MASB11: BIOSTATISTISK GRUNDKURS
DATORLABORATION 1, 3 APRIL 2019
FÖRDELNINGAR, SIMULERING OCH FÖRDELNINGSANPASSNING

Syfte

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

Förberedelseuppgifter

Du måste ha arbetat ordentligt med de väsentligaste begreppen i kapitel 4.1–4.3 i kursboken. Repetera vid behov begreppen *fördelningsfunktion, sannolikhetsfunktion och täthetsfunktion* samt *normalfördelningen*. **Du skall ha gjort följande uppgifter *innan* du kommer till laborationen.**

Hemuppgift 1:

Orientera dig om `R` och de olika fönstren som dyker upp när man startar `RStudio`. Ett minimum är att läsa igenom de första sex sidorna i stencilen ”Introduktion till `R`”. Ladda gärna ner `R` och `RStudio` till din dator och kör några kommandon från ”Introduktionen”. Har du inte möjlighet till detta kan du använda datorerna i MH:230 eller MH:231 där du loggar in med hjälp av ditt Stil-konto (både `R` och `RStudio` är redan installerat där).

Hemuppgift 2:

För att träna på normalfördelningen gör **uppgift 3.81** (serumkolesterolhalt hos kvinnor) och **uppgift 3.86** (vikt hos 10-åriga flickor) i arbetsmaterialet. Fastnar du kan du få hjälp på övning 3.

Hemuppgift 3:

Det är viktigt att skilja **data** (observationerna, mätningarna från ett visst slumpmässigt fenomen) från **modellen** man gör om den slumpmässiga variationen. I kursen använder vi begreppen *empirisk fördelningsfunktion*, *frekvensfunktion (täthetsfunktion)*, *stolpdiagram*, *histogram*, *sannolikhetsfunktion* och *fördelningsfunktion*. Vilka av dessa hör till ”data” och vilka hör till ”modellen”? Det är också möjligt att para ihop begreppen, så att ett modellbegrepp motsvarar ett annat begrepp för data. Gör det!

Data	Modell

1 Datamaterial jordprov

I skogsområdet ASAs 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 `jordprov.Rdata` som innehåller de två variablerna `al` och `ca`. Öppna filen i det övre högra fönstret i RStudio. Du kan se de uppmätta halterna genom att klicka på ikonen längst till höger på raden `jordprov`. Alternativt skriver du `View(jordprov)` i Rstudios console i nedre vänstra fönstret.

1.1 Överblick av materialet

Först vill man beräkna några sammanfattande mått för data (medelvärde, min, max, standardavvikelse o.s.v.) och se några översiktsfigurer. För att nå variabeln `al` i datamaterialet `jordprov` skriver du `jordprov$al`. Kommandot `mean(jordprov$al)` dig medelvärdet av aluminiummätningarna.

Uppgift 1.1:

Gör en översiktsanalys av aluminium- och calciumhalterna genom att t.ex. använda följande funktioner i R: `summary()`, `sd()`, `hist()`, `plot()`, `boxplot()`. Pröva också `plot(jordprov$al,jordprov$ca)`.

Svar:

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 trappstegsfunktion som tar ett skutt med höjd $1/n$ för varje datapunkt. I `R` kan du få funktionen utritad genom kommandot `plot.ecdf()`.

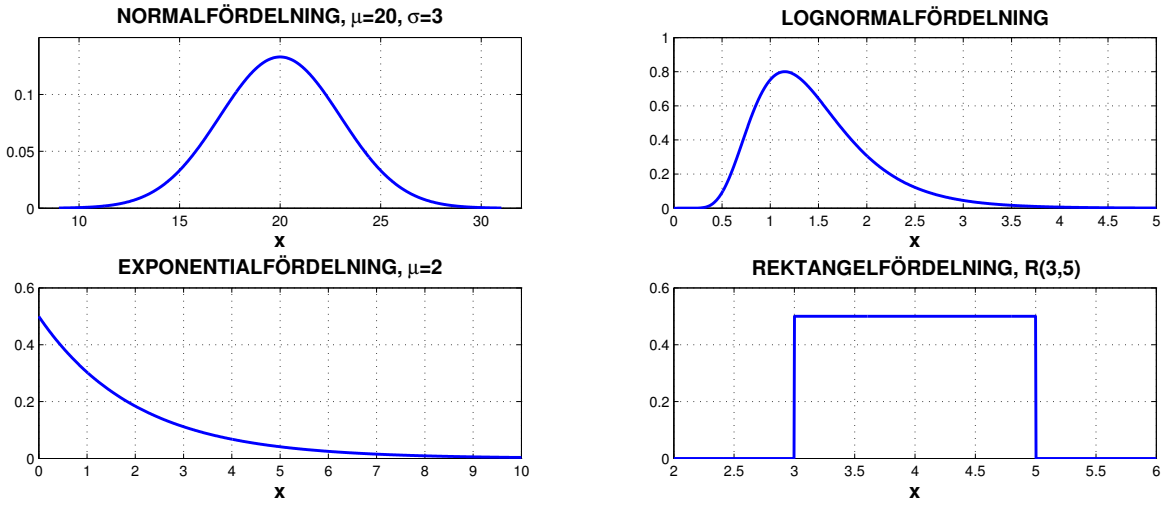
Uppgift 1.2:

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. Vilken aluminiumhalt övstigs i 70 % av mätningarna?

Svar:

1.2 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? Pröva vad som 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 t.ex. är utritad i figuren ovan.

Svar:

Uppgift 2.3:

Gör samma sak för en rektangelfördelning R(3,5), lägg de 1000 slumptalen i förslgsvis `unif1`.

Svar:

2.1 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)`.

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:

Pröva nu vad som händer då 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:

Använd metoden QQ-plot för att avgöra om dessa halter kan modelleras med en normalfördelning (`qqnorm(jordprov$al)`).

Svar:

Tidigare har du låtit R beräkna medelvärde och standardavvikelse för aluminiumhalterna. Dessa värden kan du använda som uppskattningar av väntevärdet μ och standardavvikelsen σ i den anpassade fördelningen. Om du tyckte att en normalfördelning passade bra till data **kan vi nu sätta upp en modell för våra observationer.**

Modell: X = aluminiumhalten i ett jordprov; X är normalfördelad med väntevärde "*medelvärdet för data*" och standardavvikelsen "*standardavvikelse för data*".

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.

Uppgift 3.2:

Antag att du vill beräkna sannolikheten att aluminiumhalten i ett prov överstiger 80, med andra ord du vill beräkna $P(X \geq 80)$. På övningarna har du gjort det med hjälp av räknare och/eller tabell. I `R` är kommandot för beräkning av normalfördelningens fördelningsfunktion `pnorm()`. Skriv först `?pnorm` så ser du hur kommandot ska skrivas.

Svar:_____

Uppgift 3.3:

Då du vill bestämma den Al-halt som kommer att understigas av 10 % av kommande mätningar är det en kvantil i normalfördelningen som efterfrågas. Skissa gärna fördelningen, markera kvantilen men använd `qnorm()` i `R`.

Svar:_____

4 Modell för calciumhalten

Uppgift 4.1:

Pröva med en QQ-plot om även calciumhalterna kan modelleras med en normalfördelning.

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.

Svar:_____

Uppgift 4.3:

Uppskatta sannolikheten att en calciummätning överstiger 30 mg/g. Ledning: Om $Y =$ calciumhalt söker vi $P(Y > 30)$. vilket är ekvivalent med att $P(\ln(Y) > \ln(30))$. Men om Y är lognormalfördelad gäller att $\ln(Y)$ är normalfördelad. Tänk ut hur väntevärdet och standardavvikelsen i den normalfördelningen kan skattas!

Svar: