Grunnkurs i statistikk og sannsynlighetsteori

Gunnar Taraldsen

1. august 1997
(revidert 12. juni 2006)

Innhold

Figurer

Tabeller

Del I

Innledende manøvre

Kapittel 1

Litt historikk og mengdeteori

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 Kort om statistikkfagets historiske utvikling

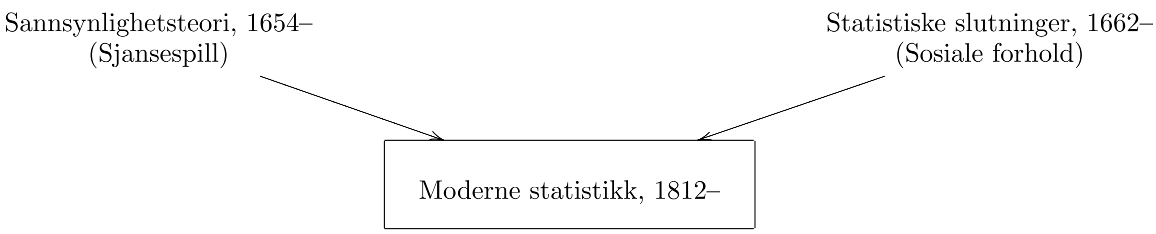
Moderne matematisk statistikk vokste frem fra sannsynlighetsteorien og beskrivende statistikk slik som indikert i figur 1.1. I det etterfølgende gis en kort gjennomgangen av statistikkfagets historikk og derved gis også et inntrykk av noen av de mange mulige anvendelsesområdene for teorien.

Egyptiske gravsteder vitner om terningspill. Dermed synes det klart at de tidligste sivilisasjoner hadde et forhold til sannsynlighetsbegrepet. Utviklingen av den matematiske teorien for sannsynlighetsberegninger kan oppsummeres ved noen av høydepunktene:

1654 Korrespondanse mellom Blaise Pascal og Pierre de Fermat anngående beregningen av sannsynligheter i forbindelse med terningspill. Dette regnes av mange som starten på den matematiske sannsynlighetsteorien.

1657 Christiaan Huygens, kjent for sine arbeider innen optikk, publiserte *De Ratiocintiis in Alea Ludo* (Beregninger i sjansespill). Dette er den første publiserte avhandling om sannsynlighet.

1713 Jacob Bernoullis avhandling *Ars Conjectandi* (Kunsten å gjette) ble publisert etter hans død. Avhandlingen inneholder blant annet et teorem for grenseverdien til en sannsynlighet knyttet til gjentagelsen av et enkelt sjanseeksperiment. Lignende grenseverdibetraktninger er fortsatt et forskningstema. Denne avhandlingen er en milepel i sannsynlighetsteoriens fødsel.



Figur 1.1: Den historiske utviklingen av moderne matematisk statistikk.

1919 Richard von Mises publiserte *Grundlagen der Wahrscheinlichkeitsrechnung* hvor den empiriske definisjonen av sannsynlighet ved relativ frekvens ble formulert.

1933 Andrei Kolmogorov ga sannsynlighetsteorien et aksiomatisk fundament ved publiseringen av *Grundbegriffe der Wahrscheinlichkeitrechnung.* Arbeidet bygde på den abstrakte teorien for mål og integrasjon som hadde blitt grunnlagt av blant annet Henri Lebesgue og Maurice Fréchet. Kolmogorovs aksiomer vil også være vårt utgangspunkt for den matematiske formuleringen av teorien.

Ordet statistikk vil i vår sammenheng benyttes i to betydninger. Ordet betegner en samling av data, som i f.eks. Statistisk Årbok, men betegner også en metode for å trekke slutninger på grunnlag av en samling med data. Statistikk i ordets første betydning har sine røtter tilbake til skrivekunstens vugge. Utviklingen av teorien for statistiske slutninger kan oppsummeres ved noen høydepunkter:

1662 John Graunt publiserte boken *Natural and Political Observations Made upon the Bills of Mortality* som var basert på fødsels- og dødsrater for befolkningen i England som var rammet av pest.

ca 1662 Sir William Petty begynte utviklingen av en teori for *the art of reasoning by figures upon things relating to government.*

1700 Edmund Halley, kjent for å ha en komet oppkalt etter seg, publiserte arbeider knyttet til dødelighetstabeller og forventet levetid. Etter disse arbeidene fremsto statistikken som en egen livskraftig vitenskap.

Det logiske fundamentet for den matematiske statistikken er gitt ved den matematiske formuleringen av sannsynlighetsteorien. En kan dermed se på sannsynlighetsteorien som en viktig del av statistikkfaget, eller alternativt se på den matematiske statistikk som en naturlig videreføring av sannsynlighetsteorien.

Utgangspunktet i sannsynlighetsteorien er å beregne sannsynligheter for visse utfall når en sannsynlighetsmodell er gitt. Ved statistiske slutninger snus dette på hodet ved at en forsøker å finne den eller de blant flere mulige sannsynlightesmodeller som beskriver gitte data best mulig. Oppdagelsen av samspillet mellom sannsynlighetsteori og statistikk kan sees på som fødselen av den moderne matematiske statistikk. En kort historikk er gitt ved:

1812 Pierre-Simon Laplace publiserte boken *Theorie Analytique de Probabilities* som oppsummerte teorien til da, og ga en mengde anvendelser av sannsynlighetsregningen. En av anvendelsene var analysen av feilobservasjoner i eksperimentalfysikk.

ca 1812 Adolphe Quetelet, en student av Laplace, reiste rundt i Europa og demonstrerte bruken av sannsynlighetsmodeller i beskrivelsen av sosiale og biologiske fenomen.

1908 Francis Galton publiserte *Natural Inheritance.*

1930 Ronald Aylmer Fisher publiserte *The genetical theory of natural selection* og etablerte derved en syntese av Darwinismen og Mendelismen. Fisher er antagelig den enkeltperson som har hatt størst innflytelse på utviklingen av den matematiske statistikken i dette århundre.

1.2 Litt mengdeteori

Abstrakt mengdeteori står sentralt i Kolmogorovs formulering av sannsynlighetsteorien. I det følgende vil vi ikke gjøre noe forsøk på å bygge opp mengdeteorien aksiomatisk, men vil ta som gitt at leseren har en viss fortrolighet med manipulering med mengder. Hovedinnholdet her blir da å fastlegge noen notasjons- og språklige konvensjoner.

Paranteser av type $\{\cdots\}$ benyttes for å beskrive mengder. Noen eksempler på endelige mengder er gitt ved $\{a, b, c\}$ og $\{1, 2, \ldots, n\}$. Den tomme mengden $\emptyset$ er også endelig, den har ingen elementer.

Vi har $\emptyset = \{\omega \,|\, \omega \neq \omega\}$. Symbolet $|$ leses i denne sammenhengen som *slik at* og brukes generelt til definisjon av mengder. Mengden av de naturlige tall betegnes med $\mathbb{N} = \{1, 2, 3, \ldots\}$, og mengden av de reelle tall betegnes med $\mathbb{R} = \{x \,|\, -\infty < x < \infty\}$.

En mengde I er tellbar dersom den er i en-en korrespondanse med en endelig mengde eller mengden av de naturlige tall. Dette betyr at vi kan liste opp alle elementene i I ved $i_1, i_2, \ldots$. Dette burde forklare navnet. Bortsett fra mengden $\mathbb{R}$ av reelle tall er dermed alle overnevnte mengder tellbare. Definisjon: **tellbar**

Vi skriver $a \in A$ dersom a er et element i mengden A. Symbolet $\in$ leser vi som *er et element i*. Symbolet $\notin$ leser vi tilsvarende som *er ikke et element i*. To eksempler er gitt ved $b \in \{a, b, c\}$ og $d \notin \{a, b, c\}$. Definisjon: **$\in$, element i**

En mengde A er en delmengde av en mengde B dersom alle elementer i A også er elementer i B. Vi benytter notasjonen $A \subset B$ for dette. Vi har da at $a \in A \Rightarrow a \in B$. Tegnet $\subset$ leses som *er en delmengde av*. Tegnet $\Rightarrow$ leses som *medfører at*. Tre eksempler er gitt ved $\{1, 100, 14\} \subset \mathbb{N}$, $\emptyset \subset \mathbb{R}$ og $\{a, b, c\} \subset \{a, b, c\}$. Merk: Den vanligste metoden for å bevise at $A = B$ er gitt ved å vise at $A \subset B$ og $B \subset A$. Definisjon: **$\subset$, delmengde**

Anta at A_ι er en mengde for hvert element ι i en mengde I. Da er $\{A_\iota\}$, $\iota \in I$, en indeksert familie av mengder. Et eksempel er gitt ved $A_\iota = [\iota, \infty)$ for $\iota \in \mathbb{R}$, $\{A_\iota\} = \{[\iota, \infty)\}$, $\iota \in \mathbb{R}$. Definisjon: **$\{A_\iota\}$, indeksert familie**

Unionen og snittet av mengdene i familien $\{A_i\}$ er gitt ved henholdsvis

$$\bigcup_{i \in I} A_i = \{\omega \,|\, \exists i,\ \omega \in A_i\} \quad og \quad \bigcap_{i \in I} A_i = \{\omega \,|\, \forall i,\ \omega \in A_i\}.$$

Definisjon: **$\cup$, union** **$\cap$, snitt**

Tegnet $\exists$ leses som *det eksisterer en*. Tegnet $\forall$ leses som *for alle*. Unionen er dermed mengden av elementer i Ω som er inneholdt i minst en A_i og snittet er mengden av elementer som er inneholdt i alle A_i. Her har vi antatt at alle mengdene A_i er delmengder av en mengde Ω. To eksempler er gitt ved $\cup_{\iota \in \mathbb{R}} [\iota, \infty) = \mathbb{R}$ og $\cap_{\iota \in \mathbb{R}} [\iota, \infty) = \emptyset$. I noen sammenhenger er det underforstått eller uviktig hva indeksmengden I er og union og snitt skrives kortere på formen $\cup_i A_i$ eller $\cap_i A_i$.

Komplementet til A i Ω er gitt ved $A^c = \{\omega \,|\, \omega \notin A\}$, dvs komplementet til A er mengden av alle elementer i Ω som ikke er elementer i A. F eks gir $\Omega = \{1, 2, 3, 4\}$, $A = \{1, 3\}$ at $A^c = \{2, 4\}$. Definisjon: **$(\cdots)^c$, komplement**

Differansen mellom A og B er mengden av alle elementer som er inneholdt i A men ikke i B. Formelt noteres dette som $A \setminus B = \{\omega \,|\, \omega \in A, \omega \notin B\}$. Det følger at $A \setminus B = A \cap B^c$ og at $A^c = \Omega \setminus A$. Et eksempel er gitt ved $\mathbb{R} \setminus \mathbb{N} = (-\infty, 0] \cup (\cup_{n=1}^{\infty} (n-1, n))$. Definisjon: **$\setminus$, differans**

To mengder A og B sies å være disjunkte dersom de ikke har noen felles elementer, dvs dersom $A \cap B = \emptyset$. F eks er mengdene gitt ved intervallene $(0, 1)$ og $(1, 2)$ disjunkte. Definisjon: **disjunkt**

En union av to disjunkte mengder noteres som $A \uplus B$. Operasjonen $\uplus$ er kun definert for disjunkte mengder. Når vi skriver $A \uplus B$, så er det underforstått at mengdene A, B er disjunkte. Denne konvensjonen er sammenlignbar med at vi ofte skriver $f(x)$ for verdien til en funksjon f i et punkt x uten å eksplisitt skrive ned at vi antar at x er et element i definisjonsmengden til f. Notasjonen $\uplus_i A_i$ benyttes tilsvarende for unionen av en vilkårlig familie disjunkte mengder. Dette siste betyr at $i \neq j \Rightarrow [A_i \cap A_j = \emptyset]$. Et eksempel er gitt ved Definisjon: **$\uplus$, disjunkt union**

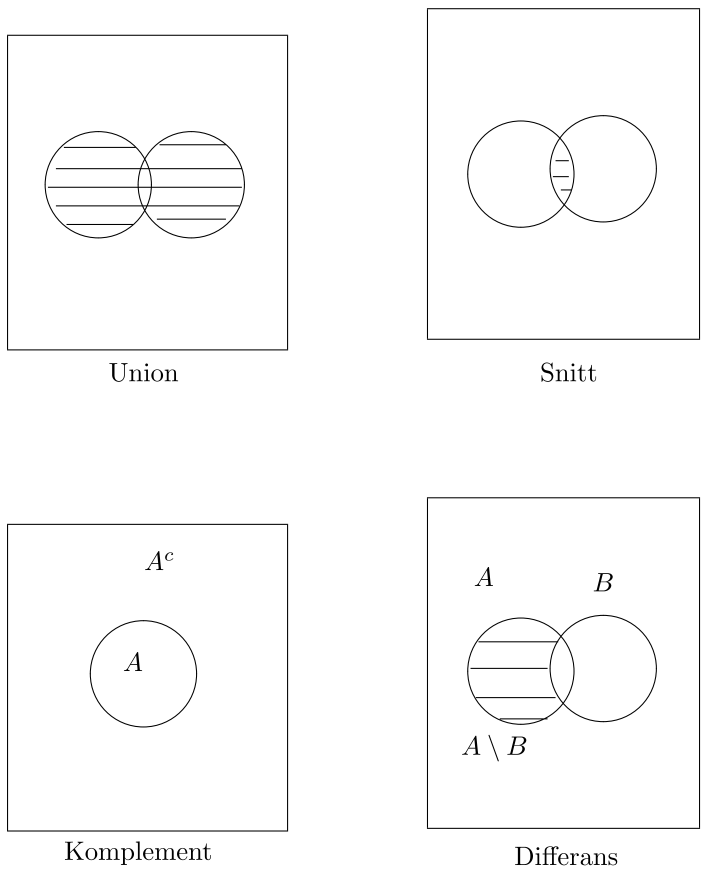
$$\biguplus_{t \in \mathbb{R}} \{t\}.$$

Relasjoner mellom mengder og operasjoner på mengder som gjelder generelt vil også måtte gjelde for delmengder av et rektangel i planet. Dette gir muligheten til illustrasjon av sammenhenger som ellers kan virke uoversiktlige i Venn-diagram. Noen eksempler er gitt i figur 1.2.

1.3 Oppgaver

1 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?



Figur 1.2: Venn-diagram.

2 La $\mathbb{Q}_+$ være mengden av alle positive rasjonale tall, dvs $\mathbb{Q} = \{m/n \,|\, m, n \in \mathbb{N}\}$. Er $\mathbb{Q}_+$ tellbar? Er mengden av rasjonale tall tellbar?

3 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?

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

5 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

Sjanse-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

Følgende definisjoner bygger på mengdeteorien og er utgangspunktet for Kolmogorovs matematiske formulering av sannsynlighetsteorien.

Definisjon: **Utfallsrom** **Utfall**

> Et utfallsrom er en ikke-tom mengde. Et utfall er et element i utfallsrommet.

Et eksempel på et utfallsrom Ω er gitt ved enhetsintervallet, dvs $\Omega = [0, 1]$. I dette eksemplet er $\omega = 1/2$ et utfall, mens $\omega = -1$ er ikke et utfall. Et annet eksempel er gitt ved utfallsrommet $\Omega = \mathbb{N}$. I så fall er ikke $\omega = 1/2$ et utfall, mens $\omega = -1$ er et utfall.

Definisjon: **Hendelsesfamil** **hendelse**

> En familie av delmengder av utfallsrommet er en hendelsesfamilie dersom
>
> 1. Komplementet til en hendelse er en hendelse.
> 2. En tellbar union av hendelser er en hendelse.
>
> En hendelse er et medlem av hendelsesfamilien.

Merk: En familie $\mathcal{E}$ av delmengder er det samme som en mengde av delmengder, dvs at dersom $A \in \mathcal{E}$ så er A en mengde. Spesielt betyr dette at en hendelse er en delmengde av utfallsrommet.

De to aksiomene for en hendelsesfamilie $\mathcal{E}$ kan skrives

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

hvor det siste aksiomet skal gjelde for enhver tellbar familie $\{A_i\}$ av hendelser.

Begrepet hendelsesfamilie er identisk med begrepet σ-algebra av mengder. Denne siste betegnelsen er den vanlige i generell integrasjonsteori. Vi benytter betegnelsen hendelsesfamilie fordi dette ligger nærmere språkbruken ellers i sannsynlighetsteorien.

Det følger av aksiomene at den tomme mengden er en hendelse fordi $\emptyset = \cup_{i\in\emptyset} A_i$, og dermed er utfallsrommet selv også en hendelse. Et tellbart snitt av hendelser er en hendelse fordi $\cap_i A_i = [\cup_i A_i^c]^c$.

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

Eksempel 2.1.1 (Eksperiment: Tre myntkast i rekkefølge) Resulatet av et enkelt myntkast er M (mynt) eller K (krone). Utfallsrommet for eksperimentet gitt av tre myntkast i rekkefølge er $\Omega = \{MMM, MMK, MKM, MKK, KMM, KMK, KKM, KKK\}$ og består dermed av 8 enkeltutfall. Vi lar hendelsesfamilien $\mathcal{E}$ være familien av alle delmengder. La $E = \{MKK, KMK, KKM, KKK\}$ og anta at utfallet av eksperimentet ble $\omega = MKK$. Da inntraff hendelsen E, dvs vi fikk $\omega \in E$. Å si at hendelsen E inntraff er det samme som å si at vi fikk flere krone enn mynt.

Et alternativt utfallsrom i dette eksemplet er $\tilde{\Omega} = \{000, 001, 010, 011, 100, 101, 110, 111\}$ ved å identifisere 0 med mynt og 1 med krone. Fordelen med dette valget er at alle de mulige utfallene kommer fram ved å telle binært.

Eksempel 2.1.2 (Eksperiment: Gjentatt myntkast inntil resultatet er krone) Ut-fallsrommet er $\Omega = \{K, MK, MMK, MMMK, \ldots\}$ og vi kan la hendelsesfamilien $\mathcal{E}$ være familien av alle delmengder.

Eksempel 2.1.3 (Eksperiment: Levetiden til en lyspære) La eksperimentet være å måle tiden det tar før en nyinnkjøpt lyspære brenner ut. Det naturlige utfallsrommet er mengden av alle positive reelle tall, dvs $\Omega = (0, \infty)$. Det er naturlig å kreve at mengden $E = [a, \infty)$ er en hendelse for alle vilkårlige $a \geq 0$ fordi dette er hendelsen at levetiden er større enn eller lik a. Dette kravet gir at enhver union av intervall, at alle åpne delmengder og at alle lukkede delmengder av Ω er hendelser. Den naturlige familien $\mathcal{E}$ av hendelser er den minste familien av hendelser som inneholder alle overnevnte hendelser.

Definisjon: **Borel-mengde**

I det foregående eksemplet var den naturlige familien $\mathcal{E}$ av hendelser lik den minste familien av hendelser som inneholder alle åpne mengder. Dette er det samme som at $\mathcal{E}$ er familien av Borelmengder i Ω. Eksistensen og entydigheten til slike familier behandles i videregående kurs. Her nøyer vi oss med å fastslå at det finnes delmengder av Ω som ikke er Borelmengder. Dette gjelder ikke bare for $\Omega = \mathbb{R}$ (talllinjen), men også for $\Omega = \mathbb{R}^2$ (planet), $\Omega = \mathbb{R}^3$ (rommet) og for $\Omega = \mathbb{R}^n$ (generelt euklidsk rom). Dersom en insisterer på at alle delmengder skal være hendelser i disse eksemplene så leder dette til problemer når alle disse skal tilordnes en sannsynlighet.

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. Spesielt vil aldri ordet eksperiment bli definert som et matematisk begrep.

> Med et eksperiment mener vi enhver prosess hvor de mulige resultatene av eksperimentet kan identifiseres med elementer i en mengde. Denne mengden er utfallsrommet for eksperimentet.

Identifiseringen av utfallsrommet for et eksperiment er første skritt i den matematiske beskrivelsen av eksperimentet.

2.2 Sannsynlighet og statistiske eksperiment

La Ω være et utfallsrom med en hendelsesfamilie $\mathcal{E}$. En fordeling P for Ω er en funksjon slik at enhver hendelse E er tilordnet en sannsynlighet $P(E) \geq 0$ med $P(\Omega) = 1$ og slik at addisjonsregelen

$$P(\biguplus_i A_i) = \sum_i P(A_i)$$

gjelder for enhver tellbar familie $\{A_i\}$ av disjunkte hendelser.
Et sannsynlighetsrom er et utfallsrom Ω utstyrt med en hendelsesfamilie $\mathcal{E}$ og en fordeling P.

Definisjon: **fordeling, addisjons-regelen, sannsynlig-hetsrom**

En fordeling er dermed en funksjon $P : \mathcal{E} \to [0, \infty)$ som oppfyller to krav. Dersom normaliseringskravet $P(\Omega) = 1$ fjernes, så sies P å være et mål. En fordeling er dermed et normalisert mål. Dersom 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.

Ligningen $\sum_{i \in \emptyset} PA_i = 0$ gir

$$\boxed{P(\emptyset) = 0}$$

Spesialtilfellet $P(\biguplus_{C \in \{A,B\}} C) = PA + PB$ gir

$$\boxed{P(A \uplus B) = PA + PB}$$

Inklusjonen $A \subset B$ gir $B = A \uplus (B \setminus A)$, så $PB = PA + P(B \setminus A) \geq PA$, dvs P er monotont voksende:

$$\boxed{A \subset B \Rightarrow PA \leq PB}$$

Spesialtilfellet $A \subset \Omega$ av det foregående gir

$$\boxed{PA \leq 1}$$

En generell union kan skrives som en disjunkt union ved $A \cup B = (A \setminus B) \uplus (A \cap B) \uplus (B \setminus A)$. Dette gir $P(A \cup B) = P(A \setminus B) + P(A \cap B) + P(B \setminus A)$. Ved $B = (B \setminus A) \uplus (A \cap B)$ og addisjonsregelen følger $P(B \setminus A) = PB - P(A \cap B)$, og $P(A \setminus B) = PA - P(A \cap B)$ bevises tilsvarende. Innsetting gir den mer generelle addisjonsregelen

$$\boxed{P(A \cup B) = PA + PB - P(A \cap B)}$$

I mange problem kan det være enklere å beregne sannsynligheten til den komplementære hendelsen. Likheten $P(\Omega) = P(A \uplus A^c) = PA + PA^c$ gir at sannsynligheten til hendelsen deretter kan finnes ved:

$$\boxed{P(A) = 1 - P(A^c)}$$

En fordeling benyttes for å beskrive graden av usikkerhet i sjanse-eksperiment.

Vi ser på noen eksempler for å belyse dette.

Eksempel 2.2.1 (Sjanse-Eksperiment: 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 2.2.2 (Sjanse-Eksperiment: 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 2.2.3 (Sjanse-Eksperiment: 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 sjanse-eksperiment.

Et sjanse-eksperiment er et eksperiment hvor det er knyttet usikkerhet til utfallet av eksperimentet.

Dersom resultatet av et eksperiment er ω, så sier vi at hendelsen E har inntruffet dersom $\omega \in E$. En mulig modell for et sjanse-eksperiment er gitt ved at utfallsrommet er et sannsynlighetsrom. Fordelingen P gir sannsynligheten $P(E)$ for at hendelsen E inntreffer. Sannsynligheten tolkes i dette tilfellet som en vurdering av rimeligheten for at hendelsen E inntreffer.

En sannsynlighetsmodell for et sjanse-eksperiment er gitt ved en 1-1 korrespondanse mellom de mulige utfallene i eksperimentet og utfallene i et sannsynlighetsrom.

Bortsett fra fotballkamp eksemplet, så kan alle overnevnte eksperiment gjentas et vilkårlig antall ganger.

Et statistiske eksperiment er et eksperiment som kan gjentas et vilkårlig antall ganger.

Definisjon: **relativ hyppighet**

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

$$P_n(A) = \frac{nA}{n}.$$

Det følger at P_n er en fordeling (bevis dette!).

Definisjon: **statistisk lovmessig eksperiment**

Et eksperiment er et statistisk lovmessig eksperiment dersom grenseverdien

$$PE = \lim_{n \to \infty} P_n(E),$$

eksisterer for alle hendelser E. Det følger at dette definerer en endelig additiv fordeling P. Det kan også bevises (Vitali-Hahn-Saks teoremet) at P er en fordeling. Denne unike fordelingen er den empiriske fordelingen. Konklusjonen er da:

I et statistisk lovmessig eksperiment er enhver hendelse E tilordnet en entydig empirisk sannsynlighet $P(E)$. Vi sier i så fall at hendelsen E inntreffer i $P(E) \cdot 100\%$ av eksperimentene når eksperimentet gjentas flere ganger.

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. I praksis vil observasjonen av en statistisk lovmessighet være gitt ved at den relative frekvensen til en hendelse synes å nå en grenseverdi når eksperimentet gjentas mange ganger. De filosofiske sidene knyttet til hva vi egentlig mener med at eksperimentet kan gjentas et vilkårlig antall ganger lar vi vile.

2.3 Oppgaver

6 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?

7 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".

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

9 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".

10 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?

11 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 $P(A^c \cap B)$.

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

13 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.

14 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.

15 Vis at den relative hyppigheten er en fordeling.

Vis at den endelige addisjonsregelen gjelder for en empirisk fordeling.

16 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.

Anta at $\{\mathcal{E}_i\}$ er familien av hendelsesfamilier i Ω som inneholder alle åpne mengder. Vis at $\mathcal{G}$ er familien av Borel-mengder.

Kapittel 3

Sannsynlighetstetthet

Tyche, in Greek religion, the goddess of chance, with whom the Roman Fortuna was later identified; a capricious dispenser of good and ill fortune. The Greek poet Hesiod called her the daughter of the Titan Oceanus and his consort Tethys; other writers attributed her fatherhood to Zeus, the supreme god. She was also associated with the more beneficent Agathos Daimon, a good spirit, protective of individuals and families, and with Nemesis, who, as an abstraction, represented punishment of overprosperous man and so was believed to act as a moderating influence. She was often shown winged, wearing a crown, and bearing a sceptre and cornucopia; but she also appeared blindfolded and with various devices signifying uncertainty and risk. Among her monuments was a temple at Argos, where the legendary Palamedes is said to have dedicated to her the first set of dice, which he is supposed to have invented.

Britannica Online (1997)

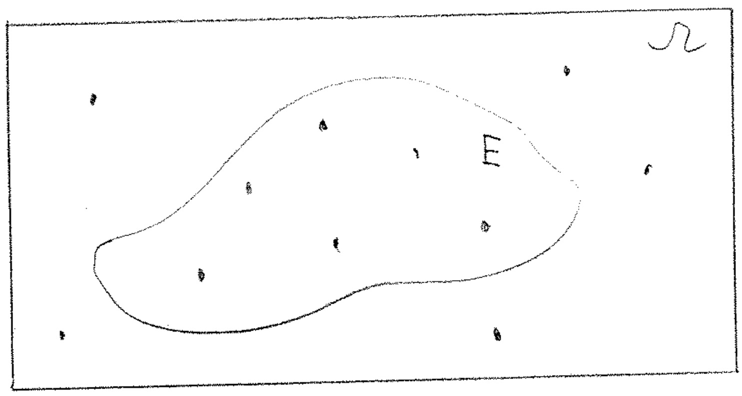
Fortuna, in Roman religion, goddess of chance or lot who became identified with the Greek Tyche; the original Italian deity was probably regarded as the bearer of prosperity and increase. As such she resembles a fertility deity, hence her association with the bounty of the soil and the fruitfulness of women. Frequently she was an oracular goddess consulted in various ways regarding the future. Fortuna was worshiped extensively in Italy from the earliest times. At Praeneste her shrine was a well-known oracular seat, as was her shrine at Antium. Fortuna is often represented bearing a cornucopia as the giver of abundance and a rudder as controller of destinies, or standing on a ball to indicate the uncertainty of fortune.

Britannica Online (1997)

En funksjon f er en tetthet for fordelingen P med hensyn på μ dersom sannsynligheter kan beregnes ved integrasjon som ved

$$P(A) = \int_A f(\omega)\mu(d\omega).$$

Vi skal se på tilfellet hvor $\mu A =$ antall punkter i hendelsen A. I dette tilfellet er integralet over definert ved summasjon og vi sier at f er en diskret tetthet. Et annet tilfelle er gitt ved at utfallsrommet er tallinjen og $\mu A =$ størrelsen til A, hvor spesielt $\mu[a, b] = b - a$. Dersom f er stykkevis kontinuerlig er integralet over gitt ved Riemann-integralet. I dette tilfellet er f en tetthet for $\mathbb{R}$. Et poeng med innføringen av tettheter er at det kan være enklere å spesifisere en fordeling P ved en tetthet f fremfor å spesifisere P direkte.



Figur 3.1: Atomene til en diskret fordeling og en hendelse.

3.1 Diskrete fordelinger

Definisjon: **diskret tetthet**

La Ω være et utfallsrom. *En funksjon $f : \Omega \to [0, \infty)$ er en diskret sannsynlighetstetthet dersom*

$$\sum_{\omega \in \Omega} f(\omega) = 1$$

Et utfall ω er et atom dersom $f(\omega) > 0$. Det kan vises at det overstående gir det finnes høyst et tellbart antall atomer $\omega_1, \omega_2, \ldots$.

Definisjon: **diskret fordeling**

En fordeling P er en diskret fordeling dersom det finnes en diskret sannsynlighetstetthet f slik at

$$P(A) = \sum_{\omega \in A} f(\omega).$$

Spesielt gir dette at $P\{\omega\} = f(\omega)$. Dersom f er en diskret sannsynlighetstetthet, så kan P defineres ved formelen over. En diskret fordeling for Ω er karakterisert av atomene $\omega_1, \omega_2, \ldots$ og sannsynligheten $p_i = P\{\omega_i\}$ til hvert atom ω_i. Dette er illustrert i figur 3.1.

Eksempel 3.1.1 (Tettheten tilsvarende to myntkast) Utfallsrommet er $\Omega = \{MM, MK, KM, KK\}$, og det er rimelig å anta at alle utfallene er like sannsynlige. Dette gir $P\{\omega\} = 1/4$. Definer f ved $f(\omega) = P\{\omega\}$. Det følger at f er en diskret sannsynlighetstetthet og at $P(A) = \sum_{\omega \in A} f(\omega)$. Dermed er P en diskret fordeling.

Definisjon: **diskret utfallsrom**

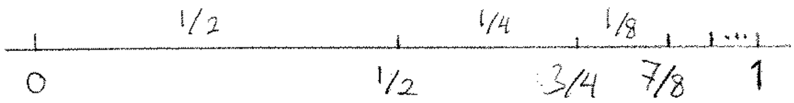
Et utfallsrom Ω er et diskret utfallsrom dersom Ω er tellbar og er utstyrt med en hendelsesfamilie slik at alle enpunktsmengder $\{\omega\}$ er hendelser. Eksperimentet over kunne beskrives ved en diskret sannsynlighetstetthet fordi utfallsrommet var diskret. Dersom Ω er et diskret utfallsrom, så følger det generelt at P er diskret. Det neste eksemplet viser at P kan være diskret uten at Ω er diskret.

Eksempel 3.1.2 (Endelig empirisk tetthet ved pilkast) Utfallsrommet er $\Omega = \{BOM\} \cup \{(x, y) \mid x^2 + y^2 \leq R^2\}$. Her er x, y pilens horisontale og vertikale koordinater med origo $(0, 0)$ i sentrum av blinken som har radius R. Anta at en person har kastet n ganger med resultat $\omega_1, \ldots, \omega_n$. En mulig fordeling for det neste pilkastet er gitt ved $PA = nA/n$, hvor nA er antall ganger A inntraff i de foregående pilkastene. Med andre ord er nA antall j slik at ω_j er i A. Formelt kan dette skrives som $nA = \sum_{i=1}^{n} 1_A(\omega_i)$, hvor 1_A er indikatorfunksjonen til A. Tettheten er gitt ved

$$f(\omega) = \frac{n\{\omega\}}{n} = \sum_i \frac{1_{\omega_i}(\omega)}{n}.$$

Definisjon: **diskret uniform fordeling**

En diskret fordeling er uniform dersom det kun er et endelig antall ulike atomer $\omega_1, \ldots, \omega_n$ og $P\{\omega_i\} = p$ uavhengig av i. Det følger at $p = 1/n$. Dette er den viktigste diskrete fordelingen. I gode tilfeller kan symmetriargumenter benyttes for å konkludere at en uniform fordeling er rimelig.



Figur 3.2: Den geometriske rekken $1/2 + 1/4 + 1/8 + \cdots = 1$.

F eks vil vi anta at det er 1/6 sannsynlighet for å få en 5'er i et terningkast dersom terningen er symmetrisk, dvs dersom det ikke er noen forskjell på de seks sidene på terningen. Det neste eksemplet viser at uniforme fordelinger kan benyttes til utledning av andre fordelinger.

Eksempel 3.1.3 (Myntkast inntil resultatet er krone) Det diskrete utfallsrommet $\Omega = \{K, MK, MMK, \ldots\}$ bør være selvforklarende. Vi vil nå finne en fordeling ved å finne sannsynligheten til hvert enkeltutfall. Vi antar at mynten er symmetrisk og konkluderer at $P\{K\} = 1/2$. Konklusjonen $P\{MK\} = 1/4$ følger fordi hver av de fire utfallene MM, MK, KM, KK i eksperimentet hvor en mynt kastes 2 ganger er like sannsynlige ved symmetri med hensyn til kastrekkefølgen og myntens to sider. Utfallet MK i dette siste eksperimentet gir sannsynligheten for utfallet MK i Ω fordi betingelsene for dette utfallet er like i de to eksperimentene. Tilsvarende kan sannsynligheten for hvert enkeltutfall i Ω finnes ved sannsynligheten for et tilsvarende utfall i et eksperiment med 2^n like sannsynlige utfall. Dette gir generelt at $P\{\omega\} = 1/2^n$ hvor n er antall myntkast tilsvarende ω. Vi kan kontrollere at P er normalisert ved $P\Omega = 1/2 + 1/4 + 1/8 + \cdots = 1$. Dette summasjonsresultatet er illustrert i figur 3.2. Generelt er summen av en geometrisk rekke $a + ak + ak^2 + \cdots = a/(1-k)$ dersom $|k| < 1$. Vi kan nå regne ut sannsynligheten for mer kompliserte hendelser. F eks er sannsynligheten for at vi må kaste mynten et odde antall ganger før vi får krone lik 2/3:

$$P\{K, MMK, MMMMK, \ldots\} = \sum_{i=0}^{\infty} \left(\frac{1}{2}\right)^{2i+1} = \frac{1}{2} \sum_{i=0}^{\infty} \left(\frac{1}{4}\right)^{i} = \frac{1}{2} \frac{1}{1-1/4} = 2/3.$$

3.2 Kontinuerlige fordelinger

La utfallsrommet være $\mathbb{R}$ = "de reelle tall". *Fordelingen P er kontinuerlig dersom* Definisjon: **kontinuerlig fordeling**

$$P(A) = \int_A f(x)dx.$$

Generelt defineres $\int_A f(x)dx = \int 1_A(x)f(x)dx$ hvor 1_A er indikatorfunksjonen til A. Funksjonen f er sannsynlighetstettheten til sannsynlighetsfordelingen P. Kravet $P(A) \geq 0$ gir at $f \geq 0$. Normaliseringen $P\Omega = 1$ gir normaliseringsbetingelsen

$$\int_{-\infty}^{\infty} f(x)dx = 1,$$

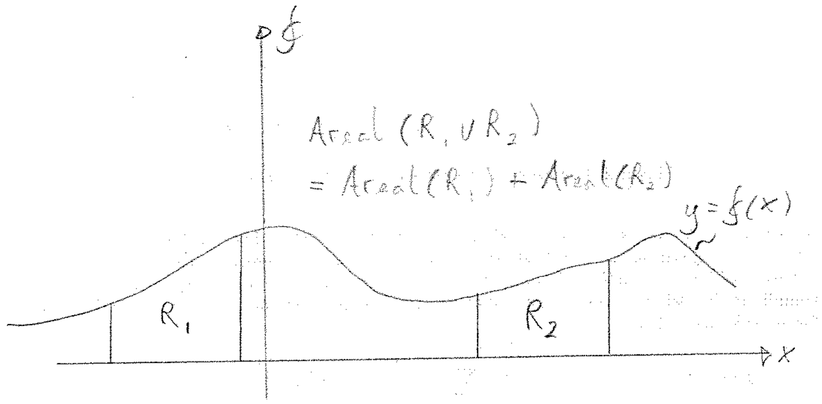
dvs f er normalisert. *Generelt er en tetthet f for tallinjen en funksjon $f : \mathbb{R} \to [0, \infty)$ som er normalisert.* Det er en-en korrespondanse mellom kontinuerlige tettheter og kontinuerlige fordelinger. Hendelsesfamilien tilsvarende P er familien av alle Borel-mengder, dvs den minste hendelsesfamilien som inneholder alle intervall. Definisjon: **tetthet for** $\mathbb{R}$

Det at en kontinuerlig tetthet definerer en fordeling er en egenskap til integralet. Addisjonsregelen kan bevises ved

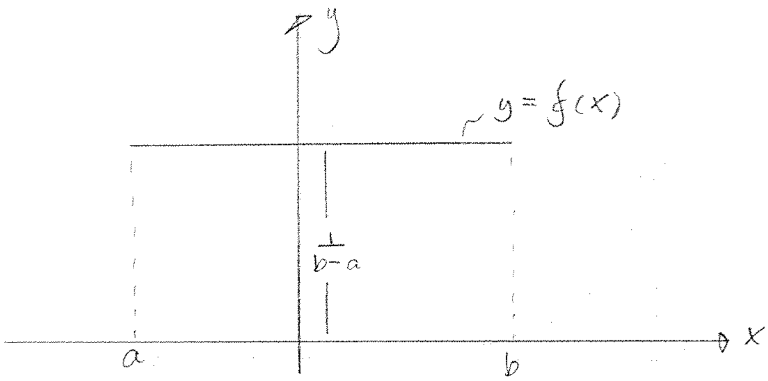
$$P(\uplus_{i=1}^{\infty} E_i) = \int_{\uplus_{i=1}^{\infty} E_i} f(x)dx = \int 1_{\uplus_{i=1}^{\infty} E_i}(x)f(x)dx = \int \sum_{i=1}^{\infty} 1_{E_i}(x)f(x)dx = \sum_{i=1}^{\infty} \int 1_{E_i}(x)f(x)dx$$
$$= \sum_{i=1}^{\infty} P(E_i).$$

Den endelige addisjonregelen er illustrert i figur 3.3.

Noen tettheter er så viktige at de har fått egne navn. Vi vil nevne tre eksempler her.



Figur 3.3: Addisjonregelen for en kontinuerlig tetthet tilsvarer addisjon av areal under grafen.



Figur 3.4: Tettheten til en uniform fordeling.

Definisjon: **uniform fordeling**

Tettheten til den uniforme fordelingen på $[a, b]$ *er gitt ved (se figur 3.4)*

$$f(x) = \frac{1}{b-a},\ a \leq x \leq b.$$

Det er her underforstått at $f(x) = 0$ når $x \notin [a, b]$.

Definisjon: **standard normal-fordeling**

Tettheten til standard normalfordelingen er gitt ved (se figur 3.5)

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}.$$

Definisjon: **eksponensial-fordelin-gen**

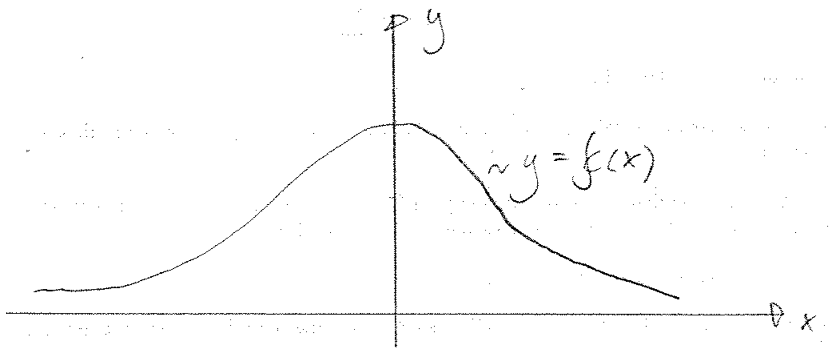
Tettheten til eksponensialfordelingen er gitt ved (se figur 3.6)

$$f(x) = \frac{1}{\lambda} e^{-\frac{x}{\lambda}}, x > 0. \tag{3.1}$$

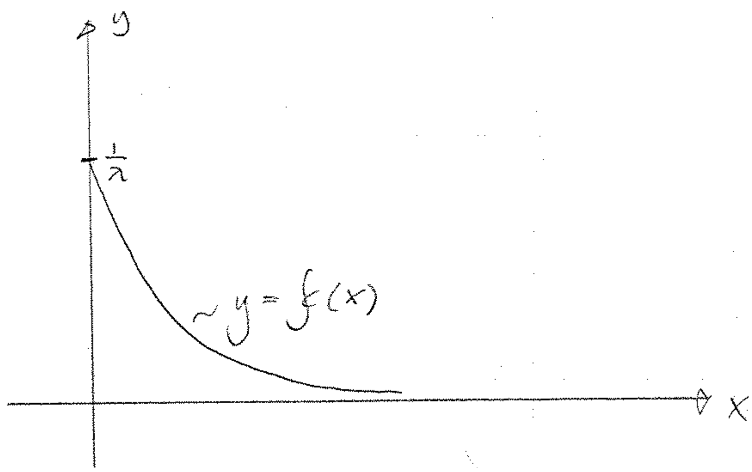
Det er her underforstått at $f(x) = 0$ når $x \leq 0$.

Eksempel 3.2.1 (Fri veilengde) Den frie veilengden er avstanden et molekyl tilbakelegger mellom to kollisjoner med andre molekyler i en gass. Det kan argumenteres for at den frie veilengden er eksponensialfordelt med parameter λ. Vi definerer

$$\text{midlere fri veilengde } = \int x f(x) dx = \int_0^\infty x \frac{1}{\lambda} e^{-x/\lambda} dx \stackrel{delv.int.}{=} \lambda,$$



Figur 3.5: Tettheten til standard normalfordeling.



Figur 3.6: Tettheten til en eksponensialfordeling.

dvs parameteren λ er den midlere frie veilengden. Vi kan nå f eks beregne sannsynligheten for at den frie veilengden er mindre eller lik halvparten av den midlere frie veilengden:

$$P(\text{fri veilengde} \leq \lambda/2) = \int_0^{\lambda/2} \frac{1}{\lambda} e^{-x/\lambda} dx = \left[\frac{1}{\lambda} (\text{-}\lambda) \mathrm{e}^{\text{-x}/\lambda} \right]_0^{\lambda/2} = 1 - e^{-1/2} \approx \underline{\underline{0,39}}.$$

3.3 Oppgaver

17 La $\lambda \geq 0$. Vis at

$$f(n) = \frac{1}{1+\lambda} \left(\frac{\lambda}{1+\lambda} \right)^n, n = 0, 1, 2, \ldots$$

definerer en diskret tetthet for utfallsrommet $\mathbb{R}$. Er dette utfallsrommet diskret? Hva med tilfellet $\lambda = 0$?

18 La f være en diskret sannsynlighetstetthet. Vis at da er P definert ved

$$P(A) = \sum_{\omega \in A} f(\omega)$$

en sannsynlighetsfordeling.

19 Vis at dersom utfallsrommet er diskret, så må fordelingen også være diskret. Gjelder det motsatte?

20 Anta at P er den uniforme fordelingen på et utfallsrom Ω. Finn tettheten til P når det er gitt at Ω inneholder 1001 elementer. Hva med tilfellet når $\Omega = \mathbb{N}$?

21 Vis at $1/2 + 1/4 + 1/8 + \cdots = 1$.

22 La k være en konstant. Finn P(Personen lever mer enn 60 år) når tettheten for personens alder er

$$f(t) = k \cdot t^2 (100 - t)^2, 0 \leq t \leq 100.$$

Hva verdien til k?

23 Finn sannsynligheten for at det er mellom 50 og 100 dager mellom to gruveulykker når det er gitt at tiden mellom to gruveulykker er eksponensialfordelt med parameter $\lambda = 241$ dager.

24 Vis at $1_{\uplus_i A_i} = \sum_i 1_{A_i}$.

25 La utfallsrommet være tallinjen $\mathbb{R}$. Vis at det er en-en korrespondanse mellom kontinuerlige fordelinger og tettheter for $\mathbb{R}$.

Kapittel 4

Betinget sannsynlighet og uavhengighet

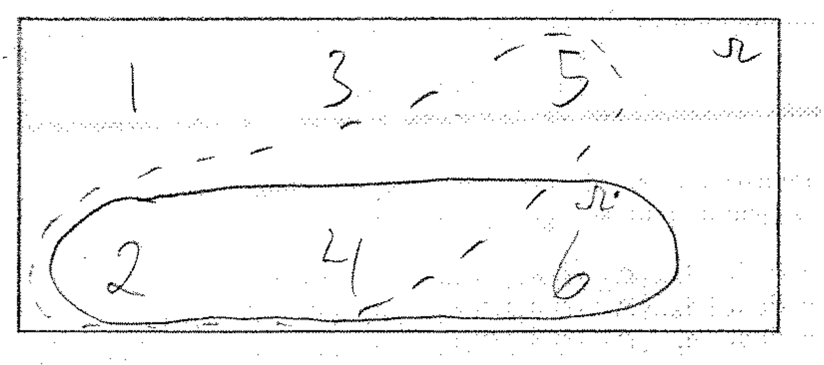
The concept of mutual independence of two or more experiments holds, in a certain sense, a central position in the theory of probability.
A.N. Kolmogorov (1933)

Uavhengighet av hendelser er et meget viktig begrep. Begrepet er tildels utgangspunktet for at sannsynlighetsteorien har utviklet seg til noe mer enn det som finnes i den grunnleggende integrasjonsteorien. To hendelser A, B som har positiv sannsynlighet er uavhengige dersom den betingede sannsynligheten til A gitt B er lik sannsynligheten til A. Med dette som utgangspunkt må vi først si litt om betingede sannsynligheter.

4.1 Betinget sannsynlighet

Eksempel 4.1.1 (Terningkast med informasjon) En terning kastes og det er oppgitt at resultatet er et partall. Noen rimelige vurderinger er da gitt ved $P\{1\} = 0$ (1 er ikke et partall.), $P\{2\} = 1/3$ (Det finnes 3 mulige partall.), og $P\{2, 4, 5\} = 2/3$ (Hendelsen inneholder 2 av de 3 mulige partallene.). Figur 4.1 kan være en hjelp til å forstå den siste vurderingen.

Anta at vi har et statistisk lovmessig eksperiment, dvs sannsynligheten for en hendelse er det samme som det relative antall ganger hendelsen inntreffer når eksperimentet gjentas. Sannsynligheten $P(A \,|\, \Omega')$ for at A inntreffer gitt at Ω' inntreffer er gitt ved å kun telle utfallene innenfor



Figur 4.1: Terningkastet resulterte i at Ω' inntraff.

Ω'. Dette leder til

$$P(A \mid \Omega') = \lim \frac{n(A \cap \Omega')}{n\Omega'} = \lim \frac{\frac{n(A\cap\Omega')}{n}}{\frac{n\Omega'}{n}} = \frac{P(A \cap \Omega')}{P(\Omega')}.$$

Definisjon: **Betinget sannsynlighet**

Denne regningen motiverer en generell definisjon. *Sannsynligheten til A gitt B defineres generelt ved*

$$\boxed{P(A \mid B) = \frac{P(A \cap B)}{P(B)},\ P(B) \neq 0.}$$

Definisjonen av den betingede sannsynligheten $P(A \mid B)$ gjelder generelt når P er en fordeling. Når B fikseres gir dette en ny fordeling Q ved $Q(A) = P(A \mid B)$. Vi vil bevise at Q er en fordeling. Vi har $Q(A) \geq 0$ og $Q(\Omega) = P(\Omega \cap B)/P(B) = 1$. Addisjonsregelen følger fordi $(\uplus_i A_i) \cap B = \uplus_i (A_i \cap B)$

$$Q(\uplus_i A_i) = \frac{P((\uplus_i A_i) \cap B)}{P(B)} = \frac{P(\uplus_i (A_i \cap B))}{P(B)} = \frac{\sum_i P(A_i \cap B)}{P(B)} = \sum_i \frac{P(A_i \cap B)}{P(B)}$$
$$= \sum_i Q(A_i).$$

Eksempel 4.1.2 (Kortstokk) Et kort trekkes fra en kortstokk. Hva er sannsynligheten for at det er et kløverkort gitt at det er en konge?

Svar a): Utfallsrommet består av de fire kongene, og en av disse er kløver. Sannsynligheten er dermed 1/4.

Svar b): Det opprinnelige utfallsrommet kan også brukes ved

$$P(\text{kløver} \mid \text{konge}) = \frac{P(\text{kløver og konge})}{P(\text{konge})} = \frac{\frac{1}{52}}{\frac{4}{52}} = \frac{1}{4}.$$

Eksempel 4.1.3 (Levealder) Anta at levealderen til en person er gitt ved tettheten

$$f(t) = 3 \cdot 10^{-9} \cdot t^2 (100 - t)^2,\ 0 \leq t \leq 100.$$

Hva er sannsynligheten for at en person dør med en alder mellom 80 og 85 år gitt at levetiden er > 70 år? Denne betingede sannsynligheten kan beregnes ved

$$P(80 < \text{levetid} < 85 \mid \text{levetid} > 70) = \frac{P(80 < \text{levetid} < 85 \cap \text{levetid} > 70)}{P(\text{levetid} > 70)}$$
$$= \frac{P(80 < \text{levetid} < 85)}{P(\text{levetid} > 70)} = \frac{\int_{80}^{85} f(t)dt}{\int_{70}^{100} f(t)dt} \approx 0,19$$

Definisjonen av $P(A \mid B)$ har en viktig omskrivning

$$\boxed{P(A \cap B) = P(A \mid B)P(B).}$$

Mer generelt kan sannsynligheten til snittet av n hendelser finnes ved produktet av n sannsynligheter ved gjentatt bruk av regneregelen over

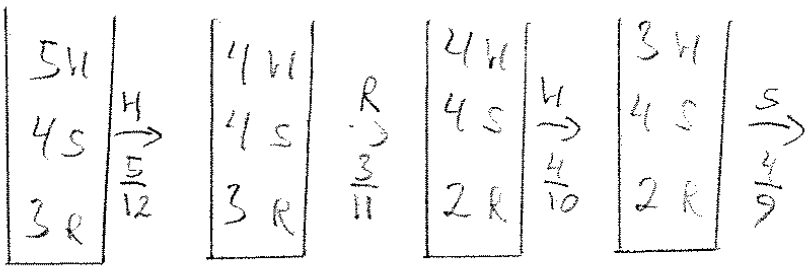
$$\begin{aligned} P(A_n \cap \cdots \cap A_1) &= P(A_n \mid A_{n-1} \cap \cdots \cap A_1)P(A_{n-1} \cap \cdots \cap A_1) \\ &= P(A_n \mid A_{n-1} \cap \cdots \cap A_1)P(A_{n-1} \mid A_{n-2} \cap \cdots \cap A_1)P(A_{n-2} \cap \cdots \cap A_1) \\ &= P(A_n \mid A_{n-1} \cap \cdots \cap A_1)P(A_{n-1} \mid A_{n-2} \cap \cdots \cap A_1) \cdots P(A_2 \mid A_1)P(A_1). \end{aligned}$$

Vi tar med et eksempel på bruk av denne regneregelen.

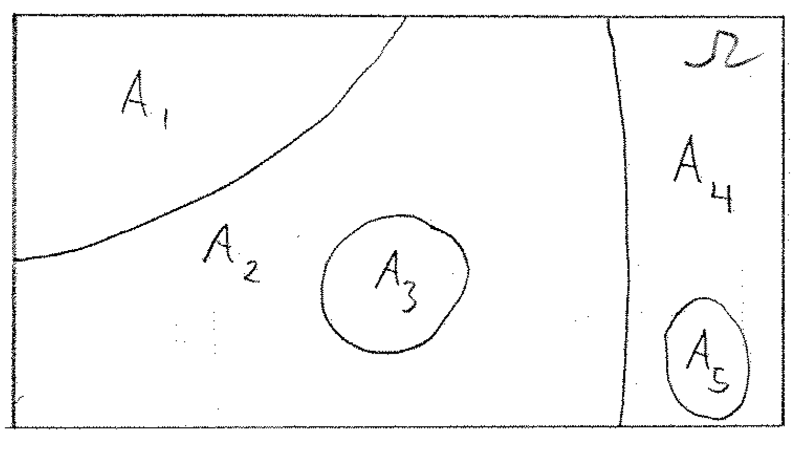
Eksempel 4.1.4 (Snittet av 4 hendelser ved betingede sannsynligheter) En urne inneholder 5 hvite brikker, 4 sorte brikke og 3 røde brikker. Hva er sannsynligheten for å trekke sekvensen (hvit, rød, hvit, sort) når 4 brikker trekkes uten tilbakelegning?

Definer hendelsene A_1 = "hvit trekkes i 1.trekning", A_2 = "rød trekkes i 2.trekning", A_3 = "hvit trekkes i 3.trekning" og A_4 = "sort trekkes i 4.trekning". Den søkte sannsynligheten er gitt ved (se figur 4.2)

$$\begin{aligned} P(A_1 \cap A_2 \cap A_3 \cap A_4) &= P(A_1)P(A_2 \mid A_1)P(A_3 \mid A_1 \cap A_2)P(A_4 \mid A_1 \cap A_2 \cap A_3) \\ &= \frac{5}{12}\frac{3}{11}\frac{4}{10}\frac{4}{9} \approx 0,02. \end{aligned}$$



Figur 4.2: Betingede sannsynligheter ved trekking av 4 brikker fra en urne.



Figur 4.3: En partisjon av Ω.

Definisjon: **partisjon**

En familie $\{A_i\}$ av hendelser er en partisjon (= oppdeling) av Ω dersom $\Omega = \uplus_i A_i$. En partisjon med 5 hendelser er illustrert i figur 4.3. Dersom $\{A_i\}$ er en tellbar partisjon av Ω, så følger det at

$$P(B) = P(B \cap \Omega) = P(B \cap (\uplus_i A_i)) = P(\uplus_i (B \cap A_i)) = \sum_i P(B \cap A_i)$$
$$= \sum_i P(B \,|\, A_i) P(A_i).$$

Innsetting av dette gir Bayes formel:

$$P(A_j \,|\, B) = \frac{P(A_j \cap B)}{P(B)} = \frac{P(B \,|\, A_j) P(A_j)}{\sum_i P(B \,|\, A_i) P(A_i)}.$$

Formelen gir sannsynligheten for A_j gitt B ved sannsynlighetene for B gitt A_i og sannsynlighetene for A_i. Legg merke til at betingelsene er snudd!

Eksempel 4.1.5 (Valg) Et valg mellom 1 republikaner og 3 demokrater foregår ved at 1 av demokratene velges ut først i et primærvalg. Sannsynligheten for at demokrat 1, 2, eller 3 vinner primærvalget estimeres til å være henholdsvis 35%, 40% og 25%. Hva er sannsynligheten for at republikaneren blir sittende i Kongressen dersom republikaneren med sannsynlighet 40%, 35% og 60% vinner et valg mot henholdsvis demokrat 1, 2 og 3?

La B være hendelsen at republikaneren beholder sitt sete etter valget. Hendelsene A_1, A_2, A_3 tilsvarende at henholdsvis demokrat 1, 2 eller 3 vinner primærvalget gir en partisjon av utfallsrommet. Dermed har vi

$$P(B) = \sum_i P(B \,|\, A_i) P(A_i) = 0,40 \cdot 0,35 + 0,35 \cdot 0,40 + 0,60 \cdot 0,25 \approx 43\%.$$

Eksempel 4.1.6 (Varsellampe) Hva er sannsynligheten for at bilen du kjører har for lite olje gitt at oljelampen lyser rødt og indikerer at det er for lite olje?

Vi antar at sannsynligheten for at oljelampen lyser ved for lite olje er 99%. Videre antar vi at oljelampen i 2% av tilfellene lyser uten grunn. Tilslutt antar vi at det er 10% sjanse for at det er for lite olje på bilen.

Sannsynligheten for at oljelampen lyser er gitt ved

$$P(\text{varsel}) = P(\text{varsel og nok olje}) P(\text{nok olje}) + P(\text{varsel og lite olje}) P(\text{lite olje}).$$

Dette gir den søkte sannsynligheten

$$P(\text{lite olje} \,|\, \text{varsel}) = \frac{P(\text{lite olje og varsel})}{P(\text{varsel})} = \frac{P(\text{varsel} \,|\, \text{lite olje}) P(\text{lite olje})}{P(\text{varsel})}$$
$$= \frac{0,99 \cdot 0,10}{0,02 \cdot 0,90 + 0,99 \cdot 0,10} \approx 85\%.$$

4.2 Uavhengighet

Definisjon: **uavhengighet**

Hendelsene i familien $\{A_i\}$ er uavhengige dersom vi har den generelle produktregelen

$$P(A_{i_1} \cap \cdots \cap A_{i_n}) = P(A_{i_1}) \cdots P(A_{i_n}), \ i_j \neq i_k \text{ for } j \neq k.$$

Det viktigste tilfellet er gitt ved at hendelsene A, B er uavhengige dersom $P(A \cap B) = P(A)P(B)$, eller ekvivalent dersom $P(A \,|\, B) = P(A \cap B)/P(B) = P(A)$ i tilfellet hvor $P(B) \neq 0$.

> Dersom A og B er uavhengige er sannsynligheten til A gitt B lik sannsynligheten til A.

Eksempel 4.2.1 (Uavhengige myntkast) Utfallsrommet tilsvarende n myntkast er gitt ved $\Omega = \{k_1 \ldots k_n \,|\, k_j = K, M\}$. Hver av de 2^n mulige utfallene er like sannsynlige ved symmetri. Dermed er $P(A) = \sum_{\omega \in A} 2^{-n}$. La $A_1 =$ "K i 1.kast" og $A_2 =$ "M i 2.kast". Da er $P(A_1) = 2^{n-1}/2^n = 1/2$ og tilsvarende er $P(A_2) = 1/2$ og $P(A_1 \cap A_2) = 1/4$. Dette gir $P(A_1 \cap A_2) = P(A_1)P(A_2)$ og det er bevist at hendelsene A_1 og A_2 er uavhengige. Det kan bevises tilsvarende at dersom $A_1, \ldots, A_n$ er hendelser tilsvarende resultatene i henholdsvis myntkast nummer $1, \ldots, n$, så er hendelsene $A_1, \ldots, A_n$ uavhengige.

		Varsel		
		Sol	Overskyet	Regn
Været	Sol	30%	5%	5%
	Overskyet	4%	20%	2%
	Regn	10%	4%	20%

Tabell 4.1: Værmelding i forhold til tilsvarende observert vær.

Mynteksemplet ga et eksempel hvor vi kunne bevise at de gjentatte forsøkene var uavhengige. I det neste eksemplet er en antagelse om uavhengighet utgangspunktet for beregning av en sannsynlighet.

Eksempel 4.2.2 (Uavhengige lyskryss) På vei til jobben passerer Eva 4 lyskryss. Anta at hvert av lyskryssene lyser grønt 40 sekunder i løpet av hvert minutt. Anta videre at avstanden mellom lyskryssene er så stor at hvert av lyskryssene er uavhengige. Sannsynligheten for at Eva må stoppe minst tre ganger er da

$$\begin{aligned} P\{RRRR, GRRR, RGRR, RRGR, RRRG\} &= P\{RRRR\} + 4P\{RRRG\} \\ &= (20/60)^4 + 4(20/60)^3(40/60) = 1/9. \end{aligned}$$

4.3 Oppgaver

26 Tabell 4.1 gir sannsynligheter for snittet av hendelser definert av henholdsvis været og den tilsvarende værmeldingen til et meterologisk institutt. Hva er sannsynligheten for at det skal bli sol? Hva er sannsynligheten for regn dersom det var varslet sol? Hvor ofte gjetter instituttet galt? Hvordan ville du gå frem for å lage en tabell tilsvarende Tabell 4.1?

27 La fordelingen P være eksponensialfordelingen med parameter λ. Vis at

$$P((s+t,\infty)\,|(t,\infty)) = P(s,\infty).$$

Hva betyr dette resultatet dersom fordelingen tilsvarer tiden mellom gruveulykker?

28 Hva er sannsynligheten for sekvensen "sort, sort, rød, hvit, hvit" når 5 brikker trekkes uten tilbakelegning fra en urne inneholdende 6 hvite, 4 sorte og 5 røde brikker?

29 En gambler har tre kort; det 1. er rødt på begge sidene, det 2. er blått på begge sidene, det 3. er rødt på den ene siden og blått på den andre siden. Et kort trekkes tilfeldig og legges på bordet. Anta at kortet som vises er rødt. Gambleren vedder Kr 1000 på at den andre siden av kortet er rødt? Bør du vedde i mot?

30 En bøyd mynt gir krone med dobbel så stor sannsynlighet som den gir mynt. Dersom den viser krone, så trekkes en brikke fra en urne med 3 hvite og 4 røde brikker. Dersom mynten viser mynt, så trekkes en brikke fra en urne med 6 hvite og 3 røde brikker. Hva er sannsynligheten for at mynten viste mynt, når en hvit brikke ble trukket?

31 En sort og en hvit terning kastes. Se på hendelsene $A =$ "odde tall på den sorte terningen", $B =$ "odde tall på den hvite terningen" og $C =$ "summen på terningene er odde". Vis at hver av disse hendelsene er parvis uavhengige. Er hendelsene A, B, C uavhengige?

32 To terninger kastes. Hva er sannsynligheten for at den ene terningen viser dobbelt så mange øyne som den andre?

33 Følgende problem ble stilt til Pascal av de Méré i 1654: Hvor mange ganger må to terninger kastes for at sannsynligheten for å få minst en dobbel 6'er er større enn 1/2?

34 Vis at $(\uplus_i A_i) \cap B = \uplus_i (A_i \cap B)$.

35 La A være en hendelse slik at $P(A) > 0$. Vis at familien av alle hendelser inneholdt i A er en hendelsesfamilie i A. Vis at A blir et sannsynlighetsrom med fordeling Q når vi definerer $Q(B) = P(B)/P(A)$. Hva er forskjellen på Q og den betingede sannsynligheten gitt A?

36 Formuler og bevis Bayes teorem.

37 Vis ved å anta uavhengighet at sannsynligheten for at du må kaste en mynt 10 ganger før du får krone er lik $(1/2)^{10}$. Kan du gi et argument som ikke bygger på uavhengighet?

38 Skriv opp alle betingelser som skal være oppfylt for at hendelsene A_1, A_2, A_3 er uavhengige. Er hendelsene $\emptyset, \emptyset, \emptyset$ uavhengige?

39 Anta at A og B er uavhengige. Er da A og B^c uavhengige?

Kapittel 5

Kombinatorisk sannsynlighet

In the study of simple games of chance, sampling procedures, occupancy and order problems, etc., we are usually dealing with finite sample spaces in which the same probability is attributed to all points. To compute the probability of an event A we have then to divide the number of sample points in A ("favorable cases") by the total number of sample points ("possible cases"). This is facilitated by a systematic use of a few rules which we shall now proceed to review.

W. Feller (1950)

Kombinatorikk er en gammel gren av matematikken. F eks har Det gamle testamentet referanser til kombinatoriske problem. Avhandlingen *Dissertatio de arte combinatoria* (1660) skrevet av Wilhelm Leibniz kan regnes som den første faglige publikasjonen i dette feltet. Hovedproblemet i kombinatorikken er å telle antall mulige kombinasjoner. Vi vil ha nytte av kombinatorikken for f eks å få beregnet antall mulige utfall i en hendelse.

5.1 Multiplikasjonsregelen og permutasjoner

I noen tilfeller er det mulig å illustrere de mulige utfallene med et tre-diagram som i det neste eksemplet.

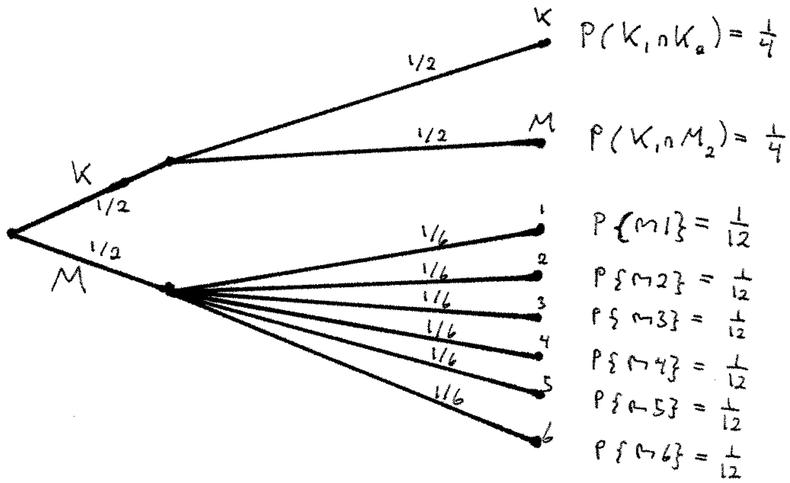
Eksempel 5.1.1 (Mynt + Terning) La et eksperiment være gitt ved at vi først kaster en mynt. Dersom resultatet er krone, så kastes mynten en gang til. Dersom resultatet var mynt, så kastes i stedet en terning. Tre-diagrammet i figur 5.1 illustrerer de mulige utfallene. Sannsynligheten til enkeltutfallene kan beregnes ved produktet av sannsynlighetene langs grenene. Sannsynligheten for f eks utfallet $m1$ er

$$\begin{aligned} P(m1) = & P(m \text{ i første deleksperiment}) \cdot \\ & P(1 \text{ i andre deleksperiment} \mid m \text{ i første deleksperiment}) = \frac{1}{2}\frac{1}{6} = \frac{1}{12}. \end{aligned}$$

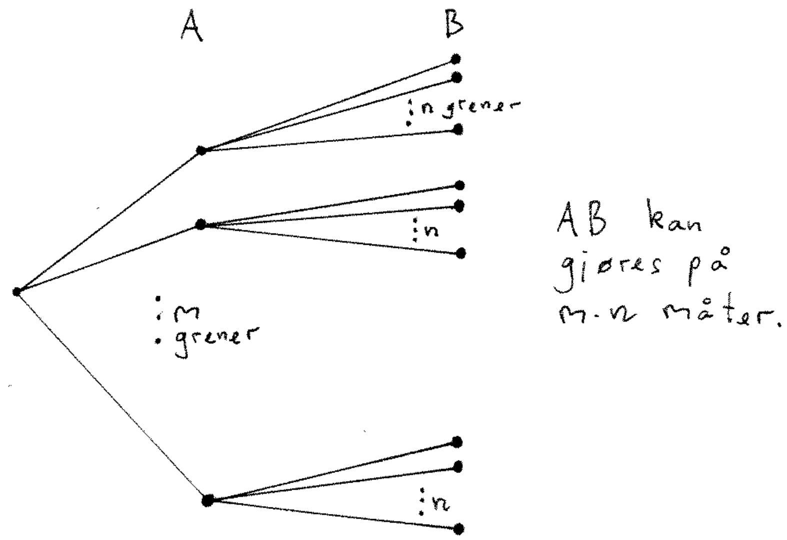
Se på et eksperiment gitt ved at vi først utfører et deleksperiment med m mulige utfall. Deretter utføres et nytt deleksperiment. Som i Eksempel 5.1.1 kan hvilket deleksperiment som utføres her være avhengig av resultatet i det foregående. Problemet er så å avgjøre hvor mange mulige utfall eksperimentet har. Generelt kan dette fort bli en uoversiktlig oppgave. Et tilfelle er imidlertid enkelt: Dersom det andre deleksperimentet alltid har n mulige utfall, så har eksperimentet $m \cdot n$ mulige utfall. Dette er illustrert i figur 5.2 og 5.3. Multiplikasjonsregelen er elementær, men fortjener å utheves fordi den er så viktig.

> Multiplikasjonsregelen: Anta at operasjon A og B kan gjøres på hhv m og n måter. Da kan operasjonen AB (A etterfulgt av B) gjøres på $m \cdot n$ måter.

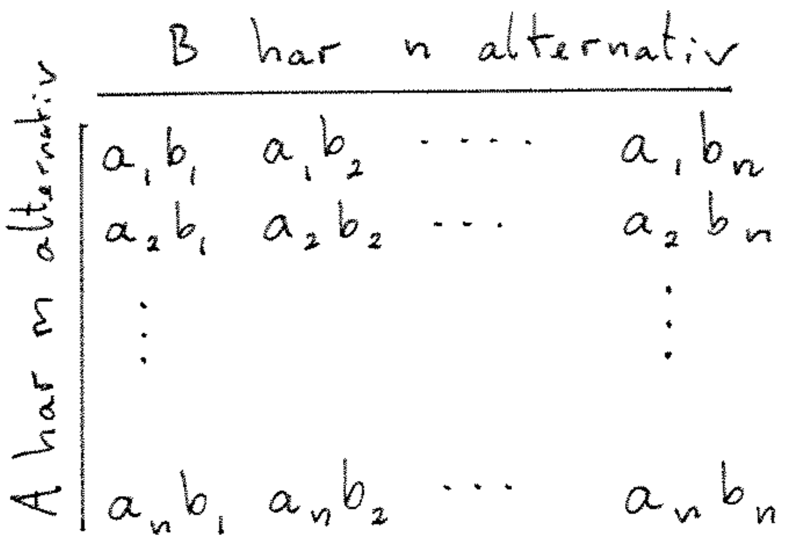
Vi valgte å formulere multiplikasjonsregelen uten referanse til sannsynlighetesteorien fordi kombinatorikken også mer generelt står på egne ben uten referanse til sannsynlighet. Ved induksjon



Figur 5.1: Tre-diagram for et spesielt Mynt + Terning eksperiment.



Figur 5.2: Tre-diagram for multiplikasjonsregelen.



Figur 5.3: Matrise for multiplikasjonsregelen.

kan det vises at operasjonen $A_1 \cdots A_n$ kan gjøres på $m_1 \cdots m_n$ måter dersom hver A_i kan gjøres på m_i måter. Dette er den generelle multiplikasjonsregelen.

Eksempel 5.1.2 (Passord) Anta at et passord er på formen bokstav-bokstav-siffer-siffer. Første operasjon er her å velge en bokstav blant 29 mulige, og de resterende operasjonene er tilsvarende. Multiplikasjonsregelen gir da at antall passord blir $29 \cdot 29 \cdot 10 \cdot 10 = 67600$.

Å permutere betyr å bytte om. F eks er alle permutasjoner av *abc* gitt ved *abc*, *acb*, *bac*, *bca*, *cab*, *cba*. Dette viser at det finnes 6 permutasjoner av tre elementer. Dette kunne vi ha konkludert ved hjelp av multiplikasjonsregelen uten å ha skrevet opp alle permutasjonene. Operasjonen A_1 er å velge a, b eller c, dvs den kan gjøres på tre måter. Operasjonen A_2 er å velge en av de to gjenværende bokstavene, og operasjonen A_3 er å velge den siste bokstaven. Multiplikasjonsregelen gir da at operasjonen $A_1A_2A_3$ kan gjøres på $3 \cdot 2 \cdot 1 = 6$ måter, dvs det finnes 6 permutasjoner av 3 elementer.

Dette kan generaliseres. *Vi definerer* $n!$ *(les: n fakultet) ved* Definisjon: **n!**

$$n! = n \cdot (n-1) \cdots 1$$

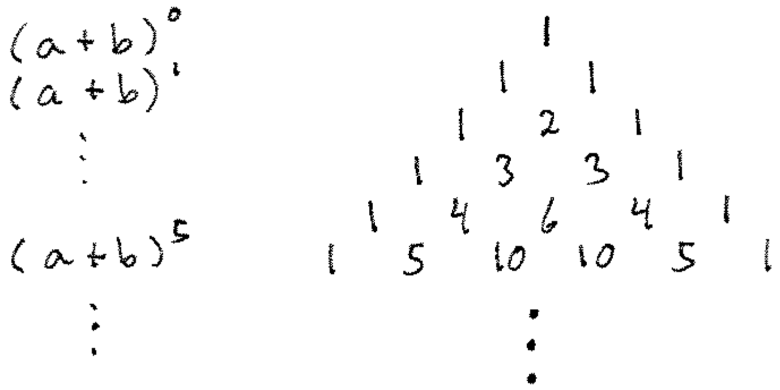
og spesielt $0! = 1$. Multiplikasjonsregelen gir at det finnes $n!$ permutasjoner av n elementer $a_1, a_2, \ldots, a_n$: Operasjon A_1 er å velge første element blant n mulige. Når dette er gjort velges andre element blant de $n-1$ gjenværende element. Generelt kan operasjon A_j gjøres på $n+1-j$ måter hvor $j = 1, \ldots, n$. Spesielt kan da operasjon A_n gjøres på 1 måte fordi det bare er et element igjen å velge. Konklusjonen er da at operasjonen $A_1 \cdots A_n$ kan gjøres på $n \cdot (n-1) \cdots 2 \cdot 1$ måter. Dette betyr at vi kan velge en permutasjon på $n!$ måter.

Det finnes $n!$ mulige permutasjoner av n elementer.

La A_i være som i det foregående. Dersom vi bare gjør operasjon A_1 til og med operasjon A_k, så er resultatet en permutasjon av lengde k fra n elementer. Operasjonen $A_1 \cdots A_k$ kan gjøres på

$$n(n-1) \cdots (n+1-k) = \frac{n(n-1) \cdots (n+1-k) \cdot (n-k) \cdots 2 \cdot 1}{(n-k) \cdots 2 \cdot 1} = \frac{n!}{(n-k)!}$$

måter. Vi har dermed bevist at:



Figur 5.4: Pascals trekant.

> Det finnes $n!/(n-k)!$ permutasjoner av lengde k fra n elementer.

F eks finnes det $4 \cdot 3 = 12$ permutasjoner av størrelse 2 fra mengden $\{1, 2, 3, 4\}$. Dette kunne også vært funnet ved å liste opp alle disse, dvs $12, 13, 14, 21, 23, 24, 31, 32, 34, 41, 42, 43$. Denne siste metoden blir fort uhåndterlig når n og m er store.

5.2 Binomialkoeffisienter og kombinatoriske sannsynligheter

Pascals trekant i figur 5.4 er gitt ved at hvert tall er gitt ved summen av tallet opp til venstre og tallet opp til høyre. Dette gjelder for alle tallene bortsett fra tallet 1 på toppen (linje 0) når alle linjene fylles ut med 0 der det ikke er skrevet noe i Pascals trekant. Ved hjelp av Pascals trekant kan en beregne koeffisientene i $(a+b)^n$. F eks gir linje 5 i Pascals trekant at

$$(a+b)^5 = b^5 + 5ab^4 + 10a^2b^3 + 10a^3b^2 + 5a^4b + a^5.$$

Definisjon: $\binom{n}{k}$

Tallene i Pascals trekant er binomialkoeffisientene definert ved

$$\binom{n}{k} = \frac{n!}{(n-k)!k!}.$$

Pascals trekant kan da skrives som i figur 5.5. Spesielt gir figur 5.4 og figur 5.5 at $\binom{5}{3} = 10$. Systemet som gir hvordan en linje i Pascals trekant kan regnes ut fra linjen over er gitt ved den kombinatoriske identiteten

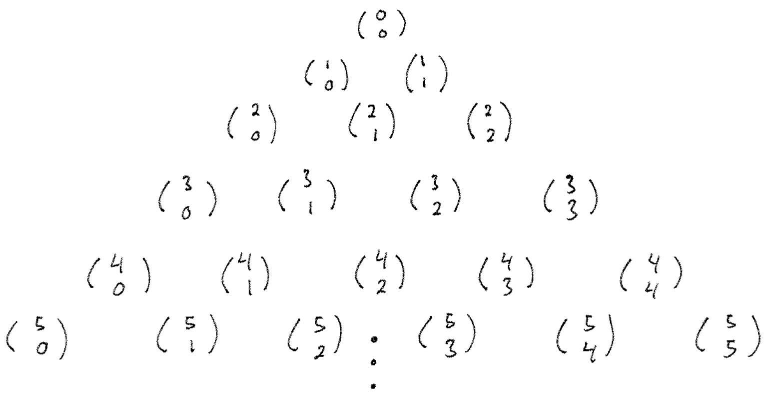
$$\binom{n+1}{k+1} = \binom{n}{k} + \binom{n}{k+1}.$$

Identiteten bevises ved

$$\binom{n}{k} + \binom{n}{k+1} = \frac{n!}{k!(n-k)!} + \frac{n!}{(k+1)!(n-(k+1))!}$$

$$= \frac{n!(k+1)}{k!(k+1)(n-k)!} + \frac{n!(n-k)}{(k+1)!(n-(k+1))!(n-k)}$$

$$= \frac{n!(k+1+n-k)}{(k+1)!(n-k)!} = \frac{(n+1)!}{(k+1)!(n-k)!} = \binom{n+1}{k+1}.$$



Figur 5.5: Binomialkoeffisientene i Pascals trekant.

Denne kombinatoriske identiteten kan benyttes for å gi et induksjonsbevis av binomialteoremet:

$$(a+b)^n = \sum_{k=0}^{n} \binom{n}{k} a^k b^{n-k}.$$

Eksempel 5.2.1 (Fikset mynt) La p være sannsynligheten for å få krone i et enkelt myntkast. Vi vil vise at sannsynligheten for å få m kronesider ved n myntkast er

$$\binom{n}{m} p^m (1-p)^{n-m}.$$

La Ω være mengden av alle mulige sekvenser $\omega = \omega_1 \ldots \omega_n$ hvor ω_i er lik krone eller mynt. Produktregelen gir at utfallsrommet Ω har 2^n mulige enkeltutfall. Når vi antar at hvert myntkast er uavhengig av de andre, så følger det at

$$P\{\omega \,|\, \omega_1 = \cdots = \omega_m = \text{krone}, \omega_{m+1} = \cdots = \omega_n = \text{mynt}\} = p^m (1-p)^{n-m},$$

og mer generelt er dette sannsynligheten til hvert enkeltutfall med m kronesider. Vi må vise at det finnes $\binom{n}{m}$ slike enkeltutfall. Dette er det samme som at det skal finnes $\binom{n}{m}$ ikke-ordnede utvalg av størrelse m fra mengden $\{1, \ldots, n\}$ hvor det at j blir trukket tilsvarer at $\omega_j = \text{krone}$. Antall ordnede utvalg av størrelse m er lik antall permutasjoner av lengde m, dvs lik $n!/(n-k)!$. Dermed er antall ikke-ordnede utvalg lik binomialkoeffisienten $n!/[(n-k)!k!]$.

Det foregående eksemplet inneholder to resultater som fortjener å utheves.

Antall ordnede og ikke-ordnede utvalg av størrelse m fra n elementer er gitt ved hhv

$$\frac{n!}{(n-m)!} \text{ og } \binom{n}{m}.$$

Fordelingen som ble utledet i Eksempel 5.2.1 er binomialfordelingen. *Binomialfordelingen er definert ved den diskrete tettheten* Definisjon: **binomialfordelingen**

$$f(k) = \binom{n}{k} p^k (1-p)^{n-k}, \; 0 \le p \le 1, \; k = 0, \ldots, n.$$

5.3 Oppgaver

40 Vis ved induksjon at operasjonen $A_1 \cdots A_n$ kan gjøres på $m_1 \cdots m_n$ måter dersom hver A_i kan gjøres på m_i måter.

41 Hvor mange ledd er det i ekspansjonen av

$$(a+b+c)(d+e+f)(x+y+u+w) \ ?$$

Hvilke av leddene aeu, cdx, bef, xvw finnes i ekspansjonen?

42 Stirlings formel gir at

$$n! \approx \sqrt{2\pi} n^{n+1/2} e^{-n}.$$

Bruk Stirlings formel til å approksimere 30! ved først å beregne en approksimasjon til ln(30!). Sammenlign svaret med den eksakte verdien 265252859812268935315188480000000.

43 Skriv opp alle permutasjoner av elementene i mengden $\{1, 2, 3, 4\}$. Skriv også opp alle permutasjoner av lengde 2. Hvordan stemmer disse to listene med multiplikasjonregelen?

44 Skriv opp alle ordnede utvalg av størrelse 4 og av størrelse 2 fra mengden $\{1, 2, 3, 4\}$. Skriv også opp de tilsvarende ikke-ordnede utvalg. Hvordan stemmer listene med multiplikasjonregelen?

45 Fire menn og fire kvinner skal plasseres i en rad med 8 stoler. Hvor mange plasseringer finnes det? Hvor mange plasseringer finnes det når det skal sitte en mann i annenhver stol?

46 Hvor mange linjestykker kan trekkes mellom 5 punkter A, B, C, D, E i planet?

47 La M være en mengde. En permutasjon p av elementene i M defineres ved at det er en bijektiv funksjon $p : M \to M$. Dette betyr at p er injektiv ($m_1 \neq m_2 \Rightarrow p(m_1) \neq p(m_2)$) og surjektiv ($\{p(m) \mid m \in M\} = M$), eller med andre ord at p har en invers p^{-1} som er definert på hele M. Skriv opp alle permutasjoner p av elementene i $M = \{a, b, c\}$.

48 Prøv å gi definisjoner av ordnede og ikke-ordnede utvalg av størrelse n fra en generell mengde M. Vis at definisjonene er rimelige i tilfellet $n = 2$ og $M = \{1, 2, 3\}$.

49 Bruk den kombinatoriske identiteten

$$\binom{n+1}{k+1} = \binom{n}{k} + \binom{n}{k+1}$$

til å bevise binomialteoremet ved induksjon.

50 Bevis at binomialkoeffisientene er naturlige tall.

51 Vis at $\{1, \ldots, n\}$ har $\binom{n}{m}$ delmengder av størrelse m. Bruk dette til å gi et kombinatorisk bevis av binomialteoremet.

52 Vis at tettheten til binomialfordelingen er en tetthet for utfallsrommet $\mathbb{R}$.

53 Et spill er gitt ved at n personer kaster en mynt. Dersom en spiller får et resultat forskjellig fra alle de andre, så er denne spilleren slått ut av spillet. Hva er sannsynligheten for at en person blir slått ut?

54 A somewhat inebriated conventioneer finds himself in the embarrassing predicament of being unable to predetermine whether his next step will be forward or backward. What is the probability that after hazarding n such maneuvers he will have stumbled forward a distance of r steps?

Hint: La x og y være hhv antall skritt forover og bakover. Da er $x + y = n$ og $x - y = r$.

55 X-ray shows that 6 of Muffy's 32 teeth have cavities. Unfortunately, her dentist is a bit myopic and drills 6 of her teeth at random. What is the probability that fewer than half of the cavities are properly drilled?

56 Følgende oppgave (noe modernisert her) ble stilt til Sir Isac Newton av vennen Samuel Pepys i 1693: Hva er sannsynlighetene for å få minst en 6'er når 6 terninger kastes, for å få minst to 6'ere når 12 terninger kastes, og for å få minst tre 6'ere når 18 terninger kastes?

Kapittel 6

Observatorer og tilfeldige variable

"A random variable is a quantity whose values are determined by chance." What does that mean? The word "quantity" suggests magnitude - numerical magnitude. Ever since rigor has come to be demanded in mathematical definitions, it has been recognized that the word "variable", particularly a variable whose values are "determined" somehow or other, means in precise language a function. Accordingly a random variable is a function: a function whose numerical values are determined by chance. This means, in other words, that a random variable is a function attached to an experiment - once the experiment has been performed the value of the function is known.

P.R. Halmos (1950)

Teorien for tilfeldige variable og vektorer står sentralt i sannsynlighetsteorien. I Kolmogorovs formulering av sannsynlighetsteorien er enhver tilfeldig størrelse assosiert til et eksperiment representert ved en funksjon definert på utfallsrommet til eksperimentet. Spesielt vil et tilfeldig tall identifiseres med en funksjon som tar reelle verdier. I sannsynlighetsteorien kalles en slik funksjon for en tilfeldig variabel. Generelt gir Kolmogorovs formulering en metode for utledning av nye sannsynlighetsrom fra et underliggende sannsynlighetsrom.

6.1 Observatorer og fordelingen til en observator

La Ω være utfallsrommet for et eksperiment. I noen situasjoner er det ikke selve utfallet ω som er interesant, men snarere en egenskap $T(\omega)$ til ufallet. Dersom f eks ω er en tilfeldig utvalgt person fra en befolkning, så er det alderen $T(\omega)$ som er relevant for å finne aldersfordelingen i befolkningen. Vi velger å diskutere et annet eksempel mer i detalj.

Eksempel 6.1.1 (Antall kron i myntkast med n mynter) Et enkelt myntspill er gitt ved at en person velger mynt eller krone for deretter å kaste n mynter. Dersom personen valgte krone, så vinner han alle myntene med kronesiden opp, og tilsvarende for mynt alternativet. La K være antall kron i myntkastet. Utfallsrommet $\Omega_K = \{0, 1, 2, \ldots, n\}$ gir de mulige verdiene til K. Vi vil nå vise hvordan fordelingen P_K til dette eksperimentet er gitt ved symmetriantagelser i et tenkt underliggende eksperiment. Anta at myntene er nummerert. Da er utfallsrommet Ω mengden av de 2^n enkeltutfall som fremkommer ved at mynt 1 kan ta verdien mynt eller krone, mynt 2 kan ta verdien mynt eller krone, osv. Dersom myntene er identiske og rettferdige, så følger det at hvert enkeltutfall i dette underliggende eksperimentet er like sannsynlige. Dette gir da $P\{\omega\} = 2^{-n}$. Dersom resultatet av myntkastet er ω, så er antall kron k en funksjon K av ω, dvs $k = K(\omega)$. Antall ω som gir $k = K(\omega)$ er lik antall ikke-ordnede utvalg av størrelse k fra $\{1, \ldots, n\}$, dvs lik $\binom{n}{k}$. Direkte argument: 1.kroneside kan plasseres på n plasser, 2.kroneside kan plasseres på $n-1$ plasser osv gir $n \cdot (n-1) \cdots (n-k+1)$ muligheter, men rekkefølgen vi plasserer kronesidene i endrer ikke ω, så vi må dividere med $k!$. Dette gir

$$P_K\{k\} = P\{\omega \,|\, K(\omega) = k\} = \sum_{\substack{\omega \\ K(\omega)=k}} P\{\omega\} = \binom{n}{k} 2^{-n}.$$

Dermed har vi funnet fordelingen P_K til K. Konklusjonen er at K er binomialfordelt med parameter $p = 1/2$. Denne utledningen har vi gjort før. Det nye her er at utledningen tar utgangspunkt i funksjonen K.

Definisjon: $(T \in A)$

I det foregående eksemplet så vi at vi kunne finne fordelingen til en egenskap ved utfallet i et eksperiment. Dette kan gjøres helt generelt. *Når T er en funksjon $T : \Omega \to \Omega_T$, og A er en delmengde av Ω_T, så defineres*

$$\boxed{(T \in A) = \{\omega \,|\, T(\omega) \in A\}}.$$

I konvensjonell matematisk notasjon er $(T \in A) = T^{-1}(A)$. Notasjonen $(T \in A)$ har imidlertid den fordelen at den har naturlige generaliseringer til tilfeller med flere betingelser. Et eksempel er gitt ved

$$(T \in A, (U \in B \text{ eller } V \in C))$$

som er definert lik

$$\{\omega \,|\, T(\omega) \in A, (U(\omega) \in B \text{ eller } V(\omega) \in C)\}.$$

I konvensjonell matematisk notasjon er overnevnte mengde gitt ved det noe mindre leselige utrykket

$$(T^{-1}(A)) \cap ((U^{-1}(B)) \cup (V^{-1}(C))).$$

Definisjon: **observator**

La Ω og Ω_T være utfallsrom utstyrt med hver sin hendelsesfamilie. *En observator for Ω med verdier i Ω_T er en funksjon $T : \Omega \to \Omega_T$ slik at $(T \in A)$ er en hendelse i Ω for enhver hendelse A i Ω_T.* Det er vanlig å la være å spesifisere Ω. Formuleringen "La S, T, U være observatorer" skal forstås som at det er gitt et underliggende utfallsrom slik at S, T, U alle er observatorer for dette utfallsrommet. Vi vil konvensjonelt kalle dette utfallsrommet for Ω. Mengdene Ω_S, Ω_T, Ω_U der S, T, U tar sine verdier vil vanligvis spesifiseres. La P være en fordeling for Ω og la T være en observator for Ω med verdier i Ω_T. *Fordelingen P_T til observatoren T er definert ved*

Definisjon: P_T

$$\boxed{P_T(A) = P(T \in A).}$$

Vi beviser at P_T er en fordeling for Ω_T:

1. Fordi T er en observator, så er $(T \in A)$ en hendelse for hver hendelse A. Dermed er $P_T(A) = P(T \in A) \geq 0$ veldefinert for alle hendelser A i Ω_T.

2. Fordelingen P_T er normalisert fordi $P_T(\Omega_T) = P(T \in \Omega_T) = 1$ pga $(T \in \Omega_T) = \Omega$.

3. Identiteten
$$(T \in \uplus_i A_i) = \uplus_i (T \in A_i)$$
gir $P_T(\uplus_i A_i) = P(\uplus_i (T \in A_i)) = \sum_i P_T A_i$. Dette viser at P_T oppfyller addisjonsloven.

Definisjon: **enkel observator**

En observator T er enkel dersom den tar et endelig antall verdier $t_1, \ldots, t_n$. Fordelingen til T er da gitt ved den diskrete tettheten

$$f_T(t) = P_T\{t\} = P(T = t) = P\{\omega \,|\, T(\omega) = t\}, \;\; t = t_1, \ldots, t_n.$$

Dette resultatet har en naturlig generalisering.

> Fordelingen til en observator som tar et tellbart antall verdier er diskret.

6.2 Tilfeldige variable, vektorer og uavhengighet

Tallinjen $\mathbb{R}$ er antagelig det utfallsrommet som benyttes oftest, direkte eller som en byggekloss. Dersom ikke annet er sagt så er det alltid underforstått at tallinjen er utstyrt med den minste hendelsesfamilien som inneholder alle intervall. Familien av hendelser i $\mathbb{R}$ er dermed lik familien av Borel-mengder i $\mathbb{R}$. La det være underforstått at et underliggende utfallsrom er gitt. *En tilfeldig*

Definisjon: **tilfeldig variabel**

variabel er en observator som tar reelle tall som verdier. Dersom Ω er det underliggende utfallsrommet, så betyr dette spesielt at en tilfeldig variabel X er en funksjon $X : \Omega \to \mathbb{R}$. Antagelsene i definisjonen sikrer også at sannsynligheten

$$P(a \leq X \leq b) = P_X[a, b]$$

er veldefinert for alle intervall $[a, b]$. Dette følger fordi $(a \leq X \leq b) = (X \in [a, b])$ er en hendelse i Ω fordi intervallet $[a, b]$ er en hendelse i $\mathbb{R}$. Mer generelt har vi følgende resultat.

Fordelingen til en tilfeldig variabel er en fordeling for tallinjen.

La Q være en fordeling for $\mathbb{R}$. *Fordelingsfunksjonen G til Q er definert ved $G(x) = Q(-\infty, x]$.* Fordelingsfunksjonen har følgende egenskaper:

Definisjon: **fordelingsfunksjon**

1. For $x \in \mathbb{R}$ er $0 \leq G(x) \leq 1$.
2. $\lim_{x \to -\infty} G(x) = 0$ og $\lim_{x \to \infty} G(x) = 1$.
3. G er monotont voksende.
4. G er kontinuerlig fra høyre.

En fordelingsfunksjon for tallinjen er definert til å være en funksjon som har egenskapene over. Det kan bevises at dersom G er en fordelingsfunksjon, så finnes en unik fordeling Q for tallinjen som har G som fordelingsfunksjon. Utgangspunktet for beviset er sammenhengen

$$Q(a, b] = G(b) - G(a).$$

Beviset finnes i mer videregående bøker i sannsynlighetsteori eller generell integrasjonsteori. Konklusjonen er verdt å utheve.

Det er 1-1 korrespondanse mellom mengden av fordelinger og mengden av fordelingsfunksjoner for tallinjen.

Et hovedpoeng med innføringen av fordelingsfunksjoner er at det typisk er lettere å spesifisere en fordeling ved å angi fordelingsfunksjonen fremfor å definere fordelingen direkte. Dette var også poenget med innføringen av tettheter, men fordelen med fordelingsfunksjonene er at disse er definert for enhver fordeling for tallinjen. På figur 6.1 er det skissert tre mulige fordelingsfunksjoner. *Fordelingsfunksjonen F_X og tettheten f_X til en tilfeldig variabel X er ved definisjon lik hvv fordelingsfunksjonen og tettheten til fordelingen P_X til X.* Dette betyr at

Definisjon: F_X, f_x

$$F_X(x) = P(X \leq x).$$

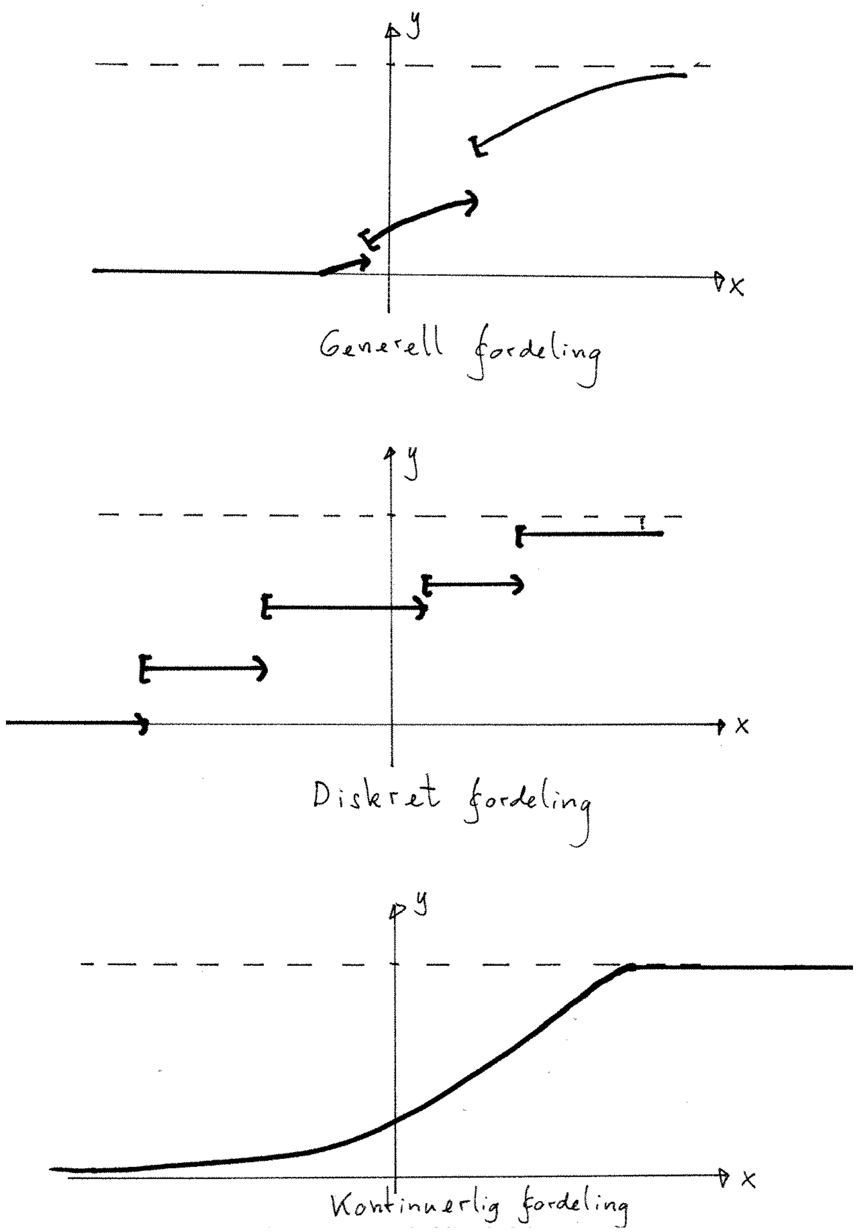
Dersom X har en kontinuerlig fordeling, så gir analysens fundamentalteorem at

$$\boxed{f_X = F_X'.}$$

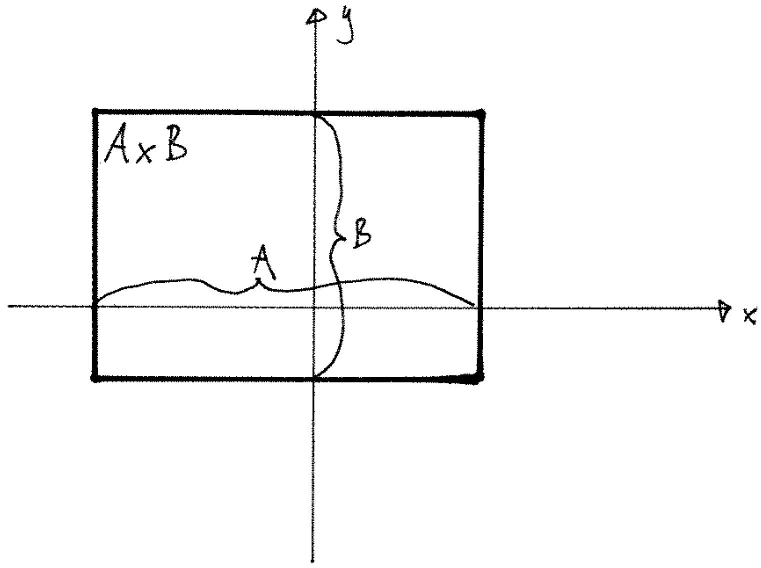
Vi skal se at dette resultatet er meget anvendelig for beregning av tettheter til tilfeldige variable.

La X, Y være tilfeldige variable. Da kan vi definere en observator $T = (X, Y)$ ved $T(\omega) = (X(\omega), Y(\omega))$. For hvert utfall ω er da $T(\omega)$ en vektor i vektorrommet $\mathbb{R}^2$, dvs et punkt i planet. Dette har en naturlig generalisering. *En tilfeldig vektor er en observator som tar verdier i et vektorrom.* Spesielt betyr dette at $T(\omega)$ er en vektor for alle utfall ω i Ω. Hendelsesfamilien i vektorrommet skal være valgt slik at dersom S og T er to tilfeldige vektorer med verdier i samme vektorrom, så definerer lineærkombinasjonen $\alpha S(\omega) + \beta T(\omega)$ en ny tilfeldig vektor $U = \alpha S + \beta T$. Dersom ikke annet er sagt, så er hendelsesfamilien til $\mathbb{R}^2$ gitt som den minste hendelsesfamilien som inneholder alle rektangler.

Definisjon: **tilfeldig vektor**



Figur 6.1: Tre typer fordelingsfunksjoner.



Figur 6.2: Produktet av to intervall er et rektangel.

La $T_1, \ldots, T_n$ være observatorer. Da kan vi definere en observator $T = (T_1, \ldots, T_n)$ ved $T(\omega) = (T_1(\omega), \ldots, T_n(\omega))$. *Simultanfordelingen til observatorene $T_1, \ldots, T_n$ er ved definisjon lik fordelingen til observatoren $T = (T_1, \ldots, T_n)$.* Spesielt er simultanfordelingen til to tilfeldige variable X, Y gitt ved

Definisjon: **simultanfordeling**

$$P_{X,Y}(A \times B) = P(X \in A, Y \in B).$$

Produktet $A \times B$ av to mengder er definert ved

Definisjon: $A \times B$

$$A \times B = \{(a, b) \,|\, a \in A, b \in B\}.$$

Spesielt er $A \times B$ et rektangel dersom A og B er intervall. Dette er illustrert i figur 6.2. *Den simultane fordelingsfunksjonen $F_{X,Y}$ til X, Y er definert ved*

Definisjon: $F_{X,Y}$

$$F_{X,Y}(x, y) = P_{X,Y}((-\infty, x] \times (-\infty, y]) = P(X \leq x, Y \leq y).$$

La h være en funksjon definert på mengden

$$A = \{(x, y) \,|\, a \leq x \leq b,\ f(x) \leq y \leq g(x)\}$$

i planet. En slik mengde kalles y-enkel fordi enhver linje paralell med y-aksen snitter mengden i et intervall. Et eksempel er vist i figur 6.3. Integralet over en x-enkel mengde kan beregnes ved itterert integrasjon

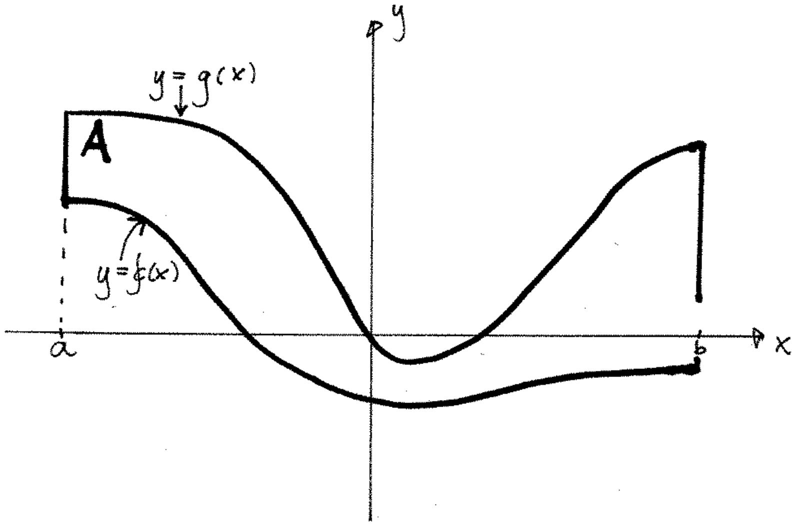
$$\iint_A h(x, y)dxdy = \int_{x=a}^{b} [\int_{y=f(x)}^{g(x)} h(x, y)dy]dx.$$

Noen mengder er både x-enkle og y-enkle og dette gir en verdifull frihet i valg av integrasjonsrekkefølge. Mer generelt kan integralet over mengder i planet beregnes ved å splitte opp mengden i mengder som er enkle.

To tilfeldige variable X, Y er simultant kontinuerlig fordelt dersom (X, Y) er kontinuerlig fordelt, dvs dersom det finnes en $f_{X,Y}$ slik at

Definisjon: **simultant kontinuerlig fordelt**

$$P_{X,Y}(A) = \iint_A f_{X,Y}(x, y)dxdy.$$



Figur 6.3: En y-enkel mengde.

Funksjonen $f_{X,Y}$ er sannsynlighetstettheten til (X, Y), og simultantettheten til X, Y. Det følger at

$$f_{X,Y} = \frac{\partial^2}{\partial x \partial y} F_{X,Y}.$$

Dersom vi kjenner fordelingen til (X, Y) kan vi finne fordelingen til X og til Y. Fordelingen til X er gitt ved

$$P_X(A) = P(X \in A, Y \in \mathbb{R}) = P_{X,Y}(A \times \mathbb{R}).$$

Definisjon: **marginal-fordeling**

Fordelingen til Y finnes tilsvarende. *I denne sammenhengen kalles fordelingen P_X for marginal-fordelingen til X.*

Dersom X, Y er simultant kontinuerlig fordelt, så er X og Y kontinuerlig fordelt. Sannsynlighetstettheten til X, den marginale tettheten til X, er gitt ved

$$f_X(x) = \int_{y=-\infty}^{\infty} f_{X,Y}(x, y) dy$$

Dette bevises ved derivasjon av

$$F_X(x) = P(X \leq x) = P(X \leq x, Y \in \mathbb{R}) = \int_{t=-\infty}^{x} \int_{y=-\infty}^{\infty} f_{X,Y}(t, y) dy dt.$$

Tettheten til Y finnes tilsvarende.

Dersom $X_1, \ldots, X_n$ er tilfeldige variable, så er $T = (X_1, \ldots, X_n)$ en tilfeldig vektor i $\mathbb{R}^n$. Simultanfordelingen til $X_1, \ldots, X_n$ er $P_{X_1,\ldots,X_n} = P_T$, dvs lik fordelingen til $(X_1, \ldots, X_n)$. De foregående definisjonene for X, Y generaliseres tilsvarende.

Definisjon: **uavhengige observatorer**

Observatorene T_1 og T_2 er uavhengige dersom $P(T_1 \in A_1, T_2 \in A_2) = P(T_1 \in A_1) P(T_2 \in A_2)$ gjelder for alle hendelser A_1, A_2. Dette kan generaliseres. *En familie $\{T_j\}$ av observatorer er en familie av uavhengige observatorer dersom produktregelen*

$$P(T_{j_1} \in A_1, \ldots, T_{j_n} \in A_n) = P(T_{j_1} \in A_1) \cdots P(T_{j_n} \in A_n)$$

gjelder for distinkte $j_1, \ldots, j_n$. Følgende konsekvens er spesielt viktig:

> Dersom de tilfeldige variablene $X_1, \ldots, X_n$ er uavhengige med tettheter f_{X_i}, så er
>
> $$f_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = f_{X_1}(x_1) \cdots f_{X_n}(x_n).$$

Vi har tidligere definert uavhengighet for hendelser. Det kan vises at familien $\{A_j\}$ av hendelser er uavhengig hvis og bare hvis familien $\{1_{A_j}\}$ av indikatorvariable er uavhengig.

Vi konkluderer med en definisjon som vil bli brukt mye i det etterfølgende. *Observatorene* $T_1, \ldots, T_n$ *er et tilfeldig utvalg av størrelse n fra fordelingen* Q *dersom de er uavhengige og* $P_{T_1} = \cdots = P_{T_n} = Q$. Dersom Q har en tetthet f, så vil vi si at $T_1, \ldots, T_n$ er et tilfeldig utvalg fra tettheten f. Definisjon: **tilfeldig utvalg**

6.3 Oppgaver

57 Finn fordelingen P_Y til observatoren Y for utfallsrommet $\Omega = \{(1, 2, 0), (2, 1, 3), (4, 1, 1)\}$ når $Y(x, y, z) = x + y + z$. Er Y en tilfeldig variabel?

58 La $T : \Omega \to \Omega_T$. Vis at (i) $(T \in A^c) = (T \in A)^c$ og (ii) $(T \in \cup_i A_i) = \cup_i (T \in A_i)$.

Dette beviser at familien av mengder A slik at $(T \in A)$ er en hendelse er en hendelsesfamilie i Ω_T.

59 Bevis at $(T \in \uplus_i A_i) = \uplus_i (T \in A_i)$.

60 Vis at fordelingen til en observator som tar et tellbart antall verdier er diskret.

61 Hvorfor er $[0, 1] \cup [10, 15]$ en hendelse i utfallsrommet $\mathbb{R}$?

62 Anta at 5 personer, inkludert deg selv og en venn, stiller seg tilfeldig opp på en rekke. La Y være antall personer mellom deg og vennen. Finn tettheten til den tilfeldige variabelen Y.

63 La Y være antall ess en pokerspiller mottar blant de fem første kortene han blir tildelt. Tegn opp fordelingsfunksjonen til Y.

64 Hva er en tilfeldig variabel X? Hva betyr det at F_X er fordelingsfunksjonen til X?

65 Tegn grafen til en fordelingsfunksjon hvor den tilhørende fordelingen hverken er kontinuerlig eller diskret.

66 Prøv å bevise at fordelingsfunksjonen F_X til en tilfeldig variabel har egenskapene

1. For $x \in \mathbb{R}$ er $0 \leq F_X(x) \leq 1$.
2. $\lim_{x \to -\infty} F_X(x) = 0$ og $\lim_{x \to \infty} F_X(x) = 1$.
3. F_X er monotont voksende.
4. F_X er kontinuerlig fra høyre.

67 En tilfeldig variabel Y har en fordelingsfunksjon gitt ved

$$F_Y(y) = \begin{cases} 0 & y < 0 \\ y^2 & 0 \leq y < 1 \\ 1 & y \geq 1 \end{cases}$$

Finn tettheten f_Y til Y. Beregn sannsynligheten $P(\frac{1}{2} < Y \leq \frac{3}{4})$.

68 Hva er en tilfeldig vektor? Vi skal senere definere hva som menes med et konfidensintervall J. Spesielt er et konfidensintervall et tilfeldig intervall. Prøv å gi en definisjon av hva som menes med at J er et tilfeldig intervall.

69 Hva menes med en x-enkel mengde? Tegn en x-enkel mengde i planet. Hva tror du er definisjonen av en z-enkel mengde i rommet?

70 La X være en tilfeldig variabel med kontinuerlig fordeling. Definer $Y = 2X$. Er da X, Y simultant kontinuerlig fordelt?

71 La en diskret tetthet være definert ved

$$f_{X,Y}(x, y) = cxy,\ (x, y) \in \{(1, 1), (2, 1), (2, 2), (3, 1)\}.$$

Finn konstanten c. Finn den marginale tettheten til X.

72 La X og Y være hhv antall konger og antall damer som trekkes når fire kort trekkes fra en kortstokk. Finn tettheten $f_{X,Y}$.

73 Se på eksperimentet gitt ved at en mynt og en terning kastes samtidig. La X være antall kronesider som vises og la Y være antall øyne som vises på terningen. Definer utfallsrommet Ω tilsvarende dette eksperimentet. Hvordan er funksjonene X og Y definert når disse skal være observatorer for Ω? Finn $F_{X,Y}(1, 2)$ og gi en tolkning av dette tallet. Hvordan vil grafen til $F_{X,Y}$ se ut?

74 Finn f_X og f_Y når

$$f_{X,Y} = \frac{1}{x},\ 0 < y < x,\ 0 < x < 1.$$

75 Vis at dersom

$$f_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = f_{X_1}(x_1) \cdots f_{X_n}(x_n),$$

så er $X_1, \ldots, X_n$ uavhengige.

76 Anta at levetiden X (i timer) til en lyspære har en tetthet

$$f_X(x) = \frac{1}{1000} e^{-\frac{x}{1000}}.$$

Finn simultantettheten f_{X_1,X_2,X_3,X_4} tilsvarende levetidene til fire lyspærer innstallert i et kontor. Hva er sannsynligheten for at alle fire fortsatt lyser etter 1050 timer?

77 Hva betyr det at $X_1, \ldots, X_m$ er et tilfeldig utvalg fra binomialfordelingen med parametre $p = 1/2$ og $n = 10$? Finn simultantettheten til $X_1, \ldots, X_m$.

78 Finn konstanten k når

$$f_{X,Y}(x, y) = k,\ 0 \leq x, y \leq 1,\ x + y \leq 1$$

er simultantettheten til X, Y. Gi et geometrisk argument som viser at X og Y ikke er uavhengige.

Kapittel 7

Forventningsverdi

The relative frequency of the repetition is the "measure" of probability, just as the length of a column of mercury is the "measure" of temperature.

R. von Mises (1950)

La $x_1, \ldots, x_n$ være resultatet vi får ved å gjenta et statistisk forsøk n ganger. Selv om resulatet x_i varierer fra gang til gang, så vil noen typer eksperiment være statistisk lovmessige ved at middelverdien

$$\bar{x} = \frac{x_1 + \cdots + x_n}{n}$$

konvergerer mot et tall μ når n vokser. Dette tallet μ er eksperimentets empiriske forventningsverdi.

Anta at eksperimentet over kan modeleres ved et tilfeldig utvalg $X_1, \ldots, X_n$ fra en fordeling P. Det kan da bevises at

$$\bar{X}(\omega) = \frac{X_1(\omega) + \cdots + X_n(\omega)}{n}$$

konvergerer mot forventningsverdien $\mathrm{E}(X)$ når n vokser. Dette konvergensteoremet kalles store talls lov. I det følgende vil forventningsverdien $\mathrm{E}(X)$ defineres og drøftes. Spesielt vil det framgå at $\mathrm{E}(X)$ ikke avhenger av ω. Ved identifiseringen $X_i(\omega) = x_i$ følger det derfor at store talls lov gir en forklaring på den observerte lovmessigheten til $\bar{x}$ ved at $\mu = \mathrm{E}(X)$.

Dersom vi lar $Y_i = 1_A(X_i)$, så gir store talls lov at det relative antall ganger $(= \bar{y})$ hendelsen $(X_i \in A)$ inntreffer konvergerer mot sannsynligheten til denne hendelsen $(= \mathrm{E}(1_A(X)) = P(X \in A))$. Store talls lov kan dermed sees som en forklaring på R. von Mises tolkning av den observerte relative frekvensen som en måling av sannsynligheten til en hendelse.

7.1 Definisjon av forventningsverdi og substitusjonsformelen

La X være en tilfeldig vektor for utfallsrommet Ω utstyrt med fordelingen P. Spesielt kan X være en tilfeldig variabel.

Forventningsverdien $\mathrm{E}\,X$ *til* X *er definert ved integralet* Definisjon: EX

$$\boxed{\mathrm{E}[X] = \int X(\omega) P(d\omega).}$$

Notasjonen $\mathrm{E}(X)$ er mao en kort skrivemåte for integralet av X med hensyn på fordelingen for utfallsrommet. I det følgende vil vi indikere hvordan det generelle integralet over er definert.

Dersom X *tar verdiene* $x_1, \ldots, x_n$, *så er integralet definert ved* Definisjon: $\int X(\omega)P(d\omega)$

$$\int X(\omega) P(d\omega) = \sum_i x_i P(X = x_i).$$

Dermed har vi gitt mening til forventningsverdien til en enkel tilfeldig vektor. Spesielt følger det at

$$\boxed{P(A) = \mathrm{E}(1_A)}$$

Generelt kan integralet defineres ved

$$\int X(\omega)P(d\omega) = \lim \int X_i(\omega)P(d\omega),$$

hvor $X_1, X_2, \ldots$ er en følge enkle tilfeldige vektorer som konvergerer mot X. Fordi hver X_i er enkel er dermed det generelle integralet definert som en grense av endelige summer.

For beregning av forventningsverdier er det spesielt to tilfeller som er viktige. Dersom P er kontinuerlig med en tetthet f, så kan det bevises at

$$\mathrm{E}[X] = \int X(\omega)f(\omega)d\omega.$$

Dersom P er diskret med tetthet f, så kan det bevises at

$$\mathrm{E}[X] = \sum X(\omega)f(\omega).$$

Disse to tilfellene viser at den generelle definisjonen av forventningsverdi generaliserer både summasjon og Riemann-integralet.

Metoden med substitusjon for beregning av Riemann-integralet har en generalisering. La P være en fordeling for utfallsrommet Ω, og la T være en observator for Ω med verdier i utfallsrommet Ω_T. Anta tilslutt at ϕ er en tilfeldig vektor for utfallsrommet Ω_T. Det kan da bevises at følgende formel for variabelskifte holder generelt

$$\boxed{\int \phi(T(\omega))P(d\omega) = \int \phi(t)P_T(dt).}$$

Et bevis av substitusjonformelen for tilfellet at P er diskret er gitt ved

$$\sum_{\omega} \phi(T(\omega))P\{\omega\} = \sum_t \sum_{T(\omega)=t} \phi(T(\omega))P\{\omega\} = \sum_t \phi(t) \sum_{T(\omega)=t} P\{\omega\} = \sum_t \phi(t)P(T = t).$$

Spesielt har vi benyttet at dersom P er diskret, så er P_T diskret med tetthet

$$f_T(t) = P_T\{t\} = P(T = t) = \sum_{T(\omega)=t} P\{\omega\}.$$

Konsekvensen av det generelle tilfellet er meget sentral i sannsynlighetsteorien og verdt å understreke.

> Forventningsverdien til en funksjon av en observator er lik integralet av funksjonen mhp fordelingen til observatoren.

Tilfellet $T = X$ og $\phi(X) = X$ gir spesielt

$$\mathrm{E}\,X = \int x P_X(dx)$$

Formelen for substitusjon betyr at det i mange sammenhenger er unødvendig å spesifisere det underliggende sannsynlighetsrommet Ω. Det er tilstrekkelig å spesifisere fordelingen til T. Derfor er det lett å glemme hva en observator er. Vi repeterer:

En observator T er en funksjon $T : \Omega \to \Omega_T$ slik at $(T \in A)$ er en hendelse for enhver hendelse A.

Fordelingen P_T er kun én av egenskapene til T. Det er f eks umulig å avgjøre om T er uavhengig av en gitt hendelse A i Ω dersom vi bare kjenner fordelingen til T. Et annen demonstrasjon er gitt ved den sedvanlige antagelsen om at $X_1, \ldots, X_n$ er et tilfeldig utvalg. Antagelsen innebærer at alle X_i'ene har lik fordeling, men antagelsen innebærer slett ikke at variablene er like.

Definisjon: E_T

Vi innfører notasjonen

$$\mathrm{E}_T(\phi) = \int \phi(t) P_T(dt).$$

Her er integralet på høyresiden veldefinert fordi P_T er en fordeling. Da kan substitusjonsformelen formuleres uten integraltegn.

Substitusjonsformelen: $\mathrm{E}(\phi(T)) = \mathrm{E}_T(\phi)$.

7.2 Forventningsverdier av summer og produkt

En meget viktig egenskap til forventningsverdier er at de kan tas ledd for ledd, dvs

$$\boxed{\mathrm{E}(aX + bY) = a\,\mathrm{E}\,X + b\,\mathrm{E}\,Y.}$$

Egenskapen er så viktig at den fortjener et eget navn. Vi sier at operatoren E er lineær. Navnet operator forklares ved at E "opererer" på en tilfeldig vektor X, og gir en vektor $\mathrm{E}\,X$. Regneregelen over er som om vi multipliserer med E fra venstre. Dette er en av grunnene til at vi tillater oss å skrive $\mathrm{E}\,X$ fremfor å benytte funksjonsnotasjonen $\mathrm{E}(X)$.

I det overstående er det underforstått at $Z = aX + bY$ er definert ved $Z(\omega) = aX(\omega) + bY(\omega)$. For at $\mathrm{E}(aX + bY)$ skal gi mening, må det derfor bevises at Z er en tilfeldig vektor. Vi utelater beviset her.

Anta at P er diskret. Da kan vi skissere beviset av at E er lineær:

$$\mathrm{E}(aX + bY) = \sum (aX(\omega) + bY(\omega))P\{\omega\} = a \sum X(\omega)P\{\omega\} + b \sum Y(\omega))P\{\omega\} = a\,\mathrm{E}\,X + b\,\mathrm{E}\,Y.$$

Dersom X og Y er uavhengige, så gjelder produktregelen: $\mathrm{E}(XY) = \mathrm{E}(X)\,\mathrm{E}(Y)$.

La $T = (X, Y)$ og $\phi(t) = \phi(x, y) = xy$. Uavhengighet gir $P_{X,Y}(A \times B) = P_X(A)P_Y(B)$ og en skisse av beviset av produktregelen er gitt ved

$$\begin{aligned} \mathrm{E}(XY) = \mathrm{E}\,\phi(T) = \mathrm{E}_T\,\phi = \int \phi(t)P_T(dt) &= \int \phi(x, y)P_{X,Y}(dxdy) \\ = \int y[\int xP_X(dx)]P_Y(dy) &= \int y \ \mathrm{E}\,X \ P_Y(dy) = \mathrm{E}\,X \ \ \mathrm{E}\,Y. \end{aligned}$$

Definisjon: $\mathrm{Var}\,X$

Variansen til en tilfeldig vektor X er definert ved

$$\mathrm{Var}\,X = \mathrm{E}(X - \mathrm{E}\,X)^2.$$

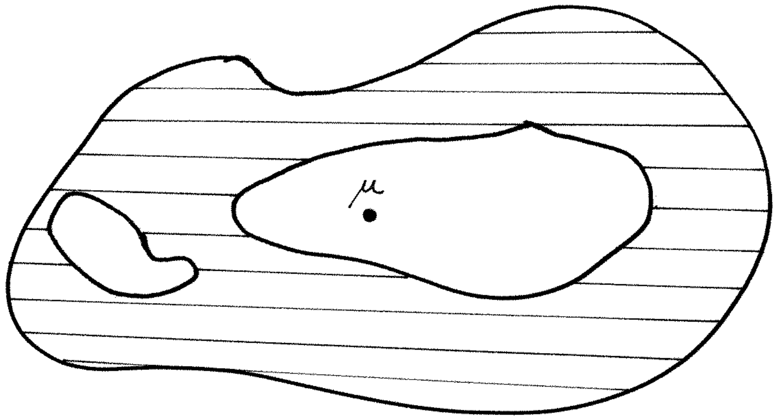
Notasjonskonvensjonene $\mu_X = \mathrm{E}\,X$ og $\sigma_X = \sqrt{\mathrm{Var}\,X}$ er vanlige. Størrelsen σ_X er standardavviket til X.

Alternativt kan variansen beregnes ved formelen

$$\boxed{\mathrm{Var}\,X = \mathrm{E}\,X^2 - \mu_X^2.}$$

Beviset er gitt ved

$$\mathrm{Var}\,X = \mathrm{E}(X - \mu_X)^2 = \mathrm{E}(X^2 - 2X\mu_X + \mu_X^2) = \mathrm{E}\,X^2 - 2\,\mathrm{E}\,X\mu_X + \mu_X^2 = \mathrm{E}\,X^2 - \mu_X^2.$$



Figur 7.1: Tyngdepunkt til en fordeling.

Generelt gjelder regneregelen

$$\boxed{\operatorname{Var} aX = a^2 \operatorname{Var} X.}$$

Dersom X og Y er uavhengige, så gir produktregelen at

$$\boxed{\operatorname{Var}(X+Y) = \operatorname{Var} X + \operatorname{Var} Y.}$$

Vi overlater til leseren å bevise de to sistnevnte regnereglene.

Eksempel 7.2.1 (Middelverdi) La X være enkel med verdier $x_1, \ldots, x_n$ og $P(X = x_i) = 1/n$ (uniform fordeling). Da er forventningsverdien μ_X lik middelverdien $\overline{x}$:

$$\mu_X = \operatorname{E} X = \sum_{i=1}^{n} x_i P(X = x_i) = \frac{x_1 + \cdots + x_n}{n} = \overline{x}$$

Det fremgår av dette at forventningsverdi kan sees som en generalisering av middelverdi. Variansen beregnes tilsvarende

$$\sigma_X^2 = \operatorname{Var} X = \operatorname{E} X^2 - \mu_X^2 = \frac{x_1^2 + \cdots + x_n^2}{n} - \mu_X^2 = \overline{x^2} - \overline{x}^2$$

Dersom definisjonen benyttes direkte finnes

$$\operatorname{Var} X = \operatorname{E}(X - \mu_X)^2 = \frac{(x_1 - \overline{x})^2 + \cdots + (x_n - \overline{x})^2}{n} = \overline{(x - \overline{x})^2}$$

For vilkårlige $x_1, \ldots, x_n$ har vi altså at

$$\overline{(x - \overline{x})^2} = \overline{x^2} - \overline{x}^2.$$

7.3 Geometrisk tolkning av forventningsverdi og varians

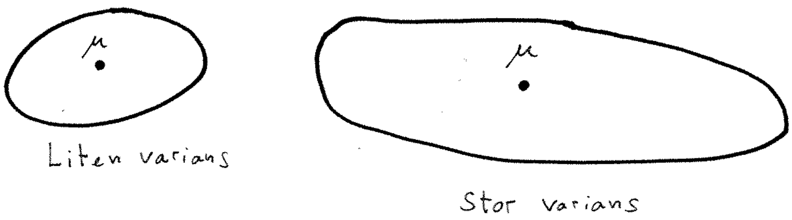
La ρ være en massetetthet i planet. Tyngdepunktet til et legeme beskrevet av denne massetettheten er definert ved

$$x_t = \frac{\int\int (x_1, x_2)\rho(x_1, x_2) dx_1 dx_2}{\int\int \rho(x_1, x_2) dx_1 dx_2}$$

Geometrisk sier tyngdepunktet spesielt noe om hvor legemet er som på figur 7.1 For en sannsynlighetstetthet er totalsannsynligheten (jf totalmassen) lik 1. Analogien med massetetthet gir at forventningsverdien $\operatorname{E} X$ kan sees på som tyngdepunktet til fordelingen P_X.

Tilsvarende kan variansen $\operatorname{Var} X$ sees på som treghetsmomentet til fordelingen P_X om tyngdepunktet. Som på figur 7.2 følger det at variansen $\operatorname{Var} X$ er et mål på spredningen til fordelingen P_X.

Merk: $P(X = \mu_X) = 0$ er mulig, også når X er diskret fordelt. Dette kan sammenlignes med at tyngdepunktet til et legeme godt kan ligge i et hull i legemet som på figur 7.1.



Figur 7.2: Variansen = Treghetsmoment om tyngdepunktet til en fordeling.

7.4 Beregning av forventningsverdi og varians

Følgende eksempler viser at beregning av forventningsverdier kan gjøres på andre måter enn ved direkte integrasjon eller summasjon.

Eksempel 7.4.1 (Forventningsverdi og varians til en Bernoulli variabel) *La Z være Bernoullifordelt med parameter p, dvs $P(Z = 1) = p$ og $P(Z = 0) = (1 - p)$.* Da er

Definisjon: **Bernoulli-fordelt**

$$\mathrm{E}\, Z = \sum_{z=0}^{1} z P(Z = z) = p$$

$$\mathrm{Var}\, Z = \sum_{z=0}^{1} z^2 P(Z = z) - p^2 = p(1 - p)$$

Det første integralet kunne beregnes direkte ved definisjonen fordi Z er en enkel variabel. Det andre integralet ble beregnet ved substitusjonsformelen, dvs variabelen ω ble erstattet av variabelen $z = Z(\omega)$.

Eksempel 7.4.2 (Forventningsverdi og varians beregnet ved variabelskifte) La X være binomialfordelt med parametre (n, p), dvs

$$f_X(x) = \binom{n}{x} p^x (1 - p)^{n-x}, x = 0, \ldots, n.$$

Definer variabelen $Y = Z_1 + \cdots + Z_n$, hvor $Z_1, \ldots, Z_n$ er et tilfeldig utvalg fra en Bernoullifordeling P_Z med parameter p. Det kan da bevises at Y er binomialfordelt med parametre (n, p). Dette gir

$$\mathrm{E}\, X = \mathrm{E}\, Y = \mathrm{E}(Z_1 + \cdots + Z_n) = n\, \mathrm{E}\, Z = np$$

$$\mathrm{Var}\, X = \mathrm{Var}\, Y = \mathrm{Var}(Z_1 + \cdots + Z_n) = n\, \mathrm{Var}\, Z = np(1 - p)$$

Poenget her er igjen substitusjonsformelen som gir at forventningsverdien til $\phi(X)$ er lik forventningsverdien til $\phi(Y)$ når X og Y har lik fordeling.

Eksempel 7.4.3 (Forventningsverdi beregnet ved normaliseringen) La X være binomialfordelt med parametre (n, p). Da er

$$\begin{aligned}\mathrm{E}\, X &= \sum_{x=0}^{n} x \binom{n}{x} p^x (1 - p)^{n-x} = \sum_{x=1}^{n} x \binom{n}{x} p^x (1 - p)^{n-x} = pn \sum_{x=1}^{n} \binom{n-1}{x-1} p^{x-1} (1 - p)^{n-x} \\ &= pn \sum_{k=0}^{n} \binom{n-1}{k} p^k (1 - p)^{n-x} = pn.\end{aligned}$$

Den siste likheten følger fordi sannsynlighestestettheten til en binomialfordelt variabel med parametre $(n - 1, p)$ er normalisert. Dette normaliseringstrikset kan også benyttes ved beregning av enkelte integral.

Den momentgenererende funksjonen M_X og den karakteristiske funksjonen ϕ_X er definert ved

Definisjon: M_X, ϕ_X

$$M_X(t) = \mathrm{E}\, e^{tX} \quad , \quad \phi_X(t) = \mathrm{E}\, e^{itX}, i = \sqrt{-1}.$$

Den karakteristiske funksjonen ϕ_X eksisterer alltid, mens eksistens av den momentgenererende funksjonen M_X avhenger av fordelingen til X. All informasjon om fordelingen til X er inneholdt i hver av disse funksjonene.

Fordelingene til to tilfeldige variable er like hvis og bare hvis de karakteristiske funksjonene er like.

Dette gir en metode for beregning av fordelinger som vi skal se eksempler på senere. Derivasjon gir

$$M'_X(0) = \left(\frac{d}{dt}\,\mathrm{E}(e^{tX})\right)_{t=0} = \big(\mathrm{E}(Xe^{tX})\big)_{t=0} = \mathrm{E}\,X.$$

Ved å derivere n ganger finnes mer generelt

$$\boxed{\mathrm{E}\,X^n = M_X^{(n)}(0) \quad , \quad \phi_X^{(n)}(0) = \mathrm{E}(iX)^n.}$$

Eksempel 7.4.4 (Derivasjon av den momentgenererende funksjon) La X være binomialfordelt med parametre (n,p). Binomialteoremet gir

$$M_X(t) = \mathrm{E}\,e^{tX} = \sum_{x=0}^{n} e^{tx}\binom{n}{x}p^x(1-p)^{n-x} = \sum_{x=0}^{n}\binom{n}{x}(e^t p)^x(1-p)^{n-x} = (e^t p + 1 - p)^n$$

Dermed har vi enda en alternativ beregning av forventningsverdi og varians:

$$\mathrm{E}\,X = M'_X(0) = (n(e^t p + 1 - p)^{n-1}p)_{t=0} = np, \quad \mathrm{Var}\,X = M_X^{(2)}(0) - (np)^2 = \cdots = np(1-p)$$

7.5 Chebyshevs ulikhet

Chebyshevs ulikhet

$$\boxed{P(|X-\mu_X| \geq t\sigma_X) \leq 1/t^2}$$

gjelder generelt for enhver tilfeldig vektor X. Ulikheten gir at sannsynligheten for at utfallet i utfallsrommet Ω_X faller lenger enn t ganger standardavviket unna forventningsverdien er begrenset av $1/t^2$. Dette er dermed en presisering av i hvilken forstand standardavviket måler spredningen av utfallene.

Chebyshevs ulikhet bevises ved følgende beregning:

$$P(|X-\mu_X| \geq t\sigma_X) = \mathrm{E}\,1_{(|X-\mu_X|\geq t\sigma_X)} \leq \mathrm{E}((\frac{X-\mu_X}{t\sigma_X})^2 1_{(|X-\mu_X|\geq t\sigma_X)}) \leq \mathrm{E}((\frac{X-\mu_X}{t\sigma_X})^2 = 1/t^2.$$

Ulikhetene over følger av at E er en monotont voksende operator, dvs

$$\boxed{Y \leq Z \;\Rightarrow\; \mathrm{E}\,Y \leq \mathrm{E}\,Z}.$$

Anta at $X_1,\ldots,X_n$ er et tilfeldig utvalg fra en fordeling P_X. Vi kan da beregne

$$\mathrm{E}\,\overline{X} = \frac{\mathrm{E}\,X_1 + \cdots + \mathrm{E}\,X_n}{n} = \mu_X$$

$$\mathrm{Var}\,\overline{X} = \frac{\mathrm{Var}\,X_1 + \cdots + \mathrm{Var}\,X_n}{n^2} = \frac{\sigma_X^2}{n}$$

Spesielt følger det at

$$\mathrm{Var}\,\overline{X} = \mathrm{E}(\overline{X}-\mu_X)^2 \to 0$$

når n vokser, dvs at middelverdien $\overline{X}$ i en viss forstand konvergerer mot forventningsverdien $\mathrm{E}\,X$. Denne typen konvergens kalles L^2 konvergens, og vi har bevist at

$$\boxed{\lim_{n\to\infty}\frac{X_1+\cdots+X_n}{n} = \mathrm{E}\,X \quad (L^2- \text{ konvergens}).}$$

Dermed har vi bevist en svak versjon av store talls lov.

Innsetting av $t = \epsilon/\sigma$ gir en reformulering:

$$P(|X-\mu_X| \geq \epsilon) \leq \frac{\sigma_X^2}{\epsilon^2}.$$

Reformuleringen av Chebyshevs ulikhet gir at vi også har en annen type konvergens

$$P(\left|\overline{X}-\mu_X\right| > \epsilon) \to 0.$$

Denne typen konvergens kalles konvergens i sannsynlighet. Konklusjonen er dermed at Chebyshevs ulikhet gir en svak versjon av store talls lov.

7.6 Oppgaver

79 Vis at $P(A) = \mathrm{E}\,1_A$ ved å observere at 1_A er en enkel tilfeldig variabel.

80 En urne inneholder 5 røde og 4 hvite brikker. La X være antall røde brikker vi får når vi trekker 3 brikker med tilbakelegning. Finn forventningsverdien $\mathrm{E}\,X$ og variansen $\mathrm{Var}\,X$. Gi en tolkning av $\mathrm{E}\,X$. På hvilken måte er substitusjonsformelen benyttet i beregningene over?

81 La T være en enkel observator. Bevis substitusjonsformelen $\mathrm{E}\,\phi(T) = \mathrm{E}_T\,\phi$ ved å bruke definisjonen av integralet av en enkel tilfeldig vektor.

82 Anta at X er uniformt fordelt på intervallet $[a, b]$. Beregn $\mathrm{E}\,X$.

83 Vis at E er lineær ved å anta at P er kontinuerlig.

84 Ti terninger kastes. Finn forventningsverdien til summen av antall øyne.

85 La $X = 3X_1 + 4X_2 + 2X_3$ hvor

$$f_{X_1}(x) = \binom{6}{x}(1/3)^x(2/3)^{6-x} \quad x = 0, 1, \ldots, 6$$

$$f_{X_2}(x) = \frac{e^{-4}4^x}{x!} \quad x = 0, 1, 2, \ldots$$

$$f_{X_3}(x) = 5e^{-5x} \quad x > 0.$$

Finn $\mathrm{E}\,X$.

86 La X være antall kort som må deles ut før det kommer et ess. Finn $\mathrm{E}\,X$ og gi en tolkning av resultatet.

87 Anta at $f_{X,Y}(x, y) = e^{-x-y}$, $x, y > 0$. Finn $\mathrm{E}(X + Y)$.

88 Anta at $X_1, \ldots, X_n$ er uavhengige og kontinuerlig fordelte variable. Vis at $\mathrm{E}(X_1 \cdots X_n) = \mathrm{E}(X_1) \cdots \mathrm{E}(X_n)$. Vis at $\phi(X_1, X_2)$ og $\psi(X_3)$ er uavhengige tilfeldige variable.

89 La $f_X(x) = 2/x^3$, $x > 1$. Vis at X ikke har endelig varians.

90 La μ være en vektor. Definer en tilfeldig vektor X ved $X(\omega) = \mu$. Vis at X er uavhengig av enhver observator.

91 Anta at X og Y er uavhengige. Bevis regneregelen $\mathrm{Var}(aX + bY) = a^2\,\mathrm{Var}\,X + b^2\,\mathrm{Var}\,Y$.

92 La $\mathrm{E}\,X = \mu$ og la $\mathrm{Var}\,X = \sigma^2$. Vis at

$$\mathrm{E}\left(\frac{X-\mu}{\sigma}\right) = 0 \;\text{ og }\; \mathrm{Var}\left(\frac{X-\mu}{\sigma}\right) = 1.$$

93 La X være en enkel tilfeldig vektor med verdier $x_1, \ldots, x_n$ i rommet $\mathbb{R}^3$. Gi en tolkning av $\mathrm{E}\,X$ og $\mathrm{Var}\,X$ når fordelingen P_X tolkes som massefordelingen til et system av partikler med posisjoner $x_1, \ldots, x_n$.

94 La X_1, X_2 være uavhengige binomiske variable med parametre hhv (n_1, p_1) og (n_2, p_2). Finn $\mathrm{E}(4X + 6Y)$ og $\mathrm{Var}(4X + 6Y)$.

95 La X være uniformt fordelt på intervallet $[0, 2]$. Finn den momentgenererende funksjonen M_X. Beregn $\operatorname{E} X^r$ direkte og ved bruk av M_X for $r = 0, 1, \ldots$. Benytt resultatet sammen med binomialteoremet og regn ut $\operatorname{E}(X - \operatorname{E} X)^6$.

96 Anta at X er Poisson-fordelt, dvs

$$f_X(x) = \frac{e^{-\lambda}\lambda^x}{x!} \quad x = 0, 1, 2, \ldots$$

Finn M_X og benytt denne til å beregne $\operatorname{E} X$ og $\operatorname{Var} X$.

97 Anta at $f_X(x) = e^{-x}, x > 0$. Bruk Chebyshevs ulikhet til å finne en øvre skranke for sannsynligheten for at X tar en verdi mer enn 2 standardavvik unna forventningsverdien. Beregn denne sannsynligheten og kommenter resultatet.

Kapittel 8

Beregning av tettheter og betingede tettheter

This task would have been a rather hopeless one before the introduction of Lebesgue's theories of measure and integration. However, after Lebesgue's publication of his investigations, the analogies between measure of a set and probability of an event, and between integral of a function and mathematical expectation of a random variable, became apparent.

A.N. Kolmogorov (1933)

Beregningen av fordelingen til en observator for et eksperiment er et hovedproblem i sannsynlighetsteorien. I enkle tilfeller kan dette gjøres direkte ved definisjonen, f eks ved kombinatoriske argument. Generelt er dette imidlertid et vanskelig problem, og vi skal se på flere metoder.

Teorien for betingede sannsynligheter kan utvides til en teori for betingede forventningsverdier. Vi skal se hvordan dette kan gjøres i noen tilfeller ved hjelp av betingede tettheter.

8.1 Funksjoner av observatorer

En observator S er basert på en observator T dersom S er en funksjon av T, dvs dersom det finnes en funksjon ϕ slik at Definisjon: **basert på**

$$S(\omega) = \phi(T(\omega)).$$

Vi benytter notasjonen $S = \phi(T)$. Fordelingen til S er gitt av fordelingen til T ved formelen

$$\boxed{P_S A = P_T(\phi \in A)}$$

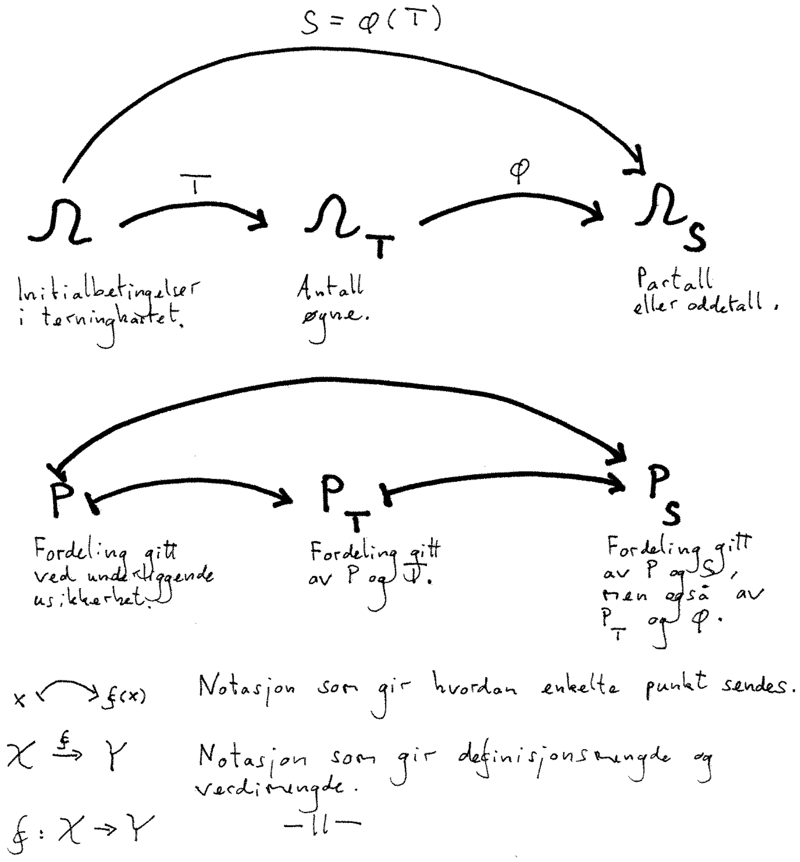
Likheten over kan begrunnes ved

$$\begin{aligned}(S \in A) &= \{\omega \,|\, \phi(T(\omega)) \in A\} = \{\omega \,|\, T(\omega) \in \{t \,|\, \phi(t) \in A\}\} = \{\omega \,|\, T(\omega) \in (\phi \in A)\} \\ &= (T \in (\phi \in A))\end{aligned}$$

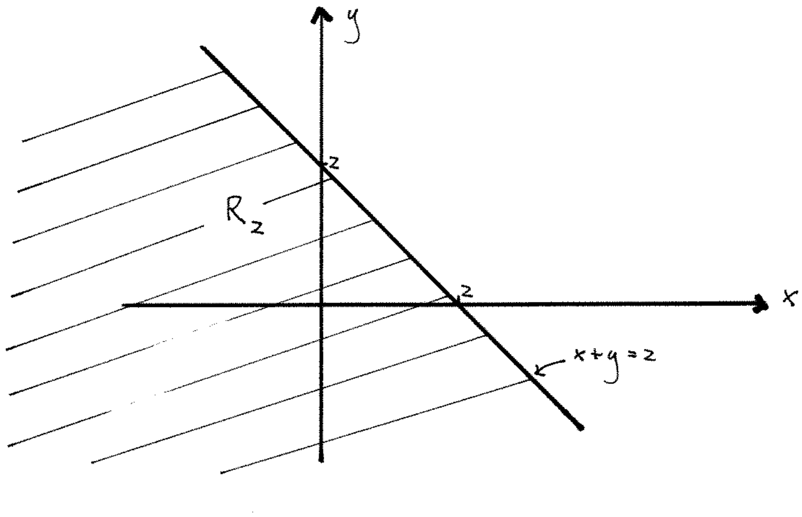
Konklusjonen er:

> Fordelingen til en funksjon av en observator er gitt av funksjonen og fordelingen til observatoren.

Poenget i det overstående er at selv om P_S er definert ved hjelp av P og S, så er P_S entydig gitt av P_T og ϕ. Vi trenger derfor ikke kjenne P for å beregne P_S! Resulatetet er illustrert i figur 8.1 sammen med en antydet modell for terningkast. Resultatet tilsvarer at forventningsverdien $\mathrm{E}\,\phi(T) = \mathrm{E}_T\,\phi$ kan beregnes når E_T er kjent uten at E er kjent. Faktisk er resultatet for fordelingene et spesialtilfelle av substitusjonsformelen for forventningsverdier.



Figur 8.1: Funksjon av en observator.



Figur 8.2: Integrasjonsområdet ved konvolusjon.

8.2 Beregning av tettheter

Anta at $S = \phi(T)$ er kontinuerlig fordelt. Da er tettheten til S gitt ved

$$\boxed{f_S(s) = F'_S(s) = \frac{d}{ds} P_T(\phi \leq s).}$$

Dette er metoden med fordelingsfunksjonen. Metoden med fordelingsfunksjonen er ofte den mest anvendelige metoden for å finne en tetthet.

Eksempel 8.2.1 (Tettheten til $aX + b$) Vi vil finne tettheten til $Y = aX + b$ når vi antar at X er kontinuerlig fordelt. Vi antar $a > 0$. Tettheten finnes ved

$$f_Y(y) = \frac{d}{dy} P(aX + b \leq y) = \frac{d}{dy} \int_{\{x \mid ax+b \leq y\}} f_X(x)dx = \frac{d}{dy} \int_{-\infty}^{\frac{y-b}{a}} f_X(x)dx$$

$$= \frac{d}{dy} F_X(\frac{y-b}{a}) = f_X(\frac{y-b}{a}) \frac{1}{a}.$$

Tilfellet $a < 0$ kan behandles tilsvarende. Tilsammen har vi resultatet

Dersom $Y = aX + b$, så er

$$f_Y(y) = f_X(\frac{y-b}{a}) \frac{1}{|a|}.$$

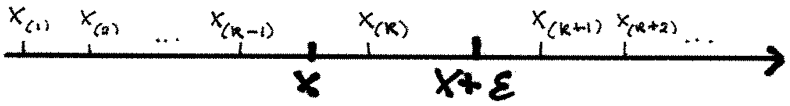
Spesielt gir beviset også at Y er kontinuerlig fordelt.

Eksempel 8.2.2 (Tettheten til $X + Y$ og konvolusjon.) Vi vil finne tettheten til $Z = X + Y$ når vi antar at X, Y er simultant kontinuerlig fordelt. Vi beregner fordelingsfunksjonen

$$F_Z(z) = P(X + Y \leq z) = \iint_{R_z} f_{X,Y}(x, y)dxdy \overset{figur\ 8.2}{=} \int_{x=-\infty}^{\infty} \left(\int_{y=-\infty}^{z-x} f_{X,Y}(x, y)dy \right) dx.$$

Merk: Det kan være til stor hjelp å lage en skisse av integrasjonsområdet ved beregningen av dobbelintegral. Ved derivasjon finner vi

$$f_Z(z) = F'_Z(z) = \frac{d}{dz} \int_{x=-\infty}^{\infty} \left(\int_{y=-\infty}^{z-x} f_{X,Y}(x, y)dy \right) dx = \int_{x=-\infty}^{\infty} f_{X,Y}(x, z - x)dx.$$



Figur 8.3: Et trinomisk utfall.

Dersom vi antar uavhengighet, så finner vi

$$f_Z(z) = \int_{x=-\infty}^{\infty} f_X(x) f_Y(z-x)dx \overset{def}{=} (f_X \star f_Y)(z).$$

Definisjon: $f_X \star f_Y$

Konvolusjonsproduktet $f_X \star f_Y$ er definert av integralet over. Vi kan da formulere et viktig resultat

> Dersom X, Y er uavhengige og kontinuerlig fordelt, så er $X+Y$ kontinuerlig fordelt og tettheten er gitt ved konvolusjonen
>
> $$f_{X+Y} = f_X \star f_Y.$$

Eksempel 8.2.3 (Ordningsobservatoren) Ordningsfunksjonen r er definert ved at $r(x_1, \ldots, x_n) = (t_1, \ldots, t_n)$ hvor $t_1, \ldots, t_n$ er $x_1, \ldots, x_n$ skrevet i stigende rekkefølge. Et konkret eksempel: Dersom $x = (1, 2, 1, 3, -1)$, så er $t = r(x) = (-1, 1, 1, 2, 3)$.

Definisjon: X_{min}, $X_{(j)}$, X_{maks}

La $X_1, \ldots, X_n$ være tilfeldige variable. *Ordningsobservatoren $T = (X_{(1)}, \ldots, X_{(n)})$ er definert ved $T = r(X_1, \ldots, X_n)$ hvor r er ordningsfunksjonen.* Spesielt er da $X_{(n)} = X_{\text{maks}} = \text{maks}\{X_1, \ldots, X_n\}$ og $X_{(1)} = X_{\text{min}} = \text{min}\{X_1, \ldots, X_n\}$.

Vi vil finne marginaltetthetene til alle $X_{(i)}$, $i = 1, \ldots, n$, når vi antar at $X_1, \ldots, X_n$ er et tilfeldig utvalg fra en kontinuerlig tetthet f. I utgangspunktet er den søkte tettheten gitt ved

$$f_{X_{(i)}}(x) = \lim_{\epsilon \to 0} \frac{P(x < X_{(i)} \leq x + \epsilon)}{\epsilon}.$$

Vi vil beregne denne grenseverdien ved å erstatte telleren med en sannsynlighet som kan beregnes kombinatorisk.

Definer variabelen

$$Y_i = \begin{cases} a & X_i \leq x, \\ b & x < X_i \leq x + \epsilon, \\ c & x + \epsilon < X_i. \end{cases}$$

Spesielt er variabelen Y_i basert på X_i. La $p_a = P(Y_i = a)$, $p_b = P(Y_i = b)$ og $p_c = P(Y_i = c)$. Videre defineres variabelen

$$N_a = \text{ antall } Y_i\text{'er som er lik } a = \#\{Y_i \mid Y_i = a\},$$

og N_b, N_c defineres tilsvarende. Uavhengigheten til Y_i'ene og et kombinatorisk argument gir

$$P(N_a = n_a, N_b = n_b, N_c = n_c) = {p_a}^{n_a} {p_b}^{n_b} {p_c}^{n_c} \frac{n!}{n_a! n_b! n_c!}.$$

Definisjon: **trinomisk fordeling**

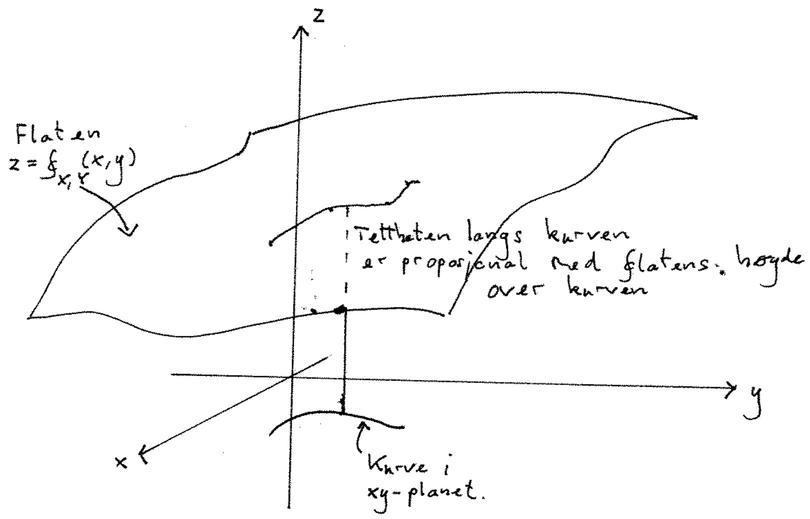
Dette er en generalisering av det binomiske tilfellet. *Vi sier at (N_a, N_b, N_c) er trinomisk fordelt. Vi skriver dette kort som $(N_a, N_b, N_c) \sim$ trinomisk(n, p_a, p_b, p_c).*

I grensen $\epsilon \to 0$ er hendelsen $(x < X_{(k)} \leq x + \epsilon)$ lik hendelsen $(N_a = k-1, N_b = 1, N_c = n-k)$. Dette er illustrert på figur 8.3. Sannsynligheten for denne hendelsen er

$$P(N_a = k-1, N_b = 1, N_c = n-k) = {p_a}^{k-1} {p_b}^{1} {p_c}^{n-k} \frac{n!}{(k-1)!1!(n-k)!} =$$

$$P(X_i \leq x)^{k-1} P(x < X_i \leq x + \epsilon)^1 P(x + \epsilon < X_i)^{n-k} \frac{n!}{(k-1)!(n-k)!} =$$

$$F(x)^{k-1} (F(x+\epsilon) - F(x))(1 - F(x+\epsilon))^{n-k} \frac{n!}{(k-1)!(n-k)!}.$$



Figur 8.4: Geometrisk tolkning av betinget tetthet.

L'Hopitals regel gir tilslutt

$$
\begin{aligned}
f_{X_{(k)}}(x) &= \lim_{\epsilon \to 0} \frac{d}{d\epsilon} P(x < X_{(k)} \le x + \epsilon) \\
&= \lim_{\epsilon \to 0} \frac{d}{d\epsilon} F(x)^{k-1}(F(x+\epsilon) - F(x))(1 - F(x+\epsilon))^{n-k} \frac{n!}{(k-1)!(n-k)!} \\
&= \frac{n!}{(k-1)!(n-k)!} f(x) F(x)^{k-1} (1 - F(x))^{n-k}.
\end{aligned}
$$

La $(X_{(1)}, \ldots, X_{(n)})$ være ordningsobservatoren til et tilfeldig utvalg $X_1, \ldots, X_n$ fra en kontinuerlig fordeling med tetthet f. Da er

$$f_{X_{(k)}}(x) = \frac{n!}{(k-1)!(n-k)!} f(x) F(x)^{k-1} (1 - F(x))^{n-k}.$$

8.3 Betingede tettheter

Anta at X, Y har en simultan tetthet $f_{X,Y}$. Geometrisk kan da $z = f_{X,Y}(x, y)$ tolkes som en flate i rommet, og tettheten i punktet (x, y) er høyden til denne flaten som illustrert på figur 8.4. Anta så at vi kun er interessert i utfall (x, y) som ligger på en kurve. Intuitivt bør tettheten tilsvarende disse utfallene være proposjonal med høyden til flaten $z = f_{X,Y}(x, y)$ over punktene (x, y) som ligger på kurven. Spesielt kan kurven være en horisontal linje gitt ved at y er konstant.

Det foregående motiverer en definisjon. *Den betingede tettheten $f_{X \mid Y}$ til X gitt Y er definert ved* Definisjon: $f_{X \mid Y}$

$$f_{X \mid Y}(x \mid y) = \frac{f_{X,Y}(x, y)}{f_Y(y)}.$$

Det følger at tettheten til $X \mid Y = y$ (les: X gitt $Y = y$) er en sannsynlighetstetthet for alle y. Beviset for det kontinuerlige tilfellet er gitt ved:

1. Tettheten er større eller lik 0.

$$f_{X \mid Y}(x \mid y) = \frac{f_{X,Y}(x, y)}{f_Y(y)} \ge 0.$$

2. Tettheten er normalisert.

$$\int f_{X\,|\,Y}(x\,|\,y)dx = \int \frac{f_{X,Y}(x,y)}{f_Y(y)}dx = \frac{\int f_{X,Y}(x,y)dx}{f_Y(y)} = \frac{f_Y(y)}{f_Y(y)} = 1.$$

Vi har tidligere sett at produktregelen $P(A \cap B) = P(A\,|\,B)P(B)$ er nyttig. Den følgende produktregelen for tettheter er ikke mindre viktig.

Produktregelen: $f_{X,Y}(x,y) = f_{X\,|\,Y}(x\,|\,y)f_Y(y)$.

Notasjonen av indeksene kan i noen tilfeller bli for tungvint. Det er derfor konvensjonelt å utelate indeksene. Produktregelen kan f eks skrives på formen $w(x,y) = w(x\,|\,y)w(y)$, hvor det underforstås at argumentene gir hva w står for. Her er mao $w(x,y) = f_{X,Y}(x,y)$, $w(x\,|\,y) = f_{X,Y}(x\,|\,y)$ og $w(y) = f_Y(y)$.

Eksempel 8.3.1 (Betinget tetthet) La N være antall biler innlevert på et verksted i løpet av en dag. Vi antar at

$$f_N(n) = P(N = n) = \frac{\lambda^n}{n!}e^{-\lambda}, \quad \lambda \geq 0, \quad n = 0, 1, \ldots .$$

Dette er tettheten til Poisson fordelingen. Vi vil finne fordelingen til antall oljeskift X i løpet av en dag.

Nåre det er gitt at det er levert inn n biler, så er sannsynligheten for x oljeskift gitt ved

$$f_{X\,|\,N}(x\,|\,n) = P(X = x\,|\,N = n) = \binom{n}{x}p^x(1-p)^{n-x},$$

når p er sannsynligheten for at det skal utføres oljeskift på en bil som er innlevert. Formelen gir også spesialtilfellet $P(X = 0\,|\,N = 0) = 1$.

Dette gir simultanfordelingen til X, N

$$f_{X,N}(x,n) = f_{X\,|\,N}(x\,|\,n)f_N(n) = \binom{n}{x}p^x(1-p)^{n-x}\frac{\lambda^n}{n!}e^{-\lambda}.$$

Dermed kan vi regne ut tettheten til X

$$f_X(x) = \sum_n f_{X,N}(x,n) = \sum_n \binom{n}{x}p^x(1-p)^{n-x}\frac{\lambda^n}{n!}e^{-\lambda} = \frac{(\lambda p)^x}{x!}e^{-\lambda p}$$

Konklusjonen er dermed at X også er Poisson fordelt. Modellen gir at forventningsverdien til antall biler N i løpet av en dag er λ, mens forventningsverdien til antall oljeskift X i løpet av en dag er λp.

Helt generelt er det mulig å definere en betinget forventning $\mathrm{E}(X\,|\,Y = y)$. Vi nøyer oss med tilfellet hvor det finnes en betinget tetthet. Det kan da bevises at

$$\boxed{\mathrm{E}(\phi(X)\,|\,Y = y) = \int \phi(x)f_{X\,|\,Y}(x\,|\,y)dx.}$$

8.4 Oppgaver

98 En urne inneholder 4 røde og 3 hvite brikker. Identifiser utfallsrommet Ω og fordelingen P for eksperimentet gitt ved at to brikker trekkes fra urnen. La X være antall røde brikker som blir trukket. Finn fordelingen til X og $Y = 2X - 1$. Hvordan er funksjonen $X : \Omega \to \mathbb{R}$ definert?

99 Temperaturen T i en viss kjemisk reaksjon varierer fra eksperiment til eksperiment, men synes å være godt beskrevet av tettheten

$$f_T(t) = te^{-t^2/2} \quad t > 0,$$

når temperaturen er målt i grader Fahrenheit. Temperaturen kan regnes om til grader Celsius ved formelen

$$Y = \frac{5}{9}(T - 32).$$

Finn tettheten til temperaturen målt i grader Celsius.

100 La (X, Y) være uniformt fordelt på enhetskvadratet, dvs $f_{X,Y}(x, y) = 1, \;\; 0 < x, y < 1$. Finn tettheten til $X+Y$, XY og X/Y. Hint: Tegn en figur for integrasjonsområdet i beregningen av f eks $P(X + Y \leq z)$.

101 Anta at X er uavhengig av Y, Z. Vis at X er uavhengig av $Y + Z$.

102 La $X_1, \ldots, X_4$ være 4 uavhengige målinger tilsvarende at et eksperiment med tetthet

$$f(x) = \frac{1}{3} e^{-x/3} \quad x > 0$$

gjøres 4 ganger. Finn sannsynlighetene $P(X_4 \leq 5,1)$ og $P(X_{\text{maks}} \leq 5,1)$.

103 Anta at tiden du bruker på å stå i kø når du går i banken er uniformt fordelt mellom 0 og 10 minutter. Hvis du skal i banken 4 ganger i løpet av den neste måneden, hva er så sannsynligheten for at din nest lengste ventetid er mindre enn 5 minutter?

104 Anta at $f_{Y \,|\, X}(y \,|\, x) = \frac{2y+4x}{1+4x}$ og $f_X(x) = \frac{1}{3}(1 + 4x)$ hvor $0 < x, y < 1$. Finn marginaltettheten f_Y til Y.

105 La $f_{X,Y}(x, y) = 2$ for $0 < x < y < 1$. Finn den betingede sannsynligheten $P(0 < X < \frac{1}{2} \,|\, Y = \frac{3}{4}) = \mathrm{E}(1_{(0,\frac{1}{2})}(X) \,|\, Y = \frac{3}{4})$.

106 Bevis formelen $P_{\phi(T)} A = P_T(\phi \in A)$ ved hjelp av substitusjonsformelen for beregning av forventningsverdier.

107 La f være tettheten til X. Illustrer den geometriske sammenhengen mellom grafen til f og grafen til tettheten g til $X + a$.

108 Konvolusjonsproduktet mellom to funksjoner er definert ved integrasjon slik som i teksten. Hvordan vil du begrunne at $f \star g = g \star f$?

109 Benytt $(X_{\text{maks}} \leq x) = (X_1 \leq x, \ldots, X_n \leq x)$ til å beregne tettheten til X_{maks} når $X_1, \ldots, X_n$ er et tilfeldig utvalg fra en tetthet f. Sammenlign med resultatet i teksten.

110 Produktregelen for tettheter har en generalisering til tilfellet med n tilfeldige variable $X_1, \ldots, X_n$. Prøv å formulere og bevise en slik generalisering. Hvordan stemmer formelen når variablene er uavhengige?

111 Prøv å formulere og bevise en generalisering av Bayes teorem ved betingede tettheter.

112 Dersom Y har en kontinuerlig fordeling, så kan det vises generelt at

$$\mathrm{E}(X \,|\, Y = y) = \frac{\frac{d}{dy}\,\mathrm{E}(X 1_{(Y \leq y)})}{f_Y(y)}.$$

Vis at denne formelen gir formelen i teksten når vi antar at X, Y er simultant kontinuerlig fordelt.

113 La $X_1, \ldots, X_n$ være uavhengige og uniformt fordelt på intervallet $[0, 1]$. Beregn $\mathrm{E}(X_1 \,|\, X_{\text{maks}} = x)$ ved formelen i forrige oppgave.

Kapittel 9

Poissonfordelingen og de små talls lov

I en viss forstand er den uniforme fordelingen, poissonfordelingen og normalfordelingen de tre viktigste fordelingene i sannsynlighetsteorien.

Den uniforme fordelingen er viktig av symmetrigrunner, og er koplet til den klassiske definisjonen av sannsynlighet. Klassiske sannsynligheter beregnes kombinatorisk med utgangspunkt i utfall som er like sannsynlige, dvs med utgangspunkt i en uniform modell. Spesielt leder kombinatoriske argument til binomialfordelingen.

Innholdet i de små talls lov er at poissonfordelingen fremkommer som en grense av binomialfordelingen. Fordelingen er knyttet til mange anvendelser, f eks i forbindelse med ventetidsproblemer.

Vi skal senere vise at normalfordelingen fremkommer som en grense hvor utgangspunktet er et tilfeldig utvalg fra en vilkårlig fordeling. Dette er sentralgrenseteoremet. Det er ikke uten grunn at G.Polya i 1920 refererte til teoremet som *den zentralen Grenzwertsatz der Wahrscheinlichkeitsrechnung.*

9.1 Poissonfordelingen og Poisson-punktprosesser

En tilfeldig variabel X er poissonfordelt med parameter $\lambda \geq 0$ dersom Definisjon: Poisson$(\cdot)$

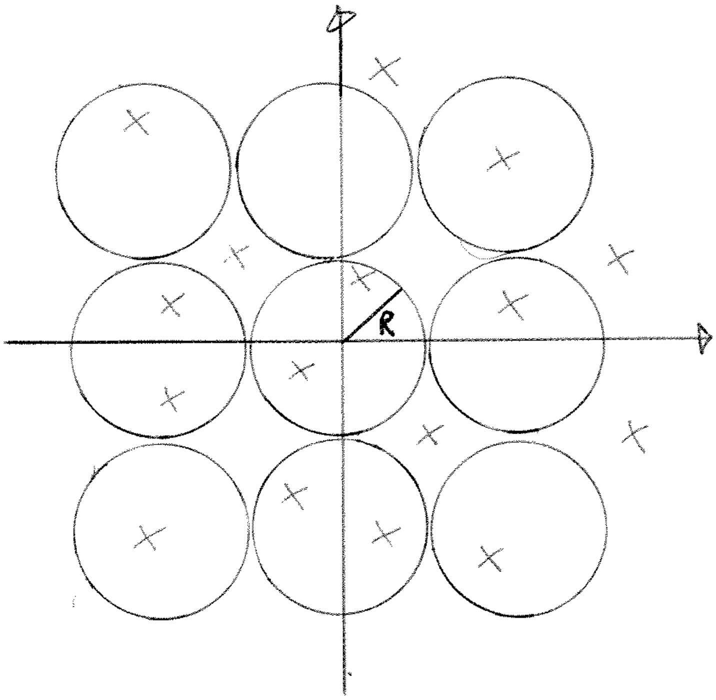
$$P(X = x) = \frac{\lambda^x}{x!}e^{-\lambda} \qquad x = 0, 1, \ldots \tag{9.1}$$

Dette noteres kort som $X \sim$ Poisson(λ). *Tegnet "$\sim$" leses som "er fordelt som" og brukes også mer generelt. F eks betyr $X \sim Y$ at X og Y har lik fordeling.* Definisjon: $X \sim Y$

Poissonfordelingen kan benyttes til modellering av f eks følgende situasjoner:

1. Telling av antall skjønnhetsfeil X (feil i overflaten) på en tallerken.
2. Telling av antall røde blodlegemer X i en blodprøve av angitt størrelse.
3. Telling av antall anrop X på en telefonsentral i løpet av en dag.
4. Telling av antall α-partikler X som sendes ut fra en radioaktiv kilde i løpet av et gitt tidsintervall.
5. Telling av antall biler X som passerer et gitt sted på veien i løpet av et gitt tidsintervall.

I de to første situasjonene kan usikkerheten innføres ved at en tenker seg et tilfeldig utvalg fra en stor beholdning. I de tre siste situasjonene er usikkerheten innebygget. En annen forskjell er at de tre siste situasjonene er knyttet til tid, mens de to første situasjonene er knyttet til rom.



Figur 9.1: Skjønnhetsfeil tilfeldig fordelt på tallerkener. Hver feil er gitt ved et kryss.

Poissonfordelingen kan utledes som fordelingen til en observator for et underliggende sannsynlighetsrom gitt ved en Poisson-punktprosess. Vi vil ikke gjøre dette, men ser i stedet på to eksempler. Det første eksemplet er geometrisk. Det andre eksemplet er knyttet til tid og leder også til to nye fordelinger: eksponensialfordelingen og gammafordelingen.

Eksempel 9.1.1 (Antall skjønnhetsfeil) La X være antall skjønnhetsfeil på en tallerken som velges tilfeldig fra en stor beholdning. Det er da klart at X kan ta verdier $x = 0, 1, 2, \ldots$, men det er usikkert hvilken verdi som tas. Vi vil anta at vi kan identifisere hver skjønnhetsfeil med et punkt på tallerkenen. Vi kan tenke oss en underliggende modell for å finne fordelingen til X. Plasser tallerkenene som på figur 9.1 og tenk deg at feilene drysses tilfeldig utover tallerkenene. La $Q = (Q_1, Q_2, \ldots)$ være de tilfeldige posisjonene til feilene. Dette betyr at hver Q_i er en observator, dvs Q er et eksempel på en stokastisk prosess. Antall feil i tallerkenen som er plassert i origo er da en funksjon ϕ av posisjonene gitt ved

$$X = \phi(Q) = \text{ antall } n \text{ slik at } |Q_n| \leq R.$$

Modellen kan spesifiseres ved å anta at samlingen $Q_1, Q_2, \ldots$ av tilfeldige punkter i planet er en Poisson-punktprosess. Da følger det at $X \sim \text{Poisson}(\lambda \pi R^2)$. Parameteren λ angir forventet antall feil per areal. Forventet antall feil per tallerken blir $\lambda \pi R^2$, dvs proposjonal med tallerkenens areal πR^2. Vi vil ikke behandle teorien for Poisson-punktprosesser i mer detalj.

Eksempel 9.1.2 (Telling av biler) Anta at vi står ved en vei og skal registrere biler som passerer. Tidspunktet T_1 for første bilpassering er da usikkert. Det samme gjelder tidspunktet T_n for n'te bilpassering. Vi vil anta at antall biler

$$N_t = \#\{n \,|\, 0 < T_n \leq t\}$$

som passerer i løpet av tidsintervallet $(0, t]$ oppfyller $N_t \sim \text{Poisson}(\lambda t)$. Parameteren λ angir forventet antall bilpasseringer per tid. Det følger at T_1 er eksponensialfordelt med parameter λ. Mer generelt følger det at T_n er gammafordelt med parametre n, λ. Vi vil utdype disse påstandene i resten av kapittelet.

9.2 Fordelinger knyttet til en Bernoulli-forsøksrekke

Definisjon: Bernoulli($\cdot$)

En tilfeldig variabel er bernoullifordelt med parameter p, $0 \leq p \leq 1$, dersom

$$P(X = x) = p^x (1 - p)^{(1-x)} \qquad x = 0, 1. \tag{9.2}$$

Vi sier kort at $X \sim \mathrm{Bernoulli}(p)$.

Dette er det samme som at $P(X = 0) = 1 - p$ og at $P(X = 1) = p$. Spesielt vil enhver indikatorvariabel 1_A være bernoullifordelt.

Den momentgenererende funksjonen

$$M_X(t) = \mathrm{E}\, e^{tX} = e^{t0}(1-p) + e^{t1}p = (1-p) + pe^t$$

gir forventningsverdi og varians

$$\mathrm{E}\, X = M_X'(0) = (pe^t)_{t=0} = p, \quad \mathrm{Var}\, X = \mathrm{E}\, X^2 - (\mathrm{E}\, X)^2 = M_X''(0) - p^2 = p - p^2,$$

som også kunne vårt funnet enkelt direkte.

Definisjon: **Bernoulli-forsøksrekke**

En Bernoulli-forsøksrekke er en rekke uavhengige like forsøk hvor det kun er to mulige utfall i hvert forsøk. En modell for en Bernoulli-forsøksrekke er gitt ved et tilfeldig utvalg $X_1, X_2, \ldots$ av Bernoulli-variable. I en suksess/fiasko modell tenker vi på resultatet $X_j = 1$ som suksess i forsøk nummer j. Resultatet 0 svarer tilsvarende til fiasko. Vi skal vise at en Bernoulli-forsøksrekke gir opphav til den binomiske-, den geometriske- og den negativt binomiske fordelingen.

Definisjon: binomisk$(\cdot)$

En tilfeldig variabel X er binomisk fordelt med parametre n, p hvor $n = 0, 1, \ldots$ og $0 \leq p \leq 1$ dersom

$$P(X = x) = \binom{n}{x} p^x (1-p)^{(1-x)} \qquad x = 0, 1, \ldots, n. \tag{9.3}$$

Vi sier kort at $X \sim \mathrm{binomisk}(n, p)$. Den momentgenererende funksjonen finnes ved bruk av binomialteoremet:

$$M_X(t) = \sum_{x=0}^{n} e^{tx} \binom{n}{x} p^x (1-p)^{n-x} = \sum_{x=0}^{n} \binom{n}{x} (e^t p)^x (1-p)^{n-x} = ((1-p) + pe^t)^n.$$

La $X_1, \ldots, X_n$ være et tilfeldig utvalg fra en Bernoulli(p) fordeling. Vi minner om at dette betyr at for alle n er $X_n \sim \mathrm{Bernoulli}(p)$ og variablene er uavhengige. Det kan vises kombinatorisk at $(X_1 + \cdots + X_n) \sim \mathrm{Bernoulli}(n, p)$. Et alternativ bevis er gitt ved

$$M_{X_1 + \cdots + X_n}(t) = \mathrm{E}\, e^{tX_1 + \cdots + tX_n} = \mathrm{E}\, e^{tX_1} \cdot \cdots \cdot \mathrm{E}\, e^{tX_n} = ((1-p) + pe^t)^n.$$

Dette er tilstrekkelig fordi en fordeling er entydig gitt av den tilhørende momentgenererende funksjonen. Vi kan ikke konkludere at X er lik summen av Bernoulli-variablene X_i, men substitusjonsteoremet gir

$$\mathrm{E}\, X = \mathrm{E}(X_1 + \cdots + X_n) = np, \quad \mathrm{Var}\, X = \mathrm{Var}(X_1 + \cdots + X_n) = n(p - p^2),$$

fordi forventningsverdien til en tilfeldig variabel kun avhenger av variabelens fordeling.

Vi vil utheve hovedpoenget:

> Antall suksesser i n uavhengige forsøk med suksessannsynlighet p er binomisk(n, p) fordelt.

Definisjon: geometrisk$(\cdot)$

En variabel N er geometrisk fordelt med parameter p, $0 < p \leq 1$, dersom

$$P(N = n) = p(1-p)^{n-1} \qquad n = 1, 2, \ldots. \tag{9.4}$$

Vi sier kort at $N \sim \mathrm{geometrisk}(p)$. Tilfellet $p = 1$ gir $P(N = 1) = 1$.

Den momentgenererende funksjonen finnes ved den geometriske rekken

$$M_N(t) = \mathrm{E}\, e^{tN} = \sum_{n=1}^{\infty} e^{tn} p(1-p)^{n-1} = pe^t \sum_{n=1}^{\infty} \left(e^t(1-p)\right)^{n-1} = \frac{pe^t}{1 - e^t(1-p)}.$$

Den momentgenererende funksjonen er kun definert for $e^t(1-p) < 1$. Funksjonen kan brukes til å finne

$$\mathrm{E}\, N = 1/p, \ \mathrm{Var}\, N = (1-p)/p^2.$$

Vi skal nå se at dette f eks kan tolkes som at forventet antall myntkast for å få krone er 2.

La $X_1, X_2, \ldots$ være et tilfeldig utvalg fra en Bernoulli(p) fordeling med $p > 0$. Definer variabelen

$$N = \min\{n \,|\, X_n = 1\}.$$

I en suksess/fiasko modell betyr dette at N er antall forsøk som skal til for å oppnå suksess. Uavhengigheten gir

$$\begin{aligned} P(N = n) &= P(X_1 = 0, \ldots, X_{n-1} = 0, X_n = 1) \\ &= P(X_1 = 0) \cdots P(X_{n-1} = 0) \cdot P(X_n = 1) = p(1-p)^{n-1}. \end{aligned}$$

Vi har dermed bevist at $N \sim$ geometrisk(p). Dersom et forsøk gjøres hver tidsenhet, f eks et forsøk hvert minutt, så kan vi identifiserer n med et tidspunkt. Da får vi følgende reformulering:

> Antall tidsenheter til suksess er geometrisk fordelt.

Definisjon: **negativt-** binomisk(·)

En variabel X er negativt binomisk fordelt med parametre r, p hvor $r > 0$ og $0 < p \leq 1$ dersom

$$P(X = x) = \binom{x + r - 1}{x} p^r (1-p)^x \qquad x = 0, 1, 2, \ldots. \tag{9.5}$$

Vi sier kort at $X \sim$ negativt-binomisk(r, p).

I tilfellet $y > 0$ og $x = 0, 1, \ldots$ er binomialkoeffisienten definert ved

$$\binom{y}{x} = \frac{y \cdot (y-1) \cdots (y - x + 1)}{x} \qquad \binom{y}{0} = 1 \tag{9.6}$$

Se på en suksess/fiasko modell gitt av tilfeldig utvalgte Bernoulli-variable $X_1, X_2, \ldots$. La Y_r være antall forsøk som skal til for å oppnå r suksesser. Da gir uavhengigheten at

$$\begin{aligned} P(Y_r = y) &= P(r - 1 \text{ suksesser i } y - 1 \text{ forsøk og suksess i } y\text{'te forsøk}) \\ &= \binom{y-1}{(y-1) - (r-1)} p^{r-1} (1-p)^{y-r} \cdot p. \end{aligned} \tag{9.7}$$

Av overstående følger det at $(Y_r - r) \sim$ negativt-binomisk(r, p). Ved å identifisere y med et tidspunkt som tidligere følger det at

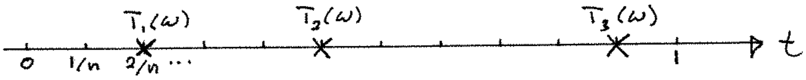
> Antall tidsenheter for å oppnå r suksesser minus r tidsenheter er negativt binomisk fordelt.

Konstruksjonen over med utgangspunkt i uavhengige Bernoulli-variable viser også at dersom $X_1, X_2, \ldots, X_r$ er et tilfeldig utvalg fra en geometrisk(p) fordeling, så er

$$[(X_1 - 1) + \cdots + (X_r - 1)] \sim \text{negativt-binomisk}(r, p). \tag{9.8}$$

9.3 Fra binomisk fordeling til poissonfordelingen

I mange situasjoner er det ikke naturlig å anta at tiden er diskret. Et eksempel er gitt ved registrering av biler. Det er ikke naturlig å kun registrere suksesser, dvs bilpasseringer, hvert minutt. Det vil være bedre å gjøre registreringen hvert sekund. Det mest naturlige synes imidlertid å være å tillate registreringer ved vilkårlige tidspunkter. Dette leder til en grensebetrakting hvor



Figur 9.2: De små talls lov. Når $n \to \infty$ er antallet punkter poissonfordelt. Samlingen $T_1, T_2, \ldots$ av punkter er da en Poisson-punktprosess T.

tiden mellom registreringene går mot 0 mens forventet antall registreringer per tidsenhet holdes konstant. Det er mulig å følge denne ideen gjennom den foregående diskusjonen skritt for skritt.

Et mulig første skritt er å se på grensen tilsvarende den geometriske fordelingen. Resultatet er en overgang fra den geometriske fordelingen til eksponensialfordelingen.

Definisjon: eksponential$(\cdot)$

Vi skriver $X \sim$ eksponential(λ), $\lambda > 0$, dersom X er kontinuerlig fordelt med tetthet

$$f_X(x) = \lambda e^{-\lambda x} \qquad x > 0. \tag{9.9}$$

Ved analogi med den geometriske fordelingen er følgende resultat ventet:

Tiden til første suksess er eksponensialfordelt.

Sammenlign dette med at du står og venter på bussen. Det er suksess når bussen kommer. Parameteren λ er lik forventet antall busspasseringer per tidsenhet.

Det andre skrittet er å se på summen av r uavhengige eksponensialfordelte variable som gir tiden til r suksesser. Vi velger i stedet en annen rute, som er lettere å generalisere til flere dimensjoner.

La $X_n \sim$ binomisk(n, p). Se på X_n som antall suksesser i tidsintervallet $(0, 1]$. Vi skal nå se på hva som skjer når n vokser. Mer detaljert ser vi på en oppdeling av intervallet i n like store intervaller. I høyre endepunkt av hvert av de n intervallene plasseres et punkt med sannsynlighet p. Dersom et punkt plasseres tilsvarer dette suksess. En mulig realisering er vist på figur 9.2. Når vi antar at plasseringen av punktene er uavhengige, så følger det at totalantallet er binomisk fordelt. Det forventede antall suksesser pn skal holdes konstant, så vi setter $pn = \lambda$. Vi antar altså at

$$X_n \sim \text{binomisk}(n, \lambda/n).$$

Ved å benytte

$$e^{-\lambda} = \lim_{n \to \infty} (1 - \lambda/n)^n$$

følger det at

$$P(X_n = x) = \frac{n!}{x!(n-x)!}(\lambda/n)^x (1-\lambda/n)^{n-x} = \frac{\lambda^x}{x!}(1-\lambda/n)^n \frac{n!}{(n-x)!n^x}(1-\lambda/n)^{-x} \to \frac{\lambda^x}{x!}e^{-\lambda},$$

når $n \to \infty$. Dette er de små talls lov. Vi vil reformulere resultatet.

Vi sier at X_n konvergere mot X i fordeling dersom Definisjon: $X_n \xrightarrow{L} X$

$$\lim_{n \to \infty} F_{X_n}(x) = F_X(x)$$

for alle x hvor $F_X(x)$ er kontinuerlig. Vi skriver kort $X_n \xrightarrow{L} X$. Bokstaven "L" står her for Lov (event. Law) i betydningen at fordelingen sees på som en lov for resultatet av et eksperiment.

De små talls lov har nå følgende formulering: La $X_n \sim$ binomisk$(n, \lambda/n)$ og la $X \sim$ Poisson(λ). Da vil $X_n \xrightarrow{L} X$ når n vokser. Dette er en konsekvens av at hvert av leddene i summen som gir sannsynligheten på venstresiden konvergerer mot de tilsvarende leddene på høyresiden. Vi gjentar hovedresultatet.

La X være poissonfordelt med intensitet λ. La X_n være binomisk fordelt tilsvarende n forsøk og suksessansynlighet λ/n. Da konvergerer X_n mot X i fordeling når n vokser.

De små talls lov kalles ofte også for Poisson-approksimasjonen. Dette skyldes at binomiske sannsynligheter tilsvarende store n verdier kan beregnes ved at de er tilnærmet lik den tilsvarende Poisson-sannsynligheten. Mer generelt kan forventningsverdien til en funksjon av en binomisk variabel approksimeres. Dette følger fordi det kan vises at konvergens i fordeling er ekvivalent med at

$$\lim \mathrm{E}\,\phi(X_n) = \mathrm{E}\,\phi(X),$$

for enhver begrenset og kontinuerlig ϕ.

9.4 Fra poissonfordelingen til gammafordelingen

Anta at $X \sim \text{Poisson}(\lambda)$. Poisson-approksimasjonen i foregående avsnitt rimeliggjør

$$\begin{aligned}
M_X(t) &= \lim_{n\to\infty} M_{X_n}(t) = \lim_{n\to\infty} ((1-\lambda/n) + e^t\lambda/n)^n = e^{\lambda(e^t-1)}, \\
\mathrm{E}\,X &= \lim_{n\to\infty} \mathrm{E}\,X_n = \lim_{n\to\infty} n\lambda/n = \lambda, \\
\mathrm{Var}\,X &= \lim_{n\to\infty} \mathrm{Var}\,X_n = \lim_{n\to\infty} n(\frac{\lambda}{n} - (\frac{\lambda}{n})^2) = \lambda.
\end{aligned}$$

Disse resultatene kan bevises mer direkte fra fordelingen til X.

Vi har tidligere sett at den geometriske og negativt binomiske fordelingen gir en beskrivelse av hhv ventetiden til 1 suksess eller r suksesser. Utgangspunktet var da at antall suksesser i et diskretisert tidsintervall var binomisk fordelt. Vi vil nå se på det tilsvarende kontinuerlige tilfellet hvor den binomiske fordelingen er erstattet av poissonfordelingen. Vi skal vise at ventetiden til r suksesser er gammafordelt. Tilfellet $r = 1$ gir eksponensialfordelingen.

Definisjon: gamma$(\cdot)$

En tilfeldig variabel X er gammafordelt med parametre $r > 0$, $\lambda > 0$ dersom X har tetthet gitt ved

$$f_X(x) = \frac{\lambda^r}{\Gamma(r)} x^{r-1} e^{-\lambda x} \qquad x > 0. \tag{9.10}$$

Vi skriver kort at $X \sim$ gamma(r, λ).

Definisjon: $\Gamma(\cdot)$

Gammafunksjonen Γ er definert ved

$$\boxed{\Gamma(r) = \int_0^\infty x^{r-1} e^{-x}\, dx.}$$

Den viktigste egenskapen til gammafunksjonen er at den generaliserer fakultetsfunksjonen:

$$n! = \Gamma(n+1) \qquad n = 0, 1, \ldots. \tag{9.11}$$

Dette følger av at

$$\Gamma(1) = 1$$

og rekursjonsformelen

$$\Gamma(r+1) = r\Gamma(r) \qquad r > 0. \tag{9.12}$$

Rekursjonsformelen kan bevises ved en delvis integrasjon. Spesialtilfellet

$$\Gamma(1/2) = \sqrt{\pi}$$

Tabell 9.1: Fordelinger knyttet til poissonfordelingen.

Tolkning	Type	Verdier	Parametre	Tetthet $f_X(x)$	$M_X(t)$	E X	Var X
Suksess?	Bernoulli(p)	x = 0, 1	$0 \leq p \leq 1$	$p^x(1-p)^{(1-x)}$	$(1-p)+pe^t$	p	$p(1-p)$
Antall suksesser	binomisk(p)	x = 0,..., n	$0 \leq p \leq 1$	$\binom{n}{x}p^x(1-p)^{(1-x)}$	$((1-p)+pe^t)^n$	np	$np(1-p)$
Tid til suksess	geometrisk(p)	$x = 1, 2, \ldots$	$0 < p \leq 1$	$p(1-p)^{x-1}$	$\frac{pe^t}{1-e^t(1-p)}$ $t < \ln \frac{1}{1-p}$	$1/p$	$(1-p)/p^2$
Tid til r suksesser minus r	negativt binomisk(r, p)	x = 0, 1, ...	$r > 0$ $0 < p \leq 1$	$\binom{x+r-1}{x}p^r(1-p)^x$	...	$\frac{r(1-p)}{p}$	$\frac{r(1-p)}{p^2}$
Antall suksesser	Poisson(λ)	$x = 0, 1, \ldots$	$\lambda \geq 0$	$\frac{\lambda^x}{x!}e^{-\lambda}$	$e^{\lambda(e^t-1)}$	λ	λ
Tid til suksess	eksponential(λ)	$x > 0$	$\lambda > 0$	$\lambda e^{-\lambda x}$	$\frac{\lambda}{\lambda-t}$	$1/\lambda$	$1/\lambda^2$
Tid til r suksesser	gamma(r, λ)	$x > 0$	$r, \lambda > 0$	$\frac{\lambda^r}{\Gamma(r)}x^{r-1}e^{-\lambda x}$	$\left(\frac{\lambda}{\lambda-t}\right)^r$	r/λ	r/λ^2

kan benyttes sammen med rekursjonsformelen til beregning av $\Gamma(r)$ for halvtallige r.

Vi vil nå utlede fordelingen til tidspunktet T_r for suksess nummer r. La W_t være antall suksesser i intervallet $(0, t]$. Da er

$$(T_r \leq t) = (W_t \geq r).$$

I ord: At tidspunktet for r'te suksess er mindre eller lik t er ekvivalent med at det er minst r suksesser i intervallet $(0, t]$. Ved å anta $W_t \sim$ Poisson(λt), så finner vi at $T_r \sim$ gamma(r, λ) fordi

$$\begin{aligned}
f_{T_r}(t) &= \frac{d}{dt}P(T_r \leq t) = \frac{d}{dt}P(W_t \geq r) = \frac{d}{dt}\sum_{w=r}^{\infty} e^{-\lambda t}\frac{\lambda^w t^w}{w!} \\
&= \sum_{w=r}^{\infty} -\lambda e^{-\lambda t}\frac{\lambda^w t^w}{w!} + \sum_{w=r}^{\infty} e^{-\lambda t}\frac{\lambda^w w t^{w-1}}{w!} \\
&= \sum_{w=r}^{r} e^{-\lambda t}\frac{\lambda^w w t^{w-1}}{w!} = \frac{\lambda^r}{(r-1)!}e^{-\lambda t}t^{r-1}
\end{aligned}$$

Antagelsen over betyr spesielt at forventet antall suksesser per tidsenhet er konstant lik λ. Vi har nå funnet den kontinuerlige beskrivelsen av ventetider:

> Ventetiden til r'te suksess er gammafordelt.

Det kan vises at summen av r tilfeldig valgte variable fra en eksponential(λ) fordeling er gamma(r, λ) fordelt. Dette kan tolkes som at fordelingen til neste suksess er eksponensialfordelt uavhengig av forhistorien. Resultatet gir også at familien av gammafordelte variable reproduserer seg selv ved at dersom $X_1 \sim$ gamma(r_1, λ), $X_2 \sim$ gamma(r_2, λ) og X_1, X_2 er uavhengige, så gjelder $X_1 + X_2 \sim$ gamma($r_1 + r_2, \lambda$).

En komprimert oppsummering av innholdet i dette kapitlet er gitt i tabell 9.1.

9.5 Oppgaver

114 Vis at en indikatorvariabel er bernoullifordelt.

115 Vis at en sum av uavhengige likt fordelte indikatorvariable er binomialfordelt ved et kombinatorisk argument.

116 Et ungt par planlegger å fortsette å få barn inntil de får en jente. Hva er deres forventede familiestørrelse? Gi en tolkning av svaret.

117 Vis at en geometrisk variabel X er uten hukommelse i den forstand at

$$P(X = n + k \mid X > n) = P(X = k).$$

118 Finn den momentgenererende funksjonen til en negativt binomisk variabel X. Beregn $\mathrm{E}\, X$ og $\mathrm{Var}\, X$.

119 Anta at X_1, X_2 er et tilfeldig utvalg fra en negativt-binomisk(r, p) fordeling. Vis at $X_1 + X_2 \sim$ negativt-binomisk$(2r, p)$. Gi en tolkning av resultatet ved å se på en sum av uavhengige geometriske variable.

120 En bildekkmaskin som er feiljustert gir et defekt bildekk med sannsynlighet $0,15$. Maskinen går hver dag inntil 3 defekte bildekk er produsert før den blir stoppet og justert. Hva er sannsynligheten for at maskinen produserer 5 eller flere bildekk før den blir stoppet? Hvor mange bildekk vil den produsere i gjennomsnitt før den blir stoppet?

121 Astronomer estimerer at 100 billioner stjerner i melkeveien har minst en planet i bane rundt seg. La p være sannsynligheten for at det finnes intelligent liv i et slikt solsystem. Hvor liten kan p være for at det skal være minst 50 % sjanse for at det skal finnes intelligent liv i en de andre solsystemene i melkeveien?

122 Regn ut den momentgenererende funksjonen til en Poisson-variabel. Benytt denne til beregning av forventning og varians.

123 Anta at X er poissonfordelt med $P(X = 1) = P(X = 2)$. Finn $P(X = 4)$.

124 En ung dame er svært populær blant de mannlige medstudentene. I gjennomsnitt mottar hun 4 telefoner hver kveld. Hva er sannsynligheten for at antall telefoner hun mottar i morgen er mer enn standardavviket pluss gjennomsnittet?

125 Anta at $X \sim$ Poisson(λ) og at den betingede fordelingen til Y gitt $X = x$ er binomisk med parametre x, p. Vis at $Y \sim$ Poisson(λp).

126 Anta at det i et visst land forekommer gjennomsnitlig $2,5$ flyulykker i året. Hva er sannsynligheten for at det skjer minst 4 flyulykker neste år? Hva er sannsynligheten for at tidsrommet mellom to flyulykker er mindre enn tre måneder?

127 Finn fordelingen til $X + Y$ når $X \sim$ Poisson(λ) og $Y \sim$ Poisson(μ) er uavhengige.

128 Vis at $\Gamma(\frac{7}{2}) = \frac{15\sqrt{\pi}}{8}$.

129 Vis at tettheten til gammafordelingen er normalisert.

130 Beregn den momentgenererende funksjonen til gammafordelingen.

131 Anta at $X_1, \ldots, X_n$ er uavhengige. Vis at gammafordelingen reproduserer seg selv ved at $X_i \sim$ gamma(r_i, λ) gir $\sum_i X_i \sum$ gamma$(\sum_i r_i, \lambda)$. Gi en tolkning av resultatet ved å se på en sum av uavhengige eksponensialfordelte variable.

132 Anta at en værstasjon på Svalbard har 3 vindmålere, hvor en er i drift og to er reserve. Hver av vindmålerne har en levetid som er eksponensialfordelt med forvented levetid på 1000 timer. Finn fordelingen til tiden det tar før alle er defekte. Hva er den forventede levetiden til vindmålingssystemet?

Kapittel 10

Normalfordelingen og det viktigste grenseteoremet

Dersom $X_1, \ldots, X_n$ er et tilfeldig utvalg fra en fordeling som har forventningsverdi μ og varians σ^2, så er summen $X_1 + \cdots + X_n$ tilnærmet normalfordelt. Mer presist gjelder

$$\lim_{n\to\infty} P\left(\frac{X_1 + \cdots + X_n - n\mu}{\sqrt{n}\sigma} \le z\right) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-\frac{t^2}{2}}\, dt.$$

Dette er det viktigste grenseteoremet i statistikken. Resultatet kalles sentralgrenseteoremet, som er en litt freidig oversettelse av navnet G.Polya ga teoremet i sin artikkel *Über den zentralen Grenzwertsatz der Wahrscheinlichkeitsrechnung und das Momentenproblem* i Mathematischen Zeitschrift (8), 1920. Høyresiden er fordelingsfunksjonen til en standard normalfordelt variabel. Teoremet forklarer meget generelt hvorfor summen av mange uavhengige tilfeldige variable har en tendens til å være godt beskrevet av normalfordelingen.

10.1 Egenskaper til normalfordelingen

En tilfeldig variabel X er normalfordelt med parametre μ, σ hvor $-\infty < \mu < \infty$ og $\sigma > 0$ dersom Definisjon: $\mathrm{N}(\cdot)$

$$P(X \le x) = \frac{1}{\sqrt{2\pi}} \frac{1}{\sigma} \int_{-\infty}^{z} e^{-\frac{(t-\mu)^2}{2\sigma^2}}\, dt \qquad -\infty < x < \infty. \tag{10.1}$$

Vi skriver kort at $X \sim \mathrm{N}(\mu, \sigma^2)$.

For videre beregninger trenger vi

$$\int_{-\infty}^{\infty} e^{-\frac{x^2}{2}}\, dx = \sqrt{2\pi}.$$

Dette vises ved omforming til et dobbeltintegral som kan beregnes ved polarkoordinater $x = r\cos\theta$, $y = r\sin\theta$:

$$\int_{-\infty}^{\infty} e^{-\frac{x^2}{2}}\, dx \cdot \int_{-\infty}^{\infty} e^{-\frac{y^2}{2}}\, dy = \int_{y=-\infty}^{\infty} (\int_{-\infty}^{\infty} e^{-\frac{x^2}{2}}\, dx) e^{-\frac{y^2}{2}}\, dy = \int_{y=-\infty}^{\infty} \int_{x=-\infty}^{\infty} e^{-\frac{x^2+y^2}{2}}\, dx dy$$
$$= \iint_{\mathbb{R}^2} e^{-\frac{r^2}{2}}\, dA = \int_{\theta=0}^{2\pi} \int_{r=0}^{\infty} e^{-\frac{r^2}{2}} r dr d\theta = 2\pi(-e^{-r^2/2})|_0^{\infty} = 2\pi$$

Anta at Y er standard normalfordelt, dvs anta at $Y \sim \mathrm{N}(0, 1)$. Da følger det at normaliseringen er i orden

$$P(-\infty < Y < \infty) = \frac{1}{\sqrt{2\pi}} \frac{1}{1} \int_{-\infty}^{\infty} e^{-\frac{(y-0)^2}{2\cdot 1^2}}\, dy = 1.$$

Den momentgenererende funksjonen finnes ved

$$M_Y(t) = \mathrm{E}\, e^{tY} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{ty} e^{-\frac{y^2}{2}} dy = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{(y-t)^2 - t^2}{2}} dy = e^{\frac{t^2}{2}}$$

Dermed kan forventning og varians finnes

$$\mathrm{E}\, Y = M_Y'(0) = e^{t^2/2} t \mid_{t=0} = 0, \ \mathrm{Var}\, Y = (M_Y''(0))^2 - (\mathrm{E}\, Y)^2 = e^{t^2/2}(t^2+1) \mid_{t=0} = 1$$

Vi vil nå vise at $Y \sim N(0,1)$ gir at $X = \sigma Y + \mu \sim N(\mu, \sigma^2)$. I regningen under foretas variabelskiftet $u = \sigma y + \mu$:

$$P(X \le x) = P(\sigma Y + \mu \le x) = P(Y \le \frac{x-\mu}{\sigma}) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\frac{x-\mu}{\sigma}} e^{-\frac{x^2}{2}} dx = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{(u-\mu)^2}{2\sigma^2}} du$$

Dette viser at $X \sim N(\mu, \sigma^2)$. Resultatet er viktig fordi det gjør det mulig å benytte tabulerte verdier for standard normalfordelingen til å beregne tilsvarende verdier for generelle normalfordelinger. Her vil vi benytte resultatet for å finne den momentgenererende funksjonen til en generell $Z \sim N(\mu, \sigma^2)$:

$$M_Z(t) = \mathrm{E}\, e^{tZ} = \mathrm{E}\, e^{tX} = \mathrm{E}\, e^{t\sigma Y + t\mu} = e^{t\mu} M_Y(t\sigma) = e^{t\mu + \frac{t^2\sigma^2}{2}}.$$

Vi kan også enkelt vise at valget av symboler for parametrene i normalfordelingen er velbegrunnet

$$\mathrm{E}\, Z = \mathrm{E}\, X = \mathrm{E}(\sigma Y + \mu) = \mu, \ \mathrm{Var}\, Z = \mathrm{Var}\, X = \mathrm{Var}(\sigma Y) + \mathrm{Var}(\mu) = \sigma^2.$$

Definisjon: $\mathrm{N}(\mu, 0)$

Vi har foreløpig ikke tillatt tilfellet $\sigma = 0$. *Vi utvider den tidligere definisjonen ved å skrive* $Z \sim \mathrm{N}(\mu, 0)$ *dersom* $P(Z = \mu) = 1$. Resultatet $\sigma Y + \mu \sim N(\mu, \sigma^2)$ er da et spesialtilfelle av følgende:

> Dersom X, Y er uavhengige og normalfordelt, så er lineærkombinasjonen $\alpha X + \beta Y$ normalfordelt.

Anta at $X \sim N(\mu_X, \sigma_X^2)$, $Y \sim N(\mu_Y, \sigma_Y^2)$, og at X, Y er uavhengige. Vi beviser at $X + Y \sim N(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2)$. Beviset følger ved å se på den momentgenererende funksjonen

$$M_{X+Y}(t) = \mathrm{E}\, e^{t(X+Y)} \stackrel{uavh}{=} \mathrm{E}\, e^{tX}\, \mathrm{E}\, e^{tY} = e^{t\mu_X + \frac{t^2\sigma_X^2}{2}} e^{t\mu_Y + \frac{t^2\sigma_Y^2}{2}} = e^{t(\mu_X \mu_Y) + \frac{t^2(\sigma_X^2 + \sigma_Y^2)}{2}}.$$

10.2 Bevis av det viktigste grenseteoremet i statistikken

Definisjon: **Normalisert utgave**

Generelt er den normaliserte utgaven Y *av en tilfeldig variabel* X *gitt ved*

$$Y := (X - \mathrm{E}\, X)/\sqrt{\mathrm{Var}\, X}.$$

Egenskapene $\mathrm{E}\, Y = 0$ og $\mathrm{Var}\, Y = 1$ begrunner navnet. En variabel X har en normalisert utgave dersom den har endelig forventning og varians.

Det viktigste grenseteoremet i statistikken er:

> La $X_1, \ldots, X_n$ være et tilfeldig utvalg fra en fordeling. Anta at den normaliserte utgaven Z_n av $X_1 + \cdots + X_n$ eksisterer. Da vil
>
> $$\lim_{n \to \infty} P(Z_n \le z) = P(Z \le z), \ \text{hvor } Z \sim \mathrm{N}(0,1)$$

En reformulering av det overstående er gitt ved å si at Z_n konvergerer i fordeling mot Z, eller enda kortere $Z_n \xrightarrow{L} Z$.

Det kan vises at dersom $M_{Z_n}(t) \to M_Z(t)$, så vil $Z_n \xrightarrow{L} Z$. Dette resulatet vil vi ta som gitt, og vår oppgave er da å bevise at $M_{Z_n}(t) \to e^{t^2/2}$.

Vi har $\mathrm{E}(X_1 + \cdots + X_n) = n\mu_X$ og $\mathrm{Var}(X_1 + \cdots + X_n) = n\sigma_X^2$. Den normaliserte utgaven av summen blir da

$$Z_n = \frac{X_1 + \cdots + X_n - n\mu_X}{\sqrt{n}\sigma_X}$$

Dette gir

$$\begin{aligned} M_{Z_n}(t) &= \mathrm{E}\, e^{tZ_n} = \mathrm{E}\, e^{t\frac{X_1+\cdots+X_n-n\mu_X}{\sqrt{n}\sigma_X}} = \mathrm{E}\, e^{\frac{t}{\sqrt{n}}\frac{X_1-\mu_X}{\sigma_X}+\cdots+\frac{t}{\sqrt{n}}\frac{X_n-\mu_X}{\sigma_X}} \overset{uavh}{=} \mathrm{E}\, e^{\frac{t}{\sqrt{n}}\frac{X_1-\mu_X}{\sigma_X}} \cdots \mathrm{E}\, e^{\frac{t}{\sqrt{n}}\frac{X_n-\mu_X}{\sigma_X}} \\ &= (\mathrm{E}\, e^{\frac{t}{\sqrt{n}}W})^n. \end{aligned}$$

Oppgaven er dermed redusert til å bevise at

$$\left(\mathrm{E}\, e^{\frac{t}{\sqrt{n}}W}\right)^n \to e^{t^2/2}.$$

når det antas at $\mathrm{E}\, W = 0$ og $\mathrm{Var}\, W = 1$.

En mulighet er å bruke rekkeutviklingen $e^x = 1 + x + x^2/2 + \cdots$ til eksponensialfunksjonen:

$$(\mathrm{E}\, e^{\frac{t}{\sqrt{n}}W})^n = (\mathrm{E}(1 + \frac{t}{\sqrt{n}}W + \frac{t^2}{2n}W^2 + \cdots))^n = (1 + 0 + \frac{t^2}{2n}1 + \cdots)^n \to e^{t^2/2}.$$

I det overstående kan det by på problemer å begrunne de enkelte overgangene.

Et bevis via to gangers bruk av L'Hopitals regel er en annen mulighet:

$$\begin{aligned} \lim_{n\to\infty} \ln(\mathrm{E}\, e^{\frac{t}{\sqrt{n}}W})^n &= \lim_{n\to\infty} \frac{\ln(\mathrm{E}\, e^{\frac{t}{\sqrt{n}}W})}{1/n} = \lim_{x\to 0} \frac{\ln(\mathrm{E}\, e^{xtW})}{x^2} \\ &= \lim_{x\to 0} \frac{(\mathrm{E}\, e^{xtW}tW)/(\mathrm{E}\, e^{xtW})}{2x} = \lim_{x\to 0} \frac{1}{2} \frac{\mathrm{E}(e^{xtW}t^2W^2)(\mathrm{E}\, e^{xtW}) - (\mathrm{E}\, e^{xtW}tW)^2}{(\mathrm{E}\, e^{xtW})^2} = t^2/2 \end{aligned}$$

I bevisene over er det stilltiende antatt at alle de momentgenererende funksjonenen finnes. Et mer generelt gyldig bevis er gitt ved å ta utgangspunkt i den karakteristiske funksjonen $\phi_Z(t) = M_Z(it)$ i stedet. Begge bevisene vi ga kan modifiseres til dette tilfellet.

I beregningene er det benyttet et ombytte av henholdsvis summasjon og derivasjon med forventningsverdi. Begge disse ombyttene kan begrunnes, men det kreves mer verktøy enn vi kan gi i denne framstillingen.

10.3 Khi-kvadrat-, Fisher- og Student-t-fordelingene

Vi vil se på tre fordelinger som er avledet fra normalfordelingen. Disse fordelingene vil vi senere bruke i statistikkdelen av denne fremstillingen.

Vi gir her bare en oppsummering av hvordan disse fordelingene fremkommer.

La Z være standard normalfordelt. Da vil $Y = Z^2$ være khi-kvadratfordelt med 1 frihetsgrad.

La $Z_1, \ldots, Z_n$ være et tilfeldig utvalg fra en $N(0,1)$ fordeling. Da er Definisjon: ${\chi^2}_n(\cdot)$

$$Y = Z_1^2 + \cdots + Z_n^2$$

khi-kvadratfordelt med n frihetsgrader. Vi skriver kort $Y \sim {\chi^2}_n$.

La $U \sim {\chi^2}_m$, $V \sim {\chi^2}_n$ og anta at U, V er uavhengige. Da er variabelen $F = (U/m)/(V/n)$ Fisher-fordelt med m og n frihetsgrader. Vi skriver kort $F \sim$ Fisher(m,n). Definisjon: Fisher$(\cdot)$

La $Z \sim N(0,1)$, $V \sim {\chi^2}_n$ og anta at Z, V er uavhengige. Da er variabelen $T = Z/\sqrt{V/n}$ Student-t-fordelt med n frihetsgrader. Vi skriver kort at $T \sim$ Student-t(n). Definisjon: Student-t$(\cdot)$

Metoden med fordelingsfunksjonen og metoden med den momentgenererende funksjonen kan brukes for å finne tetthetene til ${\chi^2}_n$-, Fisher(m,n)- og Student-t(n)-fordelingene.

10.4 Oppgaver

133 Beregn integralet

$$I = \frac{1}{\sqrt{2\pi}} \int_{1.25}^{2.50} e^{-\frac{t^2}{2}} dt$$

ved bruk av tabell.

134 La $X \sim$ binomisk$(100, 0.4)$. Beregn $P(38 < X < 43)$ ved hjelp av sentralgrenseteoremet. Hvorfor kan sentralgrenseteoremet benyttes? Kan Poisson-approksimasjonen benyttes her?

135 Flyselskapene A og B gir identisk service på to flyavganger med samme avgangstid, dvs sannsynligheten for at en passasjer velger selskap A er 1/2. Anta at selskapene konkurerer om en kundemasse på 400 potensielle passasjerer. Selskap A har en kapasitet på 230 passasjerer, og selger en bilett til enhver som vil kjøpe. Hva er sannsynligheten for at selskap A får for mange passasjerer?

136 Diameteren til styrestangen i en viss sportsbil må være mellom 1.480 og 1.500 cm (inklusive endepunktene) for å være brukbar. Fabrikken produserer med en diameter som er normalfordelt med forventning 1.495 cm og standardavvik 0.005 cm. Hvor mange prosent blir ubrukelige?

137 La X være en tilfeldig variabel med momentgenererende funksjon

$$M_X(t) = e^{3t+8t^2}.$$

Beregn $P(-1 \leq X \leq 9)$ ved bruk av tabell.

138 Blodtrykket (systolisk) til 18 år gamle kvinner er normalfordelt med forventning 120 mm Hg og standardavvik 12 mm Hg. Hva er sannsynligheten for at en tilfeldig valgt 18 år gammel kvinne har et blodtrykk over 150 ? Mellom 110 og 130?

139 La $X_1, \ldots, X_9$ være et tilfeldig utvalg fra en $\mathrm{N}(2, 4)$ fordeling. La $Y_1, \ldots, Y_4$ være et uavhengig tilfeldig utvalg fra en $\mathrm{N}(1, 1)$ fordeling. Beregn $P(\overline{X} \geq \overline{Y})$.

140 Hvor stort må et tilfeldig utvalg fra en normalfordeling være for at utvalgets forventning $\overline{X}$ med sannsynlighet større enn 90% skal ligge mindre enn et standardavvik unna fordelingens forventning μ? Mindre enn to standardavvik unna?

141 En heis i idretts-seksjonen ved et statsuniversitet i USA skal ha en kapasitet på 2400 pounds. Anta at 10 fotballspillere går på heisen i 20 etasje. Hva er sannsynligheten for at det vil mangle 10 spillere ved neste trening når vekten til fotballspillerne ved universitetet er normalfordelt med forventning 220 pounds og standardavvik 20 pounds?

142 En student bruker minibankkortet 100 ganger mellom hver månedlig utskrift fra banken. I stedet for å notere de eksakte beløpene, så avrunder han beløpet til nærmeste 10 krone. La Y_j være avrundingsfeilen tilhørende transaksjon nummer j, og anta at Y_j er uniformt fordelt i intervallet $[-5, 5]$.

Benytt sentralgrenseteoremet til å approksimere sannsynligheten for at avrundingsfeilen blir større (i absoluttverdi) enn 50 kroner i løpet av en måned.

Sammenlign resultatet med det som finnes ved bruk av Chebyshevs ulikhet.

143 Anta at en terning kastes 100 ganger. Estimer sannsynligheten for at summen av resultatene er større enn 370.

144 La $\{X_j\}$ være et tilfeldig utvalg av størrelse n fra en Poisson(λ) fordeling. Vis at da er

$$X_1 + \cdots + X_n \sim \text{Poisson}(n\lambda).$$

Begrunn at en Poisson-variabel med stor parameter er tilnærmet normalfordelt.

145 Et elektronikkfirma mottar gjennomsnittlig 50 bestillinger i uken på en elektronisk brikke. Anta at antall bestillinger er poissonfordelt. Approksimer (ved sentralgrenseteoremet) sannsynligheten for at et lager på 60 brikker er for lite for en uke.

146 Anta at $X \sim {\chi^2}_n$. Vis at $\mathrm{E}\, X = n$ og $\mathrm{Var}\, X = 2n$ når det er gitt at dette er sant for $n = 1$.

147 La $Z_1, \ldots, Z_4$ være et tilfeldig utvalg fra N$(1, 1)$-fordelingen. Anta at

$$\sum_j Z_j^2$$

beskriver resultatet av et eksperiment. Når dette eksperimentet gjentas 100 ganger, hvor mange ganger venter du at resultatet ligger mellom 2 og 3?

148 Vis at dersom $X \sim \text{Fisher}(m, n)$, så er $1/X \sim \text{Fisher}(n, m)$.

149 Kvantilen $t_{\alpha,\nu}$ er definert ved $P(T > t_{\alpha,\nu}) = \alpha$ hvor $T \sim$ Student-t(ν). Finn $t_{0.05,24}$ i en tabell.

150 Anta at $T \sim$ Student-t(ν). Vis at $-T \sim$ Student-t(ν). Hva kan du konkludere om tettheten til T?

151 Dersom X, Y er uavhengige og normalfordelt, så er lineærkombinasjonen $\alpha X + \beta Y$ normalfordelt. Bruk dette til å bevise at $\sigma Y + \mu \sim N(\mu, \sigma^2)$ når det antas at Y er standard normalfordelt.

152 Prøv å gi et bevis for sentralgrenseteoremet i et mer generelt tilfelle. Hint: Behold regningen i beviset som er gitt, men modifiser antagelsene slik at variablene i summen kan ha ulike fordelinger.

Kapittel 11

Noen andre viktige fordelinger

11.1 Khi-kvadrat-, Fisher F- og student t-fordelingene

Vi vil se på tre fordelinger som er avledet fra normalfordelingen. Disse fordelingene vil vi senere bruke i statistikkdelen av denne fremstillingen.

Vi begynner med å gi en oppsummering av hvordan disse fordelingene fremkommer.

La Z være standard normalfordelt. Da vil $Y = Z^2$ være khi-kvadratfordelt med 1 frihetsgrad. La mer generelt $Z_1, \ldots, Z_n$ være et tilfeldig utvalg fra en $N(0, 1)$ fordeling. Da er $Y = Z_1^2 + \cdots + Z_n^2$ khi-kvadratfordelt med n frihetsgrader. Vi skriver kort $Y \sim \chi_n^2$.

La $U \sim \chi_m^2$, $V \sim \chi_n^2$ og anta at U, V er uavhengige. Da er variabelen $F = (U/m)/(V/n)$ Fisher fordelt med m og n frihetsgrader. Vi skriver kort $F \sim \text{Fisher}(m, n)$.

La $Z \sim N(0, 1)$, $V \sim \chi_n^2$ og anta at Z, V er uavhengige. Da er variabelen $T = Z/\sqrt{V/n}$ Student t fordelt med n frihetsgrader. Vi skriver kort at $T \sim \text{Student-t}(n)$.

Metoden med fordelingsfunksjonen og metoden med den momentgenererende funksjonen vil bli brukt for å finne sannsynlighetstetthetene til khi-kvadrat-, Fisher F- og Student t-fordelingene.

Del III

Statistikk

Kapittel 12

Statistiske modeller og punktestimering

Most studies and experiments, scientific or industrial, large scale or small, produce data whose analysis is the ultimate object of the endeavor. Mathematical statistics deals with situations in which the data can be thought of as the outcome of a random experiment.

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

En sannsynlighetsmodell kan benyttes til å beregne fordelingen til observasjonene i et eksperiment. Utgangspunktet for modellen er antagelser om en underliggende usikkerheten, dvs det antas at den underliggende usikkerheten er gitt ved én fordeling.

I en statistisk analyse er problemstillingen snudd. Med utgangspunkt i en statistikk er hensikten å trekke konklusjoner om antagelsene i en underliggende modell. En statistisk modell er gitt ved en antagelse om at den underliggende usikkerheten er beskrevet av en av flere mulige fordelinger. Familien av mulige fordelinger indekseres av en parametermengde, dvs til hver parameter svarer det en fordeling.

Forventningsverdien til resultatet av et eksperiment blir da en funksjon av parameteren i modellen. En type statistisk konklusjon er gitt ved et estimat for forventningsverdien beregnet ved hjelp av gitte data. Estimatet er med andre ord en funksjon av de gitte data.

Generelt er en estimator en funksjon av en observator. Fordi observatoren er en tilfeldig størrelse følger det at estimatoren selv også er en tilfeldig størrelse. Fordelingen til en estimator kan benyttes til en vurdering av om estimatoren er god. F eks sies en estimator å være forventningsrett dersom forventningsverdien er lik estimanden.

En gitt statistikk modelleres ofte ved å anta at den er resultatet av en tilfeldig trekning fra én av flere mulige fordelinger. Vi skal bevise at et tilfeldig utvalgs forventning og varians er eksempler på forventningsrette estimatorer for fordelingens forventning og varians.

KJENT: EKSPERIMENT, SANNSYNLIGHETSMODELL, UTFALLSROM, FORDELING, OBSERVATOR, TILFELDIG UTVALG

NYTT: STATISTISK MODELL, PARAMETERMENGDE, ESTIMATOR, ESTIMAT, ESTIMAND, FORVENTNINGSRETT, UTVALGETS FORVENTNING, UTVALGETS VARIANS

12.1 Statistiske modeller kontra sannsynlighets modeller

Anta at i en urne med 100 kuler, så er 37 hvite. Sannsynlighetene for de forskjellige mulige trekningene kan beregnes ved å spesifisere en sannsynlighetsmodell. F eks kan en slik modell begrunne at sannsynligheten for å trekke to kuler som er hvite er lik $(37/100) \cdot (36/99)$.

Anta så at antall hvite kuler er en ukjent størrelse θ. To kuler trekkes tilfeldig og resultatet er to hvite kuler. En statistisk modell kan settes opp for å trekke konklusjoner om θ med utgangspunkt i observasjonene. Vi skal se på konklusjoner av tre typer:

Punktestimering: Med utgangspunkt i observasjonene beregner vi et estimat w for den θ-verdien som passer best med observasjonene, f eks $w = 100 \cdot (1+1)/2$. Dette er punktestimering, og konklusjonen er at vi gjetter på at $\theta = 100$.

Intervallestimering: Vi finner et estimat s for et område hvor vi finner de θ verdiene som passer best med observasjonene, f eks $s = \{50, 51, \ldots, 100\}$. I tilfellet hvor θ er et reelt tall, så velger vi oftest å se på et område gitt ved et intervall $s = [a, b]$. Dette er intervallestimering, og konklusjonen i eksemplet er at vi gjetter på at $\theta \in [50, 100]$.

Hypotesetesting: Utgangspunktet er en gitt hypotese. Hypotesen kan f eks være at det er minst 50 hvite kuler. Konklusjonen er enten at hypotesen forkastes eller at den aksepteres, dvs det er bare to mulige konklusjoner. Dersom hypotesen aksepteres i eksemplet, så er konklusjonen at den gitte hypotesen $\theta \in \{50, 51, \ldots, 100\}$ er sann.

Definisjon: $\{P^\theta\}$, Θ

En statistisk modell er gitt ved en familie $\{P^\theta\}$ av fordelinger for et utfallsrom Ω. Parametermengden Θ er mengden av mulige verdier til parameteren θ.

Dette er en generalisering av den tilsvarende sannsynlighetsmodell hvor vi kun så på én fordeling P for Ω. Modellen er fullstendig spesifisert dersom vi for hver parameter θ har spesifisert fordelingen P^θ.

Vi gjentar hovedpoenget

> Forskjellen på en statistisk modell i forhold til en sannsynlighetsmodell er gitt ved innføringen av parametermengden og den tilhørende familien av mulige fordelinger. En statistisk modell brukes til å trekke konklusjoner om en mulig fordeling ved hjelp av data, mens en sannsynlighetsmodell brukes til å trekke konklusjoner om de mulige observasjonene når en fordeling er gitt.

Fordi utfallsrommet Ω er beholdt ved overgangen til en statistisk modell, så kan definisjonene av de sentrale begrepene i sannsynlighetsmodellen overføres direkte. Spesielt er en observator T en funksjon $T : \Omega \to \Omega_T$. Fordelingen P_T^θ til T er gitt ved

$$P_T^\theta(A) = P^\theta(T \in A).$$

Resultatet er en familie $\{P_T^\theta\}$ av fordelinger for Ω_T. I en statistisk modell har dermed hver observator en familie av mulige fordelinger.

12.2 Punktestimering

La $\{P^\theta\}$, $\theta \in \Theta$, være en statistisk modell. En estimator W for parameteren θ er en observator for Ω med verdier i en mengde Ω_W som inneholder parametermengden Θ. Det som skiller W fra en generell observator er altså tilleggsantagelsen $\Omega_W \supset \Theta$.

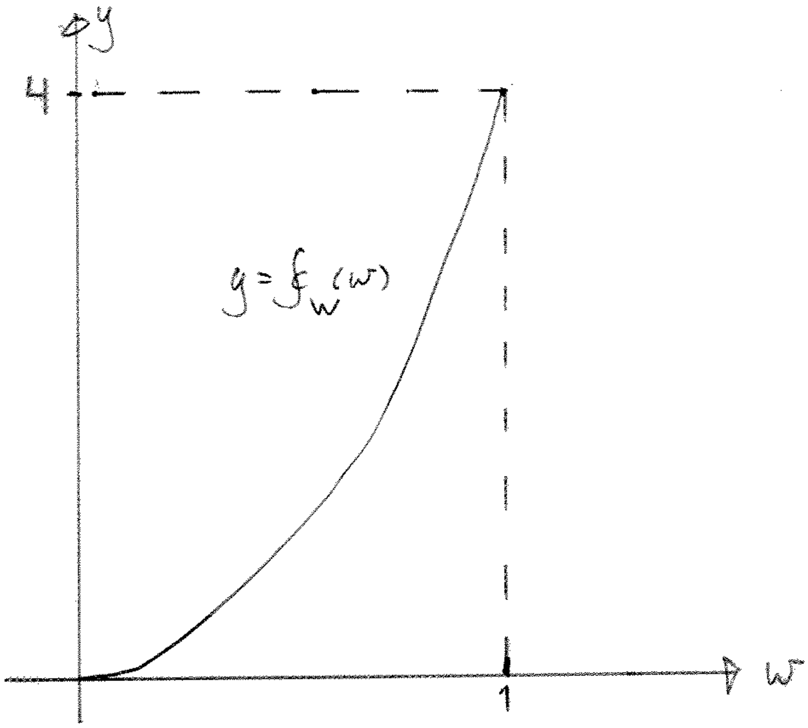
Eksempel 12.2.1 (Estimat for θ i en uniform fordeling) Anta at $X = (X_1, \ldots, X_n)$ hvor X_1, ..., X_n er et tilfeldig utvalg fra den uniforme fordelingen på intervallet $[0, \theta]$. Vår oppgave er å finne en estimator for θ. Uavhengigheten (tilfeldig utvalg!) gir at fordelingen P_X^θ til X er gitt av simultantettheten

$$f_X(x; \theta) = f_{X_1}(x_1; \theta) \cdots f_{X_n}(x_n; \theta).$$

Den uniforme fordelingen er gitt ved tettheten

$$f_Y(y; \theta) = \frac{1}{\theta}; \ \theta > 0; \ 0 \le y \le \theta.$$

Antagelsen om at vi har et utvalg fra den uniforme fordelingen betyr at $f_{X_i} = f_Y$ for alle i. Denne siste antagelsen kunne også vært skrevet som $X_i \sim \text{uniform}[0, \theta]$. Fordelingen P_X^θ er dermed fullstendig spesifisert av antagelsene. Parametermengden i denne modellen er $\Theta = \{\theta \,|\, \theta > 0\}$.



Figur 12.1: Tettheten til observatoren W ved $n = 4$ og $\theta = 1$

En mulig estimator for θ er gitt ved

$$W = \text{maks}\{X_1, \ldots, X_n\}.$$

Her er $W = \phi(X)$ hvor $\phi(x) = \text{maks}\{x_1, \ldots, x_n\}$. Dermed er W en funksjon av X, dvs W er basert på X.

Dersom utfallet er ω, så er estimatet lik $w = W(\omega) = \phi(X(\omega))$. Når utfallet $x_1 = X_1(\omega), \ldots, x_n = X_n(\omega)$ er gitt, så kan estimatet $w = \text{maks}\{x_1, \ldots, x_n\}$ beregnes. Poenget med dette siste er at beregningen kan gjøres uten at utfallet ω er kjent. Fordi W er basert på X kan w beregnes når x er gitt.

La $X_{(1)}, \ldots, X_{(n)}$ være ordningsobservatorene. Vi har da at $W = X_{(n)}$. Vi har tidligere funnet fordelingen til hver enkelt $X_{(i)}$. Spesielt er fordelingen til W gitt ved tettheten

$$f_W(w) = f_{X_{(n)}}(w) = n f_Y(w) F_Y(w)^{n-1}$$

Her er $F_Y(w) = \int_0^w f_Y(y)dy = w/\theta$ for $0 \leq w \leq \theta$. Dette gir

$$f_W(w) = \frac{n}{\theta^n} w^{n-1};\ \theta > 0;\ 0 \leq w \leq \theta.$$

Vi kan skissere tettheten til W for hver θ og n verdi. Et eksempel er gitt på figur 12.1. Spesielt gir en skisse av tettheten et inntrykk av hvordan fordelingen er lokalisert i forhold til verdien θ som skal estimeres.

Et numerisk lokaliseringsmål er gitt av forventningsverdien til W

$$\mathrm{E}^\theta W = \int_0^\theta w \cdot \frac{n}{\theta^n} w^{n-1} = \frac{n}{(n+1)\theta^n} w^{n+1} \Big|_0^\theta = \frac{n}{n+1}\theta.$$

Fordi $\mathrm{E}^\theta W \neq \theta$ sier vi at W ikke er forventningsrett. Vi kan lage en estimator V som er forventningsrett:

$$V = \frac{n+1}{n} W = \frac{n+1}{n} \text{maks}\{X_1, \ldots, X_n\}$$

gir

$$\mathrm{E}^\theta V = \frac{n+1}{n} \mathrm{E}^\theta W = \frac{n+1}{n} \frac{n}{n+1} \theta = \theta.$$

Mer generelt kan vi være interessert i en estimator for en funksjon $\tau(\theta)$ av parameteren θ.

La Θ være parametermengden i en statistisk modell. En estimand τ er en funksjon $\tau : \Theta \to \Theta_\tau$. Definisjon: **estimand**

Et eksempel på en estimand er gitt ved $\tau(\theta) = \theta$, hvor $\Theta_\tau = \Theta$. Et annet eksempel er gitt ved forventningsverdien til en tilfeldig variabel, dvs $\mu = \mathrm{E}\, X$. Notasjonen betyr at $\mu(\theta) = \mathrm{E}^\theta X$

for alle θ, og $\Theta_\mu = \mathbb{R}$. Denne konvensjonen tilsvarer at likhet $f = g$ mellom funksjoner betyr at $f(x) = g(x)$ for alle x i definisjonsområdet $D_f = D_g$. Konvensjonen med å utelate θ vil bli brukt også i det følgende.

Definisjon: **estimator** *En estimator er en observator. La $W : \Omega \to \Omega_W$ være en observator og la $\tau : \Theta \to \Theta_\tau$ være en estimand. Da er W en estimator for estimanden τ dersom $\Theta_\tau \subset \Omega_W$.*

Kravet $\Theta_\tau \subset \Omega_W$ er et meget svakt krav, men det gir dog en sammenheng mellom estimatoren og estimanden. Det er et svakt krav fordi W ikke trenger å ta alle verdier i Ω_W, og τ trenger ikke ta alle verdier i Θ_τ. F eks kan $\Theta_\tau = (0, \infty)$, $\Omega_W = \mathbb{R}$ og $W(\omega) = -10$, dvs estimatoren gir alltid estimatet -10. Denne estimatoren bommer alltid på estimanden τ som i dette tilfellet alltid er positiv.

Forventningsretthet er et kriterium som sikrer at en estimator "sikter på blinken".

Definisjon: **forventningsrett** *Estimatoren W er forventningsrett for estimanden τ dersom*

$$\boxed{\mathrm{E}\, W = \tau}$$

I tilfellet $W = (W_1, \ldots, W_n)$ for tilfeldige variable W_i betyr dette at $\mathrm{E}^\theta\, W_j = \tau_j(\theta)$ for alle θ og j. Det antas i det overstående at Ω_W er et vektorrom slik at forventningsverdien $\mathrm{E}\, W$ er definert.

Det er ønskelig at en estimator er forventningsrett fordi dette sikrer at estimatorens fordeling er lokalisert omkring verdien som skal estimeres. Vi har følgende reformulering

> En estimator er forventningsrett dersom tyngdepunktet til estimatorens fordelingen er lik estimanden.

I analysen av en gitt statistikk $x_1, \ldots, x_n$ er det vanlig å anta at statistikken er en realisering av et tilfeldig utvalg, dvs at

$$(x_1, \ldots, x_n) = (X_1(\omega), \ldots, X_n(\omega))$$

hvor $X_1, \ldots, X_n$ er et tilfeldig utvalg. Forventningen og variansen til utvalget er to meget brukte estimatorer.

Definisjon: **$\overline{X}$, S^2** *Anta at $X_1, \ldots, X_n$ er et tilfeldig utvalg. Utvalgets forventningsverdi er*

$$\boxed{\overline{X} = \frac{X_1 + \cdots + X_n}{n}}$$

og for $n \geq 2$ er utvalgets varians

$$\boxed{S^2 = \frac{(X_1 - \overline{X})^2 + \cdots + (X_n - \overline{X})^2}{n - 1}.}$$

Både $\overline{X}$ og S^2 er tilfeldige variable. De er basert på utvalget $X_1, \ldots, X_n$ og estimatene $\overline{x}$ og s^2 kan dermed beregnes når en statistikk $x_1, \ldots, x_n$ er gitt. Forventningsverdien μ og variansen σ^2 til fordelingen $P^\theta_{X_j}$ er begge funksjoner av θ, dvs de er estimander. Dermed er $\overline{X}$ og S^2 estimatorer for hhv μ og σ^2.

Grunnen til at vi dividerer med $n - 1$ i definisjonen av S^2 er gitt av det generelle resultatet:

> Et tilfeldig utvalgs forventningsverdi og varians er forventningsrette estimatorer for hhv forventningsverdi og varians.

Beviset for $\overline{X}$ bruker ikke antagelsen om uavhengighet

$$\mathrm{E}\, \overline{X} = \mathrm{E}\, \frac{X_1 + \cdots + X_n}{n} = \frac{\mathrm{E}\, X_1 + \cdots + \mathrm{E}\, X_n}{n} = \frac{\mu + \cdots + \mu}{n} = \mu.$$

Uavhengigheten gjør at vi også kan beregne variansen

$$\operatorname{Var}\overline{X} = \operatorname{Var}\frac{X_1 + \cdots + X_n}{n} = \frac{\operatorname{Var}X_1 + \cdots + \operatorname{Var}X_n}{n^2} = \frac{1}{n}\sigma^2.$$

Merk: Variansen til estimatoren $\overline{X}$ går mot null når utvalgets størrelse vokser. Dette tolkes som at presisjonen til estimatoren er bedre jo større utvalget er.

Beviset for at S^2 er forventningsrett er litt mer komplisert og bygger på formelen

$$\boxed{S^2 = \frac{n}{n-1}\left(\overline{X^2} - \overline{X}^2\right).}$$

Denne beregningsformelen gir ofte også den enkleste måten å beregne S^2. Beviset av beregningsformelen følger av identiteten

$$\overline{(x-\overline{x})^2} = \overline{x^2} - \overline{x}^2$$

som er et spesialtilfelle av beregningsformelen $\operatorname{Var}Z = \operatorname{E}Z^2 - (\operatorname{E}Z)^2$ for variansen. I det overstående er

$$\overline{X^2} = (X_1^2 + \cdots + X_n^2)/n.$$

Beregningsformelen gir følgende resultat.

> Et utvalgs forventningsverdi og varians er en funksjon av summen og kvadratsummen til utvalget.

Likheten $\operatorname{Var}Z = \operatorname{E}Z^2 - (\operatorname{E}Z)^2$ gir oss også at $\operatorname{E}X_i^2 = \operatorname{Var}X_i + (\operatorname{E}X_i)^2 = \sigma^2 + \mu^2$, som gir

$$\operatorname{E}\overline{X^2} = \frac{(\sigma^2+\mu^2) + \cdots + (\sigma^2+\mu^2)}{n} = \sigma^2 + \mu^2.$$

Forventningsverdien til $\overline{X}^2$ finnes ved

$$\operatorname{E}\overline{X}^2 = \operatorname{Var}\overline{X} + (\operatorname{E}\overline{X})^2 = \frac{1}{n}\sigma^2 + \mu^2.$$

Beviset av at S^2 er forventningsrett kan dermed konkluderes ved innsetting av de foregående resultater i beregningsformelen for S^2

$$\begin{aligned}\operatorname{E}S^2 = \operatorname{E}\frac{n}{n-1}[\overline{X^2} - \overline{X}^2] &= \frac{n}{n-1}[\operatorname{E}\overline{X^2} - \operatorname{E}\overline{X}^2] \\ = \frac{n}{n-1}[\sigma^2 + \mu^2 - \frac{1}{n}\sigma^2 - \mu^2] &= \frac{n}{n-1}[\sigma^2 - \frac{1}{n}\sigma^2] = \sigma^2.\end{aligned}$$

Eksempel 12.2.2 (Estimering av forventning og varians til en Bernoulli fordeling) La $X_1, \ldots, X_n$ være et tilfeldig utvalg fra en Bernoulli(p)-fordeling. Vi antar at $n \geq 2$. Fordelingen har forventningsverdi p og varians $p(1-p)$. Både forventningsverdien og variansen er funksjoner av parameteren p og er dermed estimander. Identifiser dette med en suksess/fiasko modell, dvs hendelsen $(X_j = 1)$ tolkes som suksess i forsøk j. La Y være antall suksesser. Da følger det at Y/n er en forventningsrett estimator for parameteren p. Dette er en konsekvens av at p er forventningsverdien til fordelingen og av at $Y = \sum X_i$.

En annen forventningsrett estimator for p er gitt av X_1. Estimatoren X_1 er ikke basert på Y: Selv om vi har gitt antall suksesser y, så kan vi ikke generelt avgjøre om forsøk nummer 1 var en suksess eller ikke. Både estimatoren Y/n og X_1 er imidlertid basert på $X = (X_1, \ldots, X_n)$. I følge kritererier som skal diskuteres senere vil det følge at X_1 er en dårligere estimator enn Y/n.

Vi kan også finne en forventningsrett estimator for variansen som er basert på Y. Fordi Bernoulli-variablene oppfyller $X_i^2 = X_i$ finner vi at utvalgets varians er basert på Y:

$$S^2 = \frac{n}{n-1}[\overline{X^2} - \overline{X}^2] = \frac{n}{n-1}[\overline{X} - \overline{X}^2] = \frac{n}{n-1}\frac{Y}{n}(1 - \frac{Y}{n}).$$

Det finnes også andre posisjonsmål for en fordeling, f eks medianen. Dette betyr at forventningsretthet bare er et av mange mulige mål for at posisjonen til estimatoren stemmer overens med verdien som skal estimeres.

12.3 Oppgaver

153 Anta at i en urne med 100 kuler, så er 37 hvite. Sannsynligheten for å trekke to kuler som er hvite er lik $(37/100) \cdot (36/99)$.

Begrunn dette på 3 måter ved å sette opp 3 forskjellige sannsynlighetsmodeller. Eventuelle tilfeldige variable som benyttes skal spesifiseres.

Hint: Produktregel for betingede sannsynligheter, ordnet utvalg og ikke-ordnet utvalg.

154 Anta at to kuler trekkes tilfeldig fra en urne med 100 kuler. Resultatet er to hvite kuler.

Sett opp en fullstendig spesifisert statistisk modell og estimer antall hvite kuler i urnen.

Finn variansen til estimatoren.

155 Hva er forskjellen (og likheten) mellom områdeestimering og hypotesetesting?

156 Forklar hva som menes med at forventningsverdien $\mathrm{E}\, X$ til en tilfeldig variabel X er en estimand τ. Er i så fall X en forventningsrett estimator for τ?

157 La $Y_1, \ldots, Y_4$ være et tilfeldig utvalg fra uniform$(0, \theta)$-fordelingen. En estimator for θ er gitt ved $W = Y_{\text{maks}}$. Finn sannsynligheten for at W ikke er lenger enn $1/8$ unna θ. Hva er sannsynligheten når $\theta = 1$?

158 La $Y_1, \ldots, Y_n$ være et tilfeldig utvalg fra eksponential$(1/\theta)$-fordelingen. En estimator for θ er gitt av $W = nY_{\text{min}}$. Finn tettheten f_W til W.

159 Ved et sykehus nær et lager for radioaktivt avfall fra et kjernekraftverk så registreres antall pasienter per år som får diagnosen kronisk leukemi. Resultatet de siste 12 årene er 2, 1, 0, 0, 1, 1, 1, 4, 1, 0, 2, 3. Anta at antall forekomster per år er beskrevet av en poissonfordeling. Estimer den relevante parameteren.

160 La $Y_1, \ldots, Y_n$ være et tilfeldig utvalg fra en fordeling med forventningsverdi μ. Under hvilke betingelser er estimatoren

$$W = \phi(Y_1, \ldots, Y_) = \sum_{j=1}^{n} a_j Y_j$$

forventningsrett for μ?

161 Anta at 14, 10, 18, 21 er resultatet av en trekning fra en uniform$(0, \theta)$-fordeling. Finn en forventningsrett estimator W for θ basert på den tredje ordningsobservatoren $Y_{(3)}$. Hva er verdien til W for utvalget som ble trukket? Gi et eksempel som viser at estimatoren W ikke er en god estimator.

162 La $X \sim$ uniform$(0, \theta)$ og definer $\tau(\theta) = \theta^2$. Finn en forventningsrett estimator W for τ.

Hint: La W være basert på X^2.

163 En estimator W_n er asymptotisk forventningsrett for τ dersom

$$\lim_{n \to \infty} \mathrm{E}^{\theta}\, W_n = \tau(\theta).$$

La $W_n = Y_{\text{maks}}$ hvor $Y_1, \ldots, Y_n$ er et tilfeldig utvalg fra uniform$(0, \theta)$-fordelingen. Er W_n forventningsrett for τ når $\tau(\theta) = \theta$? Er W_n asymptotisk forventningsrett?

164 Fremfor å kreve forventningsretthet, så kan en estimators fordeling sies å være riktig sentrert dersom estimatorens median er lik estimanden. En slik estimator kalles median-forventningsrett. Estimatoren $[(n+1)/n]Y_{\text{maks}}$ er forventningsrett for θ når $Y_1, \ldots, Y_n$ er et tilfeldig utvalg fra uniform$(0, \theta)$-fordelingen. Er den median-forventningsrett?

165 En estimator W er en beste lineær forventningsrett estimator for τ dersom

1. W er en lineær funksjon av observasjonene $X_1, \ldots, X_n$.
2. $\mathrm{E}\, W = \tau$.
3. $\mathrm{Var}\, W \leq \mathrm{Var}\, W^*$ for alle estimatorer W^* som oppfyller de to overnevnte krav.

Vis at et tilfeldig utvalgs forventning er en beste lineær forventningsrett estimator for fordelingens forventningsverdi.
Hint: La $W = a_1 X_1 + \cdots + a_n X_n + c$. Vis at $c = 0$. Partiellderiver mhp a_j'ene for å vise at $a_j = 1/n$.

166 Bevis at

$$\overline{(x - \overline{x})^2} = \overline{x^2} - \overline{x}^2$$

167 Bevis at S^2 er en forventningsrett estimator for σ^2.
Hva er antagelsene i beviset?

168 Bevis at $X^2 = X$ dersom X er en Bernoulli-variabel.

169 Hva betyr det at W er en observator for utfallsrommet Ω?
Hva betyr det at W er en estimator?
Hva betyr det at τ er en estimand?
Hva betyr det at W er en estimator for τ?
Hva betyr det at W er en estimator for τ basert på observatoren $(X_1, \ldots, X_n)$?

Kapittel 13

Nøyaktighet, konsistens og tilstrekkelighet

I do not know of any explicit statement of the properties, consistency, efficiency and sufficiency, which may characterize estimates prior to my 1922 paper. I had noted the functional peculiarity of sufficiency in an earlier paper, 1920, in the Monthly Notices of the Astronomical Society.

R.A.Fisher (1940)

Hensikten med en estimator er å beregne et estimat. Estimatet er vår gjetning på en ukjent størrelse i modellen. Nøyaktigheten til estimatoren sørger for at i en tenkt gjentagelse av eksperimentet, så vil de tilsvarende beregnede estimatene ligge tett samlet i nærheten av den ukjente størrelsen. Konsistensen til estimatoren sørger for at estimatet beregnet ved hjelp av et økende antall eksperiment konvergerer mot den ukjente estimanden. Spesielt svarer dette intuitivt til at dersom fordelingen til eksperimentet er kjent, dvs som om eksperimentet er gjentatt uendelig mange ganger, så gir estimatoren en riktig beregningsformel.

Tilstrekkelighet er kanskje det viktigste nye begrepet ved overgangen fra en sannsynlighetsmodell til en statistisk modell. Intuitivt er en observator tilstrekkelig for et eksperiment dersom verdiene $t_1, t_2, \ldots$ til observatoren inneholder samme informasjon om eksperimentet som verdiene $\omega_1, \omega_2, \ldots$ som eksperimentet gir. Informasjon om eksperimentet betyr her informasjon om parameteren i den statistiske modellen for eksperimentet. I praktiske eksempler kan ω tilsvare 100 tall, mens t er gitt ved 2 tall. Det sier seg selv at en slik datareduksjon kan gjøre analysen svært mye enklere.

KJENT: STATISTISK MODELL, PARAMETER, ESTIMATOR, FORVENTNINGSRETT, TSJEBYSJEVS ULIKHET, TETTHET, BETINGET FORDELING

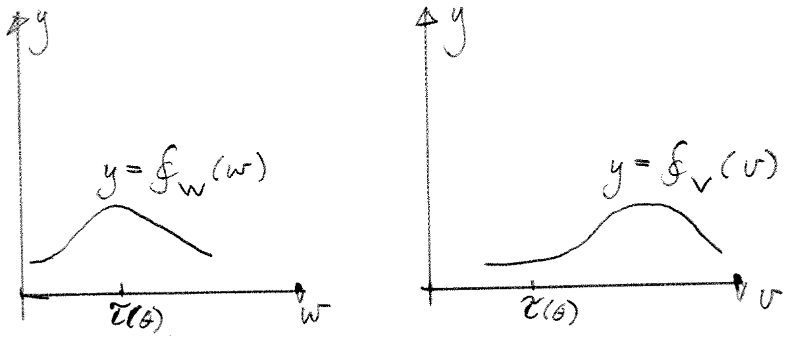
NYTT: NØYAKTIGHET, KONSISTENS, KJENT PARAMETER, TILSTREKKELIGHET

13.1 Nøyaktigheten til en punktestimator

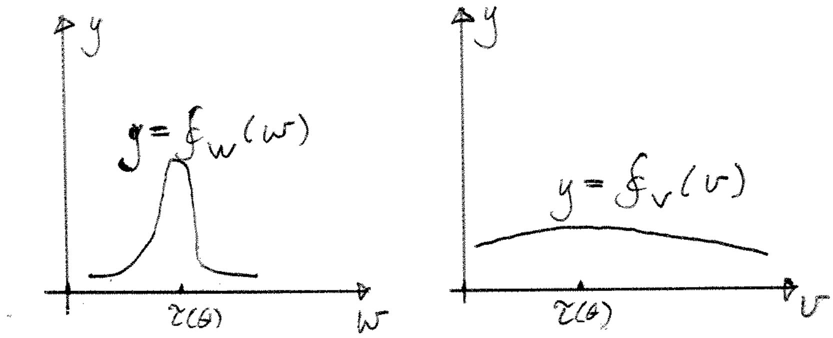
Vi begynner med en oppsummering av de viktigste definisjonene fra forrige kapittel. Utgangspunktet er en statistisk modell $\{P^\theta\}$ med en parametermengde Θ. En estimator W er en observator. Anta at $\tau : \Theta \to \Theta_\tau \subset \Omega_W$. Da er W en estimator for τ. Anta at utfallet er ω. Da gir estimatoren W estimatet $w = W(\omega)$. Estimatoren W er forventningsrett (unbiased) dersom $\mathrm{E}^\theta W = \tau(\theta)$ for alle θ.

Anta at V, W er tilfeldige variable med tettheter som på figur 13.1. På figurene fremgår det at W's fordeling er lokalisert omkring $\tau(\theta)$, mens dette ikke gjelder V's fordeling. Dermed vil vi konkludere at W er en bedre estimator for $\tau(\theta)$ enn V. Dette er et geometrisk argument.

En analytisk begrunnelse er gitt ved at vi foretrekker en forventningsrett estimator fremfor andre estimatorer som ikke er forventningsrette. En estimator W er jo forventningsrett for τ



Figur 13.1: Fordelingen til estimatoren W har bedre posisjon enn fordelingen til V i forhold til $\tau(\theta)$.



Figur 13.2: Fordelingen til estimatoren W har mindre spredning enn fordelingen til V.

dersom τ er tyngdepunktet til W's fordeling.

> Tenk et øyeblikk på W som om den beskriver resulatet av at vi skyter på blink. Vi forsøker å treffe τ. At W er forventningsrett kan sammenlignes med at vi sikter på riktig punkt.

Anta at V, W er tilfeldige variable med tettheter som på figur 13.2. Her foretrekkes W fremfor V fordi fordelingen til W har mindre spredning enn fordelingen til V. Dette kan reformuleres som at W er mer nøyaktig enn V. Vi vil bruke variansen til W som et mål for spredningen til W.

> Tenk igjen på W som om den beskriver resultatet av at vi skyter på blink etter τ. At W har liten varians kan sammenlignes med at de fleste skuddene treffer i nærheten av blinken.

Definisjon: **nøyaktighet**

Anta at V og W er forventningsrette estimatorer for τ. Da er W nøyaktigere enn V dersom

$$\operatorname{Var}^\theta W < \operatorname{Var}^\theta V \text{ for alle } \theta.$$

Definisjon: **relativ nøyaktighet**

Den relative nøyaktigheten til W i forhold til V er $\operatorname{Var} W / \operatorname{Var} V$.

Her er det, som alltid i en statistisk modell, underforstått at Var avhenger av parameteren θ. Konvensjonen med å unnlate å skrive opp θ avhengigheten brukes også i andre sammenhenger. Begrepet "relativ effisiens" er synonymt med relativ nøyaktighet. Alternativt benyttes også "relativ effektivitet".

Eksempel 13.1.1 (Mer nøyaktig estimator) Anta at $Y_i \sim$ eksponential$(1/\theta)$ og at $Y_1, \ldots, Y_n$ er uavhengige. Her er $\mathrm{E}\, Y_i = \theta$. Dette gir at $W = \overline{Y}$ og $V = Y_1$ er forventningsrette for θ. Vi beregner variansene ved

$$\mathrm{Var}\, V = \mathrm{Var}\, Y_1 = \theta^2, \ \mathrm{Var}\, W = \mathrm{Var}\, \overline{Y} = \mathrm{Var}\, \frac{Y_1 + \cdots \ Y_n}{n} = \frac{1}{n^2} \cdot n\theta^2 = \frac{1}{n}\theta^2.$$

Dette gir at $\mathrm{Var}\, W < \mathrm{Var}\, V$ når $n \geq 2$, dvs W er mer nøyaktig enn V når $n \geq 2$.

Vi velger følgende definisjon:

En estimator W for τ er en beste estimator dersom W er forventningsrett og $\mathrm{Var}\, W \leq \mathrm{Var}\, V$ *for alle forventningsrette estimatorer V.*

Definisjon: **beste estimator**

En beste estimator er dermed en forventningsrett estimator som har minst like liten varians som alle andre forventningsrette estimatorer. Som tidligere skal ulikheten $\mathrm{Var}\, W \leq \mathrm{Var}\, V$ tolkes som at

$$\mathrm{Var}^\theta\, W \leq \mathrm{Var}^\theta\, V \ \text{ for alle } \theta.$$

13.2 Konsistente estimatorer

La W_n være en estimator for τ. Da er W_n konsistent dersom

$$W_n \xrightarrow{P} \tau(\theta) \text{ for alle } \theta.$$

Definisjon: **konsistens**

Notasjonen $\xrightarrow{P}$ betyr konvergens i sannsynlighet. Dette betyr at for alle $\epsilon > 0$ vil

$$P(|W_n - \tau| < \epsilon) \to 1 \text{ når } n \to \infty.$$

I det overstående er både P og τ avhengig av θ og det er underforstått at påstanden skal gjelde for alle θ. En estimator W_n er mao konsistent for τ dersom W_n konvergerer i sannsynlighet mot τ.

Dersom W_n er forventningsrett for τ, så gir Tsjebysjevs ulikhet at

$$P(|W_n - \tau| < \epsilon) > 1 - \frac{\mathrm{Var}\, W_n}{\epsilon^2}.$$

Dermed har vi resultatet:

Dersom W_n er forventningsrett og $\mathrm{Var}\, W_n \to 0$, så er W_n konsistent.

Det finnes eksempler som viser at det motsatte ikke er tilfelle: W_n kan være konsistent uten at $\mathrm{Var}\, W_n \to 0$,

Et meget viktig eksempel på en konsistent estimator er gitt av et tilfeldig utvalgs forventningsverdi $\overline{X}$. Her er $W_n = \overline{X}$ hvor n er utvalgets størrelse. Det følger at W_n er en konsistent estimator for forventningsverdien $\mu = \mathrm{E}^\theta\, X_i$ dersom variansen $\sigma^2 = \mathrm{Var}^\theta\, X_i < \infty$. Begrunnelsen er at $\mathrm{Var}\, W_n = \sigma^2/n \to 0$. Resultatet er viktig i forhold til tolkningen av sannsynlighetsteorien og fortjener et eget navn.

Store talls lov (svak versjon) sier at et tilfeldig utvalgs forventningsverdi $\overline{X}$ er en konsistent estimator for forventningsverdien.

13.3 Tilstrekkelige observatorer

R. A. Fisher ga følgende definisjon av tilstrekkelighet:

En observator er tilstrekkelig dersom den inneholder all relevant informasjon.

Dette er ikke en matematisk definsjon. Formuleringen gir rom for flere mulige presiseringer. Vi velger følgende:

Definisjon: **tilstrekkelig** *Anta at X er en observator med tetthet $f_X(x;\theta)$. Observatoren T er tilstrekkelig for θ i forhold til X dersom $T = \phi(X)$ og*

$$f_X(x;\theta) = g(x) \cdot h(\phi(x);\theta).$$

Idé: g er uavhengig av θ, så denne er gitt uten at θ er kjent. All informasjon om θ-avhengigheten er i funksjonen h. Konklusjoner om θ finnes via konklusjoner om h, og denne avhenger kun av x gjennom $\phi(x)$.

All informasjon om θ inneholdt i X er dermed også inneholdt i $T = \phi(X)$.

Mer generelt defineres T til å være tilstrekkelig for θ i forhold til X dersom fordelingen til X gitt T er uavhengig av parameteren θ. Det kan vises at dette er i samsvar med definisjonen vi valgte. Den elementære definisjonen av betinget sannsynlighet er imidlertid ikke generell nok til at vi kan gjennomføre beviset her. Begrepet "suffisient" er synonymt med tilstrekkelighet.

Eksempel 13.3.1 (Tilstrekkelig observator for p) Vi skal vise at antall suksesser Y i uavhengige Bernoulli-forsøk $X_1, \ldots, X_n$ er en tilstrekkelig observator for suksess-sannsynligheten p. Vi har at

$$\begin{aligned} f_X(x;p) &= p^{x_1}(1-p)^{1-x_1} \cdots p^{x_1}(1-p)^{1-x_1} = p^{\sum x_i}(1-p)^{n-\sum x_i} \\ &= p^y(1-p)^{n-y} = g(x)h(\phi(x);p), \end{aligned}$$

hvor $g(x) = 1$, $y = \phi(x) = \sum x_i$, og $h(y;p) = p^y(1-p)^{n-y}$. Dermed har vi bevist at $Y = \phi(X)$ er tilstrekkelig.

Eksempel 13.3.2 (Tilstrekkelig observator for μ og for (μ, σ^2)) La $X_1, \ldots, X_n$ være et tilfeldig utvalg fra en $N(\mu, \sigma^2)$-fordeling.

Anta at σ er kjent. Dette betyr at parametermengden er

$$\Theta = \{\mu \,|\, -\infty < \mu < \infty\}.$$

Vi skal finne en tilstrekkelig estimator for X og μ. Tettheten til X er

$$\begin{aligned} f_X(x;\mu) &= \frac{1}{\sqrt{2\pi}}\frac{1}{\sigma}e^{-\frac{(x_1-\mu)^2}{2\sigma^2}} \cdots \frac{1}{\sqrt{2\pi}}\frac{1}{\sigma}e^{-\frac{(x_n-\mu)^2}{2\sigma^2}} \overset{!}{=} \left(\frac{1}{\sqrt{2\pi}}\frac{1}{\sigma}\right)^n e^{-\frac{n\overline{x^2}-2\mu n\overline{x}+n\mu^2}{2\sigma^2}} \\ &= \left(\frac{1}{\sqrt{2\pi}}\frac{1}{\sigma}\right)^n e^{-\frac{n\overline{x^2}}{2\sigma^2}} \cdot e^{-\frac{-2\mu n\overline{x}+n\mu^2}{2\sigma^2}}. \end{aligned}$$

Overgangen $\overset{!}{=}$ følger ved at

$$\sum(x_i-\mu)^2 = \sum x_i^2 - 2\mu\sum x_i + n\mu^2 = n\overline{x^2} - 2\mu n\overline{x} + n\mu^2.$$

Vi har faktorisert f_X i en faktor som er uavhengig av μ og i en faktor som kun avhenger av x gjennom $\overline{x}$. Dermed kan vi konkludere at utvalgets forventning $\overline{X}$ er en tilstrekkelig observator for μ når variansen σ^2 er kjent.

Anta så at både variansen og forventningsverdien er ukjente, dvs parametermengden er

$$\Theta = \{(\mu, \sigma^2) \,|\, \sigma > 0, -\infty < \mu < \infty\}.$$

Da er $T = (\overline{X}, S^2)$ en tilstrekkelig observator. Beviset følger av foregående regning når vi i tillegg bruker at

$$S^2 = \frac{n}{n-1}\left(\overline{X^2} - \overline{X}^2\right) \quad \text{som gir} \quad \overline{X^2} = \frac{n-1}{n}S^2 + \overline{X}^2.$$

Definisjon: **kjent parameter** *Anta at parameteren i en statistisk modell er på formen $\theta = (\theta_1, \theta_2)$. Når θ_2 er kjent betyr dette at parameterrommet er redusert fra Θ til*

$$\Theta_1 = \{\theta_1 \,|(\theta_1, \theta_2) \in \Theta\}.$$

Vi har et generelt tilstrekkelighetsprinsipp:

Dersom T er en tilstrekkelig observator, så skal estimatoren være basert på T.

Det er viktig å merke seg at overstående skal gjelde for alle tilstrekkelige observatorer.

Eksempel 13.3.3 (Diskvalifisering av en estimator og datareduksjon) La $X_1, \ldots, X_n$ være et tilfeldig utvalg fra en Bernoulli(p)-fordeling. Da er $\overline{X}$ og X_1 begge forventningsrette estimatorer for p basert på det tilfeldige utvalget. Estimatoren $\overline{X}$ er en tilstrekkelig observator, så denne inneholder all informasjon om p. Estimatoren X_1 diskvalifiseres fordi X_1 ikke er basert på $\bar{X}$. Derfor foretrekker vi estimatoren $\overline{X}$ fremfor X_1.

Enkeltobservasjonen $\bar{x}$ (antall suksess delt på n) inneholder dermed like mye informasjon om p som de n observasjonene $x_1, \ldots, x_n$. Dette viser at en passelig valgt tilstrekkelig observator kan lede til **datareduksjon**. Observatoren $X = (X_1, \ldots, X_n)$ er også tilstrekkelig. Vi velger å trekke slutninger om p med utgangspunkt i $\bar{x}$ fremfor x fordi det er enklere med bare ett tall.

Det kan virke utrolig at all informasjonen er samlet i et tall. Forklaringen er antagelsen om at observasjonene er gitt ved en så enkel modell. Dersom det f eks ikke er uavhengighet mellom X_i'ene, så er fortsatt X tilstrekkelig, men statusen til $\overline{X}$ er uklar.

Idéen om tilstrekkelighet er en av de mest elegante i statistikken. Vi har sett på to viktige konsekvenser. Den ene konsekvensen er en metode som generelt diskvalifiserer enkelte estimatorer. Den andre konsekvensen er at det i noen tilfeller er mulig å oppnå en kraftig reduksjon av datamengden som skal analyseres. Begge konsekvensene er imidlertid avhengig av **antagelsen** om riktig statistisk modell, og denne antagelsen kan være gal!

13.4 Oppgaver

170 La $Y_1, \ldots, Y_n$ være et tilfeldig utvalg fra eksponential($1/\lambda$)-fordelingen. La $W_1 = \overline{Y}$, $W_2 = Y_1$ og $W_3 = nY_{min}$. Vis at $W_1 \sim \text{gamma}(n, n/\lambda)$. Vis at W_i'ene er forventningsrette estimatorer for λ. Finn de relative nøyaktighetene til W_i'ene. Hvilken estimator er best?

171 Vi har diskutert to mulige estimatorer X_1 og $\overline{X}$ for den binomiske parameteren p. Sammenlign disse.

172 La $Y_1, \ldots, Y_n$ være et tilfeldig utvalg fra uniform($0, \theta$)-fordelingen. Vis at Y_{maks} er en konsistent estimator for θ. Anta at $\theta = 1$. Hvor stor må n være for at $P(|Y_{\text{maks}} - \theta| < 0,04) \geq 0,95$? Er Y_1 en konsistent estimator?

173 Vis at et tilfeldig utvalgs forventning er en konsistent estimator for forventningsverdien.

174 La $X \sim$ eksponential($1/\lambda$) og la $Y_1, \ldots, Y_n$ være et tilfeldig utvalg fra fordelingen til $X + \theta$. Vis at Y_{min} er en tilstrekkelig observator for θ når $\lambda = 1$.

Er $\overline{Y}$ en tilstrekkelig observator for λ? Hva betyr dette??

175 La X_1, X_2, X_3 være uavhengige Bernoulli-variable med $P(X_j = 1) = p$. Er $X_1 + 2X_2 + 3X_3$ en tilstrekkelig observator for p i forhold til X_1, X_2, X_3? I forhold til $\sum_j X_j$?

176 Gi et eksempel som viser at bruken av en tilstrekkelig observator kan gi datareduksjon. Betyr dette at de opprinnelige dataene er overflødige?

177 Anta at W_n konvergerer i sannsynlighet mot τ. Følger det da at W_n er en konsistent estimator for τ? Er W_n forventningsrett?

178 Hva betyr det at W er en beste estimator for τ?

179 La $\{P^\theta\}$ være en statistisk modell hvor alle fordelingene P^θ er diskrete.

Vis at dersom T er tilstrekkelig for θ i forhold til X, så er fordelingen $P_{X \mid T}$ til X gitt T uavhengig av θ.

Vis at dersom fordelingen $P_{X \mid T}$ til X gitt T er uavhengig av θ, så er T tilstrekkelig for θ i forhold til X.

180 La $Y_1, \ldots, Y_n$ være et tilfeldig utvalg fra eksponential$(1/\theta)$-fordelingen. En estimator for θ er gitt av $W_1 = nY_{\min}$. En annen estimator er gitt ved utvalgets forventning, dvs $W_2 = \overline{Y}$. Beregn $\operatorname{Var} W_1$ og $\operatorname{Var} W_2$. Hvilken estimator er mest presis?

Hint: Et av integralene som må beregnes kan beregnes ved integralet i definisjonen av gammafunksjonen.

Kapittel 14

Metoder for å finne estimatorer

Much modern work may be considered as stemming from a very remarkable paper of Fisher (1922) which gave, among other things, the idea of sufficiency. Fisher's later work on the foundations of statistical inference stressed likelihood and conditionality and his views on the bases are set out in his last book (Fisher, 1973); see also particularly his papers of 1925 and 1934.

D.R. Cox and D.V. Hinkley (1974)

I enkle tilfeller finnes det et naturlig estimat for en ukjent størrelse i en statistisk modell. F eks er naturlige estimatorer for en fordelings momenter gitt av et tilfeldig utvalgs momenter. Dette motiverer momentmetoden hvor et estimat finnes ved å kreve likhet mellom fordelingens første momenter og dataenes momenter. Momentmetoden gir ofte en estimator uten for mye regnearbeid.

Rimelighetsmetoden regnes for å være teoretisk bedre fundert. Utgangspunktet er rimelighetsfunksjonen til de gitte dataene, dvs tettheten til observatoren sett på som funksjon av parameteren i modellen. Et mulig prinsipp i en statistisk analyse er gitt ved at enhver konklusjon kun skal avhenge av rimelighetsfunksjonen til de gitte data. Spesielt må da et estimat være en funksjon av rimelighetsfunksjonen. Dette er spesielt tilfelle for en rimelighetsestimator, som ved definisjon er gitt ved å velge den parameterverdien som maksimaliserer rimelighetsfunksjonen. Et rimelighetsestimat er dermed gitt ved at det maksimaliserer rimeligheten, dvs sannsynligheten, til de gitte observasjonene. Intuitivt synes dette å være et godt valg.

KJENT: STATISTISK MODELL, ESTIMATOR, TILSTREKKELIG OBSERVATOR

NYTT: MOMENTESTIMATOR, RIMELIGHETSFUNKSJONEN, RIMELIGHETSESTIMATOR

14.1 Innledning

Så langt har vi definert hva en estimator er, og sett på noen egenskaper en estimator bør ha. Nå skal vi presentere to metoder som kan brukes til å finne estimatorer. Vi starter med en repetisjon av noen sentrale begrep.

En estimator er en observator W. Når dataene fra eksperimentet foreligger, så benyttes estimatoren til å beregne estimatet w.

Spesielt har estimatoren W en tilhørende fordeling. Denne kan brukes til å vurdere om estimatoren sikter riktig og om estimatene har liten spredning. Dette vurderes kvantitativt ved å se på om W er forventningsrett, $\mathrm{E}\, W = \tau$, og ved å se på om variansen $\mathrm{Var}\, W$ er liten.

I tillegg har vi diskutert konsistens, $W_n \xrightarrow{P} \tau$, og tilstrekkelighet. En observator $T = \phi(X)$ er tilstrekkelig for θ i forhold til X dersom vi har en faktorisering

$$f_X(x;\theta) = g(x)h(t;\theta), \quad \text{hvor } t = \phi(x).$$

Begrepet ga:

1. En metode for å diskvalifisere enkelte estimatorer.

2. I enkelte tilfeller en kraftig reduksjon av datamengden.

Et tilfeldig utvalgs forventning og varians er forventningsrette estimatorer for forventningen og variansen. Dette gjelder uansett hvilken statistisk modell som er valgt. Estimatorer for andre estimander vil typisk avhenge av modellen.

Vi skal se på momentmetoden (*method of moments*) og rimelighetsmetoden (*maximum likelihood method, sannsynlighetsmaksimeringsmetoden*). Momentmetoden gir oss en estimator ved løsning av et ligningssystem. Den kan brukes når f eks rimelighetsmetoden blir for komplisert.

14.2 Momentmetoden

La en statistisk model $\{P^\theta\}$ være gitt. Anta at $x_1, \ldots, x_n$ er resultatet av en trekning fra fordelingen til X.

Definisjon: $\tilde{\tau}$

Momentestimatet $\tilde{\tau}$ for estimanden $\tau : \Theta \to \Theta_\tau$ er gitt ved at $\tilde{\tau} = \tau(\tilde{\theta})$ hvor

$$\boxed{\mathrm{E}^{\tilde{\theta}}\, X^j = \overline{x^j} \qquad j = 1, \ldots, k.}$$

Ved løsning av ligningene velges k, hvis det er mulig, slik at ligningssystemet har en unik løsning. Det typiske er at $\theta = (\theta_1, \ldots, \theta_m)$, hvor hver θ_j er et reelt tall, og man velger $k = m$.

En reformulering av momentmetoden er gitt ved:

> Momentestimatoren finnes ved å kreve at fordelingens første momenter er lik verdien til et tilfeldig utvalgs første momenter.

Med $\tilde{\theta} = (\tilde{\theta}_1, \tilde{\theta}_2)$ betyr dette f eks at oppgaven er redusert til å løse to ligninger med to ukjente

$$\mathrm{E}^{\tilde{\theta}_1, \tilde{\theta}_2}\, X = \frac{1}{2}(x_1 + x_2) \qquad \mathrm{E}^{\tilde{\theta}_1, \tilde{\theta}_2}\, X^2 = \frac{1}{2}({x_1}^2 + {x_2}^2).$$

Eksempel 14.2.1 (Momentmetoden) Anta at $X_1, \ldots, X_n$ er et tilfeldig utvalg fra tettheten

$$f_X(x; \theta) = (\theta + 1) y^\theta \qquad 0 < x < 1.$$

Vi skal finne momentestimatoren for $\theta > -1$. Fordi vi bare har et ukjent tall trenger vi bare 1.moment, dvs forventningsverdien

$$\mathrm{E}^\theta\, X = \int_0^1 x(\theta + 1) x^\theta dx = \frac{\theta + 1}{\theta + 2} x^{\theta + 2} \Big|_0^1 = \frac{\theta + 1}{\theta + 2}$$

Vi trenger bare bruke den første momentligningen:

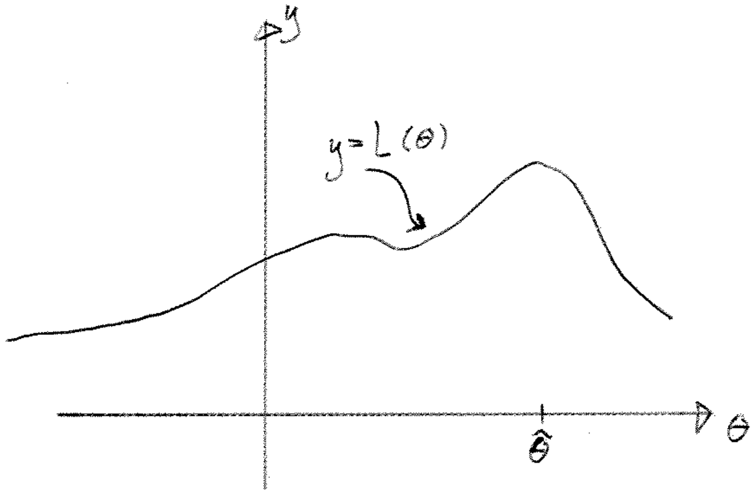
$$\mathrm{E}^\theta\, X = \frac{\theta + 1}{\theta + 2} = \overline{x} \qquad \text{som gir}$$

$$\theta + 1 = \overline{x}(\theta + 2),\ \theta(1 - \overline{x}) = 2\overline{x} - 1,\ \theta = \frac{2\overline{x} - 1}{1 - \overline{x}}.$$

Momentestimatoren er dermed

$$W = \frac{2\overline{X} - 1}{1 - \overline{X}}.$$

Merk: Metoden ga en estimator uten mye regning. I utgangspunktet var det vanskelig å gjette på mulige kanditater.



Figur 14.1: Rimelighetsestimat

14.3 Rimelighetsmetoden

La $\{P^\theta\}, \theta \in \Theta$, være en statistisk modell. Anta at observatoren X har en tetthet f_X. Denne tettheten er avhengig av θ fordi fordelingen P_X^θ til X er avhengig av θ.

Definisjon: L_X^x

Rimelighetsfunksjonen $L_X^x : \Theta \to \mathbb{R}$ til verdien x for observatoren X er definert ved

$$L_X^x(\theta) = f_X(x; \theta).$$

Notasjonen $L_X(\theta; x)$ brukes også. I de fleste tilfeller vil vi forenkle notasjonen ved å utelate en eller begge av indeksene X og x.

Det typiske tilfellet er at X er en vektor med diskret eller kontinuerlig fordeling. Betegnelsen "likelihood funksjonen" (the likelihood function) benyttes synonymt.

La x være en gitt observasjon. Da er $\hat{\tau}$ et rimelighetsestimat for estimanden $\tau : \Theta \to \Theta_\tau$ dersom $\hat{\tau} = \tau(\hat{\theta})$ hvor $\hat{\theta}$ er et globalt maksimalpunkt for rimelighetsfunksjonen L^x, dvs dersom Definisjon: $\hat{\tau}$

$$L^x(\theta) \leq L^x(\hat{\theta}) \qquad \textit{for alle } \theta.$$

Den intuitive ideen er gitt ved:

Vi velger den parameterverdien som gjør observasjonen så rimelig, dvs sannsynlig, som mulig.

Et eksempel er gitt i figur 14.1.

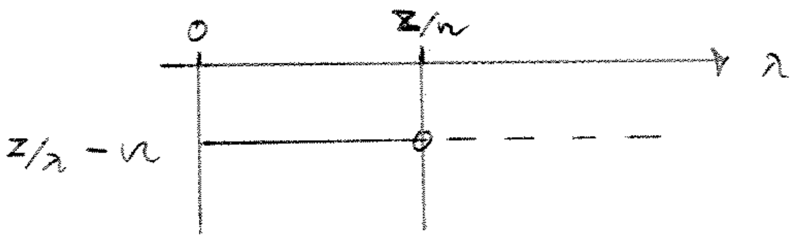
Et rimelighetsestimat $\hat{\tau}$ avhenger av observasjonen x, dvs $\hat{\tau} = \phi(x)$. En rimelighetsestimator for τ er da gitt ved $\phi(X)$. Definisjon: **rimelighetsestir**

Rimelighetsmetoden vil oftest være vanskeligere å gjennomføre enn momentmetoden. Spesielt er dette tilfellet når $\Theta \subset \mathbb{R}^k$ og dimensjonen k er stor.

Advarsel: I litteraturen brukes noen ganger $\hat{\theta}$ også som betegnelse på rimelighetsestimatoren.

Eksempel 14.3.1 (Rimelighetsestimator) La $X_1, \ldots, X_n$ være et tilfeldig utvalg fra tettheten

$$f(x; \lambda) = \frac{\lambda^x}{x!} e^{-\lambda} \qquad x = 0, 1, 2, \ldots$$



Figur 14.2: Fortegnsskjema

hvor $\lambda \geq 0$. Dette betyr spesielt at vi antar $X_i \sim \text{Poisson}(\lambda)$. Rimelighetsfunksjonen er

$$\begin{aligned} L(\lambda) = f_X(x;\lambda) &\overset{uavh.}{=} f_{X_1}(x_1;\lambda)\cdots f_{X_n}(x_n;\lambda) = \frac{\lambda^{x_1}}{x_1!}e^{-\lambda}\cdots\frac{\lambda^{x_n}}{x_n!}e^{-\lambda} \\ &= \frac{1}{x_1!\cdots x_n!}\cdot\lambda^z e^{-\lambda n} \quad \text{hvor } z = x_1+\cdots+x_n. \end{aligned}$$

Merk: Regningen over viser også at $Z = X_1 + \cdots + X_n$ er en tilstrekkelig observator.

Rimelighetsestimatet kan nå finnes ved å maksimalisere L. Fordi ln er strengt voksende er dette ekvivalent med å maksimalisere $\ln L$. Pga faktoriseringen over ser vi at det er tilstrekkelig å maksimalisere $\ln(\lambda^z e^{-\lambda n})$:

$$0 = \frac{d}{d\lambda}(z\ln\lambda - \lambda n) = z/\lambda - n \;\text{ gir } \lambda = z/n.$$

Fortegnsskjemaet for den deriverte i figur 14.2 gir at det kritiske punktet z/n er et maksimalpunkt. Konklusjonen er at rimelighetsestimatoren faller sammen med utvalgets forventning

$$\overline{X} = Z/n.$$

Følgende observasjoner er viktige:

1. Det kan lønne seg å maksimalisere $\ln L$.
2. Det kan lønne seg å faktorisere L i en θ-avhengig og en θ-uavhengig del. Da finnes også en eventuell tilstrekkelig observator.
3. Dersom T er en tilstrekkelig observator, så vil rimelighetsestimatoren automatisk være basert på T.

Eksempel 14.3.2 (Rimelighetsestimator for μ, σ^2) La $X_1, \ldots, X_n$ være et tilfeldig utvalg fra en $N(\mu, \sigma^2)$-fordeling. Her er $\Theta = \{(\mu,\sigma^2) \mid \mu \in \mathbb{R},\ \sigma > 0\}$. Uavhengigheten gir rimelighetsfunksjonen

$$L(\mu,\sigma^2) \overset{uavh}{=} \left(\frac{1}{\sqrt{2\pi}}\right)\left(\frac{1}{\sigma^2}\right)^{n/2} e^{-\frac{n}{2\sigma^2}(\overline{x^2}-2\mu\overline{x}+\mu^2)}.$$

Vi må maksimalisere

$$u(\mu,\sigma^2) = \ln\left(\left(\frac{1}{\sigma^2}\right)^{n/2} e^{-\frac{n}{2\sigma^2}(\overline{x^2}-2\mu\overline{x}+\mu^2)}\right) = -\ln(\sigma^2)\ n/2 - \frac{n}{2\sigma^2}(\overline{x^2}-2\mu\overline{x}+\mu^2).$$

Dette er en funksjon av to variable, og en metode er å løse ligningssystemet

$$\frac{\partial u}{\partial\mu} = 0, \quad \frac{\partial u}{\partial\sigma^2} = 0.$$

I vårt spesielle tilfelle er det imidlertid klart (hvorfor?) at maksimalpunktet må være slik at $\overline{x^2} - 2\mu\overline{x} + \mu^2$ er minimalisert. Regningen

$$\overline{x^2} - 2\mu\overline{x} + \mu^2 = \overline{x^2} - \overline{x}^2 + (\overline{x}-\mu)^2$$

gir at minimalpunktet er $\mu = \overline{x}$. Oppgaven er dermed redusert til å minimalisere

$$-\ln(\sigma^2)\ n/2 - \frac{n}{2\sigma^2}(\overline{x^2} - \overline{x}^2)$$

Ved $t = \sigma^2$ og derivasjon følger

$$\begin{aligned} 0 &= \frac{d}{dt}\Big(-\ln(t)\ n/2 - \frac{n}{2t}(\overline{x^2}-\overline{x}^2)\Big) = -1/t\cdot n/2 + t^{-2}\frac{n}{2}(\overline{x^2}-\overline{x}^2) \qquad |\cdot t^2 \\ 0 &= -tn/2 + \frac{n}{2}(\overline{x^2}-\overline{x}^2) \end{aligned}$$

Det kritiske punktet er dermed $\sigma^2 = \overline{x^2} - \overline{x}^2$ og et fortegnsskjema gir at dette er et maksimalpunkt. Ved sammenligning med beregningsformelen for utvalgets varians S^2 kan vi konkludere at rimelighetsestimatoren for (μ,σ^2) er

$$W = \left(\overline{X}, \frac{n-1}{n} S^2\right)$$

14.4 Oppgaver

181 En kriminolog søker igjennom FBI arkiver for å undersøke forekomsten av en sjelden type fingeravtrykk. Antallet forekomster i 6 arkiver, hver med 100 000 fingeravtrykk, ble funnet til å være hhv 3, 0, 3, 4, 2, 1. Estimer den relevante parameteren med momentmetoden. Er estimatoren forventningsrett? Begrunn valget av statistisk modell.

182 Finn momentestimatorene for μ, σ^2 basert på et tilfeldig utvalg av størrelse n fra $N(\mu, \sigma^2)$-fordelingen. Er estimatoren for μ konsistent?

183 Vis at $\overline{X^k}$ er en momentestimator for $\mathrm{E}\, Y^k$ når $X_1, \ldots, X_n$ er et tilfeldig utvalg fra fordelingen til Y. En del av oppgaven er å spesifisere eventuelle forutsetninger.

184 La $y_1, \ldots, y_n$ være resultatet av en trekning fra en Bernoulli(p)-fordeling. Finn rimelighetsestimatet $\hat{p}$ for parameteren p. Er rimelighetsestimatoren for variansen forventningsrett?

185 En ung student fører statistikk over hvor mange kvinner han må invitere på en forestilling på kino i forhold til hvor mange som aksepterer. Fem kinobesøk var resultatet av å måtte invitere hhv 3, 6, 4, 2, og 9 kvinner. Gi en begrunnelse for at statistikken kan sees som resultatet av en trekning fra en geometrisk fordeling. Kommenter antagelsene i modellen. Estimer forventningsverdien og variansen til fordelingen ved utvalgets varians, ved momentmetoden, og ved rimelighetsmetoden.

186 La T være en tilstrekkelig observator for parameteren θ i forhold til observatoren X. Vis at rimelighetsestimatoren for θ er basert på T. Er estimatoren basert på X?

187 I løpet av en 8 dagers periode er avgangstidene til en buss observert til å være 5:15, 5:21, 5:14, 5:23, 5:29, 5:17, 5:15, 5:18. Anta at avgangstiden er uniformt fordelt i intervallet $[\theta_1, \theta_2]$. Estimer θ_1 og θ_2. Beregn også rimelighetsestimatet for den forventede avgangstiden.

188 La $Y_1, \ldots, Y_n$ være et tilfeldig utvalg fra tettheten

$$f(y \,|\, \theta) = (\theta + 1) y^\theta \qquad 0 < y < 1$$

Finn momentestimatoren og rimelighetsestimatoren for θ. Beregn de tilhørende estimatene når resultatet av en trekning er 0.5, 0.4 og 0.9.

Beregn også estimatorer og estimat for fordelingens forventning og varians ved de to metodene.

189 La $f(x) = 1/2$ for $x = \theta$ og la $g(x) = 1$ for $\theta < x < \theta + 1/2$. Definer $f(x), g(x)$ lik 0 for andre x verdier. En fordeling Q for tallinjen er da definert ved

$$Q(A) = \sum_{x \in A} f(x) + \int_A g(x) dx.$$

Vis at fordelingens forventnig er $\mu = \theta + 1/8$. Tegn grafen til fordelingens fordelingsfunksjon F.

Estimer θ ved momentmetoden basert på verdier fra en tilfeldig trekning fra fordelingen.

Fordelingen Q har en diskret del og en kontinuerlig del. Approksimer den kontinuerlige delen med g_n gitt ved

$$g_n(x) = 1/(2n) \qquad x = \theta + 1/(2n), \theta + 2/(2n), \ldots, \theta + 1/2.$$

Bruk denne approksimasjonen til å finne en rimelighetsestimator for θ basert på et tilfeldig utvalg. Hva skjer i grensen $n \to \infty$?

190 Hensikten med denne oppgaven er å gi en metode, betingingsmetoden, som med utgangspunkt i en forventningsrett estimator gir en forventningsrett estimator basert på en tilstrekkelig observator. I noen tilfeller gir metoden en beste estimator. Metoden virker generelt, men vi ser kun på det diskrete tilfellet.

La $\{P^\theta\}$ være en statistisk modell hvor alle fordelingene P^θ er diskrete. Anta at W er en forventningsrett estimator for τ.

Vis at dersom T er en tilstrekkelig observator for θ, så er $\mathrm{E}^\theta(W \mid T = t)$ et estimat for τ basert på observasjonen t.

Vis at estimatoren over er forventningsrett for τ.

La $X_1, \ldots, X_n$ være et tilfeldig utvalg fra en Bernoulli(p)-fordeling. Vis at X_1 er en forventningsrett estimator for p. Vis at $T = X_1 + \ldots + X_n$ er en tilstrekkelig observator for p. Bruk betingingsmetoden til å finne en forventningsrett estimator basert på T for p.

191 La $X_1, \ldots, X_n$ være et tilfeldig utvalg fra uniform$(0, \theta)$-fordelingen. Vis at X_{maks} er en tilstrekkelig observator for θ.

Dersom $X_{\text{maks}} = t$, så er $X_1 = t$ med sannsynlighet $1/n$ og ellers uniformt fordelt over $(0, t)$ med den gjenværende sannsynligheten $(n-1)/n$. Begrunn dette.

Benytt den forventningsrette estimatoren $2X_1$ til å finne en forventningsrett estimator for θ basert på T ved hjelp av betingingsmetoden.

Kapittel 15

Intervallestimering og konfidensintervall

The earliest example of confidence intervals appears to occur in the work of Laplace (1812), who points out how an (approximate) probability statement concerning the difference between an observed frequency and a binomial probability p can be inverted to obtain an associated interval for p. Other examples can be found in the work of Gauss (1816), Fourier (1826), and Lexis (1875). However, in all these cases, although the statements made are formally correct, the authors appear to consider the parameter as the variable which with the stated probability falls in the fixed confidence interval. The proper interpretation seems to have been pointed out for the first time by E. B. Wilson (1927). About the same time two examples of exact confidence statements were given by Working and Hotelling (1929) and Hotelling (1931). A general method for obtaining exact confidence bounds for a real-valued parameter in a continuous distribution was proposed by Fisher (1930), who however later disavowed this interpretation of his work.

E. L. Lehmann (1986)

Fremfor å gjette på verdien til en ukjent størrelse kan det være hensiktsmessig å gjette på et område som inneholder den ukjente størrelsen. Når denne gjetningen skal basere seg på en observator, dvs være en funksjon av en observator, så ender vi opp med at gjetningen må være gitt av en tilfeldig mengde. Hvert utfall i den tenkte gjentagelsen av eksperimentet gir en mengde, dvs et områdeestimat for estimanden.

Dersom områdeestimatet inneholder estimanden i minst 95% av gjentagelsene i det lange løp, så sies områdeestimatoren å være en 95% områdeestimator. Ekvivalent sier vi at estimatoren har et konfidensnivå lik 95%. Når nivået er gitt er det ønskelig at det beregnede områdeestimatet er så lite som mulig.

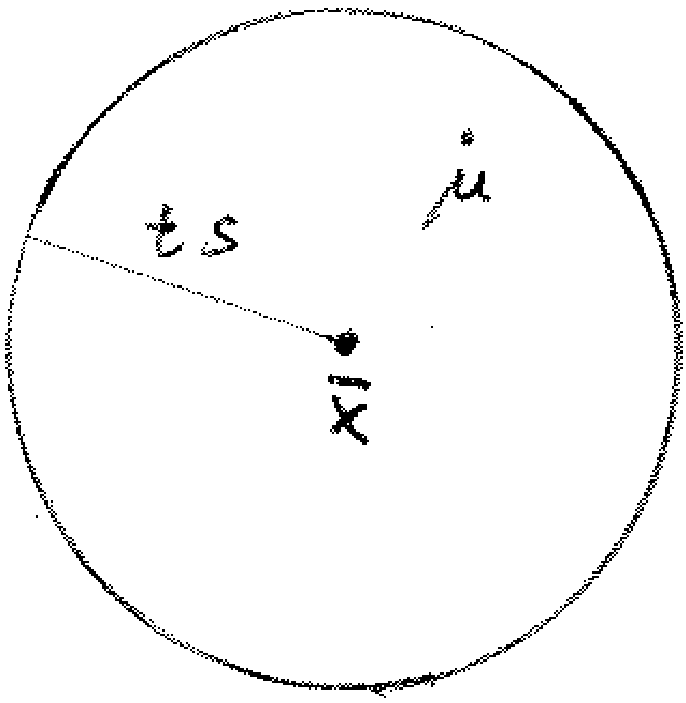
Vi vil se på flere metoder for å utlede og beregne områdeestimatorer. De fleste metodene gir et områdeestimat som et intervall, dvs ved en intervallestimator.

KJENT: statistisk modell, observator, estimator, estimand, tilfeldig utvalg, Tsjebysjevs ulikhet, normalfordeling, Students t-fordeling

NYTT: områdeestimator, tilfeldig mengde, tilfeldig intervall, intervallestimator, konfidensmengde, konfidensnivå, konfidenskoeffisient, ensidig konfidensintervall, intervall-invertering, pivot observator

15.1 Område- og intervallestimatorer

La $\{P^\theta\}$, $\theta \in \Theta$, være en statistisk modell for utfallsrommet Ω. En punktestimator W for en estimand $\tau : \Theta \to \Theta_\tau$ er en observator $W : \Omega \to \Omega_W$ hvor $\Theta_\tau \subset \Omega_W$.



Figur 15.1: Mengdeestimat for $\mu = (\mu_1, \mu_2)$

Punktestimatoren gir en gjetning på verdien til estimanden med utgangspunkt i observasjonene. En områdeestimator gir en gjetning på et område, dvs en mengde, som skal inneholde verdien til estimanden. Vi velger følgende definisjon:

Definisjon: $\Omega_{(V)}$

La $\Omega_{(V)}$ være en mengde. Anta at $V(\omega) \subset \Omega_{(V)}$ for alle utfall ω. Da er V en områdeestimator i $\Omega_{(V)}$ dersom

$$(x \in V) \stackrel{def}{=} \{\omega \,|\, x \in V(\omega)\}$$

er en hendelse for alle punkter x i $\Omega_{(V)}$.

Hvert utfall ω gir en mengde $V(\omega)$ inneholdt i $\Omega_{(V)}$. Usikkerheten i ω gir en tilsvarende usikkerhet for $V(\omega)$. Dette betyr at V kan oppfattes som en tilfeldig mengde. Antagelsen "($x \in V$) er en hendelse" i definisjonen sikrer at sannsynligheten $P(x \in V)$ for at V inneholder et fiksert punkt x er definert.

Definisjon: $\Theta_\tau \subset \Omega_{(V)}$

La τ være en estimand med verdier i Θ_τ. En områdeestimator V i $\Omega_{(V)}$ er en områdeestimator for τ dersom $\Theta_\tau \subset \Omega_{(V)}$.

Når eksperimentet er gjort, så brukes V til å regne ut områdeestimatet $v = V(\omega)$ for τ. Konklusjonen er at vi gjetter på at $\tau \in v$.

Eksempel 15.1.1 (Mengdeestimator) La $X_1, \ldots, X_n$ være et tilfeldig utvalg fra en fordeling. En områdeestimator V for fordelingens forventningsverdi μ er gitt av

$$V = \{y \,|\, |y - \overline{X}| \leq tS\},$$

hvor $\overline{X}$ og S^2 er utvalgets forventning og varians, og t er et tall. Et mulig mengdeestimat er vist i figur 15.1.

Definisjon: $V = [X, Y]$

La V være en områdeestimator. Dersom $V(\omega)$ er et lukket intervall for alle ω, så er V en intervallestimator.

Vi beviser at et tilfeldig intervall $V = [X, Y]$, der X, Y er tilfeldige variable, er en intervallestimator. Dette beviset ved:

$$(w \in V) = (w \in [X, Y]) = (X \leq w \leq Y) = (X \leq w \text{ og } w \leq Y) = (X \leq w) \cap (w \leq Y),$$

som er en hendelse når X, Y er tilfeldige variable.

Metodene vi ser på tar som utgangspunkt at vi kan se på dataene som resultatet av et statistisk eksperiment. Grovt sett vurderes en metode til å være god dersom den gir et godt resultat i de fleste av de tenkte gjentagelsene av eksperimentet. F eks er en punktestimator god dersom den gir et estimat i nærheten av estimanden i de fleste av eksperimentene. Tilsvarende er en områdeestimator god dersom den gir et estimert området som inneholder estimanden i de fleste av eksperimentene. Dette motiverer en definisjon.

Definisjon: **konfidensmengde**

La α være et reelt tall. En områdeestimator V for en estimand τ er en konfidensmengde for τ med konfidensnivå α dersom

$$P(\tau \in V) \geq \alpha.$$

I så fall er V en $\alpha 100\%$ områdeestimator for τ.

I det overstående krever vi ikke likhet fordi venstresiden avhenger av parameteren via P og τ, mens konfidensnivået α er uavhengig av parameteren.

Ved gjentagelse av et statistisk eksperiment vil et $\alpha 100\%$ områdeestimat inneholde estimanden i minst $\alpha 100\%$ av tilfellene i det lange løp.

Det er ønskelig at konfidensnivået er så høyt som mulig. Det vanlige er å sammenligne områdeestimatorer med et gitt nivå, og blant disse foretrekkes den som gir det minste området, hvis en slik finnes.

La $\alpha' \leq \alpha$. Da vil et konfidensområde V med nivå α også være et konfidensområde med nivå α' fordi

$$P(\tau \in V) \geq \alpha \geq \alpha'.$$

Dette betyr at det svarer uendelig mange konfidensnivå til et konfidensområde. Dette motiverer en definisjon.

Definisjon: **konfidenskoeffisient**

Konfidenskoeffisienten α til en områdeestimator V for en estimand τ er den største nedre skranken for mengden

$$\{P^{\theta}(\tau(\theta) \in V) \mid \theta \in \Theta\}.$$

Definisjonen betyr spesielt at V har et konfidensnivå α dersom α er mindre eller lik konfidenskoeffisienten. Intuitivt er konfidenskoeffisienten det beste konfidensnivået til områdeestimatoren. Til en områdeestimator finnes det mange konfidensnivå, men kun én konfidenskoeffisient. I praksis kan ofte konfidenskoeffisienten finnes som minimalverdien til funksjonen $g : \Theta \to \mathbb{R}$, som er definert ved

$$g(\theta) = P^{\theta}(\tau(\theta) \in V).$$

Advarsel: Konvensjonene for hva som menes med konfidensnivå og konfidenskoeffisient varierer fra forfatter til forfatter.

15.2 Metoder for å finne konfidensintervall

La W være en punktestimator for τ. En intuitivt rimelig intervallestimator for τ er da gitt ved

$$V = [W - X, W + Y],$$

hvor X og Y er tilfeldige variable som tar positive verdier. En mulighet er å la $X = Y = a$. Konstanten a kan bestemmes av det ønskede konfidensnivået. Tsjebysjevs ulikhet gir en variant av dette.

Eksempel 15.2.1 (Konfidensområde fra Tsjebysjevs ulikhet) La W være en forventningsrett estimator for τ. Anta at $s > 0$ er en konstant slik at

$$\mathrm{Var}^{\theta} W \leq s^2$$

for alle θ. Da gir Tsjebysjevs ulikhet at

$$P(|W - \tau| < ts) > 1 - \frac{\mathrm{Var}\, W}{t^2 s^2} \geq 1 - \frac{s^2}{t^2 s^2} = 1 - \frac{1}{t^2}.$$

Konklusjonen er at

$$V = \{w \,|\, |W - w| \leq ts\}$$

er en $(1 - \frac{1}{t^2})100\%$ områdeestimator for τ. Dette følger fordi

$$(\tau \in V) = (|W - \tau| \leq ts).$$

Spesielt finnes et 90% konfidensområde ved å velge t slik at $1 - \frac{1}{t^2} = 0,90$.

Vi ønsker å få et områdeestimat som er så lite som mulig, dvs vi må velge s så liten som mulig under betingelsen $\mathrm{Var}^{\theta} W \leq s^2$. Spesielt ser vi at dersom W er en nøyaktig estimator, dvs dersom variansen er liten, så vil metoden gi et lite områdeestimat.

Utledningen over gjelder uansett hvilken fordeling vi har. Når fordelingstypen er gitt kan vi bruke bedre ulikheter enn Tsjebysjevs ulikhet.

I noen tilfeller kan et konfidensintervall finnes enklere ved å finne to ensidige konfidensintervall. Anta at

$$P(W_- \leq \tau) \geq \alpha_-, \qquad P(W_+ \geq \tau) \geq \alpha_+, \qquad P(W_- \leq W_+) = 1.$$

Da følger det at

$$P(W_- \leq \tau \leq W_+) = 1 - P(W_- \geq \tau) - P(W_+ \leq \tau) \geq \alpha_- + \alpha_+ - 1,$$

dvs $[W_-, W_+]$ er et konfidensintervall med nivå $\alpha_- + \alpha_+ - 1$.

Definisjon: **intervall-invertering**

I det følgende vil vi beskrive metoden med intervall-invertering. Metoden gir alltid en intervallestimator med ønsket nivå med utgangspunkt i en tilfeldig variabel.

La X være en tilfeldig variabel. I noen tilfeller kan en finne en $a(\theta)$ og en $b(\theta)$ slik at

$$P(a \leq X \leq b) \geq 1 - \alpha,$$

ved å finne $a \leq b$ slik at

$$P(a > X) = P(X > b) \leq \alpha/2.$$

Dersom X har en kontinuerlig fordeling, så finnes alltid en unik a og en unik b ved å kreve likhet. Tallet b er da ikke annet enn $\alpha/2$ kvantilen til fordelingen til X. Dette er illustrert i figur 15.2. Generelt finnes alltid en unik største a og en unik minste b slik at ulikhetene holder. Fordelingen til X avhenger av parameteren θ, så a og b må nødvendigvis avhenge av θ.

Den doble ulikheten

$$a(\theta) \leq x \leq b(\theta)$$

gir generelt et områdeestimat

$$v = \{\theta \,|\, a(\theta) \leq x \leq b(\theta)\}.$$

Dersom a, b er valgt slik som over, så følger det at den tilsvarende estimatoren V har konfidensnivå lik $1 - \alpha$ fordi

$$(\theta \in V) = (a(\theta) \leq X \leq b(\theta)).$$

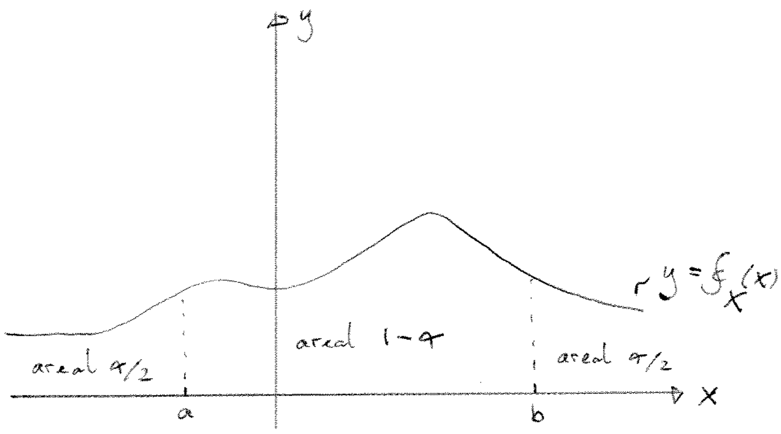
Dersom $\Theta \subset \mathbb{R}$ finnes også et unikt minste intervallestimat som inneholder v.

En ulempe med den foregående metoden er at den generelt resulterer i mye regnearbeid fordi fordelingen til X avhenger av θ. Denne observasjonen motiverer en definisjon.

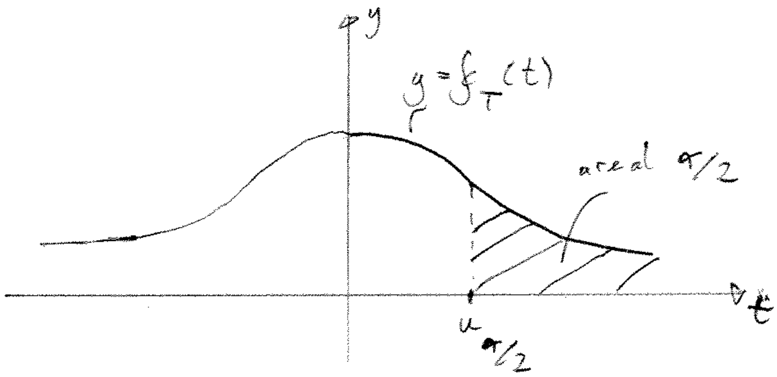
Definisjon: T^{θ}

La $\{P^{\theta}\}$, $\theta \in \Theta$, være en statistisk modell. En pivot observator T er en funksjon som gir en observator T^{θ} for hver parameterverdi θ slik at fordelingen til T^{θ} er uavhengig av θ.

Det neste eksemplet viser at metoden med intervall-invertering kan forenkles ved bruk av en pivot variabel.



Figur 15.2: $P(a \leq X \leq b) = 1 - \alpha$ når $P(a > X) = P(X > b) = \alpha/2$.



Figur 15.3: Kvantilen $u_{\frac{\alpha}{2}}$ til standard normalfordelingen.

Eksempel 15.2.2 (Normalfordeling med kjent varians) La $X_1, \ldots, X_n$ være et tilfeldig utvalg fra $N(\mu, \sigma^2)$-fordelingen med σ kjent. Dette betyr at $\Theta = \{\mu \mid \mu \in \mathbb{R}\}$. Da er μ den eneste parameteren.

Vi vet at $\overline{X}$ er en forventningsrett estimator for μ. Videre er $\operatorname{Var} \overline{X} = \frac{1}{n}\sigma^2$. Fordi en sum av normalfordelte variable er normalfordelt følger det at $\overline{X} \sim N(\mu, \frac{1}{n}\sigma^2)$. Dette betyr at

$$T^\mu := \frac{(\overline{X} - \mu)\sqrt{n}}{\sigma} \sim N(0, 1),$$

dvs T er en pivot variabel. Dette gir

$$P(T > u_{\frac{\alpha}{2}}) = \frac{\alpha}{2},$$

hvor $\frac{\alpha}{2}$ kvantilen $u_{\frac{\alpha}{2}}$ til standard normalfordelingen kan finnes i tabell. Fra figur 15.3 følger det at

$$P(|T| \leq u_{\frac{\alpha}{2}}) = 1 - \alpha.$$

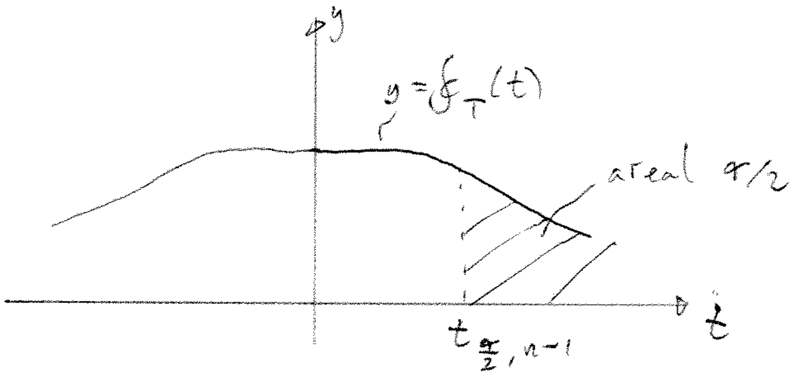
Vi har følgende ekvivalente ulikheter

$$|T| \leq u_{\frac{\alpha}{2}}, \quad \left| \frac{(\overline{X} - \mu)\sqrt{n}}{\sigma} \right| \leq u_{\frac{\alpha}{2}}, \quad |\overline{X} - \mu| \leq \frac{u_{\frac{\alpha}{2}}\sigma}{\sqrt{n}}, \quad \overline{X} - \frac{u_{\frac{\alpha}{2}}\sigma}{\sqrt{n}} \leq \mu \leq \overline{X} + \frac{u_{\frac{\alpha}{2}}\sigma}{\sqrt{n}}.$$

Konklusjonen er dermed at

$$\boxed{\left[\overline{\mathrm{X}} - \frac{\mathrm{u}_{\frac{\alpha}{2}}\sigma}{\sqrt{\mathrm{n}}}, \ \overline{\mathrm{X}} + \frac{\mathrm{u}_{\frac{\alpha}{2}}\sigma}{\sqrt{\mathrm{n}}}\right]}$$

er en $(1 - \alpha)100\%$ intervallestimator for μ.



Figur 15.4: Kvantilen $t_{\frac{\alpha}{2},n-1}$ til Student-t$(n-1)$-fordelingen.

Eksempel 15.2.3 (Intervallestimator for μ når σ^2 er ukjent) La $X_1, \ldots, X_n$ være et tilfeldig utvalg fra N(μ, σ^2)-fordelingen. Dette betyr at

$$\Theta = \{(\mu, \sigma^2) \,|\, \mu \in \mathbb{R}, \sigma > 0\}.$$

Vi skal finne et $(1-\alpha)100\%$-konfidensintervall for μ. Utgangspunktet er at (godta dette!)

$$\frac{(\overline{X}-\mu)\sqrt{n}}{S} \sim \text{Student-t}(n-1).$$

Dette gir en pivot variabel som kun avhenger av parameteren (μ, σ^2) gjennom estimanden μ! Deretter kan vi gjøre som i forrige eksempel:

$$P(\frac{(\overline{X}-\mu)\sqrt{n}}{S} > t_{\frac{\alpha}{2},n-1}) = \frac{\alpha}{2},$$

hvor $\frac{\alpha}{2}$ kvantilen $t_{\frac{\alpha}{2}}$ til Student-t$(n-1)$-fordelingen kan finnes i tabell. Fra figur 15.4 følger det at

$$P(\left|\frac{(\overline{X}-\mu)\sqrt{n}}{S}\right| \leq t_{\frac{\alpha}{2},n-1}) = 1-\alpha.$$

Vi har følgende ekvivalente relasjoner

$$\left|\frac{(\overline{X}-\mu)\sqrt{n}}{S}\right| \leq t_{\frac{\alpha}{2},n-1}, \quad |\overline{X}-\mu| \leq \frac{t_{\frac{\alpha}{2},n-1}S}{\sqrt{n}}, \quad \overline{X} - \frac{t_{\frac{\alpha}{2},n-1}S}{\sqrt{n}} \leq \mu \leq \overline{X} + \frac{t_{\frac{\alpha}{2},n-1}S}{\sqrt{n}}.$$

Vi konkluderer at

$$\boxed{[\overline{X} - \frac{t_{\frac{\alpha}{2},n-1}S}{\sqrt{n}}, \overline{X} + \frac{t_{\frac{\alpha}{2},n-1}S}{\sqrt{n}}]}$$

er en $(1-\alpha)100\%$ intervallestimator for μ.

Når dataene er gitt, kan vi beregne intervallestimatet

$$[\overline{x} - \frac{t_{\frac{\alpha}{2},n-1}s}{\sqrt{n}}, \overline{x} + \frac{t_{\frac{\alpha}{2},n-1}s}{\sqrt{n}}].$$

Dette estimatet vil oftest være et større intervall enn det vi fikk når σ var kjent. Dette er ikke uventet fordi både μ og σ^2 var ukjente i dette siste tilfellet.

Student-t$(n-1)$-fordelingen ligner på normalfordelingen, men grafen til tettheten er litt flatere. Dette gir som resultat at $t_{\frac{\alpha}{2},n-1} > u_{\frac{\alpha}{2}}$. Dette gir en alternativ forklaring på hvorfor estimatet med ukjent varians oftest er større. Når n vokser vil Student-t$(n-1)$-fordelingen nærme seg $N(0,1)$-fordelingen. Spesielt vil $\lim t_{\frac{\alpha}{2},n-1} = u_{\frac{\alpha}{2}}$.

15.3 Oppgaver

192 Gi et talleksempel som viser at intervallestimatet beregnet ved antatt ukjent varians for forventningsverdien i en normalfordeling kan bli smalere enn intervallestimatet beregnet ved antatt kjent varians.

193 Hva tror du er definisjonen av hva som menes med et ensidig konfidensintervall? Vis hvordan to ensidige konfidensintervall kan gi et konfidensintervall med ønsket nivå.

194 Vi har tidligere gitt en generell definisjon av hva en estimator er. Gi en begrunnelse for at en intervallestimator V er en estimator ved å identifisere et utfallsrom Ω_V slik at V tar verdier i Ω_V. Identifiser en hendelsesfamilie i Ω_V.

195 La $Y_1, \ldots, Y_n$ være et tilfeldig utvalg fra uniform$(0, \theta)$-fordelingen. Finn en 90% intervall-estimator ved å bruke metoden med intervall-invertering på Y_{maks}. Bruk også metoden på $2Y_1$, og kommenter resultatet.

196 Vis at $(\tau \in V)$ er en hendelse når V er en områdeestimator for estimanden τ. Avhenger denne hendelsen av parameteren i modellen?

197 Gi en begrunnelse for kravet $\Theta_\tau \subset \Omega_{(V)}$ i definisjonen av at V er en områdeestimator for τ.

198 La $X \sim \mathrm{N}(\mu, \sigma^2)$. Finn h slik at $P(\mu - h < X < \mu + h) = 0.50$.

199 La $Y_1, \ldots, Y_n$ være et tilfeldig utvalg fra $\mathrm{N}(\mu, \sigma^2)$-fordelingen. Presiser hvordan μ kan oppfattes som en estimand i denne modellen.

200 Bevis at

$$V = \{y \mid |y - \overline{X}| \leq tS\}$$

gir en områdeestimator V for en fordelings forventningsverdi μ. Her er V basert på et tilfeldig utvalg fra fordelingen.

Finn konfidenskoeffisienten til V når fordelingen er bernoullifordelingen med parameter p.

201 Finn $t_{0.15,24}$-kvantilen til Student-t-fordelingen. Sammenlign med den tilsvarende kvantilen for normalfordelingen.

202 Beregn $P(-0.883 < T < 1.383)$ når T er Student-t fordelt med 9 frihetsgrader. Hvordan er fordelingen til T definert?

203 De ansatte ved 15 sykehus deltok i en undersøkelse vedrørende bieffekter til en gitt medisinering. Ved de 15 sykehusene var resultatet at hhv 5.8, 5.3, 4.5, 3.9, 4.6, 5.4, 7.9, 8.2, 6.9, 5.7, 4.6, 6.3, 8.4, 4.6, 7.3 prosent av pasientene fikk bieffekter etter en slik medisinering.

Beregn et 95% konfidensintervall for den virkelige middelverdien representert av disse tallene. Spesifiser den statistiske modellen med en begrunnelse.

Beregn et 95% konfidensintervall ved hjelp av Tsjebysjevs ulikhet ved å anta at variansen er begrenset av 2 ganger utvalgets varians.

204 La $\overline{X}$ være forventningsverdien til et tilfeldig utvalg fra $\mathrm{N}(\mu, \sigma^2)$-fordelingen. Beregn konfidenskoeffisienten til intervallestimatoren

$$V = [\overline{X} - \frac{u_{\frac{\alpha}{2}}\sigma}{\sqrt{n}}, \overline{X} + \frac{u_{\frac{\alpha}{2}}\sigma}{\sqrt{n}}]$$

for forventningen μ når variansen σ er kjent.

Er V en intervallestimator når σ er ukjent?

205 Gjør oppgave 7.6.14, s. 346, i Larsen-Marx.

206 La $Y_1, \ldots, Y_n$ være et tilfeldig utvalg fra $\mathrm{N}(\mu, \sigma^2)$-fordelingen hvor σ^2 er kjent. Utled en $(1 - \alpha) \cdot 100\%$ intervallestimator for μ. Hvor stor må n være for at lengden til konfidensintervallet skal være mindre enn L?

207 Resultatene av et eksperiment er 4.3, 5.8, 8.4, 3.7, 5.2 og 5.1. Det antas at variansen er 1.50 og at tallene kan sees som en trekning fra en normalfordeling. Beregn et 90% og et 50% konfidensintervall for forventningsverdien til fordelingen.

208 La $Y_1, \ldots, Y_n$ være et tilfeldig utvalg fra uniform$(0, \theta)$-fordelingen. Finn et 90% konfidensintervall for θ basert på estimatoren $(n+1)Y_{\text{min}}$. Finn også et basert på Y_{maks}. Hvilken intervallestimator er best?

209 Kan det finnes informasjon i et punktestimat som "forsvinner" i et intervallestimat?

210 Utbrudd av matforgiftninger er ofte resultatet av bakterier i salater. I 1967 undersøkte helserådet i New York 220 tunfisk salater markedsført av ulike leverandører og utsalg. Totalt ble 179 funnet å være uegnet i følge deres kriterium. Finn et 95% konfidensintervall for p, dvs den virkelige andelen av dårlige tunfisk salater i New York. Gi en begrunnelse for antagelsene i modellen.

211 Anta at 3 uavhengige Bernoulli-forsøk resulterte i 1 suksess. Beregn et 95% konfidensintervall for suksess-sannsynligheten ved hjelp av en eksakt intervallestimator.

212 For et tilfeldig utvalg fra en $\mathrm{N}(\mu, \sigma^2)$-fordeling kan det vises at utvalgets forventning $\overline{Y}$ og utvalgets varians S^2 er uavhengige observatorer.

Bruk dette til å begrunne at

$$P(\frac{|\overline{Y} - \mu|}{\sigma}\sqrt{n} < z_{\alpha/2},\ \chi^2_{1-\alpha/2,n-1} < \frac{(n-1)S^2}{\sigma^2} < \chi^2_{\alpha/2,n-1}) = (1-\alpha)^2.$$

Benytt overstående til å finne et $100 \cdot (1-\alpha)$ konfidensområde i planet for parameteren $\theta = (\mu, \sigma^2)$.

213 Yrkespiloter risikerer å få svekket hørsel pga støy i cockpiten. Lydnivået målt for 18 kommersielle flygninger ble målt til å være hhv 74, 77, 80, 82, 82, 85, 80, 75, 75, 72, 90, 87, 73, 83, 86, 83, 83, 80 (dB). Finn et 95% konfidensintervall for forventningen μ og for variansen σ^2 ved å anta at støyen er normalfordelt. Finn et 95% konfidensområde for (μ, σ^2). Tegn en figur som gir området.

Kapittel 16

Konfidensintervall: Flere eksempler

KJENT: statistisk modell, konfidensintervall, konfidensnivå

NYTT: konfidensintervall for varians, approksimativt konfidensintervall

16.1 Oppsummering

Vi starter med en repetisjon av de viktigste begrepene. En intervallestimator V er gitt ved to tilfeldige variable A, B ved $V = [A, B]$. Dette er et α-konfidensintervall for τ dersom

$$P^{\theta}(A \leq \tau(\theta) \leq B) \geq \alpha \quad \text{for alle } \theta.$$

I så fall er α konfidensnivået til V. Konfidenskoeffisienten til V er den største α slik at at ligningen er oppfylt.

Vi vil peke på to kriterier for vurdering av konfidensintervall.

1. Vi ønsker til et gitt konfidensnivå å finne et konfidensintervall som er så lite som mulig.
2. Dersom T er en tilstrekkelig observator, så skal V være basert på T.

Merk: Dette siste punktet skal gjelde for alle tilstrekkelige observatorer. Det siste punktet er strengt tatt kun en gjentagelse av det generelle tilstrekkelighetsprinsippet som gjelder for enhver estimator, og mer generelt for enhver statistisk slutning.

Vi har ikke beskrevet noen metode som generelt gir et konfidensintervall for en generell estimand, men snarere pekt på noen muligheter.

Intuitivt er en god mulighet å se på konfidensintervall med utgangspunkt i et intervall sentrert om en forventningsrett estimator. Lengden til intervallet kan bestemmes av det ønskede konfidensnivået. Vi har sett at Tsjebysjevs ulikhet kan brukes i denne forbindelsen ved å velge intervall-lengden proposjonal med en øvre skranke for standardavviket til estimatoren. Spesielt ser vi da at en nøyaktig estimator gir opphav til et nøyaktig, dvs lite, konfidensintervall. Dette er mest av teoretisk interesse. I konkrete modeller kan en bruke bedre ulikheter som gir tilsvarende bedre intervallestimatorer.

La det være gitt at et $(1-\alpha)$-områdeestimat for parameteren θ skal finnes med utgangspunkt i en tilfeldig variabel X. Generelt kan vi da finne a og b slik at

$$P(a \leq X \leq b) \geq 1 - \alpha.$$

Dersom X har kontinuerlig fordeling kan til og med likhet oppnås ved å bestemme a og b ved

$$P(X < a) = P(X > b) = \alpha/2.$$

Her er det klart at a og b er funksjoner av parameteren θ. Et $(1-\alpha)$-områdeestimat for θ er da gitt ved

$$V = \{\theta \,|\, a(\theta) \leq X \leq b(\theta)\}.$$

Dette er en generelt anvendelig metode, men valget av variabelen X er ikke opplagt. Et konfidensområde for en generell estmand τ er gitt av

$$\tau(V) = \{\tau(\theta) \,|\, \theta \in V\},$$

men det er ikke gitt at dette er det beste valget. Denne metoden refererer vi til som intervall-inverteringsmetoden.

En tredje mulighet er gitt ved en variasjon av intervall-inverteringsmetoden. Den tilfeldige variabelen X erstattes av en pivotvariabel Y, dvs en familie $\{Y^\theta\}$ av tilfeldige variable slik at fordelingen til Y^θ er uavhengig av parameteren θ. Et $(1-\alpha)$-områdeestimat for θ er da gitt ved

$$V = \{\theta \,|\, a \leq Y^\theta \leq b\},$$

hvor a og b velges slik at

$$P(a \leq Y \leq b) = 1 - \alpha.$$

Poenget er her at fordelingen til Y er uavhengig av parameteren, så a og b kan velges uavhengig av parameteren. Et annet poeng er at en i noen tilfeller kan finne en pivotvariabel som avhenger av parameteren θ gjennom estimanden τ, dvs $Y^\theta = Z^{\tau(\theta)}$. I så fall sier vi at Z er en pivotvariabel for τ. Et $(1-\alpha)$-områdeestimat for τ er da gitt ved

$$V = \{t \,|\, a \leq Z^t \leq b\}.$$

16.2 Eksempler

Eksempel 16.2.1 (Intervall-invertering) La $Y_1, \ldots, Y_n$ være et tilfeldig utvalg fra uniform$(0,\theta)$-fordelingen. Da er $W = \frac{n+1}{n} Y_{maks}$ forventningsrett for θ. I tillegg er W en tilstrekkelig observator. Tettheten er

$$f_W(w) = \frac{n^{n+1}}{(n+1)^n \theta^n} w^{n-1}.$$

Vi vil finne et konfidensintervall med nivå $\alpha = 90\%$ ved å finne a, b slik at $P(a \leq W \leq b) = \alpha$.

Merk: $[a, b]$ er ikke det søkte konfidensintervallet!

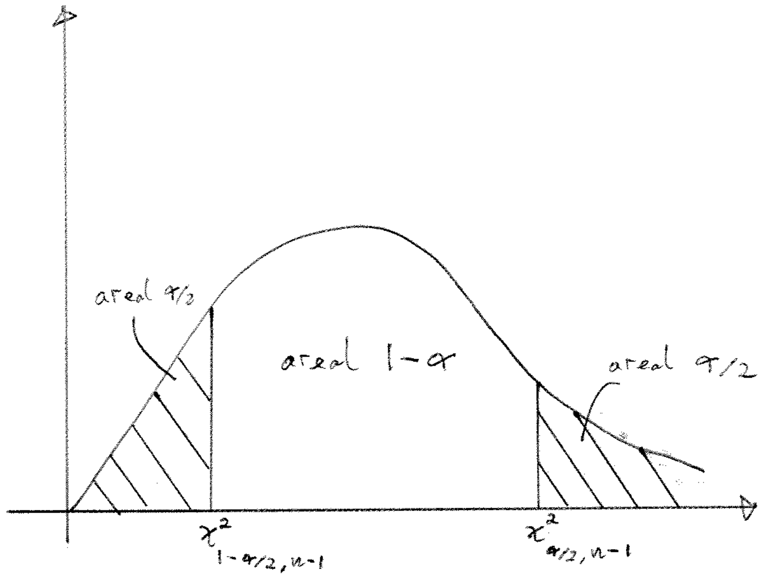
Vi forenkler litt ved å finne a, b slik at $P(W < a) = P(W > b) = (1-\alpha)/2 = 5\%$:

$$\begin{aligned} P(W < a) &= \int_0^a f_W(w)dw \\ &= \int_0^a \frac{n^{n+1}}{(n+1)^n\theta^n} w^{n-1} dw = \frac{n^{n+1}}{(n+1)^n\theta^n} \frac{a^n}{n} = 5\% \\ \Rightarrow a(\theta) = \sqrt[n]{5\%}\frac{n}{n+1}\theta & \end{aligned}$$

$$\begin{aligned} P(W > b) &= 1 - \int_0^b f_W(w)dw = 5\% \\ \Rightarrow b(\theta) = \sqrt[n]{95\%}\frac{n}{n+1}\theta & \end{aligned}$$

Vi inverterer intervallet, dvs løser den dobble ulikheten mhp parameteren θ

$$\begin{aligned} a \leq W \leq b \\ \sqrt[n]{5\%}\frac{n}{n+1}\theta \leq W \leq \sqrt[n]{95\%}\frac{n}{n+1}\theta \\ \frac{n}{n+1}\frac{W}{\sqrt[n]{95\%}} \leq \theta \leq \frac{n}{n+1}\frac{W}{\sqrt[n]{5\%}}. \end{aligned}$$



Figur 16.1: Kvantiler i khi-kvadrat fordelingen.

Konklusjonen er at

$$V = [\frac{n}{n+1}\frac{W}{\sqrt[n]{95\%}}, \frac{n}{n+1}\frac{W}{\sqrt[n]{5\%}}]$$

er et konfidensintervall med nivå 90%.

Merk: V er en estimator. Når vi setter inn verdier fra et utført eksperiment, så finnes intervallestimatet

$$v = [\frac{n}{n+1}\frac{w}{\sqrt[n]{95\%}}, \frac{n}{n+1}\frac{w}{\sqrt[n]{5\%}}].$$

Eksempel 16.2.2 (Pivot variabel) La $Y_1, \ldots, Y_n$ være et tilfeldig utvalg fra $N(\mu, \sigma^2)$-fordelingen. Vi vil finne et $(1-\alpha)$-konfidensintervall for σ^2. I følge tilstrekkelighetsprinsippet vil en god estimator nødvendigvis være en funksjon av $\overline{Y}$ og S^2 fordi $T = (\overline{Y}, S^2)$ er en tilstrekkelig observator. Her er

$$\frac{(n-1)S^2}{\sigma^2}$$

en pivotvariabel for σ^2. Fordelingen er uavhengig av (μ, σ^2) og gitt ved

$$\boxed{\frac{(n-1)S^2}{\sigma^2} \sim \chi^2{}_{n-1}.}$$

Beviset av dette utelates her. Dette gir

$$P\Big(\frac{(n-1)S^2}{\sigma^2} \geq \chi^2{}_{\frac{\alpha}{2},n-1}\Big) = \frac{\alpha}{2}.$$

Kvantilen $\chi^2{}_{\frac{\alpha}{2},n-1}$ til khi-kvadrat fordelingen finnes i tabell. Fra figur 16.1 følger det at

$$P\Big(\chi^2{}_{1-\frac{\alpha}{2},n-1} \leq \frac{(n-1)S^2}{\sigma^2} \leq \chi^2{}_{\frac{\alpha}{2},n-1}\Big) = 1-\alpha$$

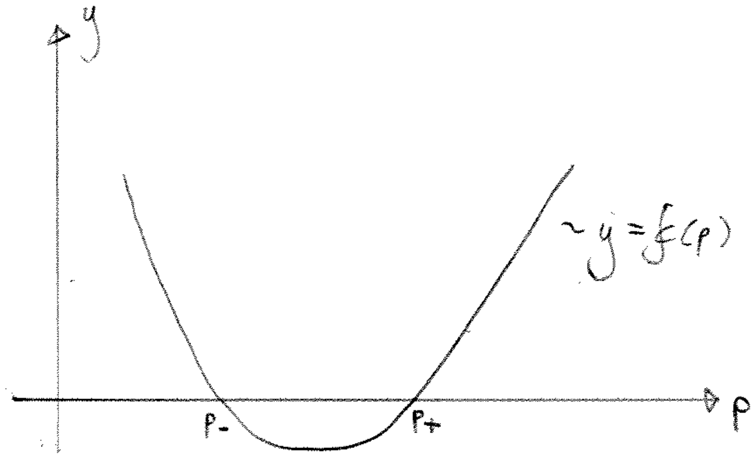
Vi løser den ene ulikheten mhp σ^2:

$$\frac{(n-1)S^2}{\sigma^2} \leq \chi^2{}_{\frac{\alpha}{2},n-1}, \quad \frac{(n-1)S^2}{\chi^2{}_{\frac{\alpha}{2},n-1}} \leq \sigma^2.$$

Dette gjøres tilsvarende med den andre ulikheten. Konklusjonen er at vi har funnet $(1-\alpha)$-konfidensinterallet

$$V = [\frac{(n-1)S^2}{\chi^2{}_{\frac{\alpha}{2},n-1}}, \frac{(n-1)S^2}{\chi^2{}_{1-\frac{\alpha}{2},n-1}}]$$

for σ^2. Om μ er gitt blir løsningen den samme.



Figur 16.2: Intervallestimat for p.

Eksempel 16.2.3 (Tilnærmede konfidensintervall) La Y være binomisk(n, p) fordelt. Vi vil finne et konfidensintervall for p. Vi tar utgangspunkt i den normaliserte variabelen

$$\frac{\frac{Y}{n} - p}{\sqrt{\frac{p(1-p)}{n}}},$$

som har forventning 0 og varians 1. Sentralgrenseteoremet ($Y \sim X_1 + \cdots + X_n$!) gir tilnærmingen

$$P\left(-u_{\frac{\alpha}{2}} \leq \frac{\frac{Y}{n} - p}{\sqrt{\frac{p(1-p)}{n}}} \leq u_{\frac{\alpha}{2}}\right) \doteq 1 - \alpha$$

for store n. Vi finner et tilnærmet konfidensintervall ved å ta utgangspunkt i at variabelen har $N(0, 1)$ fordeling.

Et intervallestimat finnes ved å løse ulikheten

$$\frac{|\frac{y}{n} - p|}{\sqrt{\frac{p(1-p)}{n}}} \leq u_{\frac{\alpha}{2}}$$

med hensyn på p. Kvadrering og ordning gir

$$p^2\left(1 + \frac{{u_{\frac{\alpha}{2}}}^2}{n}\right) - p\left(\frac{2y}{n} + \frac{{u_{\frac{\alpha}{2}}}^2}{n}\right) + \left(\frac{y}{n}\right)^2 \leq 0.$$

Figur 16.2 gir at nullpunktene $p_\pm$ til 2.grads polynomet $f(p)$ på venstresiden gir intervallestimatet $v = [p_-, p_+]$. Løsningene er

$$p_\pm = \frac{\left(\frac{y}{n} + \frac{{u_{\frac{\alpha}{2}}}^2}{2n}\right) \pm \frac{u_{\frac{\alpha}{2}}}{\sqrt{n}}\sqrt{\frac{y}{n}(1 - \frac{y}{n}) + \frac{{u_{\frac{\alpha}{2}}}^2}{4n}}}{1 + \frac{{u_{\frac{\alpha}{2}}}^2}{n}}$$

Vi har antatt at n er stor så vi kan approksimere en gang til ved å neglisjere leddene

$$\frac{{u_{\frac{\alpha}{2}}}^2}{2n}, \quad \frac{{u_{\frac{\alpha}{2}}}^2}{4n}, \quad \frac{{u_{\frac{\alpha}{2}}}^2}{n}.$$

Dette gir det tilnærmede konfidensintervallestimatet

$$v = [\frac{y}{n} - u_{\frac{\alpha}{2}}\sqrt{\frac{y}{n}(1 - \frac{y}{n})\frac{1}{n}}, \frac{y}{n} + u_{\frac{\alpha}{2}}\sqrt{\frac{y}{n}(1 - \frac{y}{n})\frac{1}{n}}]$$

Formelen over er lik den vi har funnet tidligere for normalfordelingen. Den kjente σ^2 er erstattet av estimatet $y/n \cdot (1 - y/n) \cdot 1/n$.

16.3 Oppgaver

214 En tilfeldig trekning fra en $\mathrm{N}(\mu, \sigma^2)$ fordeling gir resultatet 249, 254, 243, 268, 253, 269, 287, 241, 273, 306, 303, 280, 260, 256, 278, 344, 304, 283, 310.

Estimer μ og σ^2.

Finn 95% konfidensintervall for μ og σ^2.

Begrunn at et 95% konfidensintervall for σ finnes ved å ta kvadratroten av intervallgrensene for 95% konfidensintervallet for σ^2.

215 En tilfeldig trekning fra en binomisk$(10, p)$ fordeling ga resultatet 3. Finn et eksakt 95% konfidensintervall for p.

Sammenlign dette med det approksimative 95% konfidensintervallet som følger ved tilnærming med normalfordeling.

Formuler grenseteoremet som er relevant i dette tilfellet, og forklar hvorfor det kan brukes her.

Prøv å gi en definisjon av hva som menes med et approksimativt 95% konfidensintervall.

216 Stikkordene "Tsjebysjevs ulikhet", "intervall-invertering", "pivotvariabel" og "approksimativt konfidensintervall" antyder fire mulige metoder for å finne konfidensintervall. Gi en kortfattet fremstilling, med talleksempel og utledning, av disse metodene.

For tilfellet "approksimativt konfidensintervall" skal du lage et eksempel hvor Poisson-approksimasjonen benyttes.

Det er en fordel om alle de fire metodene blir benyttet på det samme tallmaterialet.

217 Er tilsrekkelighet et relevant begrep i forbindelse med intervallestimering?

Begrunn svaret så godt du kan.

218 Bevis at

$$\tau(V) = \{\tau(\theta) \mid \theta \in V\}$$

er et konfidensområde med nivå α for en generell estmand τ dersom V er et konfidensområde med nivå α for θ.

Kapittel 17

Hypotesetesting

I chose the term 'null hypothesis' without particular regard for its etymological justification but by analogy with a usage, formerly and perhaps still current among physicists, of speaking of a null experiment, or a null method of measurement, to refer to a case in which a proposed value is inserted experimentally in the apparatus and the value is corrected, adjusted, and finally verified, when the correct value has been found; because the set-up is such, as in the Wheatstone Bridge, that a very sensitive galvanometer shows no deflection when exactly the right value has been inserted.
The governing consideration physically is that an instrument made for direct measurement is usually much less sensitive than one which can be made to kick one way or the other according to wether too large or too small a value has been inserted.

R. A. Fisher (1958)

I en gitt statistisk modell kan en hypotese, dvs en antagelse om en ukjent størrelse, identifiseres med en delmengde av parametermengden. Dette er analogt med at en hendelse identifiseres med en delmengde av utfallsrommet.

En hypotesetest basert på en observator er gitt ved at nullhypotesen forkastes hvis og bare hvis observatoren gir en verdi i det kritiske området. Beslutningsregelen i en hypotesetest kan dermed identifiseres med en observator og en hendelse, dvs det kritiske området, i verdiområdet til observatoren.

Koeffisienten til en hypotesetest er den minste øvre skranken for sannsynligheten for å feilaktig forkaste nullhypotesen. I et statistisk lovmessig eksperiment betyr dette at når en test med koeffisient 5% benyttes på gjentagelser av eksperimentet, så vil nullhypotesen forkastes feilaktig i ikke mer enn 5% av tilfellene i det lange løp.

KJENT: statistisk modell, statistisk lovmessig eksperiment, parameter, observator, hendelse, binomisk(n, p), $\mathrm{N}(\mu, \sigma^2)$

NYTT: hypotese, kritisk verdi, kritisk område, testobservator, forkastningsregel, signifikansnivå, signifikanskoeffisient, enkel hypotese

17.1 Terningen på tiltalebenken

Så langt har vi konkludert analysen med et estimat, dvs enten et punktestimat eller et intervallestimat. Estimatet er vår gjetning på verdien til en ukjent størrelse. Intervallet er vår gjetning på et område som inneholder den ukjente størrelsen.

Metodene er mulig å vurdere ved hjelp av sannsynlighetsteori. F eks vil et 90%-konfidensintervall være slik at sannsynligheten for at intervallet inneholder den ukjente størrelsen er minst 90%. Her er det viktig å merke seg at det er intervallet som er den tilfeldige størrelsen.

I en hypotesetest er problemstillingen enklere, dvs det finnes bare to mulige svar. En test gir en prosedyre for å avgjøre om en hypotese er riktig eller gal på grunnlag av de observerte data. Igjen kan sannsynlighetsteorien benyttes til å vurdere egenskapene til testen. Vi konkretiserer diskusjonen med et eksempel.

Eksempel 17.1.1 (Hypotese i terningkast) Anta at vi kaster en terning 18 ganger og registrerer at 12 av kastene resulterer i odde antall øyne. På dette grunnlaget er det nærliggende å tro at det kanskje er mer sannsynlig å få et odde antall øyne enn et like antall øyne i et terningkast. For å ta stilling til dette kan vi sette opp en statistisk modell.

La Y være antall terningkast som gir et odde antall øyne når en terning kastes 18 ganger. Da er det rimelig å anta at Y er en tilfeldig variabel med verdier i $\Omega_Y = \{0, 1, \ldots, 18\}$. Dersom vi antar at terningkastene er uavhengige og at

$$p = P(\text{odde antall øyne i et enkeltkast}),$$

så følger det at

$$Y \sim \text{binomisk}(18, p).$$

I utgangspunktet er det grunn til å anta at terningen er symmetrisk. Vi formulerer dette som at nullhypotesen er gitt ved $p = 1/2$. Den alternative hypotesen er at det er mer sannsynlig å få et odde antall øyne enn et like antall øyne i et terningkast, dvs at $p > 1/2$.

Parameterrommet i vår modell er dermed $\Theta = \{p \,|\, 1/2 \leq p \leq 1\}$. En vanlig notasjon for det foregående er gitt ved

$$\begin{aligned} H_0 :& \quad p = 1/2, \\ H_1 :& \quad p > 1/2. \end{aligned} \tag{17.1}$$

Vi ser at hypotesene er gitt ved en partisjon av parameterrommet, dvs

$$\Theta = \Theta_0 \uplus \Theta_1 = \{1/2\} \uplus (1/2, 1].$$

Dette gjelder generelt.

> I en statistisk modell er hypotesene i en hypotesetest spesifisert ved en oppdeling av parameterrommet i en del tilsvarende nullhypotesen og en del tilsvarende den alternative hypotesen.

La Y betegne antall terningkast av 18 som gir et odde antall øyne. Intuitivt synes det rimelig å forkaste nullhypotesen $p = 1/2$ dersom eksperimentet gir en uventet høy Y verdi. Mer konkret er en mulig beslutningsregel for denne hypotesetesten gitt ved at H_0 forkastes dersom $Y \geq y^*$. Verdien y^* kalles den kritiske verdien til testobservatoren Y.

Vi velger y^* så stor at hendelsen

$$(Y \geq y^*), \quad \text{dvs forkastning av } H_0,$$

er svært lite sannsynlig når H_0 er sann. Et mulig valg er gitt ved å velge y^* slik at

$$P(H_0 \text{ forkastes} \,|\, H_0 \text{ er riktig}) = P(Y \geq y^* \,|\, p = 1/2) \leq 5\%.$$

I så fall er testens signifikansnivå lik 5%. Dersom testen gir forkastning av nullhypotesen, dvs dersom $y \geq y^*$, så sier vi at den alternative hypotesen er bevist "hevet over enhver tvil". Signifikansnivået er en presisering av hva som menes med "hevet over enhver tvil".

> I et statistisk lovmessig eksperiment vil en hypotesetest med nivå α feilaktig forkaste nullhypotesen i ikke mer enn $\alpha \cdot 100\%$ av tilfellene i det lange løp.

For testen med $y^* = 11$ finner vi at

$$P(H_0 \text{ forkastes} \,|\, H_0 \text{ er riktig}) = \sum_{y=11}^{18} \binom{18}{y} (\frac{1}{2})^y (\frac{1}{2})^{18-y} \approx 0.24,$$

og vi sier at signifikanskoeffisienten til testen er 24%. Distinksjonen mellom signifikansnivå og koeffisient for en test tilsvarer distinksjonen mellom konfidensnivå og koeffisient for et konfidensintervall. Enhver $\alpha \geq 0.24$ kvalifiserer som et nivå for testen, og vi har spesielt bevist at testen har signifikansnivå 25%. Fordi vi observerte $y = 12$ ser vi at denne testen vil gi forkastning av nullhypotesen, men et 25% nivå kan vel neppe sies å tilsvare at H_1 er bevist "hevet over enhver tvil".

Tenk et øyeblikk på terningen som den tiltalte i en rettsak. Hypotesen H_1 er tiltalen og nullhypotesen H_0 tilsvarer at terningen er uskyldig. Et nivå på 25% betyr at rettsystemet tillater at terningen blir uskyldig dømt i opptil 25% av tilfellene. Et slikt rettssystem vil neppe oppfattes som rettferdig av folk flest.

Vi ledes derfor til å se på en test med en større kritisk verdi y^*. For $y^* = 15$ finner vi at

$$P(H_0 \text{ forkastes} \,|\, H_0 \text{ er riktig}) = \sum_{y=15}^{18} \binom{18}{y} (\frac{1}{2})^y (\frac{1}{2})^{18-y} \approx 0.004,$$

dvs signifikanskoeffisienten er lik 4 ‰ . Dette gir at testen har et nivå på 5 ‰ , som er for lite til å være av praktisk interesse. Denne testen gir at nullhypotesen beholdes fordi $y = 12$ ikke er inneholdt i mengden $\{15, 16, 17, 18\}$, dvs i testens kritiske område.

Tenk igjen på terningen som den tiltalte i en rettsak. Et nivå på 5 ‰ fører til at det skal svært mye til for at terningen skal bli dømt. Selv med en skriftlig tilståelse, fingeravtrykk og 10 øyevitner er det ikke gitt at terningen blir dømt. Et slikt rettssystem vil heller ikke oppfattes som rettferdig av folk flest.

Testen med $y^* = 13$ har et nivå på 5% fordi

$$P(H_0 \text{ forkastes} \,|\, H_0 \text{ er riktig}) = \sum_{y=13}^{18} \binom{18}{y} (\frac{1}{2})^y (\frac{1}{2})^{18-y} \approx 0.048,$$

som f eks finnes ved oppslag i en statistisk tabell. Dette signifikansnivået synes rimelig. Fordi observasjonen $y = 12$ ikke er med i det kritiske området $\{13, \ldots, 18\}$ har vi følgende konklusjon:

Med et signifikansnivå på 5% forkaster denne testen ikke nullhypotesen.

Spesifiser alltid signifikansnivået i konklusjonen på en hypotesetest.

17.2 Hypotese, forkastningsregel og signifikansnivå

La $\{P^\theta\}$, $\theta \in \Theta$, være en statistisk modell for utfallsrommet Ω.

Nullhypotesen er gitt av en delmengde Θ_0 av parameterrommet Θ. Den alternative hypotesen er gitt av komplementmengden $\Theta_1 = \Theta \setminus \Theta_0$. Definisjon: Θ_0, Θ_1

Tradisjonelt betegner H_0 nullhypotesen og H_1 den alternative hypotesen. Nullhypotesen H_0 er riktig hvis og bare hvis $\theta \in \Theta_0$. En hypotese kan dermed identifiseres med en delmengde av parameterrommet. Dette kan sammenlignes med at en hendelse identifiseres med en delmengde av utfallsrommet.

En forkastningsregel er gitt ved en testobservator og et kritisk område for testobservatoren. En testobservator W er en observator. Et kritisk område K_W for W er en hendelse i verdiområdet til observatoren. Dersom eksperimentet gir at $W(\omega) \in K_W$, så forkastes nullhypotesen. Definisjon: K_W

En forkastningsregel kan også kalles en beslutningsregel, men begrepet beslutningsregel kan benyttes mer generelt. F eks er beslutningsregelen tilsvarende en intervallestimator for en estimand τ gitt ved at dersom eksperimentet gir et intervallestimat $[x, y]$, så er konklusjonen at vi beslutter at $x \leq \tau \leq y$.

Kort fortalt er en beslutningsregel i en hypotesetest gitt ved en observator $W : \Omega \to \Omega_W$ og en hendelse $K_W \subset \Omega_W$. Dersom hendelsen $(W \in K_W)$ inntreffer, så forkastes nullhypotesen.

En hypotesetest er fullstendig spesifisert av nullhypotesen, den alternative hypotesen og en forkastningsregel. Definisjon: **Hypotesetest**

I definisjonen av en hypotesetest inngår kun utfallsrommet og parameterrommet for eksperimentet, mens en mulig statistisk lovmessighet ikke spiller noen rolle. Familien $\{P^\theta\}$ av de mulige fordelingene kan benyttes ved en vurdering av en gitt test.

La W og K_W være hhv testobservatoren og det kritiske området til en hypotesetest hvor nullhypotesen er gitt av parameterdelmengden Θ_0. Testen har signifikansnivå α dersom Definisjon: α

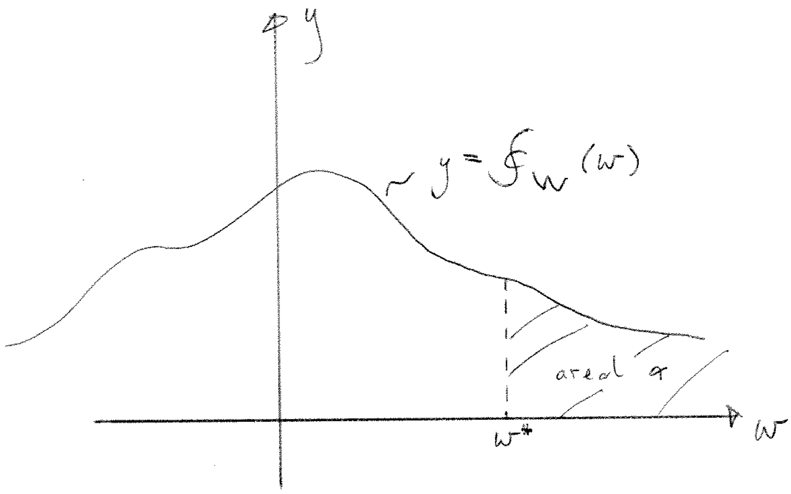
$$P^\theta(W \in K_W) \leq \alpha \text{ for alle } \theta \in \Theta_0.$$

Signifikanskoeffisienten til testen er lik den minste av de α verdiene som oppfyller ulikheten.

Konvensjonen med å utelate parameteren θ er vanlig også i denne sammenhengen. Vi vedtar at ulikheten

$$P(H_0 \text{ forkastes} \,|\, H_0) \leq \alpha$$

betyr eksakt det samme som foregående ulikhet. Denne notasjonen er intuitivt rimelig og leses som



Figur 17.1: Kritisk område $w \geq w^*$ for en test med signifikanskoeffisient α.

"sannsynligheten for at H_0 forkastes gitt H_0 er mindre enn eller lik α".

Her kan det og bemerkes at testen gir at hendelsen $(W \in K_W)$ er identisk med hendelsen $(H_0$ forkastes$)$.

I praksis kan ofte signifikanskoeffisienten til en test finnes som maksimalverdien til funksjonen $f : \Theta_0 \to \mathbb{R}$ definert ved

$$f(\theta) = P^\theta(W \in K_W).$$

Definisjon: $\Theta_0 = \{\theta_0\}$

Nullhypotesen er enkel dersom den er gitt ved en enpunktsmengde, dvs dersom

$$\Theta_0 = \{\theta_0\}.$$

I dette spesielle tilfellet er signifikanskoeffisienten lik funksjonsverdien $f(\theta_0)$.

17.3 Utledning av en test med et gitt nivå

I mange situasjoner er det ønskelig å utlede en test med et gitt signifikansnivå α. Vi gir et enkelt eksempel på hvordan dette kan gjøres.

Anta at testen skal baseres på en tilfeldig variabel W med et kritisk område av typen

$$K_W = \{w \,|\, w \geq w^*\}.$$

Dersom vi videre antar at W er kontinuerlig fordelt gitt nullhypotesen, og at W kun har én mulig fordeling gitt nullhypotesen, så er den kritiske verdien $w^* = \tilde{w}$, hvor $\tilde{w}$ er gitt av

$$P(W \geq \tilde{w} \,|\, H_0) = \alpha.$$

Dermed er den kritiske verdien ikke annet enn α kvantilen til fordelingen til W. Dette er illustrert i figur 17.1.

Mer generelt gir ligningen en $\tilde{w}(\theta)$ for hver θ som er mulig under nullhypotesen. Dersom vi velger

$$w^* \geq \tilde{w}(\theta) \text{ for alle } \theta,$$

så følger det at

$$P^\theta(W \geq w^*) \leq P^\theta(W \geq \tilde{w}(\theta)) = \alpha, \quad \text{for alle } \theta \in \Theta_0.$$

En unik w^* kan finnes som den minste øvre skranke for $\tilde{w}(\theta)$. I praksis kan ofte w^* finnes som maksimalverdien til funksjonen $\tilde{w}$.

I eksemplene så langt har forkastningsområdet vært på formen

$$[w^*, \infty).$$

I terningeksemplet så var dette intuitivt motivert. I andre eksempler kan det naturlige forkastningsområdet ha en annen form, f eks

$$[w^*, \infty),\ [w_1, w_2],\ \{w \mid |w - w_0| \geq w_r\} \text{ etc.}$$

17.4 Oppgaver

219 Definisjonen av et kritisk område K_W for W er at det er en hendelse i verdiområdet til observatoren. Prøv å gi en begrunnelse for at definisjonen velges slik.

Hint: Hva med $(W \in K_W)$?

220 Gi en begrunnelse for at i et statistisk lovmessig eksperiment vil en hypotesetest med nivå α feilaktig forkaste nullhypotesen i ikke mer enn $\alpha \cdot 100\%$ av tilfellene i det lange løp. Hva betyr det at koeffisienten til testen er lik α?

221 La $Y \sim$ binomisk$(18, p)$. Konstruer en test av hypotesen $p = 1/2$ mot alternativet $p > 1/2$ med signifikansnivå 10%. Det kritiske området skal være på formen $K_Y = \{y^*, \ldots, 18\}$ hvor y^* er så liten som mulig. Beregn også signifikanskoeffisienten til denne testen. Hva er konklusjonen på testen dersom vi observerer $y = 12$?

Hint: En statistisk tabell kan være nyttig.

222 La $Y \sim$ binomisk(n, p).

Anta at $n = 12$. Finn en 5% forkastningregel for å teste $H_0 : p = 1/2$ mot $H_1 : p > 1/2$.

Anta at $n = 14$. Finn en 5% forkastningregel med testobservator $W = n - Y$ for å teste $H_0 : p = 3/5$ mot $H_1 : p < 3/5$.

Anta at $n = 20$. Finn en 1% forkastningregel for å teste $H_0 : p = 0{,}3$ mot $H_1 : p > 0{,}3$.

223 Gjør oppgave 6.2.3 i Larsen-Marx.

224 Gjør oppgave 6.2.4 i Larsen-Marx.

225 Gjør oppgave 6.2.5 i Larsen-Marx.

226 Gjør oppgave 6.2.7 i Larsen-Marx.

227 Gjør oppgave 6.2.8 i Larsen-Marx.

228 Gjør oppgave 6.2.9 i Larsen-Marx.

229 Gjør oppgave 6.2.10 i Larsen-Marx.

230 Gjør oppgave 6.2.11 i Larsen-Marx.

231 Anta at nullhypotesen er enkel. Det er gitt at en hypotesetest skal basere seg på en tilfeldig variabel W med et kritisk område av typen $K_W = [a, b]$. Hvordan vil du bestemme $[a, b]$ slik at testen får en konfidenskoeffisient lik α når du antar at W er kontinuerlig fordelt? Hva med tilfellet når W er diskret fordelt? Hvilken rolle spiller den alternative hypotesen her?

Kapittel 18

Styrkefunksjonen og rimelighetsmetoden for hypotesetester

Enhver beslutningsregel basert på tilfeldige størrelser gir feil beslutning med en viss sannsynlighet. Regelen er statistisk god dersom sannsynligheten for feil er liten.

I en hypotesetest er det to mulige feil. Regelen kan forkaste nullhypotesen selv om denne er riktig, eller regelen kan beholde nullhypotesen selv om alternativet er riktig. Disse feilene kalles hhv type-1 og type-2 feil. Nivået til en hypotesetest begrenser sannsynligheten for type-1 feil, og velges typisk lik 5% eller 1% fordi sannsynligheten for type-1 feil ønskes liten.

Når nivået α er gitt ønsker en å minimalisere sannsynligheten for type-2 feil. Denne sannsynligheten avhenger imidlertid av parameteren i den statistiske modellen, og for parameterverdier i nærheten av nullhypotesen er denne sannsynligheten omtrent lik $1 - \alpha$. Dette betyr at det typisk er umulig å ha en liten sannsynlighet for både type-1 og type-2 feil.

Styrkefunksjonen til en test gir sannsynligheten for at nullhypotesen forkastes som funksjon av parameteren. Verdien til styrkefunksjonen for parameterverdier tilsvarende den alternative hypotesen gir teststyrken. Dette betyr at teststyrken er lik 1 minus sannsynligheten for type-2 feil. En test er uniformt bedre enn en annen dersom den har større teststyrke, dvs dersom den har mindre sannsynlighet for type-2 feil.

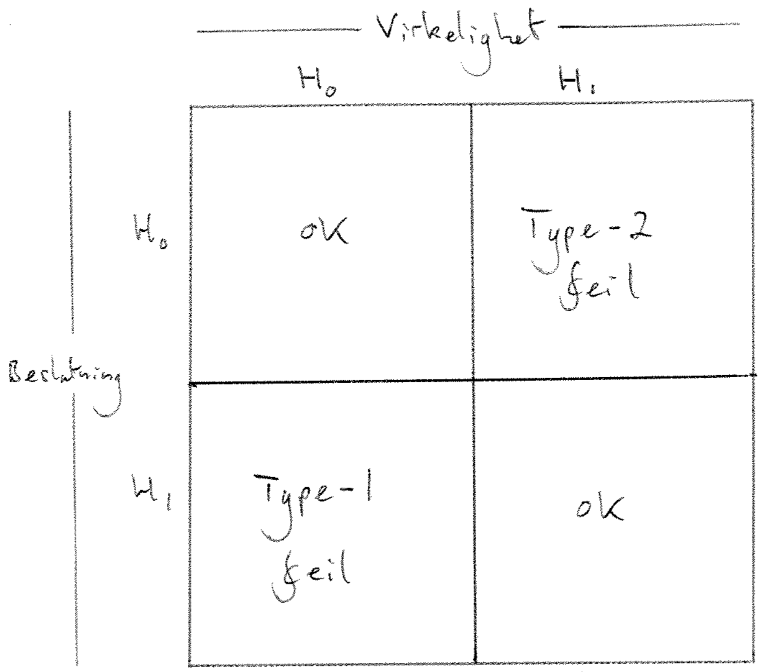
La l_0 være en minste øvre skranke for verdiene til rimelighetsfunksjonen tilsvarende nullhypotesen, og la l være en minste øvre skranke for alle verdiene til rimelighetsfunksjonen. Forholdet l_0/l er den relative rimeligheten til nullhypotesen. Den relative rimeligheten avhenger av rimelighetsfunksjonen, som igjen er gitt av en observasjon. Rimelighetsobservatoren er den tilhørende tilfeldige variabelen gitt ved at observasjonene er tilfeldige. Rimelighetstesten er gitt ved å forkaste nullhypotesen dersom den relative rimeligheten til nullhypotesen er mindre enn eller lik den kritiske rimeligheten. Den kritiske rimeligheten kan velges slik at testen får riktig nivå.

KJENT: tilstrekkelighet, hypotese, kritisk verdi, kritisk område, testobservator, forkastningsregel, signifikansnivå, signifikanskoeffisient, enkel hypotese, rimelighetsfunksjon

NYTT: type-1 feil, type-2 feil, styrkefunksjon, teststyrke, uniformt bedre test, relativ rimelighet, rimelighetsobservator, rimelighetstest

18.1 Innledning

Vi starter med en rask oppsummering av tidligere hovedpunkter. En hypotese H_0 er gitt ved en delmengde $\Theta_0 \subset \Theta$. En beslutningsregel er gitt ved en testobservator W og et kritisk område



Figur 18.1: Klassifisering av feil i en hypotesetest.

$K_W \subset \Omega_W$. H_0 forkastes dersom $w = W(\omega) \in K_W$. Denne testen har signifikansnivå α dersom

$$P(H_0 \text{ forkastes} \,|\, H_0) = P^\theta(W \in K_W \,|\, \theta \in \Theta_0) \leq \alpha.$$

Dersom eksperimentet er statistisk lovmessig betyr dette at testen feilaktig forkaster H_0 i ikke mer enn $\alpha \cdot 100\%$ av tilfellene i det lange løp.

Den alternative hypotesen har spillt en liten rolle så langt. Vi skal se at den alternative hypotesen kommer frem fra kulissene når vi skal sammenligne hypotesetester med et gitt nivå.

Vi har heller ikke gitt noen generell metode for utledning av hypotesetester. Dette skal vi rette på ved å presentere rimelighetstesten. Rimelighetstesten har samme status for hypotesetesting som rimelighetsestimatoren har for estimering.

18.2 Klassifisering av feil og styrkefunksjonen

Definisjon: **type-1 feil**

Dersom en test feilaktig forkaster nullhypotesen, så er dette en type-1 feil. Med denne definisjonen kan definisjonen av nivået til en test reformuleres ved at testen har nivå α hvis og bare hvis

$$P(\text{type-1 feil}) \leq \alpha.$$

Dermed har vi:

> Nivået til en hypotesetest begrenser sannsynligheten for å gjøre en type-1 feil.

Definisjon: **type-2 feil**

Det finnes en annen mulig feil. *Dersom en test ikke forkaster nullhypotesen, selv om den alternative hypotesen er riktig, så er dette en type-2 feil.* De mulige feiltypene er illustrert i figur 18.1.

Notasjonen indikerer en usymmetri mellom nullhypotesen og den alternative hypotesen. En type-1 feil sees på som mer alvorlig enn en type-2 feil. Dette kommer til uttrykk ved at vi først kontrollerer sannsynligheten for type-1 feil ved å velge ønsket nivå.

Sannsynligheten for type-2 feil er gitt ved

$$P(\text{type-2 feil}) = P(H_0 \text{ forkastes ikke} \,|\, H_1) = P^\theta(W \notin K_W \,|\, \theta \in \Theta_1).$$

Gitt et signifikansnivå, dvs gitt en begrensning på sannsynligheten for å gjøre type-1 feil, så ønsker vi at sannsynligheten for type-2 feil skal være så liten som mulig. Dette kan benyttes til sammenligning av tester med et gitt signifikansnivå. Tradisjonelt gjøres dette ved hjelp av testenes styrkefunksjoner.

La $\{P^\theta\}$, $\theta \in \Theta$, være en statistisk modell. La videre W og K_W være testobservatoren og det kritiske området for en hypotesetest. *Styrkefunksjonen*

Definisjon: $\beta(\theta)$

$$\beta : \Theta \to [0, 1]$$

defineres ved

$$\beta(\theta) = P^\theta(W \in K_W).$$

Styrkefunksjonen gir dermed sannsynligheten for å forkaste nullhypotesen som en funksjon av parameteren. Symbolet γ benyttes også i noen tilfeller for å betegne styrkefunksjonen. Vi har her valgt β fordi dette synes mest tradisjonelt, og symbolet "β" ligner på symbolet "p" som kan tas som en indikasjon på den engelske betegnelsen "power-function".

Styrkefunksjonen β inneholder all informasjon som skal til for en vurdering av testen. F eks har testen et nivå α hvis og bare hvis

$$\beta(\theta) \leq \alpha \text{ for alle } \theta \in \Theta_0,$$

og koeffisienten til testen er den minste øvre skranken for mengden $\{\beta(\theta) \,|\, \theta \in \Theta_0\}$.

Teststyrken $\beta(\theta)$ er definert for alle parameterverdier θ tilsvarende den alternative hypotesen, dvs for $\theta \in \Theta_1$. Dette er det samme som at teststyrken er restriksjonen av styrkefunksjonen β til Θ_1. Vi bemerker at teststyrken er gitt ved

Definisjon: **teststyrken**

$$P^\theta(W \in K_W \,|\, \theta \in \Theta_1) = 1 - P^\theta(W \notin K_W \,|\, \theta \in \Theta_1) = 1 - P(\text{type-2 feil}).$$

Teststyrken er dermed lik 1 minus sannsynligheten for type-2 feil. Dermed er det klart at det er ønskelig med så stor teststyrke som mulig for å minimalisere sannsynligheten for type-2 feil. Dette motiverer et generelt sammenligningskriterium for tester.

La A og B betegne to tester med likt signifikansnivå og la β_A og β_B være de tilhørende styrkefunksjonene. *Testen A er uniformt bedre enn testen B dersom*

Definisjon: **uniformt bedre test**

$$\beta_A(\theta) \geq \beta_B(\theta) \textit{ for alle } \theta \in \Theta_1.$$

Adverbet "uniformt" er her en understrekning av at ulikheten skal gjelde for alle parameterverdier tilsvarende den alternative hypotesen, og ikke bare for en enkelt parameterverdi. Definisjonen kan reformuleres ved bruk av definisjonen av teststyrke:

En test er uniformt bedre enn en annen test dersom den har større teststyrke.

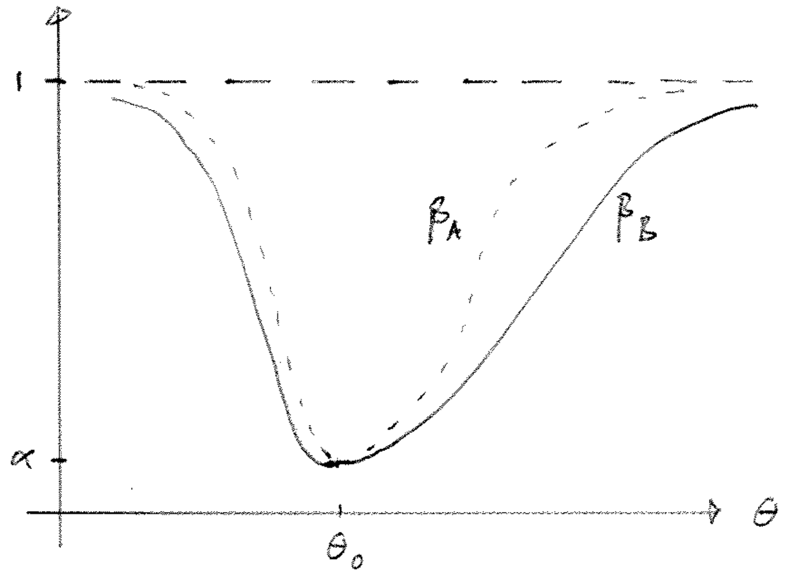
Figur 18.2 viser et eksempel på to mulige styrkefunksjoner. Som regel er styrkefunksjonen kontinuerlig. Dette gir spesielt at teststyrken tvinges til å være liten for parameterverdier i nærheten av nullhypotesens parameterverdier. Dette er forklaringen på at vi ikke kan klare å kontrollere både sannsynligheten for type-1 og type-2 feil.

18.3 Rimelighetstesten

Anta som før at en statistisk modell med parameterrom Θ er gitt, og la Θ_0 være parametermengden tilsvarende nullhypotesen H_0. La L være rimelighetsfunksjonen tilsvarende en observasjon x.

Den relative rimeligheten $\lambda(x)$ til hypotesen H_0 gitt observasjonen x er definert lik

Definisjon: $\lambda(x)$



Figur 18.2: Styrkefunksjonene viser at test A er uniformt bedre enn test B.

$$\lambda(x) = \frac{\sup\{L(\theta) \,|\, \theta \in \Theta_0\}}{\sup\{L(\theta) \,|\, \theta \in \Theta\}}.$$

Her er telleren $\sup\{L(\theta) \,|\, \theta \in \Theta_0\}$ lik det minste tallet l_0 som oppfyller

$$L(\theta) \leq l_0 \text{ for alle } \theta \in \Theta_0,$$

og tilsvarende for nevneren. Generelt betegner $\sup M$ den minste øvre skranken for mengden M.

Anta at rimelighetsestimatet $\hat{\theta}$ finnes, dvs anta at det finnes et globalt maksimalpunkt $\hat{\theta}$ for L. Anta at $\hat{\theta}_0$ er rimelighetsestimatet tilsvarende at parametermengden Θ er erstattet av parametermengden Θ_0. Da følger det at den relative rimeligheten er

$$\boxed{\lambda(x) = \frac{L(\hat{\theta}_0)}{L(\hat{\theta})}}.$$

Metodene vi har brukt for å bestemme rimelighetsestimat er pga dette også viktige i forbindelse med konstruksjon av en rimelighetstest.

Definisjon: $\Lambda = \lambda(X)$

La $\lambda(x)$ være den relative rimeligheten til hypotesen H_0 gitt observasjonen x av verdien til observatoren X. Rimelighetsobservatoren Λ til hypotesen H_0 basert på X er definert ved

$$\Lambda = \lambda(X).$$

Intuitivt synes det rimelig å forkaste nullhypotesen dersom den har liten relativ rimelighet. Dette motiverer definisjonen av den generelle rimelighetstesten.

Definisjon: **rimelighets-test**

Fikser et tall λ^* og la Λ være rimelighetsobservatoren til nullhypotesen basert på observatoren X. En rimelighetstest med kritisk rimelighet λ^* er gitt ved å forkaste nullhypotesen dersom $\Lambda \leq \lambda^*$. Denne entydige testen er rimelighetstesten for nullhypotesen basert på X.

Den kritiske rimeligheten λ^* kan bestemmes slik at testen får et ønsket nivå α. Når α er gitt betyr dette at λ^* velges så stor som mulig slik at det kritiske området $[0, \lambda^*]$ blir så stort som mulig. Dette betyr at λ^* kan velges lik den største nedre skranke for mengden av $1 - \alpha$ kvantiler tilsvarende mengden av de mulige fordelinger til rimelighetsobservatoren Λ.

Det vanligste tilfellet er at en rimelighetstest baseres på et tilfeldig utvalg, dvs X er spesielt en vektor med dimensjon lik størrelsen til utvalget. Formuleringen av rimelighetstesten stiller imidlertid ingen krav til observatoren X. Et godt valg vil være å velge en observator X som er tilstrekkelig for parameteren i forhold til observasjonene. Det kan vises at en rimelighetstest basert på en tilstrekkelig observator er ekvivalent med rimelighetstesten basert på observasjonene selv. Dette viser at en rimelighetstest automatisk oppfyller tilstrekkelighetsprinsippet.

Eksempel 18.3.1 (Rimelighetstest for uniform fordeling) La $Y_1, \ldots, Y_n$ være et tilfeldig utvalg fra uniform$(0, \theta)$-fordelingen. Vi vil teste nullhypotesen $\theta = \theta_0$ mot alternativet $\theta < \theta_0$. Parameterrommet er da

$$\Theta = \{\theta \,|\, \theta \leq \theta_0\}.$$

Nullhypotesen er enkel og gitt ved

$$\Theta_0 = \{\theta_0\}.$$

Rimelighetsfunksjonen til observasjonene $y_1, \ldots, y_n$ er

$$L(\theta) = f(y_1 \,|\, \theta) \cdots f(y_n \,|\, \theta) = 1/\theta^n \qquad 0 \leq y_j \leq \theta,$$

og $L(\theta) = 0$ ellers. Rimelighetsestimatet $\hat{\theta}_0$ for θ tilsvarende parametermengden Θ_0 er trivielt gitt ved

$$\hat{\theta}_0 = \theta_0.$$

Rimelighetsestimatet $\hat{\theta}$ for θ tilsvarende parametermengden Θ er gitt ved

$$\hat{\theta} = y_{\text{maks}},$$

fordi dette er et maksimalpunkt for L. Tilsammen finner vi den relative rimeligheten

$$\lambda(y) = \frac{1/\theta_0^n}{1/y_{\text{maks}}^n} = \left(\frac{y_{\text{maks}}}{\theta_0}\right)^n \qquad y_{\text{maks}} \leq \theta_0.$$

Dersom $y_{\text{maks}} > \theta_0$, så er $\lambda(y) = 0$ fordi dette gir $L(\theta_0) = 0$. Notasjonen over er konsistent med konvensjonen å kun spesifisere funksjonen der verdien er ulik 0, slik vi også gjør for tettheter til tilfeldige variable.

Rimelighetsobservatoren er dermed gitt ved

$$\Lambda = \left(\frac{Y_{\text{maks}}}{\theta_0}\right)^n \cdot 1_{(Y_{\text{maks}} \leq \theta_0)}.$$

Med en kritisk rimelighet $\lambda^* \leq 1$ finner vi at sannsynligheten for å feilaktig forkaste nullhypotesen er

$$P(\Lambda \leq \lambda^*) = P(Y_{\text{maks}} \leq \theta_0 \sqrt[n]{\lambda^*}) = P(Y_1 \leq \theta_0 \sqrt[n]{\lambda^*}) \cdots P(Y_n \leq \theta_0 \sqrt[n]{\lambda^*}) = \lambda^*.$$

Konklusjonen er dermed at koeffisienten til testen er lik den kritiske rimeligheten!

En ekvivalent α nivå test med Y_{maks} som testobservator har et kritisk område

$$[0, \theta_0 \sqrt[n]{\alpha}].$$

Dette følger fordi

$$(0 \leq \Lambda \leq \alpha) = (0 \leq Y_{\text{maks}} \leq \theta_0 \sqrt[n]{\alpha}).$$

18.4 Oppgaver

232 La p være parameteren tilsvarende en Bernoulli forsøksrekke. Anta at $H_0 : p = p_0$ blir testet mot alternativet $H_1 : p > p_0$.

Dersom H_0 forkastes med et nivå 0,05, vil da H_0 nødvendigvis forkastes i en test med et nivå 0,01?

Dersom H_0 forkastes med et nivå 0,01, vil da H_0 nødvendigvis forkastes i en test med et nivå 0,05?

Begrunn svarene!

233 Oppgave 6.3.3 i Larsen-Marx.

234 Oppgave 6.3.4 i Larsen-Marx.

235 La $X \sim \text{uniform}(0, \theta)$ være testobservatoren for en test med et kritisk område

$$[0,\ 0{,}1] \cup [1{,}9,\ \infty).$$

Begrunn at dette er et rimelig valg av kritisk område. Finn signifikanskoeffisienten α tilsvarende nullhypotesen $\theta = 2$. Skisser styrkefunksjonen β for denne testen når den alternative hypotesen er gitt ved $\theta \neq 2$. Hva er teststyrken for $\theta = 2{,}5$?

236 En urne inneholder 10 brikker. Vi ønsker å teste

$$H_0 : \text{ halvparten av brikkene er hvite}$$

mot alternativet

$$H_1 : \text{ mer enn halvparten av brikkene er hvite.}$$

Testen utføres ved å trekke tre brikker uten tilbakelegning og H_0 forkastes dersom minst to av disse er hvite.

Sett opp en statistisk modell hvor parameteren θ er antall hvite brikker i urnen. Finn parameterrommet tilsvarende denne testen. Finn signifikanskoeffisienten til testen. Finn teststyrken $\beta(\theta)$ for $\theta = 6$ og $\theta = 7$. Skisser styrkefunksjonen til testen.

237 Oppgave 6.3.8 i Larsen-Marx.

238 Oppgave 6.3.9 i Larsen-Marx.

239 La $X_1, \ldots, X_n$ være et tilfeldig utvalg fra geometrisk(p)-fordelingen. Finn rimelighetsobservatoren Λ for å teste hypotesen $p = p_0$ mot alternativet $p \neq p_0$. Kan du konstruere en test med nivå α?

240 La $X_1, \ldots, X_n$ være et tilfeldig utvalg fra eksponential(λ)-fordelingen. Finn rimelighetsobservatoren W for å teste hypotesen $\lambda = \lambda_0$ mot alternativet $\lambda \neq \lambda_0$. Kan du konstruere en test med nivå $\alpha = 0{,}05$?

241 La $X_1, \ldots, X_n$ være et tilfeldig utvalg fra $\mathrm{N}(\mu, 1)$-fordelingen. Finn rimelighetsobservatoren Λ for å teste hypotesen $\mu = \mu_0$ mot alternativet $\mu \neq \mu_0$. Kan du konstruere en test med nivå $\alpha = 0{,}01$?

242 Anta at observatoren T er tilstrekkelig for parameteren θ i forhold til observatoren X. Bevis at rimelighetsobservatoren basert på T er lik rimelighetsobservatoren basert på X.

Hint: Bruk faktoriseringen av rimelighetsfunksjonen som tilstrekkeligheten gir.

243 Finn rimelighetsobservatoren Λ basert på Y_{maks} fra et tilfeldig utvalg av størrelse n fra uniform$(0, \theta)$-fordelingen tilsvarende en test av hypotesen $\theta = \theta_0$ mot alternativet $\theta < \theta_0$.

Kommenter resultatet i forhold til eksemplet i teksten.

La $\theta_0 = 1$ og bestem den kritiske rimeligheten slik at testen har koeffisient lik 0,05. Skisser styrkefunksjonen β for denne testen.

Erstatt Y_{maks} med Y_1 i det foregående. Skisser styrkefunksjonen γ for denne testen.

Er den første testen uniformt bedre enn den siste testen? Er den første testen bedre enn den siste på andre vis?

Kapittel 19

Eksempler på konstruksjon av rimelighetstester

Eksempel 19.0.1 (Test for forventning med ukjent varians) Anta at $X_1, \ldots, X_n$ er et tilfeldig utvalg fra $\mathrm{N}(\mu, \sigma^2)$-fordelingen. Parameterrommet er

$$\Theta = \{(\mu, \sigma^2) \mid \mu \in \mathbb{R}, \sigma \geq 0\}.$$

Vi vil teste $H_0 : \mu = \mu_0$ mot alternativet $H_1 : \mu \neq \mu_0$.

Nullhypotesen er gitt ved

$$\Theta_0 = \{(\mu, \sigma^2) \mid \mu = \mu_0, \sigma \geq 0\}.$$

Denne nullhypotesen er ikke enkel, dvs ikke gitt av et enkelt punkt i parameterrommet. Geometrisk er nullhypotesen gitt ved en stråle slik som indikert i figur 19.1.

Vi vil bruke rimelighetsmetoden til konstruksjon av en test. Den relative rimeligheten til H_0 er

$$\lambda(x) = \frac{\text{“rimeligheten til } H_0\text{”}}{\text{“total rimelighet”}} = \frac{L(\hat{\theta}_0)}{L(\hat{\theta})},$$

hvor $\hat{\theta}_0$ og $\hat{\theta}$ er rimelighetsestimatene tilsvarende Θ_0 og Θ.

Beregning av $\hat{\theta}_0$:

$$L = f_X(x \mid \theta) = f(x_1 \mid \theta) \cdots f(x_n \mid \theta) = \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} e^{-\frac{1}{2}\left(\frac{x_1 - \mu_0}{\sigma}\right)^2 - \cdots - \frac{1}{2}\left(\frac{x_n - \mu_0}{\sigma}\right)^2} \tag{19.1}$$

$$\begin{aligned}
\ln L &= \frac{n}{2} \ln\left(\frac{1}{2\pi\sigma^2}\right) - \frac{1}{2\sigma^2}\left((x_1 - \mu_0)^2 + \cdots + (x_n - \mu_0)^2\right) \\
&= -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\left(n\overline{x^2} - 2\mu_0 n\overline{x} + n\mu_0^2\right) \\
&\overset{t=\sigma^2}{=} -\frac{n}{2}(\ln(2\pi) + \ln(t)) - \frac{n}{2t}\left(\overline{x^2} - 2\mu_0\overline{x} + \mu_0^2\right) \\
0 &= \frac{d}{dt}\ln L = -\frac{n}{2}t^{-1} + t^{-2}\frac{n}{2}\left(\overline{x^2} - 2\mu_0\overline{x} + \mu_0^2\right) \\
\sigma^2 &= t = \overline{x^2} - 2\mu_0\overline{x} + \mu_0^2 = \overline{x^2} - \overline{x}^2 + (\overline{x} - \mu_0)^2
\end{aligned} \tag{19.2}$$

Regningen over gir rimelighetsestimatet

$$\hat{\sigma}_0^2 = \overline{x^2} - \overline{x}^2 + (\overline{x} - \mu_0)^2$$

for variansen når forventningsverdien er kjent. Verdien til rimelighetsfunksjonen blir

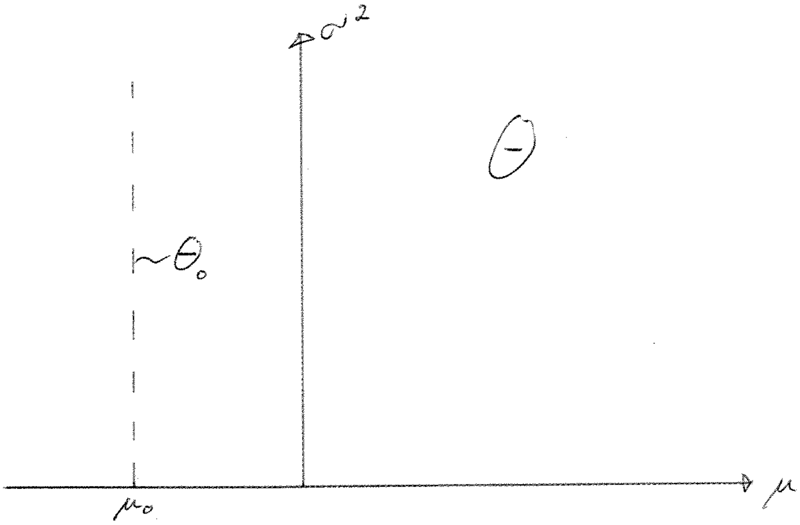
$$l_0 = L(\hat{\theta_0}) = \left(\frac{1}{2\pi}\right)^{n/2}\left(\frac{1}{\overline{x^2} - \overline{x}^2 + (\overline{x} - \mu_0)^2}\right)^{n/2} e^{-n/2}$$

En tilsvarende regning gir rimelighetsestimatet

$$\hat{\theta} = (\overline{x}, \overline{x^2} - \overline{x}^2)$$

når både forventningsverdi og varians er ukjent. Verdien til rimelighetsfunksjonen blir

$$l = L(\hat{\theta}) = \left(\frac{1}{2\pi}\right)^{n/2}\left(\frac{1}{\overline{x^2} - \overline{x}^2}\right)^{n/2} e^{-n/2}$$



Figur 19.1: Nullhypotesen er ikke enkel.

Den relative rimeligheten til H_0 blir da

$$\begin{aligned}\lambda(x) = l_0/l &= \left(\frac{\overline{x^2}-\overline{x}^2}{\overline{x^2}-\overline{x}^2+(\overline{x}-\mu_0)^2}\right)^{n/2} \\ &= \left(\frac{\overline{x^2}-\overline{x}^2+(\overline{x}-\mu_0)^2}{\overline{x^2}-\overline{x}^2}\right)^{-n/2} = \left(1+\frac{(\overline{x}-\mu_0)^2}{\overline{x^2}-\overline{x}^2}\right)^{-n/2} = \left(1+\left(\frac{\sqrt{n}}{\sqrt{n-1}}\,\frac{\overline{x}-\mu_0}{s}\right)^2\right)^{-n/2},\end{aligned} \tag{19.3}$$

hvor

$$s^2 = \frac{n}{n-1}(\overline{x^2}-\overline{x}^2)$$

er utvalgets varians.

Gitt nullhypotesen $\mu=\mu_0$ så er variabelen

$$T = \sqrt{n}\frac{\overline{X}-\mu_0}{S}$$

Student-t$(n-1)$-fordelt. Rimelighetsobservatoren

$$\Lambda = \lambda(X) = \left(1+\frac{1}{n-1}T^2\right)^{-n/2}$$

har dermed også en fordeling som er uavhengig av parameteren $\theta \in \Theta_0$! En rimelighetstest med koeffisient α er dermed gitt ved å velge den kritiske rimeligheten λ^* lik $1-\alpha$ kvantilen til fordelingen til Λ.

En ekvivalent test med T som testobservator er imidlertid mer praktisk fordi kvantilene til T er tabulert. Forkastning ved $\Lambda \le \lambda^*$ er ekvivalent med forkastning ved $|T| \ge t^*$ ved passende valg av t^*. Figur 19.2 gir at vi får en test med koeffisient α ved å forkaste nullhypotesen når

$$|T| \ge t_{\alpha/2,n-1},$$

hvor $t_{\alpha/2,n-1}$ er $\alpha/2$ kvantilen til Student-t$(n-1)$-fordelingen.

Eksempel 19.0.2 (Test for forventning med kjent varians) Anta at $X_1,\ldots,X_n$ er et tilfeldig utvalg fra N(μ,σ^2)-fordelingen hvor variansen σ^2 er kjent. Parameterrommet er

$$\Theta = \{\mu \,|\, \mu \in \mathbb{R}\} = \mathbb{R}.$$

Vi vil teste $H_0 : \mu = \mu_0$ mot alternativet $H_1 : \mu \ne \mu_0$.

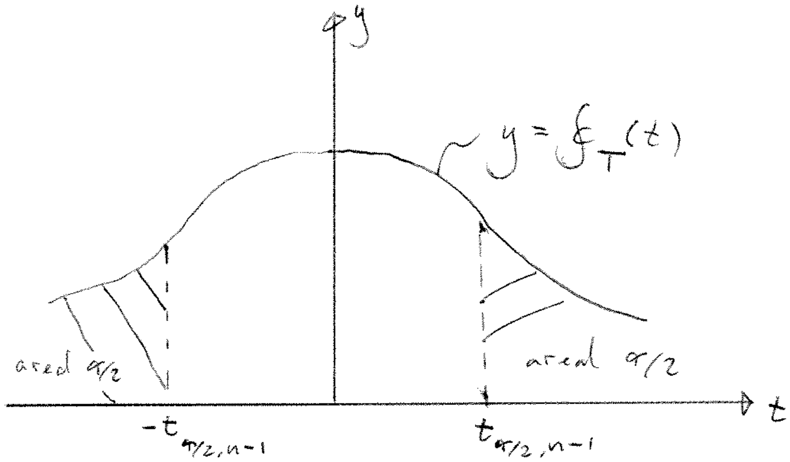
Nullhypotesen er enkel og gitt ved

$$\Theta_0 = \{\mu_0\}.$$

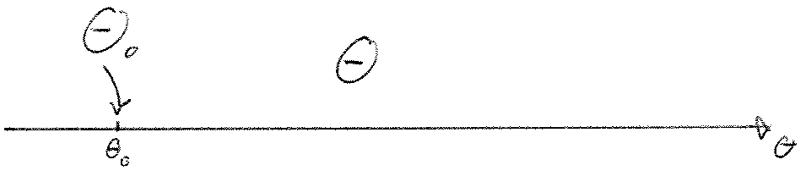
Geometrisk er nullhypotesen gitt ved et punkt slik som indikert i figur 19.3.

Ved beregning av rimelighetsestimatene følger det at den relative rimeligheten er:

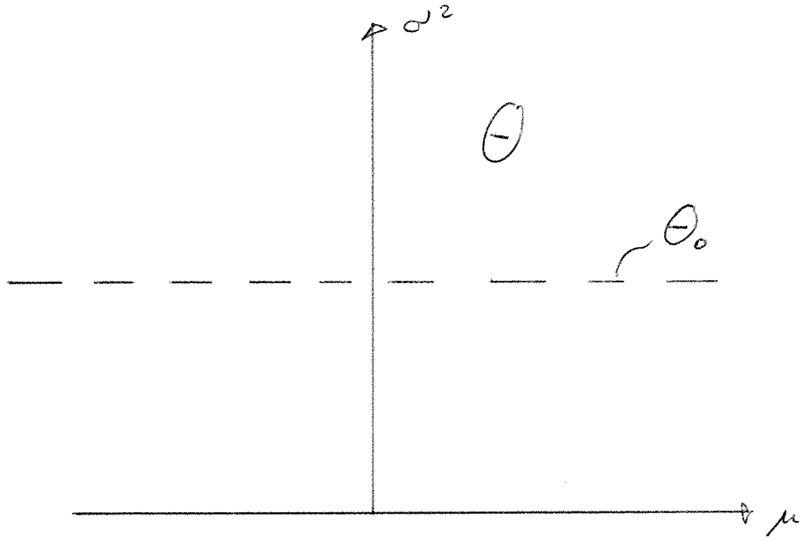
$$\lambda(x) = \frac{L(\hat{\theta}_0)}{L(\hat{\theta})} = \frac{\exp\left[-\frac{n}{2\sigma^2}(\overline{x^2}-2\mu_0\overline{x}+\mu_0^2)\right]}{\exp\left[-\frac{n}{2\sigma^2}(\overline{x^2}-2\overline{x}^2+\overline{x}^2)\right]} = e^{-\frac{n}{2\sigma^2}(\overline{x}-\mu_0)^2}.$$



Figur 19.2: $P(|T| \geq t_{\alpha/2,n-1}) = \alpha/2 + \alpha/2 = \alpha$.



Figur 19.3: Nullhypotesen er enkel.



Figur 19.4: Nullhypotesen er gitt ved en linje i parameterrommet.

Fordi

$$\lambda(x) \leq \lambda^*$$

er ekvivalent med

$$\left|\sqrt{n}(\overline{x} - \mu_0)/\sigma\right| \geq u^*$$

for passende u^* følger det at vi kan benytte

$$U = \sqrt{n}(\overline{X} - \mu_0)/\sigma$$

som testobservator. Fordelen med U fremfor rimelighetsobservatoren Λ er at

$$U \sim \mathrm{N}(0,1)$$

under antagelsen $\mu = \mu_0$. Med testobservator U og kritisk område

$$\{u \,|\, |u| \geq u_{\alpha_2}\}$$

følger det at vi har en test med koeffisient α når $u_{\alpha/2}$ er $\alpha/2$ kvantilen til standard normalfordelingen.

Eksempel 19.0.3 (Test for varians) Anta at $X_1, \ldots, X_n$ er et tilfeldig utvalg fra $\mathrm{N}(\mu, \sigma^2)$-fordelingen. Parameterrommet er

$$\Theta = \{(\mu, \sigma^2) \,|\, \mu \in \mathbb{R}, \sigma \geq 0\}.$$

Vi vil teste $H_0 : \sigma^2 = {\sigma_0}^2$ mot alternativet $H_1 : \sigma^2 = {\sigma_0}^2$.

Nullhypotesen er gitt ved

$$\Theta_0 = \{(\mu, \sigma^2) \,|\, \mu \in \mathbb{R}, \sigma^2 = {\sigma_0}^2\}.$$

Denne nullhypotesen er ikke enkel, dvs ikke gitt av et enkelt punkt i parameterrommet. Geometrisk er nullhypotesen gitt ved en linje slik som indikert i figur 19.4.

Rimelighetsestimatene for θ tilsvarende Θ_0 og Θ er gitt ved

$$\hat{\theta}_0 = (\overline{x}, {\sigma_0}^2) \qquad \hat{\theta}_0 = (\overline{x}, \overline{x^2} - \overline{x}^2).$$

Den relative rimeligheten til nullhypotesen blir dermed

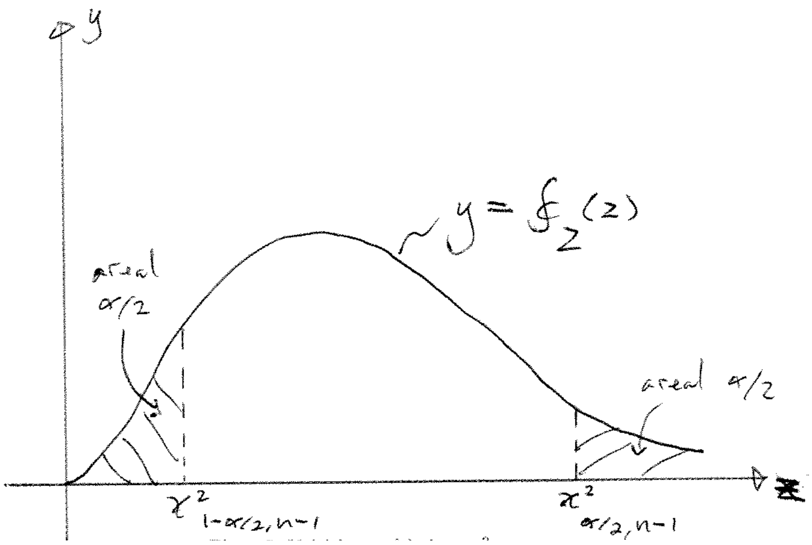
$$\lambda(x) = \frac{\left(\frac{1}{{\sigma_0}^2}\right)^{n/2} \exp\left(-\frac{n}{2}\frac{\overline{x^2} - \overline{x}^2}{{\sigma_0}^2}\right)}{\left(\frac{1}{\overline{x^2} - \overline{x}^2}\right)^{n/2} \exp\left(-\frac{n}{2}\frac{\overline{x^2} - \overline{x}^2}{\overline{x^2} - \overline{x}^2}\right)} = \left(\frac{\overline{x^2} - \overline{x}^2}{{\sigma_0}^2}\right)^{n/2} e^{-\frac{n}{2}\left(\frac{\overline{x^2} - \overline{x}^2}{{\sigma_0}^2} + 1\right)}$$

Rimelighetsobservatoren blir

$$\Lambda = \lambda(X) = Z^{n/2} e^{-\frac{n}{2}(Z+1)},$$

hvor

$$Z = \frac{\overline{X^2} - \overline{X}^2}{{\sigma_0}^2}.$$



Figur 19.5: Kritisk område i en χ^2-test.

Under nullhypotesen $\sigma^2 = {\sigma_0}^2$ følger det at

$$Z \sim \chi^2(n-1),$$

dvs spesielt er fordelingen til Z uavhengig av $\theta \in \Theta_0$. Dette gir at fordelingen til rimelighetsobservatoren Λ også er uavhengig av $\theta \in \Theta_0$! En test med koeffisient α er dermed gitt ved å velge den kritiske rimeligheten lik $1-\alpha$ kvantilen til Λ. Kvantilene til Λ kunne godt ha vært tabulert for forskjellige n, men dette er ikke vanlig i mindre tabellverk.

Et alternativ er å la Z være testobservatoren. Det er klart at Λ er liten dersom Z er liten (pga $Z^{n/2}$ faktoren) eller dersom Z er stor (pga $\exp -\frac{n}{2}(Z+1)$ faktoren). Dette motiverer å velge et kritisk område på formen $[0,a] \cup [b,\infty)$. Dette gir en test med koeffisient α dersom

$$P_Z([0,a] \cup [b,\infty)) = \alpha.$$

Konvensjonelt velges det kritiske området slik at hvert av intervallene får like stor sannsynlighet, dvs det kritiske området velges lik

$$[0, \chi^2_{1-\alpha/2,n-1}] \cup [\chi^2_{\alpha/2,n-1}, \infty).$$

Dette er illustrert i figur 19.5 . Denne χ^2-testen har koeffisient α, men den er ikke ekvivalent med rimelighetstesten, dvs det er ikke en rimelighetstest.

19.1 Oppgaver

244 Vis ved et talleksempel at χ^2-testen ikke er en rimelighetstest.

Mer konkret: Finn utfall x og y slik at den relative rimeligheten til H_0 oppfyller $\lambda(x) \leq \lambda(y)$, mens χ^2-testen forkaster nullhypotesen for y, men ikke for x.

Hint: Beregn verdiene til den relative rimeligheten tilsvarende intervallendepunktene i χ^2-testen for et passende eksempel.

245 Bruk sentralgrenseteoremet til å begrunne at testene vi har gitt for tilfeldig utvalg fra normalfordelingen gir approksimative tester dersom vi tar et tilstrekkelig stort utvalg fra en vilkårlig fordeling med endelig varians.

Begrunn de tre tilfellene: test av forventning med ukjent varians, test av forventning med kjent varians, test av varians hver for seg.

246 Oppgave 7.5.4 i Larsen-Marx.

247 Oppgave 7.5.5 i Larsen-Marx.

248 Oppgave 7.5.6 i Larsen-Marx.

249 Oppgave 7.5.9 i Larsen-Marx.

250 Oppgave 7.5.10 i Larsen-Marx.

251 Oppgave 7.5.13 i Larsen-Marx.

252 Oppgave 7.5.14 i Larsen-Marx.

253 Oppgave 7.5.15 i Larsen-Marx.

254 Oppgave 7.5.17 i Larsen-Marx.

255 Oppgave 7.7.1 i Larsen-Marx.

256 Oppgave 7.7.2 i Larsen-Marx.

257 Oppgave 7.7.3 i Larsen-Marx.

258 Oppgave 7.7.7 i Larsen-Marx.

259 Oppgave 7.7.5 i Larsen-Marx.

260 Oppgave 7.7.9 i Larsen-Marx.

Del IV

Appendiks

Tillegg A

Faginformasjon 1997

S101 Sannsynlighet og statistikk I

Faginformasjon sommer og høst 1997

Faglærer:

Gunnar Taraldsen, rom 303 B, Lade.
Tlf 735 91454, e-post gunnar@matstat.unit.no, hjemmeside http://www.matstat.unit.no/~gunnar.

Lærebok:

R.J. Larsen og M.L. Marx, *An Introduction to Mathematical Statistics and Its Applications*, andre utgave, Prentice-Hall, 1986. G. Taraldsen, *Notat 1-18*, 1997. J. Samseth og A. Thorvaldsen, *Statistiske tabeller og formler*, Tapir, 1996.

Forelesninger:

Man-Fre klokken 9.15–13.00 i perioden 30.juni til 25.juli 1997 på Lade i seminarrom SL11. Første forelesning er mandag 30. juni klokken 12.15–14.00.

I tillegg kommer to helgesamlinger (tors-lør) på tilsammen 6 dager. Tidspunktene for helgesamlingene er 30.oktober til 1.november 1997 og 27.november til 29.november 1997.

Forelesningsplan:

Uke 27: Eksperiment. Sannsynlighet. Kombinatorikk.
Uke 28: Tilfeldige variable. Spesielle fordelinger.
Uke 29: Punkt- og intervallestimering.
Uke 30: Hypotesetesting.

Helgesamlingene benyttes til oppsummering og kommentarer til øvingsoppgavene.

Øvinger:

Innleveres og godkjennes per brev i løpet av høsten. Det vil bli gitt 12 øvinger, og 6 av disse må være godkjent for adgang til eksamen.

Pensum:

Forelesningene, notat 1-18, øvingene og kapittel 1–7 i læreboken, unntatt avsnittet 5.6 og appendiksene 4.1, 5.1, 7.1–2.

Skriftlig eksamen:

Mandag 1. desember 1997, klokken 9-15.
Hjelpemidler: 4 håndskrevne A4-sider med instituttets stempel, kalkulator, *Statistiske tabeller og formler* (Tapir 1996).

Gunnar Taraldsen
Institutt for matematiske fag - NTNU, seksjon LADE
Trondheim 1. november 1997

Tillegg B

Eksamen desember 1997

Norges teknisk–naturvitenskapelige universitet
Institutt for matematiske fag

Faglig kontakt under eksamen: Gunnar Taraldsen (735 91454)

S101 Sannsynlighet og statistikk 1

Mandag 1. desember 1997
Kl. 9-15

Tillatte hjelpemidler: Fire håndskrevne A4-sider med instituttets stempel, kalkulator, *Statistiske tabeller og formler* (Tapir 1996)

Sensurdato: 22. desember 1997

Oppgave 1 *Lyskryss, del 1*

Øyvind kjører bil til jobben hver dag og har begynt å ergre seg over at han stadig vekk får rødt lys i det første lyskrysset han kommer til. Han bestemmer seg for å undersøke dette nærmere.

a) I løpet av 20 dager får han rødt lys 8 ganger. Estimer sannsynligheten p for rødt lys.

b) Estimer standardavviket til estimatoren for p ved hjelp av en forventningsrett estimator for variansen.

c) Beregn et tilnærmet 68% intervallestimat for p ved hjelp av sentralgrenseteoremet. Hint: Du kan anta at $\Phi(1) = 84\%$.

d) Beregn et 68% intervallestimat for p ved hjelp av lineær interpolasjon i tabellen med binomiske sannsynligheter.

Oppgave 2 *Lyskryss, del 2*

a) Øyvind tar tiden på hvor lenge han må vente de gangene han får rødt lys. Observasjonene hans er 28s, 10s, 5s, 15s, 7s, 25s, 29s, 22s. Anta at disse 8 observasjonene er et tilfeldig utvalg fra uniform$(0, a)$-fordelingen. Diskuter kort for og i mot denne antagelsen. Diskusjonen skal innledes med en liste over mulige årsaker til tilfeldighetene.

b) Estimer a ved hjelp av en forventningsrett og tilstrekkelig estimator W. Estimatoren skal være basert på den største observerte verdien.

c) Vis at W er forventningsrett.

d) Vis at W er tilstrekkelig.

Oppgave 3 *Lyskryss, del 3*

a) En hypotese er $a = 30s$, og et alternativ er $a > 30s$. Gir observasjonene grunnlag for å si at alternativet er bevist med et 5% signifikansnivå når testobservatoren er den største observerte verdien?

b) Hva er det kritiske området til testen? Hva blir det kritiske området for en test med den første observerte verdien som testobservator og hva er konklusjonen på denne testen?

c) Tegn en nøyaktig figur med grafene til styrkefunksjonene til de to testene i samme koordinatsystem. Kommenter i figurteksten.

d) Finn sannsynligheten for en type-2-feil for de to testene dersom $a = 80s$ ved hjelp av figuren. Marker på figuren hvordan svaret finnes.

e) Hva betyr denne sannsynligheten for type-2-feil i praksis dersom Øyvinds eksperiment er statistisk lovmessig? Hva betyr denne sannsynligheten dersom eksperimentet ikke er statistisk lovmessig?

Oppgave 4 *Lyskryss, del 4*

a) La R være hendelsen gitt ved at Øyvind får rødt lys. La T være tiden Øyvind bruker i lyskrysset, dvs spesielt er $T = 0$ dersom det lyser grønt. Vis at

$$P(T \leq t) = P(T \leq t|R) \cdot p + P(T \leq t|R^c) \cdot (1-p)$$

med utgangspunkt i Kolmogorovs aksiomer og definisjonen av betinget sannsynlighet.

b) Vis at fordelingsfunksjonen til T er gitt ved

$$F(t) = \begin{cases} 0 & t < 0 \\ pt/a + (1-p) & 0 \leq t \leq a \\ 1 & t > a \end{cases}$$

Gi en begrunnelse for at T hverken er kontinuerlig eller diskret fordelt.

Oppgave 5 *Busser*

Magnar sitter på kontoret og lager løsningsforslag til en gammel eksamen i statistikk. Han ergrer seg over at ikke instituttkontoret hadde et løsningsforslag, og kjeder seg fordi han har laget slike løsningsforslag mer en én gang tidligere. I stedet for å telle sauer, så teller han 16 busser som passerer på veien utenfor kontorvinduet i løpet av 4 timer.

a) Estimer sannsynligheten for at det ikke passerer en eneste buss i løpet av 15 minutter ved å anta at antallet er poissonfordelt.

b) Spesifiser antagelser som gir at antallet busser er binomisk fordelt og bruk dette til å begrunne antagelsen om poissonfordeling.

c) Magnar har erfart mange ganger at dersom han kommer 10 minutter etter at bussen skulle ha gått, og det står andre som bekrefter at bussen ikke har kommet ennå, så er det stor sannsynlighet for at bussen snart kommer. Begrunn at denne erfaringen er i strid med antagelsen om poissonfordeling.

d) Magnar lager en bedre modell ved å anta at det med sikkerhet kommer én buss i hvert 15 minutters intervall gitt ved rutetabellen til bussen. Han antar at tettheten tilsvarende bussavgangen i et slikt intervall er lineær med størst tetthet i venstre endepunkt og null tetthet i høyre endepunkt. I tillegg antar han at bussavgangene i hvert 15 minutters intervall er uavhengige. Magnar beregner forventet ventetid som funksjon av tidspunktet han kommer til bussholdeplassen og han beregner forventet ventetid i tilfellet hvor han går til bussholdeplassen ved et tilfeldig tidspunkt. Han finner også et optimalt tidspunkt. Hva er Magnars tre resultater?

Des - 97

1a La $X_i = [\text{rødt lys dag } i] \sim \text{bernoulli}(p)$. Da er $E\overline{X} = p$, dvs $\overline{X}$ er en forv. rett estimator for p. Et estimat for p er da $\overline{x} = \frac{8}{20} = 40\%$

b.) Fra $ES^2 = \sigma^2 = p\cdot(1-p)$ og $\operatorname{Var}\overline{X} = \frac{1}{n}\sigma^2$ følger det at $\frac{1}{n}S^2 = \frac{1}{n-1}(\overline{X^2} - \overline{X}^2) = \frac{1}{n-1}(\overline{X} - \overline{X}^2)$ er en forv. rett estimator for $\operatorname{Var}\overline{X}$. Et estimat for standardavviket til $\overline{X}$ er dermed

$$S/\sqrt{n} = \sqrt{\frac{1}{19}\cdot\frac{8}{20}\cdot\frac{12}{20}} \approx 0{,}11239 \approx 11{,}2\%$$

c.) Det er 68% sannsynlighet for at intervallet $[\overline{X} - \text{SDEV}\,\overline{X}, \overline{X} + \text{SDEV}\,\overline{X}]$ inneholder p dersom* $\overline{X} \sim N(p, (\text{SDEV}\,\overline{X})^2)$ fordi $P(|\overline{X} - p| > \text{SDEV}\,\overline{X})$ $= 2\cdot P\left(\frac{\overline{X}-p}{\text{SDEV}\,\overline{X}} > 1\right) = 2\cdot(1 - \Phi(1)) = 32\%$.

*Begrunnes ved sentralgrenseteoremet. Et tilnærmet 68% intervallestimat er gitt ved grensene $\overline{x} \pm s/\sqrt{n}$, dvs $[28{,}8\%; 51{,}2\%]$

d.) Metoden på s.279 i Larsen&Marx gir at et eksakt intervall $[p_1, p_2]$ er gitt ved $P(Y \geq 8; p_1) = P(Y \leq 8; p_2) = 16\%$ hvor $Y \sim \text{binomisk}(20, p_i)$. Lineær interpolasjon gir p_i:

$P(Y \leq 7; p_1) = 84\%$, $p_1 = 0{,}25 + \frac{0{,}30 - 0{,}25}{0{,}772 - 0{,}898}(0{,}84 - 0{,}898) \approx 0{,}2730$

$P(20 - Y \leq 11; p_2) = 84\%$, $1 - p_2 = 0{,}40 + \frac{0{,}50 - 0{,}40}{0{,}748 - 0{,}943}(0{,}84 - 0{,}943)$ $\approx 0{,}4528$, $p_2 \approx 0{,}5472$): $[27{,}3\%; 54{,}7\%]$

2a
- Avreisetidspunkt er tilfeldig.
- Kjøretiden varierer pga trafikken
- Det er nærliggende å anta at Øyvind kommer til lyskrysset ved et tidspunkt som er normalfordelt pga en sum av tilfeldigheter. Dette kan benyttes til å sette opp en modell med 3 parametre. Denne modellen vil inneholde den uniforme modellen som et grensetilfelle tilsvarende stor varians i normalfordelingen.

} Stor variasjon på sekundskala!

- Den uniforme modellen foretrekkes her fordi den er enkel, og naturlig. Mot: For enkel, og derfor lite fleksibel.

b.) $P(Y_{maks} \le y) = (y/a)^n$

$$E\,Y_{maks} = \int_0^a y \cdot n\,(y/a)^{n-1} \cdot 1/a \, dy = \frac{n}{n+1} \cdot a$$

$$W = \frac{n+1}{n} Y_{maks}, \quad w = \frac{9}{8} \cdot 29\,s \approx 32{,}625\,s \approx \underline{\underline{33\,s}}$$

c.) $E\,W = \frac{n+1}{n} \cdot E\,Y_{maks} = \underline{\underline{a}}$): $\underline{\underline{W \text{ er forventningsrett for } a}}$

d.) $$f(\vec{y}\,; a) = \prod_{i=1}^{n} \frac{1}{a}[0<y_i<a] = \underbrace{\left(\prod_{i=1}^{n}[0<y_i]\right)}_{\text{uavh. av } a} \cdot \underbrace{\frac{1}{a^n}\left[W < \frac{n+1}{n}a\right]}_{\text{avh. av } \vec{y} \text{ via } W.}$$

gir at W er tilstrekkelig.

d.) $f(\vec{y};a) = \prod_{i=1}^{n} \frac{1}{a}[0<y_i<a] = \left(\prod_{i=1}^{n}[0<y_i]\right)\cdot \frac{1}{a^n}\cdot[y_{maks}<a]$

uavh. av a — avhenger av a gjennom y_{maks}

Dermed er y_{maks} tilstrekkelig.
Omskrivningen $[y_{maks}<a] = [w < \frac{n+1}{n}a]$ viser at også w er tilstrekkelig.
(Strengt tatt: Y_{maks} og W er estimatorer og det er disse, ikke estimatene y_{maks}, w, som er tilstrekkelige.)

3 a

$P(Y_{maks} \ge 29s) = 1 - P(Y_{maks} \le 29s) = 1-\left(\frac{29s}{30s}\right)^8$
$\approx 23,75\% > 5\%$

): Alternativet er ikke bevist med signifikansnivå 5%

(Alternativt kan det kritiske området i b.) benyttes!)

b.) $5\% = 1-\left(\frac{y}{30s}\right)^8$ gir $y = 30s\cdot\sqrt[8]{95\%} \approx 29,808$

$[29,81s;\infty)$ $\qquad$ $[30s\cdot 95\%;\infty) \approx [28,50s;\infty)$

H_0 beholdes også for denne testen

c.) $P(Y_{maks}\ge 29,81s; a) = 1-\left(\frac{29,81}{a}\right)^8 = \beta(a)$

$P(Y_1 \ge 28,50s; a) = 1-\left(\frac{28,50s}{a}\right) = \gamma(a)$

a	30	33	40	50	75	100
β	4,96%	55,66%	90,48%	98,40%	99,94%	99,99%
γ	5,00%	13,64%	28,75%	43%	62%	71,5%

c.)

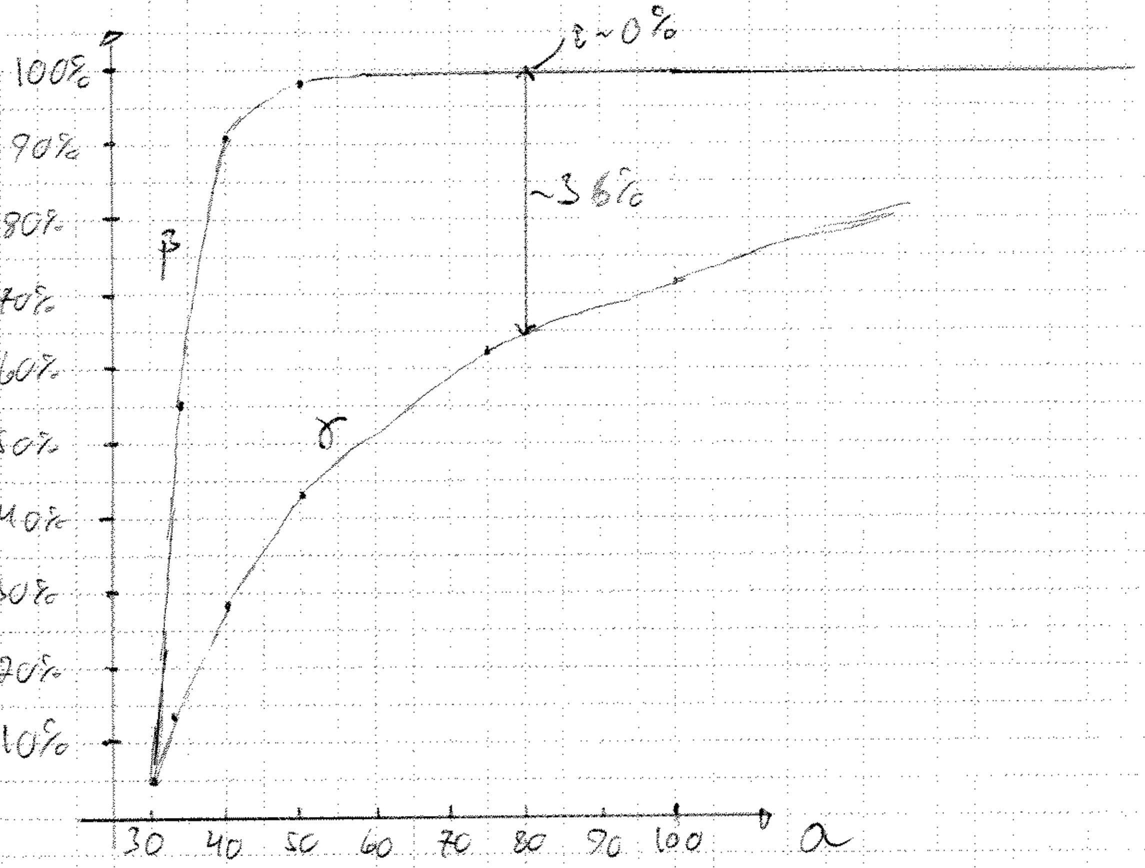


Fig. 1 Testen med testobservator Y_{maks} er uniformt bedre enn testen med testobservator Y_1.

d.) Sannsynligheten for type-2 feil er hhv (omtrent) 0% og 36% fra fig. 1.

e.) Statistisk tolkning: I det lange løp vil prosedyren gi type-2 feil i (høyst) 1% (34%) av tilfellene (relativ frekvens → 1% (34%)).
Generelt: Vår vurdering er at prosedyren gir type-2 feil med 1% (34%) sannsynlighet. Dette er en konsekvens av vurderingene av sannsynligheter som ligger til grunn for modellen.

4a $P(T \le t) = P\left([(T \le t) \cap R] \cup [(T \le t) \cap R^c] \right)$

$$\overset{\text{Kolmogorov}}{\underset{\text{Add. regel}}{=}} P((T \le t) \cap R) + P((T \le t) \cap R^c)$$

$$\overset{\text{Def. av bet.}}{\underset{\text{sannsynl.}}{=}} P(T \le t \mid R) \cdot P(R) + P(T \le t \mid R^c) \cdot P(R^c)$$

$$= P(T \le t \mid R) \cdot p + P(T \le t \mid R^c) \cdot (1-p)$$

b.) $F(T \le t) = P(T \le t \mid R) \cdot p + P(T \le t \mid R^c) \cdot (1-p)$

$$= \begin{cases} 0 & t < 0 \quad (\text{da er } P(T \le t) = 0!) \\ p \cdot t/a + (1-p) & 0 \le t \le a \quad (\text{unif. ved rødt og } T = 0 \text{ ved grønt}) \\ 1 & t > a \quad (P(T \le a) = 1) \end{cases}$$

T er ikke kontinuerlig fordelt fordi F ikke er deriverbar i 0.

T er ikke diskret fordelt fordi F ikke er en trappefunksjon.

5a Poissonparameteren tilsvarende 15 minutter estimeres av $\hat{\lambda} = \frac{16}{4\text{timer}} : 15\text{minutter} = 1$. Dette gir estimatet

$$\frac{\hat{\lambda}^0}{0!} e^{-\hat{\lambda}} = \underline{\underline{e^{-1}}} \approx 0{,}3679 \underline{\underline{\approx 37\%}}$$

b.) Del opp 15 min. intervallet i n like store intervall. La X_i = [Det kom en buss i intervall i], $i = 1, \dots, n$. Anta at X_i'ene er et tilfeldig utvalg fra bernoulli(p)-fordelingen. Det følger da at $\sum X_i \sim$ binomisk(n,p). Dersom n er stor nok er det rimelig å anta at det maksimalt kommer én buss i hvert intervall, og dette gir at antall busser $= \sum X_i \sim$ binomisk(n,p). Vi har $E\sum X_i = np = \lambda$ ved $p = \lambda/n$. Ved $n \to \infty$ følger $\binom{n}{x} \cdot (\lambda/n)^x \cdot (1-\lambda/n)^{n-x} \to \frac{\lambda^x}{x!} \cdot e^{-\lambda}$ (Poisson-approksimasjonen) som gir $\sum X_i \xrightarrow{L} Y \sim$ Poisson(λ).

c.) Poissonantagelsen fører til at ventetiden er eksponensialfordelt med forventning (estimert lik 15 min) uavhengig av når Magnar kommer til busholdeplassen, og spesielt uavhengig av om bussen har gått en gang i løpet av de 10 minuttene før Magnar kom til bussholdeplassen. Dette er i strid med Magnars erfaring i situasjonen gitt i oppgaven.

d.) Tettheten til avgangstiden T er $f(t) = 2 - 2t$, $0 \le t \le 1$, hvor $t=1$ svarer til 15 minutter. Forv. ventetid dersom Magnar kommer ved tidspunkt $s=0$ er $ET = \int_0^1 t \cdot (2-2t) = \frac{1}{3}$ (dvs 5 min.). La V være Magnars ventetid. Magnar når enten bussen i det intervallet han ankommer, eller så når han bussen en gang i neste intervall: $EV = E(V \mid T \ge s) \cdot P(T \ge s)$
$+ E(V \mid T < s) \cdot P(T < s)$.

5d (fortsatt) Fordelingen gitt $T \geq s$: $g(t) = f(t)/P(T\geq s)$, $s \leq t \leq 1$. $E(V|T\geq s)\cdot P(T\geq s) = \int_s^1 (2-2t)\cdot t\,dt = \frac{1}{3} - s^2 + \frac{2}{3}s^3$.

$P(T<s) = \int_0^s (2-2t)\,dt = 2s - s^2$. $E(V|T<s) = (1-s) + \frac{1}{3}$.

Tilsammen $EV = \frac{1}{3} + \frac{8}{3}s - \frac{13}{3}s^2 + \frac{5}{3}s^3$, $0 \leq s \leq 1$.

Tilfeldig tidspunkt tilsvarer uniform S, som gir

$E(E[V|S]) = \frac{1}{3} + \frac{4}{3} - \frac{13}{9} + \frac{5}{12} = \frac{23}{36} \approx 0{,}64$, dvs 9 min og 35 sek. ventetid ved tilfeldig valgt tidspunkt.

Forventet ventetid når Magnar ankommer ved tidspunkt u: $EV = 15\text{min}\cdot\left(\frac{1}{3} + \frac{8}{3}\cdot\frac{u}{15\text{min}} - \frac{13}{3}\cdot\left(\frac{u}{15\text{min}}\right)^2 + \frac{5}{3}\cdot\left(\frac{u}{15\text{min}}\right)^3\right)$

$$= \left(5 + 40\cdot\frac{u}{15\text{min}} - 65\cdot\left(\frac{u}{15\text{min}}\right)^2 + 25\cdot\left(\frac{u}{15\text{min}}\right)^3\right)\text{min}$$

Tidspunkt for ekstremverdi til EV:

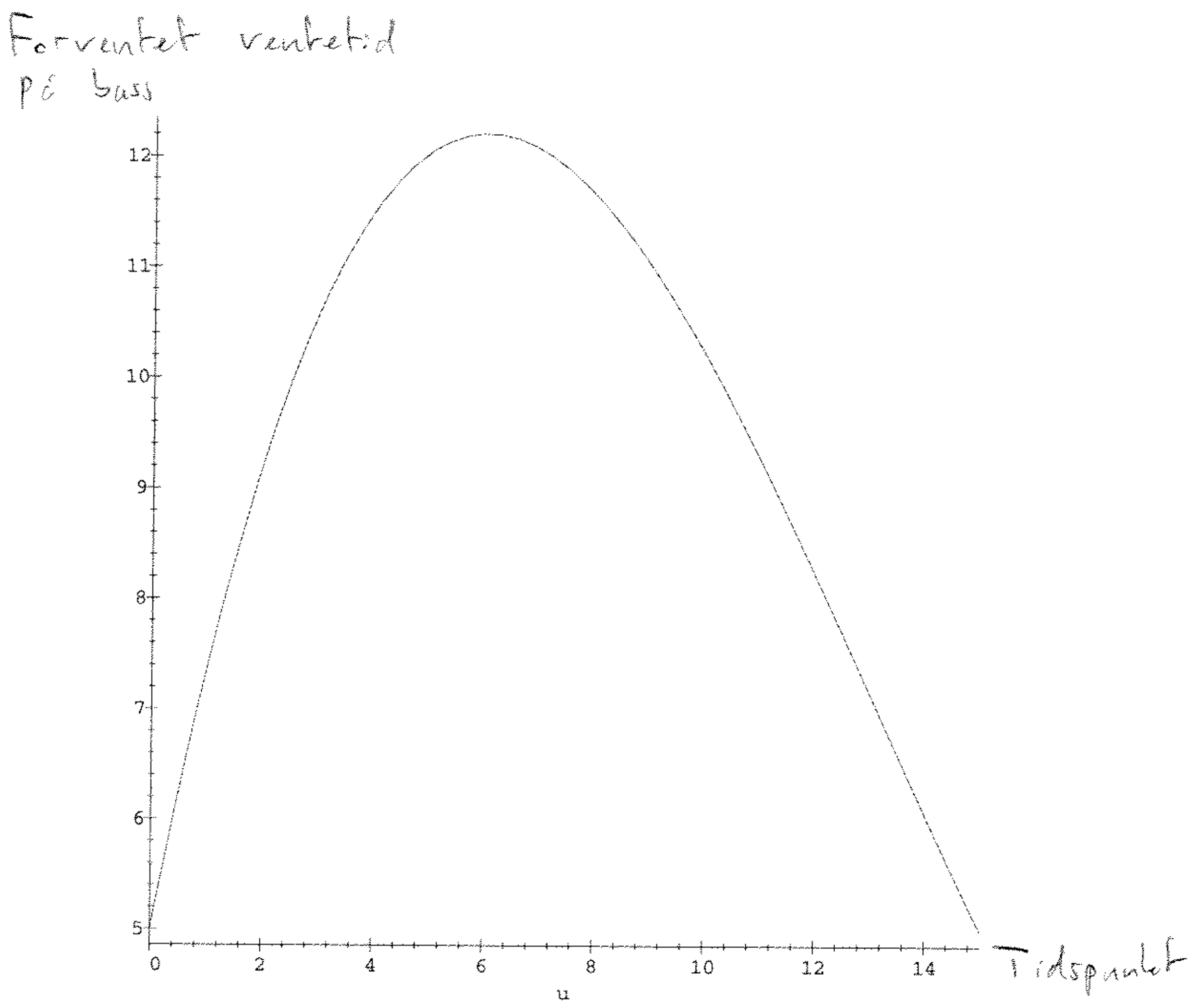
$0 = \frac{8}{3} - \frac{26}{3}s + \frac{15}{3}s^2$, $s = \frac{\overset{13}{\cancel{26}} \pm \sqrt{\overset{13}{\cancel{26}}^2 - \cancel{4}\cdot 15\cdot 8}}{\cancel{2}\cdot 15} = \frac{13 \pm 7}{15} = \begin{cases} 2/5 \\ 4/3 \text{ ikke aktuell} \end{cases}$

$EV_{s=2/5} = \frac{1}{3} + \frac{8}{3}\cdot\frac{2}{5} - \frac{13}{3}\cdot\left(\frac{2}{5}\right)^2 + \frac{5}{3}\cdot\left(\frac{2}{5}\right)^3 = \frac{61}{75}$

): Dersom du kommer 6 min etter tidspunktet hvor bussen skulle ha gått, så blir den forventede ventetid maksimal, og lik 12 min 12 sek.

Det optimale tidspunktet er gitt ved $s=0$, dvs Magnar konkluderer at han bør møte opp til det tidspunktet hvor bussen skal gå. Da er ventetiden minimal, og lik 5 min.

```
> ?plot
> plot(5 + 40*(u/15) - 65*(u/15)^2 + 25*(u/15)^3, u=0..15);
```




```
>
```

Tillegg C

Eksamen juni 1997

Norges teknisk–naturvitenskapelige universitet
Institutt for matematiske fag

Faglig kontakt under eksamen: Gunnar Taraldsen
73 59 14 54

S 101 SANNSYNLIGHET OG STATISTIKK I

Fredag 6. juni 1997
Kl. 9-15
Tillatte hjelpemidler: 4 håndskrevne A4-sider; instituttets lommekalkulator
Statistiske tabeller og formler (Tapir)
Sensurfrist: Fredag 27. juni 1997

Oppgave 1 (Jente eller gutt?)

a) La Y være antall jenter født i n fødsler. Gi en begrunnelse for at $Y = Y_1 + \cdots + Y_n$, hvor Y_i-ene er uavhengige Bernoulli-variable.

b) La p være sannsynligheten for at en nyfødt unge er en jente. Gi et kombinatorisk argument for at $P(Y = y) = \binom{n}{y} p^y (1-p)^{n-y}$.

c) Når n er stor, så er Y tilnærmet normalfordelt med $E(Y) = np$ og $\text{Var}(Y) = np(1-p)$. Formuler sentralgrenseteoremet og vis at dette gir denne tilnærmelsen.

d) Statistisk årbok fra 1994 oppgir at det ble født 27708 jenter og 29154 gutter i Norge i perioden 1986–1990. Benytt tilnærmelsen over til å lage en test for hypotesen $H_0\colon p = 0.50$ mot alternativet $H_1: p < 0.50$ med et signifikansnivå $\alpha = 1\%$. Hva er konklusjonen på testen ut fra de gitte dataene?

e) Finn det tilnærmede kritiske området for testobservatoren Y for testen.

f) Hva menes med type-I og type-II feil i testen? Hva er sannsynligheten for type-I feil i testen?

g) Tegn grafen til styrkefunksjonen γ til denne testen ved å beregne den tilnærmede sannsynligheten $\beta(p)$ for type-II feil for $p = 0$, $p = 0.49$, $p = 0.495$, $p = 0.50$.

Oppgave 2 (Potetavling)

Statistisk årbok oppgir en potetavling på 2424, 2575, 2275, 2662 (kg per dekar) i årene 1989–92.

a) Argumenter for at avlingen er normalfordelt.

b) Estimer forventningsverdien μ og standardavviket σ med de konvensjonelle estimatorene $\bar{X}$ og S.

c) Vis at estimatoren S^2 for variansen er forventningsrett.

d) Vis at $T_{n-1} = \sqrt{n}(\bar{X} - \mu)/S$ er (Student) t-fordelt med $n-1$ frihetsgrader. Du kan bruke resultater fra *Statistiske tabeller og formler.*

e) Hva menes med et 90%-konfidensintervall for μ?

f) Utled og beregn et 90%-konfidensintervall for μ.

g) Anta nå at estimatet for σ er lik σ. Beregn et 90%-konfidensintervall for μ.

h) Kommenter forskjellen på de to intervallestimatene.

Oppgave 3 **(Tilfeldige variable)**

I hele denne oppgaven skal vi la $S = [0,1]$ være utfallsrommet for et eksperiment og anta at $P([a,b]) = b - a$ når $[a,b]$ er en hendelse, d.v.s. P er den uniforme fordelingen. Vi skal videre anta at X, Z_A og Z_B er tilfeldige variable i dette eksperimentet.

a) Finn sannsynlighetstettheten f tilsvarende P og skriv ned ligningen som gir sammenhengen mellom f og P.

b) La $A = [0, 1/2]$ og $B = [0, 1/4] \cup [1/2, 3/4]$. Vis at A og B uavhengige hendelser.

c) Definer Z_A ved at den tar verdien 1 dersom utfallet s av eksperimentet er i A og verdien 0 ellers. Definer Z_B tilsvarende fra B. Hvilken kjent fordeling har $Z = Z_A + Z_B$?

d) Finn den simultane fordelingsfunksjonen f_{Z_A,Z_B} til Z_A og Z_B.

e) Finn og skisser den kumulative fordelingsfunksjonen F_Z til Z.

f) I resten av oppgaven antar vi at X er normalfordelt med $E(X) = 0$ og $\mathrm{Var}(X) = 1$. Skriv ned en formel som gir den kumulative fordelingsfunksjonen F_X. Hva er sammenhengen mellom F_X og fordelingsfunksjonen f_X?

g) Definer hva det betyr at X er en tilfeldig variabel i dette eksperimentet.

h) Skriv ned ligningen som gir sammenhengen mellom X og P. Intervallet $[0, 1]$ skal inngå i denne ligningen.

2/7-97

Juni 97

I.

a) $Y_i = 1$ dersom fødsel # i resulterer i jente, og $P(Y_i=1) = p$. Hver fødsel antas uavhengig.

b.) $p^y \cdot (1-p)^{n-y}$ er sannsynligheten for at fødsel # $1, \ldots, y$ resulterer i jente og fødsel # $1+y, \ldots, n$ resulterer i gutt.

Det finnes $n!$ permutasjoner av disse fødslene. Permutasjonene innenfor de y jentefødslene endrer ikke utfallet, og tilsv. for de $n-y$ guttefødslene.

): $n! / (y!(n-y)!)$ er antall mulige fødsler med y jenter.

$$P(Y=y) = \sum_{\substack{\#\text{ fødselsek.} \\ \text{med } y \text{ jentef.}}} p^y \cdot (1-p)^{n-y} = \binom{n}{y} \cdot p^y \cdot (1-p)^{n-y}$$

c.)
$$\left.\begin{aligned} EY &= \sum EY_i = n \cdot p \\ \operatorname{Var} Y &\underset{\text{Uavh.}}{=} n \cdot \operatorname{Var} Y_i = n \cdot [EY_i^2 - (EY_i)^2] \\ &= n \cdot [p - p^2] = n \cdot p \cdot (1-p) \end{aligned}\right\} \text{Unødvend}$$

Y_i'ene er uavh. med identisk fordeling med $EY_i = p$ og $\operatorname{Var} Y_i = p \cdot (1-p)$.

$$\lim_{n\to\infty} P\left(a \le \frac{Y_1 + \cdots + Y_n - n \cdot p}{\sqrt{n} \cdot \sqrt{p \cdot (1-p)}} \le b\right) = \int_a^b \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{1}{2}x^2} dx$$

M. a. o. er $Z = \dfrac{Y - n \cdot p}{\sqrt{n} \cdot \sqrt{p \cdot (1-p)}}$ tilnærmet

$\sim N(0, 1)$.

Da er $Y = \sqrt{n} \cdot \sqrt{p \cdot (1-p)} \cdot Z + np$ tilnærmet normalfordelt med $E[Y] = np$ og $\operatorname{Var} Y = n \cdot p \cdot (1-p)$.

d.) $P\left(\dfrac{Y - n \cdot \frac{1}{2}}{\sqrt{n \cdot \frac{1}{2} \cdot \frac{1}{2}}} \leq -u_{1\%} \right) = 1\%$

Tabell gir $u_{1\%} = 2{,}326$.

Dataene $\dfrac{27708 - \frac{1}{2} \cdot (27708 + 29154)}{\sqrt{\frac{1}{4} \cdot (27708 + 29154)}} = -6{,}064$

H_0 forkastes

e.) $\dfrac{y - np}{\sqrt{n \cdot p \cdot (1-p)}} \leq -u_{1\%}$ $\quad (p = \frac{1}{2})$

$y \leq np - u_{1\%} \cdot \sqrt{n \cdot p(1-p)} \approx 28153{,}7$

$y \leq 28153 = y_c$ er kritisk område

$(n/2 = 28431)$

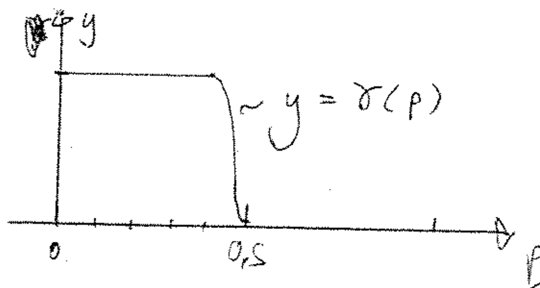
f.) Type I feil: H_0 er riktig, men forkastes.

Type II feil: H_1 er riktig, men H_0 forkastes ikke.

$P(\text{Type I feil}) = \alpha = 1\%$

g.) Sanns. for type II feil er $P(Y > y_c)$, d.v.s. sanns. for at $\frac{Y - np}{\sqrt{n \cdot p \cdot (1-p)}} > \frac{y_c - np}{\sqrt{n \cdot p \cdot (1-p)}} = y_p$

p	0	0,1	0,2	0,3	0,4	0,5	0,45	0,48	0,49
y_p	∞	314	236	102	46	−2,33	21,6	7,21	2,44
β	0	0	0	0	0	99%	0	0	0,007
γ	1	1	1	1	1	1%			0,993



0,495

0,05

0,48

0,52

II.

a) Antall poteter som resulterer fra en avling er binomialfordelt ved at dette er summen av Bernoulli-var. tilsv. hver mulig potet. Antallet er stort, så en asymptotisk normalfordeling følger. En faktor gir omregning til kg/dekar, så avlingen i kg/dekar er (tilnærmet) normalfordelt.

b)

$\hat{\mu} \approx 2484$ kg/dekar

147,69

$$\hat{\sigma} = \sqrt{\frac{1}{3} \cdot [(2424-2484)^2 + \cdots]} \approx 170,5 \text{ kg/}$$

29 082

c) $S^2 = \frac{1}{n-1} \cdot \sum (X_i - \bar{X})^2 = \frac{1}{n-1} \cdot [\sum (X_i^2 - \bar{X}$

$E X_i^2 = \sigma^2 + \mu^2$, $E \bar{X}^2 = \frac{\sigma^2}{n} + \mu^2$

$E S^2 = \frac{1}{n-1} \cdot \sum (\sigma^2 - \frac{1}{n} \sigma^2) = \sigma^2 \cdot \frac{1}{n-1} \cdot n \cdot \frac{n-1}{n}$

$\therefore E S^2 = \sigma^2$

4/9

d.) Det holder å vise at

$$T_{n-1} = \frac{U}{\sqrt{Z}} \cdot \sqrt{n-1} \quad \text{hvor } U \sim N(0,1)$$
$$Z \sim \chi^2_{n-1}$$
U, Z uavh.

$$Var(\bar{X} - \mu) = \frac{1}{n^2} \cdot n \cdot \sigma = \frac{\sigma}{n}$$

$$(\bar{X} - \mu)/\sqrt{\frac{\sigma}{n}} \sim N(0,1)$$

$$(n-1) \cdot S^2/\sigma^2 \sim \chi^2_{n-1}$$ gir påstanden

e. & f.

$$\bar{X} \pm t_{\alpha/2, n-1} \cdot \frac{S}{\sqrt{n}}$$ [2283,6 , 2684,4]

$t_{5\%, 3} = 2,35$ $\quad P(T_{n-1} > t_{\alpha/2, n-1}) = \alpha/2$ etc. for utledn.

Det er 90% sannsynlig at

$$\mu \in [\bar{X} - t_{\alpha/2, n-1} \cdot \frac{S}{\sqrt{n}}, \bar{X} + t_{\alpha/2, n-1} \cdot \frac{S}{\sqrt{n}}]$$

g) $\sigma = 170,5$, $\bar{X} \pm u_{5\%} \cdot \frac{\sigma}{\sqrt{n}}$ [2343,7 , 2624,3]

$u_{5\%} = 1,645$

h.) Det siste intervallet er mindre fordi det er mindre usikkerhet p.g.a. tilleggsoppl. i g.. t-fordelingen er flatere enn normalf.

5/8

III.

a) $f(s) = \begin{cases} 1 & 0 \leq s \leq 1 \\ 0 & \text{ellers} \end{cases}$ ← husk å nevne $f : S \to \mathbb{R}$!

$$P[a,b] = \int_a^b f(s)\,ds$$

b) $P(A \cap B) = P(A) \cdot P(B) \Leftrightarrow$ uavh.

$P(A \cap B) = P[0, 1/4] = 1/4$

$P(A) \cdot P(B) = \frac{1}{2} \cdot \frac{1}{2} = 1/4$

$\therefore$ A og B er uavh.

g) $X : [0,1] \to \mathbb{R}$

(X er en reell funksjon definert på $S = [0,1]$.)

[$X^{-1}(a,b)$ må også alltid være en hendelse, men siden vi ikke ...]

6/8

c.) Z_A og Z_B er uavh. Bernoulli-variabler, så $Z = Z_A + Z_B$ er binomialfordelt med $p = 1/2$ og $n = 2$.

d.) $$f_{Z_A, Z_B}(0,0) = P(Z_A = 0 \,\&\, Z_B = 0) = 1/4$$
$$= f_{Z_A, Z_B}(z_A, z_B)$$

$(z_A, z_B) \in \{(0,0), (0,1), (1,0), (1,1)\}$

e.) $$F_Z(z) = P(Z_A + Z_B \le z)$$

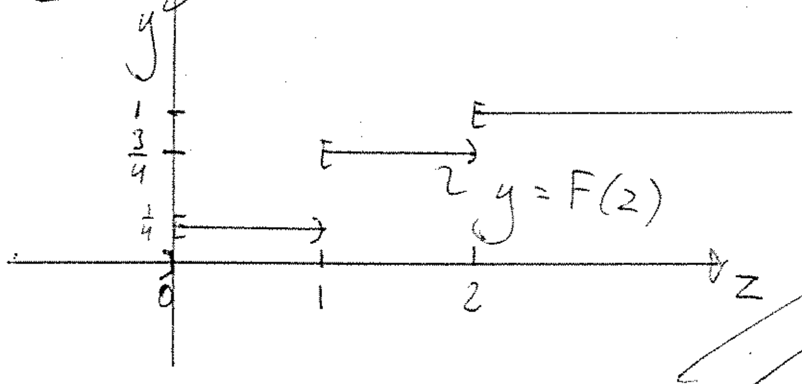
$F_Z(z) = 0 \quad z < 0$, $0 \le z < 1$: $F_Z(0) = 1/4$

$1 \le z < 2$: $F_Z(z) = F_Z(1) = 1/4 + 1/4 + 1/4 = 3/4$

$F_Z(z) = 1$

(Kontroll: $f_Z(z) = \binom{2}{z} \cdot \frac{1}{2}^z \cdot \frac{1}{2}^{2-z}$

$f_Z(0) = \frac{1}{4}$, $f_Z(1) = 2 \cdot \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{2}$, $f_Z(2) = \frac{1}{4}$)



h.)

$$P(a \le X \le b)$$

$$P\{s \in [0,1] \mid a \le X(s) \le b\} = \frac{1}{\sqrt{2\pi}} \int_a^b e^{-\frac{1}{2}x^2} \, dx$$

f.)

$$F_X(x) = P(X \le x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{1}{2}t^2} \, dt$$

$$f_X(x) = F_X'(x)$$

SLUTT

x	0	0,4	0,7	1,4
F_X	0,5	0,655	0,816	0,919

0	0,5	1,0	1,5	2,0
0,5	0,691	0,841	0,933	0,977