

STATISTISKE METODER I MEDISIN OG HELSEFAG

Odd O. Aalen (red.)

Arnoldo Frigessi

Tron Anders Moger

Ida Scheel

Eva Skovlund

Marit B. Veierød



GYLDENDAL

Odd O. Aalen (red.)
Arnoldo Frigessi Tron Anders Moger
Ida Scheel Eva Skovlund Marit B. Veierød

Statistiske metoder i medisin og helsefag



GYLDENDAL

© Gyldendal Norsk Forlag AS 2018

1. utgave, 2006

2. utgave, 2018

ISBN: 978-82-05-53601-2

Omslagsdesign: Gyldendal Akademisk

Layout og sats: Matematisk Sats/Arve Michalsen

Alle henvendelser om boken kan rettes til

Gyldendal Akademisk

Postboks 6730 St. Olavs plass

0130 Oslo

akademisk@gyldendal.no

www.gyldendal.no/akademisk

Det må ikke kopieres fra denne boken i strid med åndsverkloven eller avtaler om kopiering inngått med KOPINOR, interesseorgan for rettighetshavere til åndsverk. Kopiering i strid med lov eller avtale kan medføre erstatningsansvar og inndragning, og kan straffes med bøter eller fengsel.

Forord til 1. utgave

Denne boken tar sikte på å dekke behovet for en grunnleggende innføring i statistikk og sannsynlighetsregning. Den er særlig rettet mot lesere innenfor medisin og helsefag, men kan også være godt egnet for lesere med en annen bakgrunn. Mange av metodene som gjennomgås, er felles for de fleste anvendelser av statistikk, og eksemplene som benyttes er stort sett av allmenn interesse.

Boken har sitt utgangspunkt i *Innføring i statistikk med medisinske eksempler*, som kom på Ad Notam Gyldendal i 1994 (2. utg. 1998), men er en fullstendig omarbeiding av denne. Mye av stoffet er omorganisert og bearbeidet, og det er dessuten flere nye kapitler. *Eva Skovlund* har skrevet avsnitt 8.8 om ikke-parametriske metoder og kapittel 9 om kliniske forsøk, *Marit Veierød* har skrevet kapittel 10 om epidemiologiske grunnbegreper og design, *Tron Anders Moger* har skrevet kapittel 12 om logistisk regresjon og kapittel 13 om levetidsanalyse, og *Arnoldo Frigessi* og *Ida Scheel* har skrevet kapittel 14 om bayesiansk statistikk. Det øvrige stoffet er skrevet av undertegnede.

Mange av de metodene en i praksis vil støte på ved anvendelse av statistikk, er dekket i boken, og med de nye kapitlene gir boken en langt mer helhetlig innføring enn tidligere i statistiske metoder rettet mot medisinske anvendelser.

Boken gir også en innføring i sannsynlighetsregning frem til de mest grunnleggende fordelingene. Det gis stort sett fullstendige utledninger frem til og med binomisk fordeling, men senere er det ikke lagt særlig vekt på matematiske begrunnelser fordi det vil gjøre fremstillingen altfor komplisert. Det gis i stedet eksempler og intuitive forklaringer, og en del resultater er demonstrert ved simulering. Ønsket er å få frem det praktiske innholdet i ideene. Matematiske forklaringer er flere steder lagt til egne avsnitt som kan hoppes over uten å bryte kontinuiteten.

En bærende ide i boken er at leseren skal lære såpass mye sannsynlighetsregning at han kan skjønne hva P -verdier og konfidensintervaller egentlig er. Disse begrepene støter en på overalt i forskningslitteraturen, ikke minst i medisinen, og forståelsen er ofte dårlig.

Boken inneholder mange oppgaver, og de fleste av dem er gitt som eksamen i medisinsk statistikk ved Universitetet i Oslo. Vi er takknemlige for at andre lærere har gitt oss tillatelse til å bruke deres oppgaver, og vil særlig nevne Thore Egeland, Petter Laake, Knut Liestøl, Magne Thoresen og Lars Walløe. De aller fleste oppgavene tar utgangspunkt i reelle problemstillinger og gir eksempler som er et meget viktig

supplement til det øvrige stoffet i boken. Særlig eksamensoppgavene legger vekt på diskusjon og på at beregningene settes inn i en større sammenheng. Det er gitt tallfasit til alle oppgavene.

Det er opprettet en nettside for boken, der supplerende stoff vil bli lagt inn etter hvert, se <http://www.med.uio.no/imb/stat/statbok/>.

Jeg vil takke Nina Gunnes og Jon Michael Gran for hjelp med manus. Til slutt vil jeg takke forlaget ved Ellen Aspelund og Lisbeth Opøien for godt samarbeid om boken. Og ikke minst en takk til Arve Michaelsen som har gjort en stor innsats med den krevende layouten av et matematisk manus.

Oslo, desember 2005, Odd O. Aalen (red.)

Forord til 2. utgave

I denne utgaven er det gjort flere viktige endringer som bringer stoffet up to date. Det er særlig i regresjonsstoffet at større endringer er gjort, og det er også lagt inn noe nytt oppgavestoff. De tre fundamentale regresjonsmetoder i statistikk – lineær, logistisk og Cox – er beskrevet på en lett tilgjengelig måte i kapitlene 11 til 13. I kapittel 10 er det tatt inn litt moderne stoff om kausal inferens. I tillegg er det i alle kapitler gjort mindre endringer for oppdatering og modernisering av teksten. Forfatterne har vært ansvarlig for de samme kapitlene som beskrevet i forordet til 1. utgave, med unntak av at Tron Anders Moger også har revidert kapittel 11 i denne utgaven.

Jeg vil takke forlaget ved Jørgen Jahr Glomnes for meget godt og effektivt samarbeid om boken.

Oslo, desember 2017, Odd O. Aalen

Innhold

1	Tallenes verden	11
1.1	Vi er omgitt av tall	11
1.2	Tallene er viktige	12
1.3	Sannsynlighet og usikkerhet	13
1.4	Data i medisin	13
1.5	Data er usikre	15
1.6	Statistikk som bilde av virkeligheten	16
1.7	Dataprogrammer for statistisk analyse	18
2	Hvordan forstå og beskrive tall	19
2.1	Presentasjon av data	19
2.2	Hvor ligger sentrum i fordelingen?	23
2.3	Hvor mye varierer tallene?	26
2.4	Grupperte data	28
2.5	Søylediagram, kakediagram og tidsserieplott	31
2.6	Sammenlikning av grupper ved hjelp av punktdiagram og box-plott	33
2.7	Beskrivelse av samvariasjon	34
	Oppgaver til kapittel 2	36
3	Tilfeldighet og sannsynlighet	41
3.1	Begrepet tilfeldighet	41
3.2	Begrepene utfallsrom og begivenhet	44
3.3	Begrepet sannsynlighet	48
3.4	Regneregler for sannsynlighet	50
3.5	Gunstige/mulige-metoden	53
3.6	Betinget sannsynlighet	54
3.7	Stokastisk uavhengighet	56
3.8	Genetikk og sannsynlighetsregning	59
3.9	Total sannsynlighet og Bayes' lov	60
3.10	Anvendelser av Bayes' lov i medisin	63
3.11	Kombinatorikk	68
	Oppgaver til kapittel 3	73

4	Binomisk fordeling: Introduksjon til hypotesetesting	89
4.1	Sannsynlighetsfordelinger for tellevariabler	89
4.2	Binomisk fordeling	93
4.3	Introduksjon til hypotesetesting	100
	Oppgaver til kapittel 4	106
5	Normalfordelingen	111
5.1	Normalfordelingens to roller	111
5.2	Sannsynlighetsfordelinger for målevariabler	111
5.3	Normalfordelingen	115
5.4	Normaltilnærming til binomisk fordeling	118
	Oppgaver til kapittel 5	122
6	Statistisk analyse av binomiske tellevariabler	125
6.1	Estimering av en sannsynlighet. Hva er et konfidensintervall?	125
6.2	Testing av nullhypotese om en sannsynlighet	129
6.3	Sammenlikning av sannsynlighet i to grupper	131
6.4	Er forskjellen mellom gruppene reell?	134
6.5	Tabellanalyse: Khi-kvadrattest	137
	Oppgaver til kapittel 6	140
7	Poisson-fordelingen: Fordelingen for de sjeldne forekomster	157
7.1	Poisson-fordelingen: Egenskaper og beregninger	157
7.2	De uavhengige forekomster: Poisson-prosessen	161
7.3	Anvendelser av Poisson-fordelingen	164
7.4	Normaltilnærming til Poisson-fordelingen	167
7.5	Sammenlikning av to Poisson-variabler	169
	Oppgaver til kapittel 7	170
8	Sammenlikning av gjennomsnitt: t-tester og ikke-parametriske metoder	179
8.1	Egenskaper ved gjennomsnitt	179
8.2	Estimert standardfeil. Studentfordelingen	186
8.3	Konfidensintervall for forventningen	190
8.4	Testing av hypotese om forventningen	191
8.5	Analyse av pardata	192
8.6	Sammenlikning av to gjennomsnitt	194
8.7	Litt om de forutsetningene vi har gjort	196
8.8	Ikke-parametriske metoder	197
	Oppgaver til kapittel 8	203

9	Kliniske forsøk	213
9.1	Kliniske forsøk på mennesker	213
9.2	Protokoll	214
9.3	Forsøksplaner	214
9.4	Randomisering og blinding	215
9.5	Valg av effektvariabler	216
9.6	Bestemmelse av antall pasienter	217
9.7	Å vise likeverdighet av behandlinger	222
9.8	Statistisk analyse	224
9.9	Generalisering	225
9.10	Intern og ekstern validitet	226
9.11	Kvalitet av kliniske utprøvnings	227
	Oppgaver til kapittel 9	227
10	Epidemiologiske grunnbegreper og design	231
10.1	Introduksjon	231
10.2	Insidensrate og prevalens, viktige mål for sykdomsforekomst	232
10.3	Eksposering, risikofaktor	235
10.4	Konfundering og multivariable statistiske metoder	235
10.5	De forskjellige typene epidemiologiske studier	239
10.6	Feilkilder i epidemiologiske studier	246
10.7	Validitet	246
10.8	Kausalitet (årsakssammenheng)	247
	Oppgaver til kapittel 10	247
11	Lineær regresjon	251
11.1	Tilpasning av en rett linje	251
11.2	Korrelasjon	255
11.3	Illustrasjon: Hva er normal lungefunksjon?	256
11.4	Multippel lineær regresjon	258
11.5	Interaksjoner i regresjonsanalyse	260
11.6	Regresjonsanalyse ved hjelp av statistikkprogram	260
11.7	Noen kommandoer for bruk av SPSS i regresjon	267
	Oppgaver til kapittel 11	268
12	Logistisk regresjon	271
12.1	Introduksjon	271
12.2	Hva er logistisk regresjon?	271
12.3	Logistisk regresjon i praksis	273
12.4	Logistisk regresjon i SPSS	275
	Oppgaver til kapittel 12	276

13	Levetidsanalyse	281
13.1	Introduksjon	281
13.2	Sensurering	282
13.3	Overlevelseskurver, median levetid og Kaplan–Meier-plott	283
13.4	Log-rang-test for overlevelse i to grupper	287
13.5	Cox-regresjon	288
13.6	Cox-regresjon i praksis	290
13.7	Levetidsanalyse i SPSS	291
	Oppgaver til kapittel 13	292
14	Bayesiansk statistikk	297
14.1	Introduksjon	297
14.2	Et kritisk blikk på frekventistisk statistikk	298
14.3	Den bayesianske metoden for å lære fra data	300
14.4	Apriorifordelingen	302
14.5	Fra apriorifordeling til aposteriorifordeling via Bayes' lov: matematiske detaljer	305
14.6	Diskusjon	306
A	Diverse oppgaver	309
B	Oppgaver med løsninger	321
C	Metodekart	327
D	Litteratur	329
E	Fasit	331
F	Statistiske tabeller	343
	Om forfatterne	347
	Register	349

Tallenes verden

1.1 Vi er omgitt av tall

Vår kunnskap og våre oppfatninger om verden uttrykkes på mange måter. Mange fag er preget av en verbal tradisjon, der begreper, ideer og resonnementer er formulert i ord. På den annen side har vi den kunnskap som kommer via tall. Alle har vi lært om tall på skolen, og vi bruker dem i det daglige. Vi støter på dem overalt: når vi betaler i butikken, når vi måler feberen, når vi følger en oppskrift og baker en kake. Tallene i slike sammenhenger er greie og entydige; Vi vet alle hva 100 g sukker vil si (iallfall hvis vi har en vekt tilgjengelig!).

Vi er imidlertid også omgitt av tall som har en langt mer kompleks fortolkning. I Aftenposten var det en gang følgende oppslag på førstesiden: «90 dør av Oslo-luften hvert år.» Dette er resultat av en større undersøkelse som konkluderte med at luftforurensning tar 90 liv. Ved første blick kan det virke som et slikt tall er greit å forholde seg til, akkurat som en teller opp at det er 90 mennesker i auditoriet. Imidlertid skjønner en ved litt ettertanke at tallet 90 i Aftenposten ikke er et resultat av noen enkel telling. Man kan telle antallet som dør i Oslo i løpet av et år, men man kan ikke telle antallet som dør av luftforurensning. Ingen lege skriver en slik diagnose på dødsattesten, fordi det ikke lar seg fastslå på noen som helst sikker måte i det enkelte tilfellet. Tallet 90 er altså ikke resultat av noen telling. Snarere er det kommet frem ved bruk av kompliserte matematiske modeller. Slike modeller gir alltid usikre svar og beror på en serie med antakelser som er diskutabile. Det tilsynelatende enkle tallet 90 dekker altså over en kompleks virkelighet og har lite til felles med de tallene vi normalt møter i vår hverdag.

Når avisen skriver at Arbeiderpartiet har en oppslutning på 28,4 %, er dette et noe enklere tall. Utvilsomt er det resultatet av en telling; gallup-instituttet har f.eks.

intervjuet 1000 personer og spurt hva de ville stemme på hvis det var valg i morgen. Imidlertid er det klart at tallet har en annen svakhet, det er jo bare basert på en bitte liten brøkdel, mindre enn en promille, av den stemmeberettigede befolkning. Dette gir en usikkerhet i tallet, kanskje 2–3 % i hver retning. Selv om galluptallet er et prosenttall har det derfor en helt annen karakter enn når vi får vite at buksen i butikken er nedsatt med 30 %.

Tall i media presenteres ofte med lite innsikt. For eksempel ble det i en melding fra NTB i begynnelsen av 1997 meddelt følgende: «Åtte prosent av de mannlige og 14 prosent av de kvinnelige legene røyker.» Dette ble i en avis slått opp stort på følgende måte: «22 prosent av legene røyker daglig.»

En stor overskrift i en avis forkynte følgende: «Kvinner ser mest fjernsyn.» Videre sies det at «for første gang siden fjernsynet kom til Norge, ser kvinner mer på TV enn menn», og det vises til en undersøkelse utført av Markeds- og Mediainstituttet. Ser en nærmere på tallene, oppdager en imidlertid at kvinnene ser fjernsyn 144 minutter per dag, mens mennene ser 142 minutter. For det første er jo denne forskjellen svært liten. For det andre er den høyst usikker av flere grunner: Det er bare et lite utvalg av personer (kanskje noen hundre) som er spurt. Det er også meget vanskelig å anslå hvor lenge en ser på fjernsyn per dag (bare forsøk det for deg selv). I beste fall er det et usikkert anslag for hver person som er med i undersøkelsen. Gjennomsnittene som blir oppgitt, er derfor høyst usikre tall, og det er følgelig ikke noe grunnlag for å påstå at kvinner ser mer på fjernsyn enn menn.

Tallene omgir oss, de kryr i avisen, de kastes mot oss fra TV-skjermen og på Internett. Mange av disse tallene skjuler langt mer enn man tenker over når man ser dem. Det er en stor utfordring å forstå hva kvantitativ informasjon egentlig betyr. Det hersker en utbredt «numerisk analfabetisme» med dårlig forståelse av tallenes begrensninger.

1.2 Tallene er viktige

Tall er ikke bare plagsomme størrelser som kommer forstyrrende inn i ordenes verden. De gir oss essensiell kunnskap. Et dramatisk eksempel på dette er det følgende: Gjennom 1980-årene så en i Norge og mange andre land en sterk stigning i forekomsten av plutselig spedbarnsdød. Som navnet sier, innebærer dette at et spedbarn dør plutselig uten at man ved obduksjon finner noen årsak. Denne økningen var et mysterium inntil det dukket opp undersøkelser fra enkelte land som viste at det kunne være en sammenheng med den måten spedbarna ble lagt i sengen på. Utover i 80-årene ble det nemlig gitt sterke anbefalinger om at spedbarn skulle ligge på magen, da skulle de ikke kveles i sitt eget gulp. Nå oppdaget man at barn som lå på ryggen, hadde mye mindre risiko for å dø. En begynte derfor også i Norge å anbefale at barna ble lagt på ryggen eller i sideleie, og dette hadde en umiddelbar og dramatisk effekt. I 1989 var det 142 tilfeller av plutselig spedbarnsdød, i 1991 bare 51 tilfeller, altså ca. en tredjedel. I dag er det

under fem tilfeller av plutselig spebarnsdød hvert år. Det som overbeviste leger og helsesøstre om faren ved å legge spedbarn på magen, var nettopp tallene over dødsfall og hvordan de forandret seg. Et særdeles uheldig råd som gjennom en årrekke hadde blitt gitt til nybakte foreldre, ble altså avslørt ved bruk av tall.

1.3 Sannsynlighet og usikkerhet

Når vi går til legen, vil vi gjerne tro at han vet hva han gjør. Får vi medisin, regner vi med at den virker og iallfall ikke er farlig for oss. Blir vi sendt til røntgenundersøkelse, regner vi med at resultatet er riktig enten testen er positiv (tegn på sykdom) eller negativ (ikke tegn på sykdom). At det ikke nødvendigvis er så enkelt, blir vi likevel stadig minnet om gjennom massemedia. I Dagbladet var en førsteside viet oppslaget: «92 døde av vanlig medisin». Dette var antallet pasienter i Norge som i løpet av ett år fikk en så alvorlig bivirkning av sin medisin at de døde. Det var til dels snakk om alminnelig brukte medisiner.

Et annet eksempel: NRK Dagsrevyen hadde for noen år siden et intervju med en kvinne som hadde fått fjernet et bryst fordi legene trodde det var kreft i det. Senere viste det seg at det var feil. Kvinnen hadde ikke kreft, men hadde fått store plager på grunn av brystet hun hadde mistet.

Som i all annen virksomhet er det altså også i medisinsk behandling en risiko for alvorlige feil eller utilsiktede bivirkninger. Nøkkelordet her er risiko, som er nært beslektet med sannsynlighet. Det siste kjenner vi fra mange sammenhenger: Leverer vi f.eks. en utfyllt lottorekke, er sannsynligheten eller sjansen for å vinne ca. én på fem millioner.

Statistikk og sannsynlighetsregning er hjelpemidler til å forstå tall og usikkerheter. Faget dreier seg om hvordan vi samler inn tall og informasjon og hvordan vi analyserer dem slik at vi forstår hva tallene sier. Registreringer eller informasjon som foreligger som tall, kalles data. Dataene kan være resultat av målinger langs en skala, f.eks. av et menneskes høyde, eller de kan oppstå fra tellinger, f.eks. ved at en teller opp antall røykere i en gruppe personer. Mye av den kunnskapen vi har om verden rundt oss, er basert på systematiske målinger og tellinger. I statistikk tar vi for oss metoder for å organisere og analysere slike data.

1.4 Data i medisin

Både i klinisk medisin og i medisinsk forskning består mye av virksomheten i å samle inn opplysninger og kunnskap. Legen måler pasientens puls eller kolesterol, han spør pasienten om vedkommende har smerter eller er depriment. Ofte foreligger slike opplysninger i form av tall, f.eks. finner kanskje legen at pasientens puls er 72 slag per

minutt, eller at kolesterolnivået er 6,8 mmol per liter. Andre ganger er opplysningene mer upresise: Kanskje får legen vite at pasienten har ganske sterke smerter i korsryggen. En slik opplysning kan selvfølgelig ikke uten videre uttrykkes som et tall.

Hvis data er resultater av tellinger eller målinger, vil de være gitt som tall. I moderne medisin spiller slike talldata en stadig økende rolle. Legen tar f.eks. en blodprøve av pasienten og sender den til et laboratorium for analyse. Han får så tilbake en liste med tall som angir mengden av diverse stoffer i blodet, telling av røde og hvite blodceller osv. Denne listen, som kanskje kan være meget lang, skal være et hjelpemiddel for legen til å finne ut om pasienten lider av en bestemt sykdom. Legen må derfor kunne forholde seg til slike talldata og vite litt om hvordan han skal hankses med dem. Videre spiller data og statistikk en svært viktig rolle i den systematiske utprøvingen av medisinsk behandling. En stiller i dag meget strenge krav til vurdering av nye behandlinger. Disse kravene innebærer omhyggelige registreringer og målinger og grundig analyse av de fremkomne data. Dette må legen ha god kjennskap til hvis han skal kunne ta stilling til verdien av forskjellige medisinske behandlingsformer.

Store deler av medisinen er naturvitenskapelig, i den forstand at opplysninger og data er resultat av målinger med apparater på samme måte som i andre naturvitenskaper, og at de studeres etter liknende prinsipper. Mye av medisins fremskritt er knyttet til introduksjonen av naturvitenskapelige metoder, som derfor har fått en stadig sterkere posisjon. Bruk av statistikk er særlig egnet i forbindelse med denne formen for vitenskapelig metode.

Selv om data ikke naturlig foreligger som tall og er direkte resultat av måling eller telling, vil en ofte ha behov for å forsøke å uttrykke dem som tall likevel. En kan f.eks. være interessert i å sammenlikne to behandlinger mot smerter i korsryggen (f.eks. behandling hos kiropraktor kontra behandling med akupunktur). La oss si at ti pasienter får den ene behandlingen og ti den andre, og at en etter en måned ønsker å sammenlikne de to gruppene av pasienter med hensyn til hvor vondt de fremdeles har i ryggen. Et menneskes opplevelse av smerte kan jo ikke måles med et apparat. Likevel kan sammenlikningen mellom to slike grupper bli mye enklere hvis en kan finne tallmessige uttrykk for den smerten de føler. En kan da regne ut et gjennomsnitt i hver gruppe og på den måten finne ut hvilken av gruppene som har klart seg best.

En måte å tallfeste subjektive inntrykk som smerter på er ved hjelp av en *visuell analogskala* (VAS). Pasienten skal krysse av på et sted på skalaen som han selv mener svarer til den smerten han føler (se figur 1.1). En kan så måle avstanden fra venstre endepunkt til det stedet pasienten krysser av, og bruke dette som et tallmessig mål for pasientens smerte.

I motsetning til måling ved hjelp av apparater i naturvitenskapelig forstand vil bruken av slike analogskalaer ha et langt mer subjektivt preg. Å måle smerte kan aldri bli objektivt på samme måte som det å måle feber ved hjelp av et termometer. En kan derfor rette mange innvendinger mot bruken av analogskalaer. Likevel spiller de en



Figur 1.1 Visuell analogskala

meget viktig rolle i medisinsk forskning.

Ønsket om å tallfeste subjektive menneskelige opplevelser dukker også opp i andre sammenhenger i medisinen. I psykiatri studerer en behandling mot depresjon. En har utviklet metoder, dels basert på intervjuer, for å måle hvor depriment et menneske er, og uttrykke dette ved en tallindeks. Et velkjent eksempel på en slik indeks fra et annet område er intelligenskvotienten, som er et tallmessig uttrykk for et menneskes problemløsende evne. Slike indekser spiller en meget viktig rolle i deler av medisinen, men en må alltid ha for øye at tallet bare vil være en blek avskygning av den menneskelige virkelighet en ønsker å måle. Tallene og indeksene er alltid en reduksjon av virkeligheten bakenfor.

1.5 Data er usikre

Hvis en spør om et menneskes kjønn, vil dette være en opplysning med svært liten usikkerhet. (Ved meget sjeldne sykelige tilstander kan det være vanskelig å fastslå kjønn på et nyfødt barn). Selv om enkelte data derfor er svært sikre, vil de fleste være beheftet med en viss usikkerhet.

Hvis en måler pulsen til en pasient og finner at den er 72 slag per minutt, vil en alltid kunne stille en del spørsmålstegn ved dette tallet. For det første kan en spørre seg hvordan målingen er skjedd. Har en talt pulsslag over et helt minutt, eller har en kanskje bare talt over 15 sekunder og så multiplisert med 4? Kanskje har en talt galt? Er det tale om en pasient som er nervøs ved et legebesøk og derfor har høyere puls enn normalt? Har pasienten akkurat gått opp en trapp og dermed fått økt pulsen? Selv om ingen slike spesielle forhold er til stede, vil en likevel som regel finne at en ny måling av pulsen gir et noe annet resultat enn den første. Pulsen varierer over tid og er ikke en konstant størrelse. Tallet 72 kan godt være et temmelig usikkert uttrykk for hvilken puls pasienten har i sin alminnelighet.

Slik vil de aller fleste ting som registreres, være beheftet med usikkerhet. Hvis f.eks. en pasient besøker legen fordi han er depriment, vil den enkelte leges holdning og personlighet kunne være avgjørende for om pasientens sinnsstemning kommer klart frem. Kanskje føler pasienten trygghet og tør å fortelle om sine problemer, kanskje gjør han det ikke.

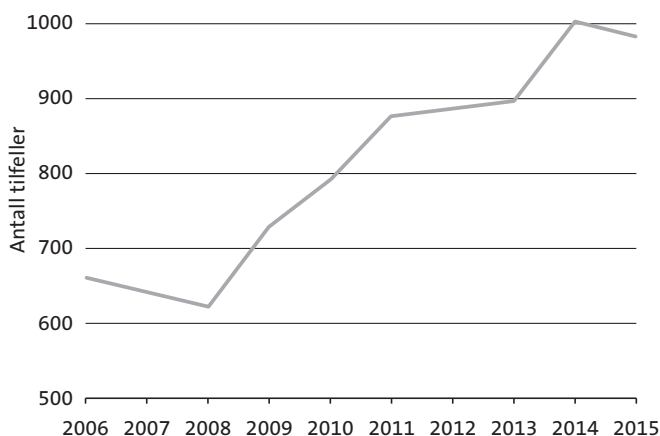
En annen form for usikkerhet skyldes målefeil. Hvis f.eks. en rekke studenter ut-

fører samme kjemiske laboratorieforsøk, vil de nesten alltid komme til noe forskjellig resultat. Dette er en alminnelig erfaring, og slik variasjon vil også være til stede ved forsøk som utføres av spesialister. Et særlig problem utgjør måleusikkerhet ved kliniske laboratorieprøver av pasienter. I verste fall kan en feil her føre til gal diagnose og behandling.

Faget statistikk er et verktøy som bl.a. nyttes til å forstå og analysere usikkerheten i data. En kan beregne hvilke konsekvenser usikkerheten har, og innrette seg slik at en kan kontrollere dens innflytelse. Dette kommer vi tilbake til i senere kapitler.

1.6 Statistikk som bilde av virkeligheten

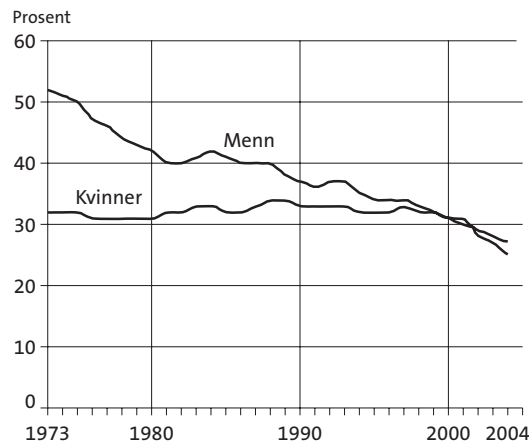
Det er viktig å få et bilde av hvordan sykdomsrisiko endrer seg over tid. For mange kreftformer har en sett en betydelig økning over de senere år. Et eksempel er gitt på figur 1.2, som viser endringen i forekomst av føflekkreft blant kvinner i Norge. Økningen over en tiårsperiode er fra rundt 600 til 1000 tilfeller per år. Dette er den kreftformen som øker mest for tiden. Man antar det å utsette seg for sterk sol uten tilstrekkelig beskyttelse er en viktig årsaksfaktor.



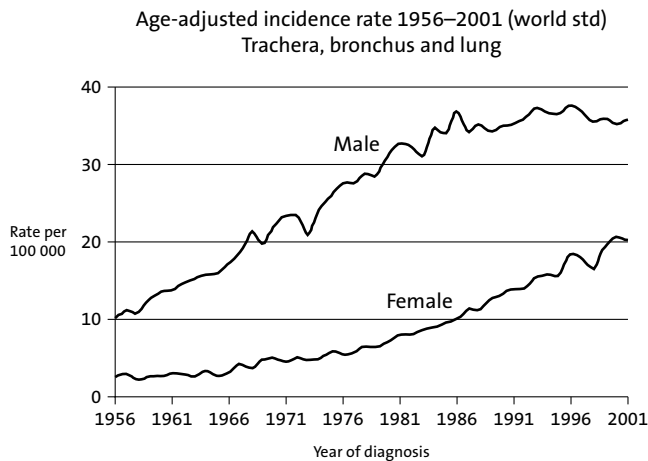
Figur 1.2 Forekomst av føflekkreft blant kvinner i Norge, 2006–2015

De negative effektene av røyking har vært et sentralt tema i medisinen i mange år. På figur 1.3 ser du utviklingen i antall røykere, som har falt sterkt over mange år blant menn, og nå også begynner å falle blant kvinner. Utviklingen i lungekreft ser du på figur 1.4. Til tross for den gunstige utviklingen i antall røykere kan man bare så vidt begynne å spore et fall i insidensen av lungekreft for menn. For kvinner stiger insidensen stadig sterkt. Forklaringen på dette er at lungekreft utvikler seg over lang

Andel som røyker daglig blant kvinner og menn 16–74 år, 1973–2004. Treårig glidende gjennomsnitt. Prosent



Figur 1.3 Utvikling i antall røykere over tid. Figur fra Statistisk sentralbyrå



Figur 1.4 Insidens av lungekreft blant menn og kvinner. Figur fra Krefregisteret

tid. En gunstig utvikling med hensyn til røyking vil derfor først kunne spores etter mange år. Her ser vi hvordan de statistiske oversiktene er et essensielt element i vår forståelse av sykdommen.

Uttrykket Big Data er nylig blitt veldig populært. Med Big Data tenker vi på alle de enorme datamengdene som i dag er blitt tilgjengelige, og som ikke bare er kolossale i størrelse, men ofte mye mer uorganiserte enn det man har vært vant til. Helsefeltet

er fullt av store datasett med potensial for å hente ut mye ny medisinsk kunnskap. Et eksempel er genetiske data, der f.eks. kartlegging av hele det humane genom gir enorme datamengder for hvert enkelt individ, med 3,2 milliarder basepar. Ideen er å koble dette sammen med kunnskap om sykdomsrisiko og behandlingseffekter slik at hver person kan få den optimale forebygging og behandling. I tillegg til genene kommer også data om metylering, som bestemmer i hvilken grad genene er «skrudd på». Videre er det feltet proteomikk, som dreier seg om de rundt 400 000 proteinene som genene koder for.

Et annet spørsmål er om personenes medisinske historie, nedfelt i medisinske journaler og registre, kan benyttes til å finne ut hvilke behandlinger som virker best. På disse og mange andre felt er store datamengder samlet og er i prinsippet tilgjengelig for statistiske analyser.

1.7 Dataprogrammer for statistisk analyse

Det eksisterer en rekke dataprogrammer som kan brukes til å utføre alle de analysene du finner i denne boken, samt mange andre statistiske analyser som ikke blir omtalt her. Et godt, omfattende og brukervennlig program er SPSS (Statistical Package for the Social Sciences; programmets internett-adresse: www.spss.com). Et noe videregående program er STATA (www.stata.com), som er godt egnet til medisinske problemstillinger, og som nyter en økende internasjonal popularitet.

Profesjonelle statistikere bruker gjerne programmet R (www.r-project.org). Dette er noe mer avansert enn de andre programmene som er nevnt, men brukes også i økende grad av andre fagfolk enn statistikere når en har store og kompliserte datamengder.

Hvordan forstå og beskrive tall

2.1 Presentasjon av data

Når en samler inn store mengder av data, er det avgjørende at de blir presentert på en hensiktsmessig måte. Vi skal se på forskjellige måter å gjøre dette på.

Tabell 2.1 Alderen til hver av de nye medisinstudentene i Oslo ett år. Alderen er beregnet per 31.12.84 og angitt med to desimaler (eksakt alder). Rekkefølgen nedover er alfabetisk. Antall studenter er 150.

20,53	24,92	22,78	21,10	20,51	24,96	22,75	23,63	21,11	31,50
20,92	22,58	24,58	21,72	24,41	21,24	22,53	20,25	24,96	28,15
19,62	31,71	33,29	26,09	21,05	21,72	20,37	21,80	22,60	21,90
22,33	27,88	21,74	19,03	21,45	23,15	21,51	20,34	19,64	23,06
21,70	25,15	23,65	20,11	21,62	21,56	20,67	21,68	19,51	21,55
25,53	21,61	19,68	21,90	19,21	20,98	19,82	21,77	22,11	21,93
21,76	23,05	21,91	23,80	21,42	20,40	20,37	26,14	20,45	23,50
21,82	20,40	22,66	21,04	20,53	22,72	26,75	22,72	22,75	21,68
21,13	23,89	24,75	19,80	22,07	23,42	21,78	22,85	23,30	21,38
20,82	21,80	23,01	26,15	22,88	22,62	27,47	23,02	20,75	25,18
21,45	20,80	20,15	21,86	21,91	26,98	24,10	34,15	26,08	21,12
29,63	22,84	22,57	20,72	21,50	21,23	21,53	23,45	23,06	21,33
21,94	21,78	24,71	28,07	21,13	26,73	20,42	19,90	21,29	23,62
18,60	25,54	23,10	25,56	23,77	22,67	22,18	21,72	20,75	23,85
27,90	28,38	20,81	21,21	22,89	21,88	21,75	20,70	25,39	24,12

La oss tenke oss at vi er interessert i alderen til et kull studenter. Da kan vi finne ut hva alderen til hver enkelt student er, og samle disse opplysningene. Resultatet for et kull medisinstudenter, stilt opp i alfabetisk rekkefølge, er gitt i tabell 2.1. For å illustrere visse beregningsmetoder er alderne gitt på en litt usedvanlig måte, nemlig med to desimaler (beregnet ut fra fødselsdatoen). Alderen 23,77 år betyr at vedkommende per 31.12.84 var $23 + 77/100$ år gammel.

En oppstilling av en masse data, som i tabell 2.1, er ikke særlig egnet til å gi noe overblikk. Det er vanskelig å få noe som helst inntrykk av hvordan alderne fordeler seg. En må derfor finne frem til bedre måter å stille opp tallene på.

Tabell 2.2 Hyppighetstabell over medisinstudentenes alder. De kumulative relative hyppighetene er beregnet som summen av de relative hyppighetene opp til og inklusive den aktuelle klassen. (At totalsummen blir 100,1 istedet for 100,0, skyldes avrundingsfeil.)

Klasse- intervaller	Alder i år	Hyppighet	Relativ hyppighet	Kumulativ rel. hypp.
18,00–18,99	18	1	0,7 %	0,7 %
19,00–19,99	19	9	6,0 %	6,7 %
20,00–20,99	20	23	15,3 %	22,0 %
21,00–21,99	21	46	30,7 %	52,7 %
22,00–22,99	22	20	13,3 %	66,0 %
23,00–23,99	23	18	12,0 %	78,0 %
24,00–24,99	24	9	6,0 %	84,0 %
25,00–25,99	25	6	4,0 %	88,0 %
26,00–26,99	26	7	4,7 %	92,7 %
27,00–27,99	27	3	2,0 %	94,7 %
28,00–28,99	28	3	2,0 %	96,7 %
29,00–29,99	29	1	0,7 %	97,4 %
30,00–30,99	30	0	0,0 %	97,4 %
31,00–31,99	31	2	1,3 %	98,7 %
32,00–32,99	32	0	0,0 %	98,7 %
33,00–33,99	33	1	0,7 %	99,4 %
34,00–34,99	34	1	0,7 %	100,1 %
Totalt		150	100,1 %	

Hvis vi skal kunne danne oss et klarere bilde av studentenes alder, kan vi sette opp en *hyppighetstabell*. Dette kan f.eks. skje ved at vi teller opp antallet som er 18 år gamle, antallet som er 19 år gamle, osv. Dette er hyppighetene av de forskjellige utfall. Resultatet er gitt i tabell 2.2. I tabellen finner du også *relative hyppigheter* i prosent,