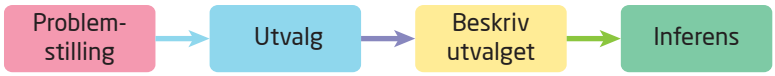
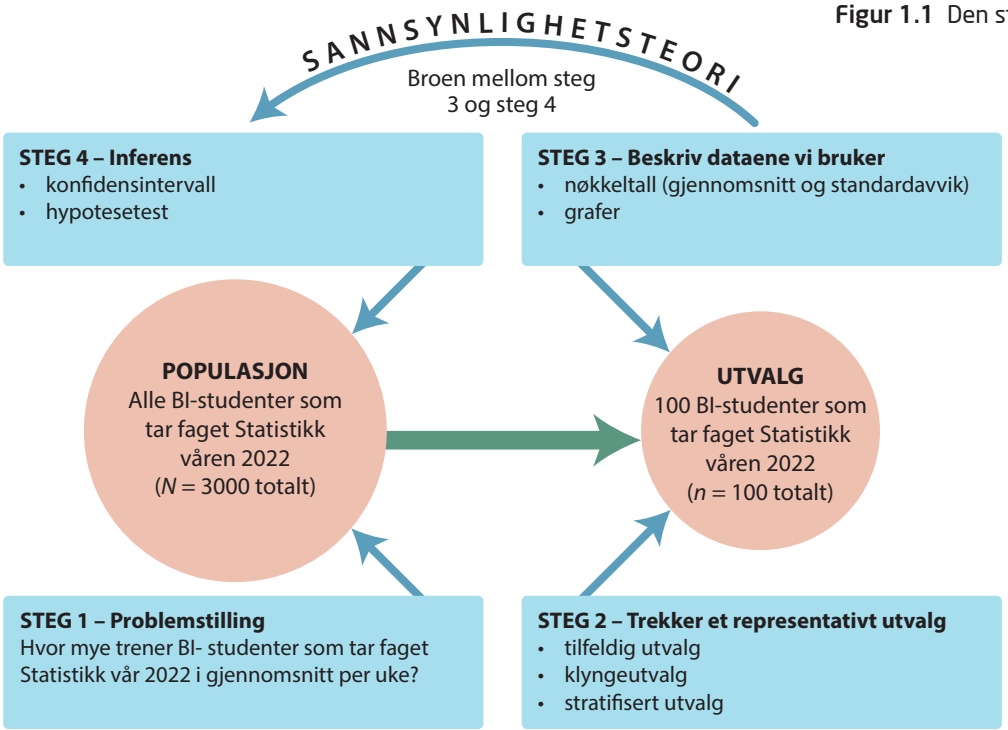
1.3.2 De fire stegene (elementene) i en statistisk analyse



Hovedmålet med denne boken er å gjøre deg i stand til å gjennomføre en statistisk analyse på egenhånd. Målet med en statistisk analyse er å få kunnskap om populasjonen selv om vi ikke har muligheten til å spørre alle i populasjonen. Den typen statistisk undersøkelse vi skal jobbe med i boken foregår i fire hovedsteg (studer figur 1.1, Den statistiske analysen, nøye når du leser om de fire stegene nedenfor):

Figur 1.1 Den statistiske analysen



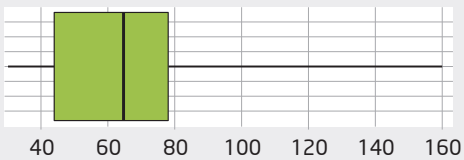
1) Vi definerer en problemstilling – se figuren over

I en statistisk analyse må vi først definere problemstillingen vi ønsker svar på. For eksempel ønsker vi å finne ut hvor mye BI-studenter, som tar faget Statistikk våren 2022, trener i gjennomsnitt per uke. MERK: Problemstillingen gjelder faget alltid populasjoen, som her er alle BI-studenter som tar ~~faget~~ Statistikk våren 2022.

I kapittel 3 vil vi se nøye på dette første elementet i en statistisk analyse.

eksemplet med budsjettstørrelse for tippeligaklubbene. Minimumsverdien er her 30, og maksimumsverdien er 159. Vi har dermed følgende fem tall: 30, 44, 64.5, 78, 159. Denne versjonen av boksplott er vist på figur ~~6.9~~ 5.9. Den midterste boksen går fra Q_1 til Q_3 og viser spennet for den midterste halvdelen av verdiene. Så går det to streker ut fra boksen, en ned fra Q_1 til minimumsverdien og en opp fra Q_3 til maksimumsverdien.

Figur 5.9 Vanlig boksplott for tippeligabudsjett



Utbryter: En observasjon som skiller seg ut fra flertallet. På engelsk: «outlier».

Den andre typen boksplott tar også med spesielle observasjoner som kalles *utbrytere*. En utbryter er en «vill» observasjon, en verdi som skiller seg ut ved å ligge langt over eller under storparten av verdiene. Slike verdier kan skyldes eksepsjonelle individer, målefeil eller feil inntasting av data. I noen tilfeller velger vi å fjerne utbryterobservasjonene, mens vi beholder dem i andre tilfeller. Uansett er det nyttig å skanne dataene for slike ekstreme observasjoner. En vanlig måte å definere utbrytere på er å si at det er verdier som ligger lavere enn en gitt avstand ned fra Q_1, eller høyere enn den samme avstanden opp fra Q_3. Hvor langt under nedre kvartil eller over øvre kvartil må observasjonen være for å kalles en utbryter? Jo, en vanlig regel er å bruke 1.5 ganger bredden på boksen i boksplottet. Denne bredden er $Q_3 - Q_1$ og kalles *interkvartilbredden*, forkortet *IK*. En observasjon som ligger høyere enn $Q_3 + 1.5 \cdot IK$, er med andre ord en utbryter. Og tilsvarende er en observasjon som ligger lavere enn $Q_1 - 1.5 \cdot IK$, en utbryter.

Interkvartilbredden: Avstanden mellom øvre og nedre kvartil: $IK = Q_3 - Q_1$

EKSEMPEL 5.2

For budsjettene i tippeligaen fant vi at nedre kvartil var $Q_1 = 44$ millioner kroner, og øvre kvartil var $Q_3 = 78$ millioner kroner. Interkvartilbredden er da $IK = 78 - 44 = 34$ millioner kroner. Vi regner ut 1.5 ganger IK som $1.5 \cdot 34 = 51$. For tippeligaen var $Q_3 = 78$ millioner kr, så utbrytergrensen er $78 + 51 = 129$ millioner kr i øvre ende og $44 - 51 = -7$ millioner kr i nedre ende. Det er altså ingen utbrytere for de lave budsjettverdiene. Men for høye budsjettverdier ser vi at det er en klubb (Rosenborg, RBK) som har et budsjett på 159 millioner kr. Rosenborg representerer altså en utbryter i forhold til budsjettall.

Partisjon av utfallsrommet: En oppdeling av utfallsrommet i hendelser som er disjunkte, og der unionen av hendelsene er hele utfallsrommet.

Videre, husk at komplementet til A skrives A^c og er alt i Ω som ikke er i A, derfor må $A \cup A^c = \Omega$; alt som er og ikke er i en mengde må jo være alt. Oppdelingen av Ω i to deler som ikke overlapper, her A og A^c, som til sammen er lik Ω kalles en partisjon av utfallsrommet Ω. Siden komplementet A^c er den motsatte hendelsen av A, altså «ikke A», må A og A^c være disjunkte hendelser. Merk at $A \cap \Omega = A$, siden $A \subseteq \Omega$ må hele A være inneholdt i Ω, og snittet $A \cap \Omega$ må derfor bli hele A. Vi kan derfor skrive A på følgende litt kompliserte måte (hvorfor vi gjør dette blir snart klart):

$$A = A \cap \Omega = A \cap (B \cup B^c) = (A \cap B) \cup (A \cap B^c)$$

Denne likheten kan også sees fra et Venn-diagrammet i figur 7.4.

Siden B og B^c er disjunkte, er $(A \cap B)$ og $(A \cap B^c)$ også disjunkte. Dette følger siden $A \cap B$ er i B, og $A \cap B^c$ er i B^c, og vi vet at B og B^c har ingenting felles. Se også Venn-diagrammet. Vi har derfor

$$\begin{aligned} P(A) = P(A \cap \Omega) = P(A \cap (B \cup B^c) &\overset{\text{(regel 4)}}{=} P\Big((A \cap B) \cup (A \cap B^c)\Big) \\ &= P(A \cap B) + P(A \cap B^c) \end{aligned}$$

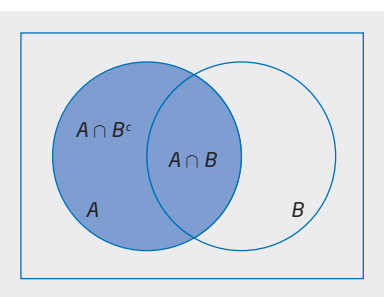
Loven om total sannsynlighet er en utvidelse av denne regelen. For å vise denne trenger vi å dele utfallsrommet i en *partisjon*, av n deler, det vil si i hendelsene $B_1, B_2, \ldots, B_n$ som til sammen dekker hele utfallsrommet, men som ikke overlapper hverandre.

EKSEMPEL 7.14

La A og B være disjunkte hendelser. Hva må C være for at samlingen A, B og C skal danne en partisjon?

Svar: $C = (A \cup B)^c$

Figur 7.4 Venn-diagram for $A = (A \cap B) \cup (A \cap B^c)$



Regel 9: Loven om total sannsynlighet

La A være en hendelse og $B_1, B_2, \ldots, B_n$, en partisjon:

$$P(A) = P(A \cap B_1) + P(A \cap B_2) + \ldots + P(A \cap B_n)$$

Se også figur 7.3.

Før vi går videre skal vi merke oss at vi får en annen variant av loven om total sannsynlighet dersom vi slår denne sammen med multiplikasjonsregelen $P(A \cap B) = P(A)P(A \mid B)$ (regel 7 side 132).

P(A) skal erstattes av $P(B_1)$

P(A) skal erstattes av $P(B_2)$

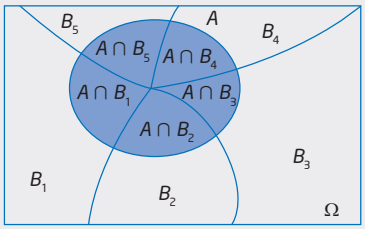
P(A) skal erstattes av $P(B_n)$

Regel 10: Loven om total sannsynlighet

Hvis A er en hendelse og B_1, B_2, …, B_n er en partisjon av utfallsrommet, gjelder:

$$P(A) = \cancel{P(A)}P(A|B_1) + \cancel{P(A)} \cdot P(A|B_2) + \cdots + \cancel{P(A)} \cdot P(A|B_n)$$

m total sannsynlighet illustrert med partisjon i fem deler



EKSEMPEL 7.15

Du jobber med en nettbutikk som selger elektronikk, og du ønsker å analysere sannsynligheten for at en kunde kjøper en ny type dyr bærbar datamaskin fra en ny produsent. For å gjøre det litt enkelt, antar vi at det kun finnes to typer kunder: (A) de som kjøper dyre produkter og (B) de som ikke gjør det. Anta videre at du vet at 60 % av tidligere kunder har kjøpt dyre produkter (kategori A) og 40 % av kundene har ikke gjort det (kategori B). Vi kan derfor anta at sannsynligheten for at en ny kunde vil kjøpe et dyrt produkt er 0.60, og for at en som ikke gjør det er den 0.40.

Produsenten av den dyre bærbare datamaskinen har gjort en markedsundersøkelse og funnet ut at for kunder som kjøper dyre produkter (kategori A) er sannsynligheten 0.70 for å kjøpe den dyre bærbare datamaskinen, og for kunder fra kategori B er sannsynligheten 0.30.

Vi kan nå bruke loven om total sannsynlighet til å finne ut sannsynligheten for at en tilfeldig kunde kjøper den dyre bærbare datamaskinen:

$$\begin{aligned} P(\text{kjøper dyr datamaskin}) &= P(\text{kjøper dyr datamaskin} \mid A) \cdot P(A) \\ &\quad + P(\text{kjøper dyr datamaskin} \mid B) \cdot P(B) \end{aligned}$$

Der $P(\text{kjøper dyr datamaskin} \mid A)$ er sannsynligheten for å kjøpe gitt at kunden er fra kategori A, den er 0.70, og $P(A)$ er sannsynligheten for at en tilfeldig kunde er fra kategori A som er 0.60, etc. Setter vi inn de resterende verdiene har vi:

$$\begin{aligned} P(\text{kjøper dyr datamaskin}) &= 0.70 \cdot 0.60 + 0.30 \cdot 0.40 \\ &= 0.42 + 0.12 \\ &= 0.54 \end{aligned}$$

Sannsynligheten for at en tilfeldig kunde kjøper den dyre bærbare datamaskin er derfor 0.54.

$$P(\text{vinne ved å beholde}) = P(\text{bil bak dør 1} \mid \text{åpner dør 3})$$

$$= \frac{P(\text{bil bak dør 1} \cap \text{åpner dør 3})}{P(\text{åpner dør 3})}$$

$$= \frac{1/6}{1/3+1/6} = \frac{1/6}{1/2} = \frac{2}{6} = \frac{1}{3}$$

7.6

Sannsynligheten for $P(\text{åpner dør 3})$ finner vi ved å lese av siste kolonne på figur ~~7.4~~. Det er to muligheter for at dette inntreffer. Enten er bilen bak dør 2 (med sannsynlighet $\frac{1}{3}$), og programlederen må derfor åpne dør 3 (med sannsynlighet 1). Dette har derfor sannsynligheten $\frac{1}{3} \cdot 1 = \frac{1}{3}$. Den andre muligheten er at bilen er bak dør 1 (med sannsynlighet $\frac{1}{3}$), og programlederen åpner dør 3 (med sannsynlighet $\frac{1}{2}$ fordi han velger tilfeldig mellom dørene 2 og 3). Dette har derfor sannsynligheten $\frac{1}{3} \cdot \frac{1}{2} = \frac{1}{6}$. De to disjunkte hendelsene har total sannsynlighet lik $1/3 + 1/6 = 2/6 + 1/6 = 3/6 = 1/2$. Av figur ~~7.4~~ ser vi videre at sannsynligheten er $\frac{1}{6}$ for hendelsen «bil bak dør 1» ∩ «åpner dør 3» = «bilen er bak dør 1, og programlederen åpner dør 3».

7.6

Vi analyserer nå dørbyttet. Vi leser av figur ~~7.4~~ at sannsynligheten er $\frac{1}{3}$ for hendelsen «bilen er bak dør 2, og programlederen åpner dør 3», og får

7.6

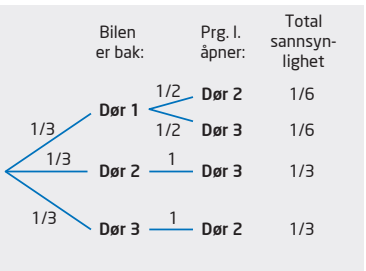
$$P(\text{vinne ved å bytte}) = P(\text{bil bak dør 2} \mid \text{åpner dør 3})$$

$$= \frac{P(\text{bil bak dør 2} \cap \text{åpner dør 3})}{P(\text{åpner dør 3})}$$

$$= \frac{1/3}{1/3+1/6} = \frac{1/3}{1/2} = \frac{2}{3}$$

Som nevnt er framgangsmåten helt tilsvarende hvis spilleren starter med å velge dør 2 eller 3. Argumentet viser at vi øker sjansen for å vinne dersom vi bytter dør.

Figur 7.6 Utfall og sannsynligheter hvis spilleren velger dør 1 fra start



7.11 Oppgaveløsninger

Løsning på oppgave 7.1

a. $P(A|B) = P(A \cap B)/P(B) = (1/5)/(1/2) = 0.4$

b. Bayes: $P(A|B) = P(B \cap A)/P(B) = (1/5)(1/5)/(1/2) = 0.08$

c. $P(A|B) = P(A \cap B)/P(B) = (P(A) + P(B) - P(A \cup B))/P(B)$
$= (1/5 + 1/2 - 1/5)/(1/2) = 1$

Løsning på oppgave 7.2

Sannsynligheten er $(12-3)(12-2)(12-1)/12^3 \approx 0.57$. Alternativ utregningsmetode:
$\frac{11}{12} \cdot \frac{10}{\cancel{11}} \cdot \frac{9}{\cancel{10}}$.

12 12

Løsning på oppgave 7.3

a. $P(A|B) = \dfrac{P(A)P(B|A)}{P(B)} = \dfrac{0.001 \cdot 0.99}{0.021} \approx 0.047$

siden

$P(B) = P(B|A)P(A) + P(B|A^c)P(A^c) = 0.99 \cdot 0.001 + (1-0.98) \cdot 0.999 \approx 0.021$

b. $P(A|B) = \dfrac{P(A)P(B|A)}{P(B)} = \dfrac{0.1 \cdot 0.99}{0.12} \approx 0.85$

siden

$P(B) = P(B|A)P(A) + P(B|A^c)P(A^c) = 0.99 \cdot 0.1 + (1-0.98) \cdot 0.90 \approx 0.12$

Løsning på oppgave 7.4

a. $P(B|A) = P(A \cap B)/P(A) = (1/5)/(1/4) = 0.8$

b. $P(C|A) = P(A \cap C)/P(A) = (P(A) + P(C) - P(A \cup C))/P(A) =$
$= (1/4 + 2/5 - 1/2)/(1/4) = 0.6$

Løsning på oppgave 7.5

La A = «ostehøvelen er lagd på den gamle maskinen», B = «ostehøvelen er lagd på den nye maskinen» og D = «ostehøvelen er defekt»

a. $P(D) = P(D|A)P(A) + P(D|B)P(B) = 0.80 \times 0.05 + 0.20 \times 0.01 = 0.042$

b. $P(A|D) = \dfrac{P(A \cap D)}{P(D)} = \dfrac{P(D \cap A)}{P(D)}$

$= \dfrac{P(D|A)P(A)}{P(D)} = \dfrac{0.80 \times 0.05}{0.042} \approx 0.95$

Løsning på oppgave 7.6

$P(\text{London} \cup \text{Stockholm}) = 0.1 + 0.7 - 0.01 = 0.79$

9.7 Oppgaveløsninger

$0.59^{59} \cdot 0.41^{41}$

Løsning på oppgave 9.1

a. $\mathrm{E}(X) = 0 \cdot 0.59 + 1 \cdot 0.41 = 0.41$

b. $\mathrm{Var}(X) = (0 - 0.41)^2 \cdot 0.59 + (1 - 0.41)^2 \cdot 0.41 = 0.2419$

c. Binomisk forsøksrekke med $n = 100$ og $p = 0.59$ gir $\binom{100}{59} \cdot$ ~~$0.5959 \cdot 0.4141$~~ $= 0.081$

Løsning på oppgave 9.2

a. $p(2) = 0.15$ siden summen av sannsynlighetene skal være 1.

b. $P(X \geq 2) = p(2) + p(3) = 0.15 + 0.05 = 0.2$.

c. $\mathrm{E}(X) = 0 \cdot 0.6 + 1 \cdot 0.2 + 2 \cdot 0.15 + 3 \cdot 0.05 = 0.65$. I det lange løp selger han i snitt 0.65 leiligheter per dag.

d. $\mathrm{Var}(x) = (0 - 0.65)^2 \cdot 0.6 + (1 - 0.65)^2 \cdot 0.2 + (2 - 0.65)^2 \cdot 0.15 + (3 - 0.65)^2 \cdot 0.05 = 0.8275$

e. Binomisk forsøksrekke med $n = 5$ og $p = 0.2$ gir $\binom{5}{3} \cdot 0.2^3 \cdot 0.8^2 = 0.0512$

Løsning på oppgave 9.3

a. $P(X = 9) = 0.046$.

b. $P(X > 1) = 1 - 0.301 = 0.699$

c. Det er klart mest sannsynlig å få de lave sifrene 1, 2, 3 og sifrene blir mindre og mindre sannsynlige jo større de er. Derfor forventer vi å få et lavt tall, som er en del mindre enn 5.

d. Store talls lov sier at gjennomsnittet vil stabilisere seg mot forventningen $\mathrm{E}(X) = 1 \cdot 0.301 + \cdots + 9 \cdot 0.05 = 0.046 = 3.441$

e. $\mathrm{Var}(x) = (1 - 3.441)^2 \cdot 0.301 + \cdots + (9 - 3.441)^2 \cdot 0.046 = 6.061$ så standardavviket vil stabilisere seg mot $\sigma = 2.462$.

Løsning på oppgave 9.4

a. $\mathrm{E}(X) = 0.22$ og $\mathrm{E}(Y) = 0.17$. Det forventes færrest hendelser på plattform B.

b. $\mathrm{E}(X + Y) = 0.22 + 0.17 = 0.39$.

c. $\mathrm{Var}(X) = (0 - 0.22)^2 \cdot 0.8 + (1 - 0.22)^2 \cdot 0.18 + (2 - 0.22)^2 \cdot 0.02 = 0.2116$ så standardavviket til X er $\sigma_X = 0.46$. Liknende får vi $\mathrm{Var}(Y) = 0.3211$ og $\sigma_Y = 0.567$. Det er størst variasjon i antall hendelser på plattform B.

d. Vi kan addere variansen pga. uavhengigheten. $\mathrm{Var}(X + Y) =$ ~~$0.2116 + 0.567$~~ og vi får $\sigma_{X+Y} = \sqrt{0.5327} = 0.7299$.

e. Hvis plattformene ligger nær hverandre så blir de utsatt for samme vær, og dette kan påvirke risikoen for hendelser på lik måte på begge plattformene. Da vil X og Y være avhengige. En annen skjult variabel kan være økonomien til operatøren. Dersom økonomien er stram, kan sikkerheten lide på begge plattformene, slik at antall hendelser samvarierer.

3^2

Løsning på oppgave 9.5

Vi bruker at $\text{Var}(X) = \sigma_X^2 = 4$ og at $\text{Var}(Y) = \sigma_Y^2 = 25$.

a. $\text{E}(Z) = 3 \cdot 10 = 30$ og $\text{Var}(Z) = \cancel{32} \cdot 4 = 36$.

b. $\text{E}(Z) = 2 \cdot (-2) + 1 = -3$ og $\text{Var}(Z) = 2^2 \cdot 25 = 100$.

c. $\text{E}(Z) = 2 \cdot 10 - 2 + 1 = 19$ og $\text{Var}(Z) = 2^2 \cdot 4 + 25 = 41$.

Løsning på oppgave 9.6

a. $P(X < Y) = P(X = 0, Y = 1) = 0.2$.

b.

x	0	1	2
$p(x)$	0.4	0.2	0.4

og

y	0	1
$p(y)$	0.6	0.4

c. $\text{E}(X) = 1$ og $\text{E}(Y) = 0.4$.

d. $\text{Var}(X) = 0.8$ og $\text{Var}(Y) = 0.24$

Dette resultatet er intuitivt, siden p er andelen av forsøkene som ender med suksess i det lange løp, og n er antall forsøk vi utfører, forventer vi å få $n \cdot p$ suksesser en binomisk forsøksrekke.

I en binomisk fordeling forventer vi $n \cdot p$ suksesser.

Variansen til en binomisk fordelt tilfeldig variabel X finner vi ved å bruke variansregelen på side 193 og addere alle variansene $p \cdot (1 - p)$, siden de binomiske forsøkene er uavhengige. Dette gir oss

$$\text{Var}(X) = \text{Var}(X_1) + \text{Var}(X_2) + \ldots + \text{Var}(X_n) = n \cdot p(1 - p)$$

Forventning og varians for binomisk fordeling

Hvis X er en binomisk fordelt variabel hvor $0 \leq p \leq 1$ er sannsynligheten for suksess i hvert av n forsøk, så er

$$\text{E}(X) = n \cdot p \quad \text{og} \quad \text{Var}(X) = n \cdot p \cdot (1 - p)$$

EKSEMPEL 10.8

For X binomisk fordelt med $n = 25$ og $p = 0.2$ så er:

$$\text{E}(X) = 25 \cdot 0.2 = 5 \quad \text{og} \quad \text{Var}(X) = 25 \cdot 0.2 \cdot (1 - 0.2) = 4$$

10.3 Hypergeometrisk fordeling

10.3

Den neste typen diskrete tilfeldige variabler vi skal studere, kan illustreres med krukka i figur 10.5. Vi trekker et visst antall baller fra krukka, *uten tilbakelegging*, og *teller opp* hvor mange av disse kulene som er røde. Den tilfeldige variabelen er altså

$X =$ «antall røde baller når vi trekker n baller uten tilbakelegging»

Sannsynlighetsfordelingen til X kan utledes ved å bruke gunstige delt på mulige. Hvert utvalg er jo like sannsynlig, så det er lett å telle opp antall mulige utvalg totalt. Men hvordan teller vi antall gunstige utvalg? Hvor mange utvalg inneholder det ønskede antallet røde baller? Jo, vi teller først antall utvalg med ønsket antall røde baller, og multipliserer så dette med antall utvalg med de resterende grønne ballene av det totale antallet grønne baller. Det er lettere å forklare dette med et eksempel:

Se side 107 for gunstige delt på mulige.

Forventning og varians for hypergeometrisk fordeling

Hvis X er en hypergeometrisk fordelt variabel, beskrevet av N, n, K og k som over, da er:

$$\mathrm{E}(X) = n \cdot \frac{K}{N} \quad \text{og} \quad \mathrm{Var}(X) = n \cdot \frac{K}{N} \cdot \frac{N-K}{N} \cdot \frac{N-n}{N-1}$$

EKSEMPEL 10.13

For X hypergeometrisk fordelt med $N = 52$, $n = 5$ og $K = 13$, så er:

$$\mathrm{E}(X) = 5 \cdot \frac{13}{52} = 1.25 \quad \text{og} \quad \mathrm{Var}(X) = 5 \cdot \frac{13}{52} \cdot \frac{52-13}{52} \cdot \frac{52-5}{52-1} \approx 0.86$$

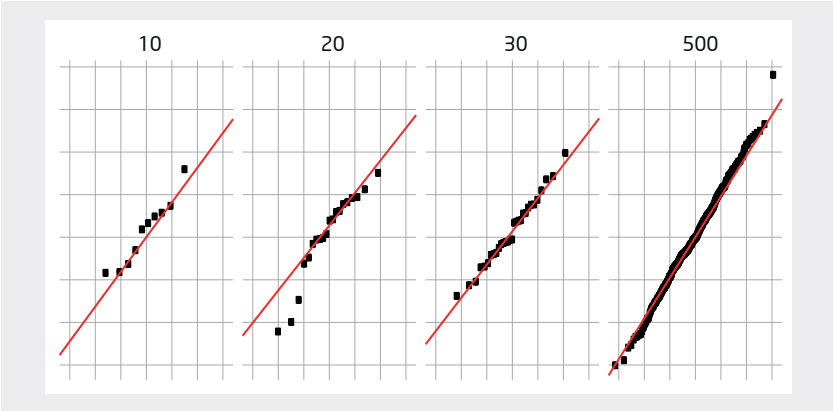
10.3.1 En anvendelse med hypergeometrisk fordeling (*)

Vi har hittil illustrert hypergeometrisk fordeling med enkle eksempler. Nå tar vi et mere praktisk eksempel.

Vi skal estimere hvor stor populasjon av abbor vi har i et tjern. Dette er et viktig, siden en lav bestand gjør at abboren er truet. Dersom det er høy sannsynlighet for at bestanden er mindre enn 100 individer, blir det fiskeforbud i en toårsperiode slik at bestanden kan bygge seg opp igjen. For å undersøke abborbestanden fanger viltforvalteren totalt 30 abbor i tjernet, som merkes og slippes ut igjen. En stund senere fanger viltforvalteren 30 abbor på nytt. Av de tretti hun fanget i andre runde, har syv merke fra første runde.

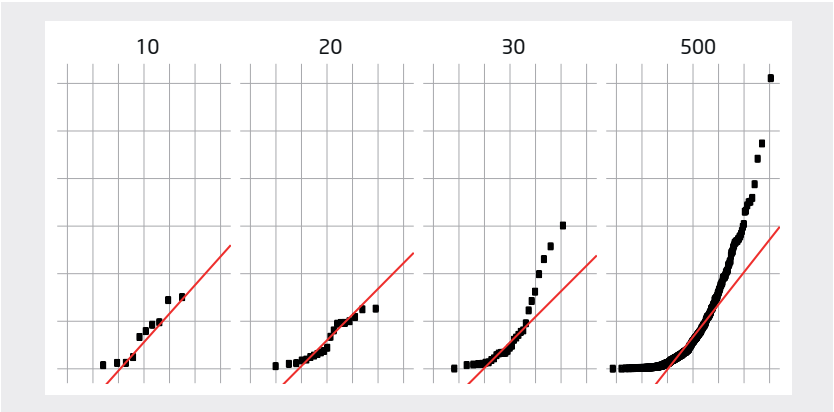
Vi kan nå bruke hypergeometrisk fordeling til å anslå den totale bestanden i tjernet. Den totale bestanden, som er ukjent, er tallet N i den hypergeometriske fordeling. Vi ønsker altså å anslå størrelsen til N. Vi kjenner antall gunstige, $K = 30$ (det er fisken vi har merket), $n = 30$ (antall fisk fanget i andre runde) og $k = 7$, som er antall fisk merket av de totalt $n = 30$ som ble fanget i andre runde. Vi regner ut formelen for hypergeometrisk fordeling for ulike verdier av N (med $K = 30$, $n = 30$ og $k = 7$). Anslaget vårt på den totale bestandsstørrelsen er den verdien N som gir høyest sannsynlighet for å gjenfange syv av tretti fisk. Vi bruker datamaskin til å regne ut formelen for mange verdier av N. Resultatet er vist på figur ~~10.6~~. Det er $N = 128$ som gir høyest sannsynlighet for å gjenfange 7 av 30, og dette er vårt beste anslag for N basert på informasjonen vi har tilgjengelig. Ifølge dette estimatet trenger vi derfor ikke innføre fiskeforbud.

10.4



Figur 11.11 QQ-plott for normalfordelte observasjoner. De fire diagrammene viser tilfeldige utvalg med størrelsene n = 10, 20, 30 og 500.

Vi kan sammenlikne de normalfordelte observasjonene fra figur 11.11 med QQ-plott der utvalgene er trukket fra en fordeling som *ikke* er normalfordelt. Se figur 11.12.



Figur 11.12 QQ-plott for observasjoner som ikke er normalfordelte. De fire panelene viser tilfeldige utvalg med størrelsene n = 10, 20, 30 og 500.

I de to største utvalgene ser vi klart at QQ-plottet ikke er lineært. For en utvalgsstørrelse på $n = 500$ er det hevet over enhver tvil at dataene ikke kan være nær normalfordelte. Men for små utvalg, $n = 10$ og $n = 20$, er det ikke store forskjellen mellom figurene ~~3 og 4~~. Dette innebærer at for svært små utvalg, la oss si med $n < 30$, er det vanskelig å avgjøre normalfordeling ved hjelp av QQ-plott.

16.5 Oppsummering av begreper og formler

- Vi antar vi trekker et tilfeldig utvalg fra en normalfordelt variabel med utvalgsstørrelse n, og σ estimeres med utvalgsstandardavviket s. Metodene i kapittel 14 og 15 kan da benyttes, med den modifikasjonen at

$$t = \frac{\bar{x} - \mu_0}{s / \sqrt{n}}$$

er t-fordelt med $n - 1$ frihetsgrader, hvis $n \geq 30$, eller når $n < 30$ og utvalgsfordelingen (det vil si et histogram av utvalget) ikke er helt uforenlig med en normalfordeling.
- t-fordelingen har en tetthetsfunksjon som ser ganske lik ut som tetthetsfunksjonen til den standard normalfordelte variabelen, men har litt tykkere haler.
- Hvis n er stor nok (over 100) er det svært små forskjeller på om vi bruker t-metodene eller antar at t er standard normalfordelt.
- Formelen for konfidensintervall for μ med nivå c er

$$\bar{x} \pm t^* \frac{s}{\sqrt{n}}$$

der t^* er tallet slik at $P(t \geq t^*) = \frac{1-c}{2}$, som kan finnes fra tabell C.
- Hypotesetester følger samme oppskrift som i kapittel 15, men t-fordelingen erstatter Z. Vi har:

Ikke fotskrift her

Oppskrift for hypotesetester når $H_0 : \mu = \mu_0$

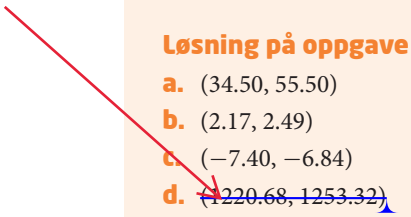
- Regn ut $T_{obs} = \frac{\bar{x}_{obs} - \mu_0}{\left(\frac{s}{\sqrt{n}}\right)}$.
- Når $H_A : \mu \neq \mu_0$: p-verdien er $2 \cdot P(t < -|T_{obs}|)$.
- Når $H_A : \mu > \mu_0$: p-verdien er $P(t > T_{obs})$.
- Når $H_A : \mu < \mu_0$: p-verdien $P(t < T_{obs})$.

I alle tilfellene er t i sannsynlighetene t-fordelt med $n - 1$ frihetsgrader, som kan slås opp i tabell C eller finnes via kalkulator.

erstatt T_{obs} med t

16.7 Oppgaveløsninger

(1224.1,1249.89)



Løsning på oppgave 16.1

a. $t^\star_{0.05}\,(23) = 1.714$

b. $t^\star_{0.01}\,(3) = 4.541$

c. $t^\star_{0.1}\,(700) \simeq t^\star_{0.1}\,(500) = 1.283$

Løsning på oppgave 16.2

a. 0.862

b. 0.968

c. 0.531

Løsning på oppgave 16.3

a. (34.50, 55.50)

b. (2.17, 2.49)

c. $(-7.40, -6.84)$

d. ~~(1220.68, 1253.32)~~

Løsning på oppgave 16.4

a. $\bar{x} = (101.234 + 99.766) / 2 = 100.5$

b. $(101.234 - 99.766) / 2 = 0.734$

c. $t^\star_{\alpha/2} \cdot s / \sqrt{n} = 0.734$, Det vil si at $t^\star_{\alpha/2} = 0.734 \cdot \sqrt{22} / 2.0 = 1.721$. Hvis vi ser i tabell C (Vedlegg) for 21 frihetsgrader, finner vi at det tilsvarer $t^\star_{0.05}$. Siden halearealet er 5 %, er konfidensnivået 90 %.

Løsning på oppgave 16.5

a. $H_0 : \mu = 50$ mot $H_A : \mu > 50$

b. $t = (51.3 - 50) / (14.2 / \sqrt{287}) = 1.551$

c. $p = 0.061$, $t^\star_{0.05}\,(200) = 1.653$

d. Ikke nok støtte for H_A. Behold H_0.

Løsning på oppgave 16.6

a. $H_0 : \mu = 500$ mot $H_A : \mu \neq 500$

b. $t = (492.429 - 500) / (7.390 / \sqrt{7}) = -2.711$

c. $t^\star_{0.025}\,(6) = 2.447$, $p = 2 \cdot 0.0175 = 0.035$

d. Nei, vi forkaster H_0.

17.4 Oppsummering av begreper og formler

- En andel kan også skrives som et gjennomsnitt. Derfor vet vi mye om hvordan vi skal gjøre inferens med andeler allerede.
- Populasjonsandel er p, utvalgsandel er $\hat{p}$. Fra sentralgrenseteoremet er $\hat{p}$ cirka fordelt som $N\left(p, \sigma / \sqrt{n}\right)$ der man kan regne ut at $\sigma = \sqrt{p(1-p)}$.
- Når vi har minst fem suksesser og fem fiaskoer kan vi finne vi et konfidensintervall for p med nivå α med formelen $\hat{p} \pm z^{\star} SE(\hat{p})$ der $SE\left(\hat{p}\right) = \sqrt{\dfrac{\hat{p}(1-\hat{p})}{n}}$ og $z^{\star}$ er slik at $P\left(Z < z^{\star}\right) = \dfrac{\alpha}{2}$. For det vanlige nivået $\alpha = 5\ \%$ er $z^{\star} = 1.96$.
- For hypotesetester om p med $H_0 : p = p_0$, bruker vi at σ er kjent til å være $\sigma = \sqrt{p(1-p)} = \sqrt{p_0(1-p_0)}$ når H_0 er sann. Vi bruker derfor test-observatoren

$$T_{obs} = \frac{\hat{p} - p_0}{\sqrt{\dfrac{p_0(1-p_0)}{n}}}$$

- Oppskrift for hypotesetester når $H_0 : p = p_0$ når n er så stor at både $np_0 \geq 5$ og $n(1-p_0) \geq 5$:
  - Regn ut T_{obs} som over
  - Når $H_A : \mu \neq \mu_0$: p-verdien er $2 \cdot P\left(Z < -\left|T_{obs}\right|\right)$.
  - Når $H_A : \mu > \mu_0$: p-verdien er $P\left(Z > T_{obs}\right)$.
  - Når $H_A : \mu < \mu_0$: p-verdien $P\left(Z < T_{obs}\right)$.

  I alle tilfellene brukes $Z \sim N(0,1)$, og sannsynlighetene kan finnes

erstatt T_{obs} med t

18.4 Oppsummering av begreper og formler

- Uavhengige utvalg: To utvalg A og B der observasjonene i A ikke danner naturlige par med observasjonene i B
- Relaterte utvalg: To utvalg A og B der hver observasjon i utvalg A danner et par med en observasjon i utvalg B. To relaterte utvalg har derfor samme størrelse.
- 95 % konfidensintervall for $\mu_1 - \mu_{2:}$

$$\bar{x}_1 - \bar{x}_2 \pm t^*_{0.025} \cdot SE_{\bar{x}_1 - \bar{x}_2}$$

Der $SE_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$

- For å teste $H_0 : \mu_1 = \mu_2$ mot $H_A : \mu_1 \neq \mu_2$ bruker vi testobservatoren

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

- For differansen i andeler, $p_1 - p_2$, i to utvalg er 95 % konfidensintervallet

$$\hat{p}_1 - \hat{p}_2 \pm 1.96 \cdot SE_{\hat{p}_1 - \hat{p}_2}$$

- Der $SE_{\hat{p}_1 - \hat{p}_2} = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$

bruker vi (ikke nytt kulepunkt)

- For å teste $H_0 : p_1 = p_2$
- ~~Bruker~~ vi testobservatoren $z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})(\frac{1}{n_1} + \frac{1}{n_2})}}$

der $\hat{p}$ er den sammenslåtte andelen

Kommaet i $P(X = x, Y = y)$ står for «og», og «og» i sannsynlighet står for snitt. Det vil si

$$P\big(X = x, Y = y\big) = P\big(X = x \text{ og } Y = y\big) = P\big(\{X = x\} \cap \{Y = y\}\big)$$

EKSEMPEL 21.2

For å finne simultanfordelingen til X, Y i eksempel 21.1 må vi regne ut $P(X = x, Y = y)$ for $x = 0, 1, 2$ og $y = 0{,}1$, altså seks tall i alt (seks mulige kombinasjoner av X og Y).

Vi har $2^2 = 4$ mulige utfall av myntkastet og to kombinasjoner av X og Y som umulig kan inntreffe – til sammen 6 kombinasjoner. La M og K stå for kron og mynt. Vi kan ha MM, MK, KM, KK. Alle fire utfall har samme sannsynlighet, altså 1/4 La oss samle verdiene til X og Y som vi da får i en tabell.

Myntkast	Verdi av X	Verdi av Y	Sannsynlighet
MM	0	0	¼
MK	1	0	¼
KM	1	1	¼
KK	2	1	¼

Vi kan nå lese av alle sannsynlighetene til X og Y. Vi får

$$P(X=0, Y=0) = P(\text{MM}) = 1/4$$

fra tabellens øverste rad. Tilsvarende får vi

$$P(X=1, Y=0) = P(\text{MK}) = 1/4$$
$$P(X=1, Y=1) = P(\text{KM}) = 1/4$$
$$P(X=2, Y=1) = P(\text{KK}) = 1/4$$

Ingen andre kombinasjoner av X, Y har noen sannsynlighet for å skje. Vi får dermed følgende tabell for simultanfordelingen til X og Y. Tallene inni tabellen gir sannsynligheten for $P(X = x, Y = y)$.

	$x = 0$	$x = 1$	$x = 2$
$y = 0$	¼	¼	0
$y = 1$	0	¼	¼

Bortsett fra at vi har eksplisitt definert variablene X og Y, så er eksemplet over en direkte anvendelse av teorien fra kapittel 6 og 7. Så hvorfor er vi interessert i simultanfordelinger?

For det første gir simultane sannsynlighetsfordelinger enkle utregningsregler for å beregne samvariasjon, som er et viktig fokus i mange studier.

For det andre er sannsynlighetsregning basis for de fleste økonomiske og finansielle modeller, modeller i markedsføring, psykologi, kort sagt for model-

Avslutningsvis legger vi merke til at vi også kan regne ut variansen til en differanse ved formelen over. Vi viser i seksjon 21.4.2 at

$$\mathrm{Var}(X-Y)=\mathrm{Var}(X)+\mathrm{Var}(Y)-2\mathrm{Cov}(X,Y)$$

21.4.1 Utregning som gir formelen for varians av en sum av to tilfeldige variabler

Vi viser nå $\mathrm{Var}(X+Y)=\mathrm{Var}(X)+\mathrm{Var}(Y)+2\mathrm{Cov}(X,Y)$. Vi husker fra kapittel 9 at en alternativ formel for variansen til en variabel la oss si W, er $\mathrm{Var}(W)=E(W^2)-(E(W))^2$. Dette gir

$$\mathrm{Var}(X+Y)=E\left((X+Y)^2\right)-\left(E(X+Y)\right)^2$$

Vi bruker regneregler fra seksjon 9.2. Siden $E(X+Y)=E(X)+E(X)$, har vi

$$\left(E(X+Y)\right)^2=\left(E(X)+E(Y)\right)^2=E(X)^2+E(Y)^2+2E(X)E(Y)$$

Vi har også at $(X+Y)^2=X^2+2XY+Y^2$. Dette gir

$$\begin{aligned}&E\left((X+Y)^2\right)-\left(E(X+Y)\right)^2\\&\quad=E\left(X^2+2XY+Y^2\right)-\left(E(X)^2+E(Y)^2+2E(X)E(Y)\right)\\&\quad=E\left(X^2+2XY+Y^2\right)-E(X)^2-E(Y)^2-2E(X)E(Y)\end{aligned}$$

Forventningen til en sum er summen til forventningene (dette følger fra å bruke $E(X+Y)=E(X)+E(X)$ flere ganger). Derfor er

$$\begin{aligned}E\left(X^2+2XY+Y^2\right)&=E\left(X^2\right)+E(2XY)+E\left(Y^2\right)\\&=E\left(X^2\right)+2E(XY)+E\left(Y^2\right)\end{aligned}$$

der vi har brukt regneregelen $E(aW)=aE(W)$ for $a=2$ og $W=XY$. Vi får derfor

$$\begin{aligned}&E\left(X^2+2XY+Y^2\right)-E(X)^2-E(Y)^2-2E(X)E(Y)\\&=E\left(X^2\right)+2E(XY)+E\left(Y^2\right)-E(X)^2-E(Y)^2-2E(X)E(Y)\\&=\left(E\left(X^2\right)-E(X)^2\right)+\left(E\left(Y^2\right)-E(Y)^2\right)+2\left(E(XY)-E(X)E(Y)\right)\end{aligned}$$

Y

de røde parentesene skal ikke være her

EKSEMPEL 22.7

Regresjonstabellen for USD

	Estimate	Std. Error	*t* value	Pr(>|*t*|)
(Intercept)	1.508	0.282	5.354	0.000
EUR	0.791	0.025	?	0.000

Skriv opp formelen for regresjonslinja. Regn ut t for $\hat{\beta}_1$.

Svar: $\text{USD} = 1.508 + 0.791 \cdot \text{EUR} \cdot t = 0.791/0.025 = 31.6$

22.4.2 Konfidensintervall for regresjonsmodellen

Ved å snu om på brøkene for testobservatoren i likning (22.5) får vi følgende formel for et 95 % konfidensintervall:

$$\hat{\beta}_1 \pm t^*_{0.025} \cdot \cancel{SE}$$

$\mathrm{SE}(\hat{\beta}_1)$

Her er $t^*_{0.025}$: kritisk verdi fra tabell, bruk $(n - 2)$ frihetsgrader.

EKSEMPEL 22.8

Vi ser videre på mattescore-eksemplet. Vi har $n = 100$ studenter i utvalget, og regresjonstabellen er

	Estimate	Std. Error	*t* value	Pr(>|*t*|)
(Intercept)	10.447	8.453	1.236	0.220
Kartleggingsscore	0.901	0.160	5.650	0.000

Et 95 % konfidensintervall for $\hat{\beta}_1$ er

$$0.901 \pm t^*_{0,025} \cdot 0.160$$

der vi kan slå opp $t^*_{0,025}$ i tabell (eller bedre: bruke kalkulator). Vi har $100 - 2 = 98$ frihetsgrader. Konfidensintervallet er

$$0.901 \pm 1.987 \cdot 0.160 \rightarrow 0.901 \pm 0.318$$

Vi er 95 % sikre på at stigningstallet β_1 ligger i intervallet (0.584, 1.218).

22.4 Inferens for β_0 og β_1

Vi vet at regresjonslinja er en nyttig oppsummering av en lineær trend. Vi tar nå dette et steg videre og generaliserer til populasjonen som utvalget er trukket fra. I forrige seksjon så vi på forutsetningene vi må gjøre for at inferensen skal fungere. For det første må gjennomsnittet til y være bestemt lineært av x:

$$\mu_{y|x} = \beta_0 + \beta_1 x$$

For det andre må ε ha et konstant standardavvik σ_ε – og enten må vi ha et stort utvalg, eller ε må være normalfordelt.

Korrelasjon $\leftrightarrow \beta_1 \neq 0$.

Det aller viktigste spørsmålet i regresjon er om x og y har samvariasjon i populasjonen. Vi tester dette ved å se om stigningstallet β_1 er forskjellig fra null. For hvis $\beta_1 = 0$, er jo $\mu_{y|x} = \beta_0 + \beta_1 x = \beta_0 + 0x = \beta_0$. Altså endrer ikke $\mu_{y|x}$ seg når x endrer seg, og det er ingen samvariasjon annet enn tilfeldig variasjon fra ε. Dersom $\beta_1 > 0$, har vi en positiv samvariasjon i populasjonen, siden $\mu_{y|x} = \beta_0 + \beta_1 x$ da er likningen for en rett linje med positivt stigningstall. Tilsvarende vil $\beta_1 < 0$ bety at vi har en negativ samvariasjon i populasjonen. Med andre ord kan vi utføre en test for samvariasjon ved å teste hypotesene

$$H_0 : \beta_1 = 0 \quad \text{mot} \quad H_A : \beta_1 \neq 0$$

Som vanlig bruker vi en testobservator som måler spriket mellom observert og forventet verdi:

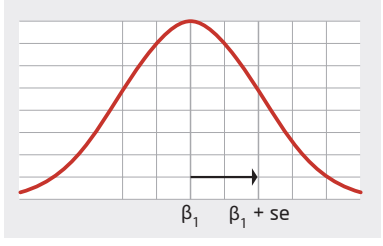
$$t = \frac{\hat{\beta}_1 - 0}{SE} \qquad (22.5)$$

$SE(\hat{\beta}_1)$

$t = \dfrac{\text{observert} - \text{forventet}}{\text{standardfeil}}$

Det kan vises at hvis antagelsene om modellen for enkel lineær regresjon er oppfylt, er denne testobservatoren nær t-fordelt med $n - 2$ frihetsgrader. Her er SE standardfeilen til $\hat{\beta}_1$, som viser hvor mye $\hat{\beta}_1$ varierer fra utvalg til utvalg rundt den sanne verdien β_1, som skissert på figuren i margen. Den samme logikken holder for den andre parameteren, $\hat{\beta}_0$, som har sin egen SE.

Figur 22.13 Utvalgsfordelingen til $\hat{\beta}_1$



Hvorfor trekker vi ifra 2 i stedet for 1 for frihetsgraden, slik vi gjorde for t-test av gjennomsnittet? Grunnen er at vi har estimert *to* størrelser: $\hat{\beta}_0$ og $\hat{\beta}_1$. Frihetsgradene i en modell er n minus antall estimerte parametere i modellen. Siden enkel regresjon inneholder to parametere, er frihetsgraden $n - 2$.

$n - 2$ frihetsgrader i enkel regresjon.

22.4.1 Regresjonstabell fra statistikkprogram

I praksis bruker vi statistisk programvare til å regne ut $\hat{\beta}_0$ og $\hat{\beta}_1$ og deres standardfeil SE. Uansett hvilket program vi bruker, mottar vi resultatet i form av en eller flere tabeller. I regresjonsmodeller estimeres flere parametre og

EKSEMPEL 22.13

En finansanalytiker har estimert regresjonsmodellen for valutakurser basert på data fra de siste $n = 173$ dagene:

$$\text{USD} = 1.508 + 0.791 \cdot \text{USD}$$

Hun tenker seg et scenario for morgendagen, $\text{EUR}_{ny} = 11.1$, og stiller seg følgende spørsmål: Prognosen for USD er USD $= 1.508 + 0.791 \cdot 11.1 = 10.29$ NOK. Men hvor sikker er jeg på dette tallet? Jeg trenger et intervall som med 95 % sikkerhet vil inneholde morgendagens USD, i en situasjon der $\text{EUR}_{ny} = 11.1$.

$(\widehat{\beta_1})$

Slike intervaller for individuelle y-observasjoner i en subpopulasjon definert av x kalles *prediksjonsintervaller*. Det er lett å forveksle prediksjonsintervaller med konfidensintervallet for $\mu_{y|x}$, som er et tradisjonelt konfidensintervall for en parameter. Prediksjonsintervallet er konstruert slik at det vil inneholde den faktiske y-observasjonen 95 % av gangene, så lenge regresjonsmodellens antagelser er oppfylt. I tillegg må feilleddene være nøyaktig normalfordelte. Formelen for prediksjonsintervallet er

$$\hat{y} \pm t^*_{0.025} \cdot s_e \cdot \sqrt{1 + \frac{1}{n} + \frac{SE(b_1)^2 \cdot (x_{ny} - \bar{x})^2}{s_e^2}}$$

der $t^*_{0.025}$ har $n - 2$ frihetsgrader.

Vi illustrerer formelen med et eksempel som er mer komplisert enn det ser ut til, siden valutakurser varierer med tid på en komplisert måte. Utvalget vårt består derfor ikke av tverrsnittsdata, men innebærer tidsavhengighet. For å forsvare bruk av prediksjonsintervall i et slikt tilfelle kreves det teori som vi ikke dekker her. Vi tillater oss likevel å bruke dette eksemplet som en relevant anvendelse av prediksjonsintervall.

EKSEMPEL 22.14

Scenarioet for morgendagen: $\text{EUR}_{ny} = 11.1$. I utvalget over $n = 173$ dager er gjennomsnittet $\widehat{\text{EUR}} = 11.308$. Regresjonstabellen er

	Estimate	Std. Error	*t* value	Pr(>|*t*|)
(Intercept)	1.508	0.282	5.354	0.000
EUR	0.791	0.025	31.780	0.000

22.8 Oppsummering av begreper og formler

- Regresjonsmodellen $\mu_{y|x} = \beta_0 + \beta_1 x$ sier at den forventete verdien til en *responsvariabel* y avhenger lineært av verdien til *prediktorvariabelen* x
- Modellen krever 1) at spredningsdiagrammet av y mot x har en lineær trend og 2) at feilleddene ε_i har konstant varians (homoskedastisitet) og er normalfordelte (alternativt at utvalget er stort).
- I praksis kontrollerer vi homoskedastisitet ved å sjekke at *residualene* $e_i = y_i - \hat{y}_i$ har samme varians uavhengig av x ved hjelp av et residualplott. Hvorvidt feilleddene er normalfordelte sjekker vi i praksis ved hjelp av et QQ-plott.
- *Modellanslaget* $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_1$ kan brukes til å predikere y når x er kjent. Verdiene $\hat{\beta}_0$ og $\hat{\beta}_1$ beregnes fra minste kvadraters metode.
- Hypotesetester for β_0 og β_1 baseres på t-fordeling med $n - 2$ frihetsgrader. Viktigst er det å teste om y og x har lineær samvariasjon:
  $H_0 : \beta_1 = 0, H_A : \beta_1 \neq 0$
  Vi utfører testen ved å finne p-verdi i regresjonstabell fra software.
  Eller ved å regne ut testobservatoren $t = \dfrac{\hat{\beta}_1 - \beta_1}{SE}$
  der $\hat{\beta}_1$ er den estimerte verdien til β_1. Standardfeilen SE finner vi i regresjonstabellen. Tilsvarende gjelder for å teste hypoteser om β_0.
- Et 95 % konfidensintervall for β_1 er $\hat{\beta}_1 \pm t^*_{0.025} \cdot SE$ der vi finner standardfeilen SE i regresjonstabellen fra software. Vi bruker $n-2$ frihetsgrader i t-fordelingen.
- Selv om x og y har samvariasjon (β_1 er ikke 0) så er det ikke sikkert at x forårsaker y. Det kan hende at samvariasjonen er spuriøs, og at det er en ukjent variabel som påvirker både x og y.

$\mathrm{SE}(\hat{\beta}_1)$

f. $t = -0.312$, behold H_0.

g. Prognosen for Arild er $\hat{y} = 37.378 + 2.906 \cdot 8 = 60.626$.

h. $60.626 \pm 1.987 \cdot 21.89 \cdot \sqrt{1 + \frac{1}{100} + \frac{1.189^2 \cdot (8 - 6.82)^2}{21.89^2}}$,

slik at prediksjonsintervallet blir (16.82, 104.43).

Løsning på oppgave 22.6

a. 0.5809, 0.9838, 0.4861, 1.6474, 1.0549, 0.5098, 1.126

b. 0.2191, 0.1162, 0.0139, 0.1526, 0.0451, −0.1098, −0.426

c. $s_e = \sqrt{\frac{(0.2191)^2 + (0.1162)^2 + \cdots + (-0.426)^2}{7 - 2}} = 0.237$

d. Budsjettet har standardavviket $s_x = 1.778$, og $SE(b_1) = \frac{0.237}{1.778 \cdot \sqrt{7 - 1}} = 0.054$

Løsning på oppgave 22.7

a. $\widehat{Fødselsvekt} = 3395.53 - 14.57 \cdot Sigg$

b. $\hat{\beta}_1$ forteller oss at når man røyker 1 mer røyk under graviditeten så synker den gjennomsnittlige fødselsvekten til ungen med 14.57 gram.
$\hat{\beta}_0$ forteller oss at dersom man ikke røyker under graviditeten så vil gjennomsnittlig fødselsvekt til ungen være 3395.53 gram

c. Et 99 % konfidensintervall for β_1 er $\hat{\beta}_1 \pm t_{0,005} \cdot SE(\hat{\beta}_1 \Leftrightarrow -14.57 \pm 2.579 \cdot 2.565 \Leftrightarrow -14.57 \pm 6.62\ (-21.2\,, -7.95)$. M.a.o. er vi er 99 % sikre på at fødselsvekten til et barn synker når mor røyker under graviditeten sammenliknet med om hun ikke røyket, og at nedgangen er et sted mellom 7.95 og 21.2 gram per ekstra røyk man røyker i gjennomsnitt under graviditeten. Vi benyttet t-fordelingen med $1388 - 2 = 1386$ frihetsgrader i denne oppgaven.

d. Vi leser av i tabell 1 at den tosidige p-verdien når vi tester H_0: $\beta_1 = 0$ mot H_1: $\beta_1 < 0$ er mindre enn 0.0001, som betyr at vår ensidige test har en p-verdi mindre enn $0.0001/2 = 0.00005$, som er atskillig lavere enn signifikansnivået $\alpha = 0.05$. Vi forkaster derfor H_0: $\beta_1 = 0$

e. Vi har tiltro til at fødselsvekten for barn av mødre som røyket under graviditeten (i USA i 1988) synker med økt antall røykte sigaretter.