MATEMATISK STATISTIKK

MOMENTER I EN FORELESNINGSSERIE FOR

SENSORISK STUDIEGRUPPE

VÅREN 1977

ved

KJELL STEINSHOLT

FORORD

Dette korte kurset i statistikk, er selvsagt ikke på noen måte uttømmende. Det er ment å gi informasjon om en del begrep som vi bruker i det daglige arbeidet med sensorisk analyse. Kanskje er det ikke pedagogisk riktig, men rimeligvis nødvendig at det legges en del vekt på å vise hvor forsiktig vi må være i tolkningen av de resultatene vi kommer fram til. Jeg ser det også som vesentlig at det gis innføring i utregning av en del parametre, som vi kanskje i dag kan få f.eks. ved enkle lommekalkulatorer.

Vi bruker også daglig en rekke tabeller for å finne ut hvorvidt resultater som vi beregner kan kalles for **signifikante** eller ei. En viss innføring i beregningsgrunnlaget for disse tabellene og for de statistiske størrelsene som testes, vil vel rimeligvis være av en viss nytte og kunne gi større tillit (eller mistillit) til det vi finner.

Utvikling av en rekke formler vil - la oss si av tidshensyn - bli utelatt.

En del av deltakerne vil trolig være svært interessert i forsøksopplegg, men i første rekke tror jeg det vil være fornuftig bare perifert å berøre problemer innenfor dette veldige området. Muligens kan det seinere bli tid til noen diskusjonstimer om hvordan det materialet vi skal bedømme best kan skaffes til veie. Mange i gruppa er vel primært interessert i selve bedømmelsesmetodene, **utregningen av** resultatene fra bedømmelser og hvordan disse kan tolkes. Dette kollokviet vil da bli ledet med dette som hovedmålsetning.

OM FORSØK OG FORSØKSPLANLEGGING

Den tankevirksomheten som fører fra observasjoner (i vårt tilfelle bedømmelser) til et eller annet utsagn, kaller vi induksjon. Er vi svært heldige eller dyktige, kan en ved induksjon komme fram til en generell regel, en naturlov. Det kan hevdes at en utforming av utsagn eller regler på basis av observasjoner, er forskernes sak. Men de som skal bruke resultatene bør også ha en god peiling på den måten en går fram på ved induksjonen, slik at de med kritisk sans kan studere de informasjonene de får forelagt.

Det er da tre ting som leseren av rapporter o.l. har et rimelig krav på å få vite, og det er:

1. Hvor stammer observasjonene fra
2. Hvordan har en skaffet til veie informasjonene
3. Hvordan er de behandlet som grunnlag for et gitt utsagn.

Det er i prinsippet to forskjellige måter å samle opplysninger på. Den ene måten er å lage et systematisk eksperiment (f.eks. ved å bruke forskjellig stabilisatormengder i is-krem, og bedømme kvaliteten). Den andre er å gå ut i naturen og observere data som allerede er der (f.eks. telling av antall kronblader på en prestekrave). Forskjellen i de to oppleggene kan da også beskrives som en forskjell i graden av den kontrollen vi har over forsøket. I det systematiske forsøket kan vi holde i hvert fall en del av "bakgrunns-støy" i forsøket borte, men dette slår til med full tyngde i det ikke-eksperimentelle forsøket. Selv om vi bruker de samme metodene for induksjonen, vil det være enklest å komme fram til et relativt holdbart utsagn for det systematiske forsøket.

Nå er det dessverre slik at en selv i de best organiserte laboratorie-eksperimenter får noe forskjellige resultater når vi gjentar et forsøk ved det vi kan kalle "identiske samme forsøksbetingelser". En del effekter, bl.a. målefeil,

kan vi ikke kontrollere. Vi vil da alltid få en variasjon som vi må ta hensyn til i vår induksjon. Særlig store variasjoner må vi vente innen sensorikken, hvor det er menneskelige måleapparater med alle deres muligheter for variasjon både mellom individer og innenfor samme individ ved forskjellig tidspunkt. Bakgrunnsstøyen kan i mange tilfelle fullstendig overdøve den musikken som en kunne ha glede og nytte av.

GJENTAK OG UNIVERS

Innenfor statistikken kan vi si at vi har kunnskap av to slag. Vi har observasjon(er) fra et enkelt forsøk og disse observasjonene kan vi si vi vet med sikkerhet selv om de kan ha sine mangler. Har vi målt pH i en ost til 5,2, så er dette en faktisk opplysning. Vi kan også ha kunnskaper som vi har laget oss på grunnlag av en rekke enkelt-observasjoner. Her bør en imidlertid alltid være usikker fordi det ikke kan gis noe bevis for riktigheten av kunnskap som vi har skaffet oss empirisk. Vi kan imidlertid gjøre denne usikkerheten svært liten ved riktig valg av forsøksbetingelser som bl.a. kan bestå i å ha et tilstrekkelig antall enkeltresultat som grunnlag for de konklusjonene vi har dratt. Hvert enkelt resultat kalles da et gjentak, og alle mulige gjentak kaller vi for et univers.

Det kan ofte være vanskelig å få et klart bilde av det universet vi egentlig tenker å slutte noe om. Ofte er et slikt univers ubegrenset. Et forsøk på laboratoriet vil være et gjentak av et univers som består av uendelig mange tenkte gjentak. En smaksbedømmelse foretatt av tre personer som f.eks. er ment å være representativt for Norges befolkning, vil være et univers av alle de kombinasjoner av tre personer vi kan få laget av befolkningen, og dette blir et stort tall, men er likevel endelig.

Et univers kan også deles opp i sub-universer. F.eks. kan barn bety alle barn i verden, setter vi nasjonalitet foran, har vi foretatt en innskrenkning, og setter vi at barna skal være fra et enkelt sted, har vi foretatt enda en innsnevring. Det vil da være lettere på grunnlag av observasjoner å kunne si noe med en bestemt sikkerhet, f.eks. om barnas sunnhetstilstand i det lille universet. Vi har imidlertid tapt på almengyldigheten av vårt utsagn.

Det er vel også rimelig at utsagnet vårt blir sikrere desto fler gjentak vi har fra et bestemt univers. Med et sampel eller et utvalg forstår vi da de antall gjentak som vi har til rådighet for observasjoner.

LITT OM REGNING MED SUMMETEGN

Vi tenker oss at vi har ei rekke med n tall karakterisert ved en størrelse X. Denne rekka kan da beskrives som

$$X_1, X_2, \text{---------} X_i, \text{----------} X_n$$

Hvis vi har bruk for summen av de enkelte leddene, bruker vi å skrive denne slik:

$$\Sigma X_i \quad \text{eller bare } \Sigma X$$

som betyr at vi skal sette inn alle tall fra 1 til n for i og summere. På samme måten kan vi skrive summen av en funksjon av X, f.eks.

$$\Sigma (X_i)^2$$

som betyr at vi skal suksesivt sette i lik alle tall fra 1 til n, kvadrere for hver gang og så summere alle kvadratene.

En fullstendig angivelse bør inkludere nedre og øvre verdi for i og j plassert under og over summetegnet. Dette er utelatt for å forenkle skrivingen.

Hvis vi nå tenker oss at vi har flere rekker med X-er hvor hver rekke består av n tall, og at tallene er ordnet under hverandre i m antall rekker, så kan summen skrives:

$$\Sigma \Sigma X_{ij}$$

hvor da j er det alminnelige leddet av rekker. Uttrykket betyr at vi først setter j = 1): vi betrakter første rekke og summerer alle tall i rekka fra 1 til n. Så går vi over i neste rekke og lar j være 2 mens vi summerer fra 1 til n i denne rekka, osv. Dette kan da skrives slik:

$$\Sigma\ \Sigma\ X_{ij} = X_{11}+X_{21}\text{ ------}+X_{i1}\text{---}+X_{n1}+X_{12}+X_{22}+\text{--}X_{i2}+\text{--}X_{n2}$$

$$+X_{ij}+\text{---}X_{nm}$$

Det sier seg da selv at

$$\Sigma\ \Sigma X_{ij} = \Sigma\ \Sigma X_{ji}$$

fordi faktorenes orden er likegyldig i summering.

Vi kan utvide modellen til å være X-er eller faktorer karakterisert ved X-er, plassert i ei kasse. X-ene kan da karakteriseres ved nummer for kollonne, rekke og plass bakover i kassa. Summen av alle X-ene blir da:

$$\Sigma\ \Sigma\ \Sigma\ X_{ijk}$$

hvor vi også kan bytte indeksene uten å endre summen.

Vi kan utvide antall indekser så mye vi vil, og summetegnene blir brukt på samme måten. Videre vil de reglene som vi setter opp for et summetegn også ha gyldighet om det er flere tegn med i bildet.

Når alle X er like, så vil

$$\Sigma X = (X + X\text{ -------}X) = nX$$

Har vi flere kjennetegn (X,Y og Z) ved elementene (hvert ledd i rekka, i kollonnene og i kassa kalles et element) og disse skal summeres, vil

$$\Sigma\ (X_i+Y_2+Z_i) = X_1+Y_1+Z_1+X_2+Y_2+Z_2+\text{------}$$

$$X_i+Y_i+Z_i+\text{-----}X_n+Y_n+Z_n = \Sigma X_i + \Sigma Y_i + \Sigma Z_i$$

Hvis vi har en konstant under summetegnet kan denne settes utenfor summetegnet

$$\Sigma aX_i = aX_1 + aX_2 + ----aX_i + aX_n = a\Sigma X_i$$

FUNKSJONER

Foretar vi en omregning av en størrelse X etter en bestemt oppskrift slik at det for hver X bare fremkommer en ny størrelse Y, kaller vi den nye størrelsen for en funksjon av X og betegner den f(X).

For eksempel kan vi sette

$$Y = f(X) = X^2$$

idet vi her for hver X skal kvadrere denne og får da ett, og bare ett tall Y.

OBSERVASJONER

Vi tenker oss at vi plukker et sampel epler fra et epletre og bestemmer f.eks. vekt, vitamininnhold, protein-, sukker- og askeinnhold i hvert av de n eplene i samplet. Vi vil da oppdage at det er en viss variasjon mellom eplene. Ingen av våre målemetoder er helt nøyaktige slik at vi der har en grunn til variasjoner, men vi må også gå ut fra at det er en reell variasjon mellom eplene. Hvis hensikten var å finne f.eks. vekten av epler på treet eller vekten av epler generelt, vil observasjonene våre bare være tilnærmete tall. De riktige tallene kaller vi verdien av en random (eller tilfeldig) variabel, og observasjonene kan bare danne grunnlag for estimering av denne verdien.

Andre eksempler på random variable kan være antall kronblader på prestekraver, det kan være antall barn i familien, det kan være fettinnhold i melk, osv.

Noen random variable kan bare ha bestemte tallverdier s.s. barn i en familie, antall kronblader osv. Disse kaller vi diskrete random variable. Andre derimot kan tenkes å ha alle mulige verdier mellom en øvre og nedre grense. Disse kaller vi kontinuerlige random variable.

Hvis vi skal finne vekta på et bestemt eple, så observerer vi egentlig ikke en random variabel, men en enkelt størrelse. Målefeil gir også en viss variasjon i slike observasjoner hvis eplet veies mange ganger og vekta er nøyaktig nok.

Nå kan også eplene karakteriseres ved f.eks. fargen; mennesker kan karakteriseres ved øyefarge, kjønn, fysisk kondisjon osv. Slike karakteristikker kaller vi for konstante kjennetegn. Konstante kjennetegn kalles også for kvalitative i motsetning til de random variable som ofte betegnes for kvantitative.

Hvis vi imidlertid teller opp antall epler som er grønne, gule og røde, så er de absolutte eller relative tallene (ofte gitt i prosent) random variable.

FREKVENSFORDELING

Det er en stående regel i utarbeiding av rapporter at originalobservasjonene ikke skal presenteres. Med mange slike observasjoner vil en presentasjon bli fullstendig uoversiktlig og til liten glede for leseren. En enkel måte å gi en oversiktlig fremstilling av et originalmateriale på, er å ordne materialet i en frekvensfordeling. I tabell 1 og 2 er det ført opp en frekvensfordeling for henholdsvis en diskret random variabel, og en kontinuerlig random variabel ordnet i klasser. I presentasjonen kan en da enten bruke de absolutte tall eller relative frekvenser. I det siste tilfellet er det nødvendig også å gi antallet observasjoner da det selvsagt ikke er uten betydning for leseren

om det ligger 10 eller 1000 observasjoner til grunn for de frekvenser som er beregnet. Frekvensfordelingen kan også gis i form av et søylediagram.

Tabell 1.

Antall grisunger x	Frekvens z
2	1
3	1
4	4
5	6
6	17
7	20
8	30
9	35
10	51
11	52
12	39
13	45
14	21
15	7
16	5
	n=334

Tabell 2.

Vekten av 144 frosker i gram.

34,0	40,0	40,0	35,5	33,5	34,0	31,0	41,0	26,0	38,5	23,0	28,5
33,5	34,0	33,5	37,0	32,0	35,5	35,5	35,5	31,0	25,5	39,0	26,5
34,0	37,5	28,9	30,2	32,0	28,5	28,2	33,9	28,1	29,0	29,2	25,7
29,5	27,0	28,0	26,6	32,4	34,0	33,1	30,0	27,9	30,4	23,5	31,3
31,8	27,6	34,7	29,5	26,5	38,0	21,0	37,0	35,0	34,0	33,0	37,0
27,2	33,5	25,5	35,0	33,8	32,9	32,9	38,5	28,7	36,7	31,0	38,5
30,5	29,0	30,0	27,5	28,2	32,0	40,1	28,2	30,5	33,5	26,5	32,4
30,5	29,5	27,0	32,0	32,0	31,5	32,5	27,5	34,5	30,0	33,2	26,2
29,0	24,2	31,2	32,0	33,0	33,0	31,7	33,2	26,5	26,5	32,5	27,3
31,0	28,0	28,5	32,2	30,2	29,5	32,0	28,0	31,0	29,5	26,5	35,5
32,5	29,5	23,2	29,5	37,5	26,8	28,2	30,8	30,2	29,2	35,0	34,2
30,8	27,5	25,5	31,0	27,5	31,0	21,8	27,8	25,5	29,0	27,0	30,0

Vekten av frosker ordnet i klasser

Klasse	Antall	Kumulert antall	Relativ frekvens	Kumulert relativ frekvens
21,0-22,9	2	2	0,014	0,014
23,0-24,9	4	6	0,028	0,042
25,0-26,9	17	23	0,118	0,160
27,0-28,9	25	48	0,174	0,334
29,0-30,9	26	74	0,180	0,515
31,0-32,9	26	100	0,180	0,694
33,0-34,9	21	121	0,146	0,840
35,0-36,9	9	130	0,062	0,964
39,0-40,9	4	143	0,028	0,992
41,0-42,9	1	144	0,007	0,999

STATISTISKE HJELPESTØRRELSER (OBSERVATØRER)

Det kan ofte være praktisk å konsentrere originalmaterialet i noen få funksjoner av observasjonene. Det er oppstilt enkelte ønskemål for disse, bl.a. bør de:

a) være veldefinerte og entydige
b) være en funksjon av alle observasjonene
c) ha enkle matematiske egenskaper
d) være lite influert av tilfeldige feil
e) være lette å regne ut.

Hjelpestørrelsen er som regel en sentralverdi som observasjonene grupperer seg om, og en verdi som angir spredningen rundt denne sentralverdien.

Den mest brukte sentralverdien er det aritmetiske gjennomsnittet eller middeltallet. Dette betegnes som kjent $\bar{X}$ og er definert ved

$$\bar{X} = \frac{1}{n} \Sigma X_i$$

Summen av differensene mellom observasjonene og middeltallet er 0

$$\Sigma(X_i - \bar{X}) = \Sigma X_i - \Sigma \bar{X} = n\bar{X} - n\bar{X} = 0$$

fordi

$$\Sigma X_i = n\bar{X}$$

etter definisjonen, og

$$\Sigma \bar{X} = n\bar{X}$$

fordi $\bar{X}$ kan betraktes som en konstant.

Nå skal en være oppmerksom på at $\bar{X}$ kan gi misvisende resultat. Tar en f.eks. bakterietellinger hvor forskjeller mellom paralleller er ganske store, vil det kanskje ikke være noen

mening i middeltallet. Beregning av middeltall for en avledet størrelse som ikke har lineær sammenheng med X, vil som regel gi et galt resultat og en skal også være svært varsom med å beregne middeltall for prosenter.

En annen sentralverdi som er litt benyttet, er medianen, som er definert ved at antallet observasjoner som er større og mindre enn denne, er like.

Som mål for spredningen rundt middeltallet, brukes vanligvis middelavviket, s_x, som er roten av sampel variansen, V.

$$s_x^2 = V_x = \frac{\Sigma(X-\bar{X})^2}{n-1}$$

hvor da n er antall observasjoner. Dette er en god karakteristikk bl.a. fordi kvadratsummen av avvikene fra middeltallet er mindre enn kvadratsummen av avvikene fra et hvilket som helst annet tall:

$$\Sigma(X-C)^2 = \Sigma(X-\bar{X}+X-C)^2 =$$
$$\Sigma(X-\bar{X})^2 + \Sigma(2(X-\bar{X})(\bar{X}-C) + \Sigma(\bar{X}-C)^2 =$$
$$\Sigma(X-\bar{X})^2 + (\bar{X}-C)\,\Sigma(X-\bar{X}) + \Sigma(\bar{X}-C)^2 =$$
$$\Sigma(X-\bar{X})^2 + n(\bar{X}-C)^2$$

s har dessuten samme dimensjon som $\bar{X}$.

Med de moderne regnemaskinene er det ikke noe problem å få regnet ut middelavviket. Ellers kan $\Sigma(X-\bar{X})^2$ regnes ut på en enkel måte idet:

$$\Sigma(X-\bar{X})^2 = \Sigma(X^2-2X\bar{X}+\bar{X}^2) =$$
$$\Sigma X^2-2\bar{X}\Sigma X+n\bar{X}^2 =$$
$$\Sigma X^2-2\bar{X}\Sigma X+\bar{X}\Sigma X =$$
$$\Sigma X^2-\frac{(\Sigma X)^2}{n}$$

Ved lineær transformering av X til en ny størrelse X^1 vil gjennomsnittet og sampelvariansen bli transformert på samme måte:

$$X^1 = aX + b$$
$$\Sigma X^1 = a\Sigma X + nb$$
$$\frac{\Sigma X^1}{n} = \frac{a\Sigma X}{n} + b$$
$$\bar{X}^1 = a\bar{X} + b$$

$\Sigma(X-X)^2$ forkortes ofte til $\mathbf{x}$ hvor da $\mathbf{x}$ er avviket fra gjennomsnittet.

Det kan vises at en ved hjelp av s og $\bar{X}$ kan avgrense et område fra $\bar{X}$-3s til $\bar{X}$+3s hvor praktisk talt alle observasjonene ligger innenfor, uansett hvor skjev fordelingen av observasjonene er.

Det kan i enkelte tilfelle være hensiktsmessig å innføre en dimensjonsløs størrelse, **va**riasjonskoeffisienten, definert ved:

$$\text{V koef.} = \mathbf{s}_x/\bar{X}$$

Hvis en bruker medianen som sentralverdi, kan spredningen karakteriseres ved kvartilintervalet:

$$Q = \tfrac{1}{2}(Q_3 - Q_1)$$

hvor 1/4 av alle observasjonene er mindre enn Q_1 og 1/4 større enn Q_3. Observasjonene kan også inndeles i 10 like grupper, desiler eller 100, centiler.

Innenfor rutinemessig statistisk kvalitetskonstoll brukes av og til variasjonsbredder d.e. avstand mellom største og minste verdi, som mål for spredningen. Den store svakheten er at en her bare har brukt to observasjoner ved beregningen.

GRUNNFORMLER I KOMBINATORIKKEN

Vi tar utgangspunkt i at vi har n elementer nummerert i en bestemt rekkefølge, $a_1 - - a_i - - - a_n$. Et element kan her

velges ut på n forskjellige måter. Hvis vi skal finne antall kombinasjoner av to eller flere elementer, må vi først presisere hva vi vil mene med forskjellig.

Antall forskjellige ordninger, permutasjoner, av n elementer er $n!$

Eksempel: Du kan ordne 4 bokstaver ABCD på $4 \cdot 3 \cdot 2 = 24$ forskjellige måter.

1. Etter at element a_i er trukket ut, legger vi det tilbake igjen. Kombinasjon som $a_i a_1$ er da mulig.

2. Hvis vi ikke legger elementene tilbake, vil bare kombinasjonen av typen $a_i a_j$ hvor $i \neq j$ kunne forekomme.

3. Betrakter vi $a_i a_j$ og $a_j a_i$ som forskjellige idet vi tar hensyn til rekkefølgen, taler vi om ordnede utvalg.

4. Er rekkefølgen uten interesse, vil $a_i a_j$ og $a_j a_i$ ikke være forskjellige. Vi har da et uordnet utvalg.

Før vi går videre, vil vi definere følgende uttrykk:

$n!$ (n fakultet) $= n(n-1)$ -----$2 \cdot 1$

$n^{(r)}$ (n i r faktoriell) $= \frac{n!}{(n-r)!} = n(n-1)$ -----$(n-r+1)$

$\binom{n}{r}$ (n over r) $= \frac{n!}{r!(n-r)!}$

Det kan vises at antall forskjellige kombinasjoner av r elementer fra n kan regnes ut på denne måten.

	Tilbakelegging	Ikke tilbakelegging
Ordnet	n^r	$n^{(r)}$
Ikke ordnet	$\binom{n+r-1}{r}$	$\binom{n}{r}$

En annen viktig regel er denne:

n elementer kan inndeles i m grupper med henholdsvis r_1, r_2 ----r_m elementer i hver på

$$\frac{n!}{r_1!, r_2!---r_m!}$$

måter.

Tabellene $\binom{n}{r}$ kalles binomiske koeffisienter fordi de forekommer i binomialformelen.

$$(p + q)^n = \Sigma \binom{n}{x} p^x q^{n-x}$$

På tilsvarende måte kalles koeffisientene

$$\frac{n!}{r_1!r_2! = r_m!}$$

for multinomiske fordi de kan dannes ved utregning av

$$(p_1 + p_2 + -----p_m)^n$$

ordnet etter potenser av p_1, p_2 ------p_m

Ellers er det en viktig regel som sier at hvis vi har n_1 elementer fra a_1---a_{n_1}, n_2 elementer b_1--b_{n_2} og n_r elementer c_1-----c_{n_r}, er det mulig å danne $n_1 \cdot n_2 \cdot n_r$ r tupler $(a_{i_1}, b_{i_2} \cdots i_r)$ sammensatt av et element fra hver gruppe.

Eksempel: Ei kvinne med 4 lange skjørt, 4 bluser og 3 par sko, vil trolig prøve alle kombinasjonene 4·4·3 = 48 før hun bestemmer seg for hva hun skal ha på til festen.

ELEMENTER I SANNSYNLIGHETSREGNINGEN

Vi har tidligere sett litt på noen metoder som vi bruker for å ordne og karakterisere en serie observasjoner. Ved

hjelp av visse karakteristika som $\bar{X}$ og s kan vi skaffe oss kunnskaper om det utvalget av gjentak (sampel) som observasjonene gjelder. Imidlertid må samplet oppfattes som tilfeldig valgt fra et univers av sampler, og det er som oftest dette universet vi ønsker å kunne si noe om. Dette får vi ved å generalisere kunnskapene om samplet ved induksjon. Vi kan imidlertid aldri si noe absolutt om universet, men vi kan uttale oss med bestemte sikkerhetskriterier. Til dette har vi bruk for sannsynlighetsregning. Kombinatorikken så vi litt på tidligere og de reglene som ble nevnt der, nyttes ofte innenfor sannsynlighetsregningen. En del nye regler skal nevnes i det følgende:

Den klassiske definisjonen på sannsynlighet var laget ut fra hazardspillernes synspunkt og behov, og kan stilles opp slik: Hvis en begivenhet A, definert i et univers U av begivenheter, kan inntreffe i n_A av n tilfeller hvor alle begivenhetene er like mulige, så er sannsynligheten for begivenheten A:

$$P(A|U) = P(A) \text{ definert som } n_A/n.$$

For klassikerne var sannsynlighetene like store for alle enkelttilfeller idet dette er tilfelle for terninger og for kort (uniform sannsynlighetsmodell).

En mer generell definisjon kan være: Vi antar at universet U er en uendelig rekke av begivenheter, hver med sitt kjennetegn. Vi tar et n-sampel fra dette universet, og her opptrer kjennetegnet A n_A ganger. Dersom n vokser, vil n_A/n gå mot en grense som vi kaller sannsynligheten for A.

$$P(A|U) = \lim_{n \to \infty} (n_A/n)$$

Hvis A og B er to begivenheter som utelukker **hverandre slik** at de ikke kan inntreffe samtidig, så er

$$P(A \text{ eller } B|U) = \frac{n_A+n_B}{n} = P(A\ U) + P(B\ U)$$

Satsen kan generaliseres til å gjelde alle begivenheter som er mulige, og er da

$$P(A \text{ eller } B \text{ eller } C|U) = \frac{n_A+n_B+\cdots+n_C}{n} =$$

$$P(A|U) + P(B|U) + \cdots + P(C|U) = 1$$

fordi summen av alle uavhengige begivenheter som kan inntreffe, er n. Hvis da $\bar{A}$ betegner begivenheten ikke A, så er:

$$P(A|U) + P(\bar{A}|U) = 1$$
$$P(A|U) = 1 - P(\bar{A}|U)$$

Dette kan illustreres ved såkalte Venn-diagrammer.

Hvis A og B er to begivenheter som kan inntreffe samtidig og vi med $n_{AB} = n_{BA}$ forstår antall tilfeller hvor både A og B har inntruffet, så er:

$$P(A \text{ og } B|U) = \frac{n_{AB}}{n} = \frac{n_A}{n} \cdot \frac{n_{BA}}{n_A} = \frac{n_B}{n} \cdot \frac{n_{AB}}{n_B}$$

Sannsynligheten er da sannsynligheten for A multiplisert med sannsynligheten for B i et delunivers hvor A har inntruffet - eller lik sannsynligheten for B multiplisert med den betingede sannsynlighet for A.

$$P(A \text{ og } B|U) = P(A|U)\ P(B|A, U)$$

Hvis sannsynligheten for B er den samme enten A har inntruffet eller ikke, sier vi at begivenheten B er stokastisk uavhengig av A.

$$P(B|A, U) = P(B|\bar{A}, U) = P(B|U)$$

Når A og B er gjensidig uavhengige vil

$$P(\text{A og B}|U) = P(A|U)P(B|U)$$

Hvis A og B utelukker hverandre, er

$$P(\text{A og B}|U) = 0$$

Hvis A og B ikke utelukker hverandre, kan vi få A og/eller B. Blant de tilfellene hvor A har inntruffet, inngår også en del tilfeller hvor både A og B har inntruffet. Disse er da tatt med 2 ganger i P(A|U), men skal bare forekomme en gang i uttrykket P(A og/eller B|U). Vi får da regelen:

Sannsynligheten for at A og/eller B skal inntreffe, er lik sannsynligheten for at A skal inntreffe pluss sannsynligheten for at B skal inntreffe, minus sannsynligheten for at både A og B skal inntreffe.

Alle disse reglene kan lett utvides til flere variable.

FORDELINGSFUNKSJON

Diskrete random variable

Vi tenker oss at vi har et utvalg på n gjentak av et univers (U) og at vi for hvert gjentak har observert et diskret variabel X. Vi har sett hvordan vi kan ordne observasjonsrekkene i en frekvensrekke hvor frekvensen til X_i er f.eks. z_i. Da vil selvfølgelig

$$z_1 + z_2 + \text{----} z_m = n \qquad \text{og}$$

$$z_1/n + z_2/n + \text{---}z_m/n = 1$$

når antall frekvenser er m. Til hver x_i vil det være en relativ frekvens $\mathbf{X}_i/n$ og også en sannsynlighet $P(X_1 U)$. Vi kan da si at sannsynligheten er en funksjon av X og sette

$$P(X|U) = f(X)$$

Denne funksjonen kaller vi for fordelingsfunksjonen for X i universet. Da de verdiene X kan ha utelukker hverandre, må

$$\Sigma P(X|U) = \Sigma f(X) = 1$$

Multipliserer vi de relative frekvensene $\frac{z_i}{n}$ med X_i, får vi:

$$\frac{z_1X_1}{n} + \frac{z_2X_2}{n} + \text{----}\frac{z_mX_n}{n} = \Sigma\frac{z_mX_i}{n} = \bar{X}$$

Tilsvarende definerer vi forventningen μ, for **fordelings-**funksjonen

$$\mu = E(\mathbf{X}) = \Sigma f(\mathbf{X})X$$

Forskjellen mellom $\bar{X}$ og μ, ligger i at $\bar{X}$ er middelet av de aktuelle observasjonene. Den eksakte verdien for forventningen kan vi bare finne hvis fordelingsfunksjonen er nøyaktig kjent. Etter den første likningen vil imidlertid frekvensen nærme seg sannsynligheten når $\mu \to \infty$ og $X \to \mu$. Vi sier da at $\bar{X}$ er en forventringsrett estimator av μ.

Vi definerte tidligere middelavviket for samplet. For fordelingsfunksjonen har vi en tilsvarende størrelse, standardavviket definert ved

$$\sigma^2 = \text{Var}(\mathbf{X}) = \Sigma f(\mathbf{X})(X-\mu)^2$$

Fører vi inn den relative frekvensen $f_i = z_i/n$ i formelen for middelavviket, får vi:

$$s^2 = \Sigma\frac{(X-\bar{X})^2}{n-1} = \Sigma\frac{z(X-\bar{X})^2}{n-1} =$$

$$\Sigma\frac{nf(X-\bar{X})^2}{n-1} = \frac{n}{n-1}\,\Sigma f(X-\bar{X})^2$$

Hvis $n \to \infty$, går $\frac{n}{n-1} \to 1$,

$f \to f(X)$

$\bar{X} \to \mu$

$\Sigma f(X-\bar{X}) \to \Sigma f(X)(X-\mu)^2$

og $s^2 \to \sigma^2$

Som nevnt kjenner vi meget sjelden til fordelingsfunksjonen, og de karakteristikka vi beregner er derfor bare estimater. Det er imidlertid utviklet en del fordelingsfunksjoner som en da kan undersøke om passer i de enkelte tilfeller.

DEN HYPERGEOMETRISKE FORDELING

Vi betrakter et vareparti på N enheter med M defekte. Sannsynligheten for å finne en defekt når enheten tas tilfeldig ut, er M/N = p. Tar vi ut en enhet til, vil sannsynligheten for at denne er defekt bli:

$$\frac{M-1}{N-1} = \frac{\frac{M}{N}-\frac{1}{N}}{1-\frac{1}{N}} = \frac{p-\frac{1}{N}}{1-\frac{1}{N}}$$

Sannsynligheten for å få defekte enheter i de x første trekk og da n - x normale enheter i de neste, er produktet av sannsynlighetene for hver av trekkene, og altså:

$$\frac{p(p-\frac{1}{N})\text{---}(p-\frac{x-1}{N})\;\; q\;\; (q-\frac{1}{N}\;\text{-----}\;(q-\frac{n-x-1}{N})}{1(1-\frac{1}{N})\;\;\text{-----}\;(1-\frac{n-1}{N})}$$

Kombinasjonen x ganger defekt og n-x ganger normal av n antall enheter er ifølge kombinatorlæren $\binom{n}{x}$

Følgelig blir sannsynligheten for å få x defekte enheter i N-samplet når p er sannsynligheten i første trekk og q=1-p:

$$P_H(x|N,p,n) = \binom{n}{x} \frac{p(p-\frac{1}{N}) \text{ --- } (p-\frac{x-1}{N})q\ (q-\frac{1}{N}) \text{ --- } (q-\frac{n-x-1}{N})}{1(1-\frac{1}{N}) \text{ --- } (1-\frac{n-1}{N})}$$

Hvis nå $N \to \infty$, vil funksjonen gå mot

$$f(x) = \binom{n}{x} p^x q^{n-}$$

som kalles binomialfordelingen, som vi også kan komme fram til ved trekk og tilbakelegging.

Den binomiale fordelingsfunksjonen vil gi sannsynligheten for x gjentak med kjennetegnet E og (n-x) gjentak med kjennetegnet $\bar{E}$ i et tilfeldig utvalg på n gjentak. Vi bruker da denne innen sensorikken ved partest, triangeltest og liknende opplegg hvor det er gitt to muligheter (lik eller ulik), (søtest eller minst søt) osv., p er her henholdsvis 0,5 og 0,33.

Det har vist seg at den også er brukbar til å beskrive enkelte frekvensfordelinger, f.eks. arrstråler hos valmuer. Parametrene p og n må da estimeres. (En betegner estimater for de virkelige parametrene med tegnet ^ over symbolet for parameteren).

Det finnes en rekke andre fordelingsfunksjoner for diskrete random variable. I sensorikken kan vi få bruk for den multinomiale fordelingsfunksjonen. Den gir fordelingen for en serie uavhengige kjennetegn med hver sin sannsynlighet.

$$f(x_1,x_2, \text{ ---- } x_m) = \Sigma \frac{n!}{x_1!\ x_2!\ x_m!}\ p_1^{x_1}\ p_2^{x_2} \text{ ---} p_m^{x_m}$$

KONTINUERLIGE FORDELINGSFUNKSJONER

Variable som teoretisk kan ha alle mulige verdier, kan ikke så lett ordnes etter relative frekvenser. Vi kan imidlertid ordne slike observasjoner i klasser og gjøre klasseintervalet så lite vi bare vil. For en slik klasse kan vi operere med en frekvens z_i og også med en relativ frekvens z_i/n. Til denne relative frekvensen svarer på samme måte som for de diskrete variable, en sannsynlighet $p(X_i|U)$ for en verdi av X i klassen. Fordelingsfunksjonen for X er i dette tilfellet en kontinuerlig funksjon som tilfredsstiller kravene om at arealet av flaten som er avgrenset av kurven, x-aksen og ordinatene for grensene av klassen er $P(X_i|U)$ og $\int P(X_i|U) = 1$.

I stedet for Σ-tegn bruker vi $\int$-tegnet og definerer forventningen og standardavvik som henholdsvis:

$$E(x) = \mu = \int f(X)X\,dX \text{ og } E(X)$$
$$Var(x) = \sigma^2 = \int f(X)(X-\mu)^2\,dX$$

Den viktigste av de kontinuerlige fordelingsfunksjonene er den normale eller Gauss-Laplace fordelingslov. Selv om fordelingen ser komplisert ut

$$(X) = \frac{1}{\sigma\sqrt{2\pi}}\ e$$

så er parametrene μ og σ gamle kjente.

hvor μ og δ er gamle, kjente størrelser. Det grafiske bildet er en "klokkeformet" kurve med ett maksimum for $X = \mu$.

Hvis vi stykker opp kurven ved å starte ved μ og gå $\frac{1}{2}\sigma$ for hver klasse, finner vi de sannsynlighetene for **X** som er vist i tabell 3. Bare den ytterste klassen er her

større enn $\frac{1}{2}\sigma$. Vi kan også finne sannsynligheten for $(X-\mu)$ større enn en bestemt verdi.

Tabell 3.

Klassegrenser for den normale fordelingsfunksjonen.

Nedre	Øvre	P(X U)
μ	$\mu + 0{,}5\sigma$	0,19146
$\mu + 0{,}5\sigma$	$\mu + \delta$	0,14988
$\mu + \sigma$	$\mu + 1{,}5\sigma$	0,09185
$\mu + 1{,}5\sigma$	$\mu + 2\delta$	0,04406
$\mu + 2\sigma$	$\mu + 2{,}5\sigma$	0,01654
$\mu + 2{,}5\sigma$	$\mu + 3\delta$	0,00486
$\mu + 3\delta$	$\mu + 3{,}5\sigma$	0,00112
$\mu + 3{,}5\sigma$	$\mu + 4\sigma$	0,00020
$\mu + 4\sigma$	∞	0,00003

Sannsynligheten for $X - \mu = a\sigma$ er satt i tabell (4). Ut fra slike beregninger kan vi også sette opp nødvendig a-verdi for valgte sannsynligheter. Vi ser at verdier av $X > 3\sigma$ er meget sjelden.

Den normale fordelingsfunksjonen er svært viktig, ikke minst fordi fordelingsfunksjoner som brukes i praktisk statistiske metoder ofte har utgangspunkt i at de observerte variable er normalfordelt. Det er særlig i tilfeller hvor den random variable tydeligvis må være bestemt av et stort antall årsaker som er omtrent likeverdige, at normalfordelingen passer som modell. De fleste observasjoner fra biologien bør derfor passe så noenlunde inn i normalfordelingen.

Hvis vi ikke bare har ett random sampel av gjentak, men f.eks. r sampler med n gjentak, så vil vi kunne regne ut r gjennomsnitt og r sampelvarianser. Disse beregnede verdier vil

også variere. Hvis fordelingsfunksjonen for den observerte variable er normal, så kan det vises at $\bar{X}$ er normalfordelt med

$$E(\bar{X}) = \mu \quad \text{og}$$
$$\text{Var}(\bar{X}) = \sigma^2/n$$

Det kan også vises at fordelingsfunksjonen for $V = s^2$ er

$$f(V) = K \cdot V^{\frac{1}{2}(n-3)} \, e^{\frac{(n-1)V}{2\sigma^2}}$$

og at

$$E(V) = \sigma^2$$

$$\text{Var}(V) = \frac{2}{n-1}\sigma^2$$

Det kan vises at uansett fordelingsfunksjon så er $\bar{X}$ og V forventningsrette estimatorer av μ og σ^2, og også uavhengig av hverandre. Når fordelingsfunksjonen for X ikke er normal, vil imidlertid fordelingsfunksjonen for $\bar{X}$ og for V være anderledes og som regel ukjent.

Vi ser imidlertid av formlene at $(\bar{X})$ vil avta proporsjonalt med n. Dette betyr at nøyaktigheten av et gjennomsnitt også øker proporsjonalt med roten av antall observasjoner.

Tabell 4.

Sannsynligheten, P, for $X - \mu \geq a\sigma$ ved valgte a-verdier for den normale fordelingsfunksjonen.

a	P
0,5	0,61708
1,0	0,31732
1,5	0,13362
2,0	0,04550
2,5	0,01242
3,0	0,00270
3,5	0,00046
4,0	0,00006

Det er noe diskusjon om betegnelsen på størrelsen $s_{\bar{x}}$ og s/n . Gjennomsnittets middelavvik er en korrekt, men tungvint betegnelse, noen kaller den også for middelfeilen.

En rekke statiske metoder har som forutsetning at den tilfeldig variable har normal fordelingsfunksjon. Imidlertid vil selv relativt store avvik fra denne forutsetningen som regel ikke gi større vanskeligheter. Metodene blir da betegnet som robuste.

TEST-FUNKSJONER

Students-t.

Vi forutsetter at den observerte tilfeldig variable har normal fordelingsfunksjon. I et tilfeldig utvalg på n gjentak, har vi skaffet oss n observasjoner og beregnet gjennomsnitt og middelavvik. Setter vi

$$t = \frac{\bar{X} - \mu}{s/\sqrt{n}}$$

så er også t en tilfeldig variabel, men den er ikke normalfordelt med mindre n er et meget stort antall. Den engelske statistikeren GOSSET (1876-1937) som utga sine arbeider under pseudonymet "Student" utledet fordelingsfunksjonen for t, og denne er da kjent som Student's t.

$$f(t) = \frac{K}{(t^2+f)^{\frac{1}{2}(f+1)}}$$

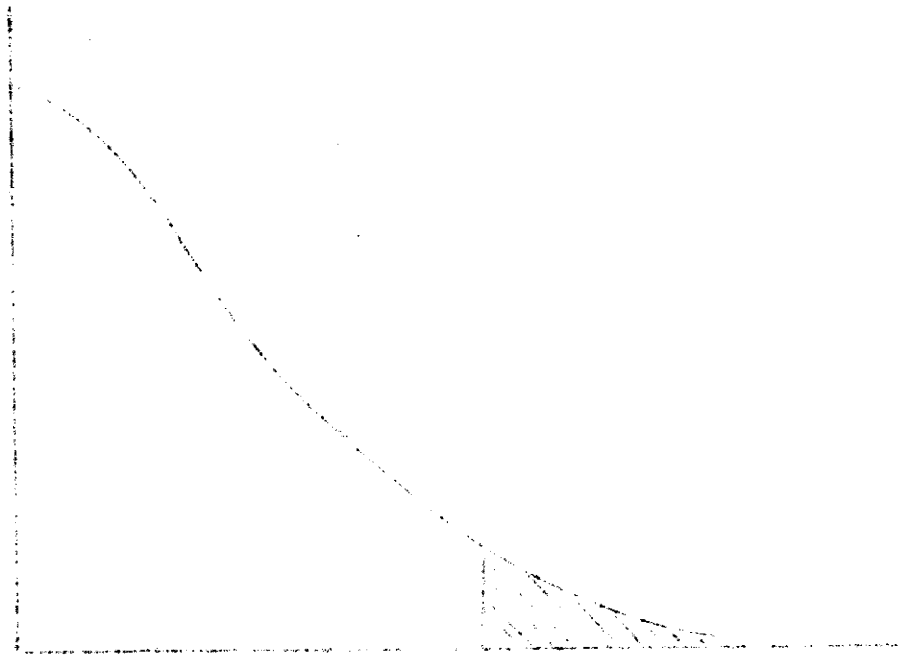
hvor K er en konstant. Parameteren f har blitt kalt antall frihetsgrader og kan vanligvis beregnes som antallet av de variable som har mulighet til fritt å kunne variere. Har vi f.eks. n observasjoner og regner ut $\bar{X}$, så har n-1 av observasjonene mulighet til å variere. Den siste er imidlertid bestemt fordi $\bar{X}$ da er bestemt.

Funksjonen er symetrisk om t=0 og faller sammen med den normale når $f \to \infty$.

Som regel er vi ikke interessert i fortegnet for t, men bare i tallverdien.

$$t = \frac{|\bar{X} - \mu|}{s/\sqrt{n}}$$

Student´s t fordeling. Tosidig test.



Hver verdi av t' vil da være den dobbelte av t hvis vi bare bruker positive $(\bar{X} - \mu)$-verdier.

Fra figuren over t-fordelingen ser en at arealet av den skraverte flata er lik sannsynligheten for at tallverdien av t skal være større eller lik en størrelse a, $(p(t' \geq a))$. Vi kaller dette for et to-sidig test i motsetning til en-sidige test hvor en også er interessert i fortegnet for t. For det en-sidige testet er derfor sannsynligheten bare halvparten av sannsynligheten for det to-sidige testet for samme a-verdi.

Nå er t en funskjon av antall frihetsgrader, og det må da a også være. Det er derfor beregnet tabeller for a-verdier ved valgte sannsynligheter (P) og bestemte antall frihets-grader (f).

OM ESTIMERING

En av de viktigste oppgavene i praktisk statistisk arbeid er å estimere gode tilnærmelsesverdier for størrelser vi ikke kjenner. Innenfor all naturvitenskap er vi interessert i å utvikle metoder som gir oss mest mulig presise mål for ukjente størrelser. F.eks. kan vi være interesserte i bedring i metodikken ved analyse av næringsmidler med hen-syn til fett, protein, sukker osv. Ved slike estimeringer har vi feil av to typer. Feil som slår ut i en retning, kaller vi systematiske. Disse kan skyldes feil ved måle-instrumentene eller metodene. F.eks. ved bestemmelse av protein i melk, tar vi som regel utgangspunkt i N-analyser og beregner proteininnholdet ut fra det, selv om vi vet at en del av dette N-innholdet er ikke-protein-nitrogen. Slike systematiske feil kan være grunn til mange feil-slutninger, men ved fornuftig analyse av metodikk, bør de kunne elimineres. Tilbake er da de tilfeldige feilene som

kan slå i begge retninger. Disse kan gjøres mindre ved f.eks. å ta i bruk bedre metoder, og arbeide mer nøyaktig. Imidlertid kan en neppe helt slippe bort fra slike feil. I teorien oppfattes da slike feil som forårsaket av et stort antall uavhengige faktorer som virker i begge retninger og hver feil kan da betraktes som en sum av uavhengig småfeil. Det kan under slike forutsetninger vises at den tilfeldige feilen er en normalfordelt tilfeldig variabel med forventning = 0. Vi forutsetter at vi har n uavhengige bestemmelser av proteininnholdet i ei melkeprøve. Gjennomsnittet $\bar{X}$ er da en forventningsrett estimator av proteininnholdet P. Hvis vi fortsatte med å analysere flere n-utvalg, ville vi få en variasjon også av $\bar{X}$-verdiene fordi $\bar{X}$ er en tilfeldig variabel. Det beste estimat vi kan få av proteininnholdet ut fra vårt utvalg, er å sette $P = \bar{X}$. Hvor godt vi har estimert P ved dette, kommer av mange ting hvor metodens nøyaktighet for proteinbestemmelse og nøyaktigheten ved arbeidet vil slå sterkt ut. Vi kan da snakke om presisjonen ved analysen. Når estimatoren er $\bar{X}$, vil $s/\sqrt{n}$ si noe om presisjonen, og resultatet presenteres da ofte som

$$P = \bar{X} \pm s/\sqrt{n}$$

Kan vi med en bestemt sannsynlighet si hvor stort det intervallet som P må ligge i, er, kaller vi dette tallområdet for et <u>konfidensintervall</u>. Forutsetter vi at fordelingsfunksjonen for den observerte tilfeldig variable er normal, så vil

$$t = \frac{\bar{X} - C}{s\sqrt{n}}$$

være t fordelt med n - 1 frihetsgrader.

Utregnede tabeller kan da gi oss sannsynligheten for $t \geq a$ og da er $Q = 1 - P$ sannsynligheten for $t \leq a$ og sannsynligheten for

$$|\bar{X} - C| \leq a \cdot s/\sqrt{n}$$

når vi bare regner med tallverdien av $(\bar{X} - C)$.

Denne ulikheten kan omskrives til

$$\bar{X} \quad as/\sqrt{n} \leqq C \leqq \bar{X} + as/\sqrt{n}$$

og vi har da konfidensgrensen for C med konfidenssannsynligheten Q. Når vi derfor sier at C ligger innenfor dette intervallet, er det 1 - Q sannsynlighet for at dette er feil.

Vi kan på samme måte beregne konfidenssannsynligheter for differansen mellom to observasjonsrekker idet differansen $\bar{X}_1 - \bar{X}_2$ er da en forventningsrett estimator av

$$C - (\bar{X}_1 - \bar{X}_2 = \bar{d})$$

Det er da en forutsetning at observasjonene X_1 og X_2 er tilnærmet normalt fordelt med $\sigma_1 = \sigma_2 = \sigma$. For $n_1 \neq n_2$ kan det vises at

$$\bar{d} - as\sqrt{\frac{n_1+n_2}{n_1\ n_2}} \leqq C \leqq \bar{d} + as\sqrt{\frac{n_1+n_2}{n_1\ n_2}}$$

hvor

$$S^2 = \frac{\Sigma(X_1-\bar{X}_1)^2 + \Sigma(X_2-\bar{X}_2)^2}{n_1 + n_2 - 2}$$

med $f = n_1 + n_2 - 2$.

TESTING AV HYPOTESER

I tabell 6 er det gjengitt observasjoner av tørrstoffinnholdet pr. rute for to sorter formargkål i n=6 blokker. Med en konfidenssannsynlighet på 0,95 som gir en a=2,571 finner vi konfidensgrensen.

$$\bar{d} - as/\sqrt{n} = 9,50 - 8,42 = 1,08$$
$$\bar{d} + as/\sqrt{n} = 9,50 + 8,42 = 17,92$$

Hvis vi aksepterer innholdet av utsagnet om dette konfidenintervallet selv om det er 5% sannsynlighet for at det ikke er riktig, vil det si bl.a. at $\bar{d}$ ikke kan være 0. Hvis vi på forhånd hadde fremsatt en hypotese om at det ikke var noen forskjell mellom sortene, ja så måtte vi **forkaste** hypotesen. **Påstanden om** at $\mu_1 - \mu_2 = 0$, kaller vi en 0-hypotese og testing av slike hypoteser har blitt en viktig del av den statistiske metodikken.

Hvis vi imidlertid krever at konfidenssannsynligheten skal være så stor som 0,99, vil konfidensgrensene for eksemplet være -3,7 og 22,7 og nullhypotesen kan ikke forkastes.

Ved en hypotese forstår vi som oftest en regel som kan forklare et fenomen eller en gruppe av fenomener, funnet ved spekulasjon over utførte observasjoner. Hypotesen testes ved konfrontasjoner.

Null-hypotesen er imidlertid ikke grunnlagt på observasjoner. Den er bare et arbeidsgrunnlag som gjør det mulig å utforme teknikken for å undersøke påviselig forskjeller. Det er da vanligvis ingen god mening i å akseptere en nullhypotese. En konkluderer da også med at nullhypotesen ikke kan forkastes på bestemte sannsynlighetsnivåer, og ikke med at den aksepteres.

Tabell 6.
Observasjoner for kg tørrstoff pr. rute av to sorter formargkål.

Gjentak	1. sort	2. sort
1	65	48
2	49	45
3	63	42
4	57	48
5	45	39
6	52	52

Testing av en 0-hypotese går da ut på at vi anser denne for treffende. På dette grunnlaget utleder vi så regler for hva vi kan vente oss av observasjonene. Hvis så disse ikke stemmer med reglene, slutter vi at det er noe i veien med 0-hypotesen.

I praksis bør vi være oppmerksomme på at hvis vi har få gjentak, så bør vi ta forbehold når vi konkluderer. Særlig er det da vanskelig å forkaste 0-hypotesen. Med svært mange gjentak kan det ofte være tilfelle at en må forkaste 0-hypotesen selv om differansene er så små at de ikke har noen praktisk interesse.

Kji-kvadrat og sammenhengen mellom noen vanlig brukte fordelingsfunskjoner

Vi lar sannsynligheten for et kjennetegn E i universet U være P(E/U), og da kan en beregne sannsynligheten for z gjentak i et tilfeldig sampel på n gjentak ved binomialloven. E(z) er da $\mu = np$ og $\sigma = npq$, hvor $q = 1-p$.

Det kan da vises at

$$U = \frac{z-np}{npq}$$

er standardnormalfordelt ($\mu=0$, $\sigma = 1$) når n er et relativt stort tall. Dette kan da brukes til å teste hypoteser om p.

En har imidlertid funnet at U^2 er en bedre testvariabel, og denne har fått betegnelsen Kji-kvadrat (X^2).

Den vanlige formelen for X^2 er da:

$$\chi^2 = \frac{(z-np)^2}{npq} = \frac{(z-np)^2}{np} + \frac{((n-z)-nq)^2}{nq}$$

eller generelt

$$\chi^2 = \Sigma(f-F)^2/F$$

hvor f er den verdien som er observert og F er den verdien som en venter å finne.

Tetthetsfunksjonen for denne er:

$$f(\chi^2) = \frac{1}{2^{\frac{1}{2}f}-(\frac{1}{2}f)} \quad e^{-\frac{1}{2}\chi^2} (\chi^2)^{\frac{1}{2}f-1}$$

hvor f er antall frihetsgrader. χ^2-fordelingen er beregnet for forskjellige antall frihetsgrader og ført opp i tabeller hvor P i tabellhodene betyr sannsynligheten for $\chi^2 \geq a$, og a er verdiene i tabellen.

For fordelingen kan det utledes følgende viktige regler:

1. Dersom Y_1, Y_2 ---Y_n er uavhengig og har Kji-kvadrat-fordeling med henholdsvis f_1, f_2 ---f_n frihetsgrader, så er summen ΣY_i Kjikvadrat-fordelt med Σf_i frihetsgrader.

2. Dersom $U_1 = U_2 + U_3$ er Kji-kvadratfordelt med f_i frihetsgrader og U_2 også er Kji-kvadratfordelt med f_2 frihetsgrader og U_2 og U_3 er stokastisk uavhengig, så er U_3 Kji-kvadratfordelt med $f_3 = f_1 - f_2$.

3. Hvis U har en standardnormalfordeling og V^2 en uavhengig Kji-kvadratfordeling med f frihetsgrader, så har

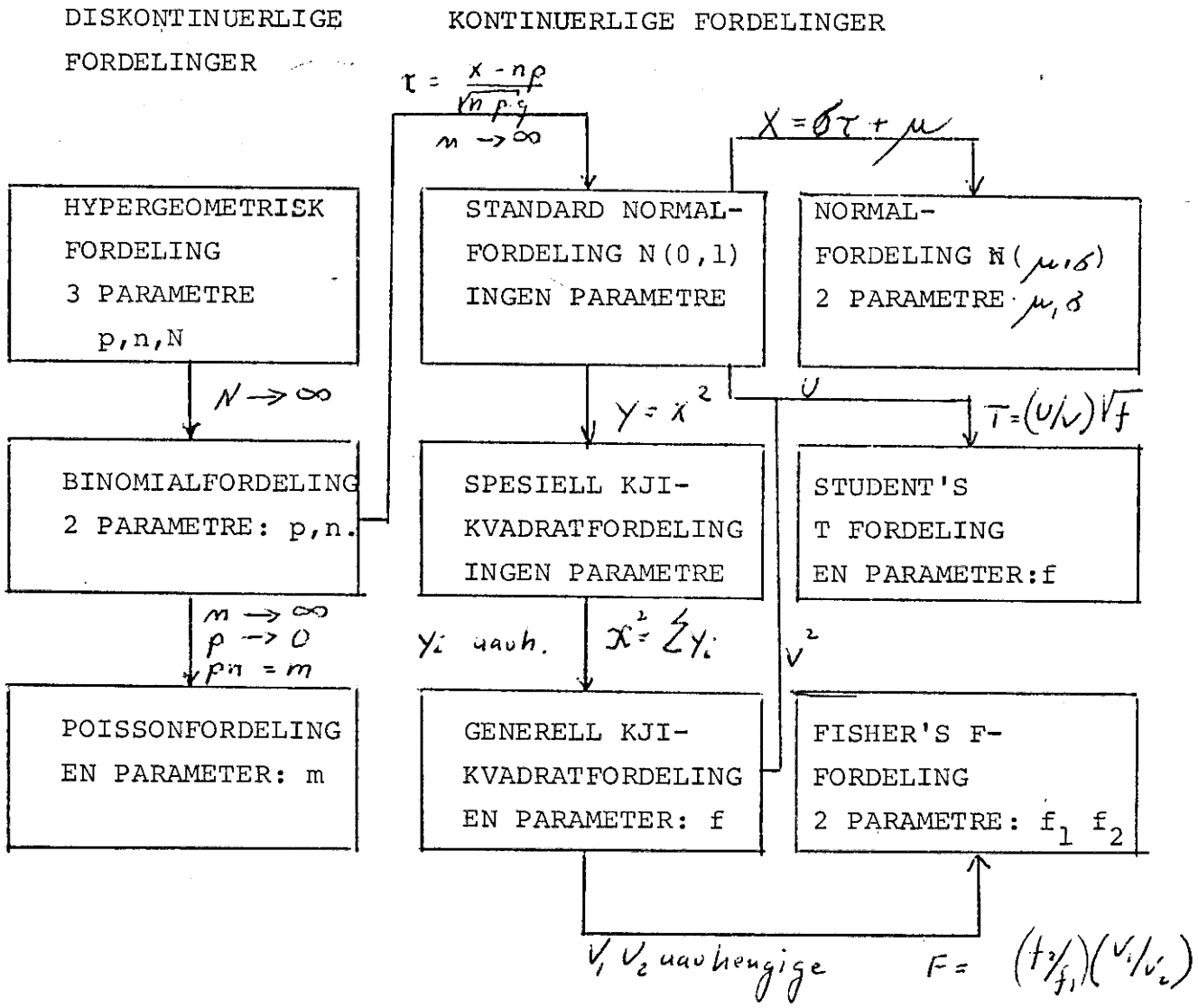
$$t = \frac{U}{V}\sqrt{f}$$

en Student's t-fordeling med f frihetsgrader.

4. Dersom V_1 og V_2 er uavhengige og har Kji-kvadrat-fordelinger med henholdsvis f_1 og f_2 frihetsgrader, så har

$$F = \frac{f_2}{f_1} \frac{V_1}{V_2}$$

Fisher's F-fordeling med f_1 og f_2 frihetsgrader.


DISKONTINUERLIGE FORDELINGER
KONTINUERLIGE FORDELINGER
$\tau = \frac{x - np}{\sqrt{npq}}$
$n \to \infty$
$X = \sigma\tau + \mu$
HYPERGEOMETRISK FORDELING
3 PARAMETRE
p,n,N
STANDARD NORMAL-FORDELING N(0,1)
INGEN PARAMETRE
NORMAL-FORDELING N(μ,σ)
2 PARAMETRE μ,σ
$N \to \infty$
$Y = X^2$
U
$T = (U/V)\sqrt{f}$
BINOMIALFORDELING
2 PARAMETRE: p,n.
SPESIELL KJI-KVADRATFORDELING
INGEN PARAMETRE
STUDENT'S T FORDELING
EN PARAMETER: f
$n \to \infty$
$p \to 0$
$pn = m$
Y_i uavh.
$X^2 = \sum Y_i$
V^2
POISSONFORDELING
EN PARAMETER: m
GENERELL KJI-KVADRATFORDELING
EN PARAMETER: f
FISHER'S F-FORDELING
2 PARAMETRE: f_1 f_2
V_1 V_2 uavhengige
$F = (f_2/f_1)(V_1/V_2)$

Det er derfor en nøye sammenheng mellom de forskjellige funksjonene vi vanligvis bruker i praktisk statistikk. (Se figuren).

I praksis brukes Kji-kvadrattestet særlig til hypotesetesting for diskrete kvalitative kjennetegn. Forutsetter en at 0-hypotesen er treffende, kan det vises at

$$E(\chi^2) = f \text{ og } Var(\chi^2) = 2f$$

Hvis derfor det χ^2 vi beregner, ligger langt fra antall frihetsgrader, er det trolig noe i veien med 0-hypotesen vår.

Student's-t

Tidligere har vi sett på hvordan en kan bruke fordelingsfunksjonen

$$t = \frac{\bar{X}-\mu}{s/\sqrt{n}}$$

til å bestemme konfidensinterval.

Vi kan ta for oss 0-hypotesen $\mu = 0$. Da μ ikke går inn i formelen for fordelingsfunksjonen, vil også

$$t = \frac{\bar{X}}{s/\sqrt{n}}$$

Variasjonsområdet for tallverdien av t strekker seg fra 0 til ∞ (tosidig test). Deler vi kurven for t opp i to deler, ved $t = a$ på en slik måte at sannsynligheten for en t-verdi i området B er 0,05, så er selvsagt sannsynligheten for en t-verdi i området A = 0,95. Finner vi en beregnet t-verdi i området B, så bør vi se med all mulig skepsis på utgangspunktet, nemlig nullhypotesen. Forkaster vi 0-hypotesen selv om den er treffende, sier vi at vi har gjort en feil

av type I. Sannsynligheten for denne er da i dette tilfellet 5%. Området B kaller vi også forkastningsområdet for 0-hypotesen.

Vi kan også bruke t-testet til å teste en 0-hypotese om forventningen for differanser : $E(\bar{d}) = \mu_1 - \mu_2 = 0$

$$t = \frac{|\bar{X}_1 - \bar{X}_2|}{s}\sqrt{\frac{n_1 n_2}{n_1+n_2}}$$

hvor s er

$$\frac{\Sigma(X_1-\bar{X}_1)^2 + \Sigma(X_2-\bar{X}_2)^2}{n_1+n_2-2}$$

Formelen gjelder under forutsetning av at fordelingen for den observerte random variable i de to universene ikke avviker svært mye fra den normale og at de to standardavvikene ikke er alt for ulike.

Når $n_1 = n_2 = n$, vil formelen bli

$$t = \frac{|\bar{d}|}{s_2/\sqrt{n}}$$

under 0-hypotesen $\mu_d = 0$, idet en oppfatter d som en tilfeldig variabel.

F-test og variansanalyser

En-veis gruppering.

0-hypotesen $\mu_1 - \mu_2 = 0$ kan som nevnt testes med t-test. Hvis vi har observasjoner av samme random variable fra forskjellige universer, så kan det være av interesse å teste en 0-hypotese som går ut på at forventningene for X er den samme i universene, altså

$$\mu_1 = \mu_2 = \text{-------} \mu_n$$

Som illustrasjon kan vi ta et eksempel fra OTTESTAD hvor observasjonene X = vekten i kg av 56 dager gamle grisunger i 5 kull (K). Grisungene hadde samme far, men fem mødre (T_1-T_5). Antall grisunger pr. kull varierte, og vi vil betegne antall pr. kull med n_j. Kullene betegner vi med indeks j. Vi tenker oss at kullene representerer hvert sitt univers, og vi er interessert i om forventningen for den observerte variable er forskjellig i noen av universene. 0-hypotesen går da ut på at forventningene i de K-universene er like. **(Tabellen står på side 36).**

Det testet som brukes for dette formålet er blitt kalt for F-test etter FISHER. For k = 2 faller F- og T-test sammen, og $F = t^2$.

La oss nå si at vi har k grupper (5) observasjoner. I den j-te gruppa har vi n_j observasjoner (4-6) og hver observasjon betegnes X_{ij}.

For hver gruppe kan vi beregne et gjennomsnitt ($\bar{X}_j$) og en varians (V_j). Da er

$$\bar{X}_j = \frac{1}{n_j} \sum_{i=1}^{i=n_j} X_{ij} = S_j/nj$$

hvor S_j er summen av observasjonene i gruppa.

$$V_j = s_j^2 = \frac{1}{n_j - 1} \Sigma(X_{ij} - \bar{X}_j)^2$$

Vi kan selvsagt også regne ut et gjennomsnitt for alle observasjonene

$$\bar{X} = \frac{i}{N} \Sigma \Sigma X_{ij} = \frac{1}{N} \Sigma S_j = \frac{1}{N} \Sigma n_j \bar{X}_j$$

hvor $N = \Sigma n_j$

Nå definerer vi en varians som vi kaller for variansen innen gruppene og regner den ut som et veid gjennomsnitt av gruppevariansen.

$$V_R = \frac{1}{N-k} \Sigma(n_j-1)\, V_j$$

hvor n_j-1 er antall frihetsgrader for gruppevariansen, og

$$\Sigma(n_j-1) = \sum_{j=1}^{k} n_j - \sum_{j=1}^{k} 1 = N - k$$

Fører vi inn V_j får vi

$$V_R = \frac{1}{N-k} \Sigma\, \Sigma(X_{ij}-\bar{X}_j)^2$$

Nå definerer vi en ny varians for gruppegjennomsnittene ved:

$$V_T = \frac{1}{k-1} \Sigma n_j(\bar{X}_j-\bar{X})^2$$

Med like mange observasjoner i alle gruppene vil

$$V_R = \frac{1}{k} \Sigma V_j$$

$$V_T = \frac{n}{k-1} \Sigma(\bar{X}_j-\bar{X})^2$$

Ved å innføre S_j for summen i den j-te klassen og S for totalsummen, er:

$$\Sigma\, \Sigma(X_{ij}-\bar{X}_j)^2 = \Sigma\, \Sigma X_{ij}^2 - \Sigma S_j^2/n_j \quad \text{og}$$
$$\Sigma n_j\, (\bar{X}_j-\bar{X})^2 = \Sigma S_j^2/n_j - S^2/N$$

Summen av disse er:

$$\Sigma\, \Sigma(X_{ij}-\bar{X})^2 = \Sigma\Sigma X_{ij}^2 - s^2/n = \Sigma\Sigma(X_{ij}-\bar{X})^2$$

Tellerne til V_R og V_T kan vi altså finne ved sisse formlene.

Utregnet fra vårt eksempel finner vi $V_R = 5,19$ og $V_T = 22,73$. Ved testing av 0-hypotesen bruker en nå

$$F = V_T/V_R$$

med k - 1 og N - k frihetsgrader.

Fordelingsfunksjonene for F er beregnet under forutsetning av at den variable er normalfordelt og at standardavviket er det samme for de k universene. Selve tetthetsfunksjonen for F er ganske komplisert, og jeg tar den ikke med her. Under forutsetning av at μ_j er konstant, har de to variansene samme forventning. En stor F betyr derfor at vår 0-hypotese ikke er treffende. Det er imidlertid vist at F-testet er ganske robust : det kan brukes selv om fordelingen for X avviker en del fra normalfordelingen.

Det er utarbeidet tabeller over F-fordelingen med nedre verdier for kritiske områder på 0,1, 1 og 5% nivået for bestemte antall frihetsgrader. I grisungeforsøket vil F-verdien 4,38 gi signifikans på 5% nivået. For å kunne slutte noe om at det f.eks. er arvelighetsfaktoren hos mora som har ført til resultatet; eller om det har vært forskjeller i foring, stell eller liknende, må forsøket være godt planlagt. Vi må imidlertid ha det klart for oss at den store F-verdien kan ha inntruffet ved reine skjære slumpen.

Tabell 7.

a) Vekten av 56 dager gamle grisunger i fem kull med forskjellige mødre (T_i), men med samme far (etter OTTESTAD)

T_1	T_2	T_3	T_4	T_5
12	16	11	15	17
18	17	8	16	17
16	12	9	11	14
13	10	12	14	14
		10		13
		10		

n_j	4	4	6	4	5
S_j	59	55	60	56	75
$\bar{X}_j$	14,75	13,75	10,00	14,00	15,00
V_j	7,58	10,92	2,00	4,67	3,50

b) Bedømmelse av 4 iskremprøver med 5 dommere

Dommer	A	B	C	D	SUM
1	3,5	2,5	4,0	3,5	13,5
2	4,0	2,0	4,0	4,0	14,0
3	3,5	3,0	4,5	3,0	14,0
4	4,0	2,0	2,5	3,0	11,5
5	4,0	2,5	4,0	3,0	13,5
Sum	19,0	12,0	19,0	16,5	66,5

2-veis gruppering

Som eksempel på et opplegg av denne typen, har jeg tatt en bedømmelse av iskrem foretatt av 5 dommere. Vi må her ta visse forbehold fordi det kan diskuteres om verdiene for bedømmelsene er diskrete eller kontinuerlige. Da skalaen bare dekker 6 verdier, er det klart at fordelingen ikke kan være normal. SNEDECOR OG COCHAN har diskutert skalaer med begrensede verdier, og anbefalt at t-test og dermed trolig også F-testet kan brukes med en viss korrigering for diskontinuitet, uten at vi skal diskutere dette noe nærmere.

I vårt tilfelle kan vi betrakte dommerne som "blokker", hvor da hver enkelt dommer burde fått prøvene i tilfeldig rekkefølge og uavhengig av rekkefølgen for de andre dommerne. Dette er imidlertid upraktisk i en bedømmelsessituasjon, og virkningen kan diskuteres i hvert enkelt tilfelle.

Vi kan nå tenke oss at vi har k alternativer av forsøksfaktoren (iskremprøvene) bedømt av n dommere. Vi har da

totalt N = nk observasjoner, f.eks. bedømmelse av smak.
Det kan nå beregnes tre gjennomsnitt.

$\bar{X} = S/N$ (totalgjennomsnittet)
$\bar{X}_j = S_j/n$ (gjennomsnittene for prøvene)
$\bar{X}_i = S_i/k$ (" for blokker)

Det kan nå vises at

$$\Sigma\Sigma(X_{ij} - \bar{X})^2 = k\Sigma(\bar{X}_i - \bar{X})^2 + n\Sigma(\bar{X}_j - \bar{X})^2 + \Sigma\Sigma(X_{ij} - \bar{X}_j - \bar{X}_i + \bar{X})^2$$

og av disse kan vi beregne 3 varianser:

$$V_{Blokk} = \frac{k}{n-1} \Sigma(\bar{X}_i - \bar{X})^2$$

$$V_{Prøver} = \frac{n}{k-1} \Sigma(\bar{X}_j - \bar{X})^2$$

$$V_{Rest} = \frac{1}{(n-1)(k-1)} \Sigma\Sigma(X_{ij} - \bar{X}_j - \bar{X}_i + \bar{X})^2$$

med frihetsgrader henholdsvis (n-1), (k-1) og (n-1)(k-1)

Vanligvis beregnes kvadratsummen ved følgende formler:

$$\Sigma\Sigma(X_{ij} - \bar{X})^2 = \Sigma\Sigma X_{ij}^2 - S^2/N$$
$$k\Sigma(\bar{X}_i - \bar{X})^2 = \frac{1}{k}\Sigma S_i^2 - S^2/N$$
$$n\Sigma(\bar{X}_j - \bar{X})^2 = \frac{1}{n}\Sigma S_j^2 - S^2/N$$

Vi stiller etter dette opp en 0-hypotese om at

$$\mu_1 = \mu_2 = \mu_k$$

som kan testes ved F-testet.

Det resultatet vi kommer fram til, kan vel helst betraktes som gyldig for de dommerne vi har brukt, dvs. dommerne er

et tilfeldig utvalg. Hvorvidt resultatet kan gis mer generell gyldighet, vil være et spørsmål som er åpent for diskusjon.

Variansanalysen kan utvides til å omfatte flere faktorer uten at selve utregningsmåten blir vesentlig endret. En kan da få fram mulige krysseffekter mellom de variable. Dette krever at faktorene er såkalt ortogonale. Ved testing av flere varianser mot samme varians for rest, vil F-verdiene bli interkorrelert. OTTESTAD har kommet fram til en enkel metodikk for korrigering av dette, uten at vi skal komme nærmere inn på dette her.

REGRESJON

Hittil har vi bare sett på en observasjon i et sampel. Vi skal nå se litt på to variable og hvordan en av disse kan være avhengig av den andre.

I tabell A har jeg gjengitt en rekke observasjoner for prosent protein i melk og prosent fett, formoltiter og optisk tetthet. Ved å betrakte tallene nærmere, kan vi se at at det er en viss samvariasjon, og det skulle bl.a. indikere at det burde foreligge en mulighet til å beregne prosent protein ut fra en av de øvrige observasjonene for samme prøve.

I det følgende skal vi bare se på såkalt lineær regresjon, enda det finnes selvsagt en rekke andre modeller som kan beskrive samvariasjon. Vi betegner nå den observerte uavhengige variable med **X** og den avhengige som **Y**. Vi kan da tenke oss at universet av X-er er delt opp i underuniverser hvor hvert har en konstant X. Hvis det er sammenheng mellom X og Y, vil forventningen for Y være avhengig av X og dermed være en funksjon av X.

$$E(Y/UX) = f(X)$$

Denne funksjonen kaller vi for regresjonsfunksjonen for Y med hensyn på X.

Tilsvarende kan vi dele universet etter Y-verdien og får:

$$E(X/UY) = g(Y)$$

I lineær regresjon gjør vi tre forutsetninger for samvariasjon mellom Y og X:

1. For hver X er det en normal fordeling av Y hvor den observerte Y er tilfeldig, tatt ut. Vi kan ha flere Y-er for hver X.

2. Universet av Y-er for hver X har et middel, μ, som ligger på linjen

$$\mu = \alpha + \beta(X - \bar{X}) = \alpha + \beta x$$

hvor da α og β er parametre.

3. Standardavviket for Y rundt middelet $(\alpha + \beta x)$ er konstant selv om X varierer.

Den matematiske modellen er da

$$Y = \alpha + \beta x + \epsilon$$

hvor ϵ er en tilfeldig variabel $N(0, \sigma_{x.y})$ α er gjennomsnittet i universet når X = 0. β er stigningsforholdet for regresjonslinja (forandringen i Y pr. økning i x.) ϵ er da den tilfeldige variasjonen regresjonslinja med forventning 0.

$$\epsilon = Y - \hat{\alpha} - \hat{\beta}x$$

og gir feilen for regresjonslinja for paret (X, Y) når α og $\hat{\beta}$ står for estimater av α og β. Ved minste kvadraters metode kan en finne minimum for summen av kvadratene av feilene.

$$\Sigma(Y-\hat{\alpha}-\hat{\beta}x)^2 = \Sigma(Y-\hat{\alpha}) - \hat{\beta}x)^2 =$$
$$\Sigma(Y^2-2Y\hat{\alpha}+\alpha^2-2Y\hat{\beta}x+2\hat{\alpha}x\ \hat{\beta}x^2)$$

Partiell derivering med hensyn på β, gir:

$$\Sigma(2Yx+2\hat{\alpha}x+2\hat{\beta}x^2)$$
$$\Sigma(Yx+\hat{\alpha}x+\hat{\beta}x^2) =$$
$$\Sigma Yx+\hat{\alpha}\Sigma x+\hat{\beta}\Sigma x^2 =$$
$$\Sigma Yx+\hat{\beta}\Sigma x^2 = 0$$

$$\Sigma Yx = \Sigma(Y-\bar{Y})+\bar{Y})x = \Sigma((Y-\bar{Y})x+\bar{Y}\Sigma x = \Sigma yx$$

$$\hat{\beta} = \frac{\Sigma xy}{\Sigma x^2}$$

Partiell derivering m.h. p. $\hat{\alpha}$:

$$\Sigma(-2Y + 2\hat{\alpha} + 2\hat{\beta}x)$$

$$\Sigma(\hat{\alpha} + \hat{\beta}x - Y), \qquad \Sigma x = 0$$

$$\Sigma - \Sigma Y =$$

$$n\hat{\alpha} - \Sigma Y = 0$$

$$\hat{\alpha} = \bar{Y}$$

Etter dette er $\hat{Y}$ for hver x:

$$\hat{Y} = \bar{Y} - \frac{\Sigma xy}{\Sigma x^2}\, x$$

Avvikene for hver observert Y blir:

$$Y - \hat{Y} = Y - \bar{Y} - bX = y - bx$$

idet $\hat{\beta}$ settes lik b og $\alpha = \bar{Y}$

Da er

$$\Sigma(y-bx) = 0$$

og $\Sigma(y-bx)^2 =$

$$\Sigma y^2 - \Sigma 2byx + \Sigma b^2 x^2 =$$

$$\Sigma y^2 - 2b\Sigma xy + b^2 \Sigma x^2 =$$

$$\Sigma y^2 - \frac{2(\Sigma xy)^2}{\Sigma x^2} + \frac{(\Sigma xy)^2}{\Sigma x^2} =$$

$$\Sigma y^2 - \frac{(\Sigma xy)^2}{\Sigma x^2} = \Sigma d_{y.x}^2$$

Middelkvadrat-avviket fra regresjonen er:

$$s_{y.x}^2 = \Sigma d_{y.x}^2/(n-2)$$

hvor (n-2) er antall frihetsgrader og sampel standard avviket fra regresjonen

$$s_{y.x} = \sqrt{s_{y.x}^2}$$

Sampelstandard avviket for regresjonskoeffisenten er

$$s_b = s_{y.x}\sqrt{\Sigma x^2}$$

og denne kan testes for signifikans ved at

$$t = (b-\beta)/s_b \text{ med } (n-2) \text{ frihetsgrader.}$$

En kan da også beregne konfidensinterval for β, idet

$$b-t_\alpha s_b \leq \beta \leq b+t_2 s_b$$

Feilen ved å bruke estimatet

$$\hat{Y} = \hat{Y}+bx$$

i stedet for

$$\mu = \alpha + bx$$

vil være

$$\hat{Y}-\mu = (\hat{Y}-\alpha)+(b-\beta)x$$

Nå er

$$Y=\alpha+\beta x+\varepsilon$$

Divideres med n og summeres, får vi:

$$\bar{Y} = \alpha+\varepsilon$$

Da er

$$\bar{Y}-\mu=\hat{\varepsilon}+(b-\beta)x$$

Her har ε variansen $\sigma_{y.x}^2/n$

og b er fordelt om β med variansen $\sigma_{y.x}^2/\Sigma x^2$

Da er $\sigma_{\hat{Y}}^2 = \sigma_{y.x}^2(\frac{1}{n}+\frac{x^2}{\Sigma x^2})$

og sampel standard avviket for $\hat{Y}$

$$S_{\hat{Y}} = S_{y.x} \sqrt{(1/n) + (x^2/\Sigma x^2)}$$

Hvis vi ønsker å estimere en $\hat{Y}$ ved en bestemt X-verdi, vil

$$S_{\hat{Y}} = S_{y.x} \sqrt{1 + (1/n) + (x^2/\Sigma x^2)}$$

KORRELASJON

Korrelasjonskoeffisient er også et mål for en mulig sammenheng mellom 2 variable. Den defineres ved

$$r = \frac{\frac{\Sigma x_1 x_2}{n-1}}{\sqrt{\frac{\Sigma(x_1^2)\ (\Sigma(x_2))^2}{(n-1)\ (n-1)}}} = \frac{\Sigma x_1 x_2}{\sqrt{\Sigma x_1^2 . \Sigma x_2^2}}$$

$\Sigma(x_1 x_2)/(n-1)$ kalles for covariansen, og vi ser at r er kvotienten mellom covariansen og de to sampel standardavvikene for x_1 og x_2.

r er en dimensjonsløs størrelse idet teller og nevner har samme dimensjon.

Regresjonskoeffisientene når x_1 er avhengig variabel er $\Sigma(x_1 x_2)/\Sigma x_2^2$ og når x_2 er avhengig variable $\Sigma(x_1 x_2)/\Sigma x_1^2$. Sammenhengen mellom samplets regresjons- og korrelasjonskoeffisienter er derfor:

$$b_{12} \cdot b_{21} = r^2$$

Vi har tidligere sett at

$$\Sigma d_{y.x}^2 = \Sigma y^2 - (\Sigma xy)^2/\Sigma x^2$$

Fra formelen for r ser vi at

$$(\Sigma xy)^2 = r^2 \Sigma x^2 \Sigma y^2$$

og derfor er

$$\Sigma d_{y.x}^2 = (1-r^2)\Sigma y^2$$

Her må begge sider av likhetstegnet være positive, og dette betyr at $r^2 \leq 1$ eller $-1 \leq r \leq +1$.

Av denne formelen kan vi videre se at vi har fått redusert summen av kvadratavviket med $r^2 \Sigma y^2$ ved å innføre en forklaringsvariabel. Multipliserer vi r^2 med 100 så får vi % reduksjon i summen av kvadratavvikene.

Universets korrelasjonskoeffisient betegnes ρ og er definert ved

$$\rho = \text{cov}\,(X_1 X_2)/\sigma_1\,\sigma_2$$

Testing av hypotesen $\rho = 0$ kan gjøres ved å bruke t-tabellen i det

$$t = b_{y.x}/s_b = r\sqrt{n-2}/\sqrt{1-r^2}$$

med (n-2) frihetsgrad. For andre hypoteser om ρ kan vi ikke bruke t-testet.

FISHER har foreslått at en tranformerer r til en størrelse z etter

$$z = \tfrac{1}{2}[\log_e(1+r) - \log_e(1-r)]$$

som da er tilnærmet normalfordelt med

$$\sigma_z = \frac{1}{\sqrt{n-3}}$$

Det er da utarbeidet tabeller over z for forskjellige r-verdier. De kritiske verdiene for valgte forkastningsnivå tas ut av tabell for standard normal fordeling og:

$$z_r - (z_{0,01})\sigma_z = z = z_r + (z_{0,01})\sigma_z$$

og de grensene vi finner for z overføres til ρ ved tabell.

Skal vi teste hyposten om at to sampel korrelasjonskoeffisienter er tatt tilfeldig fra samme univers, så regner vi først ut differansen i z-verdiene for begge r.

Så beregner vi middelavviket som er

$$d_{z_1-z_2} = \sqrt{\sigma_{z_1}^2 + \sigma_{z_2}^2}$$

og dividerer

$$d_{z_1-z_2} / \sqrt{\sigma_{z_1}^2 + \sigma_{z_2}^2}$$

Sannsynligheten for den verdien vi da finner eller større verdier finner vi i tabell over standard normal fordelingen.

Vi har fra før at

$$s_y^2 = \frac{\Sigma y^2}{n-1}$$

hvor y er observasjoner gitt i avviksvariable.

Nå er

$$s_{y.x}^2 = \frac{(1-r^2)\Sigma y^2}{n-2}$$

Da er

$$\frac{s_{y.x}^2}{s_y^2} = \frac{(1-r^2)\ (n-1)}{(n-2)} \approx 1-r^2$$

når n er et tilstrekkelig stort tall.

Vi ser her at r^2 er den delen av s_y^2 som er forklart ved regresjonen, mens $(1-r^2)$ er den delen som ikke kan forklares av regresjonen. Den samme definisjonen gjelder for andre former for sammenhenger med en eller flere uavhengig variable. Korrelasjonskoeffisienten kalles da for multipel og betegnes med r^2. Med flere uavhengig variable har vi også korrelasjons-koeffisienter mellom den avhengig variable og hver og en av de uavhengige i det vi antar at de andre uavhengige holdes konstant. Disse kalles for partielle korrelasjonskoeffisienter.

RANGERINGSTEST

To rangeringer

Problemstillingen kan her være at en skal rangere to egenskaper og finne ut om det er sammenheng mellom dem. Det kan også være at en er interessert i om to dommere rangerer en egenskap forskjellig. Det kan endelig også være at en er interessert i hvordan en dommer reproduserer sitt rangeringsresultat for to identiske serier.

Som eksempel kan vi se på den siste problemstillingen i det jeg har latt en rekke studenter rangere 7 forskjellige pølser tilsatt soya-protein. En tilfeldig, utvalgt rangerte pølsene på denne måten ved to rangeringer:

Pølse kode:	A	B	C	D	E	F	G
1. rangering	7	3	6	1	5	4	2
2. "	1	4	3	2	7	6	5

Hvordan skal vi så finne ut om vedkommende student var flink til å reprodusere sin rangering.

Pølsene var selvsagt kodet forskjellig ved de to rangeringene.

Generelt kan vi la hvert par i den første rekka få en x-score, a_{ij}, hvor den eneste betingelsen er at $a_{ij} = -a_{ji}$. Hvert par i den andre rekka får en y-score, b_{ij}, hvor betingelsen er at $b_{ij} = -b_{ji}$. Det defineres så en generallisert korrelasjon koeffisient r hvor

$$r = \frac{\Sigma a_{ij}\, b_{ij}}{\sqrt{a_{ij}^2 \cdot b_{ij}^2}}$$

KENDALL's KORRELASJONS KOEFFISIENT

Utgangspunktet ved denne metoden er at et par får +1 hvis $p_i > p_j$ hvor da p_i er rangeringsnummeret for prøven i første rangering, og -1 hvis $p_i < p_j$.

Altså er $a_{ij} = +1$ når $p_i > p_j$

$a_{ij} = -1$ " $p_i < p_j$

og tilsvarende for b_{ij}. Hvis nå paret er rangert likt i de to rangeringene, vil produktet $a_{ij}b_{ij}$ bli +1, i motsatt tilfelle blir produktet -1.

Lager vi en sum, S, ved å summere antall par som har lik rekkefølge i de to rangeringene og trekke fra antall par som har motsatt rekkefølge, vil vi få at

$$2\,S = \Sigma a_{ij}\, b_{ij}$$

fordi hvert par under summetegnet er tatt med to ganger ($a_{ij}\, b_{ij}$ og $a_{ji}\, b_{ji}$). $a_{ij}^2 = b_{ij}^2 = n(n-1)$ fordi summene bare er summen av de 2 n parene, i det kvadratet av hvert element er +1.

Da blir

$$\Gamma = \frac{2\,S}{n(n-1)} = \tau = \frac{S}{\binom{n}{2}}$$

som kalles for KENDALL's rangeringskoeffisient. S kan variere mellom + og $-\binom{n}{2}$ og τ derfor mellom 1 og -1.

For enhelhets skyld ordner vi 1. rangering fra 1 til n, og får for vårt eksempel:

Pølse kode.	D	G	B	F	E	C	A
1. rangering	1	2	3	4	5	6	7
2. "	2	5	4	6	7	3	1

Vi ser nå først på prøve D. I første rangering var den rangert foran alle. I andre rangering var den rangert som nr. 2.

Alle tall som er høyere en D's rangeringsnummer til høyre for D, representerer da prøver som er dårligere enn D ved parvis sammenlikning og vil få +1 i score, mens tall mindre enn D til høyre for D vil gi -1 i score, fordi 2. rangering for paret da er for-

skjellig fra 1. rangering. Slik kan vi gå videre til prøve G og ved samme **resonement** finner vi antall par som er likt og ulikt rangert i de to seriene. Dette kan settes opp slik:

Pølse kode:	D	G	B	F	E	C	A	
Likt rangert (+1)	5	2	2	1	0	0	0	10
Ulikt rangert (-1)	1	3	2	2	2	1	0	11
							S =	-1

$$\tau = \frac{-2}{7 \cdot 6} = n\frac{-2}{42} = -0,05$$

som selvsagt ikke er signifikant på 5 % nivået.

SPEARMAN's KORRELASJONSKOFFISIENT

Vi lar nå

$$a_{ij} = p_i - p_j$$

og

$$b_{ij} = q_1 - q_j$$

hvor p_i og q_i er rangeringsnummeret for den ite prøven i henholdsvis 1. og 2. rangering. Både p_i og q_i går fra 1 til n og derfor er

$$\Sigma(p_i - p_j)^2 = \Sigma(q_i - q_j)^2$$

Da er

$$r = \frac{\Sigma p_i - p_j)(q_i - q_j)}{\Sigma(p_i - p_j)^2}$$

$$\sum_{i,j=1}^{n} (p_i - p_j)(q_i - q_j) = \sum_{i=1}^{n}\sum_{j=1}^{n} p_i q_i + \sum_{i=1}^{n}\sum_{j=1}^{n} p_j q_j$$

$$\sum_{i=1}^{n}\sum_{j=1}^{n} (p_i q_j + p_j q_i) = 2n \sum_{i=1}^{n} p_i q_i - 2 \sum_{i=1}^{n} p_i \sum_{j=1}^{n} q_j =$$

$$2n \sum_{i=1}^{n} p_i q_i - \tfrac{1}{2}n^2(n+1)^2$$

$(p_i - q_i)$ er differensen, d, mellom rangeringsnummerene for den ite prøven i de to rangeringene og

$$S(d^2) = \sum_{i=1}^{n} (p_i - q_i)^2 = 2 \Sigma p_i^2 - 2 \Sigma p_i q_i$$

Vi får da at:

$$\Sigma(p_i - p_j)(q_i - q_j) = 2n\Sigma p_i^2 - \tfrac{1}{2}n^2(n+1)^2 - nS(d^2)$$

men $$\sum_1^n p_i^2 = \frac{1}{6}(n+1)(2n+1)n$$

og

$$\Sigma(p_i-p_j)(q_i-q_j) = \tfrac{1}{6}n^2(n^2-1) - nS(d^2)$$

Nevneren for r er:

$$\Sigma(p_i-p_j)^2 = 2n\Sigma p_i^2 - 2\,\Sigma p_i p_j =$$

$$2n\,\Sigma p_i^2 - 2(\Sigma p_i)^2 =$$

$$\frac{2}{6}n^2(n+1)(2n+1) - \frac{2}{4}(n+1)^2n^2 =$$

$$\frac{n^2}{3}(2n^2+2n+n+1) - \frac{n^2}{2}(n^2+2n+1) =$$

$$\frac{4n^4}{6} + \frac{6n}{6} + \frac{2n^2}{6} - \frac{3n^4}{6} - \frac{6n^3}{6} - \frac{3n^2}{6} =$$

$$\frac{n^4}{6} - \frac{n^2}{6} = \frac{1}{6}(n^2-1)n^2$$

og

$$r = \frac{\frac{1}{6}n^2(n^2-1) - nS(d^2)}{\frac{1}{6}n^2(n^2-1)} =$$

$$1 - \frac{6S(d^2)}{n^3-n} = \rho$$

Som et eksempel på beregning av SPEARMAN's korrelasjonskoeffisient ta vi en annen students resultat ved gjentatt rangering av 7 pølser.

Pølse, kode	A	B	C	D	E	F	G
1. rangering	3	5	7	2	1	6	4
2. rangering	2	6	7	3	1	5	4
differense (d)	1	-1	0	-1	0	1	0

$$\rho = 1 - \frac{6 \cdot 4}{7(49-1)} = 1 - \frac{24}{7 \cdot 48} = 0{,}9286$$

som etter tabellen over fordelingen av ρ er signifikant på 5 % nivået

Sammenheng mellom flere rangeringer.

Vi antar at n objekter rangeres etter én egenskap av m dommere.

Objekt	1	2	n
1. dommer	r_{11}	r_{12}	r_{1n}
i. dommer	r_{i1}	r_{i2}	r_{in}
m. dommer	r_{m1}	r_{m2}	r_{mn}

Hvis alle dommerne rangerer på samme måten, vil summen for de enkelte objektene bli m, 2m ..... nm. Alle summene blir $m(1+2+....n) = m.n(n+1)/2$ og middelet $m(n+1)/2$.

Avvikene fra gjennomsnittet blir:

$m-m(n+1)/2 = -m(n-1)/2$

$2m-m(n+1)/2 = -m(n-3)/2$

$mn-m(n+1)/2 = m(n-1)/2$

Kvadratsummen av avvikene blir:

$$m^2(n^3-n)/12$$

som er den største verdien denne kan få. For observasjonene vil en finne en kvadratsum = S, og KENDALL definerte da

$$W = S/(m^2(n^3-n)/12) = 12\ S/m^2(n^3-n)$$

W vil ligge mellom 0 og 1.

For alle studentene i bedømmelsen av pølser, fikk vi dette resultatet for første gangs rangering.

Pølse kode:	A	B	C	D	E	F	G	
Student 1.	3	6	7	4	1	2	5	
" 2.	2	6	7	5	1	4	3	
" 3	6	2	7	4	1	5	3	
" 4	4	6	5	2	1	7	3	
" 5	2	6	7	4	3	5	1	
" 6	4	6	2	3	1	7	5	
" 7	3	6	7	1	2	4	5	
" 8	2	7	4	6	1	5	3	
" 9	2	5	7	3	1	6	4	
" 10	6	7	3	2	4	1	5	
" 11	5	6	2	7	1	4	3	
" 12	3	5	7	6	2	1	4	
" 13	7	3	6	1	5	4	2	
" 14	3	6	7	1	2	5	4	
" 15	6	5	2	4	7	1	3	
" 16	4	7	5	6	1	2	3	
" 17	3	6	7	1	2	4	5	
" 18	3	5	7	2	1	6	4	
Sum	68	100	99	62	37	73	65	504

Summen kan kontrolleres ved

$$(18 \cdot 7 \cdot 8)/2 = 504$$

Gjennomsnittet er: $(18 \cdot 8)/2 = 72$

Summen av kvadratavvikene blir:

$$(72-68)^2 + (72-100)^2 + \ldots . (72-65)^2 = 2904$$

$$W = 12 \cdot 2904/18^2(7^3-7) = 0,3201$$

Denne verdien kan testes ved

$$F = \frac{(m-1)W}{1-W} = \frac{17 \cdot 0,3201}{0,6799} = 8,00$$

med $(n-1-2/m) \approx 6$ og $(m-1)(n-1-2/m) \approx 102$ frihetsgrader.

F-verdien er derfor signifikant på 1 % nivået.

SCHEFFÉ-TEST

SCHEFFÉ har utarbeidet en variansanalyse for parvise test, hvor dommeren bare skal gi en verbal uttalelse om graden av forskjell mellom prøvene. Testet egner seg trolig godt i forbrukeranalyse fordi fremgangsmåten er grei å forstå.

Vi har også brukt SCHEFFE-testet ganske mye ved bedømmelse av produkter. Forutsetningen er da at vi har et relativt lite antall produkter som skal bedømmes, ellers vil bedømmelsen ta lang tid og også kreve større mengder av hvert produkt.

Antall par vokser naturligvis sterkt med antall prøver idet det mulige antallet, M, kan beregnes ved:

$$M = \tfrac{1}{2}m(m-1)$$

hvor m er antall prøver. Da prøvene også bør bedømmes i begge rekkefølger vil f.eks. 6 prøver kreve 30 par-bedømmelser.

Vi kan forøvrig gå rett på et eksempel, også hentet fra kurset i sensorisk kvalitetskontroll hvor studentene skulle bedømme 4 øltyper mot hverandre.

Testingssjemaet fremgår av figur 1 hvor da prøvene alltid hadde betegnelsen A og B i de 12 parene som ble servert i tilfeldig rekkefølge.

Resultatene for hver dommer ble samlet i et skjema som tabell (2)og (ij I tabellen har prøvene fått nummer 1-4, bedømmelsesrekkefølge (i,j) og enkelte dommerresultat x_{ijk} er summert for hvert utsagn som nå har blitt overført til tall. "A foretrekkes sterkt framfor B" gis 3 poeng osv. til "A foretrekkes sterkt framfor A" som får 3 poeng.

Totalvariasjonen er:

$$S_t = 3^2(18+4)+2^2(43+21)+1^2(28+30) = 512$$

med $2rM = 2 \cdot 14 \cdot 6$ frihetsgrader

(r=antall dommere, M=antall par).

Den 1. prøvens gjennomsnitts preferense over den andre er gitt ved

$$\hat{\mu}_{ij} = \sum_{k=1}^{r} X_{ijk}/r$$

Variasjonen som skyldes forskjell i preferanse er

$$S_\mu = r \sum_{i=1}^{m} \sum_{i=j}^{m} \hat{\mu}_{ij}^2$$

$$S_\mu = [14\ 0{,}501^2 + 1{,}143^2 + \ldots (-0{,}713^2] = 153{,}7$$

med $2\ M = 2 \cdot 6$ frihetsgrader.

Kvadratsummen for feilen blir:

$$S_e = \sum_{i=1}^{m} \sum_{j=1}^{m} \sum_{k=1}^{r} (X_{ijk} - \hat{\mu}_{ij})^2$$

og det kan vises at

$$S_e = S_t - S_\mu$$

$$S_e = 512 - 153{,}7 = 358{,}3$$

Hvis dommerne hadde vært konsistente i sin bedømmelse ville vi ventet at

$$\hat{\mu}_{ij} = - \hat{\mu}_{ji}$$

Når dette ikke alltid slår til, kan det komme av tilfeldigheter, men det kan også skyldes at serveringsrekkefølgen har hatt en systematisk betydning.

Vi kan da for det første regne ut den gjennomsnittlige preferanse i over j for de to parene (i,j) og (j,i).

Denne er :

$$\hat{\Pi}_{ij} = \tfrac{1}{2}(\hat{\mu}_{ij} - \hat{\mu}_{ji})$$

og kvadratsummen for denne er:

$$S_\Pi = 2r \sum_{i<j} \hat{\Pi}_{ij}$$

$$S_\Pi = 2 \cdot 14 [(-0{,}321)^2 + (0{,}358)^2 + \ldots 1{,}250^2)] = 107{,}5$$

med $M = 6$ frihetsgrader.

Som nevnt kan vi vente at

$$\hat{\mu}_{ij} = -\hat{\mu}_{ji} \quad \text{eller}$$
$$\hat{\mu}_{ij} + \hat{\mu}_{ji} = 0$$

Avviket fra dette kan skyldes en mulig effekt av serveringsrekkefølgen, δ_{ij}.

Denne estineres ved

$$\hat{\delta}_{ij} = \tfrac{1}{2}(\hat{\mu}_{ij} + \hat{\mu}_{ji})$$

og kvadratsummen

$$S_\delta = 2r\sum_{i<j} \delta_{ij}^2$$

Det kan da vises at

$$S_\delta = S_\mu - S_\pi$$

For eksempelet blir dette

$$S_\delta = 153{,}7 - 107{,}5 = 46{,}2$$

med M = 6 frihetsgrader.

Nå tenker vi oss at det finnes parametre $\alpha_1, \alpha_2 \ldots \alpha_m$ som karakteriser de m prøvene slik at den gjennomsnittlige preferensen kan forklares som en diferens mellom de tilsvarende parametrene.

$$\pi_{ij} = \alpha_i - \alpha_j$$

hvor da

$$\hat{\alpha}_i = (\sum_{j=1}^{m} \pi_{ij})/m$$

og $$\sum \hat{\alpha}_i = 0$$

Kvadratsummen for α-ene blir:

$$S_\alpha = 2rm\sum_{i=1}^{m} \hat{\alpha}_i^2$$

For eksempel blir:

$$\hat{\alpha}_1 = (-0,321 - 0,358 + 0,572)/4 = -0,027$$

$$\hat{\alpha}_2 = (0,321 - 1,143 + 0,643)/4 = -0,045$$

$$\hat{\alpha}_3 = (0,358 + 1,143 + 1,250)/4 = 0,688$$

$$\hat{\alpha}_4 = (-0,572 - 0,643 - 1,250)/4 = -0,616$$

og derfor

$$S_\alpha = 2 \cdot 14 \cdot 4 \left[-0,027^2 + (-0,045)^2 + (-0,616)^2\right] = 95,9$$

med m-1 = 3 frihetsgrader.

Endelig kan det forekomme at den gjennomsnittlige preferensen π_{ij} ikke helt stemmer med differensen mellom de tilsvarende α-ene. Forskjellen her kaller en for avviket fra subtraktivitet og den betegnes med γ_{ij}.

$$\pi_{ij} = \alpha_i - \alpha_j + \gamma_{ij}$$

$$\hat{\gamma}_{ij} = \hat{\pi}_{ij} - \alpha_i + \alpha_j$$

og

$$S_\gamma = 2r \sum_{i<j} \hat{\gamma}_{ij}$$

Det kan da vises at

$$S_\gamma = S_\pi - S_\alpha$$

om i vårt eksempel blir:

$$S_\gamma = 107,5 - 95,9 = 11,6$$

med M-m-1 = 3 frihetsgrader.

Variasjonsanalysen er satt opp i tabell 3.

For å finne ut hvilke av α-ene som er forskjellig fra de andre beregner en ved en modifikasjon av TUKEY's HONESTHY SIGNIFICANT DIFFERENCE test følgende målestokk.

$$\gamma_\varepsilon = q_{1-\varepsilon} \sqrt{S^2/2rm}$$

hvor ε = forkastningsnivået.

q_1-ε finnes i tabell over studentisert variasjonsvidde for m=4 prøver og for feilens frihetsgrader (156)

S^2 = middelkvadrat for feilen

r = antall prøver

På 5 % nivået får vi:

$$\gamma_{0,05} = \pm 3{,}63\sqrt{2{,}297/2 \cdot 14{,}4} = \pm 0{,}520$$

og differensene α-ene imellom kan sammenliknes med denne verdien

Her ble:

$\alpha_1 - \alpha_2 = 0{,}018$		/ ingen signifikantforskjell /
$\alpha_1 - \alpha_3 = -0{,}715$		/ 3 bedre enn 1/
$\alpha_1 - \alpha_4 = 0{,}589$		/ 1 bedre enn 4/
$\alpha_2 - \alpha_3 = 0{,}733$		/ 3 bedre enn 2/
$\alpha_2 - \alpha_4 = 0{,}571$		/ 2 bedre enn 4/
$\alpha_3 - \alpha_4 = 1{,}304$		/3 bedre enn 4/

alt på 5 % nivået.

Konklusjon:

3 ble preferert før de øvrige

4 ble bedømt som dårligere enn de andre. (Dette var Brigg).

Det er også av interesse å merke seg at rekkefølgeeffekten var signifikant i det det ølet en smakte på først fikk bedre bedømmelse enn en kunne vente.

CHEFFÉ - TEST

Tabell 1.

Navn: .....

Du får utdelt et par prøver, 12 ganger. Den ene av prøvene i hvert par er A, den andre B. Du skal smake på prøvene og bestemme deg for et av de følgende svarene og sette kryss rett ut for dette:

	1	2	3	4	5	6	7	8	9	10	11	12
A foretrekkes sterkt framfor B												
A " moderat " B												
A " svakt " B												
A og B er likeverdige												
B foretrekkes svakt framfor A												
B " moderat " A												
B " sterkt " A												

Tabell 3. ANALYSESKJEMA

Årsak	Kvadratsum	Frihetsgrad	Middelkv.	F-verdi
Hovedeffekt	95,9	3	32,0	*** 13,93
Avvik fra subtraktivitet	11,6	2	3,9	
Gjennomsnittlig preferanse	107,5	6	17,9	
Rekkefølgeeffekt	46,2	6	7,7	*** 3,35
Gjennomsnitt	153,7	12	12,81	
Feil	358,3	156	2,297	
Total	512,0	168		

Tabell 2. UTREGNINGSSKJEMA

Servering nr.	Par (i,j) (j,i)	Scoringsfrekvens (x_{ijk}) 3	2	1	9	-1	-2	-3	Total score	$\hat{\mu}_{ij}$	$\hat{\pi}_{ij}$	$\hat{\delta}_{ij}$	$\hat{\gamma}_{ij}$
11	1,2	0	3	4	5	1	1	0	7	+0,501	-0,321	+0,822	-0,339
9	2,1	2	5	3	1	3	0	0	16	+1,143			
10	1,3	1	1	3	3	4	2	0	0	0	-0,358	+0,358	+0,357
5	3,1	2	1	6	2	2	1	0	10	+0,715			
12	1,4	1	8	1	1	1	2	0	15	+1,071	+0,572	+0,500	-0,017
6	4,1	1	2	1	3	5	2	0	-1	-0,072			
1	2,3	0	2	1	1	5	5	0	-10	-0,714	-1,143	+0,429	-0,410
4	3,2	3	7	0	3	1	0	0	22	+1,572			
7	2,4	2	6	2	0	2	2	0	14	+1,000	+0,643	+0,357	-0,071
8	4,2	0	0	5	3	4	1	1	-4	-0,286			
3	3,4	4	6	2	1	1	0	0	25	+1,786	+1,250	+0,537	-0,054
2	4,3	2	2	0	1	1	5	3	-10	-0,713			
	Total	18	43	28	24	30	21	4					

VARIANSANALYSE. 2-veis gruppering.

Som et eksempel på bruk av et slikt opplegg for bedømmelse, har jeg tatt en ostebedømmelse for 10 oster av 13 dommere etter skalaen 0-5. Utregningen er rett fram. Det skal imidlertid pekes på at det i høyeste grad kan diskuteres om fordelingen er normal og hvorvidt skalaen er kontinuerlig eller diskontinuerlig.

Når resultatet, som i eksemplet, blir at det er signifikant forskjell mellom prøvene, er det selvsagt av interesse å finne ut hvilke prøver det er forskjell mellom. Ved sammenlikningen må en da ta hensyn til at en kan ha fått et stort tall eller et lite tall på grunn av slump. Sannsynligheten for dette blir større desto fler prøver en har å sammenlikne. Dette tas hensyn til i Tukey's honestly significant difference i det det tas en verdi ut fra en tabell over studentisert variasjonsvidde:

$$q = (x_n - x_1)S_v$$

og denne blir multiplisert med middelfeil. I dette tilfellet er antall prøver 10 og feilens frihetsgrader 108 slik at det tallet en kan sammenlikne differensene med på 5 % nivået, blir:

$$\pm 4{,}56 \sqrt{0{,}3472/13} = \pm 0{,}7452$$

Hvis vi imidlertid på forhånd hadde vært interessert i f.eks. differensen mellom 1 og 11, så ville

HSD = least sign.difference =

$$\pm 2{,}72\sqrt{0{,}3472/13} = \pm 0{,}4445$$

Sammenlikninger kan ellers utføres ved at en på forhånd setter opp det en kaller for ortogonale sammenlikninger, uten at en skal gå nærmere inn på dette her. SNEDECOR og COCHRAN gir en utførlig omtale av slike sammenlikninger.

Prøve

Dommer	1	2	4	6	11	13	16	17	19	20	
A	4	4	4	3	4	4	4	4	3	3	37
B	3	3	4	4	4	4	3	4	3	4	36
C	4	5	3	3	4	4	3	4	4	3	37
D	4	4	5	3	5	4	3	3	3	4	38
E	4	4	3	3	4	4	3	4	4	3	36
F	3	3	4	3	4	4	2	4	2	4	33
G	4	4	4	4	4	4	4	4	3	3	38
H	4	4	3	4	4	4	4	4	3	4	38
I	4	4	4	4	5	3	3	4	3	3	37
J	4	3	4	4	4	4	3	4	3	3	36
K	3	5	3	2	5	3	2	4	3	3	33
L	3	5	4	2	4	3	3	5	4	4	37
M	4	4	4	3	4	4	3	3	4	2	35
	48	52	49	42	55	49	40	51	42	43	471
$\bar{x}$	3,69	4,00	3,77	3,23	4,23	3,77	3,08	3,92	3,23	3,31	

C = 1706,4692

Årsak	f.gr.	kv.s.	m.kv.	F
Total	129	58,5308		
Dommere	12	3,4308		
Prøver	9	17,6077	1,9564	5,6269
Feil	108	37,4923	0,3472	

HSD = $\pm$ 4,56 $\sqrt{0,3472/13}$ = $\pm$ 0,7452

11 > 6, 16, 19, 20

2 > 6, 16, 19

17 > 16

NØDVENDIG ANTALL DOMMERE FOR AT 0,1, 0,5 og 1,0 POENG SKAL VÆRE EN SIGNIFIKANT FORSKJELL MELLOM GJENNOMSNITT (PÅ 5 % - NIVÅET).

Differense		
0,1	$n > 0,3472/0,0224^2$	$n \geq 692$
0,5	$n > 0,3472/0,1119^2$	$n \geq 28$
1,0	$n > 0,3472/0,2237^2$	$n \geq 7$

($4,47 \sqrt{0,3472/n} \leq$ 0,1, 0,5 eller 1,0)

Variansanalyse av smakspoeng for forskjellige tilsetninger til næringsmidler.

Vi tar som eksempel en nylig utført forsøk med yoghurt-iskrem, hvor vi ville undersøke effekten av fettinnhold, sukkerinnhold og innhold av fettfritt melketørrstoff på en rekke faktorer, bl.a. på smaken.

Våre teknologiske kunnskaper gir oss grunnlaget for utvelgelse av mengder (nivåer) av tilsetningsstoffene. Nivåene bør velges slik at de dekker de mulighetene en regner med er aktuelle. De bør videre være ekvidistante ,dvs. at det er lik avstand mellom tilsetningsmengdene. Dette vil lette arbeidet i den videre utregninge svært mye !

Valg av nivåer med smakspoeng som gjennomsnitt av 5 dommere, går fram av tabellen (Hvis vi hadde vært interessert i om dommerne reagerer forskjellig på tilsetningene, kunne vi regnet med dommerne som sub-plot, noe vi ikke har gjort her).

Gjennomsnittslig smakspoeng for yoghurt-is (soft) ved forskjellige tilsetningsmengder av smørfett, sukker og fettfritt melketørrstoff. 5 dommere.

Ft %	Sukker%	Smørfett				
		0	3,5	7,0	10,5	Sum Ft
8	12	2,8	3,4	3,6	3,6	
8	14	2,5	3,7	3,6	3,7	
8	16	3,2	3,8	3,2	3,8	40,9
10	12	2,5	3,3	4,1	3,8	
10	14	3,1	4,0	3,5	3,8	
10	16	2,4	2,8	2,9	3,1	39,3
12	12	2,9	3,9	4,0	3,7	
12	14	2,4	2,8	2,9	3,2	
12	16	2,5	3,1	2,7	2,8	36,9
Sum fett		24,3	30,8	30,5	31,5	117,1

Totalvariasjonen i tabellen er

$$2,8^2 + 2,5^2 + \ldots\ldots 3,2^2 + 2,8^2 - 117,1^2/36 = 9,2297$$

og korreksjonsfaktoren, $K = 117,1^2/36 = 380,9003$

Vi setter opp tre under-tabeller hvor bare to ingredienser varierer:

Ft

	Sukker%			Fett%				
	12	14	16	0	3,5	7,0	10,5	
8	13,4	13,5	14,0	8,5	10,9	10,4	11,1	40,9
10	13,7	14,4	11,2	8,0	10,1	10,5	10,7	39,3
12	14,5	11,3	11,1	7,8	9,8	9,6	9,7	36,9
	41,6	39,2	36,3	24,3	30,8	30,5	31,5	117,1

Sukker%	Fett%				
	0	3,5	7,0	10,5	
12	8,2	10,6	11,7	11,1	41,6
14	8,0	10,5	10,0	10,7	39,2
16	8,1	9,7	8,8	9,7	36,3
	24,3	30,8	30,5	31,5	117,1

En ser umiddelbart at det er mange muligheter for kontroll av utregningene !

Vi beregner så variasjonene p. gr. av det fettfrie melketørrstoffet:

$(40,9^2 + 39,3^2 + 36,9^2)/12 - K = 0,6755$

og p. gr. av sukkeret:

$(41,6^2 + 39,2^2 + 36,3^2)/12 - K = 1,1739$

Kryssvirkningen mellom ffmt og sukker:

$(13,4^2 + \ldots\ldots 11,1^2)/4 - K - 0,6755 - 1,1739 = 2,1128$

Virkningen av fett:

$(24,3^2 + \ldots 31,5^2)/9 - K = 3,7253$

Kryssvirkning ffmt og fett:

$(8,5^2 + \ldots\ldots 9,7^2) - K - 0,6755 - 3,7253 = 0,1356$

Kryssvirkning sukker og fett:

$(8,2^2 + \ldots\ldots 9,7^2)/3 - K - 1,1739 - 3,7253 = 0,7572$

Variansanalysen settes opp slik:

Årsak	Frihetsgrad	Kvadratsum	Middelkvadrat	F
Total	35	9,2297		
Ffmt	2	0,6755	0,3378	$6,24^{o}$
Sukker	2	1,1739	0,5870	$10,85^{x}$
Fett	3	3,7253	1,2418	$22,95^{xx}$
Ft x sukker	4	2,1128	0,5282	$9,76^{x}$
Ft x fett	6	0,1356		
Sukker x fett	6	0,7572		
Feil	12	0,6494	0,0541	

F-verdiene er testet etter OTTESTAD og viser at fettet er signifikant på 1 % nivået mens sukkeret og krysseffekten sukker/Ft . er signifikant på 5 % nivået.

Det kan også være av interesse å se hvordan poengene er avhengig av sukker- og fettilsetningen. Sukkeret er variert i 3 nivåer og en har her mulighet for en lineær og/eller en kvadratisk sammenheng. For å teste dette transformeres tilsetningsmengdene til henholdsvis

- 1, o og +1 for det lineære systemet og til

+ 1, - 2, + 1 for det kvadratiske systemet. Disse transformasjonene forutsetter at tilsetningsmengden er ekvidistante. Transformeringen fra x til de nye x^{1} går da i det lineære tilfellet etter

$$x^{1} = (x-\bar{x})/d$$

hvor d er avstanden mellom tilsetningsmengdene. Den transformerte variable x^{11} i det kvadratiske tilfellet er:

$$x^{11} = 3((x-\bar{x})/d)^{2} - 2$$

For ortogonialitet må videre summen av de tranformerte variable være o og summen av produktene av de tranformerte variable også 0.

Vi regner da ut produktene av de tranformerte variable og summen av poeng for hver variable:

%sukker	12	14	16	Kryssprodukt	Kodet regresjons-koef.	Kv.sum
Sum	41,6	39,2	36,3			
Lineær	-1	0	+1	-5,3	-0,2208	1,1704
Kvadratisk	+1	-2	+1	-0,5	-0,0069	0,0035
						1,1739

En ser at kvadratsummen for sukker er spaltet opp i en lineær og kvadratisk komponent, og at det bare er den lineære som har betydning. Den kodede regresjonslikningen blir da:

$$\hat{Y} = -0,2208\, x^{1}$$

og ompodet $\hat{Y} = -0,2208\,(S-14)/2 + 3,2528$

$$\hat{Y} = -0,1104S + 4,7984$$

Vi foretar samme oppspaltning for fettets virkning. Her har vi imidlertid 4 nivåer og derfor også muligheter for en kubisk effekt. De ortogonale transformasjonene blir da noe mer kompliserte:

$$x^{1} = 2(x-\bar{x})/d$$

$$x^{11} = ((x^{1})^{2} - 5)/4$$

$$x^{111} = (5(x^{1})^{3}/12) - 41x^{1}/12$$

og de ortogonale funksjonene

lineær	-3	-1	+1	+3
kvadratisk	1	-1	-1	1
kubisk	-1	+3	-1	+1

Vi får da:

% fett	0	3,5	7	10,5	Kryss produkt	Kodet regresjons koef.	Kv.sum	F
Sum	24,3	30,8	30,5	10,5				
Lineær	-3	-1	+1	+3	21,3	0,1183	2,5205	46,59
Kvad.	1	-1	-1	1	-5,5	-0,1528	0,8403	15,53
Kub.	-1	+3	-3	+1	8,1	0,0450	0,3645	6,74
							3,7253	

Her er både den lineære og den kvadratiske komponenten signifikant.

Kodet regresjonslikning blir:

$$\hat{Y} = 0,1183x^{1} - 0,1528x^{11}$$

og omkodet

$$\hat{Y} = 3,2528 + 0,1183\,[2(F-5,25)/3,5] - 0,1528\,[2(F-5,25)/3,5]^2$$

som gir:

$$\hat{Y} = -0,0125F^2 + 0,1986F + 2,7461$$

Den signifikante kryssvirkningen mellom Ft og sukker kom av at poengene ved et lavt innhold av sukker økte med innholdet av . Ved middel sukkertilsetning økte poengene fra 8 til 10 % Ft men avtok for 12 % ffmt. For 16 % sukker fikk den laveste tilsetningen av Ft høyest poeng. Forholdet ble det samme når Ft betraktes som konstant mens sukkertilsetningen varierer.

For å undersøke om effekten av Ft øker lineært eller kvadratisk med økning i sukker og omvendt, går vi til tabellen over kryssvirkningen og beregner lineær og kvadratisk effekt av Ft for hver sukkermengde.

Sukker%	12	14	16	
Lineær Ft_L	1,1	-2,2	-2,9	-4,0
Kvadratisk Ft_K	0,5	-4,0	2,7	-0,8

Ft_L	1,1	-2,2	-2,9	Sammenl.	Kvadratsum
Lineær $Ft_L.S_L$	-1	0	1	-4,0	1.0000
Kvadratisk $Ft_L.S_K$	1	-2	1	2,6	0,1408

Ft_K	0,5	-4,0	2,7		
Lineær $Ft_K S_L$	-1	0	1	2,2	0,1008
Kvadratisk $Ft_K S_K$	1	-2	1	11,2	0,8711
					2,1127

Vi finner derfor at både den dobbelte lineære komponenten $Ft_L S_L$ og den dobbelte kvadratiske $Ft_K S_K$ er signifikante. Dette betyr at forklaringen av variasjonen i smakspoengene er svært komplisert.

I en eventuell regresjonslikning bør vi ha med fettinnholdet, sukkerinnholdet, kvadratet av fettinnholdet, sukker x fettfritt m. tørrstoff og kvadratet av sukkerinnholdet x kvadratet av innholdet av fettfritt melketørrstoff. Ved omkoding av den siste størrelsen vil en også få med det fettfrie melketørrstoffet og kvadratet av dette og også kvadratet av sukkerinnholdet. Dessuten vil en få med 1. og 2. grads produktene av sukker og fettfritt melketørrstoff. Likningen blir derfor svært komplisert, og hvis en vil fram til den bør en bruke EDB.