3.5 非参数假设检验

在实际问题中，人们往往不知道总体分布的类型，或者知之甚少，而又需要对总体的分布做出某种判断，这是一类不同于参数假设检验的统计推断问题．通常做法是先假定总体服从某种分布，再根据由总体抽取的样本来检验假设，并做出判断．这种以总体的分布形式为假设对象的假设检验称为非参数假设检验.

3.5.1 皮尔逊 χ^2 拟合检验

拟合检验全称拟合优度检验，它反映了实际数据与原假设分布之间拟合的优劣程度．

3.5.1.1 分布的 χ^2 检验法

设总体 X 是仅取 r 个可能值的离散型随机变量，

不失一般性，设其概率函数为

$\mathrm{P}\{X=a_i\}=p_i, i=1,2,\cdots,r$，其中 $\sum_{i=1}^{r} p_i = 1$，

简记为$(p_1, p_2, \cdots, p_r)$.

$(X_1, X_2, \cdots, X_n)$ 是取自总体 X 的样本，$x_1, x_2, \cdots, x_n$ 是样本值，$n_i, i=1,2,\cdots,r$ 表示 $x_1, x_2, \cdots, x_n$ 中取值为 a_i 的个数，即样本中出现事件 $\{X=a_i\}$ 的频数，显然，$n_1, n_2, \cdots, n_r$ 都是样本的函数，所以它们也是随机变量，且有 $\sum_{i=1}^{r} n_i = n$ 和 $(n_1, n_2, \cdots, n_r)$ 服从多项分布，即概率分布为

$$\frac{n!}{n_1!\ n_2!\ \cdots n_r!} p_1^{n_1} p_2^{n_2} \cdots p_r^{n_r} = n! \prod_{i=1}^{r} \frac{p_i^{n_i}}{n_i!}$$

频率是概率的反映，如果总体的概率分布确实是

$$P\{x=a_i\}=p_i, i=1, 2, \cdots, r,$$

当 n 充分大时，实际频数 n_i 与理论频数 np_i 之间的差异将越来越小．

$$\frac{n_i}{n} \to p_i,$$

皮尔逊首先提出用下面的统计量

$$\chi^2 = \sum_{i=1}^{r} \frac{(n_i - np_i)^2}{np_i} \qquad (3.5.1)$$

来衡量它们的差异程度，这个统计量又称为皮尔逊统计量.

如果 $(p_1, p_2, \cdots, p_r)$ 是总体 X 服从的真实分布，$n_i \approx np_i$，统计量 χ^2 值偏小，否则它有偏大的趋势.

定理 3.5.1（皮尔逊定理）

当 $(p_1, p_2, \cdots, p_r)$ 是总体的真实概率分布时，由式 (3.5.1) 定义的统计量 χ^2 的渐近分布是自由度为 $r-1$ 的 χ^2 分布．

当要检验 H_0：$p_i = p_{i0}$，$i=1, 2, \cdots, r$ 时，

对于给定的检验水平 α，如果 $\chi^2 \geqslant \chi^2_{1-\alpha}(r-1)$，拒绝 H_0

$$\chi^2 = \sum_{i=1}^{r} \frac{(n_i - np_{i0})^2}{np_{i0}}.$$

即认为试验结果与原假设有显著差异.

皮尔逊统计量也可用来检验总体是否服从某个给定的分布函数 $F_0(x)$，

设 $(X_1, X_2, \cdots, X_n)$ 是取自分布为 $F(x)$ 的样本，欲检验假设

$$H_0: F(x)=F_0(x)\ [F_0(x)\text{是某个已知的分布}].$$

选取 $r-1$ 个实数 $-\infty<y_1<\cdots<y_{r-1}<\infty$，它们将实数轴分为 r 个区间

$(-\infty, y_1]$，$(y_1, y_2]$，$(y_2, y_3]$，$\cdots$，$(y_{r-1}, +\infty)$，

记
$$\left.\begin{aligned} p_1 &\triangleq F(y_1) \\ p_i &\triangleq F(y_i)-F(y_{i-1}),\ i=2, 3, \cdots, r-1 \\ p_r &\triangleq 1-F(y_{r-1}) \end{aligned}\right\} \quad (3.5.2)$$

用 n_i 表示样本值落在第 i 个区间中的个数，

则 $(n_1, n_2, \cdots, n_r)$ 服从多项分布，

当 $H_0: F(x)=F_0(x)$ 成立时

$$\chi^2 = \sum_{i=1}^{r} \frac{(n_i - np_{i0})^2}{np_{i0}} \quad (3.5.3)$$

的渐近分布为 $\chi^2(r-1)$.因此，可以按上述方法对假设 H_0 进行检验，式(3.5.3)中的 p_{i0} 是将 F_0 代替式(3.5.2)中的 F 算得的 p_i 值.

例 3.5.1 在某盒子中存放有白球和黑球，现作下面这样的试验：用返回抽取方式从盒中摸球，直到摸取的是白球为止，记录下抽取的次数，重复进行如此的试验 100 次，其结果如下：

抽取次数	1	2	3	4	≥5
频数	43	31	15	6	5

试问该盒中的白球与黑球的个数是否相等？$(\alpha=0.05)$

解 设 X 表示首次出现白球所需的摸球次数，则 X 服从几何分布

$$P\{x=k\}=(1-p)^{k-1}p,\ k=1, 2, \cdots$$

其中，p 表示从此盒中任取一球取到白球的概率，

则问题可归结为检验假设

$$H_0: P\{X=k\}=(1-p_0)^{k-1}p_0,\ k=1, 2, \cdots,$$

其中，$p_0=\frac{1}{2}$.

若 H_0 成立，则有 $p_{10}=P\{X=1\}=\frac{1}{2}$，$p_{20}=P\{X=2\}=\frac{1}{4}$，

$$p_{30}=P\{X=3\}=\frac{1}{8},\quad p_{40}=P\{X=4\}=\frac{1}{16},$$

$$p_{50}=P\{X\geqslant 5\}=\sum_{k=5}^{\infty}\frac{1}{2^k}=\frac{1}{16}$$

将 p_{i0} 及已知的实际频数代入式(3.5.3)，得统计量值

$$\chi^2=\frac{(43-50)^2}{50}+\frac{(31-25)^2}{25}+\frac{(15-12.5)^2}{12.5}+\frac{(6-6.25)^2}{6.25}+\frac{(5-6.25)^2}{6.25}$$
$$=3.2$$

查 χ^2 分布表得 $\chi^2_{0.95}(4)=9.488$，由于 $3.2<9.488$，

因此，认为试验结果与假设无显著差异，即认为盒中白球与黑球个数相等.

若 H_0 中给出的理论分布 $F_0(x)$ 含有未知参数.

欲检验假设，$H_0: F(x)=F_0(x; \theta_1, \theta_2, \cdots, \theta_m)$.

$$\chi^2 = \sum_{i=1}^{r} \frac{(n_i - np_{i0})^2}{np_{i0}} \quad (3.5.3)$$

不能计算出 p_{i0}

设 $\hat\theta_1, \hat\theta_2, \cdots, \hat\theta_m$ 分别为 H_0 成立时未知参数 $\theta_1, \theta_2, \cdots, \theta_m$ 的极大似然估计，

$$\text{记 } \hat\theta=(\hat\theta_1, \hat\theta_2, \cdots, \hat\theta_m),$$

$$\hat{p}_{10}=F_0(y_1;\hat{\theta}_1,\cdots,\hat{\theta}_m)$$
$$\hat{p}_{i0}=F_0(y_i;\hat{\theta}_1,\cdots,\hat{\theta}_m)-F_0(y_{i-1};\hat{\theta}_1,\cdots,\hat{\theta}_m),\quad i=2,\cdots,r-1$$
$$\hat{p}_{r0}=1-F_0(y_{r-1};\hat{\theta}_1,\cdots,\hat{\theta}_m)$$

得统计量 $$\chi^2=\sum_{i=1}^{r}\frac{(n_i-n\hat{p}_{i0})^2}{n\hat{p}_{i0}} \qquad (3.5.4)$$

当 n 充分大时，χ^2 统计量的近似分布为

$\chi^2(r-m-1)$

m 为 $F_0(x)$ 中待估参数个数.

例 3.5.2 在例 1.3.2 中，曾经从直方图上粗略地看出总体 X 服从正态分布，现在对此组数据用 χ^2 拟合优度检验.（$\alpha=0.10$）.

解 欲检验假设 H_0：$X\sim N(\mu,\sigma^2)$

先算出 H_0 成立的前提下，μ 和 σ^2 的 MLE 值，

$$\hat{\mu}=\frac{1}{100}\sum_{i=1}^{100}x_i=343.83,$$

$$\hat{\sigma}^2=\frac{1}{100}\sum_{i=1}^{100}(x_i-\overline{x})^2=4.04^2.$$

把实数轴划分成若干区间，一般要求区间的划分尽可能关于 $\overline{x}$ 的值对称，且注意 $n\hat{p}_{i0}$ 不要太小，一般不小于 5，

现在，这批数据最小 $x_{(1)}=332$，最大 $x_{(100)}=358$，

取 $r=6$，区间的划分及计算值如下.

组号	区　间	$\hat{p}_{i0}$	$n\hat{p}_{i0}$	n_i	$(n_i-n\hat{p}_{i0})^2/n\hat{p}_{i0}$
1	$(-\infty,337.5]$	0.0582	5.82	5	0.1155
2	$(337.5,340.5]$	0.1479	14.79	12	0.5263
3	$(340.5,343.5]$	0.2620	26.20	32	1.2840
4	$(343.5,346.5]$	0.2773	27.73	30	0.4853
5	$(346.5,349.5]$	0.1738	17.33	12	1.6654
6	$(349.5,\infty)$	0.0808	8.03	9	0.1048

$$\begin{aligned}\hat{p}_{30}&=P\{340.5<X\leqslant 343.5\}\\&=\Phi\left(\frac{343.5-343.83}{4.04}\right)-\Phi\left(\frac{340.5-343.83}{4.04}\right)\\&=\Phi(-0.08)-\Phi(-0.82)\\&=0.2620\end{aligned}$$

$$\chi^2=\sum_{i=1}^{6}\frac{(n_i-n\hat{p}_{i0})^2}{n\hat{p}_{i0}}=3.8998,$$

查表得 $\chi^2_{0.90}(3)=6.25$. $r=6$，$m=2$，$\alpha=0.10$，

由于 $3.8998<6.25$，因此接受 H_0，

即可以认为罐头内装食品净重服从正态分布.

例 3.5.3 某电话交换台，在 100min 内记录了每分钟被呼叫的次数 x_i，整理后其结果如下

（n_i 是出现呼叫次数 x_i 的次数）

x_i	0	1	2	3	4	5	6	7	8	9
n_i	0	7	12	18	17	20	13	6	3	4

问：可以认为呼叫次数 X 的分布为泊松分布吗？

解 欲检验假设 H_0：$X \sim P(\lambda)$.

λ 未知，在 H_0 成立的条件下，求出 λ 的 MLE 值为

$$\hat{\lambda} = \bar{x} = \frac{1}{100}(1\times 7 + 2\times 12 + \cdots + 9\times 4) = 4.33$$

算出理论概率 $\hat{p}_{i0} = \frac{\hat{\lambda}^i}{i!}e^{-\hat{\lambda}}$，$i=0, 1, 2, \cdots$，

进而算出 $n\hat{p}_{i0}$，列于下表

x_i	0	1	2	3	4	5	6	7	8	≥9
$\hat{p}_{i0}$	0.013	0.057	0.123	0.178	0.193	0.167	0.121	0.074	0.040	0.034
$n\hat{p}_{i0}$	1.3	5.7	12.3	17.8	19.3	16.7	12.1	7.4	4.0	3.4

合并后为 $x\leqslant 1$，$x=2$，…，$x\geqslant 8$，共 8 组，计算得

$$\chi^2 = \sum_{i=1}^{8}\frac{(n_i - n\hat{p}_{i0})^2}{n\hat{p}_{i0}} = \frac{(7-7.0)^2}{7.0} + \frac{(12-12.3)^2}{12.3} + \cdots + \frac{(7-7.4)^2}{7.4}$$

$$= 1.289$$

现在 $r=8$，$m=1$，取 $\alpha=0.05$，查表得 $\chi^2_{0.95}(6)=12.6$.

由于 $1.289 < 12.6$，故接受 H_0，即认为 $X\sim P(\lambda)$.

3.5.1.2 独立性的 χ^2 检验法

假定一个二维总体（X, Y）. 将 X 和 Y 的取值范围分别分成 r 个和 q 个互不相交的区间 A_1，A_2，…，A_r 和 B_1，B_2，…，B_q.

设从总体中抽取一个容量为 n 的样本 (x_1, y_1)，(x_2, y_2)，…，(x_n, y_n)，n_{ij} 表示样本值中 x 落于 A_i，而其 y 落于 B_j 中的个数 $(i=1, 2, \cdots, r;\ j=1, 2, \cdots, q)$.

记 $n_{i\cdot} \triangleq \sum_{j=1}^{q} n_{ij}$，$n_{\cdot j} \triangleq \sum_{i=1}^{r} n_{ij}$. $n = \sum_{i=1}^{r}\sum_{j=1}^{q} n_{ij}$.

X \ Y	B_1	B_2	…	B_q	$n_{i\cdot}$
A_1	n_{11}	n_{12}	…	n_{1q}	$n_{1\cdot}$
A_2	n_{21}	n_{22}	…	n_{2q}	$n_{2\cdot}$
⋮	⋮	⋮	…	⋮	⋮
A_r	n_{r1}	n_{r2}	…	n_{rq}	$n_{r\cdot}$
$n_{\cdot j}$	$n_{\cdot 1}$	$n_{\cdot 2}$	…	$n_{\cdot q}$	n

提出假设 H_0：总体（X, Y）的分量 X 和 Y 相互独立.

记 $p_{ij} \triangleq P\{X\in A_i, Y\in B_j\}$，

$p_{i\cdot} \triangleq P\{X\in A_i\}$，

$p_{\cdot j} \triangleq P\{Y\in B_j\}$，

$i=1, 2, \cdots, r;\ j=1, 2, \cdots, q$

$$p_{i\cdot} = \sum_{j=1}^{q} p_{ij}$$

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

$$\sum_{i=1}^{r} p_{i\cdot} = \sum_{j=1}^{q} p_{\cdot j} = 1$$

在 H_0 成立的条件下，有 $p_{ij} = p_{i\cdot}\, p_{\cdot j}$，

即要检验 H_0：$p_{ij} = p_{i\cdot}\, p_{\cdot j}$，$i=1, 2, \cdots, r$；
$j=1, 2, \cdots, q$，

在 H_0 成立的条件下，$p_{i\cdot}$ 和 $p_{\cdot j}$ 的极大似然估计分别为

$$\hat{p}_{i\cdot} = n_{i\cdot}/n, i=1, 2, \cdots, r;$$

$$\hat{p}_{\cdot j} = n_{\cdot j}/n, j=1, 2, \cdots, q$$

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

当 H_0 成立时，χ^2 统计量的分布渐近于自由度为 $(r-1)(q-1)$ 的 χ^2 分布.

$$rq-(r+q-2)-1=(r-1)(q-1).$$

设 $(X_1, Y_1), (X_2, Y_2), \cdots, (X_n, Y_n)$ 为取自总体 (X, Y) 的样本，检验假设

$$H_0: X \text{ 与 } Y \text{ 相互独立}.$$

计算步骤如下.

(1) 将 X 和 Y 的观测值范围分别分成 r 个和 q 个互不相交的区间，这样就组成了 rq 个互不相交的矩形区域.

(2) 求出样本值落入各小矩形的实测频数.

(3) 当 H_0 成立时，构造统计量.

$$\chi^2 = n\sum_{i=1}^{r}\sum_{j=1}^{q}\frac{(n_{ij} - n_{i\cdot}n_{\cdot j}/n)^2}{n_{i\cdot}n_{\cdot j}}$$

当 n 充分大时，χ^2 渐近于 $\chi^2[(r-1)(q-1)]$ 分布.

在检验水平 α 下，

当 χ^2 值 $\geqslant \chi^2_{1-\alpha}[(r-1)(q-1)]$ 时，

拒绝 H_0，否则接受 H_0.

例 3.5.4 某研究所推出一种感冒特效新药，为证明其疗效，选择 200 名患者做疗效试验，将他们分为两组，分别不服药或服药，观察 3 日后痊愈的情况，得出下列数据.

Y / X	痊愈者	未痊愈者	合计
未服药者	48	52	100
服药者	56	44	100
合计	104	96	200

问新药是否确有明显疗效？（$\alpha = 0.25$）

解 H_0：新药无明显疗效

每个患者考察两个指标：X 表示是否服药，

Y 表示是否痊愈；

X 取两个“值”（即服药状况）：未服药、服药，

Y 也取两个“值”（即痊愈状况）：痊愈、未痊愈.

新药是否有明显疗效的问题，实际上就是服不服这种药是否影响到患者的痊愈，因而就是 X 与 Y 之间是否相互独立的问题.

Y / X	痊愈者	未痊愈者	合计
未服药者	48	52	100
服药者	56	44	100
合计	104	96	200

$n = 200$，$n_{11} = 48$，$n_{12} = 52$，$n_{21} = 56$，$n_{22} = 44$，

$$n_{1\cdot} = n_{\cdot 2} = 100, \quad n_{\cdot 1} = 104, \quad n_{\cdot 2} = 96$$

$$\chi^2 = n\sum_{i=1}^{r}\sum_{j=1}^{q}\frac{(n_{ij} - n_{i\cdot}n_{\cdot j}/n)^2}{n_{i\cdot}n_{\cdot j}} \approx 1.282;$$

对于 $\alpha = 0.25$，查表得 $\chi^2_{0.75}(1) = 1.323$，

由于 $1.323 > 1.282$. 故接受 H_0，即认为这种感冒新药并无明显疗效.

作业

P91: 20, 21, 23,25