Orientering
fra Miljøstyrelsen

Nr. 7 1989

Industrielle forurenings-undersøgelser

Miljøstyrelsen

Industrielle forurenings-undersøgelser

Indførelsen af miljømæssigt og økonomisk bedst tilgængelig teknologi forudsætter ofte, at ny viden erhverves målrettet fra data. Dette kræver statistisk dataplanlægning og -analyse. Rapporten redegør for de generelle principper, begreber og retningslinjer, som undersøgelser må respektere: Den trinvise gennemførelse af undersøgelser, datagrundlagets kvalitet og kvantitet, den iterative dataanalyse, den korrekte fortolkning af statistiske nøglebegreber samt betydningen af sober rapportering.

Miljøministeriet

Miljøstyrelsen

Strandgade 29, 1401 København K, tlf. 31 57 83 10

Pris kr. 90.- inkl. 22% moms

ISSN nr. 0107-2722
ISBN nr. 87-503-7469-9

Orientering fra Miljøstyrelsen

1985
Nr. 1 : Badevand og strandkvalitet
Nr. 2 : Luftforurening fra motorkøretøjer
Nr. 3 : Emneordsliste til databasen MILJØLITT
Nr. 4 : Aktuelle emner : NPO-handlingsplan og tilsyn m.v.
Nr. 5 : Miljøafgifter
Nr. 6 : Miljøkreditrådets årsberetning 1984
Nr. 7 : Forbrug og forurening med arsen, crom, cobalt og nikkel
Nr. 8 : Kemikalieaffald - fritagelse for aflevering
Nr. 9 : Oprensningsteknik ved oliespild på land

1986
Nr. 1 : Oversigt over godkendte bekæmpelsesmidler 1986
Nr. 2 : Om dioxinhandlingsplan, tilsyn og andet
Nr. 3 : Badevand og strandkvalitet
Nr. 4 : Miljøkreditrådets årsberetning 1985
Nr. 5 : Genanvendelsesrådets årsberetning 1985
Nr. 6 : Miljølovsrevision og ministerredegørelser
Nr. 7 : Mikrobiologiske vand- og miljøundersøgelser 1985

1987
Nr. 1 : Oversigt over godkendte bekæmpelsesmidler 1987
Nr. 2 : Badevand og strandkvalitet
Nr. 3 : Miljøstraffesager I
Nr. 4 : Miljøgodkendelser og forsøgsanlæg
Nr. 5 : Vandmiljøhandlingsplan og tilsynsundersøgelse
Nr. 6 : Genanvendelsesrådets årsberetning 1986
Nr. 7 : Miljøstraffesager II

1988
Nr. 1 : Oversigt over godkendte bekæmpelsesmidler 1988
Nr. 2 : Badevand
Nr. 3 : Fælleskommunale affaldsselskaber
Nr. 4 : Genanvendelsesrådets årsberetning 1987
Nr. 5 : Miljøtilsyn 1987

1989
Nr. 1 : Den kommunale affaldsplan
Nr. 2 : Miljøstraffesager III
Nr. 3 : Skærpede udstødningsnormer for biler
Nr. 4 : Oversigt over godkendte bekæmpelsesmidler
Nr. 5 : Badevand
Nr. 6 : Kemikaliefri ukrudtsbekæmpelse i grønne områder
Nr. 7 : Industrielle forureningsundersøgelser

Orientering fra Miljøstyrelsen
Nr. 7 1989

Industrielle forurenings-undersøgelser

Planlægning og analyse af data

Lars Pallesen

Miljøministeriet
Miljøstyrelsen

Indhold

Forord

Motivationen for denne orientering fra Miljøstyrelsen er at angive anvisninger på, hvorledes miljøundersøgelser af forurenende industrielle anlæg gennemføres på effektiv vis.

Den statistiske dataplanlægning og dataanalyse i sådanne miljøstudier vil ofte udgøre det saglige grundlag for indførelse af miljømæssig og økonomisk bedre teknologi, herunder også renere teknologi.

Rådet vedrørende genanvendelse og mindre forurenende teknologi har med en bevilling fra »Udviklingsprogrammet for Renere Teknologi« ønsket at tilvejebringe en rapport, der er udformet som en håndbog i dataplanlægning og -analyse. Rapporten beskriver det teoretiske grundlag for afgørelsen af, hvorvidt én teknologi kan karakteriseres som mere fordelagtig end en anden. Videre giver rapporten anvisning på, hvorledes man i praksis griber spørgsmålet an i konkrete undersøgelser.

Udover sit primære sigte er det imidlertid også håbet, at den foreliggende rapport vil kunne fungere som lærebog og opslagsværk for de mange teknikere, der til daglig arbejder mere generelt med miljøundersøgelser.

Rapportens forfatter er civilingeniør Lars Pallesen, Ph. D., som i en lang årrække har arbejdet med miljøproblemer i teori og praksis, herhjemme og i udlandet. Han er tidligere institutbestyrer for Instituttet for Matematisk Statistik og Operationsanalyse (IMSOR) ved Danmarks Tekniske Højskole og arbejder nu i industrien.

Kontoret for Renere Teknologi
Miljøstyrelsen

1. Introduktion

Baggrund og sigte

Nærværende rapport er en udredning af dataplanlægning og dataanalyse i miljøstudier specielt med henblik på forurenende industrielle anlæg.

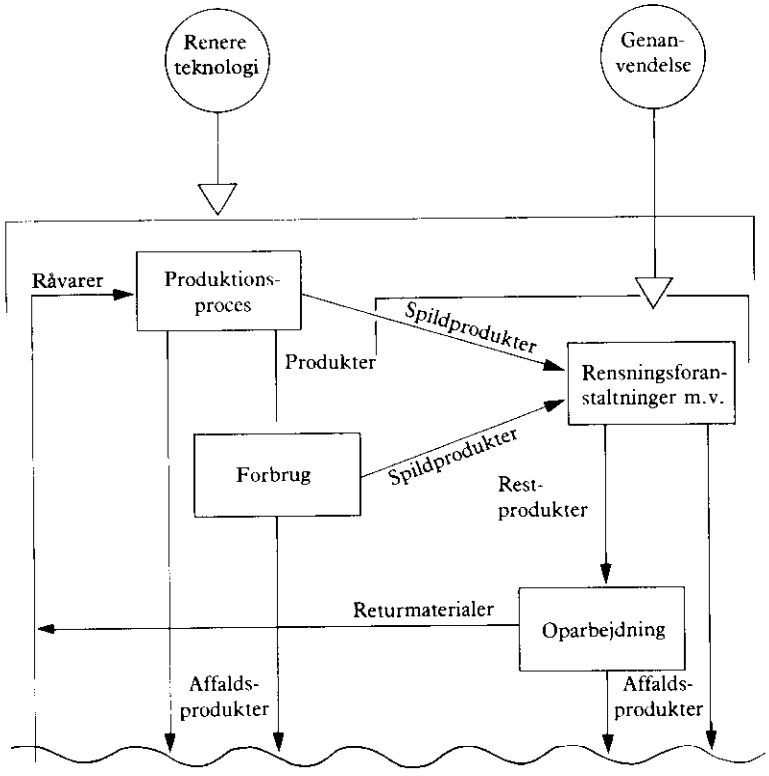
Den industrielle og samfundsmæssige udvikling har betydet, at erhvervelsen af nødvendig viden angående miljøforurening har svært ved at holde trit med det erkendte behov for beslutninger angående vejledninger og indgreb.

Sådanne beslutninger bør i størst muligt omfang bygge på viden, indsigt og fakta, og i mindst muligt omfang på fornemmelser, sporadiske observationer eller usubstantieret frygt.

Der eksisterer således et presserende behov for at den læreproces, hvori relevant viden opbygges, effektiviseres mest muligt; d.v.s. at de observerede eller eksperimentelle data, som ny bedre indsigt skal udspringe af, må være planlagt så de indeholder mest mulig relevant information og bliver analyseret så mest muligt af deres informationsindhold bliver ekstraheret.

Data og information

Data er ikke information. Data kan indeholde information. Planlægningsopgaven er at økonomisere, så man får mest information, ikke nødvendigvis flest data, fra en given undersøgelses dataindsamlende fase.



Figur 1.1
Produktionsprocessen set i sammenhængen renere teknologi og genanvendelse.

Dataanalysen må indrettes mod maximering af relevant information fra det analyserede datagrundlag. Ikke nødvendigvis en belysning af flest mulige spørgsmål, men målrettet klarlægning af den information, som det er undersøgelsens formål at fremskaffe.

Fortolkning

Information skal fortolkes, og fortolkninger kan lede til konklusioner og evt. beslutninger. Det er vigtigt at skelne mellem den analytiske objektivitet og fortolkningens formålsbestemthed. En sober analyse vil hyppigt indebære mulighed for flere fortolkninger, hvoraf nogle eller alle har plads i den konklusion, hvorfra beslutninger må træffes. Dette skyldes, at der normalt vil være elementer af tvivl til stede, hidrørende fra måleusikkerhed, modelapproksimationer, begrænsninger i datamaterialets omfang, begrænsede generaliseringsmuligheder m.v. En dataanalyse kan således ikke blot f.eks. indeholde effekt- eller belastningsestimater, men må også på passende vis angive den usikkerhed, hvormed estimaterne er behæftede.

Usikkerhed

Formidling

Erhvervelsen af viden i en undersøgelse vil sjældent have til primært formål at oplyse og uddanne undersøgelsens ophavsmænd. Undersøgelsens dokumentation må være af en sådan art, at dens informationsindhold formidles. Rapporter over undersøgelser må således være eksplicitte i informationspræsentationen, d.v.s. dragne konklusioner klart må refereres til rapportens informationsindhold, og der må redegøres for usikkerhedsmomenter i kvantitativ form, når det er muligt.

Det er nærværende projekts formål, at udrede disse spørgsmål og give normative anvisninger på hvorledes de beskrevne elementer kan og bør indgå i rapporter over miljøundersøgelser, specielt med henblik på forurenende industrielle anlæg.

Produktionsprocessen

En fabrikationsproces kan opfattes som et system, der af råvarer fremstiller produkter med positiv værdi, nemlig salgsprodukterne. Desuden fremstilles restprodukter med realiserbar værdi, samt produkter med negativ værdi, nemlig affaldsprodukter. De sidste er undertiden blevet opfattet som restprodukter uden realiserbar værdi, d.v.s. med værdien nul. I det omfang de er forurenende d.v.s. forvolder skade og/eller kræver modforanstaltninger bliver deres reelle værdi negativ i en total økonomisk vurdering.

Set i sammenhængen renere teknologi og genanvendelse, kan produktstrømme afbildes som vist i figur 1.1, der stammer fra Miljøstyrelsen (1986). Ved justeringer (optimering) og ændringer i produktionsprocessen og/eller indførelse af helt ny produktionsteknologi, herunder ny type råvare eller ligefrem et substituerende produkt, kan de negative produkters mængde og skadelighed påvirkes i gunstig retning. Ved indførelse af renere teknologi kræver den relative karakter of ordet »renere« en nøjere klassifikation: renere end alle andre teknologier, renere end eksisterende teknologi, eller blot renere end nogen som helst teknologi. Tales der om renere end betragtede alternativer, må det klargøres hvor dækkende disse er for det nuværende teknisk-økonomiske stade.

Relativ renhed

Ønskelige forbedringer er imidlertid typisk ledsaget af betydelige omkostninger: nyinvesteringer, efteruddannelse af medarbejdere, produktionstab under omstilling m.v., samt en vis usikkerhed angå-

ende ændringernes effekt, såvel hvad angår de forureningsmæssige, som de økonomiske aspekter.

Af hensyn til både miljø og produktivitet må en fornuftbetonet omlægning bygge på information, som realistisk belyser de påtænkte ændringers effekt og også specificerer usikkerheden ved effektestimationen.

Informationsstrøm

En sådan information kan også opfattes som en produktstrøm fra fabrikationsprocessen. Procesovervågning og -optimering bør normalt være en permanent delaktivitet af produktionsprocessen (se f.eks. Box & Draper, 1969). Effekten af mulige større eller mindre ændringer i processen kan bedst vurderes ud fra eksperimenter hermed. Undertiden kan laboratorieforsøg fungere som substitut for den egentlige proces af økonomiske, sikkerhedsmæssige og andre grunde.

Statistikkens rolle

Information er ikke gratis, og prisen afhænger ikke kun af datamængden, men også af planlægningen og analysen af det datasæt, som informationen ekstraheres fra. For at uddrage mest information af data må analysen normalt foretages ved statistiske metoder. Endnu vigtigere er det dog, at planlægningen også foregår ud fra statistiske principper. Dette skyldes ikke blot, at et velplanlagt forsøg oftest er enkelt at analysere og leder til klare konklusioner, medens en uheldig eller manglende plan kan give datasæt, der kræver stor ekspertise at analysere under anvendelse af komplicerede og tidskrævende metoder. Den vigtigste grund til at planlægningen er så afgørende, ligger i det faktum, at selv den mest avancerede statistiske teknik aldrig kan uddrage mere information, end der er til stede i det aktuelle datasæt. Brug af en avanceret analyse, er ingen garanti for at man når frem til sikre, entydige konklusioner. Derimod er det i høj grad i planlægningsfasen at informationsindholdet af data fastlægges. Det gælder nemlig slet ikke at informationsindholdet er proportionalt med antallet af målinger. F.eks. kan to lige store, men forskelligt planlagte forsøgsserier give effektestimatorerne præcisioner, der ligger størrelsesordener fra hinanden.

Planlægning

Statistisk bistand

Det skal ikke benægtes at mange problemstillinger bedst belyses ved forsøg, hvis planlægning og analyse kræver aktiv bistand af en statistiker. Imidlertid må det bedste ikke stå i vejen for det gode, og mange problemstillinger kan belyses fuldt tilfredsstillende med enkle forsøgsplaner og dertil hørende simple analyser. Nærværende udredning giver anvisninger herpå.

Det er et væsentligt bidrag fra en given undersøgelses medarbejder(e) at forstå den foreliggende problemstilling så godt, at man bliver i stand til hensigtsmæssigt at udnytte de ideer og simple analyseteknikker, som vil blive gennemgået. Kapitel 2 er tænkt som en hjælp hertil. I dette kapitel fremlægges visse generelle synspunkter og principper. Det velplanlagte forsøg vil ofte producere data, der er egnede til grafisk repræsentation; d.v.s. at en ikke-misledende kvalitativ analyse kan – og bør – foretages ved hjælp af egnede plot. Kapitel 3 handler herom.

Udredningens indhold

Nøglebegreber

Kapitel 4 indeholder en diskussion og fortolkning af nøglebegreber, som hyppigt optræder i kvantitativ statistisk analyse og argu-

mentation. Udgangspunktet er en forklaring på den rolle som statistisk inferens spiller i dataanalysen. Herefter følger en nøjere gennemgang af inferensens hovedelementer, statistisk estimation og test, og deres forbindelse til antalsplanlægningen af en undersøgelses målinger.

Beregninger

I kapitel 5 gennemgås beregningsgangen for seks vigtige typeeksempler på kvantitativ statistisk analyse, herunder det centrale spørgsmål om, hvorledes antallet af målinger fastlægges i en konkret undersøgelse. Kapitlet afsluttes med en summarisk omtale af visse generaliseringer, med henvisning til egnet litteratur. Kapitel 6 indeholder rapportens konklusion, som er et resumé og en afgrænsning af rapportens indhold.

Eksempler

Hvor der i rapporten er benyttet taleksempler, er disse kunstigt konstruerede og udelukkende benyttet til at demonstrere mekanikken i afbildninger og udregninger. Dette tillader en kortere form, end hvor egentlige eksempler gennemregnes, og hvor datas oprindelse må inddrages; – da det i virkeligheden er meningsløst at planlægge og analysere data uden forståelse af den reelle verden bag data. Desuden gives der i de enkelte afsnit henvisninger til speciallitteraturen.

Disposition

Lidt uden for rapportens hovedlinie ligger Appendix A som indeholder et diskuteret forslag til disposition af en rapport over en teknisk undersøgelse. Endelig indeholder Appendix B en alfabetisk ordnet liste af forklaringer på begreber, ord og fagudtryk, der har speciel betydning i dataplanlægning og -analyse.

Ordforklaring

Statistisk førstehjælp

Statistisk forsøgsplanlægning og analyse er et stort emne, hvis beherskelse kræver såvel teoretisk indlæring som praktisk erfaring, hvilket netop udgør den professionelle statistikers uddannelse. Nærværende udredning kan opfattes som en slags statistisk førstehjælp.

En lægelig førstehjælpsbog erstatter ikke behovet for at søge lægelig hjælp ved sygdom eller ulykke. På den anden side går man selvfølgelig ikke til læge, blot man har fået en hudafskrabning eller en slem forkølelse. Det er de statistiske paralleller til den slags hyppigt forekommende, enkle men særdeles mærkbare problemer som udredningen handler om. Sund levevis har her sin parallel i god forsøgsplanlægning. Det vel gennemførte eksperimentelle arbejde kan i mange tilfælde forebygge behovet for senere at skulle have »repareret« analysen, hvad der ikke altid kan lade sig gøre.

Nærværende udredning overflødiggør med andre ord ikke behovet for statistisk konsultation hverken i planlægnings- eller analysefasen, men den skulle klargøre, hvornår egentlig hjælp er påkrævet, og hvornår man kan klare sig uden. Det vigtigste budskab er imidlertid, at en undersøgelses sundhed under alle omstændigheder bevidst skal plejes.

Sundhed

2. Statistisk analyse, model og data

Statistisk nødvendighed

En vis forståelse for statistisk tankegang og metode er nødvendig for alle ingeniører og andre, der seriøst beskæftiger sig med procesforbedringer i miljømæssige såvel som andre henseender. Grunden hertil fremgår af følgende korte fem-trins argument (jf. Hunter, 1981):

1) Det er urimeligt at forestille sig, at en aktuel produktion på alle måder foregår så ideelt, at forbedringer i form af procesoptimering, produktudvikling og/eller teknologifornyelse på forhånd kan udelukkes.
2) Forbedringer kræver nødvendigvis ændringer, ellers bevares selvsagt status quo.
3) Rationelle ændringer kan ikke foregå alene på grundlag af fornemmelser, strøtanker, eller fordomme, men må bygge på fakta, d.v.s. data.
4) For at undgå spild, må fremskaffelse og fortolkning af data foretages så effektivt som muligt.
5) Statistisk tankegang og metode handler netop om det i punkt 4) krævede.

2.1. Den empiriske læreproces

Læreprocessen

Indsigt og viden, der er baseret på fakta, stammer – direkte eller indirekte - fra observerede data. Når viden ud over den eksisterende ønskes erhvervet, må den empiriske læreproces fortsættes, d.v.s. at nye data må indsamles og deres informationsindhold ekstraheres. Det gælder om, at læreprocessen forløber så effektivt som muligt, så erhvervelsen af ny viden foregår hurtigt og økonomisk. Hertil kan statistisk tankegang yde et stort bidrag. Når et datasæt skal fortolkes, er det en stor hjælp, hvis de væsentlige aspekter ved data kan udnyttes gennem en model.

Model

Har man for eksempel foretaget en lang række måling Y_i, $i = 1,2,\cdots,n$ af en bestemt miljøvariabel og fundet en vis usystematisk variation omkring et niveau, kan man som model forestille sig, at Y_i er normalfordelt omkring en middelværdi, μ, med variansen, σ^2. Dette skrives

$$Y_i \varepsilon N(\mu,\sigma^2), \tag{2.1}$$

eller alternativt

$$Y_i = \mu + e_i, \tag{2.2}$$

Fejl

hvor den statistiske fejl eller støj, e_i, er normalfordelt

$$e_i \varepsilon N(O, \sigma^2). \tag{2.3}$$

Det kan også tænkes, at μ, der er Y_i's forventningsværdi, på specificeret måde afhænger af omstændigheden, X_i, som man kender. D.v.s.

$$Y_i = \mu_i + e_i, \tag{2.4}$$

hvor

$$\mu_i = f(X_i). \tag{2.5}$$

$f(X_i)$ siges at være den deterministiske model, eller blot modellen. Modelsammenhængen mellem X og Y kan være specificeret i form, men med en ukendt modelkonstant Θ, f.eks. en proportionalitetskonstant:

$$\mu_i = f(X_i) = \Theta X_i \tag{2.6}$$

Parametre

Sådanne ukendte konstanter kaldes i statistisk sprogbrug modelparametre, eller blot parametre. Mere generelt kan μ_i afhænge af flere omstændigheder $X_{i,j}$ $j = 1,\cdots,m$, og/eller flere parametre Θ_k, $k = 1,2,\cdots,p$. D.v.s.

$$Y_i = f(X_{i,1},X_{i,2},\cdots,X_{i,m};\Theta_1,\Theta_2,\cdots,\Theta_k) + e_i = f(\underset{\sim}{X}_i,\underset{\sim}{\Theta}) + e_i, \tag{2.7}$$

hvor bølgelinien angiver en vektor. Y_i består således af en systematisk eller deterministisk del $f(\underset{\sim}{X}_i,\underset{\sim}{\Theta})$, og en tilfældig eller stokastisk del, e_i.

Estimation

Ved konstruktion af modellen f, gælder det om, at mindst muligt af Y_i kommer til at optræde som tilfældig støj, medens mest muligt af Y_i indgår i den systematiske model, som kan forklares, forstås og benyttes til forudsigelser. Muligheden for at justere de ukendte parameterværdier $\underset{\sim}{\Theta}$, udnyttes derfor til at få modellen $f(\underset{\sim}{X}_i,\underset{\sim}{\Theta})$ til at stemme så godt som muligt overens med observationerne, Y_i, $i = 1,2,\cdots,n$. De værdier for $\underset{\sim}{\Theta}$, der skaber denne bedste overensstemmelse, kaldes estimaterne eller de estimerede parametre, og betegnes $\underset{\sim}{\hat{\Theta}}$. I den tilpassende eller fittede model indsættes $\underset{\sim}{\hat{\Theta}}$ på $\underset{\sim}{\Theta}$'s plads, hvorved modellen kan prediktere Y-værdier for givne omstændigheder $\underset{\sim}{X}$. For omstændighederne $\underset{\sim}{X}_i$, $i = 1,\cdots,n$ svarende til de observerede Y_i-værdier, kaldes disse predikterede, fittede eller tilpassede værdier:

$$\hat{Y}_i = f(\underset{\sim}{X}_i,\underset{\sim}{\hat{\Theta}}) \quad , \quad i = 1,2,\cdots,n \ . \tag{2.8}$$

Residualer

Forskellen mellem disse model-værdier $\hat{Y}_i$ og de faktiske observerede værdier Y_i, kaldes residualerne,

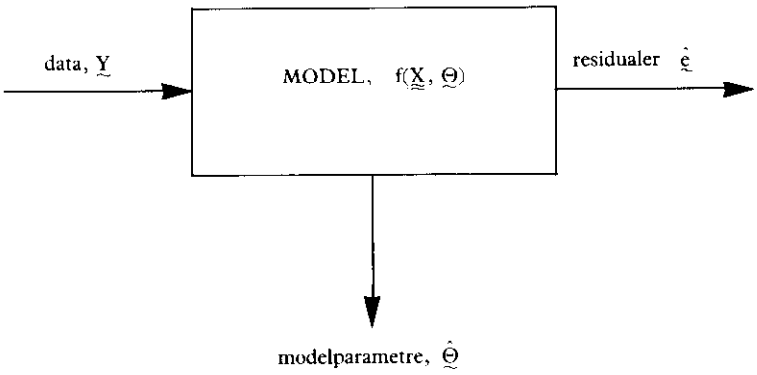
$$r_i = Y_i - \hat{Y}_i \ , \ i = 1,2,\cdots,n \ . \tag{2.9}$$

Residualerne er den del af Y_i-erne, som modellen er ude af stand til at

forklare, og det ses, at r_i-erne estimerer de statistiske støjbidrag i (2.7):

$$\hat{e}_i = Y_i - f(\underset{\sim}{X}_i, \hat{\underset{\sim}{\Theta}}) = r_i. \tag{2.10}$$

Hvis den betragtrede model f giver en god repræsentation af den datagenererende proces, og man ved estimation af parametrene $\underset{\sim}{\Theta}$ sørger for, at modellen kommer i bedst mulig overensstemmelse med de observerede data $\underset{\sim}{Y}$, så skulle residualerne $\underset{\sim}{r}$ som støjestimater ikke indeholde information.



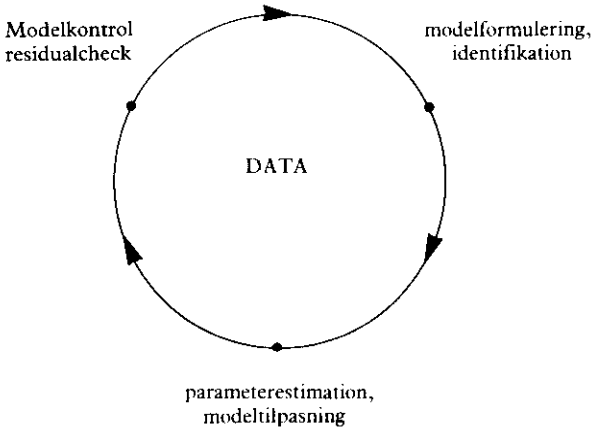
Figur 2.1.
Den statistiske model som filter af information fra støj.

Filter

I denne formulering kan ekstraktion af information fra et datasæt $\underset{\sim}{Y}$ illustreres af figur 2.1. Den betragtede model f virker som et filter, der separerer information fra støj. Parameterestimationen repræsenterer den erhvervede information, som fortolkes i lyset af modellen f.

Iterativ analyse

Viser en analyse af residualerne, at der stadig er noget at lære fra disse, er dette et tegn på, at modelformen f kan forbedres. Analysen af residualerne optræder således som en modelkontrol. Indses det, at en oprindeligt foreslået model er forkert eller utilstrækkelig, kan heri ligge ny indsigt. Nu må en ny model formuleres, dens parametre estimeres, og dens tilpasning checkes. Det kan så udløse en ny iteration, som illustreret i figur 2.2, indtil konvergens er opnået omkring en rimeligt beskrivende model.

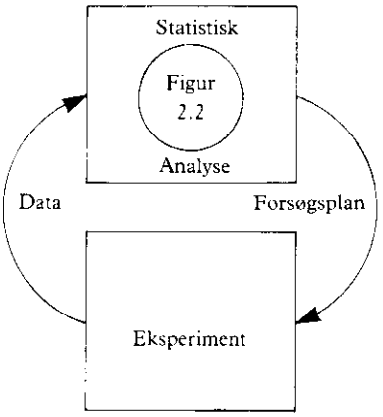


Figur 2.2.
Den iterative karakter af statistisk dataanalyse

Iterativ modelbygning

At en model er databeskrivende betyder ikke at den nødvendigvis er korrekt i betydningen sand, blot at den beskriver det aktuelle datasæt på en måde, der ikke afslører mangler. At den er beskrivende behøver heller ikke at betyde, at den er tilstrækkeligt god for en bestemt anvendelse, ofte benævnt adækvat. Det til rådighed stående datasæt kan nemlig være af en sådan beskaffenhed, at en tilstrækkeligt omfattende og præcis model ikke lader sig opstille og estimere. I så fald må yderligere data insamles. Ud fra den herskende viden, d.v.s. den aktuelt fittede model, kan nye forsøg planlægges, der bedst belyser og forbedrer modellens svage punkter. Disse nye data indgår herefter i en fornyet statistisk analyse, der kan afsløre behov for yderligere data o.s.v. Figur 2.3 illustrerer denne ydre iteration, som fortsætter indtil situationen er tilstrækkeligt oplyst, inden for gældende tids- og budgetsmæssige begrænsninger.

Figur 2.3
Iterativt undersøgelsesforløb vekslende mellem analyse og forsøg.



2.2 Datakvalitet og -kvantitet

Datagrundlag

Undersøgelser, påtænkte eller udførte, kan bygge på et datagrundlag af forskelligt omfang og kilde. En klassifikation, som den der er vist i figur 2.4, kan medvirke til at klargøre en undersøgelses rette indplacering i en mere global videnopbygning.

Figur 2.4.
Typer af undersøgelser

datamængde	egne, nye data	andres kendte data
ingen	(a) idé-oplæg	(b) idé-udredning
sparsom	(c) pilot-projekt	(d) motiverende udredning
tilstrækkelig	(e) egentligt projekt	(f) sammenfattende udredning

Der er ingen af de seks kategorier af undersøgelser, der a priori er bedre end andre. Inden for hver kategori er der mulighed for undersøgelser af større eller mindre valør, fagkundskab og datakvalitet. Hvis imidlertid opsætning og ordvalg i rapporteringen ikke reflekterer datagrundlaget, kan det afføde misforståelser, frustrationer og/eller fejlbeslutning hos læseren, - det være sig en rekvirent, klient eller anden type bruger.

En teknisk rapport bør i princippet ikke »sælge« læseren et budskab, en beslutning, eller en snæver konklusion, men præsentere undersøgelsens resultater i en neutral og forståelig form. Læseren kan så selv uddrage sit budskab eller træffe sin beslutning, måske under indtryk af argumentation fra anden kilde eller af helt andre hensyn. De seks kategorier skal kort karakteriseres:

Idéoplæg

(a) En undersøgelse kan godt gennemføres uden at bygge på data overhovedet. Den kan være en påpegelse af en potentiel forureningskilde, en idé til løsning af et miljøproblem, en bekymring for en teknologisk udvikling belyst ved parallelle eksempler på forurening eller ulykker inden for beslægtet område, eller andet. En sådan type undersøgelse kan karakteriseres som et idéoplæg, et debatpapir eller lignende. Da det ikke uden data er muligt at drage statistisk valide konklusioner, må sprogbrug, der antyder en statistisk støtte, undgås, og den spekulative karakter af undersøgelsen bør ikke skjules.

Idéudredning

(b) En anden helt acceptabel type undersøgelse, der heller ikke bygger på data, er sammenfatning og systematisering af andres tidligere type (a) undersøgelser; idéudredning kan man kalde den. Tydelig angivelse af referencer er nødvendige, ikke blot af hæderlighedshensyn over for forfatterne, men så læseren får klarhed over forbindelsen mellem rapportens og referencernes konklusioner. Igen må konklusionernes spekulative karakter fremgå af ordvalget, uagtet referencernes antal og notabilitet.

Pilotprojekt

(c) En undersøgelse byggende på et mindre, nyt materiale kan karakteriseres som et pilotprojekt, og data bør gengives fuldt ud i rapporten. Jf. figur 2.3, og de ledsagende bemærkninger, kan det være mest fordelagtigt at gennemføre et mindre projekt før et evt. senere større, som herved kan planlægges bedre. En tilsyneladende nedslående konklusion om manglende tiltro til nytten af yderligere undersøgelser, kan være overordentlig værdifuld. Den tilsyneladende spildte pilotundersøgelse, kan have sparet den langt større undersøgelses fiasko. Der er heller intet galt eller overraskende i, at undersøgelser af denne type ikke kan konkludere entydigt; og en evt. inkonklusivitet må ikke skjules verbalt.

Motiverende udredning

(d) En undersøgelse uden egne nye data, men byggende på andres sparsomme, spredte data, kan enten have karakter af en sammenfletning og bearbejdelse af tidligere konklusioner; eller der kan være tale om en ny analyse af andres data, evt. fra flere kilder. Igen må rapporten redegøre for, om der er tale om det ene eller det andet; eller, hvis begge dele udføres, klart separere de to typer

indsats. Undersøgelsen kan karakteriseres som motiverende for yderligere undersøgelser.

Egentligt projekt

(e) En undersøgelse med en så tilstrækkelig datamængde, at den kan besvare de stillede spørgsmål tilfredsstillende, må gøre dette på en statistisk sund måde. Data bør generelt gengives i rapporten eller i appendix. Man kan her tale om en egentlig undersøgelse eller udredning eller om et egentligt projekt.

Sammenfattende udredning

(f) Når en problemstilling er velbelyst, specielt fra flere sider, kan det være af interesse at foretage en udredning, der samler og/eller bearbejder resultaterne; eller, når det er berettiget, foretager statistisk reanalyse af datasæt fra evt. flere kilder. I rapporteringen må der klart sondres mellem hvilke konklusioner, der er gengivelser fra referencerne, og hvilke der er nye, udsprunget af bearbejdelsen.

Andre undersøgelser

Der er tydeligvis tale om en fremskridende rækkefølge i disse undersøgelsestyper: (a)-er kan danne grundlag for (b), (b) for (c), (c) for (d), (d) for (e), og (e) for (f). Men vejen (a) over (c) til (e) kan også være naturlig, og andre muligheder eksisterer. De nævnte seks kategorier er meget skematisk opdelte, og det kan være berettiget at gennemføre undersøgelser, der i henseende til datamængde (figurens lodrette retning) ligger mellem kategorierne eller ud over »tilstrækkelig«. Derimod må det generelt frarådes at rapportere undersøgelser, der blander egne og andres data (figurens vandrette retning). Egne, nye resultater bør rapporteres separat, en evt. påfølgende rapport af typen (d) eller (f) kan samle resultaterne.

Hvor alle seks kategorier i figur 2.4 potentielt kan rumme undersøgelser af høj kvalitet, er der tale om kvalitetsforskelle mellem de fire kategorier i figur 2.5. De fire kategorier fremkommer som kombinationerne af aktiv vs. passiv, og systematisk vs. tilfældig dataindsamling.

Aktiv dataindsamling

Aktiv dataindsamling indebærer at forsøgsomstændighederne, X_i, $i=1,\cdots,n$, er manipuleret på systematisk vis. Et tilfældigt sæt forsøgsomstændigheder er altid kritisabelt; hvor tilfældigt betyder på slump, efter bekvemmelig eller anden subjektiv indskydelse. En sådan plan kan føre til misledende statistiske konklusioner, og selv hvor det ikke sker, vil et element af tvivl altid stå tilbage. En systematisk manipulering af forsøgsomstændighederne betyder her et statistisk planlagt forsøg, herunder også brug af randomisering.

Randomisering

Figur 2.5.
Datakvalitet

data	tilfældig	systematisk
aktiv	kritisabelt forsøg	statistisk forsøgsplan
passiv	kritisabel dataindsamling	moniterings-plan

Randomisering er en objektiv tilfældighed, som fastlægges på forhånd, og ikke ændres under forsøgets gennemførelse, uagtet f.eks. hvor ubekvem den måtte vise sig. Randomiseringen drejer sig normalt om rækkefølgen af de forsøg, som planen indebærer, men kan f.eks. også omhandle allokering af laboranter til de forskellige prøver, og andre sagen uvedkommende faktorer. Uden randomisering kan disse ellers påvirke undersøgelsens resultater systematisk. Randomisering af rækkefølge/allokering kan praktisk udføres f.eks. ved mønt- eller terningkast, ved brug af tabeller over tilfældige tal (se f.eks. Beyer, 1968) eller ved computerrandomisering.

Passiv dataindsamling

Passiv dataindsamling betyder, at den studerede proces blot observeres uden påvirkning, og målinger indsamles som en monitering af et forløb. Dette kan enten ske på tilfældig vis, hvilket som før altid er kritisabelt; eller på en systematisk vis, f.eks. på forud fastsatte tidspunkter. Disse tidspunkter kan være fastsat med regelmæssige intervaller eller på randomiserede tidspunkter. Denne form for systematisk tilfældighed er – som før – på forhånd fastlagt objektivt ved en passende sandsynlighedsfordeling og ændres ikke subjektivt undervejs p.g.a. bekvemmelighed eller andet.

Systematiske planer

Kun systematiske planer er acceptable, de aktive er egentlige statistiske forsøgsplaner, de passive bør snarere opfattes som moniteringsplaner. Hvor det overhovedet lader sig gøre, vil aktive planer normalt være stærkt at foretrække for passive, når sigtet med undersøgelserne er innovativt og ikke blot deskriptivt.

Kausalitet

Konklusioner fra en undersøgelse hvor dataindsamlingen er foretaget passivt, om end systematisk, vil kun kunne forventes gyldige for en fortsat passiv fremtid. De fundne niveauer, variationer, sammenhænge m.v. er kompakte beskrivelser af datamaterialet. Men årsagssammenhængen to variable imellem, kan der f.eks. intet sikkert siges om: Måske den ene påvirker den anden, eller den anden påvirker den ene, eller måske de i virkeligheden blot samvarierer på grund af påvirkning fra en fælles tredje variabel. Faglig indsigt kan i nogle situationer gøre en bestemt årsagssammenhæng plausibel. Kausalitet kan imidlertid kun demonstreres med statistisk sikkerhed af en dataanalyse baseret på planlagte forsøg med aktiv indgriben i den betragtede proces.

Denne konstatering understreger, at en undersøgelses fremadrettede værdi afhænger af, hvordan data er indsamlet, et forhold der ikke altid fremgår af datamaterialet selv, men må oplyses eksplicit og afspejles i formulering af bl.a. konklusioner.

2.3 Repræsentativitet af undersøgelser

Bemærkningerne om aktiv vs. passiv dataopsamling er en del af en større problematik angående repræsentativiteten af en undersøgelse.

Som hovedregel gælder det, at en undersøgelses resultater kun er gyldige under tilsvarende omstændigheder som disse, hvorunder den er udført. Til en generalisering herudover knytter sig en usikkerhed,

Generalisering

som afhængig af dataplanlægningen kan være af ukendt systematisk

karakter, eller af usystematisk stokastisk karakter. En systematisk usikkerhed kan undertiden vurderes subjektivt, med større eller mindre held bl.a. afhængigt af den vurderendes sagkyndighed. En ikke-systematisk eller stokastisk usikkerhed kan derimod estimeres statistisk objektivt. En undersøgelses resultater kan være behæftet med usikkerhed af begge typer, således at man ved generalisering i forskellige retninger oplever forskellige typer generaliseringsusikkerhed. Denne skelnen kan konkretiseres ved et realistisk, tænkt eksempel.

Anlæg vs. laboratorium

Vi forestiller os den situation, at en kendt produktionsproces giver anledning til luftforurening med et bestemt spildstof, som dannes ved en sidereaktion under syntesen af salgsproduktet. Man ønsker at undersøge om brug af en alternativ katalysator vil formindske dannelsen af forureningen, og har nu valget mellem at foretage forsøget under kontrollerede laboratorieomstændigheder eller på det arbejdende anlæg. I produktionsprocessen er variationen i udbytte og forurening relativt stor p.g.a. variationer i råmaterialer, aflejringer i reaktor, temperaturvariation, m.v. Udføres undersøgelsen i laboratoriet, kan disse variationskilder næsten helt elimineres, og man kan derfor fra ret få målinger med stor præcision afgøre effekten af katalysatorskift. Men hvorvidt et sådant resultat kan overføres til industriel praksis, belyses derimod ikke af denne undersøgelse. Resultatet gælder for katalysatorskift under de givne laboratorieomstændigheder. Det kan tænkes, at der i råstofferne til fuldskalaprocessen findes en urenhed som forgifter den i laboratoriet foretrukne katalysator, så dens gunstige virkning formindskes eller udebliver. Der kan med andre ord eksistere en forskel mellem anvendelse i laboratorium og industri. Denne forskels eksistens eller størrelse kan ikke belyses statistisk fra data, men må skønnes mere eller mindre subjektivt af sagkyndige. Der er grelle eksempler på, at denne type skøn har været helt forkerte, men ofte går det også godt. Under alle omstændigheder er og bliver undersøgelsens repræsentativitet statistisk ubelyst.

Subjektive skøn

Hvis man modsat foretager målinger med de alternative katalysatorer i fuldskalaanlægget, vil ikke blot omkostningerne for måling normalt være langt højere, men variationen, støjen, vil også være langt større. D.v.s. at undersøgelsens konklusioner kommer til at fremstå med en meget ringere præcision. Dette forhold kan overfladisk set opmuntre til laboratorieforsøg fremfor fuldskalaforsøg; men hvor omkostningsargumentet kan være tungtvejene er præcisionsargumentet ret beset forkert: Den større præcision er opnået på bekostning af en systematisk fejl af ukendt størrelse. Denne fejl er i fuldskalamålinger konverteret til en stokastisk fejl, som kan behandles statistisk.

Konvertering af fejl

Er der tale om ikke ét, men mange procesanlæg, er det tilsvarende risikabelt at bygge en anbefaling om katalysatorskift på målinger fra ét anlæg. Den estimerede effekt gælder ikke med sikkerhed for andre anlæg, hvor lokale forhold kan spille ind. Ved generalisering fra et anlæg til et andet, eller til flere andre, kan der igen forekomme systematiske fejl af ukendt størrelse. Man kan måske subjektivt

Flere anlæg

skønne deres størrelse, eller deres manglende realitet – bekvemt, men risikabelt. Den statistisk sunde løsning er imidlertid at foretage målinger på flere anlæg.

Skal et mindre antal anlæg være repræsentative for alle anlæg, må udvælgelsen ske randomiseret, d.v.s. ved tilfældig, blind udtrækning. Kun herved vil den estimerede variation mellem anlæg være repræsentativ for alle anlæg. Foretages et egentligt valg af anlæg, f.eks. af bekvemmelighedshensyn eller for at få forskellige typer repræsenteret, kommer undersøgelsen igen til kun at repræsentere netop disse anlæg. Anlægsforskellene kan være oplysende, men må

»Fixed effect«

statistisk set regnes som såkaldte »fixed effects«, d.v.s. faste deterministiske forskelle. Og variationen til øvrige, ikke målte anlæg forbliver ukendt. Ved randomisering, på den anden side, har alle forskelle samme chance for at komme til at indgå i estimationen af variationen mellem anlæg, hvorfor disse forskelle er konverteret til statistisk analyserbar stokastisk variation. Statistisk set regnes variation mellem

»Random effect«

anlæg nu som en såkaldt »random effect«.

Man kan undertiden med fordel lægge begrænsninger i randomiseringen, f.eks. hvis der eksisterer to klart forskellige anlægstyper. Her kan man beslutte, at hver gruppe skal repræsenteres med et i forvejen fastsat antal anlæg, som så specificeres individuelt ved randomisering. Sådanne planer kræver imidlertid større statistisk indsigt at planlægge og analysere.

Andre forskelle

Betragtninger af ovennævnte type har naturligvis gyldighed ud over det skitserede eksempel. Der kan være tale om andre typer produktionsændringer eller teknologiskift; og foruden variation mellem anlæg, kan man tænke på variation over tid (for samme anlæg), variation mellem prøvetagningsudstyr, hvis flere muligheder kunne tænkes anvendt, variation mellem analyserende laboratorium m.v.

»Alt andet lige«

Ved fortolkning af statistiske modeller benyttes ofte og med rette vendingen, »alt andet lige«, når effekten af en variabelændring beskrives. Det betyder imidlertid ikke, at det er god eksperimentel fremgangsmåde, kun at ændre én påvirkende variabel ad gangen og se dens effekt for sig, før man overvejer andre. Er man f.eks. både interesseret i anlægsforskelle, prøveudtagningsforskelle og tidsvariation, er det mest effektivt at planlægge et samlet forsøg, hvor alle disse faktorer indgår, fremfor at belyse dem separat. Igen bliver der let tale om forsøg, hvor planlægning og analyse kræver større statistisk kyndighed.

Repræsentativ præcision

Det skal understreges, at inddragelsen af flere meget forskelligartede anlæg i et måleprogram ikke nødvendigvis øger det totale antal krævede målinger for at estimere en (katalysator-)effekt med given nøjagtighed. Effekten estimeres for hvert anlæg som forskellen mellem forureningen ved de to katalysatorer, og nøjagtigheden af disse anlægsdifferenser er ikke påvirket af eventuelle forskelle i generelt niveau mellem anlæggene. Et gennemsnit af differenser bliver således lige så nøjagtigt bestemt, som hvis det samme antal målinger havde været udført på ét anlæg. Men nu ved man, at differensestimatet er repræsentativt. Kun hvis differenserne virkelig er forskellige fra anlæg til anlæg, kan præcisionen påvirkes, men i så tilfælde ville

målinger på ét anlæg netop ikke være repræsentative, og variationen i effekt bliver nu af særlig interesse.

På samme måde vil effektestimater, der skal repræsentere et længere tidsrum ikke nødvendigvis kræve flere målinger for at opnå en given præcision, selvom der måtte indtræde store variationer hen over tiden. Man skal igen sørge for, at målinger på de alternative katalysatorer foretages i par eller grupper, såkaldte blokke, under mere homogene omstændigheder, end der kan opretholdes blokkene imellem. D.v.s. at man over den længere tidsperiode foretager mindre måleserier, inden for hvilke målebetingelserne er nogenlunde ens, medens forskelle måleserierne imellem må betragtes som fordelagtige i relation til undersøgelsens repræsentativitet.

Blokke

Inden for hver måleserie eller blok foretages målinger med begge katalysatorer i randomiseret orden, og en effektdifferens udregnes. Disse differenser er upåvirkede af niveauforskelle mellem blokkene og giver ved gennemsnitsdannelse anledning til samme statistiske præcision, som hvis ingen blokforskelle fandtes. Dette kunne faktisk ske, hvis der ingen tidsvariation fandtes, eller hvis undersøgelsen var gennemført på mindre repræsentativ måde, hvad angik tidsaspektet.

Blokforskelle

Meningen med repræsentativitet i relation til aktiv vs. passiv dataindsamling kan konkretiseres ved et andet tænkt eksempel. Har man f.eks. fundet en statistisk model for luftforureningen fra en produktionsproces, og modellen udsiger at forureningen er særlig kraftig, når indetemperaturen i fabrikken er høj (ca. 25°C) og mindre når den er lav (ca. 20°C), så kunne man overveje at installere køling i fabrikken. Forestiller vi os, at produktionsprocessen er varmeudviklende, at fabrikshallen i øvrigt er uopvarmet, og at data er passivt opsamlede, kan den højere temperatur og den højere forurening begge skyldes en større produktion, så køling altså ingen effekt ville have, uagtet at modellen viste en klar sammenhæng. Det er den ubevidste udeladelse i modellen af den virkeligt påvirkende variabel, der giver anledning til den kraftige tilsyneladende sammenhæng. En sådan overset, influerende variabel kaldes undertiden en »lurking variabel«.

»Lurking variabel«

Modellen byggende på passivt opsamlede data, kan altså ikke uden videre forventes at bevare gyldighed til konsekvensprediktion af aktive indgreb. Dens repræsentativitet er indskrænket til et fortsat passivt forløb. Her kunne den f.eks. anvendes til udnyttelse af temperaturmålinger, som substitutmål for den skabte luftforurening.

Substitutmål

Ønsker man en model til aktiv brug, må den baseres på aktivt indsamlede data. Her vil man i planlægningen sørge for, at temperatur og produktivitet ikke er koblet, men bliver varieret uafhængigt af hinanden ved temporær opstilling af varme/køleanlæg. Kun hvis man på forhånd vidste, at temperaturen ingen effekt kunne have, ville dette besværlige eksperimentelle arrangement kunne udelades. Og i konsekvens heraf måtte temperaturvariablen så heller ikke optræde i modellen.

Afkobling

Med andre ord: hvis man vil bestemme den kausale effekt af en

variabel X på en miljøvariabel Y, så må man aktivt variere X efter en egnet plan og ikke slå sig til tåls med, at X selv uden aktiv manipulation alligevel varierer. Variationsområdet for X skal vælges så tilpas stort, at man ved brug af en resulterende model hvori X indgår, ikke unødigt tvinges til en ekstrapolation. Her bliver repræsentativiteten igen mere tvivlsom end ved interpolation. X bør altså som hovedregel varieres over det område, der har anvendelsesinteresse.

Ekstrapolation

Repræsentativitet

Problemstillingen omkring repræsentativitet giver rigeligt at bekymre sig om, og næppe nogen praktisk undersøgelse kan tage højde for alle potentielle faldgruber. Det vigtigste er,

- at undersøgelsen belyser et så bredt interesseområde af omstændigheder som muligt,
- at man ved passende blokning opnår god præcision, og
- at man ved bevidst brug af randomisering sikrer en statistisk troværdig vurdering af den usikkerhed, der knytter sig til undersøgelsens resultater.

2.4 Vigtige undersøgelsestyper

Mængden af alternative muligheder for forsøgsplaner med tilhørende analyse, kan forekomme uoverskueligt stor, specielt hvis man som ikke-statistiker konsulterer originallitteratur eller opslagsværker med sådanne planer.

Overblik

Fokuserer man imidlertid på et begrænset antal væsentlige undersøgelsestyper, kan der nok skabes et vist overblik og gives anvisninger, så den statistiske side af disse undersøgelser kan gennemføres på en helt acceptabel måde. Følgende undersøgelsestyper vil blive belyst:

1) *Middelværdi og varians*
   Denne type undersøgelse drejer sig om, hvorledes man estimerer et repræsentativt niveau, eller middelværdi for en procesvariabel, finder usikkerheden på estimatet og statistisk tester om middelværdien kan have en given værdi, den statistiske usikkerhed taget i betragtning. Herunder gives en statistisk beskrivelse af processens fluktuation omkring middelværdien. Dette behandles i afsnit 3.1 og 5.1.

2) *Forskel mellem to middelværdier*
   Denne type undersøgelse tager sigte mod at estimere en forskel mellem to alternative procespåvirkninger eller typer teknologi, eller procesændring versus standardmetode, etc. Usikkerheden i estimatet angives og fortolkes, herunder vises hvorledes statistisk test udføres for hypotetiske forskelsværdier, specielt nul, d.v.s. ingen forskel. Dette behandles i afsnit 3.2 og 5.2.

3) *Forskel mellem én middelværdi og flere alternativer*
I denne type undersøgelse estimeres effekterne af flere alternative påvirkninger over for en fælles reference, typisk den umodificerede standardproces. Estimaternes usikkerhed findes og deres signifikans testes. Dette behandles i afsnit 3.3 og 5.3.

4) *Forskel mellem flere sideordnede middelværdier*
I denne type undersøgelse er der ingen naturlig reference, og analysen drejer sig om at eftervise og estimere forskelle på en række alternative påvirkninger eller teknologier etc. Herunder testes statistisk om observerede forskelle er signifikante andre variationskilder taget i betragtning. Dette behandles i afsnit 3.3 og 5.4.

5) *Korrelation*
Ved denne type undersøgelse skal korrelationen mellem to variable beskrives, og det skal anføres hvorledes man statistisk tester, om den observerede korrelation blot kan skyldes tilfældig fluktuation. Dette behandles i afsnit 3.2 og 3.5.

6) *Regression*
Den type undersøgelse sigter mod at opstille en kausal model, der forklarer en given variabel som funktion af en anden variabel. Den bedste regressionslinie findes og fortolkes statistisk. Dette behandles i afsnit 3.2 og 5.6.

En række andre undersøgelser herunder generaliseringer af de 6 behandlede, omtales kursorisk i afsnit 5.7.

Enkelhed

Mange værdifulde undersøgelser dækkes umiddelbart af disse seks behandlede typer, eller kan omskrives, så de kommer til det. Fordelen ved at formulere en undersøgelses sigte på simpel vis er ikke blot at dens gennemførelse og analyse forenkles, med deraf følgende reduceret mulighed for at noget går galt, men også formidlingen af resultaterne lettes.

Hvor en undersøgelse ikke lader sig gennemføre efter et enkelt, statistisk overskueligt program, bør det seriøst overvejes at inddrage en statistisk konsulent i arbejdet. Dette gælder generelt for særligt kostbare undersøgelser, hvor forsøgsplaner og analyser af mere skræddersyet tilsnit, kan give anledning til forbedringer af væsentlig konsekvens for præcision, repræsentativitet, og/eller økonomi.

2.5 Gennemførelse af undersøgelser

Gennemførelse af en succesrig undersøgelse indeholder en række mere eller mindre veldefinerede trin, hvor dataplanlægning og -analyse spiller en rolle. Disse trin kan afgrænses på forskellig måde; nedenstående er brugbar i nærværende sammenhæng.

Formulering

a) Opgavens formål og omfang skal specificeres, herunder tidsplan,

budget og afrapporteringens form. Denne specifikation er nødvendig, for at der ikke senere skal opstå tvivl om, hvorvidt undersøgelsen har levet op til forventningerne, samt hvorvidt eventuelle afvigelser herfra er velbegrundede eller ubegrundede, bevidste eller ubevidste, acceptable eller uacceptable.

Formuleringen må gerne være vag på de punkter hvor undersøgelsens resultater skal give mulighed for ny indsigt: resultater bør ikke foruddiskonteres. Loves der på forhånd (for) meget, bliver undersøgelsen fra starten handicappet af den unødvendigt øgede risiko for at skuffe. Formelt kan rapporten blive tynget af uinteressante undskyldninger og bortforklaringer, eller i værste fald en dækken over de uindfriede forventninger.

Derimod bør formuleringen ikke være vag hvad angår angrebsvinkel, benyttede metoder, tanker vedrørende forsøgsplan og målemetoder etc. Det kan også anbefales, at man angiver den usikkerhed hvormed undersøgelsens resultater kan forventes belyst, når der tages hensyn til det budgetterede antal målinger, den forventede måleusikkerhed, og andre skønnede variationskilder (f.eks. variationer mellem anlæg eller over tid). En sådan vurdering kræver statistiske overvejelser: hvilken nøjagtighed kan forventes med et givet måleinstrument, og hvilket antal målinger skønnes nødvendigt for at opnå en given nøjagtighed. En præcis opgaveformulering beskytter både den udførende og den rekvirerende part mod senere skuffelser.

Usikkerhedsvurdering

Realisme i rekvirentens forventninger kan dermed sikres; – hvis udbyttet på forhånd synes for magert i forhold til omkostningerne, er det bedre slet ikke at sætte nogen undersøgelse i gang, eller måske iværksætte en mindre, forberedende undersøgelse. Tilsvarende er det bedre for den udførende part at afslå udførelsen af et undersøgelsesprojekt, hvortil der stilles forventninger, som er opskruede i forhold til, hvad udføreren mener kan leveres.

Planlægning

b) Når projektet er sat i værk må udføreren nøjere planlægge udførelsen af måleprogrammet, gerne som en sekventiel plan der tillader analyse og justering undervejs. Det skal afgøres, om der skønnes behov for særligt sagkyndig rådgivning i indledningsfasen, f.eks. en kemiker, måleteknikcr, lægc eller statistiker. Det er normalt billigere at betale sig fra potentielle fejl og forglemmelser, end senere at skulle rette en kuldsejlet undersøgelse op.

Statuspunkter

c) Der bør fastlægges statuspunkter undervejs, så det kan checkes, at undersøgelsen skrider planmæssigt frem, at budget og frister overholdes, og at resultaterne lever op til opgaveformuleringens forventninger. Det betyder specielt, at dataanalysen ikke bør vente til alle målinger er foretaget, men analyseres undervejs. Det er ofte hensigtsmæssigt at rekvirenten ved statustidspunkterne informeres om projektets hidtidige forløb. Trods omhyggelig projektformulering kan der optræde misforståelser som her let ryddes af vejen. Det kan også på grundlag af tidlige resultater vise sig, at projektet kan forbedres ved en drejning, som kræver

rekvirentens accept. Viser der sig store uforudsete vanskeligheder, så projektet ikke kan gennemføres planmæssigt, er det i reglen også klogt at finde en udvej i forhandling med rekvirenten, så tidligt i forløbet som muligt (øget – men måske alligevel tabsgivende – budget, længere frister, afbrydelse af projektet, eller andet).

Afrapportering

d) Når projektet nærmer sig afrapporteringstidspunktet, foretages den statistiske dataanalyse under hensyntagen til, hvorledes målingerne faktisk er foretaget – specielt om de er aktive eller passive – og med nøje angivelse af på hvilke områder randomisering har været foretaget. Konklusioner udformes med angivelse af de fundne usikkerheder på resultaterne, uagtet hvad man havde håbet eller forventet. Hvor det er muligt, bør rapportudkast sendes til gennemsyn hos rekvirent og eventuelle rådgivere, der opfordres til at checke for konsistens, fejl og forglemmelser. Påpegede afvigelser fra projektbeskrivelsen eller nye ideer bør ikke føre til andet end højst redaktionelle ændringer: rapporten skal dokumentere det foretagne arbejde, og ikke hvad man måske tidligere havde forestillet sig eller nu kunne ønske sig.

Komplet rapport

Afrapporteringen bør i størst muligt omfang være selvforklarende og komplet. Man bør undgå at henvise til baggrundsmateriale, f.eks. datalister, som er vanskelige at skaffe sig adgang til, og som i værste fald kan forsvinde, hvis de ikke publiceres. De for en statistisk reanalyse nødvendige tal, oplysninger og forklaringer må indgå i rapporten eller dens bilag/appendices. Dette kan være værdifuldt for fremtidigt videnopbyggende arbejde ligesom det – berettiget – promoverer troværdighed.

3. Grafisk dataanalyse

Kvalitativ analyse

Før en kvantitativ statistisk dataanalyse indledes, bør en kvalitativ vurdering foretages. Hermed menes en grafisk baseret analyse til afsløring af hovedtræk i datamaterialet. I mange tilfælde vil den grafiske analyse i virkeligheden give svar på de væsentlige spørgsmål, som den kvantitative analyse dernæst kan bekræfte på en mere autoritativ måde.

Væsentlighed

Den grafiske analyse må ikke betragtes nedladende. Udover i sig selv at være oplysende, kan den inspirere til, hvorledes den kvantitative analyse skal gribes an. Derved bliver den ofte en forudsætning for, at den kvantitative analyse ikke bliver flakkende eller afsporet, men målrettet kommer til at koncentrere sig om de væsentlige aspekter i data. En effekt, der er fundet ved kvantitativ analyse og som ikke kan erkendes grafisk, må betragtes med skepsis. Grafiske analyser har en vigtig rolle at spille i afsløring af en kvantitativ analyses forglemmelser, fejl og bommerter.

Residualanalyse

Både en indledende og en sideløbende grafisk dataanalyse er i realiteten nødvendige for en effektiv kvantitativ dataanalyse. Jf. afsnit 2.1 er analysen af residualer en vigtig del af den iterative modelbygning og residualanalyse kan ofte tilfredsstillende udformes som rent grafisk analyse. Det vil sige, at ikke blot data, men også residualer på forskellige trin i analysearbejdet skal plottes på passende måde. Videre kan den grafiske analyse visuelt afsløre, hvor oplysende datapunkter mangler, en vigtig information for videre dataplanlægning. Endelig kan den formidle de kvantitative hovedkonklusioner.

Klare figurer

I samme grad den grafiske analyse skal tages alvorligt, må den udføres seriøst. Afbildninger af data må ikke benyttes som udsmykning af en ellers tung tekst, alt for udspekulerede, humoristiske eller fortænkte afbildninger bør undgås. De fremstillede figurer må ikke i deres form kunne inspirere til over- eller fejlfortolkning, eller opmuntre til tænkepause fra rapportens væsentlige spørgsmål. Tværtimod skal de søge at fange læserens interesse for undersøgelsens problematik, og give et reelt billede af, hvor meget eller hvor lidt data kan sige herom. De skal udføres sobert, hvilket blandt andet betyder, at

- data skal afbildes på en klar og entydig måde, så muligheden for fejlfortolkning minimeres;
- så mange relevante oplysninger som muligt skal direkte fremgå af afbildningerne;
- afbildningerne skal muliggøre visuelle sammenlignigner mellem forskellige dele eller aspekter af data;
- afbildningerne skal kunne læses på flere detaljeringsniveauer, d.v.s. de skal både vise de store linier og tillade inspektion af finere strukturer;
- generelt er enkelhed i afbildningen at foretrække: ekstra linier, skraveringer, farver, etc., der ikke i sig selv tilføjer ny information, bør normalt undgås;

- hvor en variabels størrelse afbildes som en højde, bør perspektiv eller anden pynt undgås, som f.eks. kunne lede tanken hen på proportionalitet med overflade, rumfang eller vægt;
- i afbildninger med akser bør akserne krydse hinanden i deres nulpunkter, når det er muligt;
- akseinddelinger bør være ækvidistante, eller på anden måde gennemskueligt systematisk, f.eks. logaritmisk.

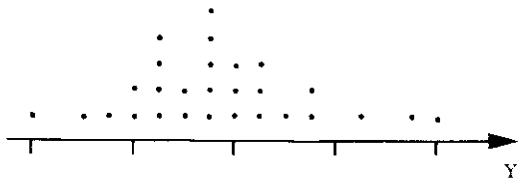
Der er i litteraturen foreslået et utal af alternative måder at afbilde data på, hver med deres oplagte og påståede fordele og ulemper. Gode referencer er Tukey (1977), Champers *et al.* (1983), duToit *et al.* (1986) og specielt Tufte (1983). Standard computerprogrammer byder også på et stort antal muligheder. Kun i sjældne tilfælde er det særligt afgørende, præcis hvilken afbildningsteknik man benytter blandt sobre muligheder. Det afgørende er, at man overhovedet foretager den grafiske analyse samvittighedsfuldt. I nedenstående afsnit omtales et udvalg af gode muligheder.

3.1 Én variabel

Prikdiagram

En simpel og effektiv måde at få overblik over et mindre datamateriale på, er ved hjælp af et prikdiagram. Diagrammet består, som vist i figur 3.1, af at hver måling repræsenteres ved en prik ud for målеværdien på den vandrette akse, her kaldet Y-aksen, da vi i afsnit 2.1 betegnede målinger med Y. Hvis antallet af målinger er større end omkring 10-15, er det i reglen nødvendigt for at bevare klarhed at stable punkter med næsten samme værdi. Dette gøres ved at afrunde målingerne, og afbilde punkter med samme afrundede værdi over på hinanden.

Figur 3.1.
Prikdiagram.

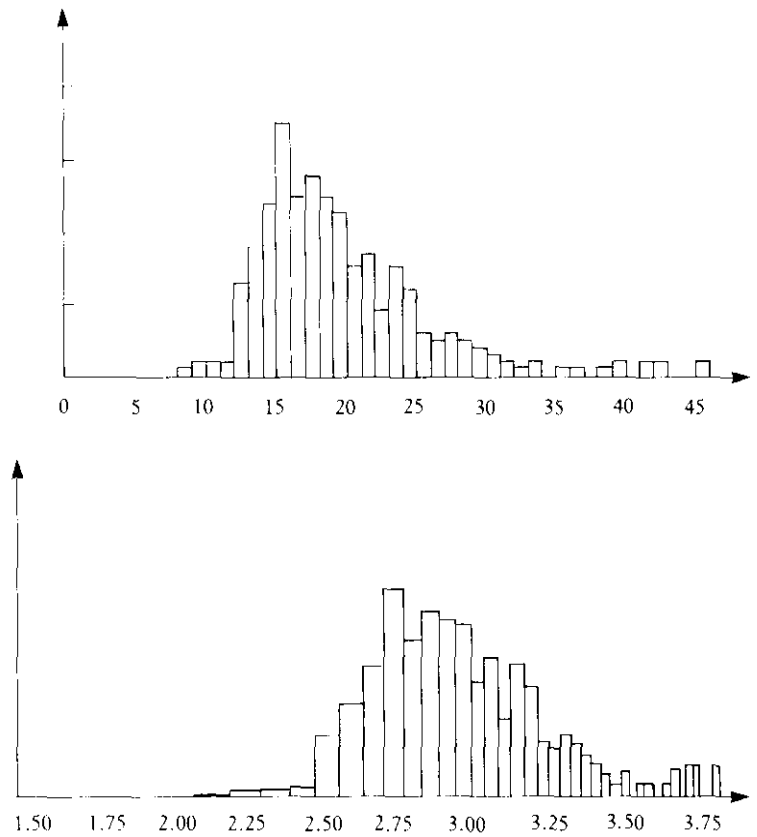


Et prikdiagram giver et godt indtryk af målingernes generelle størrelse, deres niveau, samt deres variation. Prikdiagrammer vil ofte kunne afsløre grove datafejl, f.eks. kommafejl, kopieringsfejl m.v., så disse fejl kan rettes før en kvantitativ dataanalyse. Afvigende datapunkter behøver ikke være fejl, men kan være særligt interessante. Deres tidlige afsløring giver mulighed for, om ønsket, at søge dem dubleret i dataplanlægningen. Et prikdiagram kan også give et indtryk af hvorvidt målingernes fordeling er skæv, topuklet eller besidder andre markante karakteristika.

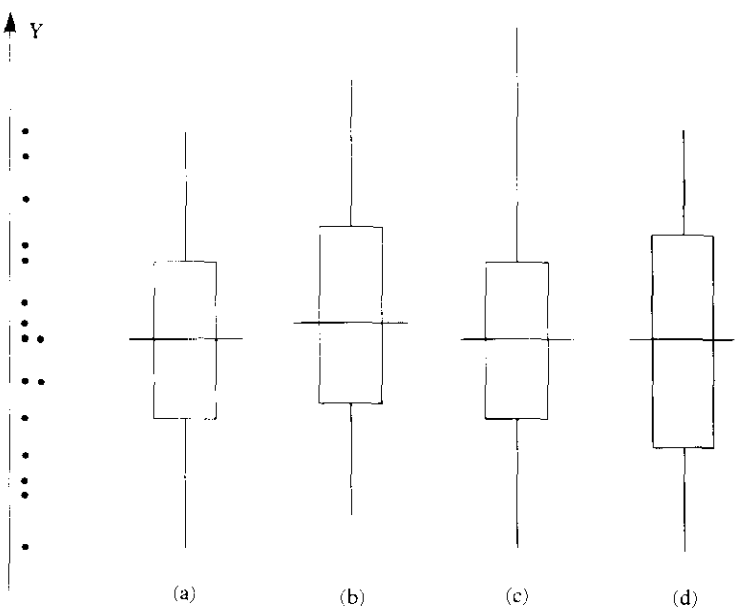
Histogram

Er antallet af målinger større end omkring 20-30, er et histogram i reglen at foretrække for et prikdiagram. I histogrammet afbildes antallet af målinger i valgte intervaller som søjler, således at søjlernes

Figur 3.2.
Histogrammer over samme datamateriale, søjlearealerne er proportionale med observationsantal. (a) original skala (b) logaritmisk skala med samme intervalgrænser.



Figur 3.3.
Boxplot for fire datasæt. Prikdiagrammet til venstre viser datasæt (a).



areal er proportional med antallet af målinger. Ofte vælges intervallerne af samme længde, hvorved søjlehøjderne bliver proportionale med antallet af målinger. Antallet af intervaller må hverken vælges for stort eller for lille, da man så mister henholdsvis overblik eller finstruktur. Følgende tommerfingerregel giver i reglen gode histogrammer:

$$\text{Antal intervaller} = 1 + 3{,}3 \log_{10} (\text{antal målinger}) \qquad (3.1)$$

Det nøjagtige antal intervaller er ikke kritisk, man kan øge eller sænke i forhold til (3.1) efter bekvemmelighed. Har man f.eks. 80 målinger findes $1 + 3{,}3\log_{10}(80) = 7{,}28$. Man vil nu afhængig af en praktisk akseinddeling vælge et antal intervaller mellem måske 5 og 10.

Histogrammet giver overblik på samme måde som prikdiagrammet. Specielt kan man som vist i figur 3.2a afsløre en betydelig skævhed i målingernes fordeling. Tyngdepunktet for fordelingen ligger til venstre i forhold til variationsområdets midte; eller, sagt på anden måde, fordelingens højre hale er længere og tungere end den venstre. Sådan en skævhed mod højre kan ofte elimineres ved en logaritmetransformation, d.v.s. at vi i stedet for at arbejde med de oprindelige målinger benytter deres logaritmer.

Fordelingens form

Transformeret form

Figur 3.2b er et histogram for de logaritmetransformerede data. Man har udregnet $Z_i = \log Y_i$, $i = 1,2,\cdots,n$, og så opfattet de transformerede værdier Z_i-erne, som data fremfor rådata Y_i-erne. Der er nu tale om en praktisk taget symmetrisk fordeling. Det bemærkes, at de oprindelige intervalinddelinger fra figur 3.2a er bevaret i figur 3.2b. Herved bliver intervallerne i 3.2b af forskellig længde, og da det er søjlernes arealer, der skal være proportionale med deres respektive antal målinger, bliver højderne det ikke. Dataplot eller residualplot kunne i andre situationer pege på fordele ved andre transformationer.

Boxplot

Et tredje plot, der giver godt overblik over et talsæt, er »boxplot«, som der er vist 4 eksempler på i figur 3.3. Et »boxplot« defineres af fem punkter: placering af »kassens« top og bund, dens gennemskærende linie og endepunkterne for de liniestykker, der stikker ud af kassens ender. Kassens bredde har ingen betydning. Disse fem punkter refererer til en vist eller underforstået lodret akse, i virkeligheden den samme Y-akse som før, blot stillet lodret. Det øverste endepunkt markerer den største måling. Kassens top markerer grænsen mellem de 25% største målinger og de 75% mindste, d.v.s. den øvre 25% fraktil, eller den nedre 75% fraktil, ofte blot betegnet 75% fraktilen. Kassens gennemskæring markere 50% fraktilen, eller medianen. Kassens bund markerer 25% fraktilen (den nedre), medens det nedre linieendepunkt markerer den mindste måling.

Tolkning

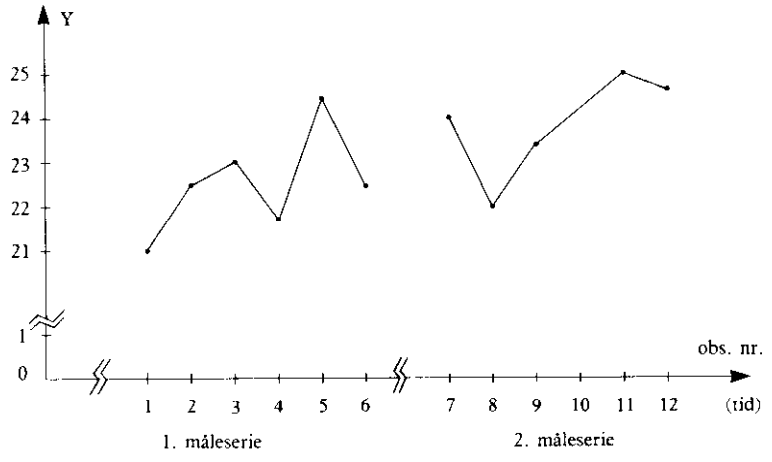
Datas niveau, spredning og skævhed træder tydeligt frem i boxplot. Fordelingen afbildet i figur 3.3a er symmetrisk. 3.3b er skæv mod høje værdier (mod højre), 3.3c ligner 3.3a bortset fra den meget høje maximalværdi (en fejlmåling)? Også spidshed (topstejlheden) af fordelingen fornemmes tydeligt. I boxplottet 3.3d er de fire afstande

mellem de fem markerende punkter ca. lige store, hvilket reflekterer en jævn fordeling. For en normalfordeling vil der også være symmetri omkring medianen, men liniestykkerne vil tilsammen være længere end kassen. Der forekommer forskellige varianter af boxplottet, nogle computerprogrammer markerer de ekstreme målinger separat og lader liniestykkerne fra kassen stoppe ved sidste måling før de ekstreme værdier.

Tidsplot

Et fjerde vigtigt plot af måleværdier, er et tids- eller sekvensplot. Her er måleaksen igen lodret, medens tids- eller nummereringsaksen er vandret. Overblikket i et sådant diagram fremmes ofte af at forbinde punkterne, som det er sket i figur 3.4. Er målingerne ikke

Figur 3.4.
Tidsplot af 11 observationer målt i to serier; observation nummer 10 mangler.

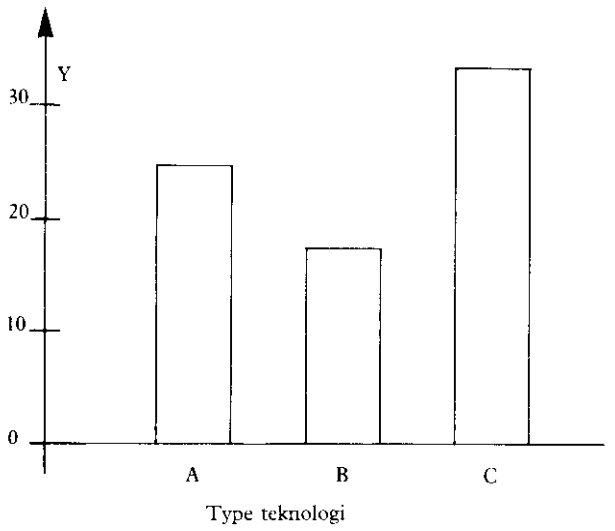


taget med samme tidsafstand må dette fremgå af plottet. I figur 3.4 er målingerne taget i to omgange og i anden omgang mangler den fjerde måling. Når det er muligt, bør måleaksen (den lodrette Y-akse) skære tidsaksen i sit nulpunkt. Det må klart markeres, f.eks. som gjort i figur 3.4, hvis sådan skæring ikke er praktisk muligt. Dette kan hænde hvis variationen er lille i forhold til niveauet, og variationen ønskes tydeligt afbildet, eller hvis Y-aksen er logaritmisk, eller af anden grund.

Søjlediagram

Et hyppigt forekommende plot som grafisk minder om histogrammet, men som oftest har mere til fælles med tidsplottet, er søjlediagrammet. Her vises målingernes størrelse som søjlehøjde, d.v.s. refererende til en lodret måleakse. Er den vandrette akse en sekvens- eller tidsakse, er plottet ækvivalent med tidsplottet. Imidlertid er det nu et ufravigeligt krav, at måleaksen skæres af tidsaksen i sit nulpunkt, d.v.s. at søjlernes fod markerer nulpunktet. Ellers bliver søjlediagrammet vildledende. På den vandrette akse kan også blot være markeret målesteder, procestype, eller anden kategorisering, hvor rækkefølgen fra venstre mod højre er arbitrær. Målingerne behøver ikke være kontinuerte måletal, men kan f.eks. også være antal (antal pumpestop, antal arbejdsskader etc.). For at skelne søjlediagrammet fra histogrammet, bør der være afstand mellem søjlerne som i figur 3.5.

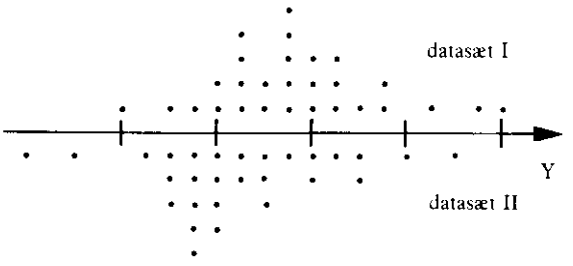
Fig. 3.5
Søjlediagram.



3.2 To variable

Behov for plot med to variable kan opstå på forskellig måde. Der kan dels være tale om samme type måling foretaget på to alternative måder (to typer teknologi; en proces med og uden indgreb; to tider, etc.), dels om to forskellige variable målt på samme procesanlæg,etc. I den første situation, er det hyppigt oplysende at kunne sammenligne prikdiagrammer, histogrammer eller boxplot for de to situationer.

Figur 3.6.
Prikdiagram med to datasæt.



To plot på én akse

For prikdiagrammernes vedkommende tegnes de som nævnt ovenfor, men kan hensigtsmæssigt afbildes på de to sider af samme akse, som vist i figur 3.6. Selv en ret lille forskydning eller variationsforskel afsløres let af det menneskelige øje.

Helt tilsvarende for histogrammet, her afbilder man også de to plot ryg mod ryg. Undertiden foretrækkes det, at lade måleaksen være lodret, som vist i figur 3.7. Boxplottene er umiddelbart egnede til denne type sammenligning, jf. figur 3.3.

X-Y plot

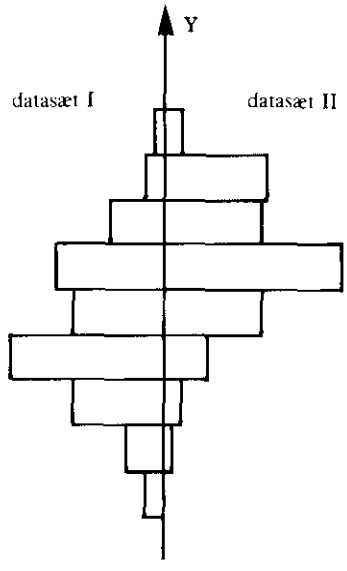
I den anden situation, hvor en eventuel relation mellem to variable ønskes afsløret, er et sædvanligt X-Y-plot egnet, figur 3.8. Grafisk udføres det som et tidsplot, idet man dog ikke bør forbinde punkterne. Igen, hvis det er muligt, bør akserne skære hinanden i deres nulpunkter, og afvigelser herfra må markeres tydeligt. Akseinddelingerne må være ækvidistante, eller på anden måde enkelt og

systematisk gennemskuelige (f.eks. logaritmiske skalaer). Er de to variable målt i samme enhed (f.eks. °C) bør akseinddelinger tilstræbes ens. Er der tale om en årsag og virkning sammenhæng, bør årsagsvariablen afbildes ud ad den vandrette akse, og virkningen op ad den lodrette. Er der tale om to passivt observerede variable, er det et godt princip, at akseenhederne skaleres således, at variationerne i begge retninger er lige store.

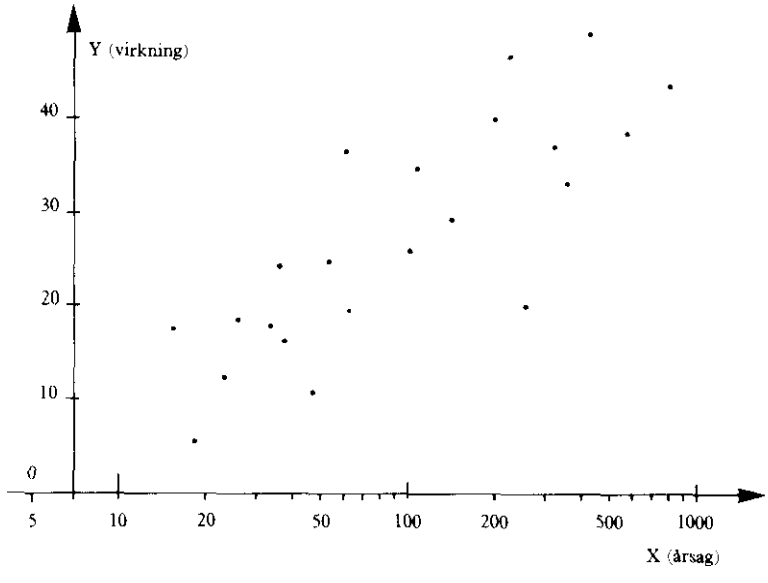
Tidsplot

Tidsplot med to variable kan enten udføres som illustreret i figur 3.9, hvor to tidsplot, et for X-variablen og et for Y-variablen, er tegnet oven i hinanden med fælles tidsakse. Er måleenheden for X

Figur 3.7.
Histogram med to datasæt.

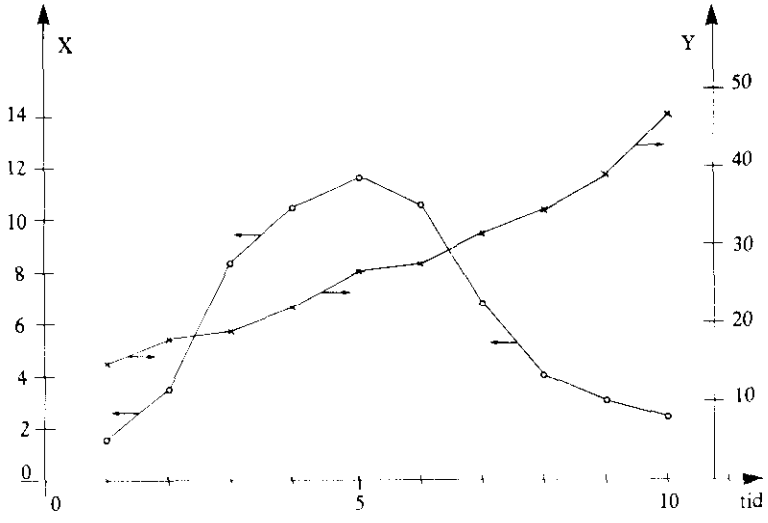


Figur 3.8.
X-Y diagram



og Y forskellige, kan de to kurver af forbundne punkter som vist refereres til forskellige lodrette akser.

Tidsaspektet kan også belyses ved at nummerere punkter i X-Y plottet, figur 3.8, enten ved at anføre nummeringen ved siden af punktet, eller benytte tal som markering fremfor blot en prik eller et kryds. Er antallet af punkter stort kan man gruppere punkterne så de tidligste punkter markeres med »0«, de næste med »1«, o.s.v. op til »9«.



Figur 3.9.
Tidsplot for to variable med to måleakser.

3.3 Flere variable

Ved sammenligning af målinger stammende fra forskellige oprindelser (flere typer behandling, flere slags teknologi, flere anlæg, flere tidspunkter, etc.) kan de individuelle prikdiagrammer eller histogrammer bedst sammenlignes, når man sørger for, at måleakserne har samme inddelinger; men der findes næppe nogen særlig fiks grafisk løsning på denne sammenligning. Box-plottene er i reglen mest velegnede her, arrangeret som figur 3.3.

Mange X-Y plot

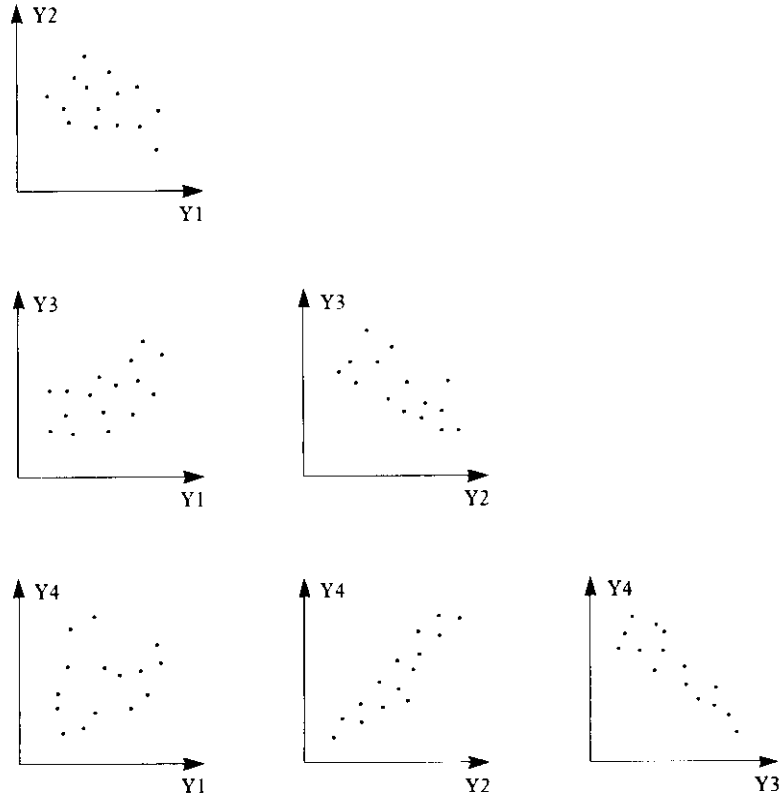
Ved afbilding af flere variable mod hinanden kan det give et godt overblik at arrangere alle kombinationerne af X-Y plot som vist i figur 3.10 for fire variable, Y1, Y2, Y3 og Y4.

Plot med mange variable bliver let uoverskuelige, men kan ved talentfuldt valg af afbildningsteknik udføres med stor klarhed; mange inspirerende eksempler herpå er givet i speciallitteraturen, f.eks. i Tufte (1983). De fleste udnytter særlige strukturer i data, hvorfor anvisninger af mere generel karakter næppe kan gives i oversigtsform.

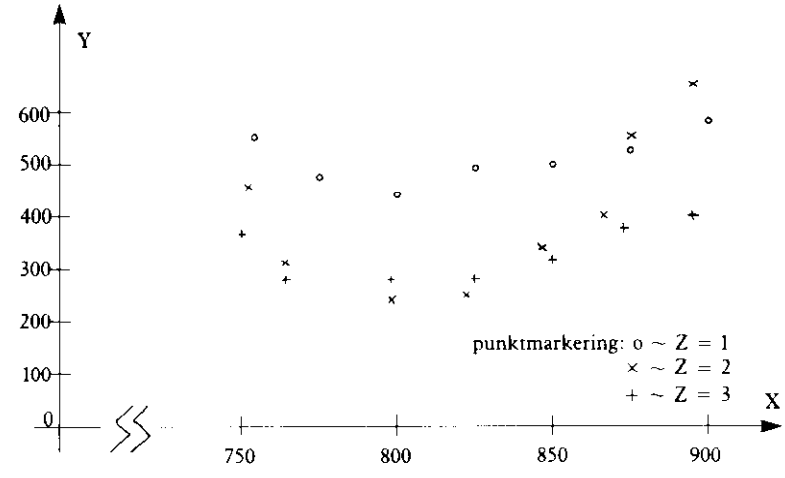
Diskret variabel

For tre variable hvoraf den ene er eller gøres diskret med et begrænset antal værdier/kategorier, kan X-Y plot udføres med forskellig punktmarkering afhængig af den diskrete variabel. I figur 3.11 kan f.eks. Y være en støvmængde i skorstensrøg, X forbrændingstemperatur, og Z være alternative rensningsforanstaltninger, eller måske luftoverskud som indreguleres på 3 forskellige niveauer.

Figur 3.10.
Opsætning af alle X-Y plot mellem Y1, Y2, Y3 og Y4.

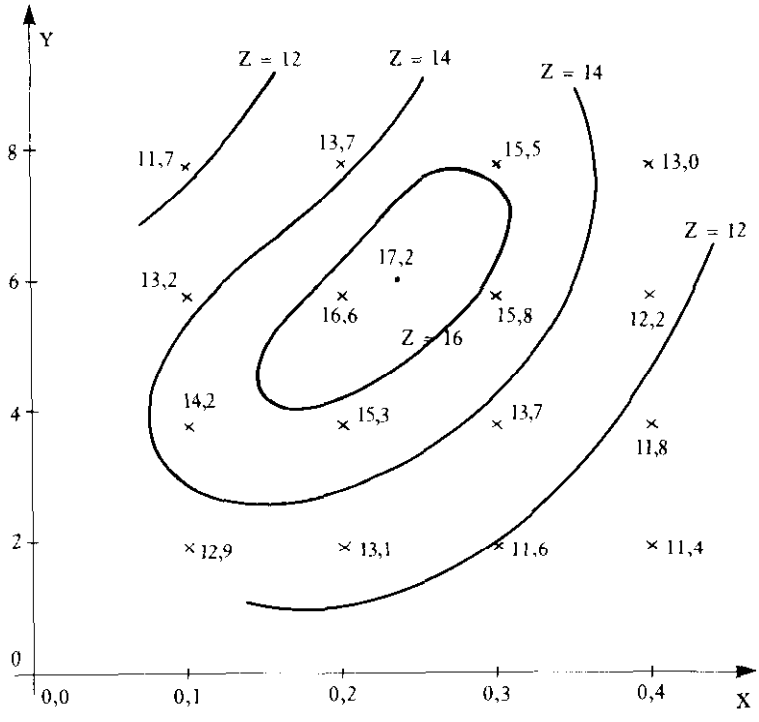


Figur 3.11.
X-Y plot med Z som diskret variabel.



Konturplot

For tre kontinuerte variable kan konturplot som i figur 3.12, vise én variabels afhængighed af to andre. De markerede punkter er målinger og de indtegnede konturlinier er »højdelinier« fra et tilpasset polynomium, eller anden tilpasset flade. Det beregnede maksimum er også markeret.



Figur 3.12.
Konturplot. De indtegnede højdekurver markerer fladen der er tilpasset de markerede punkter.

4. Kvantitativ statistisk dataanalyse

Inferens

De to vigtigste kvantitative værktøjer i statistisk dataanalyse er statistisk parameterestimation og statistisk hypotesetestning. Disse discipliner er hovedingredienser i statistisk inferens, d.v.s. kvantitativ statistisk slutning eller dragen lære af data.

Den store tilgængelighed af computerprogrampakker, som kan udføre estimation og testning gør selve beregningsarbejdet trivielt, således at så at sige enhver kan producere en »statistisk analyse« af et næsten vilkårligt datasæt. Denne trafik har næppe ensidigt bedret statistikkens omdømme; og hvor det er sket, ikke altid berettiget.

Fejlfortolkning

Computerprogrammerne afhjælper ikke behovet for at forstå, hvad metoderne kan bruges til, og hvad de ikke må misbruges til. Fejlfortolkning af statistiske analyser og brug af forkert statistisk teknik er desværre stærkt udbredt også blandt ellers respekterede eksperter; men stor fagkundskab inden for én profession indebærer ikke automatisk samme sikkerhed i udøvelsen af en anden.

Mange af de almindeligt forekommende fejl er af fortolkningsmæssig karakter, eller bunder dybere i en manglende forståelse for, hvad statistisk inferens egentlig er. Dette berettiger nærværende kapitel.

Overfladisk set har statistik meget til fælles med matamatik, i og med at begge fag i en vis forstand handler om tal. Matematik som fag er opbygget omkring logiske følgeslutninger, resultater er derfor rigtige eller forkerte, konklusioner er beviste eller ikke-beviste, og fejlkonklusioner er udtryk for brist i udførelse af det matematiske arbejde. Matematik er deduktiv af natur. Man slutter fra det generelle til det specielle, fra teori til praksis, og med mindre man begår fejl, får man definitionsmæssigt ret.

Matematisk deduktion

Statistisk induktion

Statistik derimod er induktiv. Man slutter fra det specielle til det generelle, fra data til model, fra praksis til teori. Data kan kun give en begrænset information om det system der studeres. Ved statistisk inferens søger man at slutte sig til, hvorledes den datagenererende proces (det generelle) mon er indrettet, siden den producerer de observerede data (det specielle). Fra nødvendigvis ukomplette og fejlbehæftede data, kan man aldrig være absolut sikker på, om ens inducerede model er korrekt. Muligheden for fejlslutning er til stede selv når statistikkens spilleregler overholdes. Analyseret ved sunde statistiske metoder kan samme datasæt derfor give anledning til forskellige fortolkningsmuligheder. Dette faktum er i en negativ betydning blevet opfattet som ulogisk, hvilket kun kan skyldes en fejlagtig opfattelse af, at statistik burde være deduktiv. Men statistikkens væsen er altså induktiv, og heraf flyder netop dens særegne anvendelighed i læreprocessen, i forskning og udvikling.

Spillerum for fortolkning

Usikkerhed hidrørende fra begrænsede og støjbehæftede data tillader ikke vilkårlige fortolkningsmuligheder, men et vist spillerum hvis størrelse statistikken oplyser om. Derfor er det essentielt for god statistisk praksis, så eksplicit som muligt at angive usikkerhedsmomenter, d.v.s. rimelige spillerum for fortolkningsmuligheder. Er disse ubrugeligt vidtfavnende må flere data indsamles, og indsnæv-

ringen foregår hurtigst ved planlagte forsøg. Ofte kan man på forhånd skønne sig til, hvor mange observationer der er nødvendige for at undersøgelsen kan få interesse. Det spørgsmål behandles også nedenfor.

4.1 Statistisk hypotesetestning

Statistisk testning handler om, at en hypoteses plausibilitet skal vurderes i lyset af data. Hypotesen, (også kaldet nul-hypotesen), gives ved testningen fordel af al tvivl i data, d.v.s. at hvis den ikke decideret modsiges af data, vil den ifølge almindelig statistisk sprogbrug blive »accepteret«, og kun ellers »forkastet«. Det ville formentlig lede til færre misforståelser at erstatte udtrykket »accepteret« med »ikke-forkastet«, idet den udeblevne forkastelse kun betyder at hypotesen kan være sand for så vidt angår det aktuelle datasæt, ikke at den er det. Man skal derfor være påpasselig med at formulere og teste nulhypoteser. En hypotese, der på forhånd er utroværdig, bliver ikke mere sand af, at den er »accepteret« ved statistisk test.

Accept af hypotese

Undlade test

I de fleste sammenhænge bør man simpelthen undlade at udføre test, som ikke på forhånd synes rimelige. Gør man det alligevel, må man ikke blot erindre sig selv om den korrekte fortolkning af »accept«, men også i sin rapportering sikre, at læseren ikke kan misledes. Det er en form for uhæderlighed, inkompetence, eller unyttig indforståethed at formulere en rapporttekst, så den kan læses derhen, at en »accepteret« hypose er statistisk bevist sand.

Ideen bag testning

Ideen bag hypotesetestning er mere detaljeret følgende: Man tager udgangspunkt i en nulhypotese H_0, som repræsenterer en mulig sandhed for, hvorledes data er genereret. Altså en speciel version af den statistiske model. Heroverfor står en alternativ hypotese H_1, typisk mere kompliceret og generel; altså en bredere statistisk model. Der udregnes nu en teststatistisk T, som er et tal, der efter en given regneforskrift findes fra data. Regneforskriften er konstrueret således, at teststatistikkens sandsynlighedsfordeling er velkendt under forudsætning af, at nulhypotesen H_0 er sand. Det vil sige, at fik man en mængde datasæt, alle genereret ifølge H_0-modellen, så ville T-erne udgøre den kendte fordeling, referencefordelingen. Regneforskriften er videre søgt konstrueret således, at hvis teststatistikken T udregnes på grundlag af data, der ikke stammer fra H_0-modellen, men fra H_1-modellen, så vil T-værdien afvige mest muligt fra referencefordelingens typiske værdier.

Teststatistik

Referencefordeling

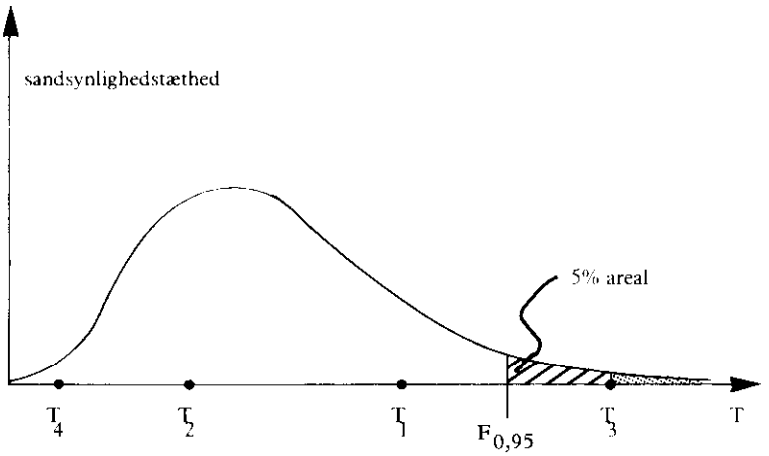
Testningen går nu ud på at sammenligne den aktuelle T-værdi med referencefordelingen. Falder T i et typisk område, er der altså intet tegn på, at T ikke stammer fra denne fordeling. Der er med andre ord ikke kastet tvivl på H_0, og man »accepterer« derfor H_0. Falder T utypisk i forhold til referencefordelingen, d.v.s. i et værdiområde, der er sjældent eller helt usandsynligt, tages det som udtryk for, at T ikke tilhører denne fordeling. T må altså stamme fra en anden fordeling, det vil sige at man må forkaste H_0.

Signifikans

Et kvantitativt mål for utroværdigheden af nulhypotesen H_0 er signifikansniveauet. Signifikansniveauet måler hvor ekstremt teststatistikken falder i forhold til referencefordelingen.

Hvis fordelingen i figur 4.1 er T's referencefordeling, viser den, hvorledes T varierer fra datasæt til datasæt, når disse alle er genereret under H_0; d.v.s. af en proces eller et system, der følger nulhypotesens model. Videre er regneforskriften for T indrettet således, at hvis data ikke stammer fra H_0-modellen, så vil T tendere mod større værdier. I figur 4.1 ligger de mulige teststørrelser T_1 og T_2 typisk i forhold til referencefordelingen, de kaster ikke tvivl om H_0. T_3 derimod ligger ret højt, og er altså en sjælden hændelse under H_0, og kaster derfor tvivl over H_0 's sandhed.

Figur 4.1.
Teststatistikken T's referencefordeling.



Niveau α

En test siges at være udført på signifikansniveauet α (f.eks. lig 5%), hvis man vælger en grænse i referencefordelingens sandsynlighedsmasse så brøkdelen α ligger i forkastelsesområdet. På figuren er α valgt lig 5%, d.v.s. at grænsen mellem T-værdier der fører til accept/forkastelse af H_0, ligger på fordelingens 95% fraktil ($F_{1-\alpha} = F_{0,95} = F_{95\%}$). Det betyder, at i tænkte gentagne anvendelser af testen, ville man i 5% af tilfældene forkaste H_0, selv hvis denne hypotese var sand. Dette er naturligvis en fejl, endda en grov fejl ud fra princippet om, at H_0 skal have fordel af al tvivl. Fejlen benævnes type I fejl og begrænses ved at vælge en lav værdi for α. Hyppigst ser man α valgt lig 5%, men der er intet statistisk, der anbefaler netop denne værdi; valget bør foretages ud fra den faglige sammenhængs behov.

Type I fejl

En udregnet teststatistik med værdien T_3 fører altså til forkastelse af H_0 på $\alpha = 5\%$ signifikansniveau. Eller, sagt på anden måde, data viser signifikant afvigelse fra H_0 på niveauet $\alpha = 5\%$. Værdier som T_1, T_2 og T_4 fører derimod til »accept« af H_0 på signifikansniveau $\alpha = 5\%$. Det bemærkes, at T_4 i realiteten er en usædvanlig værdi i relation til nulhypotesens referencefordeling. Som teststatistikken var konstrueret, førte afvigelse fra H_0 imidlertid generelt til større værdier for T, hvorfor T_4 ikke skal føre til forkastelse. Den lave værdi udtrykker usædvanlig god overensstemmelse mellem data og H_0-model.

p-værdi

Frem for at benytte en fast α-værdi, kan man i en testsituation angive på hvilket niveau hypotesen netop vipper mellem accept og forkastelse. Dette niveau kaldes »p-værdien« og findes som referencefordelingens sandsynlighedsmasse, der ligger mere ekstremt end den observerede teststatistik. p-værdien for T_3 i figur 4.1 er $p \simeq 0{,}02 = 2\%$. Med teststatistikken T_3 vil hypotesen altså forkastes for alle niveauer $\alpha > p \simeq 2\%$. For T_1 er p-værdien klart større, ca. 0,15; for T_2 og T_4 findes henholdsvis $p \simeq 0{,}75$ og $p \simeq 0{,}99$.

Til trods for, at T_1, T_2 og T_4 ligger i acceptområdet for teststatistikken, kan de være udregnet fra data, der ikke stammer fra H_0-modellen. Det er en væsentlig pointe, at fordelingen af T givet at H_0 er usand (H_1 er sand), ikke fremgår af figur 4.1, og at dette spørgsmål slet ikke kommer ind i billedet ved beslutning om accept eller forkastelse af H_0. For nogle modeltyper kan man udregne chancen for, at H_0 accepteres, selv hvor H_0 er usand. Det er sandsynligheden for at T falder i acceptområdet for H_0's referencefordeling, når T i virkeligheden stammer fra en anden sandsynlighedsfordeling, nemlig i overensstemmelse med H_1. Den fejl der herved begås, kaldes fejl af type II, og regnes mindre alvorlig end type I, på grund af det grundlæggende synspunkt, der favoriserer H_0. Risikoen for at begå fejl af type II betegnes undertiden β.

Type II fejl

Styrken 1-β

Ved styrken af en signifikanstest, forstår man dens ønskelige evne til at forkaste H_0, når H_1 er sand. Styrken er lig 1-β. Grunden til at α ikke reduceres til en ekstremt lille værdi – i figur 4.1 ved at flytte forkastelsesgrænsen langt mod højre – er, at β herved vil vokse uforholdsmæssigt. Uagtet om H_1 skulle være sand, vil der opstår stor sandsynlighed for at T falder i det stærkt udvidede acceptområde, d.v.s. at næsten ethvert praktisk forekommende datasæt vil føre til accept af H_0. Med andre ord, testens styrke til at forkaste H_0, når H_0 faktisk er forkert, bliver urimeligt lille.

En statistisk tests styrke afhænger ikke kun af valget af α, men også af teststatistikkens konstruktion og datamængden. Megen grundforskning har drejet sig om at finde så stærke statistiske test som muligt, men der er en øvre grænse herfor. Benyttet rigtigt og på egnede datasæt, er de i computerprogrammer almindeligt tilgængelige testmetoder så gode, som de kan skaffes.

Datamængde

Hvad angår datamængden, skal man være klar over, at styrken af enhver statistisk test bliver ringe, hvis datamaterialet er småt. Det betyder, at et spinkelt datamateriale vil have tendens til at give accept at en foreslået nulhypotese, uanset om den er sand. En »accepteret« hypotese er således langt fra bevist, når man har med spinkle datamængder at gøre. Ja, en næsten vilkårlig hypotese kan »accepteres« blot datamængden gøres tilstrækkeligt lille og usikker. Disse statistiske synspunkter strider ikke mod sund fornuft, men står i et vist modsætningsforhold til, hvordan statistiske test undertiden misbruges.

Når det er muligt, bør styrken af en statistisk test angives foruden det benyttede signifikansniveau. Desværre er der ikke udbredt tradition herfor. En anden forbedring over almindelig praksis ville være angivelsen af p-værdien foruden eller – på længere sigt – i stedet for blot signifikansniveauet α.

Visse teststatistikker har ikke en ensidigt rettet tendens til at blive større hvis H_0 er usand, men tenderer enten mod særligt store eller særligt små værdier afhængigt af afvigelser fra H_0. I disse tilfælde testes med en øvre og en nedre grænse, der tilsammen afskærer en sandsynlighedsmasse af størrelsen α. Sådanne test kaldes tosidede test, modsat den overfor forklarede ensidede test. Men iøvrigt er tankegangen ens for de to slags test.

Tosidet test

For klarhedens skyld er der overfor talt om T's fordeling under H_0 og H_1, som om der fandtes én fordeling i hvert tilfælde. I mange testsituationer vil såvel H_0 som H_1 kunne give anledning til hele klasser af fordelinger. De væsentlige statistiske ideer ændres imidlertid ikke heraf. En mere teoretisk detaljeret gennemgang kan bl.a. findes i Rao (1973), Hogg & Craig (1970), eller på dansk i Conradsen (1984).

Test overbrugt

Signifikanstest er nok en noget overbrugt statistisk teknik. Mest misledende i situationer hvor nulhypotsen ikke er specielt troværdig a priori, og hvor den modtager en ufortjent statistisk velsignelse i form af en korrekt fundet, men overfortolket accept fra et svagt datamateriale.

I virkeligheden er en forkastelse af en nulhypotese et mere konklusivt udsagn end en accept. Forkastelsen betyder, at hypotesen er i modstrid med data, så man *må* forkaste på det givne signifikansniveau. En accept derimod siger kun, at man *kan* acceptere den formulerede nulhypotese; den udelukker ikke, at man også ville kunne aceptere andre mulige nulhypoteser på samme datagrundlag.

Asymmetri i test

Den markante asymmetri mellem accept og forkastelse, er en konsekvens af asymmetrien mellem nulhypotesen H_0 og den alternative hypotese H_1. Det svarer helt til problemstillingen i en retsal: Er den anklagede skyldig indtil bevist uskyld, eller uskyldig indtil bevist skyld? I de fleste sager er nulhypotesen at den anklagede er uskyldig. Kan intet bevises, hverken for eller imod, frikendes vedkommende, som accept af nulhypotesen.

Formulering af hypotese

Ved meningsfuld brug af statistisk signifikanstestning må man bevidst formulere nulhypotesen således, at det bliver sagligt korrekt eller mest oplysende, at tvivl kommer denne hypotese til fordel. Ved indførelse af ny, delvis uprøvet teknologi, som forventes mere miljøvenlig end den gamle, vil det være interessant at formulere en nulhypotese H_0, om at den gamle teknologi er mest miljøvenlig. Herefter vil man ved signifikanstest håbe at kunne forkaste denne hypotese og helst på et lavt signifikansniveau α (en lille α-værdi er ensbetydende med stærk signifikans). Dermed bliver det konklusivt statistisk bevist at teknologiskiftet er miljømæssigt fordelagtigt. En accept af H_0 må derimod ikke fortolkes som bevis på at den gamle teknologi faktisk er mest miljøvenlig. For at kunne drage denne konklusion med statistisk sikkerhed, må en nulhypotese om at den nye teknologi er bedst forkastes ved statistisk test. Man kan komme i den situation,

Modstridende accepter

at begge modstridende nulhypoteser accepteres ved brug af samme datasæt. Dette betyder, at datasættet er utilstrækkeligt til at afgøre hvilken teknologi, der er miljømæssigt bedst.

Overbevisende accept

Signifikanstest kan også være berettiget, når hypotesen, der ønskes testet, på forhånd synes plausibel. En accept vil virke overbevisende hvis data er planlagt således, at hypotesen sættes på en virkelig prøve. Der er f.eks. ikke meget at hente i accepten af en nulhypotese som siger, at temperaturen hvorunder en proces forløber ingen effekt har på dannelsen af giftigt spildstof, hvis temperaturen kun har været varieret meget lidt under måleforsøget. Har man derimod aktivt varieret den over et stort temperaturinterval, bliver accepten langt mere overbevisende.

4.2 Statistisk parameterestimation

Statistisk estimation handler om fra data at bestemme en eller flere parametre, d.v.s. de ukendte konstanter i den statistiske model, jf. afsnit 2.1. Estimation begrænser sig ikke til at finde en enkelt talværdi pr. parameter, men må indeholde en specifikation af den usikkerhed, der knytter sig til dette estimat.

Ukendte parametre

Den statistiske model formuleres således, at spørgsmål og uklarheder, som den eksperimentelle undersøgelse skal belyse, udtrykkes i parametre. Det kunne f.eks. være ændringer i miljøbelastning ved indførelse af ny teknologi, hvor belastningen pr. produktions-batch ved gammel teknologi kunne kaldes μ_0 og ved ny teknologi μ_1. De udførte målinger af belastningen er

$$\begin{cases} Y_{0,i} = \mu_0 + e_{0,i} & i = 1,2,\cdots,n_0, \\ Y_{1,i} = \mu_1 + e_{1,i} & i = 1,2,\cdots,n_1, \end{cases} \tag{4.1}$$

hvor $e_{0,i}$ og $e_{1,i}$ er statistisk fejl, som måske kan antages normalfordelt med middelværdi 0 og varians henholdsvis σ_0^2 og σ_1^2. Parameteren af interesse bliver

$$\delta = \mu_1 - \mu_0, \tag{4.2}$$

og den statistiske opgave er nu at estimere δ. Denne analyses beregninger behandles i afsnit 5.2, for nærværende skal de principielle punkter trækkes frem.

Punktestimat

Dataanalysens estimerede værdi for δ, punktestimatet, betegnes $\hat{\delta}$. Desuden leverer den statistiske dataanalyse en estimeret standardafvigelse for $\hat{\delta}$, som benævnes $\hat{\sigma}_\delta$. Standardafvigelsen er et eksplicit usikkerhedsmål og kan opfattes som den naturlige målestok ved angivelsen af intervallet omkring $\hat{\delta}$ for plausible δ-muligheder. Det skulle være såre heldigt, om $\hat{\delta}$ præcist rammer δ. Det interessante spørgsmål er, om δ kan ligge så langt fra estimatet $\hat{\delta}$, at det får væsentlige konsekvenser i vurderingen af den nye versus den gamle teknologi.

Konfidensinterval

Til vurdering heraf konstrueres et konfidensinterval af formen

$$[\hat{\delta} - f\hat{\sigma}_\delta \; ; \; \hat{\delta} + f\hat{\sigma}_\delta]. \tag{4.3}$$

2. Statistisk analyse, model og data

Konfidensgrad

Statistisk nødvendighed

En vis forståelse for statistisk tankegang og metode er nødvendig for alle ingeniører og andre, der seriøst beskæftiger sig med procesforbedringer i miljømæssige såvel som andre henseender. Grunden hertil fremgår af følgende korte fem-trins argument (jf. Hunter, 1981):

1) Det er uhensigtsmæssigt at forestille sig, at en aktuel produktion på alle måder foregår så ideelt, at forbedringer i form af procesoptimering, produktudvikling og/eller teknologifornyelse på forhånd kan udelukkes.

2) Forbedringer kræver nødvendigvis ændringer, ellers bevares selvsagt status quo.

3) Rationelle ændringer kan ikke foregå alene på grundlag af formodninger, strømninger, eller fordomme, men må bygge på fakta, d.v.s. data.

4) For at undgå spild, må fremskaffelse og fortolkning af data foretages så effektivt som muligt.

5) Statistisk tankegang og metode handler netop om det i punkt 4) krævede.

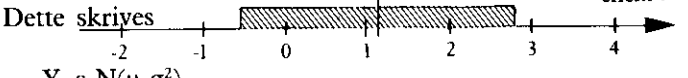
2.1. Den empiriske læreproces

Læreprocessen

Indsigt og viden, der er baseret på fakta, stammer – direkte eller indirekte - fra observerede data. Når viden ud over den eksisterende ønskes erhvervet, må den empiriske læreproces fortsættes, d.v.s. at nye data må indsamles og deres informationsindhold ekstraheres. Det gælder om, at læreprocessen forløber så effektivt som muligt, så erhvervelsen af ny viden foregår hurtigt og økonomisk. Hertil kan statistisk tankegang yde et stort bidrag. Når et datasæt skal fortolkes, er det en stor hjælp, hvis de væsentlige aspekter ved data kan udnyttes gennem en model.

Model

Har man for eksempel foretaget en lang række måling Y_i, i = 1,2,···,n af en bestemt miljøvariabel og fundet en vis usystematisk variation omkring et niveau, kan man som model forestille sig, at Y_i er normalfordelt omkring en middelværdi, μ, med variansen, σ^2. Dette skrives



$$Y_i \varepsilon N(\mu,\sigma^2), \qquad (2.1)$$

I praksis kendes naturligvis kun ét interval, nemlig det der er konstrueret fra de aktuelle data. Og hvorvidt netop dette omslutter δ vides ikke; men det kan opfattes som stammende fra en mængde konfidensintervaller hvoraf andelen $1-\alpha$ (95%) gør. Vi nærer derfor en konfidens eller tiltro af størrelsen $1-\alpha$ (95%) på, at vort interval omslutter δ.

Fejl
Konfidens og sandsynlighed

Det er fristende at opfatte konfidensen som en sandsynlighed, og der er uenighed blandt fagstatistikere om en sådan sprogbrug er velegnet, uheldig eller formastelig. Under alle omstændigheder skal

Figur 4.2.
95% konfidensinterval for δ (skraveret). De vandrette stregmarkeringer viser konfidensintervaller fra forskellige datasæt fra samme proces. (Den lodrette punkterede linie illustrerer placering af den ukendte sande effekt δ = 1,5 med en ukendt, sand standardafvigelse 0,8 på estimaterne $\hat{\delta}$).

man holde sig for øje, når man står med et udregnet, fastlagt konfidensinterval, at den sande værdi for δ også ligger fast. δ falder ikke på tilfældig vis enten udenfor eller indenfor, δ har en fast beliggenhed.

Opfatter vi i figur 4.2 det nederste konfidensinterval som det aktuelt fundne, afstikker det grænserne for plausible værdier for δ i lyset af den aktuelle undersøgelses datasæt. Hvor plausibelt intervallet er udtrykkes kvantitativt ved konfidensgraden.

Konfidens og signifikans

Lykkeligvis er der en nær forbindelse mellem signifikans og konfidens. Det forholder sig således, at signifikanstest kan udføres ved hjælp af konfidensintervaller. Ønsker man at teste på signifikansniveau α, om δ kunne have den hypotetiske værdi δ_0, kan testen udføres ved at se om δ_0 ligger inden for eller uden for et $(1-\alpha)$ - konfidensinterval. Ligger δ_0 indenfor, accepteres hypotesen H_0: $\delta = \delta_0$; ligger δ_0 udenfor forkastes hypotesen H_0, til fordel for H_1: $\delta \neq \delta_0$.

Test ved konfidensinterval

Eksempelvis ses det i figur 4.2, at værdien $\delta_0 = 0$ ligger i konfidensintervallet, så en hypotese om at de to typer teknologi er miljømæssigt identiske, ville altså blive accepteret. Men det ville en hypotese om en gunstig effekt af den nye teknologi på f.eks. $\delta - 1$ eller $\delta = 2$ også blive. $\delta = 3$ ville blive afvist, som for optimistisk, medens skeptikeren, der påstod at den nye teknologi er et tilbageskridt på måske $\delta = -0{,}5$, også kan få sin hypotese accepteret.

Et $(1-\alpha)$–konfidensinterval for δ definerer de hypotetiske værdier for δ, der accepteres ved en signifikanstest på niveau α. Området udenfor dækker δ-værdier som $\hat{\delta}$ er signifikant forskellig fra.

Konklusivitet

I eksemplet figur 4.2 er datamaterialet øjensynligt ikke særligt konklusivt. Dette faktum ville ikke komme eksplicit frem ved f.eks. blot at signifikansteste hypotesen H_0: $\delta = 0$, jf. afsnit 4.1. Estimatet $\hat{\delta}$ med dets omkringliggende konfidensinterval, giver klar besked: Bedste bud på effekten er omkring $\delta = 1{,}2$, men den kan være over dobbelt så stor, og det kan også være, at den slet ikke findes.

Vigtighed og signifikans

Da H_0: $\delta = 0$ ikke kan afvises, er effekten ikke signifikant, men den kan alligevel godt være vigtig. Spørgsmålet er her, om en effekt af størrelsesordnen $\delta = 1{,}2$ har nogen miljømæssig interesse. Hvis den har, kan effekten være vigtig, den er blot muligvis skjult i usikre data. Er $\delta = 1{,}2$ uden reel miljømæssig interesse, og skal δ ligge væsentligt over 3-4 for at den bliver det, så er det eftervist, at ingen vigtig effekt kan skjule sig.

Modsat kan undersøgelser også afsløre stærkt signifikante effekter, der er uden egentlig betydning. Det sker hvis $\delta = 0$ ligger langt fra $\hat{\delta}$ målt i standardafvigelser $\hat{\sigma}_{\hat{\delta}}$, og konfidensintervallet omkring $\hat{\delta}$ kun omslutter værdier af ringe praktisk betydning. Den slags resultater optræder typisk ved datarige undersøgelser, hvor konfidensintervaller placeret tæt ved nul kan være så smalle at de alligevel ekskluderer nul.

Konstruktion af konfidensintervaller

Konfidensintervaller konstrueres for parametre i mange almindeligt benyttede statistiske modeller efter ovenanførte læst. Også i tilfælde hvor de konstrueres på en derfra mere eller mindre afvigende måde, forbliver fortolkningen af intervallerne den samme. Den nøjagtige fremgangsmåde ved konstruktionen behøver brugere næppe

bekymre sig meget om i praksis. Intervallerne udregnes automatisk af mange statistiske computerprogrammer, subsidiært kan deres formler slås op f.eks. i Beyer (1968), Conradsen (1984), eller i en anden statistisk lærebog.

4.3 Bestemmelse af antal målinger

Antalsplanlægning

Før man iværksætter en eksperimental undersøgelse, bør man statistisk overveje hvor mange målinger, der skal foretages for at få belyst den givne problemstilling tilstrækkeligt. Antalsplanlægningen er en vigtig del af forsøgsplanlægningen. Det er spild af resourcer at udføre et unødigt stort antal målinger; et for lille måleprogram er imidlertid også spild af resourcer. Et meget restriktivt budget er ingen undskyldning for at springe antalsplanlægningen over. Er budgettet for småt til at kunne udføre det antal målinger, som skønnes nødvendige for at producere brugbar information; -ja, så er det mere økonomisk klogt slet ikke at måle overhovedet og spare alle pengene.

Ved antalsplanlægningen må der skelnes mellem to situationer:

A: Estimationssituationen, hvor realiteten af den studerede effekt δ ikke betvivles, men bestemmelse af dens størrelse er undersøgelsens formål.

B: Testsituationen, hvor det primære formål med undersøgelsen er at fastslå om den studerede effekt overhoved med statistisk sikkerhed eksisterer.

Antal til estimation

I situation A vil antalsplanlægning dreje sig om at udregne et skøn på det antal målinger, som er nødvendige for at standardafvigelsen $\hat{\sigma}_{\hat{\delta}}$ af effektestimatet $\hat{\delta}$ er tilfredsstillende lille. Faktoren f i udtrykket (4.3) afhænger kun ubetydeligt af antal målinger med mindre dette er meget lille. Størrelsen af konfidensintervallet for effekten δ omkring estimatet $\hat{\delta}$, er derfor praktisk taget proportionalt med stanardafvigelsen..

Antal til test

I situation B må der ikke blot tages hensyn til konfidensintervallets umiddelbare længde for med rimelig sikkerhed at få etableret effektens signifikans. Som det er illustreret i figur 4.2, er selve placeringen af konfidensintervallet omkring den ukendte sande værdi δ også en kilde til usikkerhed på planlægningstidspunktet. På forhånd er det selvfølgelig uvist, om det kommer til at ligge forskudt til højre eller til venstre. I relation til figuren betyder en forskydning til venstre, at der kræves et kortere konfidensinterval for at etablere signifikans, end hvis intervallet lå på sin forventede middelplacering, nemlig centralt omkring δ.

Der må altså planlægges efter en standardafvigelse som ikke blot forventes at etablere signifikans hvis konfidensintervallet ligger centralt, men også gør det i en stor procentdel af andre på forhånd sandsynlige placeringer. Dette behov for sikkerhedsmargin stiller krav om en mindre standardafvigelse svarende til flere observationer.

Skøn før målinger

Beregninger

Eksempler

Disposition

Ordforklaring

Statistisk førstehjælp

Sundhed

5. Statistiske beregninger

En række vigtige problemstillinger kan ved behændighed og faglig indsigt i undersøgelsens teknologiske indhold (kemi, reaktionskinetik, forbrændingsteknologi, toksikologi, etc.), udformes statistisk simpelt. Nedenfor gennemgås seks typer statistisk analyse, med omtale af estimation, signifikanstest og planlægningsaspekter.

Det er ikke sigtet at præsentere en minilærebog i statistisk regneteknik. Formålet er at klargøre idé og fortolkning af disse simple metoder, hvis udregninger kan udføres af så at sige alle de statistiske programpakker/-moduler som udbydes til selv små computere. Afsnit 5.7 generaliserer metoderne, og der gives henvisninger til anbefalet litteratur.

5.1 Middelværdi og varians

Behov for statistisk bestemmelse af et niveau eller en middelværdi optræder hyppigt. Man kan ønske at estimere middelemissionen fra et forbrændingsanlæg over en vis periode, eller middelemissionen fra mange anlæg, eller det ønskes bestemt, hvor stor middelkoncentrationen af en giftig forbindelse er i processpildevand, eller middeltørretiden for en overfladebehandling hvorunder afdunstningen er særlig kraftig, o.s.v.

Model

Den statistiske analyse kan tage udgangspunkt i modellen

$$Y_i = \mu + e_i, \quad i = 1,2,\dots,n \qquad (5.1)$$

hvor Y_i er den i-te observation af målevariablen, μ er middelværdien, og e_i er den statistiske fejl, eller støj, ved den i-te af ialt n målinger. Estimation af parameteren μ er nu helt konkret formålet med de statistiske beregninger.

Det er bekvemt, hvis man om fejlen e_i med god approksimation kan antage, at den er normalfordelt med middelværdi 0 og konstant varians σ^2, samt at e_i-erne er statistisk uafhængige. En middelværdi på nul er nødvendig, da en afvigelse herfra ville være en systematisk fejl, som ville blive fanget af μ. Altså en fejl man ikke kunne afsløre og behandle statistisk, men som ville påvirke estimationen af μ med et fast ukendt bidrag. Er denne antagelse opfyldt, kan gennemsnit af målingerne benyttes som punktestimat for μ.

$$\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i \qquad (5.2)$$

Middelværdi og gennemsnit

Det er ikke helt standardiseret, men anbefalelsesværdig sprogbrug, at sondre mellem middelværdi, som betegnelse for den ukendte sande værdi μ, og gennemsnit som betegnelse for den kendte, databestemte værdi $\bar{Y}$.

Antagelsen om konstant varians for alle e_i-erne, d.v.s. for alle

Der er ikke en af de seks kategorier af undersøgelser, der a priori er bedre end andre. Inden for hver kategori er der mulighed for undersøgelser af større eller mindre valør, fagkundskab og datakvalitet. Hvis imidlertid opsætning og ordvalg i rapporteringen ikke reflekterer datagrundlaget, kan det afføde misforståelser, frustrationer og/eller fejlbeslutning hos læseren, - det være sig en rekvirent, klient eller anden type bruger.

Idé og fortolkning

En teknisk rapport bør i princippet ikke »sælge« læseren et budskab, en beslutning, eller en snæver konklusion, men præsentere undersøgelsens resultater i en neutral og forståelig form. Læseren kan så selv uddrage sit budskab eller træffe sin beslutning, måske under indtryk af argumentation fra anden kilde eller af helt andre hensyn. De seks kategorier skal kort karakteriseres:

Idéoplæg

(a) En undersøgelse kan godt gennemføres uden at bygge på data overhovedet. Den kan være en påpegelse af en potentiel forureningskilde, en idé til løsning af et miljøproblem, en bekymring for en teknologisk udvikling belyst ved parallelle eksempler på forurening eller ulykker inden for beslægtet område, eller andet. En sådan type undersøgelse kan karakteriseres som et idéoplæg, et debatpapir eller lignende. Da det ikke uden data er muligt at drage statistisk valide konklusioner, må sprogbrug, der antyder en statistisk støtte, undgås, og den spekulative karakter af undersøgelsen bør ikke skjules.

Idéudredning

(b) En anden helt acceptabel type undersøgelse, der heller ikke bygger på data, er sammenfatning og systematisering af andres tidligere type (a) undersøgelser; idéudredning kan man kalde den. Tydelig angivelse af referencer er nødvendige, ikke blot af hæderlighedshensyn over for forfatterne, men så læseren får klarhed over forbindelsen mellem rapportens og referencernes konklusioner. Igen må konklusionernes spekulative karakter fremgå af ordvalget, uagtet referencernes antal og notabilitet.

Pilotprojekt

(c) En undersøgelse byggende på et mindre, nyt materiale kan karakteriseres som et pilotprojekt, og data bør gengives fuldt ud i rapporten. Jf. figur 2.3, og de ledsagende bemærkninger, kan det være mest fordelagtigt at gennemføre et mindre projekt før et evt. senere større, som herved kan planlægges bedre. En tilsyneladende nedslående konklusion om manglende tiltro til nytten af yderligere undersøgelser, kan være overordentlig værdifuld. Den tilsyneladende spildte pilotundersøgelse, kan have sparet den langt større undersøgelses fiasko. Der er heller intet galt eller overraskende i, at undersøgelser af denne type ikke kan konkludere entydigt; og en evt. inkonklusivitet må ikke skjules verbalt.

Motiverende udredning

(d) En undersøgelse uden egne nye data, men byggende på andres sparsomme, spredte data, kan enten have karakter af en sammenfatning og bearbejdelse af tidligere konklusioner, eller der kan være tale om en ny analyse af andres data, evt. fra flere kilder. Igen må rapporten redegøre for, om der er tale om det ene eller det andet, eller, hvis begge dele udføres, klart separere de to typer

målingerne, Y_i i = 1,2,···,n, betyder, at hver målings usikkerhed omkring μ på forhånd er lige stor. Den forudsætning er nødvendig, ikke blot for at estimatet (5.2) udnytter datainformationen effektivt, men mere afgørende for at usikkerheden på $\hat{\mu}$ kan estimeres statistisk.

Eftersom målingerne Y_i er summen af et fast bidrag μ og en stokastisk afvigelse e_i med varians σ^2, bliver Y_i's varians Var(Y) også lig σ^2. Og standardafvigelsen på enkeltmålingen bliver definitionsmæssigt

$$\text{std}(Y) = \sigma_Y = \sigma \, . \tag{5.3}$$

Hvis videre målingerne er statistisk uafhængige, d.v.s. at støjbidragene e_i er det, bliver variansen på gennemsnittet

$$\text{Var}(\bar{Y}) = \sigma^2_{\bar{Y}} = \frac{1}{n^2} \sum_{i=1}^{n} \text{Var}(Y_i) = \frac{1}{n} \sigma^2 \quad , \tag{5.4}$$

og standardafvigelsen på estimatoren $\hat{\mu}$, også kaldet $\hat{\mu}$'s standardfejl, bliver

$$\text{std}(\hat{\mu}) = \sigma_{\hat{\mu}} = \text{std}\ \bar{Y} = \sigma_{\bar{Y}} = \sigma/\sqrt{n} \quad . \tag{5.5}$$

(Egentlig burde indeks μ til σ være $\hat{\mu}$; hvilket vi dog af hensyn til enkelhed i notationen ikke skal benytte.)

Uafhængighed

Udtrykket (5.5) kan give en helt forkert standardafvigelse for $\hat{\mu}$, hvis der ikke er statistisk uafhængighed mellem målingerne. Manglende uafhængighed kan optræde, hvis målingerne tages meget tæt op ad hinanden i tid eller sted, og hvis passende randomisering er undladt.

Ønskes f.eks. middelemissionen for et forbrændingsanlæg over et døgn, må målingerne ikke foretages inden for et kort tidsrum på f.eks. 5 minutter, hvor nabomålinger kan minde utypisk meget om hinanden og derfor lede til underestimation af den virkelige varians. (Se også afsnit 2.3 om repræsentativitet). Et bedre arrangement består i at foretage målingerne jævnt over døgnet med ens, i forvejen fastsat tidsinterval; eller endnu bedre randomiseret, det vil sige på i forvejen fastsatte, tilfældige tidspunkter. Ønskes repræsentativitet over flere døgn, må flere døgn inddrages, og det bedste billede af den sande variation fås, hvis målingerne fordeles over flest mulige døgn.

Statistisk fejl

Normalt kendes variansen σ^2 på den statistiske fejl ikke på forhånd, men må også estimeres. Det er vigtigt at slå fast, at den statistiske fejl, e_i, ikke er synonym med målefejlen, f.eks. i måleapparaturet eller i den kemiske analyse af den udtagne prøve. e_i indbefatter alle variationskilder, der bidrager til variabiliteten mellem repræsentative målinger, herunder – i det aktuelle eksempel – døgnvariationer og forskelle mellem dage, inhomogenitet i røgen p.g.a. ikke-perfekt turbulens, variation som skyldes mere eller mindre tilfældige ændringer i forbrændingstemperatur eller iltoverskud, etc.

Den statistiske fejl e_i må altså erkendes som en sum af et antal varianskomponenter:

$$e_i = e_i(a) + e_i(b) + e_i(c) + e_i(d) + \cdots \tag{5.6}$$

hvor uafhængighed af bidragene betyder at

$$\sigma^2 = \text{Var } e_i = \text{Var } e_i(a) + \text{Var } e_i(b) + \cdots \tag{5.7}$$

Disse varianskomponenter kan anskues som faldende i to grupper, den procesbetingede, såkaldt naturlige variabilitet og målefejlen:

$$e_i = e_i(n) + e_i(m) \quad , \tag{5.8}$$

hvor tilsvarende til (5.7) det gælder, at

$$\sigma^2 = \text{Var } e_i = \text{Var } e_i(n) + \text{Var } e_i(m) \, . \tag{5.9}$$

Naturlig variabilitet

Den naturlige variabilitet $e_i(n)$ er udtryk for, at den betragtede målevariabel rent faktisk varierer under de omstændigheder, som man i undersøgelsen ønsker omfattet af hensyn til repræsentativitet.

Målevariabilitet

Målevariabiliteten $e_i(m)$ er et udtryk for at selve måleprocessen er usikker, at man altså ved gentagne observationer under helt ens omstændigheder ville få varierende måleværdier.

Undertiden kendes på forhånd måleusikkerheden for målingerne i en undersøgelse. Den kan være kendt fra tidligere, tilsvarende kemiske laboratorieanalyser af koncentrationer f.eks., eller måske erfaring med et givet måleinstrument. (Prøvetagningsfejlen, som også er en potentiel kilde til måleusikkerhed, må ikke glemmes i denne forbindelse). Det er derfor ikke helt ualmindeligt at se den påståede kendte måleusikkerhed anført, og dens varians anvendt som variansen på den totale statistiske fejl, e_i. Dette er klart ukorrekt, og kan lede til meget grove fejlkonklusioner i den statistiske analyse, – specielt når den naturlige variabilitet er den dominerende komponent i (5.9), hvad man hyppigt kommer ud for i processer uden for laboratoriet.

Skøn over varians

Den statistiske analyse må baseres på et realistisk skøn over $\text{Var}(e_i) = \sigma^2$, og dette opnås sikrest fra måledata selv. Det er generelt ønskeligt af hensyn til undersøgelsens præcision, at σ^2 er lille, men for så vidt angår den naturlige komponents bidrag, kan man ikke nedbringe den. Man kan underestimere den, og derved snyde sig selv og/eller andre, men den naturlige variabilitet er et produkt af de repræsentative omstændigheder, som man ønsker analysen skal omfatte.

Variansreduktion

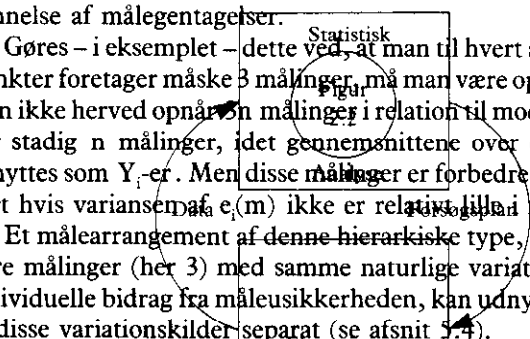
I det tænkte eksempel kan man selvfølgelig nedbringe e_i ved at eliminere en eventuel dag-til-dag varianskomponent. Dette gøres ganske enkelt ved at udføre alle målingerne på én dag, så den tilsvarende komponent i (5.6) nu ikke varierer. Den er derimod blevet til et fast bidrag, nemlig denne dags særlige afvigelse fra middelværdien over alle dage, og dette bidrag kommer nu til at indgå i μ. Estimationen af μ bliver herefter behæftet med en ukendt

Systematisk afvigelse

Iterativ modelbygning

Reduktion af målefejl

Figur 2.3
Iterativt undersøgelsesforløb vekslende mellem analyse og forsøg.



Reel variabilitet

Datagrundlag

Figur 2.4.
Typer af undersøgelser

datamængde	egne, nye data	andres kendte data
ingen	(a)	(b)
pilot	(c)	(d)
egentligt projekt	(e)	(f)

Standardafvigelser

Normalfordeling

Konfidensinterval

Som nævnt i afsnit 4.2, afhænger faktoren f af intervallets ønskede konfidensgrad $(1-\alpha)$, og kun forholdsvis lidt af observationsantallet, som analysen bygger på. Mere specifikt går denne afhængighed på, hvor mange observationer estimatet $\hat{\sigma}_\mu$ bygger på. Men derudover afhænger f også af, hvilken statistisk fordeling $\hat{\mu}$ har. Er e_i normalfordelt bliver $\hat{\mu}$ det også, og i så fald kan f slås op i en tabel over t-fordelingen med n-1 frihedsgrader. Uden at gå nøjere ind på hverken t-fordelingen eller begrebet frihedsgrad, skal det anføres at for given konfidensgrad $(1-\alpha)$, skal man som f-faktor vælge $(1-\alpha/2)$-fraktilen i en t-fordeling med n-1 frihedsgrader. For et 95% konfidensinterval, bliver f = 2,26 ; 2,09 ; 2,02 og 1,96 for henholdsvis n = 10, 20, 40, og ∞ . n = ∞ svarer til at σ^2 er eksakt kendt på forhånd. Disse forskelle i f-faktorer er ikke af stor praktisk betydning. Imidlertid gælder det videre, at selv hvis e_i-erne afviger en del fra normalfordelingen, vil $\hat{\mu}$ alligevel tendere mod en normalfordeling. Benyttes således alligevel de f-faktorer, som egentlig bygger på en normalfordelingsantagelse, produceres stadig konfidensintervaller af approksimativt samme konfidensgrad.

f-faktor

Dette er ikke ment som et modargument for transformation af data, hvis en grafisk analyse (afsnit 3.1) skulle pege på fordele herved. Mange miljøvariable kan fordelagtigt logaritmetransformeres før en videre analyse. Men konfidensintervaller er robuste med hensyn til afvigelse fra normalfordelingsantagelsen modsat antagelserne om konstant varians og statistisk uafhængighed.

Robusthed

Beregningsarbejdet ledende til (5.12) er enkelt nok til håndregning, men de fleste statistiske computerprogrammer klarer direkte det hele, inklusive tabelopslag af f for specificeret α. Fortolkningen af intervallet hjælper et sådant program imidlertid ikke med (se afsnit 4.2); ej heller fortolkningen af, hvad det er for en middelværdi μ, man har estimeret, og hvilke variationskilder, der er blevet afspejlet i usikkerhedsangivelsen. Dette sidste emne er til gengæld omtalt detaljeret her.

Test

Ønsker man at teste om μ i lyset af data kan have en given hypotetisk værdi, kan dette afgøres ved inspektion af konfidensintervallet (5.12) som forklaret i afsnit (4.2). Specielt udføres en test på signifikationsniveau α om hvorvidt μ's sande værdi kan være nul, ved at se om værdien nul ligger i eller uden for det konstruerede $100(1-\alpha)\%$ konfidensinterval. Indenfor giver accept, udenfor giver forkastelse. Mange computerprogrammer udfører test for $\mu = 0$ med specifik angivelse af p-værdi (afsnit 4.1).

Eksempel 5.1(a)

Følgende observationer af en målevariabel tænkes foretaget på 27 parallelle produktionsenheder: 64, 85, 82, 104, 72, 72, 64, 84, 80, 72, 72, 68, 75, 68, 78, 64, 80, 66, 78, 72, 65, 80, 72, 72, 74, 78, 64. Middelværdien ønskes estimeret og fortolket.

Man udregner gennemsnit og variansestimat, jf. (5.2) og (5.10): $\hat{\mu} = 74{,}259$ *og* $\hat{\sigma}^2 = 76{,}353 = 8{,}74^2$.

Middelværdien estimeres til omkring 74,259 og enkeltmålingerne har en standardafvigelse på cirka 8,74 omkring middelværdien. Da der kun er

foretaget én måling pr. produktionsenhed vides ikke om denne variation målingerne imellem skyldes statistisk måleusikkerhed eller naturlig variabilitet mellem produktionsenhederne, eller en blanding af begge dele.

Usikkerheden på middelværdiestimatet $\hat{\mu}$, *fås jf. (5.11) til* $\hat{\sigma}_{\mu} = 1{,}68$. *Et 95% konfidensinterval for middelværdien* μ, *fås af (5.12) med* $f = 2{,}048$ *(fra en t-tabel med* $\alpha = 5\%$ *og antal frihedsgrader lig 26). Man finder intervallet [70,815; 77,703], d.v.s. at middelværdien* μ *med konfidensgrad 95% ligger i dette interval.*

Enhver fremsat hypotese om, at μ*'s sande værdi er et givet tal fra dette interval ville blive accepteret på signifikansniveau* $\alpha = 5\%$*; f.eks. ville såvel* $\mu = 71$ *som* $\mu = 77$ *blive accepteret. Derimod ville* $\mu = 0$ *blive forkastet ikke blot på niveau* $\alpha = 5\%$ *men også på lavere niveauer* α, *blot større end p-værdien som i dette tilfælde er* $p \ll 0{,}0001\%$.

Antalsplanlægning: Estimation

Antalsplanlægning med henblik på estimation (situation A i afsnit 4.3) foretages således:

Man må beslutte sig til, hvor nøjagtig man ønsker μ estimeret ved specifikation af en ønsket værdi for estimatets standardafvigelse, σ_{μ} $(= \sigma_{\bar{Y}})$ Dernæst må man gætte, hvilken statistisk usikkerhed målingerne Y_i er behæftede med, det kan f.eks. ske ved at vurdere de enkelte relevante led i (5.7).

Fra den vurdering findes nu overslagsmæssigt

$$n = [\sigma/\sigma_{\mu}]^2 \tag{5.13}$$

Hvis f.eks. det gættes at $\sigma = 0{,}2$ og ønskes at $\sigma_{\mu} = 0{,}1$, må n være mindst lig $(0{,}2/0{,}1)^2 = 4$. Et så lille antal målinger vil ikke give noget sikkert estimat af σ v.h.a. (5.11). Af den grund bør man overveje at vælge n større end det skønnede minimalantal på 4, måske 6.

Antalsplanlægning: Test

Antalsplanlægning med henblik på signifikanstest (situation B i afsnit 4.3) foretages således:

Som overfor gættes σ . Dernæst specificere hvilken hypotese man, efter målingerne er indsamlet, ønsker testet: $H_0 : \mu = \mu_0$. Endelig besluttes hvilken afvigelse fra $\mu = \mu_0$ man med stor sikkerhed $(1-\beta)$ ønsker at kunne finde signifikant på niveau α. Denne afvigelse af interesse kaldes δ_0, således at målingerne fra en proces med et sandt niveau eller middelværdi på $\mu = \mu_0 + \delta_0$ skal planlægges i antal med henblik på at afvise at $\mu = \mu_0$.

»Værste fald«

Vi forestiller os af hensyn til forklaringen nedenfor, at δ_0 er positiv; en negativ δ_0 giver samme resultat. Er Y_i -ernes sande middelværdi $\mu_0 + \delta_0$, vil denne middelværdi blive estimeret med standardafvigelsen $\sigma_{\mu} = \sigma/\sqrt{n}$. Det betyder at med sandsynligheden $1-\beta$ vil vi fra data finde en $\hat{\mu}$ så

$$(\mu_0 + \delta_0) - z_{1-\beta}\sigma_{\mu} < \hat{\mu}, \tag{5.14}$$

hvor $z_{1-\beta}$ er $1-\beta$ fraktilen i en normalfordeling som findes i tabel. Tallet på venstre side af (5.14) er altså med sandsynligheden $1-\beta$ en »i værste fald«- værdi.

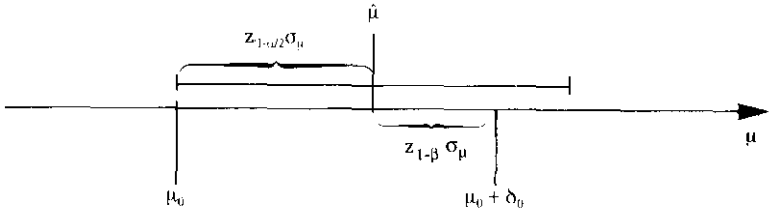
Findes faktisk $\hat{\mu}$ lig denne »værste fald«-værdi, vil konfidensintervallet for μ, ligning (5.12), ligge symmetrisk omkring denne værdi. Intervallets nedre grænse ligger afstanden $z_{1-\alpha/2}\ \sigma_{\mu}$ til venstre for $\hat{\mu}$, idet f i (5.12) er sat lig $1-\alpha/2$ fraktilen af en normalfordeling, svarende til at σ_{μ} betragtes kendt. For at give forkastelse af H_0: $\mu = \mu_0$, skal konfidensintervallets nedre grænse ligge til højre for μ_0, d.v.s. at

$$\mu_0 < \hat{\mu} - f\sigma_{\mu} \tag{5.15}$$

$$<=> \mu_0 < (\mu_0 + \delta_0 - z_{1-\beta}\sigma_{\mu}) - z_{1-\alpha/2}\sigma_{\mu} \quad . \tag{5.16}$$

Figur 5.1 illustrerer forholdene, der leder til uligheden (5.16). Af (5.16) findes at

$$(z_{1-\beta} + z_{1-\alpha/2})\sigma_{\delta} < \delta_0 \quad . \tag{5.17}$$



Figur 5.1.
Principtegning for antalsplanlægning, således at H_0: $\mu = \mu_0$ med sandsynlighed $1-\beta$ vil blive forkastet på niveau α hvis $\mu_0 + \delta_0$. $\hat{\mu}$ markerer et med sandsynlighed $1-\beta$ værste fald estimat; omkring $\hat{\mu}$ er indtegnet et $100(1-\alpha)\%$ konfidensinterval, som netop ekskluderer μ_0.

Da $\sigma_{\delta}^2 = \sigma/n$, findes heraf det nødvendige antal målinger

$$n > (z_{1-\beta} + z_{1-\alpha/2})^2\ [\sigma/\delta_0]^2. \tag{5.18}$$

Gættes for eksempel igen at $\sigma = 0{,}2$ og besluttes det sagkyndigt at $\delta_0 = 0{,}3$ med $\alpha = 5\%$ og $\beta = 20\%$, betyder dette, at man med 80% sandsynlighed vil vise en forskel på $\delta_0 = 0{,}3$ signifikant på 5% niveauet med antallet af målinger

$$n > (z_{0{,}80} + z_{0{,}975})^2 [\sigma/\delta_0]^2 = (0{,}842 + 1{,}96)^2 (0{,}2/0{,}3)^2 = 3{,}48. \tag{5.19}$$

Et så lavt antal målinger som 4 ville dog føre til at f-faktoren i (5.12) øgedes mærkbart. Under indtryk heraf kunne man øge n til f.eks. 6.

Eksempel 5.1(b)
Antalsplanlægning foretages normalt før data er opsamlet, men kan også blive aktuel efter at målinger er foretaget, i så fald med henblik på yderligere målinger. Ønskes μ i dette eksempel estimeret med en standardafvigelse på f.eks. $\sigma_{\mu} = 0{,}8$ findes jf. (5.13) $n = (8{,}74/0{,}8)^2 = 119{,}3$, dvs. at yderligere knap 100 observationer er nødvendige. Dette kan kun lade sig gøre, hvis der findes dette yderligere antal produktionsenheder. Måles derimod gentagne gange på de samme produktionsenheder, bliver nærværende analyse ikke korrekt. Herom mere i eksempel 5.4.

5.2 Forskel mellem to middelværdier

Model

Interessen for statistisk at bestemme forskellen mellem to middelværdier opstår i et væld af situationer, hvor effekten af en påtænkt procesjustering, en ændring i teknologi eller et andet indgreb skal undersøges. Den statistiske analyse kan tage udgangspunkt i modellen

$$\begin{cases} Y_{0,i} = \mu_0 + e_{0,i} & i = 1,2,\cdots,n_0 \quad , \\ Y_{1,i} = \mu_1 + e_{1,i} & i = 1,2,\cdots,n_1 \quad , \end{cases} \tag{5.20}$$

hvor $Y_{0,i}$-erne er målinger foretaget under standardbetingelser, og $Y_{1,i}$-erne de alternative målinger. Da parametrene μ_0 og μ_1 måler niveauerne eller middelværdierne under de to betingelser, bliver det deres differens

$$\delta = \mu_1 - \mu_0 \quad , \tag{5.21}$$

som den statistiske analyse skal estimere inklusive kvantitativ usikkerhedsangivelse.

For hver af de to betingelser svarer modellen (5.20) helt til modellen (5.1). Derfor er de i afsnit 5.1 gjorte bemærkninger fuldt gyldige også i nærværende analyse, idet visse tilføjelser må gøres. Da det er effekten af et specificeret indgreb, der skal estimeres, er det afgørende ikke at få denne effekt sammenblandet med andre mulige ændringer. Det betyder ikke, at den betragtede proces nødvendigvis skal holdes kunstigt stabil hvad angår andre naturligt varierende påvirkninger; det ville kunne gå ud over repræsentativiteten, jf. diskussionen i afsnit 2.3. Derimod må det sikres, at disse variationskilder ikke ensidigt påvirker den ene betingelse, hvorved en systematisk fejl vil opstå i estimationen af δ. Således er det en dårlig idé først at foretage alle n_0 målinger under standardbetingelserne, og dernæst alle de n_1 alternative målinger. En opadgående tidstrend f.eks. ville her føre til overestimation af δ. Dette er en hovedårsag til, at en stor mængde allerede eksisterende data angående standardbetingelserne, er meget lidt værd i en statistisk analyse som nærværende. Det er i reglen bedst og altid sikrest, at se bort fra sådanne historiske data og indsamle nye, hvor rækkefølgen af de to betingelser kan randomiseres. Lader randomisering sig ikke udføre (evt. med passende restriktioner, f.eks. blokvis) er de nedenfor anførte beregninger uegnede, men en mere kompliceret såkaldt interventionsanalyse kan undertiden anvendes, se Pallesen *et al.* (1985) og Booman *et al.* (1987).

Naturlig variation

Trend

Interventionsanalyse

δ estimeres ved

$$\hat{\delta} = \hat{\mu}_1 - \hat{\mu}_0 \, , \tag{5.22}$$

hvor $\hat{\mu}_0$ og $\hat{\mu}_1$ ligesom i (5.2) er gennemsnit, henholdsvis $\bar{Y}_0$ og $\bar{Y}_1$, således at, jf. (5.4) og (5.5),

$$\begin{cases} \text{Var } \hat{\mu}_0 = \sigma^2/n_0 \quad , \\ \text{Var } \hat{\mu}_1 = \sigma^2/n_1 \quad , \end{cases} \tag{5.23}$$

Samme varians

hvor den antagelse er gjort, at $Y_{0,i}$ og $Y_{1,i}$ d.v.s. $e_{0,i}$ og $e_{1,i}$, har samme varians σ^2. For at sikre at denne antagelse er opfyldt, må dels den naturlige variabilitet dels måleusikkerheden være den samme i de to tilfælde. Samme naturlige variabilitet sikres ved randomiseringen, med mindre selve indgrebet gør, at effekten af de andre påvirkninger ændres. I så fald har indgrebet foruden en eventuel virkning δ på niveauet også en indflydelse på σ^2. Benyttes samme måleteknik under de to betingelser, sikres ens måleusikkerhed, med mindre måleusikkerheden er væsentligt forskellig på de to niveauer. En betydningsfuld forskel mellem varianserne på $Y_{0,i}$-erne og $Y_{1,i}$ kan hyppigt afsløres grafisk som omtalt i afsnit 3.2, og variansinhomogenitet kan i givet fald tit afhjælpes ved en passende transformation, se evt. Box *et al.* (1978). Visse computerprogrammer udfører en signifikanstest (F-test) for hypotesen H_0: $\text{Var}(e_{0,i}) = \text{Var}(e_{1,i})$; en sådan test fortolkes helt efter retningslinierne i afsnit 4.1.

Fra (5.22) og (5.23) fås, at

$$\text{Var}(\hat{\delta}) = \text{Var}(\hat{\mu}_0) + \text{Var}(\hat{\mu}_1) = \sigma^2 \left[\frac{1}{n_0} + \frac{1}{n_1}\right] \quad . \tag{5.24}$$

$\hat{\delta}$'s standardafvigelse kan nu findes heraf, idet σ^2 må estimeres fra observationerne. I slægtskab med (5.10) benyttes sædvanligvis estimatet

$$\hat{\sigma}^2 = \frac{1}{n_0 + n_1 - 2}\left[\sum_{i=1}^{n_0} (Y_{0,i}-\bar{Y}_0)^2 + \sum_{i=1}^{n_1} (Y_{1,i}-\bar{Y}_1)^2\right] \quad , \tag{5.25}$$

så

$$\text{std}(\hat{\delta}) = \hat{\sigma}_\delta = \hat{\sigma}\left[\frac{1}{n_0} + \frac{1}{n_1}\right]^{\frac{1}{2}} \quad . \tag{5.26}$$

Konfidensinterval

Parallelt til (5.12) bliver konfidensintervallet for σ nu

$$[\hat{\delta} - f\hat{\sigma}_\delta \; ; \; \hat{\delta} + f\hat{\sigma}_\delta] \quad , \tag{5.27}$$

hvor faktoren f er defineret som ved (5.12). Også signifikanstest for særlige, hypotetiske værdier for δ (specielt H_0: $\delta = 0$) udføres og fortolkes som omtalt for μ i afsnit 5.1

Eksempel 5.2(a)

Ud over de i eksempel 5.1 foretagne observationer forestiller man sig målinger foretaget på 27 andre produktionsenheder, som er underkastet en modifikation, hvis formål skulle være at nedbringe den målte variabels størrelse. De målte værdier er: 64, 84, 80, 84, 80, 80, 72, 80, 70, 66, 76, 68, 72, 66, 64, 74, 64, 66, 88, 66, 57, 81, 73, 58, 75, 70, 70.

Den nye middelværdi og varians estimeres til $\hat{\mu} = 72{,}148$ *og* $\hat{\sigma}_1^2 = 65{,}202 = 8{,}07^2$.

I forhold til før (nu givet index »0«) er målt et gennemsnitligt fald fra 74,259 til 72,148, et estimeret fald på $\hat{\delta} = \hat{\mu}_1 - \hat{\mu}_0 = -2{,}111$.

Spørgsmålet er, om denne værdi er signifikant eller kan skyldes tilfældigheder. De fleste computerprogrammer vil foretage en F-test til belysning af, om σ_0^2 *og* σ_1^2 *kan antages ens. Teststørrelsen vil være 1,17 (eller 1/1,17 afhængigt af programmeringen) som ville blive refereret til en F(26,26)-fordeling. En p-værdi på* $p = 0{,}692$ *ville i dette tilfælde blive udregnet. Det betyder at hypotesen* H_0: $\sigma_0 = \sigma_1$ *accepteres ved test på ethvert signifikansniveau* $\alpha < p = 0{,}629$; *specielt accepteres på niveau* $\alpha = 0{,}05 = 5\%$.

Antages varianserne ens udregnes et fælles variansestimat jf. (5.25): $\hat{\sigma}^2 = 70{,}78 = 8{,}41^2$.

$\hat{\delta}$*'s standardafvigelse findes jf. (5.26) til std(*$\hat{\delta}$*)* $= 2{,}290$. *Til et 95% konfidensinterval, jf. (5.27), benyttes* $f = 2{,}007$ *således at intervallet bliver: [-6,707 ; 2,485] . Dette interval indeholder* $\delta = 0$, *som altså ikke kan forkastes i lyset af data; en uønsket opadgående effekt af modifikationen, f.eks.* $\delta = 2{,}0$, *kan heller ikke afvises på niveau* $\alpha = 5\%$, *selvom den estimerede effekt var negativ. På den anden side kan effekten altså også være betydeligt mere gunstig end* $\hat{\delta} = -2{,}111$, *måske* $\delta = -6{,}0$. *Mange computerprogrammer afgør spørgsmålet vedrørende mulig accept af hypotesen* H_0: $\delta_0 = 0$, *ved en t-test, der bliver udregnet som* $T = \hat{\delta}/std(\hat{\delta}) = -0{,}922$. *Denne t-værdi refereret til en t-fordeling med 52 frihedsgrader giver en p-værdi på 0,361. Dette betyder at* H_0: $\delta_0 = 0$ *kan accepteres ved signifikanstest på alle niveauer* $\alpha < p = 0{,}361$, *specielt* $\alpha = 5\%$, *men altså også* $\alpha = 10\%$, *eller 20%.*

Denne analyse er valid uanset om de sidste 27 målinger er fra samme procesenheder som de første. En god forsøgsplan ville bestå af randomiseret at udvælge et antal anlæg og så – igen randomiseret – underkaste halvdelen den modifikation, hvis effekt ønskes belyst. Herefter måles alle enheder samtidigt eller i randomiseret orden. (Vedrørende parret analyse se nedenfor).

Antalsplanlægning: Estimation

Antalsplanlægning med henblik på estimation (situation A i afsnit 4.3) foretages som før, således at $\hat{\delta}$'s standardafvigelse opnår en ønsket størrelse, σ_{δ}. Ud fra et kvalificeret gæt på σ's værdi findes fra (5.24) at n_0 og n_1 skal vælges så

$$\left[\frac{1}{n_0} + \frac{1}{n_1}\right] \leq \left[\sigma_{\delta}/\sigma\right]^2 \quad . \tag{5.28}$$

Denne betingelse kan opfyldes for mange kombinationer af n_0 og n_1. Er omkostningerne ens for de to slags målinger, vælges $n_0 = n_1$, og (5.28) fører da til

$$n_0 = n_1 > 2(\sigma/\sigma_{\delta})^2. \tag{5.29}$$

Hvis for eksempel δ ønskes bestemt med en standardafvigelse σ_{δ} på 0,1 og den statistiske fejls standardafvigelse σ er 0,2, da skal antallet af målinger være mindst $n_0 = n_1 = 2(0{,}2/0{,}1)^2 = 8$; altså ialt

16 målinger. Er $Y_{1,i}$-erne imidlertid faktoren c gang dyrere end $Y_{0,i}$-erne pr. måling (hvor c principielt kan være mindre end 1), da opnås samme nøjagtighed billigere ved at vælge

$$\begin{cases} n_0 = (1 + c^{\frac{1}{2}})\,(\sigma/\sigma_\delta)^2 \quad , \\ n_1 = n_0/c^{\frac{1}{2}} \quad . \end{cases} \tag{5.30}$$

Hvis for eksempel $Y_{1,i}$-erne koster 4 gang så meget som $Y_{0,i}$-erne, og man ønsker δ bestemt med $\sigma_\delta = 0{,}1$ medens $\sigma = 0{,}2$, da skal måleprogrammet planlægges med mindst følgende antal målinger: $n_0 = (1 + 4^{\frac{1}{2}})\,(0{,}2/0{,}1) = 12$ og $n_1 = 12/4^{\frac{1}{2}} = 6$. Det krævede antal målinger er nu 18 mod 16 ovenfor, men med den nye omkostningsfordeling bliver programmet alligevel billigere at gennemføre.

Antalsplanlægning: Test

Antalsplanlægning med henblik på signifikanstest (situation B i afsnit 4.3) foretages som for m i afsnit 5.1. Man finder parallelt til (5.17), at kravet til n_0 og n_1 skal findes fra betingelsen

$$(z_{1-\beta} + z_{1-\alpha/2})\sigma_\delta < \delta_0, \tag{5.31}$$

hvor δ_0 specificeres som den (mindste) effekt af reel interesse, der - hvis den er sand - med sandsynligheden $1-\beta$ skal føre til forkastelse af hypotesen $H_0: \delta = 0$ på signifikantsniveau α.

Indsættes (5.26) i (5.31) fås nu kravet

$$\left[\frac{1}{n_0} + \frac{1}{n_1}\right] \leq (\delta_0/\sigma)^2 \,/\, (z_{1-\beta} + z_{1-\alpha/2})^2 \quad , \tag{5.32}$$

Med en $Y_{1,i}$ omkostning pr. måling, der er c gange så stor som $Y_{0,i}$'s, leder (5.32) til

$$\begin{cases} n_0 = (1 + c^{\frac{1}{2}})\,(z_{1-\beta} + z_{1-\alpha/2})^2\,(\sigma/\delta_0)^2 \quad , \\ n_1 = n_0/c^{\frac{1}{2}} \quad . \end{cases} \tag{5.33}$$

Hvis f.eks. $c = 1$ og $\sigma = 0{,}2$, medens det forlanges at en reel effekt på $\delta_0 = 0{,}3$ skal findes signifikant på $\alpha = 5\%$ niveauet med sandsynlighed $1-\beta = 80\%$, da findes $n_0 = (1+1^{\frac{1}{2}})(0{,}842 + 1{,}960)^2\,(0{,}2/0{,}3)^2 = 2 \cdot 3.49 \simeq 7$, og $n_1 = 7/1^{\frac{1}{2}} = 7$. Er derimod $c = 4$ findes $n_0 = 10{,}47$ og $n_1 = 5{,}23$, som afrundes passende

Eksempel 5.2 (b)

Valget af 2 gange 27 målinger ovenfor i eksempel 5.2 kan være fremkommet fra (5.30) med c = 1. Havde man på forhånd skønnet σ = ca. 10 og ønsket $\sigma_\delta < 3$ ville man finde $n_1 = n_0 = (1+1)(10/3)^2 = 22{,}22$. Da forhåndsskønnet på σ er usikkert, kan man have ønsket at indbygge en vis sikkerhedsmargin. Dette gøres ved at øge n_0 (og n_1) passende. F.eks. kunne man have vurderet, at σ næppe ville være større end ca. 11. Gentages beregningen ved brug af denne forsigtige værdi, findes $n_1 = n_0 = (1+1)(11/3)^2 = 26{,}89 \simeq 27$.

Parret analyse

I visse situationer kan den naturlige variabilitet være af en sådan størrelse og karakter, at en mere præcis estimation af et indgrebs effekt opnås ved en såkaldt parret statistisk analyse.

Måleprogrammet må være planlagt parret, modsat den ovenfor beskrevne uparrede metode. Alle sædvanlige statistiske computerprogrampakker, kan udføre begge typer analyser.

Stor naturlig variabilitet

Fordelen ved den parrede analyse opstår i situationer, hvor den naturlige variabilitet er stor i forhold til den søgte eller forventede effekt, som til trods herfor alligevel er af praktisk interesse. For eksempel kunne døgnvariation i emissionen fra et anlæg være så stor i forhold til effekten af en overvejet ændring i produktionen, at denne effekt vanskeligt lod sig estimere med rimelig præcision p.g.a. den store naturlige variabilitet. Den store variabilitet kunne også skyldes fluktuationer i råvarekvaliteten. Eller, som et andet eksempel, afdunstningen fra en lakeret overflade aftager måske så meget i løbet af tørreperioden at en opnået generel sænkning af emissionsmængden fra overfladen tilsvarende er relativt lille i forhold til den naturlige variabilitet (her et fald) over tiden.

I sådanne situationer kan man med fordel parre målingerne $(Y_{0,i}, Y_{1,i})$ $i = 1,2\cdots, n$, således at hvert par bliver målt under så homogene betingelser som muligt. For eksempel på samme tidspunkt af døgnet, med brug af ens råvarekvalitet, eller på samme tidspunkt i tørreforløbet. If. diskussionen om repræsentativitet i afsnit 2.3, skal ingen bestræbelser gøres for at udligne forskelle parrene imellem.

Differenserne af de parrede målinger dannes nu,

$$Z_i = Y_{1,i} - Y_{0,1} \qquad i = 1,2,\cdots,n \quad . \tag{5.34}$$

Det ses, at den statistiske model for Z_i -erne bliver

$$Z_i = \delta + f_i \qquad i = 1,2,\cdots,n \quad , \tag{5.35}$$

Differenser analyseres

hvor δ er den søgte effekt af indgrebet, og f_i bliver den gældende statistiske fejl, hvor en betydelig del af e_i-ernes (Y_i-ernes) naturlige variabilitet er udbalanceret, vel at mærke uden at repræsentativiteten er sat over styr. Z_i-erne opfattes nu som datamaterialet og analyseres som beskrevet i afsnit 5.1. Den parrede analyse forudsætter ikke at $Y_{0,i}$ -erne og $Y_{1,i}$ -erne har samme varians.

Eksempel 5.2 (c)

Stammer tallene i eksempel 5.1 og 5.2 ovenfor fra en parret forsøgsplan, har man altså først ved randomisering udvalgt 27 procesenheder, og målt på disse modificeret og umodificeret i randomiseret rækkefølge. Matcher rækkefølgen af de to sæt af 27 målinger findes differenserne: 0, -1, -2, -20, 8, 8, 8, -4, -10, -6, 4, 0, -3, -2, -14, 10, 16, 0, 10, -6, -8, 1, 1, -14, 1, -8, 6.

Af disse differenser Z_i findes $\bar{Z} = -2,111 = \hat{\delta}$ (som før), men $\hat{\sigma}^2 = 65,718 = 8,11^2$, som er variansen af differenserne. Større niveauforskelle procesenhederne imellem påvirker ikke dette variansestimat, derimod reelle forskelle i effekten af modifikationen fra enhed til enhed, plus naturligvis måleusikkerhed m.v. Man finder, jf. (5.11), $\hat{\sigma}_{\delta} = 8,11/\sqrt{27} = 1,560$, og et 95 % konfidensinterval for δ [-5,319 ; 1,097] jf. (5.12) med

$f = 2{,}056$. Dette interval er snævrere end før, men omslutter stadig $\delta = 0$, som således kan accepteres på niveau $\alpha = 5\%$. De fleste computerprogrammer udregner med henblik på test af netop denne hypotetiske værdi teststørrelsen $T = \hat{\delta}/std(\hat{\delta}) = -1{,}353$, som refereret til en t-fordeling med 26 frihedsgrader giver en p-værdi på 0,188. D.v.s. at H_0:$\delta = 0$ kan accepteres på alle signifikansnivauer $\alpha < p = 0{,}188$, specielt $\alpha = 5\%$.

5.3 Forskel mellem én middelværdi og flere alternativer

I mange undersøgelser opnås bedst udnyttelse af data, ved at studere effekterne af flere alternative indgreb over for en fælles reference. Det kan være en række alternative optimeringsforslag på en arbejdende proces, en række alternative nye teknologier over for den nuværende teknologi, etc. Den statistiske analyse kan tage udgangspunkt i modellen

Model

$$\begin{cases} Y_{0,i} = \mu_0 + e_{0,i} & i = 1,2,\cdots,n_0 \quad , \\ Y_{k,i} = \mu_k + e_{k,i} & i = 1,2,\cdots,n_k \quad , \end{cases} \tag{5.36}$$

$$\text{for } k = 1,2,\cdots,p \ ,$$

som blot er en udvidelse af modellen (5.20) i afsnit 5.2. Kommentarerne til (5.20) er fuldgyldige her og skal ikke gentages. Idet μ_0 er den fælles referencemiddelværdi, går den statistiske analyse ud på at estimere differenserne

Analyse

$$\begin{cases} \delta_1 = \mu_1 - \mu_0 \\ \delta_2 = \mu_2 - \mu_0 \\ \cdot \\ \cdot \\ \cdot \\ \delta_p = \mu_p - \mu_0 \quad , \end{cases} \tag{5.37}$$

og herunder bestemme δ-ernes standardafvigelser og konfidensintervaller. Dette kan foretages fuldstændig som beskrevet i afsnit 5.2.

En enkelt, ikke afgørende, modifikation ligger i, at σ^2 kan estimeres fra alle data:

$$\hat{\sigma}^2 = \frac{1}{(n_0-1)+(n_1-1)+\cdots+(n_p-1)} \sum_{k=0}^{p} (n_k-1)\hat{\sigma}_k^2$$

$$= \frac{1}{n_0+n_1+\cdots+n_p-p-1} \left[\sum_{k=0}^{p} \sum_{i=1}^{n_k} [Y_{k,i}-\bar{Y}_k]^2 \right] \ . \tag{5.38}$$

f-faktoren i konfidensintervallet (5.27) skal nu findes i tabellen over t-fordelingen ud for et større antal frihedsgrader, nemlig

$n_0+n_1+\cdots n_p$-p-1, normalt en ubetydelig justering af f.

Det eneste principielt nye, der kommer ind i billedet, er det forhold, at de samme $Y_{0,i}$-er benyttes i alle $\hat{\delta}$-estimaterne. Det har – ikke overraskende – betydning for antalsplanlægningen, hvor man i forhold til det i afsnit 5.2. fundne, bør øge n_0 på bekostning af n_1, $n_2,\cdots,n_p$.

Antalsplanlægning: estimation

I antalsplanlægningen med henblik på signifikanstest kan ligning (5.33) fortsat benyttes blot man på c's plads indsætter Σc_k, og i øvrigt benytter at $n_k = n_1$, $k = 2,3,\cdots,p$.

$$\frac{1}{n_0} + \frac{1}{n_k} \leq [\sigma_{\delta_k}/\sigma]^2 \quad , \quad k = 1,2,\cdots,p. \tag{5.39}$$

Er prisen for en $Y_{k,i}$-måling, faktoren c_k gange så høj som en $Y_{0,i}$-måling, opfyldes (5.39) mest økonomisk ved at vælge

$$\begin{cases} n_0 = \left[1 + \left[\sum_{k=1}^{p} c_k\right]^{\frac{1}{2}}\right][\sigma/\sigma_\delta]^2, \\ n_k = n_0/\left[\sum_{k=1}^{p} c_k\right]^{\frac{1}{2}} . \end{cases} \tag{5.40}$$

Ønskes for eksempel, som i afsnit 5.2, at $\sigma_\delta = 0{,}1$ og $c_k = 1, k = 1,\cdots,p$ med $p = 4$ og σ på forhånd skønnet lig 0,2; da bliver $n_0 = (1+4^{\frac{1}{2}})(0{,}2/0{,}1)^2 = 12$ og $n_k = 12/(4)^{\frac{1}{2}} = 6$, $k = 1,2,3,4$. Er derimod $c_k = 4$, for $k = 1,2,3,4$, bliver det mest økonomiske valg at sætte $n_0 = 20$ og $n_k = 5$.

Antalsplanlægning: test

I antalsplanlægningen med henblik på signifikanstest kan ligning (5.33) fortsat benyttes blot man på c's plads indsætter Σc_k, og i øvrigt benytter at $n_1 = n_k$, $k = 2,3,\cdots,p$.

"Nestet" analyse

Hvis man for eksempel regner med omstændighederne fra afsnit 5.2 ($1\text{-}\beta = 80\%, \alpha = 5\%, \sigma = 0{,}2, \delta_0 = 0{,}3$) med $c_k = 1, k = 1,\cdots,p$ og $p = 4$, finder man $n_0 = (1 + 4^{\frac{1}{2}})3{,}49 = 10{,}5$ og $n_k = 10{,}5/4^{\frac{1}{2}} = 5{,}25$. Er derimod $c_k = 4, k = 1,2,3,4$, findes $n_0 = (1+16^{\frac{1}{2}})3{,}49 = 17{,}5$ og $n_k = 4{,}4$. Disse antal må rundes passende, og måleprogrammet må bygge på en randomiseret afvikling af målingerne. Specielt må målingerne fra hver af de alternative omstændigheder ikke systematisk klumpes i grupper, men blandes mellem hinanden og mellem referencemålingerne $Y_{0,i}$.

Eksempel 5.3

I forlængelse af eksemplerne 5.1 og 5.2 forestiller vi os at 3 alternative modifikationer er ønsket undersøgt over for umodificerede produktionsenheder.

Er præcisionskravene som i 5.2, d.v.s. at man forsigtigt vurderer at σ næppe er større end ca. 11 og man ønsker $\sigma_\delta \leq 3$, kan disse krav tilfredsstilles som ovenfor med $n_0 = n_1 = n_2 = n_3 = 27$.

Imidlertid er det mere økonomisk at udnytte (5.40) og med $c_1 = c_2 = c_3 = 1$ finde $n_0 = (1+(3)^{\frac{1}{2}})(11/3)^2 = 36{,}73 \simeq 37$ og $n_1 = n_2 = n_3 = 37/(3)^{\frac{1}{2}}$

= 21,36 ≃ 22. Det totale antal målinger bliver herved reduceret fra 4 × 27 = 108 til 37 + 3 × 22 = 103 med samme præcision til følge. En så lille reduktion kan vurderes ret ligegyldig andre forhold taget i betragtning, herunder administration af måleprogrammets gennemførelse. I det konkrete tilfælde forestiller vi os, at man af sådanne hensyn har valgt $n_0 = n_1 = n_2 = n_3 = 27$ *og fundet* $\hat{\mu}_0 = 74{,}259$; $\hat{\sigma}_0^2 = 76{,}353$; $\hat{\mu}_1 = 74{,}148$; $\hat{\sigma}_1^2 = 65{,}202$; $\hat{\mu}_2 = 71{,}391$; $\hat{\sigma}_2^2 = 103{,}239$; $\hat{\mu}_3 = 75{,}874$; $\hat{\sigma}_3^2 = 78{,}781$.

Et samlet variansestimat findes af (5.38) som $\hat{\sigma}^2 = (\hat{\sigma}_0^2 + \hat{\sigma}_1^2 + \hat{\sigma}_2^2 + \hat{\sigma}_3^2)/4 = 80{,}894 = 8{,}99^2$ *da* $n_0 = n_1 = n_2 = n_3$.

Modifikationseffekterne estimeres til $\hat{\delta}_1 = \hat{\mu}_1 - \hat{\mu}_0 = -2{,}111$; $\hat{\delta}_2 = \hat{\mu}_2 - \hat{\mu}_0 = -2{,}868$; *og* $\hat{\delta}_3 = \hat{\mu}_3 - \hat{\mu}_0 = 1{,}615$. *Alle tre estimater har samme estimerede standardafvigelse,* $std(\hat{\delta}) = 2{,}447$, *jf. (5.26). Til 95% konfidensintervaller for* δ_1, δ_2 *og* δ_3 *benyttes i (5.27) med* $f = 1{,}984$. *Dette er en lidt mindre værdi end i eksemplet 5.2(a) ovenfor, da* σ *er estimeret på grundlag af flere data (med* $4 \times (27-1) = 104$ *frihedsgrader mod før* $2 \times (27-1) = 52$ *frihedsgrader). 95% konfidensintervallerne for* δ_1, δ_2 *og* δ_3 *er henholdsvis [-6,966; 2,744], [-7,723; 1,987] og [-3,240; 6,470].*

Alle tre intervaller omslutter værdien $\delta = 0$, *hvorfor ingen af de undersøgte modifikationer har vist sig signifikante på* $\alpha = 5\%$ *niveauet.*

Det bemærkes, at konfidensintervallet for δ, *er lidt bredere her end ovenfor i eksempel 5.2(a). Den ændrede længde skyldes to ting, dels er f-faktoren ændret, dels er* $\hat{\sigma}_\delta$ *reestimeret. f-faktoren mindskes når variansestimatet baseres på flere data og herved bliver sikrere. Variansestimatets egen værdi kan gå i begge retninger som følge af større datagrundlag. I nærværende eksempel øges variansestimatet mere end ændringen i f-faktoren kunne ophæve.*

Svarende til eksempel 5.2(c) kunne man også her have valgt at udføre måleprogrammet således, at hver produktionsenhed blev underkastet alle tre alternative modifikationer, og at hver enhed således gav anledning til fire målinger inklusive en umodificeret måling. De fire målinger måtte for hver produktionsenhed udføres i randomiseret orden, og den statistiske analyse herefter foregå som parrede analyser efter samme mønster som gennemgået i eksempel 5.2(c).

5.4 Forskel mellem flere sideordnede middelværdier

I nogle undersøgelser er man i en situation, som den i afsnit 5.3 omtalte, bortset fra at alle betragtede middelværdier er niveauer for på forhånd sideordnede alternativer. Der er ingen standardproces eller sædvanlig teknologi hvoroverfor indgreb og ændringer skal vurderes, alle muligheder indgår på lige fod.

Model

I denne situation kan den statistiske analyse tage udgangspunkt i modellen

$$Y_{k,i} = \mu_k + e_{k,i} \qquad \begin{matrix} i = 1,2,\cdots,n_k \\ k = 1,2,\cdots,p \end{matrix} \quad , \tag{5.41}$$

meget lig (5.36), eller i den dermed ækvivalente model

$$Y_{k,i} = \mu + \alpha_k + e_{k,i}, \tag{5.42}$$

hvor index er som ovenfor. μ er en generel middelværdi, α_k er det k'te alternativs afvigelse fra μ, med begrænsningen $\Sigma\alpha_k = 0$ (eller $\Sigma n_k \alpha_k = 0$) , da ellers μ er ubestemmelig.

Individuel analyse

Analyseres i første omgang ud fra (5.41), vil de individuelle middelværdier estimers og evt. testes over for hypotetiske værdiforslag, som omtalt i afsnit 5.1. Dog kan estimatet af den statistiske støj baseres på alle data efter retningslinierne i ligning (5.38). Ud fra hensyn til denne analyse kan antalsplanlægningen foretages som aftalt i afsnit 5.1.

Man kan også være interesseret i at estimere forskelle mellem de alternative middelværdier, og dette kan i givet fald gøres som gennemgået i afsnit 5.2; antalsplanlægningen kan gennemføres i overensstemmelse hermed. Brug af test for forskelle mellem middelværdier, d.v.s. test om differenserne δ-erne er nul, bliver imidlertid nu noget dubiøs. Problemet er, at selv i situationer, hvor ingen reel forskel eksisterer mellem middelværdierne, vil der p.g.a. den statistiske støj forekomme δ-estimater af varierende størrelse. Udvælges nu blandt de mange muligheder den største, vil den vise større signifikans end berettiget, netop fordi den er specielt udvalgt blandt mange. I statistikken betegnes dette fænomen selektion. Ønsker man at foretage alle parvise sammenligninger må man benytte de statistiske teknikker, som går under navnet »multiple comparison«, se f.eks. Hichs (1964) eller Box *et al.* (1978).

Selektion

En anden fremgangsmåde består i først at teste statistisk, om alle p μ_k-ere set under ét, er signifikant forskellige efter data at dømme. Nulhypotesen $H_0 : \mu_1 = \mu_2 = \cdots = \mu_p$ kan testes ved brug af en ensidet variansanalyse. (Brugen her af ordet »ensidet« har ingen forbindelse med brugen af samme ord i forbindelse »ensidet test«.) Er μ_k-erne signifikant forskellige, kan man estimere deres værdier individuelt eller estimere særligt interessante differenser, som nævnt ovenfor. Man kan undlade videre analyse, hvis der ikke er signifikant forskel, og hvis man tager dette som udtryk for, at μ_k-erne bør antages ens. Men jf. diskussionen i afsnit 4.1, betyder manglende signifikans ikke, at μ_k-erne er statistisk bevist ens. En sagkyndig vurdering i den konkrete undersøgelse kan gå ud på, at forskelle – af helt eller delvis ukendt størrelse – må eksistere, hvorfor man ser bort fra testresultatet, og foretager estimation trods alt. I så tilfælde ville det være konsistent helt at undlade signifikanstestning.

Variansanalyse

Testning af $H_0 : \mu_1 = \mu_2 = \cdots = \mu_p$ ved variansanalyse udføres af standardcomputerprogrammer, idet output i reglen refererer til modelformuleringen (5.42), så $H_0 : \alpha_1 = \alpha_2 = \cdots = \alpha_p$. Uden at gå i detaljer foretages en udregning af to middelkvadratsummer: MS_b og MS_w . MS_w giver et variansestimat af den statistiske støj σ^2, og udregnes i stil med (5.38), som en gennemsnitlig afvigelseskvadrat-

sum inden for (»within«) hver gruppe af målinger. MS_b er også en gennemsnitlig afvigelseskvadratsum, men dannes mellem (»between«) gruppegennemsnittene, $\bar{Y}_1$, $\bar{Y}_2$, ... $\bar{Y}_p$. MS_b indeholder således to varianskomponenter, nemlig den statistiske støj σ^2, plus et bidrag fra forskellene mellem α_k -erne. Er α_k -erne ens, estimerer MS_w og MS_b begge σ^2 og deres forhold MS_w / MS_b udgør en teststatistik, som efter principperne diskuteret i afsnit 4.1, refereres til en F-fordeling med frihedsgraderne (p-1) og $(n_1 + n_2 + \cdots + n_p - p)$. En utypisk stor ($MS_b$ / MS_w -)værdi fører ved denne F-test til forkastelse af H_0, α_k -erne er altså signifikant forskellige.

Eksempel 5.4(a)

Vi forestiller os, at tallene i eksempel 5.3 ovenfor stammede fra en situation, hvor de fire muligheder for procesenhedernes arbejdsmåde på forhånd er sidestillede. En ensidet variasanalyse vil nu producere følgende variansanalyseskema, med de i computer-output typisk forekommende engelske betegnelser:

SOURCE	SS	DF	MS	F
BETWEEN	336,44	3	112,148	1,386
WITHIN	8412,95	104	80,894	
ADJ. TOTAL	8749,39	107		

MS-værdien i WITHIN-linien er MS_w, og estimerer σ^2 (samme værdi som i eksempel 5.3). MS-værdien i BETWEEN-linien er MS_b som er udregnet på grundlag af forskellene mellem de fire $\hat{\mu}$-er. Tallet under F er MS_b / MS_w, og givet H_0-hypotsen, at der ikke er reel forskel på procesenhederne i de fire modifikationer, skulle dette tal stamme fra en F-fordeling med frihedsgraderne 3 og 104. Den hertil svarende p-værdi er udregnet til 0,2512 , d.v.s. at hypotesen kan accepteres ved signifikanstest på niveau $\alpha < 0{,}2512$, herunder på niveau $\alpha = 5\%$. Der er altså ikke signifikant forskel på de fire modifikationer af procesenhederne. En fælles middelværdi μ, estimeres som gennemsnittet af alle observationer til $\hat{\mu} = 73{,}418$ og kan statistisk belyses som omtalt i afsnit 5.1. Skulle man trods alt tro på, at der må være forskel på de fire modifikationer, kan deres individuelle middelværdier estimeres separat, og belyses statistisk som omtalt i afsnit 5.1.

»Fixed« vs. »random«

Denne ensidede variansanalyse har i sammenhæng med nærværende udredning flere anvendelser. I forlængelse af diskussionen om »fixed« versus »random« effekt i afsnit 2.3 om repræsentativitet, ses det, at i modelformen (5.42) kan α_k-erne opfattes som udtryk for disse effekter. Ovennævnte F-test kan om ønsket benyttes til signifikanstest af såvel »fixed« som »random« effekter. I »fixed« effekt tilfældet følger det direkte af ovenstående forklaring. For »random« effekt tilfældet ændres fortolkningen, men ikke testens beregning. I dette tilfælde opfattes α_k-erne som bidrag til den naturlige variabilitet i stil med udtrykkene (5.6) og (5.8) og denne naturlige variabilitet kan nu estimeres. Dette foregår enkelt hvis

$n_1 = n_2 = \cdots = n_k = n$ ellers ikke. Kaldes α_k-erne varians $\sigma^2(\alpha)$, estimerer MS_b nemlig $n\sigma^2(\alpha) + \sigma^2$, således at

$$\hat{\sigma}^2(\alpha) = (MS_b - MS_w)/n \quad . \tag{5.43}$$

Gentagne målinger

Det ses specielt, at denne type variansanalyse ved hensigtsmæssig forsøgsplanlægning kan estimere den naturlige varians og målevariansen, som diskuteret i afsnit 5.1. I eksemplet der, hvor 3 målinger tænktes foretaget på hvert måletidspunkt, bliver i nærværende notation n = 3 og p lig antallet af måletidspunkter (dér kaldt n).

Når $n_1 = n_2 = \cdots = n_p \geq 2$, vil usikkerheden i estimationen af $\sigma^2(\alpha)$ primært afhænge af antallet af grupper, p. Sættes f.eks. p til omkring 10, vil $\hat{\sigma}(\alpha)$ være behæftet med en usikkerhed af størrelsesordenen 30% (approksimativ standardafvigelse på $\hat{\sigma}(\alpha)$). Antalsplanlægning med henblik på variansanalysens F-test og varianskomponentestimation, kan udføres ud fra tabeller i Beyer (1968) og Owen (1962); men skal ikke omtales her.

»Nestet« analyse

Større hierarkier af varianskomponenter med tilhørende »nestet« variansanalyse er bl.a. omtalt i Box *et al.* (1978) og Hicks (1964). Eksempler på udnyttelse af variansanalyser i forureningssammenhæng findes i Pallesen (1987).

Eksempel 5.4(b)

Vi forestiller os, at man af hensyn til forbedret præcision i den eksperimentelle undersøgelse har foretaget 3 målinger på hver af de 4 × 27 = 108 procesenheder. Det vil sige, at der er foretaget ialt 3 × 108 = 324 målinger, som kan belyse måleusikkerhed, naturlig variabilitet mellem procesenheder (med samme modifikation) og forskel mellem modifikationer. Det ville være forkert at opfatte de 81 målinger pr. modifikation som sideordnede da de ikke stammer fra 81 men kun 27 procesenheder; den statistiske analyse i eksempel 5.4(a) kan derfor ikke anvendes. Derimod kan man i computerprogrammer specificere måleprogrammets »nestede« eller »hierarkiske« karakter og få følgende variansanalyseskema:

SOURCE	SS	DF	MS
BETWEEN MAIN GROUPS	1009,32	3	336,44
BETWEEN SUBGROUPS WITHIN MAIN GROUPS	25238,85	104	242,68
WITHIN SUBGROUPS	1377,17	216	6,375
TOTAL	27625,34	323	

MS-værdien i linien WITHIN SUBGROUPS estimerer måleusikkerheden, som er fundet ud fra 3 gentagelser for hver produktionsenhed. Denne måleusikkerhed er estimeret til $Var(e_i(m)) = 6{,}376 = 2{,}5^2$*; altså en standardafviglse på ca. 2,5 . MS-værdien i linien BETWEEN SUBGROUPS WITHIN MAIN GROUPS estimerer målevariansen plus et evt. bidrag fra den naturlige variation mellem produktionsenheder; mere specifikt estimeres* $Var(e_i(m)) + 3\ Var(e_i(n))$*, hvor tretallet stammer fra*

antallet af gentagelser for hver produktionsenhed. Under en nulhypotese om at produktionsenhederne er ens, så al statistisk varians skyldes måleusikkerhed, vil brøken $MS_{bw}/MS_w = 242{,}68/6{,}376 = 38{,}06$ *stamme fra en F-fordeling med 104 og 216 frihedsgrader. Til værdien 38,06 for denne F-statistik vil et computeroutput angive en p-værdi på 0,00000 , altså må nulhypotesen forkastes på alle signifikansniveauer større end i hvert tilfælde* $\alpha = 0{,}00001$*, specielt på niveau* $\alpha = 5\%$ *og* $\alpha = 1\%$*. Der er altså en betydelig naturlig veriation mellem produktionsenheder. Denne variation kan estimeres som* $Var(e_i(n)) \simeq (243{,}68 - 6{,}376)/3 = 79{,}10 = 8{,}89^2$*, altså betydeligt større end målevariansen* $Var(e_i(m)) = 6{,}376 = 2{,}5^2$*. Jf. diskussionen i afsnit 5.1 er den statistiske fejl summen af disse to bidrag. Var den naturlige variation nul, kunne man nøjes med at måle gentagne gange på samme produktionsenhed for hver modifikation. Her hvor den naturlige variation er stor, giver gentagne målinger på samme enhed kun ringe forbedring i total præcision, og måleusikkerheden, som estimeres fra disse gentagelser, giver intet billede af den virkelige statistiske usikkerhed,* σ^2*. Tages én måling pr. produktionsenhed bliver* $\sigma^2 = Var(e_n) + Var(e_m) \simeq 79{,}10 + 6{,}376 = 85{,}48 = 9{,}24^2$*, altså en standardafvigelse på ca. 9,24. Tages 3 målinger på hver produktionsenhed, og benyttes gennemsnittet af disse data til belysning af modifikationsforskellene, bliver den statistiske usikkerhed på disse data* $\hat{\sigma}^2 = Var(e_n) + (1/3)Var(e_m) \simeq 79{,}10 + 6{,}376/3 = 81{,}225 = 9{,}01^2$*, altså en standardafvigelse på ca. 9,01 . Selv et ubegrænset antal gentagne målinger på samme produktionsenhed, kan ikke reducere* σ^2 *ned under* $Var(e_n) \simeq 79{,}10 = 8{,}89^2$*.*

I det betragtede eksempel, hvor der er foretaget 3 gentagelser, vurderes signifikansen af eventuelle forskelle mellem de 4 modifikationer ved at betragte F-statistikken $MS_w/MS_{bw} = 336{,}44/242{,}68 = 1{,}386$ *og referere denne størrelse til en F-fordeling med 3 og 104 frihedsgrader. Hertil svarer en p-værdi på 0,2512, d.v.s. at en nulhypotese om ingen reel forskel på de fire modifikationer, kan accepteres på alle signifikansniveauer* $\alpha <$ *0,2512, herunder på niveau* $\alpha = 5\%$*. Det bemærkes, at nærværende statistiske analyse af modifikationsforskelle hvor alle 324 observationer i det beskrevne måleprogram benyttes, giver nøjagtigt samme resultat som analysen i eksempel 5.4(a) hvor 108 observationer indgik. Som beskrevet i nærværende eksempel var de 108 observationer netop gennemsnit af 3 gentagelser, altså fra samme rådata. Havde man derimod fejlagtigt opfattet de* 3×27 *observationer for hver modifikation som 81 parallelle målinger for hver modifikation og udført beregningerne som i eksempel 5.4(a), var analysen gået galt. Således specificeret ville et computerprogram have fundet* $MS_w = (25238{,}85 + 1377{,}17)/(104 + 216) = 83{,}175$*; og F-statistikken* $MS_b/MS_w = 4{,}045$ *ville refereret til en F-fordeling med 3 og 320 frihedsgrader give en p-værdi på 0,00763. Man ville altså have draget den helt misledende konklusion, at der var stærkt signifikant forskel på modifikationerne* ($\alpha = 1\%$)*.*

5.5 **Korrelation**

Mange undersøgelser tager sigte på at etablere hvorvidt, der er sammenhæng mellem visse målevariable og, hvis en sammenhæng eksisterer, beskrive og udnytte den kvantitativt.

Det kan tænkes at visse procesomstændigheder, kontrollerbare eller ikke-kontrollerbare, påvirker en koncentration af et giftigt reststof i processpildevandet, eller emissionen af luftforurenende flygtige forbindelser eller andet.

Korrelations koefficient

I sådanne situationer tales ofte i en løs forstand om korrelation mellem to varibale, X og Y. I statistisk dataanalyse betyder ordet korrelation det samme som korrelationskoefficient, og har en bestemt definition. For et antal observationspar defineres den observerede eller empiriske korrelationskoefficient som

$$r_{XY} = \frac{\sum_i (X_i-\bar{X})\,(Y_i-\bar{Y})}{(\sum_i (X_i-\bar{X})^2 \sum_i (Y_i-\bar{Y})^2)^{\frac{1}{2}}} \tag{5.44}$$

Ved den teoretiske eller sande korrelationskoefficient ρ_{XY} forstås det tal, som man ville finde i (5.44) hvis antallet af observationer var ubegrænset. Opfattet som et estimat af denne sande ρ_{XY}, skrives r_{XY} ofte $\hat{\rho}_{XY}$. Ækvivalent med (5.44) kan man derfor skrive

$$\hat{\rho}_{XY} = \frac{\hat{\sigma}^2_{XY}}{\hat{\sigma}_X\,\hat{\sigma}_Y} \quad , \tag{5.45}$$

Kovarians

hvor $\hat{\sigma}_X$ og $\hat{\sigma}_Y$ er de estimerede standardafvigelser af henholdsvis X_i-erne og Y_i-erne udregnet som i (5.10), og kovariansen, Covar (X,Y) = σ^2_{XY}, udregnet som

$$\hat{\sigma}^2_{XY} = \frac{1}{n-1} \sum_{i=1}^{n} (X_i-\bar{X})\,(Y_i-\bar{Y}) \quad . \tag{5.46}$$

Kovariansen er et mål for hvor meget de to variable kovarierer eller samvarierer. Det ses ud af udtrykket (5.46), at hvis der til en X_i-værdi større end $\bar{X}$ hører en Y_i-værdi større end $\bar{Y}$, giver dette et positivt bidrag til kovariansen; det samme er tilfældet med sammenhørende (X_i,Y_i), som er mindre end $(\bar{X},\bar{Y})$. Omvendt hvis afvigelserne fra $\bar{X}$ og $\bar{Y}$ går i modsat retning for henholdsvis X_i og Y_i.

Korrelationen eller korrelationskoefficienten er i realiteten lig kovariansen, normeret med standardafvigelserne for X og Y. ρ_{XY} bliver således en dimensionsløs størrelse uberørt af den benyttede måleenhed. Det skal bemærkes, at korrelationskoefficienten er symmetrisk i X og Y, d.v.s. $\rho_{XY} = \rho_{YX}$. Videre ses det, at $-1 \le \rho_{XY} \le 1$, hvor ρ_{XY} lig +1 eller -1 betyder perfekt korrelation, d.v.s. at X og Y varierer i henholdsvis eksakt fase eller modfase.

Model

Som statistisk model for korrelationsanalysen kan vi opskrive

$$\begin{cases} X_i = \mu_X + e_{X,i} \\ Y_i = \mu_Y + e_{Y,i} \end{cases} \quad i = 1,2,\cdots,n \quad , \tag{5.47}$$

Denne model ligner overfladisk (5.20), men observationerne X_i og Y_i

er i (5.47) parrede og $e_{X,i}$ og $e_{Y,i}$ er ikke statistisk uafhængige, ej heller har de nødvendigvis samme varians. Korrelationen mellem X og Y kan ses som korrelationen mellem e_X og e_Y.

Transformation

Figur 3.8 viser en situation, hvor der er positiv korrelation mellem X og Y . På denne figur er det i virkeligheden log(X) der er afbildet, således at figuren giver det visuelle indtryk af korrelationen mellem log(X) og Y . Var plottet udført med X og Y ville ikke blot det visuelle indtryk ændres, men den udregnede r_{XY}-værdi er også forskellig fra $r_{\log(X)Y}$. Korrelationen mellem to variable er således ikke invariant over for ikke-lineære transformationer. Almindeligvis bør man vælge en transformation af X og Y, således at deres individuelle fordelinger bliver approximativt symmetriske. I forhold til korrelationen af rådata, kan denne blive større eller mindre, men værdien bliver lettere at fortolke, når approximativ symmetri er til stede.

Korrelationsbegrebet er særligt vigtigt i forbindelse med passivt (men systematisk) indsamlede data. Som diskuteret i afsnit 2.2 kan man ved en passivt observeret sammenhæng aldrig med sikkerhed statistisk afgøre retningen af årsag og virkning. Dette afspejles vel i korrelationsbegrebets symmetriske natur, som derfor er et hensigtsmæssigt kvantitativt mål for graden af samvariation i samhørende målepar, d.v.s. en bivariat fordeling.

Bivariat fordeling

Korrelation mellem to variable X og Y er ikke kun naturbestemt. Den påvirkes dels af den valgte transformation, et valg der altid foretages om end undertiden ubevidst ved valg af f.eks. måleinstrument (en pH-måling er f.eks. den negative logaritme til brintionkoncentrationen).

Udvalgte data

Desuden afhænger korrelationen af måleprogrammets planlægning. Vælges det f.eks. kun at betragte Y-værdier over et mere begrænset interval end vist i figur 3.8, bliver korrelationen af punkterne i det vandrette bånd som intervallet definerer som hovedregel mindre end uden denne begrænsning. Dette til trods for at relationerne mellem X og Y ikke ændres af sorteringen. Korrelationsangivelser bør normalt ikke omhandle udvalgte delmængder af data, ret vilkårlige og misledende tal kan resultere.

Forsøgsdata

For aktivt indsamlede data, hvor man efter en bestemt forsøgsplan har varieret f.eks. X og så observeret resulterende Y-er, vil korrelationen mellem X og Y ikke kun afhænge af hvorledes X påvirker Y , men i høj grad af hvorledes X varieres. Stort variationsinterval for X har tendens til at give høje numeriske værdier for korrelationen, og vice versa. Korrelationsangivelser er derfor i sådan sammenhæng næsten meningsløse, og man bør her i stedet opstille en simpel lineær model v.h.a. regressionsanalyse, se afsnit 5.6.

Varianskomponenter

Endelig afhænger korrelationen af hvilke naturlige variationskilder, der er omfattet af måleplanen, samt af måleusikkerhedens variationsbidrag. Hvad angår den naturlige variation, må en observeret korrelation fortolkes i relation til netop de inddragne variationskomponenter, (se evt. sektion 2.3 og 5.1); korrelationen mellem samme variable kan mere generelt eller mere specielt være helt anderledes. Hvis f.eks. korrelationen mellem luftfugtighed og tørretid i en bestemt overfladebehandlingsfabrik er fundet, kan korrelationen være

større eller mindre, hvis tal fra mange fabrikker analyseres, eller hvis beregningerne udførtes på tal fra kun et enkelt produkt i fabrikken.

Målefejl

Hvad angår målefejl, vil disse have en tendens til at sløre den underliggende sande korrelation mellem de betragtede variable. Spaltes den statistiske variation e_X og e_Y op i deres komponenter, naturlig variabilitet og måleusikkerhed jf. (5.8), får (5.47) formen

$$\begin{cases} X_i = \mu_X + e_{X,i}(n) + e_{X,i}(m) \\ Y_i = \mu_Y + e_{Y,i}(n) + e_{Y,i}(m) \end{cases} \tag{5.48}$$

Uden målefejl overhovedet ville korrelationen mellem X og Y være lig korrelationen mellem den naturlige variation i X og Y, altså mellem $e_X(n)$ og $e_Y(n)$. Denne underliggende korrelationskoefficient kaldes $\rho_{XY}(\max)$. Det kan vises, at denne korrelation fortyndes af måleusikkerhed ifølge formlen

$$\rho_{XY} = \rho_{XY}(\max)\,(R_X R_Y)^{\frac{1}{2}} \quad , \tag{5.49}$$

»Reliability«

hvor R_X og R_Y er målingernes pålidelighed (»reliability«), defineret ved

$$\begin{cases} R_X = \dfrac{\sigma_X^2(n)}{\sigma_X^2(n) + \sigma_X^2(m)} \quad , \\ R_y = \dfrac{\sigma_Y^2(n)}{\sigma_Y^2(n) + \sigma_Y^2(m)} \quad . \end{cases} \tag{5.50}$$

Nævneren i disse udtryk er den samlede statistiske varians i målingerne, jf. (5.9). Hvis den totale varians har en standardafvigelse på f.eks. 0,2 og måleusikkerheden har en standardafvigelse på 0,1 , bliver pålideligheden $(0{,}2^2\text{-}0{,}1^2)/(0{,}2)^2 = 0{,}75$ eller 75%. Gælder dette tal for både X- og Y-målinger, bliver $\rho_{XY} = 0.75 \times \rho_{XY}(\max)$. Det betyder, at selv hvis den underliggende korrelation er perfekt ($= \pm 1$) vil man p.g.a. måleusikkerheden i praksis finde en korrelationskoefficient på numerisk kun omkring 0,75. Ved større datasæt planlagt så separate estimater af varianskomponenterne er mulig, som omtalt sidst i afsnit 5.4, vil man dog kunne udregne et estimat af den underliggende korrelation $\rho_{XY}(\max)$.

Regressionslinier

For korrelerede variable kan man opstille en regressionslinie til prediktion af den ene variabel på grundlag af den anden. Denne linie går altid gennem punktet $(\bar{X},\bar{Y})$, som er tyngdepunktet for prikkerne i et X-Y diagram; men dens hældning afhænger af om den er ment til et prediktere Y fra X eller X fra Y:

$$\hat{Y}_i(X_i) = \bar{Y} + b_{Y|X}\,(X_i\text{-}\bar{X}) \tag{5.51}$$

$$\text{med } b_{Y|X} = \hat{\rho}_{XY}\,\frac{\hat{\sigma}_Y}{\hat{\sigma}_X} \quad ,$$

medens

$$\hat{X}_i(Y_i) = \bar{X} + b_{X|Y}(Y_i - \bar{Y})$$

$$\text{med } b_{X|Y} = \hat{\rho}_{XY} \frac{\hat{\sigma}_X}{\hat{\sigma}_Y} \quad . \tag{5.52}$$

X indgår på samme måde i (5.51) som Y gør i (5.52) og vice versa, helt i pagt med at korrelationsbegrebet er symmetrisk i X og Y.

To hældninger

Men det ses også at $b_{Y|X} \neq (b_{X|Y})^{-1}$ med mindre $\rho_{XY} = \pm 1$, hvor punkterne ligger perfekt på en ret linie.

Det forhold, at der findes to og ikke én regressionslinie, er ikke en inkonsistens, men en konsekvens af datapunkternes spredning. Denne kan dels skyldes måleusikkerhed, men derudover også at den naturlige variabilitet ikke er perfekt korreleret.

Linierne (5.51) og (5.52) giver i deres respektive retninger prediktioner, som er uden systematisk fejl. Brug af en fælles linie ville introducere sådan fejl i den ene eller begge retninger. (5.51) og (5.52) udregnes af almindelige regressionsprogrammer til regnemaskiner, men det er altså ikke ligegyldigt, hvilken variabel man i programmet gør til den »afhængige« og hvilken til den »uafhængige«.

Prediktionsfejl

Prediktionsfejlene ved at benytte (5.51) og (5.52) har henholdsvis varianserne

$$\text{Var } Y_i|X_i = \sigma_Y^2(1 - \rho_{XY}^2) \quad , \tag{5.53}$$

og

$$\text{Var } X_i|Y_i = \sigma_X^2(1 - \rho_{XY}^2) \quad . \tag{5.54}$$

Uden kendskab til den anden variabel vil den bedste prediktion af X_i og Y_i være henholdsvis $\bar{X}$ og $\bar{Y}$ med de tilhørende prediktionsvarianser σ_X^2 og σ_Y^2. I begge tilfælde sker der altså en variansreduktion ved udnyttelse af den information som ligger i kendskab til den anden variabel. Faktoren $(1 - \rho_{XY})$ kvantificerer denne information. Formlerne (5.53) og (5.54) gælder kun når et så tilstrækkeligt stort antal (X,Y)-par er indsamlet, at μ_X, μ_Y, σ_X^2, σ_Y^2 og ρ_{XY} kan estimeres essentielt uden fejl. For passivt observerede variable er det ofte økonomisk muligt at opnå et så stort antal, at antagelsen er rimelig. I denne forbindelse spiller det også en rolle, at prediktionerne $\hat{Y}_i(X_i)$ og $\hat{X}_i(Y_i)$ kun benyttes for passivt observerede nye prediktorer X_i og Y_i henholdsvis. Jf. diskussionen i afsnit 2.3 og slutningen af afsnit 2.2, kan statistiske analyseresultater baseret på passivt observerede data kun benyttes troværdigt i fremtidige anvendelsessituationer, der også er passive. Herunder undgås også meget ekstreme værdier for prediktorerne, som ville gøre (5.53) og (5.54) til særligt ringe approximationer. Er datamaterialet lille, eller voves en ekstrapolerende modelbrug ud over det passivt oplevede, må afsnit 5.6's formler benyttes.

Computerprogrammer til korrelationsanalyse, udregner foruden

Test for $\rho_{XY} = 0$

punktestimater for ρ_{XY}, i reglen også en p-værdi for hypotesen $H_0: \rho_{XY} = 0$. I nogle tilfælde opgives også konfidensinterval for ρ_{XY}.

Konfidensintervaller for ρ_{XY} opstilles på grundlag af en transformation (Fisher's transformation) af parameteren ρ_{XY}. Først udregnes statistikken

$$Z_F(r_{XY}) = \frac{1}{2}\log_e\left[\frac{1 + r_{XY}}{1 - r_{XY}}\right] \quad , \tag{5.55}$$

som er approximativt normalfordelt med middelværdi $Z_F(\rho_{XY})$ og variansen $1/(n-3)$. Et $1-\alpha$ konfidensinterval for $Z_F(\rho_{XY})$ er

$$[Z_F(r_{XY}) - (n-3)^{-\frac{1}{2}} z_{1-\alpha/2};\ Z_F(r_{XY}) + (n-3)^{-\frac{1}{2}} z_{1-\alpha/2}] \quad , \tag{5.56}$$

hvor $z_{1-\alpha/2}$ er tabelværdien for $1-\alpha/2$ fraktilen af en standard normalfordeling. Intervalgrænserne i (5.56) transformeres dernæst tilbage til den originale skala ved at benytte (5.55) inverst, herved fremkommer konfidensintervallet for ρ_{XY}. Dette interval fortolkes som diskuteret i afsnit 4.2.

Test for hvorvidt en observeret korrelation i realiteten blot kan skyldes tilfældigheder og ikke har nogen basis i virkeligheden, kan afgøres ved at teste nulhypotesen $H_0 : \rho_{XY} = 0$, jf. afsnit 4.1.

Antalsplanlægning

Ved korrelationsanalyse er antalsplanlægning sjældent så afgørende et aspekt som ved undersøgelsestyperne omtalt i afsnittene 5.1 til 5.4. Alligevel kan det undertiden være af interesse at danne sig et indtryk af hvilken præcision, der kan forventes fra et datamateriale af given størrelse.

Fra observationsantallet n beregnes den approximative standardafvigelse for $Z_F(r_{XY})$ som $(n-3)^{-\frac{1}{2}}$. For en mulig ρ_{XY}-værdi findes den tilsvarende $Z_F(\rho_{XY})$-værdi. Dette tal tillægges og fratrækkes $(n-3)^{-\frac{1}{2}}$ og disse grænser tilbagetransformeres ved invers brug af (5.55). De herved fundne værdier giver approximativt plus/minus én standardafvigelse omkring den mulige ρ_{XY}-værdi. Beregningerne kan gentages for andre mulige ρ_{XY}-værdier. Standardafvigelserne afhænger af ρ_{XY} og er ikke helt lige store i plus- og minusretningen. Dette svarer til A-situationen i antalsplanlægning.

B-situationen, hvor antalsplanlægningen handler om signifikanstest, vil i givet fald sigte på at forkaste, at den sande korrelation kan være nul. Man ønsker med sandsynlighed $1-\beta$ sikret, at en sand korrelationskoefficient på ρ^0_{XY} fører til forkastelse af $H_0 : \rho_{XY} = 0$ på niveau α . Kravet til n er

$$n > 3 + (z_{1-\beta} + z_{1-\alpha/2})^2/(Z_F(\rho^0_{XY}))^2. \tag{5.57}$$

Korrelationsanalyse er et større emne og der kan bl.a. henvises til Hogg and Graig (1970) vedrørende den bivariate normalfordeling, til Box *et al.* vedrørende modelbygning og fortolkning, og i øvrigt til regressionsreferencerne anført sidst i afsnit 5.6.

Eksempel 5.5

Fra $n = 83$ sammenhørende værdier for (X,Y) har man beregnet $\bar{X} = 11{,}41$, $\bar{Y} = 35{,}14$, $\hat{\sigma}_X^2 = 1{,}1368 = 1{,}066^2$, $\hat{\sigma}_Y^2 = 5{,}1141 = 2{,}261^2$ og $\hat{\sigma}_{XY} = 1{,}5716$. Heraf findes fra (5.45) at $\hat{\rho}_{XY} = 0{,}6518$, og af (5.51) og (5.52) henholdsvis $b_{YX} = 1{,}3824$ og $b_{XY} = 0{,}3073$.

Ønskes Y_i predikteret fra kendskab til at den sammenhørende X_i er lig 12, findes $\hat{Y}_i(X_i = 12) = 35{,}14 + 1{,}3824(12 - 11{,}41) = 35{,}9556$. Ønskes omvendt X_i estimeret fra kendskab til $Y_i = 36{,}5$ findes $\hat{X}_i(Y_i = 36{,}5) = 11{,}41 + 0{,}3073(36{,}5 - 35{,}14) = 11{,}8279$. At en Y-værdi på 36,5 gav $\hat{X}_i = 11{,}8279 < 12$ til trods for at korrelationen mellem X og Y er klart positiv, skyldes at der er tale om to forskellige regressionslinier i de to prediktionsretninger.

Fra (5.53) og (5.54) findes henholdsvis $Var(Y_i|X_i) = 5{,}1141(1 - 0{,}6518^2) = 2{,}9414 = 1{,}715^2$ og $Var(X_i|Y_i) = 1{,}1368(1 - 0{,}6518^2) = 0{,}6538 = 0{,}8086^2$. Fra $Z_F(r_{XY} = 0{,}6518) = 0{,}7784$ med standard afvigelsen $(83-3)^{-\frac{1}{2}} = 0{,}1118$, kan et 95% konfidensinterval for $Z(\rho_{XY})$ findes som $[0{,}7784 - f \times 0{,}1118 \; ; \; 0{,}7784 + f \times 0{,}1118] = [0{,}5593 ; 0{,}9976]$ med $f = z_{1-\alpha/2} = 1{,}960$. Disse Z-intervalgrænser transformeres ved invers brug af (5.55) tilbage til r-intervalgrænserne $[0{,}5075 ; 0{,}7606]$, som således er et approximativt 95% konfidensinterval for ρ_{XY}. Intervallet omslutter ikke $\rho_{XY} = 0$, denne hypotese forkastes derfor på niveau $\alpha = 5\%$. De fleste computerprogrammer vil producere en p-værdi for hypotesen $\rho_{XY} = 0$, her $p < 0{,}00001$, (ofte i computeroutput angivet $p = 0{,}00000$), så korrelationen er stærkt signifikant.

5.6 Regression

Det er et meget hyppigt formål med en undersøgelse at etablere en modelsammenhæng mellem en styrbar variabel og en anden. Man kunne være interesseret i at finde sammenhængen mellem dannet spildprodukt i en kemisk reaktion som funktion af temperatur, mellem opholdstid i en sedimentationstank og det udledte spildevands indhold af forurenede partikler, mellem luftoverskud i en forbrændingsproces og emission af giftige forbindelser, etc.

Model

Den statistiske analyse kan tage udgangspunkt i den simple lineære model

$$Y_i = \beta_0 + \beta_1 X_i + e_i \quad , \quad i = 1,2,\cdots n \quad , \tag{5.58}$$

med e_i-erne uafhængigt normalfordelt med middelværdi 0 og varians σ^2. Denne type model er særligt velegnet i situationer, hvor der er foretaget en aktiv forsøgsplanlægning vedrørende omstændigheden X_i i de n forsøg.

Den aktive planlægning i kombinationen med passende randomisering sikrer, som omtalt i afsnit 2.2 og 2.3, at demonstrationen af en signifikant statistisk sammenhæng kan tolkes som en kausal virkning af X på Y.

Transformation

Før modellen (5.58) fittes, d.v.s. parametrene estimeres, må et

X-Y-plot tegnes for at se, om der er mening i den påtænkte regressionslinie. Undertiden kan regressionen gøres mere fornuftig ved transformation af Y_i-erne primært med henblik på stabilisering af variansen, således at punktspredningen omkring regressionslinien er ca. ens overalt. Plot af residualer mod X_i-erne kan her være afslørende.

Transformation af X_i udføres med henblik på at linearisere sammenhæng mellem X og Y. Findes f.eks. at et plot af Y mod X^{-1} er klart mere lineært end et af Y mod X, opfattes de reciproktransformerede X-er blot som X-data. Eller skrevet på anden måde, modellen

$$Y_i = \beta_0 + \beta_1(X_i^{-1}) + e_i \tag{5.59}$$

er i statistisk forstand en lineær model, i og med at højresiden er lineær i β-parametrene. (5.59) omfattes derfor af regressionsanalysen.

Estimation

Parametrene β_0 og β_1 estimeres i computerprogrammer ved

$$\begin{cases} b_1 = \hat{\beta}_1 = \hat{\rho}_{XY} \quad \hat{\sigma}_Y/\hat{\sigma}_X \quad , \\ b_0 = \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1\bar{X} \quad . \end{cases} \tag{5.60}$$

hvor $\hat{\rho}_{XY}$ er defineret i (5.45), og $\hat{\sigma}_X$ og $\hat{\sigma}_Y$ er de estimerede standardafvigelser, udregnet som i (5.10). Det bemærkes, at (5.60) er identisk med regressionslinien (5.51) selvom dataomstændighederne og fortolkningen er forskellig.

Beregninger

De sædvanlige computerprogrammer giver et estimat af variansen på den statistiske fejl σ^2, giver estimerede standardafvigelser på b_0 og b_1, samt fortæller om deres værdi er signifikant forskellig fra nul. Normalt vil man i første omgang hæfte sig ved hældningen. Er den ikke signifikant forskellig fra nul, kan det altså ikke fra data afvises, at forbindelsen fra X til Y slet ikke eksisterer.

Forsøgsplanlægning

Af hensyn til forsøgsplanlægningen er det vigtigt at notere sig udtrykket for variansen på hældningsestimatet:

$$\mathrm{Var}(\hat{\beta}_1) = \sigma^2 / \sum_{i=1}^{n}(X_i - \bar{X})^2 \quad . \tag{5.61}$$

Heraf ses, at ikke blot antallet af observationer, men også deres placering er af betydning for b_1's præcision. Jo større spredning des større præcision. På den anden side er der også grænser for hvor ekstremt X_i-erne bør lægges; langt uden for området af praktisk interesse kan være meningsløst. Dette skyldes, at den lineære model i praksis må ses som en approksimation – en Taylor ekspansion om man vil – til en mere kompliceret men ukendt virkelig sammehæng; og den lineære approksimation bliver gradvis dårligere, jo mere man strækker gyldighedsområdet. Inden for X_i-området af interesse bør man heller ikke lægge alle målinger i intervalendepunkterne, da en afvigelse fra linearitet så ikke kan afsløres. En mere jævn fordeling er generelt at foretrække.

Antalsplanlægning

Antalsplanlægning med henblik på estimation af hældningen

(situation A) sker i samspil med punktplaceringen. For at opnå en standard-afvigelse på $\hat{\beta}_1$ mindre end $\sigma^0_{\beta_1}$ kræves ifølge (5.61) at

$$\sum_{i=1}^{n} (X_i-\bar{X})^2 > (\sigma/\sigma^0_{\beta_1})^2 \quad . \tag{5.62}$$

Det samme gør sig gældende i antalsplanlægningen, med henblik på hypotesetestning (situation B). Ønskes med sandsynlighed 1-β at hypotesen $H_0:\beta_1=\beta_1^0$ afvises på niveau α, hvis den sande værdi for β_1 er β_1^*, må

$$\sum_{i=1}^{n} (X_i-\bar{X})^2 > (z_{1-\beta}+z_{1-\alpha/2})^2 \left[\frac{\sigma}{\beta_1^0 - \beta_1^*}\right]^2 \quad , \tag{5.63}$$

hvor z-erne er de fraktiler i standardnormalfordelingen, som er angivet ved indexeringen.

Ved prediktion af ny Y-værdi for en given gerne udvalgt X-værdi X_{ny}, benyttes

$$\hat{Y}(X_{ny}) = b_0 + b_1X_{ny}. \tag{5.64}$$

Prediktionsfejl

Dette estimat er uden systematisk fejl, men behæftet med den stokastiske usikkerhed som skyldes at regressionslinien (5.64) er fastlagt fra et begrænset datamateriale.

$$\text{Var } \hat{Y}(X_{ny}) = \sigma^2\left[\frac{1}{n} + (X_{ny}-\bar{X})^2/\sum_{i=1}^{n}(X_i-\bar{X})^2\right] \quad . \tag{5.65}$$

Det fremgår af (5.65) at usikkerheden er mindst i midten omkring $\bar{X}$, og vokser kvadratisk med X_{ny}'s afstand herfra.

Foretages en ny måling Y_{ny} under betingelsen X_{ny}, vil denne jf. (5.58) være behæftet med en statistisk fejl e_{ny} med varians σ^2. Det betyder at Y_{ny}'s varians omkring $\hat{Y}$ er

$$\text{Var } Y_{ny}|X_{ny} = \sigma^2 + \text{Var } \hat{Y}(X_i) = \tag{5.66}$$

$$\sigma^2\left[1 + \frac{1}{n} + (X_{ny} - \bar{X})^2/\sum_{i=1}^{n} (X_i-\bar{X})^2\right] \quad .$$

hvor σ^2 estimeres ved

$$\hat{\sigma}^2 = \hat{\sigma}^2_Y(1 - r^2_{XY}), \tag{5.67}$$

svarende til (5.53), som altså er den nedre grænse for prediktionsvariansen.

Regressionsanalyse er et stort og veludviklet fagområde, med et væld af god litteratur, f.eks. Draper & Smith (1981), Neter *et al.* (1983) og Brook & Arnold (1985).

Eksempel 5.6

Vi forestiller os et planlagt forsøg gennemført hvor X_i er manipuleret på planlagt vis og de tilsvarende Y_i-værdier er målt. Vi forestiller os n = 83 observationerspar og i øvrigt samme talværdier som i eksempel 5.5, (specielt er $\Sigma(X-\bar{X})^2 = (83-1)\hat{\sigma}^2 = 93,22$).

Regressionslinien bliver $\hat{Y}(X) = \hat{\beta}_0 + \hat{\beta}_1 X$, med $\hat{\beta}_1 = 1,3824$ og $\hat{\beta}_0 = 35,14 - 1,3824 \times 11,41 = 19,3668$ identisk med prediktionslinien for Y fra X i eksempel 5.5. Specielt bliver $\hat{Y}(12) = 35,9556$. En ny Y-observation ved X=12 vil ifølge (5.66) have en varians omkring sin middelværdi på $Var(Y_{ny}|X=12) = \sigma^2(1 + 1/83 + (12 - 11,41)^2/93,22) = \sigma^2(1,0158) \simeq 5,1141(1 - 0,6518^2)\,1,0158 = 2,9414 \times 1,0158 = 2,9878 = 1,73^2$.

Hældningens usikkerhed estimeres fra (5.61) til $Var\hat{\beta}_1 = 2,9414/93,22 = 0,03155$, hvoraf et $100(1-\alpha)\%$ konfidensinterval for β_1 udregnes som $[1,3824 - z_{1-\alpha/2} \times 0,03155; 1,3824 + z_{1-\alpha/2} \times 0,03155] = [1,3206; 1,4442]$, hvor α er sat lig 5%, hvilket giver $z_{0,95} = 1,960$.

De fleste computerprogrammer vil producere en t-statistik for hypotesen $\beta_1 = 0$; p-værdien bliver i dette tilfælde praktisk taget lig nul ($p < 0,00001$), d.v.s. hældningen er stærkt signifikant.

5.7 Generaliseringer

Generel lineær model

Metoderne omtalt i afsnittene 5.1 til 5.6 har en række generaliseringer af stor anvendelighed i dataplanlægning og -analyse. De kan omfattes af den generelle lineære model i flere variable:

$$Y_i = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \cdots + \beta_k X_{i,k} + e_i \quad . \tag{5.68}$$

Faktorforsøg

X-erne kan være forskellige omstændigheder hvis indflydelse på Y ønskes belyst. Forsøgsplanen må nu indrettes så effekten af disse influerende omstændigheder, også kaldet faktorer, kan estimeres med bedst mulig præcision. Generelt er det uøkonomisk at undersøge én faktor ad gangen; og faktorforsøgsplaner findes, som angiver hvorledes flere variable med passende systematisk manipulation kan belyses i et enkelt evt. sekventielt opbygget forsøgsprogram. Metoder til planlægning og udførelse af faktorforsøg er velbeskrevet i Box *et al.* (1978) og Brøndum & Monrad (1979 og 1986).

Modelformen (5.68) omfatter også situationer hvor Y_i's afhængighed af en eller flere variable er ulineær, f.eks. kvadratisk. Måske kunne udbyttet Y af en proces afhænge af pH og temperatur på approximativt kvadratisk vis:

$$\begin{aligned} Y_i = {} & \beta_0 + \beta_1(\text{pH}) + \beta_2(\text{temp}) + \beta_3(\text{pH})^2 + \beta_4(\text{temp})^2 \\ & + \beta_5(\text{pH})(\text{temp}) + e_i \quad . \end{aligned} \tag{5.69}$$

Det ses, at (5.69) er en speciel version af (5.68) med $k = 5, X_1 = (\text{pH}), X_2 = (\text{temp}), X_3 = (\text{pH})^2$, $X_4 = (\text{temp})^2$ og $X_5 = (\text{pH})$

(temp). Statistiske metoder til forsøgsplanlægning og -analyse med henblik på beskrivelse og optimering af procesvariable (Y), betegnes i faglitteraturen som response surface metoder. Disse metoder findes velbeskrevet i Myers (1971) og Box & Draper (1987). Specielt med henblik på procesoptimering i et arbejdende anlæg giver Box & Draper (1969) nyttige anvisninger.

Response surface

Polynomial regression

En anden specialversion af (5.68) er den polynomiale regressionsmodel:

$$Y_i = \beta_0 + \beta_1 Z_i + \beta_2 Z_i^2 + \cdots + \beta_k Z_i^k + e_i. \qquad (5.70)$$

Denne modelform kan beskrive en ikke-lineær sammenhæng, men omfattes dog af de statistiske metoder for den generelle lineære model, da indtrykket (5.70) er lineært i parametrene $\beta_0, \beta_1, \cdots, \beta_k$. Den statistiske analyse af regressionsmodeller er velbeskrevet i Brook & Arnold (1985), Draper & Smith (1981), og Neter *et al.* (1983).

Balancerede blokforsøg

Der er også statistisk mulighed for at udnytte fordelene af en parret analyse som omtalt i afsnit 5.2 i situationer, hvor flere alternativer ønskes sammenlignet. Dette kan ske ved at udføre et balanceret blokforsøg, som analyseres ved hjælp af en tosidet variansanalyse. Denne og andre mere avancerede metoder er behandlet bl.a. i Hicks (1964), i Peng (1967), i Brøndom & Monrad (1979 & 1986) og i Box *et al.* (1978) Sidstnævnte er en særligt god generel reference til forsøgsplanlægning. Gutman *et al.* (1971) giver en mere bred introduktion til anvendt statistik, og som minilærebog kan henvises til Bowen & Weisberg (1980).

6. Konklusion og resumé

Behov for ny viden

Den hastige industrielle og samfundsmæssige udvikling bevirker, at erhvervelsen af nødvendig ny viden angående miljøforurening og ny teknologi har svært ved at holde trit med erkendte behov for vejledninger, indgreb, procesoptimering og indførelse af bedre teknologi.

Læreprocessen og statistik

Effektiv erhvervelse af information kræver statistisk planlagt dataindsamling og dataanalyse. Den empiriske læreproces går ud på fra eksisterende viden at opstille en matematisk model hvis ukendte konstanter, parametrene, estimeres på grundlag af data, hvorefter den således tilpassede model testes for acceptabel overensstemmelse med data. Herefter kan modellen modificeres, reestimeres, gentestes, remodificeres, o.s.v., indtil rimelig overensstemmelse mellem model og data er opnået. På dette grundlag kan nye data planlægges indsamlet, og den iterative analyse af det øgede datamateriale gennemføres, o.s.v. Den erhvervede viden udtrykkes knapt og overskueligt i den tilpassede model, dens form og parameterestimater.

Datagrundlag

Enhver undersøgelse må redegøre for sit datagrundlag, dets størrelse og kvalitet. Om data er egne eller andres, planlagte eller tilfældige, aktivt eller passivt indsamlede. Kun systematiske planer giver pålidelige data. Aktivt indsamlede data, d.v.s. med bevidst manipulerede omstændigheder, kan understøtte konklusioner vedrørende effekter af aktiv indgriben, medens passivt indsamlede data kun med sikkerhed kan belyse en fortsat passiv fremtid. Som hovedregel er data kun repræsentative for de bevidst varierede omstændigheder hvorunder indsamlingen er sket.

Trin i undersøgelser

En undersøgelses gennemførelse bør styres gennem følgende trin: a) Formulering af opgavens formål og omfang, dens angrebsvinkel, forsøgsplan og målemetoder, samt skønnede usikkerheder på resultater. b) Detailplanlægning af undersøgelsen efter igangsættelsen, herunder vurdering af behov for ekspertbistand. c) Gennemførelse af undersøgelsen med statuspunkter hvor kursændringer kan foretages i samarbejde med rekvirenten. d) Statistisk analyse og afrapportering, hvor undersøgelsens faktisk opnåede resultater sobert præsenteres i en selvforklarende og komplet form.

Kvalitativ dataanalyse

Den statistiske dataanalyse må bestå af en kvalitativ og en kvantitativ analyse. Den første kan betegnes grafisk dataanalyse, hvor data afbildes så informationsindholdet formidles på en klar og tankevækkende måde. Ofte er den grafiske dataanalyse en forudsætning for at den efterfølgende kvantitative dataanalyse bliver sikker og målrettet, samt at dens konklusioner kan blive præsenteret forståeligt.

Kvantitativ dataanalyse

Den kvantitative dataanalyse består hovedsageligt af parameterestimation og signifikanstestning. Parameterestimationen er væsentlig, da de spørgsmål og uklarheder undersøgelsen skal besvare i den statistiske model udtrykkes som parametre, det vil sige ukendte

Estimation

konstanter. Ved estimation bestemmes parametrene fra data, inklusive angivelse af den tilbageværende usikkerhed.

Test

Signifikanstestning handler om kvantitativ vurdering af en foreslået hypoteses plausibilitet set i lyset af data. Beregningerne forbun-

det med statistisk analyse foretages let ved hjælp af passende computerprogrammer, resultaternes fortolkning derimod forbliver dataanalytikerens ansvar.

Antal forsøg

Valg af passende antal forsøg i en undersøgelse bør foretages bevidst ud fra vurdering af datas forventede statistiske usikkerhed i sammenhæng med undersøgelsens ønskede følsomhed. Denne kan udtrykkes enten ved præcision af et parameterestimat eller sandsynlighed for forkastelse af en given ukorrekt hypotese.

Undersøgelsestyper

Seks undersøgelsestyper er belyst: 1) middelværdi og varians, 2) forskelle mellem to middelværdier, 3) forskel mellem én middelværdi og flere alternativer, 4) forskel mellem flere sideordnede middelværdier, 5)korrelation og 6) regression. Mange andre undersøgelsestyper er af betydelig praktisk interesse, disse er bl.a. beskrevet i den refererede litteratur, men deres effektive udnyttelse kan kræve medvirken af en fagstatistiker.

Der er hermed søgt redegjort for de generelle principper, begreber og retningslinier for dataplanlægning og dataanalyse, som sunde undersøgelser må respektere, specielt i miljøstudier over forurenende, industrielle anlæg.

Appendix A:

Disposition af rapport over undersøgelse.

Der findes en stor speciallitteratur om udfærdigelse af tekniske rapporter, bl.a. Cooper (1964), som også giver mange yderligere referencer. Desuden kan henvises til Bowen & Weisberg (1980) kapitel 13, og til Ehrenberg (1982) som giver mange gode råd.

I nærværende appendix skal gives et konkret forslag til disposition af en teknisk rapport, samt visse stilistiske hovedregler.

Den endelige udformning af en given rapport vil naturligvis afhænge af dens indhold og karakter, samt af forfatterens evner og intentioner som stilist. Nedenstående rapportdisposition tilsikrer en vis naturlig orden og klarhed, der ikke så sjældent savnes i dokumentationen af tekniske undersøgelser:

Resumé
Indledning
Metoder
Data
Kvalitativ dataanalyse
Kvantitativ dataanalyse
Diskussion
Konklusion
Appendices
Referenceliste

Denne disposition skal opfattes som et muligt udgangspunkt hvorfra man kan og ofte bør afvige.

Abstract

Resumeet, også synonymt kaldet »abstract«, skal være et meget kort sammendrag af rapportens indhold og konklusioner. De vigtigste konklusioner skal fremhæves, der må ikke gemmes »spændende overraskelser« til selve rapporten. Resumeet bør normalt være så kort, at det let kan stå på en enkelt (A4-) side sammen med rapporttitel, forfatter, udgivelsestidspunkt m.v. Resten af rapporten udformes som om resumeet ikke eksisterede, det vil bl.a. sige, at der ikke må henvises til resumeet. Omvendt bør resumeet heller ikke referere til specifikke afsnit, formler eller sider i rapporten.

Executive summary

Foruden dette helt korte resumé skrives undertiden også et udvidet resumé, tit benævnt »executive summary«, som i meget kort form gennemgår det væsentlige i hele rapporten med tilstrækkelig detaljering af resultater og konklusioner, så undersøgelsens essentielle informationsindhold kan formidles til læseren. Henvisninger til selve rapporten kan her være på sin plads, så læseren let kan lokalisere de afsnit i rapporten som ønskes detaljeret og/eller underbygget yderligere. Selve rapporten fra indledning og fremefter må heller ikke påvirkes af det udvidede resumé; der er intet forkert i at læseren i indledningen efter at have læst resumé og udvidet resumé måske for tredje gang støder på samme udtryk, synspunkt eller beskrivelse.

Indledning

Indledningen eller introduktionen indeholder beskrivelsen af undersøgelsens formål, dens udgangspunkt og angrebsvinkel, samt normalt et litteraturstudium, d.v.s. en kort gennemgang af tidligere arbejder inden for området med referencer til disses dokumentation.

Metoder

Afsnittet »metoder« indeholder teoretiske og/eller praktiske værktøjer som er benyttet i undersøgelsen, specielt hvor disse afviger fra tidligere tilsvarende undersøgelsers. For eksempel ny målestrategi eller nyt måleudstyr, særlig statistisk analysemetode, modificeret teoretisk model for årsags-virkning sammenhæng, eller andet. Meget lange eller specialiserede beskrivelser bør dog lægges i appendix sidst i rapporten; og i afsnittet anføres så blot det for kontinuiteten nødvendige plus en henvisning til dette appendix.

Data

I det påfølgende anførte afsnit »data«, gives foruden en listning af undersøgelsens datamateriale, en detaljeret redegørelse for datas tilblivelse: forsøgsomstændigheder, tider, steder, målemetoder, enheder, etc. Er datamaterialet meget omfattende kan hele dette afsnit eller dele af det lægges i appendix. Under alle omstændigheder bør materialet fremlægges så detaljeret og komplet, at enhver kan reproducere rapportens udregnede resultater, og det skal helst også muliggøre en alternativ analyse, som andre ideer eller en senere udvikling kunne inspirere til.

Kvalitativ dataanlyse

Afsnittet »kvalitativ dataanalyse« vil indeholde relevanta plot af data. Denne grafiske analyse kommenteres, vigtige kvalitative aspekter fremhæves, og den påfølgende kvantitative analyse motiveres. Indebærer den grafiske analyse et meget stort antal plot, hvoraf kun nogle er egentlig afslørende for datas informationsindhold, kan øvrige plot gengives i appendix.

Kvantitativ dataanalyse

Den kvantitative statistiske analyse gennemgås herefter. Så vidt muligt bør man stræbe efter en neutral præsentation af dataanalysen, gerne med en vis overflod af resultater. En fyldig dataanalyse giver mulighed for at alternative synspunkter end forfatterens kan belyses af den foretagne undersøgelse. Dele af en meget omfattende analyse bør overvejes placeret i appendix.

Diskussion

Herefter følger det diskuterende afsnit, hvor forfatteren fremfører de fortolkningsmuligheder, som man under arbejdet er nået frem til. Det forklares hvorledes de stemmer overens med andre resultater eller tidligere formodninger, der afstikkes grænser for rimelige fortolkningsmuligheder, og disse sættes i perspektiv over for forureningsmæssige, økonomiske eller andre relevante konsekvenser. Har en sekventiel forsøgsstrategi været anvendt forklares det, hvorfor man undervejs foretog de aktuelle valg; og i det hele taget redegøres for de vurderinger, skøn og beslutninger, som har påvirket undersøgelsens konkrete udførelse. Ligeledes redegøres kort for de dele af undersøgelsen, hvor arbejdet ikke gav resultater, for spor der blev forfulgt men opgivet, for ideer der uden succes blev afprøvet. Alt dette arbejde kan meget vel repræsentere en overvældende del af hele undersøgelsens indsats, men bør kun omtales kort i selve rapporten. Skønner man behov for en fyldigere omtale må dette ske i særligt appendix. Rapporten skal omhandle undersøgelsens resultater ikke være en logbog over arbejdets forløb.

Konklusion

Herefter følger rapportens konklusion. Den opsummerer undersøgelsens resultater herunder fortolkningsmuligheder og mulige handlingskonsekvenser. Længere subjektive vurderinger og spekulationer bør henlægges til afsnittet »diskussion«. Endelig bør ubesvarede spørgsmål trækkes frem, herunder nye probemstillinger som undersøgelsen rejser. Konklusionen bør skrives ret kort, med mindre man ønsker her at give et resumé af hele undersøgelsen. Dette kan være berettiget hvis et »udvidet resumé« ikke findes.

Appendices, referencer og bilag

Endelig kommer appendices samt en referenceliste, evt. kommenteret. Undertiden knyttes bilag til en rapport; ønsker man at skelne mellem appendices og bilag er de sidste mere perifere. Appendices er fuldgyldige dele af rapporten, medens bilag indeholder baggrundsmateriale, referencer o.lign.

Brug af appendix

Om materiale anbringes i appendix eller i selve rapporten er et spørgsmål om at gøre denne læservenlig; materialets vigtighed er ikke et kriterium. Som hovedregel bør alt det materiale lægges i appendix, som ikke er nødvendigt for rapportens kontinuitet. Herunder også svært tilgængelige og specialiserede rapportdele, som kan svække opmærksomheden omkring rapportens hovedlinie.

I en konkret undersøgelses rapportering vil et eller flere af de foreslåede afsnit kunne undlades, måske fordi de bliver så små at inkorporering i andre føles mere naturlig. Særlige yderligere afsnit kan der selvsagt også byde sig anledning til. Undlades afsnittet »diskussion«, som klart indeholder forfatterens tanker og vurderinger, må disse, optrædende i andre mere objektive afsnit, entydigt fremstå som subjektive. Dette kan gøres med stor klarhed og give en fin kompakt stil; men mindre sikre forfatterens forsyndelser på dette punkt er almindelige, hvilket skaber forvirring eller ligefrem vildleder.

Subjektivitet

Rapporten skrives kun helt undtagelsesvist i læserækkefølgen. Ofte kan appendix skrives først, dernæst selve rapporten med indledningen til sidst, og resumé (plus evt. udvidet resumé) til allersidst. Man må normalt være indstillet på at skrive en rapport igennem mange gange, herunder slette og/eller omarrangere rækkefølgen af længere tekstdele. Selv meget rutinerede tekniske skribenter klarer sig sjældent med mindre end 3-4 gennemskrivninger af alle afsnit i en rapport. Det er ligeledes en almindelig erfaring at hver gennemskrivning giver anledning til at mere materiale/flere resultater/bedre forståelse bliver inkorporeret på et færre antal sider, som så oven i købet bliver mere lettilgængelige.

Skriveteknik

Mange erfarne tekniske skribenter benytter en skriveteknik, hvor alle tabeller og figurer udformes først, undertiden blot som råskitser, disse sættes op i et rapportskelet med afsnitsoverskrifter, underoverskrifter og stikord. Den første redisponering af rapporten kan nu ske inden ret megen tekst er skrevet. Herefter indføjes tekstafsnittene, i en rækkefølge der ikke nødvendigvis er læserækkefølgen.

Andre forfattere skriver (dikterer) blot løs, uden nøje tanke for den endelige disponering af rapporten. Teksten »klippes og klistres« herefter, skrives igennem igen, o.s.v. Få stræber efter at frembringe

et acceptabelt manuskript i første omgang. Det har vist sig en alt for hæmmende og tidsforbrugende procedure.

Stil

Det klæder tekniske rapporter at stil og sætningsopbygning er enkel, uden kunstfærdige konstruktioner, særprægede ordvalg og slangagtige vendinger. Fyldeord og gentagelser bør luges ud med hård hånd; kun overfladisk kan de bidrage til en flydende stil, de trætter, sinker og kan irritere læseren. Stave- og banale regnefejl bør undgås; de undergraver rapportens troværdighed mere generelt.

Appendix B: Ordforklaring

I dette appendix gives en kort forklaring på begreber, ord og fagudtryk, der har speciel betydning i dataplanlægning og dataanalyse. De tilsvarende engelske ord er angivet i kantet parentes, de optræder dels i faglitteratur om emnet, dels benyttes de typisk i standard computerprogrammer.

Antalsplanlægning[sizing (of experiment)], planlægning af hvor mange målinger et måleprogram skal indeholde, så det kan belyse sin problemstilling tilstrækkeligt konklusivt.

Blok [block], mindre del af et totalt måleprogram. Måleomstændighederne inden for blokke er mere ens end mellem blokke, d.v.s. end for måleprogrammet totalt set.

Deduktion [deduction], slutning fra det generelle til det specielle, fra model til data, fra terori til praksis, omfatter matematiske ræsonnementer.

Empirisk [empirical], erfaringsmæssigt, d.v.s. på basis af praktiske observationer, modsat teoretiske overvejelser.

Estimation [estimation], bestemmelse af parameters værdi ud fra data.

Fejl [error], 1) måleusikkerhed eller naturlig variation som betyder at gentagne observationer ikke giver helt samme måleværdi. 2) Fejl af type I(II), den fejl der begås ved at forkaste en sand hypotese (acceptere en falsk hypotese).

Forsøgsplan [(experimental) design], plan over den systematiske variation i omstændighederne hvorunder et måleprogram skal gennemføres.

Forventningsværdi [expected value] synonymt med middelværdi.

Gennemsnit [average], summen af et antal målinger divideret med deres antal.

Hypotesetest [test of hypothesis], se test.

Induktion [induction], slutning fra det specielle til det generelle, fra data til model, fra praksis til terori, omfatter statistiske ræsonne menter.

Inferens [inference], læren om hvorledes man drager statistiske slutninger; omfatter estimation og hypotesetest.

Konfidens(grad) [(degree of) confidence], kvantitativt mål for troværdigheden af at en given parameter ligger i et opgivet interval, konfidensintervallet.

Konfidensinterval [Confidence interval], interval som med given konfidens(grad) indeholder den sande værdi af en parameter.

Korrelation(skoefficient) [Correlation (coefficient)], kvantitativ mål for samvariation mellem to variable. Målet af en dimensionsløs størrelse, d.v.s. det afhænger ikke af måleskala.

Middelværdi [mean], teoretisk gennemsnit af alle målinger som kunne tænkes foretaget af en variabel, f.eks. en målestørrelse, eller en funktion af en eller flere målestørrelser.

Model [model], matematisk udtryk der beskriver en sammenhæng af

interesse. Typisk indeholder modellen ukendte, justerbare parametre.

Niveau [level],1) Udtryk for den generelle størrelse(sorden) af målinger, en middelværdi kan benyttes som mål for niveau. 2) Ved signifikanstest er niveauet, kaldet α, lig chancen for at begå fejl af type I.

Nulhypotese [Null hypothesis], i test en hypotese H_0, der får fordel af al tvivl.

p-værdi [p value], ved signifikanstest er p-værdien grænsen mellem de niveauer α, der ville føre til accept, og de der ville føre til forkastelse af nulhypotesen H_0.

Parameter [parameter], ukendt konstant som indgår i statistisk model.

Prediktor [predictor], variabel som benyttes til forudsigelse af en anden variabel, typisk i regressionsmodeller.

Regression [regression], model der knytter sammenhængen fra en eller flere prediktorvariable til en variabel afhængig af disse.

Sandsynlighed [probability], den statistiske betydning af ordet svarer godt til den dagligdags betydning; undertiden dog indsnævret til situationer, hvor man kan tale om relativ hyppighed af hændelser.

Signifikans [significance], i test tales om signifikant afvigelse fra nulhypotesen når denne må forkastes. Et indgreb eller en ændring er signifikant eller viser signifikans, når en hypotese om ingen effekt forkastes ved test. Signifikansens niveau er lig det niveau, hvorpå testen bliver udført.

Signifikanstest [significance test], synonymt med test.

Spredning [spread], bruges normalt synonymt med standardafvigelse.

Standardafvigelse [standard deviation] kvadratroden af variansen.

Standardfejl [standard error], kan benyttes synonymt med standardafvigelse.

Statistik [1) statistics; 2)statistic], 1) læren om planlægning og fortolkning af fejlbehæftede data; 2) størrelse udregnet på grundlag af data, normalt for at kondensere datas informationsindhold. F.eks. er et gennemsnit $\bar{Y}$ en statistik.

Stokastisk [stochastic], varierende uforudsigeligt efter en kendt eller ukendt fordelingslov.

Støj [noise], synonymt med fejl, forklaring 1).

Test [test], statistisk teknik hvorved plausibiliteten af en nulhypotese afprøves i lyset af data. Se også nulhypotese, niveau og p-værdi.

Transformation [transformation], ændring af målingers talværdier ifølge en regneforskrift, med henblik på anvendelse af de transformerede værdier i den statistiske analyse. Regneforskriften kan f.eks. være logaritmen eller reciprokværdien.

Type I(II) fejl [type I(II) error], se fejl, forklaring 2).

Uafhængig [independent], 1) uafhængig variabel, er synonymt med prediktor; 2) (statistisk) uafhængig betyder at den statistiske fejl, hvormed to eller flere variable er behæftede, intet har med hinanden at gøre, at de ikke har nogen fælles komponent.

Varians [*variance*], *kvantitativt mål for variationen af stokastiske* variable, f.eks. målinger eller statistikker. Variansen er den forventede værdi af variablens kvadrerede afvigelse fra sin middelværdi.

Referencer

Beyer, William H. (1968): »Handbook and Tables for Probability and Statistics«, *2nd ed.*, The Chemical Rubber Co., Cleveland, Ohio, U.S.A.

(Omfattende opslagsværk med bl.a. tabeller og forsøgsplaner, men næsten uden vejledning.)

Booman, K.A.; Pallesen, L. & Berthouex, P.M. (1987): »Intervention Analysis to Estimate Phosphorus Loading Shifts«, Systems Analysis in Water Quality Management, M.B. Beck editor, Pergamon Press, Oxford, England.

(Artikel med avanceret statistisk analyse af effekten fra lovgivningsmæssigt miljøindgreb.)

Bowen, Bruce D. & Weisberg, Herbert F. (1980): »An Introduction to Data Analysis«. W.H. Freeman & Comp., San Francisco.

(Lettilgængelig minilærebog i anvendt statistik.)

Box, G.E.P. & Draper, N.R. (1987): »Empirical Model-Building and Response Surfaces«, John Wiley, New York.

(Meget komplet og detaljeret gennemgang af empirisk modelbygning.)

Box, G.E.P. & Draper, N.R. (1969): »Evolutionary Operation«, John Wiley, New York.

(Lettilgængelig indføring i statistisk metode til løbende forbedring af arbejdende produktionsproces.)

Box, G.E.P.; Hunter, W.G. & Hunter, J.S. (1978): »Statistics for Experimenters«, John Wiley, New York.

(Fremragende generel lærebog i forsøgsplanlægning og dataanalyse; velegnet til selvstudium.)

Brook, Richard J. & Arnold, Gregory C. (1985): »Applied Regression Analysis and Experimental Design«, Marcel Dekker Inc., New York.

(En blandt mange lærebøger i anvendt statistik med henblik på forsøgsplanlægning og analyse.)

Brøndum, L. & Monrad, J.D. (1979): »Statistisk forsøgsplanlægning I«, 2. rev. udg., Den private Ingeniørfond, København.

(Første del af omfattende dansk lærebog i statistisk forsøgsplanlægning.)

Brøndum, L. & Monrad, J.D. (1986): »Statistisk forsøgsplanlægning II«, Den Private Ingeniørfond, København.

(Anden del af omfattende dansk lærebog i statistisk forsøgsplanlægning.)

Chambers, John M.; Cleveland, Williams S; Kleiner, Beat & Tukey, Paul A. (1983): »Graphical Methods for Data Analysis«, Wadsworth International Group, Belmond, California, and Duxbury Press, Boston.

(God, rimeligt kort, og overskuelig gennemgang af grafisk dataanalyse.)

Conradsen, Knut (1984): »En Introduktion til Statistik«, 5. udgave, IMSOR, DTH, Lyngby.

(Glimrende dansk lærebog i statistik, god til opslag, men vanskelig til selvstudium.)

Cooper, Bruce M. (1984): »Writing Technical Reports«, Penguin Books.

(Udmærket billigbog med råd og vejledning om udformning af tekniske rapporter.)

Draper, N.R. and Smith, H. (1981): »Applied Regression Analysis«, 2nd ed., John Wiley, New York.

(Omfattende bog om statistisk dataanalyse med regressionsmetoder.)

Ehrenberg, A.S.C. (1982): »Writing Technical Papers or Reports«, The American Statistician, 36, pp 326-329.

(God koncis vejledning i hvorledes tekniske rapporter bør skrives.)

Guttman, Irwin; Wilks, S.S. & Hunter, J. Stuart (1971): »Introductory Engineering Statistics«, 2nd ed., John Wiley, New York.

(God generel lærebog i anvendt statistik.)

Hicks, C.R. (1964): »Fundamental Concepts in the Design of Experiments«, Holt, Rinehart and Winston, New York.

(Lærebog i forsøgsplanlægning som dækker dataanalysen af mange mere komplicerede planer. Giver dog kun lidt vejledning i opstilling af egnede planer.)

Hogg, Robert V. & Craig, Allen T. (1970): »Introduction to Mathematical Statistics«, 3rd ed., The Macmillan Company, London.

(Forholdsvis let tilgængelig lærebog i teoretisk statistik.)

Hunter, W.G. (1981), personlig kommunikation.

Miljøstyrelsen (1986): »Udviklingsprogram for renere teknologi (1986-1989)«, september 1986.

(Programoplæg som har udløst nærværende undersøgelse.)

Moses, Lincoln E. (1986): »Think and Explain with Statistics«, Addison-Wesley Publishing Company, Reading, Massachusetts, U.S.A.

(En bredt skrevet introduktion til statistisk tankemåde og analyse.)

Myers, Raymond H. (1971): »Response Surface Methodology«, Allyn and Bacon Inc., Boston.
(Kort og god gennemgang af forsøgsplanlægning og analyse til empirisk modelbygning.)

Neter, John; Wasserman, William & Kutner, Michael H. (1983): »Applied Linear Regression Models«, Richard D. Irwin Inc., Homewood, Illinois, U.S.A.
(God og omfattende lærebog om regressionsanalyse.)

Owen, D.B. (1962): »Handbook of Statistical Tables«, Addison-Wesley, Reading, Massachusetts, U.S.A.
(Tabelsamling uden nævneværdig vejledning.)

Pallesen, Lars (1987): »Statistical Assessment of PCDD and PCDF Emission Data«, Waste Management & Research, 5, pp 367-379.
(Eksempel på større statistisk analyse af et datasæt fra miljøstudium.)

Pallesen, Lars; Berthouex, P.M. & Booman, Keith (1985): »Environmental Intervention Analysis: Wisconsin's Ban on Phosphate Detergents«, Water Research, 19, pp 353-362.
(Statistisk effektestimation af miljøindgreb.)

Peng, K.C. (1967): »The Design and Analysis of Scientific Experiments«, Addison-Wesley Publishing Company, Reading, Massachusetts, U.S.A.
(Lærebog i forsøgsplanlægning, omtaler mange typer planer, men giver ikke større vejledning i hvornår hvilke skal benyttes.)

Rao, C.R. (1973): »Linear Statistical Inference and its Applications«, 2nd ed., John Wiley, New York.
(Lærebog i teoretisk statistik, autoritativ men ret svært tilgængelig.)

duToit, S.H.C.; Steyn, A.G.W. & Stumpf, R.H. (1986): »Graphical Exploratory Data Analysis«, Springer-Verlag, New York.
(God gennemgang af grafisk dataanalyse.)

Tufte, Edward R. (1983): »The Visual Display of Quantitative Information«, Graphics Press, Cheshire, Connecticut, U.S.A.
(Fremragende bog om grafisk præsentation af data.)

Tukey, John W. (1977): »Exploratory Data Analysis«, Addison-Wesley Publishing Company, Reading, Massachusetts, 1977.
(Bredt skrevet bog om kvalitativ dataanalyse.)

Registreringsblad

Udgiver: Miljøstyrelsen, Strandgade 29, 1401 København K.

Serietitel, nr.: Orientering fra Miljøstyrelsen, nr. 7 1989

Udgivelsesår: 1989

Titel: Industrielle forureningsundersøgelser

Undertitel: Planlægning og analyse af data

Engelsk titel: Planning and Analysis of Data in Studies of Industrial Processes towards Environmental Improvements

Forfatter(e) og/eller udførende institution(er):
Lars Pallesen

Resumé:
Indførelsen af miljømæssigt og økonomisk bedst tilgængelig teknologi forudsætter ofte, at ny viden erhverves målrettet fra data. Dette kræver statistisk dataplanlægning og -analyse. Rapporten redegør for de generelle principper, begreber og retningslinier, som undersøgelser må respektere: Den trinvise gennemførelse af undersøgelser, datagrundlagets kvalitet og kvantitet, den iterative dataanalyse, den korrekte fortolkning af statistiske nøglebegreber, specielt hvad angår parameterestimation og hypotesetestning, samspillet mellem udførelsen af kvalitativ og kvantitativ dataanalyse samt betydningen af sober rapportering.

Standardiserede emneord (efter MDS-liste):
miljøteknik, miljøvenlig teknologi, statistik, økonomisk planlægning

Frie emneord:
parameterestimation, hypotesetestning

ISBN: 87-503-7469-9

ISSN: 0107-2722

Pris: 90,- **(inkl. 22% moms)**

Format: AS5

Sideantal: 88

Md./år for redaktionens afslutning: juni 1988

Oplag: 2.000

Andre oplysninger:

Tryk: Grafisk Service, Vejdirektoratet, København

Trykt på miljøpapir