模型的建立和遇到的问题

数 33 赵丰

December 26, 2016

不考虑各个位点的相关性，可以尝试将原 RNA 问题看成如下的问题：

为比较小明和小刚学习能力的强弱，可以收集他们历次考试的成绩和他们为备考所花费的时间的统计数据，设小明的成绩用随机变量 X 表示，小明历次成绩是 X 的随机抽样，记为 $x_1, x_2, ...x_N$, 小明的备考所花的时间用随机变量 Y 表示，历次备考时间分别为 $y_1, y_2, ...y_N$, 小刚历次成绩 X' 的样本值分别为 $x'_1, x'_2, ..x'_N$, 备考时间 Y' 的样本值分别为 $y'_1, y'_2, ..y'_N$.

一般可认为每个人的成绩除以他为这次考试付出的时间可用来衡量他的学习能力的强弱, 定义 $Z = \frac{X}{Y}, Z' = \frac{X'}{Y'}$, 则随机变量 Z 和 Z' 反映了小明或小刚学习能力的强弱。记 $z_i = \frac{x_i}{y_i}, z'_i = \frac{x'_i}{y'_i}, i = 1, 2, ..N$,$\bar{z}$ 是 Z 的样本均值, 即 $\bar{z} = \frac{1}{N}\sum_{k=1}^{n} z_i$. 同理规定 $\bar{z'}$ 的含义。通过比较 $\bar{z}$ 和 $\bar{z'}$ 的相对大小，可以比较出两人学习能力的强弱。但为了减少比较的误差，还要考虑样本数据的方差对判断结果的影响。设 $S_z^2 = \frac{1}{N-1}\sum_{k=1}^{n}(z_i - \bar{z})^2$, 则 S_z^2 为 Z 的样本方差, 同理规定 $S_z'^2$ 的含义，为书写方便，用 $\vec{x}$ 表示 $(x_1, x_2, ...x_N)$，同理规定 $\vec{y}, \vec{x'}, \vec{y'}$ 的含义;

构造的统计量为:

$$\tilde{\lambda}(\vec{x}, \vec{y}, \vec{x'}, \vec{y'}) = \frac{\bar{z} - \bar{z'}}{\sqrt{S_z^2/N + S_z'^2/N}} \tag{1}$$

现假设小明的成绩 X 服从 $N(\mu_1, \sigma_1^2)$, 他他为该次考试付出的时间 Y 服从 $N(\mu_2, \sigma_2^2)$，小刚的成绩 X' 和付出时间 Y' 也是不相关的正态随机变量。为比较两人学习能力的强弱，现在做一个假设检验：

H0: 两人学习能力没有明显差异, 即 $Z = Z'$

在对 X, Y, X', Y' 做出正态性和独立性的假设下，Z 的 pdf(概率密度函数) 可以借助标准正态的分布函数显示地表示出来 (详见 Appendix

1)，但统计量

$$\lambda = \frac{\bar{Z} - \bar{Z}'}{\sqrt{S_Z^2/N + S_Z'^2/N}} \tag{2}$$

的分布无法显式地写出。因此我们进一步地假设 Y 和 Y' 的方差很小 $(\sigma_2 - > 0)$，这时 Y 和 Y' 可以近似用它们的均值代替，Z 近似服从 $N(\frac{\mu_1}{\mu_2}, (\frac{\sigma_1}{\mu_2})^2)$ 的正态分布, 关于这一点的具体论证，详见 Appendix 2. 在这种假设下，假设检验是比较两个正态分布的均值是否相等，这是著名的 **Behrens-Fisher Problem**.

1 Appendix 1:Z 的 pdf 的推导

已知 $X\ N(\mu_1, \sigma_1^2), Y\ N(\mu_2, \sigma_2^2)$，X 与 Y 独立, 求 $Z = \frac{X}{Y}$ 的概率密度函数。根据两个随机变量商的密度公式：如果随机变量 (ξ, η) 有联合概率密度函数 $p(x, y)$, 则它们的商 $\frac{\xi}{\eta}$ 的密度函数为

$$p(z) = \int_{-\infty}^{+\infty} |y| p(zy, y) dy \tag{3}$$

(X, Y) 的联合密度函数为

$$p_{\xi/\eta}(x, y) = \frac{1}{2\pi\sigma_1\sigma_2} exp(-\frac{(x-\mu_1)^2}{2\sigma_1^2} - \frac{(y-\mu_2)^2}{2\sigma_2^2}) \tag{4}$$

记

$$K(z) = \frac{z\mu_1}{\sigma_1^2} + \frac{\mu_2}{\sigma_2^2} \tag{5}$$

$$I(z) = \frac{z^2}{\sigma_1^2} + \frac{1}{\sigma_2^2} \tag{6}$$

$$T = \frac{\mu_1^2}{2\sigma_1^2} + \frac{\mu_2^2}{2\sigma_2^2} \tag{7}$$

将 (4) 式代入 (3) 式，并将 e 指数对 y 配方得

$$p_Z(z) = \frac{1}{2\pi\sigma_1\sigma_2} exp(-T + \frac{K(z)^2}{2I(z)}) \int_{-\infty}^{+\infty} |y| exp(\frac{-I(z)}{2}(y - \frac{K(z)}{I(z)})^2) dy \tag{8}$$

对于 ye^{-y^2} 型的积分，可利用奇偶性和分部积分显式地计算出结果，对于 $exp(-y^2)$ 型的积分，可利用标准正态的分布函数 $\Phi(x)$ 表示结果。

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} exp(-\frac{u^2}{2}) du \tag{9}$$

最后可推导出 Z 的 pdf 为

$$p_Z(z) = \frac{1}{\pi\sigma_1\sigma_2} exp(-T + \frac{K(z)^2}{2I(z)})(\frac{exp(-\frac{K(z)^2}{2I(z)})}{I(z)} + \frac{\sqrt{2\pi}K(z)}{I(z)^{3/2}}(\Phi(\frac{K(z)}{\sqrt{I(z)}}) - \frac{1}{2})) \tag{10}$$

2 具体论证当 $(\sigma_2 -> 0)$，$p_Z(z)$ 趋近于 $N(\frac{\mu_1}{\mu_2}, (\frac{\sigma_1}{\mu_2})^2)$ 的 pdf

在 (10) 式的基础上，当 $(\sigma_2 -> 0)$ 时，$T(z) -> +\infty$，由于 $T(z)$ 在负指数位置上，故最右边括号中的第一项迅速衰减为 0。对于右边括号中的第二项，由 (5),(6) 式可知，$\frac{K(z)}{\sqrt{I(z)}}$ 关于 $\frac{1}{\sigma_2}$ 无穷大的阶为 $O(\frac{1}{\sigma_2})$, 故

$$\lim_{\sigma_2 \to 0}(\Phi(\frac{K(z)}{\sqrt{I(z)}}) - \frac{1}{2}) = \frac{1}{2} \tag{11}$$

$$\text{also } \lim_{\sigma_2 \to 0} \frac{K(z)}{\sigma_2 I(z)^{3/2}} = \mu_2 \tag{12}$$

$$\begin{aligned} \text{and } \lim_{\sigma_2 \to 0}(-T + \frac{K(z)^2}{2I(z)}) &= \frac{\mu_2^2}{2\sigma_2^2}(\frac{1}{\frac{\sigma_2^2 z^2}{\sigma_1^2} + 1} - 1) + \frac{\mu_2 z \mu_1}{2(\sigma_2^2 z^2 + \sigma_1^2)} \\ &+ \frac{\frac{z\mu_1\mu_2}{\sigma_1^2}}{2(\frac{\sigma_2^2 z^2}{\sigma_1^2} + 1)} - \frac{\mu_1^2}{2\sigma_1^2} = -\frac{(\mu_2 z - \mu_1)^2}{2\sigma_1^2} \end{aligned} \tag{13}$$

整理（11-13）式可得

$$\lim_{\sigma_2 \to 0} p_Z(z) = \frac{1}{\sqrt{2\pi}\frac{\sigma_1}{\mu_2}} exp(-\frac{(z - \frac{\mu_1}{\mu_2})^2}{2(\frac{\sigma_1}{\mu_2})^2}) \tag{14}$$

结论得证。

3 下一步的打算

1. 学习 **Behrens-Fisher problem** 的一种解决方法。

4 参考文献

References

[1] 生物信息学