Øvelsesgang 3: Sandsynligheder

Basketballspillere, der kaster bolden i kurv efter kurv efter hinanden, beskrives som spillere med en varm hånd. Fans og spillere har længe troet på fænomenet om den varme hånd. Det vil sige, at antagelsen, om at hver scoring er uafhængigt af tidligere scoring, ikke er opfyldt. Videnskabelige analyser viser derimod, at successive scoringer er uafhængige hændelser. Vi kan tænke på et andet eksempel, hvor udfaldet af en terning er uafhængigt af tidligere udfald af en terning. Vi taler om, at en terning ikke kan huske.

Det vi vil fokusere på her er:

1) hvad vil det sige, at begivenheder er uafhængige henholdsvis at begivenheder er afhængige hændelser

2) at lære at simulere i R

3) at sammenligne en simulering med faktiske data. Dette kan anvendes til at vurdere, om fænomenet om den varme hånd kan være en reel mulighed.

Hvordan man gemmer koder

Klik på File -> New -> R Script.

Herved åbnes et tomt dokument over konsollen. Nu er det muligt at kopiere og indsætte koder her og gemme koderne. Dette er en god måde at holde styr på koder samtidigt med, at vi nu kan genbruge dem senere. Når koderne skal anvendes igen, så kan vi kopiere dem og indsætte dem i konsollen, for efterfølgende at trykke på knappen Kør.

Vi kan også gemme dette script. Første gang vi trykker på Gem, vil RStudio bede om et filnavn. Efter angivelse af filnavnet, kan filen ses under fanen Filer i nederste højre hjørne. Filen kan altid åbnes ved at trykke på filnavnet.

At komme i gang

Vi vil nu se på data omhandlende Kobe Bryant fra Los Angeles Lakers præstation mod Orlando Magic i NBA-finalen i 2009, hvor han fik titlen ”Most Valuable Player”. Mange tilskuere mente, at han viste en varm hånd. Nu ser vi på de første rækker af data fra dette spil.

```
download.file("http://www.openintro.org/stat/data/kobe.RData", destfile = "kobe.RData")
load("kobe.RData")
head(kobe)
```

I denne dataramme viser hver række et skud udført af Kobe Bryant. I kolonnen med navnet Kobe, vises et H, hvis han scorer point, og et M, hvis han rammer ved siden af nettet.

Vi får:

H M | M | H H M | M | M | M

For at bekræfte dette, brug følgende kommando

```
kobe$basket[1:9]
```

Inden for de ni skudforsøg er der seks sæt, som er adskilt med en "|" ovenfor. Deres længder er en, nul, nul, to, nul, nul, nul og nul.

1. Hvad betyder en stribelængde på 1, dvs. hvor mange scorede point og misere er der i en streng på 1? Hvad med en streglængde på 0?
Den brugerdefinerede (en funktion, der ikke er prædefineret i R) funktion calc_streak, som blev indlæst med data, kan bruges til at beregne længderne af alle strenge og derefter til at se på fordelingen.

```
kobe_streak <- calc_streak(kobe$basket)
barplot(table(kobe_streak))
```

Da vores variabel er en diskret variabel vil et kassediagram vælge.

---

2. Beskriv denne fordeling af Kobes streng længde fra NBA finalen fra 2009. Hvad er den typiske streng længde? Hvad er den længste streng længde?

Hvad skal vi sammenligne med?

Vi har vist, at Kobe flere gange havde flere mål i træk. Er det nok til at vurdere, om han har flere varme hænder? Hvad kan vi sammenligne dem med?
Lad os vende tilbage til ideen om uafhængighed for at besvare disse spørgsmål. To processer er uafhængige, hvis resultatet af den ene proces ikke påvirker resultatet af den anden proces. Hvis hvert skud, som en spiller laver, er en uafhængig proces, vil det at have lavet et mål eller at have skudt forbi ikke påvirke sandsynligheden for, at vi laver et nyt mål eller misser andet skud.
En varm hånd vil bevirke, at skud ikke er uafhængige af hinanden. Smed andre ord, så vil modellen med den varme hånd bevirke, at hvis spilleren laver et mål ved sit første skyd, så vil han have en større sandsynlighed for at lavet et mål ved hans andet skud.

Nu antager vi, at modellen med den varme hånd gælder. I løbet af sin karriere er procentdelen af den tid, hvor Kobe laver mål være 45% eller formuleret med statistiske fagtermer:

$$P(Mål\ 2 = 1) = 0{,}45$$

Nu antager vi at han har en varm hånd, og at når han først har lavet et mål, så vil sandsynligheden for, at han laver et mål efterfølgende være 60%:

$$P(Skud\ 2 = 1|Skud\ 1 = 1) = 0{,}6$$

hvor Skud = 1 betyder, at han laver mål, mens skud = 0 betyder, at han ikke laver mål.

Med andre ord, så er sandsynligheden for at der laves mål anden gang større end sandsynligheden for at der laves mål første gang. Dette indikerer, at modellen om den varme hånd gælder. Hvis modellen ikke gjalt, så ville der ikke være nogen forskel i sandsynlighederne.

Hvis modellen om den varme hånd ikke er gælder så ville følgende gælde:

$$P(Skud\ 2 = 1|Skud\ 1 = 1) = 0{,}45$$

Det vil sige, at udfaldet af første skud ikke påvirker udfaldet af andet skud.

Simuleringer i R

Selvom vi ikke har statistik over mål, så er det nemt at simulere i R. I en simulering indstiller du grundreglerne for en tilfældig proces, og derefter bruger programmet R tilfældige tal til at generere et resultat, der overholder disse regler. Som et simpelt eksempel kan vi simulere, at vi anvender en fair mønt (P(udfald af plat) = P (udfald af krone) =1)med følgende.

$$P(Plat) = P(Krone) = \frac{1}{2}$$

```
outcomes <- c("heads", "tails")

sample(outcomes, size = 1, replace = TRUE)
```

Udfaldet af vektoren kan tænkes således: Vi har en hat med sedler med to mulige udfald: Krone og Plat. Vi udtager nu en seddel fra hatten. Dette gøres flere gange, og når vi har gjort det mange gange, så vil vi forvente, at 50 % af udfaldene er Krone, mens 50 % er Plat. Bemærk, at når vi skriver replace = TRUE, betyder dette, at det er et forsøg med tilbagelægning. Den seddel, som vi trækker, lægges tilbage i hatten. Skrives derimod replace = FALSE, så udføres forsøget uden tilbagelægning. Vi gør forsøget 100 gange (size = 100)

Den resulterende vektor vil vi kalde sim_fair_coin. (Simulering med fair mønt).

Vi skriver følgende kommando:

```
sim_fair_coin <- sample(outcomes, size = 100, replace = TRUE)
```

For at se resultaterne af denne simulering skrives navnet på objektet og derefter en tabel, der viser antallet af Krone og Plat.

```
sim_fair_coin

table(sim_fair_coin)
```

Nu laver vi et nyt eksperiment, Vi anvender en falsk mønt eller en unfair mønt, hvor sandsynligheden af Krone er 20 %, mens sandsynligheden for Plat er 80%. Dette gøres i R ved at tilføje et argument prob, som giver en vektor med de to sandsynlighedsvægte.

```
sim_unfair_coin <- sample(outcomes, size = 100, replace = TRUE, prob = c(0.2, 0.8))
```

prob = c (0.2, 0.8) angiver, at udfaldet for de to elementer i vektoren er 0,2 henholdsvis 0,8.

---

3. I simuleringen, hvor mange gange får vi krone?

Hvis du vil lære mere om prøve eller enhver anden funktion, skal du huske, at du altid kan tjekke dens hjælpefil.

```
?sample
```

Simulering uafhængige scoringer

Hvis vi simulerer en basketball spiller, hvor sandsynligheden for at score point mål er 50 %, skriver vi følgende kommando:

```
outcomes <- c("H", "M")

sim_basket <- sample(outcomes, size = 1, replace = TRUE)
```

For at lave en valid sammenligning mellem Kobe og vores simulerede uafhængige spiller, skal vi sammenligne udfaldet af skud hos disse to.

---

4. Kør en simulering med 133 skud. Tildel outputtet fra denne simulering til et nyt objekt kaldet sim_basket.
   Bemærk, at vi har navngivet den nye vektor sim_basket, det samme navn, som vi gav den forrige vektor, hvilket afspejler en optagelsesprocent på 50%. I denne situation overskriver R det gamle objekt med det nye objekt, så sørg altid for, at du ikke har brug for oplysningerne i en gammel vektor, før du tildeler dets navn igen.
   Med resultaterne af simuleringen gemt i sim_basket, har vi de nødvendige data til at sammenligne Kobe med vores uafhængige spiller. Vi kan nu sammenligne Kobe's data med vores simulerede data.

```
kobe$basket

sim_basket
```

Opgave

Sammenlign Kobe med en uafhængig spiller

Brug calc_streak til at beregne længden af mål for sim_basket.

1. Beskriv fordelingen af antallet mål. Hvad er det typiske antal mål for den uafhængige simulerede spiller, der har en scoringsprocent på 45%? Hvad er det gennemsnitlige antal mål med en scoringsprocent på 45%?
2. Hvis vi skulle køre simuleringen for den uafhængige spiller en anden gang, hvordan kan vi så forvente, at fordelingen vil være i forhold til fordelingen fra ovenstående spørgsmål? Præcis det samme? Noget lignende? Helt forskelligt? Forklar begrundelsen.
3. Hvordan sammenlignes Kobe Bryants fordelingen af mål med fordelingen for den simulerede spiller? Når vi sammenligner disse, understøttes modellen med den varme hånd? Forklar.