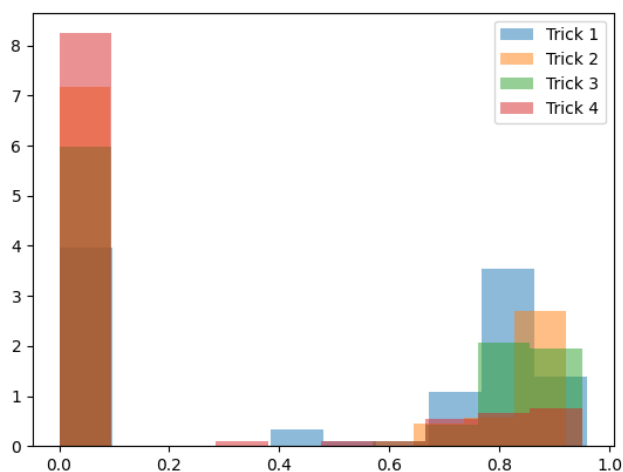


1 Uppvärmning

Figure 1: Histogram av betyg skalad mellan 0 och 1



Låt B vara betyg för en skateboardåkare och trick. Vi vill skatta $P(B > 0.6 | B > 0) = \frac{P(B > 0 | B > 0.6)P(B > 0.6)}{P(B > 0)} = \frac{P(B > 0.6)}{P(B > 0)}$ som $\tilde{P}(B > 0.6 | B > 0) = \frac{\sum_i \sum_j^{96} \text{trick}_{ij} \mathbf{1}_{\{[0.6, 1]\}}}{\sum_i \sum_j^{96} \text{trick}_{ij} \mathbf{1}_{\{[0, 1]\}}} \approx 0.96$ Det här stämmer med utseendet på ???. När man plottar run 2 mot run 1 ser de ut att ha jätte svag korrelation ??

2 En frekventistisk modell

Anm 1 Vår model för X_i är följande

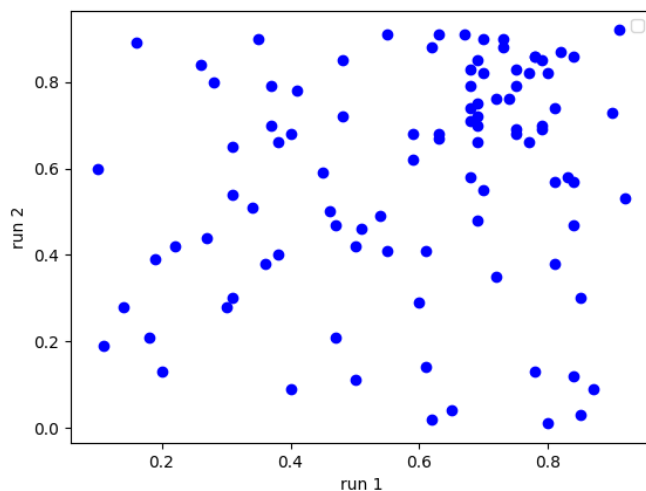
$$X_i = \begin{cases} 0 & \text{om } V_i = 0 \\ Z_i & \text{om } V_i = 1 \end{cases}$$

där $V_i \sim \text{Ber}(\theta_i)$ och $Z_i \sim \text{Beta}(\alpha_i, \beta_i)$ det här är ekvivalent med att säga

$$V_i = \mathbf{1}_{\{x \neq 0\}}(X_i) \text{ och } Z_i = X_i | (V_i = 1)$$

eftersom det här är bara en transformation av stokastiska variabler ger stickprov från X_i oss ett stickprov för Z_i och V_i

Figure 2: Spridningsdiagram mellan run 1 och run 2



(a) Skatta θ_i

Låt $x_{i[n]} = (x_{i1}, x_{i2}, \dots, x_{in})^T$ vara vår stickprov från samtliga trick skateboardåkaren i utförde.

$$L(\theta_i, \alpha_i, \beta_i | x_{i[n]}) = \prod_{j=1}^n f_{x_i}(x_{ij}) = \prod_{j=1}^n (1 - \theta_i) \mathbf{1}_{\{x=0\}}(x_{ij}) + \theta_i f_{Z_i}(x_{ij}) \mathbf{1}_{\{x \neq 0\}}(x_{ij}) \quad (1)$$

\Longleftrightarrow

$$L(\theta_i, \alpha_i, \beta_i | x_{i[n]}) = (1 - \theta_i)^{n-m} \theta_i^m \prod_{j=1}^n (f_{Z_i}(x_{ij}) \mathbf{1}_{\{x \neq 0\}}(x_{ij}) + \mathbf{1}_{\{x=0\}}(x_{ij})) \quad (2)$$

där $m = \sum_{j=1}^n \mathbf{1}_{\{x \neq 0\}}(x_{ij})$ alltså hur många gånger x_i inte är noll (gångar tävlaren i landade tricket). Nu tar vi log likelihoodfunktionen.

$$\implies \log(L) = (n - m) \log(1 - \theta_i) + m \log(\theta_i) + \sum_{j=1}^n \log(f_{Z_i}(x_{ij}) \mathbf{1}_{\{x \neq 0\}}(x_{ij}) + \mathbf{1}_{\{x=0\}}(x_{ij})) \quad (3)$$

$$\Longleftrightarrow \partial_{\theta_i} \log(L) = \frac{m - n}{1 - \theta_i} + \frac{m}{\theta_i} = 0 \quad (4)$$

$$\Longleftrightarrow \frac{m - n\theta_i}{\theta_i(1 - \theta_i)} = 0 \Longleftrightarrow \hat{\theta}_i = \frac{m}{n} \quad (5)$$

MLE för bernoulli fördelningens V_i parameter $\hat{\theta}_i = \operatorname{argmax}_{\theta \in \Omega} L(\theta_i | v_{i[n]}) = \bar{v}_i$ skulle ge oss samma resultat. Eftersom vi kan transformera stickprovet $x_{i[n]} \rightarrow v_{i[n]}$ med ?? $v_i = \mathbf{1}_{\{x \neq 0\}}(x_i)$. vilket betyder att $m = \sum_{j=1}^n v_i$ och därmed får ?? att sammanfalla med MLE av bernoulli fördelningen.

(b) skatta α_i och β_i

Observera att från ?? $\sum_{j=1}^n \log(f_{Z_i}(x_{ij}) \mathbf{1}_{\{x \neq 0\}}(x_{ij}) + \mathbf{1}_{\{x=0\}}(x_{ij})) = \sum_{j=1}^n \log(f_{Z_i}(x_{ij}) \mathbf{1}_{\{x \neq 0\}}(x_{ij}))$ eftersom $\log(1) = 0$. Vi vet att

$\operatorname{argmax}_{\alpha, \beta \in \Omega} \log(L) = \operatorname{argmax}_{\alpha, \beta \in \Omega} \sum_{j=1}^n \log(f_{Z_i}(x_{ij}) \mathbf{1}_{\{x \neq 0\}}(x_{ij}))$ vilket är ekvivalent med $\operatorname{argmax}_{\alpha, \beta \in \Omega} \log(L(\alpha, \beta | z_{i[k]}))$ för att z stickprovet innehåller alla trick som landade $z_{i[k]} = (z_{i1}, \dots, z_{ik})^T = \{x_{ij} \in x_{i[n]} : x_{ij} \neq 0\}$
Vi ska alltså bara maximera log-likelihood av beta fördelningens paramtrerna givet data from Z_i

$$\begin{aligned} & \begin{cases} \partial_{\alpha} \log(L(\alpha, \beta | z_{i[k]})) = \sum_{j=1}^k \partial_{\alpha} \log(f(z_{ij})) = 0 \\ \partial_{\beta} \log(L(\alpha, \beta | z_{i[k]})) = \sum_{j=1}^k \partial_{\beta} \log(f(z_{ij})) = 0 \end{cases} \\ & \therefore \partial_{\alpha} \log(f(z_{ij})) = \partial_{\alpha} \log\left(\frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha) \cdot \Gamma(\beta)} \cdot z_{ij}^{\alpha-1} \cdot (1 - z_{ij})^{\beta-1}\right) \\ & = \partial_{\alpha} (\log \Gamma(\alpha + \beta) - \log \Gamma(\alpha) - \log \Gamma(\beta) + (\alpha - 1) \log z_{ij} + (\beta - 1) \log(1 - z_{ij})) \\ & = \partial_{\alpha} \log(f(z_{ij})) = \psi(\alpha + \beta) - \psi(\alpha) + \log z_{ij} \text{ där } \psi = \Gamma'/\Gamma \\ & \quad (\text{Vi gör liknande för } \partial_{\beta} \log f(z_{ij})) \\ & \Rightarrow \begin{cases} \partial_{\alpha} \log L = k\psi(\alpha + \beta) - k\psi(\alpha) + \sum_{j=1}^k \log(z_{ij}) = 0 \\ \partial_{\beta} \log L = k\psi(\alpha + \beta) - k\psi(\beta) + \sum_{j=1}^k \log(1 - z_{ij}) = 0 \end{cases} \end{aligned}$$

Det går dock inte att lösa ML skattningen analytisk härifrån. Numeriska metoder som newton rhapsion eller gradient descent behövs för att skatta vår ML skattning. Att göra så medför sig en del problem som ökar systematiska felet genom numerisk fel. Vi behåller riktighet i punktskattningen genom att använda moment metoden istället. Vi utgår från följande system ekvationerna

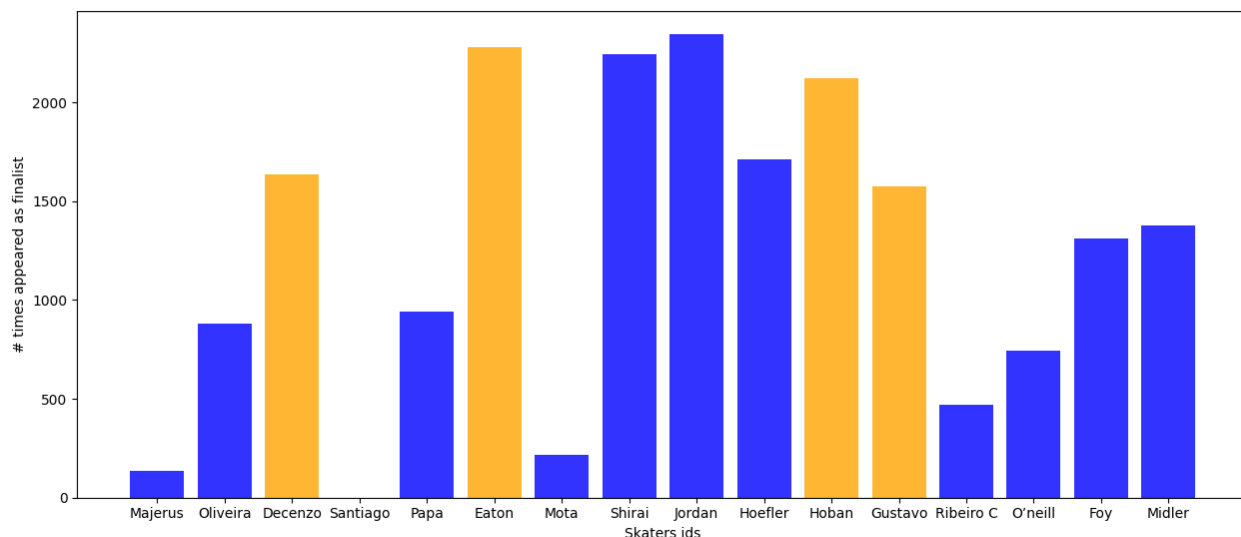
$$\begin{aligned} & \begin{cases} M_1(\mathbf{Z}_i) = E[Z_i] \\ M_2(\mathbf{Z}_i) = \operatorname{Var}[Z_i] + E[Z_i]^2 \end{cases} \iff \begin{cases} \Leftrightarrow M_1 = \frac{\alpha}{\alpha + \beta} \\ M_2 = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)} + \left(\frac{\alpha}{\alpha + \beta}\right)^2 \end{cases} \\ & \quad \therefore S^2 = \frac{1}{n} \sum (Z_k - \bar{Z})^2 = M_2 - M_1^2 \\ & \iff \begin{cases} \Leftrightarrow \alpha = \beta \frac{M_1}{1 - M_1} \\ S^2 = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)} \end{cases} \iff \begin{cases} \tilde{\alpha} = M_1 \left(\frac{M_1(1 - M_1)}{S^2} - 1 \right) \\ \tilde{\beta} = (1 - M_1) \left(\frac{M_1(1 - M_1)}{S^2} - 1 \right) \end{cases} \\ & \quad = \begin{cases} \tilde{\alpha}_i = \bar{Z}_i \left(\frac{\bar{Z}_i(1 - \bar{Z}_i)}{S^2} - 1 \right) \\ \tilde{\beta}_i = (1 - \bar{Z}_i) \left(\frac{\bar{Z}_i(1 - \bar{Z}_i)}{S^2} - 1 \right) \end{cases} \end{aligned}$$

För vissa skatebordåkare kan man inte få en punktskattningen eftersom det finns endast en datapunkt z_1 för vissa skatebordåkare. Stickprovsvariansen i sådana fall blir lika med noll. Stickprovsvariansen representerar hur osäkert vi är om en skatebordåkares prestation på en trick. Intuivt sett Vi gör valet här att skatta stickprovsvariansen som $S_i^2 \approx \bar{S}^2$ dvs vi tar medelvärde av samtliga varianser.

(c) Model för Y_i

Y_i ska vara run betyget för åkaren i . Eftersom $Y_i \in (0, 1]$ (varje deltagare får betyg större än 0) kommer bernoulli delen försvinna. Så vi antar att $Y_i \sim \operatorname{Beta}(\alpha_i, \beta_i)$. Vi använder samma metod som vi använde för att skatta α, β för X .

Figure 3: Frequency appearing in final \mathbf{W}



(d) Simulering

Total betyget för varje deltagare beräknas som summan av deras två största trick betyg och största run betyg. Vi kan beskriva i termer av stokastiska variabler. Låt O_i vara total betyg för deltagare i . Låt $Q_{i,\text{först}} = \max(X_{i1}, X_{i2}, X_{i3}, X_{i4})$ och

$Q_{i,\text{andra}} = \max(\min(X_{i1}, X_{i2}), \min(X_{i1}, X_{i3}), \min(X_{i1}, X_{i4}), \min(X_{i2}, X_{i3}), \min(X_{i2}, X_{i4}), \min(X_{i3}, X_{i4}))$.

Vi vill simulera total betyget för varje deltagare i som $O_i = Q_{i,\text{först}} + Q_{i,\text{andra}} + \max(Y_{i1}, Y_{i2})$ De som fick de 4 högsta betyg får delta i finalen. Vi simulerar 5000 LCQ:ar. Det ger oss en följd av stokastisk sets $\mathbf{W}_1, \dots, \mathbf{W}_{5000}$. Python har redan libraries för att generera stickprov från beta och bernoulli fördelningar. Så vi slipper använda box muller, inverse metoden, eller dylikt. Jag skapade en frequency bar graph för att visualisera simuleringen och fick detta i en körning ???. De som har markerats i orange är de som faktiskt vann den verkliga LCQ:en. Typvärdet på \mathbf{W} innehöll oftast Hoban, Eaton, Jordan, och Shirai. Typvärdets frekvens vara kring 50 gger dvs 1%. Frekvensen av när \mathbf{W} innehöll samtliga verkliga vinnare (nämligen Gustavo, Hoban, Eaton, Decenzo) vara när 16-20 gger dvs mindre än 0.

3 En bayesiansk modell

(a) Apriori fördelnigar

$$f_{\theta_i}(\theta_i) \propto 1$$

$$f_{\alpha_i, \beta_i}(\alpha_i, \beta_i) = \frac{\lambda^\theta}{\Gamma(\theta)} (\alpha_i + \beta_i + 1)^{\theta-1} e^{-\lambda(\alpha_i + \beta_i + 1)} (\alpha_i + \beta_i)^{-1}$$

$$\text{Vi antar } \Theta_i \perp\!\!\!\perp A_i, B_i \text{ för alla } i \implies f_{\theta_i, \alpha_i, \beta_i}(\theta_i, \alpha_i, \beta_i) = f_{\theta_i}(\theta_i) f_{\alpha_i, \beta_i}(\alpha_i, \beta_i)$$

(b) Aposteriori för X_i och skattning

$$f_{\theta_i, \alpha_i, \beta_i | x_i}(\theta_i, \alpha_i, \beta_i | x_i) \propto f_{\theta_i}(\theta_i) f_{\alpha_i, \beta_i}(\alpha_i, \beta_i) f_{x_i | \theta_i, \alpha_i, \beta_i}(x_i | \theta_i, \alpha_i, \beta_i)$$

Vårt mål är att använda aposteriorin för att skatta $E[\Theta_i, A_i, B_i | \mathbf{X}_i = \mathbf{x}_i]$. För att göra detta använder vi metropolis algoritmen för att generera ett stickprov och ta stickprovsmedelvärdet

$$\begin{pmatrix} \theta_{i0}, \theta_{i1} \dots \theta_{i5000} \\ \alpha_{i0}, \alpha_{i1} \dots \alpha_{i5000} \\ \beta_{i0}, \beta_{i1} \dots \beta_{i5000} \end{pmatrix} \longrightarrow \begin{pmatrix} \bar{\theta}_i \\ \bar{\alpha}_i \\ \bar{\beta}_i \end{pmatrix}$$

Aposteriorin är vår målfördelning men vi kommer använda $L_p(\theta_i, \alpha_i, \beta_i) \propto \log(f_{\theta_i, \alpha_i, \beta_i | \mathbf{x}_i}(\theta_i, \alpha_i, \beta_i | \mathbf{x}_i))$ istället. Data fördelning $f_{\mathbf{x}_i | \theta_i, \alpha_i, \beta_i}(\mathbf{x}_i | \theta_i, \alpha_i, \beta_i)$ i aposteriorin är samma som likelihoodfunktionen $L(\theta_i, \alpha_i, \beta_i | \mathbf{x}_i)$

(c) Aposteriori för Y_i

(d) Simulering

(e)

4 En bayesiansk modell med en hierarki

(a) Apriori fördelning för θ

(b) Aposteriori för X_i

(c) Aposteriori för Y_i

(d) Simulering

(e) Grafisk model

5 Diskussion