属性数据分析模拟试卷

参考 2023、2024 年题型与难度；范围限于复习提纲：Ch1–4、Ch6.1、Ch8

满分 100 分

一、填空/选择题（共 **25** 分）

1)（3 分）设 $(Y_1, Y_2, Y_3) \sim \text{Mult}(N; \pi_1, \pi_2, \pi_3)$。令 $A = Y_1 + Y_2$, $B = Y_2 + Y_3$，则 $\text{Cov}(A, B) =$_____，$\rho(A, B) =$_____。

2)（3 分）Fisher 精确检验的 P 值 p 是否为统计量？_____（填“是”或“否”）。采用规则 $p < 0.05$ 拒绝原假设时，第一类错误率 _____ 0.05（填“等于/不大于/不小于/不能确定”）。

3)（2 分）在病例–对照研究中，若前瞻性关系满足 $\text{logit}\, P(Y = 1 \mid x) = \alpha + \beta^\top x$，通常可以相合估计的是 _____，通常不能直接相合估计的是 _____。

4)（3 分）广义线性模型由三部分构成：_____、_____、_____。二项分布的典则联系函数是 _____。

5)（2 分）Poisson 回归 $\log \mu_i = \log t_i + \alpha + \beta x_i$ 中，$\log t_i$ 称为 _____；在暴露量相同且其他变量不变时，x 增加 1，均值/率乘以 _____。

6)（3 分）配对二分数据表中，不一致对计数为 n_{10} 与 n_{01}。McNemar 检验统计量为 _____；在原假设下，给定 $n^* = n_{10} + n_{01}$ 后，$n_{10} \mid n^* \sim$ _____。

7)（3 分）名义三类别响应以第 3 类为基线：$\log(\pi_1/\pi_3) = \eta_1$, $\log(\pi_2/\pi_3) = \eta_2$。则 $\pi_1 =$_____，$\pi_2 =$_____，$\pi_3 =$_____。

8)（2 分）Poisson GLM 出现超散布时，常用 $\hat{\phi} =$_____ 估计散布参数，并将常规模型标准误乘以 _____。

9)（2 分）2×2 表中 $\text{OR} = 1$ 等价于 _____。在疾病罕见时，OR 近似 _____；病例–对照研究通常能稳健估计 _____，而不能直接估计 _____。

10)（2 分）采用条件似然处理高维冗余参数的关键步骤是：_____；然后 _____。

二、解答题（共 **75** 分）

1)（10 分）某疾病检测试剂的灵敏度为 0.92，特异度为 0.96。

(i) 若人群患病率为 0.005，求阳性预测值 PPV，并说明阳性结果是否足以支持大规模确诊决策。

(ii) 若人群患病率为 0.10，再次求 PPV。简要解释患病率为何会显著影响 PPV。

2)（10 分）构造一个具体数值例子说明 Simpson 悖论：在青年组和老年组内，甲地死亡率都高于乙地；但不分层合并后，乙地总体死亡率高于甲地。请给出表格并解释原因。

3)（15 分）两组随访数据如下，设事件数服从独立 Poisson 分布，并建立 $\log \mu_i = \log t_i + \alpha + \beta x_i$，其中 $x = 1$ 表示处理组，$x = 0$ 表示对照组。

组别	x	事件数 y	暴露量 t
对照组	0	24	1200
处理组	1	18	1800

(i) 求 $\hat{\alpha}, \hat{\beta}$，并解释 $\exp(\hat{\beta})$。

(ii) 构造率比 $\exp(\beta)$ 的近似 95% Wald 置信区间。

(iii) 用 Wald 检验检验 $H_0 : \beta = 0$，给出统计量和近似 P 值。

4)（15 分）二值解释变量 X 与二值响应 Y 的独立样本列联表如下：

	$Y = 1$	$Y = 0$	合计
$X = 1$	45	55	100
$X = 0$	25	75	100

考虑 logistic 模型 $P(Y = 1 \mid X = x) = \operatorname{expit}(\alpha + \beta x)$。

(i) 求 $\hat{\alpha}, \hat{\beta}$，并给出 $\mathrm{OR} = \exp(\beta)$ 的估计及 95% Wald 置信区间。

(ii) 在 $H_0 : \beta = 0$ 下求得分检验统计量，并说明其与 2×2 独立性 Pearson 卡方检验的关系。

5)（10 分）某名义三分类响应 $Y \in \{A, B, C\}$ 以 C 为基线，拟合模型

$$\log \frac{\pi_A}{\pi_C} = -1 + 0.4x, \qquad \log \frac{\pi_B}{\pi_C} = 0.5 - 0.2x.$$

(i) 求 $x = 2$ 时三个类别的拟合概率。

(ii) 解释两个斜率系数的含义。

(iii) 若去掉 x 后模型离差为 138.6，完整模型离差为 128.8，检验 x 的整体作用。

6)（15 分）同一批 $n = 200$ 个对象接受两种二分判断方法 X 与 Y，配对结果如下：

	$Y = 1$	$Y = 0$	合计
$X = 1$	60	28	88
$X = 0$	12	100	112

(i) 估计两个边缘成功率之差 $P(X = 1) - P(Y = 1)$，并给出近似 95% Wald 置信区间。

(ii) 作 McNemar 检验，给出统计量与近似 P 值。

(iii) 写出配对条件 logistic 模型中只由不一致对贡献的条件似然，并估计配对优势比及其 95% Wald 置信区间。

答案与解析

一、填空/选择题

1) 因 $A = N - Y_3,\ B = N - Y_1$，

$$\mathsf{Cov}(A, B) = \mathsf{Cov}(Y_3, Y_1) = -N\pi_1\pi_3.$$

又 $\mathsf{Var}(A) = N(1-\pi_3)\pi_3,\ \mathsf{Var}(B) = N(1-\pi_1)\pi_1$，故

$$\rho(A, B) = -\frac{\pi_1\pi_3}{\sqrt{\pi_3(1-\pi_3)\pi_1(1-\pi_1)}} = -\sqrt{\frac{\pi_1\pi_3}{(1-\pi_1)(1-\pi_3)}}.$$

2) 是；不大于。Fisher 精确检验在离散分布下通常偏保守，拒绝概率不超过名义水平。

3) 可相合估计斜率参数 β 以及由斜率给出的暴露优势比；通常不能直接相合估计截距 α、发病概率或风险比。

4) 随机部分、系统部分、联系函数；二项分布典则联系为 logit。

5) offset（偏置项/暴露量偏置项）；乘以 $\exp(\beta)$。

6)

$$W = \frac{(n_{10} - n_{01})^2}{n_{10} + n_{01}}, \qquad n_{10} \mid n^* \sim B(n^*, 1/2).$$

7)

$$\pi_1 = \frac{e^{\eta_1}}{1 + e^{\eta_1} + e^{\eta_2}}, \quad \pi_2 = \frac{e^{\eta_2}}{1 + e^{\eta_1} + e^{\eta_2}}, \quad \pi_3 = \frac{1}{1 + e^{\eta_1} + e^{\eta_2}}.$$

8) $\hat{\phi} = X^2/df$；乘以 $\sqrt{\hat{\phi}}$。

9) 独立；RR；OR；RR 或风险本身。

10) 找到冗余参数的充分统计量；给定该统计量构造条件似然并对感兴趣参数极大化。

二、解答题

1) 由 Bayes 公式，

$$\mathsf{PPV} = \frac{p \cdot \mathsf{sens}}{p \cdot \mathsf{sens} + (1-p)(1-\mathsf{spec})}.$$

(i) 当 $p = 0.005$，

$$\mathsf{PPV} = \frac{0.005 \times 0.92}{0.005 \times 0.92 + 0.995 \times 0.04} \approx 0.104.$$

即阳性者中真正患病比例约为 10.4%，假阳性占多数，不适合仅凭阳性结果作大规模确诊决策，应复核或结合更高特异度检测。

(ii) 当 $p = 0.10$，

$$\mathsf{PPV} = \frac{0.10 \times 0.92}{0.10 \times 0.92 + 0.90 \times 0.04} \approx 0.719.$$

患病率越高，阳性样本中真阳性来源越多；患病率很低时，即使特异度较高，庞大的非患病人群仍会产生大量假阳性。

2) 一个例子如下：

	甲地			乙地		
年龄层	死亡	总数	死亡率	死亡	总数	死亡率
青年	99	9900	1.0%	5	1000	0.5%
老年	10	100	10.0%	720	9000	8.0%
合计	109	10000	1.09%	725	10000	7.25%

分层看，甲地在青年和老年两层的死亡率都高于乙地；但甲地样本几乎都在低风险的青年层，乙地样本大多在高风险的老年层，因此合并后乙地总体死亡率更高。这是年龄构成混杂导致的 Simpson 悖论。

3) (i) 两组率分别为

$$\hat{\lambda}_0 = \frac{24}{1200} = 0.020, \qquad \hat{\lambda}_1 = \frac{18}{1800} = 0.010.$$

故

$$\hat{\alpha} = \log(0.020) = -3.912, \qquad \hat{\beta} = \log(0.010) - \log(0.020) = \log(0.5) = -0.693.$$

$\exp(\hat{\beta}) = 0.5$，表示处理组事件发生率估计为对照组的一半。

(ii) 对两组 Poisson 率比，

$$\widehat{SE}(\hat{\beta}) = \sqrt{\frac{1}{18} + \frac{1}{24}} = 0.312.$$

因此

$$\beta:\ -0.693 \pm 1.96(0.312) = [-1.304, -0.082],$$

$$\exp(\beta):\ [e^{-1.304}, e^{-0.082}] = [0.271, 0.921].$$

(iii)

$$Z = \frac{-0.693}{0.312} = -2.22, \qquad Z^2 = 4.94.$$

近似 P 值为 $P(\chi_1^2 \geq 4.94) \approx 0.026$。在 5% 水平下拒绝 H_0，处理组率显著较低。

4) (i)

$$\hat{\alpha} = \log \frac{25}{75} = -1.099,$$

$$\hat{\beta} = \log \frac{45/55}{25/75} = 0.898.$$

优势比估计为

$$\widehat{\mathrm{OR}} = e^{0.898} = \frac{45 \times 75}{55 \times 25} = 2.455.$$

其对数标准误为

$$\widehat{SE}\{\log(\widehat{\mathrm{OR}})\} = \sqrt{\frac{1}{45} + \frac{1}{55} + \frac{1}{25} + \frac{1}{75}} = 0.306.$$

故 95% Wald 区间为

$$\exp\{0.898 \pm 1.96(0.306)\} = [1.347, 4.473].$$

(ii) 在 H_0 下，公共成功概率估计为 $\hat{\pi} = 70/200 = 0.35$。得分为

$$U_\beta = 45 - 100(0.35) = 10,$$

校正掉截距后的信息量为

$$I_{\beta\beta\cdot\alpha} = \frac{100 \times 100}{200}\hat{\pi}(1-\hat{\pi}) = 11.375.$$

故

$$S = \frac{U_\beta^2}{I_{\beta\beta\cdot\alpha}} = \frac{100}{11.375} = 8.79.$$

该统计量恰好等于此 2×2 表独立性检验的 Pearson 卡方统计量，近似 P 值为 $P(\chi_1^2 \geq 8.79) \approx 0.0030$。

5) (i) 当 $x = 2$ 时，

$$\eta_A = -1 + 0.4(2) = -0.2, \qquad \eta_B = 0.5 - 0.2(2) = 0.1.$$

分母

$$1 + e^{-0.2} + e^{0.1} = 2.924.$$

故

$$\hat{\pi}_A = 0.280, \qquad \hat{\pi}_B = 0.378, \qquad \hat{\pi}_C = 0.342.$$

(ii) x 增加 1，类别 A 相对基线 C 的 odds 乘以 $e^{0.4} = 1.49$；类别 B 相对基线 C 的 odds 乘以 $e^{-0.2} = 0.82$。

(iii) 完整模型比约简模型少 2 个参数，故

$$\Delta D = 138.6 - 128.8 = 9.8, \qquad df = 2.$$

近似 $P = P(\chi_2^2 \geq 9.8) = e^{-9.8/2} \approx 0.00745$，说明 x 对名义响应的整体作用显著。

6) (i)

$$\hat{d} = \hat{P}(X=1) - \hat{P}(Y=1) = \frac{28-12}{200} = 0.080.$$

配对差的方差估计为

$$\widehat{\mathrm{Var}}(\hat{d}) = \frac{1}{n}\left\{\frac{n_{10}+n_{01}}{n} - \hat{d}^2\right\} = \frac{1}{200}(0.2 - 0.0064) = 0.000968,$$

标准误为 0.0311。95% Wald 区间为

$$0.080 \pm 1.96(0.0311) = [0.019, 0.141].$$

(ii)

$$W = \frac{(28-12)^2}{28+12} = 6.40.$$

近似 $P = P(\chi_1^2 \geq 6.40) \approx 0.0114$，在 5% 水平下拒绝两个边缘成功率相同的原假设。

(iii) 条件 logistic 中，只有不一致对提供关于 β 的信息。若 β 表示 X 相对 Y 的 log odds，则

$$L_c(\beta) = \left(\frac{e^\beta}{1+e^\beta}\right)^{28}\left(\frac{1}{1+e^\beta}\right)^{12}.$$

因此

$$\hat{\beta} = \log \frac{28}{12} = 0.847, \qquad \widehat{\mathrm{OR}}_{\mathrm{paired}} = \frac{28}{12} = 2.333.$$

标准误为

$$\widehat{SE}(\hat{\beta}) = \sqrt{\frac{1}{28} + \frac{1}{12}} = 0.345.$$

95% Wald 区间为

$$\exp\{0.847 \pm 1.96(0.345)\} = [1.187, 4.589].$$