第三章 单样本推断问题

第一节　符号检验和分位数推断

假设总体$F(M_e)$，M_e是总体的中位数，对于假设检验问题：

$$H_0: M_e = M_0 \leftrightarrow H_1: M_e \neq M_0$$

M_0是待检验的中位数取值

定义，$S^+ = \sum_{i=1}^{n} I(X_i > M_0)$，$S^- = \sum_{i=1}^{n} I(X_i < M_0)$，则$s^+ + s^- = n^{'}, K = \min\{s^+, s^-\}$

在零假设情况下$S^+, S^- \sim b(n^{'}, 0.5)$，在显著性水平为 α 的拒绝域为

$$\{K < c\}$$

其中 k是满足 $2P(b(n^{'}, p = 0.5) < c) \leq \alpha$ 的最大的 c值。

$$p = 2P(b(n^{'}, p = 0.5) < k)$$

例3.1. 假设某地16座预出售的楼盘均价，单位(百元/平方米)如下表所示：

36 32 31 25 28 36 40 32

41 26 35 35 32 87 33 35

楼盘均价是否为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
```

符号检验问题

$$H_0 : M_e \le M_0 \leftrightarrow H_1 : M_e > M_0 \qquad p = P_{binom}(S^- \le s^- \mid n', p = 0.5)$$

$$H_0 : M_e \ge M_0 \leftrightarrow H_1 : M_e < M_0 \qquad p = P_{binom}(S^+ \le s^+ \mid n', p = 0.5)$$

$$H_0 : M_e = M_0 \leftrightarrow H_1 : M_e \ne M_0$$

$$p = 2\min\left\{P_{binom}(S^+ \le s^+ \mid n', p = 0.5), P_{binom}(S^+ \ge s^+ \mid n', p = 0.5)\right\}$$

大样本结论

当n较大时 $S^+, S^- \sim N(\frac{n'}{2}, \frac{n'}{4})$：

$$Z = \frac{S^+ - n'/2}{\sqrt{n'/4}} \to N(0,1), n \to \infty$$

双边： $H_0 : M_e = M_0 \leftrightarrow H_1 : M_e \neq M_0$ ， p-值 $2P(N(0,1) > |z|)$

左侧： $H_0 : M_e \leq M_0 \leftrightarrow H_1 : M_e > M_0$ ， p-值 $P(N(0,1) > z)$

右侧： $H_0 : M_e \geq M_0 \leftrightarrow H_1 : M_e < M_0$ ， p-值 $P(N(0,1) < z)$

当n不够大的时候可用修正公式进行调整。

$$S^+, S^- \sim N(\frac{n'}{2}, \frac{n'}{4})$$

双边： $H_0 : M_e = M_0 \leftrightarrow H_1 : M_e \neq M_0$ ，

若$S^+ > n'/2$， $Z = \dfrac{S^+ - n'/2 - 1/2}{\sqrt{n'/4}}$ p-值 $2P(N(0,1) > z)$

若$S^+ < n'/2$， $Z = \dfrac{S^+ - n'/2 + 1/2}{\sqrt{n'/4}}$ p—值:$2P(N(0,1) < z)$

左侧：$H_0 : M_e \geq M_0 \leftrightarrow H_1 : M_e < M_0$， p-值 $P(N(0,1) < z = \dfrac{S^+ - n'/2 + 1/2}{\sqrt{n'/4}})$

右侧：$H_0 : M_e \leq M_0 \leftrightarrow H_1 : M_e > M_0$， p-值 $P(N(0,1) > z = \dfrac{S^+ - n'/2 - 1/2}{\sqrt{n'/4}})$

置信区间

根据顺序统计量构造中位数置信区间：

$$P(X_{(i)} \le M \le X_{(j)}) = \sum_{k=i}^{n} C_n^k (\frac{1}{2})^n - \sum_{k=j+1}^{n} C_n^k (\frac{1}{2})^n \quad \forall 1 \le i < j \le n$$

采用Neyman原则选择最优置信区间，首先找出置信度大于 $1-\alpha$ 的所有区间 $[X_{(i)}, X_{(j)}], i < j$，然后再从中选择区间长度最小的一个。对于大样本，可以用近似正态分布求置信区间。

例 3.3: 下表 3.3 是 16 名学生在一项体能测试上的成绩，求 Neyman 置信区间。

表 3.3. 体能测试上的成绩

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

表 3.4. 体能测试上成绩的置信区间

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

符号检验在配对样本比较的运用

配对样本$(x_1,y_1), (x_2,y_2) ,\dots (x_n,y_n)$

将 $x_i < y_i$ 记为 “+”, $x_i > y_i$ 记为 “-” , $x_i = y_i$ 记为 “0”，记P_+ 为 “+”比例， P_- 为 “-”比例，

那么假设检验问题：

$$H_0:P_+=P_- \leftrightarrow H_1:P_+ \neq P-$$

可以用符号检验。

- 例3.4 如右表是某种商品在12家超市促销活动前后的销售额对比表，用符号检验分析促销活动的效果如何？

连锁店	促销前销售额	促销后销售额	符号
1	42	40	+
2	57	60	-
3	38	38	0
4	49	47	+
5	63	65	-
6	36	39	-
7	48	49	-
8	58	50	+
9	47	47	0
10	51	52	-
11	83	72	+
12	27	33	-

p分位数的符号检验

总体分布 F 的 p 分位数 $\xi_p = F^{-1}(p)$。

不同的分位数定义：

定义：ξ_p 满足 $F(\xi_p) \geq p, F(\xi_p -) \leq p$，这种定义下 ξ_p 不唯一。

（或者：$\xi_p = F^{-1}(p) = \sup\{x : F(x) \leq p\}$，或者

$\xi_p = F^{-1}(p) = \inf\{x : F(x) \geq p\}$）

假设总体$F(M_p)$，M_p是总体的p分位数，对于假设检验问题：

$$H_0 : M_p = M_0 \leftrightarrow H_1 : M_p \neq M_0$$

M_0是待检验的p分位数取值。

$$S^+ = \sum_{i=1}^{n} I(X_i > M_0) \quad S^- = \sum_{i=1}^{n} I(X_i < M_0) \quad s^+ + s^- = n'$$

在零假设情况下 $S^+ \sim b(n', 1-p), S^- \sim b(n', p)$

在显著性水平α下，若 $P(b(n', 1-p) \le s^+) \le \alpha / 2$或者$P(b(n', 1-p) \ge s^+) \le \alpha / 2$则拒绝原假设。

例 3.5：求例 3.1 的 3/4 分位数为否为 40，此时，假设检验问题是：

$$H_0 : M_{0.75} = 40 \leftrightarrow H_1 : M_{0.75} \neq 40$$

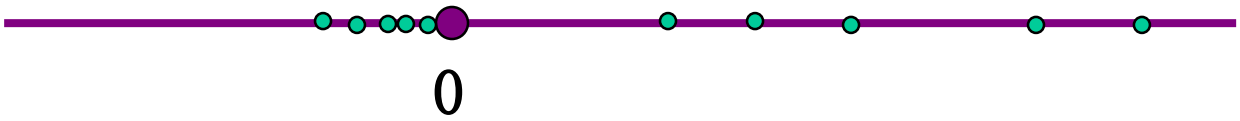
$S^+ = 2, S^- = 14$, 计算 $\mathrm{P}_{binom(16,0.75)}(\min\{S^+, S^-\} < 2) = 0$，因而拒绝零假设，认为 3/4 分位数不是 40。

Wilcoxon符号秩检验

基本概念及性质

对称分布的中心一定是中位数，在对称分布情况下，中位数不唯一，研究对称中心比中位数更有意义。

例：下面的数据中，0是对称中心吗？



定理 3.1：$X \sim F(\theta)$ 关于 θ 对称，总体的对称中心是总体的中位数之一。

证明：对于对称分布 X 有 $X-\theta$ 与 $\theta-X$ 有相同的分布：

$$\forall x, \mathrm{P}(X-\theta>x)=\mathrm{P}(\theta-X<x)$$

$P(X>\theta)=P(X<\theta)$，$P(X<\theta)+P(X>\theta)+P(X=\theta)=1$

$\Rightarrow P(X\geq\theta)=P(X\leq\theta)\geq 1/2$，即 θ 是总体的中位数。

Wilcoxon符号秩检验原理以及性质

首先设样本绝对值$|X_1|,|X_2|,\cdots,|X_n|$的顺序统计量$|X|_{(1)},|X|_{(2)},\cdots,|X|_{(n)}$如果数据关于0点对称，那么对称中心两侧的数据疏密程度应该一致，正数在取绝对值以后的样本中的秩应该和负数在绝对值样本中的秩和相近。

用R_j^+表示$|X_j|$在绝对值样本中的秩，反秩D_j由$|X_{D_j}|=|X|_{(j)}$定义。

$W_j=S(X_{D_j})=1_{(0,+\infty)}(X_{D_j})$ 表示 X_{D_j} 的符号， $R_j^+S(X_j)$ 称为符号秩统计量。

Wilcoxon符号秩统计量定义为：

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

举例

例 3.11: 如样本值为： $9, 13, -7, 10, -18, 4$ 则相应的统计量值为

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

$R=(R_1,R_2,\cdots,R_n)$ 为样本 $X_1,X_2,\cdots,X_n$ 的秩， R 在 $\Re=\{\{1,\cdots，\ n\}$的所有排列$\}$ 上均匀分布， R 取任意一组值 $(r_1,r_2,\cdots,r_n)$ 的概率都是 $1/n!$，其中 $(r_1,r_2,\cdots,r_n)$ 是 $(1,2,\cdots,n)$ 的任意一个排列。由此可见，秩统计量的分布与总体分布是没有关系的，所以秩方法是非参数方法。

由于 R 服从均匀分布，所以单个的秩 R_i $(i=1,2,\cdots,n)$ 也服从均匀分布：

$$P(R_i=r)=1/n，\ r=1,2,\cdots,n$$

定理 对任意的 $i=1,2,\cdots,n$ ， 都有

$$E(R_i)=(n+1)/2$$

$$D(R_i)=(n^2-1)/12$$

证明：对任意的$i=1,2,\cdots,n$，都有

$$E(R_i)=\sum_{r=1}^{n} r\cdot P\left(R_i=r\right)=\sum_{r=1}^{N} r\Big/n=(\mathrm{n}+1)/2$$

$$E(R_i^2)=\sum_{r=1}^{n} r^2\cdot P\left(R_i=r\right)=\sum_{r=1}^{N} r^2\Big/n=(n+1)(2n+1)/6$$

所以

$$D(R_i)=E\left(R_i^2\right)-\left(E(R_i)\right)^2=(n+1)(2n+1)/6-\left((n+1)/2\right)^2$$

$$=(n^2-1)/12$$

R_i 和 R_j $(i \neq j)$ 的联合分布也是均匀分布：

$$P(R_i = r_1, R_j = r_2) = 1/\big(n(n-1)\big), \quad r_1 \neq r_2$$

定理 对任意的 $1 \le i < j \le n$，都有

$$Cov(R_i, R_j) = -(n+1)/12$$

证明： 对任意的 $1 \le i < j \le n$，都有

$$E(R_i R_j) = \sum_{r_1 \neq r_2} r_1 r_2 P\big(R_i = r_1, R_j = r_2\big) = \sum_{r_1 \neq r_2} r_1 r_2 \Big/ \big(n(n-1)\big)$$

由于 $\sum_{r_1 \neq r_2} r_1 r_2 = \left(\sum_r r\right)^2 - \left(\sum_r r^2\right) = \big(n(n+1)/2\big)^2 - n(n+1)(2n+1)/6$

$$= n(n+1)(3n+2)(n-1)/12$$

所以 $E(R_i R_j) = (n+1)(3n+2)/12$

$$Cov(R_i, R_j) = E\big(R_i R_j\big) - E\big(R_i\big) E\big(R_j\big) = (n+1)(3n+2)/12 - \big((n+1)/2\big)^2$$

$$= -(n+1)/12$$

Wilcoxon符号秩统计量的性质

定理3.2 如果零假 $H_0:\theta=0$ 设成立，那么$S(X_1),S(X_2),\cdots,S(X_n)$ 独立于$(R_1^+,R_2^+,\cdots,R_n^+)$

证明： $X_1,\cdots,X_n$ 是样本, $(S(X_i),|X_i|),i=1,\cdots,n$ 独立同分布，

$R^+=(R_1^+,R_2^+,\cdots,R_n^+)$ 是 $(|X_1|,\cdots,|X_n|)$ 的函数，若 $S(X_i)$ 与 $|X_i|$ 独立，

则 $R^+=(R_1^+,R_2^+,\cdots,R_n^+)$， $S(X_1),\cdots,S(X_n)$ 独立。

$$P(S(X_i)=1,|X_i|\le x)=P(0<X_i\le x)=F(x)-F(0)=F(x)-0.$$

$$=\frac{F(x)-F(-x)}{2}=P(S(X_i)=1)P(|X_i|\le x)$$

记$\Re=\{\{1,\cdots,\ n\}$的所有排列$\}$, $\forall r\in\Re, d=d(r)$为r所对应的反秩，则

d$:\Re\to\Re$为一对一的。

$R^+=(R_1^+,R_2^+,\cdots,R_n^+)$在$\Re$的所有排列中均匀分布，

故反秩D=$(D_1,D_2,\cdots,D_n)$:在$\Re$的所有排列中均匀分布。

定理3.3 如果零假设$H_0:\theta=0$ 成立，那么 $S(X_1),S(X_2),\cdots,S(X_n)$ 独立于$(D_1,D_2,\cdots,D_n)$

定理3.4 如果零假设$H_0:\theta=0$ 成立，那么 $W_1,W_2,\cdots,W_n$ 独立同分布，$P(W_i=0)=P(W_i=1)=1/2$

$W^+=\sum_{j=1}^{n}R_j^+S(X_j)=\sum_{j=1}^{n}jS(X_{Dj})=\sum_{j=1}^{n}jW_j$ 与 $\sum_{j=1}^{n}jS(X_j)$ 同分布.

$P(W^+=d)=t_n(d)/2^n$,其中$t_n(d)$ 表示从 $1,\cdots,n$ 中取若干个数和为 d 的取法数。

$t_n(1)=t_n(2)=1,\ t_n(3)=2,\ t_n(4)=2,\ t_n(5)=3,\ t_n(6)=4$

性质 $E(W^+)=\sum_{j=1}^{n} jE(W_j)=\sum_{j=1}^{n} jE(S(X_j))=\frac{n(n+1)}{4}$

$$D(W^+)=\sum_{i,j}\operatorname{cov}(iW_i,jW_j)=\sum_{i=1}^{n}\tfrac{1}{4}i^2=n(n+1)(2n+1)\big/24$$

在原假设 H_0 为真时，W^+ 的分布关于 $\frac{n(n+1)}{4}$ 对称。

对称性 $S(X_j)\underline{d}\ \ 1-S(X_j)$，且 $S(X_1),S(X_2),\cdots,S(X_n)$ 与 $(R_1^+,R_2^+,\cdots,R_n^+)$ 独立，故

$$W^+=\sum_{j=1}^{n}R_j^+S(X_j)\ \ \underline{d}\ \ \sum_{j=1}^{n}R_j^+\left(1-S(X_j)\right)=n(n+1)/2-W^+,$$

所以 W^+ 的分布关于 $\frac{n(n+1)}{4}$ 对称。

Wilcoxon符号秩检验步骤：

1. 计算 $|X_i - M_0|$

2. 找出 $|X_i - M_0|$ 的秩，打结时取平均秩。

3. 令 W^+ 表示和 $X_i - M_0 > 0$ 对应的 $|X_i - M_0|$的秩和，令 W^- 表示和 $X_i - M_0 < 0$ 对应的 $|X_i - M_0|$ 的秩和。

4. 双边检验$H_0 : M = M_0 \leftrightarrow H_1 : M \neq M_0$, 取 $W = \min(W^+, W^-)$ ，当W很小时拒绝零假设；对$H_0 : M \leq M_0 \leftrightarrow H_1 : M > M_0$ ，取 $W = W^-$ ；对 $H_0 : M \geq M_0 \leftrightarrow H_1 : M < M_0$，取 $W = W^+$ 。

5. 根据W的值查Wilcoxon符号秩检验分布表。对n很大的时候，可以采用正态近似。

Wilcoxon符号秩统计量渐近分布

在小样本情况下可以计算Wilcoxon符号秩统计量的精确分布。在大样本情况下可以使用正态近似：

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

计算出Z值以后，查正态分布表对应的p-值，如果p-值很小，则拒绝零假设。

在小样本情况下，用连续性修正公式：

$$Z = \frac{W^+ - n(n+1)/4 \pm 0.5}{\sqrt{n(n+1)(2n+1)/24}} \to N(0,1)$$

Wilcoxon符号秩统计量有结分布

如果数据有g个结，在大样本情况下可以使用正态近似公式如下：

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

计算出Z值以后，查正态分布表对应的p-值，如果p-值很小，则拒绝零假设：

$$Z=\frac{W^{+}-n(n+1)/4\pm 0.5}{\sqrt{n(n+1)(2n+1)/24-\sum_{i=1}^{g}(\tau_i^3-\tau_i)/48}}\sim N(0,1)$$

假设有τ个观察值，在它们互不相等时它们的秩分别为 $r+1, r+2, \cdots, r+\tau$。由于这τ个秩的平均为

$$\frac{(r+1)+(r+2)+\cdots+(r+\tau)}{\tau}=r+(\tau+1)/2$$

所以当这τ个观察值相等，形成一个长度为τ的结时，它们的秩都是 $r+(\tau+1)/2$。由此可见，在这τ个观察值相等时它们秩的和与平方和分别为

$$\big(r+(\tau+1)/2\big)+\cdots+\big(r+(\tau+1)/2\big)=\tau\big(r+(\tau+1)/2\big)$$

$$\big(r+(\tau+1)/2\big)^2+\cdots+\big(r+(\tau+1)/2\big)^2=\tau\big(r+(\tau+1)/2\big)^2$$

而在这τ个观察值互不相等时它们秩的和与平方和分别为

$$(r+1)+(r+2)+\cdots+(r+\tau)=\tau\big(r+(\tau+1)/2\big)$$

$$(r+1)^2+(r+2)^2+\cdots+(r+\tau)^2=\tau r^2+r\tau(\tau+1)+\tau(\tau+1)(2\tau+1)/6$$

这τ个观察值相等和互不相等时，它们秩的和是相等的，它们秩的平方和是有差别的，

$$\tau\left[r+(\tau+1)/2\right]^2=(r+1)^2+(r+2)^2+\cdots+(r+\tau)^2-(\tau^3-\tau)/12$$

秩平均之后其平方和减少了$(\tau^3-\tau)/12$，其减少的量仅与观察值的个数τ，也就是结的长度有关，而与结的起点，即$r+1$无关。

在观察值相等和互不相等时，它们秩的和没有差别，为

$$\sum_{i=1}^{n}a(i)=1+2+\cdots+n=n(n+1)/2$$

观察值相等和互不相等时，它们秩的平方和有差别，所以秩平均之后其平方和为

$$\sum_{i=1}^{n}a^2(i)=1^2+2^2+\cdots+n^2-\sum_{j=1}^{g}(\tau_j^3-\tau_j)/12$$

$$=n(n+1)(2n+1)/6-\sum_{j=1}^{g}(\tau_j^3-\tau_j)/12$$

其中g为样本数据中结的个数，而τ_j是第j个结的长度，$j=1,2,\cdots,g$。

从前面的式子可以求得有结秩取平均时W^+的期望和方差。

性质. 在总体的分布关于原点 0 对称，有结秩取平均时，

$$E(W^+) = n(n+1)/4$$

$$D(W^+) = n(n+1)(2n+1)/24 - \sum_{j=1}^{g}(\tau_j^3 - \tau_j)/48$$

所以有结秩取平均后符号秩和检验统计量W^+的期望没有变化，方差有变化，应进行修正。

有结秩取平均后的符号秩和检验统计量W^+也有渐近正态性，

$$w^+ \to N\left(n(n+1)/4,\ n(n+1)(2n+1)/24 - \sum_{j=1}^{g}(\tau_j^3 - \tau_j)/48\right)$$

Wilcoxon符号秩检验导出Hodges-Lemmann估计

定义： 简单随机样本 $X_1, X_2, \cdots, X_n$ ，计算其中任意两个数的平均，称为Walsh平均，即 $\left\{ X_u^{'} : X_u^{'} = \frac{X_i + X_j}{2}, i \le j \right\}$

定理： Wilcoxon符号秩统计量 W^+ 可表示为：

$$W^+ = \sum_{i \le j} 1_{(0,+\infty)}\left(\frac{X_i + X_j}{2}\right) = \sum_{i \le j} S(X_i + X_j)$$

定义： 假设 $X_1, X_2, \cdots, X_n$ 独立同分布于 $F(X-\theta)$， 当 F 对称时，对称中心可以由Walsh平均的中位数来估计：

$$\hat{\theta} = median\{\frac{X_i + X_j}{2}, i \le j\}$$

作为 θ 的Hodges-Lemmann估计。

构造θ的区间估计

$$W^+ = \sum_{i=1}^{n} R_i^+ S(X_i)$$

$$R_i^+ S(X_i) = \sum_{j \le R_i} S(X_{(R_i)} + X_{(j)})$$

若 $X_i < 0$ ， $j \le R_i$ 时 ， $X_{(j)} + X_{(R_i)} < 0 \quad \left(X_{(j)} < X_{(R_i)} = X_i < 0\right)$ ，

$$0 = R_i^+ S(X_i) = \sum_{j \le R_i} S(X_{(R_i)} + X_{(j)})\,;$$

若 $X_i > 0$， $S(X_i) = 1$， $j \le R_i$时，且$\left|X_{(j)}\right| < X_{(R_i)}$时， $X_{(j)} + X_{(R_i)} > 0$，

$\left|X_{(j)}\right| > X_{(R_i)}$时， $X_{(j)} + X_{(R_i)} < 0$，

此时有 $R_i^+ = \sum_{j=1}^{n} I(\left|X_j\right| \le \left|X_i\right| = X_{(R_i)}) = \sum_{j \le R_i} S(X_{(R_i)} + X_{(j)})$，

$$W^+=\sum_{i=1}^{n}R_i^+S(X_i)=\sum_{i=1}^{n}\sum_{j\le R_i}S(X_{(R_i)}+X_{(j)})=\sum_{i=1}^{n}\sum_{j\le i}S(X_{(i)}+X_{(j)})$$

$$=\sum_{i=1}^{n}\sum_{j<i}S(X_{(i)}+X_{(j)})+\sum_{i=1}^{n}S(X_{(i)})$$

$$=\sum_{i=1}^{n}\sum_{j<i}S(X_i+X_j)+\sum_{i=1}^{n}S(X_i)$$

$$=\sum_{i=1}^{n}\sum_{j\le i}S(X_i+X_j)$$

构造θ的区间估计

样本$X_1,X_2,\cdots,X_n\sim F(x-\theta),F(x)$关于0对称。令$Z_i=X_i-\theta$，

$W^+=\sum_{i=1}^{n}\sum_{j\le i}S(\mathrm{Z}_i+Z_j)=\sum_{i=1}^{n}\sum_{j\le i}S(\frac{X_i+X_j}{2}-\theta)$分布已知，故

$\exists W_1,W_2$使得$P(W_1\le W^+\le W_2)\ge 1-\alpha$，即$n(n+1)/2$个$\frac{X_i+X_j}{2}$中，

$W_1\le\sum_{j\le i}I\left\{\frac{X_i+X_j}{2}>\theta\right\}\le W_2$，故$P(V_{(N-W_2)}\le\theta\le V_{(N-W_1+1)})\ge 1-\alpha$，$(V_{(N-W_2)},V_{(N-W_1+1)})$为$\theta$的置信度为$1-\alpha$的置信区间，其中$V_{(k)}$为$N=n(n+1)/2$个$\frac{X_i+X_j}{2}$的第k个顺序统计量。

正态计分检验

检验原理以及计算:

基本思想是把升幂排列的秩 R_i 用对应的正态分位$\Phi^{-1}(R_i/(n+1))$点替代，为了保证秩为正的，用变化的式子：

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

其中$S(i)$就是第 i 个数据的正态记分。

计算步骤

对假设检验问题：$H_0: M = M_0$单边或者双边。

1. 计算 $|X_i - M_0|$

2. 找出 $|X_i - M_0|$ 的秩，打结时取平均秩。

3. 用正态记分代替绝对秩：

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

$$sign(x) = 1\times 1_{(0,+\infty)}(x) + (-1)\times 1_{(-\infty,0)}(x) + 0\times 1_{\{0\}}(x)$$

记 $W = \sum_{i=1}^{n} s_i$ ，构造统计量：$T = \dfrac{W}{\sqrt{\sum_{i=1}^{n} S_i^2}}$

4. T有近似的正态分布,对右侧检验，当T大的时候，考虑拒绝零假设。

表 3.12　亚洲十国新生儿死亡率 (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 $M_0 = 16$)

$H_0 : M \geq 34 \Longleftrightarrow H_1 : M < 34$			
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
$W = 3.197,\ T^+ =$ 0.409			
p- 值 = $\Phi(T^+) = 0.659,$			
结论：不能拒绝 H_0 (水平 $\alpha < 0.659$)			

$H_0 : M \leq 16 \Longleftrightarrow H_1 : M > 16$			
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
$W = 6.384\ \ T^+ = 1.844$			
p- 值 $= 1 - \Phi(T^+) = 0.033,$			
结论：$M > 16$ (水平 $\alpha \geq 0.033$)			

§2 分布拟合检验

设总体X的实际分布函数为$F(x)$, 它是未知的.

$X_1, X_2, \cdots, X_n$ 为来自总体 X的样本.

根据这个样本来检验总体X的分布函数$F(x)$

是否等于某个给定的分布函数 $F_0(x)$, 即检验假设

$$H_0 : F(x) = F_0(x), \qquad H_1 : F(x) \neq F_0(x)$$

注意: 若总体 X 为离散型的, 则 H_0 相当于总体 X 的分布律为

$$P\{X = x_i\} = p_i, \quad i = 1,2,\cdots$$

若总体 X 为连续型的, 则 H_0 相当于总体 X 的概率密度为 $f(x)$.

（1）若 H_0 中 X 的分布函数 $F(x)$ 不含未知参数.

记 Ω 为 X 的所有可能取值的全体，将 Ω 分为k个

两两互不相交的子集 $A_1, A_2, \cdots, A_k$

以 $f_i\ (i = 1,2,\cdots,k)$ 表示样本观察值 $x_1, x_2, \cdots, x_n$

中落入 A_i 的个数,

=> 在n次试验中, 事件 A_i发生的频率为 f_i/n

另一方面, 当H_0为真时，可以根据H_0所假设的 X 的分布函数来计算 $p_i = P(A_i).$

定理1 （皮尔逊）当H_0为真且n充分大时, 统计量

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

近似服从 $\chi^2(k-1)$ 分布.

由定理1, 若给定显著性水平α, 则前述假设检验问题的拒绝域为

$$\chi^2 \geq \chi_\alpha^2(k-1)$$

选取统计量 $$\sum_{i=1}^{k} h_i \left(\frac{f_i}{n} - p_i \right)^2$$

来度量样本与H_0中所假设的分布的吻合程度，h_i是给定的常数。

一般选取$h_i = n / p_i$，则上述统计量变成

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

Pearson 定理的证明：

$k=2$ 时，

$$\chi^2=\frac{(v_1-np_1)^2}{np_1}+\frac{(v_2-np_2)^2}{np_2}=\frac{(v_1-np_1)^2}{np_1}+\frac{(v_2-np_2)^2}{np_2}$$

$$=\frac{(v_1-np_1)^2}{np_1}+\frac{(n-v_1-n(1-p_1))^2}{n(1-p_1)}$$

$$=\frac{(v_1-np_1)^2}{np_1}+\frac{(v_1-np_1)^2}{n(1-p_1)}=\frac{(v_1-np_1)^2}{np_1(1-p_1)}$$

记 $v_1=\sum_{i=1}^{n}\xi_i$ ， $\xi_i\sim b(1,p_1)$ i.i.d., $\frac{v_1-np_1}{\sqrt{np_1(1-p_1)}}\to N(0,1)$ ， 故

$\chi^2\sim\chi^2(1)$。

对一般的 k，记 $u_i = \dfrac{v_i - np_i}{\sqrt{np_i}}$，若 H_0 为真，则

$v = (v_1, v_2, \cdots, v_k) \sim multinomial(n; p_1, p_2, \cdots, p_k)$ 多项分布，

$$P(v_1 = n_1, \cdots, v_k = n_k) = \frac{n!}{n_1! \cdots n_k!} p_1^{n_1} \cdots p_k^{n_k}, \sum_i n_i = n$$，特征函数

$$\phi_{(v_1, v_2, \cdots, v_k)}(t_1, t_2, \cdots, t_k) = E(e^{it'v}) = (p_1 e^{it_1} + p_2 e^{it_2} + \cdots + p_k e^{it_k})^n$$ （练习）

$$\phi_{(u_1, u_2, \cdots, u_k)}(t_1, t_2, \cdots, t_k) = E(e^{it'u}) = E(\exp\{i \sum_{j=1,\cdots,k} t_j u_j\})$$

$$= E(\exp\{i \sum_{j=1,\cdots,k} t_j \frac{v_j - np_j}{\sqrt{np_j}}\})$$

$$= E(\exp\{i \sum_{j=1,\cdots,k} \frac{t_j}{\sqrt{np_j}} v_j\}) \exp\{-i\sqrt{n} \sum_j t_j \sqrt{p_j}\}$$

$$= \exp\{-i\sqrt{n} \sum_j t_j \sqrt{p_j}\} (p_1 e^{it_1/\sqrt{np_1}} + p_2 e^{it_2/\sqrt{np_2}} + \cdots + p_k e^{it_k/\sqrt{np_k}})^n$$

$$\ln \phi_{(u_1,u_2,\cdots,u_k)}(t_1,t_2,\cdots,t_k)$$

$$= -i\sqrt{n}\sum_{j=1,\cdots,k} t_j\sqrt{p_j} + n\ln(\sum_{j=1}^{k} p_j e^{it_j/\sqrt{np_j}})$$

$$= -i\sqrt{n}\sum_{j} t_j\sqrt{p_j} + n\ln[\sum_{j=1}^{k} p_j(e^{it_j/\sqrt{np_j}}-1)+1]$$

$$= -i\sqrt{n}\sum_{j} t_j\sqrt{p_j} + n[\sum_{j=1}^{k} p_j(e^{it_j/\sqrt{np_j}}-1)] - \frac{n}{2}[\sum_{j=1}^{k} p_j(e^{it_j/\sqrt{np_j}}-1)]^2 + o(1)$$

$$= -i\sqrt{n}\sum_{j} t_j\sqrt{p_j} + n\left[\sum_{j=1}^{k} p_j\left(it_j/\sqrt{np_j} + \frac{1}{2}(it_j/\sqrt{np_j})^2\right)\right]$$

$$-\frac{n}{2}\left[\sum_{j=1}^{k} p_j\left(it_j/\sqrt{np_j} + \frac{1}{2}(it_j/\sqrt{np_j})^2\right)\right]^2 + o(1)$$

$$= -i\sqrt{n}\sum_j t_j\sqrt{p_j} + n\left[\sum_{j=1}^{k} p_j\left(it_j/\sqrt{np_j} - \frac{t_j^{\,2}}{2np_j}\right)\right]$$

$$-\frac{n}{2}\left[\sum_{j=1}^{k} p_j\left(it_j/\sqrt{np_j} - \frac{t_j^{\,2}}{2np_j}\right)\right]^2 + o(1)$$

$$= -\frac{1}{2}\sum_{j=1}^{k} t_j^{\,2} - \frac{n}{2}\left[\frac{1}{\sqrt{n}} i\sum_{j=1}^{k}\sqrt{p_j}t_j - \frac{1}{2n}\sum_{j=1}^{k} t_j^{\,2}\right]^2 + o(1)$$

$$= -\frac{1}{2}\sum_{j=1}^{k} t_j^{\,2} - \frac{1}{2}\left[\sum_{j=1}^{k}\sqrt{p_j}t_j\right]^2 + o(1)$$

$$= -\frac{1}{2}(t'\left(I - (\sqrt{p_1},\cdots,\sqrt{p_k})'(\sqrt{p_1},\cdots,\sqrt{p_k})\right)t + o(1)$$

$$u=(u_1,\cdots,u_k)'\sim N\left(0,I-(\sqrt{p_1},\cdots,\sqrt{p_k})'(\sqrt{p_1},\cdots,\sqrt{p_k})\right)$$

$I-(\sqrt{p_1},\cdots,\sqrt{p_k})'(\sqrt{p_1},\cdots,\sqrt{p_k})$ 为投影矩阵，故

$$\chi^2=u'u\sim\chi^2\left(r(I-(\sqrt{p_1},\cdots,\sqrt{p_k})'(\sqrt{p_1},\cdots,\sqrt{p_k}))\right)$$(练习),

$$r(I-(\sqrt{p_1},\cdots,\sqrt{p_k})'(\sqrt{p_1},\cdots,\sqrt{p_k}))=k-\sum_{j=1}^{k}p_j=k-1$$

（2）若H_0中X 的的分布函数含有未知参数.

此时, 首先在假设下利用样本求出未知参数的最大似然估计, 以估计值作为参数值, 然后再根据 H_0中所假设的 X 的分布函数 $F(x)$求出 p_i的估计值

$$\hat{p}_i = \hat{P}(A_i)$$

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

中以 $\hat{p}_i$ 代替 p_i , 得到统计量

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

定理2（皮尔逊）当 H_0 为真且 n 充分大时, 统计量

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

近似服从 $\chi^2(k-r-1)$ 分布, 其中r是 X 的分布函数 $F(x)$包含的未知参数的个数.

若给定显著性水平α,则前述假设检验问题的拒绝域为

$$\chi^2 \geq \chi_\alpha^2(k-r-1)$$

注意： 运用 χ^2 检验法检验总体分布, 把样本数据进行分类时,

（1）大样本, 通常取 $n \geq 50$

（2）要求各组的理论频数 $np_i \geq 5$ 或 $n\hat{p}_i \geq 5$

（3）一般数据分成7到14组. 有时为了保证各组

$$np_i \geq 5$$

组数可以少于7组

例1　孟德尔在著名的豌豆杂交实验中，用结黄色圆形种子与结绿色皱形种子的纯种豌豆作为亲本进行杂交，将子一代进行自交得到子二代共556株豌豆，发现其中有四种类型植株

Y^-R^-	Y^-rr	yyR^-	$yyrr$	总计
（黄圆）	（黄皱）	（绿圆）	（绿皱）	
315株	101株	108株	32株	556株

试问这些植株是否符合孟德尔所提出的 9:3:3:1 的理论比例 $(\alpha = 0.05)$

解 检验假设

H_0：这些植株符合 $9:3:3:1$ 的理论比例.

H_1：这些植株不符合 $9:3:3:1$ 的理论比例.

由 $9:3:3:1$ 的理论比例可知

$$p_1=\frac{9}{16},\quad p_2=\frac{3}{16},\quad p_3=\frac{3}{16},\quad p_4=\frac{1}{16}$$

由n=556, 得

$$np_1=312.75,\quad np_2=104.25$$

$$np_3=104.25,\quad np_4=34.75$$

而 $f_1 = 315, \quad f_2 = 101, \quad f_3 = 108, \quad f_4 = 32, \quad k = 4,$

计算得 $$\chi^2 = \sum_{i=1}^{k} \frac{(f_i - np_i)^2}{np_i} = 0.47.$$

由 α=0.05 , 自由度 $k-1=4-1=3,$ 查 χ^2 分布表得

$$\chi^2_{0.05}(3) = 7.815 \quad => \chi^2 < \chi^2_{0.05}(3)$$

=>在α=0.05下接受 H_0

=>这些植株是符合孟德尔所提出的 9:3:3:1

的理论比例

例2 某农科站为了考察某种大麦穗长的分布情况，在一块实验地里随机抽取了100个麦穗测量其长度，得到数据如下（单位: cm)

6.5	6.4	6.7	5.8	5.9	5.9	5.2	4.0	5.4	4.6
5.8	5.5	6.0	6.5	5.1	6.5	5.3	5.9	5.5	5.8
6.2	5.4	5.0	5.0	6.8	6.0	5.0	5.7	6.0	5.5
6.8	6.0	6.3	5.5	5.0	6.3	5.2	6.0	7.0	6.4
6.4	5.8	5.9	5.7	6.8	6.6	6.0	6.4	5.7	7.4
6.0	5.4	6.5	6.0	6.8	5.8	6.3	6.0	6.3	5.6
5.3	6.4	5.7	6.7	6.2	5.6	6.0	6.7	6.7	6.0
5.5	6.2	6.1	5.3	6.2	6.8	6.6	4.7	5.7	5.7
5.8	5.3	7.0	6.0	6.0	5.9	5.4	6.0	5.2	6.0
5.3	5.7	6.8	6.1	4.5	5.6	6.3	6.0	5.8	6.3

试检验大麦穗长是否服从正态分布?(α=0.05)

解 检验假设

H_0：X的概率密度为 $f(x)=\dfrac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

μ,σ^2是未知的，所以应首先估计μ,σ^2

μ,σ^2 的最大似然估计为

$$\hat{\mu}=\bar{x}=5.921,\qquad \hat{\sigma}^2=\frac{n-1}{n}s^2=0.6034^2$$

把X可能取值的全体 $\Omega=[3.95,\quad 7.55]$ 划分为

k=12个互不重叠的小区间:

$A_1=[3.95,\quad 4.25]$, $A_2=(4.25,\quad 4.55]$, $\cdots$

$A_{12}=(7.25,\quad 7.55]$

=>大麦穗长的频数、频率分布表

A_i	频数 f_i	频率 f_i / n	累计频率
3.95~4.25	1	0.01	
4.25~4.55	1	0.01	
4.55~4.85	2	0.02	0.09
4.85~5.15	5	0.05	
5.15~5.45	11	0.11	0.20
5.45~5.75	15	0.15	0.35
5.75~6.05	28	0.28	0.63
6.05~6.35	13	0.13	0.76
6.35~6.65	11	0.11	0.87
6.65~6.95	10	0.10	
6.95~7.25	2	0.02	1.00
7.25~7.55	1	0.01	
合计	100	1.00	

由 $\hat{\mu} = \bar{x} = 5.921, \quad \hat{\sigma}^2 = \frac{n-1}{n} s^2 = 0.6034^2$

$\Rightarrow X \sim N(5.921, \quad 0.6034^2)$

由此可计算 $\hat{p}_i = \hat{P}(A_i)$, 若 $A_i = (t_{i-1}, \quad t_i]$, 则

$$\hat{p}_i = \Phi\left(\frac{t_i - \hat{\mu}}{\hat{\sigma}}\right) - \Phi\left(\frac{t_{i-1} - \hat{\mu}}{\hat{\sigma}}\right)$$

$$= \Phi\left(\frac{t_i - 5.921}{0.6034}\right) - \Phi\left(\frac{t_{i-1} - 5.921}{0.6034}\right)$$

$\hat{p}_i, \chi^2$ 的值见下表

χ^2 的计算表

组号	分组	频数 f_i	$\hat{p}_i$	$n\hat{p}_i$	$(f_i - n\hat{p}_i)^2 / n\hat{p}_i$
1	3.95~5.15	9	0.09976	9.976	0.09549
2	5.15~5.45	11	0.1174	11.74	0.04664
3	5.45~5.75	15	0.172	17.2	0.2814
4	5.75~6.05	28	0.1935	19.35	3.8668
5	6.05~6.35	13	0.1779	17.79	1.28972
6	6.35~6.65	11	0.1258	12.58	0.19844
7	6.65~7.55	13	0.10963	10.963	0.37849
合计		100	0.99599	99.599	6.15698

由 k=7, r=2, 得自由度 k-r-1=4, 查表得

$$\chi^2_{0.05}(4) = 9.488$$

而

$$\chi^2 = 6.15698 < 9.488$$

=>接受原假设, 即在检验水平α=0.05下，下可认为大麦的穗长服从正态分布

$$X \sim N(5.921, \quad 0.6034^2)$$

例:

例 3.15: 调查某美发店上半年各月顾客数量如表所示:

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为均匀分布？

解: 假设检验问题:

H_0 : 各月顾客数符合均匀分布$1:1$ (即各月顾客比例$p_i = p_0 = \frac{1}{6}, \forall i = 1, \cdots, 6$)

H_1 : 各月顾客数不符合$1:1$ (即即各月顾客比例$p \neq p_0 = \frac{1}{6}, \exists i = 1, \cdots, 6$)

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

上述 $E_i = np_i = 150 \times \frac{1}{6} = 25, i = 1, \cdots, 6.$

由 (3.7) 式得：

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

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 $p = \frac{1}{6}$ 。

例 3.16: 调查某农作物根部蚜虫的分布情况，调查结果如下表所示，问蚜虫在某农作物根部分布是否为泊松分布 (Poisson distribution) 。

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：

$$H_0 \quad : \quad \text{蚜虫在农作物根部的分布是泊松分布}$$
$$H_1 \quad : \quad \text{蚜虫在农作物根部的分布不为泊松分布}$$

若蚜虫在农作物根部的分布为泊松分布，则分布列为：

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

其中 λ 是泊松分布的期望，是未知的，需要用观测值估计，其估值如下

$$\hat{\lambda} = \bar{x} = (0 \times 10 + 1 \times 24 + \cdots + 5 \times 1)/50 = 1.3$$

因而

$$\hat{p_0} = \frac{e^{-1.3}(1.3)^0}{0!} = 0.2725$$

$$\hat{p}_1 = \frac{e^{-1.3}(1.3)^1}{1!} = 0.3543$$

$$\hat{p}_2 = \frac{e^{-1.3}(1.3)^2}{2!} = 0.2303$$

$$\hat{p}_3 = \frac{e^{-1.3}(1.3)^3}{3!} = 0.0998$$

$$\hat{p}_4 = \frac{e^{-1.3}(1.3)^4}{4!} = 0.0324$$

$$\hat{p}_5 = \frac{e^{-1.3}(1.3)^5}{5!} = 0.0107$$

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

经验分布函数的性质

定理 1: 令 $X_1, \cdots, X_n \sim F$, 经验分布函数 $\hat{F}_n$ 有以下性质:

1. 任意固定点 x,

$$\mathrm{E}(\hat{F}_n(x)) = F(x) \quad \mathrm{Var}(\hat{F}_n(x)) = \frac{F(x)(1-F(x))}{n}.$$

于是, $MSE(\hat{F}_n(x)) = \frac{F(x)(1-F(x))}{n} \to 0$, 因此 $\hat{F}_n(x) \to^{p} F(x)$.

2. (Glivenko-Cantelli) $\sup |\hat{F}_n(x) - F(x)| \to^{a.s.} 0$.

3. (Dvoretzky-Kiefer-Wolfowitz(DKW)inequality) 对任意的 $\epsilon > 0$,

$$P\left(\sup |\hat{F}_n(x) - F(x)| > \epsilon\right) \leq 2e^{-2n\epsilon^2}.$$

Kolmogorov-Smirnov检验

Kolmogorov-Smirnov检验将样本经验分布和理论分布的比较，检验样本是否来自于该理论分布。假设检验问题：

H_0 :样本来自所给分布$\mathbf{F_0(x)}$

H_1 :样本不是来自该分布

假设样本的经验分布函数为 $F_n(x)$

因此,当H0为真时,$F_n(x)$和$F_0(x)$的偏差应该很小
引入Kolmogorov统计量

$$D_n = \sup_{-\infty<x<\infty} |F_n(x) - F_0(x)|$$

D_n 的计算

$F_0(x)$是单调非降函数，$\hat{F}(x)$是单调非降的阶梯函数，$|F_n(x)-F_0(x)|$的上确界可在n个$X_{(i)}$处找。

$\hat{F}_n(X_{(i)})=\frac{i}{n}, \hat{F}_n(X_{(i)}-)=\frac{i-1}{n}$，$\hat{F}_n(x)$右连续，故

$$D_n=\sup_{-\infty<x<\infty}|F_n(x)-F_0(x)|$$

$$=\max_{1\le i\le n}\left\{\left|F_0(X_{(i)})-\frac{i-1}{n}\right|\vee\left|F_0(X_{(i)})-\frac{i}{n}\right|\right\}$$

由格列汶科定理 $P\{\lim\limits_{n\to\infty}\sup\limits_{-\infty<x<\infty}|F_n(x)-F(x)|=0\}=1$

定理3.4.1.(Kolmogorov)

设总体X的分布函数F(x)连续,$X_1,\cdots,X_n$是取自总体X的样本,则当H_0为真时,有

$$\lim_{n\to\infty}P\{D_n<\frac{\lambda}{\sqrt{n}}\}=k(\lambda)$$

其中$k(\lambda)=\begin{cases}0, & \lambda\le 0\\ \sum\limits_{k=-\infty}^{\infty}(-1)^k\exp\{-2k^2\lambda^2\}, & \lambda>0\end{cases}$

当$D>D_\alpha$时，拒绝原假设。

注:1)渐近分布函数k（λ）与$F_0(x)$无关

2)不适用于离散总体

例题

检验是否可以认为下列10个数是来自正态分布$N(0,1)$的随机数:

0.4855, -0.0050, -0.2762, 1.2765, 1.8634, -0.5226, 0.1034, -0.8076, 0.6804, -2.3646

i	$x_{(i)}$	$F_0(x_{(i)})$	$(i-1)/n$	i/n	δ_i
1	-2.3646	0.0090	0	0.1	0.0910
2	-0.8076	0.2096	0.1	0.2	0.1096
3	-0.5226	0.3006	0.2	0.3	0.1006
4	-0.2762	0.3912	0.3	0.4	0.0912
5	-0.0050	0.4980	0.4	0.5	0.0980
6	0.1034	0.5412	0.5	0.6	0.0588
7	0.4855	0.6863	0.6	0.7	0.0863
8	0.6804	0.7519	0.7	0.8	0.0519
9	1.2765	0.8991	0.8	0.9	0.0991
10	1.8634	0.9688	0.9	1.0	0.0688

$D_n = 0.1096$ $D_{10,0.10} = 0.36866$

$k(1.23) \approx 0.9, k(1.36) \approx 0.95, k(1.63) \approx 0.99$

表 16. Kolmogorov 检验临界值 $P(D_n \geqslant d_\alpha) = \alpha$

n \ α	0.20	0.10	0.05	0.02	0.01
1	0.90000	0.95000	0.97500	0.99000	0.99500
2	0.68377	0.77639	0.84189	0.90000	0.92929
3	0.56481	0.63604	0.70760	0.78456	0.82900
4	0.49265	0.56522	0.62394	0.68887	0.73424
5	0.44698	0.50945	0.56328	0.62713	0.66853
6	0.41037	0.46799	0.51926	0.57741	0.61661
7	0.38148	0.43607	0.48342	0.53844	0.57581
8	0.35831	0.40962	0.45427	0.50654	0.54179
9	0.33910	0.38746	0.43001	0.47960	0.51332
10	0.32260	0.36866	0.40925	0.45662	0.48893
11	0.30829	0.35242	0.39122	0.43670	0.46770
12	0.29577	0.32815	0.37543	0.41918	0.44905
13	0.28470	0.32549	0.36143	0.40362	0.43247
14	0.27481	0.31417	0.34890	0.38970	0.41763
15	0.26588	0.30397	0.33760	0.37713	0.40420
16	0.25778	0.29472	0.32733	0.36571	0.39201
17	0.25039	0.28627	0.31796	0.35528	0.38086
18	0.24360	0.27851	0.30963	0.34569	0.37062
19	0.23735	0.27136	0.30143	0.33685	0.36117
20	0.23156	0.26473	0.29408	0.32866	0.35241

例 3.18:35 位健康男性在未进食前的血糖浓度如下，试测验这组数据是否来自均值 $\mu = 80$，标准差为 $\sigma = 6$ 的正态分布。

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					

```
ks.test(healthy,pnorm,80,6)

        One-sample Kolmogorov-Smirnov test

data:  healthy
D = 0.1481, p-value = 0.4264
alternative hypothesis: two-sided
```

练习 1：

证明 Pearson 定理 $k=2$ 时，

$$\chi^2=\frac{(v_1-np_1)^2}{np_1}+\frac{(v_2-np_2)^2}{np_2}\sim\chi^2(1)$$

记 $v_1=\sum_{i=1}^{n}\xi_i$ ， $\xi_i\sim b(1,p_1)$ i.i.d., 证明

$$\frac{v_1-np_1}{\sqrt{np_1(1-p_1)}}\to N(0,1)$$ ，即证明 Laplace 中心极限定理。

练习 2

证明多项分布 $v=(v_1,v_2,\cdots,v_k)\sim multinomial(n;p_1,p_2,\cdots,p_k)$，

$$P(v_1=n_1,\cdots,v_k=n_k)=\frac{n!}{n_1!\cdots n_k!}p_1^{n_1}\cdots p_k^{n_k},\sum_i n_i=n$$

的特征函数为

$$\phi_{(v_1,v_2,\cdots,v_k)}(t_1,t_2,\cdots,t_k)=E(e^{it'v})=(p_1e^{it_1}+p_2e^{it_2}+\cdots+p_ke^{it_k})^n$$

练习 3

$$u=(u_1,\cdots,u_k)'\sim N\left(0,I-(\sqrt{p_1},\cdots,\sqrt{p_k})'(\sqrt{p_1},\cdots,\sqrt{p_k})\right)$$

证明： $I-(\sqrt{p_1},\cdots,\sqrt{p_k})'(\sqrt{p_1},\cdots,\sqrt{p_k})$ 为投影矩阵，

$$\chi^2=u'u\sim\chi^2\left(r(I-(\sqrt{p_1},\cdots,\sqrt{p_k})'(\sqrt{p_1},\cdots,\sqrt{p_k}))\right),$$

$$r(I-(\sqrt{p_1},\cdots,\sqrt{p_k})'(\sqrt{p_1},\cdots,\sqrt{p_k}))=k-\sum_{j=1}^{k}p_j=k-1$$

练习4（选做） 当 H_0 为真且 n 充分大时, 统计量

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

近似服从 $\chi^2(k-r-1)$ 分布, 其中r是 X 的分布函数 $F(x)$包含的未知参数的个数.

练习 5：

盒子里放有白球和黑球，现做如下试验：用返回抽样方式从盒子中摸球，直到摸取的球是白球为止，记录下抽取的次数。如此重复试验 200 次，其结果如下：

摸球次数	1	2	3	4	≥ 5
相应的频数	85	60	30	14	11

试检验盒子中白球和黑球个数是否相等 $(\alpha = 0.05)$。

作业：

王静龙《非参数统计分析》书上习题三的第 2、5 题，习题四的 1、2、3、4 题