临床实验生物统计学方法 R 语言实例

朱帅

2025-07-25

Contents

1 前言

生物统计学方法 R 代码分享

pdf 版本

罕见病疫苗有效性样本量估计

我们首先导入后面会用到的包，后续会用到 Hadley Wickham et al.[1] 的 dplyr 的%>% 管道函数和 Hao Zhu[2] 的 kableExtra 输出表格。

```
library(dplyr)
library(kableExtra)
```

1.1 疫苗有效性 (Vaccine Efficacy)

当我们检验一款新的疫苗是否有效时，通常会使用疫苗有效性 π(Vaccine Efficacy) 这个指标。让 P_1, P_2 分别为安慰剂组和接种疫苗组的发病率；N_1, N_2 分别为安慰剂组和接种疫苗组的群体总数；X, Y 分别为安慰剂和疫苗组的病例人数。则 $X \overset{i.i.d}{\sim} \text{Binomial}(N_1, P_1)$, $Y \overset{i.i.d}{\sim} \text{Binomial}(N_2, P_2)$。

疫苗有效性为：

$$\pi = 1 - (P_2/P_1)$$

原假设和备择假设为：

$$H_0 : \pi \leq \pi_0 \text{ versus } \quad H_1 : \pi > \pi_0$$

如果一个疾病发病率较低，我们就需要更多受试者参加实验，在这种情况下 X, Y 可以被近似成为独立的泊松分布，也就是 $X \overset{i.i.d}{\sim} \text{Poisson}(\lambda_1)$, $Y \overset{i.i.d}{\sim} \text{Poisson}(\lambda_2)$ with $\lambda_1 = N_1 \cdot P_1, \lambda_2 = N_2 \cdot P_2$.

Y 给定 T 的条件概率分布为

$$Y|T \sim \text{Binomial}(T, \theta) = \binom{T}{k} \theta^k (1-\theta)^{T-k} \text{ with } T = X + Y, \theta = \frac{\lambda_1}{\lambda_1 = \lambda_2}$$

此时原假设和备择假设为：

$$H_0 : \theta \geq \theta_0 \text{ versus } \quad H_1 : \theta < \theta_0 \text{ where } \theta_0 = \frac{1-\pi_0}{2-\pi_0}$$

- 计算的 P value 公式为

$$p = \Pr[Y \leq Y_{\text{obs}} \mid Y \sim \text{Binomial}(T, \theta_0)] = \sum_{k=0}^{Y_{\text{obs}}} \binom{T}{k} \theta_0^k (1-\theta_0)^{T-k}$$

[1] *Dplyr: A Grammar of Data Manipulation* (2023), https://dplyr.tidyverse.org.
[2] *kableExtra: Construct Complex Table with Kable and Pipe Syntax* (2024), http://haozhu233.github.io/kableExtra/.

- 计算 Statistical power 公式为

$$1-\beta = \Pr[Y \le Y_c \mid Y \sim \text{Binomial}(T, \theta_1)] = \sum_{k=0}^{Y_c} \binom{T}{k} \theta_1^k (1-\theta_1)^{T-k}$$

- 计算临床实验所需样本量公式为

$$N_2 = T/[(2-\pi_1)/P_1]$$

计算样本量所需的变量为 π_0, π_1, α, 期望统计功效 (1-β) 和安慰剂组发病率 (P_1)。

详细推导过程可以看 Ivan S. F. Chan and Norman R. Bohidar[3] 和 Matthew M Loiacono[4]

1.2 R 代码示例

```
exact_conditonal_test <- function(T, alpha, power, theta0, theta1) {
  Y_c <- qbinom(alpha, size = T, prob = theta0)-1
  p_value <- pbinom(Y_c, size = T, prob = theta0)
  power <- pbinom(Y_c, size = T, prob = theta1)
  return(c(T, Y_c, power, p_value))
}

ect_sample_size <- function(T, pi0, pi1, incidence, alpha, power){
  theta0 <- (1-pi0) / (2-pi0)
  theta1 <- (1-pi1) / (2-pi1)
  table_t <- as.data.frame(do.call(rbind,
                                   lapply(T, exact_conditonal_test,
                                          alpha = alpha, theta0= theta0, theta1 = theta1)),
                           .name_repair = "unique")
  colnames(table_t) <- c('T', 'Y_c', 'power', 'p-value')

  for (n in 1:nrow(table_t)){
    if (table_t$power[n] >= power && all(table_t$power[n:nrow(table_t)] >= power)) {
      min_n <- table_t$T[n]
      break
    }
  }
```

[3]"Exact Power and Sample Size for Vaccine Efficacy Studies," *Communications in Statistics - Theory and Methods* 27, no. 6 (1998): 1305–22, https://doi.org/10.1080/03610929808832160.

[4]*SAMPLE SIZE ESTIMATION AND POWER CALCULATIONS FOR VACCINE EFFICACY TRIALS FOR EXCEEDINGLY RARE DISEASES*, n.d.

```r
  result <- list(
    T_table = table_t,
    T = min_n,
    text = paste0('The min value of T achieve power of ', power, ' is ', min_n)
  )
  class(result) <- 'result'
  result
}

T2N <- function(T_value, incidence, pi1, dropout_rate){
  N2 <- T_value/((2-pi1)*incidence)/(1-dropout_rate)
  cat('The sample of vaccine group considering the drop out rate:', N2)
}
```

1.2.1 例 1

Chan and Bohidar[5] 在其论文中详细阐述了 exact conditional 方法的理论推导，并提供了不同样本量下统计功效（power）与显著性水平（significance level）的对应表格。为验证该方法的计算准确性，我们可以调用 `ect_sample_size()` 函数进行实证分析。具体参数设置：病例范围 33 到 40，$\pi_0 = 0.2, \pi_1 = 0.8, P_1 = 0.006, \alpha = 0.025$, 目标统计功效为 95%。将这些参数输入函数后，即可获得相应的样本量估计结果。

[5]"Exact Power and Sample Size for Vaccine Efficacy Studies."

Table II
Sample Size Determination Using the Exact Conditional Test Based on Poisson Assumption

Total Number of Cases (T)	Critical Value (Y_c)	Exact Power (%)	Exact Level (%)
33	8	91.4	1.36
34	9	95.4	2.44
35	9	94.5	1.79
36	9	93.4	1.30
37	10	96.5	2.28
38	10	95.8	1.68
39	10	95.0	1.23
40	11	97.4	2.11

Note: The hypothesis H_0: $\pi \leq 0.2$ versus H_1: $\pi > 0.2$ is tested at the nominal 2.5% level. A true efficacy of 0.8 is assumed under the alternative.

```
T <- 33:40 # the number of T
pi0 <- 0.2 # null hypothesis efficacy
pi1 <- 0.8 # True efficacy under alternative hypothesis
alpha <- 0.025 # type I error
incidence <- 0.006 # placebo incidence rate
power <- 0.95
res1 <- ect_sample_size(T, pi0, pi1, incidence, alpha, power)
kbl(res1$T_table)%>%
  kable_styling(bootstrap_options = "striped", full_width = F, position = "left")
```

T	Y_c	power	p-value
33	8	0.9139690	0.0136117
34	9	0.9540856	0.0244451
35	9	0.9449925	0.0178969
36	9	0.9347919	0.0129998
37	10	0.9653937	0.0227940
38	10	0.9584044	0.0168288
39	10	0.9504998	0.0123313
40	11	0.9738542	0.0211901

代码中的 incidence 就是 P_1，其他参数与上面提及的保持一致。可以看到输出的表格中样本量，critical value，statistical power 和 p value 与论文中的表格完全一致。我们打印出能

够达到 95% statistical power 的病例数

```
res1$T
#> [1] 37
```

打印结果显示 $\mathbf{T = 37}$，与论文中的结果一致。接下来，利用公式 (1.1) 计算疫苗组所需的样本量。为了方便重复使用，我将该计算过程封装成了函数 `T2N()`。

在本示例中，未考虑受试者脱落情况，因此将 `dropout_rate` 参数设为 0。计算结果显示，疫苗组的样本量应不少于 **5138.889**。将该值乘以 2，即可得出疫苗组与安慰剂组的总样本量应不少于 **10277.78**，进一后为 **10278**。

该结果与论文完全一致，说明我们的算法实现是正确的。

```
T2N(37, incidence, pi1, dropout_rate = 0)
#> The sample of vaccine group considering the drop out rate: 5138.889
```

1.2.2 HRV-三期

> For the primary hypothesis, HRV was considered efficiency of > 55% (Li et al., 2014; Mo et al., 2017) against any severity of RVGE caused by G1, G2, G3, G4, G8, G9 serotype; the incidence rate of RVGE in two rotavirus seasons was 5%, the ratio of subjects in HRV and placebo group was 1:1; and at least 73 cases of acute gastroenteritis (AGE) of any severity caused by G1, G2, G3, G4, G8, G9 serotype of rotavirus are expected to be observed; HRV had a protective efficacy > 70% for severe RVGE, the cumulative incidence rate of severe RVGE in two consecutive rotavirus seasons was estimated at 1% (Chen et al., 2019; Liu et al., 2020; Zhang et al., 2020); About 20% dropout rate was considered. The sample size was calculated as 6400 subjects using the exact condition method of Chan and Bohidar under the assumption of large sample Poisson distribution (Chan and Bohider, 1998).

Zhiwei Wu et al.[6] 这篇论文原本是我们希望复现样本量计算的参考文献。然而在复现过程中发现，文中并未明确给出所需的关键参数 π_1、α 以及期望的统计功效，因此无法准确计算所需样本量。因此，我们决定后续不再以该论文作为参考。

接下来的示例将以 HRV-三期这篇论文的参考文献为依据，该研究同样是关于轮状病毒（rotavirus）疫苗有效性的临床试验。

[6] "Efficacy, Safety and Immunogenicity of Hexavalent Rotavirus Vaccine in Chinese Infants," *Virologica Sinica* 37, no. 5 (2022): 724–30, https://doi.org/10.1016/j.virs.2022.07.011.

1.2.3 RV5

Zhaojun Mo et al.[7] 这篇论文将作为我们后续进行疫苗有效性检验样本量计算的参考文献。文中提供了以下参数：15% 的脱落率，$\pi_0 = 0$、$\pi_1 = 0.6$、$P_1 = 0.02$、显著性水平 $\alpha = 0.025$，以及期望的统计功效为 80%。

代入上述参数后，可得所需的最小病例数为 47。为了方便分组，我们取双数为 48。接着，使用 `T2N()` 函数计算疫苗组所需的样本量，结果为至少 2016.807，向上取整后为 2020。该样本量结果与论文中的报告一致，验证了我们算法的正确性。

> The study design was RVGE case driven. Under the following assumptions: an 85% evaluability rate, a 60% true efficacy against severe RVGE, and a true underlying attack rate of 2% for severe RVGE, a total of 48 targeted severe RVGE cases, of which no more than 16 are in the RV5 group, would be able to demonstrate the efficacy of RV5 against severe RVGE with 1-sided $\alpha = 0.025$ and power≈80%. To accrue the targeted severe RVGE cases, a total of 4040 participants with 2020 per vaccination group were sufficient. This sample size would be able to demonstrate the efficacy of RV5 against any-severity RVGE with 1-sided $\alpha = 0.025$ and power >90%, from which 100 targeted any-severity RVGE cases would be accrued.

```r
T <- 40:50 # the number of T
pi0 <- 0 # null hypothesis efficacy
pi1 <- 0.6 # True efficacy under alternative hypothesis
alpha <- 0.025 # type I error
incidence <- 0.02 # placebo incidence rate
power <- 0.8
res_rv5 <- ect_sample_size(T, pi0, pi1, incidence, alpha, power)
kbl(res_rv5$T_table)%>%
  kable_styling(bootstrap_options = "striped", full_width = F, position = "left")
```

[7] "Efficacy and Safety of a Pentavalent Live Human-Bovine Reassortant Rotavirus Vaccine (RV5) in Healthy Chinese Infants: A Randomized, Double-Blind, Placebo-Controlled Trial," *Vaccine* 35, no. 43 (2017): 5897–904, https://doi.org/10.1016/j.vaccine.2017.08.081.

T	Y_c	power	p-value
40	13	0.7692914	0.0192387
41	13	0.7363326	0.0137666
42	14	0.8052771	0.0217793
43	14	0.7757295	0.0157697
44	15	0.8362319	0.0243834
45	15	0.8100042	0.0178489
46	15	0.7819032	0.0129480
47	16	0.8396107	0.0199930
48	16	0.8146130	0.0146525
49	17	0.8650285	0.0221921
50	17	0.8429717	0.0164196

```
res_rv5$T
#> [1] 47
```

```
T2N(48, incidence, pi1, dropout_rate = 0.15)
#> The sample of vaccine group considering the drop out rate: 2016.807
```

Reference

Chan, Ivan S. F., and Norman R. Bohidar. "Exact Power and Sample Size for Vaccine Efficacy Studies." *Communications in Statistics - Theory and Methods* 27, no. 6 (1998): 1305–22. https://doi.org/10.1080/03610929808832160.

Loiacono, Matthew M. *SAMPLE SIZE ESTIMATION AND POWER CALCULATIONS FOR VACCINE EFFICACY TRIALS FOR EXCEEDINGLY RARE DISEASES*. n.d.

Mo, Zhaojun, Yi Mo, Mingqiang Li, et al. "Efficacy and Safety of a Pentavalent Live Human-Bovine Reassortant Rotavirus Vaccine (RV5) in Healthy Chinese Infants: A Randomized, Double-Blind, Placebo-Controlled Trial." *Vaccine* 35, no. 43 (2017): 5897–904. https://doi.org/10.1016/j.vaccine.2017.08.081.

Wickham, Hadley, Romain François, Lionel Henry, Kirill Müller, and Davis Vaughan. *Dplyr: A Grammar of Data Manipulation*. 2023. https://dplyr.tidyverse.org.

Wu, Zhiwei, Qingliang Li, Yan Liu, et al. "Efficacy, Safety and Immunogenicity of Hexavalent Rotavirus Vaccine in Chinese Infants." *Virologica Sinica* 37, no. 5 (2022): 724–30. https://doi.org/10.1016/j.virs.2022.07.011.

Zhu, Hao. *kableExtra: Construct Complex Table with Kable and Pipe Syntax*. 2024. http://haozhu233.github.io/kableExtra/.