Definisjoner

Symboler

Kapittel 1

Statistikk og sannsynlighet?

Some people hate the very name of statistics, but I find them full of beauty and interest. Whenever they are not brutalized, but delicately handled by the higher methods, and are warily interpreted, their power of dealing with complicated phenomena is extraordinary. They are the only tools by which an opening can be cut through the formidable thicket of difficulties that bars the path of those who pursue the Science of Man.

F.Galton (1908)

1.1 Statistikk over alt

Hva er statistikk? Ordet *statistikk* brukes med mange ulike betydninger. Det brukes for faget statistikk. Det brukes også om et av de mest sentrale begrep i statistikkfaget.

Definisjon 1.1 (*Statistikk*) En statistikk er en funksjon av data.

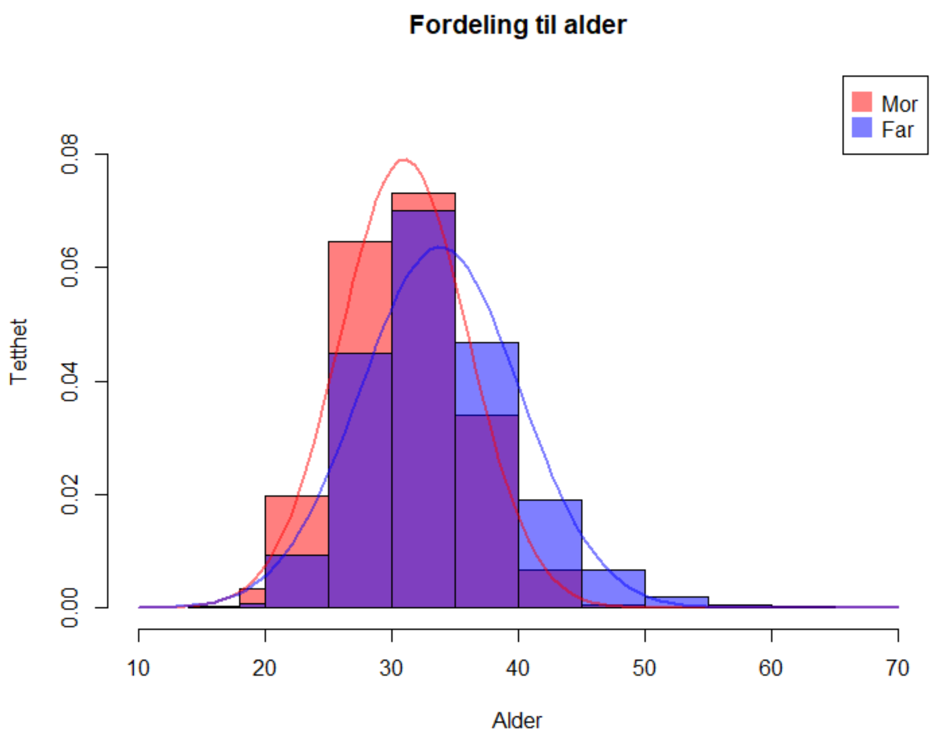
Observasjoner gir oss data. Mengden data som er tilgjengelig har eksplodert med utviklingen av datamaskinen. Ordet statistikk brukes også i betydningen *statistikkfaget* . Statistikk er vitenskapen hvor følgende spørsmål - og deres samspill - står sentralt.

- Hvilke data skal observeres?
- Hvordan lære fra data?
- Hva er gode data?

Som menneske vet du at det er en usikkerhet knyttet til alle observasjoner. Dermed er det en usikkerhet knyttet til alle data. Utgangspunktet i teoretisk statistikk er å modellere usikkerheten ved å assosiere sannsynligheter med data. Fordi en statistikk er en funksjon av data, så er statistikken selv også data med sine egne assosierte sannsynligheter. I teoretisk statistikk vil egenskapene til denne sannsynlighetsfordelingen avgjøre om det er en god eller dårlig statistikk. Dette er bokens tema, og vil bli forklart i mye større detalj etterhvert. I første omgang vil vi se på noen eksempel.

Alder	≤ 17	18-19	20-24	25-29	30-34	35-39	40-44	45-49	50-54	55-59	≥ 60	Uoppgitt
# Mor	56	366	5456	17748	20106	9336	1848	154 (45+)	0	0	0	0
# Far	21	96	2536	12189	19025	12701	5163	1809	565	185	82	698

Tabell 1.1: Antall fødsler i 2018 etter alder og kjønn.



Figur 1.1: Histogram og fordeling for alder for mor og far.

Eksempel 1.1 (Alder til mor og far ved fødselstidspunkt) Medisinsk fødselsregister inneholder opplysninger om fødsler i Norge. Tabell 1.1 gir antall fødsler i 2018 i ulike alderskategorier for kvinner og menn. Tallene i tabellen er en statistikk. Denne statistikken er en funksjon av de underliggende data for fødsler som er registrert av Medisinsk fødselsregister. Dette er da et eksempel på at en statistikk er en funksjon av data.

Figur 1.1 gir *normaliserte histogram* ut fra data i Tabell 1.1. De er laget ved opptelling i intervallene slik som i figuren på forsiden til boken, men høyden er normalisert slik at totalarelaet under kurven er lik 1. Tillegg A.2 viser hvordan grafene er laget.

Figur 1.1 inneholder også grafen til to kontinuerlige funksjoner som tilnærmer histogrammene. Funksjonene er også normalisert slik at arealet under grafen er lik 1. Histogrammene og grafene er funksjoner av de innsamlede data. Dette viser at en statistikk ikke behøver å være en tabell med tall. En statistikk kan med andre ord også være en graf.

Hva kan vi lære av disse statistikkene? Er de gode statistikker? Gode på hvilken måte? Svaret på disse spørsmålene avhenger av hva som ønskes undersøkt. *Grafene er statistikker* som er gode her ved at det kan være lettere å se sammenhenger geometrisk. De inneholder imidlertid mindre informasjon enn det som er gitt i Tabell 1.1. Grafene kan ikke brukes til å gjennvinne tallene i tabellen. På samme måte inneholder tabellen mindre informasjon enn det som er gitt i de underliggende data registrert i Medisinsk fødselsregister. Det er nettopp denne reduksjonen som kan gjøre en statistikk god. Grafene gir oss raskt et bilde av fordelingen til alder til mor og far. Dette bildet er ikke så lett å se direkte ved å stirre på alle tallene i Medisinsk fødselsregister

Hvor er så usikkerheten blitt av? Tallene gir at det ble registrert $N = 55070$ fødsler i Norge i 2018. Av disse var det $N_m = 56 + 366 + 5456 = 5878$ som hadde en mor yngre enn 25 år og $N_f = 21 + 96 + 2536 = 2653$ som hadde en far yngre enn 25 år. Disse registrerte tallene er det ingen usikkerhet i. Dersom en i stedet ser på antall faktiske fødsler, så er det et usikkert antall.

Helene ble født i 2001. Tallene gir $\hat{p}_m = N_m/N = 5878/55070 = 11\%$ som en approksimasjon til sannsynligheten p_m for at Helene blir mor før hun er 25 år. Hva betyr dette? Her er det mange og betydelige usikkerheter. En begrunnelse for påstandene kan gis ved å lage en statistisk modell for alder X til Helene ved hennes første fødsel. Parameteren vi vil si noe om er sannsynligheten p_m for at alderen $X < 25$ år. Symbolet X brukes her for en reell størrelse som det er assosiert usikkerhet med. Slike størrelser kalles *stokastiske variabler* .

Det antas at fødselstallene i 2018 er representative for Helene. Hver av de $N = 55070$ observasjonene sees som en tilfeldig trekning av et tall x. Resultatet er $x_1, x_2, \ldots, x_N$ tall som er registrert i Medisinsk fødselsregister. Hendelsen $x_i < 25$ år inntreffer i $\hat{p}_m = N_m/N = 5878/55070 = 11\%$ av tilfellene. Vi sier da at den observerte *relative hyppigheten* er 11%. Sannsynligheten p_m er, dersom modellen er rett, lik grensen for den tilsvarende relative hyppigheten når antall observasjoner går mot uendelig. Dette er på samme måte som at relativt antall Mynt i en serie myntkast kan gå mot 50% når antall myntkast går mot uendelig. I vårt tilfelle er antall observasjoner $N = 55070$, og vi forventer dermed at $\hat{p}_m$ er en god approksimasjon til p_m.

Histogrammet for Mor i Figur 1.1 er laget slik at arealet under kurven opp til alder 25 år er 11%. Arealet under den kontinuerlige kurven opp til alder 25 år er da også tilnærmet lik 11% fordi kurven approksimerer histogrammet. Kurven er grafen til en *normalfordeling* definert av funksjonen

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \mathrm{e}^{-\frac{(\mathrm{x}-\mu)^2}{2\sigma^2}} \tag{1.1}$$

hvor μ er *forventningsverdi* og σ er *standardavvik* .

En *statistisk modell* for alderen X er gitt ved å anta at sannsynlighetene assosiert med X er gitt av normalfordelingen i ligning (1.1). Sannsynligheten p_m for at alderen $X < x_m = 25$ år er da en *parameter* definert ved

$$p_m = \int_{-\infty}^{x_m} f(x)\, dx \tag{1.2}$$

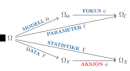
Hvordan er så den kontinuerlige kurven for Mor i Figur 1.1 laget? Problemet er at modellparametrene μ og σ er ukjente. Dette er på samme måte som at sannsynligheten p_m er ukjent. Dersom tallene $x_1, x_2, \ldots, x_N$ registrert i Medisinsk fødselsregister var kjente for oss, så kunne vi bruke *empirisk middel* $\overline{x}$ og *empirisk standardavvik* s som estimat for μ og σ. Empirisk middel $\overline{x}$ og *empirisk varians* s^2 er definert ved ligningene

$$\overline{x} = \frac{\sum_{i=1}^{N} x_i}{N} = \frac{x_1 + \cdots + x_N}{N} \tag{1.3}$$

$$s^2 = \frac{\sum_i (x_i - \overline{x})^2}{N-1} = \frac{(x_1 - \overline{x})^2 + \cdots + (x_N - \overline{x})^2}{N-1} \tag{1.4}$$

Tillegg A.2 gir en mulig metode basert på data som i Tabell 1.1. Resultatet er estimatene $\tilde{\mu}_m = 31.01$ og $\tilde{\sigma}_m = 5.05$. Disse kan brukes til å beregne et estimat $\tilde{p}_m$ ved integralet i ligning (1.2). Dette integralet kan ikke løses analytisk, men må beregnes ved numerisk integrasjon. Dette er gjort ved metoden i Tillegg A.1. Resultatet er estimatet $\tilde{p}_m = 12\%$.

Diskusjonen av dette eksempelet kan gjøres mye lenger. Det er mange åpne spørsmål. Er estimatet $\tilde{p}_m$ basert på en normalfordeling bedre enn estimatet $\hat{p}_m$ beregnet mer direkte? Hva menes med bedre? Disse, og flere spørsmål, er tema for videre diskusjon i boken. Noen av disse kan du selv arbeide med i oppgavene på side 9-.



Definisjon 1.1 gir en av betydningene til begrepet *statistikk*. En statistikk er en funksjon av data. Funksjonen og verdien til funksjonen er to forskjellige ting. Begrepet *statistikk* brukes om begge tolkningene. Denne konvensjonen skaper ofte forvirring for studenter og brukere av statistikk, men den er hensiktsmessig. Den samme konvensjonen brukes om følgende like sentrale begrep.

Definisjon 1.2 (*Parameter*) En parameter er en funksjon av modellen.

Det er gitt eksempel både eksplisitt og implisitt på både parametre og statistikker i Eksempel 1.1. Det overlates til deg å lage flere eksempel ut fra dette. Mange flere eksempel vil bli gitt senere.

1.2 Litt mengdeteori

Abstrakt mengdeteori står sentralt i Kolmogorovs formulering av sannsynlighetsteorien. I det følgende vil vi ikke gjøre noe forsøk på å bygge opp mengdeteorien aksiomatisk, men vil ta som gitt at leseren har en viss fortrolighet med manipulering med mengder. Hovedinnholdet her blir da å fastlegge noen notasjons- og språklige konvensjoner.

Paranteser av type $\{\cdots\}$ benyttes for å beskrive mengder. Noen eksempler på endelige mengder er gitt ved $\{a, b, c\}$ og $\{1, 2, \ldots, n\}$. Den tomme mengden $\emptyset$ er også endelig, den har ingen elementer. Vi har $\emptyset = \{\omega \,|\, \omega \neq \omega\}$. Symbolet $|$ leses i denne sammenhengen som *slik at* og brukes generelt til definisjon av mengder. Mengden av de naturlige tall betegnes med $\mathbb{N} = \{1, 2, 3, \ldots\}$, mengden av heltall betegnes med $\mathbf{Z} = \{\ldots, -1, 0, 1, \ldots\}$, og mengden av de reelle tall betegnes med $\mathbb{R} = \{x \,|\, -\infty < x < \infty\}$. De rasjonale tallene er $\mathbb{Q} = \{m/n \,|\, m, n \in \mathbf{Z}\}$.

En mengde I er *tellbar* dersom den er i en-en korrespondanse med en endelig mengde eller mengden av de naturlige tall. Dette betyr at vi kan liste opp alle elementene i I ved $i_1, i_2, \ldots$. Dette burde forklare navnet. Bortsett fra mengden $\mathbb{R}$ av reelle tall er dermed alle overnevnte mengder tellbare.

Vi skriver $a \in A$ dersom a er et element i mengden A. Symbolet $\in$ leser vi som *er et element i.* Symbolet $\notin$ leser vi tilsvarende som *er ikke et element i.* To eksempler er gitt ved $b \in \{a, b, c\}$ og $d \notin \{a, b, c\}$.

En mengde A er en delmengde av en mengde B dersom alle elementer i A også er elementer i B. Vi benytter notasjonen $A \subset B$ for dette. Vi har da at $a \in A \Rightarrow a \in B$. Tegnet $\subset$ leses som *er en delmengde av.* Tegnet $\Rightarrow$ leses som *medfører at.* Tre eksempler er gitt ved $\{1, 100, 14\} \subset \mathbb{N}$, $\emptyset \subset \mathbb{R}$ og $\{a, b, c\} \subset \{a, b, c\}$. Merk: Den vanligste metoden for å bevise at $A = B$ er gitt ved å vise at $A \subset B$ og $B \subset A$.

Anta at A_ι er en mengde for hvert element ι i en mengde I. Da er $\{A_\iota\}$, $\iota \in I$, en *indeksert familie* av mengder. Et eksempel er gitt ved $A_\iota = [\iota, \infty)$ for $\iota \in \mathbb{R}$, $\{A_\iota\} = \{[\iota, \infty)\}$, $\iota \in \mathbb{R}$.

Unionen og snittet av mengdene i familien $\{A_i\}$ er gitt ved henholdsvis

$$\bigcup_{i \in I} A_i = \{\omega \,|\, \exists i,\ \omega \in A_i\} \quad \text{og} \quad \bigcap_{i \in I} A_i = \{\omega \,|\, \forall i,\ \omega \in A_i\}.$$

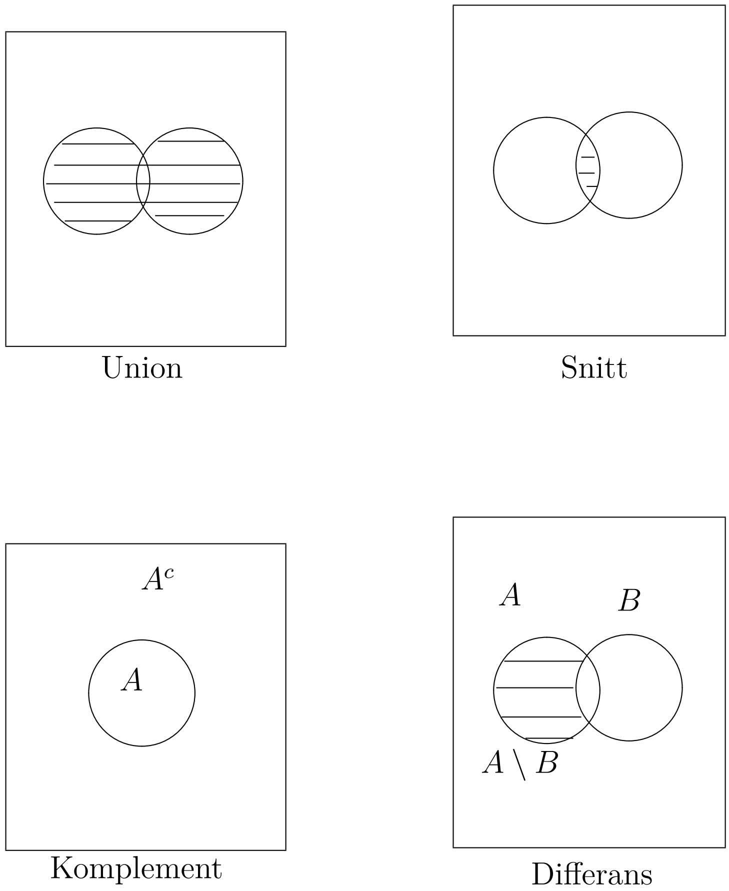
Tegnet $\exists$ leses som *det eksisterer en.* Tegnet $\forall$ leses som *for alle.* Unionen er dermed mengden av elementer i Ω som er inneholdt i minst en A_i og snittet er mengden av elementer som er inneholdt i alle A_i. Her har vi antatt at alle mengdene A_i er delmengder av en mengde Ω. To eksempler er gitt ved $\cup_{\iota \in \mathbb{R}} [\iota, \infty) = \mathbb{R}$ og $\cap_{\iota \in \mathbb{R}} [\iota, \infty) = \emptyset$. I noen sammenhenger er det underforstått hva indeksmengden I er og union og snitt skrives kortere på formen $\cup_i A_i$ eller $\cap_i A_i$.

Komplementet til A i Ω er gitt ved $A^c = \{\omega \,|\, \omega \notin A\}$. Komplementet A^c til A er mengden av alle elementer i Ω som ikke er elementer i A. F.eks. gir $\Omega = \{1, 2, 3, 4\}$, $A = \{1, 3\}$ at $A^c = \{2, 4\}$.

Symbolet $\setminus$ brukes for differansen mellom mengder. *Differansen* mellom A og B er mengden av alle elementer som er inneholdt i A men ikke i B. Formelt noteres dette som $A \setminus B = \{\omega \,|\, \omega \in A, \omega \notin B\}$. Det følger at $A \setminus B = A \cap B^c$ og at $A^c = \Omega \setminus A$. Et eksempel er gitt ved $\mathbb{R} \setminus \mathbb{N} = (-\infty, 0] \cup (\cup_{n=1}^{\infty} (n-1, n))$.

To mengder A og B sies å være *disjunkte* dersom de ikke har noen felles elementer, dvs. dersom $A \cap B = \emptyset$. F.eks. er mengdene gitt ved intervallene $(0, 1)$ og $(1, 2)$ disjunkte.

En union av to disjunkte mengder noteres som $A \uplus B$. Operasjonen $\uplus$ er kun definert for disjunkte mengder. Når vi skriver $A \uplus B$, så er det underforstått at mengdene A, B er disjunkte. Denne konvensjonen er sammenlignbar med at vi ofte skriver $f(x)$ for verdien til en funksjon f i et punkt x uten å eksplisitt skrive ned at vi antar at x er et element i



Figur 1.2: Venn-diagram.

definisjonsmengden til f. Notasjonen $\uplus_i A_i$ benyttes tilsvarende for unionen av en vilkårlig familie disjunkte mengder. Dette siste betyr at $i \neq j \Rightarrow [A_i \cap A_j = \emptyset]$. Et eksempel er gitt ved

$$\mathbb{R} = \biguplus_{t \in \mathbb{R}} \{t\}$$

Relasjoner mellom mengder og operasjoner på mengder som gjelder generelt vil også måtte gjelde for delmengder av et rektangel i planet. Dette gir muligheten til illustrasjon av sammenhenger som ellers kan virke uoversiktlige i Venn-diagram. Noen eksempler er gitt i figur 1.2.

1.3 Oppgaver

De neste oppgavene refererer til data gitt i Eksempel 1.1. Oppgaver merket med (**D**) anbefales løst med statistisk programvare på datamaskin. Eksempel på dette er gitt i Tillegg A.

1.1. Helene har ikke fått barn i 2020. Bruk denne opplysningen og Tabell 1.1 til å finne et forbedrede estimat for sannsynligheten for at hun får barn før hun er 25 år. Diskuter resultatet.

1.2. Diskuter antagelsen om at fødselstallene i 2018 er representative for Helene. Er 2018 relevant? Hva med tvillinger? Får alle barn? Bosted? Andre faktorer? Hva hvis det hadde vært oppgitt fødselstall for 1967 i stedet? (Hvis du er nysgjerrig kan du finne disse i statistikkbanken til Medisinsk fødselsregister.)

1.3. Halvor er født i 2002. Bruk dette og Tabell 1.1 til å finne en approksimasjon $\hat{p}_f$ til sannsynligheten p_f for at Halvor blir far før han er 25 år.

1.4 (**D**). Tallene i Tabell 1.1 for Far kan brukes til å begrunne estimatene $\tilde{\mu}_f = 33.84$ og $\tilde{\sigma}_f = 6.28$. Bruk dette, og en antagelse om normalfordeling, til å beregne et estimat $\tilde{p}_f$ til sannsynligheten p_f for at Halvor blir far før han er 25 år.

1.5 (**D**). Kan du estimere sannsynligheten for at Halvor blir far etter at han er 50 år? Hva med før han er 25 år, men etter at han er 15 år? Hva med Helene? Kommenter svarene og antagelsene.

1.6. Kan du finne og beskrive en metode til å estimere forventningsverdi μ_f og standardavvik σ_f ut fra tallene i Tabell 1.1?

1.7. La N være antall registrerte fødsler i 2018 og la M være antall faktiske fødsler. Er det en sammenheng mellom M og N? Hvorfor er det en usikkerhet knyttet til M? Kan du tenke deg et eksperiment som sier noe om sannsynligheten for at en faktisk fødsel blir registrert? Har alder til mor eller far noen betydning for denne sannsynligheten? Kan du tenke deg et eksperiment for å måle dette?

1.8. Gi 5 ulike eksempel på en parameter knyttet til en modell for Helene.

1.9. Gi 5 ulike eksempel på statistikker knyttet til Tabell 1.1 ved å definere funksjonene.

De neste oppgavene er omhandler elementær teori og notasjon for mengder.

1.10. La $\Omega = \mathbb{R}^2 = \{(x, y) \,|\, x, y \in \mathbb{R}\}$, dvs Ω er mengden av punkter i planet. Illustrer delmengdene $A = \{(x, y) \,|\, x = 5\}$, $B = \{(x, y) \,|\, x \geq 5\}$, $C = \{(x, y) \,|\, x \geq y^2\}$, $D = C \setminus B$, $E = \{(x, y) \,|\, x \geq y\}$, $F = C \cap E$, $G = F \cup A$, $H = \mathbb{N}^2 = \{(x, y) \,|\, x, y \in \mathbb{N}\}$. Er H tellbar?

1.11. La $\mathbb{Q}_+$ være mengden av alle positive rasjonale tall. Er $\mathbb{Q}_+$ tellbar? Er mengden av rasjonale tall tellbar? Svarene må begrunnes!

1.12. Kontroller om følgende regneregler gjelder: $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$, $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$, $(A \cup B) \cap (C \cup D) = (A \cap C) \cup (A \cap D) \cup (B \cap C) \cup (B \cap D)$, $(A \cup B)^c = A^c \cap B^c$, $(A \cap B)^c = A^c \cup B^c$.

Illustrer konklusjonene med Venn-diagram.

Kan regnereglene generaliseres til union/snitt av vilkårlige familier av mengder?

1.13. La $A_n = [0, 1/n]$. Finn $B = \cup_{n=1}^{10} A_n$, $C = \cap_{n=1}^{10} A_n$, $D = \cap_{n=1}^{\infty} A_n$, $E = \cup_{n=1}^{\infty} A_n$.

1.14. La A, B være delmengder av Ω. Vis at $\Omega = A \uplus A^c$.
Vis at $\Omega = (A \cup B)^c \uplus (A \cap B) \uplus (A \setminus B) \uplus (B \setminus A)$. Illustrer dette i et Venn-diagram.
Kan du generalisere dette til tilfellet hvor vi tar utgangspunkt i tre delmengder A, B, C?
Hva med tilfellet med n delmengder, eller et tellbart antall delmengder?

Tillegg A

R kode

Mange av figurene i boken er laget ved *R kode* som finnes i dette tillegget. Venables and Ripley (2002) gir en mye mer fullstendig introduksjon til R kommandoer. Koden som er vedlagt kan for eksempel kjøres i *RStudio* , i *ESS* pakken i Emacs, eller ved å klippe inn koden i et vindu i en nettleser som har åpnet `https://rdrr.io/snippets/` (testet i desember 2019). Noen av oppgavene i boken krever bruk av datamaskin, og eksemplene her kan være til hjelp.

A.1 Sannsynligheter fra fordelinger

```
# Sannsynlighet for X < nAge naar X er normalfordelt
nAge=25 # Alder
nMu=31.01 # Forventningsverdi
nSigma=5.05 # Standardavvik
nPT = pnorm(nAge, nMu, nSigma) # Kumulativ sannsynlighet
print(nPT)
```

A.2 Histogram og fordeling

```
# install.packages("MASS")
library(MASS)

aMorR = rep(c(17,18.5,22.5,27.5,32.5,37.5,42.5,47.5),
  c(56, 366, 5456, 17748, 20106, 9336, 1848, 154))

aFarR = rep(c(17,18.5,22.5,27.5,32.5,37.5,
  42.5,47.5,52.5,57.5,62.5),
  c(21, 96, 2536, 12189, 19025, 12701, 5163,1809, 565, 185, 82))

# Fordeling for mor # truehist gir ikke "add"?
hist(aMorR, breaks=c(14,18,20,25,30,35,40,45,50),
  xlim=c(10,70), ylim=c(0,0.09),
  col=rgb(1,0,0,0.5), xlab="Alder",
```

```
  ylab="Tetthet", main="Fordeling␣til␣alder" )

# Fordeling for far
hist(aFarR, breaks=c(14,18,20,25,30,35,40,45,50,55,60,65),
  xlim=c(10,70), col=rgb(0,0,1,0.5), add=T)

# Add legend
legend("topright", legend=c("Mor","Far"), col=c(rgb(1,0,0,0.5),
  rgb(0,0,1,0.5)), pt.cex=2, pch=15 )

# Plott av tetthet
xfit<-seq(10,70,length=100)
yfit<-dnorm(xfit,mean=mean(aMorR),sd=sd(aMorR))
lines(xfit, yfit, col=rgb(1,0,0,0.5), lwd=2)
yfit<-dnorm(xfit,mean=mean(aFarR),sd=sd(aFarR))
lines(xfit, yfit, col=rgb(0,0,1,0.5), lwd=2)
```