第五章　非参数统计基础

§5.1　非参数检验

在许多实际问题中，常常事先并不知道总体的分布类型，这就要根据抽样的样本所提供的信息，对总体分布的各种假设进行检验。称总体分布未知时所进行的假设检验为非参数假设检验。本届主要介绍检验分布类型和列联表独立性的常用非参数方法——检验法，通常也称为χ^2拟合检验法。

1. 总体分布的拟合检验：

Pearson检验法亦称为χ^2拟合检验法，用于检验假设总体服从某个预先给定的分布$F_0(x)$。

考虑总体分布的检验问题

$$H_0: F(x) = F_0(x)$$

假设分布函数$F_0(x)$形式已知，但包含γ个未知参数，用参数估计法给出未知参数估计。

基于频率稳定于概率，可直观理解χ^2拟合检验思想：

把随机试验结果的全体Ω分成k个互不相容的事件：

$$A_1, A_2, \cdots, A_k$$

在假设H_0下，记$p_i = P\{A_i\}$，A_i发生的频率为f_i/n，其中f_i表示事件A_i在n次试验中发生的次数。

对于给定的分法$A_1, A_2, \cdots, A_k$，根据大数定律，当样本容量n越来越大时，频率f_i/n稳定于概率p_i。

频率f_i/n和p_i之间的差异程度可以反映出$F_0(x)$是否为总体的真实分布。

Pearson统计量：$\chi^2=\sum_{i=1}^{k}\frac{(f_i-np_i)^2}{np_i}$

H_0成立时，χ^2有取偏小值的趋势，所以可以将χ^2的大小作为检验的标准。

为了进行检验，还需要知道χ^2的概率分布，Pearson定理就是给出了它的渐近分布。

定理5.1.1（Pearson定理）：若样本容量n充分大$(n \geq 50)$，则无论总体服从何种分布$F_0(x)$，统计量

$$\chi^2 = \sum_{i=1}^{k} \frac{(f_i - np_i)^2}{np_i}$$

总是近似地服从自由度为$k-1$的χ^2分布，其中$F_0(x)$完全确定，不含任何未知参数。

具体检验过程如下：

（1）将$(-\infty, +\infty)$分成k个互不相交的区间$A_i = (a_i, a_{i+1}], i = 1, \ldots, k$，其中$a_1, a_{k+1}$可分别取$-\infty, +\infty$。区间的划分方法应视具体情况而定。

（2）计算概率$p_i = P\{X \in A_i\} = P\{a_i < X \leq a_{i+1}\} = F_0(a_{i+1}) - F_0(a_i)$并计算$np_i$，称为理论频数。

（3）计算样本$x_1, x_2, \ldots, x_n$落在$(a_i, a_{i+1}]$中个数f_i，称为实际频数。

（4）计算检验统计量的值

$$\chi^2 = \sum_{i=1}^{k} \frac{(f_i - np_i)^2}{np_i} \cong \chi^2(k-1)$$

（5）对给定的α，查临界值$\chi^2_{1-\alpha}(k-1)$

（6）若$\chi^2 \geq \chi^2_{1-\alpha}(k-1)$，则拒绝$H_0$；否则接受$H_0$

在许多实际问题中，假设H_0只指定了分布$F_0(x;\theta_1,\cdots,\theta_\gamma)$的具体函数形式，而它包含$\gamma$个独立的未知参数。

处理此情形的一种自然做法是：在假设H_0成立的条件下，先求出γ个参数$\theta_1,\cdots,\theta_\gamma$的极大似然估计$\widehat{\theta_1},\cdots,\widehat{\theta_\gamma}$，然后代入原式，有$\hat{p}_i=\hat{p}\{X\in A_i\}=F_0(a_{i+1};\widehat{\theta_1},\cdots,\widehat{\theta_\gamma})-F_0(a_i;\widehat{\theta_1},\cdots,\widehat{\theta_\gamma}),i=1,\cdots,k$

用$\hat{p}_i$替换原式中p_i，有检验统计量

$$\chi^2 = \sum_{i=1}^{k} \frac{(f_i - n\hat{p}_i)^2}{n\hat{p}_i}$$

可以证明当样本量n充分大时，上述χ^2近似地服从$\chi^2(k-\gamma-1)$。

对给定的α，若$\chi^2 \geq \chi^2_{1-\alpha}(k-\gamma-1)$，则拒绝$H_0$；否则接受$H_0$。

注5.1.2：

χ^2拟合检验是在极限意义下获得，所以在使用时样本容量n必须足够大，同时还要求np_i不能太小。由实际应用的经验知，通常要求$n \geq 50, np_i > 5$，且最好$np_i > 10$。否则，应适当合并当初所划分区间，使np_i满足上述要求。

例题5.1.3：在某种铀的实验中，每隔一定时间观察一次由某种铀放射的到达计数器上的a粒子数X，共观察了100次，其结果如下：

i	0	1	2	3	4	5	6	7	8	9	10	11	$\sum$
f_i	1	5	16	17	26	11	9	9	2	1	2	1	100

其中f_i是观察到有i个a粒子数的次数。从理论上知道X应服从Poisson分布：

$$P\{X=i\}=\frac{\lambda^i}{i!}e^{-\lambda},\ i=0,1,2,\ldots$$

试问在显著性水平$a = 0.05$下，理论上的结果是否符合实际？

解：考虑假设检验问题

$$H_0: P\{X = i\} = \frac{\lambda^i}{i!}e^{-\lambda}$$

当H_0成立时，由极大似然法可得λ的估计为$\hat{\lambda} = \bar{x} = 4.2$。由于总体$X$是离散型的，所以

$$\hat{p}_i = P\{X = i; \hat{\lambda}\} = \frac{\hat{\lambda}^i}{i!}e^{-\hat{\lambda}},\ i = 0, 1, 2\ldots$$

所以理论频数为$n\hat{p}_i = 100\frac{4.2^i}{i!}e^{-4.2}$，将其计算结果列于下表中。

f_i	$n\hat{p}_i$	$f_i - n\hat{p}_i$	$(f_i - n\hat{p}_i)^2$	$(f_i - n\hat{p}_i)^2/n\hat{p}_i$
1	1.5	−1.8	3.24	0.415
5	6.3			
16	13.2	2.8	7.84	0.594
17	18.5	−1.5	2.25	0.122
26	19.4	6.6	43.56	2.245
11	16.3	−5.3	28.09	1.723
9	11.4	−2.4	5.7	0.505
9	6.9	2.1	4.41	0.609
2	3.6			
1	1.7			0.014
2	0.7	−0.3	0.09	
1	0.3			
$\sum$				6.257

把理论频数小于5的组合并，新的一组内理论频数 $n\hat{p}_i \geq 5$，其并组情况在下表中第二列用大括号表示

f_i	$n\hat{p}_i$	$f_i - n\hat{p}_i$	$(f_i - n\hat{p}_i)^2$	$(f_i - n\hat{p}_i)^2 / n\hat{p}_i$
1	1.5	−1.8	3.24	0.415
5	6.3			
16	13.2	2.8	7.84	0.594
17	18.5	−1.5	2.25	0.122
26	19.4	6.6	43.56	2.245
11	16.3	−5.3	28.09	1.723
9	11.4	−2.4	5.7	0.505
9	6.9	2.1	4.41	0.609
2	3.6			
1	1.7			0.014
2	0.7	−0.3	0.09	
1	0.3			
$\sum$				6.257

并组后得到$k=8$，并且在计算概率时，估计了一个参数，从而统计量χ^2的自由度为$8-1-1=6$：

$$\chi^2_{1-\alpha}(k-\gamma-1)=\chi^2_{0.95}(6)=12.592$$

因此，拒绝域为 $W=\{\chi^2\geq 12.592\}$。而$\chi^2=6.257\notin W$，所以在显著性水平$\alpha=0.05$下可以接受H_0，认为简单样本是来自Poisson分布的总体，且$\hat{\lambda}=4.2$

1.2. 二维列联表的独立性检验：

在实际问题中，常常要考察二维总体，或一维总体中个体的两个指标X与Y。

需要检验指标X与Y之间是否相互独立，这可由χ^2拟合检验来完成，这种检验也称列联表独立性检验

本节下面的内容仅限于讨论二维列联表独立性检验，类似的可以推广到三维或三维以上的列联表独立性检验。

例题5.1.4：为了调查慢性气管炎与吸烟习惯的关系在50岁以上的人群中抽查了339名，得到二维列联表习惯上称为2 × 2列联表，如下所示：

	患慢性气管炎	未患慢性气管炎	合　计
吸　烟	43	162	205
不吸烟	13	121	134
合　计	56	283	339

若用X和Y分别表示调查对象的两个指标：是否吸烟和是否患慢性气管炎，则它们分别有两种状态：

“吸烟”和“不吸烟”，“患慢性气管炎”和“未患慢性气管炎”。

通常也称指标所取的状态为水平。这样，判断患慢性气管炎与吸烟习惯的关系，实际上就是要判断两个指标X和Y是否独立。

一般地，设有两个指标X与Y，需要检验的假设为H_0：X与Y相互独立。

为此，将两个指标X与Y的取值范围分别分成r, s个互不相交的类$A_1, A_2, \ldots, A_r$和$B_1, B_2, \ldots, B_s$。

从总体中抽取容量为n的简单样本，用n_{ij}表示样本中既属于A_i类，又属于B_j类的个体数，称其为频数

令$n_{i\cdot}=\sum_{j=1}^{s}n_{ij}$，$n_{\cdot j}=\sum_{i=1}^{r}n_{ij}$，显然$n=\sum_{i=1}^{r}\sum_{j=1}^{s}n_{ij}$。将$rs$个数$n_{ij}$列成二维列表，如下表所列，称为$r\times s$列联表。

Y / X	B_1	$\cdots$	B_j	$\cdots$	B_s	$\sum$
A_1	n_{11}	$\cdots$	n_{1j}	$\cdots$	n_{1s}	$n_{1\cdot}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
A_i	n_{i1}	$\cdots$	n_{ij}	$\cdots$	n_{is}	$n_{i\cdot}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
A_r	n_{r1}	$\cdots$	n_{rj}	$\cdots$	n_{rs}	$n_{r\cdot}$
$\sum$	$n_{\cdot 1}$	$\cdots$	$n_{\cdot j}$	$\cdots$	$n_{\cdot s}$	n

如果记

$$p_{ij} = P\{X \in A_i, Y \in B_j\},\ \ i = 1, 2, \cdots, r; j = 1, 2, \cdots, s$$

$$p_{i\cdot} = P\{X \in A_i\},\ \ i = 1, 2, \cdots, r$$

$$p_{\cdot j} = P\{Y \in B_j\},\ \ j = 1, 2, \cdots, s$$

则显然有

$$p_{i\cdot} = \sum_{j=1}^{s} p_{ij},\ \ p_{\cdot j} = \sum_{i=1}^{r} p_{ij},\ \ \sum_{i=1}^{r} p_{i\cdot} = 1,\ \ \sum_{j=1}^{s} p_{\cdot j} = 1$$

当假设H_0成立时，有$p_{ij} = p_{i\cdot}p_{\cdot j}$，所以列联表中独立性检验实际上是检验假设

$$H_0 : p_{ij} = p_{i\cdot}p_{\cdot j},\ i = 1, 2, \cdots, r; j = 1, 2, \cdots, s$$

此时可以用χ^2拟合检验法来进行检验。

在H_0中，$r+s$个数$p_{1\cdot}, p_{2\cdot}, \ldots, p_{r\cdot}$和$p_{\cdot 1}, p_{\cdot 2}, \ldots, p_{\cdot s}$属于未知参数。

这$r+s$个参数中仅有$r+s-2$个独立。

不妨设$p_{r\cdot}=1-\sum_{i=1}^{r-1} p_{i\cdot}$及$p_{\cdot s}=1-\sum_{j=1}^{s-1} p_{\cdot j}$，即$r+s-2$个独立参数是$p_{i\cdot}(i=1,2,..,r-1)$及$p_{\cdot j}(j=1,2,..,s-1)$

要建立检验式H_0的检验统计量，需要求出这些未知参数的极大似然估计，分别为：

$$\hat{p}_{i\cdot}=\frac{n_{i\cdot}}{n},\ i=1,2,\ldots,r. \qquad \hat{p}_{\cdot j}=\frac{n_{\cdot j}}{n},\ j=1,2,\ldots,s$$

检验统计量可以取为

$$\chi^2=\sum_{i=1}^{r}\sum_{j=1}^{s}\frac{(n_{ij}-n\hat{p}_{ij})^2}{n\hat{p}_{ij}}=\sum_{i=1}^{r}\sum_{j=1}^{s}\frac{(n_{ij}-n\hat{p}_{i\cdot}\hat{p}_{\cdot j})^2}{n\hat{p}_{i\cdot}\hat{p}_{\cdot j}}$$

把$\hat{p}_{i\cdot}$和$\hat{p}_{\cdot j}$代入，可得检验式中假设H_0的检验统计量为

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{s} \frac{\left(n_{ij} - \frac{n_{i\cdot}n_{\cdot j}}{n}\right)^2}{\frac{n_{i\cdot}n_{\cdot j}}{n}}.$$

由于假设H_0中的独立参数个数是$r+s-2$，所以当假设H_0成立时，我们给出的检验统计量χ^2近似地服从自由度为$rs-(r+s-2)-1=(r-1)(s-1)$的χ^2分布。

对给定的α，检验的拒绝域为：

$$W=\{(x_1, x_2, \ldots, x_n): \chi^2 \geq \chi^2_{1-\alpha}((r-1)(s-1))\}.$$

即当$\chi^2 \geq \chi^2_{1-\alpha}((r-1)(s-1))$时，拒绝假设$H_0$，认为$X$与$Y$不相互独立，否则认为$X$与$Y$相互独立。

对上述例题，给定显著性水平$\alpha = 0.01$，由于$r = 2, s = 2$，查表知

$$\chi^2_{1-\alpha}\big((r-1)(s-1)\big) = \chi^2_{0.99}(1) = 6.635$$

所以拒绝域为$W = \{\chi^2 \geq 6.635\}$

而检验统计量值$\chi^2 = 7.4688 \in W$

因而拒绝原假设H_0，即认为慢性气管炎与吸烟有密切关系。

§5.2　秩和检验

1. 无结点数据的秩及性质：

定义5.2.1：设样本$X_1, X_2, \cdots, X_n$是取自总体X的简单随机样本，$X_1, X_2, \cdots, X_n$中不超过X_i的数据个数$R_i = \sum_{j=1}^{n} \mathbb{I}(X_j \leq X_i)$，称$R_i$为$X_i$的秩，$X_i$是第$R_i$个顺序统计量，$X_{(R_i)} = X_i$。令$R = (R_1, R_2, \cdots, R_n)$，$R$是由样本产生的统计量，称为秩统计量。

定理5.2.2：对于简单随机样本，$\boldsymbol{R=(R_1,R_2,\cdots,R_n)}$等可能取$\boldsymbol{(1,2,\cdots,n)}$的任意$\boldsymbol{n!}$个排列之一，$\boldsymbol{R}$在由$\boldsymbol{(1,2,\cdots,n)}$的所有可能的排列组成的空间上是均匀分布，即对于$\boldsymbol{(1,2,\cdots,n)}$的任意排列$\boldsymbol{(i_1,i_2,\cdots,i_n)}$有

$$\boldsymbol{P(R=(i_1,i_2,\cdots,i_n))=\frac{1}{n!}}$$

上述定理5.2.2给出的是$R_1, R_2, \cdots, R_n$联合分布。类似地，每一个R_i在空间$\{1, 2, \cdots, n\}$上有均匀分布；每一对(R_i, R_j)在空间$\{(r, s)|r, s = 1, 2, \cdots, n; r \neq s\}$上有均匀分布。以推论的形式表示如下：

推论5.2.3：对于简单随机样本，对任意$r, s = 1, 2, \cdots, n$，$r \neq s$及$i \neq j$，有

$$P(R_i = r) = \frac{1}{n},\ P(R_i = r, R_j = s) = \frac{1}{n(n-1)}$$

推论5.2.4：对于简单随机样本，

$$E(R_i) = \frac{n+1}{2}$$

$$Var(R_i) = \frac{(n+1)(n-1)}{12}$$

$$Cov(R_i, R_j) = -\frac{n+1}{12}$$

证明：

$$E(R_i) = \sum_{i=1}^{n} i \cdot \frac{1}{n} = \frac{n+1}{2}$$

$$Var(R_i) = \sum_{i=1}^{n} i^2 \cdot \frac{1}{n} - [E(R_i)]^2$$
$$= \frac{n(n+1)(2n+1)}{6} \cdot \frac{1}{n} - \frac{(n+1)^2}{4}$$
$$= \frac{(n+1)(n-1)}{12}$$

$$Cov(R_i, R_j) = E(R_i - E(R_i))E(R_j - E(R_j))$$
$$= \sum_{i,j,i\neq j} ((i - \frac{n+1}{2})(j - \frac{n+1}{2}) \cdot \frac{1}{n(n-1)})$$

$$Cov(R_i, R_j)$$
$$= [\sum_{i=1}^{n}\sum_{j=1}^{n}(i - \frac{n+1}{2})(j - \frac{n+1}{2}) - \sum_{j=1}^{n}(j - \frac{n+1}{2})^2] \cdot \frac{1}{n(n-1)}$$
$$= -\frac{n+1}{12}$$

这些结果说明，对于独立同分布样本来说，秩的分布和总体分布无关。

2. 有结数据的秩：

在许多情况下，数据有重复数据，称数据中存在结(tie)。结的定义如下

定义5. 2. 5：设样本$X_1, X_2, \cdots, X_n$取自总体X的简单随机抽样，将数据排序后，相同的数据点组成一个“结”，称重复数据的个数为结长。

假设有样本量为7的数据：

3.8 3.2 1.2 1.2 3.4 3.2 3.2

其中有4个结，$x_2 = x_6 = x_7 = 3.2$，结长3；$x_3 = x_4 = 1.2$，结长2；$x_1 = 3.8$和$x_5 = 3.4$的结长都为1。如有重复数据，则将数据从小到大排序后，$(R_1, R_2) = (1, 2)$，也可以等于$(2, 1)$，则秩不唯一。一般采用秩平均法处理有结数据的秩。

定义5.2.6：将样本$X_1, X_2, \ldots, X_n$从小到大排序后，若$X_{(1)} = \cdots = X_{(\tau_1)} < X_{(\tau_1+1)} = \cdots = X_{(\tau_1+\tau_2)} < \cdots < X_{(\tau_1+\tau_2+\cdots+\tau_{g-1}+1)} = \cdots = X_{(\tau_1+\tau_2+\cdots+\tau_g)}$，$g$是样本中结的个数，$\tau_i$是第$i$个结的长度，$(\tau_1, \tau_2, \cdots, \tau_g)$是$g$个正整数，$\sum_{i=1}^{g} \tau_i = n$，称$(\tau_1, \tau_2, \cdots, \tau_g)$为结统计量。

第i组样本的秩都相同，是第i组样本原秩的平均：

$$r_i = \frac{1}{\tau_i} \sum_{k=1}^{\tau_i} (\tau_1 + \cdots + \tau_{i-1} + k) = \tau_1 + \cdots + \tau_{i-1} + \frac{1+\tau_i}{2}$$

例5.2.7：样本数据为12个数，其值、秩和结统计量（用τ_i表示，为第i个结中的观测值数量）如下表所示

观测值	2	2	4	7	7	7	8	9	9	9	9	10
秩	1.5	1.5	3	5	5	5	7	9.5	9.5	9.5	9.5	12

其中有6个结，每个结长分别为2，1，3，1，4，1。

3. 秩和检验法：

设从总体$F(x)$与$G(x)$中分别抽取了容量为n_1, n_2的样本，$X_1, X_2, \cdots, X_{n_1}$和$Y_1, Y_2, \cdots, Y_{n_2}$，欲检验假设$H_0: F(x) = G(x)$。

把两个样本的观测数据合在一起按照大小次序排列并统一编号，规定每个数据在排列中所对应的序数称为该数的秩，对于相同的数值则用它们序数的平均值（必要时四舍五入）来作秩。

将容量较小样本的各观测值秩之和记为T，以T作为统计量。如果H_0成立，即$F(x)$与$G(x)$差异不显著，则统计量T就不应该太大或太小。威尔科克逊给出了统计量T的临界值T_1和T_2，使

$$P(T_1 < T < T_2) = 1 - \alpha$$

若$T_1 < T < T_2$，则接受H_0，认为$F(x)$与$G(x)$差异不显著。否则拒绝H_0，认为$F(x)$与$G(x)$差异显著。

例5.2.8：由$F(x)$与$G(x)$获得两组样本：

x_i	2.36	3.14	7.52	3.48	2.76	5.43	6.54	7.41
y_i	4.38	4.25	6.54	3.28	7.21	6.54		

试检验假设$H_0: F(x) = G(x)$，$(\text{取}\alpha = 0.05)$。

将两组样本合在一起由小到大排列，统一编号，并计算出相应的秩，列于下表：

编号	1	2	3	4	5	6	7	8	9	10	11	12	13	14
x_i	2.36	2.76	3.14		3.48			5.43	6.54				7.41	7.52
y_i				3.28		4.25	4.38			6.54	6.54	7.21		
秩	1	2	3	4	5	6	7	8	10	10	10	12	13	14

表中第9、10、11号数值相同，其秩取其和之平均值10，其中y_i一组容量较小，于是统计量$T = 4 + 6 + 7 + 10 + 12 = 49$，对于检验水平$\alpha = 0.05$，$n_1 = 8$，$n_2 = 6$，查表得到$T_1 = 32$，$T_2 = 58$，由于$T_1 < T < T_2$，故接受$H_0$，即认为$F(x)$与$G(x)$无显著差异。

秩和检验表只列到$n_1, n_2 \leq 10$的情形，当其大于10时，统计量T近似服从正态分布：

$$T \sim N\left(\frac{n_1(n_1+n_2+1)}{2}, \frac{n_1 n_2(n_1+n_2+1)}{12}\right)$$

于是可用U检验法，这时选统计量为：

$$U = \frac{T - \frac{n_1(n_1+n_2+1)}{2}}{\sqrt{\frac{n_1 n_2(n_1+n_2+1)}{12}}} \sim N(0,1)$$

则在检验水平α下拒绝域为$W = \{|u| > u_{1-\frac{\alpha}{2}}\}$

4.符号检验法：

为比较甲乙两种酒的优劣，让N个品酒人分别品尝评分,形成配对样本的成对比较模型。设X_i、Y_i分别为第i个品酒人对甲乙酒的评分。记$Z_i = X_i - Y_i, i = 1, \cdots, N$,如果假定$Z_i \sim N(\mu, \sigma^2)$，则甲、乙两酒是否有优劣的问题将转化为原假设$H_0: \mu = 0 \leftrightarrow H_1: \mu \neq 0$的检验问题，这就是样本t检验问题，但是在一些情况下，不一定有根据假定Z_i服从正态分布。所以提出一个替代的方法：对每个评酒人的评分给出一个符号

$$S_i = \begin{cases} +, Z_i > 0 \\ -, Z_i < 0 \\ 0, Z_i = 0 \end{cases}$$

即品酒人给以“+”号表示他认为“甲酒优于乙酒”,另两个符号的意义类推。如此，得到N个符号$S_1, \cdots, S_N$.检验问题:H_0:甲乙两种酒一样好$\leftrightarrow$ H_1:甲乙两种酒不一样的检验就建立在试验结果的这N个符号的基础上,故称为符号检验(sign test)。从统计模型而言,符号检验其实是二项分布参数检验的一个特例。

符号检验的具体方法：

1）小样本方法:记N个试验结果$S_1,\cdots,S_N$中“+”号有n_+个，“−”号有n_-个,其余为0，n=n_++n_-。若H_0成立(两种酒一样好)，则n个非0结果中出现“+”的概率θ=1/2，n_+服从b(n,1/2)。若两种酒确有优劣，则每个结果出现“+”号的概率0≠1/2。若记X=n_+，则问题转化为检验问题 X~ b(n,θ),0≤θ≤1,要检验$H_0:\theta=\frac{1}{2}\leftrightarrow H_1:\theta\neq\frac{1}{2}$.一个水平为α的检验的否定域为$\{X=n_+\geq c$或$X\leq d\}$，其中c和d

的值由下式确定：$\sum_{i=c}^{n}\binom{n}{i}\left(\frac{1}{2}\right)^{n} \leq \frac{\alpha}{2}, d = n - \diamond$ c，c值通过查表获得，d由公式计算。

一个更恰当的方法是计算检验的p值，在此，令由符号$S_1, \cdots, S_N$算得的X=n_+的具体值为x_0，记$x_0^{'} = min\{x_0, n - x_0\}$,则检验的p值为$p = \sum_{i=0}^{x_0^{'}}\binom{n}{i}\left(\frac{1}{2}\right)^{n} + \sum_{i=n-x_0^{'}}^{n}\binom{n}{i}\left(\frac{1}{2}\right)^{n}$.若n为偶数，而 x_0=n/2,则取 p值为 p=1。p值越接近 1，则H_0越可信。例如,给定检验水平α，则当p<α时否定H_0，当p≥α时接受H_0。

本表列出了满足 $P(n_+ \geqslant c) \leqslant \alpha$ 的临界值 c

n \ α	0.01	0.05	0.10	n \ α	0.01	0.05	0.10
5				18	15	13	13
6			6	19	15	14	13
7		7	6	20	16	15	14
8	8	7	7	21	17	15	14
9	9	8	7	22	17	16	15
10	10	9	8	23	18	16	16
11	10	9	9	24	19	17	16
12	11	10	9	25	19	18	17
13	12	10	10	26	20	18	17
14	12	11	10	27	20	19	18
15	13	12	11	28	21	19	18
16	14	12	12	29	22	20	19
17	14	13	12	30	22	20	20

n \ α	0.01	0.05	0.10	n \ α	0.01	0.05	0.10
31	23	21	20	41	29	27	26
32	24	22	21	42	29	27	26
33	24	22	21	43	30	28	27
34	25	23	22	44	31	28	27
35	25	23	22	45	31	29	28
36	26	24	23	46	32	30	28
37	26	24	23	47	32	30	29
38	27	25	24	48	33	31	29
39	28	26	24	49	34	31	30
40	28	26	25	50	34	32	31

2）大样本方法：若n很大时，根据中心极限定理，若H_0成立且n→∞时，有：$U=\frac{X-E(X)}{\sqrt{D(X)}}=\frac{X-n/2}{\sqrt{n/4}}=\frac{2X-n}{\sqrt{n}}\xrightarrow{\mathcal{L}}N(0,1)$，因此前面的检验问题水平近似为α的检验否定域是$\{X:|U|>u_{\alpha/2}\}$，其中$u_{\alpha/2}$为标准正态分布的上侧$\alpha/2$分位数。有时检验目的是从“甲不优于乙”和“甲优于乙”中选择一个，以前者为原假设，则检验问题可表示为$X\sim b(n,\theta),0\le\theta\le1$,而$H_0':\theta\le\frac{1}{2}\leftrightarrow H_1':\theta>\frac{1}{2}$。当n充分大时,可用大样本方法,其检验水平近似为α的检验的否定域是$\{X:U>u_\alpha\}$

5.符号秩和检验法：

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

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:$\{W^+ \le d$或$W^+ \ge c\}$

在小样本场合$(n \le 20)$，后面附表给出了$P(W^+ \ge c) \le \alpha$

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$\left\{|(W^+)^*| \ge u_{1-\frac{\alpha}{2}}\right\}$

附表 11　　符号秩和检验临界值

$P\ (W^{+}\geqslant C)\leqslant \alpha$

n	α			
	0.01	0.025	0.05	0.10
4	11	11	11	10
5	16	16	15	13
6	22	21	19	18
7	28	26	25	23
8	35	33	31	28
9	42	40	37	35
10	50	47	45	41
11	59	56	53	49
12	68	65	61	57
13	79	74	70	65
14	90	85	80	74
15	101	95	90	84
16	113	107	101	94
17	126	119	112	105
18	139	131	124	116
19	153	144	137	128
20	167	158	150	141

例5.2.9：某一产品有两种牌号，请用户进行评判其优劣，现选了13个用户，对这两种牌号的产品分别评分，结果如下：

i（用户）	1	2	3	4	5	6	7	8	9	10	11	12	13
x_i（甲）	55	32	41	50	60	48	39	45	48	46	52	45	44
y_i（乙）	45	37	43	55	44	50	43	46	51	47	55	46	41

试在$\alpha=0.10$水平上能否认为用户对两种产品评价一致?

解:　可以认为用户对甲产品的评分为总体X，有分布F(x)，对乙产品的评分为总体Y，有分布G(y)，对两者评价一致即表示$F(x)=G(x)$。现要检验假设H_0：$F(x)=G(x)$

用符号秩和检验，在n=13，α=0.10时，由附表查得c=74，

从而$d=\frac{13\times14}{2}-74=17$，则拒绝域为$\{W^+\leq 17$或$W^+\geq 74\}$

为求出样本对应的W^+的统计量的值，将其计算列成下表。

表 7.16　　W^+的计算

i	x_i	y_i	V_i	$\lvert W_i\rvert=\lvert x_i-y_i\rvert$	R_i
1	55	45	1	10	12
2	32	37	0	5	10
3	41	43	0	2	4.5
4	50	55	0	5	10
5	60	44	1	16	13
6	48	50	0	2	4.5
7	39	43	0	4	8
8	45	46	0	1	2
9	48	51	0	3	6.9
10	46	47	0	1	2
11	50	55	0	5	10
12	45	46	0	1	2
13	44	41	1	3	6.5

从而$W^+=12+13+6.5=31.5$，由于样本落在接受域内，因而可以认为用户对两种产品评价无明显差异。

若对本例用大样本检验，则在$\alpha=0.10$时，$u_{0.95}=1.96$，则拒绝域为$\{|(W^+)^*|\geq 1.96\}$

现由样本求得$E(W^+)=\frac{13\times 14}{4}=45.5$，$Var(W^+)=\frac{13\times 14\times 27}{24}=204.75$，从而$(W^+)^*=\frac{31.5-45.5}{\sqrt{204.75}}=-0.98$，由于$|(W^+)^*|<1.96$，故样本仍落在接受域内，结论同上。

§5.3　核密度估计

1. 目标

统计问题：给定一组独立同分布样本$X_1, X_2, \cdots, X_n$，

如何估计其密度函数$f(x)$?

2. 动机

直方图对于所有发生在$(t_{k-1}, t_k]$中所有样本都同等地对待。简单起见我们考虑$[0,1]$区间，如果有两个样本取值$0^+, 1^-$，那么我们应该认为前者表达的信息是分布在0周围有密度，而后者应该是1周围有密度，而不能把两个完全同等的看待。

因此，对于直方图，给定宽度δ，我们认为样本X_i应该反映区间$[X_i - \delta/2, X_i + \delta/2]$的信息，例如我们定义：$K_\Delta(x) = \mathbb{I}(-\delta/2 \le x \le \delta/2)/\delta$

那么对应的估计为：

$$f_n(x) = \frac{1}{n}\sum_{i=1}^{n} K_\delta(X_i - x)$$

进一步我们把宽度δ看成参数，直接定义$K(x)=\mathbb{I}(|x|\leq 1)/2$，得到了一般的核密度估计形式：

定义5.3.1：

$$\hat{f}_h(x)=\frac{1}{n}\sum_{t=1}^{n}\frac{1}{h}K(\frac{X_t-x}{h})=\int K_h(\mu-x)d\hat{F}(\mu)$$

其中$K_h(\cdot)=K(\cdot/h)$称为核函数，h是窗口参数。

3. 核函数

核函数应该满足的条件为：

- 对于任意x，$K(x) \geq 0$；
- $\int K(x)dx = 1$

因此，核函数是一个密度函数。反之，密度函数可以成为核函数。

实际中，考虑到使用核函数的目的，一般的核函数应该是单峰的、对称的。

常用的核函数如下：

$$K(\mu)=\frac{1}{2}\mathbb{I}(|\mu|\leq 1)$$

$$K(\mu)=(1-\mu)\mathbb{I}(|\mu|\leq 1)$$

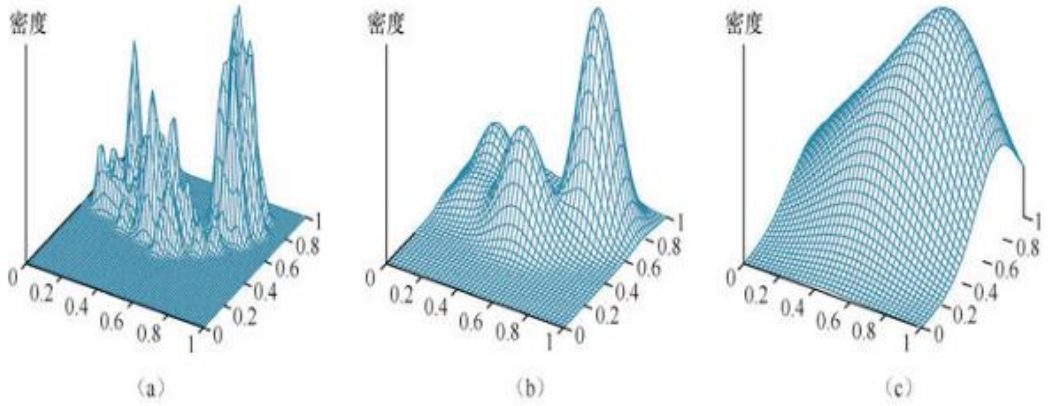
$$K(\mu)=\frac{3}{4}(1-\mu^2)\mathbb{I}(|\mu|\leq 1)$$

$$K(\mu)=\frac{1}{\sqrt{2\pi}}e^{-\mu^2/2}$$

$$K(\mu)=\frac{\pi}{4}cos(\frac{\pi}{2}\mu)\mathbb{I}(|\mu|\leq 1)$$

$$K(\mu)=1/(e^{\mu}+2+e^{-\mu})$$

例5.3.1 为了在密度估计中应用核函数，假设每个数据点都将生成一个与自己相关的密度函数。举个例子来说，我们可以采用在每个维度上标准差均为w的球形高斯核。那么对于查询点x，我们给出的密度估计值为数据核函数的均值：

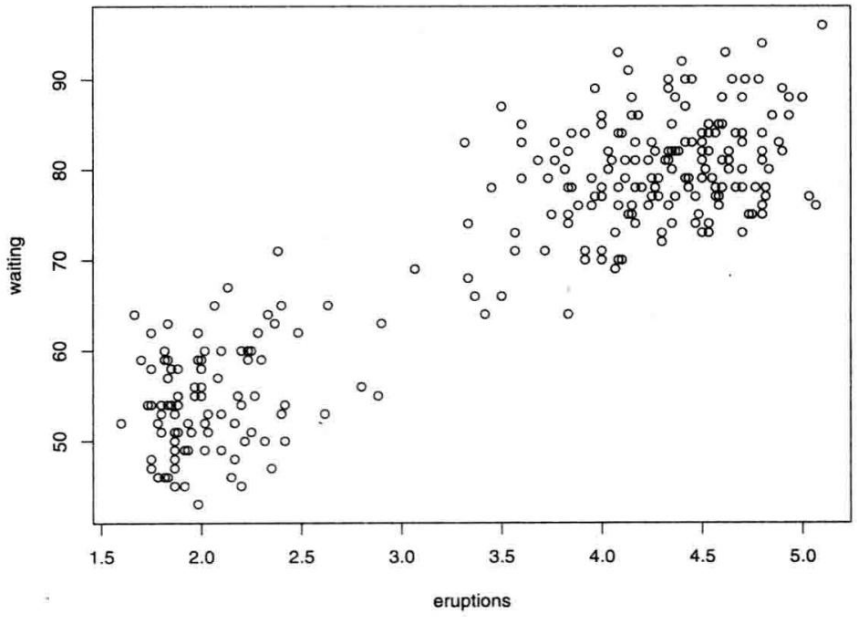


$$P(\mathbf{x}) = \frac{1}{N}\sum_{j=1}^{N}\mathcal{K}(\mathbf{x},\mathbf{x}_j)\,.$$

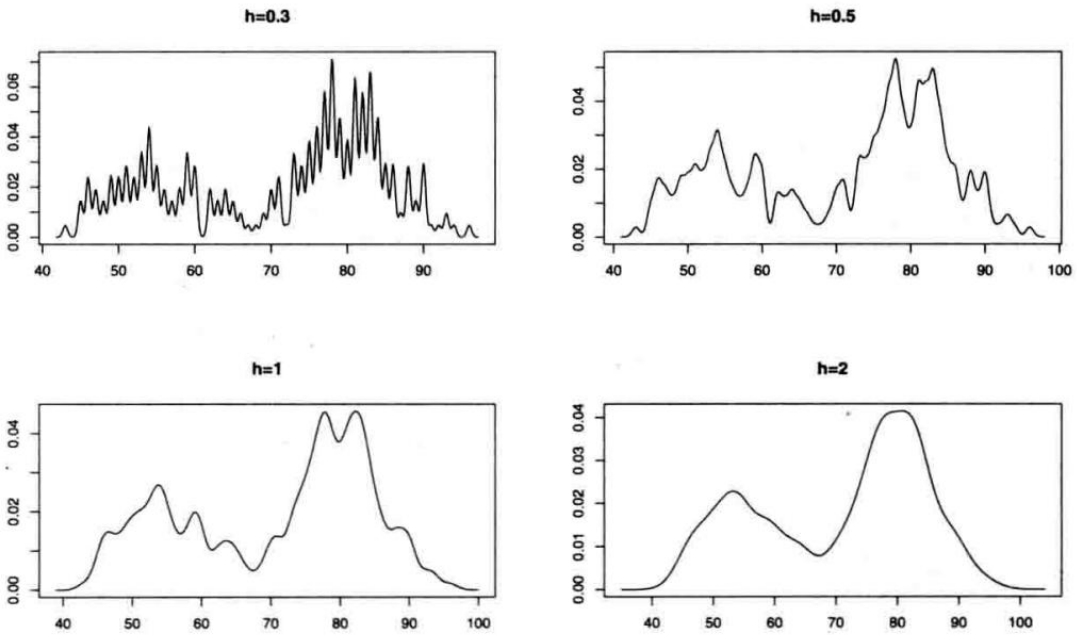
$$\mathcal{K}(\boldsymbol{x},\boldsymbol{x}_j) = \frac{1}{(w^2\sqrt{2\pi})^d}\mathrm{e}^{-\frac{D(\boldsymbol{x},\boldsymbol{x}_j)^2}{2w^2}}$$

使用核函数进行密度估计，所用数据为图2b中的数据，分别采用了w = 0.02、0.07和0.20的高斯核。其中w = 0.07的结果最接近真实情况

例5.3.2：在美国黄石国家公园有一个间歇式喷泉，由于它的喷发保持较明显的规律性，人们称之为老忠实(Old Faithful)。下图为其喷发持续时间和间隔时间的散点图：

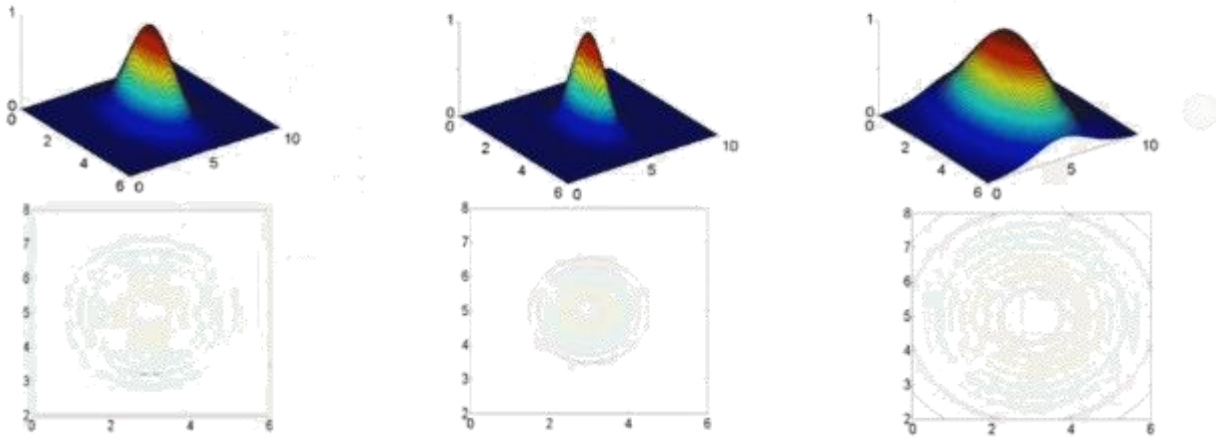


下图为对老忠实喷泉的间隔时间所作的核估计。其中h取了四个不同的值：$h = 0.3, 0.5, 1, 2$。从图上反映带宽对图形的影响。此处的核函数为标准正态密度函数。



例5.3.3：高斯核函数$f_1 = exp\left(-\frac{\|x-l^{(1)}\|^2}{2\sigma^2}\right)$，假设有两个特征值$x_1$和$x_2$，第一个标记点是$l^{(1)}$其位于$(3,5)$，即$l^{(1)} = \begin{bmatrix}3\\5\end{bmatrix}$，当改变参数$\sigma^2$，当$\sigma^2 = 0.5$，发现核函数看起来还是相似的，只是这个突起的宽度变窄了，等值线图也收缩了一些。相反地如果增大了 σ^2的值，例如假设$\sigma^2 = 3$，突起的宽度变宽，等值线图也变得更加平坦。

$\sigma^2 = 1$　　$\sigma^2 = 0.5$　　$\sigma^2 = 3$



4. 窗宽的经验选取

对于$K(\mu)=\frac{1}{\sqrt{2\pi}}e^{-\mu^2/2}$，$h_{opt}\approx 1.06\hat{\sigma}n^{-\frac{1}{5}}$

对于$K(\mu)=\frac{3}{4}(1-\mu^2)\mathbb{I}(|\mu|\leq 1)$，$h_{opt}\approx 2.34\hat{\sigma}n^{-\frac{1}{5}}$

5. 多元密度估计

多元密度估计是一元的推广，对于多元数据，同样可以有多元的核估计。假定$\boldsymbol{x} \in \mathbb{R}^{\boldsymbol{d}}$，则多元密度估计可以为：$\hat{\boldsymbol{f}}_{\boldsymbol{h}}(\boldsymbol{x}) = \frac{\boldsymbol{1}}{\boldsymbol{n}} \sum_{\boldsymbol{i}=\boldsymbol{1}}^{\boldsymbol{n}} \frac{\boldsymbol{1}}{\boldsymbol{h}^{\boldsymbol{d}}} \boldsymbol{K}\left(\frac{\boldsymbol{x}-\boldsymbol{x}_{\boldsymbol{i}}}{\boldsymbol{h}}\right)$

此处的$\boldsymbol{h}$不一定对所有维度都一样，每一个维度可以选择独立的$\boldsymbol{h}$，核函数应满足:

$$\int_{\mathbb{R}^{\boldsymbol{d}}} \boldsymbol{K}(\boldsymbol{x}) \boldsymbol{d}\boldsymbol{x} = \boldsymbol{1}$$