

## 作业 1

1. 分析下面两个因果论断为什么是错误的:
  - (a) 美国大萧条时期,文化程度较高的群体失业率较低。所以教育程度越高,失业机会越小。
  - (b) 化石分析发现:地域分布广的物种,其生存年代更为久远。所以一个物种如果在地域上分布广泛,则难以绝迹。
2. (Cross-over 设计) 医学研究中的临床试验 (clinical trial) 是随机化控制试验,该方法随机地安排处理 (treatment, 比如新药) 和对照 (control, 比如旧药),以期从总体中随机抽取的两组研究对象在各个方面上比较均衡或相似。但是在某些情况下(比如研究对象比较少),随机化可能不能很好地平衡两组(即在某些指标上两组差异较大)。为了避免这种情况的发生,有时也采用 cross-over 设计方法,即每个研究对象(病人)先接受一种药物治疗,一段时间后再接受另外一种药,次序随机决定。所以 cross-over 设计中每个研究对象的对照是其自身,从而可以有效地避免很多因素(比如性别、生活习惯、职业、教育背景等等)的干扰。但 cross-over 设计也有较大的局限性,试讨论该设计可能存在的问题。
3. 为期半年的减肥训练班结束的时候,体重超标最严重的几人的减重效果都较明显,这是否意味着训练有效?假设只有 3/4 的学员坚持到训练班活动结束后,为了衡量训练效果,假设我们比较这些坚持到底的学员训练开始和结束时候的体重,并以成对 t 检验考察训练效果的显著性。你认为这种检验方法是否恰当?如果不恰当,应该如何检验训练效果?
4. 霍乱传播案例数据(课件 1 案例 2, P27)提供了除 Southwark & Vauxhall 和 Lambeth 之外其它供水公司客户的霍乱死亡率(下表“Rest of London”),试检验 SV, Lambeth, Rest of London 客户死亡率是否相同(提示:  $3 \times 2$  列联表的 Pearson 卡方)。

Table 2. Death rate from cholera by source of water. Rate per 10,000 houses. London. Epidemic of 1854. Snow's table IX.

	No. of Houses	Cholera Deaths	Rate per 10,000
Southwark & Vauxhall	40,046	1,263	315
Lambeth	26,107	98	37
Rest of London	256,423	1,422	59

5. HIP 乳腺癌临床试验案例（课件 1 案例 1, p21）列出了处理组（含接受筛查的 Screened 组和拒绝筛查的 Refused 组）和对照组的总人数（第 2 列），乳腺癌死亡人数（第 3 列）、比率（第 4 列），因其它疾病死亡的人数（第 5 列）、比率（第 6 列）。通过该数据你能否提供 Screened 组和 Refused 组存在差异的证据？

Table 1. HIP data. Group sizes (rounded), deaths in 5 years of followup, and death rates per 1000 women randomized.

	Group size	Breast cancer No.	Breast cancer Rate	All other No.	All other Rate
Treatment					
Screened	20,200	23	1.1	428	21
Refused	10,800	16	1.5	409	38
Total	31,000	39	1.3	837	27
Control	31,000	63	2.0	879	28

6. 假设  $a \sim B(n_1, p_1), c \sim B(n_0, p_0)$ , 记  $\hat{p}_1 = a/n_1, \hat{p}_0 = c/n_0$  分别是  $p_1, p_0$  的样本估计。零假设为  $H_0: p_1 = p_0$  (齐一性), 记  $\hat{p} = (a + c)/n$  为零假设成立时  $p_1 = p_0$  的估计, 其中  $n = n_1 + n_0$ 。  $H_0$  的两样本 z-检验为

$$z = \frac{\hat{p}_1 - \hat{p}_0}{\sqrt{(1/n_1 + 1/n_0)\hat{p}(1 - \hat{p})}}$$

另一方面, 记  $b = n_1 - a, d = n_0 - c, m_1 = a + c, m_0 = b + d$ , 以列联表的形式表示上述数据:

		$y$		
		1	0	
$x$	1	$a$	$b$	$n_1$
	0	$c$	$d$	$n_0$
		$m_1$	$m_0$	$n$

则该列联表独立性检验的 Pearson 卡方统计量为

$$X^2 = \frac{n(ad - bc)^2}{n_0 n_1 m_0 m_1}$$

试验证等于  $X^2 = z^2$  (即: 齐一性检验等价于独立性检验)。