*(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 Hàn Quốc 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

* 190 người đi vào từ TQ, trong số này có 26 ca dương tính
* 170 người đi vào từ Hàn QUốc, trong số này có 7 ca dương tính

|  |  |  |
| --- | --- | --- |
|  | Khách TQ | Khách Hàn |
| Số người vào | 190 | 170 |
| Số ca dương tính | 26 | 18 |
| Tỉ lệ nhiễm bệnh | 0.139 | 0.106 |

Câu hỏi có thực sự tỉ lệ nhiễm bênh từ khách TQ (13.9%) cao hơn khách Hàn (10.6%) không?

$$P\_{china} = I\_{china} / N\_{china}$$

$$P\_{korea} = I\_{korea} / N\_{korea}$$

$$\texttt{Diff} = P\_{china} – P\_{korea}$$

$$\texttt{SE(Diff)} = \sqrt{\frac{P\_{china} (1-P\_{china})} {N\_{china}} + \frac{P\_{korea}(1-P\_{korea})} {N\_{korea}}}$$

Chi-square test:

|  |
| --- |
| dt <- as.table(rbind(c(190, 170), c(26, 18)))  dimnames(dt) <- list(type = c("Total", "Infected"),  party = c("China","Korea"))  (Xsq <- chisq.test(dt)) |

|  |
| --- |
| Pearson's Chi-squared test with Yates' continuity correction  data: dt  X-squared = 0.39996, df = 1, p-value = 0.5271 |

Quả thực không thể nói được gì nhiều với kết quả này đặc biệt là khi p-value khá lớn

Bayesian A/B test:

|  |
| --- |
| library(rstan)  # The Stan model as a string.  model\_string <- "  data {  // Number of tourists  int nChina;  int nKorea;  // Number of infected  int infectedChina;  int infectedKorea;  }  parameters {  real<lower=0, upper=1> rateChina;  real<lower=0, upper=1> rateKorea;  }  model {  rateChina ~ uniform(0, 1);  rateKorea ~ uniform(0, 1);  infectedChina ~ binomial(nChina, rateChina);  infectedKorea ~ binomial(nKorea, rateKorea);  }  generated quantities {  real rate\_diff;  rate\_diff = rateChina - rateKorea;  }  "  data\_list <- list(nChina = 190, nKorea = 170, infectedChina = 26, infectedKorea = 18)  # Compiling and producing posterior samples from the model.  stan\_samples <- stan(model\_code = model\_string, data = data\_list) |

stan\_samples

|  |
| --- |
| Inference for Stan model: 42248369f7698c1be5963dd22e806558.  4 chains, each with iter=2000; warmup=1000; thin=1;  post-warmup draws per chain=1000, total post-warmup draws=4000.  mean se\_mean sd 2.5% 25% 50% 75% 97.5% n\_eff Rhat  rateChina 0.14 0.00 0.02 0.10 0.12 0.14 0.16 0.19 3766 1  rateKorea 0.11 0.00 0.02 0.07 0.09 0.11 0.13 0.16 3447 1  rate\_diff 0.03 0.00 0.03 -0.04 0.01 0.03 0.05 0.10 3676 1  lp\_\_ -138.68 0.02 0.94 -141.31 -139.04 -138.38 -138.02 -137.76 1658 1  Samples were drawn using NUTS(diag\_e) at Tue Mar 10 10:55:01 2020.  For each parameter, n\_eff is a crude measure of effective sample size,  and Rhat is the potential scale reduction factor on split chains (at  convergence, Rhat=1). |

traceplot(stan\_samples)



# Export the samples to a data.frame for easier handling.

posterior <- as.data.frame(stan\_samples)

bayesplot::mcmc\_intervals(posterior, pars = c("rateChina", "rateKorea", "rate\_diff"), point\_est = "mean", prob = 0.8, prob\_outer = 0.95, color\_scheme = "blue")



posterior <- as.data.frame(stan\_samples)

sum(posterior$rate\_diff > 0) / length(posterior$rate\_diff)

|  |
| --- |
| [1] 0.81775 |



Bayesvl: