1. Galton 的 "回归效应".

   (a) 略. Happy reading!

   (b) 根据线性回归参数 a, b 的计算公式,

$$a = \frac{\langle ts\rangle_{\mathcal{D}} - \langle t\rangle_{\mathcal{D}} \langle s\rangle_{\mathcal{D}}}{\langle t^2\rangle_{\mathcal{D}} - \langle t\rangle_{\mathcal{D}}^2},$$
$$b = \frac{\langle t^2\rangle_{\mathcal{D}} \langle s\rangle_{\mathcal{D}} - \langle t\rangle_{\mathcal{D}} \langle ts\rangle_{\mathcal{D}}}{\langle t^2\rangle_{\mathcal{D}} - \langle t\rangle_{\mathcal{D}}^2}. \tag{1}$$

   其中 (均值号的下标 $\mathcal{D}$ 略去不写, 认为数据集规模足够大, 可以复现原题表述的正态分布),

$$\langle t\rangle = \langle t\rangle = \mu, \tag{2}$$
$$\left\langle (t - \langle t\rangle)^2 \right\rangle = \left\langle t^2 \right\rangle - \langle t\rangle^2 = \sigma^2, \tag{3}$$
$$\left\langle (s - \langle s\rangle)^2 \right\rangle = \left\langle s^2 \right\rangle - \langle s\rangle^2 = \sigma^2, \tag{4}$$
$$\langle (t - \langle t\rangle)(s - \langle s\rangle)\rangle = \langle ts\rangle - \langle t\rangle \langle s\rangle = r\sigma^2. \tag{5}$$

   于是,

$$a = \frac{r\sigma^2}{\sigma^2} = r,$$
$$b = \frac{(\sigma^2 + \mu^2)\mu - \mu(r\sigma^2 + \mu^2)}{\sigma^2} = \mu(1 - r). \tag{6}$$

   现在, 求解得到的回归模型为

$$\hat{s} = rt + \mu(1 - r), \tag{7}$$

   式中, $\hat{s}$ 代表身高预测值. 下面讨论回归效应.

   - 从全体平均而言, 注意到

$$\langle \hat{s}\rangle = r\langle t\rangle + \mu(1 - r) = \mu = \langle t\rangle, \tag{8}$$

     所以, 预测得到的身高. (原题表述不太准确, 考虑到这一层面即可得满分)

   - 考虑那些 $t > \mu$ 的数据点. 记 $\mathcal{D}' = \{(t, s) \in \mathcal{D} : t > \mu\}$. 此时, 不等式

$$\langle \hat{s}\rangle_{\mathcal{D}'} = r\langle t\rangle_{\mathcal{D}'} + \mu(1 - r) < \langle t\rangle_{\mathcal{D}'} \tag{9}$$

     成立, 这是因为 $\langle t\rangle_{\mathcal{D}'} > \mu$. 所以, 在高于平均身高 μ 的群体 $\mathcal{D}'$ 中, 儿子的平均身高 $\langle \hat{s}\rangle_{\mathcal{D}'}$ 矮于父亲的平均身高 $\langle t\rangle_{\mathcal{D}'}$. 这就是 Galton 发现的回归效应.

   (c) 这里 s 和 t 是对称的, 上述的一切结果将 s, t 交换地位后依然成立. 回归模型为

$$\hat{t} = as + b, \tag{10}$$

   而同样地: 全体样本点上, 父亲与儿子有着相等的平均身高; 在高于平均身高的儿子群体上, 父亲的平均身高矮于他们的平均身高.

   (d) 这是一个 "文字游戏"! 这两句话同时成立, 但各自所考察的 (用于计算均值的) 样本集 (阅读材料中的 "总体") 存在区别. "父亲平均矮于儿子" 是在样本集 $\mathcal{D}'' = \{(t, s) \in \mathcal{D} : s > \mu\}$ 上成立, "儿子平均矮于父亲" 是在样本集 $\mathcal{D}' = \{(t, s) \in \mathcal{D} : t > \mu\}$ 上成立.

2. 这枚铜钱 “是否” 为狄青钱分别记为 H 与 $\bar{H}$, 单次 “正面朝上” 为 E. 此时, 先验分布与 (单次试验的) 似然函数为

$$p(H) = p(\bar{H}) = \frac{1}{2}, \tag{11}$$

$$p(E|H) = 1, \tag{12}$$

$$p(E|\bar{H}) = \frac{1}{2}. \tag{13}$$

(a) 记 $E_1 := E^3$ 代表连续抛掷 3 次都为正面朝上. 由于各次试验是条件独立的, 我们有

$$p(E_1|H) = p^3(E|H) = 1, \tag{14}$$

$$p(E_1|\bar{H}) = p^3(E|\bar{H}) = \frac{1}{8}. \tag{15}$$

所以,

$$\begin{aligned} p(H|E_1) &= \frac{p(E_1|H)p(H)}{p(E_1|H)p(H) + p(E_1|\bar{H})p(\bar{H})} \\ &= \frac{1 \times \frac{1}{2}}{1 \times \frac{1}{2} + \frac{1}{8} \times \frac{1}{2}} \\ &= \frac{8}{9}. \end{aligned} \tag{16}$$

(b) 记 $E_2 := E^4$ 代表连续抛掷 4 次都为正面朝上. 同理, 有

$$\begin{aligned} p(H|E_2) &= \frac{1 \times \frac{1}{2}}{1 \times \frac{1}{2} + \frac{1}{16} \times \frac{1}{2}} \\ &= \frac{16}{17}. \end{aligned} \tag{17}$$

读者不难算出其通式

$$p(H|E^n) = \frac{2^n}{2^n + 1}. \tag{18}$$

(c) 记 $E_3 := E^3\bar{E}$, 代表 “正正正反” 的试验结果. 由于

$$p(E_3|H) = p^3(E|H)p(\bar{E}|H) = 0, \tag{19}$$

所以 $p(H|E_3) = 0$. 这是显然的, 狄青钱不可能反面朝上.