



FOLDNES | GRØNNEBERG | HERMANSEN | WELLÉN

STATISTIKK OG DATAANALYSE

En moderne innføring

2. UTGAVE

CAPPELEN DAMM AKADEMISK

Statistikk og dataanalyse

**Njål Foldnes,
Steffen Grønneberg,
Gudmund Horn Hermansen og
Einar Christopher Wellén**

Statistikk og dataanalyse

En moderne innføring

2. utgave

CAPPELEN DAMM AKADEMISK

Forord

Statistikk som et verktøy til å forstå verden

Velkommen til statistikkfaget! Her skal du lære om hvordan du kan analysere data og bruke det til å ta bedre avgjørelser.

Vi får stadig tilgang til mer og mer data, og etterspørselen etter folk som kan tolke kvantitativ informasjon, er økende.

Denne boken er en introduksjon til hvordan man kan bruke tall og data til å finne ut hvordan verden er skrudd sammen.

Verden blir mer og mer kvantitativ – det samles inn mer og mer digitalisert informasjon. Ta for eksempel Wal-Mart, en amerikansk kjøpesenterkjede med en årlig omsetning på størrelse med Norges brutto nasjonalprodukt. Wal-Mart lagrer informasjon om mer enn én million handler hver eneste time! Alle slags organisasjoner lagrer data om sin virksomhet, og de beste bruker dette for alt det er verdt for å maksimere verdiskapingen sin. Her hjemme samler Statistisk sentralbyrå inn store mengder kvantitative data fra alle samfunnsområder. Denne informasjonen blir brukt til å analysere sammenhenger og trender. Dette er viktig informasjon til politikere og andre beslutningstakere.

Med moderne datateknologi kan du selv gjøre kvantitativt arbeid og oppdage nye sammenhenger. Noen få (men smarte) tastevalg i regneark som Excel eller statistisk programvare som JMP, R eller SPSS kan gi ny informasjon som kan hjelpe din bedrift. Etterspørselen etter folk som kan tolke og «knuse» tall kommer garantert til å vokse også i årene framover.

Vi som underviser i statistikk vet godt at mange studenter har et anstrengt forhold til tall og matematikk. Mange studenter i samfunnsfagene føler seg distansert fra tall og regning, og mange har dessverre negative erfaringer fra skolen. Om dette skulle gjelde deg, også – prøv likevel å være åpen! I denne boken kreves det ingen ekstraordinære matematikkferdigheter. Likevel viker vi på ingen måte unna dette viktige faget, for målet vil alltid være å forstå dataene, og da trenger vi noe matematikk, rett og slett.

I denne boken legger vi vekt på tenkning og forståelse. Du vil lære å tolke resultatene av statistisk analyse i den sammenhengen du vil bruke dem i.

Kvantitativ informasjon: Informasjon som omhandler tall og målinger. I motsetning til kvalitativ informasjon. Alt som kan måles og telles er kvantitativ informasjon. Kvantitative metoder er altså matematiske og statistiske teknikker som hjelper oss å forstå tallmateriale.

Formler og formelle statistiske prosedyrer vil være en del av kurset, men til syvende og sist er det din forståelse av analysen som skal være grunnlaget for avgjørelser du tar.

Enkelte overskrifter i boken er merket med (*). Dette betyr at teksten som følger er fordypningsstoff, og ikke behøver å leses grundig ved første gjennomgang av boken.

Slik er boken organisert

Boken har fire hovedbolker. Det kan være greit å få en oversikt over disse bolkene før man begynner å lese. De første tre bolkene fokuserer på statistisk analyse av én variabel mens den siste og fjerde bolken handler om analyse av to eller flere variabler.

Vi starter i **bolke 1** med å introdusere hovedideene i en statistisk analyse. Hovedmålet med boken er å få deg til å utføre slike analyser på egenhånd. Vi presenterer en skjematisk oversikt over det vi kaller de fire elementene i en statistisk analyse. Dette gir et godt utgangspunkt for å lese boken. Du vil også bli gjort kjent med de mest grunnleggende begrepene i statistikk og herunder bli kjent med de ulike variabeltypene.

Vi tar deg deretter gjennom de tre første stegene i en statistisk analyse. Det første steget er å definere problemstillingen du ønsker å løse. Deretter lærer vi deg å trekke et representativt utvalg (andre steget i en statistisk analyse). Til slutt tar vi deg gjennom det tredje steget i en statistisk analyse, som handler om å beskrive data fra utvalget med nøkkeltall (f.eks. gjennomsnitt) og grafer (f.eks. histogram).

Det siste steget er inferens, som betyr hvordan vi kan gå fra informasjon om utvalget til å si noe om hele populasjonen utvalget er trukket fra. Dette steget kommer først i bolke 3, og er det mest sofistikerte steget. For å kunne gjøre og forstå inferens må vi først studere sannsynlighet. Dette gjøres i bolke 2 av boken.

Bolke 2 tar for seg sannsynlighetsregning og består av regneregler og bruk av matematiske teknikker. Her er det altså mange formler. Vi lærer regler om sannsynligheten av hendelser, og vi introduserer begrepet tilfeldig variabel, og dens sannsynlighetsfordeling. Kunnskap om sannsynlighetsregning er nødvendig for å kunne gå fra steg 3 (beskrivende statistikk) til steg 4 (inferens) i en statistisk analyse, men vi lærer også sannsynlighetsregning fordi vi ofte trenger å bruke sannsynlighetsregning direkte på problemstillinger innen økonomi, finans, markedsføring og andre fag.

Bolk 3 handler om inferens: hvordan kan vi gå fra informasjon om utvalget til å si noe om hele populasjonen som utvalget er trukket fra. Inferens er steg 4 og det siste steget i den statistiske analysen. Når vi utfører inferens svarer vi på problemstillingen fra steg 1 i den statistiske analysen. For å kunne utføre inferens kreves kjennskap til sannsynlighetsregning (broen mellom steg 3 og steg 4 i en statistisk analyse) og bruk av nøkkeltall fra den beskrivende analysen (steg 3 i en statistisk analyse). Inferens gjøres ved å beregne konfidensintervaller og utføre hypotesetester.

I **bolk 4** lærer vi å analysere sammenhengen mellom to variable. Vi studerer samvariasjon mellom to variable via simultane sannsynlighetsfordelinger, regresjon og khikvadrattester. Kan du regresjon, har du et godt grunnlag for å studere mer avanserte statistiske modeller i videregående kurs.

Kapitteloversikt

BOLK 1 INTRODUKSJON TIL STATISTIKK

- Kapittel 1 Introduksjon til statistikk 17
- Kapittel 2 Variabler 36
- Kapittel 3 Element 1 i en statistisk analyse 45
- Kapittel 4 Element 2 i en statistisk analyse 51
- Kapittel 5 Element 3 i en statistisk analyse 65

BOLK 2 SANNSYNLIGHETSTEORI. BROEN MELLOM ELEMENT 3 OG ELEMENT 4 I EN STATISTISK ANALYSE

- Kapittel 6 Grunnleggende sannsynlighetsteori 95
- Kapittel 7 Generell sannsynlighetsregning 125
- Kapittel 8 Tilfeldige variable 154
- Kapittel 9 Forventning og varians til tilfeldige variabler 177
- Kapittel 10 Diskrete tilfeldige variable 205
- Kapittel 11 Kontinuerlige tilfeldige variabler 223
- Kapittel 12 Utvalgsfordelinger og sentralgrenseteoremet 260

BOLK 3 ELEMENT 4 I EN STATISTISK ANALYSE - INFERENS (KONFIDENSINTERVALLER OG HYPOTESETESTER)

- Kapittel 13 En introduksjon til inferens 293
- Kapittel 14 Konfidensintervaller 314
- Kapittel 15 Hypotesetesting: Grunnleggende teori 329
- Kapittel 16 Inferens for et gjennomsnitt 398
- Kapittel 17 Inferens for en andel 418
- Kapittel 18 Inferens for å sammenlikne to grupper 434
- Kapittel 19 Å sammenlikne andeler for en kategorisk variabel – khikvadrattest for sannsynligheter 457

BOLK 4 SAMVARIASJON. LINEÆR REGRESJON

- Kapittel 20 Samvariasjon mellom to variable 471
- Kapittel 21 En introduksjon til simultane sannsynlighetsfordelinger 509
- Kapittel 22 Enkel regresjon 536
- Kapittel 23 Samvariasjon for to kategoriske variable 570

Vedlegg 583

Stikkordregister 592

Innhold

BOLK 1 INTRODUKSJON TIL STATISTIKK

KAPITTEL 1

Introduksjon til statistikk 17

- 1.1** Hva er statistikk og hvorfor er det viktig 18
- 1.2** Variabler: Envariabelstatistikk og flervariabelstatistikk 19
- 1.3** Elementene i en statistisk analyse – boken oppsummert i et bilde 19
- 1.4** Datamaskinens rolle i statistisk analyse 27
- 1.5** Veien videre etter denne boken 30
- 1.6** Oppsummering av begreper og formler 32
- 1.7** Oppgaver 33
- 1.8** Oppgaveløsninger 35

KAPITTEL 2

Variabler 36

- 2.1** Variabler 37
- 2.2** Kategoriske og kvantitative variabler 37
- 2.3** Målenivå 38
- 2.4** Variabler i gråsonen mellom kategorisk og kvantitativ 40
- 2.5** Oppsummering av begreper og formler 42
- 2.6** Oppgaver 43
- 2.7** Oppgaveløsninger 44

KAPITTEL 3

Element 1 i en statistisk analyse 45

- 3.1** Definer en problemstilling 46
- 3.2** Fra påstand til problemstilling 46
- 3.3** Oppsummering 48
- 3.4** Oppgaver 49
- 3.5** Oppgaveløsninger 50

KAPITTEL 4

Element 2 i en statistisk analyse 51

- 4.1** En analogi for å trekke utvalg 52
- 4.2** Tilfeldig utvalg 53
- 4.3** Klyngeutvalg 54
- 4.4** Stratifisert utvalg 56
- 4.5** Om gode og dårlige utvalgsmetoder 59
- 4.6** Oppsummering av begreper og formler 61
- 4.7** Oppgaver 62
- 4.8** Oppgaveløsninger 64

KAPITTEL 5

Element 3 i en statistisk analyse 65

- 5.1** Bruk av grafer for å beskrive data 65
- 5.2** Bruk av tall til å oppsummere data 76
- 5.3** Oppsummering av begreper og formler 85
- 5.4** Oppgaver 86
- 5.5** Oppgaveløsninger 90

BOLK 2 SANNSYNLIGHETSTEORI. BROEN MELLOM ELEMENT 3 OG ELEMENT 4 I EN STATISTISK ANALYSE

KAPITTEL 6

Grunnleggende sannsynlighets- teori 95

- 6.1** Hvorfor trenger vi sannsynlighetsregning i statistikk? 96
- 6.2** Hva er en sannsynlighet? 97
- 6.3** Tilfeldige forsøk og utfallsrommet 101
- 6.4** Mengdelære 105
- 6.5** De første sannsynlighetsmodellene 106
- 6.6** Telling, permutasjoner og kombinatorikk 111

- 6.7 Forklaring av hvorfor «gunstige delt på mulige» holder (*) 120
- 6.8 Oppsummering av begreper og formler 121
- 6.9 Oppgaver 122
- 6.10 Oppgaveløsninger 124

KAPITTEL 7

Generell sannsynlighetsregning 125

- 7.1 Addisjonsregelen 126
- 7.2 Betinget sannsynlighet 128
- 7.3 Multiplikasjonsregelen 131
- 7.4 Uavhengighet 133
- 7.5 Loven om total sannsynlighet 137
- 7.6 Bayes' formel 140
- 7.7 En anvendelse av sannsynlighetsteori – DNA-testing (*) 142
- 7.8 En anvendelse av sannsynlighetsteori – Monty Hall (*) 144
- 7.9 Oppsummering av begreper og formler 147
- 7.10 Oppgaver 148
- 7.11 Oppgaveløsninger 151

KAPITTEL 8

Tilfeldige variable 154

- 8.1 Forskjellen på en variabel og en tilfeldig variabel 154
- 8.2 En tilfeldig variabel er en representant for en populasjon 156
- 8.3 En tilfeldig variabel har alltid en sannsynlighetsfordeling 157
- 8.4 Summetegn med indeksnotasjon 157
- 8.5 Sannsynlighetsmodeller til diskrete tilfeldige variable 159
- 8.6 Sannsynlighetsmodeller til kontinuerlige tilfeldige variable 163
- 8.7 Hvorfor $P(X = x) = 0$ for en kontinuerlig tilfeldig variabel (*) 170
- 8.8 Oppsummering av begreper og formler 171
- 8.9 Oppgaver 172
- 8.10 Oppgaveløsninger 175

KAPITTEL 9

Forventning og varians til tilfeldige variable 177

- 9.1 De store talls lov 178
- 9.2 Forventning til en tilfeldig variabel 182
- 9.3 Varians og standardavvik til en tilfeldig variabel 187
- 9.4 Hvorfor gjelder regnereglene for forventning og varians? (*) 195
- 9.5 Oppsummering av begreper og formler 198
- 9.6 Oppgaver 200
- 9.7 Oppgaveløsninger 203

KAPITTEL 10

Diskrete tilfeldige variable 205

- 10.1 Bernoulli-fordeling 205
- 10.2 Binomisk fordeling 206
- 10.3 Hypergeometrisk fordeling 211
- 10.4 Poisson-fordeling 215
- 10.5 Oppsummering av begreper og formler 218
- 10.6 Oppgaver 219
- 10.7 Oppgaveløsninger 221

KAPITTEL 11

Kontinuerlige tilfeldige variable 223

- 11.1 Normalfordelingen 223
- 11.2 Flere egenskaper ved normalfordelte variable(*) 234
- 11.3 t -fordelingen 238
- 11.4 Uniform sannsynlighetsfordeling 239
- 11.5 Eksponentialfordelingen 241
- 11.6 Kvantiler 243
- 11.7 Normalfordelingen i praksis 247
- 11.8 Oppsummering og viktigste formler 251
- 11.9 Oppgaver 253
- 11.10 Oppgaveløsninger 256

KAPITTEL 12

Utvalgsfordelinger og sentralgrenseteoremet 260

- 12.1 Å tilpasse en sannsynlighetsfordeling ved å trekke fra en tilfeldig variabel 261
- 12.2 Om utvalgsfordelingen til gjennomsnittet 270
- 12.3 Forventning og varians til en sum av tilfeldige variabler 275
- 12.4 Forventning og varians til gjennomsnittet 277
- 12.5 Sentralgrenseteoremet 279
- 12.6 Oppsummering av utvalgsfordelingen til gjennomsnittet 284
- 12.7 Oppsummering av begreper og formler 285
- 12.8 Oppgaver 287
- 12.9 Oppgaveløsninger 289

BOLK 3 ELEMENT 4 I EN STATISTISK ANALYSE - INFERENS (KONFIDENSINTERVALLER OG HYPOTSETESTER)

KAPITTEL 13

En introduksjon til inferens 293

- 13.1 Estimatorer og deres usikkerhet 294
- 13.2 Hvordan estimere en populasjonsparameter? 296
- 13.3 Estimator for populasjonsandel 298
- 13.4 Estimatorer for varians og standardavvik 299
- 13.5 Notasjon for estimatorer 301
- 13.6 Hvorfor det er nyttig å ha kjennskap til utvalgsfordelinger: Broen mellom beskrivende statistikk og inferens, et dataeksempel 303
- 13.7 Oppsummering av begreper og formler 309
- 13.8 Oppgaver med løsningsforslag 311
- 13.9 Oppgaveløsninger 313

KAPITTEL 14

Konfidensintervaller 314

- 14.1 Introduksjon 314
- 14.2 Konfidensintervaller for populasjonsgjennomsnitt – grunnleggende teori 319
- 14.3 Hva påvirker konfidensintervallet? 321
- 14.4 Mer om antagelsene som ligger bak konfidensintervallet (*) 324
- 14.5 Hvorfor $\mu - a \leq \bar{x} \leq \mu + a$ er det samme som $\bar{x} - a \leq \mu \leq \bar{x} + a$ (*) 324
- 14.6 Oppsummering av begreper og formler 326
- 14.7 Oppgaver 327
- 14.8 Oppgaveløsninger 328

KAPITTEL 15

Hypotesetesting: Grunnleggende teori 329

- 15.1 Hvorfor hypotesetesting er viktig: To praktiske eksempler 330
- 15.2 Introduksjon til tosidige hypotesetester 331
- 15.3 Tosidige hypotesetester og testobservatoren 333
- 15.4 p -verdier for tosidige hypotesetester 343
- 15.5 Utvalgsfordelingen til p -verdier for tosidige tester 346
- 15.6 To vanlige misforståelser om p -verdier 356
- 15.7 En oppsummering av tosidige hypotesetester 357
- 15.8 Ensidige hypotesetester for populasjonsgjennomsnitt 358
- 15.9 Formelle hypotesetester, beslutninger, og type I- og type II-feil 372
- 15.10 Hva skal vi velge som nullhypotese og alternativ hypotese? 382
- 15.11 Forholdet mellom konfidensintervaller og tosidige hypotesetester 384
- 15.12 Mer teori om teststyrke og utvalgsstørrelse (*) 386
- 15.13 Oppsummering av begreper og formler 390
- 15.14 Oppgaver 394
- 15.15 Oppgaveløsninger 396

KAPITTEL 16

Inferens for et gjennomsnitt 398

- 16.1 *t*-fordelingen 399
- 16.2 Er *t*-metodene robuste? 402
- 16.3 Konfidensintervall for populasjonsgjennomsnitt 403
- 16.4 Hypotesetest for et gjennomsnitt 407
- 16.5 Oppsummering av begreper og formler 412
- 16.6 Oppgaver 413
- 16.7 Oppgaveløsninger 416

KAPITTEL 17

Inferens for en andel 418

- 17.1 Utvalgsfordelingen til utvalgsandelen 419
- 17.2 Konfidensintervall for en andel 421
- 17.3 Hypotesetest for en andel 424
- 17.4 Oppsummering av begreper og formler 428
- 17.5 Oppgaver 429
- 17.6 Oppgaveløsninger 431

KAPITTEL 18

Inferens for å sammenlikne to grupper 434

- 18.1 Relaterte og uavhengige utvalg 435
- 18.2 Sammenlikne to gjennomsnitt 436
- 18.3 Sammenlikning av andeler i to grupper 444
- 18.4 Oppsummering av begreper og formler 450
- 18.5 Oppgaver 451
- 18.6 Oppgaveløsninger 454

KAPITTEL 19

Å sammenlikne andeler for en kategorisk variabel - khikvadrattest for sannsynligheter 457

- 19.1 Observerte og forventete verdier 458
- 19.2 Test for sannsynligheter 460
- 19.3 Oppsummering av begreper og formler 465
- 19.4 Oppgaver 466
- 19.5 Oppgaveløsninger 468

BOLK 4 SAMVARIASJON. LINEÆR REGRESJON

KAPITTEL 20

Samvariasjon mellom to variable 471

- 20.1 Introduksjon: Fra en til flere variabler 471
- 20.2 Observere eller eksperimentere? Mer om datainnhenting når man har mer enn én variabel 472
- 20.3 Samvariasjon mellom to kategoriske variabler 476
- 20.4 Grafer som viser samvariasjon for to variabler 478
- 20.5 Samvariasjon mellom to kvantitative variabler 484
- 20.6 Korrelasjon 485
- 20.7 Om rette linjer 491
- 20.8 Minste kvadratets metode og regresjonslinja 494
- 20.9 Tolkning og prognose 497
- 20.10 Oppsummering av begreper og formler 500
- 20.11 Oppgaver 501
- 20.12 Oppgaveløsninger 506

KAPITTEL 21

En introduksjon til simultane sannsynlighetsfordelinger 509

- 21.1 Simultanfordelinger for to tilfeldige variabler 510
- 21.2 Fordelingen og forventningen til en funksjon av to tilfeldige variabler 517
- 21.3 Samvariasjon mellom to tilfeldige variabler 521
- 21.4 Variansen til en sum av to tilfeldige variabler 525
- 21.5 Oppsummering av begreper og formler 528
- 21.6 Oppgaver 529
- 21.7 Oppgaveløsninger 532

KAPITTEL 22

Enkel regresjon 536

- 22.1** Regresjonsmodellen:
Gjennomsnittet til y avhenger av x 537
- 22.2** Når er den enkle lineære regresjonsmodellen rimelig å bruke? 539
- 22.3** Regresjonens standardfeil og standardfeilen til $\hat{\beta}_1$ 547
- 22.4** Inferens for β_0 og β_1 549
- 22.5** Prediksjonsintervall (*) 557
- 22.6** Årsakssammenheng eller skjulte variabler? 559
- 22.7** Veien videre: Multiplere lineær regresjon 561
- 22.8** Oppsummering av begreper og formler 563
- 22.9** Oppgaver 564
- 22.10** Oppgaveløsninger 568

KAPITTEL 23

Samvariasjon for to kategoriske variable 570

- 23.1** Er variablene uavhengige, eller samvarierer de? 571
- 23.2** Observerte og forventete verdier i krysstabellen 572
- 23.3** Khikvadrattesten for samvariasjon mellom kategoriske variabler 573
- 23.4** Oppsummering av begreper og formler 577
- 23.5** Oppgaver 578
- 23.6** Oppgaveløsninger 580

VEDLEGG 583**Tabell A** 584**Tabell B** 586**Tabell C** 588**Tabell D** 590**STIKKORDREGISTER** 592