Sannsynlighetsregning og statistikk med anvendelser

NOST (Norwegian Open source Statistics Textbooks)

24. september 2021

Forord

Denne boken er basert på Jostein Lillestøls lærebok "Sannsynlighetsregning og statistikk med anvendelser" som blant annet ble brukt i grunnkurs i statistikk ved Norges Handelshøyskole i perioden 1977–2008. Kildekoden til boken, tilgjengelig fra `https://github.com/nsfnost/Lillestol`, er en del av NOST (Norwegian Open source Statistics Textbooks) under en CC BY-NC-SA 4.0 lisens. Dette innebærer at boken kan modifiseres (i henhold til lisensen), og tilpasses spesielle kursbehov. Da det forventes at mange ulike personer i fremtiden vil bidra til læreboken er forfatteren av boken NOST. Fordi kildekoden er under versjonskontrol vil det alltid være mulig å spore hvilken deler som kommer fra Lillestøls lærebok. Det oppfordres til at forordet ikke endres og alltid inkluderes i boken.

Forord til Lillestøl (1997, 5. utg.)

Dette er en elementær innføring i sannsynlighetsregning og statistikk, med utgangspunkt i matematikken fra videregående skole. Fremstillingen gjør ikke bruk av differensial og integralregning, og burde være tilgjengelig også for lesere med bare 1MA. Ved mange universiteter og høyskoler gis i dag innføringskurser i sannsynlighetsregning og statistikk første studieår, og boken kan være aktuell som lærebok der.

Boken prøver å oppdra leseren til modelltankegang, dvs. at praktiske problemer studeres innen en teoriramme, der konklusjoner bare har gyldighet i relasjon til de forutsetninger som gjøres. Boken er ingen kokebok for de metoder som oftest anvendes i praksis, ut fra oppfatningen at det ikke er en metode alene som overføres til nye situasjoner, men argumentasjon for metoden.

Boken er delt i to deler. Første del (kapitlene 1-8) gir en teoretisk innføring i sannsynlighetsregning og statistikk. Annen del (kapitlene 9-16) omfatter ulike emner for videre studium, der de ulike emner kan leses uavhengig av hverandre.

Første del utgjør den kjerne av stoff som er ment å gi leseren et grunnlag for å forstå sannsynlighetsteoretiske og statistiske argumenter anvendt i praksis, og eventuelt kunne utføre enklere argumenter på egen hånd. Dette stoff danner også utgangspunkt for kommunikasjon med fagstatistiker og for videre studier.

Så langt det er råd forsøker vi i første del å gi presise argumenter for metodene. Selv om teori for kontinuerlige modeller faller utenfor vår matematiske ramme, kommer vi ikke utenom å gi en grunnleggende innføring i normalfordelingen og sentrale statistiske argumenter og metoder knyttet til denne.

I annen del av boken presenteres ulike problemområder der statistisk tankegang kommer til anvendelse. Framstillingen er til dels mindre systematisk enn første del. Vi peker på muligheter, men viser til spesiallitteratur for mer detaljert informasjon.

Advarsel : Av plasshensyn er mange av de datamaterialer som er knyttet til eksempler og oppgaver mindre enn det man i praksis ønsker seg for å belyse problemstillingene. På den annen side tar teorien nettopp sikte på å vurdere påliteligheten av konklusjoner basert på begrenset informasjon.

Ved innlæringen av statistiske metoder og dataanalyse i praksis, vil en kunne dra nytte av en statistisk programpakke. Her anbefales MINITAB, men også andre pakker, som STATGRAPHICS, SPSS, SYSTAT, DATA DESK og JMP, kan tjene formålet. Det typiske regneark (som EXCEL) er ikke like tjenlig. Jeg har ikke funnet det hensiktsmessig å la en bestemt regneteknologi prege boken, men gir noen eksempler på typiske (engelske) utskrifter.

Boken bruker følgende systematikk: Kapitler og avsnitt nummereres med tallkombinasjoner, eksempelvis betyr 2.3 tredje avsnitt i Kapittel 2. Eksempler og oppgaver nummereres fortløpende innen hvert kapittel. Ved referanse til eksempel eller oppgave i et annet kapittel brukes tallkombinasjon, eksempelvis vil Oppgave 4.15 være Oppgave nr. 15 i Kapittel 4. En del stoff er $\star$-merket og kan hoppes over uten å miste sammenhengen. Det samme er tilfellet med enkelte oppgaver som sprenger grensene for det stoff som er gjennomgått. Ved slutten av boka finnes tre appendiks med terminologi, formler og tabeller.

Et mulig pensum for et ett-semesters kurs i sannsynlighetsregning og statistikk med 2 vekttall vil kunne hentes fra kapitlene 1-8 i Del I, $\star$-merket stoff unntatt. Trengs beskjæring er ett eller flere av avsnittene 3.6, 5.7, 6.6, 7.6, 7.7, 8.5 og 8.6 nærmest. Pensum for et 2 vekttalls kurs i sannsynlighetsregning alene, vil kunne omfatte kapitlene 1-6 i Del I, inkludert $\star$-merket stoff og eventuelt supplert med første del av Kapittel 16.

Et 4 vekttalls kurs i sannsynlighetsregning og statistikk vil kunne omfatte kapitlene 1-8 i del I, samt 4-5 utvalgte kapitler fra Del II. Et alternativ er færre kapitler fra Del II, med fordypning i $\star$-merket stoff fra Del I eller med supplerende opplæring i bruk av statistisk programvare.

Jeg takker Henrik Dahl for mange inspirerende samtaler om faglige og pedagogiske spørsmål forut for førsteutgaven og ved flere revisjoner. Takk går også til Geir Egil Eide.

I 5. utgave er det kommet til ett nytt avsnitt: 8.7 Resampling og en rekke nye oppgaver. Flere kapitler er reorganisert og forenklet, bl.a. ved at en del formelle begrunnelser er flyttet til sist i avsnittet og $\star$-merket.

Bergen, 30 september 1996
Jostein Lillestøl

Innhold

Del I

Grunnleggende innføring

Kapittel 1

Introduksjon

1.1 Modell-tenkning

Denne boken omhandler teori knyttet til iakttakelser eller observasjoner, og anvendelser av slik teori. Det kan være alt fra iakttakelse i en tilrettelagt eksperimentsituasjon til iakttakelse av et fenomen i våre omgivelser, i naturen eller samfunnet. Det kan være iakttakelser som tar sikte på å gi ny kunnskap, eller innsikt som er ønsket for å kunne treffe bedre beslutninger, i privatlivet eller i en jobbsituasjon.

Ny innsikt får vi raskest og sikrest dersom vi kombinerer våre iakttakelser med det vi allerede vet om fenomenet. Dette kan skje ved en såkalt *modell*. Bruk av modeller som et forenklet bilde av virkeligheten er kjent fra de fleste fagområder. Blant fordelene ved å bruke en modell nevner vi:

- En god modell hjelper oss i tenkningen ved å skape orden og oversikt. Bare de elementer som er vesentlige for vårt formål er med. En modell kan både inneholde inneholde viten og hypoteser som hittil ikke er bekreftet.

- Ved å basere våre resonnementer på en modell, kan vi undersøke de logiske konsekvenser av modellen, sammenholde disse med nye iakttakelser. Dette kan ofte lede til alternative og bedre modeller.

- Modelltenkning tillater ulike syn, idet en modell alltid innbyr til en diskusjon av de forutsetninger og hypoteser som er gjort, og til tolkninger av iakttakelser basert på modellen.

Hva er så en god modell? Realisme og enkelhet vil vanligvis trekke i hver sin retning. Utfordringen er å finne et kompromiss, som samsvarer med det formål modellen skal tjene (brukbarhet). Som et eksempel la oss se på tre ulike modeller for jordens form: 1. plan, 2. kule, 3. kule som er flattrykt ved polene. Modell 1 var lenge brukbar for de fleste praktiske formål, men ble erstattet av modell 2. En geofysiker vil for sitt formål anse at denne modellen er for urealistisk, og isteden bruke modell 3. En navigatør, i likhet med folk flest, vil imidlertid kunne forsatt være tjent med modell 2 evt. 1. Vi vil derfor aldri si at en modell er den riktige modell. Vi begrenser oss til

Figur 1.1: Status av en modell

å si at en modell ser ut til å være mer velegnet enn andre modeller. Praktisk erfaring er det eneste som kan gi en pekepinn om dette.

Modeller blir brukt i alt fra vitenskap til spekulativ virksomhet. Dersom en i vitenskapen hevder at visse observerte fenomener kan tilfredsstillende forklares ved hjelp av en bestemt modell, så kan andre interesserte studere fenomenet for å etterprøve dette. Har en modell overlevet mange etterprøvinger, tas dette til inntekt for at modellen er brukbar og har gitt oss felles innsikt. I økonomisk virksomhet kan gode modeller gi bedre beslutninger, og målbare gevinster.

Det er vanlig å skille mellom to hovedtyper av modeller, *deterministiske modeller* og *stokastiske modeller*. En deterministisk modell betrakter en iakttakelse eller et utfall som entydig bestemt ut fra gitte eksperimentbetingelser. Dersom modellen er realistisk vil vi, såframt vi har godt kjennskap til disse, kunne bruke modellen til å forutsi utfallet med stor grad av sikkerhet. Mange iakttakelser er av en slik natur at det ikke lar seg gjøre å finne en brukbar deterministisk modell. La oss illustrere dette med et eksempel :

Anta at vi skyter en terning ut av en kanon utover en horisontal flate. Dersom vi bare er interessert i å iaktta posisjonen hvor terningen treffer bakken, så er det lett å lage en brukbar deterministisk modell basert på Newtons andre lov, dvs. kraft er lik masse ganger akselerasjon K=m·a. Dersom vi bare er interessert i hvilken side av terningen som vender opp når den har falt til ro, vil det være verdiløst å prøve å lage en deterministisk modell (selv om en slik eksisterte) basert på Newtons bevegelseslover. Likevel vil det være ønskelig å lage en slags modell for eksperimentet, helst en modell som tar hensyn til forestillingen om at alle 6 sider på terningen har samme sjanse for å vende opp. Modeller som er ikke-deterministiske og inneholder forestillinger om usikkerhet og variasjon, kaller vi *stokastiske modeller*. Ordet stokastisk er avledet av det greske ord for tilfeldighet.

Typisk for mange fenomener i praksis er usikkerhet og variasjon, enten ved at vi har begrenset informasjon, eller ved at det fenomen vi studerer selv varierer. I denne sammenheng skal modeller blant annet hjelpe oss til å

- finne sammenhenger som vi ikke så tidligere,
- relatere virkninger og mulige årsaker,
- forstå variasjon,
- gjøre usikkerhetsvurderinger.

Mange praktiske og teoretiske problemer innenfor en rekke fagområder lar seg best analysere ved hjelp av stokastiske modeller. Dette betyr ikke at vi aksepterer at naturen spiller hasard, men at en stokastisk modell kanskje er det eneste av praktisk nytte. Stokastiske modeller er i dag et viktig verktøy i mange vitenskaper såvel som i teknologi og økonomisk virksomhet. Heldigvis er en stor del av den grunnleggende teori for slike modeller felles for anvendelser innen de fleste fag- og anvendelsesområder, og denne boka gir innblikk i sentrale elementer i slik teori.

1.2 Sannsynlighet

Sannsynlighetsregning er en gammel disiplin, den første tid motivert ut fra ønsket om å analysere veddemål knyttet til myntkast, terningsspill, kortspill etc. Det er kjent at Fermat, Pascal og Huygens syslet med sannsynlighetsbegrepet på midten av 1600-tallet, Cardano allerede hundre år før. To milepeler i klargjøring av sannsynlighetsbergrepet var utgivelsen av Jacob Bernoullis arbeid “Ars conjectandi” (Kunsten å gjette) i 1713 og deMoivres “Doctrine of chances” i 1718. I 1763 lanserte Thomas Bayes et subjektivistisk sannsynlighetsbegrep i sitt “Essay towards solving a problem in doctrine of chance”.

Etterhvert dukket behovet for å snakke om og regne med sannsynligheter opp på stadig flere områder. Dette krevde et bedre teoretisk fundament. Russeren Kolmogorovs aksiomatiserte sannsynlighetsbegrepet i et arbeid fra 1933, og dette ga støtet til intens utvikling av sannsynlighetsteori på resten av 1900-tallet. Parallelt med dette fikk subjektivistiske idéer sin renessanse anført av Ramsey, De Finetti og Savage med flere.

La oss ved hjelp av noen enkle eksempler prøve å nærme oss sannsynlighetsbegrepet på rent intuitivt grunnlag.

Eksempel 1 : Myntkast
Vi kaster et kronestykke og observerer enten kron eller mynt. Spørsmål: Hva er sjansen for henholdsvis kron og mynt? De aller fleste vil vel svare at sjansen for hvert av utfallene er 1/2. Noen vil kanskje uttrykke seg annerledes og si at sjansen er "fifty-fifty" eller oddsene er 1 : 1, men dette er i realiteten det samme. La oss avkreve begrunnelser: Person A begrunner sitt utsagt med å si at hvert av de 2 mulige utfall må ha samme sjanse for å inntreffe, og sjansen for kron må derfor være lik 1/2. Person B begrunner sitt utsagn med å si at dersom en knipser mynten et stort antall ganger, vil kron opptre i omtrent halvparten av forsøkene. A sier at han har benyttet sin teoretiske abstraksjonsevne vedrørende konstruksjonen av et kronestykke, mens B sier at hun har brukt sin praktiske erfaring i myntkast. De to begrunnelsene er derfor fundamentalt forskjellige, men begge kan aksepteres. De fleste vil vel si at det skulle være unødvendig å gjøre noen praktiske eksperimenter i det hele tatt for å kunne konkludere at sjansene er "fifty-fifty".

De første famlende forsøk på å definere et sannsynlighetsbegrep tok derfor utgangspunkt i idéen om like mulige utfall (A's argument). Dette var tilfredsstillende så lenge en bare var interessert i å anvende det på spill, men allerede neste eksempel viser at et slikt sannsynlighetsbegrep er for snevert.

Eksempel 2 : Kast med tegnestift
Ta en tegnestift og kast den ut på bordet. Spørsmål: Hva er sjansen p for at den faller til ro med spissen opp? Vi ser at det også er to mulige utfall av dette eksperimentet. Argumentet A benyttet i Eksempel 1 lyder nå: De to utfallene er like sannsynlige, ergo er sjansen 1/2 for at spissen peker opp. Dette argumentet har vel liten appell her, de fleste vil nok avvise det. Det er ikke lett å se hvordan vi ved å granske tegnestiftens konstruksjon, skal kunne si noe om sjansene for de to utfallene. For å skaffe seg litt innsikt i tegnestiftens egenskaper, tenker vi oss at A og B hver for seg har kastet tegnestiften n=100 ganger. La oss betegne resultatet at stiften faller til ro med spissen opp med r_1 og resultatet at den faller til ro med spissen ned med r_2. A har observert $n(r_1) = 47$, $n(r_2) = 53$, mens B har observert $n(r_1) = 42$, $n(r_2) = 58$ ($n(\cdot)$ betyr "antall ganger"). Den relative hyppighet av hvert av resultatene ble

For A:

$$h(r_1) = \frac{n(r_1)}{n} = \frac{47}{100} = 0.47 \ , \ h(r_2) = \frac{n(r_2)}{n} = \frac{53}{100} = 0.53$$

For B:

$$h(r_1) = \frac{n(r_1)}{n} = \frac{42}{100} = 0.42 \ , \ h(r_2) = \frac{n(r_2)}{n} = \frac{58}{100} = 0.58$$

Dersom andre kaster den samme tegnestiften et stort antall ganger (f.eks. 100), ville hver trolig observere relative hyppigheter nær opp til disse. Hvilken erfaring kan en så i fall trekke av våre eksperimenter? :

- Dersom vi utfører en rekke forsøksserier med mange forsøk i hver serie, så vil den relative hyppighet av hvert forsøksresultat være omtrent den samme i hver av forsøksseriene såframt forsøksbetingelsene ellers er de samme.

På bakgrunn av sine erfaringer uttalte A at sjansen for at tegnestiften faller til ro med spissen opp er $p = 0.47$, mens B påsto at sjansen var mindre, nemlig $p = 0.42$. For ikke å bli uvenner på grunn av en tegnestift, ble A og B enige om å kaste stiften 300 ganger til i fellesskap. Det viste seg at spissen pekte opp i 225 av de i alt $n = 500$ kastene, de fikk altså

$$h(r_1) = \frac{n(r_1)}{n} = \frac{225}{500} = 0.450 \ , \ h(r_2) = \frac{n(r_2)}{n} = \frac{275}{500} = 0.550$$

På bakgrunn av sine samlede erfaringer anslår A og B at sjansen for at spissen peker opp er $p = 0.45$. Dersom de fortsatte å kaste tegnestiften, er det lite trolig at det noen gang blir nødvendig å endre dette anslaget i særlig grad. Denne p oppfatter en derfor som en egenskap ved tegnestiften. Selv om den valgte verdi av p er funnet som en relativ hyppighet for et stort antall gjentakelser av eksperimentet, tyder all erfaring på at denne verdien også har relevans for et rasjonelt menneskes holdning til et lite antall kast, sogar et enkelt kast. Den "sanne verdi av p" vil aldri kunne bli eksakt kjent. Likevel viser det seg fruktbart å bygge modeller basert på forestillingen om en slik verdi. En nærmere analyse som vi vil bli i stand til å utføre senere vil vise at vi, selv med $n = 500$, ikke kan være trygg for at siste sifferet i $p = 0.45$ vil være "korrekt", n er faktisk ikke stor nok til det.

Enkelte forsøker å definere sannsynligheten p for et bestemt utfall av et eksperiment som grenseverdien for den relative hyppighet av utfallet i en serie av n uavhengige gjentakelser av eksperimentet, under de samme betingelser, når n går mot uendelig. Neste eksempel viser at dette er for snevert.

Eksempel 3 : Fødsel og død
Eva skal føde for første gang, og spør seg selv om sjansene for henholdsvis

gutt eller pike. Hun har hittil forestilt seg at sjansene er "fifty-fifty", men leser en dag i avisen av hyppigheten av guttefødsel og pikefødsel i Norge for tiden er henholdsvis 0.514 og 0.486. Disse hyppighetene kan ikke uten videre betraktes som en egenskap ved Eva, hun utfører jo sitt "eksperiment" for første gang (sml. tegnestifteksperimentet). Likevel mener Eva at tallene også er relevant for henne, og andre i samme situasjon. Med litt velvillighet vil en kanskje si at tallene ovenfor kan tolkes som relative hyppigheter i et stort antall gjentakelser av samme eksperiment, og at de uttrykker en egenskap ved den kjønnsbestemmende mekanisme hos mennesker. Som i forrige eksempel kan det være naturlig å definere sannsynligheter som idealiserte relative hyppigheter. Imidlertid ser vi at en slik definisjon innebærer visse vansker. For det første er det umulig å gjenta dette eksperimentet et ubegrenset antall ganger. Vi får også problemer med hvor langt vi kan tøye begrepet samme eksperiment. Tenk bare igjennom følgende: Hva er sjansen p for at den nyfødte overlever sin 70-årsdag? En rekke faktorer som helsetjeneste, kosthold og livsstil varierer, og dette gjør definisjonen vanskelig. Likevel er det uhyre fruktbart å tenke seg modeller med forestillinger om en "sann" p, og det finnes statistiske metoder som kan gi brukbare anslag for denne p, som kan gi grunnlag for fornuftige beslutninger.

Eksempel 4 : Snø og ski
Ski er en kjær julegave for småbarn, også i Bergen. Bergenske sportshandlere har erfart følgende problem : Dersom det faller snø før jul, resulterer dette i voldsom etterspørsel etter barneski i ukene før jul. Dersom julesnøen uteblir, er etterspørselen minimal, og heller ikke når snøen kommer vil salget nå de store høyder. Anta at bestillinger til fabrikkene må være inne før 1. august. Problemet for vår sportshandler vil være hvor mange skipar han bør bestille. Et naturlig spørsmål å stille seg da er: Hva er sjansen for julesnø i år? For å svare på dette spørsmålet kan sportshandleren selvsagt ty til nedtegnede data over nedbørforholdene i Bergen før jul tidligere år. Det kan også være aktuelt å gjøre bruk av andre opplysninger, for eksempel langtidsvarsler fra Meteorologisk institutt.

Eksempel 5 : Bookmaking
En bookmaker tilbyr veddemål omkring utfallet av cupfinalen i fotball mellom Lyn og Brann. Anta at han er kommet fram til at odds er 3 til 1 i favør av Brann. Dette betyr at han tror at sjansen for at Brann vinner er 3/4, mens sjansen for at Lyn vinner er 1/4. For å komme fram til dette resultat har han vurdert lagenes prestasjoner i de siste kampene de har spilt, de siste innbyrdes oppgjør. Videre har han tatt i betraktning lagenes formkurve, an-

tall menn på sykelisten, lagenes lagmoral for øyeblikket. Det er ikke grenser for momenter som kan trekkes inn, og en bookmakers vurdering av odds vil derfor være svært subjektiv.

I alle disse 5 eksemplene har det vært naturlig å snakke om sjanser eller sannsynligheter. Men vi har ikke kommet til klarhet i hvordan et fruktbart sannsynlighetsbegrep bør defineres, spesielt ikke dersom vi ønsker at slike problemer som er nevnt i Eksempel 4 og 5 skal falle innenfor rammen av begrepet. Det har ofte stått strid om hvor langt en kan tøye sannsynlighetsbegrepet. Noen forfekter et subjektivt eller personlig sannsynlighetsbegrep, dette i motsetning til forestillingen om at en sannsynlighet bør være noe objektivt ved det fenomen som studeres, uavhengig av personen som beregner denne sannsynligheten. Begge syn har sine fordeler og ulemper, men vi skal ikke ta stilling her. Vi skal innføre sannsynlighetsbegrepet aksiomatisk, dvs. vi stiller opp de minstekrav som må være oppfylt for at den struktur vi arbeider med skal kunne kalles en sannsynlighetsmodell. Dette teorigrunnlaget legges Kapittel 2, og utvikles videre i etterfølgende kapitler.

1.3 Statistikk

Dersom man ber en statistiker av i dag om å gi en kort definisjon av sitt fag, er det mulig at svare er:

Statistikk = Idéer og metoder for å forstå variasjon

I en norsk fremmedordbok leser vi imidlertid at statistikk er:

1. Systematisk innsamling og tabellering av tallmessige forhold.
2. Vitenskap om belysningen av visse masseforeteelser i samfunnet bygd på tallmessige oppgaver (f.eks. fødsler, dødsfall, forbrytelser osv.).

Dette samsvarer kanskje med hva folk legger i ordet statistikk, men la oss se hvordan ordet etter hvert har fått en videre betydning.

Går vi tilbake i historien for å finne opprinnelsen til fagfeltet statistikk, nevnes ofte et arbeid av engelskmannen John Graunt som ble publisert i 1662 med tittelen “Natural and political observations ... made upon the bills of mortality”. På 1700 og 1800 tallet ble det i flere land vakt en interesse for å foreta systematisk innsamling og bearbeiding av data om befolkning, handel, jordbruk osv., eller kort og godt om rikets tilstand. Ordet statistikk er da også avledet av det latinske ord for tilstand (sml. det engelske "state"). Også i dag er dette en viktig del av det statistiske arbeidsfeltet, og fagfolk kaller dette for *deskriptiv* (eller *beskrivende*) *statistik*k.

Det har imidlertid ingen hensikt å samle inn data med mindre de brukes til noe. Etter hvert ble statistikerne interessert i generelle prinsipper for tolkning av statistiske data, om hvordan slike data kan benyttes til å slutte noe ut over dataene selv, ofte kalt *statistisk inferens*. Statistikk utviklet seg som eget fag *teoretisk statistikk*. De første forløpere til moderne statistisk teori finner vi hos matematikerne Gauss og Laplace på begynnelsen av 1800-tallet. Disse laget en teori for analyse av målefeil ved fysiske observasjoner. Det neste avgjørende puff kom i de første tiår etter 1900. Sentrale personer i pionértiden da statistikk utviklet seg som eget fag var Karl Pearson, Ronald A. Fisher og Jerzy Neyman. Utviklingen skjøt raskt fart, og den skjedde i kontakt med såvel matematikken, som naturvitenskapene og samfunnsvitenskapene. Fra 1950-årene begynte mange statistikere å oppfatte fornålet med sitt fag som det å etablere prinsipper for beslutning under usikkerhet. I dag er få så ambisiøse, men ønsker en å fremheve dette aspektet brukes gjerne betegnelsen *statistisk beslutningsteori*.

Statistikkfaget har vært i kontinuerlig utvikling gjennom hele 1900-tallet, med tilfang av stadig nye metoder. Praktiske problemer fra jordbruk, industri og økonomi ledet ofte til ny teori som alle kunne dra nytte av, og som derfor har blitt felleseie for mange ulike fagområder og anvendelser. Statistikk som fag er blitt kalt "The servant of all sciences" og "The accounting discipline for scientific credibility".

Moderne statistisk teori er i hovedsak basert på stokastiske modeller og sannsynlighetsregning. Mange aktuelle analysemetoder er uten slik teoretiske forankring. Det gjelder spesielt metoder for såkalt *eksplorativ dataanalyse*, ved slutten av 1990-tallet markedsført under merkelappen "data mining". Dette felt er ikke dekket i denne boken.

De følgende eksempler tar sikte på å gi leseren et innblikk i noen sentrale statistiske problemstillinger. Vi håper at disse kan gi leseren en viss motivasjon for å gi seg i kast med statistisk teori, samt forstå nødvendigheten av basiskunnskaper i sannsynlighetsregning. Lesere som primært er interessert i sannsynlighetsregning kan gå direkte løs på Kapittel 2, og eventuelt heller lese dette stoffet som en opptakt til Kapittel 7 og 8.

Eksempel 6 : Måleproblemer - tilfeldig variasjon
En bedrift er interessert i kvaliteten av artikler produsert etter en nyutviklet metode, og som et eksperiment prøveproduseres i alt n=50 artikler, hvoretter kvaliteten måles. Kvalitet kan f.eks. være den belastning i kilo som artiklen

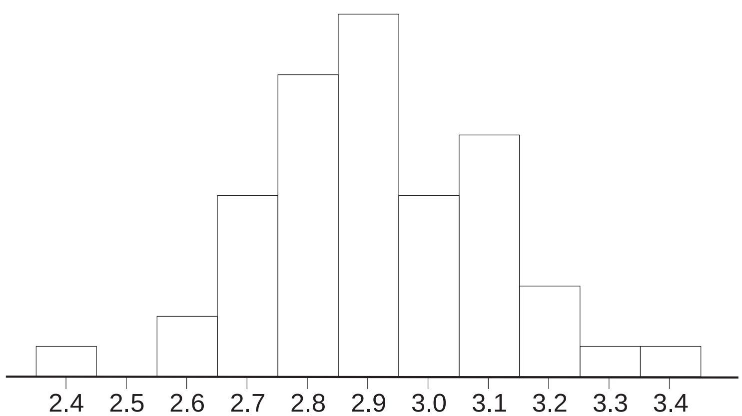
tåler inntil den går i stykker. Resultatet ble (med en desimal)

2.9	2.7	3.1	2.8	2.8	2.7	2.9	3.0	2.6	3.1
3.1	3.0	2.8	2.9	2.7	2.9	3.0	2.6	3.2	2.9
2.7	2.8	3.1	3.1	2.9	2.9	2.8	3.4	3.0	2.8
3.0	2.9	2.7	3.2	2.8	3.1	2.4	2.9	3.2	2.9
3.1	3.3	2.9	2.9	2.8	2.7	2.8	2.8	3.1	3.0

For å skaffe oversikt over dette tallmaterialet lages en *hyppighetstabell*

Verdi	2.4	2.5	2.6	2.7	2.8	2.9	3.0	3.1	3.2	3.3	3.4
Hyppighet	1	0	2	6	10	12	6	8	3	1	1
Relativ hyp.	0.02	0.00	0.04	0.12	0.20	0.24	0.12	0.16	0.06	0.02	0.02

Tabellen gir utrykk for en kvalitetsfordeling: kvaliteten kan variere, men noen verdier er mer typiske enn andre. Mer instruktivt er å lage et *histogram*, se Figur 1.2. Her er tegnet søyler svarende til hver av kategoriene i hyppighetstabellen, med areal proporsjonalt med de respektive hyppigheter, se Figur 1.2.



Figur 1.2: Histogram

En størrelse av betydelig interesse vil være den gjennomsnittlige kvalitet, som i dette tilfellet er 2.914. I tillegg til gjennomsnittet ønskes ofte et mål for variasjonen i kvalitet. Et slikt mål er variasjonsbredden, som her er 1.0. Et alternativt mål er standardavviket (defineres nedenfor), som her beregnes til 0.191. Dette gir i større grad enn variasjonsbredden uttrykk for et "typisk" avvik fra gjennomsnittet.

La oss generelt tenke oss et observasjonsmateriale bestående av n observasjoner

$$X_1, X_2, X_3, \ldots, X_n$$

Gjennomsnittet (middeltallet) av observasjonene er da gitt ved

$$\overline{X} = \frac{1}{n}(X_1 + X_2 + X_3 + \cdots + X_n) = \frac{1}{n}\sum_{i=1}^{n} X_i$$

mens *standardavviket* til observasjonene er gitt ved

$$S_X = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2}$$

S_X^2 kalles ofte for *variansen* til observasjonene. Presentasjon av hyppighetstabeller, histogrammer, gjennomsnitt og standardavvik osv. tar sikte på å få fram typiske trekk ved et tallmateriale. Disse og andre grafiske og numeriske presentasjonsformer faller under betegnelsen beskrivende statistikk.

I mange situasjoner er vi interessert i å gjøre noe mer enn bare å beskrive tallmaterialet, vi ønsker kanskje å bruke resultatene i planlegging, som grunnlag for prognoser eller beslutninger. Da melder det seg en rekke spørsmål, og la oss diskutere noen av disse med utgangspunkt i eksemplet ovenfor:

Vi ønsker typisk å uttale oss om kvaliteten av artikler dersom vi fortsetter å produsere etter samme metode. I denne forbindelse er det naturlig å tenke seg at metoden generelt kan tillegges en "sann" kvalitetsfordeling, dvs. at "i det lange løp" er andelen av de ulike kvaliteter bestemt av metoden selv. Det viser seg fruktbart å tenke på disse andelene som sannsynligheter, og kvalitetsfordelingen som en såkalt *sannsynlighetsfordeling*. Slike sannsynligheter kan danne utgangspunkt for ulike beregninger, og for dette formål kan vi ha nytte av å teori for sannsynlighetsregning. Sannsynligheter kan blant annet brukes til å si noe om kvaliteten ved produksjon av et gitt antall artikler, f.eks. sjansen for at et parti på 10 artikler ikke inneholder artikler under en viss kvalitet. Slik sannsynlighetsregning forutsetter at de "sanne" sannsynlighetene er kjente. I praksis er dette sjelden tilfelle, og spørsmålet er da hvordan vi kan bruke informasjonen fra prøveproduksjonen til å trekke konklusjoner om de teoretiske sannsynlighetene.

En problemstilling som har spesiell interesse er følgende: Vi tenker oss at produksjonsmetoden kan karakteriseres ved et kvalitetsnivå μ som gir uttrykk for gjennomsnittlig kvalitet "i det lange løp". Hver observasjon i prøveproduksjonen oppfattes som en observasjon av dette nivå, men som pga. tilfeldigheter kan avvike noe fra dette, såkalt *tilfeldig variasjon.* Sannsynlighetsfordelingen kan brukes som et uttrykk for sjansene for at en enkelt observasjon faller i de ulike kvalitetskategorier, og kvalitetsnivået μ oppfattes som *forventet* kvalitetsnivå. I praksis vil forventningen μ ofte være ukjent, og formålet med å foreta observasjoner, er bl.a. å skaffe informasjon om denne. Svært ofte blir middeltallet $\overline{X}$ brukt som anslag på forventningen μ.

Av spørsmål som en kunne ønske å få besvart nevner vi: Hvilke feilmarginer må man regne med dersom vi anslår forventningen på denne måten? Her hadde vi femti observasjoner, men hva dersom vi hadde bare fem? Hvor mange observasjoner må vi innhente dersom vi ønsker en viss feilmargin? Gir observasjonene grunnlag for å påstå at forventningen er mindre enn (større enn) 3.0? Hva er sjansen for feilaktig å påstå noe slikt osv. For å besvare spørsmålene trengs statistisk teori. Vi formulerer da de antakelser vi er villig til å gjøre om observasjonene i en statistisk modell, og utleder våre konklusjoner ut fra denne. En aktuell modell i denne forbindelse er den såkalte *målemodellen* som blir diskutert i Kapittel 8. En del grunnleggende begreper som trengs blir utviklet i Kapitlene 3-6.

Eksempel 7 : Stikkprøver - pålitelighet

Vi ønsker å finne ut noe om tilbøyeligheten til å bære på kontanter for en gruppe studenter. Det er tungvint å innhente opplysninger fra alle studentene (populasjonen). I steden trekkes en stikkprøve (et utvalg) på et mindre antall studenter, og vi ønsker å trekke konklusjoner om tilbøyeligheten blant alle studentene på grunnlag av opplysninger fra utvalget. Hvilken pålitelighet kan vi tillegge generelle konklusjoner basert på slike stikkprøver?

La oss se på en tenkt situasjon der populasjonen består av et kull på N=216 studenter. Anta at kronebeløpene til disse studentene er (rekkefølgen er likegyldig) :

80	93	65	328	10	28	113	476	57	54	13	21
232	56	49	35	5	$\underline{0}$	280	37	376	24	36	51
26	155	62	237	532	19	56	112	40	136	53	263
97	206	4	286	134	8	$\underline{145}$	500	164	41	178	$\underline{285}$
91	113	262	47	6	188	48	26	29	45	94	100
620	79	306	59	55	136	29	175	113	35	40	383
17	452	110	8	109	$\underline{273}$	114	46	170	481	38	719
25	0	381	139	153	36	242	59	78	72	380	87
192	57	123	81	13	30	$\underline{0}$	51	184	98	0	$\underline{312}$
771	27	40	99	90	206	214	12	317	119	152	11
0	19	113	59	31	37	22	14	73	310	81	341
17	86	63	7	212	45	110	72	637	46	118	26
$\underline{51}$	312	59	86	110	145	74	$\underline{8}$	$\underline{411}$	71	450	23
176	253	21	98	43	36	145	60	427	78	83	85
119	9	276	34	12	180	277	185	27	13	412	64
53	53	78	291	85	578	370	479	85	0	110	174
$\underline{7}$	176	130	109	49	345	49	156	111	78	91	373
55	222	91	846	19	64	62	274	318	185	200	502

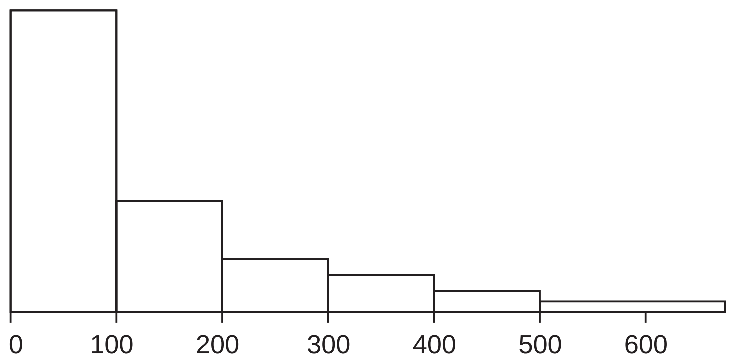
Disse tallene er i utgangspunktet ukjente for oss, men la oss likevel se litt nærmere på dem, for å forstå hva som kan hende ved analyse av stikkprøver.

La oss først se på hyppighetsfordelingen av pengebeløpene. Siden antall ulike beløp som forekommer er svært stort, får vi bedre oversikt dersom vi grupperer beløpene , f.eks. slik

Beløp	0-99	100-199	200-299	300-399	400-499	500-
Hyppighet	121	43	19	16	8	9
Relativ hyp.	0.56	0.20	0.09	0.07	0.04	0.04

Vi kunne alternativt bruke en finere gruppering, men den vi har valgt fanger likevel opp de karakteristiske trekk ved materialet. Et histogram vil i denne situasjonen se ut som i Figur 1.3.

Vi ser at denne fordelingen er nokså ulik den i foregående avsnitt, den er skjev med hovedmassen av kronebeløpene mellom 0 og 100. Gjennomsnittlig kronebeløp beregnes her til 144.20, mens standardavviket er beregnet til 155.20. Det gjennomsnittlige beløp er altså langt over det beløp som majoriteten har på seg, som skyldes at enkelte studenter bærer på relativt store beløp. Vi ser at betydelig variasjon mellom studentene innbyrdes kommer til uttrykk ved et relativt høyt standardavvik. I en slik situasjon vil alternative beskrivende mål ha interesse. Et slikt er *medianen*, som er det kronebeløp



Figur 1.3: Histogram grupperte data

som er slik at like mange studenter har mer og mindre enn dette beløp. I tallmaterialet ovenfor er kr. 85.50 en median.

La oss nå tenke oss at situasjonen er slik at tallene ovenfor alle er ukjente for oss, og at vi ønsker å finne ut noe om tilbøyeligheten til å bære med seg penger ut fra en stikkprøve på, la oss si n=10 tilfeldig utvalgte studenter. Anta at studentene er nummerert fra 1 til 216 (rekkefølgen er likegyldig), og at vi ved hjelp av en upartisk mekanisme trekker ut ti studenter (vi kan f.eks. gjøre bruk av en terning). Resultatet ble de studenter som er understreket i tabellen, og deres kronebeløp er:

Beløp :	0	145	285	273	0	312	51	8	411	7

Anta at vi primært er interessert i gjennomsnittlig pengebeløp blant alle studentene. Det synes rimelig å bruke gjennomsnittet i utvalget som anslag på gjennomsnittet i hele populasjonen. Med vårt utvalg beregner vi gjennomsnittet til 149.2, og spørsmålet er i hvilken grad dette gir pålitelig informasjon om det virkelige gjennomsnittet.

At usikkerheten ved en stikkprøve av den gitte størrelsesorden kan være betydelig, kan bekreftes ved å ta en ny stikkprøve på 10 studenter. Denne gangen fikk vi et gjennomsnitt i utvalget på 107.5. Det faktum at vårt opprinnelige anslag 149.2 er nær gjennomsnittet i populasjonen 144.2 beror derfor på flaks, det er betydelig risiko for at stikkprøven gir et anslag som er mye mer avvikende.

I praksis vet vi jo ikke hva gjennomsnittet eller medianen i populasjonen er, og vi kan da stille spørsmål av typen: Hva er sjansen for at et anslag basert på en stikkprøve er høyst 50 kroner feil? etc. Det viser seg at svaret på slike spørsmål vil avhenge av stikkprøvens størrelse, populasjonens størrelse og struktur, spesielt har variabiliteten av verdiene i populasjonen betydning. For å vurdere hvordan alt dette påvirker påliteligheten av stikkprøver, trenger vi teoretiske kunnskaper. En teoretisk modell i denne forbindelse er den såkalte *lotterimodellen* som beskrives i Kapittel 8. Den forutsetter at stikkprøven er et såkalt *tilfeldig utvalg*, et begrep som tas opp i Kapittel 3. Mer omfattende teorier for utvalgsundersøkelser tas opp i Kapittel 14 og 15.

Eksempel 8 : Samvariasjon - korrelasjon
I mange statistiske undersøkelser foreligger sammenhørende observasjoner av to eller flere variable, og formålet er å klargjøre en eventuell sammenheng mellom disse. Vi vil her ta utgangspunkt i et eksempel med to variable:

Ved avsluttende siviløkonomeksamen ved NHH våren 1967 var det 91 studenter oppe til eksamen i samfunnsøkonomi (S) og bedriftsøkonomi (B). Karakterene var (karaktersystem: 0 til 9, med 9 som beste karakter):

S	B	S	B	S	B	S	B	S	B	S	B	S	B
8	6	5	5	5	4	6	5	5	8	7	7	6	6
3	5	5	6	5	6	3	6	3	6	7	5	7	5
7	6	6	4	4	4	5	4	3	6	4	5	6	6
4	4	3	4	6	6	8	6	6	5	6	5	7	6
5	4	3	4	7	6	5	5	4	5	4	4	6	6
4	6	8	6	5	5	7	5	4	4	4	4	5	6
5	5	5	7	3	6	7	5	5	6	5	6	4	3
6	3	8	7	6	5	6	7	8	8	6	6	7	5
3	3	6	5	3	6	7	3	9	5	5	4	4	5
7	5	6	5	3	6	3	5	4	5	4	6	4	4
3	5	5	5	4	4	7	8	5	6	8	5	4	6
3	4	4	6	7	3	7	8	4	4	5	5	5	5
4	3	7	5	7	5	5	6	4	2	3	6	1	4

En bedre oversikt over materialet får vi dersom vi lager en todimensjonal hyppighetstabell. Denne uttrykker den *simultane hyppighetsfordeling* av samfunnsøkonomi og bedriftsøkonomikarakter for denne gruppen studenter. Summeres linjevis og kolonnevis får vi de *marginale hyppighetsfordelinger* for henholdsvis samfunns- og bedriftsøkonomi.

Samf.-	Bedriftsøkonomi										
øk.	0	1	2	3	4	5	6	7	8	9	Sum
0											0
1					1						1
2											0
3				1	3	3	7				14
4			1	2	8	4	4				19
5					4	7	7	1	1		20
6				1	1	6	5	1			14
7				2		8	3	1	2		16
8						1	3	1	1		6
9						1					1
Sum	0	0	1	6	17	30	29	4	4	0	91

Hyppighetstabellen gir et visst inntrykk av (positiv) samvariasjon mellom de to karakterene, studenter med god karakter i samfunn har også gjennomgående bra karakter i bedrift og omvendt. På den annen side finnes også enkelte "spesialister". Vi kunne ønske oss et numerisk mål for graden av samvariasjon mellom de to karakterene i tallmaterialet.

La oss først innføre den nødvendige notasjon: Vi tenker oss generelt n observasjonspar

$$(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$$

Vi kan da beregne $\overline{X}$, S_X og $\overline{Y}$, S_Y som er gjennomsnittet og standardavviket for henholdsvis X-observasjonene og Y-observasjonene. Som mål for samvariasjonen mellom de sammenhørende X og Y-observasjoner kan brukes den såkalte *kovariansen* mellom X-observasjonene og Y-observasjonene gitt ved

$$S_{XY} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}) \cdot (Y_i - \overline{Y})$$

Et mer hensiktsmessig mål er imidlertid *korrelasjonskoeffisienten* gitt ved

$$R_{XY} = \frac{S_{XY}}{S_X \cdot S_Y}$$

Det kan vises at vi alltid har at

$$-1 \leq R_{XY} \leq 1$$

En positiv R_{XY} uttrykker positiv samvariasjon, dvs. de store (små) verdiene av Y hører gjennomgående sammen med de store (små) verdiene av X. En negativ R_{XY} utrykker det omvendte. Graden av samvariasjon (egentlig lineær samvariasjon) er større jo større tallverdien av R_{XY} er.

I vårt eksempel lar vi (X_i, Y_i) være karakteren i samfunnsøkonomi og bedriftsøkonomi for den i'te student, i=1,2, . . . ,91.

Gjennomsnitt:	$\overline{X}$	=	5.18	$\overline{Y}$	=	5.18
Standardavvik:	S_X	=	1.61	S_Y	=	1.19
Kovarians:	S_{XY}	=	0.55			
Korrelasjonskoeffisient:	R_{XY}	=	0.29			

Som ventet er korrelasjonskoeffisienten positiv, men kanskje ikke så stor som vi hadde ventet oss. Til sammenligning kan det nevnes at korrelasjonskoeffisienten mellom gymnaspoeng og hovedkarakter til siviløkonomeksamen for de samme studentene var 0.51.

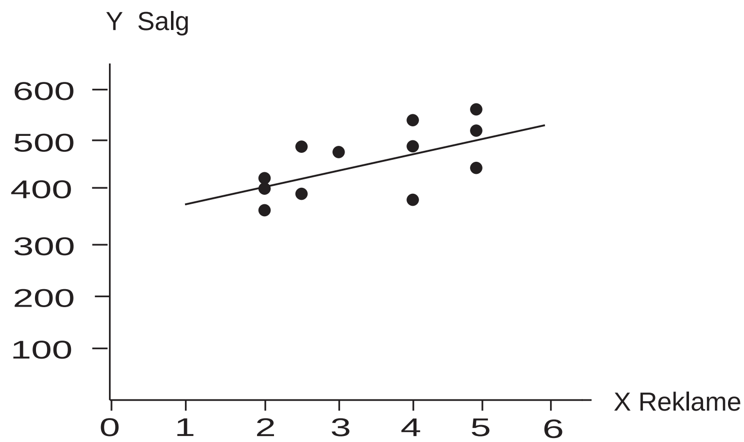
Slike korrelasjonskoeffisienter kan f.eks. brukes til å sammenligne graden av samvariasjon for etterfølgende studentkull. De foretatte beregninger er her bare av beskrivende karakter. I enkelte sammenhenger brukes korrelasjonsberegninger til å trekke generelle konklusjoner, f.eks. om en populasjon på grunnlag data fra et utvalg. Da trengs statistisk teori for å vurdere påliteligheten, samt å forstå de begrensninger som er knyttet til korrelasjonsanalyse. Dette blir tatt opp i Kapittel 9.

Eksempel 9 : Regresjon - prediksjon
En bedrift har observert sammenhørende verdier av reklameinnsats X (i tusen kroner) og solgt kvantum Y i ulike etterfølgende perioder. Resultatet ble :

Periode:	1	2	3	4	5	6	7	8	9	10	11	12
X:	4.0	2.0	2.5	2.0	3.0	5.0	4.0	2.0	5.0	4.0	2.5	5.0
Y:	385	400	395	365	475	440	490	420	560	525	480	510

Disse observasjonene kan oversiktlig presenteres i et såkalt *spredningsdiagram* som i Figur 1.4.

Vi ser at reklameinnsatsen nok påvirker solgt kvantum, men at det er betydelig variasjon som må skyldes andre forhold (tilfeldighet?). Spredningsdiagrammet innbyr til å bruke en rett linje $Y = a + b \cdot X$ til å beskrive



Figur 1.4: Spredningsdiagram

hovedtendensen i tallmaterialet. Istedenfor å tegne linjen på øyemål, beregnes ofte den såkalte *minste kvadraters regresjonslinjen* for Y mhp. X. Denne linjen er lagt slik at kvadratsummen av alle vertikale avstander fra punktene til linjen er minst mulig, og kan uttrykkes ved

$$Y = \overline{Y} + R_{XY} \cdot \frac{S_Y}{S_X}(X - \overline{X})$$

Den kan beregnes straks vi kjenner observasjonenes gjennomsnitt, standardavvik og korrelasjon. Vi ser at linjen er stigende eller fallende alt etter som R_{XY} er positiv eller negativ. Er R_{XY} nær null gir dette uttrykk for at graden av lineær samvariasjon er liten. For øvrig merker vi oss at linjen går gjennom punktet $(\overline{X}, \overline{Y})$, og at stigningstallet til linjen er

$$b = R_{XY} \cdot \frac{S_Y}{S_X} = \frac{S_{XY}}{S_X^2}$$

som gjerne kalles *regresjonskoeffisienten* for Y mhp. X. I vårt tallmateriale finner vi $\overline{X} = 3.42, \overline{Y} = 453.75, S_X = 1.17, S_Y = 59.34$ og $R_{XY} = 0.635$, slik at minste kvadraters regresjonslinjen blir

$$Y = 453.75 + 32.2 \cdot (X - 3.42) = 343.63 + 32.2 \cdot X.$$

Eksempelvis blir "beregnet" solgt kvantum for reklameinnsats $X = 3.0$ og $X = 6.0$ henholdsvis $Y = 440.2$ og $Y = 536.8$. Regresjonskoeffisienten blir her et uttrykk for endringen i solgt kvantum ved å øke annonseutgiften med

1 enhet, dvs. tusen kroner. Minste kvadraters regresjonslinjen for Y mhp. X tar sikte på å belyse "forventet" Y for gitt X, og kan ut over det rent beskrivende tenkes brukt til prognoseformål. Merk at selv om hovedtendensen i materialet er lineær, kan variasjonen omkring linjen være betydelig. Et relevant spørsmål er derfor: Hvilke feilmarginer må vi regne med dersom vi trekker generelle slutninger basert på minste kvadraters regresjonslinjen, f.eks. bruker den til å predikere Y for gitt X? Slike spørsmål kan neppe besvares uten at vi gjør en del antakelser om observasjonene: Vi tenker oss en "sann" regresjonslinje som svarer til forventet Y for gitt X. På grunn av tilfeldigheter kan hver observasjon falle utenfor linjen, ovenfor eller nedenfor. Den beregnede minste kvadraters regresjonslinjen er et forsøk på å anslå denne sanne linjen.

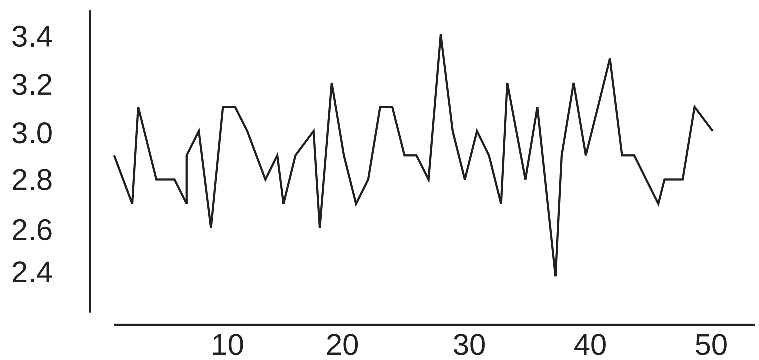
En teori som kan være oss til hjelp i disse spørsmål blir utviklet i Kapittel 8 og videreutviklet i Kapittel 12. Siden tallmaterialet var reklame og salg i etterfølgende perioder, er det også aktuelt å trekke inn tidsaspektet i analysen: Viser salgstallene også en tidstrend i tillegg til reklame-effekten? Dersom f.eks. salgstallene er kvartalsvise data over 3 år, er det noen sesongvariasjoner i salget? Disse spørsmål kan diskuteres i lys av den teori som blir utviklet i Kapittel 12 og 13.

I ingen av eksemplene ovenfor, med unntak av det siste, var rekkefølgen av observasjonene tillagt noen særskilt betydning. I mange situasjoner av praktisk interesse er nettopp rekkefølgen i fokus. Det gjelder i første rekke ved observasjonen av tekniske, økonomiske og sosiale prosesser over tid. Slike datamaterialer kalles gjerne *tidsrekker*.

Eksempel 10 : Tidsrekke - prosesskontroll
En bedrift tar regelmessige stikkprøver fra en produksjonsprosess, med sikte på å overvåke eventuell variasjon i kvalitet. Det tas 10 prøver hver dag, som hver gir en målt kvalitet. Denne kan plottes i et *tidsplott* straks målingen foreligger. Etter en uke har vi derfor n=50 observasjoner. Anta for enkelhets skyld at disse er gitt som i Eksempel 6, der hver linje inneholder en dags observasjoner i naturlig rekkefølge. Tidsplottet ser da ut som i Figur 1.5.

Plottet gir inntrykk av at kvaliteten varierer tilfeldig omkring nivået 2.9 gjennom hele uken. Dette fordi ingen dag (eller del av dag) ser ut til å peke seg ut som annerledes enn de andre, og det heller ikke ser ut til at en eller flere observerte gode (evt.dårlige) prøver gir kunnskap om hvor neste prøve plasserer seg i forhold til gjennomsnittsnivået. Merk at dette spørsmål ikke ble avklart i Eksempel 6.

Når prosessen varierer tilfeldig (uavhengighet) under like forhold hele tiden (tidshomogenitet), sier vi at prosessen er i "statistisk kontroll". Forstå-



Figur 1.5: Tidsplott

else av slike prosesser er nødvendig for å analysere mer kompliserte prosesser fra virkeligheten. For å overvåke produksjonssystemer og administrative systemer, trengs metoder til å avgjøre om en prosess er i kontroll eller ikke. For økonomiske tidsrekker trengs kjennskap til ulike former for avhengighet, når en skal lage prognoser og vurdere påliteligheten av disse. Dette er spørsmål som tas opp i Kapittel 13 og 15 påggrunnlag av kunnskaper fra Kapittel 1-8.

Det er ikke alltid lett å gjennomskue tilfeldighetenes rolle ved statistiske undersøkelser og i studiet av prosesser som omgir oss. For å oppnå en grunnforståelse kan simuleringseksperimenter være et nyttig hjelpemiddel. Myntkast er den enkleste form for slik eksperimentering, og er godt egnet til å klarlegge ulike sider ved tilfeldighetsbegrepet, som kan komme oss til nytte også i mer praktiske situasjoner (se Oppgave 14).

Teoretisk statistikk er i hovedsak å lage og utnytte modeller som knytter observasjonene til det fenomen vi er interessert i. Siden vi typisk ønsker å vurdere risikoen for feilkonklusjoner, vil dette som regel være en stokastisk modell. Studium av slike modeller kan også gi forståelse av hva slags observasjoner som bør innhentes, samt en vurdering av ulike analysemetoder. Hva slags modeller som er aktuelle i en gitt situasjon, vil kunne variere betydelig med det problemområde det er tale om, men det er likevel mulig å gi visse basiskunnskaper av almen karakter. Selv om de situasjonene som er tatt som eksempel ovenfor kan synes relativt enkle, kan en neppe vente å finne endelig svar på alle spørsmål innenfor teorirammen i en elementær lærebok. Kunnskaper herfra gjør det imidlertid mulig å orientere seg mot spesiallitteratur.

1.4 Oppgaver

1. Ta for deg en mynt, knips den n=100 ganger og observer antall kron. Er det noen grunn til å anta at den ikke er rettferdig?

2. Ta for deg en tegnestift og kast den ut på bordet gjentatte ganger og observer om den faller til ro med spissen i været eller ikke. Beregn hyppigheten av spiss i været etter n=10,20,30,...,100 kast. Ser det ut til at hyppigheten stabiliseres etter hvert? Tør du uttale deg noe om sjansen for at spissen peker i været ved et enkelt kast? Overlat stiften til en annen og be ham/henne gjenta forsøksserien. Sammenlign de erfaringer som gjøres.

3. Per og Pål kaster to vanlige mynter og vil observere antall kron. De er uenige om sjansen for å få resultatene r_1=ingen kron, r_2=en kron og r_3=to kron. Per foreslår at sjansene er henholdsvis $\frac{1}{3}, \frac{1}{3}, \frac{1}{3}$, mens Pål foreslår $\frac{1}{4}, \frac{1}{2}, \frac{1}{4}$. Hvem støtter du? Hvis du er usikker forsøk med litt myntkasting.

4. Vurder sjansen for regn i morgen. Hvordan skulle et utsagn om at sjansen for regn er 60% eventuelt kunne tolkes?

5. I veiledningen til en meningsmåling om tilslutningen til politiske partier står det at feilmarginer man må regne med er 2-3% hver vei, størst for de største partiene. Hva legger du i denne veiledning?

6. Etter jordskjelvet i San Francisco-området høsten 1989, ble det uttalt at sannsynligheten for et større jordskjelv før år 2000 var 50%. Hva legger du i denne opplysningen? Vil du oppfatte et tilsvarende utsagn om oljeutblåsning i Nordsjøen annerledes?

7. Følg med i en dagsavis i en uke, og let etter ord som sannsynlig og sannsynlighet, samt eventuelle betraktninger knyttet til sannsynligheter. Er det grunn til å tro at ordene er brukt i en mer presis betydning enn den dagligdagse? Les Lottokommentarene i en tabloidavis, for å sjekke om journalisten har en adekvat forståelse av tilfeldighetens rolle.

8. Forsøk pilkast mot en sirkulær skive med poengskala, f.eks. fra 0 til 5. Noter resultatet i hvert av n=30 kast. Lag hyppighetsfordeling og tegn histogram. Beregn gjennomsnitt og standardavvik. Gjenta eksperimentet med en lengre avstand fra skiven. Hva erfarer du?

9. Bruk dataene i Eksempel 7 til å vinne egen erfaring om påliteligheten av stikkprøver. For å velge ut studenter kan du bruke en terning. La f.eks. utfallet av tre etterfølgende kast bestemme nr. på henholdsvis boks, linje og søyle i tabellen. Gjør gjentatte forsøk med ulike stikkprøvestørrelser, f.eks. n=5 og n=10. I tillegg til å anslå gjennomsnittsverdien i populasjonen kan du også prøve å anslå medianverdien i populasjonen, f.eks. ved å bruke medianen i stikkprøven (den midterste når n er oddetall, og gjennomsnittet av de to midterste når n er partall).

10. Spør noen av dine venner (8-10 er nok) om deres høyde (X) og vekt (Y). Beregn korrelasjonskoeffisienten mellom X og Y. Tegn observasjonene i et spredningsdiagram, og legg inn minste kvadraters regresjonslinjen for Y mhp. X. Kommenter.

11. Undersøk tilgjengelige lommeregnere om muligheten for lettvint beregning av gjennomsnitt, standardavvik og eventuelt kovarians (f.eks. med egne taster). Det kan vises at

$$S_X^2 = \frac{1}{n} \sum_{i=1}^{n} X_i^2 - \overline{X}^2$$

$$S_{XY} = \frac{1}{n} \sum_{i=1}^{n} X_i Y_i - \overline{X} \cdot \overline{Y}$$

Kan dette forenkle utregningene i forhold til de opprinnelige formlene?

12. Ved en bedrift er det observert forurensingsgrad for ulike aktivitetsnivåer (f.eks. antall ovner i drift) ved 12 tidspunkter. Resultatet ble :

Aktivitetsnivå :	3	5	2	4	2	1	3	1	2	5	4	3
Forurensingsgrad :	4.7	6.4	3.5	5.4	3.9	2.6	5.9	4.1	5.1	4.6	6.1	3.0

Plott forurensingsgrad mot aktivitetsnivå. Beregn beskrivende mål. Beregn regresjonslinjen, og tegn den inn i plottet. Vurder muligheten for å predikere forurensingsgrad.

13. For en produksjonsprosess er det observert kvalitet for en stikkprøve tatt fra n=50 etterfølgende perioder, jfr. Eksempel 10. Resultatet ble (med en desimal) :

2.9	3.0	3.1	3.0	2.9	3.1	3.4	3.0	3.0	2.9
2.9	3.0	2.9	3.0	3.1	2.7	2.6	2.7	2.7	2.4
2.7	2.6	2.7	2.8	2.7	2.8	2.8	2.8	2.9	2.8
2.8	2.9	2.9	2.9	2.8	2.8	2.8	2.9	2.9	2.8
3.2	3.1	3.1	3.1	3.2	3.3	3.1	2.9	3.2	3.1

Lag et tidsplott og vurder om prosessen er i kontroll. Hvis ikke, tenk over mulige årsaker.

14. Folk legger ofte merke til spesielle resultater i gjentatte eksperimenter, som de tviler på er rene tilfeldigheter, eller handler som om det ikke var det. Eksempler på beslektede situasjoner er:

   - Mange kron på rad i myntkast.
   - Fravær av sekser i terningkast (LUDO)
   - Samme siffer (eller fravær av et siffer) i mange Lottoomganger på rad.

- Mange utskiftninger av lyspærer rett etter hverandre.

Bruk myntkastresultatene fra Oppgave 1 til å undersøke tilfeldigheters rolle ved å studere

(a) hvor ofte lange serier med kron forekommer

(b) hyppigheten av kron etter en kron, to kron osv.

Vil du etter seks røde i rulett (i) satse rødt ("i siget"), (ii) satse sort ("rødt hell er oppbrukt"), (iii) satse likegyldig eller (iv) la være å satse?

15. Nedenfor er gitt en typisk utskrift fra statistisk programvare der beskrivende mål for dataene i Eksempel 7 er beregnet.

```
>> READ 'eks1.7' 'Beloep'
>> DESCRIBE 'Beloep'
           N   Mean  Median  StDev  Min  Max
 Beloep  216  143.7   85.52  155.2    0  846
```

(a) Bruk tilgjengelig programvare til å reprodusere disse resultatene.

(b) Gir din programvare flere beskrivende mål enn de som er med her?

(c) Bruk din programvare til å beregne beskrivende mål for dataene i Eksempel 8, herunder korrelasjonen mellom karakterene i de to fagene.

Oppgaven forutsetter at filer med dataene er tilgjengelige. Hvis dette ikke er tilfelle, løs oppgaven med et utvalg på 10 tilfeldig valgte observasjoner. som leses inn for hånd.

16. Programvare gir typisk mulighet for tilfeldige trekninger, her illustrert med uttrekning av 10 studenter blant populasjonen på 216 i Eksempel 7.

```
>> SET 'Populasjon'; DATA 1:216.
>> SAMPLE 10 from 'Populasjon' put into 'Utvalg'
>> PRINT 'Utvalg'
47 12 118 200 77 33 115 10 127 151
```

(a) Bruk tilgjengelig programvare til å lage en populasjon med studentnumre fra 1 til 216, og trekk så et utvalg på 10 numre.

(b) Hvis vi bare var interessert i beløpene, kan vi isteden trekke fra en populasjon med disse. Hvordan kan du få til det?

Merknad: Kommandoer og utskrifter i denne boken ligner de vi finner i Minitab (som anbefales), men er redigert noe for at leseren lettere skal finne sammenhengen med teksten i boken forøvrig.

Kapittel 2

Diskrete sannsynlighetsmodeller

Dette kapitlet gir en innføring i sannsynlighetsmodeller, grunnleggende begreper og enkle regneregler for analyse basert på slike modeller. [1] For å ha en rød tråd gjennom temaene, vil vi tenke på et eksperiment med usikkert utfall, men idéene kan overføres til iakttakelser i andre sammenhenger.

2.1 Eksperimenter med diskret utfallsrom

> **Definisjon :** Mengden av alle mulige utfall av et eksperiment kalles *utfallsrommet* for eksperimentet.

Vi krever av et utfallsrom at det skal være fullstendig og skillende, dvs. at ett og bare ett av utfallene i utfallsrommet inntreffer når eksperimentet utføres. Vi skal symbolisere et utfallsrom med Ω. Som generelt symbol for utfall bruker vi u. Ω er altså en mengde, og u er element i Ω, vi skriver $u \in \Omega$ hvor symbolet $\in$ betyr "er element i", Dersom antall utfall i utfallsrommet er enten endelig eller tellbart uendelig, sier vi at utfallsrommet er *diskret*. I dette kapitlet skal vi begrense oss til å studere eksperimenter med diskret utfallsrom. I slike tilfeller kan vi liste opp (nummerere) alle de mulige utfallene av eksperimentet, for eksempel slik : $u_1, u_2, u_3, \ldots$ Utfallsrommet er da

[1] Appendiks A gir en oversikt over noe av den terminologi som vil bli brukt, og en ordliste norsk-engelsk.

$$\Omega = \{u_1, u_2, u_3, \ldots\}$$

Når vi skal skrive ned et utfallsrom i en konkret situasjon, bør vi representere hvert utfall med egnede symboler, for eksempel med tall eller bokstaver.

Eksempel 1 : Et myntkast
Et egnet utfallsrom for et myntkast er $\Omega = \{K, M\}$ hvor bokstavene K og M betegner henholdsvis kron og mynt. Nå kunne det selvsagt tenkes at mynten ble stående på høykant. For å fange inn denne muligheten kan en bruke utfallsrommet $\Omega = \{K, M, H\}$, men $\Omega = \{K, M\}$ kan anses som fullstendig for alle praktiske formål.

Eksempel 2 : Et terningkast
Vi kaster en terning og ønsker å observere antall øyne. Et eget utfallsrom er $\Omega = \{1, 2, 3, 4, 5, 6\}$. Tallet 6 representerer 6 øyne etc.

Eksempel 3 : Loddtrekning
Navnene på de m studentene som er til stede på en forelesning skrives ned på m lapper. De m lappene legges i en hatt og blandes godt. En lapp skal trekkes ut tilfeldig, og navnet på lappen registreres. Vi lar utfallsrommet være $\Omega = \{N_1, N_2, \ldots, N_m\}$ hvor $N_1, N_2, \ldots, N_m$ er de m navnene.

Eksempel 4 : Kvalitetskontroll
En produsert artikkel kvalitetskontrolleres og klassifiseres i en av to kategorier intakt (i) eller defekt (d). Utfallsrom $\Omega = \{i, d\}$. Dersom intakte artikler sorteres i tre kvaliteter a, b og c, vil $\Omega = \{a, b, c, d\}$ være et egnet utfallsrom. Vi merker oss at utfallsrommet til et eksperiment ikke nødvendigvis er entydig fastlagt ut fra eksperimentet, men velges etter sitt formål.

Eksempel 5 : To myntkast
En mynt kastes to ganger. Tre synspunkter kan gjøres gjeldene:

- Det er to mulige utfall : Kastene viser likt (ℓ), eller ulikt (u).
- Det er tre mulige utfall : Antall kron er enten 0, 1 eller 2.
- Det er fire mulige utfall : Begge kastene viser kron, begge viser mynt, det første kastet viser kron og det andre mynt,det første kastet viser mynt, og det andre kron.

Vi har derfor minst tre mulige utfallsrom å velge mellom: $\Omega_1 = \{\ell, u\}$, $\Omega_2 = \{0, 1, 2\}$ eller $\Omega_3 = \{KK, KM, MK, MM\}$. Av disse tre utfallsrommene vil en vanligvis foretrekke det siste, blant annet fordi det, uten å være altfor komplisert, er mest detaljert, og derfor kan fange opp flere problemstillinger enn de to andre.

Det er langt fra et mål i seg selv å lage et mest mulig detaljrikt utfallsrom. I Eksempel 1 kunne vi i tillegg til resultatene kron eller mynt registrere med hvilken fart mynten traff bordet, innfallsvinkelen, myntens plassering på bordet etc. Alle slike ting er irrelevant i gambling og for en gambler vil utfallsrommet $\Omega = \{K, M\}$ være detaljrikt nok. Generelt kan man si at det er opp til brukeren å finne et utfallsrom som passer til formålet. Utfallsrommet er referanserammen man velger å arbeide innenfor.

Merknad I Snorre's saga om Olav den hellige (94) forteller Torstein Frode en historie der norske- og svenskekongen var enige om å løse en tvist om et landområde ved kast med to terninger. Etter gjentatte dobbel seksere vant til slutt norskekongen, da den ene terningen ble kløvd på midten og viste (til sammen) syv!

Eksempel 6 : Tre talere
På et møte skal tre talere A, B og C holde hvert sitt innlegg, og rekkefølgen skal bestemmes ved loddtrekning. Som utfallsrom foreslås $\Omega = \{ABC, ACB, BAC, BCA, CAB, CBA\}$, hvor ABC betegner talerekkefølgen først A, deretter B og sist C, etc. Utfallsrommene har i alt 6 mulige utfall.

Eksempel 7 : Myntkast inntil første kron
Vi skal kaste en mynt helt til vi får kron. Et mulig utfallsrom er $\Omega = \{K, MK, MMK, MMMK, \ldots\}$. I denne situasjon vil det være nok å registrere antall kast som trengs for å få en kron, et alternativt utfallsrom er derfor $\Omega = \{1, 2, 3, 4, \ldots\}$. Begge utfallsrom er tellbart uendelige. Ethvert endelig utfallsrom vil være utilstrekkelig.

2.2 Diskrete sannsynlighetsmodeller

Definisjon : Dersom vi har et eksperiment med diskret utfallsrom $\Omega = \{u_1, u_2, u_3, \ldots\}$, og en funksjon p som tilfredstiller

A1. $0 \leq p(u) \leq 1$ for alle u

A2. $p(u_1) + p(u_2) + p(u_3) + \cdots = 1$

sier vi at (Ω, p) utgjør en *diskret sannsynlighetsmodell* for eksperimentet. Funksjonen p vil vi kalle en *sannsynlighetsfunksjon*, og $p(u)$ vil vi kalle *sannsynligheten* for utfallet u.

Vi krever altså at sannsynligheten for hvert mulig utfall skal være et reellt tall mellom 0 og 1 (begge skranker inkludert), og at summen av sannsynlighetene for alle de mulige utfall skal være 1. La oss gi en motivering for denne definisjonen: Anta at vi har gjentatt vårt eksperiment i alt n ganger, og at vi har observert $u_1, u_2, u_3, \ldots$ henholdsvis $n(u_1), n(u_2), n(u_3), ...$ ganger. Da er opplagt $0 \leq n(u) \leq n$ for alle utfall u, og $n(u_1) + n(u_2) + n(u_3) + \cdots = n$. Dividerer vi her med n får vi at de relative hyppighetene $h(u) = n(u)/n$ har egenskapene $0 \leq h(u) \leq 1$ for alle u, og $h(u_1) + h(u_2) + h(u_3) + \cdots = 1$. Siden vi ønsker et sannsynlighetsbegrep som også skal betjene forestillingen om sannsynligheter som idealiserte relative hyppigheter (se Eksempel 1.1 og 1.2), må vi i hvert fall kreve A1 og A2. [2]

Som vi ser gir ikke vår definisjon noen anvisning på hvordan vi eksplisitt skal beregne sannsynlighetene for hvert mulig utfall, enhver funksjon $p : \Omega \to R$ som tilfredsstiller A1 og A2 er tillatt. Modellbyggeren står fritt til å velge en sannsynlighetsfunksjon som samsvarer med de forestillinger han/hun har om eksperimentet.

Eksempel 8 : Et myntkast

Utfallsrom $\Omega = \{K, M\}$. La oss skrive $p = p(K)$ og $q = p(M)$. Her må ifølge krav A1 både p og q være tall mellom 0 og 1, og ifølge krav A2 må alltid $p + q = 1$, dvs. $q = 1 - p$. Forestillingen om at myntkastet er rettferdig

[2] En del viktige formler er gitt navn A1, A2, A3, E1, E2, etc. og er samlet gruppevis i Appendiks B.

uttrykkes i modellen ved å sette $p = q = 1/2$. Dersom det er grunn til å tro mynten er "skjev", kan vi eksperimentere med mynten og basere modellen på den erfaring vi får. (Se Eksempel 1.2 og Oppgave 1.1).

Eksempel 9 : Et terningkast
Utfallsrom $\Omega = \{1, 2, 3, 4, 5, 6\}$. Forestillingen om en rettferdig terning svarer i en modell til at alle de 6 utfallene er like sannsynlige, dvs. $p(1) = p(2) = p(3) = p(4) = p(5) = p(6) = p$. På grunn av A2 må derfor $p+p+p+p+p+p = 6 \cdot p = 1$, som gir at $p = 1/6$. Dersom vi ikke uten videre vil anta at terningen er rettferdig, kan vi skaffe kunnskap ved å eksperimentere med den. Anta at vi i løpet av $n = 1000$ kast observere $n(1) = 194, n(2) = 166, n(3) = 177, n(4) = 181, n(5) = 150, n(6) = 132$. På bakgrunn av denne erfaring er følgende sannsynlighetsfunksjon aktuell (se også Oppgave 8):

$$\begin{array}{lll} p(1) = 0.19, & p(2) = 0.17, & p(3) = 0.18 \\ p(4) = 0.18, & p(5) = 0.15, & p(6) = 0.13 \end{array}$$

Vårt eksperiment har imidlertid ikke bevist at terningen er falsk. Erfaring viser at de tilfeldige variasjoner i terningkast er store, og resultatet kan derfor godt skyldes tilfeldigheter. Er det observerte resultat tilstrekkelig ekstremt til å vrake modellen om rettferdighet? Statistisk teori besvarer slike spørsmål.

Eksempel 10 : Loddtrekning
Utfallsrom $\Omega = \{N_1, N_2, \ldots, N_m\}$. En rettferdig trekning er en trekning der alle lapper har samme sannsynlighet for å bli trukket ut. En modell for en rettferdig trekning er derfor gitt ved

$$p(N_1) = p(N_2) = \cdots = p(N_m) = 1/m$$

Eksempel 11 : To myntkast
Utfallsrom $\Omega = \{KK, KM, MK, MM\}$. En mulig modell antar at alle fire utfall er like sannsynlige, dvs. $p(KK) = p(KM) = p(MK) = p(MM) = 1/4$.

Eksempel 12 : Myntkast inntil første kron
Utfallsrom $\Omega = \{1, 2, 3, 4, \ldots\}$. Siden $\frac{1}{2} + (\frac{1}{2})^2 + (\frac{1}{2})^3 + (\frac{1}{2})^4 + \cdots = 1$, er sannsynlighetsfunksjonen p gitt ved $p(1) = \frac{1}{2}$, $p(2) = (\frac{1}{2})^2$, $p(3) = (\frac{1}{2})^3$,

$p(4) = (\frac{1}{2})^4, \ldots$, dvs. $p(n) = (\frac{1}{2})^n$, $n = 1, 2, 3, 4, \ldots$ en aktuell kandidat. Vi skal senere vise at rimelige antagelser om eksperimentet nettopp leder til denne modellen. Merk at denne modellen innebærer at muligheten for å vente i det uendelige på kron ikke er tilstede, dvs. er tildelt sannsynlighet null.

2.3 Begivenheter

> **Definisjon :** Enhver delmengde A av utfallsrommet Ω kalles en *begivenhet*. De utfall begivenheten A er sammensatt av kalles *gunstige utfall* for A. Vi sier at begivenheten A har inntruffet hvis og bare hvis et av de gunstige utfall for A har inntruffet.

Begrepet begivenhet er innført fordi vi ofte ønsker å gi en mindre detaljrik beskrivelse av et eksperimentresultat enn den hvert utfall gir. En begivenhet A kan beskrives enten ved at vi lister opp de utfall A er sammensatt av, eller ved at vi gir en verbal beskrivelse i form av en egenskap som kjennetegner de og bare de utfall som er med i A.

Merk at ifølge vår definisjon er utfallsrommet Ω også en begivenhet. Ω er nødt til å inntreffe, og vi gir derfor Ω navnet *den sikre begivenhet*. Den tomme mengde $\emptyset$ er også en begivenhet. Siden intet utfall er gunstig for $\emptyset$, kalles denne for *den umulige begivenhet*. En begivenhet som består av bare ett utfall kaller vi en *enkel begivenhet*, en begivenhet som består av mer enn ett utfall kaller vi en *sammensatt begivenhet*. Vi vil vanligvis benytte store bokstaver fra begynnelsen av alfabetet som symboler på begivenheter.

Eksempel 13 : To myntkast
Utfallsrom $\Omega = \{KK, KM, MK, MM\}$. Her er en liste over alle tenkelige delmengder av Ω :

$B_1 = \{KK\}, B_2 = \{KM\}, B_3 = \{MK\}, B_4 = \{MM\}, B_5 = \{KK, KM\}$,
$B_6 = \{KK, MK\}, B_7 = \{KK, MM\}, B_8 = \{KM, MK\}, B_9 = \{KM, MM\}$,
$B_{10} = \{MK, MM\}, B_{11} = \{KK, KM, MK\}, B_{12} = \{KK, KM, MM\}$,
$B_{13} = \{KK, MK, MM\}, B_{14} = \{KM, MK, MM\}, B_{15} = \Omega, B_{16} = \emptyset$.

Det er i alt $16 = 2^4$ ulike delmengder (begivenheter) i Ω. Eksempler på verbale beskrivelser: B_5 = første kast gir kron, B_9 = annet kast gir mynt, B_4 = ingen kron, B_7 = begge kast gir det samme, B_{11} = minst et kast gir kron. Her er B_1 til og med B_4 enkle begivenheter, mens B_5 til og med B_{15} er sammensatte begivenheter. Eksempelvis vil tre kron være den umulige begivenhet.

Gitt en diskret sannsynlighetsmodell (Ω, p) for et eksperiment. Sannsynligheten $p(u)$ for hvert utfall $u \in \Omega$ er da fastlagt. Vi ønsker også å snakke om sannsynligheten til en vilkårlig begivenhet A i utfallsrommet, kall denne for $P(A)$.

Definisjon : *Sannsynligheten til begivenheten* A er summen av sannsynlighetene for de gunstige utfall for A, dvs. med symboler

A3. $$P(A) = \sum_{u \in A} p(u)$$

hvor summasjonen går over alle utfall med i A.

Symbolkombinasjonen $\sum_{u \in A}$ betyr summen av de ledd hvor u er med i A, eksempelvis dersom $A = \{u_1, u_2, u_4, u_7\}$, så er

$$P(A) = \sum_{u \in A} p(u) = p(u_1) + p(u_2) + p(u_4) + p(u_7)$$

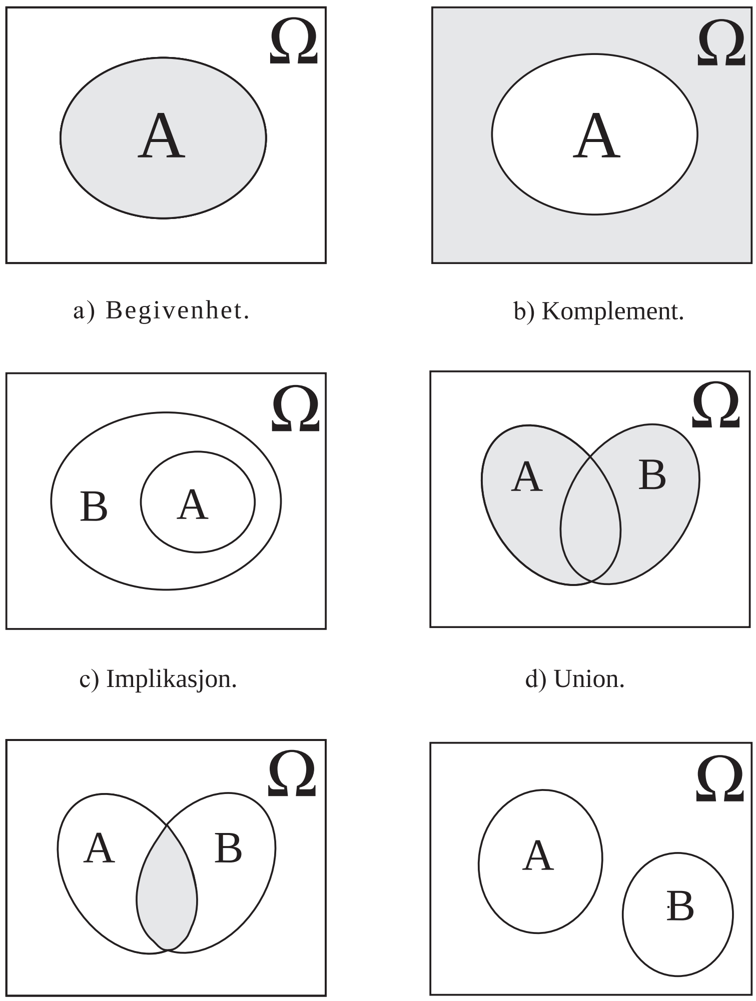
Eksempel 14 : To myntkast
Modellen som er foreslått i Eksempel 11 gir eksempelvis $P(B_5) = p(KK) + p(KM) = \frac{1}{4} + \frac{1}{4} = \frac{1}{2}$, $P(B_9) = p(KM) + p(MM) = \frac{1}{4} + \frac{1}{4} = \frac{1}{2}$, $P(B_4) = p(MM) = \frac{1}{4}$, $P(B_7) = p(KK) + p(MM) = \frac{1}{4} + \frac{1}{4} = \frac{1}{2}$, $P(B_{11}) = p(KK) + p(KM) + p(MK) = \frac{1}{4} + \frac{1}{4} + \frac{1}{4} = \frac{3}{4}$.

Eksempel 15 : Ventetid til første kron i myntkast
Utfallsrom $\Omega = \{1, 2, 3, 4, \ldots\}$. La $A = \{4, 5, 6, 7, 8, \ldots\}$ = første kron kommer i fjerde kast eller senere og $B = \{1, 3, 5, 7, 9, \ldots\}$ = første kron kommer etter et odde antall kast. Den foreslåtte modell gir $P(A) = p(4) + p(5) + p(6) + \cdots = (\frac{1}{2})^4 + (\frac{1}{2})^5 + (\frac{1}{2})^6 + \cdots = (\frac{1}{2})^4 \cdot (1 - \frac{1}{2})^{-1} = \frac{1}{8}$ og $P(B) = p(1) + p(3) + p(5) + \cdots = \frac{1}{2} + (\frac{1}{2})^3 + (\frac{1}{2})^5 + \cdots = \frac{1}{2} \cdot (1 - \frac{1}{4})^{-1} = \frac{2}{3}$ ved bruk av summeformelen for en uendelig geometrisk rekke. [3]

[3] En uendelig geometrisk rekke er en sum av form $a + ar + ar^2 + ar^3 + \cdots$ dvs. der etterfølgende ledd alltid er r ganger det foregående. Dersom r i tallverdi er mindre enn 1, blir summen av rekken lik $a(1-r)^{-1}$, dvs. første ledd dividert med en minus forholdstallet.

Vi merker oss at dersom $E = \{u\}$ er en enkel begivenhet, så gir definisjonen at $P(E) = p(u)$. Vi er derfor fristet til å la være å skille mellom et utfall u og den tilhørende enkle begivenhet $E = \{u\}$, og å benytte samme symbol, stor P, for alle sannsynligheter.

For å gjøre det mulig å regne med begivenheter skal vi nå innføre mengdeoperasjonene union, snitt og komplement.



a) Begivenhet.

b) Komplement.

c) Implikasjon.

d) Union.

e) Snitt.

f) Disjunkthet.

Figur 2.1: Elementer i mengdelære

> **Definisjon :** *Unionen* av begivenhetene A og B skriver vi $A \cup B$, og er den begivenhet som består av alle utfall som enten er med i A eller med i B (eller med i både A og B).

Vi ser at begivenheten $A \cup B$ har inntruffet hvis og bare hvis enten A eller B (eller begge) har inntruffet. Derfor blir $A \cup B$ ofte kalt "begivenheten A eller B".

> **Definisjon :** *Snittet* av begivenhetene A og B skriver vi $A \cap B$, og er den begivenhet som består av de utfall som både er med i A og med i B.

Vi ser at begivenheten $A \cap B$ har inntruffet hvis og bare hvis både A og B har inntruffet. Derfor blir $A \cap B$ ofte kalt "begivenheten A og B". Dersom A og B ikke har noe utfall felles, er $A \cap B$ lik den umulige begivenhet $\emptyset$. I så fall sier vi at A og B er *disjunkte begivenheter*. To begivenheter som er disjunkte kan altså ikke begge inntreffe.

> **Definisjon :** *Komplementet* til begivenheten A, skriver vi $\overline{A}$, og er den begivenhet som består av alle utfall i utfallsrommet som ikke er med i A.

Vi ser at begivenheten $\overline{A}$ har inntruffet hvis og bare hvis begivenheten A ikke har inntruffet. Derfor blir $\overline{A}$ ofte kalt "begivenheten ikke-A". [4]

> **Definisjon :** Vi sier at A er *delmengde* av B, og skriver $A \subset B$, dersom ethvert utfall i A også er med i B.

Når $A \subset B$, så vil, dersom A har inntruffet, også B ha inntruffet. Vi sier at begivenheten A *impliserer* (medfører) begivenheten B. De innførte begreper illustreres gjerne ved å tenke seg begivenheter (mengder) representert ved områder i planet, såkalte *Venn-diagrammer*, se figurene 1-6.

---

[4] Mange foretrekker isteden skrivemåten A^c for å unngå at $\overline{A}$ forveksles med notasjon for gjennomsnitt.

I Figur 2.1a er begivenheten A skravert, i Figur 2.1b begivenheten $\overline{A}$. Figur 2.1c illustrerer utsagnet $A \subset B$. I Figur 2.1d er begivenheten $A \cup B$ skravert, og i Figur 2.1e er $A \cap B$ skravert. Figur 2.1f illustrerer utsagnet $A \cap B = \emptyset$.

Eksempel 16 : Et terningkast
Utfallsrom $\Omega = \{1, 2, 3, 4, 5, 6\}$.$A = \{4, 5, 6\}$ = minst fire øyne, $B = \{3, 6\}$ = antall øyne delelig med tre, $C = \{1, 2\}$ = høyst to øyne, er alle begivenheter. Her er $\overline{A} = \{1, 2, 3\}$ = høyst tre øyne, $A \cup B = \{3, 4, 5, 6\}$ = minst tre øyne, $A \cap B = \{6\}$ = seks øyne, $B \cap C = \emptyset$ og $C \subset \overline{A}$.

Av og til trenger vi å snakke om unioner og snitt av mer enn to begivenheter: La $B_1, B_2, \ldots, B_n$ være n begivenheter. Den begivenhet som består av de utfall som er med i minst en av B_i'ene, kaller vi *unionen* av de n begivenhetene, og vi skriver $B_1 \cup B_2 \cup \ldots \cup B_n$ eller kortere $\cup_{i=1}^{n} B_i$. Den begivenhet som består av de utfall som er med i alle B_i'ene kaller vi *snittet* av de n begivenhetene, og vi skriver $B_1 \cap B_2 \cap \ldots \cap B_n$ eller kortere $\cap_{i=1}^{n} B_i$. På tilsvarende måte defineres union og snitt av en uendelig følge av begivenheter, med skrivemåte henholdsvis $\cup_{i=1}^{\infty} B_i$ og $\cap_{i=1}^{\infty} B_i$.

Eksempel 17 : Rakettoppskyting
I et forskningsprogram skal det skytes opp n raketter, hver oppskyting blir enten suksess eller fiasko. Anta at vi i en modell for resultatet av hele programmet, har definert n begivenheter $A_1, A_2, \ldots, A_n$, hvor A_i er begivenheten at oppskyting nr. i er suksess. Her blir

$B = A_1 \cup A_2 \cup \cdots \cup A_n$ begivenheten at minst en oppskyting er suksess,
$C = A_1 \cap A_2 \cap \cdots \cap A_n$ begivenheten at alle oppskytingene er suksess,
$D = \overline{A}_1 \cup \overline{A}_2 \cup \cdots \cup \overline{A}_n$ begivenheten at minst en oppskyting er fiasko, og
$E = \overline{A}_1 \cap \overline{A}_2 \cap \cdots \cap \overline{A}_n$ begivenheten at alle oppskytingene er fiasko.

Vi innser lett at $\overline{B} = E$ og $\overline{C} = D$.

De sammenhenger vi oppdager i Eksempel 17 gjelder generelt og kalles *De Morgans lover*. Disse lyder altså

De Morgans lover:

M1. $\overline{A_1 \cup A_2 \cup \cdots \cup A_n} = \overline{A}_1 \cap \overline{A}_2 \cap \cdots \cap \overline{A}_n$

M2. $\overline{A_1 \cap A_2 \cap \cdots \cap A_n} = \overline{A}_1 \cup \overline{A}_2 \cup \cdots \cup \overline{A}_n$

Forsøk å bli overbevist om at M1 og M2 gjelder generelt, for eksempel ved å gi en verbal (ikke, og, eller) fortolking av venstre- og høyresiden i hver likhet, eller ved å tegne Venn-diagrammer for tilfellet med bare to begivenheter (Oppgave 15).

For senere referanser nevner vi til slutt i dette avsnittet at en følge $B_1, B_2, B_3, \ldots$ av begivenheter sies å være *disjunkte* dersom $B_i \cap B_j = \emptyset$ for alle $i \neq j$, dvs. dersom aldri to av begivenhetene kan inntreffe samtidig. Illustrer at de tre begivenhetene B_1, B_2 og B_3 er disjunkte med et Venn-diagram!

2.4 Egenskaper ved sannsynligheter

Av de tre aksiomene A1, A2 og A3 for sannsynligheter kan vi nå avlede en rekke fundamentale egenskaper: Først to egenskaper som innses direkte:

E1. $0 \leq P(A) \leq 1$ for alle begivenheter A

E2. $P(\Omega) = 1$

eller sagt med ord: En sannsynlighet er alltid et reelt tall mellom 0 og 1 (begge skranker inkludert), og sannsynligheten for den sikre begivenhet er 1. Så følger den viktige

Addisjonssetningen for disjunkte begivenheter:

E3. $P(A \cup B) = P(A) + P(B)$ når $A \cap B = \emptyset$

eller sagt med ord: Dersom begivenhetene A og B er disjunkte (ikke kan inntreffe samtidig), så er sannsynligheten for begivenheten A eller B lik sannsynligheten for A pluss sannsynligheten for B.

Begrunnelse: Siden A og B ikke har noe utfall felles blir

$$\begin{aligned} P(A \cup B) &= \text{sum av sannsynlighetene for alle utfall i } A \cup B \\ &= \text{sum av sannsynlighetene for utfallene i } A \\ &+ \text{sum av sannsynlighetene for utfallene i } B \\ &= P(A) + P(B). \end{aligned}$$

Siden $A \cup \overline{A} = \Omega$ og $A \cap \overline{A} = \emptyset$, får vi av E2 og E3 at $1 = P(\Omega) = P(A \cup \overline{A}) = P(A) + P(\overline{A})$. Dette gir

E4.	$P(\overline{A}) = 1 - P(A)$

Siden komplementet til Ω er $\emptyset$ får vi spesielt av E2 og E4 at

E5.	$P(\emptyset) = 0$

dvs. sannsynligheten for den umulige begivenhet er null. Merk at E4 og E5 også kan fås av A3 direkte. En mengdefunksjon P som har egenskapene E1, E2 og E3 (og derfor også E4 og E5) kalles ofte et *sannsynlighetsmål*.

E3 og E4 er svært nyttige når en skal løse sannsynlighetsteoretiske oppgaver i praksis. Ofte kan en begivenhet som en ønsker å beregne sannsynligheten for, uttrykkes som en disjunkt union av begivenheter med kjente sannsynligheter, eller som et komplement av en begivenhet med kjent sannsynlighet. Her er et eksempel som illustrerer begge bruksmåter:

Eksempel 18 : Maskindeler
Gitt at en masseprodusert maskindel blir for lang med sannsynlighet 0.08, for kort med sannsynlighet 0.04. Vi ønsker å beregne sannsynligheten for C = hverken for kort eller for lang. Vi ser at begivenhetene A = for lang og B = for kort er disjunkte, og at $\overline{C} = A \cup B$. Vi får da

$$\begin{aligned} P(C) &= 1 - P(\overline{C}) = 1 - P(A \cup B) \\ &= 1 - (P(A) + P(B)) \\ &= 1 - (0.08 + 0.04) = 0.88 \end{aligned}$$

Av og til vil vi trenge følgende generalisering av E3:

Den generelle addisjonssetningen :

E6. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

Tar vi summen av sannsynlighetene for alle utfall i $A \cup B$, ved å legge sammen sannsynlighetene for utfallene i A og i B hver for seg, ser vi av Figur 2.1d at utfallene i $A \cap B$ blir tatt med to ganger, og vi korrigerer derfor for dette ved å trekke fra $P(A \cap B)$. Dersom $A \cap B = \emptyset$, vil $P(A \cap B) = P(\emptyset) = 0$, og vi har fått tilbake addisjonssetningen for disjunkte begivenheter.

Eksempel 19 : Nøkkelpersonene
Arnesen og Bjørnsen er de eneste som har nøkkel til pengeskapet i sitt firma. Anta definert begivenhetene A = Arnesen er til stede, B = Bjørnsen er til stede, C = begge er til stede, D = minst en er til stede og E = ingen er til stede. Gitt at $P(A) = 0.80$, $P(B) = 0.70$ og $P(C) = 0.56$. Kan vi nå finne $P(D)$ og $P(E)$? Vi ser at $C = A \cap B$, $D = A \cup B$ og $E = \overline{A} \cap \overline{B} = \overline{D}$. Vi får $P(D) = P(A \cup B) = P(A) + P(B) - P(A \cap B) = 0.80 + 0.70 - 0.56 = 0.94$ og $P(E) = P(\overline{D}) = 1 - P(D) = 1 - 0.94 = 0.06$.

Vi kan generalisere E3 i en annen retning, nemlig

Dersom $A_1, A_2, \ldots, A_n$ er n disjunkte begivenheter, så er

E7. $P(A_1 \cup A_2 \cup \cdots \cup A_n) = P(A_1) + P(A_2) + \cdots + P(A_n)$

Dette resultatet kan vises på samme måte som E3.

2.5 Uniforme sannsynlighetsmodeller

La $\Omega = \{u_1, u_2, \ldots, u_m\}$ være et utfallsrom med et endelig antall utfall m.

Definisjon : En sannsynlighetsmodell der alle m utfall har samme sannsynlighet kaller vi en *uniform sannsynlighetsmodell*.

La oss se hva som generelt karakteriserer slike modeller : Siden nå $P(u_1) = \cdots = P(u_m) = p$, får vi av A2 at

$$1 = P(u_1) + P(u_2) + \cdots + P(u_m) = p + p + \cdots + p = m \cdot p$$

Herav følger at $p = 1/m$, dvs. i en uniform modell er sannsynligheten p for hvert utfall lik en dividert med antall mulige utfall. La oss finne sannsynligheten for en vilkårlig begivenhet A. Anta at A består av i alt g gunstige utfall. Ifølge A3 blir

$$P(A) = \sum_{u \in A} P(u) = \sum_{u \in A} \frac{1}{m} = g \cdot \frac{1}{m} = \frac{g}{m}$$

Av dette kan vi trekke en vidtrekkende konklusjon, nemlig

> **Regelen om "gunstige på mulige" :** I en uniform sannsynlighetsmodell kan sannsynligheten for enhver begivenhet A finnes ved å dividere antall gunstige utfall for A med antall mulige utfall i utfallsrommet Ω.

Dette viser at uniforme sannsynlighetsmodeller vil være spesielt enkle å arbeide med. Dersom det eksperimentet vi studerer inneholder visse symmetrier, kan det lønne seg å utnytte dette ved valg av utfallsrom og modell. Vi har sett eksempler på uniforme modeller i eksemplene 8,9,10 og 11 ovenfor.

Eksempel 20 : Korttrekning
En vanlig kortstokk inneholder 52 kort: 13 spar, 13 hjerter, 13 ruter og 13 kløver. Den stokkes godt, og et kort trekkes tilfeldig. Her er det rimelig å velge en uniform modell. Det er i alt $m = 52$ mulige utfall, og hvert utfall får da sannsynlighet 1/52. Vi er interessert i sannsynlighetene for begivenhetene A = rødt kort, B = hjerter, C = hjerter dame, D = dame og E = honnør (ess, konge, dame, knekt). Antall gunstige utfall for disse begivenheter er henholdsvis 26, 13, 1, 4 og 16. Vi får derfor
$P(A) = \frac{26}{52} = \frac{1}{2}, \quad P(B) = \frac{13}{52} = \frac{1}{4}, \quad P(C) = \frac{1}{52}$
$P(D) = \frac{4}{52} = \frac{1}{13}, \quad P(E) = \frac{16}{52} = \frac{4}{13}$

Eksempel 21 : To terningkast
En terning skal trilles to ganger (eller to terninger en gang). Vi ønsker bl.a. å uttale oss om sannsynligheten for at sum øyne på de to terningene blir syv. Velger vi $\Omega = \{2, 3, 4, \ldots, 11, 12\}$ som utfallsrom (utfall er sum øyne), og antar uniform modell, vil hvert utfall få tildelt sannsynlighet 1/11. Vår

intuisjon sier imidlertid at det må være mye lettere å oppnå 7 enn 12, og denne modellen er trolig svært urealistisk. Det er ikke uten videre lett å innse hvilke sannsynligheter vi bør tildele hvert utfall i dette utfallsrommet. La oss i stedet velge et utfallsrom Ω med følgende utfall:

(1,1)	(1,2)	(1,3)	(1,4)	(1,5)	(1,6)
(2,1)	(2,2)	(2,3)	(2,4)	(2,5)	(2,6)
(3,1)	(3,2)	(3,3)	(3,4)	(3,5)	(3,6)
(4,1)	(4,2)	(4,3)	(4,4)	(4,5)	(4,6)
(5,1)	(5,2)	(5,3)	(5,4)	(5,5)	(5,6)
(6,1)	(6,2)	(6,3)	(6,4)	(6,5)	(6,6)

(1,6) betegner første terning viser 1, den andre 6 osv. Her er det av symmetrigrunner, såframt terningen(e) er rettferdig(e), rimelig å bruke en uniform modell. Hvert utfall får da tildelt sannsynlighet 1/36. Her blir begivenheten A = sum øyne lik syv = $\{(1,6),(2,5),(3,4),(4,3),(5,2),(6,1)\}$. Som vi ser er det 6 gunstige utfall for A. Bruker vi regelen om "gunstige på mulige" gir modellen at sannsynligheten for A er lik $P(A) = g/m = 6/36 = 1/6$. På samme måte kan vi nå besvare en rekke andre sannsynlighetsteoretiske spørsmål vedrørende dette eksperimentet (Oppgave 18).

I dette avsnittet har vi altså sett at beregning av sannsynligheter i en uniform modell kan reduseres til et telleproblem. Som vi snart skal se, vil det ikke alltid være like lett å foreta opptellingen som i eksemplene ovenfor. I mer kompliserte problemstillinger vil ofte de kombinatoriske metodene som vi skal presentere i neste kapittel være til hjelp.

2.6 Oppgaver

1. Eva skal føde. Diskuter valg av utfallsrom.
2. Familien som skal flytte inn i leiligheten ved siden av Per har et barn som skal begynne på samme skole som han. Per er spent på kjønn og klassetrinn (1 til 6). Foreslå et egnet utfallsrom når vi ønsker å holde rede på både kjønn og klassetrinn.
3. Foreslå et egnet utfallsrom for en to-barns-familie.
4. En student velges ut fra en gruppe av studenter og blir spurt om hvor mange søsken han/hun har. Foreslå et passende utfallsrom.
5. Antall oppringninger til et sentralbord i et gitt tidsrom skal behandles som et stokastisk fenomen. Foreslå et passende utfallsrom.

6. En urne inneholder tre kuler, en rød, en hvit og en blå. To kuler trekkes ut etter tur. Foreslå et egnet utfallsrom når

   (a) vi tar omsyn til rekkefølgen av de uttrukne kulene.
   (b) vi ikke tar omsyn til dette.
   Anta at den uttrukne kule blir lagt tilbake i urnen før annen trekning foretas. Besvar (a) og (b) også under denne forutsetningen.

7. Fire personer skal avgjøre hvilket produkt x eller y de liker best. Foreslå et egnet utfallsrom når vi ønsker å holde rede på hvem som valgte hva.

8. En terning hvor 1 og 6 er motstående sider er fikset ved å plassere et lite lodd inne i terningen ved midten av sekser- siden. Lag en sannsynlighetsmodell for et terningkast med terningen når n = 1000 kast ga de resultatene som er gitt i Eksempel 9, og vi antar at utfallene 2,3,4,5 fortsatt er like sannsynlige.

9. En falsk terning er laget ved at enersiden er høvlet ned litt. Lag en sannsynlighetsmodell for et terningkast med denne terningen når n = 1000 kast ga følgende resultat

$$\begin{array}{lll} n(1) = 209 & n(2) = 157 & n(3) = 136 \\ n(4) = 165 & n(5) = 142 & n(6) = 191 \end{array}$$

   og hvor en tar hensyn til terningens form.

10. Et eksperiment har utfallsrom

$$\Omega = \{u_1, u_2, u_3, u_4, u_5, u_6\}$$

   Anta at begivenhetene A, B og C er gitt ved
   $A = \{u_1, u_3, u_5\}$, $B = \{u_2, u_4\}$ og $C = \{u_3, u_6\}$
   Skriv ned begivenhetene
   $A \cup B$, $A \cap B$, $A \cup C$, $A \cap C$, $B \cup C$ og $B \cap C$
   samt begivenhetene
   $\overline{A}$, $\overline{B}$, $\overline{C}$, $A \cup \overline{C}$, $A \cap \overline{C}$ og $\overline{A} \cap C$
   Finner du noen disjunkte begivenheter?

11. Over utfallsrommene i Oppgave 10 er definert en funksjon $p : \Omega \to R$

$$\begin{array}{lll} p(u_1) = 0.3 & p(u_2) = 0.2 & p(u_3) = 0.2 \\ p(u_4) = 0.1 & p(u_5) = 0.1 & p(u_6) = 0.1 \end{array}$$

   (a) Avgjør om dette utgjør en sannsynlighetsmodell.
   (b) Finn sannsynligheten for begivenhetene $A, B, C, A \cap B$ og $A \cap C$.
   (c) Finn sannsynligheten for begivenhetene $\overline{A}, \overline{B}, \overline{C}, A \cup B$ og $A \cup C$
   både ved direkte beregning og ut fra resultatet i (b) ved bruk av regnereglene E3, E4 og E6 i teksten.

12. Foreslå et egnet utfallsrom for en trebarnsfamilie. Skriv ned følgende begivenheter

    $A =$ minst en gutt
    $B =$ annet barn er gutt
    $C =$ alle barn har samme kjønn

    Bruk en uniform modell og beregn sannsynlighetene for begivenhetene A,B og C. Diskuter hvorvidt en uniform modell vil være realistisk.

13. Se på situasjonen med de tre talere i Eksempel 6. Anta uniform modell og finn sannsynligheten for

    (a) A taler først (b) A taler sist. (c) A taler før C.

14. Foreslå et utfallsrom for fire myntkast. Skriv ned begivenhetene

    $A =$ to kron og to mynt
    $B =$ minst tre mynt
    $C =$ ingen kron før mynt

    (a) Skriv ned begivenhetene $A \cup B, A \cap B, A \cap C$ og $\overline{C}$.
    (b) Er A og B disjukte begivenheter? Enn A og C?
    (c) Anta uniform modell, og beregn sannsynligheten for begivenhetene.

15. Illustrer De Morgans lover i tilfellet med to begivenheter ved hjelp av et Venn-diagram.

16. Vis at dersom begivenheten A medfører B så er $P(A) \leq P(B)$.

17. Avgjør om regnereglene for sannsynligheter er oppfylt dersom

    (a) $P(A) = 0.8, \quad P(B) = 0.4, \quad P(A \cap B) = 0.1$
    (b) $P(A) = 0.6, \quad P(B) = 0.2, \quad P(A \cup B) = 0.9$

18. I situasjonen med to terningkast og uniform modell finn sannsynligheten for begivenhetene

    $B =$ sum øyne er minst syv
    $C =$ sum øyne er odde tall
    $D =$ annet kast viser mer enn første kast.

19. I situasjonen med to terningkast og uniform modell finn sannsynligheten for begivenhetene

    $F_i =$ første kast viser i øyne
    $G_j =$ annet kast viser j øyne
    Kommenter resultatet.

20. Bruk et Venn-diagram til å vise at

    $$A = (A \cap B) \cup (A \cap \overline{B})$$

    Gitt at $P(A \cap B) = 0.5$ og $P(A \cap \overline{B}) = 0.2$. Hva blir $P(A)$?

21. Avgjør om følgende utsagn er riktige eller gale.
    (a) Hvis A og B er disjunkte, B og C er disjunkte, så er også A og C.
    (b) Hvis A og $B \cup C$ er disjunkte, så er A og B disjunkte.
    (c) Hvis $A \cap B \cap C = \emptyset$, så er A, B og C disjunkte.

22. Hvilket utfallsrom kan legges til grunn i Eksempel 19?, enn i Eksempel 17?

23. Betrat tre terningkast og anta uniform modell. Finn sannsynligheten for at
    (a) sum øyne er lik ni
    (b) sum øyne er lik ti
    (c) tredje kast viser minst like mye som summen av de to første.

24. Bruk et Venn diagram til å forklare at

$$\begin{aligned} P(A \cup B \cup C) &= P(A) + P(B) + P(C) \\ &\quad -P(A \cap B) - P(A \cap C) - P(B \cap C) \\ &\quad +P(A \cap B \cap C) \end{aligned}$$

25. Forklar at følgende situasjoner vanskelig kan beskrives med et diskret utfallsrom
    (a) Ventetiden til en lyspære går i stykker.
    (b) Tidspunkter for oppringning til et sentralbord.
    (c) Ankomst og ferdigbetjening av kunder i ved en skranke.

26. Programvare gir typisk mulighet for simulering av utfall fra eksperimenter med kjente sannsynligheter, her illustrert ved simulering av 10 kast med en rettferdig terning.

```
>> RANDOM 10 'Resultat'; INTEGERS 1 to 6
>> PRINT 'Resultat'
3 1 6 1 3 3 5 1 6 2
```

    Bruk din foretrukne programvare til å gjøre det etter. Kan du også få det til på din lommeregner?

Kapittel 3

Kombinatorikk og utvalgsmodeller

Vi lar sannsynlighetsteori hvile et øyeblikk, for å kunne presentere noen effektive tellemetoder knyttet til utvalg. Dette kalles ofte *kombinatorikk*, og kommer oss til nytte i forbindelse med såkalte *utvalgsmodeller*.

3.1 Utvalg

Følgende argumentasjon forekommer ofte i det følgende:

> **Multiplikasjonsregelen :** En jobb skal utføres i to trinn. Trinn 1 kan utføres på m_1 ulike måter. For hver av de m_1 ulike måtene å utføre trinn 1, er det m_2 ulike måter å utføre trinn 2. Det er da $m_1 \cdot m_2$ ulike måter å utføre hele jobben.

Denne regelen kan åpenbart utvides til å gjelde jobber med mer enn to trinn.

Eksempel 1 : En middag
Hvor mange menyer er det mulig å komponere når vi på en restaurant har valget mellom 6 forretter, 10 hovedretter og 4 desserter? For hver av de 6 mulige måter å velge forrett er det 10 mulige måter å velge hovedrett. Det er derfor $6 \cdot 10 = 60$ mulige måter å velge forrett og hovedrett, og for hver av disse 60 er det 4 mulige måter å velge dessert. Det er derfor i alt $6 \cdot 10 \cdot 4 = 240$ måter å komponere menyen på.

Mange praktiske problemstillinger kan tilbakeføres til følgende generelle situasjon: Gitt en populasjon (dvs. en samling av objekter) bestående av N elementer. Vi ønsker å trekke et utvalg på s elementer fra populasjonen. Hva som menes med ordet utvalg må imidlertid presiseres nærmere, det fins nemlig fire hovedtyper av utvalg:

1. Ordnet utvalg med tilbakelegging.
2. Ordnet utvalg uten tilbakelegging.
3. Uordnet utvalg uten tilbakelegging.
4. Uordnet utvalg med tilbakelegging.

Et ordnet utvalg er et utvalg hvor elementene er ordnet i rekkefølge, mens et uordnet utvalg ikke tar hensyn til rekkefølgen. Trekning med tilbakelegging betyr at et uttrukket element blir lagt tilbake i populasjonen før neste trekning.

Eksempel 2 : Utvalg på to
Anta at vår populasjon består av de tre bokstavene a, b, og c (dvs. $N = 3$). Anta at vi trekker et utvalg på $s = 2$ elementer uten tilbakelegging. Vi skal symbolisere ordnede utvalg med runde parenteser (), og uordnede utvalg med krøllparenteser $\{\}$. Vi har i alt 6 mulige ordnede utvalg på 2 elementer:

$$(a, b), \quad (a, c), \quad (b, c), \quad (b, a), \quad (c, a), \quad (c, b)$$

mens vi har bare 3 mulige uordnede utvalg på 2 elementer:

$$\{a, b\}, \quad \{a, c\}, \quad \{b, c\}$$

Dersom trekningen foregår med tilbakelegging har vi i alt 9 mulige ordnede utvalg på 2 elementer:

$$(a, a), \quad (a, b), \quad (a, c), \quad (b, a), \quad (b, b), \quad (b, c), \quad (c, a), \quad (c, b), \quad (c, c)$$

mens vi har bare 6 mulige uordnede utvalg på 2 elementer:

$$\{a, a\}, \quad \{a, b\}, \quad \{a, c\}, \quad \{b, b\}, \quad \{b, c\}, \quad \{c, c\}$$

Merk altså at $\{a, b\}$ og $\{b, a\}$ regnes som samme utvalg, mens (a, b) og (b, a) regnes som to ulike utvalg, i (a, b) er a første element og b annet element, i (b, a) er b første element og a annet element.

Eksempel 3 : Utvalg på tre
Gitt en populasjon på $N = 4$ elementer, a, b, c og d. Vi skal trekke et utvalg på $s = 3$ elementer uten tilbakelegging. Det foreligger i alt 4 mulige uordnede utvalg

$$\{a,b,c\}, \quad \{a,b,d\}, \quad \{a,c,d\}, \quad \{b,c,d\}$$

mens det foreligger følgende mulige ordnede utvalg:

$$\begin{array}{llll}
(a,b,c), & (a,b,d), & (a,c,d), & (b,c,d) \\
(a,c,b), & (a,d,b), & (a,d,c), & (b,d,c) \\
(b,a,c), & (b,a,d), & (c,a,d), & (c,b,d) \\
(b,c,a), & (b,d,a), & (c,d,a), & (c,d,b) \\
(c,a,b), & (d,a,b), & (d,a,c), & (d,b,c) \\
(c,b,a), & (d,b,a), & (d,c,a), & (d,c,b)
\end{array}$$

Vi ser at for et gitt uordnet utvalg på tre elementer fins 6 mulige måter å ordne de tre elementene på, dvs. det er i alt $4 \cdot 6 = 24$ ulike ordnede utvalg.

La oss stille spørsmålet generelt: Hvor mange ulike utvalg på s elementer fra en populasjon på N elementer fins det? Vi skal besvare dette spørsmålet for hver av de tre første hovedtyper av utvalg ovenfor. Uordnet utvalg med tilbakelegging forekommer så sjelden i praksis at vi utelater det.

3.2 Ordnede utvalg. Permutasjoner

> Antall ulike *ordnede* utvalg med tilbakelegging på s elementer fra en populasjon på N elementer er N^s.

Begrunnelse: Vi tenker oss de s elementene utvalgt etter tur. Når vi etter hver utvelgelse legger tilbake det valgte elementet, er det hele tiden N mulige valg av neste element. Vi har N mulige valg av første element, for hvert av disse N mulige valg, er det N mulige valg av annet element, og det er derfor $N \cdot N = N^2$ mulige måter å velge de to første elementene. For hver av disse N^2 måtene å velge de to første elementene, er det N måter å velge det tredje, og det er derfor $N^2 \cdot N = N^3$ mulige måter å velge de tre første elementene. Slik kan vi fortsette, og tankerekken ender slik: For hver av de N^{s-1} måter å velge de $s - 1$ første elementene, er det N måter å velge det siste elementet, og det er derfor $N^{s-1} \cdot N = N^s$ mulige utvalg på s elementer.

Eksempel 4 : Tipperekker
En enkeltrekke består av en markering H (hjemmeseier), U (uavgjort) eller B (borteseier) for hver av de 12 kampene på tippekupongen. Hvor mange ulike enkeltrekker fins det? En enkeltrekke kan ses på som et ordnet utvalg med tilbakelegging på $s = 12$ elementer fra populasjonen H, U, B. Følgelig er det 3^{12} ulike enkeltrekker.

Eksempel 5 : Registreringsskilt
Anta at et registreringsskilt for motorkjøretøyer består av to bokstaver fra alfabetet (fra A til og med Å) og fem sifre (fra 0 til og med 9). Hvor mange registreringsskilt er det mulig å lage på denne måten? For hver av de 29^2 måtene å velge ut de to bokstavene på, er det 10^5 mulige måter å velge ut de 5 sifrene på, dvs. i alt $29^2 \cdot 10^5$ mulige registreringsskilt. Dersom vi ikke godtar skilt hvor alle sifrene er null, blir antallet $29^2 \cdot (10^5 - 1)$. I praksis brukes ikke Æ, Ø, Å og enkelte andre bokstaver og heller ikke visse sifferkombinasjoner.

Antall ulike *ordnede* utvalg *uten* tilbakelegging på s elementer fra en populasjon på N elementer er $(N)_s$ hvor

$$(N)_s = N(N-1)(N-2)\cdots(N-s+1)$$

Symbolet $(N)_s$ uttales gjerne "N i s faktoriell".

Begrunnelse : [1] Vi tenker oss de s elementene valgt ut etter tur. Det er N mulige valg av første element. For hvert av disse N mulige valg av første element er det, siden et element nå er fjernet, $N-1$ mulige valg av annet element. Det er derfor $N\cdot(N-1)$ mulige måter å velge de to første elementene på. For hver av disse $N(N-1)$ måtene er det, siden to elementer er fjernet, $N-2$ mulige måter å velge det tredje elementet på. Det er derfor $N(N-1)(N-2)$ mulige måter å velge de tre første elementene på. Slik kan vi fortsette. Før det siste elementet skal velges, er det fjernet $s-1$ elementer, og det er derfor $N-(s-1) = N-s+1$ mulige måter å velge det siste elementet på.

Symbolet $(N)_N$ kan tolkes som antall ulike *permutasjoner* av N elementer, dvs. antall mulige måter å ordne N elementer på. Vi ser at $(N)_N = N(N-1)(N-2)\cdots 3\cdot 2\cdot 1$. Denne størrelse betegnes vanligvis med $N!$ og

[1] Enkelte foretrekker å bruke symbolet $N^{(s)}$ istedenfor $(N)_s$.

N	s=1	2	3	4	5	6	7	8	9	10
1	1									
2	2	2								
3	3	6	6							
4	4	12	24	24						
5	5	20	60	120	120					
6	6	30	120	360	720	720				
7	7	42	210	840	2 520	5 040	5 040			
8	8	56	336	1 680	6 720	20 160	40 320	40 320		
9	9	72	504	3 024	15 120	60 480	181 440	362 880	362 880	
10	10	90	720	5 040	30 240	151 200	604 800	1 814 400	3 628 800	3 628 800

Tabell 3.1: Antall ordnede utvalg $(N)_s$

vi uttaler "N fakultet".

Eksempel 6 : Styret
I en forening på 10 medlemmer er det 4 styreverv: formann, viseformann, sekretær og kasserer. Hvor mange styresammensetninger finnes det? Vi ser at problemet er det samme som å bestemme antall ordnede utvalg på $s = 4$ elementer fra en populasjon på $N = 10$ elementer (uten tilbakelegging). Vi har følgelig $(N)_s = (10)_4 = 10 \cdot 9 \cdot 8 \cdot 7 = 5040$ mulige styrer.

Eksempel 7 : Talelisten
En representant for hvert av 8 politiske partier skal innlede i en debatt. Antall mulige talerekkefølger er åpenbart lik antall mulige måter å permutere $N = 8$ elementer, dvs. $N! = 8! = 40320$ mulige rekkefølger.

Nedenfor følger en tabell over antall ordnede utvalg $(N)_s$ for s og N fra 1 til og med 10.

3.3 Uordnede utvalg

Antall ulike *uordnede* utvalg *uten* tilbakelegging på s elementer fra en poulasjon på N elementer betegner vi $\binom{N}{s}$ og uttaler "N over s". Her blir

$$\binom{N}{s} = \frac{(N)_s}{s!}$$

Begrunnelse: En annen måte å telle opp antall mulige ordnede utvalg uten tilbakelegging på s elementer fra en populasjon på N elementer er følgende: Velg først ut de s elementene som skal være med i utvalget. Dette kan gjøres på $\binom{N}{s}$ mulige måter. For hver av disse er det $s!$ mulige måter å ordne de s elementene på, altså $\binom{N}{s} \cdot s!$. Herav følger påstanden.

Innsetter vi uttrykket for $(N)_s$ fra forrige avsnitt får vi regneformelen

$$\binom{N}{s} = \frac{N(N-1)(N-2)\cdots(N-s+1)}{s!}$$

Dersom vi multipliserer teller og nevner med $(N-s)!$ får vi alternativt

$$\binom{N}{s} = \frac{N!}{s!(N-s)!}$$

Av denne formelen ser vi følgende symmetri

$$\binom{N}{s} = \binom{N}{N-s}$$

Dette kan vi også innse direkte: Istedenfor å velge ut de s elementene som skal utgjøre utvalget, velger vi ut de $N-s$ elementene som ikke skal være med. Siden $\binom{N}{N} = 1$ ser vi at dersom vi definerer

$$\binom{N}{0} = 1 \text{ og } 0! = 1$$

så oppnår vi at de to siste formlene gjelder for $s = 0$ og $s = N$ også.

Eksempel 8 : Delegasjonen
I en forening på 10 medlemmer skal 4 velges ut som utsendinger til et møte. Hvor mange ulike delegasjoner finnes? Her er det bare spørsmål om hvem som kommer med i utvalget, det er ingen ordning som i Eksempel 6. Problemet er å finne antall uordnede utvalg på $s = 4$ elementer fra en populasjon på $N = 10$ elementer (uten tilbakelegging). Det søkte antall er

$$\binom{N}{s} = \binom{10}{4} = \frac{10 \cdot 9 \cdot 8 \cdot 7}{1 \cdot 2 \cdot 3 \cdot 4} = 210$$

Eksempel 9 : Poker
En korthånd i poker kan oppfattes som et uordnet utvalg på 5 kort fra en kortstokk på 52 kort. Antall mulige korthender blir derfor

$$\binom{N}{s} = \binom{52}{5} = \frac{52 \cdot 51 \cdot 50 \cdot 49 \cdot 48}{1 \cdot 2 \cdot 3 \cdot 4 \cdot 5} = 2\,598\,960$$

For antall uordnede utvalg gjelder følgende identitet

$$\binom{N}{s} = \binom{N-1}{s-1} + \binom{N-1}{s} \quad (0 < s < N)$$

Begrunnelse: Antall ulike utvalg på s elementer fra en populasjon på N elementer kan alternativt telles opp slik: Fiksér et av elementene i populasjonen. Antall ulike utvalg hvor dette elementet er med er $\binom{N-1}{s-1}$, mens antall ulike utvalg hvor det ikke er med er $\binom{N-1}{s}$.

Vi ser at Tabell 3.2 er i overensstemmelse med denne formelen: Hvert tall inne i tabellen er lik summen av tallet som står rett ovenfor og tallet som står ovenfor til venstre. Som øvelse kan leseren føye en ny linje til tabellen. Denne tabellen kalles ofte *Pascals trekant*. En mer omfangsrik tabell er Tabell C.1 i Appendiks C.

3.4 Utvalgsmodeller

I dette avsnitt skal vi studere et par enkle modeller for trekning av utvalg som ivaretar forestillingen om en rettferdig trekningsmåte.

N	s=0	1	2	3	4	5	6	7	8	9	10
0	1										
1	1	1									
2	1	2	1								
3	1	3	3	1							
4	1	4	6	4	1						
5	1	5	10	10	5	1					
6	1	6	15	20	15	6	1				
7	1	7	21	35	35	21	7	1			
8	1	8	28	56	70	56	28	8	1		
9	1	9	36	84	126	126	84	36	9	1	
10	1	10	45	120	210	252	210	120	45	10	1

Tabell 3.2: Antall uordnede utvalg $\binom{N}{s}$

> **Definisjon :** Med et *tilfeldig ordnet utvalg med tilbakelegging* på s elementer fra en populasjon på N elementer menes et ordnet utvalg trukket på en slik måte at alle N^s mulige ordnede utvalg har samme sannsynlighet for å bli trukket.

I denne definisjonen ligger det implisitt at utfallsrommet for trekningen består av de N^s mulige ordnede utvalg, og at modellen for trekningen er uniform, dvs. hvert utvalg har sannsynlighet $1/N^s$ for å bli trukket. Siden modellen er uniform kan sannsynligheten for enhver begivenhet beregnes ved å bruke regelen om "gunstige på mulige".

Eksempel 10 : Tipperekker
Anta at en tipperekke blir trukket ut tilfeldig, dvs. slik at alle 3^{12} mulige tipperekker er like sannsynlige. Hva er da sannsynligheten for begivenheten A= en rekke med bare H'er og B= en rekke uten U'er. Siden vi har uniform modell får vi

$$P(A) = \frac{g}{m} = \frac{1}{3^{12}},\ P(B) = \frac{g}{m} = \frac{2^{12}}{3^{12}}$$

Merk at en tipperekke uten U'er kan oppfattes som et ordnet utvalg på 12 elementer med tilbakelegging fra en populasjon på 2 elementer, nemlig

$\{H, B\}$. Vurder om dette er en realistisk modell for 12 rette i tipping.

Eksempel 11 : Fem terningkast
Utfallet av 5 kast med en terning kan oppfattes som et ordnet utvalg med tilbakelegging på $s = 5$ elementer fra populasjonen $\{1, 2, 3, 4, 5, 6\}$. Vi har da $m = 6^5$ mulige utfall. Dersom terningen er rettferdig, er det rimelig å anta at alle disse $m = 6^5$ mulige utfall er like sannsynlige, dvs. uniform modell. La A være begivenheten at ingen av kastene ga samme antall øyne. Vi får

$$P(A) = \frac{g}{m} = \frac{(6)_5}{6^5} = \frac{5}{54}$$

fordi antall utfall hvor ingen kast viser samme antall øyne er lik antall ordnede utvalg uten tilbakelegging på 5 elementer fra populasjonen på 6 elementer.

Eksempel 12 : Fødselsmåneder
Hva er sannsynligheten for begivenheten A at det i en gruppe på 6 personer finnes to eller flere som har fødselsdag i samme måned. La oss anta at alle $m = 12^6$ mulige ordnede utvalg av fødselsmåneder for de 6 personene er like sannsynlige

$$P(A) = 1 - P(\overline{A}) = 1 - \frac{(12)_6}{12^6} = 0.777$$

Diskuter om den brukte modell er realistisk.

Når vi i det følgende snakker om ordnede og uordnede utvalg, vil vi, dersom ikke annet nevnes, mene utvalg uten tilbakelegging.

> **Definisjon :** Med et *tilfeldig ordnet utvalg uten tilbakelegging* på s elementer fra en populasjon på N elementer, menes et ordnet utvalg trukket på en slik måte at alle $(N)_s$ ordnede utvalg har samme sannsynlighet for å bli trukket.

Utfallsrommet består her av de $(N)_s$ mulige ordnede utvalg, og siden vi har antatt uniform modell, har hvert mulig utvalg sannsynlighet $1/(N)_s$.

Eksempel 13 : Styret
I en forening med 10 medlemmer blir de 4 styrevervene formann, viseformann, sekretær og kasserer fordelt ved loddtrekning. Vi ønsker å finne sannsynlighetene for følgende begivenheter

$$\begin{aligned}
A &= \text{Foreningens eldste blir formann} \\
B &= \text{Foreningens eldste blir sekretær} \\
C &= \text{Foreningens eldste blir med i styret} \\
D &= \text{De 4 yngste medlemmene utgjør styret.}
\end{aligned}$$

En rettferdig trekning betyr at alle $m = (10)_4 = 5040$ mulige styrer er like sannsynlige. Vi får da

$$P(A) = \frac{g}{m} = \frac{(9)_3}{(10)_4} = \frac{1}{10}, \quad P(B) = \frac{g}{m} = \frac{(9)_3}{(10)_4} = \frac{1}{10}$$

$$P(C) = \frac{g}{m} = \frac{4 \cdot (9)_3}{(10)_4} = \frac{4}{10}, \quad P(D) = \frac{g}{m} = \frac{4!}{(10)_4} = \frac{1}{210}$$

Eksempel 14 : Slalåm
Et slalåmrenn har 15 deltakere og startrekkefølgen avgjøres ved loddtrekning. Hva er sannsynligheten for følgende begivenheter

$$\begin{aligned}
A &= \text{Favoritten starter sist} \\
B &= \text{Favoritten starter blant de 5 siste} \\
C &= \text{Begge nordmenn starter blant de 5 første.}
\end{aligned}$$

Ved en rettferdig trekning blir alle $m = 15!$ mulige startrekkefølger like sannsynlige. Vi får

$$P(A) = \frac{g}{m} = \frac{14!}{15!} = \frac{1}{15}$$

$$P(B) = \frac{g}{m} = \frac{5 \cdot 14!}{15!} = \frac{1}{3}$$

$$P(C) = \frac{g}{m} = \frac{(5)_2 \cdot 13!}{15!} = \frac{2}{21}$$

Begrunnelser: Skal favoritten starte sist er det 14! mulige måter å ordne startrekkefølgen for de andre på. For hver av de 5 aktuelle startnumrene for favoritten, er det 14! mulige ordninger av andre. For hver av de $(5)_2$ aktuelle startnumrene for de to nordmennene, er det 13! mulige måter å ordne de andre på.

> **Definisjon :** Med et *tilfeldig uordnet utvalg uten tilbakelegging* på s elementer fra en populasjon på N elementer menes et uordnet utvalg trukket på en slik måte at alle $\binom{N}{s}$ mulige uordnede utvalg har samme sannsynlighet for å bli trukket.

Utfallsrommet består her av alle de $\binom{N}{s}$ mulige uordnede utvalg og hvert av disse har fått tildelt sannsynlighet $1/\binom{N}{s}$ (uniform modell).

Eksempel 15 : Poker
Hva er sannsynligheten for begivenheten A = den utdelte korthånd består av 5 spar (sparflush). Anta at alle $m = \binom{N}{s}$ mulige korthender er like sannsynlige.

$$P(A) = \frac{g}{m} = \frac{\binom{13}{5}}{\binom{52}{5}} = \frac{1\,287}{2\,598\,960} = 0.000495$$

fordi antall gunstige korthender for begivenheten A er lik antall uordnede utvalg på 5 kort fra de 13 sparene i kortstokken.

Eksempel 16 : Delegasjonen
I en forening på 10 medlemmer er det 6 menn og 4 kvinner. En delegasjon på 4 medlemmer skal velges ut ved loddtrekning. Vi ønsker å beregne sannsynlighetene for følgende begivenheter

$$\begin{aligned}
A &= \text{Delegasjonen består av bare kvinner} \\
B &= \text{Delegasjonen består av bare menn} \\
C &= \text{Foreningens eldste er med i delegasjonen} \\
D &= \text{Delegasjonen består av 2 kvinner og 2 menn} \\
E &= \text{Delegasjonen består av 1 kvinne og 3 menn.}
\end{aligned}$$

Vi antar rettferdig trekning, dvs. at alle $\binom{N}{s}$ mulige (uordnede) utvalg er like sannsynlige, og får

$$P(A) = \frac{g}{m} = \frac{1}{\binom{10}{4}} = \frac{1}{210}$$

$$P(B) = \frac{g}{m} = \frac{\binom{6}{4}}{\binom{10}{4}} = \frac{15}{210} = \frac{1}{14}$$

$$P(C) = \frac{g}{m} = \frac{\binom{9}{3}}{\binom{10}{4}} = \frac{84}{210} = \frac{4}{10}$$

$$P(D) = \frac{g}{m} = \frac{\binom{4}{2}\cdot\binom{6}{2}}{\binom{10}{4}} = \frac{6\cdot 15}{210} = \frac{3}{7}$$

$$P(E) = \frac{g}{m} = \frac{\binom{4}{1}\cdot\binom{6}{3}}{\binom{10}{4}} = \frac{4\cdot 20}{210} = \frac{8}{21}$$

Begrunnelser: En delegasjon av bare menn er et utvalg på 4 fra de 6 mennene i foreningen, det er i alt $\binom{6}{4}$ slike. Antall ulike delegasjoner med 2 kvinner og 2 menn er gitt ved at for hver av de $\binom{4}{2}$ måtene å velge ut de 2 kvinnene som skal være med, er det $\binom{6}{2}$ måter å velge ut de 2 mennene, i alt $\binom{4}{2}\cdot\binom{6}{2}$ slike delegasjoner etc.

3.5 Noen setninger om tilfeldige utvalg

> 1. Et tilfeldig uordnet utvalg (uten tilbakelegging) kan oppnås ved å trekke et tilfeldig ordnet utvalg og se bort fra rekkefølgen av elementene i utvalget.

Begrunnelse: Det er i alt $s!$ ordnede utvalg som gir det samme uordnede utvalg (antall måter å ordne de s elementene i utvalget). Ved tilfeldig ordnet trekningsmåte blir derfor sannsynligheten for et bestemt uordnet utvalg lik

$$\frac{s!}{(N)_s} = \frac{1}{\binom{N}{s}}$$

dvs. samme sannsynlighet som et bestemt tilfeldig uordnet utvalg.

> 2. Ved trekning av et tilfeldig ordnet utvalg har hvert av de N elementene i populasjonen samme sannsynlighet $1/N$ for å opptre på en bestemt av de s plassene i utvalget. Dette kalles ofte for *ekvivalensloven* for et tilfeldig ordnet utvalg.

Begrunnelse: La $A_i =$ det bestemte elementet er på i'te plass i utvalget. Siden antall gunstige utvalg for begivenheten A_i er lik antall måter å besette de $s-1$ andre plassene i utvalget, får vi

$$P(A_i) = \frac{g}{m} = \frac{(N-1)_{s-1}}{(N)_s} = \frac{1}{N}$$

Eksempel 17 : Lat elev
I en klasse på 20 velger læreren ut 5 elever som etter tur skal regne hver sin av 5 hjemmeoppgaver på tavlen. En elev har ikke gjort oppgave nr. 4. Sannsynligheten for at han blir bedt om å regne denne oppgaven er 1/20. Se også Eksempel 13.

3. Dersom det trekkes et tilfeldig utvalg (uordnet eller ordnet) på s elementer fra en populasjon på N elementer, så har hvert element i populasjonen sannsynlighet s/N for å bli med i utvalget.

Begrunnelse: La A = Det bestemte elementet er med i utvalget. Problemets natur sammen med setning 1 ovenfor medfører at den søkte sannsynlighet må være den samme for begge typer utvalg, og det er derfor likegyldig om vi resonnerer uordnet eller ordnet. For illustrasjonens skyld tar vi likevel med begge resonnementene. Tilfeldig uordnet utvalg: Siden antall gunstige utvalg for begivenheten A er lik antall måter å velge ut de $s-1$ andre elementene som skal være med i utvalget, får vi

$$P(A) = \frac{g}{m} = \frac{\binom{N-1}{s-1}}{\binom{N}{s}} = \frac{s}{N}$$

Tilfeldig ordnet utvalg: Vi kan skrive

$$A = A_1 \cup A_2 \cup \cdots \cup A_s$$

hvor A_i = Det bestemte elementet opptrer på i'te plass i utvalget. Vi vet fra setning 2 ovenfor at $P(A_i) = 1/N$ for alle i, og addisjonssetningen for disjunkte begivenheter gir derfor

$$P(A) = P(A_1) + P(A_2) + \cdots + P(A_s) = \frac{1}{N} + \frac{1}{N} + \cdots + \frac{1}{N} = \frac{s}{N}$$

Eksempel 18 : Lat lærer
En elev har levert inn 10 oppgaver til retting hvorav en oppgave er galt løst. Læreren velger tilfeldig 4 oppgaver som han retter. Sannsynligheten for at den gale oppgaven er med i utvalget er 4/10. Se også Eksempel 13 og Eksempel 16.

15380	30458	24235	37387	55965	05730	34338
31196	55337	12100	48218	77918	96825	16175
32060	74661	85245	60211	55321	87577	82319
61696	28679	48462	29023	31904	08143	27340
45595	03178	00973	06210	27249	31618	63034
17382	39463	85125	52023	41381	72824	81201

Tabell 3.3: Tilfeldige tall

3.6 Tilfeldige tall

I avsnittene ovenfor har vi sett problemstillinger der det er tale om å trekke et tilfeldig utvalg fra en gitt populasjon. Hvordan man realiserer en slik trekning i praksis, har vi bare flyktig berørt, bl. a. gir de setningene som ble presentert i forrige avsnitt en viss teoretisk innsikt som kan brukes til å realisere trekninger. Et nyttig hjelpemiddel i praksis er tabeller over såkalte *tilfeldige tall*, det er grovt sagt en følge av en-sifrede tall, der hvert tall er en trekning blant sifrene 0, 1, 2,..., 9 (forhåpentligvis) med samme sannsynlighet. Tallene i følgen blir trukket uavhengig av hverandre (om begrepet uavhengighet se Kapittel 4.5). En modell for et tilfeldig tall er altså

Siffer	0	1	2	3	4	5	6	7	8	9
Sannsynlighet	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{1}{10}$

For å kontruere en tabell over tilfeldige tall trenger vi en mekanisme som, ut fra rene fysiske betraktninger eller generell erfaring, gir sifre som alle er like sannsynlige, og som genererer nye tall uavhengig av de som er generert før. I praksis gjøres dette på en regnemaskin (dator), men selv lommeregnere gir idag muligheter for å generere slike tall. Det finnes også publisert bøker av tilfeldige tall generert på dette viset. Det er visse filosofiske problemer forbundet med hva som egentlig er et tilfeldig tall, men de systemer som er tilgjengelig i dag kan betraktes som brukbare for de fleste praktiske formål. La oss presentere et utsnitt av en tabell over tilfeldige tall generert ved en regnemaskin:

For å lette oversikten har vi gruppert tallene fem og fem. Selv om tabellen inneholder tilfeldige en-sifrede tall kan den også brukes til å plukke ut tilfeldige to-sifrede tall, idet to suksessive sifre kan oppfattes som en rettferdig trekning fra alle 100 to-sifrede tall 00, 01, 02, ..., 97, 98, 99 (dvs. alle har samme sannsynlighet 1/100). Likeens kan tre suksessive sifre oppfattes som et tilfeldig tre-sifret tall, dvs. et tall generert slik at alle 1000 tre-sifrede tall 000, 001, 002, ..., 997, 998, 999 har samme sannsynlighet 1/1000 osv. Ved bruk av en slik tabell kan en ta tallene i rekkefølge horisontalt, vertikalt eller på skrå. Følgen av tall som fremkommer skal, dersom generatoren er brukbar, kunne oppfattes som tilfeldig uansett. En ting må vi imidlertid passe på, ved gjentatt bruk av slik tabell, må vi starte på nytt sted hver gang, og helst velge startsted tilfeldig, ellers vil en lett kunne komme til å misbruke idéen om tilfeldige tall. I praksis har en tabell over tilfeldige tall minst to bruksmåter: Den ene er å trekke et utvalg fra en endelig populasjon, den andre er å bruke tabellen til å generere data som brukes i forbindelse med en gitt modell for et system til å studere egenskapene ved systemet, det siste kalles *simulering*. La oss her se på den første bruksmåten:

Eksempel 19 : Tilfeldige utvalg
Vi har en gruppe på $N = 74$ personer. Fra denne skal trekkes et tilfeldig utvalg på $s = 5$ personer. Vi kan da nummerere personene slik 00, 01, ...,73. La oss (tilfeldig) velge å bruke siste linje i tabellen, vi får da følgende to-sifrede tall

17 38 23 94 63 85 12 55 20 ...

Her ser vi at tallene 94 og 85 ikke svarer til noen person og disse neglisjeres. Vårt utvalg på 5 personer består derfor av personene med nr. 17, 38, 23, 63, 12. Det er intuitivt opplagt, og kan vises formelt, at dersom alle 100 to-sifrede tall som genereres er like sannsynlige så vil alle 74 tillatte to-sifrede tall som framkommer være like sannsynlige, dvs. at vi har en rettferdig trekning. Det utvalg vi har fått ovenfor er et (forhåpentligvis) tilfeldig ordnet utvalg. Et tilfeldig uordnet utvalg får vi ved å se bort fra rekkefølgen av personene. I dette eksemplet er det tale om utvalg uten tilbakelegging. Skulle samme to-sifrede tall dukke opp flere ganger, tar vi det med bare en gang og går videre til neste to-sifrede tall. Dette endrer ikke på det forhold at de fem tall vi ender opp med kan anses å være resultat av en rettferdig trekning.

3.7 ⋆Grupperinger

Problemet med å trekke et utvalg (uordnet uten tilbakelegging) på s elementer fra en populasjon på N elementer kan også sees på som å dele populasjonen i to grupper, gruppe nr. 1 med s elementer og gruppe nr. 2 med $N - s$ elementer. Det følger da at det er

$$\binom{N}{s} = \frac{N!}{s!(N-s)!}$$

ulike måter å sette sammen de to gruppene på.

La oss se på et tilsvarende problem der vi skal dele en populasjon i tre grupper: En populasjon på N elementer skal deles i tre grupper, gruppe nr.1 med s_1 elementer, gruppe nr.2 med s_2 elementer og gruppe nr.3 med $N - s_1 - s_2$ elementer. I analogi med skrivemåten for to grupper lar vi $\binom{N}{s_1, s_2}$ betegne antall mulige måter å fordele N elementer på tre grupper med henholdsvis s_1, s_2 og $N - s_1 - s_2$ elementer i gruppe nr. 1, 2 og 3. Vi har følgende formel

$$\binom{N}{s_1, s_2} = \binom{N}{s_1} \cdot \binom{N - s_1}{s_2}$$

En mer symmetrisk formel er

$$\binom{N}{s_1, s_2} = \frac{N!}{s_1! s_2! (N - s_1 - s_2)!}$$

Begrunnelse: Det er i alt $\binom{N}{s_1}$ mulige valg av de s_1 elementer fra populasjonen på N som skal utgjøre gruppe nr. 1. For hvert av disse er det $\binom{N-s_1}{s_2}$ mulige valg av de s_2 elementene som skal utgjøre gruppe nr. 2 (disse må velges blant de $N - s_1$ elementene som ikke allerede er blitt plassert i gruppe nr. 1). Følgelig er det

$$\binom{N}{s_1} \cdot \binom{N - s_1}{s_2}$$

mulige måter å velge ut henholdsvis s_1 elementer til gruppe nr. 1 og s_2 elementer til gruppe nr. 2. De $N - s_1 - s_2$ elementene som er igjen utgjør så gruppe nr. 3, slik at det søkte antall er nettopp utrykket ovenfor. Dette uttrykk kan skrives

$$\binom{N}{s_1} \cdot \binom{N - s_1}{s_2} = \frac{N!}{s_1!(N - s_1)!} \cdot \frac{(N - s_1)!}{s_2!(N - s_1 - s_2)!}$$

Ved å stryke $(N - s_1)!$ i teller og nevner får vi den alternative formel.

Eksempel 20 : Gruppearbeid

I en skoleklasse på 15 elever skal det utføres gruppearbeid. Elevene skal deles i tre grupper med 6 elever i gruppe nr. 1, 5 elever i gruppe nr. 2 og 4 elever i gruppe nr. 3. Dette svarer til situasjonen ovenfor med $N = 15, s_1 = 6, s_2 = 5$ og $N - s_1 - s_2 = 4$. Antall ulike sammensetninger av de tre gruppene er derfor

$$\binom{15}{6, 5} = \binom{15}{6} \cdot \binom{9}{5} = 5005 \cdot 126 = 630\,630$$

Uttrykket $\binom{N}{s_1, s_2}$ var i første omgang tenkt definert for s_1, s_2 slik at $s_1, s_2 > 0$ og $s_1 + s_2 < N$, men vi ser at det er meningsfylt å sette

$$\binom{N}{s_1, 0} = \binom{N}{s_1}, \quad \binom{N}{0, s_2} = \binom{N}{s_2}$$

og

$$\binom{N}{s_1, s_2} = \binom{N}{s_1} = \binom{N}{s_2} \text{ når } s_1 + s_2 = N$$

slik at $\binom{N}{s_1,s_2}$ når $s_1 + s_2 = N$ gjerne kan brukes til å betegne antall måter å fordele N elementer i to navngitte grupper, med henholdsvis s_1 elementer i gruppe nr. 1 og s_2 elementer i gruppe nr. 2. Merk at

$$\binom{N}{0, 0} = \binom{N}{N, 0} = \binom{N}{0, N} = 1$$

Problemstillingen ovenfor kan åpenbart generaliseres til å dele en populasjon i et vilkårlig antall grupper: En populasjon på N elementer skal deles i r navngitte grupper, med henholdsvis s_1 elementer i gruppe nr. 1, s_2 elementer i gruppe nr. 2,..., s_{r-1} elementer i gruppe nr. $(r-1)$ og $N - s_1 - s_2 - \cdots - s_{r-1}$ elementer i gruppe nr. r. Antall ulike slike grupperinger er (med symbolikk valgt i analogi med det foregående)

$$\binom{N}{s_1, s_2, \cdots s_{r-1}} = \binom{N}{s_1}\binom{N - s_1}{s_2}\binom{N - s_1 - s_2}{s_3} \cdots \binom{N - s_1 - s_2 - \cdots - s_{r-2}}{s_{r-1}}$$

En mer attraktiv formel er kanskje

$$\binom{N}{s_1, s_2, \cdots s_{r-1}} = \frac{N!}{s_1! s_2! \cdots (N - s_1 - s_2 - \cdots - s_{r-1})!}$$

Eksempel 21 : Bridge

Antall initiale spillesituasjoner i Bridge med 13 kort til hver av spillerne Nord, Øst, Syd og Vest er

$$\binom{52}{13, 13, 13} = \frac{52!}{13!13!13!13!} = \frac{52!}{(13!)^4}$$

som er et tall med 29 sifre.

I visse sammenhenger kan det komme på tale å lage modeller for en gruppering hvor en ivaretar forestillingen om at grupperingen foretas på en rettferdig måte. Vi antar da en uniform modell der alle mulige grupperinger antas å være like sannsynlige.

Eksempel 22 : Gruppearbeid

La situasjonen være som i Eksempel 20. Vi antar at alle $\binom{15}{6,5}$ mulige fordelinger av de 15 elevene på de tre gruppene med henholdsvis 6, 5 og 4 elever i gruppe nr. 1, 2 og 3 er like sannsynlige. Vi ønsker å finne sannsynligheten til begivenhetene

A	$=$	Per kommer i gruppe nr. 1.
B	$=$	Kameratene Per, Pål og Espen kommer alle i gruppe nr. 1.
C	$=$	Per, Pål og Espen kommer i henholdsvis gruppe nr. 1, 2 og 3.

Siden vi har en uniform modell, kan disse sannsynlighetene regnes ut ved å bruke regelen om "gunstige på mulige". Vi får

$$P(A) = \frac{g}{m} = \frac{\binom{14}{5,5}}{\binom{15}{6,5}} = \frac{\frac{14!}{5!5!4!}}{\frac{15!}{6!5!4!}} = \frac{6}{15}$$

$$P(B) = \frac{g}{m} = \frac{\binom{12}{3,5}}{\binom{15}{6,5}} = \frac{\frac{12!}{3!5!4!}}{\frac{15!}{6!5!4!}} = \frac{6 \cdot 5 \cdot 4}{15 \cdot 14 \cdot 13} = \frac{4}{91}$$

$$P(C) = \frac{g}{m} = \frac{\binom{12}{5,4}}{\binom{15}{6,5}} = \frac{\frac{12!}{5!4!3!}}{\frac{15!}{6!5!4!}} = \frac{6 \cdot 5 \cdot 4}{15 \cdot 14 \cdot 13} = \frac{4}{91}$$

Begrunnelse: Antall grupperinger hvor Per er i gruppe nr. 1 er lik antall måter å fordele de 14 andre elevene på de tre gruppene med henholdsvis 5,5 og 4 elever i gruppe nr. 1,2 og 3. Antall grupperinger hvor Per, Pål og Espen alle er i gruppe nr. 1 er lik antall måter å fordele de 12 andre elevene på de tre gruppene med henholdsvis 3, 5 og 4 elever i gruppe nr. 1, 2 og 3. Antall grupperinger hvor Per, Pål og Espen kommer i henholdsvis gruppe nr. 1, 2 og 3 er lik antall måter å fordele de 12 andre elevene på de tre gruppene med henholdsvis 5, 4 og 3 elever i gruppe nr. 1, 2 og 3.

En anvendelse med en litt annen karakter har vi i følgende eksempel.

Eksempel 23 : Oppstilling
20 rekrutter, hvorav 7 fra Østlandet, 5 fra Vestlandet, 4 fra Sørlandet og 4 fra Nord-Norge stiller opp, f. eks. i matkø. Hvor mange ulike køkonstellasjoner finnes det når vi bare tar omsyn til hjemstavn og ikke person? Her er det nyttig å nummerere plassene i køen 1, 2, 3,..., 19, 20. En konstellasjon kan da oppfattes som en gruppering av tallene 1, 2, 3,..., 19, 20 i fire navngitte grupper, Ø gruppen, V gruppen, S gruppen og N gruppen, med henholdsvis 7, 5, 4 og 4 tall i hver gruppe (merk at her er populasjonen tallene 1, 2,..., 20 og ikke mengden av rekrutter). Det finnes derfor i alt

$$m = \binom{20}{7,5,4}$$

mulige ulike konstellasjoner. Anta at de 12 første rekruttene får kjøttkaker, mens resten må nøye seg med lappskaus. Antar vi at alle m mulige konstellasjoner er like sannsynlige (diskuter hvorvidt dette er realistisk), kan vi beregne sannsynligheter. Eksempelvis la

A = Blant de 12 første finnes 6 Østlendinger, 3 Vestlendinger, 2 Sørlendinger og 1 fra Nord-Norge.
B = Alle Østlendingene er blant de 12 første.

Da blir ifølge regelen om "gunstige på mulige"

$$P(A) = \frac{g}{m} = \frac{\binom{12}{6,3,2} \cdot \binom{8}{1,2,2}}{\binom{20}{7,5,4}} = \frac{93\,139\,200}{6\,983\,776\,800} = 0.013$$

$$P(B) = \frac{g}{m} = \frac{\binom{12}{7} \cdot \binom{13}{0,5,4}}{\binom{20}{7,5,4}} = \frac{71\,351\,280}{6\,983\,776\,800} = 0.010$$

Forklaring: Vi skal først gruppere de 12 første plassnumrene med 6 til Ø, 3 til V, 2 til S og 1 til N. Dette kan gjøres på $\binom{12}{6,3,2}$ måter. For hver av disse kan vi gruppere de 8 siste plassnumrene med 1 til Ø, 2 til V, 2 til S og 3 til N på $\binom{8}{1,2,2}$ måter, ialt $\binom{12}{6,3,2} \cdot \binom{8}{1,2,2}$ måter som er gunstige for begivenheten A. For begivenheten B kan de 7 plassnumre blant de 12 første som skal være Ø velges ut på $\binom{12}{7}$ måter. For hver av disse er det $\binom{13}{0,5,4}$ måter å gruppere de 13 ledige numre slik at 0 gis til Ø, 5 til V, 4 til S og 4 til N. En alternativ framgangsmåte i dette eksemplet ville være å identifisere de 20 rekruttene ved navn (eller nr.) og tenke seg at alle 20! ordninger er like sannsynlige. Ved opptellig i telleren må en da ta hensyn til at ulike ordninger kan gi samme konstellasjon.

3.8 Oppgaver

1. Handelsreisende Hansen skal reise hjemmefra a til b og så til c, og deretter tilbake til a.
   Fra a til b er det 4 reisemåter: fly, buss, båt, tog.
   Fra b til c er det 2 reisemåter: buss, båt.
   Fra c til a er det 3 reisemåter: fly, buss, båt.
   Hvor mange reisemåter foreligger i alt?

2. Kursen på en aksje observeres på etterfølgende hverdager og noteres enten oppgang (+), status quo (0) eller nedgang (−). Hvor mange utfall finnes det dersom det observeres i 20 dager? Enn i 30 dager?

3. Et produkt produseres i serier på 12 enheter. Hver enhet kan klassifiseres som intakt (i) eller defekt (d).
   (a) Hvor mange ulike produksjonsserier finnes det?
   (b) Hvor mange serier er slik at ingen intakt kommer etter defekt?
   Anta at intakte enheter sorteres i tre kvaliteter a, b og c.

(c) Hvor mange ulike produksjonsserier finnes nå?
(d) Hvor mange serier er slik at ingen er defekte?
(e) Hvor mange serier er slik at de 6 siste produksjonsnumre er defekte?

4. En visesangerinne har 10 sanger på repertoaret. Ved en konsert skal hun synge 6 av disse. Hvor mange ulike programvalg finnes det
(a) når vi tar hensyn til rekkefølgen av sangene,
(b) når vi ikke tar hensyn til rekkefølgen av sangene.

5. I en skoleklasse bestående av 10 jenter og 12 gutter skal det oppføres et skuespill som har 3 jenteroller og 4 gutteroller. Hvor mange rollelister finnes det?

6. Hvor mange ulike signal kan vi sende ved å sette flagg i rekkefølge dersom
(a) vi har 4 flagg av ulik farge,
(b) vi har 5 flagg av ulik farge,
(c) vi har 2 røde flagg og 4 flagg til av ulik farge.

7. På hvor mange måter kan 7 personer plasseres ved et rundt bord dersom
(a) vi tar hensyn til personenes plassering i forhold til rommet ellers,
(b) vi ikke tar hensyn til dette,
(c) den eneste kvinne i laget skal sitte nærmest døren.

8. n land har diplomatiske forhold.
(a) Hvor mange diplomatiske forhold?
(b) Hvor mange ambassadører?

9. I en skoleklasse bestående at 10 jenter og 12 gutter skal 4 elever utføre et bestemt gruppearbeid. Hvor mange ulike sammensetninger av denne gruppen finnes det når
(a) det velges blant alle elevene,
(b) alle 4 skal være jenter,
(c) alle 4 skal være gutter,
(d) det skal være 2 gutter og 2 jenter med.

10. Ti personer skal avgjøre hvilket av to produkter x eller y de liker best.
(a) Hvor mange utfall er det for dette eksperimentet når vi holder rede på hvem som valgte hva?
(b) Hvor mange av disse utfall er slik at 5 personer foretrekker x og 5 foretrekker y?
(c) Hvor mange utfall er slik at 6 personer foretrekker x og 4 foretrekker y?

11. Ti personer skal avgjøre hvilket av tre produkter x, y og z de liker best.
(a) Hvor mange utfall finnes det i alt?
(b) Hvor mange utfall er slik at z aldri blir foretrukket?
(c) Hvor mange utfall er slik at x blir foretrukket oftere enn y og z tilsammen?

(d) Hvor mange utfall er slik at 5 personer foretrekker x, 3 y og 2 z?

Ved løsningen av oppgavene nedenfor bør man først fastlegge utfallsrom og sannsynlighetsmodell før de enkelte punkter besvares.

12. Hvor mange tall med seks sifre finnes det når hvert siffer kan være et av tallene 0 til 9 (eksempelvis betyr 004902 tallet 4902). Hva menes med et tilfeldig seks-sifret tall?
    Finn sannsynligheten for at et tilfeldig seks-sifret tall
    (a) ikke starter med null,
    (b) ikke inneholder nuller,
    (c) hverken inneholder nuller eller niere,
    (d) inneholder bare odde sifre,
    (e) har verdi mindre enn 5000,
    (f) inneholder seks forskjellige sifre.
    (g) har tverrsum 3.

13. Se på situasjonen i Oppgave 10. Gitt at produktene i virkeligheten er identiske (bare forskjellig emballasje). Finn sannsynlighetene for begivenhetene nedenfor under forutsetning av at de ti personene velger ut tilfeldig det produkt de "liker best".
    (a) 5 liker prudukt x og 5 liker produkt y,
    (b) produkt x likes best av minst 6 personer,
    (c) et av produktene likes best av minst 7 personer.

14. Se på situasjonen i Oppgave 11. Anta at de ti personene velger sitt foretrukne produkt tilfeldig. Finn sannsynligheten for at
    (a) et av produktene blir foretrukket av minst 5,
    (b) 5 personer foretrekker x, 3 personer y og 2 personer z.

15. Til en vinsmakekonkurranse foreligger i alt 20 rødvinsmerker, hvorav 8 er Bordeaux viner. Hver deltaker får i alt 5 smaksprøver og skal for hver besvare spørsmål om opprinnelse, årgang etc. Anta at utvalget og rekkefølgen av vinene for hver deltaker skjer ved en (rettferdig) loddtrekning. Finn sannsynligheten for at en deltaker får smake:
    (a) den eldste vinen,
    (b) de to dyreste vinene,
    (c) bare Bordeaux viner,
    (d) Bordeaux vin først,
    (e) Bordeaux vin først og sist.
    (f) aldri to Bordeaux viner etter hverandre.

16. Seks trommeslagere plasseres i en rekke på en tilfeldig måte. Hva er sannsynligheten for at
    (a) Per har ytterplass,
    (b) både Per og Pål har ytterplass,
    (c) hverken Per eller Pål har ytterplass,
    (d) Per og Pål står ved siden av hverandre.

17. I et musikkorps er det 12 trompetister som er plassert 6 og 6 i de to siste rekkene. Anta at plasseringen avgjøres ved loddtrekning. Hva er sannsynligheten for at Hans og Grete står
(a) bak hverandre,
(b) i samme rekke,
(c) ved siden av hverandre,
(d) begge i ytterplass.

18. En konkurranse i et ukeblad består i at man skal "matche" fire filmstjerner med fire barnebilder. Under forutsetning av at matchingen skjer tilfeldig finn sannsynligheten for
(a) ingen rette,
(b) en rett,
(c) to rette,
(d) tre rette,
(e) fire rette.

19. Et produksjonsparti på 12 enheter inneholder 4 defekte enheter. Tre enheter velges ut tilfeldig for kontroll. Hva er sannsynligheten for at
(a) alle er intakte,
(b) en er defekt,
(c) minst en er defekt,
(d) høyst to er defekte.

20. Et produksjonsparti på 24 enheter inneholder 6 defekte. Hvor stort utvalg må trekkes for at sannsynligheten for å få minst en defekt er minst
(a) 1/2, (b) 4/5, (c) 9/10.

21. I en bedrift med 30 ansatte, hvorav 24 er arbeidere og 6 funksjonærer, skal velges en komité på 4 til å forberede julebordet. Dersom utvelgelsen foregår tilfeldig, hva er sannsynligheten for at utvalget består av
(a) ingen funksjonærer,
(b) 3 arbeidere og en funksjonær,
(c) minst en funksjonær.

22. I en avdeling i en bedrift er ansatt 10 kvinner og 10 menn. Fraværstallene for siste år blir ordnet etter stigende rekkefølge og kjønn blir notert f. eks. MKMM ... MM (i dette tilfellet er den som har minst fravær mann, nest minst kvinne etc.). Anta at fraværstilbøyeligheten for menn og kvinner i virkeligheten er den samme, og finn sannsynligheten for
(a) akkurat 5 kvinner og 5 menn blant de 10 med mest fravær,
(b) minst 6 kvinner blant de 10 med mest fravær,
(c) minst 4 kvinner blant de 5 med mest fravær.

23. I en bedrift er det 20 ansatte. Bedriftsklubben har luftet tanken om at arbeidstiden i sommerhalvåret skal starte kl. 7.30 mot kl. 8.00 nå. Det viste seg at 8 var tilhenger av det tidligere alternativ (T), mens 12 var tilhenger av det

sene (S). Man har samtidig notert seg alder og reisetid for de ansatte. Gjør de forutsetninger som trengs (diskuter om disse er realistiske) for å kunne beregne sannsynligheten for at
(a) ingen av de 3 yngste favoriserer det tidlige alternativ,
(b) høyst en blant de 5 yngste favoriserer det tidlige alternativ,
(c) blant de 6 med lengst reisetid var det flertall for det tidlige alternativ.

24. I en bedrift med 30 ansatte hvorav 24 er arbeidere (16 menn og 8 kvinner) og 6 funksjonærer (4 menn og 2 kvinner) skal det dannes et utvalg på 5 medlemmer hvorav 3 skal velges fra gruppen arbeidere og 2 fra funksjonær gruppen. Utvelgelsen skjer ved loddtrekning. Finn sannsynligheten for at utvalget vil bestå av:
(a) bare menn,
(b) minst en kvinne,
(c) minst en kvinnelig arbeider,
(d) minst en kvinne og en mann fra hver gruppe.

25. To familier skal flytte inn i samme blokk som Per. Per vet at det er et barn i hver familie som skal begynne på samme skole som ham, og er derfor spent på kjønn og klassetrinn (1 til 6). Foreslå egnet utfallsrom, anta uniform modell og finn sannsynligheten for
(a) minst en gutt,
(b) minst en førsteklassing,
(c) minst en gutt på trinn 3 eller lavere.
Vil samme modell være realistisk i samme grad dersom det dreier seg om en familie med to skolebarn?

26. En undersøkelse blant alle 200 studentene i et hybelhus ga som resultat

	Har TV	Har ikke TV	Sum
Har PC	50	30	80
Har ikke PC	80	40	120
Sum	130	70	200

(a) Finn sannsynlighetene for at en tilfeldig utvalgt student har
(i) PC (ii) TV (iii) både PC og TV (iv) PC gitt TV.

(b) Finn sannsynlighetene for at 2 tilfeldig utvalgte studenter har
(i) begge PC (ii) bare en PC (iii) en PC og en TV (iv) begge PC gitt minst en har PC.

(c) Det siste punktet under (a) og (b) krever muligens en ekstra antakelse. Forklar!

27. Hva menes med en tilfeldig korthånd på 13 kort i bridge. Finn sannsynligheten for at en tilfeldig korthånd inneholder nøyaktig
(a) bare røde kort,

(b) 4 honnører,
(c) ingen honnører,
(d) ett ess,
(e) to ess.

28. En pokerhånd på fem kort trekkes ut tilfeldig fra en kortstokk. Hva er sannsynligheten for begivenhetene
(a) et par (to kort med samme verdi og tre kort med andre verdier),
(b) to par (to ganger to kort med samme verdi og et kort med annen verdi),
(c) 3 like (tre kort med samme verdi og to kort med andre verdier),
(d) 4 like (fire kort med samme verdi og et kort med annen verdi),
(e) flush (alle fem kort av samme slag enten spar, hjerter, ruter eller kløver).

29. En kortstokk på 52 kort blandes godt, og kortene legges ut ett etter ett fra toppen. Finn sannsynligheten for at første ess kommer
(a) som femte kort,
(b) før kort nummer ni.

30. $\star$ Betrakt produktet $(a+b)^n = (a+b)(a+b)\cdots(a+b)$ hvor n er ikke negativt heltall. Bruk kombinatorisk argument til å vise *Newtons binominalformel* :

$$\begin{aligned}(a+b)^n &= b^n + \binom{n}{1}ab^{n-1} + \binom{n}{2}a^2b^{n-2} + \cdots + \binom{n}{n-1}a^{n-1}b + a^n \\ &= \sum_{k=0}^{n}\binom{n}{k}a^k b^{n-k}\end{aligned}$$

31. $\star$ La situasjonen være som i Eksempel 22. Finn sannsynligheten for begivenhetene

E = Per, Pål og Espen kommer i samme gruppe,
F = Per, Pål og Espen kommer i forskjellig gruppe,
G = Per og Pål er i samme gruppe, mens Espen er i en annen gruppe,
H = Per har minst en kamerat med seg i sin gruppe.

32. $\star$ I Eksempel 23 finn sannsynligheten for at
(a) minst fem fra Østlandet står blant de 12 første i køen,
(b) det er like mange fra hver landsdel blant de 12 første i køen.

33. $\star$ I et Bridgespill hvor kortene stokkes godt, finn sannsynligheten for at
(a) Sparene fordeler seg med i spar til Nord, j spar til Øst, k spar til Syd og l spar til Vest $(i + j + k + l = 13)$.
(b) Essene fordeler seg med i ess til Nord, j ess til Øst, k ess til Syd og l ess til Vest $(i + j + k + l = 4)$.
Hvilken av "sitsene" 4000, 3100, 2200, 2110 og 1111 er mest sannsynlig (uansett spiller).

34. Undersøk din foretrukne programvare om antallet av ulike typer utvalg lar seg lett beregne, i første rekke $(N)_s$ og $\binom{N}{s}$. Disse er ofte tilgjengelige på

lommeregnere som egne taster, med betegnelsen P for "permutations"og C for "combinations". Undersøk din lommeregner.

35. Hvordan vil du med din programvare simulere tilfeldig fordeling av korthender med 13 kort til hver av de 4 spillerne ved et bridgebord?

Kapittel 4

Betinget sannsynlighet og uavhengighet

Vi skal i dette kapitlet prøve å klarlegge begrepene betinget sannsynlighet og uavhengighet, samt etablere de viktigste regnereglene knyttet til disse begrepene. Dette utvider vår verktøykasse for stokastisk modellbygging og sannsynlighetsberegning.

4.1 Betingede sannsynlighetsmodeller

Vi har et eksperiment, eller annen form for iakttakelse, med utfallsrom

$$\Omega = \{u_1, u_2, u_3, \ldots\}$$

og sannsynligheter gitt ved

$$P(u_1), P(u_2), P(u_3), \ldots$$

La oss tenke oss at eksperimentet er utført og at vi blir gitt den opplysningen at en bestemt begivenhet B har inntruffet. Dette betyr at et eller annet av de gunstige utfall for B har inntruffet, vi vet bare ikke hvilket. Det er nå klart at den opprinnelige modellen ikke nødvendigvis passer lenger. For det første vet vi nå at en rekke utfall, nemlig alle utfall i $\overline{B}$, ikke kan ha inntruffet. For det andre er det rimelig at sannsynligheten for utfall som fortsatt er tenkelig bør endres. Vi ønsker derfor å lage en modifisert modell hvor vi tar hensyn til den tilleggsinformasjon at begivenheten B har inntruffet. [1]

[1] Vi antar hele tiden at begivenheten B er slik at $P(B) > 0$.

Definisjon : Den betingede sannsynlighet $P_B(u)$ for utfallet u gitt begivenheten B er

$$\begin{aligned} P_B(u) &= 0 \qquad \text{for } u \in \overline{B} \\ &= \tfrac{P(u)}{P(B)} \quad \text{for } u \in B \end{aligned}$$

Dette betyr at vi tildeler sannsynlighet null til de utfall som ikke lenger er aktuelle , mens vi lar det innbyrdes forhold mellom sannsynlighetene for de utfall som fremdeles er aktuelle være det samme som før. Det er lett å sjekke at denne konstruksjonen oppfyller kravene A1 og A2 til en sannsynlighetsmodell:

$$\text{A1:} \qquad 0 \leq P_B(u) \leq 1 \quad \text{for alle} \quad u \in \Omega$$

$$\text{A2:} \quad \sum_{u \in \Omega} P_B(u) = \sum_{u \in B} \frac{P(u)}{P(B)} = \frac{1}{P(B)} \sum_{u \in B} P(u) = \frac{1}{P(B)} \cdot P(B) = 1$$

Merknad. I den betingede modell gitt B kunne vi bruke B som utfallsrom istedenfor Ω. Dette kan se ut som en forenkling, men i praksis ønsker man ofte å studere flere betingede modeller samtidig, og da er det greitt å beholde Ω som felles referanseramme.

Eksempel 1 : Et terningkast

Gitt modellen for et kast med en rettferdig terning:

Utfall:	1	2	3	4	5	6
Sannsynlighet:	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

Etter et kast med terningen opplyses det at den ikke viste seks øyne. Vi vet demed at begivenheten $B = \{1, 2, 3, 4, 5\}$ må ha inntruffet. Vår intuisjon sier nå at den betingede modell gitt B bør se slik ut:

Utfall:	1	2	3	4	5	6
Sannsynlighet:	$\frac{1}{5}$	$\frac{1}{5}$	$\frac{1}{5}$	$\frac{1}{5}$	$\frac{1}{5}$	0

dvs. de 5 utfall som fortsatt er aktuelle bør fortsatt ha samme sannsynlighet. Siden $P(B) = 5/6$ ser vi av definisjonen ovenfor at

$$P_B(u) = \frac{P(u)}{P(B)} = \frac{1/6}{5/6} = \frac{1}{5} \quad \text{for alle} \quad u \in B$$

som altså stemmer med vår intuisjon. Anta isteden at terningen er falsk og konstruert slik at sjansen for sekser er redusert (som medfører at sjansen for ener er øket). Anta at en brukbar modell for et kast med denne terningen er gitt ved

$$\begin{array}{lll} P(1) = 0.19 & P(2) = 0.17 & P(3) = 0.17 \\ P(4) = 0.17 & P(5) = 0.17 & P(6) = 0.13 \end{array}$$

I dette tilfelle blir $P(B) = 0.87$, slik at den betingede modell blir

Utfall:	1	2	3	4	5	6
Sannsynlighet:	$\frac{0.19}{0.87}$	$\frac{0.17}{0.87}$	$\frac{0.17}{0.87}$	$\frac{0.17}{0.87}$	$\frac{0.17}{0.87}$	0

La nå A være en vilkårlig begivenhet i utfallsrommet Ω. Ifølge A3 bør den betingede sannsynlighet for begivenheten A gitt B være summen av de betingede sannsynligheter for de utfall som er gunstige for A, dvs.

$$P_B(A) = \sum_{u \in A} P_B(u)$$

Med denne definisjonen vil vi få de samme regneregler for betingede sannsynligheter som for vanlige sannsynligheter (E1-E7 i kapittel 2.4). I tillegg har vi noen nye regneregler. Den første er:

$$P_B(A) = \frac{P(A \cap B)}{P(B)}$$

Begrunnelse : Del summen ovenfor opp i to deler, de utfall i A som også er med i B, og de utfall i A som også er med i $\overline{B}$. Ifølge definisjonen av betinget sannsynlighet, er alle ledd i den siste summen null, mens i den første summen blir fellesfaktoren $1/P(B)$ multiplisert med en sum, som ifølge A3 er $P(A \cap B)$. Forsøk selv å formalisere begrunnelsen.

Dette betyr at enhver sannsynlighet i den betingede modell gitt B kan beregnes ut fra sannsynligheter i den opprinnelige modell. I praksis er det vanlig å bruke en annen skrivemåte for betingede sannsynligheter. Istedenfor $P_B(A)$ skriver man $P(A \mid B)$ hvor symbolkombinasjonen $A \mid B$ uttales "A gitt B", og vår fundamentale formel lyder derfor

$$\boxed{\text{E8.} \qquad P(A \mid B) = \frac{P(A \cap B)}{P(B)}}$$

Eksempel 2 : Korttrekning
Et kort trekkes tilfeldig fra en kortstokk. La

A = Sort kort
B = Honnør (ess, konge, dame eller knekt)

Her blir

$$P(A) = \frac{g}{m} = \frac{26}{52}\ ,\ P(B) = \frac{g}{m} = \frac{16}{52}\ ,\ P(A \cap B) = \frac{g}{m} = \frac{8}{52}$$

slik at

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{8/52}{16/52} = \frac{1}{2}$$

mens

$$P(B \mid A) = \frac{P(A \cap B)}{P(A)} = \frac{8/52}{26/52} = \frac{4}{13}$$

Eksempel 3 : Poker
Anta at alle $m = \binom{52}{5}$ mulige korthender er like sannsynlige, og la

A = korthånden består av 5 spar
B = korthånden består av 5 sorte kort

Her blir $P(B) = \binom{26}{5}/\binom{52}{5}$ og $P(A \cap B) = P(A) = \binom{13}{5}/\binom{52}{5}$, slik at

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{\binom{13}{5}}{\binom{26}{5}} = \frac{1\,287}{65\,780} = 0.0196$$

Eksempel 4 : To terningkast.
Anta at alle $m = 36$ mulige par av øyne er like sannsynlige. La

A = Begge terninger viser 6 øyne
B = Minst en terning viser 6 øyne
C = Sum øyne er minst 10

Vi har $P(A) = 1/36$, $P(B) = 11/36$ og $P(C) = 6/36$, og finner at

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{1/36}{11/36} = \frac{1}{11}$$

$$P(B \mid C) = \frac{P(B \cap C)}{P(C)} = \frac{5/36}{6/36} = \frac{5}{6}$$

Det overlates til leseren å beregne flere betingede sannsynligheter :
$P(B \mid A)$, $P(C \mid B)$, $P(C \mid A)$ og $P(A \mid C)$.

4.2 Regneregler for betingede sannsynligheter

Tar vi utgangspunkt i den fundamentale formelen E8 og multipliserer med $P(B)$ på begge sider av likhetstegnet, får vi

E9. $\quad P(A \cap B) = P(B) \cdot P(A \mid B)$

Selv om denne formelen uttrykker samme saksforhold som E8, poengterer den en annen bruksmåte: Når $P(B)$ og $P(A \mid B)$ er kjent kan $P(A \cap B)$ beregnes. Dette vil vi dra nytte av senere ved beregning av sannsynligheter. Dersom vi lar A og B bytte rolle i E8 får vi $P(B \mid A) = P(A \cap B)/P(A)$. Innsettes her uttrykket for $P(A \cap B)$ fra E9, får vi

E10. $\quad P(B \mid A) = \dfrac{P(B) \cdot P(A \mid B)}{P(A)} \quad$ **(Bayes lov)**

som er en viktig regneregel i praksis. En annen viktig regneregel fås ved følgende betraktninger. Anta at

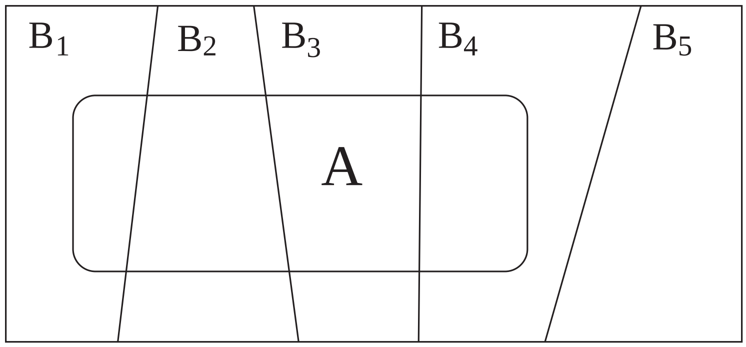
$$\Omega = B_1 \cup B_2 \cup \cdots \cup B_r$$

er en disjunkt oppsplitting av den sikre begivenhet (se Figur 4.1). La A være en vilkårlig begivenhet. Vi kan da skrive

$$A = (A \cap B_1) \cup (A \cap B_2) \cup \cdots \cup (A \cap B_r)$$

Siden dette er en disjunkt union medfører E7 at

E11. $\quad P(A) = P(B_1) \cdot P(A \mid B_1) + P(B_2) \cdot P(A \mid B_2) + \cdots + P(B_r) \cdot P(A \mid B_r)$



Figur 4.1: Beregning av sannsynlighet ved betinging

I mange problemstillinger hvor $P(A)$ ikke lar seg beregne direkte, er det mulig å splitte opp den sikre begivenhet på en slik måte at sannsynlighetene på høyresiden av E11 er kjent eller kan beregnes. Vi sier da at $P(A)$ blir beregnet ved betinging (mhp. $B_1, B_2, \dots B_r$). Ved beregning av nevneren i Bayes lov er det ofte aktuelt med betinging. Siden $\Omega = B \cup \overline{B}$ får vi et viktig spesialtilfelle av E11:

$$P(A) = P(B) \cdot P(A \mid B) + P(\overline{B}) \cdot P(A \mid \overline{B})$$

I mange situasjoner er det naturlig å konstruere sannsynlighetsmodellen ved å gjøre antakelser om noen av de betingede sannsynligheter i modellen. Når man skal løse sannsynlighetsteoretiske problemer i praksis, hender det ofte at man lar være å gi en fullstendig beskrivelse av sannsynlighetsmodellen, man nøyer seg med å gjøre tilstrekkelig mange ad hoc antakelser om modellen til at de søkte sannsynligheter lar seg beregne. Man må da selvsagt sørge for at de valgte antagelser ikke står i motstrid med hverandre. Nyttige formler i denne forbindelse er nettopp E9, E10 og E11. Sannsynligheten på venstresiden lar seg bestemme ved å gjøre antagelser om sannsynlighetene på høyresiden. Såkalte urnemodeller kan bidra til å klargjøre begrepene.

Eksempel 5 : To urner
Gitt to urner, den første med 4 hvite og 2 sorte kuler, den andre med 2 hvite og 2 sorte kuler. Først velges tilfeldig en av urnene, deretter trekkes tilfeldig en kule fra den valgte urne. Vi ønsker å beregne sannsynligheten for at den valgte kule er hvit. Definer begivenhetene

U_i = Urne nr. i velges, $i = 1, 2$.

$H =$ Hvit kule trekkes

Tilfeldig valg av urne og tilfeldig trekning fra urnen betyr at

$$P(U_1) = P(U_2) = \frac{1}{2}$$

$$P(H \mid U_1) = \frac{4}{6}, \quad P(H \mid U_2) = \frac{2}{4}$$

Ifølge regneformlene ovenfor får vi

$$P(U_1 \cap H) = P(U_1) \cdot P(H \mid U_1) = \frac{1}{2} \cdot \frac{4}{6} = \frac{1}{3}$$

$$P(U_2 \cap H) = P(U_2) \cdot P(H \mid U_2) = \frac{1}{2} \cdot \frac{2}{4} = \frac{1}{4}$$

slik at

$$P(H) = P(U_1 \cap H) + P(U_2 \cap H) = \frac{1}{3} + \frac{1}{4} = \frac{7}{12}$$

mens sannsynligheten for sort kule (S) blir $P(S) = 5/12$. Dette kan enten regnes ut på samme måte eller enklere ved å bruke at $P(S) = 1 - P(H)$. Gitt at den uttrukne kule er hvit. Sannsynligheten for at den kom fra urne nr. 1 blir

$$P(U_1 \mid H) = \frac{P(U_1 \cap H)}{P(H)} = \frac{1/3}{7/12} = \frac{4}{7}$$

mens sannsynligheten for at den kom fra urne nr. 2 blir $P(U_2 \mid H) = 3/7$. Dette kan enten regnes ut på samme måte eller enklere ved å bruke at $P(U_2 \mid H) = 1 - P(U_1 \mid H)$. Det kan vi fordi U_1 og U_2 er komplementære begivenheter (dvs. $U_2 = \overline{U_1}$) og betingede sannsynligheter oppfyller samme regneregler som vanlige sannsynligheter.

Eksempel 6 : To trekninger
Gitt en urne med 2 hvite og 3 sorte kuler. Først trekkes tilfeldig en kule fra urnen, deretter trekkes en ny kule tilfeldig fra de gjenværende kuler i urnen. Definer følgende begivenheter

E = Hvit kule i første trekning
F = Hvit kule i annen trekning

Ordet tilfeldig betyr at ved hver trekning har alle kulene i urnen samme sannsynlighet for å bli trukket ut, dvs.

$$P(E) = \frac{2}{5} \qquad P(\overline{E}) = \frac{3}{5}$$
$$P(F \mid E) = \frac{1}{4} \qquad P(F \mid \overline{E}) = \frac{2}{4}$$

Vi får derfor

$$P(E \cap F) = P(E) \cdot P(F \mid E) = \frac{2}{5} \cdot \frac{1}{4} = \frac{1}{10}$$
$$P(\overline{E} \cap F) = P(\overline{E}) \cdot P(F \mid \overline{E}) = \frac{3}{5} \cdot \frac{2}{4} = \frac{3}{10}$$

og følgelig

$$P(F) = P(E \cap F) + P(\overline{E} \cap F) = \frac{1}{10} + \frac{3}{10} = \frac{2}{5}$$

Videre blir

$$P(E \cap \overline{F}) = P(E) \cdot P(\overline{F} \mid E) = \frac{2}{5} \cdot \frac{3}{4} = \frac{3}{10}$$
$$P(\overline{E} \cap \overline{F}) = P(\overline{E}) \cdot P(\overline{F} \mid \overline{E}) = \frac{3}{5} \cdot \frac{2}{4} = \frac{3}{10}$$

og følgelig

$$P(\overline{F}) = P(E \cap \overline{F}) + P(\overline{E} \cap \overline{F}) = \frac{3}{10} + \frac{3}{10} = \frac{3}{5}$$

som vi kunne innsett direkte fra formelen $P(\overline{F}) = 1 - P(F)$. En alternativ måte å studere denne problemstillingen på er å anta at alle $(5)_2 = 20$ ordnede utvalg av 2 kuler er like sannsynlige. Det overlates til leseren å påvise at dette leder til samme sannsynligheter. Gitt at annen trekning ga hvit kule. Sannsynligheten for at også første trekning ga hvit kule blir

$$P(E \mid F) = \frac{P(E \cap F)}{P(F)} = \frac{1/10}{2/5} = \frac{1}{4}$$

mens sannsynligheten for ikke hvit (sort) kule blir 3/4. Dette kan enten regnes ut på samme måte eller ved $P(\overline{E} \mid F) = 1 - P(E \mid F)$.

Eksempel 7 : Tennis

I herretennis spilles best av 5 sett, dvs. at kampen vunnet straks en av spillerne har vunnet 3 sett. To spillere A og B som rangeres likt spiller mot hverandre. Hva er sannsynligheten for at A vinner kampen dersom

(a) A har vunnet to sett og B ett

(b) A har vunnet de to første sett

(c) A har vunnet første sett

Vi får ved å betinge mhp. hvem som vinner neste sett

$$\begin{aligned} p_a &= \frac{1}{2}\cdot 1 + \frac{1}{2}\cdot\frac{1}{2} = \frac{3}{4} \\ p_b &= \frac{1}{2}\cdot 1 + \frac{1}{2}\cdot p_a = \frac{7}{8} \\ p_c &= \frac{1}{2}\cdot p_b + \frac{1}{2}\cdot\frac{1}{2} = \frac{11}{16} \end{aligned}$$

Vi har her brukt at når begge har vunnet like mange sett, så er sannsynligheten 1/2 for at A vinner. Vi ser at når vi regner ut sannsynlighetene i denne rekkefølgen, kan vi bruke resultatet av den foregåenede beregning i den neste.

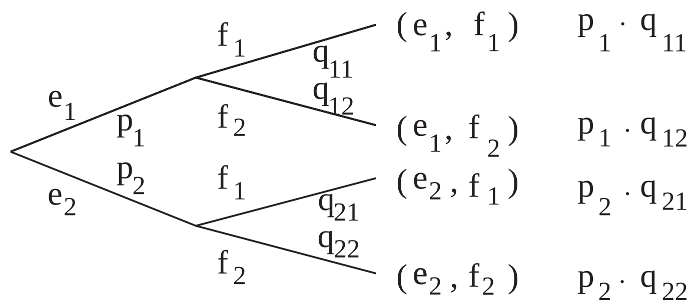
Eksemplet har en viss historisk interesse som et av de første sannsynlighetsteoretiske problemer som ble studert i det vitenskapelige miljø, riktignok i en annen forkledning: Hva er en rettferdig deling av potten når et spill må avbrytes før det er ferdig? Etter flere feilaktige løsninger på 1400-tallet, ga Pascal, Fermat og Huygens hver sin korrekte løsning på 1600-tallet. Pascal og Fermat brukte kombinatoriske argumenter, mens løsningen ovenfor er i Huygens' ånd, selv om han ikke brukte begrepet sannsynlighet.

Leseren kan selv forsøke å utvikle idéen videre til å løse "best av 7-spill", slik tilfellet er i Stanley Cup i ishockey.

De tre eksemplene ovenfor kan oppfattes som *trinnvise eksperimenter*, dvs. at utfallet av hvert trinn definerer betingelsene for neste trinn. I Eksempel 5 var første trinn valg av urne, annet trinn valg av kule fra den valgte urne. I Eksempel 6 var første trinn valg av en kule, annet trinn valg av kule blant de gjenværende i urnen etter at den første er trukket. I Eksempel 7 vil hvert sett være et nytt trinn.

Et trinnvis eksperiment kan beskrives ved et såkalt *sannsynlighetstre*. I tilfellet med to trinn og to utfall på hvert trinn ser et tre ut som i Figur 4.2.

Et sannsynlighetstre leses fra venstre mot høyre. Vi starter med et forgreningspunkt (roten av treet) som svarer til første trinn. Ut fra dette går grener, en for hvert mulig utfall av første trinn (her e_1 og e_2), på hver sin gren skriver vi de respektive sannsynligheter (her p_1 og p_2). Langs hver av disse grener nås et nytt forgreningspunkt, som svarer til annet trinn. Fra



Figur 4.2: Sannsynlighetstre

et slikt punkt går nye grener, en for hvert mulig utfall av annet trinn (her f_1 og f_2), og på hver gren skriver vi de sannsynligheter (q_{i1} og q_{i2}) som er aktuelle for den forgrening vi er på ($i = 1$ eller 2), alt etter utfallet av de foregående trinn. De ulike mulige hendelsesforløp av eksperimentet er påført ved enden, dette svarer til de mulige utfall av hele eksperimentet. Til hvert utfall svarer altså en sti gjennom treet fra roten til ytterste kvist.

Vi finner sannsynligheten for hvert utfall ved å multiplisere sammen sannsynlighetene langs den sti som svarer til dette utfallet. Disse sannsynlighetene kan så påføres ved enden av treet, merk at alle disse må summere seg opp til en. Sannsynligheten for en bestemt begivenhet A kan nå finnes ved å summere de sannsynlighetene som svarer til stier som er gunstige for begivenheten A.

Et sannsynlighetstre kan være et nyttig hjelpemiddel til å gi en oversiktlig beskrivelse av mer kompliserte sannsynlighetsteoretiske situasjoner, dvs. trinnvise situasjoner med mer enn to trinn og med flere (ofte varierende antall) utfall på hvert trinn, se Eksempel 9 nedenfor Slik trebeskrivelse viser seg å være spesielt nyttig ved analyse av sekvensielle beslutningsproblemer under usikkerhet (se Kapittel 16).

Vi har utelatt en formell begrunnelse for at den trinnvise konstruksjon ovenfor virkelig gir en sannsynlighetsmodell, og at de betingede sannsynligheter kan gjenfinnes fra denne. Det er ikke vanskelig, men krever en viss formalisme, som i figuren. Leseren kan heller prøve å gjennomskue dette ved å lage et sannsynlighetstre, konstruere modell med utfall og tilhørende sannsynligheter og verifisere beregningen i Eksempel 5 og 6 (se Oppgave 19).

I Kapittel 3 studerte vi bl.a. ordnede utvalg uten tilbakelegging. Dette kan alternativt studeres som et trinnvis eksperiment, og det er kanskje slik trekningen lettest gjennomføres i praksis. Det er lett å verifisere at tilfeldig utvelgelse på hvert trinn leder til et tilfeldig utvalg i den forstand vi definerte det i Kapittel 3. Nedenfor følger argumentasjonen for tilfellet med utvalg på to elementer, som lett kan utvides trinnvis til flere. Stoffet er $\star$-merket og kan overspringes uten fare for å miste tråden.

$\star$ Gitt en populasjon på N elementer nummerert fra 1 til N. Først trekkes tilfeldig et element, deretter trekkes tilfeldig et nytt element fra de gjenværende $N-1$ elementer. La e_i bety at første trekning gir element nr. i og f_j bety at annen trekning gir element nr. j. Siden hver trekning foregår tilfeldig har vi følgende delmodeller

Trinn 1:	Utfall:	e_1	e_2	$\ldots$	e_N
	Sannsynlighet:	$\frac{1}{N}$	$\frac{1}{N}$	$\ldots$	$\frac{1}{N}$

Dersom trinn 1 ga e_i som resultat har vi følgende modell for trinn 2:

Trinn 2:	Utfall:	f_1	f_2	$\ldots$	f_i	$\ldots$	f_N
	Sannsynlighet:	$\frac{1}{N-1}$	$\frac{1}{N-1}$	$\ldots$	0	$\ldots$	$\frac{1}{N-1}$

Som utfallsrom for to-trinns eksperimentet bruker vi de ulike kombinasjonene (e_i, f_j), og to-trinnsmodellen er dermed definert ved

$$\begin{aligned} P((e_i, f_j)) &= 0 \text{ for j=i} \\ &= \frac{1}{N} \cdot \frac{1}{N-1} \text{ for } j \neq i \end{aligned}$$

Siden $\frac{1}{N} \cdot \frac{1}{N-1} = \frac{1}{(N)_2}$, ser vi at denne to-trinnsmodellen er ekvivalent med en modell hvor vi antar at alle $(N)_2$ mulige ordnede utvalg uten tilbakelegging er like sannsynlige. Den eneste forskjell er at den foreliggende modell nytter et større utfallsrom og tildeler sannsynlighet null til de utfall hvor samme element forekommer i begge trekninger.

La oss se hvordan vi i praksis ville beregnet sannsynligheten for begivenheten F_j = annen trekning gir element nr. j. La E_i = første trekning gir element nr. i. Vi kan anta at $P(E_i) = 1/N$ og $P(F_j \mid E_i) = 0$ for $j = i$ og $1/(N-1)$ for $j \neq i$. Siden $\Omega = E_1 \cup E_2 \cup \cdots \cup E_N$ er en disjunkt oppsplitting av den sikre begivenhet følger at

$$P(F_j) = \sum_i P(E_i)P(F_j \mid E_i) = \sum_{i \neq j} \frac{1}{N} \cdot \frac{1}{N-1} = \frac{1}{N}$$

Sml. ekvivalensloven for tilfeldig ordnet utvalg.

4.3 Subjektive sannsynligheter

I de fleste av problemstillingene vi har studert til nå har vi kunnet fastlegge sannsynligheter med en viss grad av objektivitet. Vi har i det innledende kapitlet antydet at det finnes sannsynlighetsteorier som tillater *subjektive sannsynligheter* (også kalt personlige sannsynligheter) i motsetning til oppfatningen om at sannsynligheter er noe objektivt ved det fenomen som studeres.

Vi vil nedenfor presentere en del eksempler hvor de sannsynligheter som inngår nødvendigvis vil være av subjektiv natur. Temaet subjektive sannsynligheter er omfattende og bør vies et eget kapittel (Kapittel 16). Vi ønsker imidlertid her å belyse visse aspekter ut fra det kjennskap vi til nå har fått om regning med sannsynligheter, samtidig får vi belyst regnereglene ytterligere.

La oss tenke oss at du er villig til, på et mer eller mindre subjektivt grunnlag, å evaluere sannsynligheten for to eller flere begivenheter. Dersom du ønsker at de vanlige (og naturlige?) regnereglene for sannsynligheter skal gjelde, kan du ikke velge verdiene på disse sannsynlighetene fritt, du må følge visse spilleregler. La oss anta at du først skal evaluere sannsynligheten for en bestemt begivenhet A. Enhver verdi i intervallet fra null til en er tillatt, og la oss si at ditt valg er $P(A)$. Dette valg kan kanskje virke besynderlig på oss andre, men så lange ditt valg ikke har konsekvenser for oss, er det ingen grunn til å intervenere. La oss nå tenke oss du skal evaluere sannsynligheter i forbindelse med to begivenheter A og B som studeres i sammenheng. I denne situasjon kan du evaluere en rekke sannsynligheter f. eks. $P(A)$, $P(B \mid A)$, $P(B \mid \overline{A})$ osv. Legg merke til at når de tre første sannsynligheter er spesifisert så har du, dersom de vanlige regnereglene for sannsynligheter skal gjelde, også fastlagt $P(B)$, $P(\overline{A})$, $P(\overline{B})$, $P(A \mid B)$, $P(A \mid \overline{B})$, $P(\overline{A} \mid B)$, $P(\overline{A} \mid \overline{B})$. Kan hende har du evaluert $P(B)$ eller noen av de andre direkte og oppdager at dette ikke stemmer overens med det du får ved beregning ut fra de tre første sannsynlighetene. Du bør kanskje da vurdere om ikke en eller flere av dine sannsynligheter bør modifiseres.

Eksempel 8 : Fotball
Cupfinalen i fotball mellom Lyn og Brann forestår. Anta at du skal inngå et veddemål om utfallet, og at du i denne forbindelse vurderer sannsynligheten for at Brann vinner (B) til å være $P(B) = p$. Ved ankomst til banen ser du at det er sjenerende sol og vind slik at banevalget kan ha en viss innflytelse på utfallet av kampen. La A være begivenheten at Brann ved banevalget får den gunstige banehalvdel i første omgang. Siden banevalget avgjøres ved

myntkast vil du antagelig sette $P(A) = 1/2$. Anta at du setter $P(B \mid A) = q$ og $P(B \mid \overline{A}) = r$. Vi vil anta at $q > p > r$. For at dine vurderinger skal samsvare i henhold til regnereglene må p, q og r være slik at

$$p = \frac{1}{2}(q + r)$$

Kan hende dine vurderinger ikke samsvarer med dette, men at du er villig til å revurdere situasjonen. Det må skje raskt og da er en mulighet å bruke en veid sum p, q og r som et nytt "anslag" for $P(B)$, kanskje velger du gjennomsnittet $(p + q + r)/3$. Eksempelvis dersom $p = 0.70$, $q = 0.80$ og $r = 0.50$ slik at $(q + r)/2 = 0.65$, får vi $(p + q + r)/3 = 0.67$. Det kan vises at denne metoden til å modifisere de opprinnelige sannsynlighetene i en viss forstand svarer til at alle de tre første vurderingene brukes på nytt og tillegges like stor vekt, avgjørende her er at $P(A) = \frac{1}{2}$ er noenlunde udiskutabelt. Implisitt blir da også $P(B \mid A)$ og $P(B \mid \overline{A})$ modifisert, med de tall vi brukte ovenfor til henholdsvis 0.82 og 0.52. Siden veddemålet skal inngås før banevalget, er det $P(B)$ som er relevant, og vi sløyfer derfor en nærmere redegjørelse. Et alternativt scenario kan være at du innser at den opprinnelige p-verdi var noe optimistisk, men at du fester lit til q og r. I så fall bør den modifiserte verdi av $P(B)$ være $(q + r)/2$, i talleksemplet lik 0.65, dvs. ikke mye forskjellig fra 0.67.

Hensikten med dette eksemplet var å vise at vurdering av subjektive sannsynligheter krever betydelig omtanke, kombinert med teoretisk innsikt. Eksperimenter har vist at det i praksis er lett å gjøre vurderinger som ikke "henger sammen".

La oss illustrere noen andre poenger med nye eksempler.

Eksempel 9 : Gullfisken
Per har fått beskjed om å mate gullfisken mens foreldrene er på ferie. På forhånd vurderer Per's far at sannsynligheten for at sønnen glemmer å mate fisken er 1/4. Hvis sønnen husker å mate fisken er sjansen for at den overlever ferien lik 9/10, men hvis han glemmer å mate fisken er sannsynligheten bare 1/2. Da foreldrene kom hjem var gullfisken død, hva er da sannsynligheten for at Per har glemt å mate den? La

B = Gullfisken døde
A = Per glemte å mate gullfisken

Vi har nå at $P(A) = 1/4, P(B \mid A) = 1/2, P(B \mid \overline{A}) = 1/10$, og vi ønsker å beregne $P(A \mid B)$. Vi kan her benytte Bayes lov.

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

Størrelsene i telleren er kjente, mens nevneren beregnes ved

$$P(B) = P(A) \cdot P(B \mid A) + P(\overline{A}) \cdot P(B \mid \overline{A}) = \frac{1}{4} \cdot \frac{1}{2} + \frac{3}{4} \cdot \frac{1}{10} = \frac{1}{5}$$

Vi får derfor

$$P(A \mid B) = \frac{1/4 \cdot 1/2}{1/5} = \frac{5}{8}$$

I dette eksemplet vil alle sannsynligheter som Pers far har benyttet i sine beregninger nødvendigvis måtte være av subjektiv natur. I slike situasjoner blir $P(A) = 1/4$ gjerne omtalt som *apriorisannsynligheten* for begivenheten A, mens $P(A \mid B) = 5/8$ er *aposteriorisannsynligheten* for A (gitt B).

Eksempel 10 : En reise
Hansen skal på et møte på den andre siden av fjellet, og han akter å forsøke å krysse fjellovergangen med egen bil. Skulle veien vise seg å være blokkert av snø, kan han snu og ta toget (som kjører i tunnel), og likevel nå fram i tide, dersom intet uhell inntreffer. Han ønsker spesielt å evaluere sannsynligheten for å komme for sent (C), men finner det lettere å evaluere denne gitt det som skjedde underveis: veien åpen (A) eller ikke, uhell (B) eller ikke, samt evaluere sannsynlighetene for hver av disse mulighetene. Hansens vurderinger gis her i form av et sannsynlighetstre, se Figur 4.3.

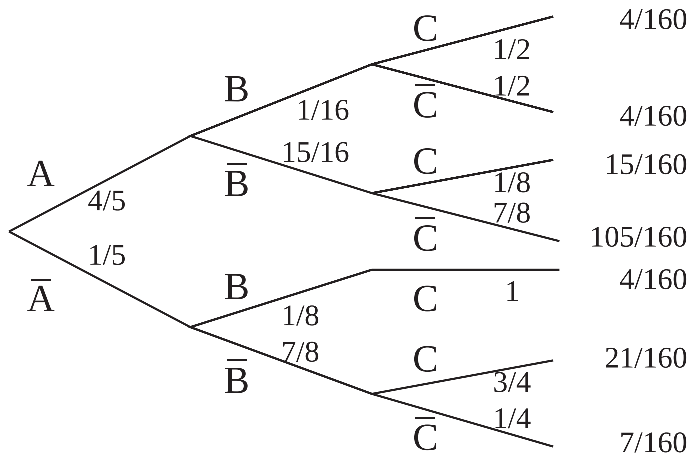
Her er eksempelvis $P(A) = 4/5$, $P(B \mid A) = 1/16$ og $P(C \mid A \cap B) = 1/2$. Ved å multiplisere sammen sannsynlighetene langs grenene finner vi tallene til høyre for treet, eksempelvis er (jfr. Oppgave 10):

$$P(A \cap B \cap C) = P(A) \cdot P(B \mid A) \cdot P(C \mid A \cap B) = \frac{4}{5} \cdot \frac{1}{16} \cdot \frac{1}{2} = \frac{4}{160}$$

Vi finner at

$$P(C) = \frac{4}{160} + \frac{15}{160} + \frac{4}{160} + \frac{21}{160} = \frac{11}{40}$$

Et teoretisk grunnlag for subjektiv sannsynlighet er utviklet i vårt århundre som en integrert del av generell beslutningsteori. For en beslutningstaker som



Figur 4.3: Hansens sannsynligheter

står overfor beslutninger under usikkerhet kan man, ved å anta at han/hun oppfyller visse adferdsforutsetninger, utlede at han/hun vil (bør) handle som om han/hun regnet med sannsynligheter som følger de vanlige regnereglene for slike. Disse sannsynlighetene vil ifølge sin natur måtte være subjektive. Dette tema blir tatt opp til grundigere diskusjon i Kapittel 16.

4.4 Uavhengige begivenheter

La A og B være to begivenheter. Sannsynligheten for A er $P(A)$ (ubetinget), mens den betingede sannsynlighet for A gitt B er $P(A \mid B)$. I enkelte situasjoner viser det seg at begivenhetene A og B er slik at disse to sannsynlighetene er like, dvs. at $P(A \mid B) = P(A)$. Dette kan tolkes slik at det faktum at B har inntruffet ikke påvirker sjansen for at A skal inntreffe, og det er naturlig å si at begivenheten A er uavhengig av begivenheten B. På samme måte dersom $P(B \mid A) = P(B)$ er det naturlig å si at B er uavhengig av A. Av E8, E9 og E10 følger det imidlertid at følgende tre utsagn er ekvivalente (såframt $P(A) > 0$ og $P(B) > 0$)

(i) $P(A \mid B) = P(A)$

(ii) $P(B \mid A) = P(B)$

(iii) $P(A \cap B) = P(A) \cdot P(B)$

Siden det viste seg at uavhengighet er et symmetrisk begrep, er det naturlig å velge en symmetrisk definisjon:

> **Definisjon :** Begivenhetene A og B sies å være *uavhengige* dersom $P(A \cap B) = P(A) \cdot P(B)$

Eksempel 11 : Korttrekning
Et kort trekkes tilfeldig fra en kortstokk. La

$A =$ Sort kort
$B =$ Honnør

Her er $P(A) = 1/2$, $P(B) = 4/13$ og $P(A \cap B) = 8/52 = 2/13$. Vi ser at $P(A \cap B) = P(A) \cdot P(B)$, og vi har dermed vist at begivenhetene A og B er uavhengige. Dette samsvarer vel med vår intuisjon.

Eksempel 12 : To terningkast
Anta en modell hvor alle $m = 36$ mulig utfall er like sannsynlige, og la

$E_i =$ første kast gir i øyne, $i = 1, 2, \ldots, 6$
$F_j =$ annet kast gir j øyne, $j = 1, 2, \ldots, 6$

Her er $P(E_i \cap F_j) = 1/36$, mens $P(E_i) = 6/36 = 1/6$ og $P(F_j) = 6/36 = 1/6$. Vi ser at $P(E_i \cap F_j) = P(E_i) \cdot P(F_j)$, slik at vi kan konkludere med at E_i og F_j er uavhengige. Det kan vises at dersom E og F er to vilkårlige begivenheter som refererer seg til henholdsvis første og annet kast, så er de uavhengige. Vi sier at de to kastene utgjør to *uavhengige deleksperimenter.*

Vi har også bruk for begrepet uavhengighet for mer enn to begivenheter:

> **Definisjon :** De n begivenheter $A_1, A_2, \ldots, A_n$ sies å være *uavhengige* dersom sannsynligheten for snittet av ethvert utvalg av disse begivenhetene er lik produktet av sannsynlighetene.

Vi merker oss f. eks. at dersom tre begivenheter A, B og C skal være uavhengige, så må

1. $P(A \cap B \cap C) = P(A) \cdot P(B) \cdot P(C)$
2. $P(A \cap B) = P(A) \cdot P(B), P(A \cap C) = P(A) \cdot P(C),$
   $P(B \cap C) = P(B) \cdot P(C)$

I dagligtale, ikke minste i sport og spill, brukes ofte begrepet “odds”. En begivenhet A har odds lik $O(A) = P(A)/P(\overline{A})$, mens for gitt B er odds for A endret til $O(A \mid B) = P(A \mid B)/P(\overline{A} \mid B)$. I Eksempel 8 med gullfisken blir $O(A) = 1 : 3$, mens $O(A \mid B) = 5 : 3$ og $O(A \mid \overline{B}) = 5 : 27$. Merk at uttrykt ved odds blir den tilhørende sannsynlighet $P = O/(1 + O)$. I totalisatorspill betyr et tilbudt odds hvor mange ganger innsatsen spilleren får ved vinst.

Et mål for grad av avhengighet er det såkalte *odds-forholdet* med verdier mellom 0 og ∞:

$$OR(A, B) = \frac{O(A \mid B)}{O(A \mid \overline{B})} = \frac{O(B \mid A)}{O(B \mid \overline{A})}$$

Dette er 1 hvis og bare hvis A og B er uavhengige, og større enn/mindre enn 1 alt ettersom A er mest sannsynlig sammen med B eller $\overline{B}$. Et alternativt mål er $G = (OR-1)/(OR+1)$, med verdi mellom -1 og +1, der 0 svarer til uavhengighet. I Eksempel 9 blir $OR(A, B) = 5 : 3/5 : 27 = 1 : 1/1 : 9 = 9$, mens $G = (9 - 1)/(9 + 1) = 0.8$.

4.5 Uavhengige eksperimenter og produktmodeller

I forrige avsnitt så vi hvordan vi i en modell kunne sjekke om begivenheter er uavhengige. I dette avsnitt skal vi se hvordan begrepet uavhengighet kan brukes den motsatte veien til konstruksjon av modeller: Har man forestillinger om at visse begivenheter er uavhengige, er det naturlig at disse forestillinger kommer til uttrykk i modellen. I praksis støter en ofte på situasjoner hvor et eksperiment kan betraktes som sammensatt av to eller flere deleksperimenter, som man forestiller seg ikke avhenger av hverandre. Det er da naturlig å anta at deleksperimentene er uavhengige, dvs. at en samling begivenheter som hører til hvert sitt deleksperiment er uavhengige.

Eksempel 13 : To terningkast
La oss som i Eksempel 11 definere

E_i = første kast viser i øyne
F_j = annet kast viser j øyne

Dersom hvert av kastene skjer med en rettferdig terning, er det rimelig å sette

$$P(E_i) = \frac{1}{6} \quad \text{for} \quad i = 1, 2, \ldots, 6$$

$$P(F_j) = \frac{1}{6} \quad \text{for} \quad j = 1, 2, \ldots, 6$$

Antar vi at de to kastene er uavhengige, får vi

$$P(E_i \cap F_j) = P(E_i) \cdot P(F_j) = \frac{1}{6} \cdot \frac{1}{6} = \frac{1}{36}$$

Selv om dette gir oss den samme modell for de to kastene som den vi hadde i Eksempel 12, er det viktig å merke seg den prinsipielle forskjell: I Eksempel 12 startet vi med en totalmodell for de to kastene og påviste at begivenheter vedrørende hvert sitt kast var uavhengige. I Eksempel 13 starter vi med sannsynligheter for begivenheter vedrørende hvert sitt kast, og benytter så forestillingen om uavhengige kast til å beregne sannsynligheten for begivenheter vedrørende begge kast. For å illustrere at dette virkelig representerer et nytt hjelpemiddel ved konstruksjon av sannsynlighetsmodeller, kan vi se på to kast med en falsk terning: Anta at det (f. eks. på bakgrunn av empirisk erfaring) er rimelig å sette

$$P(E_1) = 0.19, P(E_2) = P(E_3) = P(E_4) = P(E_5) = 0.17, P(E_6) = 0.13$$

og likeens

$$P(F_1) = 0.19, P(F_2) = P(F_3) = P(F_4) = P(F_5) = 0.17, P(F_6) = 0.13$$

Antar vi at de to kastene er uavhengige, får vi

$$\begin{array}{ccccccc}
P(E_1 \cap F_1) & = & P(E_1) \cdot P(F_1) & = & 0.19 \cdot 0.19 & = & 0.0361 \\
P(E_1 \cap F_2) & = & P(E_1) \cdot P(F_2) & = & 0.19 \cdot 0.17 & = & 0.0323 \\
P(E_2 \cap F_2) & = & P(E_2) \cdot P(F_2) & = & 0.17 \cdot 0.17 & = & 0.0289 \\
\vdots & & \vdots & & \vdots & & \vdots \\
P(E_6 \cap F_6) & = & P(E_6) \cdot P(F_6) & = & 0.13 \cdot 0.13 & = & 0.0169
\end{array}$$

Eksempel 14 : Brannalarm

Sannsynligheten for at en bestemt type brannalarm skal virke ved brann er 0.9. Hva er sannsynligheten for alarm (A) dersom man installerer tre slike alarmer som virker uavhengig av hverandre? La

$$A_i = \text{Alarm nr. } i \text{ virker}, \; i = 1, 2, 3.$$

Vi ser at

$$A = A_1 \cup A_2 \cup A_3 \quad \text{mens} \quad \overline{A} = \overline{A_1} \cap \overline{A_2} \cap \overline{A_3}$$

Det at alarmene virker uavhengig av hverandre tolkes dithen at $\overline{A_1}$, $\overline{A_2}$ og $\overline{A_3}$ blir uavhengige begivenheter. Vi får derfor

$$\begin{aligned} P(A) &= 1 - P(\overline{A}) = 1 - P(\overline{A_1} \cap \overline{A_2} \cap \overline{A_3}) \\ &= 1 - P(\overline{A_1}) \cdot P(\overline{A_2}) \cdot P(\overline{A_3}) \\ &= 1 - 0.1^3 = 1 - 0.001 = 0.999 \end{aligned}$$

Eksempel 15 : Myntkast inntil første kron
La oss for $i = 1, 2, 3, \ldots$ definere

$$A_i = \text{Kast nr. } i \text{ er mynt}$$

Med en rettferdig mynt setter vi $P(A_i) = \frac{1}{2}$ for $i = 1, 2, \ldots$. La B_n = første kron inntreffer i n'te kast. Vi får

$$B_n = A_1 \cap A_2 \cap \cdots \cap A_{n-1} \cap \overline{A_n}$$

Antar vi at kastene er uavhengige, får vi

$$\begin{aligned} P(B_n) &= P(A_1 \cap A_2 \cap \cdots \cap A_{n-1} \cap \overline{A_n}) \\ &= P(A_1) \cdot P(A_2) \cdots P(A_{n-1}) \cdot P(\overline{A_n}) \\ &= \frac{1}{2} \cdot \frac{1}{2} \cdots \frac{1}{2} \cdot \frac{1}{2} = (\frac{1}{2})^n \end{aligned}$$

et resultat som ble annonsert i Eksempel 2.12.
En modell konstruert som i eksemplene ovenfor kalles en *produktmodell*. Det kreves egentlig en formell begrunnelse for at dette kan gjøres på en logisk konsistent måte. Vi antyder dette som $\star$-merket stoff.

$\star$ Gitt et eksperiment som består av to deleksperimenter. Anta at en brukbar modell for første eksperiment er gitt ved:

Mulige utfall:	e_1	e_2	$\ldots$	e_m	
Sannsynligheter:	p_1	p_2	$\ldots$	p_m	slik at $\sum_{i=1}^{m} p_i = 1$

mens en brukbar modell for annet eksperiment er gitt ved :

Mulige utfall:	f_1	f_2	$\ldots$	f_n	
Sannsynligheter:	q_1	q_2	$\ldots$	q_n	slik at $\sum_{j=1}^{n} q_j = 1$

En modell for hele eksperimentet bruker utfallsrommet:

$$\begin{array}{rcccc}\Omega = & \{(e_1, f_1), & (e_1, f_2), & \ldots, & (e_1, f_n), \\ & (e_2, f_1), & (e_2, f_2), & \ldots, & (e_2, f_n), \\ & . & . & & . \\ & . & . & & . \\ & . & . & & . \\ & (e_m, f_1), & (e_m, f_2), & \ldots, & (e_m, f_n)\}\end{array}$$

Antar vi at de to deleksperimentene er uavhengige, er det rimelig å tildele utfallet (e_i, f_j) en sannsynlighet lik produktet av de faktorene

$$P((e_i, f_j)) = p_i q_j \quad i = 1, 2, \ldots, m, \ j = 1, 2, \ldots, n.$$

Vi ser at dette definerer en sannsynlighetsmodell, fordi

$$\sum_i \sum_j P((e_i, f_j)) = \sum_i \sum_j p_i q_j = (\sum_i p_i) \cdot (\sum_j q_j) = 1 \cdot 1 = 1$$

Det overlates til leseren å sjekke at sannsynlighetene fra deleksperimentene kan gjenvinnes på en konsistent måte fra totalmodellen. og at konstruksjonen kan utvides til et vilkårlig antall deleksperimenter. Vanskelig blir dette først når denne idéen skal generaliseres til utfallsrom som ikke er endelige. Det var slike grunnleggende problemer sannsynlighetsteoretikere ryddet opp i på 1900-tallet.

4.6 Binomiske forsøksrekker

Mange praktiske problemstillinger kan tilbakeføres til følgende situasjon:

Binomisk forsøksrekke. Det utføres en rekke forsøk (eksperimenter, iakttakelser), der

1. Hvert forsøk resulterer enten i suksess (S) eller fiasko (F).
2. Sannsynligheten for suksess er den samme i alle forsøk.
3. Forsøkene er uavhengige.

Anta at vi utfører n binomiske forsøk og la oss betegne sannsynligheten for suksess i hvert forsøk med p. Vi ønsker nå å beregne sannsynligheten for at vi i løpet av de n forsøkene observerer X suksesser: Siden forsøkene er uavhengige blir sannsynligheten for utfallet

$$SS \cdots SFF \cdots F$$

dvs. først x suksesser og deretter $n-x$ fiaskoer lik

$$p \cdot p \cdots p \cdot (1-p) \cdots (1-p) = p^x (1-p)^{n-x}$$

På tilsvarende måte blir sannsynligheten for ethvert annet spesifisert utfall bestående av x suksesser og $n-x$ fiaskoer også lik $p^x(1-p)^{n-x}$ (faktorenes orden er likegyldig). Siden det er $\binom{n}{x}$ måter å velge ut de x forsøkene som skal gi suksess, er det $\binom{n}{x}$ gunstige utfall for begivenheten at vi får x suksesser, og hvert av disse har sannsynlighet $p^x(1-p)^{n-x}$. Følgelig blir sannsynligheten for x suksesser i løpet av de n forsøkene lik

$$\binom{n}{x} p^x (1-p)^{n-x}$$

Denne formelen gjelder for $x = 0, 1, 2, \ldots, n$.

Eksempel 16 : Ti myntkast
Dette kan betraktes som en binomisk forsøksrekke med $n = 10$ forsøk, hvor kron tilsvarer suksess. Antar vi rettferdig mynt får vi at sannsynligheten for i alt x kron i løpet av de ti myntkastene blir

$$\binom{10}{x} (\frac{1}{2})^x (1 - \frac{1}{2})^{10-x} = \binom{10}{x} (\frac{1}{2})^{10}$$

Eksempelvis blir sannsynligheten for bare 3 kron

$$\binom{10}{3} (\frac{1}{2})^{10} = \frac{120}{1024}$$

Eksempel 17 : Ti terningkast
La oss anta at sekser betyr suksess. Situasjonen kan betraktes som en binomisk forsøksrekke med $n = 10$ forsøk og med suksess-sannsynlighet $1/6$. Sannsynligheten for x seksere i løpet av de ti kastene blir

$$\binom{10}{x} (\frac{1}{6})^x (1 - \frac{1}{6})^{10-x}$$

Eksempelvis blir sannsynligheten for 7 seksere

$$\binom{10}{7}(\frac{1}{6})^7(1-\frac{1}{6})^3$$

mens sannsynligheten for minst 7 seksere blir

$$\sum_{x=7}^{10}\binom{10}{x}(\frac{1}{6})^x(1-\frac{1}{6})^{10-x}$$

Eksempel 18 : Produksjonsprosess
En produksjonsserie består av $n = 100$ artikler. Hver artikkel klassifiseres som enten defekt eller intakt. Anta at for hver artikkel så er sannsynligheten for at den er defekt lik 0.05. Anta videre uavhengighet. Vi har dermed en binomisk forsøksrekke. Sannsynligheten for x defekte artikler i produksjonsserien blir

$$\binom{100}{x}0.05^x(1-0.05)^{100-x}$$

Eksempelvis blir sannsynligheten for 5 defekte

$$\binom{100}{5}0.05^5 0.95^{95}$$

mens sannsynligheten for høyst 5 defekte blir

$$\sum_{x=0}^{5}\binom{100}{x}0.05^x 0.95^{100-x}$$

Det finnes tabeller over binomiske sannsynligheter, som vil redusere det praktiske regnearbeid ved beregning av slike, se Tabell C.2 og i Appendiks C. Vi vil ta opp binomiske forsøk til bredere diskusjon i Kapittel 6.2, og en nærmere veiledning i bruken av tabeller utstår til da. En annen interessant problemstilling i forbindelse med binomiske forsøksrekker er å finne sannsynligheten for at første suksess kommer i n'te forsøk. I en situasjon hvor en observerer en binomisk forsøksrekke inntil første suksess kan det være hensiktsmessig å tenke seg de mulige utfall i utfallsrommet Ω som

$$S,\ \ FS,\ \ FFS,\ \ FFFS, \ldots$$

dvs. første suksess i første forsøk, første suksess i annet forsøk etc. Siden forsøkene er antatt å være uavhengige betyr det at disse utfallene tildeles sannsynlighetene

$$p,\ \ (1-p)p,\ \ (1-p)(1-p)p,\ \ (1-p)(1-p)(1-p)p, \ldots$$

med andre ord sannsynligheten for at første suksess kommer i n'te forsøk blir derfor

$$(1-p)^{n-1} \cdot p \quad \text{for} \quad n = 1, 2, 3, \ldots$$

Ved å benytte summeformelen for en geometrisk rekke, ser vi at disse sannsynlighetene summerer seg opp til en, slik at vi dermed har fått etablert en sannsynlighetsmodell. Legg merke til at dette innebærer at sannsynligheten for at vi observerer i det uendelige uten å få suksess er null, dette utfallet er følgelig heller ikke tatt med som et mulig utfall. Sannsynligheter av typen ovenfor blir gjerne kalt for *geometriske sannsynligheter*.

Vi har sett eksempel på slike sannsynligheter i Eksempel 15 hvor problemet var ventetiden til første kron i myntkast [2], i dette tilfellet er p=1/2 . Et annet eksempel vil være ventetiden til første sekser i terningkast, i dette tilfellet er $p = 1/6$. Sannsynligheten for at første sekser kommer i n'te kast er derfor

$$(\frac{5}{6})^{n-1} \cdot \frac{1}{6} \quad \text{for} \quad n = 1, 2, 3, \ldots$$

Enda et eksempel vil være ventetiden til første defekte artikkel i en produksjonsprosess som gjennomgående i det lange løp gir 5% defekte. Dersom vi antar uavhengighet med hensyn til defekt og intakt for de ulike produksjonsnumre, vil også denne situasjon falle innenfor rammen av teorien ovenfor, nå med $p = 0.05$. Sannsynligheten for at første defekt kommer som produksjonsnummer n er derfor:

$$0.95^{n-1} \cdot 0.05 \quad \text{for} \quad n = 1, 2, 3, \ldots$$

Vi vil komme tilbake til geometriske sannsynligheter i Kapittel 6.7.

[2] I dette eksemplet brukte vi riktignok en annen men likeverdig formalisering ved løsningen av problemet, vi ønsker å få vist begge.

4.7 ⋆Paradokser

Sannsynlighetsregning rommer en rekke paradokser som kan ha praktiske konsekvenser. Vi gir her to eksempler på statistiske problemstillinger med fare for feilkonklusjoner, der innsikt i betingede sannsynligheter kan klargjøre situasjonen.

Eksempel 19 : Representasjonsfellen
Betrakt følgende utsagn

> "Folkegruppe B utgjør 10 % av befolkningen, men utgjør 15% av dem som har et bestemt gode".

Utsagnet inviterer leseren til å konkludere at folkegruppe B er favorisert, noe som kan være forhastet. Fellen kan lett blottlegges ved betraktninger rundt betingede sannsynligheter.

La A være begivenheten at en tilfeldig person i befolkningen har godet. Det vil være relevant å sammenligne $P(A \mid B)$ for ulike folkegrupper, men utsagnet ovenfor sammenligner i stedet $P(B \mid A)$ med $P(B)$. Dersom $P(B \mid A)$ er større (mindre) enn $P(B)$, kan vi ikke slutte mer enn at $P(A \mid B)$ er større (mindre) enn den tilsvarende sannsynlighet for minst en annen gruppe B.

La oss ta et talleksempel begrenset til tre grupper B_1, B_2 og B_3, og la sannsynlighetene for at en tilfeldig person tilhører hver av gruppene være

$$P(B_1) = 0.40 \quad P(B_2) = 0.10 \quad P(B_3) = 0.50$$

Anta videre at betingede sannsynligheter $P(A \mid B_i)$ er gitt ved to situasjoner i tabellen nedenfor, og at vi fokuserer på B_2. Vi beregner derfor $P(B_2 \mid A)$ ved Bayes lov (sjekk beregningen).

	$P(A \mid B_1)$	$P(A \mid B_2)$	$P(A \mid B_3)$	$P(B_2 \mid A)$
Situasjon 1	0.1	0.5	0.6	0.14
Situasjon 2	0.1	0.2	0.6	0.06

Vi ser at i situasjon 1 kommer gruppe 2 godt ut ved sammenligningen $0.14 > 0.10$, mens det i virkeligheten fins en langt større gruppe (B_3) som er bedre stillet, men også en gruppe (B_1) som er dårligere stillet. Kommenter selv situasjon 2.

Utsagn som er beslektet med det som er gitt i eksemplet forekommer ofte i media, oftest ubevisst, men er også brukt bevisst i meningspåvirkning. Eksempler på situasjoner: diskriminering mellom folkegrupper, sykdomsforekomst i ulike sosialgrupper, uhellsforekomst i ulike grupper (f.eks. av bilister), reklame og politisk kampanjevirksomhet og forbrukerspørsmål, f.eks. produktansvar.

Eksempel 20 : Diskriminering?
Betrakt følgende medieoppslag :

> "Ved høstens opptak tok UCB-universitetet opp 44 % av de mannlige søkere og 35 % av de kvinnelige."

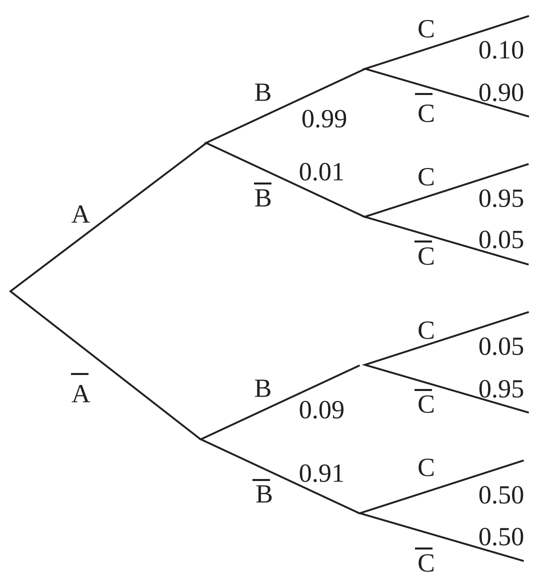
Dette kunne skyldes at andelen kvinner med dårlig opptaksgrunnlag var større enn blant de mannlige. Dersom det kan godtgjøres at dette var likelig fordelt, er det vel grunnlag for å hevde at diskriminering har funnet sted? Ikke ubetinget! Anta at opptak skjer ved søknad til de enkelte institutter. I et forsøk på å finne ut hvilke som eventuelt diskriminerer, viste det seg at ingen gjorde det. Andelen opptatte kvinner var omtrent lik mennenes ved de fleste, og ved et par institutter var andelen opptatte kvinner faktisk større. Forklaringen på dette tilsynelatende paradoks var at kvinnene hadde i større utstrekning søkt til institutter der det var vanskelig å komme inn.
Den beskrevne situasjon er et eksempel på det såkalte *Simpson's paradoks*. En formulering i sannsynlighetsteoretiske termer kan bidra til å klargjøre dette.

Simpson's paradoks :
La $\Omega = B_1 \cup B_2 \cup \ldots \cup B_r$ være en disjunkt oppsplitting av den sikre begivenhet. Anta at

$$P(C \mid A \cap B_i) \geq P(C \mid \overline{A} \cap B_i) \quad \text{for alle i}$$

Det er likevel mulig atP$(C \mid A) < P(C \mid \overline{A})$.

La oss i Figur 4.4 illustrere at dette er mulig i den enkleste situasjonen der $\Omega = B \cup \overline{B}$.



Figur 4.4: Simpsons paradoks

Her er betingelsen oppfylt, mens

$$\begin{aligned} P(C \mid A) &= P(B \mid A) \cdot P(C \mid A \cap B) + P(\overline{B} \mid A) \cdot P(C \mid A \cap \overline{B}) \\ &= 0.99 \cdot 0.10 + 0.01 \cdot 0.95 = 0.11 \\ P(C \mid \overline{A}) &= P(B \mid \overline{A}) \cdot P(C \mid \overline{A} \cap B) + P(\overline{B} \mid \overline{A}) \cdot P(C \mid \overline{A} \cap \overline{B}) \\ &= 0.09 \cdot 0.05 + 0.91 \cdot 0.50 = 0.46 \end{aligned}$$

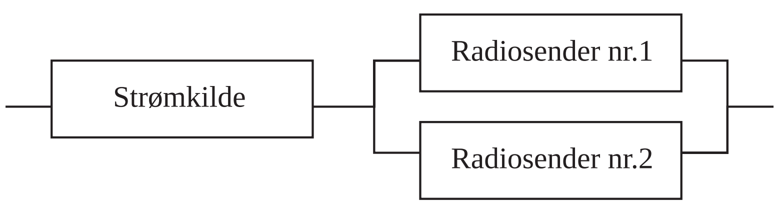
For å knytte forbindelsen til eksemplet over, la C=opptatt, B_i=søker til institutt nr.i og A=kvinnelig søker.

Simpson's paradoks opptrer i mange forkledninger, og det gjelder å være på vakt! Greier du denne : Er det mulig at totalbeskatningen er øket fra et år til et annet, mens beskatningen i alle inntektsgrupper har gått ned?

4.8 ⋆Pålitelighetsanalyse

Sannsynlighetsregning har de senere år i stigende grad blitt tatt i bruk ved studier av systempålitelighet. Vi gir her tre enkle eksempler, hver med sin egen form for systembeskrivelse, henholdsvis *logisk diagram*, *begivenhetstre* og *feiltre*.

Eksempel 21 : Systempålitelighet : Logisk diagram
Betrakt følgende enkle situasjon, der vi har en strømkilde og to radiosendere. For å sende signal må strømkilden og minst en av senderne virke. Dette er illustrert i diagrammet i Figur 4.5 med sannsynligheter påført hver enhet for at den virker.



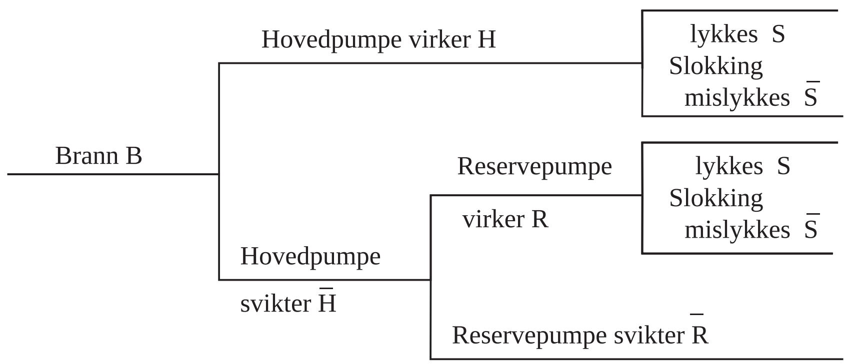
Figur 4.5: Logisk diagram

Anta at enhetene virker/ikke virker uavhengig av hverandre. Da blir

$$\begin{aligned} P(S \cap (B_1 \cup B_2)) &= P(S) \cdot P(B_1 \cup B_2) \\ &= P(S) \cdot (P(B_1) + P(B_2) - P(B_1 \cap B_2) \\ &= P(S) \cdot (P(B_1) + P(B_2) - P(B_1) \cdot P(B_2)) \\ &= 0.999 \cdot (0.99 + 0.99 - 0.99 \cdot 0.99) \\ &= 0.999 \cdot 0.9999 = 0.9989 \end{aligned}$$

Eksempel 22 : Systempålitelighet : Begivenhetstre

En begivenhet B som representerer et uhell, f.eks. brann, utløser aktiviteter for å bringe situasjonen under kontroll, f.eks. bruk av brannslokningsutstyr. Anta at situasjonen er som beskrevet i begivenhetstreet i Figur 4.6.



Figur 4.6: Begivenhetstre

Sannsynligheten for at brann oppstår som ikke slokkes (A) blir da

$$\begin{aligned} P(A) &= P(B) \cdot P(H \mid B) \cdot P(\overline{S} \mid B \cap H) \\ &+ P(B) \cdot P(\overline{H} \mid B) \cdot P(R \mid B \cap \overline{H}) \cdot P(\overline{S} \mid B \cap \overline{H} \cap R) \\ &+ P(B) \cdot P(\overline{H} \mid B) \cdot P(\overline{R} \mid B \cap \overline{H}) \end{aligned}$$

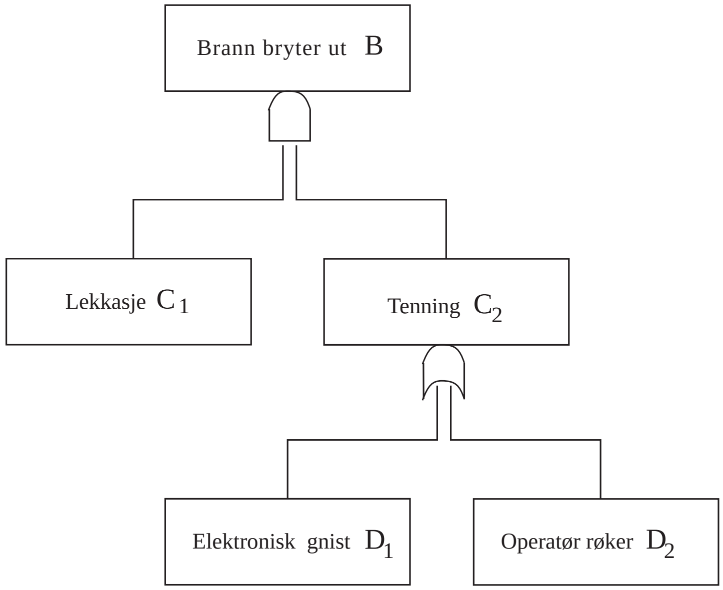
Kompliserte systemer vil omfatte mange muligheter for feil eller uhell, og hver kan tenkes analysert på tilsvarende måte, ofte med forenklende antakelser om uavhengighet. Sannsynligheten for feil/uhell av et eller annet slag i tidsperioden beregnes oftest ved å summere enkeltsannsynlighetene, fordi dette i hvert fall ikke undervurderer totalsannsynlighetene (hvorfor?). En må imidlertid alltid stille spørsmålene

- Er alle tenkelige uhellstyper tatt med?
- Er enkeltsannsynlighetene realistiske?
- Er evt. forenklende antakelser om uavhengighet realistiske?

Eksempel 23 : Systempålitelighet : Feiltre
Ofte studeres de betingelser som må være tilstede for at uhell skjer, noe som ofte kan beskrives ved et feiltre som beskrevet i Figur 4.7.

Legg merke til at dette er tilbakeskuende, mens begivenhetstreet ser forover. Vi ser at

$$\begin{aligned} P(B) &= P(C_1) \cdot P(C_2) \\ &= P(C_1) \cdot P(D_1 \cup D_2) \\ &\approx P(C_1) \cdot (P(D_1) + P(D_2)) \end{aligned}$$



Figur 4.7: Feiltre

4.9 Oppgaver

1. I Oppgave 2.10-11 finn følgende betingede modeller (i) gitt A (ii) gitt B (iii) gitt C. Finn deretter $P(A \mid B)$, $P(A \mid C)$, $P(B \mid A)$, $P(C \mid A)$, $P(A \mid \overline{B})$, $P(A \mid \overline{C})$ og $P(C \mid \overline{A})$. Finn også $P(A \mid A \cup B)$, $P(A \mid A \cup C)$, $P(A \mid A \cap B)$, $P(A \mid A \cap C)$, $P(A \mid A \cup \overline{C})$, $P(A \mid A \cap \overline{C})$, $P(A \mid \overline{A} \cap C)$ og $P(A \mid \overline{A} \cup B \cup C)$. Kommenter.

2. I Oppgave 2.12 finn sannsynlighetene $P(A \mid B)$, $P(B \mid A)$, $P(C \mid A)$ og $P(A \mid C)$. Kommenter.

3. I Oppgave 2.18 finn $P(B \mid C), P(C \mid B), P(B \mid D), P(D \mid B)$.

4. Regn ut noen betingede sannsynligheter etter eget valg for situasjonene i Oppgave 3.16 og 3.17.

5. Anta at $P(A) = 0.3, P(B) = 0.2$ og $P(A \cup B) = 0.4$. Finn $P(A \mid B)$ og $P(B \mid A)$.

6. I Oppgave 3.25 finn den betingede sannsynligheten for minst en førsteklassing gitt minst en gutt på trinn 3 eller lavere.

7. For situasjonen i Oppgave 3.27 finn de betingede sannsynligheter for
(a) 4 honnører gitt bare røde kort,
(b) bare røde kort gitt 4 honnører,
(c) ett ess gitt bare røde kort,
(d) ett ess gitt 4 honnører,
(e) 4 honnører gitt et ess.
Er noen av begivenhetene som omtales i Oppgave 3.26 uavhengige ?

8. Vis at dersom $A \subset B$ så er

$$P(A \mid B) = \frac{P(A)}{P(B)}$$

9. Vis at dersom A og B er disjunkte så er

$$P(A \mid A \cup B) = \frac{P(A)}{P(A) + P(B)}$$

10. Vis at $P(A \cap B \cap C) = P(A) \cdot P(B \mid A) \cdot P(C \mid A \cap B)$
Hint: Bruk E9 to ganger.

11. Et parti på 26 artikler blir undersøkt ved en inspeksjonsplan som foregår i to trinn: Først trekkes tilfeldig 8 artikler. Dersom en eller flere er defekte forkastes hele partiet. Dersom alle er i orden trekkes 8 nye artikler tilfeldig fra de resterende i partiet og partiet forkastes dersom en eller flere av disse er defekte. Finn sannsynligheten for å akseptere et parti med 3 defekte.

12. I spillet Russisk Rulett velger spilleren tilfeldig en av tre revolvere, hver med 5 kamre. Antall tomme kamre er henholdsvis 4, 3 og 2. Kamrene snurres rundt og revolveren avfyres mot tinningen. Hva er sannsynligheten for at spilleren overlever spillet? Gitt at han overlevet, hva er sannsynligheten for at skuddet ble avfyrt med den revolver med flest tomme kamre?

13. Sannsynligheten for at en gift mann har sett et bestemt teaterstykke er 0.4, mens sannsynligheten for at en gift kvinne har sett stykket er 0.5. Sannsynligheten for at konen har sett stykket gitt at mannen har sett det er 0.9. For et vilkårlig valgt ektepar finn sannsynligheten for at
(a) begge ektefeller har sett stykket,
(b) mannen har sett stykket gitt at konen har sett det,
(c) minst en av ektefellene har sett stykket.

14. En bedrift har to maskiner A og B som kan brukes til å produsere samme artikkel. A står for 80% av produksjonen, mens B står for 20%. Man har erfart at A leverer gjennomgående 5% defekte, mens B leverer 1% defekte. Hva er sannsynligheten for at den skriver seg fra maskin A?

15. Anta at sannsynligheten for at Per vil leve om 20 år er 0.6, mens sannsynligheten for Pål er 0.9. Finn sannsynligheten for
(a) begge lever,

(b) ingen lever,
(c) minst en lever.

16. Det foreligger to urner, hver med 6 kuler. Urne nr. 1 har 4 hvite og 2 sorte kuler, mens Urne nr. 2 har 3 hvite og 3 sorte kuler. En urne velges tilfeldig og to kuler trekkes tilfeldig fra denne, begge var hvite. Hva er sannsynligheten for at kulene kommer fra Urne nr. 1 når
(a) trekningen skjer med tilbakelegging,
(b) uten tilbakelegging.

17. I en "multiple choice" test er det n svaralternativer for hvert spørsmål. Anta at sannsynligheten for at Per vet svaret på et bestemt spørsmål er p. Dersom han ikke vet svaret, antar vi at han krysser av tilfeldig. Hva er sannsynligheten for at Per visste svaret gitt at han krysset av riktig?

18. La situasjonen være som beskrevet i Eksempel 10. Finn sannsynlighetene for
(a) veien stengt gitt for sent fremme,
(b) uhell underveis gitt for sent fremme,
(c) veien åpen gitt tidsnok fremme og intet uhell underveis.

19. Lag et sannsynlighetstre i hver av følgende situasjoner
(a) Eksempel 5, (b) Eksempel 6, (c) Eksempel 9.

20. Per knipser to rettferdige mynter. Gitt at minst en viste kron. Hva er sannsynligheten for at begge viste kron? Løs oppgaven under forutsetningene
(a) Per ser resultatet av begge myntene, men gir deg ufullstendig informasjon.
(b) Den ene mynten trillet tilfeldigvis under sofaen.
Blir svaret det samme for (a) og (b)?

21. Tre fanger A, B og C blir informert om at en av dem er valgt ut tilfeldig for å bli henrettet, mens de to andre blir satt fri. Fange A spør vokteren om han ikke i hemmelighet kan fortelle hvem av de to medfangene som skal settes fri, siden han allerede vet at en av disse skal settes fri. Vokteren nekter å svare fordi han mener at når A vet hvem av kameratene som blir satt fri, vil hans egen sjanse for å bli henrettet øke fra 1/3 til 1/2 fordi A nå er en blant to fanger hvorav en skal henrettes. Hva mener du om vokterens argumentasjon?

22. En person er anholdt som mistenkt for en bestemt forbrytelse. Man venter på rapport fra kriminalteknisk laboratorium angående blodspor på åstedet. På grunnlag av det foreliggende bevismateriale vurderer inspektør Snusen at (apriori) sannsynligheten for at anholdte er skyldig er lik p. Beskjed kommer så om at anholdtes blodtype samsvarer med blodsporene på åstedet som stammer fra gjerningsmannen. Anta at andelen av befolkning som har denne blodtypen er lik r. Finn (aposteriori) sannsynligheten for at anholdte er skyldig uttrykt ved p og r.

23. En rettferdig mynt knipses to ganger. La
A = første kast viser kron,

B = annet kast viser kron,
C = kastene viser samme resultat.
Er A og C uavhengige begivenheter? Enn B og C? Er A, B og C uavhengige?

24. Vis at dersom A og B er uavhengige, så er også A og $\overline{B}$ uavhengige.

25. En arbeider skal opplæres til å utføre en vanskelig arbeidsoperasjon. Man har erfart at 10% av nye arbeidere greier å utføre operasjonen tilfredsstillende i første forsøk, 20% i andre forsøk og 30% i tredje forsøk. Anta uavhengighet mellom forsøkene og finn sannsynligheten for en ny arbeider i løpet av 3 forsøk vil lykkes
(a) alle tre ganger, (b) ingen ganger,
(c) minst en gang, (d) akkurat en gang.

26. La situasjonen være som i forrige oppgave, men anta at 25% av de nye arbeiderne har lært arbeidsoperasjonen i sitt tidligere arbeid, for disse er det 80% sjanse for at de lykkes i hvert av de tre forsøkene. Finn sannsynlighetene for begivenhetene i (a), (b), (c) og (d) for en ny arbeider når vi ikke vet noe om hans tidligere arbeidsbakgrunn.

27. Per har fått 8 kroner for å kjøpe et brød, men er falt for fristelsen å kjøpe en tyggegummi til 1 krone. For å "redde" situasjonen har han funnet en venn som er villig til å knipse mynt og krone. Innsatsen er 1 krone i første omgang, taper Per, dobles innsatsen og slik fortsetter man inntil Per enten har vunnet tilbake den krona som gikk til tyggegummi, eller blitt blakk. Hva er sannsynligheten for at Per kommer hjem med brød?

28. Et spill mellom to spillere består av spilleomganger, den som først vinner fire omganger vinner spillet. Anta at sannsynligheten for at A vinner første omgang er 1/2, og at i derpå følgende omganger er oddsene 2:1 i favør av den som vant siste omgang. Finn sannsynligheten for at A vinner etter
(a) etter 4 omganger, (b) etter 5 omganger, (c) før eller senere.

29. En bedrift bruker en maskin som alltid justeres ved ukeslutt (5 dagers uke), men kan kreve ekstra justering i løpet av uken, som foretas etter endt arbeidsdag. Man har erfart at maskinen på en dag da det er henholdsvis 1, 2, 3, 4 dager siden den ble justert sist, vil trenge ny justering ved arbeidsdagens slutt med sannsynligheter henholdsvis 0.10, 0.12, 0.15, 0.20. Finn sannsynligheten for
(a) ingen ekstra justering i løpet av uken,
(b) første justering skjer på onsdag,
(c) justering må foretas på onsdag,
(d) justering ble foretatt på onsdag gitt at det var minst en ekstra justering.

30. Løs foregående oppgave dersom vi isteden antar at sannsynligheten er 0.10 for at maskinen krever justering etter endt arbeidsdag uavhengig av hvor lenge det er siden den justert sist.

31. ⋆I en bedrift er det 3 avdelinger, i avd. 1 er det 6 kvinner, 3 menn, i avd. 2 er det 4 kvinner, 8 menn og i avd. 3 er det 3 kvinner og 9 menn. To av de ansatte skal sendes på et kurs og man ønsker ikke at begge skal sendes på et kurs og man ønsker ikke at begge skal komme fra samme avdeling. Derfor velges først tilfeldig hvilke to avdelinger som skal være representert, deretter velges en representant tilfeldig fra hver av de utvalgte avdelingene. Finn sannsynligheten for at
(a) begge kursdeltakere er kvinner,
(b) en kvinne og en mann reiser på kurs.
Påvis at trekningsmåten ikke er rettferdig i den forstand at ikke alle ansatte har samme sjanse for å komme med. Det blir foreslått at man isteden bør trekke en person tilfeldig blant alle ansatte, og deretter trekke en tilfeldig blant de ansatte i de to avdelinger som til nå ikke er representert. Er dette en rettferdig trekningsmåte? Hvis ikke, anvis en slik (hvor fortsatt hver avdeling bare skal være representert med en person).

32. Kursen på en aksje observeres på etterfølgende hverdager og det observeres enten oppgang (+), status quo (0) eller nedgang (−). Anta at sannsynligheten for 0 er 0.6 mens sannsynligheten for + og − er begge 0.2. Anta videre at kursendringen på en bestemt hverdag er uavhengig av tidligere kursendringer (diskuter hvorvidt dette er realistisk). Finn sannsynligheten for at det i løpet av 10 dager
(a) er status quo de fem første dager,
(b) er akkurat fem dager med status quo,
(c) ikke er kursfall,
(d) er fem kursendringer.

33. Anta at endringer i en aksjeindeks er observert sammen med renteendringer på 1000 tidspunkter med følgende resultat:

	Kursnedgang	Kursoppgang	Sum
Rentenedgang	60	440	500
Renteoppgang	420	80	500
Sum	480	520	1000

Finn ulike betingede hyppigheter som belyser sammenhengen mellom renteendring og kursendring i perioden. Kan disse tolkes som sannsynligheter?

34. En forsikringsagent vurderer sjansen for at en kontrakt inngås ved første besøk hos kunden til 10%, som øker til 40% ved andre besøk. Agenten har besøkt samme 3 personer på to etterfølgende dager. Hva er sannsynligheten for at ingen kontrakt ble inngått?

35. Et foretak regner at sjansen for å få en bestemt kontrakt er 45% dersom hovedkonkurrenten ikke byr på samme kontrakt, men bare 25% dersom denne byr. Anta at sjansen for at konkurrenten byr er 40%. Hva da sjansen for å få kontrakten?

36. Et produkt vurderes av produktutviklere til å ha 60% sjanse for å gi positivt dekningsbidrag, mens markedførere tror at sjansen bare er 40%. Før beslutningen om lansering blir produktet testet av et forbrukerpanel. Anta at dersom testen er positiv/negativ er det henholdsvis 75% og 25% sjanse for fortjeneste. Bedriften vil satse dersom det er mer enn 50% sjanse for positivt dekningsbidrag. Hva er din konklusjon dersom testen ble positiv evt. negativ?

37. En bedrift vil lansere et nyutviklet produkt i januar neste år, og regner med en sjanse på 60% for at produktet skal bli en suksess, såfremt konkurrenten ikke kommer før med et tilsvarende produkt. I så fall er sjansen redusert til 30%. Hvor stor må sjansen for at konkurrenten lanserer sitt produkt først være for at sjansen for egen suksess går under 50%?

38. To terninger trilles 5 ganger i alt. Hva er sannsynligheten for
(a) minst en sekser i alle 5 omganger,
(b) akkurat tre av omgangene ga minst en sekser,
(c) minst tre av omgangene ga minst en sekser.

39. En tipper regner med at sannsynligheten er 0.4 for at han tipper en enkelt kamp rett (altså noe bedre enn ren gjetting). Finn sannsynligheten for at han med en enkeltrekke på 12 kamper tipper
(a) akkurat 4 rette,
(b) høyst 4 rette,
(c) minst 10 rette.
Løs denne oppgaven også under forutsetningen av at han mener at sjansen for korrekt tips er 0.6.

40. En "multiple choice" test består av 10 spørsmål hvert med 4 svaralternativer. Hva er sannsynligheten for at en student får minst 5 rette svar dersom hun gjetter på alle 10 spørsmålene?

41. Avgjør om følgende situasjon med rimelighet kan sies å være binomiske forsøksrekker.
(a) En falsk mynt knipses gjentatte ganger og kron eller mynt observeres.
(b) En terning trilles gjentatte ganger og antall øyne noteres.
(c) To terninger trilles gjentatte ganger og det observeres om sum øyne er mer eller mindre enn syv.
(d) Kjønn til etterfølgnde barn født på en fødestue observeres.
(e) Regn eller ikke på etterfølgende dager.
(f) Regn eller ikke på etterfølgende søndager.
(g) På etterfølgende hverdager observeres om avisgutten kommer før eller etter kl. 16.00 med kveldsavisen.
(h) På suksessive dager observeres om morgenflyet fra Oslo til Bergen er fullt eller ikke.
(i) På suksessive mandager observeres om morgenflyet fra Oslo til Bergen er fullt eller ikke.
(j) Det observeres om en buss kommer i rute, eller etter ruten på suksessive stoppesteder.

(k) En rekke personer skal avgjøre hvilket av to produkter de liker best.
(l) En pakke med 20 artikler inneholder 5 defekte. Artikler trekkes ut tilfeldig en etter en og klassifiseres som defekt eller intakt.
(m) Som (l) men anta at artikkelen legges tilbake før ny trekning foretas.
(n) For en rekke forsøkspersoner observeres om et medikament har positiv, ingen eller negativ effekt.
(o) I en innsjø er det satt ut en del fisk som er merket. For hver fisk som tas opp observeres om fisken er merket eller ikke.

42. 10 studenter skal delta i en øl-test for å rangere lettøl fra 5 av landets bryggerier som blir servert uten merkeidentifikasjoner som sort A, B, C, D og E. Anta at lettøl fra disse 5 bryggeriene smaker nøyaktig likt. Finn sannsynligheten for at
(a) Alle deltakere velger rekkefølgen ACDBE.
(b) Akkurat 3 av deltakerne velger ACDBE.
(c) Akkurat 5 av deltakerne mener E er bedre enn B.
(d) Ingen av bryggeriene får mer enn to stemmer på første plass.
(e) Minst to deltakere velger samme rekkefølge.

43. En terning trilles gjentatte ganger. Finn sannsynligheten for at
(a) første sekser kommer i sjette kast,
(b) første sekser kommer før sjette kast,
(c) første sekser kommer etter sjette kast,
(d) annen sekser kommer i sjette kast,
(e) annen sekser kommer før sjette kast,
(f) det trengs minst tre kast for at sum øyne er seks,
(g) sum øyne før første sekser er høyst lik seks.

44. $\star$Et system observeres på tidspunktene $t = 0, 1, 2, 3, \ldots$. Det kan være i en av to tilstander og starter i tilstand 1. I tilstand 1 er sannsynlighetene henholdsvis p og $1 - p$ for at systemet fortsatt er i denne tilstand/skifter tilstand ved neste tidspunkt. De tilsvarer sannsynligheter for tilstand 2 er henholdsvis q og $1 - q$.
(a) Finn et uttrykk for sannsynligheten for at systemet er i tilstand 1 etter 3 tidsenheter.
(b) Gitt at systemet er i tilstand 1 etter 3 tidsenheter, hva er sannsynligheten for at det har vært innom tilstand 2 underveis.
Et system som beskrevet ovenfor blir ofte kalt en *Markov kjede*.

Kapittel 5

Stokastiske variable

5.1 Innledende definisjoner

La utgangspunktet være et eksperiment med et gitt utfallsrom. [1] Formålet med observasjon behøver ikke nødvendigvis være å få vite akkurat hvilket av de mulige utfall som inntreffer. For mange problemstillinger er en mindre detaljrik beskrivelse av eksperimentresultatet tilstrekkelig. Ofte er det bekvemt med en tallmessig beskrivelse, dvs. vi er interessert i en tallstørrelse som er entydig gitt ved utfallet av eksperimentet, og som delvis karakteriserer dette, eksempelvis

- I to terningkast kan det hende at vi bare er interessert i summen av øynene, og ikke resultatet av hvert enkelt kast.
- I kvalitetskontroll er vi ofte interessert i antall defekte artikler i et utvalg, ikke hvilke artikler som var defekte eller ikke.
- I en binomisk forsøksrekke er vi som oftest interessert i antall suksesser, ikke hvilke forsøk som ga henholdsvis suksess og fiasko.

Ved å trekke ut slike tallstørrelser kan vi ofte bearbeide eksperimentresultatet på en mer hensiktsmessig måte.

[1] I dette kapitlet kommer en mengde nye begreper, som det ikke alltid er like lett å motivere, fruktene kommer senere. Det er ikke vanskelig, men det krever konsentrasjon: Lukk døra, slå av CD-spilleren og stump røyken!

Definisjon : Gitt et eksperiment med utfallsrom Ω.
En *stokastisk variabel* er en funksjon med utfallsrommet Ω som definisjonsmengde og verdimengde blant reelle tall R.

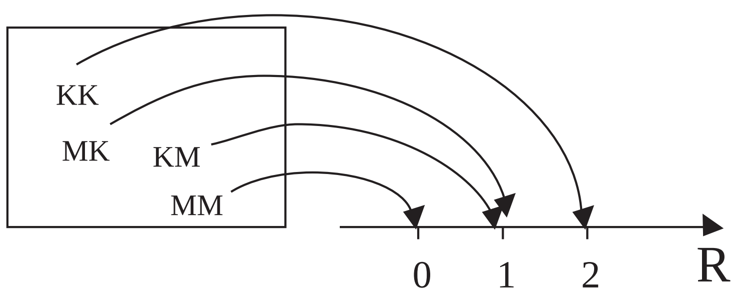
Stokastiske variable betegnes som oftest med store bokstaver fra slutten av alfabetet f.eks. X, Y, Z, U, V, W etc. [2]) La X være en stokastisk variabel (s.v.). X er da en forskrift som til ethvert utfall u i utfallsrommet tilordner et reellt tall $X(u)$:

$$X : u \to X(u)$$

Eksempel 1 : To myntkast
Anta at utfallsrommet er $\Omega = \{\text{MM, MK, KM, KK}\}$. La X være en stokastisk variabel definert ved

u	:	MM	MK	KM	KK
$X(u)$	:	0	1	1	2

Vi ser at X representerer antall kron i de to kastene. Den stokastiske variable X er illustrert i Figur 5.1.



Figur 5.1: En stokastisk variabel

Eksempel 2 : To terningkast
Anta at utfallsrommet er $\Omega = \{(1,1),(1,2),...,(6,6)\}$. La S være en stokastisk variabel definert ved

[2]I moderne matematikk er det ikke god tone å kalle en funksjon for en variabel, men det er uhensiktsmessig å endre en allerede innarbeidet notasjon på dette punkt.

$$S((1,1)) = 2,\ S((1,2)) = 3,\ \ldots,\ S((6,6)) = 12$$

dvs. til utfallet (i,j) tilordnes det reelle tallet $i+j$. Vi ser at S representerer summen av øynene i de to kastene.

Eksempel 3 : Tombola
I en tombola er det N lodd nummererte fra 1 til N. Et lodd trekkes tilfeldig, og trekkes lodd nr. i får man en gevinst av verdi v_i (som kan være positiv, null eller negativ). La utfallsrommet være $\Omega = \{1, 2, ..., N\}$, og la Y være den stokastiske variabel som til utfallet i tilordner verdien v_i. Vi ser at Y representerer verdien av det uttrukne lodd.

Eksempel 4 : Indikator
La A være en begivenhet i et utfallsrom Ω. Dersom X er en stokastisk variabel definert ved at

$$X(u) = \begin{cases} 1 & \text{for alle } u \in A \\ 0 & \text{for alle } u \in \overline{A} \end{cases}$$

sier vi at X er en *indikatorvariabel*, den indikerer om begivenheten A har inntruffet eller ikke.

Eksempel 5 : Ventetid til første kron
Anta at vi har valgt $\Omega = \{$K, MK, MMK, MMMK,$\ldots\}$ som utfallsrom. La N være en stokastisk variabel definert ved

$$\begin{array}{cccccccc} u & : & \text{K} & \text{MK} & \text{MMK} & \text{MMMK} & \ldots \\ N(u) & : & 1 & 2 & 3 & 4 & \ldots \end{array}$$

Da vil N representere ventetiden til første kron.

I anvendelser hvor utfallsrommet selv er en delmengde av de reelle tall, kan det komme på tale å bruke identitetsfunksjonen som stokastisk variabel, dvs. definere X ved

$$X(u) = u \ \text{ for alle } u \in \Omega$$

eksempelvis vil X = antall øyne i et terningkast kunne oppfattes som den stokastiske variable som til utfallet "i øyne" tilordner det reelle tall i.

5.2 Sannsynlighetsfordeling

La X være en stokastisk variabel definert i et utfallsrom Ω. Med symbolkombinasjonen $(X = x)$ vil vi forstå den begivenhet at den stokastiske variable X antar verdien x, eller presist

$$(X = x) = \{u; X(u) = x\}$$

dvs. gunstige utfall for begivenheten $(X = x)$ er de utfall u som er slik at $X(u) = x$. La oss betegne de mulige [3] verdiene av X med $x_1, x_2, ..., x_r$. Vi vil ofte være interessert i sannsynlighetene for de r begivenhetene

$$(X = x_1), (X = x_2), \ldots, (X = x_r)$$

I denne forbindelse innfører vi følgende begrep:

> **Definisjon :** *Sannsynlighetsfordelingen* til X er en oppregning av de mulige verdier av X sammen med sannsynligheten for at X antar disse verdiene.

Sannsynlighetsfordelingen kan gis i tabellform slik

x	x_1	x_2	$\cdots$	x_r
$P(X = x)$	$P(X = x_1)$	$P(X = x_2)$	$\cdots$	$P(X = x_r)$

eller ved en funksjon p definert for de mulige verdier av X slik at

$$p(x) = P(X = x)$$

Vi sier at p er *(punkt)fordelingsfunksjonen* til X.

Eksempel 6 : To myntkast
Dersom X = antall kron, får vi

$$(X = 0) = \{MM\}, \;\; (X = 1) = \{KM, MK\}, \;\; (X = 2) = \{KK\}$$

[3] For enkelhets skyld antar vi bare et endelig antall mulige verdier. Dette vil vi ofte gjøre i det følgende uten at dette er nødvendig, resultatene gjelder også med visse reservasjoner dersom vi har et tellbart uendelig antall mulige verdier.

I en modell hvor alle 4 utfall er like sannsynlige blir sannsynlighetsfordelingen til X

x	0	1	2
$P(X=x)$	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$

Eksempel 7 : To terningkast
Dersom $S =$ sum øyne i de to kastene får vi

$$(S=2)=\{(1,1)\},\ (S=3)=\{(1,2),(2,1)\},\ldots,(S=12)=\{(6,6)\}$$

I en modell hvor alle 36 mulige utfall er like sannsynlige blir sannsynlighetsfordelingen til S

s	2	3	4	5	6	7	8	9	10	11	12
$P(S=s)$	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

Vi merker oss at

$$\Omega = (X=x_1)\cup(X=x_2)\cup\cdots\cup(X=x_r)$$

er en disjunkt union. Følgelig blir

$$1 = P(\Omega) = P(X=x_1)+P(X=x_2)+\cdots+P(X=x_r)$$

Fordelingsfunksjonen $p(x)$ til X har derfor følgende egenskaper

1. $0 \le p(x) \le 1$ for alle x
2. $p(x_1)+p(x_2)+\cdots+p(x_r)=1$

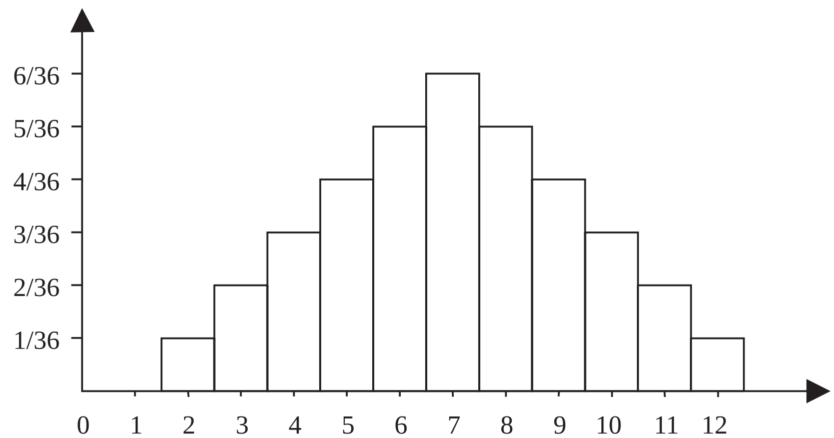
Vi ser at dette stemmer i Eksempel 6 og 7.

Merknad. Ut fra en gitt sannsynlighetsmodell definerer den stokastiske variable X en ny sannsynlighetsmodell med mengden av de mulige verdier av X som utfallsrom, dvs. et utfallsrom bestående av reelle tall

$$\Omega_X = \{x_1, x_2, ..., x_r\}$$

og sannsynlighetsfunksjon lik fordelingsfunksjonen $p(x)$ til X.

Dersom man ønsker å billedliggjøre sannsynlighetsfordelingen til X tegner man ofte et såkalt *histogram*. Dette er en figur bestående av søyler som hver er sentrert i en mulig verdi av X med areal som er proporsjonalt med



Figur 5.2: Et histogram

sannsynligheten for at X antar denne verdien. I eksemplet med to terningkast vil histogrammet til sum øyne S se ut som i Figur 5.2.

I mange situasjoner der man er interessert i en stokastisk variabel X, er sannsynlighetsfordelingen $p(x)$ helt eller delvis ukjent. For å skaffe informasjon om $p(x)$ kan det være aktuelt å utføre det eksperimentet som modellen er ment å beskrive gjentatte ganger, og observere verdien av X hver gang. Vi teller deretter opp antall ganger vi observerte de ulike mulige verdier av X, og tegner en figur med søyler som hver er sentrert i en mulig verdi av X, med areal proporsjonalt med antall ganger X har antatt denne verdi. En slik figur kalles et *empirisk histogram*. Et empirisk histogram gir oss informasjon om formen på den sannsynlighetsfordelingen $p(x)$ (dvs. om histogrammet) som svarer til en noenlunde realistisk modell for eksperimentet. Påliteligheten av denne informasjon vil selvsagt avhenge av hvor mange ganger vi gjentar eksperimentet, jfr. tanken om sannsynlighet som idealisert hyppighet i det lange løp (se Oppgave 42 og Eksempel 1.6).

La X være stokastisk variabel. Det er da naturlig å innføre følgende skrivemåter

$$(X \le x) = \{u; X(u) \le x\}, \quad (X > x) = \{u; X(u) > x\} \quad \text{etc.}$$

dvs. $(X \le x)$ betyr begivenheten at X antar en verdi mindre enn eller lik x, mens $(X > x)$ betyr begivenheten at X antar en verdi større enn x etc. Anta at sannsynlighetsfordelingen til X er gitt ved fordelingsfunksjonen $p(x) = P(X = x)$. Sannsynligheten for begivenheten $(X \le x)$ blir lik

$$F(x) = P(X \leq x) = \sum_{k \leq x} p(k)$$

dvs. vi summerer sannsynlighetene $P(X = k)$ for alle k som er $\leq x$. Funksjonen $F(x) = P(X \leq x)$ definert for alle reelle x kalles den *kumulative fordelingsfunksjonen* til X. En kumulativ fordelingsfunksjon $F(x)$ har generelt egenskapene (i) $F(x)$ ikke avtagende (ii) $0 \leq F(x) \leq 1$ (iii) $F(-\infty) = 0$, $F(\infty) = 1$. For diskrete modeller som vi studerer her, vil den være en trappefunksjon med sprang i hver av de mulige verdier av X og trinnhøyde lik sannsynligheten for at X antar verdien, se Oppgave 38.

Sannsynlighetsfordelingen til X kan alternativt beskrives ved å angi den kumulative fordeelingsfunksjon istedenfor punktfordelingsfunksjonen. Kumulative sannsynligheter er også hensiktsmessige talloppslag og ved beregninger foretatt av programvare, idet vi ofte er interessert i om variablen X er mindre enn, større enn, eller har verdi i et intervall. Da er det greitt å vite at $P(X > x) = 1 - F(x)$ og $P(a < X \leq b) = F(b) - F(a)$.

5.3 Forventning

Gitt et eksperiment med sannsynlighetsmodell

$$\begin{array}{ccccc} u_1 & u_2 & u_3 & \cdots & u_m \\ P(u_1) & P(u_2) & P(u_3) & \cdots & P(u_m) \end{array}$$

La X være en stokastisk variabel.

Definisjon : Med *forventningen* til X menes

$$EX = X(u_1)P(u_1) + X(u_2)P(u_2) + \cdots + X(u_m)P(u_m)$$

eller mer kompakt skrevet [a])

$$\text{F1.} \qquad EX = \sum_u X(u)P(u)$$

der summasjonen går over alle utfallene i utfallsrommet.

[a]Den engelske betegnelsen er "expectation", derav symbolet EX.

Vi ser at forventningen er definert som en veid sum hvor verdien av X for de mulige utfall veies med de respektive sannsynlighetene for disse utfallene. At begrepet forventning har en sentral plass i sannsynlighetsteori og statistikk vil bli klart etter hvert. La oss som en første motivasjon vise at forventning kan tolkes som "et idealisert gjennomsnitt i det lange løp":

Anta at vi har gjentatt vårt eksperiment i alt n ganger, og at vi har observert utfallene $u_1, u_2, \ldots, u_m$ henholdsvis $n(u_1), n(u_2), \ldots, n(u_m)$ ganger. Gjennomsnittlig X-verdi i løpet av de n forsøk blir

$$\frac{1}{n}(X(u_1)n(u_1) + X(u_2)n(u_2) + \cdots + X(u_m)n(u_m))$$

som ved å innføre de relative hyppighetene $h(u) = n(u)/n$ kan skrives

$$X(u_1)h(u_1) + X(u_2)h(u_2) + \cdots + X(u_m)h(u_m)$$

I Kapittel 2.2 motiverte vi sannsynlighetsbegrepet ved at en sannsynlighet $P(u)$ bl.a. kunne svare til "en idealisert relativ hyppighet i løpet av et stort antall gjentakelser av eksperimentet". Dersom modellen ovenfor er realistisk vil, såfremt n er tilstrekkelig stor,

$$h(u_1) \approx P(u_1),\ h(u_2) \approx P(u_2), \ldots,\ \ h(u_m) \approx P(u_m)$$

Dersom vi i uttrykket ovenfor erstatter alle relative hyppigheter $h(u)$ med den tilsvarende "ideale" sannsynlighet $P(u)$ får vi

$$X(u_1)P(u_1) + X(u_2)P(u_2) + \cdots + X(u_m)P(u_m)$$

som nettopp er definisjonen på forventningen til X.

Merk at forventningen til X er et modellteoretisk begrep. EX kan beregnes straks vi har en modell for eksperimentet, før det eventuelt utføres. Selv om begrepet her er motivert ved forestillingen om et "gjennomsnitt i det lange løp", tyder erfaring på at en forventningsverdi også har relevans for folks holdning til et enkelt eksperiment.

Eksempel 8 : Et terningkast
La $X =$ antall øyne. Med den vanlige modellen får vi

$$EX = 1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + 3 \cdot \frac{1}{6} + 4 \cdot \frac{1}{6} + 5 \cdot \frac{1}{6} + 6 \cdot \frac{1}{6} = \frac{7}{2}$$

Merk at en forventningsverdi ikke behøver å være en mulig verdi, ordet forventning må altså ikke tas altfor bokstavelig. Anta at det som gevinst utbetales like mange kroner som terningen viser øyne. X kan da tolkes som gevinsten i kroner. Forventet gevinst blir altså kr 3.50. Dette kan tolkes dithen at ved gjentatte terningkast blir gjennomsnittlig gevinst pr. kast i det lange løp kr 3.50. Forventningsverdien har imidlertid relevans for et enkelt kast, de fleste er vel enig i at spilleren bør avkreves en innsats på kr 3.50 for at spillet skal være "rettferdig" (både for spiller og banken).

Eksempel 9 : Tombola
En tombola inneholder N nummererte lodder med verdier som vi betegner henholdsvis $v_1, v_2, \ldots, v_N$ (se Eksempel 3). Vi trekker et lodd tilfeldig og lar Y = verdien av det uttrukne lodd. Vi får

$$EY = \sum_{i=1}^{N} Y(i)P(i) = \sum_{i=1}^{N} v_i \cdot \frac{1}{N} = \frac{1}{N}\sum_{i=1}^{N} v_i = \overline{v}$$

dvs. forventet verdi er lik den gjennomsnittlige verdi av loddene i tombolaen, her betegnet med $\overline{v}$.

La X være en stokastisk variabel med sannsynlighetsfordeling gitt ved $p(x) = P(X = x)$. Dersom EX skal beregnes, er det nok å kjenne sannsynlighetsfordelingen til X, vi har nemlig følgende formel [4]

F2. $$EX = \sum_x xP(X = x) = \sum_x xp(x)$$

hvor summasjonen går over de mulige verdier av X.

Denne formelen er en direkte følge av definisjonen F1: For hver mulig verdi x av X grupperer vi sammen leddene for de utfall u som er slik at $X(u) = x$. I disse ledd er x en felles faktor, og summen av sannsynlighetene for de utfallene som gir verdien x er $P(X = x)$.

Eksempel 10 : To myntkast
Forventet antall kron EX kan beregnes ut fra definisjonen (se Eksempel 1)

$$\begin{aligned} EX &= X(MM)P(MM) + X(MK)P(MK) \\ &+ X(KM)P(KM) + X(KK)P(KK) \\ &= 0 \cdot \frac{1}{4} + 1 \cdot \frac{1}{4} + 1 \cdot \frac{1}{4} + 2 \cdot \frac{1}{4} = 1 \end{aligned}$$

[4]Vi skal vise en rekke setninger om forventning og vil betegne dem F1, F2, F3 etc.

eller bruk av sannsynlighetsfordelingen til X (se Eksempel 6) og formel F2.

$$\begin{aligned} EX &= 0 \cdot P(X=0) + 1 \cdot P(X=1) + 2 \cdot P(X=2) \\ &= 0 \cdot \frac{1}{4} + 1 \cdot \frac{1}{2} + 2 \cdot \frac{1}{4} = 1 \end{aligned}$$

Eksempel 11 : To terningkast
La situasjonen være som i Eksempel 7. Vi får

$$\begin{aligned} ES &= \sum_s sP(S=s) \\ &= 2 \cdot \frac{1}{36} + 3 \cdot \frac{2}{36} + 4 \cdot \frac{3}{36} + \cdots + 11 \cdot \frac{2}{36} + 12 \cdot \frac{1}{36} = 7 \end{aligned}$$

Her ville en beregning direkte ut fra definisjonen bli svært tungvint.

Eksempel 12 : Ventetid til første kron
Antall kast N inntil første kron i myntkast kan oppfattes som en stokastisk variabel med sannsynlighetsfordeling gitt ved funksjonen (se Eksempel 5 og Kapittel 4.6)

$$p(n) = P(N=n) = (\frac{1}{2})^n \quad n = 1, 2, 3, \ldots$$

Ifølge summeformelen for en geometrisk rekke får vi

$$\sum_{n=1}^{\infty} (\frac{1}{2})^n = \frac{\frac{1}{2}}{1-\frac{1}{2}} = 1$$

slik at dette virkelig er en sannsynlighetsfordeling. Her blir

$$EN = \sum_{n=1}^{\infty} n \cdot p(n) = \sum_{n=1}^{\infty} n \cdot (\frac{1}{2})^n = 2$$

At denne summen er 2 følger av kjente regneformler i matematisk analyse. For øvrig er dette et spesialtilfelle av et mer generelt resultat som blir vist i Kapittel 6.7.

La oss til slutt i dette avsnittet gjøre noen betraktninger om såkalte rettferdige spill:

> Et spill der forventet gevinst er null kalles et *rettferdig* spill.

Eksempelvis vil myntkast hvor man vinner 10 kroner dersom mynten viser kron, men taper det samme beløp dersom den viser mynt være et rettferdig spill, fordi forventet gevinst her blir

$$EG = 10 \cdot \frac{1}{2} + (-10) \cdot \frac{1}{2} = 0$$

Eksempel 13 : Gambling
I et spill vinner spilleren et beløp lik innsatsen dersom en mynt viser kron, men taper innsatsen dersom den viser mynt. Spilleren har bestemt seg for å spille inntil han får kron, dog ikke mer enn tre omganger. Hva er forventet gevinst ved å spille dette spillet dersom

(1) Innsatsen er kr.10 i hver omgang

(2) Innsatsen er kr.10 i første omgang, men dobles for hver ny omgang i et forsøk på å ta igjen det tapte.

Antar vi rettferdig mynt og uavhengig kast får vi følgende modell

Utfall:	K	MK	MMK	MMM
Sannsynlighet:	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{8}$
Gevinst G_1 :	10	0	-10	-30
Gevinst G_2 :	10	10	10	-70

Forventet gevinst i de to situasjonene blir:

$$\begin{aligned} EG_1 &= 10 \cdot \frac{1}{2} + 0 \cdot \frac{1}{4} + (-10) \cdot \frac{1}{8} + (-30) \cdot \frac{1}{8} = 0 \\ EG_2 &= 10 \cdot \frac{1}{2} + 10 \cdot \frac{1}{4} + 10 \cdot \frac{1}{8} + (-70) \cdot \frac{1}{8} = 0 \end{aligned}$$

I dette spillet utgjør hver spilleomgang et rettferdig spill, og de funne resultater synes å indikere at et spill med rettferdige spilleomganger ikke kan vendes til spillerens fordel ved å spille spesielle systemer, spillet er og blir rettferdig.

5.4 Setninger om forventning

La i det følgende k betegne en konstant og X en stokastisk variabel. Vi kan da definere nye stokastiske variable, f. eks. slik:

$$\begin{array}{lll} Z = k & \text{dvs. } Z(u) = k & \text{for alle utfall } u \\ Z = k + X & \text{dvs. } Z(u) = k + X(u) & \text{for alle utfall } u \\ Z = kX & \text{dvs. } Z(u) = kX(u) & \text{for alle utfall } u \end{array}$$

Vi har da følgende setninger:

F3.	$E(k) = k$
F4.	$E(k + X) = k + EX$
F5.	$E(kX) = kEX$

Setningene er nokså opplagte, men la oss illustrere dem med konkrete eksempler før vi tar de formelle begrunnelser .

Eksempel 14 : Forventning
La Z være underskudd i kroner ved å spille et sjansespill. Betrakt tre ulike spill : Det første spillet har en innsats på k kroner, og uansett utfall gir spillet ingen flere utgifter eller inntekter. Da er $Z = k$ og setningen F3 sier at forventet underskudd er lik innsatsen. Det andre spillet har også en fast utgift på k kroner, men med en ekstra utgift X som er stokastisk, dvs. avhengig av utfallet på sjansespillet (en negativ verdi vil bety inntekt). Da er $Z = k + X$, og setningen F4 sier at forventet underskudd er lik fast utgift pluss forventet ekstra utgift. I det tredje spillet er alle utgiftene (inntektene) stokastiske. Dersom X er underskudd dollar, er underskudd i kroner lik $Z = kX$ der k er lik dollarkursen på evalueringsdatoen. Setning F5 sier at forventet underskudd i kroner er lik dollarkursen ganger forventet underskudd i dollar.

Begrunnelser : (formelle, men gir god øvelse i bruk av definisjoner)

$$\begin{aligned} E(k) &= \sum_u kP(u) = k\sum_u P(u) = k \cdot 1 = k \\ E(k+X) &= \sum_u (k + X(u))P(u) \end{aligned}$$

$$
\begin{aligned}
&= k\sum_u P(u) + \sum_u X(u)P(u) = k + EX \\
E(kX) &= \sum_u kX(u)P(u) = k\sum_u X(u)P(u) = k \cdot EX
\end{aligned}
$$

De to siste setningene er spesialtilfeller av en mer generell setning om forventningen til Z dersom Z er en lineær funksjon av X. (se Oppgave 27). I en rekke sammenhenger trenger vi å beregne forventningen til en variabel Z som er avledet av X på en mer komplisert (f.eks. ikke-lineær) måte, eksempelvis $Z = X^2, Z = \sqrt{X}, Z = lnX$ etc. Den generelle definisjon på en slik variabel er som følger:

> **Definisjon :** Gitt en stokastisk variabel X.
> Med $Z = \phi(X)$ menes den stokastiske variable gitt ved
>
> $Z(u) = \phi(X(u))$ for alle utfall u.

La nå X være en stokastisk variabel med sannsynlighetsfordeling gitt ved $p(x) = P(X = x)$, og la $Z = \phi(X)$. Nå er det klart at

$$EZ = \sum_z zP(Z = z)$$

hvor summasjonen går over alle mulige verdier av Z. For å beregne EZ etter denne formelen trenger vi imidlertid å kjenne sannsynlighetsfordelingen til Z. I noen situasjoner er det lett å finne denne fordelingen ut fra kjennskapet til fordelingen til X. I så fall er saken grei. I mange situasjoner vil det være brysomt å forsøke å finne fordelingen til Z. Dette er heller ikke nødvendig, noe som følgende setning viser:

> F6. $$E\phi(X) = \sum_x \phi(x)p(x)$$
>
> hvor summasjonen går over de mulige verdier av X.

I likhet med formel F2 vises F6 ved å bruke definisjonen på forventning F1 på den avledede variable $\phi(X)$ og deretter gruppere sammen alle utfall u som gir samme verdi av X. Setningen viser at for å beregne $E(\phi(X))$, er det nok å kjenne sannsynlighetsfordelingen til X, eksempelvis blir

$$E(X^2) = \sum_x x^2 p(x)$$

et resultat som kommer til nytte senere. Den generelle setningen F6 kan brukes til å gi alternative bevis for F3, F4 og F5 (Oppgave 27).

Eksempel 15 : Forventet nytte
Anta at gevinsten i kroner ved et prosjekt er en stokastisk variabel X med kjent sannsynlighetsfordeling. Når prosjektet skal jevnføres med andre aktuelle prosjekter, kan en beregne forventet gevinst EX. Imidlertid kan et prosjekt med lavere forventet gevinst gjerne bli foretrukket, i fall dette innebærer en mindre risiko. En beslutningstakers holdning til risiko kan fanges opp med en såkalt nyttefunksjon. La derfor $U = \phi(X)$ være "nytten" av gevinsten X, der funksjonen ϕ er slik at prosjekter rangeres i henhold til deres forventede nytte $EU = E\phi(X)$. Anta at valget står mellom to prosjekter med gevinst X_1 og X_2 med sannsynlighetsfordelinger

x	0	1	4
$p_1(x)$	0.1	0.4	0.5

x	0	4	9
$p_2(x)$	0.5	0.4	0.1

Dersom nyttefunksjonen er $\phi(X) = \sqrt{X}$, får vi forventet nytte ved hjelp av formel F6 slik:

$$\begin{aligned} EU_1 &= E\sqrt{X_1} = \sum_x \sqrt{x} \cdot p_1(x) = 0 \cdot 0.1 + 1 \cdot 0.4 + 2 \cdot 0.5 = 1.4 \\ EU_2 &= E\sqrt{X_2} = \sum_x \sqrt{x} \cdot p_2(x) = 0 \cdot 0.5 + 2 \cdot 0.4 + 3 \cdot 0.1 = 1.1 \end{aligned}$$

En person med denne nyttefunksjonen vil derfor foretrekke prosjekt nr.1 framfor prosjekt nr.2, selv om forventet gevinst $EX_1 = 2.4$ er mindre enn $EX_2 = 2.5$.

Med utgangspunkt i to stokastiske variable X og Y kan vi definere nye stokastiske variable, for eksempel vil

$$\begin{aligned} Z &= X + Y \quad \text{bety at} \quad Z(u) = X(u) + Y(u) \quad \text{for alle } u \\ Z &= X \cdot Y \quad \text{bety at} \quad Z(u) = X(u) \cdot Y(u) \quad \text{for alle } u \end{aligned}$$

Følgende setning er svært viktig i praksis

F7. $E(X+Y) = EX + EY$

dvs. forventningen til en sum lik summen av forventningene. Begrunnelse:

$$\begin{aligned} E(X+Y) &= \sum_u (X(u)+Y(u))P(u) \\ &= \sum_u X(u)P(u) + \sum_u Y(u)P(u) = EX + EY \end{aligned}$$

På samme måte kunne en kanskje tro at $E(X \cdot Y) = EX \cdot EY$, dvs. forventningen til et produkt er lik produktet av forventningene. Dette er generelt ikke tilfelle. Vi kommer tilbake til dette i Kapittel 5.6.

Eksempel 16 : To terningkast
La oss definere følgende stokastiske variable

$$\begin{aligned} X &= \text{antall øyne i første kast} \\ Y &= \text{antall øyne i annet kast} \\ S &= \text{sum øyne i de to kastene.} \end{aligned}$$

Vi kan da skrive $S = X + Y$, og setning F7 gir derfor

$$ES = E(X+Y) = EX + EY$$

Med den vanlige modellen får vi (se Eksempel 8)

$$EX = EY = \frac{7}{2} \quad \text{og følgelig blir} \quad ES = \frac{7}{2} + \frac{7}{2} = 7$$

et resultat vi allerede kjenner fra Eksempel 11, men som der krevde betydelig mer regnearbeid.

Vi trenger følgende generalisering av formel F7:

La $X_1, X_2, ..., X_n$ være n stokastiske variable. Da er

F8. $E(X_1+X_2+\cdots+X_n) = EX_1+EX_2+\cdots+EX_n$

Denne formelen kan vises på samme måte som F7. Vi kan faktisk vise en enda mer generell formel, som har formlene F3 - F8 (unntatt F6) som spesialtilfeller. Den lyder

F9. $E(k_0 + k_1X_1 + k_2X_2 + \cdots + k_nX_n) =$
$k_0 + k_1EX_1 + k_2EX_2 + \cdots + k_nEX_n$

hvor $k_0, k_1, k_2, ..., k_n$ betegner konstanter.

Forventningen til en lineær kombinasjon av stokastiske variable er altså lik den samme lineære kombinasjon av forventningene. Den praktiske nytten av formlene F7, F8 og F9 ligger i følgende : Anta at man ønsker å beregne forventningen til en bestemt stokastisk variabel, og at det viser seg vanskelig å gjøre dette direkte. En mulighet er da å forsøke å uttrykke den stokastiske variable som en lineærkombinasjon av andre stokastiske variable med kjente forventninger.

5.5 Varians

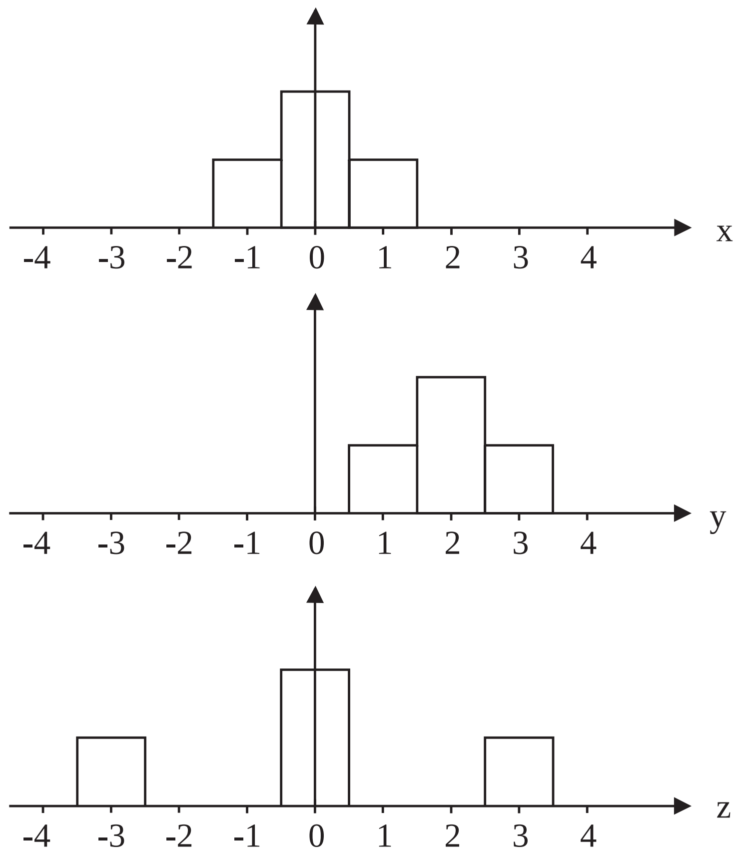
Forventningen til en stokastisk variabel kan tolkes som tyngdepunktet eller sentret i den tilhørende sannsynlighetsfordeling. I det følgende vil vi også ha bruk for et mål for spredningen i sannsynlighetsfordelingen. La oss sammenligne tre stokastiske variable X, Y og Z med sannsynlighetsfordelinger gitt ved

x	-1	0	1
$P(X = x)$	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$

y	1	2	3
$P(Y = y)$	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$

z	-3	0	3
$P(Z = z)$	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$

La oss se på de tilhørende histogrammer i Figur 5.3.



Figur 5.3: Histogrammer for X, Y og Z

Sammenlignet med fordelingen til X, ser vi at fordelingen til Y har en annen forventning men samme spredning, mens fordelingen til Z har samme forventning men større spredning. Litt løst kan vi si at et eventuelt mål for spredning i en sannsynlighetsfordeling må ta sikte på å uttrykke et slags "sannsynlig eller forventet avvik fra forventningsverdien". La oss gjøre noen betraktninger i denne forbindelse:

Forventningen EX til en stokastisk variabel X blir ofte betegnet med μ_X eller kortere μ dersom det er klart hvilken stokastisk variabel det dreier seg om. Av setning F4 i forrige avsnitt følger nå at

$$E(X - \mu) = 0$$

dvs. forventet avvik fra forventningsverdien (målt med fortegn) er alltid null, og er således ubrukelig som spredningsmål. En mulighet er nå å regne alle avvik positive, dvs. beregne tallverdien av avvikene og bruke

$$E \mid X - \mu \mid$$

som spredningsmål. Det viser seg imidlertid at dette spredningsmål er matematisk lite bekvemt, og heller ikke særlig fruktbart. Følgende spredningsmål vil vise seg å være fruktbart:

Definisjon : Med *variansen* til en stokastisk variabel X menes forventningen til $(X - \mu)^2$ hvor $\mu = EX$, og vi skriver

$$\text{V1.} \qquad varX = E(X - \mu)^2$$

Dersom fordelingen til X er godt sentrert om sin forventning μ vil $(X - \mu)^2$ med stor sannsynlighet bli liten og følgelig blir $E(X - \mu)^2$ liten. Omvendt, dersom fordelingen er lite sentrert om forventningen, vil $(X - \mu)^2$ med stor sannsynlighet bli stor og følgelig blir $E(X - \mu)^2$ stor. Setter vi $\phi(X) = (X - \mu)^2$ og bruker setning F6, får vi følgende beregningsformel for varians

$$\text{V2.} \qquad varX = \sum_x (x - \mu)^2 p(x)$$

Av denne formelen går det fram at en varians alltid er et ikke-negativt tall, og er null hvis og bare hvis den stokastiske variable er en konstant (lik μ). I innledningseksemplet får vi $EX = EZ = 0$ og $EY = 2$, og følgelig

$$\begin{aligned}
varX &= (-1-0)^2 \cdot \tfrac{1}{4} + (0-0)^2 \cdot \tfrac{1}{2} + (1-0)^2 \cdot \tfrac{1}{4} &&= \tfrac{1}{2} \\
varY &= (1-2)^2 \cdot \tfrac{1}{4} + (2-2)^2 \cdot \tfrac{1}{2} + (3-2)^2 \cdot \tfrac{1}{4} &&= \tfrac{1}{2} \\
varZ &= (-3-0)^2 \cdot \tfrac{1}{4} + (0-0)^2 \cdot \tfrac{1}{2} + (3-0)^2 \cdot \tfrac{1}{4} &&= \tfrac{9}{2}
\end{aligned}$$

Ved å kvadrere ut $(X - \mu)^2$ og bruke setningene om forventning, kan vi få en alternativ formel for varians (merk at μ er en konstant).

$$\begin{aligned}
E(X-\mu)^2 &= E(X^2 - 2\mu X + \mu^2) = E(X^2) + E(-2\mu X) + E(\mu^2) \\
&= E(X^2) - 2\mu EX + \mu^2 = E(X^2) - 2\mu^2 + \mu^2 \\
&= E(X^2) - \mu^2
\end{aligned}$$

Vi får derfor følgende setning

$$\text{V3.} \qquad varX = E(X^2) - (EX)^2$$

som gir oss regneformelen

$$\text{V4.} \qquad varX = \sum_x x^2 p(x) - (EX)^2$$

I mange situasjoner er det lettere å beregne varians ved denne formelen enn ved definisjonen.

Eksempel 17 : Et terningkast
Vi har tidligere vist at forventningen til X = antall øyne er lik $EX = 7/2$. Variansen til X blir derfor

$$\begin{aligned} varX &= E(X - \frac{7}{2})^2 = \\ &= (1 - \frac{7}{2})^2 \cdot \frac{1}{6} + (2 - \frac{7}{2})^2 \cdot \frac{1}{6} + \cdots (6 - \frac{7}{2})^2 \cdot \frac{1}{6} = \frac{35}{12} \end{aligned}$$

som alternativt kan beregnes ved

$$\begin{aligned} varX &= E(X^2) - (\frac{7}{2})^2 \\ &= 1^2 \cdot \frac{1}{6} + 2^2 \cdot \frac{1}{6} + \cdots 6^2 \cdot \frac{1}{6} - (\frac{7}{2})^2 = \frac{91}{6} - \frac{49}{4} = \frac{35}{12} \end{aligned}$$

Eksempel 18 : To terningkast
Vi har tidligere vist at forventningen til S = sum øyne er lik $ES = 7$. Variansen til S blir derfor

$$\begin{aligned} varS &= E(S - 7)^2 \\ &= (2 - 7)^2 \cdot \frac{1}{36} + (3 - 7)^2 \cdot \frac{2}{36} + \cdots + (12 - 7)^2 \cdot \frac{1}{36} = \frac{35}{6} \end{aligned}$$

som alternativt kan beregnes slik

$$\begin{aligned} varS &= E(S^2) - (7)^2 \\ &= 2^2 \cdot \frac{1}{36} + 3^2 \cdot \frac{2}{36} + \cdots + 12^2 \cdot \frac{1}{36} - 7^2 = \frac{35}{6} \end{aligned}$$

Eksempel 19 : Tombola
La situasjonen være som i Eksempel 9, hvor vi trakk et lodd tilfeldig fra en tombola med N lodder med verdier henholdsvis $v_1, v_2, ..., v_N$. Vi husker at forventningen til Y = verdien av det uttrukne lodd ble $EY = \overline{v} = \frac{1}{N}\sum_{i=1}^{N} v_i$. Variansen til Y blir

$$varY = E(Y-\overline{v})^2 = \sum_{i=1}^{N}(v_i-\overline{v})^2 \cdot \frac{1}{N} = \frac{1}{N}\sum_{i=1}^{N}(v_i-\overline{v})^2$$

dvs. lik verdienes gjennomsnittlige kvadrerte avvik fra gjennomsnittet.[5]) Alternativt får vi

$$varY = E(Y^2) - (EY)^2 = \frac{1}{N}\sum_{i=1}^{N} {v_i}^2 - (\overline{v})^2.$$

Det siste uttrykket kan også fås ved å kvadrere ut uttrykket ovenfor og summere (god øvelse, se også Oppgave 1.11).

Vi skal merke oss følgende setninger om varians:

V5.	$var(k) = 0$ (k konstant)
V6.	$var(X+k) = varX$
V7.	$var(kX) = k^2 varX$

Begrunnelser : Brukes setningene F3, F4 og F5 i forrige avsnitt, får vi

$$\begin{aligned}
var(k) &= E(k-Ek)^2 = E(k-k)^2 = E(0) = 0\\
var(X+k) &= E(X+k-E(X+k))^2 = E(X+k-(EX+k))^2\\
&= E(X-EX)^2 = varX\\
var(kX) &= E(kX-E(kX))^2 = E(kX-kEX)^2 = E(k(X-EX))^2\\
&= E(k^2(X-EX)^2) = k^2E(X-EX)^2 = k^2varX
\end{aligned}$$

Vi vil også trenge en formel for variansen til en sum. Det er kanskje grunn til å tro at $var(X+Y)$ er lik $varX + varY$. Dette er imidlertid generelt ikke tilfellet. Vi skal diskutere dette i neste avsnitt.

[5]Denne kvadratsum brukes ofte til å karakterisere spredningen av observasjonene i et tallmateriale og kalles da empirisk varians.

Kvadratroten av $varX$ kalles *standardavviket* til X og betegnes ofte $\sigma(X)$, σ_X eller kortere σ dersom det er klart hvilken stokastisk variabel det er tale om, altså

$$\sigma = \sqrt{varX}$$

Med denne skrivemåten blir altså $\sigma^2 = varX$. Det er klart at begrepet standardavvik i og for seg er overflødig; kjenner vi variansen, kjenner vi samtidig standardavviket og omvendt. Standardavviket til X har imidlertid den egenskap at det har samme dimensjon som X og $\mu = EX$, slik at spredningen målt ved σ kan avsettes på x-aksen (f. eks. i histogrammet) til hver side av μ som et slags "sannsynlige avvik". Med de to alternative begrepene varians og standardavvik oppnår vi også en større grad av fleksibilitet, noen saksforhold kan uttrykkes mest oversiktlig med varianser, andre med standardavvik. La X være en stokastisk variabel hvor $\mu = EX$ og $\sigma^2 = varX$. Da blir

$$Z = \frac{X - \mu}{\sigma}$$

kalt *den standardiserte variable til* X. Bruker vi regnereglene for forventning og varians ser vi at

$$EZ = E(\frac{X - \mu}{\sigma}) = \frac{1}{\sigma}E(X - \mu) = \frac{1}{\sigma}(EX - \mu) = 0$$

$$varZ = var(\frac{X - \mu}{\sigma}) = (\frac{1}{\sigma})^2 var(X - \mu) = \frac{1}{\sigma^2}varX = 1$$

En standardisert variabel har altså alltid forventning null og varians (og standardavvik) lik 1. Vi vil komme tilbake til bruken av standardiserte variable i neste kapittel. Begrepene forventning og varians vil komme til nytte senere i Kapittel 6 i forbindelse med tilnærmet beregning av visse sannsynligheter. Disse begreper spiller også en sentral rolle i utviklingen av statistisk teori.

5.6 Simultane fordelinger, uavhengighet og kovarians

La X og Y være to stokastiske variable definert i forhold til det samme utfallsrommet Ω. Vi skal nå interessere oss for begivenheter av typen

$$(X = x) \cap (Y = y)$$

og spesielt sannsynlighetene for slike begivenheter. [6]

> **Definisjon :** *Den simultane sannsynlighetsfordeling* til X og Y er en oppregning av de mulige verdipar (x, y) for (X, Y) sammen med sannsynlighetene for at (X, Y) antar disse verdipar. Den simultane sannsynlighetsfordeling for X og Y kan beskrives ved å angi funksjonen
>
> $$p(x, y) = P((X = x) \cap (Y = y))$$
>
> definert for de mulige verdipar (x, y). Funksjonen $p(x, y)$ vil vi kalle *den simultane fordelingsfunksjonen* for X og Y.

Eksempel 20 : Tre myntkast
Betrakt de to stokastiske variable X = antall sideskift og Y= antall kron, dvs.

u	:	KKK	KKM	KMK	MKK	KMM	MKM	MMK	MMM
$X(u)$	:	0	1	2	1	1	2	1	0
$Y(u)$	:	3	2	2	2	1	1	1	0

Her blir eksempelvis

$$\begin{aligned}
(X = 0) \cap (Y = 0) &= \{\text{MMM}\} \\
(X = 0) \cap (Y = 1) &= \emptyset \\
(X = 1) \cap (Y = 1) &= \{\text{KMM}, \text{MMK}\}
\end{aligned}$$

Den simultane fordelingen for X og Y er da gitt ved tabellen:

$x \backslash y$	0	1	2	3	Sum
0	1/8	0	0	1/8	2/8
1	0	2/8	2/8	0	4/8
2	0	1/8	1/8	0	2/8
Sum	1/8	3/8	3/8	1/8	1

[6] Man pleier å kalle (X, Y) for en stokastisk vektor, og (x, y) en realisasjon av denne vektoren.

For hvert tallpar (x, y) viser tabellen $P((X = x) \cap (Y = y))$. Ved å summere hver linje i tabellen får vi søylen til høyre. Dette gir oss sannsynlighetsfordelingen til X. Ved å summere hver søyle i tabellen får vi linjen nederst i tabellen. Dette gir oss sannsynlighetsfordelingen til Y. Begge summeres til en.

Når den simultane sannsynlighetsfordeling til X og Y er kjent, kan vi beregne sannsynlighetsfordelingene til X og Y hver for seg. Disse kalles nå *marginale fordelinger:*

La $y_1, y_2, ..., y_k$ være de mulige verdier av Y. Vi kan da skrive

$$\begin{aligned}(X = x) &= ((X = x) \cap (Y = y_1)) \cup ((X = x) \cap (Y = y_2)) \cup \ldots \\ &\quad \cup ((X = x) \cap (Y = y_k))\end{aligned}$$

Siden denne unionen er disjunkt, får vi

$$\begin{aligned}P(X = x) &= P((X = x) \cap (Y = y_1)) + P((X = x) \cap (Y = y_2)) + \cdots \\ &\quad + P((X = x) \cap (Y = y_k)) \\ &= p(x, y_1) + p(x, y_2) + \cdots + p(x, y_k)\end{aligned}$$

Betegner vi den marginale fordelingsfunksjonen til X med $p_1(x)$, får vi altså formelen

$$p_1(x) = \sum_y p(x, y)$$

hvor summasjonen går over de mulige verdier av Y. På tilsvarende måte, dersom vi betegner den marginale fordelingsfunksjonen til Y med $p_2(y)$, får vi formelen

$$p_2(y) = \sum_x p(x, y)$$

hvor summasjonen går over de mulige verdier av X. Marginale fordelingsfunksjoner kan altså utledes fra den simultane fordelingsfunksjon ved å "summere bort" den andre variablen (se Eksempel 20).

Det er ofte aktuelt å studere simultane fordelinger til mer enn to stokastiske variable: Med den *simultane fordelingsfunksjon* til $X_1, X_2, ..., X_n$ menes funksjonen

$$p(x_1, x_2, ..., x_n) = P((X_1 = x_1) \cap (X_2 = x_2) \cap ... \cap (X_n = x_n))$$

Resultatene ovenfor kan generaliseres, dvs. marginale sannsynligheter finnes av simultane sannsynligheter ved å "summere bort" de andre variablene.

Definisjon : Dersom begivenhetene $(X = x)$ og $(Y = y)$ er uavhengige, dvs.

$$P((X = x) \cap (Y = y)) = P(X = x) \cdot P(Y = y),$$

for alle mulige verdipar (x, y), sier vi at X og Y er *uavhengige stokastiske variable.*

Av definisjonen følger at X og Y er uavhengige hvis og bare hvis

$$p(x, y) = p_1(x) \cdot p_2(y)$$

dvs. den simultane fordelingsfunksjon er lik produktet av de marginale. Vi ser at i Eksempel 16 er X og Y uavhengige, mens i Eksempel 20 er X og Y ikke uavhengige, dvs. avhengige. Mer generelt sier vi at $X_1, X_2, ..., X_n$ er *uavhengige stokastiske variable* dersom

$$p(x_1, x_2, ..., x_n) = p_1(x_1)p_2(x_2) \cdots p_n(x_n)$$

hvor $p_i(x_i)$ betegner den marginale fordelingsfunksjon til X_i. Begrepet uavhengighet blir som oftest brukt konstruktivt i den forstand at når man kjenner de marginale fordelingsfunksjonene til $X_1, X_2, ..., X_n$ og er villig til å anta at at $X_1, X_2, ..., X_n$ er uavhengige stokastiske variable, så er den simultane fordelingsfunksjonen gitt ved formelen ovenfor.

Eksempel 21 : Solgt kvantum
Et firma antar at solgt kvantum pr. dag ved et av sine to utsalg er en stokastisk variabel X med sannsynlighetsfordeling

x	0	1	2
$P(X = x)$	0.6	0.3	0.1

For senere bruk beregner vi $EX = 0.5$ og $varX = 0.45$. Solgt kvantum pr. dag ved det andre utsalget antas å være en stokastisk variabel Y med samme sannsynlighetsfordeling som X. Antas uavhengighet blir den simultane fordeling for (X, Y):

$x \backslash y$	0	1	2	Sum
0	0.36	0.18	0.06	0.6
1	0.18	0.09	0.03	0.3
2	0.06	0.03	0.01	0.1
Sum	0.6	0.3	0.1	1.0

dvs. hver sannsynlighet inne i tabellen er beregnet som produktet av de respektive marginale sannsynligheter. Ut fra dette kan vi bl. a. finne sannsynlighetsfordelingen til samlet solgt kvantum $S = X + Y$. Vi får

s	0	1	2	3	4
$P(S = s)$	0.36	0.36	0.21	0.06	0.01

Diskuter antakelsen om uavhengighet!

Anta at X og Y er to stokastiske variable med simultan fordelingsfunksjon $p(x, y)$. La oss studere en ny stokastisk variabel Z avledet av X og Y ved at $Z = \phi(X, Y)$. Vi ønsker å beregne $EZ = E\phi(X, Y)$. Nå er det klart at $EZ = \sum zP(Z = z)$, hvor summasjonen går over alle mulige verdier av Z. For å beregne EZ etter denne formelen trenger vi imidlertid å kjenne sannsynlighetsfordelingen til Z. I noen situasjoner er det lett å finne denne fordelingen ut fra kjennskapet til den simultane fordeling til X og Y. Som regel kommer vi raskere fram til svaret ved å bruke følgende formel

$$\text{F10.} \qquad E\phi(X, Y) = \sum_{(x,y)} \phi(x, y)p(x, y)$$

hvor summasjonen går over de mulige verdipar (x, y).

Begrunnelsen er analog med den for formel F6: Bruk definisjonen av forventning for den avledede variable, og grupper så sammen alle utfall som gir samme verdipar (x, y). Vi ser at for å beregne forventningen til $Z = \phi(X, Y)$ er det nok å kjenne den simultane fordeling til X og Y: For alle mulige verdipar (x, y) beregnes verdien $\phi(x, y)$ som veies med sannsynligheten $p(x, y)$ hvoretter resultatene summeres.

Formelen ovenfor kan nyttes til å gi et alternativt bevis for setningen om forventningen til en sum (Oppgave 28). Vi kan nå også vise et annet viktig resultat, nemlig

F11. $\quad E(X \cdot Y) = EX \cdot EY$ såframt X og Y er uavhengige.

Begrunnelse : Vi bruker F10 med $\phi(X, Y) = X \cdot Y$, og bruker uavhengigheten:

$$
\begin{aligned}
E(X \cdot Y) &= \sum_{(x,y)} xy \cdot p(x,y) = \sum_{(x,y)} xy \cdot p_1(x)p_2(y) \\
&= \sum_y \sum_x xy \cdot p_1(x)p_2(y) = \sum_y yp_2(y)(\sum_x xp_1(x)) \\
&= (\sum_x xp_1(x))(\sum_y yp_1(y)) = EX \cdot EY
\end{aligned}
$$

Det må understrekes at ved avhengighet er det ikke tilfelle at forventningen til et produkt er lik produktet av forventningene.

Eksempel 22 : Kjøp - salg
Et firma supplerer sitt lager av en bestemt reservedel ved bestilling (fra utlandet) en gang hvert halvår. Antall enheter X som må bestilles neste gang betraktes som en stokastisk variabel med forventning $EX = 2.5$. Prisen pr. enhet på bestillingstidspunktet er også en stokastisk variabel Y med forventning $EY = 2000$ (i kroner). Utgiften ved neste bestilling blir $Z = X \cdot Y$. Her er det rimelig å anta at X og Y er uavhengige, slik at forventet utgift blir $EZ = EX \cdot EY = 2.5 \cdot 2000 = 5000$. Anta at det isteden dreier seg om solgt kvantum X av en artikkel og Y er markedspris slik at $Z = X \cdot Y$ er salgsinntekten. Dersom en ønsker å beregne forventet salgsinntekt, er det ikke uten videre rimelig å anta at salget er uavhengig av markedspris, og da kan formel F11 ikke benyttes.

Vi husker at variansen til en stokastisk variabel X kunne oppfattes som et mål for spredning eller variasjon. I det følgende vil vi også trenge et mål for samvariasjon mellom to stokastiske variable X og Y. Følgende begrep viser seg å være fruktbart:

Definisjon : Med *kovariansen* mellom to stokastiske variable X og Y menes

C1. $$cov(X, Y) = E((X - EX)(Y - EY))$$

Dersom $cov(X, Y) = 0$, sier vi at X og Y er *ukorrelerte.*

La oss gi en motivering for denne definisjonen: Dersom den simultane sannsynlighetsfordeling til (X, Y) er slik at store (små) verdier av X hører sammen med store (små) verdier av Y, vil produktet $(X - EX) \cdot (Y - EY)$ bli positivt med stor sannsynlighet. Følgelig blir forventningen av dette produktet positivt, og dermed blir kovariansen mellom X og Y positiv. Omvendt dersom små (store) verdier av X gjennomgående hører sammen med store (små) verdier av Y vil produktet $(X - EX) \cdot (Y - EY)$ bli negativt med stor sannsynlighet. Følgelig blir forventningen negativ, og dermed blir kovariansen mellom X og Y negativ. Dersom det er liten grad av samvariasjon mellom X og Y så blir kovariansen omkring null. Vi vil finne en alternativ formel for kovarians. Betrakt

$$(X - EX)(Y - EY) = X \cdot Y - X \cdot EY - Y \cdot EX + EX \cdot EY$$

Tar vi forventningen på begge sider får vi, siden EX, EY og $EX \cdot EY$ er konstanter, at

$$E(X - EX)(Y - EY) = E(X \cdot Y) - EX \cdot EY - EX \cdot EY + EX \cdot EY,$$

Vi har derfor

$$\boxed{\text{C2.} \qquad cov(X, Y) = E(X \cdot Y) - EX \cdot EY}$$

som gir oss følgende beregningsformel

$$\boxed{\text{C3.} \qquad cov(X, Y) = \sum_{(x,y)} xy \cdot p(x, y) - EX \cdot EY}$$

Ofte er det lettere å bruke denne formelen ved beregning av kovarians enn definisjonen. Dersom X og Y er uavhengige stokastiske variable husker vi at $E(X \cdot Y) = EX \cdot EY$. Innsettes dette i C2 ser vi at i dette tilfellet blir $cov(X, Y) = 0$. Vi har derfor vist at

$$\boxed{\text{C4.} \qquad p(x, y) = p_1(x) \cdot p_2(y) \text{ medfører } cov(X, Y) = 0}$$

dvs. uavhengighet medfører ukorrelerthet. At det omvendte ikke nødvendigvis er tilfelle viser bl. a. Eksempel 20 der vi finner at kovariansen mellom X og Y er null, men variablene er likevel avhengige.

Eksempel 23 : Solgt kvantum
For situasjonen i Eksempel 21 blir kovariansen mellom de solgte kvanta X og Y pga. uavhengigheten lik null. Anta istedenfor uavhengighet at den simultane fordeling for X og Y er gitt ved

$x\backslash y$	0	1	2	Sum
0	0.40	0.16	0.04	0.6
1	0.16	0.10	0.04	0.3
2	0.04	0.04	0.02	0.1
Sum	0.6	0.3	0.1	1.0

Vi ser at de marginale fordelingene er de samme som i Eksempel 21, og vi husker at

$$EX = EY = 0.5 \qquad varX = varY = 0.45$$

I dette tilfellet er det imidlertid positiv samvariasjon mellom X og Y. Vi har nemlig

$$E(X \cdot Y) = 0 \cdot 0 \cdot 0.40 + 0 \cdot 1 \cdot 0.16 + \cdots + 2 \cdot 2 \cdot 0.02 = 0.34$$

og ifølge C2 blir da $cov(X, Y) = 0.34 - 0.5 \cdot 0.5 = 0.09$, dvs. positiv. Anta at den simultane fordeling for X og$Y$ isteden er gitt ved

$x\backslash y$	0	1	2	Sum
0	0.33	0.19	0.08	0.6
1	0.19	0.09	0.02	0.3
2	0.08	0.02	0.00	0.1
Sum	0.6	0.3	0.1	1.0

Nå er samvariasjonen negativ, og beregning av kovariansen gir $cov(X, Y) = -0.08$. For ytterligere diskusjon se Oppgave 30.

Nå er tiden inne for å presentere formelen for variansen til en sum:

$$\text{V8.} \qquad var(X+Y) = varX + varY + 2cov(X,Y)$$

Begrunnelse : Av definisjonen på varians følger at

$$\begin{aligned} var(X+Y) &= E((X+Y)-E(X+Y))^2 = E((X-EX)+(Y-EY))^2 \\ &= E((X-EX)^2+(Y-EY)^2+2(X-EX)(Y-EY)) \\ &= E(X-EX)^2+E(Y-EY)^2+2E((X-EX)(Y-EY)) \end{aligned}$$

hvor vi har benyttet de kjente regnereglene for forventning.

Av formelen ovenfor kan vi trekke en viktig konklusjon

V9. $var(X+Y) = varX+varY$ når X og Y er ukorrelerte

dvs. ukorrelerthet (og derfor spesielt uavhengighet) medfører at variansen til en sum er lik summen av variansene.

Eksempel 24 : To terningkast
Med den tidligere benyttede modell er det klart at X = antall øyne i første kast og Y = antall øyne i annet kast er uavhengige stokastiske variable. Vi husker (se Eksempel 16) at $varX = varY = 35/12$. Følgelig blir variansen til S = sum øyne

$$varS = var(X+Y) = varX + varY = \frac{35}{12} + \frac{35}{12} = \frac{35}{6}$$

som stemmer med det vi fant med betydelig mer regnearbeid i Eksempel 17.

Formel V8. kan generaliseres (Oppgave 34).

$$\text{V10.} \qquad var(\sum_{i=1}^{n} X_i) = \sum_{i=1}^{n} var(X_i) + \sum_{i \neq j} cov(X_i, X_j)$$

dvs. variansen til en sum er lik summen av variansene pluss summen av alle mulige par av kovarianser. Merk at siden $cov(X_1,X_2) = cov(X_2,X_1)$ etc., vil to og to av disse kovariansene være like. Vi merker oss at

$$\boxed{\text{V11.}\quad var(\sum_{i=1}^{n} X_i) = \sum_{i=1}^{n} var(X_i) \quad \text{såframt} \quad cov(X_i, X_j) = 0}$$

Denne formelen gjelder spesielt når $X_1, X_2, ..., X_n$ er uavhengige.

Kovarians som mål for samvariasjon lider av den svakhet at det er avhengig av måleenhetene for variablene, f.eks. om et beløp er målt i kroner eller øre. Dette unngår vi ved å bruke *korrelasjonskoeffisienten* definert ved

$$\boxed{\text{C5.}\qquad \rho(X,Y) = \frac{cov(X,Y)}{\sqrt{varX}\cdot\sqrt{varY}}}$$

Korrelasjonskoeffisienten har egenskapen

$$-1 \leq \rho(X,Y) \leq +1.$$

Dette er det modellteoretiske motstykke til den empiriske korrelasjonskoeffisienten definert i Kapittel 1 og begrepene knyttes sammen i Kapittel 9.

5.7 Betingede sannsynlighetsfordelinger

La (X,Y) ha simultan fordelingsfunksjon

$$p(x,y) = P((X=x)\cap(Y=y))$$

og la de marginale fordelingsfunksjonene være henholdsvis

$$p_1(x) = P(X=x) \quad \text{og} \quad p_2(y) = P(Y=y).$$

La nå x være slik at $p_1(x) > 0$, og la oss beregne betingede sannsynligheter gitt at $X = x$. Vi får

$$P(Y=y \mid X=x) = \frac{P((X=x)\cap(Y=y))}{P(X=x)} = \frac{p(x,y)}{p_1(x)}$$

Dette blir kalt *den betingede sannsynlighetsfordeling for Y gitt $X = x$*. På tilsvarende måte blir, når $p_2(y) > 0$

$$P(X=x \mid Y=y) = \frac{P((X=x)\cap(Y=y))}{P(Y=y)} = \frac{p(x,y)}{p_2(y)}$$

kalt *den betingede sannsynlighetsfordeling for* X *gitt* $Y = y$. Betingede sannsynlighetsfordelinger har samme egenskaper som vanlige sannsynlighetsfordelinger (Oppgave 37). Forventningen til Y beregnet i den betingede fordeling gitt $X = x$ skriver vi

$$E(Y \mid X = x)$$

og kaller *den betingede forventning* for Y gitt $X = x$. På tilsvarende måte betyr $E(X \mid Y = y)$ den betingede forventning for X gitt $Y = y$.

Eksempel 25 : Betinget forventning
Anta at (X, Y) har simultan fordeling som i Eksempel 20. Da blir den betingede fordeling for Y gitt $X = 1$

y	0	1	2	3
$P(Y = y \mid X = 1)$	0	1/2	1/2	0

mens den betingede fordeling for X gitt $Y = 1$ er

x	0	1	2
$P(X = x \mid Y = 1)$	0	2/3	1/3

Vi ser at $E(Y \mid X = 1) = 3/2$, mens $E(X \mid Y = 1) = 4/3$

Det er lett å vise (Oppgave 37) at X og Y er uavhengige hvis og bare hvis den betingede fordeling for Y gitt $X = x$ er lik den marginale fordeling for Y for alle verdier av x.

5.8 ⋆Mer om forventning

Vi vil avslutte dette kapitlet med å presentere en del supplerende stoff om sannsynlighetsfordelinger og forventning. Vi vil gi noen eksempler på anvendelser av teorien ovenfor, samtidig som vi gjør noen spredte betraktninger over emner som kan øke vår teoretiske innsikt. Disse er, i rekkefølge, konvolusjon, setningen om dobbeltforventning, prediksjon med kvadratisk tap, muligheten for at en forventning er uendelig, samt St. Petersburg spillet.

Eksempel 26 : Samlet etterspørsel
En vare etterspørres av tre kunder, og daglig etterspurt kvantum fra disse oppfattes som tre stokastiske variable X_1, X_2 og X_3. Disse antas å være uavhengige med samme sannsynlighetsfordeling gitt ved

x	0	1	2	3
$P(X=x)$	0.4	0.3	0.2	0.1

Vi ser at $EX = 0 \cdot 0.4 + 1 \cdot 0.3 + 2 \cdot 0.2 + 3 \cdot 0.1 = 1.0$. Samlet daglig etterspørsel etter varen blir

$$S = X_1 + X_2 + X_3.$$

Vi ønsker å finne sannsynlighetsfordelingen til S. Den viser seg å bli

s	0	1	2	3	4	5	6	7	8	9
$P(S=s)$	0.064	0.144	0.204	0.219	0.174	0.111	0.056	0.021	0.006	0.001

Vi ser eksempelvis at

$$\begin{aligned} P(S \geq 6) &= P(S=6) + P(S=7) + P(S=8) + P(S=9) \\ &= 0.056 + 0.021 + 0.006 + 0.001 = 0.084. \end{aligned}$$

Forventningen til S blir

$$ES = 0 \cdot 0.064 + 1 \cdot 0.144 + 2 \cdot 0.204 + \cdots + 9 \cdot 0.001 = 3.0$$

men dette følger enklere av at

$$ES = EX_1 + EX_2 + EX_3 = 1.0 + 1.0 + 1.0 = 3.0.$$

Utledning av sannsynlighetsfordelingen:

$(S=0)$	inntreffer på en måte:	0+0+0
$(S=1)$	inntreffer på tre måter:	0+0+1, 0+1+0, 1+0+0
$(S=2)$	inntreffer på seks måter:	0+0+2, 0+2+0, 2+0+0
		0+1+1, 1+0+1, 1+1+0
$(S=3)$	inntreffer på ti måter:	0+0+3, 0+3+0, 3+0+0
		0+1+2, 1+0+2, 2+0+1
		0+2+1, 1+2+0, 2+1+0
		1+1+1
$\vdots$		
$(S=9)$	inntreffer på en måte:	3+3+3.

Eksempelvis blir (legg merke til hvor vi bruker antakelsen om uavhengighet)

$$\begin{aligned} P(S=1) &= P((X_1=0) \cap (X_2=0) \cap (X_3=1)) \\ &+ P((X_1=0) \cap (X_2=1) \cap (X_3=0)) \end{aligned}$$

$$
\begin{aligned}
&+ \quad P((X_1 = 1) \cap (X_2 = 0) \cap (X_3 = 0)) \\
&= \quad P(X_1 = 0) \cdot P(X_2 = 0) \cdot P(X_3 = 1) \\
&+ \quad P(X_1 = 0) \cdot P(X_2 = 1) \cdot P(X_3 = 0) \\
&+ \quad P(X_1 = 1) \cdot P(X_2 = 0) \cdot P(X_3 = 0) \\
&= \quad 3 \cdot 0.4^2 \cdot 0.3 = 0.144.
\end{aligned}
$$

Det overlates til leseren å verifisere de andre tallene i tabellen. Selv om dette var et relativt enkelt eksempel, ser vi at det medfører et betydelig regnearbeid.

Problemer som dreier seg om summer av uavhengige stokastiske variable med samme sannsynlighetsfordeling dukker opp i mange anvendelser, og er blitt tatt opp på prinsipielt grunnlag under betegnelsen *konvolusjon*. Det er utviklet spesielle teknikker for å løse slike problemer, men det vil føre for langt å komme inn på disse her. En type problemer av en litt annen karakter enn forrige eksempel har vi i det følgende.

Eksempel 27 : Etterspørsel i leveringstiden
La X være daglig etterspørsel etter en vare. Anta at X har sannsynlighetsfordeling gitt ved

x	0	1	2	3
$P(X = x)$	0.4	0.3	0.2	0.1

Vi ser at forventet daglig etterspørsel er

$$EX = 0 \cdot 0.4 + 1 \cdot 0.3 + 2 \cdot 0.2 + 3 \cdot 0.1 = 1.$$

Anta at etterspørselen på suksessive dager $X_1, X_2, X_3, \ldots$ kan oppfattes som uavhengige stokastiske variable med samme sannsynlighetsfordeling som X. Anta videre at leveringstiden ved bestilling av varen er en stokastisk variabel N som antar verdier blant de naturlige tall, som er uavhengig av X_i'ene og med sannsynlighetsfordeling

n	1	2	3
$P(N = n)$	0.25	0.50	0.25

La S_N være etterspørselen i leveringstiden. Da er

$$S_N = X_1 + \cdots + X_N.$$

Her vil antall ledd i summen være stokastisk, i motsetning til i Eksempel 26. Vi ønsker å finne sannsynlighetsfordelingen til S_N. Den viser seg å bli

s	0	1	2	3	4	5	6	7	8	9
$P(S = s)$	0.196	0.231	0.226	0.180	0.094	0.048	0.019	0.005	0.001	0.000

Vi ser eksempelvis at

$$\begin{aligned} P(S_N \geq 5) &= P(S_N = 5) + \cdots + P(S_N = 9) \\ &= 0.048 + 0.019 + 0.005 + 0.001 + 0.000 = 0.073. \end{aligned}$$

Forventningen til S_N blir (dersom vi bruker alle desimalene)

$$ES_N = 0 \cdot 0.196 + 1 \cdot 0.231 + 2 \cdot 0.226 + \cdots + 9 \cdot 0.000 = 2.0.$$

Vi ser at $ES_N = EN \cdot EX$, og vi skal vise nedenfor av dette generelt er tilfelle i slike situasjoner det her er tale om. En begrunnelse for sannsynlighetsfordelingen ovenfor krever, med de teknikker vi har til rådighet her, betydelig regnearbeid. De ulike sannsynlighetene kan finnes ved å betinge med hensyn på lengden av leveringstiden, eksempelvis (legg merke til hvor vi bruker antakelsene om uavhengighet)

$$\begin{aligned} P(S_N = 2) &= P(N = 1)P(S_N = 2 \mid N = 1) + P(N = 2)P(S_N = 2 \mid N = 2) \\ &\quad + P(N = 3)P(S_N = 2 \mid N = 3) \\ &= \frac{1}{4}P(X_1 = 2) + \frac{1}{2}P(X_1 + X_2 = 2) + \frac{1}{4}P(X_1 + X_2 + X_3 = 2) \end{aligned}$$

Ifølge opplysningene er $P(X_1 = 2) = 0.2$, mens

$$\begin{aligned} P(X_1 + X_2 = 2) &= P((X_1 = 0) \cap (X_2 = 2)) \\ &+ P((X_1 = 2) \cap (X_2 = 0)) \\ &+ P((X_1 = 1) \cap (X_2 = 1)) \\ &= P(X_1 = 0) \cdot P(X_2 = 2) \\ &+ P(X_1 = 2) \cdot P(X_2 = 0) \\ &+ P(X_1 = 1) \cdot P(X_2 = 1) \\ &= 2 \cdot 0.4 \cdot 0.2 + 0.3 \cdot 0.3 = 0.25 \end{aligned}$$

På tilsvarende måte finner vi (se også Eksempel 26)

$$P(X_1 + X_2 + X_3 = 2) = 0.204.$$

Innsatt ovenfor gir dette

$$P(S_N = 2) = \frac{1}{4} \cdot 0.2 + \frac{1}{2} \cdot 0.25 + \frac{1}{4} \cdot 0.204 = 0.226$$

Det overlates til leseren å verifisere de andre sannsynlighetene i tabellen ovenfor.

Situasjonen ovenfor er et eksempel på en såkalt *forgreningsprosess*. Slike problemstillinger dukker opp i mange anvendelser, og det er derfor utviklet generelle teorier for analyse av slike, noe som det vil føre for langt å komme inn på her.

Vi vil nå presentere *setningen om dobbeltforventning* som kan være et nyttig hjelpemiddel i mange anvendelser: La X og Y være to stokastiske variable med simultan sannsynlighetsfordeling $p(x,y) = P((X = x) \cap (Y = y))$. Vi kan da beregne $E(Y \mid X = x)$. Denne forventning vil generelt være en funksjon av x, vi skriver

$$\phi(x) = E(Y \mid X = x).$$

Betrakt den stokastiske variable $\phi(X)$. Denne stokastiske variable blir ofte betegnet $E(Y \mid X)$. La oss studere forventningen til denne variable.

$$\begin{aligned} E\phi(X) &= \sum_x \phi(x)p_1(x) = \sum_x E(Y \mid X = x)p_1(x) \\ &= \sum_x (\sum_y yp_2(y \mid x))p_1(x) = \sum_x \sum_y yp(x,y) \\ &= \sum_y y \sum_x p(x,y) = \sum_y yp_2(y) = EY \end{aligned}$$

der $p_2(y \mid x)$ er den betingede fordeling for Y gitt $X = x$. Vi har dermed vist at

$$\boxed{\text{F12.} \qquad EE(Y \mid X) = EY}$$

og det er dette som kalles setningen om dobbeltforventning. Forventningen til Y kan altså beregnes som en veid sum over x av $E(Y \mid X = x)$ med vekter lik sannsynlighetene for at $X = x$. I praksis kan setningen om dobbeltforventningen utnyttes på følgende måte:

I situasjoner der forventningen EY skal beregnes, men dette ikke er lett å gjøre direkte, kan en ofte finne en egnet stokastisk variabel X, beregne den betingede forventning av Y gitt X, og så bruke regelen om dobbeltforventning. La oss se på et par slike anvendelser :

Eksempel 28 : Spedisjon
Et vareparti skal sendes fra A til B med en av tre mulige spedisjonsfirmaer nummerert 1, 2, og 3. Av erfaring vet man at tiden det tar kan variere. La

X = nummer på det spedisjonsfirma som brukes.

Y = tiden i døgn fra innlevering til mottak .

Anta at en vurderer situasjonen dithen at

y	1	2	3
$P(Y = y \mid X = 1)$	0.5	0.4	0.1
$P(Y = y \mid X = 2)$	0.3	0.4	0.3
$P(Y = y \mid X = 3)$	0.2	0.6	0.2

Anta videre, basert på erfaring med pristilbud og service etc. at det er like stor sjanse for at oppdraget blir gitt til firma nr. 1 som til de andre to tilsammen, mens de to andre firmaene er like sannsynlige. Hva blir forventet leveringstid (før det er fastlagt hvem oppdraget blir gitt til). Vi ser at

$$E(Y \mid X=1) = 1.6 \quad E(Y \mid X=2) = 2.0 \quad E(Y \mid X=3) = 2.0$$

og bruker vi setningen om dobbeltforventning blir

$$\begin{aligned} E(Y) &= E(Y \mid X=1)P(X=1) + E(Y \mid X=2)P(X=2) \\ &\quad + E(Y \mid X=3)P(X=3) \\ &= 1.6 \cdot 0.50 + 2.0 \cdot 0.25 + 2.0 \cdot 0.25 = 1.8. \end{aligned}$$

Det overlates til leseren å vise at

y	1	2	3
$P(Y=y)$	0.375	0.450	0.175

slik at vi igjen kan finne

$$EY = 1 \cdot 0.375 + 2 \cdot 0.450 + 3 \cdot 0.175 = 1.8.$$

I dette eksemplet er det nok omtrent det samme arbeid ved begge utregningsmetoder, men det finnes eksempler der utregning uten bruk av setningen om dobbeltforventning blir mye mer komplisert.

En spesielt viktig anvendelse oppstår i følgende situasjon: La $X_1, X_2, X_3, ...$ være stokastiske variable som alle har samme forventning, kall den EX. La n være gitt naturlig tall og la

$$S_n = X_1 + X_2 + \cdots + X_n.$$

Av formel F8 følger at

$$\boxed{\text{F13.} \qquad ES_n = n \cdot EX.}$$

Anta nå at N er en stokastisk variabel som antar verdier blant de naturlige tall og som er uavhengige av X_i'ene. Betrakt

$$S_N = X_1 + X_2 + \cdots + X_N.$$

Her er, i motsetning til S_n, antall ledd i summen stokastisk. Følgende formel synes rimelig

$$\boxed{\text{F14.} \qquad ES_N = EN \cdot EX.}$$

Begrunnelse : Vi ser at

$$E(S_N \mid N = n) = E(S_n \mid N = n) = ES_n = nEX.$$

Nest siste likhet er en følge av at N og X_i'ene er uavhengige, og følgelig vil S_n og N være uavhengige. Vi har derfor

$$E(S_N \mid N) = NEX.$$

Ifølge setningen om om dobbeltforventning blir

$$ES_N = E(E(S_N \mid N)) = E(N \cdot EX) = EX \cdot EN$$

fordi EX er en konstant som kan settes utenfor det ytterste forventningstegnet. Under forutsetningen av at X_i'ene er uavhengige stokastiske variable med samme forventning og varians, finnes det også regneformler for variansen til summer av typen ovenfor. I tilfellet med sum med et gitt antall ledd

$$\boxed{\text{V12.} \qquad varS_n = nvarX.}$$

I tilfellet med et stokastisk antall ledd

$$\boxed{\text{V13.} \qquad varS_N = EN \cdot varX + varN \cdot (EX)^2.}$$

Det vil føre for langt å vise denne formel her. Setning F14 og V13 kan blant annet nyttes til å finne forventning og varians til etterspørsel i leveringstiden i Eksempel 27 uten først å foreta omfattende beregninger for å etablere sannsynlighetsfordelingen. En annen anvendelse av setningen om dobbeltforventning er:

Eksempel 29 : Ventetid til første kron
La N = ventetiden til første kron i myntkast. En alternativ måte å beregne forventningen til N på, er som følger. La

$$\begin{aligned} X &= 1 \;\text{ dersom første kast er kron} \\ &= 0 \;\text{ dersom første kast er mynt.} \end{aligned}$$

Da blir

$$\begin{aligned} EN &= E(N \mid X=0)P(X=0) + E(N \mid X=1)P(X=1) \\ &= \frac{1}{2}(E(N \mid X=0) + E(N \mid X=1)). \end{aligned}$$

Men nå er

$$\begin{aligned} E(N \mid X=0) &= 1 + EN \\ E(N \mid X=1) &= 1 \end{aligned}$$

fordi forventet ventetid til kron, gitt første kast mynt, må være lik en (det første kastet) pluss forventet ventetid fra dette kast til første kron, som er EN. Vi får derfor

$$EN = \frac{1}{2}(1 + EN + 1) = 1 + \frac{1}{2}EN$$

Løses denne ligning med hensyn på EN ser vi at $EN = 2$.

Vi vil nå se på et interessant aspekt ved begrepene forventning og varians, nemlig *prediksjon med kvadratisk tapsfunksjon:*

Et eksperiment skal utføres. Eksperimentet medfører bl. a. at en stokastisk variabel Y antar en bestemt verdi y. Anta at sannsynlighetsfordelingen $P(Y = y)$ til Y er kjent, slik at EY og $varY$ også er kjent. La $\mu = EY$. La oss tenke oss at vi, før eksperimentet blir utført, blir bedt om å tippe hvilken verdi på Y vi kommer til å observere. La oss anta kvadratisk tap, dvs. dersom vi tipper verdien z og observerer verdien y, så taper vi beløpet $(y - z)^2$. Før eksperimentet er utført kan vi oppfatte tapet som en stokastisk variabel $T = (Y - z)^2$ og forventet tap blir

$$ET = E(Y - z)^2.$$

Vi ønsker å velge z slik at vi minimerer forventet tap. Det viser seg at da må vi velge $z = \mu$ og forventet tap er da lik $varY$. For å innse dette vil vi omforme uttrykket for forventet tap.

$$\begin{aligned} E(Y-z)^2 &= E(Y - \mu + \mu - z)^2 \\ &= E((Y-\mu)^2 + 2(Y-\mu)(\mu - z) + (\mu - z)^2) \\ &= E(Y-\mu)^2 + 2(\mu - z)E(Y - \mu) + E(\mu - z)^2. \end{aligned}$$

Vi har først trukket fra og lagt til μ, deretter kvadrert ut uttrykket og brukt regnereglene for forventning. I det siste uttrykket er det først leddet lik variansen, mens det midterste leddet er lik null. Vi får

$$E(Y - z)^2 = varY + (\mu - z)^2$$

og vi ser at dette er større enn eller lik $varY$, med likhet hvis og bare hvis z velges lik μ.

Anta nå at vi kan observere en stokastisk variabel X som gir delvis informasjon om Y, f.eks. ved at den simultane fordeling til X og Y, $P((X = x) \cap (Y = y))$ er kjent. Dersom vi får vite at $X = x$, vil det være den betingede fordeling for Y gitt $X = x$ som er relevant for oss når vi skal foreta prediksjonen. Med kvadratisk tap vil det være optimalt å predikere forventningen i denne betingede fordelingen, dvs. at vi gir prediksjonen $Z = E(Y \mid X = x)$. Forventet tap blir da variansen i den betingede fordelingen.

Eksempel 30 : Uendelig forventning
La X være en stokastisk variabel som antar verdier blant de naturlige tall. Anta at sannsynligheten for at X antar verdien x er omvendt proporsjonal med kvadratet av verdien x, dvs. at X har punktsannsynlighet av form

$$p(x) = P(X = x) = k \cdot \frac{1}{x^2} \quad x = 1, 2, \ldots$$

der k er en konstant slik at $\sum p(x) = 1$. [7])
Forventningen til X blir

$$EX = \sum_{x=1}^{\infty} xp(x) = k \sum_{x=1}^{\infty} x \frac{1}{x^2} = k \sum_{x=1}^{\infty} \frac{1}{x} = \infty$$

Summen representerer ikke noe endelig tall, den er uendelig.

Dette eksemplet har lært oss at dersom en stokastisk variabel X kan anta et uendelig antall verdier, er det mulig at forventningen til X ikke eksisterer som noe endelig tall. Dette betyr bl.a. at man i enkelte setninger for forventning må ta reservasjon om at de forventninger som inngår eksisterer, dersom setningen også skal gjelde i situasjoner med uendelig antall mulige verdier. Eksempelvis, for at setningen $E(X + Y) = EX + EY$ skal gjelde i slike situasjoner, må vi anta at EX og EY eksisterer, bl. a. for å unngå uttrykk av form $\infty - \infty$ som er en meningsløshet.
Vi går videre til

Eksempel 31 : Petersburg spillet
Et kasino tilbyr et spill hvor spilleren for en innsats på k kroner får knipse en mynt inntil første kron inntreffer. Dersom spilleren får første kron i n'te omgang utbetales 2^n kroner til spilleren. Lar vi N = ventetid til første kron, vil altså gevinsten være en stokastisk variabel

$$G = 2^N - k.$$

Siden sannsynlighetsfordelingen til N er gitt ved

$$p(x) = (\frac{1}{2})^n \quad n = 1, 2, \ldots$$

[7]Lesere med gode kunnskaper i matematisk analyse vil lett kunne finne ut at $k = 6/\pi^2$.

vil forventet gevinst bli (bruk formel F6)

$$
\begin{aligned}
EG &= \sum_{n=1}^{\infty}(2^n - k)p(n) = \sum_{n=1}^{\infty}(2^n - k)(\frac{1}{2})^n \\
&= -k + (1 + 1 + 1 + \cdots) = \infty
\end{aligned}
$$

Vi ser at forventningen til G ikke eksisterer som noe endelig tall. Dette innebærer bl. a. at det ikke finnes noen innsats k, samme hvor stor den er, som gjør dette spillet til rettferdig spill i den forstand vi studerte i Eksempel 13. ($EG = 0$). Analysen ovenfor forutsetter egentlig at kasinoet har uendelige ressurser (og spilleren uendelig med tid til rådighet). La oss tenke oss at spillets regler er slik at spilleren kan spille inntil 20 omganger for innsatsen. Det meste kasinoet risikerer å utbetale etter et spill er derfor 2^{20}=1 048 476 kroner. Nå er

$$
G = \begin{cases} 2^N - k & \text{for } N \leq 20 \\ 2^{20} - k & \text{for } N > 20 \end{cases}
$$

Forventet gevinst blir nå isteden

$$
\begin{aligned}
EG &= \sum_{n=1}^{20}(2^n - k)p(n) + (2^{20} - k)P(N > 20) \\
&= \sum_{n=1}^{20}(2^n - k)(\frac{1}{2})^n + (2^{20} - k)(1 - \sum_{n=1}^{20}(\frac{1}{2})^n) \\
&= -k + 21
\end{aligned}
$$

slik at 21 kroner vil være den innsats som gjør spillet rettferdig. I et kasino med store ressurser er det altså mulig å etablere rettferdige spill med mulighet for å gjøre en formue selv med en beskjeden innsats, men merk at sannsynligheten for moderat (selv positiv) gevinst blir relativt liten. Dette og andre eksempler viser at forventningsverdi ikke i alle situasjoner kan brukes som et mål for verdien av et spill for en spiller, evt. verdien av et investeringsobjekt for en investor. For eksempel vil de færreste være villige til å satse et stort beløp dersom sjansen for å mangedoble dette er liten, f. eks. investere 10 000 kroner med sjanse på 1/10 000 for å få tilbake 1 milliard selv om forventet gevinst her er 90 000 kroner.

5.9 Oppgaver

1. En modell og to stokastiske variable X og Y er gitt i tabellen

u	u_1	u_2	u_3	u_4	u_5	u_6
$P(u)$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{2}{8}$	$\frac{2}{8}$	$\frac{1}{8}$	$\frac{1}{8}$
$X(u)$	0	0	1	2	3	4
$Y(u)$	1	2	-1	1	-2	0

Finn sannsynlighetsfordelingene til henholdsvis X og Y. Finn også sannsynlighetsfordelingene til følgende stokastiske variable $X+Y, X-Y, X^2, Y^2$ og X^2+Y^2.

2. En modell og to stokastiske variable X og Y er gitt i tabellen

u	u_1	u_2	u_3	u_4	u_5
$P(u)$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{2}{8}$	$\frac{3}{8}$
$X(u)$	1	2	2	1	4
$Y(u)$	4	4	4	2	1

Vis at X og Y har samme sannsynlighetsfordeling. Følger det herav at $S = X+Y$ og $V = X+X = 2X$ har samme sannsynlighetsfordeling?

3. Regn ut forventningen til de stokastiske variable som er omtalt i Oppgave 1 og 2, både ut fra definisjonen av forventning og ut fra de respektive sannsynlighetsfordelinger. Der det er mulig illustrerer også hvordan de etablerte regneregler for forventning kan nyttes for å forenkle regnearbeidet.

4. Tegn histogrammer for de stokastiske variable omtalt i Oppgave 1 og 2.

5. Kast en terning 100 ganger og observer antall øyne i hvert kast, og beregn gjennomsnittlig antall øyne i de 100 kastene. Sammenlikn dette med forventet antall øyne i et kast 7/2.

6. For en falsk terning har man funnet ut at det er rimelig å sette sannsynligheten for 1, 2, 3, 4, 5 og 6 øyne lik henholdsvis 0.19, 0.17, 0.17, 0.17, 0.17 og 0.13. (se Oppgave 2.8). La X være antall øyne i et kast. Beregn EX og $varX$.

7. La antall øyne i første og annet kast med terningen i forrige oppgave være henholdsvis X og Y, og la $S = X+Y$ være sum øyne. Finn sannsynlighetsfordelingen til S og beregn ut fra denne ES og $varS$. Kunne disse størrelsene ha vært funnet på en enklere måte?

8. Det utføres tre myntkast med en rettferdig mynt. La X være antall kron. Finn sannsynlighetsfordelingen til X, og beregn EX og $varX$.

9. La situasjonen være som i Oppgave 3.15, og la X være antall Bordauxviner som en deltaker får smake. Finn sannsynlighetsfordelingen til X og beregn EX.

10. I Oppgave 3.18 hva blir forventet antall rette?

11. La situasjonen være som i Oppgave 3.19, og la X være antall defekte i utvalget. Finn sannsynlighetsfordelingen til X og beregn EX.

12. La situasjonen være som i Oppgave 3.21, og la X være antall funksjonærer i utvalget. Finn sannsynlighetsfordelingen til X og beregn EX.

13. La situasjonen være som i Oppgave 4.25, og la X være antall ganger en ny arbeider lykkes i løpet av tre forsøk. Finn sannsynlighetsfordelingen til X og beregn EX. Besvar denne oppgaven også under forutsetningene i Oppgave 4.26.

14. La situasjonen være som i Oppgave 4.29, og la X være ventetiden til første justering. Finn sannsynlighetsfordelingen til X og beregn EX. La Y være antall ekstra justeringer i løpet av uken. Finn sannsynlighetsfordelingen til Y og beregn EY. Besvar denne oppgaven også under forutsetningene i Oppgave 4.30.

15. La situasjonen være som i Oppgave 4.31, og la X være antall kvinner på kurs. Finn sannsynlighetsfordelingen til X og beregn EX. La Y være antall menn på kurs, og finn EY. Løs også oppgaven under de alternative trekningsmåtene som er angitt.

16. La situasjonen være som i Oppgave 4.34, og la X være antall rette tippetegn. Finn EX og $varX$ under begge forutsetninger som er angitt.

17. En pakning inneholder 10 artikler, hvorav 3 er defekte. Det trekkes tilfeldig et utvalg på 4 artikler. La X være antall defekte artikler i utvalget. Finn sannsynlighetsfordelingen til X. Beregn EX og $varX$.

18. En produksjonsprosess gir defekt artikkel med sannsynlighet $p = 0.05$. Det produseres 4 artikler. Anta uavhengighet. La X være antall defekte artikler. Finn sannsynlighetsfordelingen til X. Beregn EX og $varX$.

19. En pakning inneholder 5 artikler hvorav en er defekt. For å finne ut hvilken som er defekt trekkes tilfeldig artikler fra pakningen, og disse blir undersøkt en etter en. La X være antall artikler som må undersøkes. Finn sannsynlighetsfordelingen til X. Beregn EX og $varX$.

20. $\star$Det utføres gjentatte terningkast med en rettferdig terning. La X være antall kast som er nødvendig for at sum øyne skal være minst 6. Finn sannsynlighetsfordelingen til X. Beregn EX og $varX$.

21. I et spill er innsatsen 1 krone hvoretter spilleren får rett til 3 terningkast. Dersom spilleren ikke får noen sekser, er innsatsen tapt. Dersom han får en sekser, får han 2 kroner tilbake (innsatsen + en krone), ved to seksere får han 4 kroner tilbake og ved tre seksere får han 8 kroner. Er dette et rettferdig spill? Hvis ikke hvor stort beløp må spilleren få tilbake ved tre seksere for å gjøre spillet rettferdig.

22. La X være antall kroner i tre myntkast. To spill er tilbudt

(a) gir utbetaling $Y = 2X - 2$

(b) gir utbetaling $Z = (X-1)^2$.

Hvilket spill vil du foretrekke?
Lønner det seg å spille dersom innsatsen er 2 kroner?

23. Du triller terning og kan etter hvert kast velge om du vil stoppe og innkassere et kronebeløp lik sum øyne hittil. Haken er at straks du får en ener mister du alt. Når lønner det seg å stoppe? Hva er en rettferdig innsats for å få spille?

24. En bedrift har 8 ansatte som utfører et arbeid med en viss risiko for skade, og bedriften ønsker å tegne en forsikring på kr 100 000 pr ansatt. Sannsynligheten for at en arbeider blir skadet i løpet av et år vurderes til 0.0001. Hva er en "rettferdig" årspremie for bedriften (vi ser bort fra omkostninger, fortjeneste etc.).

25. La $Y \geq X$, dvs. $Y(u) \geq X(u)$ for alle u. Vis at da er $EY \geq EX$.

26. La X være stokastisk variabel som antar verdiene $1, 2, \ldots N$, hver med sannsynlighet $1/N$. Finn EX og $varX$.

27. La $X_1, X_2, \ldots X_n$ være uavhengige stokastiske variable, alle med samme forventning μ og samme varians σ^2. La

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Vis at $E(\overline{X}) = \mu$ og $var(\overline{X}) = \sigma^2/n$

28. Følgende gir god øvelse i å regne med forventning, varians og kovarians :

(a) Bruk F3, F4 og F5 til å vise at

$$E(a + bX) = a + bEX.$$

Bruk F6 til å gi et alternativt bevis for denne formelen.

(b) Bruk V6 og V7 til å vise at

$$var(a + bX) = b^2 varX.$$

(c) Vis at standardavviket har egenskapen

$$\sigma(a + bX) = \mid b \mid \sigma(X).$$

der $\mid b \mid$ betegner tallverdien av b.

(d) Vis at kovarians har egenskapen

$$cov(a + bX, c + dY) = b \cdot d \cdot cov(X, Y).$$

29. (a) Bruk F10 til å vise at

$$E(bX + dY) = bEX + dEY.$$

(b) Vis at når X og Y er uavhengige så er

$$var(bX + dY) = b^2 varX + d^2 varY.$$

Er formelen $var(X - Y) = varX - varY$ korrekt?

30. (a) Vis at korrelasjonskoeffisienten $\rho(X, Y)$ har egenskapen at for positive konstanter b og d er

$$\rho(a + bX, c + dY) = \rho(X, Y).$$

Hva skjer dersom enten b eller d (eller begge) er negative?

(b) Vis at korrelasjonskoeffisienten har egenskapen

$$-1 \leq \rho(X, Y) \leq +1.$$

Hint: Vis først at

$$var(bX + Y) = b^2 varX + 2bcov(X, Y) + varY.$$

Gjør bruk av denne formel med $b = -cov(X, Y)/varX$.

(c) Hva kan sies om X og Y dersom
(i) $\rho(X, Y) = 0$ (ii) $\rho(X, Y) = 1$ (iii) $\rho(X, Y) = -1$.

(d) Uttrykk formel V8 og V10 standardavvik og korrelasjonskoeffisienter.

31. La situasjonen være som i Eksempel 23.

(a) Finn sannsynlighetsfordelingen til samlet solgt kvantum S i hvert av tilfellene med positiv og negativ samvariasjon.

(b) Beregn variansen til S i begge tilfeller direkte ut fra (a) og ved V8..

(c) Sammenlign (a) og (b) med tilfellet uavhengighet (Eksempel 21).

(d) Beregn korrelasjonskoeffisienten mellom X og Y i de tre tilfellene uavhengighet, positiv og negativ samvariasjon.

32. Finn følgende korrelasjonskoeffisienter

(a) mellom X og Y i Oppgave 1.

(b) mellom X og Y i Oppgave 2.

(c) mellom $X + Y$ og $X - Y$ i Oppgave 1.

33. Avkastingen pr. investert krone i hvert av to investeringsobjekter antas å være stokastiske variable med forventning og standardavvik gitt ved

	Forventning	Standardavvik
Investering nr. 1 ...	0.10	0.08
Investering nr. 2 ...	0.15	0.12

En investor har bestemt seg for å plassere halvparten av sin kapital i hvert av objektene. Finn forventning og standardavvik til avkasting pr. investert krone dersom korrelasjonskoeffisienten mellom avkastingene i hvert av objektene er

(i) 0 (ii) 0.5 (iii) -0.5

Formuler problemstillingen generelt og diskuter mulige konsekvenser for investering under usikkerhet.

34. Et firma har to bilselgere A og B. Antall solgte biler pr. uke for de to antas å være uavhengige stokastiske variable X og Y med sannsynlighetsfordeling henholdsvis

x	0	1	2
$P(X=x)$	0.3	0.4	0.3

y	0	1	2
$P(Y=y)$	0.5	0.4	0.1

(a) Finn den simultane fordeling til X og Y.

(b) Finn sannsynligheten for at de en uke selger like mange biler.

(c) Finn sannsynlighetsfordelingen til $S = X + Y$.

(d) Gitt at det en uke ble solgt to biler. Hva er sannsynligheten for at A solgte begge?

35. (a) Vis formel V10 i teksten, dvs.

$$var(\sum_{i=1}^{n} X_i) = \sum_{i=1}^{n} var(X_i) + \sum_{i \neq j} cov(X_i, X_j)$$

(b) Vis at

$$cov(\sum_{i=1}^{n} X_i, \sum_{j=1}^{m} Y_j) = \sum_{i=1}^{n} \sum_{j=1}^{m} cov(X_i, Y_j)$$

36. $\star$(a) Kan X og X^2 være uavhengige?

(b) Kan X og X^2 være ukorrelerte?

(c) Vis at dersom X og Y er uavhengige så er også $U = f(X)$ og $V = g(Y)$ uavhengige.

37. (a) Vis at betingede sannsynlighetsfordelinger har samme egenskaper som vanlige sannsynlighetsfordelinger (se Kapittel 5.2).

(b) Vis at X og Y er uavhengige hvis og bare hvis den betingede fordeling for Y gitt $X = x$ er lik den marginale fordeling for Y.

38. Finn den kumulative fordeling til følgende stokastiske variable

    (a) Antall øyne X i et myntkast.

    (b) Antall kron X i to myntkast.

    (c) Sum øyne S i to terningkast.

    (d) De stokastiske variable X og Y i Oppgave 1.

    (e) De stokastiske variable X, Y, S og V i Oppgave 2.

    Tegn figurer for å illustrere disse kumulative fordelingene.

39. $\star$Fordelingen til en stokastisk variabel X sies å være *symmetrisk* om a dersom

    $$P(X = a + x) = P(X = a - x) \text{ for alle } x.$$

    Vis at dersom dette er tilfelle, så må

    $$EX = a \quad \text{(såframt forventningen eksisterer).}$$

    Hint: Vis først at $X - a$ og $a - X$ må ha samme fordeling og følgelig samme forventning.

40. $\star$(a) Vis at $E(X^2) \geq (EX)^2$.

    (b) La X ha fordelingen

x	a/A	b/B
$P(X = x)$	$A/(A + B)$	$B/(A + B)$

    der a, b, A og B er positive tall.
    Bruk resultatet i (a) til å vise ulikheten

    $$\frac{(a + b)^2}{A + B} \leq \frac{a^2}{A} + \frac{b^2}{B}$$

    Denne ulikheten kan generaliseres til

    $$\frac{(a + b + c + \ldots)^2}{A + B + C + \ldots} \leq \frac{a^2}{A} + \frac{b^2}{B} + \frac{c^2}{C} + \ldots$$

    og har flere anvendelser, bl.a. i Kapittel 14.

41. $\star$La X_1, X_2 og X_3 være uavhengige med samme sannsynlighetsfordeling $p(x)$.

    (a) Finn sannsynlighetsfordelingen til $S = X_1 + X_2$ dersom $p(x)$ er gitt ved

    (i)

x	0	1	2
$P(X = x)$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$

    (ii)

x	0	1	2
$P(X = x)$	$\frac{2}{5}$	$\frac{1}{5}$	$\frac{2}{5}$

    (iii)

x	0	1	2
$P(X = x)$	$\frac{3}{6}$	$\frac{2}{6}$	$\frac{1}{6}$

Hint: Gå fram som i Eksempel 25.

(b) Finn sannsynlighetsfordelingen til $T = X_1 + X_2 + X_3$ i tilfellene (i), (ii) og (iii).

(c) Tegn histogrammer og kommenter disse (se også Kapittel 6.5 om normaltilnærming).

42. (a) Kast en terning gjentatte ganger og observer antall øyne. Tegn et empirisk histogram etter
(i) $n = 20$ kast (ii) $n = 100$ kast.

(b) Kast to terninger gjentatte ganger og observer sum øyne. Tegn et empirisk histogram etter
(i) $n = 20$ kast (ii) $n = 100$ kast.

(c) Kast tre terninger gjentatte ganger og observer sum øyne. Tegn et empirisk histogram etter
(i) $n = 20$ kast (ii) $n = 100$ kast. Kommenter resultatene.

Denne oppgaven gjøres enklest ved å simulere terningkast på PC.

Kapittel 6

Noen sannsynlighetsfordelinger

Dette kapitlet handler om sentrale diskrete sannsynlighetsfordelinger som dukker opp i mange ulike situasjoner. Vi legger vekt på å beskrive de forutsetninger som gir opphav til fordelingene, samt studere de viktigste egenskapene, bl.a. forventning og varians. Vi vil oppøve bruk av hjelpemidler for beregning av sannsynligheter, herunder tabeller og tilnærmingsmetoder. Spesielt viktig er tilnærmingen til normalfordelingen. [1]

6.1 Indikatorfordelingen

La I være en indikatorvariabel, dvs. I kan anta verdiene 0 eller 1. (Se Eksempel 5.4). Lar vi p betegne sannsynligheten for at I antar verdien 1, ser sannsynlighetsfordelingen slik ut

i	0	1
$P(I=i)$	$1-p$	p

Her blir

$$E(I) = 0 \cdot (1-p) + 1 \cdot p = p$$

En indikatorvariabel har spesielt den egenskapen at $I^2 = I$. Følgelig får vi

$$\begin{aligned} var(I) &= E(I^2) - (EI)^2 = EI - (EI)^2 \\ &= EI(1-EI) = p(1-p) \end{aligned}$$

[1] Dette er en såkalt kontinuerlig fordeling. I Kapittel 8.4 gis en kort omtale av egenskaper ved slike fordelinger.

Indikatorvariable er i første rekke et hjelpemiddel i studiet av andre viktige sannsynlighetsfordelinger, spesielt i forbindelse med beregning av forventning og varians. La oss illustrere denne mulighet med et eksempel.

Eksempel 1 : Ulike øyne i terningkast
n terninger trilles. Hva blir forventet antall forskjellige øyne som terningene viser? Definer i alt seks indikatorvariable $I_1, I_2, \ldots, I_6$ hvor

$$\begin{aligned} I_j &= 1 \text{ hvis minst en terning viser } j \text{ øyne} \\ &= 0 \text{ hvis ingen terning viser } j \text{ øyne} \end{aligned}$$

Antall forskjellige øyne X kan da skrives

$$X = I_1 + I_2 + \cdots + I_6$$

Nå blir $EI_j = P(I_j = 1) = 1 - P(I_j = 0) = 1 - (5/6)^n$ for alle j, fordi sannsynligheten for at en terning ikke viser j øyne er $5/6$ og de ulike terningene representerer uavhengige kast. Vi får

$$\begin{aligned} EX &= EI_1 + EI_2 + \cdots + EI_6 \\ &= 6 \cdot (1 - (5/6)^n) \end{aligned}$$

For $n = 3$ gir dette $EX = 2.53$, for $n = 6$ blir $EX = 3.99$, mens for $n = 12$ blir $EX = 5.33$. Her ville en direkte beregning av EX ved først å finne sannsynlighetsfordelingen til X bli nokså brysom.

6.2 Den binomiske fordeling

La oss følge opp de tre forutsetningene for en binomisk forsøksrekke fra Kapittel 4.6 :

Binomisk situasjon : Det utføres n binomiske forsøk, dvs.

1. Hvert forsøk gir enten suksess eller fiasko.
2. Sannsynligheten p for suksess er den samme i hvert forsøk.
3. Forsøkene er uavhengige.

Antall suksesser X i løpet av de n forsøkene observeres.

Vi er interessert i sannsynlighetsfordelingen til X. Denne ble faktisk allerede utledet i Kapittel 4.6, og kalles *den binomiske fordeling.* [2]

Binomisk fordeling : Dersom X har fordelingsfunksjon

$$p(x) = P(X = x) = \binom{n}{x} p^x (1-p)^{n-x} \;;\quad x = 0, 1, \ldots, n,$$

sier vi at X er *binomisk fordelt* med parametre (n,p).

Vi har behov for enkle regneformler for forventning og varians til binomiske variable. Disse er

Dersom X er binomisk fordelt (n,p), har vi at

$$EX = n \cdot p \quad varX = n \cdot p \cdot (1-p)$$

La oss prøve å begrunne formlene. Ved bruk av den vanlige regneformelen for forventning får vi

[2] Merk at siden dette er en sannsynlighetsfordeling må

$$1 = \sum_x p(x) = \sum_{x=0}^{n} \binom{n}{x} p^x (1-p)^{n-x}$$

som også kan bekreftes ved å anvende den såkalte Newtons binomialformel på uttrykket $(p + (1-p))^n$ (se Oppgave 3.29).

$$EX = \sum_x xp(x) = \sum_{x=0}^{n} x \binom{n}{x} p^x (1-p)^{n-x}$$

Dette uttrykk lar seg ikke lett forenkle, og vi forsøker derfor en annen angrepsmåte. Definer for $j = 1, 2, \ldots, n$

$$\begin{aligned} I_j &= 1 \text{ dersom j'te forsøk er suksess} \\ &= 0 \text{ dersom j'te forsøk er fiasko} \end{aligned}$$

Vi ser at vi kan skrive

$$X = I_1 + I_2 + \cdots + I_n$$

Siden forsøkene er uavhengige, blir $I_1, I_2, \ldots, I_n$ uavhengige stokastiske variable, og siden hver I_j er en indikatorvariabel, blir

$$EI_j = P(I_j = 1) = p \quad \text{og} \quad varI_j = p(1-p)$$

Ved å bruke kjent setning om forventning får vi

$$\begin{aligned} EX &= E(I_1 + I_2 + \cdots + I_n) \\ &= EI_1 + EI_2 + \cdots + EI_n \\ &= p + p + \cdots + p = np \end{aligned}$$

Videre blir på grunn av uavhengigheten

$$\begin{aligned} varX &= var(I_1 + I_2 + \cdots + I_n) \\ &= varI_1 + varI_2 + \cdots + varI_n \\ &= p(1-p) + p(1-p) + \cdots + p(1-p) = np(1-p) \end{aligned}$$

Som en illustrasjon anvender vi teorien ovenfor på eksemplene 15, 16 og 17 i Kapittel 4.6. Disse dreide seg om henholdsvis myntkast, terningkast og en produksjonsprosess:

$X =$ antall kron i 10 myntkast er binomisk fordelt $(n = 10, p = \frac{1}{2})$:

$$P(X = x) = \binom{10}{x} (\frac{1}{2})^x (1 - \frac{1}{2})^{10-x}$$

Vi får $EX = n \cdot p = 10 \cdot \frac{1}{2} = 5$ og $varX = np(1-p) = 10 \cdot \frac{1}{2} \cdot \frac{1}{2} = \frac{5}{2}$.

X = antall seksere i 10 terningkast er binomisk fordelt $(n = 10, p = 1/6)$:

$$P(X = x) = \binom{10}{x}(\frac{1}{6})^x(1 - \frac{1}{6})^{10-x}$$

Vi får $EX = n \cdot p = 10 \cdot \frac{1}{6} = \frac{5}{3}$ og $varX = np(1-p) = 10 \cdot \frac{1}{6} \cdot \frac{5}{6} = \frac{25}{18}$.

X = antall defekte blant 100 produserte artikler er, med defektprosent på 5, binomisk fordelt $(n = 100, p = 0.05)$:

$$P(X = x) = \binom{100}{x}0.05^x(1 - 0.05)^{100-x}$$

Vi får $EX = 100 \cdot 0.05 = 5$ og $varX = 100 \cdot 0.05 \cdot 0.95 = 4.75$.

Til hjelp ved beregning av binomiske sannsynligheter er det utarbeidet tabeller. I Appendiks C har vi Tabell C.2 som gir oss $P(X = x)$ når X er binomisk fordelt (n, p) for n-verdier opp til 10 og de fem p-verdiene $0.05, 0.1, 0.2, 0.3$ og 0.4 samt Tabell for $p = 0.5$ og n helt opp til 30. For p-verdier og n-verdier som ikke dekkes av disse tabellene, må vi ofte ty til større tabellverker. Disse gir som regel de såkalte kumulative sannsynligheter $P(X \leq x)$ istedenfor $P(X = x)$. Det finnes slike tabeller for n helt opp til 1000 og p-verdier i sprang på 0.01. For store verdier av n vil vi i Kapittel 6.5 utvikle en metode for tilnærmet beregning av $P(X \leq x)$.

Eksempel 2 : Eksamen

En eksamen består av åtte spørsmål hvor en i hvert spørsmål har valget mellom fem svar hvorav ett er riktig. For å bestå eksamen kreves minst tre riktige svar. Hva er sannsynligheten for at en elev som er helt uforberedt, og som velger sine svar tilfeldig, ikke består eksamen? Vi betrakter de åtte spørsmålene som åtte uavhengige forsøk, hvor sannsynligheten for suksess (dvs. riktig svar) er $p = 1/5 = 0.2$ i hvert forsøk. Lar vi X betegne antall riktige svar, blir X binomisk fordelt $(n = 8, p = 0.2)$:

$$P(X = x) = \binom{8}{x}(\frac{1}{5})^x(1 - \frac{1}{5})^{8-x}$$

Av Tabell C.2 med $n = 8$ og $p = 0.2$ finner vi

x	0	1	2	3	4	5	6	7	8	sum
$P(X = x)$	0.1678	0.3355	0.2936	0.1468	0.0459	0.0092	0.0011	0.0001	0.0000	1.0000

Sannsynlighetene er her gitt med fire desimalers nøyaktighet, eksempelvis er sannsynligheten for åtte riktige svar positiv, men så liten at det ikke syns i fjerde desimal. Vi får nå at sannsynligheten for å stryke blir

$$\begin{aligned} P(X \leq 2) &= P(X = 0) + P(X = 1) + P(X = 2) \\ &= 0.1678 + 0.3355 + 0.2936 = 0.7969 \end{aligned}$$

Sannsynligheten for å bestå finnes enklest slik:

$$P(X \geq 3) = 1 - P(X \leq 2) = 1 - 0.7969 = 0.2031$$

men kan også finnes ved å summere i tabellen. Merk at $EX = np = 8 \cdot 0.2 = 1.6$ og $varX = np(1-p) = 1.28$. Dette kan selvsagt også (mer tungvint) finnes ved direkte beregning ut fra tabellen ovenfor.

Eksempel 3 : Ti myntkast
Lar vi X = antall kron i ti myntkast, ser vi av en binomisk tabell med $n = 10$ og $p = 0.5$ (Tabell C.3) at

x	0	1	2	3	4	5
$P(X = x)$	0.0010	0.0098	0.0439	0.1172	0.2051	0.2461

Her merker vi oss at tabellen ikke direkte gir sannsynlighetene for 6, 7, 8, 9 og 10 kron, men dette er heller ikke nødvendig. Lar vi X' = antall mynt i ti myntkast, er det klart at også X' er binomisk fordelt $(n = 10, p = 0.5)$. Vi får derfor

$$\begin{aligned} P(X = 6) = &\quad P(X' = 4) = \quad 0.2051 \\ P(X = 7) = &\quad P(X' = 3) = \quad 0.1172 \quad \text{osv.} \end{aligned}$$

slik at disse sannsynlighetene også er gitt ved tabellen.

Eksempel 4 : Levende mus
En biolog driver forsøk med levende mus, og vet av erfaring at gjennomsnittlig ca. 80% av musene overlever slike eksperimenter. Hvor mange mus må hun starte med dersom sannsynligheten skal være minst 0.98 for at 5 eller flere mus overlever eksperimentet. Anta hun starter med n mus, hver mus representerer et forsøk med to utfall, den overlever eller dør. Det er nå rimelig å anta at sannsynligheten for at en mus overlever er 0.80. Videre antar vi at død/overlever for de ulike mus er uavhengige begivenheter (en noe mer tvilsom antakelse). Med disse antakelsene blir antall mus som overlever X binomisk fordelt $(n, p = 0.8)$. Vi skal altså bestemme n slik at

$$P(X \geq 5) \geq 0.98$$

Vi trenger derfor $P(X = x)$ for $p = 0.8$ og ulike n. Nå ser vi at Tabell C.2 ikke dekker tilfellet $p = 0.8$. Vi skal imidlertid greie oss likevel : Betrakt isteden $X' =$ antall mus som dør. Siden $X' = n - X$, kan vi skrive
$P(X = x) = P(X' = n - x)$. Vi ser at $X' = n - X$ blir binomisk fordelt $(n, p = 0.2)$, og følgelig kan vi likevel finne $P(X = x)$ ved hjelp av denne tabellen. La oss forsøke om det er nok å starte med syv mus. Med $n = 7$ og $p = 0.2$ finner vi av Tabell C.2

x	0	1	2	3	4	5	6	7
$P(X = x)$	0.0000	0.0004	0.0043	0.0287	0.1147	0.2753	0.3670	0.2097

Vi ser at $P(X \geq 5) = 0.2753 + 0.3670 + 0.2097 = 0.8520$. Syv mus er altså for lite. På tilsvarende måte finner vi at med $n = 8$ blir $P(X \geq 5) = 0.9437$ som også er for lite. La oss derfor prøve $n = 9$. Vi finner av Tabell C.2 følgende sannsynligheter :

x	0	1	2	3	4	5	6	7	8	9
$P(X = x)$	0.0000	0.0000	0.0003	0.0028	0.0165	0.0661	0.1762	0.3020	0.3020	0.1342

Her blir $P(X \geq 5) = 0.9805$, dvs. ni mus er tilstrekkelig.

6.3 Den hypergeometriske fordeling

Betrakt følgende problemstilling :

Den hypergeometriske situasjon :
Gitt en populasjon bestående av N objekter hvorav

M	er spesielle	(f.eks. defekte)
$N - M$	er vanlige	(f.eks. intakte)

Fra populasjonen trekkes et tilfeldig uordnet utvalg på n elementer og antall spesielle elementer Y i utvalget noteres.

Vi ønsker å finne sannsynlighetsfordelingen til Y. Ifølge forutsetningen er alle $\binom{N}{n}$ mulige utvalg like sannsynlige. Antall gunstige utvalg for begivenheten at vi får y spesielle elementer i utvalget finnes slik: De y spesielle elementene i utvalget kan velges blant de M spesielle elementene i populasjonen på $\binom{M}{y}$ måter. For hver av disse $\binom{M}{y}$ måtene er det $\binom{N-M}{n-y}$ ulike måter å velge ut de $n-y$ vanlige elementene i utvalget. Følgelig er det $\binom{M}{y} \cdot \binom{N-M}{n-y}$ gunstige utvalg for begivenheten $(Y = y)$. Sannsynlighetsfordelingen til Y er derfor gitt i følgende definisjon :

Hypergeometrisk fordeling :
Dersom Y har fordelingsfunksjon

$$p(y) = P(Y = y) = \frac{\binom{M}{y} \cdot \binom{N-M}{n-y}}{\binom{N}{n}}$$

sier vi at Y er *hypergeometrisk fordelt* (N, M, n).

Vi har behov for enkle regneformler for forventning og varians til hypergeometrisk fordelte variable. Disse er

Dersom Y er hypergeometrisk fordelt (N, M, n), så er

$$EY = n \cdot \frac{M}{N} \quad varY = \frac{N-n}{N-1} \cdot n \cdot \frac{M}{N} \cdot (1 - \frac{M}{N})$$

La oss prøve å begrunne formlene. Ved bruk av den vanlige regneformelen for forventning får vi

$$EY = \sum_y yp(y) = \sum_y y \frac{\binom{M}{y} \cdot \binom{N-M}{n-y}}{\binom{N}{n}}$$

men dette uttrykk lar seg ikke lett forenkle. La oss isteden prøve en annen angepsmåte, ved bruk av indikatorer: Vi vet at et tilfeldig (uordnet) utvalg kan oppnås ved å trekke et tilfeldig ordnet utvalg og så se bort fra rekkefølgen av elementene. La oss derfor definere

$$\begin{aligned} I_j &= 1 \text{ dersom det j'te element i utvalget er spesielt} \\ &= 0 \text{ dersom det j'te element i utvalget er vanlig} \end{aligned}$$

Vi kan da skrive Y som en sum av indikatorvariable

$$Y = I_1 + I_2 + \cdots + I_n$$

Ifølge ekvivalensloven for et tilfeldig ordnet utvalg, har hvert av de N elementene i populasjonen samme sannsynlighet for å komme på j'te plass i utvalget (se Kapittel 3.5). Herav følger at

$$EI_j = P(I_j = 1) = \frac{M}{N}$$

$$varI_j = EI_j(1 - EI_j) = \frac{M}{N}(1 - \frac{M}{N})$$

Ved å bruke kjent setning om forventning får vi

$$\begin{aligned} EY &= E(I_1 + I_2 + \cdots + I_n) \\ &= EI_1 + EI_2 + \cdots + EI_n \\ &= \frac{M}{N} + \frac{M}{N} + \cdots \frac{M}{N} = n \cdot \frac{M}{N} \end{aligned}$$

Når det gjelder variansen er problemet ikke så enkelt fordi $I_1, I_2, \ldots I_n$ ikke er uavhengige stokastiske variable, og formelen om at variansen til en sum er lik summen av variansen kan derfor ikke nyttes. Vi må trekke inn kovarianser, og de som er interessert i å følge den videre argumentasjon, finner den som $\star$ -merket stoff ved slutten av avsnittet.

La oss belyse forholdet mellom den hypergeometriske fordeling og den binomiske fordeling :

Dersom de n trekningene isteden skjer med tilbakelegging, er det klart at vi har en binomisk situasjon: Hver trekning gir enten et spesielt eller vanlig objekt, sannsynligheten for et spesielt objekt er M/N i hver trekning og trekningene er uavhengige. I dette tilfellet blir Y binomisk fordelt $(n, p = M/N)$, og da blir $EY = n \cdot M/N$ og $varY = n \cdot M/N \cdot (1 - M/N)$. Sammenligner vi dette med resultatene ovenfor, ser vi at forventningen er den samme i de to situasjonene. Variansen er forskjellig, men den eneste forskjellen er faktoren $(N - n)/(N - 1)$. Denne kan oppfattes som en korreksjonsfaktor som må være med i den hypergeometriske situasjon når trekningene skjer uten tilbakelegging. For eksempel dersom $n = N$, dvs. alle elementene i populasjonen trekkes ut, er det klart at $Y = M$, dvs. konstant og da blir $varY = 0$. Vi ser at dette stemmer med formelen for $varY$ ovenfor, idet korreksjonsfaktoren $(N - n)/(N - 1)$ nå blir lik null. I en situasjon hvor utvalget på n er lite i forhold til populasjonen N, er det klart at det spiller liten rolle om trekningene skjer med eller uten tilbakelegging, det er liten sjanse for å trekke samme objekt to eller flere ganger. Ved trekning uten tilbakelegging blir trekningene "tilnærmet uavhengige", og vi kan konkludere:

Binomisk tilnærming :
Dersom Y er hypergeometrisk fordelt (N, M, n), og n er liten i forhold til N blir

$$Y \text{ tilnærmet binomisk fordelt } (n, \tfrac{M}{N})$$

Dette resultatet kan ofte nyttes ved tilnærmet beregning av sannsynligheter i forbindelse med den hypergeometriske situasjon, når tabeller over den hypergeometriske fordeling ikke lenger strekker til.

La oss se på noen situasjoner som naturlig leder til en hypergeometrisk fordeling:

Eksempel 5 : Delegasjonen
I en forening med 10 medlemmer er det 6 menn og 4 kvinner. En delegasjon på 4 medlemmer velges ut ved loddtrekning. La Y = antall kvinner i delegasjonen. Dersom vi antar at alle $\binom{10}{4}$ mulige delegasjoner er like sannsynlige, følger at Y blir hypergeometrisk fordelt $(N = 10, M = 4, n = 4)$, dvs.

$$P(Y = y) = \frac{\binom{4}{y} \cdot \binom{6}{4-y}}{\binom{10}{4}} \;;\;\; y = 0, 1, 2, 3, 4.$$

Her blir $EY = n \cdot \frac{M}{N} = 4 \cdot \frac{4}{10} = 1.6$, mens

$$varY = \frac{N-n}{N-1} \cdot n \cdot \frac{M}{N} \cdot (1 - \frac{M}{N}) = \frac{6}{9} \cdot 4 \cdot \frac{4}{10} \cdot (1 - \frac{4}{10}) = 0.64$$

Ved utregning av de binomiske koeffisientene eller bruk av tabell, får vi (Se også Eksempel 3.16)

y	0	1	2	3	4
$P(Y = y)$	0.07143	0.38095	0.42857	0.11429	0.00476

Leseren kan selv beregne EY og $varY$ ut fra tabellen, og sjekke at resultatet ovenfor stemmer.

Eksempel 6 : Delegasjonen
Anta at foreningen isteden har 100 medlemmer hvorav 40 er kvinner og 60 er menn. En delegasjon på 4 medlemmer trekkes ut tilfeldig. Lar vi $Y =$ antall kvinner i delegasjonen, får vi at Y er hypergeometrisk fordelt $(N = 100, M = 40, n = 4)$, dvs.

$$P(Y = y) = \frac{\binom{40}{y} \cdot \binom{60}{4-y}}{\binom{100}{4}} \; ; \;\; y = 0, 1, 2, 3, 4.$$

Her blir $EY = n \cdot \frac{M}{N} = 4 \cdot \frac{40}{100} = 1.6$, som i forrige eksempel, mens $varY = \frac{N-n}{N-1} n \cdot \frac{M}{N} \cdot (1 - \frac{M}{N}) = 0.92$, dvs. noe større enn i forrige eksempel. Beregning av sannsynlighetene ovenfor blir noe tungvint, og det er derfor laget tabeller for hypergeometriske sannsynligheter for ulike N,M og n. Fra en slik tabell henter vi for tilfellet $(N = 100, M = 40, n = 4)$.

y	0	1	2	3	4
$P(Y = y)$	0.12436	0.34907	0.35208	0.15118	0.02331

Leseren kan sjekke EY og $varY$ ved beregning ut fra tabellen. For å dekke alle mulige verdier av N,M og n for N helt opp til 100, blir en slik tabell raskt en uhåndterlig bok på 500 sider eller så. I mange situasjoner vil det være tilstrekkelig med en tilnærmet beregning av hypergeometriske sannsynligheter. Det kan vi gjøre her: Siden $n = 4$ er liten i forhold til $N = 100$ kan vi, av det som er sagt ovenfor, slutte at Y er tilnærmet binomisk fordelt $(n = 4, p = M/N = 0.4)$. Av en tabell over binomiske sannsynligheter finner vi

y	0	1	2	3	4
$P(Y = y) \approx$	0.1296	0.3456	0.3456	0.1536	0.0256

Om dette kan betraktes som en brukbar tilnærmelse, vil avhenge av formålet med beregningen. Anta at det isteden var 1000 medlemmer hvorav 400 kvinner i foreningen. Antall kvinner i en delegasjon på $n = 4$ blir nå hypergeometrisk fordelt $(N = 1000, M = 400, n = 4)$, og følgelig tilnærmet binomisk fordelt $(n = 4, p = M/N = 0.4)$, dvs. samme tilnærmede fordeling som ovenfor. I denne situasjonen blir tilnærmelsen betydelig bedre enn ovenfor, det viser seg at tallene er korrekte inntil tredje desimal.

$\star$ Variansen til Y kan beregnes ved hjelp av formelen V10.

$$varY \;\; = \;\; var(\sum_{j=1}^{n} I_j)$$

$$= \sum_{j=1}^{n} varI_j + \sum_{i \neq j} cov(I_i, I_j)$$

Det er nå lett å vise at for alle $i \neq j$ blir

$$E(I_i \cdot I_j) = P((I_i = 1) \cap (I_j = 1)) = \frac{M(M-1)}{N(N-1)}$$

og følgelig blir

$$\begin{aligned} cov(I_i, I_j) &= E(I_i \cdot I_j) - EI_i \cdot EI_j \\ &= \frac{M(M-1)}{N(N-1)} - (\frac{M}{N})^2 \\ &= -\frac{1}{N-1} \cdot \frac{M}{N} \cdot (1 - \frac{M}{N}) \end{aligned}$$

Siden vi i alt har $n(n-1)$ kovariansledd i summen ovenfor, får vi

$$\begin{aligned} varY &= n \cdot \frac{M}{N}(1 - \frac{M}{N}) + n \cdot (n-1) \cdot (-\frac{1}{N-1} \cdot \frac{M}{N} \cdot (1 - \frac{M}{N})) \\ &= \frac{N-n}{N-1} \cdot \frac{M}{N} \cdot (1 - \frac{M}{N}) \end{aligned}$$

6.4 Poissonfordelingen

Følgende sannsynlighetsfordeling er viktig for en rekke anvendelser:

Definisjon : Dersom X har fordelingsfunksjon gitt ved

$$p(x) = P(X = x) = \frac{\lambda^x}{x!} e^{-\lambda} \; ; \quad x = 0, 1, 2, \ldots$$

sier vi at X er *Poissonfordelt* med parameter $\lambda (\lambda > 0)$. [a] [b]

[a] Oppkalt etter matematikeren Poisson (1781-1840).

[b] Her er $e = 2.7183...$ grunntallet i det naturlige logaritmesystem.

En del situasjoner hvor Poissonfordelingen dukker opp vil bli omtalt senere i dette avsnittet. La oss først studere noen egenskaper ved denne fordelingen: For det første merker vi oss at dette er et eksempel på en fordeling definert over en (tellbar) uendelig mengde av verdier. At det virkelig er en fordeling ser vi ved å summere $p(x)$ over disse verdiene: [3]

$$\sum_{x=0}^{\infty} p(x) = \sum_{x=0}^{\infty} \frac{\lambda^x}{x!} e^{-\lambda} = e^{-\lambda} \sum_{x=0}^{\infty} \frac{\lambda^x}{x!} = e^{-\lambda} \cdot e^{\lambda} = 1$$

Vi har følgende "regneformler", som viser at parameteren λ kan tolkes som forventningen (og variansen) i fordelingen :

> Dersom X er Poissonfordelt med parameter λ, så er
>
> $$EX = \lambda \quad varX = \lambda$$

La oss begrunne formelen for forventningen :

$$\begin{aligned} EX &= \sum_{x=0}^{\infty} xp(x) = \sum_{x=0}^{\infty} x \frac{\lambda^x}{x!} e^{-\lambda} = \sum_{x=1}^{\infty} \frac{\lambda^x}{(x-1)!} e^{-\lambda} \\ &= \lambda e^{-\lambda} \sum_{x=1}^{\infty} \frac{\lambda^{x-1}}{(x-1)!} = \lambda e^{-\lambda} \sum_{x=0}^{\infty} \frac{\lambda^x}{x!} = \lambda e^{-\lambda} e^{\lambda} = \lambda \end{aligned}$$

På tilsvarende måte kan det vises at også $varX = \lambda$.

La oss se på et par eksempler på Poissonfordelinger :

$\lambda = 1 : P(X = x) = \frac{1}{x!} e^{-1}$; $x = 0, 1, 2, \ldots$

x	0	1	2	3	4	$\cdots$
$P(X = x)$	e^{-1}	e^{-1}	$\frac{1}{2}e^{-1}$	$\frac{1}{6}e^{-1}$	$\frac{1}{24}e^{-1}$	$\cdots$

$\lambda = 2 : P(X = x) = \frac{2^x}{x!} e^{-2}$; $x = 0, 1, 2, \ldots$

x	0	1	2	3	4	$\cdots$
$P(X = x)$	e^{-2}	$2e^{-2}$	$2e^{-2}$	$\frac{4}{3}e^{-2}$	$\frac{2}{3}e^{-2}$	$\cdots$

[3] Her og nedenfor brukes følgende formel fra matematisk analyse:

$$e^{\lambda} = \sum_{x=0}^{\infty} \frac{\lambda^x}{x!} \quad \text{for alle reelle } \lambda$$

Poissonfordelinger for ulike verdier av λ finnes tabellert i de fleste statistiske tabellverker, som regel for λ-verdier i området fra 0.1 til 10.0, se Tabell C.4 i Appendiks C.

Poissonfordelingen blir ofte brukt som modell i situasjoner hvor det er tale om antall hendelser i et gitt tidsrom eller antall objekter i et gitt område.

Eksempel 7 : Reservedeler

Et firma holder en bestemt reservedel på lager, og har erfart at det siste halve året er blitt etterspurt gjennomsnittlig en pr. dag, men variasjonene fra dag til dag er betydelige. La oss anta at antall etterspurte artikler X på en gitt dag er Poissonfordelt med forventning $\lambda = 1$. Av Tabell C.4 finner vi :

x	0	1	2	3	4	5	6	7
$P(X = x)$	0.3679	0.3679	0.1839	0.0613	0.0153	0.0031	0.0005	0.0001

hvor sannsynligheten for 8 eller flere etterspurte deler er så liten at det “ikke syns” selv med fire desimaler. Vi ser f.eks. at sannsynligheten for at minst en del etterspørres er $1 - 0.3679 = 0.6321$, mens sannsynligheten for minst 2 deler er $1-(0.3679+0.3679) = 0.2642$. Sjekk tabelloppslagene med utregning av formelutrykkene ovenfor!

Eksempel 8 : Bakterier

I drikkevann fra en bestemt kilde har man erfart at det er gjennomsnittlig 2 bakterier pr. volumenhet. Det tas en tilfeldig prøve på en volumenhet. La X være antall bakterier i denne prøven, og la oss anta at X er Poissonfordelt med forventning $\lambda = 2$. Av Tabell C.4 finner vi (sml. formeluttrykk ovenfor)

x	0	1	2	3	4	5	6	$\cdots$
$P(X = x)$	0.1353	0.2707	0.2707	0.1804	0.0902	0.0361	0.0120	$\cdots$

Vi ser f. eks. at sannsynligheten for å finne akkurat to bakterier er 0.2707, mens sannsynligheten for å finne mer enn to blir $1 - (0.1353 + 0.2707 + 0.2707) = 0.3233$.

Poissonfordelingen brukes også ofte som en tilnærmelse til den binomiske fordeling. Vi har nemlig

> **Poissontilnærmelsen :**
> Dersom X er binomisk fordelt (n, p), hvor n er stor og p liten, så er X tilnærmet Poissonfordelt med forventning $\lambda = n \cdot p$

Når n er stor og p er liten gjelder altså at

$$p(x) = \binom{n}{x} p^x (1-p)^{n-x} \approx \frac{(np)^x}{x!} e^{-np}$$

For å antyde hva som menes med stor og liten, ser vi på to eksempler der $\lambda = np = 2$.

Binomisk	x	0	1	2	3	4	$\cdots$
$n = 10$ $p = 0.2$	$p(x)$	0.1074	0.2684	0.3030	0.2013	0.0881	$\cdots$
$n = 100$ $p = 0.02$	$p(x)$	0.1353	0.2702	0.2734	0.1823	0.0902	$\cdots$

De tilsvarende Poissonsannsynligheter for $\lambda = np = 2$ var

Poisson	x	0	1	2	3	4	$\cdots$
$\lambda = 2$	$p(x)$	0.1353	0.2707	0.2707	0.1804	0.0902	$\cdots$

Som vi ser fungerer tilnærmelsen betydelig bedre i det siste eksemplet. Det er ikke praktisk å gi noen tommelfingerregler om hvor stor n og hvor liten p bør være for å få en brukbar tilnærmelse. Hva som skal anses brukbart vil avhenge av det problem som skal løses. Som en rettesnor kan likevel anføres at $\lambda = np$ i alle fall bør ligge i området fra 0.1 til 10.0 når Poissontilnærmelsen brukes.

Eksempel 9 : Forsikring
I befolkningen blir gjennomsnittlig 20 personer pr. 10 000 rammet av en bestemt type skade i løpet av et år. Et forsikringsselskap har $n = 1000$ personer som er forsikret mot slik skade. La X være antall utbetalinger for slik skade i løpet av et år. Under rimelige antakelser (hvilke ?) blir X binomisk fordelt $(n = 1000, p = 0.002)$. Siden n er stor mens p er liten, blir X tilnærmet Poissonfordelt med forventning $\lambda = np = 2$. Av tabellen ovenfor kan vi derfor hente følgende tilnærmede sannsynligheter:

Ingen utbetaling	0.1353
To utbetalinger	0.2707
Høyst to utbetalinger	0.1353+0.2707+0.2707=0.6767
Minst tre utbetalinger	1− 0.6767=0.3233

En Poissonmodell vil ofte være aktuell i situasjoner der det foreligger et "medium" som kan inneholde visse "partikler", og X er antall partikler i et utsnitt av mediet. Mediet kan f. eks. tenkes å være den reelle tallinjen, planet eller rommet. Parameteren λ er da en konstant som kan tolkes som gjennomsnittlig antall partikler pr. enhet av mediet. I Eksempel 7 er mediet tallinjen (tiden), partiklene tidspunktene for innkommet bestilling og λ gjennomsnittlig antall bestillinger pr. tidsenhet (dag). I Eksempel 8 er mediet rommet (vannet), partiklene bakteriene og λ gjennomsnittlig antall bakterier pr. volumenhet. Det kan vises at en Poissonmodell er realistisk dersom følgende tre betingelser er oppfylt:

1. Partiklene opptrer uavhengig av hverandre i ikke-overlappende utsnitt av mediet.

2. Sannsynligheten for at det finnes en partikkel i et lite utsnitt av mediet er proporsjonal med størrelsen av utsnittet.

3. Sannsynligheten for mer enn en partikkel i et lite utsnitt av mediet er liten sammenlignet med sannsynligheten for akkurat en partikkel.

Disse tre antakelsene kan med rimelighet hevdes å være oppfylt i situasjonene beskrevet i Eksempel 7 og 8. I mange anvendelser vil utsnittet av mediet ikke nødvendigvis være lik enheten. Dersom λ fortsatt betegner antall partikler pr. enhet vil, under betingelsene ovenfor, antall partikler X i et utsnitt på t enheter av mediet være Poissonfordelt med forventning $\lambda \cdot t$, dvs.

$$P(X = x) = \frac{(\lambda t)^x}{x!} e^{-\lambda t} \ , \quad x = 0, 1, 2, \ldots$$

For diskusjon av betingelsene 1 - 3 i konkrete situasjoner se Oppgave 10.

$\star$ Vi skal til slutt gi en kort begrunnelse for at Poisson-sannsynligheter framkommer som grense for binomiske sannsynligheter : Sett $\lambda = n \cdot p$ i uttrykket for binomiske sannsynligheter og omform dette som følger

$$\begin{aligned}
\binom{n}{x} p^x (1-p)^{n-x} &= \frac{n(n-1)\cdots(n-x+1)}{x!} \left(\frac{\lambda}{n}\right)^x \left(1 - \frac{\lambda}{n}\right)^{n-x} \\
&= \frac{\lambda^x}{x!} \left(1 - \frac{\lambda}{n}\right)^n \frac{1 \cdot (1 - \frac{1}{n}) \cdot (1 - \frac{2}{n}) \cdots (1 - \frac{x-1}{n})}{(1 - \frac{\lambda}{n})^x}
\end{aligned}$$

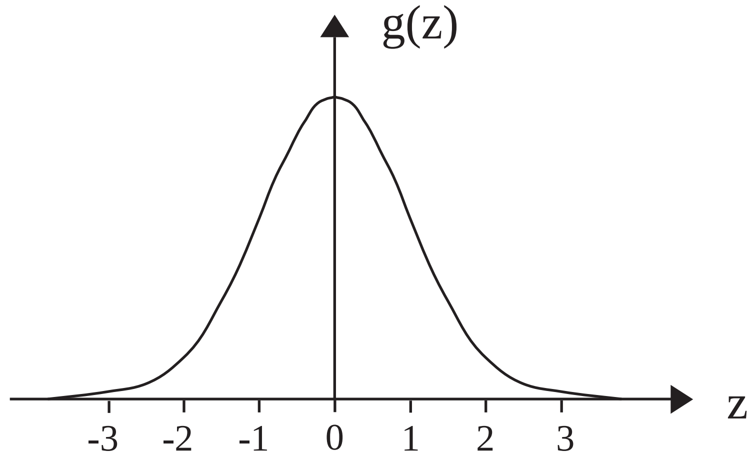
Hold λ fast og la n gå mot uendelig, dvs. p gå mot null. For gitt x vil tredje faktor gå mot 1, mens annen faktor ifølge velkjente resultater i matematisk analyse går mot $e^{-\lambda}$, som gir Poissonsannsynligheten. Dette er en slags teoretisk begrunnelse for at tilnærmelsen ovenfor kan være brukbar for stor n og liten p.

6.5 Normaltilnærming.

Betrakt følgende funksjon:

$$g(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}$$

Det grafiske bildet av denne funksjonen er gitt i Figur 6.1. Denne funksjonen har en bemerkelsesverdig evne til å dukke opp om og om igjen i statistisk teori, og er gitt navnet *normalkurven.* [4]



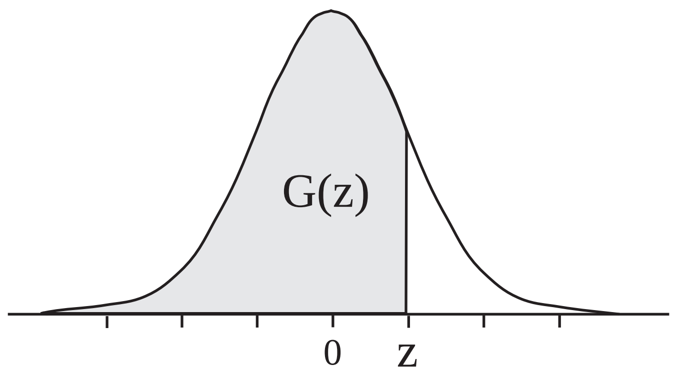
Figur 6.1: Normalkurven

Vi ser at normalkurven har klokkefasong, og er symmetrisk om null. Det kan også vises at det totale areal som avgrenses av kurven og z-aksen er 1. I statistiske tabellverker finner vi tabulert (se Tabell C.5 og C.6 i Appendiks C)

$$G(z) = \text{ Arealet under normalkurven til venstre for } z.$$

Dette er illustrert i Figur 6.2.

[4] I litteraturen brukes også betegnelsen Gauss-kurven etter matematikeren Gauss (1777 - 1855)



Figur 6.2: Arealer under Normalkurven

Vi ser at arealet under normalkurven til høyre for z blir $1 - G(z)$, mens arealet under normalkurven mellom u og v blir $G(v) - G(u)$ når $u \leq v$. Siden normalkurven er symmetrisk om null må vi ha

G1. $\quad G(-z) = 1 - G(z)$

Den siste formelen innebærer at det er nok å tabellere $G(z)$ for positive argumenter. La oss presentere et utsnitt av Tabell C.5:

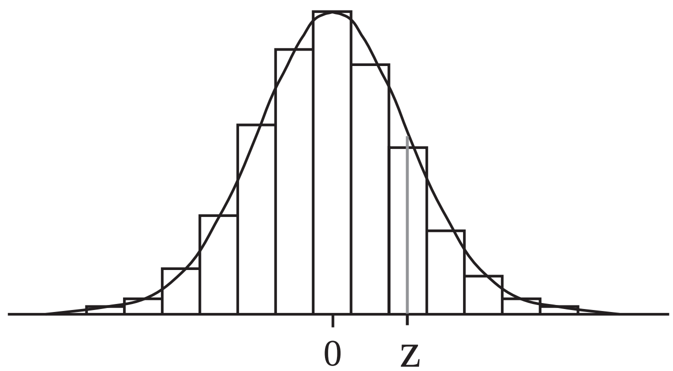
z	0	0.5	1.0	1.5	2.0	3.0	$\cdots$	∞
$G(z)$	0.500	0.692	0.841	0.933	0.977	0.999	$\cdots$	1.000

Vi skal nå se at normalkurven kan brukes som et hjelpemiddel ved tilnærmet beregning av sannsynligheter: La X være en stokastisk variabel med forventning μ og varians σ^2. Betrakt den standardiserte variable

$$Z = \frac{X - \mu}{\sigma}$$

Anta at histogrammet til Z kan tilnærmes med normalkurven, som illustrert i Figur 6.3. Dette vil bl. a. bety at $P(Z \leq z) \approx G(z)$.
Siden $P(X \leq x) = P(\frac{X-\mu}{\sigma} \leq \frac{x-\mu}{\sigma}) = P(Z \leq \frac{x-\mu}{\sigma})$, får vi formelen

G2. $\quad P(X \leq x) \approx G(\frac{x - \mu}{\sigma})$



Figur 6.3: Normaltilnærming

Med en fordeling som kan tilnærmes med normalkurven er det, for tilnærmet beregning av sannsynligheter, nok å kjenne forventning og standardavvik. Andre typer sannsynligheter kan henføres til denne grunnleggende formlen. Ved å bruke $P(X > x) = 1 - P(X \leq x)$ og $P(a < X \leq b) = P(X \leq b) - P(X \leq a)$, får vi f.eks.

G3. $$P(X > x) \approx 1 - G(\frac{x-\mu}{\sigma})$$

G4. $$P(a < X \leq b) \approx G(\frac{b-\mu}{\sigma}) - G(\frac{a-\mu}{\sigma})$$

I hvilke situasjoner kan vi rettferdiggjøre bruk av normaltilnærming? Følgende situasjon er spesielt viktig:

Dersom X kan skrives som en sum av et tilstrekkelig stort antall uavhengige stokastiske variable med samme sannsynlighetsfordeling, så vil histogrammet til $Z = \frac{X-\mu}{\sigma}$ kunne tilnærmes med normalkurven.

Begrunnelsen for dette ligger delvis i følgende setning:

Sentralgrensesetningen :
Anta at $X_1, X_2, \ldots, X_n$ er uavhengige stokastiske variable med samme sannsynlighetsfordeling, og la

$$S_n = X_1 + X_2 + \cdots + X_n.$$

Når n vokser mot uendelig vil [a]

$$P(\frac{S_n - ES_n}{\sigma(S_n)} \leq z) \to G(z)$$

[a]Forutsett at forventning og varians eksisterer (jfr. Kapittel 5.8).

Under forutsetningene i setningen vil en derfor kunne vente at

$$P(\frac{S_n - ES_n}{\sigma(S_n)} \leq z) \approx G(z) \quad \text{for tilstrekkelig stor } n.$$

Sentralgrensesetningen har hatt en fundamental betydning i utviklingen av moderne sannsynlighetsteori, og den har også vidtrekkende konsekvenser for statistisk praksis. Setningen tilskrives gjerne Laplace, men et spesialtilfelle av setningen var kjent av De Moivre [5]). Senere er det utviklet mer generelle versjoner av setningen, bl. a. behøver de stokastiske variable som inngår ikke nødvendigvis ha samme sannsynlighetsfordeling, visse typer avhengighet tillates også, men det vil føre for langt å komme inn på disse forhold her. Et bevis for setningen vil også falle utenfor rammen av denne framstilling. Vi vil isteden forsøke å gi leseren en viss forståelse av hvordan et slikt resultat kan oppstå, ved å se på et eksempel:

Anta at den felles sannsynlighetsfordeling til X_i'ene er gitt ved $p(x)$ (Se Oppgave 5.41).

x	0	1	2
$p(x)$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$

Sannsynlighetsfordelingen til $S_2 = X_1 + X_2$ er da

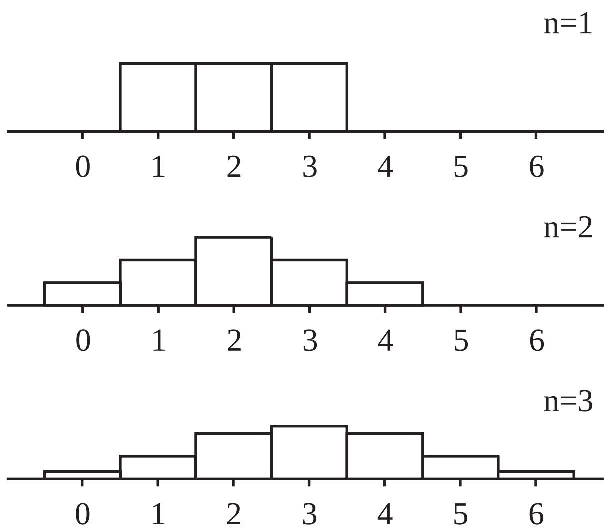
s	0	1	2	3	4
$P(S_2 = s)$	$\frac{1}{9}$	$\frac{2}{9}$	$\frac{3}{9}$	$\frac{2}{9}$	$\frac{1}{9}$

[5]De Moivre fransk matematiker (1667 - 1754). Laplace fransk matematiker (1759 - 1827).

mens sannsynlighetsfordelingen til $S_3 = X_1 + X_2 + X_3$ blir

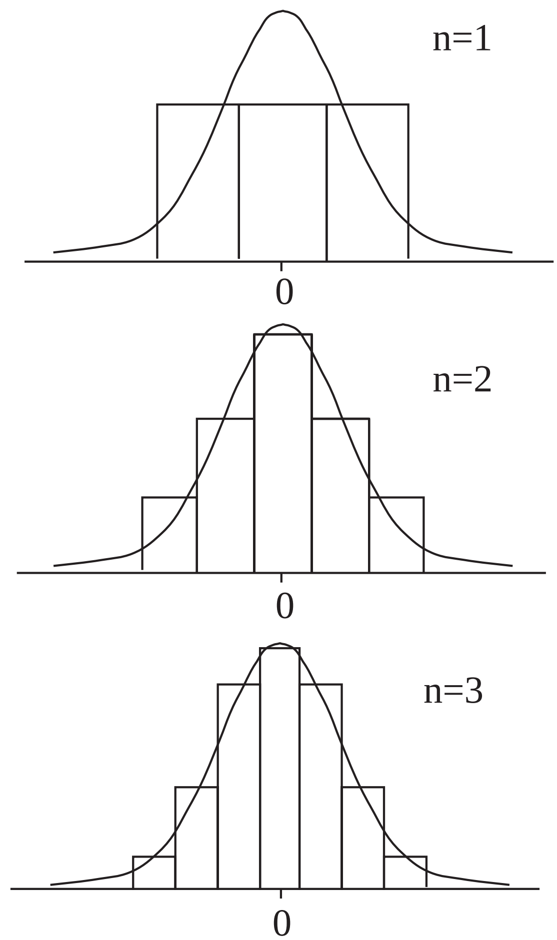
s	0	1	2	3	4	5	6
$P(S_3 = s)$	$\frac{1}{27}$	$\frac{3}{27}$	$\frac{6}{27}$	$\frac{7}{27}$	$\frac{6}{27}$	$\frac{3}{27}$	$\frac{1}{27}$

Histogrammene til S_1, S_2 og S_3 er derfor som vist i Figur 6.4.



Figur 6.4: Histogrammet til S_n

Vi har her startet med uavhengige variable med histogram som slett ikke er klokkeformet, men allerede for $n = 3$ viser histogrammet til S_n trekk av klokkefasong. Vi ser imidlertid at ettersom n vokser vil histogrammene bli utflytende og det vanskeliggjør sammenligninger. Dette er en av grunnene til at vi standardiserer, slik at vi kan sammenligne fordelinger med samme forventning og varians. Histogrammene til de standardiserte variable ser ut som i Figur 6.5.

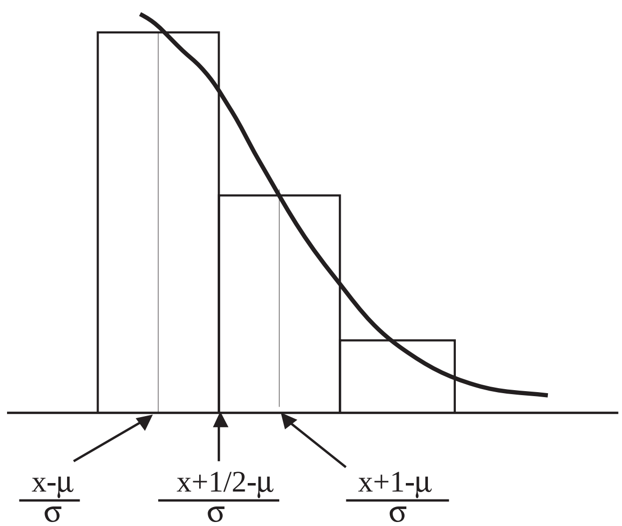


Figur 6.5: Histogrammer til de standardiserte S_n

Vi ser at tendensen mot en klokkeformet kurve er åpenbar, og dersom vi fortsetter med $n = 4, 5, \ldots$, vil vi ytterligere få bekreftet dette. For stor n vil søylene i histogrammet bli smale, men til gjengjeld blir det mange av dem, og omrisset av histogrammet kan etter hvert ikke skjeldnes fra en kontinuerlig kurve, som altså viser seg å være normalkurven. Det oppsiktsvekkende er at det er den samme kurven som framkommer uansett hvilken sannsynlighetsfordeling vi starter med. Selv dersom vi starter med en fordeling av en helt annen karakter enn i eksemplet ovenfor vil histogrammene til en sum av uavhengige variable med denne fordeling nærme seg normalkurven, i den forstand som er gjort presist i sentralgrensesetningen. Imidlertid kan en vente at tilnærmingen skjer noe langsommere dersom vi starter med en U-formet fordeling (Se Oppgave 5.41 (ii)) eller med en skjev fordeling (Se Oppgave 5.41 (iii)).

Før vi ser på anvendelser av normaltilnærming vil vi gjøre noen merknader:

La X være en stokastisk variabel med forventning μ og varians σ^2 som er slik at den standardiserte variable $Z = (X - \mu)/\sigma$ har histogram som kan tilnærmes med normalkurven. I en rekke anvendelser vil X være en heltallsvariabel, dvs. X kan bare anta verdier blant de hele tall. I slike tilfeller vil vi kunne gi en enda bedre tilnærming enn den vi har i formelen $P(X \leq x) \approx G((x - \mu)/\sigma)$. La oss studere et utsnitt av histogrammet til Z sammen med et utsnitt av normalkurven.



Figur 6.6: Normaltilnærming for heltallsvariable

I Figur 6.6 er tegnet to søyler sentrert i to mulige verdier av Z, nemlig $(x - \mu)/\sigma$ og $(x + 1 - \mu)/\sigma$ der x er heltall. Siden $P(X \leq x) = P(Z \leq (x - \mu)/\sigma)$ er lik summen av arealene av søylen over $(x - \mu)/\sigma$ og søylene til venstre for $(x - \mu)/\sigma$, ser vi av figuren at vi får en bedre tilnærming dersom vi, istedenfor å beregne arealet under normalkurven til venstre for $(x - \mu)/\sigma$, beregner arealet til venstre for midtpunktet mellom de to mulige verdiene, nemlig $(x + \frac{1}{2} - \mu)/\sigma$. Disse betraktningene har gitt oss følgende tilnærmingsformel

$$\text{G5.} \qquad P(X \le x) \approx G(\frac{x + \frac{1}{2} - \mu}{\sigma})$$

Dette kalles *normaltilnærming med heltallskorreksjon* og formelen brukes bare i situasjoner der X er en heltallsvariabel. Formlene G3 og G4 ovenfor kan i denne situasjon modifiseres tilsvarende.

Eksempel 10 : Terningkast
Vi utfører n uavhengige kast med en vanlig terning og lar

$$X_i = \text{antall øyne i } i\text{'te kast } i = 1, 2, \ldots, n$$

Her blir $X_1, X_2, \ldots, X_n$ uavhengige stokastiske variable med samme sannsynlighetsfordeling. Vi vet fra før at $EX_i = 7/2$, mens $varX_i = 35/12$. Sum øyne i n kast blir

$$S_n = X_1 + X_2 + \cdots + X_n$$

Her blir $ES_n = n \cdot \frac{7}{2}$, mens $varS_n = n \cdot \frac{35}{12}$, slik at når n går mot uendelig vil

$$P(\frac{S_n - n \cdot \frac{7}{2}}{\sqrt{n \cdot \frac{35}{12}}} \le z) \to G(z)$$

Vi har tidligere tegnet histogrammet til S_n i tilfellet $n = 2$. Dette hadde form av en trekant. Det samme vil være tilfelle for den standardiserte variable. Allerede for $n = 3$ vil histogrammet få den karakteristiske klokkefasongen (Oppgave 23). La antall kast $n = 4$. Da blir $ES_4 = 14$ og $varS_4 = 35/3 = (3.4)^2$. La oss beregne sannsynligheten for at sum øyne blir høyst 10. Eksakt beregning gir (tidkrevende opptelling)

$$P(S_4 \le 10) = \frac{203}{6^4} = \frac{203}{1296} = 0.137$$

Tilnærmet beregning (med heltallskorreksjon) gir

$$\begin{aligned} P(S_4 \le 10) \approx G(\frac{10 + 0.5 - 14}{\sqrt{35/3}}) &= G(-1.02) = 1 - G(1.02) \\ &= 1 - 0.846 = 0.154 \end{aligned}$$

dvs. slett ikke verst. Spørsmålet om hva som kan betraktes som en brukbar tilnærmelse må vi imidlertid avgjøre i hver konkret anvendelse, og avhenger selvsagt av formålet med beregningen.

Eksempel 11 : Binomiske sannsynligheter
Anta at X er binomisk fordelt (n, p). Vi vet da at vi kan skrive

$$X = I_1 + I_2 + \cdots + I_n$$

hvor $I_1, I_2, \ldots, I_n$ er indikatorer som oppfyller kravene til uavhengighet og samme fordeling i sentralgrensesetningen. Siden $EX = np$ og $varX = np(1-p)$, kan vi slutte at når n går mot uendelig vil

$$P(\frac{X - n \cdot p}{\sqrt{np(1-p)}} \leq z) \to G(z)$$

La X = antall kron i løpet av 10 myntkast. Da er X binomisk fordelt ($n = 10, p = \frac{1}{2}$). Vi ønsker å beregne sannsynligheten for høyst 4 kron. Eksakt regning gir

$$P(X \leq 4) = \frac{386}{1024} = 0.377$$

Tilnærmet regning gir

$$\begin{aligned} P(X \leq 4) \approx G(\frac{4 + \frac{1}{2} - 10 \cdot \frac{1}{2}}{\sqrt{10 \cdot \frac{1}{2} \cdot \frac{1}{2}}}) &= G(-0.317) = 1 - G(0.317) \\ &= 1 - 0.624 = 0.376 \end{aligned}$$

Anta at det isteden er tale om $n = 100$ myntkast, og vi ønsker å beregne sannsynligheten for høyst 40 kron. Siden de fleste tabeller over den binomiske fordeling ikke går så langt som til $n = 100$, er vi nødt å ty til normaltilnærmelse. Vi får

$$\begin{aligned} P(X \leq 40) \approx G(\frac{40 + 0.5 - 50}{\sqrt{100 \cdot \frac{1}{2} \cdot \frac{1}{2}}}) &= G(-1.9) = 1 - G(1.9) \\ &= 1 - 0.971 = 0.029 \end{aligned}$$

For beregning av binomiske sannsynligheter ved hjelp av normaltilnærming vil vi gi en *tommelfingerregel:* Normaltilnærming gir som regel godt resultat når $np(1-p) \geq 10$, brukbart for $np(1-p) \geq 5$.

Eksempel 12 : Hypergeometriske sannsynligheter
Anta at Y er hypergeometrisk fordelt (N, M, n). Vi har sett at vi kan skrive

$$Y = I_1 + I_2 + \cdots + I_n$$

der $I_1, I_2, \ldots, I_n$ er indikatorer med samme sannsynlighetsfordeling, men som er avhengige, slik at kravet i sentralgrensesetningen ikke er oppfylt. Dersom n er liten i forhold til N, vet vi at $I_1, I_2, \ldots, I_n$ er tilnærmet uavhengige. I en slik situasjon kan vi, dersom vi ikke kommer fram på annen måte, likevel bruke normaltilnærming med brukbart resultat. La oss ta et konkret eksempel:
I en forening på $N = 1000$ medlemmer er det $M = 500$ kvinner og $N - M = 500$ menn. Det trekkes et tilfeldig utvalg på $n = 100$ medlemmer som inviteres til jubileumsmiddag. La Y være antall kvinner i utvalget. Vi ønsker å beregne sannsynligheten for at det inviteres færre kvinner enn menn, dvs. høyst 49 kvinner, samt sannsynligheten for at det inviteres høyst 45 kvinner. Vi ser at Y er hypergeometrisk fordelt $(N = 1000, M = 500, n = 100)$. Her blir $EY = n \cdot \frac{M}{N} = 50$ og $varY = \frac{N-n}{N-1} \cdot n \cdot \frac{M}{N} \cdot (1 - \frac{M}{N}) = 22.5$ Vanlige tabeller strekker ikke til. Siden n er stor, men likevel relativt liten i forhold til N, forsøker vi normaltilnærming:

$$\begin{aligned}
P(Y \leq 49) \approx G(\frac{49 + 0.5 - 50}{\sqrt{22.5}}) &= G(-0.105) = 1 - G(0.105) \\
&= 1 - 0.542 = 0.458 \\
P(X \leq 45) \approx G(\frac{45 + 0.5 - 50}{\sqrt{22.5}}) &= G(-0.949) = 1 - G(0.949) \\
&= 1 - 0.829 = 0.171
\end{aligned}$$

Av et større tabellverk finner vi de eksakte sannsynligheter, men de er ikke nevneverdig forskjellig fra de tilnærmede.

Eksempel 13 : Poissonsannsynligheter
Anta at X er Poissonfordelt med forventning λ. Dersom λ er stor kan vi benytte normaltilnærming. Dette er ikke like lett å begrunne som i de to foregående eksemplene. Det faktum at Poissonsannsynligheter kan oppnås som grense for binomiske sannsynligheter for voksende n, og at binomiske sannsynligheter kan tilnærmet beregnes ved bruk av normalkurven for stor n, antyder at normaltilnærming kan være aktuelt også her. Et annet argument: I situasjoner der λ er heltall kan vi skrive (se Oppgave 37)

$$X = X_1 + X_2 + \cdots + X_\lambda$$

der X_i'ene er uavhengige og X_i Poissonfordelt med forventning 1. Vi kan med referanse til sentralgrensesetningen slutte at når λ går mot uendelig vil

$$P(\frac{X-\lambda}{\sqrt{\lambda}} \leq z) \to G(z)$$

Vi ser at Tabell C.4 ikke går lenger enn til $\lambda = 10.0$. For $\lambda \geq 10.0$ vil normaltilnærming gi brukbart resultat for de fleste praktiske formål. La oss ta et eksempel: Anta at antall ordrer på en bestemt vare pr. dag er Poissonfordelt med forventning $\lambda = 10.0$. Da blir eksempelvis

$$P(X \leq 8) \approx G(\frac{8+0.5-10}{\sqrt{10}}) = G(-0.4743) = 0.318$$

mens det korrekte resultat er 0.333.

Noen sluttkommentarer om normaltilnærming:
Anta fortsatt at histogrammet til $Z = (X-\mu)/\sigma$ kan tilnærmes med normalkurven, og la $A(k)$ være arealet under normalkurven mellom $-k$ og k. Dette kan beregnes ved

$$A(k) = G(k) - G(-k) = G(k) - (1 - G(k)) = 2G(k) - 1.$$

Vi får

$$\boxed{\text{G6.} \qquad P(\mid X-\mu \mid < k\cdot\sigma) \approx A(k)}$$

dvs. at sannsynligheten for at X faller innenfor k ganger standardavviket fra sin forventning er entydig (tilnærmet) gitt ved k. La oss gi noen verdier

k	0.5	1.0	2.0	3.0
$A(k)$	0.38	0.68	0.95	0.99

Ved å sette $k = d/\sigma$ i formelen ovenfor får vi en alternativ formel

$$\boxed{\text{G7.} \qquad P(\mid X-\mu \mid < d) \approx A(\frac{d}{\sigma})}$$

Vi ser at når standardavviket σ er kjent, vil vi kunne beregne tilnærmet sannsynligheten for at X faller innenfor et bestemt antall enheter d fra sin forventning, dvs. σ kan direkte relateres til risiko.

Eksempel 14 : Tegnestiften
La sannsynligheten være p for at en tegnestift faller med spissen opp ved et kast, og la X være antall ganger den peker opp i n uavhengige kast, slik at X er binomisk fordelt (n, p). Betrakt sannsynligheten

$$P(\mid \frac{X}{n} - p \mid < k \cdot \sigma(\frac{X}{n})) \quad \text{der} \quad \sigma(\frac{X}{n}) = \sqrt{\frac{p(1-p)}{n}}.$$

Ifølge teorien ovenfor blir denne sannsynligheten tilnærmet lik $A(k)$. Det vil være instruktivt å beregne σ for noen verdier av n og p. Siden p trolig ligger i området omkring 0.5, nøyer vi oss med noen få p-verdier [6]

p	$n =$	100	500	1 000	10 000
0.4		0.049	0.021	0.015	0.0049
0.5		0.050	0.022	0.016	0.0050
0.6		0.049	0.021	0.015	0.0049

Av denne tabellen (og tabellen ovenfor) kan vi f.eks. slutte at når $n = 1000$ er sannsynligheten for at hyppigheten avviker fra sannsynligheten p med mindre enn 0.016 tilnærmet lik 0.68, mens sannsynligheten for at hyppigheten avviker mindre enn $2 \cdot 0.016 = 0.032$ er tilnærmet lik 0.95. Dette eksemplet belyser fra en teoretisk synsvinkel de forhold som ble diskutert i Eksempel 2 i Kapittel 1, hvor vi søkte å motivere sannsynlighetsbegrepet.

6.6 Tsjebysjeffs ulikhet. Store talls lov.

I slutten av forrige avsnitt så vi at standardavviket er svært informativt i situasjoner hvor normaltilnærming er brukbar. Følgende setning viser imidlertid at standardavviket til en viss grad også er informativt i mer generelle situasjoner.

[6] Merk at σ varierer lite i området omkring $p = 0.5$, for øvrig vil p alltid gi samme σ som $p' = 1 - p$.

Tsjebysjeffs ulikhet : [a]
La X være en stokastisk variabel med forventning μ og standardavvik σ. Da er

$$P(\mid X-\mu \mid < k\cdot\sigma) \geq 1-\frac{1}{k^2} \quad (\text{for } k\cdot\sigma > 0).$$

[a]Oppkalt etter den russiske matematiker Tsjebysjeff (1821 - 1894).

Dette viser at uansett sannsynlighetsfordelingen til X så gir ulikheten en nedre skranke $1-1/k^2$ for sannsynligheten for avvik i tallverdi mindre enn k ganger standardavviket fra forventningen. Denne skranken er for noen verdier av k

k	1	2	3	4	5	$\cdots$
$1-1/k^2$	0.00	0.75	0.88	0.94	0.96	$\cdots$

For en gitt sannsynlighetsfordeling vil vanligvis de eksakte sannsynlighetene være betydelig høyere, men tabellen gir altså garanterte verdier som kan nyttes også i situasjoner hvor vi ikke kjenner sannsynlighetsfordelingen i detalj.
Begrunnelse for Tsjebysjeffs ulikhet: Vi ser at det er nok å vise at

$$P(\mid Z \mid < k) \geq 1-1/k^2$$

der $Z=(X-\mu)/\sigma$ er den standardiserte variable til X. Definer en ny stokastisk variabel Y ved at

$$\begin{aligned} Y &= 0 \quad \text{dersom} \quad \mid Z \mid < k \\ &= k^2 \quad \text{dersom} \quad \mid Z \mid \geq k \end{aligned}$$

Da er $Y \leq Z^2$ og følgelig også $EY \leq EZ^2 = varZ = 1$. Men nå er

$$EY = 0\cdot P(\mid Z \mid < k) + k^2 P(\mid Z \mid \geq k).$$

Følgelig blir $k^2 P(\mid Z \mid \geq k) \leq 1$, eller med andre ord

$$P(\mid Z \mid \geq k) \leq 1/k^2$$

som er ekvivalent med det søkte resultat.

Eksempel 15 : Tennsatsen
En produsent lager tennsatser og har erfart at brenntiden til en tennsats kan

oppfattes som en stokastisk variabel X med forventning 5.0 og standardavvik 0.2. Da er eksempelvis

$$P(\mid X - 5.0 \mid < 1.0) \geq 0.96,$$

uansett hvilken sannsynlighetsfordeling X har.

En variant av Tsjebysjeffs ulikhet får vi ved å sette $d = k \cdot \sigma$, dvs.

$$P(\mid X - \mu \mid < d) \geq 1 - \frac{\sigma^2}{d^2} \quad (\text{for } d > 0).$$

En interessant anvendelse av denne har vi i den binomiske situasjon:

Anta at X er binomisk fordelt (n, p), og la oss anvende den siste ulikheten på den relative hyppighet av suksesser i de n forsøk. Siden $E(X/n) = p$ og $var(X/n) = p(1-p)/n$, får vi

$$P(\mid \frac{X}{n} - p \mid < d) \geq 1 - \frac{p(1-p)}{nd^2}.$$

Vi ser at den garanterte nedre skranke for sannsynligheten for at den relative hyppighet av suksesser avviker fra sannsynligheten p med mindre enn d, vokser med n for gitt $d > 0$. Vi ser at vi faktisk kan konkludere at

$$P(\mid \frac{X}{n} - p \mid < d) \to 1 \quad \text{når} \quad n \to \infty$$

uansett $d > 0$. I denne forbindelse sies ofte at den relative hyppighet *konvergerer* mot p i *sannsynlighet.* Dette resultat omtales ofte som *Bernoullis lov om de store tall*[7], og er en slags teoretisk bekreftelse på at sannsynligheter kan framkomme som idealiserte relative hyppigheter i det lange løp (Se Kapittel 1).

6.7 ⋆Den geometriske fordeling

Anta at vi har en binomisk forsøksrekke med suksess-sannsynlighet p, og la

$$N = \quad \text{ventetiden til første suksess.}$$

I Kapittel 4.6 fant vi at

[7]Oppkalt etter Jakob Bernoulli (1654 - 1705).

$$p(n) = P(N = n) = (1 - p)^{n-1} \cdot p \ ; \ n = 1, 2, 3, \dots$$

Denne sannsynlighetsfordelingen kalles *den geometriske fordeling*. Ifølge summeformelen for en geometrisk rekke får vi

$$\sum_n p(n) = p \sum_{n=1}^{\infty} (1 - p)^{n-1} = p((1 - (1 - p))^{-1} = 1$$

som viser at vi virkelig har en sannsynlighetsfordeling. Vi ønsker å beregne forventningen til N. Vi får

$$EN = \sum_n np(n) = p \sum_{n=1}^{\infty} n(1 - p)^{n-1} = \frac{1}{p}$$

Begrunnelse: La for enkelhets skyld $q = 1 - p$ og betrakt

$$\begin{aligned}
(1 - q) \sum_{n=1}^{\infty} nq^{n-1} &= \sum_{n=1}^{\infty} nq^{n-1} - \sum_{n=1}^{\infty} nq^n \\
&= \sum_{n=1}^{\infty} nq^{n-1} - \sum_{n=2}^{\infty} (n - 1)q^{n-1} \\
&= 1 + \sum_{n=2}^{\infty} q^{n-1} = 1 + q(1 - q)^{-1} = p^{-1}
\end{aligned}$$

Eksempler på anvendelser av dette resultat er: Forventet ventetid til første kron i myntkast (p=1/2) blir 2. Forventet ventetid til første sekser i terningkast (p=1/6) blir 6. Forventet ventetid til første defekt når defektsannsynligheten er $p = 0.05$ blir 20.

Variansen til en geometrisk fordelt stokastisk variabel N viser seg å bli $varN = q/p^2$. Denne utledning er noe mer komplisert, og siden resultatet ikke har så stor praktisk interesse, utelater vi begrunnelsen her. Vi vil isteden rette oppmerksomheten mot en interessant egenskap ved den geometriske fordeling. La oss først finne sannsynligheten for at ventetiden til første suksess er mer enn k

$$P(N > k) = \sum_{n=k+1}^{\infty} p(n) = p \sum_{n=k+1}^{\infty} (1 - p)^{n-1} = (1 - p)^k$$

Gitt at ventetiden er minst k, da er sannsynligheten for at vi må vente n tidsenheter til før første suksess:

$$\begin{aligned}
P(N = k + n \mid N > k) &= \frac{P(N = k + n)}{P(N > k)} \\
&= \frac{p(1 - p)^{k+n-1}}{(1 - p)^k} = p(1 - p)^{n-1}
\end{aligned}$$

som jo er det samme som $P(N = n)$. dette betyr at gjenstående ventetid er upåvirket av hvor lenge vi allerede har ventet, vi sier at den geometriske fordeling er uten minne. Dette er kanskje ikke så overaskende når vi tenker på at vi har utledet den geometriske fordeling fra en serie med uavhengige binomiske forsøk. Det kan vises at den geometriske fordeling er den eneste diskrete fordeling som er uten minne.

6.8 ⋆Den multinomiske fordeling

Vi husker at en binomisk forsøksrekke var en rekke uavhengige forsøk der hvert forsøk hadde to mulige utfall og der sannsynlighetene for disse utfallene var den samme i alle forsøkene. I dette avsnittet vil vi se på en tilsvarende situasjon, men hvor det kan foreligge mer enn to mulige utfall i hvert forsøk. [8]) Det utføres en rekke eksperimenter (forsøk) hvor:

1. Hvert forsøk gir et av m mulige utfall $u_1, u_2, \ldots, u_m$.
2. Sannsynlighetene for hvert av de mulige utfallene er de samme i alle forsøk.
3. Forsøkene er uavhengige.

Denne situasjon kaller vi en *multinomisk forsøksrekke.*

Vi ser at tilfellet $m = 2$ svarer til binomisk forsøksrekke. Tilfellet $m = 3$ kalles ofte en trinomisk forsøksrekke. Eksempelvis har vi tidligere sett at i en løpende produksjon hvor hver produsert artikkel blir klassifisert som intakt (i) eller defekt (d) kan betraktes som en binomisk forsøksrekke. Dersom intakte artikler klassifiseres i tre kvaliteter (a, b og c) kan produksjonen betraktes som en multinomisk forsøksrekke med $m = 4$ mulige resultater i hvert forsøk. Vi må da være villige til å anta uavhengighet og samme sannsynlighet for de ulike kvalitetskategorier for alle forsøk.

Anta at vi utfører n multinomiske forsøk og anta at sannsynligheten for at utfallene $u_1, u_2, \ldots, u_m$ observeres i et enkelt forsøk er henholdsvis $p_1, p_2, \ldots, p_m$ hvor $p_1 + p_2 + \cdots + p_m = 1$. Vi ønsker å beregne sannsynligheten for at

$$\begin{aligned}
&u_1 \text{ observeres } x_1 \text{ ganger}\\
&u_2 \text{ observeres } x_2 \text{ ganger}\\
&\vdots\\
&u_m \text{ observeres } x_m \text{ ganger, der } x_1 + x_2 + \cdots + x_m = n.
\end{aligned}$$

La derfor for $i = 1, 2, \ldots, m$

$$X_i = \quad \text{antall ganger resultatet } u_i \text{ observeres.}$$

[8]I det følgende forutsettes det at leseren er kjent med stoffet i Kapittel 3.7 om gruppering.

Det viser seg at den simultane sannsynlighetsfordelingen til $X_1, X_2, \ldots, X_m$ er gitt ved sannsynlighetsfunksjonen:

$$\begin{aligned} p(x_1, x_2, \ldots, x_m) &= P((X_1 = x_1) \cap (X_2 = x_2) \cap \cdots (X_m = x_m)) \\ &= \frac{n!}{x_1! x_2! \cdots x_m!} p_1^{x_1} p_2^{x_2} \cdots p_m^{x_m} \end{aligned}$$

for alle heltall $x_1, x_2, \ldots, x_m$ slik at $x_1 + x_2 + \cdots + x_m = n$, mens $p(x_1, x_2, \ldots, x_m) = 0$ for alle andre $x_1, x_2, \ldots, x_m$. Dette kalles *den multinomiske sannsynlighetsfordeling*, og vi sier at $(X_1, X_2, \ldots, X_m)$ er multinomisk fordelt med parametre $(n, p_1, p_2, \ldots, p_m)$.

En begrunnelse for dette resultatet kan gis i analogi med den vi ga i den binomiske situasjon: Vi vil først finne ut hvor mange utfallssekvenser som er slik at u_1 forekommer x_1 ganger, u_2 forekommer x_2 ganger, $\ldots, u_m$ forekommer x_m ganger. Vi tenker oss de n forsøkene nummerert fra 1 til n og det søkte antall kan finnes ved å telle opp antall måter å velge ut x_1 plasser (numre) som skal gis symbolet u_1, x_2 plasser som skal gis symbolet $u_2, \ldots, x_{m-1}$ plasser som skal gis symbolet u_{m-1}, mens de resterende $x_m = n - (x_1 + x_2 + \cdots + x_{m-1})$ plassene gis symbolet u_m. Dette er likeverdig med å spørre om antall måter å dele opp populasjonen bestående av tallene $1, 2, \ldots, n$ i m benevnte grupper med henholdsvis x_1 elementer i gruppe nr.1 (u_1-gruppen), x_2-elementer i gruppe nr.2 (u_2-gruppen), $\ldots, x_m$ elementer i gruppe nr.m (u_m-gruppen). I Kapittel 3.7 viste vi at dette antall var lik

$$\binom{n}{x_1, x_2, \ldots, x_{m-1}} = \frac{n!}{x_1! x_2! \cdots x_m!}.$$

På grunn av uavhengigheten og forutsetningen om samme sannsynlighet i alle n forsøk, følger at hver av disse utfallssekvensene har sannsynlighet som er et produkt av n faktorer, hvor x_1 faktorer er lik p_1, x_2 faktorer lik $p_2, \ldots, x_m$ faktorer lik p_m. Det er bare rekkefølgen av faktorene som er forskjellig for de ulike utfallssekvensene. Alle får derfor sannsynlighet

$$p_1^{x_1} p_2^{x_2} \cdots p_m^{x_m}$$

som viser at den oppgitte formel er korrekt.

Merknad. Den multinomiske fordeling slik den er angitt ovenfor er en simultan fordeling for m stokastiske variable $X_1, X_2, \ldots, X_m$. Egentlig er dimensjonen bare $m - 1$ fordi $X_m = n - (X_1 + \cdots + X_{m-1})$ og $p_m = 1 - (p_1 + \cdots + p_{m-1})$, slik at fordelingen er fullstendig beskrevet ved at

$$\begin{aligned} p(x_1, \ldots, x_{m-1}) &= P((X_1 = x_1) \cap \cdots \cap (X_{m-1} = x_{m-1})) \\ &= \frac{n!}{x_1! \cdots x_{m-1}!(n - (x_1 + x_2 + \cdots + x_{m-1}))!} \\ &\cdot \; p_1^{x_1} \cdots p_{m-1}^{x_{m-1}} (1 - (p_1 + \cdots + p_{m-1}))^{n-(x_1+\cdots+x_{m-1})} \end{aligned}$$

Vi ser at den første skrivemåten gir en mer attraktiv formel, og som regel brukes den, unntatt er tilfellet $m = 2$ hvor den binomiske fordeling vanligvis skrives som en fordeling i en dimensjon.

Eksempel 16 : Tolv terningkast
Vi utfører $n = 12$ terningkast. Hvert kast har seks mulige utfall, et øye (u_1) to øyne (u_2), ..., seks øyne (u_6). En rettferdig terning svarer til at $p_1 = p_2 = \cdots = p_6 = 1/6$. Antar vi at kastene er uavhengige, er vi i en multinomisk situasjon. La for $i = 1, 2, \ldots, 6$

$$X_i = \quad \text{antall ganger } i \text{ øyne observeres.}$$

Da blir sannsynlighetsfordelingen til $X_1, X_2, \ldots, X_6$

$$p(x_1, x_2, \ldots, x_6) = \frac{12!}{x_1!x_2!\cdots x_6!}(\frac{1}{6})^{x_1} \cdot (\frac{1}{6})^{x_2} \cdots (\frac{1}{6})^{x_6}$$

for heltall $x_1, x_2, \ldots, x_6$ slik at $x_1 + x_2 + \cdots + x_6 = 12$. Sannsynligheten for at alle øyne forekommer to ganger blir eksempelvis

$$p(2, 2, \ldots, 2) = \frac{12!}{(2!)^6}(\frac{1}{6})^{12} = 0.0034$$

Eksempel 17 : Ti kast med to mynter
Vi utfører $n = 10$ omganger med myntkast med to mynter. I hver omgang observerer vi resultatet ingen kron (u_1), en kron og en mynt (u_2) eller to kron (u_3). En rettferdig mynt svarer til at $p_1 = 1/4, p_2 = 1/2$ og $p_3 = 1/4$. Antar vi at omgangene er uavhengige er vi i en multinomisk situasjon. La X_i være antall ganger u_i observeres $(i = 1, 2, 3)$. Da blir sannsynlighetsfordelingen til X_1, X_2 og X_3

$$p(x_1, x_2, x_3) = \frac{10!}{x_1!x_2!x_3!}(\frac{1}{4})^{x_1} \cdot (\frac{1}{2})^{x_2} \cdot (\frac{1}{4})^{x_3}$$

for heltall x_1, x_2, x_3 slik at $x_1 + x_2 + x_3 = 10$. Sannsynligheten for at vi observerer 3 omganger med begge kron, 4 omganger med en hver og 3 omganger med begge mynt er derfor

$$p(3, 4, 3) = \frac{10!}{3!4!3!}(\frac{1}{4})^3 \cdot (\frac{1}{2})^4 \cdot (\frac{1}{4})^3 = 0.0641$$

Eksempel 18 : Tipperekke
En tipperekke i fotball består av $n = 12$ kamper, hver kamp har som mulige utfall hjemmeseier (H), uavgjort (U) eller borteseier (B). I løpet av en lengre tidsperiode er det observert ca. 45% hjemmeseire, 25% uavgjorte og 30% borteseire. (Tallene er

konstruerte.) La X_1, X_2 og X_3 betegne henholdsvis antall hjemmeseire, uavgjorte og borteseire i kommende spilleuke. Er vi villige til å anta forutsetningene om uavhengighet og samme sannsynlighet for alle kampene på kupongen (diskuter om dette er realistisk), er det rimelig å arbeide ut fra at (X_1, X_2, X_3) er multinomisk fordelt $(n = 12, p_1 = 0.45, p_2 = 0.25, p_3 = 0.30)$ dvs.

$$p(x_1, x_2, x_3) = \frac{12!}{x_1!x_2!x_3!} 0.45^{x_1} \cdot 0.25^{x_2} \cdot 0.30^{x_3}$$

Ut fra dette vil vi kunne finne sannsynlighetene for ulike tegnfordelinger. Eksempelvis blir

$$p(6, 2, 4) = \frac{12!}{6!2!4!} 0.45^6 \cdot 0.25^2 \cdot 0.30^4 = 0.0582$$

Diskuter påstanden: "Det lønner seg å satse på bestemte tegnfordelinger i tipping da noen tegnfordelinger har vist seg å forekomme oftere enn andre".

Anta at $(X_1, X_2, \ldots, X_m)$ er multinomisk fordelt med parametre $(n, p_1, p_2, \ldots, p_m)$. Da vil , for $i = 1, 2, \ldots, m, X_i$ være binomisk fordelt med parametre (n, p_i). Dette innses direkte ved å notere resultatet u_i som suksess og ikke-u_i som fiasko. Dette viser at

$$EX_i = np_i$$

$$varX_i = np_i(1 - p_i).$$

Det er imidlertid klart at $X_1, X_2, \ldots, X_m$ ikke kan være uavhengige. Det viser seg at

$$cov(X_i, X_j) = -np_ip_j \quad \text{for} \quad i \neq j$$

dvs. som rimelig er negativt korrelerte. Dette innses enklest slik: La oss notere suksess dersom u_i eller u_j inntreffer, dette har sannsynlighet p_i+p_j. Antall suksesser $X_i + X_j$ blir da binomisk fordelt med parameter $(n, p_i + p_j)$, slik at

$$var(X_i + X_j) = n(p_i + p_j)(1 - (p_i + p_j)).$$

Men nå er ifølge formel V8

$$var(X_i + X_j) = varX_i + varX_j + 2cov(X_i, X_j),$$

slik at

$$\begin{aligned} cov(X_i, X_j) &= \frac{1}{2}[n(p_i + p_j)(1 - (p_i + p_j)) \\ &\quad -np_i(1 - p_i) - np_j(1 - p_j)] = -np_ip_j. \end{aligned}$$

Kunnskaper om den multinomiske fordeling er nyttig i en rekke forbindelser både i rene sannsynlighetsteoretiske anvendelser og i statistisk praksis, bl.a. ved test av om en gitt sannsynlighetsmodell er realistisk, og ved analyse av såkalte kategoridata.

6.9 Oppgaver

1. La X være binomisk fordelt (n, p). Bruk tabeller til å finne følgende sannsynligheter:
   (a) $P(X = 3)$, $P(X \leq 2)$ og $P(X \geq 4)$ når $n = 5$ og $p = 0.2, 0.4, 0.5$ og 0.6,
   (b) $P(X = 6)$, $P(X \leq 4)$ og $P(X \geq 8)$ når $n = 10$ og $p = 0.2, 0.4, 0.5$ og 0.6,
   (c) $P(X = 12)$, $P(X \leq 8)$ og $P(X \geq 16)$ når $n = 20$ og $p = 0.5$.
2. La X være binomisk fordelt (n, p). Hvilken verdi er mest sannsynlig dersom
   (a) n=5 og p=0.1, 0.2, 0.3, 0.4 og 0.5,
   (b) n=10 og p=0.1, 0.2, 0.3, 0.4 og 0.5.
3. En produksjonsprosess gir 5% defekte når den er i "kontroll". De 10 siste artiklene som produseres hver dag kontrolleres og antall defekte X noteres. Maskinene justeres dersom antall defekte er 3 eller større. Hva er sannsynligheten for justering dersom
   (a) prosessen fortsatt er i kontroll,
   (b) prosessen er ute av kontroll og gir nå gjennomgående 10% defekte.
4. En fabrikant av syltetøy vil undersøke om en ny konserveringsmetode gir endret smak på produktet. Ti personer får smake på 5 smaksprøver hvorav bare en er produsert med den nye metode, og de blir bedt om å velge ut en prøve som de mener smaker ulikt de andre. Finn sannsynligheten for at antall korrekte identifikasjoner av det nye produkt er

   (a) høyst to (b) minst tre (c) minst fire

   under forutsetning av at smaken er uendret og smakerene velger tilfeldig. Hva er da forventet antall korrekte identifikasjoner?
5. Betrakt et tilfeldig n-sifret tall.
   (a) Hva menes med et slikt tall?
   (b) Hva blir forventet antall nuller?
   (c) Hva blir forventet antall ulike sifre?
6. Et produksjonsparti på N artikler inneholder M defekte. Et tilfeldig utvalg på n artikler trekkes ut og antall defekte Y noteres. Lag tabeller over sannsynlighetsfordelingen til Y i følgende situasjoner:
   (a) N=10 M=2 n=2 (b) N=10 M=2 n=3

(c) N=10 M=2 n=4 (d) N=10 M=3 n=2

(e) N=10 M=3 n=3 (f) N=10 M=3 n=4.

Hva er forventet antall defekte i utvalget i hver av disse situasjonene?

7. (a) I en pokerhånd på 5 kort finn forventet antall
   (i) ess (ii) spar (iii) sorte kort

   (b) I en bridgehånd på 13 kort finn forventet antall
   (i) ess (ii) spar (iii) sorte kort (iv) honnører

8. Et produksjonsparti består av N artikler hvorav M er defekte. Det trekkes ut tilfeldig n artikler for kontroll. Bruk tilnærming til binomisk fordeling til å finne tilnærmet sannsynligheten for at utvalget inneholder minst to defekte i følgende situasjoner:

(a)	N=100	M=5	n=5	(b)	N=100	M=5	n=10
(c)	N=100	M=5	n=15	(d)	N=1000	M=50	n=5
(e)	N=1000	M=50	n=10	(f)	N=1000	M=50	n=15.

   I et større tabellverk over hypergeometriske sannsynligheter finner vi følgende eksakte sannsynligheter

   (a) 0.0190 (b) 0.0769 (c) 0.1609.

   Kommenter resultatene ovenfor i lys av dette.

9. La Y være hypergeometrisk fordelt (N, M, n). Vis at vi kan la M og n bytte plass og likevel ha samme fordeling. Forklar hvordan dette kan utnyttes ved binomisk tilnærming for, i visse situasjoner, å oppnå et bedre resultat.

10. Diskuter om en Poissonmodell kan være realistisk i følgende situasjoner:

    (a) oppringninger til et sentralbord,

    (b) emisjon av radioaktive partikler,

    (c) bakterier i drikkevann,

    (d) tidspunkt for "blow out" i oljeboring i Nordsjøen,

    (e) fotballspillere på en bane,

    (f) trykkfeil i en tekst,

    (g) tidspunkt for ulykker på en arbeidsplass,

    (h) tidspunkt for hver personskade på en arbeidsplass,

    (i) solgte biler i gitt tidsrom,

    (j) selvmord i New York City.

11. La X være Poissonfordelt med forventning λ.

    (a) Hva for verdi(er) av X er mest sannsynlig dersom
    (i) λ=0.5 (ii) λ=1.0 (iii) λ=5.0.

(b) Finn sannsynligheten for at X er høyst en dersom
(i) λ=0.5 (ii) λ=1.0 (iii) λ=5.0.

(c) Finn sannsynligheten for at X er minst tre dersom
(i) λ=0.5 (ii) λ=1.0 (iii) λ=5.0.

12. En fabrikant av frokostgryn ønsker at det skal være minst en rosin i gjennomsnittlig 90% av spiseskjeene fra en pakke. Hva må forventet antall rosiner pr. spiseskje være for å oppnå dette? Anta at innholdet av rosiner er planlagt ut fra dette. Hva er da sannsynligheten for at en tilfeldig spise-skje inneholder 5 eller flere rosiner?

13. Anta at Anne har dobbelt så stor uhellsrisiko som Berit i trafikken pr. kjøretime, men kjører bare halvparten så mye. Hvilken rolle spiller dette mht. sannsynlighetsfordelingen for antall årlige uhell?

14. En bilforretning har to bilselgere A og B. A selger gjennomsnittlig en bil i uken, mens B selger gjennomsnittlig to biler i uken. Anta Poissonfordeling og finn sannsynligheten for at i løpet av en gitt uke

    (a) A selger minst to biler

    (b) B selger minst to biler.

    Anta uavhengighet (vurder dette) og finn sannsynligheten for at

    (c) både A og B selger minst to biler

    (d) A og B selger minst fire biler tilsammen

    (e) A og B selger samme antall biler

    (f) B selger flere biler enn A.

15. La X være binomisk fordelt (n, p). Sammenlign de eksakte sannsynligheter med Poissontilnærmelsen i følgende situasjoner

    (a) n=6 p=0.05, 0.10, 0.20

    (b) n=10 p=0.05, 0.10, 0.20

    (c) n=20 p=0.05, 0.10, 0.20

    Eksakte sannsynligheter i (c) krever mer enn Tabell C.2 i Appendiks C.

16. En maskin lager komponenter slik at i det lange løp er gjennomsnittlig 1 av 200 defekte.

    (a) Hva er sannsynligheten for at en eske med 50 komponenter inneholder høyst en defekt.

    (b) Hva er sannsynligheten for, blant 100 esker i en kartong, at ingen eske har mer enn en defekt.

17. En revisjonspopulasjon består av N bilag hvorav M er beheftet med feil. Det trekkes en stikkprøve på n bilag og antall bilag med feil i utvalget Y noteres. I revisjonslitteraturen antas Y ofte å være tilnærmet Poissonfordelt med forventning $\lambda = n \cdot M/N$.

    (a) Under hvilke forutsetninger er dette rimelig?

    (b) Bruk Poissontilnærming til å beregne sannsynligheten for feilfri stikkprøve når $N = 1000, M = 5, n = 100$

    (c) Foreslå en bedre tilnærming (Hint: Se Oppgave 9).

18. Det utføres n uavhengige forsøk, hvert forsøk gir enten suksess eller fiasko. Sannsynlighetene for suksess i forsøkene er ikke de samme, men henholdsvis $p_1, p_2, \ldots, p_n$. La X være antall suksesser. Drøft utsagnene

    (a) Dersom p_i'ene ikke er mye forskjellige vil X være tilnærmet binomisk fordelt $(n, \overline{p})$ der

    $$\overline{p} = (p_1 + p_2 + \cdots + p_n)/n.$$

    (b) Dersom n er stor og alle p_i'ene små vil X være tilnærmet Poissonfordelt med forventning

    $$\lambda = n \cdot \overline{p} = p_1 + p_2 + \cdots + p_n.$$

    Konstruer selv eksempler der disse to forhold kan anvendes.

19. En undersøkelse tyder på at 3% av utestående fordringer av en bestemt type ikke lar seg drive inn. Et firma har 200 slike fordringer. Hva er sannsynligheten for at firmaet ikke klarer å drive inn (i) mer enn 4 (ii) mer enn 6 ?

20. Anta at X er binomisk fordelt (n, p).

    (a) Vis at for den relative hyppighet X/n gjelder at

    $$E(\frac{X}{n}) \quad var(\frac{X}{n}) = \frac{p(1-p)}{n}$$

    Anta at Y er hypergeometrisk fordelt (N, M, n).

    (b) Vis at for den relative hyppighet Y/n gjelder at

    $$E(\frac{Y}{n}) \qquad var(\frac{Y}{n}) = \frac{N-n}{N-1} \cdot \frac{1}{n} \cdot \frac{M}{N} \cdot (1 - \frac{M}{N})$$

    Hva skjer med variansene i (a) og (b) når n vokser?

21. Bruk Tabell C.5 til å beregne følgende

(a) Arealet under normalkurven til venstre for
(i) 1.15 (ii) -0.63 (iii) 0.875 (iv) -1.424.

(b) Arealet under normalkurven til høyre for
(i) 0.72 (ii) -2.15 (iii) 0.537 (iv) -1.325.

(c) Arealet under normalkurven mellom
(i) 0.65 og 2.10 (ii) -0.32 og 1.17 (iii) -1.370 og 0.785.

Gi beste tilnærmede svar for tilfellene med mer enn to desimaler.

22. Bruk Tabell C.5 og C.6 i Appendiks C til å finne z slik at arealet under normalkurven

(a) til venstre for z er 0.7324

(b) til høyre for z er 0.0516

(c) til venstre for z er 0.2578

(d) mellom $-z$ og z er 0.6424

(e) til venstre for z er 0.95

(f) mellom $-z$ og z er 0.50.

Gi beste tilnærmede svar når det spørres etter et areal som ikke direkte kan avleses i tabellen.

23. Betrakt situasjonen med gjentatte terningkast beskrevet i Eksempel 10. Tegn histogrammene til S_n og til den standardiserte variable til S_n i tilfellene (a) n=2 (b) n=3. Kommenter.

24. La X være binomisk fordelt (n, p). Tegn histogrammene til X og til den standardiserte variable til X i tilfellene

(a) $n = 4$ og $p = 0.2$, 0.5

(b) $n = 9$ og $p = 0.2$, 0.5

(c) $n = 16$ og $p = 0.5$.

Kommenter resultatene.

25. Anta at X er binomisk fordelt (n, p). Bruk normaltilnærming til å utføre følgende beregninger:

(a) Beregn $P(X \leq 1)$, $P(X > 3)$ og $P(X = 4)$ når
(i) $n = 10$, $p = 0.5$ (ii) $n = 10$, $p = 0.4$ (iii) $n = 10$, $p = 0.2$.

(b) Beregn $P(X \leq 7)$, $P(X \geq 9)$ og $P(9 < X \leq 12)$ når
(i) $n = 30$, $p = 0.5$ (ii) $n = 30$, $p = 0.4$ (iii) $n = 30$, $p = 0.2$.

(c) Beregn $P(X < 40)$, $P(X > 35)$ og $P(20 \leq X \leq 30)$ når (i) $n = 100$, $p = 0.5$ (ii) $n = 100$, $p = 0.4$ (iii) $n = 100$, $n = 0.2$.

En binomisk tabell gir følgende eksakte sannsynligheter

(a) (i) 0.0108, 0.8281, 0.2051 (ii) 0.0463, 0.6178, 0.2508
(iii) 0.3758, 0.1209, 0.0881

(b) (i) 0.0026, 0.9919, 0.1594 (ii) 0.0435, 0.9060, 0.4021
(iii) 0.7608, 0.1287, 0.0580.

Kommenter de resultater som er funnet ovenfor i lys av dette.

26. La X være binomisk fordelt $(n = 100, p = 0.2)$. Bruk normaltilnærming til å utføre følgende beregninger:

    (a) Finn heltall k slik at $P(X \leq k)$ er tilnærmet lik
    (i) 0.50 (ii) 0.90 (iii) 0.95.

    (b) Finn heltall k slik at $P(X \geq k)$ er tilnærmet lik
    (i) 0.80 (ii) 0.95 (iii) 0.99.

    (c) Finn heltall k slik at $P(\mid X - 20 \mid \leq k)$ er tilnærmet lik
    (i) 0.50 (ii) 0.80 (iii) 0.95.

27. La X være antall kron i n myntkast. Bruk normaltilnærming til å finne ut omtrent hvor stor n må være for at

$$P(\mid \frac{X}{n} - 0.50 \mid < 0.01)$$

skal bli minst lik 0.95.

28. Anta at man ved en pålitelig test har funnet at spiringsevnen til en viss type frø gjennomgående er 95%. Frøene blir solgt i pakker på 200 og det garanteres i reklamen 90% spiring. Hva er sannsynligheten for at en tilfeldig valgt pakke vil ha en spiringsprosent på under 90?

29. I en forbrukerundersøkelse får $n = 100$ deltakere hver smake tre sjokoladebiter hvorav en er med et nytt tilsetningsstoff. Deltakerne blir bedt om å velge ut en av bitene som smaker ulikt de andre. Under forutsetning av at det nye fabrikatet ikke smaker forskjellig fra det gamle, finn sannsynligheten for at antallet som identifiserer det nye produkt er

    (a) minst 30 (b) minst 35 (c) minst 40.

30. En bedrift har et rapportskjema som i 40% av tilfellene er mangelfullt utfylt. Hva er sannsynligheten for at antall mangelfullt utfylte blant 100 skjemaer er (i) mer enn 45 (ii) mellom 45 og 55 ?

31. En professor skal gi en forelesning for et studentkull på $n = 250$ studenter. Han skal dele ut forelesningsnotater, men regner med at sannsynligheten for at en tilfeldig student møter opp er 0.80. Hvor mange eksemplarer må han ta med når han ønsker at sannsynligheten for nok eksemplarer til alle fremmøtte skal være

    (a) minst 0.90 (b) minst 0.95 (c) minst 0.99.

32. La X være antall arbeidsuhell i løpet av et år for en arbeider i en bedrift. Anta at X har sannsynlighetsfordeling

x	0	1	2	3	4
$P(X=x)$	0.50	0.20	0.15	0.10	0.05

(a) Finn EX og $varX$.

La Y være antall år blant de 5 neste som en tilfeldig arbeider går fri uhell.

(b) Finn sannsynlighetsfordelingen til Y, og beregn $P(Y \leq 1)$.

Anta at bedriften har 10 arbeidere. La Z være antall arbeidere som er innblandet i uhell kommende år.

(c) Finn sannsynlighetsfordelingen til Z, og beregn $P(Z \geq 8)$.

La S være antall uhell ved bedriften kommende år.

(d) Finn tilnærmet sannsynligheten $P(S \geq 8)$.

Diskuter de forutsetninger som må gjøres for å løse (b) (c) og (d).

33. Et bilutleiefirma disponerer i alt 3 kombi varebiler som leies ut for en dag av gangen. Antall slike biler S som leies ut pr. dag antas å ha sannsynlighetsfordeling

x	0	1	2	3
$P(X=x)$	0.1	0.2	0.3	0.4

(a) Finn EX og $varX$.

La $X_1, X_2, X_3, \ldots$ være antall utleide biler på etterfølgende dager. Anta uavhengighet (diskuter om dette er realistisk).

(b) Finn sannsynlighetsfordelingen til $X_1 + X_2$ og beregn $P(X_1 + X_2 \leq 4)$.

(c) Finn sannsynligheten for at høyst 55 biler blir utleid i løpet av en måned (30 dager).

34. La Y være en stokastisk variabel som er hypergeometrisk fordelt $(N = 100, M, n)$. Gjør bruk av normaltilnærming til å beregne

(a) $P(Y \leq 4)$, $P(Y \geq 6)$ og $P(Y = 5)$, når $M = 50, n = 10$

(b) $P(Y \leq 1)$, $P(Y \geq 3)$ og $P(Y = 2)$, når $M = 20, n = 10$

(c) $P(Y = 0)$, $P(Y \geq 2)$ og $P(Y = 1)$, når $M = 10, n = 10$

(d) $P(Y \leq 9)$, $P(Y \geq 11)$ og $P(Y = 10)$, når $M = 50, n = 20$

(e) $P(Y \leq 3)$, $P(Y \geq 5)$ og $P(Y = 4)$, når $M = 20, n = 20$

(f) $P(Y \leq 1)$, $P(Y \geq 3)$ og $P(Y = 2)$, når $M = 10, n = 20$

I et større tabellverk for den hypergeometriske fordeling finner vi følgende eksakte sannsynligheter:

(a) 0.3703, 0.3703, 0.2593 (b) 0.3603, 0.3188, 0.3182
(c) 0.3305, 0.2615, 0.4080 (d) 0.4016, 0.4016, 0.1969
(e) 0.3916, 0.3647, 0.2437 (f) 0.3630, 0.3188, 0.3182

Hvilken lærdom kan vi trekke av dette?

35. I en storkommune er det $N = 10\ 000$ stemmeberettigede personer. En avstemning om ølsalg i dagligvareforretningene er nær forestående. For å spå utfallet trekkes et utvalg på n personer som får valget mellom å svare Ja eller Nei. La Y være antall Ja-svar i utvalget. Anta at det i populasjonen i virkeligheten er M Ja-tilhengere og $N - M$ Nei-tilhengere.

   (a) Drøft hvorvidt det er rimelig å anta at Y er hypergeometrisk fordelt.

   Man vil spå at Ja vinner avstemningen dersom $Y/n > 1/2$, at Nei vinner dersom $Y/n < 1/2$.

   (b) Finn sannsynligheten for riktig spådom i situasjonene
   (i) M=5100 n=50, n=100 og n=200
   (ii) M=5200 n=50, n=100 og n=200.

   (c) For en situasjon der $M = 5100$ hvor stor må n være for at sannsynligheten for å spå korrekt skal bli
   (i) minst 0.90 (ii) minst 0.95 (iii) minst 0.99.

   (d) Diskuter eventuelle feilkilder som kan redusere verdien av en slik opinions-undersøkelse med sikte på å forutsi utfallet av avstemningen.

36. La X_1 og X_2 være uavhengige og binomisk fordelt med parametre henholdsvis (n_1, p) og (n_2, p).

   (a) Gi en begrunnelse for at $X_1 + X_2$ er binomisk fordelt $(n_1 + n_2, p)$.

   Et stafettlag i skiskyting består av fire personer som, hva skyteferdigheter angår, vurderes å være like gode. Anta at det i liggende skyting skytes på 5 ballonger med 5 skudd. For hver ballong som ikke er truffet må det gås en strafferunde. Anta at treffsannsynligheten er $p = 0.8$.

   (b) Finn sannsynligheten for at første startende må gå minst en strafferunde.

   (c) Finn sannsynligheten for at de to første startende tilsammen må gå minst to runder.

   (d) Finn sannsynligheten for at hele laget må gå høyst tre strafferunder i liggende skyting.

37. La X_1 og X_2 være uavhengige og Poissonfordelte med forventninger henholdsvis λ_1 og λ_2.

(a) Gi en heuristisk begrunnelse for at $X_1 + X_2$ er Poissonfordelt med forventning $\lambda_1 + \lambda_2$.

En kontordame skriver et manuskript og en regner at hun gjennomsnittlig lager en trykkfeil pr. femte side. Anta derfor at antall feil X på en side er Poissonfordelt med forventning $\lambda = 0.2$.

(b) Finn sannsynligheten for at et manuskript på to sider er feilfritt.

(c) Finn sannsynligheten for at et manuskript på fem sider er feilfritt.

38. Et flyselskap regner med at sannsynligheten for at en person som har forhåndsbestilt plass møter fram ved flyavgang er 0.90. Anta at det brukes fly med 120 seter.

(a) Finn forventet antall ledige plasser i flyet når alle plassene er bestilt.

Anta at det er blitt overbooket med 5%.

(b) Hva er sannsynligheten for at alle frammøtte som har bestilt får plass.

(c) Hva er sannsynligheten for at 5 frammøtte stand-by passasjerer også får plass.

(d) Hvor mange kan høyst overbookes dersom en krever at sannsynligheten for at alle som har bestilt får plass skal være minst 0.99.

Anta at det er overbooket med dette antall.

(e) Hva blir forventet antall ledige plasser i flyet.

Diskuter de antakelser som gjøres ved løsning av oppgaven.

39. En bedrift har 30 maskiner av en bestemt type der en vital del må skiftes ut etter en viss levetid. Det er foreslått et alternativt program for vedlikeholdet av disse maskinene som man håper kan forlenge levetiden. Av de 30 maskinene velges ut tilfeldig 10 som får vedlikehold etter den nye metode, mens de andre får vedlikehold etter den vanlige metoden. Etter en tid noteres det antall maskiner i de to grupper der utskifting har funnet sted siden programmet startet. De nødvendige opplysninger er gitt i tabellen.

	Ikke utskiftet	Utskiftet	Sum
Vanlig metode	x	x	x
Ny metode	x	3	10
Sum	x	12	30

(a) Fyll inn de manglende tall i x-feltene i tabellen.

La Y være antall maskiner blant de som er vedlikeholdt etter ny metode der utskifting har funnet sted. Anta at den nye metoden i virkeligheten ikke forandrer levetiden i forhold til den vanlige metoden slik at det ville være 12 som måtte utskiftes uansett vedlikeholdsmetode.

(b) Forklar at under denne antakelse vil Y være hypergeometrisk fordelt $(N = 30, M = 12, n = 10)$. Finn tilnærmet $P(Y \leq 3)$ og $P(Y = 3)$.

Bedriften ser at forskjellen i utskiftningsprosent er 15% i favør av den nye metoden. Man lurer på om det derfor er grunn til å påstå at den nye metoden er bedre enn den gamle, eller om det observerte resultat med rimelighet kan forklares ved at metodene er likeverdige, og at mange maskiner med lang levetid tilfeldigvis kom i gruppen som fikk vedlikehold etter den nye metode.

(c) Kan en av de sannsynligheter som er funnet under (b) gi en pekepinn om dette, i så fall hvilken.

40. Et akuttmottak ved et sykehus kan uten problemer ta seg av 3 tilfeller pr. time. I gjennomsnitt kommer ett tilfelle pr. time.

    (a) Hva er sannsynligheten for at det kommer (i) høyst 3 (ii) mer enn 3.

    (b) Hva er antakelsen ved beregningen i (a), og hvilke forhold kan gjøre den urealistisk?

41. I et prestegjeld er det gjennomsnittlig 15 dødsfall i året. En teologistudent var sommervikar i 5 uker, og opplevde 7 begravelser. Hvor sannsynlig var dette? Diskuter omstendigheter som kan gjøre beregningen urealistisk og/eller villedende?

42. Sannsynligheten for at en solgt videospiller krever reparasjon som dekkes av garantien er 2%. Anta at slik reparasjon i gjennomsnitt koster leverandøren 200 kroner. Et år selges 2000 spillere. Hva er sannsynligheten for høyst 50 garantireparasjoner? Hvor store kostnader er en 90% sikker på ikke vil overskrides?

43. En bankautomat blir fyllt med 200 000 kroner hver fredag rett før stengetid, og skal helst ikke gå tom i løpet av helgen. Man har erfart at dette skjer i gjennomsnitt hver fjerde uke. Anta at etterspørselen er normalfordelt med forventning μ og standardavvik $\sigma = 0.2 \cdot \mu$.

    (a) Hva er en rimelig verdi på μ under disse forutsetninger?

    (b) Hvor mye må en fylle i dersom en skal bli 99% sikker på å dekke etterspørselen?

44. $\star$En spesialdel med forholdsvis lite forbruk skal produseres bare en gang for lager. Delen skal brukes i et anlegg som antas å bli foreldet og utrangert etter et visst antall år. Når dette vil skje vites ikke med sikkerhet, men det regnes med sannsynligheter på 0.5, 0.3 og 0.2 for at det vil skje etter henholdsvis 5,6 og 7 år. Årlig forbruk ventes å være Poissonfordelt med forventning 1 enhet. Delen koster kr. 400,- og har ingen skrapverdi. Kostnaden dersom lageret tømmes før anlegget er utrangert (stockout) settes til kr. 20 000,-.

    (a) Hvor mye bør produseres for å minimere forventede kostnader?

(b) Hva blir forventet antall ubrukte deler?

45. $\star$Etterspørselen pr. uke (X) etter et tidsskrift i en sentrumskiosk antas å ha sannsynlighetsfordeling

x	0	1	2	3	4	5
$P(X = x)$	0.1	0.2	0.3	0.2	0.1	0.1

(a) Finn EX og $varX$.

Innkjøpsprisen er kr. 10,- og salgsprisen er kr. 15,-. Utsolgte eksemplarer blir returnert, og man får refundert kr. 5,- pr. eksemplar. La Y være bruttofortjenesten pr. uke dersom kiosken kjøper inn q eksemplarer.

(b) Vis at vi kan skrive

$$\begin{aligned} Y &= 10 \cdot X - 5q \quad \text{for} \quad X < q \\ &= 5 \cdot q \qquad\quad \text{for} \quad X \geq q. \end{aligned}$$

(c) Finn EY for tilfellene $q = 0, 1, 2, \ldots, 5$.

(d) Vis at det innkjøpskvantum som gir størst forventet bruttofortjeneste er $q = 2$.

Kiosken vurderer at den lider et goodwill tap på kr. 2,- for hver kunde som ikke får dekket behovet.

(e) Vis at dersom dette tas med i analysen er $q = 3$ optimalt.

La Z være antall ganger i løpet av et år (52 uker) at kiosken blir utsolgt.

(f) Hva blir sannsynlighetsfordelingen til Z når $q = 3$ er innkjøpt kvantum hver uke.

(g) Finn tilnærmet sannsynligheten for at kiosken er utsolgt minst 20 av ukene i et år.

46. $\star$To maskiner A og B er nyjustert på mandag. Ventetiden til neste justering antas å være uavhengige geometrisk fordelte variable med forventning henholdsvis 2 og 3 dager. Finn sannsynligheten for at

(a) A trenger justering i løpet av uken (5 dager),

(b) B trenger justering i løpet av uken,

(c) begge trenger justering i løpet av uken,

(d) minst en maskin trenger justering i løpet av uken,

(e) neste justering for begge maskiner skjer på samme dag.

(f) neste justering for B skjer før A.

47. $\star$En bedrift produserer en artikkel som sorteres i tre kvaliteter A_1, A_2 og A_3. Man har erfart at produksjonen gir i gjennomsnitt 60% av kvalitet A_1, 30% av kvalitet A_2 og 10% av kvalitet A_3. I en dagsproduksjon på $n = 15$ artikler la antallet i de tre kvalitetskategorier være h.h.v. X_1, X_2 og X_3.

    (a) Skriv ned den simultane fordelingsfunksjon $p(x_1, x_2, x_3)$ til X_1, X_2 og X_3 og beregn $p(9, 4, 2)$.

    (b) Finn $P(X_1 \geq 10)$ og $P(X_1 + X_2 \geq 13)$.

    (c) Finn også $P(X_1 \geq 10 \mid X_1 + X_2 = 13)$.

48. $\star$I en situasjon hvor $n = 10$ personer skal velge den av tre smaksprøver som de liker best, finn sannsynligheten for at minst 5 personer foretrekker en og samme prøve når prøvene i virkeligheten er identiske.

49. $\star$En produksjonsprosess gir gjennomstående 85% feilfrie artikler, 10% av artiklene kan repareres, mens 5% må vrakes. I en dagsproduksjon på $n = 100$ artikler la antall som repareres være X_1 og antall som vrakes være X_2. Anta at kostnadene ved å reparere en artikkel er 15 kroner, mens tapet ved å vrake en artikkel er 100 kroner. La L være tapet som skyldes defekter i dagsproduksjonen.

    (a) Utrykk L ved X_1 og X_2 og beregn EL og $varL$.

    (b) Bruk normaltilnærming til å beregne sannsynligheten for at tapet som skyldes defekter i dagsproduksjonen overstiger 1000 kroner.

50. $\star$Et vurderingssystem for styrken av en korthånd i bridge går ut på følgende: Hvert ess teller fire poeng, hver konge tre poeng, hver dame to poeng og hver knekt teller ett poeng, alle andre kort teller null. Et kort velges tilfeldig, og la X være poengverdien av dette.

    (a) Finn sannsynlighetsfordelingen til X og beregn EX og $varX$.

    En korthånd på tretten kort velges tilfeldig, og la S være sum poeng på hånden.

    (b) Beregn ES og $varS$.

    (c) Beregn eksakt eller tilnærmet sannsynligheten for at hånden har tretten poeng eller mer (åpning).

51. Programvare gir muligheter for beregning av sannsynligheter og kumulative sannsynligheter for de vanligste fordelinger, herunder binomisk, Poisson og normal. Eksempler på dette er

```
>> pdf 5; binomial 10 0.5.
    x   P( X = x)
 5.00     0.2461
>> cdf 5; binomial 10 0.5.
    x   P( X <= x)
 5.00      0.6230
>> pdf 2; Poisson 1.
    x   P( X = x)
 2.00     0.1839
>> cdf 2; Poisson 1.
    x   P( X <= x)
 2.00      0.9197
>> cdf 3; normal 0 1
      x   P( X <= x)
 3.0000      0.9987
>> invcdf 0.95; normal 0 1.
 P( X <= x)       x
     0.9500   1.6449
```

Hva er det som er beregnet her? Hva betyr PDF, CDF og INVCDF? Se om du kan reprodusere resultatene med din programvare. Se også om du kan simulere observasjoner fra de samme fordelingene.

Kapittel 7

Statistisk inferens: Sannsynligheter og andeler

7.1 Innledning

En stokastisk modell som inneholder nok antakelser til at vi kan beregne sannsynligheten for enhver begivenhet i modellen, kalles en *helspesifisert modell.* De modeller vi har studert ovenfor i kapitlene om sannsynlighetsregning har alle vært helspesifiserte. I enkelte situasjoner kan vi være villige til å gjøre en del antakelser, men ikke nok til at modellen blir helspesifisert. Vi taler da om en *delvis spesifisert modell.* I praksis støter vi ofte på problemstillinger hvor vi, med utgangspunkt i en delvis spesifisert modell, ønsker å utføre det eksperimentet som modellen er ment å beskrive, for om mulig å finne ut noe om de uspesifiserte elementene i modellen.

Eksempel 1 : Produksjonsprosess
En produksjonsprosess skal produsere $n = 100$ artikler. Vi er villige til å gjøre følgende antakelser om prosessen:

1. Hver produsert artikkel blir enten defekt eller intakt.
2. Sannsynligheten p for defekt er den samme for alle artikler.
3. Uavhengighet.

Dvs. at vi har en binomisk forsøksrekke. Dersom vi, f. eks. ut fra tidligere erfaring, er villige til å anta at $p = 0.10$, har vi en helspesifisert modell. Det er da et rent sannsynlighetsteoretisk problem å beregne sannsynligheten for en hvilken som helst begivenhet som måtte interessere oss. Lar vi for eksempel X betegne antall defekte artikler, vet vi at

$$X \text{ er binomisk fordelt } (n = 100, p = 0.10).$$

Uten noen slik antakelse om p, har vi bare en delvis spesifisert modell. Vi har da ikke nok antakelser til å kunne beregne sannsynligheter. Nå er

$$X \text{ binomisk fordelt } (n = 100, p)$$

hvor p er uspesifisert, dvs. (mer eller mindre) ukjent. Vi kan da ønske oss nærmere opplysninger om p. For dette formål kan vi gjennomføre produksjonen av de $n = 100$ artiklene. På basis av den delvis spesifiserte modellen og det observerte produksjonsresultat, kan vi kanskje gi et brukbart anslag for p (*estimering*). Kan hende er problemstillingen den at man ønsker å gå over til en ny produksjonsmetode. For den tradisjonelle produksjonsmetoden er erfaringsvis $p = 0.10$, mens for den nye metoden er p ukjent. På grunnlag av en prøveproduksjon på $n = 100$ artikler ønsker man å fastslå om det er noen grunn til å påstå at den nye metoden gir bedre kvalitet enn den gamle, dvs. påstå at $p < 0.10$ (*hypotesetesting*). Slike problemer er ikke lenger rene sannsynlighetsteoretiske problemer, de er såkalte *statistiske inferensproblemer*.

Eksempel 2 : Utvalgsundersøkelse
En forening har N =1000 medlemmer hvorav M menn og $N - M$ kvinner. Det trekkes et tilfeldig utvalg på n=100 medlemskort. Lar vi Y betegne antall menn i utvalget, vet vi at

$$Y \text{ er hypergeometrisk fordelt } (N = 1000, M, n = 100).$$

Dersom antall mannlige medlemmer er kjent, f.eks. M=500, er modellen helspesifisert, og sannsynligheten for begivenheter vedrørende utvalget kan beregnes. For eksempel kan vi beregne sjansen for at det er flere menn enn kvinner i utvalget. I fall M er ukjent, har vi bare en delvis spesifisert modell. Formålet med å trekke et tilfeldig utvalg av medlemskort er kanskje nettopp å finne ut noe om M, f.eks. klarlegge om det er flere menn enn kvinner i foreningen. Vi har da ikke lenger et rent sannsynlighetsteoretisk problem, men et statistisk inferensproblem.

Vi kan si at sannsynlighetsregning og statistisk inferens er komplementære disipliner:

Sannsynlighetsregning: Med utgangspunkt i en helspesifisert modell for et eksperiment kan vi, før eksperimentet utføres, beregne sannsynligheter for begivenheter vedrørende eksperimentet.

Statistisk inferens: På bakgrunn av en delvis spesifisert modell for et eksperiment kan vi, på basis av det observerte utfall av eksperimentet, si noe om de uspesifiserte elementer i modellen.

Selv om vi i dette kapitlet i hovedsak er interessert i statistisk inferens om sannsynligheter (Eksempel 1) eller andeler (Eksempel 2), vil vi innføre en del generell teori som også danner en vesentlig del av grunnlaget for neste kapittel.

7.2 Estimering

Anta at vi har en delvis spesifisert modell for et eksperiment, der det uspesifiserte element vi er interessert i kan uttrykkes ved en parameter θ hvis verdi er ukjent. I Eksempel 1 kan f.eks. sannsynligheten p være den ukjente parameter, dvs. $\theta = p$. I Eksempel 2 kan det f.eks. være andelen a av menn i populasjonen, dvs. $\theta = a = M/N$. Vi ønsker på grunnlag av det observerte utfall av eksperimentet å komme med et anslag, i fagterminologien ofte kalt et *estimat*, for den ukjente parameteren θ. I Eksempel 1 virker det rimelig å bruke den relative hyppighet av defekte i prøveproduksjonen som estimat for den ukjente defektsannsynligheten. I Eksempel 2 vil andelen menn i utvalget være et rimelig estimat for andelen menn i populasjonen. Et slikt estimat kan treffe mer eller mindre godt, og vi kan ikke på forhånd være sikker på hvor godt, estimater vil vanligvis være underlagt tilfeldigheter. For å illustrere dette poenget ser vi på et lett gjennomskuelig eksempel.

Eksempel 3 : Myntkast
Finn et kronestykke og forsøk å anslå sannsynligheten p for at et myntkast gir kron på grunnlag av $n = 10$ myntkast (glem for en stund at du vet at Norges Bank leverer mynter med $p = 1/2$). Mitt kronestykke ga bare 3 kron, og med denne opplysningen alene velger jeg å estimere kronsannsynligheten med den relative hyppigheten av kron, dvs. med $3/10 = 0.3$. I denne situasjonen er det rimelig å anta at antall kron er binomisk fordelt $(n = 10, p)$, og i fall

mynten er rettferdig var sjansen for at nettopp denne estimerte verdi skulle slå til lik ca. 0.117. De mulige verdier med tilhørende sannsynligheter finner vi av Tabell C.3:

Estimert verdi:	0.0	0.1	0.2	0.3	0.4	0.5	0.6	...
Sannsynlighet:	0.001	0.010	0.044	0.117	0.205	0.246	0.205	...

Når mynten er rettferdig ($p = 0.5$) vil 0.5 riktignok være den estimerte verdi som har størst sjanse for å slå til, men risikoen for store feilanslag er betydelig. Tilsvarende ser vi av Tabell C.2 at dersom vi hadde en falsk mynt med kronsannsynlighet p=0.4, vil 0.4 være det mest sannsynlige estimat, med omtrent samme risiko for feilanslag. For generelt å redusere risikoen for feilanslag må vi selvfølgelig utføre et større antall myntkast. Undersøker vi f.eks. tilfellet $n = 100$ vil, dersom mynten er rettferdig, fortsatt 0.5 være det estimat som har størst sannsynlighet, mens sannsynligheten for å få et estimat som avviker mye er redusert. Sannsynligheten for at estimert verdi blir akkurat 0.5 er riktignok bare 0.08, men i tilfellet $n = 100$ er det langt flere mulige verdier omkring 0.5. Eksempelvis blir sannsynligheten for at anslaget faller i intervallet fra 0.45 til 0.55 lik 0.73 (se også Oppgave 6.24). Samme vurderinger kan gjøres dersom mynten er falsk.

Definisjon : En *estimator* $\hat{\theta}$ for en ukjent parameter θ er en stokastisk variabel, der realisert verdi kan brukes som estimat (anslag) for θ.

Et estimat for parameteren θ er brukbart når det er nær den faktiske verdi av θ. Vi bør derfor velge en estimator $\hat{\theta}$ som har stor sannsynlighet for å gi et estimat nær θ, uansett hvilken verdi θ måtte finne på å ha. Dette innebærer at sannsynlighetsfordelingen (ofte kalt *sampling-fordelingen*) til $\hat{\theta}$ bør være konsentrert omkring den ukjente θ. Vi har flere muligheter for å uttrykke et slikt ønske, eksempelvis

1. $E \mid \hat{\theta} - \theta \mid$ bør være liten
2. $E(\hat{\theta} - \theta)^2$ bør være liten
3. $P(\mid \hat{\theta} - \theta \mid < d)$ bør være stor.

Her vil vi kalle $\mid \hat{\theta} - \theta \mid$ for estimeringsfeilen. Nr. 1 ønsker seg liten forventet estimeringsfeil, mens nr. 2 ønsker seg liten forventet kvadrert estimeringsfeil.

Nr. 3 ønsker, for gitt $d > 0$, at sannsynligheten for estimeringsfeil mindre enn d skal være stor. I praksis viser det seg at det første kriteriet er lite velegnet, og vi skal derfor studere de to siste kriteriene. Følgende identitet er viktig (se Oppgave 14).

$$E(\hat{\theta} - \theta)^2 = var\hat{\theta} + (E\hat{\theta} - \theta)^2$$

Vi ser at om forventet kvadratfeil $E(\hat{\theta} - \theta)^2$ skal bli liten, må følgende to betingelser være oppfylt:

1. $E\hat{\theta}$ må avvike lite fra θ.
2. $var\hat{\theta}$ må være liten.

I mange situasjoner er det rimelig å velge en estimator $\hat{\theta}$ som er slik at $E\hat{\theta} = \theta$, og vi innfører derfor følgende begrep:

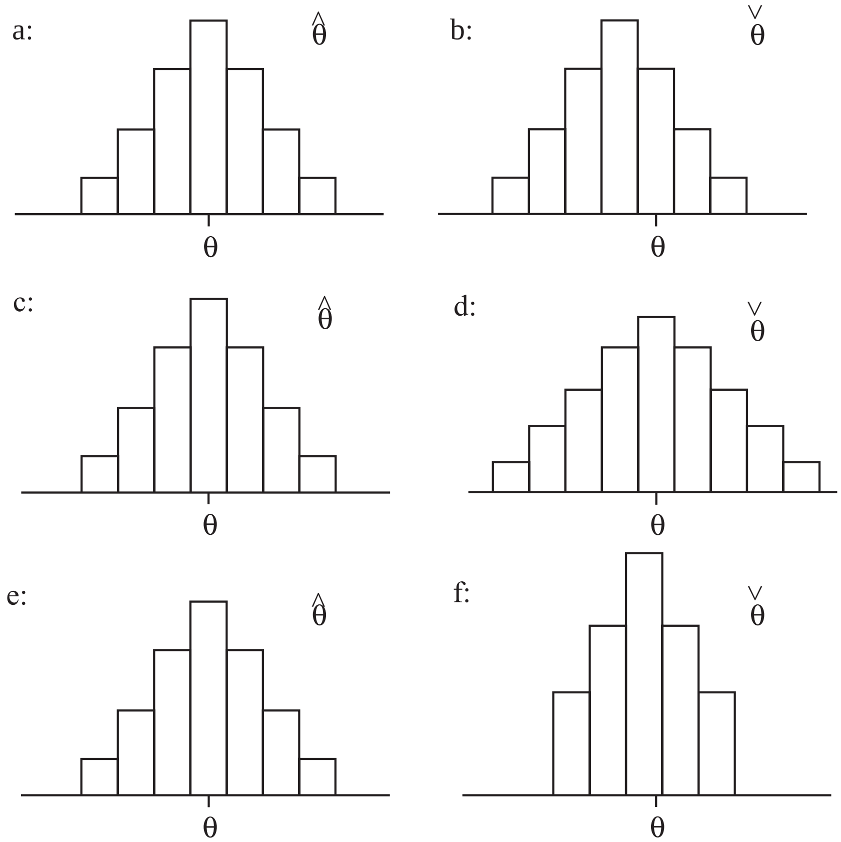
> **Definisjon :** En estimator $\hat{\theta}$ sies å være *forventningsrett* for θ dersom $E\hat{\theta} = \theta$ uansett θ. En estimator som ikke er forventningsrett vil vi kalle *forventningsskjev*.

Vi ser at med en forventningsrett estimator $\hat{\theta}$, blir $E(\hat{\theta} - \theta)^2 = var\hat{\theta}$, og en sammenligning av to eller flere forventningsrette estimatorer kan derfor skje ved å sammenligne deres varianser. Vi velger da den estimator som har minst varians.

La oss illustrere betraktningene ovenfor med eksempler: Anta at vi har valget mellom to estimatorer kalt $\hat{\theta}$ og $\check{\theta}$ med sannsynlighetsfordelinger (samplingsfordelinger) som på Figur 7.1.
I Figur 7.1a–b er $\hat{\theta}$ forventningsrett, mens $\check{\theta}$ er forventningsskjev. Siden $\hat{\theta}$ og $\check{\theta}$ har samme varians, vil vi foretrekke $\hat{\theta}$ framfor $\check{\theta}$. I Figur 7.1c–d er $\hat{\theta}$ og $\check{\theta}$ begge forventningsrette, men $\hat{\theta}$ foretrekkes framfor $\check{\theta}$ fordi $\hat{\theta}$ har mindre varians. I Figur 7.1e–f er $\hat{\theta}$ forventningsrett, mens $\check{\theta}$ ser ut til å være forventningsskjev. Allikevel kan det tenkes at $\check{\theta}$ er å foretrekke framfor $\hat{\theta}$, fordi den systematiske skjevheten muligens er oppveid av en mindre varians.

I svært mange situasjoner vil det være estimatorer som naturlig byr seg fram, og som viser seg å være forventningsrette.

Eksempel 4 : Binomisk situasjon
La X være antall defekte artikler i en produksjon på n artikler. Med de antakelser vi gjorde i Eksempel 1 fikk vi at



Figur 7.1: Samplingfordelinger til estimatorer $\hat{\theta}$ og $\check{\theta}$

$$X \quad \text{er binomisk fordelt } (n, p)$$

der sannsynligheten p for at en artikkel er defekt antas å være ukjent. Vi ønsker å estimere p. En rimelig estimator for p vil være

$$\hat{p} = \frac{X}{n}$$

dvs. hyppigheten av defekte blant de n produserte artiklene. Vi finner

$$E(\hat{p}) = E(\frac{X}{n}) = \frac{1}{n}EX = \frac{1}{n}np = p$$

slik at estimatoren $\hat{p}$ er forventningsrett. Videre blir

$$var(\hat{p}) = var(\frac{X}{n}) = (\frac{1}{n})^2 varX = (\frac{1}{n})^2 np(1-p) = \frac{p(1-p)}{n}$$

Vi ser at variansen til $\hat{p}$ avtar med n, og at usikkerheten ved å estimere p med $\hat{p}$ vil kunne gjøres så liten man ønsker ved å velge n tilstrekkelig stor.

Eksempel 5 : Hypergeometrisk situasjon
La oss igjen studere problemstillingen fra Eksempel 2 med en forening på N =1000 medlemmer, hvorav M er menn og $N-M$ kvinner. La Y = antall menn i et utvalg på 100 medlemskort. Vi hadde at Y er hypergeometrisk fordelt ($N = 1000, M, n = 100$) hvor M er ukjent. Vi ønsker å estimere andelen av menn i foreningen $a = M/N$. En rimelig estimator er

$$\hat{a} = \frac{Y}{n} = \text{ andelen av menn i utvalget.}$$

Her blir

$$E(\hat{a}) = E(\frac{Y}{n}) = \frac{1}{n}EY = \frac{1}{n}n\frac{M}{N} = \frac{M}{N} = a$$

slik at estimatoren er forventningsrett. Vi får videre

$$\begin{aligned} var(\hat{a}) &= var(\frac{Y}{n}) = (\frac{1}{n})^2 varY = (\frac{1}{n})^2 \cdot \frac{N-n}{N-1} \cdot n\frac{M}{N} \cdot (1 - \frac{M}{N}) \\ &= \frac{N-n}{N-1} \cdot \frac{1}{n} \cdot \frac{M}{N}(1-\frac{M}{N}) = \frac{N-n}{N-1} \cdot \frac{1}{n}a(1-a) \end{aligned}$$

Vi ser at variansen avtar med n. Når $n = N$ blir variansen null, vi har da undersøkt alle medlemskort og oppnådd full sikkerhet.

7.3 Rapportering, tolking og planlegging

Vi skal estimere en parameter θ. Anta at vi som estimeringsmetode har valgt en forventningsrett estimator $\hat{\theta}$. Som vårt estimat for θ rapporterer vi observert verdi av $\hat{\theta}$. Vi bør også rapportere et mål for påliteligheten av metoden. En mulighet er å oppgi variansen til estimatoren $\hat{\theta}$, eller alternativt standardavviket $\sigma(\hat{\theta})$ (kvadratroten av variansen). Her vil vi foretrekke å rapportere standardavviket, bl.a. fordi dette har samme dimensjon som parameteren vi skal estimere. Det er vanlig å rapportere slik:

$$\text{estimat} \pm \text{standardavvik til estimatoren.}$$

Standardavviket til estimatoren blir ofte kalt *standardfeilen* og rapportmetoden *standardfeilmetoden*. Standardfeilen gir en indikasjon om i hvilken grad vi risikerer at det funne estimat avviker fra den verdi vi ønsker estimert. Estimater med tilhørende standardavvik (standardfeil) forekommer ofte rutinemessig i vitenskapelige arbeider og utredninger, og en leser med litt innsikt i statistisk teori vil kunne vurdere utsagnskraften i resultatene.

I situasjoner der sannsynlighetsfordelingen til estimatoren $\hat{\theta}$ tilnærmes med normalkurven, vil standardavviket $\sigma(\hat{\theta})$ være spesielt nyttig. Vi har da ifølge resultatene i Kapittel 6.5

$$P(\mid \hat{\theta} - \theta \mid < d) \approx A(\frac{d}{\sigma(\hat{\theta})})$$

der $A(k)$ er arealet under normalkurven mellom $-k$ og k.

Vi ser at kjennskap til standardavviket $\sigma(\hat{\theta})$ setter oss i stand til å beregne (tilnærmet) sannsynligheten for at estimeringsfeilen $\mid \hat{\theta} - \theta \mid$ er mindre enn d enheter, for enhver valgt $d > 0$. Formelen kan alternativt skrives

$$P(\mid \hat{\theta} - \theta \mid < k \cdot \sigma(\hat{\theta})) \approx A(k)$$

slik at sannsynligheten for en estimeringsfeil på mindre enn k ganger standardavviket er entydig gitt ved k. Ut fra en tabell over $G(k)$ (arealet under normalkurven til venstre for k) kan vi lage følgende tabell over $A(k) = 2G(k) - 1$ for utvalgte verdier av k

k	1.0	1.645	1.96	2.0	2.58	3.0
$G(k)$	0.8413	0.95	0.975	0.9772	0.995	0.9987
$A(k)$	0.6826	0.90	0.95	0.9544	0.99	0.9974

Vi ser at sannsynligheten for estimeringsfeil på mindre enn 1 $\times$ standardavviket er ca. 0.68, at sannsynligheten for estimeringsfeil på mindre enn 2 $\times$ standardavviket er ca. 0.95, mens sannsynligheten for estimeringsfeil på mindre enn 3 $\times$ standardavviket er svært nær 1.00. For senere bruk har vi også gitt de verdier av k som gir eksakt sannsynlighetene 0.90, 0.95 og 0.99.

De generelle betraktningene ovenfor kan brukes i en rekke ulike praktiske situasjoner, så som den binomiske og den hypergeometriske situasjon. I noen av disse er det mulig å beregne eksakt sannsynlighetene for estimeringsfeil av en gitt størrelse, men det er ofte ikke bryet verdt å gjøre dette, da vi som oftest bare er interessert i et grovt anslag (størrelsesorden) på sjansen for estimeringsfeil. I andre situasjoner hvor normaltilnærmelse ikke er rimelig, vil rapportering av standardavviket likevel være informativt, idet man via Tsjebysjeffs ulikhet får nedre skranker på sjansene for estimeringsfeil på mindre enn k ganger standardavviket (se Kapittel 6.6).

I Eksemplene 4 og 5 så vi at standardavviket til estimatoren avhang av den parameter vi skulle estimere, og dette er typisk for svært mange situasjoner. Vi må derfor ofte nøye oss med å anslå dette standardavviket. Noen ganger vil det være mulig å utnytte erfaring om usikkerheten i lignende situasjoner fra tidligere (prospektivt anslag), men som regel er vi henvist til å estimere standardavviket $\sigma(\hat{\theta})$ utfra de foreliggende observasjoner (retrospektivt anslag). I slike situasjoner vil det være naturlig å rapportere

$$\text{estimat} \pm \text{estimert standardavvik for estimatoren}$$

Vi kunne, i analogi med ovenfor, ønske å tolke dette med sannsynlighetsutsagn av typen: Sannsynligheten for estimeringsfeil på mindre enn $1 \times$ estimert standardavvik er ..., sannsynligheten for estimeringsfeil på mindre enn $2 \times$ estimert standardavvik er ... etc. La $S(\hat{\theta})$ være den valgte estimator for standardavviket $\sigma(\hat{\theta})$. Generelt ønsker vi derfor å kunne beregne

$$P(\mid \hat{\theta} - \theta \mid < k \cdot S(\hat{\theta}))$$

i det minste med brukbar tilnærmelse. Siden det ligger en viss risiko i å erstatte $\sigma(\hat{\theta})$ med $S(\hat{\theta})$, må vi vente at denne sannsynligheten, for gitt k, blir noe mindre enn den vi hadde når $\sigma(\hat{\theta})$ var kjent, nemlig $A(k)$. I situasjoner der vi estimerer på grunnlag av et stort antall observasjoner, vil $S(\hat{\theta})$ gi et såpass presist anslag for $\sigma(\hat{\theta})$ at det ikke spiller noen vesentlig rolle om $\sigma(\hat{\theta})$ blir erstattet med et estimat. Da kan tilnærmelsen $A(k)$ fortsatt gi en pekepinn om faren for estimeringsfeil av størrelse $k \times$ (estimert) standardavvik. På den annen side finnes viktige praktiske problemstillinger der vi kan beregne sjansene for estimeringsfeil mer nøyaktig, ja endog helt eksakt, også når standardavviket er estimert og selv med få observasjoner (se Kapittel 8.4).

Eksempel 6 : Produksjonsprosess

La situasjonen være som i Eksempel 1. Det prøveproduseres $n = 100$ artikler etter en ny produksjonsmetode. Antall defekte artikler X antar vi er binomisk fordelt $(n = 100, p)$. Vi velger $\hat{p} = X/n$ som estimator for defektsannsynligheten p. Standardavviket til denne estimatoren er

$$\sigma(\hat{p}) = \sqrt{\frac{p(1-p)}{n}}$$

Her er p den ukjente som skal estimeres, og standardavviket kan derfor ikke beregnes eksakt. Anta at vi har erfart at den gamle produksjonsmetoden gjennomgående i det lange løp ga 10% defekte, svarende til $p = 0.10$ og et standardavvik på $\sqrt{0.10 \cdot 0.90/100} = 0.03$. Anta at vi er overbevist om at den nye metoden i hvert fall ikke er dårligere, dvs. ikke har større p-verdi enn den gamle. Nå ser vi at en mindre p-verdi enn 0.10 også betyr mindre standardavvik, og 0.03 vil derfor være et pessimistisk anslag for standardavviket ved den nye produksjonsmetoden. La oss tenke oss at vi blant de 100 produserte artiklene etter den nye metoden fant 5 defekte. Vårt estimat for p blir derfor $5/100 = 0.05$. Det synes da rimelig å rapportere defektsannsynligheten slik

$$0.05 \quad \pm \quad 0.03.$$

Siden tilnærming til normalkurven kan brukes i binomiske situasjoner (merk at her blir $np(1-p)$ lik 9.0 i tilfellet p=0.10 og lik 4.75 i tilfellet p=0.05 dvs. noenlunde akseptabelt ifølge vår tommelfingerregel), vil det være naturlig å tolke dette slik: Det rapporterte standardavvik 0.03 forteller at estimeringsmetoden garanterer ca.:

68% sjanse for at estimeringsfeilen er mindre enn 0.03,
95% sjanse for at estimeringsfeilen er mindre enn 0.06.

Disse sannsynlighetene er bare omtrentlig siden de er basert på tilnærming til normalkurven. Eksakte sannsynligheter kan finnes ved hjelp av binomiske tabeller, men de avviker ikke nevneverdig fra de som er angitt her. Siden defektprosenten i prøveproduksjonen med den nye metode ble 5%, dvs. halvparten av den vi kjenner for den gamle metoden, kan vi spørre om vi dermed har "bevist" at den nye produksjonen gir bedre resultat enn den gamle. Det rapporterte standardavvik forteller at en estimeringsfeil på mer enn 0.06 ikke er utenkelig (5% sjanse), og det kan derfor godt tenkes at p fortsatt er 0.10 og det observerte resultat skyldes tilfeldigheter. Dette er spørsmål som vi vil få belyst fra en annen synsvinkel i et senere avsnitt om hypotesetesting (se også Oppgave 26).

La oss isteden tenke oss at situasjonen var slik at vi ikke har noen forhåndsinformasjon om defekt-sannsynligheten ved den nye produksjonsmetoden. Rapporteringen av standardavviket $\sigma(\hat{p}) = \sqrt{p(1-p)/100}$ til den brukte estimator skaper nå visse problemer. Vi kan selvfølgelig være veldig pessimistiske og rapportere den største verdi som kan oppnås, det er 0.05, som inntreffer for $p = 0.5$ (se Eksempel 9), slik at defektprosenten blir rapportert som

$$\text{estimat} \ \pm \ 0.05.$$

Anta at vi observerte 4 defekte blant de 100 prøveproduserte artiklene, dvs. at vårt estimat blir $4/100 = 0.04$. Det ser ut til at den nye produksjonsmetoden gir en lav defektprosent svarende til et mye lavere standardavvik enn det rapporterte, og vi føler at rapporten her ikke bringer fram den utsagnskraft som ligger i det observerte resultat (eksempelvis kan p aldri bli negativ). En mulighet vil være å anslå $\sigma(X/100)$ på grunnlag av det observerte resultat ved å erstatte p i formelen med det oppnådde estimat for p. Vi får da $\sqrt{0.04 \cdot 0.96/100}$ og det synes rimelig å rapportere

$$0.04 \ \pm \ 0.02.$$

Tilnærming til normalkurven vil nå lede til følgende sannsynlighetsutsagn:

68% sjanse for estimeringsfeil mindre enn 0.02 etc.

Vi bør imidlertid være litt varsomme her. Det er en viss fare for at den underliggende p er så liten at normaltilnærming er betenkelig, spesielt når vi også har erstattet standardavviket med et anslag forbundet med en viss usikkerhet. Slike betenkeligheter ville ikke oppstå i samme grad dersom det observerte antall defekte var noe høyere (hvorfor?) eller dersom antall prøveproduserte artikler var større (hvorfor?).

Eksempel 7 : Meningsmåling
I Lillevik er det ikke Vinmonopol. En del tilhengere av pol vil foreta et utvalg blant de stemmeberettigede for å lodde stemningen i byen. Dersom denne er gunstig vil de fremme forslag om en folkeavstemning. Anta at det er N=10 000 stemmeberettigede hvorav M er for pol, mens $N - M$ er mot pol. Man ønsker å estimere andelen av tilhengere $a = M/N$. Det velges ut tilfeldig $n = 100$ personer, og blant disse viser det seg at antall tilhengere Y er lik 56. Vi er i en hypergeometrisk situasjon. Estimatoren $\hat{a} = Y/n$ har standardavvik

$$\sigma(\hat{a}) = \sqrt{\frac{N-n}{N-1} \cdot \frac{1}{n} \cdot a(1-a)}.$$

Her er a ukjent, men fra tidligere vet man at brøkdelen a er ca. 0.5. Dette vil gi et standardavvik på ca. 0.05. Vårt estimat for a blir 0.56, med denne verdi innsatt i formelen for standardavviket får vi også ca. 0.05 (funksjonen $a(1-a)$ viser liten variasjon omkring $a = 0.5$). Vi kan derfor trygt rapportere

$$0.56 \quad \pm \quad 0.05.$$

Siden n er stor, men likevel liten i forhold til N, kan vi tolke denne rapporten i lys av normaltilnærmelsen:

68% sjanse for estimeringsfeil på mindre enn 0.05 etc.

Man bør derfor ikke føle seg helt trygg på at flertallet av de stemmeberettigede er for pol. La oss tenke oss en situasjon med like mange tilhengere som motstandere, dvs. M=5 000. Sannsynligheten for å få det observerte resultat eller et som går enda mer i favør av tilhengerne blir da:

$$\begin{aligned} P(Y \geq 56) &= 1 - P(Y \leq 55) \approx 1 - G(\frac{55 + 0.5 - 50}{5}) \\ &= 1 - G(1.1) = 1 - 0.8643 = 0.1357 \end{aligned}$$

Det observerte resultat vil derfor, selv under forutsetning om at de to grupper er like sterke, ikke være oppsiktsvekkende.

Eksempel 8 : Ulykkesrisiko

La oss studere antall nordmenn som rammes av en viss type sjelden ulykke i løpet av et år. Vi kjenner ikke antall personer under risiko, men tror at det er noenlunde konstant fra år til år. Antall som rammes et bestemt år antar vi er en stokastisk variabel X som er Poissonfordelt med forventning λ (jfr. argumentene i Kapittel 6.4). Brukes X som estimator for λ, blir dens standardavvik $\sigma(X) = \sqrt{\lambda}$, som lar seg estimere med $\sqrt{X}$. Dersom vi et år observerte 9 ulykker, anslås standardavviket til 3, og vi rapporterer at forventet antall ulykker i løpet av et år, under de rådende forhold, er

$$9 \quad \pm \quad 3.$$

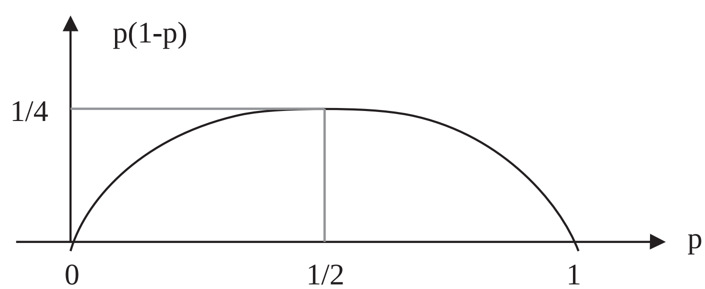
Igjen kan vi bruke 68% - 95%-regelen basert på normaltilnærming (jfr. Eksempel 6.13) for å få et grovt intrykk av faren for feilkonklusjoner. Vi ser at konklusjoner angående ulykkesrisikoen basert på kun dette året vil være nokså usikre. Dersom vi vet at gjennomsnittlig antall ulykker over en del år har vært 5 pr. år, vil siste år med nesten dobbelt så mange ulykker, neppe gi grunnlag for å hevde en generell økning av ulykkesrisikoen.

En annen grunn til å interessere seg for standardavviket til estimatorer ligger i muligheten for å planlegge vår statistiske undersøkelse, slik at vi oppnår et på forhånd ønsket presisjonsnivå.

Eksempel 9 : Binomisk situasjon
Anta at X er binomisk fordelt (n, p). Vi ønsker å estimere p ved bruk av estimatoren $\hat{p} = X/n$. Anta at vi står fritt til å velge n slik at estimatoren får en ønsket presisjon uttrykt ved et ønsket standardavvik Δ. For den binomiske situasjon betyr det at $var\hat{p} = p(1-p)/n = \Delta^2$. Følgelig må

$$n = \frac{p(1-p)}{\Delta^2}$$

slik at, for gitt Δ, vil den korrekte n være avhengig av p via proporsjonalitetsfaktoren $p(1-p)$. Det er instruktivt å tegne en figur som gir faktoren $p(1-p)$ som funksjon av p i intervallet [0,1].



Figur 7.2: Funksjonen $p(1-p)$

Vi ser at $p(1-p) = 1/4-(p-1/2)^2$, slik at funksjonen oppnår sin største verdi 1/4 for $p = 0.5$. Derfor er alltid $n \leq 1/(4\Delta^2)$. Velger vi antall observasjoner $n = 1/(4\Delta^2)$, ser vi at vi er garantert det ønskede presisjonsnivå uansett hva p måtte være. Eksempelvis dersom vi ønsker $\Delta = 0.01$, får vi $n = 1/(4\Delta^2) = 2500$.

I mange situasjoner i praksis har en imidlertid en viss peiling på størrelsen av p på forhånd. Dersom vi, som i Eksempel 6, studerer antall defekte X i en prøveproduksjon og vi vet fra før at slik produksjon gir om lag 10% defekte, er det rimelig å velge n ved å bruke formelen ovenfor med $p = 0.10$. Dersom vi ønsker $\Delta = 0.01$ får vi $n = 0.10(1 - 0.10)/0.01^2 = 900$, dvs. betydelig lavere enn den "pessimistiske" n ovenfor. Dersom problemet isteden dreier seg om estimering av en kjønnsproporsjon, f.eks. proporsjonen av hunkalver hos en bestemt kvegrase, og vi vet at denne er nær 0.50, er det klart at vi er nødt til å velge n=2500 for å oppnå at $\Delta = 0.01$.

Dette eksemplet skulle klart vise betydning av å bruke *a* priori informasjon til å planlegge størrelsen av eksperimentet. Å ta observasjoner koster både tid og penger. Vi ønsker ikke å gjøre mer enn nødvendig for å oppnå en ønsket presisjon. På den annen side ønsker vi ikke å komme i den situasjon at hele undersøkelsen viser seg å være verdiløs fordi vi i utgangspunktet planla for få observasjoner. La oss se på nok et eksempel.

Eksempel 10 : Hypergeometrisk situasjon
Anta at Y er hypergeometrisk fordelt (N, M, n). vi ønsker å estimere $a = M/N$ ved bruk av estimatoren $\hat{a} = Y/n$. Anta at vi står fritt til å velge n slik at estimatoren får et ønsket standardavvik Δ. Vi skal derfor ha (se Eksempel 5)

$$var\hat{a} = \frac{N-n}{N-1} \cdot \frac{1}{n} \cdot a(1-a) = \Delta^2$$

Løser vi dette med hensyn på n får vi

$$n = \frac{a(1-a)}{\frac{a(1-a)}{N} + \Delta^2 \cdot \frac{N-1}{N}} \approx \frac{a(1-a)}{\Delta^2}$$

Tilnærmingen fås ved å la $N \to \infty$ i den eksakte formel, og gjelder derfor kun for stor N. Beregningen med den tilnærmede formel er enklere, og gir en øvre skranke n', svarende til uendelig populasjon, for det nødvendige antall obsevasjoner n. Det er lett å sjekke at n kan (tilnærmet) uttrykkes som

$$n = n' \cdot \frac{1}{1 + \frac{n'}{N}}$$

Vi kan tolke brøken som en korreksjonsfaktor, som reduserer det nødvendige antall observasjoner, som følge av at populasjonen er endelig. For å få et inntrykk av populasjonsstørrelsens rolle, ser vi på et regneeksempel der $n' =$

2500, som bl.a. er tilfelle for $\Delta = 0.01$ og $a = 0.5$ (jfr. Eksempel 9). Vi får da

N	100	1 000	10 000	100 000	1 000 000	∞
n	96	714	2 000	2 439	2 494	2 500

For små populasjoner vil n utgjøre en betydelig andel av denne. Ettersom populasjonen øker, vil n ikke øke i samme takt, og når populasjonen er stor, vil n være en beskjeden andel av denne. Heri ligger noe av grunnlaget for moderne meningsmålinger. I Eksempel 7 der Lillevik hadde 10 000 stemmeberettigede, er det nødvendig med en stikkprøve på 2 000 for å oppnå et standardavvik på 0.01 for estimert andel poltilhengere. I et land med 10 mill. stemmeberettigede, vil en landsomfattende meningsmåling foran et presidentvalg med to kandidater høyst kreve en stikkprøve på 2 500 for å oppnå samme presisjon.

7.4 Konfidensgrenser

I mange situasjoner ønsker en å bruke observasjoner til å sette grenser for verdien av ukjent modellparameter θ. Dersom det er knyttet en pålitelighetsgaranti til beregning av grensene, vil vi kalle dem *konfidensgrenser*. Vi taler om den nedre konfidensgrense $\underline{\theta}$ og den øvre konfidensgrense $\overline{\theta}$ for θ, som sammen definerer et intervall $[\underline{\theta}, \overline{\theta}]$.

> **Definisjon :** Et *konfidensintervall* for en ukjent parameter θ er et intervall med grenser som, med en gitt sannsynlighet c kalt *konfidensnivået*, omslutter θ.

Det er opp til brukeren å velge konfidensnivå, og dette er bl.a. bestemmende for lengden av konfidensintervallet. Vi kan ønske oss et kort intervall (stor presisjon) med høyt konfidensnivå (stor sikkerhet), men det er klart at et lengre intervall kan garanteres med større sikkerhet enn et kort. Et konfidensintervall kan oppfattes som mengden av plausible parameterverdier i lys av den foreliggende informasjon. Økes informasjonen vil en kunne "skrelle vekk" ytterligere verdier, slik at en får et kortere intervall med den samme sikkerhet. Et valgt kompromiss mellom krav til presisjon og sikkerhet vil derfor kunne bestemme omfanget av observasjoner. I forrige avsnitt ga vi en metode for tilnærmet beregning av konfidensintervaller: I situasjoner der vi har en forventningsrett estimator $\hat{\theta}$ for θ med kjent standardavvik $\sigma(\hat{\theta})$, og med fordeling som kan tilnærmes med normalkurven, vet vi at sannsynligheten for at $\hat{\theta}$ avviker fra den ukjente θ med høyst $k \cdot \sigma(\hat{\theta})$, er tilnærmet lik arealet under normalkurven mellom $-k$ og k. Følgelig bestemmer

$$\hat{\theta} \;\; \pm \;\; k \cdot \sigma(\hat{\theta})$$

et konfidensintervall for θ med konfidensnivå c tilnærmet lik $A(k)$, og tabellen i forrige avsnitt gir derfor tilnærmet c for ulike valg av k, som kan oppfattes som en sikkerhetsfaktor. I de fleste situasjoner er imidlertid $\sigma(\hat{\theta})$ ukjent og må estimeres. Konfidensgarantien vil da bli noe tvilsom.

Statistisk teori kan imidlertid anvise beregningsmåter for konfidensgrenser med eksakt konfidensgaranti for mange aktuelle modeller. Beregningsmåten er imidlertid spesifikk for den enkelte modell og ofte nokså komplisert. Tabeller vil være et alternativ, men blir ofte altfor omfangsrike når en skal dekke spekteret av aktuelle konfidensnivåer og observasjonsstørrelser. For en del aktuelle situasjoner, så som den binomiske og hypergeometriske, vil beregning av eksakte konfidensgrenser kunne tilbys som del av statistisk programvare. I Poissonsituasjonen er imidlertid tabellene overkommelige, og i Tabell C.12 i Appendiks C gir vi konfidensgrenser for forventningen i Poissonfordelingen for ulike konfidensnivåer.

Eksempel 11 : Ulykkesrisiko
La situasjonen være som i Eksempel 8, hvor vi ønsker et konfidensintervall for forventet antall ulykker pr. år λ, basert på observert antall et bestemt år X, som antas Poissonfordelt. Dersom vi ønsker konfidensnivå lik 95%, og observerer $X = 9$ ulykker, gir Tabell C.12 eksakt nedre og øvre konfidensgrense for λ lik henholdsvis 4.12 og 17.08. Merk at når disse grensene ikke ligger symmetrisk om estimatet 9, skyldes det at Poissonfordelingen er skjev, mens normaltilnærming gir de symmetriske grensene 9 ± 6. Hvis vi kjenner det totale antall nordmenn n under risiko, vil en kunne dividere de funne tall med n, og få estimat og konfidensgrenser for sannsynligheten for at en tilfeldig person under risiko rammes i løpet av et år. Et forsikringsselskap med m forsikrede vil kunne multiplisere disse tall igjen med m, og få estimat og konfidensgrenser for forventet antall utbetalinger i løpet av et år.

Tabell C.12 har en rekke anvendelsesmuligheter: I binomiske situasjoner der suksess-sannsynligheten p er liten og n er stor, vet vi at antall suksesser X er tilnærmet Poissonfordelt med forventning $\lambda = n \cdot p$. Vi kan da få tilnærmede konfidensgrenser for p ut fra Tabell C.12, ved å dividere grensene for λ med n. Tilsvarende vil en i hypergeometriske situasjoner der andelen spesielle $a = M/N$ er liten, og n er stor, men liten i forhold til N, ha at antall spesielle i utvalget Y er tilnærmet Poissonfordelt med forventning $\lambda = n \cdot a = n \cdot M/N$. Vi kan da få tilnærmede konfidensgrenser for a ved å dividere grensene for λ med n, samt grenser for M ved å multiplisere disse igjen med N. I slike skjeve situasjoner gir Poissontilnærmingen langt mer pålitelige grenser enn de som er basert på normaltilnærming, som imidlertid gir best resultat i symmetriske situasjoner ($p \approx 0.5$). Dersom observerte antall er så stort at det ikke dekkes av Tabell C.12, vil normaltilnærmingen likevel kunne gi tilfredstillende resultat.

I enkelte situasjoner er en bare interessert i den ene konfidensgrensen. Det vil være tilfellet når en kun er opptatt av hvorvidt den ukjente θ ikke er for høy, f.eks. under en faregrense. Det er da nok å beregne en øvre konfidensgrense, slik at vi har et ensidig konfidensintervall $(-\infty, \overline{\theta}]$. Risikoen for at intervallet ikke fanger opp den ukjente θ er da bare knyttet til den øvre grensen, og vil typisk være halvert i forhold til den tosidige risikoen $r = 1 - c$. Oppslag i tabeller for tosidig konfidensnivå c svarer derfor til et ensidig konfidensnivå lik $c' = 1 - r/2 = (1 + c)/2$. I Eksempel 11 vil 15.71 være en øvre konfidensgrense for λ med ensidig konfidensnivå på 95% (ved bruk av Tabell C.12 og 90% tosidig nivå).

Eksempel 12 : Revisjon
I forbindelse med vurdering av den interne kontroll hos en klient gransker revisor utbetalinger mht. avvik fra avtalte kontroller (signaturer etc.). En andel utbetalinger med kontrollavvik på 1% anses normalt, mens mer enn 5% anses urovekkende. Revisor gransker et tilfeldig utvalg på n=200 utbetalinger blant et mye større antall registrerte utbetalinger, og observerer Y utbetalinger med kontrollavvik. Siden n er stor og andelen utbetalinger med kontrollavvik a er lite, er det rimelig å anta at Y er tilnærmet Poissonfordelt med forventning $\lambda = n \cdot a$. La oss si at vi observerer $Y = 4$, dvs. 2% med avvik i utvalget. Av Tabell C.12 finner vi en øvre konfidensgrense på 9.15 for λ med konfidensnivå 95%. Dette gir en øvre grense lik 9.15/200=0.04525 for a, og revisor ser ingen grunn til nærmere gransking. Merk imidlertid at dersom a virkelig er 0.01, vil det være en betydelig risiko for at konfidensgrensen faller over “faregrensen” på 0.05, som da krever nærmere gransking. I følge Tabell C.12 skjer det når $Y = 5$ eller mer, som har sannsynlighet 0.0526 (se Tabell C.4 med $\lambda = 200 \cdot 0.01 = 2$). For å redusere denne risikoen må utvalget økes noe.

7.5 Hypotesetesting

Mange statistiske undersøkelser tar sikte på å belyse fremsatte hypoteser, og eventuelt klargjøre om hypotesene bør forkastes. Vi vil fortsatt anta at vi har en delvis spesifisert modell for det fenomen som studeres. Innenfor denne rammen vil en hypotese være et utsagn om de ukjente elementene i modellen, dvs. som oftest uspesifiserte modellparametre.

> **Definisjon :** En *hypotesetest* er en observasjonsbasert metode for å klargjøre om en gitt hypotese vedrørende en statistisk modell, kalt *nullhypotesen* H_0, bør forkastes til fordel for en gitt *alternativ hypotese* H_A.

H_0 og H_A vil alltid utelukke hverandre gjensidig samt utfylle alle muligheter som foreligger. Dette betyr at forkastning av nullhypotesen H_0 medfører aksept av den alternative hypotesen H_A. Oftest vil nullhypotesen være at vi stiller oss reservert til en framsatt påstand, som da vil være den alternative hypotesen. Vi stiller oss imidlertid åpne for at observasjoner kan peke såpass i favør av den alternative hypotesen at vi velger å gå inn for denne. Teorien for hypotesetesting tar sikte på å etablere metoder som har innebygget visse garantier mot feilslutninger.

> **Definisjon :** En *testmetode* er definert ved en stokastisk variabel T avledet av observasjonene, kalt *testobservatoren*, og et *forkastningsområde*, dvs. de verdier av T som betyr at nullhypotesen H_0 skal forkastes.

Vi vil vanligvis kunne velge blant flere testmetoder, og spørsmålet er da hvilke retningslinjer vi skal legge til grunn for vårt valg. Vi vil her fokusere på risikoen for feilslutninger.

To typer feilslutninger kan forekomme. På den ene side kan vi komme til å forkaste H_0 når H_0 i virkeligheten er riktig, dette kalles *forkastningsfeil*, mens vi på den annen side kan komme til å la være å forkaste en hypotese som burde ha vært forkastet, dette kalles *godtakingsfeil*. De mulige situasjonene som kan oppstå er

	"Virkeligheten"	
Slutning	H_0 riktig	H_A riktig
Ikke forkast H_0	Korrekt slutning	Godtakingsfeil
Forkast H_0	Forkastningsfeil	Korrekt slutning

I denne forbindelse innfører vi :

$$\begin{aligned} \alpha &= \text{Sannsynligheten for forkastningsfeil } (\alpha \text{ -risiko}) \\ \beta &= \text{Sannsynligheten for godtakingsfeil } (\beta \text{ -risiko}) \end{aligned}$$

Vi ønsker oss selvsagt en testmetode som innebærer at risikoen for begge typer feilslutninger er liten, men, som vi skal se, vil disse to ønsker trekke i hver sin retning. I praksis må vi derfor foreta en rimelig avveining mellom de to ønskene, eller alternativt finne ut hvor omfattende undersøkelsen må være for at begge risikoer er redusert til et akseptabelt nivå. Slike avveininger kan gjøres ved hjelp av følgende begrep:

Definisjon : *Styrkefunksjonen* til en test angående en parameter θ er sannsynligheten for å forkaste H_0 betraktet som funksjon av θ, betegnet her med $\Pi(\theta)$. Dersom

$$\max_{\theta \in H_0} \Pi(\theta) = \alpha$$

sier vi at testen har *signifikansnivå* α, mens $\Pi(\theta)$ beregnet for en θ i H_A kalles *styrken* i dette alternativet.

Definisjonen av signifikansnivå innebærer at $\Pi(\theta) \leq \alpha$ for alle θ i nullhypotesen, med likhet for minst en θ, typisk for grensetilfellet mellom nullhypotese og alternativ.

La oss illustrere de grunnleggende idéer omkring hypotesetesting ved hjelp av et konstruert eksempel.

Eksempel 13 : Testing i binomisk modell
En mann påstår at han har evner til å forutsi utfallet av myntkast. Han kan riktignok ikke spå riktig utfall hver gang, og en rimelig tolking av påstanden vil være at hans sjanse for å spå riktig utfall av et myntkast er større enn $\frac{1}{2}$, som er sjansen for å spå riktig for en person uten slike evner, og som er henvist til å velge sin "spådom" tilfeldig blant de to mulighetene krone eller mynt. For å se om det er noen grunn til å gi mannen rett i sin påstand, lar vi mannen få anledning til å spå utfallet av et antall myntkast som utføres med en rettferdig mynt. Det er rimelig å anta at sannsynligheten p for riktig spådom er den samme i hvert kast. Anta at vi stiller oss tvilende til mannens påstand. Vi formulerer da den hypotese at mannen ikke er bedre enn folk flest til å spå utfallet av myntkast, dvs. vår hypotese er at $p = \frac{1}{2}$. Som vårt alternativ til hypotesen har vi at mannen har spådomsevner, dvs. at $p > \frac{1}{2}$. Anta at vi gir mannen anledning til n=10 myntkast. La X være antall riktige spådommer, og anta at de 10 spådommene representerer uavhengige forsøk med hensyn til utfallene suksess og fiasko. Vi kan da slutte at X er binomisk fordelt $(n = 10, p)$ der p er ukjent. På grunnlag av observert verdi av X skal vi avgjøre om det er noen grunn til å forkaste hypotesen, og følgelig måtte gi mannen rett. Vi skal altså teste

$$H_0 : p = \frac{1}{2} \quad \text{mot} \quad H_A : p > \frac{1}{2}$$

Mange riktige spådommer taler i favør av at mannen har spådomsevner, og det er derfor rimelig å bruke X som testobservator og forkaste H_0 når X er tilstrekkelig stor. Vår testmetode blir derfor:

$$\text{Forkast } H_0 \text{ og påstå } H_A \text{ dersom } X \geq k$$

der k er en passende valgt konstant. Problemet er å velge k slik at vi får en brukbar test. Forkastningsområdet vil i dette eksemplet være alle mulige verdier av X større enn eller lik k, som vi vil gi navnet testens *kritiske verdi*.

La oss først belyse risikoen for forkastningsfeil, dvs. påstå at mannen har spådomsevner (H_A) når han i virkeligheten ikke har det (H_0). Sannsynligheten for dette (signifikansnivået) blir

$$\alpha = P_{H_0}(X \geq k)$$

der P_{H_0} betegner sannsynlighet beregnet under forutsetning av at H_0 er riktig. Av Tabell over den binomiske fordeling $(n = 10, p = 0.5)$ får vi

x	$\cdots$	6	7	8	9	10
$P_{H_0}(X = x)$	$\cdots$	0.2051	0.1172	0.0439	0.0098	0.0010

Av denne tabellen finner vi signifikansnivået α for testen for ulike valg av den kritiske verdi k

k	$\cdots$	6	7	8	9	10
$\alpha = P_{H_0}(X \geq k)$	$\cdots$	0.3770	0.1719	0.0547	0.0108	0.0010

Dersom vi velger en metode som forkaster H_0 når $X \geq 8$ betyr det at sannsynligheten for å påstå spådomsevner når slike ikke er til stede blir ca. 5%. Velger vi isteden å være litt mer skeptiske og forkaste H_0 først når $X \geq 9$ blir denne sannsynligheten redusert til ca. 1%. La oss så se på risikoen for godtakingsfeil, dvs. la være å påstå spådomsevner hos en mann som i virkeligheten har det. Sannsynligheten for å la være å påstå H_A når H_A er riktig skriver vi

$$\beta = P_{H_A}(X < k).$$

Denne sannsynlighet vil avhenge av hvilken $p > 0.5$ i alternativet H_A vi ser på. La oss interessere oss for alternativet $p = 0.8$, nedenfor kalt H_1. For den binomiske fordeling $(n = 10, p = 0.8)$ finner vi av Tabell C.2, for ulike valg av kritisk verdi k

k	$\cdots$	6	7	8	9	10
$\beta = P_{H_1}(X < k)$	$\cdots$	0.0328	0.1209	0.3222	0.6242	0.8926

Anta at vi insisterer på at sannsynligheten α for å påstå spådomsevner når slike ikke er til stede skal være høyst ca. 0.05. Vi må da velge $k = 8$ som kritisk verdi. Dette gir $\alpha = 0.0547$. Vi ser imidlertid at dette valg medfører at såpass gode spådomsevner som svarende til $p = 0.8$ vil ha sannsynlighet på hele $\beta = 0.3222$ for å ikke bli avslørt. Vår spåmann vil kunne hevde at denne testen ikke vil gi ham en rimelig sjanse til å vise sine ferdigheter. Vi ser imidlertid at dersom vi reduserer β, f.eks. ved å isteden velge $k = 7$ som kritisk verdi, vil α måtte øke. Omvendt vil reduksjon av α medføre økning av β. Med det gitte antall observasjoner, er det ikke mulig å finne en testmetode som samtidig gjør α og β liten. Dersom vi imidlertid foretar flere observasjoner, og dermed får mer informasjon, vil trolig risikoen for feilkonklusjoner kunne reduseres. La oss forsøke med $n = 30$ observasjoner. Nå blir X binomisk fordelt $(n = 30, p)$. Av en større tabell over den binomiske fordeling finner vi (alternativt: bruk normaltilnærming)

$$P_{H_0}(X \geq 20) = 0.0495 \qquad P_{H_1}(X < 20) = 0.0257$$

Med $n = 30$ observasjoner og $k = 20$ som kritisk verdi, får vi en test som oppfyller kravet om at signifikansnivået α skal være høyst 0.05 samtidig som β er tilsvarende liten. Vi har ovenfor vurdert testmetoden ut fra alternativet H_1 $(p = 0.8)$. De samme betraktninger kan gjøres gjeldende for et vilkårlig annet valgt alternativ $p > 0.5$. Anta nå at vi har valgt å gjøre $n = 30$ observasjoner og fastlagt testmetoden til

$$\text{Forkast } H_0 \text{ og påstå } H_A \text{ når } X \geq 20.$$

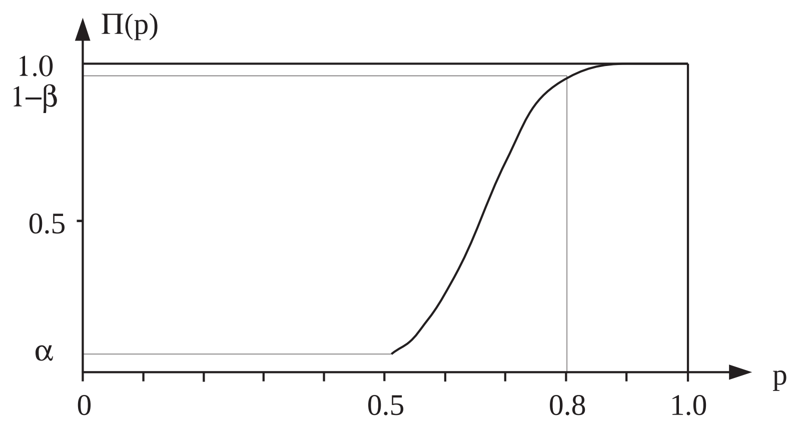
Testmetodens egenskaper kan oppsummeres i styrkefunksjonen til testen, som her er sannsynligheten for å forkaste nullhypotesen betraktet som funksjon av p. Vi skriver

$$\Pi(p) = P(X \geq 20).$$

Av en binomisk tabell finner vi for ulike p

p	0.5	0.6	0.7	0.8	0.9	1.0
$\Pi(p)$	0.0495	0.2915	0.7304	0.9743	0.9999	1.0000

$\Pi(p)$ angir testens evne til å avsløre at $p > 0.5$ for de ulike p (> 0.5). Vi sier at $\Pi(p)$ er testens styrke i alternativet p (> 0.5). På Figur 7.3 finner vi igjen $\Pi(0.8) = 1 - \beta$ og $\Pi(0.5) = \alpha$ (signifikansnivået).



Figur 7.3: Styrkefunksjonen for testen

Vi har her sett på situasjonen der nullhypotesen $p = p_0$ ble testet mot alternativet $p > p_0$. I andre situasjoner vil $p < p_0$ være det aktuelle alternativ, se Eksempel 1. Dette gjøres på tilsvarende måte, og beskrives ikke i detalj her, se Oppgave 31. Forøvrig får problemet samme form som beskrevet ovenfor ved å formulere hypotesen om $q = 1 - p$ isteden. I atter andre situasjoner er det aktuelle alternativ $p \neq p_0$, som for eksempel når en vil teste om en mynt er rettferdig ($p = 1/2$ mot $p \neq 1/2$). I dette tilfellet forkastes H_0 både når X er stor og liten, se Oppgave 32 for detaljer.

Merknad. Mange foretrekker å formulere binomisk testing med den standardiserte variable

$$Z = \frac{X - np_0}{\sqrt{np_0(1-p_0)}}$$

som testobservator. Denne er tilnærmet standardnormalfordelt når nullhypotesen er riktig, og n er tilstrekkelig stor.

I praksis er en interessert i å planlegge antall observasjoner n slik at en har oppfylt en gitt α-risiko og β-risiko. Det er brysomt å prøve og feile som i eksemplet, og en ferdig formel basert på normaltilnærming er:

$$n = (\frac{k_\alpha \cdot \sqrt{p_0(1-p_0)} + k_\beta \cdot \sqrt{p_1(1-p_1)}}{p_1 - p_0})^2$$

der k_α er punktet som har areal under normalkurven lik α til høyre for seg. Den kritiske verdi kan så beregnes ved

$$k = np_0 \pm \frac{1}{2} \pm k_\alpha \cdot \sqrt{np_0(1-p_0)}$$

der en bruker minus eller plusstegnet alt ettersom alternativet til nullhypotesen går i negativ eller positiv retning. I det tosidige tilfellet må en bruke $\alpha/2$ istedenfor α i formelen og bruke to kritiske verdier (begge fortegnene).

Eksempel 14 : Prosesskontroll

Anta at vi studerer en løpende produksjonsprosess der den enkelte artikkel kan enten klassifiseres som defekt eller intakt. Antar vi at sannsynligheten p for at en enkelt artikkel er defekt er den samme for alle artikler samt at kvaliteten av artiklene er uavhengig av hverandre, så er vi i en binomisk situasjon. Vi antar at når prosessen er "i kontroll" så er av erfaring defektsannsynligheten på et bestemt nivå $p_0 = 0.10$, og dersom prosessen løper ut av kontroll så ytrer det seg kun ved at defektsannsynligheten er økt, f.eks.

innstiller seg på et nytt nivå $p_1 > p_0$. Dette kan skyldes maskinslitasje e.l. La oss tenke oss at situasjonen er slik at ved arbeidsdagens slutt undersøkes de n sist produserte artiklene, antall defekte X observeres og det avgjøres om det er grunn til å tro at prosessen er ute av kontroll slik at maskinene bør justeres før produksjonen tar til igjen neste morgen. Under de gitte forutsetninger er X binomisk fordelt (n, p) der p er ukjent. Situasjonen kan formuleres som et hypotesetestingsproblem der nullhypotese og alternativ er gitt ved

$$\begin{aligned} H_0 :&\quad p = p_0 \quad \text{(prosessen er i kontroll)} \\ H_A :&\quad p > p_0 \quad \text{(prosessen er ute av kontroll).} \end{aligned}$$

Som testobservator brukes X og vi

$$\begin{gathered} \text{forkaster } H_0 \text{ og påstår } H_A \text{ dersom } X \geq k \\ \text{forkaster ikke } H_0 \text{ dersom } X < k. \end{gathered}$$

der k er en passende valgt konstant. Forkastning og ikke-forkastning vil her i praksis bety henholdsvis justering og ikke-justering av produksjonsprosessen. Anta at vi ønsker styrke 0.95 i alternativet $p_1 = 0.20$. Problemstillingen er som ovenfor med $\alpha = 0.05$ og $\beta = 0.05$, slik at $k_\alpha = k_\beta = 1.645$. Formelen ovenfor gir da $n = 133$, som innsatt gir $k = 19.5$. Siden k må være et heltall velger vi $k = 20$. Da blir α noe mindre enn 0.05 på bekostning av β som blir noe større enn 0.05. Ved valg av $k = 19$ blir forholdet omvendt.

I praksis blir statistisk prosesskontroll utført ved hjelp av et såkalt kontrolldiagram, der data plottes regelmessig, slik at verdifull informasjon som ligger i rekkefølgen av observasjonene også kan blottlegges, mer om dette i Kapittel 15.

Vi har i dette avsnittet diskutert testing av hypoteser om sannsynligheter i binomiske situasjoner. Framgangsmåten i forbindelse med testing av hypoteser om andeler i hypergeometriske situasjoner er helt analog, og tas ikke opp spesielt her.

La oss oppsummere og utdype noen av de innførte begreper: Etter at en modell for vårt eksperiment er etablert og nullhypotese og alternativ er fastlagt, står en overfor valg av testobservator. I enkelte situasjoner er det opplagt hva som bør velges (som i eksemplene ovenfor), i andre situasjoner kan det foreligge flere muligheter. Anta at en testobservator T er valgt, en må da fastlegge forkastningsområdet, dvs. hvilke verdier av T (store, små eller både store og små) som bør bety forkastning av nullhypotesen. Dette vil avhenge av problemstillingen, dvs. hvilken alternativ hypotese som foreligger. Ønsket om at testen skal ha et bestemt signifikansnivå og styrke, er så bestemmende for antall observasjoner og testens kritiske verdi.

Signifikansnivået α til en test angir sannsynligheten for forkastningsfeil, dvs. å forkaste nullhypotesen og dermed påstå den alternative hypotese når nullhypotesen er riktig, mens $\beta = 1-\Pi$, der Π er styrken (styrkefunksjonen) til testen, angir sannsynligheten for godtakingsfeil, dvs. å la være å påstå den alternative hypotese når denne er riktig.

Det må understrekes at de risikoer vi er villige til å tolerere må komme til uttrykk på forhånd, før vi ser på observasjonene. Når vi så fastlegger testmetoden i henhold til våre ønsker, får vi en garanti på testmetodens pålitelighet. Eksempelvis kan signifikansnivået α tolkes slik at ved gjentatt bruk av samme metode i lignende situasjoner vil vi, gjennomsnittlig i det lange løp, oppleve forkastningsfeil i $100 \cdot \alpha\%$ av tilfellene. Dersom testmetoden ikke er fastlagt på forhånd, er det lett å la seg påvirke av observasjonene ved valg av metode. I verste fall kan observatøren være forutinntatt i situasjonen, og ved å titte på observasjonene kan han, ved å "velge" signifikansnivået tilstrekkelig stort (eller lite), oppnå den konklusjon han ønsket seg. Dette er juks!

Når metoden er fastlagt kan vi ta fram observasjonene (eventuelt innhente disse), finne den realiserte verdi av testobservatoren T, kall den t. Vi kan nå trekke vår konklusjon: forkastning av nullhypotesen dersom t er en verdi i det, på forhånd fastlagte, forkastningsområdet. I motsatt fall konkluderer vi at de foreliggende observasjoner ikke gir grunnlag for å forkaste nullhypotesen.

Det at en nullhypotese er forkastet betyr ikke at vi har bevist, i streng forstand, at nullhypotesen er feil. Heller ikke betyr ikke-forkastning at vi har bevist at nullhypotesen er riktig. Utsagn som "det er statistisk bevist at ..." bør alltid tas med en klype salt. Som regel foreligger det en viss risiko for feilkonklusjoner og disse bør gjøres kjent.

Det kan knyttes en forbindelse mellom bruken av konfidensintervallet og hypotesetesting: Gitt en nullhypotese angående en parameter θ. Beregn et konfidensintervall for θ med konfidensnivå c, ensidig eller tosidig alt ettersom alternativet til nullhypotesen er det. Forkastes nullhypotesen dersom denne ligger utenfor konfidensintervallet, har vi en test med signifikansnivå $\alpha = 1-c$.

Vi har ovenfor lagt vekt på å bruke teorirammen for hypotesetesting til å diskutere de to typer feil i sammenheng. Det virker både naturlig og nødvendig i beslutningssituasjoner som kan gis en godta-forkast formulering. Slike problemer kan også studeres i en videre beslutningsteoretisk ramme, der også subjektiv kunnskap om ukjente parametre kan inngå. Dette aspekt forfølges i Kapittel 16.

I resten av dette avsnittet skal vi diskutere ulike sider ved hypotesetesting som objektivt hjelpemiddel til å vurdere og formidle utsagnskraften i et observasjonsmateriale, slik det ofte gjøres i utredninger og vitenskapelige rapporter. På den ene side har vi tester som tar sikte på å vurdere graden av tilpasning observasjonene har til en bestemt modell, kalt *føyningstester*. Eksempler på dette fins i Kapittel 7.7 og i Kapittel 9. På den annen side har vi tester som angår et enkelt aspekt ved modellen. I mange situasjoner dreier det seg om å teste nullhypotesen at det ikke er noen sammenheng av et bestemt slag, ofte formulert som et utsagn om en parameter, mot alternativet at sammenheng foreligger. Testen skal gi grunnlag for å vurdere hvorvidt en mulig observert sammenheng er *signifikant*, dvs. ikke kan skyldes tilfeldigheter. Eksempler på dette fins i Kapittel 8, 10, 11 og 12.

I praksis ser en, nesten konvensjonelt, brukt signifikanssnivåene 1%, 2.5% og 5% (spesielt i vitenskapelige anvendelser), og utfallet av en test rapporteres ofte slik "det observerte resultat er signifikant (evt. ikke signifikant) på 5% nivå". Dette betyr at med signifikanssnivå 5% vil observasjonene medføre at nullhypotesen blir (evt. ikke blir) forkastet. Noen har etablert den praksis å rapportere et resultat som er signifikant på 5% nivå, men ikke på 2.5% nivå med *, et resultat som er signifikant på 2.5% nivå, men ikke på 1% nivå med ** og et resultat som er signifikant på 1% nivå med ***. Mer vanlig er det å rapportere resultatet av en test ved den såkalte P-verdien.

> **Definisjon :** Sannsynligheten for at testobservatoren skulle gi en verdi minst like mye i favør av H_A som det vi observerte, beregnet under forutsetning av at H_0 er riktig, kalles *P-verdien* til det observerte resultat.

P-verdien gir en indikasjon på hvor utsagnskraftig det observerte resultat er i favør av den alternative hypotesen, jo mindre P-verdi desto mindre trolig er det at nullhypotesen er riktig. Vi har følgende sammenheng mellom P-verdi og signifikansnivå:

$$\boxed{P \leq \alpha \Longleftrightarrow \quad H_0 \text{ forkastes ved valgt signifikansnivå } \alpha}$$

Dette betyr at P er det minste signifikansnivå som tillater at hypotesen forkastes. Ved lesing av en rapport med P-verdier, bør en på forhånd gjøre seg opp en mening om hvilken risiko for forkastningsfeil en er villig til å løpe, dvs. bestemme seg for α. Dersom den rapporterte P-verdi er mindre enn eller lik den valgte α, vil det si at det observerte resultat medfører forkastning av nullhypotesen for vedkommende person.

La oss begrunne påstanden ovenfor i situasjonen der store verdier av testobservatoren T peker mot H_A, slik at vi forkaster H_0 når $T \geq k$, der k er kritisk verdi. Med signifikanssnivå α er k bestemt ved at $\alpha = P_{H_0}(T \geq k)$. Hvis observert verdi av T er t, blir P-verdien til det observerte resultat $P = P_{H_0}(T \geq t)$, og vi ser at $t \geq k \Leftrightarrow P \leq \alpha$. Tilsvarende begrunner vi resultatet dersom det er små (eller både store og små) verdier av T som peker mot H_A.

Til slutt noen kritiske kommentarer: Signifikanstesting blir ofte utført som en bevisstløs pliktøvelse, med betydelig fare for å villede seg selv og andre, bl.a. på grunn av ensidig fokusering på forkastningsfeil, gjennom å tilpasse testen et gitt signifikanssnivå eller beregning av P-verdi. De to vanligste feilgrep er:
1. En betydningsfull sammenheng forblir uoppdaget fordi rapportert ikkesignifikans skyldes for få observasjoner, og dermed for liten teststyrke.
2. En rapportert signifikant sammenheng eller effekt oppfattes feilaktig som betydningsfull. Selv den mest betydningsløse forskjell fra nullhypotesen kan avsløres, dersom vi har mange nok observasjoner.
Begge feilgrep kan unngås ved å utføre styrkebetraktninger, idet en da må ta stilling til hva som anses for å være betydningsfulle forskjeller.

7.6 Analyse av forskjeller

Man ønsker ofte å sammenligne samme modellparameter θ i to ulike situasjoner eller for to populasjoner

- defektsannsynligheten p ved to produksjonsmetoder
- andelen a med et bestemt standpunkt i to populasjoner

La generelt θ_1 og θ_2 være parameteren for hver av de to situasjonene. Vi kan da være interessert i å anslå forskjellen $\theta_1 - \theta_2$ eller teste hypotesen at parametrene er like $\theta_1 = \theta_2$. Anta at det foreligger uavhengige observasjoner fra hver av de to situasjonene/populasjonene. Det er da klart at

Med $\hat{\theta}_1$ og $\hat{\theta}_2$ uavhengige forventningsrette estimatorer for hhv. θ_1 og θ_2 blir

$$E(\hat{\theta}_1 - \hat{\theta}_2) = \theta_1 - \theta_2 \ , \quad \sigma(\hat{\theta}_1 - \hat{\theta}_2) = \sqrt{\sigma^2(\hat{\theta}_1) + \sigma^2(\hat{\theta}_2)}$$

Det betyr at $\hat{\theta}_1 - \hat{\theta}_2$ blir en forventningsrett estimator for $\theta_1 - \theta_2$ med standardavvik som kan beregnes ut fra standardavviket til hver av estimatorene $\hat{\theta}_1$ og $\hat{\theta}_2$. Dette standardavviket må i praksis oftest estimeres med

$$S(\hat{\theta}_1 - \hat{\theta}_2) = \sqrt{S^2(\hat{\theta}_1) + S^2(\hat{\theta}_2)}$$

der $S(\hat{\theta}_1)$ og $S(\hat{\theta}_2)$ er estimatorer for hhv. $\sigma(\hat{\theta}_1)$ og $\sigma(\hat{\theta}_2)$. I situasjoner av praktisk interesse, som binomiske og hypergeometriske situasjoner, vil fordelingen til $\hat{\theta}_1 - \hat{\theta}_2$ etter standardisering kunne tilnærmes med normalkurven. Feilmarginer konstruert etter standardfeilmetoden kan derfor vurderes på vanlig tilnærmet måte, såfremt antall observasjoner ikke er for lite.

Eksempel 15 : To produksjonsmetoder
Anta at det produseres etter to metoder og man finner X_1 defekte blant n_1 produserte etter metode 1, X_2 defekte blant n_2 produserte etter metode 2. Antar vi at defektsannsynligheten ved de to metodene er hhv. p_1 og p_2, er det rimelig å anta at X_1 er binomisk fordelt (n_1, p_1) og X_2 binomisk fordelt (n_2, p_2), samt at X_1 og X_2 er uavhengige. Siden $\hat{p}_1 = X_1/n_1$ og $\hat{p}_2 = X_2/n_2$ er forventningsrette estimatorer for hhv. p_1 og p_2 blir

$$\hat{p}_1 - \hat{p}_2 \ \text{ forventningsrett for } \ p_1 - p_2$$

og standardavviket til denne estimatoren blir

$$\sigma(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

som i praksis estimeres med $S(\hat{p}_1 - \hat{p}_2)$ definert ved at p_1 og p_2 erstattes med $\hat{p}_1$ og $\hat{p}_2$ i formelen ovenfor.

Anta at det ble produsert $n_1 = n_2 = 100$ artikler med hver metode, og at antall defekte ble henholdsvis $X_1 = 9$ og $X_2 = 5$. Vi får da

$$\hat{p}_1 - \hat{p}_2 = 0.09 - 0.05 = 0.04$$

mens det tilhørende standardavvik anslås til

$$S(\hat{p}_1 - \hat{p}_2) = \sqrt{0.09 \cdot 0.91/100 + 0.05 \cdot 0.95/100} = 0.036$$

Det er derfor ingen grunn til å påstå at de to metodene generelt har ulik defektsannsynlighet.

Hypergeometriske situasjoner der en ønsker å estimere forskjellen på andelene spesielle elementer i to populasjoner, f.eks. andelen defekte i to produksjonspartier, ut fra tilfeldige utvalg, kan behandles på samme måte som i eksemplet. Eneste forskjell blir korreksjonsfaktorer for endelige populasjoner av typen $(N-n)/(N-1)$, som kan utelates dersom utvalgene utgjør en liten andel av populasjonene.

I spørreskjemaundersøkelser trekkes ofte et utvalg fra en større populasjon, mens en ved analysen av resultatene ofte bryter ned på undergrupper, f. eks. kjønn, som ønskes sammenlignet. Antallet i hver gruppe er da ikke gitt på forhånd. Likevel er det naturlig å rapportere standardavvik etter formlene ovenfor. Dette forutsetter at vi har et tilfeldig utvalg fra populasjonen, som etter gruppering kan oppfattes som et tilfeldig utvalg fra hver sin gruppe.

Eksempel 16 : Spørreskjemaundersøkelse
I en spørreskjemaundersøkelse ble n=400 tilfeldige personer fra en populasjon spurt om deres holdning (ja/nei) til et bestemt spørsmål. Svarene ble gruppert på kjønn og resultatet ble

	Ja	Nei	Sum
Mann	95 (43.2%)	125 (56.8%)	220 (100%)
Kvinne	70 (38.9%)	110 (61.1%)	180 (100%)
Sum	165	235	400

Estimert Ja-forskjell mellom Mann - Kvinne

$$0.432 - 0.389 = 0.043$$

Estimert standardavvik

$$\sqrt{0.432 \cdot 0.568/220 + 0.389 \cdot 0.611/180} = 0.049$$

Her er antatt at populasjonen er så stor at korreksjonsfaktor for endelig populasjon er unødvendig.

Vi er ofte interessert i å teste om det er forskjell mellom to sannsynligheter

$$H_0 : p_1 = p_2 \quad \text{mot} \quad H_A : p_1 \neq p_2 \;\; (\text{ evt.} > \text{eller} <)$$

En mulig testobservator er

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

der $\hat{p} = (X_1 + X_2)/(n_1 + n_2)$, dvs. at nevneren er en estimator for $\sigma(\hat{p}_1 - \hat{p}_2)$ som forutsetter at $p_1 = p_2$. Det kan vises at når H_0 er riktig, så kan fordelingen til U tilnærmes med normalkurven, slik at kritiske verdier eller P-verdier for testen kan beregnes tilnærmet.

Med dataene i Eksempel 15 blir $U = 1.63$, som gir $P = 0.102$ for tosidig alternativ og $P = 0.051$ for ensidig alternativ.

U-testen er forøvrig likeverdig med den såkalte kjikvadrattesten for uavhengighet i 2×2-tabell presentert i Kapittel 9. En alternativ test er Fisher-Irwins test beskrevet i Kapittel 10.

7.7 En test for realisme av en modell.

I Kapittel 1 introduserte vi begrepet sannsynlighetsmodell, og ønsket at slike skulle forklare empiriske fenomener. Vi har i dette kapitlet sett hvordan vi kan teste utsagn om enkeltsannsynligheter basert på data. Vi skal nå se hvordan vi kan teste om en gitt modell er realistisk.

Eksempel 17 : Rettferdig terning?
En modell for en rettferdig terning er

Utfall (u_i):	1	2	3	4	5	6
Sannsynligheter (p_i):	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

Vi har en vanlig terning og ønsker å teste om den er rettferdig. For dette formål ble terningen kastet $n = 240$ ganger og vi observerte antall ganger terningen viste henholdsvis $1, 2, ..., 6$ øyne. Vi fikk

Utfall (u_i):	1	2	3	4	5	6
Hyppighet (X_i):	32	44	36	45	34	49

Forventet antall kast med i øyne, dersom terningen er rettferdig, blir $EX_i = np_i = 240 \cdot 1/6 = 40$. Er de observerte antall forenlig med dette, eller så avvikende at modellen bør forkastes? Vi trenger en testobservator som sammenfatter graden av avvik.

La oss diskutere problemstillingen i generelle vendinger: Gitt et eksperiment med et endelig antall utfall hvor det foreligger en sannsynlighetsmodell med utfall og sannsynligheter gitt ved

$$\begin{array}{lccccc} & u_1 & u_2 & u_3 & \dots & u_m \\ H: & p_1 & p_2 & p_3 & \dots & p_m \quad (\sum_{i=1}^m p_i = 1). \end{array}$$

Det råder tvil om modellen er realistisk og den er foreløpig satt opp som en arbeidshypotese H. For å få et inntrykk av om modellen er realistisk, utføres eksperimentet n ganger uavhengig av hverandre og under de samme betingelser. La $X_1, X_2, \dots, X_m$ være antall ganger vi observerer henholdsvis $u_1, u_2, \dots, u_m$ $(\sum_{i=1}^m X_i = n)$.

Under de gitte forutsetninger og modell H, ser vi at X_i er binomisk fordelt (n, p_i) for $i = 1, 2, \dots, m$. Dette innses ved å betrakte u_i som suksess; vi har da n binomiske forsøk med suksessannsynlighet p_i. Følgelig er $EX_i = np_i$ og $varX_i = np_i(1 - p_i)$. Et egnet mål på hvor godt observasjonene i det store og hele samsvarer med modellen er

$$Q = \sum_{i=1}^{m} \frac{(X_i - np_i)^2}{np_i}$$

Dersom modellen er realistisk vil X_i være forholdsvis nær sin forventning np_i, i hvert fall gjennomgående nærmere enn når modellen er urealistisk. En stor verdi på Q vil derfor indikere at modellen ikke er realistisk, spørsmålet er bare hvor stor Q må være for at en slik tvil om at modellen er realistisk synes berettiget [1]. La oss beregne forventningen til Q ut fra forutsetningene ovenfor

$$\begin{aligned} EQ &= \sum_{i=1}^{m} \frac{1}{np_i} E(X_i - np_i)^2 = \sum_{i=1}^{m} \frac{1}{np_i} varX_i \\ &= \sum_{i=1}^{m} \frac{1}{np_i} np_i(1 - p_i) = \sum_{i=1}^{m} (1 - p_i) = m - 1 \end{aligned}$$

[1] I uttrykket for Q har vi i hvert ledd dividert med forventningen np_i, bl.a. fordi et gitt avvik $X_i - np_i$ bør tillegges mer vekt når np_i er liten enn når np_i er stor.

Det kommer derfor neppe på tale om å så tvil om modellen er realistisk før Q er noe større enn $m-1$.

Anta at observasjonene er slik at den beregnede verdi av Q er tallet q. Vi ønsker et mål på hvor oppsiktsvekkende det observerte resultat er i lys av modellen. Et slikt mål vil være

$$P = P_H(Q \geq q)$$

beregnet på grunnlag av modellen H. P er altså sannsynligheten, med den gitte modell, for å få det observerte resultat eller et resultat som taler enda mer i disfavør av modellen. Dersom denne P-verdi er liten, er det observerte resultat overraskende og det kan indikere at modellen H er urealistisk. Hvor liten P-verdi som skal til for at vi forkaster denne modellen, er vårt eget valg, og vil avhenge av omstendighetene, bl.a. den bruk vi hadde tenkt å gjøre av modellen.

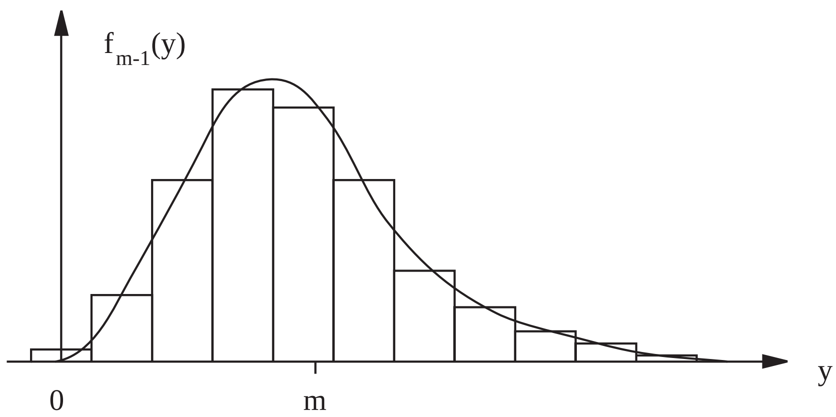
For å beregne P-verdien trenger vi altså å kjenne sannsynlighetsfordelingen til Q under forutsetningene ovenfor. Den eksakte fordeling er generelt vanskelig å etablere, vi må nøye oss med følgende tilnærmelse. [2]

> **Kjikvadrattilnærmelsen:** Sannsynlighetsfordelingen (histogrammet) til Q vil, under forutsetning av at modellen H gjelder, når n vokser, nærme seg en kontinuerlig kurve, kalt *kjikvadratkurven* med $m-1$ frihetsgrader. [a]
>
> [a]Av historiske årsaker: $m-1$ er antall frie parametre; når $m-1$ p_i'er er gitt, er også den siste bestemt.

Det matematiske uttrykk for kurven utelates her. Vi nøyer oss med å fastslå at kurven ikke avhenger av sannsynlighetene $p_1, p_2, ..., p_m$ i modellen, men bare avhenger av antall utfall m. Kurvens tyngdepunkt vil forskyves mot høyre for økende m, som synes rimelig ettersom $EQ = m-1$.
I Figur 7.4 har vi inntegnet et tenkt histogram for Q og tilhørende kjikvadratkurve. Lar vi $\overline{F}_{m-1}(q)$ betegne arealet til høyre for q under kjikvadratkurven med $m-1$ frihetsgrader, vil derfor for store n

$$P_H(Q \geq q) \approx \overline{F}_{m-1}(q).$$

[2]Denne tilnærmelsen tilskrives Karl Pearson (1857-1936), som var en av pionerene i moderne statistisk teori.



Figur 7.4: Kjikvadrattilnærmelsen

For å kunne bruke dette resultatet i praksis trenger vi å ha tabellert $\overline{F}_\nu(q)$ for ulike q og ulike frihetsgradtall ν . En slik tabell er Tabell C.7 i Appendiks C. Vi kunne alternativt ha tabellert $F_\nu(q)$ som er arealet til venstre for q. Da er $\overline{F}_\nu(q) = 1 - F_\nu(q)$, men i forbindelse med kjikvadratkurver trenger en i praksis oftest de høyresidige arealene. Merk at siden kjikvadratkurver må tabelleres for en rekke frihetsgradtall $\nu = 1, 2, 3, \ldots$, kan hver tabell nødvendigvis ikke være så omfattende som tilfellet var med normalkurven. For frihetsgradtall større enn 5 har vi måttet nøye oss med såkalte fraktiltabeller, dvs. angi de punkter q som har bestemte arealer til høyre for seg (Tabell C.8). For kjikvadrattilnærmelsen vil graden av tilnærming avhenge av n og til en viss grad av p_i'ene i modellen, en ser ofte anbefalt at tilnærmelsen er brukbar for de fleste praktiske formål dersom $np_1, np_2, \ldots, np_m$ alle er større enn 5, men også i situasjoner hvor dette ikke er tilfelle vil tilnærmelsen gi en pekepinn på størrelsesorden av sannsynlighetene.

Eksempel 17 : Rettferdig terning? (fortsatt)
Med de data som er gitt ovenfor beregner vi testobservatoren

$$Q = \sum_{i=1}^{m} \frac{(X_i - np_i)^2}{np_i} = \sum_{i=1}^{6} \frac{(X_i - 40)^2}{40} = 5.95$$

Dersom terningen er rettferdig, er $EQ = 5$ og det observerte resultat er ikke særlig mye større enn dette. Under denne forutsetningen kan sannsynlighetsfordelingen til Q tilnærmes med kjikvadratkurven med $6 - 1 = 5$ frihetsgrader. Ifølge Tabell C.7 blir $P(Q \geq 5.95) \approx 0.31$, dvs. sannsynligheten for å observere et resultat som det gitte, eller et som er enda mer avvikende fra "det forventede" $40, 40, \ldots, 40$ målt med Q, er ca. lik 0.31. Det observerte resultat kan derfor ikke sies å være tilstrekkelig oppsiktsvekkende til at det gir grunnlag for å påstå at terningen er falsk.

Eksempel 18 : Reklamasjoner
Et varemagasin har erfart at siste år kom det gjennomsnittlig en reklamasjon pr. åpningsdag. Man ønsker å teste om det er rimelig å anta at antall reklamasjoner på en gitt dag er Poissonfordelt med forventning 1. For dette formål noteres antall reklamasjoner hver dag i en periode på $n = 50$ åpningsdager. Alle opplysninger som trengs er gitt i tabellen

u_i		X_i	p_i	np_i
0		15	0.368	18.4
1		21	0.368	18.4
2		12	0.184	9.2
3		2	0.061	3.0
4	eller flere	0	0.019	1.0

Utfallene $u_i (i = 1, 2, \ldots, 5)$ representerer de ulike antall reklamasjoner som kan inntreffe i løpet av en dag, der vi har gruppert sammen 4 eller flere for å få et endelig og rimelig antall grupper. X_i'ene er antall dager som vi observerte de ulike utfall. De forventede antall dager dersom Poissonmodellen holder er np_i, der de teoretiske Poissonsannsynlighetene p_i er hentet fra Tabell C.4. I dette tilfellet blir

$$Q = \sum_{i=1}^{5} \frac{(X_i - np_i)^2}{np_i} = 3.18$$

Dersom Poissonmodellen benyttes vil $EQ = 4$. Her er observert Q noe mindre og det observerte resultat gir derfor intet grunnlag for å vrake Poissonmodellen. Her kan fordelingen til Q tilnærmes med kjikvadratkurven med $5 - 1 = 4$ frihetsgrader og sannsynligheter for det gitte resultat eller et mer avvikende vil faktisk være mer enn 50% når Poissonmodellen benyttes.

Det er også mulig å teste om en Poissonmodell er realistisk uten antatt forventning. Denne må i så fall estimeres ut fra dataene ved gjennomsnittet. Framgangsmåten ved testingen er ellers den samme, men frihetsgradtallet er en mindre.

I de to eksemplene ovenfor ga det observerte tallmateriale ingen grunn til å vrake den oppsatte modell. Vi kunne ønske oss et mer omfattende materiale for å etterprøve modellene, og det kan ikke utelukkes at dette vil kunne føre til at modellene vrakes på et senere tidspunkt. En må imidlertid være klar over at i mange situasjoner vil en modell være brukbar for sitt formål selv om den er noe grov. Et meget stort tallmateriale vil derfor lett kunne gi forkastning av en grov modell også når denne avviker ubetydelig, for praktiske formål, fra en mer realistisk modell. Dette er ikke ment som en advarsel mot å samle inn ytterligere materiale for å etterprøve modeller, snarere tvert imot. Man bør bare være bevisst hvilket formål modellen skal tjene og selv foreta rimelige avveininger med hensyn til realisme og brukbarhet av den oppsatte modell, jfr. de betraktninger som ble gjort i Kapittel 1.

⋆Med den terminologi vi har innført ovenfor vil generelt $X_1, X_2, \ldots, X_m$ være multinomisk fordelt med parametre $(n, p_1, p_2, \ldots, p_m)$ se Kapittel 6.8, og problemstillingen er å teste en hypotese H om disse parametrene. Som testobservator brukes Q og testmetoden består i at hypotesen forkastes når $Q \geq k$, der k er en passende valgt konstant. Ønskes signifikansnivå α må k velges slik at $P_H(Q \geq k) = \alpha$. For øvrig vil forkastning være ekvivalent med at P-verdien er mindre enn α. Med dette kan vi altså fastlegge sjansen for å forkaste hypotesen (modellen) når denne er riktig (realistisk). På den annen side ønsker vi å studere sjansen for å la være å forkaste en hypotese som er feil (modell som åpenbart er urealistisk). Med andre ord ønsker vi studere styrkefunksjonen til testen. Dette er et relativt komplisert problem, bl. a. fordi en slik funksjon vil avhenge av alle $p_1, p_2, \ldots, p_m$, og vi kan ikke gå inn på det her. Vi skal imidlertid ikke underslå at det finnes andre testmetoder som er velegnet dersom man har spesielle alternativer til hypotesen i tankene. Kjikvadrattesten som vi har presentert ovenfor er imidlertid relativt enkel å utføre og benyttes ofte i praksis.

Den problemstilling som er presentert i dette avsnittet svaret til å teste en helspesifisert hypotese om parametrene i en multinomisk modell. Problemstillingen kan utvides til å teste delvis spesifiserte hypoteser i denne modellen. Dette kan bl.a. brukes til å teste om observasjoner er forenlig med en bestemt fordelingstype, og ved analyse av krysstabellerte kategoridata (se Kapittel 9), som er ofte forekommer ved meningsmålinger og markedsundersøkelser.

7.8 Oppgaver

1. I en binomisk situasjon med suksessannsynlighet p observeres antall suksesser X i løpet av n forsøk.

(a) Finn en forventningsrett estimator for fiaskosannsynligheten $q = 1 - p$, og beregn dennes varians.

(b) Finn en forventningsrett estimator for forskjellen $r = q - p = 1 - 2p$, og beregn dennes varians.

2. I en hypergeometrisk situasjon der antall spesielle elementer i populasjonen er M, observeres antall spesielle elementer Y i et utvalg på n elementer.

   (a) Finn en forventningsrett estimator for M og angi dens varians.

   (b) Finn en forventningsrett estimator for antall vanlige elementer i populasjonen $N - M$ og angi dens varians.

3. La X være antall suksesser i løpet av n binomiske forsøk. Som estimator for suksessannsynligheten p brukes $\hat{p} = X/n$. Finn $P(\mid \hat{p} - p \mid \leq 0.1)$ i følgende situasjoner

$$\begin{array}{llll} \text{(a)} & p = 0.5 & \text{og} & n = 5,\ 10,\ 25,\ 50,\ 100 \\ \text{(b)} & p = 0.4 & \text{og} & n = 5,\ 10,\ 25,\ 50,\ 100 \\ \text{(c)} & p = 0.2 & \text{og} & n = 5,\ 10,\ 25,\ 50,\ 100. \end{array}$$

   Bruk normaltilnærming i situasjoner der binomiske tabeller ikke rekker.

4. (a) Ta for deg et kronestykke, og knips den n ganger. Estimer sannsynligheten for kron p og oppgi feilmarginer dersom (i) $n = 10$ (ii) $n = 100$.

   (b) Ta for deg en tegnestift, og kast den n ganger. Estimer sannsynligheten for at den faller med spissen i været p og oppgi feilmarginer dersom (i) $n = 10$ (ii) $n = 100$.

   (c) Ta for deg en terning, og kast den n ganger. Estimer sannsynligheten for å få sekser p og oppgi feilmarginer dersom (i) $n = 10$ (ii) $n = 100$.

5. Finn fram en pose erter og tell opp et visst antall, f. eks. $N = 100$. La en venn merke en del av ertene før alle legges i en boks, uten at du får vite hvor mange som er merket. Forsøk så å anslå andelen og antallet merkede erter ut fra et utvalg på $n = 10$ erter. Gjenta eksperimentet med et annet blandingsforhold. Forsøk også andre stikkprøvestørrelser, f.eks. $n = 25$. Vurder i hvert tilfelle risikoen for estimeringsfeil. Er det mulig å besvare spørsmålene dersom du heller ikke vet antallet kuler i boksen?

6. Betrakt populasjonen av $N = 216$ studenter i Eksempel 1.7. Estimer på grunnlag av en stikkprøve andelen studenter
   (i) med mindre enn 10 kroner på seg,
   (ii) med mer enn 100 kroner på seg.
   Vurder påliteligheten av estimatet. Forsøk ulike stikkprøvestørrelser, f. eks. $n = 10$ og $n = 25$. Angående trekningsmåten, se Oppgave 1.9.

7. To estimatorer $\hat{\theta}$ og $\check{\theta}$ for en parameter θ er foreslått. Hvilken vil du velge dersom deres sannsynlighetsfordelinger er henholdsvis

k	$\theta - 2$	$\theta - 1$	θ	$\theta + 1$	$\theta + 2$
$P(\hat{\theta} = k)$	0.1	0.2	0.4	0.2	0.1
$P(\check{\theta} = k)$	0.2	0.0	0.6	0.0	0.2

8. Anta X_1 og X_2 forventningsrette estimatorer for henholdsvis θ_1 og θ_2.

   (a) Lag med utgangspunkt i X_1 og X_2 forventningsrette estimatorer for (i) $\theta_1 + \theta_2$ (ii) $\theta_1 - \theta_2$ (iii) $(\theta_1 + \theta_2)/2$.

   Anta at X_1 og X_2 er uavhengige med samme varians σ^2.

   (b) Finn variansen til estimatorene i (a).

9. Anta at X_1 og X_2 er stokastiske variable med forventninger $EX_1 = \theta_1 + \theta_2$ og $EX_2 = \theta_1 - \theta_2$.

   (a) Lag med utgangspunkt i X_1 og X_2 forventningsrette estimatorer for θ_1 og θ_2.

   Anta at X_1 og X_2 er uavhengige med samme varians σ^2.

   (b) Finn variansen til estimatorene i (a).

10. La $\hat{\theta}$ være en forventningsrett estimator for θ med fordeling som kan tilnærmes med normalkurven.

    (a) Dersom $\sigma(\hat{\theta})$ er 0.6, finn tilnærmet $P(\mid \hat{\theta} - \theta \mid \leq 0.4)$.

    (b) Dersom $\sigma(\hat{\theta})$ er 0.8, finn d slik at $P(\mid \hat{\theta} - \theta \mid \leq d) \approx 0.85$.

    (c) Dersom vi krever $P(\mid \hat{\theta} - \theta \mid \leq 0.7) \geq 0.80$, hvor liten må standardavviket til $\hat{\theta}$ i så fall være.

11. En bedrift masseproduserer en artikkel. Ut fra erfaring fra de senere ukers produksjon antar man at defektsannsynligheten er p_0. Det utføres så en overhaling av produksjonsutstyret, og etter denne antas defektsannsynligheten å være p (mindre enn p_0); p ønskes estimert.

    (a) Hvor mange artikler n må minst undersøkes for at standardavviket til estimatoren skal bli høyst 0.01 dersom
    (i) $p_0 = 0.05$ (ii) $p_0 = 0.10$ (iii) $p_0 = 0.15$.

    (b) Finn i hvert av disse situasjonene tilnærmet sannsynligheten for at estimatoren skal avvike høyst 0.015 enheter fra den korrekte p.

12. En bedrift masseproduserer artikler. Man har erfart at defektprodusenten vanligvis ikke overstiger a_o. En dag er det produsert N artikler, hvorav antall defekte M er ukjent. Man ønsker å estimere brøkdelen $a = M/N$ av defekte artikler på grunnlag av et tilfeldig utvalg fra dagsproduksjonen.

    (a) Hvor mange artikler n er det rimelig å velge ut dersom en ønsker en estimator med standardavvik høyst lik 0.01 i tilfellene

(i)	$N = 100$	$a_0 = 0.05,\ 0.10,\ 0.15$
(ii)	$N = 500$	$a_0 = 0.05,\ 0.10,\ 0.15$
(iii)	$N = 1000$	$a_0 = 0.05,\ 0.10,\ 0.15$
(iv)	$N = 10000$	$a_0 = 0.05,\ 0.10,\ 0.15$

(b) Finn i hvert av disse tilfellene tilnærmet sannsynligheten for estimatoren skal avvike høyst 0.015 enheter fra den korrekte a.

13. Forskjellen i defektsannsynlighet med to produksjonsmetoder $p_1 - p_2$ ønskes anslått ut fra andelen defekte blant henholdsvis n_1 og n_2 produserte artikler etter de to metodene.

 (a) Finn standardavviket til estimatoren i tilfellet at $p_1 = p_2 = p$

 (b) Finn tilnærmet sannsynligheten for at estimatoren gir verdi mellom -0.1 og 0.1 dersom $n_1 = n_2 = 20$ og
 (i) $p_1 = p_2 = 0.5$ (ii) $p_1 = p_2 = 0.3$.
 Hint: Bruk normaltilnærmelse, jfr. Kapittel 7.6..

 (c) Hvor mange artikler av hvert slag må produseres for at standardavviket skal bli høyst 0.05?

14. Vis formelen

$$E(\hat{\theta} - \theta)^2 = var\hat{\theta} + (E\hat{\theta} - \theta)^2$$

 Hint: Skriv $(\hat{\theta} - \theta)^2 = (\hat{\theta} - E\hat{\theta} + E\hat{\theta} - \theta)^2$ og kvadrer ut.

15. La X være binomisk fordelt (n, p). Som estimator for p er foreslått $\tilde{p} = (X + 1)/(n + 2)$. Vis at

 (a) $E\tilde{p} = p + \frac{1-2p}{n+2}$

 (b) $E(\tilde{p} - p)^2 = \frac{1}{(n+2)^2}(np(1-p) + (1-2p)^2)$

 Sammenlign denne estimatoren med den vanlige $\hat{p} = X/n$ når $n = 5$ og når $n = 50$ (Tegn figur!).

16. La X være binomisk fordelt (n, p). En estimator for p er $p_0 = \frac{1}{2}$, dvs. det observerte resultat blir neglisjert og $\frac{1}{2}$ oppgis som estimat uansett. Sammenlign denne estimator med $\hat{p} = X/n$ og $\tilde{p} = (X + 1)/(n + 2)$.

17. En bedrift består av to avdelinger. Man ønsker å estimere sannsynligheten for at en vilkårlig valgt arbeider utsettes for en viss type arbeidsuhell i løpet av et år, denne antas å være den samme lik p i begge avdelinger. Anta at det siste år var n_i arbeidere ved avdelingen nr. i, hvorav X_i ble utsatt for uhell $(i = 1, 2)$. To estimatorer er foreslått

$$\hat{p} = \frac{X_1 + X_2}{n_1 + n_2} \quad \text{og} \quad \tilde{p} = \frac{1}{2}(\frac{X_1}{n_1} + \frac{X_2}{n_2})$$

(a) Påvis at begge er forventningsrette.

(b) Beregn variansen til estimatorene.

(c) Vis at $\hat{p}$ er å foretrekke framfor $\tilde{p}$ uansett hva n_1 og n_2 er.

Hint: Påvis at $\frac{1}{n_1+n_2} \leq \frac{1}{4}(\frac{1}{n_1} + \frac{1}{n_2})$

18. En bedrift har et antall på 200 av en bestemt artikkel på lager, men bare 100 av disse er det plass til innendørs. Etter lengre tid på lager vil en del artikler ikke lenger oppfylle et gitt kvalitetskrav, og ved en lagerinspeksjon ønsker man å estimere det totale antall mindreverdige artikler $M = M_1 + M_2$ på lager. (M_1 inne, M_2 ute.) Tre framgangsmåter vurderes som alle består i å undersøke 40 artikler.

    (a) De 40 velges tilfeldig blant alle 200 og antall mindreverdige Y noteres.

    (b) Det velges tilfeldig 20 artikler inne og 20 ute og antall mindreverdige Y_1 og Y_2 noteres.

    (c) Som (b) men det velges 10 inne og 30 ute.

    Foreslå forventningsrette estimatorer for $M = M_1 + M_2$ i tilfellene (a), (b) og (c) og angi deres varianser. Diskuter hvorvidt en av framgangsmåtene er å foretrekke.

19. Et varehus ønsker å studere holdningen til bakgrunnsmusikk i salgslokalene, og vil ikke innføre det med mindre det er nokså sikkert at minst halvparten av kundene vil like det. I et utvalg på 800 kunder sa 424 at de ville like bakgrunnsmusikk. Hva er din konklusjon?

20. Med sikte på å klarlegge tilslutningen til en rekke politiske partier hos en befolkning på, la oss si, 3 millioner stemmeberettigede personer er det valgt ut tilfeldig $n = 2500$ personer som har avgitt svar m.h.t. partipreferanse (i praksis vil en neppe kunne gjennomføre et rent tilfeldig utvalg, men anta for enkelhets skyld at dette er gjort her). Av disse 2500 personene ga 1132 personer sin tilslutning til A-partiet, 570 til B-partiet, 298 til C- partiet,. . . , mens 109 personer ga sin tilslutning til K- partiet.

    (a) Rapporter disse resultatene.

    (b) Hvilke feilmarginer må man regne med ved tolking av disse tallene.

    (c) Hvilke feilmarginer må man regne for disse partiene dersom undersøkelsen istedet var basert på $n = 1800$ personer (jfr. Oppgave 1.4).

21. (a) Ved estimering av en sannsynlighet p drøft bruk av

$$\hat{p} \pm k \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

som konfidensintervall for p med tilnærmet konfidensnivå $c = A(k)$.

(b) Ved estimering av forskjellen mellom to sannsynligheter $p_1 - p_2$ drøft bruk av

$$\hat{p}_1 - \hat{p}_2 \pm k \cdot \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

som konfidensintervall med konfidensnivå $c = A(k)$.

Drøft også (a) og (b) dersom det dreier seg om estimering av andeler basert på utvalg fra en endelig populasjon.

22. Et markedsføringsfirma studerer sjansen p for at en person i en bestemt kundegruppe er kjent med innholdet av en kampanjeannonse. Anta at det blant $n = 400$ tilfeldige personer var 175 som kjente innholdet.

    (a) Lag et konfidensintervall med tilnærmet konfidensnivå 95% for sannsynligheten p.

    Anta at det blant $n_1 = 210$ tilfeldige kvinner var 99 som kjente innholdet, mens det blant $n_2 = 190$ tilfeldige menn var 76 som kjente innholdet.

    (b) Lag et konfidensintervall med tilnærmet konfidensnivå 95% for forskjellen mellom sannsynlighetene $p_1 - p_2$ for kvinner og menn.

    (c) Vil en kunne rettferdiggjøre intervallet i (b) dersom det isteden dreide seg om et tilfeldig utvalg på 400 personer hvor det viste seg at 210 var kvinner og 190 menn?

23. Ledelsen i en fagforening har sendt ut informasjon til sine 2500 medlemmer, der 1300 er kvinner og 1200 menn. Ut fra en spørreundersøkelse ønsker ledelsen å anslå andelen medlemmer a som har satt seg inn i informasjonen. Besvar spørsmålene i Oppgave 22 med andeler istedenfor sannsynligheter, når tallene ellers er de samme.

24. La X være binomisk (n, p) der p er ukjent. Lag tilnærmede tosidige konfidensintervaller for p basert på Poissontilnærming i tilfellene

    (a) n=100 og X = 1, 2, 5 og 10

    (b) n=200 og X = 1, 2, 5 og 10

    når vi ønsker tilnærmet konfidensnivå 95%.
    Lag også intervaller basert på normaltilnærming og vurder brukbarheten av garantien i hvert tilfelle.

25. Et firma er interessert i defektsannsynligheten p for en bestemt artikkel produsert av en underleverandør. I en prøveproduksjon på n artikler var X defekte. Lag en øvre konfidensgrense for p med (ensidig) konfidensnivå 95% basert på Poissontilnærming i tilfellene

    (a) $n = 100$ og $X = 1$ (b) $n = 200$ og $X = 1$ (c) $n = 200$ og $X = 2$

Dersom defektsannsynligheten er ca. 0.005, hvor stor må n være for at ventet øvre grense er 0.01

26. La situasjonen være som beskrevet i Eksempel 6 der man ønsket å finne ut om den nye produksjonsmetoden gjennomgående gir lavere defektprosent enn den gamle.

    (a) Formuler situasjonen som et hypotesetestingsproblem og rapporter resultatet når 5 av 100 artikler viste seg å være defekte.

    (b) Er resultatet signifikant (dvs. gir det grunn til å påstå forbedring) dersom en har valgt signifikansnivå 5%? Enn 10%?

27. Formuler situasjonen i Eksempel 7 som et hypotesetestingsproblem.

    (a) Er det observerte resultat signifikant dersom signifikansnivået er valgt lik 10%.

    (b) Finn den kritiske verdi dersom en ønsker 5% signifikansnivå.

28. En pedagog påstår at minst halvparten av ungdomsskoleelevene tilbringer mer tid foran TV skjermen pr. uke enn på skolen. I en studie som omfattet 300 ungdomsskoleelever var dette tilfelle for 130 av dem? Er det grunnlag for å slutte seg til pedagogens påstand? Hva er rimelig å bruke som nullhypotese i denne situasjon?

29. En bedrift ønsker å markedsføre en bestemt vare i to varianter, den ene beregnet for diabetikere, men man ønsker at smaken skal være så godt som identisk. For å teste produktene har man latt $n = 200$ personer få smake tre smaksprøver, den ene av disse er diabetikervaren, og hver blir bedt om å peke ut en av prøvene som de mener smaker annerledes enn de to andre. Det viste seg at 76 personer korrekt identifiserte diabetikervaren.

    (a) Formuler situasjonen som et hypotesetestingsproblem og rapporter resultatet.

    (b) Hvilken konklusjon trekker bedriften dersom man har bestemt seg for et signifikansnivå på 5%. Enn 10%.

30. Frø av en bestemt type har vanligvis spiresannsynlighet lik 0.95. Ved en planteskole frykter en at forurensing i omgivelsene medfører at frøene deres har redusert spireevne. For å anslå den ukjente spiresannsynligheten p for disse frøene såes n frø og antall som spirer X observeres.

    (a) Anslå p dersom

    $$\text{(i) } n = 25 \text{ og } X = 22 \qquad \text{(ii) } n = 100 \text{ og } X = 91.$$

    Hvilke feilmarginer må vi regne med i hvert av disse anslagene?

    (b) Gir datamaterialet grunnlag for å påstå at spireevnen er dårligere enn den vanlige 0.95 i tilfellene (i) og (ii)?

(c) Drøft valget av n i (a) og (b).

(d) Drøft alternative opplegg for å sammenligne disse frøene med andre typer frø mht. spireevne.

31. La X være binomisk fordelt (n, p). Vi ønsker å teste $H_0 : p = p_0$ mot $H_A : p < p_0$

(a) Forklar at det er rimelig å forkaste H_0 og påstå H_A når X er liten, og lag en test med signifikansnivå α i tilfellene

$$\begin{array}{rllll} \text{(i)} & n{=}10, & p_0 = 0.5 & \text{og} & \alpha = 0.10 \\ \text{(ii)} & n{=}30, & p_0 = 0.5 & \text{og} & \alpha = 0.10 \\ \text{(iii)} & n{=}30, & p_0 = 0.2 & \text{og} & \alpha = 0.10 \end{array}$$

(b) I tilfelle (iii) observeres $X = 3$. Hva blir konklusjonen? Beregn også P-verdien til det observerte resultat og forklar hva dette innebærer.

(c) Hvilken styrke har testene i (i) og (ii) mot alternativet $p = 0.3$? Skisser styrkefunksjonen for de to tilfellene.

(d) Hvor mange observasjoner må vi ha for at styrken mot alternativet $p = 0.3$ skal bli ca. 0.90 når $p_0 = 0.5$?

32. La X være binomisk fordelt (n, p). Vi ønsker å teste

$$H_0 : p = p_0 \quad \text{mot} \quad H_A : p \neq p_0.$$

(a) Forklar at det er rimelig å forkaste H_0 og påstå H_A når $\mid X - np_0 \mid$ er stor.

(b) Lag en test med signifikansnivå α i tilfellene

$$\begin{array}{rllll} \text{(i)} & n = 10, & p_0 = 0.5 & \text{og} & \alpha = 0.10 \\ \text{(ii)} & n = 30, & p_0 = 0.5 & \text{og} & \alpha = 0.10 \\ \text{(iii)} & n = 30, & p_0 = 0.2 & \text{og} & \alpha = 0.10. \end{array}$$

(c) Skisser styrkefunksjonene for testene i (b).

(d) Et kronestykke blir knipset $n = 100$ ganger og antall kron X ble observert til 62. Forklar at P-verdien til det observerte resultat blir $P(\mid X - 50 \mid \geq 12)$ med $p = 0.5$. Beregn denne og forklar hva resultatet innebærer.

33. Et firma med tre avdelinger A, B, C studerer effekten av opplæring for å bedre påliteligheten av ordreregistreringer. Resultatet ble

	Før opplæring			Etter opplæring		
	A	B	C	A	B	C
Antall registreringer	258	220	293	409	332	420
Antall med feil	56	35	59	68	20	76

Test om opplæring har hatt effekt ved hver enkelt avdeling og samlet.

34. En produsent av frokostgryn leverer pakninger i tre størrelser, liten, medium og stor, og har erfart at disse etterspørres i forholdet 1 : 3 : 2 når alle størrelser har samme eksponering. Produsenten har laget en variant av produktet med rosiner, og prøvelanserer dette i et område.

Pakning	Liten	Medium	Stor	Totalt
Antall	240	575	385	1200

Tyder resultatet på at etterspørselen etter det nye produktet fordeler seg på samme måte?

35. En artikkel masseproduseres og sorteres i tre kvaliteter a, b og c. Et omfattende observasjonsmateriale viser at med den produksjonsmetode som er i bruk, fordeler artiklene seg gjennomsnittlig i lange løp med 25% av kvalitet a, 62% av kvalitet b og 13% av kvalitet c. Man overveier å gå over til en ny produksjonsmetode, og det prøveproduseres $n = 50$ artikler med den nye metoden og disse kvalitetssorteres. Resultatet ble at blant de 50 ble 21 av kvalitet a, 25 av kvalitet b og 4 av kvalitet c. Tyder dette på at kvalitetsfordelingen ved den nye metoden er en annen enn med den gamle?

36. Ta for deg de 100 første tallene i en tabell over tilfeldige tall, evt. fra en siffergenerator på en lommeregner eller PC. Tell opp antall ganger hvert siffer forekommer, og utfør en test for å avsløre om siffergeneratoren ikke er i orden.

37. Testing av uavhengighetsantakelsen i binomiske forsøk med $p = \frac{1}{2}$ kan skje ved å telle opp antall såkalte følger K. Eksempelvis består følgende serie med n=20 myntkast av ialt K=7 følger (her adskilt med vertikale streker).

$$|MMMMMMM|KKK|M|K|MMMM|KK|MM|$$

Det kan vises at

$$EK \approx \frac{n+2}{2} \quad varK \approx \frac{n-1}{4}$$

og at normaltilnærming kan brukes dersom n ikke er for liten (bør helst være større enn 25). Bruk datamaterialet ovenfor til å teste hypotesen at mynten er rettferdig og uten minne (uavhengighet).

38. $\star$Anta at X har sannsynlighetsfunksjon $p_\theta(x)$, som avhenger av en ukjent parameter θ. Den såkalte *sannsynlighetsmaksimeringsestimatoren* bestemmes ved den verdi av θ som gjør den observerte X mest sannsynlig. Denne kan finnes ved å maksimere $L_x(\theta) = p_\theta(x)$ som funksjon av θ, den såkalte rimelighetsfunksjonen. Finn sannsynlighetsmaksimeringsestimatoren dersom

    (a) X er binomisk fordelt (n, p)

(b) X Poissonfordelt med forventning λ.

39. Ta for deg en statistisk programpakke, og finn ut om den kan brukes til å (i) lage konfidensintervaller for sannsynligheter (ii) teste hypoteser om sannsynligheter (iii) planlegge antall observasjoner for å få en test med ønsket nivå og styrke.

Kapittel 8

Statistisk inferens: Forventninger og gjennomsnitt

8.1 Målemodellen

En problemstilling som dukker opp i mange sammenhenger er: En ukjent størrelse μ skal måles, men vi har erfart at gjentatte målinger av μ vanligvis ikke gir samme resultat, selv om vi forsøker å måle under de samme forsøksbetingelser hver gang. Dette skyldes ofte en ufullkommen metode, enten ved at måleverktøyet bare måler med en begrenset nøyaktighet, eller andre forstyrrende faktorer som ligger utenfor vår kontroll. I noen situasjoner kan nøyaktig måling av μ være praktisk (endog teoretisk) umulig. Vi sier at målingene er foretatt med *målefeil*. Ved å foreta gjentatte målinger håper vi at vi kan redusere den usikkerhet som er forbundet med en enkelt måling. De første statistiske teorier på 1700-tallet [1] tok nettopp sikte på å legge et teoretisk grunnlag for behandling av målefeil.

Vi skal i dette avsnittet studere en enkel statistisk modell som har vist seg fruktbar ved analyse av usikkerhet i målefeil. Selv om modellen er gitt navnet *målemodellen*, har den langt bredere anvendelse og er i dag en del av grunnlaget for usikkerhetsvurderinger ved observasjoner innen de fleste fagområder.

Siden vi skal ha en modell for måling under usikkerhet, er det rimelig å formulere en stokastisk modell: En måling av en ukjent størrelse μ oppfattes

[1] Nøkkelpersonene var matematikerne Laplace (1749-1827) og Gauss (1777-1855).

som en realisasjon av en stokastisk variabel X med forventning μ. Foretar vi n gjentatte målinger av μ, kan vi oppfatte disse som realisasjoner av n stokastiske variable $X_1, X_2, \ldots, X_n$ alle med samme forventning μ. Det er ofte rimelig å anta at usikkerheten i hver måling er den samme. Vi uttrykker dette i modellen ved å anta at de stokastiske variable $X_1, X_2, \ldots, X_n$ har den samme varians, kall den σ^2. Videre er det ofte rimelig å anta at målingene ikke avhenger av hverandre. Vi uttrykker dette i modellen ved å anta at $X_1, X_2, \ldots, X_n$ er uavhengige stokastiske variable (se Kapittel 5.6). Vi har derfor:

Målemodellen: $X_1, X_2, \ldots, X_n$ er n uavhengige stokastiske variable der

$$EX_1 = EX_2 = \cdots = EX_n = \mu$$

$$varX_1 = varX_2 = \cdots = varX_n = \sigma^2.$$

I denne modellen er μ og σ^2 parametre hvis verdier vanligvis er ukjente. Hensikten med å foreta de n målingene er nettopp å få informasjon om μ og helst også om σ^2, dersom denne også er ukjent. Eksempler på situasjoner hvor målemodellen kan være realistisk er:

- Gjentatte veiinger av en gjenstand med sann vekt μ. σ^2 uttrykker da usikkerheten i hver veiing.

- Alkoholkonsentrasjonen i n blodprøver tatt på samme person. μ er den "sanne promille", og σ^2 uttrykker usikkerheten i en enkelt blodprøve.

- Måling av strekkstyrken av n nylonsnører av en bestemt type. Her er μ gjennomsnittlig strekkstyrke av slike nylonsnører, mens σ^2 uttrykker variasjonen i strekkstyrke fra snøre til snøre.

- Bestemmelse av kvikksølvkonsentrasjonen i n lagesild oppfisket fra Mjøsa. Her uttrykker μ den gjennomsnittlige konsentrasjon av kvikksølv hos lagesild i Mjøsa, mens σ^2 uttrykker variasjonen i konsentrasjonen.

Merk at de to første eksemplene dreier seg om usikkerhet i form av målefeil, mens det tredje eksemplet dreier seg om usikkerhet ved at de målte objekter kun er representanter fra en større, prinsipielt uendelig, populasjon med en viss innbyrdes variasjon. Det samme gjelder det siste eksemplet. I de to siste eksemplene vil variansen σ^2 også kunne inneholde en viss komponent av direkte målefeil, men denne vil vanligvis være neglisjerbar i forhold til de individuelle variasjoner hos måleobjektene. [2]

I alle eksemplene kan den aktuelle variable måles på en prinsipielt kontinuerlig skala, selv om resultatet i praksis måles på en diskret skala, f.eks. kvalitet (strekkstyrke) av et produkt målt i kilo med to desimaler. Målemodellen tillater imidlertid diskrete variable, f.eks antall registrerte feil ved et produkt.

I praksis ser man ofte data på en vurderingsskala, f.eks kvalitet vurdert i kategorier fra 1 til 7. Da er det ikke lenger sikkert at analyse basert på målemodellen er meningsfull. Kan hende reflekterer tallene bare såkalte *ordinale* egenskaper ved kvaliteten, at 2 er bedre enn 1 osv. Det har da liten mening å bruke begrepet forventet kvalitet. Det kan imidlertid være grensetilfeller der det kan diskuteres om ikke målemodellen likevel kan brukes. Betrakt følgende to situasjoner:

1. En kontrollør vurderer n produkter i den løpende produksjon.
2. Et produkt eller en tjeneste blir vurdert av n kunder.

I den første situasjonen kan en nok tenke seg at retningslinjer for skalering er gitt, slik at målemodellen kan brukes. Den andre situasjonen er mer problematisk, spesielt fordi ulike kunder ikke nødvendigvis legger det samme i kategoribeskrivelsene, f.eks. 1=helt ubrukelig, 2=nesten ubrukelig osv. Det fins en omfattende litteratur om vurderingsskalaer (hovedsakelig med utgangspunkt i psykologi) som søker å løse dette problem, enten ved å etablere skaleringsmetoder slik at målemodellen likevel kan brukes, eller ved å basere analysen på alternative modeller.

La oss i første omgang studere problemet å estimere μ på basis av de n observasjonene $X_1, X_2, \ldots X_n$. I praksis har det ofte vært vanlig å bruke gjennomsnittet av observasjonene

$$\hat{\mu} = \overline{X} = \frac{1}{n}(X_1 + X_2 + \cdots + X_n)$$

som estimator for den felles forventning μ. Denne estimatoren har bl.a. følgende egenskaper

$$\boxed{E(\overline{X}) = \mu \qquad var(\overline{X}) = \frac{\sigma^2}{n}}$$

[2] Eksempel 6 i Kapittel 1 vil også kunne motivere deg for teorien i dette avsnittet, og anbefales lest (evt. repetert) nå.

Vi ser at $\overline{X}$ er en forventningsrett estimator for μ, med varians som avhenger av variansen til enkeltobservasjonene og antall observasjoner.

Begrunnelse: Merk de overganger der vi benytter modellantakelser

$$
\begin{aligned}
E(\overline{X}) &= E(\frac{1}{n}(X_1 + X_2 + \cdots + X_n)) = \frac{1}{n}E(X_1 + X_2 + \cdots + X_n) \\
&= \frac{1}{n}(EX_1 + EX_2 + \cdots + EX_n) = \frac{1}{n}(\mu + \mu + \cdots + \mu) \\
&= \frac{1}{n} \cdot n \cdot \mu = \mu. \\
var(\overline{X}) &= var(\frac{1}{n}(X_1 + X_2 + \cdots + X_n)) = \left(\frac{1}{n}\right)^2 var(X_1 + X_2 + \cdots + X_n) \\
&= \left(\frac{1}{n}\right)^2 (varX_1 + varX_2 + \cdots + varX_n) \\
&= (\frac{1}{n})^2(\sigma^2 + \sigma^2 + \cdots + \sigma^2) = (\frac{1}{n})^2 n\sigma^2 = \frac{\sigma^2}{n}.
\end{aligned}
$$

Ved vurdering av påliteligheten til $\overline{X}$ som estimator for μ bruker vi standardavviket til $\overline{X}$. Vi har

Kvadratrotloven :

$$\Delta = \sigma(\overline{X}) = \frac{\sigma}{\sqrt{n}}.$$

Estimatoren kan gjøres så presis man ønsker ved å ta et tilstrekkelig stort antall observasjoner. Imidlertid sier kvadratrotloven at vi, for å halvere usikkerheten målt ved standardavviket, må firedoble antall observasjoner.

Siden $\overline{X}$ oppfyller betingelsene i sentralgrensesetningen, kan vi bruke normaltilnærming såframt n ikke er for liten (se Kapittel 6.5). Vi får

$$P(\mid \overline{X} - \mu \mid < k \cdot \sigma(\overline{X})) \approx A(k)$$

På samme måte som i forrige kapittel, kan dette brukes til å lage sannsynlighetsutsagn om estimeringsfeilen. Også her er det rimelig å rapportere

$$\text{estimat} \pm \text{standardavviket til estimatoren}$$

som i denne situasjonen blir

$$\overline{X} \pm \sigma(\overline{X}) \quad \text{der} \quad \sigma(\overline{X}) = \frac{\sigma}{\sqrt{n}}$$

Eksempel 1 : Alkotest
En bestemt metode til å måle alkoholkonsentrasjon i blod er forbundet med en betydelig usikkerhet. Følgelig tas det gjentatte målinger. Anta at en enkelt måling kan oppfattes som realisasjon av en stokastisk variabel X med forventning den "sanne"alkoholkonsentrasjon μ i promille og med standardavvik lik σ. Fra tidligere erfaring har det vist seg at $\sigma = 0.20$ uansett hvilken μ vi har. La oss anta at vi har tatt $n = 3$ målinger hos en person rett etter hverandre, og betegn disse med X_1, X_2, X_3. Vi ønsker på dette grunnlag å estimere μ. Dersom det er rimelig å anta at observasjonene er uavhengige av hverandre, har vi et tilfelle av målemodellen. Anta at observasjonene var

$$0.40 \quad 0.65 \quad 0.75$$

Her blir $\overline{X} = \frac{1}{3}(0.40 + 0.65 + 0.75) = 0.60$ og $\sigma(\overline{X}) = 0.20/\sqrt{3} = 0.115$, og vi rapporterer derfor at alkoholkonsentrasjonen er

$$0.60 \quad \pm \quad 0.115$$

Jeg tror ikke jeg ville ta sjansen på å sikte vedkommede på dette grunnlag (i hvertfall ikke dersom forsvareren har litt greie på statistikk).

For å kunne angi standardavviket $\sigma(\overline{X})$ til estimatoren må vi kjenne standardavviket σ til enkeltmålingene. I mange situasjoner er imidlertid σ helt eller delvis ukjent. Vi må da nøye oss med et anslag for σ. Noen ganger kan man utnytte informasjon fra tidligere, om tilsvarende målinger foretatt med om lag samme presisjon. Uten slik informasjon er vi henvist til å anslå σ på grunnlag av de foreliggende observasjonene. La oss forsøke å finne en brukbar estimator for σ^2 på grunnlag av observasjonene $X_1, X_2, \ldots, X_n$. Vi har

$$\sigma^2 = E(X_1 - \mu)^2 = E(X_2 - \mu)^2 = \cdots = E(X_n - \mu)^2$$

Herav følger at $E(\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2) = \sigma^2$. Hadde μ vært kjent ville derfor

$$\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2$$

være en forventningsrett estimator for σ^2. Når μ er ukjent erstatter vi μ med sin estimator $\overline{X}$ i dette uttrykket, og håper på at

$$S_X^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

er en brukbar estimator for σ^2. Denne størrelsen kalles ofte for *den empiriske varians*. Vi ser at den empiriske varians gir uttrykk for spredningen av de n observasjonene. Den er det empiriske motstykke til det modellteoretiske begrep varians. [3] Kvadratroten av S_X^2, kalt S_X, omtales ofte som *det empiriske standardavvik*. Den empiriske varians kan også uttrykkes på andre måter, som i en gitt situasjon kan være mer velegnet for beregningsformål (se f.eks. Oppgave 1.11).

Vi vil studere den foreslåtte estimator litt nærmere, og spør først om S_X^2 er en forventningsrett estimator for σ^2. Det viser seg at (se $\star$-merket stoff til slutt i avsnittet)

$$E(S_X^2) = \frac{n-1}{n}\sigma^2$$

slik at den foreslåtte estimator ikke er forventningsrett. Faktoren $(n-1)/n$ nærmer seg 1 når n vokser, og skjevheten gjør seg derfor bare gjeldende for små n. Vi sier at estimatoren er asymptotisk forventningsrett. Det funne resultat viser imidlertid at skjevheten kan korrigeres ved å dividere med $n-1$ istedenfor n. Det er vanlig å bruke den korrigerte kvadratsummen som estimator for σ^2 med betegnelsen S^2, og å omtale denne som den empiriske varians. Noen legger mindre vekt på forventningsretthet, og lar være å foreta korreksjon. Vi ser at ulik praksis bare slår ut for små n.

Anta at vi har valgt å bruke $\overline{X}$ som estimator for μ, og at σ på forhånd er ukjent, slik at vi må bruke våre observasjoner også til å estimere σ. Anta at vi har valgt den forventningsrette

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

som estimator for σ^2. Det er da vanlig å bruke $S = \sqrt{S^2}$ som estimator for σ, og videre bruke

$$S(\overline{X}) = \frac{S}{\sqrt{n}}$$

som estimator for standardavviket $\sigma(\overline{X}) = \sigma/\sqrt{n}$ til $\overline{X}$. Resultatet av estimeringen kan vi da rapportere slik

$$\overline{X} \quad \pm \quad S(\overline{X})$$

[3] I situasjoner der det ikke er behov for å skille mellom (teoretisk) varians og empirisk varians sløyfes gjerne ordet empirisk (Jfr. Kapittel 1.3.)

Sannsynligheten for en estimeringsfeil mindre enn k ganger estimert standardavvik blir

$$P(\mid \overline{X} - \mu \mid < k \cdot S(\overline{X}))$$

Siden det ligger en viss risiko i å erstatte σ med S, må vi vente at denne sannsynligheten, for gitt k, blir noe mindre enn den vi hadde når σ var kjent, nemlig $A(k)$. Imidlertid vil vi, dersom n ikke er for liten, kunne vente at S ligger nær σ, slik at tilnærmelsen $A(k)$ fortsatt kan gi en pekepinn om faren for estimeringsfeil. For ytterligere kommentarer, se Kapittel 8.5.

Eksempel 2 : Strekkstyrke
Vi ønsker å undersøke en bestemt type nylonsnøre mht. strekkstyrken, og har innkjøpt 10 snører av denne typen. La $X_1, X_2, \ldots, X_{10}$ være strekkstyrken i kg for de 10 snørene. Anta at målemodellen er realistisk, og at $\mu = EX_i$ gir uttrykk for gjennomsnittlig strekkstyrke for snører av denne typen, mens $\sigma^2 = varX_i$ gir uttrykk for tilfeldig variasjon fra snøre til snøre. Forutsetningen om uavhengighet kan være tvilsom dersom snørene er valgt ut på en spesiell måte, f.eks. dersom de er produsert rett etter hverandre. Snører av dårlig kvalitet kan da komme samlet, f.eks. ved at maskinene er dårlig justert, eller at råstoffet tilfeldigvis var av dårligere kvalitet enn vanlig. Anta at observasjonene er gitt ved

9.36	9.75	9.23	10.32	10.07	9.68	9.96	9.70	10.15	9.68

Vi får $\overline{X} = 9.79$ og $S = 0.34$, slik at $S(\overline{X}) = S/\sqrt{10} = 0.11$. Vi rapporterer at gjennomsnittlig strekkstyrke er 9.79 ± 0.11.

Et interessant spørsmål i forbindelse med dette eksemplet vil være: Anta at vi plukker ut et nytt snøre og belaster det med 10 kg. Hva er sannsynligheten for at det ryker. La X være strekkstyrken til dette snøret. Vi skal altså finne $P(X < 10.0)$. En beregning basert på estimert forventning 9.79 og estimert standardavvik 0.34 og normaltilnærmelse, vil gi en sannsynlighet på 0.73. De estimerte verdier er åpenbart nokså usikre, og bør helst ikke brukes i “eksaktesannsynlighetsberegninger av denne typen, med mindre de er basert på et langt større erfaringsmateriale (se likevel Kapittel 16.5).

Vi har sett at presisjonen av estimatet vil øke med antall observasjoner. Dersom vi ønsker en viss presisjon må vi sørge for å ta tilstrekkelig mange observasjoner. På den annen side vil for mange observasjoner være sløsing med ressurser (tid, evt. penger). Ønsker vi at estimatoren skal ha standardavvik Δ, er antall observasjoner gitt ved formelen

$$n = \frac{\sigma^2}{\Delta^2}.$$

For å bruke formelen trenger vi å kjenne σ, dvs. usikkerheten i hver enkelt observasjon. Noen ganger er denne kjent fra tidligere observasjoner foretatt med om lag samme usikkerhet. Uten slik informasjon kan man foreta en forundersøkelse basert på et mindre antall observasjoner. Vi vil som regel være fornøyd med å bestemme størrelsesorden av n, og et grovt anslag for σ vil derfor være tilstrekkelig. Under alle omstendigheter vil litt forsøksplanlegging være bedre enn ingen forsøkplanlegging.

La oss så gå over til å studere testing av hypoteser om forventninger i målemodeller.

Eksempel 3 : Testing av forbedring
Anta vi etter omfattende målinger har anslått at den gjennomsnittlige kvikksølvkonsentrasjon hos en fiskeart i en innsjø er 17.2 enheter. Etter dette er det iverksatt restriksjoner på utslippet av avfall, og etter en tid ønsker vi å undersøke om det er noen påviselig reduksjon i kvikksølvkonsentrasjon. Variasjonen i konsentrasjon fiskene i mellom har hittil svart til et standardavvik σ omtrent lik 1.5. Feilen ved måling av den enkelte fisk er neglisjerbar i forhold til dette. Det er bestemt å undersøke $n = 9$ fisk. Problemet kan formuleres som et testproblem i en målemodell.

I generelle termer studerer vi følgende problem for målemodellen: La $X_1, X_2, \ldots, X_n$ være n uavhengige observasjoner fra samme sannsynlighetsfordeling hvor forventningen μ er ukjent, mens variansen σ^2 antas kjent. Vi ønsker å teste om μ er lik en gitt verdi μ_0 mot alternativet at μ er mindre enn verdien μ_0, dvs.

$$H_0 : \mu = \mu_0 \quad \text{mot} \quad H_A : \mu < \mu_0.$$

Ved valg av testmetode vil vi ta utgangspunkt i

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

som vi vet er en brukbar estimator for μ. En rimelig testmetode vil være å forkaste H_0 og påstå H_A når $\overline{X}$ er liten i forhold til μ_0, hvor liten vil avhenge av hvilke risker for feilkonklusjon vi er villig til å løpe. Som testobservator er det hensiktsmessig å bruke

$$Z = \frac{\overline{X} - \mu_0}{\sigma(\overline{X})} = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}}$$

og testmetoden lar vi være:

$$\text{Forkast } H_0 \text{ og påstå } H_A \text{ dersom } Z \leq k$$

der k er en passende valgt konstant. Anta at vi ønsker signifikansnivå α. Vi må da velge k slik at

$$P_{H_0}(Z \leq k) = \alpha$$

For bestemmelse av k må vi derfor kjenne sannsynlighetsfordelingen til Z under forutsetning av at H_0 er riktig. Eksakt bestemmelse er ikke mulig uten ekstra forutsetninger om de stokastiske variable i modellen (mer om dette i Kapittel 8.4). Her må vi nøye oss med tilnærmet beregning: Anta at n er så stor at vi kan forsvare å bruke normaltilnærming (se Kapittel 6.5). Under forutsetning av at H_0 er riktig vil histogrammet til Z kunne tilnærmes med normalkurven (fordi da er Z den korrekte standardisering av $\overline{X}$). Vi får derfor

$$P_{H_0}(Z \leq k) \approx G(k)$$

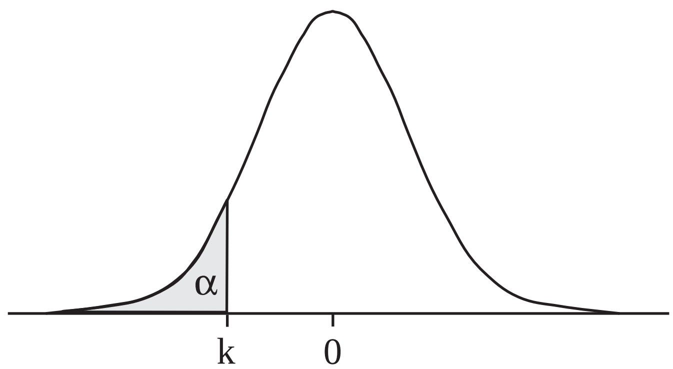
Ved å bestemme k av ligningen $G(k) = \alpha$, får vi derfor en test med tilnærmet signifikansnivå α, se Figur 8.1. Eksempelvis dersom vi krever $\alpha = 0.05$, blir $k = -1.645$, mens dersom vi krever $\alpha = 0.01$, blir $k = -2.326$.
Merk at testmetoden kan alternativt skrives som

$$\text{Forkast } H_0 \text{ og påstå } H_A \text{ dersom } \overline{X} \leq \mu_0 - k_\alpha \cdot \sigma/\sqrt{n}$$

der k_α det positive tall som har areal α til høyre for seg under normalkurven (her er $k = -k_\alpha$). Vi ser at hvor mye mindre $\overline{X}$ må være enn μ_0 før vi våger å forkaste H_0, vil avhenge av antall observasjoner n, usikkerheten i hver enkelt observasjon σ og signifikansnivået α.

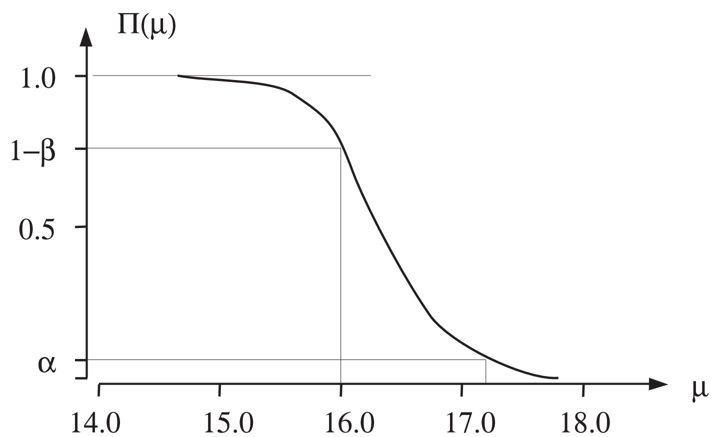
Vi vil også være interessert i styrken til testen, dvs. evnen til å avsløre at $\mu < \mu_0$, dersom μ i virkeligheten er slik. Styrkefunksjonen blir

$$\begin{aligned}
\Pi(\mu) &= P(\text{forkaste } H_0) = P(Z \leq -k_\alpha) \\
&= P(\overline{X} \leq \mu_0 - k_\alpha \cdot \sigma/\sqrt{n}) \\
&= P(\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \leq \frac{\mu_0 - \mu}{\sigma/\sqrt{n}} - k_\alpha) \\
&\approx G(\frac{\mu_0 - \mu}{\sigma/\sqrt{n}} - k_\alpha)
\end{aligned}$$



Figur 8.1: Bestemmelse av kritisk verdi

Vi har her brukt normaltilnærming på sannsynlighetsfordelingen til $(\overline{X}-\mu)/(\sigma/\sqrt{n})$, som er den standardiserte variable til $\overline{X}$ uansett hva μ er. Vi ser at $\Pi(\mu)$ er en avtagende funksjon av μ og at $\alpha = \Pi(\mu_0)$. Generelt vil styrken avhenge av α (via k_α) og av μ, σ og n (via størrelsen $(\mu_0-\mu)/(\sigma\sqrt{n})$).



Figur 8.2: Styrkekurve

Eksempel 3 : Testing av forbedring (fortsatt)
I Eksempel 3 var $\mu_0 = 17.2$ og $\sigma = 1.5$. Med $n = 9$ og $\alpha = 0.05$ blir $k = -1.645$, som svarer til å forkaste H_0 når $\overline{X} \leq 16.38$. Styrkefunksjonen til denne testen blir

$$\Pi(\mu) \approx G(\frac{16.38-\mu}{0.5})$$

La oss tabellere denne for noen μ-verdier:

μ	14.5	15.0	15.5	16.0	16.5	17.0	17.2	17.5
$\Pi(\mu)$	0.9999	0.9971	0.9604	0.7749	0.4032	0.1065	0.0500	0.0130

Styrkekurven er skissert i Figur 8.2.

Eksempelvis ser vi at $\Pi(16.0) = 0.7749$, dvs. at dersom μ er redusert til 16.0, er det en sannsynlighet på $1 - 0.7749 = 0.2251$ for at vi ikke oppdager at det har foregått en reduksjon i det hele tatt (godtakingsfeil). Vi merker oss forøvrig at for $\mu > 17.2$ blir $\Pi(\mu) < \alpha$, slik at α blir en øvre skranke for forkastningsfeil også når nullhypotesen er status quo eller forverring, og vi ønsker å se om det er noen grunn til å påstå forbedring, dvs. at situasjonen er å teste $H_0 : \mu \geq \mu_0$ mot $H_A : \mu < \mu_0$.

I dette eksemplet var antall observasjoner n gitt. En test som oppfyller både et gitt krav til forkastningsfeil α og godtakingsfeil β, krever planlegging av n. En formel basert på normaltilnærming er:

$$n = (\frac{k_\alpha + k_\beta}{\mu_1 - \mu_0})^2 \cdot \sigma^2$$

der μ_1 er det valgte alternativ, der en ønsker en gitt β-risiko. Her er k_α og k_β punkter med de angitte arealer til høyre for seg under normalkurven. For anvendelse av formelen se Oppgave 5.

Noen merknader om testing i målemodellen:

1. I enkelte situasjoner er det aktuelt å teste nullhypotesen $\mu = \mu_0$ (eventuelt $\mu \leq \mu_0$) mot alternativet $\mu > \mu_0$, se Oppgave 6. Da forkaster en nullhypotesen når Z er større enn en kritisk verdi k, som i dette tilfellet bli $k = k_\alpha$. Formlen for planlegging av antall observasjoner er den samme. Vi overlater til leseren å utarbeide detaljene, se Oppgave 9.

2. I noen situasjoner kan det være aktuelt med såkalt to-sidig test, dvs. at alternativet til nullhypotesen $\mu = \mu_0$ er $\mu \neq \mu_0$, se Oppgave 8. Da forkaster en nullhypotesen når $\mid Z \mid$ er større enn en kritisk verdi k, dvs. både når Z er liten og når Z er stor. I dette tilfellet blir $k = k_{\alpha/2}$. Samme planleggingsformel kan (tilnærmet) brukes for n, men erstatt α med $\alpha/2$. Leseren kan selv utarbeide detaljene, se Oppgave 9.

3. Vi har i målemodellen hittil bare studert tester for forventning μ for gitt varians σ^2. Man kan også formulere hypotesetestingsproblemer for variansen i målemodellen. Et praktisk eksempel vil være en produksjonsprosess

som, når den er i kontroll, gir artikler av en gitt forventet kvalitet μ_0, mens variansen σ^2 er gitt lik σ_0^2. Kan hende løper prosessen ut av kontroll ved at variansen har økt til et nivå større enn σ_0^2, dvs. prosessen gir mer varierende kvalitet enn før. Situasjonen kan studeres som et hypotesetestingsproblem om σ^2 hvor nullhypotesen er $\sigma^2 = \sigma_0^2$ og alternativet er $\sigma^2 > \sigma_0^2$. Vi har ikke etablert nok teori til å behandle dette problem.

4. Vi har sett på testing av forventning i målemodellen når variansen σ^2 var kjent. I de fleste praktiske situasjoner vil imidlertid σ^2 være ukjent. Vi har ikke etablert nok teori til å gjennomføre en grundig diskusjon av denne problemstillingen, men siden den er så viktig, vil vi gjengi de viktigste resultatene i Kapittel 8.4.

5. Målemodellen kan legges til grunn for ulike betraktninger i forbindelse med prosesstyring. I denne sammenheng er en både interessert i endringer i forventning (nivå) og varians (spredning). Det praktiske verktøyet er *kontrolldiagrammer*, der observasjonene plottes i naturlig rekkefølge, og grenser for den naturlige variasjon tegnes inn, se Kapittel 15. Regler for "å gripe inn"ved avvikende mønstre kan da etableres basert på sannsynlighetsbetraktninger, som gir visse garantier mot feilaktig å gripe inn osv.

$\star$ Estimatoren $\hat{\mu} = \overline{X}$ har en interessant egenskap: Den er minste kvadraters estimatoren for forventningen μ i målemodellen, som innebærer at den minimerer kvadratsummen

$$Q(\mu) = \sum_{i=1}^{n} (X_i - \mu)^2$$

Dette er en direkte følge av likheten (se Oppgave 27 (a))

$$\sum_{i=1}^{n} (X_i - \mu)^2 = \sum_{i=1}^{n} (X_i - \overline{X})^2 + n(\overline{X} - \mu)^2.$$

som viser at

$$Q(\mu) \geq \sum_{i=1}^{n} (X_i - \overline{X})^2.$$

med likhet hvis og bare hvis μ velges lik $\hat{\mu} = \overline{X}$.

Dette blir ofte sett på som en gunstig egenskap ved $\overline{X}$, den føyer tallmaterialet på en bra måte, i den forstand at den gjennomsnittlige kvadrerte avstand fra observasjonene til estimatet for μ blir lite. Minimeringen av kvadratsummen ovenfor er en anvendelse av et generelt prinsipp for utledning av estimatorer kalt *minste kvadraters metode*, som spiller en sentral rolle i statistisk teori. Imidlertid bør det advares mot å legge for stor vekt på at en estimator er minste kvadraters estimator. Det fins andre generelle prinsipper for utledning av estimatorer, og en estimator bør alltid vurderes i lys av den modell vi er villig til å anta. For ytterligere kommentarer se Kapittel 8.4.

Formelen ovenfor kan også brukes ved beregning av forventningen til den empiriske varians. Vi har

$$S_X^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2 - (\overline{X} - \mu)^2.$$

Dette gir

$$E(S_X^2) = E(\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2) - E(\overline{X} - \mu)^2 = \sigma^2 - \sigma^2/n = (n-1)/n \cdot \sigma^2.$$

8.2 Lotterimodellen.

I praksis støter en ofte på situasjoner hvor vi ønsker å trekke slutninger om en gitt endelig populasjon på grunnlag av et utvalg av elementer fra populasjonen. La oss nevne noen eksempler:

- Vi kan finne ut noe om ungdommens røykevaner i en gitt aldersgruppe i Bergen ved å spørre et utvalg av ungdommer.
- Vi kan finne ut noe om årsinntekten til et kull siviløkonomer 5 år etter eksamen ved å ta for oss et utvalg av slike.
- Vi kan anslå volumet av trevirket i en skog ved å måle et utvalg av trærne i skogen.

En slik framgangsmåte kalles ofte for en *utvalgsundersøkelse*. Grunnen til at man foretar en utvalgsundersøkelse framfor å innhente opplysninger om alle elementene i populasjonen, kan være at det siste er tidkrevende og/eller kostbart. Det finnes derfor en god del teoretisk litteratur om ulike utvalgsmodeller og utvalgsmetoder. Slike utvalgsmodeller danner et felles teoretisk grunnlag for en rekke ulike anvendelser, eksempelvis intervjuundersøkelser, meningsmålinger, markedsundersøkelser, kvalitetskontroll, stikkprøverevisjon etc. Vi skal her se på en enkel (men nyttig) utvalgsmodell som vi i det følgende vil kalle *lotterimodellen*. Først litt notasjon:

Gitt en populasjon bestående av N elementer. Til hvert element er knyttet en verdi v. Vi tenker oss elementene nummerert fra 1 til N, og lar v_i betegne v-verdien til element nr.i. Fra populasjonen trekkes et ordnet utvalg på n elementer (uten tilbakelegging). La Y_j betegne v-verdien til det j'te uttrukne element i utvalget ($j = 1, 2, \ldots, n$).

I de fleste anvendelser spiller rekkefølgen av elementene i populasjon og utvalg ingen rolle, og nummereringen er bare av praktisk natur.

Lotterimodellen: Når verdiene $Y_1, Y_2, \ldots, Y_n$ er bestemt som et *tilfeldig ordnet utvalg* fra $\{v_1, v_2, \ldots, v_N\}$. sier vi at observasjonene oppfyller betingelsen i lotterimodellen.

På grunnlag av v-verdien til de n elementene utvalget $Y_1, Y_2, \ldots, Y_n$, ønsker vi å slutte noe om v-verdiene til de N elementene i populasjonen $v_1, v_2, \ldots, v_N$. Spesielt vil vi være interessert i å estimere den gjennomsnittlige v-verdi i populasjonen

$$\overline{v} = \frac{1}{N}\sum_{i=1}^{N} v_i.$$

Som estimator er det rimelig å velge det tilsvarende gjennomsnitt i utvalget:

$$\overline{Y} = \frac{1}{n}\sum_{j=1}^{n} Y_j.$$

Vi har da følgende resultat

Dersom $Y_1, Y_2, \ldots, Y_n$ oppfyller betingelsen i lotterimodellen, gjelder

$$E(\overline{Y}) = \overline{v} \qquad var(\overline{Y}) = \frac{N-n}{N-1}\cdot\frac{\sigma^2}{n}$$

der σ^2 er *variansen i populasjonen* definert ved

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(v_i - \overline{v})^2$$

Vi ser at $\overline{Y}$ er en forventningsrett estimator for $\overline{v}$, med varians som avhenger av spredningen av v-verdier i populasjonen målt med σ^2, samt av populasjonsstørrelsen N og utvalgsstørrelsen n.

Begrunnelse: Dersom trekningen isteden hadde foregått med tilbakelegging, vil Y_j'ene ha samme fordeling og være uavhengig av hverandre, dvs. oppfylle kravene i målemodellen. Regnereglene for forventning og varians gir da

$$E(\overline{Y}) = \overline{v} \qquad var(\overline{Y}) = \frac{\sigma^2}{n}$$

Tar en omsyn til avhengigheten i regneformlene, får vi resultatet ovenfor. Begrunnelsen finnes nedenfor som $\star$-merket stoff. Faktoren $(N-n)/(N-1)$ kan oppfattes som en korreksjonsfaktor som må være med når trekningen skjer uten tilbakelegging [4]. Vi ser at variansen avtar for økende utvalgsstørrelse n. Når $n = N$ blir variansen lik null. Dette er rimelig fordi nå er hele populasjonen med i utvalget og $\overline{Y} = \overline{v}$, dvs. all usikkerhet om $\overline{v}$ er nå fjernet. Som før vil vi rapportere

$$\text{estimat} \pm \text{standardavvik}.$$

Igjen ønsker vi å relatere standardavviket til risikoen for estimeringsfeil, f.eks. gi sannsynlighetsutsagn. Nå kan $\overline{Y}$ riktignok skrives som en sum av stokastiske variable med samme sannsynlighetsfordeling, men på grunn av avhengigheten, er ikke alle forutsetningene for bruk av normaltilnærmelse oppfylt. Vi bør derfor utvise en viss varsomhet når vi bruker standardavviket til å vurdere risikoen for estimeringsfeil. Dersom utvalgsstørrelsen n er stor nok, men likevel liten i forhold til N, blir avhengigheten liten og normaltilnærmelsen kan gi brukbart resultat. 68%−95% regelen kan da gi en viss pekepinn.

Eksempel 4 : Lakseoppdrett
Et forsøk med oppdrett av laks består av $N = 50$ laks som er klekket på samme tidspunkt. Vi er interessert i gjennomsnittsvekten av laksene etter en viss tid. Det er tidkrevende og brysomt å fange opp alle laksene og veie dem, så vi nøyer oss med å plukke ut tilfeldig $n = 10$ laks for veiing. Vi lar $v_1, v_2, \ldots, v_{50}$ betegne (de ukjente) vektene til de 50 laksene i dammen, og lar $Y_1, Y_2, \ldots, Y_{10}$ betegne vektene av de uttrukne laksene. Vi velger å bruke $\overline{Y}$ som estimator for $\overline{v}$, og vet at $\overline{Y}$ er forventningsrett med varians $\frac{4}{49}\sigma^2$, slik at standardavviket blir $\sigma(\overline{Y}) = \frac{2}{7}\sigma$. Her er σ ukjent, men fra et tidligere forsøk vet vi at σ er om lag 0.50. Standardavviket til estimatoren anslås derfor på forhånd til 1/7=0.143. Anta at de 10 veiingene ga følgende resultat i kg.

1.39	2.41	2.45	2.52	2.01	2.91	2.10	2.05	2.79	2.27

Her blir $\overline{Y}$=2.290 og vi rapporterer derfor at gjennomsnittsvekten av de $N = 50$ laksene på dette tidspunktet er

$$2.290 \;\; \pm \;\; 0.143$$

[4]Enkelte lærebøker oppgir faktoren til $1 - n/N$, som bare er tilnærmet riktig for stor N, mens andre sløyfer faktoren helt.

Dersom vi ikke hadde noen forhåndsinformasjon om σ^2, er vi henvist til å anslå σ^2 ut fra de foreliggende observasjonene.

La oss generelt se på problemet med estimering av σ^2: Det synes rimelig å bruke

$$S_Y^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \overline{Y})^2 \quad \text{som estimator for} \quad \sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(v_i - \overline{v})^2$$

dvs. som estimat for variansen i populasjonen bruker vi den tilsvarende empiriske varians i observasjonene. Det kan vises at (se Oppgave 34)

$$E(S_Y^2) = \frac{N}{N-1}\cdot\frac{n-1}{n}\sigma^2 \approx \frac{n-1}{n}\sigma^2 \quad \text{for store } N.$$

Estimatoren S_Y^2 er derfor ikke forventningsrett, men kan om ønskelig korrigeres. Siden N vanligvis er stor, ellers hadde vi neppe foretatt noen utvalgsundersøkelse, vil det som regel være nok å dividere kvadratsummen med $n-1$ istedenfor n, som før kalt S^2.[5]

I eksemplet ovenfor gir tallmaterialet S^2=0.190 som estimat for σ^2, som gir S=0.436 som estimat for σ, dvs. ikke langt fra det vi hadde stipulert.

La oss se på muligheten for å drive forsøksplanlegging. Anta at vi ønsker at estimatoren $\overline{Y}$ skal få et gitt standardavvik Δ. Ut fra formelen

$$\frac{N-n}{N-1}\cdot\frac{\sigma^2}{n} = \Delta^2$$

kan vi for gitt Δ, σ og N bestemme den utvalgsstørrelsen n som er nødvendig for å oppnå dette. Løser vi dette mhp. n, får vi

$$n = \frac{\sigma^2}{\frac{\sigma^2}{N} + \Delta^2\cdot\frac{N-1}{N}} \approx \frac{\sigma^2}{\Delta^2}$$

Tilnærmingen fås ved å la $N \to \infty$ i den eksakte formel, og gjelder derfor bare for stor N. Beregningen med den tilnærmede formel er enklere, og gir en øvre skranke n', svarende til uendelig populasjon, for det nødvendige antall observasjoner n. Det er lett å sjekke at n kan (tilnærmet) uttrykkes som

$$n = n'\cdot\frac{1}{1+\frac{n'}{N}}$$

[5] Sammenligner vi med det analoge resultat for målemodellen i Kapittel 8.1 ser vi at $N/(N-1)$ er en korreksjonsfaktor som korrigerer for at utvalg skjer uten tilbakelegging.

Vi kan tolke brøken som en korreksjonsfaktor, som reduserer det nødvendige antall observasjoner, som følge av at populasjonen er endelig, sml. Eksempel 7.10 for den hypergeometriske situasjon.

Dersom vi i Eksempel 4 ovenfor ønsker Δ lik 0.10 må vi, dersom vi stipulerer at σ er ca. lik 0.50, øke utvalgsstørrelsen til $n = 17$ (sjekk utregningen!).

Noen sluttkommentarer om lotterimodellen kontra målemodellen:
Lotterimodellen dreier seg om trekning av et utvalg fra en gitt avgrenset populasjon, mens målemodellen dreier seg om uavhengige gjentatte observasjoner av samme størrelse (i vid forstand). La oss ta et eksempel i grenseland: Anta at vi vil bestemme konsentrasjonen av kvikksølv hos en fiskeart i Mjøsa. Vi har en avgrenset populasjon av slike fisk, men antallet N er her ukjent. Vi er derfor avskåret fra å bruke lotterimodellen i studiet av gjennomsnittsinnholdet av kvikksølv hos fiskene i Mjøsa. Vi kan isteden bruke målemodellen, kvikksølvinnholdet hos slike fisk varierer tilfeldig omkring et (gjennomsnittlig") nivå μ, og hver målt fisk representerer en måling av dette nivå. En målemodell er enklere å arbeide med enn lotterimodellen, og noen ganger er man fristet til å bruke en målemodell selv om lotterimodellen er mer realistisk: I lotterisituasjoner der utvalget n er lite i forhold til N, er det liten forskjell på om vi trekker med eller uten tilbakelegging. Målemodellen kan da oppfattes som en tilnærmelse til lotterimodellen. Selv om vi kjente antallet N av fiskearten i Mjøsa ville vi, siden N er svært stor, antakelig velge målemodellen framfor den tilsvarende lotterimodell. Vi ser for øvrig at konsekvensen ved å estimere på grunnlag av en målemodell i en lotterisituasjon, vil være at vi oppgir et noe for høyt standardavvik.

$\star$ En begrunnelse for forventningen og variansen til gjennomsnittet i utvalget i lotterimodellen er ikke vanskelig ut fra den teori vi allerede har: Ved hver trekning er sannsynligheten $1/N$ for å trekke element nr. i. Vi har derfor

v	v_1	v_2	$\cdots$	v_N
$P(Y_j = v)$	$\frac{1}{N}$	$\frac{1}{N}$	$\cdots$	$\frac{1}{N}$

Følgelig har $Y_1, Y_2, \ldots, Y_n$ alle samme sannsynlighetsfordeling. For alle $j = 1, \ldots, n$ får vi derfor

$$EY_j = \sum_{i=1}^{N} v_i \cdot \frac{1}{N} = \frac{1}{N} \sum_{i=1}^{N} v_i = \overline{v}$$

$$varY_j = \sum_{i=1}^{N} (v_i - \overline{v})^2 \frac{1}{N} = \frac{1}{N} \sum_{i=1}^{N} (v_i - \overline{v})^2 = \sigma^2.$$

Dette gjelder uansett om trekningen foregår med eller uten tilbakelegging, det siste pga. ekvivalensloven for ordnede utvalg (se Kapittel 3.5). Ved trekning uten tilbakelegging er imidlertid $Y_1, Y_2, \ldots, Y_n$ avhengige.

Av regnereglene for forventning følger at

$$E(\sum_{j=1}^{n} Y_j) = \sum_{j=1}^{n} EY_j = \sum_{j=1}^{n} \overline{v} = n \cdot \overline{v}.$$

Følgelig blir $E\overline{Y} = \overline{v}$, som altså gjelder både ved trekning med og uten tilbakelegging. Vi minner så om følgende generelle formel (se Kapittel 5.6)

$$var(\sum_{j=1}^{n} Y_j) = \sum_{j=1}^{n} varY_j + \sum_{j \neq k} cov(Y_j, Y_k).$$

Dersom trekningen foregår med tilbakelegging, slik at Y_j'ene er uavhengige, faller alle kovariansledd bort, og vi får variansformlen uten korreksjonsledd.

Nå har vi imidlertid $n(n-1)$ kovariansledd som må være like store, kall hver kovarians for c. Vi kan derfor skrive

$$var(\sum_{j=1}^{n} Y_j) = n \cdot \sigma^2 + n(n-1)c.$$

Denne formelen gjelder for $n = 1, 2, \ldots, N$. I tilfellet $n = N$ er all usikkerhet fjernet, fordi vi da vet at $\sum Y_j$ er lik konstanten $\sum v_i$. Variansen til summen er i dette tilfelle lik null, og vi har derfor

$$0 = N \cdot \sigma^2 + N(N-1)c$$

Løser vi denne identitet m.h.p. c får vi at

$$c = -\frac{1}{N-1} \cdot \sigma^2.$$

Innsettes dette i formelen for $var(\sum Y_j)$, får vi

$$var(\sum_{j=1}^{n} Y_j) = n \cdot \sigma^2 + n(n-1)(-\frac{1}{N-1}\sigma^2) = \frac{N-n}{N-1} \cdot n \cdot \sigma^2.$$

Av dette følger at

$$var(\overline{Y}) = var(\frac{1}{n}\sum_{j=1}^{n} Y_j) = (\frac{1}{n})^2 var(\sum_{j=1}^{n} Y_j) = \frac{N-n}{N-1} \cdot \frac{\sigma^2}{n}$$

som viser den annonserte variansformel.

8.3 En enkel regresjonsmodell

I praksis finner vi ofte iakttakelser som faller innenfor rammen av følgende tankegang: Med en gitt *stimulus* t observeres en *respons* X. Vi vil oppfatte X som en stokastisk variabel, mens t oppfattes som en variabel som ikke er underlagt tilfeldigheter. Eksempler på dette kan være

- Avlingskvantum X for et gitt gjødningskvantum t.
- Surhetsgraden X i et vann for et gitt røkutslipp t.
- Solgt kvantum X av en vare for gitt reklameinnsats t.

I en del slike situasjoner er det rimelig å anta at forventningen til X er en lineær funksjon av t slik at vi kan skrive

$$EX = \gamma + \beta t.$$

Her er γ og β størrelser som nærmere karakteriserer sammenhengen mellom stimulus og respons. Vi ser at γ er forventet respons dersom stimulus $t = 0$. Videre vil $\beta = 0$ svare til at responsen ikke avhenger av stimulus, mens jo større β er, desto større er responsen for en gitt stimulus $t > 0$. Det er vanlig å kalle linjen $x = \gamma + \beta t$ for *regresjonslinjen for* X m.h.p. t og koeffisienten β kalles *regresjonskoeffisienten for* X m.h.p. t. I praksis vil som regel både γ og β være ukjente. For å få informasjon om γ og β kan vi gjenta eksperimentet, la oss si n ganger, hvor vi observerer responsen for en rekke ulike verdier av stimulus. Vi observerer

Stimulus:	t_1,	t_2,	$\ldots$,	t_n
Respons:	X_1,	X_2,	$\ldots$,	X_n

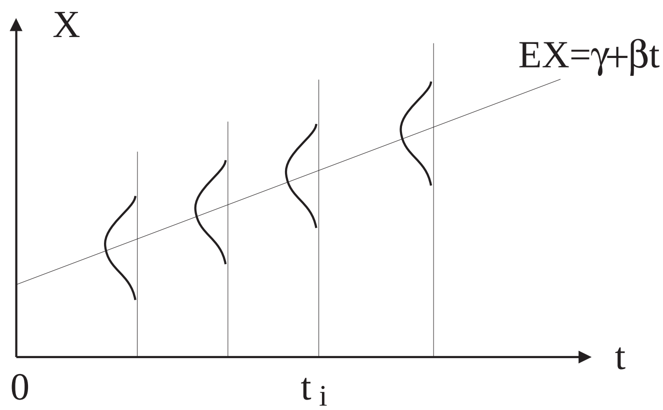
dvs. X_i er observert respons med stimulus lik t_i. Vi har altså antatt at $EX_i = \gamma + \beta t_i$ for $i = 1, 2, \ldots, n$. Det er ofte rimelig å anta at usikkerheten ved hver måling av responsen er den samme. Vi kan uttrykke dette i en modell ved å anta at de n stokastiske variable $X_1, X_2, \ldots, X_n$ alle har den samme varians, kall den σ^2. Videre er det ofte rimelig å anta at responsen ikke avhenger av hverandre. Vi uttrykker dette i en modell ved å anta at $X_1, X_2, \ldots, X_n$ er uavhengige stokastiske variable. En oppsummering av de antakelsene som er gjort, gir oss

Regresjonsmodellen:
$(t_1, X_1), (t_2, X_2), \ldots, (t_n, X_n)$ er n observasjonspar, hvor $X_1, X_2, \ldots, X_n$ er uavhengige stokastiske variable, slik at

$$\begin{aligned} EX_i &= \gamma + \beta t_i \quad \text{for} \quad i = 1, 2, \ldots, n \\ varX_i &= \sigma^2 \quad \text{for} \quad i = 1, 2, \ldots, n \end{aligned}$$

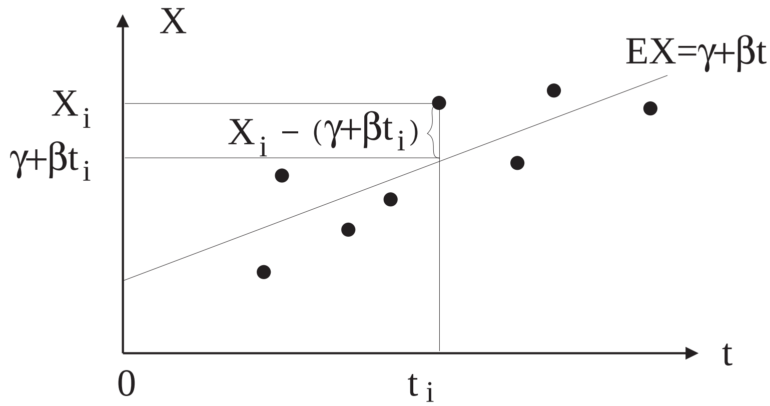
Vi kan alternativt si at $X_i = \gamma + \beta t_i + U_i$, der U_i'ene er uavhengige avviksledd med forventning 0 og samme varians σ^2.

I denne modellen er γ og β ukjente parametre som vi ønsker å estimere på basis av de n observasjonsparene $(t_1, X_1), (t_2, X_2), \ldots, (t_n, X_n)$. Med andre ord ønsker vi å estimere regresjonslinjen. Som oftest vil også σ^2 være ukjent, og da vil vi også være interessert i å estimere denne parameteren.



Figur 8.3: Tilfeldige avvik fra regresjonslinje

Regresjonsmodellen er illustrert i Figur 8.3, der vi har tegnet inn regresjonslinjen $x = \gamma + \beta t$ dvs. forventningen til X for ulike t. Vi ser at regresjonskoeffisienten β er vinkelkoeffisienten til denne linjen, mens γ er det punkt der linjen skjærer den vertikale aksen. Parameteren σ i modellen uttrykker den naturlige tilfeldighet i X omkring sin forventning, dvs. omkring regresjonslinjen. I Figur 8.4 er tegnet inn et observasjonsmateriale, dvs. sammenhørende verdier av (t_i, X_i), samt illustrert et observert avvik fra regresjonslinjen $U_i = X_i - (\gamma + \beta t_i)$. I en konkret praktisk situasjon kan vi plotte observasjonsmaterialet på denne måten og den figur som framkommer kalles *spredningsdiagram*. Imidlertid er regresjonslinjen ikke kjent, og problemet består i å legge inn en linje som best mulig føyer observasjonsmaterialet i en viss forstand. Dette kan selvfølgelig gjøres på øyemål, men det viser seg i praksis at andre metoder er bedre.



Figur 8.4: Regresjonslinje med observasjoner

Det er praktisk å innføre $\alpha = \gamma + \beta\bar{t}$ som parameter istedenfor γ, der $\bar{t}$ betegner gjennomsnittet av de n t-verdiene. Vi ser at α kan tolkes som forventet respons når stimulus settes på gjennomsnittsverdien. Vi har altså

$$EX_i = \alpha + \beta(t_i - \bar{t}) \quad \text{for } i = 1, 2, \ldots, n.$$

Vi ønsker nå å finne brukbare estimatorer for α og β. Mye brukt er de såkalte *minste kvadraters estimatorene*, som er de estimatorene $\hat{\alpha}$ og $\hat{\beta}$ som minimerer uttrykket

$$Q(\alpha, \beta) = \sum_{i=1}^{n} (X_i - \alpha - \beta(t_i - \bar{t}))^2 \quad \text{m.h.p. } \alpha \text{ og } \beta.$$

Dette innebærer at vi bestemmer den linje som føyer tallmaterialet best, i den forstand at den samlede kvadrerte vertikale avstand fra observasjonsparene til linjen er minst mulig. Det kan vises at minste kvadraters estimatorene er gitt ved (se $\star$- merket stoff til slutt i avsnittet)

$$\begin{aligned} \hat{\alpha} &= \overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \\ \hat{\beta} &= \frac{\sum_{i=1}^{n}(X_i - \overline{X})(t_i - \bar{t})}{\sum_{i=1}^{n}(t_i - \bar{t})^2}. \end{aligned}$$

Vi skriver $M = \sum_{i=1}^{n}(t_i - \bar{t})^2$, og merker oss at $\sum_{i=1}^{n}(t_i - \bar{t}) = 0$. Estimatoren $\hat{\beta}$ kan derfor uttrykkes som

$$\hat{\beta} = \frac{1}{M}\sum_{i=1}^{n}(t_i - \bar{t})X_i.$$

Vi ser at både $\hat{\alpha}$ og $\hat{\beta}$ er lineære funksjoner av observasjonene $X_1, X_2, \ldots, X_n$. Estimatorene har følgende egenskaper:

$$\boxed{\begin{array}{ll} E(\hat{\alpha}) = \alpha & \qquad var(\hat{\alpha}) = \dfrac{\sigma^2}{n}. \\ E(\hat{\beta}) = \beta & \qquad var(\hat{\beta}) = \dfrac{\sigma^2}{M}. \end{array}}$$

Som i målemodellen følger dette direkte av regnereglene for forventning og varians, og anbefales som øvelse; fasit kommer nedenfor som $\star$-merket stoff.

Vi ser at både $\hat{\alpha}$ og $\hat{\beta}$ er forventningsrette estimatorer. Videre ser vi at estimatorene kan gjøres så presise man måtte ønske ved å øke n og M. For å øke M må vi unngå at t_i-verdiene klumper seg sammen om samme verdi, de bør velges spredt innenfor det område vi tar sikte på å bruke modellen.

Minste kvadraters regresjonslinjen vil vi skrive

$$\hat{X} = \hat{\alpha} + \hat{\beta} \cdot (t - \bar{t})$$

eller alternativt

$$\hat{X} = \hat{\gamma} + \hat{\beta} \cdot t$$

der $\hat{\gamma} = \hat{\alpha} - \hat{\beta}\bar{t}$ kan oppfattes som en estimator for parameteren γ. $\hat{X}$ kan oppfattes som en estimator for den sanne regresjonslinjen idet

$$E\hat{X} = \alpha + \beta \cdot (t - \bar{t}) = \gamma + \beta \cdot t.$$

$\hat{X}$ kan i praksis også tenkes brukt til å predikere responsen X for gitt stimulus t, og blir i slik sammenheng kalt en *prediktor*.

Eksempel 5 : Vekstforsøk

I et vekstforsøk har man observert en del sammenhørende verdier av kvantum (t) av et bestemt gjødningsstoff og avlingens størrelse (X). Anta at resultatet ble

Gjødning (gr./m^2):	10	20	30	40	50	60	70
Avling (kg/m^2):	1.0	1.3	1.5	1.7	2.0	2.0	2.2

I en modell for dette eksperimentet vil det være naturlig å betrakte de observerte sju avlingskvanta som realisasjoner av sju stokastiske variable $X_1, X_2, \ldots, X_7$, idet avkastningene rimeligvis ikke er entydig bestemt ved de tilhørende gjødningskvanta $t_1, t_2, \ldots, t_7$, men vil også avhenge av utenforliggende forhold, såsom varierende jordsmonn, nedbør i vekstperioden og sol i vekstperioden. Erfaring tyder imidlertid på at forventet avling som en lineær funksjon av gjødningskvantumet kan være en brukbar antakelse, i hvert fall innenfor det kvantumsområde det her er tale om. Dette kan bekreftes ved å tegne et spredningsdiagram. En antakelse om uavhengighet og samme varians virker også rimelig. Tallmaterialet gir

$$\bar{t} = 40 \quad \overline{X} = 1.67 \quad \hat{\beta} = 0.01964.$$

Minste kvadraters regresjonslinje er derfor gitt ved

$$\begin{aligned} \hat{X} &= 1.67 + 0.01964 \cdot (t - 40) \\ &= 0.886 + 0.01964 \cdot t \end{aligned}$$

Den funne regresjonslinje kan brukes som en prediktor. Eksempelvis vil $t = 45$ gi prediksjonen $0.886 + 0.01964 \cdot 45 = 1.77$, mens $t = 80$ gir prediksjonen $0.886 + 0.01964 \cdot 80 = 2.46$. Dette forutsettes selvsagt at prediksjonen gjelder for forhold som svarer til den tidligere observerte situasjon. Eksempelvis vil det med det observerte grunnlag være urimelig å forsøke å predikere avkastningen for et gjødningskvantum på $t = 180$, idet en ikke kan vente at linearitetsantakelsen holder over et så stort område.

I dette eksemplet var de tilfeldige avvik fra regresjonslinjen små (tyder på liten σ), slik at estimater og prediksjoner tydeligvis blir ganske pålitelige, selv om antall observasjoner er lite. I andre situasjoner kan observasjonene vise stor spredning (stor σ), og da kan estimater og prediksjoner inneholde et betydelig element av usikkerhet (se f.eks. Eksempel 1.9). Vi må derfor være i stand til å gi vurderinger av påliteligheten. Sentralt i slike vurderinger er standardavviket til observasjonene σ. I noen situasjoner er dette kjent fra tidligere erfaring, men som regel er σ ukjent, og spørsmålet er derfor om det lar seg gjøre å estimere σ ut fra de foreliggende observasjoner. Siden $\sigma^2 = var X_i = E(X_i - (\alpha + \beta(t_i - \bar{t})))^2$, ser vi at

$$\frac{1}{n} \sum_{i=1}^{n} (X_i - (\alpha + \beta(t_i - \bar{t})))^2$$

har forventning σ^2. Denne størrelsen kunne brukes som estimator for σ^2 hadde det ikke vært for at α og β er ukjente. Men vi har jo estimert α og β med henholdsvis $\hat{\alpha}$ og $\hat{\beta}$. Betrakt derfor

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - (\hat{\alpha} + \hat{\beta}(t_i - \bar{t})))^2.$$

S^2 er et gjennomsnittsmål for observasjonenes vertikale avstand fra den estimerte linjen, og er på mange måter en naturlig estimator for σ^2. Det kan vises at (se $\star$-merket stoff)

$$ES^2 = \frac{n-2}{n}\sigma^2.$$

Estimatoren er ikke forventningsrett, men vi ser at, dersom vi i definisjonen av S^2 dividerer med $n-2$ istedenfor n, får vi en forventningsrett estimator for σ^2 (for store n gjør det liten forskjell). Det er nå naturlig å bruke $S = \sqrt{S^2}$ som estimator for standardavviket σ.

Dersom vi ønsker å angi usikkerheten ved å bruke $\hat{X}$ som prediktor for X, tar vi utgangspunkt i prediksjonsfeilen $\hat{U} = \hat{X} - X$. Under forutsetningene i vår regresjonsmodell blir $E\hat{U} = 0$, dvs. forventet prediksjonsfeil lik null, mens (se Oppgave 35)

$$\sigma(\hat{U}) = \sigma\sqrt{1 + \frac{1}{n} + \frac{(t-\bar{t})^2}{M}}$$

Uttrykket avhenger av σ såvel som n, M og t, og kan forstås slik: Annet og tredje ledd i summen skyldes usikkerhet ved estimeringen av hhv. konstantleddet og regresjonskoeffisienten, mens første ledd skyldes prediksjonsusikkerhet som er der selv om all estimeringsusikkerhet er fjernet. Dersom n og M er store, ser vi at standardavviket til prediksjonsfeilen er tilnærmet lik σ. Anslaget S for σ vil kunne danne grunnlag for risikovurderinger mht. prediksjonsfeilen. 68% - 95% regelen vil kunne gi en viss pekepinn, men kan være litt for optimistisk.

I Eksempel 5 finner vi at $S = 0.0824$. Dersom vi sier at vi kan predikere avlingen med feilmargin på $2 \cdot 0.08 = 0.16$ med 95% sikkerhet, er dette litt optimistisk. Med bare 7 observasjoner vil S være et relativt upålitelig estimat for σ, rotuttrykket ikke være neglisjerbart og tilnærmingen til normalkurven vil være mindre god. Vi skal se senere at en 95% garanti krever en feilmargin på ca 0.21 for tilfellet $t = 40$, økende til 0.24 for $t = 70$.

Dersom vi ønsker å angi usikkerheten ved å bruke $\hat{\beta}$ som estimator for regresjonskoeffisienten β, kan vi i tillegg rapportere standardavviket til $\hat{\beta}$, eller eventuelt et estimat for dette. Dersom σ er kjent på forhånd rapporterer vi

$$\hat{\beta} \;\; \pm \;\; \sigma(\hat{\beta}) \quad \text{der} \quad \sigma(\hat{\beta}) = \frac{\sigma}{\sqrt{M}}$$

Dette kan tolkes i lys av $68\% - 95\%$ regelen, idet histogrammet til den standardiserte variable til $\hat{\beta}$, under rimelige tilleggsforutsetninger, vil kunne tilnærmes med normalkurven. [6] Dersom σ ikke er kjent, men estimert med S, rapporterer vi isteden

$$\hat{\beta} \;\; \pm \;\; S(\hat{\beta}) \quad \text{der} \quad S(\hat{\beta}) = \frac{S}{\sqrt{M}}$$

Også her vil $68\% - 95\%$ regelen kunne gi en viss pekepinn, men dette kan være litt for optimistisk, spesielt dersom vi har få observasjoner. I Eksempel 5 finner vi $M = 2800$, slik at $S(\hat{\beta}) = 0.0824/\sqrt{2800} = 0.00156$, og da kan vi rapportere regresjonskoeffisienten som 0.01964 ± 0.00156. Det betyr at tredje desimal er usikker, og det synes unødvendig å dra med seg hele fem desimaler.

I mange sammenhenger ønsker en teste hypoteser om regresjonskoeffisienten i en regresjonsmodell. I første omgang antar vi σ er kjent. La oss se på den situasjon at vi ønsker å teste om β har en bestemt verdi β_0 mot alternativet at β er mindre enn β_0, dvs.

$$H_0 : \beta = \beta_0 \;\; \text{mot} \;\; H_A : \beta < \beta_0.$$

Ved valg av testmetode vil vi ta utgangspunkt i estimatoren $\hat{\beta}$. En rimelig testmetode vil være å forkaste H_0 og påstå H_A når $\hat{\beta}$ er liten i forhold til β_0. En hensiktsmessig testobservator er

$$Z = \frac{\hat{\beta} - \beta_0}{\sigma(\hat{\beta})} = \frac{\hat{\beta} - \beta_0}{\sigma/\sqrt{M}}$$

og testmetoden lar vi være:

$$\text{Forkast } H_0 \text{ og påstå } H_A \text{ dersom } Z \leq k.$$

Anta at vi ønsker en test med signifikansnivå α. Vi må da velge k slik at

$$P_{H_0}(Z \leq k) = \alpha.$$

[6] Dette gjelder når n og M er stor (følger av en generalisering av sentralgrensesetningen) og/eller når observasjonene selv er normalfordelt (se avsnitt 8.4).

For eksakt bestemmelse av k må vi kjenne sannsynlighetsfordelingen til Z under forutsetning av at H_0 er riktig. Eksakt bestemmelse er ikke mulig uten å gjøre flere antakelser i modellen (se Kapittel 8.4). Vi må her nøye oss med normaltilnærming. Histogrammet til den standardiserte variable til $\hat{\beta}$ kan tilnærmes til normalkurven når n er stor og samtidig M er stor. Bruker vi normaltilnærming får vi

$$P_{H_0}(Z \leq k) \approx G(k)$$

Ved å bestemme k av ligningen $G(k) = \alpha$, dvs. punktet som har areal α under normalkurven til venstre for seg, får vi en test med tilnærmet signifikansnivå α. Eksempelvis dersom vi krever $\alpha = 0.10$ blir $k = -1.28$. Testmetoden kan alternativt skrives (med notasjon og argumentasjon som i Eksempel 3):

$$\text{Forkast } H_0 \text{ og påstå } H_A \text{ dersom } \hat{\beta} \leq \beta_0 - k_\alpha \cdot \sigma/\sqrt{M}.$$

Styrkefunksjonen for testen blir

$$\begin{aligned}
\Pi(\beta) &= P(\text{forkaste } H_0) = P(Z \leq -k_\alpha) \\
&= P(\hat{\beta} \leq \beta_0 - k_\alpha \sigma/\sqrt{M}) = P\left(\frac{\hat{\beta} - \beta}{\sigma/\sqrt{M}} \leq \frac{\beta_0 - \beta}{\sigma/\sqrt{M}} - k_\alpha\right) \\
&\approx G\left(\frac{\beta_0 - \beta}{\sigma/\sqrt{M}} - k_\alpha\right)
\end{aligned}$$

Vi ser at $\Pi(\beta)$ er en avtagende funksjon av β. Om ønskelig kan vi bruke styrkefunksjonen til å lage en test som oppfyller både et gitt nivå og styrkekrav.

Eksempel 6 : Testing av gradvis forbedring
En bedrift har nettopp satt i verk en forbedret rensemetode før avfallsvannet slippes ut i en innsjø. Man er bl.a. interessert i å undersøke om den nye metoden gradvis kan lede til en redusert konsentrasjon av et giftstoff i innsjøen. Anta at man på den første i hver måned i 11 måneder har hentet opp en vannprøve og måler dens giftkonsentrasjon. Lar vi tidsenheten være en måned kan resultatene presenteres slik

t_i :	1	2	3	4	5	6	7	8	9	10	11
X_i :	15.1	17.4	16.0	15.5	13.2	16.7	14.3	13.7	15.2	11.5	12.7

Ut fra tidligere erfaring med slike målinger mener man at standardavviket σ ved en enkelt måling er 1.5. Situasjonen kan med rimelighet analyseres innenfor rammen av en regresjonsmodell og problemet kan formuleres som å teste en hypotese om regresjonskoffisienten i en slik modell. Det at den nye rensemetoden overhodet ikke har innvirkning på giftkonsentrasjonen svarer til $H_0 : \beta = 0$ og forbedring svarer til $H_A : \beta < 0$, idet vi antar at den nye rensemetoden i hvert fall ikke kan forverre situasjonen. Anta at man er villig til å bruke et signifikansnivå på α=0.10. Her blir $\bar{t}$=6.0, $M = \sum(t_i - \bar{t})^2$=110 og $\hat{\beta} = -0.368$. Den realiserte verdi av Z blir -2.573 som er mindre enn den kritiske verdi $k = -1.28$. Følgelig forkaster vi nullhypotesen og påstår at den nye rensemetoden har gitt positiv virkning, om virkningen svarte til forventningen eller har praktisk betydning, f.eks. for fisket, er en annen sak. Hvilke innvendinger kan reises mot denne analysen?

Vi har ovenfor studert testing når alternativet til nullhypotesen er av typen $\beta < \beta_0$. I andre situasjoner vil alternativet kunne være $\beta > \beta_0$. Framgangsmåten er da tilsvarende, og detaljene overlates til leseren. Det samme gjelder dersom alternativet er $\beta \neq \beta_0$. Ovenfor har vi forutsatt at standardavviket σ er kjent. Dersom σ er ukjent og er estimert med S, blir teorien noe mer problematisk. Mer om dette i neste avsnitt.

Vi hoppet over begrunnelsen for forventningsrett, varians og minste kvadraters egenskapen ovenfor. Det kommer her som $\star$-merket stoff.

$\star$ Regnereglene for forventning og varians gir (merk at $\sum_{i=1}^n (t_i - \bar{t}) = 0$)

$$\begin{aligned}
E(\hat{\alpha}) &= E(\frac{1}{n}\sum_{i=1}^n X_i) = \frac{1}{n}\sum_{i=1}^n EX_i \\
&= \frac{1}{n}\sum_{i=1}^n (\alpha + \beta(t_i - \bar{t})) = \alpha \\
E(\hat{\beta}) &= E(\frac{1}{M}\sum_{i=1}^n (t_i - \bar{t})X_i) = \frac{1}{M}\sum_{i=1}^n (t_i - \bar{t})EX_i \\
&= \frac{1}{M}\sum_{i=1}^n (t_i - \bar{t})(\alpha + \beta(t_i - \bar{t})) = \beta. \\
var(\hat{\alpha}) &= var(\frac{1}{n}\sum_{i=1}^n X_i) = (\frac{1}{n})^2 \sum_{i=1}^n varX_i \\
&= (\frac{1}{n})^2 \sum_{i=1}^n \sigma^2 = (\frac{1}{n})^2 \cdot n\sigma^2 = \frac{\sigma^2}{n}
\end{aligned}$$

$$\begin{aligned} var(\hat{\beta}) &= var(\frac{1}{M}\sum_{i=1}^{n}(t_i-\bar{t})X_i) = (\frac{1}{M})^2\sum_{i=1}^{n}(t_i-\bar{t})^2 varX_i \\ &= (\frac{1}{M})^2\sum_{i=1}^{n}(t_i-\bar{t})^2\cdot\sigma^2 = \sigma^2\cdot(\frac{1}{M})^2\cdot M = \frac{\sigma^2}{M}. \end{aligned}$$

Følgende formel er nyttig. Begrunnelsen krever en viss evne til å manipulere kvadratsummer (Se Oppgave 33 (b))

$$\sum_{i=1}^{n}(X_i-\alpha-\beta(t_i-\bar{t}))^2 = \sum_{i=1}^{n}(X_i-\hat{\alpha}-\hat{\beta}(t_i-\bar{t}))^2 + n(\hat{\alpha}-\alpha)^2 + M(\hat{\beta}-\beta)^2.$$

Herav følger bl. a. at

$$Q(\alpha,\beta) \geq \sum_{i=1}^{n}(X_i-\hat{\alpha}-\hat{\beta}(t_i-\bar{t}))^2$$

med likhet hvis og bare hvis α og β velges lik henholdsvis $\hat{\alpha}$ og $\hat{\beta}$. Dette viser at $\hat{\alpha}$ og $\hat{\beta}$ virkelig er minste kvadraters estimatorene for α og β. Formelen kan også brukes til beregning av ES^2. Tar vi forventning på begge sider av likhetstegnet får vi

$$\begin{aligned} n\sigma^2 &= nES^2 + nvar\hat{\alpha} + Mvar\hat{\beta} \\ &= nES^2 + n\cdot\frac{\sigma^2}{n} + M\cdot\frac{\sigma^2}{M}. \end{aligned}$$

Herav ser vi at

$$ES^2 = \frac{n-2}{n}\sigma^2.$$

8.4 Normalitetsantakelse og t-kurven

Den målemodellen som ble presentert i Kapittel 8.1 inneholdt ingen eksplisitte forutsetninger om sannsynlighetsfordelingene til hver av de stokastiske variable som inngikk i modellen, og vi kunne derfor heller ikke si noe eksakt om sannsynlighetsfordelingen til gjennomsnittet $\overline{X}$ av disse. Ved bruk av normaltilnærmelsen, var det likevel mulig å etablere tilnærmede sannsynlighetsutsagn om estimeringsfeilen når $\overline{X}$ ble brukt som estimator, samt etablere tester med tilnærmet signifikansnivå basert på $\overline{X}$ som testobservator, i hvertfall i situasjoner der variansen σ^2 var kjent. Tilsvarende kommentarer gjelder for den enkle regresjonsmodellen i Kapittel 8.3.

I praksis ser en imidlertid ofte brukt målemodeller og regresjonsmodeller med en tilleggsforutsetning om sannsynlighetsfordelingen til de stokastiske variable (observasjonene), nemlig at disse selv er normalfordelte. Dette er et eksempel på en såkalt *kontinuerlig sannsynlighetsmodell*, der hele den reelle tallinjen er utfallsrom, og sannsynligheter beregnes eksakt som arealer under en modellkurve. Vi har begrenset oss til å bygge opp generell teori for diskrete modeller, og som vi har sett, har det vært mulig å studere en lang rekke interessante sannsynlighetsteoretiske og statistiske problemstillinger uten å trekke inn kontinuerlige modeller. Innføring av kontinuerlige modeller åpner imidlertid nye muligheter, og en rekke problemer lar seg f.eks. bare beskrive hensiktsmessig med en kontinuerlig modell. Vi skal ikke her utvide det generelle teorigrunnlag med kontinuerlige modeller, men bare nøye oss med å gi et innblikk i hvilke konsekvenser en antakelse om eksakt normalitet har i forbindelse med målemodellen og regresjonsmodellen.

Først hva menes med at en stokastisk variabel X er normalfordelt:

> **Definisjon :** Vi sier at X er *normalfordelt* med forventning μ og varians σ^2 dersom alle sannsynligheter i forbindelse med den standardiserte variable
>
> $$Z = \frac{X - \mu}{\sigma}$$
>
> kan beregnes eksakt som arealer under normalkurven
>
> $$g(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}$$
>
> I så fall skriver vi at X er fordelt $N(\mu, \sigma^2)$.

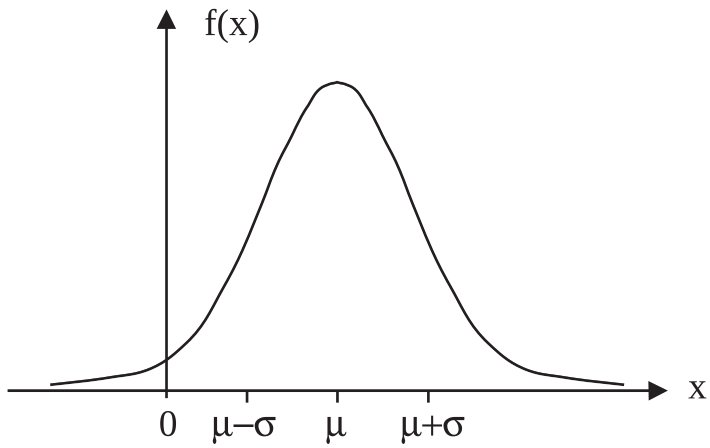
Da blir eksempelvis

$$P(X \leq x) = P\left(Z \leq \frac{x - \mu}{\sigma}\right) = G\left(\frac{x - \mu}{\sigma}\right)$$

der $G(z)$ er arealet under normalkurven $g(z)$ til venstre for z. I en normalmodell er derfor siste likhet å oppfatte som eksakt, ikke som tilnærming. Sannsynlighetsfordelingen til X er i virkeligheten beskrevet ved kurven i Figur 8.5, hvor μ angir symmetripunktet, og σ er et mål for spredningen omkring dette.

Kurven er bestemt ved funksjonen

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



Figur 8.5: Normalfordelingen $N(\mu, \sigma^2)$

og alle sannsynligheter i forbindelse med X kan finnes ved å beregne arealer under denne kurven. Som vi har sett kan dette alltid føres tilbake til å beregne arealer under normalkurven $g(z)$. Denne svarer til $N(0,1)$, dvs. normalfordelingen med forventning 0 og varians 1, den såkalte *standardnormalfordelingen*.

I fagterminologien blir $f(x)$ kalt *sannsynlighetstettheten* til X. Generelt for slike gjelder at de er ikke-negative og avgrenser et totalt areal lik 1. En kontinuerlig sannsynlighetsmodell er en modell som kan beskrives ved en sannsynlighetstetthet og der sannsynligheter er gitt ved arealer under tettheten. For regning med kontinuerlige modeller trengs ferdighet i integralregning, eksempelvis er

$$P(a \leq X \leq b) = \int_a^b f(x)dx$$

og forventningen til X er definert ved

$$EX = \int_{-\infty}^{\infty} xf(x)dx$$

I forbindelse med normalfordelte variable kan det vises en rekke setninger, spesielt viktig er denne:

> En stokastisk variabel som kan skrives som en lineær funksjon av uavhengige normalfordelte variable er selv normalfordelt.

Dette har en rekke viktige konsekvenser bl.a.:

For målemodellen: Dersom vi føyer til en antakelse om at observasjonene $X_1, X_2, \ldots, X_n$ alle er normalfordelte, medfører dette at $\overline{X}$, som er en lineær funksjon av observasjonene, også er normalfordelt, med forventning μ og varians σ^2/n.

For regresjonsmodellen: Dersom vi føyer til en antakelse om at observasjonene $X_1, X_2, \ldots, X_n$ alle er normalfordelte, medfører dette at $\hat{\alpha}$ og $\hat{\beta}$, som begge er lineære funksjoner av observasjonene, også er normalfordelt, med forventninger α og β og varianser henholdsvis σ^2/n og σ^2/M.

Dette betyr at i en målemodell eller en regresjonsmodell hvor normalitet inngår som en av forutsetningene i modellen, vil de sannsynlighetsutsagn som ble etablert i avsnittene 8.1 og 8.3 som tilnærmede utsagn, være å oppfatte som eksakte. Dette gjelder for de problemstillinger der variansen til observasjonene σ^2 er kjent. Mer oppsiktsvekkende er at det er mulig å etablere eksakte sannsynlighetsutsagn i situasjoner der σ^2 er ukjent og må estimeres på grunnlag av observasjonene. La oss utdype dette:

I målemodellen med normalitetsantakelse vil den stokastiske variable

$$Z = \frac{\overline{X} - \mu}{\sigma(\overline{X})} = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$$

være en nøkkelstørrelse i forbindelse med inferens om μ. Den vil være fordelt $N(0, 1)$ og danner i situasjoner der σ er kjent et utgangspunkt for konstruksjon av sannsynlighetsutsagn. Dersom σ er ukjent er Z ubrukelig til dette. Man kan isteden ta utgangspunkt i den stokastiske variable

$$T = \frac{\overline{X} - \mu}{S(\overline{X})} = \frac{\overline{X} - \mu}{S/\sqrt{n}}$$

dvs. en analogi til Z der σ er erstattet med en estimator S. Det har vært vanlig å bruke $S = \sqrt{S^2}$ der

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2$$

som er en forventningsrett estimator for σ^2 i målemodellen. I regresjonsmodellen med normalitetsantakelse vil den stokastiske variable

$$Z = \frac{\hat{\beta} - \beta}{\sigma(\hat{\beta})} = \frac{\hat{\beta} - \beta}{\sigma/\sqrt{M}}$$

være fordelt $N(0,1)$ og danner i situasjoner der σ er kjent et utgangspunkt for konstruksjon av sannsynlighetsutsagn ved inferens om β. Dersom σ er ukjent kan man isteden ta utgangspunkt i

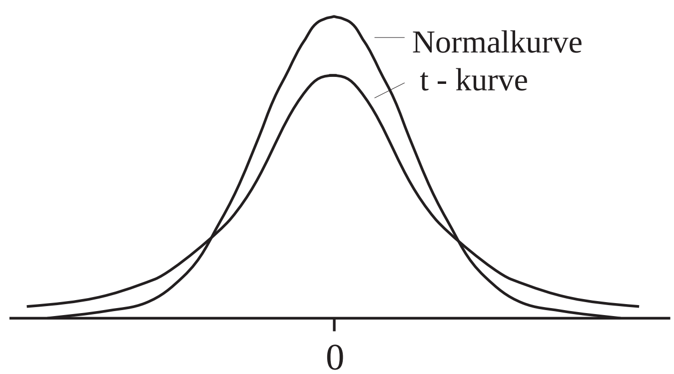
$$T = \frac{\hat{\beta} - \beta}{S(\hat{\beta})} = \frac{\hat{\beta} - \beta}{S/\sqrt{M}}$$

dvs. σ er erstattet med en estimator S. Det er vanlig å bruke $S = \sqrt{S^2}$ der

$$S^2 = \frac{1}{n-2} \sum_{i=1}^{n} (X_i - \hat{\alpha} - \hat{\beta}(t_i - \bar{t}))^2$$

som er en forventningsrett estimator for σ^2 i regresjonsmodellen.

Dersom vi kjente sannsynlighetsfordelingen til T i situasjonene ovenfor, kunne vi bruke T til å lage sannsynlighetsutsagn på tilsvarende måte som med Z i situasjonen med σ kjent. Under forutsetning av at observasjonene i modellen er normalfordelte kan det vises at T, i begge situasjoner, har en kontinuerlig modellkurve som er klokkeformet og symmetrisk om null, en såkalt t-kurve. Den ligner normalkurven, men har noe tyngre “halerenn denne, se Figur 8.6.



Figur 8.6: t-kurve og normalkurve

Det matematiske uttrykk for t-kurven er relativt komplisert og utelates her. Kurvens eksakte form vil avhenge av n, med svært tunge haler for liten

n og mer og mer lik normalkurven ettersom n vokser. Dette lar seg forklare ved at når σ blir erstattet med S vil dette være årsak til større risiko for verdier fjernt fra symmetripunktet, men når n er stor vil S gi et meget presist anslag for σ, slik at det spiller mindre rolle om σ er erstattet med S.

I målemodellen brukes den såkalte t-fordeling med $n-1$ frihetsgrader, i regresjonsmodellen t-fordelingen ned $n-2$ frihetsgrader. Frihetsgradtallet uttrykker da graden av usikkerhet som skyldes at σ er estimert: stort frihetsgradtall, liten usikkerhet.

Arealer under t-kurver er tabellert og er lett tilgjengelig i tabellverker. Slike tabeller kan, i situasjoner der σ er ukjent, brukes på samme måte som vi har brukt arealer under normalkurven i situasjonen med kjent varians. Siden det trengs en t-tabell for hvert frihetsgradtall, er t-tabellene vanligvis mindre detaljerte enn de tilsvarende normaltabeller. Vanligvis gis punkter som avgrenser bestemte arealer til høyre eller venstre for seg, såkalte fraktiltabeller, se Tabell C.9 i Appendiks C. Med andre ord har vi at

$$P(Z \leq k) = G(k) \quad \text{mens} \quad P(T \leq k) = G_m(k)$$

der $G_m(k)$ betegner arealet til venstre for k under t-kurven med m frihetsgrader, med $m = n-1$ for målemodellen og $m = n-2$ for regresjonsmodellen. Videre er

$$P(\mid Z \mid \leq k) = A(k) \quad \text{mens} \quad P(\mid T \mid \leq k) = A_m(k)$$

der $A_m(k)$ betegner arealet mellom $-k$ og k under t-kurven. På grunn av symmetrien har vi som i tilfellet med normalkurven at

$$A_m(k) = 2G_m(k) - 1$$

For målemodellen betyr dette at

$$P(\mid \overline{X} - \mu \mid \leq k \cdot S(\overline{X})) = A_{n-1}(k)$$

dvs. vi kan i prinsippet konstruere eksakte sannsynlighetsutsagn vedrørende $\overline{X}$ som estimator for μ av typen: Sannsynligheten for at $\overline{X}$ avviker fra μ med høyst k ganger anslått standardavvik er $A_{n-1}(k)$. For regresjonsmodellen gjelder tilsvarende at

$$P(\mid \hat{\beta} - \beta \mid \leq k \cdot S(\hat{\beta})) = A_{n-2}(k).$$

Her er det også mulig å lage sannsynlighetsutsagn om estimeringsfeilen ved $\hat{X}$ som estimator for forventet respons $EX = \gamma + \beta t$ for en gitt t. Vi har nemlig

$$P(\mid \hat{X} - (\gamma + \beta t) \mid \leq k \cdot S(\hat{X})) = A_{n-2}(k)$$

Endelig er det mulig å lage sannsynlighetsutsagn om prediksjonsfeilen ved å bruke $\hat{X}$ som prediktor for en ny observasjon X for gitt t:

$$P(\mid \hat{X} - X \mid \leq k \cdot S(\hat{X} - X)) = A_{n-2}(k)$$

I disse formlene inngår estimerte standardavvik for hhv. prediktor og prediksjonsfeil, jfr. Oppgave 35.

$$\begin{aligned} S(\hat{X}) &= S \cdot \sqrt{\frac{1}{n} + \frac{(t - \bar{t})^2}{M}} \\ S(\hat{X} - X) &= S \cdot \sqrt{1 + \frac{1}{n} + \frac{(t - \bar{t})^2}{M}} \end{aligned}$$

Merk at begge avhenger av t, og øker ettersom t avviker mer og mer fra $\bar{t}$.

Vi har altså konstatert at antakelse om normalitet åpner nye muligheter, men under hvilke omstendigheter vil en normalitetsantakelse i tillegg til de andre antakelsene i en modell være realistisk? I praksis finner vi to typer argumenter. Enten brukes et teoretisk argument om at verdien av hver stokastisk variabel i modellen er bestemt av en rekke faktorer som virker uavhengig av hverandre, og en referanse til sentralgrensesetningen støtter da opp om en normalitetsantakelse. Eksempelvis kan høyden til en utvalgt 16-åring tenkes å være bestemt av en rekke arvemessige og miljømessige faktorer som summeres opp. Alternativt brukes et empirisk argument om at praktisk erfaring fra lignende situasjoner har vist at den sannsynlighetsfordeling det er tale om har klokkefasong som til forveksling ligner normalfordelingen. Hvorfor ikke uttrykke slik forhåndsviten eksplisitt i modellen?

I mange situasjoner er det imidlertid ikke rimelig å trekke veksler på noen av disse argumentene, og da bør en selvsagt ikke inkludere normalitet blant forutsetningene i modellen. I praksis ser en dessverre ofte benyttet metoder som forutsetter normalitet uten at dette rettferdiggjøres, ja endog er brukeren bevisst. Det finnes metoder til å sjekke om et tallmateriale med rimelighet kan sies å være generert av en normalfordeling.

La oss nå kommentere visse optimale egenskaper ved de metoder som er foreslått i Kapittel 8.1 og 8.3, og deretter komme med visse reservasjoner:

For estimering av forventningen μ i målemodellen kan det vises at estimatoren $\overline{X}$ har følgende egenskaper:

1. $\overline{X}$ er den forventningsrette estimator for μ som har minst varians blant de forventningsrette estimatorer som er lineære funksjoner av observasjonene $X_1, X_2, \ldots, X_n$ (Gauss-Markov's setning).

2. $\overline{X}$ er den forventningsrette estimator μ som har minst varians dersom vi er villig til å anta at sannsynlighetsfordelingen til observasjonene er normal.

Vi kan imidlertid innvende: Under 1. begrenser en seg til klassen av lineære forventningsrette estimatorer. Kan hende finnes det en estimator utenfor denne klassen som er bedre. Under 2. antas det at observasjonene har en såkalt normalfordeling. Kan hende dette er en urealistisk antakelse.

Det finnes også teoretiske resultater som sier at ved testing av hypoteser om μ i målemodellen, så vil testmetoder basert på $\overline{X}$ være optimale i en viss forstand. Videre finnes teori som sier at ved estimering og testing i den enkle regresjonsmodellen, så vil metoder basert på $\hat{\alpha}$ og $\hat{\beta}$ slik de er definert i Kapittel 8.3 være optimale i en viss forstand.

Med utgangspunkt i de ideer som er skissert ovenfor er det med årene utviklet omfattende generell teori basert på forutsetninger om normalitet (og ukjent varians), bl.a. teorien for lineære normale modeller som i dag regnes som klassisk statistisk teori.[7] Videre er det utviklet brukerorienterte spesialområder såsom variansanalyse og regresjonsanalyse, se Kapittel 11 og 12.

8.5 Mer om måleproblemer

Ved de fleste praktiske anvendelser av målemodellen, vil nok variansen til observasjonene være ukjent. Dersom vi kan rettferdiggjøre at observasjonene er normalfordelte, vil en kunne dra nytte av teorien i forrige avsnitt til å konstruere konfidensgrenser og tester med eksakte sikkerhetsgarantier, selv når antall observasjoner er lite. Bruker vi metoder basert på modeller der variansen er kjent, forutsetter vi egentlig at vi allerede har anslått denne i en lignende situasjon, med et stort antall observasjoner som grenser til sikkerhet, og at vi er villig til å anta at variabiliteten er den samme i den foreliggende situasjon.

Dette kan være en betenkelig praksis, fordi omstendighetene kan være endret uten at vi er klare over det. En varians som i modellen er gitt en urealistisk verdi, kan føre til feiltolkning av tallmaterialet. For å unngå dette er det anbefalt at variansen anslås ut fra tallmaterialet selv, også når det

[7] Som grunnlegger av denne teori regnes Ronald A. Fisher (1890-1962).

foreligger betydelig forhåndsinformasjon. I målemodellen med normalitetsantakelse om observasjonene bestemmer

$$\overline{X} \pm k \cdot S(\overline{X})$$

et konfidensintervall for forventningen μ med konfidensnivå $c = A_{n-1}(k)$ (arealet mellom $-k$ og k under t- kurven med $n-1$ frihetsgrader). Ønsker vi å teste hypotesen $H_0 : \mu = \mu_0$ mot en alternativ hypotese, kan vi bruke den såkalte t-observatoren

$$T = \frac{\overline{X} - \mu_0}{S(\overline{X})}$$

som, dersom H_0 er riktig, er t-fordelt med $n-1$ frihetsgrader. Forkastningsområdet og bestemmelse av kritisk verdi avhenger av den alternative hypotesen, likeledes eventuell beregning av P-verdi. Vi har i hovedsak tre tilfeller:

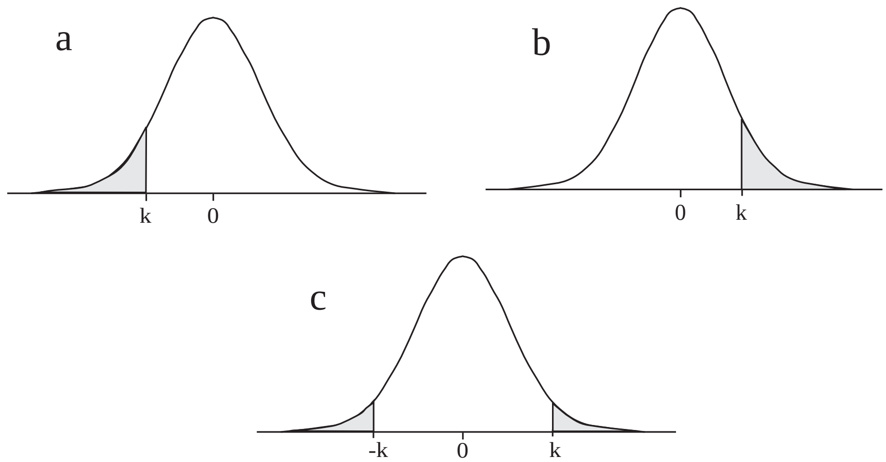
Tilfelle	Alternativ	Forkastningsområde	P-verdi
a.	$\mu < \mu_0$	$T \leq k$	$P_{H_0}(T \leq t)$
b.	$\mu > \mu_0$	$T \geq k$	$P_{H_0}(T \geq t)$
c.	$\mu \neq \mu_0$	$\mid T \mid \geq k$	$P_{H_0}(\mid T \mid \geq t)$

Den kritiske verdi k bestemmes slik at arealet under t-kurven svarende til forkastningsområdet er lik det ønskede signifikansnivå, se Figur 8.7a-c.

Vi husker at et konfidensintervall kan tolkes som mengden av plausible verdier av μ. I tilfellet c der alternativet er tosidig, kan en alternativt bruke konfidensintervallet ved hypotesetesting. Dersom nullhypotesen faller utenfor konfidensgrensene forkastes hypotesen. Dette er ekvivalent med t-testen med signifikansnivå $\alpha = 1 - c$, der c er konfidensnivået. Vi kan også bruke konfidensgrensene når alternativet er ensidig. Da forkastes hypotesen når denne er utenfor den relevante konfidensgrensen (den øvre i tilfelle a og nedre tilfelle b). I det tilfellet er signifikansnivået isteden $\alpha = (1-c)/2$, dvs. $c = 1 - 2\alpha$.

Eksempel 7 : Forurensing
La situasjonen være som i Eksempel 3, med følgende observerte verdier av kvikksølvkonsentrasjonen

14.8 15.4 16.1 14.7 17.5 15.9 13.4 16.0 15.7



Figur 8.7: Kritiske verdier for t-testen

Vi finner at $\overline{X}$=15.5, S=1.138, slik at $S(\overline{X}) = S/\sqrt{9} = 0.38$. Antas observasjonene normalfordelte vil eksakte konfidensintervaller for forventet konsentrasjon μ ha form

$$15.50 \quad \pm \quad k \cdot 0.38$$

der sikkerhetsfaktoren k finnes i Tabell C.9 med frihetsgradtall $\nu = 9 - 1 = 8$ og det ønskede konfidensnivå. Eksempelvis krever konfidensnivået 0.95 at $k = 2.31$, mens $k = 1.96$ var nok i tilfellet med kjent varians. Konfidensnivået lik 0.90 krever $k = 1.86$, mot $k = 1.645$ i tilfellet med kjent varians, se Oppgave 2. La oss se på en typisk t-analyseutskrift:

```
 >> READ 'eks8.7' X
>> TINTERVAL 0.90 X
    N    Mean  StDev  SeMean   90.0 % ConfInt

 X  9  15.500  1.138   0.379  ( 14.79, 16.21 )
>> TTEST 17.2 X
Test of MU = 17.200 vs MU not = 17.200
    N    Mean  StDev  SeMean      T  P-value

 X  9  15.500  1.138   0.379  -4.48   0.0021
```

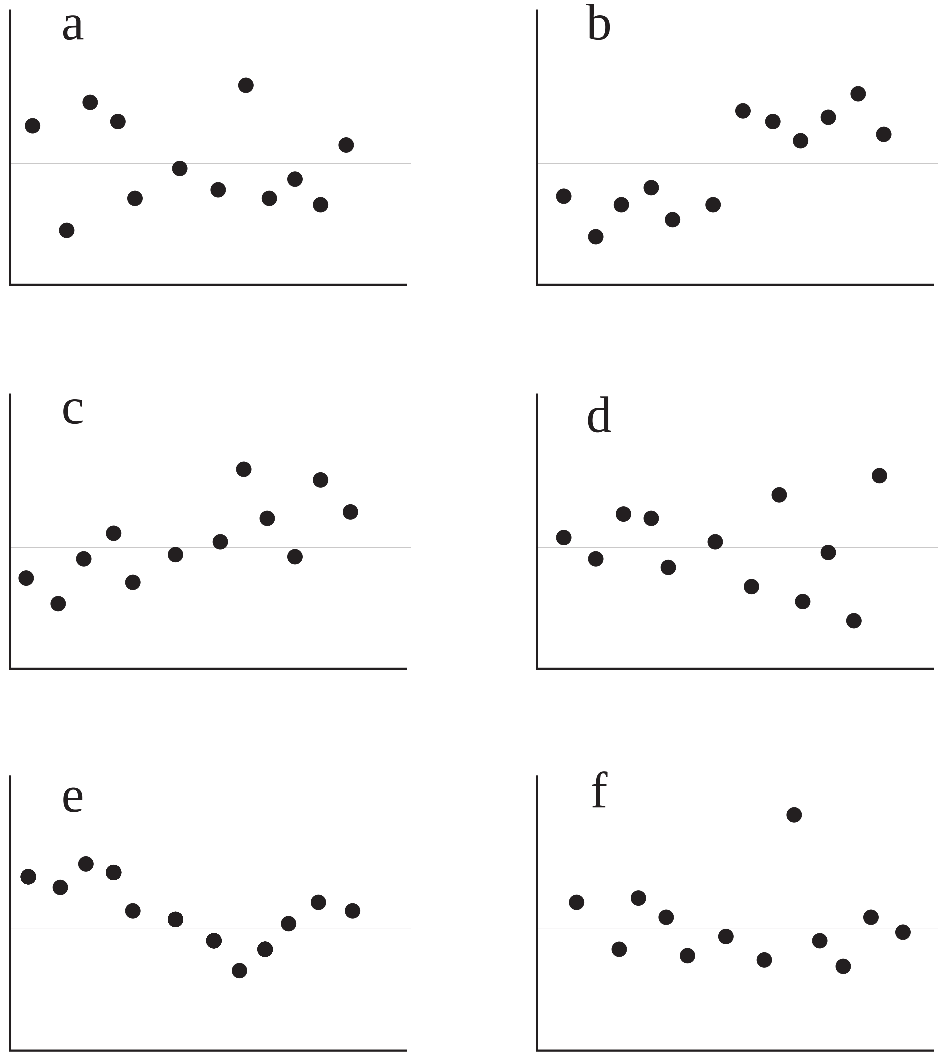
Her er data lest fra en fil 'eks8.7' og lagret som en søyle X. Vi ber om et konfidensintervall med konfidensnivå 0.90, som gir intervallet [14.79, 16.21], i tillegg til de beskrivende mål. Her er

Mean=gjennomsnittet av observasjonene
StDev=standardavviket til observasjonene
SeMean=anslått standardavvik til gjennomsnittet (ofte kalt standardfeilen). Leseren kan selv sjekke at beregningen samsvarer med teorien ovenfor. Videre ber vi om å få utført en t-test av hypotesen $\mu = 17.20$, og får, i tillegg til de beskrivende mål, beregnet T og P-verdien. Med 5% nivå forkastes hypotesen ($P < 0.05$). Dersom bare en eventuell reduksjon i kvikksølvkonsentrasjonen er aktuell eller ønskes fastslått, slik at alternativet til nullhypotesen er ensidig, må den oppgitte P-verdi halveres, dvs. $P = 0.00105$. Programvare gir vanligvis anledning til å spesifisere dette.

Egen kommando for t-test er egentlig unødvendig, idet vi kan sjekke om nullhypotesen ligger i t-intervallet eller ikke. Konfidensnivå 90% svarer til 10% signifikansnivå i det tosidige tilfellet og 5% i det ensidige tilfellet.

Planlegging av antall observasjoner i forbindelse med T-metoder krever betydelig mer kompliserte regninger enn tilfellet var for de tilsvarende Z-metodene. Grove betraktninger med antatt kjent varians kan likevel være til noen hjelp, men statistiske programpakker bør kunne gi støtte her. Hittil har dette aspekt vært neglisert i de fleste pakker, men den følge at t-intervaller og t-testen misbrukes i betydelig grad i praksis. Ukritisk bruk av konfidensintervaller med konvensjonelle konfidensnivåer $(90\%, 95\%, 99\%)$ innebærer ofte en ensidig fokusering på sikkerheten på bekostning av presisjonen (konfidensintervallets lengde), og neglisjering av planleggingsfasen. Anvendt i forkast-aksept situasjoner innebærer dette en neglisjering av risikoen for akseptfeil. Ukritisk hypotesetesting med konvensjonelle signifikansnivåer eller beregning av P-verdier, innebærer ofte at risikoen for godtakingsfeil i forbindelse med vesentlige avvik fra nullhypotesen ikke er vurdert, igjen et planleggingsspørsmål.

Det hender ofte at en er usikker på om forutsetningene for bruk av målemodellen er oppfylt, og en kan ønske seg metoder som forsøker å stille en diagnose ut fra observasjonene selv. Forutsetningene var: (i) samme forventning (ii) samme varians (iii) uavhengighet og eventuelt (iv) normalitet. Det finnes formelle metoder der hver forutsetning tas som en nullhypotese, og der en tilhørende test er egnet til å avsløre spesielle alternativer til denne. Vi vil ikke gå inn på slike metoder her, men heller antyde bruken av mer uformelle grafiske metoder, sammen med en viss intuisjon. Ofte kan brudd på forutsetningene $(i) - (iii)$ avsløres ved å se på observasjonene gruppevis hvis de er tatt på ulikt tidspunkt, ulikt sted eller av ulike personer. Dersom det er en naturlig rekkefølge av observasjonene er det lurt å plotte dem som funksjon av observasjonsnummeret. Eksempler på dette gis i Figur 8.8a-f.



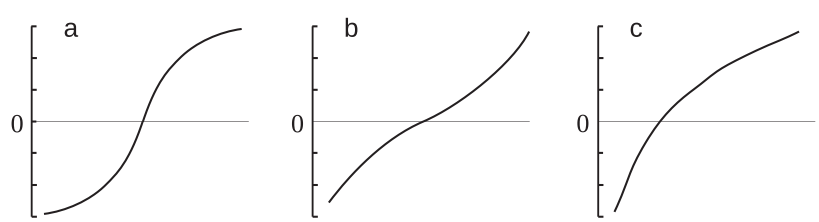
Figur 8.8: Diagnose-plott

I Figur 8.8a ser observasjonene ut til å variere tilfeldig omkring en forventning markert med den horisontale linje, og at (i) - (iii) er oppfylt. I Figur 8.8b ser observasjonene ut til å falle i to grupper med ulik forventning, slik at (i) ikke er oppfylt, men (ii) godt kan være det. Teori for denne situasjonen finnes i Kapittel 11. I Figur 8.8c ser det ut til at forventningen øker med observasjonsnummeret, men at (ii) og (iii) godt kan være oppfylt. Dette innbyr til en regresjonsmodell, dersom det er meningsfylt. I Figur 8.8d ser det ut til at variansen øker, slik at (ii) ikke er oppfylt. I Figur 8.8e ser det ut til at etterfølgende observasjoner er korrelerte, slik at (iii) ikke er oppfylt. Teori for denne situasjon finnes i Kapittel 13. Figur 8.8f antyder en “vill”observasjon som bryter med forutsetningen om normalitet, og som eventuelt må korrigeres eller fjernes før t-metoden brukes.

En indikasjon på om normalitet (iv) er en rimelig antakelse, får vi ved å gruppere observasjonene, og lage et histogram, se Figur 1.2 i Eksempel 1.6. Det kan imidlertid være vanskelig å skille en normalfordeling fra andre symmetriske fordelinger. For dette formål kan derfor et såkalt normalskår-plott være et hensiktsmessig diagnoseredskap. De n observasjonene $X_1, X_2, \ldots, X_n$ ordnes først i stigende rekkefølge. Den i'te observasjon talt nedenfra, kall den $X_{(i)}$, tildeles så en såkalt normalskår $Y_{(i)}$ bestemt ved [8]

$$G(Y_{(i)}) = \frac{i}{n+1} \quad i = 1, 2, \ldots, n$$

Vi har da bestemt en normalskår Y_i til hver observasjon X_i, og vi kan plotte tallparet (X_i, Y_i) for $i = 1, 2, \ldots, n$ i et spredningsdiagram. Dersom den observerte fordeling av X_i'ene kan tilnærmes med en normalfordeling, vil punktene i normalskårplottet ligge noenlunde på linje. Figur 8.9a-c illustrerer formen på plottet for noen typiske avvik fra normalitet. Dersom den observerte fordeling er symmetrisk, men har “tyngre halerenn normalfordeling, vil punktene anta en S-form (Figur 8.9a), mens ved “lettere halerenn normalfordelingen, vil punktene anta en omvendt S (Figur 8.9b). Dersom den observerte fordeling er skjev med lang hale mot høyre, vil plottet ha form som i Figur 8.9c.



Figur 8.9: Avvik fra normalitet

[8]Noen statistiske programpakker med innlagt normalskårfunksjon bytter om de to aksene, og bruker kanskje en annen likeverdig definisjon.

Nå vil, som vi har antydet ovenfor, normalitet kunne være en urealistisk antakelse, men man trodde lenge at de metodene som hadde vist seg å være optimale under normalitetsantakelser, også var uten alvorlige konkurrenter i situasjoner der den underliggende sannsynlighetsfordeling ikke nødvendigvis var normal. Dette viste seg ikke alltid å være tilfellet. I de senere tiår har statistikerne derfor i stigende grad rettet oppmerksomheten mot en rekke alternative metoder, såkalte *robuste metoder*. Disse har den egenskapen at dersom den underliggende fordeling virkelig er normal så taper en noe, men ikke mye, i forhold til de "optimale"metodene, mens på den annen side kan gevinsten være betydelig dersom fordelingen avviker noe fra normalfordelingen.

La oss i forbindelse med målemodellen kort antyde en situasjon hvor estimatoren $\overline{X}$ får alvorlige konkurrenter. Anta at det er en viss sjanse for "ville "observasjoner, dvs. verdier langt fra den μ vi skal estimere. I et gjennomsnitt hvor alle observasjonene er gitt samme vekt, vil en "vill"verdi lett kunne ødelegge den tendens som de andre observasjonene samlet gir. En mulig måte å redusere virkningen av ville observasjoner, er å ordne observasjonene i stigende rekkefølge, stryke den minste og den største observasjonen, og ta gjennomsnittet av de resterende $n-2$ observasjonene. Mer generelt kunne vi stryke de k minste og de k største observasjonene og ta gjennomsnittet av de resterende. Et slikt gjennomsnitt kalles et *trimmet gjennomsnitt*. Den mest ekstreme form for trimming får vi når vi stryker alle observasjonene unntagen den midterste (evt. de to miderste). Vi sitter da igjen med den såkalte *medianen* av observasjonene. Det kan vises at i visse situasjoner vil medianen være å foretrekke framfor alle andre estimatorer. I Eksempel 1 ser vi at medianen av observasjonene er 0.65, mens i Eksempel 2 er medianen 9.725. Dersom man i en gitt situasjon velger å rapportere medianen (eller et annet trimmet gjennomsnitt) bør også dette følges av (estimert) standardavvik.

Formler for standardavvik til trimmede gjennomsnitt er ofte kompliserte, vanskelig å utlede, og avhenger typisk av den underliggende fordeling. Det fins en alternativ tilnærming, som gjør det mulig å anslå standardavvik i ikke-standard situasjoner, uten å kjenne de fordelingsteoretiske egenskapene i detalj. Metoden kalles *resampling*, og er blitt mulig først med dagens raske regneteknologi. En kort beskrivelse er gitt i avsnitt 8.7.

8.6 Mer om regresjonsproblemer

I de fleste praktiske situasjoner der en foretar statistiske analyser basert på en regresjonsmodell, er det lite rimelig å anta at variansen til observasjonene er kjent. Dersom vi er villige til å anta at observasjonene er normalfordelte, vil derfor teorien i Kapittel 8.4 kunne hjelpe oss et stykke videre. Mange av dagens regresjonsanalyseprogrammer gir også utskrifter hvor fortolkningen i betydelig grad er knyttet til normalitetsantakelsen.

I regresjonsmodellen med normalitetsantakelse om observasjonene gir

$$\hat{\beta} \;\; \pm \;\; k \cdot S(\hat{\beta})$$

et konfindensintervall for regresjonskoeffisienten β med konfidensnivå $c = A_{n-2}(k)$, dvs. lik arealet mellom $-k$ og k under t-kurven med $n-2$ frihetsgrader. Ønsker vi å teste hypotesen $H_0 : \beta = \beta_0$ mot en alternativ hypotese kan vi bruke t- observatoren

$$T = \frac{\hat{\beta} - \beta_0}{S(\hat{\beta})}$$

som, dersom H_0 er riktig, er t-fordelt med $n-2$ frihetsgrader. Forkastningsområdet og bestemmelse av kritisk verdi avhenger av den alternative hypotesen, likeledes beregning av P-verdi. Vi kan bruke den samme skjematiske oppstilling og figurer som for målemodellen i forrige avsnitt. I noen situasjoner er det aktuelt å lage konfidensintervaller og teste hypoteser om konstantleddet. Dette kan gjøres på tilsvarende måte som for regresjonskoeffisienten. Resultatene i Kapittel 8.4. gir videre at

$$\hat{X} \;\; \pm \;\; k \cdot S(\hat{X})$$

er et konfidensintervall for forventningen EX for en gitt t, mens

$$\hat{X} \;\; \pm \;\; k \cdot S(\hat{X} - X)$$

er et såkalt prediksjonsintervall for en ny uavhengig observasjon for en gitt t. Sikkerhetsfaktoren k bestemmes i begge tilfeller slik at arealet mellom $-k$ og k under t-kurven med $n-2$ frihetsgrader er lik det ønskede konfidensnivå. Regresjonsanalyseprogrammer vil som regel kunne utføre beregningen av ønskede estimerte standardavvik, enten automatisk eller som en opsjon, slik at formlene fra Kapittel 8.4. kan unngås. Merk imidlertid at

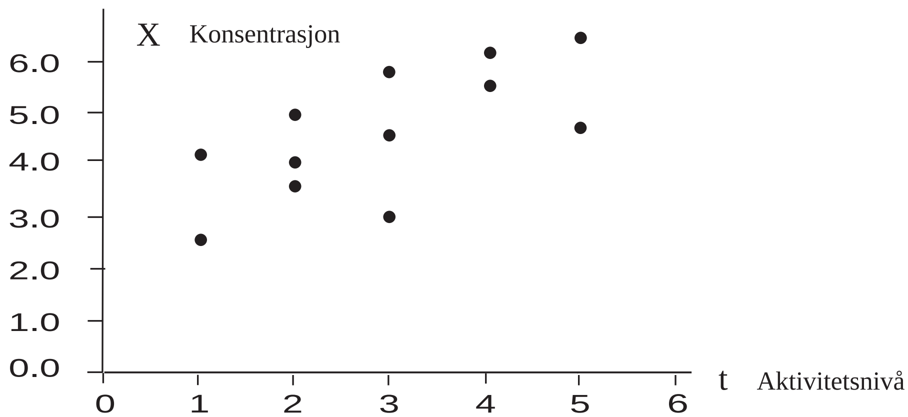
$$S(\hat{X} - X) = \sqrt{S^2 + (S(\hat{X}))^2}$$

som gir mulighet for å beregne den tredje størrelsen straks de to øvrige er kjente.

Eksempel 8 : Aktivitet - Avfall
Konsentrasjonen av et avfallsstoff i en elv måles regelmessig. I tillegg til en viss "naturligkonsentrasjon kommer avfallet fra en bedrift, og forventet bidrag fra denne antas i hvert fall ikke å avta med aktivitetsnivået. Følgende observasjoner på 12 ulike tidspunkter foreligger:

Aktivitetsnivå (t):	3	5	2	4	2	1	3	1	2	5	4	3
Konsentrasjon (X) :	4.7	6.4	3.5	5.4	3.9	2.6	5.9	4.1	5.1	4.6	6.1	3.0

Spredningsdiagrammet for observasjonene ser ut som i Figur 8.10.



Figur 8.10: Spredningsdiagram

En utskrift fra et analyseprogram for lineær regresjon gir typisk:

```
 >> READ 'eks8.8' t X
>> REGRESS X ON t
X = 2.91 + 0.584 t (S=0.9728)
                         St.Dev.  T-ratio=
 Variable  Coefficient  of Coef.  Coef/S.D      P
                2.9060    0.6810      4.27  0.002
        X       0.5837    0.2127      2.74  0.021
Degrees of Freedom for T-test = 10
R-squared = 43.0% (Adjusted for d.f. = 37.2%)
```

Vi ser at regresjonskoeffisienten på 0.5837 er såpass stor i forhold til sitt anslåtte standardavvik på 0.2127, at det er grunn til å hevde at konsentrasjonen øker med aktivitetsnivået i bedriften. Med 10 frihetsgrader blir nødvendig sikkerhetsfaktor svarende til 95 % konfidensnivå lik k=2.228, som gir konfidensintervallet $0.5837 \pm 2.228 \cdot 0.2127$, dvs. [0.110 , 1.058]. Intervallet omfatter ikke null.

Merk at de T-verdier som skrives ut svarer til nullhypotesen at koeffisienten er null, og at P-verdi er gitt for tosidig alternativ. P-verdien for situasjoner med ensidig alternativ er halvparten, i vår situasjon blir det 0.0105. Dette betyr at med valgt signifikansnivå på 5%, vil nullhypotesen bli forkastet, men ikke dersom signifikansnivået er valgt lik 1%. Dersom nullhypotesen skulle være noe annet enn null, er vi nødt til å regne ut T-brøken selv, og slå opp P-verdi i tabell.

Estimert standardavvik omkring regresjonslinjen ser vi er S = 0.9728, som gir en grov indikasjon på typiske prediksjonsfeil ved å bruke regresjonslinjen som prediktor. Utskriften inneholder også opplysninger om hvor stor del av variasjonen i konsentrasjonen av avfall som blir forklart ved variasjonen i aktivitetsnivået (R-squared). Dette blir forklart nærmere i Kapittel 12.

Et regresjonsprogram vil som regel kunne gi mer detaljerte opplysninger om bruken av $\hat{X}$ som estimator eller prediktor for gitt t :

```
>> PREDICT X for t=1 2 3 4 5
           StD     StD       Conf.Int.        Pred.Int.
t   Fit    Fit   PredErr        95%              95%
1  3.490  0.495   1.091   (2.386,4.593)  (1.057,5.922)
2  4.073  0.342   1.031   (3.311,4.835)  (1.775,6.371)
3  4.657  0.281   1.013   (4.030,5.284)  (2.400,6.914)
4  5.241  0.363   1.038   (4.431,6.050)  (2.926,7.555)
5  5.824  0.525   1.105   (4.655,6.994)  (3.361,8.288)
```

Her er gitt estimater/prediksjoner for aktivitetsnivåene 1, 2, 3, 4, 5, med tilhørende standardavvik for henholdsvis estimat og prediksjonsfeil. I tillegg gis konfidensintervaller for forventet konsentrasjon og prediksjonsintervaller for ny observerbar konsentrasjon svarende til 95% sikkerhet, som fortsatt er beregnet på grunnlag av sikkerhetsfaktoren k=2.228 for tilfellet med 10 frihetsgrader. Leseren kan selv sjekke beregningen, samt beregne tilsvarende intervaller for andre sikkerhetsnivåer.

Merk at dersom programmet ikke gir annet enn $\hat{X}_i$ og $S(\hat{X}_i)$, kan de øvrige størrelser beregnes enkelt ved formlene ovenfor.

Bruken av t-tabellen er i dette eksemplet knyttet til en regresjonsmodell med normalfordelte observasjoner. Selv om spredningsdiagrammet gir en viss pekepinn på om de ulike forutsetningene er oppfylt, kan vi ønske oss ytterligere hjelpemidler til å sjekke disse. Sentralt her er de observerte *residualene*, som er de enkelte observasjonenes vertikale avstand fra den estimerte regresjonslinjen. Vi skriver

$$\hat{U}_i = X_i - \hat{X}_i \quad i = 1, 2, \ldots, n$$

Disse kan oppfattes som "observasjoner" av de forventede avvik $U_i = X_i - EX_i$, $i = 1, 2, \ldots, n$, som, under forutsetningene i regresjonsmodellen: (i) har samme forventning lik null (ii) har samme varians (iii) er uavhengige og (iv) er normalfordelte.

Ulike former for plott av de observerte residualene vil kunne avsløre avvik i disse forutsetningene. En kan plotte residualene som funksjon av observasjonsnummeret, og de samme betraktninger som i Kapittel 8.5 kan gjøres, jfr. Figur 8.8a-f. Alternativt kan en plotte residualene mot høyresidevariablen t.

Ved vurdering av unormalt store residualer, må disse ses i forhold til sitt standardavvik. Siden S estimerer standardavviket σ til de sanne residualene, kan en se de observerte residualer i forhold til S. Det viser seg imidlertid at standardavviket til de observerte residualene avhenger av n og t, og er mindre enn σ. For ikke å undervurdere residualene mht. avvik, bør standardavviket til de observerte residualene estimeres med [9]

$$S(\hat{U}_i) = \sqrt{S^2 - (S(\hat{X}_i))^2}$$

som er tilnærmet lik S når n er stor. Grovt sagt vil en observasjon være påfallende i forhold til en normalmodell, dersom residualen i tallverdi er noe større enn 2 ganger sitt standardavvik. Det kan derfor være hensiktsmessig å beregne de såkalte *standardiserte residualer* $\hat{U}_i/S(\hat{U}_i)$. For å bekrefte rimeligheten av en normalmodell, er det også aktuelt å lage et normalskår-plott for (de standardiserte) residualer, jfr. Kapittel 8.5 og Figur 8.9a-c.

Eksempel 8 : Aktivitet - Avfall (fortsatt)
I tillegg til resultatene ovenfor bør regresjonsprogrammer kunne gi følgende opplysninger (f.eks. ved bruk av en subkommando):

[9] Til forskjell fra formelen for estimert standardavvik til prediksjonsfeilen av en ny observasjon, blir residualstandardavviket mindre når n er liten og ettersom t avviker fra sitt gjennomsnitt. Dette henger sammen med at ved få observasjoner, vil de mest avvikende i t-retningen få forholdsvis størst innflytelse på bestemmelsen av linjen $\hat{X}$, slik at tilpasningen her ofte kan bli "for god til å være sann".

```
>> FITS
Obs     t      X    Fit  StD.Fit   Resid  StD.Res  St.Res
  1  3.00  4.700  4.657    0.281   0.043    0.931    0.05
  2  5.00  6.400  5.824    0.525   0.576    0.819    0.70
  3  2.00  3.500  4.073    0.342  -0.573    0.911   -0.63
  4  4.00  5.400  5.241    0.363   0.159    0.902    0.18
  5  2.00  3.900  4.073    0.342  -0.173    0.911   -0.19
  6  1.00  2.600  3.490    0.495  -0.890    0.837   -1.06
  7  3.00  5.900  4.657    0.281   1.243    0.931    1.33
  8  1.00  4.100  3.490    0.495   0.610    0.837    0.73
  9  2.00  5.100  4.073    0.342   1.027    0.911    1.13
 10  5.00  4.600  5.824    0.525  -1.224    0.819   -1.49
 11  4.00  6.100  5.241    0.363   0.859    0.902    0.95
 12  3.00  3.000  4.657    0.281  -1.657    0.931   -1.78
```

Her er tabellert observasjonene, tilpasset verdi med standardavvik, residualene med standardavvik, samt standardiserte residualer. Vi ser at ingen av disse er så store at det gir grunn til å forkaste normalmodellen. Programvare markerer ofte avvikende observasjoner automatisk, samt observasjoner der t'en er i utkanten av observasjonsområdet, som i stor grad kan påvirke bestemmelsen av regresjonslinjen. Vær imidlertid oppmerksom på at dersom standardiserte residualer større enn 2 markeres som avvikende, vil en få ca. 5% “avvikereselv om observasjonene er normalfordelt.

8.7 ⋆Resampling

I dette kapitlet har vi sett hvordan statistiske modeller kan sette oss i stand til å vurdere usikkerheten ved estimater for parametre i modellene. De situasjoner vi har tatt opp er typiske, men mange aktuelle problemer faller likevel utenfor. Det kan være at vi er interessert i andre parametre enn gjennomsnitt og forventninger, med estimator og pålitelighet som er vanskelig å utlede. Det kan være at situasjonen krever andre, og ofte mer kompliserte modeller, der estimeringsmetoden ikke ligger like klart i dagen. Statistisk teori kan riktignok by på generelle prinsipper for konstruksjon av estimatorer, med pålitelighet som lar seg beregne tilnærmet.[10] Problemet er at hver ny situasjon ofte krever sitt, som er utenfor rekkevidde for den vanlige bruker.

Ved inngangen til 1990-årene har ny og raskere regneteknologi gitt oss nye muligheter. En av disse går under navnet *resampling*. Vi skal forklare teknikken med et forholdsvis enkelt eksempel, nemlig estimering i lotterimodellen:

La oss som før tenke oss at vi estimerer gjennomsnittet i populasjonen med gjennomsnittet i utvalget, men at vi ikke kjenner formelen for standardavviket til denne estimatoren.

Anta at vi har et utvalg på n elementer fra populasjonen på N elementer. Resampling består her i å trekke m gjentatte utvalg på n elementer med tilbakelegging fra utvalget på n elementer, og så beregne gjennomsnittet i hvert av disse utvalgene. En har da m tall, og kan beregne det empiriske standardavvik av disse, som brukes som et anslag på standardavviket til den opprinnelige estimator.

Generelt la $\hat{\theta} = \hat{\theta}(X_1, X_2, \ldots, X_n)$ være estimator for parameteren θ basert på observasjonene $X_1, X_2, \ldots, X_n$. La disse observasjonene utgjøre en populasjon hvorfra en trekker n med tilbakelegging. Dette gjentas m ganger. Vi har

$$(X_1^{(i)}, X_2^{(i)}, \ldots, X_n^{(i)}) \qquad i = 1, 2, \ldots, m$$

og beregner resampel-estimatene

$$\hat{\theta}^{(i)} = \hat{\theta}(X_1^{(i)}, X_2^{(i)}, \ldots, X_n^{(i)}) \qquad i = 1, 2, \ldots, m$$

Endelig beregnes

$$S^*(\hat{\theta}) = \sqrt{\frac{1}{m} \sum_{i=1}^{m} (\hat{\theta}^{(i)} - \hat{\theta})^2}$$

der $\hat{\theta}$ er det opprinnelige estimat, eller evt. gjennomsnittet av resampelestimatene. Under forholdsvis svake forutsetninger om modellen, vil $S^*(\hat{\theta})$ være et brukbart estimat for $\sigma(\hat{\theta})$ uansett.

Eksempel 9 : Stikkprøver - Pålitelighet

La situasjonen være som i Eksempel 1.7, der vi tok en stikkprøve på $n = 10$ tilfeldig utvalgte studenter (uten tilbakelegging) fra en populasjon på $N = 216$ studenter. Disse hadde følgende kontantbeløp på seg:

[10] Et slikt prinsipp er sannsynlighetsmaksimeringsmetoden, se Oppgave 7.38.

Beløp:	0	145	285	273	0	312	51	8	411	7

Gjennomsnittet 149.2 er vårt anslag på gjennomsnittet i hele populasjonen. Standardavviket til estimatoren blir anslått til 48.6 ut fra formlene i avsnitt 8.2. Vi gjorde isteden resampling med $m = 1000$ gjentak av utvalg med tilbakelegging av størrelse $n = 10$, og beregnet de $m = 1000$ gjennomsnitt. Vi fikk da 46.3, altså ikke altfor mye forskjellig fra anslaget ovenfor.

Anta at vi isteden var interessert i å anslå medianen i populasjonen. Dette anslår vi med medianen i utvalget, som ovenfor er 98.0. Vi har nå ikke kjennskap til formel for standardavvik til median, og bruker derfor resampling. Vi beregnet medianer i hvert av de $m = 1000$ utvalg, og deretter standardavviket til disse. Resultatet ble 94.8, som er vårt anslag på standardavviket til estimatet for medianen.

8.8 Oppgaver

1. Ta for deg en statistisk programpakke, og finn ut i hvilken grad den dekker metodene i dette kapitlet. Bruk pakken i de oppgaver der den er et tjenlig hjelpemiddel.

2. Se på situasjonen i Eksempel 1 med alkotesten.

   (a) Lag sannsynlighetsutsagn om estimeringsfeilen.

   (b) Lag et konfidensintervall (se Oppgave 3) dersom vi ønsker konfidensnivå tilnærmet lik (i) 0.90 (ii) 0.95 (iii) 0.99 .

   (c) Hvor mange flere observasjoner må vi ha for å
   (i) redusere standardavviket til estimatoren til 0.09,
   (ii) halvere lengden av konfidensintervallene.

3. Anta målemodellen der μ er ukjent og σ kjent.

   (a) Forklar at for enhver $k > 0$ vil

   $$\left[\overline{X} - k \cdot \frac{\sigma}{\sqrt{n}}, \overline{X} + k \cdot \frac{\sigma}{\sqrt{n}}\right]$$

   være et konfidensintervall for μ med konfidensnivå tilnærmet lik $A(k)$.

   (b) Finn k slik at konfidensnivået c er ca. lik (i) 0.90 (ii) 0.95 (iii) 0.99.

   (c) Hva kan sies dersom observasjonene selv er normalfordelte?

4. Vis og fortolk følgende resultater i målemodellen

   (a) $P(\mid \overline{X} - \mu \mid < k \cdot \frac{\sigma}{\sqrt{n}}) \geq 1 - \frac{1}{k^2}$ for alle $k > 0$

(b) $P(|\overline{X}-\mu| < d) \to 1$ når $n \to \infty$ $(d > 0)$.
Hint: Anvend Tsjebysjeffs ulikhet (Kapittel 6.6) på $\overline{X}$.

(c) $P(\frac{\overline{X}-\mu}{\sigma/\sqrt{n}} < z) \to G(z)$ når $n \to \infty$.
Hint: Betrakt sentralgrensesetningen (Kapittel 6.5).

(d) $P(\overline{X} \le k) \approx G(\frac{k-\mu}{\sigma/\sqrt{n}})$ for store n
Er det aktuelt med heltallskorreksjon her?

5. La situasjonen være som i Eksempel 3 med testing av forbedring.

   (a) Vis at for å lage en test med signifikansnivå ca. 0.05 og med teststyrke 0.90 for alternativet $\mu = 16.0$, så trengs det $n = 14$ observasjoner.

   (b) Hva blir da styrken i alternativet $\mu = 15.5$?

6. En bedrift produserer en type nylonsnøre, og basert på lengre tids erfaring mener man at strekkstyrken i kg til et tilfeldig snøre kan oppfattes som en stokastisk variabel med forventning $\mu = 9.40$ og standardavvik $\sigma = 0.40$. Det er hevdet at råstoff fra en annen leverandør gir bedre forventet kvalitet ($\mu > 9.40$). Det er undersøkt $n = 10$ slike snører (anta fortsatt $\sigma = 0.40$).

   (a) Formuler situasjonen som et hypotesetestingsproblem og lag en testmetode med signifikansnivå ca 5%. Hvilken styrke har alternativet $\mu = 10.0$, og hva innebærer dette?

   (b) Utfør testen og angi konklusjonen når datamaterialet er som gitt i Eksempel 2. Regn også ut P-verdien for det observerte resultat og forklar hva dette innebærer.

   (c) For dataene i Eksempel 2 ble σ estimert til 0.34. Er det grunn til å tro at $\sigma = 0.40$ var en urealistisk antakelse? Hvilken konsekvenser har det at σ er stipulert for høyt (evt. for lavt)?

7. Anta at strekkstyrken X av et tilfeldig nylonsnøre fra produksjonen er normalfordelt (se avsnitt 8.4) med forventning μ og standardavvik $\sigma = 0.40$. Finn sannsynligheten for at snøret tåler en belastning på 8 kg i tilfellene (i) $\mu = 9.40$ (råstoff 1) (ii) $\mu = 9.70$ (råstoff 2)

8. En bedrift produserer bokser av metall som ifølge spesifikasjonene skal ha diameter 20.0 cm. Det er viktig at maskinen som former og sveiser sylinderen er korrekt justert, ellers blir det problemer ved monteringen av bunnen, for liten eller for stor diameter er like lite ønskelig. Av tidligere erfaring vet man at når prosessen er riktig justert, er standardavviket σ til diameteren til en tilfeldig sylinder lik 0.1. Før en lengre produksjonsserie settes i gang prøveproduseres $n = 4$ sylindre og deres diametre måles nøyaktig. Det viste seg at $\overline{X} = 20.2$ og at estimert standardavvik samsvarte godt med tidligere erfaring.

   (a) Formuler situasjonen som et hypotesetestingsproblem. En ønsker 5% signifikansnivå.

(b) Gir det observerte resultat grunnlag for å påstå at maskinen ikke er korrekt justert?

(c) Hva er sannsynligheten med den gitte test å avsløre en maskin som gir forventet diameter 20.1 cm? Enn 19.9 cm?

(d) Hvor mange sylindre må prøveproduseres for å være 99% sikker på å avsløre maskinen i (c)?

9. Betrakt målemodellen der μ er ukjent og σ kjent.

(a) Forklar at dersom vi ønsker å teste $H_0 : \mu = \mu_0$ mot $H_A : \mu > \mu_0$, er det rimelig å forkaste H_0 når $Z \geq k$. Forklar hvordan vi bestemmer k. Vis at et tilnærmet uttrykk for styrkefunksjonen til testen er

$$\Pi(\mu) \approx G\left(\frac{\mu - \mu_0}{\sigma/\sqrt{n}} - k\right)$$

(b) Forklar at dersom vi ønsker å teste $H_0 : \mu = \mu_0$ mot $H_A : \mu \neq \mu_0$, er det rimelig å forkaste H_0 når $\mid Z \mid \geq k$. Forklar hvordan vi bestemmer k. Vis at tilnærmet uttrykk for styrkefunksjonen er

$$\Pi(\mu) = 1 - \left(G\left(\frac{\mu - \mu_0}{\sigma/\sqrt{n}} + k\right) - G\left(\frac{\mu - \mu_0}{\sigma/\sqrt{n}} - k\right)\right).$$

(c) Skisser grovt styrkefunksjonene i (a) og (b) som funksjon av $\theta = (\mu - \mu_0)/(\sigma/\sqrt{n})$

(d) Vis formlene for nødvendig stikkprøvestørrelse gitt i teksten.

10. To målinger X_1 og X_2 av samme størrelse μ antas uavhengige med forventning μ, og med varianser henholdsvis σ_1^2 og σ_2^2.

(a) Vis at alle estimatorer av form

$$T = aX_1 + (1-a)X_2$$

der a er en valgt konstant $(0 \leq a \leq 1)$, er forventningsrette for μ.

(b) Vis at $var(T) = a^2\sigma_1^2 + (1-a)^2\sigma_2^2$.

(c) Sammenlign tre slike estimatorer
$T_1 = (X_1 + X_2)/2$, $T_2 = (X_1 + 2X_2)/3$ og $T_3 = X_2$.
Hvilken av disse bør foretrekkes dersom
(i) $\sigma_1^2 = \sigma_2^2$ (ii) $\sigma_1^2 = 3\sigma_2^2$ (iii) $\sigma_1^2 = 10\sigma_2^2$.

(d) $\star$ Dersom vi kjenner σ_1^2 og σ_2^2 så er den beste estimator av formen ovenfor

$$\hat{\mu} = \frac{(1/\sigma_1^2)X_1 + (1/\sigma_2^2)X_2}{(1/\sigma_1^2) + (1/\sigma_2^2)}$$

Hint: Ulikheten i Oppgave 5.40 kan nyttes.

11. Vi vil foreta en undersøkelse blant $N = 216$ studenter i et kull for å finne ut noe om tilbøyeligheten til å bære på kontanter.

    (a) Anslå gjennomsnittsbeløpet $\overline{v}$ for alle studenter på en bestemt tirsdag på grunnlag av en stikkprøve på $n = 10$ studenter den dagen. Trekk studentene fra tabellen i Eksempel 1.7, f.eks. ved å benytte en terning. Vurder påliteligheten av anslaget.

    Anta at bakgrunnen for vår interesse var et ran som ble begått samme dag, og at vi finner stikkprøven såpass interessant at vi innhenter opplysningene for alle $N = 216$ studenter, dvs. slik de er gitt i Eksempel 1.7.

    (b) Hvor godt traff du med stikkprøven?

    Uken etter planlegges en ny stikkprøveundersøkelse for å se om studentene er blitt mer forsiktige.

    (c) Hvor stor stikkprøve trengs nå for å anslå gjennomsnittet med tilhørende standardavvik på ca 30 kroner?

    (d) Anta at det observerte gjennomsnitt i denne stikkprøven ble 98.50 kroner. Gir dette grunnlag for å påstå større forsiktighet?

12. En tømmeroppkjøper er interessert i å overta et parti tømmer bestående av $N = 1000$ stokker. For å finne ut hva hun skal tilby for partiet burde hun måle volumet av hver av stokkene $v_1, v_2, \ldots, v_N$, men hun finner at dette er for tidkrevende. Isteden velger hun ut n stokker tilfeldig og måler disse. La Y_i være volum (i m^3) av i'te målte stokk $i = 1, 2, \ldots, n$.

    (a) Finn en forventningsrett estimator for det totale volum $v = v_1 + v_2 + \cdots + v_N$. Angi estimatorens varians uttrykt ved N, n og $\sigma^2 = \sum_{i=1}^{N}(v_i - \overline{v})^2/N$.

    (b) Tømmeroppkjøperen kjøpte året før et lignende parti tømmer fra samme distrikt. Da målte hun alle stokkene og det viste seg at $\overline{v} = 0.600$ og $\sigma^2 = 0.0025$. Han ønsker estimat med standardavvik $\Delta = 3$. Hvor mange stokker bør hun i så fall måle?

    (c) Finn tilnærmet sannsynligheten for at oppkjøperens anslag for det totale volum avviker mer enn $6m^3$ fra den korrekte verdi, samt sannsynligheten for at anslaget er mer enn $3m^3$ for stort. Hint: Bruk normaltilnærmelse.

13. En bedrift produserer artikler med kvalitet som varierer rundt et nivå μ (Jfr. Eksempel 1.6). Kvaliteten av en tilfeldig artikkel oppfattes som en stokastisk variabel X med forventning $\mu = 2.90$ og standardavvik $\sigma = 0.21$. Anta normalitet.

    (a) Finn sannsynligheten for at X er
    (i) mindre enn 2.8 (ii) mindre enn 2.4.

La $\overline{X}$ være gjennomsnittlig kvalitet for 4 artikler. Anta uavhengighet (diskuter dette).

(b) Finn sannsynligheten for at $\overline{X}$ er
(i) mindre enn 2.8 (ii) mindre enn 2.4.

Diskuter antakelsen om normalitet i lys av tallmaterialet i Eksempel 1.6.

14. La $X_1, X_2, \ldots, X_n$ være kvaliteten av n artikler der forventet kvalitet μ er ukjent. Anta at X_i'ene er uavhengige normalfordelte med kjent standardavvik $\sigma = 0.21$.

(a) Finn sannsynligheten for at $\overline{X}$ avviker fra μ med
(i) høyst 0.21 (ii) høyst 0.42 (iii) høyst 0.50
i tilfellene $n = 1, 4, 9$.

(b) Hvor mange artikler må vi observere for at
(i) standardavviket til $\overline{X}$ skal bli høyst 0.05,
(ii) sannsynligheten for at $\overline{X}$ avviker høyst 0.10 fra den sanne μ blir minst 0.99.

I hvilken grad er normalitetsantakelsen nødvendig.
Hva kan sies i (a) og (b) dersom σ i virkeligheten er større (mindre) enn 0.21?
Hva kan sies dersom vi overhodet ikke vet noe om σ?.

15. Anta at den stokastiske variable T er T-fordelt med ν frihetsgrader.

(a) Bestem k slik at $P(T > k) = \alpha$ dersom
(i) ν=5, α=0.01, 0.05, 0.10
(ii) ν=10, α=0.01, 0.05, 0.10

(b) Bestem k slik at $P(\mid T \mid > k) = \alpha$ dersom
(i) ν=5, α=0.01, 0.05, 0.10
(ii) ν=10, α=0.01, 0.05, 0.10

(c) Bestem k slik at $P(T < k) = \alpha$ dersom
(i) ν=5, α=0.05, 0.975
(ii) ν=10, α=0.05, 0.975

Sammenlign svarene med det du får ved å anta at T er standardnormalfordelt.

16. Betrakt målemodellen med normalitetsantakelse der forventningen μ og variansen σ^2 er ukjent. Anta at μ estimeres med $\overline{X}$ og at det tilhørende standardavvik estimeres på vanlig måte.

(a) Finn tilnærmet sannsynligheten for en estimeringsfeil på høyst k ganger estimert standardavvik i følgende situasjoner
(i) n=2, k=1,2,3,4 (ii) n=6, k=1,2,3,4 (iii) n=15, k=1,2,3,4

Anta at vi bruker et konfidensintervall for μ av form

$$\left[\overline{X} - k \cdot \frac{S}{\sqrt{n}}, \overline{X} + k \cdot \frac{S}{\sqrt{n}}\right]$$

der konfidensnivået er c.

(b) Bestem k i følgende situasjoner
(i) n=2, c=0.90, 0.95, 0.99
(ii) n=6, c=0.90, 0.95, 0.99
(iii) n=15, c=0.90, 0.95, 0.99

Sammenlign resultatene i (a) og (b) med de tilsvarende for situasjonen at σ er kjent.

17. En bedrift framstiller et produkt med et visst proteininnhold som angis på emballasjen. Som råstoff brukes bl.a. soyamel som innblandes etter vekt. Man har erfart at konsentrasjonen av proteiner i råstoffet kan variere en del, noe som kan skyldes varierende vekstforhold eller lagringsforhold (fuktighet etc.). Bedriften har lagret et parti soyamel i sekker og ønsker å anslå den gjennomsnittlige proteinkonsentrasjonen μ i gram pr. vektenhet. For dette formål tas en prøve fra hver av n utvalgte sekker. La $X_1, X_2, \ldots, X_n$ være proteinkonsentrasjonen i gram pr. vektenhet i disse prøvene. Disse antas uavhengige normalfordelte med forventning μ og standardavvik σ. Observasjonene ble

23.7 20.6 23.2 21.3 21.9 24.0 26.3 24.8 21.2 22.7 23.6 26.8 21.5

(a) Anslå μ og σ og vurder påliteligheten av anslagene.

(b) Lag et konfidensintervall for μ med konfidensnivå 0.99.

18. La situasjonen være som i Oppgave 14. Det er produsert fire artikler, og kvaliteten av disse ble

3.1 3.4 2.9 3.0

Lag et konfidensintervall for μ med konfidensnivå lik 0.95 når (a) σ kjent lik 0.21 (b) σ ukjent.
Gir resultatene grunnlag for å påstå at μ ikke er 2.90?

19. Under normale forhold gir produksjonsprosessen i foregående oppgave artikler med forventet kvalitet $\mu = 2.90$. Ved arbeidsdagens slutt måles kvaliteten av de $n = 9$ siste artiklene. Resultatet ble

2.4 2.9 2.1 2.7 2.6 3.3 2.5 2.6 2.5

Gir dette grunn til å påstå at maskinen er "ute av stille"(dvs. $\mu < 2.90$)? Lag en ensidig test med 5% signifikansnivå.

20. Bedriften i foregående oppgave vurderer et alternativt produksjonsopplegg som muligens kan forbedre kvaliteten ($\mu > 2.90$). Det prøveproduseres $n = 9$ artikler og kvaliteten ble

3.4 2.9 3.7 3.1 3.2 2.5 3.3 3.2 3.3

Gir dette grunnlag for å påstå forbedring? Lag en ensidig test med 5% signifikansnivå.

21. Det er produsert n=9 artikler med hver av to produksjonsmetoder. Kvaliteten ble så målt for hver artikkel, og resultatet ble

Metode 1 :	2.9	2.7	3.4	2.8	2.8	3.3	2.6	3.0	2.8
Metode 2 :	2.4	2.9	2.1	2.7	2.6	3.3	2.5	2.6	2.5

Anslå forskjellen i forventet kvalitet med de to metodene, og anslå standardavviket til estimatet. Vurder om det er grunn til å anta at en av metodene gir høyere forventet kvalitet. Hint : Bruk resultatet i avsnitt 7.6.

22. En bedrift framstiller en bestemt type brød. Vektene i gram av tilfeldig valgte brød antas å være uavhengige og normalfordelte $N(\mu, \sigma^2)$. Forventningen μ vil avhenge av deigen, innstillingen av maskinen som porsjonerer ut deigen, samt steketiden. Ifølge forskriftene skal denne typen brød veie minst 750 gram. Anta at en bestemt produksjonsplan innebærer at $\mu = 760$ og $\sigma = 10$.

    (a) Hva er sannsynligheten for at et tilfeldig valgt brød er undervektig?

    (b) Anta at tre brød kjøpes. Finn følgende sannsynligheter
    (i) minst ett er undervektig
    (ii) det tyngste brødet veier mer enn 780 gram.

    (c) Hvor mange brød må kjøpes for at sannsynligheten for å få minst ett undervektig er 0.90?

    (d) Hvor stor må μ minst være for å oppnå at sannsynligheten for at et tilfeldig valgt brød er undervektig blir høyst 0.001.

23. Bedriften i forrige oppgave vurderer en alternativ produksjonsplan der hverken μ eller σ på forhånd er kjent. Det produseres et parti og $n = 4$ brød velges ut tilfeldig for veiing. Resultatet ble

    782 758 775 769

    (a) Rapporter resultatet, og lag et konfidensintervall for μ med nivå 0.95.

    For å være på den sikre siden ønsker bedriften en plan svarende til $\mu = 780$.

    (b) Utfør en T-test for hypotesen $\mu = 780$ mot alternativet $\mu < 780$ når vi ønsker et signifikansnivå på 5%.

    (c) Sier teorien ovenfor noe om styrkefunksjonen til T-testen?

24. En forsker har forut for en større undersøkelse formulert 10 hypoteser som hun ønsker å få testet på grunnlag av det innsamlede tallmateriale. For hver test bruker hun signifikansnivå 0.05. Finn sannsynligheten for at

    (a) minst en hypotese blir feilaktig forkastet.

(b) minst tre hypoteser blir feilaktig forkastet.

Hvilket signifikansnivå måtte hun bruke ved hver test for at sannsynligheten for minst en feilaktig forkastning skal være høyst 0.10. Kommenter resultatene.

25. Følgende utsagn har forekommet i forbindelse med hypotesetesting

    (a) Signifikansnivået er sannsynligheten for at nullhypotesen er riktig.
    (b) Signifikansnivået er på grunnlag av observasjonene beregnet til 0.147.
    (c) P-verdien er sannsynligheten for at nullhypotesen er riktig.
    (d) P-verdien er sannsynligheten for å forkaste en riktig nullhypotese.

    Alle disse utsagn tyder på begrepsforvirring, forklar!

26. I målemodellen kan uavhengighetsantakelsen testes ved å ta utgangspunkt i at hver observasjon har sannsynlighet 0.5 for å falle hhv. over og under medianen (jfr. Oppgave 7.37). Bruk dette til å lage en test for uavhengighet i tallmaterialet i Eksempel 1.6 og Oppgave 1.13.

27. Et slankefirma reklamerer at kvinner som er mer enn 20 kg overvektige, vil kunne gå ned minst 5 kg i løpet av en diettperiode på 4 uker. Drøft mulige presiseringer av dette utsagnet. I et forsøk med dietten viste det seg at blant 64 kvinner i denne kategorien ble vektreduksjonen i gjennomsnitt 4 kg, med et beregnet standardavvik på 2 kg. Gir dette støtte for rimelige tolkninger av reklamen.

28. En vare ønskes levert i henhold til gitte spesifikasjoner som, avhengig av varens art, kan være vanskelig å sjekke ved levering. Eksempelvis gjennomsnitt og variasjon i vekten av et parti fiskeyngel til et oppdrettsanlegg. Godtaking av et parti kan derfor tenkes skje ved veiing av en stikkpøve, og der prinsippene for veiingen er innarbeidet i leveringskontrakten. Diskuter hvordan en kunne formulere dette i en slik kontrakt.

29. Et supermarkedkjede har satt som vilkår at minst 100 enheter må i gjennomsnitt selges hver uke for at en bestemt merkevare skal være med i sortementet. I en prøveperiode er produktet satt ut i 10 utvalgte butikker, og salget viste seg å bli

Butikk nr.	1	2	3	4	5	6	7	8	9	10
Salg	109	92	113	86	124	97	131	114	106	121

    Situasjonen er blitt analysert som et hypotesetestingsproblem basert på målemodellen. Diskuter dette i følgende situasjoner:

    (a) Observasjonene er for 10 tilfeldige uker i samme butikk og beslutningen gjelder denne.
    (b) Observasjonene er for samme tilfeldige uke i 10 butikker og beslutningen gjelder for disse samlet.

(c) Observasjonene er for samme tilfeldige uke i et utvalg av 10 tilfeldige butikker i kjeden, og beslutningen gjelder samlet for alle i kjeden.

30. Gitt følgende sammenhørende verdier av en stimulusvariabel t og tre responsvariable X, Y og Z.

t:	10	8	13	9	11	14	6	4	12	7	5
X:	804	695	758	881	833	996	724	426	1084	482	568
Y:	914	814	874	877	926	810	613	310	913	726	474
Z:	746	677	1274	711	781	884	608	539	815	642	573

Besvar punktene (a) - (d) etter tur for alle tre responsvariable, og kommenter resultatene.

(a) Plott respons mot stimulus i et spredningsdiagram og vurder om den lineære regresjonsmodellen er aktuell.

(b) Beregn beskrivende mål som gjennomsnitt, standardavvik og korrelasjonskoeffisient.

(c) Finn minste kvadraters regresjonslinjen, og test om regresjonskoeffisienten er positiv. Vurder risikoen for prediksjonsfeil.

(d) Studer hvordan en residualanalyse vil belyse disse dataene.

31. Studer Eksempel 1.9. i lys av det du har lært om regresjon og prediksjon i dette kapitlet.

(a) Samsvarer formelene på tross av ulik notasjon?

(b) Diskuter forutsetningene i en evt. regresjonsmodell.

(c) Test om reklamen generelt påvirker salget.

(d) Prediker salg for annonseutgifter på
(i) kr. 2.000,- (ii) kr. 4.000,- (iii) kr. 6.000,-
Vurder risikoen for prediksjonsfeil.

32. La situasjonen være som i Eksempel 6, men anta at variansen σ^2 til observasjonene ikke er kjent, og at analysen isteden bygger på T-observatoren.

(a) Estimer σ^2 på grunnlag av observasjonene som er gitt i eksemplet.

(b) Utfør en T-test for å finne ut om den nye rensemetoden gir forbedring. Rapporter P-verdien til det observerte resultat, og forklar hva denne innebærer.

33. (a) Vis at for enhver konstant μ gjelder at

$$\sum_{i=1}^{n}(X_i-\mu)^2 = \sum_{i=1}^{n}(X_i-\overline{X})^2 + n(\overline{X}-\mu)^2.$$

Spesialiser resultatet til tilfellene $\mu = 0$ og $\mu = EX$.
Hint: Skriv $(X_i-\mu)^2 = ((X_i-\overline{X}) + (\overline{X}-\mu))^2$ kvadrer ut og summer.

(b) $\star$ Vis at for alle konstanter α og β gjelder at

$$\sum_{i=1}^{n}(X_i-\alpha-\beta(t_i-\bar{t}))^2 = \sum_{i=1}^{n}(X_i-\hat{\alpha}-\hat{\beta}(t_i-\bar{t}))^2+n(\hat{\alpha}-\alpha)^2+M(\hat{\beta}-\beta)^2$$

der $\hat{\alpha} = \overline{X}$, $\hat{\beta} = \frac{1}{M}\sum_{i=1}^{n}(t_i-\bar{t})X_i$ og $M = \sum_{i=1}^{n}(t_i-\bar{t})^2$.

34. $\star$ Betrakt lotterimodellen i avsnitt 8.2. Vis at

$$ES^2 = \frac{N}{N-1}\cdot\frac{n-1}{n}\sigma^2$$

$$\text{der} \quad S^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i-\overline{Y})^2 \quad \text{og} \quad \sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(v_i-\overline{v})^2.$$

Hint: Bruk resultatet i Oppgave 33 (a) og gjør som i tilfellet med målemodellen.

35. Betrakt regresjonsmodellen i avsnitt 8.3.

(a) Anta at minste kvadraters regresjonslinjen

$$\hat{X} = \hat{\alpha} + \hat{\beta}(t-\bar{t})$$

brukes som estimator for forventningen til X for en gitt t.
Vis at (Hint: se (d))

$$E\hat{X} = EX \qquad var\hat{X} = \sigma^2(1/n + (t-\bar{t})^2)/M$$

(b) Anta at $\hat{X}$ brukes som prediktor av ny uavhengig observasjon X for gitt t. La $U = \hat{X} - X$ være prediksjonsfeilen. Vis at

$$EU = 0 \qquad varU = \sigma^2(1 + 1/n + (t-\bar{t})^2)/M$$

(c) Drøft risikoen for prediksjonsfeil som funksjon av σ, n, t og M. Hvordan kan en gå fram for å rapportere og tolke usikkerheten i denne prognosemodellen?

(d) $\star$Vis at

$$cov(\hat{\alpha}, \hat{\beta}) = 0$$

som er en av grunnene til reformuleringen av modellen i avsnitt 8.3.

Del II

Emner for videre studier

Innledning

Annen del av denne boka inneholder supplerende stoff som kan leses etter interesse og behov. De enkelte kapitlene er, så langt det er mulig, bygd opp slik at de kan leses uavhengig av hverandre. Vi vil i første rekke ta sikte på å gi en innføring i en del statistiske problemområder av betydelig praktisk interesse. I et avsluttende kapittel vil vi ta opp analyse av beslutningsproblemer under usikkerhet.

I Kapittel 7 og 8 presenterte vi en del av de sentrale begreper i forbindelse med statistisk inferens. Som illustrasjoner studerte vi noen inferensproblemer for en del enkle men svært viktige modeller. Dette stoff utgjør det som enhver bruker av statistiske metoder bør kjenne til. Leseren skulle nå være i stand til å gjennomføre enklere statistiske resonnementer på egen hånd. Mange problemområder krever imidlertid mer raffinerte modeller. Disse vil i svært mange tilfeller være generaliseringer i ulike retninger av de modellene som er presentert ovenfor. De ervervede kunnskaper kan derfor betraktes som et utgangspunkt for videre studier i metoder som har relevans for ens eget interesseområde. De vil også være et brukbart grunnlag for kommunikasjon med en fagstatistiker.

De emner som blir tatt opp i de følgende kapitler tas ikke opp i full bredde. For å kunne gjøre dette måtte vi ha utvidet teorikunnskapene ytterligere, noe som antakelig ville stille leseren på en tålmodighetsprøve. Vi vil nøye oss med smakebiter på aktuelle problemstillinger og analysemetoder med den hensikt å gjøre leseren oppmerksom på disse. For en grundigere veiledning må vi vise til spesiallitteratur om de ulike emner. Det er også en rekke emner som faller utenfor den rammen vi har valgt her.

La oss avslutte denne innledningen med noen generelle betraktninger : Formålet med en statistisk undersøkelse vil som regel være å etablere kunnskaper om sammenhenger og årsaksforhold, og vi ønsker helst å kunne trekke generelle konklusjoner, dvs. konklusjoner som har verdi ut over dataene selv. Kanskje ønsker vi å etablere ny teori eller kanskje ønsker vi å få bekreftet eller avkreftet allerede eksisterende teori.

En effektiv analyse av statistiske data, utover det rent beskrivende, vil være avhengig av den forhåndsviten vi har om hvordan dataene er framkommet, dvs. vi bør støtte oss til en modell. Da først vil en være i stand til å vurdere påliteligheten av de analysemetoder som brukes og de konklusjoner som trekkes ved hjelp av disse. En blir da fort klar over at det ikke er likegyldig hvordan våre data er innsamlet. De bør samles inn på en slik måte at det ikke medfører altfor vanskelige tolkningsproblemer, bl.a bør man i størst mulig grad eliminere mulige feilkilder. Muligheten for å gjøre dette vil variere med det anvendelsesområde det er tale om, ytterpunktene er på den ene side laboratorieforsøk der data innsamles under veldefinerte forhold, på den annen side data som har fanget vår interesse, og der omstendighetene ved datainnsamlingen ikke er kjent i detalj.

Det er vanskelig, for ikke å si umulig, å lage en enhetlig teori som kan fange opp alle problemstillinger innen disse ytterpunkter, dvs. lage en "kokebok"med standardmetoder. Hvert praktisk problem kan ha sine særtrekk som gjør at det ikke uten videre kan presses inn i den ramme som anvises av en lærebok.

En effektiv dataanalyse vil som regel være avhengig av det teorigrunnlag som fins på vedkommende fagområde, spesielt vanskelig er situasjonen på områder der teorigrunnlaget er usikkert, diskutabelt eller mangler helt. Ofte kan data være forenlig med flere ulike teorier, og statistikk alene kan derfor ikke gi svar uten reservasjoner. "Facts from figureskan være uoppnåelig.

De fleste statistiske inferensmetoder som er tilgjengelige i litteraturen er utviklet til bruk under veldefinerte forutsetninger, og ansvar for å tenke over om disse med rimelighet kan sies å være oppfylt ligger hos brukeren. Det syndes nok mye her, spesielt ved bruk av de "standardmetodersom er lett tilgjengelige i statistisk programvare. Generelt råd : Bruk ikke metoder basert på forutsetninger som du ikke forstår rekkevidden av, søk heller assistanse av en fagstatistiker, helst allerede før data er innsamlet.

Kapittel 9

Analyse av samvariasjon

Mange statistiske undersøkelser tar sikte på å avdekke samvariasjon og mulige årsakssammenhenger mellom variable. Vi har to hovedtyper variable: *kategorivariable* og *målevariable*. I en undersøkelse som angår individer er kjønn, boregion og holdning (negativ, indifferent, positiv) alle kategorivariable, mens reaksjonsevne, alder og puls er målevariable. Ofte nyttes andre betegnelser: Kategorivariable kalles gjerne *kjennetegn* i spørreskjemaundersøkelser og *faktorer* i andre sammenhenger. I litteraturen findeles ofte variabeltypene ytterligere. Kategorivariable deles i tre typer: Variable med to kategorier (som kjønn) kalles *dikotome variable*, de med flere uordnede kategorier (som bosted) kalles *nominale variable*, de med flere ordnede kategorier (som holdning) kalles *ordinale variable*.

Vi vil i dette kapitlet ta for oss analyse av samvariasjon, først for kategorivariable og avslutningsvis for målevariable. Vi vil også diskutere farer for slutningsfeil, både når det gjelder tilfeldigheters rolle og påståtte årsakssammenhenger. For målevariable blir dette tema også bli belyst i Kapittel 12.

9.1 Kryssklassifiserte hyppigheter

Anta at vi utfører en undersøkelse der vi registrerer antall objekter eller individer som faller i hver av flere mulige kategorier, bestemt ved at observasjonene kryssklassifiseres mhp. to eller flere kjennetegn (kategorivariable).

Eksempel 1 : Kunder
I en undersøkelse omkring kundene i et varehus kan man tenke seg å studere en rekke kjennetegn for ekspederte kunder, eksempelvis kjønn, yrkes-

status, inntekt, kundeforhold, varenes kostende og betalingsform. For de ulike kjennetegn kan vi tenke oss følgende kategorier: kjønn (mann, kvinne), yrkesstatus (yrkesaktiv eller ikke), inntekt (lav, høy), kundeforhold (fast, tilfeldig), varenes kostende (lav, middels, høy), betalingsform (kontanter, kort, annen). Brukes denne listen av kategorier for hvert kjennetegn vil hver kunde kunne klassifiseres i en av $2 \cdot 2 \cdot 2 \cdot 2 \cdot 3 \cdot 3 = 144$ mulige kategorier. I praksis kan en være interessert i å bruke flere (eller færre) kategorier for noen av kjennetegnene, eksempelvis inntekt (lav, middels, høy), betalingsform (kontanter, annen). Vi kan også være interessert i flere, færre eller andre kjennetegn. Hvilke kjennetegn og kategorier som brukes, vil avhenge av formålet med undersøkelsen, samt muligheten for og kostnadene ved å innhente de nødvendige opplysninger.

Resultatet av undersøkelser med krysklassifiserte hyppigheter gis ofte i form av såkalte *kontingenstabeller*. Den enkleste situasjonen har vi når det foreligger to kjennetegn som hvert innbyr til klassifisering i to kategorier. Våre data er da gitt i form av en 2×2 tabell.

Eksempel 2 : Kortbruk
La situasjonen være som i Eksempel 1. I alt n=250 utvalgte kunder har fylt ut et spørreskjema som skal sette oss i stand til å studere eventuelle sammenhenger mellom en rekke kjennetegn. La oss først klassifisere m.h.t. kjennetegnene betalingsform og kjønn. Anta at annet betalingsmiddel enn kontanter i dette eksemplet alltid er kort, og at resultatet er i Tabell 9.1.

	Kontanter	Kort	Sum
Kvinne	85	55	140
Mann	50	60	110
Sum	135	115	250

Tabell 9.1: Data

Ønsker vi en tabell over de relative hyppigheter kan vi dividere alle tall i tabellen med antall observasjoner n=250, som gir Tabell 9.2.
Av dette kan det være fristende å konkludere at kundekretsen består av flest kvinner og at kundekretsen viser større tilbøyelighet til å bruke kontanter enn kort. Av interesse er også de linjevise relative hyppighetene i Tabell 9.3 (med to desimaler), som antyder at korttilbøyeligheten er betydelig mindre blant kvinner enn blant menn.

	Kontanter	Kort	Sum
Kvinne	0.34	0.22	0.56
Mann	0.20	0.24	0.44
Sum	0.54	0.46	1.00

Tabell 9.2: Relative hyppigheter

	Kontanter	Kort	Sum
Kvinne	0.61	0.39	1.00
Mann	0.45	0.55	1.00
Sum	0.54	0.46	1.00

Tabell 9.3: Linjevise relative hyppigheter

La oss imidlertid trå varsomt. Før vi i eksemplet framsetter generelle konklusjoner angående kundekretsen, bør vi foreta en vurdering om det observerte resultat kan skyldes tilfeldigheter og/eller om det kan være andre forhold som tilslører bildet. Til dette kreves teori.

9.2 En modell for toveis-klassifikasjoner

La oss betrakte våre observasjoner som utfall av et eksperiment, der hvert utfall er beskrevet ved to eller flere kjennetegn (faktorer), med to eller flere kategorier (nivåer) for hvert kjennetegn. Den enkleste situasjon er et såkalt 2×2 klassifikasjon, dvs. vi har to kjennetegn som begge tillater klassifisering i to kategorier, slik at eksperimentet har $2 \cdot 2 = 4$ mulige utfall. Et mulig 2×2 eksperiment vil være dersom vi i Eksempel 1 observerer en tilfeldig kunde og registrerer kjønn og betalingsform (kontanter, annen). Dersom vi registrerer kjønn, yrkesstatus, betalingsform (kontanter, kort, annen), har vi et $2 \times 2 \times 3$ eksperiment osv.

Vi vil i dette avsnittet se på en modell for toveis-klassifikasjoner. Slike situasjoner forekommer ofte i praksis, og kunnskap om denne modellen vil også kunne være til hjelp for forståelsen av mer kompliserte situasjoner, f.eks. ved analyse av delproblemer som angår bare to kjennetegn. La oss bruke følgende notasjon:

Gitt to kjennetegn e og f, der e klassifiseres i r kategorier $e_1, e_2, \ldots, e_r$ og f klassifiseres i s kategorier $f_1, f_2, \ldots, f_s$. Vi har da $r \cdot s$ mulige utfall (e_i, f_j) $i = 1, 2, \ldots r$; $j = 1, 2, \ldots s$. La p_{ij} være sannsynligheten for utfallet

	f_1	f_2	Sum
e_1	p_{11}	p_{12}	$p_{1\cdot}$
e_2	p_{21}	p_{22}	$p_{2\cdot}$
Sum	$p_{\cdot 1}$	$p_{\cdot 2}$	1

Tabell 9.4: Modell for 2×2 eksperiment

(e_i, f_j) . En slik $r \times s$ situasjon kan oversiktlig presenteres ved en toveistabell med r rader og s søyler. For 2×2 situasjonen med 4 utfall ser tabellen ut som Tabell 9.4.

La $p_{i\cdot}$; $i = 1, 2, \ldots, r$ være de marginale sannsynligheter for de r kategoriene for kjennetegnet e, og $p_{\cdot j}$; $j = 1, 2, \ldots, s$ være de marginale sannsynligheter for de s kategoriene for kjennetegnet f. I tabellen finnes disse ved å summere henholdsvis radene og søylene. Merk at prikken markerer at vi har summert p_{ij} over de fotskrifter som kan være på prikkens plass, dvs. $p_{1\cdot} = p_{11} + p_{12}$ osv.

En mulig hypotese vil være at de to kjennetegnene er uavhengige, ifølge definisjonen av uavhengighet skjer dette når

$$p_{ij} = p_{i\cdot} \cdot p_{\cdot j}$$

for alle $i = 1, 2, \ldots, r$ og $j = 1, 2, \ldots, s$, dvs. når sannsynlighetene for alle utfall i tabellen er lik produktet av de tilhørende marginale sannsynlighetene (se Kapittel 4.5 om uavhengighet og produktmodeller).

De betingede sannsynligheter for kjennetegnet f gitt at e_i observeres, er for $i = 1, 2, \ldots, r$ gitt ved

$$P(f_j \mid e_i) = P((e_i, f_j))/P(e_i) = p_{ij}/p_{i\cdot} \quad j = 1, 2, \ldots, s$$

dvs. hver rad divideres med marginaltallet til høyre. Ved uavhengighet blir $P(f_j \mid e_i) = P(f_j) = p_{\cdot j}$, slik at alle disse radbrøkene er lik marginalraden i bunnen av tabellen. Tilsvarende blir hver søyle ved divisjon av marginaltallet i bunnen av tabellen lik marginalraden til høyre. Konkret betyr det at uansett hvilken kategori vi observerer for det ene kjennetegnet, så vil alle sannsynligheter for det andre være uendret.

Dersom betingelsen for uavhengighet ikke er oppfylt, sier vi at kjennetegnene er avhengige eller viser *samvariasjon*. Samvariasjon kan anta mange ulike former. Dersom kategoriene er ordinale, dvs. kan ordnes i en naturlig rekkefølge, kan vi tale om positiv samvariasjon. Dette betyr at kategorikombinasjoner høy/høy og lav/lav gjennomgående har større sannsynlighet enn høy/lav og lav/høy for de to kjennetegn. Negativ samvariasjon betyr at kombinasjonene høy/lav og lav/høy gjennomgående har større sannsynlighet enn høy/høy og lav/lav.

Positiv samvariasjon betyr grovt sagt at det i toveistabellen er mer sannsynlighet rundt diagonalen fra nordvest til sørøst enn rundt diagonalen fra sørvest til nordøst. I 2×2 situasjonen kan alltid kjennetegnene oppfattes som ordinale (hvorfor?), og avhengighet kan uttrykkes ved det såkalte *kryssproduktforholdet* $K = p_{11}p_{22}/p_{12}p_{21}$, som ved uavhengighet er lik 1, og ved positiv (negativ) samvariasjon er større (mindre) enn 1.

La oss nå tenke oss at eksperimentet er utført n ganger uavhengig av hverandre, og at vi observerer antall ganger hvert utfall inntreffer. La

$$X_{ij} = \text{antall ganger utfallet } (e_i, f_j) \text{ observeres.}$$

Resultatene av et slikt $r \times s$ eksperiment kan oversiktlig presenteres ved en toveis kontingenstabell med r rader og s søyler. For 2×2 situasjonen med 4 utfall ser tabellen ut som Tabell 9.5.

	f_1	f_2	Sum
e_1	X_{11}	X_{12}	$X_{1\cdot}$
e_2	X_{21}	X_{22}	$X_{2\cdot}$
Sum	$X_{\cdot 1}$	$X_{\cdot 2}$	n

Tabell 9.5: Observasjoner av 2×2 eksperiment

De marginale antall i hver av kategoriene for kjennetegnet e betegner vi $X_{i\cdot}$; $i = 1, 2, \ldots, r$, og de marginale antall i hver av kategoriene for kjennetegnet f betegner vi $X_{\cdot j}$; $j = 1, 2, \ldots, s$. I tabellen finnes disse ved å summere henholdsvis radene og søylene. Prikken markerer at vi har summert over de fotskrifter som kan være på prikkens plass, i tabellen $X_{1\cdot} = X_{11}+X_{12}$ osv., slik at notasjonen samsvarer med den for sannsynlighetene.

De naturlige estimatorene for utfallssannsynlighetene i toveis-tabellen er de tilsvarende observerte relative hyppigheter, dvs.

$$\hat{p}_{ij} = X_{ij}/n$$

og for de marginale sannsynlighetene

$$\hat{p}_{i\cdot} = X_{i\cdot}/n \qquad \hat{p}_{\cdot j} = X_{\cdot j}/n$$

Av forutsetningen om n uavhengige observasjoner følger , ved å la utfallet (e_i, f_j) bety suksess, at X_{ij} er binomisk fordelt (n, p_{ij}). Dette betyr at $EX_{ij} = np_{ij}$, som viser at estimatorene ovenfor er forventningsrette.

Estimatorer for de ulike betingede sannsynligheter gir seg også, henholdsvis de linjevise og kolonnevise relative hyppigheter (jfr. Tabell 9.3).

Dersom vi antar at kjennetegnene e og f er uavhengige, vil en mer naturlig estimator for p_{ij} være

$$\check{p}_{ij} = \hat{p}_{i\cdot} \cdot \hat{p}_{\cdot j}$$

og hvis $\check{p}_{ij}$ gjennomgående ikke er mye forskjellig fra $\hat{p}_{ij}$, er det grunn til å tro at kjennetegnene er (tilnærmet) uavhengige. Dette er altså tilfelle når de relative hyppighetene (Jfr. Tabell 9.2) er tilnærmet lik produktet av de tilhørende marginale hyppighetene, dvs. når

$$\frac{X_{ij}}{n} \approx \frac{X_{i\cdot}}{n} \cdot \frac{X_{\cdot j}}{n}$$

For at situasjonen i Eksempel 2 skal beskrives ved modellen ovenfor, må vi oppfatte observasjonene for de 250 kundene som $n = 250$ uavhengige realisasjoner av et 2×2 eksperiment, med samme sannsynlighet for at hver (tilfeldig) kunde faller i de ulike kategoriene. Dette er rimelig dersom kundene utgjør et tilfelig utvalg fra en stor kundekrets. De relative hyppighetene i Tabell 9.2 vil da kunne oppfattes som estimater for de teoretiske sannsynlighetene i 2×2-modellen. Vi ser at multiplikasjonsegenskapen ovenfor langt fra er oppfylt i dette tilfellet, som igjen antyder at kjennetegnene kjønn og kortbruk kan være avhengige.

Under de forutsetningene som er nevnt ovenfor, vil den simultane fordeling til observasjonene i kontingenstabellen (Tabell 9.5) være såkalt multinomisk fordelt (jfr. Kapittel 6.8). Ytterligere egenskaper ved estimatorene og mulige aktuelle testmetoder kan derfor studeres på grunnlag av denne fordeling.

I de fleste anvendelser ønsker vi å trekke konklusjoner om en avgrenset populasjon på N elementer (ofte individer) som kan klassifiseres mhp. to kjennetegn e og f. Anta at det i populasjonen er N_{ij} individer som vil kunne klassifiseres i kategori (e_i, f_j). Andelen slike individer er N_{ij}/N. Vi antar at disse andelene er ukjente og at vi trekker et utvalg på n individer fra populasjonen med sikte på å anslå andelene, evt. si noe om samvariasjonen mellom kjennetegnene i populasjonen. Ved trekning med tilbakelegging vil en kunne betrakte hver trekning som et forsøk der vi har sannsynlighet $p_{ij} = N_{ij}/N$ for utfallet (e_i, f_j), slik at modellen ovenfor gjelder. Når slike utvalg i praksis trekkes uten tilbakelegging blir modellen en tilnærmelse, men dersom n er liten i forhold til N (som regel er dette tilfelle), er forskjellen uten betydning, og modellen blir derfor brukt også i slike situasjoner istedenfor mer kompliserte (flervariable hypergeometriske) modeller. Ofte er populasjonen ikke klart avgrenset, noe som antakelig vil være tilfelle med kundekretsen i Eksempel 2.

Eksempel 3 : Avstemming
I en undersøkelse om aktiviteten blant medlemmer i en yrkesorganisasjon er det utsendt spørreskjemaer til et tilfeldig utvalg av medlemmene og i alt n=200 utfylte skjemaer foreligger til analyse. Vi er bl.a. interessert i kjennetegnet kjønn og om vedkommende avga stemme over siste tariff-forslag eller ikke. Anta at resultatet ble som i Tabell 9.6.

	Stemte	Stemte ikke	Sum
Mann	86	34	120
Kvinne	64	16	80
Sum	150	50	200

Tabell 9.6: Observasjoner

For å undersøke om kjennetegnene kjønn og stemmetilbøyelighet med rimelighet kan sies å være uavhengige, kan vi se på Tabell 9.7 der vi har beregnet de relative hyppighetene, samt påført de tilhørende produkter av de marginale hyppighetene i parentes.

	Stemte	Stemte ikke	Sum
Mann	0.43 (0.45)	0.17 (0.15)	0.60
Kvinne	0.32 (0.30)	0.08 (0.10)	0.40
Sum	0.75	0.25	1.00

Tabell 9.7: Relative hyppigheter

Vi ser at de observerte hyppigheter ikke er mye forskjellig fra produktet av de marginale hyppighetene, og det synes lite rimelig å påstå at stemmetilbøyeligheten avhenger av kjønn (på tross av at de linjevise relative hyppighetene viser at 72% av mennene og hele 80% av kvinnene stemte). Dersom vi aksepterer tanken om uavhengige kjennetegn, vil det være naturlig på grunnlag av de observerte data å bruke tallene i parentes som sannsynligheter i en eventuell modell. Disse sannsynlighetene skal da representere sjansene for at et tilfeldig valgt medlem faller i de ulike kategoriene.

9.3 Testing av uavhengighet og mål for samvariasjon

I praksis vil to kjennetegn bli betraktet som uavhengige dersom produktet av de marginale hyppighetene ikke avviker altfor mye fra hyppighetene av hvert utfall. I motsatt fall betraktes kjennetegnene som avhengige. Dette betyr samvariasjon, men ikke nødvendigvis noe direkte årsaksforhold. Spørsmålet er imidlertid hvor store avvik vi kan tilskrive tilfeldigheter og likevel betrakte de to kjennetegnene som uavhengige. Det er vanskelig å gi noe entydig svar på dette, det vil bl.a. avhenge av hvilken bruk vi vil gjøre av resultatet. I en del situasjoner med en viss avhengighet mellom kjennetegnene, vil avhengigheten være såpass liten at det er hensiktsmessig å holde på uavhengighet, fordi dette er en enklere modell, og det ikke har praktiske konsekvenser å anta noe annet.

La oss formulere situasjonen som et hypotesetestingsproblem, der nullhypotesen er

$$H_0: \text{ Kjennetegnene } e \text{ og } f \text{ er uavhengige}$$

mot alternativet at kjennetegnene er avhengige. Mulig testobservator er:

$$Q = \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{(X_{ij} - n\hat{p}_{i\cdot}\hat{p}_{\cdot j})^2}{n\hat{p}_{i\cdot}\hat{p}_{\cdot j}}.$$

Q kan oppfattes som mål for graden av avvik mellom de observerte hyppigheter X_{ij} og forventningen for tilfellet at kjennetegnene er uavhengige $np_{ij} = np_{i\cdot}p_{\cdot j}$, som vi estimerer med $n\hat{p}_{i\cdot}\hat{p}_{\cdot j}$.

Store verdier av Q indikerer at kjennetegnene er avhengige, og spørsmålet er hvor stor Q må være for at vi forkaster nullhypotesen. Dersom situasjonen kan beskrives med toveis-modellen fra forrige avsnitt og nullhypotesen om uavhengighet er riktig, kan det vises at Q for stor n har fordeling som kan tilnærmes med kji-kvadratkurven med frihetsgradtall $\nu = (r-1)\cdot(s-1)$ (Se Kapittel 7.7).

Eksempel 4 : Vedtektsendring
En større organisasjon vil undersøke holdningen til en foreslått vedtektsendring blant medlemmene, spesielt er man interessert i om kjennetegnene alder

og synspunkt er uavhengige eller ikke. I alt er n=180 tilfeldig utvalgte medlemmer spurt om saken. Resultatene er lagret som to kolonner i en datafil 'eks9.4' med følgende kodetall for de to variablene : ALDER (1=under 30 år, 2=30 - 50 år, 3=over 50 år), SYNSPUNKT (1=negativ, 2=indifferent, 3=positiv). En analyse ga dette :

```
 >> READ 'eks9.4' 'ALDER' 'SYN'
>> TABLE 'ALDER' 'SYN' ; CHISQUARE ;
EXPECTED
           SYN 1  SYN 2  SYN 3           Total
 ALDER 1      12     11     29              52
            13.0   17.9   21.1
 ALDER 2      11     35     34              80
            20.0   27.6   32.4
 ALDER 3      22     16     10              48
            12.0   16.5   19.5
   Total      45     62     73             180
 TABLED : OBSERVED and EXPECTED COUNTS (below)
CHISQUARE (D.F. = 4) = 24.8 (P-VALUE =
0.0001)
```

Leseren kan bekrefte resultatene ved å bemerke at

$$
\begin{aligned}
Q &= (12-13.0)^2/13.0 + (11-17.9)^2/17.9 + (29-21.1)^2/21.1 \\
&+ (11-20.0)^2/20.0 + (35-27.6)^2/27.6 + (34-32.4)^2/32.4 \\
&+ (22-12.0)^2/12.0 + (16-16.5)^2/16.5 + (10-19.5)^2/19.5 \\
&= 24.8
\end{aligned}
$$

P-verdien til det observerte resultat blir (frihetsgrader $(3-1)\cdot(3-1)$=4)

$$P = P_{H_0}(Q \geq 24.8) = 0.0001$$

Merk at Tabell C.7 i Appendiks C ikke går tilstrekkelig langt, men gir at $P < 0.001$. Uansett er dette så lite at det er grunn til å påstå at holdning til vedtektsendringen viser sammenheng med alder. Merk at denne metoden ikke forteller hva denne sammenheng består i, vi må da studere dataene nærmere. Tallene indikerer imidlertid at holdningen til endringen går i negativ retning med økende alder.

Vi ser at uttrykket som ble brukt til å teste uavhengighet ovenfor har følgende form

$$Q = \sum_{k=1}^{m} \frac{(O_k - E_k)^2}{E_k}$$

der vi summerer over et antall kategorier m, og der O_k og E_k er henholdsvis observert og (anslått) forventet antall i kategori nr. k. Vi husker at den Q som ble brukt til å teste om en modell var realistisk i Kapittel 7.7 hadde samme struktur. I virkeligheten er begge spesialtilfeller av en generell teori for testing av hypoteser i såkalte multinomiske modeller (Se Kapittel 6.8).

For mange interessante hypoteser i slike modeller vil testobservatorer av formen Q ha sannsynlighetsfordeling som kan tilnærmes med kjikvadratkurven med frihetsgradtall bestemt ut fra situasjonen. Tester basert på Q kalles derfor *kjikvadrattester.* Generelt vil graden av tilnærming øke med n, men sannsynlighetene for utfall i hver enkelt kategori bør heller ikke være for liten. Som en tommelfingerregel brukes at alle E_k bør være større enn 5.

I situasjoner der antall frihetsgrader er 1, slik tilfellet er i 2×2 situasjonen, viser det seg at tilnærmelsen til kjikvadratkurven kan forbedres. Istedenfor Q ovenfor brukes

$$Q = \sum_{k=1}^{m} \frac{(\mid O_k - E_k \mid - 0.5)^2}{E_k}.$$

Dette er en heltallskorreksjon analog med den vi bruker i normaltilnærmelsen, i litteraturen kalles dette *Yates korreksjon.* Slik korreksjon har ingen betydning og kan sløyfes når $n > 50$.

I teorien ovenfor har vi forutsatt at n var et gitt antall, mens alle de andre tallene i den observerte tabell var tilfeldig bestemt. I mange situasjoner er den ene marginalen gitt på forhånd. Det gjelder f.eks. dersom vi i Eksempel 2 bestemte oss for å spørre ut 140 kvinner og 110 menn istedenfor å velge 250 kunder tilfeldig. Situasjonen kan da betraktes som inferens om forskjeller mellom sannsynligheter i to binomiske forsøksrekker, en teori som ble utviklet i Kapittel 7.6. Testen som ble omtalt der er imidlertid likeverdig med kjikvadrattesten, som derfor også kan brukes til å teste hypotesen om at to sannsynligheter for to grupper observasjoner er like.

Merk at kjikvadrattesten kan også brukes når det er tale om sannsynligheter for mer enn to kategorier for mer enn to grupper, f.eks. dersom vi i Eksempel 4 bestemte oss for å spørre et gitt antall i hver aldersgruppe

Siste kjøp før kampanje	Siste kjøp etter kampanje Vårt merke (1)	Annet merke (2)	Sum
Vårt merke (1)	100	35	135
Annet merke (2)	65	800	865
Sum	165	835	1000

Tabell 9.8: Merkebytte

(f.eks. 60 i hver), for så å teste om sannsynlighetene for de ulike holdninger var like i de tre aldersgruppene.

Analyse av situasjoner der alle marginalene i en 2×2 tabell er gitte, blir tatt opp i Kapittel 10.3 i tilknytning til randomiserte eksperimenter. I slike situasjoner kan vi begrunne enda en test som alternativ til kjikvadrattesten i 2×2 situasjonen.

I praksis studeres ofte endringer i holdninger eller adferd fra før og etter et tiltak, og det er spørsmål om tiltaket virket.

Eksempel 5 : Merkebytte
La oss anta at det er gjennomført en reklamekampanje for en merkevare, og at data foreligger fra et forbrukerpanel på n=1000 personer som er spurt før og etter kampanjen (Tabell 9.8).

Det vil alltid være en viss overgang mellom merker. At kampanjen ikke har noen virkning, kan oppfattes slik at sjansen for at en tilfeldig person bytter fra vårt merke, er den samme som å bytte til vårt merke. Med vår tidligere notasjon for 2×2-tabeller, er nullhypotesen $p_{12} = p_{21}$, mens alternativet $p_{12} \neq p_{21}$ betyr endring, med $p_{21} > p_{12}$ i favør av vårt merke.

En egnet testmetode for denne situasjonen, *McNemars test*, er basert på testobservatoren [1]

$$Q = \frac{(\mid X_{12} - X_{21} \mid -1)^2}{X_{12} + X_{21}}$$

Denne er tilnærmet kjikvadratfordelt med 1 frihetsgrad når nullhypotesen er riktig. Nullhypotesen forkastes når Q er stor, og våre data gir

$$Q = \frac{(\mid 35 - 65 \mid -1)^2}{35 + 65} = 8.41$$

Dette gir $P = P_H(Q \geq 8.41) = 0.004$, som viser at nullhypotesen forkastes selv med forholdsvis lavt signifikansnivå. Vi kan derfor med rimelighet anta

[1] Her er tallet 1 en korreksjon av samme type som Yates korreksjon.

at den overvekt av overganger til vårt merke som er tilstede i våre data, ikke beror på tilfeldigheter. Om det er en varig effekt er en annen sak.

Det er ofte ønskelig å kunne beregne et mål for styrken og om mulig også retningen (positiv eller negativ) av en observert samvariasjon mellom to kjennetegn. Dette vil ofte være tilfellet i større undersøkelser der det ikke er plass til å gjengi alle krysstabeller fullt ut. En kan også ønske å sammenligne graden av samvariasjon i flere slike tabeller. Videre kan det hende at selv om en samvariasjon er statistisk signifikant, dvs. hypotesen om uavhengighet er forkastet, så er likevel samvariasjonen uten praktisk betydning. Eksempelvis kan vi ha et ønske å predikere et kjennetegn på grunnlag av observasjon av det andre kjennetegnet, og det kan kreve en høy grad av samvariasjon for å være meningsfylt. Merk forøvrig at det er først når vi har med ordinale kjennetegn å gjøre, at det er aktuelt å tale om en retning på samvariasjonen.

Det viser seg at det ikke fins noe allment akseptert mål for samvariasjon mellom kjennetegn, selv ikke i den enkle 2×2 situasjonen. Dersom vi i en gitt situasjon ønsker å beregne graden av samvariasjon, står vi derfor overfor et valg mellom flere ulike mål, som hver kan ha sine fortrinn når det gjelder å uttrykke spesielle typer samvariasjon. Et mål som er mye brukt for 2×2-tabeller er:

$$G = \frac{X_{11}X_{22} - X_{12}X_{21}}{X_{11}X_{22} + X_{12}X_{21}}$$

Målet kalles ofte for *(den empiriske) Gamma-koffisienten.* For en teoretisk motivering se Oppgave 11. Felles for dette og en del andre mål er at det antar verdier mellom -1 og $+1$. En verdi nær 0 indikerer at de to kjennetegnene er uavhengige (eller nær uavhengige). Jo større tallverdien av G er, desto større er graden av samvariasjon. Positiv verdi indikerer positiv samvariasjon, negativ verdi indikerer negativ samvariasjon. Vi ser fortegnet til G er bestemt av telleren. Denne er positiv når $X_{11}X_{22}$ er større enn $X_{12}X_{21}$ som i en viss forstand betyr overvekt av observasjoner på diagonalen fra nordvest til sørøst. Dette indikerer at kategoriene for de to kjennetegnene har tendens til å forekomme oftest med samme indeks. Omvendt vil $X_{11}X_{22}$ mindre enn $X_{12}X_{21}$ bety overvekt av observasjoner langs den andre diagonalen fra sørvest til nordøst og indikere at kjennetegnene har tendens til å forekomme oftest med motsatt indeks.

Eksempel 6 : Samvariasjon
Anvendes det foreslåtte mål på samvariasjon på tallene i Eksempel 2 får vi

$$G = \frac{85 \cdot 60 - 50 \cdot 55}{85 \cdot 60 + 50 \cdot 55} = 0.299$$

dvs. en moderat positiv samvariasjon.

Målet G for samvariasjon i 2×2-tabeller lar seg lett generalisere til toveistabeller med ordinale kjennetegn med mer enn to kategorier. Et annet mål for samvariasjon i en 2×2-tabell er:

$$R = \frac{X_{11}X_{22} - X_{12}X_{21}}{\sqrt{X_{1\cdot} \cdot X_{2\cdot} \cdot X_{\cdot 1} \cdot X_{\cdot 2}}}$$

Det kan vises at den empiriske korrelasjonskoeffisienten mellom de to variable en får ved å innføre indikatorvariable for de to kjennetegnene nettopp er R. [2] Følgelig vil også R anta verdier mellom -1 og $+1$, og med samme fortolkning som G ovenfor (merk at telleren er den samme for begge mål).

Mange mener at R er mindre egnet som mål for samvariasjon av indikatorvariable, bl.a. fordi variasjonsområdet for R innsnevres ettersom marginalfordelingene til de to variable avviker mer og mer. Det ligger derfor en viss fare i å rapportere R alene, idet liten R-verdi ikke nødvendigvis betyr at kjennetegnene er så godt som uavhengige, men kan reflektere det faktum at marginalene er svært avvikende (kan sjekkes dersom originaltabellen foreligger). Forøvrig kan det vises at det i 2×2-situasjonen er en enkel sammenheng mellom R og kjikvadratobservatoren. Vi har nemlig $Q = nR^2$ (se Oppgave 13). Ikke-forkasting av hypotesen om uavhengighet (liten Q) betyr som kjent ikke at vi har bevist at kjennetegnene er uavhengige, men snarere at det ikke er tilstrekkelig grunnlag til å påstå det motsatte. I denne forbindelse vil svært avvikende marginaler gi liten informasjon om eventuell avhengighet.

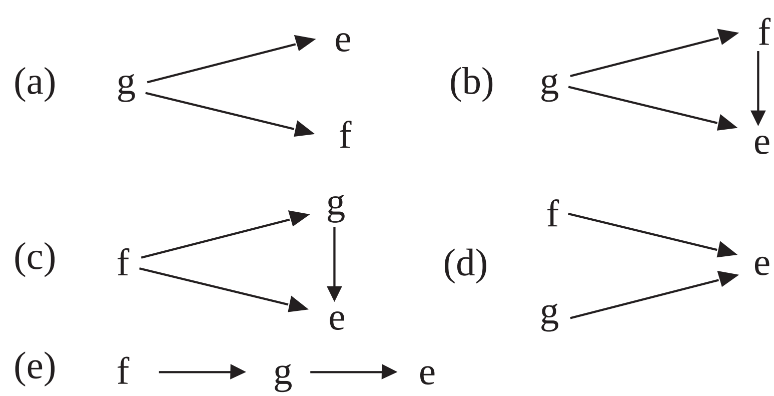
9.4 Tolking av årsaksforhold

Når vi kryssklassifiserer data mhp. to eller flere kjennetegn kan statistisk teori hjelpe oss med å påvise eventuell samvariasjon mellom kjennetegnene. Ofte ønsker vi noe mer, nemlig å uttale oss om årsaksforhold. Statistisk teori kan imidlertid ikke alene beviseet bestemt årsaksforhold. Til dette kreves også innsikt i det fagfelt som dataene er hentet fra. Statistisk metode kan imidlertid bidra med å gjøre kjent typiske feller.

[2] For definisjon av empirisk korrelasjonskoeffisient, se Kapittel 1.3 og Kapittel 9.5.

I praksis må data ofte innhentes ved spørreskjemaer som returneres i varierende grad. Idéelt sett burde antall observasjoner n være et fast tall valgt på forhånd, og ikke underlagt tilfeldigheter. I de fleste situasjoner kan man likevel betrakte n som fast tall i analysen. Man må likevel ha klart for seg hva et såkalt *bortfall* innebærer. I Eksempel 3 kan det tenkes at det å stemme og å sende inn spørreskjemaer er kjennetegn som viser positivt samsvar (ansvarsbevissthet), slik at personer som stemte er overrepresentert blant de som sendte inn skjemaer. Dette vil man eventuelt kunne sjekke ved å sammenligne med stemmeprosenten ved selve avstemningen. Imidlertid er det vel liten grunn til å tro at tilbøyeligheten til å sende inn skjemaer adskiller seg vesentlig for kvinner og menn, slik at dette forhold ikke er avgjørende for den analysen vi har gjort. I andre situasjoner kan det imidlertid være umulig å gjette seg til konsekvensene av et eventuelt bortfall slik at det ikke er mulig å trekke generelle konklusjoner ut fra observasjonene. [3]

Ved tolking av en observert samvariasjon mellom to kjennetegn, bør en ha et våkent øye for alternative kjennetegn som kan forklare det observerte resultat. Konklusjoner basert på sammenligning av to kjennetegn alene kan lett bli misvisende, især dersom en forsøker å uttale seg om årsaksforhold. Allerede når vi utvider diskusjonen til tre kjennetegn blir tolkningsmulighetene mange.



Figur 9.1: Ulike årsaksforhold

For å klarlegge noen muligheter kan vi se på Figur 9.1, som illustrerer hvordan et tredje kjennetegn g kan påvirke sammenhengen mellom to

[3]I undersøkelser som omfatter brede lag av folket, viser det seg ofte at kvinner er bedre respondenter enn menn, spesielt er eldre kvinner på landet bedre enn yngre menn i byen.

kjennetegn e og f. Pilene indikerer årsaksforhold som har en bestemt retning. Eksempelvis illustrerer Figur 9.1a den mulighet at kjennetegn g påvirker begge kjennetegnene e og f, mens disse ikke påvirker hverandre innbyrdes. Dette betyr at en observert samvariasjon mellom e og f ikke gir uttrykk for noen årsakssammenheng, den kan forklares ved den "utenforliggendefaktor g. I Figur 9.1b har vi i tillegg den muligheten at kjennetegn g også påvirker e via f, og hvor også f muligens har en egenvirkning på e. Tolk selv Figur 9.1c-e.

Dersom man, ut fra observasjoner, har funnet samvariasjon (eller mangel på samvariasjon) mellom e og f, bør man være forsiktig med å trekke generelle konklusjoner ut fra dette alene. Ved innføring av et tredje kjennetegn g kan situasjonen fortone seg noe annerledes. La oss se på en konkret problemstilling:

Eksempel 7 : Kortkort

	Konfeksjon			Elektro		
	Kont.	Kort	Sum	Kont.	Kort	Sum
Kvinne	70	40	110	15	15	30
Mann	25	25	50	25	35	60
Sum	95	65	160	40	50	90

Tabell 9.9: Kortbruk-data

	Konfeksjon	Elektro	Sum
Kvinne	110	30	140
Mann	50	60	110
Sum	160	90	250

Tabell 9.10: Kortbruk-marginale data

La situasjonen være som i Eksempel 2. Av Tabell 9.3 ser vi at tilbøyeligheten til å bruke kort er betydelig mindre blant kvinner enn blant menn. Gir dette grunn til å påstå at kvinner er mindre "kortbevissteenn menn? Neppe! Dette innser vi raskt dersom vi bringer andre kjennetegn inn i diskusjonen. Varehuset består av to avdelinger, konfeksjon og elektrisk utstyr. La oss tenke oss at tallmaterialet er klassifisert mhp. de tre kjennetegnene betalingsform (e), kjønn (f) og avdeling (g). Anta at de nødvendige data er gitt i Tabell 9.9.

	Konfeksjon	Elektro	Sum
Kontant	95	40	135
Kort	65	50	115
Sum	160	90	250

Tabell 9.11: Kortbruk-marginale data

[4] Ut fra denne tabell får vi to nye 2×2 tabeller, nemlig Tabell 9.10 og 9.11 i tillegg til Tabell 9.1.

Disse resultatene rommer mange tolkningsmuligheter. Mest påfallende er at kjennetegnene kjønn og avdeling viser betydelig grad av samvariasjon (Tabell 9.10). Vurderer vi så samvariasjonen mellom kjennetegnene kjønn og betalingsmåte innen hver avdeling (Tabell 9.9), ser vi at kjennetegnet avdeling i noen grad kan forklare den samvariasjon vi observerte i Tabell 9.1. Følgende forklaring er mulig: I konfeksjonsavdelingen handler mesteparten av kundene for mindre beløp hvor tilbøyeligheten til å bruke kort er liten, mens elektroavdelingen selger hovedsakelig større forbruksgjenstander hvor kortbruk er mer vanlig. I en situasjon hvor kvinnene viser større tilbøyelighet til å være kunde i konfeksjonsavdelingen enn i elektroavdelingen vil Tabell 9.1 alene ikke være særlig informativ. De observerte data kan forklares ut fra årsaksskjemaet gitt i Figur 9.1c, der kjennetegnet kvinne (f) påvirker betalingsform (e) via avdeling (g), men trolig også direkte. Det siste er altså en effekt som er til stede selv etter at vi har tatt hensyn til forskjeller som skyldes avdeling. Betyr en slik effekt at kvinner likevel må sies å være mindre kortbevisste enn menn? Ikke ubetinget! Ytterligere innsikt kan vi oppnå ved å utvide diskusjonen til enda flere kjennetegn, mest interessant er kanskje yrkesstatus. Hvis flere menn enn kvinner er yrkesaktive, og disse får lønn via lønnskonto i bank (og dermed kort opp i hendene), vil dette kunne forklare en stor del av forskjellen mellom mann og kvinne. Årsakskjeden er kanskje som i Figur 9.1e hvor kjennetegnet yrkesstatus er g. Et utsagn om at kvinner, fordi de er kvinner, viser mindre tilbøyelighet til kortbruk enn menn, vil imidlertid lett kunne misforstås. Dersom en sammenlignet kategoriene yrkeskvinne og yrkesmann m.h.t. kortbruk, kan det tenkes at det ikke er noen forskjell i det hele tatt, ja endog at sammenhengen går i motsatt retning, dvs. yrkeskvinnene er mer “kortbevissteenn mennene. For ytterligere spekulasjoner se Oppgave 4.

[4] Ved å legge Elektrotabellen over Konfeksjonstabellen får vi en tredimensjonal $2 \times 2 \times 2$ tabell.

Ved analyse av samvariasjon mellom to kjennetegn og innføring av et tredje kan man oppleve følgende:

- En observert samvariasjon mellom to kjennetegn kan "forklares bort" ved innføring av et tredje kjennetegn.
- En observert samvariasjon mellom to kjennetegn er fremdeles til stede etter at et tredje kjennetegn forklarer en del av samvariasjonen mellom de to første.
- Det er samvariasjon mellom de to kjennetegn også etter at et tredje kjennetegn innføres, men samvariasjon går nå i motsatt retning.
- To kjennetegn som vurdert for seg synes uavhengige, viser samvariasjon dersom man tolker observasjoner i relasjon til et tredje kjennetegn.

La oss gi eksempler på hver av disse, for korthets skyld utelater vi konkrete tall.

Eksempel 8 : Storker og fødsler
I Danmark observerte man før i tiden at hyppigheten av storker var større i områder med mange barn enn i områder med få barn. Få vil likevel påstå at det var noe årsaksforhold mellom kjennetegn barn (e) og stork (f). En utenforliggende faktor kan være folketetthet (g). En stor folketetthet medfører på den ene siden mange barn, på den andre siden mange hus med piper som gir plass for hekkende stork. Dette tankemønster svarer til Figur 9.1a. Folketetthet er muligens en forklarende årsak.

Eksempel 9 : Røking og lungekreft
Det er observert positiv samvariasjon mellom lungekreft (e) og røking (f). Dersom man innfører urbaniseringsgrad som et tredje kjennetegn (g), viser det seg at hyppigheten av lungekreft er større blant røkere enn ikke-røkere både i områder med lav og høy urbanisering. Videre viser det seg at urbane røkere viser høyere tendens til lungekreft enn ikke-urbane røkere. Det samme er tilfelle for ikke-røkere. Dette synes å indikere at både kjennetegnet røking og urbanisering kan påvise lungekreft. Det observerte resultat kan muligens forklares ved Figur 9.1d eller kanskje Figur 9.1b. En av grunnene til at det gikk lenge før røking ble allment akseptert som en vanlig årsak til lungekreft var diskusjonen om hvorvidt det ikke fantes andre utløsende faktorer som kunne forklare den påviste samvariasjon.

Eksempel 10 : Vær og avling
I et tallmateriale hvor kjennetegnene avling (e), og temperatur (f) og nedbør (g) er observert i suksessive år (kategorier lav, høy) fant man en negativ samvariasjon mellom temperatur og avling. Dette var noe overraskende idet man forestilte seg at varm vekstperiode var godt for årsveksten. Nå viste det seg at nedbør og avling hadde positiv samvariasjon, mens nedbør og temperatur hadde negativ samvariasjon. Disse observasjoner kan forklares ved årsakskjeden i Figur 9.1c, slik at høy temperatur har positiv direkte effekt på avling, men høy temperatur hører som regel sammen med lite regn, som i sin tur gir dårlig avling. Den observerte negative samvariasjon mellom temperatur og avling kan skyldes at denne indirekte effekten av temperatur er større enn den direkte. Ser man på årene med gjennomgående høy og lav nedbør hver for seg, viste det seg at temperatur og avling i begge tilfeller hadde positiv samvariasjon, altså for en gitt nedbørmengde er høy temperatur av det gode. Faktum synes å være at i naturen er de to faktorene nedbør og temperatur konkurrerende, mens de ikke vil være det i et drivhus, hvor vi for å oppnå høy avling både har høy temperatur og høy fuktighet.

Eksempel 11 : Alder og musikkinteresse
I et tallmateriale er kjennetegnene interesse for klassisk musikk (e) og alder (f) observert for et visst antall personer (kategorier lav, høy). Det viste seg at samvariasjon var såpass liten at vi er tilbøyelig til å si at alder ikke har noe å si for interessen. Dette kan imidlertid tilsløre interessante forhold. La oss i tillegg studere kjennetegnet utdanning (g) (kategori lav, høy). Da viste det seg at dersom en betraktet personene med "lav"og "høy"utdanning hver for seg, så hadde interesse og alder negativ samvariasjon for den første kategorien, positiv samvariasjon for den andre. Når de to kategoriene slås sammen har disse effektene oppveiet hverandre og forledet om til å tro at alder ikke påvirker interesse for klassisk musikk. Det er derfor likevel mulig at det foreligger en årsakskjede som den i Figur 9.1d.

Vi har med de foregående eksempler ønsket å vise at tilsynelatende enkle datamaterialer kan gi rom for flere tolkningsmuligheter, og at det foreligger betydelig fare for feiltolkninger dersom man studerer samvariasjon mellom to kjennetegn isolert. Det samme gjelder for parvis sammenligning av en rekke kjennetegn. Ideelt sett burde vi studere alle de kjennetegn vi mener har betydning for problemet i sammenheng. Vi skjønner imidlertid at med fire kjennetegn eller mer blir tolkningsmulighetene svært mange og analysen av dataene blir, dersom vi skal gå fram som i Eksempel 7, snart uoversiktlig.

Et tjenlig verktøy for analyse av kategoridata er basert på såkalte *logli-*

neære modeller. Slike gir mulighet for analyse på en systematisk oversiktlig måte, slik at vi best mulig er i stand til å fange opp særtrekk i datamaterialet, samt å avgjøre om disse kan skyldes tilfeldigheter eller ikke.

9.5 Samvariasjon av målevariable

En problemstilling av betydelig praktisk og prinsipiell interesse er: Det foreligger en populasjon, og man er interessert i sammenhengen mellom to variable i denne populasjonen. Det er ofte ikke hensiktsmessig å observere sammenhørende verdier for alle elementene i populasjonen. I steden trekkes et utvalg. En kan da foreta korrelasjonsberegninger for observasjonene i utvalget og spørsmålet er i hvilken grad disse kan danne utgangspunkt for generelle konklusjoner om korrelasjonen mellom de to variable i hele populasjonen. La oss nevne konkrete eksempler:

En økonom ønsker å studere sammenhengen mellom inntekt og sparing i en bestemt yrkesgruppe (f.eks. registrerte revisorer) på grunnlag av et utvalg fra yrkesgruppen. En psykolog ønsker å studere i hvilken grad to tester (f.eks. en skriftlig intelligenstest og en praktisk test) gjennomgående måler det samme. Hun anvender testen på et utvalg av elever.

I det første eksemplet har vi en avgrenset populasjon der inntekt og sparing kan tallfestes for alle personene i populasjonen, men vi kjenner tallene bare for personene i utvalget. I det andre eksemplet er populasjonen ikke avgrenset på samme vis, men kan i en viss forstand oppfattes som alle de som disse testene kan tenkes anvendt på. Selv om det aldri vil kunne foreligge tall fra tester av alle disse, er det fruktbart å tenke seg en "sannsimultan testskår fordeling. Denne fordeling blir ofte tenkt på som en sannsynlighetsfordeling, den angir sjansen for at en tilfeldig utvalgt person har en bestemt kombinasjon av testskårer på de to testene. Hvilke konklusjoner kan vi trekke om korrelasjonen i denne teoretiske fordelingen på grunnlag av den observerte korrelasjonen mellom testskårene for personene i utvalget? Vi er interesserte i dette spørsmålet fordi den teoretiske korrelasjonen gir uttrykk for den samvariasjon vi kan vente oss ved mer almen anvendelse av testene.

Vi tenker oss at (X, Y) er et stokastisk variabelpar med en simultan sannsynlighetsfordeling, der samvariasjonen uttrykkes ved korrelasjonskoeffisienten $\rho = \rho(X, Y)$ (se Kapittel 5.6). På grunnlag av n uavhengige observasjoner fra denne fordelingen

$$(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$$

beregner vi den empiriske korrelasjonskoeffisienten $R = R_{XY}$ gitt ved (se

Eksempel 1.8).

$$R_{XY} = \frac{S_{XY}}{S_X \cdot S_Y} = \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}) \cdot (Y_i - \overline{Y})}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2} \cdot \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \overline{Y})^2}}$$

Vi vil kunne bruke R som en estimator for ρ, og som testobservator til å teste hypoteser om ρ. Spesielt ønsker vi å teste hypotesen om ukorrelerte variable, dvs.

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

Forat ikke tilfeldigheter skal spille oss et puss, bestemmer vi en kritisk verdi k slik at vi påstår H_A først når $\mid R \mid \geq k$. Ønsker vi at testen skal ha et bestemt signifikansnivå, trenger vi å vite sannsynlighetsfordelingen til R når H_0 er riktig. Det er utarbeidet tabeller for bestemmelse av k (avhengig av n) for gitte signifikansnivåer. Disse forutsetter at den simultane sannsynlighetsfordeling til (X, Y) er såkalt *binormal*. Dette betyr at både X og Y er normalfordelte og at også alle betingede fordelinger for Y gitt X (og X gitt Y) er normalfordelte. I en slik fordeling medfører $\rho = 0$ at X og Y er uavhengige, noe som ikke gjelder generelt (se Kapittel 5.6). Under denne forutsetningen kan vi isteden ta utgangspunkt i

$$T = \frac{R}{\sqrt{1 - R^2}}\sqrt{n-2}$$

som, dersom H_0 er riktig, er t-fordelt med $n - 2$ frihetsgrader, en fordeling vi allerede har tabeller for.

Eksempel 12 : To tester
Det observeres en score på to tester; en for verbal forståelse (X) og en for analytisk evne (Y). Anta at disse er observert for 9 individer:

Individ (i) :	1	2	3	4	5	6	7	8	9
Score (X_i) :	28	33	37	31	32	26	35	36	30
Score (Y_i) :	30	23	34	39	33	28	29	38	24

Her blir korrelasjonskoeffisienten R=0.345, dvs. moderat positiv (lineær) samvariasjon (tegn spredningsdiagram!) Vi ønsker å teste om disse data gir grunnlag for å påstå at korrelasjonen er ulik null i den populasjon som individene er trukket fra, evt. alle som testen er tenkt brukt på. Vi får

$$T = \frac{0.345}{\sqrt{1 - 0.345^2}}\sqrt{9-2} = 0.97$$

som langt fra er tilstrekkelig til å påstå at $\rho \neq 0$ i den tenkte bivariate populasjonsfordelingen. t-tabellen med 7 frihetsgrader krever kritisk verdi lik 2.365 for en tosidig test med 5% signifikansnivå. En $R = 0.345$ vil være signifikant ulik null først når antall observasjoner er 33 eller flere.

Merknad. Dersom (X, Y) er binormalt fordelt gjelder at

$$ER \approx \rho \qquad \sigma(R) \approx \frac{1}{\sqrt{n-2}}$$

et resultat som kan brukes til å lage tilnærmede konfidensintervaller for ρ.

Avslutningsvis gir vi et par reservasjoner mot ukritisk bruk av korrelasjonskoeffisienter: I utgangspunktet måler R bare graden av lineær samvariasjon for observasjonene. Generalisering til den underliggende fordeling (populasjon) krever antakelser (binormalitet) som er vanskelig å begripe, enn si sjekke, og som derfor ofte overses. I mange situasjoner vil en være tjent med å bruke en såkalt *rangkorrelasjonskoeffisient,* der observasjonene er erstattet med ranger. En slik fanger opp monoton ikke-lineær samvariasjon, og hypotesetesting kan skje med mindre restriktive forutsetninger enn binormalitet. For videreføring av dette tema må vi vise til annen litteratur.

I situasjoner der en har en rekke variable hvor en på forhånd har liten kunnskap om mulige sammenhenger, er det lett å ty til korrelasjonsberegninger på observasjonene for å skaffe seg innsikt, eventuelt avsløre årsakssammenhenger. Dette er en praksis som ikke bør følges uten en viss innsikt i de mulige feiltolkninger som er tilstede: Korrelasjonsberegninger kan aldri bevise årsakssammenhenger mellom to eller flere variable. Observerte korrelasjoner mellom to variable kan endre karakter når de studeres i lys av andre variable (partiell korrelasjon). Her gjelder de samme reservasjoner som ved slutten av forrige avsnitt.

9.6 Oppgaver

1. Vis at i modellen for et 2×2 eksperiment betyr uavhengighet at

$$\begin{aligned} p_{11}/(p_{11}+p_{12}) &= p_{21}/(p_{21}+p_{22}) \\ p_{11}/(p_{11}+p_{21}) &= p_{12}/(p_{12}+p_{22}) \end{aligned}$$

Gi konkret fortolkning av hver likhet og vis at begge er ekvivalent med

$$p_{11}p_{22}/p_{12}p_{21} = 1$$

Forklar at dersom vi erstatter likhetene med ulikheten større enn (mindre enn), så svarer dette til positiv (negativ) samvariasjon.

2. Test hypotesen om uavhengighet mellom kjønn og stemmetilbøyelighet i Eksempel 3 ved å bruke den beskrevne kjikvadratmetoden.

3. (a) Test hypotesen om uavhengighet i de tre marginale 2×2 tabellene i Eksempel 7. (Tabellene 9.1, 9.10 og 9.11).

   (b) Test hypotesen om uavhengighet mellom kjønn og betalingsmåte innen hver avdeling (Tabell 9.9).

4. Diskuter hvorvidt kunnskaper om følgende forhold kan ha interesse for å vurdere tallene i Eksempel 7 om kortbruk:

   (a) varenes kostende.

   (b) mann og hustru handler gjennomgående sammen i Elektroavdelingen.

   (c) menn handler gjennomgående stort i Konfeksjonsavdelingen.

5. En bedrift har i lengre tid produsert en bestemt artikkel, men har den senere tid erfart et økende antall reklamasjoner. I et første forsøk på å skaffe seg innsikt i forholdet har man tatt for seg de $n = 100$ siste reklamasjonene. Disse blir klassifisert etter to kjennetegn reklamasjonstype A, B eller C og produksjonskjede 1, 2 eller 3. Resultatet ble

	Type A	Type B	Type C	Sum
Kjede 1	29	8	9	46
Kjede 2	9	4	6	19
Kjede 3	10	8	17	35
Sum	48	20	32	100

   (a) Test hypotesen om at reklamasjonstype og produksjonskjede er uavhengige kjennetegn. Bruk 5% signifikansnivå.

   (b) Dersom hypotesen forkastes, forsøk å tolke tallene så langt det synes rimelig.

6. To bedrifter produserer samme artikkel som de leverer under hvert sitt varemerke. Begge bedrifter leverer tre kvalitetssorteringer A, B og C, men produksjonen foregår etter to ulike metoder. Bedriftene vurderer nå å markedsføre varen under samme varemerke, men før dette gjøres vil en gjerne ha rede på om de to bedriftene leverer varer med ulik kvalitetsfordeling (hvorfor?). For å klargjøre dette har man valgt ut 50 tilfeldige artikler fra hver bedrift som blir kvalitetssortert. Resultatet ble:

	Fabrikk 1	Fabrikk 2	Sum
Kvalitet A	20	12	32
Kvalitet B	19	31	50
Kvalitet C	11	7	18
Sum	50	50	100

Gir dette materialet grunnlag for å påstå at de to bedrifter produserer ulik kvalitetsfordeling? Velg 5% signifikansnivå.

7. Ved en eksamen på en høyskole leverte n=210 studenter inn sin besvarelse ved en eksamen i statistikk. Etter sensuren sammenholdes opplysninger om bestått/ikke bestått med linje fra videregående skole . Resultatet ble (tallene er konstruerte)

	Linje 1	Linje 2	Linje 3	Sum
Bestått	74	58	38	170
Ikke bestått	16	12	12	40
Sum	90	70	50	210

Test hypotesen om at prestasjonen er uavhengig av linje. Beregn P-verdien til det observerte resultat og avgi konklusjon dersom vi krever 1% signifikansnivå.

8. En studentforening arrangerer ekskursjon til et ølbryggeri for de nye studentene. Av n=250 studentene deltok i alt 100, og disse fordelte seg etter bosted slik:

	Østpå	Sørpå	Vestpå	Nordpå	Sum
Deltok	38	16	25	21	100
Deltok ikke	57	29	40	24	150
Sum	95	45	65	45	250

Tyder dette på at interessen er uavhengig av bosted? Utfør en test med signifikansnivå 5%.

9. En butikk har reklamert i lokalavisene for en tilbudsvare som også tilbys ved reklamearrangement inne i butikken. I alt n=600 kunder er blitt intervjuet når de forlater butikken og spurt om de har kjøpt tilbudsvaren eller ikke (kjennetegn e), om de har lest avisreklamen for tilbudsvaren eller ikke (g), om de regner seg for fast kunde eller ikke (f). Anta at resultatet ble

	Fast kunde		Tilfeldig kunde	
	Kjøpt	Ikke kjøpt	Kjøpt	Ikke kjøpt
Lest reklame	85	39	98	57
Ikke lest	10	16	52	243

Analyser tallmaterialet med sikte på å belyse ulike sammenhenger mellom kjennetegnene. Hvilke(n) av årsakskjedene i Figur 9.1 er forenlig med dette tallmaterialet?

10. $\star$ Betrakt to indikatorvariable I og J med simultan sannsynlighetsfordeling

$$P[(I=i)\cap(J=j)] = p_{ij}$$

for i=0,1 og j=0,1. Vis at korrelasjonskoeffisienten mellom I og J er gitt ved (se Oppgave 5.29)

$$\rho = \frac{p_{00}p_{11} - p_{01}p_{10}}{\sqrt{(p_{00}+p_{01})(p_{10}+p_{11})(p_{00}+p_{10})(p_{01}+p_{11})}}$$

Bruk dette til å motivere det mål for samvariasjon i en 2×2 tabell som er gitt i teksten.

11. I modellen for 2×2 eksperimenter betrakt kryssproduktforholdet $K = p_{11}p_{22}/p_{12}p_{21}$. Forklar at den såkalte Gammakoeffisienten

$$\gamma = \frac{K-1}{K+1} = \frac{p_{11}p_{22} - p_{12}p_{21}}{p_{11}p_{22} + p_{12}p_{21}}$$

kan tjene som et mål for samvariasjon. Hva betyr det at
(a) γ =0 (b) γ =-1 (c) $\gamma = +1$.

12. Beregn gammakoeffisienten G for samvariasjon i følgende kontingenstabeller:

(a)	20	20	(b)	30	20	(c)	20	30
	30	30		20	30		30	20
(d)	20	10	(e)	10	5	(f)	10	40
	30	40		40	45		40	50

Kommenter resultatene.

13. (a) Vis at for 2×2 tabeller gjelder sammenhengen $Q = nR^2$ mellom kjikvadratobservatoren Q og korrelasjonsmålet R gitt i teksten.
    (b) Beregn R for situasjonene (a) - (f) i Oppgave 12.

14. Test om beregnet korrelasjonskoeffisient mellom karakteren i bedriftsøkonomi og samfunnsøkonomi i Eksempel 1.8 er signifikant forskjellig fra null. Forklar at det er rimelig å benytte en ensidig test her. Hva er populasjonen det generaliseres til? Kunne en isteden ha brukt kjikvadrattest?

15. Vis at minste kvadraters regresjonslinjen for Y mhp. X faller sammen med minste kvadraters regresjonslinjen for X mhp. Y hvis og bare hvis $R_{XY} = \pm 1$.

16. Ved en skole avholdes eksamen i to obligatoriske fag hvor kunnskaper i hvert av fagene antas å ha en viss overføringsverdi til det andre. Anta at 10 studenter deltok ved begge eksamener, og deres karakterer målt på en 0 til 9 skala ble

X:	4	2	8	5	6	7	0	5	6	7
Y:	4	0	5	4	2	6	1	5	7	6

   (a) Tegn spredningsdiagram.
   (b) Beregn $\overline{X}, \overline{Y}, S_X, S_Y, S_{XY}$ og R_{XY}.
   (c) Bestem minste kvadraters regresjonslinje for Y mhp. X og X mhp. Y, og tegn begge inn i spredningsdiagrammet.
   (d) To studenter hadde karakterene X=2 og X=7 ved den første eksamen, men møtte ikke til den andre pga. sykdom. Gi en prediksjon for deres respektive prestasjoner dersom de hadde møtt.
   (e) To studenter møtte ikke til den første eksamen pga. sykdom, men fikk karakter Y=2 og Y=7 ved den andre. Gi en prediksjon for deres prestasjoner dersom de hadde møtt.
   (f) Drøft om det er rimelig å bruke to ulike linjer i (d) og (e) til prognoseformål, eller om det ville være bedre å bruke en felles linje fastlagt ut fra spredningsdiagrammet.

17. Gitt n observasjonspar (X_i, Y_i) $i = 1, 2, \ldots, n$. Anta at det er foretatt en lineær skalaendring av observasjonene, dvs. at vi isteden observerer (V_i, W_i) $i = 1, 2, \ldots, n$, der $V_i = a + bX_i$ og $W_i = c + dY_i$ (Jfr. Oppgave 5.29).

   (a) Vis at $S_V^2 = b^2 \cdot S_X^2$ og $S_W^2 = d^2 \cdot S_Y^2$
   (b) Vis at $S_{VW} = bd \cdot S_{XY}$
   (c) Vis at $R_{VW} = \pm R_{XY}$ med positivt tegn dersom b og d har samme fortegn.

Kapittel 10

Kontrollerte eksperimenter

10.1 Innledning

De mest gunstige omstendigheter for innsamling av et tallmateriale vil være et *kontrollert eksperiment*. Med dette menes at data samles inn ved et eksperiment med veldefinerte eksperimentbetingelser, som i de fleste tilfeller er valgt av eskperimentator selv.

Et kontrollert eksperiment vil som regel gi langt mer pålitelige konklusjoner enn et mer eller mindre tilfeldig innsamlet tallmateriale. Ved en god planlegging kan vi nemlig eliminere en rekke feilkilder. Et kontrollert eksperiment gir også grunnlag for sannsynlighetsbetraktninger og for vurdering av usikkerheten knyttet til konklusjonene fra analysen av tallmaterialet. Som regel vil man også kunne forankre dataanalysen i modeller som allment kan aksepteres. En viktig egenskap ved mange kontrollerte eksperimenter er at de kan gjentas, også av andre, under tilnærmet de samme eksperimentbetingelser, slik at eventuelle konklusjoner kan etterprøves.

Muligheten for å utføre kontrollerte eksperimenter vil variere med det fagområde det er tale om. De beste mulighetene har vi i naturvitenskapene, i første rekke fysikk, biologi og kjemi, samt i teknikk og medisin. I et laboratorium er det ofte mulig å finne et eksperimentopplegg (design) og en analysemetode som er skreddersydd for det problem vi ønsker å belyse. I individorienterte fag som pedagogikk og psykologi er også mulighetene gode. Det samme gjelder visse områder relatert til økonomisk virksomhet, f.eks. i produktutvikling, produksjonsplanlegging og i studier av konsumentadferd. I samfunnsvitenskaper som sosiologi og samfunnsøkonomi er mulighetene mindre, men her finnes eksempler på observasjonssituasjoner, gitt navnet *kvasieksperimenter*, der en i noen grad kan rettferdiggjøre bruk av de samme modeller og analysemetoder som brukes under mere ideelle omstendigheter.

På en rekke fagfelter har en ingen mulighet for å influere på datamaterialet, og er henvist til rene observasjonsstudier. I slike situasjoner vil en rekke statistiske analysemetoder vanskelig kunne rettferdiggjøres. Spørsmålet er om det finnes andre analysemetoder, eller om det rett og slett er urimelig eller umulig å trekke generelle konklusjoner utfra observasjonene, slik at en analyse må begrenses til det rent beskrivende. Spesielt vanskelig er studier der observasjonene tas på ulike tidspunkter, f.eks. for å avdekke effekt av sosiale og politiske tiltak. Det kan da være vanskelig å eliminere muligheten for at eventuelle endringer skyldes andre forhold. En annen vanskelighet oppstår når de foreliggende data er fra en snevrere populasjon enn den vi ønsker å generalisere til.

Eksempelvis, en av grunnene til at det gikk såpass lang tid før røking som årsak til lungekreft ble allment akseptert i vitenskapelige kretser, var at man lenge bare hadde historiske data å holde seg til, og en retrospektiv analyse rommet altfor mange muligheter for feiltolkninger.

Vi vil illustrere en del poenger vedrørende kontrollerte eksperimenter ved et eksempel:

Eksempel 1 : Vaksine

Betrakt en situasjon der barna i et distrikt ofte lider av en bestemt sjelden sykdom. En ny og utprøvd vaksine blir introdusert og over en femårsperiode observeres en betydelig nedgang i det antall barn som angripes av sykdommen. Det er da fristende å trekke den konklusjon at nedgangen skyldes vaksinen. Dette kan være noe forhastet, idet femårsperioden kan tenkes falle sammen med en generell bedring i kosthold, hygiene og helsetjeneste.

Det hadde vært et noe sikrere grunnlag å trekke konklusjoner på, dersom man hadde gitt noen barn vaksine og andre ikke, den siste gruppen utgjør da en såkalt kontrollgruppe. Utover dette behandles de to grupper likt, eksempelvis tilbys de samme helsetjeneste. Det beste ville være om utvelgelsen skjer tilfeldig. Da sikrer man at de som vaksineres ikke utgjør en spesiell gruppe m.h.t. andre kjennetegn, noe som kan vanskeliggjøre tolkningen etterpå, f.eks. ved at de som ble vaksinert i utgangspunktet hadde gjennomgående god helse. En eventuell betydelig forskjell mellom de to gruppene etter en viss periode vil da med rimelighet kunne tilskrives vaksinen og ikke andre utenforliggende forhold.

I dette eksemplet er det altså metodologiske grunner for å utføre vaksinasjonsprogrammet i form av et kontrollert eksperiment. Det kan imidlertid være konkurrerende hensyn å ta, f.eks. rent økonomiske eller det moralske ved å ikke gi vaksinen straks til alle som kunne trenge den. Vi må da imidlertid ha for øye at alternative metoder kan medføre at man selv etter lengre tids bruk, ikke er i stand til å gi en dokumentasjon av vaksinens virkninger. Det kreves god innsikt i vedkommende fagområde kombinert med betydelig statistisk kløkt, dersom slike data skal kunne analyseres på en måte som står for kritikk, men statistisk teori kan også være til hjelp her.

Vi vil nedenfor se på et par sentrale problemstillinger i forbindelse med kontrollerte eksperimenter.

10.2 Sammenlignende eksperimenter og randomisering

I praksis vil en ofte være interessert i å klarlegge virkningen av en eller flere alternative behandlinger som kan anvendes på en viss type objekter, f.eks. ulike medikamenter anvendt på pasienter, ulike læremetoder anvendt på elever, ulike gjødninger anvendt på planter, ulike dietter ved foring av griser eller ulike bearbeidelsesmåter av et industriprodukt. Vi vil her se på situasjoner der en ønsker å sammenligne to behandlinger, den ene kan tenkes å være en ny behandling B som man ennå ikke fullt ut kjenner virkningen av, den andre kan være den tradisjonelle behandling A (som muligens består i ikke å behandle i det hele tatt).

Vi vil tenke oss at et visst antall objekter velges ut og gis behandling B, og deres responser blir så observert. Selv om vi mener å kjenne virkningen av den tradisjonelle behandlingen svært godt, vil det være hensiktsmessig å inkludere objekter som gis behandling A i eksperimentet, fordi vi da sikrer at sammenligningen av de to behandlingene skjer under de samme betingelser. Vi har dermed en *behandlingsgruppe* (B) og en *kontrollgruppe* (A), og et eksperiment av denne typen kalles ofte et *sammenlignende eksperiment*. Vi har tidligere antydet at de objekter som skal gis en bestemt behandling bør velges ut ved loddtrekning, dette kalles gjerne *randomisering*. Flere ulike former for randomisering er tenkelig, vi vil nedenfor studere to hovedtyper, *komplett randomisering og randomisering innen blokker*.

Ved komplett randomisering foreligger et gitt antall objekter som deltakere i eksperimentet, fra disse trekkes et tilfeldig utvalg som gis behandling B, mens de resterende gis behandling A. Ved randomisering innen blokker tenker vi oss isteden at objektene er inndelt i flere grupper, kalt blokker, og at det trekkes et tilfeldig utvalg fra hver blokk som gis behandling B, mens de resterende i hver blokk gis behandling A. Vi ser at vi med den siste framgangsmåten kan sikre at visse typer objekter er med både blant de som gis behandling A og B.

Den observerte respons i et kontrollert eksperiment kan være av to typer, enten gruppert respons (f.eks. i kategorier av typen suksess-fiasko, ja-nei, høy-lav) eller gradert respons (f.eks. målinger i meter, kilo, sekunder etc.).

Vi vil illustrere noen typiske problemstillinger for de situasjonene som er skissert ovenfor med eksempler. La oss starte med å studere gruppert respons:

10.3 Gruppert respons

Eksempel 2 : Verktøy-randomisert
Ved produksjon av en bestemt artikkel kan en vanskelig arbeidsoperasjon enten utføres for hånd (metode 1) eller med et spesialverktøy (metode 2). Blant N=21 arbeidere velges ut tilfeldig n=11 arbeidere som skal utføre arbeidet etter metode 2, mens de 10 andre skal bruke metode 1. Vi er bl.a. interessert i å teste om det er grunn til å påstå at metode 2 gir bedre resultat enn metode 1. For hver deltaker i eksperimentet observeres det om vedkommende lykkes med arbeidsoperasjonen i første forsøk eller ikke. Resultatene var

	Lykkes	Lykkes ikke	Sum
Metode 1	6	4	10
Metode 2	8	3	11
Sum	14	7	21

Vi ser at metode 2 kom relativt gunstig ut, idet 40% av deltakerne ikke lyktes med metode 1, mens det tilsvarende tall for metode 2 ble ca. 27%. De observerte forskjeller i arbeidsresultatet kan enten forklares ved at metode 2 virkelig er bedre enn 1, eller ved at de er likeverdige og at det observerte resultat skyldes tilfeldigheter (vi antar at metode 2 ikke under noen omstendighet er dårligere enn 1). La oss som nullhypotese H_0 bruke at metodene er likeverdige. Som testobservator bruker vi antallet Y blant de som bruker metode 2 som ikke lykkes. Det er små verdier av Y som gir grunnlag for å forkaste nullhypotesen til fordel for metode 2. Dersom nullhypotesen er riktig er Y hypergeometrisk fordelt ($N = 21$, $M = 7$, $n = 11$). Dette forklarer vi slik: Dersom de to metodene er likeverdige, vil det være $M = 7$ som ikke lykkes i første forsøk uansett hvilken metode de brukte (om de lykkes eller ikke skyldes altså ikke metoden, men andre forhold). Den observerte verdi av Y vil derfor være bestemt ved trekningen, nemlig ved hvilke arbeidere blant de $M = 7$ spesielle som tilfeldigvis ble trukket ut til å arbeide etter metode 2.

Beregner vi P-verdien til det observerte resultat, får vi

$$P_{H_0}(Y \leq 3) = \sum_{y=0}^{3} \frac{\binom{7}{y}\binom{14}{11-y}}{\binom{21}{11}} = 0.438$$

Denne er såpass høy at det ikke er noen rimelig grunn til å påstå at metode 2 virkelig gir bedre resultat enn metode 1.

Vi ser at i dette eksemplet hadde randomiseringen en rolle utover det å sikre rettferdighet for de to metodene. Den satte oss også i stand til å finne den eksakte sannsynlighetsfordelingen til testobservatoren Y når nullhypotesen er riktig. Analysemetoden som er benyttet i dette eksemplet kalles *Fisher-Irwins test*, og er aktuell i en rekke situasjoner. La oss imidlertid se på den tilsvarende problemstilling der det ikke er foretatt noen randomisering.

Eksempel 3 : Verktøy-ikke randomisert
En bedrift produserer en artikkel på to produksjonssteder 1 og 2. På det ene stedet (2) er det anskaffet spesialverktøy, og før man bestemmer seg for å anskaffe dette på det andre produksjonsstedet, vil man gjennomføre et eksperiment som omfatter alle arbeiderne på begge produksjonsstedene, henholdsvis 10 ved sted 1 og 11 ved sted 2, som alle får arbeide under samme produksjonsforhold. Anta, for korthets skyld, at data er de samme som i Eksempel 2. Denne situasjonen faller imidlertid utenfor rammen av den modell vi brukte i forrige eksempel, fordi det her ikke foreligger noen randomisering.

Det foreliggende problem kan imidlertid studeres slik: Vi kan se på produksjonsresultatet for de to produksjonsstedene som to binomiske forsøksrekker med henholdsvis n_1=10 og n_2=11 forsøk i hver og med suksesssannsynlighet henholdsvis p_1 og p_2. Disse tolkes som sannsynligheten for at en tilfeldig arbeider på vedkommende arbeidsplass lykkes (se for øvrig Oppgave 6.16). Nullhypotesen om at de to produksjonsstedene holder samme kvalitet kan formuleres som $p_1 = p_2$, mens alternativet at sted 2 gir bedre kvalitet kan formuleres som $p_2 > p_1$. La X_1 og X_2 være antall som lykkes ved de to produksjonsstedene. Antar vi at produksjonsresultatene for de ulike arbeiderne er uavhengige, vil X_1 og X_2 være uavhengige stokastiske variable som er binomisk fordelte henholdsvis (n_1, p_1) og (n_2, p_2).

Det kan gis argumenter for at testing av hypotesen om ingen forskjell kan utføres som i forrige eksempel, dvs. betrakte alle marginalene som gitte tall. Som testobservator kan brukes en hvilken som helst av de fire størrelsene inn i tabellen. Merk at når alle marginalene er gitte vil en av disse bestemme de tre andre (jfr. Oppgave 6.35). Testingen foregår da som om den valgte testobservator, under nullhypotesen, var hypergeometrisk fordelt (N, M, n) der M og n er de respektive marginaler, eksempelvis er X_1 hypergeometrisk fordelt (N=21, M=14, n=10).

Selv om de to eksemplene ovenfor er beslektet, og vi har foreslått samme analysemetode for begge, er det en prinsipiell forskjell m.h.t. de konklusjoner det er mulig å trekke. Anta at datamaterialet var slik at det var grunnlag for å forkaste hypotesen om samme kvalitet til fordel for produksjon med spesialverktøy. I det første eksemplet vil det, p.g.a. randomiseringen, være grunnlag for å påstå at spesialverktøy generelt gir bedre produksjonsresultat, mens det i det andre eksemplet ikke er grunnlag for å trekke en så vidtrekkende konklusjon, med mindre vi på forhånd visste at de to arbeidsplassene var like m.h.t. kvaliteten av produksjonen uten spesialverktøy.

En alternativ testmetode basert på normaltilnærming er U-testen som vi studerte i Kapittel 7.6, som er ekvivalent med en kjikvadrattest som beskrevet i Kapittel 9.3.

La oss så se på et eksempel av typen randomiserte blokker. For enkelhets skyld betrakter vi en situasjon med bare to objekter innen hver blokk, slike situasjoner forekommer ofte i praksis og kalles *parvise sammenligninger*.

Eksempel 4 : Konservering
En bedrift som produserer syltetøy vurderer en ny metode for konservering som det påstås gir bedre smak enn den som brukes nå. Bedriften ønsker å finne ut om det er grunnlag for dette, og det utføres derfor et kontrollert eksperiment. Det er mulig at råstoffet kan variere fra koking til koking, og det synes derfor rimelig å dele hver koking i to deler hvorav den ene konserveres etter den nye metode, den andre etter den gamle. Utvelgelsen bør skje ved loddtrekning, det sikrer en rettferdig behandling dersom kvaliteten innen hver koking skulle variere. Dette gjentas for i alt ti kokinger. I denne situasjon har vi altså ti blokker, hver blokk består av to kvanta syltetøy.

La vår nullhypotese være at de to metodene er likeverdige, mens alternativet er at den nye metoden gir bedre produkt. Når syltetøyet er ferdig for salg, avgjør bedriftens smaksekspert på grunnlag av en rekke smaksprøver hvilken av de to typer syltetøy innen hver koking som hun mener smaker best, anta for enkelhets skyld at uavgjort ikke forekommer. La X være antall ganger det nye syltetøy var best, store verdier av X gir grunnlag for å forkaste nullhypotesen. Når nullhypotesen er riktig vil X være binomisk fordelt (n=10, p=0.5), fordi da er det tilfeldig hvilket av de to produkter som smakseksperten velger ut som best. Anta at den observerte verdi av X er 7. P-verdien til det observerte resultat er da

$$P = P_{H_0}(X \geq 7) = 0.1719,$$

som neppe gir grunn til å påstå at det nye produkt er bedre enn det gamle. Den analysemetode vi har brukt her går ofte under navnet *tegntesten*.

For å undersøke om det nye konserveringsmidlet gir bedre smak er det foreslått et alternativt eksperiment, nemlig at de to typer syltetøy blir prøvd ut på et panel bestående av et visst antall konsumenter. Kan hende har bedriften erfaring for at kvaliteten av produktet er svært jevn både innen hver koking og fra koking til koking, slik at dette utelukkes som eventuelle feilkilder ved undersøkelsen. Imidlertid tenker man seg at dersom det er forskjell i smak for de to produktene så kan det slå ut ulikt fra person til person, idet smak og behag er forskjellig. I denne situasjonen har vi en blokk bestående av to prøver for hver person, og vi vil, for hver person, trekke lodd hvilken av de to prøver som vedkommende får smake først. Vi unngår da eventuelle feil som skyldes at personer, i en situasjon hvor de ikke smaker forskjell, systematisk sier at det første (evt. siste) produktet er best. Analysen av resultatene kan foregå på samme måte som ovenfor. I en situasjon der vi ikke kan utelukke feilkilder av den typen vi skisserte først, vil vi kan hende utføre et større eksperiment der hver deltaker i panelet får smake flere smaksprøver fra ulike kokinger. Analysen av et slikt eksperiment blir noe mer komplisert.

10.4 Gradert respons-Rangtester

La oss så se på noen situasjoner hvor den observerte respons er gradert.

Eksempel 5 : Verktøy-randomisert
La situasjonen være som beskrevet i Eksempel 2, men anta isteden at det er forbrukt tid for arbeidsoperasjonen som er av interesse. Det er påstått at spesialverktøy (metode 2) gjennomgående forkorter arbeidstiden. Blant de N=21 arbeidere velges ut tilfeldig n=11 som utfører arbeidet etter metode 2, de resterende etter metode 1. Resultatet ble (i sekunder)

Metode 1:	45	48	61	52	48	63	52	54	50	58	
Metode 2:	47	59	43	50	45	45	49	41	47	95	50

For å karakterisere tendensen i dette tallmaterialet kunne vi bruke gjennomsnittstidene for de to metodene. Imidlertid ser vi at den nest siste observasjon for metode 2 er svært avvikende [1] og vil trekke gjennomsnittet opp på et nivå som ikke er typisk for observasjonene som helhet. Det synes mer rimelig å bruke medianen til å karakterisere materialet. Den er for metode 1 lik 52 og for metode 2 lik 47, slik at spesialverktøy ser ut til å kunne redusere arbeidstiden. Spørsmålet er om forskjellen er stor nok til at vi avviser at den skyldes tilfeldigheter. Vi stiller da opp nullhypotesen at de to metodene er like m.h.t. tidsforbruk og alternativet at metode 2 er raskere. En metode å teste denne hypotesen på er å ordne tallmaterialet i stigende rekkefølge og understreke observasjoner svarende til metode 2, slik

$$\begin{array}{ccccccccccc} \underline{41} & \underline{43} & \underline{45} & 45 & \underline{45} & \underline{47} & \underline{47} & 48 & 48 & \underline{49} & \underline{50} \\ 50 & \underline{50} & 52 & 52 & 54 & 58 & \underline{59} & 61 & 63 & \underline{95} & \end{array}$$

Vi tildeler så de 21 observasjonene rang fra 1 til 21 i henhold til rekkefølgen. Sum rang for metode 2-observasjonene blir

$$W = 1 + 2 + 3 + 5 + 6 + 7 + 10 + 11 + 13 + 18 + 21 = 97$$

Merk at det kan oppstå et problem ved at to eller flere observasjoner har samme verdi, og dette må løses på en, for de to metodene, rettferdig måte.

Det er grunn til å forkaste nullhypotesen når W er tilstrekkelig liten, fordi dette svarer til at metode 2-observasjonene er gjennomgående små. For å avgjøre om det observerte resultat gir grunn til forkasting må vi beregne P-verdien

$$P = P_{H_0}(W \leq 97).$$

Slike sannsynligheter kan, som følge av randomiseringen, beregnes eksakt, og det er utarbeidet tabeller som gjør det mulig å avlese disse direkte. Vi vil her ikke gå nærmere inn på teorien for eksakt beregning, ei heller på bruken av slike tabeller, og må derfor henvise leseren til annen litteratur. Den skisserte metode går i litteraturen under navnet *Wilcoxons test for to utvalg*. Imidlertid kan det vises at dersom N ikke er altfor liten, f.eks. minst 15, og utvalget ikke er altfor skjevt m.h.t. antallet i hver gruppe, så vil sannsynlighetsfordelingen til W kunne tilnærmes til normalkurven. Vi vil da trenge at såframt nullhypotesen er riktig (se Oppgave 13)

[1] Ved nærmere undersøkelse viste det seg at observasjonen skyldtes et arbeidsuhell som ikke hadde noe med bruk av spesialverktøy å gjøre.

$$EW = \frac{n(N+1)}{2} \quad varW = \frac{n(N-n)(N+1)}{12}$$

Brukes dette i den foreliggende situasjon, får vi EW=121 og $varW$=201.67, slik at

$$P = P_{H_0}(W \leq 97) \approx G(\frac{97-121}{14.2}) = G(-1.69) = 0.0455$$

Materialet gir altså rimelig grunn til å påstå at metode 2 er raskere (signifikant på 5% nivå). I praksis vil man trenge de eksakte tabellene som er nevnt ovenfor bare for spesielt små N.

Over til en situasjon med parrede sammenligninger:

Eksempel 6 : Verktøy-randomiserte blokker
La situasjonen være som i forrige eksempel, men anta isteden at det deltar n=10 arbeidere som utfører arbeidsoperasjonen to ganger, en med hver metode. Her utgjør hver arbeider en blokk bestående av to forsøk, og for hver arbeider trekkes det lodd om hvilken metode som skal brukes først. Det kan ikke utelukkes at det er betydelige individuelle variasjoner i arbeidstempo fra arbeider til arbeider. Med dette eksperimentopplegget blir imidlertid de to metodene satt opp mot hverandre for hvert individ, og dette burde gi et sikrere grunnlag for konklusjoner vedrørende forskjeller mellom metodene. Det er derfor trolig at opplegget her er å foretrekke framfor opplegget i forrige eksempel. Imidlertid kan det tenkes forhold der opplegget ikke kan brukes, f.eks. dersom en arbeider som er opplært til å bruke spesialverktøy ikke lenger er i stand til å utføre arbeidsoperasjonen uten spesialverktøy like godt som før. Anta at dette ikke er tilfellet her. Resultatet av eksperimentet ble (i sekunder)

Arbeider nr.:	1	2	3	4	5	6	7	8	9	10
Metode 1:	48	53	52	57	43	83	59	71	40	61
Metode 2:	45	42	58	50	41	47	53	66	45	53

For å karakterisere tendensen i dette tallmaterialet kan vi først beregne forskjellen i forbrukt tid for hver arbeider. Denne er

Forskjell:	3	11	−6	7	2	36	6	5	−5	8

Den gjennomsnittlige forskjell er 6.7 sekunder, mens medianforskjellen er 5.5. Spørsmålet er om denne forskjell er stor nok til å se bort fra at den skyldes tilfeldigheter. La nullhypotese og alternativ være som i forrige eksempel. En måte å teste hypotesen på, er å ordne forskjellene etter stigende tallverdi slik (merk den rettferdige behandling av de sammenfallende tallverdier)

$$2 \quad 3 \quad 5 \quad -5 \quad -6 \quad 6 \quad 7 \quad 8 \quad 11 \quad 36$$

Vi tildeler disse tall rang fra 1 til 10 i henhold til rekkefølgen. Sum rang for de negative differanser blir

$$V = 4 + 5 = 9.$$

Det er grunn til å forkaste nullhypotesen når V er tilstrekkelig liten fordi dette betyr få og små negative differenser, dvs. metode 2-observasjonene er gjennomgående mindre enn de tilhørende metode 1-observasjonene. For å avgjøre om det observerte resultat gir grunn til forkastning må vi beregne P-verdien

$$P = P_{H_0}(V \leq 9).$$

Slike sannsynligheter kan, som følge av randomiseringen, beregnes eksakt, og det er utarbeidet tabeller hvor vi kan avlese disse direkte, vi viser igjen til annen litteratur. Den skisserte metode går i litteraturen under navnet *Wilcoxons test for parrede observasjoner* . Det kan vises at dersom n ikke er altfor liten, så vil sannsynlighetsfordelingen til V kunne tilnærmes med normalkurven. Vi vil da trenge at såframt nullhypotesen er riktig blir (se Oppgave 14)

$$EV = \frac{n(n+1)}{4} \quad varV = \frac{n(n+1)(2n+1)}{24}$$

Brukes dette i den foreliggende situasjon får vi $EV = 27.5$ og $varV = 96.25$ slik at

$$P = P_{H_0}(V \leq 9) \approx G(\frac{9 - 27.5}{9.81}) = G(-1.89) = 0.029$$

mens det eksakte tall ifølge tabell er 0.032. [2] Materialet gir altså rimelig grunn til å påstå at metode 2 er raskere (signifikant på 5% nivå).

[2] Bruk av heltallskorreksjon gir forøvrig enda bedre tilnærming.

For problemstillingene i de to siste eksemplene har vi skissert enkle analysemetoder basert på ranger. Det finnes også andre metoder. Mest brukt er kanskje metoder basert på gjennomsnittsbetraktninger på observasjonene direkte. Disse forutsetter som regel en såkalt variansanalysemodell, og dette er temaet for Kapittel 11. For eksakt beregning av P-verdier kreves da i tillegg forutsetning om normalitet av de opprinnelige observasjonene, noe som ikke ser ut til å være tilfellet for materialene i Eksempel 5 og 6. En vil kanskje kunne tro at en kaster bort mye informasjon ved å erstatte de opprinnelige tall men ranger. Dette er ikke tilfelle, og i tillegg har slike metoder den gunstige egenskap at eventuelle ekstreme observasjoner, som muligens skyldes uhell, ikke får avgjørende betydning for konklusjonen. Metoden er hva statistikerne kaller *robust*.

En kommentar om alternative eksperimentopplegg: Ved randomiserte blokker forsøker en å velge blokker med objekter som er mest mulig homogene, men gjerne med betydelig variasjon mellom blokkene. Da får en prøvd de ulike behandlingsmetoder direkte mot hverandre under mest mulig like forhold. En vil derfor som regel foretrekke dette eksperimentopplegg framfor komplett randomisering dersom begge er mulig å realisere. Eksempelvis vil en psykolog ved sammenligning av to læremetoder gjerne arbeide med en rekke par av en-eggede tvillinger, men slike er ikke alltid lette å oppdrive, og han må kan hende være foruten.

Vi har her bare tatt for oss eksperimenter med en faktor med to kategorier: metode 1 og metode 2. Eksperimenter med flere faktorer og med flere kategorier for hver faktor, er diskutert i Kapittel 11 og 15.

Merknad: De rangmetodene som er presentert i dette avsnittet forutsetter ingen spesiell sannsynlighetsfordeling for den graderte respons, de er såkalte fordelingsfrie metoder. Metodene er i praksis også brukt i situasjoner der observasjonene selv er på en rangeringsskala, dvs. er ordinale (se innledningen i Kapittel 9). Dette er ofte tilfelle i markedsforskning, der konsumenter bedømmer iht. en vurderingsskala, for eksempel en 7-delt skala fra Dårlig (-3) via Middels (0) til Godt (3). Metoden er ikke like godt teoretisk forankret i slike situasjoner, og et lite antall ordinale kategorier leder til mange sammenfallende ranger. God programvare bør imidlertid holde orden på dette, slik at reelle sannsynlighetsgarantier kan gis under visse forutsetninger.

10.5 Oppgaver

1. En sykehuslege tror at en bestemt behandling kan forlenge livet til personer som har hatt hjerteattakk. Diskuter muligheten for et forsøksopplegg der eventuelle forskjeller i levetid med rimelighet kan tilskrives behandlingen og ikke andre utenforliggende forhold.

2. Det påstås at man ved sprøyting med visse kjemikalier fra fly kan øke sjansen for at regnskyer avgir regn. Diskuter de problemer som er knyttet til å kunne avgjøre dette. Foreslå et eksperimentopplegg som kan gi et grunnlag for pålitelige konklusjoner.

3. Man ønsker å utføre en undersøkelse om bruk av selvinstruerende (programmert) lærestoff kontra vanlig klasseromsundervisning. En gruppe elever tilbys de to alternativer og får velge selv. Etter endt undervisning tar begge grupper samme prøve og resultatene sammenlignes. Diskuter eventuelle betenkelige sider ved et slikt forsøksopplegg, foreslå et alternativt opplegg. Diskuter også muligheten av opplegg som tar omsyn til elevenes motivasjon (evt. uvillighet) til å delta etter et bestemt opplegg, eller overhodet.

4. La situasjonen være som i forrige oppgave. Blant 24 elever velges ut tilfeldig 12 elever som får selvinstruerende undervisning, resten får vanlig undervisning. Man ønsker å vite om det er grunn til å påstå at selvinstruerende undervisning gir dårligere resultat enn vanlig undervisning. Resultatet av den etterfølgende prøve ble

	Bestått	Ikke bestått	Sum
Ny undervisning	9	3	12
Vanlig undervisning	11	1	12
Sum	20	4	24

   Formuler situasjonen som et hypotesetestingsproblem, beregn P-verdi og avgi konklusjon dersom signifikansnivået er valgt lik 5%.

5. En bedrift vurderer to lagringsmetoder, den ene er noe mer kostbar enn den andre. Et produksjonsparti på 400 artikler deles i to og hver halvpart lagres etter hver sin metode. Blant de 200 som ble lagret etter den rimeligste metode måtte 15 vrakes etter en viss tid, mens bare 10 måtte vrakes av de som var lagret etter den kostbare metoden. Gir dette grunnlag for å påstå at den rimelige metoden er dårligere enn den kostbare metoden? Hvilken forutsetning bygger analysen på?

6. Et gartneri ønsker å undersøke om en viss type frø har økt spireevne dersom det pakkes i en ny type emballasje (1) i forhold til vanlig emballasje (2). Ta sannsynligheten for at et tilfeldig frø spirer som et uttrykk for spireevnen. Det tas $n_1 = n_2 = 50$ frø av hver sort og disse sås under de samme forhold. Av de frø som var lagret i den nye emballasje spirte 45 og av de som ble lagret i

den gamle emballasjen spirte 40. Estimer forskjellen i spireevne og rapporter resultatet. Test hypotesen om at de to spireevnene er like, beregn P-verdien til det observerte resultat og gi konklusjonen dersom signifikansnivået er valgt lik 5%.

7. Man ønsker å undersøke om en ny type solbadolje A er bedre enn et konkurrerende produkt B. I alt 12 personer har sagt seg villige til å delta i et eksperiment. Hver rygg "deles i toden ene del smøres inn med A, den andre del med B, hvilken del som får A skjer ved loddtrekning. Etter en lengre periode i solen observeres grad av solbrenthet på en skala fra 0 til 3. Anta at resultatet var

Person nr.:	1	2	3	4	5	6	7	8	9	10	11	12
Med A:	2	1	1	2	0	2	1	0	2	1	1	1
Med B:	3	2	1	1	1	3	0	1	2	1	3	2
Forskjell:	−	−	0	+	−	−	+	−	0	0	−	−

Ved analysen velger en å se bort fra de personer der forskjellen er null. La X være antall positive differenser.

(a) Forklar at dersom produktene i virkeligheten er likeverdige vil X være binomisk fordelt (n=9, p=0.5).

(b) Hvordan kan eventuelle forskjeller mellom A og B forklares?

(c) Formuler situasjonen som et hypotesetestingsproblem. Hva blir konklusjonen dersom det valgte sinifikansnivå var 5%?

8. 15 skiløpere i Norgeseliten er samlet i treningsleir. Det er nettopp markedsført en ny type skismøring (A) og langrennstreneren ønsker å teste denne mot en type smøring (B) som han vet er velegnet for dagens føre. Treneren velger ut tilfeldig 7 løpere som smører med A, mens resten smører med B. Deretter starter alle løpere i et 15 km langrenn med fellesstart. Resultatet ble i minutter og sekunder

A:	48.13	47.59	48.09	47.43	48.31	48.01	49.03	
B:	48.06	47.45	49.27	48.36	48.54	49.14	48.20	48.26

Dersom den nye smøringen er signifikant bedre enn den gamle ønsker man å gjøre dette kjent blant konkurranseløpere. Beregn tilnærmet P-verdi. Hva blir konklusjonen dersom signifikansnivå 5% er valgt? Enn 10%?

9. Se igjen på problemstillingen i Oppgave 3 og 4. Anta at hver prøve blir gradert på en karakterskala fra 0 til 9, hvor 0 og 1 betyr ikke bestått. Anta at resultatet ble

Ny undervisning:	1	8	5	6	5	3	1	0	3	2	6	4
Vanlig undervisning:	0	2	5	7	4	4	9	7	5	5	3	6

Vil du på dette grunnlag påstå at selvinstruerende undervisning gir dårligere resultat.

10. Det blir påstått at ingrediensen Luriol, når den blir blandet i bensinen, gir en høyere toppfart for biler. For å undersøke dette blir følgende undersøkelse gjennomført: Ti bilfirmaer har hver stillet til rådighet to biler som begge er av samme modell og standard. Det trekkes lodd hvilken av de to bilene av hver modell som skal få Luriol. Det blir så arrangert en fartsprøve. Resultatet ble i km/t.

Modell:	1	2	3	4	5	6	7	8	9	10
Uten Luriol:	136	94	127	125	133	124	127	138	103	113
Med Luriol:	140	108	128	128	126	137	139	147	112	108

Uten en test som tar sikte på å klargjøre om Luriol har den påståtte effekt. Vi ønsker signifikansnivå 5%.

11. Betrakt igjen problemstillingen fra Oppgave 3, men anta nå at de 24 elevene er gruppert to og to slik at de to innen hvert par står noenlunde likt i faget før undervisningen tar til. Fra hvert par velges tilfeldig en som får selvinstruerende undervisning, den andre får vanlig undervisning. Anta at karakterene på prøven etter undervisningsperioden ble

Par nr.:	1	2	3	4	5	6	7	8	9	10	11	12
Ny undervisning:	3	1	4	5	7	3	6	4	5	7	6	4
Vanlig undervisning:	5	3	4	4	7	5	7	5	6	5	8	6

Vil du på dette grunnlag påstå at selvinstruerende undervisning gir dårligere resultat. Bruk 5% signifikansnivå.

12. Her følger en utskrift fra analyse av dataene i Eksempel 5.

```
 >> READ 'eks10.5' X Y
>> WILCOXON2 X Y
-- W = 134.0 (expected = 121)
-- P = 0.0968 (twosided adjusted for ties)
```

Forklar den tilsynelatende forskjell fra resultatet i teksten. Reproduser resultatet med din foretrukne programvare (testen kalles også Mann-Whitney).

Ved løsning av de to neste oppgavene trengs formlene:

$$1 + 2 + 3 + \cdots + k = k(k+1)/2$$
$$1^2 + 2^2 + 3^2 + \cdots + k^2 = k(k+1)(2k+1)/6$$

13. $\star$Betrakt en lotterisituasjon hvor elementene i populasjonen har verdiene $1, 2, 3, \ldots, N$, dvs. $v_1 = i$.

    (a) Vis at i dette tilfellet blir

    $$\overline{v} = \frac{1}{2}(N+1) \quad \text{og} \quad \sigma^2 = \frac{(N+1)(N-1)}{12}$$

    (b) Bruk dette til å vise formlene for EW og $varW$ som trengs til Wilcoxons test for to utvalg.

14. $\star$La $I_1, I_2, \ldots, I_n$ være uavhengige indikatorvariable slik at

    $$P(I_j = 1) = P(I_j = 0) = \frac{1}{2}$$

    og la

    $$V = I_1 + 2I_2 + \cdots + nI_n.$$

    (a) Vis at

    $$EV = \frac{n(n+1)}{4} \quad \text{og} \quad varV = \frac{n(n+1)(2n+1)}{24}$$

    (b) Vis at testobservatoren V som brukes i Wilcoxons test for parrede observasjoner har form som ovenfor, der I_j indikerer om observasjonen som har fått rang j er negativ eller positiv.

Kapittel 11

Variansanalyse

11.1 Innledning

I Kapittel 8.1 studerte vi analyse av data som i vid forstand kunne oppfattes som gjentatt observasjon av samme ukjente størrelse. I praksis foreligger ofte observasjoner der vi har supplerende opplysninger, f.eks. at observasjonene kan henføres til gitte grupper. Vi kan da være interessert i eventuelle gruppeforskjeller. Slike problemstillinger forekommer ofte i forbindelse med kontrollerte eksperimenter (se Kapittel 10), men også i andre sammenhenger.

Vi skal derfor ta opp generaliseringer av målemodellen som går under navnet *variansanalysemodeller*. Dette er modeller som forklarer observasjonene ut fra variasjon i en eller flere kategorivariable, her kalt *faktorer*. Til hver faktor hører to eller flere kategorier ("nivåer"), og kategorikombinasjonen bestemmer gruppetilhørigheten til observasjonen. I modellene gjøres antakelser om hvordan faktorene bidrar til den forventede respons. Vi vil kunne være interessert i hvilken kombinasjon av nivåer på de ulike faktorer som gir størst forventet respons, eventuelt om enkelte av faktorene er uten betydning. Eksempler på dette kan være:

- Innsatsfaktorer og produksjonsresultat.
- Markedsføringstiltak og salg.
- Kostholdsfaktorer og ytelse (evt. reduksjon av risiko for skade/død).

Vi vil gi et kort innblikk i aktuelle modeller og analysemetoder knyttet til konkrete eksempler. Eksakte pålitelighetsgarantier forutsetter at responsene er uavhengige og normalfordelte, og er knyttet til t-fordelingen, som vi ble kjent med i Kapittel 8, og en ny fordeling kalt F-fordelingen. Dersom dataene er samlet inn ved et randomisert eksperiment (se Kapittel 10), vil garantiene gjelde tilnærmet selv uten normalitetsantakelse. Uttømmende begrunnelser for disse metodene krever matematiske kunnskaper ut over det vi har forutsatt til nå, men en grundigere behandling er imidlertid lett tilgjengelig ellers i litteraturen.

11.2 Konstant-effekt og to-utvalgsmodellen

Vi er ofte interessert i å studere virkningen og eventuelt forskjeller i responsen ved to ulike framgangsmåter, eksempelvis to behandlingsmåter for individer, to produksjonsmåter for artikler etc. Vi skaffer oss derfor observasjoner for hver av framgangsmåtene. Hvilken modell som er aktuell vil avhenge av omstendighetene, vi vil her studere to vesensforskjellige situasjoner, som best illustreres ved eksempler.

Eksempel 1 : Spesialverktøy
En vanskelig arbeidsoperasjon kan utføres uten eller med spesialverktøy. Vi ønsker å anslå den tidsforskjell som forventes ved de to metodene.[1] Anta at bedriften har latt $n = 10$ arbeidere utføre arbeidsoperasjonen en gang med hver metode. La X_i og Y_i være forbrukt tid for i'te arbeider ved henholdsvis uten og med spesialverktøy ($i = 1, 2, \ldots, n$). Vi vil anta at $X_1, X_2, \ldots, X_n, Y_1, Y_2, \ldots, Y_n$ er uavhengige stokastiske variable med forventninger henholdsvis

$$EX_i = \alpha_i + \delta \quad EY_i = \alpha_i \quad (i = 1, 2, \ldots, n)$$

og varianser

$$varX_i = \sigma_1^2 \quad varY_i = {\sigma_2}^2 \quad (i = 1, 2, \ldots, n)$$

I denne modellen blir δ å oppfatte som tillegget (evt. fradraget) i forventet tid som skyldes at spesialverktøy ikke brukes, og det vil være av interesse å estimere δ, eventuelt teste hypoteser om δ på grunnlag av observasjonene. Merk at modellen tillater at de ulike arbeiderne har ulikt arbeidstempo (dvs. ulike α_i), men det kreves altså at tillegget (fradraget) δ er det samme for alle arbeiderne. Videre ser vi at variasjonen i arbeidstempo tillates å være ulik for de to metodene, men for hver metode kreves det at den er samme

[1] Denne situasjonen ble også brukt som gjennomgangseksempel i Kapittel 10 om kontrollerte eksperimenter.

for alle arbeiderne. Disse antakelsene kan selvsagt diskuteres. Den mest interessante er kanskje at tillegget δ antas konstant, og det har gitt modellen navnet *konstant-effekt modellen.*

Vi vil nå studere inferensproblemer for modellen presentert i eksemplet. Betrakt forskjellen i respons for hver arbeider

$$Z_i = X_i - Y_i \quad (i = 1, 2, \dots, n).$$

Ifølge antakelsene i modellen blir

$$\begin{aligned} EZ_i &= E(X_i - Y_i) = EX_i - EY_i = \alpha_i + \delta - \alpha_i = \delta \\ varZ_i &= var(X_i - Y_i) = varX_i + varY_i = \sigma_1^2 + \sigma_2^2 \end{aligned}$$

Vi ser nå at $Z_1, Z_2, \dots, Z_n$ er uavhengige stokastiske variable med samme forventning δ og med samme varians $\sigma^2 = \sigma_1^2 + \sigma_2^2$, dvs. oppfyller betingelsene i den enkle målemodellen fra Kapittel 8.1, og estimering av δ er derfor redusert til et velkjent problem basert på observasjonene $Z_1, Z_2, \dots, Z_n$. En aktuell estimator for δ er

$$\hat{\delta} = \overline{Z} = \frac{1}{n}\sum_{i=1}^{n} Z_i = \overline{X} - \overline{Y}.$$

Denne er forventningsrett med varians σ^2/n. Vanligvis er σ^2 ukjent, men en forventningsrett estimator er som før

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n} (Z_i - \overline{Z})^2$$

Som estimator for standardavviket $\sigma(\hat{\delta}) = \sigma/\sqrt{n}$, bruker vi $S(\hat{\delta}) = S/\sqrt{n}$, Estimert δ med tilhørende standardavvik blir derfor

$$\overline{Z} \pm \frac{S}{\sqrt{n}}$$

Eksempel 1 : Spesialverktøy (fortsatt)
La oss illustrere dette med data for Eksempel 1. Anta at observasjonene (i sekunder) er gitt ved

X_i:	48	53	52	57	43	63	59	51	40	61
Y_i:	45	42	58	50	41	47	53	46	45	53

Differensene i forbrukt tid blir

$$Z_i: \quad 3 \quad 11 \quad -6 \quad 7 \quad 2 \quad 16 \quad 6 \quad 5 \quad -5 \quad 8$$

Utregningen gir $\overline{Z} = 4.7$ og $S = 6.7$, og for anslått forventet reduksjon i forbrukt tid med spesialverktøy rapporterer vi 4.7 ± 2.1 . Sett i forhold til det estimerte standardavvik, ser det ut til at reduksjonen er signifikant, dvs. neppe skyldes tilfeldigheter. Om reduksjonen er av praktisk betydning for bedriften er en annen sak.

Eksakte sannsynlighetsutsagn om estimeringsfeilen i konstanteffekt modellen kan ikke gjøres uten ytterligere forutsetninger. Dersom vi er villig til å anta at observasjonene er normalfordelte, kan man ta utgangspunkt i den stokastiske variable

$$T = \frac{\overline{Z} - \delta}{S/\sqrt{n}}$$

som da er t-fordelt med $n-1$ frihetsgrader (se Kapittel 8.4). Ved hjelp av en tabell over arealer under t-kurver kan vi derfor lage sannsynlighetsutsagn. Eksempelvis med 9 frihetsgrader er sannsynligheten for estimeringsfeil på høyst 2.26 ganger estimert standardavvik lik 0.95 (se Tabell C.9).

Det kan også være aktuelt å teste hypoteser om δ, f.eks. teste hypotesen $\delta = \delta_0$, der δ_0 er et spesifisert tall. Hypotesen om ingen forskjell mellom de to metodene svarer til å sette $\delta_0 = 0$. En mulig testobservator er T, der vi for δ bruker den spesifiserte δ_0. Under forutsetning av normalfordelte observasjoner vil denne T, dersom nullhypotesen er riktig, være t-fordelt med $n-1$ frihetsgrader. Forkastingsområdet vil avhenge av om problemstillingen er ensidig eller to-sidig. I eksemplet vil det første være tilfelle dersom vi på forhånd vet at spesialverktøy i hvert fall ikke øker forventet forbrukt tid. Ved hjelp av t-tabellen kan vi derfor etablere tester med ønsket signifikansnivå, eventuelt beregne P-verdier.

En viktig grunn til at konstant-effekt modellen er rimelig i Eksempel 1, er at responsen for de to framgangsmåtene blir observert for samme individ. Vi vil nå se på en problemstilling der dette ikke er tilfelle.

Eksempel 2 : Spesialverktøy

Anta at problemstillingen er som i Eksempel 1, men at bedriften isteden har latt $n_1 = 10$ arbeidere utføre arbeidsoperasjonen etter metode 1 (uten spesialverktøy) og $n_2 = 11$ andre arbeidere etter metode 2 (med spesialverktøy). La oss nummerere arbeiderene innen hver gruppe, og la X_j være forbrukt tid for j'te arbeider i gruppe nr.1 ($j = 1, 2, \ldots, n_1$) og Y_j være forbrukt tid for j'te arbeider i gruppe nr. 2 ($j = 1, 2, \ldots, n_2$). Anta at observasjonene (i sekunder) er gitt ved:

X_j:	45	48	61	52	48	63	52	54	50	58	
Y_j:	47	59	43	50	45	45	49	41	47	52	50

Vi vil basere dataanalysen på følgende modell:

To-utvalgsmodellen :

$X_1, X_2, \ldots, X_{n_1}, Y_1, Y_2, \ldots, Y_{n_2}$ er uavhengige stokastiske variable med forventninger henholdsvis

$$EX_j = \mu_1 \qquad EY_j = \mu_2$$

og varianser henholdsvis

$$varX_j = \sigma_1^2 \qquad varY_j = \sigma_2^2$$

Vi ser at to-utvalgsmodellen er sammensatt av to enkle målemodeller, en for hver gruppe, med tilleggsantakelsen at observasjonene i de to gruppene er uavhengige. Sammenligner vi modellen for dette eksemplet med den i Eksempel 1, ser vi at her oppfattes observasjonene som målinger av et felles forventet prestasjonsnivå for hver metode. De individuelle variasjoner fra arbeider til arbeider er nå å oppfatte som del av de tilfeldige variasjonene omkring dette nivået. En vil derfor vente at variansene σ_1^2 og σ_2^2 vil være betydelig større her enn i Eksempel 1. Vi vil være interessert i å estimere forventet tidsforskjell mellom de to metodene, i denne modellen uttrykt ved $\delta = \mu_1 - \mu_2$. Siden $\overline{X}$ og $\overline{Y}$ er forventningsrette estimatorer for henholdsvis μ_1 og μ_2, ser vi at

$$\hat{\delta} = \overline{X} - \overline{Y} = \frac{1}{n_1}\sum_{j=1}^{n_1} X_j - \frac{1}{n_2}\sum_{j=1}^{n_2} Y_j$$

er forventningsrett estimator for $\delta = \mu_1 - \mu_2$. Estimatoren har varians

$$var\hat{\delta} = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} = (\frac{1}{n_1} + \frac{1}{n_2})\sigma^2$$

Den siste likheten forutsetter at variansen i de to gruppene er like, dvs. $\sigma_1^2 = \sigma_2^2 = \sigma^2$, en antakelse som ofte gjøres i praksis, delvis fordi teorien blir enklere. I dette tilfellet kan nemlig

$$S^2 = \frac{1}{n_1 + n_2 - 2}(\sum_{j=1}^{n_1} (X_j - \overline{X})^2 + \sum_{j=1}^{n_2} (Y_j - \overline{Y})^2)$$

brukes som forventningsrett estimator for fellesvariansen σ^2, og standardavviket til estimatoren estimeres med $S(\hat{\delta})$, der S erstatter σ i formelen for $\sigma(\hat{\delta})$. Estimert δ med tilhørende standardavvik blir derfor

$$\overline{X} - \overline{Y} \pm S \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

Med de gitte observasjonene får vi $\overline{X} = 53.1$, $\overline{Y} = 48.0$ og $S = 5.40$, slik at vår rapport om differensen mellom forventningene δ blir 5.10 ± 2.36. Sammenholdes den estimerte reduksjon i forbrukt tid med de estimerte standardavvik, ser det ut til at reduksjonen er signifikant, dvs. neppe kan skyldes tilfeldigheter.

Vi kan heller ikke i denne modellen gi eksakte sannsynlighetsutsagn om estimeringsfeilen uten ytterligere forutsetninger. Dersom vi er villige til å anta at observasjonene er normalfordelte, kan man ta utgangspunkt i den stokastiske variable

$$T = \frac{\hat{\delta} - \delta}{S \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

som da er t-fordelt med $n_1 + n_2 - 2$ frihetsgrader. Ved hjelp av en t-tabell kan vi derfor lage sannsynlighetsutsagn. Eksempelvis med 19 frihetsgrader, er sannsynligheten for en estimeringsfeil på høyst 2.09 ganger estimert standardavvik lik 0.95 (se Tabell C.9).

Også for denne modellen er det aktuelt å teste hypotesen $\delta = \delta_0$, der δ_0 er et spesifisert tall, eksempelvis null. En mulig testobservator er T, der vi for δ bruker den spesifiserte δ_0. Under forutsetning av normalfordelte observasjoner vil denne T, dersom nullhypotesen er riktig, være t-fordelt med $n_1 + n_2 - 2$ frihetsgrader, og ved hjelp av t-tabellen kan vi derfor etablere tester med ønsket signifikansnivå, eventuelt beregne P-verdier.

Merknad. Teorien ovenfor forutsetter at vi er villig til å anta at variansene σ_1^2 og σ_2^2 er like. Uten denne antakelsen må vi estimere variansene i hver gruppe for seg:

$$S_1^2 = \frac{1}{n_1 - 1} \sum_{j=1}^{n_1} (X_j - \overline{X})^2 \qquad S_2^2 = \frac{1}{n_2 - 1} \sum_{j=1}^{n_2} (Y_j - \overline{Y})^2$$

og deretter beregne en estimator for $\sigma(\hat{\delta})$ der σ_1 og σ_2 erstattes med henholdsvis S_1 og S_2. Teorien er nå langt mer komplisert, men gode tilnærmingsmetoder er implementert i programvare, slik at brukeren ikke behøver å gjøre antakelsen om like varianser, dersom denne er urealistisk.

```
 >> READ 'eks11.2' X Y
>> TWOSAMPLE X Y; POOLED S.
      N   Mean  StDev  SE Mean
 X   10  53.10   5.92      1.9
 Y   11  48.00   4.90      1.5
Pooled Standard Deviation = 5.40
Standard Error Difference of Means = 2.36
95 % Conf.Int. for EX - EY : (0.2, 10.0)
T-test EX = EY (VS. NEQ): T=2.16 DF=19
P=0.044
```

En typisk utskrift fra en programpakke er gitt ovenfor, og gir i hovedsak de tidligere omtalte størrelser. De individuelle standardavvik er ikke så mye forskjellige at antakelsen om felles verdi er urimelig. Sist i utskriften er resultatet av T-testen for $H_0 : \delta = 0$ mot $H_A : \delta \neq 0$. Dersom vi krever 5% signifikansnivå, kan vi forkaste hypotesen om lik forventet tid ($P < 0.05$). For begge modellene ovenfor kan det vises, dersom vi antar at observasjonene er normalfordelte, at de metoder som er foreslått er optimale i en viss forstand for inferens om δ. I mange situasjoner vil det imidlertid være urealistisk å anta normalitet. Det gjelder spesielt i situasjoner der det er en viss mulighet for “ville”observasjoner. I eksemplene ovenfor kan det tenkes at slike observasjoner oppstår ved et arbeidsuhell hos en arbeider, som ikke er forårsaket av den arbeidsmetode som brukes. Det er likevel ikke tilrådelig å luke ut eventuelt avvikende observasjoner på rent subjektiv grunnlag, kan hende skyldes avviket likevel metoden. Det finnes imidlertid alternative metoder for sammenligning av to framgangsmåter, med gunstige egenskaper, som bl.a. unngår at “ville”observasjoner får avgjørende innflytelse på de konklusjoner som trekkes.

For konstant-effekt modellen i Eksempel 1, vil medianen av differensene være en alternativ estimator for δ. En alternativ testmetode vil være Wilcoxon-testen for parrede observasjoner. For modellen med to forventninger i Eksempel 2, vil differensen mellom medianene være en alternativ estimator for σ, mens Wilcoxon testen for to utvalg kan være en alternativ testmetode. De nevnte testmetoder er basert på ranger, og er beskrevet i Kapittel 10.4. De forutsetter ikke normalitet, og kan rettferdiggjøres i mer generelle situasjoner enn de som er beskrevet i Kapittel 10.4. Selv når observasjonene er normalfordelte, så taper en lite i forhold til de "optimale" metodene, mens gevinsten kan være betydelig for store avvik fra normalitet.

11.3 En-faktor modellen

En møter ofte situasjoner der en ønsker å sammenligne mer enn to grupper, definert ved flere kategorier for en faktor, som vi nedenfor betegner A.

Eksempel 3 : Produksjonsmetoder
En bedrift vurderer i alt $r = 3$ ulike produksjonsmetoder for et produkt. Faktoren A er her produksjonsmetode. For å sammenligne metodene m.h.t. kvaliteten av det ferdige produkt, produseres 5 enheter med hver metode og kvaliteten av disse tallfestes. Resultatet ble:

Produksjonsmetode nr. 1 :	4.7	3.5	3.3	4.2	3.6
Produksjonsmetode nr. 2 :	3.2	4.2	3.3	3.9	3.0
Produksjonsmetode nr. 3 :	3.1	2.9	2.2	3.0	2.8

Vi skal bruke følgende generelle notasjon : Det produseres n_i artikler med metode nr. i $(i = 1, 2, \ldots, r)$, og Y_{ij} er målt kvalitet av den j'te produserte artikkel med metode nr.i. $(j = 1, 2, \ldots, n_i; i = 1, 2, \ldots, r)$. Vi vil basere dataanalysen på en modell der vi antar at alle disse stokastiske variable er uavhengige, og der forventningen og variansen er den samme for observasjonene med samme metode, men muligens ulik fra metode til metode, dvs.

$$EY_{ij} = \mu_i \quad varY_{ij} = \sigma_i^2 \;\; j = 1, 2, \ldots, n_i.$$

Parameteren μ_i er altså forventet kvalitet med metode nr.i. Ofte kan det være rimelig å anta at variansen er den samme for de ulike metodene, kall i så fall den felles verdi for σ^2.

Som estimator for kvalitetsnivåene for de ulike metodene, kan vi bruke de respektive gjennomsnitt

$$\overline{Y}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} Y_{ij} \quad j = 1, 2, \ldots, n_i,$$

som er forventningsrette med varianser henholdsvis σ_i^2/n_i. Variansene σ_i^2 for hver metode kan vi estimere på vanlig måte, kall estimatorene S_i^2. Har vi antatt en felles σ^2, kan vi utnytte alle observasjonene ved estimeringen. En forventningsrett estimator for σ^2 er da

$$S^2 = \frac{1}{n-r}\sum_{i=1}^{r}\sum_{j=1}^{n_i}(Y_{ij}-\overline{Y}_i)^2 = \frac{1}{n-r}\sum_{i=1}^{r}(n_i-1)S_i^2$$

der $n = \sum_{i=1}^{r} n_i$ er det totale antall observasjoner. Vi kan rapportere forventet kvalitet for hver metode slik

$$\overline{Y}_i \pm \frac{S}{\sqrt{n_i}} \quad i = 1, 2, \ldots, r.$$

Vi kan være interessert i å vurdere eventuelle forskjeller i forventet kvalitet. I en situasjon med mer enn to forventninger, kan vi sammenligne to og to, og rapportere forskjeller som i Eksempel 2. For metode nr.i og nr.k :

$$\overline{Y}_i - \overline{Y}_k \pm S\sqrt{\frac{1}{n_i}+\frac{1}{n_k}}$$

Eksakte konfidensintervaller kan i begge tilfeller lages med sikkerhetsfaktor hentet fra t-tabellen med $n-r$ frihetsgrader (forutsetter normalfordelte observasjoner).

Ved sammenligning av en rekke forventninger kan det være lurt å ta utgangspunkt i hypotesen at alle forventningene er like, dvs.

$$H_A : \mu_1 = \mu_2 = \cdots = \mu_r.$$

og teste den mot alternativet at minst to forventninger er ulike. Hvis denne hypotesen ikke forkastes, er det liten grunn til å legge vekt på gruppeforskjeller, de kan like godt skyldes tilfeldigheter. Vi forutsetter at alle observasjonene har samme varians σ^2. Uten denne antakelsen er det vanskelig å trekke generelle konklusjoner. Dersom nullhypotesen H_A var riktig, kunne vi bruke gjennomsnittet av alle observasjonene

$$\overline{Y} = \frac{1}{n}\sum_{i=1}^{r}\sum_{j=1}^{n_i}Y_{ij} = \frac{1}{n}\sum_{i=1}^{r}n_i\overline{Y}_i$$

til å anslå den felles forventning. Dersom gruppegjennomsnittene $\overline{Y}_1, \overline{Y}_2, \ldots, \overline{Y}_r$ avviker lite fra det totale gjennomsnitt $\overline{Y}$, er det ingen grunn til å forkaste nullhypotesen, mens store avvik tyder på at nullhypotesen ikke er riktig. Et mål som sammenfatter graden av avvik er

$$Q_A = \sum_{i=1}^{r} n_i(\overline{Y}_i - \overline{Y})^2$$

Her er n_i brukt som vekter i summen, bl.a. fordi et bestemt avvik basert på et stort antall observasjoner bør tillegges større vekt enn et tilsvarende basert på et lite antall observasjoner. Dersom Q_A er stor gir dette grunn til å forkaste nullhypotesen. Stor i forhold til hva? Vi må åpenbart se Q_A i forhold til den naturlige tilfeldige variasjon anslått med S^2. Betrakt kvadratsummene

$$Q_0 = \sum_{i=1}^{r}\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y_i})^2 \qquad Q = \sum_{i=1}^{r}\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y})^2$$

som måler observasjonenes variasjon omkring sine respektive gruppegjennomsnitt og omkring det totale gjennomsnitt. Det kan vises at

$$Q = Q_A + Q_0$$

som tolkes slik:

$$\begin{aligned} \text{Total variasjon} \quad &= \quad \text{variasjon forklart ved faktor } A \\ &+ \quad \text{uforklart (tilfeldig) variasjon.} \end{aligned}$$

Vi merker oss at

$$S^2 = \frac{Q_0}{n - r}.$$

Her er $n - r$ det såkalte frihetsgradtall knyttet til Q_0. For å gjøre Q_A sammenlignbar med S^2, må vi også dividere Q_A med sitt frihetsgradtall, som er $r - 1$. En egnet testobservator er derfor forholdet [2]

$$F = \frac{Q_A/(r-1)}{Q_0/(n-r)}$$

Dersom forholdet F er stort gir dette grunn til å forkaste nullhypotesen om at alle forventningene er like. Dette er en såkalt F-*observator*. Under forutsetning av normalfordelte observasjoner er sannsynlighetsfordelingen til F,

[2]Motivasjon av F : Ved å bruke forhold og ikke differens oppnår vi at testobservatoren ikke avhenger av eventuelle endringer av måleenheten, f.eks. fra gram til kilo, centimeter til meter. Divisjon med frihetsgradtallene er rimelig fordi $EQ_A/(r-1) \geq \sigma^2$, med likhet hvis og bare hvis nullhypotesen er riktig, mens $EQ_0/(n-r) = \sigma^2$ uansett om nullhypotesen er riktig eller ikke, noe som samtidig indikerer rimeligheten av å forkaste H_A når F er stor.

dersom nullhypotesen er riktig, en såkalt *Fisherfordeling* eller kortere en F-*fordeling*. Denne fordeling dukker også opp i en rekke andre sammenhenger, den er grundig studert av statistikere, og det er utarbeidet tabeller som gir arealer under F-kurver, se Tabell C.10 og C.11 i Appendiks C. Ved hjelp av slike tabeller kan vi lage tester tilpasset et gitt signifikansnivå, eventuelt beregne P-verdier. I den foreliggende situasjon dreier det seg om F-fordelingen med frihetsgradtall $(r-1, n-r)$. La oss se på utskrift fra en analyse av dataene ovenfor. [3]

```
 >> READ 'eks11.3' Y A
>> ANOVA Y A ; MEANS A.
 Analysis of Variance Table
 Source    DF          SS      MS      F       P
 Factor A   2       2.929   1.465   6.15   0.015
 Error     12       2.860   0.238
 Total     14       5.789

A N  Mean StDev 95% Conf.Int.Means (Common S=0.488)
1 5 3.860 0.577                (-------*-------)
2 5 3.520 0.507           (-------*-------)
3 5 2.800 0.354 (-------*-------)
                -+---------+---------+---------+-
                2.40      3.00      3.60      4.20
```

Utskriften har i hovedsak gitt oss størrelsene nevnt ovenfor, og starter med den såkalte *variansanalyse-tabellen* (ANOVA-tabellen). I denne har vi kvadratsummer (SS=Sum of Squares), frihetsgradtall (DF=Degrees of Freedom) og forholdene (MS=SS/DF=Mean Sum of Squares) og beregnet $F = 6.15$ med tilhørende $P = 0.015$. Vi ser at vi kan forkaste hypotesen om like forventede kvaliteter med 5% signifikansnivå, men ikke med 1% signifikansnivå.

Vi ser forøvrig at antakelsen om felles varians ikke er urimelig, se også Oppgave 18. Dersom hypotesen om like forventninger forkastes, innbyr utskriften til å sammenligne konfidensintervallene parvis, med sikte på rangering av metodene. Metode nr.1 peker seg ut som bedre enn metode nr.3, mens metode nr.2 hverken kan påstås bedre enn nr.3 eller dårligere enn nr.1.

[3] Dataene leses fra en fil med kolonnefelter, ett for responsen Y, og ett for faktoren A, der kodetall 1,2,3 brukes for kategoriene.

Det er ingen grunn til å ekskludere metode nr.2 fra den videre vurdering, spesielt ikke dersom den er billigere enn metode nr.1. Den observerte forskjell kan godt skyldes tilfeldigheter, og man må i alle fall vurdere om den estimerte forskjell har praktisk betydning. I denne forbindelse ville det kanskje være bedre å estimere alle forventede forskjeller med tilhørende feilmarginer, etter formelen ovenfor.

Det er imidlertid et problem knyttet til alle slike parvise betraktninger, idet vi ikke har noen totalgaranti mht. risiko for feilkonklusjoner for de $r(r-1)/2$ sammenligninger som gjøres (her 3). Det finnes teorier for *multiple sammenligninger*, som kan gi slike garantier, som medfører bruk av noe større sikkerhetsfaktorer enn t-tabellen foreskriver. En rekke forslag til metode for multippel sammenligning tilbys (Fisher, Tukey, Dunnett etc.), som avviker noe mht. hvilke garantier som gis. Slike metoder gis ofte som opsjoner i aktuell programvare for variansanalyse.

11.4 To-faktor modeller

Vi utvider nå diskusjonen til observasjoner gruppert ved to faktorer, nedenfor kalt A og B.

Eksempel 4 : Produksjonsmetoder og råstoff
I Eksempel 3 studerte vi virkningen av ulike produksjonsmetoder på kvaliteten av et produkt. Anta at det også er ulike muligheter m.h.t. bruk av råstoff. Vi har da to faktorer, produksjonsmetode (A) og råstoff (B). Anta at det er r=3 produksjonsmetoder og s=4 ulike råstoffkvaliteter, vi har da i alt $r \cdot s = 12$ mulige faktorkombinasjoner. La oss i første omgang tenke oss at vi produserer en artikkel for hver faktorkombinasjon, og måler kvaliteten av hver av disse. La Y_{ij} være kvaliteten av artikkelen produsert med metode nr.i med råstoff nr.j, ($i = 1, 2, \ldots, r; j = 1, 2, \ldots, s$). Anta at resultatene ble som følger:

Y_{ij}	j=1	2	3	4	Gj.snitt
i= 1	5.4	4.3	3.7	3.2	4.15
2	6.3	6.0	5.7	4.5	5.63
3	5.3	5.8	4.8	4.0	4.98
Gj.snitt	5.67	5.37	4.73	3.90	4.92

Vi har her beregnet den gjennomsnittlige kvalitet som er observert for de ulike metoder (linjevis) og de ulike råstoffer (kolonnevis), samt det totale gjennomsnitt, som her ble 4.92. Det kan se ut som metode nr.2 og råstoff nr.1 er den beste kombinasjonen. Vi trenger imidlertid er nærmere vurdering av usikkerheten ved en slik konklusjon.

Vi vil lage en modell for eksperimentet ovenfor som tar omsyn til at de ulike produksjonsmetodene kan gi ulik kvalitet for ulike typer råstoff, dvs. at forventet kvalitet EY_{ij} er en funksjon både av i og j. La oss skrive $EY_{ij} = \mu_{ij}$. Nå kan det være rimelig å anta at μ_{ij} kan splittes opp i to komponenter, en som skyldes produksjonsmåte og en som skyldes råstoff, og at disse to faktorene virker additivt, dvs. at vi kan skrive

$$EY_{ij} = \mu + \alpha_i + \beta_j \quad (i = 1, 2, \ldots, r; j = 1, 2, \ldots, s)$$

Her er μ et gitt kvalitetsnivå, mens α_i kan oppfattes som et tillegg (evt. fradrag) i kvalitet som skyldes produksjonsmåte nr.i, og som gjør seg gjeldende uansett råstoff, og β_j kan oppfattes som et tillegg i forventet kvalitet som skyldes råstoff nr.j, og som gjør seg gjeldende uansett produksjonsmåte. Vi kan uten tap av generalitet anta $\sum_i \alpha_i = 0$ og $\sum_j \beta_j = 0$. Kvalitets/nivået μ blir da å oppfatte som et gjennomsnitt tatt over de ulike metoder og ulike råstoff. I tillegg til antakelsene ovenfor vil vi anta at alle Y_{ij}'ene er uavhengige med samme varians σ^2.

I modellen er $\mu, \alpha_1, \alpha_2, \ldots, \alpha_r, \beta_1, \beta_2, \ldots, \beta_s$ og σ^2 ukjente parametre. Vi ønsker å estimere disse parametrene, samt teste hypoteser om dem. La oss innføre følgende notasjon for søyle- og radgjennomsnitt:

$$\overline{Y}_{\cdot j} = \frac{1}{r}\sum_{i=1}^{r} Y_{ij}, \quad \overline{Y}_{i\cdot} = \frac{1}{s}\sum_{j=1}^{s} Y_{ij}$$

og for det totale gjennomsnitt

$$\overline{Y} = \frac{1}{rs}\sum_{i=1}^{r}\sum_{j=1}^{s} Y_{ij} = \frac{1}{r}\sum_{i=1}^{r}\overline{Y}_{i\cdot} = \frac{1}{s}\sum_{j=1}^{s}\overline{Y}_{\cdot j}$$

Prikkene angir altså hvilken indeks som er summert bort. Følgende estimatorer virker rimelige for sine respektive parametre

$$\begin{aligned}
\hat{\mu} &= \overline{Y} \\
\hat{\alpha}_i &= \overline{Y}_{i\cdot} - \overline{Y} \quad (i = 1, 2, \ldots, r) \\
\hat{\beta}_j &= \overline{Y}_{\cdot j} - \overline{Y} \quad (j = 1, 2, \ldots, s)
\end{aligned}$$

Som estimator for forventet kvalitet μ_{ij} brukes da

$$\hat{\mu}_{ij} = \hat{\mu} + \hat{\alpha}_i + \hat{\beta}_j = \overline{Y} + (\overline{Y}_{i\cdot} - \overline{Y}) + (\overline{Y}_{\cdot j} - \overline{Y})$$

Det er lett å vise at estimatorene ovenfor alle er forventningsrette for sine respektive parametre. Deres varianser kan også finnes, men de utelates her (Oppgave 15). Størrelsen $\hat{\mu}_{ij}$ kan tenkes brukt som estimator for forventningen μ_{ij}. I lys av situasjonen i eksemplet tolkes denne slik: Vi tar utgangspunkt i det totale gjennomsnitt $\overline{Y}$, til dette legges to korreksjonsfaktorer, den ene korrigerer for at produksjonsmetode nr.i er brukt, den andre korrigerer for at råstoff nr. j er brukt. Differensen

$$Y_{ij} - \hat{\mu}_{ij} = Y_{ij} - \overline{Y}_{i\cdot} - \overline{Y}_{\cdot j} + \overline{Y}$$

kan tolkes som det avvik som ikke lar seg forklare ved den antatte modell, og som tilskrives tilfeldigheter. Betrakt kvadratsummen av alle slike avvik

$$Q_0 = \sum_{i=1}^{r} \sum_{j=1}^{s} (Y_{ij} - \overline{Y}_{i\cdot} - \overline{Y}_{\cdot j} + \overline{Y})^2$$

Det kan vises at $EQ_0 = (r-1) \cdot (s-1)\sigma^2$, slik at

$$S^2 = \frac{1}{(r-1)(s-1)} Q_0$$

er en forventningsrett estimator for σ^2. Som i tilfellet med en-faktormodellen i forrige avsnitt kan vi splitte kvadratsummen for den totale variasjon

$$Q = \sum_{i=1}^{r} \sum_{j=1}^{s} (Y_{ij} - \overline{Y})^2$$

i flere kvadratsummer. Her blir

$$Q = Q_A + Q_B + Q_0$$

der Q_0 er definert ovenfor og

$$Q_A = s \cdot \sum_{i=1}^{r} (\overline{Y}_{i\cdot} - \overline{Y})^2 \qquad Q_B = r \cdot \sum_{j=1}^{s} (\overline{Y}_{\cdot j} - \overline{Y})^2$$

Dette tolkes slik

$$\begin{aligned}\text{Total variasjon} &= \text{variasjon forklart ved faktor } A\\ &+ \text{variasjon forklart ved faktor } B\\ &+ \text{uforklart (tilfeldig) variasjon.}\end{aligned}$$

Dette indikerer metoder til å teste interessante hypoteser. Hypotesen $H_A : \alpha_1 = \alpha_2 = \cdots = \alpha_r = 0$ svarer til at faktor A overhodet ikke innvirker på forventningen, mens $H_B : \beta_1 = \beta_2 = \cdots = \beta_s = 0$ svarer til at faktor B ikke innvirker på forventningen. Den første hypotesen forkastes dersom Q_A er stor i forhold til Q_0, den siste dersom Q_B er stor i forhold til Q_0. Egnede testobservatorer til å teste hypotesene H_A og H_B er henholdsvis

$$F_A = \frac{Q_A/(r-1)}{Q_0/(r-1)(s-1)} \quad F_B = \frac{Q_B/(s-1)}{Q_0/(r-1)(s-1)}$$

som, dersom observasjonene er normalfordelte og de respektive nullhypoteser holder, er F-fordelt med frihetsgradtall henholdsvis
$((r-1),(r-1)\cdot(s-1))$ og $((s-1),(r-1)\cdot(s-1))$.

Ved analysen av de data som er gitt ovenfor, fikk vi følgende variansanalysetabell:

```
>> READ 'eks11.4' Y A B
>> ANOVA Y A B
 Analysis of Variance Table
 Source    DF          SS     MS       F       P
 Factor A   2       4.372  2.186   15.87   0.004
 Factor B   3       5.497  1.832   13.27   0.005
 Error      6       0.828  0.138
 Total     11      10.697
```

Vi ser at begge faktorene A og B er signifikante, og vi kunne i likhet med utskriften i Eksempel 3 be om mer detaljerte opplysninger om hva forskjellen mellom gruppene besto i med sikte på å rangere de ulike produksjonsmetoder og de ulike råstoffer m.h.t. kvalitet. Dette leder til problemstillinger av typen multiple sammenligninger som ble omtalt i forrige avsnitt.

Modellen ovenfor brukes også i situasjoner der vi bare er interessert i den ene faktoren, f.eks. produksjonsmetode, mens den annen faktor er en forstyrrende faktor som det lønner seg å ta omsyn til i analysen. Eksempelvis dersom råstoff varierer fra parti til parti, og vi har bare nok til å produsere tre enheter pr. parti.

Det sentrale poeng ved modellen som er presentert ovenfor, er antakelsen om additive effekter. Det er imidlertid mulig at en faktor i kombinasjon med en annen gir et bidrag til responsen ut over det rent additive, dette kalles *samspill*. Best kjent er kanskje dette fenomen i farmasi, hvor to legemidler (faktorer) som i ulike doseringer hver for seg har gunstig virkning, har en helt annen, kanskje verdiløs eller ødeleggende virkning dersom de brukes sammen. I medisin kalles dette synergistiske effekter.

Muligheten for samspill må ikke overses. Det kan være en kilde til feilkonklusjoner, og kan representere uutnyttede muligheter, se f.eks Kapittel 15.4 om kvalitetsforbedringer.

Dersom samspill ikke kan utelukkes, kreves det en annen modell, og en annen analysemetode enn den vi har presentert ovenfor. Innenfor en slik modell kan man bl. a. teste om samspill ikke er til stede, dette vil imidlertid kreve mer enn en observasjon pr. faktorkombinasjon. La oss gi et kort innblikk i modeller med samspill.

Anta at vi i Eksempel 4 har m observasjoner for hver faktorkombinasjon, og la Y_{ijk} være kvaliteten av artikkel nr.k produsert etter metode nr.i med råstoff nr.j. En modell med samspill mellom faktorene kan skrives slik

$$EY_{ijk} = \mu_{ij} = \mu + \alpha_i + \beta_j + \phi_{ij}$$

der ϕ_{ij} representerer samspillet som en korreksjon til hovedeffektene α_i og β_j til de to faktorene. Den naturlige estimator for μ_{ij} er nå gjennomsnittet $\overline{Y}_{ij}$ av de m observasjonene for faktorkombinasjonen (i, j). De øvrige parametrene estimeres i henhold til oppstillingen

$$\begin{array}{ccccccccc} \hat{\mu}_{ij} & = & \hat{\mu} & + & \hat{\alpha}_i & + & \hat{\beta}_j & + & \hat{\phi}_{ij} \\ \overline{Y}_{ij} & = & \overline{Y} & + & (\overline{Y}_{i\cdot} - \overline{Y}) & + & (\overline{Y}_{\cdot j} - \overline{Y}) & + & (\overline{Y}_{ij} - \overline{Y}_{i\cdot} - \overline{Y}_{\cdot j} + \overline{Y}) \end{array}$$

der prikker angir hvilken indeks som er summert bort i tillegg til k. I likhet med situasjonen uten samspill, kan vi splitte opp total variasjon slik

$$Q = Q_A + Q_B + Q_{AB} + Q_0$$

der det ekstra leddet Q_{AB} er variasjon forklart ved samspill mellom de to faktorene. På tilsvarende måte som ovenfor kan vi nå motivere

$$F_{AB} = \frac{Q_{AB}/(r-1)(s-1)}{Q_0/rs(m-1)}$$

som testobservator for $H_{AB} : \phi_{ij} = 0$, dvs. at det ikke er samspill mellom de to faktorene. Dersom H_{AB} er riktig, er F_{AB} F-fordelt med frihetsgradtall $((r-1)(s-1), rs(m-1))$.

Som ovenfor gis de nødvendige beregninger for å gjennomføre testen i form av en ANOVA-tabell, se Oppgave 17.

11.5 Varianskomponent-modeller

Den teori som er utviklet ovenfor kan generaliseres til modeller for mer kompliserte situasjoner med mer enn to faktorer og ulike former for samspill. Felles for disse modellene er at responsen blir dekomponert etter de faktorer man mener kan forklare denne, først tar en med de såkalte egenvirkningene av hver faktor, deretter tas med samspill mellom ulike faktorer i den grad man mener at disse kan gjøre seg gjeldende. Innenfor den valgte modell kan en så stille spørsmål om disse egenvirkningene og samspillene som en kan få besvart ved å samle inn observasjoner. De nødvendige observasjoner vil kunne avhenge av modellen.

Modeller av dette slaget kalles gjerne *dekomponeringsmodeller*, og ved dekomponering er det aktuelt med to typer komponenter, såkalte *forventningskomponenter og varianskomponenter*. Vi har ovenfor bare sett eksempler på det første. Vi betraktet der de ulike faktorkombinasjoner som gitte. Det kan imidlertid tenkes problemstillinger der faktorkombinasjoner blir bestemt som del av eksperimentet. En slik faktor er det mer rimelig å representere med en varianskomponent. La oss illustrere de to typer komponenter med et eksempel.

Eksempel 5 : Produksjonsmetoder og arbeidere
Anta at ulike produksjonsmetoder (faktor A) blir prøvd ut med arbeidere med ulike ferdigheter (faktor B). I noen situasjoner vil de arbeidere som deltar være gitt, f. eks. alle arbeidere i bedriften, eventuelt de som er villig til å delta. Vi vil anta at det ikke er samspill mellom metode og arbeider. Dette er ikke alltid realistisk, det kan f. eks. tenkes at en god metode av enkelte møtes med skepsis fordi den er ny, og disse gjør en dårlig innsats av den grunn. Vi vil derfor bruke to-faktor modellen beskrevet i avsnitt 11.4. Denne svarer til at responsen Y_{ij} for i'te metode og j'te arbeider dekomponeres slik

$$Y_{ij} = \mu + \alpha_i + \beta_j + \text{tilfeldig variasjon}$$

der α_i og β_j er konstanter, og tilfeldig variasjon er stokastiske variable med forventning null og samme varians σ^2. Inferensen gjelder nå de deltakende personer og en bør være forsiktig med å trekke konklusjoner utover disse. Imidlertid kan det tenkes at vi nettopp ønsker å trekke generelle konklusjoner om ferdigheten til arbeidere for de ulike produksjonsmetodene på grunnlag av observerte ferdigheter for et utvalg av arbeidere. Skal dette gjøres, bør de arbeidere som får prøve seg representere et tilfeldig utvalg fra den populasjon av arbeidere man ønsker å trekke konklusjoner om. I denne situasjon kan det være aktuelt å bruke en litt annen dekomponering enn ovenfor, nemlig

$$Y_{ij} = \mu + \alpha_i + B_j + \text{tilfeldig variasjon}$$

der vi har erstattet konstanten β_j med en stokastisk variabel B_j. Denne skal reflektere det forhold at arbeider nr. j har et bidrag til responsen, utover rent tilfeldige variasjoner, som er stokastisk bestemt, nemlig ved loddtrekningen av denne arbeideren fra populasjonen til å delta i eksperimentet. Det kan være rimelig å anta at alle B_j'ene er uavhengige med samme forventning, som uten tap av generalitet kan antas lik null, og med samme varians σ_0^2. Denne variansen gir uttrykk for variasjon i ferdigheter blant arbeiderne i populasjonen, uansett metode. I denne modellen vil faktor A være en forventningskomponent og faktor B en varianskomponent. Hva den siste angår, vil vi være interessert i å teste om variansen σ_0^2 er null. Dette betyr at systematisk individuell variasjon mellom arbeidere i populasjonen ikke er til stede, og at observerte forskjeller mellom arbeidere i sin helhet kan forklares som tilfeldig variasjon.

Dataanalyse på grunnlag av en modell med varianskomponenter vil i grove trekk kunne gjennomføres som i de tilsvarende modeller med forventningskomponenter dersom modellen er uten samspill. Generelt vil imidlertid dataanalysen kunne arte seg forskjellig i de to situasjonene, forskjellen består i hva vi veier kvadratsummen for effekten opp mot, dvs. nevneren i F-observatoren. I tilfellet med to faktorer er oppbyggingen av F-observatorene og frihetsgradtallene gitt ved:

			Kvadratsum i nevner		
Effekt	Frihetsgrad	Kvadratsum	Alle faste	A fast B tilf.	Begge tilf.
A	$(r-1)$	Q_A	Q_0	Q_{AB}	Q_{AB}
B	$(s-1)$	Q_B	Q_0	Q_0	Q_{AB}
AB	$(r-1)(s-1)$	Q_{AB}	Q_0	Q_0	Q_0

Når vi ønsker å undersøke virkningen av en faktor alene eller i kombinasjon med andre faktorer, vil det eksperiment vi utfører sammen med vår forhåndsviten bestemme modellen. På den annen side vil kunnskap om egenskapene ved ulike modeller være til hjelp ved valg av eksperiment. Innen faget eksperimentplanlegging står dekomponeringsmodeller sentralt i diskusjonen når det dreier seg om observasjoner av typen gradert respons.

Teorien ovenfor er i hovedsak utviklet for situasjoner der observasjonene er tatt i henhold til en plan med et valgt antall observasjoner pr. faktorkombinasjon. I samfunnsfag studeres ofte utvalgsdata fra en populasjon, der en i ettertid grupperer observasjonene, f.eks. etter kjønn og andre karakteristika. En er da interessert i å se på gruppeforskjeller mht. målte variable, og variansanalyse etter forventningskomponentmodellen peker seg ut som en mulig analysemåte. Antall observasjoner av hvert slag er imidlertid nå tilfeldig bestemt, og dette kan være et problem ved analysen.

For det første kan forutsetningen om uavhengige observasjoner være tvilsom, og en variansanalyse er da knapt mer enn en eksplorativ metode. For det andre kan det være et praktisk problem ved at tilgjengelige variansanalyseprogrammer forutsetter en balansert situasjon, dvs. like mange observasjoner pr. faktorkombinasjon. Et alternativ er da å omforme variansanalysemodellen til regresjonsmodell ved bruk av egnede indikatorvariable, og bruke teorien i Kapittel 12.

Det fins fordelingsfrie alternativer til variansanalyse med en og to faktorer, dvs. som ikke forutsetter normalfordelte observasjoner. Mest kjent er Kruskal-Wallis test og Friedmans test som er såkalte rangmetoder, der man erstatter observasjonene med deres rang etter størrelse. Slike metoder er også aktuelle dersom observasjonene selv er på en rangeringsskala, slik tilfellet ofte er i meningsmålinger, bl.a. i markedsforskning.

11.6 Oppgaver

1. Et forskerteam mener å ha laget et gjødningsprodukt (A) tilpasset hveteproduksjon som gir høyere avkastning enn det produkt (B) som fagfolk til nå har anbefalt for formålet. En frittstående institusjon har fått i oppdrag å undersøke om det er grunn til å satse på de nye produktet. Undersøkelsen er lagt opp slik: Et rektangulært jordstykke deles opp i 9 blokker med 2 like store ruter i hver blokk:

   Blokknr.

1	2	3	4	5	6	7	8	9

   Jordstykket sås deretter til med hvete. Innen hver blokk velges så en av rutene som gjødsles med A, mens den andre ruten gjødsles med B. Ved innhøstingen veies opp produsert hvetekvantum i hver av de 18 rutene. Resultatet i kg pr. rute ble:

Blokk nr.:	1	2	3	4	5	6	7	8	9
Gjødsel A:	27.2	27.9	25.8	26.6	25.4	25.1	24.0	23.2	22.5
Gjødsel B:	26.3	26.8	24.5	25.9	26.0	24.6	24.2	22.8	22.2

   (a) Bruk en konstant-effekt modell og estimer forventet forskjell i avling med A og B og rapporter resultatet.

   (b) Vurder sannsynligheten for feilestimering av ulik størrelsesorden. Tyder det observerte resultat på en reell forskjell mellom A og B?

   (c) Utfør en t-test for hypotesen om at A og B er like gode mot alternativet at A er bedre enn B. Hva blir konklusjonen dersom 5% signifikansnivå er valgt?

2. Det er foreslått to ulike produksjonsmåter for et produkt som skal tåle en viss belastning, metode nr. 1 faller noe dyrere enn metode nr. 2, men man mener at nr. 1 gir et bedre resultat. Det prøveproduseres $n_1 = n_2 = 6$ artikler etter de to metodene. Disse ble utsatt for en økende belastning og tålte en belastning i kg på henholdsvis

Metode 1:	13.2	11.7	12.3	12.9	13.4	11.2
Metode 2:	10.8	11.5	12.7	12.5	11.0	12.9

   (a) Gjør rede for valg av modell.

   (b) Estimer forskjellen i forventet kvalitet og rapporter resultatet.

   (c) Vurder sannsynligheten for feilestimering av ulik størrelsesorden. Tyder det observerte resultat på en reell forskjell mellom de to metodene.

(d) Utfør en t-test for hypotesen om at de to metodene er like gode mot alternativet at metode 1 er bedre. Hva blir konklusjonen dersom signifikansnivået er valgt lik 5%?

3. La situasjonen være som beskrevet i Eksempel 1, der vi studerte tidsforbruket ved arbeidsoperasjon med og uten spesialverktøy.

   (a) Lag konfidensintervall konfidensnivå 95% for forskjellen i forventet tid uten og med spesialverktøy.

   (b) Gjennomfør en t-test med signifikansnivå 5% for hypotesen om at forskjellen i forventet tid er null mot alternativet at spesialverktøy medfører reduksjon i forventet tid.

   (c) Kjenner du alternative testmetoder som kan brukes i denne situasjonen?

4. La situasjonen være som i Eksempel 2, der vi studerte tidsforbuket ved arbeidsoperasjon med og uten spesialverktøy.

   (a) Lag konfidensintervall konfidensnivå 95% for forskjellen i forventet tid uten og med spesialverktøy.

   (b) Gjennomfør en t-test med signifikansnivå 5% for hypotesen om at forskjellen i forventet tid er null mot alternativet at spesialverktøy medfører reduksjon i forventet tid.

   (c) Verifiser informasjonen i utskriften i teksten. Kan P-verdien i (b) finnes i utskriften.

   (d) Kjenner du en alternativ testmetode som kan brukes i den foreliggende situasjon?

5. La situasjonen være som i Oppgave 10.10.

   (a) Vurder om en konstant-effekt modell kan brukes her, i så fall

   (b) estimer forventet forskjell i toppfart med og uten Luriol og rapporter resultatet.

   (c) Utfør en t-test for å teste hypotesen om at Luriol er uten effekt mot alternativet at Luriol øker toppfarten. Hva blir konklusjonen dersom 5% signifikansnivå er valgt.

6. Samme eksamen avholdes på to læresteder. Det deltok i alt $n_1 = 10$ studenter fra sted nr. 1 og $n_2 = 8$ studenter fra sted nr. 2. Karakterene ble gitt av de samme sensorer på en skala fra 0 til 9, der 9 er beste karakter. Vi ønsker å teste om det er noen forskjell mht. forventet prestasjonsnivå hos studentene på de to lærestedene. Resultatet ble

Sted nr. 1:	2	4	0	7	6	5	5	4	3	5
Sted nr. 2:	5	6	6	4	3	7	5	2		

(a) Diskuter valg av testmetode.

(b) Er forskjellen signifikant på 5%?

7. For å teste levetiden av to typer batterier (A og B) for bruk i lommeradioer er valgt ut 12 batterier av hver type fra et parti ferske batterier fra de respektive leverandører. 24 radioer er til disposisjon og det velges for sikkerhets skyld tilfeldig hvilke radioer som blir forsynt med henholdsvis A og B batterier. Undersøkelsen ga følgende resultat mht. levetid i timer:

Type A:	7.2	11.4	14.3	13.4	9.3	14.1	16.3	9.5	2.7	6.6	18.2	11.5
Type B:	15.8	11.3	16.2	16.0	18.4	13.9	18.5	9.2	17.4	18.0	13.2	15.3

(a) Tyder tallmaterialet på at forutsetningene for bruk av t-testen for to utvalg er oppfylt?

(b) Foreslå en alternativ testmetode som bør kunne brukes.
Hint: Se Kapittel 10.4.

8. En bedrift har markedsført en ny type deodorantsåpe A, og i reklamen hevdes det at den hindrer svettelukt mer effektivt enn visse andre såper på markedet. En forbrukerorganisasjon har bedt om belegg for denne påstanden, og får tilsendt et kortfattet utdrag av en rapport. Denne beskriver bl. a. et eksperiment med 72 deltakere der, over en testperiode på 6 dager, 36 deltakere brukte A, mens resten brukte den til nå ledende såpe B. Hver dag, umiddelbart før "den daglige vaskble det foretatt en duftprøve (av duft som var 24 timer gammel). Duften ble målt på en skala fra 0 (dårligst) til 8 (best). Rapporten oppgir at gjennomsnittlig duftskår i testperioden var henholdsvis 5.89 for de med såpe A og 5.74 for de med såpe B. Konklusjonen som oppgis er at, med et signifikansnivå på 1%, så er A signifikant bedre enn B. Anta at du er bedt om å komme med en kritisk vurdering av disse opplysningene.

(a) Lag modell for eksperimentet og finn ut hva standardavviket til hver enkelt luktprøve må ha vært, og vurder om dette synes rimelig.

(b) Anta at det ikke er grunn til å tvile på opplysningene. Diskuter hvorvidt den problemformulering som er brukt er relevant i forbrukersammenheng.

9. Et firma vurderer tre nye prosedyrer for å stoppe en prosess som gå ut av kontroll. Det simuleres uhell 15 ganger, og en operatør prøver 5 ganger med hver metode. Tiden til prosessen stoppes ble målt i sekunder, med resultat

Prosedyre nr. 1:	78	63	85	52	75
Prosedyre nr. 2:	70	68	58	63	55
Prosedyre nr. 3:	82	78	67	59	87

(a) Diskuter bruk av enfaktor-modellen i denne situasjonen.

(b) Estimer forventet tid for de tre prosedyrene, og ranger dem dersom det synes rimelig.

(c) Er det grunn til å forkaste hypotesen om at det ikke er noen forskjell mellom de tre prosedyrene mht. forventet tid? Bruk 5% signifikansnivå.

(d) Foreslå en alternativ analysemåte som tar omsyn til rekkefølgen av observasjonene for hver prosedyre.
Kan det fortsatt være elementer som er forstyrrende for analysen?

10. Fire typer piggdekk prøves ved bremseprøver på en glattkjøringsbane. Bilen har samme dekktype på alle fire hjul, og etter 6 prøver med samme type, skiftes over til neste type. Bilen kjøres en gitt hastighet og bremselengden måles i meter. Resultatet ble

Dekktype nr. 1:	32	35	30	41	32	37
Dekktype nr. 2:	47	31	33	42	49	38
Dekktype nr. 3:	27	30	32	24	37	38
Dekktype nr. 4:	26	18	15	29	35	22

(a) Er det grunn til å forkaste hypotesen om at forventet bremselengde er den samme for de fire dekktypene? Bruk 5% signifikansnivå.

(b) Hvis det er grunnlag for det, sammenlign dekktypene parvis, og eventuelt ranger dem.

Anta isteden at det ble brukt to biler og tre prøver for hver dekktype før dekkskift

(c) Endre dette noe på konklusjonene ovenfor?
(De tre første observasjonene er for bil nr.1, de tre siste for bil nr.2)

(d) Er det ting du ville gjort annerledes, evt. andre faktorer som burde vært med i eksperimentopplegget?

11. Et forbrukerkontor vurderer 4 vaskemidler, og har gjennomført prøvevask av ensartet tilgriset tøy, og anbefalt dosering av vaskemidlet. Resultatet av vasken vurderes på en skala etter gitte kriterier. Det er brukt 3 ulike maskiner i testen, og for at ikke en enkelt maskins effektivitet skal påvirke resultatet, er det utført en vask for hver kombinasjon vaskemiddel/maskin. Resultater ble

	Vaskemiddel			
	1	2	3	4
Maskin 1	15	17	18	12
Maskin 2	13	16	20	7
Maskin 3	21	22	25	19

(a) Utfør en variansanalyse med to faktorer.

(b) Test hypotesen om at vaskemidlene er like gode.

(c) Ranger vaskemidlene (høy skår best).

(d) Hva ville konklusjonen ha blitt med variansanalyse med en faktor, dvs. vi ser bort fra hvilken maskin som ble brukt.

(e) Hva slags analyse ville du foretatt dersom det forelå to observasjoner pr kombinasjon vaskemiddel/maskin for (i) samme dosering (ii) to ulike doseringer.

12. La situasjonen være som i Oppgave 9 men anta isteden at fire operatører A, B, C og D får prøve alle tre prosedyrene og at det foreligger en observasjon for hver kombinasjon menneske/prosedyre. Anta at resultatet ble

	A	B	C	D
Prosedyre nr. 1	76	55	70	65
Prosedyre nr. 2	70	58	62	52
Prosedyre nr. 3	88	72	75	61

(a) Diskuter muligheten for å bruke en to-faktor modell uten samspill.

(b) Estimer parametrene i denne modellen, og lag en ANOVA-tabell.

(c) Test hypotesen om at prosedyrene er like mht. forventet tid.

Anta vi har 5 observasjoner for hver kombinasjon prosedyre og operatør.

(d) Hvilke nye muligheter for analyse åpner seg nå?

(e) Diskuter sider ved problemet som ikke er belyst ved forventet tid.

13. Forklar at det i konstant-effekt modellen er nok å anta at vi har n uavhengige observasjonspar med samme kovarians.

14. Situasjonen i Eksempel 2 kan også studeres ut fra teorien for variansanalyse med en faktor. Vis at resultatet av analysen blir det samme ved

(a) å utføre analysen.

(b) å påvise at $F = T^2$, og deretter sammenligne F-fordelingen med $(1, n-2)$ frihetsgrader med t-fordelingen med $n-2$ frihetsgrader.

15. Vis at estimatorene i to-faktormodellene er forventningsrette og beregn deres varianser.

16. Anta at vi i toutvalgs-modellen i Eksempel 2 kjente variansene σ_1^2 og σ_2^2 i de to gruppene,

(a) Hvordan ville du lage konfidensintervaller og utføre testingen dersom σ_1^2 og σ_2^2 er (i) like (ii) forskjellige ?

(b) Vis at dersom vi bare har råd til å ta n observasjoner i alt, så er det optimalt å velge

$$n_1 = n\frac{\sigma_1}{\sigma_1 + \sigma_2} \qquad n_2 = n\frac{\sigma_2}{\sigma_1 + \sigma_2}$$

Hint : Ulikhetene i Oppgave 5.40 kan nyttes.

(c) Undersøk tilgjengelig programvare mht. muligheten for å lage konfidensintervaller og utføre testingen dersom variansene er ukjente, men antatt forskjellige.

17. La situasjonen være som i Eksempel 4, men anta at vi har to observasjoner pr. kombinasjon produksjonsmetode og råstoff med resultat:

	Råstoff nr.			
	1	2	3	4
Metode 1	5.4 4.8	4.3 4.9	3.7 3.8	3.2 3.7
2	6.3 6.0	6.0 5.8	5.7 5.2	4.5 4.9
3	5.3 5.5	5.8 5.4	4.8 4.5	4.0 4.8

(a) Tolk utskriften nedenfor. Er det samspill mellom faktorene?

(b) Hvordan vil du gjøre analysen dersom vi antar fravær av samspill?

(c) Forsøk å finne programvare som gjør variansanalyse, og som gir supplerende informasjon: multiple sammenligninger osv.

```
>> READ 'eks11.4b' Y A B
>> ANOVA Y A B A*B
 Analysis of Variance Table
 Source      DF         SS     MS       F      P
 Factor A     2      7.106  3.553   34.80  0.000
 Factor B     3      7.385  2.462   24.11  0.000
 Factor A*B   6      0.514  0.086    0.84  0.563
 Error       12      1.225  0.102
 Total       23     16.230
```

18. Nedenfor er en utskrift fra en sjekk om like varianser i ulike grupper for tallmaterialet i Eksempel 3. Hva forteller denne?

```
>> Read 'Eks11.3' Y A
>> Variance check Y A; Confidence 95.0.
Homogeneity of Variance
Confidence intervals standard deviations
(Bonferroni)
   Lower     Sigma     Upper   n  Factor Levels
 0.311863  0.577061  2.22105  5  1
 0.273974  0.506952  1.95120  5  2
 0.191072  0.353553  1.36079  5  3
Bartlett's Test (assumes normality): P-value :
0.655
Levene's Test (also for non-normal): P-value :
0.673
```

Kapittel 12

Regresjonsanalyse

12.1 Lineære forklaringsmodeller

Vi er ofte interesserte i å studere sammenhengen mellom en variabel Y og en eller flere andre variable $X_1, X_2, \ldots$. Her vil Y omtales som *den avhengige variable* og de øvrige som *forklaringsvariable* (av noen misvisende kalt uavhengige variable). Formålet kan variere: Å se hvordan disse samvarierer med Y, kan forklare Y eller evt. predikere Y.

Det vil i mange situasjoner være formålstjenlig å anta tilnærmet lineær sammenheng mellom den avhengige og de forklarende variable:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_r X_r + U$$

Her er den avhengig variable Y uttrykt som en lineær funksjon av r forklaringsvariable $X_1, X_2, \ldots, X_r$ pluss et restledd U, oftest kalt *feilleddet*. Størrelsene $\beta_0, \beta_1, \beta_2, \ldots, \beta_r$ er konstanter som uttrykker betydningen av de respektive forklaringsvariable. β_0 kalles *konstantleddet*, mens $\beta_1, \beta_2, \ldots, \beta_r$ kalles *regresjonskoeffisienter*. Forklaringsvariablene selv omtales ofte som *regressorer*. En innføring i lineære forklaringsmodeller med én (sikker) forklarende variabel fins i Kapittel 8.3 og 8.6, og anbefales repetert før en leser videre.

Eksempel 1 : Avling
Vi ønsker å studere sammenhengen mellom avling (Y) og forklaringsvariablene gjødningskvantum (X_1), temperatur (X_2), fuktighet (X_3) og tid for planting (X_4). Aktuell forklaringsmodell er

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + U$$

Dersom det dreier seg om drivhus vil trolig avling forklares nesten like godt uten å ta i betraktning tidspunktet for planting. Det er da rimelig å utelate denne forklaringsvariablen, bl.a. for å få en enklere modell. Andre forklaringsvariable som kan være aktuelle, er veksttid (X_5), forbrukt kvantum av vern mot skadedyr eller råte (X_6), antall beskjæringer (X_7). Den siste vil antakelig være en indikatorvariabel, dvs. ha verdi enten null eller en.

Feilleddet i modellen omfatter den del av den avhengig variable som ikke kan forklares lineært ved de forklaringsvariable som er benyttet. Det kan tenkes å bestå av flere ulike komponenter. Disse kan være tilfeldig variasjon, dvs. at forklaringsvariablene ikke entydig bestemmer verdien av den avhengig variable. Dette kan skyldes målefeil m.h.t. de variable som inngår, eller at andre faktorer som ikke er tatt med i modellen påvirker resultatet. Det kan også tenkes at de faktorer som inngår selv innvirker på feilleddet, dette er tilfelle dersom den avhengig variable avhenger av faktorene på en mer komplisert måte enn den som modellen gir uttrykk for. I mange situasjoner er en villig til å anta at dette ikke er tilfelle, i hvert fall ikke i avgjørende grad. En modell er typisk mest anvendelig dersom feilleddet er lite i forhold til den forklarende (strukturelle) del.

Med denne antakelsen kan vi gi regresjonskoeffisientene en praktisk fortolkning, idet β_i er den endring av den avhengig variable Y som skyldes at forklaringsvariablen X_i endres med en enhet, mens alle andre forklaringsvariable holdes konstant. Modellen gir da uttrykk for at den marginale endringen i responsen er den samme for en endring på en enhet av faktoren, uansett hvilken verdi av faktoren vi tar utgangspunkt i. Videre uttrykker modellen at de ulike faktorene er additive, og at den marginale endring av responsen som skyldes endring av en faktor ikke avhenger av hvilke verdier de andre faktorene har. Modellen uttrykker med andre ord at det ikke er samspill mellom faktorene.

Betraktningene ovenfor forutsetter at til hver forklarende faktor X_i svarer en forklaringsvariabel, nemlig X_i selv. Imidlertid er antakelsen om linearitet ikke så restriktiv som dette. Det dreier seg om linearitet i koeffisientene $\beta_0, \beta_1, \beta_2, \ldots, \beta_r$, og det er ingenting i veien for å innføre forklaringsvariable som er funksjoner av de faktorer som inngår. Eksempelvis kan det tenkes at den marginale virkningen av en faktor X_i avtar med økende verdi av denne. Det er da aktuelt å prøve logaritmen til X_i eller kvadratroten av X_i som forklaringsvariabel, eventuelt i tillegg til X_i selv. Dersom en og samme faktor inngår i flere forklaringsvariable, mister vi den konkrete fortolkning av regresjonskoeffisientene som vi ga ovenfor. Ved å innføre forklaringsvariable som er en funksjon av to eller flere faktorer, kan vi også få uttrykt visse former for samspill mellom faktorene, men hvordan dette gjøres vil avhenge av omstendighetene ved hver enkelt anvendelse.

Lineære forklaringsmodeller er aktuelle fordi de er spesielt enkle å arbeide med. I mange situasjoner kan de rettferdiggjøres ut fra den aktuelle problemstilling, noen ganger kan de argumenteres med at lite er vunnet ved å bruke en mer realistisk og mer komplisert forklaringsmodell. I noen situasjoner mangler en forhåndsinnsikt til å argumentere for en bestemt forklaringsmodell. Bruk av lineære modeller kan da representere et første forsøk på å skaffe en viss innsikt i situasjonen, med sikte på å lage en realistisk modell. Det er da aktuelt å prøve en rekke forklaringsvariable for å se hvilken kombinasjon av slike som gir best forklaring. Kan hende gir ingen lineær modell tilstrekkelig god forklaring for vårt formål.

Ordet forklaring i forbindelse med bruk av lineære modeller må ikke tas altfor bokstavelig. Slike modeller pretenderer å gi en viss innsikt, ikke nødvendigvis full innsikt. Den bruk vi kan gjøre av en lineær forklaringsmodell, vil avhenge av de antakelser vi er villige til å gjøre om de størrelser som inngår i modellen, samt i hvilken grad vi "tror på"modellen.

Dersom vi i en gitt situasjon studerer en avhengig variabel Y, og har bestemt oss for en lineær forklaringsmodell med r forklaringsvariable
$X_1, X_2, \ldots, X_r$, vil de koeffisienter $\beta_0, \beta_1, \beta_2, \ldots, \beta_r$ som gir brukbar forklaring som regel være ukjente. For å få informasjon om de ukjente elementene i modellen, kan vi observere den avhengig variable for ulike verdikombinasjoner av forklaringsvariablene. Et slikt tallmateriale analyseres gjerne ved en såkalt *regresjonsanalyse.*

Formålet med en regresjonsanalyse er i første omgang å fastlegge regresjonskoeffisientene i henhold til den lineære forklaringsmodellen på en fornuftig måte. Utover dette ønsker vi å vurdere den fastlagte modellens totale forklaringsevne, samt de ulike forklaringsvariablenes absolutte og relative betydning. Slike betraktninger kan være til hjelp for eventuell senere bruk av modellen, f.eks. til prediksjoner eller som grunnlag for beslutninger.

12.2 Minste kvadraters metode

La oss tenke oss at det foreligger n observasjoner av den avhengig variable for ulike verdikombinasjoner av forklaringsvariablene, slik

$$
\begin{array}{ccccc}
Y_1 & X_{11} & X_{12} & \dots & X_{1r} \\
Y_2 & X_{21} & X_{22} & \dots & X_{2r} \\
\vdots & \vdots & \vdots & & \vdots \\
Y_n & X_{n1} & X_{n2} & \dots & X_{nr}
\end{array}
$$

Her betegner Y_i den i'te observasjon av den avhengig variable, mens X_{ij} er den i'te observasjon av forklaringsvariabel nr. j. For at data overhodet skal være til hjelp, viser det seg at antall observasjoner n må være minst lik $r+1$. I praksis blir som regel regresjonskoeffisientene i en lineær forklaringsmodell fastlagt ved *minste kvadraters metode*, dvs. $\beta_0, \beta_1, \beta_2, \dots, \beta_r$ blir bestemt slik at

$$\sum_{i=1}^{n} [Y_i - (\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_r X_{ir})]^2$$

er minst mulig. Dette innebærer at forklaringsmodellen blir fastlagt slik at kvadratsummen av de beregnede feilledd er minst mulig [1], dvs. gir gjennomgående god forklaring for de observerte data. De koeffisienter som fastlegges på denne måten betegner vi $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_r$. Formeluttrykkene er relativt kompliserte, med mindre man innfører matrisenotasjon. Da er de relativt enkle, og kan lett implementeres i programvare [2]. Programvare for slik regresjonsanalyse er idag lett tilgjengelig for brukere, og disse gir i varierende grad opplysninger for de formål som er nevnt sist i forrige avsnitt.

I praksis blir ofte resultatet av en regresjonsanalyse presentert slik:

$$
\begin{array}{ccccccccccc}
\hat{Y} & = & \hat{\beta}_0 & + & \hat{\beta}_1 X_1 & + & \hat{\beta}_2 X_2 & + & \cdots & + & \hat{\beta}_r X_r \quad (\pm S) \\
 & & & & (\pm S_1) & & (\pm S_2) & & & & (\pm S_r)
\end{array}
$$

Uttrykket uten feilledd kalles den beregnede regresjonsligning, der $\hat{Y}$ kalles den beregnede respons. Dette henger sammen med ønsket om at høyresiden i formelen skal gi uttrykk for (anslå) den forventede respons for ulike verdikombinasjoner $X_1, X_2, \dots, X_r$ av forklaringsvariablene. Regresjonsligningen vil i så fall kunne brukes for prediksjonsformål, og vi kaller derfor $\hat{Y}$ ofte *minste kvadraters prediktoren.*

Beregnet respons for de gitte observasjoner blir

[1] Metoden er illustrert grafisk i tilfellet med en forklaringsvariabel i Kapittel 8.3

[2] Vektoren av regresjonskoeffisienter kan skrives $\hat{\beta} = (X'X)^{-1}X'Y$, hvor X er datamatrisen med de forklarende variable, og der konstantleddet er representert ved en førstekolonne med enere og Y er kolonnevektoren med Y_i'ene.

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \hat{\beta}_2 X_{i2} + \cdots + \hat{\beta}_r X_{ir}$$

og forskjellen mellom disse og de tilsvarende observerte responser (de såkalte residualene) blir $\hat{U}_i = Y_i - \hat{Y}_i$. Standardavviket S til disse residualene gir uttrykk for hvor god tilpasning observasjonene har til modellen, og gir også uttrykk for en typisk prediksjonsfeil dersom ligningen brukes til å predikere responsen for en ny verdikombinasjon av de forklarende variable.

Størrelsene $S_1, S_2, \ldots, S_r$ ovenfor er standardfeil knyttet til hver regresjonskoeffisient. Det er nemlig ikke nok å kjenne den fastlagte regresjonskoeffisientens absolutte verdi for å vurdere betydningen av den tilhørende forklaringsvariabel, idet hver må ses i forhold til sin måleenhet, og viser det seg, også i forhold til måleenhetene for de andre forklaringsvariablene. Grovt sagt, jo mindre standardfeilen er i forhold til regresjonskoeffisienten, desto skarpere konklusjoner kan vi trekke om betydningen av vedkommende forklaringsvariabel. Mer presise definisjoner og fortolkninger av størrelsene ovenfor følger i neste avsnitt.

Eksempel 2 : Fyringsutgifter

Betrakt eneboliger der parafin er brukt som eneste oppvarming. Årlig forbruk (Y) antas å avhenge av boareal (X_1) og lengde på fyringssesong (X_2). Vi har følgende 12 observasjoner fra ulike steder i landet, med enheter henholdsvis 1000 liter, m^2 og måneder.

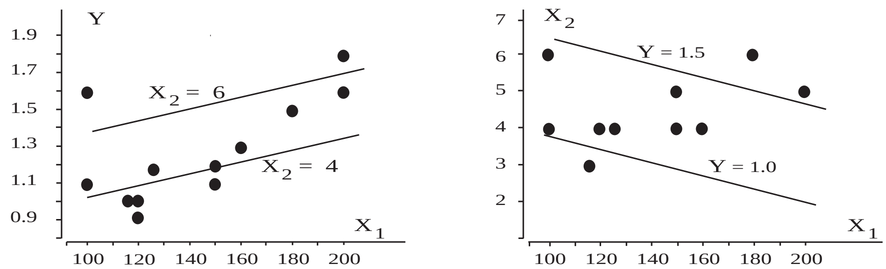
Y	1.2	1.3	1.6	1.0	1.6	1.2	1.1	1.5	1.8	1.0	0.9	1.1
X_1	150	160	100	120	200	125	100	180	200	115	120	150
X_2	5	4	6	4	5	4	4	6	5	3	4	4

Beregning av minste kvadraters prediktoren med tilhørende standardfeil for koeffisientene og standardavvik til feilleddet, ga resultatet:

$$\begin{array}{ccccccc} \hat{Y} & = & -0.0556 & + & 0.00337X_1 & + & 0.1883X_2 \quad (S = 0.167) \\ & & & & (\pm 0.00152) & & (\pm 0.0602) \end{array}$$

Data og regresjonsplanet kan illustreres i et tre-dimensjonalt plott. I to dimensjoner må vi nøye oss med å plotte to variable av gangen, med nivålinjer for utvalgte verdier av den tredje variablen, se Figur 12.1.

I den beregnede regresjonsligningen er regresjonskoeffisienten til X_1 (boareal) svært liten, men ser vi den i forhold til sin standardfeil, og til måleenheten, er den en likevel en viktig komponent i forklaringen av fyringsutgiften, om



Figur 12.1: To-dimensjonale plott

enn ikke så viktig som lengden på fyringssesongen. La oss bruke formelen til å predikere fyringsutgifter for en enebolig på 100 m^2, der fyringssesongen antas å være 5 måneder. Vi får

$$\hat{Y} = -0.0556 + 0.00337 \cdot 100 + 0.1883 \cdot 5 = 1.223$$

Den beregnede S=0.167 indikerer at prediksjonsfeilen kan bli stor. I praksis kan fyringssesongens lengde på et sted variere fra år til år, og skal prediksjonen gjøres før sesongen, ved bruk av forventet sesong, vil dette gi enda større risiko for prediksjonsfeil. Er det andre forklarende variable som kunne være av interesse i dette eksemplet?

Et annet mål for tilpasningen til modellen tar utgangspunkt i sammenhengen:

$$\sum_{i=1}^{n} (Y_i - \overline{Y})^2 = \sum_{i=1}^{n} (\hat{Y}_i - \overline{Y})^2 + \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

som tolkes slik:

$$\begin{aligned} \text{Total variasjon} \quad &= \quad \text{variasjon forklart ved } X_1, X_2, \ldots, X_r \\ & \qquad + \text{ uforklart (tilfeldig) variasjon} \end{aligned}$$

Den såkalte *determinasjonskoeffisienten* definert ved

$$R^2 = \frac{\sum (\hat{Y}_i - \overline{Y})^2}{\sum (Y_i - \overline{Y})^2} = \frac{\text{Forklart variasjon}}{\text{Total variasjon}}$$

er et mål på hvor godt Y-observasjonene lar seg forklare lineært ved X-observasjonene. Vi ser at R^2 er et tall mellom 0 og 1, som er lik 1 hvis og bare hvis $Y_i = \hat{Y}_i$ for alle i, dvs. når Y-observasjonene er fullt ut forklart av X-observasjonene. I eksemplet ovenfor er determinasjonskoeffisienten beregnet til 0.72, dvs. hele 72% av variasjonen i fyringsutgifter er forklart ved boareal og lengde på fyringssesong.

Merknad : Alternativt kan en beregne korrelasjonskoeffisienten, slik den er definert i Kapittel 1 (se også Kapittel 9.5), mellom Y-observasjonene og de tilhørende beregnede $\hat{Y}$, dvs.

$$R = R_{Y\hat{Y}}$$

R kalles den *multiple korrelasjonskoeffisienten* mellom Y og $(X_1, X_2, \ldots, X_r)$. Den har verdi mellom -1 og $+1$, med kvadrat lik determinasjonskoeffisienten (herav notasjonen ovenfor). I tilfellet med en forklarende variabel X, viser det seg at $R_{Y\hat{Y}} = R_{YX}$, dvs. den enkle korrelasjonskoeffisienten mellom Y-observasjonene og X-observasjonene.

Som regel er vi interessert i å trekke generelle konklusjoner om de faktorer som inngår, dvs. konklusjoner som har gyldighet ut over det foreliggende observasjonsmaterialet. Vi blir da ledet til å tenke på det foreliggende materialet som et av mange tenkelige materialer, og det er da naturlig å vurdere observasjonene i lys av en stokastisk modell. Dette gir mulighet for en nærmere vurdering av den analysemetode som anvendes, samt en vurdering av usikkerheten ved de konklusjoner som trekkes.

Vi vil nedenfor redegjøre for den standardmodellen som ligger til grunn for de vanlige regresjonsprogrammer, og se på hvordan de opplysninger slike programmer vanligvis gir, kan tolkes i lys av standardmodellen. Videre vil vi diskutere de enkelte forutsetningene i standardmodellen og antyde ulike omstendigheter der disse kan være urealistiske. En vanlig regresjonsanalyse basert på minste kvadraters metode kan da lett lede til feiltolkninger av observasjonene, men i slike situasjoner kan en ofte gjøre bruk av alternative regresjonsmodeller og analysemetoder.

12.3 Standardregresjonsmodellen

I den linære forklaringsmodellen vil vi gjøre følgende tilleggantakelser:

Vi vil oppfatte forklaringsvariablene $X_1, X_2, \ldots, X_r$ som sikre variable, dvs. ikke underlagt tilfeldigheter [3]. Feilleddet U vil vi imidlertid oppfatte som en stokastisk variabel som uttrykker tilfeldig variasjon. Mer konkret vil

[3]For å markere dette brukes ofte små bokstaver $x_1, x_2, \ldots, x_r$.

vi anta at U har forventning null og en bestemt ukjent varians σ^2. Dette innebærer at Y også blir en stokastisk variabel. Forventningen til Y blir

$$EY = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_r X_r$$

og dens varians σ^2. Her blir altså $\beta_0, \beta_1, \beta_2, \ldots, \beta_r$ og σ^2 ukjente parametre som vi ønsker nærmere kunnskap om.

For dette formål foretar vi n gjentatte observasjoner av den avhengig variable og de tilhørende forklaringsvariable. Vi vil anta at de n feilleddene er uavhengige av hverandre, som er ekvivalent med å anta at de avhengig variable er innbyrdes uavhengige. Med den notasjon vi har brukt før kan vi oppsummere modellen slik:

Standardregresjonsmodellen :
Gitt observasjoner $(Y_i, X_{i1}, X_{i2}, \ldots, X_{ir})$ $i = 1, 2, \ldots, n$ der

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_r X_{ir} + U_i$$

hvor $U_1, U_2, \ldots, U_n$ er uavhengige stokastiske variable med forventning null og varians σ^2.

Vi vil nedenfor vurdere konsekvensen av enda en antakelse, nemlig at feilleddene er normalfordelte, i vår modell betyr dette at de avhengige variable selv er normalfordelte. Denne modellen generaliserer den enkle regresjonsmodellen i Kapittel 8.3 til situasjoner med mer enn en forklaringsvariabel. Notasjonen avviker noe fra den vi brukte der, idet vi her bruker den notasjon som er mest vanlig ved beskrivelse av flere variable.

La $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_r$ være minste kvadraters estimatorene for parametrene $\beta_0, \beta_1, \ldots, \beta_r$. For å sikre at disse er veldefinerte og lar seg beregne, trengs en teknisk antakelse, nemlig at ingen forklaringsvariabel er overflødig ved at dens verdi kan skrives som en lineær funksjon av andre forklaringsvariable. I motsatt fall har vi såkalt *multikolinearitet* som medfører at regresjonskoeffisientene ikke er entydig bestemt. Nesten multikolinearitet er også lite ønskelig da dette gir både beregningsmessige problemer og stor usikkerhet m.h.t. konklusjoner.

Under forutsetningene i standardregresjonsmodellen blir minste kvadraters estimatorene $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_r$ forventningsrette for sine respektive parametre, dvs.

$$E\hat{\beta}_j = \beta_j \qquad j = 0, 1, \ldots, r.$$

Deres respektive varianser kan beregnes, og de har følgende form

$$var\hat{\beta}_j = \frac{\sigma^2}{M_j} \qquad j = 0, 1, \ldots, r.$$

der M_j er en funksjon av de observerte verdier av alle forklaringsvariablene. I tilfellet med en forklaringsvariabel er formelen for variansen til regresjonskoeffisienten gitt i Kapittel 8.3, men allerede for to forklaringsvariable er formlene for beregning av M_j'ene nokså kompliserte med mindre man tar i bruk matrisenotasjon. Det kan imidlertid vises at

$$M_j = (1 - R_j^2) \cdot \sum_{i=1}^{n} (X_{ij} - \overline{X}_j)^2$$

der R_j^2 er determinasjonskoeffisienten ved en regresjon med X_j som avhengig variabel og de øvrige X'er som forklarende variable. Vi ser at $var\hat{\beta}_j$ øker med R_j^2 og multikolinearitet er problematisk dersom $R_j^2 > 0.9$.

Minste kvadraters estimatorene viser seg å være lineære funksjoner av de observerte avhengig variable. Dette betyr at dersom vi er villig til å anta at observasjonene er normalfordelte, så er også disse estimatorene normalfordelte med forventning og varians gitt ovenfor. Dersom observasjonene ikke antas normalfordelte, vil estimatorene likevel være tilnærmet normalfordelte under visse forutsetninger, bl.a. må hverken n eller M_j'ene være for små.

Residualene $\hat{U}_i$ kan oppfattes som et forsøk på å anslå virkelige feilledd U_i, og kan brukes til å estimere variansene til feilleddene. Det kan vises at

$$S^2 = \frac{1}{n - (r+1)} \sum_{i=1}^{n} \hat{U}_i^2$$

er en forventningsrett estimator for σ^2. Merk at $r + 1$ er antall ukjente parametre i regresjonsuttrykket, sml. tilfellet med en forklaringsvariabel. En rimelig estimator for variansen til hver enkelt regresjonskoeffisient får vi ved å erstatte σ^2 med S^2 i formlene for variansene gitt ovenfor.

Når en skal vurdere betydningen av en rekke estimerte regresjonskoeffisienter, trengs et mål for den usikkerhet som må tillegges hver av dem. På bakgrunn av opplysningene ovenfor er det naturlig å betrakte estimert standardavvik, dvs. $S_j = S(\hat{\beta}_j) = S/\sqrt{M_j}$. Brukes vår vanlige rapporteringsmetode oppgis

$$\hat{\beta}_j \quad \pm \quad S(\hat{\beta}_j) \qquad j = 1, 2, \ldots, r.$$

Det er vanskelig å gi presise utsagn om estimeringsfeilen uten å anta at observasjonene er normalfordelte. I dette tilfellet er nemlig

$$T_j = \frac{\hat{\beta}_j - \beta_j}{S(\hat{\beta}_j)}$$

t-fordelt med $n - r - 1$ frihetsgrader. Ved hjelp av en t-tabell kan vi defor gi eksakte sannsynlighetsutsagn om estimeringsfeilen, f.eks. i form av konfidensintervaller for regresjonskoeffisientene.

Det kan også være av interesse å teste hypoteser om regresjonskoeffisientene. Mest aktuell er kanskje $H_0 : \beta_j = 0$, dvs. at forklaringsvariablen X_j overhodet ikke påvirker den avhengig variable. Som testobservator kan vi bruke T_j ovenfor, hvor vi har erstattet β_j med null. Under forutsetning av normalfordelte observasjoner vil denne testobservator, dersom nullhypotesen er riktig, være t-fordelt med $n-r-1$ frihetsgrader. Vi er dermed i stand til å lage tester med et gitt signifikansnivå, eventuelt beregne P-verdier for det observerte resultat.

Vi ser at for å gjøre de betraktninger som er skissert ovenfor, er det nok å kjenne til de estimerte regresjonskoeffisienter $\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_r$ og de tilhørende estimerte standardavvik $S_1, S_2, \ldots, S_r$. I tillegg til dette gir regresjonsprogrammer vanligvis også $T_1, T_2, \ldots, T_r$ og S (evt.S^2), og i noen tilfeller skrives også ut residualene.

Standardregresjonsmodellen gir også muligheter for å vurdere usikkerheten som ligger i å bruke den beregnede regresjonsligning som prediktor. Feilmarginer for prediksjonsfeilen blir gjerne konstruert med utgangspunkt i estimatet S for standardavviket til feilleddet. Grovt sagt vil prediksjonsfeil av størrelse opp til én gang dette standardavviket kunne inntreffe relativt ofte, feil omkring to ganger standardavviket inntreffer også en gang i blant, mens prediksjonsfeil på mer enn tre ganger standardavviket vil inntreffe svært sjelden. Dersom antall observasjoner n er relativt stort i forhold til antall forklaringsvariable, vil vanligvis 68% - 95% regelen gi en viss pekepinn, men ellers er denne nok noe optimistisk. Dersom observasjonene er normalfordelte, er det mulig å lage mer presise sannsynlighetsutsagn om prediksjonsfeilen med utgangspunkt i t-tabellen med $n-r-1$ frihetsgrader.

Standardregresjonsmodellen generaliserer teorien i Kapittel 8.3 fra en til flere forklaringsvariable. De metoder som er foreslått har tilsvarende optimale egenskaper som i tilfellet med en forklaringsvariabel, og vil skal ikke gå i detalj her.

12.4 Diskusjon

Vi vil nå drøfte hver av antakelsene i standardmodellen med sikte på å skape en viss forståelse for de muligheter og begrensninger som er knyttet til regresjonsanalyse:

Antakelsen om at forklaringsvariablene er sikre variable vil først og fremst være aktuell i situasjoner der det utføres et eksperiment med en respons som observeres for ulike verdikombinasjoner av en eller flere forklaringsvariable (stimuli, innsatsfaktorer e.l.) valgt av den som utfører eksperimentet. Vi kaller dette et eksperiment av stimulus-respons typen. Som eksempel ta et drivhuseksperiment der avling observeres for ulike kombinasjoner av gjødningskvantum, temperatur, fuktighet etc.

Antakelsen om at feilleddene er uavhengige med samme varians kan ofte rettferdiggjøres ved at de gjentatte observasjonene av responsen skjer uavhengig av hverandre for de ulike verdikombinasjoner av forklaringsvariablene, men slik at eksperimentbetingelsene ellers er de samme fra gang til gang. Dette er et eksempel på et såkalt (delvis) kontrollert eksperiment. Feilleddet kan da tenkes å omfatte usikkerhet som skyldes målefeil og/eller tilfeldig variasjon som skyldes de faktorer som ligger utenfor vår kontroll. Det er imidlertid grunn til å tenke over antakelsen om samme varians for de ulike verdikombinasjoner. Eksempelvis kan det tenkes at visse kombinasjoner av temperatur og fuktighet gjør plantene mer mottakelige for visse sykdommer enn andre, med økt risiko for sterk reduksjon av avlingen som ikke er til stede ellers.

Antakelsen om at feilleddet er normalfordelt, som er nødvendig for eksakte sannsynlighetsutsagn basert på opplysningene fra regresjonsanalysen, vil kunne brukes dersom vi enten har empirisk erfaring som støtter antakelsen, eller mener at feilleddet skyldes en rekke uavhengige faktorer som hver bidrar lite.

Antakelsen om at forventet respons er en lineær funksjon av forklaringsvariablene er alltid åpen for diskusjon. Selv om dette ikke er rimelig for alle verdikombinasjoner av forklaringsvariablene, eksempelvis for ekstremt høye gjødningskvanta og ekstremt høye og lave temperaturer, kan det være en rimelig antakelse innenfor det verdiområde vi regner med å gjøre bruk av modellen og de konklusjoner som følger av denne. Vi har tidligere antydet at linearitet ikke er så restriktivt som det kan se ut for, idet vi kan betrakte funksjoner av faktorene som forklaringsvariable. I mange situasjoner er det ikke rimelig å anta at observert respons selv kan beskrives med en lineær forklaringsmodell. Typisk eksempler er en rekke såkalte vekstmodeller. For slike situasjoner kan det tenkes alternative forklaringsmodeller som ikke er lineære. Disse blir straks vanskeligere å arbeide med både matematisk og fortolkningsmessig. I noen situasjoner kan det imidlertid tenkes at en størrelse avledet av responsen kan la seg forklare lineært. I så fall brukes denne som avhengig variabel istedenfor responsen selv. Ved eksponensiell vekst er det eksempelvis aktuelt å betrakte logaritmen til responsen som avhengig variabel. Det vil føre for langt å gå i detalj her, fordi de muligheter som byr seg vil avhenge av omstendighetene (se Oppgave 6 og 7).

Mindre avvik fra forutsetningene i standardmodellen vil neppe ha konsekvenser for eventuelle konklusjoner som trekkes på grunnlag av denne. Mindre avvik fra linearitet fanges opp av feilleddet som økt usikkerhet i tillegg til de rent tilfeldige variasjoner. Observasjon av forklaringsvariablene kan ofte innebære en viss usikkerhet i form av målefeil, men så lenge disse er små i forhold til usikkerheten i feilleddet, spiller det liten rolle om de i modellen blir betraktet som sikre variable. Mindre ulikheter i variansen fra observasjon til observasjon spiller også liten rolle. Det samme gjelder mindre avvik fra normalitet, dersom vi velger å bruke denne antakelsen.

Større avvik fra forutsetningene vil det ofte være mulig å avsløre når observasjonene analyseres. Mye kan læres ved å studere residualene, dvs. differensen mellom observert respons og "beregnet forventet respons"ut fra modellen. Ifølge modellen bør disse variere tilfeldig omkring null. De kan imidlertid vise seg å ha særtrekk som gjør det nødvendig å modifisere modellen. Eksempelvis bør det faktum at en eller flere residualer er svært store i tallverdi vies spesiell oppmerksomhet, da modellen har gitt dårlig forklaring av disse observasjonene. Vurderer vi residualene i forhold til estimert standardavvik for feilleddet, får vi en pekepinn på om avvikene kan skyldes tilfeldigheter. Dersom dette ikke synes rimelig, bør en vurdere en rekke muligheter: Avvikene kan skyldes at sannsynlighetsfordelingen til feilleddet er ikke-normal med betydelig risiko for "ville"observasjoner. Det er utviklet metoder til å teste antakelsen om normalitet på grunnlag av residualene. Dersom det er flere avvikende residualer bør en vurdere om disse har visse fellestrekk, eksempelvis om de kan forklares ved en variabel som vi har utelatt i modellen, f.eks. en variabel som er utenfor vår kontroll. Dersom de avvikende responsene er observert etter hverandre i tid, kan en stille spørsmålstegn ved antakelsen om uavhengige observasjoner.

Ovenfor har tankene kretset om et (delvis) kontrollert eksperiment av stimulus-respons typen. Ved et godt planlagt eksperiment av denne typen er mulighetene best for å realisere antakelsene i standardmodellen. I mange situasjoner er det imidlertid ikke mulig å ha full kontroll over alle aktuelle forklaringsvariable og tildele disse sikre verdier. Ofte må man nøye seg med å observere forklaringsvariablene som del av eksperimentet selv. Som eksempel ta et frilandseksperiment hvor riktignok gjødning kan tilsettes i gitte kvanta, mens temperatur og nedbør ikke kan det. Disse forklaringsvariablene kan ikke lenger oppfattes som sikre variable, men må isteden oppfattes som stokastiske variable med verdier som realiseres i løpet av eksperimentet, eksempelvis gjennomsnittsverdier for vekstperioden. I samfunnsvitenskapene vil det som regel alltid være slik at viktige forklaringsvariable er utenfor vår kontroll. Siden deres verdier ikke er gitt på forhånd, må disse representeres ved stokastiske variable i eventuell forklaringsmodell.

Den lineære forklaringsmodellen er fortsatt aktuell, men som vi ser er vi nå utenfor rammen av standardregresjonsmodellen. Heldigvis viser det seg mange av de teoretiske resultater som er utviklet for standardmodellen også gjelder i situasjonen med stokastiske forklaringsvariable under en viktig, men ofte oversett forutsetning, nemlig at forklaringsvariablene er uavhengige av feilleddet. Minste kvadraters estimatorene for regresjonskoeffisientene er fortsatt forventningsrette. Deres varianser er imidlertid noe mer kompliserte enn for standardmodellen. Det kan imidlertid gis teoretiske argumenter for at rapportering av usikkerheten ved estimeringen bør skje betinget gitt de observerte verdier av forklaringsvariablene, dvs. på samme måten som i det sikre tilfellet. Ved testing brukes også de samme metoder som for standardmodellen, idet man kan vise at testobservatorene, under normalitet, har den samme sannsynlighetsfordeling som i det sikre tilfellet. I praksis vil derfor dataanalysen kunne foregå akkurat på samme måte i de to situasjonene, den eneste forskjell er at vi i tilfellet med et kontrollert eksperiment kan velge bestemte verdikombinasjoner av forklaringsvariablene med hensikt for å redusere usikkerheten, i motsetning til å måtte ta dem som de kommer.

Eksempel 3 : Avling og vær

En bonde har de siste åtte årene dyrket samme vekst. Han har hvert år notert seg avlingen i kilo pr. arealenhet, samt nedbør i centimeter og gjennomsnittstemperatur i vekstperioden. Han har hele tiden brukt samme kvantum av en bestemt gjødning, og regner også med at jorden stort sett var av samme kvalitet og fikk samme kultivering. Anta at resultatene ble:

Avling(Y):	70	60	80	80	90	60	70	50
Nedbør (X_1):	10	20	25	20	15	15	30	25
Temperatur (X_2):	25	18	22	22	25	18	16	16

Vi vil gjøre bruk av en regresjonsmodell $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + U$ der vi antar at feilleddet er tilfeldig variasjon som skyldes en rekke uavhengige faktorer som ikke er korrelert hverken med temperatur eller nedbør. Observasjonene gir oss følgende estimerte regresjonsuttrykk

$$\begin{array}{ccccccc} \hat{Y} & = & -39.2 & + & 1.26X_1 & + & 4.15X_2 \\ & & & & (\pm 0.55) & & (\pm 0.96) \end{array}$$

Determinasjonskoeffisienten er beregnet til 0.79, slik at den lineære modellen har gitt god forklaring.

Vi ser begge regresjonskoeffisientene er positive, og vurdert i forhold til de estimerte standardavvik er begge signifikant forskjellig fra null. Dette samsvarer med de tanker bonden hadde på forhånd, for en gitt temperatur vil avlingen gjennomgående øke med nedbøren, for en gitt nedbør vil avlingen gjennomgående øke med temperaturen, dvs. egenvirkningen av begge forklaringsvariablene er positiv. Dersom denne regresjonsligningen brukes til prediksjon vil, dersom vi observerer en nedbør på 20 cm og gjennomsnittstemperatur på 20 grader C, predikert avling bli 68.9 kilo pr. arealenhet. Vi har også beregnet korrelasjonskoeffisientene

$$R_{YX_1} = -0.17 \quad R_{YX_2} = 0.76 \quad R_{X_1X_2} = -0.67$$

Hvordan kan det ha seg at den første er negativ mens den tilsvarende regresjonskoeffisienten er positiv?

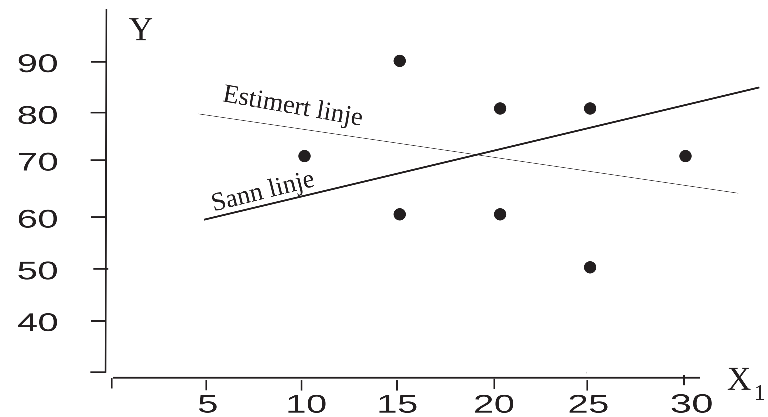
I situasjoner der en eller flere stokastiske forklaringsvariable er korrelert med feilleddet bryter teorien ovenfor sammen, og vanlig regresjonsanalyse basert på minste kvadraters metode kan lede til alvorlige feiltolkninger.

Eksempel 4 : Avling og vær

La situasjonen være som beskrevet i forrige eksempel. Anta at det ikke foreligger opplysninger om temperaturen, og nedbør (X_1) blir brukt som eneste forklaringsvariabel. Observasjonsmaterialet ovenfor gir da følgende regresjonsuttrykk

$$\begin{array}{ccccc} \hat{Y} & = & 76.6 & - & 0.33X_1 \\ & & & & (\pm 0.81) \end{array}$$

Vi ser at regresjonskoeffisienten er negativ, men så pass liten i forhold til standardavviket at vi ledes til å konkludere at nedbør er uten betydning for avlingen. Sett i forhold til resultatet i Eksempel 3, hvor konklusjonen var at avlingen økte med økende nedbør virker dette forvirrende. Følgende betraktninger kan imidlertid bidra til å forstå det som har skjedd: Vår forklaringsmodell er $Y = \beta_0 + \beta_1 X_1 + U$, der β_1 skal gi uttrykk for den systematiske endringen i avling som skyldes en økning i nedbør på en cm når alle andre faktorer ellers er de samme, og U er tilfeldig variasjon. I den grad temperatur (X_2) har betydning for avlingen, vil det komme til uttrykk i feilleddet som tilfeldig variasjon fra år til år. Nå viser nedbør og temperatur erfaringsmessig negativ samvariasjon, mye nedbør som regel sammen med lave temperaturer. Dette betyr at X_1 og U sannsynligvis ikke er uavhengige, men vil ha negativ samvariasjon. La oss illustrere mulige konsekvenser av dette med Figur 12.2.



Figur 12.2: Korrelerte feilledd

I figuren er inntegnet den "sanneregresjonslinjen som er stigende med X_1, dvs. β_1 positiv. På grunn av de forhold som er nevnt ovenfor, vil feilleddene i år med lav (høy) nedbør være gjennomgående positive (negative), dvs. som på figuren. Vi ser at enhver estimeringsmetode basert på tilpasning av en rett linje i spredningsdiagrammet, slik som minste kvadraters metode, vil feilestimere regresjonslinjen. I dette tilfellet vil β_1 bli underestimert, noe som forklarer resultatet ovenfor. Den estimerte linjen gir altså ikke et korrekt bilde av egenvirkningen av X_1 som forklaringsvariabel, men har istedenfor gitt oss et visst inntrykk av virkningen av X_1 i samspill med andre faktorer, her trolig i første rekke temperatur. Den linje som er beregnet gir altså uttrykk for at totalvirkningen av en økning av nedbøren er omtrent null, eller svakt negativ. I lys av resultatene i Eksempel 3 hvor vi fant at egenvirkningen av både nedbør og temperatur var positiv, kan dette tolkes dithen at nedbør utover det normale et bestemt år har en positiv direkte effekt på avlingen, men denne blir mer enn oppveiet av en negativ indirekte effekt som skyldes at mye nedbør gjerne følges av lave temperaturer. Den regresjonslinje som er beregnet her er derfor fortsatt aktuell dersom vi neste år skal predikere avlingen for en gitt nedbørmengde, uten kjennskap til temperaturen, eksempelvis for nedbør på 20 cm predikerer vi avlingen 70.0 kilo pr. arealenhet.

Situasjoner der forklaringsvariablene og feilleddet er korrelerte kan dukke opp i mange sammenhenger og det kan være lett å overse dette. Det som spesielt gjør situasjonen vanskelig, er at vi ofte ikke er i stand til å avgjøre dette ut fra observasjonene selv, f.eks. ved studium av residualene. Vi har tilsynelatende fått god tilpasning, men er likevel blitt lurt. Dette kan ofte være et problem i mange samfunnsfaglige undersøkelser.

En annen omstendighet som kan lede til feilkonklusjoner har vi dersom feilleddene i regresjonsmodellen ikke er innbyrdes uavhengige. Dette vil ofte kunne være tilfelle i situasjoner der observasjonene av responsen er tatt ved suksessive tidspunkter, slik at det er mulighet for en viss tidsavhengighet. Vi vil ta opp problemer av denne art i neste kapittel om tidsrekker.

12.5 Anvendelser

Siden regresjonsanalyse trolig er den mest brukte (og misbrukte) statistiske inferensmetode i praksis, vil vi ta for oss enkelte detaljer ved analysen av et tallmateriale med flere forklarende variable.

Eksempel 5 : Studieprestasjoner
For et tilfeldig utvalg på 39 ferdigutdannede siviløkonomer fra et årskull (1984) foreligger registrering av følgende variable:

X_1 : Kjønn (0=Mann, 1 = Kvinne)
X_2 : Alder (Antall år ved opptak)
X_3 : Praksis (Antall år før opptak)
X_4 : Skolepoeng fra videregående skole (75 = 5 i snitt)
X_5 : Norskkarakter
X_6 : Matematikkkarakter
X_7 : Antall år med matematikk
X_8 : Studieretning (1=Naturfaglig, 2 = Samfunnsfaglig, 3 = Språklig, 4 = Handel og kontor)
Y : Gjennomsnittskarakteren til siviløkonomeksamen (karakterskala fra 0 til 9, med 9 som topp).

Disse opplysningene er lagt inn på datafil 'eks12.5' med 39 linjer (en for hver student) og 9 kolonnefelter (ett for hver variabel). I den følgende analysen søker en å forklare "responsvariablen" Y ved $X_1, X_2, \ldots, X_7$. X_8 holdes foreløbig utenfor av grunner vi kommer tilbake til.

```
>> READ 'eks12.5' Y X1-X8
 Row    Y  X1  X2  X3  X4   X5   X6  X7  X8
   1  5.1   0  22   2  71  4.0  5.0   3   1
   2  3.6   0  21   1  67  4.0  4.0   3   1
   3  4.9   0  21   1  73  4.0  5.0   1   2
   4  6.6   0  20   1  81  5.0  6.0   2   2
. . . 39 ROWS READ
>> REGRESS Y on X1-X7
Y = 2.49 - 0.838 X1 - 0.063 X2 - 0.553 X3 - 0.0162 X4
         + 0.290 X5 + 0.884 X6 + 0.159 X7   (S=0.787)
                          St. Dev.  T-ratio =
 Variable  Coefficient  of coef.  coef/s.d.
                 2.487     5.860       0.42
       X1      -0.8377    0.3222      -2.60
       X2      -0.0633    0.1968      -0.32
       X3      -0.5533    0.3732      -1.48
       X4      -0.0163    0.0645      -0.25
       X5       0.2899    0.3981       0.73
       X6       0.8844    0.3213       2.75
       X7       0.1591    0.1858       0.86
Degrees of freedom for T-test = 31
R-squared = 49.9 % (adjusted for d.f. 38.6 %)
```

Utskriften gir

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \cdots + \hat{\beta}_7 X_7$$

med $\hat{\beta}_j$, $S(\hat{\beta}_j)$ og $T = \hat{\beta}_j / S(\hat{\beta}_j)$ i egne kolonner. Den siste gir mulighet for å teste hypotesene $\beta_j = 0$ enkeltvis (forutsetter standardmodellen med normalantakelse). En tosidig t-test med 5% signifikansnivå, gir kritisk verdi 2.04 (Tabell C.9 med frihetsgradtall $39 - 8 = 31$). Følgelig er det bare β_1 og β_6 som kan påstås ulik null. Dette betyr at det bare er X_1 (kjønn) og X_6 (matematikk-karakteren) som er statistisk signifikante variable i en forklaringsmodell der alle variable $X_1, X_2, \ldots, X_7$ inngår. Det er imidlertid en viss indikasjon på at X_3 (praksis) også kan ha en viss (negativ) betydning. Vi merker oss at forklaringsgraden målt med determinasjonskoeffisienten R^2

bare er moderat stor (49.9%), og at feilleddet har et relativt stort standardavvik (estimert til 0.787).

Vi vil få en enklere forklaringsmodell dersom variablene X_2, X_4, X_5 og X_7 utelates helt. En slik sesjon ga utskriften:

```
 >> REGRESS Y on X1 X3 X6
Y = 2.36 - 0.791 X1 - 0.616 X3 + 0.756 X6  (S=0.787)
                     St. Dev.  T-ratio =
Variable Coefficient of coef.  coef/s.d.
               2.365    1.028       2.30
      X1     -0.7910   0.2912     -2.72
      X3     -0.6159   0.2709     -2.27
      X6      0.7560   0.1907      3.97
Degrees of freedom for T-test = 35
R-squared = 45.5 % (adjusted for d.f. 40.8 %)
```

I denne analysen er alle variablene som inngår statistisk signifikante på 5%-nivå (35 frihetsgrader gir kritiske verdier 2.03 på 5%-nivå og 2.72 på 1%-nivå). Det er ikke overraskende at X_3 nå står fram som mer betydningsfull, idet praksis er korrelert med flere av de utelatte variablene, først og fremst alder og skolepoeng (mer praksis for å oppveie lave skolepoeng). Dette er bl.a. årsaken til at forklaringsgraden ikke er vesentlig redusert ved å utelate de 4 øvrige variable (determinasjonskoeffisienten er redusert fra ca. 50% til ca. 45%). Vi ser videre at standardavviket til feilleddet er omtrent det samme som i forrige analyse. Dette innebærer at usikkerheten ved prediksjoner trolig blir mindre med den siste modellen, fordi vi har færre estimerte regresjonskoeffisienter som det hefter usikkerhet ved. Størrelsen R^2 justert for frihetsgrader er definert ved

$$R^2_{just} = R^2 - \frac{r}{n-r-1}(1-R^2)$$

og er veiledende mth. hvorvidt en modell med færre forklarende variable bør foretrekkes, selv om forklaringsgraden målt med R^2 er redusert. Vi ser at den justerte R^2 er økt i den forenklede modellen (fra ca. 39% til ca. 41%), slik at denne peker seg ut.

Utskriftene ovenfor utgjør det som automatisk kommer ut av de fleste programmer. I tillegg finnes vanligvis en del opsjoner, eksempelvis for diagnostiske formål eller prediksjonsformål:

```
>> DIAGNOSTICS
                     St.Dev.
 Row    Y  Pred. Y  Pred. Y  Residual  St.Res.
  11  3.8     5.53     0.15     -1.73   -2.28*
*is observation with large standardized residual
>> PREDICT ROW 1-4
              St.Dev.     St.Dev.
 Row  Pred. Y  Pred. Y  Pred.Error
   1    4.913    0.315       0.834
   2    4.773    0.217       0.802
   3    5.529    0.149       0.787
   4    6.285    0.265       0.817
```

Ved diagnosen er her linje nr.11 plukket ut som spesiell, fordi den standardiserte residualen er stor (større en 2.0). Vedkommende student har tydeligvis en betydelig lavere prestasjon enn opptaksgrunnlaget skulle tilsi, men ikke så ekstremt at den antatte normalitet bør stå i fare. Normalitetsantakelsen kan sjekkes ved et normalskårplott av residualene, jfr. Kapittel 8.5. Ovenfor ber en også om å få predikert den avhengig variable for de verdikombinasjoner av de forklarende variable som er gitt i linje 1 til 4. Det dreier seg om å predikere prestasjonen til nye tilfeldige studenter i populasjonen med disse opptaksdata. Vi ser at estimatene for standardavvikene til prediksjonsfeilen er ca. 0.80 (litt større enn S=0.787). Denne feilmarginen kan garanteres med ca. 68% sikkerhet, mens $2 \cdot 0.80 = 1.60$ kan garanteres med ca. 95% sikkerhet. Selv om to studenter starter opp med en prognose som avviker med et helt karaktertrinn, er det derfor gode sjanser for å gjøre prognosene til skamme.

Leseren har kanskje lurt på hvordan det har seg at skolepoeng (X_4) ikke kom ut signifikant i den første analysen. Det kan forklares ved at viktige elementer i denne karakteren er med som egne variable, som muligens er korrellert med de karakterer som ikke er representert ved egne variable. Korrelasjonskoeffisientene mellom variablene er gitt ved:

```
 >> CORRELATE Y X1-X7
        Y     X1     X2     X3     X4    X5     X6
X1 -0.218
X2 -0.318 -0.246
X3 -0.309 -0.330  0.622
X4  0.556 -0.023 -0.536 -0.473
X5  0.270 -0.022 -0.218 -0.301  0.526
X6  0.553  0.078 -0.336 -0.234  0.680 0.021
X7 -0.119  0.089  0.088  0.201 -0.174 0.332 -0.374
```

Vi ser at X_4 faktisk er den variablen som er sterkest korrelert med Y. Dette betyr at selv om X_4 er uten betydning i tillegg til de øvrige variable, så er skolepoengene den variabel som alene gir best forklaring. En slik enkel regresjonsanalyse faller ut slik:

```
  >> REGRESS Y on X4
 Y = - 4.08 + 0.127 X4  (S=0.787)
                         St. Dev.  T-ratio =
 Variable  Coefficient  of coef.  coef/s.d.
                -4.080     2.300      -1.77
       X4       0.1267     0.031       4.07
Degrees of freedom for T-test = 37
R-squared = 31.0 % (adjusted for d.f. 29.1 %)
```

Vi ser ar X_4 er signifikant, den fanger opp mange relevante ting som er utelatt i modellen. Forklaringsgraden målt med R^2 er imidlertid bare 31%, og det ser ikke ut til at en tjener noe på å forenkle modellen til bare en forklarende variabel, idet den justerte R^2 bare er 29%, i motsetning 41% i modellen med tre forklarende variable.

La oss til slutt diskutere den utelatte variabel X_8. Dette er en kategorivariabel med flere enn to kategorier, og kan ikke inngå direkte i en lineær modell (hvorfor?). Dette problem kan løses ved å innføre indikatorvariable:

$$X_9 \quad = \quad 1 \text{ for naturfaglig studieretning } (= 0 \text{ ellers})$$
$$X_{10} \quad = \quad 1 \text{ for samfunnsfaglig studieretning } (= 0 \text{ ellers})$$
$$X_{11} \quad = \quad 1 \text{ for språklig studieretning } (= 0 \text{ ellers})$$
$$X_{12} \quad = \quad 1 \text{ for handel og kontor } (= 0 \text{ ellers})$$

Tre av disse variablene, f.eks. de tre siste, er nok til å representere alle muligheter. Naturfaglig studieretning er da representert ved at de tre variablene er null. En utvidet analyse ga resultatet:

```
 >> INDICATORS of X8 into X9-X12
>> REGRESS Y on X1 X3 X6 X10-X12
 Y = 2.08 -  0.783 X1  -  0.622 X3  + 0.856 X6 - 0.316 X10
         - 0.526 X11 - 0.008 X12              (S=0.780)
                       St. Dev.  T-ratio =
 Variable  Coefficient  of coef.  coef/s.d.
                 2.075    1.072       1.94
       X1      -0.7831   0.3043     -2.57
       X3      -0.6224   0.2802     -2.22
       X6       0.8561   0.2123      4.03
      X10      -0.3158   0.3173     -1.00
      X11      -0.5262   0.3872     -1.36
      X12      -0.0076   0.5224     -0.01
Degrees of freedom for T-test = 32
R-squared = 49.1 % (adjusted for d.f. 39.6 %)
```

Det ser ut til at studieretning ikke har betydning sett i sammenheng med de øvrige variable.

Forkastning av en hypotese om at en regresjonskoeffisient er null betyr at vi påstår at den er ulik null. Ikke forkasting betyr ikke nødvendigvis at vi bør påstå det motsatte, at koeffisienten er lik null, eller bør handle som om den er null. Tenk på en situasjon der vi på forhånd tror at en regresjonskoeffisient er positiv (evt. negativ). Dersom observasjonene gir en estimert koeffisient som ikke er signifikant forskjellig fra null, men likevel har det fortegn som vi på forhånd trodde, vil det virke urimelig å sette koeffisienten lik null. Den tilhørende forklaringsvariablen bør beholdes, vi har jo på sett og vis fått bekreftet våre formodninger, bare en estimert koeffisient med motsatt fortegn vil stå i motstrid til disse. På den annen side dersom vi på forhånd tror at den forklaringsvariablen det er tale om er uten betydning, og regresjonsanalysen viser en ikke signifikant regresjonskoeffisient og altså ikke er i motstrid med vår formodning om at regresjonskoeffisienten er null, virker det rimelig å fjerne denne forklaringsvariablen fra regresjonsligningen og estimere denne på nytt med de gjenværende forklaringsvariablene. I en situasjon hvor vi på forhånd ikke har bestemte formodninger om betydningen av en forklaringsvariabel, vil framgangsmåten avhenge av et visst skjønn, kanskje fjerner man en ikke-signifikant forklaringsvariabel ut fra ønsket om en enkel modell.

I situasjoner der en vet lite om mulige forklaringer av en respons kan det være fristende å forsøke en hærskare av forklaringsvariable, enten alle på en gang eller ulike kombinasjoner av slike etter tur. Gjentatte regresjonsanalyser med det samme observasjonsmaterialet med sikte på å finne en enkel forklaringsmodell med tilstrekkelig forklaringsevne har imidlertid visse betenkelige sider, bl.a. er det vanskelig, selv i lys av standardregresjonsmodellen, å vurdere påliteligheten av de konklusjoner som trekkes av den modell som til slutt blir valgt. En metode som har fått stor popularitet, er såkalt *trinnvis regresjon*. Dette foregår vanligvis slik at man først velger ut den variabel som alene gir best forklaring, og deretter trekker inn den variabel som sammen med den første variablen forbedrer forklaringens mest osv. Før eller senere kommer man til et punkt der det er lite å hente ved å innføre enda en forklaringsvariabel. Det kan da også være aktuelt å fjerne en variabel som kom inn tidligere, men som i lys av de andre variablene som er trukket inn, bidrar lite til forklaringen. Hensikten med trinnvis regresjon er nettopp å bestemme en kombinasjon av et mindre antall forklaringsvariable som kan fungere som en hensiktsmessig forklaringsmodell. For nærmere beskrivelse av trinnvis regresjon må vi vise til spesiallitteratur, og vi vil advare mot bruk av de regneprogrammer som er allment tilgjengelig uten å være innforstått med den kritikk som kan rettes mot metoden.

Ovenfor har tankene kretset om problemer der den avhengig variable er gitt ut fra omstendighetene og aktuelle forklaringsvariable velges i henhold til dette, som oftest slike variable som man på forhånd tror kan påvirke den avhengig variable, det ligger jo i navnet. Mange problemstillinger kan imidlertid ha en noe annen karakter: Vi har to eller flere variable hvor vi ikke nødvendigvis har klare forstillinger om noen sammenheng eller årsaksforhold. Hensikten med å observere sammenhørende verdier av disse variablene er kanskje nettopp å forsøke å bringe på det rene om det finnes noen sammenheng. En måte å analysere slike data på er å foreta en rekke regresjonsanalyser hvor vi etter tur lar hver av variablene fungere som avhengig variable, med ulike kombinasjoner av de andre variablene som forklaringsvariable. En slik framgangsmåte faller utenfor rammen av vanlig statistisk inferensteori, og det er derfor vanskelig å vurdere påliteligheten av de konklusjoner som trekkes på grunnlag av den modell vi ender opp med. Framgangsmåten må heller betraktes som beskrivende eksplorativ statistikk, dvs. et forsøk på å skaffe seg innsikt i et observasjonsmateriale med mange variable. For dette formål fins imidlertid andre velegnede metoder, bl.a. såkalt prinsipalkomponentanalyse.

12.6 Oppgaver

1. En regresjonsanalyse med tre forklaringsvariable er foretatt på grunnlag av $n = 20$ sammenhørende observasjoner. Den estimerte regresjonsligning ble

$\hat{Y}$	$= 10.7 +$	$1.3X_1$	$+ 1.4X_2$	$- 0.6X_3$
Est.st.avvik		(0.5)	(0.8)	(0.1)
t-verdi		(2.6)	(…)	(…)
P-verdi		($\approx$ 0.002)	(…)	(…)
Konf.int (95%)		(0.24,2.36)	(…)	(…)

   Fyll inn de korrekte tall ovenfor, og tolk resultatene.

2. For en rekke utvalgte personer foreligger følgende observasjoner vedrørende siste års sparing (Y), inntekt (X_1) (beløp i tusen) og forsørgelsesbyrde (X_2)

Familie nr	Sparing	Inntekt	Forsørgelsesbyrde
1	2.6	96	4
2	3.5	84	3
3	5.4	115	2
4	0.5	50	1
5	3.2	142	3
6	4.4	136	5
7	1.7	91	2
8	2.9	107	4
9	2.0	128	6
10	7.2	119	3
11	1.2	76	3
12	3.3	93	4

   (a) Anta lineær regresjonsmodell, estimer parametrene i denne med et regresjonsanalyseprogram og tolk resultatene.

(b) Drøft modellen. Hvilke andre variable kunne ha interesse for å forklare sparenivået hos de enkelte personer.

3. I en kystby er det foretatt taksering av eneboliger for fastsettelse av kommunale avgifter. For et utvalg eneboliger foreligger taksten (i titusen kr.) samt antall rom og alder.

Bolig nr.	1	2	3	4	5	6	7	8	9	10	11	12
Takst	39	48	43	45	35	42	36	45	57	38	58	48
Antall rom	8	10	9	10	8	11	7	10	12	8	10	9
Alder	10	5	11	8	32	14	23	15	2	18	4	8

(a) Utfør to enkle regresjonsanalyser med forklarende variabel (i) Antall rom (ii) Alder . Hvilken variabel gir alene best forklaring?
Prediker taksten for en bolig med ti rom.
Prediker taksten for en ti år gammel bolig.

(b) Utfør multippel regresjonsanalyse med begge forklarende variable, og prediker taksten for en ti år gammel bolig med ti rom.

(c) Anta at de seks siste boligene har strandrett. Utvid analysen i (b) slik at takstverdien av en strandrett kan estimeres.

(d) Diskuter ut fra residualanalyse om noen av boligene kan være "feiltaksert".

4. Ta for deg Statistisk Årbok og finn der et observasjonsmateriale som du kan ha interesse av å analysere med regresjonsanalyse. Utfør denne analysen og tolk resultatene.

5. Hvilke av følgende modeller kan analyseres ved vanlig lineær regresjonsanalyse

(a) $Y = \beta_0 + \beta_1 X^2 + U$

(b) $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + U$

(c) $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + U$

(d) $Y = \beta_0 + X^{\beta_1} + U$

(e) $Y = \beta_0 + \beta_1 logX + U$

6. Hvilke av følgende modeller kan analyseres ved regresjonsanalyse dersom vi tar en logaritmisk transformasjon av den avhengig variable

(a) $Y = \alpha e^{\beta X} + U$

(b) $Y = \alpha e^{\beta X} \cdot U$

(c) $Y = \alpha \beta^X \cdot U$

(d) $Y = \alpha X^\beta \cdot U$

(e) $Y = \alpha X_1^{\beta_1} \cdot X_2^{\beta_2} \cdot U$

Er det noen vesenforskjell på modellene i (b) og (c)?

7. En bedrift har følgende salgstall for en artikkel for de siste seks år

År (X):	1	2	3	4	5	6
Salg (Y):	125	142	176	211	244	306

Det ser ut til at salget har en vekst som er raskere enn lineær vekst.

(a) Drøft om det er rimelig å bruke en av modellene i Oppgave 6 istedenfor en lineær modell.

(b) Anta at modellen i Oppgave 6(c) er valgt. Estimer parametrene i denne modellen på grunnlag av observasjonene.

(c) Bruk den fastlagte modell til å gi en prognose for salget i år 7 og 8.

(d) Vil du feste lit til en tilsvarende prognose for år nr. 12?

8. Et eksperiment vedrørende sammenhengen mellom hastighet (km/t) og bremselengde (meter) for kjøring på et bestemt veidekke er gjennomført. Resultatet ble

Hastighet (X):	30	40	50	60	70	80	90	100	110
Bremselengde(Y):	12	19	28	39	53	68	85	106	134

Det er foreslått en modell av form

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + U.$$

(a) Estimer parametrene i modellen med et regresjonsanalyseprogram. Tolk resultatet.

(b) Gi en prognose for bremselengden ved en hastighet på 85 km/t, og ved en hastighet på 110 km/t.

9. En teknisk rapport gir opplysning om virkningen (Y) av et tilsetningsstoff som kan tilsettes i varierende kvanta (X_1) ved ulike temperaturer (X_2). Rapporten gir bl.a. følgende regresjonsligning, med tilhørende estimerte standardavvik

$$\begin{array}{ccccccc} \hat{Y} & = & 40.7 & + & 1.72X_1 & + & 0.45X_2 \\ & & & & (0.42) & & (0.11) \end{array}$$

Estimert standardavvik av feilleddet er 3.2. Nå viser det seg at rapporten bruker amerikanske måleenheter. Y er strekkstyrke målt i pounds, X_1 er volum målt i ounces og X_2 er temperatur i Fahrenheitgrader. Oversett opplysningene ovenfor til norske måleenheter når vi har følgende sammenhenger

1 pound	=	0.454 kg
1 ounce	=	29 milliliter
Fahrenheit	=	9/5 Celcius + 32

10. Ved analyse av aksjekurser for børsnoterte selskaper studeres ofte en modell av følgende form

$$R_j = \alpha_j + \beta_j \cdot R_M + U_j$$

der R_M er avkastningen i markedet ifølge en markedsindeks og R_j er avkastningen på aksjene i selskap nr.j. Parametrene a_j og β_j karakteriserer selskap nr.j i forhold til markedsindeksen, mens U_j er tilfeldig variasjon. Alle variablene antas å være stokastiske.

(a) Gjør nødvendige antakelser, og vis at

$$\begin{aligned} \beta_j &= cov(R_j, R_M)/var(R_M) \\ cov(R_j, R_k) &= \beta_j \cdot \beta_k \end{aligned}$$

(b) Anta at avkastninger er notert for n ulike tidsperioder. Hvilke betingelser bør være oppfyllt for å anslå β_j' ene ved gjentatte regresjonsanalyser.

11. I økonomi brukes ofte produksjonsfunksjoner som relaterer produksjonskvantum Q med arbeidsinnsats L og kapital K. Den såkalte Cobb-Douglas funksjonen ser slik ut:

$$Q = AL^\beta K^{1-\beta}$$

Drøft hvordan denne kan omformes til en lineær modell som kan danne utgangspunkt for estimering av β ut fra observerte data.

12. I økonomi anvendes ofte modeller som relaterer konsum C og inntekt Y, f.eks.

$$C = \gamma + \beta Y + U$$

I tillegg kommer ofte bibetingelsen $Y = C + I$, der I er investering som antas å være gitt utenfor modellen (eksogen). Vurder om antakelsen at Y er uavhengig av feilleddet U er rimelig i en slik situasjon. Hvilke problemer reiser dette for estimeringen av β?

13. La $Y_1, Y_2, \ldots, Y_n$ være uavhengige med samme varians σ^2 og med forventninger

$$EY_i = \beta x_i \quad i = 1, 2, \ldots, n$$

dvs. en standard regresjonsmodell med sikker forklaringsvariabel der konstantleddet er antatt lik null.

(a) Vis at estimatoren

$$\hat{\beta} = \frac{1}{M} \sum_{i=1}^{n} x_i Y_i \quad \text{der} \quad M = \sum_{i=1}^{n} x_i^2$$

er forventningsrett med varians $var\hat{\beta} = \sigma^2/M$.

(b) Vis at (jfr.Oppgave 8.33)

$$\sum_{i=1}^{n} (Y_i - \beta x_i)^2 = \sum_{i=1}^{n} (Y_i - \hat{\beta} x_i)^2 + M(\hat{\beta} - \beta)^2$$

og bruk dette til å vise at $\hat{\beta}$ er minste kvadraters estimatoren for β.

(c) Vis at en forventningsrett estimator for σ^2 er gitt ved

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (Y_i - \hat{\beta} x_i)^2$$

(d) Forklar at dersom Y_i'ene antas normalfordelte, så er estimatoren $\hat{\beta}$ normalfordelt.

14. En bedrift har en maskin som utporsjonerer av råstoff for videre bearbeiding. Maskinen kan innstilles på en glidende skala fra 0.0 til 3.0. Det kvantum som porsjoneres ut i de ulike posisjoner' vil avhenge av konsistensen av råstoff, og man ønsker å anslå sammenhengen for en gitt type råstoff. Resultatet ble:

Posisjon:	0.5	1.0	1.5	2.0	2.5	3.0
Kvantum i kg:	0.63	1.18	1.87	2.36	2.94	3.65

(a) Forklar at det her kan være rimelig å bruke en lineær regresjonsmodell uten konstantledd.

(b) Estimer parametrene i en slik modell og rapporter resultatet.

(c) Estimer forventet kvantum råstoff i posisjonene
(i) 0.8 (ii) 1.0 (iii) 2.2

(d) Hvilken posisjon bør maskinen stilles inn på for å lage kiloporsjoner?

15. $\star$Betrakt modellen

$$Y_i = \gamma + \beta X_i + U_i \quad i = 1, 2, \ldots, n$$

Det er isteden foreslått å basere analysen på de enkelte observasjonenes avvik fra gjennomsnittet, dvs. sette $Y_i' = Y_i - \overline{Y}$ og $X_i' = X_i - \overline{X}$.

(a) Vis at for disse observasjonene har vi en modell av form

$$Y_i^{'} = \beta X_i^{'} + U_i^{'} \quad 1, 2, \ldots, n.$$

dvs. en regresjonsmodell uten konstantledd.

(b) Vis at minste kvadraters estimatoren for β i den transformerte modell gir samme estimat for β som minste kvadraters estimatoren i den opprinnelige modell.

(c) Dersom vi gjør standardantakelsene om uavhengighet, samme varians etc. for den opprinnelige modell, vil det samme nødvendigvis gjelde for den transformerte modell?

16. Gi en kritisk vurdering av Eksempel 5 mht.

(a) de tekniske forutsetningene for analysen

(b) muligheten for å trekke generelle konklusjoner om sammenhengen mellom studieprestasjon og opptaksvariable.

Kapittel 13

Tidsrekkeanalyse og prediksjon

13.1 Innledning

Mange statistiske undersøkelser har form av at samme størrelse observeres på etterfølgende tidspunkter over en viss tid. Slike observasjoner kalles gjerne en *tidsrekke*. Eksempler på tidsrekker er

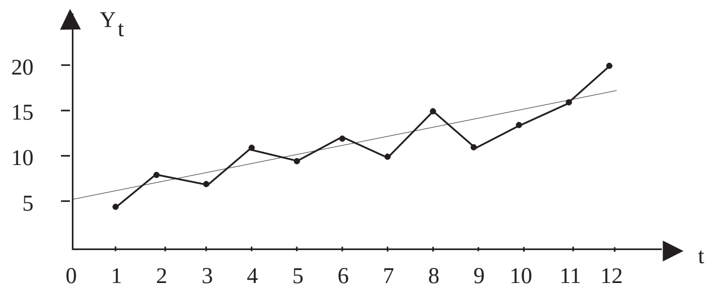
- Temperaturer målt på etterfølgende dager.
- Forurensing målt på etterfølgende dager.
- Omsetning i et varehus fra dag til dag.
- En prisindeks fra uke til uke.
- Sysselsetting fra måned til måned.
- Antall drepte i trafikken fra år til år.

En tidsrekke beskriver et utviklingsforløp, og formålet med en tidsrekkeanalyse kan være å forstå den utvikling som har funnet sted. Ofte vil det også være et underliggende ønske om å bruke de funne resultater til å treffe beslutninger. I forbindelse med tidsrekkeanalyse kommer det derfor inn et moment av *prediksjon*, vi ønsker å forutsi den videre utvikling av tidsrekken på grunnlag av de observasjoner som er gjort opp til nå.

Ved analyse av tidsrekker kan man lete etter utviklingsmønstre og typiske regulariteter. De konklusjoner som eventuelt trekkes om fremtiden av en slik analyse forutsetter at omstendighetene ikke endrer seg drastisk i forhold til den tidsperioden vi har observasjoner fra. De beslutninger som treffes på grunnlag av en tidsrekkeanalyse kan ha ulik natur. Det kan være tale om å innrette seg etter en ventet utvikling, eksempelvis når det gjelder å bestemme produksjonskvanta. Det kan også være tale om å treffe tiltak for å snu en uønsket utvikling som synes å inntreffe dersom en ikke foretar seg noe, eksempelvis ved forurensingsproblemer.

En rekke ting er typiske for tidsrekker, vi vil spesielt trekke fram fire elementer:

Vi kan ha en *trend* som gir uttrykk for langtidsutviklingen av tidsrekken. I denne forbindelse taler vi gjerne om stigende eller fallende trend.



Figur 13.1: Tidsrekke med trend og sesongkomponent

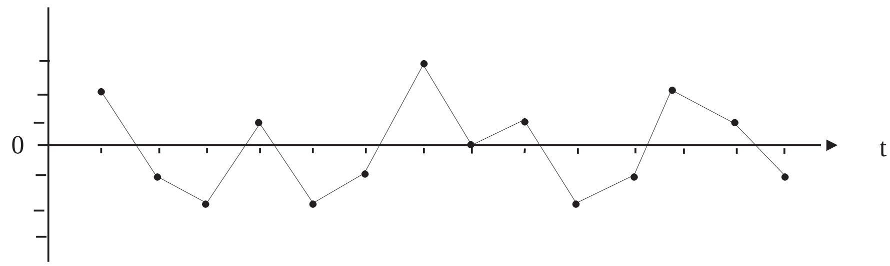
Vi kan ha *periodiske variasjoner*, dvs. variasjoner som gjentar seg med en viss regelmessighet etter et visst tidsrom. Slike periodiske variasjoner kan bestå av korttidsvariasjoner og/eller langtidsvariasjoner. Vi kjenner dette fra mange geofysiske tidsrekker, eksempelvis døgnvariasjoner og årvariasjoner, og fra økonomiske tidsrekker med årlige sesongvariasjoner og/eller lengre konjunkturvariasjoner. Det er av og til nødvendig å skille ut konjunkturvariasjoner som er eget element, bl.a. fordi slike ikke har en klart definert periodelengde i motsetning til rene sesongvariasjoner.

Det tredje element vi vil trekke fram er *autokorrelasjon*, dvs. en viss avhengighet mellom etterfølgende verdier av tidsrekken. Det er tale om positiv autokorrelasjon som betyr at store (små) verdier gjennomgående etterfølges av store (små) verdier. Det motsatte er negativ autokorrelasjon som betyr at store (små) verdier gjennomgående etterfølges av små (store) verdier. Et eksempel på positiv autokorrelasjon vil være en periodedranker som etter å ha drukket et stort kvantum den ene dagen, er tilbøyelig til å følge opp den neste. Mer seriøse eksempler finner vi i mange økonomiske tidsrekker hvor en påvirkning i en bestemt retning ofte gjør seg gjeldende en tid framover. Et eksempel på negativ autokorrelasjon kan være en families daglige innkjøp av melk. Det finnes også økonomiske tidsrekker med slik karakter, typisk er innkjøpte kvanta for lager i etterfølgende perioder, etterspørsel etter visse kostbare varer med et lite antall etterspørrere.

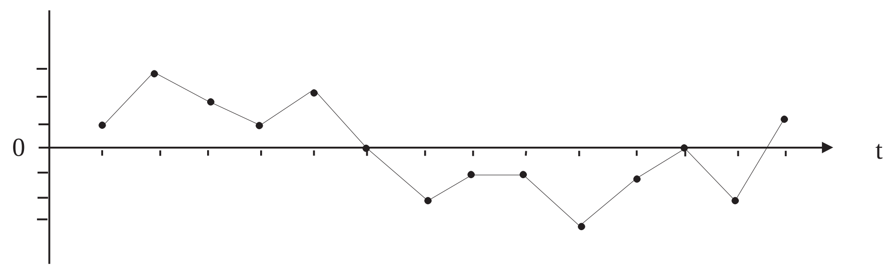
Det fjerde element som er typisk for tidsrekker er *tilfeldig variasjon*, dvs. variasjon som ikke kan forklares ut fra trend, periodisk variasjon eller autokorrelasjon.

I praksis vil disse fire elementer kunne være til stede i varierende grad, for noen tidsrekker kan enkelte elementer mangle helt. Ved tidsrekkeanalyse ønsker vi å klarlegge dette. La oss illustrere noen av disse forhold med figurer.

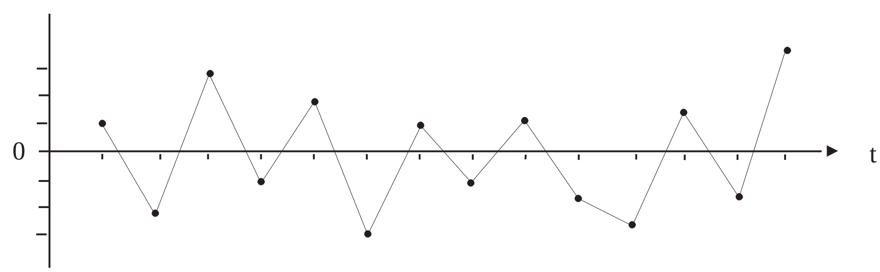
I Figur 13.1 ser vi en tidsrekke som er observert kvartalsvis over tre år, og observasjonene er representert ved punktene i figuren. Her er det åpenbart en stigende trend, antydet ved den inntegnede rette linjen. Det er også en tydelig sesongkomponent med første kvartal som lavsesong og fjerde kvartal som høysesong. Om det i tillegg er autokorrelasjon er vanskelig å si, trolig ikke. La oss nå illustrere autokorrelasjon, og for enkelhets skyld ser vi på tidsrekker uten trend og sesongvariasjon. Tidsrekkene i Figurene 13.2a, 13.2b og 13.2c viser henholdsvis ingen, positiv og negativ autokorrelasjon.



a: Ingen autokorrelasjon



b: Positiv autokorrelasjon



c: Negativ autokorrelasjon

Figur 13.2: Ulik grad av autokorrelasjon

Vi er ofte interessert i en rask avklaring på om en tidsrekke uten trend og sesong har autokorrelasjon eller ikke. En enkel test er den såkalte “runtesten som tar utgangspunkt i om verdiene ligger på oversiden (+) eller undersiden (−) av "nivåettil tidsrekken. Antall "følgerpå hver side brukes som testobservator, se Oppgave 7.37. Dette er moderat for ingen autokorrelasjon, mens det er lite (stort) for positiv (negativ) autokorrelasjon.

For tidsrekkene i Figurene 13.2a–c er antall følger henholdsvis 10, 3 og 13. Uavhengighet svarer til binomiske forsøk med $p = 0.5$ for +, og forventet antall følger og P-verdier kan dermed tilnærmet beregnes. En programvareutskrift av runtesten for den første tidsrekken (tallene er i Eksempel 5 nedenfor):

```
 >> runs 'rekke1'
Expected no. of runs = 8
Observed no. of runs = 10
P-value = 0.2231
note: n too small for good approximation
```

Vi ser at antall runs ikke avviker mye fra forventet ved uavhengighet, og at P-verdien ikke er spesielt liten, som støtter antakelsen om uavhengighet. For de to andre tidsrekkene ble P-verdien ca. 0.006, dvs. klar indikasjon på avhengighet. Merk imidlertid at korte tidsrekker gir grove tilnærminger.

13.2 Analysemetoder

La oss innføre følgende notasjon: Vi betegner den observerte tidsrekken

$$Y_1, Y_2, Y_3, \ldots, Y_n$$

slik at Y_t er den observerte verdi på tidspunkt t, og vi antar her at vi har observert fram til og med tidspunkt n. Dersom vi skal predikere tidsrekken videre framover i tiden, betegner vi prognosetallene henholdsvis

$$\hat{Y}_{n+1}, \hat{Y}_{n+2}, \hat{Y}_{n+3}, \ldots$$

slik at $\hat{Y}_t$ betegner predikert verdi av tidsrekken for tidspunkt t.

Det er vanskelig å lage en generell teori for analyse av tidsrekker som er egnet under alle omstendigheter. Det finnes derfor flere angrepsmåter som hver kan ha sine fortrinn under ulike forhold, men som også i noen grad konkurrerer om brukerens gunst. Det vil derfor være et spørsmål om valg på grunnlag av den innsikt som brukeren har om det foreliggende problem og de forhold som ønskes belyst.

Vi vil først ta opp analyse av tidsrekker der trend og sesongvariasjoner er de dominerende trekk. Dette vil ofte være tilfellet for mange tidsrekker i økonomiske sammenhenger, hvor sesongkomponenten kan bestå i variasjoner med årstiden for månedlige eller kvartalsvise data, eller variasjoner med ukedag for daglige data. Vi vil for enkelhets skyld anta at det bare er en sesongkomponent som gjentar seg periodisk, eksempelvis fra år til år. Et eksempel på en slik tidsrekke har vi i Figur 13.1, hvor en tidsrekke er observert kvartalsvis over tre år. Vi vil foreløpig holde autokorrelasjon utenfor i diskusjonen, men kommer tilbake til dette senere.

Vi vil først ta for oss en meget enkel metode for analyse av tidsrekker med trend og sesongvariasjoner. Metoden bygger på følgende modell

$$Y = T \cdot S \cdot U.$$

Vi tenker oss altså hver tidsrekkeverdi Y som et produkt av en trendverdi T, en sesongfaktor S og en tilfeldig faktor U som skyldes andre omstendigheter. Dersom vi på grunnlag av den observerte tidsrekken $Y_1, Y_2, \ldots, Y_n$ har fastlagt de tilhørende T_t, S_t, U_t for $t = 1, 2, \ldots, n$, sier vi at vi har dekomponert tidsrekken. Ved en slik dekomponering vil trenden T angis med den samme enhet som tidsrekken selv, mens sesongfaktoren S og den tilfeldige faktor U oppfattes som indekser som gir det prosentvise avvik fra trenden. Disse indeksene bør følgelig variere omkring 1 over tid. Sesongindeksen skal gi uttrykk for periodiske avvik, og gis derfor verdier som gjentar seg fra periode til periode.

Eksempel 1 : Fyringsolje
En leverandør av fyringsolje har de siste 5 1/2 år notert månedlig salg $Y_1, Y_2, \ldots, Y_{66}$. Tidsrekken (ikke vist her) har en klar trend og sesongvariasjoner med en periode på 12 måneder. I desember i fjor (måned nr.60) ble det solgt $Y_{60} = 224\ 000$ liter. Ifølge den trend som synes å foreligge “burdede ha solgt $T_{60} = 200\ 000$ liter. Nå er desember høysesong og en analyse viser at desembersalget gjennomgående er 40% over den generelle trenden, vi setter $S_{12} = S_{24} = \cdots = S_{60} = 1.40$. Nå var tilfeldigvis desember måned spesielt mild dette året og en regner at dette medførte at desembersalget dette året var 20% lavere enn vanlig, svarende til $U_{60} = 0.80$. Vi har dermed fått skrevet $Y_{60} = T_{60} \cdot S_{60} \cdot U_{60}$.

Alternativt kunne en tenke seg å dekomponere tidsrekken ut fra en lineær modell

$$Y = T + S + U$$

Her blir alle størrelser målt med de samme enheter, S og U gir da uttrykk for absolutte og ikke prosentavvik fra trenden T. Både S og U bør variere omkring null, den første periodisk.

De to modellene er beslektet, idet den første modellen $Y = T \cdot S \cdot U$ alternativt kan skrives

$$logY = logT + logS + logU$$

slik at denne modellen svarer til en lineær dekomposisjon for logaritmen til de opprinnelige størrelsene. Den første modellen er først og fremst aktuell i situasjoner der sesongutslagene er gjennomgående proporsjonale med trendens verdi, mens den lineære modellen er aktuell når sesongutslagene gjennomgående ikke avhenger av trendens verdi.

Uansett hvilken dekomponeringsmodell som velges vil det første trinn i analysen som regel være å fastlegge trenden. Dette kan gjøres ved ulike metoder. I mange situasjoner er det rimelig å prøve å tilpasse en lineær trend, dvs. betrakte trenden som en lineær funksjon av tiden. Ut fra et plott av tidsrekken kan vi, etter beste skjønn, tegne inn en rett linje som angir hovedtendensen. Mer vanlig er å bestemme minste kvadraters regresjonslinje for Y med hensyn på tiden t, dvs. at trenden blir bestemt ved

$$T = \hat{\gamma} + \hat{\beta} \cdot t$$

der $\hat{\gamma}$ og $\hat{\beta}$ er fastlagt ved minste kvadraters metode, se Kapittel 8.3. Trenden er da stigende eller synkende alt ettersom $\hat{\beta}$ blir positiv eller negativ. En $\hat{\beta}$ nær null svarer til at det ikke er noen langtidsutvikling i tidsrekken, i hvert fall ikke lineær.

Enkelte tidsrekker kan ha et utviklingsforløp som ikke er forenlig med en lineær trend, det kan da være aktuelt å bruke andre funksjonsformer, og en rekke slike er studert i litteraturen, eksponensiell vekst osv. I praksis vil det som regel være lite å hente ved å forlate tanken om lineær trend med mindre dette er åpenbart urealistisk, enten vurdert ut fra den foreliggende tidsrekken eller ut fra allment aksepterte teorier som er relevante for det foreliggende problem.

Det å bestemme langtidsutviklingen i en tidsrekke kan også oppfattes som å glatte tidsrekken, dvs. fjerne uregelmessigheter, som ifølge vår modell har karakter av sesongavvik og tilfeldige avvik fra en trend. Det finnes en rekke metoder for glatting av tidsrekker, en av disse er såkalte *glidende gjennomsnitt*.

Eksempel 2 : Glidende gjennomsnitt
Betrakt følgende utsnitt av en tidsrekke

Tidsrekke:	$\cdots$	2	6	1	5	6	4	5	9	$\cdots$
Glidende gjennomsnitt:		$\cdots$	3	4	4	5	5	6	$\cdots$	

Her er beregnet gjennomsnitt av orden 3, dvs. hvert tall i annen linje er gjennomsnittet av de tre nærmeste tall i linjen over. Vi ser at den nye rekken er en glattet versjon av den første, og gir dermed et uttrykk for hovedtendensen for øyeblikket, som her er stigende. I forbindelse med tidsrekker som observeres månedlig med sesongvariasjoner over 12 måneder, er det aktuelt å beregne glidende gjennomsnitt av orden 12. På dette viset blir sesongvariasjonene og tilfeldige variasjoner i rekken i noen grad glattet ut og det som gjenstår er i hovedsak trend. Dersom tidsrekken observeres kvartalsvis, vil glidende gjennomsnitt av orden 4 kunne brukes.

Glidende gjennomsnitt blir ofte brukt i situasjoner hvor vi ikke på forhånd har bestemte forestillinger om langtidsutviklingen. Ofte kan en slik analyse peke i retning av en lineær trend eller en annen klar trend, i så fall bør dette følges opp [1]. Imidlertid kan det også hende at slik analyse ikke avdekker noen klar trend og det er da ofte urimelig å fortsette etter den modell vi startet med. Det er da trolig andre trekk enn trend og sesongvariasjoner som er særmerkt ved tidsrekken og dette vil kreve en annen modell og andre metoder.

Vi vil i det følgende anta at vi har valgt å dekomponere en tidsrekke etter modellen $Y = T \cdot S \cdot U$. La oss tenke oss at trenden er fastlagt på grunnlag av tidsrekken $Y_1, Y_2, \ldots, Y_n$, ut fra de metoder som er skissert ovenfor eller på annen måte. La oss betegne den fastlagte trend med T_t. Vi kan da beregne $Z_t = Y_t/T_t$; $t = 1, 2, \ldots, n$. Ifølge modellen kan vi skrive $Z_t = S_t \cdot U_t$, dvs. Z_t kan oppfattes som en tidsrekke der trenden er fjernet, men der sesongvariasjoner og tilfeldige variasjoner fortsatt er til stede. Vi ønsker nå å bestemme sesongindekser, og dette kan skje ved å fjerne de tilfeldige variasjonene i tidsrekken Z_t.

La oss for enkelhets skyld anta at tidsrekken består av månedlige data med sesongvariasjoner over 12 måneder. Vi kan da ta for oss verdiene av Z_t, svarende til samme måned i alle de år som tidsrekken omfatter. Tar vi gjennomsnittet av disse tallene vil tilfeldigheter fra år til år reduseres og dette gjennomsnittet tas da som sesongindeks for vedkommende måned. Vi får da bestemt 12 sesongindekser $S_1, S_2, \ldots, S_{12}$ som gir uttrykk for sesongvariasjonen uansett år, slik at $S_1 = S_{13} = S_{25} = \cdots, S_2 = S_{14} = S_{26} = \cdots, \cdots, S_{12} = S_{24} = S_{36} = \cdots$. Disse blir ofte justert slik at gjennomsnittet over en periode er lik 1. Vi kan nå beregne $U_t = Z_t/S_t$ for $t = 1, 2, \cdots, n$, som gir uttrykk for de tilfeldig avvik fra måned til måned.

Vi ser at vi kan gjennomføre en helt analog analyse på grunnlag av den lineære modellen $Y_t = T_t + S_t + U_t$, idet vi etter tur trekker fra de størrelser som er fastlagt istedenfor å dividere.

På grunnlag av en dekomponert tidsrekke $Y_t = T_t \cdot S_t \cdot U_t$ er det naturlig å bruke følgende prediktor for framtidige verdier

$$\hat{Y}_t = T_t \cdot S_t \qquad t = n + 1, n + 2, \ldots$$

Den trend som ble observert i tidsrommet $t = 1, 2, \ldots, n$ forlenges ut i framtida, og samme sesongindeks brukes. Indeksen for tilfeldige avvik er satt lik 1 da vi ikke kan forutsi disse. Dersom vår analyse var basert på den lineære modellen, ville vi sette $\hat{Y}_t = T_t + S_t$.

Eksempel 3 : Dekomponering

En tidsrekke Y_t er observert kvartalsvis i en treårsperiode, dvs. for $t = 1, 2, \ldots, 12$, og det er denne tidsrekken som er illustrert i Figur 13.1 ovenfor. La oss utføre en analyse basert på dekomponering av form $Y = T \cdot S \cdot U$. På øyemål ser det ut til at trenden er lineær, og kan la seg tilfredsstillende representere ved $T_t = 5+1 \cdot t$. Minste kvadraters metode gir imidlertid $T_t = 4.4 + 1.1 \cdot t$, og dette er brukt i beregningene i

[1]Det knytter seg også visse reservasjoner til bruk av glidende gjennomsnitt. En avvikende observasjon kan forårsake at den nye rekke får en bølgefasong som ikke er til stede i den opprinnelige rekke, såkalt Slutsky-Yule effekt.

tabellen nedenfor. Med denne dekomponering predikeres tidsrekkens verdi for første kvartal kommende år:

$$\hat{Y}_{13} = T_{13} \cdot S_{13} = (4.4 + 1.1 \cdot 13) \cdot 0.82 = 15.3$$

og for fjerde kvartal kommende år

$$\hat{Y}_{16} = T_{16} \cdot S_{16} = (4.4 + 1.1 \cdot 16) \cdot 1.19 = 26.2$$

t	Y_t	T_t	Z_t	S_t	U_t
1	4.8	5.5	0.87	0.82	1.06
2	7.7	6.6	1.17	1.05	1.11
3	7.2	7.7	0.94	0.94	1.00
4	10.8	8.8	1.23	1.19	1.03
5	8.0	9.9	0.81	0.82	0.99
6	12.1	11.0	1.10	1.05	1.05
7	10.8	12.1	0.89	0.94	0.95
8	15.6	13.2	1.18	1.19	1.00
9	11.2	14.3	0.78	0.82	0.95
10	13.5	15.4	0.88	1.05	0.84
11	16.0	16.5	0.97	0.94	1.03
12	20.4	17.6	1.16	1.19	0.97

Ved denne form for analyse er det ikke lett å vurdere påliteligheten i eventuelle konklusjoner. Til det kreves ytterligere forutsetninger om de størrelser som inngår i modellen. Slike forutsetninger kan også sette oss i stand til å vurdere egenskapene ved den foreslåtte prognosemetode kontra andre metoder.

En alternativ framgangsmåte for analyse av trend og sesongvariasjon er å ta utgangspunkt i en lineær regresjonsmodell.

Eksempel 4 : Omsetning
En bedrift registrerer kvartalsvise omsetningstall og regner med en lineær trend og at fjerde kvartal er høysesong i forhold til de tre første kvartalene. Følgende modell er da aktuell

$$Y = \beta_0 + \beta_1 t + \beta_2 Q + U$$

der t betegner tidspunkt og Q er en indikatorvariabel som er 1 dersom observasjonen svarer til fjerde kvartal og 0 ellers, og U er tilfeldig avvik. Parametrene β_0, β_1 og β_2 er ukjente og ønskes fastlagt på grunnlag av den observerte tidsrekke. Modellen ovenfor kan bygges ut i ulike retninger, til trender som ikke er lineære i tid, og til mer enn to sesongtyper. Det siste kan skje ved å innføre flere sesongvariable. La oss bruke en modell av denne typen på problemstillingen i Eksempel 3, hvor trolig alle fire kvartaler er ulike. Modellen blir da

$$Y = \beta_0 + \beta_1 t + \beta_2 Q_2 + \beta_3 Q_3 + \beta_4 Q_4 + U$$

der Q_i er 1 dersom observasjonen svarer til i'te kvartal og 0 ellers for $i = 2, 3, 4$. Merk at det er nok med tre indikatorvariable for å beskrive forskjeller for de fire kvartalene. La oss bruke tallmaterialet til å bestemme koeffisientene ved minste kvadraters metode, betegn disse med $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_4$ (se Kapittel 12.2). Som prediktor brukes

$$\hat{Y}_t = \hat{\beta}_0 + \hat{\beta}_1 t + \hat{\beta}_2 Q_2 + \hat{\beta}_3 Q_3 + \hat{\beta}_4 Q_4.$$

Våre observasjoner gir

$$\hat{Y}_t = 3.22 + 0.96 \cdot t + 2.14 \cdot Q_2 + 1.42 \cdot Q_3 + 4.73 \cdot Q_4$$

som bl. a. gir følgende prognoser

$$\begin{aligned}
\hat{Y}_{13} &= 3.22 + 0.96 \cdot 13 + 2.14 \cdot 0 + 1.42 \cdot 0 + 4.73 \cdot 0 = 15.7 \\
\hat{Y}_{16} &= 3.22 + 0.96 \cdot 16 + 2.14 \cdot 0 + 1.42 \cdot 0 + 4.73 \cdot 1 = 23.3
\end{aligned}$$

Sesongvariasjoner har ofte en syklisk karakter i den forstand at i løpet av sesongperioden vil sesongeffekten gå gradvis fra f.eks. lavsesong, via middels sesong til høysesong, deretter via middels sesong til lavsesong igjen ved starten av neste sesongperiode. Det er mulig å lage modeller hvor en tar omsyn til dette, og dermed får ytterligere forbedret eventuelle prognoser. Tidsrekken i Eksempel 3 ser imidlertid ut til å ikke ha denne karakter, høysesong etterfølges direkte av lavsesong.

Fordelen med en analyse basert på en lineær regresjonsmodell, der alle elementene trend, sesongavhengighet og tilfeldig variasjon kommer til uttrykk, er at vi kan støtte oss på en velutviklet teori, se Kapittel 12. Betraktes de variable som inngår i modellen som stokastiske variable med nærmere angitte egenskaper, kan vi komme med presise utsagn om trenden og sesongvariablenes betydning, usikkerheten ved prediksjoner osv. Med tilgang til god programvare vil man som regel foretrekke en analyse basert på en regresjonsmodell framfor den mer klassiske analysen presentert ovenfor.

For analyse og prediksjon av tidsrekker i praksis trengs metoder som også kan ta omsyn til autokorrelasjon og andre spesielle særtrekk, herunder såkalt ikke-stasjonæritet, som ofte forekommer i økonomiske data. I litteraturen fins flere hovedretninger, og vi nevner (med populær metode i parentes):

- "Time Series Analysis and Prediction" (Box-Jenkins metode)
- "Time Series Forecasting" (Holt-Winter's metode)

Box-Jenkins metode er knyttet til den såkalte generelle ARIMA-modellen, mens Holt-Winters metode er mindre generell, men legger i steden vekt på et mer fleksibelt system for oppdatering av prognoser etterhvert som nye data kommer til.

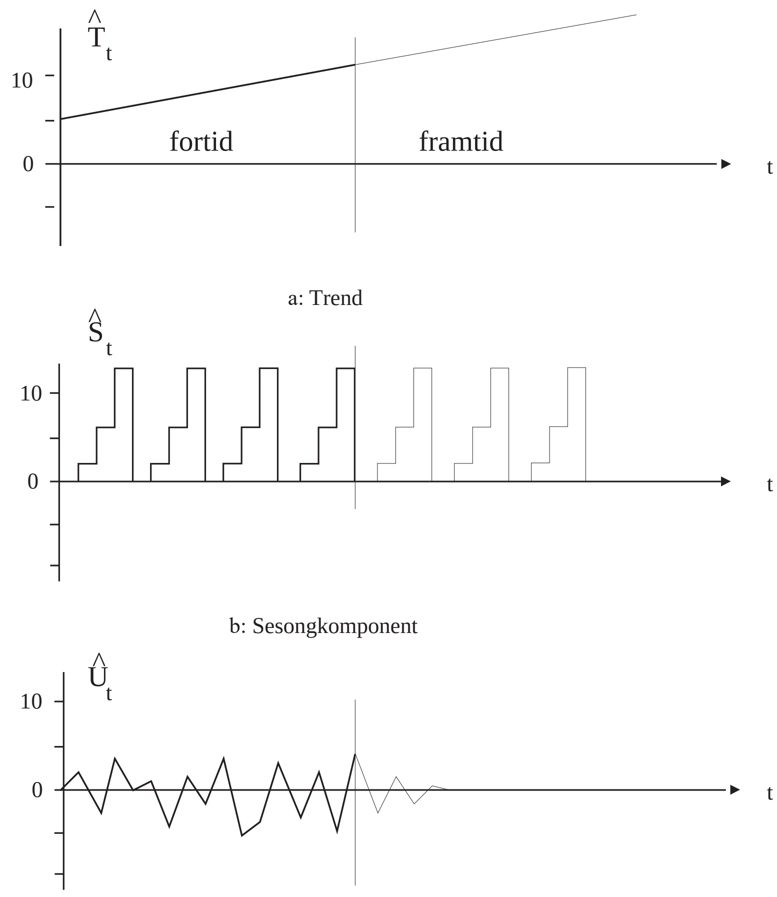
Det fins idag et rikelig tilbud på brukervennlig programvare for tidsrekkeanalyse og prediksjon, men det er vanskelig å gi generelle anbefalinger om metode. En advarsel mot bruk av tilsynelatende avanserte metoder som er automatiserte, med få krav til brukeren, er imidlertid på sin plass. Det lønner seg som regel å holde seg nær til dataene. Et problem i praksis er hvor lang tidsrekke som trengs for å bruke metoden med en viss trygghet. For en del avanserte metoder trengs gjerne mer enn 50 observasjoner, men da møter en raskt problemet at omstendighetene er endret fra de første til de siste observasjonene.

Selv om vi her ikke vil gjennomgå noe avansert system for analyse og prediksjon av tidsrekker, vil vi bruke den enkle dekomponeringsmodellen til å forklare hvordan slike systemer i prinsippet kan ta omsyn til trend, sesongvariasjon og autokorrelasjon.

Først søker en å bestemme trenden og sesongvariasjonene og fjerne disse fra tidsrekken, slik at vi sitter igjen med en rekke hvor autokorrelasjon er det dominerende trekk. På denne rekken kan vi anvende tilgjengelige metoder for analyse av autokorrelasjonen. Når de tre elementene trend, sesongavhengighet og autokorrelasjon er isolert, kan de igjen kombineres til å gi prediksjoner av framtidige verdier av tidsrekken som tar hensyn til alle tre særtrekk ved rekken.

Anta vi har en tidsrekke som vi har valgt å dekomponere i henhold til den lineære modellen $Y_t = T_1 + S_t + U_t$. Anta at en lineær trend $\hat{T}_t$ og sesongkomponenter $\hat{S}_t$ er fastlagt ut fra den observerte rekke, slik at beregnede verdier er $\hat{Y}_t = \hat{T}_t + \hat{S}_t$. De beregnede feilledd er da $\hat{U}_t = Y_t - \hat{Y}_t$. La oss illustrere de prinsipielle sider ved figurer.

I Figur 13.3a er den lineære trenden $\hat{T}_t$ videreført ut over den observerte rekke til framtid, og i Figur 13.3b er sesongkomponenten $\hat{S}_t$ videreført. Residualene $\hat{U}_t$ i Figur 13.3c viser tydelig tendens til negativ autokorrelasjon og denne tendens er videreført et kort stykke ut i fremtid. Den prognose som gis for fremtidige t er dermed gitt ved $\hat{Y}_t = \hat{T}_t + \hat{S}_t + \hat{U}_t, t = n+1, n+2, \ldots$



a: Trend

b: Sesongkomponent

c: Residualer

Figur 13.3: Trend og sesong i tidsrekke

13.3 Autokorrelasjon

Vi vil i dette avsnittet ta for oss sentrale begreper knyttet til autokorrelasjon, og noen sentrale tidsrekkemodeller for autokorrelasjon. Disse er aktuelle for analyser og prognoser i praksis, og er også et godt utgangspunkt for videre studier dersom situasjonen krever andre metoder.

Autokorrelasjon kan opptre sammen med trend og/eller sesongvariasjon, men la oss i første omgang anta at slike elementer ikke er til stede, slik at tidsrekken i hovedsak kan ha karakter som illustrert før i Figurene 13.2a–c.

La Y_t; $t = 1, 2, \ldots n$ være en tidsrekke, der Y_t'ene oppfattes som stokastiske variable med samme forventning μ og varians σ^2, men der det er tidsavhengighet.

Et teoretisk mål for denne tidsavhengigheten er gitt ved den såkalte *autokovariansfunksjonen*

$$c_t(k) = cov(Y_t, Y_{t+k})$$

Spesielt interessant er denne størrelsen dersom tidsrekken er *stasjonær*, som grovt sagt betyr at karakteren av rekken er den samme uansett hvilken tidsperiode vi betrakter. Autokovariansen avhenger da ikke av t, men bare av forskjellen i tid mellom de to tidspunktene i formelen, og vi skriver da $c(k)$. Merk at $c(0) = \sigma^2$. For å skaffe innsikt i tidsavhengigheten er det aktuelt å anslå denne autokovariansen ut fra den observerte rekke $Y_1, Y_2, \ldots, Y_n$, f.eks. ved

$$\hat{c}(k) = \frac{1}{n-k}\sum_{t=1}^{n-k}(Y_t - \overline{Y})(Y_{t+k} - \overline{Y}).$$

Et nærmere studium av den estimerte autokovariansfunksjonen kan bidra til konstruksjon av en brukbar modell for tidsavhengighet i rekken. Mer praktisk er å studere estimatet for den tilhørende *autokorrelasjonsfunksjonen*

$$r(k) = c(k)/c(0).$$

Det finnes en rekke metoder for tidsrekkeanalyse som bygger på forutsetningen om at tidsrekken er stasjonær. La oss presentere to enkle modeller av dette slaget, som også kan vise vei til mer kompliserte modeller. Først

$$Y_t = \gamma + \rho Y_{t-1} + U_t$$

I denne modellen oppfattes tidsrekkens verdi på tidspunkt t som en lineær funksjon av dens verdi på tidspunkt $t-1$ pluss et feilledd som uttrykker en tilfeldig påvirkning på tidspunkt t. Vi vil anta at disse feilleddene er uavhengige stokastiske variable med forventning null og samme varians. Parameteren ρ uttrykker graden av autokorrelasjon og må i praksis være et tall mellom -1 og $+1$, ellers vil tidsrekkens verdier kunne vokse over alle grenser: Positiv ρ uttrykker positiv autokorrelasjon, negativ ρ uttrykker negativ autokorrelasjon. Uavhengighet svarer til $\rho = 0$. Denne modellen kalles *autoregresjon av første orden*, og benevnes AR(1). Mer generelt vil en AR(p) modell uttrykke Y_t ved verdier $1, 2, \ldots, p$ tidsenheter bakover i tid. En annen type modell er

$$Y_t = \mu + \alpha U_{t-1} + U_t$$

Her oppfattes tidsrekkens verdi på tidspunkt t som en lineær funksjon av tilfeldige påvirkninger på tidspunkt t og ett trinn bakover i tid. Disse antas uavhengige stokastiske variable med forventning null og samme varians. Parameteren α uttrykker graden av avhengighet med forrige tidspunkt. Denne modellen kalles ofte *bevegelig gjennomsnitt ("moving average") av første orden*, og benevnes MA(1). Mer generelt vil en MA(q) modell uttrykke Y_t ved tilfeldige påvirkninger $0, 1, 2, \ldots, q$ tidsenheter bakover i tid. Disse to modelltyper kan kombineres til den såkalte ARMA(p,q) modellen.

Merk at tilfellet med uavhengige observasjoner, ofte kalt *ren støy* (PN for "pure noise"), fremkommer som spesialtilfeller av AR(1) og MA(1) ved å sette hhv. $\rho = 0$ og $\alpha = 0$. En annen interessant tidsrekkemodell er såkalt *"random walk"*(RW) gitt ved

$$Y_t = Y_{t-1} + U_t$$

som er grensetilfellet for AR(1) når $\rho = 1$. Her er endringene "ren støy".

For AR(1) og MA(1) modeller vil autokorrelasjonsfunksjonen ha en enkel form, som gjør at tidsrekker av dette slaget er forholdsvis enkle å identifisere. Vi har nemlig (se Oppgave 11 og 12):

Modell	Autokorrelasjonsfunksjon
PN	$r(k) = 0$ for $k > 0$
MA(1)	$r(1) \neq 0$, $r(k) = 0$ for $k > 1$
AR(1)	$r(k) = \rho^k$ for $k > 0$

For identifikasjon i praksis ser en på et såkalt ACF-plott, med estimater for $r(k)$ som funksjon av k, beregnet på grunnlag av den observerte tidsrekke, og som det derfor hefter usikkerhet ved. Et plott med små autokorrelasjoner uten klart mønster sannsynliggjør PN som modell, en klar "piggfor $k = 1$ og ellers intet mønster indikerer MA(1), mens et eksponensielt avtagende mønster indikerer AR(1). For identifikasjon av mer kompliserte modeller trenger en ytterligere kjennetegn som vi ikke skal gå inn på her.

Typisk for stasjonære tidsrekker er at autokorrelasjonen avtar forholdsvis raskt mot null. Dette er ikke tilfellet for en "random walk". En slik kan imidlertig identifiseres ved å sjekke om differensene er "ren støy".

Eksempel 5 : Autokorrelasjon
Betrakt de tre tidsrekkene i Figur 13.2a–c. Disse er:

Y_t:	2	-1	-2	1	-2	-1	3	0	1	-2	-1	2	1	-1
$Y_t^{'}$:	1	3	2	1	2	0	-2	-1	-1	-3	-1	0	-2	1
$Y_t^{''}$:	1	-2	3	-1	2	-3	1	-1	1	-2	-3	2	-2	4

Et ACF-plott av så korte rekker som dette gir ikke god mening i praksis, men hvis noe antyder et slikt at den første rekken ikke har klar autokorrelasjon, mens den andre og tredje rekken har henholdsvis positiv og negativ autokorrelasjon, med AR(1) som aktuell modell. Et mer realistisk eksempel kommer til slutt i avsnittet.

Når en egnet modelltype er identifisert, vil den som regel inneholde ukjente parametre som må estimeres ut fra tidsrekken. Dersom vi har valgt en AR(1) modell, kan vi estimere ρ med minste kvadraters estimatoren

$$\hat{\rho} = \frac{\sum_{t=2}^{n} Y_{t-1}Y_t}{\sum_{t=1}^{n} Y_t^2}$$

Formelen forutsetter (som i eksemplet) at gjennomsnittsnivået er null, se Oppgave 12.13 og 12.15.

Generelt vil estimering av ukjente parametre i ARMA-modeller være relativt komplisert, og en trenger programvare for beregningene.

Eksempel 6 : Autoregresjon
La oss estimere ρ for AR(1) modell for de tre tidsrekkene i Eksempel 5. Vi får henholdsvis

$$\rho = -0.17 \quad \rho^{'} = 0.50 \quad \rho^{''} = -0.56$$

Den første rekken har en negativ estimert koeffisient, men ikke tilstrekkelig negativ til å forlate antakelsen om uavhengighet. Koeffisientene for de to andre rekkene er i samsvar med det vi ventet oss, henholdsvis klar positiv og klar negativ autokorrelasjon.

La oss kort se hvordan en kunne predikere en AR(1) tidsrekke (anta for enkelhets skyld at forventningen er null): Med observert tidsrekke $Y_1, Y_2, \ldots, Y_n$ vil vi, dersom ρ var kjent, predikere den neste verdi av tidsrekken slik

$$\hat{Y}_{n+1} = \rho Y_n$$

Det tilhørende feilledd kan ikke forutsies ut fra tidsrekken, og blir følgelig erstattet med sin forventning null. Vi ser at for positiv autokorrelasjon har prognosen samme fortegn som den sist observerte verdien, men i tallverdi noe mindre enn denne. Ved negativ autokorrelasjon skifter prognosen fortegn i forhold til den sist observerte verdi. Legg merke til at med en første ordens modell har vi ingenting å hente ved å ta omsyn til enda tidligere verdier av tidsrekken. En prognose for tidsrekkens verdi k tidsenheter senere vil være (Oppgave 13)

$$\hat{Y}_{n+k} = \rho^k Y_n$$

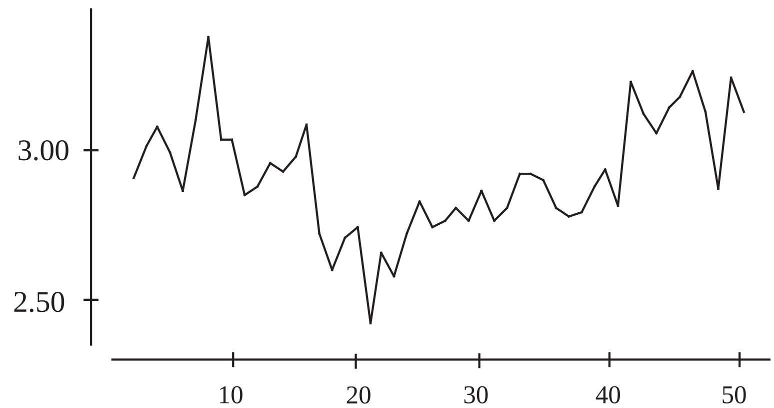
Vi ser at når k vokser vil prognosen nærme seg null, som er rimelig fordi et stykke ut i tid vil de tilfeldige variasjonene gjøre seg såpass gjeldende at det er risikabelt å være mer spesifikk.

Eksempel 7 : Prognose
Anta at de tre tidsrekkene i Eksempel 5 er observert over en lengre periode som går forut for de 14 observasjonene som er gitt i eksemplet og at en på dette grunnlag har bestemt seg for en AR(1) modell med parameter ρ henholdsvis 0.0, 0.6 og -0.6. Prognosen fra tidspunkt 2 og et skritt framover blir for de tre rekkene henholdsvis 0.0, 1.8 og 1.2, mens de faktisk observerte verdier ble henholdsvis -2, 2 og 3.

Dersom parameteren ρ ikke er kjent, men estimert lik $\hat{\rho}$ ut fra tidsrekken, kan vi bruke prediktorer av formen ovenfor, hvor ρ er erstattet med $\hat{\rho}$. Dersom tidsrekken varierer om et nivå ulik null, korrigeres formlene med det beregnede gjennomsnitt av tidsrekken.

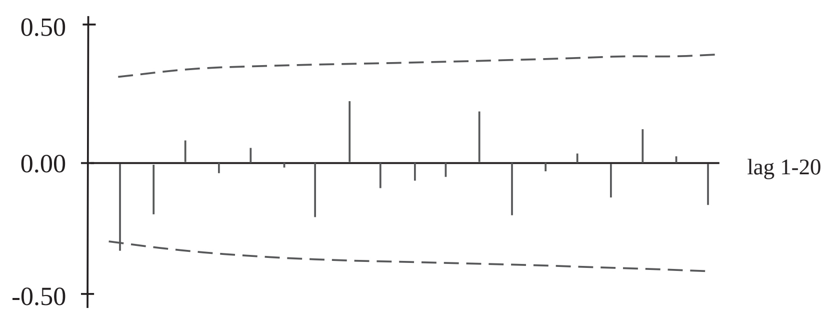
Generelt vil prognoseformler i ARMA-modeller være relativt kompliserte, og en trenger programvare for beregningene.

Eksempel 8 : Identifikasjon
La oss studere nærmere tidsrekken gitt i Oppgave 1.13. Den ser slik ut:



Figur 13.4: Tidsrekke med autokorrelasjon

Den ligner det vi ofte ser for kurser på børsnoterte aksjer. Et ACF-plott viser at autokorrelasjonene avtar sent mot null, noe som tyder på at rekken ikke er stasjonær, men at endringene fra et tidspunkt til det neste kan være det. ACF-plottet for disse ser slik ut:



Figur 13.5: Autokorrelasjonsfunksjon (estimert)

Piggen"ved "laggk=1 ligger på grensen til å være signifikant (95% konfidensgrenser er stiplet i figuren), mens de øvrige pigger ikke tillegges betydning. Aktuelle modeller for differensene er derfor PN eller MA(1). Det er naturlig å forsøke det siste, og etter estimering av modellen teste om MA-parameteren er statistisk signifikant. Dette er en såkalt ARIMA(0,1,1) modell. Her er en programvareutskrift:

```
>> READ 'oppg1.13' 'data'
>> ARIMA 0 1 1 'data'; FORECAST 2 periods.
ARIMA Model: 1 regular difference
No. of obs.: 50, after differencing 49
 Parameter  Estimate  St.Dev  t-ratio
 MA 1         0.5278  0.1224     4.31
Residuals: MS = 0.0208 (DF=48)
Statistics for model fit (Ljung-Box)
 Lag                 12           24           36
 Chisquare  6.0(DF=11)  18.5(DF=23)  33.9(DF=35)
Forecasts from period 50
                    95 Percent Limits
 Period  Forecast    Lower     Upper  Actual
   51     3.12853  2.84592   3.41115
   52     3.12853  2.81600   3.44106
```

Den estimerte modell er

$$Y_t = Y_{t-1} + 0.5278 \cdot U_{t-1} + U_t$$

Vi ser at MA-parameteren er statistisk signifikant og at modellen gir god tilpasning (små kjikvadratverdier i forhold til frihetsgradtallet). En modell der differensene er AR(1), noe som ikke er helt utelukket ut fra ACF-plottet, gir dårligere tilpasning. Utskriften gir også (like) prediksjoner for to perioder fremover, samt øvre og nedre grenser med 95% garanti.

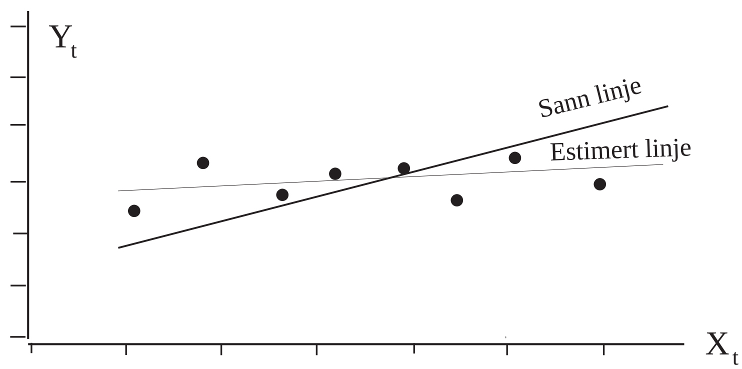
13.4 Autokorrelasjon og regresjon

I praksis ønsker en ofte å forklare og/eller predikere en tidsrekke ved andre observerte tidsrekker, i tillegg til trend og sesongvariasjon. Det gjelder ikke minst i økonomisk sammenheng, der vi observerer resultat ("output") i suksessive perioder, som dels er en følge av egen innsats ("input") i perioden og/eller tidligere perioder, og dels av omstendigheter som er utenfor vår kontroll.

Dette åpner for et bredt spekter av mulige forklaringsmodeller, som det vil føre for langt å gi en balansert fremstilling av her. Temaet er grundig behandlet i økonometrisk litteratur. Vi vil begrense oss til å gi en kort omtale av tidsrekker som forklares ved enkle regresjonsmodeller, med vekt på hvilke problemer autokorrelerte feilledd kan medføre. Dette faller utenfor rammen av standardregresjonsmodellen beskrevet i Kapittel 12.3. Vi kan derfor ikke uten videre bruke resultatene knyttet til denne. Autokorrelasjon kan medføre feiltolkning av observasjonene, bl.a. ved risiko for at regresjonskoeffisienter er grovt feilestimert på tross av at vi har fått fastlagt en modell med tilsynelatende god tilpasning til observasjonene. La oss illustrere problemet med den enkle regresjonsmodellen

$$Y_t = \gamma + \beta X_t + U_t \quad t = 1, 2, \ldots, n$$

der vi antar at feilleddet U_t er positivt autokorrelert.



Figur 13.6: Regresjon og positiv autokorrelasjon

I Figur 13.6 har vi tegnet inn den sanne regresjonslinjen med hel strek, samt en tidsrekke som svarer til at feilleddene har positiv autokorrelasjon [2], dvs. positive (negative) avvik fra regresjonslinje etterfølges gjennomgående av positive (negative) avvik. Vi har gått ut fra at det første avviket fra regresjonslinjen tilfeldigvis var positivt, og tidsrekken vil da typisk kunne ha et forløp som vist på figuren. Vi ser at analyse basert på den observerte rekken uten kunnskap om autokorrelasjon, vil gi en estimert linje som stiplet på figuren. Vi ser at vi har underestimert regresjonskoeffisienten β. Tilsvarende dersom vi tilfeldigvis startet med et negativt feilledd ville vi antakelig overestimere regresjonskoeffisienten. Det som er spesielt infamt er at vi tilsynelatende har fått god tilpasning til en rett linje, men linjen ligger altså ikke på rett sted. Et forsøk på å estimere variansen til feilleddet på grunnlag av residualene vil, med den rekken som er illustrert i figuren, totalt undervurdere denne. Vi vet heller ikke om vi har eventuelt over- eller underestimert regresjonskoeffisienten. Dersom vi hadde fortsatt å observere tidsrekken en tid ville den høyst sannsynlig krysse den sanne linjen igjen, slik at sjansen for å gå i fellen er redusert.

De forhold som her er belyst har en viss betydning for den analyse av trend og sesongvariasjon som ble foretatt i avsnitt 13.2 basert på regresjonsmodellen, og kan

[2] Her var forklaringsvariablen tiden selv, men situasjonen er i prinsippet den samme for alle forklaringsvariable med verdier som gjennomgående er korrelert med t.

også forkludre andre analysemetoder. Dersom det også er autokorrelasjon til stede, vil denne kunne feilaktig bli tolket som trend eller sesongvariasjon, selv om vi har fått god lineær tilpasning til modellen.

Det finnes ulike måter å unngå slike feller som er beskrevet ovenfor, men disse vil avhenge noe av hvilke typer tidsavhengighet man ønsker å gardere seg mot. Vi vil her kort se på situasjonen dersom man er villig til å representere feilleddet med en første ordens autoregressiv modell, dvs.

$$U_t = \rho U_{t-1} + V_t \quad t = 1, 2, \ldots, n$$

der V_t'ene er uavhengige avvik med forventning null og samme varians. Med utgangspunkt de observerte rekkene $Y_1, Y_2, \ldots, Y_n$ og $X_1, X_2, \ldots, X_n$ beregner vi

$$Y_t^{'} = Y_t - \rho Y_{t-1} \quad X_t^{'} = X_t - \rho X_{t-1}$$

Vi ser at ifølge regresjonsmodellen $Y_t = \gamma + \beta X_t + U_t$, kan vi skrive [3]

$$Y_t^{'} = \gamma(1-\rho) + \beta X_t^{'} + V_t$$

slik at vi har en regresjon som oppfyller antakelsene i standardregresjonsmodellen, der β kan estimeres på vanlig måte uten problemer.

Den metode som er foreslått ovenfor lar seg gjennomføre dersom vi kjenner ρ, men dette er som regel ikke tilfelle. Vi merker oss at regresjonskoeffisienten i den avledede modell er β uansett hvilken ρ som brukes. Det er derfor aktuelt å bruke $\rho = 1$, dvs. sette $Y_t^{'} = Y_t - Y_{t-1}$ og $X_t^{'} = X_t - X_{t-1}$. Det viser seg at faren for feilestimering av β nå er betydelig redusert, uansett hva ρ virkelig er. Et alternativ til dette vil være å estimere ρ på grunnlag av (en del av) den observerte rekke. Ut fra den autoregressive modellen for feilleddet U_t, ser det ut til at vi kan estimere ρ på samme måte som tilfellet var i avsnitt 13.3. Dette forutsetter imidlertid at feilleddene U_t lar seg observere, men det kan de jo ikke. Det beste vi kan gjøre synes å være å estimere regresjonslinjen $\hat{Y}_t = \hat{\gamma} + \hat{\beta} \cdot X_t$ på vanlig måte med minste kvadraters metode, og deretter beregne residualene $\hat{U}_t = Y_t - \hat{Y}_t$ og bruke disse istedenfor U_t ved estimering av ρ. Dette gir estimatoren

$$\hat{\rho} = \frac{\sum_{t=2}^{n} \hat{U}_{t-1}\hat{U}_t}{\sum_{t=1}^{n} \hat{U}_t^2}$$

Av Figur 13.6 ser vi imidlertid at denne antakelig vil underestimere ρ idet variasjonen omkring den estimerte linje ser mer tilfeldig ut enn omkring den sanne linjen. Det finnes metoder som korrigerer denne underestimeringen.

Vi vil ofte ved gjennomføringen av en regresjonsanalyse etter vanlige metoder være interessert i å teste om feilleddene med rimelighet kan antas uavhengige, slik at forutsetningene i standardregresjonsmodellen er oppfylt. Vår mulighet ligger i å

[3]Vi ser at $Y_t^{'}$ og $X_t^{'}$ ikke lar seg beregne for $t = 1$. Det er da vanlig å isteden sette $Y_1^{'} = \sqrt{1-\rho^2} Y_1$ og $X_1^{'} = \sqrt{1-\rho^2} X_1$, hvorfor kan vi ikke gå inn på her.

se på residualene $\hat{U}_t$ etter regresjonsanalysen. Den testen som vanligvis utføres er *Durbin-Watsons test* som forestiller seg at alternativet til hypotesen om uavhengighet er første ordens autoregresjon. Problemstillingen kan derfor oppfattes som å teste $\rho = 0$ mot alternativet $\rho > 0$ (evt. $\rho < 0$). Durbin-Watson testobservatoren er (jfr. Oppgave 17)

$$D = \frac{\sum_{t=2}^{n} \left(\hat{U}_t - \hat{U}_{t-1}\right)^2}{\sum_{t=1}^{n} \hat{U}_t^2} \approx 2(1 - \hat{\rho}).$$

Liten verdi av D gir grunn til å påstå positiv autokorrelasjon. Tabell C.13 i Appendiks C gir kritiske verdier for D svarende til 5% signifikansnivå. En D lavere enn tabellverdien d_L betyr klar forkastning av hypotesen om uavhengighet til fordel for positiv autokorrelasjon. En verdi mellom d_L og d_U gir svak indikasjon på positiv autokorrelasjon, mens en verdi mellom d_U og 2 gir klar indikasjon på at hypotesen om uavhengighet er rimelig. Med negativ autokorrelasjon som (ensidig) alternativ brukes samme rettesnor, men med $D^{'} = 4 - D$ som testobservator istedenfor D.

De forhold som er diskutert ovenfor gjelder også for regresjonsanalyse med mer enn en forklaringsvariabel, og testobservatoren for testing av uavhengige feilledd har samme form. Sannsynlighetsfordelingen til D vil avhenge av antall forklarende variable r, og i Tabell C.13 er de kritiske verdier gitt i separate kolonner for ulike r. Durbin-Watson testen med rapport av P-verdier fins som regel som opsjon i programvare.

I praksis er det nok bare i tilfellet med klar forkasting av hypotesen om uavhengighet at en standard regresjonsanalyse diskrediteres, og en ser seg om etter en alternativ metode som fanger opp autokorrelasjonen, f.eks. som antydet ovenfor. I situasjonen med svak autokorrelasjon har man ingen garanti for at en alternativ metode gir et mer pålitelig resultat, men det er et spørsmål om sunt omdømme.

Eksempel 9 : Omsetning

Vi studerer igjen tidsrekken bestående av kvartalsvise omsetningstall som er gitt i Eksempel 3, og analysert i Eksempel 4 på grunnlag av en regresjonsmodell med fire forklaringsvariable. Beregning av residualene ga som resultat (sjekk og tegn figur!)

0.62	0.42	−0.31	−0.97	0.00	1.00	−0.53
0.00	−0.63	−1.42	0.84	0.97		

Det gir $D = 1.69$ som ifølge Tabell C.13 ikke er tilstrekkelig lite til å påstå positiv autokorrelasjon. Vi føler oss derfor noenlunde trygge på den analysen som ble foretatt i Eksempel 4, og korreksjon for autokorrelasjon synes derfor ikke aktuelt.

La oss ta for oss et avsluttende mer realistisk eksempel.

Eksempel 10 : Salg-Reklame

La oss igjen ta for oss dataene fra Eksempel 1.9, der vi observerte sammenhørende verdier av reklameinnsats X (i tusen kroner) og solgt kvantum Y i etterfølgende perioder. Anta at det er kvartalsvise data over tre år :

Periode:	1	2	3	4	5	6	7	8	9	10	11	12
Kvartal:	1	2	3	4	1	2	3	4	1	2	3	4
X:	4.0	2.0	2.5	2.0	3.0	5.0	4.0	2.0	5.0	4.0	2.5	5.0
Y:	385	400	395	365	475	440	490	420	560	525	480	510

La oss tenke oss muligheten av at det i tillegg til reklameeffekten er både lineær trend og sesongvariasjon. En slik analyse, som samtidig kan indikere om det skulle være autokorrelasjon, er gitt i utskrift nedenfor.

Med reklame som eneste forklaringsvariabel var denne klart signifikant. Durbin-Watson observatoren er D=1.35, som indikerer en svak positiv autokorrelasjon (ikke signifikant). Vi ser til vår forundring at tidstrend har slått ut reklamen som forklarende variabel i den utvidede modellen. Videre ser vi at 4. kvartal er lavsesong i forhold til 1. kvartal, mens 2. og 3. kvartal er mellomsesong, som riktignok ikke er signifikant forskjellig fra 1. kvartal. Forklaringsgraden er også økt fra 40.3% til hele 90.5%, mens standardavviket til feilleddet er redusert fra S=50.23 til S=25.86. Dette indikerer at den siste modellen kan ha betydelig bedre prediksjonsevne. Endelig ser vi at Durbin-Watson observatoren er D=2.65, som indikerer en svak negativ autokorrelasjon (ikke signifikant).

```
 >> READ 'eks1.9' X Y
>> SET T = 1:12
>> SET S = 3(1:4)
>> INDICATOR S in S1 S2 S3 S4
>> REGRESS Y on X ; DW.
Y = 344 + 32.2 X (S=50.23)
                          St.Dev    T-ratio
 Variable  Coefficient  of Coef.  Coef/S.D.      P
                343.71     44.77       7.68  0.000
        X        32.21     12.40       2.60  0.027
Degrees of freedom for T-test = 10
R-squared = 40.3% (Adjusted for d.f. = 34.3%)
Durbin-Watson statistic = 1.35
>> REGRESS Y on X T S2 S3 S4 ; DW.
Y = 368 + 7.24 X + 15.2 T
       - 31.1 S2 - 41.5 S3 - 80.0 S4 (S=25.86)
                          St.Dev    T-ratio
 Variable  Coefficient  of Coef.  Coef/S.D.      P
                368.34     31.67      11.63  0.000
        X        7.241     8.319       0.87  0.418
        T       15.205     2.768       5.49  0.002
       S2       -31.12     21.68      -1.44  0.201
       S3       -41.50     24.45      -1.70  0.141
       S4       -80.04     25.73      -3.11  0.021
Degrees of freedom for T-test = 6
R-squared = 90.5% (Adjusted for d.f. = 82.6%)
Durbin-Watson statistic = 2.65
```

Vi bør imidlertid summe oss litt her : Studerer vi tallene litt nærmere, ser vi at det er en tendens til at reklameinnsatsen er økt over tid. Det er derfor vanskelig å avgjøre om økt salg skyldes reklameinnsats, tidstrend eller begge deler. Videre ser vi at reklameinnsatsen gjennomsnittlig var større i 1. kvartal enn de øvrige, og det er derfor vanskelig å si om det virkelig er en sesongeffekt eller om den delvis skyldes mer reklame.

Dette eksemplet viser igjen hvor vanskelig det er å trekke generelle konklusjoner i situasjoner der data ikke er planmessig innsamlet. Hvis vi virkelig ønsket å undersøke reklameeffekten alene, må vi søke å variere reklamen over tiden og for de ulike kvartaler i henhold til en plan som unngår at de forklarende variable blir korrelerte.

13.5 Oppgaver

1. Dekomponer tidsrekken i Eksempel 3 etter modellen $Y_t = T_t + S_t + U_t$ og gi prognoser for kommende år. Drøft valget av dekomponeringsmetode for denne rekken.

2. Gitt følgende tidsrekke med kvartalsvise økonomiske data over en periode på fire år

Y_t:	14.0	16.2	18.3	20.5	15.6	22.1	18.4	27.2
	18.9	22.0	24.8	26.8	16.9	24.0	26.9	27.7

   (a) Tegn denne tidsrekken i en figur, og tegn inn eventuell trend.

   (b) Dekomponer tidsrekken etter modellen $Y_1 = T_t S_t U_t$, der sesongperioden er et år.

   (c) Gi en prognose for tidsrekkens verdier neste år.

3. Anta at tidsrekken i Oppgave 2 er en fortsettelse av tidsrekken i Eksempel 3, slik at vi har observert rekken i alt over en periode på syv år. Vi vet at i skillet mellom de to observasjonsperiodene fant det sted en endring i de økonomiske forhold.

   (a) Tyder observasjonene på at trenden er forskjellig i de to periodene ? Hvis ikke anslå en felles trend for hele materialet.

   (b) Tyder observasjonene på at sesongfaktoren kan være lik for de to periodene. Fastlegg i så fall sesongfaktorene på grunnlag av hele materialet og gi nye prognoser for neste år.

4. Gitt følgende tidsrekke som er en alternativ fortsettelse av tidsrekken i Eksempel 3 :

Y_t:	12.3	13.5	16.0	20.4	15.8	17.1	18.0	27.7
	20.8	24.3	26.4	30.0	24.0	30.8	24.3	34.8

   Drøft om tidsrekken kan ha endret karakter ved tidsskillet.

5. Enkelte tidsrekker kan ved tidsskiller få et skift, dvs. at rekken med utgangspunkt i et nytt nivå i hovedsak utvikler seg som før.

   (a) Forklar at dette kan analyseres innenfor en lineær modell, f.eks. slik

$$Y_t = \beta_0 + \beta_1 t + \beta_2 Q + \beta_3 I + U$$

   der I er en indikatorvariabel med verdi 0 før tidsskillet og 1 etter.

   (b) Tolk parametrene i modellen i (a).

   (c) $\star$Utfør en analyse basert på en modell av denne typen (med tre sesongvariable) og med hele tallmaterialet fra Eksempel 3 og Oppgave 4.

6. Drøft hvilke problemer som oppstår ved analyse av tidsrekker der vi på forhånd ikke vet om eventuelle tidsskiller der rekken kan ha endret karakter.

7. Det foreligger en tidsrekke Y_t av kvartalsvise data med trend og sesongvariasjoner der vi vet at sesongkomponenten har en syklisk karakter, dvs. jevnt stigende fra lavsesong til høysesong og ned igjen over en periode på et år.

   (a) Drøft bruk av følgende modell

   $$Y_t = \beta_0 + \beta_1 t + \beta_2 sin(2\pi k/4) + U_t$$

   der k betegner kvartal nr.$k = 1, 2, 3, 4$.

   (b) Lag tilsvarende modell dersom tidsrekken er observert månedsvis.

   (c) Hvilken teknikk er aktuell for analyse av slike modeller?

8. Gitt følgende tidsrekke som antas å være uten trend og sesongvariasjoner :

Y_t:	0.9	2.2	1.2	−0.9	−0.3	−1.7	0.6	1.7	−0.8	−2.3
	−1.5	−0.7	2.1	0.6	−0.8	−1.9	−1.2	0.8	1.4	1.0

   (a) Viser denne rekken tegn til autokorrelasjon?

   (b) Estimer en første orden autoregressiv modell.

   (c) Prediker de to neste verdiene av rekken.

9. Betrakt AR(1) modellen

   $$Y_t = \gamma + \rho Y_{t-1} + U_t$$

   der U_t'ene er uavhengige med samme varians τ^2. Vis at (ubetinget) forventning og varians er gitt ved

   $$\text{(i)}\quad \mu = \gamma/(1-\rho) \qquad \text{(ii)}\quad \sigma^2 = \tau^2/(1-\rho^2)$$

   Kan du også si noe om betingede forventninger og varianser?

10. Betrakt en *"random walk"*

   $$Y_t = Y_{t-1} + U_t$$

   der U_t'ene er uavhengige med samme varians τ^2. Undersøk egenskapene: forventning, varians etc.

11. Vis at en AR(1) tidsrekke har autokovariansfunksjon med form

   $$c(k) = \rho^k c(0) \quad k = 0, 1, 2, \ldots$$

   Kommenter resultatet.

12. Vis at en MA(1) tidsrekke har autokovariansfunksjon med form

$$c(0) = (1+\alpha^2)\cdot\tau^2,\ \ c(1) = \alpha\cdot\tau^2,\ \ c(k) = 0 \quad \text{for} \quad k \geq 2.$$

    Kommenter resultatet.

13. Anta en første ordens autoregressiv tidsrekke $Y_t = \rho Y_{t-1} + U_t$ der ρ er kjent, og U_t'ene uavhengige variable med forventning null og varians σ^2.

    (a) Rettferdiggjør bruk av $\hat{Y}_{n+k} = \rho^k Y_n$ som prediktor k skritt fram i tid fra tidspunkt n.

    (b) Vis at prediksjonsfeilen $\hat{U}_{n+k} = Y_{n+k} - \hat{Y}_{n+k}$ har forventning null og varians

    $$(1+\rho^2+\rho^4+\cdots+\rho^{2(k-1)})\sigma^2$$

    Kommenter resultatet.

14. Som prognose for verdien av en tidsrekke ett skritt fram i tid, er valgt den sist observerte verdi dvs.

$$\hat{Y}_{t+1} = Y_t \quad t = 1,2,3,\ldots$$

    Finn forventningen og variansen til prediksjonsfeilen $\hat{U}_{t+1} = Y_{t+1} - \hat{Y}_{t+1}$ dersom

    (a) $Y_t = \gamma + \beta t + U_t$ (lineær trend)

    (b) $Y_t = \gamma + \rho Y_{t-1} + U_t$ (autoregresjon)

    (c) $Y_t = \mu + \alpha U_{t-1} + U_t$ (glidende gjennomsnitt)

    der U_t'ene er uavhengige variable med forventning null og varians τ^2.

15. En prognosemetode kalt *eksponensiell glatting* er gitt ved formelen

$$\hat{Y}_{t+1} = aY_t + (1-a)\hat{Y}_t$$

    dvs. prognosen på tidspunkt t for neste verdi av tidsrekken er en veid sum av observert verdi av rekken på tidspunkt t og prognosen for dette tidspunkt foretatt på $t-1$. Dersom vi starter med prognoser på tidspunkt n trengs en startverdi $\hat{Y}_n$ som velges etter beste skjønn.

    (a) Drøft valg av a.

    (b) Fyll inn prognoser og sammenlign med de virkelig observerte verdier dersom vi velger (i) $a = 0.2$ (ii) $a = 0.5$

t:	n	n+1	n+2	n+3	n+4	n+5
$\hat{Y}_t$:	0.0	(...)	(...)	(...)	(...)	(...)
Y_t:	1.0	1.0	0.0	1.0	2.0	2.0

(c) Vis at denne prognosemetoden ikke er brukbar dersom tidsrekken har negativ autokorrelasjon.

(d) Vis at denne prognosemetoden ikke gir gode resultater for tidsrekker med trend (jfr. Oppgave 14).

16. Følgende modell for autokorrelasjon er foreslått

$$Y_t = \beta Y_{t-1} + \beta(1-\beta)Y_{t-2} + \beta(1-\beta)^2 Y_{t-3} + \cdots + U_t$$

der β er tall mellom 0 og 2 som uttrykker graden av autokorrelasjon.

(a) Tolk denne modellen.

(b) Under hvilke forutsetninger er det rimelig å bruke prediktoren

$$\hat{Y}_t = \beta Y_{t-1} + \beta(1-\beta)Y_{t-2} + \beta(1-\beta)^2 Y_{t-3} + \cdots$$

(c) Vis at

$$\hat{Y}_{t+1} = \beta Y_t + (1-\beta)\hat{Y}_t$$

og kommenter resultatet i lys av prognosemetoden i forrige oppgave.

17. $Y_1, Y_2, \ldots, Y_n$ er uavhengige med forventning μ og varians σ^2. Betrakt

$$D = \frac{C^2}{S^2} = \frac{\frac{1}{n-1}\sum_{t=2}^{n}\left(Y_t - Y_{t-1}\right)^2}{\frac{1}{n-1}\sum_{t=1}^{n}\left(Y_t - \overline{Y}\right)^2}$$

Vis at $EC^2 = 2\sigma^2$ og forklar at D antar verdier mellom 0 og 4. Forklar videre at $D \approx 2$ indikerer uavhengighet, mens verdier som er noe mindre/større enn 2 indikerer hhv. positiv og negativ autokorrelasjon (av første orden).

18. Gitt følgende tidsrekke:

Y_t:	5.9	7.3	8.5	9.2	9.9	10.6	11.8	12.3	13.7	15.2
	16.7	17.4	18.7	19.2	20.1	21.2	21.8	22.5	23.7	24.4

Som modell er valgt $Y_t = \gamma + \beta t + U_t$, dvs. lineær trend uten sesongvariasjoner.

(a) Tegn tidsrekken i en figur og drøft om modellen synes rimelig.

(b) Estimer minste kvadraters regresjonslinje.

(c) Beregn residualene og drøft om feilleddene U_t kan være autokorrelerte.

(d) Utfør Durbin-Watsons test for uavhengighet.

19. Utfør flere regresjonsanalyser på dataene i Eksempel 10.

(a) med reklame og tid som forklarende variable, uten å ta omsyn til sesongvariable,

(b) med reklame i forrige periode som ekstra forklaringsvariabel.

Kommenter resultatene.

20. For å bli kjent med karakteren av ulike typer tidsrekker, kan disse simuleres med programvare på en PC. Samtidig kan en studere de relative fortrinn av ulike prognosemetoder under ulike omstendigheter. Studer mulighetene for å foreta slike analyser med den programvare du har til rådighet.

21. Undersøk tilgjengelig statistisk programvare kan tilby analyse og prediksjon av tidsrekker . Finn ut hvilke typer modeller som ligger til grunn. Undersøk også om hvilke elementer, trend, sesongvariasjon, autokorrelasjon osv. som programvaren fanger opp.

Kapittel 14

Utvalgsundersøkelser

14.1 Innledning

En utvalgsundersøkelse er karakterisert ved følgende: Vi har en populasjon bestående av N elementer (objekter eller individer), og vi ønsker informasjon om disse elementene. Dersom N er stor vil det ofte være uhensiktsmessig å observere alle elementene i populasjonen, det kan f.eks. være for dyrt, tidkrevende eller praktisk umulig. Isteden trekkes et utvalg på et mindre antall elementer n, og konklusjoner om populasjonen som helhet trekkes på grunnlag av dette utvalget.

Den enkleste problemstilling av denne typen har vi når de N elementene i populasjonen kan beskrives med verdier, henholdsvis $v_1, v_2, \ldots, v_N$, og vi ønsker å trekke slutninger om gjennomsnittet $\overline{v}$ på grunnlag av et utvalg av elementer fra populasjonen. For det tilfellet at utvalget er trukket tilfeldig vil lotterimodellen beskrevet i Kapittel 8.2 kunne brukes til å vurdere usikkerheten vedrørende de konklusjoner som trekkes.

I praksis vil en ofte ikke kunne realisere et tilfeldig utvalg. Dette kan ha forskjellige årsaker, den mest vanlige er at det vil bli for kostbart og tidkrevende å gjennomføre. Statistikere har da foreslått en rekke alternative framgangsmåter som kan komme til anvendelse under ulike omstendigheter. Felles for disse er at de bygger på modeller som er generaliseringer i ulike retninger av lotterimodellen. En framgangsmåte for å trekke et utvalg kalles av fagfolkene ofte for en *utvalgsplan* eller *design*. En design karakteriseres som god i relasjon til det foreliggende problem, dersom den lar seg praktisk gjennomføre innenfor rammen av våre ressurser, og samtidig leder til tillitvekkende konklusjoner.

Vi vil nedenfor se på noen aktuelle utvalgsplaner og analysemetoder. Vi vil bl.a. vise hvordan tilleggsinformasjon om populasjonen kan utnyttes til å oppnå mer pålitelige konklusjoner.

I forbindelse med utvalgsundersøkelser er det også aktuelt å trekke slutninger om proporsjoner, f.eks. brøkdelen av velgere som favoriserer et bestemt politisk parti. Dette kan ses på som et spesialtilfelle av teorien i dette kapitlet, hvor hvert element i populasjonen har v-verdi som enten er 0 eller 1, eksempelvis 0 dersom velgeren ikke favoriserer partiet og 1 dersom velgeren favoriserer partiet. Da er $\overline{v}$ brøkdelen av tilhengere i populasjonen. Imidlertid vil estimering av proporsjoner også romme en del aspekter som ikke er dekket av teorien nedenfor.

14.2 Stratifisering

La oss først forfølge en tankegang som kalles *stratifisering*:

Anta at vi skal estimere gjennomsnittet $\overline{v}$ av de N ukjente verdiene $v_1, v_2, \ldots, v_N$ i populasjonen. Istedenfor å velge ut n elementer tilfeldig fra hele populasjonen tenker vi oss at denne er delt opp i et visst antall delpopulasjoner, kalt *strata*, med henholdsvis $N_1, N_2, \ldots, N_k$ elementer i hver, slik at $N_1 + N_2 + \cdots + N_k = N$, og at vi trekker et tilfeldig utvalg på henholdsvis $n_1, n_2, \ldots, n_k$ elementer fra hvert stratum slik at $n_1 + n_2 + \cdots + N_k = n$. En viktig grunn for å foreta stratifisering har vi dersom det er visse typer elementer i populasjonen som vi ønsker skal være representert i utvalget, en annen grunn følger av teorien nedenfor.

Vi merker oss at vi kan skrive

$$\overline{v} = \frac{N_1}{N}\overline{v}_1 + \frac{N_2}{N}\overline{v}_2 + \cdots + \frac{N_k}{N}\overline{v}_k$$

der $\overline{v}_i$ er gjennomsnittet av v-verdiene for elementene i stratum nr. i. En rimelig estimator for $\overline{v}_i$ vil være gjennomsnittet $\overline{Y}_i$ av v-verdiene fra de utvalgte elementene fra stratum nr. i, dvs.

$$\overline{Y}_i = \frac{1}{n_i} = \sum_{j=1}^{n_i} Y_{ij}$$

der Y_{ij} betegner j'te observasjon fra stratum nr. i. Som estimator for gjennomsnittet av v-verdiene i hele populasjonen $\overline{v}$ er det rimelig å bruke

$$\tilde{Y} = \frac{N_1}{N}\overline{Y}_1 + \frac{N_2}{N}\overline{Y}_2 + \cdots \frac{N_k}{N}\overline{Y}_k.$$

Innen hvert stratum velges et tilfeldig utvalg, og vi kan her bruke resultatene som er utviklet for lotterimodellen, nemlig

$$E\overline{Y}_i = \overline{v}_i$$

$$var\overline{Y}_i = \frac{N_i - n_i}{N_i - 1} \cdot \frac{\sigma_i^2}{n_i} \approx \frac{\sigma_i^2}{n_i}$$

der tilnæmelsen er god når n_i er liten i forhold til N_i. Her er σ_i^2 variansen svarende til en enkelt observasjon fra stratum nr. i. Av dette følger at

$$E\tilde{Y} = \overline{v} \quad \text{dvs. estimatoren er forventningsrett.}$$

Vi antar at trekningene for de ulike strata skjer uavhengig av hverandre. Herav følger at

$$var\tilde{Y} \approx \frac{N_1^2}{N^2} \cdot \frac{\sigma_1^2}{n_1} + \frac{N_2^2}{N^2} \cdot \frac{\sigma_2^2}{n_1} + \cdots \frac{N_k^2}{N^2} \cdot \frac{\sigma_k^2}{n_k}$$

Det er ofte rimelig å trekke ut et antall fra hvert stratum proporsjonalt med antall elementer totalt i stratumet, dvs. velge [1]

$$n_1 = n \cdot \frac{N_1}{N}, \quad n_2 = n \cdot \frac{N_2}{N}, \ldots, \quad n_k = n \cdot \frac{N_k}{N}$$

Dette kalles *proporsjonal utvelging.* I denne situasjon forenkles variansformelen til

$$var\tilde{Y} \approx \frac{1}{n}(\frac{N_1}{N} \cdot \sigma_1^2 + \frac{N_2}{N} \cdot \sigma_2^2 + \cdots + \frac{N_k}{N} \cdot \sigma_k^2)$$

La oss sammenligne estimatoren $\tilde{Y}$ i tilfellet proporsjonal utvelging med estimatoren $\overline{Y}$ basert på vanlig tilfeldig utvalg på n elementer fra hele populasjonen. Vi husker at $\overline{Y}$ også var forventningsrett og med varians

$$var\overline{Y} \approx \frac{\sigma^2}{n}$$

For det tilfellet at gjennomsnittet og variansen er den samme for hvert stratum, $\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_k^2 = \sigma^2$, vil høyresiden i de to siste variansformlene være like. I alle andre tilfeller vil høyresiden for $var\tilde{Y}$ være mindre enn høyresiden for $var\overline{Y}$. Mer konkret viser det seg at

$$var\overline{Y} \approx var\tilde{Y} + \frac{N_1}{N}(\overline{v}_1 - \overline{v})^2 + \frac{N_2}{N}(\overline{v}_2 - \overline{v})^2 + \cdots + \frac{N_k}{N}(\overline{v}_k - \overline{v})^2$$

[1] Vi må i praksis velge $n_1, n_2, \ldots, n_k$ som de nærmeste positive heltall med sum n.

som viser at reduksjonen i varians blir større ettersom $\overline{v}_1, \overline{v}_2, \ldots, \overline{v}_k$ avviker mer og mer fra det totale gjennomsnittet $\overline{v}$.

Det er interessant å legge merke til at ved proporsjonal utvelging er estimatoren $\tilde{Y}$ også lik gjennomsnittet av de observerte verdier, og det som gir variansreduksjonen er altså at utvalgsplanen er en annen.

I praksis kan resultatene ovenfor utnyttes slik: På forhånd deles populasjonen i strata slik at elementer innen hvert stratum er mest mulig homogene, eller ekvivalent, at gjennomsnittene av v-verdiene fra stratum til stratum er avvikende. Siden vi ikke kjenner eksakte v-verdier til elementene i populasjonen må dette gjøres etter beste skjønn, f.eks. ut fra visse kjennetegn som er knyttet til disse elementene. Reduksjonen av variansen ved bruk av $\tilde{Y}$ som estimator istedenfor $\overline{Y}$ vil avhenge av i hvilken grad vi lykkes å utføre denne stratifiseringen i henhold til disse ønsker, men poenget er at vi har intet å tape ved å forsøke en stratifisert utvelging. Noen eksempler:

- Ved en undersøkelse av forbruksvaner hos en gruppe av befolkningen, kan vi stratifisere m.h.t. kjønn, bosted og/eller inntektsnivå.
- Ved undersøkelse av gjennomsnittsverdi av kjøp i et varehus, kan vi stratifisere m.h.t. hvilken avdeling de enkelte kjøp fant sted.
- Ved undersøkelse av det gjennomsnittlige antall kuer hos gårdbrukere over hele landet, kan vi stratifisere etter fylke, gårdens størrelse, om brukeren har biinntekt eller ikke.
- I revisjon kan en stratifisere etter bokført verdi.

I disse situasjonene har vi en viss tro på at v-verdiene innen hvert stratum er mer homogene enn i populasjonen som helhet. La oss regne på et konkret enkelt eksempel, i en realistisk anvendelse vil selvsagt populasjonen være mye større.

Eksempel 1 : Røking
På en arbeidsplass er det 10 røkere, 6 menn og 4 kvinner. Vi ønsker å anslå gjennomsnittlig antall sigaretter som er røkt pr. røker en bestemt dag på grunnlag av et utvalg av $n = 5$ røkere. Anta at vi deler populasjonen i to strata etter kjønn, idet vi av erfaring vet at kvinnene røker gjennomgående noe mindre enn mennene. Anta at populasjonen består av følgende tall som er ukjente for oss.

Stratum 1:	8	10	5	7	4	9
Stratum 2:	4	2	4	6		

Ved proporsjonal utvelging trekkes tilfeldig 3 menn og 2 kvinner. En tabell over tilfeldige tall lar oss observere person nr. 1, 3 og 6 fra stratum nr. 1, og person nr. 1 og 4 fra stratum nr. 2. Vi får

$$\begin{aligned} \overline{Y}_1 &= (8+5+9)/3 = 7.3 \\ \overline{Y}_2 &= (4+6)/2 = 5.0 \end{aligned}$$

slik at

$$\tilde{Y} = \frac{6}{10}\overline{Y}_1 + \frac{4}{10}\overline{Y}_2 = \frac{6}{10}7.3 + \frac{4}{10}5.0 = 6.4$$

Anta at vi isteden velger de $n = 5$ tilfeldig fra hele populasjonen. En tabell over tilfeldige tall lar oss da observere person 2, 3, 4 og 6 blant mennene og person nr. 2 blant kvinnene, slik at vi får

$$\overline{Y} = (10+5+7+9+2)/5 = 6.6$$

På grunn av teorien ovenfor vil vi foretrekke å bruke den første framgangsmåten. Vi ser for øvrig at gjennomsnittet i populasjonen virkelig er 5.9, noe som vi selvfølgelig ikke vet ut fra det utvalg som er foretatt.

Når en skal vurdere usikkerheten ved et beregnet estimat, ønsker vi å bruke standardavviket til estimatoren. Igjen må dette estimeres. Vi kan estimere variansen innen hvert stratum på vanlig måte, dvs. vi beregner

$$S_i^2 = \frac{1}{n_i}\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y}_i)^2$$

hvor Y_{ij} er den j'te observasjon fra stratum nr. 1. Med utgangspunkt i formelen for var $\tilde{Y}$ ovenfor kan vi få en estimator for denne variansen ved å erstatte σ_i^2 med S_i^2. Vi får da estimatoren

$$S^2 = \frac{N_1^2}{N^2}\cdot\frac{S_1^2}{n_1} + \frac{N_2^2}{N^2}\cdot\frac{S_2^2}{n_1} + \cdots \frac{N_k^2}{N^2}\cdot\frac{S_k^2}{n_k}$$

som i tilfellet med proporsjonal utvelging reduseres til

$$S^2 = \frac{1}{n}(\frac{N_1}{N}\cdot S_1^2 + \frac{N_2}{N}\cdot S_2^2 + \cdots + \frac{N_k}{N}\cdot S_k^2)$$

Som estimator for standardavviket brukes da S.

Man kan spørre om det er mulig å finne en estimator som reduserer variansen ytterligere i forhold til proporsjonal utvelging. Det viser seg at $var\tilde{Y}$ er minst mulig når utvalgsstørrelsen n_i velges, ikke proporsjonal med N_i, men proporsjonal med $N_i\sigma_i$. Dette kalles *optimal utvelging*. Slik utvelging krever imidlertid kunnskap om σ_i, noe som sjelden er tilgjengelig. Dersom vi vet at variabiliteten innen et bestemt stratum er stor i forhold til andre strata, kan det være lønnsomt å øke utvalgsstørrelsen fra dette i forhold til vanlig proporsjonal utvelging. Imidlertid må det ekstreme forhold til før en tjener noe vesentlig på å gjøre dette.

I noen situasjoner kan kostnadene ved å velge ut elementer variere fra stratum til stratum. Det finnes også teorier for bestemmelse av utvalgsstørrelser hvor en tar hensyn til dette.

Vi har for å forenkle teorien ovenfor tatt utgangspunkt i formler hvor korreksjonsfaktorer av typen $(N-n)/(N-1)$ er utelatt og de funne resultater gjelder da strengt tatt bare når utvalgsstørrelsene innen hvert stratum er lite i forhold til det totale antall i hvert stratum. Dette vil også være tilfellet i mange anvendelser. Imidlertid vil den konklusjon at det kan være lønnsomt å stratifisere ha mer generell gyldighet, og i situasjoner der utvalgsstørrelsene forholdsvis ikke er små, kan det være aktuelt å ta med korreksjonsfaktorer av typen ovenfor.

14.3 Klyngeutvalg

I enkelte situasjoner er det ikke hensiktsmessig å gjennomføre et rent tilfeldig utvalg, ei heller et stratifisert utvalg. Dette kan skyldes praktiske omstendigheter og/eller høye kostnader. Eksempelvis dersom en ønsker å gjennomføre en forbrukerundersøkelse som gjelder husstandene i en større by, vil det ikke være hensiktsmessig å la en intervjuer besøke utvalgte husstander spredt over et stort område. Isteden tenker man seg at byen deles opp i mindre områder, f.eks. i kvartaler. Det velges så først tilfeldig et visst antall kvartaler. Fra hvert av de utvalgte kvartaler velges så tilfeldig et visst antall husstander som intervjues. Dette kalles et *klyngeutvalg*, her er det tale om et to-trinns klyngeutvalg. På første trinn trekkes et utvalg av såkalte *primære utvalgsenheter*. Fra hver av de utvalgte primærenheter velges et antall *sekundære utvalgsenheter*, og disse danner grunnlag for den statistiske undersøkelsen.

Det viser seg at estimater basert på klyngeutvalg generelt ikke er så pålitelige som tilsvarende estimator basert på et rent tilfeldig utvalg med den samme utvalgsstørrelse. Imidlertid kan et klyngeutvalg vise seg å være mindre tidkrevende og billigere, slik at vi med gitte ressurser kan gjennomføre undersøkelsen med en betydelig større utvalgsstørrelse enn med et rent tilfeldig utvalg. Eksempelvis vil både tid og penger være spart ved å intervjue familier som bor i visse konsentrerte områder enn at alle er spredt utover et større område.

Vi vil ikke her bygge opp noen teori for klyngeutvalg. En slik teori vil avhenge av hvordan de primære utvalgsenhetene velges ut, enten at hver primær utvalgsenhet har samme sannsynlighet for å bli valgt ut, eller at sannsynligheten for å velge en bestemt primær utvalgsenhet er proporsjonal med antall sekundære utvalgsenheter som denne består av. Den siste framgangsmåten sikrer at alle sekundære utvalgsenheter har samme sannsynlighet for å bli valgt ut, eksempelvis at alle husstander i byen har samme sjanse for å komme med i undersøkelsen. Det viser seg at teorien i det siste tilfellet også blir betydelig enklere enn i det første.

Tanken om klyngeutvalg kan også utvides til utvalg som tas i mer enn to trinn, og teorier for tre-trinns utvalg er tilgjengelig i litteraturen. I praksis kan det også være aktuelt å bruke klyngeutvalg i kombinasjon med stratifisering. Kan hende er bebyggelsen i byen ulik fra distrikt til distrikt, noen bydeler med hovedsaklig eldre leiegårder, andre med nyere blokkbebyggelse og småhus, igjen andre med større frittliggende eneboliger. Det kan tenkes at forbruksmønsteret er noe ulikt i de tre typer boligområder, og for å få holdepunkter i denne retning er det ønskelig at hver type er skikkelig representert i utvalget. Vi kan da betrakte de tre typer boligområder som tre strata og innen hvert stratum foretas så et klyngeutvalg etter de retningslinjer som er beskrevet ovenfor.

Grunnen til at man på ulike trinn i en utvalgsundersøkelse anbefaler å velge utvalgsenheter tilfeldig er at det sikrer en viss rettferdighet. Dette kan bidra til å fjerne mulige feilkilder som ikke er åpenbare for de som gjennomfører undersøkelsen. Videre danner en slik framgangsmåte grunnlag for en teori som setter oss i stand til å vurdere påliteligheten av de konklusjoner som trekkes av undersøkelsen. I praksis hender det imidlertid ofte at man er nødt til å gå på akkord med målsettingen om at utvalg skal velges tilfeldig. Som regel skyldes dette omstendigheter av praktisk og/eller økonomisk natur. En bør imidlertid alltid forstå konsekvensene av den utvalgsmåte som brukes, om den kan innebære mulige feilkilder, samt hvilke usikkerhetsmarginer en må regne med i eventuelle konklusjoner. I litteraturen er beskrevet en rekke mulige framgangsmåter for gjennomføring av utvalgsundersøkelser ut fra totalvurdering av ideelle målsettinger og praktisk gjennomførbarhet. Den utvalgsplan som bør velges i en konkret situasjon vil avhenge av omstendighetene.

14.4 Estimeringsmetoder

I de foregående avsnitt har vi diskutert ulike utvalgsplaner. Med en gitt plan vil det, når observasjoner foreligger til analyse, være mulig å velge mellom ulike analysemetoder. Vi har ovenfor studert de enklest tenkelige estimatorer, og vil her se på noen alternative estimatorer som er aktuelle i situasjoner der det foreligger tilleggsinformasjon om hver enkelt utvalgsenhet. Det viser seg at disse estimatorene under visse omstendigheter, kan gi betydelig presisjonsgevinst i forhold til de vanlige estimatorene.

Anta at til hvert av de N elementene i populasjonen er knyttet to verdier, en v-verdi og en w-verdi, slik at populasjonen består av N tallpar

$$(v_1, w_1), (v_2, w_2), \ldots, (v_N, w_N)$$

Anta at summen av w-verdiene er kjent, dvs. $\overline{w}$ er kjent. Vi ønsker fortsatt å estimere gjennomsnittet $\overline{v}$ av v-verdiene på grunnlag av et utvalg på n elementer fra populasjonen. Vi observerer da både v-verdien og w-verdien for hvert element i utvalget, og spørsmålet er om vi ikke også kan utnytte w-verdiene når vi estimerer $\overline{v}$. Dette synes i hvert fall rimelig i situasjoner der vi vet at det er en viss samvariasjon mellom v og w. La oss kort skissere et par eksempler hvor dette er aktuelt:

- En kommune ønsker å anslå den gjennomsnittlige silokapasiteten for alle gårdsbruk i kommunen på grunnlag av et utvalg gårder. Man kjenner det totale dyrkede areal på alle gårdene, både de som er med og ikke er med i utvalget.

- En revisor ønsker å anslå gjennomsnittlig innestående på en rekke konti på et bestemt tidspunkt på grunnlag av et utvalg konti. Hun kjenner innestående på alle konti på samme tidspunkt året før.

Vi vil nedenfor anta at vi trekker et vanlig tilfeldig utvalg. La oss innføre følgende notasjon: Y_i betegner fortsatt v-verdien til det i'te element i utvalget, mens den tilhørende w-verdien betegner vi med X_i. Vårt observasjonsmateriale består derfor av n observasjonspar

$$(Y_1, X_1), (Y_2, X_2), \ldots, (Y_n, X_n)$$

samt gjennomsnittet av w-verdiene i hele populasjonen $\overline{w}$. En aktuell estimator for $\overline{v}$ er nå den såkalte *rateestimatoren*, gitt ved

$$R = \overline{Y} \cdot \frac{\overline{w}}{\overline{X}}$$

Denne estimatoren kan i lys av det første eksemplet ovenfor motiveres slik: Dersom tilfeldigvis de utvalgte gårdene gjennomgående har mindre dyrket areal enn gjennomsnittet $\overline{w}$ vil en kunne vente at også silokapasiteten for disse i gjennomsnitt er lavere enn $\overline{v}$, dvs. at vi underestimerer $\overline{v}$. Tilsvarende vil vi kunne overestimere $\overline{v}$ dersom de utvalgte gårder gjennomgående har dyrket areal over gjennomsnittet. Opplysning om i hvilken retning en feilestimering trolig går, får vi ved å sammenligne det observerte gjennomsnittlige dyrkede areal $\overline{X}$ med det totale gjennomsnitt $\overline{w}$, som er antatt kjent. Den vanlige estimatoren $\overline{Y}$ kan da korrigeres ved å multiplisere med faktoren $\overline{w}/\overline{X}$.

Det kan vises at rateestimatoren er tilnærmet forventningsrett med varians

$$var R \approx \frac{N-n}{N-1} \cdot \frac{1}{n} \cdot \frac{1}{N} \sum_{i=1}^{n} (v_i - \overline{v}\frac{w_i}{\overline{w}})^2$$

Tilnærmelsen er best når n og N er stor. På den annen side husker vi at

$$var\overline{Y} \approx \frac{N-n}{N-1} \cdot \frac{1}{n} \cdot \frac{1}{N} \sum_{i=1}^{n} (v_i - \overline{v})^2$$

Variansen til R vil være betydelig mindre enn variansen til $\overline{Y}$ straks v-verdiene er tilnærmet proporsjonal med de tilhørende w-verdier. Merk at for eksakt proporsjonalitet vil det tilnærmede uttrykk for $\mathrm{var}R$ reduseres til null.

En annen aktuell estimator er den såkalte *regresjonsestimatoren*

$$V = \overline{Y} + \hat{\beta} \cdot (\overline{w} - \overline{X})$$

der koeffisienten $\hat{\beta}$ er bestemt ved minste kvadraters metode anvendt på observasjonene i utvalget, dvs.

$$\hat{\beta} = \frac{\sum_{i=1}^{n}(X_i - \overline{X})Y_i}{\sum_{i=1}^{n}(X_i - \overline{X})^2}$$

For å forstå denne estimatoren trengs kanskje et tilbakeblikk på teorien i Kapittel 8.3. Tankegangen er imidlertid den samme som med rateestimatoren. De observerte w-verdier i utvalget kan avdekke eventuelle skjevheter i utvalget som kan korrigeres. Her korrigeres ut fra forestillingen om en mulig lineær samvariasjon mellom w-verdier og v-verdier. Positiv $\hat{\beta}$ vil indikere positiv samvariasjon. Dersom vi observerer $\overline{X}$ mindre enn $\overline{w}$, vil estimatoren $\overline{Y}$

antakelig underestimere $\overline{v}$. Vi ser at i dette tilfellet vil estimatoren V korrigere oppover, omvendt dersom $\hat{\beta}$ blir negativ. V vil derfor trolig gi et bedre estimat enn $\overline{Y}$. Teori indikerer at regresjonsestimatoren gjennomgående har mindre varians enn rateestimatoren i hvert fall for større utvalg. Rateestimatorer eller regresjonsestimatorer er ofte aktuelle i situasjoner der v-verdiene representerer en størrelse knyttet til elementene i populasjonen på et bestemt tidspunkt, mens w-verdiene representerer samme størrelse for disse elementene på et tidligere tidspunkt og disse antas kjent for alle elementene i populasjonen.

Når det gjelder estimering av varians for rateestimatoren og regresjonsestimatoren må vi vise til spesiallitteratur.

Eksempel 2 : Konti
La oss tenke oss $N = 10$ konti hvor innestående pr. 31.desember i to etterfølgende år var følgende (i tusen kroner):

$$\begin{array}{lcccccccccc} w_i: & 2 & 6 & 1 & 6 & 7 & 4 & 1 & 3 & 3 & 7 \\ v_i: & 3 & 9 & 2 & 5 & 9 & 7 & 1 & 5 & 4 & 7 \end{array}$$

Vi ser at $\overline{w} = 4$, og vi antar at dette er kjent. Den andre tallrekken antar vi er ukjent, og at vi ønsker å estimere $\overline{v}$ på grunnlag av et utvalg på $n = 3$ konti. En tabell over tilfeldige tall lar oss observere konto nr. 1, 3 og 9. Vi får da

$$\overline{Y} = (3 + 2 + 4)/3 = 3.0.$$

Vi ser at $\overline{X} = 2$, og sammenlignet med $\overline{w} = 4$ gir dette berettiget mistanke om at $\overline{v}$ er underestimert. Rateestimatet blir

$$R = 3.0 \cdot \frac{4}{2} = 6.0$$

mens regresjonsestimatet blir

$$V = 3.0 + 1.0 \cdot (4 - 2) = 5.0,$$

og hvilket som bør velges er et skjønnsspørsmål. I realistiske anvendelser vil selvsagt N være mye større, dette må bare oppfattes som et regneeksempel med enkle tall.

Det finnes også en rekke andre forslag til estimatorer som tar omsyn til tilleggsinformasjon. Under de fleste omstendigheter vil disse ikke være dårligere enn den vanlige estimatoren $\overline{Y}$, som ikke tar omsyn til tilleggsinformasjon, og under spesielle omstendigheter vil de være betydelig bedre. Valg av estimator vil derfor være delvis avhengig av de forestillinger vi gjør oss om den tilleggsinformasjon som foreligger. Det finnes versjoner av disse estimatorene også for andre utvalgsplaner enn rent tilfeldig utvalg.

En type utvalg av betydelig interesse er såkalt proporsjonal utvelgelse, der enhetene i utvalget trekkes med sannsynligheter proporsjonale med sine w-verdier, som antas kjent. Tanken bak dette er at store v-verdier har større innflytelse på resultatet enn små, og derfor viktigere å få med i utvalget. Dersom det er korrelasjon mellom v og w, vil vil fremgangsmåten nettopp sikre dette, og på en slik måte at egenskapene til estimeringsmetoder f.eks. rateestimatoren kan utledes. Dette er et alternativ til å bruke v-verdiene til å stratifisere i to eller flere grupper.

Vi har ovenfor omtalt noen sentrale idéer fra teorien for utvalgsundersøkelser. En rekke andre idéer har vi måtte la ligge, og interesserte må søke spesiallitteratur. Dette gjelder spesielt de praktiske sider ved gjennomføringen av utvalgsundersøkelser, intervjupraksis osv.

14.5 Estimering av populasjonsstørrelse

En problemstilling av en noe annen karakter enn ovenfor, er å anslå størrelsen av en populasjon. Dette har en viss aktualitet i forbindelse med økologi og ressursproblemer, f.eks. ved å anslå størrelsen av en laksestamme eller størrelsen av en viltbestand. Slike undersøkelser utføres gjerne ved å merke et visst antall individer fra populasjonen, slippe disse løs igjen, og senere ta et utvalg fra populasjonen, observere andelen av merkede individer, og trekke konklusjoner om populasjonens størrelse på grunnlag av dette.

La N være antall individer i populasjonen og M være antall merkede individer, slik at brøkdelen av merkede individer i populasjonen er $a = M/N$. Det trekkes så et utvalg individer, og anta at brøkdelen merkede individer i utvalget er $\hat{a}$. Det synes da rimelig å anslå populasjonens størrelse N ved

$$\hat{N} = \frac{M}{\hat{a}}$$

Nå kan en slik undersøkelse gjennomføres på ulike måter, vi vil her kort omtale to måter: Enten gjenfanges et bestemt antall individer n fra populasjonen og antall merkede individer Y observeres, eller man fanger individer inntil et bestemt antall merkede individer m er oppnådd og observerer det antall individer X som det ble nødvendig å fange i alt. Det første kalles *direkte utvelging*, det siste *invers utvelging*.

La oss først betrakte direkte utvelging: Brøkdelen av merkede individer i utvalget er da $\hat{a} = Y/n$, slik at estimatoren for populasjonsstørrelsen N kan skrives

$$\hat{N} = \frac{M \cdot n}{Y}$$

For at dette skal ha noen mening må selvsagt $Y > 0$ og vi antar at både M og n er valgt tilstrekkelig store til at dette i praksis er oppfylt. Estimatoren $\hat{N}$ er ikke forventningsrett, den stokastiske Y i nevneren kompliserer det hele. Det kan imidlertid vises at

$$E\hat{N} \approx N + (N - M)/M \approx N$$

Den siste tilnærmelsen er brukbar dersom M og n ikke velges altfor liten, i så fall vil altså $\hat{N}$ være tilnærmet forventningsrett. Videre kan det vises at variansen til $\hat{N}$ er tilnærmet lik

$$var\hat{N} \approx \frac{M^2}{a^4} \cdot \frac{N-n}{N-1} \cdot \frac{a(1-a)}{n} \approx \frac{M^2}{n} \cdot \frac{(1-a)}{a^3}$$

Den siste tilnærmelsen gjelder når n er liten i forhold til N. Dette uttrykket avhenger av a som er ukjent, men et brukbart estimat for a er $\hat{a}$. Variansen til $\hat{N}$ kan derfor estimeres ved å erstatte a med $\hat{a}$ i formelen ovenfor. Vi er dermed i stand til i en viss grad å vurdere usikkerheten ved den estimering av populasjonens størrelse som er foretatt. Teorien ovenfor bygger på en viktig forutsetning, nemlig at de merkede og de gjenfangede individene representerer et tilfeldig utvalg fra populasjonen.

Eksempel 3 : Laks

En lakseoppdretter har laks i en dam og ønsker å anslå det totale antall laks i dammen. Han fanger inn $M = 30$ laks som merkes og slippes ut igjen. Dagen etter fanger han $n = 20$ laks og observerer 4 merkede blant disse. Han estimerer derfor det totale antall laks i dammen til $\hat{N} = 30 \cdot 20/4 = 150$. Et anslag for variansen til estimatoren blir ved innsetting av $\hat{a} = 4/20 = 1/5$ lik 4500. Dersom vi rapporterer estimat standardavvik blir vår rapport for populasjonens størrelse

$$150 \quad \pm \quad 67$$

altså ikke altfor god sikkerhet.

La oss isteden betrakte invers utvelging. Brøkdelen av merkede individer blant de gjenfangede individer er da $\hat{a} = m/X$ slik at estimatoren for populasjonsstørrelsen kan skrives

$$\hat{N} = \frac{M \cdot X}{m}$$

Dersom hvert nytt gjenfanget individ er tilfeldig utvalgt fra populasjonen, viser det seg at denne estimatoren er tilnærmet forventningsrett, dvs.

$$E\hat{N} \approx N$$

Den tilhørende varians er tilnærmet gitt ved

$$var\hat{N} \approx \frac{M^2}{m} \cdot \frac{1-a}{a^2}$$

hvor tilnærmelsen er brukbar dersom m ikke er for stor i forhold til M [2] Siden a i dette uttrykket er ukjent må man i praksis estimere denne variansen ved å erstatte a med estimatet $\hat{a}$.

Eksempel 4 : Laks
Anta at lakseoppdretten i Eksempel 3 isteden har bestemt seg for å velge ut laks inntil $m = 5$ merkede er tatt opp, og det viste seg at han måtte fange 22 laks i alt for å oppnå dette. Han estimerer derfor det totale antall laks i dammen til $\hat{N} = 30 \cdot 22/5 = 132$. Et anslag for variansen til estimatoren blir ved innsetting av $\hat{a} = 5/22$ lik 2692. Dersom vi rapporterer estimert standardavvik blir vår rapport for populasjonens størrelse

$$132 \quad \pm \quad 52$$

For en nærmere redegjørelse av de relative fortrinn av de to metodene samt fastlegging av utvalgsstørrelser må vi vise til spesiallitteratur.

14.6 Oppgaver

1. En forening har $N = 700$ medlemmer i sitt kartotek, hver med sitt tresifrede medlemsnummer. Hvordan vil du gå fram for å trekke et tilfeldig utvalg på n= 20 medlemmer?

[2] Vi kan som tilnærmelse tenke oss at vi slipper løs igjen hvert merket individ som fanges, og ventetiden til hvert nytt merket individ er da geometrisk fordelt, se Kapittel 6.7, hvorav resultatet følger.

2. Man ønsker å anslå gjennomsnittlig studiegjeld $\overline{v}$ for studentene på et kull av $N = 250$ studenter på grunnlag av et tilfeldig utvalg på $n = 30$ studenter. Det viste seg at gjennomsnittet i utvalget (i tusen kroner) var 54, mens den empiriske varians var 400.

   (a) Estimer $\overline{v}$ og rapporter resultatet.

   (b) Ved et annet studiested vet man ut fra en fullstendig undersøkelse at gjennomsnittlig gjeld var 48. Gir resultatet av utvalgsundersøkelsen ovenfor grunnlag for å påstå at gjelden er høyere blant studentene der enn her?

   (c) Kunne vi på noe vis oppnå et bedre estimat for gjennomsnittlig gjeld basert på et utvalg?

3. (a) Påvis at en lotterisituasjon der de N elementene i populasjonen har verdier enten 1 eller 0 er det samme som en hypergeometrisk situasjon.

   (b) La M være 1'ere i populasjonen og påvis at

$$\overline{v} = \frac{M}{N} \quad \sigma^2 = \frac{M}{N}(1 - \frac{M}{N})$$

   (c) La Y være antall 1'ere i et utvalg på n elementer og påvis at $\overline{Y} = Y/n$. Bruk dette til å gi alternativ utledning av forventningen og variansen til estimatoren $\hat{a} = Y/n$ for estimering av $a = M/N$ i en hypergeometrisk situasjon.

   (d) Finn en forventningsrett estimator for variansen til $\hat{a}$.

4. Blant et studentkull på $N = 250$ intervjues et tilfeldig utvalg på $n = 50$ studenter. Det viste seg at 10 av disse hadde deltidsjobb.

   (a) Estimer brøkdelen av studenter på dette kullet som har deltidsjobb og rapporter resultatet.

   (b) Samme undersøkelse ble foretatt for et årskull av eldre studenter og her hadde 20 av de 50 utvalgte deltidsjobb. Estimer og rapporter.

   (c) Gir resultatene grunnlag for å påstå at de eldre studentene er mer tilbøyelig til å jobbe deltid enn de yngre?

5. I en situasjon med k=4 strata er antallet innen hvert stratum og stratumstandardavvikene gitt ved henholdsvis

N_i :	400	300	200	100
σ_i :	10	20	20	30

   (a) Hvor mange observasjoner n må velges ut i alt dersom vi bruker proporsjonal utvelging og ønsker at estimatoren skal ha et standardavvik lik Δ dersom (i) $\Delta = 1$ (ii) $\Delta = 2$ (iii) $\Delta = 5$

(b) Hvor mange observasjoner må velges ut i alt dersom vi isteden bruker optimal utvelging?

6. Ved rapportering av variansen til estimatoren $\tilde{Y}$ som brukes ved stratifisert utvelging antok vi at variansen σ_i^2 innen hvert stratum ble estimert ved

$$S_i^2 = \frac{1}{n_i} \sum_{j=1}^{n_i} (Y_{ij} - \overline{Y}_i)^2$$

(a) Er denne estimatoren forventningsrett for σ_i^2? Hvis ikke, korriger den slik at den blir forventningsrett. Hint: Se Kapittel 8.2.

(b) Gi en formel for den estimerte varians til $\tilde{Y}$ hvor vi har med korreksjonsfaktoren for endelig utvalg og eventuell korreksjon fra (a).

7. En revisor skal anslå det gjennomsnittlige beløp $\overline{v}$ som er innestående på $N = 1000$ konti ut fra et utvalg på av $n = 100$ konti som granskes nærmere. Revisoren ønsker å utnytte informasjon fra året før da han gjennomgikk alle konti. Han deler disse i fire strata etter innestående året før. Han har også beregnet standardavviket σ_i for hvert stratum dette året. Resultatet var:

Stratum nr.:	1	2	3	4
Innestående i kr.:	0-999	1000-4999	5000-9999	10000-
Antall konti (N_i):	400	300	200	100
Standardavvik(σ_i):	250	1000	1500	3000

(a) Det er foreslått å velge ut $n_1 = 10, n_2 = n_3 = n_4 = 30$ fra de fire strata. Hvorfor?
Anta at observerte gjennomsnitt og tilhørende empiriske standardavvik er

Gj.snitt ($\overline{Y}_i$)	800	3500	8000	20000
St.avvik (S_i)	200	900	1600	3500

(b) Estimer $\overline{v}$ og rapporter resultatet.

(b) Bokført verdi er 4.9 millioner kroner. Er dette urimelig?

8. Anta at revisoren i forrige oppgave isteden valgte $n = 100$ konti tilfeldig fra alle $N = 1000$ konti og at han observerte gjennomsnittlig innestående lik 5300 kr. Gjennomsnittlig innestående på de samme utvalgte konti året før var 4800 kr., mens det totale gjennomsnitt året før var 4600 kr. Gi et alternativt estimat for $\overline{v}$ på grunnlag av disse opplysningene.

9. Gitt følgende sammenhørende verdier av (v, w) for en populasjon på $N = 10$ elementer.

v_i :	2	3	4	5	2	3	2	6	1	4
w_i :	1	2	4	3	2	4	1	3	1	2

Det trekkes et tilfeldig utvalg på $n = 3$ elementer.

(a) Beregn variansen til rateestimatoren.

(b) Beregn også variansen til gjennomsnittsestimatoren.

(c) Hvor stort utvalg må trekkes for at denne skal gi samme pålitelighet som rateestimatoren for $n = 3$.

10. (a) Foreslå en estimator for variansen til rateestimatoren.

    (b) $\star$Forsøk å finne ut om den foreslåtte estimator er forventningsrett.

11. En tannlege har gjennomført pensling av tennene til $N = 1000$ elever med sikte på å redusere tannråte. Elevene går til kontroll en gang hvert halvår og tidligere hadde de gjennomsnittlig 2.2 hull pr. gang. Et halvår senere er han spent på om penslingen har gitt resultater, og han noterer seg antall hull hos de n =10 første barna som kommer til kontroll. Resultatet var

$$Y_i: \quad 2 \quad 0 \quad 3 \quad 1 \quad 1 \quad 4 \quad 0 \quad 4 \quad 2 \quad 3$$

(a) Estimer gjennomsnittlig antall hull etter et halvår med penslede tenner og rapporter resultatet.

(b) Er det grunn til å tro at penslingen har gitt resultater? Tannlegen har opplysninger om antall hull ved siste kontroll for elevene ovenfor. Disse var (i samme rekkefølge)

$$w_i: \quad 3 \quad 0 \quad 4 \quad 2 \quad 1 \quad 5 \quad 1 \quad 4 \quad 3 \quad 3$$

(c) Bruk dette til å gi et alternativt estimat.

(d) Kommenter resultatene og også opplegget til undersøkelsen.

12. En skogeier (f.eks. staten) ønsker å telle antall trær i et større område. For dette formål er området delt i et rutenett bestående av $N = 100$ ruter. Av disse velges ut $n = 10$ ruter og trærne der telles. Nå foreligger det luftfoto over vedkommende område og det er mulig ut fra dette å anslå grovt antall trær i hver rute.

    Drøft hvordan denne tilleggsinformasjon kan utnyttes ved estimering av de totale antall trær i området.

13. En skogeier ønsker å anslå det totale volum av skogen i et større område. For dette formål er området delt i et rutenett bestående av N ruter. Av disse velges ut tilfeldig n ruter, fra rute nr. i velges tilfeldig m_i trær og gjennomsnittlig volum $\overline{Y}_i$ av disse noteres (i m^3). Samtidig telles det totale antall trær M_i i denne ruten.

    (a) Foreslå en estimator for det totale volum av skogen uttrykt ved N, m_i, M_i, $\overline{Y}_i$, $i = 1, 2, \ldots, n$.

(b) Estimer skogens totale volum dersom $N = 100$, $n = 8$ $m_1 = m_2 = \cdots = m_8 = 5$ og det ble observert

Rute nr. 1 :	1	2	3	4	5	6	7	8
Antall trær :	30	50	55	35	30	40	65	10
Gj.snitt :	0.8	0.6	1.1	1.0	0.4	0.5	1.0	0.5

(c) Drøft muligheten for å trekke inn flere opplysninger som kan gi et mer pålitelig estimat.

14. $\star$Anta en situasjon med k strata med N_i elementer i stratum nr.i, der variansen er σ_i^2, $i = 1, 2, \ldots, k$.

    (a) Vis at optimal utvelging svarer til n_i proporsjonal med $N_i \cdot \sigma_i$.

    Anta at kostnadene ved hver observasjon fra stratum nr. i er c_i, $i = 1, 2, \ldots, k$. Det er foreslått å velge n_i proporsjonal med $N_i \cdot \sigma_i / \sqrt{c_i}$.

    (b) Vis at dette minimerer variansen for gitte kostnader, og også minimerer kostnadene for gitt varians. Hint: Ulikheten i Oppgave 5.40 kan brukes.

15. Ved en undersøkelse er viltbestanden i et område anslått til 2 000 - 3 000 dyr. Drøft hvordan denne informasjon kan brukes til å planlegge en undersøkelse med merking og gjenfanging for neste år, der en ønsker å anslå bestanden med en sikkerhetsmargin på pluss-minus 200 dyr.

16. Anta at en bestand av villrein skal anslås.

    (a) Drøft metoden med merking og gjenfanging i denne sammenheng.

    (b) Hvilke konsekvenser har det dersom individer dør og fødes i perioden fra merking til gjenfanging.

    (c) Foreslå eventuelt andre metoder til å estimere bestanden.

Kapittel 15

Kvalitetskontroll og forbedring

15.1 Innledning

Statistiske metoder har anvendelsesmuligheter ved kvalitetskontroll av masseproduserte artikler og overvåking av prosessser. Vi kan f.eks. være interessert i å estimere andelen defekte artikler i et større produksjonsparti på grunnlag av et utvalg av artikler (hypergeometrisk situasjon), eller estimere defektsannsynligheten i en løpende produksjonsprosess (binomisk situasjon), eller kvaliteten av artiklene i en løpende produksjon (målesituasjon). Slike problemer kan angripes ved de metodene som er utviklet i Kapittel 7 og 8.

Hensikten med kvalitetskontroll kan være å avgjøre om et gitt produksjonsparti ikke lever opp til en ønsket kvalitet, eller å avgjøre om en produksjonsprosess er ute av kontroll, dvs. sjekke om defektsannsynligheten i produksjonen viser økende tendens, eller sjekke om prosessen gir dårligere målt kvalitet enn før. Slike situasjoner kan ofte betraktes som et beslutningsproblem under usikkerhet med to mulige beslutninger. I tilfellet med produksjonspartiet, akseptere det og sende partiet til kunden eller forkaste partiet. I tilfellet med produksjonsprosessen, la den fortsette eller stoppe den for justering.

Vi finner analoge problemstillinger innen revisjon, når generelle konklusjoner trekkes på grunnlag av stikkprøver. Det kan dreie seg om konklusjoner mht. feilhyppighet og beløpsmessig feil i en gitt populasjon av objekter (f.eks. bilag), eller tilsvarende for en administrativ prosess.

Dette kan formuleres som hypotesetestingsproblemer, men det er ofte hensiktsmessig med terminologi tilpasset anvendelsesområdet. I dette kapitlet er kvalitet i produksjonssituasjoner i fokus. For bruk av tilsvarende metoder i revisjon, se Lillestøl (1996).

I de senere år har interessen innen kvalitetsarbeid gått mer over mot metoder for forbedring av prosesser, dvs. reduksjon av defekt/feilsannsynligheter og økning av forventet målt kvalitet og/eller reduksjon i variasjon i kvalitet. Vi gir en kort omtale av slike problemstillinger i et avsluttende avsnitt.

Dette kapitlet vektlegger koblingen mellom kvalitetskontroll/forbedring og sannsynlighetsteori/statistisk teori. For nærmere beskrivelse av hvordan idéene tillempes i praksis må vi vise til spesiallitteratur.

15.2 Kontroll av produksjonsparti

Vi antar at en artikkel masseproduseres. Hver artikkel kan klassifiseres som enten defekt eller intakt. Man vil selvsagt helst levere bare intakte artikler, men en undersøkelse av alle artiklene før de leveres er ofte ikke hensiktsmessig. Det kan enten være for tidkrevende og/eller dyrt eller det kan skade artiklene, kanskje også ødelegge dem (f.eks. dersom artiklene er fyrverkerisaker). Anta at artiklene produseres i partier på et gitt antall enheter. Av de grunner som er nevnt velger vi å trekke ut et visst antall enheter tilfeldig fra hvert parti og undersøke disse, telle opp antall defekte og akseptere resten av partiet dersom antall defekte i utvalget er lite. La oss innføre følgende notasjon vedrørende kvalitetskontroll av et gitt parti.

$$\begin{aligned}
N &= \text{antall partikler i partiet} \\
n &= \text{antall artikler i utvalget} \\
M &= \text{antall defekte artikler i partiet} \\
Y &= \text{antall defekte artikler i utvalget} \\
k &= \text{aksepttall} \\
\theta &= M/N \text{ brøkdelen av defekte i partiet}
\end{aligned}$$

Vår kontrollplan innebærer altså at partiet aksepteres dersom $Y \leq k$, forkastes dersom $Y > k$, hvor k er en konstant som velges slik at planen får ønskede egenskaper. Vi ønsker selvsagt ikke at et parti med få defekte skal bli forkastet, på den annen side ønsker vi heller ikke at et parti med mange defekte skal bli akseptert. Dette er motstridende krav som spiller inn når k skal velges fornuftig. Hvordan en gitt plan virker i ulike situasjoner kan studeres ved å beregne sannsynligheten for å akseptere partiet som funksjon av θ:

$$\begin{aligned} K(\theta) &= P_\theta(\text{Akseptert parti}) \\ &= P_\theta(Y \leq k) = \sum_{y=0}^{k} P_\theta(Y = y) \end{aligned}$$

hvor P_θ betegner sannsynlighet beregnet under forutsetning av at brøkdelen defekte i hele partiet er θ. Funksjonen $K(\theta)$ vil vi kalle *karakteristikken* til den gitte plan. Den vil her være en avtagende funksjon av θ. Valget av n og k vil være bestemmende for karakteristikken. For gitt n vil, uansett $\theta, K(\theta)$ øke med k, dvs. sjansen for både ønsket og uønsket aksept øker med k, og konsekvensene av dette må vurderes ved valg av k og n.

Siden utvalget på n skjer tilfeldig (uten tilbakelegging), er vi i en hypergeometrisk situasjon, dvs. Y er hypergeometrisk fordelt $(N, N\theta, n)$. Følgelig er

$$P_\theta(Y = y) = \frac{\binom{N\theta}{y} \cdot \binom{N(1-\theta)}{n-y}}{\binom{N}{n}}$$

og slike sannsynligheter kan, for ulike θ , i prinsippet beregnes direkte eller ved bruk av tabeller. Dette er imidlertid ofte brysomt. I mange situasjoner er partistørrelsen N stor og utvalgsstørrelsen n forholdsvis liten. Vi har sett i Kapittel 6.3 at vi da kan bruke binomiske sannsynligheter som tilnærmelse til hypergeometriske sannsynligheter (repeter hvorfor!), dvs.

$$P_\theta(Y = y) \approx \binom{n}{y} \theta^y (1-\theta)^{n-y}$$

Dersom θ er liten og n er stor, men fremdeles liten i forhold til N, vet vi at binomiske sannsynlighet kan tilnærmes med Poissonsannsynligheter (Se Kapittel 6.4), slik at

$$P_\theta(Y = y) \approx \frac{(n\theta)^y}{y!} e^{-n\theta}$$

Det er grunn til å merke seg at så lenge populasjonen N er stor i forhold til utvalget n, vil den eksakte populasjonsstørrelse ikke spille noen rolle, den forsvinner i den binomiske (og Poisson) tilnærmingen. Eksempelvis får vi en situasjon med $n = 10$, tilnærmet samme karakteristikk for $N = 1000$ som for $N = 10000$. I noen situasjoner kan det også komme på tale å bruke normaltilnærmelse til de hypergeometriske (binomiske) sannsynlighetene.

Eksempel 1 : Kontrollplan-karakteristikk
I en situasjon hvor $N = 8$ og vi har valgt $n = 4$ og $k = 1$ blir $\theta = M/8$, og karakteristikken blir da

$$
\begin{aligned}
K(\theta) &= \sum_{y=0}^{1} P_\theta(Y = y) \\
&= \frac{\binom{M}{0}\cdot\binom{8-M}{4}}{\binom{8}{4}} + \frac{\binom{M}{1}\cdot\binom{8-M}{3}}{\binom{8}{4}}
\end{aligned}
$$

Funksjonen er definert for $M = 0, 1, 2, \ldots, 8$, og verdiene er gitt ved (sjekk!)

M	0	1	2	3	4	5	6	7	8
$K(\theta)$	1	1	$\frac{55}{70}$	$\frac{35}{70}$	$\frac{17}{70}$	$\frac{5}{70}$	0	0	0

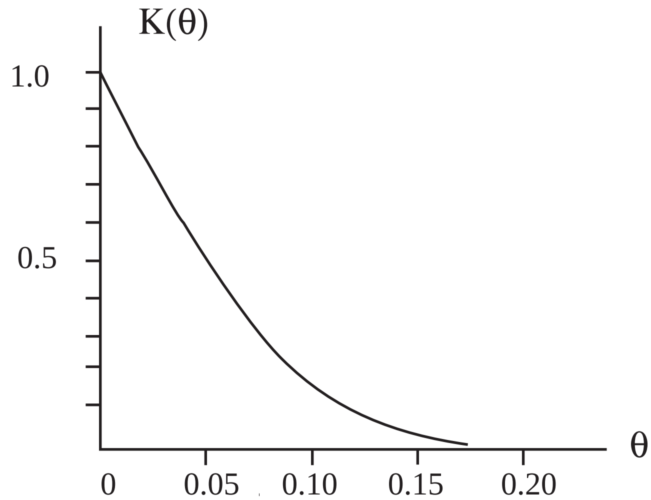
Leseren kan selv sammenligne denne plan med alternative planer, se Oppgave 1.

Eksempel 2 : Kontrollplan-karakteristikk
Et mer realistisk eksempel har vi når $N = 1000$ og $n = 50$. Anta at vi velger en plan med $k = 2$. I dette tilfellet er binomisk tilnærmelse ganske god, og Poisson tilnærmelse brukbar for liten $\theta = M/N$. Vi har følgende karakteristikk

$$
\begin{aligned}
K(\theta) &= \sum_{y=0}^{2} \frac{\binom{N\theta}{y}\cdot\binom{N(1-\theta)}{50-y}}{\binom{1000}{50}} \\
&\approx \sum_{y=0}^{2} \binom{50}{y} \theta^y (1-\theta)^{50-y} \approx \sum_{y=0}^{2} \frac{(50\theta)^y}{y!} e^{-50\theta}
\end{aligned}
$$

Det siste uttrykket er inntegnet i Figur 15.1 som om θ varierer kontinuerlig i intervallet fra 0 til 1. Vi ser at et parti med 5 % defekte har ca. 54 % sjanse for å bli akseptert, mens et parti med 10 % defekte har ca. 12 % sjanse for å bli akseptert. Leseren kan selv sammenligne denne plan med alternative planer.

Det kan være aktuelt å spesifisere andeler θ_1 og θ_2 ($\theta_1 < \theta_2$), slik at aksept er ønskelig hvis $\theta \leq \theta_1$ mens forkasting er ønskelig hvis $\theta \geq \theta_2$. I



Figur 15.1: Karakteristikk

denne forbindelse taler man ofte om

$$\begin{aligned} \text{Produsentens risiko} &= \max_{\theta\leq\theta_1}(1-K(\theta)) = (1-K(\theta_1)) \\ \text{Konsumentens risiko} &= \max_{\theta\geq\theta_2} K(\theta) = K(\theta_2) \end{aligned}$$

men disse begreper er ikke entydige, idet valget av θ_1 og θ_2 vil være delvis vilkårlig. Illustrer produsentens og konsumentens risiko i Figur 15.1!

En kontrollplan er ikke fullstendig spesifisert med mindre man også angir hva som skal foretas med aksepterte og forkastede partier. Her foreligger mange muligheter, og hva som gjøres vil i praksis avhenge av omstendighetene. Et akseptert parti kan f.eks. videresendes med eventuelle defekte som er funnet, eventuelt til redusert pris. De funne defekte kan være fjernet, eller de kan være erstattet med nye artikler som er kontrollert, evt. ikke kontrollert. Et forkastet parti kan vrakes i sin helhet, dette vil f.eks. alltid være tilfellet dersom artiklene ødelegges i kontrollen. Alternativt kan man, dersom det er økonomisk forsvarlig, foreta en fullstendig kontroll av det forkastede parti med sikte på å erstatte alle defekte artikler med intakte.

Et aktuelt begrep i forbindelse med vurdering av egenskapene til en fullstendig spesifisert kontrollplan er *forventet utgående kvalitet* $Q(\theta)$ som gir uttrykk for forventet andel defekte i leverte partier som funksjon av andel defekte i partiet før kontrollen. Begrepet er først og fremst aktuelt å bruke i situasjoner der artikler ikke ødelegges i kontrollen og defekte artikler kan erstattes.

Den enkleste situasjon har vi dersom et akseptert parti videresendes med de defekte som er funnet, mens et forkastet parti blir underlagt fullstendig kontroll der alle defekte erstattes med intakte. Forventet utgående kvalitet er da

$$Q(\theta) = Q_1(\theta) = \theta K(\theta)$$

idet sannsynligheten for å akseptere et parti med en andel θ defekte er $K(\theta)$, mens for et forkastet parti er antall leverte defekte null, og det skjer med sannsynlighet $1 - K(\theta)$.

En mer realistisk situasjon vil være at også de funne defekte i aksepterte partier erstattes med intakte. Forventet utgående kvalitet blir da isteden

$$Q(\theta) = Q_2(\theta) = \sum_{y=0}^{k} \frac{M-y}{N} P_\theta(Y=y)$$

idet, dersom vi finner y defekte og aksepterer resten av partiet ($y \leq k$), er det $M - y$ gjenværende defekte blant de N artiklene, mens dersom partiet forkastes ($y > k$), vil det leverte parti overhodet ikke inneholde defekte.

For planer av den typen vi har betraktet her, vil $Q(\theta)$ være liten når θ er liten, siden partiet da allerede er godt før kontrollen. Den er liten også når θ er stor siden vi da har stor sjanse for å forkaste partiet, foreta en fullstendig kontroll og levere feilfritt parti. Mellom disse to ytterlighetene vil typisk funksjonen $Q(\theta)$ ha et maksimum θ_0. Er dette maksimum lite vil kvaliteten av leverte partier være god uansett hvor mange defekte det opprinnelig var i partiet. Den første av de to planene ovenfor er spesielt enkel å analysere da forventet utgående kvalitet kan beregnes ut fra karakteristikken alene. For store N, slik tilfellet ofte er i praksis, vil de to formlene ovenfor tilnærmet gi samme svar.

Eksempel 3 : Utgående kvalitet
For kontrollplanen i Eksempel 1 der $N = 8$, og $n = 4$ og $k = 1$ blir forventet utgående kvalitet for de to framgangsmåtene beskrevet ovenfor henholdsvis (sjekk tallene!)

M	0	1	2	3	4	5	6	7	8
$Q_1(\theta)$	0	0.125	0.196	0.188	0.121	0.045	0	0	0
$Q_2(\theta)$	0	0.063	0.125	0.134	0.093	0.036	0	0	0

Vi ser at forventet defektprosent til forbruker som ventet alltid er lavere ved den siste framgangsmåten enn den første. Videre ser vi at maksimal forventet defektprosent er henholdsvis 19.6 % og 13.4 %. Dette inntreffer dersom antall defekte M i partiet er henholdsvis 2 og 3 (25 % og 37.5 %). I praksis vil selvsagt M variere fra parti til parti, med de nevnte verdier vil altså være de minst gunstige for forbrukeren.

Eksempel 4 : Utgående kvalitet
For kontrollplanen i Eksempel 2, der $N = 1000$, $n = 50$ og $k = 2$, får vi

θ	0.01	0.02	0.04	0.05	0.06	0.10
$K(\theta)$	0.986	0.920	0.677	0.544	0.423	0.125
$\theta K(\theta)$	0.009	0.018	0.027	0.027	0.025	0.012

Siste linje i denne tabellen er altså forventet utgående kvalitet $Q_1(\theta)$ for den første framgangsmåten. Beregnes $Q_2(\theta)$ vil denne avvike fra $Q_1(\theta)$ med høyst en enhet i siste desimal. Vi ser at partier med lav defektprosent (opp til 2 %) gir ubetydelig lavere defektprosent til forbruker, mens for defektprosent omkring 5 %, er defektprosenten til forbruker halvert, mens for partier med defektprosent på 10 % vil defektprosenten til forbruker igjen være av størrelsesorden 1 %. Vi ser at den maksimale forventede defektprosent til forbruker er ca. 2.7 %, som inntreffer for partier hvor defektprosenten er 4 - 5 %.

Den utvalgsplanen som er beskrevet ovenfor er spesielt enkel. Det finnes andre og mer kompliserte planer som kan være aktuelle i visse situasjoner. Mest aktuelt er kanskje utvalgsplaner i to trinn: Først trekkes et utvalg, og dersom antall defekte er spesielt lite aksepteres partiet. Dersom antall defekte er spesielt stort forkastes partiet, mens dersom antall defekte er moderat trekkes et nytt utvalg hvoretter beslutningen om aksept eller forkastning foretas. Fordelen ved trinnvise planer er at de kan redusere omkostningene til kontroll, f.eks. ved at et dårlig parti avsløres etter kontroll av et lite antall artikler.

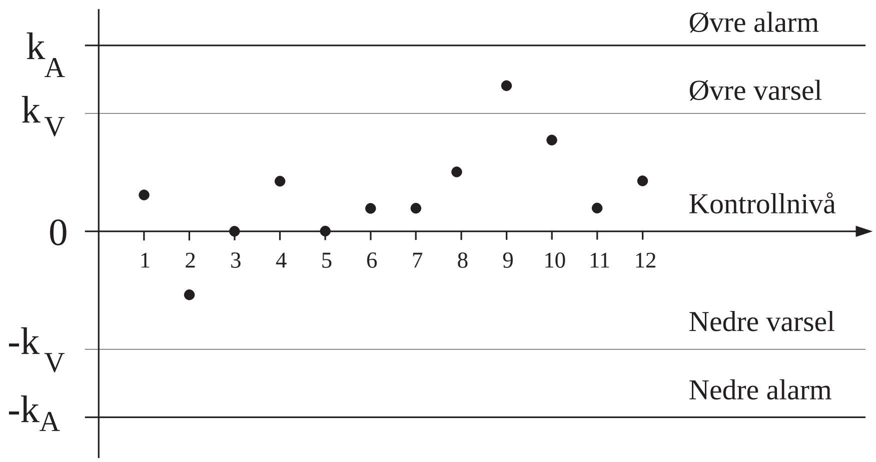
15.3 Prosesskontroll

Anta at vi har en produksjonsprosess som, når den er i kontroll, gir artikler med kvalitet som har en forventning μ_0 og standardavvik σ. Både μ_0 og σ antas kjent, de er anslått på grunnlag av et stort antall artikler i en prøveproduksjon. Når den ordinære produksjon settes i gang ønsker man å holde kvaliteten under oppsikt ved at man ved regelmessige intervaller (f.eks. hver time) kontrollerer et visst antall artikler n. Gjennomsnitts-kvaliteten av disse $\overline{X}$ noteres. Det er hensiktsmessig å plotte de gjennomsnittene som er beregnet ved suksessive tidspunkter i et *kontrolldiagram*.

Dersom forutsetningen i målemodellen er oppfylt, og prosessen fortsatt i kontroll, vil $\overline{X}$ ha forventning μ_0 og standardavvik $\sigma/\sqrt{n}$. Vi kan derfor alternativt plotte den standardiserte størrelse

$$Z = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}}$$

Dette er gjort i Figur 15.2, kalt $\overline{X}$-diagram. Betraktningene gjelder også for tilfellet $n = 1$, og da har vi et kontrolldiagram for individuelle observasjoner, et X-diagram.



Figur 15.2: Kontrolldiagram

Så lenge prosessen er i kontroll, vil Z'ene variere omkring nullnivå, ekstreme observasjoner kan tyde på at noe er skjedd. I figuren er lagt inn to såkalte alarmgrenser, straks et gjennomsnitt faller utenfor en av disse tolkes det som om prosessen en ute av kontroll, og nærmere gransking av årsaken settes i gang, eventuelt stoppes den løpende produksjon.

Vi antar at problemets natur er slik at kvaliteten bør ligge på nivået μ_0, økning eller senkning av dette er lite ønskelig eller like interessant, følgelig har vi to alarmgrenser, en på hver side.

En alarm kan være falsk i den forstand at prosessen fremdeles er i kontroll, men et gjennomsnitt falt tilfeldigvis utenfor. Problemet er derfor å legge alarmgrensene slik at man på den ene side unngår falske alarmer i rimelig grad, men på den annen side har rimelig sjanse til å oppdage relativt raskt en eventuell betydningsfull endring av kvalitetsnivået. Man finner det ofte hensiktsmessig også å legge inn to varselgrenser i figuren, en på hver side av kontrollnivået. Disse brukes gjerne slik at dersom to eller flere etterfølgende punkter faller mellom varselgrensen og den tilhørende alarmgrensen så gir dette indikasjon på at prosessen kan være ute av kontroll. Hovedspørsmålet er nå hvor alarmgrensen og eventuelt varselgrensene bør legges for at de målsettinger som er uttrykt ovenfor blir oppfylt.

Dersom produksjonsprosessen er i kontroll, slik at forutsetningen i målemodellen om uavhengighet for kvaliteten av suksessive artikler er oppfylt, vil fordelingen til Z kunne tilnærmes med normalkurven. Dersom vi legger alarmgrensene ved $\pm k_A$ vil sannsynligheten Π_A for et punkt utenfor disse (falsk alarm) være [1]

$$\begin{aligned}
\Pi_A = P(\mid Z \mid > k_A) &= 1 - P(-k_A \leq Z \leq k_A) \\
&\approx 1 - (G(k_A) - G(-k_A)) \\
&= 2(1 - G(k_A))
\end{aligned}$$

mens sannsynligheten Π_V for et punkt utenfor varselgrensene $\pm k_V$ vil på tilsvarende måte bli $\Pi_V \approx 2(1 - G(k_V))$.

Det er klart at valget av k_A og k_V er opp til brukerens skjønn med hensyn til konsekvensene av falsk alarm og følgende produksjonsstans og konsekvensen av ualarmert endring i kvaliteten, f.eks. ved at allerede produserte artikler må kasseres.

La oss i første omgang studere en plan som ikke gjør bruk av varselgrensene. En måte å vurdere kontrollsystemet på er å se på sannsynligheten for ikke-aksjon som funksjon av kvalitetsnivået μ, denne funksjonen kalles også for *karakteristikken* til kontrollsystemet. Den blir

$$\begin{aligned}
K(\mu) &= P(-k_A \leq Z \leq k_A) \\
&= P(\mu_0 - k_A \cdot \sigma/\sqrt{n} \leq \overline{X} \leq \mu_0 + k_A \cdot \sigma/\sqrt{n}) \\
&= P(\frac{\mu_0 - \mu}{\sigma/\sqrt{n}} - k_A \leq \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \leq \frac{\mu_0 - \mu}{\sigma/\sqrt{n}} + k_A) \\
&\approx G(\frac{\mu_0 - \mu}{\sigma/\sqrt{n}} + k_A) - G(\frac{\mu_0 - \mu}{\sigma/\sqrt{n}} - k_A)
\end{aligned}$$

Merk at vi nå må standardisere med μ istedenfor μ_0 og deretter bruke normaltilnærming. For korthets skyld innføres

[1] Merk ar dersom vi isteden plotter $\overline{X}$ så svarer dette til at alarmgrensene settes ved $\mu_0 \pm k_A \cdot \sigma/\sqrt{n}$, mens varselgrensene settes ved $\mu_0 \pm k_V \cdot \sigma/\sqrt{n}$.

$$\theta = \frac{\mu - \mu_0}{\sigma/\sqrt{n}}$$

dvs. θ er den standardiserte endring i kvaliteten fra kontrollnivået μ_0. Vi har da formelen

$$K(\theta) \approx G(\theta + k_A) - G(\theta - k_A)$$

For et gitt kvalitetsnivå μ kan vi derfor beregne tilhørende θ ut fra kjennskap til μ_0 og σ og n, og vi kan da beregne (eller avlese) sannsynligheten $K(\theta)$ for at systemet ikke alarmerer endring i kvalitetsnivået ved en enkelt kontroll av n artikler. Et studium av $K(\theta)$ kan gi pekepinn på hvordan k_A og (evt.n) bør velges for at kontrollplanen skal ha ønskede egenskaper. En alternativ betrakting som kan gjøres gjeldende er å se på forventet antall kontroller inntil alarm inntreffer $L(\theta)$ som funksjon av θ. Anta at prosessen nå er på kvalitetsnivå μ svarende til en standardisert endring på θ i forhold til kontrollnivået μ_0. Vi vil da ha at

$$L(\theta) = \frac{1}{1 - K(\theta)}$$

Dette er en konsekvens av at suksessive kontroller kan betraktes som en binomisk forsøksrekke der sannsynligheten for suksess (alarm) er $p = 1 - K(\theta)$. Vi vet fra før at forventet ventetid til første suksess er $1/p$. [2]

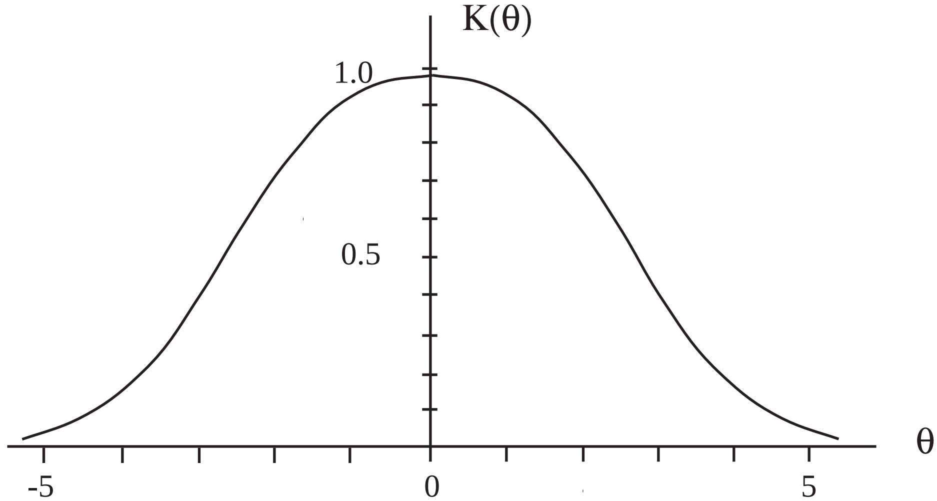
En typisk karakteristikk er inntegnet i Figur .

For den kontrollplanen vi her studerer vil bruk av $K(\theta)$ og $L(\theta)$ være ekvivalent, men i praksis mener mange at bruk av $L(\theta)$ er mer meningsfylt, fordi den sier noe om hvor mye av en gitt kvalitet som blir akseptert i gjennomsnitt før alarm inntreffer. Funksjonen $L(\theta)$ kan se ut som på Figur 15.4, merk at det har vært nødvendig å bruke logaritmisk skala på den vertikale aksen for at figuren ikke skal forsvinne "i det blå".

Vi har tidligere antydet at man også kunne gjøre bruk av varselgrensene. La oss se på følgende kontrollplan:

> Alarm dersom et punkt faller utenfor alarmgrensene, eller to suksessive punkter faller mellom varsel- og alarmgrensen på samme side.

[2] Ventetiden til første alarm er geometrisk fordelt, se Kapittel 6.7.



Figur 15.3: Karakteristikk

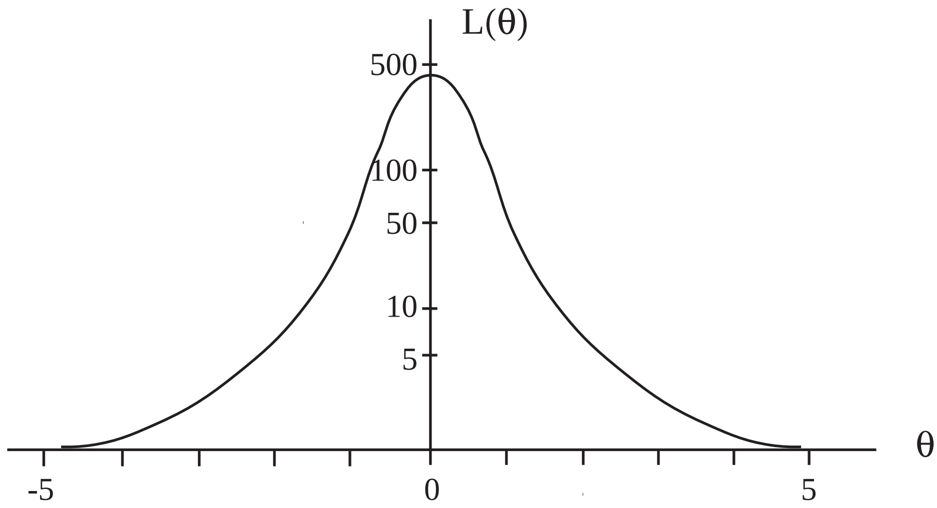
Det kan vises at forventet antall kontroller inntil alarm som funksjon av θ nå blir

$$L(\theta) = \frac{1 + K_1(\theta)}{1 - K_0(\theta) - K_0(\theta) \cdot K_1(\theta)}$$

der $K_0(\theta)$ er sannsynligheten for at et punkt faller mellom varselgrensene, og $K_1(\theta)$ er sannsynligheten for at et punkt faller mellom en varsel- og den tilhørende alarmgrense. Vi ser at $K_0(\theta)$ finnes på samme måte som $K(\theta)$ ved å bruke formelen med k_V istedenfor k_A, mens $K_1(\theta) = K(\theta) - K_0(\theta)$.

Den planen som her er foreslått kan sammenlignes med den enklere planen ovenfor ved å studere $L(\theta)$ funksjonen for de to planene. Det ser ut til at den siste planen, med fornuftig valg av k_A og k_V, vil være å foretrekke framfor den første. For samme forventede ventetid til falsk alarm når prosessen er i kontroll, vil den siste planen gjennomgående være raskere til å oppdage reelle endringer.

I praksis ser en ofte valgt $k_A = 3$ og $k_V = 2$. For en prosess som er i kontroll, vil dette svare til at sannsynligheten Π_A for et punkt utenfor alarmgrensene er ca. 0.0026, mens sannsynligheten Π_V for et punkt utenfor varselgrensene er ca. 0.0456. Noen bruker isteden $k_A = 3.09$ og $k_V = 1.96$ svarende til sannsynlighetene 0.002 og 0.05. Hva antall observasjoner n ved hver kontroll angår, brukes gjerne n i området fra 4 til 10, som oftest $n =$



Figur 15.4: Forventet tid til alarm

4 eller 5, altså et relativt lite antall observasjoner ved hver kontroll. Dette skyldes bl.a. at for å oppdage reelle endringer raskt, vil det erfaringsmessig være bedre å ta et lite antall observasjoner relativt ofte enn et større antall noe sjeldnere, mens det siste på den annen side vil være billigere.

Eksempel 5 : Prosesskontroll
Anta at vi bruker et kontrollsystem med alarm og varselgrenser gitt ved $k_A = \pm 3$ og $k_V = \pm 2$. Anta at $\mu_0 = 10.0$, $\sigma = 1.0$ og $n = 4$. Vi er spesielt interessert i at endringer av forventet kvalitet μ til 11.5 eller mer skal oppdages raskt, dette svarer til

$$\theta = \frac{\mu - \mu_0}{\sigma/\sqrt{n}} = 3$$

Ved bruk av formlene ovenfor og Tabell C.5 og C.6 i Appendiks C, får vi følgende resultater (sjekk!)

θ	0	1	2	3	4
$K(\theta)$	0.9974	0.9772	0.8413	0.5000	0.1587
$K_0(\theta)$	0.9544	0.8400	0.5000	0.1587	0.0228
$K_1(\theta)$	0.0430	0.1372	0.3413	0.3413	0.1359
$L(\theta)$	229	25	4.0	1.7	1.2

Vi ønsker å sammenligne denne plan med en plan uten varselgrenser med omlag samme $L(0)$, dvs. samme forventet ventetid til første falske alarm. Dette betyr at siden $L(0) = 1/(1 - K(0))$, må $K(0)$ for denne plan være

$$K(0) = (L(0) - 1)/L(0) = (229 - 1)/229 = 0.9956$$

Dette svarer til $k_A = 2.85$, dvs. som ventet noe mindre k_A enn for planen hvor alarm også kan inntreffe med to suksessive varsler. Med $k_A = 2.85$ får vi

θ	0	1	2	3	4
$K(\theta)$	0.9956	0.9678	0.8023	0.4404	0.1251
$L(\theta)$	229	31	5.1	1.8	1.1

Vi ser at begge planer gjør jobben med å oppdage $\theta = 3$ relativt raskt, mens planen med varselgrenser ser ut til å oppdage små endringer noe raskere.

Merk at sannsynlighetsbetraktningene ovenfor er knyttet til uavhengighet og normaltilnærming (sentralgrensesetningen). Det kreves ikke absolutt at observasjonene selv er normalfordelte. Unntatt er selvfølgelig når n er liten, herunder spesielt $n = 1$.

De kontrollplaner som er beskrevet til nå tar sikte på å avdekke evt. endring i kvalitetsnivå μ, mulige endringer i variansen σ^2 er holdt utenfor. Den teori som er utviklet forutsetter at variansen fortsatt er den samme selv etter at prosessen er løpt ut av kontroll. Man kan imidlertid studere egenskapene ved disse kontrollplanene dersom vi også tillater at prosessen løper ut av kontroll ved at variansen endres, i praksis vil den som regel alltid øke. Rent intuitivt er det vel klart at sjansen for alarm øker når variansen øker, men i hvilken grad krever teoretisk analyse.

I mange situasjoner vil eventuelle endringer i variansen være like viktig som endringer i nivået, ja endog være av primær interesse. Det er derfor utviklet egne kontrollplaner som tar sikte på å føre kontroll med spredningen av observasjonene. Vi vil komme tilbake til dette nedenfor.

Anta at kontrollproblemet isteden dreier seg om en løpende produksjonsprosess hvor artiklene kun blir klassifisert som intakte eller defekte. For kontrollformål undersøkes regelmessig n artikler av produksjonen og antall defekte noteres. Dette gjentas etter gitte tidsintervaller. Den teori vi har utviklet ovenfor gjelder i grove trekk også for denne situasjonen dersom vi plotter

$$Z = \frac{X - np_0}{\sqrt{np_0(1-p_0)}}$$

der p_0 er den defektsannsynlighet som svarer til at prosessen er i kontroll. Merk at eventuelle reduksjoner i defektsannsynligheten neppe vil ha alvorlige konsekvenser, men det vil også ha interesse å finne ut om noe slikt har skjedd, og man bruker derfor vanligvis alarmgrenser på begge sider av kontrollnivået likevel. Merk at for små n og p_0 vil det ikke være rimelig å bruke normaltilnærming ved beregning av alarmsannsynligheter, disse bør tas direkte ut fra en binomisk tabell. I slike situasjoner vil det også være aktuelt å plotte antall defekte istedenfor den standardiserte størrelsen.

Den metode for prosesskontroll som er beskrevet ovenfor kalles i litteraturen ofte for *Shewharts metode*. Det finnes andre metoder som bygger på helt andre prinsipper. Mest aktuell er kanskje den såkalte *kumulativ-sum metoden:*

La X_i for $i = 1, 2, 3, \ldots$ være observerte kvaliteter på etterfølgende tidspunkter for en produksjonsprosess. Når prosessen er i kontroll oppfattes disse som uavhengige observasjoner med forventning μ_0. Betrakt den kumulative sum

$$Y_t = \sum_{i=1}^{t} (X_i - \mu_0)$$

Dersom kvalitetsnivået for produksjonen er μ og ikke μ_0 ser vi at

$$EY_t = (\mu - \mu_0) \cdot t$$

Dette motiverer at vi kan holde prosessen under oppsikt mht. kvalitetsnivå ved å plotte Y_t som funksjon av t ettersom observasjonene foreligger. Så lenge prosessen er i kontroll, dvs. $\mu = \mu_0$, vil Y_t variere omkring null, men dersom prosessen er ute av kontroll, slik at kvalitetsnivået isteden er $\mu_1 \neq \mu_0$, vil Y_t variere omkring en linje med vinkelkoeffisient $\mu_1 - \mu_0$. Dette indikerer at så snart Y_t-plottet viser en systematisk trend forskjellig fra den horisontale linje så gir det grunn til å tro at noe er skjedd mht. kvalitetsnivået i produksjonsprosessen. Vi vil ikke her gå i detalj med hvordan alarmgrenser konstrueres, men må vise til spesiallitteratur.

En sammenligning av Shewharts metode og kumulativ-sum metoden viser at små endringer i kvalitetsnivået har større sjanse for å bli oppdaget raskere ved kumulativ-sum metoden enn ved Shewharts metode, det vil riktignok kunne ta en viss tid med begge metoder. På den annen side vil en betydelig endring i kvalitetsnivå ha størst sjanse for å bli oppdaget raskt ved Shewharts metode, da det alltid tar noe tid før en stigende eller fallende trend blir åpenbar ved plotting av den kumulative sum.

La oss til slutt kort se på hvordan vi kan føre kontroll med spredningen av kvalitet i produksjonsprosessen: Anta at variansen, når prosessen er i kontroll, er gitt lik σ_0^2. La oss tenke oss at prosessen kan gå ut av kontroll ved at variansen øker, la oss si til en ny verdi $\sigma^2 > \sigma_0^2$, en reduksjon er i praksis lite aktuelt. Igjen tenker vi oss at etter visse tidsintervaller observerer n artikler av produksjonen som kvalitetsvurderes. Som mål for spredningen av kvaliteten kan vi beregne

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

Vi vet at $ES^2 = \sigma^2$, slik at dersom en observert S^2 er stor i forhold til σ_0^2, gir dette grunn til å tro at σ^2 er større enn σ_0^2. Som kontrollstørrelse velger vi å bruke

$$W = S^2/\sigma_0^2$$

som kan plottes i et kontrollskjema etter hvert som nye stikkprøver av produksjonen foretas. Vi velger å aksjonere straks W har verdi større enn k. Sannsynligheten for ikke-aksjon, også her kalt karakteristikken, vil vi betrakte som funksjon av $\theta = \sigma/\sigma_0$ dvs. forholdet mellom det virkelige standardavvik og det "normale". Vi har altså

$$K(\theta) = P_\theta(W \leq k)$$

For å kunne gi en spesifikk beregningsformel trenger vi derfor å kjenne sannsynlighetsfordelingen til W. Dersom vi er villige til å anta at X-observasjonene oppfyller kravene i målemodellen og er normalfordelte, kan det vises at fordelingen til

$$Q = \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

er eksakt lik kjikvadratkurven med $n-1$ frihetsgrader (se Kapittel 7.7). Vi merker oss at $W \leq k$ er ekvivalent med at $Q \leq (n-1)\cdot k/\theta^2$ slik at karakteristikken blir

$$\begin{aligned} K(\theta) &= P_\theta(Q \leq (n-1)\cdot k/\theta^2) \\ &= F_{n-1}((n-1)\cdot k/\theta^2) \end{aligned}$$

der $F_{n-1}(y)$ betegner arealet til venstre for y under kjikvadratkurven med $n-1$ frihetsgrader. Vi ser at karakteristikken, for gitt n og k, som ventet er avtagende i θ. Selv om formelen ovenfor er utledet under normalitetsforutsetninger, vil de tall som regnes ut også generelt gi en viss innsikt i egenskapene ved ulike planer.

Den metode for spredningskontroll som er beskrevet her virker uansett om prosessen løper på det opprinnelige kvalitetsnivå eller har tilpasset seg et nytt nivå. I praksis vil en derfor som regel trenge å føre kontroll med både nivåendringer og spredningsendringer.

Eksempel 6 : Spredningskontroll
Anta at $n = 4$. La oss prøve to ulike valg av alarmgrense, nemlig $k = 2$ og 3. Ved hjelp av Tabell C.7 har vi funnet følgende:

	θ	1.0	1.2	1.5	2.0	2.5	3.0	4.0
k=2	$K(\theta)$	0.89	0.75	0.56	0.32	0.19	0.12	0.04
k=3	$K(\theta)$	0.97	0.90	0.73	0.48	0.30	0.20	0.10

Tabellen viser at den første planen har 11% sjanse for falsk alarm ved en kontroll, den andre bare 3%. På den annen side vil sjansen for å oppdage en dobling av standardavviket ($\theta = 2$) være henholdsvis 68% og 52% for de to planene.

Som i tilfellet med kontroll av kvalitetsnivå kan vi studere planer med både alarm og varselgrenser, eventuelt både øvre og nedre slike. Vi kan også betrakte forventet ventetid til alarm etter de samme retningslinjer som ovenfor. Litteraturen beskriver også planer for spredningskontroll som istedenfor empiriske varianser, tar utgangspunkt i variasjonsbredden dvs. differensen mellom største og minste verdi av de n X-observasjonene.

I praksis finnes mange typer kontrolldiagram og ulike tillempninger av slike. Vi har her begrenset oss til å belyse koblingen til sannsynlighetsregning og statistisk teori.

Merknad. Dersom man har grunn til å anta at en prosess er i kontroll, er det verdiløst å foreta inspeksjon av et ferdig parti som i avsnitt 15.2, med sikte på aksept eller forkastning av partiet, se Oppgave 18. Dette er et sterkt argument for at kvalitetsarbeid må rettes mot selve produksjonsprosessen.

15.4 Kvalitetsforbedring

Det beste utgangspunkt for forbedring av en prosess vil være når prosessen er i kontroll. Et kontrolldiagram som beskrevet i forrige avsnitt kan da tjene som et "lytte og diagnoseverktøy, der påfallende mønstre i observasjonene muligens kan tilskrives en spesiell årsak.

Det første trinn i kvalitetsforbedring av en prosess, vil være å identifisere slike årsaker, og deretter fjerne dem, slik at vi sitter igjen med en prosess i kontroll. En slik prosess varierer tilfeldig rundt et nivå med konstant avvikstendens. Denne variasjon har sjelden noen enkelt årsak, men er typisk forårsaket av mange små årsaker som lite kan gjøres noe med, uten å endre prosessen på en mer fundamental måte. Det vil faktisk virke mot sin hensikt å reagere på enkeltavvik fra nivået som er innenfor grensene for tilfeldige avvik.

Merknad. Grensene som prosessen i kontroll naturlig varierer mellom, kalles vanligvis *kontrollgrenser*. Slike grenser må begrepsmessig ikke forveksles med aksjonsgrenser som presentert i forrige avsnitt, ei heller med spesifikasjonsgrenser, slike som ofte uttrykkes i bransjestandarder.

Vanligvis vil nivå $\pm\ 3\times$ standardavvik kunne fungere som kontrollgrenser. Med en prosess i kontroll har vi to muligheter mht. å bedre kvaliteten: endre kvalitetsnivået og redusere variasjonen. Det vil som regel være en rekke faktorer som er bestemmende for kvalitet, så som produktdesign, maskindesign, produksjonsrutiner og råmaterialer.

Statistisk eksperimentplanlegging og analyse kan bidra til å finne bedre kombinasjoner av faktorer enn de som er i bruk til nå. Det er ofte mulig å gruppere faktorer i henhold til om de i hovedsak kan påvirke nivå eller variasjon, og det kan være et spørsmål hvilke en bør gripe fatt i først, eller om de kan studeres samtidig. I denne forbindelse kan bruk av målfunksjoner som kombinerer de to karakteristika ved prosessen være aktuelle. Mange foretrekker å angripe variabilitet først, og når denne er redusert studeres mulighet for forbedring i nivå. La oss fokusere på forbedring i nivå. Det er ofte mange faktorer som kan påvirke kvalitetsnivået. Dersom hver faktor ønskes studert på flere enn to nivåer, blir det i alt mange mulige faktorkombinasjoner, og med flere observasjoner pr. faktorkombinasjon, vil omfanget av eksperimentet bryte grenser for det som er praktisk og økonomisk forsvarlig å gjennomføre. En må da forsøke å inngå kompromisser mht. eksperimentplan, f.eks. ved å

- begrense seg til faktorer som anses mest vesentlige,
- redusere antall nivåer for hver faktor, f.eks. til 2,
- la være å observere enkelte faktorkombinasjoner,
- begrense seg til (høyst) en observasjon pr. faktorkombinasjon.

Ved alle disse forslag risikerer en å tape noe, og valg av eksperimentplan må skje med åpne øyne, slik at vesentlig informasjon har minst mulig sjanse for å unnslippe.

Fjerning av faktorer uten nærmere undersøkelse kan medføre at uventede kvalitetspåvirkende effekter forblir uoppdaget. Ved reduksjon av antall nivåer til to, vil ikke-lineære effekter ikke kunne oppdages (i første omgang). Ved å la være å observere enkelte faktorkombinasjoner risikerer en at enkelte effekter ikke lar seg identifisere, med mindre en velger spesielle eksperimentplaner, og gjør tilleggsantakelser om effektene. Ved å observere hver faktorkombinasjon bare en gang, kan en vanskelig anslå usikkerhet, med mindre en gjør bestemte antakelser om fravær av mer kompliserte effekter (samspill).

Det har i praksis vist seg at en kan lære mye, selv når en begrenser seg til to nivåer (f.eks. Lav/Høy) for hver faktor. Vi skal nedenfor gi en smakebit på fremgangsmåten i slike situasjoner.

Vi tenker oss at hver faktor kan påvirke responsen direkte ved såkalte hovedeffekter som adderes opp. I tillegg kan det tenkes at bestemte kombinasjoner av faktorer påvirker responsen ut over dette. Vi kaller dette *samspillseffekter*, se Kapittel 11.4 for mer formell diskusjon. Her brukes en noe enklere notasjon, som i tilfellet med to nivåer for hver faktor, lettere lar seg generalisere til mange faktorer.

Eksempel 7 : Et 2^2-eksperiment

Ved overflatebehandlingen av et produkt brukes et kjemikalium som vurderes tilsatt i to ulike styrkegrader (Faktor A: Lav/Høy), ved to ulike temperaturer (Faktor B: Lav/Høy). Kvaliteten av finishen måles (høyere tall jo bedre). Det er gjort 2 observasjoner for hver av de $2^2 = 4$ faktorkombinasjonene, og resultatet ble

Faktor-kombinasjon	Faktor A	B	AB	Observasjoner nr. 1	nr.2	Gjennomsnitt/ Standardavvik	
1	−	−	+	45	47	$\overline{X}_1 = 46$	$S_1 = 1.4$
2	+	−	−	40	40	$\overline{X}_2 = 40$	$S_2 = 0.0$
3	−	+	−	53	49	$\overline{X}_3 = 51$	$S_3 = 2.8$
4	+	+	+	50	53	$\overline{X}_4 = 52$	$S_4 = 1.4$

Her er faktorkombinasjonene nummerert fortløpende og er beskrevet for hver faktor med koder $-/+$ for Lav/Høy. Ved å "multipliserekodene for A og B får vi en nyttig kode for samspillet AB. Denne er $+$ når begge faktorer er på samme nivå, og $-$ når faktorene er på motsatt nivå. Observasjonene og deres gjennomsnitt og standardavvik er ført ut til høyre for hver faktorkombinasjon.

Estimatorer for ulike effekter blir:

$$\begin{array}{rll} \text{Hovedeffekt}A: & \frac{1}{2}((\overline{X}_2 - \overline{X}_1) + (\overline{X}_4 - \overline{X}_3)) & = -2.5 \\ \text{Hovedeffekt}B: & \frac{1}{2}((\overline{X}_3 - \overline{X}_1) + (\overline{X}_4 - \overline{X}_2)) & = 8.5 \\ \text{Samspill}\ \ AB: & \frac{1}{2}((\overline{X}_4 - \overline{X}_3) - (\overline{X}_2 - \overline{X}_1)) & = 3.5 \end{array}$$

Disse svarer til $+$ kategorien for hver faktor, og estimater for $-$ kategorien fås ved å skifte fortegn (summen er null). Legg merke til at alle beregninger har form

$$\frac{1}{2}(\pm\overline{X}_1 \pm \overline{X}_2 \pm \overline{X}_3 \pm \overline{X}_4)$$

der fortegnene er gitt ved de respektive $-/+$ koder for hver effekt.

Begrunnelsen for formlene ovenfor kan skje ut fra formlene i Kapittel 11.4 eller direkte. Estimatene for hovedeffekten A får vi ved å beregne gjennomsnittet av endringene fra Lav til Høy for A for de to mulige nivåer for B, og ta gjennomsnittet av disse. Tilsvarende for hovedeffekten B. Samspillet AB anslås ved å se på forskjellen i anslått endring fra Lav til Høy, for A, for henholdsvis Høy og Lav for B.

Anslagene ovenfor bør vurderes i lys av sine standardavvik. Som regel beregnes dette under antakelsen at uavhengige enkeltobservasjoner med samme standardavvik σ uansett faktorkombinasjon. Da er variansen til estimatorene ovenfor σ^2/m, der m er antall gjentatte observasjoner for hver faktorkombinasjon (i eksemplet er $m = 2$). Anslag for σ^2 blir da

$$S^2 = \frac{1}{4}(S_1^2 + S_2^2 + S_3^2 + S_4^2)$$

der S_i er beregnet standardavvik for observasjonene ved faktorkombinasjon nr. i. Estimator for standardavviket til effektestimatorene blir derfor $S/\sqrt{m}$, og vi kan rapportere

$$\text{Effektestimat} \pm \frac{S}{\sqrt{m}}$$

med vanlig grov fortolkning av feilmarginer. Ved antakelse om normalfordelte observasjoner kan en lage eksakte konfidensintervaller der en putter inn en sikkerhetsfaktor etter oppslag i t-fordelingen med $4(m-1)$ frihetsgrader.

I eksemplet ovenfor beregner vi $S = 1.72$, som gir estimert standardavvik lik 1.21. Av resultatene ser vi en klar temperatureffekt i favør av høy temperatur, og en ikke signifikant styrkeeffekt i favør av lav kjemikaliestyrke. Imidlertid ser vi en mulig samspilleffekt

$$3.5 \pm 1.21.$$

Et 95% konfidensintervall vil med 4 frihetsgrader kreve sikkerhetsfaktor k= 2.776, som gir konfidensgrenser 3.5 ± 3.36. Det kan altså se ut som at høy kjemikaliestyrke er gunstig i kombinasjon med høyere temperatur. En går da selvsagt inn for denne kombinasjonen, selv om hovedeffekten av A favoriserer lav kjemikaliestyrke. I praksis vil en helst ikke rapportere vha. hovedeffekter og samspill i situasjoner der samspill er mer enn en tredjedel av hovedeffekter.

Eksempel 8 : Et 2^3-eksperiment

Anta at vi i forrige eksempel kan tenkes å velge en av to dyseåpninger (Faktor C: Lav/Høy) i tillegg til faktorene A og B. Vi har da følgende skjema med resultater med to observasjoner pr. faktorkombinasjon.

Faktor-	Faktorer			Samspill				Gj.snitt	St.avvik
kombinasjon	A	B	C	AB	AC	BC	ABC	$\overline{X}_i$	S_i
1	−	−	−	+	+	+	−	49	1.0
2⋆	+	−	−	−	−	+	+	43	1.7
3⋆	−	+	−	−	+	−	+	52	1.7
4	+	+	−	+	−	−	−	54	2.4
5⋆	−	−	+	+	−	−	+	46	1.4
6	+	−	+	−	+	−	−	40	0.0
7	−	+	+	−	−	+	−	51	2.8
8⋆	+	+	+	+	+	+	+	52	1.4

Her vil alle estimatorer for hovedeffekter og samspill ha form

$$\frac{1}{2^{3-1}}(\pm\overline{X}_1 \pm \overline{X}_2 \pm \ldots \pm \overline{X}_8)$$

der $-/+$ velges i henhold til skjemaet ovenfor. Standardavvikene til estimatoren kan estimeres med $S/\sqrt{2m}$ (begrunnes som ovenfor), der m er antall observasjoner pr. faktorkombinasjon og

$$S^2 = \frac{1}{8}(S_1^2 + S_2^2 + \ldots + S_8^2)$$

Dersom vi antar normalfordelte observasjoner, kan sikkerhetsfaktorer beregnes ut fra t-fordelingen med $8(m-1)$ frihetsgrader. Hovedeffekter og samspill blir anslått til:

A	B	C	AB	AC	BC	ABC
-2.25	7.75	-2.25	3.75	-0.25	-0.75	-0.25

Her blir S=1.77, som gir estimert standardavvik lik 0.88. Det ser ut til at C ikke er involvert i samspill, men at C har signifikant hovedeffekt i favør av lav dyseåpning.

Man har ofte ikke anledning til flere enn en observasjon pr. faktorkombinasjon. Det er da ikke mulig å anslå S som ovenfor, men det fins måter for mer uformell betraktning av usikkerheten.

Med bare to nivåer for hver faktor vil en, selv med et moderat antall faktorer, få mange kombinasjoner. Eksempelvis gir 8 faktorer $2^8 = 256$ kombinasjoner. I praksis er det ofte bare et fåtall faktorer som påvirker responsen ofte 2 - 3 (Paretoprinsippet). Det fins eksperimentplaner, såkalte *fraksjonerte planer*, der en ikke observerer alle faktorkombinasjoner, men er likevel i stand til å avdekke alle samspill mellom to faktorer, men ikke nødvendigvis mer kompliserte samspill, som likevel sjelden forekommer i praksis. I eksemplet ovenfor er det f.eks. nok å observere faktorkombinasjonene merket med $\star$, dersom vi antar at samspill mellom alle tre faktorene ikke forekommer.

Det kan være vanskelig å få anledning til å gjennomføre eksperimenter i stor skala med prosesser som er i drift. Det er derfor utviklet eksperimentplaner og statistiske metoder for situasjoner der en trinnvis foretar små endringer som ikke medfører driftsstans, som på noe sikt kan lede til optimal faktorkombinasjon, såkalt "Evolutionary Operations".

15.5 Oppgaver

1. Studer karakteristikken til følgende alternative planer for situasjonen i Eksempel 1: (a) $n = 4$ og $k = 0.2$ (b) $n = 3$ og $k = 0.1$
   Kommenter resultatene.

2. Finn produsentens og konsumentens risiko for planene i Oppgave 1 dersom $\theta_1 = 1/8$ og $\theta_2 = 4/8$.

3. Finn forventet utgående kvalitet for hver av planene i Oppgave 1 dersom

   (a) aksepterte partier sendes som de er, mens et forkastet parti erstattes med et intakt.

   (b) alle funne defekte artikler i akseptert parti erstattes med intakte, mens et forkastet parti erstattes med et intakt.

   (c) aksepterte partier sendes som de er, mens forkastede partier ikke videresendes.

   Finn også maksimum forventet andel defekte ved hver plan.

4. Betrakt en kontrollplan der forkastning medfører at alle resterende artikler i partiet også kontrolleres. Finn forventet antall artikler som må kontrolleres i alt som funksjon av M dersom $N = 8$ og (jfr. Oppgave 1),
   (a) $n = 4$ og $k = 0, 1, 2$ (b) $n = 3$ og $k = 0, 1$

5. (a) Bestem karakteristikkene og tegn disse i en figur for følgende planer

$$\begin{array}{rllll} \text{(i)} & N = 2000 & n = 100 & \text{og} & k = 4 \\ \text{(ii)} & N = 2000 & n = 50 & \text{og} & k = 2 \\ \text{(iii)} & N = 2000 & n = 25 & \text{og} & k = 1 \end{array}$$

   (b)Bestem produsentens og konsumentens risiko for θ_1= 0.05 θ_2= 0.10
   (c) Bestem forventet utgående kvalitet for planene i (a) dersom aksepterte partier sendes som de er og hvert forkastet parti erstattes med et intakt.

6. Bestem en plan der produsentens risiko α og konsumentens risiko β er gitt ved $\alpha = \beta = 0.05$, beregnet for henholdsvis $\theta_1 = 0.01$ og $\theta_2 = 0.05$, i tilfellene
   (i) $N = 10\ 000$ (ii) $N = 1\ 000$ (iii) $N = 100$

7. En revisor har et større antall bilag av en viss type, og det er uaktuelt å gjennomgå alle disse med mindre en stikkprøve skulle avsløre betydelige uregelmessigheter.
   (a) Formuler situasjonen som et kvalitetskontrollproblem.
   (b) Finn tilnærmet karakteristikken for følgende planer
   (i) n=40, k=0, (ii) $n = 80$, $k = 1$, (iii) $n = 100$, $k = 1$.
   Er det nødvendig å kjenne N eksakt for å besvare (b)?

8. En produsent leverer partier av en vare bestående av $N = 25$ enheter. Disse kan variere noe i kvalitet og kan sorteres i to kvaliteter. Det er ikke ønskelig at et levert parti inneholder mer enn to enheter av annen sortering. Mottakeren vurderer to ulike stikkprøvemetoder:

   (a) Det trekkes 8 artikler og partiet aksepteres dersom antall defekte er høyst en.

   (b) Trekk først 5 artikler. Dersom utvalget er uten defekte, aksepteres partiet, mens dersom det er mer enn en defekt forkastes partiet. Ved akkurat en defekt trekkes nye 5 artikler og partiet aksepteres dersom dette utvalget er uten defekte, i motsatt fall forkastes det.

   Sammenlign karakteristikken til de to metodene når enheter av annen sortering spiller rollen som defekte.

9. Sammenlign begrepene i teorien for kontroll av produserte artikler med begrepene i hypotesetestingsteorien. Drøft eventuelle sammenhenger mellom
   (a) Karakteristikk og styrkefunksjon.
   (b) Konsumentens risiko og signifikansnivå.
   (c) Produsentens risiko og styrke.

10. En bestemt type snøre vil ved vanlige produksjonsforhold gjennomsnittlig tåle en belastning i kilo på 10.0, og man har erfaring for at standardavviket kan settes til 0.4. Hver time velges ut 4 snører som belastes. Gjennomsnittlig belastning som tåles blir beregnet og danner utgangspunkt for føring av kontrolldiagram. Anta først at kontrolldiagrammet har alarmgrensene $k_A = \pm 3$.

    (a) Finn forventet ventetid til falsk alarm.

    (b) Finn forventet ventetid til alarm dersom kvalitetsnivået er endres til 9.5 kg og til 9.0 kg.

    Anta at vi i tillegg bruker varselgrensen $k_V = \pm 2$. Besvar (a) og (b) dersom to etterfølgende varsler gir alarm.

11. Følgende observasjoner på 8 etterfølgende tidspunkter foreligger for situasjonen i forrige oppgave.

Tidspunkt	1	2	3	4	5	6	7	8
Obs. nr.1	9.5	10.8	10.2	10.4	10.3	9.7	9.3	9.8
2	10.4	10.5	9.1	9.6	9.4	9.1	9.0	9.2
3	10.7	9.6	9.5	10.0	9.7	9.4	9.6	10.1
4	9.8	10.3	10.4	10.6	9.1	10.1	9.9	9.5

    (a) Beregn standardiserte gjennomsnitt og plott disse i et kontrolldiagram med alarmgrenser $k_A = \pm 3$ og varselgrenser $k_V = \pm 2$.

    (b) Gir tallmaterialet grunn til å aksjonere, i så fall når?

12. En produksjonsprosess gir under vanlige omstendigheter 5% defekte. Fra den løpende produksjon velges hver halvtime $n = 100$ artikler som testes og antall defekte noteres. En dag var resultatet som følger:

Tidspunkt	1	2	3	4	5	6	7	8	9	10	11	12
Defekte	3	7	5	1	7	4	9	7	9	6	10	9

    (a) Plott resultatene i et kontrolldiagram med alarmgrenser $k_A = \pm 3$ og varselgrenser $k_V = \pm 2$.
    (b) Gir tallmaterialet grunn til å aksjonere, i så fall når?
    (c) Hva er forventet ventetid til alarm fra et tidspunkt da prosessen begynner å gi gjennomgående 10% defekte.

13. I en situasjon med kontroll av defektprosenten i en produksjonsprosess som under normale omstendigheter gir 10% defekte velges $n = 10$ artikler ut for testing hver halvtime. Man ønsker å plotte antall defekte artikler direkte i et kontrolldiagram med bare en øvre alarmlinje. Bestem denne alarmlinjen i følgende situasjoner
(a) Forventet ventetid til falsk alarm skal være ca. 75.
(b) Forventet ventetid til alarm ved dobling av defektprosenten blir ca. 3.
Hvordan kan man eventuelt få oppfylt begge ønskemål (a) og (b)?

14. Anta at vi bruker planen i Oppgave 13 (a) og at vi har observert

Tidspunkt	1	2	3	4	5	6	7	8	9	10	11	12
Defekte	1	2	0	2	1	0	2	2	1	3	1	2

    Gir tallmaterialet grunn til å aksjonere, i så fall når?

15. En bedrift vil føre kontroll med spredningen av kvalitet etter metoden beskrevet i teksten. Anta at $n = 5$.
    (a) Tegn karakteristikken $K(\theta)$ for tilfellet $k = 3$.
    (b) Hvilken verdi må k ha for at sannsynligheten for falsk alarm ved en kontroll skal være ca. 0.01?
    (c) Hva er sannsynligheten for alarm dersom standardavviket er doblet?

16. Bestem en plan for spredningskontroll av den typen som er beskrevet i teksten der forventet ventetid til falsk alarm er 200, mens forventet ventetid til alarm ved en dobling av standardavviket er 5. Finn forventet antall artikler som passerer før alarmen går dersom standardavviket er doblet.

17. Man ønsker å føre kontroll med både nivå og spredning av kvaliteten i den løpende produksjon, og regelmessig velges ut 9 artikler som kvalitetsmåles. Under normale forhold har kvaliteten til en tilfeldig valgt artikkel forventning 40.0 og standardavvik 3.0. Anta at observerte gjennomsnitt $\overline{X}$ og estimerte standardavvik S plottes direkte i hvert sitt kontrolldiagram. For begge

kontrolldiagrammer ønsker en at forventet ventetid til falsk alarm av typen forverring skal være ca. 100. Fastlegg kontrollplaner i henhold til dette. Følgende observasjoner foreligger:

$\overline{X}$:	41.7	38.5	39.2	40.6	40.9	39.3	38.7	40.4	42.1
	40.1	39.2	37.7	40.5	38.4	38.6	37.9	38.5	
S:	3.7	2.8	2.2	3.4	3.3	3.1	2.6	2.9	3.4
	3.9	3.2	4.0	4.2	3.9	4.8	4.2	4.6	

Blir det slått alarm i løpet av disse 17 tidspunktene, i så fall når, og hva slags alarm?

18. Anta at vi har en prosess som produserer artikler med konstant defektsannsynlighet p, der vi har uavhengighet over tid (prosessen er i kontroll). N etterfølgende artikler fra denne prosessen utgjør et parti, hvor det er et ukjent antall defekte M. Fra partiet trekkes n artikler, og antall defekte Y blant disse noteres.
Vis at antall gjenværende defekte i partiet $X = M - Y$ er uavhengig av Y.
Drøft konsekvensene av dette for partiinspeksjon i praksis.

19. Gjennomfør beregningene av de størrelser som er oppgitt i Eksempel 8, og lag deretter konfidensintervaller for hovedeffekter og samspill.

20. En fruktdrikk leveres i pappemballasje, og en vurderer tre ulike faktorer ved produksjonen. Et eksperiment ble utført der smak ble vurdert på en skala fra 1 til 9 av en dommer. Resultatet ble

Vokslag (A)	Luftrom (B)	Tilsetning (C)	Observasjoner	
Tynt	Intet	Lav	6	8
Tynt	Intet	Høy	8	7
Tynt	Lite	Lav	9	9
Tynt	Lite	Høy	1	2
Tykt	Intet	Lav	6	7
Tykt	Intet	Høy	6	8
Tykt	Lite	Lav	9	8
Tykt	Lite	Høy	2	3

Anslå hovedeffekter og samspill, og vurder deres signifikans, med sikte på finne den beste eller gode kombinasjoner av faktorene.

Kapittel 16

Beslutning under usikkerhet

16.1 Innledning

En beslutningstaker skal treffe en beslutning under usikkerhet. Skjematisk tenker vi oss at problemet kan beskrives slik

		Beslutninger			
Utfall	Sannsynligheter	a_1	a_2	$\cdots$	a_n
u_1	p_1	g_{11}	g_{12}	$\cdots$	g_{1n}
u_2	p_2	g_{21}	g_{22}	$\cdots$	g_{2n}
$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$
u_m	p_m	g_{m1}	g_{m2}	$\cdots$	g_{mn}

Tabell 16.1: Beslutningstabell

Her foreligger n mulige beslutninger (aksjoner) $a_1, a_2, \ldots, a_n$ og beslutningstakeren må velge en av disse. Konsekvensene av den valgte beslutning avhenger av hvilket av de m utfallene $u_1, u_2, \ldots, u_m$ som inntreffer. I tabellen betegner p_i sannsynligheten for utfallet u_i, mens g_{ij} gir uttrykk for konsekvensen dersom utfallet u_i inntreffer og beslutningstakeren har valgt beslutning a_j, i mange forbindelser er det hensiktsmessig å kalle g_{ij}'ene for gevinster.

Eksempler på beslutningsproblemer som skjematisk kan beskrives slik er mange, la oss nevne noen få:

- Hasardspill, f.eks. med en eller flere terninger. Utfallene er de ulike kombinasjoner av øyne, og beslutningene er hvilken av disse kombinasjoner som det satses på og eventuelt hvor mye.

- Et visst kvantum av en vare skal innkjøpes til lager for senere salg. Etterspørselen er usikker og utfallene representerer ulike mulige etterspurte kvanta, og beslutningene er de ulike kvanta som kan kjøpes inn for lager.
- Et prosjekt skal gjennomføres og utfallene er de ulike forløp av hele prosjektperioden, og beslutningene representerer handlingsprogrammer, dvs. en beskrivelse av hvordan en skal forholde seg til enhver situasjon som kan oppstå i prosjektperioden.

Usikkerheten i et beslutningsproblem kan være av varierende karakter. På den ene side kan den bestå i ren tilfeldighet, slik tilfellet ofte er i hasardspill. På den annen side har vi problemer der usikkerheten består i manglende informasjon om tingenes tilstand, eksempelvis dersom man vurderer å ta med noen hjem til middag, uten å vite om man har de nødvendige ingredienser for vellykket aften. I det første tilfellet kan sannsynligheter for de ulike utfall fastsettes noenlunde objektivt, i det andre ikke, en kan stille spørsmål om det i det hele tatt er rimelig å gjøre bruk av sannsynligheter her. Det finnes også problemer som inneholder elementer av begge typer usikkerhet, eksempelvis dersom vi har en vare på lager og vi kjenner etterspørselstilbøyeligheten, f.eks. uttrykt ved en sannsynlighetsfordeling, men i tillegg er det usikkerhet om hva konkurrentene gjør for å eventuelt øke sin markedsandel.

Konsekvensene for beslutningstakeren i et beslutningsproblem kan også ha varierende karakter, fra rene økonomiske konsekvenser, til konsekvenser av typen goodwill, trivsel for seg selv og eventuelt andre som beslutningstakeren velger å ta omsyn til (f.eks. samfunnets tarv). Økonomiske konsekvenser lar seg som regel tallfeste, mens andre konsekvenser lar seg vanskelig tallfeste, i hvert fall slik at de er sammenlignbare og eventuelt kan oppsummeres i et enkelt tall.

Eksempel 1 : Maskinforretning

En maskinforretning vurderer å kjøpe inn et visst antall av en bestemt landbruksmaskin til lager før sesongen. Anta at bedriften har valgt å beskrive problemet med følgende tabell.

Etter-spørsel	Sann-synlighet	Gevinster ved innkjøp 0	1	2	3
0	0.1	0	-2	-4	-6
1	0.3	-1	5	3	1
2	0.4	-2	4	10	8
3	0.2	-3	3	9	15

Sannsynlighetene i tabellen er fastlagt ut fra vurdering av behov og økonomisk evne hos potensielle kunder, eventuelt også en vurdering av konkurrenters engasjement. Her er etterspurt kvantum høyst 3, med 2 som det mest sannsynlige. Gevinstene i tabellen er fastlagt ut fra en regnskapsmessig vurdering for de kommende år (enheten i tabellen kan f.eks. være kr. 10 000). Vi ser at hver solgt maskin bidrar med $+5$ til gevinsten, hver usolgt maskin med -2 (som skyldes lagerkostnader fram til neste sesong og eventuelt salg til lavere pris). Tap av goodwill ved udekket etterspørsel bidrar med -1 pr. enhet.

La oss ta disse tallene for gitt og prøve å bestemme et fornuftig innkjøpskvantum: En mulig beslutning er å velge å kjøpe inn det mest sannsynlige etterspurte kvantum, dvs. 2 enheter. Dette ignorerer helt konsekvensene. En annen mulighet er å treffe den beslutning som, for det verst (evt.best) tenkelige utfall for beslutningen, sikrer den største gevinst. Her er dette å kjøpe inn 1 (evt.3) enheter. Dette ignorerer sannsynlighetene for de ulike utfall, og vil nok bli ansett for å være for pessimistisk (evt. for optimistisk). I faglitteraturen blir disse to beslutningene kalt henholdsvis *maximin og maximax* beslutningen. I praksis ønsker vi en beslutning som tar omsyn til både sannsynlighetene og konsekvensene. En mulighet er å beregne forventet gevinst for hver beslutning. Eksempelvis dersom det kjøpes inn 2 enheter blir forventet gevinst

$$EG = (-4) \cdot 0.1 + 3 \cdot 0.3 + 10 \cdot 0.4 + 9 \cdot 0.2 = 6.3$$

og tilsvarende for de andre. Vi får

Innkjøpt kvantum	0	1	2	3
Forventet gevinst	-1.7	3.5	6.3	5.9

Vi ser at den beslutning som gir størst forventet gevinst er å kjøpe inn 2 enheter.

Et eksempel som dette reiser en del prinsipielle spørsmål:

1. Er det rimelig å bruke sannsynligheter til å beskrive ikke bare rene tilfeldigheter, men også ukjente omstendigheter, f.eks. andre personers bevisste men ukjente valg?
2. Kan vi rettferdiggjøre bruk av et enkelt mål for konsekvensene i form av gevinster?
3. Er i så fall forventede gevinster et rimelig vurderingsgrunnlag for valg av beslutning?
4. Hvordan kan vi sammenfatte separate vurderinger av en rekke usikre momenter og separate vurderinger av en rekke ulike konsekvenser?

Svarene på disse spørsmål vil avhenge noe av karakteren av det beslutningsproblem som foreligger. En beskrivelse av beslutningsproblemer som i Tabell 16.1 kalles beskrivelse på *normalform*. I praksis vil en ofte foretrekke å beskrive problemet på såkalt *ekstensiv form*. Dette er emnet for de følgende avsnitt som også inneholder delvise svar på spørsmålene ovenfor.

16.2 Beslutningsanalyse

Vi vil ta opp en teknikk for beslutningsanalyse som har vist seg nyttig i praksis, spesielt for kompliserte beslutningsproblemer av prosjekttypen, der mange delavgjørelser må treffes over tid. Teknikken kalles *beslutningstreanalyse*, og vi illustrerer den best ved å gjennomdrøfte et noenlunde realistisk problem:

Den 10.august 1996 får elektronikkfirmaet A/S Elektron vite at A/S Kjemi trenger 1000 kontrollenheter av en bestemt type for installasjon i en ny fabrikk. A/S Kjemi har tidligere benyttet konvensjonelle enheter, prisen på disse er nå 14 000 kr. pr. enhet. A/S Elektron tror at de kan være i stand til å lage en kontrollenhet som er bedre enn den eksisterende, og overveier å bygge en prototyp på egen bekostning. A/S Kjemi har sagt seg villig til å teste prototypen, og dersom den viser seg å være bedre enn den konvensjonelle, så vil firmaet plassere hele ordren på 1000 enheter hos A/S Elektron, til pris 14 000 kr. pr. enhet. Levering og oppgjør for partiet skal finne sted 31.des. 1996.

A/S Elektron mener at kostnadene ved å utvikle prototypen er 800 000 kr., og at dersom de får ordren, kan de tenke seg to produksjonsmetoder:

Metode 1:	faste kostnader 1 200 000 kr.
	variable kostnader 10 000 kr. pr. enhet.
Metode 2:	faste kostnader 1 700 000 kr.
	variable kostnader 9 000 kr. pr. enhet.

Produksjonen etter metode 2 krever kjøp av en del tilleggsutstyr og leveringsdatoen for dette er noe usikker. A/S Elektron regner imidlertid med at dersom de velger metode 2 og går til innkjøp av utstyret, men dette ikke når fram til en fastsatt dato, kan de skifte over til metode 1 og allikevel oppfylle kontrakten. Vi regner med at de faste kostnader da blir 1 700 000 kr., dvs. tilleggsinvesteringen på 500 000 kr. "går tapt". Produksjonen av de 1000 enhetene vil delvis gå på bekostning av A/S Elektrons øvrige virksomhet. En hadde regnet med en netto inntekt på 1 000 000 kr. på andre prosjekter i tiden fram til 31.des. 1996. Dersom produksjon finner sted, regner en med at denne inntekten blir halvert til 500 000 kr. A/S Elektron regner med at arbeidet med å utvikle prototypen ikke har slike sidevirkninger. Anta at A/S Elektron's "statusden 10.aug. 1996 er
−1 000 000 kr. Med status menes et mål for foretakets tilstand i øyeblikket. Det kan for eksempel være netto likvide aktiva. A/S Elektron regner med at sannsynligheten for at de skal få ordren er 0.5, og at sannsynligheten for at bestilt tilleggsutstyr skal nå fram i tide er 0.8.

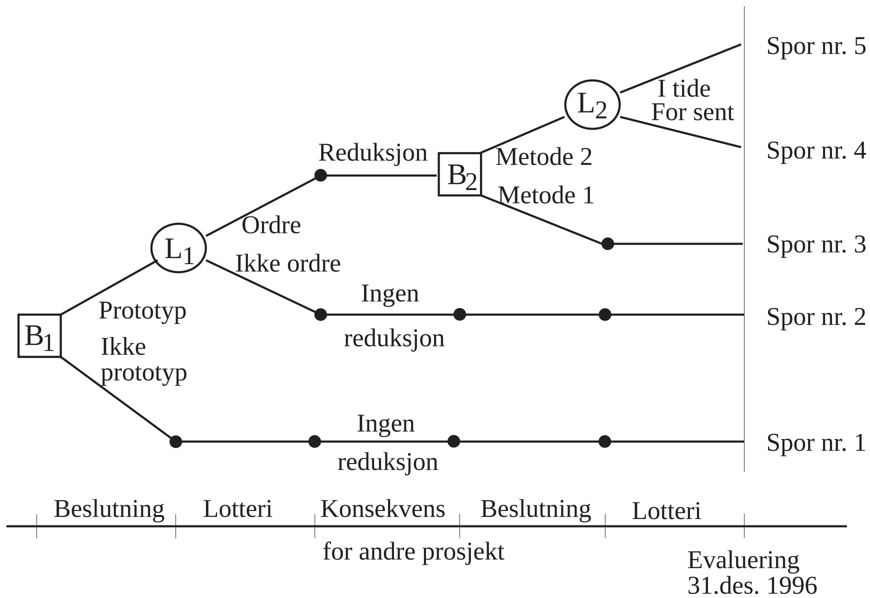
Anta at A/S Elektron ønsker å maksimere forventet fortjeneste i prosjektperioden (fram til 31.des. 1996). Som konsulenter skal vi bestemme et optimalt handlingsprogram for foretaket.

Dette er et eksempel på et beslutningsprogram under usikkerhet, av sekvensiell natur, der den første beslutning får konsekvenser for senere beslutninger. Slike beslutningsproblemer lar seg ofte beskrive oversiktlig ved et såkalt *beslutningstre*. I Figur 16.1 er det vist hvordan et beslutningstre for vårt problem kan se ut.

Vi tenker oss treet med roten på venstre side av arket med forgreninger mot høyre, hver gren mellom to forgreningspunkter svarer enten til en beslutning eller et utfall av et *"lotteri"*. Vi har to typer forgreningspunkter, markert med henholdsvis firkant og sirkel:

- betyr en beslutningssituasjon, hver gren som løper mot høyre fra dette forgreningspunktet svarer til en mulig beslutning.

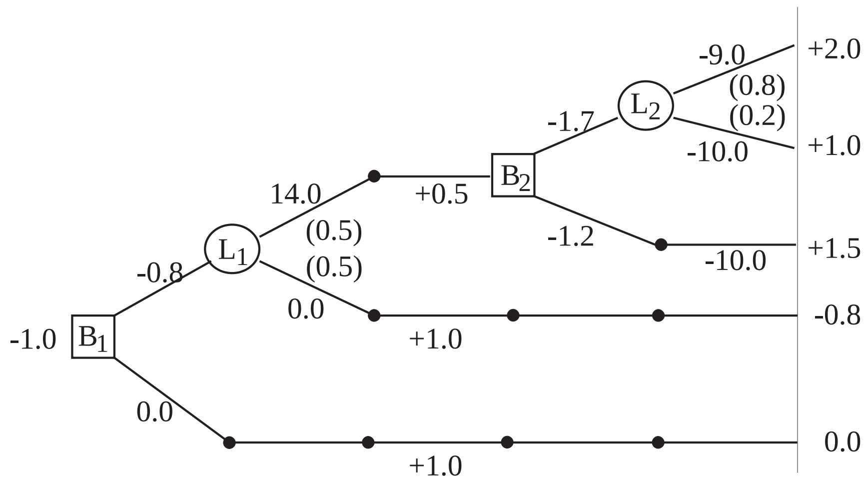
○ betyr et lotteri, hver gren som løper mot høyre fra dette forgreningspunktet svarer til et mulig utfall av dette lotteriet.



Figur 16.1: Beslutningtre A/S Elektron

Treet er organisert slik at beslutninger og utfall opptrer i logisk rekkefølge, dvs. i den rekkefølge informasjonen blir tilgjengelig. Ved et forgreningspunkt har vi til rådighet all den informasjon som ligger i grener som går forut for forgreningspunktet, dvs. grener som kan nås ved å gå mot venstre langs grenene.

La oss forsøke å rettferdiggjøre treet i Figur 16.1: Firkanten B_1 helt til venstre betegner den initiale beslutningssituasjon: skal foretaket utvikle en prototype eller ikke, svarende til henholdsvis øvre og nedre gren ut fra firkanten. Utvikles ikke prototypen er all usikkerhet fjernet, og det er heller ikke flere beslutningsproblemer, vi fører grenen fram til enden av treet (evalueringsdatoen). Velger en å utvikle prototypen står en ovenfor et lotteri L_1 med to utfall: ordre eller ikke ordre. Dersom en ikke få ordren, blir det ikke flere forgreninger, grenen føres fram til enden. Dersom en får ordren, er en i en ny beslutningssituasjon B_2, skal det produseres etter metode 1 eller metode 2. Velger en å produsere etter metode 1, er vi ved enden av treet, men velges metode 2 er vi i lotterisituasjonen L_2: blir det nødvendige tilleggsutstyret levert i tide eller ikke. Vi ser at prosjektet kan følge et av 5 ulike spor gjennom treet, et spor svarende til hvert av endepunktene nummerert fra 1 til 5.



Figur 16.2: Beslutningtre med data (Alle beløp i mill.)

I Figur 16.2 er beslutningstreet påført alle de relevante data som er gitt i teksten. Sannsynligheter er påført i parantes. Ved starten er status -1 000 000, og dersom en velger å utvikle en prototyp påløper kostnader 800 000. Dersom en får ordren blir salgsinntekten av de ferdige enheter 14 000 000 og i tillegg kommer netto 500 000 fra andre prosjekter. Produseres etter metode 1 påløper faste kostnader 1 200 000 og variable kostnader 10 000 000. For hvert spor gjennom treet får vi en betalingsstrøm som gir en bestemt fortjeneste, spor nr. 3 gir for eksempel

$$\begin{aligned}\text{Fortjeneste} &= -800\,000 + 14\,000\,000 + 500\,000 - 1\,200\,000 - 10\,000\,000 \\ &= +2\,500\,000\end{aligned}$$

Status ved enden av treet = -1 000 000 +2 500 000 = 1 500 000.
For hvert av de fem endepunktene av treet er påført status ved prosjektperiodens slutt. Det overlates til leser å sjekke alle detaljer.

Vi ser at spor nr.5 ville vært mest gunstig for foretaket, men dette kan ikke med sikkerhet realiseres, foretaket må først få ordren og tilleggsutstyret må nå fram i tide, og slik informasjon foreligger ikke på det tidspunkt foretaket skal treffe sin initiale beslutning. Et handlingsprogram må inneholde regler om hvilken beslutning som skal treffes i enhver beslutningssituasjon som kan oppstå i prosjektperioden. Vi skal finne det handlingsprogram som maksimerer forventet fortjeneste i prosjektperioden, dvs. fram til enden av treet. I dette tilfellet blir det ekvivalent med å maksimere forventet status ved enden av treet. Vi vil foretrekke å bruke dette begrep i den videre diskusjon, bl.a. fordi dette gir et uttrykk for foretakets evne til å inngå nye engasjementer ved prosjektperiodens slutt.

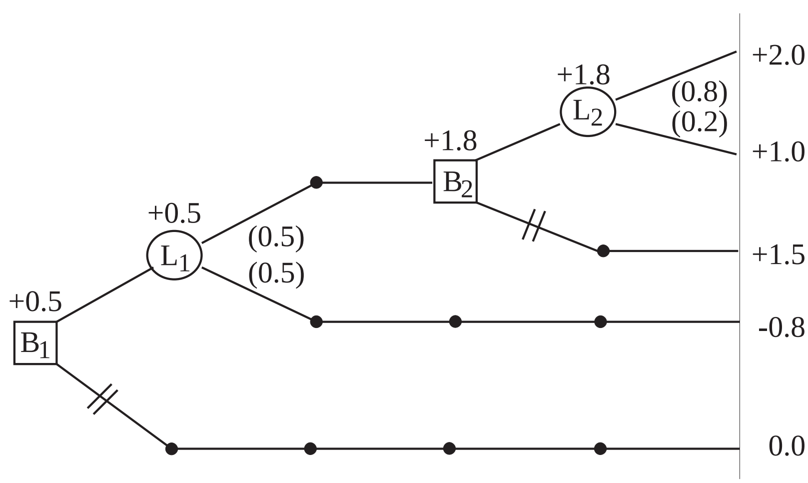
For å finne et handlingsprogram som er optimalt i den forstand vi har definert, kan vi i hovedtrekk gjøre følgende:

1. Beregn foretakets status ved hvert av de fem endepunktene av treet.

2. Beregn systematisk "verdien"av alle forgreningspunkter ved å starte i endepunktene av treet som gis verdi lik status, og gå så baklengs i treet til roten. Hver lotteriforgrening gis verdi lik forventet verdi av grenene, mens hver beslutningsforgrening får verdien til den gren som har størst verdi. Denne gren beholdes, de andre strykes.

3. Et optimalt handlingsprogram finnes ved å starte ved roten, og følge de grener som ikke er strøket ut.

Dette arbeidsprogram er utført i Figur 16.3.

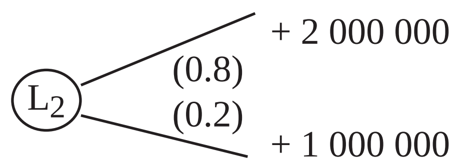


Figur 16.3: Analyse av beslutningstre (Alle beløp i mill.)

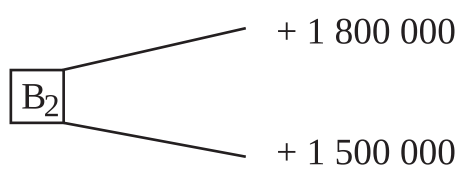
La oss illustrere hvordan beregningene utføres:
Figur 16.4 viser lotteriet L_2 som gir verdien +2 000 000 med sannsynlighet 0.8, verdien +1 000 000 med sannsynlighet 0.2. Forventet verdi blir

$$0.8 \cdot 2\,000\,000 + 0.2 \cdot 1\,000\,000 = 1\,800\,000$$

Når kriteriet er forventet sluttverdi, vil beslutningstakeren være indifferent mellom dette lotteri og et sikkert beløp på 1 800 000. Vi kan derfor tildele forgreningspunktet L_2 verdien +1 800 000. Sett fra beslutningspunktet B_2 ser situasjonen nå ut som i Figur 16.5.



Figur 16.4: Lotteri



Figur 16.5: Beslutning

Siden verdien av øvre gren er størst foretrekkes denne. Nedre gren strykes, og forgreningspunktet B_2 tildeles verdien +1 800 000. Det overlates til leseren å begrunne fortsettelsen i Figur 16.12. Vi ser at det optimale handlingsprogram er:

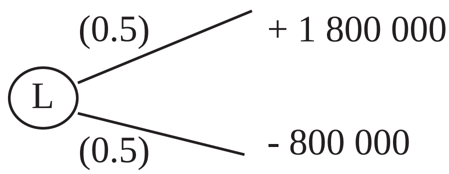
- Utvikl prototypen, dersom ordre, produser etter metode 2.

Det initiale beslutningspunkt har fått tildelt verdien +500 000. Dette er forventet sluttverdi ved bruk av det optimale handlingsprogram, dvs. forventet nettofortjeneste er +1 500 000.

Dersom foretakets målsetting er å maksimere forventet fortjeneste i perioden fram til 31.des.1996, har beslutningsproblemet nå funnet sin løsning. Et nærliggende spørsmål er imidlertid: Er forventet fortjeneste en rimelig målsetting? Kanskje det bekymrer foretaket at dersom de går inn for prosjektet, er sannsynligheten hele 0.5 for å ende opp med status −800 000, mens de helt sikkert når opp til status 0 ved å la være (risikoaversjon). På den annen side kunne det hende at foretaket desperat trengte til 2 000 000, og at det fortoner seg svært forlokkende å gamble på dette (risikovilje). I slike situasjoner vil neppe forventet fortjeneste være den relevante målsetting.

I mange problemstillinger inngår også konsekvenser av ikke-monetær karakter (f.eks. forurensing). Spørsmålet er nå om også slike situasjoner lar seg analysere ved hjelp av beslutningstrær.

I vår forventningsverdianalyse tildelte vi systematisk hvert lotteri en verdi, nemlig forventet verdi av grenene. Vi ønsker fortsatt å tildele en verdi til hvert lotteri. Denne verdi må sammenfatte beslutningstakerens holdning til både usikkerheten og konsekvensene. Dette kan gjøres ved å bestemme en såkalt *sikkerhetsekvivalent*.



Figur 16.6: Lotteri

La oss studere lotteriet L i Figur 16.6, hvor forventet verdi er

$$1\ 800\ 000 \cdot 0.5 + (-800\ 000) \cdot 0.5 = +500\ 000$$

Anta at beslutningstakeren blir tilbudt et fast beløp som erstatning for lotteriet. En beslutningstaker som bruker forventningsverdier som kriterium er indifferent mellom lotteriet og et sikkert beløp på 500 000, men villig til å selge lotteriet straks det tilbudte beløp overskrider 500 000. En beslutningstaker som har sterk aversjon mot posisjonen −800 000, vil trolig være villig til å godta et lavere beløp for å bli fritatt for den risiko lotteriet innebærer. Kan hende han er villig til å selge lotteriet for ethvert beløp større enn 300 000, men beholde lotteriet for alle beløp mindre enn 300 000. I så fall sier vi at beslutningstakerens sikkerhetsekvivalent for lotteriet er lik 300 000. Sikkerhetsekvivalenten blir et presist mål for beslutningstakerens holdning til den risiko lotteriet innebærer. Jo høyere sikkerhetsekvivalenten er, desto mer risikovillig er beslutningstakeren.

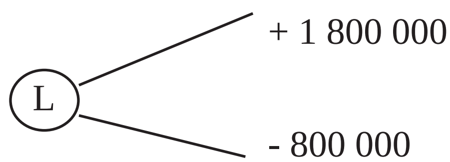
For å kunne bringe beslutningstakerens holdning til risiko inn i analysen, må vi ved hver lotteriforgrening i beslutningstreet, istedenfor forventningsverdien, notere foretakets sikkerhetsekvivalent for lotteriet. Analysen forgår da i prinsippet slik:

1. Beregn foretakets status ved hvert av endepunktene av treet.
2. Beregn systematisk verdien av alle forgreningspunkter ved å starte i endepunktene som gis verdi lik status, og gå så baklengs i treet til

roten. Hver lotteriforgrening gis en verdi lik foretakets sikkerhetsekvivalent, mens hver beslutningsforgrening får verdien til den gren som har størst verdi. Denne gren beholdes, de andre strykes.

3. Et optimalt handlingsprogram finnes ved å starte ved roten, og følge de grener som ikke er strøket ut.

Ovenfor har vi antatt at sannsynlighetene for de ulike utfall av lotteriet er gitte, vi har bevisst skjøvet under teppet at det også er problematisk å evaluere disse sannsynlighetene. I praksis vil som regel lotteriet L ovenfor fortone seg som i Figur 16.7,



Figur 16.7: Lotteri

dvs. uten objektive sannsynligheter for hvert av utfallene. Ønsker beslutningstakeren å gjøre en beslutningstreanalyse, må han være villig til å fastsette sin sikkerhetsekvivalent også for lotterier hvor sannsynlighetene ikke er objektivt gitt. Han må da foreta en simultan vurdering av sjansene for de ulike utfall og deres konsekvenser. Vurderingen vil nødvendigvis måtte bli mer eller mindre subjektiv, og det endelige valg av sikkerhetsekvivalent vil representere en beslutning fra beslutningstakerens side. Analysen foregår så i prinsippet som ovenfor.

I praksis er som regel et beslutningstre omfattende og komplisert, og for beslutningstakeren kan det fortone seg problematisk, tidkrevende, uvant og frustrerende å måtte ta stilling til en rekke fiktive lotterier, ofte av sammensatt natur, hvor både vurdering av usikkerhet og konsekvenser skal komme til uttrykk. Heri ligger også faren for at beslutningstakeren gjør seg skyldig i inkonsistente vurderinger. For å redusere disse problemer kan det være hensiktsmessig å skille beslutningstakerens vurdering av usikkerhet og konsekvenser. Videre vil det være ønskelig om beslutningstakerens holdning til konsekvenser er tilstrekkelig klarlagt på forhånd, slik at denne kan brukes ved vurderingen av ethvert lotteri som inngår i treet.

16.3 Preferanseindeks: Nytte

En beslutningstaker skal ta stilling til et eller flere prosjekter, der det i hvert kan inngå både egne og andres beslutninger samt usikre utfall som gjensidig avhenger av hverandre. Hver situasjon med usikkert utfall vil vi kalle et lotteri med gevinster, men gevinstene kan godt selv være et lotteri.

Når vi nedenfor bruker betegnelsen prospekt, kan det bety en gevinst (g), et lotteri (L) eller et helt prosjekt (P) som i forrige avsnitt. Dersom beslutningstakeren foretrekker prospekt a framfor prospekt b skriver vi $a \succ b$. Dersom beslutningstakeren er indifferent mellom prospektene a og b skriver vi $a \sim b$. Vi antar at beslutningstakeren, dersom han får seg forelagt to prospekter a og b, alltid er i stand til å avgjøre om $a \succ b$, $a \sim b$ eller $b \succ a$.

Vi kunne ønske oss en preferanseindeks π med egenskapen

$$a \succ b \Longleftrightarrow \pi(a) > \pi(b)$$

slik at ulike prospekter kan rangeres etter deres “beregnedepreferanse, og at denne indeksen tillater bruk av de vanlige regneregler for sannsynlighet og forventning. Dette er mulig dersom beslutningstakeren oppfyller følgende prinsipper for konsistent adferd:

Transitivitet:
La a, b og c være tre alternative prospekter.

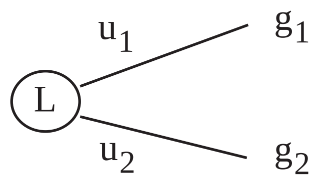
1. Dersom $a \sim b$ og $b \sim c$, så må $a \sim c$
2. Dersom $a \succ b$ og $b \succ c$, så må $a \succ c$
3. Dersom $a \succ b$ og $b \sim c$, så må $a \succ c$

Substitusjon:
Gitt et lotteri L, og la L' være det lotteri vi får ved å erstatte en av gevinstene g med en annen gevinst g'.

Dersom $g \sim g$' så må $L \sim L$'

Det kan vises at en beslutningstaker som oppfyller disse to adferdsantakelsene vil oppføre seg som om det var tilordnet sannsynligheter til alle utfall og preferanser til alle gevinster som omfattes av prospektene, slik at preferansen for de enkelte lotterier som inngår er forventet preferanse for gevinstene. Videre er helheten logisk konsistent, dvs at de vanlige regnereglene for sannsynligheter og forventninger gjelder for alle sammensatte lotterier såvel som hele prosjektet.

I litteraturen blir preferanseindeksen ofte kalt *nyttefunksjonen.* og resultatet kalles *nytteforventningsteoremet.* Dersom nyttefunksjonen er gitt vil beslutningsproblemet bestå i å maksimerere forventet nytte.



Figur 16.8: Lotteri

Dette betyr f.eks. at for en konsistent beslutningstaker som står overfor et lotteri L med to utfall u_1 og u_2, hvor gevinstene er henholdsvis g_1 og g_2 (se Figur 16.8), så vil

$$\pi(L) = \pi(g_1)p(u_1) + \pi(g_2)p(u_2)$$

dvs. nytten av lotteriet L er lik summen av nytten for gevinstene veid med de subjektive sannsynligheter for de respektive utfall eller m.a.o. lik forventet nytte. Dette kan så holdes opp mot nytten av andre lotterier eller nytten av en sikker gevinst g.

Bestemmelsen av sannsynligheter og preferanser kan i prinsippet reduseres til beslutninger av følgende to typer:

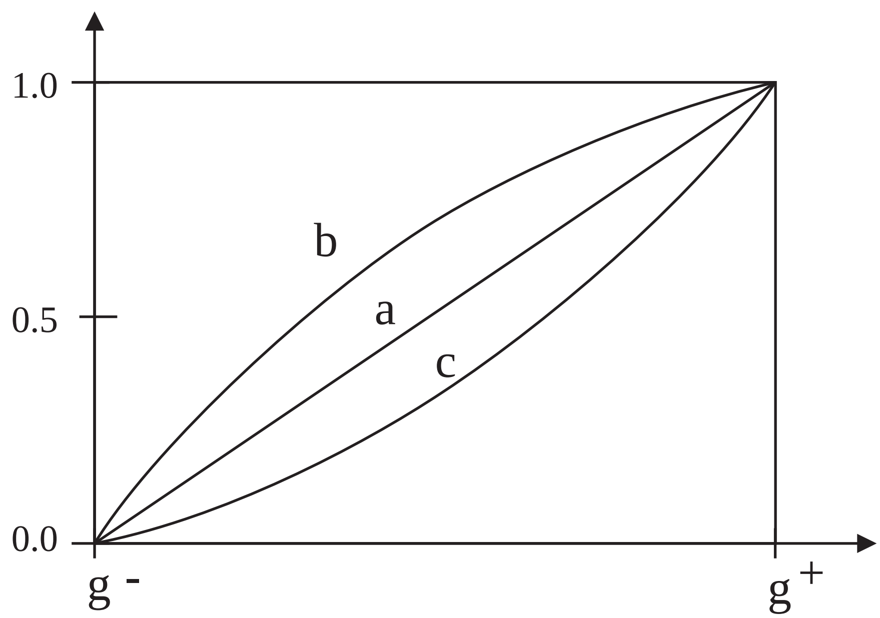
1. **Kalibrerende urne:**
   Gitt en urne med et stort antall (f.eks. 1000) kuler av ulike farger, en farge til hvert av de mulige utfall av det lotteri L som er under vurdering. Betrakt et alternativt lotteri L' som er tilfeldig trekning av en kule fra urnen, der gevinsten svarer til fargen på kulen, dvs. samme gevinster som lotteriet L. Blandingsforholdet mellom fargene varieres inntil beslutningstakeren er indifferent mellom L og L'. Sannsynlighetene for de ulike utfall er da gitt ved andelen kuler av de respektive farger.

2. **Referanselotteri:**
   Betrakt et lotteri med to monetære referansegevinster g^+ og g^-, som er slik at $g^+ \succ g \succ g^-$ for alle aktuelle gevinster g. Ved vurdering av gevinsten g kan beslutningstakeren bruke en kalibrerende urne, f.eks. med + kuler og − kuler. Blandingsforholdet mellom + kuler og − kuler

varieres inntil beslutningstakeren er indifferent mellom dette lotteriet og gevinsten g. Preferanseindeks for g er da andelen + kuler.

Det er klart at ethvert lotteri med et endelig antall utfall og tilhørende gevinster kan ved substitusjon reduseres til et lotteri om de to referansegevinstene, der andelene + og − kuler er bestemt ved formelen ovenfor. Dette er derfor et bevis for nytteforventningsteoremet i dette tilfellet. Et generelt bevis krever noe mer inngående argumentasjon, og noen tilleggsantakelser.

Dersom alle aktuelle gevinster er monetære, vil beslutningstakerens såkalte *preferansekurve* være et nyttig hjelpemiddel. Preferansekurven er preferanseindeksen $\pi(g)$ tegnet som funksjon av g. I Figur 16.9 er det tegnet inn tre ulike og karakteristiske preferansekurver.



Figur 16.9: Preferansekurver

En beslutningstaker som er indifferent mellom ethvert lotteri og dets forventede gevinst, vil ha preferansekurve lik den rette linje a (lineære preferanser), mens preferansekurvene b og c vil representere en beslutningstaker med henholdsvis risikoaversjon og risikovilje. Forsøk å forklare hvorfor!

I de fleste situasjoner kan beslutningstakerens preferansekurve avdekkes ved at å ta stilling til noen få lotterier, gjerne 50-50 lotterier (myntkast) fordi slike er lette å forestille seg. Når så tilstrekkelig mange punkter på preferansekurven er tegnet inn, trekkes en glatt kurve gjennom punktene.

Teorien ovenfor kan være nyttig ved beslutningstreanalyse, som kan. foregå slik

1. Beslutningstakeren avdekker sine subjektive sannsynligheter for utfallene av ethvert lotteri i treet.

2. Beslutningstakeren avdekker sin nytte for alle endepunktene i treet, f.eks ved avlesning på en allerede etablert preferansekurve.

3. Beregn systematisk nytten for alle forgreningspunkter i treet ved å starte i endepunktene av treet og arbeide seg baklengs til roten. Hver lotteriforgrening får nytte lik forventet nytte av grenene basert på de subjektive sannsynlighetene. Hver beslutningsforgrening får samme nytte som grenen som har størst nytte, denne gren beholdes, de andre strykes.

4. Et optimalt handlingsprogram finnes ved å starte ved roten, og følge de grener som ikke er strøket ut.

Vi tar opp igjen problemstillingen til A/S Elektron fra avsnitt 16.2: Beslutningstakeren evaluerer først sannsynlighetene for utfallene av de to lotteriene L_1 og L_2 (hjelpemiddel: kalibrerende urne). Anta at de subjektive sannsynlighetene ble som oppgitt tidligere.

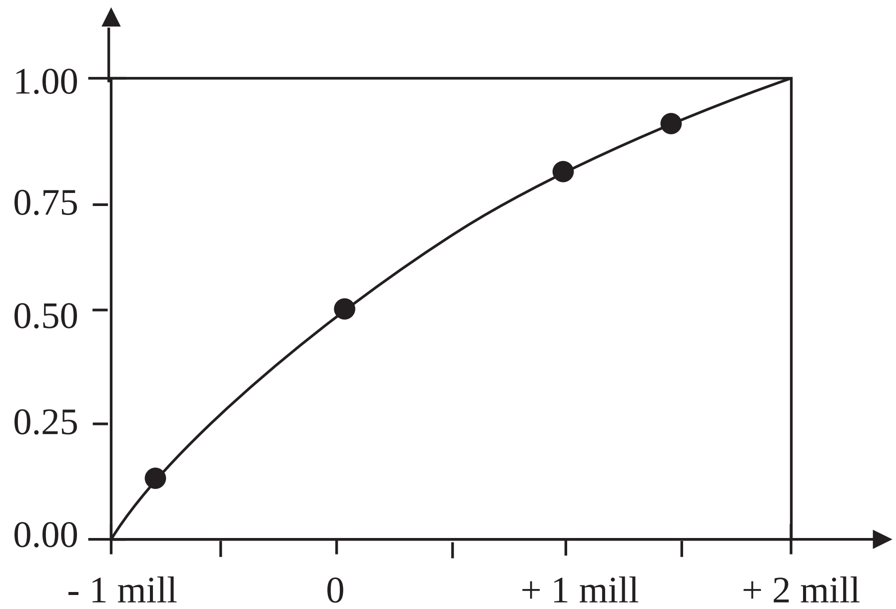
Beslutningstakerens avdekker så sin holdning til konsekvensene, her uttrykt ved status den 31.des. 1996. Som hjelpemiddel bruker vi et referanselotteri hvor referansegevinstene er valgt lik $g^- = -1\ 000\ 000$ og $g^+ = 2\ 000\ 000$. Disse skal tolkes som mulig status på evalueringsdatoen.

Beslutningstakeren har tatt stilling til de 4 gevinster -800 000, 0, 1 000 000 og 1 500 000. Dette ga opphav til følgende nytteverdier:

$$\pi(-1\ 000\ 000) = 0.00, \quad \pi(-800\ 000) = 0.14, \quad \pi(0) = 0.50$$
$$\pi(+1\ 000\ 000) = 0.80, \quad \pi(+1\ 500\ 000) = 0.91, \quad \pi(+2\ 000\ 000) = 1.00$$

Disse punktene er tegnet inn i et diagram, og preferansekurven er tegnet som en glatt kurve gjennom disse, se Figur 16.10. Diagrammet kan være nyttig som korrektiv dersom punktene avdekker et mønster som kan tyde på inkonsistens.

Vi ser at beslutningstakeren er indifferent mellom gevinsten 0 og et myntkast om de to referansegevinstene, og videre at beslutningstakeren har risikoaversjon, eksempelvis foretrekkes 0 framfor et myntkast om gevinstene



Figur 16.10: Preferansekurve A/S Elektron

-1 000 000 og $+1$ 000 000, som er et lotteri med forventning 0. Illustrer dette i figuren!

I Figur 16.11 har vi så gjennomført analysen med de funne nytteverdier. Det overlates til leseren å gjennomgå detaljene. Vi ser at optimalt handlingsprogram for foretaket er

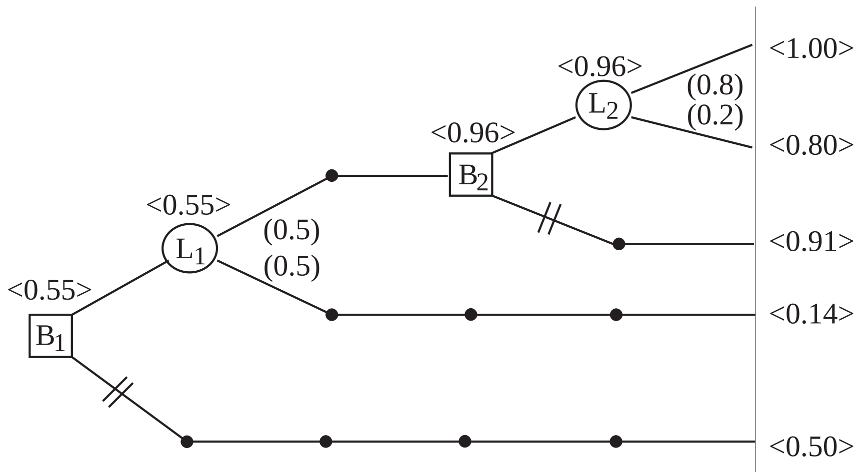
- Utvikl prototypen, dersom ordre, produser etter metode 2.

Vi ser at nytten for hele prosjektet er 0.55. Av preferansekurven ser vi at dette svarer til sikkerhetsekvivalent på 120 000 kr. Vi husker at verdien i forventningsverdianalysen var 500 000 kr., men dersom A/S Elektron har preferanse som uttrykt ovenfor, er prosjektet verdt bare 120 000 kr. Vi ser imidlertid at det optimale handlingsprogrammet er det samme etter begge kriterier.

I vårt enkle eksempel er det bare tre mulige handlingsprogrammer:

a_0: Utvikl ikke prototyp.
a_1: Utvikl prototyp, dersom ordre, produser etter metode 1.
a_2: Utvikl prototyp, dersom ordre, produser etter metode 2.

Hvert av disse handlingsprogrammene vil definere en preferanseindeks i ut-



Figur 16.11: Lotteri

gangsposisjonen. Av Figur 16.11 ser vi at

$$\pi(a_0) = 0.50, \quad \pi(a_1) = 0.525 \quad \text{og} \quad \pi(a_2) = 0.55$$

Det er lett å innse at disse indeksene kan beregnes ved, for hvert handlingsprogram, å beregne forventet preferanseindeks i sluttposisjonen, eksempelvis (se Figur 16.12)

$$\begin{aligned}
\pi(a_1) &= p(-800\,000)\pi(-800\,000) + p(0)\pi(0) \\
&\quad +p(+1\,000\,000)\pi(+1\,000\,000) + p(+1\,500\,000)\pi(1\,500\,000) \\
&\quad +p(+2\,000\,000)\pi(+2\,000\,000) \\
&= 0.5 \cdot 0.14 + 0 \cdot 0.50 + 0.1 \cdot 0.80 + 0 \cdot 0.91 + 0.4 \cdot 1.00 = 0.55
\end{aligned}$$

Her er benyttet at $p(+1\,000\,000) = 0.5 \cdot 0.2 = 0.1$ etter den vanlige regneregel for betingede sannsynligheter. En analyse av et beslutningstre kan derfor utføres ved å beregne preferanseindeksen for hvert handlingsprogram, og velge det som har størst indeks (nytte). Dette tilsvarer den analysemåten som er beskrevet i avsnitt 16.1.

Noen merknader til slutt: I mange bedriftsøkonomiske problemstillinger er det nok slik at beslutningstakerens preferanser, i det aktuelle konsekvensområde, ikke avviker svært mye fra lineære preferanser. Det innebærer at det handlingsprogram som maksimerer forventet fortjeneste som regel også er det optimale. En forventningsanalyse vil da kunne raskt gi en pekepinn på hva som er optimalt, mens den eksakte verdi av prosjektet kan avvike betydelig fra forventningsverdien, og kan først bestemmes ved at preferansene trekkes inn i analysen. For prosjekter med stor usikkerhet og med konsekvenser i et stort variasjonsområde, hvorav noen kanskje setter foretaket i likviditetsvansker, vil som regel en forventningsverdianalyse ha liten verdi, preferansene må trekkes inn i analysen.

I dette avsnittet har vi illustrert den generelle teorien ved et eksempel hvor alle konsekvensene var monetære. I praksis har vi svært ofte beslutningsproblemer hvor også ikke-monetære konsekvenser inngår, og hvor beslutningstakeren ønsker å ta disse med i analysen av problemet. Det er fullt mulig å trekke slike konsekvenser inn i en beslutningstreanalyse. Begrepene sikkerhetsekvivalent og referanselotteri er faktisk nyttige hjelpemidler til å sette en pris på slike konsekvenser.

16.4 Bayesiansk inferens

Anta at du har en problemstilling som du ønsker å betrakte som et beslutningsproblem under usikkerhet. Dersom du er villig til å følge forutsetninger om konsistent adferd av den typen som er nevnt innledningsvis i avsnitt 3, så medfører det at din holdning til usikkerheten kan la seg representere ved såkalte subjektive sannsynligheter som oppfyller de vanlige regneregler for sannsynligheter som ble utviklet i Kapittel 2 og 4.

Vi vil her ta opp en problemstilling av prinsipiell natur som har konsekvenser for statistisk praksis. Anta at et "systemkan være i en av m mulige tilstander $\theta_1, \theta_2, \ldots, \theta_m$, vi vet bare ikke hvilken tilstand systemet er i. Situasjonen dreier seg altså ikke om tilfeldighet, men om uvisshet om tingenes tilstand. Vi kan illustrere forskjellen med et eksempel.

Eksempel 2 : Falsk mynt?
I et spill brukes en mynt som kan være

$$\begin{aligned} \theta_1 &= \text{rettferdig} \\ \theta_2 &= \text{falsk med kron på begge sider} \\ \theta_3 &= \text{falsk med mynt på begge sider} \end{aligned}$$

Anta at du ikke har adgang til å granske mynten nærmere, slik at du er i villrede om myntens karakter. Anta at du er i en beslutningssituasjon, f.eks. skal velge strategi i et spill hvor mynten brukes (evt. om du skal delta overhodet). Tenk deg en rekke alternative spill der typen av mynt som brukes i hovedspillet bestemmes ved trekning av en kule fra en urne med kuler av tre farver i varierende blandingsforhold. Når blandingsforholdet er slik at du anser dette spill likeverdig med det opprinnelige har vi fått fastlagt dine subjektive sannsynligheter for de ulike muligheter. Anta at resultatet ble

$$P(\theta_1) = 0.4 \quad P(\theta_2) = 0.3 \quad P(\theta_3) = 0.3$$

Vi vil nå se på hvordan disse såkalte *apriorisannsynlighetene* modifiseres i lys av ny informasjon, f.eks. ved at du får vite utfallet av en eller flere kast med mynten. Du vil antakelig velge følgende modeller for et myntkast for de tre ulike typer mynter det er tale om

$$\begin{array}{lll} \theta_1: & P(K \mid \theta_1) = \frac{1}{2} & P(M \mid \theta_1) = \frac{1}{2} \\ \theta_2: & P(K \mid \theta_2) = 1 & P(M \mid \theta_2) = 0 \\ \theta_3: & P(K \mid \theta_3) = 0 & P(M \mid \theta_3) = 1 \end{array}$$

Vi har her brukt en skrivemåte som antyder at disse sannsynligheter oppfattes som betingede sannsynligheter gitt type mynt. Anta at du får vite at et myntkast med den for deg ukjente mynt ga kron. Dine opprinnelige sannsynligheter er nå ikke lenger relevante. Bayes lov gir for i=1, 2, 3

$$P(\theta_i \mid K) = \frac{P(\theta_i) \cdot P(K \mid \theta_i)}{P(K)}$$

hvor

$$P(K) = P(\theta_1) \cdot P(K \mid \theta_1) + P(\theta_2) \cdot P(K \mid \theta_2) + P(\theta_3) \cdot P(K \mid \theta_3)$$

I dette tilfelle blir

$$P(\theta_1 \mid K) = 0.4 \quad P(\theta_2 \mid K) = 0.6 \quad P(\theta_3 \mid K) = 0$$

som kalles *aposteriorisannsynligheter* gitt ny informasjon. Dersom du får vite at to myntkast ga to kron blir i steden aposterisannsynlighetene

$$P(\theta_1 \mid KK) = 0.25 \quad P(\theta_2 \mid KK) = 0.75 \quad P(\theta_3 \mid KK) = 0$$

Det er lett å sjekke at beregning direkte ut fra apriorisannsynlighetene og trinnvis beregning ved å modifisere aposteriorisannsynlighetene etter første kron for en ny kron gir samme resultat (se Oppgave 13).

Tankegangen i dette eksemplet kan formuleres generelt slik: Gitt m mulige tilstander $\theta_1, \theta_2, \ldots, \theta_m$ og anta at dine subjektive apriorisannsynligheter for disse er

$$P(\theta_i) \;\; ; \;\; i = 1, 2, \ldots, m$$

For å skaffe informasjon om hvilken θ som foreligger observeres en begivenhet B. Anta at vi etablerer sannsynligheter for det observerte begivenhet for hver mulig θ_i, dvs.

$$P(B \mid \theta_i) \;\; ; \;\; i = 1, 2, \ldots, m$$

Dine (subjektive) aposteriorisannsynligheter for de m mulige tilstandene er da gitt ved Bayes lov

$$P(\theta_i \mid B) = \frac{P(\theta_i) \cdot P(B \mid \theta_i)}{P(B)} \;\; ; \;\; i = 1, 2, \ldots, m.$$

Det kunne også være av interesse å ta opp problemstillinger der "systemetkan være i en av et uendelig antall tilstander:

Eksempel 3 : Falsk mynt?
I et spill brukes en mynt som kan være falsk, la θ være sannsynligheten for kron. Apriori kan θ være en vilkårlig verdi i intervallet 0 til 1, dvs. det foreligger et (overtellbart) uendelig antall muligheter. Dette kan betraktes som en generalisering av problemstillingen i Eksempel 2 hvor kun 3 θ-verdier, nemlig $\frac{1}{2}$, 1 og 0, var apriori mulige. Vi kan nå ønske å gå fram på analog måte: Formulere apriorisannsynligheter for θ, skaffe informasjon om θ ved å eksperimentere med mynten; av dette finne aposteriorisannsynligheter for θ. Vi ser at problemstillingen krever omgang med kontinuerlige sannsynlighetsfordelinger som fordeler sannsynlighet for θ i intervallet fra 0 til 1. Før vi ser nærmere på dette vender igjen oppmerksomheten mot inferensproblemet generelt.

Vi betraktet i kapitlene 7-15 statistiske inferensproblemer som det å trekke slutninger om en parameter θ i en delvis spesifisert modell på grunnlag av observasjoner, la oss realisert verdi x av en stokastisk variabel X. La oss anta at modellen kan beskrives ved at sannsynlighetsfordelingen til X

$$p(x \mid \theta)$$

er gitt som funkjon av θ, hvor θ har et gitt mulighetsområde [1]. Forøvrig antas θ å være ukjent, og den valgte inferensmetode tar sikte på å fungere uansett

[1] Vi har her brukt symbolene for diskrete fordelinger men, som vi skal se senere, vil vi anvende analog tankegang i tilfellet med kontinuerlige fordelinger.

hva den sanne verdi av θ er. De metoder vi har studert tar ikke omsyn til eventuell apriori viten om θ, utover eventuelt for å planlegge observasjonsomfanget. En mulighet for å innlemme apriori viten om θ i analysen vil være å spesifisere en sannsynlighetsfordeling

$$p(\theta)$$

som reflekterer ens, mer eller mindre subjektive mening om hvilke θ-verdier som er sannsynlige. Denne sannsynlighetsfordelingen kalles *apriorifordelingen* til θ, fordi den oppsummerer vår viten om θ før vi observerer X. Vi ønsker å modifisere vår viten om θ i lys av den observerte x. Dette kan gjøres ved å bruke Bayes lov (se Kapittel 4.2) til å beregne den betingede fordeling for θ gitt $X = x$. Vi får

$$p(\theta \mid x) = \frac{p(\theta) \cdot p(x \mid \theta)}{p(x)}$$

Her er $p(x)$ den marginale fordeling til X gitt ved

$$p(x) = \sum_{\theta} p(\theta) \cdot p(x \mid \theta)$$

$p(\theta \mid x)$ kalles *aposteriorifordelingen* til θ, fordi den oppsummerer vår viten om θ etterat obsevasjonene foreligger. Ønsker vi et estimat for θ, kan vi bruke forventningen i aposteriorifordelingen. En estimator konstruert på denne måten blir ofte kalt en *Bayes estimator*, og denne måten å trekke slutninger på kalles *Bayes inferens*.

Eksempel 4 : Binomisk situasjon
En leverandør leverer råstoff til en produsent som masseproduserer en artikkel. Det kan leveres to kvaliteter av råstoffet, og man regner at kvalitet 1 gir 5% vrak og kvalitet 2 gir 10% vrak. Det er gjort avtale med leverandøren at kvalitet 1 skal leveres og betalingen er deretter. Et parti råstoff er mottatt, men mottaker er ikke trygg på at det er av avtalt kvalitet. For å klargjøre dette undersøkes $n = 10$ ferdige enheter og antall vrakede noteres. Anta at antall vrakede X blant de undersøkte er binomisk fordelt $(n = 10, \theta)$, der θ er enten 0.05 eller 0.10. Nå er

$$p(x \mid \theta) = \binom{10}{x} \theta^x (1-\theta)^{10-x}$$

Anta at kjøper apriori tror at sjansen for lureri eller ikke er "fifty-fifty". Dette svarer til apriorifordelingen

θ	0.05	0.10
$p(\theta)$	0.5	0.5

Bruker vi Bayes formel får vi at aposteriorifordelingen er gitt ved uttrykket

$$p(\theta \mid x) = \frac{0.5 \cdot \binom{10}{x}\theta^x(1-\theta)^{10-x}}{0.5 \cdot \binom{10}{x}0.05^x 0.95^{10-x} + 0.5 \cdot \binom{10}{x}0.10^x 0.90^{10-x}}$$

Anta at observert verdi er $x = 2$. Av en tabell over binomiske sannsynligheter får vi

$$\binom{10}{2}0.05^2 0.95^8 = 0.746 \quad \text{og} \quad \binom{10}{2}0.10^2 0.90^8 = 0.2683$$

Dette gir oss følgende aposteriorisannsynlighter for θ:

θ	0.05	0.10
$p(\theta \mid x = 2)$	0.28	0.72

Det overlates til leseren å sjekke at dersom en isteden hadde valgt apriorisannsynligheten

θ	0.05	0.10
$p(\theta)$	0.70	0.30

så ville aposteriorisannsynlighten med det samme observerte resultat bli

θ	0.05	0.10
$p(\theta \mid x = 2)$	0.47	0.53

Aposteriorisannsynligheten kan brukes som grunnlag for å foreta et valg, påstå at kvalitet 2 er levert eller ikke, eventuelt foreta en (kostbar) materialanalyse av råstoffet. Hva en bør gjøre vil kunne avhenge av de kostnadsvurderinger, noe som vi kunne spinne videre på.

Analysemåten vi har illustrert med dette eksemplet er et Bayesiansk alternativ til hypotesetesting. Vi har også Bayesianske alternativ til estimering, igjen illustrert med et binomisk eksempel. En del av den matematikk som trengs for å begrunne metoden fullt ut ligger utenfor det vi krever av leseren, og på enkelte punkter nøyer vi oss med å sannsynliggjøre resultatene.

Eksempel 5 : Binomisk situasjon - estimering
Et firma ønsker å markedsføre et produkt på landsbasis med garanti om at

ved eventuelle defekter som oppstår innen et år etter kjøpedato, vil produktet bli erstattet med et nytt. Før dette gjøres ønsker firmaet å estimere sjansen p for at en solgt enhet må erstattes. Dette blir gjort ved å markedsføre produktet i et avgrenset område over et kortere tidsrom, f.eks. en måned. Anta at det i dette tidsrom ble solgt n enheter. Et år senere vil man observere antall reklamasjoner X som fører til erstatning av produktet. Under rimelige antakelser er X binomisk fordelt (n, p), dvs. har punktsannsynlighet

$$p(x \mid p) = \binom{n}{x} p^x (1-p)^{n-x}$$

Firmaet har noe forhåndsinformasjon om p, bl.a. basert på laboratorietester av produktet og generell kjennskap til kunders tilbøyelighet til å kreve sin rett.

I slike situasjoner er det hensiktsmessig å oppsummere apriori informasjonen ved å spesifisere en kontinuerlig apriorifordeling for p over intervallet [0,1], representert ved en sannsynlighetstetthet med passende form. Bayes lov i denne situasjonen lyder

$$f(p \mid x) = \frac{f(p) \cdot p(x \mid p)}{p(x)}$$

der $p(x)$ er den ubetingende punktsannsynlighet, gitt ved integralet av telleren mhp. p over intervallet [0,1]. Det er bekvemt å spesifisere en såkalt Betafordeling med parametre (r, s) som apriorifordeling for p. Denne har formen ($\propto$ betyr proporsjonal med)

$$f(p) \propto p^{r-1}(1-p)^{s-1}$$

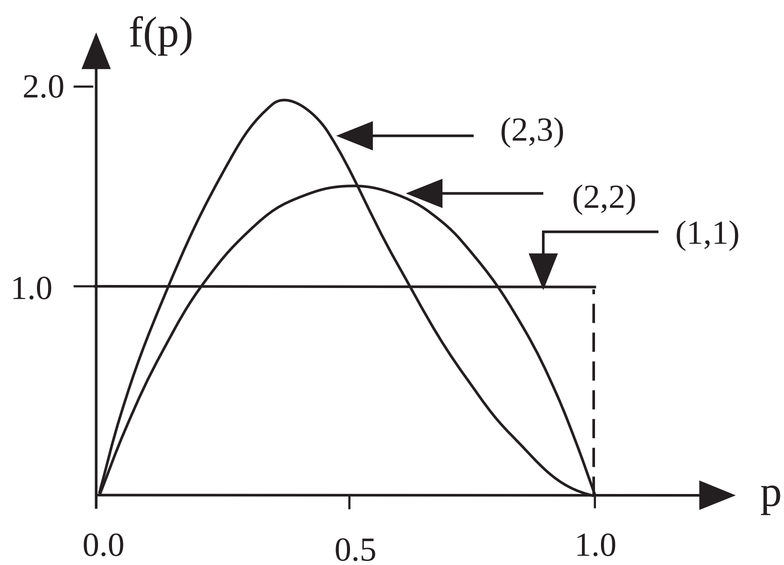
Her er r og s er positive tall som kan velges slik at fordelingen får en form som reflekter apriori kunnskapen. Det kan vises at

$$\text{Forventningen} = \frac{r}{r+s}$$

Vi ser at $r = s$ innebærer at forventningen er $1/2$. Noen ulike Betafordelinger med tilhørende forventning er gitt i tabellen

r	s	f(p)	Forventning
1	1	1	1/2
1	2	$2(1-p)$	1/3
2	1	$2p$	2/3
2	2	$6p(1-p)$	1/2
2	3	$12p(1-p)^2$	2/5
3	2	$12p^2(1-p)$	3/5
3	3	$30p^2(1-p)^2$	1/2

I figuren er tegnet tettheten for tre kombinasjoner av (r, s). Leseren kan selv tegne inn noen flere.



Figur 16.12: Apriorifordelinger: Beta

Vi ser at det er mulig å få frem en rekke ulike fordelingsformer ved å velge r og s hensiktsmessig, og i praksis skulle det være mulig å finne en som avspeiler apriorikunnskapen godt nok. Figur 16.12 antyder at variansen blir mindre og mindre ettersom r og s gjøres større. Er vi derfor nokså sikre på hva p omtrent er, velges r og s stor, og i samsvar med ens apriori forventning.

Det viser seg at aposteriorifordelingen for p gitt $X = x$ er Betafordelingen med parametre $(r + x, s + n - x)$. En god gjetning på p er forventningen i denne fordelingen som blir

$$\tilde{p} = \frac{x+r}{n+r+s}$$

Dette kalles Bayes-estimatoren for p i den binomiske situsajon (med Beta-apriorifordeling).

Anta at man finner ut at Betafordelingen med $r = 1$ og $s = 4$ gir et brukbart uttrykk for apriorikunnskapen om p. Merk at aprioriforventningen er 0.20, men enhver sannsynlighet er mulig. Anta videre at det ble solgt $n = 100$ enheter, og at etter ett år fikk man $x = 8$ reklamasjoner som medførte erstatning. Bayes-estimatet for p blir da

$$\tilde{p} = \frac{8+1}{100+1+4} = \frac{9}{105} = 0.076$$

dvs. nær det tradisjonelle estimatet, som er andelen erstatninger 0.08. Merk at med så mange som 100 observasjoner vil disse dominere over den uttrykte apriorikunnskap. For en mer skråsikker analytiker som innhenter færre observasjoner vil det være omvendt. Eksempelvis med $r = 4$, $s = 16$ og $n = 25$, $x = 2$ blir estimatet isteden korrigert fra 0.20 til 0.13, selv om aprioriestimatet såvel som observert andel er det samme som ovenfor.

Eksempel 6 : Målemodellen

Vi ønsker å estimere forventningen μ i målemodellen på grunnlag av subjektiv forhåndsvurdering og n observasjoner $X_1, X_2, \ldots, X_n$ med forventning μ og varians σ^2. På forhånd forventes μ å være μ_0, og graden av sikkerhet om dette tenkes uttrykt ved en varians σ_0^2. Som estimator for μ kan brukes

$$\tilde{\mu} = \frac{(n/\sigma^2)\overline{X} + (1/\sigma_0^2)\mu_0}{(n/\sigma^2) + (1/\sigma_0^2)}$$

dvs. en veid sum av aprioriforventningen og det observerte gjennomsnitt, med vekter som er omvendt proporsjonale med de respektive varianser. Denne estimator framkommer dersom vi som apriorifordeling bruker at μ er fordelt $N(\mu_0, \sigma_0^2)$ og antar at observasjonene er uavhengige $N(\mu, \sigma^2)$ variable. Det kan da vises at aposteriorifordelingen for μ blir
$N(\tilde{\mu}, 1/(n/\sigma^2 + 1/\sigma_0^2))$, slik at μ er forventningen i aposteriorifordelingen. Merk at vi må kjenne variansene for å bruke denne estimatoren.

Bruken av Bayes lov er sentral i denne type inferens og statistikere som bekjenner seg til denne skolen kalles derfor *Bayesianere* [2]. Det har ofte vært

[2]Oppkalt etter Thomas Bayes (1702-1761).

faglig uenighet blant statistikere i hvilken utstrekning det er rimelig å gjøre bruk av Bayesianske metoder. Hovedinnvendingen er at siden slik analyse er avhengig av en subjektiv apriorifordeling, vil to personer kunne trekke forskjellig konklusjon ut fra samme observasjoner og den samme modell. Mange statistikere mener at slik praksis ikke er tilrådelig, og søker å etablere inferensmetoder som ikke gir rom for subjektive vurderinger i form av apriorifordelinger, dvs. slik det ble gjort i Kapittel 7. Teorien i det kapitlet representerer klassisk statistisk teori. Bayesiansk teori er i hovedsak utviklet senere, selv om idéen er gammel.

Med Bayesiansk tankegang vil en rekke av de grunnleggende begreper i klassisk statistisk teori ikke lenger ha noen funksjon, eksempelvis forventningsretthet, signifikansnivå, styrke. Andre begreper får en helt annen fortolkning, eksempelvis konfidensintervall.

De statistiske problemer som er behandlet i kapitlene 7-15 kan alle tas opp fra et Bayesiansk synspunkt. Dette krever, selv i de enklere problemstillingene, matematiske kunnskaper utover det vi har forutsatt. Imidlertid har de senere års regneteknologi gjort det mulig å tilrettelegge praktiske tilnærminger, der det meste av teorien kan skjules for bruker uten særlig risiko for misbruk.

Ulike skoler statistikere har ofte stått steilt mot hverandre i metodespørsmål og en kilde til uenighet finner vi i krysningspunktet mellom objektivitet og subjektivitet. Debatten er fremdeles levende, men mange statistikere (iblant dem forfatteren) har vanskelig for å tro at et felles metodegrunnlag for alle typer problemer kan finnes, og ser derfor med velvilje på alternativer. En ting som taler til fordel for Bayesianisme er at det, i hvert fall i prinsippet, gir en enhetlig ramme for statistiske problemer, en ramme der forhåndsviten kan og må bringes inn som en integrert del av analysen. Bayesianske metoder er vel verdt et nærmere studium, slike metoder har fått økt popularitet i de senere år, selv om enkeltes forhåpninger om et Bayesiansk 21. århundre kan synes noe optimistisk.

16.5 Oppgaver

1. Gitt en situasjon med fem mulige beslutninger og tre mulige utfall der gevinstene er gitt ved

Utfall	Beslutninger a_1	a_2	a_3	a_4	a_5
u_1	1	3	−2	4	0
u_2	0	1	2	2	3
u_3	1	−2	1	−2	−1

(a) Finn maximin og maximax beslutningen.

(b) Finnes det noen beslutning som vi kan se bort fra i den videre analyse.

(c) Finn den beslutning som maksimerer forventet gevinst dersom sannsynlighetene for de ulike vurderes til

	u_1	u_2	u_3
p_i	0.1	0.6	0.3

2. La situasjonen være som i Eksempel 1 med unntak av at tap av goodwill ved udekket etterspørsel ikke forekommer. Finn den beslutning som maksimerer forventet gevinst. Hva blir forventet gevinst i dette tilfellet?

3. En bedrift vurderer om de skal utvikle et nytt produkt eller ikke. Kostnadene ved å utvikle produktet regnes til 1 mill. kr. Sannsynligheten for at dette gir et produkt som kan markedsføres vurderes til 0.7. Ved eventuell produksjon er to ulike produksjonsvolum aktuelle, stort eller lite. Etterspørselen er uviss, her antas for enkelhets skyld kun to nivåer, høy etterspørsel med sannsynlighet 0.4, lav med sannsynlighet 0.6. Netto inntekt i mill. kr. (etter fradrag av utviklingskostnader) antas å være gitt ved

Produksjon	Etterspørsel Høy	Lav
Stor	2.5	−0.5
Liten	−0.2	−0.3

(a) Lag et beslutningstre for problemstillingen der du påfører de gitte opplysninger.

(b) Finn optimalt handlingsprogram dersom bedriften vil maksimere forventet inntekt.

4. En bedrift vurderer et prosjekt der største "gevinst" g^+ er 8 mill. kr., minste gevinst g^- er −4 mill. kr. Bedriften anser en sikker gevinst på 2 mill. kr. som likeverdig med et lotteri der sannsynlighetene for g^+ og g^- er henholdsvis 0.8 og 0.2.

(a) Tegn en preferansekurve som avspeiler dette forhold.

(b) Anta at den tegnede kurve er bedriftens preferansekurve. Hva vil da bedriften foretrekke, sikre 1 mill. kr. eller et lotteri om gevinstene −1 mill. kr. og 2 mill. kr. med like sannsynligheter?

(c) Viser den tegnede kurve risikovillighet eller risikoaversjon?

5. Drøft i hvilken grad spørsmålene som ble formulert ved slutten av avsnitt 16.1 blir besvart i de etterfølgende avsnitt.

6. Finn aposteriorisannsynlighetene i Eksempel 2 når apriorisannsynlighetene er
$$P(\theta_1) = P(\theta_2) = P(\theta_3) = \frac{1}{3}$$
og vi observerer

   (a) Et myntkast som ga kron.
   (b) To myntkast som begge ga kron.
   (c) To myntkast som ga en kron og en mynt.

   Løs også oppgaven dersom apriorisannsynlighetene isteden er
$$P(\theta_1) = 0.8 \quad P(\theta_2) = 0.1 \quad P(\theta_3) = 0.1$$

7. (Apriorianalyse) Mynten i Eksempel 2 blir brukt i et spill der innsatsen er 100 kroner, mens utbetalingen er 150 kroner dersom du gjetter utfallet av et kast med mynten. Anta at de tre mulige mynter apriori anses like sannsynlige.

   (a) Tegn et beslutningstre for situasjonen.
   (b) Finn en gjetning som gir størst forventet gevinst.
   (c) Lønner det seg å spille når målsettingen er å maksimere forventet gevinst?

8. (Aposteriorianalyse) En venn har nettopp spilt spillet i forrige oppgave og opplyser at mynten viste kron.

   (a) Utnytt denne informasjon til å bygge ut beslutningstreet.
   (b) Finn den gjetning som gir størst forventet gevinst.
   (c) Lønner det seg nå å spille?

9. (Preposteriorianalyse) En ukjent har nettopp spilt spillet i Oppgave 7 to ganger. Han tilbyr følgende informasjon for salg: Betal 20 kroner for utfallet av første spill, alternativt betal 30 kroner for utfallet av begge spill.

   (a) Bygg ut beslutningstreet til å omfatte eventuell kjøp av informasjon.
   (b) Finn den optimale løsning av problemet.
   (c) Lønner det seg nå å spille?
   (d) Hvor mye vil du være villig til å betale for å få perfekt informasjon om myntens karakter?

10. Gjennomfør en alternativ løsning av Oppgave 7, 8 og 9 der vi på forhånd spesifiserer alle tenkelige handlingsprogrammer og mulige utfallssekvenser med sine respektive gevinster (jfr. Tabell 16.1).

    Hint: Gjør bruk av regning med betingede sannsynligheter og Bayes lov til å finne de nødvendige sannsynligheter.

11. Et tilbud skal sendes ut til alle medlemmene i en større bokklubb. To medarbeidere A og B er satt til å vurdere hvor stor andel som vil reflektere på tilbudet. De oppgir følgende subjektive sannsynligheter.

Andel som reflekterer	0.05	0.10	0.15
Sannsynligheter (A)	0.6	0.3	0.1
Sannsynligheter (B)	0.2	0.5	0.3

    Ledelsen velger å tillegge vurderingene til A og B like stor vekt.

    (a) Hvilke sannsynligheter bør i så fall ledelsen legge til grunn for beslutning om opplagstall?
    (b) Hva er forventet andel som reflekterer ifølge denne fordelingen.
    (c) Besvar også (a) og (b) dersom ledelsen tillegger B's vurdering dobbelt så stor vekt som A's.

12. En bedrift har bestilt 100 sekker råstoff, 50 sekker som man vet gir gjennomgående 5% defekte artikler (1) og 50 sekker av en noe billigere kvalitet (2) som gir gjennomgående 10% defekte. Ved mottak av sekkene går det ikke fram hva som er hva, de er merket A og B. Finn sannsynligheten for at sekker merket A inneholder henholdsvis kvalitet 1 og 2 dersom

    (a) Bedriften produserer n=10 artikler med råstoff fra A-sekk og en av disse ble defekt.
    (b) Bedriften deretter produserer n=10 artikler med råstoff fra B-sekk og en av disse ble defekt.

13. $\star$ Vis at dersom informasjon blir tilgjengelig i to porsjoner så gir trinnvis anvendelse av Bayes lov og direkte anvendelse med all informasjon samme resultat. Med symboler, dersom

$$\begin{aligned} p(\theta) &\rightarrow q(\theta) = p(\theta \mid x) \rightarrow q(\theta \mid y) \\ p(\theta) &\rightarrow p(\theta \mid x, y) \end{aligned}$$

14. En mynt skal brukes i et spill. Apriori er enhver verdi av kronsannsynligheten p tenkelig. Anta at vi bruker en apriorifordeling av den type som er beskrevet i Eksempel 5. La X være antall kron i n kast. Finn Bayes-estimatet for p i følgende situasjoner:

(a) $r = s = 1$, $n = 10$, $X = 4$

(b) $r = s = 5$, $n = 10$, $X = 4$

(c) $r = s = 1$, $n = 100$, $X = 40$

(d) $r = s = 5$, $n = 100$, $X = 40$

(e) $r = 4$, $s = 6$, $n = 10$, $X = 5$

15. For en gitt produksjonsserie antas strekkstyrken X til et tilfeldig tau å være normalfordelt $N(\mu, \sigma^2)$. Kvalitetsnivået μ kan variere fra en produksjonsserie til en annen og nivået til en tilfeldig serie antas apriori å være normalfordelt $N(\mu_0, \sigma_0^2)$, der $\mu_0 = 11.0$ og $\sigma_0 = 1.0$. Av en ny produksjonsserie velges n=4 tau som belastes, deres strekkstyrker ble

$$X_i: \quad 10.6 \quad 9.7 \quad 10.1 \quad 10.8$$

(a) Gi et Bayes-estimat for kvalitetsnivået μ for dette partiet dersom $\sigma = 0.4$.

(b) Finn apriorisannsynligheten og aposteriorisannsynligheten for at et tilfeldig valgt tau fra dette partiet tåler en belastning på 10.0 kr.

Besvar også oppgaven dersom σ isteden er 1.0.

Appendiks A

Terminologi

A.1 Indekser og summasjoner

Gitt en liste bestående av n tall

$$x_1, x_2, \ldots, x_n$$

slik at x_i er det i'te tallet på listen. Her kaller vi i en indeks eller fotskrift. Summen av de n tallene skriver vi

$$x_1 + x_2 + \cdots + x_n$$

Det er vanlig å la symbolet Σ bety "summen av". Vi har derfor følgende alternative skrivemåte

$$\sum_{i=1}^{n} x_i$$

som leses "summen av alle x_i'er fra $i = 1$ til $i = n$".

I denne forbindelse blir indeksen i også kalt summasjonsindeksen. Det er likegyldig hvilken bokstav som brukes, mest vanlig er likevel i, j, k, l, m og n.Vi må imidlertid passe på at vi ikke velger en summasjonsindeks som også er brukt i en annen betydning.

Symbolbruken varieres etter behov, eksempelvis

$$\sum_{i=3}^{n} x_i \quad \text{betyr} \quad \text{“summen av alle } x_i\text{'er fra } i = 3 \text{ til } i = n\text{”,}$$

$$\sum_{i=1}^{n-2} x_i \quad \text{betyr} \quad \text{“summen av alle } x_i\text{'er fra } i = 3 \text{ til } i = n - 2\text{”,}$$

$$\sum_{j=3}^{5} x_j \quad \text{betyr} \quad \text{“summen av alle } x_j\text{'ene for } j = 3, 4 \text{ og } 5\text{”},$$

$$\sum_{n=k}^{m} x_n \quad \text{betyr} \quad \text{“summen av alle } x_n \text{ fra } n = k \text{ til } n = m\text{”}.$$

I noen situasjoner vil de tall som summeres ikke nødvendigvis komme etter hverandre, vi kan da skrive

$$\sum_{i \in A} x_i$$

som leses “summen av de x_i'er med indeks som er med i mengden A”(om mendeterminologi se nedenfor). Eksempelvis dersom $A = \{1, 4, 5\}$ blir $\sum_{i \in A} x_i = x_1 + x_4 + x_5$.

Av og til går det fram av sammenhengen hvilke indekser det skal summeres over, og vi tillater oss da å skrive

$$\sum_{i} x_i$$

Symbolbruken kan også varieres på andre måter etter behov, eksempelvis

$$\sum_{i=1}^{n} 2x_i = 2x_1 + 2x_2 + \cdots + 2x_n$$

$$\sum_{i=1}^{n} (x_i + 2) = (x_1 + 2) + (x_2 + 2) + \cdots + (x_n + 2)$$

$$\sum_{i=1}^{n} x_i^2 = x_1^2 + x_2^2 + \cdots + x_n^2$$

$$\sum_{i=1}^{n} f(x_i) = f(x_1) + f(x_2) + \cdots + f(x_n)$$

$$\sum_{i=1}^{n} i x_i = 1 \cdot x_1 + 2 \cdot x_2 + \cdots + n \cdot x_n$$

$$\sum_{i=1}^{n} x_i f(x_i) = x_1 \cdot f(x_1) + x_2 \cdot f(x_2) + \cdots + x_n \cdot f(x_n)$$

$$\sum_{i=1}^{n} (x_i + y_i) = (x_1 + y_1) + (x_2 + y_2) + \cdots + (x_n + y_n)$$

Følgende regneregler er verd å merke seg

$$(a) \quad \sum_i kx_i = k\sum_i x_i \quad \text{(k konstant)}$$

$$(b) \quad \sum_i (x_i + y_i) = \sum_i x_i + \sum_i y_i$$

som gjelder uansett hvilke indekser det summeres over, bare det er de samme på begge sider av likhetstegnet. Det også verd å merke seg at

$$(c) \quad \sum_i a = a + a + \cdots + a = n \cdot a$$

der n er antall indekser det summeres over.

Disse tre regnereglene kan brukes til å forenkle mer kompliserte summeringer.

I noen situasjoner kan utgangspunktet være en uendelig tallfølge $x_1, x_2, x_3, \ldots$. For den uendelige summen

$$x_1 + x_2 + x_3 + \cdots$$

brukes vanligvis den forkortede skrivemåten

$$\sum_{i=1}^{\infty} x_i$$

Lesere med kunnskaper i matematisk analyse vil vite at slike summer ikke nødvendigvis gir mening som noe endelig tall.

I enkelte sammenhenger er det aktuelt å summere tallene i et rektangulært skjema

$$\begin{array}{cccc} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & & \vdots \\ x_{m1} & x_{m2} & \cdots & x_{mn} \end{array}$$

slik at x_{ij} betegner tallet i i'te linje og j'te søyle. Summen av alle tallene skrives

$$\sum_{i,j} x_{ij}$$

der vi summerer over alle indekspar (i, j). Alternative skrivemåter er

$$\sum_{i=1}^{m}\sum_{j=1}^{n} x_{ij} = \sum_{j=1}^{n}\sum_{i=1}^{m} x_{ij}$$

I uttrykket til venstre skjer summeringen ved først å summere over j for hver i, og deretter summere over i, dvs. vi summerer hver linje og summerer deretter linjesummene. I uttrykket til høyre skjer summeringen i omvendt rekkefølge, først summeres hver søyle og deretter summeres søylesummene. Begge framgangsmåter må gi samme resultat, og hvilken som brukes vil avhenge av omstendighetene.

Symbolbruken kan også her varieres etter behov, og regneregler kan etableres. Vi vil ikke gå i detalj, men nøye oss med et interessant resultat (la $x_{ij} = a_i \cdot b_i$)

$$(d) \qquad \sum_{i=1}^{m}\sum_{j=1}^{n} a_i \cdot b_j = (\sum_{i=1}^{m} a_i) \cdot (\sum_{j=1}^{n} b_j)$$

Summasjonsindeksen behøver ikke alltid være en fotskrift, i teksten forekommer uttrykk av form

$$\sum_{x} f(x) \qquad \sum_{x,y} f(x,y)$$

eventuelt med nærmere angivelse av hvilke x (eventuelt x og y) det skal summeres over.

A.2 Mengder og funksjoner

Med en *mengde* forstås en samling av objekter, heretter kalt *elementer*.

Vi vil benytte store bokstaver, som oftest i begynnelsen av alfabetet, til å symbolisere mengder. For elementer i mengden brukes den symbolikk som ansees som mest velegnet for det aktuelle formål. Eksempler på mengder er:

A = de ensifrede tall	= {0,1,2,... , 9}
B = de to-sifrede tall	= {00,01,02,... , 98,99}
C = de norske bokstaver	= {a,b,c,... , ø,å}
D = tegnene på en skrivemaskin	= {a,A,b,B,... ,?,!}
E = elevene i en klasse	= {NN,JL,... ,BB}
N = de naturlige tall	= {0,1,2,3,... }
Z = de hele tall	= {... ,-3,-2,-1,0,1,2,3, ... }

Klammen {} leses som “mengden av elementene”, og innenfor klammene er en beskrivelse eller eventuelt en oppramsing av disse elementene. Rekkefølgen av elementene i en slik oppramsing er likegyldig, eksempelvis er {1,2,3,} og {2,1,3} samme mengde. I de situasjoner der elementene i mengden har en naturlig rekkefølge er denne brukt for å lette oversikten. Skriver vi

$$M = \{e_1, e_2, \ldots, e_n\}$$

betyr dette mengden bestående av elementene $e_1, e_2, \ldots e_n$, rekkefølgen er likegyldig. I enkelte situasjoner er det hensiktsmessig å skrive

$$M = \{e; \text{“}e \text{ oppfyller betingelsen ...”}\}$$

som leses“mengden av alle elementer e slik at e oppfyller betingelsen ...”.

En mengde som består av et endelig antall elementer kalles en *endelig mengde*. Vi ser at mengdene A, B, C, D og E ovenfor er endelige mengder, mens N og Z er *uendelige mengder*. Andre eksempler på uendelige mengder er

$$\begin{aligned} \text{O} &= \text{alle odde tall} \\ \text{Q} &= \text{de rasjonale tall} \\ \text{R} &= \text{de reelle tall} \end{aligned}$$

En *tellbar* mengde (også kalt diskret mengde) er en mengde der elementene lar seg nummerere med naturlige tall, element nr.1, element nr.2 osv. Alle endelige mengder er tellbare, noen uendelige mengder er tellbare, andre ikke. N og O ovenfor er opplagt tellbare mengder, og det kan vises at også Z og Q er tellbare. R er derimot ikke tellbar, vi sier at de reelle tall er en *overtellbar* mengde (også kalt kontinuerlig mengde).

Symbolkombinasjonen

$$e \in M$$

leser vi “e er et element i mengden M”. Eksempelvis er $4 \in A$, $13 \in B$, $u \in C$, $? \in D$, $JL \in E$, $696 \in N$, $-4 \in Z$, $3/8 \in Q$, $\sqrt{2} \in R$.

Er to mengder A og B slik at ethvert element i A også er med i B sier vi at A er en *delmengde* (eller undermengde) av B og skriver $A \subset B$. Eksempelvis er $A \subset N, O \subset N, N \subset Q$ og $Q \subset R$.

Det er ofte bruk for å utføre regneoperasjoner på mengder. De mest vanlige regneoperasjonene er *union, snitt* og *komplement*. Det finnes en rekke regneregler som angår disse regneoperasjonene. Definisjonene og de regnereglene som er nødvendige for vårt formål er gitt i teksten.

> En regel som til ethvert element i en mengde M_1 tilordner et og bare et element i en annen mengde M_2 kaller vi en *funksjon* fra M_1 til M_2.

La f være en funksjon fra M_1 til M_2, symbolisert slik

$$f : M_1 \rightarrow M_2$$

Her kalles M_1 *definisjonsmengden* (definisjonsområdet) og M_2 *målmengden*. Når x er et vilkårlig element i M_1 vil $f(x) \in M_2$ være verdien av funksjonen f for elementet x. I denne forbindelse brukes også skrivemåten

$$f : x \rightarrow f(x)$$

som sier at f er den regel som til elementet x tilordner elementet $f(x)$.

Eksempler på funksjoner er
$M_1 = B$, $M_2 = B$ og $f =$ summen av sifrene
$M_1 = C$, $M_2 = N$ og $f =$ bokstavens nr. i alfabetet
$M_1 = E$, $M_2 = A$ og $f =$ elevens karakter på en prøve.

I mange anvendelser vil M_1 og M_2 være delmengder av de reelle tall. Eksempler på funksjoner er da

$$\begin{aligned} f&: \quad x \rightarrow x+1 \\ f&: \quad x \rightarrow x^2 \\ f&: \quad x \rightarrow x^2 - 3x + 2 \\ f&: \quad x \rightarrow logx \end{aligned}$$

Noen mengder kan hensiktsmessig beskrives ved en funksjon. Her er

$$\{x; f(x) = y\}$$

mengden av alle elementer x som tilordnes verdien y ved funksjonen f. Eksempelvis kan mengden av alle løsninger av annengradsligningen $x^2 - 3x + 2 = 0$ skrives

$$\{x; x^2 - 3x + 2 = 0\}$$

som for øvrig er lik {1,2}. Likeens betrakt

$$\{x; x^2 + 1 = 0\}$$

Siden ligningen $x^2 + 1 = 0$ ikke har noen reell løsning, er denne mengden uten elementer, vi sier at det er *den tomme mengde* symbolisert med $\emptyset$.

A.3 Det greske alfabet

Skrift	Stor	Liten	Skrift	Stor	Liten	Skrift	Stor	Liten
alfa	A	α	iota	I	ι	rho	P	ρ
beta	B	β	kappa	K	κ	sigma	Σ	σ
gamma	Γ	γ	lambda	Λ	λ	tau	T	τ
delta	Δ	δ	my	M	μ	ypsilon	Υ	υ
epsilon	E	ϵ	ny	N	ν	phi	Φ	ϕ
zeta	Z	ζ	ksi	Ξ	ξ	khi	X	χ
eta	H	η	omikron	O	o	psi	Ψ	ψ
theta	Θ	θ	pi	Π	π	omega	Ω	ω

A.4 Ordliste: Engelsk - Norsk

Autocorrelation	Autokorrelasjon
Average	Gjennomsnitt
Biased	Forventningsskjev
Binomial	Binomisk
Chi-square	Kjikvadrat
Cluster	Klynge
Confidence interval	Konfidensintervall
Contingency table	Kontigenstabell
Correlation	Korrelasjon
Covariance	Kovarians
Critical region	Forkastningsområde
Critical value	Kritisk verdi
Cumulative distribution	Kumulativ fordeling

Degrees of freedom	Frihetsgrader
Density function	Tetthetsfunksjon
Design	Forsøksplan
Disjoint	Disjunkt
Distribution	Fordeling
Error of first kind	Forkastingsfeil
Error of second kind	Godtakingsfeil
Estimate	Estimat, estimere
Estimator	Estimator
Event	Begivenhet, hendelse
Expectation	Forventning
Frequency	Hyppighet
Forecast	Prognose, prediksjon
Histogram	Histogram
Hypergeometric	Hypergeometrisk
Independence	Uavhengighet
Interaction	Samspill
Intersection	Snitt
Least squares method	Minste kvadraters metode
Mean (arithmetic)	Gjennomsnitt
Mean	Forventning
Mean square deviation	Empirisk varians, middelkvadratavvik
Median	Median
Moving average	Glidende gjennonsnitt
Normal	Normal
Operating characteristic	Karakteristikk
Outlier	Vill observasjon
Paired comparison	Parvis sammenligning
Parameter	Parameter
Population	Populasjon
Posterior probability	Aposteriori sannsynlighet
Power function	Styrkefunksjon
Predictor	Prediktor
Prior probability	Apriori sannsynlighet
Probability	Sannsynlighet
Probability measure	Sannsynlighetsmål
Quantile	Fraktil
Random	Stokastisk, tilfeldig
Random variable	Stokastisk variabel
Randomisation	Randomisering
Range	Variasjonsbredde
Rank	Rang, rangordne
Ratio estimate	Rateestimat
Regression	Regresjon
Relative frequency	Relativ hyppighet
Replacement	Tilbakelegging

Sample	Utvalg, velge ut
Sample space	Utfallsrom
Sampling	Utvelging
Sampling plan	Utvalgsplan
Seasonal variation	Sesongvariasjon
Serial correlation	Autokorrelasjon
Sign test	Tegn test
Significance level	Signifikansnivå
Significance probability	P-verdi, signifikanssannsynlighet
Significant	Signifikant
Simple event	Enkel begivenhet
Simulation	Simulering
Standard deviation	Standardavvik
Standard error	Standardavvik, standardfeil
Stationary	Stasjonær
Statistic	Observator
Statistics	Statistikk
Stochastic	Stokastisk, tilfeldig
Stratification	Stratifisering
Stratum	Stratum
Sure event	Sikker begivenhet
Survey	Undersøkelse
Test statistic	Testobservator
Tied observations	Sammenfallende observasjoner
Time series	Tidsrekke
Trend	Trend
Trial	Forsøk
Unbiased	Forventningsrett
Variance	Varians
Variance analysis	Variansanalyse
Variate	Stokastisk variabel

Av mer omfattende ordlister nevnes:

Marriott: A Dictionary of Statistical terms. 5.Ed.
Longman Scientific & Technical, Burnt Hill Harlow Essex, 1990.

Nordisk Statistisk Nomenklatur. Studentlitteratur, Lund, 1975.

Den første gir utfyllende forklaringer på engelske termer, den andre gir sammenhørende termer på engelsk, norsk og de andre nordiske språk.

Appendiks B

Formler

Aksiomer :

A1. $0 \leq P(u) \leq 1$ for alle utfall u

A2. $P(u_1) + P(u_2) + P(u_3) + \cdots = 1$

A3. $P(A) = \sum_{u \in A} P(u)$

A4. $P(u \mid B) = P(u)/P(B)$ for $u \in B$ (=0 ellers)

Egenskaper :

E1. $0 \leq P(A) \leq 1$ for alle begivenheter A

E2. $P(\Omega) = 1$

E3. $P(A \cup B) = P(A) + P(B)$ når $A \cap B = \emptyset$

E4. $P(\overline{A}) = 1 - P(A)$

E5. $P(\emptyset) = 0$

E6. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

E7. $P(A_1 \cup A_2 \cup \cdots \cup A_n) = P(A_1) + P(A_2) + \cdots + P(A_n)$ når $A_1, A_2 \cdots A_n$ er disjunkte begivenheter

E8. $P(A \mid B) = P(A \cap B)/P(B)$

E9. $P(A \cap B) = P(B) \cdot P(A \mid B)$

E10. $P(B \mid A) = P(B) \cdot P(A \mid B)/P(A)$ (Bayes lov)

E11. $P(A) = P(B_1) \cdot P(A \mid B_1) + P(B_2) \cdot P(A \mid B_2) + \cdots + P(B_r) \cdot P(A \mid B_r)$ når $\Omega = B_1 \cup B_2 \cup \cdots \cup B_r)$ er en disjunkt union

E12. $P(A \cap B) = P(A) \cdot P(B)$ når A og B er uavhengige

De Morgans lover :

M1. $\overline{A_1 \cup A_2 \cup \cdots \cup A_n} = \overline{A}_1 \cap \overline{A}_2 \cap \cdots \cap \overline{A}_n$

M2. $\overline{A_1 \cap A_2 \cap \cdots \cap A_n} = \overline{A}_1 \cup \overline{A}_2 \cup \cdots \cup \overline{A}_n$

Kombinatorikk :

K1. $(N)_s = N(N-1)(N-2)\cdots(N-s+1)$

K2. $\binom{N}{s} = \frac{(N)_s}{s!} = \frac{N!}{s!(N-s)!}$

K3. $N! = 1 \cdot 2 \cdot 3 \cdots N$

Forventning :

F1. $EX = \sum_u X(u)P(u)$

F2. $EX = \sum_x xP(X = x) = \sum_x xp(x)$

F3. $E(k) = k$

F4. $E(k + X) = k + EX$

F5. $E(kX) = kEX$

F6. $E\phi(X) = \sum_x \phi(x)p(x)$

F7. $E(X + Y) = EX + EY$

F8. $E(X_1 + X_2 + \cdots + X_n) = EX_1 + EX_2 + \cdots + EX_n$

F9. $E(k_0 + k_1X_1 + k_2X_2 + \cdots + k_nX_n)$
$$= k_0 + k_1EX_1 + k_2EX_2 + \cdots + k_nEX_n$$

F10. $E\phi(X, Y) = \sum_{(x,y)} \phi(x, y)p(x, y)$

F11. $E(X \cdot Y) = EX \cdot EY$ såframt X og Y er uavhengige.

F12. $EE(Y \mid X) = EY$

F13. $ES_n = n \cdot EX$

F14. $ES_N = EN \cdot EX$

Varians :

V1. $varX = E(X-\mu)^2$ der $\mu = EX$

V2. $varX = \sum_x (x-\mu)^2 p(x)$

V3. $varX = E(X^2) - (EX)^2$

V4. $varX = \sum_x x^2 p(x) - (EX)^2$

V5. $var(k) = 0$ (k konstant)

V6. $var(k+X) = varX$

V7. $var(kX) = k^2 varX$

V8. $var(X+Y) = varX + varY + 2cov(X,Y)$

V9. $var(X+Y) = varX + varY$ når X og Y er ukorrelerte

V10. $var(\sum_{i=1}^n X_i) = \sum_{i=1}^n var(X_i) + \sum_{i \neq j} cov(X_i, X_j)$

V11. $var(\sum_{i=1}^n X_i) = \sum_{i=1}^n var(X_i)$ når X_i'ene er ukorrelerte

V12. $varS_n = nvarX.$

V13. $varS_N = EN \cdot varX + varN \cdot (EX)^2.$

Kovarians :

C1. $cov(X,Y) = E((X-EX)\cdot(Y-EY))$

C2. $cov(X,Y) = E(X \cdot Y) - EX \cdot EY$

C3. $cov(X,Y) = \sum_{(x,y)} xy \cdot p(x,y) - EX \cdot EY$

C4. $p(x,y) = p_1(x) \cdot p_2(y)$ medfører $cov(X,Y) = 0$

C5. $\rho(X,Y) = \frac{cov(X,Y)}{\sqrt{varX} \cdot \sqrt{varY}}$

Normaltilnærming : $(\mu = EX,\, \sigma^2 = varX)$

G1. $G(-z) = 1 - G(z)$

G2. $P(X \leq x) \approx G(\frac{x-\mu}{\sigma})$

G3. $P(X > x) = 1 - P(X \leq x) \approx 1 - G(\frac{x-\mu}{\sigma})$

G4. $P(a < X \leq b) \approx G(\frac{b-\mu}{\sigma}) - G(\frac{a-\mu}{\sigma})$

G5. $P(X \leq x) \approx G(\frac{x+\frac{1}{2}-\mu}{\sigma})$ (heltallsvariabel)

G6. $P(\mid X - \mu \mid < k \cdot \sigma) \approx A(k)$

G7. $P(\mid X - \mu \mid < d) \approx A(d/\sigma)$

Tsjebysjeffs ulikhet :

T1. $P(\mid X - \mu \mid < k \cdot \sigma) \geq 1 - 1/k^2$

Sannsynlighetsfordelinger :

1. Binomisk fordeling (n, p) : $P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}$

   $EX = np, \quad varX = np(1-p)$

2. Hypergeometrisk fordeling (N, M, n) : $P(Y = y) = \frac{\binom{M}{y} \cdot \binom{N-M}{n-y}}{\binom{N}{n}}$

   $EY = n \cdot \frac{M}{N}, \quad varY = \frac{N-n}{N-1} \cdot n \cdot \frac{M}{N}(1 - \frac{M}{N})$

3. Poissonfordeling (λ) : $P(X = x) = \frac{\lambda^x}{x!} e^{-\lambda}$

   $EX = \lambda, \quad varX = \lambda$

4. Geometrisk fordeling (p) : $P(N = n) = (1-p)^{n-1} p$

   $EN = 1/p, \quad varN = (1-p)/p^2$

Empiriske mål :

1. Gjennomsnitt: $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$
2. Empirisk varians: $S_X^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$
3. Empirisk standardavvik: $S_X = \sqrt{S_X^2}$
4. Empirisk kovarians: $S_{XY} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}) \cdot (Y_i - \overline{Y})$
5. Empirisk korrelasjonskoeffisient: $R_{XY} = \frac{S_{XY}}{S_X \cdot S_Y}$

Målemodellen :

$X_1, X_2, \ldots, X_n$ uavhengige $EX_i = \mu$, $varX_i = \sigma^2$.

$$E\overline{X} = \mu, \quad var\overline{X} = \frac{\sigma^2}{n}$$

Lotterimodellen :

$Y_1, Y_2, \ldots, Y_n$ tilfeldig utvalg fra $v_1, v_2, \ldots, v_N$.

$$E\overline{Y} = \overline{v}, \quad var\overline{Y} = \frac{N-n}{N-1} \cdot \frac{\sigma^2}{n}$$

der $\quad \sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(v_i - \overline{v})^2$.

Appendiks C

Tabeller

N\s	2	3	4	5	6	7	8	9	10	11	12	13
2	1											
3	3	1										
4	6	4	1									
5	10	10	5	1								
6	15	20	15	6	1							
7	21	35	35	21	7	1						
8	28	56	70	56	28	8	1					
9	36	84	126	126	84	36	9	1				
10	45	120	210	252	210	120	45	10	1			
11	55	165	330	462	462	330	165	55	11	1		
12	66	220	495	792	924	792	495	220	66	12	1	
13	78	286	715	1 287	1 716	1 716	1 287	715	286	78	13	1
14	91	364	1 001	2 002	3 003	3 432	3 003	2 002	1 001	364	91	14
15	105	455	1 365	3 003	5 005	6 435	6 435	5 005	3 003	1 365	455	105
16	120	560	1 820	4 368	8 008	11 440	12 870	11 440	8 008	4 368	1 820	560
17	136	680	2 380	6 188	12 376	19 448	24 310	24 310	19 448	12 376	6 188	2 380
18	153	816	3 060	8 568	18 564	31 824	43 758	48 620	43 758	31 824	18 564	8 568
19	171	969	3 876	11 628	27 132	50 388	75 582	92 378	92 378	75 582	50 388	27 132
20	190	1 140	4 845	15 504	38 760	77 520	125 970	167 960	184 756	167 960	125 970	77 520
21	210	1 330	5 985	20 349	54 264	116 280	203 490	293 930	352 716	352 716	293 930	203 490
22	231	1 540	7 315	26 334	74 613	170 544	319 770	497 420	646 646	705 432	646 646	497 420
23	253	1 771	8 855	33 649	100 947	245 157	490 314	817 190	1 144 066	1 352 078	1 352 078	1 144 066
24	276	2 024	10 626	42 504	134 596	346 104	735 471	1 307 504	1 961 256	2 496 144	2 704 156	2 496 144
25	300	2 300	12 650	53 130	177 100	480 700	1 081 575	2 042 975	3 268 760	4 457 400	5 200 300	5 200 300
26	325	2 600	14 950	65 780	230 230	657 800	1 562 275	3 124 550	5 311 735	7 726 160	9 657 700	10 400 600

Tabellen gir $\binom{N}{s}$. Eksempel: $\binom{9}{3} = 84$, $\binom{24}{12} = 2\ 704\ 156$.

Tabell C.1: Binomiske koeffisienter

n	x	p = .05	p = .1	p = .2	p = .3	p = .4
2	0	.9025	.8100	.6400	.4900	.3600
	1	.0950	.1800	.3200	.4200	.4800
	2	.0025	.0100	.0400	.0900	.1600
3	0	.8574	.7290	.5120	.3430	.2160
	1	.1354	.2430	.3840	.4410	.4320
	2	.0071	.0270	.0960	.1890	.2880
	3	.0001	.0010	.0080	.0270	.0640
4	0	.8145	.6561	.4096	.2401	.1296
	1	.1715	.2916	.4096	.4116	.3456
	2	.0135	.0486	.1536	.2646	.3456
	3	.0005	.0036	.0256	.0756	.1536
	4		.0001	.0016	.0081	.0256
5	0	.7738	.5905	.3277	.1681	.0778
	1	.2036	.3280	.4096	.3602	.2592
	2	.0214	.0729	.2048	.3087	.3456
	3	.0011	.0081	.0512	.1323	.2304
	4		.0005	.0064	.0284	.0768
	5			.0003	.0024	.0102
6	0	.7351	.5314	.2621	.1176	.0467
	1	.2321	.3543	.3932	.3025	.1866
	2	.0305	.0984	.2458	.3241	.3110
	3	.0021	.0146	.0819	.1852	.2765
	4	.0001	.0012	.0154	.0595	.1382
	5		.0001	.0015	.0102	.0369
	6			.0001	.0007	.0041
7	0	.6983	.4783	.2097	.0824	.0280
	1	.2573	.3720	.3670	.2471	.1306
	2	.0406	.1240	.2753	.3176	.2613
	3	.0036	.0230	.1147	.2269	.2903
	4	.0002	.0026	.0287	.0972	.1935
	5		.0002	.0043	.0250	.0774
	6			.0004	.0036	.0172
	7				.0002	.0016
8	0	.6634	.4305	.1678	.0576	.0168
	1	.2793	.3826	.3355	.1977	.0896
	2	.0515	.1488	.2936	.2965	.2090
	3	.0054	.0331	.1468	.2541	.2787
	4	.0004	.0046	.0459	.1361	.2322
	5		.0004	.0092	.0467	.1239
	6			.0011	.0100	.0413
	7			.0001	.0012	.0079
	8				.0001	.0007
9	0	.6302	.3874	.1342	.0404	.0101
	1	.2985	.3874	.3020	.1556	.0605
	2	.0629	.1722	.3020	.2668	.1612
	3	.0077	.0446	.1762	.2668	.2508
	4	.0006	.0074	.0661	.1715	.2508
	5		.0008	.0165	.0735	.1672
	6		.0001	.0028	.0210	.0743
	7			.0003	.0039	.0212
	8				.0004	.0035
	9					.0003
10	0	.5987	.3487	.1074	.0282	.0060
	1	.3151	.3874	.2684	.1211	.0403
	2	.0746	.1937	.3020	.2335	.1209
	3	.0105	.0574	.2013	.2668	.2150
	4	.0010	.0112	.0881	.2001	.2508
	5	.0001	.0015	.0264	.1029	.2007
	6		.0001	.0055	.0368	.1115
	7			.0008	.0090	.0425
	8			.0001	.0014	.0106
	9				.0001	.0016
	10					.0001

Tabellen gir $P(X = x)$ når X er binomisk fordelt (n,p). Eksempel: $n = 8$, $p = 0.1$ gir $P(X = 2) = 0.1488$.

Tabell C.2: Binomisk fordeling

n	x	p = .5
2	0	.2500
	1	.5000
3	0	.1250
	1	.3750
4	0	.0625
	1	.2500
	2	.3750
5	0	.0312
	1	.1562
	2	.3125
6	0	.0156
	1	.0938
	2	.2344
	3	.3125
7	0	.0078
	1	.0547
	2	.1641
	3	.2734
8	0	.0039
	1	.0312
	2	.1094
	3	.2188
	4	.2734
9	0	.0020
	1	.0176
	2	.0703
	3	.1641
	4	.2461
10	0	.0010
	1	.0098
	2	.0439
	3	.1172
	4	.2051
	5	.2461
11	0	.0005
	1	.0054
	2	.0269
	3	.0806
	4	.1611
	5	.2256
12	0	.0002
	1	.0029
	2	.0161
	3	.0537
	4	.1208
	5	.1934
	6	.2256
13	0	.0001
	1	.0016
	2	.0095
	3	.0349
	4	.0873
	5	.1571
	6	.2095
14	0	.0001
	1	.0009
	2	.0056
	3	.0222
	4	.0611
	5	.1222
	6	.1833
	7	.2095
15	0	.0000
	1	.0005
	2	.0032
	3	.0139
	4	.0417
	5	.0916
	6	.1527
	7	.1964
16	0	.0000
	1	.0002
	2	.0018
	3	.0085
	4	.0278
	5	.0667
	6	.1222
	7	.1746
	8	.1964
17	0	.0000
	1	.0001
	2	.0010
	3	.0052
	4	.0182
	5	.0472
	6	.0944
	7	.1484
	8	.1855
18	0	.0000
	1	.0001
	2	.0006
	3	.0031
	4	.0117
	5	.0327
	6	.0708
	7	.1214
	8	.1669
	9	.1855
19	1	.0000
	2	.0003
	3	.0018
	4	.0074
	5	.0222
	6	.0518
	7	.0961
	8	.1442
	9	.1762
20	1	.0000
	2	.0002
	3	.0011
	4	.0046
	5	.0148
	6	.0370
	7	.0739
	8	.1201
	9	.1602
	10	.1762
21	1	.0000
	2	.0001
	3	.0006
	4	.0029
	5	.0097
	6	.0259
	7	.0554
	8	.0970
	9	.1402
	10	.1682
22	1	.0000
	2	.0001
	3	.0004
	4	.0017
	5	.0063
	6	.0178
	7	.0407
	8	.0762
	9	.1186
	10	.1542
	11	.1682
23	2	.0000
	3	.0002
	4	.0011
	5	.0040
	6	.0120
	7	.0292
	8	.0584
	9	.0974
	10	.1364
	11	.1612
24	2	.0000
	3	.0001
	4	.0006
	5	.0025
	6	.0080
	7	.0206
	8	.0438
	9	.0779
	10	.1169
	11	.1488
	12	.1612
25	2	.0000
	3	.0001
	4	.0004
	5	.0016
	6	.0053
	7	.0143
	8	.0322
	9	.0609
	10	.0974
	11	.1328
	12	.1550
26	3	.0000
	4	.0002
	5	.0010
	6	.0034
	7	.0098
	8	.0233
	9	.0466
	10	.0792
	11	.1151
	12	.1439
	13	.1550
27	3	.0000
	4	.0001
	5	.0006
	6	.0022
	7	.0066
	8	.0165
	9	.0349
	10	.0629
	11	.0971
	12	.1295
	13	.1494
28	3	.0000
	4	.0001
	5	.0004
	6	.0014
	7	.0044
	8	.0116
	9	.0257
	10	.0489
	11	.0800
	12	.1133
	13	.1395
	14	.1494
29	4	.0000
	5	.0002
	6	.0009
	7	.0029
	8	.0080
	9	.0187
	10	.0373
	11	.0644
	12	.0967
	13	.1264
	14	.1445
30	4	.0000
	5	.0001
	6	.0006
	7	.0019
	8	.0055
	9	.0133
	10	.0280
	11	.0509
	12	.0806
	13	.1115
	14	.1354
	15	.1445

Tabellen gir $P(X = x)$ der X er binomisk fordelt $(n, p = 0.5)$. Eksempel: $n = 20$ gir $P(X = 8) = 0.1201$.

Tabell C.3: Binomisk fordeling ($p = 0.5$)

x\λ	.1	.2	.3	.4	.5	.6	.7	.8	.9	1.0
0	.9048	.8187	.7408	.6703	.6065	.5488	.4966	.4493	.4066	.3679
1	.0905	.1637	.2222	.2681	.3033	.3293	.3476	.3595	.3659	.3679
2	.0045	.0164	.0333	.0536	.0758	.0988	.1217	.1438	.1647	.1839
3	.0002	.0011	.0033	.0072	.0126	.0198	.0284	.0383	.0494	.0613
4		.0001	.0003	.0007	.0016	.0030	.0050	.0077	.0111	.0153
5				.0001	.0002	.0004	.0007	.0012	.0020	.0031
6							.0001	.0002	.0003	.0005
7										.0001

x\λ	1	2	3	4	5	6	7	8	9	10
0	.3679	.1353	.0498	.0183	.0067	.0025	.0009	.0003	.0001	.0000
1	.3679	.2707	.1494	.0733	.0337	.0149	.0064	.0027	.0011	.0005
2	.1839	.2707	.2240	.1465	.0842	.0446	.0223	.0107	.0050	.0023
3	.0613	.1804	.2240	.1954	.1404	.0892	.0521	.0286	.0150	.0076
4	.0153	.0902	.1680	.1954	.1755	.1339	.0912	.0572	.0337	.0189
5	.0031	.0361	.1008	.1563	.1755	.1606	.1277	.0916	.0607	.0378
6	.0005	.0120	.0504	.1042	.1462	.1606	.1490	.1221	.0911	.0631
7	.0001	.0034	.0216	.0595	.1044	.1377	.1490	.1396	.1171	.0901
8		.0009	.0081	.0298	.0653	.1033	.1304	.1396	.1318	.1126
9		.0002	.0027	.0132	.0363	.0688	.1014	.1241	.1318	.1251
10			.0008	.0053	.0181	.0413	.0710	.0993	.1186	.1251
11			.0002	.0019	.0082	.0225	.0452	.0722	.0970	.1137
12			.0001	.0006	.0034	.0113	.0264	.0481	.0728	.0948
13				.0002	.0013	.0052	.0142	.0296	.0504	.0729
14				.0001	.0005	.0022	.0071	.0169	.0324	.0521
15					.0002	.0009	.0033	.0090	.0194	.0347
16						.0003	.0014	.0045	.0109	.0217
17						.0001	.0006	.0021	.0058	.0128
18							.0002	.0009	.0029	.0071
19							.0001	.0004	.0014	.0037
20								.0002	.0006	.0019
21								.0001	.0003	.0009
22									.0001	.0004
23										.0002
24										.0001

Tabellen gir $P(X = x)$ der X er Poissonfordelt med forventning λ. Eksempel: $\lambda = 5$ gir $P(X = 4) = 0.1755$.

Tabell C.4: Poissonfordelingen

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990
3.1	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993
3.2	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995
3.3	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997
3.4	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998
3.5	.9998	.9998	.9998	.9998	.9998	.9998	.9998	.9998	.9998	.9998
3.6	.9998	.9998	.9999	.9999	.9999	.9999	.9999	.9999	.9999	.9999
3.7	.9999	.9999	.9999	.9999	.9999	.9999	.9999	.9999	.9999	.9999
3.8	.9999	.9999	.9999	.9999	.9999	.9999	.9999	.9999	.9999	.9999
3.9	1.0000	—	—	—	—	—	—	—	—	—

Tabellen gir arealet G(z) under normalkurven til venstre for z. Eksempel: z = 1.54 gir G(z) = 0.9382.

Tabell C.5: Normalfordelingen (arealtabell)

a	z	a	z	a	z
0.50	0.0000	0.70	0.5244	0.90	1.2816
.51	0.0251	.71	0.5534	.91	1.3408
.52	0.0502	.72	0.5828	.92	1.4051
.53	0.0753	.73	0.6128	.93	1.4758
.54	0.1004	.74	0.6434	.94	1.5548
.55	0.1257	.75	0.6745	.950	1.6449
.56	0.1510	.76	0.7063	.955	1.6954
.57	0.1764	.77	0.7389	.960	1.7507
.58	0.2019	.78	0.7722	.965	1.8119
.59	0.2275	.79	0.8064	.970	1.8808
.60	0.2534	.80	0.8416	.975	1.9600
.61	0.2793	.81	0.8779	.980	2.0538
.62	0.3055	.82	0.9154	.985	2.1701
.63	0.3319	.83	0.9542	.990	2.3264
.64	0.3585	.84	0.9945	.995	2.5758
.65	0.3853	.85	1.0364	.999	3.0902
.66	0.4125	.86	1.0803	.9995	3.2905
.67	0.4399	.87	1.1264	.9999	3.7190
.68	0.4677	.88	1.1750		
.69	0.4959	.89	1.2265		

Tabellen gir z slik at arealet til venstre for z under normalkurven er lik a, dvs. a = G(z). Eksempel: a = 0.955 gir z = 1.6954.

Tabell C.6: Normalkurven (fraktiltabell)

$q \backslash \nu$	1	2	3	4	5	$q \backslash \nu$	1	2	3	4	5
0.1	.752	.951	.992	.999	.999	6.2	.013	.045	.102	.185	.287
0.2	.655	.905	.978	.995	.999	6.4	.011	.041	.094	.171	.269
0.3	.584	.861	.960	.990	.998	6.6	.010	.037	.086	.159	.252
0.4	.527	.819	.940	.983	.995	6.8	.009	.033	.079	.147	.236
0.5	.480	.779	.919	.974	.992	7.0	.008	.030	.072	.136	.221
0.6	.439	.741	.896	.963	.988	7.2	.007	.027	.066	.126	.206
0.7	.403	.705	.873	.951	.983	7.4	.006	.025	.060	.116	.192
0.8	.371	.670	.850	.939	.977	7.6	.006	.022	.055	.107	.180
0.9	.343	.638	.825	.925	.970	7.8	.005	.020	.050	.099	.168
1.0	.317	.607	.801	.910	.963	8.0	.005	.018	.046	.092	.156
1.1	.294	.577	.777	.894	.954	8.2	.004	.017	.042	.085	.146
1.2	.273	.549	.753	.878	.945	8.4	.004	.015	.038	.078	.136
1.3	.254	.522	.729	.861	.935	8.6	.003	.014	.035	.072	.126
1.4	.237	.497	.706	.844	.924	8.8	.003	.012	.032	.066	.117
1.5	.221	.472	.682	.827	.913	9.0	.003	.011	.029	.061	.109
1.6	.206	.449	.659	.809	.901	9.2	.002	.010	.027	.056	.101
1.7	.192	.427	.637	.791	.889	9.4	.002	.009	.024	.052	.094
1.8	.180	.407	.615	.773	.876	9.6	.002	.008	.022	.048	.087
1.9	.168	.387	.593	.754	.863	9.8	.002	.007	.020	.044	.081
2.0	.157	.368	.572	.736	.849	10.0	.002	.007	.019	.040	.075
2.2	.138	.333	.532	.699	.821	10.5	.001	.005	.015	.033	.062
2.4	.121	.301	.494	.663	.792	11.0	.001	.004	.012	.027	.051
2.6	.107	.273	.458	.627	.761	11.5	.001	.003	.009	.022	.042
2.8	.094	.247	.424	.592	.731	12.0	.001	.003	.007	.017	.035
3.0	.083	.223	.392	.558	.700	12.5	.000	.002	.006	.014	.029
3.2	.074	.202	.362	.525	.669	13.0	.000	.002	.005	.011	.023
3.4	.065	.183	.334	.493	.639	13.5	.000	.001	.004	.009	.019
3.6	.058	.165	.308	.463	.608	14.0	.000	.001	.003	.007	.016
3.8	.051	.150	.284	.434	.579	14.5	.000	.001	.002	.006	.013
4.0	.046	.135	.262	.406	.549	15.0	.000	.000	.002	.005	.010
4.2	.040	.123	.241	.380	.521	15.5	.000	.000	.001	.004	.008
4.4	.036	.111	.221	.355	.493	16.0	.000	.000	.001	.003	.007
4.6	.032	.100	.204	.331	.467	16.5	.000	.000	.001	.002	.006
4.8	.028	.091	.187	.308	.441	17.0	.000	.000	.001	.002	.005
5.0	.025	.082	.172	.287	.416	17.5	.000	.000	.001	.002	.004
5.2	.023	.074	.158	.267	.392	18.0	.000	.000	.000	.001	.003
5.4	.020	.067	.145	.249	.369	18.5	.000	.000	.000	.001	.002
5.6	.018	.061	.133	.231	.347	19.0	.000	.000	.000	.001	.002
5.8	.016	.055	.122	.215	.326	19.5	.000	.000	.000	.001	.002
6.0	.014	.050	.112	.199	.306	20.0	.000	.000	.000	.001	.001

Tabellen gir arealet a til høyre for q under kjikvadratkurven med ν frihetsgrader ($\nu = 1, 2, 3, 4, 5$). Eksempel: $\nu = 4$, $q = 5$ gir $a = 0.287$.

Tabell C.7: Kjikvadratkurver (arealtabell)

$\nu \backslash$ a	.995	.99	.975	.95	.90	.75	.50	.25	.10	.05	.025	.01	.005
1	0.00	0.00	0.00	0.00	0.02	0.10	0.45	1.32	2.71	3.84	5.02	6.63	7.88
2	0.01	0.02	0.05	0.10	0.21	0.57	1.39	2.77	4.61	5.99	7.38	9.21	10.6
3	0.07	0.11	0.22	0.35	0.58	1.21	2.37	4.11	6.25	7.81	9.35	11.3	12.8
4	0.21	0.30	0.48	0.71	1.06	1.92	3.36	5.39	7.78	9.49	11.1	13.3	14.9
5	0.41	0.55	0.83	1.15	1.61	2.67	4.35	6.63	9.24	11.1	12.8	15.1	16.7
6	0.68	0.87	1.24	1.64	2.20	3.45	5.35	7.84	10.6	12.6	14.4	16.8	18.5
7	0.99	1.24	1.69	2.17	2.83	4.25	6.35	9.04	12.0	14.1	16.0	18.5	20.3
8	1.34	1.65	2.18	2.73	3.49	5.07	7.34	10.2	13.4	15.5	17.5	20.1	22.0
9	1.73	2.09	2.70	3.33	4.17	5.90	8.34	11.4	14.7	16.9	19.0	21.7	23.6
10	2.16	2.56	3.25	3.94	4.87	6.73	9.34	12.5	16.0	18.3	20.5	23.2	25.2
11	2.60	3.05	3.82	4.57	5.58	7.58	10.3	13.7	17.3	19.7	21.9	24.7	26.8
12	3.07	3.57	4.40	5.23	6.30	8.44	11.3	14.8	18.5	21.0	23.3	26.2	28.3
13	3.57	4.11	5.01	5.89	7.04	9.30	12.3	16.0	19.8	22.4	24.7	27.7	29.8
14	4.07	4.66	5.63	6.57	7.79	10.2	13.3	17.1	21.1	23.7	26.1	29.1	31.3
15	4.60	5.23	6.26	7.26	8.55	11.0	14.3	18.2	22.3	25.0	27.5	30.6	32.8
16	5.14	5.81	6.91	7.96	9.31	11.9	15.3	19.4	23.5	26.3	28.8	32.0	34.3
17	5.70	6.41	7.56	8.67	10.1	12.8	16.3	20.5	24.8	27.6	30.2	33.4	35.7
18	6.26	7.01	8.23	9.37	10.9	13.7	17.3	21.6	26.0	28.9	31.5	34.8	37.2
19	6.84	7.63	8.91	10.1	11.7	14.6	18.3	22.7	27.2	30.1	32.9	36.2	38.6
20	7.43	8.26	9.59	10.9	12.4	15.5	19.3	23.8	28.4	31.4	34.2	37.6	40.0
21	8.03	8.90	10.3	11.6	13.2	16.3	20.3	24.9	29.6	32.7	35.5	38.9	41.4
22	8.64	9.54	11.0	12.3	14.0	17.2	21.3	26.0	30.8	33.9	36.8	40.3	42.8
23	9.26	10.2	11.7	13.1	14.8	18.1	22.3	27.1	32.0	35.2	38.1	41.6	44.2
24	9.89	10.9	12.4	13.8	15.7	19.0	23.3	28.2	33.2	36.4	39.4	43.0	45.6
25	10.5	11.5	13.1	14.6	16.5	19.9	24.3	29.3	34.4	37.7	40.6	44.3	46.9
26	11.2	12.2	13.8	15.4	17.3	20.8	25.3	30.4	35.6	38.9	41.9	45.6	48.3
27	11.8	12.9	14.6	16.2	18.1	21.7	26.3	31.5	36.7	40.1	43.2	47.0	49.6
28	12.5	13.6	15.3	16.9	18.9	22.7	27.3	32.6	37.9	41.3	44.5	48.3	51.0
29	13.1	14.3	16.0	17.7	19.8	23.6	28.3	33.7	39.1	42.6	45.7	49.6	52.3
30	13.8	15.0	16.8	18.5	20.6	24.5	29.3	34.8	40.3	43.8	47.0	50.9	53.7
35	17.2	18.5	20.6	22.5	24.8	29.1	34.3	40.2	46.1	49.8	53.2	57.4	60.3
40	20.7	22.2	24.4	26.5	29.1	33.7	39.3	45.6	51.8	55.8	59.3	63.7	66.8
45	24.3	25.9	28.4	30.6	33.4	38.3	44.3	51.0	57.5	61.7	65.4	70.0	73.2
50	28.0	29.7	32.4	34.8	37.7	42.9	49.3	56.3	63.2	67.5	71.4	76.2	79.5
55	31.7	33.6	36.4	39.0	42.1	47.6	54.3	61.7	68.8	73.3	77.4	82.3	85.8
60	35.5	37.5	40.5	43.2	46.5	52.3	59.3	67.0	74.0	79.1	83.3	88.4	92.0
70	43.3	45.4	48.8	51.7	55.3	61.7	69.3	77.6	85.5	90.5	95.0	100.	104.
80	51.2	53.5	57.2	60.4	64.3	71.1	79.3	88.1	96.6	102.	107.	112.	116.
90	59.2	61.7	65.6	69.1	73.3	80.6	89.3	98.6	108.	113.	118.	124.	128.
100	67.3	70.0	74.2	77.9	82.3	90.1	99.3	109.	118.	124.	130.	136.	140.

Tabellen gir q slik at arealet til høyre for q under kjikvadratkurven med ν frihetsgrader er lik a. Eksempel: $\nu = 10$, a $= 0.05$ gir q $= 18.3$.

Tabell C.8: Kjikvadratkurver (fraktiltabell)

$\nu \backslash$ a	0.25	0.10	0.05	0.025	0.010	0.005	0.001	0.0005
1	1.000	3.078	6.314	12.71	31.82	63.66	318.3	636.6
2	0.816	1.886	2.920	4.303	6.965	9.925	22.33	31.60
3	0.765	1.638	2.353	3.182	4.541	5.841	10.21	12.92
4	0.741	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	0.727	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	0.718	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	0.711	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	0.706	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	0.703	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	0.700	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	0.697	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	0.695	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	0.694	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	0.692	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	0.691	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	0.690	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	0.689	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	0.688	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	0.688	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	0.687	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	0.686	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	0.686	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	0.685	1.319	1.714	2.069	2.500	2.807	3.485	3.768
24	0.685	1.318	1.711	2.064	2.492	2.797	2.467	3.745
25	0.684	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	0.684	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	0.684	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	0.683	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	0.683	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	0.683	1.310	1.697	2.042	2.457	2.750	3.385	3.646
35	0.682	1.306	1.690	2.030	2.438	2.724	3.340	3.591
40	0.681	1.303	1.684	2.021	2.423	2.704	3.307	3.551
45	0.680	1.301	1.679	2.014	2.412	2.690	3.281	3.520
50	0.679	1.299	1.676	2.009	2.403	2.678	3.261	3.496
55	0.679	1.297	1.673	2.004	2.396	2.668	3.245	3.476
60	0.679	1.296	1.671	2.000	2.390	2.660	3.232	3.460
70	0.678	1.294	1.667	1.994	2.381	2.648	3.211	3.435
80	0.678	1.292	1.664	1.990	2.374	2.639	3.195	3.416
90	0.677	1.291	1.662	1.987	2.638	2.632	3.183	3.402
100	0.677	1.290	1.660	1.984	2.364	2.626	3.174	3.390
—	0.675	1.282	1.645	1.960	2.326	2.576	3.090	3.290

Tabellen gir t slik at arealet til høyre for t under t-kurven med ν frihetsgrader er lik a. Eksempel: $\nu = 10$, a = 0.005 gir t = 3.169.

Tabell C.9: t-kurver (fraktiltabell)

$\nu_2 \backslash \nu_1$	1	2	3	4	5	6	8	10	12	15	20	24	30
1	161	200	216	225	230	234	239	242	244	246	248	249	250
2	18.5	19.0	19.2	19.2	19.3	19.3	19.4	19.4	19.4	19.4	19.4	19.5	19.5
3	10.1	9.55	9.28	9.12	9.01	8.94	8.85	8.79	8.79	8.70	8.66	8.64	8.62
4	7.71	6.94	6.59	6.39	6.26	6.16	6.04	5.96	5.91	5.86	5.80	5.77	5.75
5	6.61	5.79	5.41	5.19	5.05	4.95	4.82	4.74	4.68	4.62	4.56	4.53	4.50
6	5.99	5.14	4.76	4.53	4.39	4.28	4.15	4.06	4.00	3.94	3.87	3.84	3.81
7	5.59	4.74	4.35	4.12	3.97	3.87	3.73	3.64	3.57	3.51	3.44	3.41	3.38
8	5.32	4.46	4.07	3.84	3.69	3.58	3.44	3.53	3.28	3.22	3.15	3.12	3.08
9	5.12	4.26	3.86	3.63	3.48	3.37	3.23	3.14	3.07	3.01	2.94	2.90	2.86
10	4.96	4.10	3.71	3.48	3.33	3.22	3.07	2.98	2.91	2.85	2.77	2.74	2.70
11	4.84	3.98	3.59	3.36	3.20	3.09	2.95	2.85	2.79	2.72	2.65	2.61	2.57
12	4.75	3.89	3.49	3.26	3.11	3.00	2.85	2.75	2.69	2.62	2.54	2.51	2.47
13	4.67	3.81	3.41	3.18	3.03	2.92	2.77	2.67	2.60	2.53	2.46	2.42	2.38
14	4.60	3.74	3.34	3.11	2.96	2.85	2.70	2.60	2.53	2.46	2.39	2.35	2.31
15	4.54	3.68	3.29	3.06	2.90	2.79	2.64	2.54	2.48	2.40	2.33	2.29	2.25
16	4.49	3.63	3.24	3.01	2.85	2.74	2.59	2.49	2.42	2.35	2.28	2.24	2.19
17	4.45	3.59	3.20	2.96	2.81	2.70	2.55	2.45	2.38	2.31	2.23	2.19	2.15
18	4.41	3.55	3.16	2.93	2.77	2.66	2.51	2.41	2.34	2.27	2.19	2.15	2.11
19	4.38	3.52	3.13	2.90	2.74	2.63	2.48	2.38	2.31	2.23	2.16	2.11	2.07
20	4.35	3.49	3.10	2.87	2.71	2.60	2.45	2.35	2.28	2.20	2.12	2.08	2.04
21	4.32	3.47	3.07	2.84	2.68	2.57	2.42	2.32	2.25	2.18	2.10	2.05	2.01
22	4.30	3.44	3.05	2.82	2.66	2.55	2.40	2.30	2.23	2.15	2.07	2.03	1.98
23	4.28	3.42	3.03	2.80	2.64	2.53	2.37	2.27	2.20	2.13	2.05	2.01	1.96
24	4.26	3.40	3.01	2.78	2.62	2.51	2.36	2.25	2.18	2.11	2.03	1.98	1.94
25	4.24	3.39	2.99	2.76	2.60	2.49	2.34	2.24	2.16	2.09	2.01	1.96	1.92
30	4.17	3.32	2.92	2.69	2.53	2.42	2.27	2.16	2.09	2.01	1.93	1.89	1.84
40	4.08	3.23	2.84	2.61	2.45	2.34	2.18	2.08	2.00	1.92	1.84	1.79	1.74
60	4.00	3.15	2.76	2.53	2.37	2.25	2.10	1.99	1.92	1.84	1.75	1.70	1.65

Tabellen gir k slik at arealet til høyre for k under F-kurven med frihetsgradtall (ν_1, ν_2) er lik 0.05. Eksempel: $(\nu_1, \nu_2) = (3,5)$ gir k $= 5.41$.

Tabell C.10: F-kurver (Øvre 5%-fraktiler)

$\nu_2 \diagdown \nu_1$	1	2	3	4	5	6	8	10	12	15	20	24	30
1	4050	5000	5400	5620	5760	5860	5980	6060	6110	6160	6210	6230	6260
2	98.5	99.0	99.2	99.2	99.3	99.3	99.4	99.4	99.4	99.4	99.4	99.5	99.5
3	34.1	30.8	29.5	28.7	28.2	27.9	27.5	27.3	27.1	26.9	26.7	26.6	26.5
4	21.2	18.0	16.7	16.0	15.5	15.2	14.8	14.5	14.4	14.2	14.0	13.9	13.8
5	16.3	13.3	12.1	11.4	11.0	10.7	10.3	10.1	9.89	9.72	9.55	9.47	9.38
6	13.7	10.9	9.78	9.15	8.75	8.47	8.10	7.87	7.72	7.56	7.40	7.31	7.23
7	12.2	9.55	8.45	7.85	7.46	7.19	6.84	6.62	6.47	6.31	6.16	6.07	5.99
8	11.3	8.65	7.59	7.01	6.63	6.37	6.03	5.81	5.67	5.52	5.36	5.28	5.20
9	10.6	8.02	6.99	6.42	6.06	5.80	5.47	5.26	5.11	4.96	4.81	4.73	4.65
10	10.0	7.56	6.55	5.99	5.64	5.39	5.06	4.85	4.71	4.56	4.41	4.33	4.25
11	9.65	7.21	6.22	5.67	5.32	5.07	4.74	4.54	4.40	4.25	4.10	4.02	3.94
12	9.33	6.93	5.95	5.41	5.06	4.82	4.50	4.30	4.16	4.01	3.86	3.78	3.70
13	9.07	6.70	5.74	5.21	4.86	4.62	4.30	4.10	3.96	3.82	3.66	3.69	3.51
14	8.86	6.51	5.56	5.04	4.69	4.46	4.14	3.94	3.80	3.66	3.51	3.43	3.35
15	8.68	6.36	5.42	4.89	4.56	4.32	4.00	3.80	3.67	3.52	3.37	3.29	3.21
16	8.53	6.23	5.29	4.77	4.44	4.20	3.89	3.69	3.55	3.41	3.26	3.18	3.10
17	8.40	6.11	5.18	4.67	4.34	4.10	3.79	3.59	3.46	3.31	3.16	3.08	3.00
18	8.29	6.01	5.09	4.58	4.25	4.01	3.71	3.51	3.37	3.23	3.08	3.00	2.92
19	8.18	5.93	5.01	4.50	4.17	3.94	3.63	3.43	3.30	3.15	3.00	2.92	2.84
20	8.10	5.85	4.94	4.43	4.10	3.87	3.56	3.37	3.23	3.09	2.94	2.86	2.78
21	8.02	5.78	4.87	4.37	4.04	3.81	3.51	3.31	3.17	3.03	2.88	2.80	2.72
22	7.95	5.72	4.82	4.31	3.99	3.76	3.45	3.26	3.12	2.98	2.83	2.75	2.67
23	7.88	5.66	4.76	4.26	3.94	3.71	3.41	3.21	3.07	2.93	2.78	2.70	2.62
24	7.82	5.61	4.72	4.22	3.90	3.67	3.36	3.17	3.03	2.89	2.74	2.66	2.58
25	7.77	5.57	4.68	4.18	3.86	3.63	3.32	3.13	2.99	2.85	2.70	2.62	2.54
30	7.56	5.39	4.51	4.02	3.70	3.47	3.17	2.98	2.84	2.70	2.55	2.47	2.39
40	7.31	5.18	4.31	3.83	3.51	3.29	2.99	2.80	2.66	2.52	2.37	2.29	2.20
60	7.08	4.98	4.13	3.65	3.34	3.12	2.82	2.63	2.50	2.35	2.20	2.12	2.03

Tabellen gir k slik at arealet til høyre for k under F-kurven med frihetsgradtall (ν_1, ν_2) er lik 0.01. Eksempel: $(\nu_1, \nu_2) = (3,5)$ gir k = 12.1.

Tabell C.11: F-kurver (Øvre 1%-fraktiler)

Observert antall	konfidens-nivå 0.80		konfidens-nivå 0.90		konfidens-nivå 0.95		konfidens-nivå 0.99	
	lav	høy	lav	høy	lav	høy	lav	høy
0	0.00	2.30	0.00	3.00	0.00	3.69	0.00	5.30
	0.11	3.89	0.05	4.74	0.03	5 57	0.01	7.43
2	0.53	5.32	0.36	6.30	0.24	7.22	0.10	9.27
3	1.10	6.68	0.82	7.75	0.62	8.77	0.34	10.98
4	1.74	7.99	1.37	9.15	1.09	10.24	0.67	12.59
5	2.43	9.27	1.97	10.51	1.62	11.67	1.08	14.15
6	3 15	10.53	2.61	11.84	2.20	13.06	1.54	15.66
7	3.89	11.77	3.29	13.15	2.81	14.42	2.04	17.13
8	4.66	12.99	3.98	14.43	3.45	15.76	2.57	18.58
9	5.43	14.21	4.70	15.71	4.12	17.08	3.13	20.00
10	6.22	15.41	5.43	16.96	4.80	18.39	3.72	21.40
11	7.02	16.60	6.17	18.21	5.49	19.68	4.32	22.78
12	7.83	17.78	6.92	19.44	6.20	20.96	4.94	24.14
13	8.65	18.96	7.69	20.67	6.92	22.23	5.58	25.50
14	9.47	20.13	8.46	21.89	7.65	23.49	6.23	26.84
15	10.30	21.29	9.25	23.10	8.40	24.74	6.89	28.16
16	11.14	22.45	10.04	24.30	9.15	25.98	7.57	29.48
17	11.98	23 61	10.83	25.50	9.90	27.22	8.25	30.79
18	12.82	24.76	11.63	26.69	10.67	28.45	8.94	32 09
19	13.67	25.90	12.44	27.88	11.44	29.67	9.64	33.38
20	14.53	27.05	13.25	29.06	12.22	30.89	10 35	34.67
21	15.38	28.18	14.07	30.24	13.00	32.10	11.07	35.95
22	16.24	29 32	14.89	31.41	13.79	33.31	11.79	37.22
23	17.11	30.45	15.72	32.59	14.58	34.51	12.52	38.48
24	17.97	31.58	16.55	33.75	15.38	35.71	13.26	39.74
25	18.84	32.71	17.38	34.92	16.18	36.90	14.00	41.00
26	19.72	33.84	18.22	36.08	16.98	38.10	14.74	42.25
27	20.59	34.96	19 06	37.23	17.79	39.28	15.49	43.50
28	21.47	36.08	19.90	38.39	18.61	40.47	16.25	44.74
29	22.35	37.20	20.75	39.54	19.42	41.65	17.00	45.98
30	23.23	38.32	21.59	40.69	20.24	42.83	17.77	47.21
31	24.11	39.43	22.44	41.84	21.06	44.00	18.53	48.44
32	25.00	40.54	23.30	42.98	21.89	45.17	19.30	49.67
33	25.89	41.65	24.15	44.13	22.72	46.34	20.08	50.89
34	26.77	42.76	25.01	45.27	23.55	47.51	20.86	52.11
35	27.66	43.87	25.87	46.40	24.38	48.68	21.64	53.32
36	28.56	44.98	26.73	47.54	25.21	49.84	22.42	54.54
37	29.45	46.08	27.59	48.68	26.05	51.00	23.21	55.75
38	30.34	47.19	28.46	49.81	26.89	52.16	24.00	56.96
39	31.24	48.29	29.33	50.94	27.73	53.31	24.79	58.16
40	32.14	49.39	30.20	52.07	28.58	54.47	25.59	59.36

Eksempel: Med observert antall lik 9 er [4.12, 17.08] et konfidensintervall for forventet antall med konfidensnivå 95%.

Tabell C.12: Konfidensgrenser for Poissonforventning

	r = 1		r = 2		r = 3		r = 4		r = 5	
n	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U
12	0.97	1.33	0.81	1.58	0.66	1.86	0.51	2.18	0.38	2.51
13	1.01	1.34	0.86	1.56	0.72	1.82	0.57	2.09	0.45	2.39
14	1.05	1.35	0.91	1.55	0.77	1.78	0.63	2.03	0.51	2.30
15	1.08	1.36	0.95	1.54	0.81	1.75	0.69	1.98	0.56	2.22
16	1.10	1.37	0.98	1.54	0.86	1.73	0.74	1.93	0.62	2.16
17	1.13	1.38	1.02	1.54	0.90	1.71	0.78	1.90	0.67	2.10
18	1.16	1.39	1.05	1.53	0.93	1.69	0.82	1.87	0.71	2.06
19	1.18	1.40	1.08	1.53	0.97	1.68	0.86	1.85	0.75	2.02
20	1.20	1.41	1.10	1.54	1.00	1.68	0.90	1.83	0.79	1.99
21	1.22	1.42	1.13	1.54	1.03	1.67	0.93	1.81	0.83	1.96
22	1.24	1.43	1.15	1.54	1.05	1.66	0.96	1.80	0.86	1.94
23	1.26	1.44	1.17	1.54	1.08	1.66	0.99	1.79	0.90	1.92
24	1.27	1.45	1.19	1.55	1.10	1.66	1.01	1.78	0.93	1.90
25	1.29	1.45	1.21	1.55	1.12	1.66	1.04	1.77	0.95	1.89
26	1.30	1.46	1.22	1.55	1.14	1.65	1.06	1.76	0.98	1.88
27	1.32	1.47	1.24	1.56	1.16	1.65	1.08	1.76	1.01	1.86
28	1.33	1.48	1.26	1.56	1.18	1.65	1.10	1.75	1.03	1.85
29	1.34	1.48	1.27	1.56	1.20	1.65	1.12	1.74	1.05	1.84
30	1.35	1.49	1.28	1.57	1.21	1.65	1.14	1.74	1.07	1.83
35	1.40	1.52	1.34	1.58	1.28	1.65	1.22	1.73	1.16	1.80
40	1.44	1.54	1.39	1.60	1.34	1.66	1.29	1.72	1.23	1.79
45	1.48	1.57	1.43	1.62	1.38	1.67	1.34	1.72	1.29	1.78
50	1.50	1.59	1.46	1.63	1.42	1.67	1.38	1.72	1.34	1.77
55	1.53	1.60	1.49	1.64	1.45	1.68	1.41	1.72	1.38	1.77
60	1.55	1.62	1.51	1.65	1.48	1.69	1.44	1.73	1.41	1.77
65	1.57	1.63	1.54	1.66	1.50	1.70	1.47	1.73	1.44	1.77
70	1.58	1.64	1.55	1.67	1.52	1.70	1.49	1.74	1.46	1.77
75	1.60	1.65	1.57	1.68	1.54	1.71	1.51	1.74	1.49	1.77
80	1.61	1.66	1.59	1.69	1.56	1.72	1.53	1.74	1.51	1.77
85	1.62	1.67	1.60	1.70	1.57	1.72	1.55	1.75	1.52	1.77
90	1.63	1.68	1.61	1.70	1.59	1.73	1.57	1.75	1.54	1.78
95	1.64	1.69	1.62	1.71	1.60	1.73	1.58	1.75	1.56	1.78
100	1.65	1.69	1.63	1.72	1.61	1.74	1.59	1.76	1.57	1.78
150	1.72	1.75	1.71	1.76	1.69	1.77	1.68	1.79	1.67	1.80
200	1.76	1.78	1.75	1.79	1.74	1.80	1.73	1.81	1.72	1.82

Eksempel: $r = 2$ forklarende variable, $n = 20$ observasjoner gir grensene $d_L = 1.10$ og $d_U = 1.54$.

Tabell C.13: Kritiske verdier Durbin-Watson testen (5% nåvi)

Av mer omfattende tabellverker anbefales:

Odeh, Owen, Birnbaum & Fisher:
Pocket Book of Statistical Tables.
Marcel Dekker, Basel 1977.

Owen: Handbook of Statistical Tables.
Addison Wesley, Reading Mass. 1962.

Pearson & Hartley: Biometrika Tables for Statisticians.
Cambridge University Press, London 1962.

Appendiks D

Fasit til oppgaver

2. $\{G1, G2, G3, G4, G5, G6, P1, P2, P3, P4, P5, P6\}$.
3. $\{GG, GP, PG, PP\}$.
4. $\{0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12\}$.
5. $\{0, 1, 2, 3, 4, 5, \ldots\}$.
6. (a) $\{rh, rb, hr, hb, bh, br\}$. (b) $\{rh, rb, bh\}$.
7. $\{xxxx, xxxy, \ldots, yyyy\}$ ialt 16 mulige utfall.
8. $p(1) = 0.19 \quad p(2) = p(3) = p(4) = p(5) = 0.17 \quad p(6) = 0.13$.
9. $p(1) = p(6) = 0.20 \quad p(2) = p(3) = p(4) = p(5) = 0.15$.
10. $A \cup B = \{u_1, u_2, u_3, u_4, u_5\}$, $A \cap B = \emptyset$, $A \cup C = \{u_1, u_3, u_5, u_6\}$, $A \cap C = \{u_3\}$, $\overline{A} = \{u_2, u_4, u_6\}$. etc
11. (a) Ja. (b) $P(A) = 0.6$, $P(B) = 0.3$, $P(A \cap B) = 0$,
    (c) $P(\overline{A}) = 0.4$, $P(A \cup B) = 0.9$ etc.
12. $\{GGG, GGP, GPG, \ldots, PPP\}$, ialt 8 utfall.
    $P(A) = 7/8$, $P(B) = 1/2$, $P(C) = 1/4$.
13. (a) 1/3 (b) 1/3 (c) 1/2.
14. $\{KKKK, KKKM, \ldots, MMMM\}$, ialt 16 mulige utfall,
    $P(A) = 6/16$, $P(B) = 5/16$, $P(C) = 5/16$ etc.
17. (a) Nei (b) Nei.

18. $P(B) = 21/36$, $P(C) = 18/36$, $P(D) = 15/36$.

19. $P(E_i) = 1/6$, $P(F_j) = 1/6$.

20. 0.7.

21. (a) galt, (b) riktig, (c) galt.

22. $\{a, b, ab, i\}$ $\{SS \dots S, SS \dots F, \dots, FF \dots F\}$

23. (a) 25/216 (b) 27/216 (c) 35/216.

Kapittel 3.

1. 24.

2. $3^{20}, 3^{30}$.

3. (a) $2^{12} = 4096$ (b) 13 (c) 4^{12} (d) 3^{12} (e) 4^6 .

4. (a) $(10)_6$ (b) $\binom{10}{6}$.

5. $(10)_3 \cdot (12)_4$.

6. (a) 4! (b) 5! (c) 6!/2.

7. (a) 7! (b) 6! (c) 6!.

8. (a) $\binom{n}{2}$ (b) $(n)_2$.

9. (a) $\binom{22}{4}$ (b) $\binom{10}{4}$ (c) $\binom{12}{4}$ (d) $\binom{10}{2} \cdot \binom{12}{2}$.

10. (a) 2^{10} (b) $\binom{10}{5}$ (c) $\binom{10}{6}$.

11. (a) 3^{10} (b) 2^{10} (c) $\sum_{k=6}^{10} \binom{10}{k} 2^{10-k}$ (d) $\binom{10}{5} \cdot \binom{5}{3}$.

12. (a) 9/10 (b) $9^6/10^6$ (c) $8^6/10^6$ (d) $5^6/10^6$
    (e) $5000/10^6$ (f) $(10)_6/10^6$ (g) $56/10^6$.

13. (a) $\binom{10}{5}/2^{10}$ (b) $\sum_{k=6}^{10} \binom{10}{k} 2^{10}$ (c) $2 \cdot \sum_{k=7}^{10} \binom{10}{k} 2^{10}$

14. (a) 0.627 (b) $\binom{10}{5} \cdot \binom{5}{3}/3^{10}$.

15. (a) $5 \cdot (19)_4/(20)_5 = 5/20$ (b) $(5)_2 \cdot (18)_3\ /(20)_5 = 1/19$
    (c) $(8)_5/(20)_5$ (d) $8 \cdot (19)_4/(20)_5 = 8/20$
    (e) $(8)_2 \cdot (18)_3/(20)_5$ (f) 0.5687.

16. (a) $2 \cdot 5!/6! = 1/3$ (b) $2 \cdot 4!/6! = 1/15$
(c) $(4)_2 \cdot 4!/6! = 2/5$ (d) $10 \cdot 4!/6! = 1/3$.

17. (a) $12 \cdot 10!/12! = 1/11$ (b) $2 \cdot (6)_2 \cdot 10!/12! = 5/11$
(c) $20 \cdot 10!/12! = 5/33$ (d) $12 \cdot 10!/12! = 1/11$.

18. (a) $9/24$ (b) $8/24$ (c) $6/24$ (d) 0 (e) $1/24$.

19. (a) $\binom{8}{3}/\binom{12}{3} = 14/55$ (b) $\binom{4}{1} \cdot \binom{8}{2}/\binom{12}{3} = 28/55$
(c) $1 - 14/55 = 41/55$ (d) $1 - 1/55 = 54/55$.

20. (a) 3 (b) 6 (c) 7.

21. (a) $\binom{24}{4}/\binom{30}{4} = 0.39$ (b) $\binom{24}{3} \cdot \binom{6}{1}/\binom{30}{4} = 0.443$
(c) $1 - 0.39 = 0.61$.

22. (a) $\binom{10}{5} \cdot \binom{10}{5}/\binom{20}{10} = 0.344$ (b) $(1 - 0.344)/2 = 0.328$
(c) $\binom{5}{4} \cdot \binom{15}{6}/\binom{20}{10} + \binom{5}{5} \cdot \binom{15}{5}/\binom{20}{10} = 0.152$

23. (a) $\binom{17}{8}/\binom{20}{8} = 0.193$ (b) $\binom{15}{5}/\binom{20}{8} + \binom{5}{1} \cdot \binom{15}{7}/\binom{20}{8} = 0.307$
(c) 0.545 (urealistisk?).

24. (a) $\binom{16}{3}\binom{8}{0}\binom{4}{2}\binom{2}{0}/\binom{24}{3}\binom{6}{2} = 0.111$
(b) $1 - 0.111 = 0.889$ (c) 0.723 (d) 0.371.

25. (a) $3/4$ (b) $11/36$ (c) $7/16$ Nei.

26. (a) (i) $80/200 = 0.40$ (ii) $130/200 = 0.65$ (iii) $50/200 = 0.25$ (iv) $50/130 = 0.38$
(b) (i) $\binom{80}{2}/\binom{200}{2} = 0.1588$ (ii) $\binom{80}{1} \cdot \binom{120}{1}/\binom{200}{2} = 0.4824$
(iii) $\binom{80}{1} \cdot \binom{30}{1}/\binom{200}{2} = 0.1206$ (iv) $\binom{80}{2}/(\binom{200}{2} - \binom{80}{2}) = 0.1888$
Antakelse at uniform modell også gjelder betinget.

27. (a) $\binom{26}{13}/\binom{52}{13} = 0.0000164$ (b) $\binom{16}{4} \cdot \binom{36}{9}/\binom{52}{13} = 0.27$
(c) $\binom{36}{13}/\binom{52}{13} = 0.0036$ (d) $\binom{4}{1} \cdot \binom{48}{12}/\binom{52}{13} = 0.44$
(e) $\binom{4}{2} \cdot \binom{48}{11}/\binom{52}{13} = 0.21$

28. (a) $\binom{13}{1} \cdot \binom{4}{2} \cdot \binom{12}{3}\binom{4}{1}\binom{4}{1}\binom{4}{1}/\binom{52}{5} = 0.42256$
(b) 0.04754 (c) 0.02113 (d) 0.00024 (e) 0.00197.

29. (a) $4 \cdot (47)_3 \cdot 48!/52! = 0.0599$ (b) $1 - (44)_4 \cdot 48!/52! = 0.499$.

31. (e) 0.075 (f) 0.264 (g) 0.221 (h) 0.5l6.

32. (a) 0.162 (b) 0.0089.

33. sits: 4000: 0.010, 3100: 0.165, 2200: 0.135,
2110: 0.584, 1111: 0.106.

Kapittel 4.

1. $P(A \mid B) = 0$, $P(A \mid C) = 2/3$, $P(B \mid A) = 0$,$P(C \mid A) = 1/3$,
$P(A \mid \overline{B}) = 6/7$, $P(A \mid \overline{C}) = 4/7$, $P(C \mid \overline{A}) = 1/4$,
$P(A \mid A \cup B) = 2/3$, $P(A \mid A \cup C) = 6/7$, $P(A \mid A \cap B) =$ udef.
$P(A \mid A \cap C) = 1$, $P(A \mid A \cup \overline{C}) = 2/3$, $P(A \mid A \cap \overline{C}) = 1$,
$P(A \mid \overline{A} \cap C) = 0$, $P(A \mid A \cup B \cup C) = 3/5$.

2. $P(A \mid B) = 1$, $P(B \mid A) = 4/7$, $P(C \mid A) = 1/7$, $P(A \mid C) = 1/2$.

3. $P(B \mid C) = 2/3$, $P(C \mid B) = 4/7$, $P(B \mid D) = 3/5$, $P(D \mid B) = 3/7$.

5. $P(A \mid B) = 1/2$, $P(B \mid A) = 1/3$.

6. $P(B \mid C) = 17/63$.

7. (a) 0.327 (b) 0.00002 (c) 0.520 (d) 0.484 (e) 0.297 Nei.

11. 0.046.

12. 3/5, 4/9.

13. $P(M \cap K) = 0.36$, $P(M \mid K) = 0.72$, $P(M \cup K) = 0.54$.

14. $P(D) = 0.042$, $P(A \mid D) = 0.952$.

15. (a) 0.54 (b) 0.04 (c) 0.96.

16. (a) 16/25 (b) 2/3.

17. np/(np+(1-p)).

18. (a) 0.568 (b) 0.182 (c) 0.938.

20. (a) 1/3 (b) 1/2.

22. $p/(p + (1 - p)r)$.

23. Ja, Ja, Nei.

25. (a) 0.006 (b) 0.504 (c) 0.496 (d) 0.398.

26. (a) 0.133 (b) 0.380 (c) 0.620 (d) 0.323.

27. 7/8.

28. (a) 4/27 (b) l0/81 (c) 1/2.

29. (a) 0.53856 (b) 0.1188 (c) 0.1414 (d) 0.3064.

30. (a) $0.9^4 = 0.66$ (b) $0.9^2 \cdot 0.1 = 0.08$ (c) 0.10 (d) 0.29.

31. (a) 0.1574 (b) 0.5185, $P(D) = 2/27$, $P(E) = P(F) = 2/36$ Nei.

32. (a) 0.0778 (b) 0.2007 (c) 0.1074 (d) 0.2007.

33. h(K- gitt R-)=0.12, h(K+ gitt R-)=0.88,
h(K- gitt R+)=0.84, h(K+ gitt R+)=0.16.

34. Avhengig av tolkning: 0.6^3 eller 0.54^3 der $0.54 = 0.90 \cdot 0.60$.

35. $0.4 \cdot 0.25 + 0.6 \cdot 0.45 = 0.37$

36. Vanskelig å kombinere opplysningene uten ytterligere antakelser!

37. $p \cdot 0.30 + (1 - p) \cdot 0.60 < 0.5$ medfører $p > 1/3$

38. (a) 0.0027 (b) 0.1377 (c) 0.1707.

39. (a) 0.2128 (0.0420) (b) 0.4382 (0.0573) (c) 0.0028 (0.0835).

40. 0.0781.

42. (a) $(1/120)^{10}$ (b) $\binom{10}{3}(\frac{1}{120})^3(1 - \frac{1}{120})^7$ (c) $\binom{10}{5}(\frac{1}{2})^{10}$
(d) $\frac{10!}{2^5}(\frac{1}{5})^{10}$ (e) $1 - \frac{(120)_{10}}{120^{10}}$.

43. (a) 0.0670 (b) 0.5981 (c) 0.3349 (d) 0.0670
(e) 0.1962 (f) 0.2778 (g) 0.3871.

44. (a) $p^3 + 2p(1 - p)(1 - q) + q(1 - p)(1 - q)$.

Kapittel 5.

1.

x	0	1	2	3	4
P(X=x)	$\frac{2}{8}$	$\frac{2}{8}$	$\frac{2}{8}$	$\frac{1}{8}$	$\frac{1}{8}$

y	-2	-1	0	1	2
P(Y=y)	$\frac{1}{8}$	$\frac{2}{8}$	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{1}{8}$

2.

x	1	2	4
P(X=x)	$\frac{3}{8}$	$\frac{2}{8}$	$\frac{3}{8}$

Nei

3. 13/8, 1/8, 14/8, 12/8, 35/8, 13/8, 48/8.

6. EX = 3.35, varX = 2.83.

8.

x	0	1	2	3
P(X=x)	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$

EX = 3/2, varX = 3/4

9. EX = 2.

10. 1.0.

11.

x	0	1	2	3
P(X=x)	14/55	28/55	12/55	1/55

EX = 1.0.

12.

x	0	1	2	3	4
P(X=x)	0.388	0.443	0.151	0.018	0.0005

EX = 0.8.

13.

x	0	1	2	3
P(X=x)	0.504	0.398	0.092	0.006

EX = 0.6.

14. EX = 3.90, EY = 0.53.

15.

x	0	1	2
P(X=x)	0.324	0.519	0.157

EX = 0.833, EY = 1.167.

16. EX = 4.8, varX = 2.88 alt. EX = 7.2, varX = 2.88

17.

x	0	1	2	3
P(X=x)	5/30	15/30	9/30	1/30

EX = 1.2, varX = 0.56.

18.

x	0	1	2	3	4
P(X=x)	0.8145	0.1715	0.0135	0.0005	0.0000

EX = 0.2, varX =0.19.

19. EX = 14/5.

20. EX = 2.16.

21. Nei, forventet gevinst = 2/216, beløp 6.

22. Det er opp til deg. Nei.

23. Stopp når $n = 11$. Forventet gevinst = 8.88.

24. 80 kr.

25. $EX = (N+1)/2 \quad varX = (N+1)(N-1)/12.$

29. (b) Nei.

31. (a)

s	0	1	2	3	4	
P(S=s)	0.40	0.32	0.18	0.08	0.02	varS = 1.08
P(S=s)	0.33	0.38	0.25	0.04	0.00	varS = 0.94

(d) (i) 0 (ii) 0.20 (iii) −0.18.

32. (a) −0.421 (b) −0.658 (c) −0.041.

33. (i) 0.125, 0.072 (ii) 0.125, 0.087 (iii) 0.125, 0.053.

34. (b) 0.34 (c)

s	0	1	2	3	4
P(S=s)	0.15	0.32	0.34	0.16	0.03

(d) 0.44.

36. (a) Ja (men bare hvis X^2 er konstant). (b) Ja.

38. (b)
$$F(x) = \begin{cases} 0 & x < 0 \\ 1/4 & 0 \le x < 1 \\ 3/4 & 1 \le x < 2 \\ 1 & x \ge 2 \end{cases}$$

41. (a)

	s	0	1	2	3	4
(i)	P(S=s)	1/9	2/9	3/9	2/9	1/9
(ii)	P(S=s)	4/25	4/25	9/25	4/25	4/25
(iii)	P(S=s)	9/36	12/36	10/36	4/36	1/36

Kapittel 6.

1. (a)	(i)	0.0512	0.9421	0.0067
(b)	(i)	0.0055	0.9672	0.0001
(c)		0.1201	0.2517	0.0059

2. (a) 0, 1, 1, 2, 2(3) (b) 1, 2, 3, 4, 5(6).

3. (a) 0.0116 (b) 0.0702.

4. (a) 0.6778 (b) 0.3222 (c) 0.1209.

5. (b) n/10 (c) $10 \cdot (1-(9/10)^n)$.

6. (a)

y	0	1	2
P(Y=y)	28/45	16/45	1/45

EY=2/5

7. (a) 5/13, 5/4, 5/2 (b) 1, 13/4, 13/2, 4.

8. (a) 0.0226 (b) 0.0862 (c) 0.1709 (d) som (a).

11. (a) 0, 1(0), 5(4) (b) 0.9098, 0.7358, 0.0404,
(d) 0.0144, 0.0803, 0.8754.

12. 2.3, 0.0838

13. Ingen.

14. (a) 0.2642 (b) 0.5940 (c) 0.1569 (d) 0.3527
(e) 0.2117 (f) 0.6056.

16. (a) 0.9739 (Poisson $\approx$ 0.9735) (b) 0.0710 ($\approx$ 0.0706).

17. (b) 0.6065.

19. 0.2851, 0.6063.

21. (a) 0.8449, 0.2643, 0.8092, 0.0772
(b) 0.2358, 0.9842, 0.2956, 0.9074
(c) 0.2309, 0.5045, 0.6984.

22. (a) 0.62 (b) 1.63 (c) -0.65 (d) 0.92 (e) l.645 (f) 0.675.

25. (a) (i) 0.0134, 0.8286, 0.2045 (ii) 0.0533, 0.6266, 0.2531
(iii) 0.3463, 0.1178, 0.0937.
(b) (i) 0.0031, 0.9912, 0.1584 (ii) 0.0468, 0.9039, 0.3982
(iii) 0.7532, 0.1269, 0.0536.
(c) (i) 0.0107, 0.9981, 0.00005 (ii) 0.4594, 0.8206, 0.0262
(iii) 1.0000, 0.00005, 0.5454.

26. (a) 19-20, 25, 26 (b) 17, 14, 11 (c) 2-3, 5, 7-8.

27. 9604.

28. 0.0003 (normaltilnærming), 0.0016 (Poissontilnærming).

29. (a) 0.7907 (0.7919 normaltilnærming) (b) 0.3981 (0.4023)
(c) 0.0966 (0.0954).
30. 0.1311, 0.1294 (0.1780 grensene inkludert)
31. (a) 208 (b) 210 (c) 214.
32. (a) EX=1.0 varX=1.5 (b) Y binomisk (5, 0.5) 0.1874
(c) Z binomisk (10, 0.5) 0.0547 (d) 0.7407.
33. (a) EX=2.0, varX=1.0 (b) $P(S \leq 4) = 0.60$ (c) 0.206.
34. (a) 0.370, 0.370, 0.260 (b) 0.339, 0.339, 0.322
(c) 0.290, 0.290, 0.420 (d) 0.402, 0.402, 0.197
(e) 0.378, 0.378, 0.244 (f) 0.339, 0.339, 0.322
35. (b) (i) $\approx$ 0.500, 0.540, 0.584 (ii) 0.556, 0.618, 0.690
(c) 3000, 4100, 5800.
36. (b) 0.672 (c) 0.624 (d) 0.411.
37. (b) 0.6703 (c) 0.3679.
38. (a) 12 (b) 0.9825 (c) 0.7337 (d) 5 (e) $\approx$7.5.
39. (b) $\approx$ 0.349 (eksakt 0.350), $\approx$ 0.227 (0.233) (c) $P(Y \leq 3)$.
40. 0.9810, 0.0190. Urealistisk ved ulykker med fler involvert.
41. $\lambda = 15 \cdot 5/52 = 1.44231$, $P(X \geq 7) = 0.0007$
42. $P(X \leq 50) = 0.9491$ (0.9474 ved Poissontilnærming)
$P(X \leq k) = 0.90$ gir $k = 47(48)$.
43. $P(X \leq 200000) = 0.75$ gir $\mu = 176\ 227$
$P(X \leq k) = 0.90$ gir $k = 221\ 398$
44. (a) 10 (b) 4.3.
45. (a) EX=2.3 varX=2.01
(c) EY: 0.00, 4.0, 6.0, 5.0, 2.0, -2.0
(f) Z binomisk (n=52, p=0.4) (g) 0.644.
46. (a) 0.969 (b) 0.996 (c) 0.965 (d) 0.999 (e) 0.250 (f) 0.250.
47. (a) 0.0047 (b) 0.4032, 0.8159 (c) 0.3224.

48. 0.6266.

49. (a) $L = 15 \cdot X_1 + 100 \cdot X_2$, EL=650, varL=48025
(b) 0.055.

50. (a)

x	0	1	2	3	4
P(X=x)	9/13	1/13	1/13	1/13	1/13

EX=10/13, varX=290/169.
(b) ES=10, varS=17.066 (c) = 0.2675.

Kapittel 7.

1. (a) $1 - X/n$, $p(1-p)/n$ (b) $1 - 2X/n$, $4p(1-p)/n$.

2. (a) $N \cdot Y/n$ (b) $N \cdot (1 - Y/n)$.

3. (a) 0.6250, 0.6563, 0.7704, ≈0.843, ≈0.954
(b) 0.3456, 0.6665, ≈0.692, ≈0.850, ≈0.959
(c) 0.4096, 0.7717, ≈0.789, ≈0.923, ≈0.988
grov tilnærming, heltallskorreksjon gir bedre.

7. Begge er forventningsrette, men $\hat{\theta}$ har minst varians.

8. (a) (i) $X_1 + X_2$ (ii) $X_1 - X_2$ (iii) $(X_1 + X_2)/2$
(b) (i) $2\sigma^2$ (ii) $2\sigma^2$ (iii) $\sigma^2/2$.

9. (a) $(X_1 + X_2)/2$, $(X_1 - X_2)/2$ (b) $\sigma^2/2$, $\sigma^2/2$

10. (a) ≈0.497 (b) 1.152 (c) < 0.546.

11. (a) (i) 475 (ii) 900 (iii) 1275 (b) (i), (ii), (iii) 0.866.

12. (a) (i) 83, 90, 93 (ii) 244, 321, 359 (iii) 322, 474, 561
(iv) 454, 826, 1131 (b) 0.866 i alle.

13. (b) (i) 0.473 (ii) 0.510. (c) 200.

17. (b) $var\hat{p} = p(1-p)/(n_1 + n_2)$, $var\tilde{p} = p(1-p)(1/4n_1 + 1/4n_2)$

18. (a) NY/n (b), (c) $N_1Y_1/n_1 + N_2Y_2/n_2$.

19. Anslag: andel 53% St.avvik 1.8%
$P(X \geq 424) \approx 0.009$ litt tvil

22. (a) 0.44 ± 0.05 (b) 0.07 ± 0.10 (c) Tja.

23. (a) 0.44 ± 0.04 (b) 0.07 ± 0.09.
24. [0.0003 , 0.0557], [0.0024 , 0.0722], [0.0162 , 0.1167], [0.0480 , 0.1839] [0.0002 , 0.0279], [0.0012 , 0.0361], [0.0081 , 0.0584], [0.0240 , 0.0920]
25. (a) 0.0474 (b) 0.0237 (c) 0.0315.
26. (a) P=0.0668 (b) ikke forkast, forkast.
27. (a) P$\approx$ 0.136 nei (b) 59.
28. Nei, H_0: $p \leq 0.5$, H_A: $p > 0.5$ (pedagogens påstand).
29. (a) P$\approx$0.09 (b) ikke forkast, forkast.
30. (a) (i) $0.88 \pm 2 \cdot 0.065$ (ii) $0.91 \pm 2 \cdot 0.029$
    (b) (i) P=0.127 (ii) P=0.063.
31. (a) (i) k=2 (ii) k=11 (iii) k=2. (b) P=0.123 (c) (i) 0.38 (ii) 0.84 (d) n=39.
32. (b) (i) $k = 3$ (ii) $k = 5$ (iii) $k = 4$ (d) $P = 0.021$
33. A: $P = 0.05$, B: $P = 0.001$, C: $P = 0.25$, Alle: $P < 0.0001$.
34. Kjikvadrattest: $Q = 9.6$, $P = 0.008$, nei.
35. $Q = 7.90$, $P = 0.02$.
37. $Z = (9 - 11)/\sqrt{19/4} = 1.83$, $P \approx 0.0672$, tja
38. (a) $\hat{p} = X/n$ (b) $\hat{\lambda} = X$

Kapittel 8.

2. (b) (i) (0.41, 0.79) (ii) (0.37, 0.83) (iii) (0.30, 0.90)
   (c) (i) n=5 (ii) n=12.
3. (b) (i) k=1.645 (ii) k=1.96 (iii) k=2.58.
5. (b) 0.995.
6. (a) Styrke 0.999 (b) Forkast (P=0.0011).
7. (a) 0.9998 (b) > 0.9999

8. (b) Ja $Z = 4.0 > 1.96$ (c) 0.516, 0.516 (d) 21.

10. (c) (i) T_1 (ii) T_2 (iii) T_3 .

11. (c) n=24 (d) Nei.

12. (a) $N\overline{Y}$ (b) 218 (c) ≈ 0.046, ≈ 0.159.

13. (a) (i) 0.317 (ii) 0.0087 (b) (i) 0.171 (ii) 0.0000.

14. (a) Tilfellet n=4: (i) 0.95 (ii) 0.9999 (iii) > 0.9999
    (b) (i) n=18 (ii) n=30.

15. (a) (i) 3.365, 2.015, 1.476 (ii) 2.764, 1.812, 1.372.
    (b) (i) 4.032, 2.571, 2.015 (ii) 3.169, 2.228, 1.812.
    (c) (i) -2.015, 2.571 (ii) -1.812, 2.228.

16. (a) (i) 0.500, 0.705, 0.795, 0.844 (ii) 0.637, 0.898, 0.970, 0.990 (iii) 0.666, 0.935, 0.990, 0.999.
    Kun enkelte av disse verdiene kan finnes med brukbar tilnærmelse i tabellen i læreboken.
    (b) (i) 6.314, 12.71, 63.66 (ii) 2.015, 2.571, 4.032
    (iii) 1.761, 2.145, 2.977.

17. (a) 23.20 ± 0.54 (b) [21.55, 24.85].

18. (a) (2.894, 3.306) (b) (2.756, 3.444).

19. $T = -2.491$, $P = 0.0187$, Ja.

20. $T = 2.491$, $P = 0.0187$, Ja.

21. 0.03, 0.143, muligens.

22. (a) 0.1587 (b) 0.4044, 0.0667 (c) 14 (d) 780.9.

23. (a) 771.0 ± 5.1, [754.8, 787.4] (b) $T = -1.76 > -2.35$.
    Ikke forkast (c) Nei, ikke med elementær teori.

24. (a) $1 - 0.95^{10} = 0.40$ (b) 0.01 (c) 0.01.
    Beregningene forutsetter at testene er uavhengige.

26. Testobservator: K = Antall følger på samme side av median.
    EK = 26 (Oppg. 7.37)
    Eks.1.6 : K = 27, P=0.772, Oppg. 1.13: K=13, P=0.0002.

27. Anslag forventning 4 ± 0.25.
5 kg er ca. $\frac{1}{2}$ standandavvik mer enn forventet.
Anslått sjanse 31%.

29. (a) Anslag daglig forventning 109.3 (95% konf.int. [99.0, 119.6])
(b) ulike forventinger, bare en observasjon av total forventning.
Umulig å anslå tilfeldighetens rolle (over tid).

30. (b) t : 9.00 3.32 X : 750 203 R = 0.816.
Samme for Y og Z (!)
(c) $\hat{X} = 300 + 50.0 \cdot t$, $P \approx 0.001$ (9 frihetsgrader).

31. (c) T = 2.60, $P \approx 0.015$ (10 frihetsgrader)
(d) $S(\hat{Y} - Y) > S = 50.23$.

32. (a) $S^2 = 1.86$ (b) $T = -2.83$, $P = 0.0099$.

Kapittel 9.

2. Q=1.78, P=0.18.

3. (a) Q=5.77, 29.3, 5.17 (b) Q=2.65, 0.53.

5. Q=11.0, P=0.027 forkast.

6. Q=5.77, P=0.056 Nei.

7. Q=1.054, P=0.059. Ikke grunn til å påstå avh. av linje.

8. Q=1.26, P=0.74. Ikke grunn til å påstå noe annet.

9. (c) synes rimelig.

12. (a) 0 (b) 0.38 (c) -0.38 (d) 0.45 (e) 0.38 (f) 1.00.

13. (a) 0 (b) 0.20 (c) -0.20 (d) 0.22 (e) 0.14 (g) 0.33.

14. T=2.86 P=0.006 (tosidig).

16. (b) 5.0, 4.0, 2.32, 2.19, 3.90 0.768
(c) $\hat{Y} = 4.0 + 0.72(X - 5.0)$, $\hat{X} = 5.0 + 0.81(Y - 4.0)$
(d) 1.84, 5.44 (e) 3.38, 7.43 (i praksis nærmeste heltall).

Kapittel 10.

4. Fisher-Irwin test, $P = 0.295$, forkast ikke
   (kjikvadrattest $P = 0.273$).
5. Fisher-Irwin, $P = 0.20$. Nei. Tilfeldig utvalg.
6. Fisher-Irwin test, $P \approx 0.13$. Kjikvadrattest $P = 0.16$.
7. (c) $P = 0.090$. Ingen grunn til å påstå A bedre.
8. Wilcoxon W=44, $P \approx 0.092$ (eksakt 0.0946).
9. Wilcoxon W=132, $P \approx 0.16$.
10. Wilcoxon V=9, $P \approx 0.033$ (eksakt 0.0322).
11. Wilcoxon V=15.5 (15-16). $P \approx 0.033$. Forkast.

Kapittel 11.

1. (a) 0.49 ± 0.20 (c) $T = 2.45 \geq 1.86$, påstå A bedre.
2. (b) 0.55 ± 0.52 (c) $T = 1.05 < 1.812$, ikke påstå 1 bedre.
3. (a) $[-0.09, 9.49]$
   (b) $T = 2.22 > 1.833$, påstå $\delta > 0$
   (c) Wilcoxon (ett utvalg) eller tegntest.
4. (a) $[0.1, 10.1]$ (b) $T = 2.16 > 1.729$, påstå $\delta > 0$
   (c) Wilcoxon (to utv.).
5. (a) noe tvilsom (b) 5.30 ± 2.33
   (c) $T = 2.27 > 1.833$ (P=0.025).
6. (b) Nei. T=0.73 (P=0.238) alt. Wilcoxon $P \approx 0.25$.
7. (a) normalitet urimelig (b) Wilcoxon to utvalg.
9. (b) 70.6, 62.8, 74.6. Estimert st.avvik 4.8.
   (c) $F = 1.58 < 3.89$, ikke påstå forskjell.
10. $F = 6.79$, $P = 0.002$. Rangering (4, 3, 1, 2).
    4 klart bedre enn 1 og 2, og 3 trolig bedre enn 2

11. (b) $F = 11.77$, $P = 0.006$ (c) (3, 2, 1, 4)
    (d) $F = 1.92$, $P = 0.205$, ingen forskjell.

12. (b) $\hat{\mu} = 67.0$, $\hat{\alpha}_1 = -0.5$, $\hat{\alpha}_2 = -6.5$, $\hat{\alpha}_3 = 7.0$,
    $\hat{\beta}_1 = 11.0$, $\hat{\beta}_2 = -5.3$, $\hat{\beta}_3 = 2.0$, $\hat{\beta}_4 = -7.7$,
    (c) F=7.75, P=0.022

Kapittel 12.

1. (a) (1.75), (-6.0), (0.10), (0.00002)(-0.30, 3.10), (-0.81,-.039).

2. $Sp = -0.812 + \underset{(0.022)}{0.054} \cdot I - \underset{(0.426)}{0.468} \cdot F$, S=1.58, $R^2 = 0.41$.

3. (a) Alder $R^2 = 0.695 > 0.576$
   $\hat{Y} = 4711$, 46.30, $S(\hat{Y} - Y) \approx 5.33$, 4.5
   (b) $\hat{Y} = 47.04$ $S(\hat{Y} - Y) \approx 4.2$
   (c) 4.23 ± 1.91 (d) evt. 5 og 11.

5. (a), (b), (c), (e), men ikke (d).

6. Ikke (a), men (b), (c), (d) og (e).

7. (b) $\hat{Y} = 101.5 \cdot (1.197)^X$ (c) 357.8, 428.3 (d) Nei.

8. (a) $\hat{Y} = 10.44 - 0.3012 \cdot X + 0.01277 \cdot X^2$
   (b) 77.1, 131.8.

9. $\hat{Y}' = \underset{(0.0066)}{25.0} + \underset{(0.090)}{0.0269} \cdot X_1' + \underset{(1.45)}{0.368} \cdot X_2'$.

14. (b) $\hat{Y} = 1.200 \cdot X$, S=0.055 (c) 0.96, 1.20, 2.64 (d) 5/6.

15. (c) Nei.

Kapittel 13.

1. 16.8, 19.9, 20.3, 24.4.

2. (a) Trend $T_t = 15.4 + 0.69 \cdot t$
   (c) 22.3, 28.1, 29.1, 33.6.

3. (a) noe mindre stigning, kan skyldes tilfeldigheter.
   Felles trend $T_t = 6.54 + 0.73 \cdot t$.
   (b) Ja, trolig sesongfaktorer ca. 0.8, 1.0, 1.0, 1.2,
   prognoser 22.2, 28.4, 29.1, 35.8.

4. Trend $T_t = -2.79 + 1.22 \cdot t$
   trolig samme helning som Eks. 3, men skift.

5. (c) $\hat{Y}_t = 2.36 + 1.07 \cdot t + 2.09 \cdot Q_1 + 0.98 \cdot Q_2 + 5.77 \cdot Q_3 - 4.20 \cdot I$.

7. (b) Bruk $sin(2\pi k/12)$ (c) regresjonsanalyse.

8. (a) Ja. (b) $\hat{Y}_t = 0.38 \cdot Y_{t-1}$ (c) 0.38, 0.14.

14. (a) β, $2\sigma^2$ (b) 0, $2\sigma^2/(1+\rho)$ (c) 0, $(\alpha^2 - \alpha + 1) \cdot 2\sigma^2$.

15. (b) (i) 0.2, 0.36, 0.29, 0.43, 0.74
    (ii) 0.5, 0.75, 0.38, 0.69, 1.34.

16. (a) Fortidens innflytelse avtar geometrisk.
    (b) når U_t 'ene er uavhengige med $EU_t = 0$.

18. (b) $\hat{X} = 5.1415 + 0.987 \cdot t$
    (c) positiv autokorrelasjon. (d) $\hat{\rho} = 0.58$.

Kapittel 14.

2. (a) Estimat 54000 st.avvik 3432 (b) $P \approx 0.04$.

3. (d) $(N-n)/N(n-1) \cdot Y/n(1-Y/n)$.

4. (a) 0.20 ± 0.05 (b) 0.40 ± 0.06 (c) Ja.

5. (a) (i) 330 (ii) 83 (iii) 13 (b) (i) 289 (ii) 72 (iii) 12.

6. (b) $\sum(N_i - n_i)/N_i(n_i - 1) \cdot S_i^2$.

7. (a) optimal utvelging (b) 4970 ± 102 (c) Nei.

8. 5079.

9. (a) 0.40 (b) 0.56 (c) 4.

10. (a) $(N-n)/(N-1) \cdot n \cdot \sum_{i=1}^{n}(Y_i - \overline{Y} \cdot X_i/\overline{X})/n$.

11. (a) 2.00 ± 0.45 (b) ikke tilstrekkelig grunn.
    (c) 1.69 ± 0.16 grunn til å påstå reduksjon.

13. (a) $N/n \cdot \sum_{i=1}^{n} M_i \overline{Y}_i$ (b) 3134.75.

Kapittel 15.

1. (a)

M	0	1	2	3	4	5	6	7	8
$70 \cdot K(\theta)$	70	35	15	5	1	0	0	0	0
$70 \cdot K(\theta)$	70	70	70	65	53	35	15	0	0

(b)

M	0	1	2	3	4	5	6	7	8
$50 \cdot K(\theta)$	56	35	20	10	4	1	0	0	0
$50 \cdot K(\theta)$	56	56	50	40	28	16	6	0	0

2. (a) (0.50, 0.01), (0.00, 0.76) Eks. 1: (0.00, 0.24)
   (b) (0.38, 0.07), (0.00, 0.50).

3.

n=4	M	0	1	2	3	4	5	6	7	8
k=0	$Q(\theta)$	0	0.063	0.054	0.027	0.003	0	0	0	0

Samme $Q(\theta)$ for (a), (b) og (c).

		M	0	1	2	3	4	5	6	7	8
(a)	k=2	$Q_1(\theta)$	0	0.125	0.250	0.348	0.379	0.313	0.161	0	0
(b)		$Q_2(\theta)$	0	0.063	0.125	0.188	0.221	0.196	0.107	0	0

(c) som (a).

4. Formel $n \cdot K(\theta) + 8 \cdot (1 - K(\theta))$ der $K(\theta)$ er gitt i 1.

5. (b) (i) 0.564, 0.024 (ii) 0.459, 0.112 (iii) 0.358, 0.271.

6. (i) n=180 k=4.

7. (b) (i) $e^{-40\theta}$ (ii) $(1 + 80\theta)e^{-80\theta}$ (iii) $(1 + 100\theta)e^{-100\theta}$
   ikke så lenge N er stor.

8. (a) $K(\theta) = \sum_{y=0}^{1} \binom{M}{y}\binom{25-M}{8-y}/\binom{25}{8}$
   (b) $K(\theta) = \binom{25-M}{5}/\binom{25}{5} + \binom{M}{1}\binom{25-M}{4}/\binom{25}{5} \cdot \binom{20-M+1}{5}/\binom{20}{5}$

10. (a) 370 (b) 3.24, 1.02 (a') 224 (b') 2.41, 1.02.

11. (b) aksjon på tidspunkt 7.

12. (b) Nei (c) 6.6.

13. (a) $k_A = 4$ (b) $k_A = 2$ (c) Øke n.

14. Nei

15. (b) k=3.325 (c) 0.505.

16. n=3 k=5.3, 15.

17. Alarm $\overline{X} \leq 37.67$, $S \geq 4.74$ (Alarm ved tid 15).

20. Vokslag (A) uten betydning.
Luftrom (B) og høy tilsetning (C) ugunstig i kombinasjon.

Kapittel 16.

1. (a) a_1 , a_4 (b) a_2 (c) a_5 .

2. a_5 . Forventet gevinst 8.6.

3. (b) utvikl., hvis produkt holder, stor produksjon.
Forventet nettoinntekt 0.19 mill.

4. (b) sikre 1 mill. (c) risikoaversjon.

6. (a) 1/3, 2/3, 0 (b) 1/5, 4/5, 0 (c) 1, 0, 0.

7. (b) K, M like gode (c) nei, forventet gevinst −25.

8. (b) gjett kron (c) ja, forventet gevinst 25.

9. (c) kjøp en obs, gjett samme utfall. Forventet gevinst 5
(d) 20.

11. (a) 0.4, 0.4, 0.2 (b) 0.09.

12. (a) 0.4485, 0.5515 (b) 0.5000, 0.5000.

14. (a) 5/12 (b) 9/20 (c) 41/102 (d) 9/22 (e) 9/20.

15. (a) 10.33 (b) 0.84, 0.95 (a') 10.44 (b') 0.84, 0.84.

Appendiks E

Slutt

Dette er slutten