<u>Background</u> For public health surveillance purposes, a study has been carried out to assess the prevalence of obesity (defined as BMI>25) among male HKU undergraduates. The study took a random sample of 185 undergraduate males in May 2023, and assessed each of their heights and weights. In total, 13 students were found to be obese.

The log-likelihood for the prevalence θ is given by $\log l(\theta) = 13log\theta + 172\log(1-\theta)$. The MLE estimates for θ is 13/185 = 0.07. By using a normal approximation, the 95% confidence interval is (0.03, 0.11).

(a) Using the likelihood ratio method, obtain a 95% confidence interval for $\boldsymbol{\theta}$.

$$2\left|\log l(\hat{\theta}) - \log l(\theta)\right| \sim \chi_1^2$$

Ans:
Using the likelihood ratio method, obtain a 95% confidence interval for θ.

$$\left|\log l\left(\frac{13}{185}\right) - \log l(\theta)\right| < 1.92$$

The 95% CI is (0.04, 0.11).

(b) Using the bootstrap method, obtain a 95% confidence interval for $\boldsymbol{\theta}$.
Ans:
The BCa bootstrap 95% CI is (0.03, 0.11).

(c) Suppose the study was also carried out in 7 other tertiary institutions. Their results are summarized below:

Institutions	#1	#2	#3	#4	#5	#6	#7
No. obese	18	21	10	11	10	17	12
No. male undergraduate	161	272	154	85	101	221	150

Estimate the overall prevalence using maximum likelihood method.
Ans:
The log likelihood for θ is

$$\sum_{i=1}^{8}[x_i log\theta + (n_i - x_i)\log(1-\theta)]$$

Estimated prevalence is 8.4%.

(d) Suppose it was hypothesized that institutions which were able to recruit more participants (e.g. n > 200) may have a different prevalence of obesity. Estimate the relative difference using the maximum likelihood method.

You may assume that the obesity prevalence is $\boldsymbol{\theta}$ for schools with fewer participants, and $\boldsymbol{k\theta}$ for schools with more participants.
Ans:

The log likelihood for k and θ is

$$\sum_{low}[x_i log\theta + (n_i - x_i)\log(1-\theta)] + \sum_{high}[x_i logk\theta + (n_i - x_i)\log(1-k\theta)],$$

or

$$\sum_{i=1}^{8}[x_i logk^{high}\theta + (n_i - x_i)\log\left(1 - k^{high}\theta\right)]$$

Estimated k is 0.87.

(e) When the sample size is large, according to maximum likelihood theory

$$\widehat{\boldsymbol{\theta}} \sim \boldsymbol{N}(\boldsymbol{\theta}, \boldsymbol{I}^{-1}(\boldsymbol{\theta})),$$

where $\boldsymbol{I}^{-1}(\boldsymbol{\theta})$ is the information matrix

$$\boldsymbol{I}(\boldsymbol{\theta}) = -\boldsymbol{E}[\frac{\partial^2 \boldsymbol{logL}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}]$$

$\frac{\partial^2 \boldsymbol{logL}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}$ **is the second derivative of the log-likelihood, also named Hessian, which can be obtained by setting "hessian=T" in the optim function in R.**

Compute the standard error for the estimated prevalence of obesity <u>in the first tertiary institution</u> and calculate its 95% confidence interval.

Ans:

The standard error for $\hat{\theta}$ is 0.019. 95% CI = (0.03, 0.11).

(f) Referring to (d), compute the 95% confidence interval for k and test the hypothesis H_0: k=1.

[Hint: use solve() to compute the inverse of a matrix]

Ans:

The standard error for $\hat{k}$ is 0.17. 95% CI = (0.54, 1.20).

Wald statistics = |(0.87-1)/0.019| < 1.96.

Do not reject H_0.

(g) Perform a likelihood ratio test for (f).

Ans:

$\log l(\hat{\theta}_0) = -384.19$, $\log l(\hat{\theta}_1) = -383.93$

Likelihood ratio statistics = 0.53 < 3.84 = $\chi_1^2(0.05)$