



Gunnar G. Løvås

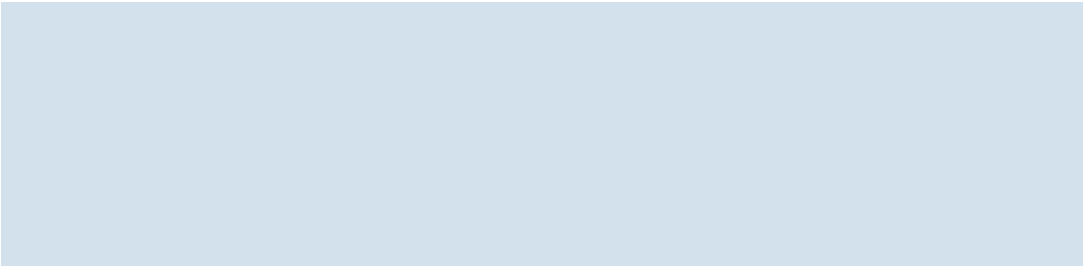
Statistikk

FOR UNIVERSITETER OG HØGSKOLER

4. utgave

UNIVERSITETSFORLAGET

Statistikk for universiteter og høyskoler



Statistikk for universiteter og høyskoler

Gunnar G. Løvås

4. utgave

Universitetsforlaget

© Universitetsforlaget, Oslo, 2020

ISBN: 9788215032924

1. utgave, 1. versjon, Universitetsforlaget, 2020

E-utgaven baserer seg på 4. trykte utgave, Universitetsforlaget 2018

Elektronisk tilrettelegging: Universitetsforlaget

Omslagsdesign: Universitetsforlaget / Sissel Tjernstad

Sats/design: Gamma grafisk AS (Vegard Brekke)

Materialet i denne publikasjonen er omfattet av åndsverklovens bestemmelser. Uten særskilt avtale med Universitetsforlaget er enhver eksemplarframstilling og tilgjengeliggjøring bare tillatt i den utstrekning det er hjemlet i lov eller tillatt gjennom avtale med Kopinor, interesseorgan for rettighetshavere til åndsverk.

www.universitetsforlaget.no

post@universitetsforlaget.no

Forord

Jeg har skrevet denne boka fordi jeg synes at statistikk og sannsynlighetsregning er nyttig og spennende, og for at du skal lese den. Du blir herved invitert med på en «reise» der vi i fellesskap skal bli kjent med statistikkfagets viktigste temaer. Boka er primært skrevet for deg som ønsker å bruke statistikk som et *verktøy* i din framtidige arbeidssituasjon. Du kan godt tenke på denne boka som en verktøykasse. Her finnes det mange generelle verktøy (skrujern, hammer og sag) som du helt sikkert får bruk for, og noen spesialverktøy (drill og vinkelsliper). Vi kommer til å bruke mest tid på de generelle verktøyene, men du bør også orientere deg om hvilke spesialverktøy som finnes.

Statistikkfaget inngår som et basisfag i de aller fleste utdanninger på høgskole- og universitetsnivå. Enten du studerer økonomifag, ingeniørfag, fiskerifag eller farmasi, så passer denne boka for deg. De ulike fagdisiplinene vektlegger ulike deler av faget, men basisverktøyene er de samme. Uansett hva du studerer, er det stor sannsynlighet for at boka inneholder det verktøyet du trenger. For å gjøre boka lettlest, interessant og underholdende har jeg valgt eksempler fra mange ulike fagområder. Boka inneholder litt matematikk, men det er viktigst at du forstår *ideene* og vet hvordan de ulike verktøyene kan brukes (og eventuelt misbrukes).

Mange studenter kan en del statistikk fra videregående skole, slik at kapittel 1 til 3 vil inneholde mye kjent stoff. Et typisk *innføringskurs* vil fokusere på kapitlene 3 til 6, som utgjør fagets kjerne og presenterer basisverktøyene, og i tillegg gi noen smakebiter fra kapittel 7 til 9 som presenterer de viktigste statistiske analyseverktøy. Boka er såpass omfattende at den også kan brukes til et *videregående kurs* der man går i dybden på flere av kapitlene 7 til 12.

Boka er betydelig oppjustert når den utkommer i sin 4. utgave. Et nytt kapittel er inkludert om Big Data, med ambisjon om å gi et overblikk over de nye muligheter og utfordringer som oppstår i en stadig mer digitalisert verden. Selv om mange snakker mye om nye begreper som kunstig intelligens og maskinlæring, handler det i bunn og grunn om analyse av data som grunnlag for å ta kloke beslutninger. Gode kunnskaper om statistikkfaget kan derfor åpne mange spennende karrieremuligheter i en tidsalder der data er det nye gullet!

Boka er også oppgradert ved at statistikkprogrammet R er inkludert, sammen med Excel, Minitab og SPSS. R er et mye brukt verktøy, godt tilpasset analyse av store datamengder.

Mange fortjener takk for at boka blir stadig bedre. Jeg vil takke Jan Terje Kvaløy, Kai Forsberg Kristensen, Trygve Almøy og Morten Helbæk, som har kommet med mange verdifulle kommentarer til boka og har bidratt med oppgaver til bokas nettside. Jeg er også stor takk skyldig til Birthe Aarekol og Kjartan Kloster Osmundsen, som har laget beskrivelsen av R og har oppdatert omtalen av de andre statistikkprogrammene. Takk til Thomas Trötscher for gode faglige innspill til kapitlet om Big Data. Takk til Richard Fossum i Finn.no, Boye Annfelt Høverstad i Statnett og Olav Magnus Linge i Zynk, som har bidratt med tekst og illustrasjoner av relevante eksempler til kapitlet om Big Data. Jeg vil spesielt takke forlagsredaktør Eli Valheim og layout-general Vegard Brekke for deres verdifulle bidrag til å gjøre boka enda bedre. En stor takk også til alle som har sendt meg kommentarer og feilmeldinger til bokas forrige utgave! Sist, men ikke minst, en stor takk til min kjære Kate, som har motivert meg til revisjonsarbeidet i en ellers hektisk hverdag!

Hvis du oppdager feil eller har kommentarer, vil jeg gjerne høre fra deg. Besøk www.nettressurser.no/statistikk og skriv en melding til meg. På forhånd takk! På nettsiden vil du også finne nyttig tilleggsstoff, inkludert datasett, flere oppgaver og noen løsningsforslag.

Jeg har skrevet læreboka for å hjelpe deg i møtet med et nytt fag. Dine medstudenter og forelesere kan også hjelpe deg, men det er *du* som må lære faget. Skal du lykkes, må du være villig til å bruke tid og krefter og ikke gi opp første gang du «møter veggen». Diskuter problemene med noen, og les en gang til. Da er det stor sjanse for at du kommer deg videre. Hvis du lykkes, vil du lære hvordan du bør forholde deg til usikker informasjon, hvordan du skal samle inn og analysere data, og hvordan du kan ta gode beslutninger. Jeg håper at du vil like både statistikkfaget og denne boka, og ønsker deg lykke til videre på reisen.

Oslo, juni 2018

Gunnar G. Løvås

Innhold

Del I Innledning

1 Hva er statistikk? 19

- 1.1 Bokas innhold 20
 - 1.1.1 Noen eksempler 20
 - 1.1.2 Historie 23
 - 1.1.3 Bokas oppbygning 25
- 1.2 Noen viktige begreper 26
 - 1.2.1 Populasjon og utvalg 26
 - 1.2.2 Variasjon 28
 - 1.2.3 Variabeltyper 29
- 1.3 Modeller 30
- 1.4 Innsamling av data 32
 - 1.4.1 Problemløsning krever data 32
 - 1.4.2 Generelt om datainnsamling 33
 - 1.4.3 Spørreundersøkelser 34
- 1.5 Oppsummering 36

2 Hva er fakta om utvalget? 37

- 2.1 Beskrivelse av kategoriske data 38
- 2.2 Beskrivelse av diskrete data 41
- 2.3 Beskrivelse av kontinuerlige data 42
 - 2.3.1 Frekvenstabell 42
 - 2.3.2 Histogram 44
 - 2.3.3 Stamme- og bladdiagram 46
 - 2.3.4 Spredningsplott 46
 - 2.3.5 Tidsserieplott 47
 - 2.3.6 Generelle diagramtips 48
 - 2.3.7 Formen på histogrammet 49

2.4	Sentralmål	50
2.4.1	Modus	50
2.4.2	Median	51
2.4.3	Gjennomsnitt	52
2.4.4	Sammenligning av modus, median og gjennomsnitt	55
2.5	Spredningsmål	56
2.5.1	Variasjonsbredde	56
2.5.2	Utvalgets varians og standardavvik	57
2.5.3	Variasjonskoeffisient	59
2.6	Skjevhet og kurtosis*	60
2.7	Grupperte data*	60
2.8	Statistikk med Excel	62
2.8.1	Legge inn data	62
2.8.2	Vanlige diagrammer	63
2.8.3	Frekvenstabeller og histogrammer	64
2.8.4	Beskrivende statistikk	65
2.9	Statistikk med Minitab	66
2.9.1	Skjermbildet	66
2.9.2	Grafikk og statistikk	68
2.10	Statistikk med SPSS	69
2.11	Statistikk med R	70
2.11.1	Skjermbildet	70
2.11.2	Grafikk og statistikk	71
2.12	Oppsummering	72
	Oppgaver til kapittel 2	72

Del II

Sannsynlighetsteori

3

Sannsynlighetsregning 79

3.1	Utfallsrom og hendelse	80
3.2	Sannsynligheten for en hendelse	83
3.2.1	Uniform sannsynlighetsmodell	84
3.2.2	Sannsynlighet basert på relativ frekvens	85
3.2.3	Subjektiv sannsynlighet for en hendelse	89
3.2.4	Eksempel på tallfesting av sannsynligheter	90
3.3	Innføring i sannsynlighetsregning	92
3.3.1	Litt mengdelære	92
3.3.2	Sannsynlighetsregningens aksiomer	94
3.3.3	De viktigste regnereglene	94
3.4	Betingede sannsynligheter	98
3.4.1	Regler for betingede sannsynligheter	100
3.4.2	Litt om Bayes	103
3.4.3	Hendelsestre	106
3.5	Uavhengige hendelser	109

3.5.1	Er hendelsene uavhengige?	110
3.5.2	Regler for uavhengige hendelser	111
3.6	Telleregler – kombinatorikk	114
3.6.1	Produktregelen	114
3.6.2	Urnemodellen	115
3.6.3	De viktigste tellereglene	116
3.6.4	Sannsynligheter basert på kombinatorikk	121
3.7	Bruk av datamaskin	123
3.8	Oppsummering	123
	Oppgaver til kapittel 3	124

4 Stokastiske variabler133

4.1	Stokastiske variabler	134
4.2	Diskrete sannsynlighetsmodeller	136
4.2.1	Sannsynlighetsfordeling	136
4.2.2	Forventningsverdi	143
4.2.3	Varsians og standardavvik	147
4.3	Kontinuerlige sannsynlighetsmodeller	149
4.4	Flere variabler samtidig	155
4.4.1	Simultanfordeling	156
4.4.2	Kovarians og korrelasjon	159
4.4.3	Uavhengighet	162
4.4.4	Hendelsestre	164
4.5	Oppsummering	166
	Oppgaver til kapittel 4	167

5 Vanlige sannsynlighetsmodeller173

5.1	Innledning	174
5.2	Binomisk modell	176
5.3	Hypergeometrisk fordeling	181
5.4	Geometrisk fordeling	184
5.5	Poissonfordelingen	185
5.6	Ekspontialfordelingen	190
5.7	Normalfordelingen	193
5.7.1	Normalfordelingen	194
5.7.2	Sjekk av normalitet	200
5.8	Sentralgrenseteoremet	202
5.9	Andre modeller*	207
5.9.1	Kjikkvadratfordelingen	208
5.9.2	Student t -fordelingen	209
5.9.3	Fisher F -fordelingen	211
5.9.4	Multinomisk fordeling	212
5.9.5	Binormalfordelingen	214

5.10	Bruk av datamaskinen	215
5.10.1	Excel	215
5.10.2	Minitab	216
5.10.3	SPSS	216
5.10.4	R	217
5.11	Oppsummering	217
	Oppgaver til kapittel 5	218

Del III Grunnleggende statistikk

6 Estimering og hypotesetesting 227

6.1	Et innledningseksempel	228
6.1.1	Hvilken vraksannsynlighet har det nye utstyret?	229
6.1.2	Hvor sikkert er estimatet?	230
6.1.3	Er det nye utstyret bedre enn det gamle?	231
6.1.4	Bør AluProd kjøpe det nye utstyret?	234
6.2	Punktestimering	234
6.2.1	Basisteori	234
6.2.2	Punktestimat for forventningsverdien μ	237
6.2.3	Punktestimat for standardavviket σ	239
6.2.4	Punktestimat for sannsynligheten p	241
6.2.5	Punktestimat for raten λ	242
6.2.6	Stratifisering*	243
6.3	Konfidensintervaller	244
6.3.1	Teori	245
6.3.2	Konfidensintervall for μ når σ er kjent (Z-intervall)	247
6.3.3	Konfidensintervall for μ når σ er ukjent (T-intervall)	248
6.3.4	Konfidensintervall for σ	251
6.3.5	Konfidensintervall for sannsynligheten p	252
6.3.6	Konfidensintervall for raten λ	254
6.4	Hypotesetesting – metode og teori	255
6.4.1	Etablering av modell og formulering av hypotesene	255
6.4.2	Testobservator og forkastingsområdets form	257
6.4.3	Valg av akseptable feilsannsynligheter	259
6.4.4	Spesifikasjon av forkastingsområde	260
6.4.5	Styrkefunksjon	262
6.4.6	Nødvendig utvalgsstørrelse	265
6.4.7	Testens konklusjon	266
6.4.8	Signifikanssannsynlighet (p -verdi)	268
6.4.9	Sammenheng med konfidensintervaller	269
6.5	Hypotesetesting i praksis	269
6.5.1	Hypotesetest av μ når σ er kjent (Z-test)	269
6.5.2	Hypotesetest av μ når σ er ukjent (T-test)	271
6.5.3	Hypotesetest av sannsynligheten p	273

6.6	Empirisk fordelingsfunksjon*	275
6.7	Bruk av datamaskinen	277
6.7.1	Excel	277
6.7.2	Minitab	277
6.7.3	SPSS	277
6.7.4	R	278
6.8	Oppsummering	278
	Oppgaver til kapittel 6	278

Del IV Statistisk analyse

7 Analyse av sammenhenger 287

7.1	Innledning	288
7.2	Korrelasjon	290
7.3	Enkel lineær regresjon	293
7.3.1	Regresjonsmodellen	293
7.3.2	Regresjonslinjen	294
7.3.3	Hvor god er modellen?	297
7.3.4	Noen statistiske egenskaper ved regresjonsmodeller	298
7.3.5	Er det en sammenheng mellom x og y ?	299
7.3.6	Hvilken respons kan vi forvente?	301
7.3.7	Hvilke observasjoner er sannsynlige?	302
7.3.8	Kontroll av forutsetningene	303
7.3.9	Vektet regresjon*	305
7.4	Multipel lineær regresjon	306
7.5	Ikke-lineær regresjon	310
7.5.1	Vanlig ikke-lineær regresjon	311
7.5.2	Polynomisk regresjon	313
7.6	Tidsrekkeanalyse – prognoser	315
7.6.1	Trend	317
7.6.2	Sesongvariasjon	318
7.6.3	Tilfeldig variasjon	319
7.6.4	Prognose	319
7.6.5	Mer avanserte modeller*	320
7.7	Bruk av datamaskinen	322
7.7.1	Excel	322
7.7.2	Minitab	323
7.7.3	SPSS	324
7.7.4	R	325
7.8	Oppsummering	325
	Oppgaver til kapittel 7	326

8	Sammenligning av grupper.....	333
8.1	Innledning	334
8.1.1	Hvor mange grupper og hvor mange observasjoner?	334
8.1.2	Randomisering	334
8.1.3	Paring og blokkdeling	335
8.1.4	Klyngeutvalg	337
8.1.5	Hva slags data har vi?	337
8.2	Litt sammenligningsfilosofi	338
8.3	<i>T</i> -test for to grupper	340
8.3.1	Uparet <i>T</i> -test	340
8.3.2	Paret <i>T</i> -test	344
8.3.3	Vurdering av paret og uparet test*	346
8.4	Varsiansanalyse for flere grupper	346
8.4.1	Noen illustrasjoner	346
8.4.2	Enveis varsiansanalyse	348
8.4.3	Samtidige konfidensintervaller	351
8.4.4	Kontroll av forutsetninger	351
8.4.5	Toveis varsiansanalyse*	352
8.4.6	Litt forsøksplanlegging*	356
8.5	Ikke-parametriske tester	357
8.5.1	Uparet Mann–Whitney–Wilcoxon-test	357
8.5.2	Paret Wilcoxon-test	360
8.5.3	Paret fortegnstest	362
8.5.4	Kruskal–Wallis’ test av flere grupper*	363
8.6	Analyse av kategoriske krysstabeller	364
8.6.1	Kjikkvadrattest	365
8.6.2	Simpsons paradoks	367
8.6.3	Modelltest	368
8.7	Bruk av datamaskinen	370
8.7.1	Excel	370
8.7.2	Minitab	371
8.7.3	SPSS	372
8.7.4	R	373
8.8	Oppsummering	374
	Oppgaver til kapittel 8	374

9 Spørreundersøkelse og datainnsamling..... 381

- 9.1 Innledning 382
- 9.2 Utforming av spørreskjema 383
- 9.3 Opplegg for innsamling 385
- 9.4 Organisering av data 389
- 9.5 Analyse av data og presentasjon av resultater 389
- 9.6 Praktisk datainnsamling og analyse 393
 - 9.6.1 Spørreundersøkelse om arealbruk 393
 - 9.6.2 Analyse av leiepriser i boligmarkedet 395
- 9.7 Fallgruver 399
- 9.8 Oppsummering 400
- Oppgaver til kapittel 9 401

10 Big Data og digitalisering..... 403

- 10.1 Innledning 404
 - 10.1.1 Dataforskeren 404
 - 10.1.2 Litt om maskinlæring 405
 - 10.1.3 Maskinlæring som konkurransesport 410
- 10.2 Organisering av data 411
 - 10.2.1 Behandling av store datamengder 411
 - 10.2.2 Tilrettelegging for analyse 413
- 10.3 Utforsking og visualisering av data 414
- 10.4 Maskinlæring og dataanalyse 419
 - 10.4.1 Regresjon 419
 - 10.4.2 Klassifikasjon 423
 - 10.4.3 Gruppering (clustering) 426
- 10.5 Fallgruver 428
- 10.6 Oppsummering 430
- Oppgaver til kapittel 10 431

Del V Andre anvendelser**11** Statistisk kvalitetsstyring435

- 11.1 Kontrolldiagrammer 436
 - 11.1.1 Shewharts $\bar{x}$ -diagram 439
 - 11.1.2 s -diagram for spredning 445
 - 11.1.3 p -diagram for defektandel 447
 - 11.1.4 Andre kontrolldiagrammer 448
- 11.2 Kapabilitet 449
 - 11.2.1 Toleransegrenser og kapabilitetsindeks 449
 - 11.2.2 Fortolkning av kapabilitetsindeks 451
- 11.3 Akseptansk kontroll 453
 - 11.3.1 Kontrollens karakteristik 454
 - 11.3.2 Akseptansk kontroll basert på defektandel 456
 - 11.3.3 Akseptansk kontroll basert på målinger 457
- 11.4 Bruk av datamaskinen 460
 - 11.4.1 Excel 460
 - 11.4.2 Minitab 460
 - 11.4.3 SPSS 461
 - 11.4.4 R 461
- 11.5 Oppsummering 461
 - Oppgaver til kapittel 11 462

12 Simulering465

- 12.1 Simulering av tilfeldige tall 466
- 12.2 Simulering av en stokastisk variabel 468
- 12.3 Simulering med datamaskinen 471
 - 12.3.1 Excel 471
 - 12.3.2 Minitab 471
 - 12.3.3 SPSS 472
 - 12.3.4 R 472
 - 12.3.5 Noen eksempler 473
- 12.4 Simulering av flere stokastiske variabler 475
 - 12.4.1 Uavhengige variabler 476
 - 12.4.2 Avhengige variabler 477
- 12.5 Simulering av funksjoner 479
- 12.6 Simulering av integral* 483
- 12.7 Oppsummering 485
 - Oppgaver til kapittel 12 486

A	Eksamensoppgaver	493
B	Ekstra bevis	525
	B.1 Bevis til kapittel 3	
	B.2 Bevis til kapittel 4	
	B.3 Bevis til kapittel 5	
	B.4 Bevis til kapittel 6	
	B.5 Bevis til kapittel 7	
	B.6 Bevis til kapittel 8	
	B.7 Bevis til kapittel 12	
C	Standard statistikkfunksjoner i Excel.....	538
	C.1 Standard statistikkfunksjoner i Excel	538
	C.2 Statistikkfunksjoner under dataanalyseverktøyet i Excel	542
D	Fasit	545
E	Statistiske tabeller.....	574
	E.1 Kumulativ binomisk sannsynlighet	575
	E.2 Kumulativ poissonfordeling	576
	E.3 Kumulativ standardnormalfordeling	577
	E.4 Standardnormalfordelingens kvantiltabell	578
	E.5 t -fordelingens kvantiltabell	579
	E.6 Kjikvadratfordelingens kvantiltabell	580
F	Stikkord	581

Del I

Innledning

1

Hva er statistikk?

- 1.1 Bokas innhold
- 1.2 Noen viktige begreper
- 1.3 Modeller
- 1.4 Innsamling av data

Moderne statistikk er noe annet enn de kjedelige tallene og tabellene du kanskje tenker på. Sannsynlighetsteorien ble utviklet på 1600–1700-tallet for å analysere vinnersannsynligheten i forskjellige pengespill. Faget har siden utviklet seg sterkt, og den generelle teorien har fått en rekke anvendelser i naturvitenskap, teknikk, medisin, økonomi og samfunnsvitenskap.

1.1

Bokas innhold

Vi bruker statistikk som hjelpemiddel når vi skal fatte beslutninger i situasjoner der noe er uforutsigbart eller tilfeldig. I en komplisert verden er det ofte slik: Vi vet om *flere mulige konsekvenser*, men vi vet *ikke sikkert* hvilken konsekvens vår beslutning vil få. Beslutninger *må* fattes selv om usikkerheten er til stede. Gjennom statistikkfaget skal vi se hvordan vi kan fatte gode beslutninger i situasjoner hvor noe er variabelt eller tilfeldig.

1.1.1 Noen eksempler

Som en liten smakebit på hva statistikkfaget dreier seg om, skal vi kort beskrive noen eksempler der faget kan være til stor nytte.

- **Legemiddeltesting.** Før nye legemidler slippes ut på markedet, må produsenten bevise at middelet virker, og at det har begrensede bivirkninger. Det blir derfor gjennomført kontrollerte forsøk, først på dyr og deretter på mennesker som har sagt seg villige til å delta i forsøket. Noen av personene får en narrepille (placebopille) slik at man kan sammenligne behandlingseffekten hos dem som fikk reell behandling, og hos dem som fikk en narrepille. Hvis legemiddelet gir gode resultater, er det bare tilfeldig, eller skyldes det faktisk at medisinen er god? Statistiske metoder kan hjelpe legene til å planlegge et godt forsøk og til å analysere resultatene etterpå.
- **Økonomiske risikovurderinger.** Svært mange bedrifter kjennetegnes av at både utgifter og inntekter er usikre: Råvareprisene varierer, rentekostnadene varierer, produksjonskostnadene er usikre, markedsprisen for sluttproduktet varierer, og man vet ikke hvor mange produkter man klarer å selge. For å sette opp et budsjett må bedriften anslå disse viktige størrelsene. Statistikkfaget kan hjelpe til med å gi et bilde av hvilket overskudd bedriften kan forvente, og hvor stor risikoen er for å gå med underskudd.
- **Aksjeanalyse og porteføljeforvaltning.** En klassisk anvendelse av statistikk er knyttet til aksjehandel, der børskursene varierer fra minutt til minutt og genererer store mengder data om verdien av alle børsnoterte selskaper. Investorer og analytikere bruker betydelige ressurser på å analysere disse dataene på jakt etter mønstre som kan indikere om det er lurt å selge eller kjøpe. Mye av dette er fånyttet når det gjelder enkeltaksjer, fordi det er tilnærmet umulig å spå den framtidige børskursen på basis av historikken. Men når det gjelder en portefølje av flere aksjer, er statistikk og sannsynlighetsregning uunnværlige verktøy som kan bidra til optimal sammensetning av en portefølje i forhold til hvor stor risiko investoren er villig til å ta.

- **Teknisk risikoanalyse.** Ved dimensjonering av bygninger (oljeplattformer, broer, boligblokker, demninger) stilles man hele tiden overfor spørsmålet: Hvor sterke skal konstruksjonene være, hvor sterk påvirkning skal de tåle? Hvor sannsynlig er det at bygningene blir utsatt for de sterkeste påkjenningene de er ment å tåle? Bør norske hus tåle jordskjelv, slik som hus i Los Angeles bør? Hva er sannsynligheten for jordskjelv i Norge i forhold til LA? Ingeniører må alltid foreta en avveining mellom hvor sterk konstruksjonen må være, og hvor mye penger man skal bruke på bygningen. I denne avveiningen kan statistikkfaget være til stor hjelp.
- **Sykdomsrisiko.** Hva er sannsynligheten for å få kreft? Er denne sannsynligheten påvirket av bestemte risikofaktorer, slik som røyking, asbest, radonstråling, kaffedrikking og lettbrus? Det fremmes en mengde resultater fra medisinsk forskning som vi må lære å forholde oss til. Ofte får slike resultater fordreide oppslag i tabloidpressen. Hva betyr egentlig resultatene? Er det sikkert at de er gyldige for hele befolkningen? Er det bare tilfeldig at forsøksresultatet ble som det ble? Slike kritiske spørsmål står helt sentralt i statistikkfaget, innenfor fagområdet *statistisk hypotesetesting*.
- **Kvalitetstesting.** Hvor mange defekte deler er det i et vareparti? Mange typer produkter lar seg ikke kvalitetsteste uten at produktene blir ødelagt. Da tvinges vi til å si noe om hele varepartiet på grunnlag av en tilfeldig utvalgt stikkprøve. Andre ganger er det mulig, men svært kostbart å teste alle produktene, slik at man ønsker å kontrollere så få produkter som mulig. Hvor mange enheter må vi teste? Hvor mange feil kan vi godta i stikkprøven og likevel akseptere varepartiet? Denne typen spørsmål er svært sentrale i mange industrier, for eksempel for produsenter og kjøpere av bildeler, flydeler og elektroniske komponenter. Sannsynlighetsregningen kan hjelpe til å gi gode svar på disse viktige spørsmålene.
- **Vedlikeholdsplanlegging.** Vedlikehold utføres vanligvis for å minimere kostnader knyttet til drift og for å tilfredsstille visse minimumskrav til sikkerhet. I norsk elforsyning brukes hvert år flere milliarder kroner på å vedlikeholde kraftstasjoner og ledninger. Hvis et kraftselskap reduserer vedlikeholdskostnadene, blir det flere feil. For å bestemme riktig nivå på vedlikeholdet må kraftselskapet vite hvilke konsekvenser feilene har, og hvor ofte de skjer. Manglende kunnskap om konsekvensene har i mange tilfeller ført til for høyt vedlikeholdsnivå – for sikkerhets skyld. En lignende problemstilling har man knyttet til *forsikring*. Vi vet hva forsikringen koster, men vi vet ikke om vi får noen nytte av den.
- **Lagerstyring.** Det koster penger å ha varer på lager, ved at kapital og areal bindes opp. Men det koster også penger å ikke ha varene på lager når man trenger dem. Det gjelder enten lageret inneholder råvarer og reservedeler til egen produksjon, eller det inneholder salgsvarer som skal leveres til kundene. Ingen vet nøyaktig når reservedelene trengs, eller nøyaktig hvor mange varer kundene vil etterspørre i neste uke. Kunnskap om kostnader og leveringstider kan sammen med statistiske data lede fram til regler for optimalt lagerhold.

- **Prosjektstyring.** Et prosjekt kan deles opp i delprosjekter. Vanligvis er det lettere å anslå varighet og kostnad for delprosjektene enn for selve hovedprosjektet. Basert på slike anslag kan sannsynlighetsregningen hjelpe oss å anslå varighet og kostnad for hele prosjektet. Et prosjekt trenger ikke være gigantisk stort for at slike vurderinger skal være nyttige.
- **Kunstig intelligens.** Bedrifter som Facebook, Amazon og Netflix samler store datamengder om sine brukere og har utviklet algoritmer som gir god innsikt i brukernes preferanser. De vet hva kundene liker, og hvordan de bør markedsføre produktene for å få maksimalt salgsvolum. Avansert statistikk er deres viktigste verktøy.
- **Valg av servicenivå.** Mange tjenesteytende virksomheter håndterer en uforutsigbar kundemasse. Man vet ikke på forhånd hvor mange kunder som vil melde sin ankomst neste dag. Problemstillingen er aktuell for matbutikken, banken eller storbedriften som skal velge antall kassepersonale, bankfunksjonærer og sentralbordbetjenter. Er det for få ansatte, blir køene for lange og irritasjonen stor. Er det for mange ansatte, blir kostnadene unødige store. Med kunnskap om typiske mønstre i kundenes ankomsttider kan man finne fram til et riktig servicenivå. Problemstillingen er like aktuell for nettbutikker som må dimensjonere kapasiteten i sin varedistribusjon. (Det er opplagt mange bedrifter som ikke har gjort leksen sin her!)
- **Juridiske spørsmål.** Et hovedprinsipp i retten er at man er uskyldig inntil det motsatte er bevist. All rimelig tvil skal komme tiltalte til gode. Med sannsynlighetsregning kan man i noen saker beregne hvor stor tvilen er, basert på visse antakelser og observasjoner. I den såkalte Landås-saken ble en hjelpepleier tiltalt for drap på en rekke pleiepasienter. Det var spesielt mange som døde når denne hjelpepleieren var på jobb. Retten besluttet likevel at hjelpepleieren skulle frifinnes, fordi den mente at det kunne være tilfeldig at så mange var døde på akkurat disse vaktene. (Retten baserte sin beslutning på sannsynlighetsberegninger presentert av forsvarerne.)
- **Spørreundersøkelse.** De fleste av oss har opplevd å få tilsendt et spørreskjema eller bli oppringt fra et meningsmålingsinstitutt. Spørreundersøkelser kan være nyttige for å kartlegge forhold som ikke lar seg registrere eller måle med vanlige «måleinstrumenter». Slike målinger brukes ofte for å kartlegge hvor tilfredse kunder og medarbeidere er, oppslutningen til politiske partier, preferanser til ulike produkter og generelle holdninger til aktuelle problemstillinger. Svært ofte er det unødvendig og altfor kostbart å spørre hele populasjonen av medarbeidere, kunder og andre involverte. Spørsmålene stilles bare til et *utvalg* fra populasjonen. Resultatene av undersøkelsen danner basis for beslutninger om tiltak som skal iverksettes (av bedriftsledelsen, partiledelsen, markedssjefen). Statistikkfaget gir svar på hvor mange som bør spørres, og hvor sikre resultatene fra undersøkelsen er.

- **Optimalisering av produksjonsprosess.** I prosessindustrien foregår produksjonen vanligvis ved at en mengde ulike råstoffer blandes sammen. Kvaliteten på sluttproduktet avgjøres av blandingsforholdene mellom disse stoffene, rekkefølgen på sammenblandingen, temperaturer, trykk, luftfuktighet osv. I kompliserte produksjonsprosesser er det ofte slik at man ikke har full kontroll på alle inputparameterne (de varierer). De produksjonsansvarlige har muligheten til å endre blandingsforhold, temperatur og trykk, men kjenner ofte ikke konsekvensene av slike endringer. For å lære mer om produksjonsprosessen er det vanlig å foreta prøveproduksjoner. Statistikkfaget kan hjelpe til med å planlegge slike prøver og analysere hvilke endringer som vil føre til bedre produktkvalitet.
- **Mønstergjenkjenning.** En tomflaskeautomat tar et bilde av flasken som settes inn i automaten. Ingen av bildene blir helt like, siden alle flaskene plasseres litt forskjellig i automaten. Automaten må deretter beslutte hva slags flaske dette er. I oljeselskapene analyserer man geologiske testdata fra borehull i Nordsjøen. Hver testmåling «avbildes» med sine karakteristikker. Ingen karakteristikker er helt like. Noen steder er det mye olje, andre steder er det lite eller ingenting. Geologene må gi en anbefaling om hvor de tror det kan finnes olje (helst mye). Oljeselskapet må deretter ta en beslutning om hvor det skal søke konsesjoner og foreta nye borer. Google og Facebook gjør noe lignende når de sorterer bildene dine i forhold til hvem som er avbildet.
- **Veddemål.** Er du villig til å satse 700 kroner på at Norge kvalifiserer seg til fotball-VM i 2026 hvis du kan vinne 1000 kroner dersom du får rett? Statistikken taler ikke til Norges fordel. Hvor stor sjanse har Norge for å kvalifisere seg? Det kan være vanskelig å svare på dette spørsmålet direkte, men basert på kampoppsettet i Norges kvalifikasjonsgruppe kan du anslå hvilke kampresultater du har størst tro på. Sannsynlighetsregningen kan hjelpe deg til å kombinere dine gjetninger på en ryddig måte, slik at du får fram en totalgjetning på hvilke sjanser Norge har.
- **Spill.** Mange spill er slik at spillerne må ta beslutninger underveis. I poker må du vurdere om du skal bytte ut kort. *Du vet hva du har, men ikke hva du får.* Ved ruletten må du velge om du skal satse på rødt eller sort. Du kan observere rulettens resultater i årevis; likevel vet du ikke hvilken farge som blir resultatet i neste runde. I yatzy må du underveis bestemme deg for om du skal samle til «hus» eller om du skal prøve å få «yatzy». Det siste er vanskeligere, men gir flere poeng. Du kan ikke vite om du vil lykkes. I alle disse situasjonene kan sannsynlighetsregningen hjelpe deg til å ta gode beslutninger.

1.1.2 Historie

Statistikk har lenge dreid seg om å samle inn store tallmengder og stille disse opp i tabeller, særlig med tanke på å belyse ulike samfunnsmessige forhold. Slik statistikk har vært et viktig hjelpemiddel for politikere og andre makthavere i tusenvis av år. Tenk bare på Josef og Maria som reiste til Betlehem for å skrive seg inn i manntallet. Alle siviliserte samfunn har etter hvert bygd opp et stort apparat for å samle inn slike opplysninger.

Arven fra romerne videreføres i Norge av Statistisk sentralbyrå (SSB). *Statistisk årbok* inneholder om lag 500 tabeller som beskriver naturressurser, sosiodemografi, sosioøkonomi, næringsøkonomi, samfunnsøkonomi og samfunnsorganisatoriske emner. Det finnes for eksempel en tabell som heter «*Hoteller og liknande overnattingsbedrifter. Overnattingar, etter gjestene sitt bustadland*». Denne tabellen baserer seg på registreringsskjemaer som samtlige hoteller, pensjonater og campingplasser må fylle ut. I skjemaet skal det opplyses om antall overnattingsgjester hver dag i sesongen, fordelt på nasjonalitet. Kan du forestille deg hvor mye arbeid som totalt sett går med til å lage denne statistikken? Årlig er det om lag 33 millioner overnattinger på ca. 900 overnattingssteder. Hovedutfordringen i denne typen datainnsamling er å motivere de som skal fylle ut registreringsskjemaene til å gjøre jobben samvittighetsfullt og riktig. Denne typen «gammeldags» tallinnsamling vil ikke være tema i denne boka.

Moderne statistikk, slik vi definerer statistikkfaget i denne boka, ble først tatt i bruk i samfunnsvitenskap, landbruksforskning og medisinsk forskning. Carl F. Gauss, Francis Galton og andre oppdaget systematiske mønstre i det som mange trodde var et ubeskrivelig kaos. Deres ideer var svært nyttige når ulike fenomener skulle beskrives. Ronald A. Fisher var blant de første som innså hvor viktig det var å planlegge eksperimenter på en systematisk måte, slik at det ble mulig å trekke gode konklusjoner etterpå. Utviklingen fram til i dag har resultert i en mengde statistiske metoder som kan benyttes i svært mange sammenhenger. I denne boka skal vi lære om de viktigste metodene, men du skal vite at det finnes flere. Mange flere.

Sannsynlighetsregningen har også en lang forhistorie, nær knyttet til terning- og kortspill. Navn som Pierre de Fermat, Blaise Pascal, Jacques Bernoulli og Thomas Bayes står sentralt i den tidlige utviklingen av det som er dagens matematisk baserte sannsynlighetsregning. Selv om utgangspunktet var nært knyttet til spill, viste det seg raskt at sannsynlighetsregningen også kunne brukes på andre områder. I Norge utga Axel Sophus Guldberg (1838–1913) boka «*Om Sandsynlighetsregningen og dens Anvendelse paa Hazardspil og Forsikringsvæsen*» så tidlig som i 1873. I denne boka ser han sannsynlighetsregningen som «*Kilden til en ny Viden-skab, som kunde erholde Anvendelser, hvorom man neppe skulde have nogen Anelse, og som griber ind i Forhold som synes at ligge udenfor al menneskelig Beregning*». Gjennom bokas mange eksempler skal vi se at Guldberg fikk rett i sin spådom: Sannsynlighetsregningen har usannsynlig mange anvendelsesområder.

På 2000-tallet har statistikkfaget opplevd en renessanse. Ny sensorteknologi og stadige framskritt innen mobile kommunikasjonsløsninger, har ført til en eksplosjon av tilgjengelige data. Både bedrifter og offentlige virksomheter samler store datamengder, som kan brukes til et mangfold av beslutninger. «Data er det nye gullet» eller den nye «oljen». New York-børsen beviser langt på vei dette. De siste års vinnere er nesten utelukkende selskaper som sitter på store, eksklusive datamengder som gir grunnlag for framtidig profitt. Velkjente eksempler er Google, Apple, Facebook og Amazon. Dette har gitt høy etterspørsel etter fageksperter som kan brukes til analyse av store datasett, og dyktige statistikere verdsettes i stadig flere bransjer og stillingskategorier.

1.1.3 Bokas oppbygning

Etter å ha lest denne boka bør du beherske de viktigste statistiske metodene slik at du kan bruke dem selv. Du skal også være i stand til å tolke statistiske opplysninger. Boka fokuserer mye på ideer og intuisjon framfor på matematikk og beviser. Hvert kapittel innledes med tekst og illustrasjoner som du kan forstå uten å måtte sette deg inn i matematiske formler.

Statistikkfaget inneholder både *statistikk* og *sannsynlighetsregning*. Dette er egentlig to ulike disipliner som griper sterkt inn i hverandre, på den måten at den ene delen ofte er en forutsetning for den andre. Vi trenger statistikk for å skaffe kunnskap til bruk i sannsynlighetsberegningene. Og vi trenger sannsynlighetsregningen for å lage gode modeller som vi bruker når vi gjør statistiske analyser. I denne boka vil vi veksle litt fram og tilbake mellom statistikk og sannsynlighetsregning.

Boka består av fem deler som bør leses i den rekkefølgen de er oppført. Mye av stoffet i del I (og noe av del II og del III) inngår i matematikkpensum for den videregående skolen. Selv om du har vært borti dette stoffet før, bør du lese igjennom stoffet for å sikre deg at definisjoner, begreper og symboler betyr det som du tror.

- I. **Innledning.** Du har allerede begynt å lese innledningen, som er ment å være en myk start på statistikkfaget. Viktige begreper skal defineres, og vi skal se hvordan forsøksdata kan presenteres.
- II. **Sannsynlighetsteori.** I denne delen skal vi for alvor gå løs på sannsynlighetsregningen, med fokus på tilfeldighet og hvilke mønstre som tross alt finnes i denne tilfeldigheten. Vi skal lære å regne ut sannsynligheten for at en hendelse vil inntreffe. Vi skal lære om tilfeldige variabler og bli kjent med de vanligste sannsynlighetsfordelingene.
- III. **Grunnleggende statistikk.** Her skal vi lære om hvordan vi kan trekke konklusjoner om hele grupper basert på opplysninger om deler av gruppen. Vi skal se hvordan statistikk kan benyttes for å teste om hypoteser er sanne.
- IV. **Statistisk analyse.** I boka tre første deler lærer du en del grunnleggende teknikker og begreper som kan brukes i ulike sammenhenger. I del IV skal vi ta disse teknikkene i bruk på nye områder. Vi skal lære å *analysere sammenhenger* mellom ulike tallstørrelser (er det en sammenheng mellom utetemperaturen og solgt mengde iskrem?) og å *sammenligne ulike grupper* (virker medisin A bedre enn medisin B?). Her skal vi også lære om spørreundersøkelser og dataanalyse, og analyse av store datamengder.
- V. **Andre anvendelser.** Vi skal lære om statistisk *kvalitetskontroll*. Til slutt skal vi se hvordan vi kan *simulere* tilfeldigheter ved hjelp av en datamaskin.

Den faglige dybden blir ikke like stor i de to siste delene av boka som i de foregående. Her er hovedhensikten å gi deg en smakebit på hva statistikken kan brukes til. Kanskje du da får lyst til å lære mer? Underveis i boka finner du enkelte avsnitt som er merket med en stjerne (*). Dette er stoff som normalt ikke vil inngå i pensum.