

Innføring i
sannsynlighetsregning
og statistikk

Per Chr. Hagen

Innføring i sannsynlighetsregning og statistikk

8. utgave

CAPPELEN DAMM AKADEMISK

© CAPPELEN DAMM AS, Oslo, 2021

ISBN 978-82-02-72759-8

1. utgave, 1. versjon 2021

E-utgaven baserer seg på 8. trykte utgave, 978-82-02-67546-2, Cappelen Damm Akademisk 2021

Denne boken ble først utgitt i 1996 av Cappelen Akademisk Forlag.

Materialet i denne publikasjonen er omfattet av åndsverklovens bestemmelser. Uten særskilt avtale med forlaget 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.

Omslagsdesign: Kristin Berg Johnsen

Sats og ombrekking: Bøk Oslo AS

Elektronisk tilrettelegging: Bøk Oslo AS

www.cda.no

akademisk@cappelendamm.no

Cappelen Damm AS er sertifisert iht. FSC®, PEFC og Miljøfyrtårn.



Symbol for bærekraftig skogbruk



Forord

Denne boken er skrevet til et innføringsemne i sannsynlighetsregning og statistikk på 5–10 studiepoeng. Et slikt emne inngår i økonomisk/administrativ utdanning, ingeniørutdanning og i flere andre tverrfaglige utdanninger.

Boken inneholder for det første det kjernepensum som inngår i ethvert grunnkurs i statistikkfaget: beskrivende statistikk (kapittel 1), grunnleggende sannsynlighetsregning (kapittel 2), stokastiske variabler (kapittel 3 og 4), estimering og konfidensintervall (kapittel 5) og hypotesetesting (kapittel 6). Også enkel regresjonsanalyse (kapittel 7.2) og kjikvadrattester (kapittel 8) vil vel de fleste mene hører med i et grunnkurs.

Boken inneholder dessuten en del emner som dels kan supplere kjernestoffet i grunnkurset og dels kan være nyttig for studentene videre i studiet. Dette gjelder kapitlene 4.2, 5.5, 6.6, 7.1, 7.3, 9 og 10. Jeg vet at det på de forskjellige institusjonene er noe ulike oppfatninger om hvilke modeller som bør behandles i grunnkurset utover kjernepensumet, og jeg har forsøkt å få med i boken de varianter jeg regner med er aktuelle. Dette gir den enkelte emneansvarlige en viss frihet i valg av temaer.

Opplegget i boken bygger på 40 års erfaring med grunnundervisning i sannsynlighetsregning og statistikk ved Norges Teknisk Høgskole (NTH, nå NTNU) og Universitetet i Sørøst-Norge (USN, tidligere Telemark distrikthøgskole (TDH) og Høgskolen i Telemark (HiT)). Framstillingen er eksempelbasert. Når nye begreper innføres, blir forklaringen av begrepene knyttet direkte opp mot eksempler som studentene kan kjenne seg igjen i.

I starten av boken dvs. i kapittel 2 og 3 er det stort sett gitt fullstendige matematiske utledninger. Det er overkommelig, og det er etter min mening nyttig å få fram hvordan sannsynlighetsregningen er bygget opp

matematisk. Fra slutten av kapittel 3 og utover er det hverken mulig eller ønskelig å ta med all matematikken. Det ville bli for komplisert og omfattende. Herfra er matematiske utledninger er bare tatt med dersom de kan bidra til å gi økt forståelse av den formel eller sammenheng en skal fram til. Stoffet er likevel preget matematisk baserte resonnementer, forklaringer og eksempler.

Ingen kan lære sannsynlighetsregning og statistikk eller andre matematiske emner bare ved å lese en lærebok eller ved å være til stede i undervisningen. Det er gjennom egen oppgaveregning en lærer mest, og det er først når en selv jobber med oppgavene, at en får testet om begrepene og metodene er forstått.

Etter hvert kapittel er det derfor mange oppgaver som studentene kan trene på. Dette er oppgaver i de temaene som er behandlet i kapitlet, ofte eksamensoppgaver. Til slutt i boken er det tatt med noen hele eksamenssett fra Universitetet i Sørøst-Norge.

Nytt i 8. utgave

8. utgave er det gjort store endringer når det gjelder struktur, grafikk og layout. Det er lagt stor vekt på å få til en utforming som skal være lettere å lese enn før. Framfor alt skal det gå klart fram hva som er definisjoner og regler, hva som er eksempler og hva som er forklarende tekst.

Det er dessuten gjort to viktige endringer i begrepsbruken for de enkleste sannsynlighetsmodellene. Elementene i et utfallsrom Ω betegnes nå som utfall (tidligere enkeltutfall), og en delmengde av Ω som består av et eller flere utfall, betegnes som en hendelse (tidligere utfall). Se kapittel 2.1 og 2.2. Det er disse betegnelsene som er blitt vanlig i norsk språk i sannsynlighetsregning og statistikk.

Takk til Arild Brandrud Næss, Handelshøyskolen, NTNU. Han har, etter å brukt boken i noen år, kommet med en rekke forslag til forbedringer, store og små. Svært mange av disse forslagene er tatt hensyn til i 8. utgave.

Eksamenssettene bak i boka er skiftet ut med eksamenssett av nyere dato. En del nye oppgaver er også kommet til i slutten av de enkelte kapitlene.

Kapitlene i tidligere utgaver om programpakken Minitab er tatt ut. Jeg har vurdert det slik at det ikke lenger er behov for å ha med bruksanvisning på Minitab eller andre statistikkprogram i en lærebok for et innføringskurs i sannsynlighetsregning og statistikk.

Bø, mai 2021

Per Chr. Hagen

Innhold

1	Innledning	13
1.1	Hva er statistikk?	13
1.2	Bruk av modeller	16
1.3	Beskrivende statistikk	18
1.4	Bruk av dataverktøy	32
	Oppgaver	33
2	Sannsynlighetsregning	35
2.1	En enkel sannsynlighetsmodell	35
2.2	Hendelser og mengdelære	42
2.3	Regneregler for sannsynlighet	46
2.4	Kombinatorikk og utvalgsmodeller	51
2.5	Betinget sannsynlighet	62
2.6	Uavhengighet og produktmodeller	72
	Oppgaver	78
3	Diskrete stokastiske variabler	89
3.1	Hva er en stokastisk variabel?	89
3.2	Sannsynlighetsfordeling og fordelingsfunksjon	91
3.3	Forventning	97
3.4	Varians	104
3.5	Flere stokastiske variabler	114
3.6	Binomisk fordeling	120
3.7	Hypergeometrisk fordeling	126
3.8	Poissonfordeling	129
	Oppgaver	134
4	Kontinuerlige stokastiske variabler	143
4.1	Oppbygging av en kontinuerlig sannsynlighetsmodell	143
4.2	Ekspontialfordelingen	156

4.3	Normalfordelingen.....	159
4.4	Tilnærming til normalfordeling	167
	Oppgaver	174
5	Estimering	187
5.1	Innledning til den statistiske analysen.....	187
5.2	Punktestimering	190
5.3	Målemodellen.....	195
5.4	Intervallestimering.....	199
5.5	Vurdering av utvalgsstørrelsen.....	208
5.6	t-fordelingen	215
	Oppgaver	218
6	Hypotesetesting	225
6.1	Hypotesetest på et eksempel	225
6.2	Hypotesetesting i målemodellen.....	230
6.3	Hypotesetesting ved normaltilnærming	235
6.4	Styrkefunksjonen	237
6.5	Signifikanssannsynlighet (<i>P</i> -verdi).....	244
6.6	t-tester	250
6.7	Tosidig test.....	253
	Oppgaver	260
7	Korrelasjon og regresjon	273
7.1	Korrelasjon.....	273
7.2	Regresjonsanalyse.....	276
7.3	Konfidensintervall og prediksjonsintervall	287
	Oppgaver	291
8	Kjikkvadrattester og kontingenstabeller	303
8.1	Kjikkvadratfordelingen.....	303
8.2	Test av modell. Helspesifisert hypotese	305
8.3	Kontingenstabeller og uavhengighetstest.....	308
	Oppgaver	318
9	Sammenligning av to grupper	327
9.1	To grupper i målemodellen.....	327
9.2	To grupper i binomisk og i hypergeometrisk modell	334
	Oppgaver	343

10 Ikke-parametriske metoder	355
10.1 Parametriske og ikke-parametriske modeller	355
10.2 Tegn testen	357
10.3 Wilcoxons ett-utvalgstest	361
10.4 Wilcoxons to-utvalgstest	366
Oppgaver	370
Eksamensoppgaver	377
Eksamen USN høst 2017	377
Eksamen USN vår 2018	379
Eksamen USN høst 2018	382
Eksamen USN vår 2019	385
Eksamen USN høst 2019	388
Eksamen USN vår 2020	390
Utvalgte svar på oppgavene	395
Eksamen USN høst 2017	405
Eksamen USN vår 2018	405
Eksamen USN høst 2018	406
Eksamen USN vår 2019	406
Eksamen USN høst 2019	407
Eksamen USN vår 2020	407
Tillegg	408
Greske bokstaver og summetegn	408
Tabell over $N(0, 1)$ -fordelingen	410
Tabell over t -fordelingen	412
Tabell over kjikvadratfordelingen	413
Stikkordregister	414

1

Innledning

1.1 Hva er statistikk?

De fleste av oss mottar hver eneste dag statistisk informasjon både fra aviser, radio og TV og fra bl.a. faglitteratur.

Eksempel på dette kan være:

- resultat av meningsmålinger om oppslutningen om de politiske partiene og om andre aktuelle samfunnsspørsmål,
- opplysninger om utviklingen på arbeidsmarkedet, om produksjonen i industrien eller i landbruket,
- data om utvikling når det gjelder aksjekurser,
- informasjon om økning eller nedgang i folketall i kommuner, fylker eller i landet som helhet,
- resultat av målinger av for eksempel forurensning, ozonlag eller polleninnhold i lufta,
- data om vær og klima. Det kan være målinger av temperatur og nedbør for siste dager og uker for å få fram informasjon om hvordan været har vært, og det kan være data for flere tiår og hundreår for å studere utviklingen av klimaet.

Slik informasjon presenteres gjerne i tabeller eller i en eller annen grafisk form.

Dette fører nok til at mange har den oppfatning at faget statistikk handler om innsamling, systematisering og presentasjon av statistisk informasjon. Dette er imidlertid bare en del av de oppgaver en statistiker arbeider med, og det er ikke den viktigste delen.

Det å presentere et tallmateriale på en hensiktsmessig og forståelig måte kaller vi *deskriptiv* eller *beskrivende* statistikk. En kort innføring i beskrivende statistikk er gitt i kapittel 1.3.

Hovedsakelig vil statistikeren arbeide med matematisk baserte teknikker for å få fram informasjon om *den virkeligheten som ligger bak tallene*. Dette kaller vi *statistisk analyse*. Typisk for de fenomener en statistiker studerer, er at de inneholder innslag av variasjon og tilfeldighet. Gjennom den statistiske analysen har man som siktemål å bringe på det rene om de effekter man observerer, kan forklares ved tilfeldig variasjon, eller om for eksempel endringer og forskjeller man observerer, må skyldes at det fins reelle forskjeller eller endringer i de fenomener som analyseres.

Når en meningsmåling viser at Arbeiderpartiet har gått tilbake for eksempel 2 % siden siste valg, vil en statistiker ikke være fornøyd med bare å legge fram dette resultatet. Han vil gjerne si noe om usikkerheten i resultatet, og vil spørre seg om dette kan skyldes at vi denne gangen tilfeldigvis fikk med færre Ap-sympatisører i meningsmålingen enn forrige gang, eller om det må skyldes at oppslutningen om Ap blant velgerne virkelig har gått ned.

Et annet eksempel som illustrerer en statistikers arbeidsmåte, kan være måling av pH-verdi eller surhetsgrad i fiskevann. Når vi måler pH-verdien i forskjellige vannprøver fra et fiskevann, vil det være en viss variasjon i de målte verdiene, både fordi pH-verdien ikke er helt den samme i alle deler av vannet og fordi selve målemetoden er noe usikker. Statistikeren vil ut fra de målingene som er gjort, anslå en verdi for surhetsgraden i fiskevannet og si noe om usikkerheten i den angitte verdien. Ved sammenligning av måledata fra et fiskevann med tilsvarende data fra et annet vann vil statistikeren vurdere om eventuelle forskjeller i surhetsgrad mellom de to vannene er *for store* til at de kan forklares ved den variasjonen vi finner innenfor ett og samme vann. Dersom dette er tilfelle, vil han konkludere med at de to vannene har forskjellig pH-verdi.

For å kunne gjennomføre statistiske analyser som det her er gitt eksempler på, trengs *sannsynlighetsregning*. Det er en matematisk basert disiplin som kan hjelpe oss til å finne ut hvor *sannsynlige* forskjellige resultat av forsøk og fenomener der tilfeldigheter spiller en rolle, er.

Sannsynlighetsregningen er, ved siden av å være et nødvendig grunnlag for den statistiske analysen, også i seg selv et nyttig redskap i mange sammenhen-

ger. Når en skal ta beslutninger under usikkerhet, for eksempel om investering i nye prosjekter, vil en ved hjelp av en sannsynlighetsmodell kunne regne på konsekvenser av de aktuelle beslutninger. Mange vil vel dessuten forbinde sannsynlighetsregningen med spill som for eksempel Lotto og tipping. Her er sannsynlighetsregningen uunnværlig når en skal vurdere vinnerjanser.

I kapitlene 2, 3 og 4 gjennomgås den grunnleggende sannsynlighetsregningen. Siktemålet med denne teorien er å bygge opp en modell for forsøket eller fenomenet som vi studerer, en modell som tar hensyn til den variasjonen og de tilfeldighetene som har innvirkning på resultatet.

Siden de metoder en statistiker arbeider med har et matematisk fundament, bruker vi ofte betegnelsen *matematisk statistikk* i motsetning til det å presentere tallmaterialet som kalles deskriptiv eller beskrivende statistikk.

Ofte ser en i media at det ut fra et tallmateriale eller ut fra resultater fra en undersøkelse eller et forsøk, trekkes konklusjoner uten at det ligger noen analyse bak. Man «synser» og uttaler seg kun fordi dette synes rimelig ut fra tallene som er observert. Men man kan ikke vite om de eventuelle forskjeller som er observert, skyldes tilfeldigheter eller om det er reelle forskjeller som er kommet fram. Man bør generelt være kritisk til medias bruk av statistisk informasjon.

Å gi et fyldestgjørende og presist svar på spørsmålet i tittelen om hva statistikk er, er like vanskelig som å gi en klar og konsis definisjon på hva økonomi eller geografi er. Det er ikke mulig å gi en kort og lettfattelig definisjon som samtidig er utfyllende. Her er likevel et forsøk på å forklare kort hva vi i denne boken vil legge i begrepet statistikk.

Definisjon 1.1 Statistikk

I faget statistikk studeres forsøk og fenomen der tilfeldigheter har innvirkning på resultatet. En utvikler og gjør bruk av matematisk baserte metoder til å få fram viten og om mulig trekke konklusjoner på grunnlag av resultatene.

Merk formuleringen «om mulig trekke konklusjoner». Av og til må man avfinne seg med at resultatene ikke gir grunnlag for å trekke noen konklusjoner på de spørsmål som er stilt.

Merk dessuten at metodespørsmålet er sentralt. Presentasjon av resultatene og eventuell utregning av viktige nøkkeltall for dataene er for så vidt også en del av arbeidet for en statistiker. Men den oppgaven vil i våre dager ofte bli utført av et dataprogram.

Statistikeren vil også delta i planleggingen av forsøksopplegget og data-innsamlingen. Grundig planlegging er viktig for å få gode data. Forsøket bør legges opp på en slik måte at innslaget av uønsket variasjon blir minst mulig, slik at de effekter en ønsker å studere, kommer best mulig fram.

I kapitlene 5–10 skal vi gi en innføring i statistisk arbeids- og tenkemåte. Vi skal gjennomgå viktige statistiske metoder, og vi skal anvende disse på en rekke vanlig forekommende problemstillinger.

Statistikk brukes som redskapsfag innen nær sagt alle fagområder. Statistisk tenkning og statistisk analyse har en selvfølgelig plass innen teknologi og realfag og innen økonomi og samfunnsfag. Studenter som tar ingeniørutdanning eller økonomiutdanning, som studerer medisin eller tar høyere landbruksutdanning vil alle i sin utdanning ha et grunnkurs i statistikk med et omfang omtrent som i denne boken. Innholdet i faget varierer lite fra utdanning til utdanning. Kjernestoffet i sannsynlighetsregning og statistikk er det samme for alle.

Statistikk er dessuten tatt i bruk av en rekke forskere i vitenskaper som for eksempel litteratur og språkfag, fag som en kanskje ikke vil forbinde med matematisk baserte metoder. Men statistikk kan være et nyttig redskap også her, for eksempel når litteraturforskere skal analysere og sammenligne forfatteres språkbruk.

1.2 Bruk av modeller

En modell er en idealisert etterligning av virkeligheten eller mer presist av den delen av virkeligheten vi ønsker å studere. Viktige eksempler er miniatyrtgaver av for eksempel skip eller bygninger og alle typer av kart, fra detaljkart over området til for eksempel en høyskole til verdenskart. Hva slags modell vi vil bruke, avhenger av de oppgaver som skal løses. Vi trenger for eksempel forskjellige typer kart om vi skal planlegge et boligområde, om vi skal gå tur i fjellet, eller om vi skal kjøre bil rundt i Europa.

Hensikten med å bruke modeller som et hjelpemiddel er at det skal bli lettere å forstå og analysere virkeligheten. Modellen vil kunne sette en i stand til å se helheten og sammenhengen mellom de enkelte deler av virkeligheten. Eksempelvis gir et kart et idealisert bilde av det terrenget vi ønsker å studere, og gjennom kartet kan en raskt og greit få informasjon som det ofte vil være svært tidkrevende og vanskelig å skaffe seg ved å studere virkeligheten direkte. Det er dessuten viktig at en i en modell kan studere forskjellige alternativer og vurdere konsekvenser av disse. På et kart vil en for eksempel kunne studere forskjellige veivalg for fjellturen eller forskjellige planløsninger for for eksempel veitraseer og bebyggelse.

Siktemålet med slik analyse i en modell er at en skal komme fram til gode veivalg, gode planløsninger osv. Generelt kan en si at målet er *gode beslutninger*.

For at de beslutninger vi tar på grunnlag av analyse i modellen, skal være gode beslutninger, må modellen være *realistisk*. Modellen må ligne på virkeligheten. Kartet må stemme med terrenget. På den annen side må modellen *ikke være for komplisert* hvis vi skal kunne forstå og analysere den på en effektiv måte. Disse to kravene til *realisme* på den ene siden og *enkelhet* på den andre siden vil være motstridende når en skal lage modellen. Økt realisme i modellen fører til økt detaljering og kompleksitet som igjen fører til at forståelsen og analysen blir vanskeligere. Når en skal velge modell, må en vurdere realisme opp mot enkelhet ut fra den oppgaven som skal løses. I praksis vil det som oftest være mulig å finne en rimelig god modell som er både tilstrekkelig realistisk og tilstrekkelig enkel for det aktuelle formål. For eksempel vil en kunne finne godt egnede kart til planleggingsformål, til fjellturen eller til bilturen i Europa.

Vi skal i statistikken verken arbeide med kart eller med miniatyrtvgaver, men med *matematiske modeller*. Det vil si at modellen er bygget opp av matematiske begrep som tall, mengder, formler og uttrykk, ligninger og funksjoner.

I for eksempel fysikk og kjemi, der en også bruker matematiske modeller, er det ofte slik at når forsøksbetingelsene er kjent, vil en kunne regne seg fram til et entydig resultat av for eksempel en kjemisk reaksjon. Vi sier at forsøket og modellen er *deterministisk*. For de fenomener og forsøk en statistiker arbeider med, er det ikke slik. Her virker tilfeldigheter inn på resultatet, og utfallet av hvert enkelt gjentak av et forsøk kan ikke forutsies.

Vi sier at forsøket og modellen er *stokastisk*. Stokastisk er et adjektiv som brukes om noe som er påvirket av tilfeldigheter.

En *stokastisk modell* vil typisk bestå av en mengde av mulige resultater for det forsøk eller fenomen vi studerer og en tabell eller en funksjon som forteller oss noe om hvor sannsynlig de forskjellige resultatene er. Vi bruker også betegnelsen *sannsynlighetsmodell*.

1.3 Beskrivende statistikk

For at informasjonen i et tallmateriale av noe størrelse skal komme fram, er det viktig at materialet legges fram på en hensiktsmessig og forståelig måte. Vi skal her kort gå gjennom en del viktige teknikker for å systematisere og presentere data. Dette kaller vi *deskriptiv* eller *beskrivende* statistikk.

Eksempel 1.1 Studenters alder

Nedenfor er gitt alder for 166 studenter i 1. årskurs av studiene i økonomi og administrasjon ved Høgskolen i Telemark.

23	40	23	20	21	32	20	21	20	20	19	22	20
35	19	19	33	21	20	21	20	25	27	26	22	25
22	21	22	23	22	42	22	23	20	19	20	27	20
24	20	21	20	22	21	20	24	20	24	22	20	23
25	19	19	22	20	31	40	23	21	22	20	21	27
21	24	26	24	22	20	31	19	19	34	20	23	21
22	19	20	38	27	23	27	20	20	29	23	21	22
21	22	22	22	20	22	20	22	27	19	22	24	19
22	27	19	24	20	21	26	21	21	20	19	22	22
36	23	23	20	25	36	22	21	19	22	40	19	22
20	27	22	19	24	23	20	19	27	21	22	20	22
20	19	25	18	21	19	22	23	21	28	22	21	21
23	23	27	22	28	22	21	20	21	27			

Vi skal bruke dette tallmaterialet til å illustrere de teknikkene vi skal gjennomgå.

Det er vanskelig å skaffe seg noe godt bilde av den informasjon som ligger i materialet om en bare ser på de 166 tallene. Vi får ikke noe skikkelig inntrykk av *aldersfordelingen* blant studentene. *Hvor stor andel av studentgruppen er for eksempel over 23 år?*

Vi vil dessuten ikke direkte ut fra tallene kunne si noe om disse studentene er gjennomgående eldre eller yngre enn en annen studentgruppe eller om spredningen i alder for disse studentene er større eller mindre enn for en annen studentgruppe. For å avgjøre slike spørsmål trenger vi det vi kaller et *sentralmål* som er et mål for alderen på en typisk student, og vi trenger et mål for *spredning* i aldersfordelingen.

Hyppighetstabell

La oss starte med å studere *fordelingen* i et tallmateriale, nærmere bestemt aldersfordelingen i eksemplet. Vi vil få et klarere bilde av studentenes alder ved å sette opp en *hyppighetstabell*. På neste side er det gitt en slik tabell for de $n = 166$ studentene. I venstre kolonne står aktuelle aldre i stigende rekkefølge. I andre kolonne står antall studenter (n_A) på hver alder. For eksempel er det $n_{23} = 15$ studenter som er 23 år. I tredje kolonne er det regnet ut hvilken *andel* (n_A/n) dette utgjør av studentgruppen. Denne andelen kalles også den *relative hyppigheten*. For eksempel utgjør 23-åringene en andel på $15/166 = 0.090$ eller 9.0 % av studentene. I siste kolonne er det regnet ut de *kumulative andelene* eller de *kumulative relative hyppighetene*. Disse finnes ved at en for hver alder summerer alle andeler til og med den aktuelle alder. Av denne kolonnen kan man for eksempel lese ut at andelen av studenter som er 23 år og yngre, er 0.741 eller 74.1 %. Følgelig er 25.9 % av studentene over 23 år.