Perkenalan Diri

Nama : Rizky Alfian Rizafahlevi

Almamater : S1 Teknik Elektro ITB, Angkatan Masuk 2008

Saya adalah pengusaha. Saya memiliki e-commerce kecil. Saya memiliki ketertarikan terhadap Mathematics, Machine Learning, dan AI. Saya memiliki cita-cita menjadi pengusaha teknologi di bidang ML/AI, sekaligus memiliki kemampuan akademis yang layak untuk melakukan riset di bidang/menggunakan bidang tersebut. Cita-cita ini sangat menguat saat saya mengalami kehilangan anak. Anak ke-2 saya, wafat di usia hampir 5 tahun pada 04 Oktober 2019, karena leukimia.

Selama pengobatan anak yang berjalan hampir 2 tahun, saya melihat tanda-tanda ketidakcakapan sebagian tenaga medis dalam menangani anak saya. Dalam decision making, tenaga medis tersebut saya lihat tidak menggunakan analisis statistic yang sufficient. Lebih ke heuristic method. Itu hipotesis saya. Tetapi hipotesis saya ini saya gali kembali, dengan saya mencari data-data terkait. Saya menemukan fakta bahwa ternyata dugaan saya benar, dari hasil riset, ditemukan bahwa tenaga medis umumnya kurang cakap dalam statistics, dan itu berimplikasi besar terhadap pengobatan pasien. Dan ini masalah global. Silakan cek https://medium.com/wintoncentre/why-doctors-are-bad-at-stats-and-how-that-could-affect-your-health-e870d551bcfe.

Oleh karena itu, saya hendak mendalami dunia ini. Saya ingin memanfaatkan AI untuk membuat company terkait AI, dan saya ingin riset, salah satunya dalam bidang medis (penanganan leukimia, etc). Salah satu langkah kecil yang saya lakukan adalah saya mengikuti kursus online untuk mendapatkan sertifikasi dari MIT. Saya mengikuti kuliah online berbayar selevel Master, MicroMaster Program in Statistics and Data Science (https://micromasters.mit.edu/ds/) dari MIT sejak 05 Februari 2020. Sekarang saya sudah lulus course pertama, yaitu Probability – The Science of Uncertainty and Data. Jadwal Final Exam belum datang, tetapi nilai saya seandainya tanpa ikut Final Exam pun sudah melewati passing grade untuk lulus. Sertifikat ready pada 30 Mei 2020. Course berikutnya akan segera mulai. Saya juga berencana akan melanjutkan kuliah sampai PhD.

Mengapa Bayesian Statistics? Sejak pertama mengenal Bayesian, saya menyukai framework statistics yang ini. Saya pikir Bayesian adalah konsep yang selaras dengan cara kerja pikiran manusia. Saya pernah membaca beberapa paper Prof. Joshua Tenenbaum terkait penerapan Bayesian dalam riset cognitive manusia (silakan cek http://web.mit.edu/cocosci/josh.html). Setelah mengikuti MicroMaster program dari MIT, saya makin menyukai Bayesian Thinking. Materi yang diberikan MIT sangat rigorous, dan challenging. Melatih kemampuan theoretical saya secara deep. Saya berhasil mendapatkan nilai 100% pada semua exercise dan problem set pada course unit tentang Bayesian Inference. Oleh karena itu, ketika mendengar tentang grup Bayesian ini dari teman, saya sangat tertarik. Apalagi ada tes-nya. Saya pikir, jika saya diterima, saya akan menemukan orang-orang yang tepat untuk saya ajak diskusi, dan menambah experience serta khazanah saya.

P-Value Hacking

p-value hacking adalah suatu tindakan *cherry picking*, yaitu ketika melakukan banyak statistical test pada data, yang dilaporkan hanya yang memiliki nilai signifikan, yaitu yang memiliki nilai $p - value < \alpha$. Dimana α adalah nilai signifikansi yang dipilih untuk pengujian.

Hal ini buruk, sebab akan mengaburkan inference yang sedang dilakukan, dan akan meningkatkan kemungkinan *false-positive*. Contoh, jika anda sedang melakukan suatu pemodelan klasifikasi (misal dengan logistic regression, selanjutnya saya sebut saja regresi agar ringkas) dari sejumlah N data, ketika semua data dipakai, anda tidak menemukan variable independent yang signifikan terhadap variable response ($p - value > \alpha$), padahal anda sudah menyusun hipotesis bahwa ada variable independent yang signifikan terhadap variable response anda. Kemudian anda coba-coba mengurangi sejumlah k data yang anda perkirakan "merusak" regresi anda, sehingga data menjadi $N - k$. Ketika melakukan regresi kembali dengan data $N - k$ tersebut, *taraa!*, anda berhasil mendapatkan ada variable independent yang signifikan. Lalu hasil analisis anda ini anda laporkan pada resume penelitian anda.

Tentunya hal tersebut menyalahi kaidah sains. Bayangkan jika model regresi anda kemudian dipakai untuk memprediksi suatu data baru oleh orang lain, maka kemungkinan terjadinya *false-positive* prediction (*over-fitting*) sangat besar. Tidak hanya itu, dalam inference, pihak lain akan menganggap variable independent anda tersebut signifikan (karena memiliki nilai $p - value < \alpha$, padahal hasil dari proses $p - value$ hacking.

Coin Tossing Problem

Untuk menganalisis kasus ini, Kita akan menggunakan 2 cara, yaitu:

1. Tes Hipotesis dengan Classical Statistics
2. Analisis menggunakan Bayesian Statistics

1. Tes Hipotesis dengan Classical Statistics

   Single coin toss adalah Bernoulli experiment, yang jika dilakukan berulang-ulang, akan terdistribusi binomial ~ B (n, p). Dimana n adalah jumlah trial (lemparan koin), dan p adalah probability kemunculan head dalam single toss.

   - X_i : Independent and Identically distributed random variables (i.i.d r.v), terdistribusi Bernoulli (p);
     $0 < p < 1$, $E[X_i] = p$, $Variance\ (X_i) = p(1-p)$. X_i adalah kemunculan head pada trial ke-i. Bernilai 1 jika muncul head, 0 jika muncul tail.
   - Untuk kasus pertama, dari eksperimen, Kita akan mengestimasi nilai p, yaitu $\hat{p}$, dengan $\bar{X}_n = \frac{X_1 + \cdots + X_n}{n} = \frac{5255}{10000} = 0.5255$ dan menggunakan *Central Limit Theorem* (CLT). Dimana,

$$Var(\bar{X}_n) = \frac{1}{n^2}\, Var(X_1 + \cdots + X_n)$$

$$Var(\bar{X}_n) = \frac{n}{n^2}\, Var(X_i) = \frac{\bar{X}_n(1 - \bar{X}_n)}{n}$$

- Mengapa menggunakan CLT? Dikarenakan dengan jumlah trial (n) yang besar, binomial distribution bisa diaproksimasi dengan normal distribution.
- Formula CLT,

$$-Z_{\frac{\alpha}{2}} \leq \frac{\bar{X}_n - \hat{p}}{\sqrt{Var(\bar{X}_n)}} \leq Z_{\frac{\alpha}{2}}$$

Dengan memasukkan nilai signifikansi $\propto = 0.05$, yang artinya rentang *Confidence Interval* yang kita inginkan adalah $1 - \alpha = 0.95 = 95\%$, didapat nilai $Z_{\frac{\alpha}{2}} = 1.96$ (dari table Z Score). Maka Formula CLT di atas menjadi:

$$-1.96 \leq \frac{\bar{X}_n - \hat{p}}{\sqrt{\frac{\bar{X}_n(1-\bar{X}_n)}{n}}} \leq 1.96$$

Dengan memasukkan nilai $\bar{X}_n$ dan n, didapat:

$$-1.96 \leq \frac{0.5255 - \hat{p}}{0.005} \leq 1.96$$

$$-0.0098 \leq 0.5255 - \hat{p} \leq 0.0098$$

$$0.5157 \leq \hat{p} \leq 0.5353$$

- Estimasi nilai p, yang diaproksimasi oleh $\hat{p}$, 95% *Confidence Interval* berada dalam rentang 0.5255 ± 0.0098.
- Kesimpulan, koin yang dipakai oleh FIFA pada tahun 1925 – 1975 *probably* bias (Interval tidak memuat p coin yang unbiased, yaitu $p = 0.5$).
- Dengan melakukan langkah-langkah yang sama untuk eksperimen ke-2, antara tahun 2010 – 2020, yang menghasilkan 78 head dari 100 kali trial, didapat

$$-1.96 \leq \frac{0.78 - \hat{p}}{0.0414} \leq 1.96$$

$$-0.0811 \leq 0.78 - \hat{p} \leq 0.0811$$

$$0.6989 \leq \hat{p} \leq 0.8611$$

- Estimasi nilai p, yang diaproksimasi oleh $\hat{p}$, 95% *Confidence Interval* berada dalam rentang 0.78 ± 0.0811.
- Kemudian, karena dengan coin yang sama terdapat 2 kali pengambilan sample (coin tossing antara tahun 1925 – 1975, dan coin tossing antara tahun 2010 – 2020), maka kita perlu menggabungkan hasil keduanya untuk menemukan final result terhadap estimasi *true mean* dari $\hat{p}$, yaitu $\hat{p}_{tot}$, dan estimasi variance dan standar deviasi gabungan.
- Penggabungan menggunakan persamaan:

$$\hat{p}_{tot} = \frac{(n_1\hat{p}_1 + n_2\hat{p}_2)}{n_1 + n_2}$$

- Dimana n_1= Jumlah trial tahun 1925 – 1975 = 10000, n_2= Jumlah trial tahun 2010 – 2020 = 100, $\hat{p}_1$= mean pada trial pertama = 0.5255, $\hat{p}_2$= mean pada trial ke-2 = 0.78. Maka,

$$\hat{p}_{tot} = 0.528$$

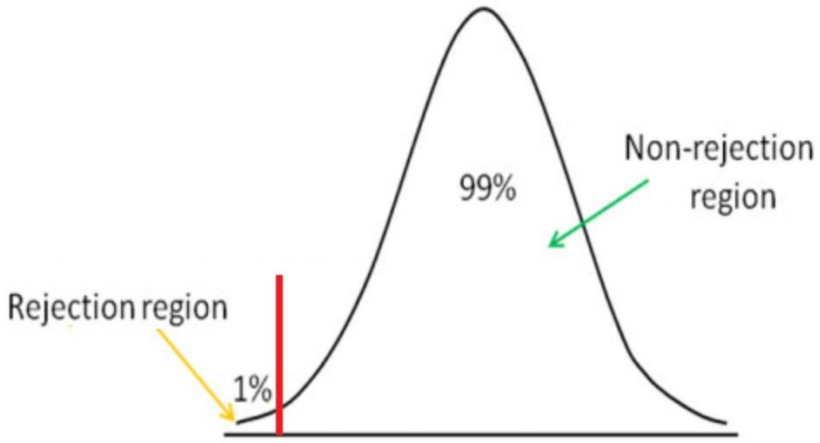
- Penghitungan variance dan standar deviasi total,

$$\hat{\sigma}^2_{tot} = \frac{n_1(\hat{\sigma}_1^2 + (\hat{p}_1 - \hat{p}_{tot})^2) + n_2(\hat{\sigma}_2^2 + (\hat{p}_2 - \hat{p}_{tot})^2)}{n_1 + n_2} = \frac{0.3125 + 6.5204}{10100} = 0.00068$$

$$\hat{\sigma}_{tot} = \sqrt{0.00068} = 0.0261$$

- Untuk mencapai kesimpulan akhir, kita gunakan tes hipotesis, dimana hipotesis yang akan kita uji adalah $H_0: \hat{p} = 0.5\ (unbiased\ coin),\ H_1: \hat{p} > H_0\ (Right\ Tail)$, dan variable yang digunakan adalah $\hat{p}_{tot} = 0.528$ dan $\hat{\sigma}_{tot} = 0.0261$.
- Untuk mempercepat perhitungan, kita gunakan Hypothesis test calculator dari link http://www.learningaboutelectronics.com/Articles/Hypothesis-testing-calculator.php#answer, didapat kesimpulan bahwa, kita reject H_0, dan kesimpulan mengarah ke H_1, yaitu coin kemungkinan biased (dengan signifikan value 0.05).

Hypothesis Testing Calculator



Select the type of Hypothesis Testing: Right-tail (μ > H0) (One tail) ▾

Enter the Null Hypothesis (H_0) Mean: 0.5

Enter the Sample Mean, $\bar{x}$: 0.528

Enter the Standard Deviation: 0.0261

Enter the Sample Size: 10100

Select the Significance Value: 5% ▾

Result:

We reject the null hypothesis and accept the alternative hypothesis. The z score of 107.81 is in the rejection area

2. Analisis menggunakan Bayesian Statistics

Untuk kasus ini, recall persamaan Bayesian:

$$f_{\Theta|X}(\theta|x) = \frac{f_{\Theta}(\theta)P_{X|\Theta}(x|\theta)}{P_X(x)}$$

Dimana,

$f_{\Theta}(\theta)$ = Prior (probability density function) kemunculan head, $\theta \in [0,1]$.

$P_{X|\Theta}(x|\theta)$ = Likelihood = Probability Event (Data), X, dalam kondisi θ tertentu.

$P_X(x)$ = Probability Event (Data), atau disebut (Probability Mass Function X), dalam semua kondisi θ.

$f_{\Theta|\mathrm{X}}\ (\theta|x)$ = Posterior (probability density function) dari θ setelah kemunculan evidence, alias data X.

a. Alur Logika:

i) Cara pertama, menggunakan perbandingan Bayes Factor.
ii) Cara Ke-2, Kita akan menggunakan evidence dari data 1925 – 1975, dari 10000 kali lemparan koin yang menghasilkan 5255 kemunculan head, untuk mengestimasi posterior distribution θ dari coin, dimana θ adalah probability kemunculan head dalam coin toss. Setelah mendapatkan posterior untuk θ dari data ini, posterior tersebut akan menjadi prior θ untuk eksperimen FIFA pada tahun 2010 – 2020, yang akan digunakan untuk mendapatkan posterior θ terbaru.

b. Analisis

i) Dengan menggunakan Bayes Factor

Untuk data 1925 – 1975:

Kita akan membandingkan likelihood X dalam 2 kondisi, $\theta = 0.5$, dan $\theta = \frac{5255}{10000} = 0.5255$

Bayes Factor (BF_{10}) $= \log\left(\frac{P_{X|\Theta}(x|\theta = 0.5255)}{P_{X|\Theta}(x|\theta = 0.5)}\right) =$
$\log\left(\frac{(0.5255^{5255})(1-0.5255)^{10000-5255}}{0.5^{10000}}\right) = -1468.4 - 1535.4 + 3010.3 = 6.5.$

Skor BF_{10} = 6.5, menunjukkan bahwa $\theta = 0.5255\ (biased)$, Extreme More Likely dibandingkan $\theta = 0.5\ (unbiased)$.

Dengan menggunakan langkah yang sama untuk data coin toss pada tahun 2010 – 2010, didapat:

$\text{BF}_{10} = \log\left(\frac{P_{X|\Theta}(x|\theta = 0.78)}{P_{X|\Theta}(x|\theta = 0.5)}\right) = 7.21.$

Skor BF_{10} = 7.21, menunjukkan bahwa $\theta = 0.78\ (biased)$, Extreme More Likely dibandingkan $\theta = 0.5\ (unbiased)$. Kalkulasi ini mensupport hipotesis bahwa coin tersebut bias dalam eksperimen pada tahun 1925 – 1975.

Kesimpulan: Dengan menggunakan analisis Bayes Factor, didapat bahwa coin FIFA tersebut Very Likely biased.

ii) Untuk cara kedua, karena kalkulasi dalam bayes sangat *computationally demanding*, maka saya menggunakan R, dan Beta Distribution. Mengapa Beta Distribution? Sebab kasus coin tossing bisa diaproksimasi dengan beta distribution. Recall persamaan beta distribution untuk prior:

$$f_{\Theta}(\theta) \sim Beta\ (\alpha, \beta) = Beta\ (1,1)$$

Saya pilih alpha, beta = 1, karena asumsi awal prior distribution uniform pada $[0,1]$

Untuk data coin tossing antara tahun 1925 – 1975, setelah dilakukan 10000 kali pelemparan dengan kemunculan head 5255 kali dan tail 4745 kali, maka pdf dari posterior menjadi:

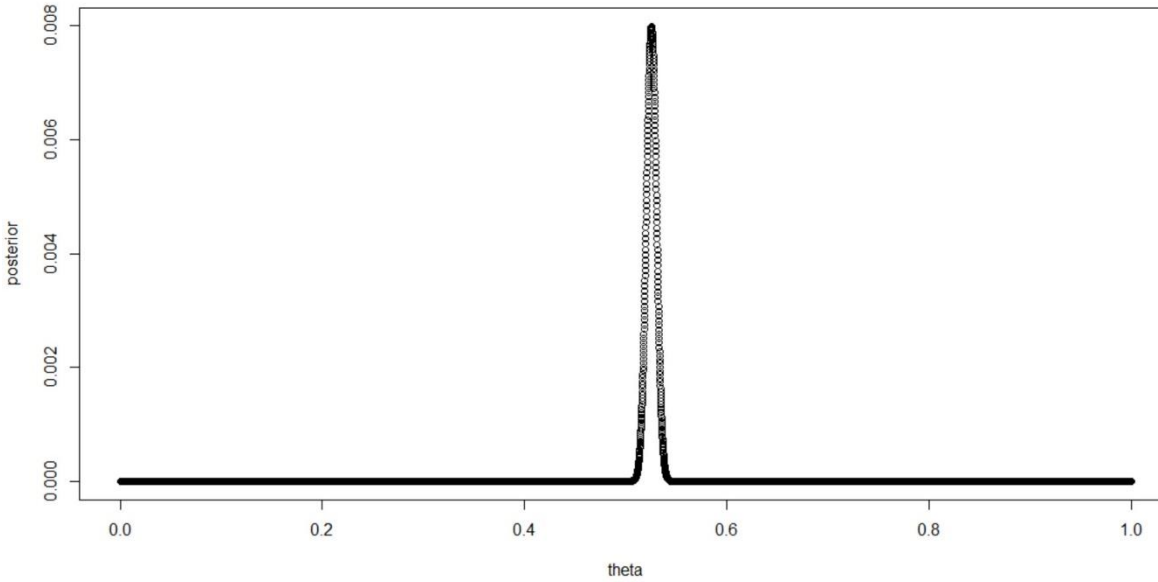
$$f_{\Theta|\text{X}}(\theta|x) \sim Beta\ (\alpha + 5255, \beta + 4745) \sim Beta\ (5256, 4746)$$

$$mean = \hat{\theta} = \frac{\alpha}{\alpha + \beta} = 0.5255$$

$$\hat{\sigma} = \sqrt{\frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}} = 0.005$$

Plot R:

```
> theta <- seq(from = 0, to = 1, by = 0.0001)
> prior <- dbeta(theta, shape1=1, shape2=1)
> likelihood <- dbinom(5255, size = 10000, prob = theta)
> numerator <- prior*likelihood
> denominator <- sum(numerator)
> posterior <- numerator/denominator
> plot(theta, posterior)
> A <- qbeta(c(0.005, 0.995), 5256, 4746)
> paste0("mean theta ", 5256/(5256+4746))
[1] "mean theta 0.525494901019796"
> paste0("proporsi posterior theta berpeluang 99% berada dalam rentang ", A[1], " sampai ", A[2])
[1] "proporsi posterior theta berpeluang 99% berada dalam rentang 0.512626030858473 sampai 0.538344621888
```



Dapat terlihat bahwa posterior distribution dari θ, $f_{\Theta|X}(\theta|x)$, normally distributed, dengan $mean = \mu = 0.5255$. Dan nilai $\hat{\theta}$ (estimasi θ) berpeluang 99% berada pada rentang $\mu \pm 0.0128 = 0.0525 \pm 0.0128$: $0.5126 \leq \hat{\theta} \leq 0.5383$.

Posterior dari percobaan pertama ini, akan menjadi prior untuk eksperimen coin trial pada tahun 2010-2020. Sehingga jika dituliskan kembali secara matematis,

$$f_{\Theta}(\theta) \sim Beta\ (5256, 4746)$$

Setelah dilakukan 10000 kali pelemparan dengan kemunculan head 78 kali dan tail 22 kali, maka pdf dari posterior menjadi:

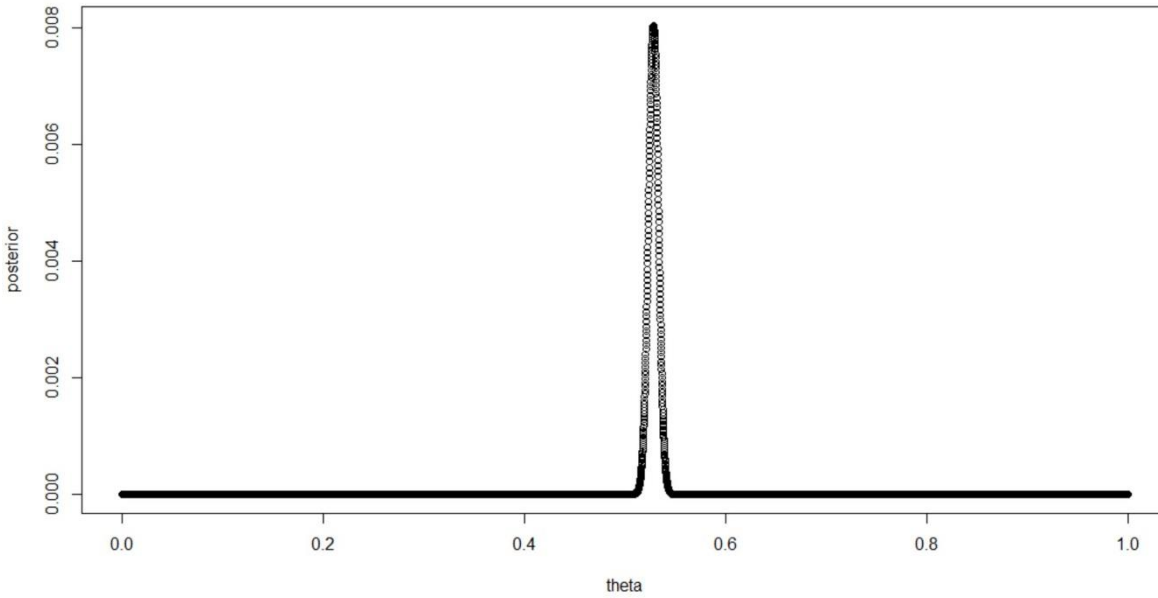
$$f_{\Theta|X}(\theta|x) \sim Beta\ (5256 + 78, 4746 + 22) \sim Beta\ (5334, 4768)$$

$$mean = \hat{\theta} = \frac{\alpha}{\alpha + \beta} = 0.528$$

$$\hat{\sigma} = \sqrt{\frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}} = 0.005$$

Plot R:

```
> theta <- seq(from = 0, to = 1, by = 0.0001)
> prior <- dbeta(theta, shape1=5256, shape2=4746)
> likelihood <- dbinom(78, size = 100, prob = theta)
> numerator <- prior*likelihood
> denominator <- sum(numerator)
> posterior <- numerator/denominator
> plot(theta, posterior)
> A <- qbeta(c(0.005, 0.995), 5334, 4768)
> paste0("mean theta ", 5334/(5334+4768))
[1] "mean theta 0.528014254603049"
> paste0("proporsi posterior theta berpeluang 99% berada dalam rentang ", A[1], " sampai ", A[2])
[1] "proporsi posterior theta berpeluang 99% berada dalam rentang 0.51521178953616 sampai 0.540795886357
```



Dapat terlihat bahwa posterior distribution dari θ, $f_{\Theta|X}(\theta|x)$, normally distributed, dengan $mean = \mu = 0.528$. Dan nilai $\hat{\theta}$ (estimasi θ) berpeluang 99% berada pada rentang $\mu \pm 0.0128 = 0.528 \pm 0.0128;\ 0.5152 \leq \hat{\theta} \leq 0.5408$.

Kesimpulan: dengan menggunakan Bayesian analisis pada bagian ii, dapat disimpulkan bahwa coin yang digunakan FIFA tersebut biased, dengan credible interval 99%.