KTH Matematik
Matematisk Statistik

SF1930 Statistisk Inlärning och dataanalys, HT 2023
Projekt

Inledning

Detta projekt syftar till att bygga en modell för att förutsäga de tävlande i säsongsfinalen i en viss sporttävling; nämligen *2022 Street League Skateboarding (SLS) Super Crown Championship*. Evenemanget består av en *sista-chans-kvaltävling* (Last Chance Qualifier (LCQ) på engelska) och finalen. Det finns åtta platser i finalen. Fyra skateboardåkare har redan kvalificerat sig till dit baserat på sin prestation under hela säsongen. Dessa (givet av efternamn) är:

Horigome Joslin Milou Ribeiro G.

LCQ:en har sexton tävlare och de fyra skateboardåkarna med högst betyg vinner de återstående fyra platserna i finalen. Skateboardåkarna som tävlar i LCQ:en (givet av efternamn) är:

Majerus	Oliveira	Decenzo	Santiago
Papa	Eaton	Mota	Shirai
Jordan	Hoefler	Hoban	Gustavo
Ribeiro C	O'neill	Foy	Midler.

I LCQ:en får varje skateboardåkare två *runs* (vilket består av 45 sekunder för att göra så många tricks som möjligt) och fyra *trick* (eng. *single trick attempts*) där skateboardåkaren försöker att bara göra ett (svårt) trick. Ett betyg mellan 0 och 10 tilldelas varje run och varje trick. Ett trick får 0 endast om skateboardåkaren inte lyckas landa tricket. *Totalbetyget* för en skateboardåkares prestation beräknas som summan av deras två högsta betyg på trick och deras högsta betyg på run. De fyra skateboardåkarna med de högsta totalbetygen får de fyra återstående platserna i finalen. Detta projekts mål är att bygga en modell för att förutsäga dessa fyra skateboardåkarna baserat på data samlade under säsongen 2022 som bestod av tre andra evenemang.

Vart och ett av de tre evenemangen som vi har data för består av ett kval och en final. I var och en av dessa två tävlingar tilldelas skateboardåkarna en ordning i vilken de turas om med de olika aktiviteterna (runs och trick). Aktiviteterna genomfördes enligt ordning: run 1, run 2, trick 1, trick 2, trick 3, trick 4. Alla tävlande genomfördes run 1 först, sedan run 2, sedan trick 1, osv. Skateboardåkarna gör varje aktivitet i samma ordning. Till exempel, om vi har tre tävlande tilldelar vi dem en ordning S_1, S_2, S_3. Tävlingen börjar med run 1 för S_1, sedan run 1 för S_2, sedan run 1 för S_3, sedan run 2 för S_1, osv.

Kvaltävlingarna har samma format som LCQ:en förutom att skateboardåkarna delas i olika *heats* i vilka de genomför aktiviteterna enligt en ordning av tävlande i det givna

heatet. Finalen består av de 8 skateboardåkarna med de högsta totalbetygen från den tillhörande kvaltävlingen. I finalerna får först varje skateboardåkare två runs och fyra trick. Därefter beräknas deras totalbetygen. De fyra skateboardåkarna med de lägsta totalbetygen elimineras. Sedan får de fyra återstående skateboardåkarna två trick till. När alla ytterligare trick är klara räknas totalbetygen om. Skateboardåkaren med det högsta totalbetyget vinner!

Datan är i form av en CSV-fil som kallas för `SLS22.csv`. CSV-filen innehåller följande kolumner:

Namn	Data typ	Beskrivning
`id`	string	skateboardåkarens efternamn
`location`	string	platsen för tävlingen
`month`	integer	tävlings-månad
`year`	integer	tävlings-år
`comp`	string	tävlings-typ (`'prelim men'` eller `'final men'`)
`heat`	integer	Heat i vilket skateboardåkaren deltog (`heat` = 1 om `comp` = `'final men'`)
`run 1`	float	betyget för run 1
`run 2`	float	betyget för run 2
`trick 1`	float	betyget för trick 1
`trick 2`	float	betyget för trick 2
`trick 3`	float	betyget för trick 3
`trick 4`	float	betyget för trick 4
`trick 5`	float	betyget för trick 5 (`'NaN'` om elimineras efter trick 4)
`trick 6`	float	betyget för trick 6 (`'NaN'` om elimineras efter trick 4)

Du kan öppna filen med hjälp av pythonpaketet Pandas[1].

```
import pandas as pd

df = pd.read_csv("SLS22.csv")
```

Detta laddar CSV-filen in i en *dataram* (*dataframe* på engelska) i Pandas. Du kan få grundläggande information om dataramen med hjälp av `df.describe()`. Du kan extrahera individuella kolumner ifrån dataramen genom att indexera den.

Till exempel ges run 1:s betyg av kolumnen `run 1`. Denna kolumn indexeras alltså med `run 1` och extraheras med `df["run 1"]`. Den extraherade kolumnen är en *dataserie* i Pandas och kan användas som en NumPy-vektor i många situationer. Stickprovsväntevärdet för `run 1` kan, till exempel, beräknas som

```
import numpy as np

mean_score_run_1 = np.mean(df["run 1"])
```

Du kan också välja vissa rader från dataramen genom att indexera med heltal. För att beräkna stickprovsmedelvärdet för run 1 för de första hundra raderna kan du använda

[1] https://pandas.pydata.org

```
# Make a new dataframe that contains the first one hundred rows.
df_100 = df[0:100]
mean_run_1_100 = np.mean(df_100["run 1"])

# Alternatively you can index both rows and columns at the same time.
mean_run_1_100 = np.mean(df[0:100]["run 1"])
```

Ett mer realistiskt scenario är om du vill extrahera en delmängd som motsvarar ett visst kolumnvärde. Till exempel, stickprovsmedelvärdet för run 1:s betyg för alla skateboardåkarna i Las Vegas evenemanget (båda kvaltävlingen och finalen) ges av

```
vegas_mask = (df["location"] = "las vegas")
mean_run_1_vegas = np.mean(df[vegas_mask]["run 1"])
```

Här skapade vi en binär mask som har en etta för alla rader där `location` kolumnen är lika med `"las vegas"` och en nolla för de andra. Att indexera dataramen med denna mask låter oss extrahera bara dessa rader. Masker kan vara kombinerade med hjälp av ”och”-operatorn: `&`. För att beräkna stickprovsmedelvärdet för run 1:s betyg för bara finalen i Las Vegas evenemanget använder vi:

```
finals_vegas_mask = vegas_mask & (df["comp"] == "final men")
mean_run_1_finals_vegas = np.mean(df[finals_vegas_mask]["run 1"])
```

Observera att parenteserna krävs ovan eftersom `&` har företräde framför `==` i Python. Det kan vara bra att ha en metod för att lägga till kolumner i en dataram vars värden i varje rad tillräknas från befintliga data. Till exempel kan vi lägga till ytterligare en kolumn till dataramen vars värde för varje rad är summan av värdena för run 1 och run 2 i den givna raden:

```
#A data series consisting of the sum of the two run scores in each row:
run_sums = df[["run 1", "run 2"]].sum(axis = 1)

#Attach the series as a new column to the dataframe:
df["run sums"] = run_sums
```

Det är också bra om vi kan manipulera värdena i en dataram. Till exempel, om vi vill byta värdena i en viss kolumn skulle vi kunna skapa en ny kolumn och använda sammalänkningsfunktionen i Pandas `concat` för att ersätta den gamla kolumnen med den nya kolumnen:

```
#Get list of column names
cols = df.columns.tolist()

#function to apply
def f(x):
        return 2*x

#Apply function to a column:
run_sums_times_two = df["run sums"].apply(f)

#Concatenate new column with the columns one would like to keep
df = pd.concat([df[cols[0,-1]], run_sums_times_two], axis = 1)
```

Du kan också tillämpa en funktion på en viss kolumn utan att skapa en ny kolumn. Vi kan också använda `numpy` för att sortera värdena i en lista

```
Test = np.array([1,1,2,5,3,3,3,4,2])
print(np.sort(Test))
```

Vi kan extrahera de unika värdena i listan samt hur ofta dessa värden visas:

```
unique_values, counts = np.unique(Test, axis = 0, return_counts = True)
print(f"{unique_values=}")
print(f"{counts=}")
```

Uppgifter

Lös följande uppgifter och skriv ned dina lösningar i en PDF-fil i form av en rapport. Tillsammans med rapporten måste du även lämna in din kod i formen av ett Python-skript eller en Jupyter notebook med bra markdown. **Du får arbetar på projektet med en (och bara en) annan student i kursen men du måste skriva din egen kod och din egen rapport. Ni får inte lämna två kopior samma rapport! Rapporten och koden ska skrivas med dina egna ord för att visa att du förstår lösningarna. Du ska också skriva namnet på den person du arbetat med i projektet på rapportens första sida.**

1. **Uppvärmning.** Följande uppgifter bör göra dig bekant med datamängden och förbereda data för användning när du bygger dina prediktiva modeller.

   (a) Alla betyg i dataramen är för närvarande tal mellan 0 och 10. Normalisera dessa värden i dataramen så att de är mellan 0 och 1.

   (b) Gör ett histogram för alla trickbetyg för trick $1-4$. Vad observerar du? Finns det ett visst värde som dyker upp oftare än de andra? Om så är fallet, hur står detta värde i jämförelse med de andra?

   (c) For varje trick $1-4$ skapa en ny kolumn med namnet `'make i'` för $i = 1, 2, 3, 4$ så att värdet av `'make i'` i en given rad är 1 om skateboardåkaren landade trick i och 0 annars.

   (d) För varje skateboardåkare skatta sannolikheten att ett trick får ett betyg som är större än 0.6 givet att skateboardåkaren landar tricket. Vad är sannolikheten att skateboardåkaren inte lyckas landa ett visst trick? Vad observerar du? Relatera dina observationer till era observationer i del (b).

   (e) Gör ett spridningsdiagram för runbetyg 1 mot runbetyg 2. Ser du någon tydligt korrelation från diagrammet?

2. **En frekventistisk modell.** Vi skulle vilja bygga en modell som kan förutsäga vilka av de 16 skateboardåkarna i LCQ:en vinner en plats i finalen. Ett sätt att göra det är att bygga en modell för varje skateboardåkare, använda modellerna för att simulera runbetyg och trickbetyg för varje skateboardåkare och kombinera simuleringarna för att simulera LCQ:en. Vi kan simulera flera LCQ:ar och extrahera de fyra skateboardåkarna med de högsta totalbetygen från var och en. Vår prediktion blir typvärdet av dessa resultat. Observera att denna modell anta att skateboardåkararnas prestationer är oberoende. För enkelheten antar vi att betyget på en viss run Y_i och betyget på ett visst trick X_i är oberoende för varje

skateboardåkare i. Vi antar även att alla trickbetyg och runbetyg är oberoende och likafördelade utfall från X_i respektive Y_i. Vi kan då börja med att specificera en modell för X_i och Y_i.

Baserat på observationerna i Uppgift 1 är en rimlig modell för X_i följande:

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

där $V_i \sim \text{Ber}(\theta_i)$, $Z_i \sim \text{Beta}(\alpha_i, \beta_i)$ och $V_i \perp\!\!\!\perp Z_i$. Det kan visas att

$$f_{X_i}(x_i|\theta_i, \alpha_i, \beta_i) = (1 - \theta_i)\mathbf{1}_{x_i=0} + \theta_i f_{Z_i}(z_i).$$

Valet $V_i \sim \text{Ber}(\theta_i)$ modellerar att en skateboardåkare får betyg 0 om och endast om de inte lyckas landa tricket, medan valet $Z_i \sim \text{Beta}(\alpha_i, \beta_i)$ modellerar att betyget för ett visst trick är delen av tricket som var "perfekt."

(a) Ge en punktskattning för varje θ_i, sannolikheten att skateboardåkaren i landar ett trick.

(b) Ge en punktskattning för parametrarna $[\alpha_i, \beta_i]^{\mathrm{T}}$ för varje skateboardåkare i. Finns det skateboardåkare för vilka din valda punktskattning inte existera? I så fall föreslå en alternativ punktskattning för dessa θ_i. Motivera dina val punktskattningar.

(c) Föreslå en modell för Y_i och ge en punktskattning för dina modells parametrar. Motivera dina val för modell och punktskattning.

(d) Använd din modell för $[X_i, Y_i]^{\mathrm{T}}$ för att simulera 5000 LCQ:ar och för varje simulering extrahera de fyra skateboardåkare $\boldsymbol{W} = [W_1, W_2, W_3, W_4]^{\mathrm{T}}$ med de högsta totalbetygen. Vad är typvärdet för $\boldsymbol{W}_1, \ldots, \boldsymbol{W}_{5000}$? De riktiga vinnarna för LCQ:en är

Gustavo Hoban Eaton Decenzo.

Hur många av de riktiga vinnarna förutsägs av typvärdet? Vad är skattade sannolikheten för de riktiga vinnarna baserat på dina simuleringar? Av typvärdet?

3. **En bayesiansk modell.** Som ett alternativ till den frekventistiska modellen utvecklad i Uppgift 2 kan vi betrakta en bayesiansk modell.

(a) Föreslå en simultan apriorifördelning för parametrarna $[\Theta_i, A_i, B_i]^{\mathrm{T}}$ för X_i där vi antar $\Theta_i \perp\!\!\!\perp A_i, B_i$ för alla i. Motivera ditt val.

(b) Generera 5000 slumpmässiga utfall från aposteriorifördelningen

$$f_{\theta_i, \alpha_i, \beta_i | \boldsymbol{X_i}}(\theta_i, \alpha_i, \beta_i | \boldsymbol{x}_i).$$

Plotta dina resulterande utfall för de marginella aposteriorifördelningarna:

$$f_{\theta_i|\boldsymbol{X}_i}(\theta_i|\boldsymbol{x}_i) \qquad \text{and} \qquad f_{\alpha_i,\beta_i|\boldsymbol{X}_i}(\alpha_i, \beta_i|\boldsymbol{x}_i).$$

Beräkna det aposteriori stickprovsmedelvärdet och den aposteriori stickprovsvariansen för varje parameter θ_i, α_i, och β_i för alla skateboardåkare.

(c) Föreslå en (simultan) aprioriförderlning för parametrarna för din modell Y_i från uppgift 2(c) och motivera ditt val. Du får anta att modellens parametrar för skateboardåkaren i är oberoende av alla andra parametrar inklusive θ_i, α_i och β_i. Generera 5000 utfall från aposteriorifördelningen (se till att spara dessa utfall!) och gör ett spridningsdiagram av resultatet. Vad är stickprovsmedelvärdet och stickprovsvariansen för var och en av dina parametrar baserat på dina utfall?

(d) Använd din bayesiansk modell för $[X_i, Y_i]^T$ för att simulera 5000 LCQ:ar genom att ta utfall från de lämpliga de aposteriori prediktiva fördelningarna. Vad är typvärdet av dina utfall $\boldsymbol{W}_1, \ldots, \boldsymbol{W}_{5000}$? Hur många av de riktiga vinnarna förutsägs? Vad är den skattade sannolikheten för de riktiga vinnarna baserat på dina utfall? Av typvärdet?

(e) I modellen i uppgift 3(d) antog vi att parametrarna $\boldsymbol{\Upsilon}_i$ för Y_i och parametrarna $\boldsymbol{\Theta}_i = [\Theta_i, A_i, B_i]^T$ för X_i är oberoende givet data (varför?). Samtidigt antog vi inte att $\Theta_i \perp\!\!\!\perp A_i, B_i$ är oberoende givet data. Låt $X_i^{(1)}, X_i^{(2)}, X_i^{(3)}, X_i^{(4)}$ betecknar skateboardåkare i:s fyra trickbetyg, låt $Y_i^{(1)}, Y_i^{(2)}$ betecknar skateboardåkare i:s två runbetyg och låt O_i betecknar deras totalbetyg. Rita en acyklisk riktad graf med så få kanter som möjligt så att den simultana fördelningen för $O_i, X_i^{(1)}, X_i^{(2)}, X_i^{(3)}, X_i^{(4)}, Y_i^{(1)}, Y_i^{(2)}, \Theta_i, A_i, B_i$ och $\boldsymbol{\Upsilon}$ är markovsk med avseende på den. Baserat på din graf kan du dra slutsatsen att den marginella aposteriorifördelningen för Θ_i, A_i, och B_i faktoriserar som

$$f_{\theta_i,\alpha_i,\beta_i|\boldsymbol{X_i}}(\theta_i, \alpha_i, \beta_i|\boldsymbol{x}_i) = f_{\theta_i|\boldsymbol{X_i}}(\theta_i|\boldsymbol{x}_i) f_{\alpha_i,\beta_i|\boldsymbol{X_i}}(\alpha_i, \beta_i|\boldsymbol{x}_i)?$$

Betrakta dina parametrarna $\boldsymbol{\Upsilon}_i$ för Y_i och parametrarna $\boldsymbol{\Theta}_i$ för X_i. Enligt din graf är vårt antagande att

$$\boldsymbol{\Upsilon}_i \perp\!\!\!\perp \boldsymbol{\Theta}_i | X_i^{(1)}, X_i^{(2)}, X_i^{(3)}, X_i^{(4)}, Y_i^{(1)}, Y_i^{(2)}$$

vettigt? Kan vi anta oberoenderelationen $\boldsymbol{\Upsilon}_i \perp\!\!\!\perp \boldsymbol{\Theta}_i | O_i$ om bara datan o_i är givet istället?

4. **En bayesiansk modell med en hierarki.** För att ta hänsyn till möjliga variationer i skateboardåkarnas prestationer mellan olika tävlingar kan vi bygga en modell som använda en hierarki. Som vi såg i föreläsningarna kan vi bygga en bayesiansk hierarki för $V_i \sim \text{Ber}(\theta_i)$ om vi grupperar utfall v_i enligt de olika tävlingarna. För enkelhets skull använder vi våra frekventisktiska punktskattningar för parametrarna α_i, β_i och parametrarna för Y_i från uppgift 2.

(a) Anta att $\Theta_i | A_i = \alpha_i, B_i = \beta_i \sim \text{Beta}(\alpha_i, \beta_i)$ och välj en lämplig simultan aprioriförderlning för $[\Theta_i, A_i, B_i]^T$. Motivera ditt val.

(b) Generera 5000 slumpmässiga utfall från den simultana aposteriorifördelningen

$$f_{A_i,B_i|\boldsymbol{X}_i}(a_i, b_i|\boldsymbol{x}_i).$$

Använd dina simuleringar för att generera 5000 slumpmässiga utfall från den marginella aposteriorifördelningen $\Theta_i | \boldsymbol{X}_i = \boldsymbol{x}_i$. Gör diagram med dina utfall för följande aposteriorifördelningar:

$$f_{\theta_i|\boldsymbol{X}_i}(\theta_i|\boldsymbol{x}_i) \qquad \text{and} \qquad f_{A_i,B_i|\boldsymbol{X}_i}(a_i, b_i|\boldsymbol{x}_i),$$

Ge skattningar för aposterioriväntevärdet och aposteriorivariansen för var och en av parametrarna. Hur jämför dessa varianser för θ_i med varianserna för θ_i beräknade för modellen i Uppgift 3?

(c) Med hjälp av dina 5000 utfall från del (b) simulera 5000 LCQ tävlingsvinnare och beräkna typvärdet av resultatet. Vilka är de respektive skattade sannolikheterna för de riktigaa vinnarna och ditt typvärde?

5. **Diskussion.** Det är alltid viktigt att reflektera över våra modellantaganden när vi gör statistisk inferens. Konkret är det viktigt att bedöma hur modellerna kan förbättras.

(a) Hur jämför resultaten (skateboardåkarna i typvärdena) av de olika modellerna? Vilka skateboardåkare är korrekt förutspådda och vilka inte är det? Ge några möjliga förklaringar till skillnaderna mellan de olika modellernas förutsägelser. Vilken modell föredrar du och varför?

(b) Hur jämför dina skattningar för θ_i i Uppgift 1 och dina skattade väntevärden och varianser för θ_i i Uppgift 3 och 4? Vad är det förväntade betyget för ett trick för varje skateboardåkare givet att tricket har lyckats landa? Vad är det förväntade runbetyget? Med tanke på de skateboardåkare som förutspås vinna enligt de olika modellerna, ger dessa statistikor några insikter om framgångsrika strategier för att vinna? (Till exempel, fungerar det att fokusera på ett bra runbetyg framför bra trickbetyg? Finns det exempel där denna strategi fungerar? Är det bra att ha bättre trickbetyg med stor varians eller lite sämre trickbetyg med mindre varians? osv.)

(c) Skatta väntevärdet och standardavvikelsen för varje skateboardåkares totalbetyg för modellerna i uppgift 3 och 4. Stödjer denna statistik dina förutsägelser? Enligt denna statistik, vad måste hända för att resultatet ska bli de riktiga vinnarna?

(d) I alla modellerna antog vi att skateboardåkarens prestationer är oberoende. Till exempel antog vi att alla V_i är oberoende. Verkar detta som ett rimligt antagande? Motivera ditt svar.

(e) I alla modeller struntade vi i ordningen som skateboardåkarna turades om. Verkar detta vara en rimlig sak att göra? Varför eller varför inte?