

Innsamling og modellering av data for analyse av militære operasjoner

Håkon Ljøgodt

Forsvarets forskningsinstitutt (FFI)

29. august 2008

FFI-rapport 2008/01591

1068

P: ISBN 978-82-464-1447-8

E: ISBN 978-82-464-1448-5

Emneord

Datainnsamling

Modellering

Meningsmålinger

Bayesianske metoder

Bootstrapmetoden

Godkjent av

Stein Malerud

Prosjektleder

Espen Skjelland

Forsknings sjef

Jan Erik Torp

Avdelingssjef

Sammendrag

Denne rapporten er skrevet som en del av prosjekt 1068: ”Metoder og modeller for analyse av freds- og lavintensitetsoperasjoner”. Hensikten med rapporten er å gi en oversikt over metoder for innsamling og modellering av data i forbindelse med analyse av militære operasjoner, inkludert freds- og lavintensitetsoperasjoner.

Beskrivelsen av metodene er kortfattet, men det henvises til aktuelle lærebøker for dem som har behov for å gå dypere inn i de ulike emnene. Det første emnet som omtales er datainnsamling med vekt på meningsmålinger siden dette er en viktig metode for å kartlegge utviklingen i konfliktområder. Deretter gjennomgås de tradisjonelle metodene for å representere data ved hjelp av statistiske fordelinger blir. Sentralt her er stegene:

- Valg av fordelingsklasse basert på tidligere erfaring og grafisk framstilling av dataene.
- Estimering av parametrene i den valgte sannsynlighetsfordelingen på grunnlag av dataene.
- Testing av den estimerte sannsynlighetsfordelingens evne til å representere dataene.

Disse stegene må om nødvendig gjentas flere ganger.

Bayesianske metoder omtales som en metode for konsistent sammenstilling av subjektiv informasjon og faktiske måleresultater. Til slutt følger en beskrivelse av Bootstrapmetoden. Dette er en datamaskinbasert metode som blant annet egner seg for beregning av konfidensintervall der dette er vanskelig med klassiske metoder.

English summary

This report has been written as a part of project 1068: 'Methods and models for the analysis of peace and low intensity operations'. The purpose of this report is to give an introduction to methods for collecting and modelling data related to military operations, including peace and low intensity operations.

The description of the methods is brief, but there are references to textbooks for those who need more information about the actual subjects. The first topic mentioned is data collection with emphasis on opinion polls since this is an important tool for assessing the desired progress in conflict areas. Then more traditional methods for representing data by means of probability distributions are briefly described. Three steps are emphasised:

- Selection of probability distribution based on experience and graphical display of the data.
- Estimation of the parameters of the selected probability distribution based on the data.
- Goodness-of-fit testing to see whether the selected probability distribution is satisfactorily representing the data.

If necessary, these steps have to be repeated.

Bayesian methods are mentioned as a tool for consistent fusion of subjective information and actual measurements. Finally, there is a description of the Bootstrap method. This is a computer intensive method well suited for estimation of confidence intervals in cases where this is difficult with classic methods.

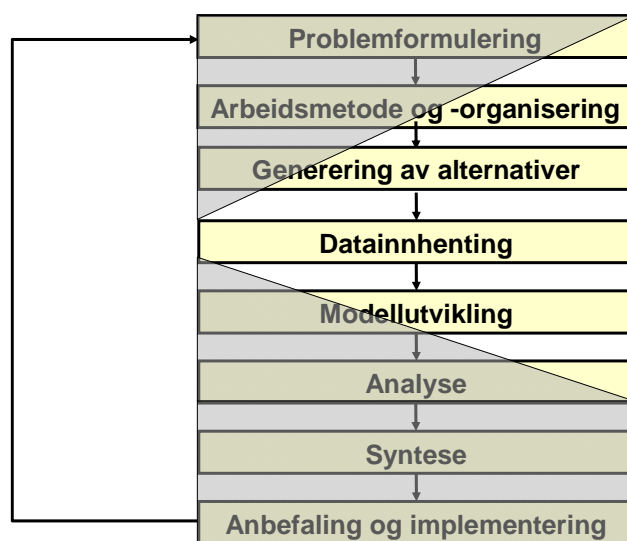
Innhold

| | | |
|----------|--|-----------|
| 1 | Innledning | 7 |
| 2 | Innsamling av data | 8 |
| 2.1 | Meningsmålinger | 9 |
| 2.1.1 | Innsamlingsmetoder | 10 |
| 2.1.2 | Utforming av spørsmål og spørreskjema | 11 |
| 2.1.3 | Måling av utvikling over tid | 12 |
| 3 | Aktuelle fordelingsklasser | 12 |
| 3.1 | Kontinuerlige fordelinger | 15 |
| 3.2 | Diskrete fordelinger | 17 |
| 4 | Parameterestimering i aktuelle fordelingsklasser | 17 |
| 5 | Tester for å undersøke om en valgt fordeling representerer dataene godt nok | 18 |
| 5.1 | Quantile-Quantile plot | 18 |
| 5.2 | Kjikkvadratføyningstesten | 19 |
| 6 | Metoder når det er få observasjoner | 20 |
| 6.1 | Bayesianske metoder | 20 |
| 6.1.1 | Bayes formel anvendt til å vurdere godheten av klassifiseringsregler | 21 |
| 6.2 | Bootstrappmetoden | 22 |
| 7 | Oppsummering | 24 |
| | Referanser | 26 |

1 Innledning

Denne rapporten er skrevet som en del av prosjekt 1068: ”Metoder og modeller for analyse av freds- og lavintensitetsoperasjoner”. Hensikten med denne rapporten er å gi en oversikt over metoder for innsamling og modellering av data i forbindelse med analyse av militære operasjoner, inkludert freds- og lavintensitetsoperasjoner. Analyse av militære operasjoner ved FFI har tradisjonelt foregått ved bruk av naturvitenskapelige metoder. Ved analyse av freds- og lavintensitetsoperasjoner vil det normalt ikke være tilstrekkelig med disse metodene alene. I tillegg til de tradisjonelle metodene vil det derfor bli gitt en beskrivelse av bayesiansk analyse og meningsmålinger som er kvantitative metoder for behandling av subjektiv informasjon.

Ved modellering er det ofte behov for å representere innsamlede data ved hjelp av sannsynlighetsfordelinger. Disse sannsynlighetsfordelingene kan så brukes til å generere input for stokastiske simuleringsmodeller eller til å beregne sannsynlighetene for begivenheter knyttet til de aktuelle stokastiske variablene. Denne rapporten er relevant for en del trinn i OA-arbeidsprosessen. Dette er illustrert i Figur 1-1 ved at de trinn som rapporten ikke behandler er skyggelagt.



Figur 1-1 De deler av OA-arbeidsprosessen som omtales i denne rapporten er fremhevet

I denne rapporten beskrives en del metoder og teknikker for å fremskaffe sannsynlighetsfordelinger som kan representere dataene. Denne prosessen vil ofte bestå av fire steg:

- Innsamling av data. Dette kan være innsamling av allerede eksisterende data eller det kan være hele prosessen med forsøksplanlegging, gjennomføring av forsøk og avsluttende bearbeiding av data.

- Valg av fordelingsklasse til å beskrive data. Ved blant annet å fremstille dataene i et histogram får man et inntrykk av hvilke fordelingsklasser som er aktuelle.
- Estimering av parametrene i den valgte fordelingsklassen. I lærebøker i statistikk vil man finne forslag til estimatorene for parametrene i de aktuelle fordelingsklassene. Programpakken i statistikk vil ofte foreta estimeringen når man anvender en "Fit-funksjon".
- Testing av hvor godt den estimerte fordelingen representerer dataene. Her er det vanlig med en visuell inspeksjon av forskjellige typer plot. Det fins også mer formelle test-observatorer som kan beregnes.

Denne standardprosessen som er skissert ovenfor forutsetter tilgang på en viss mengde data. Hvis det er vanskelig å få tak i data, kan *Subject Matter Experts* (SME) diskutere seg fram til en fordeling. Den videre modelleringen blir da innen rammen av bayesiansk statistikk. Det vil også bli omtalt en metode basert på frekventistisk (ikke-bayesiansk) statistikk for parameterestimering og usikkerhetsangivelse når man bare forutsetter identisk fordelte og uavhengige observasjoner. Denne metoden kalles "Bootstrap-metoden".

Meningsmålinger er et spesialtilfelle av utvalgsundersøkelser. Man ønsker å estimere andelen av individer i en populasjon som har en viss egenskap, uten å måtte undersøke alle individene i populasjonen. I en utvalgsundersøkelse kan man f. eks. undersøke et utvalg av trær i en skog for å estimere andelen av trær som er angrepet av barkbiller. I en meningsmåling kan man f. eks. spørre et utvalg av personer om de har tillit til politiet hvis man ønsker å estimere hvilken tillit politiet har i befolkningen.

2 Innsamling av data

Datainnsamlingen utgjør ofte en vesentlig del av arbeidet med å løse et virkelig problem. Likevel er datainnsamlingen det steget i inputmodelleringen som får minst omtale i lærebøker og som det er vanskeligst å gi en generell beskrivelse av. En viktig årsak til dette er at det i forbindelse med datainnsamling ofte dukker opp praktiske spørsmål som det kan være vanskelig å forutse. Ett moment er imidlertid felles for all datainnsamling. Før datainnsamlingen starter må man tenke nøye gjennom hvilke data som trengs for å besvare de spørsmålene man har. Man kan f. eks. ha et krav om at bredden til et konfidensintervall for en parameter man ønsker å estimere ikke må overstige en viss verdi. Ved å analysere beregningen av konfidensintervall vil en finne at dette krever at antall observasjoner må overstige et visst minimum. Siden datainnsamling ikke er gratis, må man vurdere om det opprinnelige kravet til bredden av konfidensintervallet gjør at datainnsamlingen blir for dyr. Utformingen av en datainnsamlingsplan er derfor en iterativ prosess hvor en vurderer innsamlingskostnader mot nøyaktighetskrav. Man bør også ta en siste gjennomgang av datainnsamlingsplanen for å kontrollere at de data som skal samles inn faktisk er av betydning for å kunne besvare de aktuelle spørsmålene.

I forbindelse med datainnsamling kan det ofte være nyttig å skille mellom kontrollerte eksperiment og passiv registrering av data fra et system som utvikler seg uten påvirkning av analytikerne. Dette kan illustreres med to eksempler.

2.1 Kontrollerte eksperiment

Ved valg av f. eks. håndvåpen kan en gjennomføre kontrollerte eksperiment for å se hvilket våpen som egner seg best. Man kan la et antall forsøkspersoner prøve alle de tre aktuelle håndvåpnene V_1, V_2, V_3 mot fire aktuelle mål M_1, M_2, M_3, M_4 . For at læringseffekten ikke systematisk skal påvirke resultatet, kan man for hver forsøksperson trekke rekkefølgen for skytingen med de tolv kombinasjonene av håndvåpen og mål, f. eks.: $V_2M_3, V_1M_3, \dots, V_1M_4$. Montgomery [1] gir en innføring i forsøksplanlegging.

2.2 Passiv registrering av data

Registrering av henvendelser til et legekontor illustrerer en situasjon som utvikler seg uten styring av analytikerne. I dette eksempelet er det normalt én person som tar imot innkommende samtaler, men det er kommet flere klager på lang ventetid. Det vurderes å ha to personer til å ta imot telefoner i enkelte perioder. For å vurdere når det er behov for to personer bestemmer man seg for å registrere tidspunkt for henvendelse og varighet av samtale. I den forbindelse fremkommer det en rekke spørsmål. Er det nok å registrere henvendelsene over en dag eller er det stor forskjell på ukedagene? Er det stor forskjell på antall henvendelser i f. eks. uke 24 og 42? Hvor mange må settes til å betjene telefonen i prøveperioden for å sikre at ingen avbryter henvendelsen på grunn av lang ventetid? Hva gjør man hvis man går glipp av to timer planlagt registrering? Gjør man en ny registrering på tilsvarende tid i påfølgende uke? En slik dataregistrering er kostbar. En må derfor vurdere hvorvidt et mer omfattende observasjonsmateriale er verdt de ekstra registreringskostnadene. I tillegg til drøfting av kontrollerte eksperimenter inneholder Alberts [2] også en del stoff som er relevant for passiv registrering av data.

Hvis man har et sett med levetidsdata er det viktig å undersøke om de er sensurerte. En observasjon t_i av en variabel sies å være venstresensurert hvis man bare vet at $t_i < L$ og høyresensurert hvis man bare vet at $t_i > R$ der L og R er konstanter. Hvis man er interessert i hvor mange kjøretimer en stridsvognmotor tåler før den er utslitt, vil man på et gitt tidspunkt kunne sitte med et høyresensurert datasett. For en motor som ikke er utslitt, vil man bare vite at levetiden t_i er større enn antall kjøretimer t_{i0} på innsamlingstidspunktet.

2.3 Meningsmålinger

Siden man ved utvalgsundersøkelser bare undersøker en del av populasjonen, vil estimatet man beregner være beheftet med usikkerhet. Hvis man ønsker å estimere andelen av trær som er angrepet av barkbiller i en skog, vil dette være det eneste bidraget til usikkerheten siden det er rimelig å forutsette at man ut fra en gitt definisjon kan avgjøre om et tre er angrepet av barkbiller. Ved meningsmålinger kommer andre typer usikkerheter/feilkilder i tillegg. Vedkommende som er trukket ut for intervju, kan være utilgjengelig, kan nekte å svare, kan svare slik han tror det forventes eller han kan misforstå spørsmålet.

Før de spesielle utfordringene ved meningsmålinger drøftes nærmere, kan det være naturlig å si litt om den variansen som skyldes at man bare undersøker en del av populasjonen. Anta for enkelhets skyld at man ønsker å estimere andelen p i en populasjon som har en viss egenskap. La populasjonen bestå av N individer hvorav a av disse har denne egenskapen. Man ønsker å estimere andelen $p = \frac{a}{N}$ der a er ukjent. Dette gjøres ved å trekke et tilfeldig utvalg på n individer fra populasjonen på N individer. Anta at man blant de n uttrukne individene finner at x individer har den aktuelle egenskapen. Andelen p estimeres da ved $\hat{p} = \frac{x}{n}$. I lærebøker om utvalgsundersøkelser vil en finne at et konfidensintervall for p med konfidensgrad tilnærmet lik $1 - \varepsilon$ kan uttrykkes ved

$$\left(\hat{p} - z_{1-\frac{\varepsilon}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + z_{1-\frac{\varepsilon}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \right)$$

der $z_{1-\frac{\varepsilon}{2}}$ er $(1 - \frac{\varepsilon}{2})$ -fraktilen i normalfordelingen med forventning lik 0 og varians lik 1. En

forutsetning for at denne tilnærmingen skal være god er at n ikke er for liten og at $\frac{N-n}{N-1}$ er

tilnærmet lik 1. Velges $\varepsilon = 0,05$ finner man $z_{0,975} = 1,96 \approx 2$. Dermed blir bredden på

konfidensintervallet i dette tilfellet lik $4\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$. Innsetting av noen tall vil illustrere bredden

på dette konfidensintervallet.

Hvis $\hat{p} = 0,5$ og $n = 1000$ blir konfidensintervallet lik (0,468, 0,532) og bredden lik 0,064.

Hvis $\hat{p} = 0,05$ og $n = 1000$ blir konfidensintervallet lik (0,036, 0,064) og bredden lik 0,028.

I [3] er det beskrevet alternativer til rent tilfeldig utvelgning som klyngeutvelgning og stratumdelte tilfeldig utvelgning. Under visse forutsetninger vil disse utvalgsmåtene gi mindre varians for \hat{p} enn rent tilfeldig utvelgning.

2.3.1 Innsamlingsmetoder

Et viktig moment er valg av metode for meningsmålingen. De mest kjente metodene er intervju ansikt til ansikt, intervju over telefon, utfylling av spørreskjema sendt over internett og utfylling av spørreskjema sendt via post. Egnetheten av disse metodene må vurderes blant annet opp mot kostnader, tidsbruk, svarprosent og skjevheter på grunn av at man ikke får svar fra enkelte deler av populasjonen. Det er åpenbart dyrere med en intervjuer enn med egenutfylt spørreskjema. Bruk av spørreskjema utsendt og besvart over internett gir kortest gjennomføringstid for undersøkelsen. Svarprosenten vil ofte være høyest ved telefonintervju. Man må velge metode på bakgrunn av den populasjon man vil hente informasjon om. Ønsker man å kartlegge folks erfaringer med IP-telefoni, kan det være greit å sende ut spørreskjema over internett. Ønsker man å kartlegge nordmenns holdninger til norsk EU-medlemskap er det problematisk bare å sende ut spørreskjema via internett. Dette kan imidlertid forsvares hvis tidligere undersøkelser har vist at holdningen til EU-medlemskap er den samme i gruppene "Har internetttilgang hjemme" og "Har ikke internetttilgang hjemme".

2.3.2 Utforming av spørsmål og spørreskjema

Dette er omfattende tema hvor det er vanskelig å gi noen fullstendig oppskrift. Det fins imidlertid en rekke momenter som det vil være nyttig å tenke gjennom i forbindelse med denne utformingen. I dette avsnittet vil det bli gitt eksempler på slike momenter. For en relativ fyldig fremstilling av temaet vises det til [3].

Det er viktig at oppdragsgiver og de som skal gjennomføre undersøkelsen snakker grundig sammen for å få avklart hva man ønsker å få svar på og hva det er mulig å få svar på. Etter at man er blitt enig om et utkast til spørreskjema, bør dette prøves ut på en gruppe som har mest mulig til felles med populasjonen. En slik prøve vil kunne avsløre uheldige formuleringer som fører til misforståelser eller motvilje hos dem som blir intervjuet.

Spørsmålene bør selvfølgelig være klare og entydige. ”Hva er din inntekt?” er ikke noe godt spørsmål. ”Hva er din årlige bruttoinntekt?” er en bedre formulering hvis det er dette man ønsker svar på. Ordet bruttoinntekt bør kanskje erstattes med ”inntekt før skatt er betalt”. Ordvalget må tilpasses den populasjonen en ønsker å undersøke. ”Synes du bensin- og bilavgiftene i Norge er *for høye / passe / for lave*?” er heller ikke noe godt spørsmål. Hva skal man svare hvis man mener bensinavgiften er for lav, registreringsavgiften for høy og det totale avgiftsnivået på bilhold passe?

Spørsmålene bør ikke være ledende, men dette ønsket er det ikke alltid lett å oppfylle. Følgende eksempel illustrerer dette. ”Synes du Norge gir *for mye / passe / for lite* i u-hjelp?” er et greit spørsmål. Det blir imidlertid et problem hvis den intervjuede før han svarer ønsker å vite hvor mye Norge gir. Man kan oppgi beløpet på flere ekvivalente måter som ”prosentandel av BNP”, ”totalbeløpet i milliarder kroner” eller ”i kroner per nordmann”. Måten man oppgir beløpet på vil ha innvirkning på svarfordelingen selv om måtene er ekvivalente.

Man bør i de fleste tilfellene spesifisere svaralternativene til et spørsmål. Dette gir ensartede svar som gjør den videre behandling av dataene enklere. Det gjør det også mulig for den intervjuede å vurdere svaralternativ han ellers ikke ville ha kommet på i farten. Ved mange svaralternativ kan det være behov for en samlepost ”Annet, vennligst spesifiser”. De oppførte svaralternativene bør være de man regner med vil få flest svar. Maksimalt antall svaralternativ bør ligge rundt 10. En ulempe med spesifiserte svaralternativ er at den intervjuede i noen tilfeller ikke finner svaralternativ som passer til hans oppfatning.

Når det gjelder utformingen av spørreskjemaet, bør følsomme spørsmål plasseres til slutt i tilfelle den intervjuede da velger å avbryte intervjuet. Ved et avbrudd har man sikret seg svar på de mer nøytrale spørsmålene. Beslektede spørsmål bør grupperes sammen. Antall spørsmål på et skjema kan variere, men det er viktig at det ikke tar for lang tid å besvare skjemaet. Jo lenger tid det tar å besvare skjemaet, desto lavere vil motivasjonen for å gjøre det være. Det antydes at intervju over telefon eller internett ikke bør ta mer enn et kvarter.

2.3.3 Måling av utvikling over tid

Hvis man ønsker å se hvordan oppfatninger endrer seg over tid, kan det være gunstig å intervju det samme utvalget med jevne mellomrom, f. eks. hvert år. Man må i alle fall være forsiktig med å endre formuleringen av spørsmålene. Med endring i formuleringen blir det vanskelig å skille reelle endringer i oppfatning fra de endringene som skyldes ny formulering.

Når man studerer utvikling over tid, er det viktig å ta hensyn til sesongvariasjoner. Hvis man f. eks. spør folk i Afghanistan om de har merket talibanaktivitet i nærområdet den siste uken, bør man ikke sammenligne resultatene fra et intervju i november 2007 med resultatene fra et intervju i mai 2008. En økning i rapportert aktivitet i intervjuet i mai 2008 betyr ikke nødvendigvis at Taliban har styrket seg. Det hadde gitt et riktigere bilde å sammenligne situasjonen i mai 2008 med situasjonen i mai 2007.

3 Aktuelle fordelingsklasser

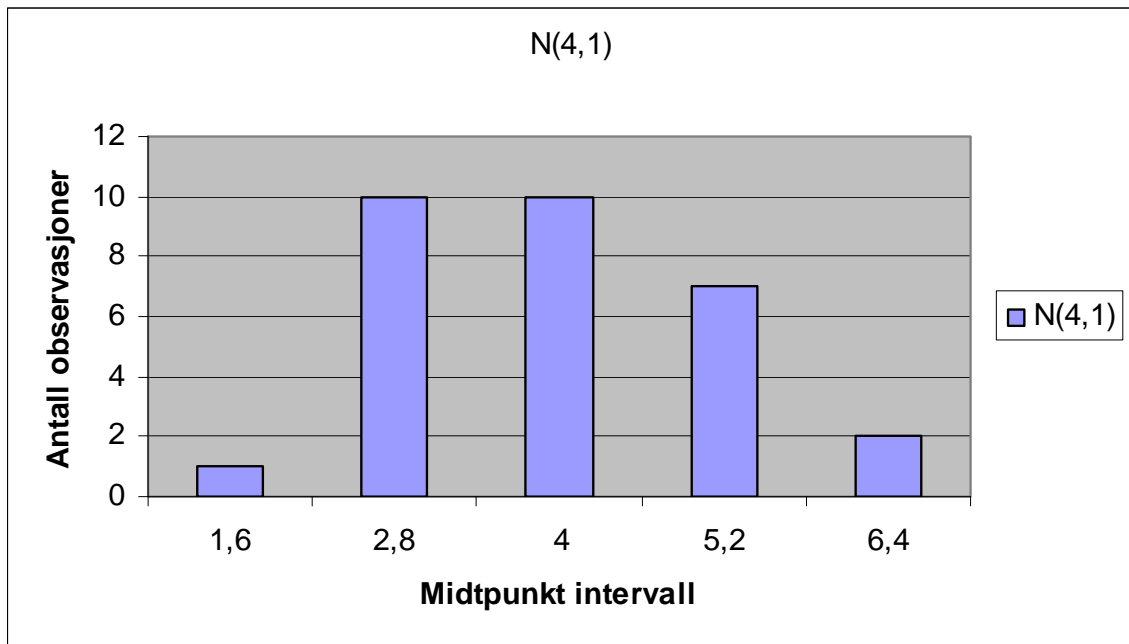
Dette kapitlet beskriver kort prosessen med å bestemme aktuelle fordelingsklasser for å representere dataene. Deretter følger noen eksempler som illustrere variasjonsmulighetene blant de kontinuerlige og diskrete fordelingene.

På bakgrunn av tidligere erfaringer med lignende data vil man i mange tilfeller ha en oppfatning av hvilke fordelingsklasser som representerer dataene på en god måte. Hvis dataene f. eks. angir tidene mellom ankomster av kunder, vil lengden av disse tidsintervallene være eksponentielt fordelt forutsatt konstant ankomstintensitet og uavhengighet mellom begivenheter i ikke overlappende tidsintervaller. Fra sentralgrenseteoremet følger det at hvis en variabel fremkommer som en sum av mange uavhengige bidrag med endelig forventning og varians, så vil fordelingen til denne summen være tilnærmet normalfordelt. Lærebøker innen de forskjellige fagfeltene (køteori, pålitelighetsteori, epidemiologi, . . .) vil ofte drøfte de mest aktuelle fordelingene innen sine områder. Law og Kelton [4] bruker ca. 100 sider på å drøfte valg av inputfordelinger.

3.1 Inspeksjon av datasettet

Før man begynner å bestemme aktuelle fordelinger vil det være nyttig å fremstille dataene i et histogram. Man vil da kunne få et godt inntrykk av formen på fordelingen. Svarene på følgende spørsmål vil i noen grad bidra til å redusere antall aktuelle fordelinger. Er fordelingen symmetrisk eller skjev? Har den lange haler? Har den tunge haler?

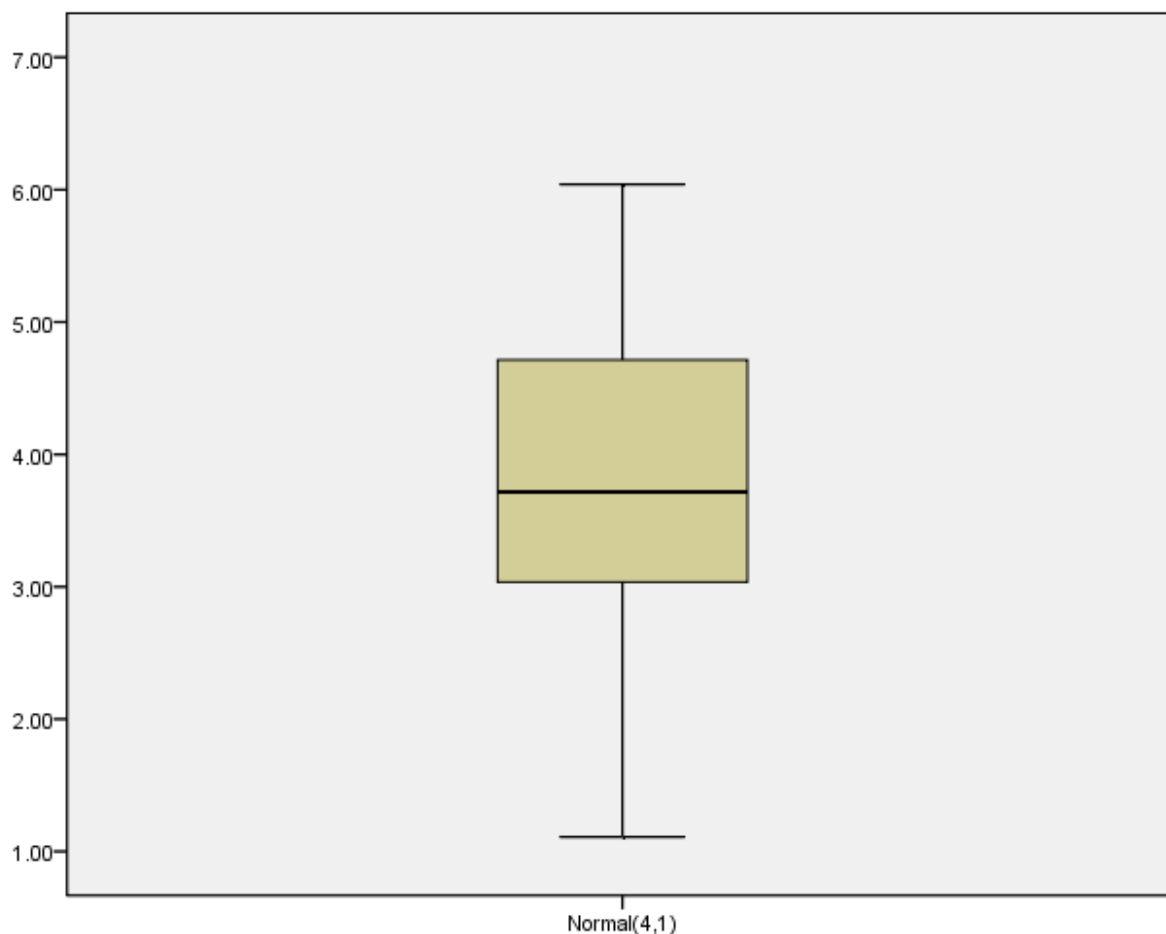
Ved fremstilling av data i et histogram må man bestemme seg for hvor mange intervallgrupper man vil ha i histogrammet. Det er ikke noe absolutt svar på dette spørsmålet, men en tommelfingerregel sier at hvis man har n observasjoner bør antall intervall, k , velges ut fra formelen $k \approx \sqrt{n}$.



Figur 3-1 Histogram av 30 observasjoner trukket fra en normalfordeling med forventning lik 4 og varians lik 1

Antall intervall i histogrammet ovenfor er valgt lik 5 siden $\sqrt{30} = 5,5$. Selv om observasjonene er trukket fra en normalfordeling, ser ikke histogrammet helt symmetrisk ut. Det visuelle inntrykket hadde vært klart mer symmetrisk hvis bare en av observasjonene i 2,8-intervallet i stedet hadde falt i 1,6-intervallet.

Fordelingen av et sett observasjoner kan også visualiseres ved et såkalt boxplot. Figur 3-2 viser et boxplot av de samme data som er fremstilt i Figur 3-1. I boxplottet markeres fem karakteristiske verdier for datasettet: minimumsverdi, første kvartil, median, tredje kvartil og maksimumsverdi. En ekstremverdi blir lett synlig i et boxplot. Når en slik ekstremverdi er påvist, er det grunn til å kontrollere om det har forekommet feil i forbindelse med målingen og registreringen av denne observerte verdien. Etter å ha laget histogram og boxplot på grunnlag av dataene, vil man som regel ha redusert antall fordelinger som kan egne seg til å representere dataene.



Figur 3-2 *Boxplot av 30 observasjoner trukket fra en normalfordeling med forventning lik 4 og varians lik 1*

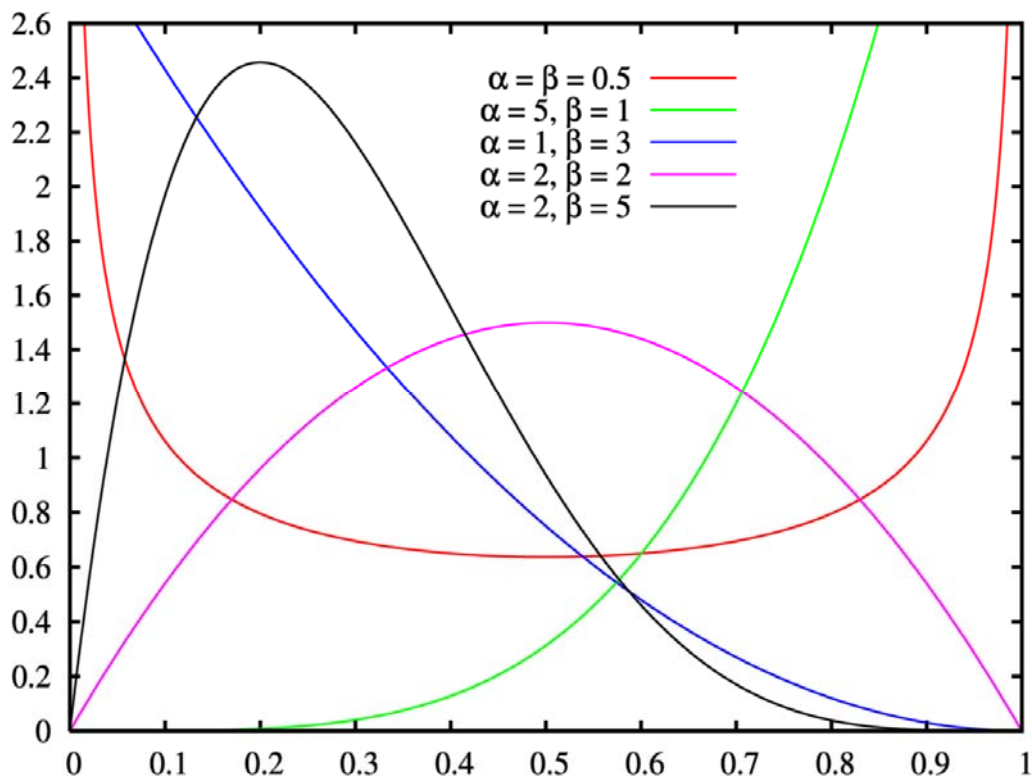
Sannsynlighetsfordelinger kan grupperes på mange måter. Hovedinndelingen er gjerne i kontinuerlige og diskrete fordelinger. I kapittel 3.2 og 3.3 vil det bli vist noen eksempler på slike fordelinger. Figurene i disse eksemplene er hentet fra Wikipedia [5] som omtaler og illustrerer et stort antall fordelinger.

Ellers grupperes fordelingene på grunnlag av verdimengden til den stokastiske variabelen. Verdimegden kan være et begrenset intervall, et intervall som er ubegrenset i den ene enden, f. eks. $[0, \infty)$, eller verdimegden kan være hele tallinjen.

Mange programpakker tilbyr en "Fit"-funksjon for bestemmelse av en fordeling for representasjon av dataene. Man bør være litt varsom med å bruke en slik "Fit"-funksjon hvis man ikke vet nøyaktig hva den gjør. Valg av fordelingsfunksjon og estimering av parametrene i den bør foregå i to trinn.

3.2 Kontinuerlige fordelinger

Blant de mest kjente kontinuerlige fordelingene er normalfordelingen, den uniforme fordelingen og eksponentialfordelingen. Som en illustrasjon av variasjonen blant de kontinuerlige fordelingene, vil en her vise noen få som ikke tilhører de mest kjente. Dette er likevel fordelinger med nyttige anvendelser. I Figur 3-3 vises sannsynlighetstettheten til betafordelingen for fem ulike valg av parametrene.



Figur 3-3 Sannsynlighetstettheter fra betafordelingen for forskjellige valg av parametrene α og β

En ser at det er mulig å få fram vidt forskjellige former av tetthetsfunksjonen ved å variere valget av parametre. Betafordelingen har intervallet $[0,1]$ som verdimengde. Eksempler på anvendelse av betafordelingen er innenfor prosjektstyringsmetoden PERT og som apriorifordeling for parameteren i en binomisk fordeling innen bayesiansk statistikk. Ved anvendelse innen prosjektstyring brukes betafordelingen til å modellere varigheten av forskjellige arbeidsoppgaver. For denne anvendelsen vil det som regel være nødvendig med en forskyvning og skalering av sannsynlighetstettheten.

Betafordelingen med parametre α og β har sannsynlighetstetthet

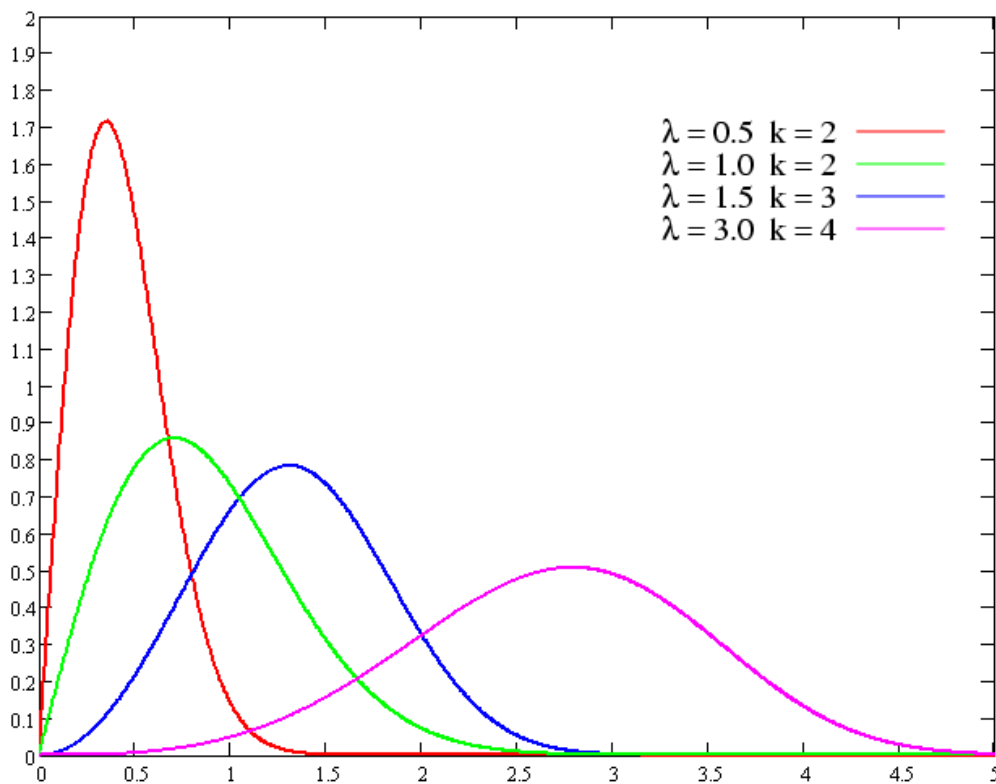
$$f(x; \alpha, \beta) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha, \beta)} \quad \text{for } 0 \leq x \leq 1$$

der betafunksjonen $B(\alpha, \beta)$ er definert ved

$$B(\alpha, \beta) = \int_0^1 y^{\alpha-1}(1-y)^{\beta-1} dy \quad \text{for reelle tall } \alpha > 0 \text{ og } \beta > 0.$$

Forventning og varians for betafordelingen er gitt ved

$$E(X) = \frac{\alpha}{\alpha + \beta} \quad \text{og} \quad \text{Var}(X) = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}$$



Figur 3-4 Sannsynlighetstettheter fra Weibullfordelingen for forskjellige valg av parametrene λ og k

I Figur 3-4 vises sannsynlighetstettheten til Weibullfordelingen for fire ulike valg av parametrene. Weibullfordelingen har intervallet $[0, \infty)$ som verdimengde. Weibullfordelingen brukes mest til modellering av overlevelsesdata. Den anvendes også til å modellere den tid det tar å utføre en arbeidsoppgave. Weibullfordelingen er under visse konvergensbetingelser også grensefordelingen for minimum av en mengde med uavhengige og identisk fordelte variable. Weibullfordelingen med parametre λ og k har sannsynlighetstetthet

$$f(x; \lambda, k) = \lambda k (\lambda x)^{k-1} e^{-(\lambda x)^k} \quad \text{for } 0 \leq x < \infty, \lambda > 0, k > 0$$

For $k=1$ reduseres Weibullfordelingen til eksponentialfordelingen. Forventning og varians for Weibullfordelingen er gitt ved

$$E(X) = \left[\Gamma\left(1 + \frac{1}{k}\right) \right] \frac{1}{\lambda} \quad \text{og} \quad \text{Var}(X) = \left[\Gamma\left(1 + \frac{2}{k}\right) - \left(\Gamma\left(1 + \frac{1}{k}\right)\right)^2 \right] \frac{1}{\lambda^2}$$

der gammafunksjonen er definert ved

$$\Gamma(x) = \int_0^{\infty} t^{(x-1)} e^{-t} dt. \quad \text{Spesielt er } \Gamma(1) = 1. \quad \text{Følgende rekursjonsformel gjelder:}$$

$$\Gamma(x) = (x-1)\Gamma(x-1). \quad \text{Dette innebærer at for heltallige } n \geq 1 \text{ så er } \Gamma(n) = (n-1)!.$$

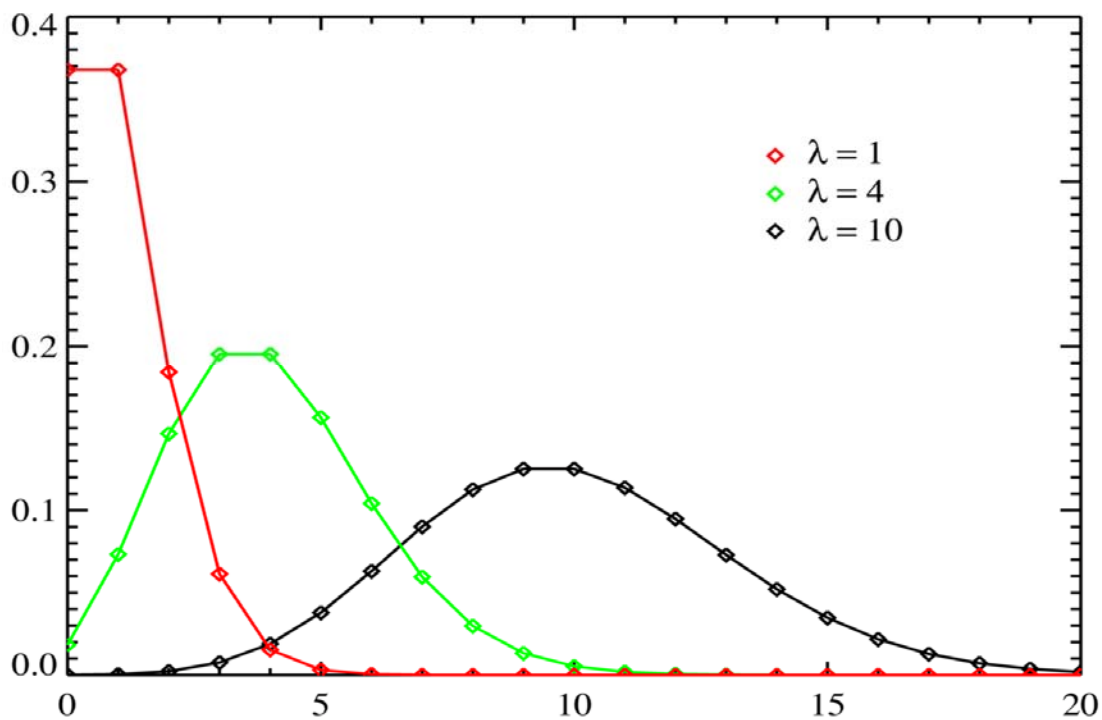
3.3 Diskrete fordelinger

Blant de mest kjente diskrete fordelingene er den binomiske, den geometriske, den hypergeometriske og Poissonfordelingen. Her vil Poissonfordelingen bli nærmere omtalt. I Figur 3-5 vises punktsannsynligheten til Poissonfordelingen for tre forskjellige valg av parameteren λ . Antall hendelser som inntreffer i et tidsintervall vil være Poissonfordelt hvis begivenhetene inntreffer med en konstant rate. Antall telefonoppringninger i en viss periode vil derfor kunne modelleres med en Poissonfordeling. Poissonfordelingen fremkommer også som en grensefordeling for den binomiske fordelingen når produktet av suksessannsynligheten og antall observasjoner holdes konstant mens man lar antall observasjoner gå mot uendelig.

Poissonfordelingen med parameter λ har punktsannsynlighet

$$f(x) = P(X = x) = \frac{e^{-\lambda} \lambda^x}{x!} \quad \text{for } \lambda > 0 \text{ og } x = 0, 1, 2, \dots$$

Forventning og varians er begge lik λ .



Figur 3-5 Punktsannsynligheter for tre Poissonfordelinger med parametre henholdsvis lik 1, 4 og 10

4 Parameterestimering i aktuelle fordelingsklasser

Når man har bestemt seg for hvilken fordelingsklasse man mener representerer dataene, må man estimere parametrene i den valgte fordelingsklassen. Sannsynlighetsmaksimeringsmetoden er den mest anvendte estimeringsmetoden. Andre vanlige estimeringsmetoder er minste kvadraters metode og momentmetoden. Estimatorene for parametrene i de vanlige fordelingsklassene finnes i mange lærebøker, se f. eks. Banks [6].

I mange tilfeller fins det eksplisitte formler for estimatorene slik at estimatene kan beregnes ved å sette de observerte verdiene inn i formlene. Weibullfordelingen er et eksempel på en fordeling hvor sannsynlighetsmaksimeringsmetoden ikke gir noen eksplisitt formel for estimatorene. I dette tilfellet må man ty til numerisk løsning av en ikke-lineær ligning for å finne estimatene for parametrene.

Som et eksempel på estimeringen, vil en nå se på normalfordelingen. La $X_1, X_2, X_3, \dots, X_n$ være n uavhengige og identisk fordelte stokastiske variable med forventning μ og varians $\sigma^2 = \tau$. Da er sannsynlighetsmaksimeringsestimatorene $\hat{\mu}$ og $\hat{\tau}$ gitt ved:

$$\hat{\mu} = \bar{X} = \frac{1}{n} \sum_{i=1}^n X_i \quad \text{og} \quad \hat{\tau} = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2$$

$\hat{\mu}$ er forventningsrett, mens $\hat{\tau}$ trenger en liten justering for å bli det.

$$\tilde{\tau} = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2 \quad \text{er en forventningsrett estimator for } \tau = \sigma^2.$$

5 Tester for å undersøke om en valgt fordeling representerer dataene godt nok

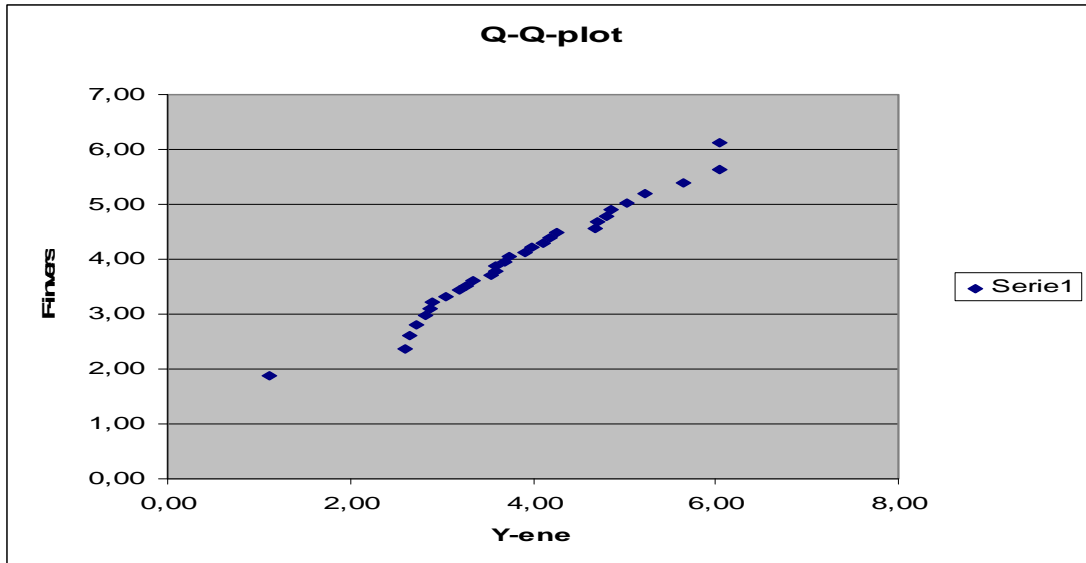
Etter at parametrene i den valgte fordelingsklassen er estimert, bør man undersøke hvor godt den konkrete fordelingen representerer dataene. En vil her omtale to metoder for å undersøke graden av tilpasning. Den ene metoden er en grafisk metode som kalles Q-Q-plot (Quantile-Quantile). Den andre metoden er kjikvadratføyningstesten som representerer en formalisert sammenligning mellom histogrammet og sannsynlighetstettheten/punktsannsynligheten. Kolmogorov-Smirnov-testen er en metode som har høyere teststyrke enn kjikvadratføyningstesten i de tilfellene den kan anvendes. Ved testing av tilpasning for diskrete fordelinger med Kolmogorov-Smirnov-testen, er det komplisert å beregne fraktilene i fordelingen til den aktuelle testobservatoren. Alle disse metodene er nærmere beskrevet i Law og Kelton [4].

5.1 Quantile-Quantile-plot

La F være en kumulativ fordelingsfunksjon (CDF på engelsk). Spørsmålet er om F beskriver fordelingen til den stokastiske variabelen X . F vil i denne sammenheng være den fordelingen som man beskrev estimeringen av i kapittel 4.

La $x_1, x_2, x_3, \dots, x_n$ være observasjoner av X . La videre $y_1, y_2, y_3, \dots, y_n$ betegne de ordnede x_i -ene, dvs. $y_1 \leq y_2 \leq y_3 \leq \dots \leq y_n$. Hvis nå $x_1, x_2, x_3, \dots, x_n$ stammer fra X med fordeling F vil

$F^{-1}\left(\frac{j-\frac{1}{2}}{n}\right)$ være nær y_j for $j=1, 2, 3, \dots, n$. Plottet av y_j mot $F^{-1}\left(\frac{j-\frac{1}{2}}{n}\right)$ vil tilnærmet ligge på en rett linje.



Figur 5-1 Q-Q-plot av 30 observasjoner fra en normalfordeling med forventning lik 4 og standardavvik lik 1

Figur 5-1 viser et Q-Q-plot av de samme dataene som tidligere er vist i Figur 3-1 og Figur 3-2. Siden man her vet hvilken fordeling som genererte observasjonene er fordelingen F i Q-Q-plottet en normalfordeling med forventning lik 4 og standardavvik lik 1. Hadde man fulgt estimeringsmåten som er beskrevet i kapittel 4, ville fordelingen F i Q-Q-plottet vært en normalfordeling med forventning lik 3,87 og standardavvik lik 1,12. Formen på et slikt plott ville ikke ha avviket meget fra det som er vist i Figur 5-1. Q-Q-plottet i Figur 5-1 illustrerer at punktene i et Q-Q-plot må avvike en del fra en rett linje før man må forkaste den anvendte fordelingen. Man kan faktisk få et avvik som i Figur 5-1 selv om man anvender den ”sanne” fordelingen F . Avvik fra en rett linje på midten av figuren er en sterkere indikasjon på dårlig tilpasset fordeling enn avvik på kantene.

5.2 Kjikvadratføyningstesten

Kjikvadratføyningstesten sammenligner overensstemmelsen mellom histogrammet over observasjonene og den estimerte sannsynlighetstettheten eller punktsannsynligheten. Kjikvadratføyningstesten gjennomføres på følgende måte. Variasjonsområdet for observasjonene deles inn i k intervaller $[c_0, c_1), [c_1, c_2), \dots, [c_{k-1}, c_k)$. La N_j være antall observasjoner i intervallet $[c_{j-1}, c_j)$, $j=1, 2, \dots, k$. La \hat{f} betegne den estimerte sannsynlighetstettheten for det kontinuerlige tilfellet og den estimerte punktsannsynligheten for det diskrete tilfellet.

$$\text{Beregn } p_j = \begin{cases} \int_{c_{j-1}}^{c_j} \hat{f}(x) dx & \text{for en kontinuerlig fordeling} \\ \sum_{c_{j-1} \leq x_j < c_j} \hat{p}(x_i) & \text{for en diskret fordeling} \end{cases}$$

La n være det totale antall observasjoner. Testobservatoren blir da:

$$\chi^2 = \sum_{j=1}^k \frac{(N_j - np_j)^2}{np_j}$$

Valget av fordeling forkastes når χ^2 er større enn øvre ε -fraktil i kjikvadratfordelingen med $k-r-1$ frihetsgrader. Her betegner r antall parametre som er estimert i den aktuelle fordelingen. Det foreligger noen retningslinjer for hvordan man deler inn variasjonsområdet for observasjonene i k intervaller. Intervallene og k bør velges slik at $np_j \geq 5$ for alle j . Intervallene behøver ikke være like lange, men størrelsen på np_j bør ikke variere for mye med j .

Det advares mot å bruke kjikvadratføyningstesten hvis det er færre enn 20 observasjoner. Med så få observasjoner skal den valgte fordelingen være svært uegnet for at den skal bli forkastet. Med et stort antall observasjoner er heller ikke testen særlig egnet da nesten alle valg av fordeling vil bli forkastet. Antall observasjoner som må til for at dette inntreffer, vil variere. Med noen hundre observasjoner er det enkelte som anbefaler å bruke den empiriske fordelingen i det videre arbeidet og dermed sløyfe hele prosessen med å representere datasettet med en kjent fordeling.

6 Metoder når det er få observasjoner

I noen tilfeller har man for få observasjoner til at man kan trekke slutninger ved hjelp av tradisjonelle metoder. I det følgende beskrives derfor bayesiansk statistikk og Bootstrapmetoden. Disse metodene er ikke primært metoder for tilfeller med få observasjoner, men de kan være til nytte i slike situasjoner.

Bayesiansk statistikk kombinerer a priori informasjon om problemet med den informasjonen som fins i de observerte data. Aprioriinformasjonen uttrykkes som regel ved apriorifordelinger som eksperter på det aktuelle området utarbeider. Hovedanvendelsen av Bootstrapmetoden synes å være estimering av varians for estimatorene som det er vanskelig å finne fordelingen til. Disse estimerte variansene kan brukes til å finne konfidensintervall for de parametrene som estimeres.

6.1 Bayesianske metoder

Bayesianske metoder har sitt navn etter Thomas Bayes (1702-1761) som var den første til å beskrive teoremet som har fått hans navn. En enkel form av Bayes' teorem uttrykt ved betingede sannsynligheter er som følger. La B_1, B_2, \dots, B_n være disjunkte begivenheter som utfyller utfallsrommet. La $P(A|B_j)$ betegne den betingede sannsynligheten for begivenheten A gitt at begivenheten B_j har inntruffet. Bayes teorem gir de betingede sannsynlighetene for B_j -ene gitt at A har inntruffet ved formelen

$$P(B_j | A) = \frac{P(A | B_j)P(B_j)}{\sum_{j=1}^n P(A | B_j)P(B_j)}.$$

Her kalles $P(B_j)$ -ene for apriorisannsynligheter mens $P(B_j|A)$ -ene kalles aposteriorisannsynligheter. Teoremet sier altså at aposteriorisannsynligheten er proporsjonal med apriorisannsynligheten og $P(A|B_j)$.

Bayesiansk statistikk skiller seg fra tradisjonell frekventistisk statistikk ved at parameterne antas å ha en sannsynlighetsfordeling. Hvis man antar at både parameterne og observasjonene er kontinuerlige, kan Bayes teorem i denne konteksten formuleres som

$$f(\theta | x_1, x_2, \dots, x_n) = \frac{L(x_1, x_2, \dots, x_n | \theta)g(\theta)}{\int L(x_1, x_2, \dots, x_n | \theta)g(\theta)d\theta}$$

Her er $g(\theta)$ aprioritettheten til parameteren θ . Den gir uttrykk for hvor sannsynlig man oppfatter ulike verdier av θ å være før man har gjort noen observasjoner. $L(x_1, x_2, \dots, x_n | \theta)$ er "likelihoodfunksjonen" for de observerte verdiene x_1, x_2, \dots, x_n når parameteren er lik θ . Aposteriorifordelingen $f(\theta | x_1, x_2, \dots, x_n)$ gir uttrykk for hvor sannsynlig man oppfatter ulike verdier av θ å være etter at man har observert verdiene x_1, x_2, \dots, x_n . Disse observerte verdiene x_1, x_2, \dots, x_n gjør at man endrer sin oppfatning av fordelingen av θ fra $g(\theta)$ til $f(\theta | x_1, x_2, \dots, x_n)$. Bayes teorem formaliserer altså hvordan en subjektiv oppfatning om θ uttrykt ved $g(\theta)$ endres i lys av observasjonene. Styrken ved den bayesianske tilnæringsmåten er at den muliggjør en konsistent sammenstilling av subjektiv informasjon og faktiske måleresultater.

For de fleste valg av apriorifordeling er det ikke mulig å beregne noen eksplisitt matematisk form på aposteriorifordelingen. Dette gjorde at Bayes teorem lenge hadde begrenset praktisk anvendelse. Tilgang på stadig mer regnekraft gjorde at det fra 1990-tallet ble praktisk mulig å beregne aposteriorifordelingene. Simuleringsmetoder som MCMC (Markov Chain Monte Carlo) har vist seg spesielt nyttige. Bolstad [7] gir en innføring i bayesiansk statistikk. Dahl [8] beskriver hvordan bayesianske metoder med MCMC er brukt til å estimere driftskostnadsvekst i Forsvaret.

6.1.1 Bayes formel anvendt til å vurdere godheten av klassifiseringsregler

I det følgende brukes et fiktivt eksempel fra Afghanistan for å illustrere anvendelsen av Bayes formel. De norske styrkene i Faryab har over tid erfart at sannsynligheten for at en tilfeldig valgt mann er talibantilhenger er 0,01. Deres erfaring tilsier videre at sannsynligheten for at en tilfeldig valgt mann er passiv talibantilhenger er 0,26 og at sannsynligheten for at han er talibanmotstander er 0,73. For å kunne plukke ut talibantilhengere for kontroll og våpenbeslag, har de merket seg en del kjennetegn som forekommer hyppigere hos talibantilhengerne enn hos de andre gruppene. La $A =$ "Mannen er pashtun med langt skjegg og har mobiltelefon."

$B_1 =$ "Mannen er talibantilhenger."

$B_2 =$ "Mannen er passiv talibantilhenger."

$B_3 =$ "Mannen er talibanmotstander."

Erfaringer fra tidligere kontroller tilsier at $P(A/B_1) = 0,6$, $P(A/B_2) = 0,1$ og $P(A/B_3) = 0,02$. Man er nå interessert i å få vite om det er stor sjanse for å treffe på en talibantilhenger hvis man plukker ut en person med kjennetegnet A . Bayes formel gir nå

$$P(B_1/A) = \frac{P(A/B_1) * P(B_1)}{P(A/B_1) * P(B_1) + P(A/B_2) * P(B_2) + P(A/B_3) * P(B_3)}$$

Ved innsetting får man:

$$P(B_1/A) = \frac{0,6 * 0,01}{0,6 * 0,01 + 0,1 * 0,26 + 0,02 * 0,73} = 0,13$$

Tilsvarende blir $P(B_2/A) = 0,56$ og $P(B_3/A) = 0,31$.

En ser at sannsynligheten for å plukke ut en talibantilhenger er 13 ganger større når man plukker ut en mann men kjennetegn A enn når man plukker ut en tilfeldig mann. Denne sannsynligheten på 0,13 er likevel lav sammenlignet med sannsynligheten for å plukke ut en talibanmotstander (0,31) eller sannsynligheten for å plukke ut en passiv talibantilhenger (0,56). Selv om sannsynligheten for kjennetegnet A er betydelig mindre i gruppen talibanmotstandere enn i gruppen talibantilhengere, vil det at motstandergruppen er så mye større gjøre at det er flere motstandere med kjennetegnet A enn det er tilhengere med kjennetegnet A . Resultatene fra Bayes formel kan her hjelpe en beslutningstager med å vurdere nytten av en foreslått klassifiseringsregel.

6.2 Bootstrapmetoden

Denne omtalen av Bootstrapmetoden er i hovedsak basert på Chernick [9]. Bootstrapmetoden er en ikke-parametrisk metode og forutsetter bare at observasjonene kommer fra identisk fordelte og uavhengige variable. Det er ingen krav om at variablene f. eks. skal være symmetriske eller normalfordelte. Bootstrapmetoden tilhører en klasse av metoder som baserer seg på gjentatte utvalg (resampling) fra de opprinnelige data. Metoden omtales ofte som beregningstung (computerintensive). Artikkelen som representerer begynnelsen for Bootstrapmetoden, ble publisert av Bradley Efron i 1979. En av de store suksessene for metoden er anvendelsen på *the two-class discrimination problem*. Klassifiseringen av et objekt som ”mål” eller ”narremål” er et typisk eksempel på et slikt problem.

Nedenfor følger en kort beskrivelse av Bootstrapmetoden med beregning av konfidensintervall som et eksempel på anvendelse. Utgangspunktet for metoden er et utvalg på n uavhengige og identisk fordelte stokastiske variabler eller vektorer X_1, X_2, \dots, X_n . Man estimerer en parameter θ med $\hat{\theta} = g(X_1, X_2, \dots, X_n)$. Man ønsker å estimere variansen for $\hat{\theta}$ og finne et konfidensintervall for θ . Blant de observerte verdiene x_1, x_2, \dots, x_n trekkes det et tilfeldig utvalg på n med tilbakelegging. La resultatet være $x_{11}, x_{12}, \dots, x_{1n}$. Det er vanlig med en utvalgsstørrelse på n , men det er ikke noe krav. Man beregner så estimatet $\tilde{\theta}_1$ ved

$\tilde{\theta}_1 = g(x_{11}, x_{12}, \dots, x_{1n})$. Stegene ”trekning av n observasjoner med tilbakelegging” og ”beregning av tilhørende estimat” gjentas k ganger. Typiske verdier for k er 500, 1000, 5000. Man får da k

estimerer for θ : $\tilde{\theta}_1, \tilde{\theta}_2, \dots, \tilde{\theta}_k$. La $\bar{\theta} = \frac{1}{k} \sum_{j=1}^k \tilde{\theta}_j$

Da kan variansen til $\hat{\theta}$ estimeres ved $\frac{1}{k} \sum_{j=1}^k (\tilde{\theta}_j - \bar{\theta})^2$

Hvis man forutsetter at $\tilde{\theta}_j$ -ene er tilnærmet normalfordelte, vil et konfidensintervall for θ med konfidensgrad på ca 95 % være gitt ved

$$\bar{\theta} \pm 2 \sqrt{\frac{1}{k} \sum_{j=1}^k (\tilde{\theta}_j - \bar{\theta})^2}$$

Man vil oppnå et mer robust konfidensintervall for θ med konfidensgrad tilnærmet lik 95 % ved å ordne estimatene $\tilde{\theta}_1, \tilde{\theta}_2, \dots, \tilde{\theta}_k$ i stigende rekkefølge og velge 2,5 % persentilen som nedre grense for konfidensintervallet og 97,5 % persentilen som øvre grense. Dette konfidensintervallet trenger ikke være symmetrisk om $\bar{\theta}$. Den angitte konfidensgraden gjelder asymptotisk, dvs. at avviket mellom den faktiske og den angitte konfidensgraden går mot null når antall observasjoner går mot uendelig.

For å gi en enkel illustrasjon av Bootstrapmetoden ble det trukket 7 observasjoner fra en normalfordeling med forventning lik 2 og varians lik 1. Resultatet angitt med to desimaler ble: 1,41, 2,88, 1,55, -0,12, 1,24, 2,62, 1,99. Målet er å bestemme et konfidensintervall for forventningen. Siden observasjonene er normalfordelte, vil man i lærebøker i statistikk finne at konfidensintervallet med eksakt konfidensgrad 0,95 er gitt ved:

$$\left(\bar{x} - t_{0,025,n-1} \frac{s}{\sqrt{n}}, \bar{x} + t_{0,975,n-1} \frac{s}{\sqrt{n}} \right)$$

der n er antall observasjoner, t angir fraktiler i Student t -

fordeling med $n-1$ frihetsgrader og s er gitt ved $s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$. Setter man inn

observasjonene ovenfor får man konfidensintervallet (1,65 - 2,45 * 0,38, 1,65 + 2,45 * 0,38) som utregnet blir (0,72, 2,58). Dette konfidensintervallet har en bredde på 1,86.

Ved bruk av Bootstrapmetoden vil man på grunn av de tilfeldige trekningene få litt forskjellige intervall ved gjentatte anvendelser på de samme observasjonene. For observasjonene ovenfor ble det foretatt tre beregninger av konfidensintervall hvor hver beregning var basert på $k = 1000$ trekninger av observasjonsvektoren. De beregnede gjennomsnittene av elementene i de utrukne vektorene ble ordnet i stigende rekkefølge og nedre grense for konfidensintervallet ble satt lik gjennomsnitt nummer 25 mens øvre grense ble satt lik gjennomsnitt nummer 975. Resultatet er vist i Tabell 6-1.

| Nedre grense | Øvre grense | Midtpunkt | Bredde |
|--------------|-------------|-----------|--------|
| 0,96 | 2,31 | 1,64 | 1,35 |
| 0,92 | 2,31 | 1,62 | 1,39 |
| 0,94 | 2,34 | 1,64 | 1,40 |

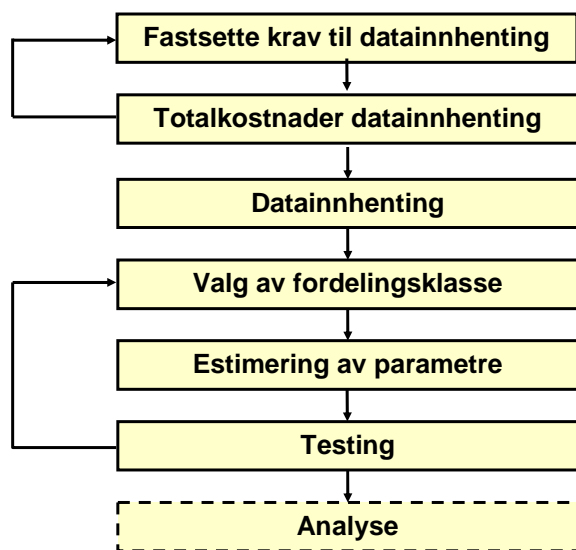
Tabell 6-1 Tre konfidensintervall med konfidensgrad tilnærmet lik 95 % beregnet med Bootstrapmetoden på et felles observasjonssett

En ser at midtpunktene for disse intervallene er litt mindre enn gjennomsnittet av observasjonene som er 1,65. Bredden av intervallene varierer litt. En større variasjon i midtpunkt og bredde vil være en indikasjon på at antall estimater, k , for forventningen er valgt for lite. Bredden på disse intervallene er bare ca 75 % av bredden på konfidensintervallet basert på forutsetningen om normalfordeling. Det tilsier at konfidensgraden for Bootstrapintervallene i dette eksempelet er noe mindre enn 95 %.

7 Oppsummering

I Figur 1-1 ble det vist hvilke deler av OA-arbeidsprosessen som behandles i denne rapporten. Figur 7-1 gir en utdypet beskrivelse av disse delene av OA-arbeidsprosessen. Rapporten tar sikte på å gi en oversikt over metoder for innsamling og modellering av data, supplert med henvisninger til kilder hvor man vil finne en grundigere behandling av de forskjellige temaene.

Rapporten understreker at man bør tenke grundig gjennom hele prosessen fra fastsettelse av datakravene til gjennomføring av analysen før man setter i gang med datainnsamling. Som vist i Figur 7-1 er dette en iterativ prosess.



Figur 7-1 Figuren utdypet de deler av OA-arbeidsprosessen som behandles i denne rapporten.

Prosjekt 1068 fokuserer på freds- og lavintensitetsoperasjoner. Ved slike operasjoner har det stor interesse å kartlegge endringer i folks holdninger. For eksempel vil økt tillit til myndighetene hos befolkningen indikere at de tiltakene som er iverksatt har hatt en ønsket virkning. Kartlegging av folks holdninger foregår som regel ved hjelp av meningsmålinger. Rapporten omtaler derfor en del viktige momenter man må ta stilling til i forbindelse med meningsmålinger.

De tradisjonelle metodene for å representere data ved hjelp av statistiske fordelinger blir kort gjennomgått. Sentralt her er stegene:

- Valg av fordelingsklasse basert på tidligere erfaring og grafisk framstilling av dataene.
- Estimering av parametrene i den valgte sannsynlighetsfordelingen på grunnlag av dataene.
- Testing av den estimerte sannsynlighetsfordelingens evne til å representere dataene.

Disse stegene er vist i Figur 7-1 som deler av en iterativ prosess.

Rapporten omtaler også bayesianske metoder fordi de gjør det mulig å sammenstille subjektiv informasjon og faktiske måleresultater på en konsistent måte. Videre er Bootstrapmetoden tatt med som en beregningstung metode som egner seg for beregning av konfidensintervaller der dette er vanskelig med klassiske metoder.

7.1 Videre arbeid

FFIs analyser av militære operasjoner innebærer som regel innsamling og modellering av data. Disse prosessene er i varierende grad dokumentert. Ved å utnytte kunnskap fra slike prosesser i en mer helhetlig tilnærming til datainnhenting, kan effektiviteten og kvaliteten i FFIs analysearbeid styrkes ytterligere. En naturlig videreføring av dette arbeidet vil derfor være å integrere erfaringer med innsamling og modellering av data i dette helhetlige og teoretisk baserte perspektivet.

Referanser

- [1] D. C. Montgomery, *Design and Analysis of Experiments* John Wiley & Sons, Inc., 2001.
- [2] D. Alberts and R. Hayes, "Code of Best Practice for Experimentation," DoD Command and Control Research Program, 2002.
- [3] L. M. Rea and R. A. Parker, *Designing and conducting survey research: a comprehensive guide* Jossey-Bass, 2005.
- [4] A. M. Law and W. D. Kelton, *Simulation Modelling and Analysis - 3rd ed.* McGraw-Hill, 2000.
- [5] Wikipedia, "http://en.wikipedia.org/wiki/Statistical_distributions."
- [6] J. Banks, J. S. Carson II, B. L. Nelson, and D. M. Nicol, *Discrete-Event System Simulation* Pearson Prentice Hall, 2005.
- [7] W. M. Bolstad, *Introduction to Bayesian Statistics* John Wiley & Sons, Inc., Hoboken, New Jersey, 2007.
- [8] F. A. Dahl, "Driftskostnadsvekst estimert med Bayesianske metoder," Forsvarets forskningsinstitutt, FFI rapport 2005/01676, 2005.
- [9] M. R. Chernick, *Bootstrap Methods* John Wiley & Sons, Inc., Hoboken, New Jersey, 2008.