KAPITTEL 7

Integrasjon

Integrasjon er et helt sentralt begrep i store deler av matematikken, samtidig som integralet har uttallige anvendelser i ulike praktiske situasjoner. Her skal vi ta for oss et par aspekter av integrasjonsbegrepet som ikke er dekket i *Kalkulus*. I seksjonene 7.1 og 7.2 skal vi se litt på hvordan datamaskiner kan brukes til å beregne integraler, både numerisk og symbolsk. Dette er viktig siden mange integraler bare kan beregnes numerisk, mens de som kan beregnes symbolsk ofte er så tidkrevende å regne ut for hånd at det er svært fordelaktig å la en datamaskin gjøre jobben. Til slutt skal vi i seksjon 7.3 se at integrasjon er et grunnleggende verktøy i sannsynlighetsregning.

7.1 Symbolsk integrasjon

I *Kalkulus*, som i alle andre bøker i grunnleggende, reell analyse, er det beskrevet en del teknikker og triks for å løse ubestemte integraler. Dette er ofte både komplisert og svært regnekrevende, og det er lett å få forståelse for sitatet av Viggo Brun som sier at *Derivasjon er et håndverk, men integrasjon er en kunst!*

Siden integrasjon ofte krever mye regning er det ikke så rart at en tidlig begynte å bruke datamaskiner for å beregne integraler, med metoder basert på de teknikkene vi kjenner fra håndregningen. Med en programmeringsomgivelse som kan håndtere funksjoner og symbolske beregninger (husk på hva vi skrev om objektorientering i seksjon 2.5) er det ikke så vanskelig å implementere mange av disse metodene, siden de er klart algoritmiske av natur. Utover på 1960-tallet fikk en på denne måten utviklet gode integrasjonsprogrammer, basert på de tradisjonelle integrasjonsteknikkene. På samme tid begynte en etterhvert å se etter helt andre måter å angripe integrasjonsproblemet på, og i 1968 publiserte R. Risch en overraskende rapport der han viste at det fins en algoritme som gir løsningen på ubestemte integraler.

Utgangspunktet er at vi har en funksjon f som vi ønsker å integrere, og en klasse $\mathbb{S}$ av funksjoner der vi ønsker å lete etter løsninger. Risch ga en algoritme som avgjør om f har en antiderivert i $\mathbb{S}$ eller ikke, og hvis det fins en antiderivert i $\mathbb{S}$ så vil algoritmen også finne denne funksjonen. Algoritmen er såpass komplisert at det bare er ganske nylig at de vanlige programsystemene for symbolsk matematikk, så som Maple og Mathematica,

har fått en rimelig fullstendig implementasjon av Risch-algoritmen.

Et grunnleggende problem med det å bruke datamaskinen til å beregne ubestemte integraler er at løsningene, om de eksisterer, ofte er så kompliserte at de er uinteressante. Selv om en løsning er komplisert kan det jo hende at den kan skrives på en ekvivalent form som er enkel, men da har vi kommet til et annet problem innen dataalgebra, nemlig forenkling av uttrykk. Dette problemet er enda vanskeligere enn integrasjonsproblemet, ikke minst fordi det er svært vanskelig å definere presist hva en forenkling av et uttrykk er.

7.2 Numerisk integrasjon

Selv om vi tar datamaskinen til hjelp er det mange ubestemte integraler som enten ikke har noen løsning som kan uttrykkes ved kjente funksjoner, eller så er løsningen så komplisert at den ikke har noen praktisk interesse. Dette betyr at det å beregne bestemte integraler bare i de enkleste tilfellene gjøres ved å finne en antiderivert og så sette inn integrasjonsgrensene. Når dette ikke fungerer er alternativet å bruke numerisk integrasjon. I *Kalkulus* er de to vanligste metodene for numerisk integrasjon beskrevet, trapesmetoden og Simpsons metode. Her skal vi se litt nærmere på hvordan disse kan implementeres effektivt på datamaskin. Dette vil illustrere noen viktige prinsipper ved numerisk programmering.

7.2.1 Implementasjon av trapesmetoden

Utgangspunktet for trapesmetoden er at vi skal finne det bestemte integralet fra a til b av den kontinuerlige funksjonen f. Som i definisjonen av integralet ved trappesummer deler vi intervallet $[a, b]$ opp i delintervaller ved hjelp av en partisjon gitt ved

$$a = x_0 < x_1 < x_2 < \cdots < x_{n-1} < x_n = b.$$

Men i steden for å tilnærme f med en funksjon som er konstant på hvert delintervall bruker vi nå en tilnærming som er en rett linje på hvert delintervall, og som på intervallet $[x_{i-1}, x_i]$ forbinder punktene $\big(x_{i-1}, f(x_{i-1})\big)$ og $\big(x_i, f(x_i)\big)$, se seksjon 8.7 i *Kalkulus*. Generelt kan delintervallene godt ha varierende bredde, men vi skal bare se på tilfellet der vi deler $[a, b]$ i n like deler, hver med bredde $\Delta x = h = (b - a)/n$. Da er $x_i = a + ih$ slik at tilnærmingen til det bestemte integralet er gitt ved

$$\int_a^b f(x)\,dx \approx \frac{h}{2}\Big(f(a) + f(b) + 2\sum_{i=1}^{n-1} f(a + ih)\Big). \tag{7.1}$$

Spørsmålet nå er bare om om vi kan stole på denne tilnærmingen? I *Kalkulus* er det gitt et feilestimat for trapesmetoden, men denne involverer den andrederiverte til f som ikke nødvendigvis er så enkel å beregne. Den vanlige teknikken består i å beregne en følge av tilnærminger med økende n og avtagende h. Denne følgen vil da konvergere mot verdien av integralet slik at når n blir stor nok kan vi bruke den aktuelle høyresiden i (7.1) som tilnærming til integralet. For å finne ut hvor stor n trenger å være er det vanlig å bruke samme teknikk som ved Newtons metode: vi stopper når to påfølgende verdier adskiller seg med mindre enn en gitt toleranse.

Dette gir oss den generelle oppskriften på det vi skal gjøre, men det er en del detaljer som må avklares. Aller først må vi bestemme oss for hvordan n skal økes fra gang til gang. I steden for å øke n med én hver gang, øker vi n slik at bredden av delintervallene halveres hver gang. Dette betyr at vi første gang bruker $h_0 = b - a$ og 2 punkter, neste gang er $h_1 = h_0/2 = (b-a)/2$ slik at vi får 3 punkter, deretter setter vi $h_2 = h_1/2 = (b-a)/4$ slik at vi får 5 punkter også videre. Etter m steg setter vi

$$h_m = h_{m-1}/2 = \frac{b-a}{2^m}$$

og bruker de $2^m + 1$ punktene

$$x_i = a + ih_m, \qquad \text{for } i = 0,\, 1,\, \ldots,\, 2^m.$$

Vi kan nå sette inn disse verdiene i (7.1) å regne ut tilnærmingen til integralet. Men for å gjøre dette på en effektiv måte må vi se litt nøyere på hva som foregår. Hvis vi kaller tilnærmingen til integralet med $2^m + 1$ punkter for T_m, så ser vi fra (7.1) at T_m er gitt ved

$$T_m = \frac{h_m}{2}\Big(f(a) + f(b) + 2\sum_{i=1}^{2^m-1} f(a + ih_m)\Big). \tag{7.2}$$

Den neste tilnærmingen T_{m+1} er gitt ved

$$T_{m+1} = \frac{h_{m+1}}{2}\Big(f(a) + f(b) + 2\sum_{i=1}^{2^{m+1}-1} f(a + ih_{m+1})\Big). \tag{7.3}$$

Hvis vi sammenligner T_m og T_{m+1} så ser vi at det er mye som er felles. Alle x_i'ene der vi beregner verdier av f i T_m er også med i uttrykket for T_{m+1}, men i tillegg får vi en ny x_i mellom hvert par av x_i'er som inngår i T_m (husk at h_{m+1} er halvparten av h_m). Dette betyr at i summen på høyre side av (7.3) er det bare de funksjonsverdiene som svarer til en odde i-verdi som er nye; alle de andre har vi beregnet før. Vi ser derfor at

$$T_{m+1} = \frac{1}{2}T_m + h_{m+1}\sum_{\substack{i=1\\ i \text{ odde}}}^{2^{m+1}-1} f(a + ih_{m+1}).$$

Denne siste summen kan vi skrive litt tydeligere. Siden vi bare skal ha odde verdier av i betyr det at det fins en j slik at $i = 2j - 1$, og når i skal variere fra 1 til $2^{m+1} - 1$ ser vi at j må variere fra 1 til 2^m. Vi har derfor

$$T_{m+1} = \frac{1}{2}T_m + h_{m+1}\sum_{j=1}^{2^m} f\big(a + (2j-1)h_{m+1}\big) \tag{7.4}$$

$$= \frac{1}{2}T_m + \frac{h_m}{2}\sum_{j=1}^{2^m} f\big(a - h_{m+1} + jh_m\big) \tag{7.5}$$

siden $h_m = 2h_{m+1}$. Basert på dette kan vi sette opp følgende kode for trapesmetoden.

Algoritme 7.1 (Trapesmetoden)**.** *La f være en funksjon som er kontinuerlig på et intervall $[a, b]$. Følgende kode vil beregne en tilnærming til integralet av f over $[a, b]$ ved hjelp av trapesmetoden. Den relative feilen vil vanligvis være omtrent* `eps` *så sant beregningene ikke stoppes av at antall iterasjoner overstiger det gitte heltallet* `nmax`*.*

```
int jmax=1, n=0, j;
double h=b-a, hg, T, Tg, xj, nyf, e;

T = 0.5*h*(f(a)+f(b));
while (n <= nmax & e > eps) {
  n = n + 1; nyf = 0.0;
  hg = h; h = 0.5*h;
  xj = a + h;
  for (j=1; j<= jmax; j++) {
    nyf = nyf + f(xj);
    xj = xj + hg;
  }
  Tg = T;
  T = 0.5*(Tg + hg*nyf);
  e = abs(Tg-T)/abs(T);
  jmax = 2*jmax;
}
```

Som tidligere har vi her brukt en Java-lignende syntaks, men hoppet over alle mulige prefiks og lignende krimskrams. Det beste er å legge koden inn i en metode som har `f`, `a`, `b`, `nmax` og `eps` som inngangsparametre og gir ut den endelige verdien av `T` (hvis vi ikke stopper fordi `n` blir større enn `nmax`) som estimat for integralet.

Som vi ser er det en del detaljer som må på plass for å få fram en rimelig presis kode. Vi begynner med å beregne en tilnærming til integralet der vi tilnærmer f med en rett linje på hele intervallet $[a, b]$ (dette svarer til tilnærmingen T_0 i diskusjonen før algoritme 7.1) slik at $h = b - a$, og lagrer denne i T. Når dette er gjort kan vi starte løkka og suksessivt halvere bredden på delintervallene. Inne i `while`-løkka begynner med å regne ut den nye verdien av h og lagre den gamle verdien av h i `hg`. Deretter summerer vi opp de funksjonsverdiene som er nye i den nye oppdelingen. Siden disse verdiene beregnes i annenhver x_i, ligger de med avstand lik den gamle intervallbredden `hg`, med start i `a+h`. Antall verdier er 1 første gang og dobles så hver gang; dette heltallet lagrer vi i `jmax`. Når verdiene er summert opp kan vi så beregne den nye tilnærmingen til integralet ved hjelp av formelen (7.5). Men før vi gjør det passer vi på å ta vare på den gamle tilnærmingen i `Tg`. Vi beregner deretter et estimat for den relative feilen fra de to siste tilnærmingene før vi oppdaterer `jmax`.

Tidligere har vi bekymret oss mye for avrundingsfeil. Når det gjelder numerisk integrasjon har vi gode nyheter i så henseende: avrundingsfeil er vanligvis ikke et problem. Den sentrale operasjonen er å summere opp verdier som for de vanligste funksjonene ikke varierer så veldig mye, det er derfor liten risiko for at vi må legge sammen to omtrent like tall med motsatt fortegn. Det som kan skape problemer med avrundingsfeil er selve

beregningen av $f(x)$, men det er et problem som ikke har noe med numerisk integrasjon å gjøre.

Vi har valgt å estimere den relative feilen, og ikke den asbolutte feilen, siden den, som vi har sett tidligere, er uavhengig av størrelsen på tallet vi beregner. Hvis vi ønsker å beregne integralet med 10 riktige siffer kan vi derfor bruke `eps` $= 0.5 \cdot 10^{-10}$. Legg merke til at vi strengt tatt bør ha med en test på om `T==0.0` før vi regner ut den relative feilen.

Det er viktig å huske på at det å estimere feilen slik vi gjør her ikke er idiotsikkert. Anta for eksempel at funksjonen vi skal integrere er

$$f(x) = 1 + (x-a)(x-c)(x-b)$$

der $c = (a+b)/2$. Da har f verdien 1 i alle de tre punktene a, c og b og det er lett å se at begge de to første estimatene for integralet blir $b-a$. Vi får derfor `e` lik `0.0` første gang vi kommer inn i det indre av while-løkka og vil derfor stoppe etter én gjennomgang med beskjed om at integralet er $b-a$, noe som åpenbart er feil. Dersom vi har mistanke om at slike funksjoner kan forekomme bør vi starte iterasjonene med m noe større enn 0 slik at sannsynligheten blir mindre for at vi åpner beregningene med å beregne 'spesielle' verdier av f.

7.2.2 Implementasjon av Simpsons metode

Trapesmetoden er basert på at vi tilnærmer f med en rett linje på hvert delintervall. Simpsons metode er litt mer raffinert og tar utgangspunkt i at vi tilnærmer f med en parabel på hvert delintervall, slik som forklart i seksjon 8.7 i *Kalkulus*. Men to punkter er ikke nok for å bestemme en parabel, vi bruker derfor også midtpunktet i hvert delintervall for å bestemme parabelen. I praksis betyr dette at vi deler opp intervallet en ekstra gang. Med Simpsons metode må vi derfor begynne med å dele opp $[a,b]$ i to delintervaller og tilnærme f med parabelen som går gjennom de tre punktene $\big(a, f(a)\big)$, $\big((a+b)/2, f((a+b)/2\big)$ og $\big(b, f(b)\big)$. Deretter halverer vi intervallene slik som for trapesmetoden.

Hvis vi gir tilnærmingen når $h = h_m = (b-a)/2^m$ navnet S_m, så har vi fra *Kalkulus* at

$$S_m = \frac{h_m}{3}\Big(f(a) + f(b) + 4\sum_{j=1}^{2^{m-1}} f\big(a + (2j-1)h_m\big) + 2\sum_{j=1}^{2^{m-1}-1} f(a+2jh_m)\Big) \qquad (7.6)$$

for $m = 1, 2, \ldots$. Estimatet S_m er altså basert på $2^m + 1$ funksjonsverdier. Vi har her spaltet opp summen ved å gruppere sammen x_i'er med odde og like verdi for i ($i = 2j-1$ og $i = 2j$), siden disse skal multipliseres med forskjellige konstanter, henholdsvis 4 og 2. Hvis vi nå halverer delintervallene og setter $h_{m+1} = h_m/2$ får vi at den neste tilnærmingen er gitt ved

$$S_{m+1} = \frac{h_{m+1}}{3}\Big(f(a) + f(b) + 4\sum_{j=1}^{2^m} f\big(a + (2j-1)h_{m+1}\big) + 2\sum_{j=1}^{2^m-1} f(a+2jh_{m+1})\Big). \qquad (7.7)$$

Funksjonsverdiene i den første summen er nye i den forstand at de ikke inngår i beregningene av S_m, mens de andre er gamle siden de også inngår i S_m. Vi legger også merke til at de 'nye' verdiene multipliseres med konstanten 4 når vi bruker dem i beregningen av S_{m+1}, mens de i senere beregninger alltid vil bli multiplisert med 2 siden de da vil være 'gamle'. Hvis vi lar r_{m+1} betegne summen av alle funksjonsverdier med odde faktor foran h_{m+1} (summen av de 'nye' funksjonsverdiene),

$$r_{m+1} = \sum_{j=1}^{2^m} f\big(a + (2j-1)h_{m+1}\big),$$

så ser vi at S_m og S_{m+1} er relatert gjennom

$$S_{m+1} = \frac{1}{2}S_m - \frac{h_m}{3}r_m + 4\frac{h_{m+1}}{3}r_{m+1}.$$

Siden $h_{m+1} = h_m/2$ kan denne formelen forenkles til

$$S_{m+1} = \frac{1}{2}S_m + \frac{h_m}{3}(2r_{m+1} - r_m). \tag{7.8}$$

På bakgrunn av dette kan vi sette opp en detaljert algoritme for Simpsons metode.

Algoritme 7.2 (Simpsons metode)**.** *La f være en funksjon som er kontinuerlig på et intervall $[a, b]$. Følgende kode vil beregne en tilnærming til integralet av f over $[a, b]$ ved hjelp av Simpsons metode. Den relative feilen vil vanligvis være omtrent* `eps` *så sant beregningene ikke stoppes av at antall iterasjoner overstiger det gitte heltallet* `nmax`*.*

```
int jmax=2, n=0, j;
double h, hg, S, Sg, xj, e, r, rg;

h=0.5*(b-a);
xj = a + h; r = f(xj);
S = h*(f(a)+4*r+f(b))/3;
while (n <= nmax & e > eps) {
  n = n + 1;
  rg = r; r = 0.0;
  hg = h; h = 0.5*h;
  xj = a + h;
  for (j=1; j<= jmax; j++) {
    r = r + f(xj);
    xj = xj + hg;
  }
  Sg = S;
  S = 0.5*Sg + hg*(2*r-rg)/3;
  e = abs(Sg-S)/abs(S);
  jmax = 2*jmax;
}
```

Kommentarene i forbindelse med trapesmetoden er også aktuelle her, men legg merke til at vi nå må starte med `jmax=2`. De observante vil kanskje synes at det ser litt skummelt ut med differansen `r-rg` i uttrykket for `S`, men husk at summen som definerer `r` er dobbelt så lang som summen som definerer `rg` så det skal mye til at `r` og `rg` er omtrent like store og dermed gir kansellering og tap av nøyaktighet.

7.2.3 Valg av metode

Med to metoder tilgjengelig for numerisk integrasjon er spørsmålet hvilken vi skal velge? Som vi vet fra *Kalkulus* er Simpsons metode mer nøyaktig enn trapesmetoden når funksjonen vi skal integrere har kontinuerlig fjerdederivert. Selv om denne metoden er litt mer regnekrevende faller derfor valget som regel på Simpsons metode. Unntaket er om funksjonen vi skal integrere ikke har så mange som 4 kontinuerlige deriverte; da er det som oftest bedre å bruke trapesmetoden.

Det bør også nevnes at det fins metoder som er enda mer nøyaktige enn Simpsons metode. Disse er basert på at vi tilnærmer f lokalt med polynomer av høyere grad enn 2. Vi kan for eksempel bruke 4 funksjonsverdier og tilnærme f med et tredjegrads polynom. Generelt kan vi bruke $k+1$ funksjonsverdier og tilnærme f med et polynom av grad k og få en metode som er svært nøyaktig for funksjoner som har $2k$ kontinuerlige deriverte. Rombergintegrasjon er en generell metode som starter med trapesmetoden og deretter på en enkel måte beregner tilnærminger basert på polynomer av stadig høyere grad, inntil det ikke er flere funksjonsverdier tilgjengelig.

7.3 Integrasjon og sannsynlighet

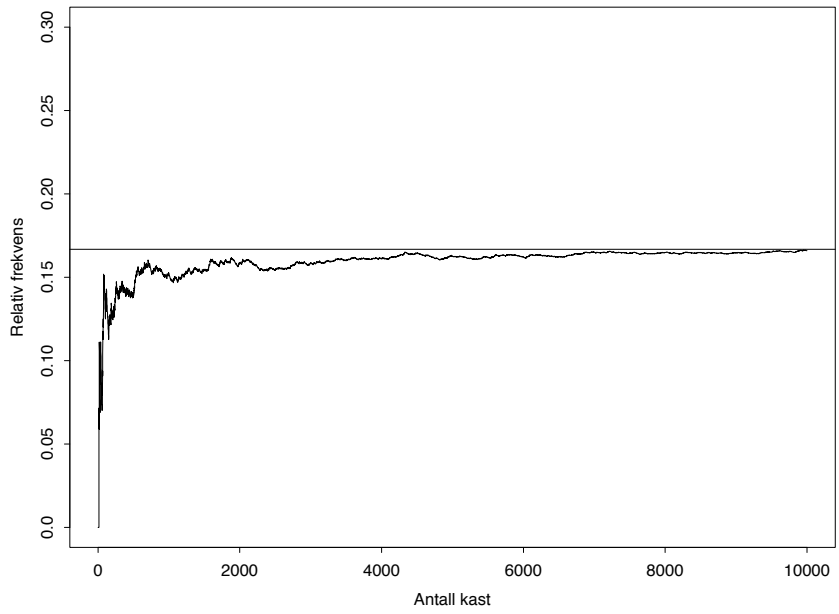
I denne seksjonen skal vi se hvordan integralregningen kan brukes til å beregne sannsynligheter, og vi skal også se nærmere på stokastisk simulering. Men aller først tar vi en rask repetisjon av sannsynlighetsbegrepet, og hva vi mener med en stokastisk (tilfeldig) variabel.

7.3.1 Hva er sannsynlighet?

På kalendere er tidspunktene for fullmåne ofte angitt, men hvordan er det mulig å vite dette på forhånd, lenge før året har begynt? Bakgrunnen er at astronomene ved hjelp av matematiske ligninger kan gi en nøyaktig beskrivelse av himmellegemenes bevegelser. Dermed kan de også regne ut nøyaktig når vi får fullmåne eller neste solformørkelse. Hendelser som fullmåne og solformørkelse kan med andre ord forutsies — de er *deterministiske*.

Når vi kaster en terning, vet vi ikke på forhånd hvor mange øyne vi får. Vi sier derfor at terningkast er et *stokastisk* (eller tilfeldig) forsøk. Et annet stokastisk forsøk er det når vi ser om et nyfødt barn er en gutt eller en jente. For heller ikke her vet vi resultatet på forhånd (hvis ikke barnets kjønn er blitt avklart i løpet av svangerskapet ved en kromosomtest eller en ultralydundersøkelse). Et kjennetegn på et stokastisk forsøk[1], er

[1] Merk at vi bruker ordet 'forsøk' i en videre betydning enn det som er vanlig ved laboratorieøvelser i biologi, fysikk og kjemi.



Figur 7.1. Relativ frekvens av seksere i 10000 terningkast.

altså at vi ikke på forhånd kan si hva resultatet vil bli, vi vet bare hvilke resultater som *kan* forekomme.

En terning har seks sider. På grunn av symmetrien til terningen har alle disse sidene like stor sjanse for å vende opp etter et terningkast. Sannsynligheten er derfor 1/6 for å få en sekser. Men hva betyr egentlig dette? Det betyr *ikke* at hvis vi kaster en terning 6 ganger, så vil vi få nøyaktig én sekser. Det *betyr* at hvis vi kaster mange ganger, vil vi få sekser i omtrent en sjettedel eller 16.7% av kastene.

For å illustrere dette har vi (ved hjelp av datamaskin) kastet en terning 10000 ganger. Etter N kast er den *relative frekvensen* av seksere gitt som antall seksere i de N første kastene dividert med N. Figur 7.1 viser den relative frekvensen som funksjon av N. Vi ser at variasjonen i den relative frekvensen er ganske stor til å begynne med, men etterhvert stabiliserer den seg nær $1/6 = 0.167$.

Hva er sannsynligheten for at et nyfødt barn er en jente? Noen vil kanskje tro at det blir født like mange gutter som jenter slik at sannsynligheten er 50%. Hvis du ser etter i Statistisk årbok (`www.ssb.no/aarbok/`) vil du se at dette ikke er tilfelle. Hvert år blir det født litt færre jenter enn gutter, og fordelingen mellom kjønnene er forholdsvis konstant fra år til år. For perioden 1994–1998 varierte andelen jenter hvert år mellom 48.3% og 48.8%. At variasjonen er så liten skyldes at det er mange fødsler — omtrent 60000 — hvert eneste år. I hele perioden 1994–1998 ble det født 299464 barn i Norge, og av disse var 145438 jenter. Den relative frekvensen av jentefødsler i femårsperioden var

derfor $145438/299464 = 0.486$.

Den relative frekvensen av seksere vil være omtrent $1/6$ når vi kaster en terning mange ganger, og den relative frekvensen av jenter blant alle nyfødte er hvert år omtrent 48.6%. Grunnlaget for begge disse utsagnene er at vi har gjentatt 'forsøkene' (kaste trening, observere kjønn til nyfødt barn) mange ganger. Denne muligheten for å gjøre mange gjentagelser av det samme 'forsøket' danner fundamentet for sannsynlighetsbegrepet, og følgende er en enkel og uformell definisjon[2] av sannsynlighet som er tilstrekkelig for våre formål:

Vi er interessert i en begivenhet (eller hendelse) A som er knyttet til et stokastisk forsøk som gjentas under like betingelser. Den relative frekvensen av begivenheten vil nærme seg en grenseverdi når forsøket gjentas mange ganger, og denne grenseverdien er sannsynligheten $\mathrm{P}(A)$ *for begivenheten A.*

Fra definisjonen av relativ frekvens ser vi at sannnsynligheten $\mathrm{P}(A)$ alltid vil være et tall i intervallet $[0, 1]$.

Rent språklig er ordet sannsynlighet knyttet til ett forsøk. Vi sier at sannsynligheten for å få sekser ved ett terningkast er $1/6$ og at sannsynligheten for at et nyfødt barn skal være en jente er 0.486. Det vi egentlig uttaler oss om er imidlertid ikke et enkelt kast eller en enkelt fødsel, men hva som vil skje i 'det lange løp'. I det lange løp vil 16.7% av terningkastene gi sekser og 48.6% av de nyfødte vil være jenter.

7.3.2 Stokastiske variable

Når vi kaster to terninger, spiller det ofte ingen rolle om vi får en firer og en femmer eller om vi får en treer og en sekser. Det som betyr noe, er at summen av antall øyne er ni. På lignende måte kan vi for en trebarns familie være interessert i hvor mange gutter og jenter det er i søskenflokken, uten at vi er interessert i kjønnet til den eldste, nesteldste eller yngste.

Det som kjennetegner disse to situasjonene, er at vi er interessert i en tallstørrelse knyttet til resultatet av et stokastisk forsøk. En slik tallstørrelse kalles en *stokastisk variabel* (eller tilfeldig variabel). Stokastiske variable betegnes gjerne med store bokstaver fra slutten av (det engelske) alfabetet, for eksempel X og Y. Ved kast med to terninger er $X =$ "sum antall øyne" en stokastisk variabel, og $Y =$ "antall gutter" er en stokastisk variabel for forsøket som består i å se hvilke(t) kjønn barna har i en tilfeldig valgt trebarns familie.

For en stokastisk variabel X er vi ofte interessert i funksjonen

$$p(x) = \mathrm{P}(X = x). \tag{7.9}$$

Denne funksjonen er definert for de x-verdiene variabelen kan anta og kalles *punktsannsynligheten* til X. Vi har $0 \leq p(x) \leq 1$ for alle x og $\sum_x p(x) = 1$ (siden X alltid må anta

[2] Vi kan ikke bruke grenseverdien av den relative frekvensen som en matematisk definisjon av sannsynlighet. Den grenseverdien vi snakker om her er empirisk (eksperimentell) og er ikke en grenseverdi i samme forstand som for en matematisk tallfølge som $\{1/n\}$. Når vi skal gi en presis matematisk definisjon av sannsynlighet, gjøres derfor dette ved å sette opp noen aksiomer som sannsynlighetsbegrepet skal tilfredsstille. Men motivasjonen for disse aksiomene kommer blant annet fra fortolkningen av sannsynlighet som relativ frekvens ved mange forsøk.

en av de mulige verdiene).

Hvis vi går tilbake til forsøkene våre, så ser vi at hvis $X =$ "sum antall øyne" ved kast med to terninger, er punktsannsynligheten gitt ved tabellen (sjekk selv)

x	2	3	4	5	6	7	8	9	10	11	12
$p(x)$	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

Hvis derimot $Y =$ "antall gutter" i en trebarns familie viser det seg at en rimelig modell er gitt ved den binomiske punktsannsynligheten

$$p(y) = \binom{3}{y} 0,514^y \, 0,486^{3-y}$$

for $y = 0$, 1, 2, 3.

De stokastiske variablene vi har sett på så langt, kan bare anta et endelig antall verdier og sies derfor å være *diskrete*. Men i mange sammenhenger har vi stokastiske variable som kan anta et kontinuerlig spekter av verdier — vi sier at vi har *kontinuerlige* stokastiske variable. Dette betyr at variabelen i prinsippet kan anta alle verdier i et intervall på tallinja (eventuelt på hele tallinja). Noen eksempler på dette er

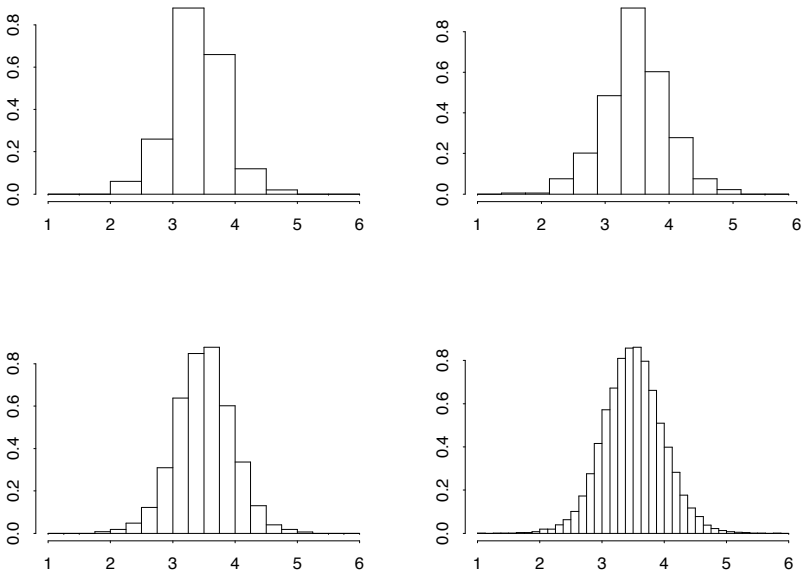
- vekten til en nyfødt jente,
- høyden til en norsk rekrutt,
- tiden mellom to oppringninger til en telefonsentral,
- endringen i en aksjekurs i løpet av en dag.

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 variable, men det viser seg å være mest hensiktsmessig å betrakte disse som kontinuerlige variable. Dette vil også være tilfellet for endringen i en aksjekurs, selv om denne ikke en gang 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.

7.3.3 Sannsynlighetstetthet – et motiverende eksempel

For en diskret stokastisk variabel X kan vi angi fordelingen til variabelen ved å oppgi sannsynligheten $\mathrm{P}(X = x)$ for alle mulige verdier av x i en tabell eller ved en formel, slik vi gjorde i eksemplene over. Hvis X er en kontinuerlig stokastisk variabel er $\mathrm{P}(X = x) = 0$ for alle x, se (7.11) nedenfor. Vi må derfor angi fordelingen til X på en annen måte.

For å se hvordan vi kan angi fordelingen til en kontinuerlig stokastisk variabel bruker vi variabelen $V =$ "vekt til nyfødt jente" som eksempel, og benytter data fra Medisinsk



Figur 7.2. 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 20000 vekter.

fødselsregister om fødselsvekter til jenter født i Norge i 1980. Vi vil bare se på "fullbårne" jenter, så vi begrenser oss til fødsler der 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 7.2. Vi bruker her klassebredde 0.5 kg, hvilket betyr at 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 det totale arealet under histogrammet mellom 2.0 kg og 3.5 kg.

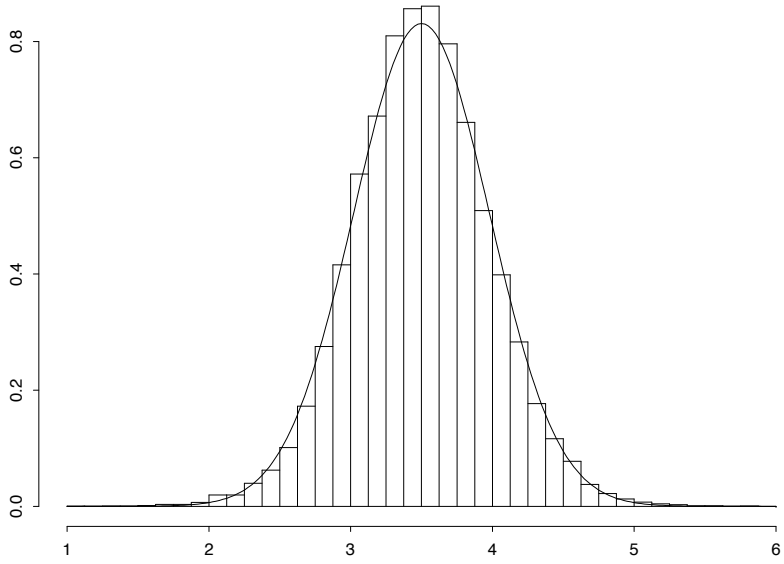
Vi ser så på et histogram til fødselsvektene av et tilfeldig utvalg av 500 jenter slik som vist øverst til høyre i Figur 7.2. Siden vi nå har flere fødselsvekter, reduserer vi klassebredden til 0.375 kg. Den relative frekvensen av fødselsvekter mellom 2.0 kg og 3.5 kg er nå summen av de relative frekvensene 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 7.2 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.

Vi ser endelig på et histogram til fødselsvektene av 20000 jenter med klassebredde 0.125 kg. Dette er gitt nederst til høyre i Figur 7.2. 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 at når vi øker antall fødselsvekter så får vi en mer nøyaktig oversikt over de relative hyppighetene 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 7.3 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 histogrammet komme nærmere og nærmere denne funksjonen, men dette er det selvsagt ikke mulig å sjekke siden vi bare har et endelig antall fødselsvekter til rådighet. En statistiker vil allikevel, som en modell, tenke seg at når antall fødselsvekter øker, vil histogrammene nærme seg en funksjon $f(v)$. Figur 7.3 viser denne funksjonen sammen med histogrammet av de 20000 fødselsvektene. Funksjonen $f(v)$ kalles *sannsynlighetstettheten* til den stokastiske variabelen V = "vekt til nyfødt jente". Ved å erstatte histogrammene med sannsynlighetstettheten $f(v)$ går vi i en viss forstand til grensen og får et histogramm med klassebredde på 0 kg, basert på uendelig mange fødsler, helt analogt med hvordan vi definerer integralet ved hjelp av trappesummer over stadig mindre intervaller.

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, i følge vår uformelle definisjon av sannsynlighet i avsnitt 7.3.1. Siden histogrammene våre vil være nær sannsynlighetstettheten $f(v)$ når vi har mange observasjoner, vil sannsyn-



Figur 7.3. Histogram av fødselsvekter for 20000 jenter med inntegnet sannsynlighetstetthet.

ligheten 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 vet vi er gitt ved integralet av $f(v)$ 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

$$\mathrm{P}(2.0 \leq V \leq 3.5) = \int_{2.0}^{3.5} f(v)\, dv.$$

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.

7.3.4 Sannsynlighetstettheter og kumulative fordelinger

I foregående avsnitt så vi at sannsynligheten for at den stokastiske variabelen $V =$ "vekt til nyfødt jente" skal anta en verdi mellom a og b er lik integralet av sannsynlighetstettheten $f(v)$ over dette intervallet. Vi vil nå se litt mer generelt på kontinuerlige stokastiske variable og lar X være en kontinuerlig stokastisk variabel med sannsynlighetstetthet $f(x)$. For at en funksjon f skal kunne kalles en sannsynlighetstetthet må den tilfredstille et par betingelser. Siden sannsynligheter aldri kan bli negative kan ikke $f(x)$ være negativ for noen x. Dessuten må vi ha $\int_{-\infty}^{\infty} f(x)dx = 1$ siden vi med full sikkerhet kan si at den stokastiske variabelen alltid vil anta en eller annen verdi på tallinjen. Som nevnt over er det slik hvis $a < b$ så er sannsynligheten for at X skal anta en verdi mellom

a og b gitt ved

$$\mathrm{P}(a \leq X \leq b) = \int_a^b f(x)dx. \tag{7.10}$$

Spesielt har vi at

$$\mathrm{P}(X = a) = \int_a^a f(x)dx = 0 \tag{7.11}$$

for ethvert tall a. Med andre ord er sannsynligheten 0 for at X skal anta en på forhånd angitt verdi a. Dette kan virke litt underlig, men ved nærmere ettertanke er det ikke så rart. Husk at den stokastiske variabelen angir resultatet av et tilfeldig 'forsøk' der resultatet er et reelt tall. Et reelt tall kan vi tenke på som et desimaltall med uendelig mange siffer til høyre for desimalkommaet, og det synes helt utenkelig at vi i et forsøk skulle kunne matche alle de uendelig mange sifrene i a.

Den *kumulative fordelingsfunksjonen* til X er gitt ved den antideriverte til sannsynlighettettheten f,

$$F(x) = \mathrm{P}(X \leq x) = \int_{-\infty}^x f(u)du. \tag{7.12}$$

Legg merke til at siden en sannsynlighetstetthet er ikke-negativ så vil den kumulative fordelingsfunksjonen alltid være en voksende funksjon. Vi kan bruke den kumulative fordelingsfunksjonen til å finne sannsynligheten for at X ligger i et intervall siden vi fra egenskaper ved integralet har relasjonen

$$\mathrm{P}(a \leq X \leq b) = F(b) - F(a).$$

Formelen (7.12) viser hvordan vi kan finne den kumulative fordelingsfunksjonen fra sannsynlighetstettheten. Ved analysens fundamentalteorem (teorem 8.3.3 i *Kalkulus*) har vi omvendt at $f(x) = F'(x)$.

Enhver ikke-negativ funksjon f_0 gir opphav til en sannsynlighetstetthet f ved formelen

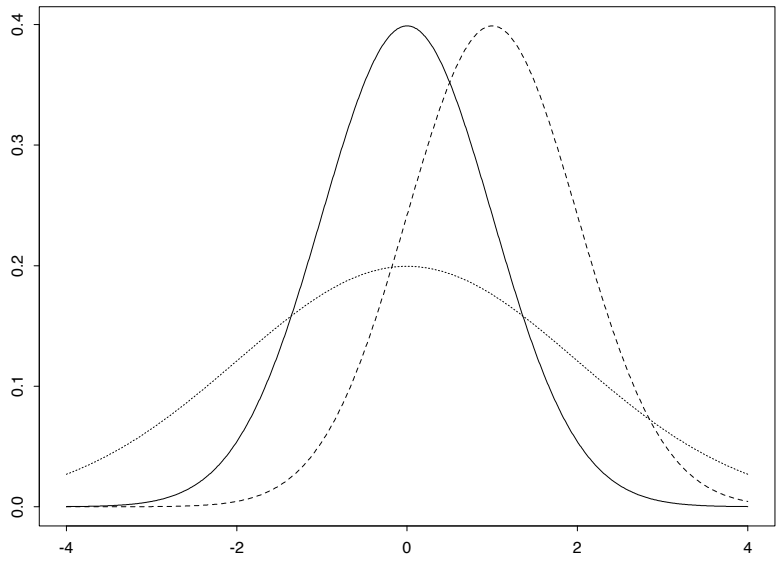
$$f(x) = \frac{f_0(x)}{\int_{-\infty}^{\infty} f_0(x)\,dx},$$

for vi ser at f, i tillegg til å være ikke-negativ slik som f_0, også har integral 1. Siden vi vet at det fins uendelig mange ikke-negative funksjoner har vi derfor et stort utvalg av sannsynlighetstettheter. På den annen side er det enkelte sannsynlighetstettheter som går igjen i mange ulike sammenhenger, og derfor er spesielt viktige. Vi skal se på tre av disse her.

Normalfordelingen Normalfordelingen spiller en sentral rolle i sannsynlighetsregningen. Den kalles også den gaussiske fordelingen etter Carl Friedrich Gauss som foreslo normalfordelingen som en modell for målefeil.

Vi sier at X er normalfordelt med forventning μ og standardavvik σ hvis sannsynlighetstettheten er gitt ved

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-(x-\mu)^2/(2\sigma^2)}. \tag{7.13}$$



Figur 7.4. Tre normalfordelingstettheter: (i) $\mu = 0$, $\sigma = 1$ (heltrukket linje); (ii) $\mu = 1$, $\sigma = 1$ (stiplet linje); (iii) $\mu = 0$, $\sigma = 2$ (prikket linje).

Figur 7.4 viser normalfordelingstettheten for tre valg av μ og σ. Legg merke til at μ gir plasseringen av toppunktet til $f(x)$, mens σ er et mål for hvor "bred" sannsynlighetstettheten er. Dette betyr at tall i nærheten av μ er de mest sannsynlige verdiene for X, mens σ forteller oss hvor stor variasjon vi kan forvente om vi genererer mange verdier fra fordelingen. Mer presist går det an å vise at hvis vi trekker mange verdier i henhold til normalfordelingen (7.13) så vil omtrent 2/3 av disse ligge innfor intervallet $[\mu - \sigma, \mu + \sigma]$. Normalfordelingen er viktig fordi mange stokastiske variable, så som målefeil, er normalfordelt. Dessuten er normalfordelingen nyttig for å tilnærme andre fordelinger.

Den sannsynlighetstettheten vi brukte i forrige avsnitt for fødselsvekten til en fullbåren jente er gitt ved (7.13) med $\mu = 3.50$ og $\sigma = 0.48$. Fødselsvekten til en jente er altså normalfordelt med forventning 3.50 kg og standardavvik 0.48 kg.

Fra (7.10) og (7.13) har vi

$$\mathrm{P}(a \leq X \leq b) = \int_a^b \frac{1}{\sigma\sqrt{2\pi}} e^{-(u-\mu)^2/(2\sigma^2)} du. \tag{7.14}$$

Med $a = 2.00$, $b = 3.50$, $\mu = 3.50$ og $\sigma = 0.48$, blir (7.14) sannsynligheten for at en jente skal ha en fødselsvekt mellom 2.0 kg og 3.5 kg. Det er ikke mulig å bestemme integralet (7.14) analytisk, så vi må bruke numeriske metoder for å beregne slike sannsynligheter[3], se oppgave 3.

[3]Det har i lang tid vært utarbeidet tabeller over den kumulative normalfordelingen med $\mu = 0$ og $\sigma = 1$ (standard normalfordelingen). Slike tabeller kan brukes til å bestemme integralet (7.14) numerisk.

Den uniforme fordelingen I avsnitt 4.3 så vi hvordan vi kan generere tilfeldige tall mellom 0 og 1. Men hva betyr det egentlig å trekke tilfeldige tall?

Siden alle relle tall mellom 0 og 1 (i prinsippet) er mulige verdier når vi trekker et tilfeldig tall, er et tilfeldig tall en kontinuerlig stokastisk variabel U som tar verdier i intervallet $(0, 1)$. Men da er sannsynligheten null for at U for eksempel er *nøyaktig* lik $1/2$ eller $\pi/4$ (husk (7.11)). At vi trekker et tilfeldig tall mellom 0 og 1 betyr derfor *ikke* at alle tall mellom 0 og 1 er like sannsynlige verdier. Det *betyr* at sannsynligheten er δ for at U skal ligge i et intervall av lengde δ, uansett hvor mellom 0 og 1 dette intervallet er plassert.

Konklusjonen er dermed at et tilfeldig tall mellom 0 og 1 svarer til en kontinuerlig stokastisk variabel U som har sannsynlighetstettheten

$$g(u) = \begin{cases} 1, & \text{hvis } 0 < u < 1, \\ 0, & \text{ellers.} \end{cases}$$

Vi sier at U er uniformt fordelt over $(0, 1)$. Fra definisjonen (7.12) ser vi at den kumulative fordelingsfunksjonen i dette tilfellet er gitt ved

$$G(u) = \begin{cases} 0, & \text{hvis } u \leq 0, \\ u, & \text{hvis } 0 < u < 1, \\ 1, & \text{hvis } u \geq 1. \end{cases} \tag{7.15}$$

Når vi trekker tilfeldige tall på en datamaskin regner vi vanligvis med flyttall slik at det bare er et endelig antall mulige tall som kan trekkes. Sannsynligheten for å få en bestemt verdi er derfor ikke 0 i dette tilfellet, så fordelingen er strengt tatt ikke kontinuerlig. Hvis vi for eksempel regner med fire gjeldende siffer, vil alle tall i intervallet $[0.49995, 0.50005)$ blir rundet av til 0.5000. Sannsynligheten for at U er 0.5000 når vi regner med fire gjeldende siffer er derfor lik $\mathrm{P}(0.49995 \leq U < 0.50005) = 0.0001$.

Eksponensialfordelingen En kontinuerlig stokastisk variabel T med sannsynlighetstetthet gitt ved

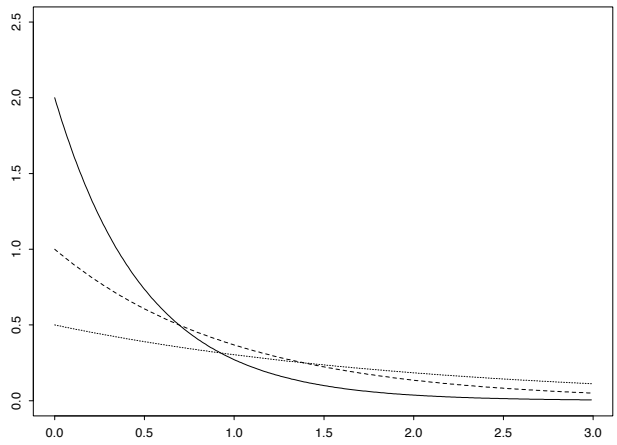
$$h(t) = \begin{cases} \lambda e^{-\lambda t} & \text{hvis } t > 0 \\ 0 & \text{hvis } t \leq 0 \end{cases}$$

sies å være eksponensialfordelt med parameter λ. Figur 7.5 viser dennne sannsynlighetstettheten for tre verdier av λ.

Eksponensialfordelingen brukes i studier av levetider for tekniske komponeneter. Den brukes også ved analyse av teletrafikk, for eksempel til å beskrive hvor lang tid det går mellom to oppringninger til en telefonsentral. I det siste tilfellet vil parameteren λ angi forventet antall oppringninger pr. tidsenhet.

7.3.5 Stokastisk simulering

I seksjon 4.3 så vi hvordan vi kan generere tilfeldige tall mellom 0 og 1. Mer presist var det vi gjorde å simulere uniformt fordelte stokastiske variable. Ofte vil vi være interessert



Figur 7.5. Tre forskjellige varianter av eksponensialfordelingen: $\lambda = 2$ (heltrukken linje), $\lambda = 1$ (grovstiplet linje), $\lambda = 0.5$ (finstiplet linje).

i å simulere stokastiske variable som har en annen fordeling. Hvis vi for eksempel ønsker å simulere trafikken til en telefonsentral, må vi generere eksponensialfordelte stokastiske variable, se oppgave 6.

Vi ønsker å generere en stokastisk variabel Y med en bestemt kumulativ fordeling $F(y)$ (for eksempel kan $F(y)$ være den kumulative eksponensialfordelingen). Vi antar at $F(y)$ er strengt voksende (bortsett fra muligens for verdier av y hvor $F(y) = 0$ eller $F(y) = 1$) slik at den omvendte funksjonen $F^{-1}(u)$ er definert for $0 < u < 1$.

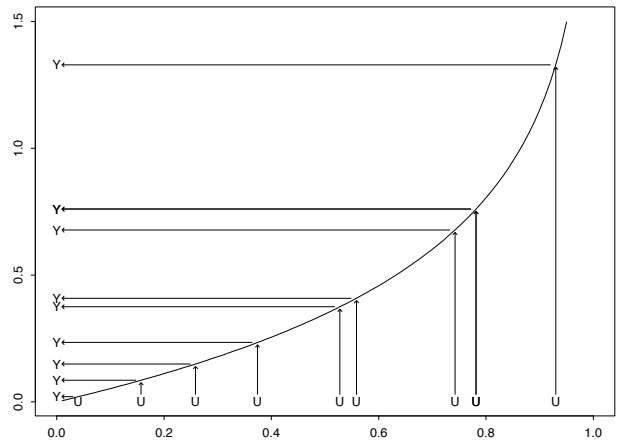
Vi antar at vi er i stand til å generere en stokastisk variabel U som er uniformt fordelt over intervallet $(0, 1)$. Vi kan da generere Y ved $Y = F^{-1}(U)$. (Merk at siden U er en stokastisk variabel, så vil også $Y = F^{-1}(U)$ være det.) For å se at dette stemmer må vi vise at Y har den ønskede kumulative fordelingen $F(y)$. Vi må altså vise at $\mathrm{P}(Y \leq y) = F(y)$. Men dette følger fra egenskaper for omvendte funksjoner,

$$\mathrm{P}(Y \leq y) = \mathrm{P}(F^{-1}(U) \leq y) = \mathrm{P}(U \leq F(y)).$$

Her er høyre side den kumulative fordelingen til U, regnet ut for $u = F(y)$. Dermed gir (7.15) at $\mathrm{P}\big(U \leq F(y)\big) = G\big(F(y)\big) = F(y)$. Altså er $\mathrm{P}(Y \leq y) = F(y)$, så Y har kumulativ fordeling $F(y)$ slik vi ønsket å vise.

For å generere kontinuerlige stokastiske variable med gitt fordeling, kan vi dermed bruke følgende framgangsmåte[4]: Først genererer vi tilfeldige tall mellom 0 og 1 slik det ble beskrevet i avsnitt 4.3 (eller med en annen metode). Deretter transformerer vi disse til den ønskede fordelingen ved hjelp av den omvendte funksjonen til den kumulative fordelingen. Figur 7.6 gir en illustrasjon av denne framgangsmåten.

[4]Denne måten å generere stokastiske variable på, fungerer fint for alle fordelinger hvor det er lett å finne den omvendte funksjonen til den kumulative fordelingen. Hvis dette ikke er tilfelle (slik det er for normalfordelingen) må en finne en tilnærmelse til den omvendte funksjonen eller benytte andre teknikker.



Figur 7.6. Simulering av 10 eksponensialfordelte variable med $\lambda = 2$.

Oppgaver

7.1 Programmer trapesmetoden og test den på integralet

$$\int_0^1 x^2\,dx = \frac{1}{3}. \tag{7.16}$$

7.2 Programmer Simpsons metode og test den på integralet (7.16).

7.3 Bruk (7.14) og numerisk integrasjon til å finne sannsynligheten for at en nyfødt jente skal veie mellom mellom 2.0 kg og 3.5 kg.

7.4 I denne oppgaven skal vi beregne den kumulative fordelingsfunksjonen $F(x)$ for normalfordelingen gitt ved sannsynlighetstettheten (7.13) når $\mu = 0$ og $\sigma = 1$ (standard normalfordeling).

a) Forklar hvorfor

$$F(x) = \frac{1}{2} + \int_0^x f(x)\,dx.$$

b) Bruk numerisk integrasjon med Simpsons metode til å bestemme den kumulative normalfordelingen $F(x)$ for $x = 0.5$, 1.0, 1.5 og 2.0. Sammenlign resultatene med det du får ved å bruke en lommeregner eller eksisterende tabeller.

c) Lag et plott av $F(x)$ på intervallet $[-3, 3]$.

7.5 En stokastisk variabel V er uniformt fordelt over (a, b) hvis den har sannsynlighetstetthet

$$f(v) = \begin{cases} 1/(b-a), & \text{hvis } a < v < b, \\ 0, & \text{ellers.} \end{cases}$$

a) Bestem den kumulative fordelingen $F(v)$ og den omvendte funksjonen til denne.

b) Hvordan kan du generere variable som er uniformt fordelt over (a, b)?

7.6 La T være eksponensialfordelt med parameter λ.

a) Bestem den kumulative fordelingen $F(t)$ og den omvendte funksjonen til denne.

b) Hvordan kan du generere eksponensialfordelte variable?

c) Skriv et program som genererer eksponensialfordelte stokastiske variable med parameter $\lambda = 2$. Ta utgangspunkt i generatoren (4.11) hvor a, c og M er gitt som i oppgave 4.6 eller en rutine i din programmeringsomgivelse som genererer uniformt fordelte tilfeldige tall i intervallet $(0, 1)$. Generer 10 eksponensialfordelte verdier.

   Den eksponensialfordelte variabelen med $\lambda = 2$ kan vi tenke oss svarer til tiden mellom to oppringninger til et sentalbord med trafikkintensitet $\lambda = 2$ samtaler per minutt. De 10 verdiene $(t_i)_{i=1}^{10}$ gir da tidsintervallet mellom 10 telefonsamtaler slik at det totale tidsintervallet vi ser på blir $t_1 + t_2 + \cdots + t_{10}$.