DIA 决赛报告：基于中心极限定理、CMH、广义线性模型、模拟实验等统计学方法的临床三期实验数据分析报告

目录

1 问题一：重估样本量

1.1 背景

在本研究中，我们希望评估研究药物相比于安慰剂在某疾病患者人群中的治疗有效性和安全性。

1.2 方法

在以下讨论中，我们仅考虑平衡设计，即两组样本量相等情况。

样本量估计 原假设和备择假设分别为：

$$H_0: p_1 - p_0 = 0$$
$$H_1: p_1 - p_0 > 0$$

设单侧检验一类错误概率为 $\alpha = 0.025$，二类错误概率为 $\beta = 0.05$，药物组 $\pi_1 = 0.40$，安慰剂组 $\pi_0 = 0.25$，设 n 为样本量（即药物组和安慰剂组人数之和）。

为了计算所需的样本量，基于中心极限定理，可以使用正态分布来近似估算两独立比例之差。基于以下公式：

$$n = 2\left(\frac{Z_\alpha + Z_\beta}{\delta}\right)^2 \times (p_1(1-p_1) + p_2(1-p_2))$$

其中：

- $Z_\alpha = 1.96$，对应于单侧 0.025 的 z 值。
- $Z_\beta = 1.28$，对应于 90% 的把握度。
- $\delta = 0.15$，校正后的应答率的差异。
- $p_1 = 0.40$，试验药物组的应答率。
- $p_2 = 0.25$，安慰剂组的应答率。

代入上述值，我们计算得到**第二阶段还需要的样本量**为 199.4

$$\begin{aligned} n &= 2\frac{(1.96+1.28)^2}{(0.15)^2}\left[(0.4(1-0.4)) + 0.25(1-0.25)\right] \\ &= 199.45 \end{aligned}$$

，向上取整得到**第二阶段还需要的样本量为** 200。

考虑期中分析一类错误矫正的样本量重估 现在，我们考虑期中分析带来的第一类错误膨胀的问题。由于题目中没有具体给出信息时间，我们假设信息时间为 t，我们选用较为保守的 O'Brien-Fleming 消耗函数，则在期中分析的时候，取校正后的显著性水平 $\alpha' = 1-\Phi(\frac{Z_{1-0.025}}{\sqrt{t}})$，剩余的计算步骤与上文相同，这里不再赘述。

2 问题二：按 CMH 统计分析方法计算总体样本量

在这个问题中，我们需要使用 Cochran-Mantel-Haenszel（CMH）方法来考虑各地区的差异。首先，我们计算每个地区的效应大小和差异：

每个地区的差异：

$$\begin{aligned} \text{亚洲:} \quad \delta_{\text{Asia}} &= 53\% - 35\% = 18\% \\ \text{欧洲:} \quad \delta_{\text{Europe}} &= 54\% - 27\% = 27\% \\ \text{美洲:} \quad \delta_{\text{America}} &= 50\% - 25\% = 25\% \end{aligned}$$

使用地区权重来计算总的预期差异：

$$\delta = 0.50 \times 18\% + 0.20 \times 27\% + 0.30 \times 25\% = 21.9\%$$

使用总的预期差异，我们现在可以得到单组的样本量计算公式，如下所示：

$$n = \left(\frac{Z_\alpha + Z_\beta}{\delta}\right)^2 \cdot (p_1(1-p_1) + p_2(1-p_2))$$

其中：Z_α 是对应于 I 类错误率的 z 值，即 1.96（单侧 0.025）。Z_β 是对应于 II 类错误率的 z 值，即 1.645（为了达到 95% 的把握度或 5% 的 II 类错误率）。δ 是预期的应答率差异，即 21.9%。p_1 是整体的预期研究药物组应答率。p_2 是整体的预期安慰剂组应答率。

首先，我们需要计算 p_1 和 p_2 的方差。考虑到应答率，我们有：

$$p_1 = 52.3\%, \quad p_2 = 30.4\%$$

方差为：

$$\text{var}(p_1) = p_1(1-p_1) = 0.523 \times 0.477 = 0.2494$$

$$\text{var}(p_2) = p_2(1-p_2) = 0.304 \times 0.696 = 0.2114$$

代入上述公式，我们得到：

$$n = \left(\frac{1.96 + 1.645}{0.219}\right)^2 \cdot (0.2494 + 0.2114) = 125.5$$

计算后，向上取整得到每组大约需要 126 名受试者。因此，总体需要的样本量为 252 名受试者。

3 问题三：利用广义线性模型求出相对风险的点估计与置信区间

相对风险的定义为

$$RR = \frac{p_1}{p_2}$$

其中 p_1 是药物组的应答率，p_2 是安慰剂组的应答率。注意，由于这里的 p 是应答率，因此，**在本模型中，当相对风险 RR 大于 1 时，我们认为药物组比安慰剂组应答率更高。**

在不考虑交互效应的前提下，只考虑治疗组（药物和安慰剂组），分层因素（体重和区域），以及连续变量年龄我们将建立的泊松回归（属广义线性回归，用 $\log(\cdot)$ 作为连接函数）模型为：

$$\log(p) = \beta_0 + \beta_1 \times Treat + \beta_2 \times Weight + \beta_3 \times Europe + \beta_4 \times America + \beta_5 \times age$$

我们利用 R 语言进行建模，得到的泊松回归结果如下：

Listing 1: 利用泊松回归计算相对风险

```r
library(readxl)
# myPath = '***'
# data <- read_exel("myPath")
data <- read_excel("D:/大三上学习资料/DIA数据科学大赛备赛资料/决赛/题目材料/DIA 数据竞赛 决赛模拟数据表.xlsx")
# 对各列重命名，否则太冗长
colnames(data) = c('id','week','weight','region','treat','y1','y2','y3','disFlg','reason','age','priorUse')
data$week = factor(data$week,
                   levels = c('week 4','week 8','week 12','week 16','week 20','week 24'))

data[which(data$weight=='Weight group 1'),]$weight = 'thin'
data[which(data$weight=='Weight group 2'),]$weight = 'fat'
```

```r
data$weight = factor(data$weight,levels=c('thin','fat'))
data$region = factor(data$region)
data$treat = factor(data$treat,levels=c('Placebo','Experimental Drug'))
data$y1 = as.logical(data$y1)
data$y2 = as.logical(data$y2)
data$y3 = as.logical(data$y3)
data$reason = factor(data$reason)
data$priorUse = factor(data$priorUse,levels=c('No','Yes'))

# 提取出所有地24周有效的数据
data24 = data[which(data$week=='week 24' & !is.na(data$y1)),]
# 建立对数回归模型
formula = 'y1~treat+weight+region+age'
model = glm(formula,family=poisson(link='log') ,data=data24)
summary(model)
# Call:
#   glm(formula = formula, family = poisson(link = "log"), data = data24)
#
# Coefficients:
#                          Estimate Std. Error z value Pr(>|z|)
# (Intercept)             -1.410182   0.472665  -2.983  0.00285 **
# treatExperimental Drug   0.518174   0.230265   2.250  0.02443 *
# weightfat               -0.051770   0.229661  -0.225  0.82165
# regionEurope             0.386521   0.277366   1.394  0.16346
# regionNorth America      0.008856   0.310710   0.029  0.97726
# age                      0.005847   0.008796   0.665  0.50623
# ---
#   Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
#
# (Dispersion parameter for poisson family taken to be 1)
#
# Null deviance: 119.57  on 169  degrees of freedom
# Residual deviance: 110.31  on 164  degrees of freedom
# AIC: 286.31
#
# Number of Fisher Scoring iterations: 5
```

由于我们对研究药物组和安慰剂组关于主要终点的相对风险 RR 感兴趣，则应该关注泊松回归模型中 treat 变量的系数 β_1。从回归结果可以看出，$\log(RR)$ 的点估计为 $\beta_1 = \log(\hat{RR}) = 0.518174$，标准误为 $se(\hat{RR}) = 0.230265$。

相对风险 RR 的点估计为：

$$\begin{aligned}\exp(\hat{RR}) &= \exp(0.518174) \\ &= 1.678959\end{aligned}$$

虽然 RR 不服从正态分布，但根据统计学知识，$\log(RR)$ 服从正态分布，那么，$\log(RR)$ 的 95% 置信区间为：

$$\begin{aligned}CI' &= (\log(\hat{RR}) - 1.96se(\hat{RR}), \log(\hat{RR}) + 1.96se(\hat{RR})) \\ &= (0.518174 - 1.96 \times 0.230265, 0.518174 + 1.96 \times 0.230265)\end{aligned}$$

从而 RR 的 95% 置信区间为：

$$\begin{aligned} CI &= (\exp(0.518174 - 1.96 \times 0.230265), \exp(0.518174 + 1.96 \times 0.230265)) \\ &= (1.069140, 2.636608) \end{aligned}$$

注意到 RR 的 95% 置信区间内不包含 0，因此我们可以认为，我们可以在 5% 的显著性水平下拒绝原假设，即我们可以认为，药物组的应答率确实高于安慰剂组，对于主要终点而言，药物组更加有效。

4 问题四：附件数据异常挖掘——辛普森悖论现象

根据分析，我们认为“辛普森悖论”现象确实存在。我们将数据分为体重偏轻组和体重偏重组，分别对两个亚组的数据，以治疗组（药物和安慰剂组），分层因素（此时只考虑体重），以及连续变量年龄作为解释变量，同样建立与问题三类似的泊松回归模型，R 代码以及回归结果如下：

Listing 2: 分别计算体重偏轻组和体重偏重组在药物组和安慰剂组变量上关于主要终点的相对风险

```
# 提取出所有地24周有效的数据
data24 = data[which(data$week=='week 24' & !is.na(data$y1)),]
# 按照体重提取两组数据
data24.fat = data24[which(data24$weight=='fat'),]
data24.thin = data24[which(data24$weight=='thin'),]
formula2 = 'y1~treat+region+age'
model = glm(formula2,family=poisson(link='log') ,data=data24.fat)
summary(model)
# Call:
#   glm(formula = formula2, family = poisson(link = "log"), data = data24.fat)
#
# Coefficients:
#                         Estimate Std. Error z value Pr(>|z|)
# (Intercept)            -2.34246    0.87781  -2.669  0.00762 **
# treatExperimental Drug  0.66076    0.37294   1.772  0.07643 .
# regionEurope            1.01890    0.55614   1.832  0.06694 .
# regionNorth America     0.58191    0.62694   0.928  0.35332
# age                     0.01288    0.01498   0.860  0.38979
# ---
#   Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
#
# (Dispersion parameter for poisson family taken to be 1)
#
# Null deviance: 48.241  on 67  degrees of freedom
# Residual deviance: 39.788  on 63  degrees of freedom
# AIC: 113.79
#
# Number of Fisher Scoring iterations: 5
model = glm(formula2,family=poisson(link='log') ,data=data24.thin)
summary(model)
# Call:
#   glm(formula = formula2, family = poisson(link = "log"), data = data24.thin)
```

```
#
# Coefficients:
#                         Estimate Std. Error z value Pr(>|z|)
# (Intercept)            -1.09842    0.55101  -1.993   0.0462 *
# treatExperimental Drug  0.43531    0.29291   1.486   0.1372
# regionEurope            0.11887    0.33737   0.352   0.7246
# regionNorth America    -0.18008    0.36632  -0.492   0.6230
# age                     0.00363    0.01097   0.331   0.7408
# ---
#   Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
#
# (Dispersion parameter for poisson family taken to be 1)
#
# Null deviance: 71.295  on 101  degrees of freedom
# Residual deviance: 67.988  on  97  degrees of freedom
# AIC: 177.99
#
# Number of Fisher Scoring iterations: 5
```

类似的，对于体重偏重组，RR 的 95% 置信区间为：

$$\begin{aligned} CI &= (\exp(0.66076 - 1.96 \times 0.37294), \exp(0.66076 + 1.96 \times 0.37294)) \\ &= (0.932205, 4.021771) \end{aligned}$$

对于体重偏轻组，RR 的 95% 置信区间为：

$$\begin{aligned} CI &= (\exp(0.43531 - 1.96 \times 0.29291), \exp(0.43531 + 1.96 \times 0.29291)) \\ &= (0.870407, 14.705477) \end{aligned}$$

注意到，对于以上两个亚组，RR **的 95% 置信区间都包含了 1，也即按照 5% 的显著性水平，不能拒绝原假设，也即无法得出药物组比安慰剂组的应答率更高的结论**。但是，如问题三所述，在将两个体重的亚组数据合并之后，我们却拒绝了原假设，得到了药物组比安慰剂组的应答率更高的结论，这就是一个辛普森悖论现象的例子，提示我们在合并不同亚组的数据时，需要非常谨慎。

5 问题五：基于模拟求出满足一致性条件的最小样本量

5.1 Python codes

Listing 3: 利用模拟方法计算满足一致性条件的该国最低样本量

```
import numpy as np
from math import *
# 下面引入修饰器numba库，可以加速模拟过程
# 如果您的电脑上没有安装numba库，注释掉该行，以及所有的@njit修饰器
from numba import njit

@njit
def simulate_response(n: int, p_treatment=0.55, p_placebo=0.30) -> ( int, bool):
    """n为样本量，按照1:1将被试分配到药物组和安慰剂组，利用模拟的方式，
```

```
    计算药物组和安慰剂组一共应答的人数，也即【疗效观测值】"""
    sig = False
    responses_control = np.random.binomial(n//2,p_placebo)
    responses_case = np.random.binomial(n//2,p_treatment)
    xbar1 = responses_control/(n//2)
    xbar2 = responses_case/(n//2)
    sig = (xbar2 - xbar1 - \
           1.96 * sqrt( (xbar1*(1-xbar1)/(n//2) + xbar2*(1-xbar2)/(n//2)) )) > 0 # 是否具有统
               计学显著性
    return responses_case + responses_control ,sig

@njit
def consistency_met(n, total_size=200, target_ratio=0.5) ->  bool:
    """假设该国的样本量为n，通过模拟的方式，计算该国的疗效观测值和总的疗效观测值
    并在总的样本具有统计学差异的前提下，判断该国的观测值是否是总体观测值的一半以上，返回布尔值
    """
    sig = False
    while not sig: # 只有在总体样本有显著性意义时才进行后续判断
        country_response, s = simulate_response(n)
        global_response, sig = simulate_response(total_size)

    # 检查是否满足一致性标准
    return country_response >= global_response * target_ratio

@njit
def find_sample_size(target_prob=0.80, total_size=200, target_ratio=0.5, simulations=1e5) ->
    int:
    """给定总的样本量total_size，通过模拟的方式，返回满足一致性条件的最小样本量
    """
    n = 2 # 因为要按照1: 1进行入组，所以在该国的样本量必为偶数
    while n <= total_size:
        met_criteria = \
             sum([consistency_met(n, total_size, target_ratio) \
                 for _ in  range(simulations)])
        prob = met_criteria / simulations
        if prob >= target_prob:
            return n
        n += 2
    return None

required_sample_size = find_sample_size(0.8,200,0.5,1e5)

print(f"需要的样本量为: {required_sample_size}")
```

5.2 程序简介与模拟结果

我们定义，疗效观测值指：在样本中在我们的模拟程序中，一共有三个函数，分别是 simulate-response，consistency-met，以及 find-sample-size。其中，find-sample-size 相当于一个接口。最核心的功能是 simulate-response 与 consistency-met。simulate-response 根据所输入的药物组和安慰剂组的应答率，分别返回服从二项分布的非负整数作为各组的应答人数，并且判断两组应答率是否具有统计学差异，将两组应答人数之和作为疗效观测值返回。

consistency-met 则判断在本次模拟中，当该国样本量为 n 时，且总的 200 名受试者关于终点显示出统计学差异的前提下，该国疗效观测值是否达到总观测值的一半。

我们将每一轮的模拟次数调为 10 万，得到使得该一致性的概率至少为 80% 的该国样本量为 112。