Definisjoner

Symboler

Del I

Innledende manøvre

Kapittel 1

Statistikk og sannsynlighet?

Some people hate the very name of statistics, but I find them full of beauty and interest. Whenever they are not brutalized, but delicately handled by the higher methods, and are warily interpreted, their power of dealing with complicated phenomena is extraordinary. They are the only tools by which an opening can be cut through the formidable thicket of difficulties that bars the path of those who pursue the Science of Man.

F.Galton (1908)

1.1 Statistikk over alt

Hva er statistikk? Ordet *statistikk* brukes med mange ulike betydninger. Det brukes for faget statistikk. Det brukes også om et av de mest sentrale begrep i statistikkfaget.

Definisjon 1.1 (*Statistikk*) En statistikk er en funksjon av data.

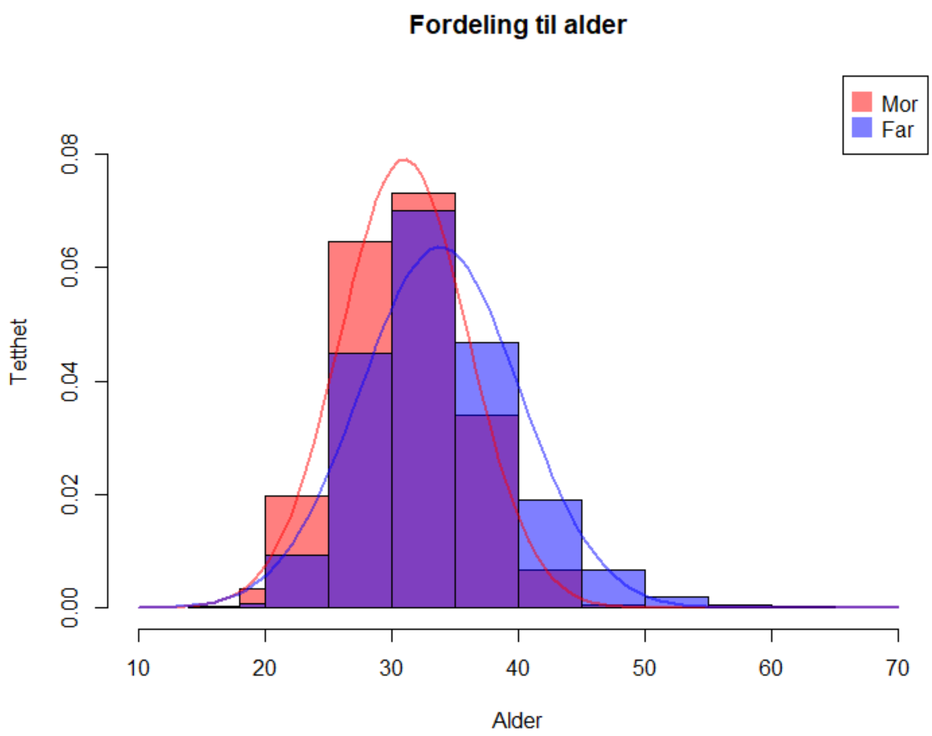
Observasjoner gir oss data. Mengden data som er tilgjengelig har eksplodert med utviklingen av datamaskinen. Ordet statistikk brukes også i betydningen *statistikkfaget* . Statistikk er vitenskapen hvor følgende spørsmål - og deres samspill - står sentralt.

- Hvilke data skal observeres?
- Hvordan lære fra data?
- Hva er gode data?

Som menneske vet du at det er en usikkerhet knyttet til alle observasjoner. Dermed er det en usikkerhet knyttet til alle data. Utgangspunktet i teoretisk statistikk er å modellere usikkerheten ved å assosiere sannsynligheter med data. Fordi en statistikk er en funksjon av data, så er statistikken selv også data med sine egne assosierte sannsynligheter. I teoretisk statistikk vil egenskapene til denne sannsynlighetsfordelingen avgjøre om det er en god eller dårlig statistikk. Dette er bokens tema, og vil bli forklart i mye større detalj etterhvert. I første omgang vil vi se på noen eksempel.

Alder	≤ 17	18-19	20-24	25-29	30-34	35-39	40-44	45-49	50-54	55-59	≥ 60	Uoppgitt
# Mor	56	366	5456	17748	20106	9336	1848	154 (45+)	0	0	0	0
# Far	21	96	2536	12189	19025	12701	5163	1809	565	185	82	698

Tabell 1.1: Antall fødsler i 2018 etter alder og kjønn.



Figur 1.1: Histogram og fordeling for alder for mor og far.

Eksempel 1.1 (Alder til mor og far ved fødselstidspunkt) Medisinsk fødselsregister inneholder opplysninger om fødsler i Norge. Tabell 1.1 gir antall fødsler i 2018 i ulike alderskategorier for kvinner og menn. Tallene i tabellen er en statistikk. Denne statistikken er en funksjon av de underliggende data for fødsler som er registrert av Medisinsk fødselsregister. Dette er da et eksempel på at en statistikk er en funksjon av data.

Figur 1.1 gir *normaliserte histogram* ut fra data i Tabell 1.1. De er laget ved opptelling i intervallene slik som i figuren på forsiden til boken, men høyden er normalisert slik at totalarelaet under kurven er lik 1. Tillegg B.2 viser hvordan grafene er laget.

Figur 1.1 inneholder også grafen til to kontinuerlige funksjoner som tilnærmer histogrammene. Funksjonene er også normalisert slik at arealet under grafen er lik 1. Histogrammene og grafene er funksjoner av de innsamlede data. Dette viser at en statistikk ikke behøver å være en tabell med tall. En statistikk kan med andre ord også være en graf.

Hva kan vi lære av disse statistikkene? Er de gode statistikker? Gode på hvilken måte? Svaret på disse spørsmålene avhenger av hva som ønskes undersøkt. *Grafene er statistikker* som er gode her ved at det kan være lettere å se sammenhenger geometrisk. De inneholder imidlertid mindre informasjon enn det som er gitt i Tabell 1.1. Grafene kan ikke brukes til å gjennvinne tallene i tabellen. På samme måte inneholder tabellen mindre informasjon enn det som er gitt i de underliggende data registrert i Medisinsk fødselsregister. Det er nettopp denne reduksjonen som kan gjøre en statistikk god. Grafene gir oss raskt et bilde av fordelingen til alder til mor og far. Dette bildet er ikke så lett å se direkte ved å stirre på alle tallene i Medisinsk fødselsregister

Hvor er så usikkerheten blitt av? Tallene gir at det ble registrert $N = 55070$ fødsler i Norge i 2018. Av disse var det $N_m = 56 + 366 + 5456 = 5878$ som hadde en mor yngre enn 25 år og $N_f = 21 + 96 + 2536 = 2653$ som hadde en far yngre enn 25 år. Disse registrerte tallene er det ingen usikkerhet i. Dersom en i stedet ser på antall faktiske fødsler, så er det et usikkert antall.

Helene ble født i 2001. Tallene gir $\hat{p}_m = N_m/N = 5878/55070 = 11\%$ som en approksimasjon til sannsynligheten p_m for at Helene blir mor før hun er 25 år. Hva betyr dette? Her er det mange og betydelige usikkerheter. En begrunnelse for påstandene kan gis ved å lage en statistisk modell for alder X til Helene ved hennes første fødsel. Parameteren vi vil si noe om er sannsynligheten p_m for at alderen $X < 25$ år. Symbolet X brukes her for en reell størrelse som det er assosiert usikkerhet med. Slike størrelser kalles *stokastiske variabler* .

Det antas at fødselstallene i 2018 er representative for Helene. Hver av de $N = 55070$ observasjonene sees som en tilfeldig trekning av et tall x. Resultatet er $x_1, x_2, \ldots, x_N$ tall som er registrert i Medisinsk fødselsregister. Hendelsen $x_i < 25$ år inntreffer i $\hat{p}_m = N_m/N = 5878/55070 = 11\%$ av tilfellene. Vi sier da at den observerte *relative hyppigheten* er 11%. Sannsynligheten p_m er, dersom modellen er rett, lik grensen for den tilsvarende relative hyppigheten når antall observasjoner går mot uendelig. Dette er på samme måte som at relativt antall Mynt i en serie myntkast kan gå mot 50% når antall myntkast går mot uendelig. I vårt tilfelle er antall observasjoner $N = 55070$, og vi forventer dermed at $\hat{p}_m$ er en god approksimasjon til p_m.

Histogrammet for Mor i Figur 1.1 er laget slik at arealet under kurven opp til alder 25 år er 11%. Arealet under den kontinuerlige kurven opp til alder 25 år er da også tilnærmet lik 11% fordi kurven approksimerer histogrammet. Kurven er grafen til en *normalfordeling* definert av funksjonen

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \mathrm{e}^{-\frac{(\mathrm{x}-\mu)^2}{2\sigma^2}} \tag{1.1}$$

hvor μ er *forventningsverdi* og σ er *standardavvik* .

En *statistisk modell* for alderen X er gitt ved å anta at sannsynlighetene assosiert med X er gitt av normalfordelingen i ligning (1.1). Sannsynligheten p_m for at alderen $X < x_m = 25$ år er da en *parameter* definert ved

$$p_m = \int_{-\infty}^{x_m} f(x)\, dx \tag{1.2}$$

Hvordan er så den kontinuerlige kurven for Mor i Figur 1.1 laget? Problemet er at modellparametrene μ og σ er ukjente. Dette er på samme måte som at sannsynligheten p_m er ukjent. Dersom tallene $x_1, x_2, \ldots, x_N$ registrert i Medisinsk fødselsregister var kjente for oss, så kunne vi bruke *empirisk middel* $\overline{x}$ og *empirisk standardavvik* s som estimat for μ og σ. Empirisk middel $\overline{x}$ og *empirisk varians* s^2 er definert ved ligningene

$$\overline{x} = \frac{\sum_{i=1}^{N} x_i}{N} = \frac{x_1 + \cdots + x_N}{N} \tag{1.3}$$

$$s^2 = \frac{\sum_i (x_i - \overline{x})^2}{N-1} = \frac{(x_1 - \overline{x})^2 + \cdots + (x_N - \overline{x})^2}{N-1} \tag{1.4}$$

Tillegg B.2 gir en mulig metode basert på data som i Tabell 1.1. Resultatet er estimatene $\tilde{\mu}_m = 31.01$ og $\tilde{\sigma}_m = 5.05$. Disse kan brukes til å beregne et estimat $\tilde{p}_m$ ved integralet i ligning (1.2). Dette integralet kan ikke løses analytisk, men må beregnes ved numerisk integrasjon. Dette er gjort ved metoden i Tillegg B.1. Resultatet er estimatet $\tilde{p}_m = 12\%$.

Diskusjonen av dette eksempelet kan gjøres mye lenger. Det er mange åpne spørsmål. Er estimatet $\tilde{p}_m$ basert på en normalfordeling bedre enn estimatet $\hat{p}_m$ beregnet mer direkte? Hva menes med bedre? Disse, og flere spørsmål, er tema for videre diskusjon i boken. Noen av disse kan du selv arbeide med i oppgavene på side 9-.

Definisjon 1.1 gir en av betydningene til begrepet *statistikk*. En statistikk er en funksjon av data. Funksjonen og verdien til funksjonen er to forskjellige ting. Begrepet *statistikk* brukes om begge tolkningene. Denne konvensjonen skaper ofte forvirring for studenter og brukere av statistikk, men den er hensiktsmessig. Den samme konvensjonen brukes om følgende like sentrale begrep.

Definisjon 1.2 (*Parameter*) En parameter er en funksjon av modellen.

Det er gitt eksempel både eksplisitt og implisitt på både parametre og statistikker i Eksempel 1.1. Det overlates til deg å lage flere eksempel ut fra dette. Mange flere eksempel vil bli gitt senere.

Eksempel 1.2 (Mikrometer, skyvelær og øyemål) Tabell 1.2 gir resultatet av å måle en lengde μ med et mikrometer, med et skyvelær (Taraldsen, 2006) og med øyemål. Tallene i tabellen er data og derfor også en en statistikk. Inspeksjon av tallene viser at resultatene fra skyvelæret tilfeldigvis også er lik avrundede verdier fra målingene med mikrometer.

En mulig modell for målingene med skyvelæret er at de er lik avrundede verdier av en underliggende verdi x. Resultatet $y = 7.5$mm i måling 1 med skyvelæret tilsvarer da en underliggende verdi x slik at $7.450\text{mm} \le x < 7.55\text{mm}$. Litt ettertanke gir at avrundingen tilsvarende skyvelæret er avrundet på samme måte som data i Tabell 1.1. Dette gjelder også resultatene fra mikrometeret, men der er det avrunding innenfor et mindre intervall.

	Lengde (mm)		
Måling	Mikrometer	Skyvelær	Øyemål
1	7.489	7.5	10
2	7.503	7.5	10
3	7.433	7.4	10
4	7.549	7.5	10
5	7.526	7.5	10
6	7.396	7.4	10
7	7.543	7.5	10
8	7.509	7.5	10
9	7.504	7.5	10
10	7.383	7.4	10

Tabell 1.2: Måling av en lengde.

Soldater	År	Forventet antall år
0	109	108.7
1	65	66.3
2	22	20.2
3	3	4.1
4	1	0.6
5-	0	0.1

Tabell 1.3: Antall prøyssiske soldater sparket i hjel av hester.

Målingene med øyemål tilsvarer avrunding innenfor et større intervall. Disse siste målingene var gjort slik at personen kun kunne gi svar i hele 1cm'ere.

Vi skal analysere tallene i Tabell 1.2 senere. Det er mange naturlige spørsmål. Hvordan estimeres lengden μ? Hva er nøyaktigheten? Usikkerheten? Hva er effekten av avrundingen? Alle data lagret på en datamaskin er med endelig nøyaktighet, og i mange tilfeller er data avrundet til kun noen få siffer. Når kan effekten av avrundingen neglisjeres? Hva hvis ikke? Kan effekten av avrunding fjernes ved et alternativt eksperiment?

Eksempel 1.3 (Bortkiewicz 1898: Das Gesetz der Kleinen Zahlen) En hest kan sparke i hjel et menneske. von Bortkiewicz (1898, p.25) studerte død til prøyssiske soldater sparket i hjel av hester. Tabell 1.3 er resultatet av å telle antall dødsfall per år for 10 kavalerikorps i 20 år. Rad 1 viser at i 109 av disse 200 kavaleriårene så var det ingen dødsfall, og forventet antall år uten dødsfall er 108.7 år. Hva betyr dette siste? Hva er usikkerheten her?

von Bortkiewicz (1898) brukte tallene i Tabell 1.3 sammen med flere andre eksempel når han formulerte det som nå kalles *Loven for de små tall* . En mulig modell er gitt ved å anta at for hver soldat, så er det hvert sekund en viss sannsynlighet for å bli sparket i hjel. Et mer positivt eksempel er gitt ved fødselstallene i Tabell 1.1. En mulig modell der er gitt ved at det hvert sekund er en viss sannsynlighet for at en kvinne føder et barn. Her er det igjen mange åpne spørsmål, og vi skal se mer på dette senere.

1.2 Oppgaver

De neste oppgavene refererer til data gitt i Eksempel 1.1. Oppgaver merket med (**D**) anbefales løst med statistisk programvare på datamaskin. Eksempel på dette er gitt i Tillegg B.

1.1. Helene har ikke fått barn i 2020. Bruk denne opplysningen og Tabell 1.1 til å finne et forbedrede estimat for sannsynligheten for at hun får barn før hun er 25 år. Diskuter resultatet.

1.2. Diskuter antagelsen om at fødselstallene i 2018 er representative for Helene. Er 2018 relevant? Hva med tvillinger? Får alle barn? Bosted? Andre faktorer? Hva hvis det hadde vært oppgitt fødselstall for 1967 i stedet? (Hvis du er nysgjerrig kan du finne disse i statistikkbanken til Medisinsk fødselsregister.)

1.3. Halvor er født i 2002. Bruk dette og Tabell 1.1 til å finne en approksimasjon $\hat{p}_f$ til sannsynligheten p_f for at Halvor blir far før han er 25 år.

1.4 (**D**). Tallene i Tabell 1.1 for Far kan brukes til å begrunne estimatene $\tilde{\mu}_f = 33.84$ og $\tilde{\sigma}_f = 6.28$. Bruk dette, og en antagelse om normalfordeling, til å beregne et estimat $\tilde{p}_f$ til sannsynligheten p_f for at Halvor blir far før han er 25 år.

1.5 (**D**). Kan du estimere sannsynligheten for at Halvor blir far etter at han er 50 år? Hva med før han er 25 år, men etter at han er 15 år? Hva med Helene? Kommenter svarene og antagelsene.

1.6. Kan du finne og beskrive en metode til å estimere forventningsverdi μ_f og standardavvik σ_f ut fra tallene i Tabell 1.1?

1.7. La N være antall registrerte fødsler i 2018 og la M være antall faktiske fødsler. Er det en sammenheng mellom M og N? Hvorfor er det en usikkerhet knyttet til M? Kan du tenke deg et eksperiment som sier noe om sannsynligheten for at en faktisk fødsel blir registrert? Har alder til mor eller far noen betydning for denne sannsynligheten? Kan du tenke deg et eksperiment for å måle dette?

1.8. Gi 3 ulike eksempel på en parameter knyttet til en modell for Helene.

1.9. Gi 3 ulike eksempel på statistikker knyttet til Tabell 1.1 ved å definere funksjonene.

De neste oppgavene refererer til data gitt i Eksempel 1.2. Oppgaver merket med (**D**) anbefales løst med statistisk programvare på datamaskin. Eksempel på dette er gitt i Tillegg B.

1.10 (**D**). Regn ut empirisk middel og empirisk standardavvik til mikrometermålingene i Tabell 1.2. Gi en intuitiv begrunnelse for at empirisk middel er et estimat for lengden μ. Hva med empirisk standardavvik? Hint: Anta at målingene er normalfordelt med forventning μ og standardavvik σ.

1.11 (**D**). Gjør som i Oppgave 1.10, men se på målingene med skyvelær og med øyemål. Diskuter resultatene.

1.12. Gi 3 ulike eksempel på en parameter knyttet til en modell for mikrometermålingen.

1.13. Gi 3 ulike eksempel på statistikker knyttet til mikrometermålingen ved å definere funksjonene.

1.14. Forklar hvordan dataene i Tabell 1.1 kan representeres som avrundede data av samme type som i Tabell 1.2.

De neste oppgavene refererer til data gitt i Eksempel 1.3. Oppgaver merket med (**D**) anbefales løst med statistisk programvare på datamaskin. Eksempel på dette er gitt i Tillegg B.

1.15. Bruk Tabell 1.3 til å estimere sannsynligheten for at ingen soldater dør av hestespark i et regiment i løpet av et år.

1.16. Estimer sannsynligheten for at det dør 3 soldater, eller færre, i et regiment i løpet av et år.

1.17. I løpet av et år i et regiment blir x soldater sparket i hjel av en hest. La $x_1, \ldots, x_N$ tilsvare observasjonene for de $N = 200$ regimentårene gitt i Tabell 1.3. De 200 observasjonene er ikke kjent, men det er allikevel mulig å beregne empirisk middel $\overline{x}$. Forklar dette og regn ut empirisk middel $\overline{x}$.

1.18. Gi 3 ulike eksempel på en parameter knyttet til en modell for dødsfallene.

1.19. Gi 3 ulike eksempel på statistikker knyttet til observasjonene av dødsfall ved å definere funksjonene.

De neste oppgavene omhandler elementær teori og notasjon for funksjoner og mengder.

1.20. Definer funksjonene exp og ln. Er det flere mulige definisjoner?

1.21. Kan du gi to rimelige definisjoner av $\ln : (0, \infty) \to \mathbb{R}$ hvor resultatet er to ulike funksjoner? Hint: Hva er domenet til ln ?

1.22. La A være en 2×2 matrise. Definer en funksjon som tilsvarer $A \mapsto A^{-1}$. Finnes det flere naturlige definisjoner? Hint: Generalisert invers for løsning av lineært ligningssystem.

1.23. La A og B være to mengder i planet som hver inneholder høyst 10 punkter. Definer en funksjon f som tilsvarer $(A, B) \mapsto A \cap B$. Hva er naturlig domene, kodomene og definisjonsmengde. Hva blir verdimengden?

1.24. La $\Omega = \mathbb{R}^2 = \{(x, y) \,|\, x, y \in \mathbb{R}\}$, dvs Ω er mengden av punkter i planet. Illustrer delmengdene $A = \{(x, y) \,|\, x = 5\}$, $B = \{(x, y) \,|\, x \geq 5\}$, $C = \{(x, y) \,|\, x \geq y^2\}$, $D = C \setminus B$, $E = \{(x, y) \,|\, x \geq y\}$, $F = C \cap E$, $G = F \cup A$, $H = \mathbb{N}^2 = \{(x, y) \,|\, x, y \in \mathbb{N}\}$. Er H tellbar?

1.25. La $\mathbb{Q}_+$ være mengden av alle positive rasjonale tall. Er $\mathbb{Q}_+$ tellbar? Er mengden av rasjonale tall tellbar? Svarene må begrunnes!

1.26. Kontroller om følgende regneregler gjelder: $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$, $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$, $(A \cup B) \cap (C \cup D) = (A \cap C) \cup (A \cap D) \cup (B \cap C) \cup (B \cap D)$, $(A \cup B)^c = A^c \cap B^c$, $(A \cap B)^c = A^c \cup B^c$.

Illustrer konklusjonene med Venn-diagram.

Kan regnereglene generaliseres til union/snitt av vilkårlige familier av mengder?

1.27. La $A_n = [0, 1/n]$. Finn $B = \cup_{n=1}^{10} A_n$, $C = \cap_{n=1}^{10} A_n$, $D = \cap_{n=1}^{\infty} A_n$, $E = \cup_{n=1}^{\infty} A_n$.

1.28. La A, B være delmengder av Ω. Vis at $\Omega = A \uplus A^c$.

Vis at $\Omega = (A \cup B)^c \uplus (A \cap B) \uplus (A \setminus B) \uplus (B \setminus A)$. Illustrer dette i et Venn-diagram.

Kan du generalisere dette til tilfellet hvor vi tar utgangspunkt i tre delmengder A, B, C?

Hva med tilfellet med n delmengder, eller et tellbart antall delmengder?

Del II

Sannsynlighetsteori

Kapittel 2

Eksperiment og Kolmogorovs aksiomer

The situations we are going to model can all be thought of as random experiments. Viewed naively, an experiment is an action which consists in observing or preparing a set of circumstances and then observing the outcomes of this situation. We add to this notion the requirement that to be called an experiment such an action must be repeatable, at least conceptually. The adjective "random" is used only to indicate that we do not, in addition, require that every repetition yield the same outcome, although we do not exclude this case. What we expect and observe in practice when we repeat a random experiment many times is that the frequency of each of the possible outcomes will tend to stabilize. This "long term frequency" is to many statisticians, including the authors, the operational interpretation of the mathematical concept of probability. In this sense, almost any kind of activity involving uncertainty, from horse races to genetic experiments, falls under the vague heading, "random experiment".

P. J. Bickel og Kjell A. Doksum (1977)

2.1 Utfallsrom og eksperiment

Kolmogorov (1933) formulerte aksiomer for sannsynlighetsteorien og brukte mengdeteori som en matematisk modell. Teori og notasjon for mengder antas kjent her. Tillegg A.2 kan konsulteres etter behov.

La Ω være en mengde og la $\mathcal{E}$ være en mengde delmengder av Ω. Dette betyr at

$$(A \in \mathcal{E}) \Rightarrow (A \subset \Omega) \tag{2.1}$$

En mengde av mengder kalles ofte en *familie av mengder* . Familien $\mathcal{E}$ er en *hendelsesfamilie* hvis

$$[E \in \mathcal{E}] \Rightarrow [E^c \in \mathcal{E}], \quad [A_i \in \mathcal{E}] \Rightarrow [\bigcup_i A_i \in \mathcal{E}], \tag{2.2}$$

hvor det siste skal gjelde for enhver tellbar familie $\{A_i \,|\, i \in I\}$ av hendelser. En *hendelse* er et medlem av hendelsesfamilien. Et *utfall* er et element i utfallsrommet.

Definisjon 2.1 (*Utfallsrom*) En ikke-tom mengde utstyrt med hendelser.

Det er kanskje en bekymring for leseren at overstående er meget generelt. Vi ser på noen konkrete eksempler for å belyse begrepene og for å vise anvendelsen i forbindelse med modellering av eksperiment.

Eksempel 2.1 (Et myntkast) Du kaster en mynt og får resultatet M (Mynt). Dette er et eksperiment med utfallsrom

$$\Omega = \{M, K\} \tag{2.3}$$

Hendelsesfamilien $\mathcal{E}$ er familien av alle delmengder av Ω. Resulatet av et enkelt myntkast er M (Mynt) eller K (Kron). Disse utfallene er inneholdt i Ω som dermed er et mulig utfallsrom for eksperimentet.

La ω være utfallet av eksperimentet. Hendelsen A har da *inntruffet* dersom $\omega \in A$. Hendelsen Ω inntreffer alltid og hendelsen $\emptyset$ inntreffer aldri. Hendelsen $A = \{M\}$ har inntruffet i tilfellet her. Et alternativt utfallsrom er

$$\Omega = \{0, 1\} \tag{2.4}$$

Her er $\omega = 0$ dersom resultatet er Mynt og $\omega = 1$ dersom resultatet er Kron. De to foregående utfallsrom er så små som mulig. Et tredje alternativt utfallsrom er

$$\Omega = \mathbb{R} \tag{2.5}$$

Her er $\omega = 0$ dersom resultatet er Mynt og $\omega = 1$ dersom resultatet er Kron. Et *utfallsrom for et eksperiment* er et utfallsrom som inneholder alle mulige utfall av eksperimentet. Utfallsrommet $\Omega = \mathbb{R}$ for dette eksperimentet skal vi bruke mye videre i boken. Hendelsesfamilien $\mathcal{E}$ er den minste [1] som inneholder alle intervall.

Eksempel 2.2 (Tre myntkast i rekkefølge) Du kaster en mynt og får resultatet MMK (Mynt,Mynt,Kron). Dette er et eksperiment med minste utfallsrom

$$\Omega = \{MMM, MMK, MKM, MKK, KMM, KMK, KKM, KKK\} \tag{2.6}$$

Det består av 8 enkeltutfall. Vi lar hendelsesfamilien $\mathcal{E}$ være familien av alle delmengder. La $E = \{MKK, KMK, KKM, KKK\}$. Utfallet av eksperimentet ble $\omega = MMK$. Da inntraff ikke hendelsen E, dvs vi fikk $\omega \in E^c$. Å si at hendelsen E inntraff er det samme som å si at vi fikk flere Kron enn Mynt, og det fikk vi ikke.

Et alternativt utfallsrom i dette eksemplet er

$$\Omega = \{000, 001, 010, 011, 100, 101, 110, 111\} \tag{2.7}$$

[1] Se Oppgave 2.4

ved å identifisere 0 med Mynt og 1 med Kron. Fordelen med dette valget er at alle de mulige utfallene kommer fram ved å telle binært. Et tredje alternativt utfallsrom er

$$\Omega = \{0,1\}^3 = \{0,1\} \times \{0,1\} \times \{0,1\} \tag{2.8}$$

Et utfall $\omega = (\omega_1, \omega_2, \omega_3)$ betyr at myntkast nummer i ga resusltat ω_i. I kast i er da $\omega_i = 0$ dersom resultatet er Mynt og $\omega_i = 1$ dersom resultatet er Kron.

Eksempel 2.3 (Gjentatt myntkast inntil Kron) Du kaster en mynt og får resultatet MMK (Mynt,Mynt,Kron). Dette kan sees som utfallet i et eksperiment hvor du kaster en mynt inntil du får Kron. Utfallsrommet er

$$\Omega = \{K, MK, MMK, MMMK, \ldots\} \tag{2.9}$$

Et alternativt utfallsrom er

$$\Omega = \mathbb{N} \tag{2.10}$$

Resultatet i eksperimentet er da $\omega = 3$. Begge utfallsrom kan være utstyrt med hendelses-familien $\mathcal{E}$ lik familien av alle delmengder.

Eksempel 2.4 (Levetid til klimaannlegg i fly) Proschan (1963) studerte levetidsdata for komponenter i klimaanlegg til en flåte med jetfly. Levetidene målt i hele timer for fly nummer 7914 ble $50, 44, 102, 72, 22, 39, 3, 15, 197, 188, 79, 88, 46, 5, 5, 36, 22, 139, 210, 97, 30, 23, 13, 14$ (Taraldsen, 2011, s.977). Utfallsrommet for første observasjon $\tilde{\omega} = \omega_1 = 50$ timer er

$$\tilde{\Omega} = \mathbb{R} \tag{2.11}$$

Hendelsesfamilien $\tilde{\mathcal{E}}$ er den minste som inneholder alle hendelser på formen $\tilde{H} = (t_1, \infty)$. Alle levetidene samlet er gitt ved enkeltutfallet $\omega = (\omega_1, \ldots, \omega_n)$ hvor $n = 24$. Utfallsrommet er da

$$\Omega = \mathbb{R}^n \tag{2.12}$$

Hendelsesfamilien $\mathcal{E}$ er den minste som inneholder alle hendelser på formen $H = (t_1, \infty) \times \cdot \times (t_n, \infty)$.

Tid (timer)	Antall	Forventet antall
0-40	166	180.5
40-80	151	145.0
80-120	132	115.4
120-160	98	92.7
160-200	73	73.8
200-240	45	59.3
240-280	53	47.2
280-320	40	37.9
320-360	23	30.2
360-400	26	24.3
400-440	24	19.3
440-480	9	16.0
480-520	9	11.6
520-560	8	9.9
560-600	9	8.0
600-700	17	13.5
700-800	8	8.2
800-	12	10.0

Tabell 2.1: Levetid til V805 radar senderør.

Eksempel 2.5 (Levetiden til V805 radar senderør) V805 senderrøret var en vanlig komponenent i mange flyradarsystemer. Davis (1952, Tabell C på s.133) studerte levetiden. Tabell 2.1 er resultatet av $n = 903$ levetider $\omega_1, \ldots, \omega_n$. Utfallsrommet for eksperimentet tilsvarende den første observasjonen $\tilde{\omega} = \omega_1$ er

$$\tilde{\Omega} = \mathbb{R} \tag{2.13}$$

Utfallsrommet for eksperimentet tilsvarende de $n = 903$ observasjonene er

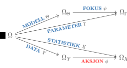
$$\Omega = \mathbb{R}^n \tag{2.14}$$

Her er det da observert et utfall $\omega = (\omega_1, \ldots, \omega_n)$. Hendelsesfamilien $\mathcal{E}$ er den minste som inneholder alle hendelser på formen $B = (t_1, \infty) \times \cdot \times (t_n, \infty)$.

Eksempel 2.6 (Gjentatt myntkast) Du kaster en mynt mange ganger og får resultatet $MMKMK \cdots$ (Mynt,Mynt,Kron,Mynt,Kron, ...). Denne uendelige serien kan identifiseres med utfallet $\omega = (0, 0, 1, 0, 1, \ldots)$. Et minste utfallsrom er

$$\Omega = \{0, 1\}^{\mathbb{N}} \tag{2.15}$$

Hendelsesfamilien $\mathcal{E}$ er den minste som inneholder alle hendelser på formen $B_n = \{\omega \,|\, \omega_1 \in I_1, \ldots, \omega_n \in I_n\}$ hvor n er vilkårlig og $I_j \in \{0, 1\}$.



De foregående eksemplene viser at utfallsrommet Ω kan være en endelig mengde, en tellbar mengde, eller en mengde som ikke er tellbar. Ordet eksperiment stammer fra det latinske *experiri* som betyr *prøve*. I det foregående og i det som følger vil ordet eksperiment bli brukt i en mer generell betydning.

Definisjon 2.2 (*Eksperiment*) Et eksperiment er en prosedyre som kan tenkes gjentatt under identiske idealiserte forhold. De ulike mulige utfall av eksperimentet skal kunne identifiseres med ulike element i et utfallsrom.

Termen *eksperiment* definert over er ikke et matematisk begrep. Et myntkast er et eksperiment. Et eksperiment er et fenomen som ligger utenfor matematikkens verden. Identifiseringen av utfallsrommet for et eksperiment er første skritt i den matematiske beskrivelsen av eksperimentet. Dette er det som er illustrert i alle de foregående eksperimentene.

2.2 Sannsynlighet og eksperiment

La Ω være et utfallsrom med en hendelsesfamilie $\mathcal{E}$.

Definisjon 2.3 (*Sannsynlighet*) En funksjon $\mathrm{P} : \mathcal{E} \to \mathbb{R}$ som oppfyller

1. $\mathrm{P}(A_1 \cup A_2 \cdots) = \mathrm{P}(A_1) + \mathrm{P}(A_2) + \cdots$ når $A_1, A_2, \ldots$ er disjunkte.
2. $\mathrm{P}(A) \geq 0$.
3. $\mathrm{P}(\Omega) = 1$.

En sannsynlighet P kalles også en *sannsynlighetsfordeling* , eller bare en *fordeling* , og også et *sannsynlighetsmål* . Tallet $\mathrm{P}(A)$ er sannsynligheten til hendelsen A. Et *sannsynlighetsrom* er et utfallsrom Ω utstyrt med en hendelsesfamilie $\mathcal{E}$ og en fordeling P.

En fordeling er en funksjon $\mathrm{P} : \mathcal{E} \to [0, \infty]$ som oppfyller to krav. Dersom normaliseringskravet $P(\Omega) = 1$ fjernes, så sies P å være et mål. En fordeling er dermed et normalisert mål. Det første kravet kalles *tellbar additivitet* . Dersom denne addisjonsregelen bare kreves for endelige familier, så sies P å være en endelig additiv fordeling. I det følgende vil vi bevise noen av de viktigste regnereglene for fordelinger. Det vil fremgå av bevisene at mange av regnereglene også er gyldige for endelig additive mål fordi det kun er den endelige addisjonsregelen som benyttes.

En sannsynlighetsfordeling benyttes for å beskrive graden av usikkerhet i eksperiment. Vi ser på noen eksempler for å belyse dette.

Eksempel 2.7 (Terningkast) Utfallsrommet for et terningkast er

$$\Omega = \{1, 2, 3, 4, 5, 6\}.$$

Ved symmetri er det rimelig å anta at hvert enkelt utfall er like sannsynlig, dvs $P\{\omega\} = p$. Sannsynligheten til en generell hendelse E er da gitt ved addisjonsregelen, dvs $PE = \sum_{\omega \in E} P\{\omega\}$. Normaliseringen gir $1 = P\Omega = \sum_{\omega=1}^{6} P\{\omega\} = 6p$ gir $p = 1/6$, og det følger at P er en fordeling, Når fordelingen er gitt kan f eks sannsynligheten for at terningkastet gir 5 eller 6 øyne beregnes ved $P\{5, 6\} = \sum_{\omega \in \{5,6\}} p = 2p = 1/3$. Videre finnes $P(\text{odde antall øyne}) = P\{1, 3, 5\} = 1/2$.

Eksempel 2.8 (To myntkast) Utfallsrommet er

$$\Omega = \{MM, MK, KM, KK\}.$$

Det er rimelig å anta at enkeltutfallene er like sannsynlige. Dermed kan en f eks beregne $P(\text{To kronesider opp}) = P\{KK\} = 1/4$, og $P(\text{Minst en myntside opp}) = 1 - P(\text{Ingen myntsider opp}) = 1 - P\{KK\} = 3/4$.

Dersom myntene er identiske og eksperimentet er å kaste de samtidig, så er utfallsrommet

$$\Omega = \{0, 1, 2\},$$

hvor enkeltutfallene teller antall myntsider opp. I dette tilfellet er en rimelig fordeling gitt ved $P\{0\} = P\{2\} = 1/4$ og $P\{1\} = 1/2$, som kan begrunnes ved likheten med eksperimentet hvor myntkastene ble utført i rekkefølge.

Eksempel 2.9 (Fotballkamp) La eksperimentet være gitt ved Norges neste hjemmelandskamp i fotball med utfallsrom

$$\Omega = \{0-0, 0-1, 1-0, 1-1, 0-3, \ldots\}.$$

En mulig vurdering er gitt ved $P(\text{Norge vinner}) = P\{1-0, 2-0, 2-1, 3-1, \ldots\} = 80\%$.

I alle eksemplene over er eksperimentene karakterisert ved et utfallsrom, og utfallet av eksperimentet er usikkert. Slike eksperiment kaller vi *sjanseeksperiment* . Dersom resultatet av et eksperiment er ω, så sier vi at hendelsen E har inntruffet dersom $\omega \in E$. En mulig modell for et sjanseeksperiment er gitt ved at utfallsrommet er et sannsynlighetsrom. Fordelingen P gir sannsynligheten $\mathrm{P}(E)$ for at hendelsen E inntreffer. Sannsynligheten tolkes i dette mest generelle tilfellet som en vurdering av rimeligheten for at hendelsen E inntreffer.

Definisjon 2.4 (*Sannsynlighetsmodell*) En sannsynlighetsmodell for et eksperiment er gitt ved at utfallsrommet er et sannsynlighetsrom.

Bortsett fra fotballkamp eksemplet, så kan alle overnevnte eksperiment gjentas et vilkårlig antall ganger. Alle kan tenkes gjentatt i et tankeeksperiment. Et sjanseeksperiment kan gjentas et vilkårlig antall ganger under idealiserte identiske betingelser. Termen *modell* brukes i stedet for sannsynlighetsmodell når betydningen er klar fra konteksten. Likeledes brukes termen *eksperiment* i stedet for sjanseeksperiment når betydningen er klar fra konteksten. Termen *statistisk eksperiment* er også i bruk. I alle tilfellene er det en modell for et eksperiment.

Hyppigheten nA til A er antall ganger hendelsen A inntreffer når et statistisk eksperiment er utført n ganger. Den *relative hyppigheten* til A er da definert ved

$$\mathrm{P}_n(A) = \frac{nA}{n} \tag{2.16}$$

Det følger at P_n er en fordeling. Dette er den *empiriske fordelingen* . Et eksperiment er et *statistisk lovmessig eksperiment* dersom grenseverdien

$$\mathrm{P}(E) = \lim_{n\to\infty} \mathrm{P}_n(E) \tag{2.17}$$

eksisterer for alle hendelser E. Det følger at dette definerer en fordeling P. Denne unike fordelingen er da den *sanne fordelingen* til eksperimentet. Vi sier i så fall at hendelsen E inntreffer i $\mathrm{P}(E) \cdot 100\%$ av eksperimentene når eksperimentet *gjentas i det lange løp* .

I praksis vil et eksperiment kun gjentas et endelig antall ganger, men vi vil analysere eksperimentet ved å tenke oss at eksperimentet kan gjentas et vilkårlig antall ganger. Observasjonen av en statistisk lovmessighet vil være gitt ved at den relative frekvensen til en hendelse synes å nå en grenseverdi når eksperimentet gjentas mange ganger.

2.3 Oppgaver

Oppgaver merket med (**T**) er teorioppgaver som kan oppfattes som vanskelige.

2.1. Finn en hendelse B som ikke har inntruffet i Eksempel 2.1. Finnes det flere?

2.2. Vis at den tomme mengden $\emptyset$ og utfallsrommet Ω er hendelser.

2.3. La $\Omega = \{1, 2, 3\}$. Er $\{\{1, 2\}, \emptyset\}$ en familie av mengder? Er $\{\{1, 2\}, \emptyset\}$ en hendelsesfamilie? Er $\mathcal{E} = \{\emptyset, \Omega, \{1, 2\}, \{3\}\}$ en hendelsesfamilie? La Ω være utstyrt med hendelsene i $\mathcal{E}$. Er 1 et utfall? Er $\{1\}$ en hendelse? Bruk dette til å begrunne at det kan finnes en delmengde av utfallsrommet som ikke er en hendelse.

2.4 (**T**). La $A = \{\mathcal{G} \mid \mathcal{G}$ er en hendelsesfamilie som inneholder alle intervall i $\mathbb{R}\}$. Vis at familien $\mathcal{E}_d$ av alle delmengder av $\mathbb{R}$ er inneholdt i A. La

$$\mathcal{E} = \bigcap_{\mathcal{G} \in A} \mathcal{G}$$

Er $\mathcal{E}$ en familie mengder? Er $\mathcal{E}$ en hendelsesfamilie? Er $\mathcal{E}$ den minste hendelsesfamilien som inneholder alle intervall?

2.5 (**T**). Vis at $\Omega = \mathbb{R}^3$ er et fjerde mulig utfallsrom for eksperimentet i Eksempel 2.2. Illustrer de mulige utfall som punkter i rommet og marker spesielt resultatet ω som ble observert etter tre myntkast. En *åpen boks* er en mengde på formen $B = I_1 \times I_2 \times I_3$ hvor I_j er et åpent intervall. La $\mathcal{E}$ være den minste hendelsesfamilien som inneholder alle åpne bokser. Vis at $\{\omega\}$ er en hendelse.

2.6. Resultatet av å kaste en mynt 4 ganger er $MMKM$. Definer et minste utfallsrom for eksperimentet. Husk: Hendelsene må også spesifideres! Vis at $\Omega = \{0,1\}^4$ og $\Omega = \mathbb{R}^4$ også er mulig utfallsrom for eksperimentet. Hendelser?

2.7. Vis at $\Omega = \mathbb{R}$ er et mulig utfallsrom i Eksempel 2.3. Hva er resultatet ω i det gitte eksperimentet? Hendelsesfamilie?

2.8. Se på Eksempel 2.4. Si med ord hva det betyr at hendelsen $\tilde{H} = (t_1, \infty)$ inntreffer. Vis at alle åpne intervall er hendelser i $\tilde{\Omega}$. Utfallet ω er oppgitt i eksempelet. Hva er ω_{10}? Si med ord hva hendelsen $H = (t_1, \infty) \times \cdot \times (t_n, \infty)$ betyr. Vis at alle åpne bokser er hendelser.

2.9 (**D**). Se på Eksempel 2.5. Forklar at data gitt i tabell 2.1 tilsvarer et utfall ω i utfallsrommet

$$\Omega = \mathbb{N}^k \tag{2.18}$$

med $k = 18$. Kan du estimere forventet levetid? Kan du tegne et normalisert histogram?

2.10. Gi eksempler som viser at utfallsrommet Ω kan være en endelig mengde, en tellbar mengde, eller en mengde som ikke er tellbar. Husk: Hendelsene må også defineres.

2.11. Hva er et eksperiment? Gi et eksempel og gi en begrunnelse.

2.12. La $A = \{[0, 1/4), [1/4, 1/2), [1/2, 1)\}$. Er A en familie mengder? Er A en hendelsesfamilie? La $\Omega = [0, 1)$ være utfallsrommet utstyrt med den minste hendelsesfamilien som inneholder A. Er $[1/2, 3/4)$ en hendelse?

2.13. Hva er utfallsrommet for et enkelt terningkast? Hva er utfallsrommet for gjentatt kast med terning inntil resultatet blir seks øyne? Hva er sannsynligheten for at terningen må kastes 2 ganger? Hva er sannsynligheten for at terningen må kastes n ganger?

2.14. En student som snart er ferdig med utdannelsen er innkalt til tre jobbintervju. Hun vil karakterisere et intervju som en suksess dersom det resulterer i en invitasjon til en omvisning på arbeidstedet. I motsatt fall karakteriseres intervjuet som en fiasko. Definer et utfallsrom Ω for dette eksperimentet. Hvilke utfall er inneholdt i hendelsen A = "andre suksess forekommer i tredje intervju"? Finn hendelsen B = "første suksess forekommer aldri".

2.15. En urne inneholder 6 brikker som er nummerert fra 1 til 6. Tre brikker trekkes fra urnen. Hvilke utfall er inneholdt i hendelsen "den nest minste verdien er 3"?

2.16. I et terningsspill (*craps*) kastes to terninger. Kasteren vinner dersom resultatet er 7 eller 11, og taper dersom resultatet er 2, 3 eller 12. Dersom resultatet er noe annet, anta 9, så fortsetter kastingen inntil resultatet er 7 (tap) eller inntil resultatet 9 fra det første kastet gjentas (vinst). Karakteriser utfallene i hendelsen "kasteren vinner med en 9".

2.17. Studentsamskipnaden viste filmen “Ny student i byen” to ganger i løpet av orienteringsdagen. Av totalt 2000 nye studenter var det 850 som så på den første forestillingen, 690 så på den andre forestillingen, mens 730 unnlot å se filmen. Hvor mange nye studenter så filmen to ganger?

2.18. La $PA = 1/3$, $PB = 1/2$ og $P(A \cup B) = 3/4$. Finn $P(A \cap B)$, $P(A^c \cup B^c)$ og $P(A^c \cap B)$.

2.19. Et eksperiment har to mulige utfall. Det ene har sannsynlighet p og det andre har sannsynlighet p^2. Finn p.

2.20. To kort trekkes i rekkefølge fra en kortstokk med 52 kort. Hva er sannsynligheten for at det andre kortet som trekkes er større enn det første kortet? Hint: Enten er kortene like store eller så er det første størst eller så er det andre størst. Disse tre hendelsene er disjunkte.

2.21. La Ω være et utfallsrom utstyrt med en hendelsesfamilie. Vis at $\emptyset$ og Ω er hendelser. Vis at et tellbart snitt av hendelser er en hendelse.

2.22. Vis at $nA = \sum_{i=1}^{n}(\omega_i \in A)$ ved å definere det som inngår. Vis at den empiriske fordelingen er en fordeling.

2.23 (**T**). La $\mathcal{E}_i$ være hendelsesfamilier i Ω. Vis at $\mathcal{F} = \mathcal{E}_1 \cap \mathcal{E}_2$ er en hendelsesfamilie. Vis at $\mathcal{G} = \cap_i \mathcal{E}_i$ er en hendelsesfamilie.

2.24. Vis at $\mathrm{P}(\emptyset) = 0$.
Vis at $\mathrm{P}(A \cup B) = \mathrm{P}(A) + \mathrm{P}(B) - \mathrm{P}(A \cap B)$.
Vis at $A \subset B$ gir $\mathrm{P}(A) \leq \mathrm{P}(B)$.
Vis at $\mathrm{P}(A) \leq 1$.
Vis at $\mathrm{P}(A^c) = 1 - \mathrm{P}(A)$.