Fördelningar, simulering och fördelningsanpassning

Datorlaboration 1

Skriv ditt namn här

2024-04-06

Syftet med dagens laboration är att du ska - träna på begreppen täthetsfunktion och fördelningsfunktion - träna på att simulera slumptal från en fördelning - träna på att använda olika grafiska metoder för att undersöka vilka fördelningar ett datamaterial kan komma från

Innehållsförteckning

Datamaterial jordprov

I skogsområdet ASA försökspark i Småland är 94 olika gropar grävda i marken och från varje grop är jordprover tagna där bland mycket annat aluminiumhalt och calciumhalt är uppmätta (mg/g). Data finns i filen `data/lab1_filer/jordprov.Rdata` som innehåller de två variablerna `al` och `ca`.

Öppna filen i RStudio, exempelvis genom att klicka på den under Files. Du kan se de uppmätta halterna genom att klicka på ikonen längst till höger på raden

`jordprov`. Alternativt skriver du `View(jordprov)` i RStudios konsoll (fönster). För att nå variabeln `al` i datamaterialet `jordprov` skriver du `jordprov$al`.

```
load("data/lab1_filer/jordprov.Rdata")
# skriv din R-kod här
```

1 Överblick av datamaterialet

Först vill man beräkna några sammanfattande mått för data (medelvärde, min, max, standardavvikelse o.s.v.) och se några översiktsfigurer. Kommandot `mean(jordprov$al)` ger dig medelvärdet av aluminiummätningarna.

Uppgift 1.1

Gör en översiktsanalys av aluminium- och calciumhalterna genom att använda följande funktioner i R: `sd()`, `summary()`, `hist()`, `boxplot()`, `plot()`. Pröva också `plot(jordprov$al,jordprov$ca)`.
Då man har mätningar, $x_1, x_2, \ldots, x_n$, fås mycket information genom att rita upp den s.k. empiriska fördelningsfunktionen (*empirical cumulative distribution function* på engelska). Datapunkterna, x_i sorteras från minsta till största. Andelen datapunkter som är mindre eller lika med x_i plottas sedan mot x_i. Det blir en växande trapp-stegs-funk-tion som tar ett skutt med höjd $1/n$ för varje data-punkt. I R kan du få funktionen utritad genom kommandot `plot.ecdf()`.

```
# skriv din R-kod här
```

Uppgift 1.2

Rita ut den empiriska fördelningsfunktionen för aluminiumhalterna. En grid läggs in i figuren om du skriver `plot.ecdf(jordprov$al,panel.first=grid())`. Använd figuren för att ta reda på hur stor andel av mätningarna som understeg 80 mg/g.

Svar:

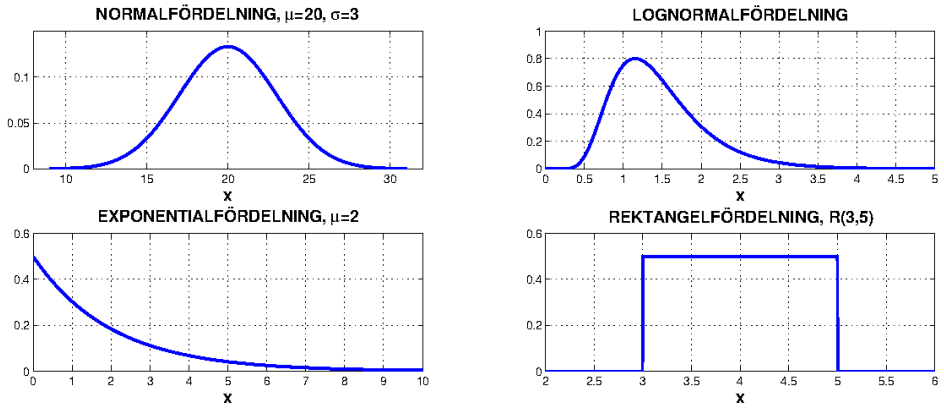
Uppgift 1.3

Vilken aluminiumhalt överstigs i 70% av mätningarna?

Svar:

Passar någon standardfördelning till mina data?

Nu vill vi, med grafiska metoder, undersöka om de två dataseten kan modelleras med några standardfördelningar. Några kontinuerliga standardfördelningar som vi stöter på i kursen är normalfördelningen, lognormalfördelningen, exponentialfördelningen och rektangelfördelningen (likformig fördelning eller på engelska uniform).



När du tittar på histogrammet för aluminiumhalter verkar det inte orimligt att de skulle vara normalfördelade, men ett histogram är oftast ett trubbigt instrument då man vill anpassa en standardfördelning till data. En mer använd metod är att rita ut data i ett så kallat fördelningspapper eller QQ-plot. För att illustrera metoden är vi hjälpta av att se hur den fungerar på stickprov där vi verkligen vet fördelningen, vi behöver alltså veta hur man skapar slumptal från olika fördelningar.

2 Simulering av slumpvariabler i R

I R finns det färdiga funktioner för simulering från respektive fördelning. Några exempel på dessa funktioner ser du i tabellen nedan.

Fördelning	Funktion	Exempel
Normal	`rnorm(antal,mean,stddev)`	`rnorm(20,3,1)`
Exponential	`rexp(antal,1/mean)`	`rexp(50,0.5)`
Rektangel	`runif(antal,min,max)`	`runif(30,-2,5)`
Binomial	`rbinom(antal,n,p)`	`rbinom(10,5,0.2)`
Poisson	`rpois(antal,mean)`	`rpois(25,4)`

Vill du t.ex. simulera 100 slumptal från en normalfördelning med väntevärde (mean) 20 och standardavvikelse (stddev) 3 och lägga dem i variabeln `norm1` gör du det genom `norm1<-rnorm(100,20,3)`.

Uppgift 2.1

Rita ut de 100 normalfördelade slumptalen i `norm1` i ett histogram. Ser histogrammet ut som den teoretiska normalfördelningen i figuren ovan? Vad händer om du i stället gör ett histogram på 1000 simulerade normalfördelade slumptal.

Svar:

Uppgift 2.2

Skapa 1000 slumptal från en exponentialfördelning med väntevärde 2 och lägg dem i variabeln `exp1`. Gör ett histogram och jämför med exponentialfördelningens täthetsfunktion som är utritad i figuren ovan.

Uppgift 2.3

Gör samma sak för en rektangelfördelning $R(3, 5)$, d.v.s. dra 1000 slumptal (kalla dem unif1), visualisera dem med ett histogram och jämför med den teoretiska täthetsfördelningen.

QQ-plot

Ofta har man frågeställningen om data i ett stickprov kan tänkas modelleras med en teoretisk standardfördelning. Detta gällde t.ex. aluminiumhalterna i marken som du studerade tidigare. En grafisk metod när man försöker anpassa sina data till en fördelning är att använda sig av en så kallad QQ-plot där Q står för kvantil (quantile). Värdena i datamaterialet jämförs med de man kunde förvänta sig från en viss teoretisk fördelning. Om data överensstämmer med de förväntade kommer punkterna i en QQ-plot att ligga utmed en rät linje. Omvänt, om QQ-plotten visar stor avvikelse från en rät linje passar inte den fördelning vi testat med till våra data. För att pröva om ett stickprov kan tänkas komma från en normalfördelning är kommandot `qqnorm(stickprovsnamn)`.

Uppgift 2.4

Pröva hur QQ-plotten ser ut då du anpassar stickprovet med 100 normalfördelade slumptal (`norm1`) till en normalfördelning genom kommandot `qqnorm(norm1)`. Vad händer när du även skriver kommadot `qqline(norm1)`?

Svar:

Eftersom vi vet att `norm1` innehåller normalfördelade slumptal bör anpassningen förstås vara god. Observera dock att man inte kan kräva en perfekt rät linje i plotten, vi har ju att göra med slumptal. En mindre avvikelse i linjens båda ändar är inte ovanligt.

Uppgift 2.5

Hur ser det ut om du försöker anpassa exponentialfördelade slumptal `exp1` eller rektangelfördelade slumptal `unif1` till en normalfördelning genom att använda kommandona `qqnorm(exp1)` respektive `qqnorm(unif1)`.

Svar:

3 Modell för aluminiumhalten

Vi tittar på aluminiumhalterna i de 94 jordproverna igen.

Uppgift 3.1

Avgör om dessa halter kan modelleras med en normalfördelning. Använd metoden QQ-plot med kommandot `qqnorm(jordprov$al)`. Verkar det vara en god anpassning?

Svar:

Tidigare har du använt R för att beräkna medelvärde (m) och standardavvikelse (s) för aluminiumhalterna. Dessa värden kan du använda som uppskattningar av väntevärdet μ och standardavvikelsen σ i den anpassade fördelningen. Givet att du tyckter att en normalfördelning passar bra till data *kan vi nu sätta upp en modell för våra observationer.*

Modell: X = "aluminiumhalten i ett jordprov"; X är normalfördelad med väntevärde μ och standardavvikelse σ. Vi skattar dessa parametrar som $\hat{\mu} = m$ ("medelvärdet för stickprovet") och $\hat{\sigma} = s$ ("standardavvikelse för stickprovet").

Uppgift 3.2

Beräkna de skattade parametrarna och för in dem i din modell genom att skriva

```
m = mean(jordprov$al)
s = sd(jordprov$al)
hist(jordprov$al,probability=TRUE)
xx <- seq(40,120,by=0.2)
lines(xx,dnorm(xx,m,s))
```

Nu kan du jämföra täthetsfunktionen från modellen med histogrammet för stickprovet. De borde vara ganska lika för att modellen ska anses vara bra.
Med denna modell kan du beräkna sannolikheter och göra förutsägelser kring framtida mätningar. Du kan t.ex. beräkna sannolikheten att en ny aluminiummätning kommer att överstiga 80 mg/g, eller bestämma den Al-halt som kommer att understigas av 10% av kommande mätningar.
Vilka värden har de skattade parametrarna?

Svar:

> Uppgift 3.3
>
> Antag att du vill beräkna sannolikheten att aluminiumhalten i ett prov överstiger 80 mg/g, med andra ord du vill beräkna $P(X > 80)$. På övningarna har du gjort det med hjälp av räknare och/eller tabell. I R är kommandot för beräkning av normalfördelningens fördelningsfunktion $P(X \leq x)$ `pnorm()`. Förklara varför den önskade sannolikheten beräknas med kommandot `1-pnorm(80,m,s)`.

Svar:

> Uppgift 3.4
>
> Den Al-halt som kommer att understigas av 10% av kommande mätningar är en kvantil i normalfördelningen. Använd R för att beräkna kvantilen med kommandot `qnorm(0.1,m,s)`.

Svar:

4 Modell för calciumhalten

> Uppgift 4.1
>
> Pröva med en QQ-plot om även calciumhalterna kan modelleras med en normalfördelning. Ser det bra ut?

Svar:

En annan standardfördelning som är vanlig för biodata är lognormalfördelningen. Mätningar kan modelleras med en lognormalfördelning om de **logaritmerade** mätningarna passar bra till en normalfördelning. Det innebär att det inte behövs någon speciell QQ-plot för denna fördelning, man kan använda `qqnorm(log(stickprovsnamn))`.

> Uppgift 4.2
>
> Pröva om calciummätningarna verkar vara lognormalfördelade. Ser det bra ut?

Svar:

Nu ska du uppskatta sannolikheten att en calciummätning överstiger 30 mg/g. Först behöver du sätta upp en modell för calciummätvärden. Antag att calciumvärden följer en lognormalfördelning. Sen använder du modellen för att göra beräkningar.

Uppgift 4.3

Om Y = calciumhalt söker vi $P(Y > 30)$, vilket är ekvivalent med att $P(\log(Y) > \log(30))$. Om Y är lognormalfördelad gäller att $\log(Y)$ är normalfördelad. Sätt upp en modell för $\log(Y)$ och skatta dess parametrar. Vilka värden har de skattade parametrarna?

Svar:

Uppgift 4.4

Beräkna sannolikheten att log av calciumhalten överstiger log(30). Visa det kommando du använder för att uföra denna beräkning.

Svar: