时所犯错误的概率至多为 α；当样本值未落入 W 时只表明现有的数据与零假设 H_0 是相容的，即未出现矛盾.这并不意味着 H_0 一定成立.此时接受 H_0 可能犯第二类错误，而犯第二类错误的概率究竟多大，并未明确显示出来(它与样本量 n 及其他因素有关)，这是与第一类错误的概率至多为已知的 α 大不相同的.在实际的研究工作中总希望得到某种明确的结论.一般总是根据已往的知识和现有的数据猜想"某个结论"可能成立.这种情形下，常把"某个结论"列为"备择假设"，而将"某个结论的否定"列为"零假设".此时，若采用检验水平为 α 的否定域，"零假设"被拒绝，则我们有理由说："某个结论"成立.(因为这时出错的概率至多为 α).

2. 临界值和 p 值

前面说过，给出一个检验法就是要给出一个否定域 W.否定域 W 通常是由一个直观上有明确意义的统计量 $\varphi(X_1, X_2, \cdots, X_n)$ 来确定.确定的方式可概括为两种.

第一种(单边情形)：

$$W = \{(x_1, x_2, \cdots, x_n): \varphi(x_1, x_2, \cdots, x_n) > \lambda\} \tag{3.1}$$

第二种(双边情形)：

$$W = \{(x_1, x_2, \cdots, x_n): \varphi(x_1, x_2, \cdots, x_n) < \lambda_1$$

或 $\varphi(x_1, x_2, \cdots, x_n) > \lambda_2\}$ $(\lambda_1 < \lambda_2)$, (3.2)

这里 λ 叫做单边情形的临界值，λ_1 和 λ_2 叫做双边情形的临界值.临界值是根据检验水平 α 来确定.通常，对于单边情形，应找 λ 满足

$$\sup_{\theta \in \Theta_0} P_\theta(\varphi(X_1, X_2, \cdots, X_n) > \lambda) = \alpha \tag{3.3}$$

这里"$\theta \in \Theta_0$."是零假设 H_0，$P_\theta(A)$ 表示参数的真值是 θ 时事件 A 的概率(下同).对于这个 λ，(3.3)式表明否定域(3.1)的精确检验水平恰好是 α.

当满足(3.3)的 λ 不存在时，应选 λ 满足

$$\sup_{\theta \in \Theta_0} P_\theta[\varphi(X_1, X_2, \cdots, X_n) > \lambda] \leqslant \alpha < \sup_{\theta \in \Theta_0} P_\theta[\varphi(X_1, X_2, \cdots, X_n) \geqslant \lambda], \tag{3.4}$$

此时，否定域的检验水平不超过 α.

对于双边情形，应选取 $\lambda_1 < \lambda_2$ 满足

$$\sup_{\theta \in \Theta_0} P_\theta[\varphi(X_1, X_2, \cdots, X_n) < \lambda_1] = \frac{\alpha}{2} \tag{3.5}$$

$$\sup_{\theta\in\Theta_0} P_\theta[\varphi(X_1,X_2,\cdots,X_n)>\lambda_2]=\frac{\alpha}{2} \tag{3.6}$$

对这样的 λ_1,λ_2，否定域(3.2)的检验水平不超过 α，当这样的 λ_1 不存在时，应选 λ_1 满足

$$\sup_{\theta\in\Theta_0} P_\theta[\varphi(X_1,X_2,\cdots,X_n)<\lambda_1]\leqslant\frac{\alpha}{2}<\sup_{\theta\in\Theta_0} P_\theta[\varphi(X_1,X_2,\cdots,X_n)\leqslant\lambda_1] \tag{3.7}$$

当满足(3.6)的 λ_2 不存在时，应选 λ_2 满足

$$\sup_{\theta\in\Theta_0} P_\theta[\varphi(X_1,X_2,\cdots,X_n)>\lambda_2]\leqslant\frac{\alpha}{2}<\sup_{\theta\in\Theta_0} P_\theta[\varphi(X_1,X_2,\cdots,X_n)\geqslant\lambda_2] \tag{3.8}$$

此时，否定域(3.2)的检验水平不超过 α.

上述根据检验水平 α 确定临界值从而获得否定域的方法，简称临界值方法.这是本书用来确定否定域的基本方法.读者容易看出，§2 中关于正态总体的假设检验的否定域都是用这个方法确定的.

例 3.1 设 $X\sim N(\mu,\sigma^2)$.未知 σ，检验假设 $H_0:\mu\leqslant\mu_0$（μ_0 是已知数）.这里 $\Theta=\{(\mu,\sigma^2):\mu$ 任意，$\sigma^2>0\}$.零假设 H_0 可表示为：$\theta\in\Theta_0$.这里 $\Theta_0=\{(\mu,\sigma^2):\mu\leqslant\mu_0,\sigma^2>0\}$，备择假设是 $H_a:\theta\in\Theta_1$，这里 $\Theta_1=\Theta-\Theta_0$.设样本是 $X_1,X_2,\cdots,X_n(n\geqslant2)$.取统计量

$$\varphi(X_1,X_2,\cdots,X_n)=\frac{\bar X-\mu_0}{\sqrt{S^2/n}},$$

这里 $\bar X=\frac{1}{n}\sum_{i=1}^{n}X_i,S^2=\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar X)^2$.从直观上看，$\varphi$ 值越大，对假设 H_0 越不利，故应取单边情形的否定域

$$\begin{aligned}W&=\{(x_1,x_2,\cdots,x_n):\varphi(x_1,x_2,\cdots,x_n)>\lambda\}\\&=\left\{(x_1,x_2,\cdots,x_n):\frac{\bar x-\mu_0}{\sqrt{S^2/n}}>\lambda\right\}\end{aligned}$$

这里 $\bar x=\frac{1}{n}\sum_{i=1}^{n}x_i,S^2=\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar x)^2$.对给定的检验水平 α，应取 λ 满足(3.3)注意　$\mu\leqslant\mu_0$ 时

$$\frac{\bar X-\mu_0}{\sqrt{S^2/n}}\leqslant\frac{\bar X-\mu}{\sqrt{S^2/n}}$$

于是

$$\sup_{\theta\in\Theta_0} P_\theta(\varphi>\lambda)=P_{\mu\sigma^2}\left(\frac{\bar{X}-\mu}{\sqrt{S^2/n}}>\lambda\right)$$

但是 $t_{n-1}=\dfrac{X-\mu}{\sqrt{S^2/n}}$ 服从 $n-1$ 个自由度的 t 分布(对一切 μ,σ),故从 t 分布临界值表可找到 λ 满足 $P(t_{n-1}>\lambda)=\alpha$(这个 λ 就是分位数 $t_{1-\alpha}(n-1)$),并且满足(3.3),我们得到否定域

$$W=\left\{(x_1,x_2,\cdots,x_n):\frac{\bar{x}-\mu_0}{\sqrt{S^2/n}}>t_{1-\alpha}(n-1)\right\}.$$

我们指出,还有一种确定否定域的方法——p 值方法.该方法可提供人们更多的信息.

我们先研究单边情形的否定域(3.1),不去考虑其中的 λ 如何由检验水平 α 来确定,而去考虑一个新的函数.设 $x_1,x_2,\cdots,x_n$ 是样本值(已知的 n 个常数).令 $p(x_1,x_2,\cdots,x_n)=\sup\limits_{\theta\in\Theta_0}P_\theta[\varphi(X_1,X_2,\cdots,X_n)\geqslant\varphi(x_1,x_2,\cdots,x_n)]$ (3.9)

这里"$\theta\in\Theta_0$"就是零假设 H_0.$p(x_1,x_2,\cdots,x_n)$是 H_0 成立的条件下统计量 φ 取值不小于 $\varphi(x_1,x_2,\cdots,x_n)$的最大概率.

定义 3.2 $p(x_1,x_2,\cdots,x_n)$(由(3.9)定义)叫做单边情形下样本值 $(x_1,x_2,\cdots,x_n)$的 p 值.

p 值有什么用呢?先看下列特性.

引理 3.1 设对给定的 $\alpha\in(0,1)$,恰有一个 λ 满足

$$\sup_{\theta\in\Theta_0}P_\theta[\varphi(X_1,X_2,\cdots,X_n)>\lambda]=\alpha \tag{3.10}$$

则 $\varphi(x_1,x_2,\cdots,x_n)>\lambda$ 的充要条件是 $p(x_1,x_2,\cdots,x_n)<\alpha$.

这个引理告诉我们,在条件(3.10)下,样本值$(x_1,x_2,\cdots,x_n)$落入检验水平为 α 的否定域的充要条件是样本值的 p 值小于 α,换句话说,当且仅当样本值的 p 值小于 α 时拒绝 H_0.这时犯第一类错误的概率的最大值是 α.这个结论告诉我们,有了样本值 $x_1,x_2,\cdots,x_n$ 后,根据统计量 φ 可计算出相应的 p 值 $p(x_1,x_2,\cdots,x_n)$,然后与检验水平 α 进行比较.当 p 值小于 α 时拒绝 H_0,当 p 值不小于 α 时不拒绝 H_0.即否定域为$\{(x_1,x_2,\cdots,x_n):p(x_1,x_2,\cdots,x_n)<\alpha\}$.这种确定否定域的方法简称为 p 值方法.p 值方法的优点

在于:不预先给定检验水平 α,从计算出的 p 值就可以知道,对一切大于这个 p 值的 α,拒绝 H_0 而引起的错误其概率不超过 α.

p 值 $p(x_1,x_2,\cdots,x_n)$也可看作是样本$(x_1,x_2,\cdots,x_n)$与零假设 H_0 相容程度的度量. p 值越大,相容程度越高;反之,p 值越小,则相容程度越低. p 值小到一定程度则认为二者不相容了,即应拒绝 H_0. 当 p 值小于 α 时认为二者不相容,这时拒绝 H_0 而引起的错误其概率不超过 α.

引理 3.1 的证明 设 $p(x_1,x_2,\cdots,x_n)<\alpha$. 则 $\sup\limits_{\theta\in\Theta_0}P_\theta[\varphi(X_1,X_2,\cdots,X_n)\geqslant\varphi(x_1,x_2,\cdots,x_n)]<\alpha$,与(3.3)式对比知 $\varphi(x_1,x_2,\cdots,x_n)>\lambda$. 反过来,设 $\varphi(x_1,x_2,\cdots,x_n)>\lambda$,则有 $\varepsilon>0$ 满足 $\varphi(x_1,x_2,\cdots,x_n)-\varepsilon>\lambda$. 于是 $p(x_1,x_2,\cdots,x_n)=\sup\limits_{\theta\in\Theta_0}P_\theta[\varphi(X_1,X_2,\cdots,X_n)\geqslant\varphi(x_1,x_2,\cdots,x_n)]\leqslant\sup\limits_{\theta\in\Theta_0}P_\theta[\varphi(X_1,X_2,\cdots,X_n)>\varphi(x_1,x_2,\cdots,x_n)-\varepsilon]<\sup\limits_{\theta\in\Theta_0}P_\theta[\varphi(X_1,X_2,\cdots,X_n)>\lambda]=\alpha$. 证毕.

给定 $\alpha\in(0,1)$,不一定有 λ 满足(3.10),即不一定存在精确检验水平恰为 α 的检验(当检验用的统计量 $\varphi(X_1,X_2,\cdots,X_n)$是离散型随机变量时常出现这种情况). 此时应考虑检验水平不超过 α 的检验. 我们可推广引理 3.1.

引理 3.2 设对给定的 $\alpha\in(0,1)$,有 λ 满足

$$\sup_{\theta\in\Theta_0}P_\theta[\varphi(X_1,X_2,\cdots,X_n)>\lambda]\leqslant\alpha<\sup_{\theta\in\Theta_0}P_\theta[\varphi(X_1,X_2,\cdots,X_n)\geqslant\lambda],\tag{3.11}$$

则 $\varphi(x_1,x_2,\cdots,x_n)>\lambda$ 的充要条件是 $p(x_1,x_2,\cdots,x_n)\leqslant\alpha$

证 很容易,设 $\varphi(x_1,x_2,\cdots,x_n)>\lambda$,则 $P_\theta[\varphi(X_1,X_2,\cdots,X_n)\geqslant\varphi(x_1,x_2,\cdots,x_n)]\leqslant P_\theta[\varphi(X_1,X_2,\cdots,X_n)>\lambda]$,于是 $p(x_1,x_2,\cdots,x_n)=\sup\limits_{\theta\in\Theta_0}P_\theta[\varphi(X_1,X_2,\cdots,X_n)\geqslant\varphi(x_1,x_2,\cdots,x_n)]\leqslant\sup\limits_{\theta\in\Theta_0}P_\theta[\varphi(X_1,X_2,\cdots,X_n)>\lambda]\leqslant\alpha$.

反之,若 $\varphi(x_1,x_2,\cdots,x_n)\leqslant\lambda$,则 $P_\theta[\varphi(X_1,X_2,\cdots,X_n)\geqslant\varphi(x_1,x_2,\cdots,x_n)]\geqslant P_\theta[\varphi(X_1,X_2,\cdots,X_n)\geqslant\lambda]$,利用(3.11)知 $p(x_1,x_2,\cdots,x_n)>\alpha$. 引理 3.2 证完.

从引理 3.2 知,若 λ 满足(3.11),则否定域 $W=\{(x_1,x_2,\cdots,x_n):\varphi(x_1,x_2,\cdots,x_n)>\lambda\}$的检验水平不超过 α,而且样本值$(x_1,x_2,\cdots,x_n)$落

入 W 的充要条件是该样本值的 p 值不超过 α. 因此,在检验假设 H_0 时,我们根据样本值计算其 p 值,当且仅当 p 值不超过 α 时拒绝 H_0. 这种利用 p 值确定否定域的方法仍叫做 p 值方法.

例 3.2 设 $X\sim N(\mu,\sigma^2)$,未知 σ,检验假设 $H_0:\mu\leqslant\mu_0$,例 2.6 已研究过此检验问题,其否定域是

$$W=\{(x_1,x_2,\cdots,x_n):\varphi(x_1,x_2,\cdots,x_n)>\lambda_0\},$$

其中 $\varphi(x_1,x_2,\cdots,x_n)=(\bar{x}-\mu_0)/\sqrt{S^2/n}$, $\lambda_0=t_{1-\alpha}(n-1)$ $\left(\bar{x}=\dfrac{1}{n}\sum_{i=1}^{n}x_i, S^2=\dfrac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2\right)$.

根据样本值 $x_1,x_2,\cdots,x_n$,我们可直接计算 p 值

$$p(x_1,x_2,\cdots,x_n)=\sup_{\theta\in\Theta_0}P_\theta[\varphi(X_1,X_2,\cdots,X_n)\geqslant\varphi(x_1,x_2,\cdots,x_n)],$$

这里 $\theta=(\mu,\sigma^2)$, $\Theta_0=\{(\mu,\sigma^2):\mu\leqslant\mu_0,\sigma^2>0\}$

易知 $p(x_1,x_2,\cdots,x_n)=P_{\mu_0\sigma^2}[\dfrac{\bar{X}-\mu_0}{\sqrt{S^2/n}}\geqslant\varphi(x_1,x_2,\cdots,x_n)]=P[T\geqslant\varphi(x_1,x_2,\cdots,x_n)]$,这里 T 是服从 $n-1$ 个自由度的 t 分布的随机变量. 因此 $p(x_1,x_2,\cdots,x_n)$可以算出. 从引理 2.1 知道,当且仅当这个 p 值小于 α 时拒绝 H_0.

例如,为了检验假设 $H_0:\mu\leqslant25$,我们有样本值 $x_1,x_2,\cdots,x_{64}$,若由此计算出 $\bar{x}=25.9$, $S^2=17.3$,则 $\varphi(x_1,x_2,\cdots,x_{64})=(\bar{x}-25)/\sqrt{S^2/64}=1.731$. 我们可计算 p 值 $p(x_1,x_2,\cdots,x_{64})=P(T\geqslant1.731)=0.042<0.05$,故对于检验水平 $\alpha=0.05$ 应拒绝 H_0.

现在我们来研究双边情形的否定域(3.2),不去考虑其中的 λ_1 和 λ_2 如何由检验水平 α 来确定,而去考虑一个新的函数. 虽然不去确定 λ_1 和 λ_2 的具体数值,但从统计量 $\varphi(X_1,X_2,\cdots,X_n)$的直观意义及(3.5)、(3.6). 我们可找到 λ_0 满足:$\lambda_1\leqslant\lambda_0<\lambda_2$. 设 $x_1,x_2,\cdots,x_n$ 是样本值,当 $\varphi(x_1,x_2,\cdots,x_n)\leqslant\lambda_0$ 时,令

$$p(x_1,x_2,\cdots,x_n)=\min\{2\sup_{\theta\in\Theta_0}P_\theta[\varphi(X_1,X_2,\cdots,X_n)\leqslant\varphi(x_1,x_2,\cdots,x_n)],1\} \tag{3.12}$$

当 $\varphi(x_1,x_2,\cdots,x_n)>\lambda_0$ 时,令

$$p(x_1,x_2,\cdots,x_n)=\min\{2\sup_{\theta\in\Theta_0}P_\theta[\varphi(X_1,X_2,\cdots,X_n)\geqslant\varphi(x_1,x_2,\cdots,x_n)],1\} \tag{3.13}$$

这里"$\theta\in\Theta_0$"就是零假设 H_0(下同.)

定义 3.3 由(3.12)和(3.13)定义的 $p(x_1,x_2,\cdots,x_n)$叫做双边情形下样本值$(x_1,x_2,\cdots,x_n)$的 p 值.

p 值的重要意义见下列引理.

引理 3.3 设对给定的 $\alpha\in(0,1)$,有惟一的 λ_1 和惟一的 λ_2 满足

$$\sup_{\theta\in\Theta_0}P_\theta[\varphi(X_1,X_2,\cdots,X_n)<\lambda_1]=\frac{\alpha}{2} \tag{3.14}$$

$$\sup_{\theta\in\Theta_0}P_\theta[\varphi(X_1,X_2,\cdots,X_n)>\lambda_2]=\frac{\alpha}{2} \tag{3.15}$$

则"$\varphi(x_1,x_2,\cdots,x_n)<\lambda_1$ 或 $\varphi(x_1,x_2,\cdots,x_n)>\lambda_2$"成立的充要条件是 $p(x_1,x_2,\cdots,x_n)<\alpha$.

引理 3.4 设对给定的 α,有 λ_1 和 λ_2 满足

$$\sup_{\theta\in\Theta_0}P_\theta[\varphi(X_1,X_2,\cdots,X_n)<\lambda_1]\leqslant\frac{\alpha}{2}<\sup_{\theta\in\Theta_0}P_\theta[\varphi(X_1,X_2,\cdots,X_n)\leqslant\lambda_1] \tag{3.16}$$

$$\sup_{\theta\in\Theta_0}P_\theta[\varphi(X_1,X_2,\cdots,X_n)>\lambda_2]\leqslant\frac{\alpha}{2}<\sup_{\theta\in\Theta_0}P_\theta[\varphi(X_1,X_2,\cdots,X_n)\geqslant\lambda_2], \tag{3.17}$$

则"$\varphi(x_1,x_2,\cdots,x_n)<\lambda_1$ 或 $\varphi(x_1,x_2,\cdots,x_n)>\lambda_2$"成立的充要条件是 $p(x_1,x_2,\cdots,x_n)\leqslant\alpha$.

引理 3.3 的证明 设 $\varphi(x_1,x_2,\cdots,x_n)<\lambda_1$,则 $\varphi(x_1,x_2,\cdots,x_n)\leqslant\lambda_0$ (这里 $\lambda_0\in[\lambda_1,\lambda_2)$),于是从(3.12)知 $p(x_1,x_2,\cdots,x_n)\leqslant2\sup_{\theta\in\Theta_0}P_\theta[\varphi(X_1,X_2,\cdots,X_n)\leqslant\varphi(x_1,x_2,\cdots,x_n)]\leqslant2\sup_{\theta\in\Theta_0}P_\theta[\varphi(X_1,X_2,\cdots,X_n)<\lambda_1-\varepsilon]$ (对某个正数 ε)$<\alpha$.

若 $\varphi(x_1,x_2,\cdots,x_n)>\lambda_2$,则有 $\varepsilon>0$ 使得 $\varphi(x_1,x_2,\cdots,x_n)>\lambda_2+\varepsilon$,于是 $p(x_1,x_2,\cdots,x_n)\leqslant2\sup_{\theta\in\Theta_0}P_\theta[\varphi(X_1,X_2,\cdots,X_n)\geqslant\varphi(x_1,x_2,\cdots,x_n)]\leqslant2\sup_{\theta\in\Theta_0}P_\theta[\varphi(X_1,X_2,\cdots,X_n)>\lambda_2+\varepsilon]<\alpha$(据(2.17)).

故只要 $\varphi(x_1,x_2,\cdots,x_n)$小于 λ_1 或大于 λ_2 则一定有 $p(x_1,x_2,\cdots,x_n)$

$<\alpha$.

另一方面，设 $p(x_1,x_2,\cdots,x_n)<\alpha$，若 $\varphi(x_1,x_2,\cdots,x_n)\leqslant\lambda_0$，则 $2\sup\limits_{\theta\in\Theta_0}P_\theta[\varphi(X_1,X_2,\cdots,X_n)\leqslant\varphi(x_1,x_2,\cdots,x_n)]<\alpha$. 从(3.14)知 $\varphi(x_1,x_2,\cdots,x_n)<\lambda_1$. 若 $\varphi(x_1,x_2,\cdots,x_n)>\lambda_0$ 则 $2\sup\limits_{\theta\in\Theta_0}P_\theta[\varphi(X_1,X_2,\cdots,X_n)\geqslant\varphi(x_1,x_2,\cdots,x_n)]<\alpha$. 利用(3.15)知 $\varphi(x_1,x_2,\cdots,x_n)>\lambda_2$. 总之，只要 $p(x_1,x_2,\cdots,x_n)<\alpha$ 就一定有 $\varphi(x_1,x_2,\cdots,x_n)<\lambda_1$ 或 $\varphi(x_1,x_2,\cdots,x_n)>\lambda_2$. 引理 3.3 证完.

引理 3.4 的证明方法是类似的，从略.

引理 3.3 和引理 3.4 告诉我们，为了检验假设 $H_0:\theta\in\Theta_0$，否定域 $W=\{(x_1,x_2,\cdots,x_n):\varphi(x_1,x_2,\cdots,x_n)<\lambda_1$ 或 $\varphi(x_1,x_2,\cdots,x_n)>\lambda_2\}$ 的检验水平不超过 α(有时精确检验水平恰好是 α)，而且在引理 3.3 的条件下，样本值 $(x_1,x_2,\cdots,x_n)$ 落入这个否定域的充要条件是样本值的 p 值小于 α. 在引理 3.4 的条件下，样本值 $(x_1,x_2,\cdots,x_n)$ 落入否定域的充要条件是样本值的 p 值不超过 α. 这种用 p 值来确定否定域的方法仍叫做 p 值方法.

双边情形下 p 值方法的优点与单边情形下 p 值方法的优点是一样的：不必预先给定检验水平，计算出样本值的 p 值 $p(x_1,x_2,\cdots,x_n)$ 后，与任给的 α 进行比较就知道何时应拒绝零假设 H_0，而且拒绝 H_0 产生的错误其概率不超过 α.

和单边情形一样，双边情形的 p 值也可看作是样本值 $(x_1,x_2,\cdots,x_n)$ 与零假设 H_0 相容程度的度量，p 值越小表明二者的相容程度越低，p 值小到一定程度就认为二者不相容了.

p 值方法有很大优点，但也有麻烦之处：要根据样本值计算出相应的 p 值. 有时这种计算还比较复杂. 好在一些常见的假设检验问题(见本章各节)里，p 值的计算程序已在流行的统计软件包(如 SAS)中给出. 使用这些软件，很容易算出 p 值.

例 3.3 设 $X\sim N(\mu,\sigma^2)$，μ,σ^2 均未知，为了检验假设 $H_0:\sigma^2=\sigma_0^2$(σ_0 是已知数)，前面已说过应使用统计量 $\varphi(X_1,X_2,\cdots,X_n)=\frac{1}{\sigma_0^2}\sum\limits_{i=1}^{n}(X_i-\bar{X})^2$，当 φ 值太小或太大时应拒绝 H_0，故应采用双边情形的否定域 $W=\{(x_1,x_2,\cdots,x_n):\varphi(x_1,x_2,\cdots,x_n)<\lambda_1$ 或 $\varphi(x_1,x_2,\cdots,x_n)>\lambda_2\}$. 如何计

算样本值$(x_1,x_2,\cdots,x_n)$的 p 值?

从直观上看,$S^2=\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2$ 可作为 σ^2 的估计值(这里 $\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i$).可见,如果 H_0 成立,则$\frac{1}{n-1}\varphi(x_1,x_2,\cdots,x_n)$应和 1 相差不大.即 $\varphi(x_1,x_2,\cdots,x_n)$应和 $n-1$ 相差不大.因而在否定域 W 中,λ_1 应小于 $n-1$,λ_2 应大于 $n-1$.取 $\lambda_0=n-1$.则 $\lambda_1<\lambda_0<\lambda_2$,注意 $\varphi(x_1,x_2,\cdots,x_n)\leqslant\lambda_0$ 的充要条件是 $S^2\leqslant\sigma_0^2$,这里 $S^2=\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2$,于是从定义 3.3 知 $S^2\leqslant\sigma_0^2$ 时

$$p(x_1,x_2,\cdots,x_n)=\min\{2\sup_{\substack{\mu\text{任意}\\ \sigma^2\leqslant\sigma_0^2}}P_{\mu,\sigma^2}[\varphi(X_1,X_2,\cdots,X_n)\leqslant\varphi_0],1\},$$

这里 $\varphi_0=\varphi(x_1,x_2,\cdots,x_n)$.

易知 $p(x_1,x_2,\cdots,x_n)=\min\{2P(\xi\leqslant\varphi_0),1\}$,这里随机变量 ξ、服从 $n-1$ 个自由度的 χ^2 分布.

类似地,当 $S^2\geqslant\sigma_0^2$ 时

$$p(x_1,x_2,\cdots,x_n)=\min\{2P(\xi)\geqslant\varphi_0),1\}.$$

根据例 3.4 中提供的 10 个数据,知 $S^2=75.7>64=\sigma_0^2$,$\varphi_0=\frac{1}{\sigma_0^2}\sum_{i=1}^{10}(x_i-\bar{x})^2=10.65$,$P(\xi\geqslant\varphi_0)=0.30$.(因为 ξ 服从 9 个自由度的 χ^2 分布),于是 p 值 $=0.6$.可见对一切 $\alpha\leqslant0.6$ 在检验水平 α 下都不应拒绝假设 $H_0:\sigma^2=\sigma_0^2$.

3. 假设检验与置信区间的联系

我们指出假设检验的接受域与置信区间有一种简单而深刻的联系.设 X 的分布函数是 $F(x,\theta)$,θ 是未知参数,$\theta\in\Theta$.$(X_1,X_2,\cdots,X_n)$是 X 的样本.对任何 $\theta_0\in\Theta$,考虑零假设 $H_0:\theta=\theta_0$,备择假设 $H_a:\theta\neq\theta_0$,设 $A(\theta_0)$是 H_0 的检验水平为 α 的接受域(即 $A(\theta_0)$的补集是检验水平为 α 的否定域),即当且仅当$(X_1,X_2,\cdots,X_n)$的值属于 $A(\theta_0)$时接受假设 H_0,且

$$P_{\theta_0}((X_1,X_2,\cdots,X_n)\overline{\in}A(\theta_0))\leqslant\alpha$$

令

$$S(x_1,x_2,\cdots,x_n)=\{\theta:(x_1,x_2,\cdots,x_n)\in A(\theta)\}\qquad(3.18)$$

则对一切 θ 有 $P_\theta[\theta\in S(X_1,X_2,\cdots,X_n)]\geqslant1-\alpha$.

由此可见，如果集合 $S(x_1, x_2, \cdots, x_n)$ 是个区间，则它就是 θ 的置信水平为 $1-\alpha$ 的置信区间. 这就是利用假设检验的接受域构造置信区间——寻找置信区间的第三个方法.

例 3.4 设 $X \sim N(\theta, 1)$，$\theta \in (-\infty, +\infty)$. $(X_1, X_2, \cdots, X_n)$ 是 X 的样本. 对任何 θ_0，考虑检验零假设 $H_0: \theta = \theta_0$（备择假设是 $H_a: \theta \neq \theta_0$），从 §1 中的讨论知，可采用接受域 $A(\theta_0) = \{(x_1, x_2, \cdots, x_n): |x - \theta_0| \leqslant c\}$. 易知 $c = \frac{1}{\sqrt{n}} z_{1-\frac{\alpha}{2}}$（$z_{1-\frac{1}{2}\alpha}$ 是 $N(0,1)$ 的 $1-\frac{1}{2}\alpha$ 分位数）时检验水平为 α. 从 (3.18) 知 $S(x_1, x_2, \cdots, x_n) = [\bar{x} - c, \bar{x} + c]$. 故 $[\bar{X} - c, \bar{X} + c]$ 是 θ 的 $1-\alpha$ 水平置信区间. $\left(\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i\right)$.

§4 两个正态总体的假设检验

§2 中讲了一个总体的检验问题，在实际工作中还常碰到两个总体的比较问题，§1 中例 1.3 就是这方面的典型.

设 $X \sim N(\mu_1, \sigma_1^2)$，$Y \sim N(\mu_2, \sigma_2^2)$，且 X，Y 相互独立. 根据实际问题的需要，我们主要是讲三类问题：

① 未知 σ_1^2, σ_2^2，但知道 $\sigma_1^2 = \sigma_2^2$，检验假设 $H_0: \mu_1 = \mu_2$.

② 未知 μ_1, μ_2，检验假设 $H_0: \sigma_1^2 = \sigma_2^2$.

③ 未知 μ_1, μ_2，检验假设 $H_0: \sigma_1^2 \leqslant \sigma_2^2$.

④ 未知 σ_1^2, σ_2^2 但知道 $\sigma_1^2 \neq \sigma_2^2$，检验假设 $H_0: \mu_1 = \mu_2$.

1. 未知 σ_1^2, σ_2^2，但知道 $\sigma_1^2 = \sigma_2^2$，检验假设 $H_0: \mu_1 = \mu_2$.

从分析具体问题开始，研究例 1.3. 为阅读方便，再把例 1.3 复述一遍.

例 4.1 （即本章例 1.3）在漂白工艺中要考察温度对针织品断裂强力的影响. 在 70 ℃ 与 80 ℃ 下分别重复作了八次试验，测得断裂强力的数据如下（单位：千克力）：

70 ℃：20.5, 18.8, 19.8, 20.9, 21.5, 19.5, 21.0, 21.2

80 ℃:17.7,20.3,20.0,18.8,19.0,20.1,20.2,19.1

究竟 70 ℃下的强力与 80 ℃下的强力有没有差别?

用 X,Y 分别表示 70 ℃与 80 ℃下的断裂强力,它们自然是独立的.根据过去的经验,可以认为 X,Y 都是服从正态分布,方差是相等的.即 $X\sim N(\mu_1,\sigma_1^2)$,$Y\sim N(\mu_2,\sigma_2^2)$,$\sigma_1^2=\sigma_2^2$(也可用下面 2 中的办法检验方差是否相等).我们的问题是检验假设 H_0:$E(X)=E(Y)$(即 $\mu_1=\mu_2$).

读者很自然想到,比较两组数据的平均数,看哪个大.经过计算知,70 ℃时的平均强力是 20.4 千克力,80 ℃时的平均强力是 19.4 千克力,相差 1 千克力,70 ℃的大.

但是能否由此就简单地下结论:"70 ℃就是比 80 ℃的强力大"呢? 还不能.这是因为产生这 1 千克力差别的原因可能有二:一个是 μ_1 与 μ_2 的差异,另一个是试验误差的影响.即使在 $\mu_1=\mu_2$ 的情形,由于试验误差的存在,也完全有可能会产生这 1 千克力的差别,特别当试验误差比较大时,更是这样.但是有了以上的分析,我们也就有了解决问题的直观想法:估出试验误差的影响大小,并将这 1 千克力跟它作某种意义上的比较.如果单单是试验误差的影响还不足以引起这 1 千克力的差异,就否定"$\mu_1=\mu_2$",否则就不否定"$\mu_1=\mu_2$".

我们现在来介绍一般的理论,然后再用到这个具体例子上去.设 $x_1,x_2,\cdots,x_n$ 是来自 X 的样本值;$y_1,y_2,\cdots,y_n$ 是来自 Y 的样本值;$X\sim N(\mu_1,\sigma_1^2)$,$Y\sim N(\mu_2,\sigma_2^2)$,$\sigma_1^2=\sigma_2^2$,要检验的假设 H_0 是 $\mu_1=\mu_2$.

很自然的想法是,研究样本平均值之差:

$$\overline{x}-\overline{y}$$

如果这个差数的绝对值很大,则不大可能 $\mu_1=\mu_2$,反之,若差数比较小,则很可能 $\mu_1=\mu_2$.当然这里的"大"与"小"是相对试验误差而言的.

和以前一样，我们先假设 $\mu_1 = \mu_2$，看产生什么后果，会不会产生不合理的现象. 不过在假设 $H_0: \mu_1 = \mu_2$ 成立的条件下，随机变量 $\bar{X} - \bar{Y}$ 的概率分布仍然算不出来，因为它的方差等于 $\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{n}$，而 σ_1^2, σ_2^2 不知道. 我们自然想到用

$$s_1^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

代替 σ_1^2，用

$$s_2^2 = \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y})^2$$

代替 σ_2^2.

经过数学研究. 可以证明随机变量

$$\tilde{T} = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2 + S_2^2}{n}}} \tag{4.1}$$

服从 $2n-2$ 个自由度的 t 分布.

于是，在假设 H_0 下，统计量

$$T = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2 + \sum_{i=1}^{n}(Y_i - \bar{Y})^2}{n(n-1)}}} \tag{4.2}$$

服从 $2n-2$ 个自由度的 t 分布. 查 t 分布临界值表得到 λ 满足

$$P\{|T| > \lambda\} = \alpha$$

现在好了，根据所给的样本值 $x_1, x_2, \cdots, x_n, y_1, y_2, \cdots, y_n$，具体算出统计量 T 的值，如果得到的 $|T|$ 的值大于 λ，则否定原来的假设 $H_0: \mu_1 = \mu_2$，反之，如果得到的 $|T|$ 的值不超过 λ，则假设 H_0 是相容的.

总之，检验的步骤与以前一样，就是要记住统计量 T 的形式，另外，查 t 分布临界值表时，自由度是 $2n-2$.

现在把上述一般理论用到例 4.1 上去.

第一步:提出待检验的假设 $H_0: E(X) = E(Y)$.

第二步:计算统计量(4.2)的值,这是最费劲的一步.现在 $n = 8, \bar{x} = 20.4, \bar{y} = 19.4$.①

$$\sum_{i=1}^{8} (x_i - \bar{x})^2 = (20.5 - 20.4)^2 + (18.8 - 20.4)^2 + \cdots + (21.2 - 20.4)^2 = 6.20$$

$$\sum_{i=1}^{8} (y_i - \bar{y})^2 = (17.7 - 19.4)^2 + (20.3 - 19.4)^2 + \cdots + (19.1 - 19.4)^2 = 5.80$$

于是

$$T = \frac{20.4 - 19.4}{\sqrt{\frac{6.20 + 5.80}{8 \times 7}}} = \frac{1}{\sqrt{\frac{12}{56}}}$$

$$= \sqrt{4.67} = 2.161$$

第三步:查 t 分布表,自由度是 $2n - 2 = 14$.

取 $\alpha = 0.05$,得临界值 $\lambda = 2.145$.

第四步:下结论.

现在 $T = 2.161 > 2.145$,故应否定假设 $H_0: E(X) = E(Y)$.换句话说,$E(X) \neq E(Y)$,即 70℃ 下的强力比 80 ℃ 下的强力显著地大.

以上介绍的是两个总体的样本容量相等的情形,但实际工作中有时样本容量并不相等.此时也可用类似的办法进行处理.

设 $X_1, X_2, \cdots, X_{n_1}$ 是来自 $N(\mu_1, \sigma^2)$ 的样本,$Y_1, Y_2, \cdots, Y_{n_2}$ 是来自总体 $N(\mu_2, \sigma^2)$ 的样本.数学上可以证明

---

① 这里,主要目的是为了讲假设检验,因而,在计算上,采用了 $\bar{x}$ 和 $\bar{y}$ 是整齐数字的数据.实际上,很难遇见这种情形.当 $\bar{x}, \bar{y}$ 不整齐时,通常多取一至二位有效数字.然后利用简化法来计算 $\sum (x_i - \bar{x})^2$ 和 $\sum (y_i - \bar{y})^2$.

$$\tilde{T}=\frac{(\bar{X}-\bar{Y})-(\mu_1-\mu_2)}{\sqrt{\sum_{i=1}^{n_1}(X_i-\bar{X})^2+\sum_{i=1}^{n_2}(Y_i-\bar{Y})^2}}\cdot\sqrt{\frac{n_1n_2(n_1+n_2-2)}{n_1+n_2}}$$

服从 n_1+n_2-2 个自由度的 t 分布. 可见,在假设 $H_0:\mu_1=\mu_2$ 下,统计量

$$T=\frac{\bar{X}-\bar{Y}}{\sqrt{\sum_{i=1}^{n_1}(X_i-\bar{X})^2+\sum_{i=1}^{n_2}(Y_i-\bar{Y})^2}}\cdot\sqrt{\frac{n_1n_2(n_1+n_2-2)}{n_1+n_2}} \tag{4.3}$$

服从 n_1+n_2-2 个自由度的 t 分布. 利用这一点,同样可对 H_0 进行检验.

上面的检验工作通常又叫平均数的显著性鉴定. 如果否定了假设 $H_0:\mu_1=\mu_2$,通常也说两个总体的平均数有"显著性差异".

例 4.2 研究口服避孕药对妇女血压的影响. 对某公司工作的 35 岁至 39 岁的非怀孕妇女,用抽查方法收集到下列数据. 有 8 人使用口服避孕药,8 人的血压(收缩压)的平均值是 132.86(单位:mmHg),标准差是 15.35;有 21 人未使用口服避孕药,血压的平均值是 127.44(单位:mmHg),标准差是 18.23,问:这两种血压平均值的差异是否"显著"?

解 我们假定使用口服避孕药的妇女的血压(收缩压)服从正态分布 $N(\mu_1,\sigma_1^2)$,不使用口服避孕药的妇女的血压(收缩压)服从正态分布 $N(\mu_2,\sigma_2^2)$,假定 $\sigma_1^2=\sigma_2^2$. 问题转化为检验假设 $H_0:\mu_1=\mu_2$.

使用统计量(4.3),现在 $\bar{X}=132.86$, $\bar{Y}=127.44$, $\sum_{i=1}^{8}(X_i-\bar{X})^2$

$=7\times(15.35)^2$，$\sum_{i=1}^{21}(Y_i-\bar{Y})^2=2.0\times(18.23)^2$（$n_1=8, n_2=21$），可算出统计量 $T=0.74$，设检验水平 $\alpha=0.05$，查 t 分布的临界值表知临界值 $\lambda=2.052$，现在 $|T|=0.74<2.052$，故认为假设 H_0 是相容的，即两个平均值无显著差异.

成对数据的比较

有些实际问题里的数据是天然成对的.这时为了检验平均数有无“显著性差异”，应该采用下面例 3.1 中用到的办法.由于相应的数学模型比较复杂，我们不进行理论上的严密论述，只希望读者了解统计方法，并注意，它不要求同方差的假定.

例 4.3 为了鉴定两种工艺方法对产品某性能指标有无显著性差异，对于九批材料用两种工艺进行生产，得到该指标的九对数据如下：

0.20 0.30 0.40 0.50 0.60 0.70 0.80 0.90 1.00

0.10 0.21 0.52 0.32 0.78 0.59 0.68 0.77 0.89

现在问：根据上述数据，能否说两种不同工艺对产品的该性能指标有显著性差异？（检验水平 $\alpha=0.05$.）

解 考查 9 对数据的差：

0.10 0.09 −0.12 0.18 −0.18 0.11 0.12 0.13 0.11

如果两种工艺对产品的该指标没有显著性差异，那么，这 9 个数可以看成是来自一个均值是 0 的总体 Z.我们假定 Z 服从正态分布（在许多实际工作中常这样做）.若 $z_1, z_2, \cdots, z_9$ 是来自 $N(0,\sigma^2)$ 的样本，则统计量

$$T=\frac{\bar{z}}{\sqrt{S^2/9}}$$

（这里 $S^2=\frac{1}{9-1}\sum_{i=1}^{9}(z_i-\bar{z})^2$）服从 8 个自由度的 t 分布.查附表 2 知

$$P\{|T|>2.306\}=0.05$$

这表明，如果两种工艺方法对产品的该指标没有显著性差异，则 $\{|T|>2.306\}$ 是一个小概率事件.当发生了小概率事件时，可以否定假设，即认为两种不同工艺方法对产品的该指标有显著性差异.现在，将 9 对数据的差代入 T 的表达式中（即在 T 中令 $z_1=0.10, z_2=0.09, \cdots, z_9=0.11$），因为 $\bar{z}=0.06, S^2=0.015$，计算得 $T=1.5$，既然 $|T|\leqslant 2.306$，没发生小概率事件，故未发现两种工艺方法对产品的该指标有显著性差异.

2. 未知 μ_1, μ_2，检验假设 $H_0: \sigma_1^2 = \sigma_2^2$.

就上面讨论的例 4.1 来说，我们认为两个总体的方差是相等的. 严格追问起来，有什么根据呢？除非已有大量经验可以预先作出判断，否则还是要根据所给的样本值，来检验 $H_0: \sigma_1^2 = \sigma_2^2$ 是否真的成立.

我们现在来研究一般性问题. 设 $X_1, X_2, \cdots, X_{n_1}$ 来自总体 $N(\mu_1, \sigma_1^2)$，$Y_1, Y_2, \cdots, Y_{n_2}$ 来自总体 $N(\mu_2, \sigma_2^2)$，且 X, Y 间相互独立. 如何检验假设 $H_0: \sigma_1^2 = \sigma_2^2$？

还是老办法. 先假设 H_0 成立(即 $\sigma_1^2 = \sigma_2^2$)，看有什么结果. 要比较 σ_1^2 与 σ_2^2，自然想到用它们的估计量来比比看. 令

$$S_1^2 = \frac{1}{n_1 - 1} \sum_{i=1}^{n_1} (X_i - \bar{X})^2$$

$$S_2^2 = \frac{1}{n_2 - 1} \sum_{i=1}^{n_2} (Y_i - \bar{Y})^2$$

取统计量 $F = \dfrac{S_1^2}{S_2^2}$，显然当 F 很大或很小时，就不能认为假设 H_0 成立.

所以，关键问题就是研究这个统计量的概率分布. 经过数学方面的研究，可以证明(见附录二定理 9)，随机变量

$$\widetilde{F} = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2}$$

的分布密度 $f(u)$ 是这样的：

$$f(u) = \begin{cases} \dfrac{\Gamma\left(\dfrac{n_1 + n_2 - 2}{2}\right)}{\Gamma\left(\dfrac{n_1 - 1}{2}\right)\Gamma\left(\dfrac{n_2 - 1}{2}\right)} \left(\dfrac{n_1 - 1}{n_2 - 1}\right)^{\frac{n_1 - 1}{2}} u^{\frac{n_1 - 1}{2} - 1} \cdot \\ \left(1 + \dfrac{n_1 - 1}{n_2 - 1} u\right)^{-\frac{n_1 + n_2 - 2}{2}} & u > 0 \\ 0 & u \leqslant 0 \end{cases}$$

这里介绍一个名词.

定义 4.1 如果随机变量 Z 的分布密度是

$$f_{n_1 n_2}(u)=\begin{cases}\dfrac{\Gamma\left(\dfrac{n_1+n_2}{2}\right)}{\Gamma\left(\dfrac{n_1}{2}\right)\Gamma\left(\dfrac{n_2}{2}\right)}\left(\dfrac{n_1}{n_2}\right)^{\frac{n_1}{2}}u^{\frac{n_1}{2}-1}\cdot \\ \left(1+\dfrac{n_1}{n_2}u\right)^{-\frac{n_1+n_2}{2}} & u>0 \\ 0 & u\leqslant 0\end{cases}$$

这里 n_1, n_2 是两个正整数,则称 Z 服从自由度为 n_1, n_2 的 F 分布,这里 n_1 叫第一自由度,n_2 叫第二自由度."自由度"名称的来源,大家可以不管,反正是密度函数的两个参数.

容易看出,在 H_0 成立的条件下,我们的统计量 F 恰好是服从自由度为 n_1-1, n_2-1 的 F 分布.

于是,对任给的"检验标准"α(通常 $\alpha=0.05$),可找到 λ_1, λ_2(查 F 分布的临界值表)满足:

$$P(F<\lambda_1)=\frac{\alpha}{2} \tag{4.4}$$

$$P(F>\lambda_2)=\frac{\alpha}{2} \tag{4.5}$$

这样的事件$\{F<\lambda_1\}$,$\{F>\lambda_2\}$都是"小概率事件".

根据所给的样本值 $x_1, x_2, \cdots, x_{n_1}; y_1, y_2, \cdots, y_{n_2}$,可算出 F 的数值,如果算出来的值小于 λ_1 或大于 λ_2,则应否定假设 $H_0: \sigma_1^2=\sigma_2^2$;反之,如果算出来的 F 的值界于 λ_1 与 λ_2 之间($\lambda_1\leqslant F\leqslant\lambda_2$),则假设"$\sigma_1^2=\sigma_2^2$"是相容的.

关于查表,有一点要补充说明的.找(4.5)中的 λ_2,可直接从表上读出,但找(4.4)中的 λ_1,却要绕一个弯子.由于

$$P\{F<\lambda_1\}=P\left\{\frac{1}{F}>\frac{1}{\lambda_1}\right\}$$

而$\frac{1}{F}$是服从自由度n_2-1, n_1-1的F分布，于是通过查表，可得λ_0满足：$P\left\{\frac{1}{F}>\lambda_0\right\}=\frac{\alpha}{2}$，这样，$\lambda_1=\frac{1}{\lambda_0}$也就得到了.

现在把上述一般理论用到例4.1的数据上去，看是否真的可认为70 ℃时的强力与80 ℃时的强力有相同的方差.

这里，$n_1=n_2=8, s_1^2=\frac{6.20}{7}, s_2^2=\frac{5.80}{7}$，于是

$$F=\frac{s_1^2}{s_2^2}=\frac{6.20}{5.80}=1.07$$

查F分布表，取$\alpha=0.05$，于是$\frac{\alpha}{2}=0.025$，查附表5，现在第一自由度是7，第二自由度也是7.得：

$$\lambda_2=4.99$$

$$\lambda_1=\frac{1}{\lambda_2}=\frac{1}{4.99}$$

现在，$\frac{1}{4.99}<1.07<4.99$.故可以认为70 ℃与80 ℃下的强力有相同的方差.

例4.4 在例4.2中我们假定了使用口服避孕药的妇女的血压的方差σ_1^2与不使用口服避孕药情形下的方差σ_2^2相等.若对这点有怀疑，应检验假设$H_0:\sigma_1^2=\sigma_2^2$.使用统计量$F=\frac{S_1^2}{S_2^2}$.现在$n_1=8, n_2=21$，故第一自由度是7，第二自由度是20，设$\alpha=0.05$.于是$\frac{\alpha}{2}=0.025$，设$\lambda_1, \lambda_2$分别满足(4.4)、(4.5)，查附表5知$\lambda_2=3.01, \lambda_1<\frac{1}{4.42}=0.226$，$H_0$的否定域是$F<\lambda_1$或$F>\lambda_2$.现在$S_1^2=(15.35)^2, S_2^2=(18.23)^2$(据例4.2中的数据)，计算出$F$的值是0.709，这个数在$\lambda_1$与$\lambda_2$之间，故不应拒绝$\sigma_1^2=\sigma_2^2$的假设.

3. 未知 μ_1,μ_2，检验假设 $H_0:\sigma_1^2\leqslant\sigma_2^2$.

这类问题在技术革新等实际工作中常遇到.

例 4.5 有两台车床生产同一种型号的滚珠.根据已有经验可以认为，这两台车床生产的滚珠的直径都服从正态分布.问题就是要比较两台车床所生产的滚珠的直径的方差.现在从这两台车床的产品中分别抽出 8 个和 9 个，测得滚珠的直径如下(单位是毫米)：

甲车床：15.0,14.5,15.2,15.5,14.8,15.1,15.2,14.8

乙车床：15.2,15.0,14.8,15.2,15.0,15.0,14.8,15.1,14.8

问：乙车床产品的方差是否比甲车床的小？

用 X,Y 分别表示甲、乙两车床的产品的直径. $X\sim N(\mu_1,\sigma_1^2)$, $Y\sim N(\mu_2,\sigma_2^2)$, X,Y 独立，我们来检验 $H_0:\sigma_1^2\leqslant\sigma_2^2$.

显然，如果 H_0 受到否定，那就是说，乙车床产品的方差比甲车床的小.反之，如果 H_0 相容，那就不能认为乙车床产品的方差比甲车床的小.

怎样检验假设 H_0 呢？我们先进行一般性讨论.

设 $X_1,X_2,\cdots,X_{n_1}$ 是来自总体 X 的样本，$Y_1,Y_2,\cdots,Y_{n_2}$ 是来自总体 Y 的样本[$X\sim N(\mu_1,\sigma_1^2)$, $Y\sim N(\mu_2,\sigma_2^2)$, X,Y 相互独立].要检验 $H_0:\sigma_1^2\leqslant\sigma_2^2$，即 $\dfrac{\sigma_1^2}{\sigma_2^2}\leqslant 1$.

自然想到用比值

$$F=\frac{S_1^2}{S_2^2} \tag{4.6}$$

作统计量.从直观上看，如果根据所给的样本值算出来的 F 值远大于 1，则有理由否定假设 H_0.因而希望能求出，在 H_0 成立的条件下，F 所服从的分布.可惜的是，由于现在 σ_1^2,σ_2^2 不知道(没有假定 $\sigma_1=\sigma_2$ 成立，因此和前面的情况不同)，F 的概率分布求不出来.然而，我们知道

$$\widetilde{F}=\frac{\frac{S_1^2}{\sigma_1^2}}{\frac{S_2^2}{\sigma_2^2}}$$

服从自由度为 n_1-1, n_2-1 的 F 分布.

查 F 分布表,可找到 λ 满足:

$$P\{\widetilde{F}>\lambda\}=\alpha$$

这就是说 $\{\widetilde{F}>\lambda\}$ 是一个"小概率事件".

如果假设 $H_0(\sigma_1^2\leqslant\sigma_2^2)$ 成立,则 $\widetilde{F}\geqslant F$,于是 $\{F>\lambda\}$ 更是一个小概率事件.

现在可以下结论了.如果根据所给的样本值算出 F 的值大于 λ,则应否定假设 H_0;反之,若算出的 F 的值不超过 λ,则假设 H_0 是相容的.

所以要检验 $\sigma_1^2\leqslant\sigma_2^2$,关键在于记住统计量(4.6)的表达式和会查 F 分布的临界值表.

现在应用上述办法到例 3.2 中去.

第一步:提出待检验的假设 $H_0:\sigma_1^2\leqslant\sigma_2^2$.

第二步:计算统计量 $F=\frac{S_1^2}{S_2^2}$ 的值. $n_1=8, n_2=9$.

$$\overline{x}=15.01, \overline{y}=14.99$$

$$\sum_{i=1}^{8}(x_i-\overline{x})^2=0.67,$$

$$\sum_{i=1}^{9}(y_i-\overline{y})^2=0.21$$

于是

$$F=\frac{s_1^2}{s_2^2}=\frac{0.67}{0.21}\times\frac{8}{7}=3.65$$

取 $\alpha=0.05$,查 F 分布表,第一自由度是 7,第二自由度是 8,

得 $\lambda = 3.50$.

现在 $F = 3.65 > 3.50$,故应否定假设 $H_0: \sigma_1^2 \leqslant \sigma_2^2$,也就是说,$\sigma_2^2 < \sigma_1^2$ 成立,乙车床产品的直径的方差比甲车床的小.

例 4.6 国外一家电视台用很长的节目进行赈灾演出,以便得到观众们的捐赠.为了了解捐赠情况,随机抽查了 25 个男士,平均捐赠额是 12.40 美元,标准差是 2.50 美元;还随机抽查了 25 个女士,平均捐赠额是 8.90 美元,标准差是 1.34 美元.问:男士捐赠额的方差是否大于女士捐赠额的方差?

解 我们假设一个男士的捐赠额 X 服从正态分布 $N(\mu_1, \sigma_1^2)$,一个女士的捐赠额 Y 服从正态分布 $N(\mu_2, \sigma_2^2)$.参数 μ_1, σ_1^2, μ_2, σ_2^2 都是未知的.我们来检验假设 $H_0: \sigma_1^2 \leqslant \sigma_2^2$.

使用统计量 $F = S_1^2/S_2^2$ (见(4.6)),现在第一自由度是 24,第二个自由度也是 24,对于 $\alpha = 0.01$,查 F 分布表,得临界值 $\lambda = 2.66$ 现在 $S_1^2 = (2.50)^2$, $S_2^2 = (1.34)^2$, $F = 3.48 > \lambda$,故应拒绝 H_0,因而可以认为,男士捐赠额的方差大于女士捐赠额的方差.

4. 未知 σ_1^2, σ_2^2 但知道 $\sigma_1^2 \neq \sigma_2^2$,检验假设 $H_0: \mu_1 = \mu_2$.

这是著名的 Behrens - Fisher 问题.其解决方法介绍如下.设 $X_1, X_2, \cdots, X_{n_1}$ 是来自 $N(\mu_1, \sigma_1^2)$ 的样本, $Y_1, Y_2, \cdots, Y_{n_2}$ 是来自 $N(\mu_2, \sigma_2^2)$ 的样本,两个样本相互独立.令 $\overline{X} = \dfrac{1}{n_1}\sum\limits_{i=1}^{n_1} X_i$, $\overline{Y} = \dfrac{1}{n_2}\sum\limits_{i=1}^{n_2} Y_i$.

$$S_1^2 = \frac{1}{n_1 - 1}\sum_{i=1}^{n_1}(X_i - \overline{X})^2, S_2^2 = \frac{1}{n_2 - 1}\sum_{i=1}^{n_2}(Y_i - \overline{Y})^2$$

易知

$$\overline{X} - \overline{Y} \sim N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$$

于是

$$\frac{\bar{X}-\bar{Y}-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}\sim N(0,1)$$

在零假设 $H_0:\mu_1=\mu_2$ 下，

$$\xi\overset{\Delta}{=}\frac{\bar{X}-\bar{Y}}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}\sim N(0,1)$$

可见$|\xi|$值太大时应拒绝 H_0，但 σ_1^2 和 σ_2^2 是未知的，ξ 不是统计量.自然想到用 S_1^2 代替 σ_1^2，S_2^2 代替 σ_2^2.于是应采用统计量

$$T=\frac{\bar{X}-\bar{Y}}{\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}} \tag{4.7}$$

当$|T|$太大时应拒绝 H_0.应指出的是在 H_0 下 T 的精确分布相当复杂(而且依赖于比值$\frac{\sigma_1^2}{\sigma_2^2}$).幸运的是，可以证明，在 H_0 下，统计量 T 近似服从 m 个自由度的 t 分布，这个 m 乃是与下列 m^* 最接近的整数：

$$m^*=\frac{\left(\frac{1}{n_1}S_1^2+\frac{1}{n_2}S_2^2\right)^2}{\frac{1}{n_1-1}\left(\frac{S_1^2}{n_1}\right)^2+\frac{1}{n_2-1}\left(\frac{S_2^2}{n_2}\right)^2} \tag{4.8}$$

利用 t 分布表，找临界值 λ 满足 $P(|T|>\lambda)=\alpha$.于是当且仅当 $|T|>\lambda$ 时拒绝 $H_0:\mu_1=\mu_2$.

类似地，也可解决 $\sigma_1^2\neq\sigma_2^2$ 时如何检验 $H_0:\mu_1\leqslant\mu_2$ 的问题.

例 4.7 研究患心脏病的父亲是否引起子女的胆固醇水平偏高的问题.随机调查了 100 个 2 至 14 岁的孩子(其父皆死于心脏病)，其胆固醇水平的平均值是 207.3，标准差是 35.6；另外，随机调查了父亲无心脏病史的 74 个 2 至 14 岁的孩子，其胆固醇水平的平均值是 193.4，标准差是 17.3，问：前者的胆固醇水平的平均

值与后者的胆固醇水平的平均值是否有显著差异?

解 设父亲死于心脏病的孩子的胆固醇水平 X 服从正态分布 $N(\mu_1,\sigma_1^2)$,父亲无心脏病史的孩子的胆固醇水平 Y 服从正态分布 $N(\mu_2,\sigma_2^2)$.这里参数 $\mu_1,\sigma_1^2,\mu_2,\sigma_2^2$ 都是未知的.我们要检验的假设是 $H_0:\mu_1=\mu_2$.

首先判别 σ_1^2 是否与 σ_2^2 相等,使用统计量 $F=\dfrac{S_1^2}{S_2^2}$,这里 S_1^2 和 S_2^2 分别是两个样本的方差.设 $\alpha=0.05$,从(4.4)、(4.5)及 F 分布临界值表,知:$\lambda_1=0.6548,\lambda_2=1.5491$.现在 $S_1^2=(35.6)^2,S_2^2=(17.3)^2,F=4.23>\lambda_2$,故可以认为 $\sigma_1^2\neq\sigma_2^2$.

为了检验 $H_0:\mu_1=\mu_2$,使用统计量(4.7),即

$$T=\frac{\bar{X}-\bar{Y}}{\sqrt{\dfrac{1}{n_1}S_1^2+\dfrac{1}{n_2}S_2^2}}\quad(n_1=100,n_2=74).$$

现在 $\bar{X}=207.3,\bar{Y}=193.4$,代入知 T 的值为 3.40.

从(4.8)知 $m^*=151.4$,于是 $m=151$.即 H_0 下统计量 T 近似服从 151 个自由度的 t 分布.设 $t_{0.975}(l)$是 l 个自由度的 t 分布的 0.975 分位数,则 H_0 下 $P[|T|>t_{0.975}(151)]=0.05$,故临界值 $\lambda=t_{0.975}(151)<t_{0.975}(120)=1.980$.现在 $|T|=3.40>1.980$.因而应拒绝 $H_0:\mu_1=\mu_2$.即可以认为胆固醇水平的平均值有显著性差异,父亲无心脏病史的孩子的胆固醇水平的平均值确实低些.

作为本节的结束,我们指出,t 分布与 F 分布有密切的关系:如果 X 服从 n 个自由度的 t 分布,则 X^2 服从自由度为 $1,n$ 的 F 分布.

证明很简单.直接计算一下.设 X 的分布密度是 $p(x)$,则从第二章知 X^2 的分布密度 $f(x)$可这样计算出来:

$$f(x)=\begin{cases}\dfrac{1}{2\sqrt{x}}[p(\sqrt{x})+p(-\sqrt{x})] & x>0\\ 0 & x\leqslant 0\end{cases}$$

现在

$$p(x)=\frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\cdot\Gamma\left(\frac{n}{2}\right)}\left(1+\frac{x^2}{n}\right)^{-\frac{n+1}{2}}$$

易知,当 $x>0$ 时,

$$p(\sqrt{x})=p(-\sqrt{x})=\frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\cdot\Gamma\left(\frac{n}{2}\right)}\left(1+\frac{x}{n}\right)^{-\frac{n+1}{2}}$$

于是,当 $x>0$ 时,

$$f(x)=\frac{2}{2\sqrt{x}}\cdot\frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\cdot\Gamma\left(\frac{n}{2}\right)}\left(1+\frac{x}{n}\right)^{-\frac{n+1}{2}}$$

$$=\frac{\Gamma\left(\frac{1+n}{2}\right)}{\Gamma\left(\frac{1}{2}\right)\Gamma\left(\frac{n}{2}\right)}\cdot\left(\frac{1}{n}\right)^{\frac{1}{2}}x^{\frac{1}{2}-1}\left(1+\frac{1}{n}x\right)^{-\frac{1+n}{2}}$$

这就表明 $f(x)$刚好是自由度为 $1,n$ 的 F 分布的密度.这就证明了,X^2 服从自由度为 $1,n$ 的 F 分布.

这样,我们手头只要有一张 F 分布的表,也能处理需要 t 分布的检验问题了.

§5 比率的假设检验

设 X 服从二点分布(伯努利分布).X 取值 1 或 0,且 $P(X=1)=p=1-P(X=0)$,这里 p 是未知的.p 就是所谓的"比率".当 1 表示成功,0 表示失败,则 p 就是成功率;当 1 表示合格,0 表示不合格,p 就是合格率;当 1 表示有效,0 表示无效,p 就是有效率.

先讨论一个总体的问题.此时对 p 的假设检验问题有下列三

个：

① $H_0: p \leqslant p_0, H_a: p > p_0$

② $H_0: p \geqslant p_0, H_a: p < p_0$

③ $H_0: p = p_0, H_a: p \neq p_0$,

这里 H_0 是要检验的零假设(其中 p_0 是已知数，$0 < p_0 < 1$)，H_a 是备择假设.

首先讨论上列问题①. 设 X 的简单随机样本是 $X_1, X_2, \cdots, X_n$，如何检验 $H_0: p \leqslant p_0$?

很自然想到 p 的估计量 $\hat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i$，则 n 比较大时 $\hat{p}$ 应与 p 很接近. 故 $\hat{p}$ 比 p_0 大很多时应拒绝 H_0，令 $S = \sum_{i=1}^{n} X_i$. 故对固定的 n，当 S 足够大时应拒绝 H_0. 所以否定域是 $\{(x_1, x_2, \cdots, x_n): \sum_{i=1}^{n} x_i \geqslant c\}$. 其中 c 是临界值，设检验水平是 α，取 c 为满足下式的最小整数.

$$\sup_{p \leqslant p_0} P_p(S \geqslant c) \leqslant \alpha \tag{5.1}$$

这里 $P_p(A)$ 表示总体的参数是 p 时事件 A 的概率(下同).

注意 S 服从二项分布，即

$$P_p(S = i) = C_n^i p^i (1-p)^{n-i} \quad (i = 0, 1, \cdots, n)$$

于是

$$P_p(S \geqslant k) = \sum_{i=k}^{n} C_n^i p^i (1-p)^{n-i}$$

$$= \frac{n!}{(k-1)!\,(n-k)!} \int_0^p u^{k-1}(1-u)^{n-k}\,du$$

($k \geqslant 1$，见第五章 §7)

可见 $P_p(S \geqslant k)$ 是 p 的增函数，于是(5.1)化为

$$P_{p_0}(S \geqslant c) \leqslant \alpha \tag{5.2}$$

我们不去求这个 c(即满足(5.2)的最小整数).而是另想办法判别事件"$S\geqslant c$"是否发生.设样本值是 $x_1,x_2,\cdots,x_n$(n 个已知数),令 $S_0=\sum_{i=1}^{n}x_i$.显然,从(5.2)知 $S_0\geqslant c$ 的充要条件是

$$\sum_{i=S_0}^{n}C_n^i p_0^i(1-p_0)^{n-i}\leqslant\alpha \tag{5.3}$$

设方程

$$\sum_{i=S_0}^{n}C_n^i p^i(1-p)^{n-i}=\alpha \tag{5.4}$$

的根为 $p_\alpha(S_0)(S_0\geqslant 1)$.此外,规定 $p_\alpha(0)=0$.可见 $S_0\geqslant c$ 的充要条件是 $p_0\leqslant p_\alpha(S_0)$.故当且仅当 $p_0\leqslant p_\alpha(S_0)$时拒绝 $H_0:p\leqslant p_0$.怎样计算 $p_\alpha(S_0)$呢?可以证明[①] $p_\alpha(S_0)$可用 F 分布的分位数表达出来,即

$$p_\alpha(S_0)=\left\{1+\frac{n-S_0+1}{S_0}F_{1-\alpha}[2(n-S_0+1),2S_0]\right\}^{-1}. \tag{5.5}$$

---

① (5.5)式的严格证明如下.令

$$\beta(x;p,q)=\frac{1}{B(p,q)}\int_0^x u^{p-1}(1-u)^{q-1}\mathrm{d}u\quad(0\leqslant x\leqslant 1)$$

$(p>0,q>0,B(p,q)=\int_0^1 u^{p-1}(1-u)^{q-1}\mathrm{d}u)$

这是参数为 p,q 的贝塔分布函数.我们首先指出下列事实:设 X 与 Y 相互独立,X 服从 m 个自由度的 χ^2 分布,Y 服从 n 个自由度的 χ^2 分布,则

$$U=\frac{X}{X+Y}$$

的分布函数是 $\beta\left(x;\frac{m}{2},\frac{n}{2}\right)$.

证明方法是:令 $V=X+Y$,直接计算 (U,V) 的联合密度,然后就可证明 U 的分布函数正好是 $\beta\left(x;\frac{m}{2},\frac{n}{2}\right)$.

其次,我们指出:设 $\beta_r(k_1,k_2)$是贝塔分布 $\beta(x;k_1,k_2)$的 r 分位数(k_1,k_2都是整数),$F_r(n_1,n_2)$是自由度为 n_1、n_2的 F 分布的 r 分位数,则

例 5.1 一种广泛使用的药其治疗慢性支气管炎的有效率是0.80.现在一家制药公司推出一种新药,声称:治疗慢性支气管炎的有效率高于0.80,且药价比广泛使用的那种药减少四分之一.为了验证新药的有效率是否高于0.80,收集了临床试验数据.从使用新药的病人中随机抽查了30人,其中该药对27人有效3人无效.问:能否认为新药的有效率高于0.80?

解 用 X 表示使用新药的效果.$X=1$ 表示有效,$X=0$ 表示无效.现在样本量 $n=30$.样本值是 $x_1,x_2,\cdots,x_{30}$,$S_0=\sum_{i=1}^{30}x_i=27$.设检验水平 $\alpha=0.05$.从(5.5)知

$$p_\alpha(S_0)=p_{0\,05}(27)=\left[1+\frac{4}{27}F_{0\,95}(8,54)\right]^{-1}$$

(接上页注)

$$\beta_r(k_1,k_2)=\left[1+\frac{k_2}{k_1}\cdot\frac{1}{F_r(2k_1,2k_2)}\right]^{-1}. \qquad (注\ 5.1)$$

实际上,取 $m=2k_1$,$n=2k_2$,$F=\frac{k_2}{k_1}\cdot\frac{X}{Y}$(这里 X 与 Y 分别服从 m 个自由度和 n 个自由度的 χ^2 分布),则 $U=\frac{X}{X+Y}=\frac{F}{\frac{k_2}{k_1}+F}$ 且 F 服从 $2k_1,2k_2$ 个自由度的 F 分布.于是

$$\beta(x;k_1,k_2)=P(U\leqslant x)=P\left(F\leqslant\frac{k_2}{k_1}\cdot\frac{x}{1-x}\right)(0<x<1)$$

在此式中令 $x=\beta_r(k_1,k_2)$,即知

$$F_r(2k_1,2k_2)=\frac{k_2}{k_1}\cdot\frac{\beta_r(k_1,k_2)}{1-\beta_r(k_1,k_2)}$$

由此可推出(注5.1)成立.

(5.4)式等价于

$$\frac{n!}{S_0!\,(n-S_0)!}\int_0^p u^{S_0-1}(1-u)^{n-S_0}\,du=\alpha$$

此方程的根 $p_\alpha(S_0)$ 正好是贝塔分布函数 $\beta(x,S_0,n-S_0+1)$ 的 α 分位数.于是从(注5.1)得到

$$p_\alpha(S_0)=\left\{1+\frac{n-S_0+1}{S_0}[F_\alpha(2S_0,2(n-S_0+1))]^{-1}\right\}^{-1}$$

再注意 $[F_r(n_1,n_2)]^{-1}=F_{1-r}(n_2,n_1)$,即知(5.5)成立.

$$=\left(1+\frac{4}{27}\times 2.13\right)^{-1}=0.76<p_0=0.80$$

故不能拒绝 $H_0: p\leqslant p_0$. 即没有理由说新药比老药有更高的有效率. 顺便说一下，若 30 人中有 28 人有效，则可计算出 $p_{0\,05}(28)=\left(1+\frac{3}{28}F_{0\,95}(6,56)\right)^{-1}=0.814>p_0$. 故此时应拒绝 H_0. 即可认为新药有更高的有效率.

现在来讨论假设 $H_0: p\geqslant p_0$ 的检验问题. 设 $X_1, X_2, \cdots, X_n$ 是简单随机样本. 很自然想到，当 $S=\sum_{i=1}^{n} X_i$ 太小时应拒绝 H_0. 对给定的检验水平 α，设 c 是满足下式的最大整数.

$$\sup_{p\geqslant p_0} P_p(S\leqslant c)\leqslant\alpha \tag{5.6}$$

可以证明 $P_p(S\leqslant c)$ 是 p 的减函数，于是(5.6)化为

$$P_{p_0}(S\leqslant c)\leqslant\alpha \tag{5.7}$$

我们不去寻找这个临界值 c，而是另想办法判别“$S\leqslant c$”是否发生. 仿效前一检验问题的处理方法，设样本值是 $x_1, x_2, \cdots, x_n$，令 $S_0=\sum_{i=1}^{n} x_i$. 不难看出 $S_0\leqslant c$ 的充要条件是 $P_{p_0}(S\leqslant S_0)\leqslant\alpha$，即

$$\sum_{i=0}^{S_0} C_n^i p_0^i (1-p_0)^{n-i}\leqslant\alpha \tag{5.8}$$

设方程

$$\sum_{i=0}^{S_0} C_n^i p^i (1-p)^{n-i}=\alpha \tag{5.9}$$

的根为 $\tilde{p}_\alpha(S_0)$. 于是 $S_0\leqslant c$ 的充要条件是 $\tilde{p}_\alpha(S_0)\leqslant p_0$（因为(5.9)中等号左边是 p 的严格减函数）. 所以当且仅当 $\tilde{p}_\alpha(S_0)\leqslant p_0$ 时应拒绝 $H_0: p\geqslant p_0$.

怎样计算 $\tilde{p}_\alpha(S_0)$ 呢？从(5.9)看出，$\sum_{i=S_0+1}^{n} C_n^i[\tilde{p}_\alpha(S_0)]^i[1-$

$\tilde{p}_\alpha(S_0)]^{n-1}=1-\alpha$，故$\tilde{p}_\alpha(S_0)$是贝塔分布$\beta(x;S_0+1,n-S_0)$的$1-\alpha$分位数.于是

$$\tilde{p}_\alpha(S_0)=\left[1+\frac{n-S_0}{S_0+1}\frac{1}{F_{1-\alpha}(2(S_0+1),2(n-S_0))}\right]^{-1},\tag{5.10}$$

这里$F_{1-\alpha}(n_1,n_2)$是自由度为n_1,n_2的F分布的$1-\alpha$分位数.(参看(注5.1)).

现在来研究$H_0:p=p_0$的检验问题.

设$X_1,X_2,\cdots,X_n$是简单随机样本，$S=\sum_{i=1}^{n}X_i$.显然，S太大或太小应拒绝$H_0:p=p_0$.故对给定的检验水平α，应取最大的整数c_1和最小的整数c_2满足：

$$P_{p_0}(S\leqslant c_1)\leqslant\frac{\alpha}{2},P_{p_0}(S\geqslant c_2)\leqslant\frac{\alpha}{2}.$$

我们不去找临界值c_1和c_2的具体数值，而是另想办法判别事件“$S\leqslant c_1$或$S\geqslant c_2$”是否发生.(当且仅当这个事件发生时拒绝H_0).设样本值是$x_1,x_2,\cdots,x_n$，$S_0=\sum_{i=1}^{n}x_i$.不难看出$S_0\leqslant c_1$的充要条件是$P_{p_0}(S\leqslant S_0)\leqslant\frac{\alpha}{2}$；$S_0\geqslant c_2$的充要条件是$P_{p_0}(S\geqslant S_0)\leqslant\frac{\alpha}{2}$.于是$S_0\leqslant c_1$的充要条件是$\tilde{p}_{\frac{\alpha}{2}}(S_0)\leqslant p_0$(参看(5.10))；$S_0\geqslant c_2$的充要条件是$p_0\leqslant p_{\frac{\alpha}{2}}(S_0)$(参看(5.5)).可见，当且仅当$p_{\frac{\alpha}{2}}(S_0)\leqslant p_0$或者$p_{\frac{\alpha}{2}}(S_0)\geqslant p_0$时应拒绝$H_0:p=p_0$.

现在来研究两个总体的比较问题.设X与Y相互独立，都服从伯努利分布.$P(X=1)=p_1=1-p(X=0)$，$P(Y=1)=p_2=1-P(Y=0)$.p_1,p_2未知.设X有简单随机样本$X_1,X_2,\cdots,X_{n_1}$，Y有简单随机样本$Y_1,Y_2,\cdots,Y_{n_2}$，考虑下列三个零假设的检验

问题：

④ $H_0: p_1 \leqslant p_2, H_a: p_1 > p_2$

⑤ $H_0: p_1 \geqslant p_2, H_a: p_1 < p_2$

⑥ $H_0: p_1 = p_2, H_a: p_1 \neq p_2$，

这里 H_0 是零假设，H_a 是备择假设.

令 $S_1 = \sum_{i=1}^{n_1} X_i, S_2 = \sum_{i=1}^{n_2} Y_i$，则 p_1 和 p_2 的估计量分别是 $\hat{p}_1 = \frac{S_1}{n_1}, \hat{p}_2 = \frac{S_2}{n_2}$，很自然想到：当 $\hat{p}_1$ 比 $\hat{p}_2$ 大得多时应拒绝 $H_0: p_1 \leqslant p_2$；当 $\hat{p}_1$ 比 $\hat{p}_2$ 小得多时应拒绝 $H_0: p_1 \geqslant p_2$；当 $\hat{p}_1$ 与 $\hat{p}_2$ 相差得多时应拒绝 $H_0: p_1 = p_2$. 这是一种定性的说法，“大得多”、“小得多”、“相差得多”都是不确切的，由于 p_1 和 p_2 都未知，临界值较难确定. 为了检验上述假设，我们给出两个检验法. 一是正态理论方法，这是大样本情形的近似方法；另一是 Fisher 精确检验法，各种情形下都可以用，但计算上比较复杂.

先介绍正态理论方法. 易知 $D(\hat{p}_1 - \hat{p}_2) = \frac{1}{n_1} p_1(1-p_1) + \frac{1}{n_2} p_2(1-p_2)$. 令

$$\xi = \frac{\hat{p}_1 - \hat{p}_2 - (p_1 - p_2)}{\sqrt{\frac{1}{n_1}\hat{p}_1(1-\hat{p}_1) + \frac{1}{n_2}\hat{p}_2(1-\hat{p}_2)}} \tag{5.11}$$

$$\eta = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{1}{n_1}\hat{p}_1(1-\hat{p}_1) + \frac{1}{n_2}\hat{p}_2(1-\hat{p}_2)}} \tag{5.12}$$

$$\zeta = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\hat{p}(1-\hat{p})}}, \tag{5.13}$$

这里 $\hat{p}_1 = \frac{1}{n_1}\sum_{i=1}^{n_1} X_i, \hat{p}_2 = \frac{1}{n_2}\sum_{i=1}^{n_2} Y_i$，

$$\hat{p}=\frac{1}{n_1+n_2}(n_1\hat{p}_1+n_2\hat{p}_2)$$

数学上可以证明，当 n_1 和 n_2 相当大(一般要求 $n_1\hat{p}_1(1-\hat{p}_1)\geqslant 5$，$n_2\hat{p}_2(1-\hat{p}_2)\geqslant 5$)时，随机变量 ξ 近似服从标准正态分布. 给定检验水平 α，设 z_α 是标准正态分布的 α 分位数，则 $P(\xi>z_{1-\alpha})\approx\alpha$. 易知，在 $p_1\leqslant p_2$ 的假设下 $\xi\geqslant\eta$，更有 $P(\eta>z_{1-\alpha})\leqslant P(\xi>z_{1-\alpha})\approx\alpha$. 于是 $\eta>z_{1-\alpha}$ 时拒绝 $H_0:p_1\leqslant p_2$，当 $\eta\leqslant z_{1-\alpha}$ 时不拒绝 H_0.

类似地，在 $p_1\geqslant p_2$ 的假设下，$\xi\leqslant\eta$，从而 $P(\eta<z_\alpha)\leqslant P(\xi<z_\alpha)\approx\alpha$. 于是当且仅当 $\eta<z_\alpha$ 时拒绝假设 $H_0:p_1\geqslant p_2$.

在假设 $p_1=p_2$ 的条件下，只要 n_1,n_2 相当大(一般要求 $n_1\hat{p}_1(1-\hat{p}_1)\geqslant 5$，$n_2\hat{p}_2(1-\hat{p}_2)\geqslant 5$)，统计量 ζ(见(5.13))近似服从标准正态分布，从而 $P(|\zeta|>z_{1-\frac{\alpha}{2}})\approx\alpha$. 这里 $z_{1-\frac{\alpha}{2}}$ 是标准正态分布的 $1-\frac{\alpha}{2}$ 分位数.

当且仅当 $|\zeta|>z_{1-\frac{\alpha}{2}}$ 时拒绝 $H_0:p_1=p_2$.

例 5.2 研究口服避孕药对年龄在 40 至 44 岁的妇女心脏的影响. 收集的资料表明，在 5000 个使用口服避孕药的妇女中三年内出现心肌梗死的有 13 人；而在 10000 个不使用口服避孕药的妇女中三年内出现心肌梗死的有 7 人. 试问：口服避孕药是否对妇女的心脏有显著的影响？

解 用 p_1 表示年龄在 40 至 44 岁的妇女由于口服避孕药导致三年内出现心肌梗死的概率，p_2 表示这个年龄段的妇女不服这种避孕药但在三年内出现心肌梗死的概率. 我们要检验的假设是 $H_0:p_1=p_2$，备择假设是 $H_a:p_1\neq p_2$. 使用统计量(5.13). 现在

$$\hat{p}_1=\frac{13}{5000}=0.0026,\ \hat{p}_2=\frac{7}{10000}=0.0007$$

$$\hat{p}=\frac{13+7}{15000}=0.00133$$

由于 $n_1\hat{p}_1(1-\hat{p}_1)=6.66\geqslant 5$, $n_2\hat{p}_2(1-\hat{p}_2)=6.70\geqslant 5$, 故可用统计量 ζ(见(5.13)). 可计算出 $\zeta=3.01$. 设检验水平 $\alpha=0.01$. 查标准正态分布的数值表知 $1-\frac{\alpha}{2}$ 分位数 $z_{0.995}=2.58$. 既然 $\zeta=3.01>2.58$. 故应拒绝 $H_0: p_1=p_2$, 即口服避孕药对 40 至 44 岁的妇女的心脏有显著影响.

现在来介绍 Fisher 精确检验法. 此时对样本量无任何限制. 先介绍操作方法, 然后介绍这个检验法是基于何种统计思想推导出来的. 设 $X_1, X_2, \cdots, X_{n_1}$ 是第一个总体的简单随机样本, $Y_1, Y_2, \cdots, Y_{n_2}$ 是第二个总体的简单随机样本, p_1 和 p_2 分别是两个总体的参数, p_1 和 p_2 均未知. 令 $S_1=\sum_{i=1}^{n_1} X_i$, $S_2=\sum_{i=1}^{n_2} Y_i$. 设两个样本的样本值分别是 $x_1, x_2, \cdots, x_{n_1}$ 和 $y_1, y_2, \cdots, y_{n_2}$. 令

$$S_1^0=\sum_{i=1}^{n_1} x_i,\ S_2^0=\sum_{i=1}^{n_2} y_i,\ t=S_1^0+S_2^0 \tag{5.14}$$

为了检验 $H_0: p_1\leqslant p_2$(备择假设是 $H_a: p_1>p_2$), 令

$$p_1(S_1^0)=\sum_{i\geqslant S_1^0} p(i), \tag{5.15}$$

这里

$$p(i)=\frac{\binom{n_1}{i}\binom{n_2}{t-i}}{\binom{n_1+n_2}{t}}\ ^{①}\quad (i=0,1,\cdots) \tag{5.16}$$

对给定的检验水平 α, 当且仅当 $p_1(S_1^0)\leqslant\alpha$ 时拒绝 $H_0: p_1\leqslant p_2$.

为了检验 $H_0: p_1\geqslant p_2$(备择假设是 $H_a: p_1<p_2$), 令

---

① $\binom{m}{i}$ 就是组合数 C_m^i. 当 $i>m$ 或 $i<0$ 时规定 $\binom{m}{i}=0$.

$$p_2(S_1^0)=\sum_{i\leqslant S_1^0}p(i) \tag{5.17}$$

($p(i)$的定义见(5.16))

对给定的检验水平 α，当且仅当 $p_2(S_1^0)\leqslant\alpha$ 时拒绝 $H_0:p_1\geqslant p_2$.

为了检验 $H_0:p_1=p_2$(备择假设是 $H_a:p_1\neq p_2$)，令

$$p_3(S_1^0)=\alpha\min\Big[\sum_{i\leqslant S_1^0}p(i),\ \sum_{i\geqslant S_1^0}p(i)\Big]. \tag{5.18}$$

对给定的检验水平 α，当且仅当 $p_3(S_1^0)\leqslant\alpha$ 时拒绝 $H_0:p_1=p_2$.

上述检验法就是 Fisher 精确检验法. 复杂之处在于要计算各个 $p(i)$，在实际计算时要利用下列递推关系式：当 $p(i)>0$ 时有

$$p(i+1)=p(i)\frac{(n_1-i)(t-i)}{(i+1)(n_2-t+i+1)} \tag{5.19}$$

这个关系式根据 $p(i)$的定义很容易验证. 后面还要介绍实际工作中采用的列联表变换法，它是根据(5.19)计算所有的 $p(i)$.

现在问：上述 Fisher 精确检验法是基于什么统计思想而导出的呢？沿用前面的记号，从数学上可以证明，如果假设 $H_0:p_1\leqslant p_2$ 成立，则在 $S_1+S_2=t$ 的条件下，$S_1\geqslant c$ 的条件概率的最大值是 $\sum_{i=c}^{n_1}p(i)$，即

$$\sup_{p_1\leqslant p_2}P_{p_1p_2}(S_1\geqslant c\mid S_1+S_2=t)=\sum_{i=c}^{n_1}p(i), \tag{5.20}$$

这里 $P_{p_1p_2}(A\mid S_1+S_2=t)$表示两个总体的参数分别是 p_1,p_2 时在 $S_1+S_2=t$ 的条件下事件 A 的条件概率，$p(i)$的定义见(5.16).

(5.20)的数学证明较长，从略. 给定 $\alpha\in(0,1)$. 设 c 是满足 $\sum_{i=c}^{n_1}p(i)\leqslant\alpha$ 的最小整数. 则在 $S_1+S_2=t$ 的条件下 $S_1\geqslant c$ 时应拒绝 $H_0:p_1\leqslant p_2$. 注意，根据样本值 $x_1,x_2,\cdots,x_{n_1},y_1,y_2,\cdots,y_{n_2}$，从(5.14)知事件“$S_1+S_2=t$”已发生，而“$S_1\geqslant c$”当且仅当 $S_1^0\geqslant c$ 时发生. 显然 $S_1^0\geqslant c$ 的充要条件是 $\sum_{i=S_1^0}^{n_2}p(i)\leqslant\alpha$. 从(5.15)知 $S_1^0\geqslant c$ 的充要条件是 $p_1(S_1^0)\leqslant\alpha$. 故当且仅当 p_1

$(S_1^0)\leqslant\alpha$ 时应拒绝 $H_0:p_1\leqslant p_2$.

类似地,数学上可以证明,如果假设 $H_0:p_1\geqslant p_2$ 成立,则在 $S_1+S_2=t$ 的条件下,$S_1\leqslant c$ 的条件概率的最大值是 $\sum_{i=0}^{c}p(i)$,即

$$\sup_{p_1\geqslant p_2}P_{p_1p_2}(S_1\leqslant c\mid S_1+S_2=t)=\sum_{i=0}^{c}p(i)$$

给定 $\alpha\in(0,1)$. 设 c 是满足 $\sum_{i=0}^{c}p(i)\leqslant\alpha$ 的最大整数. 从(5.17)知,$S_1^0\leqslant c$ 的充要条件是 $p_2(S_1^0)\leqslant\alpha$. 故当且仅当 $p_2(S_1^0)\leqslant\alpha$ 时应拒绝 $H_0:p_1\geqslant p_2$.

数学上可以证明,如果 $H_0:p_1=p_2$ 成立,则

$$P_{p_1p_1}(S_1\leqslant c_1\mid S_1+S_2=t)=\sum_{i=0}^{c_1}p(i)$$

$$P_{p_1p_1}(S_1\geqslant c_2\mid S_1+S_2=t)=\sum_{i=c_2}^{n_1}p(i)$$

取最大的整数 c_1 满足 $\sum_{i=0}^{c_1}p(i)\leqslant\frac{\alpha}{2}$. 再取最小的整数 c_2 满足 $\sum_{i=c_2}^{n_1}p(i)\leqslant\frac{\alpha}{2}$,则在 $H_0:p_1=p_2$ 成立且 $S_1+S_2=t$ 的条件下,事件"$S_1\leqslant c_1$ 或 $S_1\geqslant c_2$"的条件概率不超过 α. 可见,在 $S_1+S_2=t$ 的条件下 $S_1\leqslant c_1$ 或 $S_1\geqslant c_2$ 发生时应拒绝 $H_0:p_1=p_2$. 根据样本值 $x_1,x_2,\cdots,x_{n_1}$ 及 $y_1,y_2,\cdots,y_{n_2}$ 和(5.14)知"$S_1+S_2=t$"已经发生,故 $S_1^0\leqslant c_1$ 或 $S_1^0\geqslant c_2$ 时应拒绝 $H_0:p_1=p_2$. 显然 $S_1^0\leqslant c_1$ 的充要条件是 $2\sum_{i=0}^{S_1^0}p(i)\leqslant\alpha$,$S_1^0\geqslant c_2$ 的充要条件是 $2\sum_{i=S_1^0}^{n_1}p(i)\leqslant\alpha$. 故从(5.18)知,$S_1^0\leqslant c_1$ 或 $S_1^0\geqslant c_2$ 成立的充要条件是 $p_3(S_1^0)\leqslant\alpha$. 这表明,当且仅当 $p_3(S_1^0)\leqslant\alpha$ 时应拒绝 $H_0:p_1=p_2$.

以上叙述了导出 Fisher 精确检验法的统计思想. 下面介绍实际工作中采用的用于计算所有 $p(i)$ 的具体方法. 先引进一个定义. 设 S_1^0,S_2^0 和 t 由(5.14)给出.

称非负整数组成的矩阵

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}$$

为宜取的，若 $a+b=n_1, c+d=n_2, a+c=t$. 显然，对于给定的 n_1, n_2 及 t，宜取阵由其左上角的元素 a 所惟一确定. 左上角是 a 的阵称为 a 阵，用 $\boldsymbol{A}_a$ 来表示. 显然，若

$$\boldsymbol{A}_a = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \qquad (b \geqslant 1, c \geqslant 1)$$

则

$$\boldsymbol{A}_{a+1} = \begin{pmatrix} a+1 & b-1 \\ c-1 & d+1 \end{pmatrix}$$

从(5.19)知

$$p(a+1) = p(a)\frac{bc}{(a+1)(d+1)} \tag{5.21}$$

(5.21)比(5.19)的好处在于：公式便于记忆.

从 $p(i)$ 的定义(见(5.16))知，若 i 不满足下列不等式(5.22)时 $p(i)=0$.

$$n_0 \leqslant i \leqslant n^*, \tag{5.22}$$

这里

$$n_0 = \max(0, n_2 - t), n^* = \min(n_1, t) \tag{5.23}$$

先依次列出 $A_{n_0}, A_{n_0+1}, \cdots, A_{n^*}$，然后计算 $p(n_0)$，再利用(5.21)逐次计算 $p(n_0+1), p(n_0+2), \cdots$. 当然，对于检验 $H_0: p_1 \leqslant p_2$，只需列出 $A_{S_1^0}, A_{S_1^0+1}, \cdots, A_{n^*}$，计算出相应的 $p(S_1^0), p(S_1^0+1), \cdots, p(n^*)$；对于检验 $H_0: p_1 \geqslant p_2$，只需列出 $A_{n_0}, A_{n_0+1}, \cdots, A_{S_1^0}$，计算出相应的 $p(n_0), p(n_0+1), \cdots, p(S_1^0)$.

例 5.3 某公安局有两个专案组，在过去一年内一组接手 25 件人命案，结果侦破了 23 件，另一组接手 35 件人命案，结果侦破了 30 件，问：两个组的侦破能力有无差别？

解 设两个组的侦破率分别为 p_1, p_2，要检验的假设是 $H_0: p_1 = p_2$.

(注意，设 X, Y 都是二值随机变量，$X=1$ 表示第一组侦破成功，

$X=0$ 表示未能侦破，$p_1=P(X=1)$；$Y=1$ 表示第二组侦破成功，$Y=0$ 表示未能侦破，$p_2=P(Y=1)$）. 我们采用 Fisher 精确检验法来检验 H_0. 现在 $n_1=25, n_2=35, S_1^0=23, S_2^0=30, t=53$（参看(5.14)）. 从(5.23)知 $n_0=0, n^*=25$. 从(5.18)知

$$p_3(S_1^0)=2\min\Big[\sum_{i=0}^{23}p(i),\ \sum_{i=23}^{25}p(i)\Big].$$

从(5.16)知 $p(23)=0.252$，

从(5.21)知 $p(24)=p(23)\dfrac{2\times30}{24\times6}=0.105$

$$p(25)=p(24)\frac{1\times29}{25\times7}=0.017$$

于是 $\sum_{i=23}^{25}p(i)=0.374$，$\sum_{i=0}^{23}p(i)=1-\sum_{i=23}^{25}p(i)+p(23)=1-0.374+0.252=0.878$. 从而 $p_3(S_1^0)=2\times0.374=0.748>0.05$. 于是在检验水平 $\alpha=0.05$ 下不应拒绝 $H_0: p_1=p_2$. 换句话说，没有理由认为两个专案组在破案能力上有差别.

§6 总体的分布函数的假设检验

在许多实际工作中经常假定总体服从正态分布，而对其数字特征（期望、方差等）进行假设检验，怎么知道一个总体的概率分布是正态分布呢？

更一般地，怎么知道一个随机变量 X 的分布函数是某个给定的函数 $F(x)$ 呢？

这是个十分重要的问题. 有时根据对事物本质的分析，利用概率论的知识，可以给予回答. 但在很多情况下，只能从一大堆数据中去发现规律，判断总体的分布是什么样子.

一般说来，总是先根据样本值（一批观测数据）用第五章中所介绍的直方图法，推测出总体可能服从的分布函数 $F(x)$（或密度

函数),然后再利用本节所讲的方法来检验该总体的分布函数是否真的就是 $F(x)$.

本节的内容,就是介绍如何检验假设 $H_0: X$ 以 $F(x)$为分布函数.

先讲一般性的检验办法,然后再用到具体例子上去.

设 $X_1, X_2, \cdots, X_n$ 是来自总体 X 的样本.在实轴上取 m 个点:$t_1, t_2, \cdots, t_m (t_1 < t_2 < \cdots < t_m)$,于是把实轴$(-\infty, +\infty)$分成 $m+1$ 段,第 1 段是$(-\infty, t_1]$,第 2 段是$(t_1, t_2]$,$\cdots$,第$m+1$段是$(t_m, +\infty)$,用 ν_i 表示 $X_1, X_2, \cdots, X_n$ 中落入第 i 段的个数$(i=1, 2, \cdots, m+1)$.这 ν_i 是频数,$\frac{\nu_i}{n}$是频率.用 p_i 表示 X 取值落于第 i 段的概率.如果假设 H_0 成立,则 p_i 是可以算得出来的.

实际上

$$p_1 = P\{X \leqslant t_1\} = F(t_1)$$

$$p_i = P\{t_{i-1} < X \leqslant t_i\} = F(t_i) - F(t_{i-1}) \quad (2 \leqslant i \leqslant m)$$

$$p_{m+1} = P\{X > t_m\} = 1 - F(t_m)$$

而 $F(x)$是已知的.

根据概率和频率的关系知道,如果 H_0 成立,那么$\frac{\nu_i}{n}$与 p_i 差不多,就是说$\left(\frac{\nu_i}{n} - p_i\right)^2$ 应该比较小,于是

$$V = \sum_{i=1}^{m+1} \left(\frac{\nu_i}{n} - p_i\right)^2 \cdot \frac{n}{p_i}$$

也应该比较小才合理.这里的因子$\frac{n}{p_i}$起平衡的作用.否则,对于较小的 p_i 而言,即使$\frac{\nu_i}{n}$跟 p_i 相对来说有较大的差别,$\left(\frac{\nu_i}{n} - p_i\right)^2$ 也不会很大.

我们就取 V 作统计量,由于样本 $X_1, X_2, \cdots, X_n$ 是随机变量,

于是

$$V = \sum_{i=1}^{m+1} \frac{(\nu_i - np_i)^2}{np_i}$$

也是随机变量.要紧的是求出 V 的概率分布,否则还不能用于假设检验.

经过数学方面的研究,可以证明(由于证明较长,我们不证了).在假设 H_0 成立的条件下,V 近似地服从 m 个自由度的 χ^2 分布.样本容量 n 越大,近似得越好.

于是,给定"检验标准"α 后,查 χ^2 分布表,可找到 λ 满足:

$$P\{V>\lambda\}=\alpha$$

这样$\{V>\lambda\}$便是"小概率事件".

现在好了,可以对假设 H_0 作出判断了.如果根据所给的样本值 $x_1, x_2, \cdots, x_n$,算得 V 的值大于λ,则否定假设 H_0;否则假设 H_0 是相容的.

我们把上述检验办法用到下面例子中去,希望读者仔细看看全部推算过程.

例 6.1 某车床生产滚珠,随机抽取了 50 个产品,测得它们的直径为(单位:mm):

15.0 15.8 15.2 15.1 15.9 14.7 14.8 15.5 15.6
15.3 15.1 15.3 15.0 15.6 15.7 14.8 14.5 14.2
14.9 14.9 15.2 15.0 15.3 15.6 15.1 14.9 14.2
14.6 15.8 15.2 15.9 15.2 15.0 14.9 14.8 14.5
15.1 15.5 15.5 15.1 15.1 15.0 15.3 14.7 14.5
15.5 15.0 14.7 14.6 14.2

经过计算知道,样本均值 $\bar{x}=15.1$,样本方差是$(0.432\,5)^2$.我们问,滚珠直径是否服从正态分布 $N(15.1,(0.432\,5)^2)$?于是我们就来检验假设 H_0:滚珠直径服从 $N(15.1,(0.432\,5)^2)$.

主要工作就是根据所给的样本值，计算统计量[①]

$$V=\sum_{i=1}^{m+1}\frac{(\nu_i-np_i)^2}{np_i}$$

的值.

为了计算，先要定分点 t_i. 可采用下列办法较为方便.（与第五章中作直方图时一样！）先找出所给样本值中的最小数与最大数，取比最小数略小的数作 a，比最大数略大的数作 b，将区间 $[a,b]$ 作 $m+1$ 等分，得分点 $t_1,t_2,\cdots,t_m(a<t_1<\cdots<t_m<b)$，这就是我们所要的. 至于 m 该取多大，还是和画直方图时所说的一样.

现在有 50 个数据，最小的是 14.2，最大的是 15.9，取 $a=$

---

① 从下面的计算过程知道，这个统计量的计算一般是比较麻烦的. 在正态性检验的情形（即检验已给的数据是否来自一个正态总体），有时人们愿意使用所谓偏度—峰度检验法. 这里简单地介绍一下，以引起读者的注意. 设 X 是一随机变量，$\mu=E(X)$，$\sigma^2=D(X)$，人们称 $\gamma=\frac{E(X-\mu)^3}{\sigma^3}$ 为 X 的偏度；称

$$\delta=\frac{E(X-\mu)^4}{\sigma^4}$$

为 X 的峰度.

当 X 是正态分布时，易知 $\gamma=0,\delta=3$.

为了检验数据 $x_1,x_2,\cdots,x_n$ 是否来自一个正态总体，先计算 γ,δ 的估计量：

$$\hat{\gamma}=\frac{\hat{m}_3}{s^3},\hat{\delta}=\frac{\hat{m}_4}{s^4}$$

其中，

$$s=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2},\ \hat{m}_3=\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^3$$

$$\hat{m}_4=\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^4$$

经过数学研究知道，如果 n 充分大，则 $\hat{m}_3$ 与三阶中心矩 $E(X-\mu)^3$ 很接近，$\hat{m}_4$ 与四阶中心矩 $E(X-\mu)^4$ 很接近. 换句话说，如果 X 服从正态分布，而 n 又很大，则 $\hat{\gamma}$ 接近于 0，$\hat{\delta}$ 接近于 3. 所谓偏度—峰度检验法就是这样的：

当 $\hat{\gamma}$ 的数值不接近于 0，或者 $\hat{\delta}$ 的数值不接近于 3，则认为原总体不服从正态分布；反之（即 $\hat{\gamma}\approx0$ 且 $\hat{\delta}\approx3$），则认为原总体服从正态分布.

$14.05, b = 16.15, m = 6. t_1 = 14.35, t_2 = 14.65, t_3 = 14.95,$ $t_4 = 15.25, t_5 = 15.55, t_6 = 15.85.$

实数轴$(-\infty, +\infty)$被这些 t_i 分成了七段. 当 H_0 成立时,我们来计算 p_i 的值.

用 $F(x)$表示 $N(15.1, (0.432\ 5)^2)$的分布函数,则

$$p_1 = F(t_1)$$
$$p_2 = F(t_2) - F(t_1)$$
$$p_3 = F(t_3) - F(t_2)$$
$$p_4 = F(t_4) - F(t_3)$$
$$p_5 = F(t_5) - F(t_4)$$
$$p_6 = F(t_6) - F(t_5)$$
$$p_7 = 1 - F(t_6)$$

为了计算 $F(t_i)$的值,可利用标准正态分布函数 $\Phi(x)$的表(见附表 1). 因为

$$F(t_i) = \Phi\left(\frac{t_i - 15.1}{0.432\ 5}\right)$$

于是
$$F(t_1) = \Phi\left(\frac{14.35 - 15.1}{0.432\ 5}\right) = \Phi(-1.734\ 1)$$
$$= 1 - \Phi(1.734\ 1)$$
$$F(t_2) = \Phi\left(\frac{14.65 - 15.1}{0.432\ 5}\right) = \Phi(-1.040\ 5)$$
$$= 1 - \Phi(1.040\ 5)$$
$$F(t_3) = \Phi(-0.346\ 8) = 1 - \Phi(0.346\ 8)$$
$$F(t_4) = \Phi(0.346\ 8)$$
$$F(t_5) = \Phi(1.040\ 5)$$
$$F(t_6) = \Phi(1.734\ 1)$$

查标准正态分布函数 $\Phi(x)$的表,知

$$\Phi(1.734\ 1) = 0.958\ 6$$

$$\Phi(1.040\ 5)=0.850\ 9$$

$$\Phi(0.346\ 8)=0.635\ 5$$

故 $F(t_1)=0.041\ 4,\quad F(t_2)=0.149\ 1$

$F(t_3)=0.364\ 5,\quad F(t_4)=0.635\ 5$

$F(t_5)=0.850\ 9,\quad F(t_6)=0.958\ 6$

得 $p_1=0.041\ 4, p_2=0.107\ 7, p_3=0.215\ 4$

$p_4=0.271\ 0, p_5=0.215\ 4, p_6=0.107\ 7$

$p_7=0.041\ 4$

现在来计算统计量 V.为便于检查,列表如下.于是

i	1	2	3	4	5	6	7
p_i	0.041 4	0.107 7	0.215 4	0.271 0	0.215 4	0.107 7	0.041 4
np_i	2.070	5.385	10.770	13.550	10.770	5.385	2.070
ν_i	3	5	10	16	8	6	2
$(np_i-\nu_i)^2$	0.864 9	0.148 2	0.592 5	6.002 5	7.672 9	0.378 2	0.004 9
$\frac{(np_i-\nu_i)^2}{np_i}$	0.417 8	0.027 5	0.055 1	0.443 0	0.712 4	0.070 2	0.002 4

$$V=\sum_{i=1}^{7}\frac{(np_i-\nu_i)^2}{np_i}$$

$$=0.417\ 8+0.027\ 5+\cdots+0.002\ 4=1.728\ 4$$

取 $\alpha=0.05$,查 χ^2 分布表(自由度是 4),得临界值 $\lambda=9.49$.这里的自由度为什么不是 6 呢? 这是因为,要检验的假设"H_0:总体服从 $N(\mu,\sigma^2)$"中,μ,σ^2 是用该组样本的样本平均数 $\bar{x}$ 与样本方差 s^2 来代替的,这就要扣去 2 个自由度(严密的数学论证从略).

现在 $V=1.728\ 4<9.49$,故下结论:假设 H_0 是相容的.因此认为滚珠直径基本上是服从正态分布 $N(15.1,(0.432\ 5)^2)$的.这就解决了我们的问题.

上述检验法通称为分布函数的 χ^2 检验法.它的好处在于,不管事先给出的 $F(x)$是怎样的分布函数,都可以检验一个总体是否以它为分布函数.因而,它的应用较广.不过,对于连续型随机变量的样本而言,计算较麻烦,这从上面的例子也已看出.但是,χ^2 检验法对于离散型情形,使用起来还是很方便的.

假设 X 的分布是

$$P\{X=a_i\}=p_i \quad (i=1,2,\cdots,m+1)$$

$x_1,x_2,\cdots,x_n$是样本值.我们还是取统计量

$$V=\sum_{i=1}^{m+1}\frac{(\nu_i-np_i)^2}{np_i} \tag{6.1}$$

这里的 ν_i是 n 个样品中,a_i出现的频数($i=1,2,\cdots,m+1$).V 还是近似服从 m 个自由度的 χ^2 分布.下面举一个例子.

例 6.2 某工厂近五年来发生了六十三次事故,按星期几分类如下:

星　期	一	二	三	四	五	六
次　数	9	10	11	8	13	12

问:事故是否与星期几有关?(参看例 1.3)

解 用 X 表示这样的随机变量:若事故发生在星期 i,则 $X=i$.显然 X 的可能值是 1,2,3,4,5,6(星期日停工休息).

我们来检验假设 $H_0:P(X=i)=\frac{1}{6}(i=1,\cdots,6)$(这个假设的含义是出事故与星期几无关).

使用统计量(6.1),现在 $m=5$,$p_i=P\{X=i\}$.如果 H_0成立,则这个统计量 $V=\sum_{i=1}^{6}\left(\nu_i-\frac{n}{6}\right)^2\Big/\frac{n}{6}$,且近似服从 5 个自由度的 χ^2 分布.查附表 3 知

$$P\{V>11.07\}=0.05$$

现在 $\nu_1=9,\nu_2=10,\cdots,\nu_6=12$.算得 V 的值为 1.67,它比临界值 11.07 小.故假设 H_0是相容的,即不能认为出事故与星期几

有关.

习 题 十 七

1. 由经验知某零件重量 $X \sim N(\mu, \sigma^2)$, $\mu = 15$, $\sigma^2 = 0.05$. 技术革新后, 抽了六个样品, 测得重量为(单位:g):

14.7, 15.1, 14.8, 15.0, 15.2, 14.6

已知方差不变, 问平均重量是否仍为 15? ($\alpha = 0.05$)

2. 糖厂用自动打包机打包. 每包标准重量为 100kg. 每天开工后需要检验一次打包机工作是否正常. 即检查打包机是否有系统偏差. 某日开工后测得几包重量(单位:kg)如下:

99.3,98.7,100.5,101.2,98.3,99.7,99.5,102.1,100.5

问:该日打包机工作是否正常? ($\alpha = 0.05$;已知包重服从正态分布.)

3. 正常人的脉搏平均为 72min^{-1}, 现某医生测得 10 例慢性四乙基铅中毒患者的脉搏(单位:min^{-1})如下:

54, 67, 68, 78, 70, 66, 67, 70, 65, 69

问:四乙基铅中毒者和正常人的脉搏有无显著性差异? (已知四乙基铅中毒者的脉搏服从正态分布.)

4. 用热敏电阻测温仪间接测量地热勘探井底温度, 重复测量 7 次, 测得温度(℃):

112.0, 113.4, 111.2, 112.0, 114.5, 112.9, 113.6

而用某精确办法测得温度为 112.6(可看作温度真值), 试问用热敏电阻测温仪间接测温有无系统偏差? ($\alpha = 0.05$)

5. 某种导线, 要求其电阻的标准差不得超过 0.005(Ω). 今在生产的一批导线中取样品 9 根, 测得 $S = 0.007(\Omega)$, 设总体为正态分布. 问在水平 $\alpha = 0.05$ 下能认为这批导线的标准差显著地偏大吗?

6. 机床厂某日从两台机器所加工的同一种零件中, 分别抽若干个样测量零件尺寸, 得:

第一台机器的:6.2,5.7,6.5,6.0,6.3,5.8,5.7,6.0,6.0,5.8,6.0

第二台机器的:5.6,5.9,5.6,5.7,5.8,6.0,5.5,5.7,5.5

问:这两台机器的加工精度是否有显著性差异? ($\alpha = 0.05$)

7. 检查了 26 匹马, 测得每 100mL 的血清中, 所含的无机磷平均为

3.29mL,标准差为0.34mL,又检查了18头羊,100mL的血清中含无机磷平均为3.96mL,标准差为0.40mL.试以0.05的检验水平,检验马与羊的血清中含无机磷的量是否有显著性差异?

8. 十个失眠患者,服用甲、乙两种安眠药,延长睡眠的时间如下表所示:

	a	b	c	d	e	f	g	h	i	j
甲	1.9	0.8	1.1	0.1	−0.1	4.4	5.5	1.6	4.6	3.4
乙	0.7	−1.6	−0.2	−1.2	−0.1	3.4	3.7	0.8	0	2.0

问这两种安眠药的疗效有无显著性差异?(可以认为服用两种安眠药后增加的睡眠时间之差近似服从正态分布.)($\alpha=0.05$)

9. 比较甲、乙两种安眠药的疗效.将20个患者分成两组,每组10人;甲组病人服用甲种安眠药,乙组病人服用乙种安眠药.如服药后延长的睡眠时间分别近似服从正态分布,其数据仍如上题(自然,数据不是两两成对了),问这两种安眠药的疗效有无显著性差异?($\alpha=0.05$)

10. 在一正20面体的20个面上,分别标以数字0,1,2,…,9,每个数字在两个面上标出.为检验其匀称性,共作800次投掷试验,数字0,1,2,…,9朝正上方的次数如下:

数字	0	1	2	3	4	5	6	7	8	9
频数	74	92	83	79	80	73	77	75	76	91

问:该正20面体是否匀称?

11. 某工厂采用新法处理废水,对处理后的水测量所含某种有毒物质的浓度,得到10个数据(单位:$10^{-6}g\cdot L^{-1}$):

22, 14, 17, 13, 21, 16, 15, 16, 19, 18

而以往用老法处理废水后,该种有毒物质的平均浓度为19.问:新法是否比老法效果好?(检验水平 $\alpha=0.05$)

第七章　回归分析方法

回归分析方法是数理统计中的一个常用方法，是处理多个变量之间**相关关系**的一种数学方法.

提到变量间的关系，很容易使人想起微积分课程中所讨论的**函数关系**，即所谓确定性的关系.比如，自由落体运动中，物体下落的距离 s 与所需的时间 t 之间，就有如下的函数关系：

$$s=\frac{1}{2}gt^2 \quad (0\leqslant t\leqslant T)$$

变量 s 的值随 t 的值而定，也就是说，如果取定了 t 的值，那么，s 的值就完全确定了.

但是，世界上众多的变量间，还有另一类重要关系，我们称之为**相关关系**.比如，人的身高与体重间的关系.虽然一个人的“身高”并不能确定“体重”，但是，总的说来，身高者，体也重.我们就说，身高与体重这两个变量间具有相关关系.又如，在冶炼某钢种过程中，钢液的初始含碳量与冶炼时间这两个变量间也具有相关关系.

实际上，即使是具有确定性关系的变量间，由于实验误差的影响，其表现形式也具有某种程度的不确定性.这一点大家在做物理实验时是有体会的.

回归分析方法是处理变量间相关关系的有力工具.它不仅提供了建立变量间关系的数学表达式——通常称为经验公式——的一般方法，而且利用概率统计基础知识进行了分析讨论，从而能帮助实际工作者如何去判明所建立的经验公式的有效性，以及如何利用所得到的经验公式去达到预测、控制等目的.因此，回归分析方法得到越来越广泛的应用，而方法本身也在不断丰富、发展.

本讲义重点讨论一元回归.对于多元回归只作简要的介绍.

§1 一元线性回归

1. 经验公式与最小二乘法

在一元线性回归分析里,我们要考察的是:随机变量 Y 与一个普通变量 x 之间的联系.

对于有一定联系的两个变量:x 与 Y,在观测或实验中得到若干对数据

$$(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n)$$

的基础上,用什么方法来获得这两个变量之间(Y 对 x)的经验公式呢? 为说明问题,先看一个例子.

例 1.1 某种合成纤维的强度与其拉伸倍数有关.下面的表是 24 个纤维样品的强度与相应的拉伸倍数的实测记录.我们希望由此具体找出这两个量的关系式.

用上面的语言来说,对于两个变量 x(拉伸倍数),Y(强度),我们实测到 24 对数据:

(1.9,1.4),(2.0,1.3),(2.1,1.8),…,(9.5,8.1),(10.0,8.1).在此基础上,来找出 x,Y 的关系式.

由解析几何知识,平面上选定一直角坐标系后,这 24 对数据就分别对应到平面上的 24 个点(见图 7.1).这张图称为**散点图**.

编号	拉伸倍数 x	强度 Y (kg/mm^2)	编号	拉伸倍数 x	强度 Y (kg/mm^2)
1	1.9	1.4	7	3.5	3.0
2	2.0	1.3	8	3.5	2.7
3	2.1	1.8	9	4.0	4.0
4	2.5	2.5	10	4.0	3.5
5	2.7	2.8	11	4.5	4.2
6	2.7	2.5	12	4.6	3.5

续表

编号	拉伸倍数 x	强度 Y (kg/mm^2)	编号	拉伸倍数 x	强度 Y (kg/mm^2)
13	5.0	5.5	19	8.0	6.5
14	5.2	5 0	20	8.0	7.0
15	6.0	5.5	21	8.9	8.5
16	6.3	6.4	22	9.0	8.0
17	6.5	6.0	23	9.5	8.1
18	7.1	5.3	24	10 0	8.1

它给我们很多启示.首先,这些点虽然是散乱的,但大体上散布在某条直线的周围.也就是说,拉伸倍数与强度之间大致成线性关系:

$$\hat{y} = a + bx \tag{1.1}$$

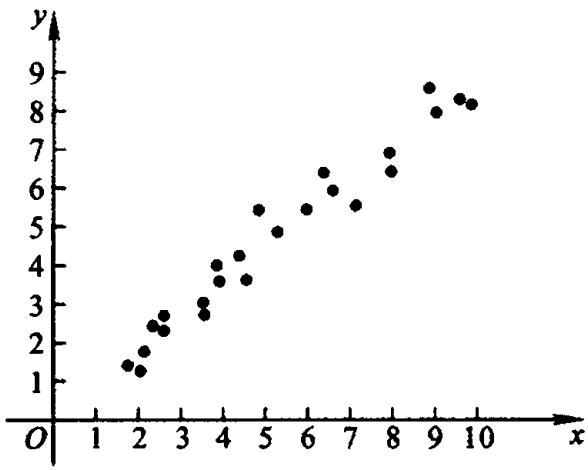

图 7.1

这里,在 y 上方加"^",是为了区别于 Y 的实际值 y.因为 Y 与 x 之间一般不具有函数关系.

至此,在散点图的启示下,经验公式的**形式已可以确定**,是所谓线性的.要完全找出经验公式,就只需确定(1.1)中的 a 和 b.这里 b 通常叫做回归系数,关系式 $\hat{y} = a + bx$ 叫做回归方程.

从散点图来看，要找出 a,b 是不困难的：在散点图上划这样一条直线，**使该直线总的来看最“接近”这 24 个点**；于是，这直线在 y 轴上的截距就是所求的 a，它的斜率就是所求的 b.

这个几何方法虽然简便，但太粗糙. 而且，对于非线性形式的问题以及多变量的问题，就几乎无法实行. 然而，它的基本思想，即“使该直线总的来看最接近这 24 个点”，却是很可取的. 问题是把这个基本思想精确化、数量化.

设给定 n 个点 $(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n)$. 那么，对于平面上任意一条直线 l：

$$y=a+bx$$

我们用数量

$$[y_t-(a+bx_t)]^2$$

来刻画点 (x_t,y_t) 到直线 l 的远近程度（读者运用解析几何知识，不难看出，$|y_t-(a+bx_t)|$ 的几何意义是点 (x_t,y_t) 沿着平行于 y 轴的方向到 l 的铅直距离，而不是沿着垂直于 l 的方向到 l 的最短距离）. 于是

$$\sum_{t=1}^{n}[y_t-(a+bx_t)]^2$$

就定量地描述了直线 l 跟这 n 个点的总的远近程度. 这个量是随不同的直线而变化的，或者说，是随不同的 a 与 b 而变化的，也就是说它是 a,b 的二元函数，记为 $Q(a,b)$：

$$Q(a,b)=\sum_{t=1}^{n}[y_t-(a+bx_t)]^2 \tag{1.2}$$

于是，要找一条直线，使该直线总的来看最“接近”这 n 个点的问题，就转化为如下的问题：

要找两个数 $\hat{a},\hat{b}$，使二元函数 $Q(a,b)$ 在 $a=\hat{a},b=\hat{b}$ 处达到最小.

由于 $Q(a,b)$ 是 n 个平方之和，所以“使 $Q(a,b)$ 最小”的原则称为平方和最小原则，习惯上称为**最小二乘原则**.

依照最小二乘原则，具体找 $\hat{a}, \hat{b}$ 的问题通常利用微积分学中的极值原理[①]，即解二元一次联立方程：

---

① 其实，用初等代数中的配方法就能圆满地解决问题. 记 $\bar{x}=\frac{1}{n}\sum_{t=1}^{n}x_t$，$\bar{y}=\frac{1}{n}\sum_{t=1}^{n}y_t$，则

$$
\begin{aligned}
Q(a,b) &= \sum_{t=1}^{n}\{(y_t-\bar{y})+[\bar{y}-(a+b\bar{x})]-b(x_t-\bar{x})\}^2 \\
&= \sum_{t=1}^{n}(y_t-\bar{y})^2+n[\bar{y}-(a+b\bar{x})]^2 \\
&\quad +b^2\sum_{t=1}^{n}(x_t-\bar{x})^2+2[\bar{y}-(a+b\bar{x})]\cdot\sum_{t=1}^{n}(y_t-\bar{y}) \\
&\quad -2b[\bar{y}-(a+b\bar{x})]\sum_{t=1}^{n}(x_t-\bar{x}) \\
&\quad -2b\sum_{t=1}^{n}(x_t-\bar{x})(y_t-\bar{y}) \\
&= \sum_{t=1}^{n}(y_t-\bar{y})^2+n[\bar{y}-(a+b\bar{x})]^2 \\
&\quad +b^2\sum_{t=1}^{n}(x_t-\bar{x})^2-2b\sum_{t=1}^{n}(x_t-\bar{x})(y_t-\bar{y}) \\
&= \sum_{t=1}^{n}(y_t-\bar{y})^2+n[\bar{y}-(a+b\bar{x})]^2 \\
&\quad +\sum_{t=1}^{n}(x_t-\bar{x})^2\left[b^2-2b\frac{\sum_{t=1}^{n}(x_t-\bar{x})(y_t-\bar{y})}{\sum_{t=1}^{n}(x_t-\bar{x})^2}\right] \\
&= \sum_{t=1}^{n}(y_t-\bar{y})^2+n[\bar{y}-(a+b\bar{x})]^2 \\
&\quad +\sum_{t=1}^{n}(x_t-\bar{x})^2\left[b-\frac{\sum_{t=1}^{n}(x_t-\bar{x})(y_t-\bar{y})}{\sum_{t=1}^{n}(x_t-\bar{x})^2}\right]^2 \\
&\quad -\frac{\left[\sum_{t=1}^{n}(x_t-\bar{x})(y_t-\bar{y})\right]^2}{\sum_{t=1}^{n}(x_t-\bar{x})^2}
\end{aligned}
$$

$$\begin{cases} \dfrac{\partial Q}{\partial a} = -2\sum_{t=1}^{n}[y_t - (a + bx_t)] = 0 & (1.3) \\ \dfrac{\partial Q}{\partial b} = -2\sum_{t=1}^{n}[y_t - (a + bx_t)] \cdot x_t = 0 & (1.4) \end{cases}$$

从(1.3)式可得

$$na = \sum_{t=1}^{n} y_t - b\sum_{t=1}^{n} x_t$$

故

$$a = \bar{y} - b\bar{x} \qquad (1.5)$$

其中 $\bar{y}, \bar{x}$ 分别是 y_t 和 x_t 的平均数. 从(1.4)式可得

$$\sum_{t=1}^{n} x_t y_t - a\sum_{t=1}^{n} x_t - b\sum_{t=1}^{n} x_t^2 = 0$$

利用(1.5)式,可由上式解得 b:

$$b = \frac{\sum_{t=1}^{n} x_t y_t - n\bar{x}\bar{y}}{\sum_{t=1}^{n} x_t^2 - n\bar{x}^2} = \frac{\sum_{t=1}^{n}(x_t - \bar{x})(y_t - \bar{y})}{\sum_{t=1}^{n}(x_t - \bar{x})^2} \qquad (1.6)$$

数学上可以证明,用(1.6)及(1.5)确定的 a, b 确实使平方和达到最小.

(接上页注)

由上式不难看出,当且仅当:

$$b = \hat{b} \triangleq \frac{\sum_{t=1}^{n}(x_t - \bar{x})(y_t - \bar{y})}{\sum_{t=1}^{n}(x_t - \bar{x})^2} \text{(当 } x_1, x_2, \cdots, x_n \text{ 不全相等)}$$

$$a = \hat{a} \triangleq \bar{y} - \hat{b}\bar{x}$$

时,$Q(a,b)$ 达最小值:

$$Q(\hat{a},\hat{b}) = \sum_{t=1}^{n}(y_t - \bar{y})^2 - \frac{\left[\sum_{t=1}^{n}(x_t - \bar{x})(y_t - \bar{y})\right]^2}{\sum_{t=1}^{n}(x_t - \bar{x})^2}$$

(当 $x_1, x_2, \cdots, x_n$ 全相等时(这在实际工作中一般不会出现),$Q(a,b)$的最小值点是 (a^*, b^*),其中 $a^* = \bar{y} - b^*\bar{x}$,$b^*$ 为任何实数.)

于是,对于给定的 n 个点$(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n)$,先按(1.6)式算出 b,再由(1.5)式算出 a,就得到了所要找的直线:

$$y=a+bx$$

(由(1.5)式不难看出,点$(\bar{x},\bar{y})$在该直线上.)也就找到了 x,Y 之间的经验公式:

$$\hat{y}=a+bx$$

对于例 1.1 的 24 个点,由(1.6),(1.5)算得

$$b=0.859,a=0.15$$

因此,得强度(Y)与拉伸倍数(x)间的经验公式:

$$\hat{y}=0.15+0.859x$$

与回归方程相应的直线称回归直线;这里回归系数 b 等于 0.859,它的意义是:拉伸倍数(x)每增加一个单位(即一倍),强度(Y)平均增加 0.859 个单位($kg\cdot mm^{-2}$).

对于经验公式的类型是线性的情况下,从上面的讨论知道,可直接用公式(1.5),(1.6)求得 a,b.然而,大量的实际问题并不属于线性的类型,怎么办呢?一个常用而简便的方法是尽可能把它们变为线性的类型.下面看两个例子.

例 1.2 在彩色显影中,根据以往的经验,形成染料光学密度 Y 与析出银的光学密度 x 之间有下面类型的关系:

$$Y\approx Ae^{B/x},B>0$$

我们希望通过一组实验数据求出未知参数 A 与 B.

虽然 Y,x 之间的关系不是线性的,但对上面的等式两边取自然对数后便得:

$$\ln Y\approx\ln A-\frac{B}{x}$$

令

$$Y^*=\ln Y$$

$$x^*=\frac{1}{x}$$

则两个新变量 x^*,Y^* 之间的关系便近似是线性的了:

$$Y^*\approx\ln A-Bx^*$$

这样，从 n 组数据 $(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n)$ 出发，按 $x_t^*=\frac{1}{x_t}$，$y_t^*=\ln y_t(t=1,2,\cdots,n)$，得 n 组新数据 $(x_1^*,y_1^*),(x_2^*,y_2^*),\cdots,(x_n^*,y_n^*)$，再用(1.5)，(1.6)得 a^*，b^*，最后，由 $\ln A=a^*$，$-B=b^*$ 就得 A，B 了.

例 1.3 炼钢厂出钢时所用的盛钢水的钢包，在使用过程中由于钢液及炉渣对包衬耐火材料的浸蚀，使其容积不断增大. 经过试验，钢包的容积(由于容积不便测量，故以钢包盛满时的钢水重量来表示)与相应的使用次数(也称包龄)的数据如下表所示. 我们希望找出它们之间的定量关系式.

使用次数(x)	容积(Y)	使用次数(x)	容积(Y)
2	106.42	11	110.59
3	108.20	14	110.60
4	109.58	15	110.90
5	109.50	16	110.76
7	110.00	18	111.00
8	109.93	19	111.20
10	110.49		

经验公式的类型是什么呢？还按例 1.1 的办法，先作散点图，(见图 7.2). 从图中看出，最初容积增加很快，以后逐渐减慢趋于稳定. 根据这个特点，我们选用双曲线

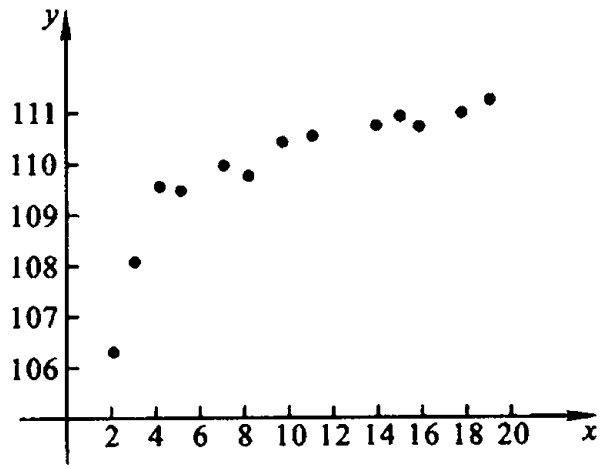

图 7.2

$$\frac{1}{y}=a+b\frac{1}{x}$$

来近似表示容积 Y 与使用次数 x 之间的关系.

显然,x,Y 间的关系不是线性的;但是,新变量 $x^*=\frac{1}{x}$,$Y^*=\frac{1}{Y}$ 之间的关系却是近似线性的.于是,对 13 组新数据 $(x_1^*,y_1^*),(x_2^*,y_2^*),\cdots,(x_{13}^*,y_{13}^*)$,用(1.5),(1.6)得 a^*,b^*,就找出了 x,Y 间的经验公式.

2. 平方和分解公式与线性相关关系

有了经验公式

$$\hat{y}=\hat{a}+\hat{b}x$$

(其中 $\hat{a}$,$\hat{b}$ 由公式(1.5),(1.6)确定)之后,是否就可用它来进行预报和控制呢? 要注意的是,我们从任意一组数据 $(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n)$ 出发,按公式(1.5),(1.6)都可建立起上述经验公式. Y 与 x 是否真的有近似的线性关系? 这还没有判明.因此,首先需要判别 x 与 Y 间是否具有线性相关关系.注意,所谓"线性相关关系"是指,Y 是否基本上随着 x 的增大而线性地增大(或线性地减小).

下面我们先来导出一个具有统计意义的分解公式.

平方和分解公式 对于任意 n 组数据 $(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n)$,恒有:

$$\sum_{t=1}^{n}(y_t-\bar{y})^2=\sum_{t=1}^{n}(y_t-\hat{y}_t)^2+\sum_{t=1}^{n}(\hat{y}_t-\bar{y})^2 \qquad (1.7)$$

其中 $$\hat{y}_t=\hat{a}+\hat{b}x_t,\quad (t=1,2,\cdots,n)$$

证 $$\begin{aligned}\Sigma(y_t-\bar{y})^2&=\Sigma[(y_t-\hat{y}_t)+(\hat{y}_t-\bar{y})]^2\text{①}\\&=\Sigma[(y_t-\hat{y}_t)^2+2(y_t-\hat{y}_t)(\hat{y}_t-\bar{y})\\&\qquad+(y_t-\bar{y})^2]\end{aligned}$$

---

① 为书写方便起见,把"$\sum_{t=1}^{n}$"简化为"Σ",下同.

但 $\sum(y_t-\hat{y}_t)(\hat{y}_t-\bar{y})=\sum[y_t-(\hat{a}+\hat{b}x_t)][\hat{a}+\hat{b}x_t-\bar{y}]$

$$\xrightarrow{\text{将 } \hat{a}=\bar{y}-\hat{b}\bar{x} \text{ 代入}}\sum[(y_t-\bar{y})-b(x_t-\bar{x})][\hat{b}(x_t-\bar{x})]$$

$$=\hat{b}[\sum(y_t-\bar{y})(x_t-\bar{x})-\hat{b}(x_t-\bar{x})^2]=0$$

这就证明了(1.7).为了说明(1.7)式的统计意义,我们先对该式中的三个平方和作下列说明.

$\sum(y_t-\bar{y})^2$ 是 $y_1,y_2,\cdots,y_n$ 这 n 个数据的偏差平方和,它的大小描述了这 n 个数据的分散程度,记作 l_{yy}.

为要了解右边的两个平方和,先来熟悉 $\hat{y}_t$.注意,$\hat{y}_t=\hat{a}+\hat{b}x_t$.由此可知,它的几何意义是:回归直线 $y=\hat{a}+\hat{b}x$ 上,其横坐标为 x_t 的点的纵坐标(见图 7.3).再注意一个事实,就是 n 个数 $\hat{y}_1,\hat{y}_2,\cdots,\hat{y}_n$ 的平均数也是 $\bar{y}$.这是因为

$$\frac{1}{n}\sum\hat{y}_t=\frac{1}{n}\sum(\hat{a}+\hat{b}x_t)$$

$$=\hat{a}+\hat{b}\frac{1}{n}\sum x_t$$

$$=\hat{a}+\hat{b}\bar{x}=\bar{y}$$

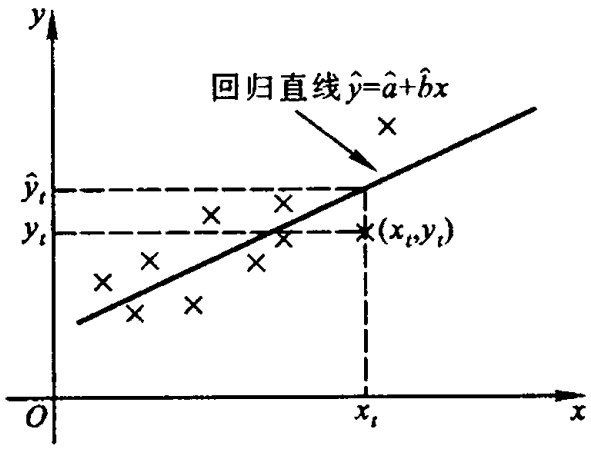

图 7.3

于是,$\sum(\hat{y}_t-\bar{y})^2$ 就是 $\hat{y}_1,\hat{y}_2,\cdots,\hat{y}_n$ 这 n 个数的偏差平方和,记作 U,它描述了 $\hat{y}_1,\hat{y}_2,\cdots,\hat{y}_n$ 的分散程度.是什么原因引起了

$\hat{y}_1,\hat{y}_2,\cdots,\hat{y}_n$的分散呢？上面已说过，$\hat{y}_t$是回归直线上的点的纵坐标，相应的横坐标是 x_t. 因此我们说，$\hat{y}_1,\hat{y}_2,\cdots,\hat{y}_n$的分散性来源于$x_1,x_2,\cdots,x_n$的分散性，而且是通过 x 对于 Y 的线性相关关系引起的. 实际上，下面的推演把问题就说得更清楚了：

$$\begin{aligned} U &= \sum(\hat{y}_t-\bar{y})^2=\sum[\hat{a}+\hat{b}x_t-(\hat{a}+\hat{b}\bar{x})]^2 \\ &= \sum\hat{b}^2(x_t-\bar{x})^2 \\ &= \hat{b}^2\cdot\sum(x_t-\bar{x})^2 \end{aligned} \tag{1.8}$$

[$\sum(x_t-\bar{x})^2$ 是 $x_1,x_2,\cdots,x_n$的偏差平方和，记作 l_{xx}，它描述了 $x_1,x_2,\cdots,x_n$的分散程度.]我们称 U 为**回归平方和**.

至于$\sum(y_t-\hat{y}_t)^2$，它就是$\sum[y_t-(\hat{a}+\hat{b}x_t)]^2$. 这在讲最小二乘原则时见到过，它也就是 $Q(a,b)$的最小值，就记作 Q. 我们称 Q 为**残差平方和**.（Q 是除了 x 对 Y 的线性影响之外的剩余因素对 $y_1,y_2,\cdots,y_n$分散性的作用，这剩余因素中包括 x 对 Y 的非线性影响及试验误差等. 因此，我们又称 Q 为**剩余平方和**，它是仅考虑 x 与 Y 的线性关系所不能减少的部分.）

有了以上对于 l_{yy}，U，Q 的分析讨论，(1.7)式的具体含义就十分清楚了，那就是 $y_1,y_2,\cdots,y_n$的分散程度（即 l_{yy}）可以分解为两部分：

$$l_{yy}=Q+U \tag{1.7′}$$

其中一部分是（来源于 $x_1,x_2,\cdots,x_n$的分散性）通过 x 对于 Y 的线性相关关系而引起的 Y 的分散性（即回归平方和 U），另一部分是剩余部分引起的 Y 的分散性（即剩余平方和 Q）.

现在我们回到本段开头提出的问题上来，回答 x，Y 间是否存在线性相关关系的问题. 一个很自然的想法是把回归平方和 U（线性影响）跟剩余平方和 Q（其他影响）进行比较.

数理统计学中，选取量

$$F\triangleq\frac{U}{Q/(n-2)} \tag{1.9}$$

来体现 x 与 Y 的线性相关关系的相对大小.

如果 F 值相当大,则表明 x 对 Y 的线性影响较大,就可以认为 x 与 Y 间有线性相关关系;反之,若 F 的值较小,则没有理由认为 x 与 Y 间有线性相关关系.

3. 数学模型与相关性检验

F 值究竟多大,才认为 x 与 Y 间有线性相关关系呢? 为了给出定量界限,也为进一步讨论其他有关问题的需要,我们对数据的结构提出下列假定:

$$\begin{aligned} Y_1 &= a + bx_1 + \varepsilon_1 \\ Y_2 &= a + bx_2 + \varepsilon_2 \\ &\cdots\cdots \\ Y_n &= a + bx_n + \varepsilon_n \end{aligned} \tag{1.10}$$

其中 $\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_n$ 是随机变量,它们相互独立,且都服从相同的正态分布 $N(0, \sigma^2)$(σ 未知).

以上假定,是我们深入讨论问题的基本假定,也是回归分析中作种种统计推断的出发点.对于(1.10)那样的表示方式,我们似乎很陌生,其实不然.比如在第六章中,我们常常说"设 $Y_1, Y_2, \cdots, Y_n$[①]是来自总体 $N(\mu, \sigma^2)$ 的样本"(它实际上是关于一个正态总体统计推断的出发点),由样本的概念(以及命题:$Y \sim N(\mu, \sigma^2) \Longleftrightarrow Y - \mu \sim N(0, \sigma^2)$),这句话就等价于:

$$\begin{aligned} Y_1 &= \mu + \varepsilon_1 \\ Y_2 &= \mu + \varepsilon_2 \\ &\cdots\cdots \\ Y_n &= \mu + \varepsilon_n \end{aligned} \tag{1.10'}$$

其中 $\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_n$ 是随机变量,相互独立,服从正态分布 $N(0, \sigma^2)$.(1.10′)和(1.10)就很相像.

---

① 为便于比较,这里不用 X_i 而用 Y_i 表示.

在(1.10)的假定下,为了判明 x 与 Y 间是否存在线性相关关系,就转化为检验下列假设 H_0:

$$H_0: b=0$$

如果由一组具体的样本值$(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$,否定了 H_0,也即判定 $b\neq 0$;联系到基本假定(1.10),判定 $b\neq 0$,也即判定 x 与 Y 间有线性相关关系.那么,什么情形下,否定 H_0 呢?

数学上可以证明,在假设 H_0 成立时,由(1.9)式提供的统计量 F 服从自由度为 $1, n-2$ 的 F 分布①.(而且,b 离开 0 越远,即 b 的绝对值越大,F 总的说来越大.)

这样,我们就得到关于相关性检验的一般程序:

(1) 计算 U, Q,从而按(1.9)得 F 值;

(2) 对于给定的检验标准 α,查自由度为 $1, n-2$ 的 F 分布的临界值表,得临界值 λ:

$$P(F>\lambda)=\alpha$$

(3) 比较(算得的)F 值与(查得的)λ 值的大小,如 $F>\lambda$,则否定假设"$H_0: b=0$",即认为 x, Y 间具有线性相关关系;否则,假设 H_0 是相容的,即没有理由认为 x, Y 间存在线性相关关系.

下面,对于上述方法作几点补充说明.

① 数学上可以证明 $\frac{Q}{\sigma^2}$ 服从 $n-2$ 个自由度的 χ^2 分布,从而 $E\left(\frac{Q}{\sigma^2}\right)=n-2$,于是 $E\left(\frac{Q}{n-2}\right)=\sigma^2$,这表明统计量 $\frac{Q}{n-2}$ 是随机项 ε 的方差的无偏估计,记它为 $\hat{\sigma}^2$ 或 s^2.

② 为了检验相关性,有的书上是通过统计量(相关系数)

---

① 在假定(1.10)之下,如果 $b=0$,则数学上可以证明:$\frac{U}{\sigma^2}$ 服从自由度为 1 的 χ^2 分布,$\frac{Q}{\sigma^2}$ 服从自由度为 $n-2$ 的 χ^2 分布,而且 Q 与 U 相互独立.由此推出统计量(1.9)服从自由度为 $1, n-2$ 的 F 分布(参见附录二的定理 8).

$$R=\frac{\sum(x_t-\bar{x})(y_t-\bar{y})}{\sqrt{\sum(x_t-\bar{x})^2\cdot\sum(y_t-\bar{y})^2}}$$

进行的.当$|R|$较大时,否定假设"$b=0$".

由(1.8)式不难验证,有

$$U=l_{yy}R^2$$

因此有

$$Q=l_{yy}(1-R^2)$$

上式表明

(i) $|R|\leqslant 1$;

(ii) 当 l_{yy} 固定时,$|R|$ 越接近 1,Q 就越小.特别地,$|R|=1$ 时,$Q=0$,即 n 个点在一条直线上;而 $R=0$ 时,$Q=l_{yy}$.

对于假设 H_0,由 F 和 R 提供的两种形式上不同的检验方法,实质上是一回事.这是因为

$$F=(n-2)\frac{U}{Q}=(n-2)\frac{l_{yy}R^2}{l_{yy}(1-R^2)}$$

$$=(n-2)\frac{R^2}{1-R^2}$$

因此,本书没有提供关于 R 的临界值表.如有需要,可按上式由 F 临界值表换算,或从参考书目[5]、[9]中查找.

③ 计算 U,Q 的公式.

直接按 $U=\sum(\hat{y}_t-\bar{y})^2$ 与 $Q=\sum(y_t-\hat{y}_t)^2$ 来计算 U,Q 是比较麻烦的.注意到(1.8),有

$$U=\hat{b}^2\sum(x_t-\bar{x})^2$$

再用 $\hat{b}=\dfrac{\sum(x_t-\bar{x})(y_t-\bar{y})}{\sum(x_t-\bar{x})^2}$ 代入上式,得

$$U=\hat{b}l_{xy} \tag{1.11}$$

其中 $l_{xy}=\sum(x_t-\bar{x})(y_t-\bar{y})$.

然后,由(1.7′)式,有

$$Q=l_{yy}-U \tag{1.12}$$

这(1.11)和(1.12)就是通常用来计算 U,Q 的计算公式.

例 1.4 炼钢基本上是个氧化脱碳的过程，钢液原来的含碳量的多少直接影响到冶炼时间的长短. 下表是某平炉 34 炉的熔毕碳(即全部炉料熔化完毕时钢液的含碳量)与精炼时间(从熔毕至出钢，冶炼所需的时间)的生产记录.

解 (1) 作散点图如下，从图看出，可直接用线性回归试一试.

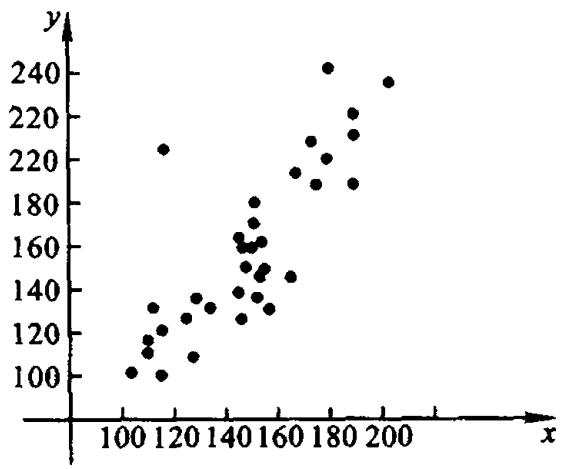

图 7.4

编号	熔毕碳 x (0.01%)	精炼时间 Y (min)	编号	熔毕碳 x (0.01%)	精炼时间 Y (min)
1	180	200	18	116	100
2	104	100	19	123	110
3	134	135	20	151	180
4	141	125	21	110	130
5	204	235	22	108	110
6	150	170	23	158	130
7	121	125	24	107	115
8	151	135	25	180	240
9	147	155	26	127	135
10	145	165	27	115	120
11	141	135	28	191	205
12	144	160	29	190	220
13	190	190	30	153	145
14	190	210	31	155	160
15	161	145	32	177	185
16	165	195	33	177	205
17	154	150	34	143	160

(2) 先算 $\bar{x},\bar{y},l_{xx},l_{yy},l_{xy}$.

$$\bar{x}=150.09,\quad \bar{y}=158.23$$

$$l_{xx}=\sum(x_i-\bar{x})^2=\sum x_i^2-\frac{1}{n}(\sum x_i)^2=25\ 462.7$$

$$l_{yy}=\sum(y_i-\bar{y})^2=\sum y_i^2-\frac{1}{n}(\sum y_i)^2=50\ 094.0$$

$$l_{xy}=\sum(x_i-\bar{x})(y_i-\bar{y})=\sum x_iy_i-\frac{1}{n}(\sum x_i)(\sum y_i)$$

$$=32\ 325.3$$

(3) 再算 $\hat{b},\hat{a};U,Q,s^2,F$.

$$\hat{b}=\frac{l_{xy}}{l_{xx}}=\frac{32\ 325.3}{25\ 462.7}=1.27$$

$$\hat{a}=\bar{y}-\hat{b}\,\bar{x}=158.23-1.27\times150.09=-32.38$$

所以,回归直线方程

$$\hat{y}=-32.38+1.27x$$

另外

$$U=\hat{b}\cdot l_{xy}=1.27\times32\ 325=41\ 053$$

$$Q=l_{yy}-U=9\ 041$$

$$s^2=Q/32=282.53,s=16.81$$

$$F=\frac{U}{Q/32}=\frac{41\ 053}{282.53}=145.3$$

(4) 相关性检验.

查自由度为 1,32 的 F 分布表得临界值:

$$\lambda=4.15(\alpha=0.05)$$

现在 $F=145.3>4.15=\lambda$,所以否定假设"$H_0:b=0$",即认为 x,Y 间存在线性相关关系;习惯上说,直线回归是显著的.(实际上,这里的 F 值还大于相应于 $\alpha=0.01$ 的临界值 $\lambda'=7.50$;此时我们称直线回归是高度显著的.)

4. 预报与控制

在第一小节末尾,我们提到过回归系数 b 的意义,就是:x 每增

加一个单位，Y 平均增加 b 个单位(当 $b<0$ 时，实际上是减少 $-b$ 个单位).这对具体工作是有一定的指导意义的.现在有了 2,3 小节的基础，我们还可以进一步来讨论预报与控制的问题.

我们还是讨论这样的情况：

$$Y=a+bx+\varepsilon$$

其中 ε 是所谓随机项，$\varepsilon\sim N(0,\sigma^2)$；所谓预报问题，就是问：$x=x_0$ 时 $Y=?$ 上面讲了从数据$(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n)$出发，利用最小二乘原则可得 a，b 的估计值 $\hat{a}$，$\hat{b}$ 及回归方程 $\hat{y}=\hat{a}+\hat{b}x$.很自然想到用

$$\hat{y}_0=\hat{a}+\hat{b}x_0$$

来预报 $Y_0=a+bx_0+\varepsilon_0$ 的值，然而实际问题还需要知道所谓预报精度.正如同我们并不满足于参数的点估计，还要求给出参数的区间估计一样.更何况这里 Y_0 是一个随机变量.数学上可以证明，只要 $\varepsilon_0,\varepsilon_1,\cdots,\varepsilon_n$($\varepsilon_1,\varepsilon_2,\cdots,\varepsilon_n$见(1.10)式)相互独立，且都服从 $N(0,\sigma^2)$，则随机变量

$$t\triangleq\frac{Y_0-\hat{y}_0}{s\sqrt{1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{l_{xx}}}}$$

服从 $n-2$ 个自由度的 t 分布.

这样，对给定的置信度 $1-\alpha$，查 $n-2$ 个自由度的 t 分布临界值表得 λ，就有

$$P\left\{\left|\frac{Y_0-\hat{y}_0}{s\sqrt{1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{l_{xx}}}}\right|\leqslant\lambda\right\}=1-\alpha \tag{1.13}$$

这里 $s=\sqrt{\frac{Q}{n-2}}$，由此得 Y_0 的置信度为 $1-\alpha$ 的置信区间：

$$\left[\hat{y}_0-\lambda s\sqrt{1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{l_{xx}}},\right.$$

$$\hat{y}_0 + \lambda s\sqrt{1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{l_{xx}}}\Big] \tag{1.14}$$

该区间以 $\hat{y}_0$ 为中点，长度为 $2\lambda s\sqrt{1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{l_{xx}}}$. 中点 $\hat{y}_0$ 随 x_0 线性地变化；其长度在 $x_0=\bar{x}$ 处最短，x_0 越远离 $\bar{x}$，长度就越长. 因此置信区间的上限与下限的曲线对称地落在回归直线两侧，而呈喇叭形(见图 7.5).

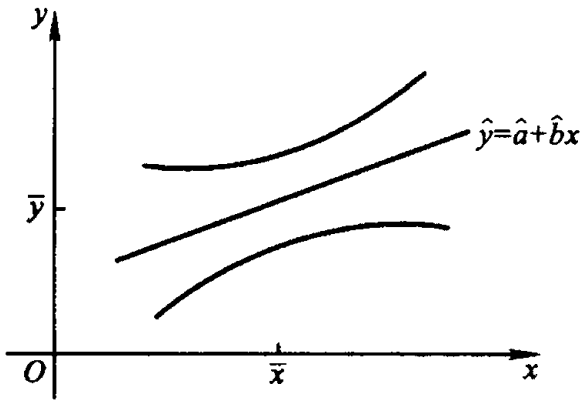

图 7.5

作为(1.14)的简化，当 n 较大，且 x_0 较接近 $\bar{x}$ 时，

$$\sqrt{1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{l_{xx}}}\approx 1$$

因此(1.14)就近似于

$$[\hat{y}_0-\lambda s,\hat{y}_0+\lambda s] \tag{1.14′}$$

又因为 n 较大时，自由度为 $n-2$ 的 t 分布接近 $N(0,1)$，所以这里的 λ，也可查正态分布表来得到. 比如，对 $\alpha=0.05$ 有 $\lambda=1.96$.

置信区间的长度直接关系到预报效果. 而我们从(1.14)，(1.14′)看到，置信区间的长度主要地被 s 的大小所决定. 因此，在预报问题中，s 是一个基本而重要的量.

例 1.5 本例是例 1.4 的继续，来讨论精炼时间的预报问题. 现测得某炉熔毕碳为 145(即 1.45%)，试估计该炉所需的精炼时间(置

信度95%).

解 这只需将(1.14)具体化即可.

$$\hat{y}_0 = \hat{a} + \hat{b}x_0 = -32.38 + 1.27 \times 145 = 151.77$$

$$s = 16.81$$

$$\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{l_{xx}}} = \sqrt{1 + \frac{1}{34} + \frac{(145 - 150.09)^2}{25\,462.7}}$$

$$= 1.015$$

查 t 分布表得

$$\lambda = 2.037$$

于是

$$\lambda s \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{l_{xx}}} = 34.76$$

得置信区间

$$[151.77 - 34.76, 151.77 + 34.76] = [117.01, 186.53]$$

如用(1.14′),置信区间是

$$[151.77 - 2.037 \times 16.81, 151.77 + 2.037 \times 16.81]$$

$$= [151.77 - 34.24, 151.77 + 34.24]$$

$$= [117.53, 186.01]$$

再如用 $\lambda = 1.96$ 代入,得置信区间是[118.82,184.72].这两个近似跟由(1.14)所得的相差无几,特别是第一个近似区间.

至于控制问题,实际上是预报问题的反问题.具体来讲,就是给出了对于 y_0 的要求,反过去找满足这种要求的相应的 x_0 的范围.解决办法是,由(1.14)式出发,将| |内的不等式按 x_0 变形,即由该不等式得一与之等价的关于 x_0 的不等式,这就最终得到 x_0 所在的范围.限于篇幅,我们就不细述了.

残差分析

利用统计量 F(见(1.9))进行线性相关性检验时,我们假定了数据的结构满足(1.10),其中随机项 $\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_n$ 独立同分布,共同分布是正态分布 $N(0,$

σ^2)(σ 未知).一个重要问题是:有了数据$(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n)$,如何判别这个假定是否成立呢?这是一个比较复杂的问题,属于回归诊断的范围.这里仅介绍一种简单易行的粗略办法.

设 $\hat{a},\hat{b}$是参数 a,b 的最小二乘估计,令 $\hat{y}_i=\hat{a}+\hat{b}x_i,\hat{\varepsilon}_i=y_i-\hat{y}_i(i=1,2,\cdots,n)$,这个 $\hat{\varepsilon}_i$叫做第 i 个残差.令

$$h_i=\frac{1}{n}+\frac{(x_i-\bar{x})^2}{l_{xx}}$$

$$\hat{\sigma}=\sqrt{\frac{Q}{n-2}}$$

$$r_i=\frac{\hat{\varepsilon}_i}{\hat{\sigma}\sqrt{1-h_i}},$$

这里 $\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i,l_{xx}=\sum_{i=1}^{n}(x_i-\bar{x})^2$,

$$Q=l_{yy}-U,l_{yy}=\sum_{i=1}^{n}(y_i-\bar{y})^2$$

$$U=(\hat{b})^2l_{xx},\bar{y}=\frac{1}{n}\sum_{i=1}^{n}y_i$$

数学上可以证明,如果(1.10)成立,则 $r_1,r_2,\cdots,r_n$近似相互独立且近似服从标准正态分布.于是,在(1.10)成立的情况下应该有 $P(|r_i|>2)\approx0.05$.换句话说,当 n 比较大时,$r_1,r_2,\cdots,r_n$中大约有$[0.05n]$个 r_i在区间$[-2,2]$之外(这里$[x]$表示不超过 x 的最大整数).若是出现这种情况,我们认为(1.10)成立.否则的话(即 $r_1,r_2,\cdots,r_n$落在$[-2,2]$之外的个数超过$[0.05n]$)应拒绝(1.10),即不能认为随机项 $\varepsilon_1,\varepsilon_2,\cdots,\varepsilon_n$服从同一分布$N(0,\sigma^2)$.

对残差$\{\hat{\varepsilon}_i,i=1,2,\cdots,n\}$的进一步分析,可提供我们许多信息.这方面的深入研究,参看[17].

§2 多元线性回归

在实际应用中,由于事物的复杂性,在很多情况下要采用多元回归方法.就方法的实质来说,多元跟一元在很多方面是相同的,只是多元回归方法更复杂些,计算量相当大.[不过,对于电子计算机,

多元回归的计算量是很小的.一般的统计软件包都有多元回归(以及逐步回归方法)的专门程序.]本讲义只列出有关结论.主要讨论线性回归.

1. 模型

设因变量 Y 与自变量 $x_1, x_2, \cdots, x_k$ 有关系式:

$$Y = b_0 + b_1 x_1 + \cdots + b_k x_k + \varepsilon$$

其中 ε 是随机项.现有 n 组数据:

$$\begin{array}{l} (y_1; x_{11}, x_{12}, \cdots, x_{1k}) \\ (y_2; x_{21}, x_{22}, \cdots, x_{2k}) \\ \cdots\cdots\cdots\cdots \\ (y_n; x_{n1}, x_{n2}, \cdots, x_{nk}) \end{array} \tag{2.1}$$

(其中 x_{ij} 是自变量 x_j 的第 i 个值,y_i 是 Y 的第 i 个观测值.)假定

$$\begin{cases} Y_1 = b_0 + b_1 x_{11} + b_2 x_{12} + \cdots + b_k x_{1k} + \varepsilon_1 \\ Y_2 = b_0 + b_1 x_{21} + b_2 x_{22} + \cdots + b_k x_{2k} + \varepsilon_2 \\ \cdots\cdots\cdots\cdots \\ Y_n = b_0 + b_1 x_{n1} + b_2 x_{n2} + \cdots + b_k x_{nk} + \varepsilon_n \end{cases} \tag{2.2}$$

其中 $b_0, b_1, \cdots, b_k$ 是待估参数;而 $\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_n$ 相互独立且服从相同的分布 $N(0, \sigma^2)$,(σ 未知).

说明:

(1) 所谓"多元"是指这里的自变量有多个,而因变量还只是一个;自变量是普通变量,因变量是随机变量.

(2) (2.1)中的诸 y 是数据,而(2.2)中的诸 Y_i 是随机变量.我们把(2.1)中的诸 y_i 当作(2.2)中的相应的 Y_i 的观测值.

(3) (2.2)表示 Y 跟 $x_1, x_2, \cdots, x_k$ 的关系是线性的.对于某些非线性的关系,可通过适当的变换化为形式上是线性的问题;比如,一元多项式回归问题(即虽然只有一个 x,但 Y 对 x 的回归式是多项式:$\hat{y} = b_0 + b_1 x + b_2 x^2 + \cdots + b_k x^k$),就可以通过变换化为多元线性回归问题.(令 $x_1 = x, x_2 = x^2, \cdots, x_k = x^k$ 就可以了.)

2. 最小二乘估计与正规方程

我们称使

$$Q(b_0,b_1,\cdots,b_k)$$

$$\triangleq \sum_{t=1}^{n}[y_t-(b_0+b_1x_{t1}+b_2x_{t2}+\cdots+b_kx_{tk})]^2$$

达到最小的 $\hat{b}_0,\hat{b}_1,\cdots,\hat{b}_k$ 为参数 $b_0,b_1,\cdots,b_k$ 的**最小二乘估计**.

可以证明,最小二乘估计也就是下列方程组的解:

$$\begin{cases} l_{11}b_1+l_{12}b_2+\cdots+l_{1k}b_k=l_{1y} \\ l_{21}b_1+l_{22}b_2+\cdots+l_{2k}b_k=l_{2y} \\ \cdots\cdots\cdots\cdots \\ l_{k1}b_1+l_{k2}b_2+\cdots+l_{kk}b_k=l_{ky} \\ b_0=\bar{y}-b_1\bar{x}_1-\cdots-b_k\bar{x}_k \end{cases} \tag{2.3}$$

其中

$$\bar{y}=\frac{1}{n}\sum_t y_t,\quad \bar{x}_i=\frac{1}{n}\sum_t x_{it},\quad i=1,2,\cdots,k$$

$$l_{ij}=l_{ji}=\sum_t(x_{it}-\bar{x}_i)(x_{jt}-\bar{x}_j),\quad i,j=1,2,\cdots,k$$

$$l_{iy}=\sum_t(x_{it}-\bar{x}_i)(y_t-\bar{y}),\quad i=1,2,\cdots,k$$

方程组(2.3)称为**正规方程**.

在多元线性回归的研究中,矩阵是一个强有力的工具.许多结论及其证明用矩阵表达出来显得简洁、清楚且便于记忆.当然,对初学者来说,要学会使用矩阵记号及其运算.下面利用矩阵论证最小二乘估计一定存在,并给出方程组(2.3)的一个重要的等价形式.

我们先用矩阵表示数据(2.1).令

$$\boldsymbol{Y}=\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix},\boldsymbol{b}=\begin{pmatrix} b_0 \\ b_1 \\ \vdots \\ b_k \end{pmatrix},\boldsymbol{E}=\begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} \tag{2.4}$$

$$\boldsymbol{X}=\begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1k} \\ x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nk} \end{pmatrix}, \boldsymbol{C}=(\boldsymbol{E} \quad \boldsymbol{X}), \tag{2.5}$$

这里 $\boldsymbol{E}$ 是分量全是 1 的 n 维列向量(n 行 1 列的矩阵).用 $\boldsymbol{A}'$表示矩阵 $\boldsymbol{A}$ 的转置,$\|\boldsymbol{a}\|$ 表示列向量 $\boldsymbol{a}$ 的长度即 $\|\boldsymbol{a}\|=\sqrt{\boldsymbol{a}'\boldsymbol{a}}$,例如 $\|\boldsymbol{E}\|=\sqrt{n}$,$\|\boldsymbol{Y}\|=\sqrt{y_1^2+y_2^2+\cdots+y_n^2}$.

令 $Q(\boldsymbol{b})=Q(b_0,b_1,\cdots,b_k)=\sum_{t=1}^{n}[y_t-(b_0+b_1x_{t1}+\cdots+b_kx_{tk})]^2$,则

$$Q(\boldsymbol{b})=\|\boldsymbol{Y}-\boldsymbol{C}\boldsymbol{b}\|^2$$

设 $\tilde{\boldsymbol{b}}=(\tilde{b}_0,\tilde{b}_1,\cdots,\tilde{b}_k)'$,则

$$\begin{aligned} Q(\boldsymbol{b}) &= \|\boldsymbol{Y}-\boldsymbol{C}\tilde{\boldsymbol{b}}+\boldsymbol{C}(\tilde{\boldsymbol{b}}-\boldsymbol{b})\|^2=(\boldsymbol{Y}-\boldsymbol{C}\tilde{\boldsymbol{b}}+\boldsymbol{C}(\tilde{\boldsymbol{b}}-\boldsymbol{b}))'(\boldsymbol{Y}-\boldsymbol{C}\tilde{\boldsymbol{b}}+\boldsymbol{C}(\tilde{\boldsymbol{b}}-\boldsymbol{b})) \\ &=(\boldsymbol{Y}-\boldsymbol{C}\tilde{\boldsymbol{b}})'(\boldsymbol{Y}-\boldsymbol{C}\tilde{\boldsymbol{b}})+(\boldsymbol{C}(\tilde{\boldsymbol{b}}-\boldsymbol{b}))'\boldsymbol{C}(\tilde{\boldsymbol{b}}-\boldsymbol{b}) \\ &\quad +2(\boldsymbol{C}(\tilde{\boldsymbol{b}}-\boldsymbol{b}))'(\boldsymbol{Y}-\boldsymbol{C}\tilde{\boldsymbol{b}}) \\ &=\|\boldsymbol{Y}-\boldsymbol{C}\tilde{\boldsymbol{b}}\|^2+\|\boldsymbol{C}(\tilde{\boldsymbol{b}}-\boldsymbol{b})\|^2-2(\tilde{\boldsymbol{b}}-\boldsymbol{b})'(\boldsymbol{C}'\boldsymbol{C}\tilde{\boldsymbol{b}}-\boldsymbol{C}'\boldsymbol{Y}) \end{aligned} \tag{2.6}$$

由此知,若 $\tilde{\boldsymbol{b}}$ 满足方程

$$\boldsymbol{C}'\boldsymbol{C}\tilde{\boldsymbol{b}}=\boldsymbol{C}'\boldsymbol{Y} \tag{2.7}$$

则 $Q(\boldsymbol{b})\geqslant\|\boldsymbol{Y}-\boldsymbol{C}\tilde{\boldsymbol{b}}\|^2=Q(\tilde{\boldsymbol{b}})$.即 $Q(\boldsymbol{b})$在 $\tilde{\boldsymbol{b}}$ 达到最小值.我们指出方程(2.7)一定有解.实际上,线性方程组(2.7)的增广矩阵

$$(\boldsymbol{C}'\boldsymbol{C} \quad \boldsymbol{C}'\boldsymbol{Y})=\boldsymbol{C}'(\boldsymbol{C} \quad \boldsymbol{Y})$$

可见增广矩阵的秩不超过矩阵 $\boldsymbol{C}$ 的秩,而 $\boldsymbol{C}'\boldsymbol{C}$ 与 $\boldsymbol{C}$ 有相同的秩,因而增广矩阵的秩与系数矩阵 $\boldsymbol{C}'\boldsymbol{C}$ 的秩相等.根据线性方程组解的存在定理,方程(2.7)一定有解.这就证明了 $b_0,b_1,\cdots,b_k$ 的最小二乘估计一定存在.

另一方面,若 $Q(\boldsymbol{b})$在 $\hat{\boldsymbol{b}}$达到最小值,设 $\tilde{\boldsymbol{b}}$ 是(2.7)的任何一个解,则从(2.6)知 $Q(\hat{\boldsymbol{b}})=Q(\tilde{\boldsymbol{b}})+\|\boldsymbol{C}(\tilde{\boldsymbol{b}}-\hat{\boldsymbol{b}})\|^2$.由于 $Q(\boldsymbol{b})$在 $\hat{\boldsymbol{b}}$达到最小值,故$\boldsymbol{C}(\tilde{\boldsymbol{b}}-\hat{\boldsymbol{b}})=0$.从而 $\boldsymbol{C}\hat{\boldsymbol{b}}=\boldsymbol{C}\tilde{\boldsymbol{b}}$.于是 $\boldsymbol{C}'\boldsymbol{C}\hat{\boldsymbol{b}}=\boldsymbol{C}'\boldsymbol{C}\tilde{\boldsymbol{b}}=\boldsymbol{C}'\boldsymbol{Y}$.即 $\hat{\boldsymbol{b}}$一定满足(2.7).

总之,为了 $\hat{b}_0,\hat{b}_1,\cdots,\hat{b}_k$ 是 $b_0,b_1,\cdots,b_k$ 的最小二乘估计,必须且只需 $\hat{\boldsymbol{b}}$

$=(\hat{b}_0,\hat{b}_1,\cdots,\hat{b}_k)'$满足方程(2.7).

下面指出方程(2.7)与方程组(2.3)是等价的(即前者的解一定是后者的解,反之亦然).从(2.5)知

$$C'C=\begin{pmatrix} n & E'X \\ XE & X'X \end{pmatrix}$$

$$C'C\begin{pmatrix} b_0 \\ b_1 \\ \vdots \\ b_k \end{pmatrix}=\begin{pmatrix} nb_0+E'X\begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_k \end{pmatrix} \\ b_0XE+X'X\begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_k \end{pmatrix} \end{pmatrix},$$

于是方程(2.7)就是下列方程组:

$$\begin{cases} nb_0+E'X\begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_k \end{pmatrix}=E'Y & (2.8) \\ b_0XE+X'X\begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_k \end{pmatrix}=X'Y & (2.9) \end{cases}$$

从(2.8)得

$$b_0=\frac{1}{n}E'Y-\frac{1}{n}E'X\begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_k \end{pmatrix} \tag{2.10}$$

代入(2.9)得

$$\left[X'X-\frac{1}{n}XEE'X\right]\begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_k \end{pmatrix}=X'Y-\frac{1}{n}XEE'Y \tag{2.11}$$

故方程(2.7)与方程组(2.10)—(2.11)等价.令 $\boldsymbol{L}=(l_{ij})_{k\times k}$,这里

$$l_{ij}=\sum_{t=1}^{n}(x_{ti}-\bar{x}_i)(x_{tj}-\bar{x}_j)$$

$$\bar{x}_i=\frac{1}{n}\sum_{t=1}^{n}x_{ti}\ (1\leqslant i\leqslant k)$$

易知
$$l_{ij}=\sum_{t=1}^{n}x_{ti}x_{tj}-\frac{1}{n}\sum_{t=1}^{n}x_{ti}\cdot\sum_{t=1}^{n}x_{tj}$$

故
$$\boldsymbol{L}=\boldsymbol{X}'\boldsymbol{X}-\frac{1}{n}\boldsymbol{X}\boldsymbol{E}\boldsymbol{E}'\boldsymbol{X}$$

类似地,$l_{iy}\triangleq\sum_{t=1}^{n}(x_{ti}-\bar{x}_i)(y_t-\bar{y})$

$$=\sum_{t=1}^{n}x_{ti}y_t-\frac{1}{n}\sum_{t=1}^{n}x_{ti}\cdot\sum_{t=1}^{n}y_t\quad\left(\bar{y}=\frac{1}{n}\sum_{t=1}^{n}y_t\right).$$

于是

$$\begin{pmatrix}l_{1y}\\l_{2y}\\\vdots\\l_{ky}\end{pmatrix}=\boldsymbol{X}'\boldsymbol{Y}-\frac{1}{n}\boldsymbol{X}\boldsymbol{E}\boldsymbol{E}'\boldsymbol{Y}$$

方程(2.11)就是

$$\boldsymbol{L}\begin{pmatrix}b_1\\b_2\\\vdots\\b_k\end{pmatrix}=\begin{pmatrix}l_{1y}\\l_{2y}\\\vdots\\l_{ky}\end{pmatrix}$$

方程(2.10)就是
$$b_0=\bar{y}-\sum_{i=1}^{k}b_i\bar{x}_i.$$

可见方程(2.7)与方程组(2.3)等价.方程(2.7)也叫正规方程,它在多元回归的研究中比方程组(2.3)更重要.

3. 平方和分解公式与 σ^2 的无偏估计

跟一元的情形类似,我们有平方和分解公式

$$l_{yy}=Q+U\tag{2.12}$$

其中

$$l_{yy}=\sum(y_t-\bar{y})^2$$

$$Q=\sum(y_t-\hat{y}_t)^2$$

$$U=\sum(\hat{y}_t-\bar{y})^2$$

而

$$\hat{y}_t=\hat{b}_0+\hat{b}_1x_{t1}+\hat{b}_2x_{t2}+\cdots+\hat{b}_kx_{tk},t=1,2,\cdots,n$$

还称 U 为回归平方和，Q 为剩余平方和.

(跟(1.11)类似，我们有

$$U=\hat{b}_1l_{1y}+\hat{b}_2l_{2y}+\cdots+\hat{b}_kl_{ky}$$

具体计算时，用这个公式是比较方便的.)

我们有

$$E[Q/(n-k-1)]=\sigma^2 \tag{2.13}$$

(实际上，可以证明 Q/σ^2 服从自由度为 $n-k-1$ 的 χ^2 分布.)记

$$\hat{\sigma}^2=Q/(n-k-1)$$

(2.13)表明，$\hat{\sigma}^2$ 是 σ^2 的无偏估计.有时 $\hat{\sigma}^2$ 也用 s^2 来记.

4. 相关性检验

跟一元的情形类似，Y 与 $x_1,x_2,\cdots,x_k$ 间是否存在线性相关关系的问题，在模型(2.2)的假定下，也就是一个假设检验的问题.要检验的是假设 $H_0:b_1=b_2=\cdots=b_k=0$.若经检验否定假设 H_0，则认为它们之间存在线性相关关系.

具体的统计量也是类似的：

$$F=\frac{U/k}{Q/(n-k-1)} \tag{2.14}$$

它是一元情形的推广[请读者将(2.14)跟(1.9)作个比较].可以证明，在(2.2)的假定以及假设 H_0 成立的情况下，(2.14)给出的统计量 F 服从自由度为 $k,n-k-1$ 的 F 分布.于是，对给定的 α，将由(2.14)算出的 F 值跟相应的临界值 λ 作比较.如 $F>\lambda$，则否定 H_0；否则 H_0 是相容的.

5. 偏回归平方和与因素主次的判别

以上几个小节的内容，纯属一元情形的推广，只是形式上复杂

些而已.而本小节是多元回归问题所特有的.

先从判别因素的主次说起.在实际工作中,我们还关心 Y 对 $x_1,x_2,\cdots,x_k$ 的线性回归中,哪些因素(即自变量)更重要些,哪些不重要.怎样来衡量某个特定因素 $x_i(i=1,2,\cdots,k)$的影响呢?我们知道,回归平方和 U 这个量,刻画了全体自变量 $x_1,x_2,\cdots,x_k$ 对于 Y 的总的线性影响.为了研究 x_k 的作用,可以这样来考虑:从原来的 k 个自变量中扣除 x_k,我们知道这 $k-1$ 个自变量 x_1, $x_2,\cdots,x_{k-1}$对于 Y 的总的线性影响也是一个回归平方和,记作 $U_{(k)}$;我们称

$$u_k \triangleq U - U_{(k)}$$

为 $x_1,x_2,\cdots,x_k$ 中 x_k 的偏回归平方和.这个偏回归平方和就可看作 x_k 产生的作用.类似地,可定义 $U_{(i)}$.

一般地,称

$$u_i \triangleq U - U_{(i)} \quad (i=1,2,\cdots,k) \tag{2.15}$$

为 $x_1,x_2,\cdots,x_k$ 中 x_i 的偏回归平方和.用它来衡量 x_i 在 Y 对 x_1, $x_2,\cdots,x_k$ 的线性回归中的作用的大小.

对于 u_i 的计算,我们有下式:

$$u_i = \frac{\hat{b}_i^2}{c_{ii}} \tag{2.16}$$

其中 c_{ii} 是矩阵 $(l_{ij})_{k\times k}$ 的逆矩阵的对角线上的第 i 个元素.

我们顺便指出,从理论上说,对于假设"$H_0:b_i=0$",可用统计量 $F_i=u_i/s^2$ 来检验.这个统计量在 H_0 成立时服从自由度为 $1,n-k-1$ 的 F 分布.实用上,如果根据观测值算出的 F_i 的数值大于 $\alpha=0.05$ 时的临界值,称变量 x_i 是显著的;而若算得的 F_i 的值还大于 $\alpha=0.01$ 时的临界值,就称 x_i 是高度显著的.当 F_i 的值很小时,就应从回归方程中将 x_i 剔除.

最后,我们指出,基于数据(2.1)检验(2.2)中随机项 $\varepsilon_1,\varepsilon_2,$

$\cdots,\varepsilon_n$是否服从正态分布$N(0,\sigma^2)$的办法也是有的，这属于残差分析的范围，本书从略．参看[17]．

例 2.1 （广告策略）．某公司为了推销商品，研究广告费用x与获得的纯利润y之间的关系，以确定最佳的广告策略．调查以往的情况，有以下数据：

x	1	1	2	2	2	3	3	4	4	4
y	14.80	15.90	20.20	20.00	18.55	22.20	20.90	21.00	18.30	20.70

x	5	5
y	16.10	14.75

（单位：万元）

试找出y与x的相关关系式并确定最优的广告费．

解 先根据数据画出散点图．

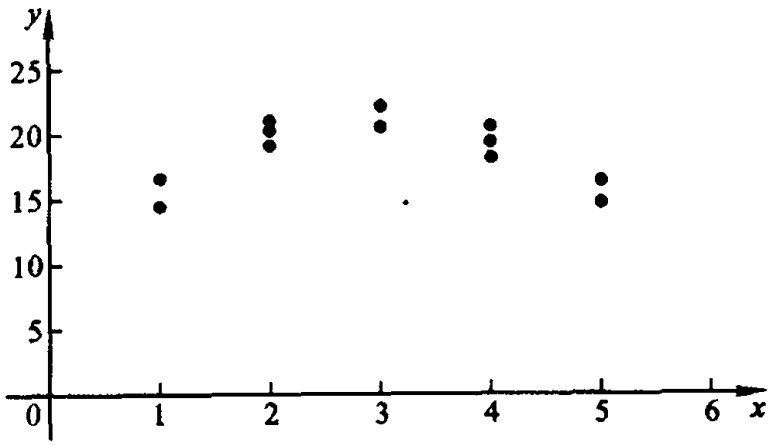

图 7.6

从图上看出y与x不是线性关系，自然想到用x的二次函数来近似y．即可认为有下列关系式：

$$y = b_0 + b_1 x + b_2 x^2 + \varepsilon$$

其中ε是随机项，$\varepsilon \sim N(0,\sigma^2)$（$\sigma$未知）令$x_1 = x, x_2 = x^2$，则上述关系式化为

$$y = b_0 + b_1 x_1 + b_2 x_2 + \varepsilon .$$

这是二元线性回归模型.从 x_1 的数据自然得到 $x_2 = x^2$ 的数据.设 x 的数据是 $x_{11}, x_{21}, \cdots, x_{n1}$,则 x_2 的数据是 $x_{12} = x_{11}^2, x_{22} = x_{21}^2, \cdots, x_{n2} = x_{n1}^2$,相应的 y 是 $y_1, y_2, \cdots, y_n$,

为了找出 b_0, b_1, b_2 的最小二乘估计,要解正规方程(2.3).

利用所给的数据,可计算出 $l_{11} = \sum_{t=1}^{n} (x_{t1} - \bar{x}_1)^2 = 22$, $l_{21} = l_{12} = \sum_{t=1}^{n} (x_{t1} - \bar{x}_1)(y_t - \bar{y}) = 132$, $l_{22} = \sum_{t=1}^{n} (x_{t2} - \bar{x}_2)^2 = 822.5$

$$(\bar{x}_1 = \frac{1}{n}\sum_{t=1}^{n} x_{t1} = 3, \bar{x}_2 = \frac{1}{n}\sum_{t=1}^{n} x_{t2} = 10.83, \bar{y} = \frac{1}{n}\sum_{t=1}^{n} y_t = 18.61)$$

$$l_{1y} = \sum_{t=1}^{n} (x_{t1} - \bar{x}_1)(y_t - \bar{y}) = \sum_{t=1}^{n} x_{t1} y_t - n\bar{x}_1 \bar{y} = 1.79$$

$$l_{2y} = \sum_{t=1}^{n} (x_{t2} - \bar{x}_2)(y_t - \bar{y}) = \sum_{t=1}^{n} x_{t2} y_t - n\bar{x}_2 \bar{y} = -33.7$$

解正规方程(2.3)得:

$$\hat{b}_0 = 7.97, \hat{b}_1 = 8.82, \hat{b}_2 = -1.46$$

回归方程是

$$\hat{y} = 7.97 + 8.82x - 1.46x^2$$

为了检验假设 $H_0: b_1 = 0, b_2 = 0$,使用统计量 F(见(2.14)).可算出 $F = 21.2$.查 F 分布表知 $F(2,9)$ 分布的 0.95 分位数是 4.26.现在 $F > 4.26$.故应拒绝 H_0.所以在检验水平 $\alpha = 0.05$ 下,上述回归方程体现了 y 与 x, x^2 的线性相关关系.利用这个回归方程可以进行预测.易知 $x = 3.02$ 时相应的 $\hat{y}$ 最大.即广告费是 3.02(万元)时纯利润最大.

***例 2.2**(生理节律模型) 为了测定一个人在 24 小时内的生理节律(例如血压(收缩压或舒张压)如何随时间而变化),一些学者提出了下列模型

$$f(t)=M+A\cos(\omega t+\phi),$$

其中 M 是基准值，A 是振幅，ϕ 是相位，ω 是角频率，例如 $\omega=360/24$. $f(t)$是所关心的生理指标. 问题是：设有观测值 $y_j=f(t_j)+\varepsilon_j(j=1,2,\cdots,n)$，这里 t_j是第j 个观测时刻，$\varepsilon_1,\varepsilon_2,\cdots,\varepsilon_n$是相互独立的随机项，$\varepsilon_j\sim N(0,\sigma^2)$($\sigma$ 未知)，如何估计 M,A,ϕ? ($0\leqslant\phi<360°$).

解 易知

$$y_j=M+A\cos\phi\cdot\cos\omega t_j-A\sin\phi\cdot\sin\omega t_j+\varepsilon_j$$

故

$$y_j=M+\beta x_j+\gamma z_j+\varepsilon_j(j=1,2,\cdots,n),$$

这里 $x_j=\cos\omega t_j$, $z_j=\sin\omega t_j$,

$$\beta=A\cos\phi,\gamma=-A\sin\phi. \tag{2.17}$$

这便化成了二元线性回归模型.

我们可利用正规方程(2.3)求出 β,γ 的最小二乘估计$\hat\beta,\hat\gamma$.

易知，$l_{11}=\sum_{j=1}^{n}(x_j-\bar x)^2$, $l_{22}=\sum_{1}^{n}(z_j-\bar z)^2$, $l_{12}=\sum_{j=1}^{n}(x_j-\bar x)(z_j-\bar z)$, $(\bar x=\frac{1}{n}\sum_{j=1}^{n}x_j,\bar z=\frac{1}{n}\sum_{j=1}^{n}z_j)$.

$$l_{1y}=\sum_{j=1}^{n}(x_j-\bar x)(y_j-\bar y)\qquad(\bar y=\frac{1}{n}\sum_{j=1}^{n}y_j)$$

$$l_{2y}=\sum_{j=1}^{n}(z_j-\bar z)(y_j-\bar y).$$

解正规方程，得

$$\hat\beta=\frac{l_{22}l_{1y}-l_{12}l_{2y}}{l_{11}l_{22}-l_{12}^2},\hat\gamma=\frac{-l_{12}l_{1y}+l_{11}l_{2y}}{l_{11}l_{22}-l_{12}^2}$$

$$\hat M=\bar y-\bar x\hat\beta-\bar z\hat\gamma,$$

从(2.17)可得到 A 和ϕ 的估计$\hat A,\hat\phi$

$$\hat A=\sqrt{(\hat\beta)^2+(\hat\gamma)^2}$$

$$\hat{\phi}=\begin{cases}360^\circ-\theta & \text{当}\hat{\beta}>0,\hat{\gamma}\geqslant 0\\ \theta & \hat{\beta}>0,\hat{\gamma}<0\\ \theta+180^\circ & \hat{\beta}\leqslant 0,\hat{\gamma}\geqslant 0\\ 180^\circ-\theta & \hat{\beta}\leqslant 0,\hat{\gamma}<0,\end{cases}$$

其中 $\theta=\arctan\left(\left|\frac{\gamma}{\beta}\right|\right)(0\leqslant\theta\leqslant 90^\circ)$,

于是有非线性回归方程

$$\hat{y}=\hat{M}+\hat{A}\cos(\omega t+\hat{\phi}) \tag{2.18}$$

这个方程是否有意义呢？要检验振幅 A 是否为零.这等价于检验 $H_0:\beta=\gamma=0$.使用统计量

$$F=\frac{U/2}{Q/(n-3)}\quad(\text{见}(2.14))$$

在 H_0 下 F 服从自由度为 $2,n-3$ 的 F 分布、若检验水平为 0.05.λ 是这个 F 分布的 0.95 分位数,则当 $F>\lambda$ 时应拒绝 H_0,从而方程(2.18)是有意义的.(若 $F\leqslant\lambda$,则不能拒绝 H_0,方程(2.18)没有意义)在计算 F 时,注意 $U=l_{1y}\hat{\beta}+l_{2y}\hat{\gamma}$,$Q=l_{yy}-U$,这里 $l_{yy}=\sum_{i=1}^{n}(y_i-\bar{y})^2$.

在实际工作中,通常观测时刻是等间隔的,$t_j=\frac{j-1}{n}(j=1,2,\cdots,n)$且 $\omega=360^\circ$.(最常见的情况是 $n=12$ 或 24)这时上面的计算公式均大为简单.实际上,$\sum_{j=1}^{n}x_j=\sum_{j=1}^{n}\cos\omega t_j=0$,$\sum_{j=1}^{n}z_j=\sum_{j=1}^{n}\sin\omega t_j=0$,$\sum_{j=1}^{n}x_jz_j=\sum_{j=1}^{n}(\cos\omega t_j)\sin\omega t_j=0$[①]. $\sum_{j=1}^{n}x_j^2=$

① 利用公式

$$\cos k\theta=\frac{\sin\left(k+\frac{1}{2}\right)\theta-\sin\left(k-\frac{1}{2}\right)\theta}{2\sin\frac{\theta}{2}}\quad(\text{当分母不是 0})$$

$$\sum_{k=0}^{n-1}\left(\frac{1+\cos 2k\theta}{2}\right)=\frac{n}{2}\quad\left(\theta=\frac{360^\circ}{n}\right)$$

同理 $\sum_{j=1}^{n} z_j^2=\frac{n}{2}$.

于是 $\hat{M}=\frac{1}{n}\sum_{j=1}^{n} y_j=\bar{y}$, $\hat{\beta}=\frac{1}{n}\sum_{j=1}^{n} x_j y_j$, $\hat{\gamma}=\frac{1}{n}\sum_{j=1}^{n} z_j y_j$, 统计量 F 为

$$F=\frac{n\hat{A}^2/2}{Q/n-3}\quad(\text{因为 } U=n(\hat{\beta})^2+n(\hat{\gamma})^2)$$

这里 $\hat{A}^2=\hat{\beta}^2+\hat{\gamma}^2$, $Q=l_{vv}-n\hat{A}^2$. 这些都是便于应用的简单公式.

习 题 十 八

1. 炼铝厂测得所产铸模用的铝的硬度 x 与抗张强度 y 数据如下:

x	68	53	70	84	60	72	51	83	70	64
y	288	293	349	343	290	354	283	324	340	286

求 y 对 x 的回归直线.

2. 检验第 1 题所得回归直线的显著性.

3. 对于第 1 题所讨论的问题, 试预报当铝的硬度 $x=65$ 时的抗张强度 y.

(接上页注)
知道

$$\sum_{k=0}^{n-1}\cos k\theta=\frac{2\sin\frac{n\theta}{2}\cdot\cos\frac{(n-1)\theta}{2}}{2\sin\frac{\theta}{2}}$$

故 $\theta=\frac{360^\circ}{n}$时, $\sum_{k=0}^{n-1}\cos k\theta=0$ $(n\geqslant 2)$

同理知 $\sum_{k=0}^{n-1}\sin k\theta=0$, $\sum_{k=0}^{n-1}\cos k\theta\cdot\sin k\theta=0$

§3 逻辑斯谛(Logistic)回归模型

在§1和§2中讨论的经典线性回归模型里,因变量(响应变量)是连续变量.在实际工作中(特别是对社会现象的研究中)常遇到因变量只取分类值尤其是只取二分类值(即0或1)的情形,这时就不能用§1和§2中的处理方法了.例如用 x 表示一个家庭的年收入,$Y=1$ 表示该家庭在一段时间内购买某种耐用消费品(例如汽车),$Y=0$ 表示不购买这种耐用消费品,我们要研究的是概率 $P(Y=1)$ 与 x 的关系.

更一般地,若随机变量 Y 只取值0或1,有若干个变量 $x_1, x_2, \cdots, x_k$ 影响 Y 的取值,我们关心的是概率 $p=P(Y=1)$ 是如何依赖于 $x_1, x_2, \cdots, x_k$ 的.

对 p 的研究等价于对 $\frac{p}{1-p}$ 的研究,因为 $\frac{p}{1-p}$ 是 p 的严格增连续函数.$\frac{p}{1-p}$ 叫做发生比或优比(odds ratio).如果有下列关系式:

$$\ln \frac{p}{1-p} = \beta_0 + \sum_{i=1}^{k} \beta_i x_i \tag{3.1}$$

(其中 $\beta_0, \beta_1, \cdots, \beta_k$ 是常数),则称二分类变量 Y 与自变量 $x_1, x_2, \cdots, x_k$ 的关系符合逻辑斯谛回归模型.这里 $p=P(Y=1)$,为了体现这个概率与 $x_1, x_2, \cdots, x_k$ 的联系,常写成 $P(Y=1 \mid x_1, x_2, \cdots, x_k)$.(3.1)有下列等价形式:

$$P(Y=1 \mid x_1, x_2, \cdots, x_k) = \frac{\exp\left\{\beta_0 + \sum_{i=1}^{k} \beta_i x_i\right\}}{1 + \exp\left\{\beta_0 + \sum_{i=1}^{k} \beta_i x_i\right\}}$$

在(3.1)中,$\beta_0, \beta_1, \cdots, \beta_k$ 通常是未知的,需要利用数据进行估

计.一旦这些参数的值确定了,(3.1)式就可用来对 p 进行预测,也可用来对各自变量的重要性进行评价.

为简单计,以下只考虑 $K=1$(即一个自变量)的情形,用 x 表示 x_1.这时(3.1)化为

$$\ln\frac{p}{1-p}=\beta_0+\beta_1 x \tag{3.2}$$

令 $p(x)=P(Y=1|x)$,则

$$p(x)=\frac{\exp\{\beta_0+\beta_1 x\}}{1+\exp\{\beta_0+\beta_1 x\}} \tag{3.3}$$

怎样估计未知参数 β_0,β_1 呢? 通常有两个办法:最大似然估计法和加权最小二乘法.

最大似然法 设有下列数据:$x=x_i$ 时 Y 的值是 y_i($i=1,2,\cdots,n$),$y_i=0$ 或 1.应注意,这里 x_i 是自变量 x 的第 i 个值,不是(3.1)中的第 i 个自变量!

显然,

$$P(Y=y_i|x_i)=[p(x_i)]^{y_i}[1-p(x_i)]^{1-y_i}$$

于是观测值$(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n)$对应的似然函数是

$$L(\beta_0,\beta_1)=\prod_{i=1}^{n}[p(x_i)]^{y_i}[1-p(x_i)]^{1-y_i}$$

于是

$$\ln L(\beta_0,\beta_1)=\sum_{i=1}^{n}y_i(\beta_0+\beta_1 x_i)-\sum_{i=1}^{n}\ln(1+e^{\beta_0+\beta_1 x_i}).$$

令

$$\frac{\partial\ln L(\beta_0,\beta_i)}{\partial\beta_i}=0\quad(i=0,1).$$得似然方程组:

$$\sum_{i=1}^{n}\left(y_i-\frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}}\right)=0$$

$$\sum_{i=1}^{n}\left(y_i-\frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}}\right)x_i=0$$

若$(\hat{\beta}_0,\hat{\beta}_1)$是似然方程组的根且$x_1,x_2,\cdots,x_n$不全相等，则似然方程组的根是惟一的，而且$(\hat{\beta}_0,\hat{\beta}_1)$是$L(\beta_0,\beta_1)$的最大值点，因而$\hat{\beta}_0,\hat{\beta}_1$分别是$\beta_0,\beta_1$的最大似然估计.(可以证明，$\ln L(\beta_0,\beta_1)$是二元严格凹函数).但应注意的是，似然方程组有时无根(例如，所有y_i都是1的情形).在SAS和SPSS等国际流行的软件包里都有计算最大似然估计$\hat{\beta}_0,\hat{\beta}_1$的程序.

加权最小二乘法 此法对数据有些特殊要求.设$x=x_i$时对Y作了n_i次观测(n_i较大)，其中事件$\{Y=1\}$发生了γ_i次($i=1,2,\cdots,m$).($x_1,x_2,\cdots,x_m$两两不同).通常用

$$z_i\triangleq\ln\frac{\gamma_i+0.5}{n_i-\gamma_i+0.5} \tag{3.4}$$

作为$\ln\frac{p(x_i)}{1-p(x_i)}$的估计值($i=1,2,\cdots,m$).

令

$$\nu_i\triangleq\frac{(n_i+1)(n_i+2)}{n_i(\gamma_i+1)(n_i-\gamma_i+1)}\qquad(i=1,\cdots,m) \tag{3.5}$$

$$\widetilde{Q}(\beta_0,\beta_1)\triangleq\sum_{i=1}^{m}\frac{1}{\nu_i}(z_i-\beta_0-\beta_1x_i)^2$$

使$\widetilde{Q}(\beta_0,\beta_1)$达到最小值的$\tilde{\beta}_0,\tilde{\beta}_1$称为$\beta_0,\beta_1$的加权最小二乘估计.这里$\frac{1}{\nu_1},\frac{1}{\nu_2},\cdots,\frac{1}{\nu_m}$就是所谓的权.可以证明加权最小二乘估计存在且惟一.令$\frac{\partial\widetilde{Q}(\beta_0,\beta_1)}{\partial\beta_i}=0(i=0,1)$，得方程组：

$$\beta_0\sum_1^m\frac{1}{\nu_i}+\beta_1\sum_{i=1}^m\frac{x_i}{\nu_i}=\sum_{i=1}^m\frac{z_i}{\nu_i}$$

$$\beta_0\sum_1^m\frac{x_i}{\nu_i}+\beta_1\sum_{i=1}^m\frac{x_i^2}{\nu_i}=\sum_{i=1}^m\frac{x_iz_i}{\nu_i}$$

解此方程组，可得加权最小二乘估计如下：

$$\hat{\beta}_0 = \frac{1}{l_1 l_3 - (l_2)^2}(l_5 l_3 - l_2 l_4) \tag{3.6}$$

$$\hat{\beta}_1 = \frac{1}{l_1 l_3 - (l_2)^2}(l_1 l_4 - l_2 l_5) \tag{3.7}$$

这里 $l_1 = \sum_{i=1}^{m} \frac{1}{\nu_i}$，$l_2 = \sum_{i=1}^{m} \frac{x_i}{\nu_i}$，$l_3 = \sum_{i=1}^{m} \frac{x_i^2}{\nu_i}$，$l_4 = \sum_{i=1}^{m} \frac{x_i z_i}{\nu_i}$，$l_5 = \sum_{i=1}^{m} \frac{z_i}{\nu i}$.

加权最小二乘估计是基于什么思想导出的呢？本来应用 $\frac{\gamma_i}{n_i - \gamma_i}$ 作为 $\frac{p(x_i)}{1-p(x_i)}$ 的估计. 为了避免分子和分母出现零，用 $\frac{\gamma_i + 0.5}{n_i - \gamma_i + 0.5}$ 作为 $\frac{p(x_i)}{1-p(x_i)}$ 的估计. 可以证明(基于概率论中的极限定理)，$z_i = \ln \frac{\gamma_i + 0.5}{n_i - \gamma_i + 0.5}$ 近似服从正态分布 $N\left(\ln \frac{p(x_i)}{1-p(x_i)}, \frac{1}{n_i p(x_i)[1-p(x_i)]}\right)$，所以 $z_i = \ln \frac{p(x_i)}{1-p(x_i)} + \varepsilon_i$，这里 ε_i 近似服从正态分布 $N(0, \Delta_i)$，$\Delta_i = \frac{1}{n_i p(x_i)[1-p(x_i)]}$. 自然用 ν_i 估计 Δ_i (ν_i 之定义见(3.5)). 利用(3.2)知

$$z_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad (i = 1, 2, \cdots, m)$$

这里 ε_i 近似服从 $N(0, \nu_i)$. 注意 $\nu_1, \nu_2, \cdots, \nu_m$ 不一定相等. 令 $\tilde{\varepsilon}_i = \frac{1}{\sqrt{\nu_i}} \varepsilon_i (i = 1, 2, \cdots, m)$，则

$$\frac{1}{\sqrt{\nu_i}} z_i = \frac{1}{\sqrt{\nu_i}}(\beta_0 + \beta_1 x_i) + \tilde{\varepsilon}_i$$

这里 $\tilde{\varepsilon}_1, \tilde{\varepsilon}_2, \cdots, \tilde{\varepsilon}_m$ 的方差相等. 仿效通常的最小二乘法的想法，应

找 β_0,β_1 使得平方和 $\sum_{i=1}^{m}\left[\frac{1}{\sqrt{\nu_i}}z_i-\frac{1}{\sqrt{\nu_i}}(\beta_0+\beta_1 x_i)\right]^2$ 达到最小.这个平方和就是上文定义的 $\widetilde{Q}(\beta_0,\beta_1)$.因而使用加权最小二乘估计是有道理的.

例 3.1(社会调查) 一个人在家是否害怕生人来?我们研究人的文化程度对此问题的影响.因变量

$$Y=\begin{cases}1 & \text{害怕}\\ 0 & \text{不害怕}\end{cases}$$

自变量 x 是文化程度,取 4 个可能的值:x_1,x_2,x_3,x_4.这里

$x_1=0$ 表示文盲,$x_2=1$ 表示小学文化程度,$x_3=2$ 表示中学文化程度,$x_4=3$ 表示大专以上(包含大专)文化程度.

根据一项社会调查报告,有下列数据:

自变量(x)	不害怕($Y=0$)人数	害怕($Y=1$)人数
0	11	7
1	45	32
2	664	422
3	168	72

我们可用逻辑斯谛(Logistic)回归模型对上述数据进行统计分析.用 $p(x)$ 表示一个人的文化程度是 x 时害怕生人的概率,即 $p(x)=P(Y=1|x)$.考虑模型(3.2),即

$$\ln\frac{p(x)}{1-p(x)}=\beta_0+\beta_1 x$$

我们用加权最小二乘法估计 β_0,β_1.

根据上面的数据,利用(3.4)和(3.5)可算出:$z_1=-0.384\ 7$,$z_2=-0.326\ 9$,$z_3=-0.451\ 5$,$z_4=-0.842\ 5$,$\nu_1=0.219\ 9$,$\nu_2=0.052\ 7$,$\nu_3=0.003\ 87$,$\nu_4=0.019\ 7$.