01\_question1

44

2022-06-02

# Q1 A trial was performed to assess the impact of an intervention on the number of times a sedentary individual engages in 10 minutes of sustained moderate-to-vigorous physically active over a week.

The study recruited participants in matched pairs (matching on age), assigning the intervention to one individual within each age pair (n total pairs).

Denote each pair using the subscript and let the number of active bouts for the person in the th pair be denoted as , , with corresponding to individual who received the intervention.

Let be the age for individuals in the th pair.

It is hypothesized that

where is an unspecified function of (age).

The investigators of the current study are uninterested in .

Assume that the distribution of age in the population follows some distribution .

If the number of pairs () observed is large and is observed densely relative to the shape of then may be estimable using the likelihood of the data.

However, if n is small or is sparsely observed, estimating both and may not be possible. This question involves deriving a conditional likelihood where you condition out the nuisance parameter(s) so that is estimable in such scenarios.

You will implement a short simulation study and apply your method to data from the study referenced above.

## (a) In the first part of this problem, we will condition out nuisance parameters () to obtain a likelihood for which depends only on the observed data.

### i. What is the likelihood of ? Your answer will involve the marginal distribution of , .

### ii. What is the likelihood of ? Identify the corresponding distribution.

The summation of two Poisson distribution is also a Poisson distribution.

### iii. What is the likelihood of ? Identify the corresponding distribution.

The distribution of is , with sample size and

The distribution for the parameter is

### iv. What is the maximum likelihood estimator of obtained from the conditional likelihood you derived in iii?

If is the MLE of , then is the MLE of ;

## (b) Propose a method for constructing an asymptotic 95% confidence interval for based on your result from (a) iv using maximum likelihood theory (you may use the observed information here).

The MLE is asymptotically normal under regularity conditions.

Hence, an asymptotic 95% confidence interval for can be calculated using

## (c) A statistician colleague of yours suggests constructing of a confidence interval using simulation. Specifically, your colleague proposes the following procedure:

Fix , the type-I error rate.

* Step 1 Choose candidate interval endpoints which is dense over a range of plausible endpoints for the confidence interval.
* Step 2 For each
  + Step 2a Simulate , using the distribution you derived in (a)iii.
  + Step 2b Calculate as the MLE you derived in (a)iv.
  + Step 2c Repeat (1)-(2) many times (e.g. 2000).
  + Step 2d Let , where ) and are the and percentiles of obtained from Steps 2a-2c, respectively.
* Step 3 Define a % CI for

Argue that your colleague’s algorithm is a valid approach for constructing a 95% confidence interval. Write a function which implements this algorithm for arbitrary: - Type-I error - Input vectors - Candidate interval endpoints - Number of simulated datasets created in step 2c

We are trying to find with a random whether we can find the interval covers the true MLE value. It is the same as using a stochastic process to find the empirical distribution of . According to Donsker’s theroem and Glivenko-Cantelli theorem, we know that the empirical process distribution convergence to the true distribution asymptotically. Hence, we can use sampling to generate a valid confidence interval.

library(tidyverse)  
## simulation function.  
gety <- function(seed,   
 n = 15) {  
 set.seed(seed)  
 X <- runif(n, min = 0, max = 1)  
 beta <- 2 \* exp(-2 \* X)   
 gamma <- 2.2  
   
 y1 <- map(gamma \* beta, ~rpois(1, .)) %>% unlist()  
 y2 <- map(beta, ~rpois(1, .)) %>% unlist()  
   
 simy <- cbind(y1 = y1, y2 = y2)  
 return(simy)  
}  
  
## simulate 5000 y1 y2 set.  
sim\_num <- 5000  
y\_ss15 <- map(1:sim\_num, ~gety(seed = ., n = 15))  
y\_ss100 <- map(1:sim\_num, ~gety(seed = ., n = 100))  
save(y\_ss15, y\_ss100, file = "bios\_qexam\_data\_1.Rdata")

load("bios\_qexam\_data\_1.Rdata")  
  
## function for question 1d  
get\_gamma <- function(y,  
 alpha = 0.05,  
 gamma\_min = 0,  
 gamma\_max = 5.5,  
 gamma\_step = 0.01,  
 N = 2000) {  
 ## sum of y1 y2  
 y1 <- y[, 1]  
 y2 <- y[, 2]  
 y12 <- y1 + y2  
 gamma\_int <- seq(gamma\_min, gamma\_max, by = gamma\_step)  
 gamma\_hat <- sum(y1) / sum(y2)  
 ## emptry matrix to store the values  
 gamma\_mle <- matrix(NA, nrow = N, ncol = length(gamma\_int))  
  
 for (j in seq\_along(1:N)) {  
 for (i in seq\_along(1:length(gamma\_int))) {  
 phat <- gamma\_int[i] / (1 + gamma\_int[i])  
 ## simulate y1 | y1+y2 2000 times  
 ya3 <- map(y12, ~rbinom(n = 1, size = .,   
 prob = phat)) %>% unlist()  
 gamma\_mle[j, i] <- sum(ya3) / sum(y12 - ya3)   
 }  
 }  
   
 ## get F(0.25) and F(0.975)  
 q025 <- apply(gamma\_mle, 2,   
 quantile, probs = alpha / 2,  
 na.rm = TRUE)  
  
 q975 <- apply(gamma\_mle, 2,   
 quantile, probs = 1 - alpha / 2,   
 na.rm = TRUE)  
 ## check whether gamma\_int is located inside F025 F975  
 gamma\_ind <- ifelse((gamma\_hat > q025) & (gamma\_hat < q975), 1, 0)  
 ## find the smallest and largest gamma\_int  
 lower <- min(gamma\_int[which(gamma\_ind == 1)])  
 upper <- max(gamma\_int[which(gamma\_ind == 1)])  
 return(c(lower, upper))  
}  
  
# get\_gamma(y\_ss100[[1]])

## parallel programming  
library(furrr)  
library(parallel)  
detectCores()  
  
## use 6 cores  
plan(multisession, workers = 6)  
finall\_5k\_N15 <- future\_map\_dfc(y\_ss15, ~get\_gamma(y = .))  
finall\_5k\_N100 <- future\_map\_dfc(y\_ss100, ~get\_gamma(y = .))  
  
save(finall\_5k\_N15, finall\_5k\_N100, finall\_per,  
 file = "bios\_qexam\_question1.Rdata")

load("bios\_qexam\_question1.Rdata")  
library(matrixStats)  
rowMeans(finall\_5k\_N15)  
rowMeans(finall\_5k\_N100)

## 

## (d) Implement a fully reproducible simulation study which assesses the performance of the confidence intervals obtained in (b) and (c). Specifically, implement a simulation study as follows:

* ,
* ,
* ,
* ,
* is an equally spaced grid on with interval length of
* Simulate datasets in Step 2c of the algorithm from (c)

Estimate 95% confidence intervals using the approaches you derived in both (b) and (c) in 5000 simulated datasets and evaluate power for testing the null hypothesis . Compare approaches on coverage probability and power. Summarize your results in 3-4 sentences as if for a statistical journal. Support your claims using 1-2 tables or figures. Provide well commented code which reproduces your simulation study and recreates your tables/figures.

The confidence interval calculated with asymptotic features 1b: sample size 15 MLE with mean 2.200, sd 0.421, and 95% confidence interval [1.375, 3.025]; sample size 100 MLE with mean 2.199, sd 0.163, and 95% confidence interval [1.879, 2.518]

This 95% confidence interval does not include value of 1, the power is nearly to 1 in both cases. As we can see in the graph below, the location of is far away from the asymptotic distribution of MLE (not included in the 95% CI), which means we have little change to fail reject the null hypothesis.

The confidence interval based on method in 1d, 1 is not included in the 95% confidence interval, the power should be larger than 0.95. Hence, we can reject the null hypothesis. The confidence interval for sample size 15 is [1.790, 3.115]; The confidence interval for sample size 100 is [1.952, 2.973]

Chart, line chart

Description automatically generated

## (e) Apply your methods to the dataset “MVPA intervention.csv”, a data matrix which contains two columns, “Y1” and “Y2”, corresponding to the number of 10 minute bouts of moderate-to-vigorous physical activity in matched pairs. Estimate and construct 95% confidence intervals using both of the approaches you derived in (b) and (c).

Write a brief, non-technical, summary of both methods as if communicating to a non-statistician, recommend and justify one approach, and conclude with an interpretation of the estimated coefficient and confidence interval.

The asymptotically normal distribution of MLE with mean 3.5 and sd 1.403, which provides the 95% confidence interval [0.749, 6.250]. This result is based on asymptotic theory with large sample size, which might be biased and misleading with small sample size.

The empirical method in 1d provides the confidence interval (1.71, 7.56). To confirm this result, a permutation test with 1000 resampling was applied. This provides similar wider confidence interval [1.825, 8.64].

With small sample size, I would not claim either method is perfect. However, I will recommend the second empirical methods. Because the sample size as 15 is small, that we cannot get a precise confident interval with asymptotic assumptions. Especially the confidence interval is not much larger for the second method. Hence, we can conclude that the mean value for is 3.5, and if we randomly generate confidence interval, there is at least 95% of the chance that [1.71, 7.56] will cover the true value. In less accurate terms, we are 95% confident that the population parameter is between [1.71, 7.56]. The later interpretation will mislead the reader to regard the true parameter value as random.

y <- read\_csv("data/MVPA\_intervention.csv")  
y1 <- y$Y1  
y2 <- y$Y2  
y1\_bar <- mean(y1)  
y2\_bar <- mean(y2)  
## mle   
mean <- y1\_bar / y2\_bar  
## sd   
sd <- (y1\_bar + y2\_bar)^3 \* y2\_bar / y1\_bar ^ 3

mean + 1.96 \* sd; mean - 1.96 \* sd

## CI with 1d

get\_gamma(y, gamma\_max = 10)

## permutation to get

set.seed(5555)  
y\_permu <- map(1:100, ~sample\_n(y, size = 15, replace = TRUE))  
  
## parallel programming  
library(furrr)  
library(parallel)  
  
## use 6 cores  
plan(multisession, workers = 6)  
finall\_per <- future\_map\_dfc(y\_permu, ~get\_gamma(y = ., gamma\_max = 10))  
rowMedians(as.matrix(finall\_per))  
## [1] 1.67 9.95

## [1] 1.825 8.640