大规模假设检验*

大规模推断讨论班

统计学的发展进步要归功于我们的科学同仁们，他们从我们的工作中获得的数据更加“原始”。二十世纪早期费舍尔对农业试验的研究推动了方差分析的发展。在 21 世纪伊始，类似的事情正在上演。一类新的“高产量”的生物医学装置，典型例子是微阵列，这种装置可以例行般地马上产生几千次实例的假设检验数据。这一点不是典型频率检验理论（以纽曼、皮尔森、费舍尔为代表）所预想的情形。这一章开始讨论大规模联立假设理论，这一理论正在统计文献中不断发展。

2.1 一个微阵列的例子

图 2.1 是一个前列腺数据的微阵列。这些数据是从 202 个测试者（其中 50 个健康正常人，52 个前列腺癌病患者）中各提取 6033 个基因的基因表达水平获得的。我们姑且不深究这些基因表达的生物学细节，而是把发现少量的“令我们感兴趣”的基因作为我们研究的主要目标。这些我们要寻找的基因的表达水平介于正常的和癌变的前列腺细胞基因表达水平之间。因而一旦确定识别了这些基因，我们便会进行进一步的调查研究以确定这些基因是否和前列腺癌病的发展具有因果关系。基于上面对前列腺数据的描述，我们可以了解到，前列腺数据是一个 6033 行 102 列的矩阵（记作 X）, 它的元素为

$$x_{ij} = \text{j 病人第 i 个基因的基因表达，} \tag{2.1}$$

其中 i=1,2,...,N;j=1,2,...,n。并且，j=1,2,...,50 表示健康人;j=51,52,...,102 表示癌症患者。令 $\overline{x}_i(1)$ 和 $\overline{x}_i(2)$ 分别表示健康人群和癌症患者人群中 x_{ij} 的平均值。那么两样本基因检测的 t 统计量为：

$$t_i = \frac{\overline{x}_i(2) - \overline{x}_i(1)}{S_i} \tag{2.2}$$

其中 s_i 是分子标准差的一个估计量

$${s_i}^2 = \frac{\sum_1^{50}(x_{ij} - \overline{x}(1))^2 + \sum_{51}^{102}(x_{ij} - \overline{x}(2))^2}{100}(\frac{1}{50} + \frac{1}{52}) \tag{2.3}$$

如何只考虑基因 i 的数据，通常我们可以用 t_i 来检验原假设

$$H_{0i} : gene\ i\ is\ \text{“}null\text{”}, \tag{2.4}$$

$i.e._{\cdot\cdot}$ 对没患癌症的和患癌症的人来说，x_{ij} 有相同的分布. 如果 $|t_i|$ 比较大则拒绝 H_{0i}. 基于标准的理论假设下，拒绝的标准通常取 5%，如果 $|t_i| > 1.98$ 则拒绝原假设, 对于自由度为 100 的学生 t 分布双侧拒绝标准为 5%.

这里我们用“z 值”代替“t 值”，讨论会比较简便；就是说，把 t_i 转化为如下形式：

$$z_i = \Phi^{-1}(F_{100}(t_i)) \tag{2.5}$$

*本文作者为大规模推断讨论班，成员：杨晓康、张洋、宋培培、张猛、刘博、朱祁恒和高磊.

这里 Φ 和 F_{100} 分别是标准正态分布和 t_{100} 分布的累积分布函数（简写为 "*cdf*"). 通常在正常抽样的假设下，如果接受 H_{0i}, 则 z_i 服从标准正态分布，

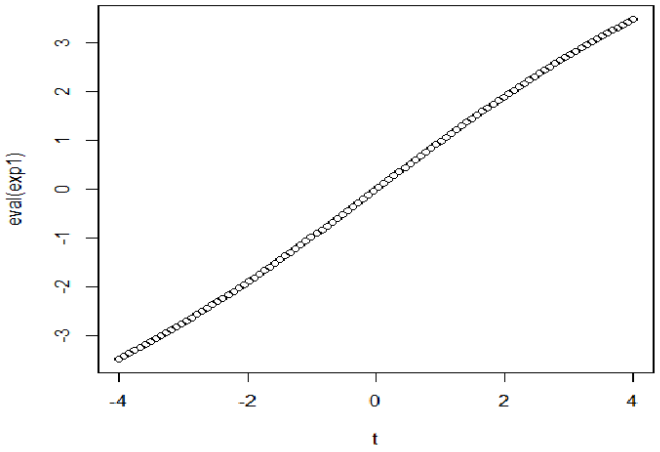
$$H_{0i} \ : \ z_i \ \sim N(0.1) \tag{2.6}$$

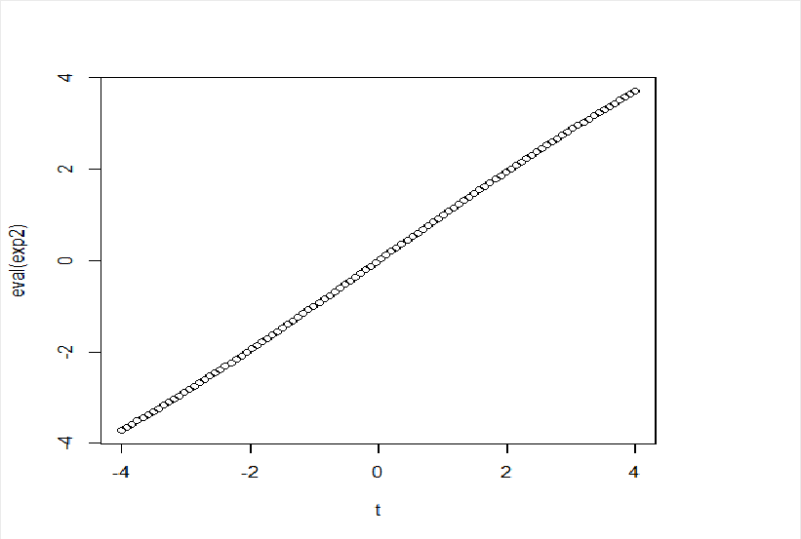
(随后的文章中称为 theoretical null). 对于 $N(0.1)$ 分布，在显著性水平为 5% 的双侧检验下，如果 $|z_i| > 1.96$ 则拒绝原假设。

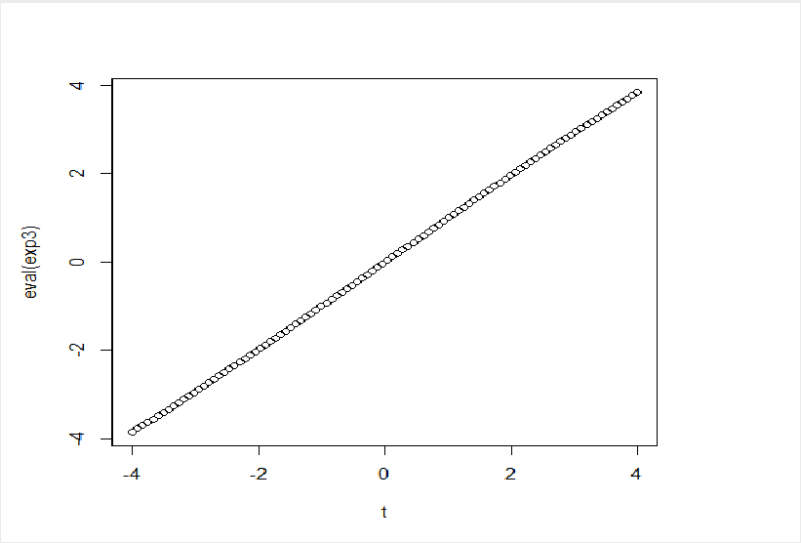
练习 2.1对于自由度 υ 等于 25 , 50 和 100 的 t 分布，在 $-4 \leq t \leq 4$ 的条件下，分别画出 $z_i = \Phi^{-1}(F_\upsilon(t))$

练习 2.1 解答：

```
t=seq(-4,4,len=100)
f1=pt(t,25)
f2=pt(t,50)
f3=pt(t,100)
exp=expression(qnorm(f1))
exp2=expression(qnorm(f2))
exp3=expression(qnorm(f3))
plot(t,eval(exp1),lty=1)
plot(t,eval(exp2),lty=1)
plot(t,eval(exp3),lty=1)
```







但是我们当然不是仅检验一个基因，我们有 6033 个基因. 图 2.1 给出了 z_i 值的直方图，与标准 $N(0.1)$ 密度曲线 $c \cdot exp\{-z^2/2\}/\sqrt{2\pi}$ 进行比较, 这条曲线的乘数 c 是为了使曲线拟合直方图的区域. 如果每一个基因 i 都接受原假设，即所有的基因都是无效的，直方图会很好的拟合这条曲线. 对研究员来说幸运的是，它并没有很好拟合，中心位置太低，两边太高. 表明有一些基因是有效的. 在多重推理的影响没有误导我们的情况下，如何独立地识别这些有效的基因是目前主要的研究课题。

多重推理一个传统的解决办法是采用 Bonferroni 约束：每次检验的显著性水平由 0.05 缩小为 0.05/6033. 这就等价于当 $|z_i| > 4.31$ 时拒绝原假设，而不是大于 1.96. 现在 6033 个原假设，其中之一被错误拒绝的总概率甚至都小于 5%. 但看图 2.1，4.31 有点过于谨慎。(只有 6 个基因是有效的）经验贝叶斯会为多重检验提供一个不那么保守的方法。

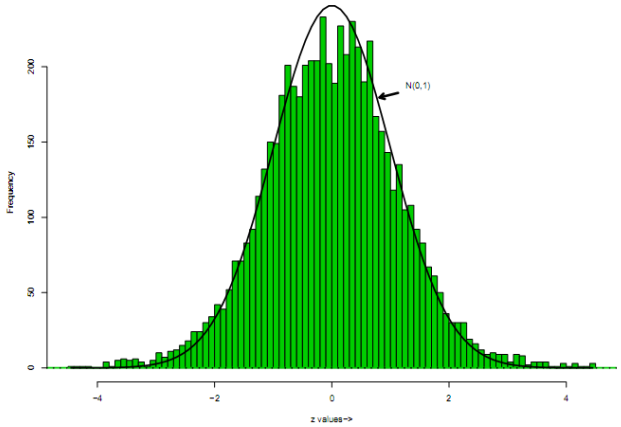


图 2.1: 前列腺数据：对可能与前列腺癌有关的 6033 个基因进行检验得到 z 值：曲线是 $N(0.1)$theoretical null。

2.2 贝叶斯方法

两分组模型提供了一个多元检验的贝叶斯框架。我们假定有 N 个记录 (前列腺基因研究)，每一个是无效的或有效的，先验概率分别为 π_0 和 $\pi_1 = 1 - \pi_0$, 同时 z 值的密度分别为 $f_0(z)$,$f_1(z)$。

$$\begin{aligned} \pi_0 &= Pr(\text{无效}) \quad f_0(z)\text{为无效时的密度} \\ \pi_1 &= Pr(\text{有效}) \quad f_1(z)\text{为有效时的密度} \end{aligned} \tag{2.7}$$

一般情况下，π_0 会比 π_1 大很多，假定

$$\pi_0 \geq 0.90 \tag{2.8}$$

这反映了大规模估计通常的目的：把拥有大量可能性的大集合减小到拥有丰富、有趣、科学前景的小集合。如果基于（2.6）的假定是合理的，那么 $f_0(z)$ 有标准正态密度，

$$f_0(z) = \varphi(z) = e^{-\frac{1}{2}z^2}/\sqrt{2\pi} \tag{2.9}$$

$f_1(z)$ 是未知的。F_0 和 F_1 分别表示 f_0 和 f_1 的概率分布，

$$F_0(Z) = \int_Z f_0(z)dz \quad \text{和} \quad F_1(Z) = \int_Z f_1(z)dz \tag{2.10}$$

由此得出混合密度函数

$$f(z) = \pi_0 f_0(z) + \pi_1 f_1(z) \tag{2.11}$$

对应于 (2.11) 式的混合概率分布函数

$$\boldsymbol{F}(\boldsymbol{Z}) = \pi_0 \boldsymbol{F}_0(\boldsymbol{Z}) + \pi_1 \boldsymbol{F}_1(\boldsymbol{Z}). \tag{2.12}$$

(在注释中一般的累积分布函数用 $\boldsymbol{F}((-\infty, z))$，但以后我们用 $\boldsymbol{z}$.) 在模型 (2.7) 下，可以得到 z 的边缘密度函数 f 以及分布函数 F。

假定我们观察到 $z \in \boldsymbol{Z}$ 并且想知道它符合 (2.7) 中的无效性还是有效性。应用贝叶斯规则得到

$$\phi(\boldsymbol{Z}) \equiv Pr\{null \mid z \in \boldsymbol{Z}\} = \pi_0 \boldsymbol{F}_0(\boldsymbol{Z})/\boldsymbol{F}(\boldsymbol{Z}) \tag{2.13}$$

这就是在给定 $z \in \boldsymbol{Z}$ 时，无效性的后验概率。由 Benjamini 与 Hochberg 提出的专业术语，我们把 $\phi(\boldsymbol{Z})$ 称为 $\boldsymbol{Z}$ 的贝叶斯错误发现率: 如果我们认为 $z \in \boldsymbol{Z}$ 是有效的，$\phi(\boldsymbol{Z})$ 就是我们做了错误的发现的概率。我们还经常把 $\phi(\boldsymbol{Z})$ 写成 $Fdr(\boldsymbol{Z})$。

如果 $\boldsymbol{Z}$ 是一个单独的点即 z_0,

$$\phi(z_0) \equiv Pr\{null \mid z = z_0\} = \pi_0 f_0(z_0)/f(z_0) \tag{2.14}$$

这就是局部贝叶斯错误发现率，还经常写成 fdr(z)。

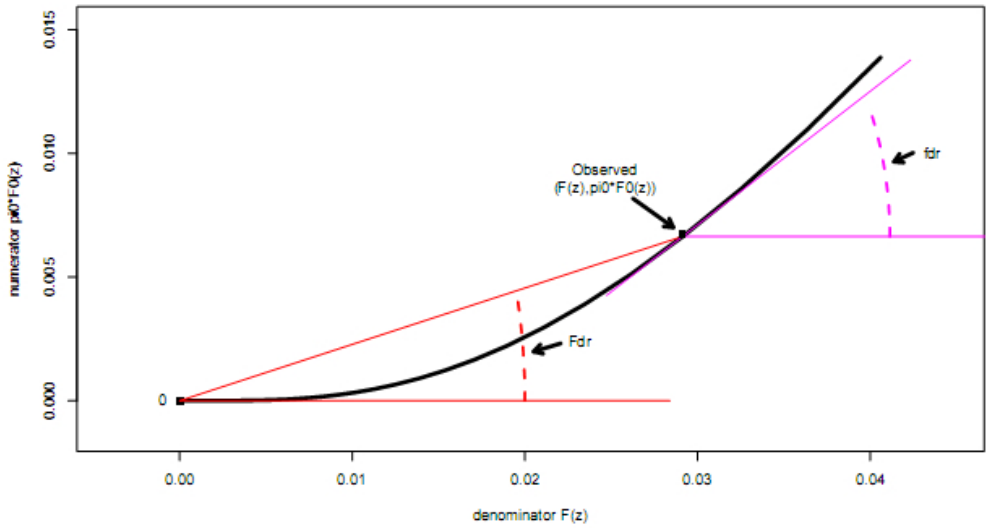


图 2.2描述了 $\boldsymbol{F}dr(z)$ 与 fdr(z) 之间的关系。$\boldsymbol{F}dr(z)$ 是连接原点与点 $(\boldsymbol{F}(z), p_0 \cdot \boldsymbol{F}_0(z_0))$ 之间割线的斜度，分别对应于 (2.16) 式中的分母与分子；fdr(z) 是在点 $(\boldsymbol{F}(z), p_0 \cdot \boldsymbol{F}_0(z_0))$ 处切线的斜度。

练习 2.2 证明

$$\boldsymbol{E}_f\{\phi(z) \mid z \in \boldsymbol{Z}\} = \phi(\boldsymbol{Z}) \tag{2.15}$$

其中 $\boldsymbol{E}_f$ 为关于边缘密度函数 f(z) 的条件期望。换句话说，$\boldsymbol{F}$ 就是在给定 $z \in \boldsymbol{Z}$ 下 fdr(z) 的条件期望。

证明:

$$\begin{aligned}\phi(\boldsymbol{Z}) =& \frac{\pi_0 \boldsymbol{F}_0}{\boldsymbol{F}(Z)} \\ =& \int_Z \frac{\pi_0 f_0(z)}{f(z)} \cdot \frac{f(z)}{\boldsymbol{F}(Z)} dz \\ =& \boldsymbol{E}_f\{\phi(z) \mid z \in \boldsymbol{Z}\}\end{aligned}$$

在给定 $z \in \boldsymbol{Z}$ 下,z 的概率密度函数为 $\frac{f(z)}{\boldsymbol{F}(Z)}$ 呢? 在应用中，z 通常是一个尾部区间，对于标准的正太累积分布函数把 $F(Z)$ 写作 $F((-\infty, z))$

$$\phi((-\infty, z)) \equiv Fdr(z) = \pi_0 F_0(z)/F(z) \tag{2.16}$$

用分子 $\pi_0 F_0(z)$ 和分母 F(z) 画图，表明 $Fdr(z)$ 和 $fdr(z)$ 分别是正割值和正切值。正如图 2.2 所示，当两者均较小时，通常表明：$fdr(z) > Fdr(z)$。

练习 2.3 假定

$$F_1(z) = F_0(z)^\gamma \quad [\gamma < 1] \tag{2.17}$$

(通常称为莱曼选择) 因此有：

$$\log\{\frac{fdr}{1-fdr}\} = \log\{\frac{Fdr}{1-Fdr}\} + \log(\frac{1}{\gamma}) \tag{2.18}$$

而且有：

$$fdr(z) \doteq Fdr(z) \tag{2.19}$$

练习 2.3 解答:

$$\frac{fdr}{1-fdr} = \frac{\pi_0 f_0/f}{1-\pi_0 f_0/f} = \frac{\pi_0 f_0}{f-\pi_0 f_0}$$

根据 2.11 式可知：$f - \pi_0 f_0 = \pi_1 f_1$, 因此上式可变为：

$$\frac{fdr}{1-fdr} = \frac{\pi_0 f_0}{\pi_1 f_1}$$

又由 2.17 式可知

$$F_1(z) = F_0(z)^\gamma$$

对上式求导可得：$f_1 = \gamma F_0^{\gamma-1} f_0$ 由此可得：$\frac{f_0}{f_1} = \frac{1}{\gamma F_0^{\gamma-1}}$ 带入 $\frac{\pi_0 f_0}{\pi_1 f_1}$ 可得：

$$\frac{\pi_0}{\pi_1 \gamma F_0^{\gamma-1}} = \frac{\pi_0 F_0}{\pi_1 \gamma F_0^{\gamma}} = \frac{\pi_0 F_0}{\pi_1 \gamma F_1} = \frac{1}{\gamma} * \frac{Fdr}{1-Fdr}$$

因此 $\log\{\frac{fdr}{1-fdr}\} = \log\{\frac{Fdr}{1-Fdr}\} + \log(\frac{1}{\gamma})$

又因为 $\frac{fdr}{1-fdr} = \frac{1}{\gamma} * \frac{Fdr}{1-Fdr}$ 把 $\frac{Fdr}{1-Fdr}$ 当做一个整体解方程可得：

$$fdr = \frac{Fdr}{\gamma + (1-\gamma)Fdr}$$

所以当 Fdr 较小时 $(1-\gamma)Fdr$ 可忽略不计，所以有

$$fdr(z) \doteq Fdr(z)$$

练习 2.4 通常从定性角度来说 2.7 式中有效密度函数 $f_1(z)$ 和 $f_0(z)$ 相比较上是一个重尾分布，为什么这表明了图 2.2 中曲线的形状?

练习 2.4 解答: 因为和 $f_0(z)$ 相比较 $f_1(z)$ 是一个重尾分布，所以累积分布函数 F_1 尾部较厚，那么 F_1 增长的比 F_0 快。在图 2.2 中，分子为 $\pi_0 F_0(z)$，分母为 $F(z) = \pi_0 F_0(z) + \pi_1 F_1(z)$, 曲线斜率为 $\frac{\pi_0 F_0(z)}{\pi_0 F_0(z) + \pi_1 F_1(z)}$。由于 F_1 增长的比 F_0 快，所以曲线斜率越来越小（??? 解释好像不对）。

2.3 经验贝叶斯估计

2.7 式中贝叶斯的两组模型涉及了 3 个变量，无效的先验概率为 π, 密度函数为 $f_0(z)$，有效密度函数 $f_1(z)$。当然，$f_0(z)$ 是已知的，如果 2.1 式中原假设成立即 $z_i \sim N(0,1)$, 那么 π_0 就是已知的。通常当 π_0 对错误观察率影响较小时，π_0 接近于 1。（在应用中，把 π_0 当做 1；第六章将讨论原假设不成立情况下 π_0 和 $f_0(z)$ 的估计值）。现在只有 $f_1(z)$ 未知，对统计学家来说，不可能知道 $f_0(z)$ 的先验信息，

然而，可以使用经验贝叶斯方法对错误观察率进行估计，另 $\overline{F}(z)$ 表示 N 个 z 值的经验分布，即

$$\overline{F}(z) = \sharp\{z_i \in Z\}/N \tag{2.20}$$

用估计的错误观察率替换 2.13 式有：当 N 较大时，我们希望 $\bar{F}(z)$ 接近 $F(z)$，$\overline{F}(z)$ 是 $F(z)$ 的一个较好的近似值。

$$\overline{Fdr}(z) = \overline{\phi}(Z) = \frac{\pi_0 F_0(Z)}{\overline{F}(Z)} \tag{2.21}$$

当 N 足够大时我们假设 $\overline{F}(Z)$ 趋向于 F(Z)，并且 $\overline{Fdr}(Z)$ 是 Fdr(Z) 的一个优质估计量。

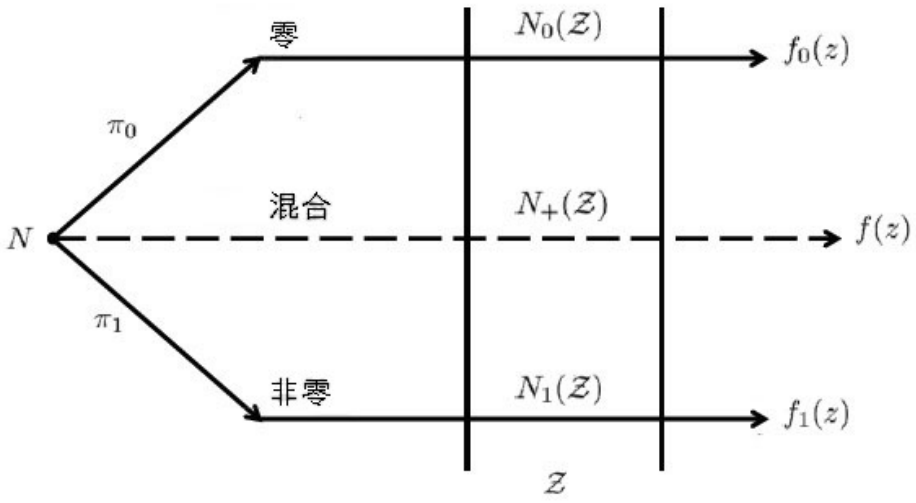


图 2.3: 两组模型（2.7）的简图；N 个 z 值分布于两侧的比例分别为 π_0 和 π_1，$N_0(Z)$ 和 $N_1(Z)$ 为 Z 中零和非零的个数；统计学家观测到来自混合密度函数 f(z)（2.11）的 $z_1, z_2, ..., z_N$，并可以看到 $N_+(Z) = N_0(Z) + N_1(Z)$。

接下来要谈论的问题是非常有意义的。图 2.3 表述了模型（2.7）：$z_1, z_2, ..., z_N$ 共 N 个值分布于零和非零两个区域，比率分别为 π_0 和 π_1。令 $N_0(Z)$ 为零值 z_i 落在 Z 的个数，类似地 $N_1(Z)$ 为非零 z_i 在 Z 的个数。我们无法观测到 $N_0(Z)$ 或 $N_1(Z)$，但我们可以观测到二者在 Z 中的总和 $N_+(Z)$，

$$N_+(Z) = N_0(Z) + N_1(Z) \tag{2.22}$$

虽然 $N_0(Z)$ 无法观测到，我们可以得到其期望值

$$e_0(Z) \equiv N\pi_0 F_0(Z) \tag{2.23}$$

在定义中我们可以令 $\overline{Fdr}(Z)$ 为

$$\overline{Fdr}(Z) = e_0(Z)/N_+(Z) \tag{2.24}$$

例子中，前列腺数据中 $N_+(Z) = 49$，z_i 值位于 Z 中，$Z = (3, \infty)$，即超过 3；$e_0(Z)$ 为

$$6033 \cdot \pi_0 \cdot (1 - \Phi(3)) \tag{2.25}$$

低于理论上的零值（2.6）。取上限为 $\pi_0 = 1$，得到 $e_0(Z) = 8.14$ 以及

$$\overline{Fdr}(Z) = 8.14/49 = 0.166 \tag{2.26}$$

注：(2.25) 中限定 π_0 为 1 并不意味着在 (2.12) 中 $\pi_1 = 0$，因为在 (2.21) 中分母 $\overline{F}(Z)$ 是考虑到 π_0 后而从数据中直接估计得到的。

实践结果为 49 例中大约有 1/6 是错误发现的：如果我们将这 49 例的报告给前列腺研究学者作为与其结论，他们大部分之后的工作就不会白费。第 4 章将验证这一想法背后的逻辑。这里我们仅认为 $\overline{Fdr}(Z)$ 为 Fdr(Z) 的一个经验贝叶斯点估计。

2.4 点估计量 $\overline{Fdr}(Z)$

2.13 式可以表示为：

$$\phi(Z) = Fdr(Z) = e_0(Z)/e_+(Z) \tag{2.27}$$

上式中 $e_+(Z) = N * F(Z)$，$e_+(Z)$ 是集合 Z 中 Z_i 的总个数的期望值，即 $E(N_+)$.

$$Fdp(Z) = N_0(Z)/N_+(Z) \tag{2.28}$$

上式中 $Fdp(Z)$ 定义为错误观察比例，它是集合 Z 中真实的错误观察比例。我们希望知道错误观察比例的个数，但是错误观察比例的个数观测不到。这里给出三个变量值来考虑,

$$\overline{Fdr}(Z) = \frac{e_0(Z)}{N_+(Z)}, \qquad \phi(Z) = \frac{e_0(Z)}{e_+(Z)}, \qquad Fdp(Z) = \frac{N_0(Z)}{N_+(Z)} \tag{2.29}$$

用下面的四个引理来讨论他们之间的关系。

引理 2.1 假定 2.23 式中定义的 $e_0(Z)$ 是给定 $N_1(Z)$ 时 $N_0(Z)$ 的条件期望相同，即 $e_0(Z) = E(N_0(Z)|N_1(Z))$。那么 $\overline{Fdr}(Z)$ 的条件期望和给定 $N_1(Z)$ 时 $Fdp(Z)$ 的条件期望满足下列不等式：

$$E\{\overline{Fdr}(Z)|N_1(Z)\} \geq \phi_1(Z) \geq E\{Fdp(Z)|N_1(Z)\} \tag{2.30}$$

其中

$$\phi_1(Z) = \frac{e_0(Z)}{e_0(Z) + N_1(Z)} \tag{2.31}$$

上式中 $\overline{Fdr}(Z) = e_0(Z)/(N_0(Z) + N_1(Z))$, 由 Jensen's 不等式可知 $E\{\overline{Fdr}(Z)|N_1(Z)\} \geq \phi_1(Z)$。当无效的 Z_i 的个数以及它的分布不依赖于 $N_1(Z)$ 时，上述假定的 $e_0(Z)$ 的条件就能满足。

注：Jensen's 不等式：设 $f(x)$ 为凹函数, 则

$$f((x_1+x_2+\ldots\ldots x_n)/n)\geq(f(x_1)+f(x_2)+\ldots\ldots+f(x_n))/n$$

若 $f(x)$ 为凸函数, 则

$$f((x_1+x_2+\ldots\ldots x_n)/n)\leq(f(x_1)+f(x_2)+\ldots\ldots+f(x_n))/n$$

称为 Jensen's 不等式。

Jensen's 不等式的加权形式为：

$f(a_1x_1+a_2x_2+\ldots\ldots+a_nx_n)\geq a_1f(x_1)+a_2f(x_2)+\ldots\ldots+a_nf(x_n)$(凹函数)

$f(a_1x_1+a_2x_2+\ldots\ldots+a_nx_n)\leq a_1f(x_1)+a_2f(x_2)+\ldots\ldots+a_nf(x_n)$(凸函数)

其中 $a_i\geq 0(i=1,2,\ldots\ldots n)$ 且 $a_1+a_2+\ldots\ldots+a_n=1$

证明 (只对凸函数加以证明)：首先对 n 是 2 的幂加以证明，用数学归纳法：假设对于 $n=2^k$Jensen's 不等式成立，那么对于 $n=2^{k+1}$ 有 $(f(x_1)+f(x_2)+\ldots\ldots+f(x_n))/n=(f(x_1)+f(x_2)+\ldots\ldots+f(x_{n/2})/(n/2)+(f(x_{n/(2+1)})+f(x)_n)/(n/2)/2\geq f(x_1+x_2+\ldots\ldots+x_(n/2))+f(x_{(n/(2+1))}+\ldots\ldots x_n)/(n/2)/2)=f((x_1+x_2+\ldots\ldots x_n)/n)$ 成立，那么对于所有的 2 的幂 Jensen's 不等式成立。

如果对于一个普通的 n，如果 n 不是 2 的幂，我们可以找到一个 k，使得 $2^k>n$, 然后我们设 $x_{(n+1)}=x_{(n+2)}=\ldots\ldots=x_{(2^k)}=(x_1+x_2+\ldots\ldots+x_n)/n$，代入 2^k 的 Jensen's 不等式结论，整理即可证明

练习 2.5 运用 Jensen's 不等式完成 2.30 式的证明

注释 在 2.30 式的关系中作了一个常见的假定：如果 $N_+(Z)=0$，则 $Fdp(Z)=0$。

引理 2.1 表明对于任意的 $N_1(Z)$ 都有 $E\{\overline{Fdr}(Z)|N_1(Z)\}\geq E\{Fdp(Z)|N_1(Z)\}$, 在这个意义上，经验贝叶斯错误发现比率是真实错误发现比例的适当有偏估计。对 $N_1Z)$ 求期望并运用 Jensen's 不等式可得：

$$\phi(Z)\geq E\{Fdp(Z)\} \tag{2.32}$$

从上式可以看出 $Fdr(Z)$ 是 $E\{Fdp(Z)\}$ 的一个上界。我们也能得到：$E\{\overline{Fdr}(Z)\}\geq E\{Fdp(Z)\}$, 但这一结论是信息不足的，因为当 $P\{N_+(Z)=0\}\geq 0$ 为正值时，$E\{Fdr\}=\infty$. 实际中我们使用 $\overline{Fdr}^{(min)}=min(\overline{Fdr}(Z),1)$ 来估计 $Fdr(Z)$, 但是一般来说 $E\{\overline{Fdr}^{(min)}\}>\phi(Z)$ 或者 $E\{\overline{Fdr}^{(min)}\}>E\{Fdp\}$。**练习 2.6** 证明 $E\{min(\overline{Fdr}(Z),2)\}\geq\phi(Z)$. 提示：过点 $(e_+(Z),\phi(Z))$ 画曲线 $(N_+(Z),\overline{Fdr}(Z))$ 的切线.

练习 2.6 解答：

标准的 Delta 方法在不要求引理 2.1 关于 $e_0(Z)$ 的条件下，采用近似的方法得到 $\overline{Fdr}(Z)$ 的均值和方差。

注：Delta 方法：Delta 方法源自这样一种方法，即在估计量的方差受到约束的情况下，由渐近正态估计得到近似分布函数。一般地说，我们可以认为 Delta 方法是相当普遍的中心极限定理。

引理 2.2$\gamma(Z)$ 表示 $N_+(Z)$ 的变异系数的平方，

$$gamma(Z) = var\{N_+(Z)\}/e_+(Z)^2 \tag{2.33}$$

那么 $\overline{Fdr}(Z)/\phi(Z)$ 有近似的均值和方差

$$frac\overline{Fdr}(Z)\phi(Z)\dot{\sim}(1+\gamma(Z),\gamma(Z)) \tag{2.34}$$

证明. 标记中去掉 Z。

$$\overline{Fdr} = \frac{e_0}{N_+} = \frac{e_0}{e_+}\frac{1}{1+(N_+-e_+)/e_+}\doteq\phi\left[1-\frac{N_+-e_+}{e_+}+\left(\frac{N_+-e_+}{e_+}\right)^2\right] \tag{2.35}$$

$(N_+-e_+)/e_+$ 的均值和方差为 $(0,\gamma(Z))$。

注： $\frac{e_0}{N_+} = \frac{e_0}{e_+}\cdot\frac{e_+}{N_+} = \frac{e_0}{e_+}\cdot\frac{e_+}{e_++N_+-e_+} = \frac{e_0}{e_+}\cdot\frac{1}{1+\frac{N_+-e_+}{e_+}}$
由麦克劳林级数可知 $\frac{1}{1+t} = 1-t+t^2+\cdots$
所以 $\frac{1}{1+\frac{N_+-e_+}{e_+}} = 1-\frac{N_+-e_+}{e_+}+\left(\frac{N_+-e_+}{e_+}\right)^2+\cdots$
且 $\phi = \frac{e_0}{e_+}$, 所以可得 2.35 式。
$E\left(\frac{N_+-e_+}{e_+}\right) = 0, var\left(\frac{N_+-e_+}{e_+}\right) = \frac{1}{e_+^2}var(N_+-e_+) = \frac{1}{e_+^2}var(N_+) = \gamma(Z)$
由 2.35 可知 $\frac{\overline{Fdr}}{\phi} = \left[1-\frac{N_+-e_+}{e_+}+\left(\frac{N_+-e_+}{e_+}\right)^2\right]$
所以可得 2.34 式。引理 2.2 得证。

由引理 2.2 显然可得：$\overline{Fdr}(Z)$ 是贝叶斯错误发生率 $\phi(Z)$ 的估计,$\overline{Fdr}(Z)$ 的精确度依赖于分母 $N_+(Z)$ 的变异性 (2.24)。如果我们为图 2.3 的两群体模型补充一个独立条件的话，便会得到更多的特定结论。

独立条件：模型 2.7 的每个 z_i 都是独立的 (2.36)

那么 $N_+(Z)$ 服从二项分布，

$$N_+(Z)\sim Bi(N,F(Z)) \tag{2.37}$$

$N_+(Z)$ 的变异系数的平方

$$\gamma(Z) = \frac{1-F(Z)}{NF(Z)} = \frac{1-F(Z)}{e_+(Z)} \tag{2.38}$$

注：若 $x\sim B(n,p)$, 则 $E(x) = np$,$D(x) = npq$. 所以有 $E[N_+(Z)] = NF(Z)$,$D[N_+(Z)] = NF(Z)[1-F(Z)]$
$\gamma(Z) = \frac{D[N_+(Z)]}{E[N_+(Z)]^2} = \frac{1-F(Z)}{NF(Z)} = \frac{1-F(Z)}{e_+(Z)}$

通常我们会对 $F(Z)$ 小的区域 Z 感兴趣，所以给定 $\gamma(Z)\doteq\frac{1}{e_+(Z)}$, 由引理 2.2 有，

$$\overline{Fdr}(Z)/\phi(Z)\dot{\sim}(1+1/e_+(Z),1/e_+(Z)) \tag{2.39}$$

在之前提到的前列腺数据中 $Z=(3,\infty)$，我们可以用 $N_+=49$ 来估计 $e_+(Z)$,$\overline{Fdr}(Z)/\phi(Z)$ 均值近似为 1.02, 标准差为 0.14。$\overline{Fdr}(Z) = 0.166$(2.26), 是贝叶斯错误发现率 $\phi(Z)$ 的近似无偏估计，

变异系数近似等于 0.14.$\phi(Z)$ 的 95% 的置信区间为 $0.166 * (1 \pm 2 * 0.14) = (0.12, 0.21)$. 所有的这些都是基于（2.36）的独立性假设，在第八章中可以看到，这仅仅是一个适度的、有风险的假定。

我们同样可以添加一个无害的假定 N 服从一个泊松分布，独立性假设依然存在，这样我们在前面的讨论中得到的整齐的结论依然是可能的。

$$N \sim Poi(\eta). \tag{2.40}$$

引理 2.3 在（2.36）,（2.40）泊松 - 独立性假设条件下有，

$$E\{Fdp(Z)\} = \phi(Z) \cdot [1 - exp(e_+(Z))]. \tag{2.41}$$

其中 $e_+(Z) = E\{N_+(Z)\} = \eta \cdot F(Z)$.

引理 2.3 证明：图 2.3 中，$N \sim Poi(\eta)$，根据泊松分布的性质有，

$$\begin{aligned} N_0(Z) &\sim Poi(\eta\pi_0 F_0(Z)) \\ N_1(Z) &\sim Poi(\eta\pi_1 F_1(Z)), \end{aligned} \tag{2.42}$$

已知，$N_+(Z)$,$N_0(Z)$ 的值有 2 种情况，所以根据条件分布的性质有，当 $N_+(Z) > 0$ 时，有

$$N_0(Z) \mid N_+(Z) \sim Bi(N_+(Z), \pi_0 F_0(Z)/F(Z)) \tag{2.43}$$

但是 $\pi_0 F_0(Z)/F(Z)) = \phi(Z)$,

$$\begin{aligned} Pr\{N_+(Z) = 0\} &= exp(e_+(Z)), \\ E\{Fdp(Z) \mid e_+(Z)\} &= E\left(\frac{e_0(Z)}{e_+(Z)} \mid e_+(Z)\right) = \phi(Z) \end{aligned} \tag{2.44}$$

上式的概率为 $1 - exp(e_+(Z))$, 很明显，$E\{Fdp(Z) \mid e_+(Z) = 0\}$ 的概率是 $exp(e_+(Z))$.

在大规模检验的应用中，有效情形的比例 π_1 通常非常接近于 1，在 Z 中 $e_+(Z)$ 可能非常小，正如（2.39）显示，$\overline{Fdr}(Z)$ 是有偏的，一个简单的修正可以去除这个有偏性，$\overline{Fdr}(Z) = e_0(Z)/N_+$.

$$\widetilde{Fdr}(Z) = e_0(Z)/(N_+) + 1 \tag{2.45}$$

引理 2.4 在（2.36）,（2.40）泊松 - 独立性假设条件下有，

$$E\{\widetilde{Fdr}(Z)\} = E\{Fdp(Z)\} = \phi(Z) \cdot [1 - exp(e_+(Z))] \tag{2.46}$$

练习 2.7 证明式 (2.46)

证明:

假定 $X \sim Poi(\lambda)$, 则

$$\begin{aligned} E\left(\frac{1}{X+1}\right) =& e^{-\lambda}\sum_{i=1}^{\infty}\frac{\lambda^k}{k!} \\ =& \frac{e^{-\lambda}}{\lambda}\sum_{i=1}^{\infty}\frac{\lambda^{k+1}}{(k+1)!} \\ =& \frac{1-e^{-\lambda}}{\lambda} \end{aligned}$$

其中利用了 $1+\lambda+\frac{\lambda^2}{2!}+...=e^{\lambda}$.

由于 $N_+ \sim Poi(\eta F(Z))$,

$$\begin{aligned} E(\widetilde{Fdr}(Z)) =& e_0 E\left(\frac{1}{N_+ + 1}\right) \\ =& \frac{e_0}{\eta F(Z)}(1-e^{-\eta F}) \\ =& \phi(1-e^{-e_+}) \end{aligned}$$

至于 $E\{\mathbf{F}dp(\boldsymbol{Z})\} = \phi(\boldsymbol{Z})\cdot[1-exp(-e_+(\boldsymbol{Z}))]$ 在定理 2.3 中已得证。

综上, 式 (2.46) 得证。

在这与其说泊松假设是必要的，不如说有了它使解决问题问题更简便了。在独立性条件下，$\widetilde{\boldsymbol{F}dr}(\boldsymbol{Z})$ 为 $\boldsymbol{E}\{\boldsymbol{F}dp(\boldsymbol{Z})\}$ 的近似无偏估计。随着 $e_+(\boldsymbol{Z})$ 的增加，$\widetilde{\boldsymbol{F}dr}(\boldsymbol{Z})$ 与 $\boldsymbol{E}\{\boldsymbol{F}dp(\boldsymbol{Z})\}$ 都以指数的速度靠近贝叶斯错误发现率 $\phi(\boldsymbol{Z})$。一般来说，当 $\phi(\boldsymbol{Z})$ 较大时，即 $\phi(\boldsymbol{Z}) \geq 10$ 时，$\widetilde{\boldsymbol{F}dr}(\boldsymbol{Z})$ 是 $\phi(\boldsymbol{Z})$ 的一个相当准确的估计量，但是当 $\phi(\boldsymbol{Z})$ 的值较小时，有偏性极大而且极具不稳定。

2.5 独立性与相关性

独立性假设在大规模检验这一领域中起到了非常重要的作用，尤其对于错误发现率理论，当然我们会在第四章中加以讨论。在实际应用中独立性假设也是一个危险的假设！

图 2.4 描述了 DTI(磁共振弥散张量成像) 数据的一部分来说明一项关于弥散张量成像的研究，当然这项研究是通过比较 6 个阅读困难的孩子与正常的孩子的大脑活动获得的。(DTI 设备测量大脑中液体的流量而形成各样式的磁共振图像。) 对 N=15443 个立体像素 (三维大脑坐标) 进行双样本检验产生了相应的 z 值，在对于有阅读困难的与正常的孩子之间没有差异这种原假设下, 每一个 $z_i \sim \boldsymbol{N}(0,1)$，就如同式 (2.6) 那样。

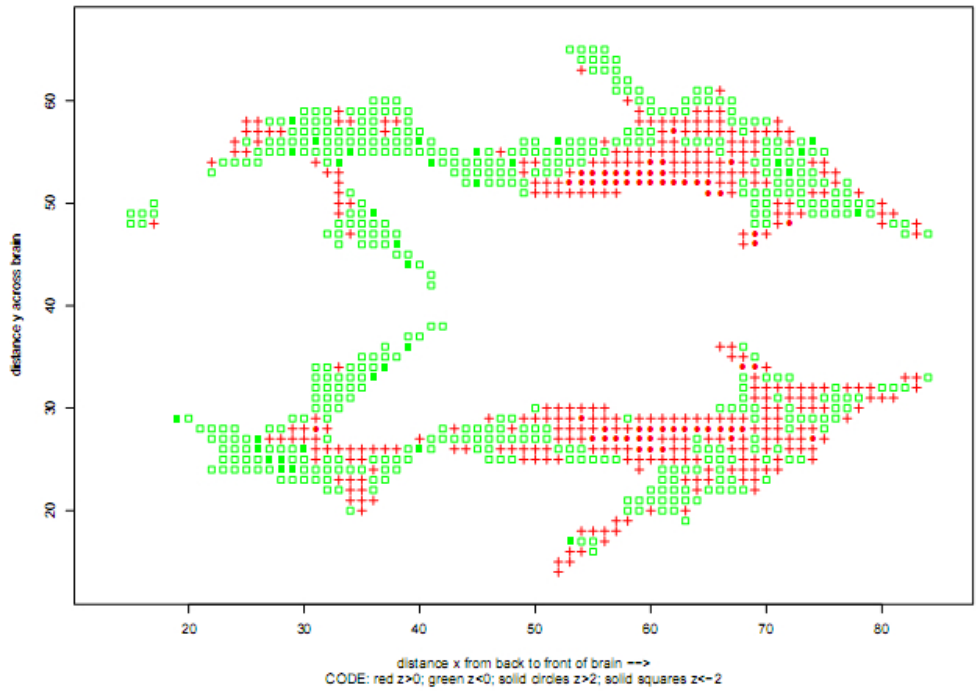


图 2.4磁共振弥散张量成像比较 6 个阅读困难与 6 个正常的孩子。对应 z 值有 N=15443 个

这张图形只显示了一个单独的大脑横截面，其中包含 15443 个立体像素中的 848 个。红色的代表 $z_i > 0$, 绿色的代表 $z_i < 0$, 实心圆代表 $z_i > 2$, 实心正方形代表 $z_i < 2$。z 值之间的这种空间的相关性是显著的: 红色之间离的很近，绿色之间也离的很近。在 x=60 附近对于实心红圆的那种对称性是尤为显著的。

对于 DTI 数据，很明显，独立性是一个不好的假定。在第八章中我们会对此深有体会。虽然对于前列腺数据，它似乎还不能算一个不好的假定，但是对于许多微阵列研究来说，独立性是一个更加糟糕的假定。把大脑几何学等价转化为微阵列进行研究不是一件容易的事。因为我们不能绘制像图（2.4）那样启发性的图片，而且相关性仍然是一个大问题。第七章将讨论相关性如何像（2.39）那样影响我们得不到精确地结果。

2.6 利用其他人的经验 II

对于贝叶斯分层模型，

$$\mu \sim g(.) \quad \text{和} \quad z \mid \mu \sim N(\mu, 1), \tag{2.47}$$

其中，g 表示 μ 的一些先验密度函数，我们将这里的“密度”涵盖了含有离散成分的 g 的概率。图（1.1）的詹姆斯 -斯坦统计量（下面均称之为 JS 统计量）为：

$$g(\mu) \sim N(M, A)。 \tag{2.48}$$

模型（2.47）也适用于联立假设检验：现在我们令

$$g(\mu) = \pi_0 \Delta_0(\mu) + (1 - \pi_0 g_1(\mu)) \tag{2.49}$$

其中 Δ_0 表示 $\mu = 0$ 时的一个脉冲函数（又称为狄克拉函数），g_1 表示备择 μ_i 的一个先验密度。（这就使得两群体模型的形式具有 $f_1(z) = \int \varphi(z-\mu)g_1(\mu)d\mu$。）对于基因 1 的前列腺数据，图（1.1）“其他的”就是所有的其他的基因，表示为 $z_2, z_3, ..., z_{6033}$. 这些可以估计先验分布（2.49）中的 π_0 和 g_1。最后，我们将先验和 $z_1 \sim N(\mu, 1)$ 相结合，通过贝叶斯理论来估计如基因 1 无差别的概率等各种数值。

如 $\overline{Fdr}(Z)$ 的完美构建能够巧妙地处理 $g(\mu)$ 的实际估计值，这将在十一章中进行更深入的讨论。这里想要表达的主要观点是基因 1 的信息正通过”学习”其他基因而不断改进。“哪些其他的”？这是一个非常重要的问题。该问题将在第十章进行讨论。

我们可以发现（2.48）中的 $g(\mu)$ 要比（2.49）中的更加顺滑，这一事实说明了在假设检验的背景下进行估计的困难。在（1.25）中，即使 N 小到 10，JS 统计量也是相当有效的。然而，如（2.39）所暗示的那样，我们需要 N 达到几百或几千才能进行精确地贝叶斯假设检验。这些“计算效率”将在第七章中进行进一步的提高。

注释

Benjamini 和 Hochberg 在 1995 年具有里程碑意义的文章中一个当前重要假设检验算法的内容引出了错误发现率，这一部分是第 4 章的主要内容。Efron,Tibshirani, Storey 和 Tusher（2001）再次在一篇关于经验贝叶斯的文章中引出了 fdr 算法，介绍了局部错误发现率。Storey (2002, 2003) 定义了“正错误发现率”，

$$pFdr(Z) = E\{N_0(Z)/N_+(Z)|N_+(Z) > 0\} \tag{2.50}$$

在公式 2.3 的表示中，并展示了如果 zi 是独立同分布的，

$$pFdr(Z) = \Phi(Z)$$

贝叶斯和经验贝叶斯微阵列技术的各种联合已经被提出，例如 Newton, Noueiry, Sarkar 和 Ahlquist (2004) 应用更为正式的贝叶斯分层模型。图 2.2 的曲线来自 Genovese 和 Wasserman (2004)，它在文中被用于推广 Benjamini-Hochberg 步骤的渐进性。Johnstone 和 Silverman (2004) 认为在原假设下 0 可能更接近于 1，而不是我们这里的结果。

双组模型非常基础以至于无法找到其提出者，但它是由 Efron (2008a) 命名并推广的。它还会在随后的章节中出现。更为专业的模型（2.47）也将会再次出现，并在第 11 章的预测理论中扮演重要的角色。在 Brown (1971) 和 Stein (1981) 的文章中，它对于深度研究多元正态均值向量估计意义重大。

前列腺癌的研究来自 Singh et al. (2002)。图 2.4 中 DTI 的数据是基于 Schwartzman, Dougherty 和 Taylor (2005) 的成果。

对于两样本对比 t 检验（或者与之相似的 Wilcoxon 检验）是合适的选择，但其他统计检验也被提出。Tomlins et al. (2005), Tibshirani 和 Hastie (2007), 以及 Wu (2007) 研究的分析理论关注随机大规模回复，这一文中用于识别基因的想法更偏向于无关效应。

对于与错误发现相关的各种概念基本都是标准化的，但可能容易混淆。这里有一个简单的专有名词表。

术语	定义
$f_0(z)$ 和 $f_1(z)$	在双组模型（2.7）中对于 z 值的原假设与被择假设的密度函数
$F_0(Z)$ 和 $F_1(Z)$	对应的概率分布函数（2.10）
f(z) 和 F(Z)	混合密度和分布（2.11）-（2.12）
Fdr(Z) 和 $\phi(Z)$	贝叶斯错误发现率 $Pr\{null\|z\in Z\}$ (2.13) 的两种命名
fdr(z)	局部贝叶斯错误发现率（2.14），也定义为 $\phi(z)$
Fdp(Z)	错误发现比率（2.28），例如在零分布中 zi 占 Z 的比重
$\overline{F}(Z)$	经验概率分布 $\sharp\{z_i\in Z\}/N$ (2.20)
$\overline{Fdr}(Z)$	用替代 F(Z)(2.12) 而得到的 Fdr(Z) 的经验贝叶斯估计值
FDR(Z)	错误发现比率 Fdp(Z) 的期望值
$N_0(Z), N_1(Z)$ 和 $N_+(Z)$	Z 中零、非零和总体 z 值的数量，参见图 2.3
$e_0(Z), e_1(Z)$ 和 $e_+(Z)$	上述各值的期望，参见（2.23）