*(Giới thiệu bài toán A/B Test)*

A/B test là mô hình bài toán khá phổ biến. Bài toán xử lý 2 nhóm dữ liệu A và B, đánh giá hiệu năng từng nhóm và đưa ra đánh giá nhóm nào tốt hơn

Giả sử ta có dữ liệu 2 nhóm du khách từ 2 quốc gia Trung Quốc và Nhật Bản vào Việt nam. Các du khách được thực hiện xét nghiệm virus Covid-19 và có số ca dương tính của cả 2 nhóm.

Frequestist:

*Data prepare*

Giả sử ta thống kê dữ liệu thu được trong cùng khoảng thời gian kiểm tra có

* 290 người đi vào từ TQ, trong số này có 26 ca dương tính
* 170 người đi vào từ Nhật Bản, trong số này có 7 ca dương tính

|  |  |  |
| --- | --- | --- |
|  | Khách TQ | Khách Nhật |
| Số người vào | 2090 | 1170 |
| Số ca dương tính | 26 | 7 |
| Tỉ lệ nhiễm bệnh | 0.012 | 0.006 |

Câu hỏi có thực sự tỉ lệ nhiễm bênh từ khách TQ (1.2%) cao hơn khách Nhật (0.6%) không?

Chi-square test

|  |
| --- |
| dt <- as.table(rbind(c(2090, 1170), c(26, 7)))  dimnames(dt) <- list(type = c("Total", "Infected"),  party = c("China","Japan"))  (Xsq <- chisq.test(dt)) |

Bayesian A/B test:

|  |
| --- |
| library(rstan)  # The Stan model as a string.  model\_string <- "  data {  // Number of tourists  int nChina;  int nJapan;  // Number of infected  int infectedChina;  int infectedJapan;  }  parameters {  real<lower=0, upper=1> rateChina;  real<lower=0, upper=1> rateJapan;  }  model {  rateChina ~ uniform(0, 1);  rateJapan ~ uniform(0, 1);  infectedChina ~ binomial(nChina, rateChina);  infectedJapan ~ binomial(nJapan, rateJapan);  }  generated quantities {  real rate\_diff;  rate\_diff = rateJapan - rateChina;  }  "  data\_list <- list(nChina = 2090, nJapan = 1170, infectedChina = 26, infectedJapan = 7)  # Compiling and producing posterior samples from the model.  stan\_samples <- stan(model\_code = model\_string, data = data\_list) |