只要证明$S(X_i)$和$|X_i|$是互相独立的即可, 事实上,

$$\begin{aligned}P[S(X_i)=1,|X_i|\leqslant x]&=P(0<X_i\leqslant x)=F(x)-F(0)=F(x)-\frac{1}{2}\\&=\frac{2F(x)-1}{2}=P[S(X_i)=1]P(|X_i|\leqslant x).\end{aligned}$$

下面的定理2.3和定理2.4平行, 读者可自己验证.

定理2.3 如果原假设$H_0:\theta=0$成立, 则$S(X_1),S(X_2),\cdots,S(X_n)$独立于$(D_1,D_2,\cdots,D_n)$.

定理2.4 如果原假设 $H_0:\theta=0$成立, 则$W_1,W_2,\cdots,W_n$是独立同分布的, 其分布为$P(W_i=0)=P(W_i=1)=\frac{1}{2}$.

证明 令$\boldsymbol{D}=(D_1,D_2,\cdots,D_n),\boldsymbol{d}=(d_1,d_2,\cdots,d_n)$,

$$\begin{aligned}&P(W_1=w_1,W_2=w_2,\cdots,W_n=w_n)\\&=\sum_d P[S(X_{D_1})=w_1,S(X_{D_2})=w_2,\cdots,S(X_{D_n})=w_n|\boldsymbol{D}=\boldsymbol{d}]P(\boldsymbol{D}=\boldsymbol{d})\\&=\sum_d P[S(X_{d_1})=w_1,S(X_{d_2})=w_2,\cdots,S(X_{d_n})=w_n]P(\boldsymbol{D}=\boldsymbol{d})\\&=\left(\frac{1}{2}\right)^n\sum_d P(\boldsymbol{D}=\boldsymbol{d})=\left(\frac{1}{2}\right)^n.\end{aligned}$$

因此有$P(W_1,W_2,\cdots,W_n)=\prod_{i=1}^n P(W_i=w_i)$ 及$P(W_i=w_i)=\frac{1}{2}$.

§2.4.2 Wilcoxon符号秩检验和抽样分布

1. Wilcoxon符号秩检验过程

假设样本点$X_1,X_2,\cdots,X_n$来自连续对称的总体分布(符号检验不需要这个假设). 在这个假定下总体中位数等于均值.它的检验目的和符号检验是一样的, 即要检验双边问题$H_0:M=M_0$或检验单边问题$H_0:M\leqslant M_0$及$H_0:M>M_0$, Wilcoxon符号秩检验的步骤如下.

(1) 对$i=1,2,\cdots,n$, 计算$|X_i-M_0|$; 它们表示这些样本点到M_0 的距离.

(2) 将上面n 个绝对值排序, 并找出它们的n个秩; 如果有相同的样本点, 每个点取平均秩.

(3) 令W^+ 等于$X_i-M_0>0$ 的$|X_i-M_0|$ 的秩的和, 而W^- 等于$X_i-M_0<0$ 的$|X_i-M_0|$ 的秩的和. 注意：$W^++W^-=n(n+1)/2$.

(4) 对双边检验$H_0 : M = M_0 \leftrightarrow H_1 : M \neq M_0$, 在原假设下, W^+ 和W^-应差不多. 因而, 当其中之一很小时, 应怀疑原假设; 在此, 取检验统计量$W = \min\{W^+, W^-\}$. 类似地, 对$H_0 : M \leqslant M_0 \leftrightarrow H_1 : M > M_0$ 的单边检验取$W = W^-$; 对$H_0 : M \geqslant M_0 \leftrightarrow H_1 : M < M_0$ 的单边检验取$W = W^+$.

(5) 根据得到的W值,查Wilcoxon符号秩检验的分布表以得到在原假设下的p值. 如果n 很大,要用正态近似得到一个与W 有关的正态随机变量Z的值, 再查表得到p值. 或直接在软件中计算得到p值.

(6)如果p值小(比如小于或等于给定的显著性水平0.05),则可以拒绝原假设. 实际上,显著性水平α可取任何大于或等于p值的数. 如果p 值较大, 则没有充分证据来拒绝原假设, 但不意味着接受原假设.

2. W^+在原假设下的精确分布

W^+在原假设下的分布并不复杂.我们举一个例子说明如何在简单情况下获得其分布. 当$n = 3$时, 绝对值的秩只有1, 2和3, 但是却有8种可能的符号排列. 在原假设下, 每一个这种排列都是等概率的(在这里, 其概率为1/8). 表2.20列出了这些可能的情况以及在每种情况下W^+的值. 可以看出, $W^+ = 3$出现了两次, 因而$P_{H_0}(W^+ = 3) = 2/8$, 其余W^+为0, 1, 2, 4, 5, 6 六个数中之一的概率为1/8.

表2.20 Wilcoxon分布列计算表

秩	符号的8种组合							
1	−	+	−	−	+	+	−	+
2	−	−	+	−	+	−	+	+
3	−	−	−	+	−	+	+	+
W^+	0	1	2	3	3	4	5	6
概率	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$

现在, 给出计算W^+概率的一般方法. 首先, $\forall j$ 有

$$E[\exp(t_j W_j)] = \frac{1}{2}\exp(0) + \frac{1}{2}\exp(t_j) = \frac{1}{2}[1 + \exp(t_j)].$$

当计算样本量为n时, W^+的母函数如下:

$$\begin{aligned} M_n(t) &= E[\exp(tW^+)] = E[\exp(t\sum jW_j)] \\ &= \prod_j E[\exp(tjW_j)] = \frac{1}{2^n}\prod_{j=1}^{n}(1 + \mathrm{e}^{tj}). \end{aligned}$$

母函数有展开式

$$M_n(t) = a_0 + a_1\mathrm{e}^t + a_2\mathrm{e}^{2t} + \cdots,$$

则$P_{H_0}(W^+=j)=a_j$. 利用指数相乘的性质, 当$n=2$ 时, 列表如下：

表2.21 n=2时Wilcoxon分布列计算表

0	1	2	3
1	1	1	1

第一行表示$M_2(t)$的各个指数幂, 第二行是这些幂对应的系数(忽略除数2^2).

当$n=3$时,我们可以从上面的表中通过移位和累加得到指数幂的系数,如下表所示(忽略除数2^3)：

表2.22 n=3时Wilcoxon分布列计算表

0	1	2	3	4	5	6	
1	1	1	1				
			1	1	1	1	+
1	1	1	2	1	1	1	

表2.22中第一行是$M_3(t)$的指数幂; 第二行是$M_2(t)$的指数幂对应的系数; 第三行是第二行的系数右移三位, 是由第三个因子的第二项(e^{3t})乘前面各项得到的, 因为是三次幂, 所以右移三位; 第四行是二、三两行的和. 由此得到$P(W^+=k)$. 类似可通过递推的方法得任意n时的W^+的分布如下：

表2.23 任意n时Wilcoxon分布列计算表

0	1	2	⋯	$\frac{n(n+1)}{2}$	
	$M_{n-1}(t)$的系数				
	(右移$\to n$位)		$M_{n-1}(t)$的系数		+
		$M_n(t)$的系数			

下面的函数dwilxonfun用来计算W^+的分布密度函数, 即$P(W^+=x)$的一个R参考程序, 其中N是样本量.

```
dwilxonfun=function(N)
{
  a=c(1,1)# when n=1 frequency of W+=1 or 0
  n=1
  pp=NULL # distribute of all size from 2 to N
  aa=NULL # frequency of all size from 2 to N
  for (i in 2:N)
  {
```

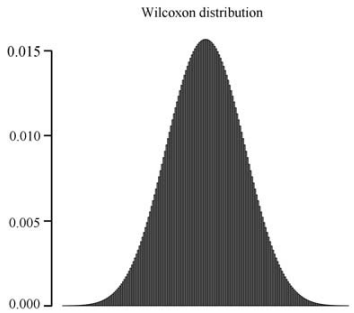

图 2.2 Wilcoxon分布图

```
    t=c(rep(0,i),a)
    a=c(a,rep(0,i))+t
    p=a/(2^i)     #density of wilcox distribut when size=N
    }
  p
 }
N=19
# sample size of expected distribution of W+
dwilxonfun(N)
```

Wilcoxon分布如图2.2所示.

3. 大样本W^+分布

如同对符号检验讨论的那样, 如果样本量太大, 则可能得不到分布表, 这时可以使用正态近似. 根据2.3节的定理, 可以得到

$$E(W^+) = E\left(\sum_{j=1}^{n} jW_j\right) = \frac{1}{2}\sum_{j=1}^{n} j = \frac{1}{2}\frac{n(n+1)}{2} = \frac{1}{4}n(n+1),$$

$$\mathrm{var}(W^+) = \mathrm{var}\left(\sum_{j=1}^{n} jW_j\right) = \frac{1}{4}\sum_{j=1}^{n} j^2 = \frac{1}{24}n(n+1)(2n+1).$$

在原假设下由此可构造大样本渐近正态统计量, 原假设下的近似计算如下：

$$Z = \frac{W^+ - n(n+1)/4}{\sqrt{n(n+1)(2n+1)/24}} \xrightarrow{\mathcal{L}} N(0,1).$$

计算出Z 值后, 可由正态分布表查出检验统计量对应的p值, 如果p值过小, 则拒绝原假设$H_0 : \theta = M_0$. 在小样本情况下使用连续性修正如下：

$$Z = \frac{W^+ - n(n+1)/4 \pm C}{\sqrt{n(n+1)(2n+1)/24}} \xrightarrow{\mathcal{L}} N(0,1).$$

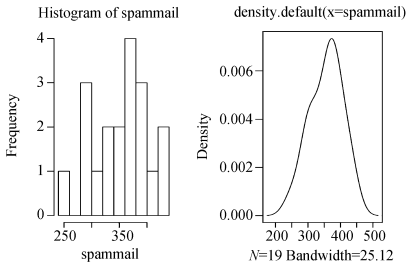

图 2.3 垃圾邮件的直方图和分布密度曲线图

当$W^+ > n(n+1)/4$时, 用正连续性修正, $C = 0.5$; 当$W^+ < n(n+1)/4$时, 用负连续性修正, $C = -0.5$.

如果数据有g个结, 在小样本情况下可以用正态近似公式：

$$Z = \frac{W^+ - n(n+1)/4 \pm C}{\sqrt{n(n+1)(2n+1)/24 - \sum_{i=1}^{g}(\tau_i^3 - \tau_i)/48}} \xrightarrow{\mathcal{L}} N(0,1).$$

在大样本情况下, 用正态近似公式：

$$Z = \frac{W^+ - n(n+1)/4}{\sqrt{n(n+1)(2n+1)/24 - \sum_{i=1}^{g}(\tau_i^3 - \tau_i)/48}} \xrightarrow{\mathcal{L}} N(0,1).$$

计算出Z值以后, 查正态分布表对应的p值. 如果p值很小, 则拒绝原假设.

下面举例说明如何应用Wilcoxon符号秩和检验, 并将它与符号检验的结果相比较, 分析在解决位置参数检验问题时各自的特点.

例2.13 为了解垃圾邮件对大型公司决策层工作影响程度, 某网站收集了19家大型公司的CEO和他们邮箱里每天收到的垃圾邮件数, 得到如下数据(单位：封)：

310 350 370 377 389 400 415 425 440 295
325 296 250 340 298 365 375 360 385

从平均意义上来看, 垃圾邮件数量的中心位置是否超出320 封?

解 首先, 我们先作数据的直方图, 如图2.3所示. 在直方图上, 没有明显的迹象表明数据的分布不是对称的, 因此采用R内置函数wilcox.test来假设检验：

$$H_0 : \theta = M_0 \leftrightarrow \quad \theta \neq M_0.$$

R程序和输出如下：

```
    wilcox.test(spammail-320)

        Wilcoxon signed rank test

data:  spammail - 320 V = 158, p-value = 0.009453 alternative
hypothesis: true location is not equal to 0
```

为方便比较, 下面采用binom.test函数进行参数位置的检验, R程序和输出如下:

```
    > suc.num<-sum(spammail>320)
    > n=length(spammail)
    > binom.test(suc.num,n,0.5)
    Exact binomial test
data:  sum(spammail > 320) and 19 number of successes = 14, number
of trials = 19, p-value = 0.06357 alternative hypothesis: true
probability of success is not equal to 0.5 95 percent confidence
interval:
 0.4879707 0.9085342
sample estimates: probability of success
             0.7368421
```

其中, suc.num表示数据中大于320的样本数. 从结果看, 虽然两个检验都拒绝了原假设, 但是wilcox.test输出的p值比binom.test小一些, 这表明在对称性的假定之下, Wilcoxon符号秩检验采用了比符号检验更多的信息, 因而可能得到更可靠的结果. 值得注意的是, 这里假定了总体分布的对称性. 如果对称性不成立, 则还是符号检验的结果更为可靠.

4. 由Wilcoxon符号秩检验导出的Hodges-Lehmann估计量

定义2.1 假设$X_1, X_2, \cdots, X_n$为简单随机样本, 计算任意两个数的平均, 将得到一组长度为$\dfrac{n(n+1)}{2}$的新的数据. 这组数据称为Walsh平均值, 即$\left\{ X'_u : X'_u = \dfrac{X_i + X_j}{2},\ i \leqslant j, u = 1, 2, \cdots, \dfrac{n(n+1)}{2} \right\}$.

定理2.5 由前面定义的Wilcoxon符号秩统计量W^+可以表示为

$$W^+ = \# \left\{ \frac{X_i + X_j}{2} > 0, \quad i \leqslant j; i, j = 1, 2, \cdots, n \right\}.$$

即W^+是Walsh平均值中符号为正的个数.

证明 记$X_{i_1}, X_{i_2}, \cdots, X_{i_p}$为$p$个正的样本点, 以原点为中心, 以$X_{i_1}$为半径画闭区间$I_1 = [-X_{i_1}, X_{i_1}]$. X_{i_1}绝对值的秩$R_{i_1}^+$等于在I_1中的样本点的个数. 注意到：I_1中样本点和X_{i_1}构成的平均值都大于0. 将这个过程对每一个样本点重复一遍, 就得到了所有秩和, 这些秩和恰好为Walsh平均值大于0的个数.

如果中心位置不是0, 而是θ, 则定义统计量如下：

$$W^+(\theta) = \#\left\{\frac{X_i + X_j}{2} > \theta, i \leqslant j; i, j = 1, 2, \cdots, n\right\}.$$

用$W^+(\theta)$作为检验$H_0 : \theta = \theta_0 \leftrightarrow H_1 : \theta \neq \theta_0$的统计量, 则这个检验是无偏检验.

定义2.2 假设$X_1, X_2, \cdots, X_n$独立同分布取自$F(x-\theta)$, 若F对称, 则定义Walsh平均值的中位数如下：

$$\hat{\theta} = \text{median}\left\{\frac{X_i + X_j}{2}, i \leqslant j; i, j = 1, 2, \cdots, n\right\},$$

并将其作为θ的Hodges-Lehmann点估计量.

例2.14 一个食物研究所在检测某种香肠的肉含量时, 随机测出如表2.24所示的数据。

表2.24 香肠肉含量（%）

62	70	74	75	77	80	83	85	88

解 假定分布是对称的, Walsh平均的数据量记为$NW, NW = \dfrac{n(n+1)}{2} = 45$, 可以用下面的R程序计算中心位置的点估计：

```
> a <-c(62,70,74,75,77,80,83,85,88)
> walsh <-NULL
> for (i in 1:(length(a)-1))
> for (j in (i+1):length(a))
> walsh <-c(walsh,(a[i]+a[j])/2)
> walsh=c(walsh,a)
> NW=length(walsh)
> median(walsh)
> 77.5
```

§2.5 估计量的稳健性评价

这一节主要介绍估计量的稳健性（robustness）评价准则。稳健性的英文是"robust", 用于评价当观测和分布发生微小变化时，估计量会不会受到太大影响。稳健性概

念首先由博克斯(Box,1953)提出，后来被汉佩尔(Hampel) 和休泊(Huber)不断发展起来。约翰.图基(Tukey,1960)给出了一个例子:假设有n 个观测$Y_i \sim N(\mu,\sigma^2)(i=1,2,\cdots,n)$，目标是估计$\sigma^2$。有两个估计量：一个估计量是$\hat{\sigma}^2 = s^2$,另一个估计量是$\tilde{\sigma}^2 = d^2\pi/2$，其中

$$d = \frac{1}{d}\sum_i |Y_i - \bar{Y}|.$$

由于$d \to \sqrt{2/\pi}\sigma$，于是$\tilde{\sigma}^2$是σ^2 的渐进无偏估计。而且可以得到ARE$(\tilde{\sigma}^2, s^2) = 0.876$. 图基进一步指出，如果$Y_i$ 以$1-\epsilon$从$N(\mu,\sigma^2)$ 中取得，以一个很小的概率ϵ 从$N(\mu, 9\sigma^2)$ 中取得，ARE会呈现出如表2.25所示的变化：

表2.25 A.R.E.随ϵ变化表

ϵ%	0	0.1	0.2	1	5
ARE$(\tilde{\sigma}^2, s^2)$	0.876	0.948	1.016	1.44	2.04

从表中可以看出，估计量s^2的优越性仅仅表现在噪声数据不足1%的场景中，这表明在实际中s^2 作为估计量对噪声数据的免疫力十分脆弱，它缺乏效率上的稳健性(robustness of efficiency)。

常用的稳健性评价准则有三类：敏感曲线（sensitivity curve）、影响函数(influence function, IF) 及其失效点(breakdown point, BP)。

1.敏感曲线

假设θ是待估分布函数F的参数，观测$\mathbf{x} = \{x_1, x_2, \cdots, x_n\}$ 来自分布F, $\hat{\theta}_n$是θ的估计量，如果增加一个异常点x,形成一个新数据$\mathbf{x} = \{x_1, x_2, \cdots, x_n, x\}$,那么估计量变成$\hat{\theta}_{n+1}$, 敏感曲线定义为：

$$S(x,\hat{\theta}) = \frac{\hat{\theta}_{n+1} - \hat{\theta}_n}{1/n+1}$$

敏感曲线用于评估估计量受外界异常干扰的能力。

例2.15 取例2.1数据的前13例数据，取异常值在[1,15] 和[60,80]之间，每次增加一个异常值，比较三种估计量样本均值。样本中位数和HL 估计量（Hodges-Lehmann）数值随异常数据的变化，制作敏感曲线分析，分析结果如图2.4 所示。

从图2.4来看，样本均值是无界的，中位数和HL估计是有界的，有界的含义是当数据受到轻度干扰时，估计值还是会稳定在一个固定值附近。

2.影响函数

敏感曲线依赖于观测数据,另一种评价估计量稳健性能的方法只依赖于分布，称为影响函数，由汉佩尔（Hampel，1974）首先提出。它表示给定分布F的一个样本，在任意点x处加入一个额外观测后对统计量T (近似或标准化)的影响。具体而

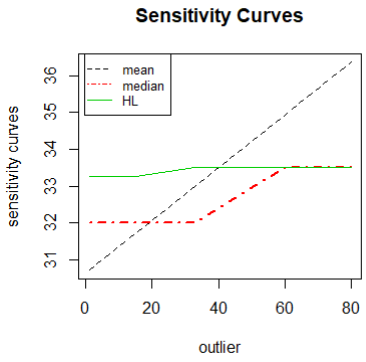

图 2.4 三类估计量的敏感曲线对比图

言，如果X以$1-\epsilon(0 \leqslant \epsilon \leqslant 1)$ 的概率来自既定分布F，而以ϵ 的概率来自另一个任意污染分布δ_x，此时的混合分布表示为：

$$F_{x,\delta} = (1-\epsilon)F + \epsilon\delta_x.$$

统计量T的影响函数就定义为：

$$IF(x,T,F) = \lim_{\epsilon\to 0} \frac{T[(1-\epsilon)F + \epsilon\delta_x] - T(F)}{\epsilon}$$

从定义来看，影响函数$IF(x,T,F)$是统计量T在一个既定分布F下的一阶导数，其中点x是有限维的概率分布空间中的坐标。如果某个统计量的IF值有界，就称该统计量对微小污染具有稳健性。

一个估计量称为稳健的，如果它的影响函数是有界的。在比例和中心都不变的情况下，样本均值的影响函数是x, 中位数的影响函数是$\text{sign}(x)$, HL 的影响函数是$F(x)-0.5$, 这么来看, 样本均值不是稳健的, 中位数和HL估计都是稳健的。

3.失效点

失效点是一种全局稳健性评价方法。一般意义下，失效点(BP)是指：原始数据中混入了异常数据时，在估计量给出错误模型估计之前，异常数据量相对于原始数据量的最大比例。失效点是估计量对异常数据的最大容忍度。汉佩尔(Hampel,1968)给出了失效点的近似求解方法。多纳赫和休伯(Donoho &Huber 1983) 提出了一种回归分析下失效点的定义，这个定义在有限样本条件下是这样给出的：

$$\epsilon_n^*(\hat{\beta}, Z) = \min\{\frac{m}{n}, \text{bias}(m, \hat{\beta}, Z) \to \infty\}.$$

其中Z为自变量与因变量组成的观测值空间，$\hat{\beta}$ 为回归估计向量，偏差函数bias 表示从Z 空间上n个观测中，将其中任意m个值做任意大小替换后(即考虑了最坏情况

下有离群点的情况)，导致回归估计不可用时所能允许的替换样本量的最小值，记为ABP(Asymptotic Breakdown Point).回归估计的失效点表示的是导致估计值$\hat{\beta}$ 无意义过失误差的额外样本量的最小比例。从定义来看，它衡量的是偏离模型分布的一个距离，超过该距离统计量就变得完全不可靠。BP值越小估计值越不稳健。

样本均值的失效点为$BP = \frac{1}{n}, ABP = 0$。因为使任意一个$x_i$变成足够大的数据之后，估计出来的均值，就不再正确了，渐近失效点为0。样本中位数的失效点数为$\frac{n-1}{2}, ABP = 0.5$。HL估计量的失效点数为$ABP = 0.29$。

§2.6 单组数据的位置参数置信区间估计

§2.6.1 顺序统计量位置参数置信区间估计

1. 用顺序统计量构造分位数置信区间的方法

在参数的区间估计中, 可以通过样本函数构造随机区间, 使该区间包括待估参数的可能性达到一定可靠性. 如果待估的参数就是分位数点m_p, 则自然想到用样本的顺序统计量构造区间估计.

令样本$X_1, X_2, \cdots, X_n$独立取自同一分布$F(x)$, $X_{(1)}, X_{(2)}, \cdots, X_{(n)}$是样本的顺序统计量, 对$\forall i < j$, 注意到：

$$P(X_i < m_p) = p, \quad \forall i = 1, 2, \cdots, n.$$

$$\begin{aligned} & P(X_{(i)} < m_p < X_{(j)}) \\ =& P\left(\text{在 } m_p \text{ 之前至少有 } i \text{ 个样本点, 在 } m_p \text{ 之前不能多于 } j-1 \text{ 个样本点}\right) \\ =& \sum_{h=i}^{j-1} \binom{n}{h} p^h (1-p)^{n-h}. \end{aligned}$$

如果能找到合适的i与j使上式大于等于$1-\alpha$,这样的$(X_{(i)}, X_{(j)})$就构成了m_p 置信度为$100(1-\alpha)\%$ 的置信区间. 当然, 为了得到精度高的置信区间, 理想结果应该是找到使概率最接近$1-\alpha$的i与j.

我们也注意到, 对$P(X_{(i)} < m_p < X_{(j)})$的计算只用到二项分布和$p$, 没有用到有关$f(x)$的具体结构, 所以总可以根据事先给定的$\alpha$, 求出满足上式的合适的$i$和$j$. 这一方法显然适用于一切连续分布, 类似这样的方法称为**不依赖于分布的统计推断方法** (distribution free).

如果我们要求的是中位数的置信区间, 那么上式简化为

$$P(X_{(i)} < m_{\mathrm{e}} < X_{(j)}) = \sum_{h=i}^{j-1} \binom{n}{h} \left(\frac{1}{2}\right)^n.$$

例2.16 表2.26所示的为16名学生在一项体能测试中的成绩, 求由顺序统计量构成的置信度为95%的中位数的置信区间.

表2.26 16名学生在一项体能测试中的成绩

82	53	70	73	103	71	69	80
54	38	87	91	62	75	65	77

解 我们将采用两步法搜索最优的置信区间.

(1) 首先确定使概率大于$1-\alpha$的所有可能区间为备选区间$(X_{(i)}, X_{(j)})(i<j)$;

(2) 从中选出长度最短的区间作为最终的结果.

第一步：所有可能的置信区间共计$\dfrac{16\times 15}{2}=120$个, 置信度95%以上的置信区间有24个, 结果如表2.27所示.

表2.27 体能测试中成绩的置信度在95%以上的置信区间

下限	上限	置信度	下限	上限	置信度
38	80	0.9615784	54	87	0.9958191
38	82	0.9893494	54	91	0.9976501
38	87	0.9978943	54	103	0.9978943
38	91	0.9997253	62	80	0.9509583
38	103	0.9999695	62	82	0.9787292
53	80	0.9613342	62	87	0.9872742
53	82	0.9891052	62	91	0.9891052
53	87	0.9976501	62	103	0.9893494
53	91	0.9994812	65	82	0.9509583
53	103	0.9997253	65	87	0.9595032
54	80	0.9595032	65	91	0.9613342
54	82	0.9872742	65	103	0.9615784

第二步：从这些区间里面找到区间长度最短的区间，为$\left(X_{(5)}, X_{(13)}\right)=(65,82)$, 置信度为0.9509583.

在例2.16中, 得到精度最优的Neyman置信区间$(X_{(5)}, X_{(13)})$，其序号不对称, 这是常见的. 在实际中, 为方便起见, 常常选择指标对称的置信区间.

具体定义为：求满足$100(1-\alpha)\%$的最大的k所构成的置信区间$(X_{(k)}, X_{(n-k+1)})$(这里k可以为0)。如果要求对称的置信区间, 则k应满足

$$1-\alpha \leqslant P(X_{(k)}<M_{\mathrm{e}}<X_{(n-k+1)})=\frac{1}{2^n}\sum_{i=k}^{n-k}\binom{n}{i}.$$

于是, 编写程序计算如下：

```
alpha=0.05
n=length(stu)
conf=pbinom(n,n,0.5)-pbinom(0,n,0.5)
for (k in 1:n)
{conf=pbinom(n-k,n,0.5)-pbinom(k,n,0.5)
 if (conf<1-alpha) {loc=k-1;break}
print(loc)
}
```

在例2.16中, 求出对称的置信区间为$(X_{(4)}, X_{(13)}) = (62, 82)$, 比例题中选出的置信区间略微长了一些.

2. 在对称分布中用Walsh平均法求解置信区间

2.4节给出了对称分布中心的Walsh平均估计方法, 自然可想到应用Walsh平均顺序统计量构造对称中心的置信区间.

定理2.6 原始数据为$X_1, X_2, \cdots, X_n \overset{\text{i.i.d.}}{\sim} F(x-\theta)$, 若$F$对称, 利用Walsh平均法可以得到$\theta$的置信区间. 首先按升幂排列Walsh平均值, 记为$W_{(1)}, W_{(2)}, \cdots, W_{(N)}$, $N = \dfrac{n(n+1)}{2}$. 则对称中心θ的$1-\alpha$置信区间为

$$(W_{(k)}, W_{(n-k+1)}),$$

式中k是满足$P(W_{(j)} < \theta < W_{(n-j+1)}) \geqslant 1-\alpha$ 的最大的j.

例2.17(数据文件chap2\scot.txt) 苏格兰红酒享誉世界, 品种繁多, 本例收集了音乐会上备受青睐的21种威士忌的储存年限(原酒在橡木桶中的储存年限), 如果假设这些年限来自对称分布, 试用Walsh平均法给出这些收藏年限中位数的置信区间. 下面给出本例的参考程序.

```
 #Walsh.AL.scot is the Walsh transform
NL=length(Walsh.AL.scot) alpha=0.05 for (k in seq(1,NL/2,1)) {
 F=pbinom(NL-k,NL,0.5)-pbinom(k,NL,0.5)
 if (F<1-alpha)
  {IK=k-1
   break
   }
} sort.Walsh.AL.scot=sort(Walsh.AL.scot)
Lower=sort.Walsh.AL.scot[IK] Upper=sort.Walsh.AL.scot[NL-IK+1]
c(Lower,Upper)
 13.50 14.75
```

与用顺序统计量求出的置信区间(7.5,19.5)比较发现, 显然Walsh平均法的结果更为精确.

§2.6.2 基于方差估计法的位置参数置信区间估计

置信区间估计中最核心的内容是求解估计量(或统计量)的方差, Bootstrap方法是常用的不依赖于分布的求解统计量$T_n = g(X_1, X_2, \cdots, X_n)$的方差的方法。在本小节中, 我们首先介绍方差估计的Bootstrap方法, 然后介绍用Bootstrap方法构造置信区间的方法.

1. 方差估计的Bootstrap方法

令$V_F(T_n)$ 表示统计量T_n的方差, F表示未知的分布(或参数), $V_F(T_n)$是分布F的函数. 比如$T_n = n^{-1}\sum_{i=1}^{n} X_i$, 那么

$$V_F(T_n) = \frac{\sigma^2}{n} = \frac{\int x^2 \mathrm{d}F(x) - \left(\int x \mathrm{d}F(x)\right)^2}{n}$$

是 F 的函数.

在 Bootstrap 方法中, 我们用经验分布函数替换分布函数F, 用 $V_{\hat{F}_n}(T_n)$估计 $V_F(T_n)$. 由于 $V_{\hat{F}_n}(T_n)$ 通常很难计算得到, 因此在Bootstrap 方法中利用重抽样的方法计算 v_{boot} 来近似 $V_{\hat{F}_n}(T_n)$. Bootstrap方法估计统计量方差的具体步骤如下：

Bootstrap 方差估计

---

(1) 从经验分布$\hat{F}_n$中重抽样 $X_1^*, X_2^*, \cdots, X_n^*$;

(2) 计算 $T_n^* = g(X_1^*, X_2^*, \cdots, X_n^*)$;

(3) 重复步骤(1), (2)共 B 次, 得到 $T_{n,1}^*, T_{n,2}^*, \cdots, T_{n,B}^*$;

(4)计算

$$v_{\text{boot}} = \frac{1}{B}\sum_{b=1}^{B}\left(T_{n,b}^* - \frac{1}{B}\sum_{r=1}^{B} T_{n,r}^*\right)^2.$$

经验分布在每个样本点上的概率密度为 $1/n$, Bootstrap方差估计所述步骤中的第(1)步相当于从原始数据中有放回地简单随机抽取 n 个样本.

由大数定律, 当 $B \to \infty$ 时, $v_{\text{boot}} \xrightarrow{\text{a.s.}} V_{\hat{F}_n}(T_n)$. T_n 的标准差 $\hat{\text{Sd}}_{\text{boot}} = \sqrt{v_{\text{boot}}}$. 下边这个关系表示了 Bootstrap方法的基本思想：

$$v_{\text{boot}} \to V_{\hat{F}_n}(T_n) \sim V_F(T_n).$$

从上面的步骤很容易得到由 Bootstrap方法对中位数的方差进行估计的基本步骤如下：

Bootstrap 中位数方差估计

给定数据 $X=(X_{(1)},X_{(2)},\cdots,X_{(n)})$

for (b in 1 to B)

$X^*_{m,b}=$ 样本量为 m、对 X 进行有放回简单随机抽样得到的样本;

$M_b=X^*_{m,b}$ 的中位数;

end for

$$v_{\text{boot}}=\frac{1}{B}\sum_{b=1}^{B}\left(M_b-\frac{1}{B}\sum_{b=1}^{B}M_b\right)^2$$

$\text{Sd}_{\text{median}}=\sqrt{v_{\text{boot}}}$

例2.18(见chap2\数据nerve.txt) 用 Bootstrap 方法对 nerve 数据估计中位数的方差,以下给出R参考程序:

```
X=nerve
Median.nerve=median(X)
TBoot=NULL
n=20
B=1000
SD.nerve=NULL
for (i in 1:B)
{
  Xsample=sample(X,n,T)
  Tboot=median(Xsample)
  TBoot=c(TBoot,Tboot)
  SD.nerve=c(SD.nerve,sd(TBoot))
}
Sd.median.nerve=sd(TBoot)
 plot(1:B,SD.nerve,col=4)
 hist(TBoot,col=3)
}
```

在以上程序中, 每次Bootstrap样本量m设为20, Bootstrap试验共进行1000次, Tboot向量中保存了每次Bootstrap样本的中位数. $B=1000$ 时, R软件计算得到中位数的抽样标准差为 $\text{Sd}_{\text{median}}=0.0052$. 我们制作了1000次对中位数进行估计的直方图和当Bootstrap试验次数增加时中位数标准差估计的变化情况图, 如

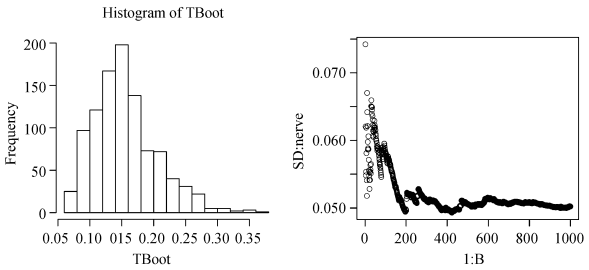

图 2.5 Bootstrap中位数估计分布和标准差变化图

图2.5所示. 从图中可以观察到, 中位数的估计抽样分布为单峰形态, 有略微右偏倾向. 当Bootstrap 试验次数增加到400以后, 中位数估计的标准差趋于稳定.

2. 位置参数的置信区间估计

(1) 正态置信区间

当有证据表明T_n 的分布接近正态分布时, 正态置信区间是最简单的一种构造置信区间的方法：

$1-\alpha$ Bootstrap 正态置信区间

$$\left(T_n - z_{\alpha/2}\hat{\mathrm{Sd}}_{\mathrm{boot}}, T_n + z_{\alpha/2}\hat{\mathrm{Sd}}_{\mathrm{boot}}\right) \tag{2.21}$$

当然, 应用这一方法的前提是T_n 的分布接近正态分布, 否则正态置信区间的精确度很低.

(2) 枢轴量置信区间

当无法确定估计量T_n的分布是否正态, 或有证据可以否定T_n的分布为正态的可能, 那么可以运用枢轴量(pivotal)的方法给出Bootstrap T_n 的置信区间. 首先回顾枢轴量的概念：一个统计量和参数θ的函数$G(T_n,\theta)$,如果$G(T_n,\theta)$的分布与θ无关, 而且是可以求得的, 那么就可以通过求解G分布的分位数, 将求θ上、下置信限的问题转化成方程组求根问题, 从而解决置信区间问题. 因此, 枢轴量是一种比较传统的求解置信区间的方法. 比如在参数推断中, 典型的枢轴量有$\dfrac{\bar{X}-\mu}{\sigma_0/\sqrt{n}} \sim N(0,1)$, $\dfrac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$.

假设θ是待估参数, $\hat{\theta}$是估计量, $\hat{\theta}-\theta$是抽样误差, 这个函数的分位点为$\delta_{\frac{\alpha}{2}}, \delta_{1-\frac{\alpha}{2}}$, 则有

$$P(\hat{\theta}-\theta\leqslant\delta_{\frac{\alpha}{2}})=\frac{\alpha}{2},$$

$$P(\hat{\theta}-\theta\leqslant\delta_{1-\frac{\alpha}{2}})=1-\frac{\alpha}{2}.$$

于是

$$P(\hat{\theta}-\delta_{1-\frac{\alpha}{2}}\leqslant\theta\leqslant\hat{\theta}-\delta_{\frac{\alpha}{2}})=1-\alpha.$$

$\hat{\theta}-\delta_{1-\frac{\alpha}{2}}$和$\hat{\theta}-\delta_{\frac{\alpha}{2}}$就是$\theta$的置信下限和置信上限. 下面只要得到$\hat{\theta}-\theta$的$\frac{\alpha}{2}$和$1-\frac{\alpha}{2}$ 分位数估计即可.

求解思路是用$\hat{\theta}$估计θ,用Bootstrap样本$\theta_j^*(j=1,2,\cdots B)$的分位点估计$\hat{\theta}$的分位点, 即用$\hat{\theta}^*_{\frac{\alpha}{2}}-\hat{\theta}$作为对$(\hat{\theta}-\theta)_{\frac{\alpha}{2}}$的估计, 用$\hat{\theta}^*_{1-\frac{\alpha}{2}}-\hat{\theta}$作为对$(\hat{\theta}-\theta)_{1-\frac{\alpha}{2}}$的估计. 即

$$\hat{\delta}_{1-\frac{\alpha}{2}}=\hat{\theta}^*_{1-\frac{\alpha}{2}}-\hat{\theta},$$
$$\hat{\delta}_{\frac{\alpha}{2}}=\hat{\theta}^*_{\frac{\alpha}{2}}-\hat{\theta}.$$

于是

$$\hat{\theta}-\hat{\delta}_{1-\frac{\alpha}{2}}=2\hat{\theta}-\hat{\theta}^*_{1-\frac{\alpha}{2}},$$
$$\hat{\theta}-\hat{\delta}_{\frac{\alpha}{2}}=2\hat{\theta}-\hat{\theta}^*_{\frac{\alpha}{2}}.$$

$1-\alpha$ Bootstrap 枢轴量置信区间

$$C_n=\left(2\hat{\theta}_n-\hat{\theta}^*_{1-\frac{\alpha}{2}},2\hat{\theta}_n-\hat{\theta}^*_{\frac{\alpha}{2}}\right). \tag{2.22}$$

(3) 分位数置信区间

假设存在 T 的一个单调变换 $U=m(T)$ 使得 $U\sim N(\phi,\sigma^2)$, 其中 $\phi=m(\theta)$, 我们不需要知道变换的具体形式, 仅知道存在这样一个变换. 令 $U_b^*=m(T_b^*)$, 因为 m 是一个单调变换, 所以有 $U^*_{B\alpha/2}=m(T^*_{B\alpha/2})$, $U^*_{B\alpha/2}\approx U-z_{\alpha/2}c$, 且 $U^*_{B(1-\alpha/2)}\approx U+z_{\alpha/2}c$, 那么有

$$\begin{aligned}P\left(T^*_{B\alpha/2}\leqslant\theta\leqslant T^*_{B(1-\alpha/2)}\right)&=P\left(m(T^*_{B\alpha/2})\leqslant m(\theta)\leqslant m(T^*_{B(1-\alpha/2)})\right)\\&=P\left(U^*_{B\alpha/2}\leqslant\phi\leqslant U^*_{B(1-\alpha/2)}\right)\\&\approx P\left(U-cz_{\alpha/2}\leqslant\phi\leqslant U+cz_{\alpha/2}\right)\\&=1-\alpha.\end{aligned}$$

满足条件的变换 m 仅在很少的情况下存在, 更一般的情况是, 我们可以用Bootstrap样本的分位点作为统计量的置信区间.

$1-\alpha$ Bootstrap 分位数置信区间

$$C_n = \left(T^*_{B\alpha/2}, T^*_{B(1-\alpha/2)}\right). \tag{2.23}$$

式中 T^*_β 是统计量 Bootstrap 的 β 分位数.

例2.19 对nerve数据中位数用三种方法构造 95%的置信区间.

解 令 $B = 1000$, $n = 20$,结果如下：

方法	95% 的置信区间
正态	$(0.058225, 0.24177)$
枢轴量	$(0.039875, 0.220000)$
分位数	$(0.084875, 0.265000)$

参考程序如下：

```
Alpha=0.05
Lcl=Median.nerve+qnorm(0.025,0,1)*Sd.median.nerve
Ucl=Median.nerve-qnorm(0.025,0,1)*Sd.median.nerve
NORM.interval=c(Lcl,Ucl)

Lcl=2*Median.nerve-quantile(TBoot,0.975)
Ucl=2*Median.nerve-quantile(TBoot,0.025)
PIVOTAL.interval=c(Lcl,Ucl)

Lcl=2*Median.nerve-quantile(TBoot,0.025)
Ucl=2*Median.nerve+quantile(TBoot,0.975)
QUATILE.interval=c(Lcl,Ucl)
```

§2.7 正态记分检验

由前面的Wilcoxon秩和检验可知, 如果$X_1, X_2, \cdots, X_n$为独立同分布的连续型随机变量,那么秩统计量$R_1, R_2, \cdots, R_n$在$1, 2, \cdots, n$上有均匀分布.

秩定义了数据在序列中数量大小的位置和序,它们与未知分布$F(x)$的n 个p 分位点一一对应. 我们知道, 分布函数是单调增函数, 秩大意味着对应分布中较大的分位点, 秩小则对应着分布中较小的分位点. 不同的分布所对应的点虽然不同, 但是序相同, 也就是说,由秩对应到不同分布的分位点之间的单调关系不变. 可见，这里分布不是本质的, 完全可以选用熟悉的分布, 比如, 用正态分布作为参照, 将秩转化为相应的正态分布的分位点, 这样, 就可以将依赖于秩的检验, 化为对分位点大

小的检验. 同时它提供了将顺序数据转化为连续数据的一种思路. 这种以正态分布作为转换记分函数, 将Wilcoxon秩检验进行改进的方法称为 **正态记分检验**.

正态记分可以用在许多检验问题中, 有多种不同的形式. 具体来说, 正态记分检验的基本思想就是把升幂排列的秩R_i用升幂排列的正态分位点代替. 比如最直接的想法是用$\Phi^{-1}(R_i/(n+1))$来代替每一个样本的值. 为了保证变换后的和为正, 一般不直接采用$\Phi^{-1}(R_i/(n+1))$ 作为记分, 而是稍微改变一下：

$$s(i) = \Phi^{-1}\left(\frac{n+1+R_i}{2n+2}\right),\ i = 1, 2, \cdots, n.$$

式中$s(i)$表示第R_i个数据的正态记分.

具体实现步骤如下.

对于假设检验问题$H_0 : M = M_0 \leftrightarrow H_1 : M \neq M_0$：

(1) 把$|X_i - M_0|(i = 1, 2, \cdots, n)$ 的秩按升幂排列,并加上相应的$X_i - M_0$的符号(成为符号秩).

(2) 用相应的正态记分代替这些秩, 如果r_i为$|X_i - M_0|$ 的秩, 则相应的符号正态记分为

$$s_i = \Phi^{-1}\left(\frac{1}{2}\left[1 + \frac{r_i}{n+1}\right]\right)\mathrm{sign}(X_i - M_0).$$

其中

$$\mathrm{sign}(X_i - M_0) = \begin{cases} 1, & X_i > M_0, \\ -1, & X_i < M_0. \end{cases}$$

用W表示所有符号记分s_i之和, 即$W = \sum_{i=1}^{n} s_i$, 正态记分检验统计量为

$$T^+ = \frac{W}{\sqrt{\sum_{i=1}^{n} s_i^2}}.$$

(3) 如果观测值的总体分布接近于正态, 或者在大样本情况, 可以认为T^+近似地有标准正态分布. 这对于很小的样本(无论是否打结)也适用. 这样可以很方便地计算p值. 实际上, 如果记$\Phi_+(x) \equiv 2\Phi(x) - 1 = P(|X| \leqslant x)$, 则有

$$\Phi_+^{-1}\left(\frac{i}{n+1}\right) = \Phi^{-1}\left[\frac{1}{2}\left(1 + \frac{i}{n+1}\right)\right],$$

大约等于$E|X|_{(i)}$. 也就是说, 它和期望正态记分相近.

(4) 当T^+大的时候, 可以考虑拒绝原假设.

例2.20 这是吴喜之(1999)书中的一个例子. 以下是亚洲10个国家1996年每1000个新生儿中的(按从小到大次序排列)死亡数(按照世界银行的“世界发展指标”, 1998)：

日本	以色列	韩国	斯里兰卡	叙利亚	中国	伊朗	印度	孟加拉国	巴基斯坦
4	6	9	15	31	33	36	65	77	88

对于新生儿死亡率的例子, 我们将考虑两个假设检验：$H_0 : M \geqslant 34 \leftrightarrow H_1 : M < 34$和$H_0 : M \leqslant 16 \leftrightarrow H_1 : M > 16$.

计算结果列在表2.28中.

为了计算T^+方便, 标出了带有$X_i - M_0$ 符号的s_i^+ 即所谓的“符号s_i^+” 它等于$\text{sign}(X_i - M_0)s_i^+$.

表2.28 亚洲10国新生儿死亡率(单位：千分之一)一例的正态记分检验
数据按$|X_i - M_0|$升幂排列(左边$M_0 = 34$, 右边$M_0 = 16$)

$H_0 : M \geqslant 34 \leftrightarrow H_1 : M < 34$			
X_i	$\lvert X_i - M_0\rvert$	符号秩	符号s_i^+
33	1	-1	-0.114
36	2	2	0.230
31	3	-3	-0.349
15	19	-4	-0.473
9	25	-5	-0.605
6	28	-6	-0.748
4	30	-7	-0.908
65	31	8	1.097
77	43	9	1.335
88	54	10	1.691
$W = 1.156, T^+ = 0.409$			
p值$= \Phi(T^+) = 0.659$			
结论：不能拒绝H_0 (水平$\alpha < 0.659$)			

$H_0 : M \leqslant 16 \leftrightarrow H_1 : M > 16$			
X_i	$\lvert X_i - M_0\rvert$	符号秩	符号s_i^+
15	1	-1	-0.114
9	7	-2	-0.230
6	10	-3	-0.349
4	12	-4	-0.473
31	15	5	0.605
33	17	6	0.748
36	20	7	0.908
65	49	8	1.097
77	61	9	1.335
88	72	10	1.691
$W = 5.217, \quad T^+ = 1.844$			
p值$= 1 - \Phi(T^+) = 0.033$			
结论：$M > 16$ (水平$\alpha < 0.033$)			

实际上, 这里也可以使用统计量$W \equiv |W^+ - W^-|$做检验. W也存在临界值表. 在原假设下的大样本正态近似统计量为

$$Z = \frac{W}{\sqrt{\sum_i R_i^2}}.$$

它的分母在没有结的情况下为$\sqrt{n(n+1)(2n+1)/6}$. 对于$H_0 : M \geqslant 34 \leftrightarrow H_1 : M < 34$, $W = \sum s_i = 1.156,\ T^+ = 0.409$, p值$= \Phi(T^+) = 0.659$; 而对于$H_0 : M \leqslant 16 \leftrightarrow H_1 : M > 16$, $W = \sum s_i = 5.217,\ T^+ = 1.844$, p值$= 1 - \Phi(T^+) = 0.033$. 这和前面

的T^+正态记分检验结果完全一样, 这种相似之处正是源于它们所代表的信息是等价的.

如定义所示, 这里的正态记分检验对应于Wilcoxon符号秩检验(统计量为W^+), 正态记分检验有较好的大样本性质. 对于正态总体, 它比许多基于秩的检验更好. 而对于一些非正态总体, 虽然结果可能不如一些基于秩的检验, 但它又比t检验要好. 表2.29列出了上述正态记分(NS^+)相对于Wilcoxon符号秩检验(W^+)在不同总体分布下的ARE值.

表2.29 正态记分相对于wilcoxon符号秩检验在不同总体下的A.R.E. 值

总体分布	均匀	正态	Logistic	重指数	Cauchy
ARE(NS^+, W^+)	$+\infty$	1.047	0.955	0.847	0.708

实际上, 在使用以秩定义的检验统计量的地方都可以把秩替换成正态记分而形成相应的正态记分统计量, 从而将顺序的数据化为定量数据进行分析.

对该例第二个检验可以使用R程序函数ns, 如下所示:

```
ns(baby,16)
$two.sided.pvalue
[1] 0.06515072
$T
[1] 1.844223
$s
[1]  0.7478586  0.9084579  0.6045853 -0.1141853 -0.2298841
-0.3487557 -0.4727891  1.0968036   1.3351777  1.6906216
```

§2.8 分布的一致性检验

在数据分析中, 经常要判断一组数据是否来自某一特定的分布, 比如对连续型分布, 常判断数据是否来自正态分布; 而对离散型分布, 常需要判断数据是否来自某一事先假定的分布, 常见的分布有二项分布、Poisson分布, 或判断实际观测与期望数是否一致. 本节我们将关注这些问题. 我们从一般到特殊, 首先考察判断实际观测与期望数是否一致, 重点介绍Pearson χ^2拟合优度检验法; 当总体均值和方差未知时, 我们将介绍两种检验数据是否偏离正态分布的常用方法: Kolmogrov-Smirnov检验法和Lilliefor 检验法.

§2.8.1 χ^2拟合优度检验

1. 实际观测数量与期望次数一致性检验

当一组数据的类型为类别数据(categorical data)时, 其中n个观测值可分为c种类别, 每一类别可计算其发生频数, 称为实际观测频数(observed frequency), 记为$O_i(i=1,2,\cdots,c)$, 表示如表2.30所示：

表2.30 实际观测频数表

类别	1	2	$\cdots$	c	总和
实测次数	O_1	O_2	$\cdots$	O_c	n

我们想了解每一类别发生的概率是否与理论分布$\{p_i, i=1,2,\cdots,c\}$ 一致. 即有如下假设检验问题：

$$H_0: \text{总体分布为 } \forall p_i, i=1,2,\cdots,c \ \ (\text{即} F(x)=F_0(x)),$$
$$H_1: \text{总体分布不为} p_i, \exists i=1,2,\cdots,c \ \ (\text{即} F(x)\neq F_0(x)).$$

若原假设成立, 则期望频数(expected frequency)应为$E_i=np_i(i=1,2,\cdots,c)$,因此可以由实际频数(O_i) 与期望频数(E_i)是否接近作为检验总体分布与理论分布是否一致的测量标准, 通常采用如下定义的Pearsonχ^2统计量：

$$\chi^2=\sum_{i=1}^{c}\frac{(O_i-E_i)^2}{E_i}=\sum_{i=1}^{c}\frac{O_i^2}{E_i}-n. \tag{2.24}$$

结论：当实际观测χ^2值大于自由度$\upsilon=c-1$的χ^2值, 即$\chi^2>\chi^2_{\alpha,c-1}$时, 则拒绝$H_0$, 表示数据分布与理论分布不符.

例2.21 调查发现某美发店上半年各月顾客数量如表2.31所示：

表2.31 上半年各月顾客数量表

月份	1	2	3	4	5	6	合计
顾客数量/百人	27	18	15	24	36	30	150

该店经理想了解各月顾客数是否服从均匀分布.

解 假设检验问题：

$$H_0: \text{各月顾客数符合均匀分布 } 1:1\left(\text{即各月顾客比例 } p_i=p_0=\frac{1}{6}, \forall i=1,2,\cdots,6\right),$$
$$H_1: \text{各月顾客数不符合 } 1:1\left(\text{即各月顾客比例 } p_i\neq p_0=\frac{1}{6}, \exists i=1,2,\cdots,6\right).$$

统计分析如表2.32所示：

表2.32 实际观测频数与期望频数汇总表

月份	1	2	3	4	5	6	合计
实际频数O_i	27	18	15	24	36	30	150
期望频数E_i	25	25	25	25	25	25	150

表2.32中，$E_i = np_i = 150 \times \dfrac{1}{6} = 25, i = 1, 2, \cdots, 6.$

由式(2.24)得

$$\begin{aligned}\chi^2 = &\frac{(27-25)^2}{25} + \frac{(18-25)^2}{25} + \frac{(15-25)^2}{25} \\ &+ \frac{(24-25)^2}{25} + \frac{(36-25)^2}{25} + \frac{(30-25)^2}{25} \\ = &12.\end{aligned}$$

结论：实测$\chi^2 = 12 > \chi^2_{0.05,6-1} = 11.07$, 接受$H_1$假设, 认为到该店消费的顾客在各月比例不相等, 即$p \neq \dfrac{1}{6}$.

2. 泊松分布的一致性检验

例2.22 调查某农作物根部蚜虫的分布情况. 调查结果如表2.33所示, 问：蚜虫在某农作物根部的分布是否为泊松分布?

表2.33 蚜虫实际株数表

每株虫数x	0	1	2	3	4	5	6以上	n合计
实际株数O_i	10	24	10	4	1	0	1	50

解 假设检验问题：

$$\begin{aligned}&H_0: \text{蚜虫在农作物根部的分布是泊松分布,} \\ &H_1: \text{蚜虫在农作物根部的分布不是泊松分布.}\end{aligned}$$

若蚜虫在农作物根部的分布为泊松分布, 则分布列为

$$P(X = x) = \frac{\mathrm{e}^{-\lambda}\lambda^x}{x!}, \quad x = 0, 1, 2, \cdots,$$

其中, λ是泊松分布的期望, 是未知的, 需要用观测值估计, 其估计值如下：

$$\hat{\lambda} = \bar{x} = (0 \times 10 + 1 \times 24 + \cdots + 6 \times 1)/50 = 1.3.$$

因而

$$\hat{p_0}=\frac{\mathrm{e}^{-1.3}\times 1.3^0}{0!}=0.2725,$$

$$\hat{p_1}=\frac{\mathrm{e}^{-1.3}\times 1.3^1}{1!}=0.3543,$$

$$\hat{p_2}=\frac{\mathrm{e}^{-1.3}\times 1.3^2}{2!}=0.2303,$$

$$\hat{p_3}=\frac{\mathrm{e}^{-1.3}\times 1.3^3}{3!}=0.0998,$$

$$\hat{p_4}=\frac{\mathrm{e}^{-1.3}\times 1.3^4}{4!}=0.0324,$$

$$\hat{p_5}=1-\hat{p}_0-\hat{p}_1-\hat{p}_2-\hat{p}_3-\hat{p}_4=0.0107.$$

根据泊松分布计算各x_i类别下的期望数$E_i=np_i(i=0,1,2,3)$,由于3、4、5、6的实际株数数量较少，此处作了合并。得表2.34：

表2.34 农作物根部蚜虫数实际株数和期望株数计算表

虫数	实际株数O_i	泊松概率p_i	期望株数E_i	$\frac{(O_i-E_i)^2}{E_i}$
0	10	0.2725	13.625	0.9644
1	24	0.3543	17.715	2.2298
2	10	0.2303	11.515	0.1993
3	6	0.1429	7.145	0.1835
总和	50			3.577

由式(2.24)得

$$\chi^2=\frac{10^2}{13.625}+\cdots+\frac{6^2}{7.145}-50=3.577<\chi^2_{0.05,2}=5.991.$$

结论：由表2.34可知, $\chi^2=3.577<\chi^2_{0.05,2}=5.991$, 不能拒绝$H_0$, 不能排除蚜虫在某农作物根部的分布不是泊松分布.

3. 正态分布一致性检验

χ^2拟合优度检验也可用于检验一组数据是否服从正态分布.

例2.23 从某地区高中二年级学生中随机抽取45位学生量得体重如表2.34所示, 问该地区学生体重(单位：kg)的分布是否为正态分布？

表2.35 45位学生体重抽样数据表（单位：kg）

36	36	37	38	40	42	43	43	44	45	48	48	50	50	51
52	53	54	54	56	57	57	57	58	58	58	58	58	59	60
61	61	61	62	62	63	63	65	66	68	68	70	73	73	75

解 假设检验问题：

H_0 : 某地区高中二年级学生体重分布为正态,

H_1 : 某地区高中二年级学生体重分布不为正态.

统计分析：将上述体重数据分为5组(class), 每组实际观测次数如表2.36所示。

表2.36 以10为间隔分组体重频数分布表（单位：kg）

体重	30∼40	40∼50	50∼60	60∼70	70∼80
次数	5	9	16	12	3

由表2.36可知, 分组数据的平均值为$\bar{X} = 54.78$;样本方差为$S^2 = 120.4040$; 样本标准差为$S = 10.9729$.其中分组均值和分组样本方差如下式计算：

$$\bar{X} = \frac{\sum_{i=1}^{K} f_i X_i}{\sum_{i=1}^{K} f_i};$$

$$S^2 = \frac{\sum_{i=1}^{K} f_i (X_i - \bar{X})^2}{\sum_{i=1}^{K} f_i}$$

其中K是组数，f_i是第i组的频数

根据正态分布计算累计概率和期望频数如表2.37所示.

表2.37 学生体重分组频数与期望频数计算表

分组	上组限b_i	实际观测频数	标准正态值 $Z_i = (b_i - \hat{\mu})/S$	累计概率 $F_0(x)$	组间概率 p_i	期望频数 $E_i = np_i$	$(O_i - E_i)^2/E_i$
30∼40	40	5	−1.35	0.0885	0.0766	3.45	0.6964
40∼50	50	9	−0.44	0.3300	0.2415	10.87	0.3217
50∼60	60	16	0.48	0.7190	0.3890	17.51	0.1302
60∼70	70	12	1.39	0.9177	0.1987	8.94	1.0474
70∼80	80	3	2.30	0.9893	0.0716	3.22	0.0150
80以上		0		1.0000	0.0107	0.48	
							2.2107

结论：由表2.37可知, 实际观测$\chi^2 = 2.2107 < \chi^2_{0.05,2} = 5.991$, 不拒绝$H_0$, 没有理由怀疑该地区高中二年级学生的体重不服从正态分布.

§2.8.2 Kolmogorov-Smirnov正态性检验

Kolmogorov-Smirnov检验法(简称K-S检验)用来检验单一简单随机样本样本$X_1, X_2, \cdots, X_n$是否来自某一指定分布$F(.)$，比如检验一组数据是否来自正态分布.这一检验方法是基于样本数据的累计分布函数与制订理论分布比较, 若两者间的差距很小, 则推论该样本取自某特定分布族.假设检验问题如下：

H_0 : 样本所来自的总体服从某特定分布,

H_1 : 样本所来自的总体不服从某特定分布.

$K-S$检验统计量如下定义：

$$D_n(x) = \max_{1\leqslant r\leqslant n}\left\{\max\left(\left|F(x_{(r)}) - \frac{r-1}{n}\right|, \left|F(x_{(r)}) - \frac{r}{n}\right|\right)\right\}$$

这里$x_{(1)} \leqslant x_{(2)} \leqslant \cdots \leqslant x_{(n)}$是样本的排序结果，Kolmogrov1933 年中给出了证明：

$$\lim_{n\to\infty} P(D_n(x) \leqslant x) = \sum_{j=-\infty}^{+\infty} (-1)^j \exp(-2nj^2x).$$

设D_n表示了$F_0(x)$与$F_n(x)$在样本上差距的最大值, 当X是连续分布时，$D_n(x)$ 可以简化。令$F_0(x)$表示待检验的分布的理论分布函数,$F_n(x)$表示样本的经验分布函数,D_n如下式所示：

$$D_n = \max_{1\leqslant r\leqslant n} \left|F_n(x_{(r)}) - F_0(x_{(r)})\right|. \tag{2.25}$$

结论：当实际观测$D_n > D_\alpha$(见附表14), 则拒绝H_0假设, 反之，则不拒绝H_0假设。

例2.24 35位健康成年男性在未进食前的血糖浓度如表2.38所示, 试检验这组数据是否来自均值为$\mu = 80$，标准差为$\sigma = 6$ 的正态分布。

表2.38 35位健康成年男性在未进食前的血糖浓度数据

87	77	92	68	80	78	84	77	81	80	80	77	92	86
76	80	81	75	77	72	81	90	84	86	80	68	77	87
76	77	78	92	75	80	78		n=35					

解 假设检验问题：

H_0 : 健康成年男性血糖浓度服从正态分布,

H_1 : 健康成年男性血糖浓度不服从正态分布.

根据正态分布计算理论分布值如表2.39所示.

表2.39 健康男性血糖浓度观测频数与理论分布对照表

血糖浓度(x)	次数(f)	累计次数(F)	$F_n(x)=F/n$	标准化值$Z=(x-\mu)/\sigma$	理论分布$F_0(x)$	D
68	2	2	0.0571	−2.00	0.0228	0.0291
72	2	4	0.1143	−1.33	0.0934	0.0209
75	2	6	0.1714	−0.83	0.2033	0.0319
76	2	8	0.2286	−0.67	0.2514	0.0228
77	6	14	0.4000	−0.50	0.3085	0.0915
78	3	17	0.4857	−0.33	0.3707	0.1150
80	6	23	0.6571	0	0.5000	0.1571
81	3	26	0.7429	0.17	0.5675	0.1754*
84	2	28	0.8000	0.67	0.7486	0.0514
86	2	30	0.8571	1.00	0.8413	0.0158
87	2	32	0.9143	1.17	0.8790	0.0353
92	3	35	1.0000	2.00	0.9772	0.0228

*该值是这一列最大值.

结论：表2.39中的$F_0(x)$是根据$Z=(x-80)/6$的标准化值查附表1而得的. 实际观测$D=\max|F_n(x)-F_0(x)|=0.1754<D_{0.05,35}=0.23$(见附表14), 故不能拒绝$H_0$, 不能说明健康成年男性血糖浓度不服从正态分布. 当样本量n较大时, 可以用$D_{\alpha,n}=1.36/\sqrt{n}$ 求得结果, 如上述$D_{0.05,35}=1.36/\sqrt{35}=0.2299=0.23$.

该例题也可以调用R中的函数ks.test解：

```
data(healthy)
ks.test(healthy,pnorm,80,5.9)
        One-sample Kolmogorov-Smirnov test
data:  healthy
D = 0.17556, p-value = 0.2309
alternative hypothesis: two-sided
```

χ^2拟合优度检验与Kolmogorov-Smirnov正态性检验都采用实际频数和期望频数之差进行检验. 它们之间最大的不同在于前者主要用于类别数据, 而后者主要用于有计量单位的连续和定量数据, χ^2拟合优度检验虽然也可以用于定量数据, 但必须先将数据分组才能获得实际的观测频数, 而Kolmogorov-Smirnov正态性检验可以直接对原始数据的n个观测值进行检验, 所以它对数据的利用较完整.

§2.8.3 Liliefor正态分布检验

当总体均值和方差未知时, Liliefor(1967)提出用样本的均值($\bar{X}$) 和标准差(S)代替总体的期望μ和标准差σ, 然后使用Kolmogorov-Smirnov正态性检验法。首先对原始数据X_i 标准化：

$$Z_i=\frac{X_i-\bar{X}}{S},i=1,2,\cdots,n$$

定义L统计量：

$$L = \max |F_n(z) - \hat{F}_0(x)|. \tag{2.26}$$

例2.25(例2.24续) 由例2.24所示的35位健康成年男性血糖浓度数据可知, 样本均值

$$\begin{aligned}\bar{X} &= (87 + 77 + \cdots + 78)/35 = 2791/35\\ &= 79.74.\end{aligned}$$

样本方差

$$\begin{aligned}S^2 &= \frac{1}{35-1}[87^2 + 77^2 + \cdots + 78^2 - 2791^2/35]\\ &= (223761 - 222562.31)/34\\ &= 1198.69/34\\ &= 35.2556,\\ S &= 5.94.\end{aligned}$$

根据正态分布计算理论估计值如表2.40所示.

表2.40 健康男性血糖浓度观测频数与期望分布对照表

血糖浓度(x)	次数(f)	累计次数(F)	$F_n(x_{(i)})$	$Z = (x - \bar{x})/S$	$\hat{F}_0(x_{(i)})$	D
68	2	2	0.0571	−1.98	0.0239	0.0332
72	2	4	0.1143	−1.30	0.0961	0.0182
75	2	6	0.1714	−0.80	0.2119	0.0405
76	2	8	0.2286	−0.63	0.2643	0.0375
77	6	14	0.4000	−0.46	0.3228	0.0772
78	3	17	0.4857	−0.29	0.3859	0.0998
80	6	23	0.6571	0.04	0.5160	0.1411
81	4	26	0.7429	0.21	0.5832	0.1597*
84	2	28	0.8000	0.72	0.7642	0.0358
86	2	30	0.8571	1.05	0.8531	0.0040
87	2	32	0.9143	1.22	0.8888	0.0255
92	3	35	1.0000	2.06	0.9803	0.0197

由表2.40可知, 实际$L = 0.1597 < L_{0.05,35} = 0.23$, 推断不能否认这些健康成年男性血糖浓度服从正态分布.

§2.9 单一总体渐近相对效率比较

假设$X_1, X_2, \cdots, X_n \overset{\text{i.i.d.}}{\sim} F(x-\theta)$,$F(x) \in \Omega_S$, 根据第1章的介绍, 只要Pitman条件满足, 我们可通过求$\mu_n'(0)$和$\sigma_n(0)$来找到一个统计量的效率C, 从而可用不同统计量的效率得到渐近相对效率(ARE). 下面根据本章定义的几个非参数统计量, 结合

参数统计中常用的统计量进行一些比较. 这里我们用$f(x)$表示$F(x)$的概率密度函数.

(1) 记符号统计量$S=\#\{X_i>0, 1\leqslant i\leqslant n\}$, 有

$$E(S)=n[1-F(-\theta)],$$

$$\mathrm{var}(S)=n[1-F(-\theta)]F(-\theta).$$

可取$\mu_n(\theta)=E(S)$及$\sigma_n^2(\theta)=\mathrm{var}(S)$, 于是有

$$\mu_n'(0)=nf(0),\quad \sigma_n^2(0)=\frac{n}{4},\quad C_S=2f(0).$$

这里C_S表示符号统计量的效率.

(2) 对Wilcoxon符号秩统计量$W^+=\sum\limits_{j=1}^{n}R_j^+S(X_j)$, 有

$$E(W^+)=np_1+n\frac{n(n-1)}{2}p_2,$$

$$\mathrm{var}(W^+)=\frac{n(n+1)(2n+1)}{24}.$$

可取$\sigma_n^2(0)=\mathrm{var}(W^+)$及

$$\mu_n(\theta)=E(W^+)=n(1-F(-\theta))+\frac{n(n-1)}{2}\int[1-F(x-\theta)]f(x-\theta)\,\mathrm{d}x,$$

有

$$\mu_n'(0)=nf(0)+n(n-1)\int f^2(x)\,\mathrm{d}x,$$

$$C_{W^+}=\sqrt{12}\int f^2(x)\,\mathrm{d}x.$$

这里C_{W^+}表示Wilcoxon符号秩统计量的效率.

(3) 对传统的t统计量, 记$\sigma_f=\int x^2f(x)\,\mathrm{d}x$. 取

$$\mu_n(\theta)=\sqrt{n}\frac{\theta}{\sigma_f},\quad \sigma_n(0)=1,$$

有$C_t=\dfrac{1}{\sigma_f}$. 这里C_t表示t统计量的效率.

由ARE的定义, $e_{12}=\dfrac{C_1^2}{C_2^2}$, 则上述三个统计量之间的ARE如下：

$$\begin{aligned}\mathrm{ARE}(S,W^+)&=\frac{C_S^2}{C_{W^+}^2}=\frac{f^2(0)}{3\left[\int f^2(x)\,\mathrm{d}x\right]^2},\\ \mathrm{ARE}(S,t)&=\frac{C_S^2}{C_t^2}=4\sigma_f^2 f^2(0),\\ \mathrm{ARE}(W^+,t)&=\frac{C_{W^+}^2}{C_t^2}=12\sigma_f^2\left[\int f^2(x)\,\mathrm{d}x\right]^2.\end{aligned}$$

因此, 对任意给定的分布, 都可计算上面的ARE, 见表2.41：

表2.41 不同分布下常用的检验ARE效率比较

分布	$U(-1,1)$	$N(0,1)$	logistic	重指数
密度	$\frac{1}{2}I(-1,1)$	$\dfrac{\exp\left(-\frac{x^2}{2}\right)}{\sqrt{2\pi}}$	$\mathrm{e}^{-x}(1+\mathrm{e}^{-x})^{-2}$	$\dfrac{\mathrm{e}^{-\lvert x\rvert}}{2}$
$\mathrm{ARE}(W^+,t;F)$	1	$\dfrac{3}{\pi}$	$\dfrac{\pi^2}{9}$	$\dfrac{3}{2}$
$\mathrm{ARE}(S,t;F)$	$\dfrac{1}{3}$	$\dfrac{2}{\pi}$	$\dfrac{\pi^2}{12}$	2

下面例子讨论了当正态分布有不同程度"污染"时, $\mathrm{ARE}(W^+,t)$的不同结果.

例2.26 假定随机样本$X_1,X_2,\cdots,X_n$来自分布$F_\varepsilon=(1-\varepsilon)\Phi(x)+\varepsilon\Phi\left(\dfrac{x}{3}\right)$. 这里$\Phi(x)$为$N(0,1)$的分布函数, 易见

$$\int f_\varepsilon^2(x)\,\mathrm{d}x=\frac{(1-\varepsilon)^2}{2\sqrt{\pi}}+\frac{\varepsilon^2}{6\sqrt{\pi}}+\frac{\varepsilon(1-\varepsilon)}{\sqrt{5\pi}},\quad \sigma_{f_\varepsilon}^2=1+8\varepsilon.$$

由上面公式得

$$\mathrm{ARE}(W^+,t)=\frac{3(1+8\varepsilon)}{\pi}\left[(1-\varepsilon)^2+\frac{\varepsilon^2}{3}+\frac{2\varepsilon(1-\varepsilon)}{\sqrt{5}}\right]^2.$$

对不同的ε, 有如表2.42所示

表2.42 不同混合结构ε 下W^+ 与t 的ARE 比较

ε	0	0.01	0.03	0.05	0.08	0.10	0.15
$\mathrm{ARE}(W^+,t)$	0.955	1.009	1.108	1.196	1.43197	1.373	1.497

从表2.41和表2.42可以看出, 只用到样本中大小次序方面信息的Wilcoxon符号秩检验和符号检验, 当总体分布F为$N(0,1)$时, 相对于t检验的效率并不算差. 当总

体分布偏离正态时, 比如在logistic分布和重指数分布下, 符号检验和W_n^+基本上都优于t检验. 可以证明, 对任何总体分布, Wilcoxon符号秩检验对t检验的渐近相对效率绝不低于0.864. 这说明, 非参数检验在使用样本的效率上不是比参数检验差很多, 有的时候甚至会更好。

之前提到, 一个检验统计量及与其关联的估计量有同样的效率. 上面的符号统计量、Wilcoxon符号秩统计量和t统计量分别相应于样本中位数、Walsh平均的中位数及样本均值, 这些都是Hodges-Lehmann估计量的特例. 一般地, 有下面的估计效率C的定理。

定理2.7 假设$\hat{\theta}$为相应于满足Pitman条件的统计量V的Hodges-Lehmann 估计量. 如果V的效率为C, 则

$$\lim_{n\to\infty} P(\sqrt{n}(\hat{\theta}-\theta)<a)=\Phi(aC).$$

即渐近地有$\sqrt{n}(\hat{\theta}-\theta)\sim N(0,C^{-2})$.

表2.43所示的为t检验(t)、符号检验(S)、Wilcoxon符号秩检验(W^+)之间的ARE范围, 其中带星号(*)的是分布为非单峰时的结果.

表2.43 t, s 和W^+的ARE 范围

	t	S	W^+
t		$(0,3);(0,\infty)*$	$\left(0,\frac{125}{108}\right)$
S	$\left(\frac{1}{3},\infty\right);(0,\infty)^*$		$\left(\frac{1}{3},\infty\right);(0,\infty)^*$
W^+	$\left(\frac{108}{125},\infty\right)$	$(0,3);(0,\infty)*$	

由表2.43可看出 $0.864=\dfrac{108}{125}<\mathrm{ARE}(W^+,t)<\infty$, 无穷大在分布为Cauchy分布时出现, 很明显, 在分布未知时, 非参数方法有很强的优越性. 在用Pitman渐近相对效率时, 要注意这个概念只对大样本适用, 并且它只局限在H_0点的一个邻域中比较.

习题

2.1 超市经理想了解每位顾客在该超市购买的商品平均件数是否为10件, 随机观察12位顾客, 得到如下数据：

顾客	1	2	3	4	5	6	7	8	9	10	11	12
件数	22	9	4	5	1	16	15	26	47	8	31	7

(1) 采用符号检验进行决策.

(2) 采用Wilcoxon符号秩检验进行决策, 比较它与符号检验的结果.

2.2 (1) 请对例题2.5的失业率对时间（月）建立线性回归模型，观察一次项的估计结果，解释该结果和例题中S_1结果的不同之处。

(2) 请根据S_3的性质$\mu(S_3|H_0) = N/6; \sigma^2(S_3|H_0) = N/12$ 将上述数据分为三段，只保留前后两段，用S_3进行检验，给出检验结果.（提示：可尝试R中library(trend)中的函数cs.test 辅助分析）。

2.3 设下表所示为拥有10万人口的某城市15年来每年因车祸的死亡率。请分别使用S_1和S_2分析死亡率是否有逐年增加的趋势?

17.3	17.9	18.4	18.1	18.3	19.6	18.6	19.2	17.7
20.0	19.0	18.8	19.3	20.2	19.9			

2.4 下表中的数据是两场篮球联赛中三分球的进球次数, 考察两场联赛三分球得分次数是否存在显著性差异。

(1) 采用符号检验;

(2) 采用配对Wilcoxon符号秩检验;

(3) 在这些数据中哪个检验更好?为什么?

三分球进球次数		
队伍序号	联赛1	联赛2
1	91	81
2	46	51
3	108	63
4	99	51
5	110	46
6	105	45
7	191	66
8	57	64
9	34	90
10	81	28

2.5 一个监听装置收到如下信号：

0 1 0 1 1 1 0 0 1 1 0 0 0 0 1 1 1 1 1 1 1 1 1 0 1 0 0 1 1 1 0 1 0 1 0 1 0 0
0 0 0 0 0 0 1 0 1 1 0 0 1 1 1 0 1 0 1 0 0 0 1 0 0 1 0 1 0 1 0 0 0 0 0 0 0 0

能否说该信号是纯粹随机干扰?

2.6 某品牌消毒液质检部要求每瓶消毒液的平均容积为500ml,现从流水线上的某台装瓶机器上随机抽取20 瓶, 测得其容量(单位：ml)为：

509	505	502	501	493	498	497	502	504	506
505	508	498	495	496	507	506	507	508	505

试检查这台机器装多装少是否随机.

2.7 六位妇女参加减肥试验, 试验前后体重(单位：lb)如下表所示, 选择适当方法判断她们的减肥计划是否成功。

妇女	1	2	3	4	5	6
试验前	174	192	188	182	201	188
试验后	165	186	183	178	203	181

2.8 (见chap2\AQI)已知一组北京市某年某月某天34个观测站的空气质量指数实地观测数据，表中列出了有关空气质量指数和级别的对应关系。问：如果要判断这一日北京市整体的空气质量，应该设计怎样的假设检验?

2.9 试给出p分位数的Bootstrap置信区间求解程序,并在nerve数据汇总求解0.75和0.25 分位数的置信区间.

2.10 以下给出的是申请进入法学院学习的学生LSAT测试成绩和GPA 成绩.

LSAT	576	635	558	578	666	580	555	661
	651	605	653	575	545	572	594	
GPA	3.39	3.30	2.81	3.03	3.44	3.07	3.00	3.43
	3.36	3.13	3.12	2.74	2.76	2.88	3.96	

每个数据点用$X_i = (Y_i, Z_i)$表示, 其中$Y_i = \mathrm{LSAT}_i$, $Z_i = \mathrm{GPA}_i$.

(1) 计算Y_i和Z_i的相关系数.

(2) 使用Bootstrap方法估计相关系数的标准误差.

(3)计算置信度为0.95的相关系数Bootstrap 枢轴量置信区间.

2.11 构造一个模拟比较4个Bootstrap置信区间的方法. $n = 50$, $T(F) = \int (x-\mu)^3 \mathrm{d}F(x)/\sigma^3$是偏度. 从分布$N(0,1)$中抽出样本$Y_1, Y_2, \cdots, Y_n$, 令$X_i = \mathrm{e}^{Y_i} (i = 1, 2, \cdots, n)$. 根据样本$X_1, X_2, \cdots, X_n$构造$T(F)$的4种类型的置信度为0.95的Bootstrap置信区间。多次重复上述步骤, 估计4种区间的真实覆盖率.

2.12 令$X_1, X_2, \cdots, X_n \sim N(\mu, 1)$. 估计$\hat{\theta} = \mathrm{e}^{\overline{X}}$是参数$\theta = \mathrm{e}^{\mu}$的MLE(极大似然估计). 用$\mu = 5$生成100个观测的数据集.

(1) 用枢轴量方法获得θ的0.95置信区间和标准差. 用参数Bootstrap方法获得θ的0.95置信区间和估计标准差. 用非参数Bootstrap方法获得θ的0.95置信区间和估计标准差. 比较两种方法的结果.

(2) 画出参数和非参数Bootstrap观测的直方图, 观察图形给出对$\hat{\theta}$分布的判断.

2.13 在白令海所捕捉的12岁的某种鱼的长度(单位：cm) 样本如下表所示：

长度/cm	64	65	66	67	68	69	70	71	72	73	74	75	77	78	83
数目	1	2	1	1	4	3	4	5	3	3	0	1	6	1	1

你能否同意所声称的12岁的这种鱼的长度的中位数总是在69∼72cm 之间?

2.14 为考察两种生产方法的生产效率是否有显著差异，随机抽取10人用方法A进行生产，抽取12人采用方法B进行生产，并记录下20人的日产量：A方法：92,69,72,40,90,53,85,87,89,88 B方法：78,95,58,65,39,67,64,75,60,80,83,96 请问两种方法的生产效率的影响不同吗?请问用wilcox.test 应该怎样设置假设，得到怎样的结果，该题目可以使用随机游程方法来解决吗?

2.15 社会学家欲了解抑郁症的发病率是否在一年内随季节的不同而不同, 他使用了来自一所大医院的病人数据, 按一年4个季节(比如：冬季=十二月、一月和二月) 依次记录过去五年中第一次被确诊为患抑郁症的病人数(单位：人)，结果如下：

春季	夏季	秋季	冬季	合计
495	503	491	581	2070

请问：发病率是否与季节有关?

2.16 运用模拟方法从标准正态分布中每次抽取样本量n=30的样本进行Wilcoxon 符号秩检验：

（1）分别在显著性水平$\alpha = 0.1, 0.05, 0.01$的条件下，基于对α 的估计结果，即经验显著性水平，得到α 的一个95% 的置信区间。

（2）将(1)中的标准正态分布变为自由度分别为1,2,3,5,10 的t分布，重新做(1) 中的分析。

2.17 两个估计量置信区间长度的平方的期望之比，是度量这两个估计量的效率高低的指标。通过10000 次模拟，每次样本量为30，分别在总体服从N(0,1) 和服从自由度为2的t分布时，比较Hodges‐Lehmann统计量和样本均值的效率（95% 置信区间）。

2.18 有一个标准化的变量X，其分布可表示为$X = (1 - I_\epsilon)Z + cI_\epsilon Z$ ，其中$0 \leqslant \epsilon \leqslant 1$，服从$n = 1$ 且成功概率为ϵ 的二项分布，Z服从标准正态分布，$c > 1$，且I_ϵ和Z 是相互独立的随机变量。当从X 的分布中抽样时，有比例为$(1 - \epsilon)100\%$ 的观测是由分布$N(0, 1)$ 生成，但有比例为ϵ100%的观测是由分布$N(0, c^2)$ 生成的，后者的观测大多为异常值。我们称X服从分布$CN(c, \epsilon)$。

（1）使用R函数中的rbinom和rnorm，自行编写一个函数，从分布ϵ100% 中抽取样本量为n的随机样本；

（2）从分布$N(0, 1)$和$CN(16, 0.25)$中各抽取样本量为100 的样本，分别制作样本直方图和箱线图，比较结果。

案例与讨论1：排球比赛中的局点

案例背景

北京时间2018年10月19日，2018年女排世锦赛半决赛，中国女排对阵意大利女排。最终苦战5 局，中国女排2-3惜败意大利无缘决赛。事后，中国队的教练组对比赛进行了复盘，试图通过统计分析的方法找到中国女排失利的主要原因以及各局比赛的转折点。对此，你有什么看法？

数据说明与约定

1．原始数据介绍

本案例所使用的原始数据分为三栏，分别记录了每回合的得失分情况、得失分时间以及该回合所属的局数。其中得失分情况由0-1 序列表示，0 代表中国队失分而1代表得分。

2．“连续游程”

游程（run）是相同的事件或符号组成的序列，我们将一段连续的得分或失分定义为“连续游程”。在排球比赛中连续游程具有重要的意义，可证明只有连续游程出现，比赛才能够分出胜负，不仅如此，连续的得分对于球队气势的提升也是十分重要的，因此我们在本题中着重研究连续游程。定义某队的“连续游程数”为该队各段连续得分（失分）游程中的回合数之和除以2。显然，一队连续游程数占优同比分占优之间存在着很强的正相关性（可自行模拟证明）。

3．“胜率”

不妨将中国队的得分与失分看作0-1马尔科夫链，有状态之间的转移矩阵：

$$\begin{bmatrix} P_{00} & P_{01} \\ P_{10} & P_{11} \end{bmatrix}$$

我们可以根据某点及其之前的比赛得失分情况，计算得到该点的转移矩阵，例如$P_{00} = \frac{n_{00}}{(n_{00}+n_{01})}$。只要转移概率均大于0，则由此转移矩阵决定的马尔科夫链是遍历的，其具有唯一的极限分布(P_0, P_1)。$P_1 = \frac{1-P_{00}}{2-P_{00}-P_{11}}$ 则代表，在此转移概率下，经过足够多的轮数后，中国队每回合得分的概率。因此，我们可以用P_1 来估计中国队目前的表现是否能使其在多回合后占据优势，不妨称P_1 为“胜率”。

问题

1. 从各局比赛两队的连续游程数角度看，意大利队相对于中国队有哪些优势？

2. 尝试对局点进行定义。根据定义，参考上述给出的数据说明与约定，使用不同方法确定各局局点，并分析所得结果。

提示

下面是提示分析流程图：为比较两支球队各自优势，除了个人技术优势以外，队伍的拼搏精神和队员的心理素质都可以通过比赛得分获得解读。分析中不仅需要看每局比分，更要整理出

各支球队的连续得分能力，尤其是关键场次局点前后队伍的表现，于是整理出如下所示的分析流程图和各局胜率图，如图2.6所示：

案例与讨论2：我们发明了趋势，趋势是我们理解的那样吗？

这里有东北地区96个气象站1961～2005年45年日平均气温和日降水量资料。有学术论文根据数据对东北地区近45年来的气候变化和突变现象进行研究，结果表明: 近45年来,东北地区年平均气温变化在2.45～5.72 摄氏度之间，年平均气温呈现显著上升趋势,1988～1989 年间发生了由低温到高温的突变;东北地区四季平均气温均呈现增高的趋势,其中冬季气温增幅最大,夏季气温增幅最小。东北地区年平均气温和季节平均气温年代际变化亦呈现明显的增高趋势,年平均气温,春季平均气温和冬季平均气温均在1981～1990 年开始变高,夏季平均气温和秋季平均气温在1991～2000 年开始变高. 1982～1983 年间发生了降水量由少到多的突变. 四季降水量变化呈现不同的趋势,其中春季和冬季降水量呈现增多的趋势,夏季和秋季降水量呈现减少的趋势。如以下8张图所示，

问题:

1. 从图2.8来看，请分析这些结果是来自于怎样的模型假设？这些模型在建模过程中产生的结论有什么问题吗？

2. 如果由你来分析这些问题，你的分析流程是怎样的？为什么？

3. 请比较S_1、S_2和S_3三种方法在改数据的趋势判断上的差异。

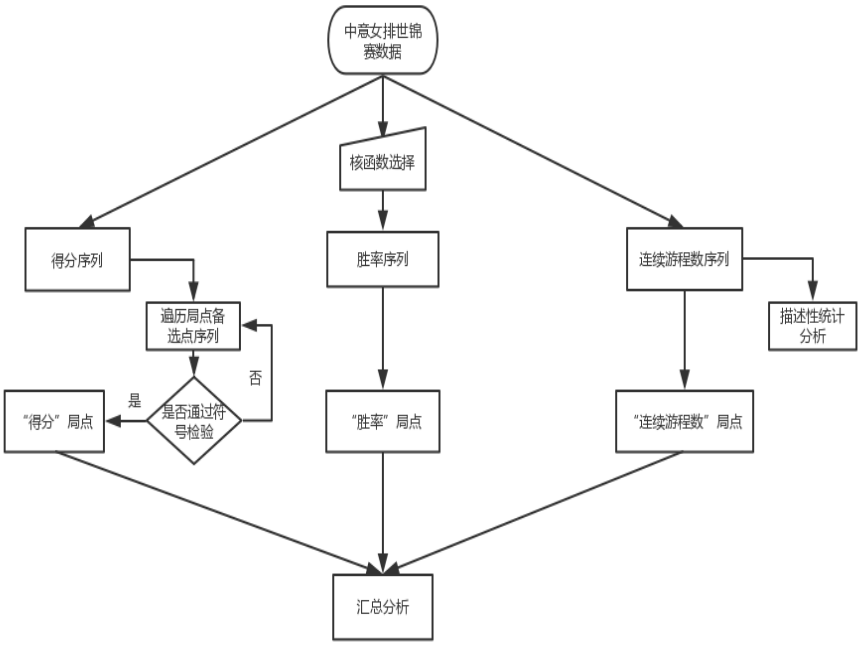

图 2.6　分析流程图

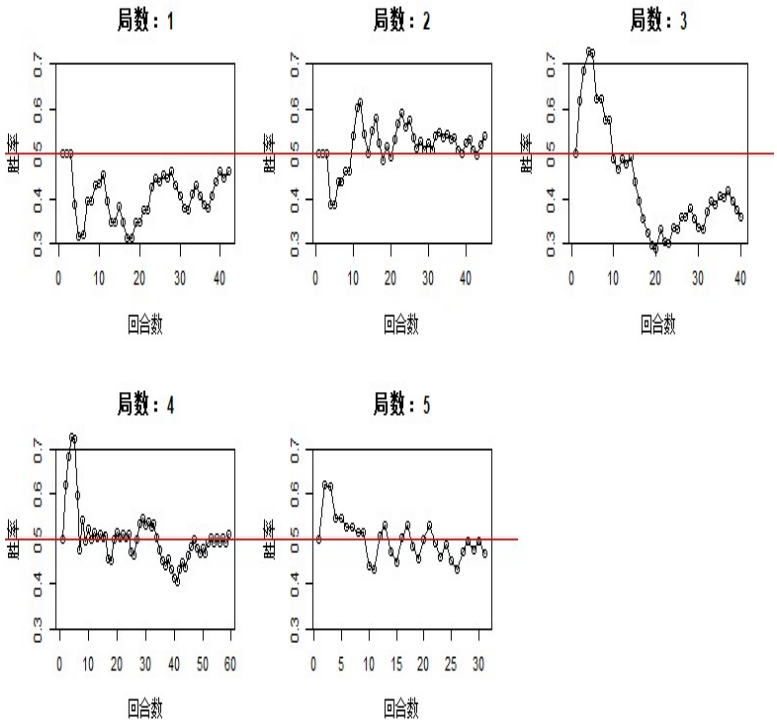

图 2.7　各局胜率表

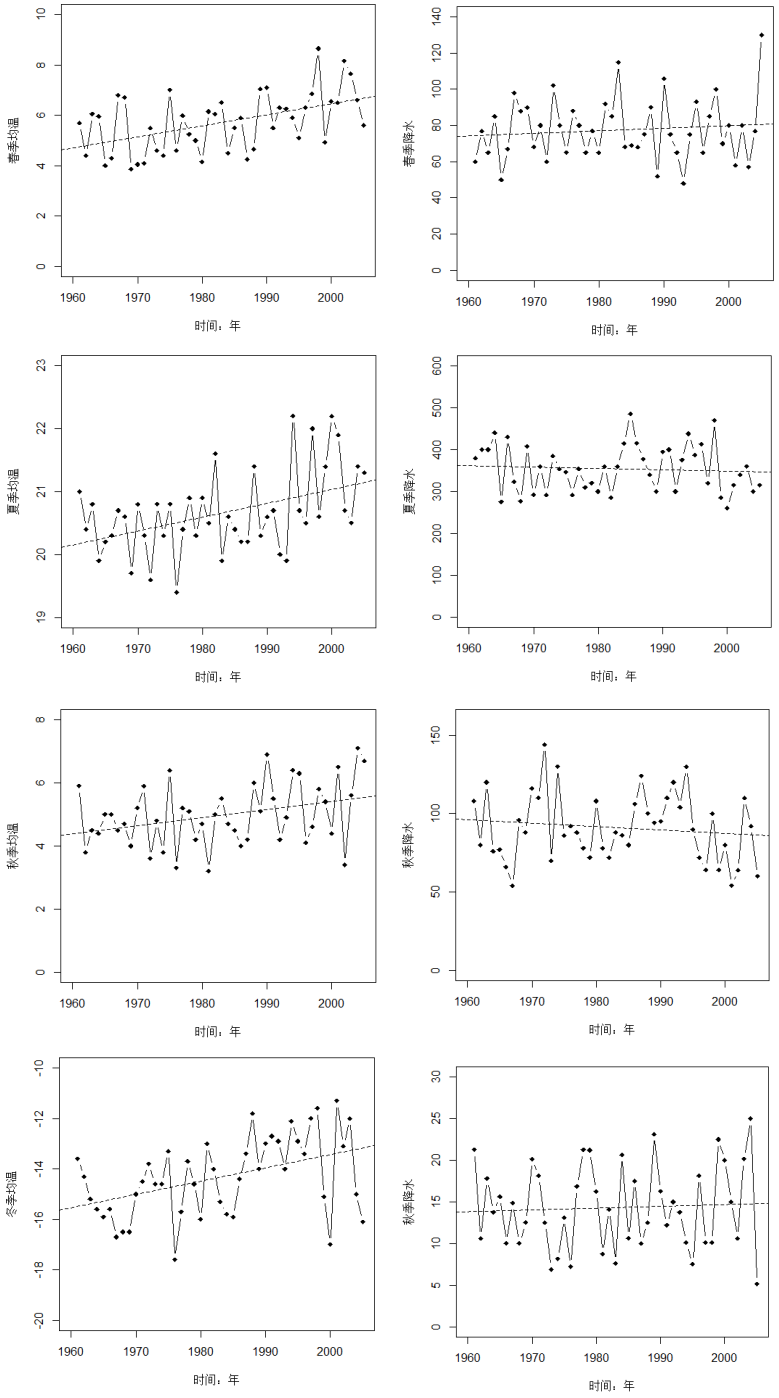

图 2.8 东北地区96个气象站1961-2005年的日平均气温和日降水量