

ALF HARBITZ

# Statistikk og sannsynlighets- regning

2. utgave



## Forord

Denne læreboka dekket rammeplanens pensumkrav på 2 vekttall statistikk for studenter ved 3-årig ingeniørutdanning på 2000-tallet. Dette utgjorde den primære målgruppen for boka. Dets behagelige matematiske nivå burde imidlertid gjøre boka aktuell også for andre, for eksempel ved universitetenes brukerkurs i statistikk.

Boka inneholder over 200 oppgaver. Ca. 80 av disse er tidligere eksamensoppgaver i statistikk gitt ved 3-årig ingeniørutdanning ved ulike høgskoler i Norge. Boka inneholder også en rekke ingeniør-relevante eksempler. Bruk av dataverktøy er illustrert ved Minitab- og Matlab-eksempler, der komplett kode er angitt. Alle nødvendige fordelingstabeller (utarbeidet ved hjelp av statistikk-pakken til Matlab) er gjengitt bakerst i boka.

En egen oppgavesamling (ca. 150 sider) med komplette løsningsforslag til alle oppgavene i læreboka er også utarbeidet av samme forfatter, og vil etter planen bli utgitt av Fagbokforlaget fra høstsemesteret 1999. Den har tittel «Statistikk og sannsynlighetsregning. Oppgaver og løsningsforslag».

Læreboka er for omfattende til at alt kan tas med i et 2-vekttalls (år 2000 terminologi) kurs. Forslag til hva som kan kuttes er listet litt lenger ned. En prioritering av Monte Carlo simulering i forhold til hypotesetesting anbefales. Dette er i tråd med SEFI-rapporten utgitt av en rekke fremtredende matematikere og statistikere i Europa vedrørende ingeniørutdanning. Her anbefales simulering som et obligatorisk emne, mens hypotesetesting er angitt som valgfritt.

En rekke enkeltpersoner har på forskjellig vis bidratt til bokas endelige utforming og innhold. Flere fagstatistikere har bidratt med gjennomlesning av deler av boka, diskusjoner og konstruktive råd til bedring som har vært tatt til følge. Blant disse er Kari Birkeland Nilsen (Høgskolen i Bergen) når det gjelder lineær regresjon, variansanalyse og generelle trekk ved boka, Øystein Evandt (Det norske Veritas) vedrørende Shewart-diagrammer og Jon Helgeland (Statens Datasentral) når det gjelder simulering. Jan Gunnar Moe (Høgskolen i Ålesund) har bidratt grundig vedrørende matematiske skriveregler, og har gitt betydelige pedagogiske bidrag til flere av kapitlene. Pedagogisk har også Carl Harbitz (Norges Geotekniske Institutt) gitt konstruktive råd og bidratt med en realistisk og mye brukt snøskredmodell vedrørende lineær regresjon. Karianne Helland har vært effektiv korrekturleser. Ellers takker jeg Fagbokforlaget for godt samarbeid,

ikke minst Ellen Sandberg for bidrag på layout-siden. En rekke andre har også bidratt på forskjellig vis, takk til dere alle.

Forslag til hva som kan kuttes i et 2 vekttalls kurs:

**Kap. 1** (*Beskrivende statistikk*). Formelkverning kan nedtones. Faren ved å ha med for få siffer i mellomregninger bør imidlertid påpekes.

**Kap. 2** (*Sannsynlighetsregning*). Man kan trolig begrense seg til å se på regneregler for to hendelser, og f.eks. kutte addisjonssetningen for 3 hendelser.

**Kap. 5** (*Kontinuerlige fordelinger*). Her kan man som et minimum trolig nøye seg med normalfordelingen og sentralgrenseteoremet. Uniform fordeling anbefales imidlertid å berøre allerede på dette stadium, på grunn av dens sentrale betydning for simulering. Dessuten er eksponensial-fordelingen verdt å vurdere, blant annet på grunn av sin enkle teoretiske form og «hukommelsesfrie» egenskap.

**Kap. 7** (*Hypotesetesting*). Vedrørende styrkefunksjon og styrkekurve kan man vurdere å nøye seg med f.eks. styrkefunksjonen for en ensidig test av  $\mu$ . Man kan også vurdere å utelate tosidige tester.

**Kap. 8** (*To populasjoner*). Hele kapitlet kan trolig sløyfes til fordel for mer Monte Carlo-simulering.

**Kap. 10** (*Variansanalyse*). Hele kapitlet kan trolig sløyfes.

**Kap. 11** (*Monte Carlo-simulering*). Her kan man nøye seg med å forklare hvorfor uavhengige tilfeldige variabler fra en gitt fordeling i prinsippet fremkommer ved å generere uavhengige uniformt fordelt variabler mellom 0 og 1 for deretter å beregne invers kumulativ fordelingsfunksjon. Deretter kan man gå nokså direkte løs på bruk av dataverktøy. Det anbefales imidlertid at hele kapitlet vektlegges, og at den meget nyttige bootstrap-teknikken tas med.

**Kap. 12** (*Shewart-diagrammer*). Dette er trolig et kapittel som faller relativt lett. Kapitlet kan egne seg til selvstudium om tiden er knapp. Innholdet er imidlertid svært viktig og akutelt i en tid da det statdig mer fokuseres på kvalitet i næringsliv og annet arbeidsliv.

Nedenfor er gjengitt en litteraturliste, blant annet over kilder som er brukt som inspirasjon i utarbeidelsen av denne boka. Dette gjelder særlig [1], som er en klassiker hva angår statistisk innføringslitteratur på engelsk. Noen eksempler og oppgaver er med tillatelse hentet herfra. Jeg vil med dette få takke nevnte kilde

og andre som har gitt tillatelse til å ta med eksempler og oppgaver i boka. [2] er en grundig og fyldig bok på norsk som omhandler svært ingenør-relaterte emner innen statistikk og sannsynlighetsregning. [3] inneholder mange interessante anvendelser vedrørende bølger og bølgebelatsninger på maritime konstruksjoner. Matematisk er [3] noe tyngre enn [1] og [4]. [4] er en god bok både faglig og pedagogisk for dem som vil gå mer grundig inn i statistisk prosesstyring og bruk av Shewart-diagrammer. Det matematiske nivået er behagelig lavt. [5] er en god og svært vid innføringsbok når det gjelder kvalitet og kvalitetsarbeid generelt. Den er lettlest på et behagelig lavt matematisk nivå. [6] er norske klassikere, særlig brukt ved statistikkundervisningen for statistikk-studenter ved Universitetet i Oslo. Matematikken er adskillig tyngre enn i de andre bøkene som er listet. Her finner du det meste av det du trenger vedrørende bevis. [7] er en klassiker hva matematiske oppslagsverk angår, og inneholder blant annet tilnærningsformlene for invers kumulativ normalfordeling benyttet i henværende bok.

## Litteraturliste

- 1) Bhattacharyya, Gouri K., og Johnson, Richard A., «Statistical Concepts and Methods», John Wiley & Sons, Inc., 1977, 640 sider.
- 2) Aven, Terje, «Pålitelighets- og risikoanalyse», Universitetsforlaget, 1991, 270 sider.
- 3) Gran, Sverre, «Lectures in Ocean Engineering – Waves and Wave Forces», A.S. Veritas Research Report No. 85-2028, 1985, 250 sider.
- 4) Wheeler, Donald J., og Chambers, David S., «Understanding Statistical Process Control», SPC Press, 1992, 410 sider.
- 5) Bergman, Bo og Klefsjö, Bengt, «Quality from Customer Needs to Customer Satisfaction», ISBN 91-44-46331, Studentlitteratur, 1994, 480 sider.
- 6) Sverdrup, Erling, «Lov og tilfeldighet», bind 1 og 2, Universitetsforlaget, 1973, 390 + 260 sider.
- 7) Abramowitz, M, og Stegun, I.A., «Handbook of Mathematical Functions», Dover Publications, Inc., New York, 1972, 1050 sider.

# Symbolliste og forkortelser

$\text{Bino}(n,p)$	binomisk fordeling med parametre $n$ og $p$
$\text{Corr}(X,Y)$	korrelasjonskoeffisient mellom $X$ og $Y$
$\text{CV}$	variasjonskoeffisient
$\text{d.f.}$	«degrees of freedom», antall frihetsgrader
$E(\cdot)$	forventnings-operator
$\text{expo}(b)$	eksponensial-fordeling med skalaparameter $b$
$F(m,n)$	$F$ -fordeling med $m$ og $n$ frihetsgrader
f.eks.	for eksempel
$\text{gamma}(b,c)$	gammafordeling med skalaparameter $b$ og formparameter $c$
h.k.	halvkorreksjon
$\text{hyp}(n,D,N)$	hypergeometrisk fordeling med parametre $n$ , $D$ og $N$
KI	konfidensintervall
$\text{Kji2}(n)$	kjikvadratfordeling med $n$ frihetsgrader
$N(\mu,\sigma)$	normalfordeling med parametre $\mu$ og $\sigma$
$N_2(\mu_1,\sigma_1;\mu_2,\sigma_2;\rho)$	binormal fordeling med parametre $\mu_1, \sigma_1, \mu_2, \sigma_2$ og $\rho$
$P^*$	signifikanssanssynlighet, også kalt $p$ -verdi
$\text{Po}(\lambda)$	Poisson-fordeling med parameter $\lambda$
$r$	empirisk korrelasjonskoeffisient
$\text{Rayl}(b)$	Rayleigh-fordeling med skalaparameter $b$
$S, s$	empirisk standardavvik
SSE	feilkvadratsum
$\text{SS}_T$	behandlingskvadratsum
$\text{std}(\cdot)$	standardavvik-operator
$\text{stud}(n)$	Students- $t$ fordeling med $n$ frihetsgrader
$t_\alpha$	øvre $\alpha$ -fraktil i $t$ -fordelingen
tn.	tilnærmet
$\text{U}[a,b]$	uniform fordeling definert på $[a,b]$
uif	uavhengige og identisk fordelte
$\text{Var}(\cdot)$	varians-operator
$\text{Weib}(b,c)$	Weibull-fordeling med skalaparameter $b$ og formparameter $c$
$\bar{X}, \bar{x}$	Overstrekning betyr empirisk middelverdi
$z_\alpha$	øvre $\alpha$ -fraktil i $N(0,1)$ -fordelingen
$\mu$	populasjons-forventning
$\sigma$	populasjons-standardavvik
$\rho$	korrelasjonskoeffisient
$\theta$	Generelt symbol for ukjent parameter
$\Phi, \phi$	$N(0,1)$ -funksjoner: $\Phi$ : kumulativ, $\phi$ : sannsynlighetstetthet
$\Gamma(x)$	Gammafunksjonen ( $\Gamma(x+1) = x!$ )

# Innhold

## 1 Beskrivende Statistikk

- 1.1 Innledning 1
- 1.2 Rådata 1
- 1.3 Rangordning av data 6
- 1.4 Grupperte data 9
- 1.5 Spredningsdiagram 18
- 1.6 Empirisk korrelasjonskoeffisient 20
- 1.7 Lineær regresjon 23
- 1.8 Oppgaver 27
- 1.9 Formelsamling 30

## 2 Sannsynlighetsregning

- 2.1 Innledning 31
- 2.2 Utfallsrom, enkeltutfall og hendelse 32
- 2.3 Sannsynlighet for en hendelse 33
- 2.4 Union, snitt og komplement 37
- 2.5 Kombinatorikk, telleregler 43
- 2.6 Betinget sannsynlighet 50
- 2.7 Uavhengige hendelser 54
- 2.8 Oppgaver 57
- 2.9 Formelsamling 62

## 3 Stokastisk variabel og sannsynlighetsfordeling

- 3.1 Innledning 63
- 3.2 Diskrete stokastiske variabler 64
- 3.3 Sannsynlighetsfordeling 66
- 3.4 Fordelingsdiagrammer 68
- 3.5 Forventning ( $\mu$ ) 69
- 3.6 Varians ( $\sigma^2$ ) og standardavvik ( $\sigma$ ) 72
- 3.7 Simultanfordeling (to variabler) 74
- 3.8 Kovarians og korrelasjon 76
- 3.9 Uavhengighet mellom to variabler 79
- 3.10 Oppgaver 81
- 3.11 Formelsamling 83

## 4 Diskrete fordelinger

- 4.1 Innledning 84

- 4.2 Binomisk fordeling 85
- 4.3 Hypergeometrisk fordeling 93
- 4.4 Poisson-fordelingen 96
- 4.5 Oppgaver 103
- 4.6 Formelsamling 109

## 5 Kontinuerlige fordelinger

- 5.1 Innledning 110
- 5.2 Sannsynlighetsmodell 111
- 5.3 Overlevelsfunksjon og feilrate 115
- 5.4 Uniform fordeling 116
- 5.5 Eksponensial-fordelingen 119
- 5.6 Gammafordelingen 121
- 5.7 Weibull-fordelingen 122
- 5.8 Rayleigh-fordelingen 124
- 5.9 Normalfordelingen 126
- 5.10 Sentralgrenseteoremet 133
- 5.11 Normaltilnærmelse til binomisk fordeling 137
- 5.12 Normaltilnærmelse til Poisson-fordelingen 139
- 5.13 Kjikvadrat-fordelingen 140
- 5.14  $t$ -fordelingen 141
- 5.15  $F$ -fordelingen 143
- 5.16 Binormal fordeling 144
- 5.17 Oppgaver 147
- 5.18 Formelsamling 158

## 6 Estimering

- 6.1 Innledning 159
- 6.2 Punktestimering av parameter 160
- 6.3 Punktestimering av
- 6.4 Punktestimering av  $\sigma^2$  163
- 6.5 Punktestimering av binomisk  $p$  163
- 6.6 Estimering ved konfidensintervall (KI) 164
- 6.7 KI for små utvalg, kjent 168
- 6.8 KI for , store utvalg 169
- 6.9 KI for  $p$ , store utvalg 169

6.10	KI for $\mu$ , små utvalg, $\sigma$ ukjent	
	171	
6.11	Oppgaver	173
6.12	Formelsamling	179
<b>7</b>	<b>Hypotesetesting</b>	
7.1	Innledning	180
7.2	Hypotesene $H_0$ og $H_1$	181
7.3	Feiltyper og styrkefunksjon	185
7.4	Valg av forkastingsområde	188
7.5	Tester med tosidig alternativ	191
7.6	Generelle trinn i hypotesetesting	
	193	
7.7	Test av $\mu$	194
7.8	Test av binomisk $p$	201
7.9	Pearsons kjikvadrat-tilpasningstest	
	203	
7.10	Oppgaver	206
7.11	Formelsamling	216
<b>8</b>	<b>To populasjoner</b>	
8.1	Innledning	217
8.2	Tilfeldiggjøring (randomisering)	
	219	
8.3	Uavhengige tilfeldige utvalg	221
8.4	Parvis sammenligning	230
8.5	Oppgaver	234
8.6	Formelsamling	238
<b>9</b>	<b>Lineær regresjon</b>	
9.1	Innledning	239
9.2	Minste kvadraters estimatorer $a^*$ og $b^*$	242
9.3	Egenskaper til $b^*$	243
9.4	Egenskaper til $a^*$	243
9.5	Estimering av $\sigma$	244
9.6	Prediksjon av $Y$ når $x$ er gitt	246
9.7	Konfidensintervall og hypotesetesting	248
9.8	Transformasjon av variabler	253
9.9	Residualer og modellsjekk	256
9.10	Oppgaver	261
9.11	Formelsamling	266

<b>10</b>	<b>Variansanalyse</b>	
10.1	Innledning	267
10.2	Sammeligning av $k$ behandlinger	269
10.3	Populasjonsmodell og inferens	275
10.4	Oppgaver	281
10.5	Formelsamling	283
<b>11</b>	<b>Monte Carlo Simulering</b>	
11.1	Innledning	284
11.2	Generering av variabler fra $F^{-1}$	
	286	
11.3	Slumptallgenerering	289
11.4	Generering av normalfordelte variable	291
11.5	Generering av binormale variabelpar	294
11.6	Generering av sammensatt variabel	297
11.7	Bootstrap	300
11.8	Oppgaver	304
11.9	Formelsamling	305
<b>12</b>	<b>Shewart-diagrammer</b>	
12.1	Innledning	306
12.2	$\bar{X}$ - $R$ -diagrammer	309
12.3	$XmR$ -diagrammer	314
12.4	Robusthet og myter	316
12.5	To grunnregler	317
12.6	Oppgaver	318
12.7	Formelsamling og tabell	319
	<b>Tabeller – binomisk</b>	320
	<b>Tabeller – Poisson</b>	325
	<b>Tabeller – normal</b>	329
	<b>Tabeller – t-fraktiler</b>	331
	<b>Tabeller – kjivadrat-fraktiler</b>	332
	<b>Tabeller – F-fraktiler</b>	333
	<b>Fasit</b>	335
	<b>Stikkord</b>	341

## Kapittel 1

# Beskrivende statistikk

## 1.1 Innledning

Beskrivende statistikk kalles også **deskriptiv** statistikk, etter det engelske ordet *descriptive*. Kapitlet omhandler de vanligste metodene for sammenfatning og presentasjon av et statistisk tallmateriale. En kan gjerne si at statistikker i tradisjonell forstand, slik de blant annet kommer til uttrykk i Statistisk Årbok, sorterer under den del av statistikken som kalles beskrivende statistikk.

Beskrivende statistikk omhandler *ikke* det vi betegner som sannsynlighetsregning, statistiske metoder og statistisk inferens. En rekke *begreper* innen beskrivende statistikk er imidlertid felles med de andre delene av statistikken. Beskrivende statistikk utgjør ikke minst derfor en naturlig start på en innføringsbok i statistikk og sannsynlighetsregning.

## 1.2 Rådata

Med **rådata** (også kalt *primærdata*) skal vi mene det opprinnelige tallmaterialet vi skal behandle, inneholdende  $n$  observasjoner  $x_1, \dots, x_n$ . Vi skal i første omgang se på beregning av følgende to svært viktige størrelser som karakteriserer tallmaterialet:

- (*Empirisk*) **middelverdi** som mål på tyngdepunkt (senter) i tallmaterialet.
- (*Empirisk*) **standardavvik** som mål på spredning rundt middelverdien.

Et annet ord for middelverdi er **gjennomsnitt**. Standardavvik er en størrelse som alltid er ikke-negativ og som får større og større verdi jo mer spredning (*avvik* fra middelverdien) det er i datamaterialet. Ordet *empirisk* betyr at tallmaterialet består av reelle data hentet fra virkelighetens verden, i motsetning til «teoretiske» data. Denne forskjellen vil komme klarere fram i senere kapitler. Formler for beregning av empirisk middelverdi og standardavvik er gitt i de neste rammer.

## Empirisk middelverdi

Vi har  $n$  tall som vi betegner  $x_1, x_2, \dots, x_n$ . Empirisk middelverdi,  $\bar{x}$ , til tallene er da definert ved følgende formel:

$$(1.1) \quad \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i = \frac{1}{n} (x_1 + x_2 + \dots + x_n)$$

der den horisontale streken over  $x$ 'en betyr middelverdi.

**NB!** Ordet gjennomsnitt er også ofte brukt, og det betyr akkurat det samme som middelverdi.

### Eks. 1.1 Beregning av middelverdi (lign. (1.1))

- a) Tre ørreter veier henholdsvis 2.1, 1.7 og 0.7 kg. Gjennomsnittsvekta, eller midlere vekt av de tre ørretene, blir:

$$\bar{x} = \frac{1}{3} \cdot (2.1 + 1.7 + 0.7) \text{ kg} = \frac{1}{3} \cdot 4.5 \text{ kg} = \underline{1.5 \text{ kg}}$$

- b) I en liten bedrift med 5 ansatte har de ansatte følgende årslønn ( i kr 1000):  
240, 180, 270, 210 og 160.

Vi skal beregne totale årlige lønnskostnader for bedriften og gjennomsnittlig lønn. La  $x_i$  betegne lønning nr.  $i$ , dvs.  $x_1 = 240$ ,  $x_2 = 180$  osv. De totale lønnskostnadene finnes ved å summere alle lønningene:

$$\sum_{i=1}^5 x_i = x_1 + x_2 + x_3 + x_4 + x_5 = 240 + 180 + 270 + 210 + 160 = 1060$$

dvs. totale lønnskostander er på kr 1 060 000

Gjennomsnittslønna  $\bar{x}$  finner vi ved å dele summen av inntekter på antall ansatte:

$$\bar{x} = \frac{1}{5} \sum_{i=1}^5 x_i = \frac{1}{5} \cdot \text{kr } 1\ 060\ 000 = \underline{\text{kr } 212\ 000}$$

- c) Gjennomsnittlig kubikkinnhold pr. tre i en homogen (ensartet) skog med 1000 trær er anslått til å være  $1,1 \text{ m}^3$  pr. tre, og vi skal anslå totalt kubikkinnhold i skogen. Vi betegner kubikkinnholdet til trærne  $x_1, \dots, x_{1000}$ , og er altså interessert i å finne summen av alle  $x$ -ene. Vi får da:

$$\bar{x} = \frac{1}{1000} \sum_{i=1}^{1000} x_i = 1,1 \text{ m}^3 \Rightarrow \sum_{i=1}^{1000} x_i = 1000 \cdot 1,1 \text{ m}^3 = \underline{1100 \text{ m}^3}$$

Totalt kubikkinnhold i skogen kan altså anslås til 1100 kubikkmeter. ☺

### Empirisk standardavvik, $s$ , og varians, $s^2$

Vi har  $n$  tall som vi betegner  $x_1, x_2, \dots, x_n$ . Empirisk standardavvik,  $s$ , til tallene, er et mål på hvor spredt tallene er rundt middelverdien,  $\bar{x}$ , og beregnes ved en av følgende 3 formler:

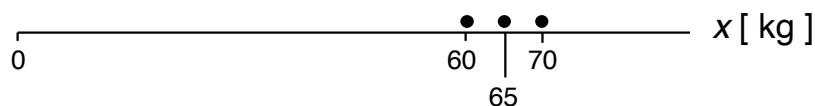
$$(1.2) \quad s = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2}$$

$$(1.3) \quad = \sqrt{\frac{1}{n-1} \left( \sum_{i=1}^n x_i^2 - n \cdot \bar{x}^2 \right)}$$

$$(1.4) \quad = \sqrt{\frac{1}{n-1} \left( \sum_{i=1}^n x_i^2 - \frac{1}{n} \left( \sum_{i=1}^n x_i \right)^2 \right)}$$

NB! Det er følgende entydige sammenheng mellom begrepene *varians* og *standardavvik*: Varians = kvadratet,  $s^2$ , av standardavviket,  $s$ .

Formlene for beregning av standardavviket  $s$  i forrige ramme kan synes noe abstrakte, så la oss prøve å se hva som kan ligge bak utgangsformelen i lign.(1.2). Vi tar utgangspunkt i dataene vist i figuren nedenfor, som viser 3 tilfeldige studentvekter:  $x_1 = 60$  kg,  $x_2 = 65$  kg og  $x_3 = 70$  kg.



Vi beregner først middelverdien, og finner at  $\bar{x} = 65$  kg. Dette stemmer vel rimelig bra med intuisjonen. Hva så med spredningen rundt middelverdien? De fleste vil vel være enige i at  $\pm 5$  kg vil være et rimelig anslag for denne spredningen. La oss prøve å bestemme en generell formel som gir et fornuftig anslag av spredningen  $\Delta x$ . Vi prøver 4 forskjellige alternativer:

**1)** Vi prøver først å ta gjennomsnittlig avvik fra middelverdien:

$$\Delta x = \frac{1}{3} \cdot \Sigma(x_i - \bar{x}) = \frac{1}{3} ((60-65) + (65-65) + (70-65)) \text{ kg} = 0 \text{ kg}$$

Dette stemmer dårlig, vi har jo helt klart en spredning forskjellig fra null.

**2)** For å komme unna fortegnsproblemet i 1), ser vi på gjennomsnittlig absoluttavvik fra middelverdien:

$$\Delta x = \frac{1}{3} \cdot \Sigma |x_i - \bar{x}| = \frac{1}{3} (|60-65| + |65-65| + |70-65|) \text{ kg} = 10/3 \text{ kg}$$

Dette stemmer brukbart med intuisjonen, men absoluttverdier er ikke særlig attraktive å jobbe med rent matematisk.

**3)** Kvadrat er en hendigere matematisk funksjon enn absoluttverdi, så vi ser på gjennomsnittlig kvadratavvik fra middelverdien:

$$\Delta x^2 = \frac{1}{3} \cdot \Sigma(x_i - \bar{x})^2 = \frac{1}{3} ((60-65)^2 + (65-65)^2 + (70-65)^2) \text{ kg}^2 = 50/3 \text{ kg}^2.$$

Benevningen  $\text{kg}^2$  er imidlertid en uhensiktsmessig og abstrakt benevning for å angi usikkerheten til data med benevning i kg.

**4)** Vi tar så kvadratrota av gjennomsnittlig kvadratavvik fra middelverdien, for å få en fornuftig benevning:

$$s = \Delta x = \left( \frac{1}{3} \cdot \Sigma(x_i - \bar{x})^2 \right)^{1/2} = (50/3)^{1/2} \text{ kg} = 4 \text{ kg}$$

som stemmer bra overens med intuisjonen.

Formelen i 4) minner svært om formelen i lign.(1.2). Eneste forskjell er at vi har brukt  $n$  ( $n = 3$  i eksemplet) i nevner, mens det er brukt  $n-1$  i lign.(1.2). Når det dreier seg om store  $n$ -verdier blir resultatet rimelig uavhengig av om vi bruker  $n$  eller  $n-1$ . I kap. 6 kommer vi tilbake til en begrunnelse for hvorfor det står  $n-1$  og ikke  $n$  i nevner i lign.(1.2).

Før vi går løs på eksempler som viser beregning av standardavvik, skal vi angi følgende merknader til beregningsformlene i lign.(1.2)-(1.4) i forrige ramme:

- 1) NB! Særlig ved bruk av lign.(1.3) og (1.4) må en passe på å bruke tilstrekkelig mange siffer i mellomregningene. Hvis vi f.eks. har en stor positiv middelverdi,  $\bar{x}$ , og et lite standardavvik,  $s$ , fremkommer  $s$  ved å subtrahere to store tall fra hverandre. Hvis disse store tallene har for få siffer, får vi lett helt gale svar.

- 2) Lign.(1.4) ovenfor er trolig den formelen som generelt er mest nyttig ved beregning av empirisk varians og standardavvik. Alle formlene (1.2), (1.3) og (1.4) er imidlertid nyttige å beherske for å beregne varians og standardavvik. Eks. 1.2, 1.3 og 1.4 viser nytten ved de enkelte formlene.
- 3) Selv på rimelig enkle lommekalkulatorer er det lagt inn beregning av middelverdi og standardavvik, ofte angitt ved samme symboler som i formlene ovenfor. På enkelte kalkulatorer kan en velge hvorvidt en skal ha  $n$  eller  $n-1$  i nevneren ved beregning av  $s$ . Vi kommer senere tilbake til denne forskjellen.

**Eks. 1.2 Beregning av standardavvik (lign.(1.2))**

$$\begin{aligned} n &= 3, \quad x_1 = 10,0, \quad x_2 = 9,9, \quad x_3 = 10,1 \\ \text{middelverdi: } \bar{x} &= \frac{1}{3} \cdot (10,0 + 9,9 + 10,1) = 10,0 \\ \text{varians: } s^2 &= \frac{1}{2} \cdot \sum_{i=1}^3 (x_i - \bar{x})^2 \\ &= \frac{1}{2} \cdot ((10,0 - 10,0)^2 + (10,0 - 9,9)^2 + (10,1 - 10,0)^2) = 0,01 \\ \text{standardavvik: } s &= \sqrt{0,01} = 0,1 \quad \odot \end{aligned}$$

**Eks. 1.3 Beregning av standardavvik (lign.(1.3))**

$$\begin{aligned} n &= 3, \quad x_1 = 0, \quad x_2 = 1, \quad x_3 = -1 \\ \text{middelverdi: } \bar{x} &= \frac{1}{3} \cdot (0 + 1 + (-1)) = 0 \\ \text{varians: } s^2 &= \frac{1}{2} \cdot \left( \sum_{i=1}^3 x_i^2 - 3\bar{x}^2 \right) = \frac{1}{2} \cdot (0^2 + 1^2 + (-1)^2 - 3 \cdot 0^2) = 1 \\ \text{standardavvik: } s &= \sqrt{1} = 1 \quad \odot \end{aligned}$$

**Eks. 1.4 Beregning av standardavvik (lign.(1.4))**

$$\begin{aligned} n &= 3, \quad x_1 = 5, \quad x_2 = 3, \quad x_3 = 8 \\ \Sigma x_i &= 5 + 3 + 8 = 16 \\ \Sigma x_i^2 &= 25 + 9 + 64 = 98 \\ \text{varians: } s^2 &= \frac{1}{3-1} \left( \sum_{i=1}^3 x_i^2 - \frac{1}{3} \left( \sum_{i=1}^3 x_i \right)^2 \right) = \frac{1}{2} \left( 98 - \frac{1}{3} \cdot 16^2 \right) = 6,33 \\ \text{standardavvik: } s &= \sqrt{6,33} = 2,52 \quad \odot \end{aligned}$$

### 1.3 Rangordning av data

Dersom vi ordner våre observasjoner,  $x_1, x_2, \dots, x_n$ , i stigende rekkefølge, skal vi bruke betegnelsen  $x_{(1)}, x_{(2)}, \dots, x_{(n)}$ .  $x_{(1)}$  er altså betegnelsen for minste verdi,  $x_{(2)}$  for nest minste, opp til  $x_{(n)}$  for største verdi. I tillegg til å beregne middelverdi og standardavvik, kan vi nå beregne følgende:

- a) **median** (midtverdi) som (robust) mål på senter,
- b) **interkvartilbredde** som (robust) mål på spredning og
- c)  **$100p$ -prosentiler** for ønsket verdi av  $p$ .

En fordel med medianen i forhold til middelverdien som mål på senter, er at medianen er lite påvirket av noen få «ekstreme» observasjonsverdier. Tilsvarende er interkvartilbredden et mer **robust** spredningsmål enn standardavviket.

#### Empirisk median (midtverdi), $m$

Her skiller vi mellom tilfellene hvor  $n$  er et **oddetall** ( $n = 3, 5, 7, 9, \dots$ ) og hvor  $n$  er et **liketall** ( $n = 2, 4, 6, 8, \dots$ ).

$n = 3, 5, 7, \dots$  (odde):  $m$  = den midterste av observasjonene etter at alle observasjonene er ordnet i stigende rekkefølge:

$$(1.5) \quad m = x_{\left(\frac{n+1}{2}\right)} \quad (n \text{ odde})$$

$n = 2, 4, 6, \dots$  (like):  $m$  = gjennomsnittet av de 2 midterste observasjonene etter at alle observasjonene er ordnet i stigende rekkefølge:

$$(1.6) \quad m = \frac{1}{2} \left( x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)} \right) \quad (n \text{ like})$$

**Eks. 1.5** Beregning av median (lign. (1.5))

$$\begin{aligned} x_1 &= 2,7, \quad x_2 = 1,9, \quad x_3 = 4,2, \quad x_{(1)} = 1,9, \quad x_{(2)} = 2,7, \quad x_{(3)} = 4,2 \\ \text{median: } m &= x_{\left(\frac{3+1}{2}\right)} = x_{(2)} = 2,7 \quad \circledcirc \end{aligned}$$

**Eks. 1.6 Beregning av median (lign.(1.6))**

$$x_1 = 2,7, \quad x_2 = 1,9, \quad x_3 = 4,2, \quad x_4 = -97,3$$

$$x_{(1)} = -97,3, \quad x_{(2)} = 1,9, \quad x_{(3)} = 2,7, \quad x_{(4)} = 4,2$$

$$\text{median: } m = \frac{1}{2}(x_{(4/2)} + x_{(4/2+1)}) = \frac{1}{2}(x_{(2)} + x_{(3)}) = 2,3 \quad \odot$$

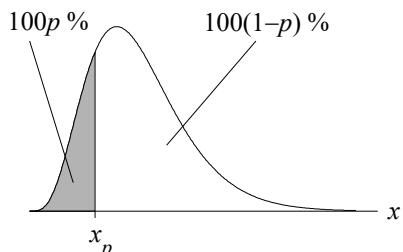
Beregn selv middelverdien,  $\bar{x}$ , i eks. 1.5 og 1.6, og du vil se at du får stor forskjell i de 2 tilfellene, p.g.a. den «ville» verdien,  $-97,3$ . Medianen,  $m$ , gir imidlertid som du ser noenlunde samme resultat i begge tilfeller. Vi sier at medianen er **robust** med hensyn til «ville» (ekstreme) verdier, dvs. lar seg lite påvirke av disse.

**100p-prosentilen,  $x_p$  (definisjon)**

100p-prosentilen er en  $x$ -verdi som er slik at minst  $100p\%$  av observasjonene er mindre eller lik, og minst  $100(1-p)\%$  av observasjonene er større eller lik denne  $x$ -verdien. En enkel beregningsmåte som tilfredsstiller definisjonen over er:

$$x_p = \begin{cases} \frac{1}{2}(x_{(np)} + x_{(np+1)}), & np \text{ heltall} \\ x_{(j)}, & np \text{ ikke heltall} \end{cases}$$

der  $j$  er minste heltall større enn  $np$ .


**Eks. 1.7 Beregning av prosentiler**

$$n = 5, \quad x_{(1)} = 1, \quad x_{(2)} = 2,3, \quad x_{(3)} = 3,4, \quad x_{(4)} = 3,7, \quad x_{(5)} = 4,9$$

Beregning av 20-prosentilen:

$$p = 0.2: \quad np = 5 \cdot 0.2 = 1 \text{ (heltall)}$$

$$\Rightarrow x_{0.2} = \frac{1}{2} \cdot \{x_{(1)} + x_{(2)}\} = \frac{1}{2} \cdot (1 + 2,3) = \underline{1,65}, \text{ dvs. 20-prosentilen er lik 1,65.}$$

Beregning av 70-prosentilen:

$p = 0,7$ :  $np = 5 \cdot 0,7 = 3,5$  (ikke heltall)  $\Rightarrow$  forhøyer  $np$  til nærmeste heltall, dvs.  $j = 4 \Rightarrow x_{0,7} = x_{(4)} = \underline{3,7}$ , dvs. 70-prosentilen er lik 3,7. ☺

### Kvartiler (definisjon):

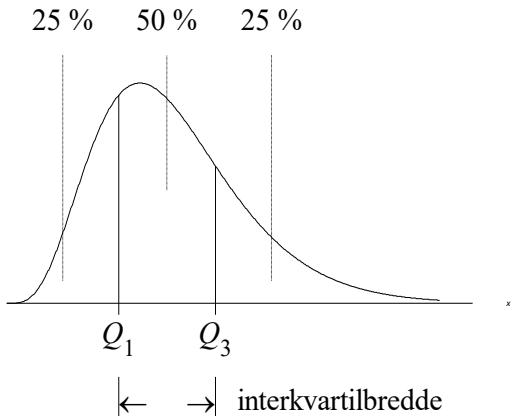
**Nedre kvartil**,  $Q_1$ , er identisk med 25-prosentilen (*en kvart* av dataene er mindre eller lik 25-prosentilen).

**Øvre kvartil**,  $Q_3$ , er identisk med 75-prosentilen (*tre kvart* av dataene er mindre eller lik 75-prosentilen).

**Medianen** er identisk med 50-prosentilen (*halvparten* av dataene er mindre eller lik medianen).

### Interkvartilbredde, $Q_3 - Q_1$

Interkvartilbredden er et (robust) mål på hvor spredt dataene er rundt medianen (midtverdien). Den finnes ved å ta differansen mellom øvre kvartil,  $Q_3 = x_{0.75}$ , og nedre kvartil,  $Q_1 = x_{0.25}$ , slik det er illustrert i figuren til høyre.



Fordelen med interkvartilbredden i forhold til standardavviket er at interkvartilbredden er lite påvirket av noen få «ville» observasjonsverdier.

**NB!** Dersom vi ikke har «ville» verdier vil normalt interkvartilbredden gi en større verdi enn standardavviket. For normalfordelingen, som er en svært sentral og viktig statistisk fordeling (kap.5), er interkvartilbredden ca. en faktor på 1.35 større enn standardavviket.

#### Eks. 1.8 Interkvartilbredde

Data:  $n = 5$ ,  $x_1 = -1,2$ ,  $x_2 = 0,7$ ,  $x_3 = 0$ ,  $x_4 = 172$ ,  $x_5 = 1,0$

Vi rangordner dataene:  $x_{(1)} = -1,2$ ,  $x_{(2)} = 0$ ,  $x_{(3)} = 0,7$ ,  $x_{(4)} = 1,0$ ,  $x_{(5)} = 172$

$$\begin{aligned} Q_1: \quad np = 5 \cdot 0.25 = 1,25 &\Rightarrow Q_1 = x_{(2)} = 0 \\ Q_3: \quad np = 5 \cdot 0.75 = 3,75 &\Rightarrow Q_3 = x_{(4)} = 1.0 \\ \text{Interkvartilbredden} &= Q_3 - Q_1 = 1.0 - 0 = \underline{1.0} \quad \text{☺} \end{aligned}$$

Det bør understrekkes at eksemplet ovenfor er et teknisk eksempel som viser hvordan vi beregner interkvartilbredden rent matematisk. I praksis bør vi generelt ha langt flere data enn  $n = 5$  for at bruk av interkvartilbredde skal være fornuftig. Imidlertid viser eksemplet hvordan interkvartilbredden effektivt undertrykker den ekstreme verdien  $x = 172$ .

## 1.4 Grupperte data

For å oppnå en mer oversiktig og informativ fremstilling av et innhentet data-materiale, er det vanlig å gruppere tallmaterialet. Det skilles mellom *to* forskjellige typer variabler:

- 1) **Diskrete** variabler: Observasjonene kan kun ha visse (diskrete) verdier som er adskilt fra hverandre (eks: antall øyne i et terningkast).
- 2) **Kontinuerlige** variabler: Observasjonene kan ha hvilke som helst verdier innenfor et begrenset eller ubegrenset definisjonsområde (eks: tiden,  $t$ ).

**NB!** I prinsippet er det ofte en «glidende» overgang fra kontinuerlige til diskrete variabler. La oss som eksempel betrakte «pers’en» på 60m til studenter ved Høgskolen i Tromsø. I prinsippet er denne tida en kontinuerlig variabel som kan ha en hvilken som helst verdi mellom, la oss si, 6s og 15s. I praksis måles imidlertid tiden på nærmeste tidel eller hundredel. Vi har da med *diskrete* verdier å gjøre (6,0s, 6,1s, 6,2s,...,15,0s i tilfelle vi har tider på nærmeste tidel). Ved tilstrekkelig fin inndeling (diskretisering) vil det ikke gi noen praktisk forskjell om vi betrakter 60m-tidene som kontinuerlige eller diskrete.

Normal *fremgangsmåte* for gruppering av et tallmateriale bestående av enkelttall kan være:

### a) Bestemmelse av maksimum og minimum

Vi finner minste og største observasjonsverdi,  $x_{\min}$  og  $x_{\max}$ .

### b) Klasseinndeling

Vi deler  $x$ -området inn i  $k$  klasser, som regel med like brede *klasseintervaller* (dvs. lik *klassebredde*). Klassene må ikke overlappe hverandre, og tilsammen må klassene dekke alle verdier fra og med  $x_{\min}$  til og med  $x_{\max}$ .

Vi etterstreber å velge klasseintervaller der *klassebredden* (øvre klassegrense minus nedre klassegrense), *klassegrenser* og *klassemidtpunktet*,  $m_i$ , blir forholdsvis pene og runde tall, eller ligger i nærheten av slike. Dette gjør vi både for å lette regnearbeidet, og for å få mest mulig brukvennlige og informative tabeller.

**NB!** Angivelse av *klassegrensene* i en tabell er ofte ikke i samsvar med de *egentlige* klassegrensene. Eks: Dersom en vektklasse er angitt som [60,70> kg, så er nedre og øvre klassegrense angitt som henholdsvis 60 og 70 kg. Normalt vil vektdata være forhøyet: Hvis en person veier f.eks. 59,5 kg, og vekta skal angis i hele kg, forhøyes vekta til 60 kg. De egentlige klassegrensene i intervallet [60,70> er derfor 59,500.. og 69,499...

Øvre *klassegrense* i en klasse er lik nedre klassegrense i neste klasse.

For entydig å angi i hvilken klasse en observasjon som ligger i grenseland mellom 2 naboklasser tilhører, benyttes 1 av følgende 3 teknikker:

1) Vi føyer på en ekstra desimal i klassegrensene i tillegg til antall desimaler i observasjonene.

2) Vi bruker forskjellig parentes ved nedre og øvre klassegrense,

Eks: < 5 , 10 ] betyr fra og med 6 til og med 10.

3) Vi lar det være et «sprang» mellom angivelse av øvre klassegrense i en klasse og nedre grense i neste klasse. *Klassebredden* for en klasse blir da forskjellen på nedre klassegrense i neste klasse og nedre klassegrense i klassen selv.

Eks: 5-9 betyr en klasse fra og med 5 til og med 9, neste klasse blir 10-14 (antar samme klassebredde), og klassebredden blir  $10 - 5 = 5$ .

### c) (Frekvens-) tabell

Vi lager en tabell med god plass til mange kolonner, og med plass til like mange rekker som det er antall klasser, pluss en summeringsrekke.

*Første kolonne* angir de ulike klasseintervallene, med de laveste verdiene først (øverst) og deretter klasser med stigende verdier.

Hva som skal stå i de neste kolonnene vil avhenge noe av oppgaven. Aktuelle kandidater er:

*Klassemidtpunkt*,  $m_i$ , dvs. midtpunktet mellom de *egentlige* klassegrensene.

*Tellekolonne:* Du merker av en strek i riktig klasserubrikk for hver av dine  $n$  observasjoner (tilsammen  $n$  streker).

*Frekvens-kolonne ( $f_i$ ):* Du angir antall observasjoner (f.eks. ved opptelling fra tellekolonnen) innenfor hver klasse. Dette kalles *klasseyfrekvensen*.

*Relativ frekvens-kolonne ( $f_i/n$ ):* Du tar klasseyfrekvensen og deler på  $n$ . Summen av alle relative frekvenser skal alltid være lik 1 ( $\pm$  avrundingsfeil).

*Kumulativ frekvens-kolonne ( $F_i$ ):* Sum av klasseyfrekvensene fra og med den første klassen til og med den klassen du ser på. I siste klasse vil kumulativ frekvens alltid være lik  $n$  (dvs. antall data).

*Relativ kumulativ frekvens-kolonne ( $F_i/n$ ):* Du tar kumulativ klasseyfrekvens og deler på  $n$ . I siste klasse skal da alltid relativ kumulativ frekvens være lik 1 ( $\pm$  avrundingsfeil).

$m_i f_i$ -kolonne: For hver klasse beregner du produktet av klassemidtpunktet,  $m_i$ , og klasseyfrekvensen,  $f_i$ . Hensikt: Beregne gruppert middelverdi.

$m_i^2 f_i$ -kolonne: For hver klasse beregner du produktet av kvadratet av klassemidtpunktet,  $m_i^2$ , og klasseyfrekvensen,  $f_i$ . Hensikt: Beregne gruppert standardavvik.

*Rektangelhøyde:* Relativ frekvens dividert på klassebredde. Hensikt: Finne høyden på hvert rektangel ved fremstilling av relativ frekvens-histogram (defineres litt senere).

#### d) Senter og spredning

På basis av de grupperte data kan vi f.eks. beregne *gruppert middelverdi*,  $x_g$ , og *gruppert standardavvik*,  $s_g$ , som mål på henholdsvis senter og spredning, etter oppskriften i neste ramme. Vi kan også beregne gruppert median,  $m_g$ , grupperte 100 $p$ -prosentiler og gruppert interkvartilbredde, slik angitt i neste ramme.

Bemerk at vi i uttrykket for gruppert standardavvik i neste ramme har  $n$  og ikke  $n-1$  i nevner. Når vi har gruppert våre data har vi allerede fjernet så mye detaljinformasjon om rådataene, at forutsetningene for å velge  $n-1$  ikke lenger er tilstede.

### Gruppert middelverdi, $\bar{x}_g$

$$(1.9) \quad \bar{x}_g = \frac{1}{n} \sum_{i=1}^k m_i \cdot f_i$$

### Gruppert standardavvik, $s_g$ :

$$(1.10) \quad s_g = \sqrt{\frac{1}{n} \sum_{i=1}^k (m_i - \bar{x}_g)^2 f_i} = \sqrt{\frac{1}{n} \sum_{i=1}^k m_i^2 f_i - \bar{x}_g^2}$$

der  $m_i$  er klassemidtpunkt og  $f_i$  er frekvens til klasse nr.  $i$  ( $i = 1, \dots, k$ ).

*Gruppert varians,  $s_g^2$ , er identisk med uttrykket under rottegnnet.*

### Gruppert median, $m_g$

$$(1.11) \quad m_g = x_1 + \frac{n/2 - F_1}{f_m} \cdot \Delta x_m$$

der

$x_1$  = nedre klassegrense i medianklassen<sup>1</sup>

$F_1$  = kumulativ frekvens i klassen *forut for* medianklassen

$f_m$  = frekvens til medianklassen

$\Delta x_m$  = klassebredden til medianklassen

<sup>1</sup>Medianklassen er den første klassen der kumulativ frekvens,  $F_m$ , er større enn  $n/2$ .

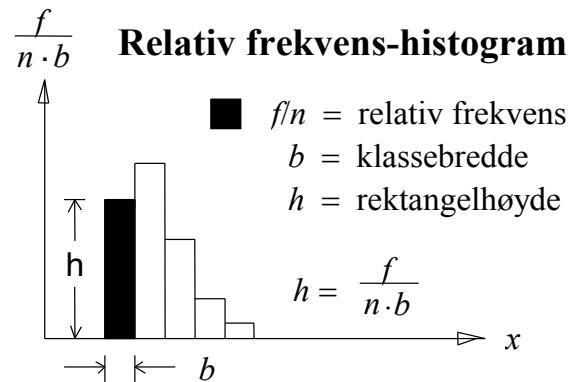
(Gruppert 100 $p$ -prosentil,  $x_{gp}$ , finnes ved å erstatte  $n/2$  i formelen ovenfor med  $np$ , og medianklassen med « $p$ -klassen». Gruppert interkvartilbredde =  $x_{g0,75} - x_{g0,25}$ ).

## e) Grafisk fremstilling

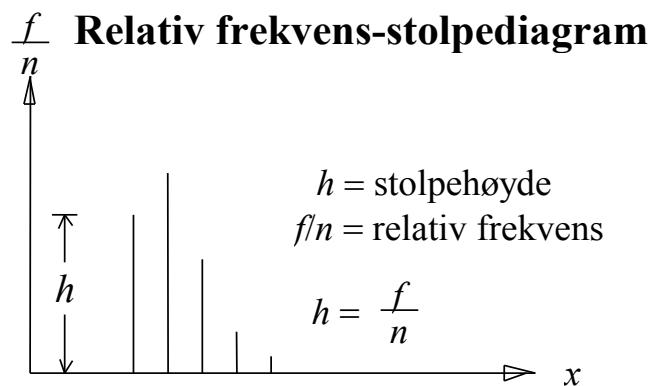
De grupperte dataene fremstilles gjerne grafisk. Vi skal her betrakte 2 vanlige grafiske presentasjonsformer: *Relativ frekvens-histogram* og *relativ frekvens-stolpediagram* (det er også mange andre ord for stolpediagram: Linjediagram,

stavdiagram, pinnediagram m.m.). Forskjellen på de 2 presentasjonsformene skulle gå fram av det følgende.

*Relativ frekvens-histogram* består av rektangler: Et rektangel pr. klasse og rektangelareal lik relativ klasefrekvens. Bredden av hvert rektangel er lik klassebredden. *Høyden av hvert rektangel* ( $y$ -verdien i  $x,y$ -diagrammet) blir derfor *relativ klasefrekvens dividert på klassebredden*.



Relativ *frekvens-stolpediagram* benyttes kun for diskrete variabler. Da tegnes en loddrett stolpe for hver diskrete verdi der vi har observasjoner. Høyden på hver stolpe ( $y$ -verdien) er normalt lik relativ frekvens.



**Eks. 1.9 Hovedeksempel, beskrivende statistikk for én variabel**

Følgende høydedata for studenter ved Høgskolen i Tromsø er gitt:

190	184	180	180	178	171	180	179	176	188
180	170	184	189	178	181	182	178	165	182
180	174	176	178	187	166	191	185	183	180
180	172	186	175	190	168	170	160	176	182
176	176								

*Oppgave*

- a) Finn middelverdi,  $\bar{x}$ , og standardavvik,  $s$ , på basis av rådataene.
- b) Rangordne dataene og bestem medianen og interkvartilbredden.
- c) Gruppér dataene og bestem gruppert middelverdi, gruppert standardavvik og gruppert median.
- d) Fremstill de grupperte høydedataene i et relativ frekvens-histogram.

*Løsningsforslag*

- a) Vi beregner først middelverdi,  $\bar{x}$ , på basis av rådataene (se lign.(1.1)):

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i = \frac{1}{42} (190 + 184 + 180 + \dots + 176) \text{ cm} = \frac{7506}{42} \text{ cm} = 178.7 \text{ cm}$$

Vi beregner så standardavviket,  $s$ , på basis av rådataene (se lign.(1.4)):

$$s^2 = \frac{1}{42-1} \left( \sum_{i=1}^{42} x_i^2 - \frac{1}{n} \left( \sum_{i=1}^{42} x_i \right)^2 \right) \text{ cm}^2 = \frac{1}{41} \left( 1343458 - \frac{7506^2}{42} \right) \text{ cm}^2 \\ \approx 49.48 \text{ cm}^2 \Rightarrow s = \sqrt{49.48} \text{ cm} = 7.0 \text{ cm}$$

- b) Vi rangordner dataene og får (stigende rekkefølge fra venstre mot høyre):

160	165	166	168	170	170	171	172	174	175
176	176	176	176	176	178	178	178	178	179
180	180	180	180	180	180	180	181	182	182
182	183	184	184	185	186	187	188	189	190
190	191								

Vi bestemmer nå medianen,  $m$ , på basis av de rangordnede dataene ovenfor. Her er  $n = 42$  som er et like tall, og vi benytter lign.(1.6):

$$m = \frac{1}{2} (x_{(42/2)} + x_{(42/2+1)}) = \frac{1}{2} (x_{(21)} + x_{(22)}) = \frac{1}{2} (180 + 180) \text{ cm} = 180 \text{ cm}$$

For å finne interkvartilbredden, må vi først finne nedre og øvre kvartil,  $Q_1$  og  $Q_3$ , respektive (se eks. 1.8).

$Q_1 = 25\text{-prosentilen (nedre kvartil):}$

$p = 0.25$ ,  $np = 42 \cdot 0.25 = 10.5$ , som ikke er et heltall. Forhøyer da til nærmeste heltall:  $j = 11$ , og får:  $Q_1 = x_{0.25} = x_{(11)} = 176$

$Q_3 = 75\text{-prosentilen (øvre kvartil):}$

$p = 0.75$ ,  $np = 42 \cdot 0.75 = 31.5$ , som ikke er et heltall. Forhøyer da til nærmeste heltall:  $j = 32$ , og får:  $Q_3 = x_{0.75} = x_{(32)} = 183$

$$\Rightarrow \text{interkvartilbredden} = Q_3 - Q_1 = 183 \text{ cm} - 176 \text{ cm} = \underline{7 \text{ cm}}$$

c) Dataene skal nå grupperes, og vi følger «oppskriften»:

$x_{\min} = x_{(1)} = 160$ ,  $x_{\max} = x_{(42)} = 191$ . Vi velger derfor *første* høyde-klassen med *nedre intervallgrense høyst* lik 160, og *siste* høydeklasse med *øvre klassegrense minst* lik 191. Vi foretar følgende skjønnsmessige valg:

Vi velger én og samme klassebredde lik 4 for alle klasser, med nederste klassegrense lik 160 og øverste klassegrense lik 191. Klassemidtpunktene blir da 161.5, 165.5, 169.5, ... osv. dersom vi antar at høydetallene er forhøyet (eks: 184.5 cm forhøyes til 185 cm). Vi kan lage følgende tabell, der tellekolonnen er fornuftig å ha dersom vi skal gruppere dataene fra råmaterialet *før* det er rangordnet. I vårt tilfelle er det lettere å fylle ut frekvenskolonnen i tabellen direkte fra de rangordnede dataene.

Høyde [cm]	fre- kvens	relativ frekv.	midt- pkt.		kum. frekv.	rel. kum.	rekt. høyde $\frac{f_i}{n\Delta x}$
$f_i$	$f_i/n$	$m_i$	$m_i f_i$	$m_i^2 f_i$	$F_i$	$F_i/n$	
160–163	1	1/42	161.5	161.5	26082.25	1	1/168
164–167	2	2/42	165.5	331.0	54780.5	3	3/168
168–171	4	4/42	169.5	678.0	114921	7	7/168
172–175	3	3/42	173.5	520.5	90306.75	10	10/168
176–179	10	10/42	177.5	1775	315062.5	20	20/168
180–183	12	12/42	181.5	2178	395307	32	32/168
184–187	5	5/42	185.5	927.5	172051.25	37	37/168
188–191	5	5/42	189.5	947.5	179551.25	42	5/168
Sum:	42	1	7519	1348062.5			

Vi finner gruppert middelverdi og standardavvik ved å benytte henholdsvis lign.(1.9) og (1.10), og benytte verdiene fra tabellen over:

$$\bar{x}_g = \frac{1}{n} \sum_{i=1}^k m_i f_i = \frac{1}{42} \cdot 7519 \text{ cm} = \underline{179.0 \text{ cm}}$$

Legg merke til at vi får nesten samme resultat som for de ugrupperte dataene, på tross av at vi har gruppert dataene i bare  $k = 8$  klasser og tatt utgangspunkt i disses midtverdier, mot opprinnelig 42 tall.

Vi beregner så gruppert standardavvik:

$$s_g^2 = \frac{1}{n} \sum_{i=1}^k m_i^2 f_i - \bar{x}_g^2 = \frac{1}{42} \cdot 1348062.5 \text{ cm}^2 - \left(\frac{7519}{42}\right)^2 \text{ cm}^2 = 47.20 \text{ cm}^2$$

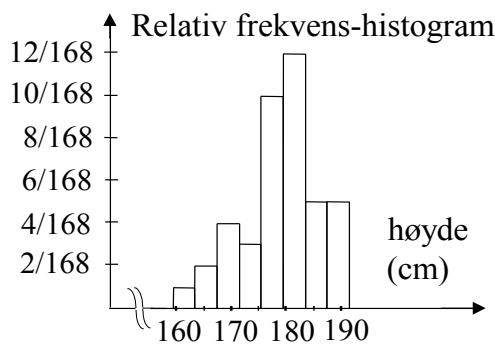
$$\Rightarrow s_g = \sqrt{47.20} \text{ cm} = \underline{6.9 \text{ cm}}$$

Også her får vi som vi ser godt samsvar med resultatene fra de ugrupperte dataene ( $s = 7,0 \text{ cm}$ ).

Deretter beregner vi gruppert median. Vi finner først medianklassen.  $n/2 = 42/2 = 21 \Rightarrow$  medianklassen er den første klassen med kumulativ frekvens større enn 21. Vi ser at dette er klasse nummer 6 ovenfra i tabellen, med kumulativ klasselfrekvens  $F_m = 32$ , klasselfrekvens  $f_m = 12$  og klassebredde  $\Delta x_m = 4$ . Videre er kumulativ klasselfrekvens i klassen forut for medianklassen lik  $F_1 = 20$ . Nedre klassegrense i medianklassen skal her regnes som  $x_1 = 179.5$ . Vi får da i henhold til lign. (1.11):

$$m_g = x_1 + \frac{n/2 - F_1}{f_m} \cdot \Delta x_m = 179.5 \text{ cm} + \frac{21-20}{12} \cdot 4 \text{ cm} = \underline{179.8 \text{ cm}}$$

Som vi ser, er også dette nær medianen bestemt på basis av ugruppert datamateriale ( $m = 180 \text{ cm}$ ).



d) Vi skal til slutt fremstille de grupperte dataene i et relativ frekvenshistogram.

Siste kolonne i tabellen viser høyden på rektanglene.

Merk at totalt areal av alle histogramstolpene blir 1, eller 100%.

Høydedataene er lagt inn på statistikkprogrampakken **Minitab**. Sammenlign resultatene fra Minitab, gjengitt nedenfor, med de som er utført ovenfor.

MTB > describe 'hoyde'.

### Descriptive Statistics

Variable	N	Mean	Median	StDev
hoyde	42	178.71	180.00	7.03

Variable	Min	Max	Q1	Q3
hoyde	160.00	191.00	175.75	183.25

MTB > histogram 'hoyde';  
 SUBC> start 161.5;  
 SUBC> increment 4.

### Character Histogram

Histogram of hoyde N = 42

Midpoint	Count
161.50	1 *
165.50	2 **
169.50	4 ****
173.50	3 ***
177.50	10 *****
181.50	12 *****
185.50	5 ****
189.50	5 ****

### MINITAB

Høydedataene er lagt inn på kolonnen kalt 'hoyde'.

«MTB >» er Minitab prompt, og det som følger etter er kommandoer lagt inn av bruker.

Hovedkommando legges inn av bruker etter «MTB >», og avsluttes med «;» dersom underkommandoer skal angis.

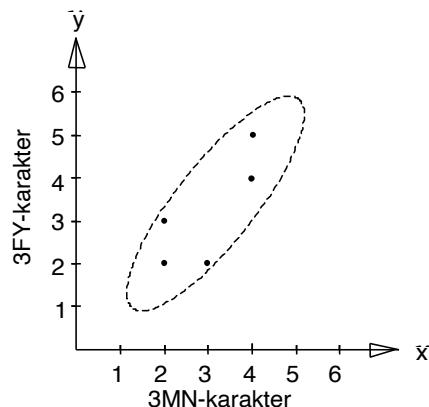
Kommandoene avsluttes med «.» tilslutt.

Utskriften til venstre er noe redigert. ☺

## 1.5 Spredningsdiagram

Hittil har vi sett på *enkelttall*. I de resterende avsnittene skal vi se på *tallpar*. La oss belyse denne forskjellen ut fra tabellen nedenfor over samhørende 3MN-(matematikk-) og 3FY-(fysikk-) karakterer til 5 tilfeldig valgte studenter. Dataene er tegnet inn i  $(x,y)$ -diagrammet til høyre, som er et eksempel på et *spredningsdiagram*.

Student nr:	3MN- karakter, $x$	3FY- karakter, $y$
1	3	2
2	2	2
3	4	5
4	2	3
5	4	4



Matematikkarakterene utgjør et tallmateriale bestående av enkelttall. Det samme gjør fysikkarakterene. Vi kan bruke teknikkene fra tidligere til å beregne f.eks. middelverdi og standardavvik til 3MN-karakterene såvel som til 3FY-karakterene. Slike mål sier imidlertid ingenting om hvordan  $x$ - og  $y$ -verdiene *samvarierer*.

Ser vi på matematikk- og fysikkarakteren til en og samme student samlet, får vi et *tallpar* for hver student. Tilsammen får vi følgende 5 tallpar:

$$(3,2), (2,2), (4,5), (2,3), (4,4)$$

Det er disse som er tegnet inn i spredningsdiagrammet ovenfor. Med tallpar kan vi belyse en del problemstillinger knyttet til hvordan  $x$ - og  $y$ -verdiene samvarierer:

- er det noen form for systematisk sammenheng mellom  $x$ - og  $y$ -verdiene?
- kan vi tallfeste i hvor stor grad det er en slik sammenheng?
- kan vi tilpasse en fornuftig funksjon som beskriver sammenhengen mellom  $x$ - og  $y$ -verdiene?
- kan vi forutsi en variabel dersom den andre er kjent?

Å studere enten  $x$ -målingene for seg selv eller  $y$ -målingene for seg selv vil ikke være til hjelp når det gjelder å svare på disse spørsmålene. Som vi ser fra spredningsdiagrammet ovenfor, er det en klar systematisk sammenheng mellom  $x$ - og  $y$ -verdiene: En student med god/dårlig karakter i 3MN ser ut til jevnt over å ha en rimelig god/dårlig karakter i 3FY (*positiv korrelasjon*). Vi har ytterligere understreket denne sammenhengen med den stiplete ellipsen.

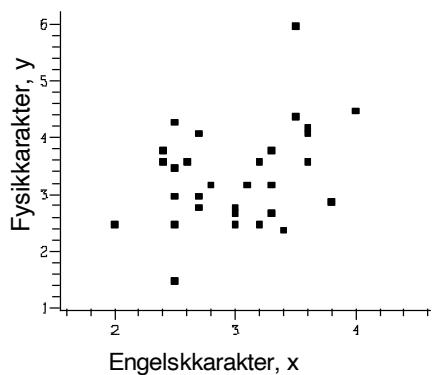
Å fremstille tallparene grafisk i form av et spredningsdiagram er et viktig første skritt i studiet av sammenhengen mellom 2 variabler. Vi avmerker (som det går fram av vårt innledningsdiagram)  $x$ -verdien langs horisontalaksen og den tilhørende  $y$ -verdien langs vertikalaksen. Parene  $(x,y)$ , som består av observasjoner (målinger), blir da plottet som grafiske punkter. Det resulterende diagram er kalt et spredningsdiagram. Ved å se på spredningsdiagrammet, kan vi få et visuelt inntrykk av en eventuell (systematisk) sammenheng mellom variablene. For eksempel kan vi observere hvorvidt punktene ligger som et bånd rundt en rett linje, rundt en krum kurve eller om de simpelthen danner en mønsterløs samling av tilsvynelatende tilfeldig spredte verdier.

**NB!** Ved tegning av spredningsdiagrammer bør skaleringen langs aksene (antall enheter pr. cm) velges slik at utstrekningen av  $x_{\text{maks}} - x_{\text{min}}$  i cm er omtrent like stor som utstrekningen av  $y_{\text{maks}} - y_{\text{min}}$  i cm.

**Eks. 1.10**
**Sammenheng mellom fysikk- og engelskkarakterer**
 $x = \text{engelskkarakter}, y = \text{fysikkarakter}$ 

Tabellen nedenfor viser sammenhengen mellom engelsk- ( $x$ ) og fysikk- ( $y$ ) karakterene til  $n = 30$  studenter.

$x$	$y$								
3,3	3,8	2,5	3,5	3,3	3,2	2,5	3,0	3,4	2,4
2,7	4,1	3,0	2,5	3,1	3,2	3,6	4,1	3,3	2,7
2,6	3,6	2,7	3,0	2,5	1,5	3,6	4,2	2,4	3,6
2,8	3,2	4,0	4,5	2,5	2,5	3,5	4,4	2,4	3,8
3,2	3,6	3,5	6,0	2,0	2,5	3,2	2,5	2,5	4,3



Spredningsdiagrammet er vist i figuren til venstre. Sørvest til nordøst-mønsteret som punktene danner indikerer en positiv sammenheng mellom  $x$  og  $y$ : Studentene med gode/ dårlige karakterer i engelsk tenderer mot å ha gode/ dårlige karakterer i fysikk. Sammenhengen mellom  $x$  og  $y$  er imidlertid opplagt ikke gitt ved noen pen matematisk funksjon. ☺

## 1.6 Empirisk korrelasjonskoeffisient

Spredningsdiagrammet gir et visuelt inntrykk av sammenhengen mellom  $x$ - og  $y$ -verdiene i et tallmateriale som består av tallpar (bivariat datasett). I mange tilfeller synes punktene å ligge i et bånd rundt en rett linje. I varierende grad vil imidlertid tilfeldige variasjoner utelukke en perfekt lineær (rettlinjet) sammenheng.

Den empiriske **korrelasjonskoeffisienten**, som vi skal betegne med  $r$ , er et mål på graden av lineær sammenheng mellom  $x$ - og  $y$ -variablene. Før vi introduserer formelen for  $r$ , skal vi angi noen viktige egenskaper ved korrelasjonskoeffisienten, og diskutere på hvilken måte den kan brukes til å måle graden av lineær sammenheng.

- a)  $r$ -verdiene ligger alltid mellom  $-1$  og  $1$ :  $-1 \leq r \leq 1$
- b) Absoluttverdien til  $r$  indikerer graden av lineær sammenheng, mens fortegnet indikerer retning. Mer presist:

**$r > 0$**  hvis mønsteret til  $(x,y)$  verdiene er et bånd som løper fra nedre venstre til øvre høyre hjørne

**$r < 0$**  hvis mønsteret til  $(x,y)$  verdiene er et bånd som løper fra øvre venstre til nedre høyre hjørne

**$r = 1$**  hvis alle  $(x,y)$  verdiene ligger eksakt på en og samme rette linje med et positivt stigningstall

**$r = -1$**  hvis alle  $(x,y)$  verdiene ligger eksakt på en og samme rette linje med et negativt stigningstall

c) En  $r$ -verdi i nærheten av null betyr at det er liten grad av lineær sammenheng.

- a)  $r = -0,9$    b)  $r = -0,5$    c)  $r = 0,0$
- 
- d)  $r = 0,9$    e)  $r = 0,5$    f)  $r = 0,0$

Figur: Eksempler på spredningsdiagram og tilhørende  $r$ -verdi

Korrelasjonskoeffisienten er nærmest null når det ikke er noe synlig mønster på sammenheng, dvs.  $y$ -verdiene synes ikke å variere i noen foretrukket retning når  $x$ -verdiene varierer. En  $r$ -verdi i nærheten av null kan også innstrefte fordi punktene ligger i et bånd rundt en kurve som er alt annet enn en rett linje. Husk at  $r$  er et mål på lineær sammenheng, og en svært bøyd kurve er langt fra lineær.

Forrige figur viser sammenhengen mellom ulike spredningsdiagram og den tilhørende  $r$ -verdien. Legg merke til at c) og f) begge tilsvarer situasjoner der  $r = 0$ . Null korrelasjon i Fig. 1.1c) skyldes fraværet av enhver sammenheng mellom  $x$  og  $y$ , mens null korrelasjon i Fig. 1.1f) skyldes at sammenhengen følger en (ikke-lineær) kurve som er mer eller mindre symmetrisk om middelverdien til  $x$ -verdiene.

### Empirisk korrelasjonskoeffisient, $r$

$r$ -verdien beregnes utifra  $n$  par av observasjoner  $(x_1, y_1), \dots, (x_n, y_n)$ , ved følgende formel:

$$r = \frac{S_{xy}}{\sqrt{S_x^2} \sqrt{S_y^2}}$$

der

$$\begin{aligned} S_x^2 &= \sum (x_i - \bar{x})^2 = \sum x_i^2 - \frac{1}{n} (\sum x_i)^2 \\ S_y^2 &= \sum (y_i - \bar{y})^2 = \sum y_i^2 - \frac{1}{n} (\sum y_i)^2 \\ S_{xy} &= \sum (x_i - \bar{x})(y_i - \bar{y}) = \sum x_i y_i - \frac{1}{n} (\sum x_i)(\sum y_i) \end{aligned}$$

og alle summer er fra  $i = 1$  til  $i = n$ .  $S_x^2$  og  $S_y^2$  er summene av kvadratiske avvik (fra middelverdien) til henholdsvis  $x$ -verdiene og  $y$ -verdiene.  $S_{xy}$  er summen av kryssproduktene mellom  $x$ -avvik og  $y$ -avvik fra de respektive middelverdier.

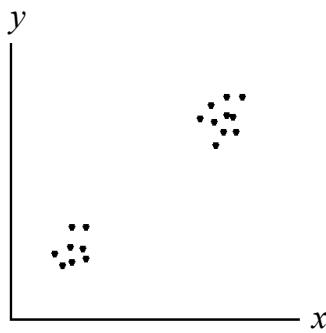
**Eks. 1.11** Beregne  $r$  for følgende  $n = 3$  par av observasjoner:  
 $(x,y) = (3,1), (1,0)$  og  $(8,2)$

$x$	$y$	$x^2$	$y^2$	$xy$	
3	1	9	1	3	
1	0	1	0	0	
8	2	64	4	16	
Tot:	12	3	74	5	19
		$= \Sigma x$	$= \Sigma y$	$= \Sigma x^2$	$= \Sigma y^2$
					$= \Sigma xy$

$$r = \frac{19 - \frac{12 \cdot 3}{3}}{\sqrt{74 - \frac{12^2}{3}} \cdot \sqrt{5 - \frac{3^2}{3}}} = 0,97$$

☺

Vi minner leseren om at  $r$  måler hvor nær mønsteret til spredningen er en rett linje. Tilfelle f) i forrige figur presenterer en stor grad av samvariasjon mellom  $x$  og  $y$ , men en som ikke er lineær. Den lave verdien til  $r$  for disse data reflekterer ikke den store graden av ikke-lineær samvariasjon.



En annen situasjon der den empiriske korrelasjonskoeffisient  $r$  ikke er god, opptrer når spredningsplottet er delt i 2 adskilte punktsamlinger. I slike tilfeller kan det være best å forsøke å bestemme den underliggende årsak. Det kan f. eks. være at en del av utvalget kommer fra en populasjon mens en annen del kommer fra én annen populasjon (Populasjonsbegrepet vil bli nærmere definert i et senere kapittel).

Figur:  $r \approx 1$ , se tekst.

## Korrelasjon og årsak

**NB!** Det kan være lett å mistolke en observert korrelasjon ( $r$  i nærheten av  $-1$  eller  $+1$ ) mellom to variabler som et årsaksforhold mellom variablene. Et klassisk eksempel er at en har observert en høy positiv korrelasjon mellom antall storker og antall barnefødsler i europeiske byer. Årsaken til dette er ikke at babyene kommer med storken, men at det er en tredje variabel som ligger og «lurer» i bakgrunnen: Størrelsen på byene. Jo større byer, jo flere storker, og jo større byer jo flere babyer. Det er altså bystørrelsen som får antall storker og antall babyer til å variere i samme retning.

Observasjonen at 2 variabler synes å samvariere i en bestemt retning medfører altså ikke nødvendigvis at det er et direkte årsaksforhold mellom variablene. Det kan være variasjonen i en tredje variabel, som forårsaker at  $x$  og  $y$  varierer i samme retning, selv om de er uten sammenheng, eller til og med har motsatt sammenheng av den som indikeres av korrelasjonskoeffisienten. Den falske korrelasjonen som fremkommer på denne måten kan vi kalle *villedende* korrelasjon. Det er mer sunn fornuft enn statistisk resonnement som skal til for å bestemme hvorvidt en observert korrelasjon kan bli tolket praktisk eller om den er villedende.

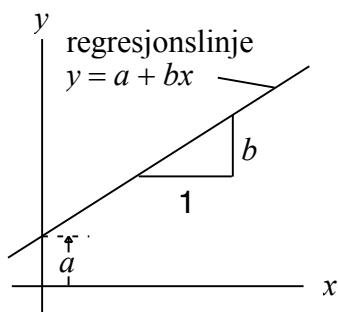
**Advarsel:** En observert korrelasjon mellom 2 variabler kan være villedende. Dvs. at den kan være forårsaket av en tredje variabel. Når vi bruker korrelasjonskoeffisienten som mål på sammenheng, må vi være oppmerksomme på muligheten for at en lurevariabel påvirker noen av variablene vi betrakter.

## 1.7 Lineær regresjon

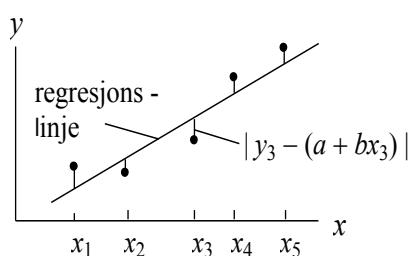
Studier av sammenhengen mellom to variabler ved målinger er ofte motivert ut fra et behov for å kunne forutsi den ene variabelen fra den andre. En leder for et jobbtreningsprogram kan ønske å studere sammenhengen mellom varigheten av treningen, og resultatet av treningen ved en påfølgende test. En skogeier kan ønske å anslå (estimere) tømmervolumet til et tre fra måling av stamme-diameteren 1 meter over bakken. En forsker innen medisin kan være interessert i å forutsi alkoholinnholdet i blod ut fra målinger fra et nylig oppfunnet pusteapparat.

I slike sammenhenger som disse, er det vanlig å la  $x$  betegne den *uavhengige* variabelen, også kalt *inn-variabelen*, og la  $y$  betegne *responsen*, eller *ut-variabelen*. Formålet er å finne hvilken form for sammenheng det er mellom  $x$  og  $y$  fra eksperimentelle data, og å bruke denne sammenhengen til å prediktere (forutsi) responsen til variabelen  $y$  (*responsvariabelen*) fra inn-variabelen,  $x$  (*prediktoren*).

Første skritt er å plotte og undersøke spredningsdiagrammet. Hvis en lineær sammenheng fremkommer, så vil beregningen av korrelasjonskoeffisienten,  $r$ , bekrefte styrken av lineær sammenheng.  $r$ -verdien indikerer hvor effektivt  $y$  kan forutsies fra  $x$  ved å tilpasse en rett linje til dataene.



En linje  $y = a + bx$  er bestemt ved to konstanter: Høyden over origo (skjæringspunktet med  $y$ -aksen),  $a$ , og stigningstallet,  $b$ , dvs. mengden  $y$  øker med når  $x$  øker med en enhet (se figur til venstre).



Vi ønsker å bestemme de verdier  $a^*$  og  $b^*$  for henholdsvis  $a$  og  $b$  som gjør at regresjonslinja  $y = a + bx$  er best mulig tilpasset våre observasjoner/data. Et mye benyttet prinsipp er her minste kvadraters metode. Prinsippet går ut på å minimere kvadratsummen

$$Q = \sum_{i=1}^n (y_i - (a + bx_i))^2$$

med hensyn på  $a$  og  $b$ , dvs. summen av kvadratene av «vertikal» avstand fra punktene til linja, se forrige figur. Matematisk partilderiverer vi  $Q$  med hensyn på  $a$  og  $b$ , setter de to ligningene vi får lik null, og løser disse med hensyn på  $a$  og  $b$ . Resultatet er gjengitt i neste ramme. Vi skal gå grundigere til verks i kap. 9 om lineær regresjon.

Eks. 1.12 viser et praktisk eksempel, som også belyser hva som menes med prediksjon ut fra den tilpassete rette linja.

### Rett linje-tilpasning, $y^* = a^* + b^* \cdot x$

Vi har et tallmateriale  $(x_1, y_1), \dots, (x_n, y_n)$  og skal tilpasse en rett linje  $y^* = a^* + b^* \cdot x$ . Minste kvadraters metode gir da følgende formler til å bestemme  $a^*$  og  $b^*$ :

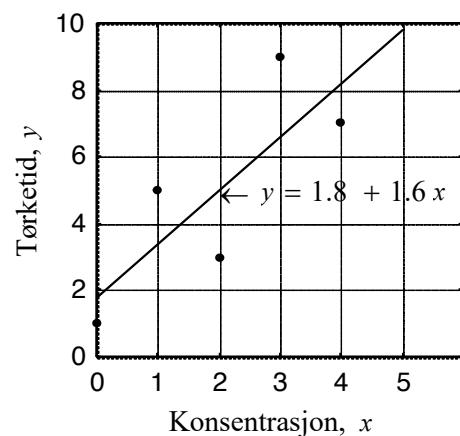
$$b^* = \frac{S_{xy}}{S_x^2}, \quad a^* = \bar{y} - b^* \cdot \bar{x}$$

der  $\bar{x}$  og  $\bar{y}$  er middelverdiene til henholdsvis  $x$ - og  $y$ -verdiene, og  $S_x^2$  og  $S_{xy}$  er definert i ramme tidligere i kapitlet.

**Eks. 1.12****Maling og tørketid.**

En kjemiker ønsker å studere sammenhengen mellom tørketiden til en maling og konsentrasjonen av en basisk oppløsning som gjør det lettere å male. Data for konsentrasjonen ( $x$ ) og den tilsvarende observerte tørketid ( $y$ ) er gitt i de 2 første kolonnene i følgende tabell:

Konsen- trasjon, $x$	Tørke- tid, $y$	$x^2$	$y^2$	$xy$
0	1	0	1	0
1	5	1	25	5
2	3	4	9	6
3	9	9	81	27
4	7	16	49	28
tot:	10	25	30	66



Figur: Data for konsentrasjon,  $x$ , og tørketid,  $y$  (i minutter), og beregninger for regresjonslinja. Spredningsdiagram og tilpasset regresjonslinje til høyre.

Spredningsdiagrammet i figuren over gir et inntrykk av en viss grad av lineær sammenheng. For å beregne  $r$  og bestemme ligningen for den tilpassede linja, beregner vi først  $\bar{x}$ ,  $\bar{y}$ ,  $S_x^2$ ,  $S_y^2$  og  $S_{xy}$  fra tabellen ovenfor:

$$\bar{x} = \frac{10}{5} = 2, \quad \bar{y} = \frac{25}{5} = 5$$

$$S_x^2 = 30 - \frac{10^2}{5} = 10, \quad S_y^2 = 165 - \frac{25^2}{5} = 40, \quad S_{xy} = 66 - \frac{10 \cdot 25}{5} = 16$$

$$r = \frac{16}{\sqrt{40 \cdot 10}} = 0.8, \quad b^* = \frac{16}{10} = 1.6, \quad a^* = 5 - 1.6 \cdot 2 = 1.8$$

Ligningen for den tilpassede linja blir da:

$$y^* = 1.8 + 1.6x$$

Linja er vist i spredningsdiagrammet i figuren ovenfor. Dersom vi ønsker å forutsi tørketiden  $y$  som tilsvarer konsentrasjonen 2.5, setter vi bare inn for  $x = 2.5$  i ligningen over og får resultatet:

Predikert tørketid for  $x = 2.5$  er  $y = 1.8 + 1.6 \cdot 2.5 = 5.8$ , dvs. 5.8 min. Grafisk finner vi denne verdien ved å lese av på  $y$ -aksen den verdien vår tilpassede linje har når  $x = 2.5$ . Merk imidlertid at dette anslaget er ganske usikkert, vi har ikke en svært sterk lineær sammenheng i dette tilfellet. ☺

I kap. 9 skal vi se nærmere på hvordan vi kan tallfeste usikkerheten til prediksjoner.

## 1.8 Oppgaver

**1.1** Beregn empirisk middelverdi og standardavvik av følgende tall:

- a)  $-1, 3, 8, 4, 13$
- b)  $-8, -9, -14, -11, 0$

**1.2** Gitt følgende tallmateriale:

1.43438 1.43443 1.43442 1.4344

a) La tallene ovenfor være  $x$ -verdier, og finn de tilsvarende  $z$ -verdier ved transformasjonen

$$z = (x - 1.4344) / 0.00001$$

b) Beregn empirisk middelverdi,  $\bar{z}$ , og standardavvik,  $s_z$ , til  $z$ -verdiene.

c) Beregn middelverdi,  $\bar{x}$ , og standardavvik,  $s_x$ , til  $x$ -verdiene ved transformasjonen  $s_x = 0,00001s_z$  og

$$\bar{x} = 1.4344 + 0.00001 \cdot \bar{z}$$

**1.3** Følg fremgangsmåten i oppgave 1.2 og beregn middelverdi og standardavvik til tallene

1221.3 1220.9 1221.7 1220.8 1221

**1.4** I en bedrift med 14 ansatte er gjennomsnittsinntekten kr 158 000 pr. ansatt. Hvor store lønnsutgifter har bedriften?

**1.5** a) Finn empirisk middelverdi og median for følgende tilfeldig målte ute-temperaturer ( $^{\circ}\text{C}$ ) som en person har notert i sin dagbok i løpet av ett år:

$x: 14, -12, 0, 16, 7$

b) En annen person har følgende måleresultater fra sin dagbok (samme sted og samme år):

$y: 14, -12, 0, 16, 7, 112$

Beregn empirisk middelverdi og median også i dette tilfellet. Ville du stole mest på middelverdien eller medianen?

**1.6** Gitt følgende  $x$ -verdier:

1, -3, 7, 12, 6

Beregn 20- og 70-prosentilen.

**1.7** Gitt  $x$ -verdiene:

21, 17, 18, 17, 22, 21, 78, 22, 24

Beregn standardavvik og interkvartilbredde. Hvilket av de to målene på spredning ville du stole mest på, dersom du i ettermiddag fikk opplyst at en av verdiene var feil?

1.8 Tabellutsnittet til Vekt [kg]
:
[10-20>
[20-30>
[30-40>
:

**1.9** Neste tabell viser daglig inntekt av CD-salg i en musikkforretning, fordelt på 4 inntektgrupper. Anta at et dags-salg på kr. 9999 kommer i den første klassen.

Videosalg (kr 1000)	Frekvens (dager)
0-9	37
10-19	148
20-29	123
30-39	57

a) Bestem klassemidtpunkt i hver klasse.

- b) Bestem gruppert middelverdi.
- c) Bestem gruppert standardavvik.
- d) Bestem gruppert median.
- e) Bestem gruppert interkvartilbredde.

### 1.10 Gitt tallparene

(0.9 , 1.1), (2.1 , 1.8) og (2.9 , 3.3)

- a) Velg 1 cm/enhet langs  $x$ - og  $y$ -aksen, og lag et spredningsdiagram.
- b) Beregn korrelasjonskoeffisienten,  $r$ .
- c) Beregn en regresjonslinje tilpasset dataene, og tegn den inn i spredningsdiagrammet.

### 1.11 Gitt følgende samhørende $x$ - og $y$ -verdier:

$x$	.68	.52	.58	.72	.62
$y$	11	5	1	4	6

- a) Beregn korrelasjonskoeffisienten,  $r$ .
- b) Beregn en regresjonslinje tilpasset dataene,  $y^* = a^* + b^*x$ .
- c) Tegn  $(x,y)$ -verdiene i et spredningsdiagram med 1 cm pr. enhet **både** langs  $x$ - og  $y$ -aksen, og tegn inn regresjonslinja. Hvorfor står linja «på tvers» av dataene?
- d) Gjør c) på nytt, men velg selv en fornuftig skaling langs aksene.

### 1.12 Størrelsen av en dyrebestand måles en gang i året. De 4 siste årene ble følgende observert:

år:	87	88	89	90
bestand:	123	237	471	982

La  $x = 1, 2, 3$  og  $4$  betegne årstallene 87, 88, 89 og 90, og la  $y$  betegne de tilsvarende bestandene.

- a) Beregn korrelasjonskoeffisienten.

- b) Utfør transformasjonen  $z = \ln y$ , og beregn  $z$ -verdiene.
- c) Bestem korrelasjonskoeffisienten for  $(x,z)$ -verdiene.
- d) Tilpass en rett linje til  $(x,z)$ -dataene.
- e) Bestem på basis av d) en regresjonsfunksjon,  $y = ce^{dx}$ , tilpasset  $(x,y)$ -dataene.

### 1.13 (Vekt-data i tabell lenger ned).

- a) Finn middelverdi og standardavvik til vekt-dataene på basis av rådataene.
- b) Rangordne vekt-dataene og bestem median og interkvartilbredde.
- c) Gruppér vekt-dataene og bestem gruppert middelverdi, gruppert standardavvik og gruppert median.
- d) Fremstill de grupperte vekt-dataene i et relativ frekvens-histogram.

### 1.14 (data i tabell lenger ned).

- a) Tegn et spredningsdiagram for samhørende vekt- og høyde-data.
- b) Beregn korrelasjonskoeffisienten for vekt- og høyde-dataene, og kommenter spredningsdiagrammet i a).
- c) Beregn korrelasjonskoeffisienten for høyde- og alders-dataene, og kommenter resultatet.

*Tabell for oppgave 1.13 og 1.14*

høyde [cm]	vekt [kg]	alder [år]	høyde [cm]	vekt [kg]	alder [år]	høyde [cm]	vekt [kg]	alder [år]
167	55	32	180	64	24	185	83	25
169	59	23	175	73	21	160	54	26
163	55	25	170	52	20	172	75	22
177	70	23	172	65	21	167	58	19
180	76	24	170	66	23	190	82	25
174	61	25	175	60	19	172	66	21
172	60	19	185	90	23	185	74	21
174	59	19	178	67	21	182	69	21
174	80	21	187	85	28	181.5	82	21
182	55	31	185	82	27	176	72	21
158	50	23	183	72	19	187	85	28
168	60	21	185	70	19			

**1.15** I et utvalg på 20 fire-barnsfamilier er antall gutter følgende:

3 2 3 2 3 1 3 3 2 3  
1 3 0 3 2 1 4 2 3 2

- a) Regn ut middelverdien for utvalget.
- b) Regn ut utvalgsstandardavviket.
- c) Sett opp en fordelingstabell for de absolutte og relative frekvensene.
- d) Bestem median og interkvartilbredde.
- e) Lag relativ frekvens stolpediagram.

**1.16** Aldersfordelingen i en bedrift er:

Alder [år]	15- 25	25- 35	35- 45	45- 55	55- 65
Frekvens	2	7	6	4	2

- a) Tegn histogram over aldersfordelingen.
- b) Beregn aritmetisk middelverdi, standardavvik og median.
- c) Avmerk de beregnede størrelsene i histogrammet.

**1.17** En måleserie resulterte i følgende klassedelte observasjonsmateriale:

intervall	frekvens
[195, 200>	2
[200, 205>	4
[205, 210>	10
[210, 215>	11
[215, 220>	9
[220, 225>	4

- a) Utvid tabellen med relative frekvenser, kumulative frekvenser og relative kumulative frekvenser.
- Lag et histogram for observasjonsmaterialet.
- b) Beregn tilnærmet aritmetisk middelverdi og standardavvik for observasjonsmaterialet.
- c) Beregn tilnærmet median og interkvartilbredde for observasjonsmaterialet.

## 1.9 Formelsamling

### Ugrupperte data

#### Empirisk middelverdi

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i = \frac{1}{n} \cdot (x_1 + \dots + x_n)$$

#### Empirisk standardavvik

$$\begin{aligned} s &= \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2} \\ &= \sqrt{\frac{1}{n-1} \left( \sum x_i^2 - \frac{1}{n} (\sum x_i)^2 \right)} \\ &= \sqrt{\frac{1}{n-1} (\sum x_i^2 - n\bar{x}^2)} \end{aligned}$$

#### Empirisk median

$$m = \begin{cases} \frac{1}{2} (x_{(n/2)} + x_{(n/2+1)}), & n \text{ like} \\ x_{(\frac{n+1}{2})}, & n \text{ odde} \end{cases}$$

#### Empirisk 100p-prosentil

$$x_p = \begin{cases} \frac{1}{2} (x_{(np)} + x_{(np+1)}), & n \text{ like} \\ x_{(j)}, & n \text{ odde} \end{cases}$$

der  $j$  er minste heltall større enn  $np$ .

#### Interkvartilbredde

$$Q_3 - Q_1 = x_{0.75} - x_{0.25}$$

### Grupperte data

#### Betegnelser

$k$  er antall klasser,  $n$  er totalt antall observasjoner,  $m_i$  er klassemidtpunkt,  $f_i$  er klassefrekvens,  $F_i$  er kumulativ klassefrekvens i klasse nr.  $i$ ,  $i = 1, \dots, k$ .

#### Gruppert middelverdi

$$\bar{x}_g = \frac{1}{n} \sum_{i=1}^k m_i \cdot f_i$$

#### Gruppert standardavvik

$$s_g = \sqrt{\frac{1}{n} \sum_{i=1}^k m_i^2 \cdot f_i - \bar{x}_g^2}$$

#### Gruppert median

$$m_g = x_l + \frac{n/2 - F_l}{f_m} \cdot \Delta x_m$$

der  $x_l$  er nedre klassegrense i medianklassen,  $F_l$  er kumulativ frekvens i klassen *forut* for medianklassen,  $f_m$  er frekvensen til medianklassen og  $\Delta x_m$  er bredden til medianklassen.

### Bivariate data

#### Summasjonsvariabler

$$S_x^2 = \sum_{i=1}^n (x_i - \bar{x})^2 = \sum x_i^2 - \frac{1}{n} (\sum x_i)^2$$

$$S_y^2 = \sum_{i=1}^n (y_i - \bar{y})^2 = \sum y_i^2 - \frac{1}{n} (\sum y_i)^2$$

$$S_{xy} = \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$$

$$= \sum x_i y_i - \frac{1}{n} \sum x_i \sum y_i = \sum x_i y_i - n \bar{x} \cdot \bar{y}$$

#### Empirisk korrelasjonskoeffisient

$$r = \frac{S_{xy}}{S_x S_y} = \frac{S_{xy}}{\sqrt{S_x^2} \cdot \sqrt{S_y^2}}$$

#### Regresjonslinje

$$y^*(x) = a^* + b^* \cdot x, \text{ der}$$

$$b^* = S_{xy} / S_x^2, \quad a^* = \bar{y} - b^* \cdot \bar{x}$$

## Kapittel 2

# Sannsynlighetsregning

## 2.1 Innledning

*Sannsynlighetsbegrepet* er et svært sentralt begrep innen statistikk og sannsynlighetsregning. Det er også et vanskelig begrep. Prøv selv, før du leser videre, å definere hva *du* mener med sannsynlighet.

Vi skal begrense oss til tilfeller der vi på en objektiv måte kan *tallfeste* sannsynligheten for at en hendelse skal inntrefte. En slik hendelse er for eksempel «å få kron i et tilfeldig myntkast med rettferdig kronestykke». De fleste vil være enige i at sannsynligheten for at dette skal skje er 50 %. Ved litt ettertanke vil de fleste også være enige i at det vi mener med 50 % i dette tilfellet er at i en lang rekke med myntkast vil ca. halvparten gi kron som utfall. Dette er tankegangen bak det **frekventistiske** sannsynlighetsbegrepet, som vi skal begrense oss til. Ordet frekventistisk henspeiler her på at man tenker seg et forsøk gjentatt en rekke ganger under identiske betingelser, og at man beregner den relative frekvensen for en gitt hendelse.

Det engelske ordet probability betyr sannsynlighet, og vi bruker bokstaven  $P$  for å betegne sannsynlighet.  $P = 0$  angir at hendelsen ikke kan inntrefte, for eksempel at Ola vil være i Norge og Malaysia samtidig.  $P = 1$  (100 %) angir hendelser som *helt sikkert* inntreffer, for eksempel at dersom Kari går så beveger bena hennes seg.

Legg merke til at vi som regel må finne ganske sære eksempler på hendelser som enten ikke kan inntrefte ( $P = 0$ ) eller er helt sikre ( $P = 1$ ). For alle andre hendelser ligger  $P$  mellom 0 og 1. Dersom vi ønsker å angi sannsynligheter i %, multipliserer vi  $P$  med 100 %.

Mange studenter, kanskje de fleste, vil synes at sannsynlighetsregningen er noe av det vanskeligste i et kurs i grunnleggende statistikk og sannsynlighetsregning. Det er måten å tenke på som er spesiell. Trenger du gjennom denne «muren», åpner det seg en verden av artige problemstillinger å gyve løs på. For de som er glad i spill, er det nok å nevne eksempler som sannsynligheten for ulike gevinst i Lotto eller ulike korthender i bridge og andre kortspill. Du vil også bli i stand til å takle ganske så sammensatte problemstillinger.

De ulike avsnittene i dette kapitlet hører nøyne sammen. Vær derfor grundig med å få med deg en mest mulig presis forståelse av begrepene underveis. Og som

ellers i matematisk statistikk: Nøl ikke før du gir deg i kast med oppgave-regning.

## 2.2 *Utfallsrom, enkeltutfall og hendelse*

Et (statistisk) *eksperiment* er prosessen med å samle data for et fenomen der utfallet varierer (statistisk variasjon). I tilknytning til eksperimenter er de viktige begrepene utfallsrom, enkeltutfall og hendelser definert som følger:

### **Utfallsrom, enkeltutfall og hendelse** (definisjon)

*Utfallsrommet* er samlingen av alle mulige forskjellige (distinkte) utfall av et eksperiment. Hvert av disse utfallene skal vi kalle *enkeltutfall*. Vi benytter symbolet  $S$  for utfallsrommet, og symbolene  $e_1, \dots, e_n$  for enkelt-utfallene i  $S$ .

En *hendelse* er en samling av enkeltutfall og utgjør et delområde av utfallsrommet  $S$ . Det er vanlig å benytte de første store bokstavene i alfabetet:  $A, B, C, \dots$  for å betegne hendelser. Vi sier at en hendelse  $A$  inntreffer når ett eller flere av enkeltutfallene i  $A$  inntreffer.

La oss belyse de begrepene vi har introdusert med noen eksempler. Vi betrakter følgende 4 eksperiment:

- Eksperiment (a): Noter *kjønnet* til de 2 første spebarn som fødes imorgen.
- Eksperiment (b): La hver av 10 personer smoke et glass Cola og et glass Cola Light, og noter *hvor mange* som foretrekker Cola Light.
- Eksperiment (c): Gi et antibiotikum til pasienter som lider av en virusinfeksjon inntil en pasient får en *antireaksjon*.
- Eksperiment (d): Mål *luftfuktigheten i prosent* i ei badstu.

Vi får følgende utfallsrom:

$S_a = \{ GG, GJ, JG, JJ \}$ ,  $G =$  Gutt,  $J =$  jente,  $GJ$  betyr *først* gutt, *så* jente, osv.  
En slik kronologisk rekkefølge fra venstre mot høyre er ganske vanlig.

$S_b = \{ 0, 1, 2, \dots, 10 \}$

$S_c = \{ A, NA, NNA, NNNA, \dots \}$ ,  $A =$  Antireaksjon,  $N =$  Ingen reaksjon

$S_d = \{ x : 0 \% \leq x \leq 100 \% \}$ , som leses: «Mengden av alle reelle verdier av  $x$  fra og med 0 % til og med 100 %»

Legg merke til at kun ett av de angitte utfall kan opptre av gangen. Vi kan imidlertid ofte beskrive en *hendelse* der flere enkeltutfall kommer innenfor definisjonen av hendelsen:

**Eks. 2.1** **Eksakt ei jente.** Betrakt eksperiment (a), og la  $A$  betegne hendelsen «eksakt ei jente». Vi får da at  $A = \{ GJ, JG \}$ . ☺

**Eks. 2.2** **Høyst 5 pasienter.** Betrakt eksperiment (c), og la  $C$  betegne hendelsen «høyst 5 pasienter får antibiotika før siste pasient får antireaksjon». Vi får da at  $C = \{ A, NA, NNA, NNNA, NNNNA \}$ . ☺

Et utfallsrom er *diskret* når det består av et *endelig* eller *tellbart uendelig* antall enkeltutfall. Utfallsrommene for eksperimentene (a), (b) og (c) er diskrete. (a) og (b) er eksempler med et endelig antall enkeltutfall, mens eksperiment (c) er et eksempel med et tellbart uendelig antall enkeltutfall. Når et utfallsrom inneholder alle tall i et intervall på den reelle tallinja er utfallsrommet *kontinuerlig*. Eksperiment (d) er et eksempel med kontinuerlig utfallsrom.

## 2.3 Sannsynlighet for en hendelse

Vi skal bruke betegnelsen  $P(A)$  for å betegne sannsynligheten for at en hendelse A skal inntrefte ( $P$  er forkortelse for det engelske ordet «Probability» som betyr sannsynlighet). Et eksempel: La  $A$  betegne hendelsen «kronsida opp i tilfeldig kast med en rettferdig mynt». Intuitivt vil de aller fleste da være med på at det er «50 % sannsynlighet for å få kron», eller: Det er «50 % sannsynlighet for at hendelsen A skal inntrefte»,  $P(A) = 0.5$  ( $0.5 = 0.5 \cdot 100 \% = 50 \%$ ).

Tenker vil litt igjennom hva vi mener med «50 % sannsynlighet» i dette tilfellet, vil de fleste være enige i at det som ligger i underbevisstheten er følgende: Den relative andelen kron i en serie med myntkast vil mer og mer nærme seg 50 % ettersom antall kast øker.

Foreløpig skal vi holde oss til følgende intuitive oppfatning av begrepet *sannsynlighet*: Den andelen  $A$  inntreffer i en serie av forsøk under identiske forhold. ( $\text{Sannsynlighet} \approx \text{antall forsøk med «gunstige» utfall} / \text{antall forsøk totalt}$  når antall forsøk er svært stor).

Vi skal først betrakte sannsynlighetsbegrepet i forbindelse med hva vi kaller en *uniform sannsynlighetsmodell*.

### **Uniform sannsynlighetsmodell** (definisjon)

Hvis et utfallsrom består av  $k$  enkeltutfall  $\{e_1, \dots, e_k\}$ , der hvert enkeltutfall inntrer med like stor sannsynlighet, så er sannsynligheten for hvert enkeltutfall lik  $1/k$ . Hvis en hendelse  $A$  består av  $m$  av disse  $k$  enkeltutfallene, består den uniforme sannsynlighetsmodellen i at vi kan sette

$$(2.1) \quad P(A) = m/k$$

Når et eksperiment innrettes slik at sannsynligheten for hvert enkeltutfall er like stor, kan vi anvende en *uniform sannsynlighetsmodell*.

I vårt myntkasteksempel har vi at  $k = 2$  (2 like sannsynlige enkeltutfall: Kron og mynt), og  $m = 1$  (kun ett enkeltutfall gir kron), dvs.  $P(A) = P(\text{«kron»}) = m/k = 1/2$ .

#### **Eks. 2.3** Terningkast med rettferdig terning.

Eksperiment: Antall øyne i ett tilfeldig kast.

Utfallsrom  $S = \{e_1, e_2, e_3, e_4, e_5, e_6\}$

Enkeltutfall:

$$\begin{aligned} e_1: & 1 \text{ øye} & e_2: & 2 \text{ øyne} & e_3: & 3 \text{ øyne} & e_4: & 4 \text{ øyne} & e_5: & 5 \text{ øyne} & e_6: & 6 \text{ øyne} \\ P(e_1) = P(e_2) = \dots = P(e_6) &= 1/6 \end{aligned}$$

Siden alle enkeltutfallene er like sannsynlige, har vi her et eksempel der en uniform sannsynlighetsmodell er riktig. Legg igjen merke til at kun ett av enkeltutfallene i utfallsrommet,  $S$ , kan opptre av gangen. Et terningkast med én terning kan for eksempel ikke gi både 2 og 5 i ett og samme kast. ☺

**Eks. 2.4**  $A = \text{«antall øyne er et like tall»}$ ,  
dvs.  $A = \{e_2, e_4, e_6\} \Rightarrow P(A) = P(e_2) + P(e_4) + P(e_6) = 3/6 = 1/2$

Siden vi her kan legge en uniform sannsynlighetsmodell til grunn, hadde det vært tilstrekkelig å telle opp det antall enkeltutfall som  $A$  består av (her:  $m = 3$ ), og dele på totalt antall enkeltutfall i utfallsrommet (her:  $k = 6$ ):  $P(A) = m/k = 3/6 = 1/2$ . ☺

Vi går nå løs på en mer generell definisjon av sannsynlighetsbegrepet knyttet til begrepet relativ frekvens, og vi skal sette opp betingelser for en sannsynlighetsmodell basert på diskrete utfallsrom.

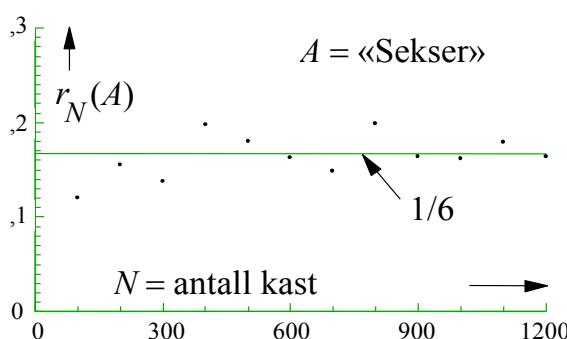
### Relativ frekvens, $r_N(A)$ (definisjon)

Relativ frekvens av en hendelse  $A$  i  $N$  forsøk («trials»),  $r_N(A)$ , er definert som følger:

$$(2.2) \quad r_N(A) = \frac{\text{Antall ganger } A \text{ inntreffer i løpet av } N \text{ forsøk}}{\text{Antall forsøk } N}$$

Her er det underforstått at forsøkene (eks: terningkast) er foretatt under identiske betingelser.

Tenk igjen på terningkasteksemplet. Si for eksempel at  $N = 6$  og  $A = \text{«sekser»}$ . Vi vil da kanskje få én sekser i løpet av de 6 forsøkene, dvs.  $r_N = 1/6$ . Det er imidlertid ikke usannsynlig at vi får 2 ( $r_N = 2/6$ ) eller ingen ( $r_N = 0$ ) seksere på 6 forsøk. Øker vi antall forsøk ( $N = 12, 24, 100, 1000, \dots$ ), forventer vi imidlertid at  $r_N$  (relativ andel seksere) vil stabilisere seg rundt  $1/6$ . En slik betraktnign legger vi til grunn for definisjonen av sannsynlighet  $P(A)$  for en hendelse  $A$  (se neste ramme).



I figuren til venstre er vist et eksempel der relativ frekvens av en hendelse  $A$  stabiliserer seg når antall ( $N$ ) forsøk øker. De svarte punktene er tilfeldige forsøk. Legg merke til at stabiliseringen går ganske sakte: Selv med over 1000 forsøk, er det en «synlig» usikkerhet rundt den teoretiske verdien  $1/6$ .

### Sannsynlighet, $P(A)$ (definisjon)

Vi bruker betegnelsen  $P(A)$  for å betegne sannsynligheten for en hendelse  $A$ , og vi skal benytte følgende definisjon:

$$(2.3) \quad P(A) = \lim_{N \rightarrow \infty} r_N(A)$$

dvs.  $P(A)$  er lik den grenseverdien som relativ frekvens  $r_N(A)$  går mot når antall eksperimenter  $N$  går mot uendelig. Det er her forutsatt at grenseverdien konvergerer mot en fast verdi.

Legg merke til den nære sammenhengen mellom begrepene *relativ frekvens* og *sannsynlighet*. Denne nære sammenhengen er svært sentral i statistikken, og vi skal kun se på det frekventistiske sannsynlighetsbegrep i denne boka.

I neste ramme har vi satt opp *betingelser* for den *sannsynlighetsmodell* vi skal anvende, basert på *diskrete* utfallsrom. Merk at vi definerer sannsynlighetene til å ligge mellom 0 og 1, hvilket tilsvarer 0 til 100 %.

### Sannsynlighetsmodell, diskrete utfallsrom

Sannsynlighet er en funksjon, definert for hendelser, som tilfredsstiller følgende betingelser:

- i) For alle hendelser,  $A$ , gjelder at  $0 \leq P(A) \leq 1$   
 $P(A) = 0$  betyr at hendelsen  $A$  helt sikkert *ikke* kan inntreffe.  
 $P(A) = 1$  betyr at hendelsen  $A$  *helt sikkert* vil inntreffe.
- ii)  $P(A)$  er summen av sannsynlighetene for at hvert av enkeltutfallene som tilhører  $A$  skal inntreffe:  

$$(2.4) \quad P(A) = \sum_{e_i \in A} P(e_i)$$

der summen skal tas over alle enkeltutfall  $e_i$  som er med i  $A$ .
- iii)  $P(S)$  er summen av sannsynlighetene til alle enkeltutfallene i hele utfallsrommet,  $S$ , og denne må være lik 1:  

$$(2.5) \quad P(S) = \sum_{e_i \in S} P(e_i) = 1$$

der summen skal tas over alle enkeltutfall  $e_i$  som er med i  $S$ .

## 2.4 Union, snitt og komplement

I dette avsnittet skal vi gjøre en del bruk av Venn-diagram. Vi kommer ikke til å gi noen formell definisjon av Venn-diagram, men det vil gjennom eksempler og oppgaver klart gå fram hva det dreier seg om. Enkelt sagt er et Venn-diagram et lukket område i planet som omslutter hele utfallsrommet,  $S$ , og der hendelser tegnes inn som lukkede delområder av utfallsrommet. Enkeltutfall markeres gjerne med punkter i diagrammet.

**Eks. 2.5** **Surströmming 1.** En tilfeldig valgt student blir tvangsført med surströmming. Studenten responderer på 2 måter. Enten liker han/hun surströmming ( $J$ ), eller så liker han/hun ikke surströmming ( $N$ ).

Eksperiment: Surströmmingen gis til 3 tilfeldig utvalgte studenter, og responsen til hver student registreres.

*Oppgave*

Lag et Venn-diagram for utfallsrommet,  $S$ , og angi følgende hendelser:

$A$ : Bare én student liker surströmming

$B$ : Første student liker surströmming

$C$ : Både første og andre student liker ikke surströmming

*Løsningsforslag*

La oss først finne alle enkeltutfallene i utfallsrommet for eksperimentet. For den første studenten er det 2 mulige resultater ( $J$  eller  $N$ ), og hvert av disse resultatene blir etterfulgt av 1 av 2 muligheter ( $J$  eller  $N$ ) for student nr.2. Hver av disse  $2 \cdot 2 = 4$  kombinerte resultatene kan bli etterfulgt av en  $J$  eller en  $N$  for den tredje studenten. Totalt får vi derfor  $n = 2 \cdot 2 \cdot 2 = 8$  mulige enkeltutfall,  $e_1, \dots, e_8$ :

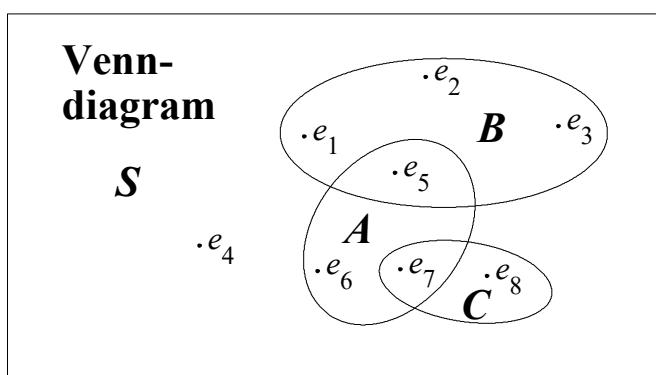
$$\begin{array}{llll} JJJ(e_1) & JJN(e_2) & JNJ(e_3) & NJJ(e_4) \\ JNN(e_5) & NJN(e_6) & NNJ(e_7) & NNN(e_8) \end{array}$$

Hendelsene  $A$ ,  $B$  og  $C$  består av enkeltutfallene:

$$A = \{ e_5, e_6, e_7 \}$$

$$B = \{ e_1, e_2, e_3, e_5 \}$$

$$C = \{ e_7, e_8 \}$$



Venn-diagrammet er vist i figuren til venstre, der rektangelet representerer utfallsrommet,  $S$ .

De ulike hendelsene omslutter akkurat de enkeltutfallene som er med i hendelsen. ☺

Vi skal nå definere 3 viktige operasjoner med hendelser, nemlig *union*, *snitt* og *komplement*:

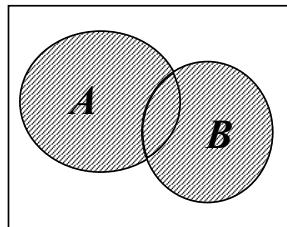
## Union, snitt og komplement

**Unionen** av 2 hendelser  $A$  og  $B$  betegnes  $A \cup B$  og er mengden av alle enkeltutfall som er med i  $A$ , eller i  $B$ , eller i både  $A$  og  $B$ .

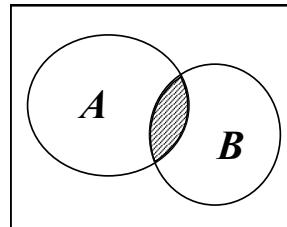
**Snittet** mellom 2 hendelser  $A$  og  $B$  betegnes  $A \cap B$ , eller kortere,  $AB$ , og er mengden av alle enkeltutfall som er med i både  $A$  og  $B$ .

**Komplementet** til en hendelse  $A$  betegnes med  $A^C$  og er mengden av alle enkeltutfall som ikke er med i  $A$ .

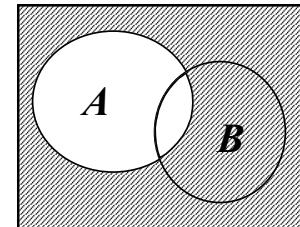
De definerte begreper er illustrert med skraverte områder nedenfor.



Union  $A \cup B$



Snitt  $AB$



Komplement  $A^C$

**Eks. 2.6** **Surströmming 2.** La hendelsene  $A$ ,  $B$  og  $C$  være definert som i eks. 2.5. Angi hvilke enkeltutfall følgende hendelser består av:

$$A \cup B, AC, BC, B^C, (A \cup B)^C, (A^C) \cup (B^C)$$

### Løsningsforslag

Fra før:  $A = \{ e_5, e_6, e_7 \}$ ,  $B = \{ e_1, e_2, e_3, e_5 \}$ ,  $C = \{ e_7, e_8 \}$

Ved å bruke definisjonene for operasjonene union, snitt og komplement får vi:

$$A \cup B = \{ e_1, e_2, e_3, e_5, e_6, e_7 \}$$

$$AC = \{ e_7 \}$$

$$BC = \emptyset \text{ (Tom mengde)}$$

$$B^C = \{ e_4, e_6, e_7, e_8 \}$$

$$(A \cup B)^C = \{ e_4, e_8 \}$$

$B^C$  er gitt ovenfor, og  $A^C = \{ e_1, e_2, e_3, e_4, e_8 \}$

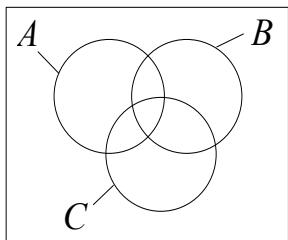
$$\Rightarrow A^C \cup B^C = \{ e_1, e_2, e_3, e_4, e_6, e_7, e_8 \} \quad \textcircled{R}$$

Vi har ovenfor betraktet snitt og union av 2 hendelser. Reglene kan lett generaliseres til flere enn 2 hendelser, f.eks.  $A \cup B \cup C$ ,  $A \cap B \cap C$  og  $(A \cup B) \cap C$ :

$$(2.6) \quad A \cup B \cup C = (A \cup B) \cup C = A \cup (B \cup C)$$

$$(2.7) \quad A \cap B \cap C = (A \cap B) \cap C = A \cap (B \cap C)$$

$$(2.8) \quad (A \cup B) \cap C = (A \cap C) \cup (B \cap C)$$



Figuren til venstre er til god hjelp for å overbevise seg om at lign.(2.6)-(2.8) er korrekte!

Et viktig spesialtilfelle når vi betrakter 2 hendelser  $A$  og  $B$  får vi når *ikke noe* enkeltutfall opptrer samtidig i de 2 mengdene.  $A$  og  $B$  sies da å være **gjensidig ekskluderende** hendelser, eller 2 hendelser som gjensidig utelukker hverandre. I dette tilfellet er  $AB = \emptyset$ , der  $\emptyset$  er betegnelsen for en tom mengde.

Vi går nå over til å se på 2 viktige *sannsynlighetslover* for hendelser:

### Sannsynlighetslover for hendelser

$$(2.9) \quad P(A^C) = 1 - P(A) \quad (\text{komplementærloven})$$

$$(2.10) \quad P(A \cup B) = P(A) + P(B) - P(AB) \quad (\text{addisjonsloven})$$

Som et spesialtilfelle av siste lov ovenfor får vi *addisjonsloven for unionen av gjensidig ekskluderende hendelser*:

$$(2.10b) \quad P(A \cup B) = P(A) + P(B)$$

når  $A$  og  $B$  er gjensidig ekskluderende hendelser ( $AB = \emptyset$ ).

Fra (2.10) kan vi utlede addisjonssetningen for 3 hendelser:

$$(2.10c) \quad P(A \cup B \cup C) = P(A) + P(B) + P(C) \\ - P(AB) - P(AC) - P(BC) + P(ABC)$$

**Eks. 2.7 Seriesystem**

Figuren nedenfor viser et seriesystem med to like og uavhengige komponenter der  $p$  = sannsynligheten for at en komponent svikter.

*Oppgave*

- Hva er sannsynligheten for at systemet funksjonerer?
- I en kjettinglenke med  $n$  løkker er sannsynligheten pr. år for at en løkke ryker konstant lik  $p$ , uavhengig av hvilken løkke vi betrakter. Hva er sannsynligheten for at kjettingen holder ett år?

*Løsningsforslag*

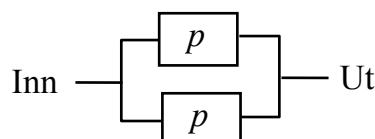
- La  $A$  betegne hendelsen at komponent 1 er OK og  $B$  at komponent 2 er OK.  
Vi får da:

$$P(\text{system OK}) = P(AB) = P(A) \cdot P(B) = (1-p) \cdot (1-p) = (1-p)^2$$

- Kjettinglenke med  $n$  løkker tilsvarer seriesystem med  $n$  komponenter. Genererer vi løsningen i a) får vi:  $P(\text{kjetting OK}) = (1-p)^n$ . ☺

**Eks. 2.8 Parallellysystem**

Figuren nedenfor viser et parallellysystem med to like og uavhengige komponenter der  $p$  = sannsynligheten for at en komponent svikter.

*Oppgave*

- Hva er sannsynligheten for at systemet funksjonerer?
- Hva er sannsynligheten for at et parallellysystem med  $n$  komponenter i parallel funksjonerer?

*Løsningsforslag*

- La  $A$  betegne hendelsen at komponent 1 er OK og  $B$  at komponent 2 er OK.  
Vi får da:

$$\begin{aligned} P(\text{system OK}) &= 1 - P(\text{begge komponenter svikter}) = 1 - P(A^C B^C) = \\ &= 1 - P(A^C) \cdot P(B^C) = 1 - p^2 \end{aligned}$$

- Med  $n$  komponenter i parallel får vi samme fremgangsmåte som i a), men må erstatte 2 med  $n$ :  $P(\text{system OK}) = 1 - p^n$  ☺

**Eks. 2.9** **Eksamensresultater i matematikk og fysikk**, karakterisert i sannsynlighetstabellen nedenfor.

*Oppgave*

Hvis en student trekkes tilfeldig fra studentmassen som ligger bak sannsynlighetene i tabellen nedenfor, hva er da sannsynligheten for a) studenten har godt resultat i minst ett av fagene, b) studenten har ikke dårlig resultat i matematikk?

		Fysikkresultater		
		Godt	Middels	Dårlig
Matematikk-resultater	Godt	.06	.11	.18
	Middels	.12	.18	.10
	Dårlig	.16	.05	.04

*Løsningsforslag*

a) Definer hendelsene  $A$  og  $B$  som følger:

$A$ : godt matematikkresultat,  $B$ : godt fysikkresultat

Vi får da:

$$P(\text{godt resultat i minst ett fag})$$

$$= P(A \cup B) = P(A) + P(B) - P(AB)$$

$$P(A) = .06 + .11 + .18 = .35$$

$$P(B) = .06 + .12 + .16 = .34$$

$$P(AB) = .06$$

$$\text{Følgelig: } P(\text{godt resultat i minst ett fag}) = .35 + .34 - .06 = .63$$

b) La  $D$  være hendelsen at studenten har dårlig matematikkresultat. Vi får da:

$$P(\text{ikke dårlig resultat i matematikk}) = P(D^C) = 1 - P(D)$$

$$P(D) = .16 + .05 + .04 = .25$$

$$\text{Følgelig: } P(\text{ikke dårlig resultat i matematikk}) = 1 - .25 = .75 \quad \odot$$

### Urnemodell

Vi tenker oss en urne som inneholder  $n$  «merkete lapper». Ved utvalg på  $k$  tilfeldige trekninger fra urna skal vi skille mellom:

- Trekning **uten** tilbakelegging, der vi aldri kan trekke én og samme lapp mer enn én gang, og trekning **med** tilbakelegging, der vi kan trekke samme lapp flere ganger.
- **Ordnede** utvalg, der rekkefølgen av trekningene er vesentlig, og **uordnede** utvalg, der rekkefølgen er uvesentlig.

Vi skal benytte de forkortete betegnelsene (uten, uordnet), (uten, ordnet), (med, uordnet) og (med, ordnet) for de fire kombinerte tilfellene vi kan ha. For eksempel betyr (med, uordnet) utvalg ved trekning med tilbakelegging der rekkefølgen av trekningene er uvesentlig.

## 2.5 Kombinatorikk, telleregler

Dersom utfallsrommet består av i alt  $k$  like sannsynlige enkeltutfall,  $P(e_1) = \dots = P(e_k) = 1/k$ , så kan vi anvende den uniforme sannsynlighetsmodellen. En hendelse,  $A$ , som består av i alt  $m$  av de  $k$  enkeltutfallene, vil da ha sannsynlighet  $P(A) = m/k$  for å inntreffe. Med denne modellen i bakhodet, skal vi studere 3 viktige telleregler:

- 1) **Potensregelen**
- 2) **Ordningsregelen**
- 3) **Kombinasjonsregelen**

Ved løsning av oppgaver i sannsynlighetsregning, vil det ofte være til hjelp å tenke seg en urne som vi trekker tilfeldige lapper fra. Før vi går løs på tellereglene ovenfor, skal vi derfor se nærmere på den urnemodellen vi skal benytte.

La oss belyse urnemodellen ved følgende eksempler:

- tilfeldig utfylling av tipperekke med 12 kamper (med, ordnet)
- antall seksere etter 5 terningkast (med, uordnet)
- tilfeldig gjetting på de 3 første lagene i eliteserien i fotball (uten, ordnet)
- tilfeldig utfylling av Lotto-kupong (uten, uordnet)

Først betrakter vi tipperekka: Vi kan her tenke oss tre lapper som er merket henholdsvis  $H$  (hjemmeseier),  $U$  (uavgjort) og  $B$  (borteseier). Først trekker vi en lapp, og lar utfallet på lappen angi hva vi skal fylle ut på 1. kamp. Så legger vi lappen tilbake, og gjentar eksperimentet 12 ganger, slik at vi får fylt ut alle kampene. Rekkefølgen er her vesentlig (det er ikke nok å ha antall  $H$ ,  $U$  og  $B$  rett, vi må også ha dem riktig plassert), så vi får et ordnet utfall. Ved å fylle ut rekke etter rekke på denne måten, vil alle kombinasjoner være like sannsynlige, og vi kan bruke vår uniforme sannsynlighetsmodell.

Terningkast kan vi overføre til vår urnemodell ved at vi tenker oss 6 lapper i urna. På hver lapp står det et heltall mellom 1 og 6, og alle lappene er forskjellige. Å kaste en terning tilsvarer da å trekke en tilfeldig lapp. Etter at vi har lest hva som står på lappen, må vi legge denne tilbake før vi trekker på nytt for å simulere neste terningkast. Dette gjentas 5 ganger, og vi teller opp hvor mange ganger vi har trukket en sekser. Rekkefølgen er her uvesentlig, det er kun antall seksere som teller, uansett når de ble trukket. Dette er derfor et eksempel på et *ordnet* utvalg med tilbakelegging.

La oss så betrakte gjetting på de 3 første lagene i eliteserien i fotball. Si at vi har 14 lag totalt. Vi tenker oss da at vi lager 14 lapper med forskjellige navn, der det på hver lapp står navnet på et av lagene. Deretter trekker vi tilfeldig 3 av lappene, *uten* å legge noen av de trukne lappene tilbake igjen. Også her får vi en *ordnet* rekkefølge – vi antar at det ikke er uvesentlig å skille mellom hvem som får gull, sølv og bronse. Ved en slik trekning vil enhver kombinasjon bli like sannsynlig og vi kan igjen bruke den uniforme sannsynlighetsmodellen.

Til slutt ser vi på Lotto-kuponen. Vi tenker oss her 34 nummererte lapper med nummer fra 1 til 34, der alle lappene har forskjellige nummer. Deretter trekker vi 7 tilfeldige lapper, *uten* å legge noen av lappene tilbake etter at de er trukket. Her er rekkefølgen vi trekker tallene i uvesentlig, så vi får en *ordnet* rekkefølge. Enhver tallkombinasjon vil på denne måten bli like sannsynlig som en hvilken som helst annen tallkombinasjon, og en uniform sannsynlighetsmodell kan anvendes.

Vi gjennomgår nå hver av reglene i den rekkefølge de er introdusert.

### Potensregelen

Vi betrakter et sammensatt eksperiment som består av  $k$  «like» deler, der utfallsrommet for hver del av eksperimentet består av  $m$  enkeltutfall. Det totale utfallet av eksperimentet er da en samling av  $k$  enkeltutfall. At de ulike eksperimentdelene er «like», betyr at det er ett og samme utfallsrom for hvert deleksperiment. For et slikt eksperiment er det i alt

$$(2.11) \quad m^k = \text{totalt antall (ordnede) utfall}$$

der

$k$  = antall «like» deleksperimenter

$m$  = antall enkeltutfall i utfallsrommet til hvert deleksperiment

**Eks. 2.10** **Tippekupong.** En tippekupong består av 12 kamper, der hver kamp ender med hjemmeseier ( $H$ ), uavgjort ( $U$ ) eller borteseier ( $B$ ). Ett utfall av eksperimentet (dvs. utfylling av en rekke på tippekupongen) kan da f.eks. være  $\{HUHBHUUHUBBH\}$ , som betyr hjemmeseier i 1. kamp, uavgjort i 2. kamp, hjemmeseier i 3. kamp, osv.

#### Oppgave

Hva er sannsynligheten for 12 rette dersom du fyller ut en rekke tilfeldig?

#### Løsningsforslag

Her er  $k = 12$  og  $m = 3$  og vi får totalt  $m^k = 3^{12} = 531441$  rekker, som alle er like sannsynlige. Kun én rekke gir 12 rette, og sannsynligheten for 12 rette med en enkeltrekke (ett eksperiment) blir da  $P(12 \text{ rette}) = 1/(3^{12}) = 1.9 \cdot 10^{-6}$ . ☺

### Ordningsregelen (permutasjonsregelen)

Antall mulige måter vi kan ordne (stokke)  $r$  elementer av totalt  $n$  elementer på har betegnelsen  $P_r^n$ , som kan leses som «antall mulige ordninger (permutasjoner) av  $r$  av  $n$  elementer». Matematisk får vi:

$$(2.12) \quad P_r^n = n \cdot (n-1) \cdots (n-r+1)$$

**Eks. 2.11** **Resultatliste.** 15 syklister deltar i en motorsykkelkonkurranse. På hvor mange måter kan, teoretisk, øverste delen av resultatlista bestående av nr. 1, 2 og 3 se ut?

*Løsningsforslag*

Vi innser at  $n = 15$ ,  $r = 3$  og får  $P_3^{15} = 15 \cdot 14 \cdot 13 = 2730$  muligheter. Resonnementet kan gå som følger: 15 forskjellige personer kan vinne. For hver av disse 15 personene er det 14 igjen som kan få 2.-plassen. Dette gir  $15 \cdot 14 = 210$  kombinasjoner. For hver av disse 210 kombinasjonene er det 13 personer som kan komme på 3. plass. Dette gir totalt  $210 \cdot 13 = 2730$  muligheter. ☺

Et spesialtilfelle av ordningsregelen får vi når  $r = n$ . I dette tilfellet får vi

$$(2.13) \quad P_n^n = n \cdot (n-1) \cdot (n-n+1) = 1 \cdot 2 \cdots n = n!$$

**$n!$**  er en spesiell og alternativ skrivemåte for  $P_n^n$ , og kalles  **$n$  fakultet**.

**Eks. 2.12** **Kø.** 6 personer står etter hverandre i kø. I hvor mange forskjellige rekkefølger kan de 6 personene stå?

*Løsningsforslag*

Her er  $n = 6$  og vi får  $6! = 6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1 = 720$  kombinasjoner. Resonnementet kan gå som følger: Tenk deg at de 6 personene er nummerert fra 1 til 6. Person 1 kan stå på 6 forskjellige plasser i køen (først, nestførst, ..., bakerst). For en gitt plassering av person 1 har person 2 fem mulige plasseringer, hvilket gir 6·5 mulige kombinasjoner for person 1 og 2. For hver av disse 30 kombinasjonene, har person 3 fire mulige plasseringer, osv. ☺

### Kombinasjonsregelen

Antall mulige måter å kombinere  $r$  av  $n$  objekter, der rekkefølgen av de  $r$  objektene innbyrdes er uvesentlig, har betegnelsen  $\binom{n}{r}$  som leses: «antall kombinasjoner for  $r$  av  $n$  elementer». Matematisk får vi at

$$(2.14) \quad \binom{n}{r} = \frac{P_r^n}{r!} = \frac{n \cdot (n-1) \cdots (n-r+1)}{1 \cdot 2 \cdots r} = \frac{n!}{r! \cdot (n-r)!}$$

Binomialformelen til venstre ovenfor har blant annet følgende egenskaper:

$$(2.15) \quad \binom{n}{r} = \binom{n}{n-r}, \quad \binom{n}{n} = \binom{n}{0} = 1, \quad \binom{n}{1} = n$$

Kombinasjonsregelen kan begrunnes som følger:

Antall måter å ordne $r$ av $n$ objekter	er det samme som	antall måter å kombinere (velge) $r$ av $n$ objekter	multiplisert med	antall måter å ordne de $r$ valgte objektene innbyrdes
$P_r^n$	=	$\binom{n}{r}$	.	$r!$

#### Eks. 2.13 Bruk av kombinasjonsregelen

##### Oppgave

Beregn alle mulige a) *kombinasjoner* og b) *ordninger* av 3 forskjellige bokstaver valgt fra de 4 bokstavene  $A, B, C$  og  $D$ .

(NB! Husk at forskjellen på kombinasjoner og ordninger er at rekkefølgen er uvesentlig når det gjelder kombinasjoner):

##### Løsningsforslag

Kombinasjoner: Ordninger (permutasjoner):

- |               |                                |
|---------------|--------------------------------|
| $\{A, B, C\}$ | $ABC, ACB, BAC, BCA, CAB, CBA$ |
| $\{A, B, D\}$ | $ABD, ADB, DAB, DBA, BAD, BDA$ |
| $\{A, C, D\}$ | $ACD, ADC, CAD, CDA, DAC, DCA$ |
| $\{B, C, D\}$ | $BCD, BDC, CBD, CDB, DBC, DCB$ |

Som vi ser har vi 4 forskjellige kombinasjoner. For hver av disse kombinasjonene, som består av 3 elementer, er det  $3! = 1 \cdot 2 \cdot 3 = 6$  mulige måter å ordne (stokke) rekkefølgen av de 3 bokstavene på. Totalt har vi  $P^{4,3} = 4 \cdot 3 \cdot 2 = 24$  mulige måter vi kan ordne  $k = 3$  av totalt  $n = 4$  bokstaver på. Ved litt refleksjon forstår vi at antall kombinasjoner av 3 bokstaver fremkommer ved å ta totalt antall ordninger,  $P_r^n$  og dividere på antall måter bokstavene i hver kombinasjon kan ordnes på, dvs  $r!$ . Vi får da den matematiske betgelsen,  $\binom{n}{r}$ , for antall kombinasjoner av  $r$  av  $n$ . ☺

**Eks. 2.14** Komité. En rådgivende komité for en fengselsreform består av 15 medlemmer. 9 er for reformen, 4 er imot reformen og 2 mener det er hipp som happ om reformen blir innført eller ikke. En reporter ønsker å velge 3 tilfeldige personer fra komiteen for å meddele deres syn i et TV-program.

#### Oppgave

- Hva er sannsynligheten for at minst 2 av de utvalgte personene vil være for reformen?
- Hva er sannsynligheten for at de 2 første personene vil være for reformen, og den tredje vil være imot?

NB! Legg merke til at i oppgave a) er rekkefølgen uvesentlig, mens dette ikke er tilfelle i oppgave b).

#### Løsningsforslag

a) 3 av 15 personer kan bli valgt ut på ialt  $\binom{15}{3} = (15 \cdot 14 \cdot 13) / (1 \cdot 2 \cdot 3) = 455$  måter, som alle er like sannsynlige (tilfeldig utvalg). La  $A_2$  og  $A_3$  betegne hendelsene «eksakt 2 for» og «eksakt 3 for». Oppgave a) består i å finne  $P(A_2 \cup A_3)$ . Disse hendelsene er gjensidig ekskluderende, og vi har da at  $P(A_2 \cup A_3) = P(A_2) + P(A_3)$ .

For å beregne  $P(A_2)$ , må vi beregne antall måter vi kan trekke 3 av 15 personer, slik at 2 av dem blir trukket blant gruppa på 9 som er for reformen, og 1 blir trukket av gruppa på 6 som er enten imot eller likegyldige:

$$\binom{9}{2} \cdot \binom{6}{1} = \frac{9 \cdot 8}{1 \cdot 2} \cdot 6 = 216$$

Ved å anvende uniform sannsynlighetsmodell får vi da at hendelsen  $A_2$  består av  $m = 216$  like sannsynlige enkeltutfall, mens utfallsrommet består av  $\binom{15}{3} = 455$  like sannsynlige enkeltutfall:

$$P(A_2) = \frac{\binom{9}{2}\binom{6}{1}}{\binom{15}{3}} = \frac{216}{455}$$

Ved tilsvarende resonnement for  $A_3$  får vi:

$$P(A_3) = \frac{\binom{9}{3}\binom{6}{0}}{\binom{15}{3}} = \frac{84}{455}$$

og den søkte sannsynlighet blir:

$$P(\text{minst 2 av 3 for}) = P(A_2) + P(A_3) = 300/455 = \underline{60/91}$$

- b) Vi har totalt  $P_3^{15} = 15 \cdot 14 \cdot 13 = 2730$  ordnede utvalg for 3 personer fra en gruppe på 15, og alle de ordnede utvalg er like sannsynlige (tilfeldig «trekning»). Antallet ordnede utvalg der de 2 første personene er for og den tredje er mot, er

$$P_2^9 \cdot P_1^4 = (9 \cdot 8) \cdot (4) = 288$$

$$\Rightarrow P(\text{første 2 for, tredje mot}) = \underline{288/2730} \quad \circlearrowright$$

La oss oppsummere dette underkapitlet med å gjengi følgende hjelpetabell som viser sammenhengen mellom hvilken type trekning vi har (uten eller med tilbakelegging, uordnet eller ordnet rekkefølge), og hvilken telleregel som kommer til anvendelse:

		Rekkefølge:	
		uordnet	ordnet
Tilbake-legging:	uten med	Kombinasjonsregel -----	Ordningsregel Potensregel

Gå nøye gjennom tabellen selv, og kontrollér at rubrikken (med,uordnet) ikke dekkes av noen av de tellereglene vi har behandlet. Vi kommer senere tilbake til dette tilfellet i forbindelse med binomisk fordeling.

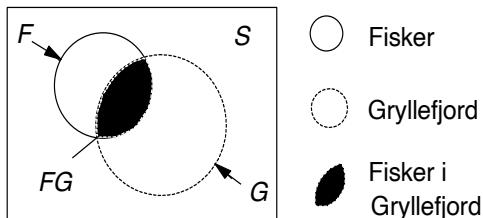
## 2.6 Betinget sannsynlighet

Sannsynligheten for en hendelse,  $A$ , må ofte modifiseres dersom en får oppgitt informasjon(er) om en annen hendelse,  $B$ . Eks: Betrakt et tilfeldig telefonnummer i Nord-Norge-katalogen(e). La  $F$  og  $G$  betegne hendelsene:

$F$ : Abonnenten er fisker

$G$ : Abonnenten bor i Gryllefjord

Utfallsrom: N-Norge-abonnenter



Situasjonen er illustrert i Venn-diagrammet til venstre. *Utfallsrommet*  $S$  består av alle telefonnumrene i N-Norge, og utgjør hele boksen. Den *heltrukne* sirkelen inneholder alle numrene der abonnenten er fisker, og den *stippled* sirkelen inneholder alle

numrene der abonnenten bor i Gryllefjord. Det *skraverte* området der de to sirklene overlapper hverandre, inneholder alle numrene der abonnenten *både* er fisker *og* bor i Gryllefjord (dvs. snitthendelsen  $FG$ ). Området som ligger utenfor begge sirklene inneholder alle numrene der abonnenten *verken* er fisker *eller* bor i Gryllefjord (dvs. hendelsen  $(F \cup G)^c$ )

Tenk deg nå at alle Nord-Norge-numrene er lagt på data, og et regnemaskinprogram sørger for at du «trekker» et nummer tilfeldig. Sannsynligheten  $P(G)$  for at nummeret skal tilhøre en abonnent i Gryllefjord vil da være antall abonnenter i Gryllefjord delt på antall abonnenter i Nord-Norge totalt.

Sett nå at programmet gir deg mulighet til først å bestemme retningsnummer (sted), og deretter å trekke tilfeldig blant alle numre med dette retningsnummeret. Si f.eks. at du ønsker å trekke blant Gryllefjord-numrene. Sannsynligheten for å trekke et nummer der abonnenten er fisker er da større enn om du trakk blant alle Nord-Norge-numrene (prosentvis flere fiskere i Gryllefjord enn i Nord-Norge totalt). Med henvisning til Venn-diagrammet ovenfor, blir sannsynligheten i dette tilfellet lik antall abonnenter i Gryllefjord som er fiskere delt på totalt antall abonnenter i Gryllefjord. Vi snakker nå om en *betinget* sannsynlighet for en hendelse  $F$  (fisker) fordi vi har *gitt* at en tilleggshendelse,  $G$  (Gryllefjord-abonnent), har inntruffet. Vi benytter i dette tilfellet en loddrett strek for å skille hendelsen vi ønsker å finne sannsynligheten for, fra de tilleggsbetingelser som er oppgitt:

$$P(F|G) = P(\text{«abonnenten er fisker gitt at abonnenten bor i Gryllefjord»})$$

NB! *Tilleggsbetingelsen*(e) er alltid oppgitt til *høyre* for den lodrette streken, mens den hendelse vi skal finne sannsynligheten til er oppgitt til *venstre* for streken.

Med henvisning til Venn-diagrammet, skjønner vi intuitivt at  $P(F|G)$  = antall abonnenter i Gryllefjord som er fiskere delt på antall abonnenter i Gryllefjord totalt. Dette er ofte en praktisk og mulig måte å bestemme betinget sannsynlighet på. I mange tilfeller er det imidlertid nødvendig, og/eller langt mer lettvint, å bruke mer indirekte metoder, som vi skal se.

Generelt gjelder følgende:

### Betinget sannsynlighet

Sannsynligheten for en hendelse  $A$  under betingelsen (tilleggsopplysningen) at en hendelse  $B$  har inntruffet, har følgende betegnelse:

$$(2.16) \quad P(A | B) \text{ som leses: «Sannsynligheten for } A \text{ gitt } B\text{»}.$$

Legg merke til at betingelsen står på *høyre* side av den lodrette streken, mens den hendelsen vi skal finne sannsynligheten til, står på *venstre* side. Følgende viktige regler gjelder for betinget sannsynlighet:

$$(2.17) \quad P(AB) = P(A) \cdot P(B | A)$$

$$(2.18) \quad P(A | B) = \frac{P(B | A) \cdot P(A)}{P(B)} \quad (\text{Bayes' formel})$$

La oss nå belyse formlene ovenfor ved å gå tilbake til telefonnummer-eksemplet ( $A = G$ ,  $B = F$ ). Ved hjelp av lign.(2.17) får vi:

$$P(F|G) = \frac{P(FG)}{P(G)}$$

som omsatt til ord kan formuleres som følger:

$P$  ( tilfeldig N-Norge-abonnent i Gryllefjord er fisker ) =  
 $P$  ( tilfeldig N-Norge-abonnent er både fisker og bor i Gryllefjord ) delt på  
 $P$  ( tilfeldig N-Norge-abonnent bor i Gryllefjord ).

Sett nå at vi var interessert i å finne  $P(G | F)$ , mens vi kjente  $P(G)$ ,  $P(F)$  og  $P(F | G)$ . Vi kunne da brukt Bayes' formel (se lign.(2.18)):

$$P(G|F) = \frac{P(F|G) \cdot P(G)}{P(F)}$$

som omsatt til ord kan formuleres som følger:

$P(\text{tilfeldig N-Norge-abonnent som er fisker bor i Gryllefjord}) =$   
 $P(\text{tilfeldig N-Norge-abonnent i Gryllefjord er fisker})$  ganget med  
 $P(\text{tilfeldig N-Norge-abonnent bor i Gryllefjord})$  delt på  
 $P(\text{tilfeldig N-Norge-abonnent er fisker})$

**Eks. 2.15** **Avissalg.** En kiosk fører statistikk over hvor mange kunder som kjøper Dagbladet ( $D$ ), hvor mange som kjøper Aftenposten ( $A$ ), og hvor mange som kjøper begge aviser. Resultat:

40 % kjøper Dagbladet, og av disse kjøper 20 % Aftenposten. 30 % kjøper Aftenposten.

### Oppgave

Hva er sannsynligheten for at en tilfeldig kunde

- a) Kjøper både Dagbladet og Aftenposten ?
- b) Kjøper verken Dagbladet eller Aftenposten ?
- c) Kjøper Dagbladet dersom vedkommende kjøper Aftenposten ?

### Løsningsforslag

Vi bruker forkortelsene  $D$  og  $A$  for henholdsvis hendelsene «tilfeldig kunde kjøper Dagbladet» og «tilfeldig kunde kjøper Aftenposten».

a)  $P(\text{«både Dagbladet og Aftenposten}}\text{»})$

$$= P(DA) = P(D) \cdot P(A | D) = 0.4 \cdot 0.2 = \underline{0.08}$$

b)  $P(\text{«Verken Dagbladet eller Aftenposten}}\text{»}) = P((D \cup A)^C) = 1 - P(D \cup A)$

$$P(D \cup A) = P(D) + P(A) - P(DA) = 0.04 + 0.3 - 0.08 = 0,62$$

$$\Rightarrow 1 - P(D \cup A) = \underline{0.38}$$

c) Ved hjelp av Bayes' formel finner vi tilslutt:

$$P(D \cap A) = P(A | D) \cdot P(D) / P(A) = 0.2 \cdot 0.4 / 0.3 = \underline{0.267} \quad \odot$$

Tilbake til det generelle: Lign.(2.17) kan generaliseres til flere enn 2 hendelser. For 3 hendelser  $A$ ,  $B$  og  $C$  får vi:

$$(2.19) \quad P(ABC) = P(A) \cdot P(B | A) \cdot P(C | BA)$$

Som huskeregel for høyre side av lign.(2.19) kan vi notere følgende: Vi begynner fra venstre i snitthendelsen  $ABC$ , dvs. vi begynner med  $P(A)$ . Så går vi til neste hendelse, nemlig  $B$ , og betinger med alt til venstre for  $B$ , dvs.  $A$ , og vi får  $P(B | A)$ . Slik fortsetter vi, og får  $C$  betinget med alt til venstre for  $C$ , nemlig  $AB$ . Den siste faktoren blir da  $P(C | AB)$ .

Merk at vi godt kunne snudd rekkefølgen i lign. (2.19), da faktorenes orden i snitthendelsen  $ABC$  er likegyldig. Eks:

$$(2.20) \quad P(CBA) = P(C) \cdot P(B | C) \cdot P(A | CB)$$

I praksis vil en ofte bruke lign. (2.19) med  $A$ ,  $B$  og  $C$  definert i kronologisk rekkefølge, dvs.  $A$  er den hendelsen som opptrer først, så kommer  $B$  og til slutt  $C$ .

**Eks. 2.16** Egg. En gårdbruker har en eske som inneholder 30 egg, derav 5 med blodflekker. Han sjekker 3 egg ved å trekke dem tilfeldig en etter en fra eska.

### Oppgave

Hva er sannsynligheten for at de første to eggene vil inneholde blodflekker og det tredje vil være uten blodflekker?

### Løsningsforslag

La  $R_1R_2H_3$  betegne snitthendelsen å trekke blodflekkete egg i de 2 første trekningene og klart egg i 3. trekning ( $R$  for rød,  $H$  for hvit). Vi bruker lign.(2.19) med  $A = R_1$ ,  $B = R_2$  og  $C = H_3$ :

$$P(R_1R_2H_3) = P(R_1) \cdot P(R_2 | R_1) \cdot P(H_3 | R_1R_2)$$

Fordi trekningene er tilfeldige, har vi at  $P(R_1) = \binom{5}{1}/\binom{30}{1} = 5/30$ . Vi ser så på  $P(R_2 | R_1)$ . Her er betingelsen  $R_1$  gitt, dvs. vi skal finne sannsynligheten for å trekke et blodflekket egg i 2. trekning gitt at det er trukket et blodflekket egg i 1. trekning. Da er 29 egg igjen, og 4 av disse er med blodflekker:  $P(R_2 | R_1) = 4/29$ . Tilsvarende resonnement gir at  $P(H_3 | R_1R_2) = 25/28$ , og vi får:

$$P(R_1R_2H_3) = (5/30) \cdot (4/29) \cdot (25/28) = \underline{\underline{25/1218}}$$

Her kunne vi også løst problemet ved å bruke reglene for ordnet utvalg uten tilbakelegging: Det er totalt  $P_3^{30}$  like sannsynlige måter å ordne  $r = 3$  av  $n = 30$  elementer uten tilbakelegging. Antallet ordnede utvalg der de 2 første eggene er flekkete og det tredje er uflekket er  $P_2^5 \cdot P_1^{25}$ . Vi får derfor:

$$P(R_1 R_2 H_3) = \left( P_2^5 \right) \cdot \left( P_1^{25} \right) / \left( P_3^{30} \right) = (5 \cdot 4) \cdot (25) / (30 \cdot 29 \cdot 28) = \underline{25 / 1218} \quad \text{😊}$$

## 2.7 Uavhengige hendelser

Begrepet *uavhengige* hendelser er et svært sentralt begrep nøyde koplet til begrepet betinget sannsynlighet:

### Uavhengige hendelser

To hendelser  $A$  og  $B$  er uavhengige hvis og bare hvis

$$(2.21) \quad P(A | B) = P(A)$$

Dersom  $P(B) \neq 0$ , er betingelsen ovenfor ekvivalent med at

$$(2.22) \quad P(AB) = P(A) \cdot P(B)$$

NB! To hendelser,  $A$  og  $B$ , som er **gjensidig ekskluderende** ( $AB = \emptyset$ ), er *aldri* uavhengige, og omvendt.

**Eks. 2.17** **To kast med en rettferdig mynt.**  $S = \{ KK, KM, MK, MM \}$ , der  $K$  = kron,  $M$  = mynt og alle enkeltutfallene er like sannsynlige. Vi definerer følgende 3 hendelser:

$A$ : Kron ( $K$ ) i første kast

$B$ : Kron i andre kast

$C$ : Kron i begge kast eller mynt i begge kast

### Oppgave

Finn ut hvilke par av hendelsene ovenfor som er uavhengige.

*Løsningsforslag*

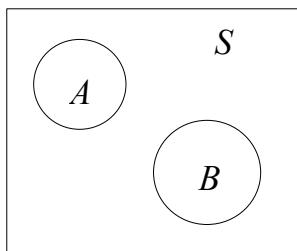
Intuitivt burde  $A$  og  $B$  være uavhengige hendelser. Det er imidlertid ikke like oppagt om  $A$  og  $C$  er uavhengige hendelser, eller om  $B$  og  $C$  er uavhengige. Ved regning finner vi (siden hvert enkeltutfall har sannsynlighet 1/4):

$$P(A) = 1/2, \quad P(AB) = 1/4 = P(A) \cdot P(B)$$

$$P(B) = 1/2, \quad P(AC) = 1/4 = P(A) \cdot P(C)$$

$$P(C) = 1/2, \quad P(BC) = 1/4 = P(B) \cdot P(C)$$

Følgelig, alle par av hendelser er uavhengige. Legg også merke til at ingen av hendelsene er gjensidig ekskluderende, både  $A$ ,  $B$  og  $C$  inneholder jo enkeltutfallet  $KK$ . Vi vil forøvrig alltid ha at gjensidig ekskluderende hendelser er avhengige. Dette er intuitivt svært logisk: Dersom en av to gjensidig ekskluderende hendelser har inntruffet, vet vi jo med 100% sikkerhet at den andre ikke kan ha inntruffet. ☺



At to hendelser er gjensidig ekskluderende fremkommer i et Venn-diagram ved at de ikke har noe overlapp. I figuren til venstre er hendelsene  $A$  og  $B$  gjensidig ekskluderende. I dette tilfellet skjønner vi at  $P(AB) = 0$ .

På tampen minner vi om 2 viktige formler:

$$1) \quad P(A) = 1 - P(A^c)$$

$$2) \quad P(A \cup B) = P(A) + P(B) - P(AB)$$

Dersom  $A$  og  $B$  er gjensidig ekskluderende, forenkles 2) til  $P(A \cup B) = P(A) + P(B)$ .

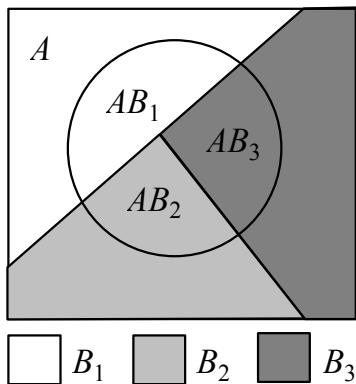
**Gjensidig ekskluderende hendelser og betinget sannsynlighet**

Vi skal ta utgangspunkt i at en hendelse  $B$  utgjør hele utfallsrommet ( $P(B) = 1$ ), og at  $B$  er inndelt i  $n$  gjensidig ekskluderende delhendelser:  $B_1, \dots, B_n$ . Dette kalles også en **partisjon** av  $S$ , og medfører at  $P(B_1 \cup B_2 \cup \dots \cup B_n) = P(B_1) + \dots + P(B_n) = 1$ . La videre  $A$  være en hendelse vi ønsker å bestemme sannsynligheten til, der de betingete sannsynlighetene  $P(A | B_1), \dots, P(A | B_n)$  er kjente. Anta også at sannsynlighetene  $P(B_1), \dots, P(B_n)$  er kjente størrelser. Da får vi følgende (husk at  $P(B) = 1$ ):

$$(2.21) \quad P(A) = P(AB) = P(A(B_1 \cup \dots \cup B_n)) = P(AB_1) + \dots + P(AB_n)$$

Å innse at lign.(2.21) er riktig, er trolig enklest ved hjelp av Venn-diagram, der et eksempel med  $n = 3$  er vist i neste figur.  $A$  er hendelsen inneholdt i sirkelen, mens  $B$  utgjør hele utfallsrommet (boksen).

Bruker vi nå lign.(2.17) på hvert av deluttrykkene til høyre i lign.(2.21) får vi:



$$\begin{aligned}
 (2.22) \quad P(A) &= P(AB) = P(AB_1) + P(AB_2) + \dots + P(AB_n) \\
 &= P(A | B_1) \cdot P(B_1) + \dots + P(A | B_n) \cdot P(B_n) \\
 &= \sum_{i=1}^n P(A | B_i) \cdot P(B_i)
 \end{aligned}$$

## 2.8 Oppgaver

**2.1** Du trekker ett lodd fra en urne med 100 lodd. 3 lodd gir kakegevinst ( $K$ ), 1 lodd gir bokgevinst ( $B$ ) og 96 lodd gir ingen gevinst ( $I$ ).

Skriv opp et passende utfallsrom,  $S$ , for det trukne loddet.

**2.2** Som i 2.1, men du trekker nå to lodd.

**2.3** En urne inneholder en rød ( $R$ ), en blå ( $B$ ) og en gul ( $G$ ) kule. Hvor mange enkeltutfall består utfallsrommet  $S$  av i følgende tilfeller:

- a) 2 trekninger uten tilbakelegging, rekkefølgen vesentlig.
- b) 2 trekninger med tilbakelegging, rekkefølgen uvesentlig.
- c) 3 trekninger uten tilbakelegging, rekkefølgen vesentlig.

**2.4** Et ruteark med  $3 \cdot 3 = 9$  ruter skal fylles ut med 5 like kryss og 4 like ringer. Hvor mange kombinasjoner får vi?

**2.5** Et brett med  $3 \cdot 3$  ruter skal dekkes med 9 forskjellige brikker. Hvor mange kombinasjoner får vi?

**2.6** En terning kastes en gang. Hendelse  $A$  betyr at terningen viser en sekser, mens hendelse  $B$  betyr at terningen viser høyst 4 øyne. Forklar hva følgende hendelser betyr, og beregn sannsynligheten for hver av dem:

- a)  $(A \cup B)^C$
- b)  $A^C \cap B^C$
- c)  $(A \cap B)^C$

d)  $A^C \cup B^C$

**2.7** Vis matematiskt at

$$\begin{aligned} P(A \cup B \cup C) &= P(A) + P(B) + P(C) \\ &\quad - P(AB) - P(AC) - P(BC) + P(ABC) \end{aligned}$$

**2.8** Gitt utfallsrommet  $S = \{e_1, e_2, e_3, e_4, e_5\}$  der alle enkeltutfallene er like sannsynlige. La videre hendelsene  $A$ ,  $B$  og  $C$  være definert som følger:  $A = \{e_1, e_2, e_4\}$ ,  $B = \{e_2, e_3\}$  og  $C = \{e_2, e_4, e_5\}$ .

- a) Tegn Venn-diagram og tegn inn hendelsene  $A$ ,  $B$  og  $C$ .
- b) Beregn  $P(A)$ ,  $P(A \cup B)$ ,  $P(AC)$ ,  $P(B^C)$ ,  $P[(A \cup B)^C]$  og  $P(A^C \cup B^C)$

**2.9** Angi telleregel og finn hvor mange enkeltutfall følgende eksperiment består av:

- a) Ett myntkast med 5 pengestykker
- b) Ett terningkast med 3 terninger
- c) Trekning av 1., 2. og 3. premie blant 7 personer med like store vinner-sjanser (trekning uten tilbakelegging).
- d) Trekning uten tilbakelegging av 3 kuler fra urne med 2 gule og 4 røde kuler.
- e) Antall kombinasjoner av 13 kort fra kortstokk med 52 forskjellige kort.

**2.10** En oversikt over hvilken utdanning ansatte ved en høyteknologisk bedrift har, er satt opp i følgende tabell:

	ingeniør-utdannet	annen utdanning
kvinner	10	30
menn	20	40

- a) Hva er sannsynligheten for at en tilfeldig ansatt er kvinne?
- b) Hva er sannsynligheten for at vedkommende er ingeniørutdannet og kvinne?
- c) Hva er sannsynligheten for at vedkommende er ingeniørutdannet når vi vet at vedkommende er kvinne?
- d) To tilfeldige ansatte snakker sammen. Hva er sannsynligheten for at ingen av dem er ingeniørutdannet?

**2.11** I en klasse på 30 elever kunne 12 danse sving ( $S$ ), 4 kunne spille gitar ( $G$ ) og det var 17 jenter ( $J$ ) i klassen. 2 av jentene spilte gitar. En av guttene og en av jentene i klassen kunne både danse sving og spille gitar. 7 av guttene kunne verken spille gitar eller danse sving.

- a) Hvor mange jenter kunne verken danse sving eller spille gitar?
- b) Hvor mange gutter kunne danse sving?

(hint: Tegn Venn-diagram over alle elevene i klassen).

**2.12** Anta at du har 20 ingredienser tilgjengelig for å lage pizzafyll.

- a) Hvor mange forskjellige pizzaer kan bli laget av
- eksakt en ingrediens?
  - eksakt to ingredienser?
  - eksakt tre ingredienser?
- b) Svar på ii) dersom to halvparter av en pizza kan være forskjellige.

**2.13** Fra en kortstokk på 52 kort gjør vi to trekking etter hverandre uten tilbakelegging. Hva er sannsynligheten for å trekke

- a) et svart og et rødt kort?
- b) ikke mer enn én honnør (ess, konge, dame, knekt)?

**2.14** På en skole med 120 elever er det 70 jenter, 20 gutter og 10 jenter synes matematikk er gøy.

- a) Hva er sannsynligheten for at en tilfeldig elev ikke syntes matematikk er gøy, når det viste seg at eleven var en gutt?
- b) Hva er sannsynligheten for at to tilfeldige elever har forskjellig kjønn og synes matematikk er gøy?

**2.15** Du trekker 13 kort fra en kortstokk på 52 kort der 16 er honnørkort. Beregn sannsynligheten for at trekningen gir:

- a) 13 kort med samme farge (hjerter, ruter, kløver eller spar).
- b) Ingen honnørkort.
- c) Bare honnørkort.

**2.16** En 6-kantet kubisk terning har to røde sider ( $R$ ) og 4 blå sider ( $B$ ). Terningen kastes 3 ganger.

- a) Bruk den generelle addisjonssetningen til å beregne sannsynligheten for minst ett rødt kast.
- b) Beregn også sannsynligheten på enklast mulig måte.

\***2.17** En tipperekke inneholder 4 kamper, hver med 3 utfall (hjemmeseier ( $H$ ), uavgjort ( $U$ ) eller borteseier ( $B$ )). Hvor mange rekker må fylles ut for å være sikret minst 3 av 4 rette?

**2.18** Sannsynligheten for at en mann er fargeblind er 0.08. Sannsynligheten for at en kvinne er fargeblind er 0.004.

- a) Hva er sannsynligheten for at en tilfeldig utvalgt person er fargeblind?

Forutsett at halvparten av befolkningen er kvinner.

- b) Det er begått en forbrytelse. Det er åpenbart at forbryteren er fargeblind. Hva er sannsynligheten for at forbryteren er en mann?

**2.19** Bladet Motor slår fast i en omfattende artikkel 2/85 at bilparken i Norge er befeftet med betydelige feil. Anta at du besøker en bruktbilforretning som har 25 biler å tilby av en bestemt årsmodeell. Av disse har to biler ingen feil, 5 har en feil, 10 har to feil, syv har tre feil og en har flere enn tre feil.

Du kjøper to biler. Hva er sannsynligheten for at du får:

- a) begge biler uten feil,
- b) én og bare én med høyst to feil,
- c) to med minst tre feil hver,
- d) to med minst tre feil tilsammen?

**2.20** Riktig eller galt?

- a) Mulige utfall når vi kaster en rettferdig mynt to ganger er: ingen «kron», en «kron» og to «kron».
- b) Hvis en rettferdig mynt har vært kastet 5 ganger med «kron» som utfall i alle 5 kast, er sannsynligheten for å få «mynt» i det 6. kastet større enn 0.5.
- c) Det er mulig at vi for to hendelser, der  $P(A) = 0.5$  og  $P(B) = 0.7$ , kan ha  $P(A \cup B) = 1.2$ .
- d)  $P(A|B)$  er alltid mindre enn  $P(B)$

e) To ikke-tomme og gjensidig ekskluderende (disjunkte) hendelser er aldri uavhengige.

f) Hvis  $P(A) = P(A|B)$ , så er  $A$  og  $B$  gjensidig ekskluderende hendelser.

g) Hvis  $P(A) = 0.4$  og  $P(B) = 0.3$  og  $P(A|B) = 0.8$ , så er  $P(AB) = 0.2$ .

**2.21** En student skal opp til eksamen i 4 fag ( $A$ ,  $B$ ,  $C$  og  $D$ ), og anslår de respektive sannsynligheter for å bestå eksamen til 0.8, 0.9, 0.7 og 0.5. Hvis vi forutsetter uavhengighet, hva er sannsynligheten for at studenten:

- a) består alle 4 eksamener,
- b) stryker i minst ett fag,
- c) hvis studenten stryker i ett (og bare ett) fag, at han/hun stryker i fag  $D$ ?

**2.22** I en klubb med 10 medlemmer skal det utpekes et styre som består av i alt 3 medlemmer.

- a) Hvor mange styrer har en å velge mellom?

Anta at et styre skal bestå av en formann, en nestformann og en sekretær.

- b) Hvor mange styrer har en nå å velge mellom?

**2.23 (E)** En medisin har 2 forskjellige «bivirkninger»: Kvalme ( $K$ ) og hodepine ( $H$ ). 20 % av brukerne blir kvalme ( $P(K) = 0.2$ ), og blant de som blir kvalme er det 30 % som også får hodepine ( $P(H|K) = 0.3$ ). 10 % av brukerne får hodepine. De som verken blir kvalme eller får hodepine får ingen bivirkninger ( $I$ ).

- a) Hva er sannsynligheten for at en tilfeldig bruker både blir kvalm og får hodepine?

- b) Er hendelsen bivirkning ( $B$ ) det samme som unionen ( $K \cup H$ ) eller snittet ( $KH$ ) av  $K$  og  $H$ ? Begrunn svaret og finn deretter  $P(B)$  for en tilfeldig bruker.

Det viser seg at røykere ( $R$ ) er betydelig mer utsatt for bivirkninger enn ikke-røykere ( $R^C$ ). Av brukerne er det 75 % som ikke røyker, og blant disse er det 4 % som får bivirkninger.

- c) Vis at sannsynligheten for å få bivirkning for en røyker er 84%.  
d) Finn sannsynligheten for at en tilfeldig bruker uten bivirkninger er ikke-røyker.

**2.24 (E)** Sannsynligheten for at et tilfeldig besøk hos tannlegen vil resultere i tann-uttrekning er 0.06, sannsynligheten for at det vil resultere i plombering er 0.23 og sannsynligheten for at det vil resultere i både tannuttrekning og plombering er 0.02.

- a) Hva er sannsynligheten for at et tilfeldig besøk hos tannlegen vil resultere i tannuttrekning, men ikke plombering?  
b) Hva er sannsynligheten for at besøket vil resultere i tannuttrekning, plombering eller begge deler?  
c) Hva er sannsynligheten for at besøket vil resultere i tannuttrekning eller plombering, men ikke begge deler?  
d) Hva er sannsynligheten for at besøket verken vil resultere i tannuttrekning eller plombering?

**2.25 (E)**  $A$  og  $B$  er to kjennetegn slik at  $P(A) = 0.25$ ,  $P(B) = p$  og  $P(A \cup B) = 0.5$ .

- a) Finn sannsynligheten  $P(A \cap B)$  uttrykt ved  $p$ .  
b) Hvilken verdi har  $p$  når  $A$  og  $B$  er uavhengige kjennetegn?

**2.26 (E)** En ny medisin mot en bestemt sykdom blir innført. Forsøk har vist at 70 % av pasientene som får medisinen blir helbredet. Dessverre kan medisinen ha bivirkninger. For en pasient som blir helbredet er sannsynligheten for bivirkninger 0.25. For en pasient som ikke blir helbredet er sannsynligheten for bivirkninger 0.1.

- a) Hva er sannsynligheten for at en tilfeldig pasient skal bli helbredet og dessuten ikke skal få bivirkninger?  
b) Hva er sannsynligheten for bivirkninger hos en tilfeldig pasient?  
c) Hva er sannsynligheten for at en pasient blir helbredet når vi vet at han ikke får bivirkninger av medisinen?

**2.27 (E)** I en klubb er 60 % av medlemmene kvinner. 50 % av kvinnene røyker og 30 % av mennene røyker. Et medlem trekkes ut på måfå.

- a) Hva er sannsynligheten for at medlemmet røyker?  
b) Hva er sannsynligheten for at medlemmet er en kvinne, forutsatt at vedkommende røyker?  
c) Hvor mange prosent av ikke-røykerne er menn?

**2.28 (E)** FBI bruker en løgndetektor som viser «skyldig» med sannsynlighet 0,92 for personer som virkelig er skyldige i en bestemt forbrytelse. Dersom løgndetektoren brukes på en

uskyldig person vil den vise «uskyldig» med sannsynlighet 0.98.

Vi definerer følgende hendelser:

- A: Personen er «skyldig».
- B: Løgndetektoren viser «skyldig».

Det velges en tilfeldig person fra en (mistenkelig) gruppe der 10 % er «skyldige».

- a) Skriv opp sannsynlighetene  $P(A)$ ,  $P(B | A)$ ,  $P(B^C | A^C)$  og beregn  $P(A^C)$ ,  $P(B^C | A)$ ,  $P(B | A^C)$  og  $P(AB)$ .

Forklar hva den siste sannsynligheten står for.

- b) Finn sannsynligheten for at løgndetektoren skal vise skyldig.
- c) Anta at detektoren viser «uskyldig» for personen. Hva er sannsynligheten for at personen i virkeligheten er «uskyldig»?

**2.29 (E)** I en urne er det 4 røde og 2 grønne kuler. Bortsett fra farge er de like. I denne oppgaven skal vi finne sannsynligheten for å trekke 2 røde og 1 grønn kule ved forskjellige trekningsprosedyrer.

- a) Bestem sannsynligheten for å trekke 2 røde og 1 grønn kule dersom trekningen er *uten* tilbakelegging.
- b) Hva blir sannsynligheten for å trekke 2 røde og 1 grønn kule dersom trekningen er *med* tilbakelegging?
- c) Trekk 3 kuler på følgende måte: Trekk først 2 *uten* tilbakelegging, registrer farge og legg dem så oppi urna igjen før 3. trekk gjøres. Hva

blir nå sannsynligheten for å få 2 røde og 1 grønn kule?

**2.30 (E)** I et bygg på 6 etasjer kommer det 5 personer inn i heisen i 1. etasje. Personene har ingenting med hverandre å gjøre, og det er like stor sannsynlighet for at hver enkelt skal gå av i etasje 2-6 (5 stopp).

- a) Hvor stor sannsynlighet er det for at alle går av i 2. etasje?
- b) Hvor stor sannsynlighet er det for at alle går av i samme etasje?
- c) Hvor stor sannsynlighet er det for at ingen skal til 6. etasje?
- d) Hva er sannsynligheten for at det går av én i hver etasje?
- e) Anta at det i starten bare var 4 personer i heisen. Hva er da sannsynligheten for at det går av høyest én i hver etasje?

## 2.9 Formelsamling

### Relativ frekvens, $r_N(A)$

$$r_N(A) = \frac{\text{Antall forsøk med } A\text{-utfall}}{\text{Totalt antall forsøk } N}$$

### Sannsynlighet, $P(A)$

$$P(A) = \lim_{N \rightarrow \infty} r_N(A)$$

### Sannsynlighetslover

$$P(A^C) = 1 - P(A)$$

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

### Potensregel

Gitt et forsøk som består av  $k$  like deleksperiment, der alle deleksperiment har samme utfallsrom bestående av  $m$  like sannsynlige enkeltutfall, og der utfallet av forsøket er en **ordnet** rekkefølge av tilfeldige utfall av de  $k$  deleksperimentene. Vi har da totalt:

$m^k$  like sannsynlige utfall.

### Ordningsregel

$r$  «nummererte» av  $n$  «nummererte» elementer kan ordnes på ialt

$$P_r^n = n \cdot (n-1) \cdots (n-r+1)$$

forskjellige og like sannsynlige måter (**ordnet** rekkefølge).

$$P_n^n = 1 \cdot 2 \cdots n = n!$$

som kalles «**n fakultet**».  $0! = 1$  pr. definisjon

### Kombinasjonsregel

$r$  «nummererte» av  $n$  «nummererte» elementer, der rekkefølgen er uvesentlig (uordnet rekkefølge), kan kombineres på

$$\binom{n}{r} = \frac{P_r^n}{n!} = \frac{n!}{r! \cdot (n-r)!}$$

forskjellige og like sannsynlige måter.

$$\binom{n}{k} = \binom{n}{n-k}, \quad \binom{n}{0} = \binom{n}{n} = 1$$

### Betinget sannsynlighet

$$P(A | B) = \frac{P(AB)}{P(B)} = \frac{P(B | A) \cdot P(A)}{P(B)}$$

der  $P(A|B)$  leses «sannsynligheten for  $A$  gitt  $B$ », og likheten mellom første og siste uttrykk ovenfor er **Bayes'** formel.

La  $B$  utgjøre hele utfallsrommet ( $P(B) = 1$ ) bestående av  $n$  gjensidig ekskluderende delhendelser (partisjon)  $B_1, \dots, B_n$ . Da gjelder:

$$P(A) = \sum_{i=1}^n P(A | B_i) \cdot P(B_i)$$

For en sammensatt snitthendelse  $ABC$  gjelder:

$$P(ABC) = P(A) \cdot P(B|A) \cdot P(C|AB)$$

### Uavhengige hendelser

To hendelser,  $A$  og  $B$ , er uavhengige hvis og bare hvis

$$P(A|B) = P(A)$$

Dersom  $P(B) \neq 0$  er ligninga over ekvivalent med at  $P(AB) = P(A) \cdot P(B)$ .

### Union

$A \cup B$  = mengden som består av alle elementer som er med i  $A$  eller i  $B$  eller i både  $A$  og  $B$ .

### Snitt

$A \cap B = AB$  = Mengden som består av alle elementer som er med både i  $A$  og i  $B$ .

### Komplement

$A^C$  = Mengden av alle elementer som ikke er med i  $A$ .

## *Kapittel 3*

# **Stokastisk variabel og sannsynlighetsfordeling**

### **3.1 Innledning**

Vi skal i dette kapitlet ta for oss de viktige begrepene **stokastisk variabel** og **sannsynlighetsfordeling** til slike variabler. Vi skal holde oss til det diskrete tilfellet. Kontinuerlige stokastiske variabler vil vi komme tilbake til senere. Svært forenklet sagt handler stokastiske variabler og sannsynlighetsfordelinger om å sette tall i stedet for bokstaver på det vi i forrige kapittel definerte som hendelser, og å se på hvilken sannsynlighet de ulike tallene er forbundet med.

En rekke *teoretiske* begreper og teknikker i dette kapitlet er analoge til de tilsvarende *empiriske* (empirisk: basert på konkrete data) begreper og teknikker fra kap.1 om deskriptiv statistikk. Én-dimensjonale sannsynlighetsfordelinger (kap. 3.3) minner svært om relativ frekvens for grupperte data. Grafisk fremstilling av én-dimensjonale sannsynlighetsfordelinger (kap. 3.4) minner svært om grafisk fremstilling av grupperte data. De teoretiske begrepene **forventning** (kap. 3.5), **varians** og **standardavvik** (kap. 3.6) er analoge til begrepene empirisk middelverdi, empirisk varians og empirisk standardavvik. Den teoretiske **korrelasjonskoeffisienten** (kap. 3.8) er analog til den empiriske korrelasjonskoeffisienten.

Litt forenklet kan vi si at teoretiske størrelser, som f.eks. forventning og varians, gir samme resultat som de tilsvarende empiriske størrelser når vi har «uendelig» mange observasjoner. I praksis har vi imidlertid et begrenset tallmateriale. Å forstå forskjellen på de empiriske og teoretiske begreper i statistikk er *svært viktig*. Dette vil forhåpentligvis gå som en rød tråd gjennom boka.

Fordelingsbegrepet er ikke entydig i den statistiske litteratur. Noen steder er begrepet sannsynlighetsfordeling knyttet til den **kumulative** fordelingsfunksjonen,  $F(x)$  (definert senere). Andre steder er begrepet knyttet til **sannsynlighets-tetthetsfunksjonen**,  $f(x)$  (også definert senere). Det er den siste varianten som vil bli benyttet her. Vi skal se både på én-dimensjonale fordelinger  $f(x)$ , og todimensjonale **simultane** fordelinger  $f(x,y)$ .

I kap. 1 så vi at den empiriske korrelasjonskoeffisienten  $r$  var i nærheten av  $\pm 1$  dersom to variabler  $x$  og  $y$  viste en sterk rettlinjet *samvariasjon*. Videre så vi at  $r \approx 0$  kunne bety at  $x$ - og  $y$ -dataene ikke viste noen sterk samvariasjon (( $x,y$ )-dataene lå «hulter til bulter» i spredningsdiagrammet), men det kunne også bety en sterk, *ikke-lineær* samvariasjon (f.eks. langs en parabel). Her skal vi se på analoge *teoretiske* fenomen (kap. 3.7 og 3.8) knyttet til *simultanfordelingen*  $f(x,y)$ . Vi skal med basis i denne definere når to variabler  $X$  og  $Y$  er *stokastisk uavhengige*, og vi skal se på den viktige forskjellen på *uavhengighet* og *ukorrelertethet* ( $X$  og  $Y$  er ukorrelerte dersom korrelasjonskoeffisienten  $\rho = 0$ ).

### 3.2 Diskrete stokastiske variabler

Det er dessverre ikke noe godt norsk ord for det engelske «stochastic». Vi skal derfor bruke ordet stokastisk. Innledningsvis kan det være nyttig å tenke på begrepet stokastisk som motsetningen til begrepet deterministisk:

stokastisk	=	uforutsigbar
deterministisk	=	forutsigbar

#### Stokastisk variabel (definisjon):

En stokastisk variabel,  $X$ , er en funksjon som er definert på et utfallsrom (utfallsrommet utgjør definisjonsområdet). For ethvert enkeltutfall,  $e$ , i utfallsrommet har  $X(e)$  en bestemt, *numerisk* verdi (dvs. et *tall*).  $X$  kan godt ha samme verdi for ulike enkeltutfall, men  $X$  kan kun ha én verdi for hvert enkeltutfall.

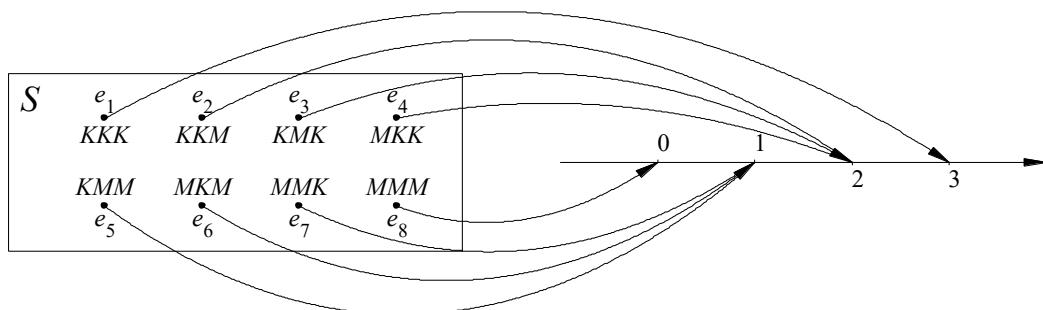


Fig. 3.1 Eks. på sammenheng mellom enkeltutfall,  $e$ , og stokastisk variabel,  $X(e) =$  antall kron ved 3 myntkast (se eks. 3.1).

**Eks. 3.1** **Tre kast med mynt.** Anta at vi kaster 3 kast med en og samme mynt. La  $K$  betegne kron og  $M$  betegne mynt. Dette er et eksperiment som er analogt til å trekke fra en urne *med* tilbakelegging. Vi ser nå på utfallsrommet over alle *ordnede* utvalg. Eksempelvis vil enkeltutfallet  $KKM$  bety kron i 1. og 2. kast, og mynt i 3. kast. Dette er en situasjon der vi kan bruke potensregelen for å finne antall enkeltutfall i utfallsrommet:  $m = 2$  (kun 2 utfall av hvert kast),  $k = 3$  (3 kast), totalt  $m^k = 2^3 = 8$  enkeltutfall. Sortert etter antall kron (nedover) får vi følgende enkeltutfall:

$$\begin{array}{llll} e_1 = KKK & e_2 = KKM & e_5 = KMM & e_8 = MMM \\ e_3 = KMK & e_6 = MKM & & \\ e_4 = MKK & e_7 = MMK & & \end{array}$$

Anta nå at det er lik sannsynlighet for kron og mynt i hvert kast, dvs.  $P(M) = P(K) = 1/2$ . Anta videre at utfallet av hvert kast er uavhengig av de andre kastene. Alle de 8 enkeltutfallene vil da være like sannsynlige, hver med sannsynlighet  $1/8$ . (Eks:  $P(KMK) = P(K) \cdot P(M) \cdot P(K) = (1/2) \cdot (1/2) \cdot (1/2) = 1/8$ ).

La en stokastisk variabel,  $X$ , være definert ved:

$$X = \text{antall kron etter 3 kast}$$

$X$  har da verdiene 3, 2, 1 og 0 i henholdsvis kolonne 1, 2, 3 og 4 ovenfor. Sammenhengen mellom enkeltutfallene til eksperimentet og den numeriske  $x$ -verdien hvert enkeltutfall tilordnes, er illustrert i figur 3.1 på forrige side. Lister vi opp de mulige verdiene for  $X$ , sammen med de tilhørende sannsynlighetene, kalles en slik tabell *sannsynlighetsfordelingen* til den stokastiske variablen,  $X$ :

Tab. 3.1 *Sannsynlighetsfordelingen til antall kron,  $x$ , i 3 myntkast*

Mulige $x$ -verdier:	0	1	2	3
Sannsynlighet, $f(x)$ :	1/8	3/8	3/8	1/8

Fra en sannsynlighetsfordeling som den ovenfor, kan vi regne ut sannsynligheten for forskjellige hendelser som avhenger av  $X$ .

$$\text{Eks: } P(X \geq 2) = P(X = 2) + P(X = 3) = 3/8 + 1/8 = 1/2.$$

$$P(0 \leq X \leq 2) = 1 - P(X = 3) = 1 - 1/8 = 7/8 \quad \text{☺}$$

I dette kapitlet skal vi kun betrakte stokastiske variabler som kan ha *adskilte* (distinkte) verdier. Slike variabler kaller vi *diskrete stokastiske variabler* (i motsetning til kontinuerlige stokastiske variabler som vi kommer til senere).

**NB!** Legg merke til at vi bruker *stor bokstav* (f.eks.  $X$ ) for å betegne en stokastisk variabel. Dette er for å understreke at vi på forhånd ikke kjenner utfallet til en stokastisk variabel. Med en gang den stokastiske variablene blir tilordnet en verdi i et eksperiment, er den ikke lenger stokastisk, og vi bruker da liten bokstav (f.eks.  $x$ ).

### 3.3 Sannsynlighetsfordeling

Vi begrenser oss, som i resten av kapitlet, til sannsynlighetsfordelingen til *diskrete* stokastiske variabler. I første omgang ser vi på én-dimensjonale sannsynlighetsfordelinger (dvs. fordelinger av én variabel), og går rett løs på en formell definisjon:

#### Sannsynlighetsfordeling (definisjon)

Sannsynlighetsfordelingen, eller enklere, fordelingen til en *diskret* stokastisk variabel,  $X$ , er en liste av alle verdier,  $x_i$ , som  $X$  kan ha, sammen med de tilhørende sannsynligheter,  $f(x_i) = P(X_i = x_i)$ . For en slik fordeling gjelder alltid at  $\sum f(x_i) = 1$ , der summen er tatt over alle  $x_i$ -verdier  $f(x)$  er definert for. Ofte vil en formel kunne erstatte en liste (se eks. 3.2), hvilket er nødvendig hvis  $X$  er uendelig tellbar.

**Eks. 3.2** **Surstrømming 3.** Si at sjansen for at en tilfeldig student i et forsøk skal svare ja ( $J$ ) på at han har smakt surstrømning er  $1/3$ , dvs.  $P(J) = 1/3$ . La oss videre betegne en negativ reaksjon med bokstaven  $N$  ( $N = \text{Nei}$ ).  $N$  er altså komplementhendelsen til  $J$ ,  $N = J^C$ . Eksperimentet går ut på å spørre en og en student inntil første ja-reaksjon. Utdalsrommet,  $S$ , blir da:

$$S = \{ J, NJ, NNJ, NNNJ, \dots \}$$

Vi lar nå  $X$  være en stokastisk variabel definert ved antall enkeltforsøk som må til i hvert eksperiment. Med andre ord,  $X$  kan ha verdiene 1, 2, 3,.... Som vi husker er en stokastisk variabel en funksjon definert på et utfallsrom. I vårt tilfelle blir  $X$  lik antall symboler i hvert enkeltutfall:  $X = 1$  tilsvarer enkeltutfallet  $J$ ,  $X = 2$  tilsvarer enkeltutfallet  $NJ$  osv. Sannsynlighetene for de ulike  $X$ -verdiene blir:

$$\begin{aligned}f(1) &= P(X=1) = P(J) = 1/3 \\f(2) &= P(X=2) = P(NJ) = P(N) \cdot P(J) = (1-P(J)) \cdot P(J) = 2/3 \cdot 1/3 \\f(3) &= P(X=3) = P(NNJ) = P(N) \cdot P(N) \cdot P(J) = (2/3)^2 \cdot 1/3\end{aligned}$$

Generelt får vi:

$$(3.1) \quad f(x) = P(X=x) = \left(\frac{2}{3}\right)^{x-1} \cdot \frac{1}{3}, \quad x = 1, 2, 3, \dots$$

I dette tilfellet kan  $X$  i prinsippet ha uendelig mange verdier (begrenset av antall studenter, riktig nok!). Det vil være umulig å sette opp en tabell over alle  $x$ -verdiene med tilhørende sannsynligheter. I dette tilfellet er det åpenbart mest rasjonelt å presentere sannsynlighetsfordelingen ved formelen ovenfor. ☺

En viktig egenskap ved en sannsynlighetsfordeling,  $f(x)$ , er at

$$(3.2) \quad \sum_{i=1}^k f(x_i) = 1$$

der summen er tatt over alle ( $k$ ) forskjellige  $x$ -verdier. La oss sjekke at dette stemmer for  $f(x)$  i eks. 3.2:

$$(3.3) \quad \sum_{i=1}^{\infty} \left(\frac{2}{3}\right)^{i-1} \cdot \frac{1}{3} = \frac{1}{3} \sum_{i=1}^{\infty} \left(\frac{2}{3}\right)^{i-1}$$

Siste sum gjenkjennes som en uendelig geometrisk rekke med første ledd  $a = (2/3)^0 = 1$  og  $k = 2/3$ :

$$(3.4) \quad \sum_{i=1}^{\infty} \left(\frac{2}{3}\right)^{i-1} = 1 \cdot \frac{1}{1 - 2/3} = 3$$

Vi skal gange summen med  $1/3$  og får da  $1/3 \cdot 3 = 1$ , som vi skulle ha!

**NB!** Vær klar over forskjellen på begrepet *sannsynlighetsfordeling*,  $f(x)$ , som er en ren *teoretisk* størrelse som *ikke* er basert på data, og begrepet *relativ frekvens*, som er en *eksperimentell* størrelse basert på data. Litt forenklet kan vi si at disse begrepene faller sammen dersom den relative frekvensen er basert på uendelig mange observasjoner. I praksis kan vi etter lang tids erfaring ha etablert en god tilnærmelse til en fordeling med basis i data.

### 3.4 Fordelingsdiagrammer

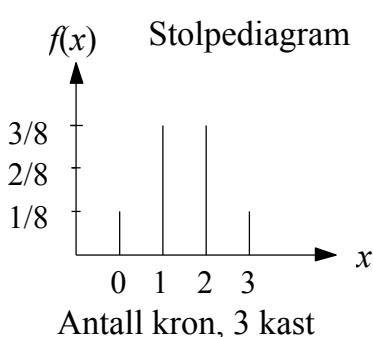
Vi skal her bare nevne 2 forskjellige grafiske fremstillingsformer som kan vise sannsynlighetsfordelinger for diskrete data:

#### 1) Sannsynlighets-stolpediagram (eng: «line diagram»):

Horisontal akse:  $x$

Vertikal akse: sannsynlighet,  $f(x)$

En heltrukken, loddrett strek (linje) ved hver  $x$ -verdi, der stolpehøyden angir sannsynligheten.



Figuren til venstre viser et stolpediagram over fordelingen gitt nedenfor ( $x$  = antall kron etter 3 kast med ekte mynt).

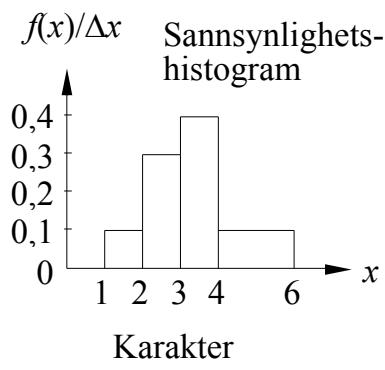
$x:$	0	1	2	3
$f(x):$	1/8	3/8	3/8	1/8

#### 2) Sannsynlighets-histogram (eng: «probability histogram»):

Horisontal akse:  $x$

Vertikal akse: sannsynlighet/rektangelbredde =  $f(x)/\Delta x$

Vertikale rektangler ved hver  $x$ -verdi, med hver  $x$ -verdi midt på rektangelbredden. Rektangelbredden er avstanden mellom 2 påfølgende  $x$ -verdier ( $\Delta x$ ).



Figuren til venstre viser et sannsynlighets-histogram over karakterfordelingen angitt nedenfor. Legg merke til at arealet av hvert rektangel er lik  $f(x)$ , slik at høyden blir  $f(x)/intervallbredde$ . Hva er strykprosenten?

$x:$	1.0-	2.0-	3.0-	4.1-
$f(x):$	0.1	0.3	0.4	0.2

Teknisk sett er det fullstendig analogi mellom stolpediagram/histogram for sannsynlighetsfordelinger og for relative frekvenser. Men husk igjen at en sannsynlighetsfordeling er teoretisk og ikke basert på data, mens relativ frekvens er eksperimentelt bestemt, dvs. på basis av data.

### 3.5 Forventning ( $\mu$ )

Det kan være nyttig å oversette begrepet forventning til «teoretisk middelverdi». Regneteknisk sett finner vi forventningsverdien akkurat som vi fant middelverdien basert på grupperte data, med den forskjell at vi erstatter relativ frekvens med sannsynlighetsfordelingen,  $f(x)$ .

Det er vanlig både i norsk og internasjonal statistisk litteratur å bruke betegnelsen  $E(X)$  (eller kortere:  $EX$ ) til å betegne forventningsverdien til  $X$  ( $E$  er forkortelse for det engelske begrepet Expectation, som betyr forventning). For å beregne forventningsverdien til en diskret stokastisk variabel,  $X$ , får vi da formelen i ramma nedenfor:

**Forventning,  $\mu = E(X)$**  (definisjon).

Vi øremarker betegnelsen  $E(X)$  til å bety forventningen til en stokastisk variabel,  $X$ , definert ved formelen

$$(3.5) \quad E(X) = \mu = \sum_{i=1}^k x_i f(x_i)$$

der  $k$  er antall forskjellige  $x_i$ -verdier. Vi har også innført den greske bokstaven  $\mu$  (tilsvarer den latinske bokstaven  $m$ ) som symbol for  $E(X)$ , hvilket også er svært utbredt i litteraturen.

**Eks. 3.3** **Stereoanlegg.**  $X$  = antall solgte enheter av type A stereoanlegg pr. uke.  $Y$  = antall solgte enheter av type B stereoanlegg pr. uke.

Data er vist i neste tabell. Her har vi 2 forskjellige stokastiske variabler,  $X$  og  $Y$ . I slike tilfeller er det vanlig å bruke subskript med stor bokstav tilknyttet fordelingsfunksjonen,  $f$ . Vi bruker samme bokstav i indeksen for å betegne den stokastiske variabel som  $f$  er fordelingen til. (Stor bokstav er brukt for å understreke at det dreier seg om en *sannsynlighetsfordeling*.)

Tab. 3.2 Sannsynlighetfordelinger for  $X$  og  $Y$ :

Type A:

$x$	0	1	2	3	4	5	
$f_X(x)$	.1	.1	.2	.3	.2	.1	
$x \cdot f_X(x)$	0	.1	.4	.9	.8	.5	$E(X) = 2.7$ (rekkesum)

Type B:

$Y$	0	1	2				
$f_Y(y)$	.23	.48	.29				
$y \cdot f_Y(y)$	0	.48	.58				$E(Y) = 1.06$ (rekkesum) ☺

En funksjon av en stokastisk variabel,  $G = g(X)$ , vil også være en stokastisk variabel. For å finne forventningsverdien til  $G$  kan vi gå fram på «vanlig» måte, dersom vi kjenner fordelingsfunksjonen til  $G$ ,  $f_G(g)$ :

$$(3.6) \quad E(G) = \sum_{i=1}^m g_i \cdot f_G(g_i)$$

der  $m$  er antall mulige forskjellige  $g$ -verdier. Vi kan imidlertid også finne  $E(G)$  på basis av fordelingsfunksjonen,  $f_X(x)$ , til  $X$ :

$$(3.7) \quad E(G) = \sum_{i=1}^k g(x_i) \cdot f_X(x_i)$$

der  $k$  er antall mulige forskjellige  $x$ -verdier ( $k$  kan være forskjellig fra  $m$ ). Den siste formelen er svært nyttig, som vi skal se, ikke minst når vi skal beregne variansen til  $X$ ,  $\text{Var}(X)$ .

Til slutt skal vi presisere at forventningsoperatoren,  $E$ , er en *lineær operator*, som innebærer at den tilfredsstiller følgende meget nyttige formel:

$$(3.8) \quad E\left(\sum_{i=1}^n a_i \cdot g(X_i)\right) = \sum_{i=1}^n E(a_i \cdot g(X_i)) = \sum_{i=1}^n a_i \cdot E(g(X_i))$$

der  $a_1, \dots, a_n$  er konstanter. Med andre ord: Forventningen til en sum er lik summen av forventningene.

<b>Eks. 3.4</b>	$x$	-1	0	1
	$f(x)$	.5	.4	.1

*Oppgave*

- a) Bestem forventningen  $E(X)$ .
- b) Bestem fordelingen til  $G = X^2$ .
- c) Bestem forventningen  $E(G)$  på to forskjellige måter.

*Løsningsforslag*

$$\begin{aligned} \text{a) } E(X) &= \sum x \cdot f(x) = (-1) \cdot 0.5 + 0 \cdot 0.4 + 1 \cdot 0.1 = \underline{0} \\ \text{b) } x = -1, 0 \text{ og } 1 \Rightarrow g = x^2 = 1, 0 \text{ og } 1: f_G(0) &= f_X(0) = \underline{0.4} \\ f_G(1) &= f_X(-1) + f_X(1) = 0.5 + 0.1 = \underline{0.6} \\ \text{c) } E(G) &= \sum g \cdot f_G(g) = 0 \cdot 0.4 + 1 \cdot 0.6 = \underline{0.6} \\ E(G) &= \sum x^2 \cdot f_X(x) = (-1)^2 \cdot 0.5 + 0^2 \cdot 0.4 + 1^2 \cdot 0.1 = \underline{0.6} \quad \odot \end{aligned}$$

**Eks. 3.5** Eksempler på operasjoner med forventningsoperatoren **E**

- i)  $E(a) = a$
- ii)  $E(bX) = bE(X)$
- iii)  $E(X+a) = E(X) + a$
- iv)  $E(a+bX) = a + bE(X)$
- v)  $E(a+bX+cX^2) = a + bE(X) + cE(X^2)$
- vi)  $E(50X-20) = 50E(X) - 20 = 50 \cdot 2.7 - 20 = 115$ , der  $E(X) = 2.7$  er hentet fra tab. 3.2.  $\odot$

**Eks. 3.6** **Pengespill.** I et pengespill er innsatsen kr. 10 for hvert spill. Det er en premie for hvert spill, og den lyder på kr. 1000. Sannsynligheten for å vinne er  $p = 0,01$  for hvert spill.

*Oppgave*

Beregn forventet gevinst,  $E(G)$ , når du spiller én gang.

*Løsningsforslag*

La  $X$  betegne antallet gevinster.  $X$  kan da ha verdiene 0 eller 1, og fordelingen  $f(x)$  til  $X$  blir som følger:

$$f(0) = P(X=0) = P(\text{ingen gevinst}) = 1-p = (1 - 0,01) = 0,99$$

$$f(1) = P(X=1) = P(\text{gevinst}) = p = 0,01$$

Sammenhengen mellom gevinst  $G$  og  $X$  blir:  $G = 1000 \cdot X$  (hvorfor?), og vi får:

$$\begin{aligned} E(G) &= E(1000 \cdot X) = 1000 \cdot E(X) = 1000 \cdot \sum x f(x) = 1000 \cdot (0 \cdot f(0) + 1 \cdot f(1)) \\ &= 1000 \cdot (0 \cdot 0,99 + 1 \cdot 0,01) = \underline{\text{kr. } 10} \quad \circledast \end{aligned}$$

### 3.6 Varians ( $\sigma^2$ ) og standardavvik ( $\sigma$ )

Vi har tidligere brukt variansbegrepet anvendt på grupperte (og ugrupperte) data. Her skal vi se på den teoretiske versjonen av variansbegrepet. For å skille disse to tilfellene, er det vanlig å benytte begrepet «empirisk varians» når vi ser på spredning av et datamateriale. Vi skiller også ved å bruke forskjellige symboler for de 2 tilfellene, akkurat som vi bruker forskjellige symboler for middelverdi og forventning.

Som symbol for empirisk standardavvik og varians brukte vi  $s$  og  $s^2$ . Her skal vi bruke den greske varianten,  $\sigma$  og  $\sigma^2$ , for henholdsvis (teoretisk) standardavvik og varians:

#### Varians, $\sigma^2$ , og standardavvik, $\sigma$ (definisjon)

$$(3.9) \quad \text{Var}(X) = \sigma^2 = E(X - \mu)^2 = E(X^2) - \mu^2$$

$$(3.10) \quad \text{std}(X) = \sigma = \sqrt{\text{Var}(X)}$$

(Var er forkortelse for varians og std er forkortelse for det engelske uttrykket for standardavvik: std = standard deviation).

NB! Varians og standardavvik kan *aldri* ha negativ verdi.

Merk det siste uttrykket for  $\text{Var}(X)$ , som ofte er *meget* nyttig å bruke. Fra forrige avsnitt husker vi jo at  $E(g(X)) = \sum g(x) \cdot f(x)$ , og vi får derfor at  $E(X^2) = \sum x^2 \cdot f(x)$ .

### Variasjonskoeffisient, $CV(x)$ (CV: «Coefficient of Variation»)

I mange tilfeller vet vi at standardavviket  $\sigma = \text{std}(X)$  til en stokastisk variabel  $X > 0$  er liten i forhold til forventningen  $\mu = E(X)$ :  $\sigma \ll \mu$ . I slike tilfeller vil variasjonskoeffisienten ofte være et godt mål på *relativ usikkerhet*:

$$(3.11) \quad CV(X) = \frac{\text{std}(X)}{E(X)} = \frac{\sigma}{\mu}$$

La oss som et eksempel tenke oss en vekt med noe unøyaktig avlesning. Etter lang tids erfaring viser det seg at standardavviket til en enkeltmåling (basert på gjentatte målinger av en kjent vekt, f.eks. et lodd) er  $\sigma = 0,0015$  kg. For en gjenstand som veier nøyaktig 1,5 kg vil da variasjonskoeffisienten være  $\sigma/\mu = 0,0015/1,5 = 0,1\%$ , og det er rimelig å si at en gjenstand på ca. 1,5 kg vil bli veiet med en relativ nøyaktighet på ca. 0,1 %.

**Eks. 3.7** **Var( $X$ )**. Vi skal finne  $\text{Var}(X)$  i eks. 3.3, og utvider tab. 3.2 ved å tilføye en  $x^2 \cdot f(x)$ -rekke:

Tab. 3.3 (utvidelse av tab. 3.2 med en  $x^2 \cdot f(x)$ -rekke):

$x$	0	1	2	3	4	5	total:	
$f(x)$	.1	.1	.2	.3	.2	.1	1	
$xf(x)$	0	.1	.4	.9	.8	.5	2.7	$= E(X) = \mu$
$x^2f(x)$	0	.1	.8	2.7	3.2	2.5	9.3	$= E(X^2)$

$$\text{Var}(X) = E(X^2) - \mu^2 = 9.3 - 2.7^2 = 2.01$$

$$\text{std}(X) = \sqrt{2.01} = 1.42 \odot$$

Noen generelle og *viktige* egenskaper ved varians og standardavvik:

#### Varians

- i)  $\text{Var}(X)$  kan aldri være negativ
- ii)  $\text{Var}(X+a) = \text{Var}(X)$ ,  $a = \text{konstant}$
- iii)  $\text{Var}(bX) = b^2 \cdot \text{Var}(X)$
- iv)  $\text{Var}(a+bX) = b^2 \cdot \text{Var}(X)$

#### Standardavvik

- i)  $\text{std}(X)$  kan aldri være negativ
- ii)  $\text{std}(X+a) = \text{std}(X)$
- iii)  $\text{std}(bX) = |b| \cdot \text{std}(X)$
- iv)  $\text{std}(a+bX) = |b| \cdot \text{std}(X)$

Til slutt skal vi definere begrepet *standardisert stokastisk variabel*,  $Z$ , samt dens egenskaper:

$$(3.12) \quad Z = \frac{X - \mu_X}{\sigma_X} \Rightarrow E(Z) = 0, \quad \text{Var}(Z) = \text{std}(Z) = 1$$

(Bevis selv at  $E(Z) = 0$  og  $\text{Var}(Z) = 1$  ved hjelp av definisjon og egenskaper til forventning og varians).

### 3.7 Simultanfordeling (2 variabler)

I et eksperiment tenker vi oss at vi kan måle verdien til 2 stokastiske variabler,  $X$  og  $Y$ , samtidig. La oss si at  $X$  kan ha verdiene  $x_1, \dots, x_k$  og at  $Y$  kan ha verdiene  $y_1, \dots, y_m$ . Vi får da totalt  $k \cdot m$  forskjellige verdipar  $(x_i, y_j)$ ,  $i = 1, \dots, k$  og  $j = 1, \dots, m$ , for  $(X, Y)$ .

La  $f(x_i, y_j)$  betegne sannsynligheten for at  $X$  og  $Y$  samtidig skal ha henholdsvis verdiene  $x_i$  og  $y_j$ , eller:

$$(3.13) \quad f(x_i, y_j) = P(X = x_i, Y = y_j)$$

Analogt med det en-dimensjonale tilfellet får vi følgende *definisjon*:

#### Simultanfordeling, $f(x, y)$

Den simultane sannsynlighetsfordelingen til 2 diskrete stokastiske variabler,  $X$  og  $Y$ , er en 2-vegs tabell som viser alle forskjellige verdier for  $X$  den ene vegen og alle forskjellige verdier for  $Y$  den andre vegen. Hver  $(x_i, y_j)$ -celle inneholder sannsynligheten  $f(x_i, y_j) = P(X = x_i, Y = y_j)$ . Celle-sannsynlighetene er ofte representert ved en formel istedet for en 2-vegs tabell, hvilket er nødvendig dersom  $X$  eller  $Y$  er uendelig tellbar.

**Eks. 3.8**

**Arbeidsfravær.**  $X$  = Antall fravær ved morgenskift  
 $Y$  = Antall fravær ved kveldsskift samme dag

I det følgende skal vi liste opp forskjellige typer informasjon vi kan få ut av en simultan sannsynlighetsfordeling, med utgangspunkt i tab. 3.4.

Tab. 3.4 Simultan sannsynlighetsfordeling til  $X$  og  $Y$ :

$x \downarrow$	$y \rightarrow$	0	1	2	3	Rekkesum, $f_X(x)$ :
0		.05	.05	.10	0	.20
1		.05	.10	.25	.10	.50
2		0	.15	.10	.05	.30
KolonneSUM, $f_Y(y)$ :		.10	.30	.45	.15	1

- a) Sannsynligheten for en *hendelse* som involverer  $X$  og  $Y$ , for eksempel hendelsen « $X + Y = 3$ ». For å beregne  $P(X + Y = 3)$  fra tab. 3.4 går vi i dette tilfellet langs «diagonalen» som består av cellene  $(x,y) = (2,1), (1,2)$  og  $(0,3)$ , og får sannsynligheten:

$$P(X + Y = 3) = 0.15 + 0.25 + 0 = 0.40$$

Tilsvarende kunne vi funnet sannsynlighetene:

$$P(X = Y) = 0.05 + 0.10 + 0.10 = 0.25$$

$$P(X = 2) = 0 + 0.15 + 0.10 + 0.05 = 0.30$$

$$P(X > Y) = 0.05 + 0 + 0.15 = 0.20$$

- b) *Sannsynlighetsfordelingen* til en *funksjon* av  $X$  og  $Y$ . La  $Z = X + Y$  betegne totalt antall fraværende på de 2 skiftene. De mulige verdiene for  $Z$  er 0, 1, 2, 3, 4 og 5. Vi viste i a) hvordan vi skulle bestemme sannsynligheten for at  $X + Y = 3$ , dvs.  $f_Z(3) = 0.40$ . Fordelingen til  $Z$  blir:

Tab.3.5 Fordelingen til  $Z = X + Y$ 

$z$	0	1	2	3	4	5	total
$f_Z(z)$	.05	.10	.20	.40	.20	.05	1.00

- c) De *marginale* sannsynlighetsfordelinger,  $f_X(x)$  og  $f_Y(y)$ . Vi finner disse simpelthen ved å ta henholdsvis rekkesum ( $f_X(x)$ ) og kolonneSUM ( $f_Y(y)$ ). Her må vi være litt obs. på om vi har  $x$  bortover og  $y$  nedover i tabellen, eller omvendt.
- d) *Forventningsverdi* og *standardavvik* til  $X$  og  $Y$ . Her går vi fram på vanlig måte ved å bruke de marginale fordelingene til  $X$  og  $Y$ . Vi går igjen tilbake til tab. 3.4:

Tab.3.6: Forventninger til  $X, X^2, Y$  og  $Y^2$  for fordelingen i tab. 3.4

$x$	0	1	2	tot	$y$	0	1	2	3	tot
$f_X(x)$	.2	.5	.3	1.0	$f_Y(y)$	.10	.30	.45	.15	1.00
$xf_X(x)$	0	.5	.6	1.1	$yf_Y(y)$	0	.30	.90	.45	1.65
$x^2f_X(x)$	0	.5	1.2	1.7	$y^2f_Y(y)$	0	.30	1.80	1.35	3.45

$$\mu_X = 1.1, \quad \sigma_X^2 = 1.7 - 1.1^2 = 0.49, \quad \sigma_X = \sqrt{0.49} = 0.70$$

$$\mu_Y = 1.65, \quad \sigma_Y^2 = 3.45 - 1.65^2 = 0.7275, \quad \sigma_Y = \sqrt{0.7275} = 0.85$$

e) Forventning av sum = sum av forventninger.

Vi etablerte under punkt b) sannsynlighetsfordelingen til  $Z = X+Y$ . På basis av denne kan vi finne  $E(Z) = \sum z f_Z(z) = .1 + .4 + 1.2 + .8 + .25 = 2.75$ . Ifølge reglene for forventning skulle vi imidlertid fått det samme ved å summere forventningene til  $X$  og  $Y$ :  $E(Z) = E(X+Y) = E(X) + E(Y) = \mu_X + \mu_Y = 1.1 + 1.65 = 2.75$ , så det stemte! ☺

### 3.8 Kovarians og korrelasjon

Kovariansen mellom  $X$  og  $Y$  er et mål på samvariasjonen til de 2 variablene, definert ved forventningen til  $(X-\mu_X) \cdot (Y-\mu_Y)$ :

$$(3.14) \quad \text{Cov}(X, Y) = E((X - \mu_X) \cdot (Y - \mu_Y)) = E(XY) - \mu_X \mu_Y$$

For å beregne  $E(XY)$  bruker vi formelen

$$(3.15) \quad E(XY) = \sum x_i y_j f(x_i, y_j)$$

der summen går over alle kombinasjoner av  $(x_i, y_j)$ , dvs. vi tar celle for celle i tabellen over den simultane sannsynlighetsfordelingen til  $X$  og  $Y$ , og for hver celle multipliserer vi  $x$ -verdien,  $y$ -verdien og sannsynlighetsverdien i tabellen. (Vi kan hoppe over de cellene der enten  $x$ -verdien,  $y$ -verdien eller sannsynligheten er null).

### Korrelasjonskoeffisienten, $\rho = \text{Corr}(X, Y)$

$\rho$  tilsvarer en standardisering av kovariansen,  $\text{Cov}(X, Y)$ :

$$(3.16) \quad \rho = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y} = \frac{\text{E}(XY) - \mu_X \mu_Y}{\sigma_X \sigma_Y}$$

og har følgende egenskaper:

- a)  $\rho$  er alltid et tall mellom  $-1$  og  $1$ . De 2 ekstreme verdiene  $+1$  og  $-1$  fås når det er en fullstendig rettlinjet sammenheng mellom  $X$  og  $Y$  (henholdsvis  $Y = a + bX$  og  $Y = a - bX$ , der  $b > 0$ ).
- b)  $\text{Corr}(a + bX, c + dY) = \text{Corr}(X, Y)$  dersom  $b$  og  $d$  har samme fortegn  
( $= -\text{Corr}(X, Y)$  dersom  $b$  og  $d$  har motsatt fortegn).

Vi kan si at korrelasjon (og kovarians) er mål på lineær samvariasjon. Dersom  $\rho$  er i nærheten av  $1$  eller  $-1$ , betyr det normalt at en rett linje vil være godt tilpasset våre  $(X, Y)$ -observasjoner med henholdsvis positivt og negativt stigningstall.

Dersom  $\rho$  derimot er i nærheten av  $0$ , er konklusjonen mer usikker. Enten vil tilfeldige  $(X, Y)$ -observasjoner ligge tilfeldig spredt i et spredningsdiagram (ingen samvariasjon), eller så vil de samvariere på en ikke-lineær måte.

(NB! Sammenhengen mellom  $\rho$  og den empiriske korrelasjonskoeffisienten,  $r$ , fra kap.1, er at  $\rho$  er teoretisk, mens  $r$  er basert på data. Når antall observasjoner går mot uendelig, vil  $r$  nærme seg  $\rho$ ).

Generelt kan vi sette opp følgende formler for forventning og varians til summen av to stokastiske variabler:

### Forventning og varians til summen av to variabler

La  $X$  og  $Y$  betegne to stokastiske variabler med forventninger  $\mu_X = E(X)$  og  $\mu_Y = E(Y)$ , varianser  $\text{Var}(X) = \sigma_X^2$  og  $\text{Var}(Y) = \sigma_Y^2$  og korrelasjonskoeffisient,  $\rho = \text{Corr}(X, Y)$ . Da gjelder generelt:

$$(3.17) \quad E(aX + bY) = a \cdot \mu_X + b \cdot \mu_Y$$

$$(3.18) \quad \begin{aligned} \text{Var}(aX + bY) &= a^2 \cdot \text{Var}(X) + b^2 \cdot \text{Var}(Y) + 2ab \cdot \text{Cov}(X, Y) \\ &= a^2 \cdot \sigma_X^2 + b^2 \cdot \sigma_Y^2 + 2ab \cdot \rho \sigma_X \sigma_Y \end{aligned}$$

#### Eks. 3.9

#### Beregning av kovarians og korrelasjonskoeffisient.

Vi betrakter følgende simultanfordeling,  $f(x,y)$ :

$x \setminus y$	0	1	2	$f_X(x)$
1	.15	.15	.00	.3
2	.35	.15	.20	.7
$f_Y(y)$	.5	.3	.2	

#### Oppgave

- Beregn korrelasjonskoeffisienten,  $\rho = \text{Corr}(X, Y)$ .
- Beregn  $\text{Var}(X - 2Y)$

#### Løsningsforslag

- Vi finner først forventningene,  $\mu_X$  og  $\mu_Y$ , og standardavvikene,  $\sigma_X$  og  $\sigma_Y$ , til henholdsvis  $X$  og  $Y$ , og finner utifra tabellen (prøv selv!) følgende verdier:

$$\mu_X = 1.7, \quad \mu_Y = 0.7, \quad \sigma_X = 0.458, \quad \sigma_Y = 0.781$$

Vi finner  $E(XY)$  ved å summere over alle cellene:

$$\begin{aligned} E(XY) &= \sum xy \cdot f(x,y) \\ &= 1 \cdot 0 \cdot 0.15 + 1 \cdot 1 \cdot 0.15 + 1 \cdot 2 \cdot 0 + 2 \cdot 0 \cdot 0.35 + 2 \cdot 1 \cdot 0.15 + 2 \cdot 2 \cdot 0.20 \\ &= 0.15 + 0.30 + 0.80 = 1.25 \end{aligned}$$

Når vi har funnet  $E(XY)$ ,  $\mu_X$  og  $\mu_Y$ , kan vi finne kovariansen,  $\text{Cov}(X, Y)$ :

$$\text{Cov}(X, Y) = E(XY) - \mu_X \mu_Y = 1.25 - 1.7 \cdot 0.7 = 0.06$$

Når vi i tillegg har standardavvikene,  $\sigma_X$  og  $\sigma_Y$ , kan vi finne korrelasjonskoeffisienten,  $\rho$ :

$$\rho = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y} = \frac{0.06}{0.458 \cdot 0.781} = \underline{0.17}$$

b) For å finne  $\text{Var}(X - 2Y)$ , benytter vi formel (3.18):

$$\text{Var}(X - 2Y) = \text{Var}(X) + 2^2 \text{Var}(Y) - 2 \cdot 2 \cdot \text{Cov}(X, Y) = \underline{2.41} \quad \circlearrowright$$

### 3.9 Uavhengighet mellom 2 variabler

Vi har tidligere definert uavhengighet mellom 2 hendelser,  $A$  og  $B$ , som ekvivalent med at  $P(AB) = P(A) \cdot P(B)$ . Vi kan betrakte den situasjon at  $X$  får en verdi,  $x_i$ , og at  $Y$  får en verdi,  $y_j$ , som de 2 hendelsene  $A$  og  $B$ . Vi får da at disse hendelsene er uavhengige hvis og bare hvis  $P(X = x_i, Y = y_j) = P(X = x_i) \cdot P(Y = y_j)$ . Vi får derfor:

#### Uavhengighet mellom 2 variabler (definisjon)

2 stokastiske variabler  $X$  og  $Y$  er uavhengige hvis og bare hvis

$$(3.19) \quad f_{X,Y}(x_i, y_j) = f_X(x_i) \cdot f_Y(y_j)$$

for alle forskjellige  $(x_i, y_j)$ -verdier i simultanfordelingen til  $X$  og  $Y$ . Vi må med andre ord kontrollere *alle* cellene, og å påse at hver celle-sannsynlighet er lik produktet av de to tilhørende rekke- og kolonne-summer (marginal-sannsynlighetene), før vi kan konkludere med at  $X$  og  $Y$  er uavhengige. Det er tilstrekkelig å påvise forskjell i én celle for å konkludere at  $X$  og  $Y$  ikke er uavhengige.

Merk forskjellen på at  $X$  og  $Y$  er *uavhengige* og at  $X$  og  $Y$  er *ukorrelerte*:

- Dersom  $X$  og  $Y$  er uavhengige, vil alltid  $\text{Cov}(X, Y)$  og  $\text{Corr}(X, Y)$  være lik null.
- Dersom  $\text{Cov}(X, Y) = 0$  er *ikke* nødvendigvis  $X$  og  $Y$  uavhengige.

Dersom én cellesannsynlighet i en simultanfordeling er forskjellig fra produktet av de tilhørende marginalsannsynligheter,  $f(x_i, x_j) \neq f_X(x_i) \cdot f_Y(y_j)$ , så er dette tilstrekkelig til å konkludere at  $X$  og  $Y$  ikke er uavhengige.

**Eks. 3.10 Ukorrelerte, men avhengige variabler**

Kontrollér selv at  $X$  og  $Y$  er ukorrelerte (dvs.  $\text{Cov}(X, Y) = 0$ ).

$x/y$	0	1	2	$f_X(x)$
0	.2	.2	.2	.6
1	.2	0	.2	.4
$f_Y(y)$	.4	.2	.4	

Fra første celle ser vi at  $f(0,0) = 0.2$ , mens  $f_X(1) \cdot f_Y(1) = 0.6 \cdot 0.4 = 0.24 \neq 0.2$ . Følgelig:  $X$  og  $Y$  er avhengige siden vi har funnet en cellesannsynlighet som er forskjellig fra produktet av de tilhørende marginalsannsynligheter. ☺

### 3.10 Oppgaver

**3.1** La  $X$  være en stokastisk variabel definert ved antall kron etter 2 kast med en rettferdig mynt. La  $K$  betegne kron og  $M$  betegne mynt i ett enkelt kast.

- a) Skriv opp utfallsrommet,  $S$ .
- b) Bestem fordelingen til  $x, f(x)$ .

**3.2** La  $X$  være antall kast inntil første kron med en rettferdig mynt. Bestem fordelingen,  $f(x)$ .

**3.3** La  $X$  være antall seksere ved ett terningkast med 3 rettferdige terninger. Finn fordelingen,  $f(x)$ .

**3.4** La  $X$  være antall kast med en rettferdig terning inntil du får en sekser. Bestem fordelingen,  $f(x)$ .

**3.5** Tegn stolpediagram for følgende fordeling,  $f(x)$ :

$x:$	0	1	2	3	8
$f(x):$	.1	.3	.4	.1	.1

**3.6** Tegn sannsynlighetshistogram for følgende fordeling,  $g(y)$ :

$y:$	-20	0	20	40	60
$g(y):$	.1	.5	.3	0	.1

**3.7** Finn forventningene  $\mu_X = E(X)$  og  $\mu_Y = E(Y)$ , der de tilhørende fordelingene er definert i henholdsvis oppgave 3.5 og 3.6. Bestem deretter  $E(2X - 5Y)$ .

**3.8** Bestem standardavvikene,  $\sigma_X = \text{std}(X)$ , og  $\sigma_Y = \text{std}(Y)$ , der fordelingene til  $X$  og  $Y$  er gitt i henholdsvis oppgave 3.5 og 3.6. Bestem også  $E(X^3)$ .

**3.9 (E)** Vi skal betrakte TILs første 2 hjemmekamper neste sesong. La  $H$  betegne hjemmeseier for TIL, la  $U$  betegne uavgjort og la  $B$  betegne borte seier for TILs motstander (TIL-tap). Enkeltutfallet  $HU$  betyr for eksempel TIL-seier i 1. kamp og uavgjort i 2. kamp.

- a) Bestem utfallsrommet  $S$  for utfallet av TILs første 2 hjemmekamper neste sesong.

Vi gjør nå følgende antagelser: Utfallene av de 2 kampene er uavhengige av hverandre, og sannsynligheten for seier, uavgjort og tap for TIL i de 2 kampene er som følger:

- 1. kamp:  $P(H) = .7$   $P(U) = .2$   $P(B) = .1$
- 2. kamp:  $P(H) = .5$   $P(U) = .2$   $P(B) = .3$
- b) Finn sannsynligheten for hvert enkeltutfall i utfallsrommet  $S$ .
- c) La  $X$  være antall TIL-seire i de første to hjemmekampene, og bestem fordelingen til  $X$ .

**3.10** Gitt følgende simultanfordeling,  $f(x,y)$ :

$y \setminus x$	0	1	2
0	0.1	0	0.2
1	0.2	0.4	0.1

- a) Bestem marginalfordelingene  $f_X(x)$  og  $f_Y(y)$ .
- b) Beregn forventning og varians til  $X$  og  $Y$ .
- c) Er  $X$  og  $Y$  uavhengige? Begrunn svaret.
- d) Beregn fordelingen til  $Z = X - Y$ .
- e) Beregn forventning og varians til  $Z$ .

**3.11** La  $\text{Var}(X) = 4$ ,  $\text{Var}(Y) = 9$  og  $\text{Cov}(X,Y) = -3$ .

- Bestem korrelasjonskoeffisienten,  $\rho(X,Y) = \text{Corr}(X,Y)$ .
- Bestem  $\text{Var}(2X+3Y)$ .
- Er  $X$  og  $Y$  uavhengige?

**3.12** I mangel av noe bedre å gjøre sitter noen studenter og kaster terning. Spillet er: En terning kastes til man får 1 eller 6.

- Finn sannsynligheten for at spillet stopper etter første kast.
- Vis at sannsynligheten for at spillet stopper etter 3 eller færre kast er  $19/27$ .
- Anta at man spiller dette spillet med en innsats på kr. 100. Man vinner kr 100 hvis antall kast som må utføres er forskjellig fra 2 og 3 og taper innsatsen ellers. Finn forventet gevinst.

**3.13 (E)** En sannsynlighetsfordeling er gitt ved:

$x$	0	1	2	3	4
$P(X=x)$	$2k$	$3k$	$k$	$3k$	$2k$

- Bestem verdien av konstanten  $k$ .
- Finn forventning og standardavvik.

**3.14 (E)** I en modell for olje- og gassreservoarer er det laget prognosenter for oljemengde og gassmengde. Vi betegner oljemengde med  $X$  og gassmengde med  $Y$ . De stokastiske variablene  $X$  og  $Y$  antas å ha følgende fordelinger (i passende enheter):

$x$	12	15	18
$P(X=x)$	$1/3$	$1/3$	$1/3$
$y$	21	27	33
$P(Y=y)$	$1/4$	$1/2$	$1/4$

- Finn forventet oljemengde  $E(X)$  og finn forventningen  $E(X/2-5)$ .
- Finn variansene  $\text{Var}(X)$  og  $\text{Var}(X/2-5)$ .
- Gitt at  $P(X=15 \cap Y=27) = 1/4$ . Er  $X$  og  $Y$  uavhengige? Hva er betinget sannsynlighet  $P(X=15 | Y=27)$ ? Er det mulig å regne ut  $E(X+Y)$  og  $\text{Var}(X+Y)$  med de gitte opplysningene?
- Drivverdigheten av olje- og gassfeltet avhenger av en rekke andre faktorer. Hvis oljemengden er hhv 12,15 og 18, regner en sannsynligheten for at feltet er drivverdig er hhv .6 .8 og .9. Bruk disse opplysningene og fordelingen til  $X$  for å beregne sannsynligheten for at feltet er drivverdig.

## 3.11 Formelsamling

### Betegnelser

$f(x)$  er sannsynlighetsfordeling til  $X$ ,  
 $f(x,y)$  er simultanfordeling til  $X$  og  $Y$ .  
 $n$  = antall forskjellige  $x$ -verdier.

### Forventning, $E(X) = \mu$ .

$$E(X) = \mu = \sum_{i=1}^n x_i f(x_i)$$

### Forventning, $E(g(X))$

La  $G = g(X)$  være en funksjon av  $X$ .  
 $E(G) = E(g(X))$  kan finnes som følger:

$$E(g(X)) = \sum_{i=1}^n g(x_i) \cdot f_X(x_i)$$

der  $f_X(x)$  er marginalfordelingen til  $x$ ,  
eller alternativt:

$$E(G) = \sum_{i=1}^m g_i \cdot f_G(g_i)$$

der  $f_G(g)$  er marginalfordelingen til  $G$  og  
 $m$  = antall forskjellige  $g$ -verdier.

### Forventning til sum

$$E \sum_{i=1}^n a_i Y_i = \sum_{i=1}^n a_i E(Y_i)$$

der  $a$ 'ene er konstanter.

### Varians, $\text{Var}(X) = \sigma^2$

$$\begin{aligned} \text{Var}(X) &= \sigma^2 = E(X - \mu)^2 \\ &= E(X^2) - \mu^2 = \sum_{i=1}^n x_i^2 f(x_i) - \mu^2 \end{aligned}$$

### Standardavvik, $\text{std}(X) = \sigma$

$$\text{std}(X) = \sigma = \sqrt{\text{Var}(X)}$$

### Marginalfordeling, $f_X(x_i)$

$$f_X(x_i) = \sum_j f(x_i, y_j)$$

der summen er tatt over alle forskjellige  $y_j$ -verdier.

### Marginalfordeling, $f_Y(y_j)$

$$f_Y(y_j) = \sum_i f(x_i, y_j)$$

der summen er tatt over alle forskjellige  $x_i$ -verdier.

### Kovarians, $\text{Cov}(X, Y)$

$$\text{Cov}(X, Y) = E(XY) - \mu_X \mu_Y$$

$$\text{der } E(XY) = \sum_{i,j} x_i y_j \cdot f(x_i, y_j)$$

og summen er tatt over alle forskjellige  $(x_i, y_j)$ -verdier og  $\mu_X = E(X)$  og  $\mu_Y = E(Y)$ .

### Korrelasjonskoeffisient, $\rho = \text{Corr}(X, Y)$

$$\rho = \text{Corr}(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y}$$

der  $\sigma_X = \text{std}(X)$  og  $\sigma_Y = \text{std}(Y)$ .

### Uavhengighet mellom $X$ og $Y$

$X$  og  $Y$  uavhengige  $\Leftrightarrow$

$$f(x_i, y_j) = f_X(x_i) \cdot f_Y(y_j)$$

for alle  $(x_i, y_j)$

### Varians til sum

$$\begin{aligned} \text{Var}(aX + bY) &\quad (a \text{ og } b \text{ konstanter}) \\ &= a^2 \text{Var}(X) + b^2 \text{Var}(Y) + 2ab \cdot \text{Cov}(X, Y) \end{aligned}$$

## Kapittel 4

# Diskrete fordelinger

### 4.1 Innledning

Med begrepet diskrete fordelinger mener vi her fordelingen til diskrete stokastiske variabler, dvs. stokastiske variabler som kun kan ha diskrete (adskilte) verdier. Vi skal kun se på fordelinger der den stokastiske variabel,  $X$ , kan innehå ikke-negative heltallsverdier ( $x = 0, 1, 2, \dots$ ). Vi har allerede fra tidligere kapitler grunnlag for å utlede de kjente fordelingene som vil bli behandlet i dette kapitlet: Den **binomiske** fordeling, den **hypergeometrisk** fordeling og **Poisson**-fordelingen. Disse fordelinger kalles også for **tellefordelinger**.

De tre fordelinger som behandles i kapitlet har viktige likheter og forskjeller. Alle fordelinger betrakter kun binære hendelser, dvs. enten måler vi en viss egenskap med en eksperimentell enhet (for eksempel: defekt), eller så måler vi den komplementære egenskapen (for eksempel: intakt, eller ikke defekt).

Den binomiske fordeling og Poisson-fordelingen tilsvarer «med tilbakelegging»-situasjoner. Når antall observasjoner,  $n$ , er stort, og sannsynligheten,  $p$ , for at en enhet innehar en gitt egenskap er svært liten, er den binomiske fordeling tilnærmet lik Poisson-fordelingen. Den hypergeometriske fordeling tilsvarer «uten tilbakelegging»-situasjoner. For små utvalgstørrelser,  $n$ , er i praksis ofte den hypergeometriske fordeling tilnærmet lik den binomiske fordeling.

Overgangen mellom diskrete og kontinuerlige stokastiske variabler kan være temmelig «glidende». I mange tilfeller bruker vi faktisk formler for kontinuerlige fordelinger som tilnärmelser til de diskrete, der tilnärmelsene er svært gode. Vi skal i neste kapittel se på normalfordelings-tilnärmelsen til den binomiske fordeling og til Poisson-fordelingen.

Alle fordelinger i dette kapitlet har utstrakt anvendelse i samfunnslivet. Gjennom de eksempler som blir vist og de oppgaver du vil jobbe med, er det å håpe at du ser nytteverdien av disse fordelingene, samt blir i stand til å anvende dem selv der det måtte bli behov.

Da vi ofte framover vil bruke begrepene **populasjon** og **tilfeldig utvalg**, skal vi starte med å definere disse *meget* sentrale begrepene, og belyse dem med noen enkle eksempler.

**Populasjon og tilfeldig utvalg** (definisjon)

Anta at vi har totalt  $N$  aktuelle eksperimentelle enheter (individer) der hver enhet har en målbar egenskap vi vil studere. **Populasjonen** består da pr. definisjon av mengden av alle de individuelle måleresultatene vi ville funnet dersom alle  $N$  enheter ble undersøkt (målt).

Formålet med en statistisk undersøkelse er å trekke slutsninger om en populasjon som helhet på basis av et begrenset utvalg av størrelse  $n < N$  fra populasjonen. Utvalget kalles et **tilfeldig utvalg** dersom enhver mulig sammensetning av de  $n$  (forskjellige) enhetene i utvalget har like stor sannsynlighet  $1/\binom{N}{n}$  for å bli valgt.

**Eks. 4.1** Registrering av kjønn*Oppgave*

Anslå prosentandel kvinner i en befolkning på  $N = 4$  millioner innbyggere på basis av registrering av kjønnet til et tilfeldig utvalg på  $n = 1000$  personer.

*Løsningsforslag*

Her består populasjonen *ikke* av innbyggerne selv, men av de 4 millioner mulige registreringer,  $S = \{ F, M, M, \dots, F \}$ , vi ville funnet dersom vi registrerte kjønnet til alle innbyggerne ( $F$  = hokjønn,  $M$  = hankjønn). Har vi et dataregister over alle innbyggerne, kan vi enkelt finne en måte å plukke ut 1000 personer på, slik at kombinasjonen av de tilsvarende kjønnsregistreringene er like sannsynlig som en hvilken som helst annen kombinasjon av 1000 kjønnsregistreringer (stikkord: slumptallgenerator). ☺

## 4.2 Binomisk fordeling

Den binomiske fordeling (også kalt binomialfordelingen) kommer til anvendelse når det er rimelig å si at en statistisk undersøkelse består av  $n$  **Bernoulli**-forsøk:

**Bernoulli-forsøk** (definisjon).

Vi har  $n$  Bernoulli-forsøk dersom:

- 1) Hvert forsøk har bare 2 utfall:  $J$  eller  $N$  ( $J$  for Ja og  $N$  for Nei).
- 2) Sannsynligheten for positivt utfall ( $J$ ) er lik i hvert eksperiment:

$$P(J) = p.$$

- 3) Utfallene av de enkelte forsøkene er uavhengig av hverandre.

Vi innser at  $n$  Bernoulli-forsøk tilsvarer en urnemodell der vi har trekning med tilbakelegging blant  $n$  lapper der  $np$  av dem er merket  $J$  og  $n(1-p)$  av lappene er merket  $N$ .

Noen eksempler på konkrete situasjoner der Bernoulli-forsøk kan være en rimelig modell, er følgende (prøv selv å definere populasjonen i hvert tilfelle, samt angi hvorvidt forsøkssituasjonen tilsvarer med eller uten tilbakelegging):

- $n$  myntkast.
- Teste en medisin på  $n$  «tilfeldige» forsøksdyr og måle hvorvidt reaksjonen er positiv ( $J$ ) eller negativ ( $N$ ).
- $n$  lodd i pengelotteriet.
- $n$  tilfeldig utfylte Lotto-kuponger.

Vi definerer nå den stokastiske variabelen,  $X$ , som følger:

**$X =$  antall positive utfall ( $J$ ) av  $n$  Bernoulli-forsøk**

Fordelingen,  $f(x)$ , til  $X$ , vil da være det vi kaller en binomisk fordeling med parametere  $n$  og  $p$ :

**Binomisk fordeling    Bino( $n,p$ )**

$$f(x) = \binom{n}{x} p^x (1-p)^{n-x}, \quad x = 0, 1, 2, \dots, n$$

der  $x$  = antall positive utfall ( $J$ ) av  $n$  Bernoulli-forsøk og  $p = P(J)$  i hvert forsøk.

**Forventning:**  $\mu = np$

**Standardavvik:**  $\sigma = \sqrt{np(1-p)}$

3 eksempler på binomiske fordelinger er vist nedenfor, med  $n = 20$  og  $p = 0.1$ , 0.9 og 0.5. Horisontalaksen er normalisert ved å dele  $x$  ( $x = 1, 2, \dots, 20$ ) på  $n = 20$ . Da ser vi at vi har «tyngdepunktet» i fordelingen ved  $x/20 \approx p$ .

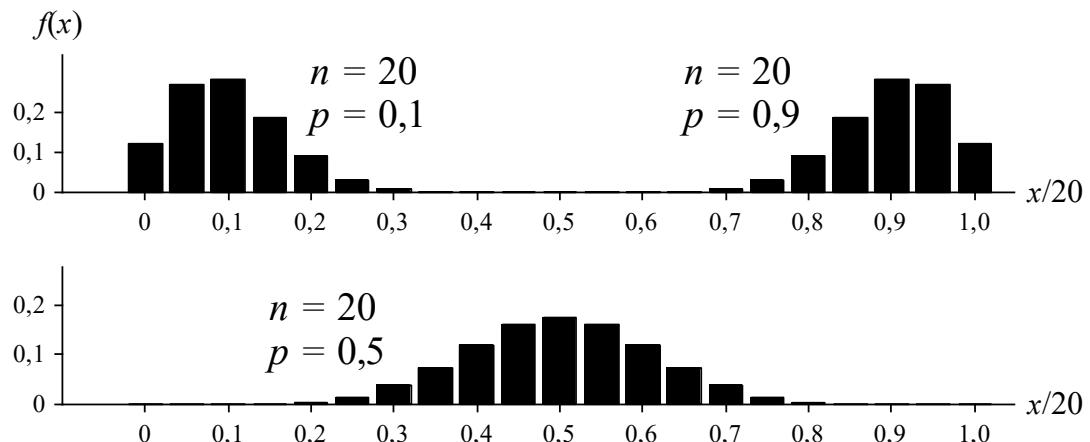


Fig. 4.1 Eksempler på binomisk fordeling med  $n = 20$ .

Noen kommentarer til fordelingene i fig. 4.1:

- For  $p = 0.1$  ser vi at fordelingen er skeiv med en «høyretung hale». Forventningen blir her  $E(X) = \mu = np = 20 \cdot 0.1 = 2$ . Dette stemmer bra med det visuelle inntrykket av fordelingen ( $x = 2$  tilsvarer  $x/20 = 2/20 = 0.1$ , som vi ser er et bra mål på tyngdepunktet i fordelingen). Variansen er her  $np(1-p) = 2 \cdot (1-0.1) = 1.8 \Rightarrow \sigma = \sqrt{1.8} \approx 1.34$ .

- For  $p = 0.9$  ser vi at fordelingen er skeiv med en «venstretung» hale, samt at fordelingen er symmetrisk i forhold til  $f(x)$  når  $p = 0.1$  med  $x/20 = 0.5$  som loddrett symmetriakse. Vi får  $E(X) = \mu = np = 20 \cdot 0.9 = 18$ . Dette stemmer bra med det visuelle inntrykket av fordelingen ( $x = 18$  tilsvarer  $x/20 = 18/20 = 0.9$ , som vi ser er et bra mål på tyngdepunktet i fordelingen). Variansen er her  $np(1-p) = 20 \cdot 0.9 \cdot (1-0.9) = 1.8 \Rightarrow \sigma = \sqrt{1.8} \approx 1.34$ , akkurat som for  $p = 0.1$ .
- For  $p = 0.5$  ser vi at fordelingen er symmetrisk om den loddrette aksen ved  $x/20 = 0.5$ . Vi ser også at fordelingen virker bredere enn for  $p = 0.1$  og  $0.9$ . La oss se om dette gir seg utslag i et høyere standardavvik:  $\text{Var}(X) = np(1-p) = 20 \cdot 0.5 \cdot (1-0.5) = 5 \Rightarrow \sigma = 2.24$ , altså betydelig større enn for de andre fordelingene. Forøvrig ser vi at også her stemmer forventningsverdien bra overens med det visuelle inntrykket av tyngdepunktet i fordelingen:  $\mu = np = 20 \cdot 0.5 = 10$ , som tilsvarer  $x/n = 10/20 = 0.5$ .

Mens vi i fig. 4.1 så på ulike binomiske fordelinger ved å variere  $p$  for en konstant  $n$ -verdi ( $n=20$ ), gir fig. 4.2 et inntrykk av hva som skjer når vi holder  $p$  konstant ( $p = 0.1$  i fig. 4.2), og øker  $n$ :

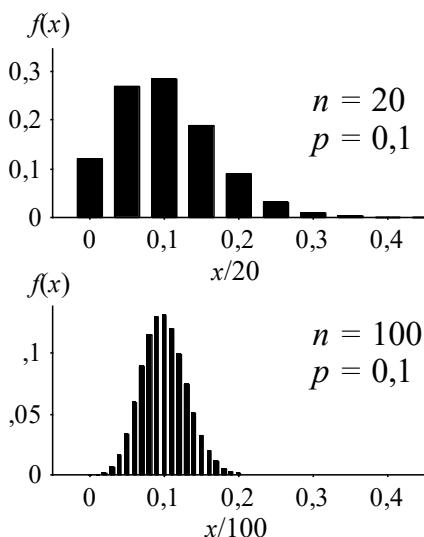


Fig. 4.2 Konstant  $p$ , økende  $n$ .

Binomisk fordeling med  $p = 0.1$ , med  $n = 20$  i øvre figur og  $n = 100$  i nedre figur. Legg merke til hvordan den øverste fordelingen er «skeiv mot høyre», mens den nederste er langt mer symmetrisk. Når  $n$  blir stor nok kan det vises at formen på den binomiske fordelingen blir svært nær den klokkeformede normalfordelingen. Dette skal vi se nærmere på i neste kapittel.

La oss nå spe på med et eksempel. Flere eksempler vil følge etter at vi har gjennomgått hvordan vi skal bruke tabeller over den binomiske fordeling.

**Eks. 4.2** **Smertestillende middel.** Et smertestillende middel har sannsynlighet  $p$  for å virke på en tilfeldig person. Vi registrerer virkningen på  $n = 3$  tilfeldig valgte personer, og lar  $J$  betegne smertestillende virkning, og

lar  $N$  betegne ikke smertestillende virkning. Videre definerer vi den stokastiske variabelen,  $X$ , som antall  $J$ -er i et tilfeldig eksperiment. Betrakter vi resultatet av hvert eksperiment som et ordnet utvalg med tilbakelegging får vi følgende  $2^3 = 8$  mulige enkeltutfall:

	$NNN$	$JNN$ $NJN$ $NNJ$	$JJN$ $JNJ$ $NJJ$	$JJJ$
$X$ -verdi:	0	1	2	3
Sannsynlighet for hver sekvens:	$p^0(1-p)^3$	$p^1(1-p)^2$	$p^2(1-p)^1$	$p^3(1-p)^0$
antall sekvenser:	$\binom{3}{0} = 1$	$\binom{3}{1} = 3$	$\binom{3}{2} = 3$	$\binom{3}{3} = 1$

La oss kommentere uttrykkene i tabellen for hver  $x$ -verdi:

**$x = 0$ :**

Dette tilsvarer at vi ikke har noen  $J$ -er, dvs. bare  $N$ -er. Vi har bare ett slikt enkeltutfall, nemlig  $NNN$ , med sannsynlighet  $P(NNN) = P(N) \cdot P(N) \cdot P(N) = (1 - P(J)) \cdot (1 - P(J)) \cdot (1 - P(J)) = (1 - p)^3 = p^0 \cdot (1 - p)^3$ . Det vil senere framgå hvorfor vi har føyd på faktoren  $p^0 = 1$  først. Sannsynlighetsfordelingen  $f(x)$  får verdien  $f(0) = \binom{3}{0} p^0 (1 - p)^3 = (1 - p)^3$

**$x = 1$ :**

Dette tilsvarer at vi har en  $J$ -verdi og dermed to  $N$ -verdier. Et eksempel er enkeltutfallet  $JNN$ , med sannsynlighet  $P(JNN) = P(J) \cdot P(N) \cdot P(N) = p^1 \cdot (1 - p)^2$ . Vi forstår at alle enkeltutfall med én  $J$  og to  $N$ -er er like sannsynlige, da faktorenes orden her er likegyldig. Når vi skal finne ut hvor mange enkeltutfall (kombinasjoner) som har én  $J$  og to  $N$ -er tilsvarer dette antall måter å plassere den ene  $J$ -en, fordi plasseringen av de to  $N$ -ene er gitt når plasseringen av  $J$ -en er bestemt. Dette tilsvarer trekning uten tilbakelegging, og ifølge kombinasjonsregelen får vi  $\binom{3}{1} = 3$  forskjellige kombinasjoner.  $f(x)$  får følgelig verdien  $f(1) = \binom{3}{1} p^1 (1 - p)^2 = 3p(1 - p)^2$

**$x = 2$ :**

Tilsvarende resonnement som for  $x = 1$  gir  $\binom{3}{2} = 3$  forskjellig måter å plassere de to  $J$ -ene, og vi får verdien  $f(2) = \binom{3}{2} p^2 \cdot (1 - p)^1 = 3p^2 \cdot (1 - p)$

**x = 3:**

Tilsvarende resonnement som for  $x = 1$  og  $2$  gir  $f(3) = \binom{3}{3} p^3 (1-p)^0 = p^3$

Vi kan nå sette opp følgende fordeling,  $f(x)$ , for eks. 4.2:

$x:$	0	1	2	3
$f(x) = P(X=x):$	$\binom{3}{0} p^0 (1-p)^3$	$\binom{3}{1} p^1 (1-p)^2$	$\binom{3}{2} p^2 (1-p)^1$	$\binom{3}{3} p^3 (1-p)^0$

Generaliserer vi eksemplet ovenfor, får vi det generelle uttrykket for den binomiske fordeling,  $f(x)$ , som ble innrammet tidligere i kapitlet:

$$f(x) = \binom{n}{x} p^x (1-p)^{n-x}, \quad x=0,1,2,\dots,n \quad \circlearrowright$$

### Bruk av binomisk tabell

Vi minner først om at en binomisk fordeling inneholder 2 parametre:  $n$  og  $p$ . For ethvert nytt valg av  $n$  og/eller  $p$ , får vi en ny fordeling. I prinsippet får vi derfor en tabell for hver kombinasjon av  $n$  og  $p$ . Vi skal her begrense omfanget av tabeller til  $n$ -verdier fra  $n = 1$  til  $n = 20$ , og 11  $p$ -verdier fra  $p = 0.05$  til  $p = 0.95$ . For større  $n$ -verdier vil det for våre formål være tilstrekkelig å bruke tilnærrelse til normalfordeling, som vi kommer til i neste kapittel.

I tabellene er det verdier for den *kumulative* fordeling,  $F(x)$ , og ikke  $f(x)$ , som er listet:

$$F(c) = P(X \leq c) = \sum_{x=0}^c f(x) = \sum_{x=0}^c \binom{n}{x} p^x (1-p)^{n-x}$$

**Eks. 4.3** Forskjell på tetthetsfunksjonen,  $f(x)$ , og kumulativ fordelingsfunksjon,  $F(x)$

$$n = 2, \quad p = 0.1 \quad \Rightarrow \quad f(x) = \binom{2}{x} \cdot 0.1^x \cdot 0.9^{2-x}, \quad x = 0,1,2$$

Beregning av $f(x)$ :	$x$	$f(x)$	$F(x)$
$f(0) = \binom{2}{0} \cdot 0.1^0 \cdot 0.9^{2-0} = 0.81$	0	0.81	0.81
$f(1) = \binom{2}{1} \cdot 0.1^1 \cdot 0.9^{2-1} = 0.18$	1	0.18	0.99
$f(2) = \binom{2}{2} \cdot 0.1^2 \cdot 0.9^{2-2} = 0.01$	2	0.01	1.00

Når vi skal finne  $F(c)$ , der  $c$  er et tall fra 0 til  $n$ , summerer vi sannsynlighetene  $f(x)$  fra og med  $f(0)$  til og med  $f(c)$ : Fra tabellen over ser vi at  $F(0) = f(0) = 0.81$ ,  $F(1) = f(0) + f(1) = 0.81 + 0.18 = 0.99$ , og  $F(2) = f(0) + f(1) + f(2) = F(1) + f(2) = 0.99 + 0.01 = 1$  (vi vil forøvrig alltid ha at  $F(n) = 1$ ).

Legg også merke til at vi kunne ha funnet  $f(x)$ -verdiene utifra  $F(x)$ -verdiene:

$$f(0) = F(0) = 0.81$$

$$f(1) = F(1) - F(0) = 0.99 - 0.81 = 0.18$$

$$f(2) = F(2) - F(1) = 1 - 0.99 = 0.01 \quad \odot$$

Før vi går løs på et tabelleksempel, skal vi liste følgende nyttige formler:

### **Nyttige formler ved bruk av binomisk tabell**

$$P(X = x) = P(X \leq x) - P(X \leq x-1)$$

$$P(a \leq X \leq b) = P(X \leq b) - P(X \leq a-1)$$

$$P(X > x) = 1 - P(X \leq x)$$

**Eks. 4.4**
**Matematikkeksamen.**

Ved en landsomfattende matematikkeksamen er det 20 % stryk.

*Oppgave*

Betrakt en bunke på 8 tilfeldig utvalgte besvarelser, og beregn sannsynligheten for følgende hendelser:

- a) Alle står.
- b) Minst halvparten stryker.
- c) 2 stryker.

*Løsningsforslag*

Vi definerer en stokastisk variabel,  $X$ , som det antall av de 8 som stryker. Dersom vi har 8 Bernoulli-forsøk, vil da  $X$  være binomisk fordelt med parametre  $n = 8$  og  $p = 0.2$ . La oss først se om vilkårene for Bernoulli-forsøk er oppfylt:

### 1) Hvert forsøk skal ha bare to utfall

Dette er opplagt tilfelle her. Hver besvarelse vil enten stryke ( $J$ ) eller stå ( $N$ ).

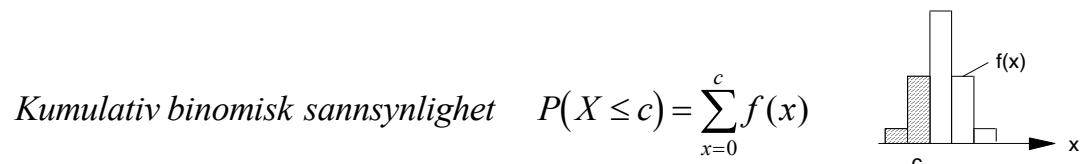
### 2) For hvert forsøk skal $P(J) = p$ være en og samme konstant

Dette er også en rimelig antakelse her, da vi forutsetter at de 8 besvarelsene utgjør et tilfeldig utvalg. En innvending her er at vi har trekning *uten* tilbakelegging, slik at  $p$  for hver trekning vil variere. Denne variasjonen vil imidlertid være neglisjerbar i dette tilfellet.

### 3) Forsøkene skal være uavhengige

For matematikkesamener bedømmer man etter en fasit, og det er rimelig å anta at bedømmelsen av en konkret besvarelse er tilnærmet uavhengig av bedømmelsen av de andre besvarelsene.

Vi går i det følgende utifra at  $X$  virkelig er binomisk fordelt med  $n = 8$  og  $p = 0.2$ , og løser oppgavene ved hjelp av følgende tabell, der de tall vi får bruk for er skyggelagt.



$c \downarrow$	$p \rightarrow$	.05	.1	.2	.3
$n = 8$	0	.663	.430	.168	.058
	1	.943	.813	.503	.255
	2	.994	.962	.797	.552
	3	1.000	.995	.944	.806

a)  $P(\text{alle står}) = P(\text{ingen stryker}) = P(X = 0) = P(X \leq 0) = \underline{0.168}$

b)  $P(\text{minst halvparten stryker}) = P(X \geq 4) = 1 - P(X \leq 3) = 1 - 0.944 = \underline{0.056}$

c)  $P(2 \text{ stryker}) = P(X = 2) = P(X \leq 2) - P(X \leq 1) = .797 - .503 = \underline{0.294} \quad \circlearrowright$

### 4.3 Hypergeometrisk fordeling

Den hypergeometriske fordeling kommer til anvendelse når vi har en situasjon som tilsvarer trekning av uordnet utvalg fra urne uten tilbakelegging. Forskjellen fra binomisk fordeling er altså at vi her har trekning *uten* tilbakelegging. Likheten er at vi for hver trekning har kun to utfall.

Et typisk eksempel der den hypergeometriske fordelingen kommer til sin rett, er stikkprøvekontroll av et vareparti, der vi for eksempel ønsker å anslå hvor stor andel av varepartiet som er defekt. Vi undersøker et antall av  $n$  forskjellige varer på tilfeldig vis, og lar  $X$  betegne antallet defekte av de  $n$  varene. Vi kan da vise at  $X$  er hypergeometrisk fordelt.

Populasjon og tilfeldig utvalg er to viktige begreper i tilknytning til hypergeometriske situasjoner. La oss belyse disse begrepene med et eksempel.

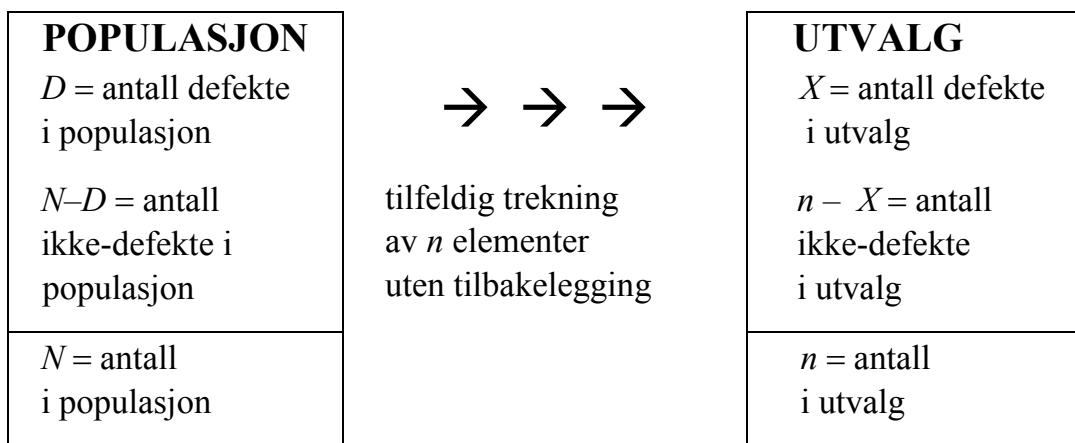
**Eks. 4.5** **Populasjon og tilfeldig utvalg.** Illustrasjon av begrepene populasjon og tilfeldig utvalg i en «hypergeometrisk» situasjon.

Et vareparti består av  $N$  enheter, og vi ønsker å anslå hvor stor andel av varepartiet som er defekt i henhold til en spesifisert måleprosedyre. La oss betegne antall defekte med  $D$ . *Populasjonen* består da av  $N$  mulige målinger, derav  $D$  med resultat «defekt» og  $N-D$  med resultat «ikke defekt».  $p = D/N$  = relativt antall defekte er videre en størrelse vi kan ønske å anslå på basis av et utvalg fra populasjonen.

Varene står stablet i like kasser, en kasse pr. enhet. Vi antar at det ikke er noen systematisk sammenheng mellom hvor en kasse er plassert og hvorvidt varen oppi kassen er defekt eller ikke. Når vi skal plukke ut et tilfeldig utvalg kan vi da gjøre det på letteste måte, og f.eks. undersøke de kassene som ligger øverst.

Dersom vi er usikre på om det er en sammenheng mellom plasseringen av en kasse og hvorvidt varen i kassen er defekt eller ikke, kan vi gjøre følgende: Kassene nummereres fra 1 til  $N$ , og vi legger  $N$  lapper merket fra 1 til  $N$  oppi en urne. Deretter trekker vi  $n$  lapper tilfeldig fra urnen, *uten tilbakelegging*, og lar de tall vi trekker bestemme hvilke kasser vi undersøker. Måleresultatene blir da et *tilfeldig utvalg* med utvalgsstørrelse  $n$  fra populasjonen med populasjonsstørrelse  $N > n$ . Kaller vi antall defekte i utvalget for  $X$ , så vil  $X/n$  være et anslag for  $D/N$ .  $X$  vil i dette tilfellet være **hypergeometrisk** fordelt. ☺

Vi kan illustrere den hypergeometriske situasjonen som følger:



La oss nå se på fordelingen til den stokastiske variabelen  $X$  = antall defekte i utvalget. Fordi vi her har trekning uten tilbakelegging, og fordi vi har uordnet utvalg (rekkefølgen har ingen betydning, da  $x$  = *antall* defekte, og i hvilken rekkefølge et gitt antall defekte blir trukket, er uvesentlig), kan vi bruke kombinasjons-regelen til å utlede den hypergeometriske fordeling. Fordelingen er gitt som følger:

<b>Hypergeometrisk fordeling</b> <b>Populasjon</b> $N$ = totalt antall i populasjon $D$ = antall defekte i populasjon	<b>hyp(<math>n,D,N</math>)</b> <b>Utvalg</b> $n$ = totalt antall i utvalg $X$ = antall defekte i utvalg
	$f(x) = P(X = x) = \frac{\binom{D}{x} \cdot \binom{N-D}{n-x}}{\binom{N}{n}}, \quad x = 0, 1, 2, \dots$
<b>Forventning:</b> $E(X) = \mu = np$ , der $p = D/N$ (antall defekte i populasjon)	
<b>Standardavvik:</b> $\text{std}(X) = \sigma = \sqrt{np(1-p) \cdot \frac{N-n}{N-1}}$	

Formelen for  $f(x)$  kan begrunnes som følger (kombinasjonsregelen): Vi har totalt  $\binom{N}{n}$  mulige måter å trekke et uordnet utvalg på  $n$  merkede enheter fra  $N$  merkede enheter. Av disse er det i alt  $\binom{D}{x}$  måter å trekke  $x$  defekte blant de totalt  $D$  defekte. For hver av disse kombinasjonene er det i alt  $\binom{N-D}{n-x}$  måter å trekke de  $n-x$  ikke defekte i utvalget fra de totalt  $N-D$  ikke defekte i populasjonen.

Legg merke til at forventningen,  $\mu = np$ , er den samme som for en binomisk fordeling, så her får ikke det faktum at vi trekker uten tilbakelegging noen betydning. Derimot ser vi at uttrykket for variansen avviker fra binomisk fordeling med en faktor  $(N-n)/(N-1)$ , som kalles «*varianskorreksjonsfaktor for endelig populasjonsstørrelse*». Vi ser også at når  $n \ll N$ , så er denne korreksjonsfaktoren nær 1, og i dette tilfellet vil den hypergeometriske fordeling være svært lik den binomiske fordeling med parametre  $n$  og  $p = D/N$ .

**Eks. 4.6** **Hypergeometrisk kuleksempel.** En urne inneholder 2 røde og 3 blå kuler. Du trekker 2 kuler tilfeldig uten tilbakelegging, og lar  $X$  betegne antall røde kuler i utvalget.

### Oppgave

- a) Bestem fordelingen til  $x$ .
- b) Finn  $P(\text{minst 1 rød kule})$ .

### Løsningsforslag

- a) Dette er en «uten + uordnet» situasjon, og  $X$  er da hypergeometrisk fordelt med følgende parametre:  $N = 5$  (antall kuler i populasjon),  $D = 2$  (antall røde kuler i populasjon) og  $n = 2$  (antall kuler i utvalg). Vi får da:

$$f(x) = P(X = x) = \frac{\binom{2}{x} \cdot \binom{5-2}{2-x}}{\binom{5}{2}} = \frac{\binom{2}{x} \cdot \binom{3}{2-x}}{\binom{5}{2}}, \quad x = 0, 1, 2$$

(NB! Husk alltid å angi definisjonsområdet.)

- b)  $P(\text{minst 1 rød kule}) = 1 - P(\text{ingen røde kuler}) = 1 - P(X = 0)$ .

$$P(X = 0) = \frac{\binom{2}{0} \cdot \binom{3}{2}}{\binom{5}{2}} = \frac{1 \cdot 3}{10} = 0.3$$

Vi får følgelig:  $P(\text{minst 1 rød kule}) = 1 - 0.3 = \underline{0.7}$  ☺

**Eks. 4.7** **Kvalitetskontroll.** En produsent godkjenner et vareparti på 1000 enheter dersom det av en stikkprøve på 10 enheter høyst er én defekt enhet.

*Oppgave*

- Beregn tilnærmet sannsynlighet for å godkjenne varepartiet hvis det inneholder 20 % defekte enheter.
- Hvor stor er defektandelen når det over lang tid med stabile forhold viser seg at 9 av 10 stikkprøver ikke inneholder noen defekte enheter?

*Løsningsforslag*

- Siden  $n = 10 \ll N = 1000$ , antar vi at det er rimelig å anvende binomisk tilnærming til den hypergeometriske situasjonen. La  $X$  betegne antall defekte i utvalget.  $X$  er da tilnærmet  $\text{Bino}(n,p)$ -fordelt med  $n = 10$  og  $p = 0.2$ . Vi får:

$$\begin{aligned} P(\text{aksepterte varepartiet}) &= P(X \leq 1) = P(X = 0) + P(X = 1) \\ &= \binom{10}{0} p^0 (1-p)^{10} + \binom{10}{1} p^1 (1-p)^9 = .8^{10} + 10 \cdot .2^1 \cdot .8^9 \approx 38 \% \end{aligned}$$

- At 9 av 10 stikkprøver ikke inneholder noen defekte kan matematisk formuleres som følger:

$$P(X = 0) = .9 \Rightarrow (1-p)^{10} = .9 \Rightarrow (1-p) = .9^{0.1} \approx 0.99 \Rightarrow p \approx 1.0 \% \quad \circlearrowright$$

## 4.4 Poisson-fordelingen

Noen typiske eksempler på stokastiske variabler som kan være Poisson-fordelte er:

- Antall svake punkt pr. meter langs en kabel.
- Antall telefoninnringninger til et sentralbord i løpet av ett minutt.
- Antall alger pr. liter i et homogent vann.
- Antall kunder som blir ekspedert pr. minutt i en bank.

Et eksempel der Poisson-fordelingen er en svært god modell, er fordelingen av antall radioaktive atomer fra et gitt radioaktivt stoff som disintegrerer i løpet av et vilkårlig langt tidsintervall. Det er da forutsatt et gitt antall  $n$  radioaktive atomer ved starttidspunktet.

Generelt er forøvrig Poisson-fordelingen en rimelig modell for mange sjeldne fenomener, der alt vi vet er gjennomsnittlig antall pr. tidsenhet eller pr. rom-

lige enhet. Poisson-fordelingen inneholder nemlig bare én parameter,  $\lambda$  ( $\ll np \gg$ ), som også er forventningsverdien,  $\mu = E(X)$ , og variansen,  $\sigma^2 = \text{Var}(X)$ :

**Poisson-fordelingen**      **Po( $\lambda$ )**

$$f(x) = P(X = x) = \frac{\lambda^x}{x!} \cdot e^{-\lambda}, \quad x = 0, 1, 2, \dots$$

**Forventning:**  $E(X) = \mu = \lambda$

**Standardavvik:**  $\text{std}(X) = s = \sqrt{\lambda}$

I kap. 4.2 så vi på den binomiske fordeling med parametre  $n$  = antall Bernoulli-forsøk og  $p = P(J)$  i hvert forsøk. Når  $n$  blir svært stor og  $p$  svært liten, vil det imidlertid i praksis by på problemer å regne ut sannsynligheter for den binomiske fordeling (tabellene blir uforholdsvis store). I dette tilfellet (stor  $n$  og liten  $p$ ) vil imidlertid den binomiske fordeling være tilnærmet lik *Poisson-fordelingen* med parameter  $\lambda = np$ , som er en enklere fordeling og hanskess med fordi den bare har en parameter,  $\lambda$  ( $\ll np \gg$ ). I tabellen bak i boka er det en liste med Poisson-tabeller som dekker  $\lambda$ -verdier fra 0,1 til 10.

En indikasjon på at Binomialfordelingen med parametre  $n$  (stor) og  $p$  (liten) er tilnærmet lik Poisson-fordelingen med parameter  $\lambda = np$ , får vi ved å sammenligne formlene for henholdsvis forventning og varians i de to fordelingene. Forventningen,  $\mu = np = \lambda$  blir lik i de to tilfellene. I en binomisk fordeling er variansen,  $\sigma^2$ , gitt ved uttrykket  $np(1-p) \approx np$  (når  $p$  er liten) =  $\lambda = \text{Var}(X)$  når  $X$  er Poisson-fordelt. Talleksempler er vist i tabellen nedenfor:

*Tabell: Sammenligning mellom binomisk sannsynlighet og Poisson-sannsynlighet.*

$np = 4$ :		Binomiske sannsynligheter:	
$n$	$p$	$X = 2$	$X = 6$
10	.4	.1209	.1115
20	.2	.1369	.1091
50	.08	.1433	.1063
100	.04	.1450	.1052
Poisson-sannsynlighet med $\lambda = 4$ :		.1465	.1042

Vi skal til slutt, før eksemplene, gjengi en formell formulering av forutsetningene for at Poisson-fordelingen skal være en rimelig modell:

### Poisson-forutsetninger

La  $J$  være en hendelse i tid eller rom som er i overensstemmelse med følgende postulater:

#### Uavhengighet

Antall ganger  $J$  forekommer i et vilkårlig tidsintervall (eller romlig intervall) er uavhengig av antall ganger  $J$  forekommer i et vilkårlig annet (disjunkt) tidsintervall (eller romlig intervall).

#### Ingen opphopning

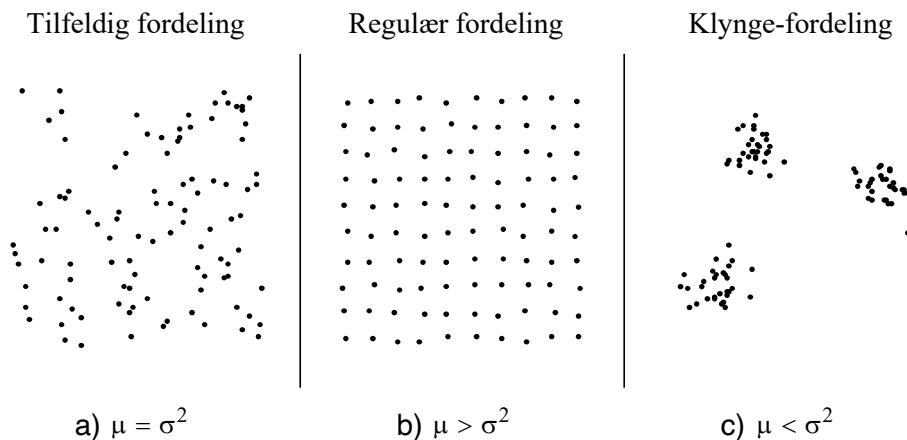
Sannsynligheten for at  $J$  kan forekomme 2 eller flere ganger samtidig (eller på samme sted) kan neglisjeres.

#### Konstant rate

Det forventede antall ganger,  $\lambda$ , som  $J$  forekommer pr. tidsintervall (eller romlig intervall) er en konstant, dvs.  $\lambda$  antas uavhengig av tid (eller rom).

$X$  = antall hendelser pr. tidsintervall (eller romlig intervall) er da **Poisson**-fordelt med fordeling som angitt i ramme på forrige side.

Et eksempel fra biologien der Poisson-fordelingen er viktig, er studier av arters romlige fordeling. Hvorvidt den romlige forekomsten av en art (antall individer pr. romlige enhet) er Poisson-fordelt kan undersøkes ved å dele det aktuelle området opp i like store romlige enheter, og telle opp antall individer innenfor hver enhet. Deretter kan man sammenligne middelverdi og varians til antall individer pr. enhet. Dersom individene er tilfeldig spredt (definert ved at Poisson-modellen er rimelig) vil forventning og varians være tilnærmet like. Dersom populasjonen opptrer i klynger (aggregert populasjon), så vil variansen være større enn forventningsverdien. Dette er illustrert i neste figur.



Figur 4.3. Romlige fordelinger av  $X =$  antall enheter (punkter) pr. arealenhet.

a)  $X$  er Poisson-fordelt (tilfeldig fordelt) med  $E(X) = \text{Var}(X)$ , b)  $X$  er regulært fordelt med  $E(X) > \text{Var}(X)$ , og c)  $X$  er klynge-fordelt med 3 klynger,  $E(X) < \text{Var}(X)$ .

#### Eks. 4.8 Bruk av Poisson-tabell.

La  $X$  være Poisson-fordelt med parameter  $\lambda = 0,90$ .

#### Oppgave

Finn følgende sannsynligheter:

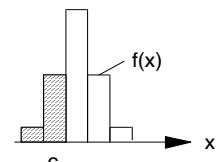
- a)  $P(X = 3)$ , b)  $P(X \geq 2)$  og c)  $P(1 < X \leq 5)$

#### Løsningsforslag

Analogt med binomisk tabell er det den kumulative fordelingen som er listet i Poisson-tabellen, dvs.  $F(c) = f(0) + f(1) + \dots + f(c)$ . Vi må derfor først omskrive de søkte sannsynlighetene til form  $P(X \leq c)$ . Vi bruker tabellen på neste side, hentet fra listen av tabeller bakerst i boka, og får (de aktuelle verdiene vi får bruk for er merket i tabellen):

Kumulativ Poisson - sannsynlighet

$$P(X \leq c) = \sum_{x=0}^c f(x)$$



$c \downarrow$	$\lambda \rightarrow$	,6	,7	,8	,9	1
0		.549	.497	.449	.407	.368
1		.878	.844	.809	.772	.736
2		.977	.966	.953	.937	.920
3		.997	.994	.991	.987	.981

Utsnitt av Poisson-tabell brukt i eks. 4.8.

- a)  $P(X = 3) = P(X \leq 3) - P(X \leq 2) = ,987 - ,937 = \underline{.050}$   
 b)  $P(X \geq 2) = 1 - P(X \leq 1) = 1 - ,772 = \underline{.228}$   
 c)  $P(2 < X \leq 5) = P(X \leq 5) - P(X \leq 2) = 1,000 - ,937 = \underline{.063} \quad \odot$

**Eks. 4.9** **Kabel.**

Antall svake punkt langs en kabel er i gjennomsnitt 6,2 pr. 100 m.

*Oppgave*

- a) Hva er sannsynligheten for å få en feilfri kabel på 10 m?  
 b) Hva er sannsynligheten for å få to feilfrie kabler på 10m hver?

*Løsningsforslag*

- a) Vi antar at antall svake punkt,  $X$ , langs kabelen er Poisson-fordelt med parameter  $\lambda = (6.2/100 \text{ m}) \cdot 10 \text{ m} = 0.62$ . Vi får:

$$P(\text{feilfri kabel}) = P(X = 0) = \frac{\lambda^0}{0!} e^{-\lambda} = e^{-0.62} \approx \underline{0.54}$$

- b) Vi antar først at vi har to «uavhengige» kabler på 10 m hver, dvs. vi antar at antall svake punkt på den ene kablen er uavhengig av antall svake punkt på den andre. La  $Y$  betegne antall kabler av de to som er feilfrie ( $Y = 0, 1$  eller  $2$ ).  $Y$  er da binomisk fordelt med parametre  $n = 2$  og  $p \approx 0.54$  (hvorfor?), og vi får:

$$P(2 \text{ feilfrie kabler}) = P(Y = 2) = \binom{2}{2} \cdot p^2 \cdot (1-p)^0 = .54^2 = \underline{0.29}$$

Anta imidlertid at vi kapper en 20 m lang kabel i to. Vi kan da beregne sannsynligheten for en feilfri kabel som er 20 m lang. Med samme Poisson-antagelse som tidligere får vi:

$$P(\text{feilfri kabel på } 20 \text{ m}) = e^{-1.24} \approx \underline{0.29} \quad \odot$$

**Eks. 4.10** **Biltrafikk.**

Vi er interessert i å undersøke biltrafikken på en vegstrekning (80 km-sone) mellom kl. 03 og kl. 04 på natta. Vegstrekningen har mye trafikk på dagtid, men lite trafikk på natta. Over lang tid har vi registrert hvor mange biler som passerer mellom kl. 03 og 04, og funnet et gjennomsnitt på 6.7 biler.

*Oppgave*

Finn sannsynligheten for følgende hendelser en tilfeldig natt mellom kl. 03 og 04:

- a) Ingen biler passerer
- b) Minst 3 biler passerer
- c) 5,6 eller 7 biler passerer

*Løsningsforslag*

Vi definerer en stokastisk variabel,  $X$ , som antall biler som passerer mellom kl. 03 og 04 en tilfeldig natt. Det er da rimelig å anta at  $X$  er tilnærmet Poisson-fordelt med parameter  $\lambda = 6.7$ . Vi omskriver de 3 hendelsene på form  $P(X \leq x)$ , bruker Poisson-tabellen med  $\lambda = 6.7$ , og får:

- a)  $P(\text{ingen biler passerer}) = P(X = 0) = P(X \leq 0) = \underline{0.001}$
- b)  $P(\text{minst 3 biler passerer}) = P(X \geq 3) = 1 - P(X \leq 2) = 1 - 0.037 = \underline{0.963}$
- c)  $P(5,6 \text{ eller } 7 \text{ biler passerer}) = P(X \leq 7) - P(X \leq 4) = .643 - .202 = \underline{.441} \quad \circlearrowright$

**Eks. 4.11** **Alger i fjord.** I en fjord er det funnet gjennomsnittlig 1,6 alger av sjeldent type pr. liter vann etter en lang rekke med undersøkelser.

*Oppgave*

Bestem sannsynligheten for følgende hendelser:

- a) Akkurat 1 alge i en tilfeldig liter vann fra fjorden.
- b) Å unngå å få algen i seg hvis man sluker 1 dl vann når man bader i fjorden.
- c) Flere enn 10 alger i en 5 liters bøtte vann fra fjorden.

*Løsningsforslag*

Vi definerer her  $X$  som antall alger pr. volumenhet (dvs. pr. liter i a), pr. desiliter i b) og pr. 5-liter i c)).  $X$  vil da være Poisson-fordelt med parameter  $\lambda = 1.6$  i a),  $\lambda = 1.6 \cdot 0.1 = 0.16$  i b) og  $\lambda = 1.6 \cdot 5 = 8$  i c). Begrunnelsen for dette kan gå som følger:

Vi tenker oss en liter sjøvann oppdelt i mange like store volumelementer. Tenker vi oss at hvert element er  $1\text{cm} \cdot 1\text{cm} \cdot 1\text{cm}$  får vi  $n = 1\ 000$  volumelementer på 1 liter. Vi lar så  $J$  betegne hendelsen at det er en alge i et tilfeldig lite volumelement.  $P(J) = p$  blir da svært liten, mens  $n$  er svært stor. Dersom vi tenker oss at vi undersøker alle de  $n$  volumelementene og forutsetter at disse tilfredsstiller kravene til  $n$  Bernoulli-forsøk, vet vi at  $X = \text{antall alger pr. liter}$  vil være tilnærmet Poisson-fordelt med forventning lik  $\lambda \approx 1.6$ .

Siden forventet antall alger er 1.6 pr. liter, må det forventede antallet være  $1.6/10 = 0.16$  alger pr. desiliter. Tilsvarende må forventet antall alger være  $1.6 \cdot 5 = 8$  alger pr. bøtte med 5 liter vann. Vi må derfor «justere»  $\lambda$ -verdien etter hvilke volummål vi ser på. Vi får:

a)  $\lambda = 1.6$ :  $P(X = 1) = P(X \leq 1) - P(X \leq 0) = .525 - .202 = \underline{.323}$

b) Siden tabellen ikke inneholder verdien  $\lambda = 0.16$ , beregner vi verdien utifra formelen for Poisson-fordelingen:

$$\lambda = 0.16: P(X = 0) = \frac{0.16^0}{0!} \cdot e^{-0.16} = \underline{0.85}$$

c)  $\lambda = 8.0$ :  $P(X > 10) = 1 - P(X \leq 10) = \underline{0.184} \quad \textcircled{\text{S}}$

## 4.5 Oppgaver

**4.1** Vi betrakter en tilfeldig valgt familie med 6 barn. Anta at sannsynligheten for «guttefødsel» er  $1/2$ , og at hvilket kjønn hvert enkelt av de 6 barna har er uavhengig av hvilket kjønn de andre barna har. Beregn følgende sannsynligheter:

- a)  $P(3$  gutter og  $3$  jenter i familien)
- b)  $P(\text{flere jenter enn gutter i familien})$

La  $X$  betegne antall gutter i familien.

- c) Beregn  $E(X)$  og  $\text{sd}(X)$ .

**4.2** I et stort lotteri gav  $5\%$  av loddene gevinst. Du kjøper  $10$  lodd. Hva er sannsynligheten for

- a)  $1$  gevinst
- b) Minst  $2$  gevinster

**4.3** En bilforretning har  $6$  nye varevogner av en bestemt type på lager.  $1/3$  av disse bilene har fabrikasjonsfeil. Du kjøper  $2$  biler og de blir tatt ut tilfeldig blant de  $6$ . La  $X$  være antall biler med fabrikasjonsfeil som du får utlevert. Sett opp fordelingen til  $x$  i tabellform og framstill den grafisk. Finn  $E(X)$  og  $\text{sd}(X)$ .

**4.4** En skole har  $30$  elever i videregående kurs II. De fordeler seg med  $20$  elever på regnskapslinjen og  $10$  elever på markedsføringslinjen. Blant disse elevene blir det gjort et tilfeldig utvalg på  $8$ . La  $X$  være antall elever i utvalget fra markedsføringslinjen.

- a) Begrunn at  $f(x) = P(X=x)$

$$= \frac{\binom{10}{x} \binom{20}{8-x}}{\binom{30}{8}}, \quad x = 0, 1, 2, \dots, 8$$

- b) Finn  $E(X)$  og  $\text{sd}(X)$
- c) Finn sannsynligheten for at  $5$  elever i utvalget kommer fra markedsføringslinjen.

**4.5** Fødselsstatistikken forteller oss at i det lange løp er en av  $60$  fødsler tvillingfødsler. Av  $500$  fødsler hva er:

- a)  $P(10$  tvillingfødsler)
- b)  $P(\text{minst } 10$  tvillingfødsler)
- c) I en liten bygd i Valdres har det i de siste årene tilsammen blitt født  $9$  gutter og en jente. Nå forventer folk i bygda seg relativt mange jenter i de nærmeste årene. Hvordan vil du som statistiker forholde deg til denne forventningen? Forklar.
- d) Hva er sannsynligheten for at det i løpet av de  $10$  neste fødslene blir født minst  $8$  jenter?

**4.6** Til en restaurant ankommer gjennomsnittlig  $10$  gjester mellom kl.  $20.00$  og  $21.00$ . Forutsett at gjestene ankommer i henhold til Poisson-fordelingen.

- a) Hva er sannsynligheten for at det ankommer akkurat  $10$  gjester i dette tidsrommet en tilfeldig kveld?
- b) Hva er sannsynligheten for at det ankommer  $12$  eller flere gjester i dette tidsrommet?

**4.7** I året 1919 utførte to engelske forskere (Greenwood og Woods) en studie over forekomsten av arbeidsulykker blant kvinner ansatt i ammunisjonsfabrikker. De tok for seg en gruppe på  $648$  kvinner og registrerte antall uhell hver kvinne var utsatt for i løpet

av en periode på 13 måneder. De fikk følgende resultat:

Antall uhell	0	1	2	3	4	5
antall kvinner	448	132	42	21	3	2

- a) Tegn stolpediagram over fordelingen og beregn gjennomsnitt og standardavvik.
- b) Tilpass en Poisson-fordeling og tegn inn den beregnede fordeling i samme diagram som den observerte. Beskriv det avviket du ser mellom de to fordelingene. Hva slags forklaring kan et slikt avvik ha?

**4.8** I en matforretning selges det gjennomsnittlig 7 litersglass med eksklusive grønne oliven pr. dag. Hvor mange slike glass bør butikken ha i hylla når den åpner om morgen for at det høyest skal være 5 % sannsynlighet for at hylla blir tom i løpet av dagen?

**4.9** Et bryggeri har funnet ut at gjennomsnittlig 0.2 % av pilsflaskene inneholder for sterke alkoholprosent. Hver dag tas det en stikkprøve på 100 flasker som kontrolleres. Dersom et visst antall  $k$  eller flere flasker viser for stor alkoholprosent stoppes og justeres prosessen der alkohol tilsettes.

- a) Hva er sannsynligheten for stans i denne prosessen hvis  $k = 1$ ?
- b) Hva må  $k$  minst være for at sannsynligheten for stans i prosessen skal være høyest 0.3 %?

**4.10** På et sentralbord kommer det inn gjennomsnittlig 2.2 innringninger pr. kvarter.

- a) Hva er sannsynligheten for at det kommer inn 10 innringninger i løpet av en time?
- b) Hva er sannsynligheten for minst 8 innringninger i løpet av en time?
- c) Hvilke forutsetninger må du gjøre for å løse a) og b)?

#### 4.11

- a) La betegnelsen  $X \sim \text{Bino}(n, p)$  bety at  $X$  er binomisk fordelt med parametre  $n$  og  $p$ . Bruk utdelte tabeller og finn de tallene som mangler nedenfor:

$$P(X \leq \cdot) = 0.724 \text{ når } X \sim \text{Bino}(12, 0.3)$$

$$P(X = \cdot) = 0.273 \text{ når } X \sim \text{Bino}(7, 0.5)$$

$$P(X = 9) = ? \text{ når } X \sim \text{Bino}(16, 0.8)$$

- b) I et lotteri er det bare 12 lodd igjen, og hele 3 av disse gir premie. Vis at sannsynligheten for å vinne er 75 % dersom du kjøper 4 lodd.
- c) La  $X$  være binomisk fordelt med parametre  $n = 300$  og  $p = 0,030$ . Finn en tilnærmet verdi for  $P(X \geq 9)$ .

**4.12** Definer  $X$  som antall seksere etter 18 kast med rettferdig terning.

- a) Begrunn hvorfor  $X$  er binomisk fordelt.
- b) Sett opp et uttrykk for fordelingen,  $f(x) = P(X = x)$ .
- c) Bestem forventning,  $\mu = E(X)$ , og standardavvik,  $\text{std}(X)$ .
- d) Bestem sannsynlighetene for
- minst én sekser
  - ingen seksere

- bare seksere
- 3 seksere
- minst 2 og høyst 4 seksere

**4.13** Ved hjelp av en Geigerteller kan en telle antall partikler som utsendes fra et radioaktivt materiale. En forsker har foretatt slike tellinger over en rekke tidsperioder, hver på 6 sekunder. Han har fått resultatene vist i neste tabell.

Antall partikler $Y$	Antall tidsperioder med $Y$ partikler
0	11
1	20
2	28
3	24
4	12
5	5
6	0
Totalt: 100	

- Lag et stolpediagram over fordelingen av antall partikler.
- Finn gjennomsnittlig antall partikler pr. tidsperiode.
- Anta at  $Y$  er Poisson-fordelt med parameter  $\mu = 2.21$ . Beregn forventet hypighet for de enkelte  $Y$ -verdier fra 0 til 6. Tegn fordelingen inn i samme diagram som du brukte under punkt a). Synes du det ser ut som om Poisson-fordelingen passer rimelig bra?

**4.14** En person (A) hevdet at han kunne smake forskjell på viner fra forskjellige produsenter innen *samme* vindistrikt. For å teste dette ble A gitt 5 smaksprøver. Han fikk *hver gang* oppgitt at prøven kom fra *en* av *to* navngitte produsenter og ble bedt om å oppgi hvilken. La  $Y$  være antall riktige svar på de 5 forsøkene.

- Anta at A bare gjetter. Forklar *kort* hvorfor punktsannsynligheten til  $Y$  da blir

$$P(Y=y) = \binom{5}{y} \cdot \left(\frac{1}{2}\right)^5, y=0,1,\dots,5$$

- Hva er sannsynligheten for at han svarer riktig samtlige ganger dersom han gjetter?
- Anta at A klarer 4 av 5 riktige. Kan vi ut fra dette være rimelig sikre på at han ikke gjetter? Begrunn svaret.

**4.15** Det er velkjent at krabber av og til mister ett eller flere av sine ben. I denne oppgaven skal vi studere et observasjonsmateriale som belyser dette fenomenet. Det er observert  $n = 1344$  tilfeldig utvalgte hankrabber. Tabellen nedenfor viser hvor mange av dem som hadde mistet henholdsvis 0,1,2,3,4, eller 5 av en på forhånd spesifisert gruppe på i alt 5 ben.

Antall ben som mangler: ↓

0	1	2	3	4	5	
1137	155	43	8	1	0	1344

Observeert frekvens: ↑ Sum ↑

- Tegn et stolpediagram av observasjonsmaterialet.

I det følgende (punktene b)-c)) skal vi undersøke om dataene støtter en hypotese om at observasjonene er binomisk fordelte med samme binomiske fordeling. La  $Y_i$  = antall ben som mangler hos  $i$ 'te krabbe. Anta at  $Y_1, \dots, Y_n$  alle har sannsynlighetsfordelingen (punktsannsynligheten)

$$P(Y_i=y) = \binom{5}{y} p^y q^{5-y}, y=0,1,2,3,4,5$$

der  $q = 1 - p$

b) Hvilken tolkning har parameteren  $p$ ?

Forklar hvorfor det ut fra dataene er rimelig å anslå  $p$  som tallet

$$\frac{155 \cdot 1 + 43 \cdot 2 + 8 \cdot 3 + 1 \cdot 4}{5 \cdot 1344}$$

Hvilken tolkning har telleren og hvilken tolkning har nevneren i denne brøken?

c) Beregn en forventet (teoretisk) frekvenstabell med  $p$  fra b). Tegn frekvensene inn på samme diagram som i a).

Det oppgis at:  $\binom{5}{0} = \binom{5}{5} = 1$ ,

$$\binom{5}{1} = \binom{5}{4} = 5, \quad \binom{5}{2} = \binom{5}{3} = 10$$

**4.16** Tabellen nedenfor angir for 200 større amerikanske byer antall dødsfall forårsaket av en bestemt sykdom i løpet av ett år.

# døde	0	1	2	3	4
# byer	93	70	26	8	3

Vi skal i denne oppgaven undersøke om en Poisson-modell passer til dataene. Det vil si at dersom  $Y$  er lik antall dødsfall i en gitt by, skal vi undersøke om  $Y$  følger en punktsannsynlighet av typen

$$P(Y=y) = \frac{\lambda^y}{y!} e^{-\lambda}, \quad y=0,1,\dots \quad (*)$$

- a) Lag et stolpediagram av dataene.  
b) Hvilken tolkning har tallet

$$\frac{0 \cdot 93 + 1 \cdot 70 + 2 \cdot 26 + 3 \cdot 8 + 4 \cdot 3}{200} ?$$

c) Estimer  $\lambda$  i punktsannsynligheten (\*) ved å sette  $\lambda$  lik tallet i b). Beregn  $P(Y=y)$  for  $y=0,1,2,3,4$  med denne verdi av  $\lambda$ .

d) Beregn det forventede (teoretiske) antall byer med 0,1,2,3 eller 4 dødsfall gitt at  $Y$  følger punktsannsynligheten (\*) med den estimerte verdi av  $\lambda$ . Lag et stolpediagram av det forventede antall på samme figur som i a).

**4.17** Fangst-gjenfangst-metoden er mye brukt for å anslå ville dyrebestanders størrelse. Den går i korthet ut på å sette ut feller flere ganger (netter). De fangete dyrene merkes og settes fri, og det er gjenfangstandelen som brukes til å anslå bestandsstørrelsen. Vi skal i denne oppgaven ikke studere selve metoden, men snarere de forutsetningene den bygger på. I den enkleste modellen antar en

- i) Et dyrs adferd blir ikke påvirket av å bli fanget.  
ii) Alle dyrene har samme sannsynlighet for å bli fanget hver gang (natt).  
a) Anta at fellene blir satt ut et stort antall netter og at sannsynligheten for at et bestemt dyr skal gå i fella en natt er liten. Forklar hvorfor en på dette grunnlag kan anta at antall ganger et bestemt dyr blir fanget,  $X$ , er (tilnærmet) Poisson-fordelt.  
b) I et lukket område hvor det var 135 (gjenkjennbare) kaniner gjennom hele fangstperioden, fikk Edwards & Ebenhardt i 1967 følgende fangstresultat:

Antall ganger fangst: ↓							
0	1	2	3	4	5	6	7
59	43	16	8	6	0	2	1

Antall kaniner: ↑                      Sum: ↑

Tilpass en Poisson-fordeling til dette materialet. Tegn den observerte fordeling og den tilpassede fordeling inn i samme stolpediagram. Er det tegn til systematiske avvik mellom observert og tilpasset modell? Hvordan vil du eventuelt karakterisere dette avviket? Diskuter kort mulige årsaker til dette avviket - hva kan være galt med antagelsene i) og ii)?

**4.18 (E)** Antallet tankbåter som kommer til en bestemt havn er Poisson-fordelt med parameter lik 2.0. Havnen kan maksimalt betjene 3 tankbåter pr. dag. De tre første fartøyene som kommer, blir betjent, mens de neste fartøyene blir omdirigert til en annen havn.

- a) Hva er sannsynligheten for at fartøyer skal bli omdirigert en bestemt dag?
- b) Hvor stor kapasitet må havnen ha om den med minst 95 % sannsynlighet skal kunne betjene samtlige tankbåter som kommer en bestemt dag?

**4.19 (E)** I en bok er det 220 trykkfeil tilfeldig spredt på bokens 200 sider. Vi antar at antall trykkfeil pr. side i boken er Poisson-fordelt. Finn sannsynligheten for at det på en tilfeldig side er:

- a) ingen trykkfeil,
- b) akkurat en trykkfeil,
- c) høyst to trykkfeil,
- d) minst to trykkfeil.

**4.20 (E)** Strålingen fra en radioaktiv kilde antas å være Poisson-fordelt. I en periode på 2 timer ble det registrert at kilden sendte ut 492  $\alpha$ -partikler.

- a) Hva er sannsynligheten for at det blir registrert nøyaktig 3 partikler i det neste minuttet?
- b) Hva er sannsynligheten for at det ikke blir registrert flere enn 3 partikler i det neste minuttet?

**4.21 (E)** Antall syklister som passerer et bestemt veikryss i løpet av en tilfeldig 2-minutters periode antas å være Poisson-fordelt med  $\lambda = 3$ .

- a) Finn sannsynligheten for at
  - i) ingen syklister passerer i løpet av 2-minutters-perioden,
  - ii) minst 3 syklister passerer i løpet av 2-minutters-perioden.
- b) Finn forventning og standardavvik til antall syklister som passerer i løpet av 2-minutters-perioden.
- c) Finn sannsynligheten for at eksakt 3 syklister passerer i løpet av en 4-minutters periode. Vi antar at hendelser i ikke-overlappende tidsintervaller er uavhengige.

**4.22 (E)** En komponent som benyttes i et elektronisk apparat har vist seg å være defekt med en sannsynlighet på 5.0 %. Vi kaller denne komponenten for «den kritiske komponenten». De øvrige komponentene i apparatet er alltid feilfrie, og ingen komponenter blir ødelagt i forbindelse med monteringen.

Bestem sannsynligheten for at firmaet skal kunne levere 75 apparater uten å bestille inn flere kritiske komponenter

dersom antall kritiske komponenter på lager er

- a) 75,
- b) 77.

Produsenten av apparatet har funnet ut at det fra en spesiell by kommer gjennomsnittlig 1.8 ordrer pr. dag.

- c) Bestem sannsynligheten for at det en dag vil komme inn eksakt 2 ordrer fra byen.

Det blir sendt varer til byen hver gang det har høpt seg opp 3 eller flere ubesørgete ordrer.

- d) En morgen er ordrelisten fra byen tom. Bestem sannsynligheten for at det vil bli sendt varer til byen før det er gått 2 dager.

Produsenten går over til å pakke appara-tene i en ny emballasje, hvor det ved hver forsendelse er en sannsynlighet på 0.10 for at det skal bli skade på emballasjen.

- e) Bestem sannsynligheten for at første skade på emballasjen vil skje i den 4. forsendelsen.
- f) Bestem sannsynligheten for at produ-senten vil oppleve emballasjeskade i løpet av de fire første forsendelsene.

**4.23** Man ønsker å undersøke utbredelsen av en bestemt sykdom i en befolkning. Til dette formål er 1000 personer trukket ut tilfeldig fra befolkningen (som totalt er mye større enn 1000). Videre har man en diagnosemetode for å avgjøre om en person er syk eller frisk. De 1000 personene blir undersøkt med metoden. La  $X$  være antall personer som blir diagnostisert til å være syke.

- a) Forklar kort hvorfor vi kan bruke binomisk fordeling for  $X$ .

Diagnosemetoden er ikke helt pålitelig: Metoden kan gi konklusjonen at en frisk person er syk og at en syk person er frisk med visse sannsynligheter. La begivenheten  $A$  bety at en tilfeldig person blir diagnostisert til å være syk. Man har ved grundige undersøkelser tidligere funnet at  $P(A | \text{personen er syk}) = 0.9$ , og at  $P(A | \text{personen er frisk}) = 0.05$ .

- b) Dersom andelen  $\pi$  av befolkningen er syke (og andelen  $1-\pi$  er friske), vis at sannsynligheten for at en tilfeldig person blir diagnostisert til å være syk er:  $p = 0.85\pi + 0.05$ .

## 4.6 Formelsamling

### Bernoulli-forsøk

Betingelser:

- 1) Hvert forsøk har 2 mulige utfall som vi betegner  $J$  og  $N$ .
- 2)  $P(J) = p =$  konstant i hvert forsøk.
- 3) Uavhengige forsøksutfall.

### Kumulativ fordeling, $F(c)$

$$F(c) = P(X \leq c) = \sum_{x=0}^c f(x)$$

### Nyttige tabellformler

$$P(X = c) = P(X \leq c) - P(X \leq c-1)$$

$$P(X > c) = 1 - P(X \leq c)$$

$$P(a \leq X \leq b) = P(X \leq b) - P(X \leq a-1)$$

### Binomisk fordeling Bino( $n,p$ )

(med tilbakelegging, uordnet rekkefølge)

$$f(x) = \binom{n}{x} p^x (1-p)^{n-x}, x = 0, 1, \dots, n$$

der  $x$  = antall  $J$ -utfall av  $n$  Bernoulli-forsøk.

$$E(X) = \mu = np$$

$$\text{std}(X) = \sigma = \sqrt{np(1-p)}$$

### Hypergeometrisk fordeling

#### **hyp( $n,D,N$ )**

(uten tilbakelegging, uordnet rekkefølge)

$$f(x) = \frac{\binom{D}{x} \binom{N-D}{n-x}}{\binom{N}{n}}, x = 0, 1, \dots$$

$N$  = antall i populasjon,

$D$  = antall «defekte» i populasjon,

$x$  = antall defekte i tilfeldig utvalg på  $n$  enheter.

$$E(X) = \mu = np = n \cdot D/N$$

$$\text{std}(X) = \sigma = \sqrt{np(1-p)} \cdot \sqrt{\frac{N-n}{N-1}}$$

### Poisson-fordelingen $\text{Po}(\lambda)$

$$f(x) = \frac{\lambda^x}{x!} \cdot e^{-\lambda}, \quad x = 0, 1, \dots$$

$$E(X) = \mu = \lambda$$

$$\text{std}(X) = \sigma = \sqrt{\lambda}$$

### Tilnærmelse

#### hypergeometrisk - binomisk

$N \gg n$  og  $D \gg n \Rightarrow$  hypergeometrisk fordeling er tilnærmet lik binomisk fordeling med parametre  $n$  og  $p = D/N$ .

### Tilnærmelse

#### binomisk - Poisson

$n \geq 10$  og  $p \leq 0.1 \Rightarrow$  binomisk fordeling med parametre  $n$  og  $p$  er tilnærmet lik Poisson-fordelingen med parameter  $\lambda = np$ .

## Kapittel 5

# Kontinuerlige fordelinger

### 5.1 Innledning

Med en kontinuerlig fordeling mener vi her sannsynlighetsfordelingen til en *stokastisk variabel* som er *kontinuerlig*. Begrepet stokastisk variabel er definert i kap. 3. Litt forenklet kan vi si at en kontinuerlig stokastisk variabel,  $X$ , er en størrelse som kan ha uendelig mange forskjellige verdier som ligger uendelig tett. Vi skal her behandle en rekke viktige fordelinger og sammenhenger mellom noen av disse. Den viktigste er **normalfordelingen** (også kalt **Gauss-fordelingen**), som behandles i kap. 5.9.

For hver fordeling skal vi innføre en forkortet betegnelse med angivelse av parametre i parentes. Eksempelvis lar vi  $N(\mu, \sigma)$  betegne en normalfordeling med forventning  $\mu$  og standardavvik  $\sigma$ . Videre minner vi om forkortelsen *uif* som betyr «uavhengige og identisk fordelte». Betegnelsen « $X_1, \dots, X_n$  er *uif*  $N(\mu, \sigma)$ » er da en forkortet skrivemåte for at « $X_1, \dots, X_n$  er uavhengige og identisk normalfordelte variabler med forventning  $\mu$  og standardavvik  $\sigma$ ».

Før vi introduserer noen av de nevnte fordelinger, skal vi i kap. 5.2 først se på en sannsynlighetsmodell for kontinuerlige variabler. Her skal vi også se på overgangen mellom diskrete og kontinuerlige variabler, som vi tidligere har påpekt er nokså «glidende», samt generelle regler for å beregne forventning og varians. Dessuten skal vi i kap. 5.3 se på begrepene **overlevelsесfunksjon** og **feilrate**, som er viktige begreper innen pålitelighetsanalyse.

Den **uniforme** fordeling er behandlet i kap. 5.4. Dette er en særdeles viktig fordeling i forbindelse med Monte Carlo-simulering og generering av «tilfeldige tall» fra gitte fordelinger. **Eksponensial**-fordelingen (kap. 5.5) er en fordeling som er realistisk i mange situasjoner fra det praktiske liv. Tre eksempler er: 1) fordeling av levetid til elektroniske komponenter, 2) fordeling av tidsintervaller mellom to påfølgende innringninger til et sentralbord, og 3) tida det tar før et radioaktivt atom disintegrerer. Eksponensialfordelingen er også sterkt koplet til Poisson-fordelingen, som vi skal se eksempel på. **Gamma**-fordelingen (kap. 5.6) er knyttet til eksponensialfordelingen ved at en sum av *uif* eksponensialfordelte variabler er gammafordelt. **Weibull**-fordelingen (kap. 5.7) er en slags generalisert eksponensial-fordeling med en ekstra (form-) parameter, og benyttes blant annet innen risiko- og pålitelighetsanalyse. Et spesialtilfelle av Weibull-fordelingen

er **Rayleigh**-fordelingen (kap. 5.8) som blant annet ofte er en realistisk fordeling for sjøbølgehøyde.

Etter en nokså grundig behandling av normalfordelingen i kap. 5.9, tar vi for oss det særdeles viktige **sentralgrenseteoremet** i kap. 5.10. Vi ser på **normaltilnærmelsen** til *binomisk* fordeling i kap. 5.11 og til *Poisson-fordelingen* i kap. 5.12. **Kjikvadrat**-fordelingen behandles deretter i kap. 5.13. Dette er en viktig fordeling, som har utstrakt anvendelse innen blant annet tester for hvor godt en modell passer til data. Den er nært knyttet til normalfordelingen ved at summen av kvadratet av  $n$  uif  $N(0,1)$ -variabler er kjikvadrat-fordelt med  $n$  frihetsgrader.  **$t$** -fordelingen behandles i kap. 5.14. Denne er knyttet til normalfordelingen ved at en standardisert sum av  $n$  uif  $N(\mu, \sigma)$ -variabler er  $t$ -fordelt med  $n-1$  frihetsgrader når vi erstatter  $\sigma$  med empirisk standardavvik,  $S$ . I kap. 5.15 behandles  **$F$** -fordelingen, som blant annet har utstrakt anvendelse innen variansanalyse. Til slutt ser vi på **binormalfordelingen** (kap. 5.16), som er en fordeling av  $(X, Y)$  der  $X$  og  $Y$  begge er normalfordelte. Kapitlet avsluttes med oppgaver i kap. 5.17 og formler i kap. 5.18.

## 5.2 Sannsynlighetsmodell

For en diskret variabel (kun heltallsverdier) var det viktig å holde rede på forskjellen mellom ulikhetssymbolene  $<$ ,  $\leq$ ,  $>$ , og  $\geq$ . Skulle vi f.eks. bruke binomisk tabell, måtte vi omskrive sannsynlighetene til  $\leq$ -uttrykk. Eks:  $P(X \geq 3) = 1 - P(X \leq 2)$ .

For en kontinuerlig fordeling faller forskjellen på  $P(X < x)$  og  $P(X \leq x)$  bort, og likeså forskjellen på  $P(X > x)$  og  $P(X \geq x)$ . Dette skyldes at  $P(X = x) = 0$  når  $X$  er kontinuerlig, fordi de mulige variabelverdiene ligger «uendelig» tett, slik at sannsynligheten for hver av dem er null. For en kontinuerlig variabel slipper vi derfor å tenke på om vi skal ha med likhetstegnet eller ikke når vi har et ulikhetstegn. Litt forsiktige må vi likevel være i de tilfeller der vi skal bruke normaltilnærmelsen til en diskret variabel, som vi skal se.

Vi skal i det følgende utdype hvorfor  $P(X = x) = 0$  når  $X$  er kontinuerlig, og vi skal se på en sannsynlighetsmodell for kontinuerlige variabler med utgangs-

punkt i beskrivende statistikk og grupperte data. La oss ta utgangspunkt i fig. 5.1.

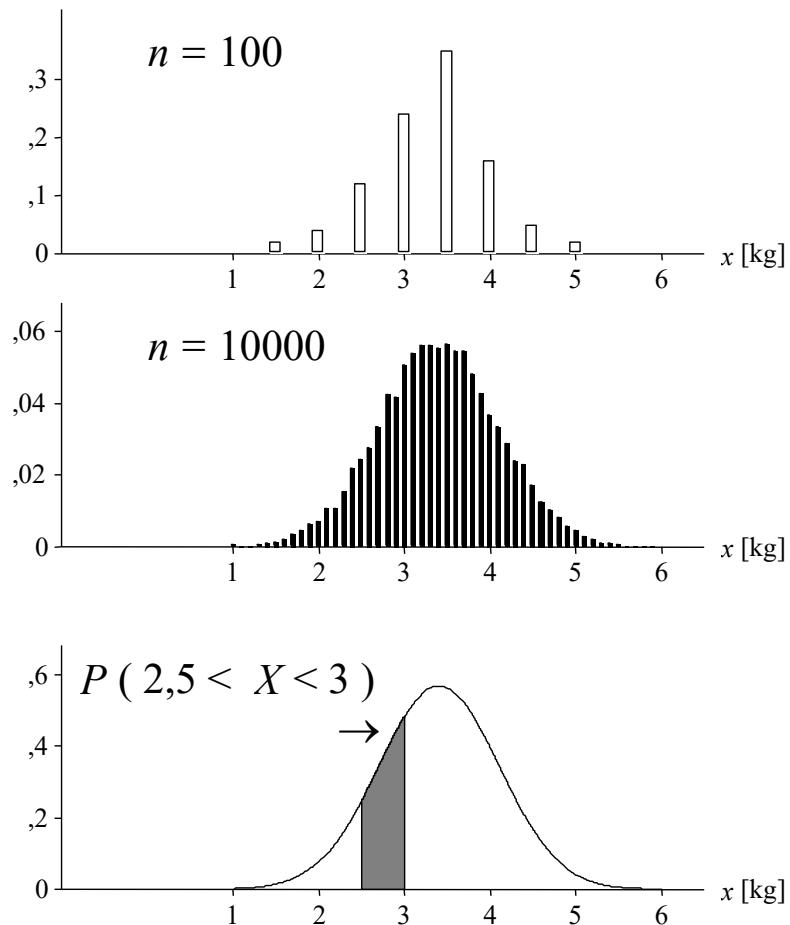


Fig. 5.1 Fra diskret til kontinuerlig fordeling.

Fødselsvektene er oppgitt på nærmeste halvkilo (øverst), nærmeste 100 gram (midten), og nærmeste gram (nederst). Allerede i midterste diagram øyner vi et nokså glatt og kontinuerlig omriss av fordelingen, og sannsynligheten (y-verdien) for hvert enkeltutfall er sterkt redusert i forhold til øverste diagram. I nederste diagram er  $X$  betraktet som en kontinuerlig variabel med kontinuerlig fordeling  $f(x)$ . Som illustrert i figuren er da sannsynlighetsbegrepet knyttet til et *intervall* i  $x$ , og ikke en bestemt  $x$ -verdi.

Med utgangspunkt i fig. 5.1 ovenfor, formulerer vi nå følgende fundamentale egenskaper til en kontinuerlig fordeling:

## Sannsynlighet, forventning og varians

En kontinuerlig (sannsynlighets-) fordeling,  $f(x)$ , er en funksjon som beskriver fordelingen av sannsynlighet til en kontinuerlig stokastisk variabel,  $X$ .  $f(x)$  kalles også **sannsynlighetstettheten** til  $X$  og har følgende generelle egenskaper:

a) Det totale arealet under kurven er 1:  $\int_{-\infty}^{\infty} f(x)dx = 1$

b) **Sannsynligheten**  $P(a \leq X \leq b) =$  areal under  $f(x)$ -kurven fra  $a$  til  $b$ :

$$(5.1) \quad P(a \leq X \leq b) = \int_a^b f(x)dx$$

c)  $f(x)$  er alltid større eller lik null

d) **Forventningen**,  $E(X)$ , er definert ved integralet

$$(5.2) \quad E(X) = \mu = \int_{-\infty}^{\infty} x \cdot f(x) dx$$

e) **Variansen**,  $\text{Var}(X) = \sigma^2$ , finnes fra en av følgende integralformler:

$$(5.3) \quad \text{Var}(X) = \sigma^2 = \int_{-\infty}^{\infty} (x - \mu)^2 \cdot f(x) dx = \int_{-\infty}^{\infty} x^2 \cdot f(x) dx - \mu^2$$

f) **Standardavviket**,  $\text{std}(X) = \sigma = \sqrt{\text{Var}(X)}$

## **P( $X = x$ ) = 0**

Dersom  $X$  er en *kontinuerlig* stokastisk variabel, så er sannsynligheten for at  $X = x$  *alltid* lik null, uansett verdi av  $x$ :

$P(X = x) = 0$  for alle mulige  $x$  når  $X$  er kontinuerlig

Til slutt skal det presiseres at de samme reglene gjelder for kontinuerlige og diskrete variabler når det gjelder forventning og varians til en sum av stokastiske variabler (se neste ramme).

### Forventning og varians til summer

La  $X$  og  $Y$  være 2 stokastiske variabler, og la  $a$  og  $b$  være to konstanter. Da gjelder følgende regler, uavhengig av om  $X$  og  $Y$  er diskrete eller kontinuerlige variabler:

$$(5.4) \quad E(aX + bY) = aE(X) + bE(Y)$$

$$(5.5) \quad \text{Var}(aX + bY) = a^2\text{Var}(X) + b^2\text{Var}(Y) + 2ab \cdot \text{Cov}(X, Y)$$

Dersom  $X$  og  $Y$  er *uavhengige* eller ukorrelerte, så er  $\text{Cov}(X, Y) = 0$ , og vi får:

$$(5.6) \quad \text{Var}(aX + bY) = a^2\text{Var}(X) + b^2\text{Var}(Y)$$

Formlene ovenfor kan enkelt generaliseres til en sum av  $n$  stokastiske variabler, slik eks. 5.1 viser.

**Eks. 5.1** **Forventning og varians til summer** (se forrige ramme).

La  $X$ ,  $Y$  og  $Z$  være 3 stokastiske variabler. Da gjelder f.eks. følgende:

- $E(X - Y - Z) = E(X) - E(Y) - E(Z)$
- $E(2X - 3Y + 5Z) = 2E(X) - 3E(Y) + 5E(Z)$
- $\text{Var}(X - Y) = \text{Var}(X) + \text{Var}(Y) - 2\text{Cov}(X, Y)$
- $\text{Var}(2X - 3Y) = 4\text{Var}(X) + 9\text{Var}(Y) - 12\text{Cov}(X, Y)$
- $E(2X^2 - 3Y^2 - 5Z^2) = 2E(X^2) - 3E(Y^2) - 5E(Z^2) \odot$

**Eks. 5.2** La  $X$  og  $Y$  være 2 kontinuerlige stokastiske variabler med følgende egenskaper:

$$E(X) = -3, \quad E(Y) = 7, \quad \text{Var}(X) = 1, \quad \text{Var}(Y) = 4, \quad \text{Cov}(X, Y) = -1$$

*Oppgave*

La  $Z = 5X - 6Y$ , og finn  $E(Z)$ ,  $\text{std}(Z)$  og  $\text{Corr}(X, Y)$ .

*Løsningsforslag*

$$E(Z) = E(5X - 6Y) = 5E(X) - 6E(Y) = 5 \cdot (-3) - 6 \cdot 7 = \underline{-57}$$

$$\text{Var}(Z) = \text{Var}(5X - 6Y) = 5^2\text{Var}(X) + 6^2\text{Var}(Y) - 2 \cdot 5 \cdot 6 \cdot \text{Cov}(X, Y)$$

$$= 25 \cdot 1 + 36 \cdot 4 - 60 \cdot (-1) = 229 \Rightarrow \text{std}(Z) = \sqrt{229} = \underline{15.13}$$

$$\text{Corr}(X, Y) = \text{Cov}(X, Y) / (\sigma_X \cdot \sigma_Y) = (-1) / (\sqrt{4} \cdot \sqrt{1}) = \underline{-0.5} \quad \circlearrowright$$

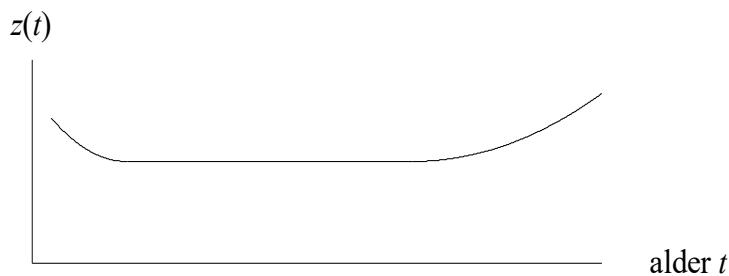
### 5.3 Overlevelsesfunksjon og feilrate

Ofte er tiden,  $t$ , den frie (uavhengige) variabelen til en fordeling. La  $T$  betegne levetiden til en enhet, f.eks. varigheten til en lyspære. Dersom vi er interessert i å studere hvor lenge lyspæra vil være, så kan størrelsen

$$(5.7) \quad R(t) = P(T > t) = 1 - P(T \leq t) = 1 - F(t)$$

være en mer naturlig størrelse enn  $F(t)$ . Av naturlige grunner kalles  $R(t) = 1 - F(t)$  **overlevelsesfunksjonen**. Har vi et visst antall nye komponenter med overlevelsesfunksjon  $R(t)$ , så vil  $R(t)$  angi forventet andel av komponentene som har overlevd ved tidspunkt  $t$  ( $t = 0$  for ny komponent).

En annen størrelse som er mye brukt innen pålitelighetsteori er **feilraten** (sviktsannsynligheten),  $z(t)$ , som er et mål på hvor hyppig det forekommer feil (svikt). Generelt vil  $z(t)$  for en komponent variere med alderen ( $t$ ), og vil typisk kunne ha badekarformen vist i neste figur.



*Fig. 5.2 Badekarform på feilraten. Avtagende forløp: Innkjøringsfase, barne-sykdommer. Stigende forløp: slitasje, alderdomstegn.*

Matematisk er feilraten  $z(t)$  definert som følger:

$$z(t) = \frac{f(t)}{R(t)}$$

Vi skal nå begrunne hvorfor den matematiske definisjonen ovenfor er et mål på hvor hyppig det forekommer feil. Vi betrakter et lite tidsintervall  $(t, t+dt)$  og antar at enheten ikke har sviktet ved tidspunkt  $t$  (dvs.  $T > t$ ). Videre betrakter vi sannsynligheten for at enheten skal svikte i tidsintervallet:

$$P(\text{enhet svikter i intervallet } (t, t+dt) \text{ gitt at enheten ikke har sviktet før } t) =$$

$$P(T \leq t+dt | T > t) = P(t \leq T \leq t+dt)/P(T > t) = [F(t+dt) - F(t)]/R(t).$$

Ut fra definisjonen av den deriverte har vi at

$$\lim_{dt \rightarrow 0} \frac{F(t+dt) - F(t)}{dt} = F'(t) = f(t),$$

slik at  $F(t+dt) - F(t) \rightarrow f(t) \cdot dt$  når  $dt \rightarrow 0$ , og vi får:

$$P(T \leq t+dt | T > t) \approx f(t)/R(t) dt = z(t) dt$$

for små verdier av  $dt$ . Med andre ord:  $z(t)$  uttrykker enhetens tilbøyelighet til å svikte ved tidspunkt  $t$ .

## 5.4 Uniform fordeling

I kap. 2 behandlet vi en uniform sannsynlighetsmodell. Et typisk eksempel der denne modellen kom til sin rett, var terningkast med rettferdig terning, der hvert av de 6 utfallene av et tilfeldig terningkast var like sannsynlige. Lar vi  $X$  betegne antall øyne ved et slikt tilfeldig terningkast, ville  $f(x) = 1/6$ ,  $x = 1, \dots, 6$ , være sannsynlighetsfordelingen til  $X$ , og dette er et eksempel på en **diskret uniform fordeling**.

Her skal vi se på en **kontinuerlig uniform** fordeling, som vi skal betegne med  $U[a,b]$ :  $U$  står for Uniform, og  $[a,b]$  angir definisjonsområdet, dvs. en variabel  $X$  som er  $U[a,b]$ -fordelt kan ha hvilke som helst verdier fra og med  $a$  til og med  $b$ . Analogt med det diskrete tilfellet blir også her  $f(x)$  lik med en konstant:  $f(x) = 1/(b-a)$ , dvs. én delt på intervallbredden til definisjonsområdet til  $x$ .

**Uniform fordeling**

Tetthetsfunksjon:

$$U[a,b]$$

$$f(x) = \frac{1}{b-a}, \quad a \leq x \leq b$$

Kumulativ fordelingsfunksjon:

$$F(x) = P(X \leq x) = \frac{x-a}{b-a}$$

Forventning:

$$\text{E}(X) = (a+b)/2$$

Standardavvik:

$$\text{std}(X) = \frac{b-a}{\sqrt{12}}$$

Når  $a = 0$  og  $b = 1$  får vi den særlig enkle fordelingen  $f(x) = 1$ ,  $0 \leq x \leq 1$ . Denne  $U[0,1]$ -fordelingen er svært sentral når det gjelder Monte Carlo-simulering (se kap. 11). Sammenhengen mellom en  $U[0,1]$ -variabel og en mer generell  $U[a,b]$ -variabel er forøvrig meget enkel, som illustrert i eks. 5.4.

De fleste programmeringsspråk, og selv enkle lommekalkulatorer, har innebygget såkalte slumptallgeneratorer, dvs. de kan frembringe tilfeldige og uavhengige  $U[0,1]$ -variabler. Ofte har generatorene kodenavn som f.eks. «rnd» eller «random» («random» er engelsk og betyr tilfeldig). En slumptallgenerator som genererer tilfeldige tall mellom 0 og 1 kan blant annet benyttes til å generere

- tilfeldige variabelverdier fra en hvilken som helst fordeling,
- tilfeldige heltall mellom 1 og 34 (til bruk i Lotto).

Sistnevnte anvendelse er illustrert i eks. 5.5. La oss først se på et eksempel på en situasjon fra virkeligheten der en uniform fordeling er realistisk.

**Eks. 5.3**

**Trafikklys.** Et trafikklys lyser rødt i 10 sekunder. En tilfeldig bilist uten biler foran seg kommer på rødt lys.

*Oppgave*

Hva blir fordelingen av ventetida  $T$  for bilisten?

*Løsningsforslag*

Vi antar at hvor lenge det har lyst rødt idet bilisten ankommer trafikklyset er uavhengig av bilisten. Det er da rimelig å anta at ventetida  $T$  er uniformt fordelt  $U[0,10]$ . Fordelingen (sannsynlighetstetthetsfunksjonen) blir da  $f(t) = 1/10$ ,  $0 \leq t \leq 10$ , der  $t$  er i sekunder. Fordelingen er her kontinuerlig, fordi ventetida er en kontinuerlig variabel.

Vi har antatt at hvor lenge det har lyst rødt idet bilisten ankommer trafikklyset er uavhengig av bilisten. Prøv å forklare hvorfor dette ikke nødvendigvis er en riktig antagelse. ☺

### Eks. 5.4 Overgang fra $U[0,1]$ - til $U[a,b]$ -fordeling.

*Oppgave*

La  $Z$  være en  $U[0,1]$ -variabel og vis at da vil  $X = a + (b-a)Z$  være en  $U[a,b]$ -variabel, der  $b > a$ .

*Løsningsforslag*

Å vise at  $X$  er en  $U[a,b]$ -variabel er ekvivalent med å vise at  $X$  har kumulativ fordelingsfunksjon  $F(x) = P(X \leq x) = (x-a)/(b-a)$ ,  $a \leq x \leq b$ . Vi får:

$$\begin{aligned} F_X(x) &= P(X \leq x) = P(a + (b-a)Z \leq x) = P(Z \leq (x-a)/(b-a)) = \\ F_Z(z = (x-a)/(b-a)) &= (x-a)/(b-a); \quad 0 \leq (x-a)/(b-a) \leq 1 \end{aligned}$$

fordi  $Z$  er en  $U[0,1]$ -variabel. Vi har dermed vist at  $F(x) = (x-a)/(b-a)$  som vi skulle. Det gjenstår å vise at dette gjelder for  $a \leq x \leq b$ . Dette følger av den doble ulikheten  $0 \leq (x-a)/(b-a) \leq 1$ . Vi deler denne i 2 ulikheter:

$$\begin{aligned} 0 \leq (x-a)/(b-a) &\Rightarrow x \geq a \\ (x-a)/(b-a) \leq 1 &\Rightarrow x - a \leq b - a \Rightarrow x \leq b \end{aligned}$$

Sammenholder vi resultatet av de to ulikhettene ovenfor får vi at  $a \leq x \leq b$ . ☺

### Eks. 5.5 Tilfeldige heltall

*Oppgave*

Forklar hvordan du kan bruke tilfeldige (uniformt fordelte) tall mellom 0 og 1 til å generere tilfeldige heltall mellom 1 og 34.

*Løsningsforslag*

Først multipliseres det tilfeldige tallet mellom 0 og 1 med 34. Deretter adderer vi 0.5, forhøyer og glemmer tallene etter komma. Eks: Anta at slumptallet er 0.631787. Vi ganger med 34 og får 21.48. Deretter legger vi til 0.5 og får 21.98, forhøyer og får 22: Tallet 22 er da et tilfeldig tall mellom 1 og 34 og er like «sannsynlig» som ethvert annet heltall mellom 1 og 34. ☺

## 5.5 Eksponensial-fordelingen

Eksponensial-fordelingen er en fordeling som er høyst realistisk i mange sammenhenger. Som eksempler kan nevnes: Fordeling av tidsintervaller mellom påfølgende innringninger til et sentralbord, samt levetiden til et radioaktivt atom. Fordelingen er også ofte brukt som levetidsfordeling i pålitelighetsanalyse. Fordelingen har konstant feilrate, og kalles derfor for en **hukommelsesfri** fordeling (se eksempel lengre ned). Eksponensial-fordelingen er dessuten interessant fordi den har en enkel teoretisk form med kun én parameter.

Videre er eksponensial-fordelingen sterkt koplet til Poisson-fordelingen, f.eks. som følger: Når fordelingen av tidsintervaller mellom påfølgende innringninger til et sentralbord er eksponensialfordelte, så vil antall innringninger i løpet av en bestemt tidsperiode, f.eks. 10 minutter, være Poisson-fordelt.

Merk at eksponensial-fordelingen ofte er knyttet til tiden som stokastisk variabel. Selv om vi har benyttet symbolet  $x$  for den frie variablene i neste ramme, vil vi ofte benytte symbolet  $t$  (og  $T$ ) i stedet for  $x$  (og  $X$ ) for den frie variablene.

### Eksponensial-fordelingen $\text{expo}(b)$

$$\text{Tetthetsfunksjon: } f(x) = \frac{1}{b} \cdot e^{-x/b}, \quad x \geq 0$$

$$\text{Kumulativ fordelingsfunksjon: } F(x) = P(X \leq x) = 1 - e^{-x/b}$$

$$\text{Forventning: } E(X) = b$$

$$\text{Standardavvik: } \text{std}(X) = b$$

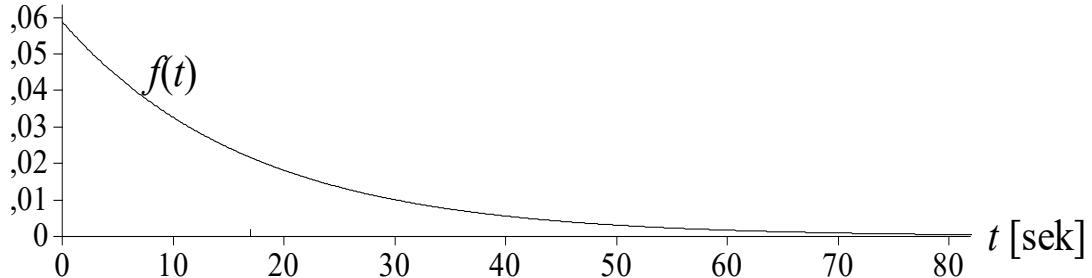
**Eks. 5.6** **Sentralbord.** På et sentralbord har en funnet at tidsintervallet mellom to påfølgende innringninger er eksponensialfordelt. Gjenomsnittlig går det etter lang tids registrering 17 sek. mellom hver innringning.

#### Oppgave

- Tegn opp fordelingen,  $f(t)$ , der  $t$  er tidsintervallet mellom to påfølgende oppringninger.
- Finn sannsynligheten for at det går mer enn ett minutt fra en tilfeldig oppringning til neste.

*Løsningsforslag*

- a) Se figur,  $f(t) = (1/17) \cdot e^{-t/17}$   
 b)  $P(T > 60) = 1 - P(T < 60) = 1 - (1 - e^{-60/17}) \approx 0.029$  ☺



**Eks. 5.7** **Sentralbord.** Tidsintervallet mellom påfølgende oppringninger til et sentralbord følger en eksponensialfordeling med parameter  $b = 15$  sek.

*Oppgave*

- a) Skriv opp fordelingen til antall oppringninger pr. minutt.  
 b) Finn sannsynligheten for at det kommer minst 10 oppringninger i løpet av et tilfeldig minutt.

*Løsningsforslag*

- a) Antall oppringninger  $Y$  i løpet av et minutt er Poissonfordelt med parameter  $\lambda = 60/15 = 4$ :  $\lambda$  = forventet antall oppringninger iløpet av 60 sekunder = 60 dividert med forventet tid i sek. mellom to påfølgende oppringninger. Følgelig:

$$f(y) = P(Y = y) = \frac{4^y}{y!} e^{-4}, \quad y = 0, 1, 2, \dots$$

- b)  $P(Y \geq 10) = 1 - P(Y \leq 9) = 1 - 0.992 = 0.8\%$  (tabell) ☺

**Eks. 5.8** **Eksponensial-fordelingen er hukommelsesfri**

*Oppgave*

Vis at feilraten  $z(t) = f(t)/R(t)$  er konstant når  $T$  er eksponensialfordelt. Forklar deretter hva som ligger i begrepet *hukommelsesfri fordeling*.

*Løsningsforslag*

$$F(t) = 1 - e^{-t/b} \Rightarrow R(t) = 1 - F(t) = e^{-t/b} \Rightarrow z(t) = \frac{\frac{1}{b} e^{-t/b}}{e^{-t/b}} = 1/b = \text{konstant.}$$

At feilraten er konstant betyr at enhetens alder ikke innvirker på den. Enheten «husker ikke» hva den har gjennomlevd. Hyppigheten av feil er uavhengig av alder. ☺

## 5.6 Gammafordelingen

**Gammafordelingen** er en anvendelig to-parameter fordeling som blant annet blir brukt innen pålitelighetsanalyse. Den får her betegnelse  $\text{gamma}(b,c)$  der  $b$  betegner skala- og  $c$  betegner formparameteren. Den er også blitt brukt i biologi som modell for biomassen av ulike byttedyrarter, som spises av et tilfeldig individ av en bestemt art (predator), innen gitte betingelser som område og tid. Det kan vises at summen av  $n$  uif  $\text{expo}(b)$ -variabler er  $\text{gamma}(b,n)$ -fordelt.

### Gammafordelingen

**gamma**( $b,c$ );  $b, c > 0$

Tetthetsfunksjon:

$$f(x) = \frac{(x/b)^{c-1} e^{-x/b}}{b\Gamma(c)}, \quad x \geq 0$$

Forventning:

$$\text{E}(X) = bc$$

Standardavvik:

$$\text{std}(X) = b\sqrt{c}$$

der  $\Gamma$  betegner gammafunksjonen:  $\Gamma(c) \equiv (c-1)! = \int_0^{\infty} u^{c-1} e^{-u} du$

**NB!** Legg merke til den nære sammenhengen mellom gammafunksjonen, definert nederst i forrige ramme, og fakultetfunksjonen, !. Når  $c$  er et heltall større enn 1, stemmer fakultetbegrepet med det vi tidligere har definert:  $\Gamma(3) = 2! = 2$ ,  $\Gamma(4) = 3! = 6$  osv. Når argumentet ikke er et heltall, defineres fakultetfunksjonen ved gammaintegralet som vist i ramma. Denne utvidete definisjonen av fakultetfunksjonen er lagt inn på enkelte lommekalkulatorer som en egen tast, hvilket gjør det lettvint å beregne verdier for gammafunksjonen.

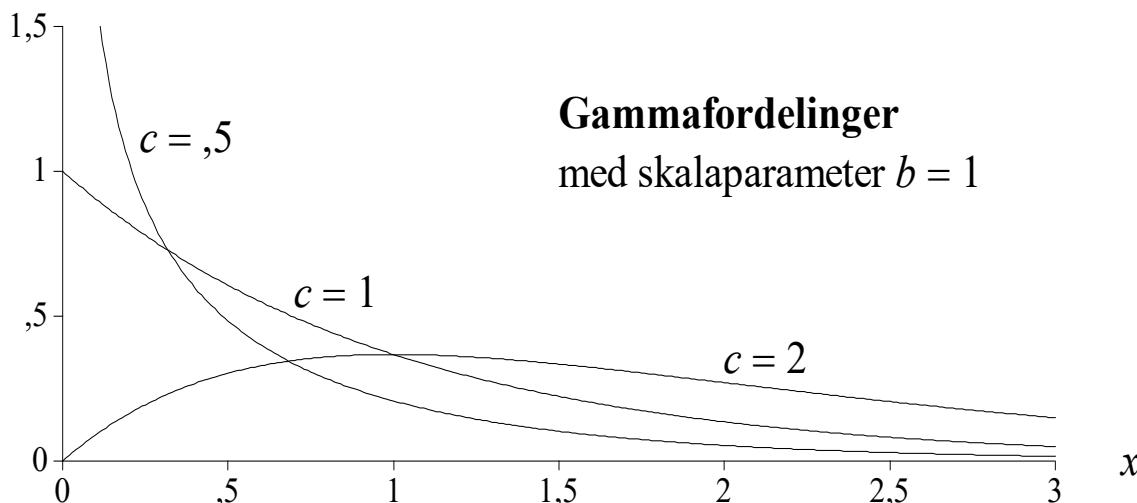


Fig. 5.3 Gammafordelingen  $f(x) = x^{c-1} e^{-x}/\Gamma(c)$  for ulike verdier av formparametren  $c$  med skalaparameter  $b = 1$ .  $c = 1$  tilsvarer eksponensial-fordelingen  $\text{expo}(b)$ .

**Eks. 5.9** **Lyspærers levetid.** En familie har 100 60 Watts lyspærer på lager. Når en pære går, skiftes den straks ut med neste. Anta at pærenes levetider  $T_1, \dots, T_{100}$  er uif og følger en eksponensialfordeling med parameter  $b = 1000$  h (h er forkortelse for «hour» på engelsk, time på norsk).

#### Oppgave

- Hva er forventet samlet levetid til de 100 pærene angitt i år?
- Hva er standardavviket til samlet levetid for de 100 lyspærene?

#### Løsningsforslag

- $E(T_1 + \dots + T_{100}) = E(T_1) + \dots + E(T_{100}) = b + \dots + b = 100b = 10^5$  h =  $10^5/(24 \cdot 365)$  år = 11,4 år
- $X = T_1 + \dots + T_{100}$  er  $\text{gamma}(b, 100) = \text{gamma}(0.114 \text{ år}, 100)$   
 $\Rightarrow \text{std}(X) = 0.114 \cdot \sqrt{100}$  år = 1.14 år ☺

## 5.7 Weibull-fordelingen

**Weibull-fordelingen** er blant annet brukt som levetidsfordeling for komponenter innen risiko- og pålitelighetsanalyse. Vi skal benytte betegnelsen  $\text{Weib}(b, c)$  som betegnelse for Weibull-fordelingen, der  $b$  er skalaparameter og  $c$  er formparameter. Både eksponensial-fordelingen (kap. 5.5) og Rayleigh-fordelingen (kap. 5.8) er spesialtilfeller av Weibull-fordelingen, henholdsvis  $\text{Weib}(b, 1)$  og  $\text{Weib}(b, 2)$ . For  $0 < c < 1$  er fordelingen monoton avtagende og

går mot uendelig når  $x$  går mot null. For  $c > 1$  skjærer Weibull-fordelingen  $y$ -aksen ved  $y = 0$ . Når  $c$  går mot uendelig blir Weibull-fordelingen mer og mer lik normalfordelingen.

### Weibull-fordelingen

**Weib**( $b,c$ );  $b, c > 0$

Tetthetsfunksjon:

$$f(x) = (c/b) \cdot (x/b)^{c-1} e^{-(x/b)^c}, \quad x \geq 0$$

Kumulativ fordelingsfunksjon:  $F(x) = P(X \leq x) = 1 - e^{-(x/b)^c}$

Forventning:  $E(X) = (b/c) \cdot \Gamma(1/c)$

Standardavvik:  $\text{std}(X) = (b/c) \sqrt{2c\Gamma(2/c) - (\Gamma(1/c))^2}$

der  $\Gamma$  betegner gammafunksjonen (se ramme for gammafordeling).

Skalaparameteren  $b$  i Weibull-fordelingen kalles også for **karakteristisk levetid**. Figuren nedenfor illustrerer hvordan formen på Weibull-fordelingen bestemmes av formparameteren  $c$ . Forandring av skalaparameteren er analogt til en strekking eller krymping av kurven langs  $x$ -aksen.

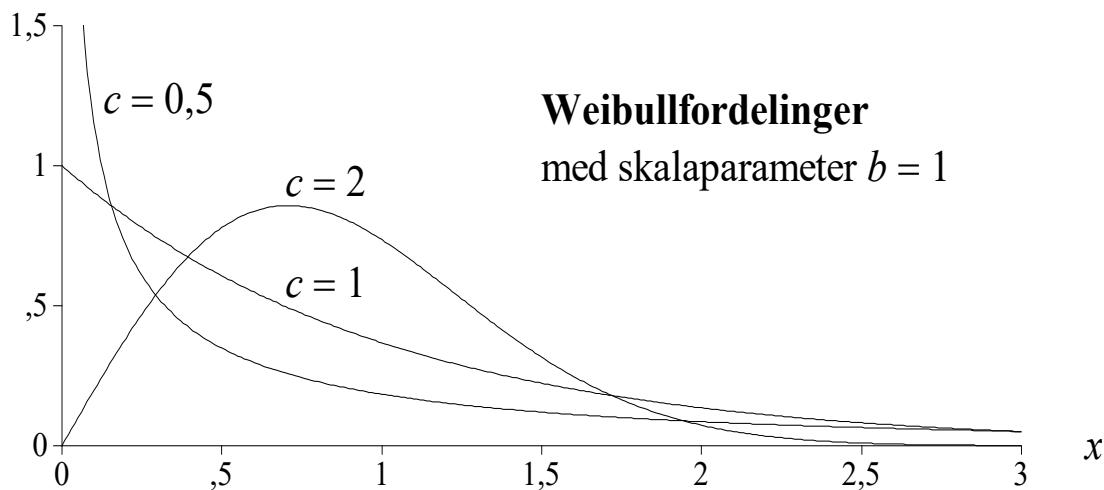


Fig. 5.4 Weibull-fordelingen  $f(x) = cx^{c-1} \exp(-x^c)$  for ulike verdier av formparametren  $c$  når skalaparameteren er  $b = 1$ .  $c = 1$  tilsvarer eksponensial-fordelingen  $\text{expo}(b)$ ,  $c = 2$  tilsvarer Rayleigh-fordelingen  $\text{Rayl}(b)$ .

**Eks. 5.10** **Levetidsfordeling.** Levetiden  $T$  til en elektronisk komponent er Weib( $b = 2$  yr,  $c = 0.5$ )-fordelt.

*Oppgave*

- Hva er sannsynligheten for at en vilkårlig komponent funksjonerer etter 5 år («yr» er engelsk forkortelse for «year»)?
- Hva er sannsynligheten for at minst én av 10 uavhengige og nye komponenter, som ble kjøpt inn samtidig, funksjonerer etter 5 år?

*Løsningsforslag*

- $p = P(T > 5) = 1 - P(T < 5) = 1 - \left(1 - e^{-(5/2)^5}\right) = e^{-\sqrt{2.5}} = 0.206$
- La  $Y$  betegne antall komponenter av de 10 som funksjonerer etter 5 år.  $Y$  er da binomisk fordelt (hvorfor?) med parametere  $n = 10$  og  $p = 0.206$ . Vi får:  
 $P(\text{minst én av 10 funksjonerer}) = P(Y \geq 1) = 1 - P(Y \leq 0) = 1 - P(Y = 0) = 1 - \binom{10}{0} \cdot p^0 \cdot (1-p)^{10-0} = 1 - 0.794^{10} = 1 - 0.100 = 90.0\% \quad \text{☺}$

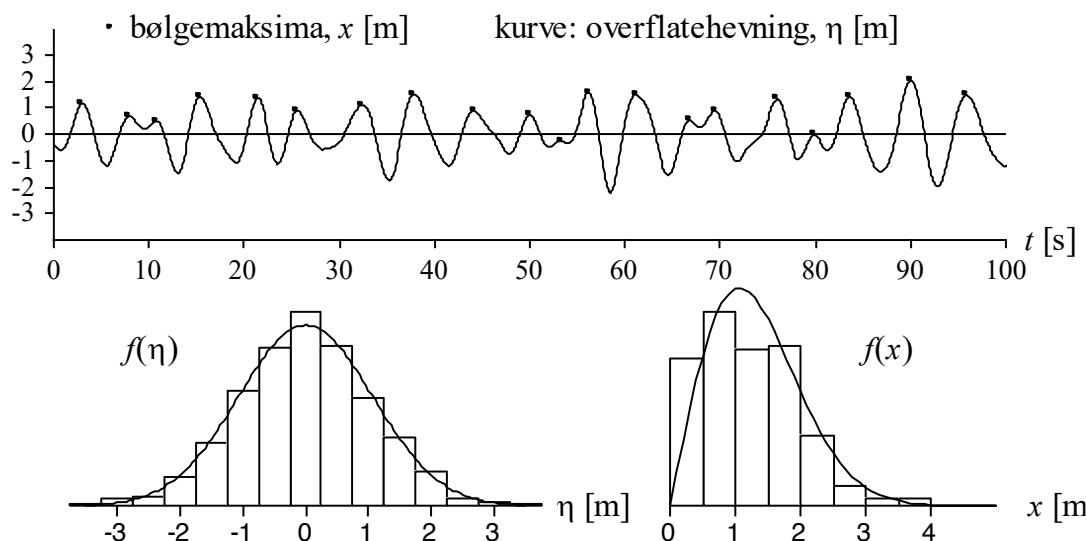
## 5.8 Rayleigh-fordelingen

**Rayleigh-fordelingen** er et spesialtilfelle av Weibull-fordelingen og benyttes ofte i forbindelse med maksima til sjøbølgenes overflatehevning og amplituden av bølgekrefter som virker på maritime konstruksjoner.

<b>Rayleigh-fordelingen</b>	<b>Rayl(<math>b</math>); <math>b &gt; 0</math></b>
Tetthetsfunksjon:	$f(x) = 2b^{-2} \cdot x \cdot e^{-(x/b)^2}, \quad x \geq 0$
Kumulativ fordelingsfunksjon:	$F(x) = P(X \leq x) = 1 - e^{-(x/b)^2}$
Forventning:	$E(X) = b\sqrt{\pi}/2$
Standardavvik:	$\text{std}(X) = \frac{b}{2}\sqrt{4 - \pi}$

**Eks. 5.11** **Fordeling av bølgemaksima.** Neste figur viser et utsnitt fra et bølgeførlop generert fra et såkalt smalbåndet JONSWAP («Nordsjø») bølgespektrum. Figuren definerer også den kontinuerlige overflatehevningen som en funksjon  $\eta(t)$  som vi kan tenke oss målt med f.eks. en bøye, og med jevne tidsintervaller. Et bølgemaksimum,  $X$ , er definert som et

lokalt maksimum på  $\eta(t)$ -kurven. Som vi ser av figuren, viser normalfordelingen en god tilpasning til  $\eta$ -dataene, og Rayleigh-fordelingen viser en god tilpasning til  $x$ -dataene. Det kan vises at dersom  $\eta$  er  $N(0, \sigma)$ -fordelt, der  $\sigma$  er standardavviket til  $\eta$ , så er  $X$   $\text{Rayl}(\sqrt{2}\sigma)$ -fordelt.



Figur 5.5 Sjøbølgers overflatehevning og maksima. Forklaring: Se tekst lengre ned.

Forklaring til fig. 5.5: Øverste diagram viser et utsnitt av en simulert tidsserie av hevning av sjøoverflata,  $\eta$ , basert på et typisk smalbåndet Nordsjø-bølgespektrum. Prikkene viser bølgemaksima. Tilpasset normalfordelingskurve,  $f(\eta)$ , og Rayleigh-fordelingskurve,  $f(x)$ , til henholdsvis hevning og bølgemaksimum ser ut til å stemme ganske bra.

### Oppgave

- Definer bølgehøyden (slik den observeres av bølgehøydeobservatører),  $H$ , til å være høydeforskjellen mellom et bølgeminimum og påfølgende bølgemaksimum,  $X$ , og anta at  $H = 2X$ . Hva blir fordelingen til  $H$  når standardavviket til overflatehevningen,  $\eta$ , er  $\sigma = 1$  m?
- Signifikant bølgehøyde er matematisk definert som  $H_s = 4\sigma$ . Hvor stor andel av bølgehøydene er større enn signifikant bølgehøyde?

### Løsningsforslag

- Siden  $X$  er Rayleigh-fordelt, må også  $H = 2X$  være Rayleigh-fordelt, siden faktoren 2 bare fører til en skalering uten å forandre formen på fordelingen. Følgelig: Siden  $X$  er  $\text{Rayl}(\sqrt{2}\sigma)$ -fordelt så er  $H$   $\text{Rayl}(2\sqrt{2}\sigma)$ -fordelt.

$$\text{b) } P(H > 4\sigma) = 1 - P(H < 4\sigma) = e^{-(4\sigma/(2\sqrt{2}\sigma))^2} = e^{-2} \approx 13.5\% \quad \odot$$

## 5.9 Normalfordelingen

Normalfordelingen, med sin karakteristiske klokkeform (Gausskurve), er den mest fundamentale statistiske fordelingen. En hovedgrunn til dette er sentralgrenseteoremet (se kap. 5.10), som sier at en sum av  $n$  uavhengige stokastiske variabler vil være tilnærmet normalfordelt, uavhengig av populasjonsfordelingen til de variabler vi summerer, sålenge  $n$  er tilstrekkelig stor. Dette danner et hovedgrunnlag for blant annet parameterestimering ved konfidensintervall (kap. 6), såvel som diverse hypotesetester (kap. 7 og 8).

Svært ofte vil også én stokastisk variabel (uten å ta en sum av stokastiske variabler) være normalfordelt. Noen eksempler på variabler som er tilnærmet normalfordelte er:

- tilfeldig studentvekt,
- tilfeldig studenthøyde,
- målefeil med diverse måleinstrumenter,
- diverse målbare prosess- og produktvariabler.

På grunn av normalfordelingens sentrale rolle er det blant annet utviklet

- en rekke tester for å teste om variabler virkelig er normalfordelte,
- gode tilnærmelsesformler til kumulativ normalfordelingsfunksjon  $F$  og invers kumulativ normalfordelingsfunksjon  $F^{-1}$  (se kap. 11),
- effektive algoritmer for å generere tilfeldige normalfordelte variabler (kap. 11).

Vedrørende tester skal vi ikke gå inn på spesielle normalitetstester her, men nøyne oss med en generell kjikvadratføyningstest i avsnitt 7.9, som også kan anvendes som normalitetstest i diverse sammenhenger. Denne benytter kjikvadrat-fordelingen som vi behandler i kap. 5.13.

La oss nå uttrykke normalfordelingen matematisk:

**Normalfordelingen**  $N(\mu, \sigma)$ 

Tetthetsfunksjon:  $f(x) = \frac{e^{-\frac{(x-\mu)^2}{2\sigma^2}}}{\sqrt{2\pi} \cdot \sigma}, \quad -\infty < x < \infty$

Forventning:  $E(X) = \mu$

Standardavvik:  $\text{std}(X) = \sigma$

Merk fra forrige ramme at vi ikke har oppgitt noe uttrykk for kumulativ fordelingsfunksjon. Det skyldes at den ikke har noe eksplisitt uttrykk, men må bestemmes ved numeriske metoder. Tabeller for kumulative  $N(0,1)$ -variabler er gitt bakerst i boka, og vi skal snart forklare bruken av slik tabell.

Noen egenskaper ved normalfordelingen er skissert i de neste figurene. Som vi ser av fig. 5.6, så er sannsynligheten ca. 68 % for at  $X$  skal ligge mindre enn ett standardavvik,  $\sigma$ , fra forventningen,  $\mu$ . Tilsvarende er sannsynligheten ca. 95 % for at  $X$  skal ligge mindre enn 2 standardavvik fra forventningen.

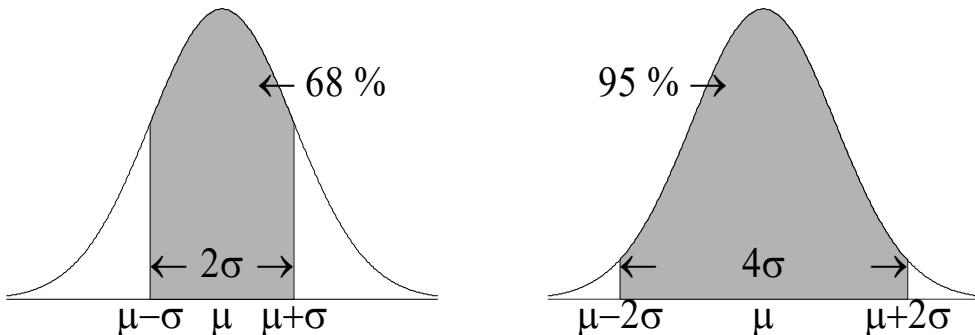


Fig. 5.6 I normalfordelingen ligger ca. 68 % av sannsynlighetsmassen innenfor  $\mu \pm \sigma$ , mens ca. 95 % ligger innenfor  $\mu \pm 2\sigma$ .

I fig. 5.7 er illustrert hvilken innflytelse standardavviket,  $\sigma$ , har på utseendet til normalfordelingen, når vi holder forventningen,  $\mu$ , konstant. Som vi ser øker bredden (spredningen) av fordelingen med økende verdi for  $\sigma$ , samtidig som maksimumshøyden til  $f(x)$  avtar.

I Fig. 5.8 er illustrert hvilken innflytelse forventningen,  $\mu$ , har på utseendet til normalfordelingen når standardavviket,  $\sigma$ , holdes konstant. Som vi ser er eneste

forskjell en forskyvning mot høyre (venstre) når forventningen øker (avtar). På figuren er  $\mu_1 < \mu_2$ .

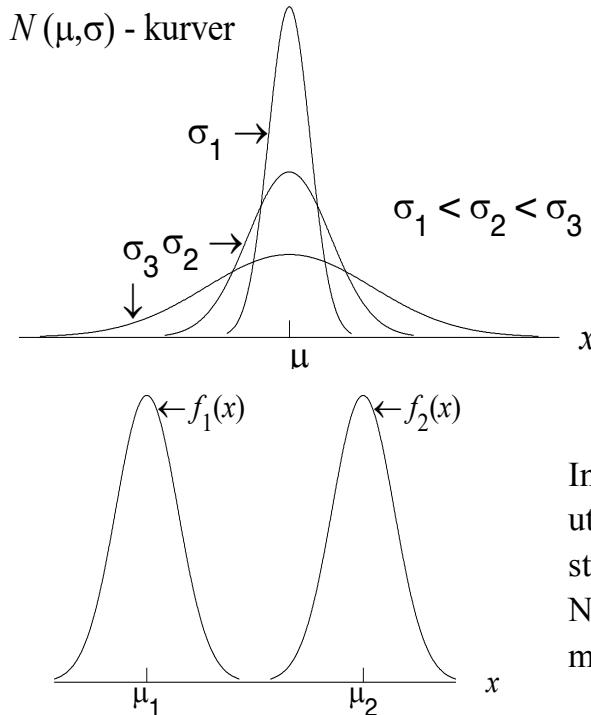


Fig. 5.7

Innflytelsen av standardavviket,  $\sigma$ , på utseendet til normalfordelingen: Når  $\sigma$  øker blir fordelingen bredere og har mindre maksimumsverdi.

Fig. 5.8

Innflytelsen av forventningen,  $\mu$ , på utseendet til normalfordelingen når standardavviket,  $\sigma$ , holdes konstant: Normalfordelingskurven beveger seg mot høyre når  $\mu$  øker.

Før vi går løs på bruk av normalfordelingstabell, skal vi introdusere følgende viktige betegnelser:

### Viktige betegnelser

Dersom en stokastisk variabel,  $X$ , er normalfordelt med forventning,  $\mu$ , og standardavvik,  $\sigma$ , så kan vi bruke betegnelsen

$$X \sim N(\mu, \sigma)$$

som leses: « $X$  er normalfordelt med forventning, my, og standardavvik, sigma», eller kortere: « $X$  er normalfordelt my sigma». Vi kan også bruke formuleringen « $X$  er en  $N(\mu, \sigma)$ -variabel». Dersom  $X$  er **standardisert**, dvs.  $\mu = 0$  og  $\sigma = 1$ , skal vi fortrinnsvis bruke symbolet  $Z$  (standardisert normalfordelt variabel):

$$Z \sim N(0, 1)$$

som ofte leses: « $Z$  er en  $N$  null én variabel».

## Bruk av tabell

Som for binomisk tabell og Poisson-tabell, er det den *kumulative* fordelingen,  $P(X \leq x)$ , som er listet i tabellen. Siden normalfordelingen er en kontinuerlig fordeling, er den kumulative fordelingen gitt ved et integral istedet for en sum:

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

En vanskelighet som kommer inn når det gjelder normalfordelingen, er at tabellen kun gjelder for en standardisert normalfordelt variabel,  $Z$ , med forventning  $\mu = 0$  og standardavvik  $\sigma = 1$ . Dette problemet skal vi komme tilbake til litt senere.

Som for de diskrete fordelingene må vi omskrive sannsynlighetene til form  $P(X \leq a)$  før vi kan bruke tabellene. Dersom  $X$  er normalfordelt og vi f.eks. skal finne sannsynligheten  $P(a < X < b)$ , må vi omskrive denne til form  $P(X < b) - P(X < a)$ , der vi har droppet å tenke på likhetstegnet. Dette er illustrert i neste figur.

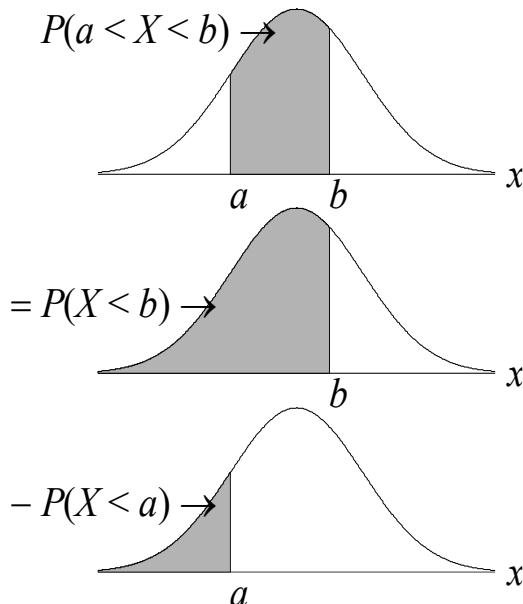


Fig. 5.9

Illustrasjon av overgang fra uttrykket  $P(a < X < b)$  til form  $P(X < b) - P(X < a)$ . Vi ser at det øverste arealet,  $P(a < X < b)$ , er lik differansen mellom det midterste arealet,  $P(X < b)$ , og det nederste arealet,  $P(X < a)$ .

La oss nå gå over til noen eksempler som viser tabellbruk. Vi starter med et eksempel på en standardisert variabel,  $Z$ . Etterpå skal vi se hvordan vi kan bruke tabellen i det generelle tilfellet at  $X$  er  $N(\mu, \sigma)$ . Vi skal innføre en spesiell betegnelse for den kumulative fordelingen,  $P(Z \leq z)$ , som inngår i normalfordelingstabellene:

### Kumulativ $N(0,1)$ -fordeling, $\Phi(z)$

I *normalfordelingstabellene* er det den *kumulative* fordelingen for en standardisert variabel,  $Z$ , som er listet, og vi skal bruke symbolet  $\Phi$  for denne kumulative fordelingen:

$$\Phi(z) = P(Z \leq z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^z e^{-\frac{x^2}{2}} dx$$

#### Eks. 5.12 Bruk av normalfordelingstabell for $N(0,1)$ -variabel.

##### Oppgave

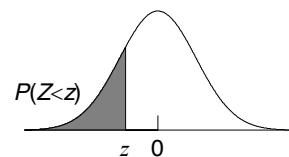
La  $Z \sim N(0,1)$ , og bestem følgende sannsynligheter:

- a)  $P(Z < -1.81)$
- b)  $P(-2.00 < Z < 0.00)$
- c)  $P(Z > 1.00)$
- d) Den  $z$ -verdi som er slik at  $P(Z < z) = 0.025$

*Løsningsforslag* (se tabellutsnitt lengre ned)

- a) Vi finner at  $P(Z < -1.81) = .0351$  direkte fra tabellen, slik illustrert nedenfor:  $z$ -verdien til og med 1. desimal står langs venstre kant nedover i tabellen, mens 2. desimal i  $z$ -verdien står i øverste rekke. Selve sannsynligheten finnes i tabellen ved «krysspeiling».

Kumulativ  $N(0,1)$ -tabell:  $\Phi(z) = \int_{-\infty}^z \frac{e^{-x^2/2}}{\sqrt{2\pi}} dx$



$z$	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
-2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
-1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
-1.8	.0359	<b>.0351</b>	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
-1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
-1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455

- b)  $P(-2.00 < Z < 0.00) = P(Z < 0.00) - P(Z < -2.00) = 0.5 - 0.0228 = .4772$   
 c)  $P(Z > 1.00) = 1 - P(Z < 1.00) = 1 - .8413 = .1587$

- d) Vi skal finne den  $z$ -verdi som er slik at  $P(Z < z) = 0.025$ . Da må vi lete i tabellen til vi finner en sannsynlighet i nærheten av 0.025, og se hvilken  $z$ -verdi det tilsvarer (bruk av tabellen «motsatt» veg). Vi finner da at  $z \approx -1.96$  ☺

### Bruk av tabell når $(\mu, \sigma) \neq (0, 1)$

Når  $X \sim N(\mu, \sigma)$  med forventning  $\mu \neq 0$  og/eller standardavvik  $\sigma \neq 1$ , kan vi ikke bruke tabellen direkte. Vi må først omskrive den sannsynligheten vi søker til «ekvivalent  $N(0,1)$ -form», slik at vi kan bruke tabellen. For å forstå denne omskrivningen kan det være nyttig å gå gjennom følgende ekvivalenser mellom følgende hendelser:

$$\begin{aligned} & a < X < b \\ & \text{(trekker fra } \mu) \\ \Leftrightarrow & a - \mu < X - \mu < b - \mu \\ & \text{(deler på } \sigma) \\ \Leftrightarrow & \frac{a - \mu}{\sigma} < \frac{X - \mu}{\sigma} < \frac{b - \mu}{\sigma} \\ & \text{(erstatter } (X - \mu)/\sigma \text{ med } Z) \\ \Leftrightarrow & \frac{a - \mu}{\sigma} < Z < \frac{b - \mu}{\sigma} \end{aligned}$$

Når to hendelser er ekvivalente, betyr det at sannsynligheten for den ene hendelsen er lik sannsynligheten for den andre hendelsen. Sammenligner vi første og siste hendelse ovenfor, får vi da:

$$P(a < X < b) = P\left(\frac{a - \mu}{\sigma} < Z < \frac{b - \mu}{\sigma}\right)$$

Den siste sannsynligheten ovenfor kan vi omskrive slik at vi kan bruke tabellen for en standardisert normalfordelt variabel,  $Z$ :

$$\begin{aligned} P\left(\frac{a - \mu}{\sigma} < Z < \frac{b - \mu}{\sigma}\right) &= P\left(Z < \frac{b - \mu}{\sigma}\right) - P\left(Z < \frac{a - \mu}{\sigma}\right) \\ &= \Phi\left(\frac{b - \mu}{\sigma}\right) - \Phi\left(\frac{a - \mu}{\sigma}\right) \end{aligned}$$

Siden  $a$ ,  $b$ ,  $\mu$  og  $\sigma$  er forutsatt å være kjente størrelser, kan vi beregne  $z$ -verdiene  $(b-\mu)/\sigma$  og  $(a-\mu)/\sigma$ . Deretter kan vi gå inn i tabellene og finne de søkte sannsynlighetene for de beregnede  $z$ -verdiene:

### **Bruk av $N(\mu,\sigma)$ -tabell når $(\mu,\sigma) \neq (0,1)$**

Gitt at  $X$  er  $N(\mu,\sigma)$ . Sannsynligheten  $P(a < X < b)$  finnes da som følger:

$$P(a < X < b) = \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)$$

Dersom  $(a-\mu)/\sigma$  er et svært negativt tall, la oss si mindre enn  $-3,5$ , så kan vi erstatte  $\Phi((a-\mu)/\sigma)$  med 0. Dersom  $(b-\mu)/\sigma$  er et stort tall, la oss si større enn  $3,5$ , så kan vi erstatte  $\Phi((b-\mu)/\sigma)$  med 1.

**Eks. 5.13** **Bruk av  $N(0,1)$  tabell når  $X$  er  $N(\mu,\sigma)$ .** La  $X$  være normalfordelt  $N(-1,1, 2.3)$ , dvs.  $\mu = -1.1$  og  $\sigma = 2.3$ .

*Oppgave*

Bestem  $P(0 < X < 5)$ .

*Løsningsforslag*

Her er  $a = 0$  og  $b = 5$ , og vi får:

$$\begin{aligned} P(0 < X < 5) &= \Phi\left(\frac{5 - (-1,1)}{2,3}\right) - \Phi\left(\frac{0 - (-1,1)}{2,3}\right) \\ &= \Phi(2.65) - \Phi(0.48) = .9960 - .6844 = \underline{.3116} \end{aligned}$$

der de 2 siste verdiene er funnet fra  $N(0,1)$ -tabell. ☺

### **Sum av normalfordelte variabler**

La  $X$  være  $N(\mu_1, \sigma_1)$  og la  $Y$  være  $N(\mu_2, \sigma_2)$ . Vi forutsetter videre at  $X$  og  $Y$  er stokastisk uavhengige variabler. Da gjelder følgende:

$$Z = X + Y \text{ er } N\left(\mu_1 + \mu_2, \sqrt{\sigma_1^2 + \sigma_2^2}\right)$$

Resultatet ovenfor kan enkelt generaliseres til en sum av  $n$  uavhengige og normalfordelte variabler:

### Sum av normalfordelte variabler

La  $X_1, X_2, \dots, X_n$  være  $n$  uavhengige normalfordelte variabler med forventninger henholdsvis  $\mu_1, \mu_2, \dots, \mu_n$  og varianser henholdsvis  $\sigma_1^2, \sigma_2^2, \dots, \sigma_n^2$ . Da vil enhver lineærkombinasjon av  $X_1, X_2, \dots, X_n$  også være normalfordelt:

$$Y = \sum_{i=1}^n a_i X_i \text{ er } N(\mu_Y, \sigma_Y) \text{- fordelt}$$

der forventning,  $E(Y) = \mu_Y$ , og standardavvik,  $\text{std}(Y) = \sigma_Y$ , er gitt ved uttrykkene:

$$\mu_Y = \sum_{i=1}^n a_i \mu_i, \quad \sigma_Y = \sqrt{\sum_{i=1}^n a_i^2 \sigma_i^2}$$

dvs.  $Y$  er normalfordelt med forventning,  $E(Y) = \sum a_i \mu_i$ , og varians,  $\text{Var}(Y) = \sum a_i^2 \sigma_i^2$ .

#### Eks. 5.14 Sum av normalfordelte, uavhengige variabler.

Vi ser på variablene  $X_1 \sim N(0,2)$ ,  $X_2 \sim N(-3,1)$  og  $X_3 \sim N(1,1)$ . Anta at  $X_1, X_2$  og  $X_3$  er uavhengige.

##### Oppgave

Definer en ny variabel  $Y = X_1 - 3X_2 - X_3$ , og bestem fordelingen til  $Y$ .

##### Løsningsforslag

$Y$  er normalfordelt med forventning,  $\mu_Y$ , og standardavvik,  $\sigma_Y$ , som følger:

$$\begin{aligned}\mu_Y &= E(X_1 - 3X_2 - X_3) = \mu_1 - 3\mu_2 - \mu_3 = 0 - 3 \cdot (-3) - 1 = 8 \\ \text{std}(Y) &= \sqrt{\text{Var}(X_1 - 3X_2 - X_3)} = \sqrt{\sigma_1^2 + 3^2 \sigma_2^2 + (-1)^2 \sigma_3^2} \\ &= \sqrt{2^2 + 9 \cdot 1^2 + 1^2} = \underline{3,74} \quad \odot\end{aligned}$$

## 5.10 Sentralgrenseteoremet

Sentralgrenseteoremet er et svært sterkt resultat som bidrar til å forklare den sentrale plass normalfordelingen har i statistikk. Før vi går løs på teoremet, skal vi innføre et par nyttige begreper det er svært viktig å skille fra hverandre, nemlig **populasjonsfordeling** og **observatorfordeling**:

### Populasjonsfordeling (definisjon)

La  $X$  være én stokastisk variabel som trekkes tilfeldig fra en populasjon. Fordelingen,  $f(x)$ , til  $X$ , kalles da populasjons-fordelingen til  $X$ . Vi skal bruke symbolet  $\mu$  for forventningen og symbolet  $\sigma$  for standardavviket i populasjonsfordelingen.

### Observatorfordeling (definisjon)

La  $X_1, X_2, \dots, X_n$  være  $n$  stokastisk uavhengige variabler, der hver av  $X$ -ene har én og samme populasjonsforventning,  $\mu$ , og samme populasjons-standardavvik,  $\sigma$ . La videre  $Y = f(X_1, X_2, \dots, X_n)$  være en vilkårlig funksjon av  $X$ -ene.  $Y$  blir da selv en stokastisk variabel, og vi skal kalle denne for en **observator** (eks:  $Y = \bar{X}$ ). Videre skal vi kalle fordelingen til  $Y$ ,  $f_Y(y)$ , for **observatorfordelingen** til  $Y$ . Observatorfordelingen vil generelt være forskjellig fra populasjonsfordelingen til  $X$ -ene.

**Eks. 5.15** **Oppdrettsanlegg.** I en mær i et oppdrettsanlegg med  $N$  fisker ønsker en å anslå fordeling av enkeltfiskenes vekt samt fiskenes gjennomsnittsvekt.

#### Oppgave

Definer populasjon, populasjonsfordeling, observator og observatorfordeling.

#### Løsningsforslag

Populasjonen er her mengden bestående av alle fiskevektene, som vi kan betegne  $x_1, x_2, \dots, x_N$ . Hadde vi kjent alle disse, gruppert dem i vektklasser og fremstilt dem i et relativ frekvens-histogram, ville vi ha fremstilt populasjonsfordelingen til enkeltfiskenes vekt. Dette er det samme som fordelingen til vekta  $X$  av én tilfeldig valgt fisk. Når vi skal anslå gjennomsnittsvekta pr. fisk, kan vi for eksempel ta gjennomsnittsvekta  $Y = \bar{X}$  av et tilfeldig utvalg på  $n \ll N$  fisker.  $Y$  blir da en observator. Fordelingen til  $Y$  blir en observatorfordeling som avhenger av utvalgets størrelse  $n$ , og som er vesentlig forskjellig fra populasjonsfordelingen til enkeltvektene.

Sentralgrenseteoremet kan formuleres på mange måter, og vi skal her gjengi en enkel versjon:

## Sentralgrenseteoremet

La  $\bar{X}$  betegne middelverdien til  $n$  stokastisk uavhengige variabler,  $X_1, \dots, X_n$ , alle med forventning,  $\mu$ , og standardavvik,  $\sigma$ . Som en tommelfingerregel vil da, når  $n$  øker, fordelingen til  $\bar{X}$  konvergere mot (nærme seg) normalfordelingen med forventning,  $\mu$ , og standardavvik,  $\sigma/\sqrt{n}$ :

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i \quad \Rightarrow \quad \bar{X} \text{ er tilnærmet } N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)\text{-fordelt}$$

Sentralgrenseteoremet (se forrige ramme) er et *sterkt* resultat. Uansett hvilken fordeling hver av  $X$ -ene har, så vil altså middelverdien til  $X$ -ene være en observator med en fordeling svært lik normalfordelingen bare  $n$  er stor nok. I praksis vil dette ofte gjelde selv for små verdier av  $n$ , særlig hvis populasjonsfordelingen(e) er symmetrisk(e). Vi skal som en forsiktig tommelfingerregel bruke grensa  $n = 20$ .

**Eks. 5.16 Middelverdifordeling.** La  $X_1, \dots, X_{100}$  være uavhengige med  $E(X_i) = 2$  og  $\text{std}(X_i) = 1$ ,  $i = 1, 2, \dots, 100$ .

*Oppgave*

Bestem tilnærmet fordeling til  $Y = \bar{X}$

*Løsningsforslag*

$X$ -ene har forventning  $\mu = 2$  og standardavvik  $\sigma = 1$ .  $\bar{X}$  vil da være tilnærmet normalfordelt med forventning,  $E\bar{X} = 2$ , og  $\text{std}(\bar{X}) = 1/\sqrt{100} = 0.1$ :

$\bar{X}$  er tilnærmet  $N(2, 0.1)$ . ☺

**Eks. 5.17 Fordeling av sum øyne i terningkast**

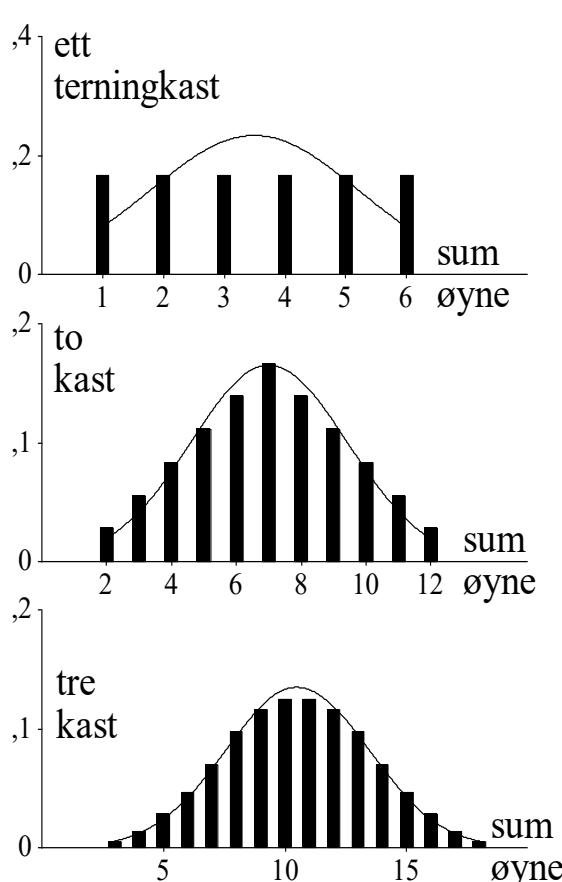
*Oppgave*

Bestem fordeling av  $Y = \text{sum øyne ved ett, to og tre terningkast}$ .

*Løsningsforslag*

Y er her henholdsvis  $X_1$ ,  $X_1 + X_2$  og  $X_1 + X_2 + X_3$ , der  $X_i$  er antall øyne i  $i$ 'te kast. Alle  $X$ -ene har her samme populasjonsfordeling, som i dette tilfellet har den enkle formen  $f(x_i) = 1/6$ ,  $x_i = 1, 2, 3, 4, 5$  eller 6. Forventning  $\mu = E(X_i)$ , og standardavvik,  $\sigma = \text{std}(X_i)$ , er henholdsvis  $\mu = 3.5$  og  $\sigma = \sqrt{35/12}$ .

Strengt tatt skulle vi ha tatt middelverdien av antall øyne etter  $n$  kast for å illustrere sentralgrenseteoremet slik vi har formulert det. I praksis hadde eneste forskjell vært at vi da langs horisontalaksen hadde hatt sum øyne delt på  $n$ , istedet for sum øyne. Forøvrig ville de eksakte og tilpassede fordelingene sett akkurat liketan ut (hadt samme form).



**Ett kast.** I dette tilfelle er sum øyne lik antall øyne i det ene kastet. Den tilpassede normalfordelingen passer som vi ser heller dårlig til omrisset av den eksakte fordelingen.

**To kast.** Omrisset av den sanne fordelingen av sum øyne blir nå en trekant. Den tilpassede normalfordelingen passer adskillig bedre til omrisset av den eksakte fordelingen enn for  $n = 1$ .

**Tre kast.** Allerede etter 3 kast, begynner sum øyne å få en fordeling med klokkeformet omriss, som likner «mistenkelig» på den tilpassede normalfordelingen.



Fig. 5.10 Illustrasjon av sentralgrenseteoremet.

## 5.11 Normaltilnærmelse til binomisk fordeling

La  $X$  være binomisk fordelt  $\text{Bino}(n,p)$ . Når størrelsen  $np(1-p)$  er et tilstrekkelig stort tall (la oss si større enn 5), så vil den binomiske fordelingen være tilnærmet lik normalfordelingen  $N(\mu, \sigma)$  med  $\mu = np$  og  $\sigma^2 = np(1-p)$ :

### Fordelingstilnærmelse binomisk – normal

Dersom  $X$  er binomisk fordelt med parametre  $n$  og  $p$  der  $np(1-p) \geq 5$ , så vil  $X$  med rimelig sikkerhet være tilnærmet normalfordelt  $N(\mu, \sigma)$  med forventning  $\mu = np$  og varians  $\sigma^2 = np(1-p)$ :

$$f(x) = \binom{n}{x} p^x (1-p)^{n-x} \approx (2\pi\sigma)^{-1/2} \cdot e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad x \geq 0$$

$$\mu = np, \quad \sigma = \sqrt{np(1-p)}$$

Når vi skal bruke normaltilnærmelsen til binomisk fordeling i praksis, må vi passe på å skrive om de sannsynligheter vi er ute etter på form  $P(X \leq x)$ , som for den binomiske fordeling. Siden  $X$  i utgangspunktet er en diskret variabel, er det *viktig* å holde rede på forskjellen på likhets- og ulikhetstegnene, før vi bruker tilnærningsformelen for normalfordelingen. Følgende nyttige regel gjelder for beregning av sannsynligheter ved hjelp av  $N(0,1)$ -tabell:

### Bruk av $N(0,1)$ -tabell

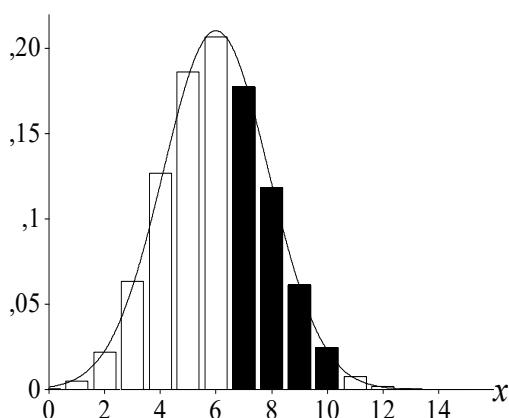
La  $X$  være binomisk med parametre  $n$  og  $p$ . Ved bruk av normaltilnærmelse skal vi da bruke følgende formel til å beregne sannsynligheten  $P(X \leq b)$ :

$$P(X \leq b) \approx \Phi\left(\frac{b - np + 0,5}{\sqrt{np(1-p)}}\right)$$

Tallet 0,5 som inngår i formelen er et «**kontinuitets-korreksjonsledd**», også kalt **halvkorreksjon** (h.k.).

**Eks. 5.18****Tilnærmelse binomisk - normal.**

La oss som et illustrerende eksempel si at vi ønsker å finne sannsynligheten  $P(7 \leq X \leq 10)$  når  $X$  er binomisk med  $n = 15$  og  $p = 0.4$ . Eksemplet er illustrert i figuren nedenfor. Når vi betrakter  $X$  som en kontinuerlig variabel, er det jo som vi ser egentlig  $P(6.5 \leq X \leq 10.5)$  vi ønsker å beregne, og det er dette forhold som gjenspeiler seg i korreksjonsleddet på 0.5.

**Histogram:**

Binomisk fordeling med parametre  $n = 15$  og  $p = 0.4$ .

**Heltrukken kurve:**

Tilpasset normalfordeling.

Fig. 5.11 Normaltilpasning til binomisk fordeling.

I dette eksemplet er  $np(1-p) = 15 \cdot 0.4 \cdot 0.6 = 3.6$ , som er *mindre* enn 5. La oss likevel se om normaltilnærmelsen skulle være brukbar i dette tilfellet. Vi får:

**Eksakt verdi (binomisk tabell):**

$$P(7 \leq X \leq 10) = P(X \leq 10) - P(X \leq 6) = .991 - .610 = .381$$

**Tilnærmet verdi (normaltilnærmelse):**

I sannsynlighetsuttrykket  $P(X \leq 10)$  er  $b = 10$ , og i sannsynlighetsuttrykket  $P(X \leq 6)$  er  $b = 6$ . Vi får derfor, ved bruk av formelen i forrige ramme:

$$P(7 \leq X \leq 10) \approx P(X \leq 10) - P(X \leq 6)$$

$$= \Phi\left(\frac{(10-6)+0.5}{\sqrt{3.6}}\right) - \Phi\left(\frac{(6-6)+0.5}{\sqrt{3.6}}\right)$$

$$= \Phi(2.37) - \Phi(0.26) \approx .991 - .603 = .388$$

Som vi ser er den tilnærmede verdien, 0.388, ikke så langt fra den eksakte, 0.381 (1.6 % avvik). Når det gjelder rimelig symmetriske og entoppede fordelinger som i dette tilfellet, kan vi nok velge en noe lavere grense enn tallet  $np(1-p) = 5$  for en brukbar tilnærmelse. ☺

## 5.12 Normaltilnærmelse til Poisson-fordelingen

Normaltilnærmelsen til Poisson-fordelingen er helt analog med normaltilnærmelsen til binomisk fordeling. Dette skyldes at Poisson-fordelingen er tilnærmet lik en binomisk fordeling med liten  $p$  og stor  $n$ . Kriteriet  $np(1-p)$  for binomisk fordeling reduseres da til kravet  $\lambda \geq 5$  for Poisson-fordelingen.

### Tilnærmelse Poisson – normal

La  $X$  være Poisson-fordelt med parameter  $\lambda \geq 5$ . Da er  $X$  tilnærmet normalfordelt  $N(\mu, \sigma)$  med  $\mu = \lambda$  og  $\sigma = \sqrt{\lambda}$ . Vi skal i dette tilfellet bruke følgende formel for å beregne tilnærmet sannsynlighet  $P(X \leq b)$  ved hjelp av  $N(0,1)$ -tabell:

$$P(X \leq b) \approx \Phi\left(\frac{b - \lambda + 0,5}{\sqrt{\lambda}}\right)$$

**Eks. 5.19**

### Tilnærmelse Poisson - normal.

La  $X$  være binomisk fordelt med parametre  $n = 1000$  og  $p = 0,01$ .

#### Oppgave

Bestem tilnærmet verdi for  $P(12 < X < 16)$ .

#### Løsningsforslag

Her er  $n \geq 10$  og  $p \leq 0,1$ , slik at  $X$  er tilnærmet Poisson-fordelt med parameter  $\lambda = np = 1000 \cdot 0,01 = 10$ . Videre er  $\lambda \geq 5$ , slik at vi kan bruke normaltilnærmelsen. Vi får:

$$\begin{aligned} P(12 < X < 16) &= P(X \leq 15) - P(X \leq 12) \\ &\approx \Phi\left(\frac{15 - 10 + 0,5}{\sqrt{10}}\right) - \Phi\left(\frac{12 - 10 + 0,5}{\sqrt{10}}\right) \\ &= \Phi(1,74) - \Phi(0,79) = .959 - .785 = \underline{0,174} \end{aligned}$$

Poisson-tabell:  $P(X \leq 15) - P(X \leq 12) = .951 - .792 = \underline{.159}$ . Om tilnærmelsen er god nok må vurderes ut fra situasjonen der tilnærmelsen skal brukes. Det er forøvrig ikke uvanlig å bruke strengere grenser enn  $\lambda \geq 5$ , f.eks.  $\lambda \geq 15$ . ☺

### 5.13 Kjikvadrat-fordelingen

Kjikvadrat-fordelingen kommer typisk til anvendelse når vi har med kvadratsummer å gjøre, og er blant annet en viktig fordeling innen variansanalyse og diverse tester for hvor godt modeller passer til data. Vi lar fordelingen ha betegnelsen  $\text{Kji2}(n)$ , der parameteren  $n$  kalles «antall frihetsgrader» (d.f. = «degrees of freedom» på engelsk). Utgangspunktet for kjikvadrat-fordelingen er fordelingen til en sum av kvadratene til  $n$  uif  $N(0,1)$ -variabler, som kan vises å være  $\text{Kji2}(n)$ -fordelt. Når antall frihetsgrader,  $n$ , går mot uendelig, blir kjikvadrat-fordelingen mer og mer lik normalfordelingen.

**Kjikvadrat-fordelingen**  $\text{Kji2}(n); n = 1, 2, \dots$

Tetthetsfunksjon: 
$$f(x) = \frac{x^{(n-2)/2} e^{-x/2}}{2^{n/2} \Gamma(n/2)}, \quad x \geq 0$$

Forventning:  $E(X) = n$ , Standardavvik:  $\text{std}(X) = \sqrt{2n}$

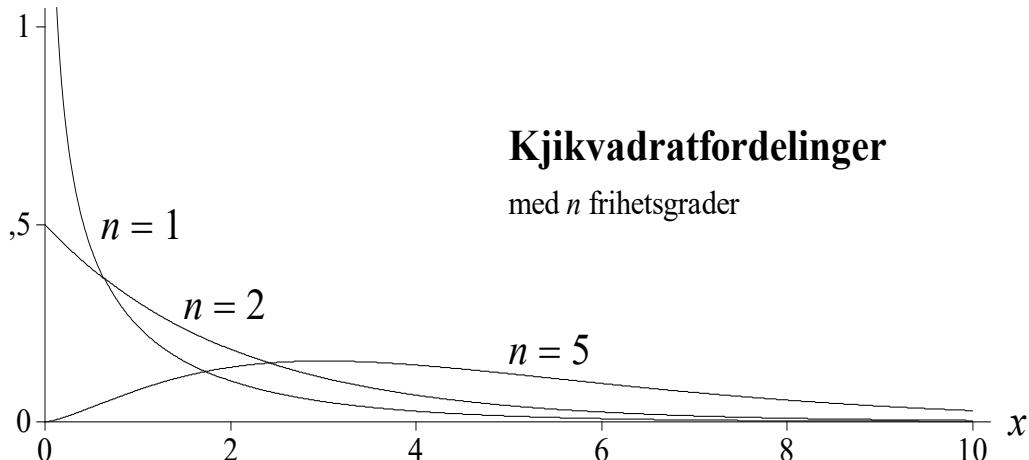


Fig. 5.12 Tre forskjellige kjikvadrat-fordelinger.  $n = 2$  tilsvarer  $\text{expo}(2)$ -fordelingen.

Følgende viktige resultat knytter kjikvadrat-fordelingen til fordelingen av empirisk varians  $S^2$  for uif  $N(\mu, \sigma)$ -variabler:

### **$S^2$ - fordeling**

La  $X_1, X_2, \dots, X_n$  være uif  $N(\mu, \sigma)$ -variabler med ukjent  $\mu$  og  $\sigma$ , og la  $\bar{X}$  betegne middelverdien til  $X$ -ene. Da vil

$$(n-1)S^2 / \sigma^2 = \sum_{i=1}^n (X_i - \bar{X})^2 / \sigma^2$$

være  $Kji2(n-1)$ -fordelt, dvs. kjikvadrat-fordelt med  $n-1$  frihetsgrader.

#### **Eks. 5.20**

**Variasjonskoeffisient.** La  $X_1, \dots, X_n$  betegne  $n$  uif  $N(\mu, \sigma)$ -variabler med ukjente parametre  $\mu$  og  $\sigma$ .

#### *Oppgave*

Beregn variasjonskoeffisienten til  $S^2$ .

#### *Løsningsforslag*

Siden  $(n-1)S^2 / \sigma^2$  er  $Kji2(n-1)$  så må  $E[(n-1)S^2 / \sigma^2] = n-1 \Rightarrow E S^2 = \sigma^2$ . Videre er  $\text{std}[(n-1)S^2 / \sigma^2] = \sqrt{2(n-1)} \Rightarrow \text{std}(S^2) = \sqrt{2/(n-1)} \cdot \sigma^2$ . Variasjonskoeffisienten,  $\text{CV}(S^2)$ , blir derfor:

$$\text{CV}(S^2) = \text{std}(S^2) / E(S^2) = \sqrt{\frac{2}{n-1}}$$

Legg merke til at variasjonskoeffisienten, som er et mål på relativ presisjon (nøyaktighet), ikke avhenger av den ukjente  $\sigma$ -verdien. ☺

## **5.14 t-fordelingen**

**t**-fordelingen med  $n$  frihetsgrader er for store  $n$  svært lik  $N(0,1)$ -fordelingen. Den kommer blant annet til anvendelse når vi har  $n$  uavhengige  $N(\mu, \sigma)$ -variabler med ukjent  $\sigma$  og relativt liten  $n$ . Vi betegner disse med  $X_1, \dots, X_n$ . Vi vet da at middelverdien  $\bar{X}$  er  $N(\mu, \sigma/\sqrt{n})$  slik at  $Z = (\bar{X} - \mu) / (\sigma/\sqrt{n})$  er  $N(0,1)$ . Erstatter vi nå  $\sigma$  med  $S$  kan det vises at

$$(5.8) \quad t = \frac{\bar{X} - \mu}{S / \sqrt{n}}$$

er  $t$ -fordelt med  $n-1$  frihetsgrader der «som vanlig»

$$S^2 = \sum_{i=1}^n (X_i - \bar{X})^2 / (n-1).$$

**t-fordelingen**  $\text{stud}(n); n = 1, 2, \dots$

$$\text{Tetthetsfunksjon: } f(x) = \frac{\left(\Gamma((n+1)/2)\right) \cdot \left(1 + (x^2/n)\right)^{-(n+1)/2}}{\sqrt{\pi n} \cdot \Gamma(n/2)}, -\infty < x < \infty$$

$$\text{Forventning: } E(X) = 0, \text{ Standardavvik: } \text{std}(X) = \sqrt{n/(n-2)}, n > 2$$

Merk at standardavviket i  $t$ -fordelingen alltid er større enn 1 og nærmer seg 1 når antall frihetsgrader,  $n$ , øker. Dette er naturlig, fordi  $\sigma$  er ukjent og må erstattes med en stokastisk variabel  $S$ , som bidrar til å øke variansen. Når  $n$  er stor, vil  $S$  være tilnærmet lik  $\sigma$  med stor sannsynlighet, og  $t$  definert ved lign.(5.2) vil være tilnærmet  $N(0,1)$ . Figuren nedenfor viser  $t$ -fordelingen med  $n = 3$  frihetsgrader sammenlignet med  $N(0,1)$ -fordelingen.

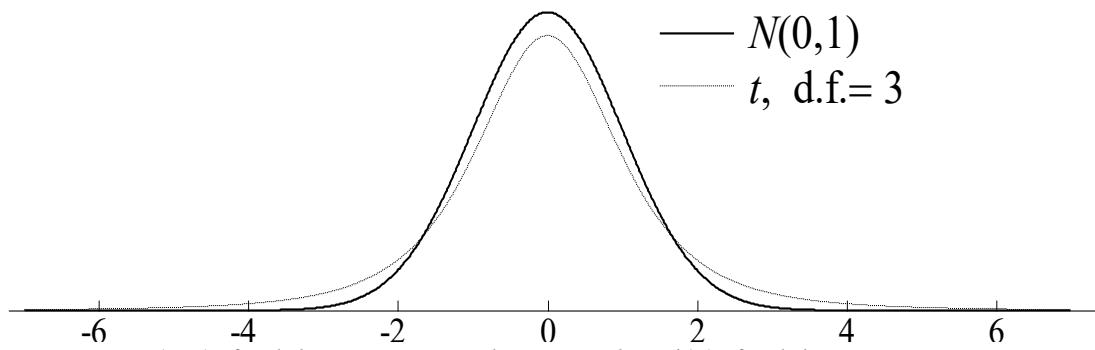


Fig. 5.13  $N(0,1)$ -fordelingen sammenlignet med  $\text{stud}(3)$ -fordelingen.

## 5.15 F-fordelingen

F-fordelingen har utstrakt anvendelse innen variansanalyse. Den er knyttet til forholdet mellom kjikvadrat-fordelte kvadratsummer:

### F-fordeling

$F(m,n)$

La  $\chi_m^2$  og  $\chi_n^2$  være to uavhengige kjikvadrat-fordelte variabler med henholdsvis  $m$  og  $n$  frihetsgrader. Da vil fordelingen til

$$F_{m,n} = \frac{\chi_m^2 / m}{\chi_n^2 / n}$$

være  $F$ -fordelingen med henholdsvis  $m$  og  $n$  frihetsgrader (d.f., «degrees of freedom»). Vi betegner denne med  $F(m,n)$ . Videre skal vi betegne øvre  $\alpha$ -fraktil i denne  $F$ -fordelingen med  $F_\alpha(m,n)$ , som tilsvarer nedre  $1-\alpha$ -fraktil.

Byttes teller og nevner får vi det viktige resultatet at

$$F_{n,m} = 1 / F_{m,n} \text{ er } F\text{-fordelt med d.f. } = (n,m).$$

NB! Merk at det antall frihetsgrader som angis først er knyttet til teller, mens det antall frihetsgrader som angis sist er knyttet til nevner.

Matematisk er  $F$ -fordelingen gitt som følger:

### F-fordelingen $F(m,n)$

Tetthetsfunksjon:  $f(x) = \frac{\Gamma((m+n)/2)(m/n)^{m/2} x^{(m-2)/2}}{\Gamma(m/2)\Gamma(n/2)(1+(m/n)x)^{(m+n)/2}}, \quad x \geq 0$

Forventning:  $E(X) = n/(n-2), \quad n > 2$

Standardavvik:  $\text{std}(X) = \sqrt{\frac{2n^2(m+n-2)}{m(n-2)^2(n-4)}}, \quad n > 4$

**Eks. 5.21** **Bruk av F-tabell.** La  $F_{4,2}$  betegne en variabel som er  $F$ -fordelt med henholdsvis 4 og 2 frihetsgrader.

*Oppgave*

- Bestem øvre 5-prosentil i  $F$ -fordelingen, dvs. bestem  $F_{.05}(4,2)$ .
- Bestem nedre 5-prosentil, dvs.  $F_{.95}(4,2)$ .

*Løsningsforslag*

Fra øvre 5-prosentil-tabellen bak i boka klipper vi følgende:

$m \rightarrow$	1	2	3	4
$n \downarrow$				
1	161,45	199.50	215.71	224.58
2	18.513	19.000	19.164	19.247
3	10.128	9.552	9.277	9.117
4	7.709	6.944	6.591	6.388

↑  
 $F_{.05}(2,4) = 6,944$

$\leftarrow F_{.05}(4,2) = 19,247$

- Vi ser direkte fra tabellen at  $F_{.05}(4,2) = \underline{19.247}$
- Siden  $F_{4,2} = 1/F_{2,4}$  finner vi nedre 5-prosentil i  $F(4,2)$ -fordelingen som den inverse av øvre 5-prosentil i  $F(2,4)$ -fordelingen. Fra tabellen får vi da:  
 $F_{.95}(4,2) = 1/6.944 = \underline{0.144} \quad \odot$

## 5.16 Binormal fordeling

Den binormale fordelingen er en kontinuerlig fordeling av et stokastisk variabelpar  $(X, Y)$  der  $X$  og  $Y$  hver for seg er normalfordelte enkeltvariabler, og der samvariasjonen mellom  $X$  og  $Y$  er entydig beskrevet ved korrelasjonskoeffisienten,  $\rho = \text{Corr}(X, Y)$ :

### **Binormal fordeling $N_2(\mu_1, \sigma_1; \mu_2, \sigma_2; \rho)$**

En binormal fordeling  $f(x,y)$  for  $(X,Y)$  er entydig gitt ved at

$$X \sim N(\mu_1, \sigma_1), \quad Y | X \sim N(\mu_2 + \rho(\sigma_2/\sigma_1)(x - \mu_1), (1 - \rho^2)^{1/2} \sigma_2)$$

eller ekvivalent

$$Y \sim N(\mu_2, \sigma_2), \quad X | Y \sim N(\mu_1 + \rho(\sigma_1/\sigma_2)(y - \mu_2), (1 - \rho^2)^{1/2} \sigma_1)$$

der  $\rho = \text{Corr}(X, Y)$ . Tetthetsfunksjonen  $f(x,y)$  kan skrives:

$$f(x,y) = \frac{\exp\left\{-\frac{1}{2} \frac{\sigma_2^2(x-\mu_1)^2 - 2\rho\sigma_1\sigma_2(x-\mu_1)(y-\mu_2) + \sigma_1^2(y-\mu_2)^2}{\sigma_1^2\sigma_2^2(1-\rho^2)}\right\}}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}},$$

$-\infty < x < \infty, \quad -\infty < y < \infty$

Grafisk får den binormale fordeling en symmetrisk klokkeform når  $X$  og  $Y$  er ukorrelerte ( $\rho = 0$ ) og begge akser er «like mye strukket». Når  $X$  og  $Y$  er korrelerte, blir utseendet mer lik en smal hatt, se neste figur.

Den binormale fordeling kan skrives på mer kompakt form som følger:

$$(5.9) \quad f(\underline{x}) = |2\pi\Sigma|^{-1} \cdot \exp(-\frac{1}{2}(\underline{x} - \underline{\mu})^T \Sigma^{-1} (\underline{x} - \underline{\mu}))$$

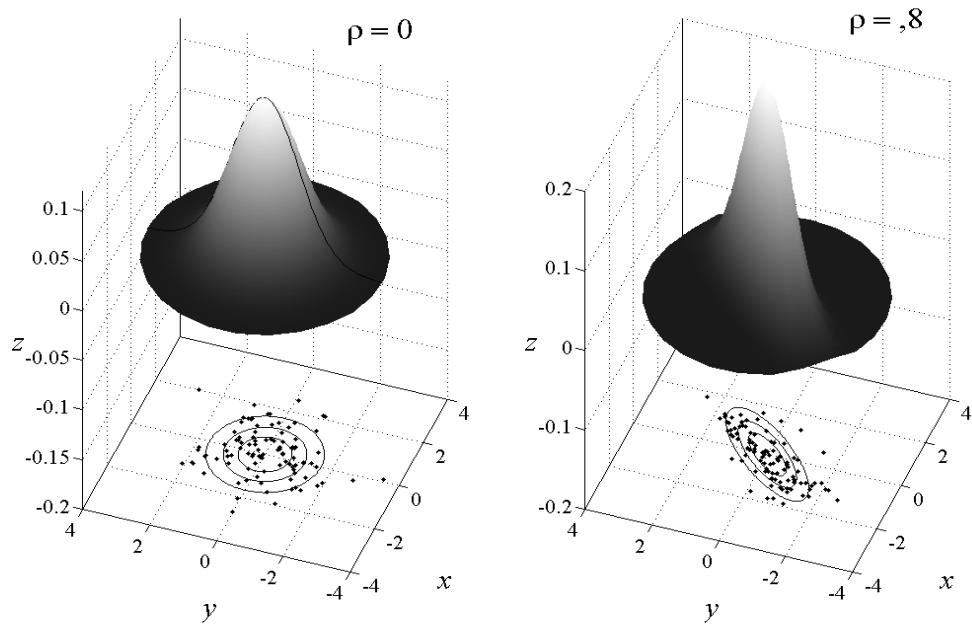
der  $\underline{x} = [x \ y]^T$ ,  $\underline{\mu} = [\mu_1 \ \mu_2]^T$ , absoluttverdi angir determinant, superskript T og  $-1$  anvendt på matriser angir henholdsvis transponert og invers, og  $\Sigma$  er kovariansmatrisen gitt som følger:

$$(5.10) \quad \Sigma = \begin{bmatrix} \text{Cov}(X,X) & \text{Cov}(X,Y) \\ \text{Cov}(Y,X) & \text{Cov}(Y,Y) \end{bmatrix} = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}$$

Kovariansmatrisen kan skrives mer kompakt som

$$\Sigma = \text{Cov}(\underline{X}, \underline{X}) = E((\underline{X} - \underline{\mu})(\underline{X} - \underline{\mu})^T) = E[\underline{X}\underline{X}^T] - E[\underline{X}]E[\underline{X}^T].$$

Det kan videre vises at eksponentuttrykket  $Q = (\underline{X} - \underline{\mu})^T \Sigma^{-1} (\underline{X} - \underline{\mu})$  er expo(2)-fordelt, hvilket kan benyttes til å konstruere tetthetskorturer (ellipser) i  $(x,y)$ -planet med gitt sannsynlighetsmasse innenfor ellipsenes avgrensninger (nest figur).



*Figur 5.14 Tre-dimensjonal fremstilling av to binormale fordelinger med konturplot med konstant tetthet inntegnet. Punktene viser 100 Monte Carlo-simulerte verdier fra de respektive fordelinger. Henholdsvis 25 %, 50 % og 75 % av sannsynlighetsmassen (volum under tetthetsflata) ligger innenfor innerste, midterste og ytterste kontur.*

**NB!** Dersom  $(X, Y)$  er binormalt fordelt og  $X$  og  $Y$  er ukorrelerte, så vil også  $X$  og  $Y$  være uavhengige stokastiske variabler. For binormale variabler er det følgelig ekvivalens mellom egenskapene ukorrelerthet og uavhengighet. Husk at dette ikke gjelder generelt. I mange tilfeller vil ikke  $X$  og  $Y$  være uavhengige selv om de er ukorrelerte.

Den binormale fordeling kan enkelt generaliseres til den  $n$ -dimensjonale multinormale fordeling med vilkårlig stor  $n$ . Den multinormale fordeling er den desidert mest utbredte multivariate fordeling på eksplisitt form. Det er faktisk svært komplisert å konstruere eksplisitte uttrykk for multivariate fordelinger, og det finnes svært få slike i litteraturen.

## 5.17 Oppgaver

**5.1** La  $X_1, X_2$  og  $X_3$  være 3 stokastisk uavhengige variabler der  $E(X_i) = (-1)^i$  og  $\text{Var}(X_i) = i$ ;  $i = 1, 2, 3$ . La  $Y$  være definert som  $Y = 2X_1 - 3X_2 - X_3$ . Bestem forventning og standardavvik til  $Y$ .

**5.2** La  $X_1$  og  $X_2$  være 2 stokastisk uvahengige  $N(0,1)$ -variabler. En tredje  $N(0,1)$ -variabel,  $X_3$ , er korrelert med både  $X_1$  og  $X_2$  ved at  $\text{Cov}(X_1, X_3) = .5$  og  $\text{Cov}(X_2, X_3) = -0.5$ .

Bestem forventning og varians til

$Y = X_1 + X_2 + X_3$ , samt korrelajonskoef- fisienten mellom  $X_1$  og  $X_3$ .

**5.3** La  $Z$  være en  $N(0,1)$ -variabel, og finn ved hjelp av tabell følgende sannsynligheter:

- a)  $P(-3 < Z < 3)$
- b)  $P(-2 < Z < 2)$
- c)  $P(-1 < Z < 1)$
- d)  $P(-5 < Z < 5)$
- e)  $P(2.11 < Z < 3.38)$
- f)  $P(-1.74 < Z < -0.27)$

**5.4** La  $Z$  være en  $N(0,1)$ -variabel, og finn ved hjelp av tabell den  $z$ -verdi som er slik at følgende likheter blir tilfredsstilt:

- a)  $P(Z < z) = .95$
- b)  $P(-z < Z < z) = .95$
- c)  $P(Z > z) = 0.5$
- d)  $P(|Z| > z) = 0.05$
- e)  $P(-z < Z < z) = 0.01$

**5.5** La  $X$  være en  $N(\mu, \sigma)$ -variabel, og bestem følgende sannsynligheter:

- a)  $P(2 < X < 3)$  når  $\mu = 2.5$  og  $\sigma = 1$

b)  $P(-5 < X < -4)$  når  $\mu = 1$  og  $\sigma = .5$

c)  $P(X > 8)$  når  $\mu = 0$  og  $\sigma = 2$

### 5.6

- a) Anta at  $X$  er  $N(0,1)$ .  
Finn  $P(-0.47 \leq X \leq 0.94)$ .

- b) Anta at  $X$  er  $N(0,1)$ .  
Finn  $P(X > 1.00)$ .

- c) Anta at  $X$  er  $N(9,9)$ .  
Finn  $P(X > 2.00)$ .

- d) Anta at  $X$  er  $N(9,9)$ .  
Finn  $P(5 \leq X \leq 11)$ .

**5.7** En tappemaskin for kartonger med 1 liter melk er innstilt på 1.004 liter. Det er kjent at forventningen for tappevolumet da er lik 1.004 liter med et standardavvik lik 0.005 liter. Man går ut fra at tappevolumet er normalfordelt. Vis at sannsynligheten for at en melkekartong inneholder mindre enn 1 liter melk er 0,2119.

**5.8** Anta at  $X$  er  $N(5,10)$ . La  $k > 0$ . Bestem  $k$  slik at  $P(5 - k < X < 5 + k) = 0.90$ .

**5.9** En bestemt sort batterier har en levetid som er tilnærmet normalfordelt med forventning 1200 dager og standardavvik 100 dager. Hvor lang tid bør garantitiden være, hvis produsenten tar sikte på at 10 prosent av batteriene skal gi grunnlag for reklamasjon fordi levetiden er for kort?

**5.10** Målefeilen til et instrument som måler blodsukkernivået er normalfordelt med forventning 0.05 og standardavvik 1.5. Dvs. ved gjentatte

forsøk vil fordelingen til differansen (målt nivå minus sant nivå) være  $N(0.05, 1.5)$ .

- a) Hvor stor prosentandel av målingene overestimerer det sanne nivået?
- b) Anta at en målefeil anses som alvorlig når den målte verdien avviker fra den sanne verdien med mer enn 2.8. Hvor stor prosentandel av målingene vil være alvorlig feil?

**5.11** En bedrift framstiller en bestemt type brød. Vekten i gram av tilfeldig valgte brød antas å være uavhengige og normalfordelt  $N(\mu, \sigma)$ . Forventningen  $\mu$  vil anhenge av deigen, innstillingen av maskinen som porsjonerer ut deigen, samt steketiden. Ifølge forskriftene skal denne type brød veie minst 750 gram.

Anta nå at en bestemt produksjonsplan innebærer at  $\mu = 760$  og  $\sigma = 10$ , hvilket vil si at et tilfeldig valgt brød er  $N(760, 10)$ .

- a) Hva er sannsynligheten for at et tilfeldig valgt brød er undervektig?
- b) Hvis 3 brød kjøpes, hva er da sannsynligheten for at minst ett er undervektig?
- c) Hvis 3 brød kjøpes, hva er sannsynligheten for at det tyngste brødet veier mer enn 780 gram?
- d) Finn sannsynligheten for at den gjennomsnittlige vekten av fem tilfeldig valgte brød er mindre enn 750 gram.

**5.12**  $X$  er binomisk fordelt med  $n = 100$  og  $p = 0,1$ .

- a) Sett opp uttrykket for beregning av  $P(X \geq 8)$  (du skal ikke regne det ut).

Finn sannsynligheten under a) tilnærmet ved:

- b) Poissonfordelingen.
- c) Normalfordelingen.

**5.13** La  $X =$  årsinntekten til en nyutdannet diplomøkonom i 1983. Anta at  $X$  er uavhengig av kjønn og tilnærmet normalfordelt med  $\mu = 120000$  kroner og  $\sigma = 20000$  kroner. Finn

- a)  $P(X < 110000)$
- b)  $P(90000 < X < 150000)$

Bitten og Birger er to av de nyutdannete diplomøkonomer 1983, som nevnt ovenfor.

- c) Hva er sannsynligheten for at minst en av dem tjener mer enn 120000?
- d) Hva er sannsynligheten for at de til sammen tjener minst 220000?

**5.14** Et tilfeldig utvalg på  $n = 100$  blir tatt fra en populasjon som har en forventning på 20 og et standardavvik på 5. Formen på populasjonfordelingen er ukjent.

- a) Hva kan du si om fordelingen til utvalgsmiddelverdien,  $\bar{X}$ ?
- b) Finn tilnærmet sannsynlighet for at  $\bar{X}$  skal overskride 20.75.

**5.15** Fyll ut med det som mangler nedenfor:

$$P(Z > ) = 0.975, \text{ når } Z \text{ er } N(0,1)$$

$$P(Z < ) = 0.050, \text{ når } Z \text{ er } N(0,1)$$

$$P(Z = 0) = ? \text{ når } Z \text{ er } N(0,1)$$

$$P(-3.8 < X < 3.8) = ?, X \text{ er } N(-3.1, 0.7)$$

**5.16** Anta at «persene» på 60-meteren til studentene i Norge er normalfordelt

med forventning  $\mu = 9.2$  sek og standardavvik  $\sigma = 0.7$  sek.

- a) Finn sannsynligheten for at en tilfeldig valgt student har «pers» på under 8.0 sek.
- b) Hva er sannsynligheten for at det blant 5 tilfeldig valgte studenter er akkurat 3 som har «pers» over 8.0 sek?

**5.17** Anta at karakterene i begynnerkurs i statistikk ved universitet og høyskoler er tilnærmet normalfordelt med forventning  $\mu = 3.2$  og standardavvik  $\sigma = 0.8$ , og at 4.1 gir stryk.

- a) Hvor stor er strykprosenten?
- b) Hva er tilnærmet fordeling til gjennomsnittskarakteren til et tilfeldig utvalg på 35 besvarelser? Er fordelingen avhengig av at populasjonsfordelingen er normalfordelt?
- c) Nevn 2 (prinsipielle) grunner til at karakterer kun kan være *tilnærmet* normalfordelt.

**5.18** Anta at sannsynligheten for at en tilfeldig valgt student drikker mer enn 5 dl melk pr. dag er 0.6. Gitt en gruppe på 70 studenter. Finn sannsynligheten (tilnærmet) for at flere enn 30, men samtidig færre enn 50 av disse drikker mer enn 5 dl melk pr. dag.

**5.19** La  $X$  angi høyden (i cm) til en tilfeldig valgt norsk soldat. Vi skal anta at  $X$  er normalfordelt  $N(\mu, \sigma)$ , med forventning  $\mu = 179$  og standardavvik  $\sigma = 4.5$ .

- a) Hva er sannsynligheten for at en tilfeldig valgt soldat er minst 190 cm?

b) Av et tilfeldig utvalg på 5 soldater. hva er sannsynligheten for at den høyeste av de 5 er minst 190 cm?

c) Hva er sannsynligheten for at gjennomsnittshøyden av 100 tilfeldige soldater er mindre enn 190 cm?

Vi ser nå på sammenhengen mellom vekt (i kg) og høyde (i cm) for soldatene illustrert ved følgende simultanfordeling  $f(x,y)$ :

Høyde [cm] →	$\leq 179$	$> 179$
Vekt [kg] ↓		
$< 70$	.25	.05
$\geq 70$	.25	.45

d) Finn sannsynligheten for at en tilfeldig valgt soldat skal være lettare enn 70 kg, gitt at han er høyere enn 179 cm.

**5.20** Du har tilgjengelig slumptall mellom 0 og 1. Hvordan vil du gå fram for å benytte dette til å simulere et terningkast med rettferdig terning?

**5.21** Vis at  $F(x) = P(X \leq x) = x$  når  $X$  er  $U(0,1)$ -fordelt.

**5.22** Vis at  $E(X) = 0.5$  og at  $\text{std}(X) = 1/\sqrt{12}$  når  $X$  er  $U[0,1]$ -fordelt.

**5.23** Levetiden til et radioaktivt atom er eksponensialfordelt med parameter  $b = 1.000$  sek. Bestem halveringstida til det radioaktive atomet, dvs. den tida det tar før det er 50% sikkert at atomet vil disintegrere.

**5.24** Et støvkorn inneholder  $10^{11}$  atomer av et radioaktivt stoff med halveringstid  $t = t_{1/2} = 1$  minutt.

- a) Bestem parameteren  $b$  i eksponentielfordelingen til levetida til et vilkårlig av de radioaktive atomene.
- b) Bestem variasjonskoeffisienten til antall radioaktive atomer som er igjen ved tidspunktet  $t = t_{1/2}$ .

**5.25 (E)** En tappemaskin for kartonger med 1 liter melk er innstilt på 1.004 liter. Vi går ut fra at tappevolumet er normalfordelt med forventning lik 1.004 liter og standardavvik lik 0.005 liter.

- a) Hva er sannsynligheten for at en tilfeldig melkekartong inneholder mindre enn 1.000 liter?
- b) Vi kjøper 5 kartonger. Hva er sannsynligheten for at disse inneholder mindre enn 5.000 liter til sammen?

**5.26 (E)** Sannsynligheten for å få bivirkninger ved bruk av en bestemt medisin er lik 0.20. Denne medisinen blir gitt til 30 pasienter, hvor forekomstene av bivirkningene er uavhengige.

- a) Bestem forventningsverdien for antall pasienter som får bivirkninger.
- b) Bestem variansen og standardavviket for antall pasienter som får bivirkninger.
- c) Hva er sannsynligheten for at høyst 8 pasienter skal få bivirkninger? Bruk normaltilnærmelse.

**5.27 (E)**  $Z$  er en tilfeldig kontinuerlig variabel som er normalfordelt med forventning  $\mu = 0$  og varians  $\sigma^2 = 1$ .

- a) Finn  $P(Z \leq 0.75)$ ,  $P(Z > -0.75)$  og  $P(-0.75 < Z \leq 0.75)$ .
- b) Bestem  $k$  slik at  $P(-k < Z \leq k) = 0.9$ .

Resistansen for en bestemt type elektriske motstander antas å være normalfordelt med forventning  $\mu = 100 \Omega$  og varians  $\sigma^2 = (2 \Omega)^2$ . Motstandene er ubrukbar dersom resistansen er mindre enn  $98 \Omega$  eller større enn  $103 \Omega$ .

- c) Hva er sannsynligheten for at en fritt valgt motstand er ubrukbar?

Motstandene pakkes i esker med 10 stykker i hver eske.

- d) Hva er sannsynligheten for at en eske inneholder høyst en ubrukbar motstand?

**5.28 (E)** Ved måling av alkoholinnholdet i blod kan måleresultatet  $X$  antas å være normalfordelt  $N(\mu, \sigma)$ .  $\mu$  [promille] er den sanne verdien av alkoholinnholdet og  $\sigma$  [promille] er et mål for analysemetodens nøyaktighet. Anta at  $\sigma = 0.04$  promille er kjent.

- a) Hva er sannsynligheten for at et tilfeldig måleresultat overstiger 0.50 promille dersom den sanne verdien for alkoholinnholdet er 0.48 promille?

Ved en promilletest tas to uavhengige blodprøver av en person som er mistenkt for promillekjøring. Domstolen foreslår følgende kriterium: Personen dømmes for promillekjøring dersom det gjennomsnittlige alkoholinnholdet  $\bar{X}$  i de to prøvene overstiger 0.53 promille.

- b) Hva er sannsynligheten for at en person som har et alkoholinnhold på 0.48 promille dømmes for promille-

kjøring dersom dette kriteriet brukes?

Advokaten hevder at rettens fremgangsmåte gir altfor stor sannsynlighet for at en uskyldig blir dømt, og man blir derfor enige om å finne et nytt kriterium:

Personen skal dømmes hvis  $\bar{X} > k$ , der  $k$  skal bestemmes slik at sannsynligheten for at en person med alkoholinnhold på 0.49 promille blir dømt, er lik 0.01.

c) Bestem  $k$ .

**5.29 (E)** En bedrift produserer syltetøy som leveres på glass. Vekten  $X$  av innholdet (nettovekten) antas å være normalfordelt  $N(\mu, \sigma)$ , der  $\mu$  avhenger av innstillingen på tappemaskinen.

Vekten  $Y$  av emballasjen (glass med lokk) antas å være  $N(0.20 \text{ kg}, 0.01 \text{ kg})$ . Et syltetøyglass med nettovekt ( $X$ ) under 0.80 kg anses for å være undervektig. Anta at  $\sigma = 0.02 \text{ kg}$  er kjent, og at maskinen er innstilt slik at  $\mu = 0.82 \text{ kg}$ .

- Hva er sannsynligheten for at et tilfeldig syltetøyglass er undervektig?
- Hva er sannsynligheten for at nøyaktig 2 av 10 glass er undervektige?
- La  $Z$  være bruttoverkten (syltetøy + emballasje) av et tilfeldig glass. Hvilken sannsynlighetsfordeling får  $Z$ ? Finn sannsynligheten for at bruttoverkten er større enn 1.0 kg.
- Man ønsker nå å justere tappemaskinen slik at 90 % av bruttoverkten overstiger 1.00 kg. Hvilken

verdi for  $\mu$  må vi da innstille tappe-maskinen på?

**5.30 (E)** En mann trenger 165 kg kunstgjødsel. Kunstgjødselen leveres i sekker á 50 kg og 10 kg. Vekten av innholdet i en 50 kg-sekk antas å være normalfordelt  $N(50, 2.0)$ . Vekten av innholdet i en 10 kg-sekk antas å være normalfordelt  $N(10, 0.4)$ . Vektene av forskjellige sekker antas å være uavhengige. Mannen bestemmer seg for å kjøpe 3 stk. 50 kg-sekker og 2 stk. 10 kg-sekker.

Vi lar  $Z$  betegne totalvekten av de 5 sekkene.

- Hvilken sannsynlighetsfordeling har  $Z$ ?
- Beregn sannsynligheten for at mannen får for lite kunstgjødsel (mindre enn 165 kg).
- Beregn sannsynligheten for at den tyngste av de tre 50 kg-sekkene veier mer enn 53 kg.
- Beregn sannsynligheten for at de tre 50 kg-sekkene i gjennomsnitt veier mer enn 51 kg.
- Hvor mange 50 kg-sekker må kjøpes for at sannsynligheten for at minst en skal veie over 53 kg er minst 0.5?

**5.31 (E)** En fabrikk produserer vareenheter med vekt  $Y$  [gram] som antas å være normalfordelt  $N(\mu, \sigma)$ . En kasseringautomat kasserer vareenheter som veier mindre enn 1168 g, og vareenheter som veier mer enn 1518 g. En har funnet ut at i det lange løp blir 12,3 % av enhetene kassert fordi de er

for lette, mens 27.7 % av artiklene blir kassert fordi de er for tunge.

- Vis at  $\sigma = 200$  og  $\mu = 1400$ .
- Hva er sannsynligheten for at en vareenhet har en vekt som er større enn 1755 g?

Vi tar ut et vareparti på 5 enheter.

- Hva er sannsynligheten for at nøyaktig  $x$  av de 5 enhetene hver har en vekt som er større enn 1755 g?
- Hva er sannsynligheten for at minst en av de fem vareenhetene har en vekt som er større enn 1755 g?
- Hva er sannsynligheten for at de fem vareenhetene totalt veier mer enn 7.5 kg?
- Bestem tilnærmet sannsynlighet for at det i et stort vareparti på 1000 enheter, høyst vil være 45 som veier over 1755 g.

**5.32 (E)** Anta at  $X$  har en standard normalfordeling og bestem:

- $P(X \leq 0.42)$  og  $P(X \leq -1.69)$
- $k$  slik at  $P(-0.3 < X \leq k) = 0.2$

Med et doseringsapparat fylles det automatiske fargepulver i poser på et transportbånd. Apparatet kan innstilles slik at vekten  $X$  av fargepulver i hver pose blir mellom 10 og 30 gram. Med apparatet innstilt på en bestemt vekt, kan  $X$  betraktes som en normalfordelt stokastisk (tilfeldig) variabel med forventning  $\mu = \text{innstilt vekt}$  og standardavvik  $\sigma = 0.5$  gram.

- Anta at innstilt vekt er  $\mu = 20.4$  gram, og at det for hver pose fylles to ganger etter hverandre for å gi samlet vekt på 40,8 gram. Hvor stor andel

av posene vil en forvente har en samlet vekt mindre eller lik 40 gram?

- Anta at innstilt vekt er  $\mu = 15$  gram, og at det for hver pose bare fylles én gang. Hvor stor er da sannsynligheten for at vektforskjellen mellom to tilfeldig valgte poser er større enn 0.7 gram?

**5.33 (E)** Vi skal få anta at vekten  $X$  i kg for en voksen mann i en bestemt befolkningsgruppe er normalfordelt med forventning  $\mu = 75$  kg og standardavvik  $\sigma = 15$  kg.

- Finn sannsynligheten for at en tilfeldig mann veier mer enn 90 kg. For en gruppe på 4 menn, finn sannsynligheten for at minst 1 veier mer enn 90 kg.
- En varsellampe i en heis begynner å lyse når den samlede vekten av personene i heisen er mer enn 800 kg. Finn sannsynligheten for at lampa vil lyse når 10 tilfeldige menn har gått inn i heisen.
- Hvis  $Y$  er vekten av en voksen kvinne i den samme befolkningsgruppen, har det vist seg at  $Y = 0.8X + 2$ , der  $X$  har samme fordeling som  $X$ . Finn  $E(Y)$ ,  $\text{std}(Y)$  og sannsynligheten for at en tilfeldig kvinne veier mindre enn 52 kg.
- 8 kvinner og 4 menn går inn i den heisen vi omtalte ovenfor. Finn sannsynligheten for at varsellampa vil lyse denne gangen.

**5.34 (E)** En maskin lager sylinderformede bolter. Det har vist seg at en bolts diameter  $D$  er normalfordelt med forventning

$\mu = 2 \text{ cm}$  og standardavvik  $\sigma = 0.1 \text{ cm}$ . En bolt anses å være ubrukbar for et bestemt formål dersom  $D$  avviker fra  $\mu$  med mer enn  $0.1 \text{ cm}$ . Diameteren til en bolt antas å være uavhengig av diameteren til de andre boltene.

- a) Hva er sannsynligheten for at en bolts diameter er større enn  $2.1 \text{ cm}$ ?
- b) Bestem sannsynligheten for at en tilfeldig bolt er brukbar.

Boltene pakkes i esker med 10 bolter i hver eske. Den gjennomsnittlige diameteren til disse 10 boltene er

$$\bar{D} = \frac{1}{10} \sum_{i=1}^{10} D_i$$

- c) Hva er sannsynligheten for at den gjennomsnittlige diameteren er mindre enn  $2.02 \text{ cm}$ ?
- d) Bestem sannsynligheten for at minst 9 av boltene i en tilfeldig eske er brukbare.

Boltene skal ned i et hull hvis diameter  $Y$  er normalfordelt med  $E(Y) = 2.2 \text{ cm}$  og  $\text{std}(Y) = 0.2 \text{ cm}$ .

- e) Hva er sannsynligheten for at en tilfeldig valgt bolt skal gå ned i et tilfeldig valgt hull?

**5.35 (E)** Levetiden  $T$  (i timer) for en viss type lyspærer betraktes som en stokastisk variabel med sannsynlighetsfordeling gitt ved frekvensfunksjonen (tethetsfunksjonen):

$$f(t) = \begin{cases} k \cdot e^{-0.001t}, & t > 0 \\ 0, & t \leq 0 \end{cases}$$

- a) Bestem  $k$  (klarer du ikke det, bruk  $k = 0.005$  i resten av oppgaven).

- b) Beregn den forventede levetiden  $m$  for slike lyspærer.
- c) Beregn sannsynligheten for at en lyspære skal ha kortere levetid enn  $m$ .
- d) En lyspære har brent i 900 timer og er fremdeles i orden. Hvor stor er sannsynligheten for at den fortsetter å brenne i minst 200 timer til?

**5.36 (E)** La  $X \sim N(0,1)$ ,  $Y = 2X+3$  og  $Z = X^2$ .

- a) Bestem  $P(X < 2)$ ,  $P(Y > 2)$  og  $P(Z < 2)$ .
- b) Bestem  $E(Y)$ ,  $\text{Var}(Y)$  og  $E(Z)$ . Finn også sannsynlighetstettheten til  $Z$ .

La  $X_1, X_2, \dots, X_{10}$  alle være uavhengige og  $N(0,1)$ -fordelte.

- c) Beregn  $P(X_1 + \dots + X_{10} > 2)$  og  $P(X_1 + \dots + X_7 < 2X_8 + 2X_9 + 2X_{10})$

**5.37 (E)** Anta at  $X$  er normalfordelt  $N(0,1)$ .

- a) Beregn  $P(X \leq 0.80)$ ,  $P(X > -0.80)$  og  $P(-1.35 < X \leq 0.95)$ .
- b) Bestem konstantene  $a$ ,  $b$ ,  $c$  slik at

$$\begin{aligned} P(X \leq a) &= 0.90, \\ P(X > b) &= 0.025, \\ P(c < X \leq 0.35) &= 0.50 \end{aligned}$$

En bedrift produserer betongsylinger av en bestemt type. Sylinderne trykkfasthet [ $\text{N/mm}^2$ ] er uavhengige normalfordelte variabler med forventning 29 og standardavvik  $0.5 \text{ N/mm}^2$ . Sylinderne er uegnet for sitt formål dersom de har trykkfasthet under  $28.4 \text{ N/mm}^2$ , og de blir i så fall kassert.

- c) Hva er sannsynligheten for at en tilfeldig sylinder blir kassert?

Hva er sannsynligheten for at en tilfeldig sylinder har en trykkfasthet på minst  $30,0 \text{ N/mm}^2$ ?

Bedriften har påtatt seg å leve 900 brukbare cylindere til en kunde. På grunn av faren for svinn produserer bedriften for sikkerhets skyld 1000 cylindere.

- d) Hvor stor er (tilnærmet) sannsynligheten for at oppdraget kan utføres? Forklar hvordan du resonnerer for å løse denne oppgaven.
- e) Hva er det minste antall cylindere bedriften må produsere hvis leveransen skal kunne gjennomføres med 99 % sannsynlighet?

**5.38 (E)** Et supermarket selger melk i kartonger på  $\frac{1}{2}$ , 1 og  $1\frac{1}{2}$  liter. Da tappingen ikke skjer med eksakt presisjon, vil vi betrakte melkevolumet i hver av kartongtypene som normalfordelte stokastiske variabler med følgende fordelinger:  $N(0,5, 0.05)$ ,  $N(1.0, 0.1)$  og  $N(1.5, 0.20)$ . Melkemengdene i de forskjellige kartongene antas å være uavhengige av hverandre.

- a) Hvor stor er sannsynligheten for at en 1-liters kartong inneholder mer enn 1.03 l?

Kunden skal ha 2.5 liter melk.

- b) Hva er sannsynligheten for at han får mindre enn 2.40 liter melk hvis han tar kartonger på 1 liter og  $1\frac{1}{2}$  liter?
- c) Hadde det vært fornuftig av kunden å velge andre kombinasjoner? Svar et skal begrunnes ved å regne ut

sannsynligheten for å få mindre enn 2.40 l melk ved andre måter å velge på.

- d) Bestem  $a$  slik at sannsynligheten for at samlet melkevolum ligger innenfor  $[6.0 - a, 6.0 + a]$  blir 0.95 hvis det kjøpes seks 1-liters melkekartonger.

**5.39 (E)** I langtids bølgestatistikk brukes ofte Weibull-fordelingen til å beskrive hvordan bølgehøyden varierer. La  $H$  betegne bølgehøyden (målt i meter) til en vilkårlig bølge. Da har  $H$  følgende fordelingsfunksjon:

$$F(h) = P(H \leq h) = 1 - e^{-\frac{h^\beta}{\alpha}}, h \geq 0$$

$\alpha$  og  $\beta$  er parametre som må estimeres ut fra de lokale forhold og observasjoner. I vårt tilfelle skal vi regne med  $\alpha = 1.12$  og  $\beta = 0.97$ .

- a) Hva er sannsynligheten for at en bølge har en høyde på over 4.5 m?
- Hva er sannsynligheten for at høyest 2 av 100 bølger er over 4.5 m høye?
- b) Bestem tilnærmet sannsynlighet for at høyest 200 av 10000 bølger er over 4.5 m høye.
- c) La oss anta at det kommer en ny bølge hvert 7. sekund på det stedet vi betrakter. Hvor mange bølger over 19 meter vil en forvente å observere i løpet av en 10 års-periode (regn med 365 dager pr. år)? Hva er tilnærmet sannsynlighet for å observere minst 4 bølger på over 19 meter i løpet av en 10 års-periode?

**5.40 (E)** Levetiden (i timer) til en bestemt type elektrisk komponent kan

betraktes som en stokastisk variabel  $X$  med sannsynlighetstetthet  $f_X(t) = \lambda e^{-\lambda t}$  for  $t \geq 0$  og  $f_X(t) = 0$  for  $t < 0$ . La  $\lambda = 0.001$  i hele denne oppgaven.

- a) Bestem  $P(X > 1000)$ . Finn også  $L_{25}$  som er slik at  $F_X(L_{25}) = \frac{1}{4}$  hvor  $F_X$  er kumulativ fordelingsfunksjon til  $X$ .

La  $X_1$  og  $X_2$  være levetiden til to elektriske komponenter av typen over. Vi forutsetter at  $X_1$  og  $X_2$  er stokastisk uavhengige. En kan da vise (det skal ikke du gjøre) at  $Y = X_1 + X_2$  har sannsynlighetstettheten  $f_Y(t) = \lambda^2 \cdot t e^{-\lambda t}$  for  $t \geq 0$ .

- b) Vis at den kumulative fordelingsfunksjonen  $F_Y$  til  $Y$  er gitt ved  $F_Y(y) = 1 - (1 + \lambda y) e^{-\lambda y}$  for  $y \geq 0$ . Bestem  $P(Y > 2000)$ .

La  $T_{\min} = \min(X_1, X_2)$ .  $T_{\min}$  er altså den maksimale tiden der begge komponentene fungerer.

- c) Vis at  $P(T_{\min} < t) = 1 - e^{-2\lambda t}$  for  $t \geq 0$ .

Hva er sannsynligheten for at begge komponentene fungerer etter 800 timer?

**5.41** En elektronisk bedrift trenger elektriske motstander i framstillingen av sine produkter. De motstandene som kjøpes inn har en resistans som er  $N(310, 10)$ .

- a) Beregn forventningsverdi og standardavvik for summen av resistansene for 5 slike motstander. Beregn sannsynligheten for at samlet resistans for de 5 motstandene overskridt 1600 ohm.

Ved en annen anledning trenger firmaet 80 slike motstander, men disse motstandene må hver ha en resistans mellom 300 og 330 ohm. Det kjøpes inn 100 motstander, som hver har resistans som er  $N(310, 10)$ .

- b) Hvor stor er sannsynligheten for å få minst 80 motstander med ønsket resistans?
- c) Motstandene er pakket i esker á 10 stykker. Hva er sannsynligheten for at en eske skal inneholde eksakt 8 motstander med resistans mellom 300 og 330 ohm?

**5.42** Tiden,  $Y$ , som går med for en dataterminal til å prosessere, editere og forandre på en storskjerm, er uniformt (rektagulært) fordelt mellom 0.75 og 2.5 sekunder.

- a) Bestem sannsynlighetstetthetsfunksjonen for  $Y$ . Tegn den. Beregn forventning,  $\mu$ , og standardavvik,  $\sigma$ , til  $Y$ .
- b) Bestem intervallet  $(\mu \pm \sigma)$  på kurven. Beregn sannsynligheten for  $(\mu - \sigma) < y < (\mu + \sigma)$ .

**5.43** En fabrikk produserer bolter. Boltenes diameter kan betraktes som en normalfordelt stokastisk variabel med forventning  $\mu = 8.20$  mm og standardavvik 0.14 mm.

- a) Finn sannsynligheten for at en tilfeldig valgt bolt har diameter mindre enn 8.40 mm.
- b) Beregn sannsynligheten for at eksakt 7 av 10 bolter har diameter mindre enn 8.40 mm.

- c) Finn sannsynligheten for at en tilfeldig valgt bolt har diameter større enn 8,40 mm, gitt at diametren er større enn 8,10 mm.

Boltene skal brukes sammen med muttere som produseres ved samme fabrikk. Mutrene har hull med diameter som antas normalfordelt med forventning 8,50 mm og standardavvik 0,12 mm. For at en mutter og en bolt skal passe sammen, må mutterens diameter være større enn boltens, men differansen må ikke overstige 0,60 mm.

- d) Finn sannsynligheten for at en tilfeldig valgt bolt skal passe sammen med en tilfeldig valgt mutter.

**5.44** Et firma produserer betongelementer med lengder 3 m og 6 m. Lengdene kan betraktes som uavhengige og normalfordelte stokastiske variabler, henholdsvis

$$X \sim N(3.00, 0.05) \text{ og } Y \sim N(6.00, 0.08).$$

- a) Beregn sannsynligheten for at et 3-meterselement skal være kortere enn 2,90 m.

Hvor stor er sannsynligheten for at et 6-meterselement er lengre enn 6,15 m, dersom vi vet at lengden er større enn 6,10 m?

- b) Et 6-meterselement settes sammen med et 3-meterselement. Beregn sannsynligheten for at den totale lengden blir mindre enn 8,90 m.

En husbygger skal dekke et spenn på 11,95 m. Hvilken av følgende kombinasjoner av elementer har størst sannsynlighet for å ha total lengde større enn 11,95 m:

- i)  $(3+3+3+3)$  m,

- ii)  $(3+3+6)$  m,  
iii)  $(6+6)$  m?

**5.45** Et meieri ønsker å framstille kartonger med kremfløte. Vektene i gram av tilfeldig valgte kartonger med kremfløte antas å være uif  $N(\mu, \sigma)$ . Ifølge forskriftene skal denne type kartonger veie minst 350 gram. Anta nå at en bestemt produksjonsplan innebærer at  $\mu = 360$  og  $\sigma = 8$ , hvilket vil si at vekten  $X$  på kartongene har fordeling  $N(360, 8)$ .

- a) Hva er sannsynligheten for at en tilfeldig valgt kartong er undervektig?  
 b) Hvis 3 kartonger kjøpes, hva er sannsynligheten for at minst en er undervektig?  
 c) Hvis 3 kartonger kjøpes, hva er sannsynligheten for at den tyngste kartongen veier mer enn 370 gram?  
 d) Bestem sannsynligheten for at den gjennomsnittlige vekten av fire tilfeldig valgte kartonger er mindre enn 350 gram.

## 5.18 Formelsamling

$f(x)$  = fordeling (tethetsfunksjon) til  $X$ .  
 $F(x) = P(X \leq x)$  = kumulativ fordelingsfunksjon til  $x$ .

$$\mu = E(X), \quad \sigma = \text{std}(X), \quad \bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$$

$$S^2 = \sum_{i=1}^n (X_i - \bar{X})^2 / (n-1)$$

$$\int_{-\infty}^{\infty} f(x) dx = 1$$

$$F(x) = \int_{-\infty}^x f(u) du$$

$$\mu = \int_{-\infty}^{\infty} x \cdot f(x) dx$$

$$\sigma^2 = \int_{-\infty}^{\infty} x^2 \cdot f(x) dx - \mu^2$$

### Sum av variabler

$$E\left(\sum_{i=1}^n a_i X_i\right) = \sum_{i=1}^n (a_i E X_i)$$

$$\text{Var}(aX + bY) = a^2 \text{Var}(X) + b^2 \text{Var}(Y) + 2ab \cdot \text{Cov}(X, Y)$$

### Uniform fordeling $U[a,b]$

$$f(x) = (b-a)^{-1}, \quad a \leq x \leq b$$

$$F(x) = (x-a)/(b-a)$$

$$\mu = \frac{1}{2} \cdot (a+b), \quad \sigma = (b-a)/\sqrt{12}$$

### Eksponensialfordeling $\text{expo}(b)$

$$f(x) = (1/b) \cdot e^{-x/b}, \quad x \geq 0$$

$$F(x) = 1 - e^{-x/b}$$

$$E(X) = \text{std}(X) = b$$

### Gammafunksjon $\Gamma(c)$

$$\Gamma(c) \equiv (c-1)! \equiv \int_0^{\infty} u^{c-1} e^{-u} du, \quad c > 0$$

### Gammafordeling gamma( $b,c$ )

$$f(x) = (x/b)^{c-1} e^{-x/b} / (b\Gamma(c)), \quad x \geq 0$$

$$E(X) = bc, \quad \text{std}(X) = b\sqrt{c}$$

### Weibullfordeling Weib( $b,c$ )

$$f(x) = (c/b) \cdot (x/b)^{c-1} e^{-(x/b)^c}, \quad x \geq 0$$

$$\mu = (b/c) \cdot \Gamma(1/c)$$

$$\sigma = (b/c) \cdot \sqrt{2c\Gamma(2/c) - \Gamma^2(1/c)}$$

### Rayleighfordeling Rayl( $b$ )

$$f(x) = 2b^{-2} \cdot e^{-(x/b)^2}, \quad x \geq 0$$

$$\mu = b\sqrt{\pi}/2, \quad \sigma = (b/2)\sqrt{4-\pi}$$

### Normalfordeling $N(\mu, \sigma)$

$$f(x) = (2\pi\sigma)^{-1/2} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad -\infty < x < \infty$$

### Bruk av $N(0,1)$ -tabell

Når  $X$  er  $N(\mu, \sigma)$  så er  $Z = (X-\mu)/\sigma \sim N(0,1)$ .  $P(X \leq b)$  finnes ved hjelp av formelen  $P(X < b) = \Phi((b-\mu)/\sigma)$  der  $\Phi$  er symbolet for kumulativ  $N(0,1)$ -fordeling, se tabell.

### Sentralgrenseteoremet

La  $X_1, \dots, X_n$  være  $n$  stokastisk uavhengige variabler, alle med forventning  $\mu$  og standardavvik  $\sigma$ . Da har vi, for tilstrekkelig stor  $n$ , at

$\bar{X}$  er tilnærmet  $N(\mu, \sigma/\sqrt{n})$ -fordelt

(eksakt normalfordeling hvis  $X$ 'ene er normalfordelte, uansett  $n$ -verdi).

### t-fordeling stud( $m$ )

La  $X_1, \dots, X_n$  være uif  $N(\mu, \sigma)$ . Da vil  $(\bar{X} - \mu) / (\sigma/\sqrt{n})$  være t-fordelt med  $m = n-1$  frihetsgrader.

### Tilnærrelse binomisk - normal

Dersom  $X$  er tilnærmet binomisk  $\text{Bin}(n, p)$  med  $np(1-p) \geq 5$  så har vi at

$$P(X \leq b) \approx \Phi\left(\frac{b - np + 0,5}{\sqrt{np(1-p)}}\right)$$

### Tilnærrelse Poisson - normal

$X$  er tilnærmet  $Po(\lambda)$  med  $\lambda \geq 5 \Rightarrow$

$$P(X \leq b) \approx \Phi\left(\frac{b - \lambda + 0,5}{\sqrt{\lambda}}\right)$$

### Kjikvadratfordeling Kji2( $n$ )

$$X \sim \text{Kji2}(n) \Rightarrow E(X) = n, \quad \text{std}(X) = \sqrt{2n}$$

# Kapittel 6

## Estimering

### 6.1 Innledning

For å få en «jordnær» tilnærming til estimeringsbegrepet, kan vi grovt si at å **estimere** vil si det samme som å **anslå**. Si for eksempel at vi ønsker å anslå (estimere) middelverdien av vektene til alle studenter i Norge ved å foreta et tilfeldig utvalg av studenter, f.eks. de som studerer ved Høgskolen i Tromsø. Kaller vi samlingen av alle vektene til studentene i Norge for en populasjon med forventning,  $\mu$ , og standardavvik,  $\sigma$ , så er det altså verdien til (parameteren)  $\mu$  vi ønsker å anslå.

Bemerk at  $\mu$  er en ukjent *konstant* som vi ville funnet dersom vi la sammen alle studentvektene i hele Norge og delte på totalt antall studenter i hele Norge (forventning = «teoretisk middelverdi»). I dette tilfellet ville vi ikke hatt behov for å anslå  $\mu$ , fordi vi ville funnet den eksakte verdien. Hvis vi derimot skal anslå  $\mu$  på basis av et begrenset utvalg fra populasjonen, vil dette anslaget variere fra utvalg til utvalg.

Si at vi ønsker å bruke middelverdien i utvalget,  $Y = \bar{X}$  som anslags-funksjon for  $\mu$ . Middelverdien kalles i dette tilfellet for en **estimator** for  $\mu$ , og estimatoren er en stokastisk variabel. Idet vi har foretatt målingene på et konkret utvalg og funnet middelverdien,  $\bar{x}$ , for utvalget, har vi fått et **estimat** (en anslagsverdi) for den ukjente forventningen,  $\mu$ , til *populasjonen som helhet*. Det er da intuitivt rimelig å si at anslaget (*estimateet*) vil bli sikrere jo større utvalget er. I forrige kapittel så vi at standardavviket til  $\bar{X}$  (som i dette tilfellet er et mål på usikkerhet til middelverdien som anslagsfunksjon) avtok omvendt proporsjonalt med kvadratrota av  $n$  ( $\text{std}(\bar{X}) = \sigma/\sqrt{n}$ ).

Når vi estimerer en forventningsverdi,  $\mu$ , ved middelverdien til et utvalg, får vi det vi kaller et **punktestimat** for  $\mu$ . Dette skyldes at middelverdien bare er en numerisk verdi, f.eks. 68,1 kg, som i seg selv ikke gir noen pekepinn på hvor usikkert anslaget er. Når vi ønsker å gi et anslag for usikkerheten, skal vi foreta noe vi kaller *estimering ved konfidensintervall*. Dette er et vanskelig og abstrakt begrep. Litt forenklet kan vi som et eksempel si at dersom vi beregner intervallet  $(\bar{x} - 2s/\sqrt{n}, \bar{x} + 2s/\sqrt{n})$ , der  $\bar{x}$  og  $s$  er middelverdi og standardavvik til de  $n$   $X$ -ene i utvalget, vil vi få et intervall som vi er «ca. 95 % sikre på» vil inneholde den sanne parameteren,  $\mu$ . Dersom  $n = 35$  ville vi kanskje typisk få intervallet (66,1 ,

70,1) som det vi kaller et tilnærmet 95 % konfidensintervall for  $\mu$ . Bredden av intervallet vil gi en indikasjon på usikkerheten til anslaget, mens senteret i konfidensintervallet gir en indikasjon på «hvor i verden»  $\mu$  ligger.

Middelverdien er ikke den eneste estimatoren vi kan bruke for å anslå forventningen,  $\mu$ . Som vi husker fra beskrivende statistikk, brukte vi også medianen som (robust) mål på senter i fordelingen. Teoretisk faller forventning og median (midtverdi) til en fordeling sammen når fordelingen er symmetrisk, som f.eks. normalfordelingen er. Når vi har valget mellom ulike estimatorer, velger vi den som er best ut ifra visse kriterier. I dette kurset skal vi begrense oss til det vi kaller **forventningsrette** estimatorer (defineres senere). Blant disse skal vi definere den estimatoren som har **minst varians** (standardavvik) som den beste.

Forventningen,  $\mu$ , er ikke den eneste parameteren vi er interessert i å estimere. Vi skal her også se på estimatorer for variansen,  $\sigma^2$ , standardavviket,  $\sigma$ , og den binomiske parameteren,  $p$ . For å skille estimatoren, som er en stokastisk variabel, fra den parameteren (konstanten) som skal estimeres, skal vi ofte bruke superskript \* for å betegne estimatoren:  $\mu^*$  for  $\mu$ -estimatoren,  $\sigma^2^*$  for  $\sigma^2$ -estimatoren og  $p^*$  for  $p$ -estimatoren. Generelt skal vi bruke symbollet  $\theta$  for en vilkårlig ukjent parameter ( $\theta = \mu$ ,  $\sigma^2$  eller  $p$  her), med estimator  $\theta^*$ .

## 6.2 Punktestimering av en parameter

La  $X_1, X_2, \dots, X_n$  utgjøre et tilfeldig utvalg fra en populasjon, der  $\theta$  er en ukjent parameter som beskriver en egenskap til populasjonsfordelingen (eks:  $\theta = \mu$  i normalfordelingen). En *estimator*,  $\theta^*$ , er da en funksjon av de stokastiske variablene i utvalget,  $\theta^* = \theta^*(X_1, X_2, \dots, X_n)$  som blir brukt til å *estimere* (anslå) den ukjente verdien til  $\theta$  på basis av utvalget. En estimator blir derved også en stokastisk variabel med en *estimatorfordeling*,  $f(\theta^*)$ , som generelt er forskjellig fra populasjonsfordelingen,  $f(x)$ . Når vi har satt inn tallene  $x_1, x_2, \dots, x_n$  fra et konkret utvalg i estimatorfunksjonen,  $\theta^*$ , får vi et tall (en anslagsverdi) som vi kaller *et estimat for  $\theta$* .

En estimators egenskaper er beskrevet ved estimatorfordelingens egenskaper. Vi skal i dette kurset betrakte følgende 2 ønskede egenskaper ved en estimator:

- 1) **Forventningsrettethet**, dvs.  $E(\theta^*) = \theta$ .
- 2) **Minst mulig varians**, dvs.  $\text{Var}(\theta^*)$  skal være minst mulig.

Det kan være nyttig å ha følgende tankemodell i hodet: Den ukjente parameteren ville vi funnet dersom antall observasjoner var uendelig stort, eller dersom vi hadde observasjoner fra alle enheter i populasjonen. Når vi i praksis kun kan ta et utvalg, prøver vi på beste måte å anslå generelle egenskaper til den populasjonen utvalget kommer fra, på basis av de observerte verdiene i utvalget.

Dersom vi har en forventningsrett estimator betyr dette følgende: Tenk deg at du foretar en rekke tilfeldige utvalg fra en populasjon. For hvert utvalg (hvert eksperiment) regner du ut et estimat (en anslagsverdi) for den ukjente populasjonsparameteren. Dersom estimatoren du bruker er forventningsrett, betyr det at gjennomsnittet av alle de estimerne du får ved gjentatte eksperimenter nærmer seg mer og mer den sanne verdien jo flere eksperimenter som blir utført.

Selv om en estimator er forventningsrett, så kan et tilfeldig estimat godt ligge langt unna den sanne verdien. Et mål på usikkerheten til en estimator får vi ved å finne standardavviket til estimatoren,  $\text{std}(\theta^*) = \sqrt{\text{Var}(\theta^*)}$ .

### Valg av estimator

Dersom 2 estimatorer,  $\theta_1^*$  og  $\theta_2^*$ , for én og samme parameter,  $\theta$ , er forventningsrette (dvs.  $E(\theta_1^*) = E(\theta_2^*) = \theta$ ), så er den estimatoren best som har *minst varians*, eller ekvivalent, minst standardavvik,  $\text{std}(\theta^*)$

## 6.3 Punktestimering av populasjonsforventningen, $\mu$

Vi skal som tidligere benytte symbolene  $\mu$  og  $\sigma$  til å betegne henholdsvis forventning og standardavvik i populasjonsfordelingen,  $f(x)$ , til en stokastisk variabel,  $X$ . Videre lar vi  $X_1, X_2, \dots, X_n$  betegne et tilfeldig utvalg av  $n$  stokastiske variabler fra denne populasjonsfordelingen, og vi lar  $\bar{X}$  betegne middelverdien til  $X$ -ene:

$$(6.1) \quad \bar{X} = \frac{1}{n} \sum_{i=1}^n X_i = \frac{1}{n} (X_1 + X_2 + \dots + X_n)$$

Fra forrige kapittel har vi at  $E(X) = \mu$ , og middelverdien,  $\bar{X}$ , er følgelig en forventningsrett estimator for  $\mu$ . Fra kap. 5 har vi også at  $\text{std}(\bar{X}) = \sigma/\sqrt{n}$ , dvs. standardavviket til middelverdien er en faktor  $\sqrt{n}$  mindre enn populasjons-

standardavviket,  $\sigma$ . Videre så vi at  $\bar{X}$  var tilnærmet normalfordelt, uansett populasjonsfordeling, bare  $n$  var stor nok (vi brukte som grense  $n \geq 20$ ).

**Eks. 6.1** **Valg av estimator.**

La  $X_1$  og  $X_2$  være 2 stokastisk uavhengige variabler fra én og samme populasjon med ukjent forventning  $\mu$  og varians  $\sigma^2$ . Følgende 2 estimatorer for  $\mu$  er gitt:

$$\mu_1^* = \frac{2X_1 + X_2}{3}, \quad \mu_2^* = \frac{X_1 + X_2}{2}$$

*Oppgave*

Vis at begge estimatorer er forventningsrette og at  $\mu_2^*$  er best.

*Løsningsforslag*

Vi viser først at de 2 estimatorene er forventningsrette:

$$\begin{aligned} E(\mu_1^*) &= E\left(\frac{1}{3}(2X_1 + X_2)\right) = \frac{1}{3}E(2X_1 + X_2) = \frac{1}{3}(2E(X_1) + E(X_2)) \\ &= \frac{1}{3}(2\mu + \mu) = \frac{1}{3} \cdot 3\mu = \mu \end{aligned}$$

$$\begin{aligned} E(\mu_2^*) &= E\left(\frac{1}{2}(X_1 + X_2)\right) = \frac{1}{2}E(X_1 + X_2) = \frac{1}{2}E(X_1 + X_2) \\ &= \frac{1}{2}(\mu + \mu) = \frac{1}{2} \cdot 2\mu = \mu \end{aligned}$$

Begge estimatorer er følgelig forventningsrette.

Så undersøker vi hvilken estimator som har minst varians. Siden  $X_1$  og  $X_2$  er uavhengige, slipper vi å tenke på kovariansen:

$$\begin{aligned} \text{Var}(\mu_1^*) &= \text{Var}\left(\frac{1}{3}(2X_1 + X_2)\right) = \frac{1}{9}\text{Var}(2X_1 + X_2) \\ &= \frac{1}{9}(4\text{Var}(X_1) + \text{Var}(X_2)) = \frac{1}{9}(4\sigma^2 + \sigma^2) = \frac{5}{9}\sigma^2 \end{aligned}$$

$$\begin{aligned} \text{Var}(\mu_2^*) &= \text{Var}\left(\frac{1}{2}(X_1 + X_2)\right) = \frac{1}{4}(\text{Var}(X_1) + \text{Var}(X_2)) \\ &= \frac{1}{4}(\sigma^2 + \sigma^2) = \frac{1}{2}\sigma^2 < \frac{5}{9}\sigma^2 \end{aligned}$$

Siden både  $\mu_1^*$  og  $\mu_2^*$  er forventningsrette estimatorer for  $\mu$ , og fordi  $\text{Var}(\mu_2^*) < \text{Var}(\mu_1^*)$ , så er  $\mu_2^*$  best. ☺

## 6.4 Punktestimering av populasjonsvariansen, $\sigma^2$

La  $X_1, \dots, X_n$  være uif med vilkårlig populasjonsfordeling og ukjent varians,  $\sigma^2$ . Vi skal da bruke  $S^2$ , gitt ved følgende formel, som estimator for  $\sigma^2$ :

$$(6.2) \quad S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2 = \frac{1}{n-1} \left( \sum_{i=1}^n X_i^2 - \frac{1}{n} \left( \sum_{i=1}^n X_i \right)^2 \right)$$

Vi skal ikke bevise det her, men dere har forutsetninger til selv å vise at  $E(S^2) = \sigma^2$ , dvs.  $S^2$  er en forventningsrett estimator for  $\sigma^2$ . Dette er grunnen til at vi bruker  $n-1$  og ikke  $n$  i nevneren foran summen, ellers ville ikke estimatoren for  $\sigma^2$  vært forventningsrett.

Dersom populasjonsfordelingen er normal  $N(\mu, \sigma)$ , har vi tidligere sett at  $S^2$ -fordelingen er nært knyttet til  $Kj_2(n-1)$ -fordelingen. Det kan i dette tilfellet også vises at  $S^2$ -estimatoren er den beste blant alle forventningsrette estimatorer for  $\sigma^2$ . Bemerk imidlertid at selv om  $S^2$  er en forventningsrett estimator for  $\sigma^2$ , så følger det *ikke* at  $S$  er en forventningsrett estimator for  $\sigma$ . Det er likevel vanlig å bruke  $S$  som estimator for  $\sigma$ , og forventningsskeivheten er liten for rimelig store  $n$ .

## 6.5 Punktestimering av binomisk parameter, $p$

$p$  er andelen i en populasjon som har en viss egenskap. Intuisjonen tilsier da at estimatoren,  $p^*$ , gitt som følger, er en fornuftig estimator for  $p$ :

$$(6.3) \quad p^* = \frac{X}{n}$$

der  $X$  er antall som har denne egenskapen i et tilfeldig utvalg på  $n$ . Dersom  $n$  er svært liten i forhold til populasjonsstørrelsen, er det rimelig å betrakte de  $n$  «trekningene uten tilbakelegging» som  $n$  uavhengige Bernoulli-forsøk.  $X$  er da tilnærmet binomisk fordelt  $Bino(n, p)$ , med forventning  $np$  og varians  $np(1-p)$ .  $p^*$  får da følgende egenskaper:

$$E(p^*) = E\left(\frac{1}{n} X\right) = \frac{1}{n} E(X) = \frac{1}{n} \cdot np = p$$

$$\text{Var}(p^*) = \text{Var}\left(\frac{1}{n} X\right) = \frac{1}{n^2} \text{Var}(X) = \frac{1}{n^2} \cdot np(1-p) = \frac{1}{n} p(1-p)$$

$$\Rightarrow \text{std}(p^*) = \sqrt{\frac{1}{n} p(1-p)}$$

Fra første ligning ovenfor ser vi at  $p^* = X/n$  er en forventningsrett estimator for  $p$ . Videre ser vi at standardavviket til  $p$ -estimatoren går som 1 over  $\sqrt{n}$ , og er dessuten avhengig av  $p$  (størst for  $p = .5$ ).

**Eks. 6.2**  $p^*$ -estimatoren

30 av 100 terningkast med «urettferdig» terning gir sekser. La  $p$  være sannsynligheten for sekser i ett tilfeldig kast.

*Oppgave*

Estimer  $p$ , og finn en tilnærmet verdi for  $\text{std}(p^*)$ .

*Løsningsforslag*

$$p^* = 30/100 = 0.3, \quad \text{std}(p^*) \approx \sqrt{p^*(1-p^*)/n} = \sqrt{0.3 \cdot 0.7 / 100} = 0.021$$

Her bestod tilnærmelsen i å sette inn  $p^*$  istedet for  $p$  i uttrykket for  $\text{std}(p^*)$ , siden vi jo ikke kjenner den ukjente  $p$ -verdien. ☺

## 6.6 Estimering ved konfidensintervall (KI)

Et punktestimat (se forrige avsnitt) gir som ordet indikerer en anslagsverdi for en ukjent parameter. Punktestimatet sier imidlertid ingenting om hvor *sikkert* anslaget er. Det gjør derimot et konfidensintervall, som vi nå skal se på.

### 100(1– $\alpha$ ) % konfidensintervall (definisjon)

La  $(1-\alpha)$  være en spesifisert sannsynlighet i nærheten av 1, og la  $L$  (for *Lower*) og  $U$  (for *Upper*) være funksjoner av  $X_1, X_2, \dots, X_n$ , slik at

$$(6.4) \quad P(L < \theta < U) = 1 - \alpha$$

Da kalles  $(L, U)$  for et 100(1– $\alpha$ ) % konfidensintervall for parameteren  $\theta$ , og  $(1-\alpha)$  kalles konfidensnivået tilknyttet intervallet. Vi skal benytte forkortelsen KI for konfidensintervall.

Konfidensintervall er et abstrakt og vanskelig begrep, selv om selve regnereglene for å beregne et konfidensintervall er overkomelige. Ikke fortvil om du synes definisjonen i forrige ramme til å begynne med kan virke noe abstrakt.

I dette kurset skal vi kun se på konfidensintervall for  $\mu$  (populasjonsforventning) og  $p$  (binomisk fordeling), dvs. konfidensintervall for  $\theta = \mu$  og  $\theta = p$ . For  $\mu$  skal vi skille mellom store ( $n \geq 30$ ) og små ( $n < 30$ ) utvalg. For store utvalg kan vi erstatte en ukjent  $\sigma$ -verdi med  $S$ , og forøvrig bruke resultatet at middelverdien er tilnærmet normalfordelt uansett populasjonsfordeling. For små utvalg skal vi benytte  $t$ -fordelingen, som tar hensyn til at vi ikke kjenner  $\sigma$ -verdien.

Når det gjelder konfidensintervall kan det være nyttig med følgende tankemodell: Dersom vi foretar en rekke gjentatte forsøk, hver gang med nye tilfeldige utvalg, kan vi konstruere en rekke konfidensintervall. Dersom vi konstruerer f.eks. 95 % konfidensintervaller, betyr det at 95 % av konfidensintervallene vil dekke den sanne, men ukjente parameteren, eller sagt på en annen måte: I 95 % av tilfellene vil den sanne parameterverdien ligge i intervallet.

For å belyse konfidensintervall-konseptet skal vi først se på et eksempel der vi ønsker å estimere populasjonsforventningen,  $\mu$ , på basis av middelverdien,  $\bar{X}$ . Vi antar at  $\bar{X}$  er normal  $N(\mu, \sigma/\sqrt{n})$  med kjent populasjons-standardavvik,  $\sigma$ . Da vet vi fra tidligere at  $Z = (\bar{X} - \mu)/(\sigma/\sqrt{n})$  er  $N(0,1)$ . Fra normalfordelingstabellen husker vi da (kanskje) at 95 % av sannsynlighets-massen lå i intervallet  $0 \mp 1,96$ :

$$\begin{aligned}
 P(-1.96 < Z < 1.96) &= .95 = 1 - .05 \\
 \Leftrightarrow P\left(-1.96 < \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < 1.96\right) &= 1 - .05 \\
 \Leftrightarrow P\left(-1.96 \cdot \frac{\sigma}{\sqrt{n}} < \bar{X} - \mu < 1.96 \cdot \frac{\sigma}{\sqrt{n}}\right) &= 1 - .05 \\
 \Leftrightarrow P\left(-\bar{X} - 1.96 \cdot \frac{\sigma}{\sqrt{n}} < -\mu < -\bar{X} + 1.96 \cdot \frac{\sigma}{\sqrt{n}}\right) &= 1 - .05 \\
 \Leftrightarrow P\left(\bar{X} + 1.96 \cdot \frac{\sigma}{\sqrt{n}} > \mu > \bar{X} - 1.96 \cdot \frac{\sigma}{\sqrt{n}}\right) &= 1 - .05 \\
 \Leftrightarrow P\left(\bar{X} - 1.96 \cdot \frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + 1.96 \cdot \frac{\sigma}{\sqrt{n}}\right) &= 1 - .05
 \end{aligned}$$

Vi ser nå at vi har fått ligninga på den formen som er oppgitt i ramma for definisjonen av et konfidensintervall, med  $\alpha = 0.05$ , den ukjente parameteren  $\mu$  «i midten» av den doble ulikheten, og med nedre grense,  $L$ , og øvre grense,  $U$ , gitt ved uttrykkene:

$$(6.5) \quad L = \bar{X} - 1.96 \cdot \frac{\sigma}{\sqrt{n}}, \quad U = \bar{X} + 1.96 \cdot \frac{\sigma}{\sqrt{n}}$$

Bemerk at  $L$  og  $U$  er stokastiske variabler som ikke inneholder ukjente parametre ( $\sigma$  er antatt kjent i eksemplet ovenfor), fordi  $\bar{X}$  er en stokastisk variabel. Konfidensintervallet  $(L, U)$  blir derved også en stokastisk variabel.

La oss nå se på et talleksempel. La  $\sigma = 22.8$ ,  $n = 100$  og si at vi ved et eksperiment finner at  $\bar{x} = 97$ . På basis av formlene ovenfor finner vi at  $L = 97 -$

$1.96 \cdot 22.8/10 = 92.5$ , og  $U = 97 + 1.96 \cdot 22.8/10 = 101.5$ , dvs.  $(L, U) = (92.5, 101.5)$  blir et 95 % KI for  $\mu$ . Resultatet kan formuleres som følger:

«Et 95 % konfidensintervall for  $\mu$  er  $(92.5, 101.5)$  beregnet fra det observerte utvalg på  $n = 100$  observasjoner.»

I eksemplet vi gikk igjennom var  $z$ -verdien  $z_{0.025} = 1.96$  lik det vi kaller *øvre 2.5 %-fraktile* i  $N(0,1)$  fordelingen, fordi 2.5 % av sannsynlighetsmassen ligger til høyre for  $z = 1.96$ . Grunnen til at vi får akkurat 2.5 %-fraktilen er at vi valgte å konstruere et symmetrisk 95 % konfidensintervall, og da blir det 2.5 % igjen «på hver side». I det generelle tilfellet konstruerer vi et  $100(1-\alpha)$  % intervall der ikke nødvendigvis  $\alpha = 0.05$ . Vi erstatter da 1,96 med  $z_{\alpha/2}$ . Betydningen av denne betegnelsen er skissert i fig. 6.1. Tab. 6.1 viser forøvrig forskjellige  $z_{\alpha/2}$ -verdier.

Tab. 6.1  $z_{\alpha/2}$ -verdier

$1-\alpha$	.80	.85	.90	.95	.99
$z_{\alpha/2}$	1.28	1.44	1.64	1.96	2.58

NB! Vær klar over at når du har beregnet et konkret konfidensintervall som inneholder 2 tall, ett for nedre intervallgrense, og ett for øvre grense, så har det *ingen mening* å snakke om sannsynligheten for at den sanne verdien  $\mu$  ligger i dette intervallet:  $\mu$  er en konstant som enten ligger i et konkret intervall med 100 % sannsynlighet, eller så ligger  $\mu$  *ikke* i intervallet med sannsynlighet 100 %. Det vi derimot kan si, er at i det lange løp vil 95 % av alle 95 % konfidensintervaller dekke (inneholde) den sanne parameteren,  $\mu$ . Det er *konfidensintervallet*, gitt ved dets grenser  $L$  og  $U$ , som er *stokastisk*, ikke parameteren  $\mu$ , som er en konstant. Denne forskjellen er det mange studenter som har problemer med, hvis det kan være noen trøst. Vi skal belyse dette ytterligere ved eksempler.

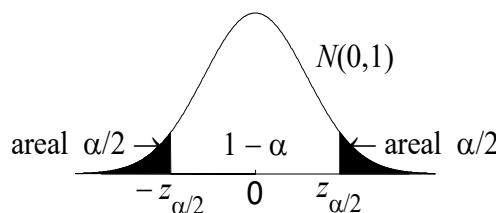


Fig. 6.1 Illustrasjon av betegnelsen  $z_{\alpha/2}$  for øvre  $\alpha/2$ -fraktile i  $N(0,1)$  fordelingen.

For å øke forståelsen av begrepet konfidensintervall, kan vi nå tenke oss at vi konstruerer stadig nye 95 % konfidensintervall basert på nye tilfeldige utvalg på  $n = 100$  fra samme populasjon. For hvert utvalg får vi da et nytt konfidensintervall, og i det lange løpet vil 95 % av disse intervallene inneholde, eller dekke, den sanne parameteren,  $\mu$ .

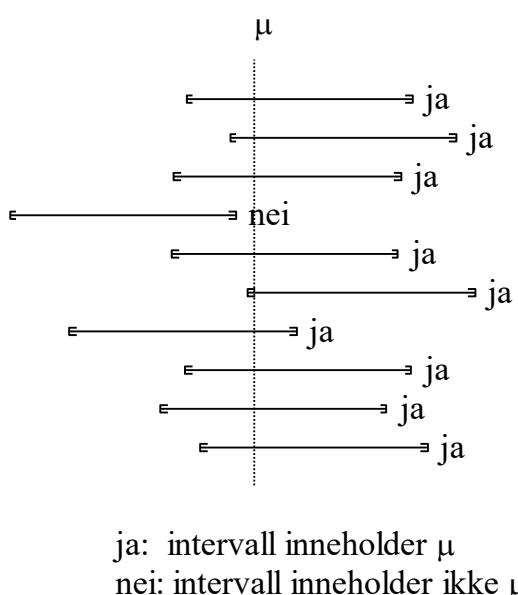


Fig. 6.2 Tolkning av konfidensintervall for  $\mu$ , illustrert ved 10 forskjellige konfidensintervall basert på 10 forskjellige utvalg.

Vi ser at 9 av 10 tilfeldige konfidensintervall inneholder (dekker) den «sanne» og ukjente  $\mu$ . Hvis dette er en trend som holder seg, dvs. ca. 90 % av svært mange tilfeldige konfidensintervall for  $\mu$  inneholder  $\mu$ , så har vi konstruert en rekke 90 % konfidensintervaller for  $\mu$ .

## 6.7 KI for $\mu$ , små utvalg, $\sigma$ kjent

Siden vi forutsetter at  $n \downarrow 20$ , kan vi anvende sentralgrenseteoremet, og anta at  $\bar{X}$  er  $N(\mu, \sigma/\sqrt{n})$  med kjent  $\sigma$ . Da kan vi, slik skissert i forrige avsnitt, beregne et  $100(1-\alpha)$  % konfidensintervall for  $\mu$  slik angitt i ramma nedenfor. Legg merke til at vi får et *eksakt*  $100(1-\alpha)$  % KI for  $\mu$  dersom  $X$ -ene er normalfordelte, uavhengig av  $n$ .

### KI for $\mu$ når $n \geq 20$ , $\sigma$ kjent

La  $\bar{X}$  betegne middelverdien av et tilfeldig utvalg på  $n$  uavhengige observasjoner fra en populasjon med populasjons-forventning  $\mu$  og populasjons-standardavvik  $\sigma$ . For  $n \geq 20$  er da et tilnærmet  $100(1-\alpha)$  % konfidensintervall for  $\mu$  gitt ved

$$(6.6) \quad \left( \bar{X} - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}, \bar{X} + z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} \right)$$

der  $z_{\alpha/2}$  er øvre  $\alpha/2$ -fraktil i  $N(0,1)$ -fordelingen. For normalfordelte  $X$ -er er intervallet et eksakt  $100(1-\alpha)$  % intervall for  $\mu$ , uavhengig av  $n$ .

**Eks. 6.3**

**KI for  $\mu$ .** Vi betrakter et tilfeldig utvalg på  $n = 100$  observasjoner fra en populasjon der  $\mu$  er ukjent og  $\sigma = 8$ . Anta at middelverdien i utvalget er funnet å være  $\bar{x} = 42.7$ .

*Oppgave*

Konstruer et 95 % konfidensintervall for  $\mu$ .

*Løsningsforslag*

Her er  $n$  så stor, at  $\bar{X}$  er normalfordelt med god tilnærmelse. Vi bruker formelverket i siste ramme og får følgende 95 % konfidensintervall for  $\mu$  (husk at 95 % KI betyr at  $\alpha = 0.05$  og  $\alpha/2 = 0.025$ , slik at vi skal bruke  $z_{0.025} = 1.96$ ):

$$\left( 42.7 - 1.96 \cdot \frac{8}{\sqrt{100}}, 42.7 + 1.96 \cdot \frac{8}{\sqrt{100}} \right) = (41.1, 44.3)$$

Resultatet kan formuleres som følger: «Et 95 % konfidensintervall for  $\mu$  er  $(41.1, 44.3)$  beregnet fra det observerte utvalg bestående av  $n = 100$  observasjoner.» ☺

## 6.8 KI for $\mu$ , store utvalg

Når  $n \downarrow 20$  sa vi at middelverdien var tilnærmet normalfordelt  $N(\mu, \sigma/\sqrt{n})$ . Når  $\sigma$  er ukjent erstatter vi denne med  $S$ . Da er imidlertid grensa  $n \geq 20$  litt snau, og vi setter grensa ved  $n = 30$  for bruk av følgende formel for et tilnærmet  $100(1-\alpha)$  % KI for  $\mu$ :

### KI for $\mu$ når $n \geq 30$ , $\sigma$ ukjent

Et tilnærmet  $100(1-\alpha)\%$  konfidensintervall for populasjons-forventningen,  $\mu$ , når  $\sigma$  er ukjent, er gitt ved:

$$(6.7) \quad \left( \bar{X} - z_{\alpha/2} \cdot \frac{S}{\sqrt{n}}, \bar{X} + z_{\alpha/2} \cdot \frac{S}{\sqrt{n}} \right)$$

der  $z_{\alpha/2}$  = øvre  $\alpha/2$ -fraktil i standard normalfordeling, og det er forutsatt *stort utvalg* ( $n \geq 30$ ).

Øvre  $\alpha/2$ -fraktil i normalfordelingen finnes fra tab. 6.1, oppgitt tidligere, eller ved bruk av *t-fordelings-tabellen* med uendelig antall frihetsgrader (d.f. =  $\infty$ ). Bruk av *t-fordelingstabell* er forklart nærmere i kap. 6.10.

Vi henviser til eks. 6.3 forrige side for utførelse i praksis. Eneste forskjell fra eks. 6.3 er at vi må bruke estimert standardavvik,  $s$ , på basis av utvalget, i stedet for  $\sigma$ , siden  $\sigma$  er ukjent.

## 6.9 KI for $p$ , store utvalg

Vi har tidligere sett at den binomiske fordelingen, med parametre  $n$  og  $p$ , var tilnærmet lik normalfordelingen for tilstrekkelig store verdier for  $np(1-p)$ . Vi satte grensa  $np(1-p) \geq 5$  for at normaltilnærmelsen kunne brukes. Merk imidlertid at vi får et problem når det gjelder konfidensintervall, fordi den ukjente  $p$ -verdien også inngår i uttrykket for variansen,  $np(1-p)$ . Vi må derfor i tillegg kreve store utvalg, dvs.  $n \geq 30$ , slik at vi kan erstatte  $p$  med  $p^*$  i uttrykket for standardavviket. Basert på normaltilnærmelsen får vi følgende uttrykk for et tilnærmet  $100(1-\alpha)\%$  konfidensintervall for  $p$ :

### KI for binomisk $p$ , $n \geq 30$

Forutsetning:  $np(1-p) \geq 5$

Et tilnærmet  $100(1-\alpha)\%$  konfidensintervall for parameteren  $p$  i den binomiske fordeling er gitt ved uttrykket

$$(6.8) \quad \left( p^* - z_{\alpha/2} \cdot \sqrt{\frac{p^*(1-p^*)}{n}}, p^* + z_{\alpha/2} \cdot \sqrt{\frac{p^*(1-p^*)}{n}} \right)$$

der  $z_{\alpha/2}$  er øvre  $\alpha/2$ -fraktil i standard normalfordeling  $N(0,1)$ , og  $p^* = X/n$ , der  $X$  er antall «sukssesser» ( $J$ -hendelser) ved  $n$  Bernoulli-forsøk.

**Eks. 6.4** I en **politisk Gallup** anslås partioppslutningen til de forskjellige politiske partier på basis av et tilfeldig utvalg på  $n = 1600$  personer.

#### Oppgave

Anslå et 95 % konfidensintervall for et parti som 160 av de 1600 som blir spurtsvarer at de vil stemme på.

#### Løsningsforslag

Her er det rimelig å si at vi har  $n = 1600$  Bernoulli-forsøk, der sannsynligheten er  $p$  for at hver person som spørres skal svare at de vil stemme på det parti vi ser på. Grunnen til dette er at  $n = 1600$  er forsvinnende lite i forhold til totalt antall stemmeberettigede, så det at vi har «trekning uten tilbakelegging» får minimal betydning. I oppgaven er det oppgitt at  $X = 160$ , og vi får da punktestimatet  $p^* = X/n = 160/1600 = 10.0\%$ . Siden  $np^*(1-p^*) = 1600 \cdot 0.1 \cdot 0.9 = 144 \gg 5$ , er det rimelig å anta at forutsetningene for å benytte formelen i forrige ramme er tilfredsstilt. Vi får da følgende 95 % KI for  $p$ :

$$\left( 0.1 - 1.96 \sqrt{\frac{0.1 \cdot 0.9}{1600}}, 0.1 + 1.96 \sqrt{\frac{0.1 \cdot 0.9}{1600}} \right) = (8.5\%, 11.5\%)$$

### 6.10 KI for $\mu$ , små utvalg, $\sigma$ ukjent

I dette tilfellet skal vi anta at tilnærmelsen blir for grov hvis vi erstatter den ukjente verdien for  $\sigma$  med estimatet  $s$  og benytter normalfordelingsfraktiler. Vi skal da bruke  $t$ -fordelingen (se kap. 5).

Bakerst i kompendiet er gjengitt en  $t$ -fordelingstabell. Bruken av denne er «omvendt» av bruken av kumulativ  $N(0,1)$ -tabell, som det vil fremgå av neste eksempel. Fremgangsmåten for å konstruere et konfidensintervall for  $\mu$  er helt analog med tilfellet for store utvalg. Eneste forskjell er at vi må huske å erstatte  $N(0,1)$ -fraktilene  $z_{\alpha/2}$  med de noe større  $t_{\alpha/2}$ -fraktilene fra  $t$ -fordelingen med  $n-1$  frihetsgrader. Vi får følgende uttrykk for et  $100(1-\alpha) \%$  konfidensintervall for  $\mu$ :

### KI for $\mu$ med ukjent $\sigma$ . Små utvalg

Forutsetninger:  $X_1, \dots, X_n$  uif  $N(\mu, \sigma)$ -variabler.

Hvis populasjonen har populasjonsfordeling  $N(\mu, \sigma)$ , der  $\sigma$  er ukjent, er et  $100(1-\alpha) \%$  konfidensintervall for  $\mu$  gitt ved

$$(6.9) \quad \left( \bar{X} - t_{\alpha/2} \cdot \frac{S}{\sqrt{n}}, \bar{X} + t_{\alpha/2} \cdot \frac{S}{\sqrt{n}} \right)$$

der  $t_{\alpha/2}$  er øvre  $\alpha/2$ -fraktil i  $t$ -fordelingen med  $n-1$  frihetsgrader (d.f. =  $n-1$ ), og  $\bar{X}$  er middelverdien av  $n$  stokastisk uavhengige  $N(\mu, \sigma)$ -variabler.

**Eks. 6.5** **Ny legering.** En ny legering er blitt foreslått brukt i en ny satellitt. Strekkstyrkemålinger er utført på 15 tilfeldig utvalgte stykker av legeringen. Middelverdi og standardavvik til de 15 stykkene er funnet å være henholdsvis 39.3 og 2.6, og sannsynlighetsfordelingen til strekkstyrke er rimelig symmetrisk.

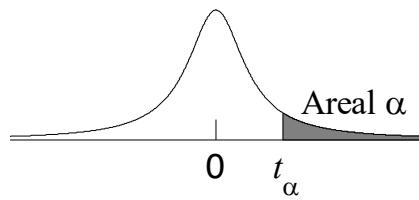
#### Oppgave

Finn et 90 % konfidensintervall for forventet strekkstyrke til legeringen.

#### Løsningsforslag

Selv om  $n = 15 < 20$  antar vi at  $\bar{X}$  er  $N(\mu, \sigma/\sqrt{n})$ , fordi det er oppgitt at sannsynlighetsfordelingen av strekkstyrke er rimelig symmetrisk. Siden  $\sigma$  er ukjent, bruker vi  $t$ -fordelingen med  $n-1 = 14$  frihetsgrader. Siden vi skal beregne et 90 % konfidensintervall, så er  $\alpha = 0.1$  og  $\alpha/2 = 0.05$ . Ved å slå opp i  $t$ -fordelingstabellen finner vi da at  $t_{0.05} = 1.761$ .

**t-fraktiler**,  $t_\alpha$ , i  
 $t$ -fordelingen med  $m$  frihetsgrader



$m$	$\alpha$					
	.25	.10	.05	.025	.01	.005
⋮	⋮	⋮	⋮	⋮	⋮	⋮
14	.692	1.345	1.761	2.145	2.624	2.977
15	.691	1.341	1.753	2.131	2.602	2.947

Fig. 6.4 Utsnitt av t-fraktiltabell bakerst i boka.

Vi setter så inn i formelen i forrige ramme ( $m = 14$ ) og får følgende 90 % konfidensintervall for  $\mu$ :

$$\left( 39.3 - 1.761 \cdot \frac{2.6}{\sqrt{15}}, 39.3 + 1.761 \cdot \frac{2.6}{\sqrt{15}} \right) = (38.12, 40.48) \quad \text{⊗}$$

Konfidensintervallet kan tolkes som følger:

Dersom forventet strekkstyrke var 38.12 eller lavere, ville det være maksimalt 5 % sannsynlighet for å oppnå den empiriske middelverdien 39.3 eller høyere verdier. Omvendt, dersom forventet strekkstyrke var 40.48 eller høyere ville det vært maksimalt 5 % sannsynlighet for å få verdien 39.3 eller lavere verdier.

## 6.11 Oppgaver

**6.1** La  $X_1$  og  $X_2$  være 2 stokastisk uavhengige variabler fra en og samme populasjon med populasjonsforventning  $\mu$  og populasjons-standardavvik  $\sigma$ .

Du skal velge en av følgende 2 estimatorer for  $\mu$ :

$$\mu_1^* = \frac{3X_1 + 7X_2}{10}, \quad \mu_2^* = \frac{X_1 + X_2}{2}$$

Vis at begge estimatorer er forventningsrette, og bestem hvilken som er best.

**6.2** La  $X_1, X_2, X_3, X_4, Y_1, Y_2$  og  $Y_3$  være 7 uavhengige stokastiske variabler med forventning  $\mu$  og standardavvik  $\sigma$ . Betrakt følgende 2 estimatorer for  $\mu$ :

$$\mu_1^* = \frac{1}{2}(\bar{X} + \bar{Y}), \quad \mu_2^* = \frac{1}{7}(4\bar{X} + 3\bar{Y})$$

der  $\bar{X}$  og  $\bar{Y}$  er middelverdiene til henholdsvis  $X$ -ene og  $Y$ -ene. Vis at begge estimatorene er forventningsrette, og at  $\mu_2^*$  er best.

**6.3** La  $\bar{X}$  være  $N(\mu, \sigma/\sqrt{n})$ .

- a) Bestem et 95 % KI for  $\mu$  når  $\sigma = 2.7$  og  $n = 16$  på basis av et utvalgs-middel på  $\bar{x} = 0.2$ .
- b) Bestem et 99 % KI for  $\mu$  når  $\sigma$  er ukjent, på basis av følgende tilfeldige  $x$ -verdier:

12.7	14.1	9.8	10.2	10.9
7.8	11.3	8.9	9.7	

- c) En medisin har ukjent helbredelses-rate,  $p$ . Bestem et 95 % KI for  $p$

basert på en undersøkelse der 72 av 100 ble helbredet. Gjør rede for de forutsetninger du gjør.

### 6.4

a) Ved en politisk gallup svarer  $x = 40$  av  $n = 1600$  at de vil stemme på et bestemt politisk parti. Beregn et tilnærmet 95 % konfidensintervall for partiets oppslutning.

b) Antall biler som passerer mellom kl. 12.00 og kl. 12.15 i en tett trafikkert envegskjøring antas å være Poissonfordelt med ukjent parameter  $\lambda$ . Ved opptelling en dag passerer det 127 biler i det aktuelle tidsrommet. Bestem på basis av denne tellingen et tilnærmet 95 % KI for  $\lambda$ .

**6.5** Et skjenkested serverer halvlitere. La  $X$  betegne det nøyaktige innholdet (volumet) i en tilfeldig halvliter. Anta at  $\text{std}(X) = 1 \text{ cl} (= 0.01 \text{ l})$  og at  $E(X) = \mu$  er ukjent.

- a) Hvor mange halvlitere må du undersøke for at bredden på et 95 % KI for  $\mu$  skal være på 1 cl?
- b) Bestem et 95 % KI for  $\mu$  på basis av en undersøkelse med middelverdi  $\bar{x} = 0.502 \text{ l}$ , og der  $n =$  det antall du fant under a).

**6.6** I en by er det  $N$  drosjer, og drosjene er nummerert fra 1 til  $N$  med nummerskilt på taket. En tilreisende står på byens flyplass og venter på drosje. Han bruker tida til å observere taknummeret til de ankommande

drosjer, og registrerer følgende nummer:

1 58 21 5 14

På basis av tallene ovenfor anslår han antallet drosjer i byen til å være 69. Han forteller dette til drosjesjåføren, som blir mektig imponert, fordi det nettopp er 69 drosjer i byen.

- a) La  $X$  betegne et tilfeldig drosjenummer. Angi populasjon og populasjonsfordeling,  $f(x)$ , i dette tilfellet.

Det kan vises at forventningen til den største av  $n$  tilfeldige nummer,  $X_{(n)} = X_{\text{maks}}$ , er gitt ved følgende uttrykk:

$$E(X_{(n)}) = \frac{n}{n+1}(N+1)$$

- b) Benytt opplysningen ovenfor til å konstruere en forventningsrett estimator,  $N^*$ , for  $N$ , som funksjon av  $n$  og  $X_{(n)}$ . Undersök om du får samme estimat som den tilreisende når du anvender din estimator på tallmaterialet til den tilreisende.

**6.7 (E)** Seks målinger ble utført for å bestemme en konstant  $m$ . Målingene kan antas å være observasjoner av en normalfordelt stokastisk variabel med forventning  $m$  og et kjent standardavvik  $s$ . Et 95 % konfidensintervall for  $m$  blir funnet til (6.15, 6.25).

- a) Finn  $s$ .

Hvor mange flere målinger må utføres for å få et konfidensintervall som har:

- b) konfidensgrad på minst 98 % og samme bredde?  
c) konfidensgrad på minst 95 % og halvparten av bredden?

**6.8 (E)** En potetmelfabrikk kjøper poteter levert i standardsekker. Av forskjellige grunner vil vekten av en sekk poteter variere fra sekk til sekk.

Vi antar at vekten av en tilfeldig valgt sekk poteter kan betraktes som normalfordelt  $N(\mu, \sigma)$ . Videre antar vi at vektene til forskjellige sekker er uavhengige.

Anta at  $\mu = 50.5$  kg og  $\sigma = 0.9$  kg.

- a) Hvor stor er sannsynligheten for at en sekk veier mindre enn 50 kg?  
b) Hvor stor er sannsynligheten for at totalvekten av 25 sekker overstiger 1255 kg?  
c) Hvor stor er sannsynligheten for at minst en av tre tilfeldig valgte sekker skal veie mindre enn 50 kg?

Anta at  $\mu$  er ukjent og at  $\sigma = 0.9$  kg. Veiing av 9 tilfeldige sekker gav følgende vekter i kg:

50.9 48.8 50.6 51.2 51.4 50.7 49.8  
49.2 51.4

- d) Bestem et 95 % konfidensintervall for  $\mu$ .  
e) Bestem også intervallet vi får der som  $\sigma$  er ukjent.  
f) Hvor mange målinger måtte vi minst gjort dersom  $\sigma = 0.9$  antas kjent og vi ville ha et 95 % konfidensintervall med en lengde mindre enn 0.9 kg?

**6.9 (E)** Omfattende målinger av surhetsgraden i en innsjø en tid tilbake, gav grunnlag for å anta at pH-verdien  $X$  for en enkeltmåling etter denne målemetoden er normalfordelt  $N(5,1)$ .

- a) Finn sannsynligheten for at gjenomsnittet  $\bar{X} = \frac{1}{9} \sum x_i$  av 9 uavhengige målinger er mindre enn 4.5.

En tid etter ble det observert et større antall døde fisk, og det ble besluttet å måle pH-verdien på nytt etter en annen målemetode. Resultatet av en enkeltmåling antas å være normalfordelt  $N(\mu, \sigma)$ , hvor  $\mu$  er «sann pH-verdi». 9 uavhengige målinger av pH-verdien ga verdiene:

4.3 3.7 4.6 3.5 3.9 4.0 3.8 4.5 3.7

- b) Estimer både forventet pH-verdi og standardavvik for en enkeltmåling ved å ta utgangspunkt i forventningsrette estimatorer for  $\mu$  og  $\sigma^2$ , og i observasjonsmaterialet ovenfor.
- c) Bestem et 90 % konfidensintervall for «sann pH-verdi» på grunnlag av de estimerte verdiene i punkt b).

**6.10 (E)** I en stor by blir det utført en meningsmåling for å undersøke om innbyggerne vil ha kristen formålsparagraf i byens barnehager. En velger tilfeldig  $n = 1225$  innbyggere og lar  $X$  være antallet blant disse, betraktet som stokastisk variabel, som vil ha en slik formålsparagraf. Undersøkelsen resulterte i verdien  $x = 735$ . La  $p$  være andelen av innbyggerne i byen som ønsker kristen formålsparagraf og velg  $p^* = X/n$  som estimator for  $p$ .

- a) Vis at  $p^*$  er forventningsrett og bestem dens varians. Begrunn at  $p^*$  er tilnærmet normalfordelt.
- b) Betrakt  $\text{Var}(p^*)$  som en funksjon i  $p$  og finn dens maksimum. Vis at  $\text{std}(p^*) \leq \frac{1}{2\sqrt{n}}$

- c) Vis at et 95 % konfidensintervall for  $p$  (tilnærmet) er gitt ved  $(p^* - 1,96 \cdot \text{std}(p^*), p^* + 1,96 \cdot \text{std}(p^*))$  og bestem et tn. 95 % KI for  $p$ .

**6.11 (E)** I et fylke er det 3 videregående skoler. Det har vært endel langtidsfravær (14 dager eller mer) blant lærerne en tid, og ledelsen vil av budsjettmessige årsaker anslå sannsynligheten  $p$  for at en tilfeldig valgt lærer vil ha minst ett langtidsfravær i løpet av en 1-årsperiode. Sannsynligheten  $p$  antas å være den samme for alle lærerne, uansett skole, og fravær for en lærer er uavhengig av fravær for en annen lærer.

Anta at det i løpet av 1 år er  $n_i$  lærere ved skole nr  $i$  ( $i = 1, 2, 3$ ), hvorav et antall på  $X_i$  lærere har minst ett langtidsfravær.  $X = \sum X_i$  blir da antall lærere ved alle skolene tilsammen som har hatt slikt fravær i perioden.

- a) Bestem fordelingen til  $X$  og til hver  $X_i$ . Skriv opp forventningen og variansen til disse.

To estimatorer er foreslått:

$$p_1^* = X/n, \text{ der } n = n_1 + n_2 + n_3 \text{ og}$$

$$p_2^* = \frac{1}{3} \sum_{i=1}^3 \frac{X_i}{n_i}$$

- b) Vis at begge estimatorene er forventningsrette og bestem  $\text{Var}(p_1^*)$ .

Vis at

$$\text{Var}(p_2^*) = \frac{p(1-p)}{9} \sum_{i=1}^3 \frac{1}{n_i}$$

For et bestemt år har vi følgende data:

$n_i$	30	40	60
-------	----	----	----

$$X_i \quad | \quad 3 \quad | \quad 3 \quad | \quad 5$$

c) Hvilken estimator vil du foretrekke?

Rapporter tilbake estimat av  $p$  ± estimert standardavvik av estimator.

**6.12 (E)** Vi skal i denne oppgaven se på behandling med et bestemt medikament på en gruppe av  $n$  pasienter. La  $X_i$  være en stokastisk variabel, slik at  $X_i=1$  dersom pasient  $i$  blir helbredet og  $X_i=0$  dersom pasient  $i$  ikke blir helbredet ( $i=1,\dots,n$ ). Anta videre at  $P(X_i=1)=p$ ,  $P(X_i=0)=1-p$  og at pasientene helbredes uavhengig av hverandre.

- a) Beregn forventning og varians for  $X_i$ .
- b) Forklar hva den stokastiske variabelen  $X=X_1+X_2+\dots+X_n$  beskriver.

Hvilken fordeling har  $X$ ?

- c) Vis at  $\bar{X}=\frac{1}{n}(X_1+X_2+\dots+X_n)$  er en forventningsrett estimator for  $p$ .

d) Begrunn at  $\bar{X}$  tn.  $N\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$

dersom  $n$  er «tilstrekkelig stor».

- e) På grunnlag av tidligere erfaring vet vi at  $p=0.6$ . 1000 pasienter blir behandlet. Hvor stor er sannsynligheten (tilnærmet) for at minst 630 personer blir helbredet?

**6.13 (E)** Ventetiden  $T$  (målt i minutter) etter tilfeldig ankomst til en viss bussholdeplass antas å være uniformt fordelt med sannsynlighetstetthet

$$f(t) = \begin{cases} 1/\theta, & 0 < t < \theta \\ 0, & \text{ellers} \end{cases}$$

der  $\theta$  er en ukjent parameter (konstant).

- a) Hva beskriver parameteren  $\theta$  i ventetidssituasjonen?  
Bestem  $P(T < \theta/4)$  og  $P(T = \theta/2)$ .
- b) Vis at forventning og varians til  $T$  er gitt ved  $E(T) = \theta/2$  og  $\text{Var}(T) = \theta^2/12$ .

La  $T_1, \dots, T_n$  være  $n$  uavhengige observasjoner med samme fordeling som  $T$ .

c) Vis at  $\theta^* = \frac{2}{n} \sum_{i=1}^n T_i$

er en forventningsrett estimator for  $\theta$ . Bestem variansen til  $\theta^*$ .

La  $T_{\max}$  betegne den største av observasjonene  $T_1, T_2, \dots, T_n$ . En kan da vise at

$$E(T_{\max}) = \frac{n\theta}{n+1},$$

$$\text{Var}(T_{\max}) = \frac{n\theta^2}{(n+2)(n+1)^2}$$

- d) La  $\theta^{**} = \frac{1}{n}(n+1)T_{\max}$ . Beregn  $\theta^*$  og  $\theta^{**}$  når følgende data er gitt:

$T_1$	$T_2$	$T_3$	$T_4$	$T_5$
7.5	1.0	4.0	12.5	8.5

Hvilken av estimatorene  $\theta^*$  og  $\theta^{**}$  er best? Begrunn svaret.

**6.14 (E)** En maskin produserer elektriske komponenter der hver av komponentene har en viss motstand (målt i ohm). På grunn av tilfeldigheter under produksjonen, vil det kunne variere litt hvor stor motstand en komponent får.

- a) Erfaring tyder på at motstandene har verdier som er normalfordelt  $N(45, 0.70)$  og at motstandene til forskjellige komponenter er uavhengige av hverandre.

Hva er sannsynligheten for at en tilfeldig valgt komponent har en motstand på mellom 44.2 ohm og 45.7 ohm?

- b) Anta i dette punkt at verdiene på motstandene er uavhengige realiseringer av den normalfordelte stokastiske variabel  $X \sim N(\mu, 0.70)$ , men at vi ikke kjenner størrelsen på  $\mu$ . Vi tar en tilfeldig stikkprøve på 6 komponenter og måler motstanden i hver av dem.

Måleresultat i ohm:

45.4 45.7 46.1 44.9 45.2 44.6

Forklar hva vi forstår med *medianen*,  $m$ , til  $X$ .

Bestem et 97 % konfidensintervall for  $m$ .

Forklar hva dette konfidensintervallet forteller oss.

- c) Anta nå at vi har fått nye opplysninger som har fått oss til å tvile på at  $X$  er normalfordelt, men at vi fortsatt har grunn til å tro at verdiene på motstandene kan ses som uavhengige realiseringer av samme stokastiske variabel  $X$ .

Bestem nå, ut fra måleresultatene ovenfor, et konfidensintervall for  $m$  med konfidensgrad så nær 97 % som mulig. Angi konfidensgraden så nøyaktig det lar seg gjøre.

Kommentér dernest forskjellen mellom lengdene av konfidensintervallene i punktene b) og c).

- 6.15 (E)** Ved en bestemt politisk meningsmåling velger en tilfeldig  $n = 1600$  personer med stemmerett, og spør om hvilket parti de ville stemme på dersom det var valg i morgen. La  $X$  være antall Arbeiderparti-velgere blant de 1600 spurte, og la  $p$  være andelen av Ap-velgere i hele velgermassen.

a) Hvilken sannsynlighetsfordeling har  $X$ ? Begrunn svaret.

b) Angi en estimator for  $p$ . Bestem estimatorens forventning og varians. Er estimatoren forventningsrett?

c) Ved undersøkelsen viste det seg at 37.2 % i utvalget ville stemme på Arbeiderpartiet. Er dette en svært sterkt indikasjon på at oppslutningen om partiet i befolkningen er større enn 36 %?

Besvar spørsmålet ved først å beregne en tilnærmet verdi for  $P(p^* > 0,372 | p = 0,36)$ .

d) Bestem et tilnærmet 95 % konfidensintervall for  $p$ . Hvor stor måtte utvalgstørrelsen  $n$  minst være for at et slikt tilnærmet 95 % konfidensintervall ikke skulle inneholde 0.36?

- 6.16 (E)** En personaldirektør i et meget stort konsern ønsket å studere sykefraværet blant de ansatte det siste året. Han valgte tilfeldig 60 ansatte. Disse hadde tilsammen 918 sykedager, og 40 av de 60 hadde vært syke i minst 18 dager. Videre ble standardavviket for antall sykedager,  $X$ , for en tilfeldig ansatt estimert til  $s = 3.8$  dager i den

samme undersøkelsen. Vi antar at  $X$  er normalfordelt.

Bestem ut fra opplysningene et 95 % konfidensintervall (intervallet i b) får konfidensgrad tilnærmet 95 %) for disse to størrelsene (i a) og b)):

a) Det forventede antall sykedager for en tilfeldig ansatt.

b) Andelen av ansatte i konsernet som hadde vært syke i minst 18 dager.

Personaldirektøren ønsket å gjøre en grundigere tilsvarende undersøkelse hvor han fortsatt var minst 95 % sikker på konklusjonen, men hvor konfidensintervallet var smalere.

c) Hvor mange personer måtte han velge dersom konfidensintervallet for det forventede antall sykedager skulle ha en lengde på høyst 1 dag?  
(Han antar nå at  $\text{std}(X) = 4$ ).

d) Hvor mange personer må velges dersom konfidensintervallet i b) ikke skal være bredere enn 0.1, uansett andelen av ansatte som hadde vært syke i minst 18 dager?

## 6.12 Formelsamling

I de følgende formler skal vi anta at  $X_1, \dots, X_n$  er uavhengige med forventning  $\mu$  og standardavvik  $\sigma$ .  $X$  uten subskript betegner enten en binomisk fordelt variabel (parametre  $n$  og  $p$ ), eller en Poisson-fordelt variabel (parameter  $\lambda$ ).

### Punktestimator for $\mu$

$$\mu^* = \frac{1}{n} \sum_{i=1}^n X_i = \bar{X}$$

$$E(\bar{X}) = \mu, \quad \text{std}(\bar{X}) = \frac{\sigma}{\sqrt{n}}$$

### Punktestimator for $\sigma^2$

$$S^2 = \frac{1}{n-1} \left( \sum_{i=1}^n X_i^2 - \frac{1}{n} \left( \sum_{i=1}^n X_i \right)^2 \right)$$

$E(S^2) = \sigma^2$ , uavhengig av populasjonsfordeling til  $X$ -ene.

$\text{Var}(S^2) = 2\sigma^4/(n-1)$  dersom  $X_1, \dots, X_n$  er uif  $N(\mu, \sigma)$ .

### Punktestimering av $p$

$X$  er  $\text{Bino}(n, p) \Rightarrow p^* = X/n$  er en forventningsrett estimator for  $p$ , med varians  $\text{Var}(p^*) = p(1-p)/n$ .

NB! Alle de følgende konfidensintervall er  $100(1-\alpha)\%$  intervall for den ukjente parameter vi betrakter.

### KI for $\mu$ , kjent $\sigma$ , $n \geq 20$

$$\left( \bar{X} - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}, \bar{X} + z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} \right)$$

der  $z_{\alpha/2}$  er øvre  $\alpha/2$ -fraktil i  $N(0,1)$ -fordelingen. Gjelder også for  $n < 20$  dersom  $X_1, \dots, X_n$  er uif  $N(\mu, \sigma)$ .

### KI for $\mu$ , ukjent $\sigma$ , $n \geq 30$

$$\left( \bar{X} - z_{\alpha/2} \cdot \frac{S}{\sqrt{n}}, \bar{X} + z_{\alpha/2} \cdot \frac{S}{\sqrt{n}} \right)$$

der  $z_{\alpha/2}$  er øvre  $\alpha/2$ -fraktil i  $N(0,1)$ -fordelingen.

### KI for $p$ , $np(1-p) \geq 5$ , $n \geq 30$

$$p^* \mp z_{\alpha/2} \sqrt{\frac{p^*(1-p^*)}{n}}$$

der  $p^* = X/n$ , og  $z_{\alpha/2}$  er øvre  $\alpha/2$ -fraktil i  $N(0,1)$ -fordelingen.

### KI for $\mu$ , ukjent $\sigma$ , $n < 30$

$$\left( \bar{X} - t_{\alpha/2} \cdot \frac{S}{\sqrt{n}}, \bar{X} + t_{\alpha/2} \cdot \frac{S}{\sqrt{n}} \right)$$

der  $t_{\alpha/2}$  er øvre  $\alpha/2$ -fraktil i  $t$ -fordelingen med  $n-1$  frihetsgrader. Dersom  $X$ -ene er uavhengige og normalfordelte har intervallet eksakt konfidensgrad på  $100(1-\alpha)\%$  uavhengig av  $n$ . For ikke-normalfordelte  $X$ -er kan nivået avvike noe for små  $n$ .

### KI for $\lambda$ , $n \geq 30$ , $\lambda \geq 5$

$$(X - z_{\alpha/2} \cdot \sqrt{X}, X + z_{\alpha/2} \cdot \sqrt{X})$$

der  $X$  er en tilfeldig realisasjon av en variabel som er Poisson-fordelt med parameter  $\lambda$ , og  $z_{\alpha/2}$  er øvre  $\alpha/2$ -fraktil i  $N(0,1)$ -fordelingen.

Grensa  $\lambda \geq 5$  vil i mange tilfeller være noe snau, og må vurderes utfra situasjonen. Det er ikke uvanlig å bruke strengere grenser, f.eks.  $\lambda \geq 15$ .

## Kapittel 7

# Hypotesetesting

### 7.1 Innledning

Vi begynner med et eksempel som illustrerer en typisk problemstilling der hypotesetesting kommer til sin rett:

#### Eks. 7.1 Test av helbredelsesrate, $p$ , til ny medisin

Erfaring har vist at helbredelsesraten for en gitt sykdom ved å bruke vanlig medisin er 60 %. En antar at raten for en ny medisin er bedre. For å undersøke dette nærmere, blir den nye medisinen prøvet på et tilfeldig utvalg på 20 pasienter, og antallet,  $X$ , som blir helbredet, blir registrert.

Hvordan skal vi bruke de eksperimentelle dataene for å besvare spørsmålene: «Gir data grunnlag for å konkludere at den nye medisinen har en høyere helbredelsesrate enn vanlig medisin? Hva er i såfall sannsynligheten for feil konklusjon?»

Legg merke til at vi ikke spør om data kan *bevise* om den nye medisinen virkelig *er* bedre enn den gamle, men at et sentralt spørsmål er sannsynligheten for feil konklusjon. Dette skyldes at vi i praksis ikke har mulighet til å prøve medisin på et så stort antall at vi kan være helt sikre i vår konklusjon.

Si nå at 14 av de 20 i utvalget blir helbredet med den nye medisinen. Dette virker som et overbevisende resultat i favør av den nye medisinen. 70 % i utvalget ble jo helbredet, i forhold til 60 % helbredelsesrate med den gamle medisinen. Vi kan imidlertid ikke være helt sikre på at den nye medisinen er best. Et utvalg på 20 personer er ikke så stort, og det er en viss sannsynlighet for at 14 eller flere skulle bli helbredet, selv om den nye medisinen har en helbredelsesrate på langt mindre enn 60 %. Vi må nøye vurdere risikoen ved å gå ut til publikum og anbefale den nye medisinen på tynt grunnlag, når vi tross alt har en medisin vi er *sikre* på at virker i 60 % av tilfellene. Det er kanskje bedre å holde på den gamle medisinen enn å ta sjansen på feilaktig å konkludere at den nye er bedre.

La oss som et annet eksempel tenke oss at 10 av de 20 blir helbredet av den nye medisinen. Det er da nærliggende å konkludere med at den nye medisinen er dårligere enn den gamle, bare 50 % ble jo helbredet i forhold til helbredelsesraten på 60 % for den gamle medisinen. Igjen kan vi imidlertid ikke være helt sikre. Det kan jo være rene skjære uflaksen som gjør at det ikke var mer

enn 10 av 20 som ble helbredet, selv om den nye medisinen skulle ha en helbredelsesrate som var større enn 60 %. Og skulle den nye medisinen virkelig vise seg å være bedre, ville det være synd å kaste den på båten på et kanskje noe tynt grunnlag. ☺

----- (eks. 7.1 fortsetter på s. 183) -----

Hypotesetestingen behandler den type problemer som er skissert ovenfor på en *kvantitativ* måte. Ikke forvil om du synes teorien i første omgang synes noe abstrakt og uforståelig. Rent abstraksjonsmessig er trolig hypotesetesting det vanskeligste emnet i et innføringskurs i statistikk. For å ha et konkret eksempel å belyse teorien ut fra skal vi undervegs vende tilbake til vårt innlednings-eksempel fra medisinens verden.

Kap. 7.2 til 7.6 er av generell karakter, mens kap. 7.7 til 7.10 tar for seg testing av parametrene  $\mu$  (populasjonsforventning) og  $p$  (binomisk parameter).

## 7.2 Hypotesene $H_0$ og $H_1$

Begrepene **nullhypotese** ( $H_0$ ) og **alternativhypotese** ( $H_1$ ) er sentrale begreper i statistisk hypotesetesting. Vi starter derfor med å gi en definisjon av disse begrepene, samt en generell definisjon av begrepet **statistisk hypotese**. Vi kommer senere til begrunnen for definisjonene av  $H_0$  og  $H_1$ .

### Statistisk hypotese (definisjon)

En statistisk hypotese er et utsagn om en populasjon. Utsagnskraften må vurderes på basis av informasjon oppnådd ved tilfeldig utvalg fra populasjonen.

### Nullhypotesen $H_0$ og alternativhypotesen $H_1$ (definisjon)

Vi betrakter en statistisk undersøkelse som har til formål å underbygge et utsagn om en populasjon, med basis i data fra et utvalg fra populasjonen.

**Nullhypotesen**  $H_0$  skal da formuleres som det «motsatte» av utsagnet vi ønsker å underbygge, mens **alternativhypotesen**  $H_1$  skal formuleres som utsagnet.  $H_0$  og  $H_1$  er komplementære hypoteser.

**Eks. 7.2** **Urettferdig terning?** Etter lang tids bruk har en mistanke om at en terning er urettferdig, fordi den viser sekser i omtrent 30 % av kastene. Vårt *utsagn* er da at terningen er urettferdig. Mer konkret kan et eksempel på en hypotese som vi ønsker å underbygge være: «Terningen har sannsynlighet  $p > 1/6$  for å vise sekser i et tilfeldig kast». Dette blir da *alternativhypotesen*  $H_1$ , mens *nullhypotesen*  $H_0$  vil være at  $p = 1/6$ .

Hva er *populasjonen* i dette tilfellet? Vi kan tenke oss at terningen kastes «uendelig» mange ganger, og at vi hver gang registrerer om det blir sekser ( $J$ ) eller ikke-sekser ( $N$ ). Populasjonen består av de resultatene vi da ville få, f.eks.:  $JJNJNNJJN\dots$  I praksis kan vi naturligvis ikke kaste uendelig mange ganger, men vi får et *tilfeldig utvalg* fra populasjonen om vi kaster terningen  $N$  ganger. Dersom f.eks.  $N = 1000$  og vi registrerer 300 seksere, skjønner vi intuitivt at vårt utsagn om urettferdig terning har stor *utsagnskraft*. Dersom imidlertid  $N = 10$  og vi registrerer 3 seksere gir ikke data sterkt støtte til vårt utsagn, da 3 av 10 seksere ikke er svært usannsynlig selv med rettferdig terning ( $p = 1/6$ ). ☺

Fordi  $H_0$  og  $H_1$  er komplementære hypoteser, kan vi sette opp følgende:

- Hypotese  $H_1$ : utsagnet er sant
- Hypotese  $H_0$ : utsagnet er usant

Ved å bruke informasjon fra utvalgs-observasjonene, må en beslutningstaker velge en av følgende 2 beslutninger eller slutninger (*inferenser*):

- Enten: Forkaste  $H_0$  og konkludere at  $H_1$  er sterkt underbygget av dataene.
- Eller: Ikke forkaste  $H_0$  og konkludere at  $H_1$  ikke er sterkt underbygget av dataene.

Prosesssen som fører til et valg mellom de 2 slutningene ovenfor, er det vi kaller å teste en statistisk hypotese. Legg merke til at vi kun har brukt begrepet «forkaste» i forbindelse med  $H_0$ . Årsaken stikker i at vi velger et «konservativt» utgangspunkt. Eksempelvis ønsker vi ikke å hive det gamle på båten med mindre vi på basis av data blir overbevist om at det nye virkelig er bedre. Vi holder med andre ord på hypotesen  $H_0$  inntil dataene gir sterkt grunn til å forkaste  $H_0$ . Grunnen til dette er at en normalt anser konsekvensene av å feilaktig kaste det gamle på båten, som langt alvorligere enn å unnlate å ta det nye i bruk, selv om det er bedre. Vi kan formulere forskjellen på  $H_0$  og  $H_1$  som følger:

Ved å teste en nullhypotese  $H_0$  mot en alternativ hypotese  $H_1$ , så er vår holdning å opprettholde  $H_0$  som sann med mindre dataene taler sterkt imot dette. I så fall forkastes  $H_0$  til fordel for  $H_1$ . Feilaktig å forkaste  $H_0$  (dvs. forkaste  $H_0$  når  $H_0$  er

sann) anses for å være en langt alvorligere feil enn å unnlate å forkaste  $H_0$  når  $H_1$  er sann.

Merk *forskjellen* på å forkaste eller ikke forkaste en *statistisk hypotese* og det å bevise eller motbevise et *matematisk utsagn*. Mens konklusjonen på et matematisk bevis/motbevis er 100 % sikker, så er en konklusjon på en statistisk hypotesetest alltid forbundet med usikkerhet.

Eks: Du vil undersøke om en terning er urettferdig og gir oftere 6 øyne enn et annet antall øyne. Du kaster 60 ganger, og får sekser i 50 av tilfellene. Det er da overveiende sannsynlig at terningen er urettferdig, men du kan ikke være 100 % sikker. Det er jo en viss sannsynlighet, om enn meget liten, for at du ville fått 50 6-ere selv om terningen var rettferdig.

**Eks. 7.1** (forts.)  **$H_0$ ,  $H_1$ , testobservator og forkastingsområde.** I lys av det vi har sagt ovenfor, forstår vi valget av  $H_0$  og  $H_1$  i vårt medisinproblem, dersom vi ønsker å sannsynliggjøre at den nye medisinen er bedre enn den gamle:

$$H_0: p \leq .6 \text{ (ny medisin ikke bedre)}$$

$$H_1: p > .6 \text{ (ny medisin bedre)}$$

Våre eksperimentelle data er angitt i form av  $X$ , antallet som blir helbredet med den nye medisinen. Før eksperimentet er utført, er  $X$  en stokastisk variabel med mulige verdier lik  $0, 1, 2, \dots, 20$ . Hver eneste verdi er fysisk mulig under både  $H_0$  og  $H_1$  (dvs. uansett om det er  $H_0$  eller  $H_1$  som er sann), slik at intet av de mulige utfall av eksperimentet kan bevise 100 % at  $H_0$  eller  $H_1$  er sann. (Bemerk igjen forskjellen fra et matematisk bevis).

Intuisjonen vår tilsier imidlertid at jo større verdier for  $X$  jo sterkere støtte for  $H_1$ . Som grunnlag for en beslutning kan vi f.eks. sette grensen på  $X = 15$ : Vi forkaster  $H_0$  (til fordel for  $H_1$ ) hvis  $X \geq 15$ , mens vi ikke forkaster  $H_0$  hvis  $X \leq 14$ . En slik regel kalles en *test* av nullhypotesen, og  $X$  blir kalt en *testobservator*. De verdier (utfall) av  $X$  som gir forkasting ( $X \geq 15$ ) kalles *forkastingsområdet* til testen. ☺

----- (eks. 7.1 fortsetter på s. 186) -----

## Spesifikasjon av en test

NB! En test er *fullstendig spesifisert* ved en *testobservator* ( $T$ ) og et *forkastingsområde* ( $R$ ), definert som følger:

### Testobservator, $T$ , og forkastingsområde, $R$

En test av  $H_0$  er et grunnlag for (be)slutning, som består i å spesifisere de verdier av en stokastisk variabel  $T$  som er slik at  $H_0$  skal forkastes.  $T$  kalles da en **testobservator**. De  $T$ -verdier som gir forkasting av  $H_0$  kalles **forkastingsområdet** for testen. En test er fullstendig spesifisert ved en testobservator og et forkastingsområde. Vi skal bruke  $T$  som generelt symbol for testobservator, og  $R$  som generelt symbol for forkastingsområde (engelsk:  $R$  for *Rejection*).

NB! Valg av symbolet  $T$  for testobservator må ikke tolkes slik at  $T$  generelt er  $t$ -fordelt. Testobservatoren vil av og til være  $t$ -fordelt og av og til følge andre fordelinger. Logikken bak å bruke  $T$  som generelt symbol for testobservator er at  $T$  er lett å huske som forkortelse for testobservator. Det kommer klart fram av den sammenhengen testobservatoren brukes i hvilken fordeling som er aktuell. Det er forøvrig vanlig å benytte betegnelsene  $z$ -test og  $t$ -test for tester der testobservatoren er henholdsvis normalfordelt og  $t$ -fordelt.

**Eks. 7.3** **Terning.** Vi vender tilbake til eks. 7.2: En terning later til å vise sekser i ca. 30 % av kastene, og vi ønsker å teste hypotesen  $H_1: p > 1/6$  mot nullhypotesen  $H_0: p = 1/6$ , der  $p$  er sannsynligheten for sekser ved et tilfeldig kast. La testobservatoren  $T$  være antall seksere etter  $N = 100$  tilfeldige kast, og si at en på forhånd bestemmer seg for å forkaste  $H_0$  dersom minst 30 av kastene gir sekser:  $R: T \geq 30$ . Testen er da fullstendig spesifisert, dvs. vi vet hva vi skal måle (antall seksere), og vi vet hvilke måleresultat som skal resultere i at  $H_0$  forkastes (minst 30 seksere) og hvilke måleresultat som skal resultere i at  $H_0$  ikke forkastes (mindre enn 30 seksere). ☺

Vi skal senere komme tilbake til kriterier for valg av testobservator og forkastingsområde.

### 7.3 Feiltyper og styrkefunksjon

Vi starter med å liste de 4 mulige situasjonene fra en test:

Tab. 7.1 De 4 mulige testsituasjonene

Testkonklusjon: ↓	Alternativer for den ukjente sannhet:	
	$H_0$ er sann ( $p \leq 0,6$ )	$H_0$ er usann ( $p > 0,6$ )
Ikke forkast $H_0$	Riktig konklusjon	Gal konklusjon (Type II feil)
Forkast $H_0$	Gal konklusjon (Type I feil)	Riktig konklusjon

Tabellen formulert i ord er som følger: Dersom  $H_0$  er sann, tar vi en riktig beslutning dersom vi ikke forkaster  $H_0$ . Forkaster vi  $H_0$ , på tross av at  $H_0$  er sann, tar vi en gal beslutning og begår en type I feil. Dersom  $H_0$  er usann ( $H_1$  er sann), tar vi en riktig beslutning dersom vi forkaster  $H_0$ . Dersom vi ikke forkaster  $H_0$ , på tross av at  $H_0$  er usann, tar vi en gal beslutning og begår en type II feil.

Med andre ord definerer vi type I og type II feil som følger:

#### Type I og type II feil (definisjon)

Type I feil: Forkaste  $H_0$  når  $H_0$  er sann

Type II feil: Ikke forkaste  $H_0$  når  $H_1$  er sann

Sannsynlighetene for de 2 typer feil har fått betegnelsene henholdsvis  $\alpha$  (les: alfa) og  $\beta$  (les: beta):

$$\alpha = P(\text{ Type I feil }) = P(\text{ forkaste } H_0 \mid H_0 \text{ sann })$$

$$\beta = P(\text{ Type II feil }) = P(\text{ ikke forkaste } H_0 \mid H_1 \text{ sann })$$

$$= 1 - P(\text{ forkaste } H_0 \mid H_1 \text{ sann })$$

NB!  $\alpha$  og  $\beta$  er generelt funksjoner av den ukjente parameteren vi skal «hypoteseteste». Vi ser videre at både  $\alpha$  og  $\beta$  er sannsynligheter for å foreta feilslutning, så vi ønsker at både  $\alpha$  og  $\beta$  skal være minst mulig. Som vi skal se, vil imidlertid en reduksjon av f.eks.  $\alpha$  normalt gå på bekostning av  $\beta$ , og omvendt.

Som mål på hvor god (sterk) en test er, innfører vi nå den såkalte *styrkefunksjonen* til en test. Vi skal øremerke symbolet  $\gamma$  til denne funksjonen. Den er generelt en funksjon av den ukjente parameteren vi hypotesetester, og er nært knyttet til definisjonene av  $\alpha$  og  $\beta$ :

### Styrkefunksjon, $\gamma(\theta)$ , og styrkekurve

Som tidligere lar vi  $\theta$  være generelt symbol for en ukjent parameter (eks:  $\theta = p$  eller  $\mu$ ). **Styrkefunksjonen**,  $\gamma(\theta)$ , til en spesifisert test av  $\theta$ , er da definert som følger:

$$\gamma(\theta) = P(\text{forkaste } H_0 \mid \theta) = \begin{cases} \alpha(\theta), & H_0 \text{ sann} \\ 1 - \beta(\theta), & H_1 \text{ sann} \end{cases}$$

Tegner vi opp kurven til  $\gamma(\theta)$  i et  $(\theta, \gamma)$ -diagram, kalles denne kurven for **styrkekurven** til testen.

**Eks. 7.1** (forts.). **Styrkefunksjon.** Vi skal nå belyse de innførte begrepene Type I og Type II feil (med sannsynligheter henholdsvis  $\alpha$  og  $\beta$ ), samt styrkefunksjonen,  $\gamma$ , i vårt medisineksempel. Vi forutsetter at de  $n = 20$  forsøkspersonene som skal prøve medisinen utgjør et tilfeldig utvalg blant et stort antall potensielle brukere. Det er da rimelig å anta at antall personer,  $X$ , som helbredes med den nye medisinen er binomisk fordelt med parametre  $n = 20$  og ukjent  $p$ . I dette tilfellet blir  $\alpha = \alpha(p)$  og  $\beta = \beta(p)$  funksjoner av  $p$  som er definert i hvert sitt område av parameteren  $p$ :  $\alpha(p)$  er definert for de  $p$ -verdier der  $H_0$  er sann ( $p \leq 0.6$ ), mens  $\beta(p)$  er definert for de  $p$ -verdier der  $H_1$  er sann ( $p > 0.6$ ).

La oss starte med å beregne styrkefunksjonen,  $\gamma(p)$ . Fra denne kan vi (se forrige ramme) beregne  $\alpha$  og  $\beta$ . La testen bestå i å forkaste  $H_0$  når  $X \geq 15$ . Vi får da:

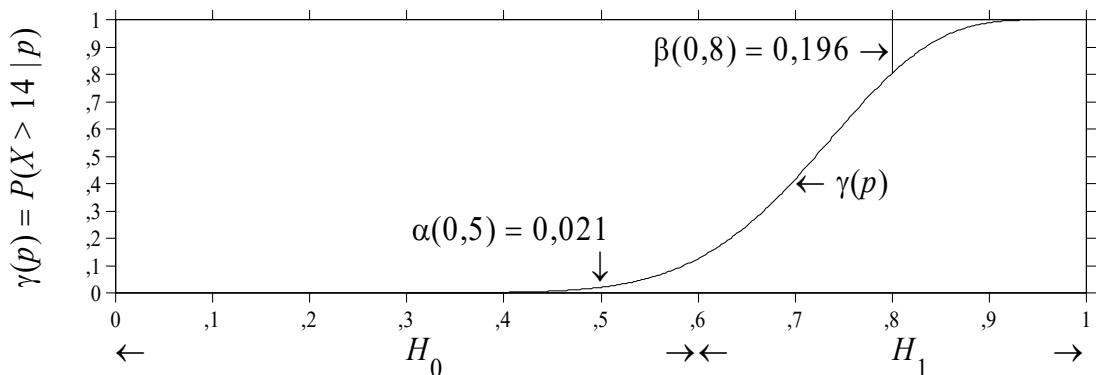
$$\gamma(p) = P(X \geq 15 \mid p) = 1 - P(X \leq 14 \mid p).$$

For ulike verdier for  $p$  finner vi  $\gamma(p)$  ved å slå opp i tabellen for den binomiske fordeling, med parametre  $n = 20$  og den aktuelle  $p$ -verdien. Eksempelvis finner vi for  $p = .5$  at  $P(X \leq 14) = .979 \Rightarrow \gamma(.5) = 1 - .979 = .021$ . Tabellen nedenfor viser flere verdier:

Tab. 7.2 Sannsynlighet for å forkaste  $H_0$  for testen  $X \geq 15$ 

$p$	.3	.4	.5	.6	.7	.8	.9
$\gamma(p)$	.000	.002	.021	.126	.416	.804	.989

NB!  $\alpha(p)$  faller sammen med  $\gamma(p)$ , mens  $\beta(p) = 1 - \gamma(p)$ , slik illustrert i Fig. 7.1. Videre er  $\alpha(p)$  kun definert under  $H_0$ , dvs. for  $p \leq 0.6$ , mens  $\beta(p)$  kun er definert under  $H_1$ , dvs. for  $p > 0.6$ .

Fig. 7.1 Styrkefunksjonen,  $\gamma(p)$ , for testen  $X \geq 15$ 

Med henvisning til fig. 7.1 ovenfor, har vi følgende kommentarer:

- Styrkekurven til en test viser hvor god (sterk) testen er, ved å vise størrelsen til feilsannsynlighetene for alle mulige realiseringer (verdier) av parameteren.
- Ordinaten («y»-verdien) til kurven i det området der  $H_0$  er sann, angir sannsynligheten for type I feil. I figuren er f.eks. vist at sannsynligheten for feilaktig å forkaste  $H_0$  dersom  $p = 0.5$  er  $\alpha(0.5) = 0.021$ .
- I det området av parameteren der  $H_1$  er sann, angir 1 minus ordinatverdien til kurven sannsynligheten for type II feil. I figuren er f.eks. vist at sannsynligheten for feilaktig å unnlate å forkaste  $H_0$  dersom  $p = 0.8$  er  $\beta(0.8) = 1 - \gamma(0.8) = 0.196$ .
- Legg merke til at  $\alpha$ -verdiene er lavere enn  $\beta$ -verdiene i nærheten av grenseområdet mellom  $H_0$  og  $H_1$  ( $p = 0.6$ ): Dette gjenspeiler at type I feil anses som mer alvorlig enn type II feil. ☺

## 7.4 Valg av forkastingsområde

**Eks. 7.1** (forts.) **Valg av  $R$ .** Når vi valgte testen  $X \geq 15$ , var dette et subjektivt valg. Vi kunne godt valgt andre forkastingsområder. Vi skal nå betrakte samme testobservator som før,  $T = X$ , men følgende forskjellige forkastingsområder betraktes:

Test A: forkast  $H_0$  når  $X \geq 15$

Test B: forkast  $H_0$  når  $X \geq 18$

Test C: forkast  $H_0$  når  $X \geq 14$

Vi har allerede studert test A. Tilsvarende fremgangsmåte for test B og C (samme binomiske tabell) gir følgende resultater:

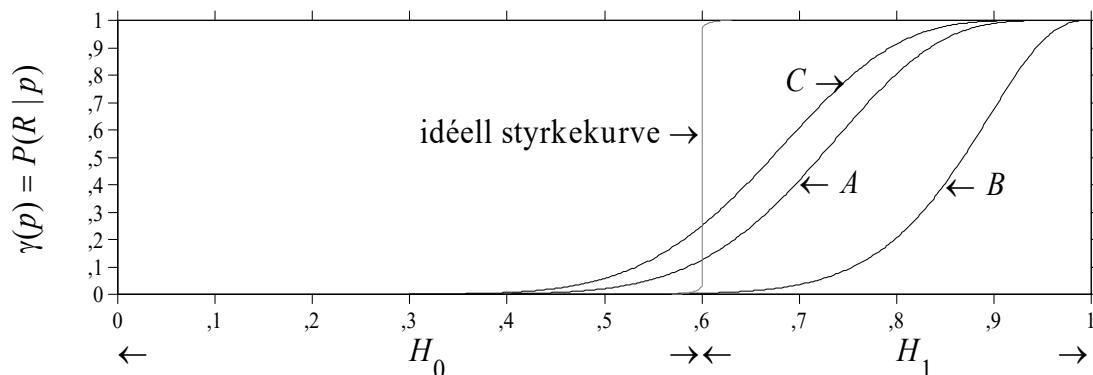


Fig. 7.2 Styrkefunksjoner,  $\gamma(p)$ , for testene A, B og C (idéell styrkekurve inntegnet).

Tab. 7.3 Forkastings-sannsynligheter for testene A, B og C

$p \rightarrow$	.3	.4	.5	.6	.7	.8	.9
(A) $\gamma(p) = P(X \geq 15)$	.000	.002	.021	.126	.416	.804	.989
(B) $\gamma(p) = P(X \geq 18)$	.000	.000	.000	.004	.035	.206	.677
(C) $\gamma(p) = P(X \geq 14)$	.000	.006	.058	.250	.608	.913	.998

Når vi sammenligner styrkekurvene til de 3 testene, observerer vi 2 viktige egenskaper:

- I hvert tilfelle opptrer den største sannsynligheten for type I feil  $\alpha(p)$  når  $p = 0.6$ , som er «i grenseland» mellom  $H_0$  og  $H_1$ . For å kontrollere type I feil er det derfor tilstrekkelig å se på sannsynligheten for forkasting av  $H_0$  for  $p$ -verdien i grensepunktet mellom  $H_0$  og  $H_1$ .

- b) Dersom en av hvilke som helst 2 tester har en mindre  $\alpha(p)$ , så er  $\beta(p)$  større enn for den andre testen. Dette viser at sannsynligheten for en feil av en type reduseres på bekostning av sannsynligheten for den andre feiltypen (man får ikke i pose og sekk).

Den idéelle styrkekurven er tegnet inn i fig. 7.2. En slik oppnås aldri i praksis, men styrkekurven vil nærme seg mer og mer den ideelle ettersom antall observasjoner,  $n$ , øker. Siden type I feil er mer alvorlig enn type II feil, er det vanlig praksis å sikre seg at  $\alpha$ -verdiene holder seg under et forutbestemt nivå. Deretter velges den test som gir minst mulig  $\beta$ -verdi.

Si at vi i undersøkelsen av den nye medisinen krever at  $\alpha$  ikke skal overskride 0,07. Da vil, som vi ser av tab. 7.3, kun test  $B$  tilfredsstille kravet. Vi kunne også valgt alternativene  $X \geq 16$ ,  $X \geq 17$ , osv. For å redusere  $\beta$  mest mulig ville vi da valgt  $X \geq 16$ . Maksimal  $\alpha$ -verdi ville i dette tilfellet blitt:

$$\alpha(.6) = P(X \geq 16 \mid p = .6) = .051$$

som kalles *signifikansnivået* til testen:

### **Signifikansnivået ( $\alpha$ ) til en test** (definisjon)

*Signifikansnivået*, eller kortere, nivået, til en test er den maksimale verdi  $\alpha(p)$  (sannsynlighet for type I feil) kan ha. Denne verdien finnes i praksis ved å beregne  $\alpha(p)$  for den (de) verdi(er) av den ukjente parameteren som skiller  $H_0$  og  $H_1$ . Vi skal bruke symbolet  $\alpha$  (uten noe argument, for ikke å forveksle med funksjonen  $\alpha(\theta)$ ) til å betegne nivået til testen.

Ved valgmulighet blant ulike tester som tilfredsstiller krav til signifikansnivå, er den testen sterkest (best) som gir lavest type II feil.

NB! Det er snarere et politisk enn et statistisk problem å bestemme hvilket signifikansnivå,  $\alpha$ , en test skal ha.

Testen  $X \geq 15$  har signifikansnivå .126, som overskridet den spesifiserte toleransen på .07. Signifikansnivået til testen  $X \geq 17$  møter toleransekravet, men har større  $\beta$ -verdier enn testen  $X \geq 16$ .

Merk at bestemmelsen av toleransekravet,  $\alpha_{\text{maks}}$ , ikke er et statistisk, men snarere et «politisk» problem, i vårt tilfelle et helsepolitisk sådant.

----- (eks. 7.1 fortsetter lenger ned på siden) -----

## Å trekke slutninger fra en test

En konklusjon fra en test kan kort og godt formuleres på en av følgende 2 måter:

- 1)  $H_0$  forkastes på signifikansnivå  $\alpha$
- 2)  $H_0$  forkastes ikke på signifikansnivå  $\alpha$

I tillegg er det god skikk å angi signifikant sannsynlighet,  $P^*$ , definert som følger:

### Signifikant sannsynlighet, $P^*$ (definisjon)

Signifikantsannsynligheten (ofte kalt **p-verdien**),  $P^*$ , til en observert verdi av testobservatoren er den minste  $\alpha$ -verdi som er slik at observasjonen leder til forkasting av  $H_0$ .

Med andre ord,  $P^*$  er sannsynligheten for den observerte verdi eller mer ekstreme verdier under  $H_0$  (eller mer presist under  $H_0$  når parameterverdi(ene) er på grensen mot  $H_1$ -området). Vi ønsker en minst mulig verdi for  $P^*$ : Jo mindre  $P^*$ -verdi, jo sterkere taler dette mot  $H_0$ , og jo mer underbygger  $P^*$ -verdien hypotesen  $H_1$ , som er den vi ønsker å underbygge.

**Eks. 7.1** (forts.) **p-verdi.** La oss gå tilbake til tab. 7.3. Anta at vi utfører test  $C$ , som har signifikansnivå  $\alpha = .25$  (svært høyt!). Si at vi utfører testen og finner at  $X = 18$ . Fra tab. 7.3 ser vi da at sannsynligheten for at hele 18 eller flere av 20 forsøkskandidater blir helbredet er bare 4 promille, dersom den sanne helbredelsesraten er  $p = 0.6$  (som den gamle medisinen). Følgelig:  $P^*$  er i dette tilfellet lik 0.004, et resultat som *sterkt* underbygger hypotesen at den nye medisinen er bedre enn den gamle. ☺☺☺

## 7.5 Tester med tosidig alternativ

Hva som menes med tosidig alternativ er illustrert i følgende eks:

$$H_0: p = 0.6 \text{ mot } H_1: p < 0.6 \text{ eller } p > 0.6 (p \neq 0.6)$$

Vi ser at  $H_1$  inneholder 2 områder for  $p$ , og ikke bare ett område som tidligere. Derav uttrykket tosidig test. Strukturen på forkastingsområdet blir derved også tosidig:

Forkast  $H_0$  dersom  $X \leq c_1$  eller  $X \geq c_2$

der grensene  $c_1$  og  $c_2 > c_1$  må bestemmes slik at  $\alpha$  holder seg under en spesifisert toleranse.

**Eks. 7.4** **Lettøl.** Et bryggeri ønsker å undersøke responsen i markedet på en ny type lettøl,  $A$ , i forhold til dagens type,  $B$ . To og to glass med type  $A$  i det ene og type  $B$  i det andre settes fram for tilfeldige kunder som smaker uavhengig av hverandre uten å vite hvilke typer som er i de to glassene. Når tilsammen  $n = 15$  av kundene har sagt at de kjenner forskjell, og hvilket de liker best, avsluttes undersøkelsen.  $X =$  antall som foretrekker type  $A$  registreres. Vi antar da at det er rimelig med en binomisk modell  $Bino(15, p)$  der  $p$  er ukjent. For enkelhets skyld skal vi anta at i hvilken rekkefølge de to ølsortene smakes ikke påvirker preferansen.

Dersom kundene ikke har noen preferanse for en av de 2 typene, så er  $p = 0.5$ . Siden vi ønsker å påvise en eventuell preferanse for  $A$  eller  $B$ , vil vi teste  $H_0: p = 0.5$  mot  $H_1: p \neq 0.5$ . Fordi ekstremt høye såvel som ekstremt lave verdier for  $X$  vil tale imot en mangel på preferanse, så bør forkastingsområdet være to-sidig. Som illustrasjon betrakter vi følgende valg av forkastingsområde:

- a) Forkast  $H_0$  dersom  $X \leq 4$  eller  $X \geq 11$
- b) Forkast  $H_0$  dersom  $X \leq 3$  eller  $X \geq 12$

Bemerk at under  $H_0$  er  $X$  binomisk fordelt med  $n = 15$  og  $p = 0.5$ . Denne fordelingen er symmetrisk (fordi  $p = 0.5$ ), og symmetriske valg for forkastingsområdet er derfor rimelige (og de er også best teoretisk).

Anta at vi ønsker en toleransegrense for sannsynligheten for type I feil på under .05. Fra tabellen over den binomiske fordeling med  $n = 15$  og  $p = 0.5$ , finner vi:

$$\text{test a): } \alpha = P(X \leq 4 \mid p = 0.5) + P(X \geq 11 \mid p = 0.5) = .059 + .059 = .118$$

$$\text{test b): } \alpha = P(X \leq 3 \mid p = 0.5) + P(X \geq 12 \mid p = 0.5) = .018 + .018 = .036$$

For å klare toleransekravet på .05 kan vi derfor bruke test b), men ikke test a). Signifikansnivået for test b) er  $\alpha = .036$ .

Når vi nå har valgt en test med tilfredsstillende signifikansnivå undersøker vi hvor godt den vil «avdekke» alternativhypotesen. Dette gjør vi ved å beregne styrkefunksjonen,  $\gamma(p)$ , for ulike verdier av  $p$ . Pr. definisjon er  $\gamma(p)$  gitt ved

$$\gamma(p) = P(\text{forkaste } H_0 \mid p) = P(X \leq 3 \mid p) + P(X \geq 12 \mid p)$$

Igjen bruker vi binomialfordelingstabellen for å finne sannsynlighetene vi trenger. For eksempel, når  $p = .4$ , så er  $P(X \leq 3) = .091$  og  $P(X \geq 12) = .002$ , slik at  $\gamma(.4) = .091 + .002 = .093$ . Andre verdier er vist i tab. 7.4, og styrkekurven er vist grafisk i fig. 7.3.

Tab. 7.4 Noen verdier til styrkefunksjonen,  $\gamma(p)$ , til test b).

$p$	.1	.2	.3	.4	.5	.6	.7	.8	.9
$P(X \leq 3)$	.944	.648	.297	.091	.018	.002	.000	.000	.000
+	+	+	+	+	+	+	+	+	+
$P(X \geq 12)$	.000	.000	.000	.002	.018	.091	.297	.648	.944
$\gamma(p)$	.944	.648	.297	.093	.036	.093	.297	.648	.944

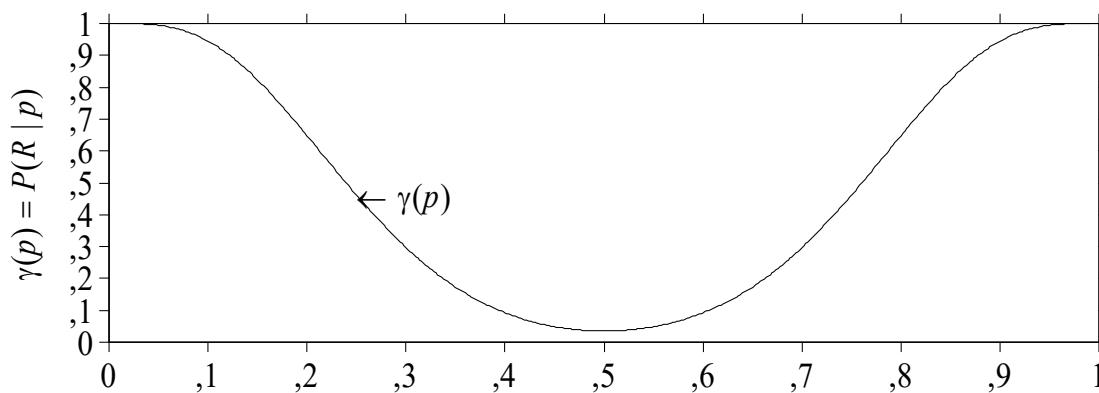


Fig. 7.3 Styrkefunksjon,  $\gamma(p)$ , for test b).

Etter å ha kjørt eksperimentet, anta at bryggeriet finner at 5 av 15 kunder foretrekker A, dvs.  $x = 5$ . Fordi denne verdien ikke er i forkastingsområdet, konkluderer bryggeriet med at  $H_0$  ikke forkastes på signifikansnivå  $\alpha = .036$ . For å beregne signifikant sannsynlighet, må vi betrakte en test der  $H_0$  forkastes

dersom  $X \leq 5$  eller  $X \geq 10$ , siden vi har observert  $X = 5$  og fordi forkastingsområdet her skal være symmetrisk. Vi ville da fått:

$$P^* = P(X \leq 5 \mid p = 0.5) + P(X \geq 10 \mid p = 0.5) = .151 + .151 = .302$$

Denne store verdien for  $P^*$  indikerer at vi mangler en sterk støtte for hypotesen at en preferanse for  $A$  eller  $B$  foreligger. ☺

## 7.6 Generelle trinn i hypotesetesting

Det er mange detaljer og begreper vi har vært igjennom. I et forsøk på å klargjøre stoffet, la oss sette opp hovedtrinnene i en hypotesetest:

- 1) Bestem en rimelig sannsynlighetsmodell på basis av de eksperimentelle data og de hypoteser det er aktuelt å fremsette. I en slik modell inngår det som regel minst en ukjent parameter,  $\theta$  (vi har sett på  $\theta = p$  i en binomisk modell  $Bino(n,p)$ ). Bestem for hvilke parameterverdier hver hypotese er sann.
- 2) Når vi vet hvilken hypotese vi ønsker å underbygge, formulerer vi  $H_0$  som den motsatte hypotesen, og  $H_1$  som den «ønskede» hypotesen. Formuler  $H_0$  og  $H_1$  ved de parameterverdier der hver av hypotesene er sanne. Eksempler kan være:

$$\begin{aligned} H_0: p &= .3 \text{ mot } H_1: p \neq .3 \text{ (tosidig alternativ)} \\ H_0: p &= .2 \text{ mot } H_1: p > .2 \text{ (ensidig alternativ)} \end{aligned}$$

- 3) Velg testobservator,  $T$ , og forkastingsområde,  $R$ .

$T$  er en funksjon av observasjonsvariablene (de eksperimentelle data), og skal konstrueres slik at dens verdi klarest mulig indikerer om observasjonene støtter  $H_1$  eller ikke. Bestem sannsynlighetsfordelingen til  $T$ . Den vil ofte være rett fram ut fra sannsynlighetsmodellen funnet i 1), og vil som regel både være en funksjon av den ukjente parameteren vi tester, innsatt verdien den ukjente parameteren har på grensen mellom  $H_0$  og  $H_1$ , og antall observasjoner.

Strukturen på  $R$  følger strukturen til  $H_1$ . Dersom f.eks.  $H_1$  er hypotesen  $p > .2$ , så vil  $R$  være på form  $T > c$  (samme form for ulikhetstegn). For å bestemme  $c$  benytter vi sannsynlighetsfordelingen til  $T$  når den ukjente parameteren har den verdi, eller de verdier, som skiller  $H_0$  og  $H_1$ . Vi finner den  $c$ -verdi som er slik at  $P(\text{forkaste } H_0) = \alpha$ , der  $\alpha$  er nivået til testen.

Et rekke tester vil imidlertid ha tilfredsstillende nivå. I tillegg til å tilfredsstille dette kravet, ønsker vi også en minst mulig sannsynlighet for type II feil med færrest mulige forsøk (minst mulig  $n$ -verdi). For å bestemme en optimal test, er styrkefunksjonen et egnet hjelpemiddel. Har vi spesifiserte krav til type II feilsannsynligheter, øker vi  $n$ -verdien helt til et optimalt valg av forkastingsområde både tilfredsstiller toleransekravet til signifikansnivå, og våre krav til type II feilsannsynligheter.

- 4) Etter at testen er skikkelig formulert, gjennomføres testen, og konklusjoner trekkes. Bestem  $T$ -verdien fra de eksperimentelle data, og bestem om  $H_0$  forkastes eller ikke. Beregn så signifikant sannsynlighet ( $p$ -verdi),  $P^*$ .

## 7.7 Test av $\mu$

La oss først kortfattet gjenta de 4 hovedtrinnene i en hypotesetest:

- 1) Bestem passende sannsynlighetsmodell.
- 2) Formuler  $H_0$  og  $H_1$  og angi hvilke verdier for  $\mu$  som hører under  $H_0$  og  $H_1$ .
- 3) Bestem testobservator,  $T$ , og forkastingsområde,  $R$ .
- 4) Utfør testen og formuler konklusjon(e) fra testen.

Vi skal anta at  $\bar{X}$  er tilnærmet  $N(\mu, \sigma / \sqrt{n})$ . Som for konfidensintervall, må vi skille mellom store utvalg ( $n \geq 30$ ) og små utvalg ( $n < 30$ ). Vi må også skille mellom tilfellene der  $\sigma$  er kjent og der  $\sigma$  er ukjent. La oss starte med situasjonen at  $\sigma$  er kjent:

### Test av $\mu$ når $\sigma$ er kjent (z-test)

Testobservator:  $T = \frac{\bar{X} - \mu_0}{\sigma / \sqrt{n}}$  som er tn.  $N(0,1)$  når  $\mu = \mu_0$

$$\begin{array}{lll} H_0: \mu \leq \mu_0 & H_0: \mu \geq \mu_0 & H_0: \mu = \mu_0 \\ H_1: \mu > \mu_0 & H_1: \mu < \mu_0 & H_1: \mu \neq \mu_0 \\ \Rightarrow R: Z > z_\alpha & \Rightarrow R: Z < -z_\alpha & \Rightarrow R: |Z| > z_{\alpha/2} \end{array}$$

der  $z_\alpha$  og  $z_{\alpha/2}$  er henholdsvis øvre  $\alpha$ -fraktil og  $\alpha/2$ -fraktil i  $N(0,1)$ -fordelingen.

### Begrunnelse for testobservatoren i forrige ramme

For å forstå tankegangen bak testobservatoren  $T$  gitt i forrige ramme, kan det først være nyttig med følgende omskrivning:

$$T = \frac{\bar{X} - \mu_0}{\sigma / \sqrt{n}} = \frac{(\bar{X} - \mu) + (\mu - \mu_0)}{\sigma / \sqrt{n}} = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} + \frac{\mu - \mu_0}{\sigma / \sqrt{n}}$$

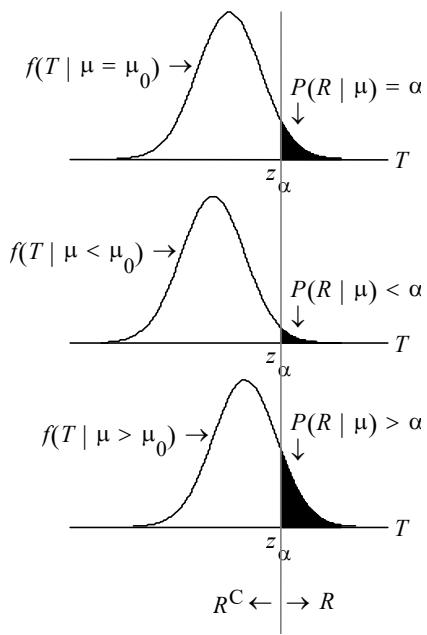
Vi kan altså erstatte  $T$  med en sum av to ledd:

$$T = Z + c$$

der første leddet  $Z = (\bar{X} - \mu) / (\sigma / \sqrt{n})$  er tn.  $N(0,1)$  uansett verdi av  $\mu$ , og andre leddet  $c = (\mu - \mu_0) / (\sigma / \sqrt{n})$  er en konstant avhengig av  $\mu$ . Når  $\mu = \mu_0$  ser vi at dette konstantleddet forsvinner. Virkningen konstanten  $c$  har på fordelingen  $f(T)$  vil være en parallelforskyvning av  $f(T)$ -kurven langs  $T$ -aksen. Dette er nærmere illustrert i figuren nedenfor. Vi begrenser oss til å betrakte testsituasjonen:

$$H_0: \mu \leq \mu_0, \quad H_1: \mu > \mu_0, \quad R: T > z_\alpha.$$

$H_0$  forkastes altså dersom  $T > \text{øvre } \alpha\text{-fraktil i } N(0,1)\text{-fordelingen}$ , og sannsynligheten for å forkaste  $H_0$  er derfor bestemt av fordelingen  $f(T)$ . Grafisk blir sannsynligheten for å forkaste  $H_0$  gitt av arealet under  $f(T)$ -kurven til høyre for  $T = z_\alpha$ , dvs. den del av kurven som ligger i forkastingsområdet  $R$ .



$\mu = \mu_0$ : I dette tilfellet er vi på grensen mellom  $H_0$  og  $H_1$ . Det er denne situasjonen som er utgangspunktet for definisjonen av testnivået  $\alpha$ , så sannsynligheten for å forkaste  $H_0$  er i dette tilfellet lik  $\alpha$ .

$\mu < \mu_0$ : I dette tilfellet er  $H_0$  sann, fordelingen til  $f(T)$  forskyves mot venstre sammenlignet med  $f(T)$  når  $\mu = \mu_0$ , og vi ser at sannsynligheten for å forkaste  $H_0$  blir mindre enn  $\alpha$ .

$\mu > \mu_0$ . I dette tilfellet er  $H_1$  sann, fordelingen til  $f(T)$  forskyves mot høyre sammenlignet med  $f(T)$  når  $\mu = \mu_0$ . Vi får økt sannsynlighet for å forkaste  $H_0$ . Kurven forskyves raskere mot høyre (sterkere test) når  $n$  øker.

**Eks. 7.5** **Behandlingstid.** På bakgrunn av utstrakt bruk i lang tid, er det kjent at behandlingstida for en sykdom ved hjelp av en (gammel) medisin har forventning på  $\mu_0 = 15$  dager og et standardavvik på  $\sigma = 3$  dager. Det blir påstått at en ny medisin kan redusere behandlingstida. Vi antar at standardavviket til behandlingstida for den nye medisinen også er  $\sigma = 3$  (Vi dropper benevningen som er dager).

### Oppgave

Undersøk om det er grunnlag for å påstå at forventet behandlingstid,  $\mu$ , for den nye medisinen er kortere enn den gamle, med bakgrunn i forsøk på et tilfeldig utvalg på  $n = 70$  pasienter. Signifikansnivået (eller kort og godt nivået) for testen settes til  $\alpha = 0.025$ . Forsøket utføres og det viser seg at  $\bar{x} = 14,1$  (dager).

### Løsningsforslag

Vi går gjennom trinn for trinn i 4-punktsoppsettet for en hypotesetest:

- 1) Vi lar  $\bar{X}$  betegne midlere behandlingstid for de 70 pasientene. Som sannsynlighetsmodell er det da rimelig å anta at  $\bar{X}$  er tilnærmet  $N(\mu, 3/\sqrt{70})$ .
- 2) Vi ønsker å vise at den nye medisinen er bedre enn den gamle, dvs. at  $\mu < \mu_0 = 15$ .  $H_0$  blir den motsatte hypotesen av den vi ønsker å vise, og vi får da:

$$H_0: \mu \geq 15 \quad H_1: \mu < 15$$

- 3) Her er  $\sigma = 3$  kjent, og vi bruker følgende testobservator (se siste ramme):

$$T = \frac{\bar{X} - 15}{3/\sqrt{70}} = \frac{\bar{X} - 15}{3} \cdot \sqrt{70}$$

Strukturen på forkastingsområdet «følger»  $H_1$ . Vi finner grunn til å forkaste  $H_0$  dersom  $T$  blir «liten nok». Vi får (se siste ramme):

$$R: T < -z_{0.025} = -1.96$$

der  $z_{0.025} = 1.96$  er øvre 2.5 %-fraktil i normalfordelingen  $N(0,1)$ , som er fordelingen til vår testobservator,  $T$ , når  $H_0$  er sann. Fraktilen kan blant annet finnes fra nederste linje i  $t$ -fordelingstabellen (d.f. =  $\infty$  antall frihetsgrader).

- 4) Vi utfører testen og finner at  $\bar{x} = 14.1$  dager. Setter vi dette tallet inn i uttrykket for testobservatoren får vi:

$$T = \frac{14.1 - 15}{3/\sqrt{70}} = -2.51 < -1.96$$

**Konklusjon:** Det er grunnlag i data på signifikansnivå 2.5 % til å forkaste  $H_0$ , og påstå at forventet behandlingstid med den nye medisinen er kortere enn med den gamle.

Vi finner til slutt signifikant sannsynlighet,  $P^*$ , dvs. det laveste testnivå vi kunne valgt og likevel fått forkasting av  $H_0$ . Dette tilsvarer å finne hvilken verdi  $\alpha$  har når øvre fraktil  $z_\alpha = 2.51$ . Da må vi slå opp i normal-fordelingstabellen med  $z = -2.51$  (fordi tabellen lister de nedre og ikke de øvre fraktiler), og finner at  $\alpha = .0060$ . Vi finner derved at  $P^* = 0.0060$ , dvs. vi kunne faktisk valgt et nivå på 0.6 % og likevel fått forkasting. Dette underbygger ytterligere hypotesen at den nye medisinen er best. ☺

**Eks. 7.6** **Styrkekurve.** Vi skal se på styrkekurven for testen i eks. 7.5. Anta at den nye medisinen i eksemplet er bedre enn den gamle, dvs. at  $\mu < 15$ . Da er  $\bar{X}$  tilnærmet  $N(\mu, 3/\sqrt{70})$  med  $\mu < 15$ . I dette tilfellet er  $(\bar{X} - \mu)/(3/\sqrt{70}) \sim N(0,1)$ , mens testobservatoren  $T = (\bar{X} - 15)/(3/\sqrt{70})$  ikke lenger er  $N(0,1)$ . Testen i eks. 7.2 bestod i å forkaste  $H_0: \mu \geq 15$  dersom  $T < -1.96$ .

### Oppgave

Finn først styrken til testen når  $\mu = 13.8$ , og bestem deretter hele forløpet til  $\gamma(p)$ .

### Løsningsforslag

Styrken til testen er sannsynligheten for å forkaste  $H_0$  når  $H_1$  er riktig, og i dette tilfellet skal vi se på  $\mu = 13.8$  (som ligger i området der  $H_1$  er riktig, siden  $13.8 < 15$ ). Siden testobservatoren,  $T$ , ikke lenger er  $N(0,1)$ , må vi foreta følgende omskrivninger:

$$\gamma(13.8) = P(\text{Forkaste } H_0 | \mu = 13.8) = P(T < -1.96 | \mu = 13.8)$$

$$= P\left(\frac{\bar{X} - 15}{3/\sqrt{70}} < -1.96 | \mu = 13.8\right)$$

(trekker fra  $\mu$  og legger til  $\mu$  i teller:  $0 = -\mu + \mu$ ):

$$= P\left(\frac{\bar{X} - 13.8 + 13.8 - 15}{3/\sqrt{70}} < -1.96 | \mu = 13.8\right)$$

(deler brøken opp i to, slik at første del blir en  $N(0,1)$ -variabel):

$$= P\left(\frac{\bar{X} - 13.8}{3/\sqrt{70}} + \frac{13.8 - 15}{3/\sqrt{70}} < -1.96 | \mu = 13.8\right)$$

(Kaller  $N(0,1)$ -variablene for  $Z$ , og separerer denne til venstre i ulikheten):

$$= P\left(Z < -\frac{13.8 - 15}{3 / \sqrt{70}} - 1.96\right)$$

(Beregner høyresiden av ulikheten og bruker  $N(0,1)$ -tabell):

$$\approx P(Z < 1.39) = \Phi(1.39) \approx 0.92$$

Ved tilsvarende regning ville vi funnet andre punkter på  $\gamma$ -kurven, f.eks.  $\gamma(13.5) = .99$  og  $\gamma(14.4) = .39$ . Den fullstendige kurven er vist i fig. 7.4.

Ved generalisering av fremgangsmåten ovenfor (erstatt 13,8 med  $\mu$ , 15 med  $\mu_0$ , 3 med  $\sigma$ , 70 med  $n$  og  $-1.96$  med  $z_{\alpha/2}$ ), ville vi funnet følgende generelle formel for styrkefunksjonen:

$$\gamma(\mu) = P(T < -z_{\alpha/2} \mid \mu) = \Phi\left(-z_{\alpha/2} - \frac{\mu - \mu_0}{\sigma / \sqrt{n}}\right) \quad \text{☺}$$

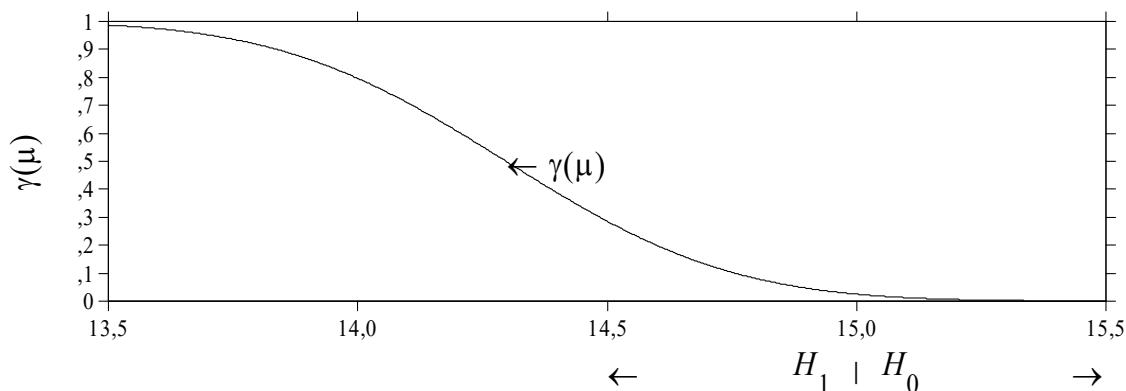


Fig. 7.4 Styrkekurven,  $\gamma(\mu)$ , for testen i eks. 7.5

Vi går nå over til å betrakte tilfellet med små utvalg ( $n < 30$ ) og ukjent  $\sigma$ . Vi må da basere oss på  $t$ -fordelingen med  $n-1$  frihetsgrader i stedet for  $N(0,1)$ -fordelingen:

**Test av  $\mu$  når  $\sigma$  er ukjent ( $t$ -test)** små utvalg,  $n < 30$

$$\text{Testobservator: } T = \frac{\bar{X} - \mu_0}{S / \sqrt{n}}$$

$T$  er tilnærmet  $t$ -fordelt med  $n-1$  frihetsgrader når  $\mu = \mu_0$

$$\begin{array}{lll} H_0: \mu \leq \mu_0 & H_0: \mu \geq \mu_0 & H_0: \mu = \mu_0 \\ H_1: \mu > \mu_0 & H_1: \mu < \mu_0 & H_1: \mu \neq \mu_0 \\ \Rightarrow R: T > t_\alpha & \Rightarrow R: T < -t_\alpha & \Rightarrow R: |T| > t_{\alpha/2} \end{array}$$

der  $t_\alpha$  og  $t_{\alpha/2}$  er henholdsvis øvre  $\alpha$ -fraktil og øvre  $\alpha/2$ -fraktil i  $t$ -fordelingen med  $n-1$  frihetsgrader.

**Eks. 7.7** Helsedirektoratet ønsker å undersøke om midlere antall bakterier pr. liter vann ved en offentlig badestrand overskridet sikkerhetsnivået på 200. Forskere har samlet 10 vannprøver på 1 liter hver og funnet følgende antall bakterier:

175	190	215	198	184
207	210	193	196	180

*Oppgave*

Gir dataene grunnlag for bekymring?

*Løsningsforslag*

Vi går igjennom vårt tidligere 4-punkts program punkt for punkt:

- 1) La  $X$  betegne antall bakterier i en tilfeldig literprøve. Her er det rimelig å anta at  $X$  er tilnærmet Poisson-fordelt, med en  $\lambda$ -verdi ikke altfor langt fra 200. Dermed er det rimelig å anta at  $X$  er tilnærmet normalfordelt  $N(\mu, \sigma)$ , der både  $\mu$  og  $\sigma$  er ukjente. Tar vi middelverdien,  $\bar{X}$ , av 10 literprøver, vil denne trolig være  $N(\mu, \sigma/\sqrt{n})$  med *svært* god tilnärmelse.
- 2) Det er ikke helt opplagt hva vi skal velge som  $H_0$  og  $H_1$  her. La oss velge en miljøvennlig profil, som går ut på at vi synes dataene gir grunn til bekymring med mindre vi får forkastet hypotesen at bakterienivået virkelig er for høyt (NB! dette er et politisk og ikke et statistisk valg). Vi formulerer derfor følgende hypoteser  $H_0$  og  $H_1$ :

$$H_0: \mu > 200 \quad H_1: \mu < 200$$

3) Testobservatoren blir som angitt i siste ramme:

$$T = \frac{\bar{X} - 200}{S / \sqrt{10}}$$

Strukturen på forkastingsområdet «følger»  $H_1$ , dvs. vi forkaster  $H_0$  dersom  $T$  får en tilstrekkelig liten verdi (se forrige ramme).  $R: T \leq -t_\alpha$ , der  $t_\alpha$  er øvre  $\alpha$ -fraktil i  $t$ -fordelingen med  $n-1 = 9$  frihetsgrader.

4) Vi velger signifikansnivå 1 %, og finner fra tabell at  $t_{01} = 2.82$ , dvs. vi forkaster  $H_0$  dersom  $T < -2.82$ . Vi utfører så testen. Først må vi finne  $\bar{x}$  og  $s$  på basis av de 10 målingene:

$$\bar{x} = 194.8 \text{ og } s = 13.14 \Rightarrow T = \frac{194.8 - 200}{13.14 / \sqrt{10}} = \frac{-5.2}{4.156} = -1.25$$

Som vi ser er  $T$  ikke mindre enn  $-2.82$ , og vi konkluderer:

«Testen gir på 1.0 % nivå ikke grunnlag for å påstå at det er mindre enn 200 bakterier pr. liter, og det synes følgelig ikke å være sterk støtte for at forskriftene er overholdt.»

Vi avslutter med test av  $\mu$  når  $\sigma$  er ukjent og utvalget er stort ( $n \geq 30$ ). I prinsippet får vi da samme fremgangsmåte som når  $\sigma$  er kjent:

### Test av $\mu$ når $\sigma$ er ukjent (store utvalg, $n \geq 30$ )

Testobservator:  $T = \frac{\bar{X} - \mu_0}{S / \sqrt{n}}$ , som er tn.  $N(0,1)$  når  $\mu = \mu_0$

$$\begin{array}{lll} H_0: \mu \leq \mu_0 & H_0: \mu \geq \mu_0 & H_0: \mu = \mu_0 \\ H_1: \mu > \mu_0 & H_1: \mu < \mu_0 & H_1: \mu \neq \mu_0 \\ \Rightarrow R: T > z_\alpha & \Rightarrow R: T < -z_\alpha & \Rightarrow R: |T| > z_{\alpha/2} \end{array}$$

der  $z_\alpha$  og  $z_{\alpha/2}$  er henholdsvis øvre  $\alpha$ -fraktil og øvre  $\alpha/2$ -fraktil i  $N(0,1)$ -fordelingen.

Data	
	C1
↓	
1	175
2	190
3	215
4	198
5	184
6	207
7	210
8	193
9	196
10	180
11	

La oss vise hvordan vi kunne utført testen i Minitab. Til venstre er vist et utsnitt som viser at dataene er lagt inn i kolonne c1.

Nedenfor er først vist koden i Minitab, og deretter utskrift.

### MINITAB

```
MTB > ttest 200 c1; # 200 er «testverdien» for  $\mu_0$ 
SUBC> alternative -1. # -1 angir at  $H_1: \mu < \mu_0$ 
```

Kommandoene ovenfor gir følgende utskrift:

#### T-Test of the Mean

Test of mu = 200.00 vs mu < 200.00

Variable	N	Mean	StDev	SE Mean	T	P
C1	10	194.80	13.14	4.15	-1.25	0.12

Vi kjenner igjen antall observasjoner,  $N = 10$ , middelverdien på 194.8, standardavviket til enkeltobservasjonene og middelverdien på henholdsvis 13.14 og 4.15, samt verdien  $T = -1.25$  til testobservatoren. I tillegg er signifikant sannsynlighet,  $P^*$  ( $p$ -verdien) = 0.12, listet. ☺

## 7.8 Test av binomisk $p$

Vi begrenser oss her til store utvalg ( $n \geq 30$ ), og antar at  $p^* = X/n$  er tilnærmet normalfordelt  $N(p, \sqrt{p(1-p)/n})$ , der  $X$  er antall J-utfall av  $n$  Bernoulli-forsøk.

Forøvrig blir framgangsmåten svært lik hypotestesting av  $\mu$ :

### Test for binomisk $p$ , store utvalg

Forutsetninger:  $n \geq 30$ ,  $np(1-p) > 5$

$$\text{Testobservator: } T = \frac{p^* - p_0}{\sqrt{p_0(1-p_0)/n}} = \frac{X - np_0}{\sqrt{np_0(1-p_0)}}$$

der  $p^* = X/n$ .  $T$  er tilnærmet  $N(0,1)$  når  $p = p_0$

$$\begin{array}{lll} H_0: & p \leq p_0 & H_0: & p \geq p_0 & H_0: & p = p_0 \\ H_1: & p > p_0 & H_1: & p < p_0 & H_1: & p \neq p_0 \\ \Rightarrow & R: T > z_\alpha & \Rightarrow & R: T < -z_\alpha & \Rightarrow & R: |T| > z_{\alpha/2} \end{array}$$

der  $z_\alpha$  og  $z_{\alpha/2}$  er henholdsvis øvre  $\alpha$ -fraktil og øvre  $\alpha/2$ -fraktil i  $N(0,1)$ -fordelingen.

#### Eks. 7.8

**Fraflytting.** I et relativt stort fylke tenker vi oss at årlig fraflytting av ungdom i perioden 1984-1988 har vært på 20 %. En gjeng statistikk-studenter ønsker å undersøke om fraflyttinga var i ferd med enten å øke eller minke fra 1989. De får tilgang til et navneregister, og plukker ut 500 tilfeldige ungdommer som bodde i kommunen ved utgangen av 1989. De finner at 91 av disse har flyttet iløpet av 1990.

#### Oppgave

Gir det observerte resultat (91 av 500 ungdommer flyttet ut) sterke indikasjoner på at utflyttingsraten ikke lenger er 20 % pr. år?

#### Løsningsforslag

Igjen går vi gjennom vårt 4-punkts skjema:

- 1) Vi lar  $p$  betegne den ukjente andelen av ungdommer som flyttet fra fylket i 1990, og vi lar  $X$  betegne antall utflyttere i utvalget på  $n = 500$ . Det antas at  $n = 500$  er så mye mindre enn totalt antall ungdom i fylket, at  $X$  med god tilnærming er binomisk fordelt  $Bino(500,p)$ . Videre ser vi ut fra observasjonen  $x = 91$  at det er rimelig å anta at  $np(1-p) > 5$ .
- 2) Vi ønsker å ha stor utsagnskraft i en konklusjon som sier at fraflyttinga har forandret seg, derfor velger vi følgende hypoteser:

$$H_0: p = 0.2 \quad H_1: p \neq 0.2$$

3) Testobservatoren blir (se forrige ramme):

$$T = \frac{X - 500 \cdot 0.2}{\sqrt{500 \cdot 0.2 \cdot 0.8}} = \frac{X - 100}{\sqrt{80}}$$

Strukturen på forkastingsområdet «følger»  $H_1$ , og vi forkaster  $H_0$  dersom  $T$  enten blir «tilstrekkelig» stor eller «tilstrekkelig» liten. Vi velger testnivå på 5 % og forkaster da  $H_0$  når  $|T| > z_{\alpha/2} = 1.96$ .

$$R: |T| > 1.96$$

4) Vi utfører testen og får:

$$T = \frac{91 - 100}{\sqrt{80}} = -1.00 \Rightarrow |T| = 1.00$$

Siden 1.00 er mindre enn 1.96, får vi ikke forkasting av  $H_0$ , og vi kan formulere følgende konklusjon: «Dataene gir på nivå 5 % ikke sterkt grunnlag for å påstå at årlig fraflytting av ungdom var forskjellig i 1990 i forhold til perioden 1984-88.» ☺

## 7.9 Pearson's kjikvadrat-tilpasningstest

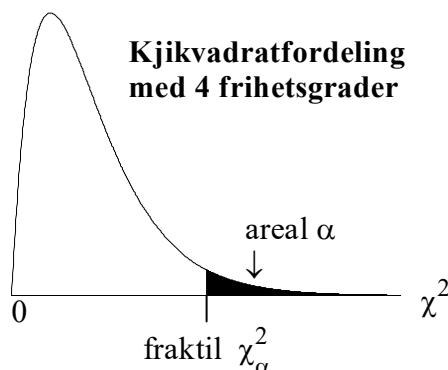
Pearson's kjikvadrat-tilpasningstest (etter Karl Pearson, 1857-1936), eller føyningstest, er en meget nyttig og anvendelig test som blant annet er mye brukt til å undersøke om det er rimelig å tilpasse en spesifisert diskret eller kontinuerlig fordeling til et gruppert tallmateriale. Testen er basert på kjikvadratfordelingen  $\chi^2$  med  $k$  frihetsgrader. Vi starter med å gi en kort beskrivelse av denne.

### Kjikvadratfordelingen med $k$ frihetsgrader

La  $Z_1, \dots, Z_k$  være  $k$  uavhengige  $N(0,1)$ -variabler, og la  $Y$  betegne kvadratsummen av de  $k$  variablene:

$$Y = \sum_{i=1}^k Z_i^2$$

Det kan da vises at  $Y$  er  $Kji2(k)$ -fordelt (se kap. 5). Et eksempel er vist i figuren nedenfor, som definerer øvre  $\alpha$ -fraktil,  $\chi_{\alpha}^2$ .



Når  $k$  er stor blir fordelingen tilnærmet normal.

Helt analogt med at  $z_\alpha$  var betegnelsen på øvre  $\alpha$ -fraktil i  $N(0,1)$ -fordelingen, lar vi  $\chi_\alpha^2$  betegne øvre  $\alpha$ -fraktil i  $\chi^2$ -fordelingen med  $k$  frihetsgrader:

$$P(Y > \chi_\alpha^2) = \alpha$$

Tabell over kjikvadratfordelingsfraktiler er gitt i tabell bakerst i boka.

### Testsituasjon

Vi tenker oss at vi har gruppert et statistisk tallmateriale  $x_1, \dots, x_n$  i  $k$  forskjellige grupper. De empiriske frekvensene (observasjonene) er  $O_1, \dots, O_k$ , dvs. vi har  $O_1$  observasjoner i gruppe 1,  $O_2$  i gruppe 2 osv., og  $O_1 + \dots + O_k = n$  = totalt antall observasjoner. Videre lar vi  $p_i$  betegne sannsynligheten for at en tilfeldig observasjon havner i gruppe nr.  $i$ , og vi lar  $E_1, \dots, E_k$  betegne de tilsvarende forventede antall observasjoner i hver gruppe:  $E_i = E(O_i) = np_i$ . Nullhypotesen  $H_0$  går da ut på å spesifisere verdiene til  $p$ -ene:

$$H_0: p_1 = p_{10}, \dots, p_k = p_{k0}.$$

Pearsons kjikvadrat-tilpasningstest er da gitt som følger:

#### Pearsons kjikvadrat-tilpasningstest

Tallmateriale:  $n$  tall gruppert i  $k$  grupper med  $O_i$  tall i gruppe nr.  $i$ .

Nullhypotese:  $H_0: p_1 = p_{10}, \dots, p_k = p_{k0}$

$$\text{Testobservator: } T = \sum_{i=1}^k \frac{(O_i - E_i)^2}{E_i} = \sum_{i=1}^k \frac{(O_i - np_{i0})^2}{np_{i0}}$$

$T$  er under  $H_0$  tn. Kji2( $k-1$ )-fordelt når alle  $E_i$ -verdier er større enn ca. 5.

Fork.område:  $R: T > \chi_\alpha^2$ , der  $\chi_\alpha^2$  er øvre  $\alpha$ -fraktil i Kji2( $k-1$ )-fordelingen.

## Kommentarer

Logikken bak testobservatoren er at dersom de antatte sannsynligheter  $p_0, \dots, p_k$  under  $H_0$  er vesentlig forskjellig fra de sanne sannsynligheter  $p_1, \dots, p_k$ , så vil dette gi seg utslag i en usannsynlig stor kvadratsum  $T$  i forhold til hva vi forventer under  $H_0$ . Forkastingsområdet blir derfor til høyre for fraktilen  $\chi_{\alpha}^2$ . Vi avslutter med et eksempel for å belyse hvordan testen utføres.

**Eks. 7.9** **Dusjforheng.** En forhandler av dusjforheng skal kjøpe inn store lager av ensfarga forheng i rødt, blått og gult. Før innkjøpet ønsker forhandleren i første omgang å teste om de tre forhengene er like populære i markedet. En rekke kunder forespørres. 42 svarer at de foretrekker det røde forhenget, 31 svarer blått og 27 svarer gult. 13 svarer at de har ingen sterk preferanse for en bestemt farge.

### Oppgave

Gir undersøkelsen grunnlag på nivå 5 % for å hevde at det er forskjellig sannsynlighet for at en tilfeldig kunde foretrekker rødt, blått eller gult forheng?

### Løsningsforslag

Vi tar kun med de kunder som har svart rødt, blått eller gult.  $H_0$  blir her:  $p_1 = p_2 = p_3 = 1/3$ . Vi har  $k = 3$  grupper. De nødvendige beregninger kan settes opp i følgende tabell:

Gruppe	Rød	Blå	Gul	Total
Frekvens O	42	31	27	100
Forventet frekvens E	33.3	33.3	33.3	99.9 <sup>1)</sup>
(O-E) <sup>2</sup> /E	2.27	0.16	1.19	$T = 3.62$

<sup>1)</sup> Summen av forventede frekvenser skal være lik summen n av observasjoner. Forskjellen mellom 99.9 og 100 skyldes avrundingsfeil.

Fra kjikvadrattabellen med  $\alpha = 0.05$  og  $k-1 = 2$  frihetsgrader finner vi at  $\chi_{\alpha}^2 = 5.99$ . Siden  $T = 3.62$  ikke er større enn  $\chi_{\alpha}^2 = 5.99$ , konkluderer vi med at det på nivå 5 % ikke er grunnlag for å hevde at de tre dusjforhengene ikke er like populære (ikke grunnlag for å forkaste  $H_0$ ). ☺

## 7.10 Oppgaver

**7.1** Gitt en normalfordeling med varians  $\sigma^2 = 400$ . Vi tar et tilfeldig utvalg bestående av 36 elementer, og beregner aritmetiske gjennomsnitt til  $\bar{x} = 159$ . Test hypotesen  $H_0: \mu = 150$  mot alternativet  $H_1: \mu > 150$  på 10 % nivået.

**7.2** Av et større vareparti kan det være feil med emballasjen på den enkelte artikkel. Vi antar at denne situasjonen kan beskrives ved hjelp av binomialfordelingen. Et tilfeldig utvalg bestående av 78 artikler gav at 35 hadde emballasjefeil. Test:

$H_0: p = 0.55$  mot  $H_1: p < 0.55$ , der  $p$  er sannsynligheten for emballasjefeil.

**7.3** Flyttet til oppgave 8.9.

**7.4**  $S$  er Poisson-fordelt med parameter  $\lambda$ . Man skal teste nullhypotesen

$H_0: \lambda = 5$  mot alternativet  $H_1: \lambda > 5$ .

- Konstruer den testen som har signifikansnivå nærmest 0.06
- Anta at  $\lambda = 8$ . Hva er sannsynligheten for feil av type II ( $\beta$ ) ved den funne test under punkt a)?

c) Finn sannsynligheten for å forkaste nullhypotesen ved den testen du fant under punkt a), når den sanne verdi av  $\lambda$  er

- 6,
- 7,
- 8,
- 9

Skisser testens styrkekurve.

**7.5** En kjøttprodusent reklamerer med at de pølsevarer de selger inneholder høyest 12 % fett. Det er mistanke om at

det faktiske fettinnholdet er høyere, og kjøttskontrollen foretar derfor en test av produsentens pølseprodukter. Sett

$X$ : Fettinnholdet i en tilfeldig produksjon og anta at  $X$  er  $N(\mu, 2)$

Vi forutsetter at erfaring har vist at fettinnholdet har et standardavvik på 2 prosentenheter. En undersøkelse av ti produksjoner gav følgende data for fettinnholdet i prosent:

13 11 15 14 17 16 11 13 12 14

$$\bar{x} = 13.6$$

- Test om fettinnholdet er 12 % mot at det er høyere. Velg signifikansnivå 0.05. Uttrykk i ord hva feil av type I i dette tilfellet innebærer.
- Anta det er riktig at produsenten fusker med fettinnholdet og at forventet fettinnhold er 14 %. Hva er sannsynligheten for feil av type II? Hva er testens styrke?
- Vi ønsker å utføre testen med et signifikansnivå på 0.01. Hva blir nå svarene på punkt b)? Sammenlign og kommenter forskjellene.
- Kjøttskontrollens statistiker synes at svaret i c) gir en for høy sannsynlighet for feil av type II. Han ønsker å øke stikkprøvens størrelse  $n$ , slik at sannsynligheten for feil av type II høyest er 0.10, men under forutsetning av at signifikansnivået fremdeles er 0.01. Hvor stor må  $n$  være for at disse krav skal være oppfylt?

**7.6** Støygrensen for tunge lastebiler er satt til 83 desibel. La  $\mu$  være midlere

støynivå for alle tunge lastebiler. Et tilfeldig utvalg på 6 tunge lastebiler ble observert med følgende støynivåer:

85.4 86.8 86.1 85.3 84.8 86.0

Det antas at desibelnivåene for de forskjellige lastebilene er uavhengige, normalfordelte med forventning  $\mu$  og ukjent varians  $\sigma^2$ .

- Gir resultatene grunnlag for å påstå at  $\mu > 83$ ? Formuler nullhypotese og alternativhypotese for å besvare dette spørsmålet. Bruk en test med signifikansnivå  $\alpha = 0.05$ .
- Konstruer et 90 % konfidensintervall for  $\mu$ .

**7.7** Antall fellingstillatelser som blir gitt for elg hver høst i Nordskogen avhenger av størrelsen,  $N$ , på elgpopulasjonen i skogen. Ett år frykter man at populasjonen er så liten at den ikke tåler beskatning, og ingen fellingstillatelser vil bli gitt hvis  $N \leq 12$ .

For å vurdere populasjonsstørrelsen har en på et tidligere tidspunkt merket 6 elger. Nå vil man fange 6 elger på nytt og telle hvor mange av dem,  $X$ , som er merket. Merk at ingen elger slippes ut igjen før antall merkede er talt opp.

- Nullhypotesen:  $H_0: N \leq 12$  skal testes mot  $H_1: N > 12$ . Som testobservator skal vi bruke  $X$ . Skal vi forkaste for store eller små verdier av testobservatoren? Svaret skal begrunnes.
- Hva slags sannsynlighetsfordeling har testobservatoren?
- Bestem forkastingsområdet for en test med signifikansnivå ca. 4 %.

**7.8** En presidentkandidat i USA påstår at han har støtte fra 60 % av velgerne. Anta at kandidatens påstand er korrekt.

- Et tilfeldig utvalg på 10 velgere tas. Hva er sannsynligheten for at minst 5 av disse støtter kandidaten?
- Anta at 100 velgere trekkes tilfeldig. Hva er sannsynligheten for at minst 50 av disse støtter kandidaten?
- For å fremlegge «statistisk bevis» for sin påstand, ber kandidaten et meningsmålingsinstitutt foreta en undersøkelse med et tilfeldig utvalg på 900 velgere. Det viser seg at av disse støtter 567 kandidaten. Er dette tilstrekkelig til å si at mer enn 60 % av velgerne støtter kandidaten? Bruk signifikansnivå 5 %.

Finn signifikanssannsynligheten  $P^*$ , og forklar hva den betyr.

**7.9** En fabrikk produserer skruer der skruediametrerne  $X$  (mm) antas å være normalfordelt med forventning  $\mu$  og standardavvik  $\sigma$ . Skruene skal normalt være 2 mm i diameter, og et standardavvik på høyst 0.04 mm er akseptabelt. Gjennom en viss tid har en observert at skruediameteren er større enn 2 mm, og en ønsker å teste påstanden at  $\mu > 2$ . Det tas et tilfeldig utvalg på 9 skruer, der disse har følgende diameter (i mm):

2.00 2.08 2.10 2.12 2.15 2.15 2.10  
2.05 2.15

- Regn ut gjennomsnittet  $\bar{x}$  og empirisk standardavvik ( $n-1$  i nevner).
- Formuler nullhypotese og alternativhypotese, og utfør testen med  $\alpha =$

5 % nivå. Testkonklusjonen formuleres med ord.

- c) Forklar hva  $\alpha$  betyr.
- d) Hva er forskjellen på sannsynlighetsfordelingen til diameteren på en tilfeldig valgt skrue, og sannsynlighetsfordelingen til gjennomsnittet av 9 tilfeldig valgte skruer?

**7.10** Vi kaster en mynt 50 ganger. I 18 av kastene ble det krone og i 32 mynt. Hvis vi bruker et 5 % signifikansnivå, er det da noen grunn til å tvile på at mynten er i orden? (Formuler  $H_0$  og  $H_1$  og utfør en passende test).

**7.11** En boks inneholder fire kuler. En kule kan være sort eller hvit. Boksen inneholder en av disse 5 kombinasjonene: Fire sorte, fire hvite, en sort og tre hvite, en hvit og tre sorte eller to sorte og to hvite. Du skal teste hypotesen at det er to av hver farge. Du velger ut to uten tilbakelegging og konkluderer med at det ikke er to av hver farge dersom begge de utvalgte er av samme farge.

Hva er i denne situasjonen:

- a) Feil av type I?
- b) Feil av type II?
- c) Hva er sannsynligheten for feil av type I?
- d) Hva er sannsynligheten for feil av type II hvis det er 3 sorte kuler og en hvit kule?

**7.12** En bedrift produserer fiskesnører. Kvaliteten på snørene måles ved bruddstyrken, som er den maksimale strekkraften snøret tåler uten å briste. Bedriften har en gjennomprøvet pro-

duksjonsprosess, hvor bruddstyrken er normalfordelt med forventning 20.00 kg og standardavvik 0.90 kg. Bedriften skal utprøve en ny og kostnadsbesparende produksjonsprosess. Den nye prosessen kan ikke produsere snører med en større forventet bruddstyrke enn 20.00 kg (som den burde være), men bedriften har mistanke om at forventningsverdien er lavere enn dette. Av produksjons-tekniske grunner kan bedriften anta at bruddstyrken ved den nye produksjonsprosessen også er normalfordelt med standardavvik 0.90 kg. La  $\mu$  betegne forventet bruddstyrke ved den nye produksjonsprosessen. Bedriften vil teste

$$H_0: \mu = 20.00 \text{ mot } H_1: \mu < 20.00$$

- a) Det innhentes en stikkprøve ved å måle bruddstyrken til 20 snører fra den nye produksjonsprosessen. Gjenomsnittlig bruddstyrke er 19.50. Utfør hypoteseprøvingen ovenfor på 5 % nivå. Hva er testens konklusjon? Finn signifikanssannsynligheten.

Anta at forventningen til bruddstyrken med den nye produksjonsprosessen er 19.50.

- b) Beregn sannsynligheten  $\beta$  for feil av type II.
- c) Bedriften mener at sannsynligheten for feil av type II i b) er for høy. Den kan redusere sannsynligheten for feil av type II ved å øke signifikansnivået  $\alpha$ . Hva blir  $\alpha$  dersom testen tilpasses slik at  $\beta = 0.10$ ?
- d) Illustrer sammenhengen mellom punktene a), b) og c) i en figur.

- e) Dersom bedriften ønsker å redusere sannsynligheten for feil av type II uten å gå på bekostning av signifikansnivået, må den øke stikkprøvens størrelse. Hvor stor må stikkprøven være dersom sannsynligheten for feil av type II skal være høyst 0.10 og signifikansnivået høyst 0.05?

**7.13** En foreleser ved BI vil undersøke om 10 minutters-pausen mellom to forelesningstimer overholdes. Tiden på 15 tilfeldig utvalgte pauser var:

14, 14, 8, 12, 13, 10, 14, 10, 11, 13, 9, 10, 15, 12, 15.

- a) Test  $H_0: \mu = 10$  mot  $H_1: \mu \neq 10$  på 5 %-nivået.  
 b) Kan du trekke noen konklusjoner av dette? I tilfelle hvilke?

**7.14** Ifølge statistisk årbok i 1982, var gjennomsnittshøyden for de i alt 34 027 vernepliktige i 1982 179.4 cm, mens det i 1981 var 33 337 vernepliktige med en gjennomsnittshøyde på 179.5 cm. Begge år var standardavviket 6.6 cm. Det har i moderne tid aldri skjedd en nedgang i gjennomsnittlig rekruthøyde. Undersøk om nedgangen denne gang er signifikant på 5 %-nivået. Diskuter mulige feilkilder.

**7.15** Fra legehold er det hevdet at 10 % av alle diagnosene er feilstilte. Anta det blir iverksatt tiltak rettet mot å forbedre diagnosestillingen. Etter at tiltakene er gjennomført, blir 200 diagnosene tilfeldig valgt ut, og det blir undersøkt om de var gale eller riktige. Resultatene viser at 13 av de 200 diagnostene var feilstilte. Test om det

har vært en signifikant nedgang i antall feilstilte diagnosene. Velg signifikansnivå 0.05.

**7.16 (E)** En bedrift fremstiller artikler med en kvalitet  $X$  som har forventning  $\mu$  og standardavvik  $\sigma$ . Hvis maskinen er riktig innstilt, skal  $\mu$  være lik 4.3. Ved en kontrollmåling av 9 tilfeldige artikler fant en følgende verdier på kvaliteten:

4.4 4.7 4.0 4.6 4.6 4.5 4.2 4.7 4.5

Anta at  $\sigma = 0.22$  er kjent.

- a) Lag et konfidensintervall med konfidensgrad 95 % for  $\mu$ . Hvilke forutsetninger må du gjøre?

Fomannen på avdelingen har en tid hatt mistanke om at  $\mu$  er større enn 4.3, og at maskinen dermed bør justeres.

- b) Formulermannens problemstilling som et hypoteseprovingsproblem, og gjennomfør testen på grunnlag av kontrollmålingene. Bruk signifikansnivå 5 %.

Bestem testens styrkefunksjon, og beregn styrken når  $\mu = 4.5$ .

Anta nå at vi ikke har tillit til at  $\sigma = 0.22$  og anser  $\sigma$  som ukjent.

- c) Test hypotesen i b) under denne forutsetningen.

- d) Bestem til slutt på grunnlag av kontrollmålingene et 99 % konfidensintervall på formen  $[0, b]$  for  $\sigma$  ( $\mu$  antas ukjent).

**7.17 (E)** En metallurg skal bestemme smeltepunktet  $\mu$  for en legering. Han vet av erfaring at gjentatte målinger av smeltepunktet kan oppfattes som uav-

hengige og normalfordelte stokastiske variabler med forventning  $\mu$  og standardavvik  $\sigma = 2.0 \text{ } ^\circ\text{C}$ .

Metallurgen utfører 8 målinger og får følgende resultater (i  $^\circ\text{C}$ ):

1468.5 1469.0 1471.0 1470.0  
1469.5 1467.5 1469.0 1472.0

- Bestem et 95 % konfidensintervall for  $\mu$  basert på de 8 målingene.
- Metallurgen synes at intervallets lengde er noe stor. Hvor mange målinger måtte han minst gjort for å få et intervall som var kortere enn  $2.0 \text{ } ^\circ\text{C}$ ?
- Metallurgen er interessert i om det er grunnlag for å hevde at  $\mu > 1468.0 \text{ } ^\circ\text{C}$ . Formuler hans problem med en passende nullhypotese og alternativ.

Konstruer testen med 5 % signifikansnivå. Hva blir konklusjonen ut fra de målte verdiene?

- Bestem styrkefunksjonen for testen i c) og lag en skisse av den. Finn styrken for  $\mu = 1470.0 \text{ } ^\circ\text{C}$ , og forklar med ord hva det betyr at styrken har denne verdien.
- Forklar kort hvilke modifiseringer vi måtte gjøre i a) og c) dersom vi ikke hadde kjent  $\sigma$ , og gjennomfør testen i c) også i dette tilfelle.

**7.18 (E)** Ved en bedrift produseres en bestemt type artikler. Sannsynligheten  $p$  for at en tilfeldig artikkel er defekt ligger normalt på ca. 0.03. Dersom defektsannsynligheten øker til mer enn 0.05, må produksjonsutstyret justeres.

a) Ved en kontroll av produksjonsutstyret undersøkte man en dag 40 tilfeldig valgte artikler. La  $X$  være antall defekte i utvalget. Som estimator for  $p$  bruker vi  $p^* = X/40$ .

Vis at  $p^*$  er forventningsrett, og bestem variansen til  $p^*$ .

- Kontrollen viste at 3 av de 40 artiklene var defekte. Estimer  $p$  og standardavviket for  $p$ .
- Gir det observerte resultatet grunn til å hevde at  $p > 0.05$ , dvs. at produksjonsutstyret bør justeres? (Signifikansnivå 5 %).

**7.19 (E)** Ved en bensinstasjon er det ukentlige salget  $X$  av bensin, normalfordelt med forventning  $\mu$  og standardavvik  $\sigma = 500$  liter.

Anta at  $\mu = 8000$  liter er kjent.

- Finn sannsynligheten for at bensinstasjonen selger over 8500 liter en uke.
- Hvor stort lager må stasjonen ha for at man skal være 95 % sikker på at lageret ikke skal tømmes i løpet av en uke?
- Anta at salget hver uke er uavhengig av salget andre uker. Finn sannsynligheten for at det totale salget i løpet av 4 uker er mindre enn 30 000 liter.

Innehaveren av bensinstasjonen har fått mistanke om at det ukentlige salget har sunket (dvs. at  $\mu < 8000$  liter), og bestemmer seg for å undersøke dette nærmere. Han noterer bensinsalget i 9 uker og finner gjennomsnittet  $\bar{x} = 7800$  liter.

- d) Er det på grunnlag av dette resultatet grunn til å tro at  $\mu < 8000$  liter? Formuler problemet som et hypotesetestingsproblem og gjennomfør testen med 5 % signifikansnivå ( $\sigma$  antas kjent lik 500 liter).
- e) Beregn testens styrke dersom  $\mu = 7500$  liter.

**7.20 (E)** La  $X$  være vekten til en tilfeldig student ved tidligere BIH.  $X$  antas normalfordelt  $N(\mu, \sigma)$  der  $\mu = 73$  kg og  $\sigma = 7$  kg.

- a) Finn sannsynligheten for at en tilfeldig student veier mellom 60 og 75 kg.

Anta at vi har et tilfeldig utvalg på 50 studenter. La  $Y$  være antallet av disse som veier mellom 60 og 75 kg.

- b) Hvilken sannsynlighetsfordeling får  $Y$ ? Finn forventning og varians til  $Y$ . Finn tilnærmet  $P(Y > 35)$ .
- c) En av heisene er beregnet å tåle vekten 800 kg. Hvor mange studenter kan ta heisen samtidig dersom vi krever at sannsynligheten for overbelastning skal bli mindre enn eller lik 0.05?

Heisselskapet syntes de fikk for mange reparasjoner på heisen på grunn av overbelastning, selv om studentene fulgte reglene for maksimalt antall personer i heisen. De mente at dette måtte skyldes at studentene veide mer enn forutsatt, dvs. at  $\mu > 73$  kg. For å kontrollere dette ble 10 tilfeldige personer kontrollveid. Tilsammen veide de 785 kg.

- d) Lag en test med 5 % signifikansnivå for å avgjøre om heisselskapet har rett.
- e) Beregn testens styrke dersom  $\mu = 80$  kg.

**7.21 (E)** Resistansen i en motstandsstråd er normalfordelt  $N(\mu, \sigma)$  med  $\mu = 23.4 \Omega$  og  $\sigma = 0.25 \Omega$ .

- a) Hva er sannsynligheten for at en motstand har en resistans større enn  $24.0 \Omega$ ?
- b) Hva er sannsynligheten for at forskjellen i absoluttverdi mellom resistansen til to motstander er større enn  $0.5 \Omega$ ?

Motstandene kommer i pakker på 100 stk. De skal brukes i enheter hvor resistansen må ligge mellom  $23.0 \Omega$  og  $24.0 \Omega$ . Bedriften har avtale med leverandøren om å returnere pakken dersom mer enn 5 stk. ikke holder mål.

- c) Beregn tilnærmet sannsynligheten for at pakken må returneres.

I praksis tar bedriften en stikkprøve på 5 stk, dersom det er mer enn 1 motstand som ikke holder sendes pakken tilbake. Anta at det i en pakke er 7 motstander som ikke holder.

- d) Bestem sannsynligheten for at bedriften sender pakken tilbake.

I en periode har bedriften måttet sende mange pakker tilbake. De mener at  $\mu > 23.4 \Omega$ . De tok derfor en stikkprøve på 10 motstander og fant følgende verdier for resistansen (målt i  $\Omega$ ):

23.2 24.7 23.8 24.2 24.8  
23.2 25.1 23.9 23.4 23.6

- e) Lag en test med 5 % signifikansnivå for å avgjøre om bedriften har rett.

**7.22 (E)** Ved bruk av et bestemt medisinsk preparat A får gjennomgående halvparten av pasientene uheldige ettervirkninger, dvs.  $p = P(\text{ettervirkninger}) = 0.5$ .

- a) Anta at 10 pasienter bruker medisinen. La  $X$  være antall pasienter som får uheldige ettervirkninger. Hvilken sannsynlighetsfordeling får  $X$ ? Finn  $E(X)$  og  $\text{Var}(X)$ . Finn sannsynligheten for at 3 eller færre får uheldige ettervirkninger.
- b) En modifisert versjon B av preparatet skal utprøves. Det påstås at de uheldige ettervirkningene nå er redusert, dvs.  $p < 0.5$ . For å sjekke denne påstanden prøves B på  $n = 10$  pasienter. Det viser seg at 3 av pasientene får ettervirkninger.

Kan vi på grunnlag av dette resultatet påstå at sannsynligheten for ettervirkninger er redusert?

Formuler spørsmålet som et hypotesetestingsproblem, og besvar ved hjelp av beregninger gjort i a) (5 % signifikansnivå).

**7.23 (E)** Produsenten av en bestemt type tabletter merker tabletene med et tall  $\mu$ , som angir den mengde aktivt stoff hver tablet skal inneholde. På grunn av ukontrollerbare variasjoner i produksjonsprosessen, vil imidlertid det eksakte innhold aktivt stoff ikke være  $\mu$ , men må oppfattes som en stokastisk variabel  $X$  med (kjent) forventning  $\mu$  og varians  $\sigma^2$ . Vi antar at  $X$  er normalfordelt.

Anta  $\mu = 100 \text{ mg}$  og  $\sigma = 2 \text{ mg}$ .

- a) En tabletter anses som defekt dersom mengden aktivt stoff avviker mer enn 4 mg fra  $\mu$ . Finn sannsynligheten for at en tilfeldig tabletter er defekt.
- b) Tablettene pakkes på brett med 10 tabletter pr. brett. La  $S$  være total mengde aktivt stoff i de 10 tabletene. Hvilken sannsynlighetsfordeling får  $S$ ? Finn sannsynligheten for at total mengde aktivt stoff i de 10 tabletene overstiger 1010 mg.
- c) Brettene selges i esker med 5 brett i hver eske. La  $Y$  være antall brett i en eske der totalmengden av aktivt stoff overstiger 1010 mg. Hvilken sannsynlighetsfordeling får  $Y$ ? Finn sannsynligheten for at minst 2 av brettene i en eske har mer enn 1010 mg aktivt stoff.
- d) Produsenten av tabletene tar daglig stikkprøver fra produksjonen for å kontrollere verdien av  $\mu$ . 10 tabletter tas tilfeldig fra dagens produksjon, og blir analysert i bedriftens laboratorium. Dersom det er grunn til å tro at  $\mu \neq 100 \text{ mg}$  må produksjonsprosessen stoppes og justeres. En dag ble følgende verdier (mg) målt for mengden av aktivt stoff:

103 98 99 104 101  
100 104 99 102 103

Er det ut fra disse verdiene grunn til å tro at  $\mu \neq 100 \text{ mg}$ ?

Formuler spørsmålet som et hypoteseprøvingsproblem og test med 5 % signifikansnivå.

- e) Bestem testens styrkefunksjon. Hva er sannsynligheten for at prosessen blir stanset og justert dersom  $\mu$  i virkeligheten er 102 mg?
- f) Hvordan ville du utført testen i d) dersom  $\sigma$  var ukjent? Utfør testingen også i dette tilfellet.

**7.24 (E)** Noen fysikere var uenige om størrelsen på lyshastigheten. En gruppe, «de ortodokse», påstod at den var 300000 km/s, mens en annen, «skeptikerne», hevdet at det var helt usannsynlig at den var «et så rundt tall», og trodde altså ikke på «de ortodokses» påstand. Man ble enige om å utføre 9 målinger, og at måleresultatene kunne antas å være uavhengige, forventningsrette og normalfordelte  $N(\mu, 300)$ .

- a) Sett opp en nullhypotese og en alternativhypotese, og forklar hva det vil si å gjennomføre en test med signifikansnivå 5 %.

- b) Formuler testen og gjennomfør den med følgende måleresultater:

300508	299150	299350
300137	299891	298592
300015	299799	298937

- c) Bestem testens styrkefunksjon og beregn styrken for  $\mu = 299700$ , 299900 og 300100. Forklar hva styrken uttrykker.

- d) Det ble uenighet om hvor egnet testen var. Man besluttet derfor å foreta nye målinger. Man krevde at testen fortsatt skulle ha signifikansnivå  $\alpha = 0.05$ , men nå skulle styrken være minst 0.95 i  $\mu = 299700$ . I mellomtiden hadde «de

ortodokse» med et annet eksperiment slått fast at lyshastigheten i hvert fall ikke var høyere enn 300000 km/s, og det godtok skeptikerne.

- Hva er en rimelig mothypotese nå?
- Hvor stor må  $n$ , antall målinger, være i den nye testen?

**7.25 (E)** Forbrukerkontoret i en bestemt by har mottatt klager på en bestemt pizzaprodusent. Denne produsenten hevder at deres store pepperonipizzaer i gjennomsnitt inneholder 60 gram pepperoni. Publikum mener at vekten av pepperoni må være betydelig lavere. En konsulent ved forbrukerkontoret får i oppdrag å utføre en hypotesetest. Vekten av pepperoni (pr. pizza) antas normalfordelt med forventningsverdi  $\mu$  og standardavvik på 15 gram. Hypotesene som skal testes er:

$$H_0: \mu = 60 \text{ mot } H_1: \mu < 60$$

Konsulenten velger ut én pizza for inspeksjon.

- a) Foreta hypotesetest med signifikansnivå  $\alpha = 0.05$ . Hva blir konklusjonen dersom den observerte pepperonivekten er 39 gram?
- b) Finn det minste signifikansnivået som gjør at vi forkaster nullhypotesen med en pepperonivekt på 39 gram.
- c) La  $\alpha = 0.05$ . Beregn styrken til testen i de tilfellene der  $\mu = 30$ ,  $\mu = 39$  og  $\mu = 50$  (gram). Skissér styrkefunksjonen. Hvilken informasjon gir styrkefunksjonen?

**7.26 (E)** I en befolkningsgruppe antar vi at fødselsvekten  $X$  (i kg) til et vilkårlig nyfødt barn er å betrakte som en normalfordelt variabel. Gjennomsnittet  $\mu$  i fordelingen antas å være 3.5 kg og standardavviket  $\sigma = 0.35$  kg. Man har en stund hatt mistanke om at den delen av befolkningsgruppen som bor i Tåkedalen, et område med mye forurensning, får barn med lavere fødselsvekt enn den øvrige delen av befolkningsgruppen. For å undersøke dette nærmere, registrerte man fødselsvekten til 1000 Tåkedal-barn over en tidsperiode. Dette resulterte i en gjennomsnittsvekt på  $\bar{x} = 3.48$  kg. Anta at vektene til Tåkedal-barna er uavhengige og  $N(\mu, 0.35)$ -fordelte.

- Foreta hypotesetest av  $H_0: \mu = 3.50$  mot  $H_1: \mu < 3.50$  med signifikansnivå på 1 %. Hva blir konklusjonen med  $\bar{x} = 3.48$  kg?
- Hvor mange barn må undersøkes hvis en ønsker at  $\mu$  med 95 % sikkerhet skal ligge innenfor et konfidensintervall med bredde 0.02?
- La oss se bort fra signifikansnivået i a). Hva er det minste signifikansnivået som gjør at en kan forkaste  $H_0$  til fordel for  $H_1$  når  $\bar{x} = 3.48$  kg?

**7.27 (E)** En brusautomat er konstruert slik at den - hvis den fungerer riktig - gir porsjoner på 218 g. På grunn av tilfeldigheter varierer imidlertid porsjonene fra gang til gang, slik at de følger en normalfordeling med standardavvik på 6,2 g. I det siste har mange kunder klaged og hevdet at automaten gir for lite brus. Ledelsen

bestemmer seg for å undersøke om klagene er berettiget, og foretar derfor kontrollveiing av 16 tilfeldig valgte porsjoner.

- Hvilken nullhypotese og mothypotese bør ledelsen teste i denne situasjonen?

Ledelsen velger å gjennomføre en test med signifikansnivå  $\alpha = 0.05$ . Forklar hva dette betyr.

- Formulér testen.

Gjennomfør testen gitt at gjennomsnittsvekten til de 16 porsjonene er 214.6 g.

Gjennomfør også testen for det tilfellet der gjennomsnittsvekten er 216.5 g.

- Beregn styrken til testen i punktet 216. Hva forteller svaret deg?

Beregn dessuten styrken til testen i punktene 212 og 214, og skissér styrkefunksjonen. Hva innebærer det at styrkefunksjonen er strengt avtagende i dette tilfellet?

- Noen i ledelsen synes at testen er for usikker og for lite følsom. De foreslår i stedet å velge signifikansnivået  $\alpha = 0.01$  og dessuten kreve at styrken av testen i punktet 216 skal være minst 0.90. Formulér den nye testen, og bestem det minste antall porsjoner vi i dette tilfellet må kontrollveie. Kan vi nå forkaste nullhypotesen hvis gjennomsnittsvekten av de kontrollveide porsjonene er 216.5 g?

**7.28** En bilfabrikk hevder at motoren i deres nyeste modell yter 135 hk i gjennomsnitt ved turtallet 4000

omdr./min., når anbefalt drivstoff benyttes. Noen bilkjøpere tror at ytelsen er lavere, og ber et frittstående bilstestefirma måle motorstyrken i henhold til fabrikkens spesifikasjoner. 16 tilfeldig utvalgte biler av denne modellen ble undersøkt. Firmaet fant at motorstyrken i gjennomsnitt var 128 hk. Anta at motorstyrken er normalfordelt med forventning  $\mu$  og standardavvik  $\sigma = 14$  hk.

- a) Formulér en passende null-hypotese og en mothypotese. Kan bilkjøperne påstå at de har rett dersom de er villige til å ta en feilrisiko på 1 %?
- b) Beregn testens styrke for  $\mu = 128$  hk.
- c) Hvor mange biler må tas med i undersøkelsen for at styrken for  $\mu = 128$  hk skal bli minst 90 %, når signifikansnivået fortsatt skal være 1 %?

## 7.11 Formelsamling

### Betegnelser

- $H_0$  = nullhypotese  
 $H_1$  = alternativ hypotese (den vi ønsker å underbygge)  
 $n$  = antall i utvalg  
 $\theta$  = hypotesetest-parameter (eks:  $\mu$  eller  $p$ ).  
 $T$  = testobservator  
 $R$  = forkastingsområde  
 $\alpha(\theta) = P(R \mid H_0) = P(\text{Type I feil})$   
 $\beta(\theta) = P(R^C \mid H_1) = P(\text{Type II feil})$   
 $\gamma(\theta) = P(R \mid \theta) = \text{styrkefunksjon}$   
 $\alpha$  = testnivå (signifikansnivå)  
 $P^*$  = signifikanssannsynlighet  
 $z_\alpha$  = øvre  $\alpha$ -fraktil i  $N(0,1)$ -fordeling.  
 $t_\alpha$  = øvre  $\alpha$ -fraktil i  $t$ -fordelingen med  $n-1$  frihetsgrader.  
 $\bar{X}$  = middelverdi:  $(X_1 + \dots + X_n) / n$   
 $S$  = empirisk standardavvik (sum fra  $i=1$  til  $i=n$ ):  

$$\sqrt{\left(\sum X_i^2 - \frac{1}{n} (\sum X_i)^2\right) / (n-1)}$$
  
 $p^*$  =  $X/n$ , der  $X$  er antall  $J$ -utfall av  $n$  Bernoulli-forsøk.  
 $O_i$  = antall observasjoner i gruppe  $i$   
 $\mu$  =  $E(X_i)$ ,  $i = 1, 2, \dots, n$   
 $\sigma$  =  $\text{std}(X_i)$ ,  $i = 1, 2, \dots, n$   
 $\chi^2$  = øvre  $\alpha$ -fraktil i  $\text{Kji}2(k-1)$ -fordelingen

**Test for  $\mu, \sigma$  kjent**

$$T = \frac{\bar{X} - \mu_0}{\sigma / \sqrt{n}}$$

$H_0: \mu \leq \mu_0$	$H_0: \mu \geq \mu_0$	$H_0: \mu = \mu_0$
$H_1: \mu > \mu_0$	$H_1: \mu < \mu_0$	$H_1: \mu \neq \mu_0$
$R: T > z_\alpha$	$R: T < -z_\alpha$	$R:  T  > z_{\alpha/2}$

### Test for $\mu, \sigma$ ukjent, $n < 30$ :

$$T = \frac{\bar{X} - \mu_0}{S / \sqrt{n}}$$

$H_0: \mu \leq \mu_0$	$H_0: \mu \geq \mu_0$	$H_0: \mu = \mu_0$
$H_1: \mu > \mu_0$	$H_1: \mu < \mu_0$	$H_1: \mu \neq \mu_0$
$R: T > t_\alpha$	$R: T < -t_\alpha$	$R:  T  > t_{\alpha/2}$

### Test for $\mu, \sigma$ ukjent, $n \geq 30$

$$T = \frac{\bar{X} - \mu_0}{S / \sqrt{n}}$$

$H_0: \mu \leq \mu_0$	$H_0: \mu \geq \mu_0$	$H_0: \mu = \mu_0$
$H_1: \mu > \mu_0$	$H_1: \mu < \mu_0$	$H_1: \mu \neq \mu_0$
$R: T > z_\alpha$	$R: T < -z_\alpha$	$R:  T  > z_{\alpha/2}$

### Test for $p, np(1-p) > 5$

$$T = \frac{p^* - p_0}{\sqrt{p_0(1-p_0)/n}} = \frac{X - np_0}{\sqrt{np_0(1-p_0)}}$$

$H_0: p \leq p_0$	$H_0: p \geq p_0$	$H_0: p = p_0$
$H_1: p > p_0$	$H_1: p < p_0$	$H_1: p \neq p_0$
$R: T > z_\alpha$	$R: T < -z_\alpha$	$R:  T  > z_{\alpha/2}$

### Pearsons kjikvadrat-tilpasningstest

$$H_0: p_1 = p_{10}, p_2 = p_{20}, \dots, p_k = p_{k0}$$

$$T = \sum_{i=1}^k \frac{(O_i - E_i)^2}{E_i} = \sum_{i=1}^k \frac{(O_i - np_{i0})^2}{np_{i0}}$$

$$R: T > \chi^2$$

## Kapittel 8

# To populasjoner

### 8.1 Innledning

Vi har valgt overskriften «to populasjoner» for enkelhets skyld. Et alternativ hadde vært «sammenligning av to behandlinger», der ordet behandling (treatment) er mye brukt i den statistiske litteraturen. Kort fortalt er forskjellen på de 2 populasjonene vi vil sammenligne at den ene populasjonen er resultatet av den ene behandlingen, og den andre populasjonen er resultatet av den andre behandlingen.

Vi starter med et belysende eksempel:

**Eks. 8.1** **EDB-system.** Si at skattemyndighetene skal innføre et nytt EDB-system, og ønsker å sammenligne brukervennligheten til 2 forskjellige systemer. Forhandlerne av de 2 systemene får holde innføringskurs av lik varighet for utvalgte medarbeidere i skatteetaten. Etterpå avlegges individuell skriftlig prøve, der poengsummen blir brukt som mål på brukervennlighet. Vi lar  $\mu_1$  betegne forventet poengsum ved system 1, og  $\mu_2$  betegne forventet poengsum med system 2. Det systemet som har høyest forventning defineres som det mest brukervennlige. ☺

Vi skal her begrense oss til å sammenligne populasjonsforventninger ved å se på konfidensintervall og hypotesetester for differansen  $\mu_1 - \mu_2$ . Fremgangsmåten blir da helt analog med konfidensintervall og hypotesetesting fra kap. 6 og 7.

Et viktig spørsmål vi skal ta opp er følgende:

### Hvordan bør vi foreta våre utvalg?

Vi skal skissere 2 forskjellige metoder å foreta utvalgene på. Metodene har sine fortrinn i ulike situasjoner. Felles for begge metoder er at utvalgene i størst mulig grad må være **tilfeldige**. Dette er nærmere utdypet i avsnitt 8.2 om tilfeldiggjøring (randomisering). Den ene metoden kaller vi **uavhengige tilfeldige utvalg** (fra 2 populasjoner), mens vi skal kalle den andre for **parvis sammenligning**. De 2 metodene er kort forklart nedenfor, med henvisning til EDB-eksemplet.

**Uavhengige tilfeldige utvalg** (fra 2 populasjoner).

Populasjon 1 består her av resultatene (dvs. poengsummene) av prøvene til alle potensielle EDB-brukere i skatteetaten, dersom alle gjennomgikk innføringskurset for system 1. Middelverdien til alle disse tenkte resultatene ville da vært identisk med populasjonsforventningen,  $\mu_1$ . I praksis må vi estimere  $\mu_1$  på basis av et begrenset utvalg fra populasjon 1.

Tilsvarende består populasjon 2 av resultatene av prøvene til alle potensielle EDB-brukere i skatteetaten, dersom alle gjennomgikk innføringskurset for system 2. Legg merke til at *individene* er de samme, selv om populasjonene er forskjellige.

Når vi foretar tilfeldige utvalg fra de 2 populasjonene, kan vi tenke oss dette gjort ved at vi trekker tilfeldige navn fra et register over alle potensielle EDB-brukere i skatteetaten. Først trekker vi f.eks. 20 tilfeldige personer som skal prøve system 1, og så 20 nye tilfeldige personer som skal prøve system 2.

Kort fortalt er en stor *fordel* med denne fremgangsmåten at vi er sikret tilfeldige utvalg fra populasjonene som helhet. Utvalgene blir dermed (statistisk sett) bredest mulig sammensatt med hensyn til alder, bosetting, kjønn og andre faktorer som kan tenkes å påvirke responsen på EDB-systemene. Vi sikrer oss også at «utenforliggende» faktorer (faktorer som ikke har noe med EDB-systemenes brukervennlighet å gjøre) ikke får noen systematisk påvirkning på resultatene.

En *ulempe* med metoden er at variansen til hvert av utvalgsmidlene kan bli unødig stor på grunn av faktorer som ikke har noe med brukervennligheten til EDB-systemet å gjøre. Det er f.eks. naturlig å anta at EDB-brukere i skatteetaten med EDB-bakgrunn vil ha høyere poengsum ved prøven enn potensielle EDB-brukere som aldri har vært borti EDB. Slike systematiske forskjeller vil i verste fall bidra til at selv om det ene EDB-systemet skulle være betydelig mer brukervennlig enn det andre systemet, så vil denne forskjellen «drukne» i andre faktorer. Denne ulempen er det man børter på ved å foreta parvis sammenligning (se nedenfor).

## Parvis sammenligning

Ved parvis sammenligning deler man de potensielle forsøksenheterne inn i **homogene (ensartede) blokker**, og så trekker man et tilfeldig par fra hver blokk. Deretter får den ene i hvert par den ene behandlingen, mens den andre i paret får den andre behandlingen. I EDB-eksemplet kan vi som et enkelt eksempel tenke oss at vi deler de ansatte i 5 alders-grupper, 2 kjønnsgrupper og 2 erfaringsgrupper (de med og uten EDB-erfaring). Dette blir tilsammen  $5 \cdot 2 \cdot 2 =$

20 blokker, og vi trekker tilsammen 20 par slik at vi får 40 forsøkspersoner som ved den andre utvalgsmetoden.

En blokk vil f.eks. være kvinner i den yngste aldersgruppa som ikke har EDB-erfaring, og det trekkes da 2 tilfeldige blant de ansatte som tilhører denne blokka. Deretter trekkes det tilfeldig hvilken av disse 2 som skal prøve system 1, mens den andre prøver system 2. Selve observasjonen blir da *differansen* mellom de 2 poengsummene, og vi får 20 slike differanser. Bemerk at det her er grunn til å tro at de to variablene i hvert par er avhengige og positivt korrelerte, mens det derimot er rimelig å anta at differansene fra hver blokk er uavhengige av hverandre.

En stor *fordel* ved parvis sammenligning er at vi får homogene (ensartede) eksperimenter innenfor hver blokk (hvert par). Forskjellen på poengsummen til 2 forsøkspersoner som prøver 2 forskjellige systemer, kan derved mer entydig knyttes til forskjell i brukervennlighet til de 2 systemene. På denne måten kan vi få ned variansen til vår estimator for forventet differanse i brukervennlighet, og derved få smalere konfidensintervall og bedre tester.

En *ulempe* ved parvis sammenligning er at vi ikke lenger har tilfeldige utvalg fra populasjonen(e) som helhet. Dessuten er antall differanser bare halvparten av antall forsøkspersoner. Dette bidrar til å øke variansen. Det kan også tenkes at det finnes faktorer vi ikke har oversikt over som kan påvirke responsen på EDB-prøvene, og at vår blokkinndeling fører til at disse faktorene påvirker resultatene på en systematisk (og uønsket) måte.

## 8.2 *Tilfeldiggjøring (randomisering)*

Vi har ofte snakket om å foreta et *tilfeldig utvalg* fra en populasjon. Begrepet ble definert allerede innledningsvis i kap. 4, men det skader ikke å gjenta definisjonen:

### Tilfeldig utvalg fra en populasjon (definisjon)

La en populasjon bestå av  $N$  enheter. Et utvalg på  $n$  forskjellige enheter fra denne populasjonen utgjør da et tilfeldig utvalg, dersom utvalget er foretatt på en slik måte at alle

$$\binom{N}{n} = \frac{N!}{n!(N-n)!}$$

mulige måter (kombinasjoner) å sette sammen utvalget på er like sannsynlige.

I tråd med definisjonen ovenfor, kan vi ved sammenligning av to behandlinger angi følgende prosedyrer for å foreta tilfeldige utvalg ved våre 2 metoder: 1) to uavhengige og tilfeldige utvalg, og 2) parvis sammenligning:

### Tilfeldiggjøring av to uavhengige utvalg ved sammenligning av to behandlinger.

Si at du har  $n$  tilfeldig valgte forsøksenheter der  $n_1$  skal få behandling 1 og  $n_2 = n - n_1$  skal få behandling 2 (dvs. ett utvalg på størrelse  $n_1$  fra populasjon 1, og ett utvalg på størrelse  $n_2$  fra populasjon 2). De to utvalgene blir da tilfeldige dersom du trekker de  $n_1$  enhetene som skal få behandling 1 fra de  $n$  mulige slik at alle  $\binom{n}{n_1}$  mulige kombinasjoner er like sannsynlige, og lar de resterende  $n_2 = n - n_1$  forsøksenheterne få behandling 2.

**Eks. 8.2** **Markedsføring.** Et firma driver markedsføring av et nytt produkt. De ønsker ved telefonoppringning å undersøke effekten, ved å spørre tilfeldige personer om de kjenner produktet *like før* og *en måned etter* en markedsføringskampanje. Anta at 1000 personer spørres før og 1000 personer spørres etter kampanjen. Si at firmaet først plukker ut 2000 personer tilfeldig fra telefonkatalogen. Deretter trekkes det 1000 tilfeldige blant de uttrukne 2000, som blir oppringt *like før* kampanjen. De resterende 1000 blir oppringt *en måned etter* kampanjen. Firmaet har da fått 2 uavhengige utvalg slik beskrevet i ramma ovenfor. ☺

### Tilfeldiggjøring av utvalg ved *parvis sammenligning* av to behandlinger.

Si at du har  $n$  forsøksenheter som er delt inn i  $n/2$  «homogene» (ensartede) blokker, 2 forsøksenheter pr. blokk. De 2 forsøksenhetene i hver blokk antas å være tilfeldig valgt fra alle mulige forsøksenheter som tilhører hver blokk. Tilfeldiggjøringen består nå (om mulig) i å slå mynt og kron om hvilken av de 2 forsøksenhetene som skal få behandling 1, og så la den andre få behandling 2.

**Eks. 8.3** **Fysikk- og matematikkarakterer.** Anta at vi skal sammenligne individuell fysikk-(3FY) og matematikk-(3MN) karakterer i videregående skole i 1992. Si at vi her plukker ut 100 tilfeldige kandidater blant alle som har avlagt eksamen i både 3FY og 3MN. Hver kandidat er nå en «blokk». Her har det imidlertid ingen mening i å snakke om å slå «mynt og kron» om rekkefølgen mellom 3FY- og 3MN-resultatet (i så fall måtte det bli trukket tilfeldig i hvilken rekkefølge kandidatene skulle avlegge de to eksamenene, hvilket er praktisk vanskelig). ☺

## 8.3 Uavhengige tilfeldige utvalg

Vi skal sammenligne 2 behandlinger på basis av 2 uavhengige utvalg, der utvalg 1 får behandling nr. 1, og utvalg 2 får behandling nr. 2. Vi lar  $X_i$  betegne responsen av behandling 1 til enhet nr.  $i$  i utvalg nr. 1. Tilsvarende lar vi  $Y_j$  betegne responsen av behandling 2 på enhet nr.  $j$  i utvalg nr. 2. Vi har da følgende eksperimentelle oppsett:

### Uavhengige tilfeldige utvalg

Utvalg:

$X_1, \dots, X_{n_1}$   
fra populasjon 1

Observatorer:

$$\bar{X} = \frac{1}{n_1} \sum_{i=1}^{n_1} X_i, \quad S_1^2 = \frac{1}{n_1 - 1} \sum_{i=1}^{n_1} (X_i - \bar{X})^2$$

$Y_1, \dots, Y_{n_2}$   
fra populasjon 2

$$\bar{Y} = \frac{1}{n_2} \sum_{i=1}^{n_2} Y_i, \quad S_2^2 = \frac{1}{n_2 - 1} \sum_{i=1}^{n_2} (Y_i - \bar{Y})^2$$

Når vi skal konstruere konfidensintervall for  $\mu_1 - \mu_2$ , eller foreta hypotesetester for  $\mu_1 - \mu_2$ , så må vi gjøre noen forutsetninger om populasjonsfordelingen(e) til hver av  $X$ -ene (populasjon 1) og til hver av  $Y$ -ene (populasjon 2). Vi skal anta følgende:

### Forutsetninger og symbol-definisjoner

- a)  $X_1, \dots, X_{n_1}$  utgjør et tilfeldig utvalg på  $n_1$  uavhengige variabler fra populasjon 1 med populasjonsforventning  $\mu_1$  og populasjons-standardavvik  $\sigma_1$ .
- b)  $Y_1, \dots, Y_{n_2}$  utgjør et tilfeldig utvalg på  $n_2$  uavhengige variabler fra populasjon 2 med populasjonsforventning  $\mu_2$  og populasjons-standardavvik  $\sigma_2$ .
- c)  $X_1, \dots, X_{n_1}$  er uavhengige av  $Y_1, \dots, Y_{n_2}$ . Med andre ord, responsmålingene fra den ene behandlingen er uavhengige av responsmålingene fra den andre behandlingen.
- d) Vi definerer  $\Delta\mu = \mu_1 - \mu_2$

Vi må som i kap. 6 og 7 også her skille mellom store og små utvalg, og vi starter med små utvalg. Som tidligere får vi her med  $t$ -fordelingen å gjøre.

## Små utvalg

### Tilleggsbetingelser for små utvalg

- a) Populasjonsfordelingen til  $X$ -ene antas å være  $N(\mu_1, \sigma)$
- b) Populasjonsfordelingen til  $Y$ -ene antas å være  $N(\mu_2, \sigma)$

(NB! Det er antatt samme  $\sigma$  i begge populasjonsfordelinger)

Hvordan skal vi så konstruere konfidensintervall og teste hypoteser for  $\Delta\mu = \mu_1 - \mu_2$ ? Analogt med at vi tidligere så på forventning og standardavvik til middelverdien fra én populasjon, ser vi nå på forventning og standardavvik til differansen mellom middelverdien til  $X$ -ene fra den ene populasjonen og middelverdien til  $Y$ -ene fra den andre:

$$E(\bar{X} - \bar{Y}) = E(\bar{X}) - E(\bar{Y}) = \mu_1 - \mu_2 = \Delta\mu$$

$\bar{X} - \bar{Y}$  er følgelig en forventningsrett estimator for  $\mu_1 - \mu_2$ . Når vi skal beregne variansen til  $\bar{X} - \bar{Y}$  trenger vi ikke å tenke på kovariansen, fordi  $\bar{X} - \bar{Y}$  er uavhengige. Fordi  $\sigma_1 = \sigma_2 = \sigma$  får vi derfor:

$$\text{Var}(\bar{X} - \bar{Y}) = \sigma^2 \left( \frac{1}{n_1} + \frac{1}{n_2} \right)$$

Videre er  $\bar{X} - \bar{Y}$  normalfordelt  $N(\mu_1 - \mu_2, \sigma (1/n_1 + 1/n_2)^{1/2})$ , og vi har derfor at

$$Z = \frac{(\bar{X} - \bar{Y}) - \Delta\mu}{\sigma \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \text{ er } N(0,1)$$

Vi kan nå benytte både  $X$ - og  $Y$ -verdiene til å estimere den felles variansen  $\sigma^2$  (og dermed det felles standardavviket  $\sigma$ ), og vi skal benytte følgende **interpolerte varians-estimator**,  $S_p^2$ , til dette formålet:

### Interpolert estimator, $S_p^2$ , for felles $\sigma^2$

små utvalg:  $n_1 < 30$  og  $n_2 < 30$

$$S_p^2 = \frac{1}{n_1 + n_2 - 2} \left( \sum_{i=1}^{n_1} (X_i - \bar{X})^2 + \sum_{j=1}^{n_2} (Y_j - \bar{Y})^2 \right) = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

La oss illustrere hvordan vi beregner interpolert standardavvik,  $s_p$ , ved et eksempel:

#### Eks. 8.4 Beregning av interpolert standardavvik, $s_p$

Utvalg fra populasjon 1:	8	5	7	6	9	7
Utvalg fra populasjon 2:	2	6	4	7	6	

Vi beregner først følgende summer:

$$\sum_{i=1}^6 x_i = 8 + \dots + 7 = 42, \quad \sum_{i=1}^6 x_i^2 = 8^2 + \dots + 7^2 = 304$$

$$\sum_{j=1}^5 y_j = 2 + \dots + 6 = 25, \quad \sum_{j=1}^5 y_j^2 = 2^2 + \dots + 6^2 = 141$$

Fra summene over finner vi at  $\bar{x} = 42/6 = 7$  og  $\bar{y} = 25/5 = 5$ . (Disse trenger vi strengt tatt ikke for å beregne  $s_p$ ). Videre får vi:

$$\sum_{i=1}^6 (x_i - \bar{x})^2 = \sum_{i=1}^6 x_i^2 - \frac{1}{6} \left( \sum_{i=1}^6 x_i \right)^2 = 304 - \frac{1}{6} \cdot 42^2 = 10$$

$$\sum_{j=1}^5 (y_j - \bar{y})^2 = \sum_{j=1}^5 y_j^2 - \frac{1}{5} \left( \sum_{j=1}^5 y_j \right)^2 = 141 - \frac{1}{5} \cdot 25^2 = 16$$

Setter inn i formelen for  $s_p^2$  (se ramme) og får:

$$s_p^2 = \frac{\sum (x - \bar{x})^2 + \sum (y - \bar{y})^2}{n_1 + n_2 - 2} = \frac{10 + 16}{6 + 5 - 2} = 2.89$$

$$\Rightarrow s_p = \sqrt{2.89} = 1.70$$

De individuelle utvalgsvariansene blir  $s_1^2 = 10/(6-1) = 2$  og  $s_2^2 = 16/(5-1) = 4$ . Vi ser at verdien for  $s_p^2$  (2.89) er nærmere  $s_1^2 = 2$  enn  $s_2^2 = 4$ . Dette skyldes at utvalgstørrelsen  $n_1 = 6$  er større enn utvalgstørrelsen  $n_2 = 5$ . ☺

Det kan vises (prøv selv!) at  $S_p^2$  er en forventningsrett estimator for  $\sigma^2$ . Analogt med tidligere kapitler kan vi nå sette opp følgende for beregning av konfidensintervall og hypotesetesting:

### Konfidensintervall for $\Delta\mu = \mu_1 - \mu_2$

(små utvalg:  $n_1 < 30$  og  $n_2 < 30$ )

Et  $100(1-\alpha)$  % konfidensintervall for  $\mu_1 - \mu_2$  er gitt ved:

$$\left( (\bar{X} - \bar{Y}) - t_{\alpha/2} \cdot S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}, (\bar{X} - \bar{Y}) + t_{\alpha/2} \cdot S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \right)$$

$$\text{der } S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

og  $t_{\alpha/2}$  er øvre  $\alpha/2$ -fraktil i  $t$ -fordelingen med  $n_1+n_2-2$  frihetsgrader (d.f.).

### Test for $\Delta\mu = \mu_1 - \mu_2$ ved små utvalg

Testobservator:  $T = \frac{\bar{X} - \bar{Y} - \delta}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$  for testene:

$$\begin{array}{lll} H_0: \Delta\mu \leq \delta & H_0: \Delta\mu \geq \delta & H_0: \Delta\mu = \delta \\ H_1: \Delta\mu > \delta & H_1: \Delta\mu < \delta & H_1: \Delta\mu \neq \delta \\ \Rightarrow R: T > t_\alpha & \Rightarrow R: T < -t_\alpha & \Rightarrow R: |T| > t_{\alpha/2} \end{array}$$

der  $R$  angir forkastingsområdet til testen med signifikansnivå  $\alpha$ . Videre er  $t_\alpha$  og  $t_{\alpha/2}$  henholdsvis øvre  $\alpha$ -fraktil og øvre  $\alpha/2$ -fraktil i  $t$ -fordelingen med  $n_1+n_2-2$  frihetsgrader.

**Eks. 8.5****Sammenligning av brukervennlighet til 2 EDB-systemer.**

Vi skal sammenligne brukervennligheten til 2 forskjellige EDB-systemer (se beskrivelse i innledningen) tenkt brukt i skatteetaten. Brukervennligheten måles ved oppnådd poengsum til individuell skriftlig prøve etter et innføringskurs. Maks. oppnåelig poengsum er 60.  $n = 25$  forsøkspersoner plukkes ut tilfeldig. Av disse trekkes et tilfeldig utvalg på  $n_1 = 13$  personer som får innføringskurs i system 1, mens de resterende  $n_2 = 12$  personer får innføringskurs i system 2. Vi lar  $X$  betegne poengsummen til en besvarelse fra utvalg 1, og lar  $Y$  betegne poengsummen til en besvarelse fra utvalg 2. Vi lar videre  $\mu_1$  og  $\sigma_1$  betegne forventning og standardavvik til poengsummen av en tilfeldig prøve for system 1. (Forventningen  $\mu_1$  vil her si den middelverdi vi ville funnet dersom alle potensielle EDB-brukere i skatteetaten hadde gjennomgått innføringskurset og avlagt prøve). Tilsvarende lar vi  $\mu_2$  og  $\sigma_2$  betegne forventning og standardavvik tilknyttet system 2. Vi får følgende resultat av prøvene:

Poengsum ( $x$ ), utvalg 1:	44, 44, 56, 46, 47, 38, 58, 53, 49, 35, 46, 30, 41
Poengsum ( $y$ ), utvalg 2:	35, 47, 55, 29, 40, 39, 32, 41, 42, 57, 51, 39

*Oppgave*

- Beregn et 95 % konfidensintervall for forskjellen i brukervennlighet for de 2 EDB-systemene, definert ved forventet forskjell på prøveresultatene.
- Gir prøveresultatene en sterk indikasjon på at det ene systemet er mer brukervennlig enn det andre?

*Løsningsforslag*

Vi utfører de nødvendige beregninger og får:

$$\text{System 1: } \bar{x} = 45.15, \quad \Sigma(x_i - \bar{x})^2 = 769.69, \quad s_1^2 = 64.0$$

$$\text{System 2: } \bar{y} = 42.25, \quad \Sigma(y_i - \bar{y})^2 = 840.25, \quad s_2^2 = 76.4$$

Fra tabellen ser vi at utvalgsstørrelsene er  $n_1 = 13$  og  $n_2 = 12$ . Interpolert varians,  $s_p^2$ , blir:

$$s_p^2 = \frac{\Sigma(x_i - \bar{x})^2 + \Sigma(y_i - \bar{y})^2}{n_1 + n_2 - 2} = \frac{767.69 + 840.25}{23} = 69.9 = 8.36^2$$

- a) Et 95 % konfidensintervall tilsvarer at  $\alpha/2 = 0.025$ . Vi slår opp i  $t$ -fordelingstabellen med 23 frihetsgrader, og finner at  $t_{0.025} = 2.069$ . Med hjelp av formelen i «KI -ramma» finner vi nå følgende:

$$\bar{x} - \bar{y} \mp t_{0.025} \cdot s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} = \\ 45.15 - 42.25 \mp 2.069 \cdot 8.35 \sqrt{\frac{1}{13} + \frac{1}{12}} = 2.90 \mp 6.92$$

dvs. (-4.02, 9.82) er et 95 % konfidensintervall for  $\mu_1 - \mu_2$ .

- b) Vi tester om dataene gir sterk grunn til å hevde at det ene systemet er mer brukervennlig enn det andre. Vi følger vårt tidligere 5-punkts skjema og får:

- 1) Sannsynlighetsmodell:  $X_1, \dots, X_{13}$  er uavhengige  $N(\mu_1, \sigma)$ -variable, og  $Y_1, \dots, Y_{12}$  er uavhengige  $N(\mu_2, \sigma)$ -variable. Dessuten er  $X$ -ene antatt å være uavhengig av  $Y$ -ene. Da kan vi bruke oppsettet i «hypoteseramma» for små utvalg.
- 2)  $H_0$ : Det er ingen forskjell i brukervennlighet mellom de 2 EDB-systemene ( $\mu_1 - \mu_2 = 0$ ).

$H_1$ : Det ene systemet er mer brukervennlig enn det andre  
( $\mu_1 - \mu_2 \neq 0$ ).

- 3) Fra «hypoteseramma» finner vi testobservator,  $T$  (med  $\delta = 0$ ), og forkastingsområde,  $R$ :

$$T = \frac{\bar{X} - \bar{Y}}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}, \quad R : |T| > t_{0.025} = 2.069$$

- 4) Vi setter inn resultatene fra forsøkene og får:

$$T = \frac{45.15 - 42.25}{8.36 \sqrt{\frac{1}{12} + \frac{1}{12}}} = \frac{2.90}{3.35} = 0.87$$

Siden  $|T| < 2.069$  kan vi formulere følgende konklusjon: «Resultatene av prøvene gir på signifikansnivå 5.0 % ikke noen sterk indikasjon på at det ene systemet er mer brukervennlig enn det andre».

Konklusjonen ovenfor kunne vi også trukket direkte fra resultatet i a), siden vårt 95 % konfidensintervall for  $\mu_1 - \mu_2$  dekket 0.

La oss se hvordan vi ville benyttet Minitab i dette tilfellet. Kode og utskrift er gjengitt nedenfor. Sammenlign selv med analysen ovenfor.

**MINITAB**

```
MTB > TwoSample 95.0 c1 c2; # Angir 95 % KI basert på data i c1 og c2  
SUBC> Alternative 0;           # Angir at vi skal teste likhet mot ulikhet  
SUBC> Pooled.                 # Angir at vi beregner pooled standardavvik
```

**Two Sample T-Test and Confidence Interval**

Two sample T for C1 vs C2

	N	Mean	StDev	SE Mean
C1	13	45.15	8.00	2.2
C2	12	42.25	8.74	2.5

95% CI for mu C1 - mu C2: (-4.0, 9.8)

T-Test mu C1 = mu C2 (vs not =): T= 0.87 P=0.39 DF= 23

Both use Pooled StDev = 8.36 ☺

**Store utvalg**

Vi skal sette som grenser  $n_1 \geq 30$  og  $n_2 \geq 30$  for at vi skal kunne bruke formlene for store utvalg. I dette tilfellet behøver vi ikke å anta at populasjonsfordelingene er normale, og vi behøver heller ikke å anta at populasjonsvariansene er like. Vi får følgende oppsett:

**Store og uavhengige utvalg** ( $n_1 \geq 30, n_2 \geq 30$ )

Et tilnærmet  $100(1-\alpha)\%$  KI for  $\mu_1 - \mu_2$  er i dette tilfellet gitt ved

$$\left( \bar{X} - \bar{Y} - z_{\alpha/2} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}, \bar{X} - \bar{Y} + z_{\alpha/2} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \right)$$

Som *testobsevator*,  $T$ , bruker vi  $T = \frac{\bar{X} - \bar{Y} - \delta}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$  for testene

$$H_0: \Delta\mu \leq \delta$$

$$H_0: \Delta\mu \geq \delta$$

$$H_0: \Delta\mu = \delta$$

$$H_1: \Delta\mu > \delta$$

$$H_1: \Delta\mu < \delta$$

$$H_1: \Delta\mu \neq \delta$$

$$\Rightarrow R: T > z_\alpha$$

$$\Rightarrow R: T < -z_\alpha$$

$$\Rightarrow R: |T| > z_{\alpha/2}$$

der  $R$  angir forkastingsområdet til testen på signifikansnivå  $\alpha$ , og  $z_\alpha$  og  $z_{\alpha/2}$  er henholdsvis øvre  $\alpha$ - og øvre  $\alpha/2$ -fraktil i  $N(0,1)$ -fordelingen.

**Eks. 8.6**

**Sammenligning av voksne menns vekt i bygd og by.** Vi tenker oss at ernæringsmyndighetene ønsker å undersøke om det er forskjell i gjennomsnittsvekta til voksne menn på bygda og i byen. 2 tilfeldige utvalg plukkes ut: Ett utvalg på størrelse  $n_1=90$  fra bygda, og ett utvalg på størrelse  $n_2 = 100$  fra byen. Anta at målingene ville gitt følgende resultat:

	Byen:	Bygda:
Utvalgsstørrelse:	90	100
middelverdi [kg]	76,4	81,2
standardavvik [kg]:	8,2	7,6

*Oppgave*

- Bestem et 98 % KI for forskjellen i forventet gjennomsnittsvekt i bygda og i byen.
- Undersøk om det er sterkt hold i data til å påstå at voksne menn jevnt over er tyngre på bygda enn i byen. Velg signifikansnivå på 2 %.

*Løsningsforslag*

La  $X$  betegne vekta til en tilfeldig voksen mann på bygda, og la  $Y$  betegne vekta til en tilfeldig voksen mann i byen. Vi lar  $\mu_1 = E(X)$  og  $\mu_2 = E(Y)$ . Fra våre data har vi at  $\bar{x} = 76,4$ ,  $\bar{y} = 81,2$ ,  $s_1 = 8,2$  og  $s_2 = 7,6$ .

- a) Et 98 % KI betyr at  $\alpha/2 = 0.01$ . Fra  $t$ -fordelingstabellen med uendelig mange frihetsgrader (nederste linje), finner vi følgende øvre 1 %-fraktil i  $N(0,1)$ -fordelingen:  $z_{0.01} = 2.33$ . Vi får derfor følgende:

$$\bar{x} - \bar{y} \mp z_{0.01} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = 76.4 - 81.2 \mp 2.33 \sqrt{\frac{8.2^2}{90} + \frac{7.6^2}{100}} = -4.8 \mp 2.7$$

dvs. vi får at  $(-7.5, -2.2)$  kg er et 98 % KI for  $\Delta\mu = \mu_1 - \mu_2$ .

- b) Siden 98 % konfidensintervallet for  $\Delta\mu$  ikke dekker 0 og bare inneholder negative verdier («tyngre menn på bygda»), vil vi få forkastet en tosidig hypotese  $H_0: \Delta\mu = 0$  mot  $H_1: \Delta\mu \neq 0$  på nivå 2 %. Siden hypotesen vi skulle teste i utgangspunktet var ensidig (vi antok at dersom det var noen forskjell så var det i retning tyngre menn på bygda), vil data i enda større grad understøtte påstanden i b), dvs. vi forventer en signifikanssannsynlighet ( $p$ -verdi) en del mindre enn 0.02. Prøv selv å gjennomføre testen. ☺

## 8.4 Parvis sammenligning

Strukturen til observasjonene ved parvis sammenligning er som følger:

### Datastruktur ved parvis sammenligning

Par:	Behandling 1:	Behandling 2:	Differanse:
1	$X_1$	$Y_1$	$D_1 = X_1 - Y_1$
2	$X_2$	$Y_2$	$D_2 = X_2 - Y_2$
:	:	:	:
$n$	$X_n$	$Y_n$	$D_n = X_n - Y_n$

Parene  $(X_1, Y_1), \dots, (X_n, Y_n)$  antas uavhengige. Vi skal benytte følgende sumobservatorer:

$$\bar{D} = \frac{1}{n} \sum_{i=1}^n D_i, \quad S_D^2 = \frac{1}{n-1} \sum_{i=1}^n (D_i - \bar{D})^2$$

Selv om parene er uavhengige, så er normalt  $X$ -en og  $Y$ -en innenfor ett og samme par avhengige av hverandre. Fordelen med parvis sammenligning er å fjerne kilder til variasjonen til  $\bar{X} - \bar{Y}$  som ikke skyldes forskjell i behandlingene, for å få en eventuell forskjell i behandlingene klarere fram.

Siden  $D$ -ene er fri for blokkeffekter (dersom hvert par antas å inneholde de samme blokkeffekter, forsvinner jo blokkeffektene når vi tar differansen), så er det rimelig å anta at  $D$ -ene utgjør et tilfeldig utvalg fra en populasjon med forventning  $\mu_D$  og varians  $\sigma_D^2$ , der:

$$\begin{aligned} E(D_i) &= E(X_i - Y_i) = \mu_D \\ \text{Var}(D_i) &= \text{Var}(X_i - Y_i) = \sigma_D^2 \end{aligned}$$

Vi får nå følgende oppsett for konfidensintervall og hypotesetesting av  $\mu_1 - \mu_2$  på basis av parvis sammenligning når vi har små utvalg ( $n < 30$ ):

### Parvis sammenligning. Små utvalg ( $n < 30$ )

Antar at differansene  $D_i = X_i - Y_i$ ,  $i = 1, 2, \dots, n$ , er uavhengige med populasjonsfordeling  $N(\mu_D, \sigma_D^2)$ . Vi har da:

Et  $100(1-\alpha)\%$  KI for  $\mu_D$  er gitt ved

$$\left( \bar{D} - t_{\alpha/2} \frac{S_D}{\sqrt{n}}, \bar{D} + t_{\alpha/2} \frac{S_D}{\sqrt{n}} \right)$$

Som testobservator,  $T$ , skal vi benytte  $T = \frac{\bar{D} - \delta}{S_D / \sqrt{n}}$  for testene

$$\begin{array}{lll} H_0: \mu_D \leq \delta & H_0: \mu_D \geq \delta & H_0: \mu_D = \delta \\ H_1: \mu_D > \delta & H_1: \mu_D < \delta & H_1: \mu_D \neq \delta \\ \Rightarrow R: T > t_\alpha & \Rightarrow R: T < -t_\alpha & \Rightarrow R: |T| > t_{\alpha/2} \end{array}$$

der  $R$  angir forkastingsområdet for testen på signifikansnivå  $\alpha$ , og  $t_\alpha$  og  $t_{\alpha/2}$  er henholdsvis øvre  $\alpha$ - og øvre  $\alpha/2$ -fraktil i  $t$ -fordelingen med  $n-1$  frihetsgrader.

Når vi har store utvalg får vi samme oppsett som ovenfor, med eneste forskjell at vi erstatter  $t_\alpha$  og  $t_{\alpha/2}$  med henholdsvis  $z_\alpha$  og  $z_{\alpha/2}$ , dvs. vi erstatter  $t$ -fordelingsfraktilene med de tilsvarende  $N(0,1)$ -fraktilene.

#### Eks. 8.7 Bivirkning av medisinsk pille

En medisinsk forsker ønsker å undersøke om en pille har den ønskede bivirkning at den reduserer blodtrykket til brukeren. Undersøkelsen

består i å måle det opprinnelige blodtrykket til 15 kvinnelige studenter. Etter regulær bruk av pillen i 6 måneder, blir blodtrykket målt igjen. Resultatet av målingene er gjengitt i neste tabell.

Her representerer hvert individ en blokk (et par) med målinger: en før bruk av pillen ( $x$ ), og en etter bruk av pillen ( $y$ ). Differansene ( $d = x - y$ ) er også gjengitt i tabellen.

*Tabell. Blodtrykksmålinger før og etter bruk av pille.*

før ( $x$ )	70	80	72	76	76	76	72	78
etter ( $y$ )	68	72	62	70	58	66	68	52
$d = x - y$	2	8	10	6	18	10	4	26
før ( $x$ )	82	64	74	92	74	68	84	
etter ( $y$ )	64	72	74	60	74	72	74	
$d = x - y$	18	-8	0	32	0	-4	10	

### Oppgave

- Bestem et 95 % konfidensintervall for  $E(D) = E(X - Y) = \mu_D$ .
- Test på nivå 1.0 % om det er grunnlag for å hevde at pillen forventes å senke blodtrykket.

### Løsningsforslag

- Vi utfører de nødvendige beregninger og får:

$$\bar{d} = \frac{1}{15} \sum_{i=1}^{15} d_i = 8.80, \quad s_D = \sqrt{\frac{1}{14} \sum_{i=1}^{15} (d_i - \bar{d})^2} = 10.98$$

Et tilnærmet 95 % KI for  $\mu_D$  er nå gitt ved

$$\left( \bar{d} - t_{0.025} \cdot \frac{s_D}{\sqrt{15}}, \quad \bar{d} + t_{0.025} \cdot \frac{s_D}{\sqrt{15}} \right) = (2.72, 14.88)$$

der vi har satt inn verdien  $t_{0.025} = 2.145$  fra  $t$ -fordelingstabellen med 14 frihetsgrader.

- For å teste om forventet blodtrykk synker på grunn av pillen, tester vi  $H_0: \mu_D = 0$  mot  $H_1: \mu_D > 0$  ( $H_1$  er den hypotese vi «ønsker» å underbygge). Vi finner følgende verdi for testobservatoren:

$$T = \frac{\bar{d}}{s_D / \sqrt{n}} = \frac{8.80}{2.84} = 3.10 > t_{0.005} = 2.977$$

Siden  $T > t_{0,005} = 2.977$ , som er øvre 0.5 %-fraktil i  $t$ -fordelingen med 14 frihetsgrader, forkaster vi  $H_0$ , dvs. vi konkluderer at det er grunnlag i data for å påstå at pillen reduserer forventet blodtrykk på signifikansnivå 5 %. Vær imidlertid oppmerksom på at jo lavere nivå som velges for testen, jo mer forsiktig må vi være med å godta forutsetningen om at  $\bar{D}$  er tilnærmet normalfordelt.

## 8.5 Oppgaver

**8.1** For å sammenligne to metoder for opplæring av industriarbeidere, blir 20 arbeidere valgt ut til et forsøk. Av disse blir 10 valgt tilfeldig for å prøve metode 1, og de 10 andre prøver metode 2. Etter opplæringen utfører arbeiderne en individuell test der forventet tid (i minutter) brukt på testen er målet på hvor god metoden er. Følgende data oppnås:

	Tid [minutter]				
Met. 1	15	20	11	23	16
Met. 2	23	31	13	19	23
Tid [minutter]					
Met. 1	21	18	16	27	24
Met. 2	17	28	26	25	28

- a) Kan du konkludere fra dataene at forventet tid er signifikant mindre etter trening med metode 1 enn med 2 ?  
(test med  $\alpha = 0.05$ )
- b) Formuler de forutsetningene du gjør om populasjonsfordelingene.
- c) Konstruer et 95 % konfidensintervall for forskjellen i forventet tid for de to metodene.

**8.2** En sosiolog ønsker å sammenligne fødselsraten til kvinner i to stammer A og B i Øst-Afrika. Fra hver stamme blir det plukket ut et tilfeldig utvalg på 100 kvinner i aldersgruppen 50-60 år, og antall barn hver kvinne har født blir registrert. Følgende frekvensfordeling oppnås:

	Antall barn				
	0	1	2	3	4
frekv. A	6	14	18	25	19
frekv. B	0	3	8	18	30
Antall barn					
	5	6	7	8	Tot
frekv. A	11	5	2	0	100
frekv. B	19	15	5	2	100

- a) Beregn middelverdi og standardavvik til hver frekvensfordeling
- b) Indikerer dataene en signifikant forskjell i forventet antall barn født av kvinner i stamme B?
- c) Konstruer et 98 % konfidensintervall for differansen mellom de to populasjons-forventningene.

**8.3** Målinger av gripestyrken til venstre- og høyre-handa til 10 kevhendte skribenter registreres:

	Person				
	1	2	3	4	5
Venstre	140	90	125	130	95
Høyre	138	87	110	132	96
Person					
	6	7	8	9	10
Venstre	121	85	97	131	110
Høyre	120	86	90	129	100

- a) Underbygger dataene påstanden at kevhendte har en større gripestyrke i venstrehånda enn i høyrehånda ?
- b) Konstruer et 90 % konfidensintervall for forventet forskjell.

**8.4** En ønsker å sammenlikne det surstoffforbruk ( $\text{mm}^3$  pr. time) en øret

har når den svømmer i henholdsvis hurtig (A) og i langsomt (B) rennende vann, og gjør et forsøk med i alt 8 nummererte ørreter som svømmer en viss tid i A og samme tid i B. For hver ørret avgjør en ved loddrekning om den først skal svømme i A eller B. I hvert enkelt forsøk blir surstoffforbruket bestemt, og resultatet blir:

		Ørret nr:			
		1	2	3	4
A	107	98	87	118	
	B	94	69	91	87
		Ørret nr:			
		5	6	7	8
A	96	125	131	106	
B	97	112	107	80	

Se på differansene  $X = A - B$ :

$$X: \quad 13 \quad 29 \quad -4 \quad 31 \quad -1 \quad 13 \quad 24 \quad 26$$

og anta at  $X$  er normalfordelt. Test om ørretens surstoffforbruk er større når den svømmer i hurtig rennende vann enn når den svømmer i langsomt rennende vann. Bruk 5 % signifikansnivå.

**8.5 (E)** To metoder A og B for måling av PH-verdier i en oppløsning skal sammenlignes. En gruppe studenter måler 10 ulike oppløsninger med begge målemetodene. En har mistanke om at de to metodene ikke gir samme gjennomsnittlige måleresultat.

Måleresultater:

Nr. (i)	1	2	3	4	5
A	8.14	7.19	6.75	7.00	7.64
B	7.91	7.48	7.31	7.36	7.62
Nr.(i)	6	7	8	9	10
A	7.01	6.98	6.30	7.52	7.40
B	7.38	7.40	6.60	7.49	7.35

Antagelser:

Metode A gir måleresultater  $X_i$  som er normalfordelt  $N(\mu_i, \sigma_1)$ .

Metode B gir måleresultater  $Y_i$  som er normalfordelt  $N(\mu_i + \Delta, \sigma_2)$ .

Differansene  $Y_i - X_i$  antas uavhengige og normalfordelte med standardavvik  $\sigma$ .

- Finn et punktestimat for  $\Delta$ .
- Bestem et 95 % konfidensintervall for  $\Delta$ .
- Finn et 99 % konfidensintervall på formen  $[0, b]$  for  $\sigma$ .

**8.6 (E)** To typer voltmetre A og B skal sammenlignes. En har mistanke om at voltmetre av typen B systematisk viser høyere verdi for spenningen enn voltmetre av type A. Seks studenter fikk hver utdelt et voltmeter av hver type og målte hver sin fritt valgte spenning med begge voltmetrerne.

La  $X_i$  være målt spenning med voltmeter av type A, og  $Y_i$  målt spenning med voltmeter av type B.

Måleresultatene ble:

Stud. nr.	1	2	3
$X_i$ [V]	5.0	2.1	6.9
$Y_i$ [V]	5.3	2.1	7.1
Stud. nr.	4	5	6
$X_i$ [V]	8.3	9.4	11.8
$Y_i$ [V]	8.2	9.7	12.0

Anta  $Z_i = Y_i - X_i$  er uif  $N(\Delta, \sigma)$ .

- a) Konstruer på grunnlag av observasjonene et 90 % konfidensintervall for  $\Delta$ . Vil du på grunnlag av observasjonene påstå at B gir høyere verdier enn A?
- b) Konstruer et 99 % konfidensintervall på formen  $[0, b]$  for  $\sigma$ .

**8.7 (E)** En PC-leverandør er interessert i å undersøke om bruk av fargeskjermer vil virke gunstig ved tekstbehandling. En gruppe på 10 personer ble bedt om å skrive inn en bestemt tekst på maskiner med svart-hvitt-skjermer. En annen gruppe på 10 personer ble bedt om å skrive inn den samme teksten mot fargeskjermer.

Tiden (i minutter) som hver av deltakerne brukte på å skrive inn teksten, ble målt.

Måleverdiene  $(x_1, \dots, x_{10})$  med svart-hvitt-skjerm, og  $(y_1, \dots, y_{10})$  med farge-skjerm er vist nedenfor:

$x:$  10.2 13.1 13.9 14.1 12.5 10.5  
13.4 14.2 14.4 13.5

$y:$  11.5 12.3 13.1 10.5 9.9 11.3  
13.4 11.0 13.2 12.8

Du kan i et av punktene nedenfor få bruk for følgende generelle konfidens-intervall:

$$\bar{x} - \bar{y} \mp t_{\alpha/2} \cdot s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

der  $t_{\alpha/2}$  er  $\alpha/2$ -fraktilen i  $t$ -fordelingen med  $n_1 + n_2 - 2$  frihetsgrader, og

$$s_p^2 = \frac{(n_1 - 1)s_x^2 + (n_2 - 1)s_y^2}{n_1 + n_2 - 2}$$

Regnehjelp:

$$\sum x_i = 129.8, \quad \sum x_i^2 = 1705, \\ \sum y_i = 119, \quad \sum y_i^2 = 1430$$

- a) Beregn  $\bar{x}$ ,  $\bar{y}$  og de observerte standardavvikene  $s_x$  og  $s_y$  for dataene over.
- b) Bestem et 95 % konfidensintervall for forskjellen mellom forventet tidsforbruk for de to gruppene. Vi antar normalfordeling i begge grupper, og uavhengighet både innen hver gruppe og gruppene imellom. Hvilken konklusjon kan du trekke på grunnlag av resultatet?

Gruppen som først skrev teksten inn mot fargeskjerm, ble dernest bedt om å skrive inn teksten på maskiner med svart-hvitt-skjerm. (Vi forutsetter at teksten er av en slik karakter at det ikke er noen fordel å ha skrevet den en gang tidligere). Måleverdiene  $(z_1, \dots, z_{10})$  er vist nedenfor i samme rekkefølge som observasjonene  $(y_1, \dots, y_{10})$ :

$z:$  12.0 12.4 13.5 11.4 12.0 11.4  
12.8 12.9 13.6 12.9

- c) Bestem et 95 % KI for forskjellen mellom forventet tidsforbruk mot svart-hvitt-skjerm ( $z_i$ ) og fargeskjerm

$(y_i)$ . Anta normalfordeling. Tolk resultat. Oppgitt:  $\bar{z}=12.44$ ,  $s_{z-y}=.84$ .

**8.8** Man ønsker å undersøke om vitamin B<sub>1</sub> er et effektivt middel til å stimulere veksten av sopp. Til dette formålet blir 10 sopper valgt tilfeldig og 5 av disse får behandling av vitamin B<sub>1</sub>. Vektene på soppene etter avsluttet forsøk var følgende:

Beh.	27	34	20	28	20
med B <sub>1</sub> :					
Ikke beh.:	18	14.5	13.5	12.5	23

- a) Beregn empirisk middelverdi, varians og standardavvik for de to gruppene.

Det antas heretter at vektfordelingen på behandlet sopp er  $N(\mu_1, \sigma)$ , og at vektfordelingen for ubehandlet sopp er  $N(\mu_2, \sigma)$ .

- b) Finn forventningsrette estimatorer for  $\mu_1 - \mu_2$  og  $\sigma^2$ .
- c) Man ønsker å finne ut om det er grunnlag for å påstå at vitamin B<sub>1</sub> har en positiv effekt. Formulér dette som et hypotesetestingsproblem og test med 1 % signifikansnivå. Hva er din konklusjon ?

**8.9** To galluper utført av Norges Markedsdata for september 1981 og januar 1982 viste blant annet følgende resultater for preferanse i %:

	Sep.81	Jan.82
Vil stemme AP	36,4	38,7
Vil stemme H	23,6	30,5
Totalt antall som oppga preferanse.	1145	1134

- a) Er endringen i andelen som vil stemme Høyre signifikant?
- b) Finn et 95 % konfidensintervall for økningen til AP. Angi signifikanssannsynligheten.

## 8.6 Formelsamling

### Uavhengige utvalg

#### Betegnelser

- $\bar{X}$  middelverdi av  $n_1 X$ -er (utvalg 1).  
 $\bar{Y}$  middelverdi av  $n_2 Y$ -er (utvalg 2).  
 $S_1$ : empirisk standardavvik, utvalg 1.  
 $S_2$ : empirisk standardavvik, utvalg 2.  
 $\mu_1$ : forventning, populasjon 1.  
 $\mu_2$ : forventning, populasjon 2.  
 $\sigma_1$ : standardavvik, populasjon 1.  
 $\sigma_2$ : standardavvik, populasjon 2.  
 $S_p^2$ : empirisk interpolert varians:  

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$
  
 $t_\alpha$ : øvre  $\alpha$ -fraktil i t-fordelingen med  $n_1 + n_2 - 2$  frihetsgrader.

#### KI og test for $\Delta\mu = \mu_1 - \mu_2$

(små utvalg,  $n_1 < 30$  og  $n_2 < 30$ ). Forutsetter  $X_1, \dots, X_{n_1}$  uif  $N(\mu_1, \sigma)$  og  $Y_1, \dots, Y_{n_2}$  uif  $N(\mu_2, \sigma)$ .

100(1- $\alpha$ ) % KI for  $\mu_1 - \mu_2$ :

$$\bar{X} - \bar{Y} \mp t_{\alpha/2} \cdot S_p \sqrt{1/n_1 + 1/n_2}$$

$$\text{Test: } T = \frac{\bar{X} - \bar{Y} - \delta}{S_p \sqrt{1/n_1 + 1/n_2}}$$

- $H_0: \Delta\mu \Leftrightarrow \delta$     $H_0: \Delta\mu \Downarrow \delta$     $H_0: \Delta\mu = \delta$   
 $H_1: \Delta\mu > \delta$     $H_1: \Delta\mu < \delta$     $H_1: \Delta\mu \neq \delta$   
 $R: T > t_\alpha$     $R: T < -t_\alpha$     $R: |T| > t_{\alpha/2}$

#### KI og test for $\Delta\mu = \mu_1 - \mu_2$ .

(store utvalg,  $n_1 \geq 30, n_2 \geq 30$ ).

Tilnærmet 100(1- $\alpha$ ) % KI for  $\mu_1 - \mu_2$ :

$$\bar{X} - \bar{Y} \mp z_{\alpha/2} \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$$

$$\text{Test: } T = \frac{\bar{X} - \bar{Y} - \delta}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

- $H_0: \Delta\mu \leq \delta$     $H_0: \Delta\mu \geq \delta$     $H_0: \Delta\mu = \delta$   
 $H_1: \Delta\mu > \delta$     $H_1: \Delta\mu < \delta$     $H_1: \Delta\mu \neq \delta$   
 $R: T > z_\alpha$     $R: T < -z_\alpha$     $R: |T| > z_{\alpha/2}$

### Parvis sammenligning

#### Betegnelser

$\bar{D}$  = middelverdi til differansene  $D_i = (X_i - Y_i), i = 1, \dots, n$  som er forutsatt uavhengige og normalfordelte.

$S_D$  = empirisk standardavvik til  $D$ -ene.

$\mu_D$  = E( $D_i$ ),  $i = 1, \dots, n$ .

$\sigma_D$  = std( $D_i$ ),  $i = 1, \dots, n$ .

$t_\alpha$  = øvre  $\alpha$ -fraktil i t-fordelingen med  $n-1$  frihetsgrader.

#### KI og test for $\mu_D$ , små utvalg.

100(1- $\alpha$ ) % KI for  $\mu_D$ :

$$\left( \bar{D} - t_{\alpha/2} \cdot \frac{S_D}{\sqrt{n}}, \bar{D} + t_{\alpha/2} \cdot \frac{S_D}{\sqrt{n}} \right)$$

$$\text{Test: } T = \frac{\bar{D} - \delta}{S_D / \sqrt{n}}$$

$H_0: \mu_D \leq \delta$     $H_0: \mu_D \geq \delta$     $H_0: \mu_D = \delta$

$H_1: \mu_D > \delta$     $H_1: \mu_D < \delta$     $H_1: \mu_D \neq \delta$

$R: T > t_\alpha$     $R: T < -t_\alpha$     $R: |T| > t_{\alpha/2}$

#### KI og test for $\mu_D$ , store utvalg

Bruker samme formler som for små utvalg, med eneste forskjell at t-fraktilene byttes ut med  $N(0,1)$ -fraktiler.

## Kapittel 9

# Lineær regresjon

## 9.1 Innledning

I kap. 1 om deskriptiv statistikk ble det beskrevet hvordan vi tilpasset en rett linje  $y = a + b \cdot x$  best mulig til samhørende  $(x, y)$  data, dvs. vi gav formler for hvilke verdier for  $a$  og  $b$  som gav best resultat basert på minste kvadraters metode. Vi skal her gå mer grundig til verks, og blant annet se på hvor sikker en slik tilpasset linje blir. Den lineære regresjonsmodellen vi skal anvende er gjengitt i ramma nedenfor. Den er den desidert mest utbredte modellen både i statistisk litteratur og i statistikk-programpakker.

### Lineær regresjonsmodell

Vi antar at vi har  $n$  samhørende  $(x, Y)$ -variabelpar, der  $x_1, \dots, x_n$  er konstanter med neglisjerbar usikkerhet mens  $Y_1, \dots, Y_n$  er stokastiske variabler. Vår lineære regresjonsmodell består i å anta en lineær sammenheng mellom  $x$  og  $E(Y|x)$ , og kan matematisk formuleres som følger:

$$Y_i = a + b \cdot x_i + E_i = \text{responsvariabel}, \quad i = 1, 2, \dots, n$$

$$y_r = a + b \cdot x = E(Y|x) = \text{regresjonslinje}$$

$a$  = **skjæringsparameter** ( $y$ -verdi der regresjonslinja skjærer  $y$ -aksen)

$b$  = **steilhetsparameter** (stigningstall) for regresjonslinja

$x_i$  = kjent  $x$ -verdi (**uavhengig variabel, prediktor**)

$E_i$  = «**Feil**»-variabel, angir  $y$ -avstand fra  $Y_i$  til regresjonslinja

Vi forutsetter at  $E_1, \dots, E_n$  er uavhengige  $N(0, \sigma)$ -variabler.  $Y_1, \dots, Y_n$  er da uavhengige og normalfordelte med forventning  $E(Y_i) = a + b \cdot x_i$  og varians  $\text{Var}(Y_i) = \sigma^2$ .

**Eks. 9.1** **Lineariteten til et elektronisk instrument** skal undersøkes ved å se om utgangsspenningen,  $V_{ut}$ , er proporsjonal med inngangsspenningen,  $V_{inn}$ . Dessuten ønsker man å finne forsterkningen,  $K = V_{ut} / V_{inn}$ . Anta «perfekt» spenningskilde som genererer  $V_{inn}$ , mens spenningsmåleren på utgangen har en målenøyaktighet på  $\pm 1\%$ .

### Oppgave

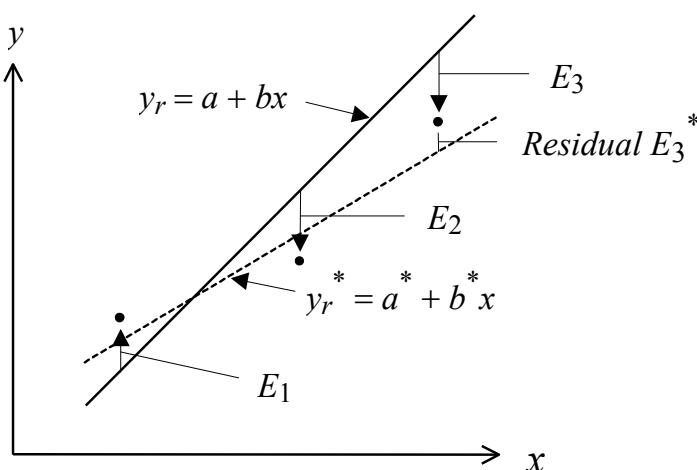
- Hva er uavhengig variabel og hva er responsvariabel?
- Hva er  $a$  i dette tilfellet?
- Hvilken instrumentegenskap tilsvarer  $b$  i den lineære regresjonsmodellen?
- Nevn eksempel på hva tilfeldige avvik fra linearitet kan skyldes.

### Løsningsforslag

- $V_{inn}$  er uavhengig variabel,  $V_{ut}$  er responsvariabel.
- $a = 0$ .
- $b = K$  i dette tilfellet.
- Måleusikkerhet til spenningsmåleren på utgangen. ☺

### Estimering

Modellen som er beskrevet inneholder generelt følgende tre ukjente parametere:  $a$ ,  $b$  og  $\sigma$ . I enkelte tilfeller vil en eller flere av disse være kjente. Hvis vi f. eks. vet at linja skal gå gjennom origo så er  $a = 0$ . De parametrerne som er ukjente må estimeres (anslås) med basis i et begrenset antall observasjoner. Vi skal bruke betegnelsene  $a^*$ ,  $b^*$  og  $S$  på estimatorene for henholdsvis  $a$ ,  $b$  og  $\sigma$ . Modell og estimatorene  $a^*$  og  $b^*$  er illustrert i Fig. 9.1 nedenfor.



Figur 9.1

Illustrasjon av «sann» regresjonslinje  $y_r = a + bx$  og estimert linje  $y_r^* = a^* + b^*x$  tilpasset et begrenset antall data (3 punkt i figuren).

### Kommentarer til figur 9.1:

- $y_r = a + b \cdot x$  er den «sanne» regresjonslinja vi ønsker å bestemme best mulig.
- De tre punktene (observasjonene) i figuren har  $y$ -avstand  $E_1, E_2$  og  $E_3$  fra den sanne regresjonslinja. Feilvariablene er uavhengige og  $N(0, \sigma)$ -fordelte.
- $y_r^* = a^* + b^* \cdot x$  er en best mulig tilpasset rett linje til de 3 observasjonene.  $y_r$  og  $y_r^*$  blir mer og mer sammenfallende ( $y_r^*$  blir bedre og bedre) med økende antall observasjoner og avtagende  $\sigma$ -verdi.
- $y$ -avstandene fra den *estimerte* regresjonslinja til observasjonene kalles **residualer**,  $E_3^*$  i figuren er et eksempel. Når antall observasjoner øker, vil jevnt over residualene nærme seg mer og mer feilvariablene, og er nyttige for å estimere  $\sigma$ .

I formlene som inngår i dette kapitlet vil følgende summasjonsformler opptrer så ofte at det er nyttig å ha spesielle betegnelser på dem:

### Summasjonsformler

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

$$\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$$

$$S_x^2 = \sum_{i=1}^n (x_i - \bar{x})^2 = \sum x_i^2 - \frac{1}{n} (\sum x_i)^2 = \sum x^2 - n(\bar{x})^2 = \sum (x_i - \bar{x}) \cdot x_i$$

$$S_y^2 = \sum_{i=1}^n (y_i - \bar{y})^2 = \sum y_i^2 - \frac{1}{n} (\sum y_i)^2 = \sum y^2 - n(\bar{y})^2 = \sum (y_i - \bar{y}) \cdot y_i$$

$$S_{xy} = \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = \sum x_i y_i - \frac{1}{n} (\sum x_i)(\sum y_i) = \sum xy - n \bar{x} \bar{y}$$

Kapitlet er disponert som følger: I kap. 9.2 skal vi angi formlene for minste kvadraters-estimatorene  $a^*$  og  $b^*$  for henholdsvis  $a$  og  $b$ . Kap. 9.3 og 9.4 behandler de statistiske egenskapene til henholdsvis  $b^*$  og  $a^*$ , som danner utgangspunkt for konfidensintervaller (KI) og tester for  $a$  og  $b$  beskrevet i kap. 9.7. I kap. 9.5 angis en estimator  $S$  for  $\sigma$  basert på residualene. Kap. 9.6 ser på de statistiske egenskapene til den estimerte regresjonslinja for en gitt  $x$ -verdi. Kap. 9.7 ser på KI og tester for  $a$ ,  $b$  og  $y_r = a + b \cdot x$ . I kap. 9.8 behandles noen vanlige ikke-lineære sammenhenger mellom  $x$  og  $y$ , og hvordan vi ved hjelp av lineariserende transformasjoner kan anvende lineær regresjonsanalyse i

slike tilfeller. Kap. 9.9 omhandler residualer og hvordan vi kan bruke disse til å undersøke om vår lineære regresjonsmodell er rimelig tilpasset våre data. Kap. 9.10 inneholder oppgaver og kap. 9.11 er en formelsamling.

## 9.2 Minste kvadraters estimatorer $a^*$ og $b^*$

Vi vil tilpasse de verdier for  $a$  og  $b$  som minimerer summen,  $Q$ , av de kvadratiske avvikene (i y-retning) mellom enkeltobservasjonene  $(x_i, y_i)$ ,  $i = 1, 2, \dots, n$ , og regresjonslinja  $y_r = a + b \cdot x$ :

$$Q = \sum_{i=1}^n (y_i - (a + b x_i))^2$$

I praksis gjøres dette ved å derivere  $Q$  med hensyn på  $a$  og  $b$ , sette de uttrykkene som fremkommer lik null, og løse de to ligningene vi får med hensyn på  $a$  og  $b$ :

$$\frac{\partial Q}{\partial a} = 0 \text{ og } \frac{\partial Q}{\partial b} = 0$$

Løsningene er det vi kaller minste kvadraters estimatorer, og vi betegner disse med  $^*$ :  $a^*$  er betegnelsen for  $a$ -estimatoren og  $b^*$  er betegnelse for  $b$ -estimatoren. Ligningene over gir følgende resultat:

### Minste kvadraters estimatorer $a^*$ og $b^*$

$$b^* = \frac{S_{xy}}{S_x^2}, \quad a^* = \bar{y} - b^* \cdot \bar{x}$$

Den tilsvarende regresjonslinja blir

$$y_r^* = a^* + b^* \cdot x = \bar{y} + b^* (x - \bar{x})$$

der definisjon av størrelsene som inngår er gitt i ramme med summasjonsformler (forrige ramme).

Legg merke til at  $y_r^*(\bar{x}) = \bar{y}$ , dvs. linja går gjennom punktet  $(\bar{x}, \bar{y})$ . Hvis man på øyemål tilpasser en rett linje gjennom  $(\bar{x}, \bar{y})$  som ser ut til å passe bra til dataene, vil man ofte komme svært nær den matematiske kurven uttrykt ved ligninga for  $y_r^*$  i forrige ramme.

### 9.3 Egenskaper til $b^*$

Vi skal nå se på de statistiske egenskapene til  $b^*$  med basis i modellen beskrevet i forrige avsnitt. Dette danner basis for intervallestimering og hypotesetesting for  $b$  beskrevet i kap. 9.5. Uttrykket for  $b^*$  kan omskrives som følger:

$$b^* = \frac{S_{xy}}{S_x^2} = \sum_{i=1}^n \left( \frac{x_i - \bar{x}}{S_x^2} \right) \cdot Y_i$$

Fra det siste uttrykket ser vi at  $b^*$  er en lineærkombinasjon av de uavhengige, normalfordelte variablene  $Y_1, \dots, Y_n$ . Følgelig er  $b^*$  en normalfordelt variabel. Ved litt regning finner vi at  $E(b^*) = b$ .  $b^*$  er følgelig en forventningsrett estimator for  $b$ . Videre finner vi følgende uttrykk for  $\text{Var}(b^*)$ :

$$\text{Var}(b^*) = \frac{\sigma^2}{S_x^2} = \frac{\sigma^2}{n} \left( \frac{1}{\frac{1}{n} S_x^2} \right)$$

der siste omskrivning er foretatt for å få fram at  $\text{Var}(b^*)$  varierer «omtrent» omvendt proporsjonalt med  $n$  fordi  $\frac{1}{n} S_x^2$  er en midlere kvadratsum som det er rimelig å anta ikke varierer særlig mye med  $n$ .

**Konklusjon:**

$b^*$  er  $N(b, \sigma/S_x)$ .

Merk at usikkerheten (standardavviket) til  $b^*$  avtar omvendt proporsjonalt med  $S_x$ , som er et mål på spredningen til våre  $x$ -variabler. Det kan med andre ord være fornuftig å spre  $x$ -verdiene godt for å få et sikkert anslag for regresjonslinjas steilhet. Vi bør imidlertid også sørge for å ha tilstrekkelig mange forskjellige  $x$ -verdier til å undersøke hvor god en lineær regresjonsmodell er på hele det aktuelle  $x$ -området.

### 9.4 Egenskaper til $a^*$

Vi ser nå på de statistiske egenskapene til estimatoren,  $a^*$ , for skjæringsparameteren  $a$ . Ved noe regning kan uttrykket for  $a^*$  omskrives som følger:

$$a^* = \bar{Y} - b^* \bar{x} = \sum_{i=1}^n \left( \frac{1}{n} - \frac{\bar{x}}{S_x^2} (x_i - \bar{x}) \right) Y_i$$

Merk at uttrykket foran  $Y_i$  i siste ligning er en ikke-stokastisk faktor som kan betraktes som en «konstant», og at hvert ledd i summen derfor blir normalfordelt, siden  $Y_i$  er forutsatt normalfordelt. Siden summen av uavhengige normalfordelte variabler er normalfordelt, blir derfor  $a^*$  normalfordelt.

Forventningen  $E(a^*)$  finnes enklest fra det første uttrykket for  $a^*$  i ligningen ovenfor:

$$E(a^*) = E(\bar{Y} - b^* \bar{x}) = E(\bar{Y}) - \bar{x} E(b^*) = a + b\bar{x} - \bar{x}b = a$$

$a^*$  er følgelig en forventningsrett estimator for  $a$ , siden  $E(a^*) = a$ . Ved noe regning finner vi følgende uttrykk for  $\text{Var}(a^*)$ :

$$\text{Var}(a^*) = \sigma^2 \left( \frac{1}{n} + \frac{(\bar{x})^2}{S_x^2} \right) = \frac{\sigma^2}{n} \left( 1 + \frac{(\bar{x})^2}{\frac{1}{n} S_x^2} \right)$$

der siste omskrivning har samme begrunnelse som siste omskrivning for  $\text{Var}(b^*)$ . Igjen ser vi at variansen er tilnærmet omvendt proporsjonal med antall observasjoner,  $n$ .

**Konklusjon:**

$$a^* \text{ er } N \left( a, \sigma \sqrt{\frac{1}{n} + \frac{(\bar{x})^2}{S_x^2}} \right).$$

Fra uttrykket for  $\text{Var}(a^*)$  ser vi at usikkerheten, i tillegg til å avta omvendt proporsjonalt med antall observasjoner,  $n$ , øker når  $|\bar{x}|$  øker. Det siste er intuitivt logisk, siden  $a$  angir  $y$ -verdien der linja skjærer  $y$ -aksen.

## 9.5 Estimering av $\sigma$

I vår modell forutsatte vi at feilvariablene  $E_1, \dots, E_n$ , som betegnet  $y$ -avstandene fra observasjonene  $Y_1, \dots, Y_n$  til den teoretiske og ukjente regresjonslinja  $y_r = a + b \cdot x$ , var uif  $N(0, \sigma)$  med ukjent standardavvik  $\sigma$ . I neste ramme er vist hvordan vi kan estimere  $\sigma$  med basis i de såkalte *residualene*. Residualene er definert som  $y$ -avstandene fra  $Y$ -observasjonene til den *estimerte* regresjonslinja  $y_r^*$  (se fig. 9.1).

Merk at selv om  $S^2$  er en forventningsrett estimator for  $\sigma^2$  så er ikke  $S$  en forventningsrett estimator for  $\sigma$ . For rimelig store  $n$  er imidlertid avviket (*bias* på engelsk) fra forventningsretthet, **forventningsskeivheten**, liten.

### Residualer, feilkvadratsum SSE og $\sigma^2$ -estimatoren $S^2$

I modellen  $Y_i = a + b \cdot x_i + E_i$ , der  $x_1, \dots, x_n$  er kjente tall,  $a^*$  og  $b^*$  er minste kvadraters estimatorer for  $a$  og  $b$ , og  $E_1, \dots, E_n$  er uif  $N(0, \sigma)$  med ukjent  $\sigma$ -verdi, er **residualene**  $E_1^*, \dots, E_n^*$  definert som følger:

$$E_i^* = Y_i - (a^* + b^* \cdot x_i), \quad i = 1, \dots, n$$

Med basis i residualene definerer vi **feilkvadratsummen SSE** («Sum of Squares of Error») som kvadratsummen av residualene:

$$\text{SSE} = \sum_{i=1}^n (E_i^*)^2 = \sum_{i=1}^n (Y_i - (a^* + b^* \cdot x_i))^2 = S_y^2 - (b^*)^2 \cdot S_x^2$$

Det kan vises at  $\text{SSE}/\sigma^2$  er kjikvadrat-fordelt med  $n-2$  frihetsgrader, følgelig er

$$S^2 = \frac{\text{SSE}}{n-2}$$

en forventningsrett estimator for  $\sigma^2$ , dvs.  $E(S^2) = \sigma^2$ . 2-tallet i nevner skyldes at vi har *to* ukjente parametre i den lineære regresjonslinja ( $a$  og  $b$ ).

### Eks. 9.2 Beregning av $a^*$ og $b^*$ og $S$

$x$	-1	0	1
$y$	0	4	3

*Oppgave*

Beregn  $a^*$ ,  $b^*$  og  $s$  fra dataene i tabellen.

*Løsningsforslag:*  $\Sigma x = 0$ ,  $\Sigma y = 7$ ,  $\Sigma x^2 = 2$ ,  $\Sigma y^2 = 25$ ,  $\Sigma xy = 3$

$$S_x^2 = \Sigma x^2 - (\Sigma x)^2/n = 2 - 0^2/3 = 2, \quad S_y^2 = \Sigma y^2 - (\Sigma y)^2/n = 25 - 49/3 = 26/3$$

$$S_{xy} = \Sigma xy - \Sigma x \cdot \Sigma y/n = 3 - 0 \cdot 7/3 = 3$$

$$b^* = S_{xy}/S_x^2 = \underline{3/2}, \quad a^* = \bar{y} - b^* \cdot \bar{x} = 7/3 - (3/2) \cdot 0 = \underline{7/3}$$

$$\text{SSE} = S_y^2 - (b^*)^2 \cdot S_x^2 = 26/3 - (3/2)^2 \cdot 2 = 26/3 - 9/2 = 25/6$$

$$s = (\text{SSE}/(n-2))^{1/2} = ((25/6)/(3-2))^{1/2} = \sqrt{25/6} \approx 2.04 \quad \odot$$

## 9.6 Prediksjon av $Y$ når $x$ er gitt

Vår regresjonslinjemodell er

$$\text{E}(Y|x) = y_r = a + bx$$

dvs. for en gitt verdi av  $x$  er forventet  $y$ -verdi lik  $a + bx$ . Siden  $a$  og  $b$  betraktes som ukjente, må disse estimeres med basis i observasjoner, og vi får den estimerte regresjonslinja

$$y_r^* = a^* + b^* x$$

Et formål med å tilpasse en regresjonslinje kan ofte være å prediktere (forutsi) forventet  $y$ -verdi for en gitt  $x$ -verdi uten å måle  $y$ -verdien. Siden  $a^*$  og  $b^*$  er stokastiske variabler blir  $y_r^*$  også en stokastisk variabel. Vi skal her se på de stokastiske egenskapene til  $y_r^*$  som er nødvendige for å bestemme hvor sikkert vi kan anslå  $y_r$  for en gitt  $x$ -verdi.

Vi setter først inn uttrykkene for  $a^*$  og  $b^*$  i uttrykket for  $y_r^*$ :

$$y_r^* = a^* + b^* x = \bar{y} + b^*(x - \bar{x}) = \sum_{i=1}^n \left( \frac{1}{n} + \frac{(x - \bar{x})}{S_x^2} \right) Y_i$$

Med samme begrunnelse som for  $a^*$  og  $b^*$ , ser vi at  $y_r^*$  blir normalfordelt. Vi får følgende forventning :

$$\text{E}(y_r^*) = \text{E}(a^* + b^* x) = \text{E}(a^*) + x\text{E}(b^*) = a + bx$$

$y_r^*$  er med andre ord en forventningsrett estimator for regresjonslinja  $y_r = a + b \cdot x$  for en gitt  $x$ -verdi. Ved noe regning finner vi følgende uttrykk for  $\text{Var}(y_r^*)$ :

$$\text{Var}(y_r^*) = \left( \frac{1}{n} + \frac{(x - \bar{x})^2}{S_x^2} \right) \sigma^2 = \frac{\sigma^2}{n} \left( 1 + \frac{(x - \bar{x})^2}{\frac{1}{n} S_x^2} \right)$$

Fra uttrykket for variansen ser vi at denne er minst for  $x = \bar{x}$  og øker symmetrisk om  $x = \bar{x}$ .

Ser vi på prediksjon av en *enkelt* observasjon  $Y|x$ , må vi ta med usikkerheten til feilreddet  $E$  som er antatt  $N(0, \sigma)$ -fordelt. Det kan da vises at variansen blir som følger:

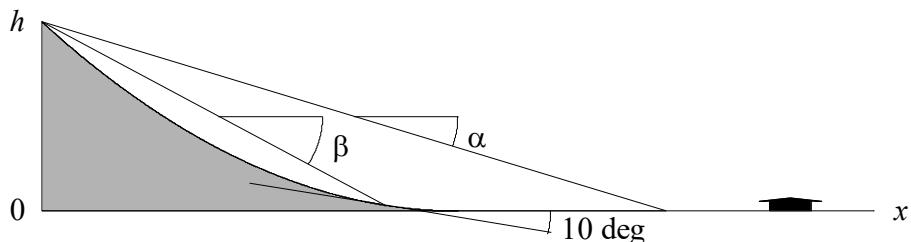
$$\text{Var}(Y|x) = \text{Var}(y_r^*) + \sigma^2 = \left( 1 + \frac{1}{n} \left( 1 + \frac{(x - \bar{x})^2}{\frac{1}{n} S_x^2} \right) \right) \sigma^2$$

Merk at for rimelig store  $n$  vil  $\text{Var}(Y|x) \approx \sigma^2$ : Usikkerheten til en enkeltobservasjon vil normalt være betydelig større enn usikkerheten til den estimerte regresjonslinja.

**Eks. 9.3** **Snøskred.** En mye brukte topografisk snøskredmodell utviklet ved NGI (Norges Geotekniske Institutt) baserer seg på en lineær regresjon mellom stoppvinkelen  $\alpha$  og «topografi»-vinkelen  $\beta$  til skredet (se forenklet skisse i figur). Modellen, som er basert på oppmåling av flere hundre ekstreme skred, er gitt som følger:

$$\alpha = 0,96 \cdot \beta - 1,4^\circ + E$$

der  $\alpha$ ,  $\beta$  og  $E$  er angitt i grader, og  $E$  er  $N(0, \sigma)$  med  $\sigma = 2,30^\circ$ .



### Oppgave

- Forklar hvorfor skredmodellen er et eksempel på den lineære regresjonsmodellen vi har sett på, og angi hva som tilsvarer  $a$ ,  $b$ ,  $x$  og  $Y$ .
- Anta at  $h = 100$  m,  $\beta = 25^\circ$  og at huset er lokalisert i posisjon  $x = 270$  m. Hva er sannsynligheten for at huset blir rammet dersom skredet går?

Gjør rede for forutsetningene du gjør.

### Løsningsforslag

- I skredmodellen har vi en topografisk kjent variabel  $\beta$ , definert som helningen på en linje mellom øvre kant av skredets løsneområde og stedet der terrenget heller  $10^\circ$ . Vinkelen  $\beta$  tilsvarer vår «ikke-stokastiske»  $x$ -variabel. Hvor langt skredet går før det stopper er imidlertid stokastisk og gitt ved vinkelen  $\alpha$ , som tilsvarer vår  $Y$ . Vår regresjonslinje er gitt ved at  $y_r = a + bx$ , slik at i skredmodellen har vi  $a = -1,4^\circ$  og  $b = 0,96$ .

b)  $\beta = 25^\circ \Rightarrow E(\alpha) = 0.96 \cdot 25^\circ - 1.4^\circ = 22.6^\circ$ , og  $\alpha$  er  $N(22.6^\circ, 2.30^\circ)$ -fordelt. Husets  $\alpha$ -vinkel er  $\text{tg}^{-1}(100/270) = 20.3^\circ$ . Vi får da:

$$P(\text{huset rammes}) = P(\alpha < 20.3^\circ) = \Phi((20.3 - 22.6)/2.30) = \Phi(-1.0) \approx 16\%$$

Vi må her forutsette at  $\beta$ -verdien er innenfor det området der den lineære regresjonsmodellen er rimelig. Videre må vi anta at antall observasjoner som ligger bak regresjonslinjen er så stort at vi kan neglisjere usikkerheten til selve regresjonslinjen og til den oppgitte verdien for  $\sigma$ . ☺

Den topografiske skredmodellen i eksemplet ovenfor er et av de mest brukte hjelpebidrifter ved utarbeidelse av faresonekart for snøskred.

## 9.7 Konfidensintervall og hypotesetesting

Vi har tidligere nevnt at  $S^2 = \text{SSE}/(n-2)$  er en forventningsrett estimator for  $\sigma^2$ , hvilket følger av at  $\text{SSE}/\sigma^2$  er  $\text{Kji}2(n-2)$  fordelt.  $a^*$ ,  $b^*$  og  $y_r^*$  er da knyttet til  $t$ -fordelingen med  $n-2$  frihetsgrader som følger:

$$\frac{b^* - b}{S/S_x}, \quad \frac{a^* - a}{S\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_x^2}}}, \quad \text{og} \quad \frac{y_r^* - a - bx}{S\sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{S_x^2}}} \quad \text{er } t_{n-2} \text{ fordelt}$$

Med basis i  $t_{n-2}$ -fordelingen kan vi helt analogt med tidligere oppsett av konfidensintervall, KI, og hypotesetesting, sette opp følgende:

### Konfidensintervall (KI) for og hypotesetesting av $a$

Et  $100(1-\alpha)\%$  KI for  $a$  er gitt ved formelen

$$\left( a^* - t_{\alpha/2} S \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_x^2}}, a^* + t_{\alpha/2} S \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_x^2}} \right)$$

Testobservator:  $T = \frac{a^* - a_0}{S \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_x^2}}}$  for testene

$$H_0: a \leq a_0$$

$$H_0: a \geq a_0$$

$$H_0: a = a_0$$

$$H_1: a > a_0$$

$$H_1: a < a_0$$

$$H_1: a \neq a_0$$

$$\Rightarrow R: T > t_\alpha$$

$$\Rightarrow R: T < -t_\alpha$$

$$\Rightarrow R: |T| > t_{\alpha/2}$$

der  $R$  angir forkastingsområdet for testen på nivå  $\alpha$ , og  $t_\alpha$  og  $t_{\alpha/2}$  er henholdsvis øvre  $\alpha$ - og  $\alpha/2$ -fraktil i  $t$ -fordelingen med  $n-2$  frihetsgrader.

### Konfidensintervall (KI) for og hypotesetesting av $b$

Et  $100(1-\alpha)\%$  KI for  $b$  er gitt ved formelen

$$(b^* - t_{\alpha/2} (S / S_x), b^* + t_{\alpha/2} (S / S_x))$$

Testobservator:  $T = \frac{b^* - b_0}{S / S_x}$  for testene

$$H_0: b \leq b_0$$

$$H_0: b \geq b_0$$

$$H_0: b = b_0$$

$$H_1: b > b_0$$

$$H_1: b < b_0$$

$$H_1: b \neq b_0$$

$$\Rightarrow R: T > t_\alpha$$

$$\Rightarrow R: T < -t_\alpha$$

$$\Rightarrow R: |T| > t_{\alpha/2}$$

der  $R$  angir forkastingområdet for testen på nivå  $\alpha$ , og  $t_\alpha$  og  $t_{\alpha/2}$  er henholdsvis øvre  $\alpha$ - og  $\alpha/2$ -fraktil i  $t$ -fordelingen med  $n-2$  frihetsgrader.

### KI for og hypotesetesting av $E(Y|x) = a + bx$

Et  $100(1-\alpha)\%$  KI for  $E(Y|x)$  er gitt ved formelen

$$\left( a^* + b^*x - t_{\alpha/2} \cdot S \cdot \sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{S_x^2}}, a^* + b^*x + t_{\alpha/2} \cdot S \cdot \sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{S_x^2}} \right)$$

Testobservator:  $T = \frac{a^* + b^*x - y_{r0}}{S \sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{S_x^2}}}$  for testene

$H_0: y_r \leq y_{r0}$	$H_0: y_r \geq y_{r0}$	$H_0: y_r = y_{r0}$
$H_1: y_r > y_{r0}$	$H_1: y_r < y_{r0}$	$H_1: y_r \neq y_{r0}$
$\Rightarrow R: T > t_\alpha$	$\Rightarrow R: T < -t_\alpha$	$\Rightarrow R:  T  > t_{\alpha/2}$

der  $R$  angir forkastingsområdet for testen på nivå  $\alpha$ , og  $t_\alpha$  og  $t_{\alpha/2}$  er henholdsvis øvre  $\alpha$ - og  $\alpha/2$ -fraktil i  $t$ -fordelingen med  $n-2$  frihetsgrader.

#### Eks. 9.4

### Konfidensintervall for og hypotesetest av regresjonsparametere.

Gitt følgende 10  $(x,y)$ -datapar der x-verdiene er konstante:

$x$	1.16	0.08	-0.70	0.06	0.26	-1.45	1.25	0.58	-0.14	-1.27
$y$	6.31	5.27	5.46	6.13	5.73	3.42	5.61	5.25	4.08	4.57

#### Oppgave

- Beregning en regresjonslinje tilpasset dataene i tabellen.
- Tegn dataene og tilpasset regresjonslinje i et spredningsdiagram.
- Beregning 95 % KI for  $a$  og for  $b$ .
- Test om  $b \neq 1$  på testnivå 5 %.
- Beregning 95 % KI for regresjonslinja gitt  $x$ , og tegn disse inn som funksjoner av  $x$  i diagrammet.

#### Løsningsforslag

Vi beregner de tradisjonelle summasjonsuttrykk og finner:

$$\bar{x} = -0.0170, \bar{y} = 5.183, S_x^2 = 7.544, S_y^2 = 7.438, S_{xy} = 5.381$$

- Med basis i verdiene over finner vi følgende verdier for  $b^*$  og  $a^*$ :

$$b^* = S_{xy}/S_x^2 = 5.381 / 7.544 = 0.713$$

$$a^* = \bar{y} - b^* \cdot \bar{x} = 5.183 - 0.713 \cdot (-0.0170) = 5.195$$

Den estimerte regresjonslinja blir da:

$$y_r^* = a^* + b^* \cdot x = 5.195 + 0.713 \cdot x$$

Når vi skal tegne denne inn i et  $(x,y)$ -diagram er det lurt å velge to  $x$ -verdier som ligger i hver sin ende av det området vi har observasjoner. Her kan vi f.eks. velge  $x = -1.5$  og  $x = 1.5$ .

- b) Observasjonene («+»-symboler) og tilpasset regresjonslinje er tegnet inn i Fig. 9.2 nedenfor, med middelverdi-punktet  $(\bar{x}, \bar{y})$  inntegnet som sirkel.
- c) 95 % KI for  $a$  og for  $b$  er gitt ved følgende uttrykk:

$$a^* \mp t_{0,025} \cdot \text{std}^*(a^*), \quad b^* \mp t_{0,025} \cdot \text{std}^*(b^*)$$

$$\text{der } \text{std}^*(a^*) = S \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_x^2}}, \quad \text{std}^*(b^*) = S / S_x$$

og  $t_{0,025}$  er øvre 2,5 %-fraktil i  $t$ -fordelingen med  $n-2 = 10-2 = 8$  frihetsgrader. Vi må følgelig bestemme  $\sigma$ -estimatoren  $S$  og  $t_{0,025}$ , resten av størrelsene som inngår har vi fra før. Fra  $t$ -tabellen bakerst i boka finner vi at  $t_{0.025} = 2.306$ . For å bestemme  $S$  beregner vi først SSE:

$$\begin{aligned} \text{SSE} &= S_y^2 - (b^*)^2 \cdot S_x^2 = 7,438 - (0.713)^2 \cdot 7.544 = 3.599 \\ \Rightarrow S &= \sqrt{\text{SSE} / (n-2)} = \sqrt{\text{SSE} / 8} = \sqrt{3.599 / 8} = 0.671 \end{aligned}$$

Vi finner da at

$$\begin{aligned} \text{std}^*(a^*) &= 0.671 \cdot \sqrt{\frac{1}{10} + \frac{(-0.0170)^2}{7.544}} = 0.212 \\ \text{std}^*(b^*) &= 0.671 / \sqrt{7.544} = 0.244 \end{aligned}$$

som gir følgende 95 % konfidensintervaller:

$$\begin{aligned} (5.195 - 2.306 \cdot 0.212, 5.195 + 2.306 \cdot 0.212) &= (4.71, 5.68) \text{ for } a \\ (0.713 - 2.306 \cdot 0.244, 0.713 + 2.306 \cdot 0.244) &= (0.15, 1.28) \text{ for } b \end{aligned}$$

- d) For å teste om  $b \neq 1$  kan vi benytte testobservatoren

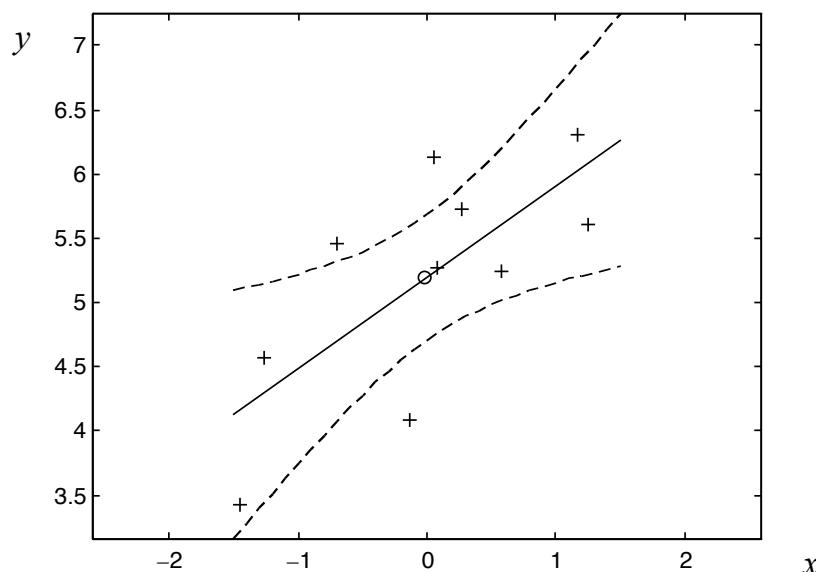
$$\begin{aligned} T &= \frac{b^* - b_0}{S / S_x} = \frac{0.713 - 1}{0.671 / \sqrt{7.544}} = -1.17 \\ \Rightarrow |T| &= 1.17 < t_{0.025} = 2.306 \end{aligned}$$

Konklusjon:  $H_0 (b \neq 1)$  forkastes ikke. Det er ikke grunnlag i data for å påstå på nivå 5 % at  $b \neq 1$

e) 95 % KI for regresjonslinja  $y_r$  når  $x$  er gitt kan bestemmes fra formelen

$$a^* + b^* \cdot x \mp t_{0.025} S \sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{S_x^2}} = \\ 5.195 + 0.713 \cdot x \mp 2.306 \cdot 0.671 \sqrt{\frac{1}{10} + \frac{(x - (-.0170))^2}{7.544}}$$

Merk at konfidensintervallene er funksjoner av  $x$ . For hver  $x$ -verdi må følgelig intervallet beregnes på nytt. Konfidensintervallene er, for enhver  $x$ -verdi, avstanden i  $y$ -retning mellom de stippled linjene i figur. 9.2 ☺



Figur 9.2 Løsning på oppgavene i eks. 9.3. De stippled linjene angir 95 % KI for regresjonslinja (heltrukken linje) for gitt  $x$ . Sirkelen angir middelverdipunktet til observasjonene (+).

## 9.8 Transformasjon av variabler

Selv om det ikke er en klar lineær sammenheng mellom  $x$  og  $y$ , vil ofte en passende transformasjon av enten  $x$  eller  $y$ , eller både  $x$  og  $y$ , føre til en lineær sammenheng mellom de *transformerte* variablene. Vi skal her se på noen av

de vanligste transformasjoner som kan være aktuelle, gjengitt i tabellen nedenfor, og beskrive hvordan vi kan anvende en lineær regresjonsanalyse i slike tilfeller.

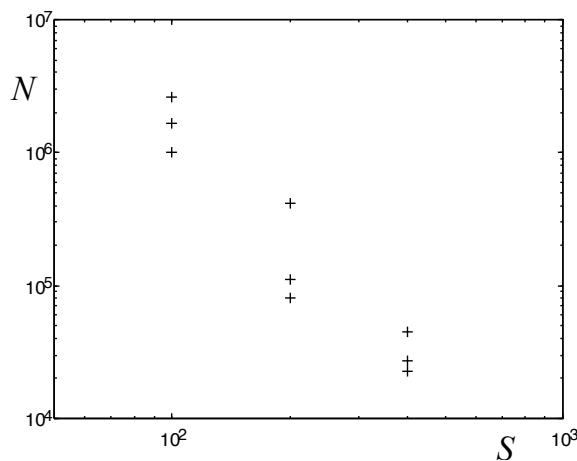
*Tabell 9.1 Noen eksempler på ikke-lineære sammenhenger mellom  $x$  og  $y$ , og tilhørende lineariserende variabeltransformasjoner  $x'$  og  $y'$  slik at  $y' = a' + b' \cdot x'$ .*

ikke-lineær modell	linearisert modell	transformasjoner
$y = f(x; a, b)$	$y' = a' + b' \cdot x$	
1) $y = a \cdot \exp(bx)$	$\log y = \log a + b \cdot x$	$y' = \log y, x' = x, a' = \log a, b' = b$
2) $y = a \cdot x^b$	$\log y = \log a + b \cdot \log x$	$y' = \log y, x' = \log x, a' = \log a, b' = b$
3) $y = 1/(a + bx)$	$1/y = a + b \cdot x$	$y' = 1/y, x' = x, a' = a, b' = b$

Fremgangsmåten er normalt som følger: Først transformeres alle de opprinnelige  $(x, y)$ -dataene til de transformerte  $(x', y')$ -verdiene. Dernest bestemmes minste kvadraters-estimatorene  $a'^*$  og  $b'^*$ . Til slutt bestemmes  $a^*$  og  $b^*$  ved invers transformasjon.

Eks:  $a' = \ln(a) \Rightarrow a'^* = \ln(a^*) \Rightarrow a^* = \exp(a'^*)$ .

**Eks. 9.5** **S-N-diagram.** I figuren under ser vi et såkalt S-N-diagram, som viser hvor mange belastninger  $N$ , med belastning  $S$ , som skulle til på 9 prøver av et konstruksjonselement i stål før materialtretthetsbrudd inntraff.  $S$  er her den uavhengige variablen, mens antall belastninger,  $N$ , til brudd inntreffer er responsvariabel. Merk at begge akser i diagrammet er logaritmiske.



Belastning [MPa]	# belastninger til brudd
$S$	$N$
100	1 666 900
100	999 900
100	2 628 400
200	418 800
200	81 400
200	110 600
400	22 500
400	44 800
400	27 200

*Fig. 9.3 S-N-diagram som viser sammenhengen mellom belastning,  $S$ , og antall belastninger,  $N$ , som skal til før brudd av et konstruksjonselement i stål.*

*Oppgave*

- Hvilken relasjon mellom  $S$  og  $N$  synes rimelig ut fra fig. 9.3?
- Tilpass en regresjonskurve  $N^*(S)$  til dataene.
- Bruk regresjonskurven til å prediktere (forutsi) antall belastninger  $N$  før brudd oppstår når  $S = 300$  MPa.
- Bestem tilnærmede 95 % KI for regresjonskurvens  $N$ -verdi,  $N_{300}$ , når  $S = 300$  MPa.

*Løsningsforslag*

$S$  og  $N$  tilsvarer henholdsvis  $x$  og  $y$  i tab. 9.1. For å få litt enklere betegnelser lar vi  $x$  og  $y$  i løsningsforslaget tilsvare  $x'$  og  $y'$  i tab. 9.1.

- Fra figuren synes en lineær sammenheng mellom  $\log S$  og  $\log N$  å være en rimelig modell. Fra tabell over transformasjoner ser vi at dette tilsvarer sammenhengen

$$N = a \cdot S^b$$

- Å tilpasse en regresjonskurve til dataene basert på modellen  $N = a \cdot S^b$ , tilsvarer å finne verdier  $a^*$  og  $b^*$  for henholdsvis  $a$  og  $b$  som gjør at kurven  $N^* = a^* \cdot S^{b^*}$  passer bra til punktene i figuren over. Vi beregner først  $a'^*$  og  $b'^*$  for den lineære modellen for de transformerte variablene:

$$\log N = \log a + b \cdot \log S = a' + b' \cdot \log S$$

For alle  $S$ - og  $N$ -verdiene i tabellen beregner vi følgelig de transformerte verdiene  $x_i = \log(S_i)$  og  $y_i = \log(N_i)$ ,  $i = 1, 2, \dots, 9$ . Så går vi fram på «vanlig» måte og finner  $a'^*$  og  $b'^*$  ved lineær regresjonsanalyse anvendt på  $x$ - og  $y$ -dataene. Tilslutt finner vi  $a^*$  og  $b^*$  ved å foreta en *omvendt* transformasjon av  $a'^*$  og  $b'^*$ :

$$a'^* = \log(a^*) \Rightarrow a^* = 10^{a'^*}, \quad b'^* = b^* \Rightarrow b^* = b'^*$$

Det kan være hensiktsmessig å sette opp de størrelsene vi trenger i en oversiktlig tabell. Grunnen til at vi tar med så mange siffer er blant annet at beregningen av SSE ellers ville blitt unøyaktig.

$S$	$N$	$x = \log S$	$y = \log N$	$x^2$	$y^2$	$x \cdot y$
100	1 666 900	2.0000	6.2219	4.0000	38.7122	12.4438
100	999 900	2.0000	6.0000	4.0000	35.9995	11.9999

100	2 628 400	2.0000	6.4197	4.0000	41.2124	12.8394
200	418 800	2.3010	5.6220	5.2947	31.6070	12.9364
200	81 400	2.3010	4.9106	5.2947	24.1142	11.2995
200	110 600	2.3010	5.0438	5.2947	25.4395	11.6058
400	22 500	2.6021	4.3522	6.7707	18.9415	11.3246
400	44 800	2.6021	4.6513	6.7707	21.6344	12.1029
400	27 200	2.6021	4.4346	6.7707	19.6654	11.5390
Summer:		$\Sigma x =$ 20.709	$\Sigma y =$ 47.656	$\Sigma x^2 =$ 48.196	$\Sigma y^2 =$ 257.326	$\Sigma xy =$ 108.091

Med basis i verdiene i tabellen over finner vi:

$$\bar{x} = \frac{1}{9} \cdot \Sigma x = \frac{1}{9} \cdot 20.709 = 2.301$$

$$\bar{y} = \frac{1}{9} \cdot \Sigma y = \frac{1}{9} \cdot 47.656/9 = 5.295$$

$$S_x^2 = \Sigma x^2 - \frac{1}{9} (\Sigma x)^2 = 48.196 - \frac{1}{9} \cdot 20.709^2 = 0.545$$

$$S_y^2 = \Sigma y^2 - \frac{1}{9} (\Sigma y)^2 = 257.326 - \frac{1}{9} \cdot 47.656^2 = 4.982$$

$$S_{xy} = \Sigma xy - \frac{1}{9} (\Sigma x \Sigma y) = 108.091 - \frac{1}{9} \cdot 20.709 \cdot 47.656/9 = -1.565$$

Videre finner vi

$$b^* = b' = S_{xy} / S_x^2 = -1.565 / 0.545 = -2.872$$

$$a' = \bar{y} - b' \cdot \bar{x} = 5.295 - (-2.872) \cdot 2.301 = 11.90$$

$$\Rightarrow a^* = 10^{a'} = 10^{11.90}$$

Den søkte regresjonsmodellen blir derfor:

$$N_r^* = 10^{11.90} \cdot S^{-2.872}, \text{ der } S \text{ er forutsatt angitt i MPa.}$$

c)  $\underline{N_r^*(300)} = 10^{11.91} \cdot 300^{-2.872} = \underline{61000}$

d) For å beregne et 95 % KI for  $N_{300}$  tar vi utgangspunkt i den transformerte modellen. Det vil si at vi antar modellen  $Y_i = a' + b'x_i + E_i'$ , der  $x_i = \log(S_i)$ ,  $Y_i = \log(N_i)$  og  $E_1', \dots, E_9'$  er antatt å være uif  $N(0, \sigma)$ -fordelte med konstant  $\sigma$ -verdi uavhengig av  $x = \log S$ . Vi lar  $x_{300} \equiv \log(300)$  og  $y_{300} \equiv \log(N_{300})$ , og beregner først et 95 % KI,  $(L', U')$ , for  $y_{300}$ . Et 95 % KI,  $(L, U)$ , for  $N_{300}$  finnes deretter ut fra følgende sammenhenger:

$$P(L' < y_{300} < U') = P(L' < a' + b' \cdot x_{300} < U') =$$

$$P(10^{L'} < a \cdot 300^b < 10^{U'}) = P(L < N_{300} < U)$$

Når vi har funnet  $L'$  og  $U'$  ser vi fra ligningene ovenfor at  $L = 10^{L'}$  og  $U = 10^{U'}$ . For å beregne  $L'$  og  $U'$  trenger vi et estimat for  $\sigma$  og øvre 2.5 %-fraktil,  $t_{0.025}$ , i  $t$ -fordelingen med  $n-2 = 9-2 = 7$  frihetsgrader. Fra tabellen bakerst i boka finner vi at  $t_{0.025} = 2.365$ . Vi finner  $s$  som følger:

$$\begin{aligned} \text{SSE} &= S_y^2 - (b^{*})^2 \cdot S_x^2 = 4.982 - (-2.872)^2 \cdot 0.545 = 0.487 \\ \Rightarrow s &= \sqrt{\text{SSE} / (n-2)} = \sqrt{0.487 / 7} = 0.264 \end{aligned}$$

Vi får følgelig at

$$a^{*} + b^{*} \cdot x_{300} \mp t_{0.025} \cdot s \cdot \sqrt{\frac{1}{9} + \frac{(x_{300} - \bar{x})^2}{S_x^2}} = (4.530, 5.042)$$

er et tilnærmet 95 % KI for  $y_{300}$ . Dette gir at

$$(10^{L'}, 10^{U'}) = (10^{4.530}, 10^{5.042}) = (33000, 110000)$$

er et tn. 95 % KI for  $N_{300}$ . ☺

## 9.9 Residualer og modellsjekk

Vi har så langt sett på modellen  $Y_i = a + b \cdot x_i + E_i$  der  $E_1, \dots, E_n$  er antatt å være uif  $N(0, \sigma)$ -fordelte med konstant og ukjent  $\sigma$ -verdi som ikke varierer med  $x$ . Nå skal vi se på metoder for å sjekke om en slik modell er rimelig. Før vi ser på analyse av residualene, skal vi se nærmere på den empiriske korrelasjonskoeffisienten,  $r$ , mellom  $x$  og  $y$ :

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \cdot \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}} = \frac{S_{xy}}{S_x S_y}$$

I kap. 1 så vi at  $-1 \leq r \leq 1$ , og at  $r$ -verdier i nærheten av  $-1$  eller  $1$  indikerte en sterk lineær sammenheng mellom  $x$  og  $y$ .  $r$ -verdien er derfor et nyttig, men utilstrekkelig mål på hvor god den lineære modellen er, og det kan være en stor fallgrube å tro at vi er «i mål» om vi får en  $r$ -verdi svært nær  $-1$  eller  $1$ . Det kan for eksempel være at datasettet vårt inneholder en **utligger** (dvs. et datapar med «ville» verdier i forhold til resten av dataene) som, dersom den var blitt fjernet eller korrigert, ville ført til en  $r$ -verdi i nærheten av null. Omvendt kan en utligger også være årsaken til at vi får en  $r$ -verdi i nærheten av null, i tilfeller der  $|r|$  er nær  $1$ , dersom utliggeren fjernes eller korrigeres.

Kombinerer vi beregning av  $r$ -verdi med et spredningsdiagram hvor  $(x,y)$ -dataene og tilpasset regresjonslinje er inntegnet, får vi en god indikasjon på hvor godt en rett linje er tilpasset data, samt at vi lett kan se om datamaterialet er «homogent» (ikke inneholder «utliggere» eller klart adskilte populasjoner).

Selv om en regresjonslinje visuelt ser ut til å passe svært godt til data, samt at  $r$ -verdien er svært nær 1 eller  $-1$ , må vi likevel være forsiktige før vi kan stole på om f.eks. konfidensintervaller og tester har riktig nivå. Formlene for KI og tester i dette kapitlet baserer seg, i tillegg til forutsetningen om at en lineær modell er rimelig, på følgende sentrale forutsetninger som bør sjekkes:

1. Uavhengighet mellom feilreddene  $E_1, \dots, E_n$ .
2. Feilreddene har konstant varians,  $\text{Var}(E_i) = \sigma^2$ ,  $i = 1, \dots, n$ , uavhengig av  $x$ -verdi.
3. Feilreddene er normalfordelte.

Vi skal nå se på noen enkle metoder basert på analyse av residualene, for å sjekke rimeligheten av de to første (og viktigste) forutsetningene over. Vedrørende normalitetsforutsetningen skal vi nøye oss med å henvise til at det finnes en rekke tester for normalitet, der blant annet Pearsons kjikvadrattest kan benyttes. Det skal imidlertid presiseres at vår lineære regresjonsmodell er rimelig robust med hensyn på avvik fra normalitetsforutsetningen. Dette skyldes at minste kvadraters-estimatorene  $a^*$  og  $b^*$  inneholder en lineær-kombinasjon av  $Y_1, \dots, Y_n$ . For rimelig store  $n$ -verdier ( $n > \text{ca. } 20$  som en forsiktig tommelfingerregel) blir da  $a^*$  og  $b^*$  tilnærmet normalfordelte ifølge sentralgrenseteoremet, selv om feilvariabelen  $E_i$  følger en fordeling som avviker betydelig fra normalfordelingen.

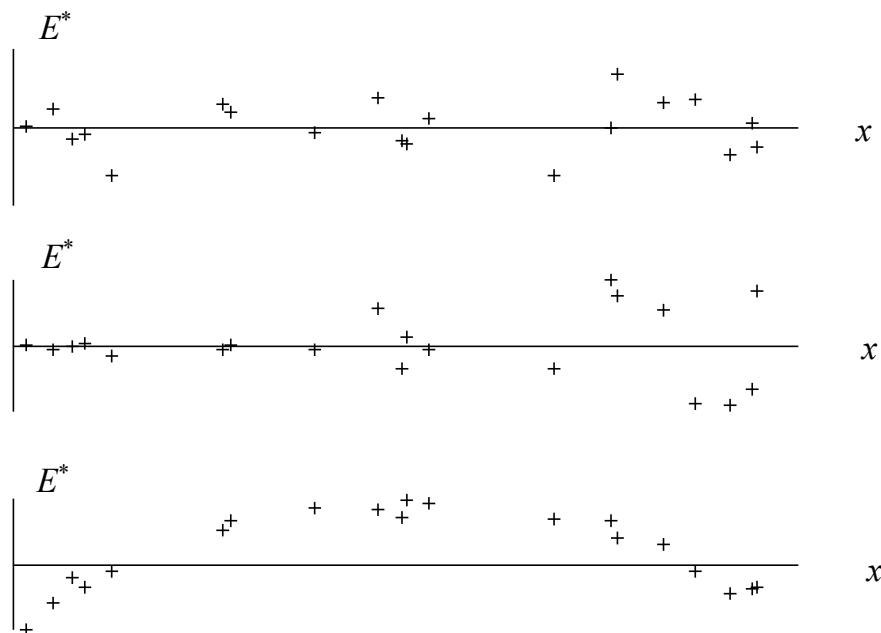
Vi skal også se at kvadratet,  $r^2$ , av den empiriske korrelasjonskoeffisienten angir hvor stor del av total kvadratsum,  $S_y^2$ , til  $y$ , som «forklares» av den lineære modellen, og hvor stor del som er «uforklart», angitt ved residual-kvadratsummen («feil»-kvadratsummen) SSE.

I kap. 9.6 vedrørende  $\sigma$ -estimatoren  $S$  definerte vi begrepet *residual*. Vi kan også beskrive residualene som følger:

$$\begin{array}{lll} Y_i & = & (a^* + b^* \cdot x_i) + (Y_i - a^* - b^* \cdot x_i) \\ \text{Observeret } y\text{-verdi} & & \text{Forklart av lineær relasjon} & \text{Residual} = \text{avvik fra lineær relasjon} \end{array}$$

Siden  $E_i^* = Y_i - a^* - b^* \cdot x_i$  kan omskrives som en lineær kombinasjon av  $Y_1, \dots, Y_n$ , som er uavhengige og normalfordelte, vil også  $E_i^*$  være

normalfordelt,  $i = 1, 2, \dots, n$ . Det kan vises at når  $n$  går mot uendelig, vil  $E_1^*, \dots, E_n^*$  gå mot uavhengige  $N(0, \sigma)$ -variabler. For rimelig store  $n$  kan vi betrakte  $E_i^*$  som en estimator for  $E_i$ . For å undersøke forutsetningene (uavhengighet og konstant varians) vedrørende  $E_1, \dots, E_n$ , ser vi derfor på om  $E_1^*, \dots, E_n^*$  ser ut til å tilfredsstille de samme forutsetninger. Vi skal her nøye oss med en helt enkel grafisk metode, som består i å plotte residualene langs vertikalaksen mot tilhørende  $x$ -verdier langs horisontalaksen. Neste figur skulle illustrere hva vi kan oppnå med dette:



Figur 9.4 Residualplott. Øverste diagram indikerer uavhengige residualer med rimelig konstant varians. Midterste diagram indikerer uavhengige residualer der variansen øker med  $x$ . Nederste diagram indikerer avhengige residualer som kan skyldes en ikke-lineær ( $x, y$ )-sammenheng.

Det øverste diagrammet av figuren viser et plott som ikke indikerer noe klart avvik fra våre forutsetninger. I midterste del av figuren ser vi imidlertid klart at forutsetningen om konstant varians er brutt, da spredningen øker med økende  $x$ -verdier. Nederste diagram viser en situasjon som indikerer sterkt avhengighet mellom suksessive residualer. Krumningen indikerer at dette kan skyldes en ikke-lineær, f.eks. kvadratisk, sammenheng mellom  $x$  og  $y$ , og at den lineære modellen trolig er lite realistisk.

La oss til slutt se på sammenhengen mellom residualene og kvadratet,  $r^2$ , av empirisk korrelasjonskoeffisient. Feilkvadratsummen SSE er kvadratsummen

av residualene, og det kan vises at den kan deles opp i følgende to additive ledd:

$$\text{SSE} = \sum_{i=1}^n (Y_i - a^* - b^* \cdot x_i)^2 = S_y^2 - r^2 S_y^2$$

Ligningen ovenfor kan omskrives og tolkes som følger:

$S_y^2$	$=$	$r^2 S_y^2$	$+$	SSE
<i>total kvadratsum</i>		<i>kvadratsum forklart av lineær relasjon</i>		<i>kvadratsum av residualer</i>

Vi ser med andre ord at  $r^2$  direkte angir hvor stor del av total kvadratsum,  $S_y^2$ , som forklares av den lineære modellen, mens SSE angir feilkvadratsummen. Jo nærmere  $r^2$ -verdien er 1, jo mindre andel bidrar feilreddet SSE til  $S_y^2$ , og jo mer tiltro vil vi ha til den lineære modellen. Fra betraktingen ovenfor kan en utvikle tester, basert på variansanalyse, for om den lineære modellen er rimelig. Dette skal vi ikke gå nærmere inn på her.

## Konklusjon

Vi oppsummerer dette underkapitlet med følgende tommelfingerregler for hvordan en kan sjekke den beskrevne lineære modellen:

1. Tegn  $(x,y)$ -dataene i et spredningsdiagram med riktig skalering ( $x/S_x$  mot  $y/S_y$  i stedet for  $x$  mot  $y$  vil automatisk gi en fornuftig skalering), og vurder på øyemål om en lineær sammenheng mellom  $x$  og  $y$  synes fornuftig. Se særlig etter om datamaterialet virker homogent og er uten «utliggere», som kun bør fjernes dersom de åpenbart er gale.
2. Dersom en lineær modell ut fra punkt 1 virker fornuftig, beregnes korrelasjonskoeffisienten,  $r$ . Dersom  $|r|$  er rimelig nær 1, f.eks.  $|r| > .9$ , er det grunn til å feste lit til den lineære modellen. Hvor streng grense vi bør sette for  $r$  avhenger av situasjonen.
3. Plott residualene som funksjon av  $x$  og se på øyemål om det er klare avvik fra forutsetningene om uavhengighet og konstant varians. Hvis en lineær modell synes fornuftig selv om en eller begge forutsetninger ikke synes å være tilfredsstilt, bør en vurdere nøye om ikke en mer avansert lineær modell (f.eks. inneholdende en modell for  $\sigma$  som funksjon av  $x$ ) bør brukes, særlig hvis et nøyaktig mål på konfidensintervaller og tester er påkrevet.

Helt til slutt minner vi leseren på at en sterk korrelasjon mellom to variabler ikke nødvendigvis indikerer noen årsak-virkning sammenheng mellom

variablene. Det kan være en tredje «lurevariabel» som gjør at to variabler er sterkt korrelerte. Dette ble behandlet i kap. 1 (stikkord: stork-baby eksempel).

## 9.10 Oppgaver

**9.1** Gitt følgende  $(x, y)$ -data:

$x$	29.0	18.6	3.5	30.4	13.5	19.6	12.6
$y$	2.48	1.66	0.44	2.77	1.94	1.82	1.09

a) Tegn verdiene inn i et spredningsdiagram.

b) Beregn  $\bar{x}$  og  $\bar{y}$ , og trekk på øyemål den rette linja gjennom  $(\bar{x}, \bar{y})$  som ser ut til å passe best mulig til data.

Anta i det følgende at  $x$ -ene er konstante, kjente tall uten statistisk usikkerhet, mens  $y$ -ene er uavhengige realiseringer av en stokastisk variabel  $Y$  som er  $N(a+bx, \sigma^2)$ -fordelt.

c) Beregn regresjonslinja  $y_r^* = a^* + b^* \cdot x$  og tegn denne inn i diagrammet. Får du bra overensstemmelse med linja du trakk på øyemål?

d) Estimer  $\sigma$ .

e) Estimer  $\text{std}(a^*)$ ,  $\text{std}(b^*)$  og  $\text{std}(y_r^* | x = 10.0)$ .

f) Anta at du har 4 uavhengige  $y$ -observasjoner i stedet for bare en for hver  $x$ -verdi. Med hvilken faktor vil  $\text{std}(b^*)$  og  $\text{std}(a^*)$  avta? Hva med  $\text{Var}(S^2)$ ?

**9.2** Sammenhengen mellom en plutselig forandring (sprang) i input til et 1. ordens reguleringssystem og output  $y$  (sprangrespons) fra systemet er

$$(1) \quad y = a(1 - e^{-t/T})$$

der  $t$  er tiden fra forandringen trer i kraft,  $T$  er en tidkonstant som er mindre jo raskere systemet er og  $a > 0$  er output når  $t$  går mot uendelig.

a) Anta at  $a$  er kjent og  $T$  ukjent.

Transformer den ikke-lineære sammenhengen i lign. (1) over til en lineær sammenheng

$$(2) \quad y' = a' + b' \cdot t$$

Hva blir sammenhengen mellom de transformerte størrelsene  $y'$ ,  $a'$  og  $b'$  og de opprinnelige:  $y$ ,  $a$  og  $T$ ?

b) Anta fremdeles at  $a$  er kjent og bestem et uttrykk for minste kvadraters-estimatoren  $b^{**}$  for  $b'$ . Hva blir den tilsvarende estimatoren  $T^*$  for  $T$ ?

c) Gitt følgende sammenhengende verdier for  $t$  [sek] og  $y$  [V]:

$t$	1	2	3	4	5
$y$	1.85	2.98	5.96	6.09	7.32
$t$	6	7	8	9	100
$y$	7.44	7.65	8.20	8.29	10.00

Undersøk om det er grunn til å anta at systemet er et 1. ordens system, og estimer i såfall tidkonstanten  $T$  for systemet. Nevn en fallgrube du kan gå i her.

**9.3** En temperaturmåler viser en bortimot perfekt lineær sammenheng mellom temperatur,  $X$ , og utslag på instrumentet,  $Y$  (alle størrelser i  $^{\circ}\text{C}$ ):

$$(3) \quad Y = X + W$$

der  $W$  er en  $N(a, \sigma^2)$ -variabel.  $W$  ved et tidspunkt  $t_1$  er uavhengig av  $W$  ved et annet tidspunkt  $t_2$ . Imidlertid viser det seg at temperaturmåleren er noe ukalibrert. For å kalibrere instrumentet, har fabrikanten i løpet av en uke målt

temperaturen ved  $0^{\circ}\text{C}$  og funnet følgende med basis i  $n = 1460$  uavhengige observasjoner:

$$\bar{y} = 0,34^{\circ}\text{C}, \quad S_y = 1,23^{\circ}\text{C}$$

- a) Hvilken av de angitte størrelser beskriver hvor ukalibrert instrumentet er?
- b) Bestem et uttrykk for minste kvadraters estimatoren  $a^*$  for  $a$  og beregn  $a^*$ -verdien med basis i fabrikantens måleresultater.
- c) Estimer standardavviket til  $a^*$  med basis i fabrikantens målinger.

**9.4** I situasjonene beskrevet nedenfor skal du, om mulig, identifisere hva som er uavhengig variabel ( $x$ ) og hva som er avhengig responsvariabel ( $y$ ):

- a) I et medisinsk forsøk undersøkes sammenhengen mellom styrke av en sovepille og antall søvntimer.
- b) I et havbruksanlegg måles fiskevekst som funksjon av fôrmengde.
- c) På en oljeplattform måles hver enkelt bølgehøyde samtidig som den resulterende variasjon i strekkspenningen i et konstruksjons-element på plattformen måles.
- d) I et tilfeldig utvalg på 1000 elever fra videregående skole sammelignes matematikk- og engelsk-karakterene på individbasis.

**9.5** Gitt følgende  $(x,y)$ -data:

$x$	-1	0	1
$y$	-1	0	1

- a) Beregn korrelasjonskoeffisienten  $r$ .

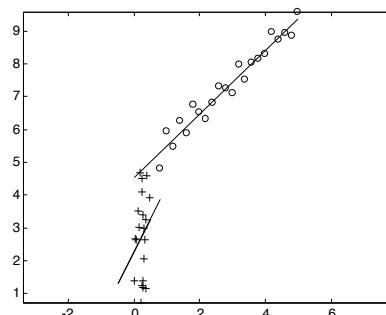
- b) Legg til to nye datapar  $(x,y) = (0,a)$  og  $(x,y) = (0,-a)$ , og vis at  $r$  nå blir  $r = (1+a^2)^{-1/2}$ . Hva går  $r$  mot når  $a$  går mot uendelig?

**9.6** Gitt følgende  $(x,y)$ -data:

$x$	-1	1	1	-1
$y$	-1	-1	1	1

- a) Beregn korr.koeffisienten  $r$ .
- b) Legg til to nye datapar  $(x,y) = (-a,-a)$  og  $(x,y) = (a,a)$ , og vis at  $r$  nå blir  $r = (1 + 2/a^2)^{-1}$ . Hva går  $r$  mot når  $a$  går mot uendelig?

**9.7** I figuren nedenfor ser dataene ut til å følge to forskjellige rette linjer. Det er tilpasset en rett linje til dataene merket «+» og en rett linje tilpasset dataene merket med sirkler.



Kan du gi en forklaring på hvorfor linja tilpasset «+»-dataene tilsynelatende gir en så dårlig tilpasning?

**9.8** Utled uttrykkene  $b^* = S_{xy}/S_x^2$  og  $a^* = \bar{y} - b^* \cdot \bar{x}$  for minste kvadraters-estimatorene  $a^*$  og  $b^*$ , ved å løse de to ligningene du får ved å derivere minste kvadratsummen

$$Q = \sum (y_i - a - bx_i)^2$$

med hensyn på  $a$  og  $b$ , og sette de deriverte lik null.

### 9.9 Vis at

$$\begin{aligned}\Sigma(x_i - \bar{x})(y_i - \bar{y}) &= \Sigma(x_i - \bar{x})y_i \\ &= \Sigma(y_i - \bar{y})x_i\end{aligned}$$

der summen går fra  $i = 1$  til  $i = n$ .

### 9.10 Vis at

$$\Sigma\left(\frac{1}{n} - \frac{\bar{x}}{S_x^2}(x_i - \bar{x})\right)^2 = \frac{1}{n} + \frac{(\bar{x})^2}{S_x^2}$$

der summen går fra  $i = 1$  til  $i = n$ .

### 9.11 Vis at

$$\text{Cov}(a^*, b^*) = \frac{-\bar{x}}{S_x^2} \sigma^2$$

Anta at vi hadde omdefinert vår regresjonsmodell som følger:

$$y_r = a + b \cdot (x - \bar{x})$$

Hva ville nå minste kvadraters-estimatorene  $a^*$  og  $b^*$  blitt, og hva ville  $\text{Cov}(a^*, b^*)$  blitt i dette tilfellet?

### 9.12 Gitt residualene

$$E_i^* = Y_i - a^* - b^* \cdot x_i, \quad i = 1, 2, \dots, n$$

der  $b^* = S_{xy}/S_x^2$  og  $a^* = \bar{Y} - b^* \cdot \bar{x}$  i den lineære regresjonsmodellen  $Y_i = a + b \cdot x_i + E_i, i = 1, 2, \dots, n$ .  $E_1, E_2, \dots, E_n$  er antatt å være uif  $N(0, \sigma)$ -variable.

Vis at

$$\text{Var}(E_i^*) = \sigma^2 \left( 1 - \frac{1}{n} \left( 1 + \frac{(x_i - \bar{x})^2}{S_x^2/n} \right) \right)$$

og at

$$\begin{aligned}\text{Cov}(E_i^*, E_j^*) &= \\ &- \frac{\sigma^2}{n} \left( 1 + \frac{(x_i - \bar{x})(y_i - \bar{y})}{S_x^2/n} \right)\end{aligned}$$

**9.13 (E)** Tabellen viser middeltemperaturen i Bergen de siste 30 år.  $x_i$  er år og  $Y_i$  er temperatur i  $^{\circ}\text{C}$ .

$x_i$	$Y_i$	$x_i$	$Y_i$
1961	7.7	1976	7.7
1964	7.3	1979	7.2
1967	7.3	1982	7.7
1970	7.4	1985	7.4
1973	7.3	1988	7.9

Verdiene i tabellen er middeltemperaturen over 3-årsperioder. Temperaturen  $7.4$   $^{\circ}\text{C}$  for 1970 betyr middeltemperaturen for årene 1969-1970-1971. (Dette har ingen betydning for bestemmelse av regresjonslinjen). De 8 første temperaturene er målt på Frediksberg, de to siste på Florida. Vi velger å ignorere dette i denne sammenheng.

#### a) Bestem regresjonslinja

$$y_r^*(x) = a^* + b^* \cdot x$$

ved en lineær regresjon. Tegn spredningsdiagram og tegn linja inn i diagrammet. Det oppgis at

$$\begin{aligned}S_x^2 &= \sum_{i=1}^{10} (x_i - \bar{x})^2 = 742.5 \\ \sum_{i=1}^{10} (x_i - \bar{x}) y_i &= 6.45\end{aligned}$$

b) Er det grunnlag for, på bakgrunn av den gitte tabellen, å hevde at temperaturen i Bergen viser en stigende tendens?

Test  $H_0: b = 0$  mot  $H_1: b > 0$

Benytt et signifikansnivå på 5 %

Det oppgis at:

$$\text{SSE} = \sum_{i=1}^{10} (Y_i - a^* - b^* x_i)^2 = 0.4530$$

**9.14 (E)** I en fremstillingsprosess for syntetiske fibre til bruk i tekstiler, inngår krymping av fibrene ved koking i trykkoker. For å finne ut hvordan temperaturen innvirker på krympningsgraden, utføres krympning av fiberprøver ved 11 forskjellige temperaturer  $x_i$  ( $^{\circ}\text{C}$ ),  $i = 1, 2, \dots, 11$ . Krympningsgraden  $Y_i$  (%) avleses hver gang. Resultatet ble:

$x_i$	120	122	124	126	128	130
$Y_i$	3.3	3.7	3.9	4.1	4.5	4.4
$x_i$	132	134	136	138	140	
$Y_i$	4.8	5.0	5.6	5.5	5.9	

Vi antar en lineær regresjonsmodell er brukbar, dvs. at

$$EY_i = a + bx_i \quad (a, b \text{ ukjente parametre})$$

Vi antar dessuten at  $Y_1, \dots, Y_{11}$  er uavhengige og normalfordelte med varians  $\sigma^2$ .

- a) Bestem den estimerte regresjonslinja  $y_r^* = a^* + b^*x$ . Tegn spredningsdiagram og tegn den estimerte regresjonslinja inn i diagrammet.
- b) Test hypotesen  $H_0: b = 0.15$  mot alternativet  $H_1: b \neq 0.15$ . Bruk 5 % signifikansnivå.

Regnehjelp:

$$\sum_{i=1}^{11} (x_i - \bar{x})^2 = 440, \quad \sum_{i=1}^{11} (x_i - \bar{x}) Y_i = 54.8$$

$$\text{SSE} = \sum_{i=1}^{11} (Y_i - a^* - b^* x_i)^2 = 0.165$$

**9.15 (E)** En forskergruppe vil

undersøke hvordan proteininnholdet i foret påvirker veksten av regnbueørret. Det ble utført forsøk med 10 grupper ørret, hvor hver gruppe fikk forskjellig andel protein,  $x_i$  (%), i foret. Gjennomsnittlig tilvekst  $Y_i$  i løpet av en gitt periode ble målt i hver gruppe. Resultatet ble:

$x_i$	18	22	26	30	34
$Y_i$	178	215	223	244	232
$x_i$	38	42	46	50	54
$Y_i$	255	248	261	264	259

- a) Bestem empirisk korrelasjonskoeffisient  $r$

$Y_1, Y_2, \dots, Y_{10}$  antas uavhengige og normalfordelte med samme ukjente varians  $\sigma^2$  og forventningsverdier  $E(Y_i) = a + bx_i$ ,  $i = 1, 2, \dots, 10$ .

- b) Bestem den estimerte regresjonslinja  $y_r^* = a^* + b^*x$  ved lineær regresjon.
- c) Tegn spredningsdiagram og tegn regresjonslinja inn i diagrammet.
- d) Finn et 95 % konfidensintervall for den sanne regresjonskoeffisienten  $b$ .

Regnehjelp:

$$S_x^2 = \sum_{i=1}^{10} (x_i - \bar{x})^2 = 1320$$

$$\sum_{i=1}^{10} (y_i - \bar{y})^2 = 6507.6$$

$$\sum_{i=1}^{11} (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{11} (x_i - \bar{x}) Y_i = 2608$$

$$\text{SSE} = \sum_{i=1}^{11} (Y_i - a^* - b^* x_i)^2 = 1354.8$$

**9.16 (E)** Tabellen viser flygods ( $x_i$ ) og tilsvarende fortjeneste ( $Y_i$ ) for 10 flyselskap (data fra 1973).

Flyselskap	flygods ( $10^9$ kg)	fortj. (mill kr)
Pan American	860	188
Flying Tiger	681	120
United	645	135
American	529	114
TWA	475	98
Seaboard	359	53
North West	246	52
Eastern	207	56
Delta	176	56
Continental	144	29

periode på 12 mndr. ( $x_1, x_2, \dots, x_{12}$ ) vært som følger (i mrd kr):

$x_i$	1	2	3	4	5	6
$Y_i$	2.0	1.9	1.8	2.1	2.0	1.9
$x_i$	7	8	9	10	11	12
$Y_i$	2.2	2.3	2.1	2.2	2.3	2.4

- a) Vis at den estimerte regresjonslinja ved en lineær regresjon er gitt ved  $y_r^* = 1.827 + 0.042 \cdot x$ , når det opplyses at

$$\sum_{i=1}^{12} (x_i - \bar{x})^2 = 143 \quad \sum_{i=1}^{12} (x_i - \bar{x}) Y_i = 6$$

- b) Lag spredningsdiagram og tegn inn den estimerte regresjonslinja.
- c) La  $b$  være stigningstallet for den samme regresjonslinja, slik at  $EY_i = a + bx_i$ . Lag et 95 % konfidensintervall for  $b$ . Oppgitt:

$$\sum_{i=1}^{12} (y_i - 1.827 - 0.042 x_i)^2 = 0.128$$

- a) Tegn spredningsdiagram.

Anta at en lineær regresjonsmodell er brukbar, dvs. at  $E(Y_i) = a + bx_i$ . Finn den estimerte regresjonslinja  $y_r^* = a^* + b^* x$ .

- b) Det blir hevdet at  $b = 0.250$ , men vi har mistanke om at  $b$  er mindre.

Test  $H_0: b = 0.250$  mot  $H_1: b < 0.250$  med 5 % signifikansnivå.

Regnehjelp:

$$\sum_{i=1}^{10} (x_i - \bar{x}) y_i = 105464$$

$$\sum_{i=1}^{10} (x_i - \bar{x})^2 = 540842$$

$$\sum_{i=1}^{10} (y_i - a' - b^* x_i)^2 = 1489$$

- 9.17 (E) Oljeinntektene  $Y_i$  fra en bestemt del av Nordsjøen har over en

## 9.11 Formelsamling

### Modell:

Observasjoner (responsvariabler):

$$Y_i = a + b \cdot x_i + E_i, i = 1, \dots, n$$

$x_1, \dots, x_n$  = kjente tall (prediktorer)

$E_1, \dots, E_n$  uif  $N(0, \sigma)$ ,  $\sigma$  ukjent

$y_r = a + b \cdot x$  = regresjonslinje

### Summasjonsformler

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

$$\bar{Y} = \frac{1}{n} \sum_{i=1}^n Y_i$$

$$S_x^2 = \Sigma(x_i - \bar{x})^2 = \Sigma x^2 - \frac{1}{n} (\Sigma x)^2$$

$$S_y^2 = \Sigma(y_i - \bar{y})^2 = \Sigma y^2 - \frac{1}{n} (\Sigma y)^2$$

$$S_{xy} = \Sigma(x - \bar{x})(y - \bar{y}) = \Sigma xy - \frac{1}{n} \Sigma x \Sigma y$$

### Empirisk korrelasjonskoeffisient, $r$

$$r = \frac{S_{xy}}{S_x S_y}$$

### Minste kvadraters estimatorer $a^*$ og $b^*$

$$b^* = S_{xy} / S_x^2 \text{ er } N(b, \sigma / S_x)$$

$$a^* = \bar{Y} - b^* \cdot \bar{x} \text{ er}$$

$$N(a^*, \sigma \sqrt{\frac{1}{n} + (\bar{x}^2 / S_x^2)})$$

### Feilkvadratsum SSE og estimator $S^2$

Residual:  $E_i^* = Y_i - a^* - b^* x_i, i = 1, \dots, n$

$$S^2 = \text{SSE} / (n - 2) \text{ der}$$

$$\text{SSE} = \Sigma(E_i^*)^2 = S_y^2 - (b^*)^2 S_x^2$$

$$E S^2 = \sigma^2, \quad \text{Var}(S^2) = 2\sigma^4 / (n - 2)$$

### Estimator $y_r^*$ for $y_r = a + b \cdot x$

$$y_r^* = a^* + b^* \cdot x \text{ er}$$

$$N\left(y_r^*, \sigma \sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{S_x^2}}\right)$$

### Estimator for prediksjon av $Y | x$

$$Y|x \text{ er } N\left(y_r^*, \sigma \sqrt{1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{S_x^2}}\right)$$

### 100(1- $\alpha$ )% konfidensintervall (KI)

for  $\theta = a, b$  eller  $y_r$ :

$$\theta^* \mp t_{\alpha/2} \cdot \text{std}^*(\theta^*)$$

der  $\theta^* = a^*, b^*$  eller  $y_r^*$ , og  $t_{\alpha/2}$  er øvre  $\alpha/2$ -fraktil i  $t$ -fordelingen med  $n-2$  frihetsgrader, og  $\text{std}^*$  betyr at vi erstatter  $\sigma$  med  $S$  i  $\text{std}(\theta^*)$ -uttrykket.

Eks:  $\theta^* = b^* \Rightarrow \text{std}^*(\theta^*) = S / S_x$ .

**Hypotesetester** for  $\theta = a, b$  eller  $y_r$ .

$$\text{Testobservator: } T = \frac{\theta^* - \theta_0}{\text{std}^*(\theta^*)}$$

Tester:

$$H_0: \theta \leq \theta_0, H_1: \theta > \theta_0 \Rightarrow R: T > t_\alpha$$

$$H_0: \theta \geq \theta_0, H_1: \theta < \theta_0 \Rightarrow R: T < -t_\alpha$$

$$H_0: \theta = \theta_0, H_1: \theta \neq \theta_0 \Rightarrow R: |T| > t_{\alpha/2}$$

der  $t_{\alpha/2}$  og  $t_\alpha$  er henholdsvis øvre  $\alpha/2$ -og øvre  $\alpha$ -fraktil i  $t_{n-2}$ -fordelingen

## **Kapittel 10**

# **Variansanalyse**

### **10.1 Innledning**

La oss starte med to eksempler der variansanalyse kan være et nyttig verktøy:

**Eks. 10.1** I en markedsundersøkelse ønsker man å sammenligne forventet varighet av 60W lyspærer fra 4 forskjellige fabrikker, som alle hevder å ha samme forventede varighet på lyspærene. Et tilfeldig utvalg på 100 lyspærer trekkes fra hver av fabrikantene. ☺

**Eks. 10.2** En gårdbruker ønsker å teste virkningen på potetavlingen av 3 forskjellige gjødslinger på en åker. Da han ikke er sikker på om åkeren har like egenskaper overalt, deler han åkeren inn i 12 deler, og bruker hver av de 3 gjødslingene på 4 forskjellige deler. Hvilke deler som får hvilken gjødsling bestemmes tilfeldig. ☺

Før vi kommenterer eksemplene, innfører vi følgende terminologi:

#### **Terminologi**

**Forsøk:** Planlagt undersøkelse for å kartlegge effekten av ulike behandlinger.

**Forsøksenheter:** De enheter vi mäter respons på.

**Forsøksfaktorer:** De forskjellige typer kontrollerbare forsøksbetingelser der effekten ønskes undersøkt.

**Faktornivåer:** De ulike nivåer forsøksfaktorene har i forsøkene.

**Behandling:** En spesifikk kombinasjon av faktornivåer for de forskjellige forsøksfaktorene som undersøkes.

**Replikater:** Ulike forsøksenheter som utsettes for samme behandling.

**Responsvariabel:** Den målevariabel/ observasjonsvariabel som angir effekten av en behandling på hver forsøksenhet.

I eks. 10.1 er forsøksenheterne de 400 lyspærene. Det er kun en forsøksfaktor, nemlig fabrikk, og i alt 4 faktornivåer (fabrikker). Vi får dermed i alt 4 behandlinger, som her er de 4 fabrikkene. Antall replikater er 100 for hver behandling (fabrikk).

I eks. 10.2 er potetene i de 12 åkerlappene forsøksenheterne. Her er det to forsøksfaktorer: åkerlapp og gjødsling. Faktoren åkerlapp har 12 forskjellige nivåer (vi antar i prinsippet at alle 12 åkerlapper kan ha forskjellige egenskaper), mens faktoren gjødsling har 3 forskjellige nivåer (3 forskjellige gjødslingstyper). Vi har totalt 12 behandlinger, eller kombinasjoner av åkerlapp og gjødsling. Måler vi avlingen i form av antall kg pr. settepote, har vi like mange replikater for hver behandling som det er settepoter pr. åkerlapp.

Når vi kun har én forsøksfaktor med  $k$  forskjellige nivåer, og vi utfører de  $k$  forskjellige behandlingene på ulike tilfeldige utvalg av forsøksenheter, får vi en-en-vegs variansanalyse. Logikken bak denne betegnelsen er at vi kun har én faktor å gruppere dataene etter med én gruppe pr. faktornivå. Har vi to faktorer, får vi en to-vegs variansanalyse, fordi vi nå har to faktorer å gruppere dataene etter. Vi skal begrense oss til en-vegs variansanalyse i dette kapitlet. Dersom vi i eks. 10.2 hadde funnet at det ikke var noen indikasjon på forskjeller i egenskapene fra åkerlapp til åkerlapp (homogen åker), kunne vi nøyet oss med å gruppere dataene kun etter gjødslingstype. Vi ville dermed fått en en-vegs variansanalyse der antall replikater for hver behandling (gjødslingstype) nå ville være antall settepoter på 4 åkerlapper.

Bemerk at i begge eksemplene har vi kun én observasjonsvariabel, nemlig varighet (f.eks. timer) i lyspæreeksemplet og avling (f.eks. kg) i poteteksemplet. Vi skal begrense oss til det en-dimensjonale tilfellet i dette kapitlet.

Variansanalyse, eller mer spesifikt analyse av variasjon rundt middelverdier, består i å dele den totale variasjonen til alle våre observasjoner i ulike komponenter. I en-vegs variansanalyse får vi én varianskomponent,  $SST$ , som er knyttet til behandlingene. Betegnelsen  $SST$  er forkortelse for «Sum of Squares due to Treatment» (kvadratsum som skyldes behandling), og vi kaller denne **behandlingskvadratsummen**. I tillegg får vi en varianskomponent,  $SSE$ , som skyldes ukontrollerbar og tilfeldig variasjon. Denne kaller vi **feilkvadratsum**, eller **residualsum**. Betegnelsen  $SSE$  er forkortelse for «Sum og Squares due to Error» (kvadratsum som skyldes feil).

I lyspæreseksemplet vil vi for hver fabrikk få en kvadratsum av de undersøkte enkeltpærenes varighet rundt middelverdien for vedkommende fabrikk, som skyldes tilfeldig og ukontrollerbar variasjon fra pære til pære. I tillegg får vi en variasjon av de ulike fabrikkmiddelverdiene rundt den totale middelverdien av alle målte lyspærer i alle fabrikker. Vår utfordring består i å undersøke om det er signifikante forskjeller på noen av de målte fabrikkmiddelverdiene, og i så fall å påpeke hvilke som skiller seg ut. For å kunne gjøre dette, må vi ha en modell for de tilfeldige variasjonene fra lyspære til lyspære. Vi skal her kun se på den enkleste og vanligste modellen: Avvikene fra forventningsverdien for en gitt behandling er uavhengige  $N(0, \sigma)$ -variabler, der  $\sigma$  er uavhengig av behandling. Jo mindre  $\sigma$ -verdi og jo flere replikater, jo lettere vil det være å påvise en signifikant forskjell på middelverdiene for en gitt forskjell på de sanne fabrikk-forventningsverdiene.

I prinsippet kunne vi sammenlignet to og to behandlinger (fabrikker) slik foreskrevet i kap. 8. Vi skal imidlertid her se hvordan vi kan bruke variansanalyse til å analysere et vilkårlig antall  $k$  behandlinger *samtidig*. En stor fordel med dette er at vi kan bruke observasjonene fra alle behandlingene til mer effektivt å «midle vekk» betydningen av den ukontrollerbare og tilfeldige variasjonen.

Kapitlet er disponert som følger: kap. 10.2 omhandler det generelle oppsettet for sammenligning av  $k$  behandlinger, inkludert en oppsummerende ANOVA-tabell. ANOVA er forkortelse for «ANalysis Of VAriants», eller variansanalyse på norsk. I kap. 10.3 innfører vi en populasjonsmodell og studerer konfidensintervall og hypotesetester. Kapitlet avsluttes med oppgaver i kap. 10.4 og formelsamling i kap. 10.5.

## 10.2 Sammenligning av $k$ behandlinger

Vi har følgende eksperimentelle oppsett: Et sett med  $k$  forskjellige behandlinger skal utføres på totalt  $n$  av  $N$  aktuelle forsøksenheter. For en gitt behandling, består populasjonen av alle mulige utfall (observasjonsverdier) de  $N$  aktuelle forsøksenhetene ville gitt hvis alle var underlagt behandlingen. For at vi skal få tilfeldige utvalg, kan vi tenke oss at vi først trekker ut  $n$  tilfeldige forsøksenheter blant de  $N$  aktuelle. Deretter trekkes  $n_1$  av disse  $n$  tilfeldig og underlegges behandling 1,  $n_2$  enheter trekkes tilfeldig fra de resterende  $n-n_1$  enheter og underlegges behandling 2, osv. Behandling  $k$  utføres på de  $n_k$  enhetene som gjenstår til slutt.

Datastrukturen for målingene er skissert i tabell 10.1 nedenfor, der  $y_{ij}$  er måling nr.  $i$  under behandling  $j$ ,  $i = 1, \dots, n_j$ ,  $j = 1, \dots, k$ . De to siste radene viser summasjonsobservatorene.

*Tabell 10.1 Datastruktur for  $k$  behandlinger basert på tilfeldige utvalg.*

	Behandling 1	Behandling 2	...	Behandling $k$
	$y_{11}$	$y_{12}$	...	$y_{1k}$
	$y_{21}$	$y_{22}$	...	$y_{2k}$
	$\vdots$	$\vdots$	...	$\vdots$
	$y_{n_1 1}$	$y_{n_2 2}$	...	$y_{n_k k}$
Middelverdier	$\bar{y}_1$	$\bar{y}_2$	...	$\bar{y}_k$
Kvadratsummer av avvik fra middel- verdier	$\sum_{i=1}^{n_1} (y_{i1} - \bar{y}_1)^2$	$\sum_{i=1}^{n_2} (y_{i2} - \bar{y}_2)^2$	...	$\sum_{i=1}^{n_k} (y_{ik} - \bar{y}_k)^2$

Vi skal nå forklare tankegangen bak variansanalysen og de tilhørende beregninger ved et numerisk eksempel. I kap. 10.3 skal vi innføre en mer generell populasjonsmodell, som gir en oppskrift for hypotesetester og konstruksjon av konfidensintervall.

**Eks. 10.3** **Elastisitetsmodul.** En stålprodusent ønsker å sammenligne elastisitetsmodulen ved 3 forskjellige ståltyper A, B og C. Standard stålplater av de ulike typene festes i den ene enden. Så måles nedbøyningen i mm ved en konstant belastning i den andre enden. Data for forsøkene er vist i neste tabell.

To sentrale problemstillinger som ønskes belyst er:

- 1) Er det noen signifikant forskjell på de tre ståltypenes elastisitetsmodul?
- 2) Kan vi konstruere konfidensintervaller for den forventede differansen mellom to og to stålkvaliteter?

Tabell 10.2 Nedbøyning i mm for 3 forskjellige stålkvaliteter

	A	B	C	rekkesum
	7	3	4	14
	9	2	6	17
	5	4	3	12
	8		3	11
	6			6
Middelverdier:	$\bar{y}_1 = 7$	$\bar{y}_2 = 3$	$\bar{y}_3 = 4$	$\bar{y} = 60/12 = 5$
Kvadrat-summer:	$\sum_{i=1}^5 (y_{i1} - \bar{y}_1)^2 = 10$	$\sum_{i=1}^3 (y_{i2} - \bar{y}_2)^2 = 2$	$\sum_{i=1}^4 (y_{i3} - \bar{y}_3)^2 = 6$	

Vi tenker oss at vi kan dekomponere variasjonene til enkeltobservasjonene rundt den totale middelverdien,  $y_{ij} - \bar{y}$ , i to komponenter: Én komponent som skyldes forskjellig middelverdi for de tre stålkvalitetene, og én komponent som skyldes tilfeldig variasjon av enkeltprøvene rundt middelverdien for hver stålkvalitet:

$$\begin{array}{lclclcl} y_{ij} & = & \bar{y} & + & (\bar{y}_j - \bar{y}) & + & (y_{ij} - \bar{y}_j) \\ \text{observa-} & = & \text{total} & + & \text{avvik p.g.a.} & + & \text{residual} \\ \text{sjon} & & \text{middel} & & \text{behandling} & & \end{array}$$

Fra dataene gitt i tabell 10.2 kan vi presentere dekomponeringen av alle observasjonene ved følgende matriser:

$$\begin{array}{lclcl} \text{Observasjoner} & & \text{Total middel} & & \text{Behandlingseffekt} & & \text{Residualer} \\ y_{ij} & = & \bar{y} & + & (\bar{y}_j - \bar{y}) & + & (y_{ij} - \bar{y}_j) \end{array}$$

$$\begin{bmatrix} 7 & 3 & 4 \\ 9 & 2 & 6 \\ 5 & 4 & 3 \\ 8 & 3 \\ 6 \end{bmatrix} = \begin{bmatrix} 5 & 5 & 5 \\ 5 & 5 & 5 \\ 5 & 5 & 5 \\ 5 & 5 \\ 5 \end{bmatrix} + \begin{bmatrix} 2 & -2 & -1 \\ 2 & -2 & -1 \\ 2 & -2 & -1 \\ 2 & & -1 \\ 2 \end{bmatrix} + \begin{bmatrix} 0 & 0 & 0 \\ 2 & -1 & 2 \\ -2 & 1 & -1 \\ 1 & & -1 \\ -1 \end{bmatrix}$$

Ser vi for eksempel på øverste element til venstre, får vi at  $7 = 5 + 2 + 0 = 7$  (stemmer!), eller  $y_{11} = \bar{y} + (\bar{y}_1 - \bar{y}) + (y_{11} - \bar{y}_1)$ .

Dersom det ikke er noen forskjell på de tre ståltypene, vil vi forvente at verdiene i «behandlings-matrisen» vil være nær null. Som et summasjonsmål på behandlingsvariasjon, summerer vi kvadratene til alle elementene i matrisen:

$$\begin{aligned} SS_T &= [2^2+2^2+2^2+2^2+2^2] + [(-2)^2+(-2)^2+(-2)^2+(-2)^2] + [(-1)^2+(-1)^2+(-1)^2] \\ &= 5 \cdot 2^2 + 3 \cdot (-2)^2 + 4 \cdot (-1)^2 = 36 \end{aligned}$$

Generaliserer vi ligningen over for behandlings-kvadratsummen,  $SS_T$ , får vi:

$$SS_T = \sum_{j=1}^k n_j (\bar{y}_j - \bar{y})^2$$

Siste matrise i dataoppsettet inneholder residualene  $y_{ij} - \bar{y}_j$ , som er avvikene fra enkeltobservasjonene til behandlingsgjennomsnittet for den behandling observasjonene er underlagt. Disse avvikene betraktes som ukontrollerbare og tilfeldige. Måleusikkerhet er et eksempel på denne type feil. Som summasjonsmål for tilfeldige feil, tar vi summen av alle kvadratene til alle elementene i «residualmatrisen»:

$$SSE = 0^2 + 2^2 + \dots + (-1)^2 + (-1)^2 = 18$$

Generelt får vi altså:

$$SSE = \sum_{j=1}^k \sum_{i=1}^{n_j} (y_{ij} - \bar{y}_j)^2$$

Til slutt ser vi på variasjonen  $(y_{ij} - \bar{y})$  til  $y_{ij}$ -verdiene rundt den totale middelverdien  $\bar{y}$ . Vi summerer som før kvadratet av alle disse avvikene, og får:

$$S_y^2 = (7-5)^2 + (9-5)^2 + \dots + (3-5)^2 + (3-5)^2 = 54$$

Legg merke til at  $S_y^2 = SS_T + SSE$  ( $36 + 18 = 54$ ), som er et generelt resultat.

### **Antall frihetsgrader** (d.f., «degrees of freedom»).

Et annet viktig aspekt med dekomponeringen er antall frihetsgrader som er knyttet til hver kvadratsum. Generelt gjelder at

Antall frihetsgrader forbundet med en kvadratsum

- = Antall elementer som kvadreres og summeres
- antall lineære betingelser som elementene må tilfredsstille

I eks. 10.3 var behandlings-kvadratsummen en sum av 3 kvadratiske ledd:  $SS_T = n_1(y_1 - \bar{y})^2 + n_2(y_2 - \bar{y})^2 + n_3(y_3 - \bar{y})^2$ . Elementene må tilfredsstille én lineær betingelse, nemlig

$$n_1(y_1 - \bar{y}) + n_2(y_2 - \bar{y}) + n_3(y_3 - \bar{y}) = 0$$

som følger av at totalmiddelet  $\bar{y}$  er et vektet gjennomsnitt av behandlings-middelverdiene:

$$\bar{y} = \frac{n_1\bar{y}_1 + n_2\bar{y}_2 + n_3\bar{y}_3}{n_1 + n_2 + n_3}$$

Antall frihetsgrader knyttet til  $SS_T$  er følgelig generelt  $k-1$ , dvs.  $3-1 = 2$  i eksemplet.

Når det gjelder residualsummen, SSE, har vi like mange kvadratledd som skal summeres som totalt antall observasjoner  $n$ , dvs. 12 i eks. 10.3. Vi har ialt  $k$  kvadratsummer, og elementene som kvadreres i disse summene summeres til null:  $\sum_i (y_{ij} - \bar{y}_j) = 0$  for  $j = 1, \dots, k$ . Dette følger direkte av definisjonen på den aritmetiske middelverdien  $\bar{y}_j$ . Vi har følgelig  $k$  lineære betingelser, og får generelt  $n-k$  frihetsgrader forbundet med residualsummen SSE,  $12-3 = 9$  frihetsgrader i eks. 10.3.

Ved tilsvarende resonnement som ovenfor, finner vi at totalkvadratsummen  $S_y^2$  er forbundet med  $n-1$  frihetsgrader, dvs.  $12-1 = 11$  i eks. 10.3. Bemerk at  $d.f.(S_y^2) = d.f.(SS_T) + d.f.(SSE)$ :  $(n-1) = (n-k) + (k-1)$ . ☺

Vi generaliserer resultatene fra eksemplet og får:

### Kvadratsummer og frihetsgrader

Ved en-vegs variansanalyse får vi følgende sammenheng mellom kvadratsummene for totalavvik ( $S_y^2$ ), behandlingsavvik ( $SS_T$ ) og residualer ( $SSE$ ), samt mellom det antall frihetsgrader d.f. disse kvadratsummene er forbundet med:

$$\sum_{j=1}^k \sum_{i=1}^{n_j} (y_{ij} - \bar{y})^2 = \sum_{j=1}^k n_j (\bar{y}_j - \bar{y})^2 + \sum_{j=1}^k \sum_{i=1}^{n_j} (y_{ij} - \bar{y}_j)^2$$

$$S_y^2 = SS_T + SSE$$

$$d.f. = n-1 \quad d.f. = k-1 \quad d.f. = n-k$$

der  $k$  er antall behandlinger og  $n = \sum n_j$  er totalt antall observasjoner.

Det er vanlig å fremstille dekomponeringen av kvadratsummene og antall frihetsgrader i en såkalt ANOVA-tabell. Denne tabellen inneholder normalt også «kvadratmidlene», som er kvadratsummene dividert med tilhørende antall frihetsgrader. Kvadratmidlene betegnes med MS («Mean Square»).

*Tabell 10.3 ANOVA-tabell for sammenligning av  $k$  behandlinger*

Kilde	Kvadratsum	d.f.	Kvadratmiddel
Behandlinger	$SS_T = \sum_{j=1}^k n_j (\bar{y}_j - \bar{y})^2$	$k-1$	$MS_T = \frac{SS_T}{k-1}$
Feil	$SSE = \sum_{j=1}^k \sum_{i=1}^{n_j} (y_{ij} - \bar{y}_j)^2$	$n-k$	$MSE = \frac{SSE}{n-k}$
Total	$S_y^2 = \sum_{j=1}^k \sum_{i=1}^{n_j} (y_{ij} - \bar{y})^2$	$n-1$	

Vi har fremdeles ikke besvart spørsmålene om hvordan vi finner ut om det er signifikante forskjeller på ulike behandlinger, og hvordan vi kan konstruere konfidensintervaller for forventet differanse mellom to og to behandlinger. Dette er tema i kap. 10.3.

### 10.3 Populasjonsmodell og inferens

For å kunne utføre formelle hypotesetester om forskjeller på effekten av ulike behandlinger, samt konstruere konfidensintervaller for forskjellen på to og to behandlinger, må vi ha en populasjonsmodell knyttet til eksperimentet. De aktuelle eksperimentelle enhetene vi tar tilfeldige utvalg fra, vil ofte være felles for de ulike behandlingene, mens populasjonen vil være knyttet til behandlingen. Med populasjon forstår vi her samlingen av alle mulige utfall (observasjoner) vi kan få for en gitt behandling dersom alle aktuelle forsøksenheter var underlagt behandlingen. Vi har derfor like mange populasjoner,  $k$ , som vi har behandlinger.

For en gitt behandling,  $j$ , skal vi anta at observasjonene  $Y_{ij}$  utgjør et tilfeldig utvalg fra en normal populasjonsfordeling med forventning  $\mu_j$ , som kan variere fra behandling til behandling, og en felles varians  $\sigma^2$ . Videre antar vi at alle observasjonene er uavhengige. Følgende nye parametre innføres:

$$\mu = \frac{1}{k} \sum_{j=1}^k \mu_j \quad (\text{gjennomsnitt av populasjonsforventningene } \mu_j)$$

$$\beta_j = \mu_j - \mu \quad (\text{effekten av behandling } j), \quad j = 1, \dots, k$$

Med de nye parametrerne kan modellen formuleres som følger:

#### Populasjonsmodell for sammenligning av $k$ behandlinger

$$Y_{ij} = \mu + \beta_j + E_{ij}, \quad i = 1, \dots, n_j, \quad j = 1, \dots, k$$

der  $\mu$  er total forventning,  $\beta_j$  er effekten av behandling  $j$ ,  $\sum_{j=1}^k \beta_j = 0$

og  $E_{ij}$ 'ene er uif  $N(0, \sigma)$ -variabler.

Nullhypotesen om ingen forventet forskjell mellom ulike behandlinger, kan nå formuleres som følger:

$$H_0: \beta_1 = \beta_2 = \dots = \beta_k = 0$$

mot alternativhypotesen

$$H_1: \beta_i \neq 0 \text{ for minst én } i\text{-verdi}$$

For å utføre en test, må vi finne en fornuftig testobservator med kjent fordeling under  $H_0$ . Dersom ingen behandlinger gir noe effekt, forventer vi at midlere behandlings-kvadratsum  $MST = SS_T/(k-1)$  skal bli liten. Midlere residual-kvadratsum,  $MSE = SSE/(n-k)$ , er en forventningsrett estimator for  $\sigma^2$ :

$$E(MSE) = E\left(\frac{SSE}{n-k}\right) = \sigma^2$$

Jo større forholdet  $MST/MSE$  er, jo mer taler dette for at  $H_0$  ikke er sann. Det kan vises at  $SS_T$  og  $SSE$  under  $H_0$  er uavhengige og kjikvadratfordelte med henholdsvis  $k-1$  og  $n-k$  frihetsgrader.  $MST/MSE$  blir dermed  $F$ -fordelt med henholdsvis  $k-1$  og  $n-k$  frihetsgrader, og vi får følgende hypotesetestoppsett:

### Hypotesetest for $k$ behandlinger

$$H_0: \beta_1 = \beta_2 = \dots = \beta_k = 0 \quad \text{mot} \quad H_1: \beta_i \neq 0 \text{ for minst én } i$$

$$\text{Testobservator: } T = \frac{SS_T / (k-1)}{SSE / (n-k)} = \frac{MST}{MSE},$$

$$\text{Forkastingsområde: } R: T > F_{\alpha}(k-1, n-k)$$

der  $n$  er totalt antall observasjoner og  $F_{\alpha}(k-1, n-k)$  er øvre  $\alpha$ -fraktil i  $F$ -fordelingen med antall frihetsgrader d.f. =  $(k-1, n-k)$ .

### Eks. 10.4 ANOVA-tabell og hypotesetest

#### Oppgave

- Konstruer ANOVA-tabellen for dataene i eks. 10.3 (stålsammensetning)
- Test  $H_0$ : Ingen behandlingseffekt, på nivå 5%

#### Løsningsforslag

- Ved å bruke tidligere beregninger i eks. 10.3 kan vi sette opp følgende ANOVA-tabell:

Tabell 10.4 ANOVA-tabell for dataene i eks. 10.3

Kilde:	kvadratsum	# frihetsgrader	kvadratmiddel	F-forhold
Behandling	$SS_T = 36$	$k-1 = 2$	$MS_T = 18$	$18/2 = 9$
Feil	$SSE = 18$	$n-k = 9$	$MSE = 2$	
Total	54	11		

b) Fra tabellen bak i boka finner vi at  $F_{0.05}(2,9) = 4.256$ . Siden  $T = 9 > 4.256$ , forkaster vi  $H_0$  og konkluderer på nivå 5 % med at minst én av de 3 stålsammensetningene påvirker elastisitetsmodulen. ☺

### Konfidensintervall for parvis sammenligning

ANOVA  $F$ -testen vi har sett på, er en metode for å undersøke om data gir grunnlag for å konkludere at det er forskjeller på de  $k$  behandlingene, med andre ord om minst én av behandlingene har noen effekt. Dersom vi forkaster  $H_0$  (ingen behandlingseffekt), bør vi gå mer spesifikt til verks med å analysere likheter og forskjeller på de ulike behandlingene. Vi skal her begrense oss til å se hvordan vi kan konstruere konfidensintervaller (KI) for forskjeller på to og to behandlinger,  $\beta_i - \beta_j$ .

Bemerk at  $\beta_j = \mu_j - \mu$ , slik at  $\beta_i - \beta_j = \mu_i - \mu_j$ . Under våre forutsetninger vil  $\bar{Y}_i - \bar{Y}_j$  være normalfordelt med forventning  $\mu_i - \mu_j = \beta_i - \beta_j$  og varians

$$\text{Var}(\bar{Y}_i - \bar{Y}_j) = \sigma^2 \left( \frac{1}{n_i} + \frac{1}{n_j} \right)$$

Videre vil  $SSE/\sigma^2$  være  $\chi^2$ -fordelt med d.f. =  $n-k$ , og uavhengig av  $\bar{Y}_i - \bar{Y}_j$ . Derfor vil forholdet

$$t = \frac{(\bar{Y}_i - \bar{Y}_j) - (\beta_i - \beta_j)}{\sqrt{\frac{SSE}{n-k}} \sqrt{\frac{1}{n_i} + \frac{1}{n_j}}}$$

være  $t$ -fordelt med  $n-k$  frihetsgrader. Vi får dermed følgende oppsett for konstruksjon av konfidensintervall for  $\beta_i - \beta_j$ :

### Konfidensintervall for behandlingsforskjell

Et  $100(1-\alpha)\%$  KI for forventet forskjell ( $\beta_i - \beta_j$ ) på effekten av to behandlinger er gitt ved

$$\bar{Y}_i - \bar{Y}_j \mp t_{\alpha/2} \cdot S \sqrt{\frac{1}{n_i} + \frac{1}{n_j}}$$

der  $S^2 = \text{MSE} = \text{SSE}/(n-k)$  og  $t_{\alpha/2}$  er øvre  $\alpha/2$  fraktil i  $t$ -fordelingen med  $n-k$  frihetsgrader.

**Eks. 10.5** Konstruksjon av **konfidensintervaller** for forventet forskjell på nedbøyning av stålplater med forskjellig sammensetning.

*Oppgave*

- Konstruer et 95 % KI for to og to av behandlings-middelverdiene i eks. 10.3.
- Er det signifikante forskjeller på noen av middelverdiene på nivå 5 % ?

*Løsningsforslag*

- Vi har  $k = 3$  forskjellige behandlinger (ståltyper) som gir opphav til  $(\bar{y}_1, \bar{y}_2, \bar{y}_3) = (7, 3, 4)$ . Fra eks. 10.3 og eks. 10.4 har vi følgende:

$$n_1 = 5, n_2 = 3, n_3 = 4, \bar{y}_1 = 7, \bar{y}_2 = 3, \bar{y}_3 = 4, \text{SSE} = 18, n-k = 9$$

Vi følger oppsettet i forrige ramme. Fra  $t$ -tabellen bak i boka finner vi at  $t_{0.05} = 2.262$  (9 frihetsgrader). Videre er  $s^2 = \text{SSE}/(n-k) = 2$  og  $s = 2^{1/2}$ . Vi får da:

$$(\bar{y}_1 - \bar{y}_2) \mp = t_{\alpha/2} \cdot s \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} = (7 - 3) \mp 2.262 \cdot \sqrt{2} \cdot \sqrt{\frac{1}{5} + \frac{1}{3}}$$

som utregnet gir at

$$(1.66, 6.34) \text{ er et 95 \% KI for } \beta_1 - \beta_2$$

Ved tilsvarende regning finner vi at

$$(0.85, 5.15) \text{ er et 95 \% KI for } \beta_1 - \beta_3$$

$$(-3.44, 1.44) \text{ er et 95 \% KI for } \beta_2 - \beta_3$$

- b) Det er signifikante forskjeller på to middelverdier på signifikansnivå 5 % dersom 95 %-intervallet ikke dekker null. Dette ser vi er tilfellet for de to første intervallene. Vi kan dermed konkludere at midlere nedbøyning for stålsammensetning A er signifikant større enn midlere nedbøyning for stålsammensetning B og C, siden tallene i intervallet er positive. ☺

### Simultane konfidensintervall

Når vi har  $k$  behandlinger vil vi generelt få totalt  $(^k, 2)$  kombinasjoner av to forskjellige behandlinger. For rimelig store  $k$  gir dette opphav til mange konfidensintervaller for differanser mellom to og to behandlinger. Det er ønskelig, men vanskelig å bestemme et konfidensnivå som gjelder for alle konfidensintervaller under ett, fordi vi har kompliserte avhengighetsforhold mellom konfidensintervallene. For et gitt simultant konfidensnivå, vil generelt bredden på hvert konfidensintervall øke når  $k$  øker.

Vi skal ikke gå videre inn på hvordan vi kan konstruere simultane konfidensintervall, men nøye oss med å påpeke at vi bør være forsiktige med å dra for sterke konklusjoner, selv om signifikante forskjeller mellom to behandlingsmiddelverdier påvises. Generelt kan det være fornuftig å begrense antall parvis sammenligninger til de som er mest interessante for undersøkelsen, og bruke resultatene til eventuelt å bestemme hvilke eksperimenter det er fornuftig å videreføre for å oppnå sikrere konklusjoner.

### Modellsjekk

Analogt med lineær regresjon, er det fornuftig å analysere residualene for å se om forutsetningen om at disse er uif  $N(0, \sigma)$ -variabler ser ut til å stemme. Vi kan for eksempel lage  $k$  horisontale akser under hverandre, en for hver behandling, og så plotte residualene for hver behandling som punkter langs hver av aksene. Vi vil da få et inntrykk av om residualene har like stor spredning for hver behandling, og om de er rimelig symmetrisk fordelt om middelverdien. Heldigvis er variansanalysen rimelig robust mot små eller moderate avvik fra forutsetningen om normalitet og konstant varians.

La oss til slutt vise hvordan vi kunne løst noen av oppgavene i eksemplene vi har sett på ved hjelp av Minitab. Merk at konfidensintervallene gjelder for hver enkelt behandling, og ikke for parvis sammenligning mellom to og to behandlinger. Det er imidlertid fullt mulig å foreta parvis sammenligning ved hjelp av Minitab.

```
MTB > aovoneway 'A' 'B' 'C';      # Dataene ligger i kolonnene kalt A, B og C.  
SUBC> gdotplot.                 # Lager punktplot av dataene
```

## One-Way Analysis of Variance

Analysis of Variance					
Source	DF	SS	MS	F	P
Factor	2	36.00	18.00	9.00	0.007
Error	9	18.00	2.00		
Total	11	54.00			

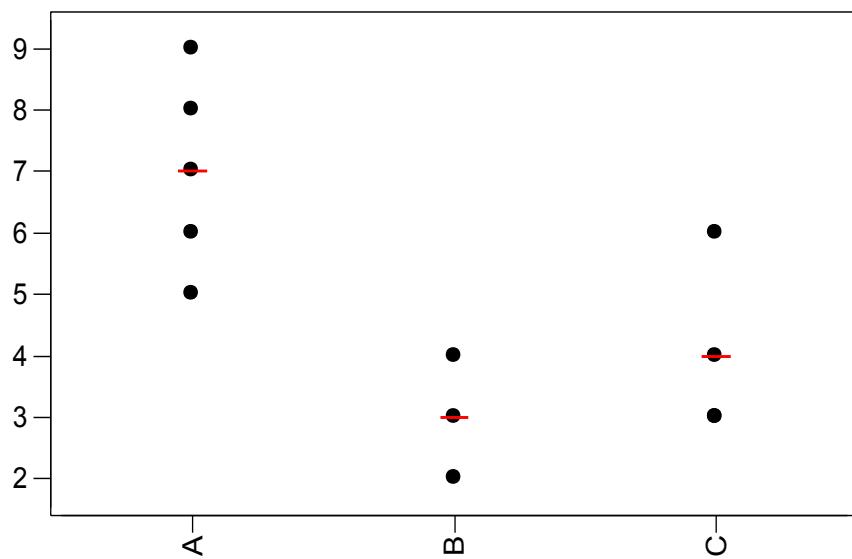
Individual 95% CIs For Mean Based on Pooled StDev					
Level	N	Mean	StDev	-----+-----+-----+-----+-----+-----	(-----*-----)
A	5	7.000	1.581		
B	3	3.000	1.000	(-----*-----)	
C	4	4.000	1.414	(-----*-----)	

Pooled StDev =	1.414	2.0	4.0	6.0	8.0
----------------	-------	-----	-----	-----	-----

## Dotplots of A - C

(group means are indicated by lines)



**Kommentar:** Fra analysen ser vi at signifikant sannsynlighet er  $P^* = 0.007$ , som gir kraftig støtte til konklusjonen om behandlingsforskjell (se eks. 10.4). Figuren over viser punktplott av dataene med middelverdien markert som en horisontal strek. Residualene kan leses som vertikalt avvik fra enkeltpunktene til middelverdiene. Minitab har forøvrig forskjellige oppsjoner for å plotte residualer som det vil føre for langt å komme inn på her.

## 10.4 Oppgaver

**10.1** En fabrikant av stålplater vil analysere stålstyrken til 3 forskjellige stålkvaliteter A, B og C, ett parti av hver type. Styrken måles i form av stålets flytegrense. Henholdsvis 100, 70 og 30 tilfeldige prøver tas fra A-, B- og C-partiet.

- a) Hva er forsøksenhetsene her?
- b) Hva er forsøksfaktorene i forsøket?
- c) Hva er faktornivåene?
- d) Hvilke behandlinger har vi?
- e) Hvilke populasjoner har vi?
- f) Hvor mange replikater er det?
- g) Hva er responsvariabelen?

**10.2** Gitt  $y_{ij}$ -dataene

	A	B	C
10	21	11	
14	22	16	
14	15	12	
14	18	20	
15		13	

- a) Beregn  $\bar{y}_1$ ,  $\bar{y}_2$ ,  $\bar{y}_3$  og  $\bar{y}$
- b) Beregn residualene og sjekk at de summerer til null.
- c) Beregn de nødvendige størrelser og sett opp ANOVA-tabell.
- d) Test om det er signifikante forskjeller mellom noen av de tre behandlings-middelverdiene på nivå 5 %.
- e) Dersom testen i d) gir forkasting, bestem hvilke behandlinger som er signifikant forskjellige basert på 95 % konfidensintervall for forskjell på to og to behandlinger.

**10.3** Gitt følgende svært forenklede ANOVA-tabell:

$$SST = 20 \quad k-1 = 4$$

$$SSE = 100 \quad n-k = 40$$

- a) Hvor mange behandlinger  $k$  og observasjoner  $n$  har vi?
- b) Hvilken  $F$ -fordeling er aktuell her?
- c) Vis at det på signifikansnivå 5 % ikke er grunnlag for å hevde at det er forskjellig effekt av behandlingene.
- d) Betyr konklusjonen i c) at det ikke er forskjellig effekt av behandlingene?

**10.4** Ytelsen til 3 forskjellige typer PC'er med samme prosessortype (Pentium 200 Pro), skal sammenlignes ved å måle CPU-tiden på en bestemt regneoperasjon. Tilfeldige PC'er fra hver av de 3 leverandørene plukkes ut og CPU-tiden måles. Resultatet i målte sekunder er gitt nedenfor:

- $\bar{y}_1 = 12.11$ ,  $\bar{y}_2 = 11.87$ ,  $\bar{y}_3 = 12.02$   
 $n_1 = 5$ ,  $n_2 = 4$ ,  $n_3 = 9$ ,  $SSE = 0.0058$
- a) Beregn de nødvendige størrelser og lag ANOVA-tabell.
  - b) Test på signifikansnivå 5 % om det er grunnlag for å konkludere at det er forskjell på noen av de tre PC-typene. Finn eventuelt hvilken PC som du kan konkludere er «signifikant best».
  - c) Ville ytelsen være avgjørende for hvilken av de 3 PC-typene du ville velge?

**10.5** 4 biltyper skal sammenlignes for å se om det er signifikante forskjeller i bensinforbruk. Tilfeldige biler plukkes

ut, og de kjører nøyaktig den samme distansen i samme hastighet. Resultatene i form av liter pr. mil er:

A	B	C	D
0.436	0.591	0.480	0.406
0.421	0.585	0.529	0.521
0.364	0.645	0.528	0.545
0.301	0.590	0.530	0.537
0.407		0.514	0.529
0.359			0.502

- a) Beregn de nødvendige størrelser og sett opp ANOVA-tabell.
- b) Test om det er signifikant forskjell på midlere bensinforbruk for de forskjellige biltypene på nivå 5 %.
- c) Tidligere tester har gitt sterke indikasjoner på at A bruker minst bensin av de fire. Gir dataene grunnlag for å konkludere med dette på nivå 5 %?

**10.6** En lege (Ola) påstår at det ikke er noen påviselig forskjell mellom 3 ulike sovepiller A, B og C. En annen lege (Kari) påstår imidlertid hårdnakket at type A er best. Et tilfeldig utvalg av villige pasienter med søvnproblemer får i løpet av tre ulike perioder tildelt de 3 forskjellige sovepillene. De vet ikke hvilket middel de får fra gang til gang, og rekkefølgen trekkes vilkårlig. Gjennomsnittlig antall søvntimer pr. natt blir registrert, og gir følgende resultat (subskript 1,2 og 3 angir henholdsvis A, B og C):

$$\bar{y}_1 = 7.34, \bar{y}_2 = 6.92, \bar{y}_3 = 7.21$$

$$n_1 = 41, n_2 = 28, n_3 = 34, SSE = 10.1$$

- a) Hvordan ville du analysert dataene for å verifisere ditt syn dersom du var Ola? Og dersom du var Kari?
- b) Gir data sterkt grunnlag for å stole mer på den ene av de to legene? I såfall hvem og hvorfor?

## 10.5 Formelsamling

### Observasjonsbetegnelser

- $Y_{ij}$  = observasjon på forsøksenhett  $i$  med behandling  $j$   
 $i = 1, \dots, n_j$  = indeks for forsøksenhett  
 $n_j$  = antall forsøksenheter underlagt behandling  $j$   
 $j = 1, \dots, k$  = indeks for behandling  
 $k$  = antall behandlinger  
 $n = n_1 + \dots + n_k$  = totalt antall observasjoner

### Modell

- $Y_{ij} = \mu + \beta_j + E_{ij}$   
 $\beta_j = \mu_j - \mu$  = effekt av behandling  $j$   
 $\mu = E(n_1 Y_{i_1} + \dots + n_k Y_{i_k})/n$  = totalforventn.  
 $E_{ij}$  er  $N(\mu_j, \sigma)$ -fordelt  
Alle  $E_{ij}$  uavhengige

### Summasjonsobservatorer

$$\bar{Y}_j = \frac{1}{n_j} \sum_i^{n_j} Y_{ij}$$

$$\bar{Y} = \frac{1}{n} (n_1 \bar{Y}_1 + \dots + n_k \bar{Y}_k)$$

Behandlingskvadratsum:

$$SS_T = \sum_{j=1}^k n_j (\bar{Y}_j - \bar{Y})^2$$

Feilkvadratsum (residualkvadratsum):

$$SSE = \sum_{j=1}^k \sum_{i=1}^{n_j} (Y_{ij} - \bar{Y}_j)^2$$

Total kvadratsum:

$$S_y^2 = \sum_{j=1}^k \sum_{i=1}^{n_j} (Y_{ij} - \bar{Y})^2 = SS_T + SSE$$

Midlere behandlingskvadratsum:

$$MS_T = SS_T/(k-1)$$

Midlere feilkvadratsum:

$$S^2 = MSE = SSE/(n-k), ES^2 = \sigma^2$$

### ANOVA-tabell for sammenligning av $k$ behandlinger:

Kvadratsum	d.f.	Kvadratmiddel	F-forhold
SS <sub>T</sub>	$k-1$	$MS_T = \frac{SS_T}{k-1}$	$MS_T/MSE$
SSE	$n-k$	$MSE = \frac{SSE}{n-k}$	
$S_y^2$			$n-1$

der d.f. = antall frihetsgrader

### F-fordeling

$F_{u,v}$  er  $F$ -fordelt med henholdsvis  $u = k-1$  og  $v = n-k$  frihetsgrader.  $\Rightarrow$

$F_{v,u} = 1/F_{u,v}$  er  $F$ -fordelt med d.f. =  $(v, u)$ .

### Hypotesetest

$H_0: \beta_1 = \dots = \beta_k = 0$  mot  $H_1: \beta_i \neq 0$  for minst én  $i$ .

Testobservator:  $T = SS_T/MSE$

$T$  er  $F(k-1, n-k)$ -fordelt under  $H_0$ .

Forkastingsområde:  $R: T > F_{\alpha}(k-1, n-k)$   
der  $F_{\alpha}(k-1, n-k)$  er øvre  $\alpha$ -fraktil i  $F$ -fordelingen med henholdsvis  $k-1$  og  $n-k$  frihetsgrader.

### Konfidensintervall

100(1- $\alpha$ ) % KI for  $\beta_i - \beta_j$ :

$$\bar{Y}_i - \bar{Y}_j \mp t_{\alpha/2} \cdot S \sqrt{\frac{1}{n_i} + \frac{1}{n_j}}$$

der  $t_{\alpha/2}$  er øvre  $\alpha/2$ -fraktil i  $t$ -fordelingen med d.f. =  $n-k$  og  $S^2 = MSE$ .

## Kapittel 11

# Monte Carlo-simulering

### 11.1 Innledning

Ordet *simulere* kommer fra latin, og kan direkte oversettes med å *etterligne*. Hva har så dette med Monte Carlo å gjøre? I denne sammenhengen forbinder vi Monte Carlo med spill, og spill har med tilfeldighet å gjøre. Vi kan litt forenklet si at Monte Carlo-simulering hovedsakelig baserer seg på numerisk generering av tilfeldige tall i en datamaskin for å etterligne og undersøke stokastiske (statistiske) prosesser og modeller. La oss først belyse dette med noen eksempler på konkrete anvendelser:

#### Lotto: tilfeldig rekke og trekning av vinnerrekke.

Vi lar datamaskinen generere uavhengige og tilfeldige tall mellom 1 og 34 inntil vi har fått 7 forskjellige tall. Den stokastiske prosessen vi etterligner i dette tilfellet er en tilfeldig utfylling av en Lotto-rekke. Dersom vi genererer 9 forskjellige tall på samme måte, lar de 7 første være vinnertall og de 2 siste være tilleggstall, har vi simulert den karakteristiske trekningen med kuler som foregår på Hamar.

#### Sannsynlighetsfordelingen til sammensatt variabel.

Som et eksempel studerer vi den stokastiske variablen  $Z = X_1 \cdot \ln(X_2) \cdot X_3^{1/2}$ , der  $X_1$ ,  $X_2$  og  $X_3$  er uavhengige stokastiske variabler. Videre antar vi at  $X_1$  og  $X_3$  er normalfordelte med kjente forventninger og varianser, mens  $X_2$  antas eksponentiellfordelt med kjent parameter  $b$ . Å bestemme sannsynlighetsfordelingen til  $Z$  rent teoretisk er komplisert, mens det som vi skal se er enkelt å generere tilfeldige variabler  $X_1$ ,  $X_2$  og  $X_3$  fra fordelingene til disse variablene, ved hjelp av Monte Carlo-simulering. For hver ny simulering av  $X_1$ ,  $X_2$  og  $X_3$  beregner vi  $Z$  og gjentar dette til vi f.eks. har simulert 1000 tilfeldige  $Z$ -verdier. Deretter kan vi f.eks. beregne middelverdi og standardavvik til  $Z$ -verdiene, samt fremstille  $Z$ -verdiene i et histogram for å få et bilde av formen på fordelingen.

#### Verifisering av sentralgrenseteoremet.

Fra kap. 5 husker vi forhåpentligvis sentralgrenseteoremet, som sa at en middelverdi av  $n$  uavhengige stokastiske variabler var tilnærmet normalfordelt uavhengig av populasjonsfordelingen(e) til de enkelte variablene vi tok middelverdien av. Betingelsen var at  $n$  var tilstrekkelig stor, og vi brukte som tommelfingerregel at  $n \geq 20$ . I mange tilfeller er dette en altfor streng grense, mens en også kan ha tilfeller der  $n$  bør være større enn 20 for at normaltilnærmelsen skal

være rimelig. Monte Carlo-simulering er et ypperlig verktøy for å verifisere hvorvidt middelverdien virkelig blir normalfordelt. La  $Y$  betegne middelverdien  $\bar{X}$  til de uavhengige variablene  $X_1, \dots, X_n$ . Vi genererer først et sett med verdier for  $X_1, \dots, X_n$  fra deres populasjonsfordeling(er). Da har vi fått generert en tilfeldig  $Y$ -verdi. Vi genererer så nye  $X$ -verdier og beregner nye  $Y$ -verdier til vi har et ønsket antall  $m$  tilfeldige  $Y$ -verdier. Det eksisterer da en rekke gode tester for å undersøke om det er rimelig å anta at  $Y$ -verdiene er normalfordelte.

Så godt som all Monte Carlo-simulering baserer seg på generering av tall i en datamaskin som idéelt sett skal være uniformt fordelte. Ofte er disse skalert til tall på intervallet fra 0 til 1, dvs. de er idéelt sett realiseringer av uavhengige  $U[0,1]$ -variabler. Selv enkle lommekalkulatorer har en slik slumptallsgenerator. Tallene som genereres er i prinsippet deterministiske, som vi skal se. De kalles derfor for *pseudorandom tall*, som på «norsk» noe fritt kan oversettes til «kvasi-tilfeldige tall».

Begrepene (sannsynlighets-) *tetthetsfunksjon*, *kumulativ (sannsynlighets-) fordelingsfunksjon* og *invers kumulativ (sannsynlighets-) fordelingsfunksjon* er helt sentrale begreper i dette kapitlet. Vi skal unnlate forstavelsen «sannsynlighets» i disse begrepene for å få litt lettere språk. Vi minner om at en kumulativ fordelingsfunksjon  $F(x)$  til en stokastisk variabel  $X$  er definert som  $F(x) = P(X \leq x)$  både for diskrete og kontinuerlige variabler. Tetthetsfunksjonen  $f(x)$  for en diskret variabel er definert som  $f(x) = P(X = x)$ , mens  $f(x)$  for en kontinuerlig variabel kan finnes ved å derivere den kumulative fordelingsfunksjonen:  $f(x) = dF(x)/dx$  for en kontinuerlig variabel  $X$ , forutsatt at den deriverte eksisterer.

Invers kumulativ fordeling har vi ikke vært borti før. Vi skal illustrere begrepet med en enkel eksponensialfordeling med parameter  $b = 1$ .

### Invers kumulativ fordelingsfunksjon $F^{-1}$

La  $T$  være eksponensialfordelt med parameter  $b = 1$ . Kumulativ fordelingsfunksjon  $F(t)$  er da gitt ved uttrykket

$$F(t) = 1 - e^{-t}$$

$F(t)$  er som vi ser en funksjon av  $t$ . Dersom vi finner  $t$  uttrykt som en funksjon av  $F$  har vi funnet den inverse kumulative fordelingsfunksjonen  $t = F^{-1}$ :

$$F(t) = 1 - e^{-t} \Rightarrow t = -\ln(1-F)$$

$F^{-1}$  er som vi skal se en svært sentral funksjon ved generering av tilfeldige variabler fra en gitt fordeling.

Disponeringen av kapitlet er som følger: I kap. 11.2 ser vi på hvordan vi kan generere tilfeldige variabler fra en hvilken som helst fordelingsfunksjon, når den inverse kumulative fordelingsfunksjonen  $F^{-1}$  er kjent. Kap. 11.3 omhandler slumptallgenerering, og angir blant annet en vel anerkjent slumptallalgoritme. I kap. 11.4 behandles ulike metoder for å generere tilfeldige normalfordelte variabler. Kap. 11.5 beskriver den binormale fordelingen og omhandler generering av tilfeldige variabelpar fra denne fordelingen (fordeling av variabelpar der hver av variablene er normalfordelte og samvariasjonen mellom dem er entydig gitt av korrelasjonskoeffisienten  $\rho$ ). Videre behandler kap. 11.6 generering av sammensatte variabler, og kap. 11.7 den fascinerende og praktisk nyttige bootstrap teknikken. Til slutt er oppgaver gitt i kap. 11.7 og en formelsamling i kap. 11.8.

## 11.2 Generering av variabler fra $F^{-1}$

### Kontinuerlige stokastiske variabler

La  $F(x) = P(X \leq x)$  betegne en vilkårlig, kontinuerlig og monoton voksende kumulativ fordelingsfunksjon. Videre, la  $R$  betegne en kontinuerlig stokastisk variabel som er uniformt fordelt  $U[0,1]$ . Vi antar at vi kjenner den inverse funksjonen  $F^{-1}$ . Vi kan da generere  $n$  uavhengige variabler  $X_1, \dots, X_n$  fra fordelingen  $F(x)$  ved følgende algoritme:

#### Generering av kontinuerlige stokastiske variabler fra $F^{-1}$

- 1) Generer  $n$  uavhengige  $U[0,1]$ -variabler  $R_1, \dots, R_n$ .
- 2) Beregn  $X_i = F^{-1}(R_i)$ ,  $i = 1, \dots, n$ , der  $F^{-1}$  er invers kumulativ fordelingsfunksjon til  $X$ -ene.
- 3)  $X_1, \dots, X_n$  er da  $n$  uavhengige stokastiske variabler fra fordelingen  $F(x) = P(X \leq x)$

Et kjapt bevis for at oppskriften i ramma virker, kan gå som følger:

$$P(X_i \leq x) = P(F^{-1}(R_i) \leq x) = P(F(F^{-1}(R_i)) \leq F(x)) = P(R_i \leq F(x)) = F(x)$$

Selv enkle kalkulatorer har innebygde rutiner for å generere uavhengige pseudorandom  $U[0, 1]$ -variabler. Algoritmen ovenfor er derfor meget nyttig og svært utbredt, selv om den krever at  $F^{-1}$  er kjent. I andre tilfeller kan det hende

at vi kjenner gode tilnærmelser til  $F^{-1}$ , se f.eks. avsn. 12.4 om generering av normalfordelte variabler.

La oss belyse bruken av det meget nyttige resultatet i ramma med et enkelt eksempel:

**Eks. 11.1 Tilfeldige eksponensialfordelte variabler fra  $\text{Exp}(0,5)$**

*Oppgave*

Generere tilfeldige og uavhengige eksponensialfordelte variabler  $T_1, \dots, T_n$  med tethetsfunksjon  $f(t) = 2e^{-2t}$ ,  $t \geq 0$ .

*Løsningsforslag*

Fordelingen har parameter  $a = 1/2$ , og kumulativ fordelingsfunksjon blir

$$F(t) = P(T \leq t) = 1 - e^{-2t}$$

Vi finner invers fordelingsfunksjon  $t = F^{-1}$  som følger:

$$\begin{aligned} e^{-2t} &= 1 - F \Rightarrow -2t = \ln(1-F) \Rightarrow \\ t &= -\frac{1}{2} \cdot \ln(1-F) \end{aligned}$$

For å generere en tilfeldig variabel  $T_i$ , skal vi nå etter oppskriften først generere en tilfeldig  $U[0,1]$ -variabel,  $R_i$ , og så sette denne inn for  $F$  i  $t$ -uttrykket ovenfor. Her kan vi imidlertid være enda kjappere, fordi  $1-F$  også er  $U[0,1]$ , slik at vi kan erstatte  $1-F$  med  $R_i$ . Løsningen på problemet blir derfor:

$$T_i = -\frac{1}{2} \cdot \ln(R_i), \quad i = 1, \dots, n$$

der  $R_1, \dots, R_n$  er uif  $U[0,1]$ -variabler. ☺

## Diskrete stokastiske variabler

La  $X$  være en diskret stokastisk variabel, dvs.  $X$  kan bare inneha visse diskrete verdier (betegelsen  $x_{(1)}$  betyr den minste  $x$ -verdien,  $x_{(2)}$  den nest minste osv.):

$$X \in \{x_{(1)} < x_{(2)} < \dots < x_{(k)}\} = D$$

der  $D$  betegner definisjonsmengden til sannsynlighetsfordelingen til  $X$ . La videre  $f_{(i)}$  og  $F_{(i)}$  betegne henholdsvis tethetsfunksjon og kumulativ fordelingsfunksjon:

$$\begin{aligned} F_{(i)} &= F(x_{(i)}) = P(X \leq x_{(i)}) \in \{F(x_{(1)}) < F(x_{(2)}) < \dots < F(x_{(k)})\} = V \\ f_{(i)} &= f(x_{(i)}) = P(X = x_{(i)}) = F(x_{(i)}) - F(x_{(i-1)}) \end{aligned}$$

der  $V$  betegner verdimengden til den kumulative fordelingsfunksjonen til  $X$ . Det blir her på den ene siden litt mer kronglete å generere tilfeldige variabler enn i det kontinuerlige tilfellet. På den andre siden er det tilstrekkelig å kjenne kumulativ fordelingsfunksjon  $F_{(i)}$ , vi behøver ikke noe uttrykk for  $F^{-1}$ . I en viss analogi med tilfellet med generering av tilfeldige kontinuerlige variabler, får vi følgende oppsett:

### Generering av tilfeldige diskrete variabler

- 1) La  $X \in \{x_{(1)} < \dots < x_{(k)}\}$  være en stokastisk variabel med kumulativ fordelingsfunksjon  $F_{(i)} = P(X \leq x_{(i)})$ ,  $i = 1, \dots, k$
- 2) La  $R_1, \dots, R_n$  betegne  $n$  uavhengige  $U[0,1]$ -variabler.
- 3) La  $F_{(i),j}$  betegne den minste kumulative verdi  $F_{(i)}$  som er større enn  $R_j$ , og la  $X_j = x_{(i)}$ ,  $j = 1, \dots, n$ .  $X_1, \dots, X_n$  vil da være  $n$  uavhengige variabler fra fordelingen til  $X$ .

Kjapt bevis for resultatet i ramma:

$$P(X_j = x_{(i)}) = P(F_{(i-1),j} < R_j < F_{(i),j}) = F(x_{(i)}) - F(x_{(i-1)}) = f(x_{(i)})$$

som gjelder for  $i = 1, 2, \dots, k$ .

La oss forlate beviset og gå løs på et eksempel:

#### Eks. 11.2 Tilstfeldige binomiske variabler fra $Bino(10, 0,5)$

##### Oppgave

Generer 3 tilfeldige variabler fra binomisk fordeling med parametre  $n = 10$  og  $p = 0.5$ .

##### Løsningsforslag

Vi genererer først 3 tilfeldige slumptall mellom 0 og 1 og får for eksempel:

$$R_1 = 0.728, R_2 = 0.890 \text{ og } R_3 = 0.129.$$

Fra binomisk tabell finner vi følgende:

$$F(5) = 0.623 \text{ og } F(6) = 0.828.$$

Følgelig er  $F(6)$  minste  $F$ -verdi større enn  $R_1 = 0.728$ , slik at  $X_1 = 6$ . Tilsvarende finner vi at  $X_2 = 7$  og  $X_3 = 3$ . ☺

### 11.3 Slumptallgenerering

Vi så i forrige avsnitt at en slumptallgenerator som kan generere tilfeldige tall mellom 0 og 1, er et sentralt verktøy når vi skal generere uavhengige variabler fra en kjent fordeling. Det finnes forskjellige måter å generere slike slumptall på. Mest utbredt i praksis er såkalte rekursive generatorer, der en og samme transformasjon anvendes suksessivt på resultatet av forrige transformasjon.

Vi skal begrense oss til å se litt på en såkalt multiplikativ kongruensgenerator, også kalt Lehmer's metode, som er på form

$$x_{n+1} = k \cdot x_n \pmod{m}$$

der  $x$ -ene og  $k$  er positive heltall. Mod  $m$  (mod for modulus) betyr at vi på høyre side dividerer produktet  $k \cdot x_n$  med  $m$  og så lar  $x_{n+1}$  være heltallsresten etter denne divisjonen. Dette henger naturlig nok sammen med at vi har et begrenset antall siffer til rådighet i en datamaskin, avgrenset av ordlengden i datamaskinen (et typisk eksempel er  $m = 2^{32}$  i en PC). Skulle dette være diffust, så kanskje neste eksempel skulle hjelpe litt på forståelsen:

**Eks. 11.3 Multiplikativ kongruensgenerator**

*Oppgave*

Bestem de 8 første tallene fra algoritmen

$$x_{n+1} = 3 \cdot x_n \pmod{10} \text{ og startverdi } x_0 = 1.$$

*Løsningsforslag*

Mod 10 betyr at vi skal ta resten etter at resultatet er dividert med 10. Det er her underforstått at  $x$ -ene er heltall. Vi får:

$$x_1 = 3 \cdot x_0 = 3 \cdot 1 = 3$$

$$x_2 = 3 \cdot x_1 = 3 \cdot 3 = 9$$

$$x_3 = 3 \cdot x_2 = 3 \cdot 9 \pmod{10} = 27 \pmod{10} = 7$$

⋮

De 8 første tallene blir: 1, 3, 9, 7, 1, 3, 9, 7. ☺

**Kommentarer til eks. 11.3.**

- Vi får en syklus på 4 tall.
- Hadde vi valgt en annen startverdi enn  $x_0 = 1$ , og/eller en annen  $k$ -verdi enn  $k = 3$ , kunne vi fått andre sykluslengder, prøv selv!
- Legg merke til at med en gang vi får generert et tall vi har fra før, så gjentar syklusen seg, på grunn av den rekursive algoritmen.
- Vi bør ikke starte med et veldig lavt tall, fordi vi da risikerer at det/de neste tallet/tallene er «dømt» til også å bli lave. De første tallene kan derved bli «avhengige».
- Dersom vi adderer 0,5 til de 8 tallene og dividerer på  $m = 10$ , får vi slumptall mellom 0 og 1. Jo større  $m$ , jo mer ville en slik operasjon gi tilnærmet uavhengige  $U[0,1]$ -variabler.

Tilbake til det generelle: Vi etterstreber generelt slumptallgeneratorer som er i stand til å generere lange sekvenser, samtidig som de genererte tallene virkelig er uavhengige og følger en  $U[0,1]$ -fordeling. For den multiplikative kongruens-generatoren vi har sett på, er sekvenslengden såvel som de statistiske egenskapene til generatoren avhengig av både  $k$ , startverdi  $x_0$  og  $m$ .

Det finnes forøvrig mange forskjellige tester som kan utføres for å teste hvor god en aktuell generator er. Vi skal ikke gå nærmere inn på dette her, men nøye oss med å gjengi en god og anerkjent slumptallgenerator:

### Anbefalt slumptallalgoritme

$$x_{(n+1)} = 69069 \cdot x_{(n)} \pmod{2^{32}}, \quad n = 0, 1, 2, \dots$$

der  $x_0$  er et tilfeldig positivt heltall større enn ca.  $\sqrt{m}$ .

Implementering av algoritmen ovenfor på datamaskin er ikke helt enkel, uten at vi skal gå nærmere inn på dette her. I praksis tar vi ikke med alle tilgjengelige siffer, og sørger ved divisjon for å få slumptall mellom 0 og 1. I praksis er det nok kun de spesielt interesserte som har bruk for å lage sin egen slumptallrutine.

Til slutt: De fleste slumptallgeneratorer har både muligheten til å starte generatoren med samme startverdi hver gang, slik at en kan få reproduksert en tallsekvens, samt muligheten til å starte generatoren med et tilfeldig tall. Vi skal

ikke gå nærmere inn på dette her. Begge muligheter kan være fornuftige, avhengig av situasjonen.

**NB!** Dersom en skal generere en lang serie med slumptall, vil det som regel være fornuftig å la generatoren få kjøre hele løpet med kun én startverdi, i motsetning til å dele opp sekvensen i delsekvenser der hver delsekvens har forskjellig startverdi.

## 11.4 Generering av normalfordelte variabler

Vi skal her se på to metoder til å generere enkeltvariabler (endimensjonalt tilfelle). I neste avsnitt skal vi se på en metode for å generere tilfeldige variabelpar  $(X, Y)$  fra en binormal fordeling, der samvariasjonen mellom  $X$  og  $Y$  er gitt ved korrelasjonskoeffisienten  $\rho$ .

Du husker naturligvis at kumulativ normalfordelingsfunksjon ikke kunne skrives som et eksplisitt enkelt funksjonsuttrykk? La oss for sikkerhets skyld gjenta formlene fra kap. 5 for tethetsfunksjon  $f(z) = \phi(z)$  og kumulativ fordelingsfunksjon  $F(z) = \Phi(z)$  for en standard  $N(0,1)$ -variabel  $Z$ , dvs. en normalfordelt variabel med forventning  $\mu = E(Z) = 0$  og standardavvik  $\sigma = \text{std}(Z) = 1$ :

$$\text{N(0,1)-fordeling: } \phi(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}, \quad \Phi(z) = P(Z \leq z) = \int_{-\infty}^z \frac{e^{-\frac{x^2}{2}}}{\sqrt{2\pi}} dx$$

For en gitt  $z$ -verdi må altså  $\Phi(z)$  løses numerisk. Enklere er det heller ikke å finne invers kumulativ fordelingsfunksjon  $F^{-1} = \Phi^{-1}$ . Her finnes det imidlertid tilnærningsfunksjoner som i praksis er gode nok i de aller fleste tilfeller. Vi skal se på en slik tilnærningsfunksjon her, gjengitt i neste ramme:

### Tilfeldige $N(\mu, \sigma)$ -variabler med basis i tilnærrelse til $\Phi^{-1}$

- 1) Generer  $n$  uavhengige  $U[0,1]$ -variabler  $R_1, \dots, R_n$ .
- 2) Forutsatt  $R_j > 0.5$ , beregn  $Z_j$  som følger:

$$Z_j = t - \frac{c_0 + c_1 t + c_2 t^2}{1 + d_1 t + d_2 t^2 + d_3 t^3}, \quad t = \sqrt{-2 \cdot \ln(1 - R_j)} \text{ der}$$

$$\begin{aligned} c_0 &= 2.515517 & c_1 &= 0.802853 & c_2 &= 0.010328 \\ d_1 &= 1.432788 & d_2 &= 0.189269 & d_3 &= 0.001308 \end{aligned}$$

- 3) For  $R_j$ -verdier mindre enn 0.5: Sett  $R_j = 1 - R_j$ , gå fram som i 2), men sett  $Z_j = -Z_j$  til slutt.
- 4) Sett  $X_j = \mu + \sigma Z_j$ .  $X_1, \dots, X_n$  vil da tilnærmet være  $n$  uavhengige  $N(\mu, \sigma)$ -variabler.

#### Eks. 11.4 Generering av $N(0,1)$ -variabel fra tilnærmet $\Phi^{-1}$ -funksjon

##### Oppgave

Beregn tilnærmet verdi for  $\Phi^{-1}(0.05)$  med basis i forrige ramme, og sammenligne med eksakt verdi fra tabell.

##### Løsningsforslag

Fordi  $R = 0.05 < 0.5$  erstatter vi først  $R = 0.05$  med  $R = 1 - 0.05 = 0.95$ , før vi «plugger inn i formlene». Vi finner  $t = 2,447747$  og  $Z_j = -1,6452$ . Fra tabell finner vi at riktig verdi skal være ca. midt mellom  $-1,64$  og  $-1,65$ , så det stemmer bra. ☺

Det finnes imidlertid andre og «eksakte» rutiner som er langt raskere enn fremgangsmåten vist i forrige ramme. En slik går som følger:

**En effektiv måte å generere uavhengige  $N(\mu,\sigma)$ -variabler på**

- 1) La  $R_1$  og  $R_2$  betegne to uavhengige  $U[0,1]$ -variabler.
- 2) Beregn  $Z_1$  og  $Z_2$  ut fra formlene

$$Z_1 = \sqrt{-2 \cdot \ln(R_1)} \cdot \cos(2\pi R_2), \quad Z_2 = \sqrt{-2 \cdot \ln(R_1)} \cdot \sin(2\pi R_2)$$

- 3)  $X_1 = \mu + \sigma \cdot Z_1$  og  $X_2 = \mu + \sigma \cdot Z_2$  vil være 2 uavhengige  $N(\mu,\sigma)$ -variabler.
- 4) Trinn 1) - 3) gjentas inntil vi har det antall  $n$  uavhengige  $N(\mu,\sigma)$ -variabler vi ønsker.

**Eks. 11.5** **Effektiv generering av  $N(\mu,\sigma)$ -variabler.** Gitt følgende 4 tilfeldige  $U[0,1]$ -variabler:  $R: 0,693 \ 0,065 \ 0,666 \ 0,554$ .

*Oppgave*

Generer 4 tilfeldige  $N(-1,4)$ -variabler med basis i de oppgitte  $U[0,1]$ -verdier

*Løsningsforslag*

En  $N(-1,4)$ -fordeling innebærer at  $\mu = -1$  og  $\sigma = 4$ . Vi betegner  $U[0,1]$ -verdiene med henholdsvis  $R_1$ ,  $R_2$ ,  $R_3$  og  $R_4$ . Lar videre  $R_1$ - og  $R_3$ -verdiene inngå i  $\ln$ -uttrykket, og lar  $R_2$ - og  $R_4$ -verdiene inngå i argumentet til cosinus og sinus. Vi får da følgende:

$$\begin{aligned} Z_1 &= \sqrt{-2\ln(0.693)} \cdot \cos(2\pi \cdot 0.065), & Z_2 &= \sqrt{-2\ln(0.693)} \cdot \sin(2\pi \cdot 0.065) \\ Z_3 &= \sqrt{-2\ln(0.666)} \cdot \cos(2\pi \cdot 0.554), & Z_4 &= \sqrt{-2\ln(0.666)} \cdot \sin(2\pi \cdot 0.554) \end{aligned}$$

som utregnet gir  $Z_1, Z_2, Z_3$  og  $Z_4 = 0.786, 0.340, -0.850$  og  $-0.300$ . Dette gir følgende 4 tilfeldige  $N(-1,4)$ -variabler:

$$X_1 = 2.14, \quad X_2 = 0.36, \quad X_3 = -4.40 \text{ og } X_4 = -2.20. \quad \circledast$$

## 11.5 Generering av binormale variabelpar

Den binormale fordeling  $N_2(\mu_1, \sigma_1; \mu_2, \sigma_2; \rho)$  ble innført i Kap. 5.14. Vi minner kort om at dette er en bivariat fordeling  $f(x,y)$ , der  $X$  er  $N(\mu_1, \sigma_1)$ ,  $Y$  er  $N(\mu_2, \sigma_2)$ , og samvariasjonen mellom  $X$  og  $Y$  er fullstendig karakterisert ved  $\text{Cov}(X, Y) = \rho \cdot \sigma_1 \sigma_2$ . Dessuten kan marginalfordelingen for  $Y | X$  angis som følger:

$$Y | X \sim N\left(\mu_1 + \rho \cdot \left(\frac{\sigma_2}{\sigma_1}\right) \cdot (x - \mu_1), \sqrt{1 - \rho^2} \cdot \sigma_2\right)$$

Benytter vi sammenhengen  $f_{X,Y}(x,y) = f_X(x) \cdot f_{Y|X}(y | x)$ , og den raske måten å generere to uavhengige  $N(0,1)$ -variabler på angitt i forrige ramme, får vi følgende oppsett for generering av binormale variabelpar:

### Generering av binormale variabelpar

- 1) La  $(X, Y)$  være binormalt fordelt  $N_2(\mu_1, \sigma_1; \mu_2, \sigma_2; \rho)$ , dvs.  $X$  er  $N(\mu_1, \sigma_1)$ ,  $Y$  er  $N(\mu_2, \sigma_2)$  og  $\rho = \text{Cov}(X, Y) / (\sigma_1 \cdot \sigma_2)$ .
- 2) La  $R_1$  og  $R_2$  være 2 tilfeldige  $U(0,1)$ -variabler.
- 3) Beregn  $X$  og  $Y$  utifra følgende formler:

$$X = \mu_1 + \sigma_1 (-2 \ln R_1)^{1/2} \cdot \sin(2\pi R_2)$$

$$Y = \mu_2 + \sigma_2 (-2 \ln R_1)^{1/2} \cdot [(1 - \rho^2)^{1/2} \cdot \cos(2\pi R_2) + \rho \cdot \sin(2\pi R_2)]$$

$(X, Y)$  vil da være et tilfeldig variabelpar fra den gitte  $N_2$ -fordelingen.

Å generere  $n$  tilfeldige variabelpar fra en binormal fordeling, er nyttig f.eks. for å få erfaring med hvordan spredningsdiagrammet for de genererte  $(x, y)$ -parene ser ut for forskjellige verdier av  $\rho$  og  $n$ . For hver simuleringsssekvens kan vi beregne den empiriske korrelasjonskoeffisienten  $r$  mellom de simulerte  $x$ - og  $y$ -verdiene, og sammenligne med hvor godt punktene i spredningsdiagrammet synes å ligge langs en rett linje. Et annet poeng er at vi hver gang kan sammenligne  $r$  med den teoretiske korrelasjonskoeffisienten  $\rho$ , som jo i simuleringen er kjent! Ved å gjenta slike eksperimenter får vi også innblikk i hvordan variasjonen i  $r$  avhenger av antall variabelpar  $n$ , og dette kan være en nyttig erfaring å få med seg. Selv for  $\rho$ -verdier i nærheten av null, kan vi slumpe til å få  $r$ -verdier i nærheten av pluss eller minus én, dersom vi simulerer få

variabelpar  $n$ . Ved å variere  $n$  i simuleringene, får vi innsikt i hvor mange observasjoner vi må ha for at  $r$ -verdiene skal være rimelig «sikre», dvs. i nærheten av  $\rho$ -verdien.

**Eks. 11.6****Binormale variabelpar.**

Gitt den binormale fordelingen  $N_2(0,1;0,1;0,9)$ ;

*Oppgave*

Generer 30 uavhengige  $(X,Y)$ -par fra den oppgitte binormale fordeling, tegn spredningsdiagram og beregn  $r$ .

*Løsningsforslag*

Fordelingen  $N_2(0,1;0,1;0,90)$  vil si at  $X \sim N(0,1)$ ,  $Y \sim N(0,1)$  og  $\rho = \text{Corr}(X,Y) = 0,90$ . Formlene i ramma ovenfor blir da:

$$X = (-2\ln R_1)^{1/2} \sin(2\pi R_2)$$

$$Y = (-2\ln R_1)^{1/2} [(1 - 0.9^2)^{1/2} \cos(2\pi R_2) + 0.9 \sin(2\pi R_2)]$$

Vi benytter Minitab. Kode og spredningsplott er gjengitt nedenfor ( $K1000 = \pi$ ).

**MINITAB**

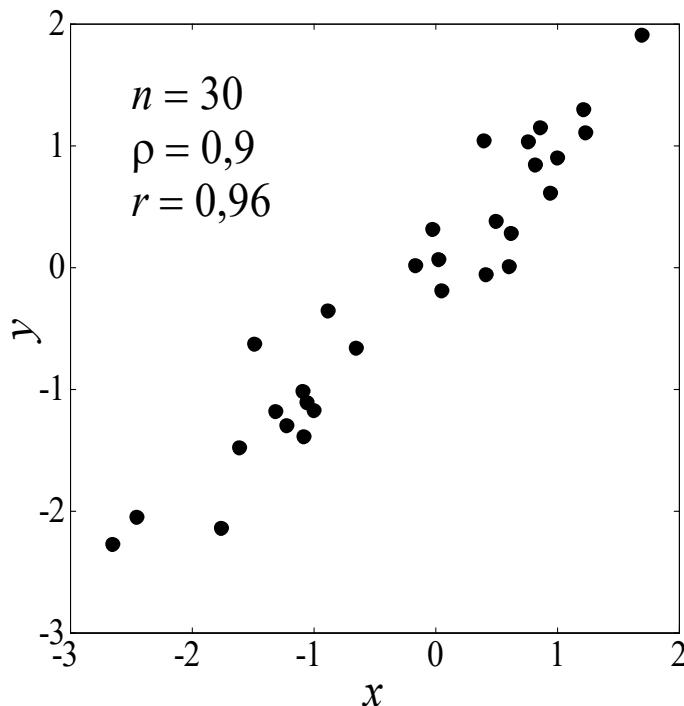
```
MTB > name c1 = 'r1=U[0,1]' c2 = 'r2=U[0,1]' c3 =' sqrt(-2ln(r1))' &
CONT> name c4 = 'cos(2pi*r2)' c5 = 'sin(2pi*r2)' &
CONT> name c6 = 'x' c7 = 'y'.           # navn på kolonne 1-7
MTB > random 30 c1 c2;                 # genererer 2*30 uavh. U[0,1]-
SUBC> uniform 0 1.                     # variabler og legger i c1 og c2
MTB > let c3 = sqrt(-2*loge(c1))       # mellomberegninger, se ramme
MTB > let c4 = cos(2*K1000*c2)         # mellomberegninger, se ramme
MTB > let c5 = sin(2*K1000*c2)         # mellomberegninger, se ramme
MTB > let k1 = 0                         # mu1 = E(X)
MTB > let k2 = 1                         # sig1 = std(X)
MTB > let k3 = 0                         # mu2 = E(Y)
MTB > let k4 = 1                         # sig2 = std(Y)
MTB > let k5 = 0.9                       # rho = corr(X,Y)
MTB > let c6 = k1 + k2*c3*c5           # tilfeldige x-variabler
MTB > let c7 = k3 + k4*c3*(sqrt(1-k5**2)*c4 + k5*c5) # tilf. y-variabler

MTB > correlation 'x' - 'y'.          # beregner korrelasjonskoeffisient
```

**Correlations (Pearson)**

Correlation of x and y = 0.958

MTB &gt; plot y\*x # lager spredningsdiagram

Figur: 30 tilfeldige variabelpar  $(X, Y)$  generert fra binormal fordeling med  $\rho = .9$ .**Eks. 11.7** Simulering av fordeling til empirisk korrelasjonskoeffisient,  $r$ .*Oppgave*

Simuler 5000 tilfeldige verdier for den empiriske korrelasjonskoeffisienten  $r$  mellom  $X$  og  $Y$ , med basis i 30 tilfeldige  $(X, Y)$ -par fra  $N_2(0,1;0,1;0,8)$ -fordelingen. Fremstill resultatet i frekvenshistogram.

*Løsningsforslag*

Vi genererer 30 tilfeldige verdier etter oppskriften i eks. 11.6, nå med  $\rho = 0.8$ , og beregner  $r$ . Dette gjentar vi 5000 ganger, og plotter de 5000  $r$ -verdiene i et histogram. Denne gangen bruker vi Matlab som dataverktøy. Kode og resultat er vist nedenfor. For dem som ikke er kjent med Matlab-kode, som er basert på vektor- og matriseformalisme: Merk hvor kompakt koden er for en såpass omfattende simulering som i dette eksemplet.

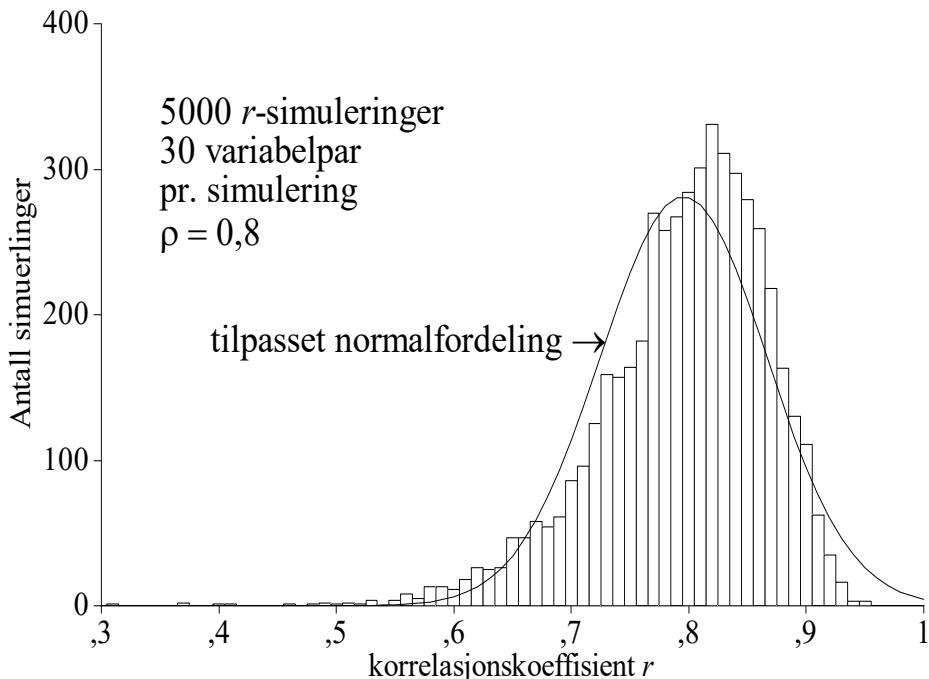
**MATLAB**

```
x = randn(30,5000); % Genererer 30*5000-matrise av uif N(0,1)-var.
```

```

y = randn(30,5000);           % Genererer 30*5000-matrise av uif N(0,1)-var.
y = 0.6*y+0.8*x;             % Beregner 30*5000-matrise av y | x -verdier
one = ones(30,1);             % hjelpevariabel av enere for riktig matrisedim.
r = sum((x-one*mean(x)).*(y-one*mean(y)));    % beregner 5000*1-matrice
r = r./std(x)./std(y)/29;      % av r-verdier
hist(r);                      % plotter histogram

```



Vi ser fra figuren at fordelingen er ganske skeiv mot venstre. Det kan vises at  $r$ -fordelingen konvergerer mot normalfordelingen når antall variabelpar blir stort nok. Tredve variabelpar er imidlertid tydeligvis ikke nok for at den symmetriske normalfordelingen skal gi en overbevisende tilpasning, og simulering er et ypperlig verktøy for å påvise dette. ☺

## 11.6 Generering av sammensatt variabel

Fra virkelighetens verden har vi ofte en deterministisk beskrivelse av sammenhengen mellom ulike variabler i form av en funksjon  $Y = f(X_1, \dots, X_n)$ , der variablene  $X_1, \dots, X_n$  som inngår, er stokastisk av natur.  $Y$  blir derved en sammensatt funksjon av stokastiske variabler, og blir derfor selv stokastisk. Vi skal belyse dette med et eksempel fra bruddmekanikkens verden.

**Eks. 11.8**
**Bruddmekanikk.**

Vi betrakter materialtretthet til en sveiseforbindelse i en oljeplattform til havs. La  $B$  betegne en bruddfunksjon som følger:

$$B = N_B - N$$

der  $N_B$  er antall belastningssykluser før materialtretthetsbrudd inntreffer, og  $N$  er det aktuelle antall belastningssykluser sveiseforbindelsen er utsatt for (gitt av sjøbølgestatistikk). Vi får altså brudd når  $B < 0$ .

Fra Det norske Veritas' regler finner vi da følgende modell for  $N_B$ :

$$N_B = \frac{a_B^{1-mn} - a_0^{1-mn}}{C\Delta\sigma^m k^m (1-mn)}$$

der

- $a_0$  = startsprekkybde
- $a_B$  = sluttsprekkybde
- $m, C$  = sprekkevekstparametre
- $n, k$  = sveisegeometriparametre
- $\Delta\sigma$  = belastningsvariasjon

Flere av variablene ovenfor er stokastiske av natur. Eksempelvis varierer lastvariasjonen med sjøbølgehøyde, og startsprekkybden vil variere med sveisekvalitet. Vi kan estimere fordelingene til disse variablene med basis i henholdsvis bølgestatistikk og plastavstøp av sveiseprøver.

Fordi  $B$  er en sammensatt funksjon av stokastiske variabler blir også  $B$  en stokastisk variabel med en fordeling med tetthetsfunksjon  $f(B)$ . Fra den deterministiske sammenhengen mellom variablene, ser vi at vi får brudd når  $B < 0$ . Vi får derfor:

$$\text{Sannsynlighet for brudd: } P_f = P(B < 0) = \int_{-\infty}^0 f(B) dB$$

Selv om vi kjenner sannsynlighetsfordelingene til enkeltvariablene som inngår, kan det imidlertid være vanskelig å bestemme  $f(B)$  teoretisk. Ved å simulere tilfeldige verdier for enkeltvariablene fra deres respektive fordelinger, og så beregne tilsvarende  $B$ -verdi, kan vi imidlertid få simulert tilfeldige verdier fra fordelingen til  $B$ . Ved å gjenta simuleringene, kan vi få simulert  $B$ -fordelingen. Vi kan også benytte slik simulering til å estimere bruddsannsynligheten  $P_f$  direkte ved rett og slett bare å telle opp antall simuleringer som gir brudd ( $B < 0$ ) og dele på totalt antall  $B$ -simuleringer. For små  $P_f$ -verdier vil simuleringen være tidkrevende. Her finnes det imidlertid teknikker basert på en kombinasjon av variabeltransformasjon og såkalt akseptanssampling-teknikk som faktisk typisk

kan føre til en effektivitetsbesparelse på en faktor 10000. Det vil føre for langt å komme nærmere inn på dette her. ☺

La oss forlate bruddmekanikkens verden, og vise et generelt oppsett for generering av tilfeldige verdier for en variabel som er sammensatt av stokastiske uavhengige enkeltvariabler:

### Generering av sammensatt variabel

- 1) La  $Y = f(X_1, \dots, X_k)$  der  $f$  er en vilkårlig (deterministisk) funksjon og  $X_1, \dots, X_k$  er  $k$  uavhengige stokastiske variabler med fordelinger  $F_1(x_1), \dots, F_k(x_k)$ .
- 2) Generer tilfeldige verdier for  $X_1, \dots, X_k$  fra deres respektive fordelinger.
- 3) Beregn  $Y = f(X_1, \dots, X_k)$  på basis av de tilfeldig genererte verdiene for  $X_1, \dots, X_k$  fra 2).  $Y$  er da en tilfeldig variabel fra fordelingen til  $Y$ .
- 4) Gjenta trinn 1)-3)  $n$  ganger. Vi får da et tilfeldig utvalg  $Y_1, \dots, Y_n$  fra fordelingen til  $Y$ .

**Eks. 11.9** **Sammensatt variabel.** La  $Z = X_1 \cdot \ln X_2 \cdot \sqrt{X_3}$  være en sammensatt variabel der  $X_1, X_2$  og  $X_3$  er uavhengige,  $X_1 \sim N(0,1)$ ,  $X_2 \sim \text{Expo}(2)$  og  $X_3 \sim N(20,1)$ .

#### Oppgave

Generer en tilfeldig verdi for  $Z$  fra dennes fordeling.

#### Løsningsforslag

Vi benytter den effektive metoden angitt i kap. 12.4 for generering av de to normalfordelte variablene og får:  $x_1 = 0.723$  og  $x_3 = 18.7$ . Videre baserer vi generering av en tilfeldig verdi for  $X_2$  fra expo(2)-fordelingen på «invers kumulativ fordelingsfunksjon»-metoden beskrevet i kap. 12.2, og får  $x_2 = 3.11$ . Vi får dermed følgende z-verdi:  $z = 0.723 \cdot \ln(3.11) \cdot (18.7)^{1/2} = 3,547$ .

#### Kommentar.

$\sqrt{X_3}$  er strengt tatt bare definert for  $X_3 \geq 0$ , mens normalfordelingen  $N(\mu, \sigma)$  er definert for  $-\infty < X_3 < \infty$ . Når  $\sigma / |\mu| < \text{ca. } 1/5$ , kan likevel normalfordelingen

være en fornuftig tilnærmelse. Dette skyldes at sannsynlighetsmassen til venstre for null blir forsvinnende liten. ☺

## 11.7 «Bootstrap»

Bootstrap er en svært anvendelig og brukervennlig simuleringsteknikk som blir stadig mer populær i takt med den rivende utviklingen innen data. En stor fordel med bootstrap er dens såkalte «**modellfrie**» natur, dvs. det er en teknikk der vi for eksempel ikke behøver å gjøre noen antakelser om fordelingen(e) til våre observasjonsvariabler. For omfattende simuleringer krever den imidlertid stor lagringskapasitet og høy ytelse (rask maskin).

Bootstrap kan blant annet benyttes til **hypotesetesting** og konstruksjon av **konfidensintervall**. En mye brukt anvendelse er å estimere **standardavviket** til kompliserte variabler. Vi skal senere se på et eksempel der de to sistnevnte anvendelsene demonstreres. La oss først forklare litt generelt om hva bootstrap går ut på.

Anta at du har et tilfeldig utvalg på 30 datapar  $(x_1, y_1), \dots, (x_{30}, y_{30})$  fra en fordeling  $f(x, y)$ , og at du ønsker å estimere standardavviket til korrelasjonskoeffisienten  $r$  mellom  $X$  og  $Y$ . Med basis i dine data har du kun én estimert  $r$ -verdi. Du gjør ingen antagelser om populasjonsfordelingen  $f(x, y)$ . Å finne en fornuftig estimator for  $\text{std}(r)$  rent matematisk ut fra formelen for  $r$  er generelt vanskelig uten å gjøre antagelser om  $f(x, y)$ . Hypotetisk kan du tenke deg at du gjør følgende: Gjenta forsøket f.eks. 100 ganger, der hvert forsøk fremskaffer 30 nye tilfeldige  $(x, y)$ -datapar fra  $f(x, y)$ . For hvert forsøk beregnes  $r$ , og du beregner til slutt standardavviket til  $r_1, \dots, r_{100}$ . I praksis må vi imidlertid som regel nøye oss med det datasettet vi har.

Ved bootstrap anvendt på situasjonen ovenfor, simulerer du dine 100 forsøk som følger: Du tenker deg en urne med 30 lapper nummerert fra 1 til 30. Så trekker du 30 lapper *med tilbakelegging*. Du lar så disse bestemme hvilke av dine ekte datapar som skal utgjøre det simulerte forsøket. Hvis du f.eks. trekker 7, 28, 3, 7, ... så vil det simulerte forsøket bestå av  $(x_7, y_7)$ ,  $(x_{28}, y_{28})$ ,  $(x_3, y_3)$ ,  $(x_7, y_7)$ , ... og du kan beregne en ny  $r$ -verdi på basis av disse. Dette kan så gjentas 100 ganger, og du kan beregne standardavviket til de 100  $r$ -verdiene som fremskaffes på denne måten. Legg merke til at for hvert «forsøk» vil ett og samme datapar gjerne gå igjen flere ganger, mens andre datapar ikke kommer med.

Bootstrap virker i første omgang som en «hokus pokus»-teknikk som er vanskelig å forstå hvorfor virker. Dette er trolig bakgrunnen for begrepet, som på norsk betyr «støvlestropp». Etter sigende skal opprinnelsen ha vært en av Rudolph Erich Raspe's historier om baron Munchhausen: Baronen hadde falt til bunn i en dyp innsjø. Mens alt syntes fortapt, kom han på idéen å løfte seg selv etter sin egen støvlestropp (analogt til uttrykket «løfte seg selv etter håret» på norsk).

Ved litt omtanke skjønner vi at bootstrap er ekvivalent med å simulere fra den empiriske kumulative fordelingen til våre data, som estimat for den sanne bakenforliggende fordeling. Når antallet observasjoner blir stort, virker dette intuitivt fornuftig. Hvor mange observasjoner som er tilstrekkelig er det imidlertid vanskelig å gi noen tommelfingerregel for, da dette vil avhenge av situasjonen.

**Eks. 11.10** **Bootstrap og diettindeks hos grønlandssel.** I forskningsfangst på sjøpattedyr tas mageprøver for å undersøke hva de spiser. Et viktig formål er å få kunnskap om hvor stor andel av viktige kommersielle arter, f.eks. sild, som beskattes. Som mål på relativ andel i seldietten av en art  $i$  kan en beregne diettindeksen

$$B_i = \frac{\text{sum av masse av art } i \text{ funnet i alle } n \text{ undersøkte sel}}{\text{sum av all byttedyrmasse funnet i alle } n \text{ undersøke sel}} \cdot 100\%,$$

$$i = 1, \dots, m$$

der vi studerer totalt  $m$  kategorier byttedyr.

### Oppgave

Del antall byttedyr i 2 kategorier:  $i = 1$  for sild og  $i = 2$  for resten. Forklar hvordan du ved bootstrap kan estimere standardavviket til diettindeksen for sild, samt konstruere et tilnærmet 95 % konfidensintervall for forventningsverdien  $b_1 = E(B_1)$ . Anta at du har data for 50 dyr.

### Løsningsforslag

Vi lar  $x_1, \dots, x_{50}$  og  $y_1, \dots, y_{50}$  betegne sildebiomassene og resten av biomassene funnet i sel 1 til 50. På matriseform kan vi da sette opp vårt datamateriale som følger (fiktive, men realistiske data):

Sel nr:	$x =$ sild [g]	$y =$ andre [g]	$bt =$ sum [g]
1	0	12	12
2	113	0	113
:	:	:	:

50		1196		23		1219
----	--	------	--	----	--	------

Merk hvor «usikre» dataene som er vist for 3 av selene ovenfor virker: 2 har nesten total dominans av sild, den tredje har ingen sild, og total biomasse varierer svært. Det siste skyldes at selen fordøyer raskt, og det er tilfeldig når i spisefasen dyrene fanges.

Det trengs flere bootstrap-simuleringer,  $n_{boot}$ , for å estimere konfidensintervall enn for å estimere standardavvik. Som forsiktige tommelfingerregler kan vi bruke  $n_{boot} = 200$  for estimering av standardavvik og  $n_{boot} = 2000$  for å estimere et 95 % konfidensintervall. Har vi lagringskapasitet og tid er det ingen ulempe å legge seg på den forsiktige siden. I dette eksemplet skal vi velge  $n_{boot} = 5000$ .

Vår fremgangsmåte blir nå som følger:

1. Trekk 50 tilfeldige heltall (med tilbakelegging) mellom 1 og 50 fra uniform fordeling, og betegn disse med  $j_1, \dots, j_{50}$ .
2. La trekningene fra 1 definere selnr., dvs. hvilke sel vi skal bruke data fra.
3. Beregn  $B_{i,boot} = \sum_{k=1}^{50} x_{j_k} / \sum_{k=1}^{50} (x_{j_k} + y_{j_k}) \cdot 100\%$
4. Gjenta punkt 1-3 ovenfor 5000 ganger.
5. Beregn standardavviket til  $(B_{i,boot})_1, \dots, (B_{i,boot})_{5000}$
6. Sorter  $(B_{i,boot})_1, \dots, (B_{i,boot})_{5000}$  i stigende rekkefølge, og konstruer konfidensintervallet  $[L, U]$ , der  $L$  er den 125. minste verdien og  $U$  er den 125. største verdien av  $B_{i,boot}$  (forutsetter en symmetrisk  $B_{sild}$ -fordeling).

Nedenfor er vist Matlabkode og resultat ved anvendelse på hele datamaterialet (kun data for 3 av selene er vist her). Histogram over de bootstrapsimulerte diettindeksene er også vist. Dataene ligger i sel-byttedyr matrisa «bio» med 50 linjer (en for hver sel) og biomasseverdier for sild i 1. kolonne. Symbolet ' betegner transponert, eksempelvis betyr bio' den transponerte av bio.

### MATLAB

```

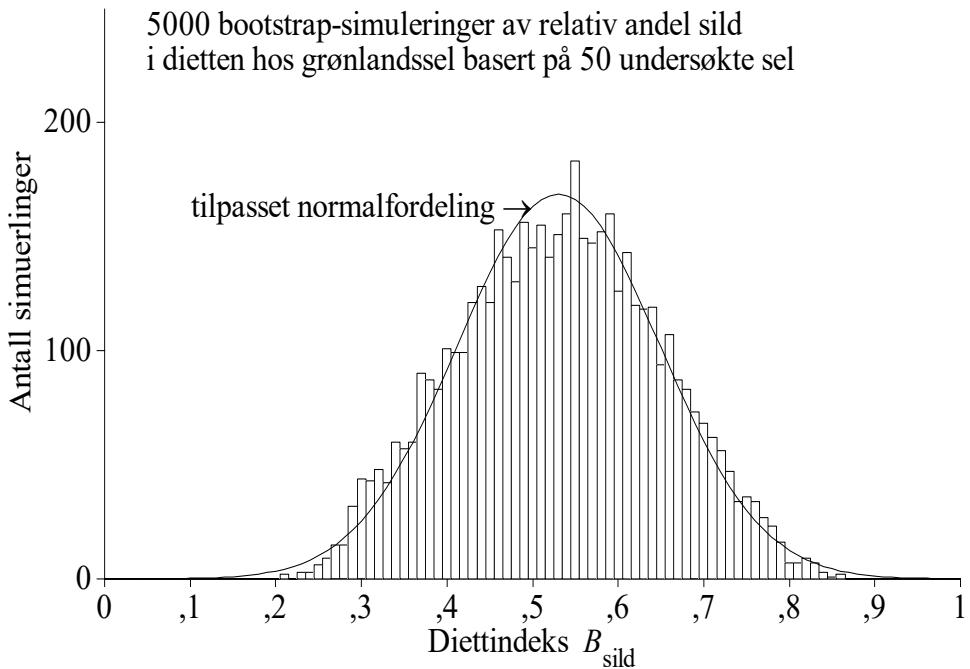
biot=sum(bio');
% sum av byttedyrmasse i hver sel
B=sum(bio)./sum(biot)*100;
% sildeindeks
iboot=fix(rand(50,5000)*50)+1; % 50*5000 indeksmatrise
bt=reshape(biot(iboot),50,5000); % 50*5000 bootstrap totalmassematrice

```

```

b=reshape(bio(iboot,1),50,5000); % 50*5000 bootstrap sildemassematrice
Bboot=sum(b)/sum(bt)*100;      % 5000 bootstrap-sildeindeks
sB=std(Bboot);                 % beregner standardavviket til Bboot
[a,b]=sort(Bboot);              % sortering i stigende rekkefølge
L=a(125);                      % simulert nedre 2,5 %-fraktil
U=a(5000-125);                 % simulert øvre 2,5 %-fraktil
[B(1) L U]                      % sildeindeks basert på ekte data og
ans =                                % simulert 95 % konfidensintervall for
      54.1055 30.7729 75.1940       % forventet sildeindeks
hist(Bboot)                         % histogramplott av Bboot

```



Merk at i dette tilfellet er det en svært bra normaltilpasning til histogrammet basert på bootstrap. Dette skyldes en kombinasjon av at vi har ganske mange dyr (50) og at diettindeksten for sild,  $B$  (54 %), er nær 0.5. Merk imidlertid at fordelingen er bred, hvilket samsvarer med det estimerte standardavviket som var på 12 %. Det estimerte 95 % konfidensintervallet for  $B_{\text{sild}}$  er [31 %, 75 %].

## 11.8 Oppgaver.

Mange av oppgavene henviser til følgende slumptall generert med  $U[0,1]$ -generator:

$$R_1 = 0.8617$$

$$R_2 = 0.3841$$

$$R_3 = 0.5595$$

$$R_4 = 0.1789$$

**11.1** La  $R$  være en  $U[0,1]$ -variabel. Hva blir oppsettet for å generere en tilfeldig  $U[-3,10]$ -variabel  $X$ ? Hva blir  $X$ -verdien på basis av  $R_1$ ?

**11.2** En eksponensialfordelt variabel  $X$  har forventning  $E(X) = 0.1$ . Bestem på enkleste måte 2 tilfeldige  $X$ -verdier på basis av  $R_1$  og  $R_2$ .

**11.3** Bestem på grunnlag av  $R_1$  til  $R_4$  fire tilfeldige binomiske variabelverdier  $X_1 - X_4$  fra binomisk fordeling med parametre  $n = 12$  og  $p = 0.3$ .

**11.4** Bestem på grunnlag av  $R_1 - R_4$  fire tilfeldige Poissonvariable fra Poissonfordelingen med parameter  $\lambda = 3$ .

**11.5** Bestem  $k$  i tetthetsfunksjonen  $f(x) = k \cdot e^{-5x}$ ,  $x \geq 0$ . Generer deretter 2 tilfeldige variabelverdier fra fordelingen på enkleste måte basert på  $R_1$  og  $R_2$ .

**11.6** Bestem syklus og sykluslengde til generatoren  $x_{n+1} = 3 \cdot x_n \pmod{10}$  med  $x_0 = 4$ .

**11.7** En frekvenstest for en  $U[0,1]$  generator gav følgende resultat ( $O_i$  = antall tall mellom  $(i-1) \cdot 0.2$  og  $i \cdot 0.2$ ,  $i = 1, 2, 3, 4$  og  $5$ ):  $O_1 = 14$ ,  $O_2 = 27$ ,  $O_3 = 19$ ,  $O_4 = 30$ ,  $O_5 = 10$ . Bestem verdi til testobservatoren og test på nivå 5% ved

hjelp av Pearson's kjikvadrat tilpasningstest.

**11.8** Generer to tilfeldige  $N(-1,2)$ -variable på basis av  $R_1$  og  $R_2$  og på basis av tilnærmelse til  $\Phi^{-1}$ .

**11.9** Generer to tilfeldige  $N(3, 0.1)$ -variable på basis av  $R_1$  og  $R_2$  og effektiv metode.

**11.10** Generer et tilfeldig variabelpar fra binormal fordeling  $N_2(0.1; 0.1; -0.5)$  med basis i  $R_1$  og  $R_2$ .

## 11.9 Formelsamling

### Betegnelser

$F(x)$  = Kumulativ fordelingsfunksjon

$f(x)$  = tethetsfunksjon

$F^{-1}$  = invers kumulativ fordelingsfunksjon

$R_1, \dots, R_n = n$  uavhengige  $U[0,1]$ -variable.

### Generering fra $F^{-1}$

$X_i = F^{-1}(R_i), i = 1, \dots, n \Rightarrow$

$X_1, \dots, X_n$  er  $n$  uavhengige variable fra kontinuerlig fordeling  $F(x)$ .

### Weibullfordelt variabel

$X_j = -b (\ln R_j)^{1/c}$  er tilfeldig variabel fra Weibullfordelingen

$$F(x) = 1 - e^{-(x/b)^c}, \quad x \geq 0.$$

$c = 1$  tilsvarer expo( $b$ )-fordelingen

$c = 2$  tilsvarer Rayleigh( $b$ )-fordelingen

### Diskret variabel

$X$  har fordeling  $F_{(i)} = P(X \leq x_{(i)})$ ,

$i = 1, \dots, m$ .  $F_{(k),j} =$  Minste av verdiene

$F_{(1)}, \dots, F_{(m)}$  som er større enn  $R_j \Rightarrow$

$x_{(k)}$  = tilfeldig  $X$ -variabel.

### $N(\mu, \sigma)$ -variabel fra $\Phi^{-1}$ -tilnærmelse

$$Z = t - \frac{c_0 + c_1 t + c_2 t^2}{1 + d_1 t + d_2 t^2 + d_3 t^3}$$

$$t = (-2\ln(1-R_j))^{1/2}$$

$$c_0 = 2.515517 \quad d_1 = 1.432788$$

$$c_1 = 0.802853 \quad d_2 = 0.189269$$

$$c_2 = 0.010328 \quad d_3 = 0.001308$$

$\Rightarrow Z$  tilfeldig  $N(0,1)$ -variabel forutsatt  $R_j > 0,5$ . For  $R_j < 0,5$ : Sett  $R_j = 1 - R_j$ , gjør som over, og erstatt  $Z$  med  $-Z$ . Da er  $X = \mu + \sigma Z$  tilfeldig  $N(\mu, \sigma)$ -variabel.

### Effektiv $N(\mu, \sigma)$ -generering

$$Z_1 = (-2 \cdot \ln R_1)^{1/2} \cdot \cos(2\pi R_2)$$

$$Z_2 = (-2 \cdot \ln R_1)^{1/2} \cdot \sin(2\pi R_2)$$

$\Rightarrow X_1 = \mu + \sigma Z_1$  og  $X_2 = \mu + \sigma Z_2$  er to uavhengige  $N(\mu, \sigma)$ -variable.

### Binormale variabelpar

$$L = (-2 \ln R_1)^{1/2},$$

$$S = \sin(2\pi R_2), C = \cos(2\pi R_2)$$

$$X = \mu_1 + \sigma_1 \cdot L \cdot S$$

$$Y = \mu_2 + \sigma_2 L [(1 - \rho^2)^{1/2} \cdot C + \rho \cdot S]$$

$\Rightarrow (X, Y)$  tilfeldig variabelpar fra binormal fordeling  $N_2(\mu_1, \sigma_1; \mu_2, \sigma_2; \rho)$

### Generering av sammensatt variabel

$Y = f(X_1, \dots, X_k)$ ,  $X_1, \dots, X_k$  uavhengige med fordelinger  $F_1(x_1), \dots, F_k(x_k)$ .

$X_1, \dots, X_k$  genereres fra sine respektive fordelinger

$\Rightarrow f(X_1, \dots, X_k)$  tilfeldig  $Y$ -variabel.

### Bootstrap

La  $x_1, \dots, x_n$  betegne et datamateriale, og la  $I_1, \dots, I_n$  betegne  $n$  uavhengige og uniformt fordelte heltall  $\in \{1, 2, \dots, n\}$ .  $\{x_{I_1}, \dots, x_{I_n}\}$  utgjør da et tilfeldig bootstrap-replikat fra opprinnelig datamateriale.

### Minitab

MTB > random 50 c1;

SUBC > normal 0 1.

Genererer 50  $N(0,1)$ -variabler og legger disse i kolonne c1. For andre fordelinger erstattes «normal» med aktuelt fordelingsnavn og «0 1» med aktuelle parametre.

## Kapittel 12

# Shewart-diagrammer

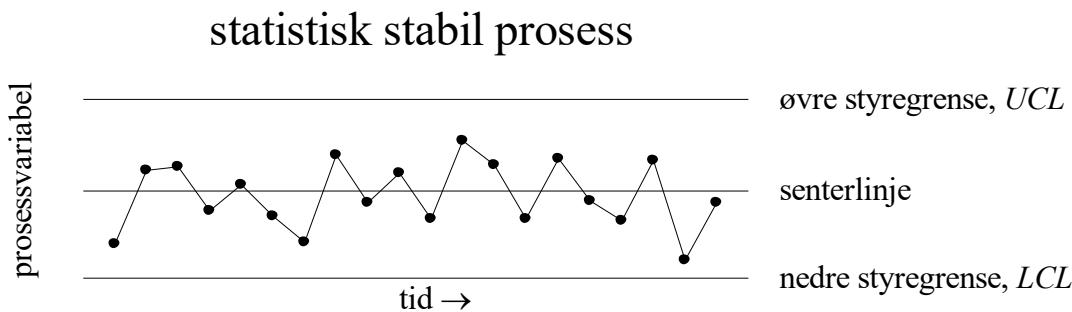
### 12.1 Innledning

Prosesser leder til varierende resultat. Egenskaper ved en masseprodusert vare vil variere fra enhet til enhet. Årsakene til variasjonen er som regel mangfoldige. Varierende materialkvalitet, skiftende temperatur, og dagsformen til en operatør kan være typiske eksempler.

Et hovedformål med statistisk prosesstyring (*Statistical Process Control*, forkortet *SPC* på engelsk) er å overvåke en prosess og slå alarm når det er rimelig sikkert at enten **nivået** eller **variasjonen** til prosessen er i ferd med å forandres. Noen ganger vil det være lett å skille ut spesielle årsaker til ustabilitet, slik som at en maskin er utslitt. I praksis vil vi imidlertid alltid stå igjen med rent tilfeldig variasjon, som ikke er lett identifiserbar eller lett å gjøre noe med. Så lenge variasjonen er liten nok og er av en slik karakter at vi har kontroll over den (statistisk stabil prosess) medfører den i liten grad en forringelse av kvalitet. Innen kvalitetsarbeid streber vi imidlertid kontinuerlig etter å redusere variasjonen til et minimum.

Vi skal her begrense oss til å studere **styrediagrammer** (*control charts* på engelsk), som er et viktig verktøy innen statistisk prosesstyring, blant annet for å overvåke en prosess. Utgangspunktet er at vi har en målbar prosessvariabel som mål på kvalitet (*process quality indicator* på engelsk). Et eksempel kan være måling av væskennivået i en fylt brusflaske før korken settes på. De fleste har sikkert opplevd at dette nivået varierer, noen ganger er flaska fylt nesten helt opp til korken, mens andre ganger når nivået knapt nok opp til flaskehalsen.

I styrediagrammet har vi oftest tid eller antall langs horisontalaksen, og de målte verdiene av prosessvariabelen langs vertikalaksen. Vi skal her begrense oss til å betrakte tiden som uavhengig variabel. Målingene blir foretatt med *like store* (ekvidistante) *tidsintervall* mellom hver gruppe av målinger. Diagrammet inneholder en **øvre styregrense**, på engelsk forkortet til UCL (*Upper Control Limit*) og en **nedre styregrense**, på engelsk forkortet til LCL (*Lower Control Limit*). Så lenge prosessvariabelen (styrevariabelen) holder seg mellom disse to grensene, sier vi at prosessen er statistisk stabil. Det bør også tegnes inn en senterlinje i styrediagrammet (se neste figur).



*Fig. 12.1 Illustrasjon av statistisk stabil prosess: Observasjonene holder seg innenfor øvre og nedre styregrensene.*

Med en gang vi får en måleverdi som er større enn UCL eller mindre enn LCL (se neste figur), slår vi «alarm». Da stoppes ofte prosessen, og vi undersøker om vi kan identifisere noen spesiell årsak til denne «ekstreme» måleverdien. Deretter prøver vi å gjøre noe som kan få prosessen stabil igjen. Fordi variasjonen er statistisk av natur, kan det imidlertid tenkes at det ikke er noe unormalt ved prosessen, og at alarmen var «falsk».

Styregrenser må ikke forveksles med **toleransegrenser/spesifikasjonsgrenser**. Sistnevnte grenser kan være krav stilt av f.eks. markedet, kunden eller en designingeniør.

**NB!** *Uansett om toleransegrensene/spesifikasjonsgrensene ligger utenfor eller innenfor styregrensene, er det styregrensene som bestemmer når en prosess skal stoppes.*

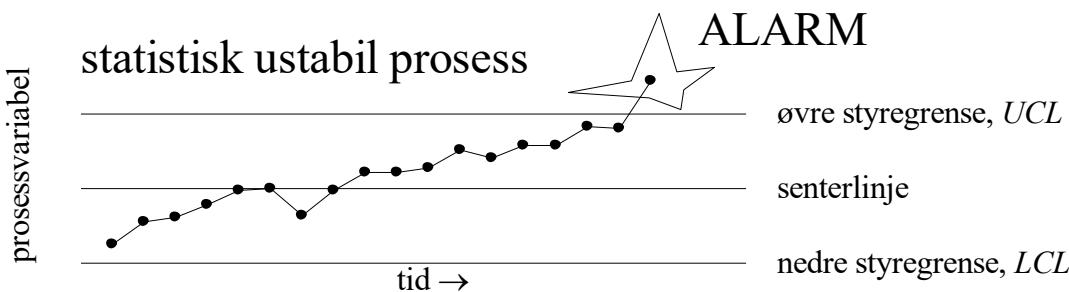
Et sentralt spørsmål er *hvor* vi skal sette øvre og nedre styregrense. To åpenbare momenter er som følger:

- 1) Setter vi styregrensene for nært senterlinjen vil vi fort få alarm dersom prosessen er i ferd med å bli ustabil, og dette er bra. Imidlertid vil vi få svært mange falske alarmer, noe som kan medføre store kostnader.
- 2) Setter vi styregrensene for langt fra senterlinjen vil vi oppnå få falske alarmer, men det kan gå lang tid før vi oppdager at en prosess er kommet helt ut av kontroll. Vi kan få situasjonen illustrert i figuren nedenfor. Dette kan også bli svært dyrt.

I praksis må vi velge grensene slik at vi balanserer fordelen med å ha prosessen under best mulig kontroll mot ulempen ved mange falske alarmer. I **Shewhart**-diagrammene settes grensene til 3 standardavvik på hver side av

senterlinjen. At faktoren 3 og ikke 2 eller et annet tall ble valgt var vel fundert, noe vi skal komme tilbake til.

Shewarts-diagrammer ble introdusert så tidlig som i 1920-årene, men ble ikke tatt i bruk i noe særlig omfang før langt senere. Det finnes mange varianter av Shewart-diagrammer, og vi skal behandle noen av de mest utbredte.



*Fig. 12.2 Illustrasjon av statistisk ustabil prosess, der alarmen går når en observasjonsverdi overskridet øvre styregrens.*

Når vi skal innføre bruk av Shewart-diagrammer, foregår dette normalt i **to faser**:

- 1) Vi foretar tilstrekkelig med målinger mens prosessen oppfører seg «normalt» til at vi kan få beregnet fornuftige styregrens.
- 2) Vi overvåker prosessen kontinuerlig, og bruker styregrensene etablert fra fase 1) som kriterier for om prosessen er statistisk stabil eller ikke.

Et viktig formål med styrediagrammene er at de aktivt skal brukes som et verktøy til en kontinuerlig forbedring av prosessen (reduksjon av variabilitet).

Kapitlet er disponert som følger: I kap. 12.2 ser vi på såkalte  $\bar{X} - R$ -diagram, eller «middelverdi-variasjonsbredde»-diagram. Her står  $\bar{X}$  for gjennomsnittet av  $n$  målinger ved hvert tidspunkt, og  $R$  står for differansen mellom den største og minste måleverdien av de  $n$  verdiene ved hvert tidspunkt. I kap. 12.3 behandles  $XmR$ -diagrammet, som er aktuelt når vi kun har én måleverdi pr. måletidspunkt ( $mR$  er forkortelse for «moving range», dvs. glidende variasjonsbredde). Deretter skal vi kommentere de gode robuste egenskapene til Shewart «3 sigma» grensene i kap. 12.4. Kapitlet avsluttes med oppgaver i kap. 12.5 og en formelsamling og tabell i kap. 12.6.

## 12.2 $\bar{X}$ -R-diagrammer

$\bar{X}$ -R-diagrammet er et sammensatt diagram som består av et  $\bar{X}$  diagram og et R-diagram. Disse tegnes over hverandre, slik at vi lett kan sammenligne middelverdier og variasjonsbredder beregnet ved samme måletidspunkt.

$\bar{X}$ -diagrammet (middelverdi-diagrammet) viser hvordan nivået til prosessen utvikler seg, mens R-diagrammet (variasjonsbredde-diagrammet) viser hvordan variasjonen til prosessen forløper. Vi skal begrense oss til å se på hvordan styregrensene i de to diagrammene beregnes basert på midlere variasjonsbredde,  $\bar{R}$ .

### $\bar{X}$ -diagram

Vi starter med å beskrive middelverdidiagrammet. Anta at vi ved hvert måletidspunkt  $t_i$  har  $n$  observasjoner av vår målevariabel, og la  $\bar{X}_i$  betegne middelverdien av de  $n$  måleverdiene. Videre antar vi at vi har  $k$  måletidspunkt,  $t_1, \dots, t_k$ . Når vi tegner middelverdiene inn i et styrediagram, velges da styregrensene som følger:

$$LCL = \bar{\bar{X}} - 3 \frac{\sigma^*}{\sqrt{n}}, \quad UCL = \bar{\bar{X}} + 3 \frac{\sigma^*}{\sqrt{n}}, \quad n \geq 1$$

der  $\bar{\bar{X}} = \frac{1}{k} \sum \bar{X}_i$  betegner middelverdien til de  $k$  gjennomsnittsverdiene funnet ved hvert måletidspunkt, og  $\sigma^*$  er en estimator for standardavviket,  $\sigma$ , til måleverdiene *ved hvert enkelt måletidspunkt*. I en statistisk stabil prosess er  $\sigma$  konstant over tid, dvs. uavhengig av måletidspunkt.

Dersom en middelverdi ikke holder seg innenfor disse grensene, stoppes prosessen for å undersøke om det er noe galt med prosessen. Dersom  $\bar{X}$  er tilnærmet normalfordelt, er sannsynligheten for falsk alarm ca. 0,027, dvs. ca. én av mellom 300 og 400 målinger fører til unødig stans i produksjonen. I praksis har det vist seg at sannsynligheten for falsk alarm ligger på omrent samme nivå, selv ved vesentlige avvik fra normalfordelingen.

Et middelverdi-diagram er illustrert i neste figur med  $n = 5$  og  $k = 20$ . Her er også enkeltobservasjonene inntegnet, for å vise at disse godt kan overskride styregrensene selv om prosessen er statistisk stabil.

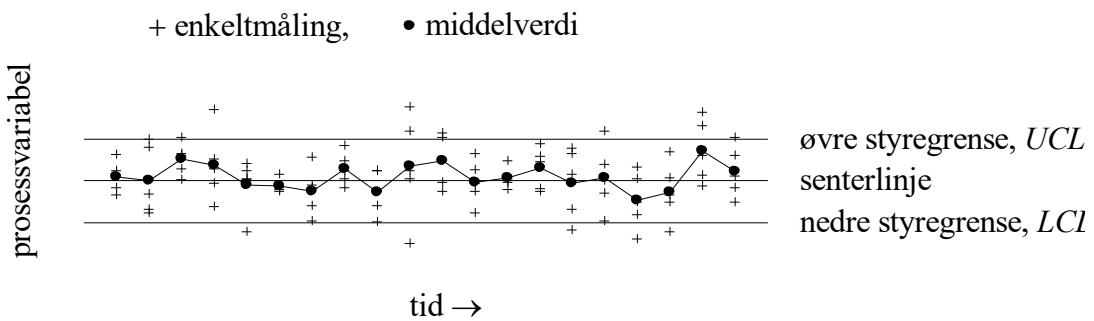


Fig. 12.3 Enkeltmålinger og middelverdier inntegnet i ett og samme diagram. Styregrensene er satt i forhold til middelverdiene. Merk at prosessen er statistisk stabil selv om mange enkeltmålinger ikke holder seg innenfor styregrensene.

Det er flere måter å beregne  $\sigma^*$  på. Merk at det er spredningen av enkeltobservasjonene rundt middelverdien ved hvert måletidspunkt som skal estimeres, og ikke spredningen rundt senterlinja. Vi skal her nøyne oss med å angi en enkel og mye brukt metode basert på de målte variasjonsbreddene,  $R_1, \dots, R_k$ . Variasjonsbredden  $R_i$  er differansen mellom største og minste målte verdi av prosessvariablen ved tidspunkt  $t_i$ . Med basis i variasjonsbreddene kan øvre og nedre styregrensene beregnes som følger:

$$UCL_{\bar{X}} = \bar{\bar{X}} + A_2 \bar{R}, \quad LCL_{\bar{X}} = \bar{\bar{X}} - A_2 \bar{R},$$

der  $\bar{R}$  er middelverdien av de  $k$  variasjonsbreddene vi har målt, og der  $A_2$  er en konstant som kun avhenger av  $n$ . Vi angir styregrensene med en indeks  $\bar{X}$  for å indikere at det er middelverdi-diagrammet vi ser på styregrensene for.  $A_2$ -verdiene finnes fra tabell i kap. 12.7, og er konstruert slik at styregrensene ovenfor er basert på forventningsrette estimatorer for  $\sigma$ , forutsatt stabil prosess og at prosessvariablene er uif og normalfordelte. Det har vist seg i praksis at disse grensene er fornuftige selv ved betydelige avvik fra normal-forutsetningen.

Hvor stor bør så  $n$  være? Ofte vil det være billigere og enklere f.eks. å foreta 10 målinger ved samme tidspunkt enn 2 målinger ved 5 forskjellige tidspunkt. Jo flere målinger man foretar på bekostning av hyppigheten til målingene, jo mindre vil man imidlertid være i stand til å fange opp raske endringer i prosessen. Den optimale  $n$ -verdien vil derfor avhenge av flere forhold, og kan variere fra prosess til prosess.

Dersom vi velger  $n = 5$ , blir middelverdien spesielt enkel å regne ut: Man summerer de 5 tallene, ganger med 2 og flytter komma en plass til venstre (deler på 10).

Eks:  $3 + 4 + 3 + 3 + 2 = 15$ ,  $15 \cdot 2 = 30$ , dvs. middelverdien er lik 3,0.

## R-diagram

Middelverdi-diagrammet brukes til å overvåke prosessnivået, mens variasjonsbredde-diagrammet (*R*-diagrammet) brukes til å overvåke prosessens **variasjon**. I prinsippet beregnes styregrensene for variasjonsbreddediagrammet på samme måte som for middelverdien med basis i 3-standardavviksgrenser. Verdiene blir imidlertid forskjellige i de to tilfellene. Husk at variasjonsbredden aldri er negativ, siden den er differansen mellom største og minste målte verdi av prosessvariabelen.

Prosessen defineres som ustabil dersom en eller flere av variasjonsbreddene ligger utenfor styregrensene. Siden middelverdi og variasjonsbredde måler to vesens forskjellige egenskaper ved prosessen, kan vi i praksis oppleve at det ene diagrammet indikerer en statistisk ustabil prosess selv om ikke det andre diagrammet gjør det. ProsesSEN defineres som «ut av kontroll» uansett hvilket av diagrammene som gir alarm.

Vi kan beregne styregrensene for variasjonsbreddediagrammet som følger:

$$UCL_R = D_4 \bar{R}, \quad LCL_R = D_3 \bar{R},$$

der styregrensene er angitt med indeks  $R$  for å indikere at det er styregrensene for variasjonsbreddediagrammet vi betrakter.  $D_3$  og  $D_4$  er konstanter som kun avhenger av  $n$ . Disse er gjengitt i tabell i kap. 12.7. Konstantene er konstruert under forutsetning av uif og normalfordelte prosessvariabler. Selv ved tildels store avvik fra denne forutsetningen har det imidlertid vist seg at sannsynligheten for falsk alarm ligger på omrent samme nivå som i normalfordelingstilfellet, dvs. den er sjeldent større enn ca. 1 %.

Når  $n \leq 5$  blir nedre 3-standardavvik-grense negativ, hvilket ikke gir noen mening i forhold til variasjonsbredden. Nedre styregrense i variasjonsbreddediagrammet settes lik 0 i disse tilfellene.

Når vi skal etablere styregrensene, velger vi en tilstrekkelig lang periode der prosessen oppfører seg «normalt». Som en tommelfingerregel bør man utføre minst 20 grupper med målinger ( $k = 20$ ), der hver gruppe består av  $n$  enkeltmålinger.

**Eks. 12.1** **Frysetørket mat.** Et firma produserer frysetørket mat i pose, der en reguleringsprosess sørger for at vektinnholdet i gram som påfylles hver pose er noenlunde konstant. For å overvåke at påfyllingsprosessen

er stabil, ønsker man å etablere styregrenser for et  $\bar{X} - R$  Shewart-diagram. Regulatoren for påfyllingsmaskinen er innstillet slik at det skal fylles 125 g i hver pose. Tabellen nedenfor viser målingene som styregrensene skal bestemmes fra.

Tab. 12.1 Posevekt for frysetørket mat. Se eks. 12.1

Kl.:	07	08	09	10	11	12	13	14	15	16
X [g]:	123.2	125.4	123.2	124.9	125.2	123.2	126.2	127.2	124.5	123.7
	121.7	125.0	123.4	123.3	126.0	123.0	125.3	126.7	123.7	124.6
	126.5	123.5	127.2	123.0	123.2	126.4	123.5	124.0	125.7	126.0
	124.2	123.6	125.1	124.6	124.3	125.0	127.0	123.1	127.2	123.7
	125.5	124.4	123.2	126.4	124.6	124.0	125.4	124.9	124.2	123.2
$\bar{X}$ [g]	124.2	124.4	124.4	124.4	124.7	124.3	125.5	125.2	125.1	124.2
R [g]	4.8	1.9	4.0	3.4	2.8	3.4	3.5	4.1	3.5	2.8
Kl.:	17	18	19	20	21	22	23	24	01	02
X [g]:	124.8	124.3	125.7	125.8	126.5	123.1	124.3	123.5	127.3	124.5
	124.9	124.9	123.6	124.7	122.6	125.4	125.6	126.6	126.1	123.3
	125.7	127.3	125.1	121.9	124.9	124.4	126.1	124.4	127.9	124.7
	124.1	124.1	124.1	125.2	124.0	125.1	128.2	122.9	125.8	126.8
	124.8	123.0	125.8	127.4	123.5	124.4	123.0	125.5	127.8	123.3
$\bar{X}$ [g]	124.9	124.7	124.9	125.0	124.3	124.5	125.4	124.6	127.0	124.5
R [g]	1.6	4.3	2.2	5.5	3.9	2.3	5.2	3.7	2.1	3.5

### Oppgave

- a) Hvilke verdier har  $k$  og  $n$ ?
- b) Bestem øvre og nedre styregrenser for et  $\bar{X} - R$ -diagram.
- c) Tegn  $\bar{X} - R$ -diagrammet med  $\bar{X}$  og  $\bar{R}$  som senterlinjer.

### Løsningsforslag

- a)  $k = 20$ ,  $n = 5$
- b) Vi finner at  $\bar{X} = 124.81$  g og at  $\bar{R} = 3.425$  g. Fra tabell bak i boka finner vi videre at  $A_2 = 0.577$ ,  $D_3 = 0$  og  $D_4 = 2.114$ . Vi får derved følgende styregrenser:

$$LCL_{\bar{X}} = \bar{X} - A_2 \bar{R} = 124.81 \text{ g} - 0.577 \cdot 3.422 \text{ g} = 122.83 \text{ g}$$

$$UCL_{\bar{X}} = \bar{X} + A_2 \bar{R} = 124.81 \text{ g} + 0.577 \cdot 3.422 \text{ g} = 126.79 \text{ g}$$

$$LCL_R = D_3 \bar{R} = 0 \text{ g}$$

$$UCL_R = D_4 \bar{R} = 2.114 \cdot 3.425 \text{ g} = 7.24 \text{ g}$$

- c) Vi benytter Minitab til å plotte  $\bar{X}$ - $R$ -diagrammet. Kode og utskrift er vist nedenfor:

**MINITAB**

```
MTB > %Rxbar ;          # kommando for x-middel – R diagram
SUBC> Rsub c1-c5;      # enkeltobservasjonene ligger i kolonne c1-c5.
SUBC> Rbar;            # angir at styregrensene baseres på R middel
SUBC> Test 1.          # angir test 1 som teller antall overskridelser
Test Results for Xbar Chart
TEST 1. One point more than 3.00 sigmas from center line.
Test Failed at points: 19
```

De aktuelle verdiene i dette tilfellet er simulert fra en stabil prosess. Alarmen i  $x$ -middel-diagrammet er derfor falsk. Bortsett fra den ene alarmen er det heller ikke noe tydelig tegn på ustabilitet fra de to diagrammene.

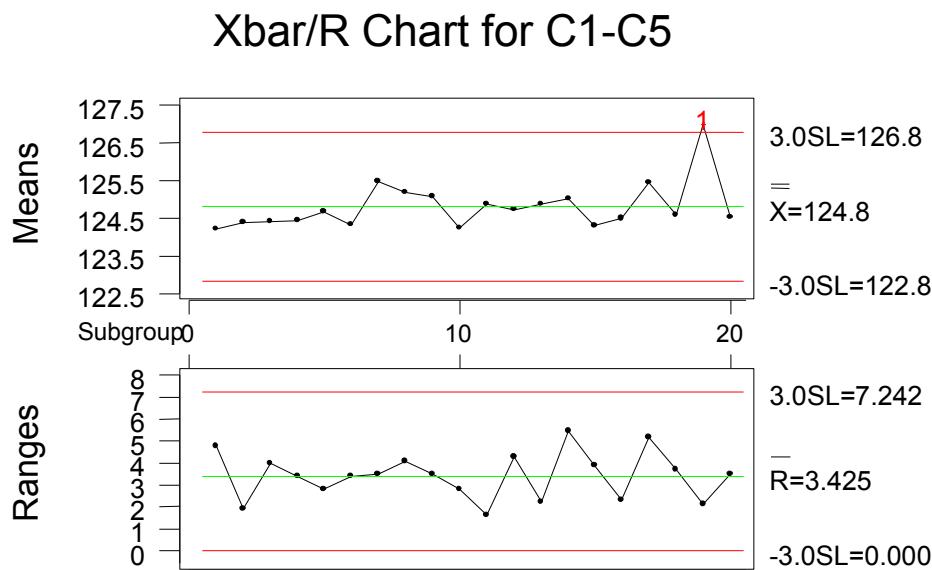


Fig. 12.4 Middelverdi-variasjonsbreddes diagram for eks. 12.1 (Minitab)

### 12.3 XmR-diagrammer

I en del tilfeller vil det av praktiske, fysiske, økonomiske eller andre grunner kun være aktuelt å foreta én enkelt observasjon ved hvert måletidspunkt ( $n = 1$ ). Et eksempel kan være temperaturmåling av væske i et rør. Vi har da ikke tilstrekkelig data til å få noe mål på spredningen ved hvert tidspunkt. Hvis prosessen er stabil kan vi da for eksempel behandle to påfølgende observasjoner som om de var foretatt ved samme tidspunkt for å bestemme den lokale variasjonsbredden.

For å fremheve den spesielle situasjonen med kun én enkelt observasjon pr. tidspunkt, benyttes forkortelsen *NP* for «Natural Process». «Tre sigma»-styregrensene i *X*-diagrammet betegnes tilsvarende *UNPL* og *LNPL* istedet for *UCL* og *LCL*, og er gitt som følger:

$$UNPL_X = \bar{X} + 2.660 \cdot \overline{mR}, \quad LNPL_X = \bar{X} - 2.660 \cdot \overline{mR}$$

For *R*-diagrammet tilsvarer situasjonen kap. 12.2 med  $n = 2$ , og vi får:

$$UCL_R = 3.268 \cdot \overline{mR}, \quad LCL_R = 0$$

Diagrammene som fremkommer som beskrevet ovenfor kalles *XmR*-diagrammer. *X*-en står her for at vi har enkeltobservasjoner, mens *mR* er en forkortelse for «moving range» som på norsk betyr glidende variasjonsbredde.

Selv om prosessen skulle være ustabil, kan fremgangsmåten ovenfor være fornuftig. Dette gjelder særlig når nivået og variasjonsmønsteret til prosessen varierer sakte i forhold til tidsdifferansen mellom to påfølgende måletidspunkt.

**Eks. 12.2 Frysetørket mat –  $XmR$ -diagram.**

*Oppgave*

Bruk Minitab til å lage et  $XmR$ -diagram med basis i de 20 første observasjonene som er gjengitt i tab. 12.1 for hvert måletidspunkt.

*Løsningsforslag*

I Minitab kalles  $XmR$ -diagram for «Imrchart», der I-en står for «individual». Nedenfor er vist kommandoer med forklaring, og diagrammet er vist deretter. Vi har ikke tatt med uskrift av tilbakemeldingen fra Minitab på den ene testen som er benyttet, da denne ikke gav alarm (ingen overskridelser).

**MINITAB**

```
MTB > %Imrchart c1;      # kommando for XmR-diagram, data i kolonne c1
SUBC> RSpan 2;          # Angir glidende variasjonsbredde for to på-
SUBC> Test 1.            # følgende observasjoner. Test 1: overskridelser
```

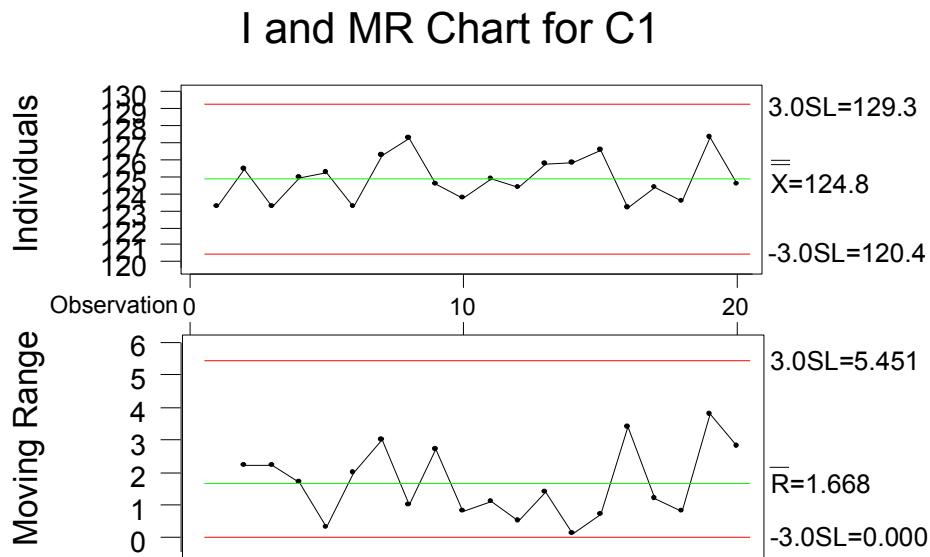


Fig. 12.5  $xmr$ -diagram for de 20 første målingene fra tab. 12.1 (Minitab)

## 12.4 Robusthet og myter

Vi har nevnt at styregrensene, dersom de bestemmes slik vi har beskrevet, er robuste med hensyn til blant annet hvilken fordeling våre prosessorvariabler måtte ha. De gode robusthetsegenskapene er såpass lite kjent at Wheeler og Chambers i sin bok (se litteraturliste) redegjør for 4 falske myter om Shewart-diagrammer. Vi skal kort se på disse, men først skal vi gjengi den såkalte «erfaringsregelen» («The empirical rule»):

**Erfaringsregelen** for homogent datasett:

- 1) Omtrent 60 % til 70 % av dataene vil være lokalisert innen en avstand på  $\pm$  ett standardavvik fra middelverdien.
- 2) Vanligvis vil 90 % til 98 % av dataene være lokalisert innen en avstand på  $\pm$  to standardavvik fra middelverdien.
- 3) Tilnærmet 99 % til 100 % av dataene vil være lokalisert innen en avstand på  $\pm$  tre standardavvik fra middelverdien.

Det er særlig den tredje delen av regelen ovenfor som har påvirket Shewarts valg av 3-standardavvik-grenser. Wheeler og Chambers gir i sin bok mange eksempler på at del 3) av regelen gjelder for fordelinger som avviker sterkt fra normalfordelingen, herunder svært skeive fordelinger. Så til mytene:

**Fire «falske» myter om Shewart-diagrammer**

- 1) Observasjonene må være normalfordelte før de kan bli plassert i et styrediagram.
- 2) Styrediagrammene virker på grunn av sentralgrenseteoremet.
- 3) Observasjonene må være uavhengige – data med autokorrelasjon passer ikke for styrediagrammer.
- 4) Prosessen må være under kontroll før data fra prosessen kan bli plottet i et styrediagram.

Myte 1 ovenfor kan «avlives» blant annet med basis i del 3 av erfaringssregelen (se ramme). Det har også som før nevnt vist seg i praksis at sannsynligheten for falsk alarm som regel er lite påvirket av avvik fra normalitetsforutsetningen.

Vedrørende myte 2, virker sentralgrenseteoremet kun på middelverdien og ikke på variasjonsbredden: Fordelingen av variasjonsbredde nærmer seg ikke mer og mer normalfordelingen jo større  $n$  er. Likevel fungerer  $R$ -diagrammene i praksis!

Myte 3 vedrører avhengighet mellom observasjonene. På tross av at påfølgende observasjoner fra en prosess som regel er autokorrelerte, har styrediagrammene vist seg å fungere i praksis. I de tilfeller der observasjonene er så sterkt korrelerte at de i alvorlig grad påvirker styregrensene, kan dette som oftest ses direkte av mønsteret til punktene i styrediagrammet. Selv om styregrensene er sterkt påvirket, vil derfor punktplottene i styrediagrammet i seg selv kunne være et nyttig verktøy.

Den siste myten sier at prosessen må være under kontroll før det er meningsfylt å plotte dataene i et styrediagram. Ifølge Wheeler og Chambers kan årsaken til denne myten være at  $\sigma$  feilaktig estimeres som standardavviket til alle observasjonene rundt den totale middelverdien. Det er **svært viktig** å huske at styregrensene baseres på «det **lokale**» standardavviket rundt middelverdien ved hvert enkelt tidspunkt.

## 12.5 To grunnregler

Uansett hvilken type Shewhart-diagrammer man skal lage, gjelder 4 grunnregler slik beskrevet i boka til Wheeler og Chambers. Vi skal her nøye oss med å gjengi de to første:

### To grunnregler for Shewhart styrediagrammer

- 1) Styregrensene i Shewhart styrediagrammer skal alltid settes i avstand  $\pm$  **tre standardavvik** fra senterlinjen.
- 2) Ved estimering av tre-sigma styregrenser må benyttet observator alltid bruke gjennomsnittet av de **lokale variasjonsmålene**.

## 12.6 Oppgaver

**12.1** Temperaturen  $X$  i 3 forskjellige lagertanker med væske som er regulert til å holde en temperatur på  $9.5^{\circ}\text{C}$ , måles ved 38 forskjellige tidspunkt. Ett og samme reguleringssystem regulerer temperaturen i de tre tankene. Gjennomsnittet av alle målingene viste en temperatur på  $9.47^{\circ}\text{C}$ . Midlere variasjonsbredde var  $0.60^{\circ}\text{C}$ .

- Bestem styregrensene for et  $\bar{X}$ -diagram.
- Bestem styregrensene for et  $R$ -diagram.
- Ved et senere tidspunkt ble gjennomsnittstemperaturen til 3 forskjellige tanker funnet å være  $10.2^{\circ}\text{C}$  mens variasjonsbredden var  $0.9^{\circ}\text{C}$ . Gir dette grunn til alarm?

**12.2** En fabrikk produserer blant annet elektriske motstander på  $10\ \Omega$ . For å undersøke om produksjonsprosessen er stabil, måles 5 og 5 motstander med jevne tidsmellomrom. Resultatene av målinger (i  $\Omega$ ) ved 20 forskjellige tidspunkt er gjengitt nedenfor.

tid	$X[\Omega]$						$\bar{X}$	$R$
1	9	9	5	9	4	7.2	5	
2	9	9	9	6	6	7.8	3	
3	10	9	16	14	17	11.6	7	
4	10	9	12	5	15	10.2	10	
5	9	9	6	9	6	7.8	3	
6	9	9	10	4	4	7.2	6	
7	10	9	8	20	11	11.6	12	
8	9	9	14	6	7	9.0	8	
9	9	9	8	9	10	9.0	2	
10	10	10	12	7	24	12.6	17	
tid	$X[\Omega]$						$\bar{X}$	$R$

11	10	10	19	5	18	12.4	14
12	9	9	10	7	7	8,	3
13	10	9	10	5	6	8,	5
14	10	9	8	14	10	10.2	6
15	10	9	5	4	7	7,	6
16	10	9	11	8	4	8,	7
17	9	9	16	17	9	12.0	8
18	9	9	10	6	7	8,	4
19	9	9	15	7	11	10.2	8
20	10	9	16	10	8	10,6	8

- En av  $X$ -verdiene ved tidspunkt 3 er feil. Finn denne og angi rett verdi.
- Angi styregrensene og konstruer et  $\bar{X}$ - $R$  diagram.
- Gir diagrammet i b) grunnlag for å undersøke prosessen nærmere?

**12.3** En bedrift produserer en vare der hver enhet skal veie 1 kg. Det har vist seg at den lokale variasjonen ved et gitt tidspunkt er neglisjerbar, men at nivået kan variere fra dag til dag. For å overvåke prosessen, måler man vekten av en vareenhet pr. dag over en periode på 20 dager, og får følgende avvik fra et kg (angitt i gram, suksessive tidspunkt fra venstre mot høyre):

13 0 -17 3 18 -1 -1 -12 23 16  
-16 7 -3 1 9 3 7 16 -12 4

- Konstruer et  $XmR$ -diagram.
- På dag 21 måles en ny verdi på 1.040 kg. Gir denne verdien grunnlag for alarm?

## 12.7 Formelsamling og tabell

### Betegnelser

$X_{ij}$  = prosessvariabel, observasjon nr.  $j$  ved tidspunkt  $i$ ,

$$i = 1, \dots, k; j = 1, \dots, n$$

$k$  = antall ekvidistante tidspunkt

$n$  = antall uavhengige observasjoner ved tidspunkt  $i$

$\mu$  =  $E(X_{ij})$  (ved stabil prosess)

$\sigma$  =  $\text{std}(X_{ij})$  (ved stabil prosess)

$s_i$  = empirisk standardavvik ved tidspunkt  $i$ .

$R_i$  = variasjonsbredde ved tidspunkt

$$i: R_i = \text{Max}(X_{ij}) - \text{Min}(X_{ij}),$$

$$j = 1, \dots, n.$$

$LCL$  = «Lower Control Limit» =

Nedre styregrense

$UCL$  = «Upper Control Limit» =

Øvre styregrense

$$\bar{X}_i = \frac{1}{n} \sum_{j=1}^n X_{ij}$$

$$\bar{\bar{X}} = \frac{1}{k} \sum_{i=1}^k \bar{X}_i$$

$$\bar{R} = \frac{1}{k} \sum_{i=1}^k R_i$$

### Styregrenser i Shewart $\bar{x}$ -diagram:

$$LCL_{\bar{X}} = \bar{\bar{X}} - A_2 \cdot \bar{R}$$

$$UCL_{\bar{X}} = \bar{\bar{X}} + A_2 \cdot \bar{R}$$

der konstanten  $A_2$  kun avhenger av  $n$  (se tabell). For  $n = 5$  er  $A_2 = 0.577$ .

### Styregrenser i Shewart $R$ -diagram:

$$LCL_R = D_3 \cdot \bar{R}$$

$$UCL_R = D_4 \cdot \bar{R}$$

der konstantene  $D_3$  og  $D_4$  kun avhenger av  $n$  (se tabell). For  $n \leq 6$  så er  $D_3 = 0$ .

### Styregrenser i $XmR$ -diagram:

$XmR$ -diagram står for diagrammer der vi kun har én enkeltobservasjon pr. måletidspunkt. Variasjonsbredden bestemmes med basis i to påfølgende observasjoner. Styregrensene blir:

$$UNPL_X = \bar{X} + 2.660 \cdot \bar{mR}$$

$$LNPL_X = \bar{X} - 2.660 \cdot \bar{mR}$$

$$UCL_R = 3.268 \cdot \bar{mR}, LCL_R = 0$$

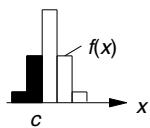
### Minimumskrav til undersøkelse av om prosessen er stabil:

$k > \text{ca. } 20$  (tommelfingerregel)

Tabell: Faktorer for bruk av  $\bar{R}$  til beregning av styregrensene.

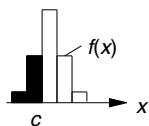
$n$	$A_2$	$D_3$	$D_4$
2	1.880	-	3.268
3	1.023	-	2.574
4	0.729	-	2.282
5	0.577	-	2.114
6	0.483	-	2.004
7	0.419	0.076	1.924
8	0.373	0.136	1.864
9	0.337	0.184	1.816
10	0.308	0.223	1.777

Kumulativ **binomisk** sannsynlighet  $P(X \leq c)$

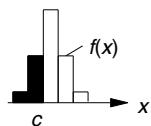


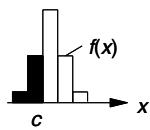
		<i>p</i>										
<i>c</i>		.05	.1	.2	.3	.4	.5	.6	.7	.8	.9	.95
<i>n</i> = 1	0	.950	.900	.800	.700	.600	.500	.400	.300	.200	.100	.050
	1	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
<i>n</i> = 2	0	.903	.810	.640	.490	.360	.250	.160	.090	.040	.010	.003
	1	.998	.990	.960	.910	.840	.750	.640	.510	.360	.190	.098
<i>n</i> = 3	0	.857	.729	.512	.343	.216	.125	.064	.027	.008	.001	.000
	1	.993	.972	.896	.784	.648	.500	.352	.216	.104	.028	.007
<i>n</i> = 4	0	.815	.656	.410	.240	.130	.062	.026	.008	.002	.000	.000
	1	.986	.948	.819	.652	.475	.313	.179	.084	.027	.004	.000
<i>n</i> = 5	0	.774	.590	.328	.168	.078	.031	.010	.002	.000	.000	.000
	1	.977	.919	.737	.528	.337	.188	.087	.031	.007	.000	.000
<i>n</i> = 6	0	.735	.531	.262	.118	.047	.016	.004	.001	.000	.000	.000
	1	.967	.886	.655	.420	.233	.109	.041	.011	.002	.000	.000
<i>n</i> = 7	0	.698	.478	.210	.082	.028	.008	.002	.000	.000	.000	.000
	1	.956	.850	.577	.329	.159	.063	.019	.004	.000	.000	.000
<i>n</i> = 8	0	.663	.430	.168	.058	.017	.004	.001	.000	.000	.000	.000
	1	.943	.813	.503	.255	.106	.035	.009	.001	.000	.000	.000

Kumulativ **binomisk** sannsynlighet  $P(X \leq c)$



Kumulativ **binomisk** sannsynlighet  $P(X \leq c)$

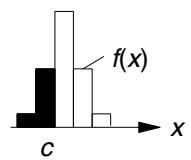


Kumulativ **binomisk** sannsynlighet  $P(X \leq c)$ 

		<i>p</i>										
		.05	.1	.2	.3	.4	.5	.6	.7	.8	.9	.95
<i>n</i> = 16		c										
<i>n</i> = 16		0	.440	.185	.028	.003	.000	.000	.000	.000	.000	.000
		1	.811	.515	.141	.026	.003	.000	.000	.000	.000	.000
		2	.957	.789	.352	.099	.018	.002	.000	.000	.000	.000
		3	.993	.932	.598	.246	.065	.011	.001	.000	.000	.000
		4	.999	.983	.798	.450	.167	.038	.005	.000	.000	.000
		5	1.000	.997	.918	.660	.329	.105	.019	.002	.000	.000
		6	1.000	.999	.973	.825	.527	.227	.058	.007	.000	.000
		7	1.000	1.000	.993	.926	.716	.402	.142	.026	.001	.000
		8	1.000	1.000	.999	.974	.858	.598	.284	.074	.007	.000
		9	1.000	1.000	1.000	.993	.942	.773	.473	.175	.027	.001
		10	1.000	1.000	1.000	.998	.981	.895	.671	.340	.082	.003
		11	1.000	1.000	1.000	1.000	.995	.962	.833	.550	.202	.017
		12	1.000	1.000	1.000	1.000	.999	.989	.935	.754	.402	.068
		13	1.000	1.000	1.000	1.000	1.000	.998	.982	.901	.648	.211
		14	1.000	1.000	1.000	1.000	1.000	1.000	.997	.974	.859	.485
		15	1.000	1.000	1.000	1.000	1.000	1.000	1.000	.997	.972	.815
		16	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	.560
<i>n</i> = 17		0	.418	.167	.023	.002	.000	.000	.000	.000	.000	.000
		1	.792	.482	.118	.019	.002	.000	.000	.000	.000	.000
		2	.950	.762	.310	.077	.012	.001	.000	.000	.000	.000
		3	.991	.917	.549	.202	.046	.006	.000	.000	.000	.000
		4	.999	.978	.758	.389	.126	.025	.003	.000	.000	.000
		5	1.000	.995	.894	.597	.264	.072	.011	.001	.000	.000
		6	1.000	.999	.962	.775	.448	.166	.035	.003	.000	.000
		7	1.000	1.000	.989	.895	.641	.315	.092	.013	.000	.000
		8	1.000	1.000	.997	.960	.801	.500	.199	.040	.003	.000
		9	1.000	1.000	1.000	.987	.908	.685	.359	.105	.011	.000
		10	1.000	1.000	1.000	.997	.965	.834	.552	.225	.038	.001
		11	1.000	1.000	1.000	.999	.989	.928	.736	.403	.106	.005
		12	1.000	1.000	1.000	1.000	.997	.975	.874	.611	.242	.022
		13	1.000	1.000	1.000	1.000	1.000	.994	.954	.798	.451	.083
		14	1.000	1.000	1.000	1.000	1.000	.999	.988	.923	.690	.238
		15	1.000	1.000	1.000	1.000	1.000	1.000	.998	.981	.882	.518
		16	1.000	1.000	1.000	1.000	1.000	1.000	1.000	.998	.977	.833
		17	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	.582
<i>n</i> = 18		0	.397	.150	.018	.002	.000	.000	.000	.000	.000	.000
		1	.774	.450	.099	.014	.001	.000	.000	.000	.000	.000
		2	.942	.734	.271	.060	.008	.001	.000	.000	.000	.000
		3	.989	.902	.501	.165	.033	.004	.000	.000	.000	.000
		4	.998	.972	.716	.333	.094	.015	.001	.000	.000	.000
		5	1.000	.994	.867	.534	.209	.048	.006	.000	.000	.000
		6	1.000	.999	.949	.722	.374	.119	.020	.001	.000	.000
		7	1.000	1.000	.984	.859	.563	.240	.058	.006	.000	.000
		8	1.000	1.000	.996	.940	.737	.407	.135	.021	.001	.000
		9	1.000	1.000	.999	.979	.865	.593	.263	.060	.004	.000
		10	1.000	1.000	1.000	.994	.942	.760	.437	.141	.016	.000
		11	1.000	1.000	1.000	.999	.980	.881	.626	.278	.051	.001
		12	1.000	1.000	1.000	1.000	.994	.952	.791	.466	.133	.006

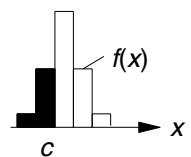


Kumulativ **Poisson** sannsynlighet  $P(X \leq c)$



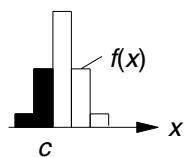
$c \downarrow \lambda \rightarrow$	.1	.2	.3	.4	.5	.6	.7	.8	.9	1
<b>0</b>	.905	.819	.741	.670	.607	.549	.497	.449	.407	.368
<b>1</b>	.995	.982	.963	.938	.910	.878	.844	.809	.772	.736
<b>2</b>	1.000	.999	.996	.992	.986	.977	.966	.953	.937	.920
<b>3</b>	1.000	1.000	1.000	.999	.998	.997	.994	.991	.987	.981
<b>4</b>	1.000	1.000	1.000	1.000	1.000	1.000	.999	.999	.998	.996
<b>5</b>	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	.999
<b>6</b>	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
$c \downarrow \lambda \rightarrow$	1.1	1.2	1.3	1.4	1.5	1.6	1.7	1.8	1.9	2
<b>0</b>	.333	.301	.273	.247	.223	.202	.183	.165	.150	.135
<b>1</b>	.699	.663	.627	.592	.558	.525	.493	.463	.434	.406
<b>2</b>	.900	.879	.857	.833	.809	.783	.757	.731	.704	.677
<b>3</b>	.974	.966	.957	.946	.934	.921	.907	.891	.875	.857
<b>4</b>	.995	.992	.989	.986	.981	.976	.970	.964	.956	.947
<b>5</b>	.999	.998	.998	.997	.996	.994	.992	.990	.987	.983
<b>6</b>	1.000	1.000	1.000	.999	.999	.999	.998	.997	.997	.995
<b>7</b>	1.000	1.000	1.000	1.000	1.000	1.000	1.000	.999	.999	.999
<b>8</b>	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
$c \downarrow \lambda \rightarrow$	2.1	2.2	2.3	2.4	2.5	2.6	2.7	2.8	2.9	3
<b>0</b>	.122	.111	.100	.091	.082	.074	.067	.061	.055	.050
<b>1</b>	.380	.355	.331	.308	.287	.267	.249	.231	.215	.199
<b>2</b>	.650	.623	.596	.570	.544	.518	.494	.469	.446	.423
<b>3</b>	.839	.819	.799	.779	.758	.736	.714	.692	.670	.647
<b>4</b>	.938	.928	.916	.904	.891	.877	.863	.848	.832	.815
<b>5</b>	.980	.975	.970	.964	.958	.951	.943	.935	.926	.916
<b>6</b>	.994	.993	.991	.988	.986	.983	.979	.976	.971	.966
<b>7</b>	.999	.998	.997	.997	.996	.995	.993	.992	.990	.988
<b>8</b>	1.000	1.000	.999	.999	.999	.999	.998	.998	.997	.996
<b>9</b>	1.000	1.000	1.000	1.000	1.000	1.000	.999	.999	.999	.999
<b>10</b>	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
$c \downarrow \lambda \rightarrow$	3.1	3.2	3.3	3.4	3.5	3.6	3.7	3.8	3.9	4
<b>0</b>	.045	.041	.037	.033	.030	.027	.025	.022	.020	.018
<b>1</b>	.185	.171	.159	.147	.136	.126	.116	.107	.099	.092
<b>2</b>	.401	.380	.359	.340	.321	.303	.285	.269	.253	.238
<b>3</b>	.625	.603	.580	.558	.537	.515	.494	.473	.453	.433
<b>4</b>	.798	.781	.763	.744	.725	.706	.687	.668	.648	.629
<b>5</b>	.906	.895	.883	.871	.858	.844	.830	.816	.801	.785
<b>6</b>	.961	.955	.949	.942	.935	.927	.918	.909	.899	.889
<b>7</b>	.986	.983	.980	.977	.973	.969	.965	.960	.955	.949
<b>8</b>	.995	.994	.993	.992	.990	.988	.986	.984	.981	.979
<b>9</b>	.999	.998	.998	.997	.997	.996	.995	.994	.993	.992
<b>10</b>	1.000	1.000	.999	.999	.999	.999	.998	.998	.998	.997

Kumulativ **Poisson** sannsynlighet  $P(X \leq c)$



$c \downarrow \lambda \rightarrow$	3.1	3.2	3.3	3.4	3.5	3.6	3.7	3.8	3.9	4
11	1.000	1.000	1.000	1.000	1.000	1.000	1.000	.999	.999	.999
12	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
$c \downarrow \lambda \rightarrow$	4.1	4.2	4.3	4.4	4.5	4.6	4.7	4.8	4.9	5
0	.017	.015	.014	.012	.011	.010	.009	.008	.007	.007
1	.085	.078	.072	.066	.061	.056	.052	.048	.044	.040
2	.224	.210	.197	.185	.174	.163	.152	.143	.133	.125
3	.414	.395	.377	.359	.342	.326	.310	.294	.279	.265
4	.609	.590	.570	.551	.532	.513	.495	.476	.458	.440
5	.769	.753	.737	.720	.703	.686	.668	.651	.634	.616
6	.879	.867	.856	.844	.831	.818	.805	.791	.777	.762
7	.943	.936	.929	.921	.913	.905	.896	.887	.877	.867
8	.976	.972	.968	.964	.960	.955	.950	.944	.938	.932
9	.990	.989	.987	.985	.983	.980	.978	.975	.972	.968
10	.997	.996	.995	.994	.993	.992	.991	.990	.988	.986
11	.999	.999	.998	.998	.998	.997	.997	.996	.995	.995
12	1.000	1.000	.999	.999	.999	.999	.999	.999	.998	.998
13	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	.999	.999
14	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
$c \downarrow \lambda \rightarrow$	5.1	5.2	5.3	5.4	5.5	5.6	5.7	5.8	5.9	6
0	.006	.006	.005	.005	.004	.004	.003	.003	.003	.002
1	.037	.034	.031	.029	.027	.024	.022	.021	.019	.017
2	.116	.109	.102	.095	.088	.082	.077	.072	.067	.062
3	.251	.238	.225	.213	.202	.191	.180	.170	.160	.151
4	.423	.406	.390	.373	.358	.342	.327	.313	.299	.285
5	.598	.581	.563	.546	.529	.512	.495	.478	.462	.446
6	.747	.732	.717	.702	.686	.670	.654	.638	.622	.606
7	.856	.845	.833	.822	.809	.797	.784	.771	.758	.744
8	.925	.918	.911	.903	.894	.886	.877	.867	.857	.847
9	.964	.960	.956	.951	.946	.941	.935	.929	.923	.916
10	.984	.982	.980	.977	.975	.972	.969	.965	.961	.957
11	.994	.993	.992	.990	.989	.988	.986	.984	.982	.980
12	.998	.997	.997	.996	.996	.995	.994	.993	.992	.991
13	.999	.999	.999	.999	.998	.998	.998	.997	.997	.996
14	1.000	1.000	1.000	1.000	.999	.999	.999	.999	.999	.999
15	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	.999
16	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
$c \downarrow \lambda \rightarrow$	6.1	6.2	6.3	6.4	6.5	6.6	6.7	6.8	6.9	7
0	.002	.002	.002	.002	.002	.001	.001	.001	.001	.001
1	.016	.015	.013	.012	.011	.010	.009	.009	.008	.007
2	.058	.054	.050	.046	.043	.040	.037	.034	.032	.030
3	.143	.134	.126	.119	.112	.105	.099	.093	.087	.082
4	.272	.259	.247	.235	.224	.213	.202	.192	.182	.173

Kumulativ **Poisson** sannsynlighet  $P(X \leq c)$

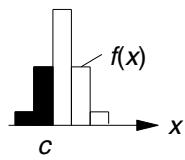


$c \downarrow \lambda \rightarrow$	6.1	6.2	6.3	6.4	6.5	6.6	6.7	6.8	6.9	7
5	.430	.414	.399	.384	.369	.355	.341	.327	.314	.301
6	.590	.574	.558	.542	.527	.511	.495	.480	.465	.450
7	.730	.716	.702	.687	.673	.658	.643	.628	.614	.599
8	.837	.826	.815	.803	.792	.780	.767	.755	.742	.729
9	.909	.902	.894	.886	.877	.869	.860	.850	.840	.830
10	.953	.949	.944	.939	.933	.927	.921	.915	.908	.901
11	.978	.975	.972	.969	.966	.963	.959	.955	.951	.947
12	.990	.989	.987	.986	.984	.982	.980	.978	.976	.973
13	.996	.995	.995	.994	.993	.992	.991	.990	.989	.987
14	.998	.998	.998	.997	.997	.997	.996	.996	.995	.994
15	.999	.999	.999	.999	.999	.999	.998	.998	.998	.998
16	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
17	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

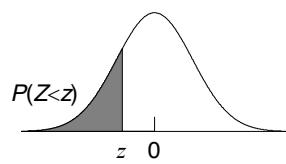
$c \downarrow \lambda \rightarrow$	7.1	7.2	7.3	7.4	7.5	7.6	7.7	7.8	7.9	8
0	.001	.001	.001	.001	.001	.001	.000	.000	.000	.000
1	.007	.006	.006	.005	.005	.004	.004	.004	.003	.003
2	.027	.025	.024	.022	.020	.019	.017	.016	.015	.014
3	.077	.072	.067	.063	.059	.055	.052	.048	.045	.042
4	.164	.156	.147	.140	.132	.125	.118	.112	.106	.100
5	.288	.276	.264	.253	.241	.231	.220	.210	.201	.191
6	.435	.420	.406	.392	.378	.365	.351	.338	.326	.313
7	.584	.569	.554	.539	.525	.510	.496	.481	.467	.453
8	.716	.703	.689	.676	.662	.648	.634	.620	.607	.593
9	.820	.810	.799	.788	.776	.765	.753	.741	.729	.717
10	.894	.887	.879	.871	.862	.854	.845	.835	.826	.816
11	.942	.937	.932	.926	.921	.915	.909	.902	.895	.888
12	.970	.967	.964	.961	.957	.954	.950	.945	.941	.936
13	.986	.984	.982	.980	.978	.976	.974	.971	.969	.966
14	.994	.993	.992	.991	.990	.989	.987	.986	.984	.983
15	.997	.997	.996	.996	.995	.995	.994	.993	.993	.992
16	.999	.999	.999	.998	.998	.998	.997	.997	.997	.996
17	1.000	1.000	.999	.999	.999	.999	.999	.999	.999	.998
18	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	.999	.999
19	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

$c \downarrow \lambda \rightarrow$	8.1	8.2	8.3	8.4	8.5	8.6	8.7	8.8	8.9	9
0	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
1	.003	.003	.002	.002	.002	.002	.002	.001	.001	.001
2	.013	.012	.011	.010	.009	.009	.008	.007	.007	.006
3	.040	.037	.035	.032	.030	.028	.026	.024	.023	.021
4	.094	.089	.084	.079	.074	.070	.066	.062	.058	.055
$c \downarrow \lambda \rightarrow$	8.1	8.2	8.3	8.4	8.5	8.6	8.7	8.8	8.9	9

Kumulativ **Poisson** sannsynlighet  $P(X \leq c)$

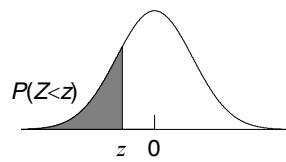


Kumulativ  $N(0,1)$ -tabell:  $\Phi(z) = P(Z < z)$

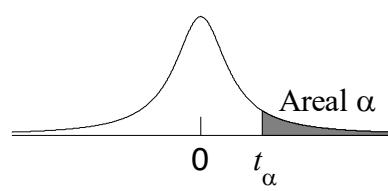


$z$	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
-3.5	.0002	.0002	.0002	.0002	.0002	.0002	.0002	.0002	.0002	.0002
-3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002
-3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003
-3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
-3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007
-3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
-2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
-2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
-2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
-2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
-2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
-2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
-2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
-2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
-2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
-2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
-1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
-1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
-1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
-1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
-1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
-1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681
-1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
-1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
-1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
-1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
-.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
-.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
-.7	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148
-.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
-.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
-.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
-.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483
-.2	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936	.3897	.3859
-.1	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4286	.4247
-.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641

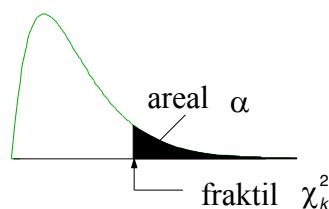
Kumulativ  $N(0,1)$ -tabell:  $\Phi(z) = P(Z < z)$



***t*-fraktiler,  $t_\alpha$ , i  
*t*-fordelingen med  $m$  frihetsgrader**

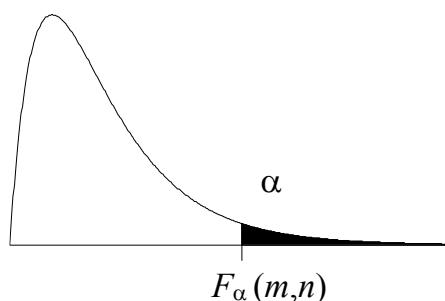


$m$	$\alpha$					
	.25	.10	.05	.025	.01	.005
1	1.000	3.078	6.314	12.706	31.821	63.657
2	.816	1.886	2.920	4.303	6.965	9.925
3	.765	1.638	2.353	3.182	4.541	5.841
4	.741	1.533	2.132	2.776	3.747	4.604
5	.727	1.476	2.015	2.571	3.365	4.032
6	.718	1.440	1.943	2.447	3.143	3.707
7	.711	1.415	1.895	2.365	2.998	3.499
8	.706	1.397	1.860	2.306	2.896	3.355
9	.703	1.383	1.833	2.262	2.821	3.250
10	.700	1.372	1.812	2.228	2.764	3.169
11	.697	1.363	1.796	2.201	2.718	3.106
12	.695	1.356	1.782	2.179	2.681	3.055
13	.694	1.350	1.771	2.160	2.650	3.012
14	.692	1.345	1.761	2.145	2.624	2.977
15	.691	1.341	1.753	2.131	2.602	2.947
16	.690	1.337	1.746	2.120	2.583	2.921
17	.689	1.333	1.740	2.110	2.567	2.898
18	.688	1.330	1.734	2.101	2.552	2.878
19	.688	1.328	1.729	2.093	2.539	2.861
20	.687	1.325	1.725	2.086	2.528	2.845
21	.686	1.323	1.721	2.080	2.518	2.831
22	.686	1.321	1.717	2.074	2.508	2.819
23	.685	1.319	1.714	2.069	2.500	2.807
24	.685	1.318	1.711	2.064	2.492	2.797
25	.684	1.316	1.708	2.060	2.485	2.787
26	.684	1.315	1.706	2.056	2.479	2.779
27	.684	1.314	1.703	2.052	2.473	2.771
28	.683	1.313	1.701	2.048	2.467	2.763
29	.683	1.311	1.699	2.045	2.462	2.756
30	.683	1.310	1.697	2.042	2.457	2.750
40	.681	1.303	1.684	2.021	2.423	2.704
60	.679	1.296	1.671	2.000	2.390	2.660
120	.677	1.289	1.658	1.980	2.358	2.617
$\infty$	.674	1.282	1.645	1.960	2.326	2.576



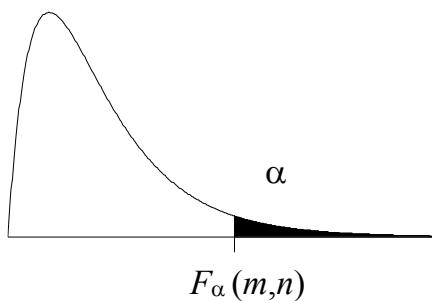
### $\alpha$ -fraktiler $\chi_{\alpha}^2$ i kjikvadratfordelingen med $k$ frihetsgrader

$k \downarrow \alpha \rightarrow$	.995	.990	.975	.950	.050	.025	.010	.005
1	0.00	0.00	0.00	0.00	3.84	5.02	6.63	7.88
2	0.01	0.02	0.05	0.10	5.99	7.38	9.21	10.60
3	0.07	0.11	0.22	0.35	7.81	9.35	11.34	12.84
4	0.21	0.30	0.48	0.71	9.49	11.14	13.28	14.86
5	0.41	0.55	0.83	1.15	11.07	12.83	15.09	16.75
6	0.68	0.87	1.24	1.64	12.59	14.45	16.81	18.55
7	0.99	1.24	1.69	2.17	14.07	16.01	18.48	20.28
8	1.34	1.65	2.18	2.73	15.51	17.53	20.09	21.95
9	1.73	2.09	2.70	3.33	16.92	19.02	21.67	23.59
10	2.16	2.56	3.25	3.94	18.31	20.48	23.21	25.19
11	2.60	3.05	3.82	4.57	19.68	21.92	24.72	26.76
12	3.07	3.57	4.40	5.23	21.03	23.34	26.22	28.30
13	3.57	4.11	5.01	5.89	22.36	24.74	27.69	29.82
14	4.07	4.66	5.63	6.57	23.68	26.12	29.14	31.32
15	4.60	5.23	6.26	7.26	25.00	27.49	30.58	32.80
16	5.14	5.81	6.91	7.96	26.30	28.85	32.00	34.27
17	5.70	6.41	7.56	8.67	27.59	30.19	33.41	35.72
18	6.26	7.01	8.23	9.39	28.87	31.53	34.81	37.16
19	6.84	7.63	8.91	10.12	30.14	32.85	36.19	38.58
20	7.43	8.26	9.59	10.85	31.41	34.17	37.57	40.00
21	8.03	8.90	10.28	11.59	32.67	35.48	38.93	41.40
22	8.64	9.54	10.98	12.34	33.92	36.78	40.29	42.80
23	9.26	10.20	11.69	13.09	35.17	38.08	41.64	44.18
24	9.89	10.86	12.40	13.85	36.42	39.36	42.98	45.56
25	10.52	11.52	13.12	14.61	37.65	40.65	44.31	46.93
26	11.16	12.20	13.84	15.38	38.89	41.92	45.64	48.29
27	11.81	12.88	14.57	16.15	40.11	43.19	46.96	49.64
28	12.46	13.56	15.31	16.93	41.34	44.46	48.28	50.99
29	13.12	14.26	16.05	17.71	42.56	45.72	49.59	52.34
30	13.79	14.95	16.79	18.49	43.77	46.98	50.89	53.67
40	20.71	22.16	24.43	26.51	55.76	59.34	63.69	66.77
50	27.99	29.71	32.36	34.76	67.50	71.42	76.15	79.49
60	35.53	37.48	40.48	43.19	79.08	83.30	88.38	91.95
70	43.28	45.44	48.76	51.74	90.53	95.02	100.43	104.21
80	51.17	53.54	57.15	60.39	101.88	106.63	112.33	116.32
90	59.20	61.75	65.65	69.13	113.15	118.14	124.12	128.30
100	67.33	70.06	74.22	77.93	124.34	129.56	135.81	140.17



**Øvre 5-prosentiler F-fordelingen m: d.f. i teller, n: d.f. i nevner**

$m \rightarrow$ $n \downarrow$	1	2	3	4	5	6	7	8	9
1	161,45	199,50	215,71	224,58	230,16	233,99	236,77	238,88	240,54
2	18,513	19,000	19,164	19,247	19,296	19,330	19,353	19,371	19,385
3	10,128	9,552	9,277	9,117	9,013	8,941	8,887	8,845	8,812
4	7,709	6,944	6,591	6,388	6,256	6,163	6,094	6,041	5,999
5	6,608	5,786	5,409	5,192	5,050	4,950	4,876	4,818	4,772
6	5,987	5,143	4,757	4,534	4,387	4,284	4,207	4,147	4,099
7	5,591	4,737	4,347	4,120	3,972	3,866	3,787	3,726	3,677
8	5,318	4,459	4,066	3,838	3,687	3,581	3,500	3,438	3,388
9	5,117	4,256	3,863	3,633	3,482	3,374	3,293	3,230	3,179
10	4,965	4,103	3,708	3,478	3,326	3,217	3,135	3,072	3,020
11	4,844	3,982	3,587	3,357	3,204	3,095	3,012	2,948	2,896
12	4,747	3,885	3,490	3,259	3,106	2,996	2,913	2,849	2,796
13	4,667	3,806	3,411	3,179	3,025	2,915	2,832	2,767	2,714
14	4,600	3,739	3,344	3,112	2,958	2,848	2,764	2,699	2,646
15	4,543	3,682	3,287	3,056	2,901	2,790	2,707	2,641	2,588
16	4,494	3,634	3,239	3,007	2,852	2,741	2,657	2,591	2,538
17	4,451	3,592	3,197	2,965	2,810	2,699	2,614	2,548	2,494
18	4,414	3,555	3,160	2,928	2,773	2,661	2,577	2,510	2,456
19	4,381	3,522	3,127	2,895	2,740	2,628	2,544	2,477	2,423
20	4,351	3,493	3,098	2,866	2,711	2,599	2,514	2,447	2,393
30	4,171	3,316	2,922	2,690	2,534	2,421	2,334	2,266	2,211
60	4,001	3,150	2,758	2,525	2,368	2,254	2,167	2,097	2,040
120	3,920	3,072	2,680	2,447	2,290	2,175	2,087	2,016	1,959
$\infty$	3,842	3,000	2,605	2,372	2,214	2,099	2,010	1,938	1,880



**Øvre 5-prosentiler F-fordelingen m: d.f. i teller. n: d.f. i nevner**

$m \rightarrow$ $n \downarrow$	10	12	15	20	24	30	60	120	$\infty$
1	241,88	243,91	245,95	248,01	249,05	250,10	252,20	253,25	254,32
2	19,396	19,413	19,429	19,446	19,454	19,462	19,479	19,487	14,496
3	8,786	8,745	8,703	8,660	8,639	8,617	8,572	8,549	8,527
4	5,964	5,912	5,858	5,803	5,774	5,746	5,688	5,658	5,628
5	4,735	4,678	4,619	4,558	4,527	4,496	4,431	4,398	4,365
6	4,060	4,000	3,938	3,874	3,841	3,808	3,740	3,705	3,669
7	3,637	3,575	3,511	3,445	3,410	3,376	3,304	3,267	3,230
8	3,347	3,284	3,218	3,150	3,115	3,079	3,005	2,967	2,928
9	3,137	3,073	3,006	2,936	2,900	2,864	2,787	2,748	2,707
10	2,978	2,913	2,845	2,774	2,737	2,700	2,621	2,580	2,538
11	2,854	2,788	2,719	2,646	2,609	2,570	2,490	2,448	2,405
12	2,753	2,687	2,617	2,544	2,505	2,466	2,384	2,341	2,296
13	2,671	2,604	2,533	2,459	2,420	2,380	2,297	2,252	2,206
14	2,602	2,534	2,463	2,388	2,349	2,308	2,223	2,178	2,131
15	2,544	2,475	2,403	2,328	2,288	2,247	2,160	2,114	2,066
16	2,494	2,425	2,352	2,276	2,235	2,194	2,106	2,059	2,010
17	2,450	2,381	2,308	2,230	2,190	2,148	2,058	2,011	1,960
18	2,412	2,342	2,269	2,191	2,150	2,107	2,017	1,968	1,917
19	2,378	2,308	2,234	2,155	2,114	2,071	1,980	1,930	1,878
20	2,348	2,278	2,203	2,124	2,082	2,039	1,946	1,896	1,843
30	2,165	2,092	2,015	1,932	1,887	1,841	1,740	1,683	1,622
60	1,993	1,917	1,836	1,748	1,700	1,649	1,534	1,467	1,389
120	1,910	1,834	1,750	1,659	1,608	1,554	1,429	1,352	1,254
$\infty$	1,831	1,752	1,666	1,571	1,517	1,459	1,318	1,221	1,000

## Fasit

De fleste svar er beregnet med Matlab, hvilket gir neglisjerbar avrundingsfeil. Forøvrig er det benyttet halvkorreksjon i de oppgaver der det er aktuelt.

- 1.1** a)  $\bar{x} = 5,40, s = 5,32$   
b)  $\bar{x} = -8,40, s = 5,22$
- 1.2** a) -2 3 2 0 b),75 2,22  
c) 1,4344075 0,0000222
- 1.3**  $\bar{x} = 1221,14 s = 0,36$
- 1.4** 2 212 000
- 1.5** a)  $\bar{x} = 5 m = 7$   
b)  $\bar{x} = 22,8 m = 10,5$   
Ville stolt mest på  $m$ .
- 1.6** -1 7
- 1.7**  $s = 19,4 Q_3 - Q_1 = 4$
- 1.8** 9,5..., 19,499..., 14,5..., 10
- 1.9** a)  $m_i$ : 5k 15k 25k 35k b) 20480  
c) 8734 d) 19831 e) 13550
- 1.10** b),949 c)  $-0,016 + 1,06x$
- 1.11** a),356 b)  $-4,834 + 16,40 \cdot x$   
d) x/y-akse: 30/1 cm/enhet
- 1.12** a)  $r = ,95$  c)  $r = ,99967$   
d)  $z = 4,10 + 0,69x$   
e)  $y = 60,4 \cdot \exp(0,69x)$
- 1.13** a)  $\bar{x} = 68,2 s = 10,9$   
b)  $m = 67 Q_3 - Q_1 = 17$
- 1.14** b),784 c)  $r = 0,11$
- 1.15** a)  $\bar{x} = 2,3$  b)  $s = 0,979$ ,  
d)  $m = 2,5 Q_3 - Q_1 = 1$
- 1.16** b) 38,57 11,25 37,50
- 1.17** b) 211,6 6,5 c) 211,8 9,7
- 2.1**  $S = \{K, B, I\}$
- 2.2**  $S = \{KK, KB, KI, BI, II\}$
- 2.3** a) 6 b) 6 c) 6
- 2.4** 126
- 2.5** 362 880
- 2.6** a), b) 5-er,  $p = 1/6$   
c), d) ethvert utfall,  $p = 1$
- 2.8** b),6 ,8 ,4 ,6 ,2 ,8
- 2.9** a)  $2^5 = 32$  (potensregel)  
b)  $6^3 = 216$  (potensregel)
- c) 210 (ordningsregel)  
d) 3 hvis uordna fargerekkefølge,  
7 hvis ordna e)  $6,4 \cdot 10^{11}$
- 2.10** a),4 b),1 c),25 d) 161/330
- 2.11** a) 9 b) 5
- 2.12** a) 20 190 1140 b) 18145
- 2.13** a) 26/51 b) 201/221
- 2.14** a) 0,6 b) 10/357
- 2.15** a)  $6,3 \cdot 10^{-12}$  b) 0,0036,  
c)  $8,8 \cdot 10^{-10}$
- 2.16** 19/27
- 2.17** 9
- 2.18** a),042 b),95
- 2.19** a) 1/300 b) 34/75 c) 7/75  
d) 259/300
- 2.20** R G G G R G G
- 2.21** a),252 b),748 c),559
- 2.22** a) 120 b) 720
- 2.23** a),06 b)  $P(KH) P(B) = ,24$  d)  
,947
- 2.24** a) 0,04 b) 0,27 c) 0,25 d) 0,73
- 2.25** a)  $p = ,25$  b)  $p = 1/3$
- 2.26** a) 21/40 b) ,205 c),660
- 2.27** a) 0,420 b) 0,714 c) 0,483
- 2.28** a),1 ,92 ,98 ,9 ,08 ,02 ,092  
b)  $P(B) = ,110$  c)  $P(A^C|B^C) =$   
,991
- 2.29** a) 3/5 b) 4/9 c) 22/45
- 2.30** a),00032 b),0016 c),3277  
d),0384 e),1920
- 3.1** a)  $S = \{KK, KM, MK, MM\}$   
b)  $f(0) = ,25 f(1) = ,5 f(2) = ,25$
- 3.2**  $f(x) = 0,5^x, x = 1,2,\dots$
- 3.3**  $f(0) = 125/216 f(1) = 75/216$   
 $f(2) = 15/216 f(3) = 1/216$
- 3.4**  $f(x) = (5/6)^{x-1} \cdot (1/6), x = 1,2,\dots$
- 3.5** Stolper (loddrette linjer) ved hver  
 $x$ -verdi med høyde  $f(x)$ .
- 3.6** Rektangler med bredde  $\Delta y = 20$   
og høyde  $g(y)/\Delta y$ .
- 3.7** 2,2 10 -45,6
- 3.8** 2,09 20,5 57,4

- 3.9** a,b)  $S = \{HH, HU, HB, UH, UU, UB, BH, BU, BB\}$  med tilhørende sannsynligheter: ,35 ,14 ,21 ,10 ,04 ,06 ,05 ,02 ,03  
c)  $f(0) = ,15$   $f(1) = ,50$   $f(2) = ,35$
- 3.10** a)  $f_X(0) = ,3$   $f_X(1) = ,4$   $f_X(2) = ,3$   
 $f_Y(0) = ,3$   $f_Y(1) = ,7$  b)  $\mu_X = 1,0$   
 $\sigma_X^2 = ,6$   $\mu_Y = ,7$   $\sigma_Y^2 = ,21$   
 c)  $X$  og  $Y$  ikke uavhengige  
 d)  $f_Z(-1) = ,2$   $f_Z(0) = ,5$   $f_Z(1) = ,1$   
 $f_Z(2) = ,2$   
 e)  $\mu_Z = ,3$   $\sigma_Z^2 = 1,01$
- 3.11** a)  $-,5$  b) 61 c) nei
- 3.12** a)  $1/3$  c)  $700/27 = 25.9$
- 3.13** a)  $1/11$  b)  $EX = 2$   $std(X) = 1,4$
- 3.14** a) 15 2,5 b) 6 1,5 c) nei  $\frac{1}{2}$  ja nei d) ,77
- 4.1** a)  $5/16$  b)  $11/32$  c)  $\mu = 3$ ,  
 $\sigma = \sqrt{3/2} = 1,22$
- 4.2** a) 0,32 b) 0,09
- 4.3**  $f(0) = 6/15$   $f(1) = 8/15$   $f(2) = 1/15$   
 $E(X) = ,667$   $std(X) = ,596$
- 4.4** a)  $8/3$  b) 1,16 c) ,049
- 4.5** a) ,107, b) ca. 0,33,  
 c) falsk forventning d) ,055
- 4.6** a) ,125 b) ,303
- 4.7** a) vertikale linjer ved  $x = 0, 1, 2, 3, 4$  og 5 med høyder lik de respektive relative frekvenser: ,691 ,204 ,065 ,032 ,005 og ,003.  
 $\bar{x} = ,465$   $s = ,83$   
 b) Poisson-tilpasning med  $\lambda^*$   
 $= \bar{x} = (1/648) \cdot \sum f_i x_i = 301/648 = ,465$  gir følgende sanns.heter: ,628 ,291 ,068 ,011 ,001 og ,0001. Avvik fra a) kan skyldes at sannsynligheten for at flere blir involvert i en ulykke samtidig ikke kan neglisjeres.
- 4.8**  $k = 13$  glass
- 4.9** a) 0,181 b)  $k = 3$
- 4.10** a) 0,116 b) 0,652  
 c) Forutsetning: Poisson-prosess med konstant  $\lambda$ . På et sentralbord vil  $\lambda$  ofte variere, f.eks. med klokkeslett.
- 4.11** a) 4, 3 og ,0197 c) ,544
- 4.12** a)  $X$  binomisk,  $n = 18$ ,  $p = 1/6$   
 b)  $f(x) = \binom{18}{x} \cdot (1/6)^x \cdot (5/6)^{18-x}$ ,  
 $x = 0,1,..,18$   
 c)  $\mu = np = 3$   $\sigma = \sqrt{5/2} = 1,58$   
 d) ,962 ,038  $10^{-14}$  ,245 ,66
- 4.13** a) stolpediagram: vertikale linjer ved  $y = 0,1,..,6$  med høyde lik de respektive relative frekvenser, f. ,11 ,20 ,28 ,24 ,12 ,05 0  
 b)  $\bar{y} = 1/100 \cdot \sum f_i y_i = 2,21$   
 c)  $\lambda^* = 2,21$  gir Poisson-sannsynligheter: ,11 ,24 ,27 ,20 ,11 ,05 ,02. Bra tilpasning.
- 4.15** c) ,82 ,17 ,01 ,0006 1,2  $\cdot 10^{-5}$  10<sup>-7</sup>
- 4.16** b) midlere antall dødsfall pr. by  
 c) ,454 ,359 ,142 ,037 ,007  
 d) 90,8 71,7 28,3 7,5 1,5
- 4.18** a) 0,1429 b) 5
- 4.19** a) ,333 b) ,366 c) ,900 d) ,301
- 4.20** a) ,1904 b) ,4142
- 4.21**  $X =$  antall syklister i perioden.  
 a) ,05 ,577 b) 3 1,73 c) ,0892
- 4.22** a) ,021 b) ,254 c) ,268 d) ,697  
 e) ,073 f) ,344
- 5.1** -4 5
- 5.2** 0 3 0,5
- 5.3** a) ,997 b) ,954 c) ,683  
 d) 1,000 e) ,017 f) ,353
- 5.4** 1,645 1,96 0 1,96 ,013
- 5.5** a) ,383 b) ,000 c) ,000
- 5.6** ,507 ,159 ,782 ,260
- 5.7** ,212
- 5.8** 16,45
- 5.9** 1070
- 5.10** ,513 ,0621
- 5.11** ,1587 ,4044 ,067 ,0127
- 5.12** b) ,780 c) ,798
- 5.13** ,309 ,866 3/4 ,76
- 5.14** tn.  $N(20, ,5)$  ,067
- 5.15** -1,96 -1,645 0 ,8413
- 5.16** a) ,0432 b) ,0164
- 5.17** a) 14 %, b)  $\bar{X}$  tn.  $N(3,2, ,135)$ ,

- c) Karakterer er *diskrete* og har *endelig* definisjonsområde.
- 5.18** ,965
- 5.19** a) ,0073 b) ,036 c) 1,000 d) ,1
- 5.23** ,693 min.
- 5.24** a) 86,56  $3,2 \cdot 10^{-6}$
- 5.25** a) 0,2119 b) 0,0368
- 5.26** a) 6 b) 4,8 2,19 c) 0,873
- 5.27** a) ,773 ,773 ,547 b) 1,645 c) ,226  
d) ,304
- 5.28** a) 0,3085 b) 0,0385 c) 0,556
- 5.29** 0,1587 0,284 0,815 0,829
- 5.30** a)  $Z$  er  $N(170, 3,48)$  b) ,076  
c) 0,187 d) 0,193 e) 10
- 5.31** b) ,038 d) ,176 e) ,132 f) ,894
- 5.32** ,663 ,0456 ,207 ,129 ,322
- 5.33** a) ,159 ,499 b) ,146 c) 62 12  
,202 d) ,465
- 5.34** ,159 ,683 ,737 ,124 ,815
- 5.35** a) ,001 b) 1000 c) ,632 d) ,819
- 5.36** a) ,977 ,691 ,843 b) 3, 4 og 1,  
 $f(z) = (2\pi z)^{-1/2} e^{-z/2}$ ,  $z \geq 0$   
c) ,264 ,5
- 5.37** a) ,788 ,788 ,740  
b)  $a = 1,28$   $b = 1,96$   $c = -1,09$   
c) ,115 ,023 d) ,074 e) 1044
- 5.38** a) ,382 b) ,327 c) 5 halvlitere  
d)  $a = ,48$
- 5.39** a) ,0215 ,636 b) ,162 c) 8,11 ,96
- 5.40** a) ,368 287,7 b) ,406 c) ,202
- 5.41** a) 1550 22,36 ,013 b) ,730 c) ,30
- 5.42** a)  $f(y) = 1/1,75$ ;  $.75 < y < 2,5$   
 $\mu = 1,625$   $\sigma = ,505$  b) ,577
- 5.43** a) ,923 b) ,031 c) ,100 d) ,896
- 5.44** a) ,023 ,288 b) ,145 alternativ 1
- 5.45** a) ,106 b) ,285 c) ,285 d) ,0062
- 6.1**  $\mu_2^*$  best (har minst varians)
- 6.3** a)  $(-1,12, 1,52)$  b)  $(8,45, 12,75)$   
c)  $,72 \mp 1,96 \cdot ,0449 = (,63, ,81)$   
Forutsetninger: uavhengige observasjoner, stor populasjon.
- 6.4** a) (1,7 %, 3,3 %) b) (105, 149)
- 6.5** a)  $n = 16$ , b)  $(49,7, 50,7)$  cl
- 6.6** a)  $S = \{1,2,\dots,N\}$   
b)  $N^* = ((n+1)/n) \cdot X_{(n)} - 1 = 69$
- 6.7** a) ,0625 b) 3 c) 18
- 6.8** a) ,289 b) ,952 c) ,641  
d) (49,86, 51,03)  
e) (49,71, 51,18) f)  $n \geq 16$
- 6.9** a) ,067 b)  $\bar{x} = 4$   $s = ,384$   
c) 90 % KI: (3,76, 4,24)
- 6.10** a)  $\frac{1}{n} p(1-p)$ , b)  $1/(4n)$ ,  
c) (,573, ,627)
- 6.12** a)  $E(X_i) = p$ ,  $\text{Var}(X_i) = p(1-p)$   
b) antall pasienter som blir helbreddet av i alt  $n$ , e) ,031
- 6.13** a) tidsavstanden mellom to på hverandre følgende bussavganger.  $P(T < \theta/4) = ,25$ ,  
 $P(T = \theta/2) = 0$  c)  $\theta^2/(3n)$   
d)  $\theta^* = 13,4$   $\theta^{**} = 15,0$   $\theta^{**}$  best
- 6.15** a)  $X \sim \text{Bino}(1600, p)$  (tilnærmet, dersom en ikke kan spørre de samme personene flere ganger).  
b)  $p^* = X/1600$ ,  $E(p^*) = p$ ,  
 $\text{Var}(p^*) = p(1-p)/1600$ ,  $p^*$  er forventningsrett c) ,160. Ingen svært sterk indikasjon.  
d) (,348, ,396),  $n = 6233$  (må da fortsatt ha  $p^* = ,372$ ).
- 6.16** a) (14,34, 16,26) b) (,547, ,786)  
c) 246 d) 385
- 7.1** Forkast  $H_0$
- 7.2** Velger  $\alpha = 0,05 \Rightarrow$  Forkast  $H_0$ ,  
( $P^* = ,046$ ).
- 7.3** Se 8.9.
- 7.4** a)  $T \geq 10$ , b)  $\beta(8) = ,717$ ,  
c)  $\lambda = 6,7,8,9$   $\gamma = ,08, ,17, ,28, ,41$
- 7.5** a) Forkast  $H_0$ . Feil av type I:  
Feilaktig konkludere med at fettinnholdet er høyere enn 12.  
b)  $\gamma(14) = ,935$   
c)  $\gamma(14) = ,798$  d)  $n \geq 14$
- 7.6** a) Forkast  $H_0$  b) (85,2, 86,3)
- 7.7** a) små verdier b)  $\text{hyp}(6,6, N)$   
c)  $R: X \leq 2$
- 7.8** a) ,834 b) ,984  
c) Forkast  $H_0$ ,  $P^* = 3,3\%$ .
- 7.9** a)  $\bar{x} = 2,1$   $s = ,051$

- b)  $H_1: \mu > 2$   $H_0: \mu \leq 2$ ,  
 $T = 3 \cdot (2,1 - 2) / 0,051 = 5,88 >$   
 $t_{0,05} = 1,860$ . Forkast  $H_0$ .
- c)  $\alpha$  = sannsynlighet for feilaktig å konkludere at forventet diameter er stor.
- d)  $X$  er  $N(\mu, \sigma)$ ,  $\bar{X}$  er  $N(\mu, \sigma/3)$
- 7.10** Grunn til å tvile på at mynten er iorden.
- 7.11** a) konkludere med at det ikke er 2 hvite og 2 sorte kuler i urnen, selv om det er det.  
 b) unnlate å konkludere med at det er forskjellig antall hvite og sorte kuler i urnen, selv om det er ulikt antall.  
 c)  $P(\text{type I feil}) = 1/3$   
 d)  $P(\text{type II feil}) = 1/2$
- 7.12** a) Forkast  $H_0$ .  $P^* = ,0065$   
 b)  $\beta(19,50) = ,20$  c)  $\alpha = ,114$   
 e)  $n = 28$
- 7.13** a) Forkast  $H_0$  b) bruk fantasien!
- 7.14**  $T = -1,966$ , Forkast  $H_0$ . Mulig feilkilde: for få siffer.
- 7.15** Forkast  $H_0$ .
- 7.16** a) 95 % KI for  $\mu$ : (4,32, 4,61).  
 Forutsetninger:  $X_1, \dots, X_9$  uif  $N(\mu, \sigma)$ .  
 b)  $H_0: \mu = 4,3$  mot  $H_1: \mu > 4,3$ .  
 $T = \frac{\bar{X} - 4,3}{0,22} \sqrt{9}, R: T > 1,64$   
 $T = 2,27 > z_{0,05} = 1,645$ .  
 Konklusjon: Forkast  $H_0$   
 $\gamma(\mu) = 1 - \Phi\left(z_{0,05} - \frac{\bar{X} - \mu}{0,22} \sqrt{9}\right)$   
 $\gamma(4,5) = 0,86$   
 c)  $\sigma$  erstattes med  $s$ ,  $T = 2,13 >$   
 $t_{0,05} = 1,86$ , d.f. = 8, dvs.  $H_0$  forkastes. d) (0, 0,52).
- 7.17** a) (1468,2, 1470,9) er et 95 % KI for  $\mu$ . b) Minst 16 målinger.  
 c)  $R: T > z_{0,05} = 1,645, T = 2,21$   
 $\Rightarrow$  Forkast  $H_0$ .  
 d)  $\gamma(\mu) = P(T > z_{0,05} | \mu) =$
- $1 - \Phi\left(z_{0,05} - \frac{\mu - \mu_0}{2} \sqrt{8}\right)$   
 $\gamma(1470) = 0,88$   
 e)  $\sigma$  erstattes med  $s$ ,  $T = 3,1 >$   
 $t_{0,05} = 1,895$ , d.f. = 7, dvs.  $H_0$  forkastes.
- 7.18** a)  $X \sim \text{Bino}(40, 0) \Rightarrow EX = 40p$  og  $\text{Var}(X) = 40p(1-p)$ . Da er  $E p^* = E(X/40) = 40p/40 = p$   
 $\text{Var}(p^*) = \text{Var}(X/40) =$   
 $\frac{1}{40^2} 40p(1-p) = p(1-p)/40$   
 b)  $p^* = 3/40 = ,075$ ,  $\text{std}^*(p^*) = \sqrt{\frac{3}{40}(1 - \frac{3}{40}) / 40} \approx ,0416$   
 c)  $P^* = P(X \geq 3 | p = ,05) = 1 - P(X \leq 2 | p = ,05) =$   
 $1 - \sum_{i=0}^2 \binom{40}{i} ,05^i \cdot ,95^{40-i} = 0,32$
- Konklusjon: Siden  $P^* > 0,05$ , er det på nivå 5 % ikke grunnlag til å hevde at  $p > ,05$ .
- 7.19** a) ,1587  
 b) lageret må være på 8823 liter  
 c) ,0228 d) Ikke forkast  $H_0$   
 e)  $\gamma(7500) = ,91$ .
- 7.20** a) ,5808 b)  $Y \sim \text{Bino}(50, ,58)$   
 $EY = ,29$   $\text{Var}(Y) = 12,17$   
 $\text{std}(Y) = \sqrt{3,49}$   $P(Y > 35) \approx ,032$   
 c)  $n = 10$   
 d)  $H_0: \mu = 73$   $H_1: \mu > 73$   
 $H_0$  forkastes. e)  $\gamma(80) = ,935$
- 7.21** a) ,0082 b) ,1573 c) ,629  
 d) ,0382 e) Forkast  $H_0$
- 7.22** a)  $\text{Bino}(10, ,5)$   $EX = 5$   $\text{Var}(X) = 2,5$   $P(X \leq 3 | H_0) = ,172$   
 b)  $P(X \leq 3 | H_0) = ,172 > ,05$   
 $H_0$  forkastes ikke
- 7.23** a)  $X$  er  $N(100, 2)$   
 $P(\text{defekt}) = ,0455$   
 b)  $S$  er  $N(1000, 2\sqrt{10})$   
 $P(S > 1010) = ,0569$   
 c)  $Y$  er  $\text{Bino}(5, ,0569)$  ,0289  
 d)  $H_0$  forkastes

- e)  $\gamma(\mu) = 1 - \Phi\left(z_{,025} - \frac{102 - 100}{2 / \sqrt{10}}\right)$   
 $+ \Phi\left(-z_{,025} - \frac{102 - 100}{2 / \sqrt{10}}\right)$   
 $\gamma(102) = ,885$
- f) t-test ville ikke gitt forkasting.
- 7.24** a)  $H_0: \mu = 300\ 000 \text{ km/s}$  mot  
 $H_1: \mu \neq 300\ 000 \text{ km/s}$   
b) Forkast  $H_0$   
c)  $\gamma(299\ 700 \text{ km/s}) = ,851$   
 $\gamma(299\ 900) = ,170$   
 $\gamma(300\ 100) = ,170$ .  
d)  $H_0$  som før,  $H_1: \mu > 300\ 000$   
Antall målinger: 11
- 7.25** a) Ikke forkast  $H_0$  b) ca. 8,1 %  
c)  $\gamma(30) = ,64$   $\gamma(39) = ,40$   
 $\gamma(50) = ,16$ . Selv dersom det gjennomsnittlig bare er 30 gram pepperoni, er det beskjedne 64 % sannsynlighet for å avsløre dette med testen.
- 7.26** a)  $H_0$  forkastes ikke, b) 4706  
c) ,035
- 7.27** a)  $H_0: \mu = 218$  mot  $H_1: \mu < 218$   
b)  $H_0$  forkastes når  $\bar{x} = 214,6$   
 $H_0$  forkastes ikke når  $\bar{x} = 216,5$   
c)  $\gamma(216) = ,36$ ,  $\gamma(212) = ,99$   
 $\gamma(214) = ,83$   
d)  $n = 126$ ,  $\bar{x} = 214,6$ , forkast  $H_0$ .
- 7.28** a)  $H_0: \mu = 135$  mot  $H_1: \mu < 135$   
Nei, ikke signifikant  
b)  $\gamma(128) = ,37$ , c)  $n = 53$
- 8.1** a) nei, b)  $X_1, \dots, X_{10}$  uif  $N(\mu_1, \sigma)$ -variable,  $Y_1, \dots, Y_{10}$  uif  $N(\mu_2, \sigma)$ -variable,  $X$ -ene uavh. av  $Y$ -ene.  
c) (-9,09, 0,69)
- 8.2** a) A: 3 1,64 B: 4,29 1,50 b) ja  
c) (-1,81, -,77) (A-B)
- 8.3** a) ja, b) (.44, 6,77)
- 8.4** Grunn til å konkludere med at forventet surstoffforbruk øker.
- 8.5** a) ,197, b) (.012, ,382)  
c) (0, ,54)
- 8.6** a) (.015, ,285). Siden nedre intervallgrense er positiv kan vi påstå at voltmetre av typen B gir høyere verdier for spenningen enn voltmetre av type A.  
b) (0, ,494).
- 8.7** a)  $\bar{x} = 12,98$   $\bar{y} = 11,90$   
 $s_x = 1,50$   $s_y = 1,23$   
b) (-,21, 2,37), ikke sign. forskj.  
c) (-,009, 1,19), ikke sign. forskj.
- 8.8** a) Behandling: 25,8 35,2 5,93  
ikke behandling: 16,3 18,33 4,28  
b)  $(\mu_1 - \mu_2)^* = 9,5$ ,  $\sigma^{2*} = 26,76$   
c) ikke signifikant vekststimulus
- 8.9** a) signifikant b) (-1,7 %, 6,3 %)
- 9.1** b) 18,17 1,74  
c)  $y_r^* = ,31 + ,08 \cdot x$  d)  $s = ,29$   
e)  $\text{std}^*(a^*) = ,25$   $\text{std}^*(b^*) = ,013$   
 $\text{std}^*(y_r^* | x=10) = ,15$   
f) faktor 2, faktor 26/5 på  $\text{Var}(S^2)$
- 9.2** a)  $\ln(1-y/a) = -t/T$   
 $y' = \ln(1-y/a)$ ,  $a' = 0$ ,  $b' = -1/T$   
b)  $b'^* = \sum t_i y'_i / \sum t_i^2$   
c)  $T^* = 4,6$  s. Setter a = 10,00 som asymptotisk verdi, og dropper punktet  $(t,y) = (100, 10,00)$  i regresjonen, da dette ligger svært langt fra de andre punktene.
- 9.3** a) a b)  $a^* = \bar{Y}$ ,  $a^* = 0,34 \text{ }^\circ\text{C}$   
c)  $\text{std}^*(a^*) = 0,032 \text{ }^\circ\text{C}$
- 9.13** a)  $y_r^* = -9,66 + ,0087x$   
b) Ikke signifikant økning.
- 9.14** a)  $y_r^* = -11,6 + ,125x$   
b) Forkast  $H_0$
- 9.15** a)  $r \approx ,89$  b)  $y_r^* = 167,2 + 1,97x$   
d) (1,14, ,2,79)
- 9.16** a)  $y_r^* = 5,90 + ,195x$   
b) Forkast  $H_0$
- 9.17** c) (.021, ,063)
- 10.1** a) forsøksenheter: stålplatene  
b) forsøksfaktorer: A, B og C  
c) faktornivåer: Kun ett pr. faktor (kvalitet).

- d) 3 behandlinger, A, B og C.  
e) 3 populasjoner, en for hver ståltype. Hver populasjon består av samlingen av de flytegrenser som ville blitt målt dersom alle prøvene i partiet ble undersøkt.  
f) 100 replikater av A, 70 av B og 30 av C.  
g) responsvariabel: flytegrense.
- 10.2** a)  $\bar{y}_1 = 13,4$   $\bar{y}_2 = 19$   $\bar{y}_3 = 14,4$   
 $\bar{y} = 15,36$   
c)  $SST = 76,81$  d.f. = 2  
 $SSE = 98,40$  d.f. = 11,  $F = 4,29$   
d)  $H_0$ : Ingen behandlingsforskjell.  $H_0$  forkastes.  
e) B har signifikant størst forventning.
- 10.3** a)  $k = 5$ ,  $n = 45$ , b)  $F_{4,40}$   
c)  $F = 2 < F_{4,40}(0,05) = 2,61$   
d) ikke nødvendigvis
- 10.4** a)  $SST = ,13$   $k-1 = 2$   
 $SSE = ,058$   $n-k = 15$ ,  $F = 167$   
b) PC<sub>2</sub> signifikant best  
c) Svært lik ytelse på tross av signifikante forskjeller. Ytelsen neppe avgjørende for valg av PC i dette tilfellet.
- 10.5** a)  $SST = ,127$  d.f. = 3  
 $SSE = ,030$  d.f. = 17,  $F = 24,0$   
b) Signifikant forskjell  
c) A signifikant best
- 10.6** a) Ola ville nøyet seg med en simultan  $F$ -test og forventet å ikke få forkastet hypotesen  $H_0$  om ingen forventet forskjell. Kari ville sammenlignet direkte A mot B og A mot C for å påvise at A er signifikant bedre enn både B og C.  
b) Ingen har rett, A signifikant bedre enn B, men ikke signifikant bedre enn C.
- 11.1**  $X = -3 + 13R$ ,  $X_1 = 8,202$   
**11.2**  $X_1 = \square 0,01488$ ,  $X_2 = \square 0,09569$   
**11.3** 5, 3, 4 og 2.
- 11.4** 5, 2, 3 og 1.  
**11.5**  $X_1 = 0,02977$   $X_2 = 0,1914$   
**11.6** syklus: 4,2,6,8,4,...  
sykluslengde: 4.  
**11.7**  $T = 14,3 < \chi^2_{,05} = 9,49$  (d.f. = 4).  
Grunnlag for å tvile på generatoren.
- 11.8** a)  $X_1 = 1,1761$   $X_2 = -1,5886$   
**11.9**  $X_1 = 2,959$   $X_2 = 3,036$   
**11.10**  $(X, Y) = (0,3631, -0,5342)$
- 12.1** a)  $LCL = 8,86$   $UCL = 10,08$   
b)  $LCL = 0$   $UCL = 1,54$   
c) grunn til alarm
- 12.2** a) Ved tidspunkt 3 skal siste verdi være 9 og ikke 17  
b)  $\bar{X}: LCL = 5,38$   $UCL = 13,57$   
R:  $LCL = 0$   $UCL = 15,01$   
c) Ja, ut fra R-diagrammet
- 12.3** a)  $\bar{X}: LCL = -35,9$   $UCL = 41,7$   
R:  $LCL = 0$   $UCL = 47,6$   
b) ikke grunn til alarm

## Stikkord

- alternativhypotese, 181, 182
- ANOVA-tabell, 269, 274, 276, 277
- Bayes' formel, 51, 52
- behandling, 217, 220, 221, 267, 268-71
- behandlingskvadratsum, 268, 270, 271
- Bernoulli-forsøk, 85, 86, 87
- betinget sannsynlighet, 50, 51, 54, 55
- binomisk fordeling, 84, 86, 87, 90, 93, 95, 97
- binomisk tabell, bruk av, 90, 91, 99
- binormal fordeling, 111, 144, 145, 294
- binormale variabelpar, 295
- blokk, 218, 232
- bootstrap, 286, 300, 301
- deterministisk, 64
- diskret fordeling, 112
- diskret stokastisk variabel, 9, 287
- eksperiment, 32, 33, 45
- eksponensial-fordeling, 110, 119, 122, 123
- empirisk, 1
- endelig utfallsrom, 33
- enkeltutfall, 32, 34, 35, 37-40
- en-vegs variansanalyse, 268
- erfaringsregelen, 316
- estimator, 159, 160, 163, 240
- estimator for  $p$ , 163, 164
- estimator for  $\mu$ , 161, 162
- estimator for  $\sigma^2$ , 245
- estimatorfordeling, 160
- faktornivå, 267, 268
- fakultet, 46
- feilkvadratsum, 245, 259, 268
- feilrate, 110, 115, 116, 120, 121
- feil-variabel, 239
- $F$ -fordelingen, 111, 143, 144
- fordeling, 63, 66, 67, 71
- forkasting, 183, 184, 190, 197, 203
- forkastingsområde, 183, 184, 193, 196, 276
- forsøk, 267
- forsøksenhetsbegrep, 267
- forsøksfaktor, 267, 268
- forventning og varians til summer, 114
- forventning, 63, 69, 71, 72, 74, 76, 77, 110, 113, 114, 128, 129, 133, 135
- forventningsrett, 160, 162
- frekvens, 11, 12
- frekventistisk sannsynlighetsbegrep, 31
- frihetsgrader, 143, 272, 273, 274, 276
- $F$ -tabell, bruk av, 144
- gammafordeling, 110, 121, 122, 123
- Gauss-fordeling, 110
- Gausskurve, 126
- generering av binormale variabelpar, 294
- generering av kontinuerlige stokastiske variabler fra  $F^{-1}$ , 286
- generering av sammensatt variabel, 297, 299
- generering av tilfeldige diskrete variabler, 288
- generering av uif  $N(\mu, \sigma)$ -variabler, 292, 293
- gjennomsnitt, 1
- gjensidig ekskluderende hendelser, 40, 54, 55
- gruppert interkvartilbredde, 11
- gruppert median, 12
- gruppert middelverdi, 11, 12, 14, 16
- gruppert standardavvik, 11, 12, 14, 16
- gruppert varians, 12
- hendelse, 32, 33
- hukommelsesfri fordeling, 119, 120
- hypergeometrisk fordeling, 84, 93, 94
- hypotesetest (se også test), 231, 248, 275, 300
- hypotesetest for k behandlinger, 276

- ikke-lineær modell, 253
- inn-variabel, 23
- interkvartilbredde (empirisk), 6, 8, 9, 12, 15
- interpolert standardavvik, 224
- interpolert varians, 223
- invers kumulativ fordelingsfunksjon, 285
- invers transformasjon, 253
- karakteristisk levetid, 123
- KI for  $a$ , 249
- KI for  $b$ , 249
- KI for  $E(Y|x) = a + bx$ , 250
- KI for  $\mu, \sigma$  kjent, 168
- KI for  $\mu$ , ukjent  $\sigma$ , små utvalg, 171
- KI for  $\mu$ , ukjent  $\sigma$ , store utvalg, 169
- KI for  $p$ , store utvalg, 170
- KI for  $\Delta\mu$ , små utvalg, 225, 227
- KI for  $\Delta\mu$ , store utvalg, 229, 230
- KI for  $\mu_D$ , små utvalg, 231, 232
- KI for  $\mu_D$ , store utvalg, 231
- kjikkvadratfordelingen, 111, 126, 140, 276
- klasse, 9, 10, 11, 13, 16
- klassebredde, 9, 10, 15, 16
- klassefrekvens, 11, 13, 16
- klassegrense, 10
- klasseintervall, 9
- klassemidtpunkt, 10, 12
- kombinasjonsregel, 43, 47, 49
- komplement, 39
- konfidensintervall for
  - behandlingsforskjell, 277, 278
- konfidensintervall (KI), 159, 164, 166, 167, 170, 172, 226, 227, 231, 248, 270, 275, 300
- kontinuerlig fordeling, 112
- kontinuerlig utfallsrom, 33
- kontinuerlig variabel, 9, 286
- kontinuitets-korreksjonsledd, 137
- korrelasjonskoeffisient (empirisk), 20, 21, 22
- korrelasjonskoeffisient, 63, 77, 78, 257, 258, 260, 296
- kovarians, 76, 77, 78
- kovariansmatrise, 145
- kumulativ fordelingsfunksjon, 63
- kumulativ frekvens, 11, 12, 16
- kvartil, 8, 15
- levetidsfordeling, 124
- liketall, 6
- lineær betingelse, 273
- lineær regresjonsmodell, 240, 247
- linearisert modell, 253
- lurevariabel, 23, 260
- marginal sannsynlighetsfordeling, 75
- Matlab, 303
- med tilbakelegging, 43
- median (empirisk), 6, 7, 11, 14, 16
- middelverdi (empirisk), 1, 2, 5, 14, 18
- middelverdidiagram, 309
- middelverdifordeling, 135
- middelverdi-variasjonsbredde»-diagram, 308
- Minitab, 17, 201, 228, 280, 295, 313, 315
- minste kvadraters estimator, 239, 242
- minste kvadraters metode, 24
- modellfri, 300
- modellsjekk, 279
- modulus, 289
- Monte Carlo-simulering, 117, 284
- multiplikativ kongruensgenerator, 289
- myter om Shewart-diagrammer, 316
- nedre klassegrense, 10, 12
- nedre kvartil, 8
- nedre styregrense, 306
- normalfordeling, 110, 126, 127, 138
- $N(0,1)$  tabell, bruk av, 128, 130-132
- normalforutsetning, 310
- normaltilnærmelse til binomisk fordeling, 137, 138
- normaltilnærmelse til Poisson-fordelingen, 139
- nullhypotese, 181, 182

- observator, 134, 135  
observatorfordeling, 134  
oddetall, 6  
ordnede utvalg, 43  
ordningsregelen, 43, 45, 49  
overlevelsесfunksjon, 110, 115
- parallellsystem, 41  
parameter, 111, 119, 120, 122, 139  
partisjon, 55  
parvis sammenligning, 217, 218, 219, 220, 221, 230, 231  
Pearsons kjikvadrat-tilpasningstest, 203, 204  
permutasjonsregel, 45  
Poisson-fordeling, 84, 96, 97, 100, 101, 120  
Poisson-tabell, bruk av, 99  
populasjon, 84, 85, 93, 94, 95, 98, 182, 275  
populasjonsfordeling, 134, 136, 284  
potensregel, 43, 45, 49  
prediksjon, 246  
prediktor, 23, 239  
primærdata, 1  
prosentil, 6, 7, 12  
prosessnivå, 306, 311  
prosessvariasjon, 306, 311  
punktestimering, 159, 160, 161, 163  
*p*-verdi, 190
- randomisering, 219  
Rayleigh-fordelingen, 111, 123, 124, 125  
regresjonskurve, 254  
regresjonslinje, 239, 248, 250, 252  
rekursiv generator, 289  
relativ frekvens, 11, 13, 14, 17, 35, 36  
relativ frekvens-histogram, 12-14, 17  
relativ frekvens-stolpediagram, 13  
relativ kumulativ frekvens, 11  
relativ usikkerhet, 73  
replikat, 267, 268  
residual, 241, 245, 258, 259, 271  
residualer og modellsjekk, 256
- residualplott, 259, 280  
residualsum, 268, 273  
responsvariabel, 23, 239, 240  
robust, 6, 7  
robusthet, 316  
rådata, 1
- sammenligning av k behandlinger, 269  
sammensatt variabel, 299  
sannsynlighet, 31, 33, 35, 36, 43, 49, 55, 113  
sannsynlighetsfordeling, 63, 65, 66, 67, 69  
sannsynlighets-histogram, 68  
sannsynlighetsmodell for diskrete utfallsrom, 37  
sannsynlighetsmodell, 110, 111, 116  
sannsynlighetstetthet, 63, 113  
sentralgrenseteoremet, 111, 126, 133, 135, 136  
sentralgrenseteoremet, 284  
seriesystem, 41  
Shewhart styrediagram, 307, 308, 312  
signifikansnivå, 189, 197  
signifikant sannsynlighet, 190, 192, 194, 197, 201  
simultan sannsynlighetsfordeling, 63, 74, 75, 78, 80  
simultane konfidensintervall, 279  
skjæringsparameter, 239  
slumptall, 290  
slumptallalgoritme, 286, 290  
slumptallgenerator, 117, 286, 289, 291  
små utvalg, 223, 227, 228, 231  
snitt, 39  
spesifikasjonsgrense, 307  
spredningsdiagram, 18, 19, 21  
spredningsdiagram, 250, 252  
standardavvik (empirisk), 1, 3, 4, 5, 14, 16, 18  
standardavvik, 63, 72, 73, 75, 110, 113, 127, 128, 129, 133, 135, 136, 300  
standardisert variabel, 128  
statistisk hypotese, 181, 182, 183

- statistisk prosesstyring, 306
- statistisk stabil prosess, 307
- steilhetsparameter, 239
- stokastisk variabel, 63, 64, 65, 66, 69, 70, 73
- stokastisk, 64
- stolpediagram, 68
- styrediagram, 306, 308
- styregrense, 308, 309
- styrkefunksjon, 186, 187, 188, 192
- styrkekurve, 186, 189, 192, 197, 198
- sum av normalfordelte variabler, 132, 133
- summasjonsformler, 241
- sykluslengde, 290
- tallpar, 18
- tellbart uendelig utfallsrom, 33
- tellefordeling, 84
- test av  $a$ , 249
- test av  $b$ , 249
- test av  $E(Y|x) = a + bx$ , 250
- test av  $\mu$ , kjent  $\sigma$ , 194, 196
- test av  $\mu$ , ukjent  $\sigma$ , små utvalg, 199
- test av  $\mu$ , ukjent  $\sigma$ , store utvalg, 200
- test av  $p$ , store utvalg, 202
- test av  $\Delta\mu$ , små utvalg, 225, 227
- test av  $\Delta\mu$ , store utvalg, 229
- test av  $\mu_D$ , små utvalg, 231, 233
- test av  $\mu_D$ , store utvalg, 231
- testobservator, 183, 184, 188, 193, 194, 196, 200, 203, 204, 276
- $t$ -fordelingen, 111, 142
- tilfeldig utvalg, 84, 85, 91, 92, 93, 220, 226
- tilfeldiggjøring, 219
- to grunnregler (Shewart-diagram), 317
- toleransegrense, 307
- tosidig test, 191
- to-vegs variansanalyse, 268
- transformasjon av variabler, 253
- transformasjoner, 253
- $t$ -test, 201, 228
- type I feil, 185
- type II feil, 185, 186
- uavhengig variabel, 239, 240
- uavhengige hendelser, 54
- uavhengige utvalg, 217, 218, 220-222
- uavhengighet mellom variabler, 64, 79
- ukorrelerhet, 64, 79
- uniform fordeling, 110, 116, 117, 118
- uniform sannsynlighetsmodell, 34, 35
- union, 39, 40
- uordnede utvalg, 43
- urnemodell, 43, 44, 86
- uten tilbakelegging, 43, 93
- utfallsrom, 32, 33, 34, 36, 45
- utligger, 257
- utvalg, 94
- ut-variabel, 23
- valg av estimator, 161, 162
- varians (empirisk), 3, 5
- varians, 63, 72, 73, 74, 77, 78, 113, 114, 133, 140, 141
- variansanalyse, 268
- varianskorreksjonsfaktor for endelig populasjonsstørrelse, 95
- variasjonsbredde, 310
- variasjonskoeffisient, 73
- Venn-diagram, 37, 38, 55, 56
- Weibull-fordelingen, 110, 122, 123
- $XmR$ -diagram, 308, 314, 315
- øvre klassegrense, 10
- øvre kvartil, 8
- øvre styregrense, 306

Statistikk og sannsynlighetsregning er spesielt tilpasset Ingeniørutdanningsrådets rammeplan for grunnlagsfaget statistikk (2 vekttall) ved høgskolenes ingeniørutdanninger. Boka inneholder også emner utover dette. Tilpasningen til ingeniørstudentene gjelder valg av eksempler og oppgaver, samt emnevalg og valg av matematisk nivå. Hvert kapittel avsluttes med oppgaver og formelsamling. Av over 200 oppgaver i boka er det ca. 80 eksamsoppgaver fra tidligere statistikkesamener for ingeniørstudenter ved forskjellige høgskoler i Norge.

Boka inneholder et fyldig kapittel om «Monte Carlo»-simulering. Et eget kapittel omhandler Shewart-diagrammer. For øvrig dekker boka de vanlige innføringsemnene som beskrivende statistikk, sannsynlighetsregning, estimering, hypotesetesting, lineær regresjon og variansanalyse. Boka burde være godt egnet også for andre enn ingeniørstudenter.

En egen bok med komplette løsningsforslag til alle oppgavene er også utarbeidet av samme forfatter, med tittel *Statistikk og sannsynlighetsregning – oppgaver og løsningsforslag*.



Alf Harbitz er forsker og statistiker ved Havforskningsinstituttets avdeling i Tromsø. Han er dr.scient. i fysikk og har arbeidet med matematisk statistikk og fysikk i næringsliv og innen universitet og høgskole fra slutten av 1970-åra.

ISBN 978-82-7674-535-1

A standard linear barcode representing the ISBN number 978-82-7674-535-1.

9 788276 745351