第五章　假 设 检 验

5.1　问题提法和基本概念

5.1.1　例子与问题提法

假设检验的概念在第四章4.1节中就曾提到了.这里我们先通过对几个常用例子的分析,总结出假设检验问题提法的形式.然后在这个基础上,引进关于假设检验的一些基本概念.

例1.1　在4.1节中我们曾提到一个在元件寿命服从指数分布的假定下,通过对抽出的若干个元件进行测试所得的数据(样本),去判定"元件平均寿命不小于5000小时"是否成立的问题.

我们把与这个问题有关的事项,用统计学的语言清楚列出如下:

1. 我们有一个总体,即所考察的那一大批元件的寿命.我们对总体分布作了一个假定,即它服从指数分布(第二章(1.20)式),该分布包含了一个未知参数 λ.

2. 我们有从该总体中抽出的样本 $X_1,\cdots,X_n$(即抽出的那 n 个元件测试出的寿命).

3. 我们有一个命题,其正确与否完全取决于未知参数 λ 之值,即"$1/\lambda \geqslant 5000$".它把参数 λ 的所有可能取的值 $0<\lambda<\infty$ 分成两部分:一部分是 $H_0=\{\lambda:\lambda \leqslant 1/5000\}$,一部分是 $H_1=\{\lambda:\lambda > 1/5000\}$. H_0 内的 λ 值使上述命题成立,而 H_1 内的 λ 值则使上述命题不成立.故我们的命题可记为:"λ 属于 H_0"或用符号写为"$\lambda \in H_0$",以至简记为"H_0".

4. 我们的任务是利用所获得的样本 $X_1,\cdots,X_n$,去判断命题"$\lambda \in H_0$"是否成立.其所以能这么做,当然是因为样本中包含了总体分布的信息,也就包含了"$\lambda \in H_0$"是否成立的信息.

在数理统计学上，把类似于上述“$\lambda \in H_0$”这种命题称为一个“假设”或“统计假设”.“假设”这个词在此就是一个其正确与否有待通过样本去判断的陈述.不要把它和通常意义相混.例如在数学上常说“假设某函数处处连续”之类的话，那是一个所讨论的问题中已被承认的前提或条件，与此处所讲的完全不同.

在数理统计学上，通用“检验”一词来代替上文的“判断”.检验一词有动词名词两种含义.动词含义是指判断全过程的操作，而名词的含义是指判断准则.例如，就本例而言一个看来合理的判断准则是：“当 $\overline{X} \geqslant C$ 时认为假设 $\lambda \in H_0$ 正确，不然就认为它不正确”（C 是一个适当的常数，以后再谈）.这就是一个检验（名词）.“认为假设正确”在统计上称为接受该假设；“认为假设不正确”在统计上称为否定或拒绝该假设.到此为止统计问题可以说已完成了：至于接受或否定假设以后如何办（如在本例中，若认为 $\lambda \geqslant 1/5000$ 不成立，该如何处理）？这不是我们要考虑的事.

以下几例的解释都与上述过程完全平行.

例 1.2 有人给我一根金条，他说其重量为 312.5 克.我现在拿到一架精密天平上重复秤 n 次，得出结果为 $X_1,\cdots,X_n$.我假定此天平上秤出的结果服从正态分布 $N(\mu,\sigma^2)$（这是一个假定，它已被承认，不是检验对象）.这时，我要检验的假设为：“$\mu = 312.5$”.在本例 σ 可以已知或未知.如果 σ 未知，则总体分布含多个参数，但假设可以只涉及其中一个.问题也可以是检验方差（当然，在方差 σ^2 未知时）：比如，人家告诉我这天平的误差方差为 $10^{-4}(g^2)$，我怀疑它是否如此，这时我可以拿一个物件在该天平上秤 n 次得 $X_1,\cdots,X_n$.利用这些数据去检验假设“$\sigma^2 = 10^{-4}$”.仍假定总体为正态分布 $N(\mu,\sigma^2)$，μ 就是那个物件之重，它可以已知（例如你拿一个其重量已经测定的物体去秤），也可以未知.

例 1.3 某工厂一种产品的一项质量指标假定服从正态分布 $N(\mu_1,\sigma^2)$.现在对其制造工艺作了若干变化，人们说结果质量起了变化或有了改进.我想通过样本来检验一下.

假定修改工艺后，质量指标仍服从正态分布，且只均值可能有

变而方差不变，即分布为 $N(\mu_2,\sigma^2)$. 我把要检验的假设定为

$$H_0:\{\mu_1 = \mu_2\}$$

或(设均值大时，质量为优)

$$H_0':\{\mu_1 \geqslant \mu_2\}$$

这要仔细解释一下.

选 H_0，是针对“质量起了变化”的说法. 由于你不能凭空说 μ_2 不等于 μ_1，我就先作假设 H_0. 如果经过检验 H_0 被否定了，则我承认质量起了变化，不然就只好仍维持 H_0. 自然，你可能辨驳说，为何不取 $\{\mu_1 \neq \mu_2\}$ 作为假设去检验? 这不也一样：你接受了它，即为质量确有变化；若否定了它，则认为无变化. 从表面上看这个提法无可非议，因为两种提法从实质上看只是表述方式不同. 但有其不可这样做的理由，这一点在以后将予以解释，现在还说不清楚.

选 H_0' 是针对“质量有了改进”的说法，与上文类似.

本例中 σ^2 可以是已知或未知，在应用上以未知的情况居多. 又“工艺变化前后质量方差一样”是一个多少有点人为的假定(一般，质量的改进也常反映在其波动变小上，即方差会小些). 如假定前后方差不一样，则得到贝伦-费希尔检验问题，这是数理统计学上的一个著名的问题，其区间估计形式已在前章讲过了.

如果不认为质量的平均值有多大问题，而问题在其方差上，则假定在工艺改变前后，质量指标的分布分别为 $N(\mu_1,\sigma_1^2)$ 和 $N(\mu_2,\sigma_2^2)$. 这时要检验的假设可以是“$\sigma_1^2=\sigma_2^2$”或“$\sigma_1^2 \leqslant \sigma_2^2$”.

例 1.4 甲、乙两位棋手下棋. 共下 n 局，甲 m 胜 $n-m$ 负(设无和局). 根据这一结果对两位棋手的技艺是否有差别下一个判断.

若以 p 记每局中甲胜的概率，则乙胜的概率为 $1-p$. 假定每局的结果独立(这很接近事实，除非其中一位或两位的心理素质差，以致已赛各局的结果显著地影响着他的情绪). 则若以 X 记在 n 局中甲胜的局数，将有 $X \sim B(n,p)$. 我们的问题可提为：检验

假设“$p=1/2$”.

例 1.5 有一颗供赌博或其他用途的骰子，怀疑它是否均匀，要用投掷若干次的结果去检验它. 若以掷出点数的概率分布来表示，所要检验的内容可表为假设

$$H_0: p_1 = p_2 = \cdots = p_6 = 1/6 \tag{1.1}$$

这里 p_i 是骰子掷出 i 点的概率. 这意味着把骰子的均匀性解释为：它掷出任何一点的机会都相同.

从以上诸例我们明确了假设检验问题的提法. 现在介绍假设检验中几个常用的名词.

1. 原假设和对立假设

在假设检验中，常把一个被检验的假设叫做原假设，而其对立面就叫做对立假设. 如在例 1.1 中，原假设为 $H_0: \lambda \leqslant 1/5000$，故对立假设为 $H_1: \lambda > 1/5000$. 在例 1.5 中，原假设 H_0 为(1.1)，而对立假设为 H_1：“$p_1, \cdots, p_6$ 不完全相同”.

原假设中的“原”字，字面上可解释为“原本有的”. 如在例 1.2，你可以说 $\mu = 312.5$ 原本就不存在问题，只因有人怀疑，才提出了也存在 $\mu \neq 312.5$ 的可能. $\mu = 312.5$ 是“原有”的而 $\mu \neq 312.5$ 是“后来的”. 这样的解释也并非处处适合(见下)，的确，对这个“原”字不必硬加一种解释.

原假设又常称为“零假设”或“解消假设”*. 这名词的含义拿例 1.3 中的假设 $H_0: \mu_1 = \mu_2$ 去看最贴切. 因为，$\mu_1 - \mu_2$ 反映工艺变化后所产生的效应. 你这个假设 H_0 把这个效应化为零了，或把这个效应“解消”了. 不难理解：在有些情况下这个名词也并非很贴切，故也有不少人不高兴用这名称.

对立假设就是与原假设对立的意思. 这个词既可以指全体，也可以指一个或一些特殊情况，例如对例 1.1，我们说对立假设是 $\lambda > 1/5000$，这是指全体. 但也可以说 $\lambda = 1.5$ 是一个对立假设，这无非是指 1.5 这个值是对立假设的一个成员. 对立假设也常称为

---

* 零假设或解消假设都是从英语 Null Hypothesis 一词而来.

"备择假设",其含义是:在抛弃原假设后可供选择的假设.

2. 检验统计量、接受域、否定域、临界域和临界值

在检验一个假设时所使用的统计量称为检验统计量. 拿例 1.1 来说,我们前已提到了一个在直观上合理的检验:当 $\overline{X}\geqslant C$ 时接受原假设,不然就否定. 这里用的检验统计量是 $\overline{X}$.

使原假设得到接受的那些样本$(X_1,\cdots,X_n)$所在的区域 A,称为该检验的接受域,而使原假设被否定的那些样本所成的区域 R,则称为该检验的否定域. 否定域有时也称为拒绝域,临界域. 如在例 1.1 中,刚才所提到的检验的接受域为

$$A=\{(X_1,\cdots,X_n):X_1+\cdots+X_n\geqslant nC\}$$

否定域为

$$R=\{(X_1,\cdots,X_n):X_1+\cdots+X_n< nC\}$$

A 与 R 互补,知其一即知其二. 定一个检验,等价于指定其接受域或否定域.

在上述检验中,C 这个值处于一个特殊的地位:$\overline{X}$ 之值一越过 C 这个界线,结论就由接受变为否定. 这个值 C 称为检验统计量 $\overline{X}$ 的临界值. 当心中明确了用什么统计量时,也可以说"检验的临界值". 例如,若心中已明确用统计量 $X_1+\cdots+X_n$,则临界值为 nC. 也可以有不止一个临界值. 如在例 1.1,若要检验的原假设改为"$\lambda=1/5000$",则一个合理的检验法是:当 $C_1\leqslant\overline{X}\leqslant C_2$ 时,接受,不然就否定. C_1,C_2 是两个适当选定的常数,它们都是临界值.

3. 简单假设和复合假设

不论是原假设还是对立假设,若其中只含一个参数值,则称为简单假设,否则就称复合假设.

如在例 1.1 中,原假设 $\lambda\leqslant1/5000$ 包含所有大于 0 而不超过 1/5000的 λ 值,它是复合的;对立假设 $\lambda>1/5000$ 也为复合. 再看例 1.2. 若 σ^2 已知,则原假设只含参数 μ 的一个值 312.5,故是一个简单假设;若 σ^2 未知,则原假设包含了所有形如

$$(312.5,\sigma^2):\sigma^2\ 任意$$

的参数值，故是复合的. 这里的要点是：在决定一个假设是简单还是复合时，要考虑到总体分布中的一切参数，而不止是直接出现在假设中的那部分参数. 如在本例，σ^2 虽则不出现在假设中，但因为它是总体分布的未知参数，故仍要考虑进来. 这种参数(如此处的 σ^2)在数理统计学上称赘余参数* 在区间估计中这个名词也常提到，例如在正态总体 $N(\mu,\sigma^2)$中，μ,σ^2 都未知，要作 μ 的区间估计，这时 σ^2 就是赘余参数.

5.1.2 功效函数

功效函数是假设检验中最重要的概念之一. 在以下将看到：同一个原假设可以有许多检验法，其中自然有优劣之分. 这区分的依据，就取决于检验的功效函数.

例 1.6 再考虑例 1.1，并设我们取定了如下的检验 Φ：

$$\Phi\text{：当 } \overline{X} \geqslant C \text{ 时接受，不然就否定} \tag{1.2}$$

如果我们使用这个检验，则原假设 $H_0:\lambda \leqslant 1/5000$ 被接受或否定，都是随机事件，因为其发生与否，要看样本 $X_1,\cdots,X_n$ 如何，而样本是随机的. 在此，原假设被否定的概率为

$$\beta_\Phi(\lambda) = P_\lambda(\overline{X} < C)$$

P_λ 的意义以前解释过，它是指事件$\{\overline{X} < C\}$的概率，是在总体分布的参数值为 λ 时去计算的. 因为(见第二章例 4.9)$2\lambda(X_1+\cdots+X_n)\sim\chi^2_{2n}$，故如以 K_{2n}记χ^2_{2n}的分布函数，则有

$$\begin{aligned}\beta_\Phi(\lambda) &= P_\lambda(X_1+\cdots+X_n < nC)\\ &= P_\lambda(2\lambda(X_1+\cdots+X_n) < 2\lambda nC)\\ &= K_{2n}(2\lambda nC)\end{aligned} \tag{1.3}$$

其值与 λ 有关，且随 λ 上升而增加. 因为 λ 愈大，离开原假设 $\lambda\leqslant 1/5000$ 就愈远，一个合理的检验法就应当用更大的概率去否定

---

* 英语 Nuisance Parameter. 也有译为“多余参数”或“讨厌参数”的，含有使问题复杂化的意味.

它.

函数(1.3)就称为检验(1.2)的功效函数.由此,提出下面一般定义:

定义 1.1 设总体分布包含若干个未知参数 $\theta_1,\cdots,\theta_k$. H_0 是关于这些参数的一个原假设,设有了样本 $X_1,\cdots,X_n$,而 Φ 是基于这些样本而对 H_0 所作的一个检验.则称检验 Φ 的功效函数为

$$\beta_\Phi(\theta_1,\cdots,\theta_k) = P_{\theta_1,\cdots,\theta_k}(\text{在检验 }\Phi\text{ 之下},H_0\text{ 被否定}) \tag{1.4}$$

它是未知参数 $\theta_1,\cdots,\theta_k$ 的函数.

容易明白:当某一特定参数值$(\theta_1^0,\cdots,\theta_k^0)$使 H_0 成立时,我们希望 $\beta_\Phi(\theta_1^0,\cdots,\theta_k^0)$尽量小(当 H_0 成立时我们不希望否定它).反之,若$(\theta_1^0,\cdots,\theta_k^0)$属于对立假设,则我们希望 $\beta_\Phi(\theta_1^0,\cdots,\theta_k^0)$尽量大(当 H_0 不成立时我们希望否定它).两个检验 Φ_1,Φ_2(同一个原假设的)哪一个更好地符合了这个要求,哪一个就更好.

由于当$(\theta_1,\cdots,\theta_k)$属于对立假设时,我们希望功效函数值 $\beta_\Phi(\theta_1,\cdots,\theta_k)$尽可能大,故在$(\theta_1,\cdots,\theta_k)$属于对立假设时,称 $\beta_\Phi(\theta_1,\cdots,\theta_k)$为检验 Φ 在$(\theta_1,\cdots,\theta_k)$处的"功效".这称呼只用于对立假设处.因为,当$(\theta_1,\cdots,\theta_k)$属于原假设时,$\beta_\Phi(\theta_1,\cdots,\theta_k)$以小为好,这时称它为"功效"就不合情理了.

5.1.3 两类错误、检验的水平

在检验一个假设 H_0 时,有可能犯以下两类(或两种)错误之一:1. H_0 正确,但被否定了;2. H_0 不正确,但被接受了.可能犯哪一类错误,要视总体分布中有关的参数值而定.如在例 1.1 中,若参数 λ 之值为 0.0001,则我们只可能犯第一种错误,而当 $\lambda=0.1$ 时,则只可能犯第二种错误.

若以 $\theta_1,\cdots,\theta_k$ 记总体分布的参数,$\beta_\Phi(\theta_1,\cdots,\theta_k)$记检验 Φ 的功效函数,则犯第一、二类错误的概率 $\alpha_{1\Phi}(\theta_1,\cdots,\theta_k)$和 $\alpha_{2\Phi}(\theta_1,\cdots,\theta_k)$,分别为

$$\alpha_{1\Phi}(\theta_1,\cdots,\theta_k)=\begin{cases}\beta_o(\theta_1,\cdots,\theta_k), & 当(\theta_1,\cdots,\theta_k)\in H_0\\ 0, & 当(\theta_1,\cdots,\theta_k)\in H_1\end{cases} \tag{1.5}$$

$$\alpha_{2\Phi}(\theta_1,\cdots,\theta_k)=\begin{cases}0, & 当(\theta_1,\cdots,\theta_k)\in H_0\\ 1-\beta_\Phi(\theta_1,\cdots,\theta_k), & 当(\theta_1,\cdots,\theta_k)\in H_1\end{cases} \tag{1.6}$$

这里 H_1 是对立假设.

在检验一个假设 H_0(对立假设 H_1)时,我们希望犯两种错误的概率都尽量小.看表达式(1.5)和(1.6),即得出我们在上段中已提到过的结论,即在选择一个检验 Φ 时,要使其功效函数 β_Φ 在 H_0 上尽量小而在 H_1 上尽量大.但这两方面的要求是矛盾的.正好像在区间估计中,你要想增大可靠性即置信系数,就会使区间长度变大而降低精度,反之亦然.在区间估计理论中,是用"保一望二"的原则解决了这个问题,即使置信系数达到指定值,在这限制之下使区间精度尽可能大.在假设检验中也是这样办:先保证第一类错误的概率不超过某指定值 α(α 通常较小,最常用的是 $\alpha=0.05$ 和 0.01,有时也用到 $0.001,0.10$,以至 0.20 等值),再在这限制下,使第二类错误概率尽可能小.

定义 1.2 设 Φ 是原假设 H_0 的一个检验,$\beta_\Phi(\theta_1,\cdots,\theta_k)$ 为其功效函数,α 为常数,$0\leqslant\alpha\leqslant 1$.如果

$$\beta_\Phi(\theta_1,\cdots,\theta_k)\leqslant\alpha, 对任何(\theta_1,\cdots,\theta_k)\in H_0 \tag{1.7}$$

则称 Φ 为 H_0 的一个水平 α 的检验,或者说,检验 Φ 的水平为 α,检验 Φ 有水平 α.

显然,若 α 为 Φ 的水平而 $\alpha_1>\alpha$,则 α_1 也是检验的水平.这样,一检验的水平并不唯一.为克服这点不方便之处,通常只要可能,就取最小可能的水平作为检验的水平.不少著作中就直接把水平定义为满足(1.7)式的最小的 α.这样做,唯一性的问题解决了,固然是好,但也有其不便之处,即有时我们只知道(1.7)成立,而无法证明 α 已达到最小,这时就不能称 α 为 Φ 的水平,不好如何称呼.因此,我们维持定义 1.2,但有这样一个默契:只要可能,尽量找最小的 α.

以上所说的叫做"固定(或限制)第一类错误概率的原则",是目前假设检验理论中一种流行的做法.你可以问:为什么不固定第二类错误概率而在这个前提下尽量减小第一类错误的概率?回答是:你这么做并非不可以,但是,大家约定统一在一个原则下,讨论问题比较方便些,这还不是主要理由.从实用的观点看,确实,在多数假设检验问题中,第一类错误被认为更有害,更需要控制.这一点将结合下一节中实例的讨论再作说明.也有些情况,确实第二类错误的为害更大,这时有必要控制这个概率,换句话说,"控制第一类错误概率"的原则也并非绝对的,可视情况的需要而变通之.

5.1.4 一致最优检验

定义 1.3 沿用定义 1.2 的记号.设 Φ 为一个水平 α 的检验,即满足(1.7).若对任何其他一个水平 α 的检验 g,必有

$$\beta_{\Phi}(\theta_1,\cdots,\theta_k) \geqslant \beta_g(\theta_1,\cdots,\theta_k),\text{对任何}(\theta_1,\cdots,\theta_k) \in H_1 \tag{1.8}$$

这里 H_1 为对立假设.则称 Φ 是假设检验问题 $H_0:H_1$ 的一个水平 α 的一致最优检验.

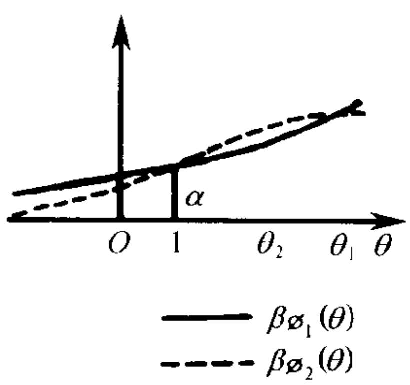

图 5.1

简单地说,水平 α 一致最优检验,就是在一切水平 α 的检验中,其功效函数在对立假设 H_1 上处处达到最大者.或者说,是在一切其第一类错误概率不超过 α 的检验中,第二类错误概率处处达到最小者.难就难在"处处"这两个字."一致最优"中的"一致",就是指这个"处处"而言.就拿两个检验 Φ_1 和 Φ_2 的比较来谈.为清楚计,不妨设原假设 H_0 为 $\theta \leqslant 1$,对立假设 H_1 为 $\theta > 1$.设 Φ_1 和 Φ_2 都是水平 α 的检验,其功效函数分别如图 5.1 中的实线和虚线所示.在对立假设 θ_1 处,β_{Φ_1} 大于 β_{Φ_2}.而在 θ_2 处则是 β_{Φ_2} 大于 β_{Φ_1}.故在

这两个检验 Φ_1, Φ_2 中,没有一个在对立假设各点处处优于另一个.由于水平 α 的检验非常多,其中能有一个一致最优者,就不是常见的情况而是较少有的例外,更确定地说,只在总体分布只依赖一个参数 θ,而原假设 H_0 是 $\theta \leqslant \theta_0$ 或 H_0 是 $\theta \geqslant \theta_0$ 的情形,且对总体分布的形式有一定的限制时,一致最优检验才存在.其他情况则是稀有的例外.在下节我们讨论一些具体检验时,将指明哪些是一致最优检验.有的情况的证明将在本章附录中给出.

由于一致最优的条件太高,在假设检验理论中也引入了另一些优良准则.这些都超出了本课程的范围之外,不能在此介绍了.

本节所讲的假设检验理论的基本概念,特别是限制第一类错误概率的原则及一致最优检验等,是 J. 奈曼(前在区间估计一节中已提到)和英国统计学家 E. S. 皮尔逊(K. 皮尔逊的儿子)合作,自 1928 年起开始引进的.基于这些概念所发展的假设检验理论,一般称之为奈曼-皮尔逊理论.从统计学的历史看,最早引进假设检验并对之作了重要贡献的统计学家,还要算我们以前多次提到过的 K. 皮尔逊和 R. A. 费歇尔.皮尔逊的工作将在本章 5.3 节中介绍.