Laboration 2

Sannolikhetsteori I

*Monir Bounadi**

2018-12-26

Inledning

Syftet med laboration 2 är att undersöka slumptalsgeneratorn i R samt att illustrera de teoretiska resultaten angående fördelning för funktioner av stokastiska variabler med hjälp av simulering. Med att simulera menar vi att generera "slumptal". Simuleringar är oftast mycket användbara i situationer där det är svårt att komma fram till resultaten teoretiskt eller om man vill skaffa sig en intuitiv förståelse för hur saker fungerar. De "slumptal" man får från datorn är egentligen inte slumpmässiga utan genereras med hjälp av en deterministisk algoritm. Oftast är dock algoritmerna så välgjorda att detta inte leder till några problem. I statistik och sannolikhetsteori vill vi oftast att slumptalen ska se ut att ha genererats av någon av de sannolikhetsfördelningar vi arbetar med—i denna laboration kommer vi att simulera från likformig fördelning, normalfördelning och exponentialfördelning.

Uppgift 1: Test av slumptalsgeneratorn i R

Vi ska i denna labb undersöka slumptalsgeneratorn i R, och vi kommer att börja med den för den likformiga fördelningen (kontinuerlig form). Innan vi går på de kod-relaterade uppgifterna vill vi att ni besvarar följande teoretiska uppgifter.

Teoretiska uppgifter

Vi delar in intervallet $[0, 1]$ i k lika stora delintervall, som vi kallar för klasser. Sedan genererar vi n observationer från en (kontinuerlig) likformig fördelning på intervallet $[0, 1]$. Antalet X_1 av dessa n observationer som hamnar i första intervallet är en slumpvariabel, liksom antalet X_2 som hamnar i andra intervallet, osv. Besvara följande frågor.

1. Ange det första och sista delintervallet på formen $[a, b]$, där a och b är intervallets gränser.
2. Vad är längden på var och ett av de k delintervallen?
3. Låt säga att vi genererar ett slumptal från den kontinuerliga likformiga fördelningen på $[0, 1]$.
   - Vad är sannolikheten att detta slumptal hamnar i det första delintervallet?
   - Vad är sannolikheten att detta slumptal inte hamnar i det första delintervallet?
   - Vilken sannolikhetsfördelning beskriver denna situation? Ange parametrarna hos fördelningen.
4. Om vi nu som i beskrivningen genererar n sådana slumptal, oberoende av varandra, vad är sannolikheten att j av dessa hamnar i det första delintervallet? Dvs, för $j = 0, 1, \ldots, n$, vad är $\mathbb{P}(X_1 = j)$? Vilken fördelning följer X_1? Ange parametrarna hos fördelningen.
5. Har antalet slumptal i de andra intervallen, dvs någon av de stokastiska variablerna $X_2, \ldots, X_k$, samma sannolikhetsfördelning som X_1? Är de oberoende?

*Tidigare versioner av Benjamin Kjelsson, Maria Deijfen, Andreas Nordvall Lagerås, Tom Britton, Jens Malmros, och OE.

6. Betrakta nu i stället andelen $Y_1 = X_1/n$ av observationerna som hamnar i första intervallet; det är också en slumpvariabel. Ange väntevärde, varians och standardavvikelse för Y_1. Tips: Börja med att härleda $\mathbb{E}(X_1)$ och $\mathrm{Var}(X_1)$.

Kod-relaterad uppgifter: Slumptalsgeneratorn för den kontinuerliga likformiga fördelningen i R

Vi ska först undersöka slumptalsgeneratorn för likformiga slumptal. Vi skapar först ett dataset (en vektor) med 100 observationer (slumptal) från en likformig fördelning på intervallet $[0, 1]$:

```
# Först ger vi ett "frö" (seed) till slumptalsgeneratorn genom funktionen set.seed
# Detta gör att jag kan få exakt samma slumptal, och således resultat, som dig
# när jag kör din kod
set.seed(19880210) # fyll i ditt egna födelsedatum. Om ni jobbar i par, välj den enas.

slumptal <- runif(100) # Skriv ?runif i Console i RStudio för mer info om denna funktion
```

Vi gör sedan ett histogram över slumptalen på detta vis:

```
hist(slumptal,
  breaks = seq(from = 0, to = 1, length.out = 10 + 1),
  main = "",
  ylab = "Antal",
  xlab = "Värde")
```

Argumentet `breaks = seq(from = 0, to = 1, length.out = 10 + 1)` säger till funktionen `hist` att göra ett histogram med staplar som avgränsas till höger och vänster av punkterna i vektorn `seq(from = 0, to = 1, length.out = 10 + 1)`. Vi visar denna vektor:

```
seq(from = 0, to = 1, length.out = 10 + 1)
```

```
##  [1] 0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0
```

Vi skriver `10 + 1` för att visa att detta leder till 10 stycken staplar. Intervallet [0, 1] delas alltså in i 10 lika stora delintervall, $[0, 0.1), [0.1, 0.2), \ldots, [0.9, 1]$, och en stapel i histogrammet visar hur många observationer i vektorn slumptal som hamnat i delintervallet som motsvaras av den stapeln.

Andelar

Hur ska man veta hur stora avvikelser från den förväntade andelen observationer i en viss klass som kan accepteras utan att slumptalsgeneratorn måste anses defekt? Storleksordningen på rimliga avvikelser kan man få fram genom att titta på (den teoretiska) standardavvikelsen: Man kan visa att avvikelser från den förväntade andelen observationer i en viss klass på 2–3 gånger standardavvikelsen inte är alltför osannolika. Avvikelser som är större än 2–3 gånger standardavvikelsen är dock mycket osannolika och man kan börja misstänka att slumptalsgeneratorn inte fungerar så bra.

Vi ska nu göra histogram över andelen observationer i klasserna (intervallen) och undersöka om de observerade andelarna tycks hålla sig inom felmarginalen (2–3 standardavvikelser från väntevärdet). Följande funktion skapar ett sådant histogram:

```
# x: vektor med slumptal (eller annan data)
prop_hist <- function(x, xlab = "Värde") {
  p <- hist(x, plot = FALSE)
  p$counts <- p$counts / sum(p$counts) # Konvertera antal till andelar
  plot(p,
    main = "Andelshistogram",
```

```
    ylab = "Andel",
    xlab = xlab)
}
```

Notera att ni behöver kopiera definitionen på denna funktion till er labbrapport innan ni kan använda den, men ni förväntas inte förklara hur funktionen fungerar och ni ska heller inte ändra i den.

Värdena i vektorn slumptal kan nu plottas genom att ha med ett stycke likt det följande i din `.Rmd`-fil:

```
prop_hist(slumptal)
```

Låt säga att du vet att när du genererar $n = 100$ slumptal så är standardavvikelsen för andelen slumptal i något av de 10 intervallen lika med $D = 0.015$ (OBS! Påhittat värde). Då kan du göra ett andelshistogram med linjer som är plus/minus t.ex. 2 standardavvikelser från andelens förväntade värde 0.1 genom att ha med ett stycke likt det följande i din `.Rmd`-fil:

```
E <- 0.1
D <- 0.015

prop_hist(slumptal)
abline(a = E, b = 0, col = "grey") # Väntevärdet
abline(a = E + 2 * D, b = 0, col = "red") # Väntevärdet + 2 standardavvikelser
abline(a = E - 2 * D, b = 0, col = "red") # Väntevärdet - 2 standardavvikelser
```

På så vis kan du lätt se om andelarna faller utanför det intervall i vilket andelen förväntas ligga för det mesta. Glöm inte att diagrammet måste numreras och ges en beskrivande text!

Vill du istället plotta andelar för n stycken nya slumptal från den kontinuerliga likformiga fördelningen mellan 0 och 1 så kan du skriva

```
set.seed(19880210) # fyll i ditt egna födelsedatum. Om ni jobbar i par, välj den enas.

prop_hist(runif(n))
```

Se till att använda `abline` med rätt standardavvikelse för det n du väljer. Notera att vi här använder `set.seed` igen, i samma kodstycke som vi använder `runif`. Du ska också göra detta, så att din analys och laborationsrapport kan återskapas av någon annan som kör din kod.

Din uppgift är nu att undersöka dessa plottar för olika värden på n:

- För ett givet värde av n, ser det ut som en likformig fördelning? Motivera varför genom att hänvisa till de teoretiska resultaten.
- Vad händer då antalet slumptal n blir stort? Experimentera gärna på din egen dator, men du behöver inte ta med alla histogram i rapporten.
- Jämför andelen observationer i klasserna med det förväntade och se om avvikelserna verkar stora (jämfört med standardavvikelsen).
- Får slumptalsgeneratorn i R godkänt?

Redovisning av uppgift 1

- Svar på frågorna i teoridelen inklusive härledning och motivering.
- Svar på och resonemang kring de avslutande frågorna i texten den kod-relaterade delen.
- Två andelshistogram som illustrerar era svar: ett för ett litet värde på n (t.ex. $n = 100$), och ett för ett (mycket) större värde på n (t.ex. $n = 1000$).

Uppgift 2: Normal- och exponentialfördelade slumptal

R kan även producera normal- och exponentialfördelade slumptal. Skriver vi

```
rnorm(n)
```

får vi en vektor med n stycken $\mathcal{N}(0,1)$-fördelade slumptal. Skriver vi istället

```
rnorm(n, mean = m, sd = s)
```

får vi istället en vektor med n stycken $\mathcal{N}(m, s^2)$-fördelade slumptal (s är standardavvikelsen).

På liknande vis kan vi få exponentialfördelade slumpvariabler genom

```
rexp(n, rate = a)
```

Här får vi alltså en vektor med n stycken exponentialfördelade slumptal med väntevärde $1/a$. För att få histogram över relativa frekvenser i olika klasser kan vi återigen använda funktionen `prop_hist` definierad ovan.

Vi kan visa två andelshistogram bredvid varandra på följande vis (exempel):

```
set.seed(19880210) # fyll i ditt egna födelsedatum. Om ni jobbar i par, välj den enas.

par(mfrow = c(1, 2))
prop_hist(rexp(20, rate = 1 / 3)) # väntevärde 3
prop_hist(rexp(1000, rate = 1 / 3)) # väntevärde 3
```

Nu är det din tur att experimentera.

- Välj själv värden på `m` och `s` för normalfördelningen, och `a` för exponentialfördelningen. Skriv i texten vilka värden du väljer och vad dessa parametrar heter för fördelningen (t.ex. väntevärde).
- Prova att göra andelsplottar för olika värden på n, för att bilda en uppfattning av vad som händer. Alla dessa plottar ska inte tas med i rapporten. Endast två per fördelning behövs, se "Redovisning av uppgift 2" nedan.
- Ser talen normal- respektive exponentialfördelade ut? Motivera ditt svar:
  - Jämför med egenskaper hos de (teoretiska) fördelningarna. Vilka egenskaper är utmärkande för normal- respektive exponentialfördelninarna?
  - Vad är det du jämför med? Fördelningsfunktionen eller täthetsfunktionen?
- Kom ihåg att använda funktionen `set.seed` med ditt födelsedatum innan du använder någon av funktionerna `rexp` eller `rnorm`. Ska du t.ex. i din rapport göra en plot av n stycken exponentialfördelade slumpvariabler med väntevärde $1/a$ så skriver du

```
set.seed(19880210) # fyll i ditt egna födelsedatum. Om ni jobbar i par, välj den enas.

prop_hist(rexp(n, rate = a))
```

Redovisning av uppgift 2

Redovisningen skall inkludera följande

- Svar på den avslutande frågan i texten ovan.
- Endast två andelshistogram sida vid sida med normalfördelade slumptal, och detsamma med exponentialfördelade slumptal

Uppgift 3: Fördelning för en funktion av stokastiska variabler

Vi ska nu använda simulerade slumptal för att få en uppfattning om fördelningen för en funktion av en stokastisk variabel.Exempel 3.48 i boken visar teoretiskt att om X är exponentialfördelad med väntevärde $1/2$ så är $V = 1 - e^{-2X}$ kontinuerligt likformigt fördelad mellan 0 och 1. Vi kan göra en simulering och se om detta verkar rimligt.

```r
set.seed(19880210) # fyll i ditt egna födelsedatum. Om ni jobbar i par, välj den enas.

x <- rexp(n, rate = 2)
v <- 1 - exp(-2 * x)

# Plotta x och v bredvid varandra
par(mfrow = c(1, 2))

prop_hist(x, xlab = "X")
prop_hist(v, xlab = "V")
```

Jämför de två histogrammen. Ser det ut som exemplets påstående gäller? Dvs, leder transformationen till en likformig fördelning? Motivera varför! Tänk på vad du redan gjort i denna labb!

Redovisning av uppgift 3

Redovisningen skall inkludera följande:

- Svar på frågan i texten ovan, inklusive motivering med hänvisning till tidigare uppgift(er) i denna laboration.
- Histogrammen med de simulerade X respektive V för ditt egen val av n.