贝叶斯统计的算例

华东理工大学机械与动力工程学院

一、英国数学家托马斯·贝叶斯（Thomas Bayes）

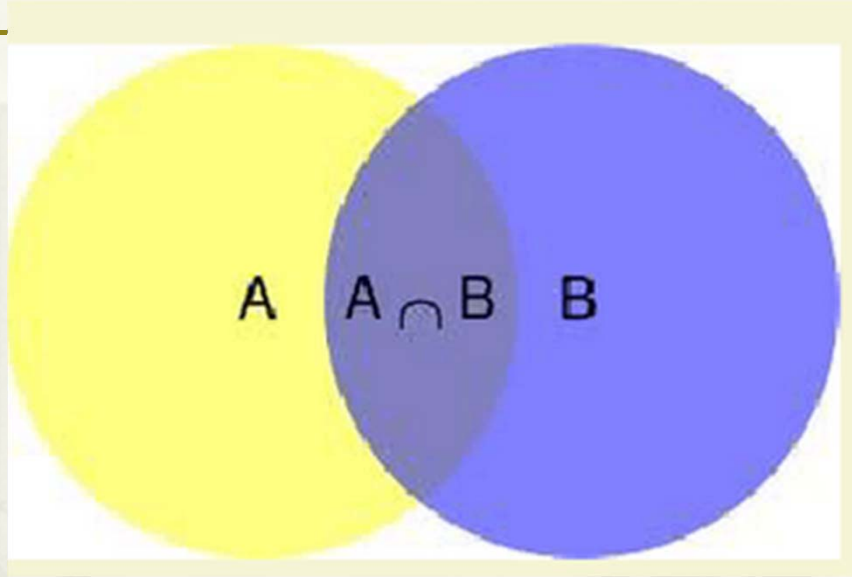
在1763年发表的一篇论文中，首先提出了这个定理。

实际上就是计算"条件概率"的公式。所谓"条件概率"（Conditional probability），就是指在事件B发生的情况下，事件A发生的概率，用P(A|B)来表示。

贝叶斯定理

根据文氏图，可以很清楚地看到在事件B

* 发生的情况下，事件A发生的概率就是P(A∩B)除以P(B)。

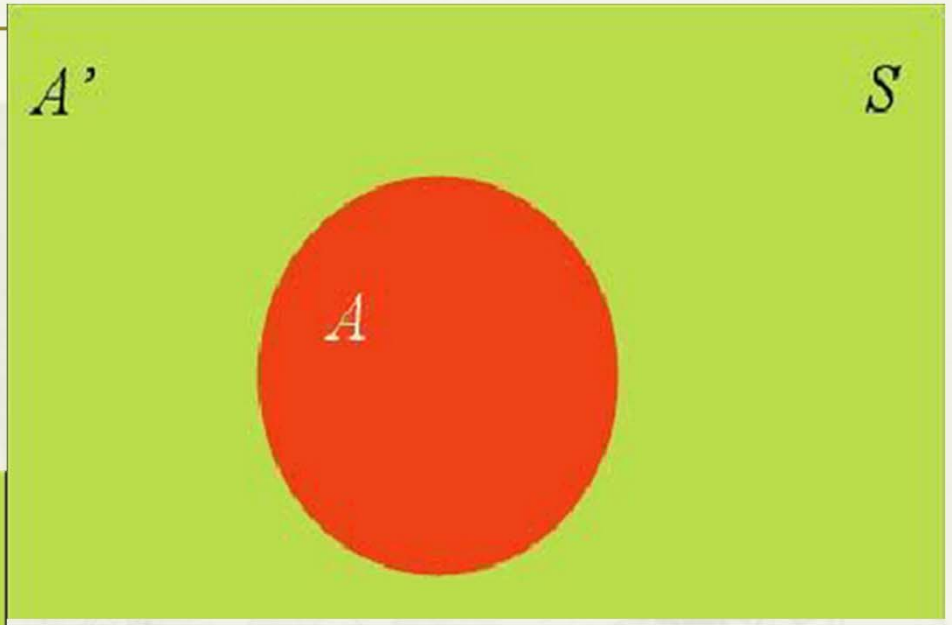


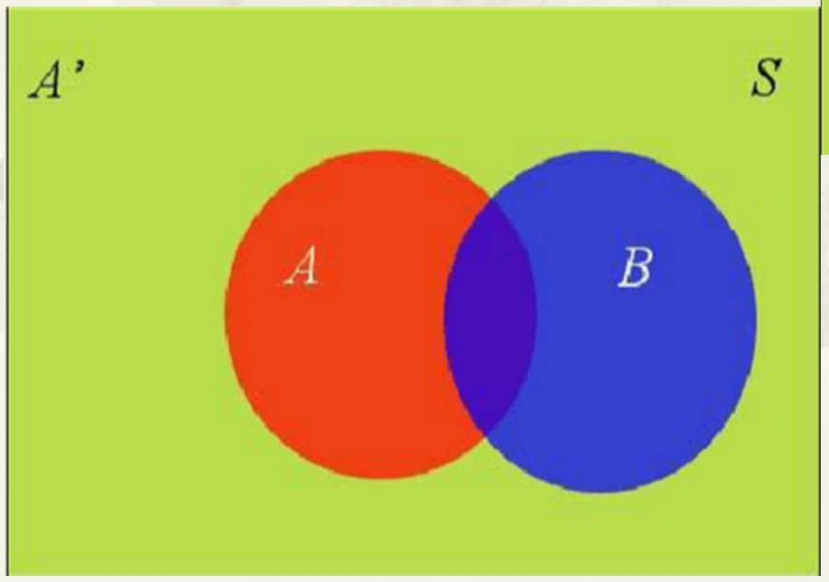
$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

二、全概率公式

假定样本空间S，是两个事件A与A’的和。





全概率公式

它的含义是，如果A和A’构成样本空间的一个划分，那么事件B的概率，就等于A和A’的概率分别乘以B对这两个事件的条件概率之和。

$$P(B)=P(B|A)P(A)+P(B|A')P(A')$$

贝叶斯推断的含义

我们把P(A)称为"先验概率"（Prior probability），即在B事件发生之前，我们对A事件概率的一个判断。P(A|B)称为"后验概率"（Posterior probability），即在B事件发生之后，我们对A事件概率的重新评估。P(B|A)/P(B)称为"可能性函数"（Likelyhood），这是一个调整因子，使得预估概率更接近真实概率。

所以，条件概率可以理解成下面的式子：

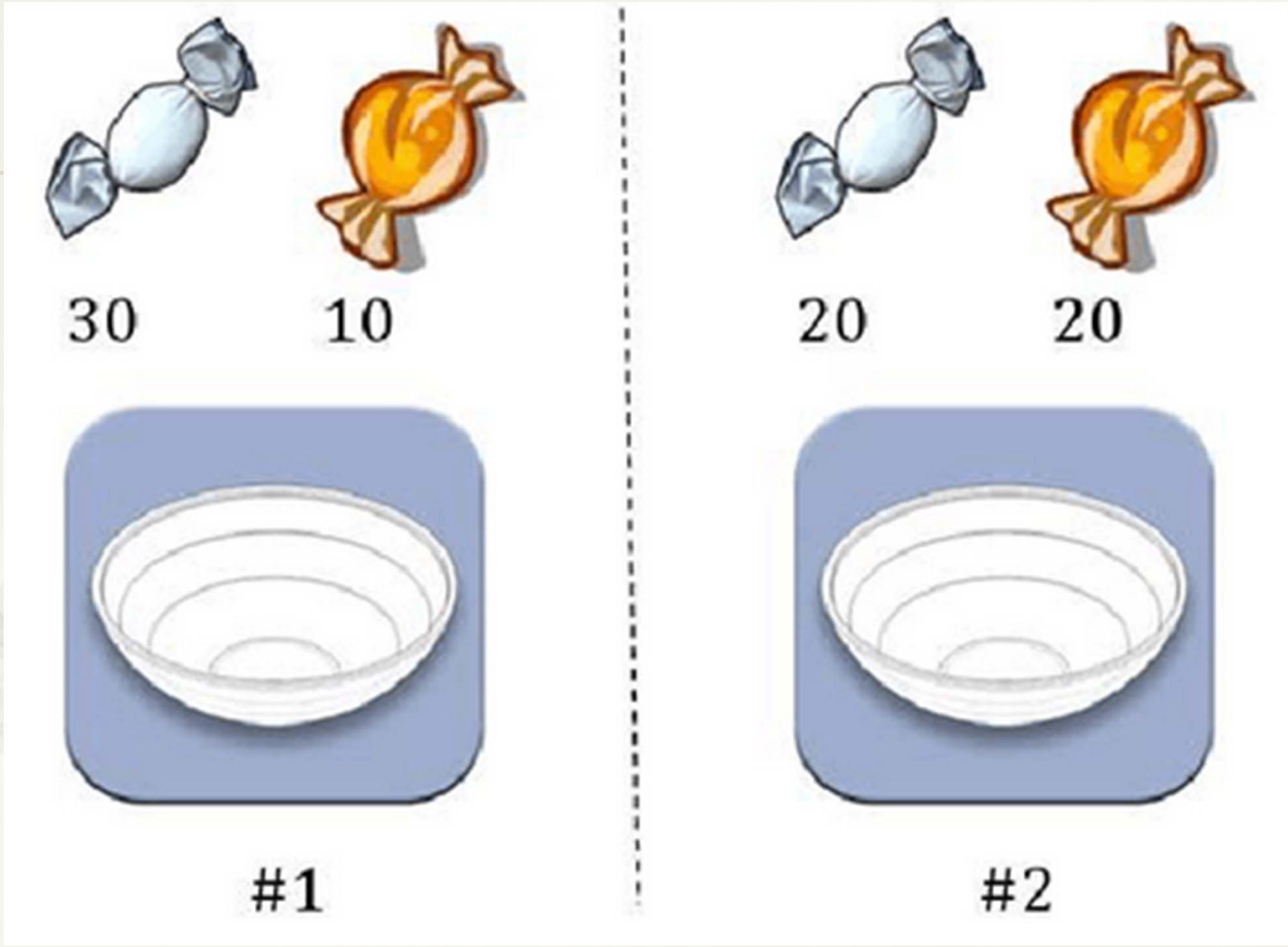
后验概率 ＝ 先验概率 x 调整因子

这就是贝叶斯推断的含义。我们先预估一个″先验概率″，然后加入实验结果，看这个实验到底是增强还是削弱了″先验概率″，由此得到更接近事实的″后验概率″。

【例子】水果糖问题

第一个例子。两个一模一样的碗，一号碗有30颗水果糖和10颗巧克力糖，二号碗有水果糖和巧克力糖各20颗。

现在随机选择一个碗，从中摸出一颗糖，发现是水果糖。请问这颗水果糖来自一号碗的概率有多大？


30
10
20
20
#1
#2

解　水果糖问题

我们假定，H1表示一号碗，H2表示二号碗。由于这两个碗是一样的，

所以P(H1)=P(H2)，也就是说，在取出水果糖之前，这两个碗被选中的概率相同。因此，P(H1)=0.5，我们把这个概率就叫做"先验概率"，即没有做实验之前，来自一号碗的概率是0.5。

解　水果糖问题

再假定，E表示水果糖，所以问题就变成了在已知 E的情况下，来自一号碗的概率有多大，即求P(H1|E)。我们把这个概率叫做"后验概率"，即在E事件发生之后，对P(H1)的修正。

根据条件概率公式，得到

$$P(H_1|E) = P(H_1)\frac{P(E|H_1)}{P(E)}$$

解　水果糖问题

已知，P(H1)等于0.5，P(E|H1)为一号碗中取出水果糖的概率，等于0.75，那么求出P(E)就可以得到答案。根据全概率公式，

$$P(E) = 0.75 \times 0.5 + 0.5 \times 0.5 = 0.625$$

将数字代入原方程，得到

$$P(H1|E) = 0.5 \times \frac{0.75}{0.625} = 0.6$$

【例】水果糖问题

* 这表明，来自一号碗的概率是0.6。也就是说，取出水果糖之后，H1事件的可能性得到了增强。

贝叶斯（Bayes）公式涉及两事件的条件概率计算

$$P(A|B)=\frac{P(A\cap B)}{P(B)}=\frac{P(B\cap A)}{P\left[(B\cap A)\cup(B\cap A^c)\right]}$$

$$=\frac{P(B|A)P(A)}{P(B|A)P(A)+P(B|A^c)P(A^c)}$$

其中，事件$B\cap A$和$B\cap A^c$是互斥的。

贝叶斯分析的简单算例1：

* 为了提高某产品的质量，公司经理考虑投资来改进生产设备，预计投资90万元，但从投资效果看，下属部门有两种意见：

θ_1：改进生产设备后，高质量产品可占90%；

θ_2：改进生产设备后，高质量产品可占70%。

经理当然希望θ_1发生，公司效益可得到很大提高，投资改进设备也合算。但是根据下属二个部门过去建议被采纳的情况，

经理认为，θ_1可信程度只有40%，

θ_2的可信程度是60%。

* 即
* $P(\theta_1)=0.4, \quad P(\theta_2)=0.6$

为了慎重起见，做了一个小规模试验，试验结果记为：

A：试制五个产品，全是高质量的产品。

* 通过计算可得到：

$P(A|\theta_1) = 0.590,$

$P(A|\theta_2) = 0.168。$

* 根据全概率公式计算得到：
* $P(A) = P(A|\theta_1)P(\theta_1) + P(A|\theta_2)P(\theta_2) = 0.337$。
*
* 可算出：

$$P(\theta_1|A) = P(A|\theta_1)P(\theta_1)/P(A) = 0.700,$$

*

$$P(\theta_1|A) = P(A|\theta_2)P(\theta_2)/P(A) = 0.300。$$

简单算例1结论

- 这表明，经理根据试验A的信息调整自己的看法，把对θ_1和θ_2的可信程度由0.4和0.6调整到0.7和0.3。
- 后者是综合了经理的主观概率和试验结果而获得的，要比主观概率更有吸引力，更贴合实际，这就是贝叶斯公式的应用。

简单算例2：

对一个烟雾探测器进行常规检查，故障探测器中有80%经历过一次用电高峰，完好探测器中有10%经历过一次用电高峰。

已知有20%的探测器已经出故障，那么经历一次用电高峰后，探测器出故障的概率是多少？

解：

- 令A=探测器出故障，
- B=探测器经历过用电高峰，
- 那么
- $P(A)=0.20$, $P(B|A)=0.80$,
- $P(B|A^c)=0.10$。

因此：

$$P\left(A\middle|B\right)=\frac{0.80\times0.20}{0.80\times0.20+0.10\times0.80}=\frac{0.16}{0.24}=0.667$$

$$P\left(A\middle|B\right)=\frac{P\left(B\middle|A\right)P\left(A\right)}{P\left(B\middle|A\right)P\left(A\right)+P\left(B\middle|A^{c}\right)P\left(A^{c}\right)}$$

令A=探测器出故障，B=探测器经历过用电高峰，那么
P(A)=0.20, P(B|A)=0.80, P(B|A^{c})=0.10。

贝叶斯公式（推广）

$$P(B_i|A)=\frac{P(A|B_i)P(B_i)}{P(A)}=\frac{P(A|B_i)P(B_i)}{\sum_{j=1}^{n}P(A|B_j)P(B_j)}$$

考虑一个实验样本，由相互独立的子集B_j所组成；

其中对于任意一个事件A，包含有事件B_j。

例1 诊断问题

一种诊断某癌症的试剂，经临床试验有如下记录：癌症病人试验结果是阳性的概率为95%，非癌症病人试验结果是阴性的概率为95%。现用这种试剂在某社区进行癌症普查，设该社区癌症发病率为0.5%，问某人反应为阳性时，该如何判断他是否患有癌症？

例1 解

设A表示“反应为阳性”的事件，B表示“被诊断者患癌症”的事件，则

$$B_1 = B, \quad B_2 = \bar{B}$$

构成一个完备事件群，由题意知

$$P(A|B_1) = 0.95, \quad P(A|B_2) = 1 - P(\bar{A}|B_2) = 1 - 0.95 = 0.05,$$

$$P(B_1) = 0.005, \qquad P(B_2) = 0.995.$$

例1 解

由贝叶斯公式易得：

$$P(B_1|A) = \frac{P(A|B_1)P(B_1)}{P(A|B_1)P(B_1)+P(A|B_2)P(B_2)}$$

$$= \frac{0.95\times 0.005}{0.95\times 0.005+0.05\times 0.995}$$

$$\approx 0.087 = 8.7\%.$$

类似可得：

$$P(B_2|A) \approx 0.913 = 91.3\%.$$

由上可知，某人患癌症的可能性很小，需要到医院做进一步检查。

例题2 在儿童智商（IQ）测验

在儿童智商（IQ）测验中，

假定智商测试结果 $X\sim N(\theta,100)$，

其中 θ 为被测试儿童智商的真值

（换而言之，如果对这个儿童做大量类似而又独立的这种测试，他的平均分数为 θ）；

例题2 在儿童智商（IQ）测验

又假设 θ 的先验分布为$N(100,225)$，易知其后验分布

$$\pi(\theta|x)\sim N\left(\mu(x),\eta^2\right),$$

其中（n=1）；

$$\mu(x)=\frac{\sigma^2}{\sigma^2+\tau^2}\mu+\frac{r^2}{\sigma^2+r^2}x$$

$$=\frac{4}{13}\times 100+\frac{9}{13}x \qquad =\frac{400+9x}{13},$$

$$\eta^2=\frac{100\times 225}{100+225}=8,32^2, \qquad \eta=8,32.$$

例题2 在儿童智商（IQ）测验

若这个儿童测试的得分为x=115，则由上述公式算的 μ (x)=110.39，故此儿童智商的真值 θ 的后验分布

$$\pi\left(\theta\middle|x\right)\sim N\left(110.39,8,39^{2}\right),$$

从而 θ 的95%可信区间为

$$(\mu(x)-1.96\eta,\mu(x)+1.96\eta)=(94.07,\quad 126.69).$$

例题2 讨论

如果不用先验信息，仅用抽样信息，则按照经典方法，由$X\sim N(\theta,100)$和该儿童测试得分$x=115$，求得θ的置信水平为0.95的置信区间为，

$$(115-10\times1.96\ ,\ 115+10\times1.96)=(95.4\ ,\ 134.6).$$

两个方法计算区间不同，而且意义不同，经典方法不能说“θ落入区间（95.4, 134.6）中的概率为0.95”，也不能说“此区间盖住θ的概率为0.95”，在束缚下，这个区间还能有什么用？这就是经典置信区间常受到批评的原因。

例题3　食物中毒

一次郊游活动中发生食物中毒的研究数据，参加郊游的320人中有304个回答了问卷。在食用的食物中，土豆沙拉和蟹肉被怀疑有问题（表3）。我们仅考虑怀疑最有问题的土豆沙拉。我们希望检验假设土豆沙拉和得病没有关系。

		食物配置		
	吃蟹肉		没有吃蟹肉	
土豆沙拉	吃了	没有吃	吃了	没有吃
得病	120	4	22	0
没得病	80	31	24	23

例题3　解：

记p_1=P（得病|吃了土豆沙拉）

和p_2=P（得病|没有吃了土豆沙拉）。

X_1表示在n_1个吃土豆沙拉中得病的人数，

X_2 表示 n_2 个没有吃土豆沙拉的人中得病的人数，则可认为X_1和X_2服从二项分布：

$$X_i \sim B(n_i, \quad p_i) \quad (i = 1, 2).$$

例题3　解：

检验土豆沙拉和得病没有没有关系等价检验

假设 H_0：　$p_1=p_2$.

取p_1和p_2的先验分布为 $B_e(\alpha_i,\beta_i)$　　$(i=1,2)$,

则 $\theta=p_1-p_2$的后验密度为

$$\mu\left(\theta\middle|X_1,X_2\right)$$

$$\propto\int_0^1\left(\theta+p_2\right)^{X_1+\alpha_1-1}\left(1-\theta-p_2\right)^{n_1-X_1+\beta_1-1}p_2^{X_2+\alpha_2-1}\left(1-p_2\right)^{n_2-X_2+\beta_2-1}dp_2.$$

例题3　解：

在本例中，样本量比较大，用逼近法计算后验分布。易知θ的后验分布渐进趋于正态分布$N(a, b^2)$，其中

$$a = p_1 - p_2, \quad b^2 = \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2},$$

$$p_1 = \frac{X_1}{n_1}, \qquad p_2 = \frac{X_2}{n_2}.$$

因此θ的100（1−α）%的可信区间为：

$$a - bz_{\alpha/2} < \theta < a + bz_{\alpha/2}$$

例题3　解：

其中，$z_{\alpha/2}$位标准正态的上$\alpha/2$分位数。对于食用了蟹肉的人来说，$X_1=120$，$n_1=200$，$X_2=4$，$n_2=35$，

因此99%可信区间为（0.337，0.635）。

而对没有食用蟹肉的人来说，$X_1=22$，$n_2=46$，$X_2=0$，$n_2=23$，

因此99%可信区间为（0.307，0.650）。

在这两种情况下，假设$\theta=0$都是落在99%区间外面，因此有很强的证据否定了零假设。

例题4 彩电寿命

彩色电视机寿命服从指数分布Exp(1/θ)，其密度函数为

$$f\left(x|\theta\right)=\theta^{-1}e^{-x/\theta}\cdot I_{(0,\infty)}(x)$$

其中：θ〉0是彩电的平均寿命。

例题4 彩电寿命

现从一批彩电中随机抽取n台进行寿命试验，试验到第r（1≤r≤n）台失效时为止，其失效时间为t1≤t2≤ …≤tn，其它n-r台彩电直至试验停止（tr）时还没有失效，这样的试验称为<u>定数截尾寿命试验</u>，所得样本（t1，… ，tr ）称为<u>截尾样本</u>。

例题4 解

假定 θ 的先验分布为逆伽玛分布 $r^{-1}(\alpha,\beta)$，求这批彩电的平均寿命及其可信下限。

解 设被抽取的n台彩电的寿命为X1，…，Xn，令$X_{(1)} \leqslant \cdots \leqslant X_{(n)}$为其次序统计量，记 $t1=X_{(1)}$，…，$tr=X_{(r)}$，可知

$$T=\sum_{j=1}^{r} t_j+(n-r)t_r$$

是 θ 的充分统计量，

例题4 解

且给定 θ 时，$2T/\theta \sim \chi^2_{2r}$, 故T的密度函数为

$$g(t|\theta)=\frac{\theta^{-r}}{\Gamma(r)}t^{r-1}e^{-t/\theta},I_{(0,\infty)}(t),\quad(\theta\geq 0).$$

θ 的先验密度和后验密度如下：

$$\pi(\theta)=\frac{\beta^{\alpha}}{\Gamma(\alpha)}\theta^{-(\alpha+1)}e^{-\beta/\theta}\cdot I_{(0,\infty)}(\theta),$$

$$\pi(\theta|t)=\frac{(t+\beta)^{\alpha}}{\Gamma(r+\alpha)}\theta^{-(r+\alpha+1)}e^{-(r+\beta)/\theta}\cdot I_{(0,\infty)}(\theta).$$

例题4 解

显然，θ 的后验分布为逆伽玛分布

$$\Gamma^{-1}(r+\alpha, t+\beta).$$

若取后验期望估计作为 θ 的贝叶斯估计，则有

$$\hat{\theta} = E(\theta|t) = \frac{\beta + t}{r + \alpha - 1}$$

例4 具体算例

设有13142台彩电寿命试验的数据，共计5369812台时，此外还有9240台彩电进行了三年现场跟踪试验。总共进行了5547810台时试验。这些试验中共有250台失效，由先验数据整理，彩电平均寿命不低于30000小时，它的10%的分位数 $\theta_{0.1}$，大约11250小时。

例4 具体算例

由此列出如下的两个方程：

$$\begin{cases} \dfrac{\beta}{\alpha-1}=30000. \\ \int_0^{11250}\pi(\theta)d\theta=0.1. \end{cases}$$

其中第一个方程式由先验分布为逆伽马分布的数学期望$E(\theta)=\beta/(\alpha-1)$确定。

例4 具体算例

在计算机上解方程组，得

$$\alpha = 1.956, \quad \beta = 2.868。$$

得到先验分布

$$\theta \sim r^{-1}(1.956, 2.868),$$

后验分布

$$\theta \mid t \sim r^{-1}(r+1.956, t+2.868)。$$

例4 具体算例

现随机抽取100台彩电，在规定条件下进行400小时寿命试验，没有一台失效，这时总的试验时间为

t=100*400 = 40000（小时），r=0。

可知彩电的平均寿命的贝叶斯估计为

$$\hat{\theta} = \frac{\beta + t}{r + \alpha - 1} = \frac{2868 + 40000}{1.956 - 1} = 44841(h)$$

例4 具体算例

θ 的$1-\eta=0.9$，可信下限为

$$\hat{\theta}_L=\frac{2(t+\beta)}{\chi_f^2(\eta)},$$

$$\alpha=1.956,\beta=2868,$$

$$t=40000,r=0,f=2(r+\alpha)=3.912.$$

例4 具体算例

$$\hat{\theta}_L = \frac{2(t+\beta)}{\chi_f^2(\eta)} = \frac{2(40000+2868)}{7.645} = 11215(\text{小时}).$$

答：彩电平均寿命接近45000小时，而平均寿命的90%可信下限约为11000小时。

例5 汽车问题

设某公路上经过的货车与客车的数量之比为2:1，货车中途停车修理的概率为0.02，客车为0.01，今有一辆汽车中途停车修理，求该汽车是货车的概率。

例5 求解

解：设 B={中途停车修理}，A1={经过的是货车}，A2={经过的是客车}，则 $B=A_1B\cup A_2B$，P(A1)=0.02；P(A2)=0.01.

$$P(A_1|B)=\frac{P(A_1)P(B|A_1)}{P(A_1)P(B|A_1)+P(A_2)P(B|A_2)}=\frac{\frac{2}{3}\times 0.02}{\frac{2}{3}\times 0.02+\frac{1}{3}\times 0.01}=0.80.$$

例6　取球的问题

* 已知甲袋中有6只红球，4只白球；乙袋中有8只红球，6只白球。求下列事件的概率：

(1)随机取一只袋，再从该袋中随机取一球，该球是红球；

(2)合并两只袋，从中随机取一球，该球是红球。

例6 解

(1)记 B={该球是红球}，$A1$={取自甲袋}，$A2$={取自乙袋}，

已知，$P(B \mid A_1) = 6/10$，$P(B \mid A_2) = 8/14$，

所以

$$P(B) = P(A_1)P(B \mid A_1) + P(A_2)P(B \mid A_2) = \frac{1}{2} \times \frac{6}{10} + \frac{1}{2} \times \frac{8}{14} = \frac{41}{70}$$

(2) $$P(B) = \frac{14}{24} = \frac{7}{12}$$

例7 全概率公式

设某工厂有两个车间生产同型号家用电器，第一车间的次品率为0.15，第二车间的次品率为0.12，两个车间的成品都混合堆放在一个仓库，假设第1,2车间生产的成品比例为2:3，今有一客户从成品仓库中随机提一台产品，求该产品合格的概率。

例7　求解：

设　B={从仓库中随机提出的一台是合格品}

Ai={提出的一台是第i车间生产的}，i=1,2

则有分解　$B=A_1B \cup A_2B$

由题意

$$P(A1)=2/5,P(A2)=3/5,P(B|A1)=0.85,P(B|A2)=0.88$$

由全概率公式

$$P(B)=P(A_1)P(B|A_1)+P(A_2)P(B|A_2)=0.4*0.85+0.6*0.88=0.868.$$

END

Thank you for your attention