Kontinuerlige stokastiske variabler

Notat til STK1100

Ørnulf Borgan
Matematisk institutt
Universitetet i Oslo

Februar 2004, 2011

Formål

I avsnitt 4.1 i *Modern Mathematical Statistics with Applications* av Devore & Berk defineres sannsynligheten for at en kontinuerlig stokastisk variabel skal anta en verdi i et intervall som integralet av en tetthetsfunksjon over intervallet. Formålet med dette notatet er å gi en motivasjon for denne definisjonen.

Innledning

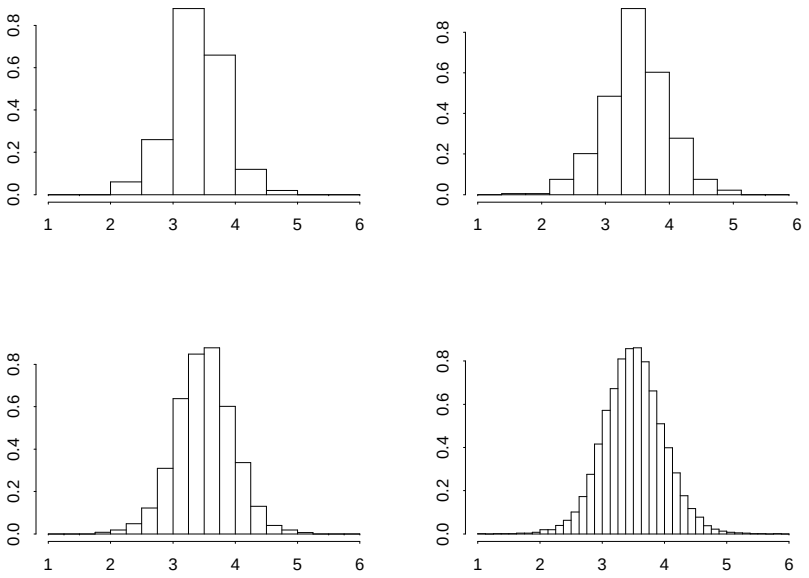
Kapittel 3 i boka til Devore & Berk tar for seg *diskrete* stokastiske variabler, dvs. variabler som kan anta et endelig antall eller et tellbart uendelig antall mulige verdier. Vi vil her se på *kontinuerlige* stokastiske variabler. Det er stokastiske variabler som kan anta alle verdier[1] i et intervall på tallinja (eventuelt på hele tallinja).

Noen eksempler på kontinuerlige stokastiske variabler er:

- vekten til en nyfødt jente
- høyden til en norsk rekrutt
- tiden mellom to oppringninger til en telefonsentral
- endringen i en aksjekurs i løpet av en dag

Sannsynlighetsfordelingen for en diskret stokastisk variabel kan gis ved punktsannsynligheten $p(x_i) = \mathrm{P}(X = x_i)$. Denne gir sannsynligheten for de ulike

[1] I praksis vil ikke vekten til en nyfødt jente bli målt mer nøyaktig enn til nærmeste tiende gram, og høyden til en rekrutt vil bare måles til nærmeste hele (eller halve) centimeter. Så selv om vekt og høyde i prinsippet kan anta alle verdier i et intervall, er det på grunn av avrunding bare et endelig antall forskjellige verdier en vil registrere. Vi kunne derfor valgt å se på vekt og høyde som diskrete stokastiske variabler som bare kan anta et endelig antall mulige verdier. Men det er mer hensiktsmessig å betrakte dem som kontinuerlige variabler som i prinsippet kan anta alle verdier i et intervall. Dette er tilsvarende som i mange andre situasjoner ved matematisk modellering: om vi benytter en diskret eller en kontinuerlig modell er ofte et spørsmål om hva som er (matematisk og/eller numerisk) mest hensiktsmessig.



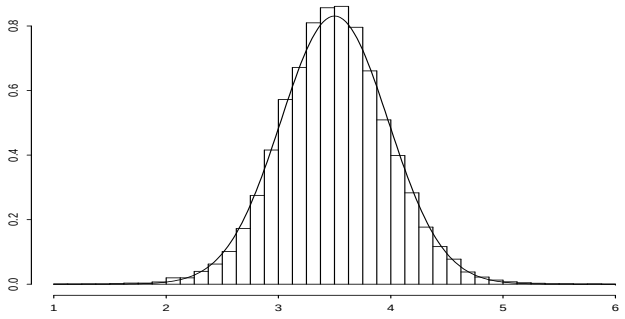
Figur 1: Histogram av fødselsvekter for "fullbårne" jenter født i Norge i 1980. Histogrammene er basert på ulike klassebredder og antall registreringer av fødselsvekter: øverst til venstre 100 vekter, øverst til høyre 500 vekter, nederst til venstre 2500 vekter og nederst til høyre 20 000 vekter.

verdiene X kan anta; jf. kapittel 3 i boka til Devore & Berk. Denne frangangsmåte kan vi ikke bruke for en kontinuerlig stokastisk variabel som kan anta alle verdier i et intervall på tallinja. Når vi skal angi sannsynlighetsfordelingen for en kontinuerlig stokastisk variabel må vi derfor gå fram på en annen måte.

Histogram og sannsynlighetstetthet

For å forklare hvordan vi kan angi sannsynlighetsfordelingen til en kontinuerlig stokastisk variabel, vil vi se på variabelen $X =$ "vekt til nyfødt jente". Til dette benytter vi data fra Medisinsk fødselsregister om fødselsvekter til jenter født i Norge i 1980. Vi ser bare på "fullbårne" jenter, så vi begrenser oss til fødsler hvor svangerskapet varte mellom 37 og 43 uker.

Vi ser først på et tilfeldig utvalg av 100 nyfødte jenter. Et histogram av fødselsvektene til disse er gitt øverst til venstre i Figur 1. Vi bruker her klassebredde 0.5 kg, dvs. vi deler inn fødselsvektene i intervaller som er 0.5 kg brede når vi lager histogrammet. Histogrammet er normert slik at *arealet* av en søyle er lik den relative frekvensen av fødselsvekter i det intervallet søylen dekker. Merk at den relative frekvensen av fødselsvekter mellom (for eksempel) 2.0 kg og 3.5 kg er summen av de relative frekvensene av fødselsvekter mellom 2.0 kg og 2.5 kg, mellom 2.5 kg og 3.0 kg og mellom 3.0 kg og 3.5 kg, altså arealet under histogrammet mellom 2.0 kg og 3.5 kg.



Figur 2: Histogram av fødselsvekter for 20 000 jenter med inntegnet sannsynlighetstetthet.

Vi ser så på et histogram til fødselsvektene av et tilfeldig utvalg av 500 jenter slik som vist øverst til høyre i Figur 1. Siden vi nå har flere fødselsvekter, reduserer vi klassebredden til 0.375 kg. Den relative frekvensen av fødselsvekter mellom 2.0 kg og 3.5 kg er nå summen av de relative frekvensene av fødselsvekter i intervallene 2.0–2.375 kg, 2.375–2.75 kg, 2.75–3.125 kg og 3.125–3.50 kg. Igjen svarer dette til arealet under histogrammet mellom 2.0 kg og 3.5 kg.

Nederst til venstre i Figur 1 har vi gitt et histogram til fødselsvektene av et tilfeldig utvalg på 2500 jenter. Her er klassebredden 0.25 kg. Vi merker at også nå er den relative frekvensen av fødselsvekter mellom 2.0 kg og 3.5 kg lik arealet under histogrammet mellom 2.0 kg og 3.5 kg.

Endelig ser vi på et histogram til fødselsvektene av 20 000 jenter. Dette er gitt nederst til høyre i Figur 1. Her bruker vi klassebredde 0.125 kg. Som i de andre tilfellene er den relative frekvensen av fødselsvekter mellom 2.0 kg og 3.5 kg lik arealet under histogrammet mellom 2.0 kg og 3.5 kg.

Vi ser fra histogrammene i Figur 1 at når vi øker antall fødselsvekter så får vi en mer nøyaktig oversikt over de relative frekvensene av ulike fødseksvekter, siden vi kan bruke mindre klassebredde når vi har mange observasjoner. Dessuten merker vi oss at histogrammene er laget slik at den relative frekvensen av fødselsvekter i et intervall er lik arealet under histogrammet over dette intervallet. Men kanskje det mest iøynefallende med plottene i Figur 1 er hvordan histogrammene ser ut til å nærme seg en underliggende "glatt" funksjon. Det er rimelig å anta at dersom vi kunne legge til stadig nye fødselsvekter så ville histogrammene komme nærmere og nærmere denne funksjonen. Men dette er det selvsagt ikke mulig å sjekke siden vi bare har et endelig antall vekter til rådighet. En statistiker vil likevel, som en modell, tenke seg at når antall fødselsvekter øker, vil histogrammene nærme seg en funksjon $f(x)$. Figur 2 viser denne funksjonen sammen med histogrammet av de 20 000 fødselsvektene. Funksjonen[2] $f(x)$ kalles *sannsynlighetstettheten* til den stokastiske variabelen X. Ved å erstatte histogrammene med sannsynlighetstettheten $f(x)$ går vi i en viss forstand til grensen og får fram et "histogram" med en "klassebredde"

[2]Funksjonen gitt på figuren er normalfordelingstettheten med $\mu = 3.50$ kg og $\sigma = 0.48$ kg; jf. avsnitt 4.3 i boka til Devore & Berk.

0 kg, basert på "uendelig mange" fødsler.

Når vi har mange fødselsvekter vil den relative frekvensen av vekter mellom 2.0 kg og 3.5 kg være nær sannsynligheten for at en jente skal ha en fødselsvekt i dette intervallet. Siden histogrammene våre vil være "nær" sannsynlighetstettheten $f(x)$ vil sannsynligheten for en fødselsvekt mellom 2.0 kg og 3.5 kg være lik arealet under sannsynlighetstettheten over intervallet fra 2.0 kg til 3.5 kg. Men dette arealet er gitt ved integralet av $f(x)$ over dette intervallet. Konklusjonen er derfor at sannsynligheten for at en nyfødt jente skal veie mellom 2.0 kg og 3.5 kg er gitt ved

$$\mathrm{P}(2.0 \leq X \leq 3.5) = \int_{2.0}^{3.5} f(x)\,dx.$$

Hvis vi er interessert i sannsynligheten for en fødselsvekt mellom a og b kan vi finne denne ved å bruke a og b som integrasjonsgrenser i steden for 2.0 og 3.5.

Dette forklarer hvorfor sannsynligheten for at en kontinuerlig stokastisk variabel X skal anta en verdi i et intervall er lik integralet av sannsynlighetstettheten over intervallet, slik det er definert i avsnitt 4.1 i boka til Devore & Berk.