中国科学技术大学管理学院统计与金融系

2020～2021 学年第二学期考试试卷

☑A 卷　　□B 卷

课程名称： 属性数据分析　**课程代码：** STAT4001.01

开课院系： 统计与金融系　**考试形式：** 半开卷

姓　名:_____　**学　号:** _____

题 号	1	2(1)	2(2)	2(3)	2(4)	2(5)	总 分
得 分							

（装订线内不要答题）

1. 选择/填空题（第 5 小题 3 分，其余每小题 2 分）

1) 记μ为响应变量的均值，g 为联系函数，X_1和X_2为预测变量，考虑如下 4 个广义线性模型：$g(\mu)=\alpha+\beta_1 x_1$；$g(\mu)=\alpha+\beta_2 x_2$；$g(\mu)=\alpha+\beta_1 x_1+\beta_2 x_2$；$g(\mu)=\alpha+\beta_1 x_1+\beta_2 x_2+\beta_{12} x_1 x_2$。则可以采用离差比较的模型共有_____对。

2) 在存在高维冗余参数η时，通常采用条件极大似然估计感兴趣参数θ，即采用给定某统计量 T 条件下的似然函数，T 应该取为_____
该条件极大似然估计在什么情况下相对于极大似然估计不会损失估计效率？

_____

3) 为研究某暴露因素对感兴趣的疾病的作用，现从感兴趣的人群中随机收集了 10 万个人的风险信息，计算结果得到相对风险（暴露因素水平 2 相对于水平 1）和优势比分别为 2.5 和 1.5，据此你可以做出哪些判断？（ AB ）

（A）该疾病比较常见；　　（B）水平 2 相对水平 1 风险更高；

（C）该疾病比较罕见；　　（D）水平 2 相对水平 1 风险更低。

4) 在用 R 软件分析有 k 个类别的属性预测变量 X 时，有两种编码方式，一种是用 factor 将 X 强行转换成属性变量，一种是自定义 $k-1$ 个哑变量，这两种方式在用似然比检验时有什么差异？_____

_____

5) 对数线性模型(XYZ, XW)对应的以 Y 为响应变量的 logistic 回归模型是(假设四个变量均为二分的，采用课本表 7.12 的记号)_____，两个模型下优势比估计结果是否一致？_____（“是”或“否”）。你的理由是：_____

_____

6) 比较课本上表 10.5 的显著性结果，请在选择模型时能给出合理的建议

_____

_____

7) 在分析课本上表 9.6 的失眠数据时，发现采用独立结构的广义估计方程法（GEE）得到的药物-时间交互效应估计标准差低于直接采用广义线性模型得到结果，其直观的原因是_____

_____

2. 解答题（5+6+10+12+12 分）

1) 2020 年 A 大学在大部分省份的高考录取平均分低于 B 大学，但根据媒体披露，A 大学的整体生源优于 B 大学，试仿照属性响应变量情形解释这一辛普森悖论形成的原因。

2) 从某方差大于均值的离散总体μ（可以取值非负整数）抽得样本 X，考虑假设检验问题$H_0: \mu = \mu_0 \leftrightarrow H_1: \mu \neq \mu_0$，试证明，基于泊松总体假定的 Wald 检验（检验统计量为 $T = (X - \mu_0)^2/\mu_0$）的实际极限第一类错误率大于名义检验水平α。

3) 为研究吸烟是否为心肌梗死的危险因素，进行了一项**病例-对照**研究，收集的数据中，心肌梗死患者中曾经吸烟和不曾经吸烟的人数分别为n_{11}和n_{21}，而正常人中曾经吸烟和不曾经吸烟的人数分别为n_{12}和n_{22}，记$n_{+1} = n_{11}+n_{21}$，$n_{+2} = n_{12}+n_{22}$。试利用δ法证明样本对数优势比 $\log\frac{n_{11}n_{22}}{n_{12}n_{21}}$ 的方差估计可取为 $1/n_{11} + 1/n_{12} + 1/n_{21} + 1/n_{22}$。【注】$n_{+1}$和$n_{+2}$是非随机的。

4) 记配对二分变量(X, Y)的 n 个独立复制为$(X_i, Y_i), i = 1, \ldots, n$。记$n_{jk} = \sum_{i=1}^{n} I(X_i = j, Y_i = k)$，$\pi_{jk} = P(X = j, Y = k)$。证明$(\pi_{12} + \pi_{21}, \pi_{12}/(\pi_{12} + \pi_{21}))$的充分统计量为$(n_{12}, n_{21})$，且$n_{12}|(n_{12} + n_{21}, n_{11}, n_{22}) \sim B(n_{12} + n_{21}, \pi_{12}/(\pi_{12} + \pi_{21}))$，并基于此条件分布推导假设检验问题$H_0: \pi_{12}=\pi_{21} \leftrightarrow H_1: \pi_{12} \neq \pi_{21}$的得分统计量。

5) 假设总体是参数为 2 和β的伽马分布（参数为α和β的伽马分布密度函数为 $\frac{\beta^\alpha}{\Gamma(\alpha)}x^{\alpha-1}e^{-\beta x}, x > 0$），记$X_1, \cdots, X_n$为从中抽得的简单随机样本。考虑假设检验问题$H_0: \beta = 1 \leftrightarrow H_1: \beta \neq 1$，试给出得分检验统计量及检验 P-值的表达式，并证明该 P 值的零极限分布为均匀分布。

1

1）5

2）η的充分统计量；总体分布（pdf 或 pmf）可以写成η的函数与θ的函数的乘积。

3）A,B

4）前者只能检验整个属性变量的效应，而后者还可以检验该属性变量任一水平的效应。

5）(X*Z)；否；尽管两个模型下真实优势比一样，但逻辑模型(X*Z)应与对数线性模型(XYZ)而不是(XYZ,XW)的估计结果一致。

6）当组内相关性非常显著时，更适宜选用 GLMM。

7）GEE 利用了独立性，但 GLM 没有，因此直观上 GEE 更有效。

2

1) 考虑只有两个省份 I 和 II 但简单情形。假设 A 大学与 B 大学在省份 I 但成绩一致高于省份 II 但成绩，且 A 大学招生指标主要放在省份 I 而 B 大学主要在省份 II，则 A 大学的总平均成绩接近于其在省份 I 的平均成绩而 B 大学的总平均成绩接近于其在省份 II 的平均成绩。从而 A 大学的总平均成绩高于 B 大学，尽管在每个省 A 大学平均成绩低于 B 大学。

2) 记原假设下总体的真实方差为 $\lambda\ \mu_0$，其中 $\lambda > 1$。则检验的第一类错误率为

$$P(T > K_1^{-1}(1-\alpha)) = P(\frac{(X-\mu_0)^2}{\lambda\ \mu_0} > \frac{K_1^{-1}(1-\alpha)}{\lambda}) \to 1 - K_1(\frac{K_1^{-1}(1-\alpha)}{\lambda}) > 1 - K_1(K_1^{-1}(1-\alpha)) = \alpha.$$

也即检验的第一类错误率的极限大于名义水平。

3) 样本优势比可以写成

$$OR = \log(n_{11}) - \log(n_{+1} - n_{11}) - \log(n_{12}) + \log(n_{+2} - n_{12})$$

因为$n_{11}|n_{+1} \sim B(n_{+1}, p_1), n_{12}|n_{+2} \sim B(n_{+2}, p_2)$, 且$n_{11} \perp n_{12}|n_{+1}, n_{+2}$，所以由$\delta$法知

$$\widehat{var}(OR) = \left(\frac{1}{n_{11}} + \frac{1}{n_{21}}\right)^2 n_{+1}\frac{n_{11}}{n_{+1}}\frac{n_{21}}{n_{+1}} + \left(\frac{1}{n_{12}} + \frac{1}{n_{22}}\right)^2 n_{+2}\frac{n_{12}}{n_{+2}}\frac{n_{22}}{n_{+2}} = \frac{1}{n_{11}} + \frac{1}{n_{21}} + \frac{1}{n_{12}} + \frac{1}{n_{22}}$$

4) 似然函数为

$$L = \prod_{j=1}^{2}\prod_{k=1}^{2} {\pi_{jk}}^{n_{jk}}$$

记$\theta = \frac{\pi_{12}}{\pi_{12}+\pi_{21}}$，$\eta = \pi_{12} + \pi_{21}$，则$\pi_{12} = \theta\eta$，$\pi_{21} = (1-\theta)\eta$，从而

$$L = {\pi_{11}}^{n_{11}}{\pi_{22}}^{n_{22}}\theta^{n_{12}}(1-\theta)^{n_{21}}\eta^{n_{12}+n_{21}}$$

所以由因子分解定理知$n_{12} + n_{21}$是(θ, η)的充分统计量。另外，由于$(n_{11}, n_{22}, n_{12} + n_{21}) \sim M(n; \pi_{11}, \pi_{22}, \eta)$，算得$n_{12}|(n_{11}, n_{22}, n_{12} + n_{21})$的分布为 $B(n_{12} + n_{21}, \theta)$。可以根据定义算得基于该条件分布的得分检验统计量为$\frac{(n_{12}-n_{21})^2}{n_{12}+n_{21}}$。

5) 根据定义容易算得得分检验统计量为 $S = \frac{(2n-\sum_{i=1}^{n} X_i)^2}{2n}$，因为在原假设之下，S 的零极限分布是自由度为 1 的卡方分布，所以 p 值$=P(\chi_1^2 > S) = 1 - K_1(S)$，其中$K_1$为自由度为 1 的卡方分布的分布函数。从而对 $0 < x < 1$，根据 S 的零极限分布是自由度为 1 的卡方分布这一事实，在零假设之下有

$$P(\text{p 值} \le x) = P(1 - K_1(S) \le x) = P(S \ge K_1^{-1}(1-x)) \to 1 - K_1(K_1^{-1}(1-x)) = x.$$

而显然 P(p 值 $\le$ x)=0，x<0；P(p 值 $\le$ x) = 1，x>1。这证明了 p 值的零极限分布为均匀分布。

中国科学技术大学管理学院统计与金融系

2021～2022 学年第二学期考试试卷

☑A 卷　　□B 卷

课程名称： 属性数据分析　**课程代码：** STAT4001.01

开课院系： 统计与金融系　**考试形式：** 半开卷

姓　名:_____　**学　号:** _____

（装订线内不要答题）

题 号	1	2(1)	2(2)	2(3)	2(4)	总 分
得 分						

1. 选择/填空题（每空 2 分）

1) 记μ为响应变量的均值，g为联系函数，X_1和X_2为预测变量，考虑如下 3 个广义线性模型：$g(\mu)=\alpha+\beta_1 x_1$；$g(\mu)=\alpha+\beta_1 x_1+\beta_2 x_2$；$g(\mu)=\alpha+\beta_1 x_1+\beta_2 x_2+\beta_{12} x_1 x_2$。则可以采用离差比较的模型共有_____对。

2) 设独立样本$Y_1,\cdots,Y_n$分布为$Y_i \sim B(n_i,\pi)$，$i=1,\ldots,n$，则$\log\frac{\pi}{1-\pi}$的极大似然估计是_____。

3) 对数线性模型(XYZ, YW, XW)对应的逻辑模型(响应变量: Y)为_____。

4) 设$Y_1,\ldots,Y_n$为独立同分布样本，其公共数学的期望和方差分别为μ和 $\varphi\mu$，则$W=\sum_{i=1}^{n}\frac{(Y_i-\bar{Y})^2}{\varphi\bar{Y}}$的极限分布是_____，据此给出φ的一个估计量：_____。

2. 解答题（5+10+15+20 分）

1) 我国已批准一些新冠病毒抗原检测试剂盒，其敏感度为 75%～98%，特意度为 95%～99%。由于我国疫情防控效果好，大部分地区的新冠病毒感染率低于10^{-6}。请问这些试剂盒是否适用于我国大部分地区？简述理由。根据公开数据测算，最近朝鲜的新冠病毒感染率保守估计超过 8%，请问这些试剂盒是否适用于朝鲜？简述理由。

2) 设二元属性变量(X,Y)的n个 iid 样本计数构成一个$I \times J$列联表，其(i,j)元素记为n_{ij}，相应概率为π_{ij}。记π_{ij}在X与Y独立的假定之下的极大似然估计为$\hat{p}_{ij}$，而无独立性假定的极大似然估计记为$\hat{\pi}_{ij}$。

a. 给出$\hat{p}_{ij}$和$\hat{\pi}_{ij}$的方差表达式；

b. 证明$\lim_{n\to\infty} \mathrm{var}(\sqrt{n}\hat{p}_{ij}) \le \lim_{n\to\infty} \mathrm{var}(\sqrt{n}\hat{\pi}_{ij})$；

c. 什么时候上式等号成立？

3) 设样本Y_1和Y_2独立且分别服从参数为π_1和π_2的泊松分布。

a. 基于$Y_1|Y_1+Y_2$的条件分布推导假设检验问题$H_0: \pi_1 = \pi_2 \leftrightarrow H_1: \pi_1 \neq \pi_2$的 Wald 检验统计量；

b. 推导π_1/π_2的近似置信系数为$1-\alpha$的 Wald 置信区间（提示：应用δ法）。

4) 考虑二元总体(X,Y)（X, $Y=1$ 或 2）抽得的 n 个简单样本(X_i,Y_i), $i=1,\ldots,n$，其相应的配对列联表数据记为$n_{jk}=\#(X_i=j,Y_i=k)$，记$\pi_{jk}=P(X=j,Y=k)$，$n^*=n_{12}+n_{21}$，$r=\pi_{12}+\pi_{21}$，$q=\pi_{12}/(\pi_{12}+\pi_{21})$。

a. 写出基于 q 的充分统计量(n_{12},n_{21})的似然函数$L(q,r)$；

b. 基于$L(q,r)$，求检验问题$H_0: q=1/2 \leftrightarrow H_0: q \neq 1/2$的得分检验统计量（注：$r$为冗余参数）；

c. 给出上述得分检验统计量的 P 值表达式，求该 P 值的零极限分布。

2021-2022《属性数据分析》期末考试参考答案

May 30, 2022

1. 填空题

(10 pt)

$$(1)3;\ (2)\log\frac{\sum_{i=1}^{n}Y_i}{\sum_{i=1}^{n}n_i-\sum_{i=1}^{n}Y_i};\ (3)(X*Z,W);\ (4)\mathcal{X}^2_{n-1},\frac{\sum_{i=1}^{n}(Y_i-\overline{Y})^2}{(n-1)\overline{Y}}.$$

注：该第（4）题假定计数数据 Y_i 发散到无穷而不是 n 发散到无穷；如认为 n 发散到无穷，结论合理也可得分。

2. 解答题

(1) (5 pt) 以平均敏感度 85% 及特异度 97% 为例，在我国大部分地区，抗原检测呈阳性者确定感染新冠病毒的比例不高于

$$\frac{0.85\times10^{-6}}{0.85\times10^{-6}+0.03\times(1-10^{-6})}\approx2.8\times10^{-5}(\text{很小}),$$

面对朝鲜地区这一比例至少为

$$\frac{0.85\times0.08}{0.85\times0.08+0.03\times0.92}\approx0.71(\text{很大}),$$

故该试剂盒适用于朝鲜但不适用于中国。

(2) 该题可假定独立性条件成立，如不假定该条件，结果合理也可得分。

a. (6 pt) 首先，

$$\hat{p}_{ij}=\frac{n_{i+}}{n}\frac{n+j}{n},\quad \hat{\pi}_{ij}=\frac{n_{ij}}{n},\quad \pi_{ij}=\pi_{i+}\pi_{+j}\ (\text{独立性假定}).$$

注意到

$$n_{i+}\sim\mathrm{B}(n,\pi_{i+}),\quad n_{+j}\sim\mathrm{B}(n,\pi_{+j}),\quad n_{ij}\sim\mathrm{B}(n,\pi_{ij}),$$

因此

$$var(n_{i+}) = n\pi_{i+}(1-\pi_{i+}), var(n_{+j}) = n\pi_{+j}(1-\pi_{+j}),$$

$$\mathbb{E}n_{i+}^2 = n\pi_{i+}(1-\pi_{i+}) + (n\pi_{i+})^2, \mathbb{E}n_{+j}^2 = n\pi_{+j}(1-\pi_{+j}) + (n\pi_{+j})^2,$$

$$var(\hat{p}_{ij}) = \frac{1}{n}\pi_{i+}\pi_{+j}[\pi_{i+}(1-\pi_{+j}) + \pi_{+j}(1-\pi_{i+})] + \frac{1}{n^2}\pi_{i+}(1-\pi_{i+})\pi_{+j}(1-\pi_{+j}).$$

另外,

$$var(\hat{\pi}_{ij}) = \frac{1}{n}\pi_{ij}(1-\pi_{ij}) = \frac{1}{n}\pi_{i+}\pi_{+j}(1-\pi_{i+}\pi_{+j}).$$

b. (3 pt) 由上两式知

$$\therefore \lim_{n\to\infty} var(\sqrt{n}\hat{\pi}_{ij}) - \lim_{n\to\infty} var(\sqrt{n}\hat{p}_{ij}) = \pi_{i+}\pi_{+j}(1-\pi_{i+})(1-\pi_{+j}) \geqslant 0.$$

c. (1 pt) 等号成立当且仅当 $\pi_{i+} = 0$ 或 1 或 $\pi_{+j} = 0$ 或 1。

(3) 以下假定 $Y_1 + Y_2$ 是给定的（非随机的)。

a. (5 pt) 首先,

$$Y_1|Y_1 + Y_2 \sim B(Y_1 + Y_2, q),$$

其中 $q = \frac{\pi_1}{\pi_1+\pi_2}$, 则 $H_0 : \pi_1 = \pi_2$ 等价于 $H_0 : q = \frac{1}{2}$。Wald 检验统计量为

$$W = \frac{(Y_1 - \frac{Y_1+Y_2}{2})^2}{\frac{1}{4}/(Y_1+Y_2)} = \frac{(Y_1-Y_2)^2}{Y_1+Y_2}.$$

b. (10 pt) 记 $\theta = \log\frac{\pi_1}{\pi_2} = \log\frac{q}{1-q}$, 则 θ 的 MLE 为 $\hat{\theta} = \log\frac{\hat{q}}{1-\hat{q}}$, 其中 $\hat{q} = \frac{Y_1}{Y_1+Y_2}$ 为 q 的 MLE, 故由 δ 法知（$\hat{q}$ 的方差为 $q(1-q)/(Y_1+Y_2)$)

$$\begin{aligned}\hat{\theta} - \theta \sim & AN\left(0, \frac{q(1-q)}{(Y_1+Y_2)}\frac{1}{[q(1-q)]^2}\right)\\ = & AN\left(0, \frac{1}{(Y_1+Y_2)q(1-q)}\right),\end{aligned}$$

$\therefore \hat{\theta} - \theta$ 的渐近方差近似为

$$\begin{aligned}\frac{1}{(Y_1+Y_2)\hat{q}(1-\hat{q})} &= \frac{1}{(Y_1+Y_2)Y_1Y_2/(Y_1+Y_2)^2}\\ &= \frac{1}{Y_1} + \frac{1}{Y_2},\end{aligned}$$

$\therefore \theta$ 的 $(1-\alpha)$Wald CI 为

$$\hat{\theta} \pm u_{\alpha/2}\sqrt{\frac{1}{Y_1}+\frac{1}{Y_2}} = \log\frac{Y_1}{Y_2} \pm \sqrt{\frac{1}{Y_1}+\frac{1}{Y_2}},$$

$\therefore \frac{\pi_1}{\pi_2} = e^\theta$ 的 $(1-\alpha)$Wald CI 为

$$\frac{Y_1}{Y_2} e^{\pm u_{\alpha/2}\sqrt{\frac{1}{Y_1}+\frac{1}{Y_2}}}.$$

(4) a. (5 pt) $\because$ 完全数据似然函数为

$$\begin{aligned}\because L(\pi_{11},\pi_{12},\pi_{21},\pi_{22}) =& \pi_{11}^{n_{11}}\pi_{12}^{n_{12}}\pi_{21}^{n_{21}}\pi_{22}^{n_{22}}\\ =& \pi_{11}^{n_{11}}\pi_{22}^{n_{22}}(1-q)^{n_{12}}[r(1-q)]^{n_{21}}\\ =& \pi_{11}^{n_{11}}\pi_{22}^{n_{22}}r^{n_{12}+n_{21}}q^{n_{12}}(1-q)^{n_{21}},\end{aligned}$$

$\therefore$ 由因子分解定理知 (n_{12}, n_{21}) 是关于 q 的充分统计量。
又

$$\because (n_{12}, n_{21}, n_{11}+n_{22}) \sim M(n, (\pi_{12}, \pi_{21}, \pi_{11}+\pi_{22})),$$

$\therefore (n_{12}, n_{21})$ 的似然函数是

$$L(q,r)(1-r)^{n_{11}+n_{22}}r^{n_{12}+n_{21}}q^{n_{12}}(1-q)^{n_{21}}.$$

b. (10 pt) 对数似然函数的一、二阶导数为

$$\begin{aligned}\dot{\ell}_1(q,r) =& \frac{\partial}{\partial q}\log L(q,r) = \frac{n_{12}}{q} - \frac{n_{21}}{(1-q)},\\ \dot{\ell}_2(q,r) =& \frac{\partial}{\partial r}\log L(q,r) = \frac{n^*}{r} - \frac{n-n^*}{1-r},\\ \ddot{\ell}_{11}(q,r) =& -\left[\frac{n_{12}}{q^2} + \frac{n_{21}}{(1-q)^2}\right],\\ \ddot{\ell}_{12}(q,r) =& \ddot{\ell}_{21}(q,r) = 0,\\ \ddot{\ell}_{22}(q,r) =& -\left[\frac{n^*}{r^2} + \frac{n-n^*}{(1-r)^2}\right],\end{aligned}$$

$\therefore H_0 : q = \frac{1}{2} \leftrightarrow H_a : q \neq \frac{1}{2}$ 的得分检验统计量为

$$\begin{aligned}S =& -\frac{[\dot{\ell}_1(\frac{1}{2},\hat{r})]^2}{\ddot{\ell}_{11}(\frac{1}{2},\hat{r}) - \ddot{\ell}_{12}(\frac{1}{2},\hat{r})\ddot{\ell}_{21}(\frac{1}{2},\hat{r})/\ddot{\ell}_{22}(\frac{1}{2},\hat{r})}\\ =& \frac{[2(n_{12}-n_{21})]^2}{4(n_{12}+n_{21})}\\ =& \frac{(n_{12}-n_{21})^2}{n_{12}+n_{21}}.\end{aligned}$$

c. (5 pt) 因为 S 的零极限分布为 $\mathcal{X}_1^2$, 故检验的 p-值为 $1-F(S)$, 其中 F 为 $\mathcal{X}_1^2$ 的分布函数，记 p-值 $1-F(S)$ 的分布函数为 $F^*(x)$。则对 $x \in [0,1]$,

$$\begin{aligned}F^*(x) =& P(1-F(S) \leqslant x) = P(S \geqslant F^{-1}(1-x))\\ \to& 1 - F(F^{-1}(1-x)) = x;\end{aligned}$$

而当 $x < 0$ 时，$F^*(x) = 0$；当 $x > 1$ 时，$F^*(x) = 1$。故 p-值的零极限分布为均匀分布 $U(0, 1)$。

中国科学技术大学管理学院统计与金融系

2022～2023 学年第二学期考试试卷

☑A 卷　　□B 卷

课程名称： 属性数据分析 **课程代码：** STAT4001.01 **开课院系：** 统计与金融系

考试形式： 半开卷 **姓　名:**_____ **学　号:** _____

（装订线内不要答题）

1. 选择/填空题（共 25 分）

1) (2 分)设(Y_1, Y_2, Y_3)服从参数为N和(π_1, π_2, π_3)的三项分布，则$Y_1 + Y_2$和$Y_3 + Y_2$之间的协方差等于_____。

2) (2 分)假设 X 是一个属性变量，什么情况下似然函数依赖于 X 的编码方式？（　　）

(A) 任何情况下都**是** (B) 任何情况下都**不是** (C) 对名义属性变量 (D)对次序属性变量

3) (2 分)在病例-对照研究下，假设二值响应变量与自变量之间的关系满足逻辑模型，则哪些参数可以有相合估计？（　　）（可多选）

(A) 所有回归参数　　(B) 没有任何参数　　(C) 只有斜率参数　　(D) 只有截距参数

4) (2 分)对风险比 RR 与优势比 OR，下列说法正确的是（　　）（可多选）

(A) |log OR| ≥ |log RR|　　(B) |log OR| ≤ |log RR|

(C) (log OR)(log RR) ≤ 0　　(D) (log OR)(log RR) ≥ 0

5) (2 分)在实际应用中，OR 比 RR 应用更广泛，其原因是_____

_____。

6) (1+2 分)对列联表的独立性检验问题，记第(i, j)格的标准化残差为e_{ij}，给定边缘计数条件下e_{ij}的极限分布为_____，无条件分布的零极限方差（　　）（可多选）

(A) 可能大于 1　　(B) 等于 1　(C) 可能小于 1

7) (1+2 分)有证据表明，吸烟与肺癌的优势比随年龄增长有增加的趋势，则是否可以用逻辑模型$\mathrm{P(Y = 1|X = x, Z = k) = expit}(\alpha_k + \beta x)$_____（填“是”或“否”）,其中Y表示肺癌发病情况（1=发病，0=正常），X表示吸烟历史（1=有吸烟史，0=无吸烟史），Z表示年龄段，理由是_____。

8) (2 分)负二项分布的概率函数为$\mathrm{f(k)} = \binom{\mathrm{k + r - 1}}{\mathrm{r - 1}} \mathrm{p^r}(1 - \mathrm{p})^{\mathrm{k}}, \mathrm{k} = 0,1,2\ldots$（p为感兴趣参数，r为冗余参数），则该分布作为指数族分布的自然参数为_____。

9) (1+2 分) 已知在无偏抽样情形下，逻辑回归与 probit 回归结果是否很接近？_____，这一

结论对有偏抽样是否仍然成立？_____（填“是”或“否”）。

10) (2 分)对$2 \times 2 \times K$列联表数据，假设模型$P(Y = 1|X = x, Z = k) = expit(\alpha_k + \beta_k x)$成立，则 Cochran-Mantel-Haenszel 检验在所有检验中是否近似最高最主要取决于什么？（只列出一条）_____。

11) (2 分)在实际问题中，何时可以采用条件似然进行统计推断？（　　）（可多选）

(A) 为简便起见　(B) 不过分损失感兴趣问题的信息　(C) 为了显得有技术含量

(D) 能严格控制 I 型错误率在名义水平

2. 解答题（共 75 分）

1) (7 分)2022 年 12 月“新十条”实施前各地封控期间的新冠病毒普筛中发现很大比例的“无症状”感染者，但在解除封控后发现大多数感染者至少都有轻症（如发烧）。试利用你在本课程所学知识合理解释这一矛盾。

2) (8 分)有证据表明，在每个年龄段，甲地的死亡率都高于乙地，但整体而言，乙地的死亡率更高。试利用你在本课程所学知识给出这一看似矛盾结果的一个合理解释。

3) (15 分)设$Y_1, \ldots, Y_n$是来自期望为μ的柏松分布的 i.i.d.样本，试先构造$\log(\mu)$的$(1-\alpha)$ Wald 置信区间，进而得到μ的$(1-\alpha)$置信区间。

4) (20 分)记随机向量(X,Y)的 i.i.d.样本为$(X_i, Y_i), i = 1, \ldots, n$，其中$X$和$Y$都是二值变量（均只能取值 0 或 1）。记$n_{jk} = \#(i: X_i = j, Y_i = k), j, k = 0,1$，并记$P(Y = 1|X = x) = expit(\alpha + \beta x)$。考虑假设检验问题$H_0: \beta = 0 \leftrightarrow H_1: \beta \neq 0$。

(i) 求H_0之下α的 MLE；

(ii) 试证明得分检验统计量（采用样本 Fisher 信息阵）恰好为拟合优度统计量$X^2 = \sum_{j,k=0}^{1} \frac{(n_{jk} - e_{jk})^2}{e_{jk}}$，其中$e_{jk} = (n_{j1} + n_{j0})(n_{1k} + n_{0k})/n$。

5) (25 分)从二元总体(X, Y)（X, Y = 1 或 2）抽得一个简单样本(X_i, Y_i), $i = 1, \ldots, n$， 记$n_{jk} = \#(i: X_i = j, Y_i = k)$，$\pi_{jk} = P(X = j, Y = k)$，$r = \pi_{12} + \pi_{21}$，$q = \pi_{12}/(\pi_{12} + \pi_{21})$.

(i) 写出似然函数$L(\pi_{11}, r, q)$；

(ii) 基于似然函数$L(\pi_{11}, r, q)$求假设检验问题$H_0: q = 1/2 \leftrightarrow H_0: q \neq 1/2$的 Wald 检验统计量$W$（注：这是含有 **2 个**冗余参数的情形，采用原假设之下的样本信息阵）；

(iii) 求W的零极限分布，据此给出相应 P 值的表达式，并求该 P 值的零极限分布。

《属性数据分析》期末 2023(A) 考试参考答案

一、 填空/选择题

1) $-N\pi_1\pi_3$.

2) B.

3) C.

4) A,D.

5) OR 在病例-对照研究中可被相合估计，但 RR 一般不能.

6) 标准正态分布; B;

7) 否; 肺癌与吸烟的优势比与年龄有关，导致相关不齐次，因此 x 的系数应依赖于 k.

8) $\log(1-p)$.

9) 是; 是.

10) $\beta_1, \ldots, \beta_K$ 是否近似相等.

11) A, B.

解答题

1) 前提假设是感染新冠病毒至少有轻症。封控前由于感染率很低，即使新冠病毒检测的准确度高，阳性预测率仍然可以比较低，导致大量的假阳性结果（归为无症状感染者）。但封控放开后，真实的感染率剧增，阳性预测率很低，从而大部分检测阳性的人有至少有轻症。

2) 当两地在不同年龄段的人口分布存在很大差异的时候会出现这一看似矛盾的结果。为简单期间，假设只有两个年龄段，设甲地在低年龄段的人口占绝大多数比例，而乙地在高年龄段的人口占绝大多数比例，则甲地和乙地总的发病率分别主要取决于低年龄段和高年龄段的发病率，这两者的大小关系一般是前者小于后者，这样就解释了总发病率甲地反而低于甲地（即使在不同年龄段上的发病率甲地都高于乙地）。

3) 似然函数为

$$L(\mu) = \prod_{i=1}^{n} \mu^{Y_i} e^{-\mu}/Y_i! = \mu^{\sum_{i=1}^{n} Y_i} e^{-n\mu} / \prod_{i=1}^{n} Y_i!,$$

易得 μ 的 MLE 为 $\bar{Y}$。因此 $\log\mu$ 的 $(1-\alpha)$Wald 置信区间为 $\log\bar{Y} \pm \hat{sd}(\log\bar{Y})u_{\alpha/2}$，其中 $\hat{sd}(\log\bar{Y}) = 1/\sqrt{n\bar{Y}}$。即 $\log\mu$ 的 $(1-\alpha)$Wald 置信区间为

$$\log\bar{Y} \pm u_{\alpha/2}/\sqrt{n\bar{Y}},$$

进而得到 μ 的 $(1-\alpha)$Wald 置信区间为

$$\bar{Y}\exp(\pm u_{\alpha/2}/\sqrt{n\bar{Y}})$$

4) 对数似然函数为

$$l(\alpha,\beta)=n_{11}(\alpha+\beta)-n_{1+}\log(1+e^{\alpha+\beta})+n_{01}\alpha-n_{0+}\log(1+e^{\alpha}),$$

易得在 $H_0:\beta=0$ 之下 α 的 MLE 为

$$\hat{\alpha}=\log\frac{n_{+1}}{n_{+0}}.$$

而

$$\begin{aligned}
\frac{\partial l(\alpha,\beta)}{\partial\alpha}=&n_{+1}-n_{1+}\frac{e^{\alpha+\beta}}{1+e^{\alpha+\beta}}-n_{0+}\frac{e^{\alpha}}{1+e^{\alpha}},\\
\frac{\partial l(\alpha,\beta)}{\partial\beta}=&n_{11}-n_{1+}\frac{e^{\alpha+\beta}}{1+e^{\alpha+\beta}},\\
\frac{\partial^2 l(\alpha,\beta)}{\partial\alpha^2}=&-n_{1+}\frac{e^{\alpha+\beta}}{(1+e^{\alpha+\beta})^2}-n_{0+}\frac{e^{\alpha}}{(1+e^{\alpha})^2},\\
\frac{\partial^2 l(\alpha,\beta)}{\partial\alpha\partial\beta}=&\frac{\partial^2\alpha l(\alpha,\beta)}{\partial\beta\partial\alpha}=\frac{\partial^2 l(\alpha,\beta)}{\partial\beta^2}=-n_{1+}\frac{e^{\alpha+\beta}}{(1+e^{\alpha+\beta})^2}.
\end{aligned}$$

故 $H_0:\beta=0$ 之下，用 $\hat{\alpha}$ 代替 α，即得到得分统计量

$$S=\frac{(n_{11}-n_{1+}n_{+1}/n)^2}{n_{1+}n_{+1}n_{+0}n_{0+}/n^3},$$

经对比可以发现 $S=X^2$。

5) 似然函数为

$$L(\pi_{11},q,r)=\pi_{11}^{n_{11}}(1-\pi_{11}-r)^{n_{22}}r^{n_{12}+n_{21}}q^{n_{12}}(1-q)^{n_{21}}.$$

上述似然函数是感兴趣的参数 q 的函数乘以冗余参数的函数，因此 $H_0:q=1/2$ 的 Wald 统计量为

$$W=-\frac{(\hat{q}-1/2)^2}{\partial^2\log L(\pi_{11},q,r)/\partial q^2|_{q=1/2}}=\frac{(n_{12}-n_{21})^2}{n_{12}+n_{21}}.$$

因为 W 的零极限分布为 χ_1^2，所以 P 值为

$$P(\chi_1^2>W)=1-F(W),$$

其中 F 为 χ_1^2 的分布函数，从而 P 值在 H_0 之下的分布函数为

$$P_{H_0}(1-F(W)\le x)=P_{H_0}(F(W)\ge 1-x).$$

注意到 W 的零极限分布是 χ_1^2，从而 $F(W)$ 的零极限分布为均匀分布，进而 $P_{H_0}(F(W)\ge 1-x)\to x, x\in(0,1)$。这表明 P 值的零极限分布为均匀分布。

中国科学技术大学管理学院统计与金融系

2023～2024 学年第二学期考试试卷

☑A 卷　　□B 卷

课程名称： 属性数据分析 **课程代码：** STAT4001.01 **开课院系：** 统计与金融系

考试形式： 半开卷 **姓　名:**_____ **学　号:**_____

（装订线内不要答题）

1. (2 分) 设 (Y_1, Y_2, Y_3) 服从参数为 N 和 (π_1, π_2, π_3) 的三项分布，则 $Y_1 + 2Y_2$ 和 $Y_3 - Y_2$ 之间的相关系数为_____。

2. (3 分) 记 Fisher 精确检验的 P 值为 p，则 p 是否为一个统计量？_____（填“是”或“否”），当 p 小于 0.05 时拒绝原假设导致的第一类错误率（　　）(单选)

   (A) 等于 0.05　　(B) 大于 0.05　　(C) 小于 0.05　　(D) 不能确定

3. (4 分) 负二项分布关于计数测度的密度函数为 $\frac{\Gamma(y+k)}{\Gamma(k)\Gamma(y+1)}\left(\frac{k}{\mu+k}\right)^k\left(1-\frac{k}{\mu+k}\right)^y, y= 0,1,2,\ldots$，其为指数族分布（至少一个未知参数）的一个条件是_____，此时自然参数为_____。

4. (3 分) 在回溯性研究中，关于 logistic 回归和 probit 回归的说法正确的有（　　）(多选)

   (A) 斜率同时等于 0 或同时不等于 0　　(B) 斜率可以一个为 0 一个不为 0

   (C) 预测结果一般接近　　(D) 预测结果一般相差较大

5. (4 分) 在实际问题中，为避免估计高维冗余参数，可以采用条件似然进行统计推断，其关键步骤是_____。

6. (8 分) 假设在两场比赛中姚明的投篮命中率均比易建联高，则易建联的平均投篮命中率是否可能比姚明更高？如果可能，请构造一个具体的数值例子加以说明。

7. (16 分) 用 D 表示个体患某疾病这一事件，E 表示个体暴露于某风险这一事件，则称

$$AR = 1 - P(D|\bar{E})/P(D)$$

为可归因风险。现随机地从人群中找 n 个暴露于风险的个体和 n 个未暴露于风险的个体，发现其中分别有 n_1 和 n_2 个人患病，设人群中暴露于风险的比例 π 已知。

(i)　　求 AR 的一个相合估计量 $\widehat{AR}$；

(ii) 求 $\widehat{AR}$ 的渐近分布，据此构造 AR 的近似水平为 $1-\alpha$ 的置信区间。

(iii) 将 AR 写成关于相对风险 $RR = P(D|E)/P(D|\bar{E})$ 的函数，则 π 未知时能否利用该关系式估计RR？简要说明理由。

8. (15) 配对随机变量 (X,Y) 的 n 个 i.i.d. 观测结果可以汇总为 2×2 列联表（总体平均表），也可以表示成 n 个 2×2 列联表（特定个体表）。

(i) 试证明相应的 McNemar 检验统计量和 CMH 检验统计量相等：

$$\frac{(n_{12}-n_{21})^2}{n_{12}+n_{21}} = \frac{[\sum_{k=1}^{n}(n_{11k}-n_{1+k}n_{+1k}/n_{++k})]^2}{\sum_{k=1}^{n} n_{1+k}\,n_{+1k}n_{2+k}n_{+2k}/[n_{++k}^2(n_{++k}-1)]},$$

其中n_{ij}为总体列联表计数，n_{ijk}为特体个体表计数，$n_{i+k}=\sum_{j=1}^{2} n_{ijk}$，$n_{+jk}=\sum_{i=1}^{2} n_{ijk}$，$n_{++k}=\sum_{i=1}^{2}\sum_{j=1}^{2} n_{ijk}$。

(ii) 上述两个检验统计量是否适用于配对病例-对照数据？简要说明理由。

9. (20 分) 考虑如下一类特殊的相依 $k\,(>1)$ 重 Bernoulli 试验：记k个试验结果为 $Y_1,\dots,Y_k$，设给定 p 时 $Y_1,\dots,Y_k$ 相互独立且都服从参数为 p 的 Bernoulli 分布，并且 p 服从均值为 μ、标准差为 $\sigma\,(>0)$ 的 beta 分布。

(i) 试求 $Y=\sum_{l=1}^{k} Y_l$ 的均值与方差。

(ii) (i)中结果与参数为 k 和 μ 的二项分布的均值和方差比较有何异同？试直观解释造成差异的原因。

(iii) 考虑将 $\boldsymbol{Y}=(Y_1,\dots,Y_k)$ 作为响应变量，X 作为自变量，对 $(\boldsymbol{Y}$,X) 的 n 个独立观测建立多类别 logit 回归模型，则相应统计推断应如何进行适当调整？

10. (25 分) 设某基因座上有两个等位基因 A 和 B（在人群中的频率分别为 p 和 $q=1-p$），记基因型 AA、AB、BB 的频率为 p^2+fpq、$2(1-f)pq$ 和 q^2+fpq，其中 f 为近交系数。从人群中抽得一个简单随机样本，统计得到样本中基因型为 AA、AB、BB 的个体数分别为 n_0、n_1、n_2。求假设检验问题 $H_0: f=0 \leftrightarrow H_1: f\neq 0$ 的似然比统计量 LR、Wald 统计量 W和得分统计量 S。

1. $$-\frac{\pi_1\pi_3+\pi_1\pi_2-2\pi_2\pi_3-2\pi_2(1-\pi_1)}{\sqrt{[\pi_1(1-\pi_1)+4\pi_2(1-\pi_2)-4\pi_1\pi_2]\,[\pi_3(1-\pi_3)+\pi_2(1-\pi_2)+\pi_2\pi_3]}}$$

2. D

3. k 固定，$\log\left(1-\frac{k}{M+k}\right)$（或 $\log\frac{M}{M+k}$）

4. B、C

5. ① 找充余参数的充分统计量，② 求给定该充分统计量时二条件似然函数并极大化之。

6. 可能。构造一例子需确保：
   ① 单场比赛中姚明三投篮命中率高于易建联，
   ② 易建联在两场比赛而平的命中率高于姚明。

7. (i) $\because P(D) = P(D|E)P(E) + P(D|\bar{E})P(\bar{E})$

$= P(D|E)\pi + P(D|\bar{E})(1-\pi)$

$\therefore \hat{AR} = 1 - \frac{n_2/n}{n_1/n\,\pi + n_2/n\,(1-\pi)} = 1 - \frac{n_2}{n_1\pi + n_2(1-\pi)}$

(ii) 记 $\hat{\pi} = (\hat{\pi}_1, \hat{\pi}_2) = (\frac{n_1}{n}, \frac{n_2}{n})$, 则

$\hat{\pi} \to AR\left(\begin{pmatrix}\pi_1\\ \pi_2\end{pmatrix}, \frac{1}{n}\begin{pmatrix}\pi_1(1-\pi_1) & 0\\ 0 & \pi_2(1-\pi_2)\end{pmatrix}\right)$.

从而由 δ 法知 $\hat{AR} \to AN(AR, \sigma^2(\pi))$, 其中

$$\sigma^2(\pi) = \frac{\pi^2 n_1 n_2\left[n_2(1-n_1/n) + n_1(1-n_2/n)\right]}{\left[(1-\pi)n_1 + \pi n_2\right]^4}$$

所以 AR 近似地以概率 $(1-\alpha)100\%$ 落在 $[\hat{AR} \pm u_{\alpha/2}\sigma(\pi)]$.

(iii) AR 在 π 未知时不可相合估计, 但 RR 可相合估计.

8. (i) 当 $n_{1+k}=n_{2+k}=1$, $n_{+1k}=2$, 其他, 根据

4种不同情形分别到下表:

n_{11k}	n_{+1k}	$n_{11k}-\frac{1}{2}n_{+1k}$	$n_{+1k}(2-n_{+1k})$	计数
1	2	0	0	n_{11}
1	1	$\frac{1}{2}$	1	n_{12}
0	1	$-\frac{1}{2}$	1	n_{21}
0	0	0	0	n_{22}

左式 $=\left[\sum_k (n_{11k}-\frac{1}{2}n_{+1k})\right]^2 \Big/ \sum_k n_{+1k}(2-n_{+1k})/4$

$=\left(\frac{1}{2}n_{12}-\frac{1}{2}n_{21}\right)^2 \Big/ \frac{1}{4}(n_{12}+n_{21})$

$=\frac{(n_{12}-n_{21})^2}{n_{12}+n_{21}}$ = 右边.

(ii) 应用, 因为检验对象是优势比, 其在病例对照研究中可相合估计.

9. (i) ∵ $Y_i|p \sim B(1,p)$, $p \sim (\mu, \sigma^2)$

∴ $EY_i = E[E(Y_i|p)] = Ep = \mu \Rightarrow EY = k\mu$.

$$\begin{aligned} Var(Y) &= E[Var(Y|p)] + Var[E(Y|p)] \\ &= E[kp(1-p)] + Var(kp) \\ &= k\mu - k(\sigma^2+\mu^2) + k^2\sigma^2 \\ &= k\mu(1-\mu) + k(k-1)\sigma^2. \end{aligned}$$

(ii) 两者均值都是 $k\mu$.

当 $k>1$ 时，Y 的方差大于 $B(k,\mu)$ 的方差 $k\mu(1-\mu)$

差异来源于 p 的随机性.

(iii) 方差估计应进行调整，可考虑引入一个散布参数.

10. 似然函数为

$$L(p,f)=(p^2+fpq)^{n_0}[2(1-f)pq]^{n_1}(q^2+fpq)^{n_2}$$

故似然比统计量为

$$LR=2\log\frac{(n_0/n)^{n_0}(n_1/n)^{n_1}(n_2/n)^{n_2}}{\tilde{p}^{2n_0}[2\tilde{p}(1-\tilde{p})]^{n_1}(1-\tilde{p})^{2n_2}}$$，其中

$\tilde{p}=\dfrac{2n_0+n_1}{2n}$ 为 H_0 之下 p 的 MLE.

记 $p_0=p^2+fpq$，$p_1=2(1-f)pq$，$p_2=q^2+fpq$，则

$\hat{p}_0=n_0/n$，$\hat{p}_1=n_1/n$，$\hat{p}_2=n_2/n$ 分别为 p_0，p_1，p_2 的 MLE.

联立 $\hat{p}_0=p+fpq$，$\hat{p}_1=2(1-f)pq$ 解得 p 与 f 的

MLE: $\hat{p}=\dfrac{2\hat{p}_0+\hat{p}_1}{2}=\tilde{p}$，$\hat{f}=\dfrac{\hat{p}_0-\hat{p}^2}{\hat{p}\hat{q}}=\dfrac{4n_0n_2-n_1^2}{(2n_0+n_1)(2n_2+n_1)}$

计算 $l(p,f)=\log L(p,f)$ 的一、二阶导函数代入 Wald

统计量表达式：

$$W=\hat{f}^2[\ddot{l}_{ff}(\hat{p},\hat{f})-\ddot{l}_{fp}^2(\hat{p},\hat{f})\ddot{l}_{pp}^{-1}(\hat{p},\hat{f})]$$

$$S=\dot{l}_f^2(\tilde{p},0)\Big/[\ddot{l}_{ff}(\tilde{p},0)-\ddot{l}_{fp}^2(\tilde{p},0)\ddot{l}_{pp}^{-1}(\tilde{p},0)]$$