STATISTIKK

STATISTISK BEHANDLING OG VURDERING AV DATA

Per Lea, Nofima

11.1 Innledning

Tidligere i boka har vi gjennomgått forskjellige metoder som nyttes ved sensorisk analyse. Videre har vi lært hvilke metoder og paneltyper som er mest hensiktsmessige i ulike bedømmelsessituasjoner. Når slike analyser er utført, vil en sitte igjen med et tallmateriale, men det er ikke uten videre gitt hvilke konklusjoner vi kan trekke på grunnlag av disse tallene.

Datasettet er alltid et utvalg – at vi endte opp med det datasettet vi virkelig gjorde, har et element av tilfeldighet i seg. Men vi ønsker vanligvis ikke å uttale oss om det konkrete utvalget vi har endt opp med; vi ønsker å generalisere resultatene. Hvis f.eks. et forsøk resulterer i at italiensk salat basert på råvarer fra leverandør A har mer gulrotsmak enn italiensk salat basert på råvarer fra leverandør B, så er det ikke noe interessant om dette gjelder bare for de 10 pakningene som inngikk i forsøket. Forsøksopplegget og den statistiske testen må være av en slik art at vi med en viss grad av sikkerhet kan si at leverandør A leverer råvarer som gir mer gulrotsmak på den italienske salaten over en lengre tid. Hvis vi ikke er villige til å generalisere ut fra det utvalget som inngår i selve forsøket, får vi en helt absurd situasjon: da må vi gjøre omfattende sammenlikninger hver eneste dag, og kanskje bytte leverandør flere ganger i uka. Mesteparten av det som produseres vil trolig gå med til testing, og dermed bli spist opp, slik at det nesten ikke blir noen ting igjen å selge!

Kan vi ikke generalisere resultatene av den sensoriske analysen vil det heller ikke være noen grunn til å utføre den!

11.2 Gangen i en statistisk test

En hypotese (nullhypotese, ofte symbolisert H_0 og uttalt «H-null») framsettes. Denne kan f.eks. være at italiensk salat produsert etter resept A, B og C alle har like mye gulrotsmak. Etter en sensorisk bedømmelse kan kanskje resultatet være at panelet har gitt salat A verdien 6.2 for gulrotsmak, B har fått 5,9, og C har fått 3,9. Når vi skal avgjøre om en slik hypotese kan forkastes eller ikke, må vi se på hvor sannsynlig det resultatet vi observerer i forsøket, faktisk er. Å beregne en slik sannsynlighet er ikke uten videre enkelt, og for å komme videre må vi gjøre noen antakelser. Vi antar rett og slett at nullhypotesen er sann, dvs. at de tre reseptene faktisk gir like mye gulrotsmak «i det lange løp». Har vi først antatt dette, betyr det at de forskjellene i middelverdier vi har observert, bare skyldes tilfeldigheter.

I disse middelverdiene inngår at vi har «midlet bort» dommerne, og også eventuelle gjentak og andre faktorer som inngår i forsøket. Siden vi nå har antatt at de tre reseptene gir like mye gulrotsmak, så kan vi også beregne sannsynligheten for at resultatene skal avvike så mye fra hverandre – eller mer - enn det vi har observert. At resept C har endt opp med en så lav verdi som 3,9, minst 2 sensoriske enheter mindre enn de to andre, kan synes mistenkelig. Men det er bare det statistiske forsøksopplegget som kan gi oss svar på om denne forskjellen er «stor nok». Eller sagt på en annen måte: hvor sannsynlig er det at vi får så store forskjeller som 2,3 og 2 enheter mellom de to høyeste og den laveste verdien? Hvis denne sannsynligheten er «liten», er det grunn til å tro at ett eller annet er «galt».

Og hva kan så være galt? – vi forutsetter selvsagt at alle beregninger og praktiske ting rundt forsøket har vært utført på en korrekt måte. Da gjenstår bare den antakelsen vi startet med, altså nullhypotesen. Hvis antakelsen om at de tre reseptene gir like mye gulrotsmak er lite sannsynlig, betyr det at antakelsen ikke er sann, det vil si at vi forkaster nullhypotesen. Det er det samme som at nullhypotesen ikke er sann, som i sin tur betyr at alternativet er sant. Alternativet i dette eksemplet er at minst to av reseptene er forskjellige.

Som «lite sannsynlig» er det vanlig å bruke grenseverdien 0,05 (5%), som vi kaller nivået for testen, eller signifikansnivået og ofte angir med p. Av historiske grunner bruker vi oftest p = 0,01, p = 0,05 eller p = 0,10. Det som avgjør om en nullhypotese kan forkastes er altså ikke bare et spørsmål om hvor store forskjeller vi har observert, men også om hvilket nivå vi ønsker å teste på.

Når vi først velger oss et nivå for en test, så betyr det at dette er den største sannsynligheten vi er villig til å akseptere for at vi forkaster en hypotese som er sann. Dette kalles ofte for «feil av type I», eller i engelsk litteratur: «Type I Error». Vi kan aldri være 100% sikker på at den konklusjonen vi trekker er sann. Det er selvsagt ønskelig at sannsynligheten for å trekke en feil konklusjon er så liten som mulig, men gjør vi nivået mindre, så betyr det også at vi må finne større forskjeller i middelverdiene før vi kan forkaste H_0.

For å vurdere om de forskjellene vi har observert ovenfor virkelig gir grunnlag for å forkaste H_0, må vi gjøre ytterligere endel antakelser. Slik vi har formulert situasjonen kan det kanskje være naturlig å bruke en Variansanalyse. Da forutsetter vi bl.a. at data er på en intervall- eller ratioskala (se kapittel 4.1). Uansett hvilken statistisk modell som velges, så kan vi skjematisk sette opp gangen i en hypoteseprøving slik:

1. Nullhypotesen og alternativet formuleres.

2. Signifikansnivået bestemmes.

3. Utsagnet i nullhypotesen antas å være sant.

4. Forsøket utføres.

5. Resultatene noteres, og de aktuelle beregningene gjøres.

6. Sannsynligheten for å få minst et så avvikende resultat som det vi fikk, beregnes.

7. Hvis sannsynligheten i punkt 6 er mindre enn det signifikansnivået vi bestemte i punkt 2, er det grunn til å betvile påstanden i punkt 3. Følgelig forkaster vi påstanden i punkt 1 - vi forkaster H_0 – og påstår at alternativet er sant

11.3 Forskjellstester

Forskjellstester kan deles inn i to grupper, generelle forskjellstester og spesifikke forskjellstester (Se kapittel 4.2).

Forskjellstester basert på den binomiske fordeling har den store fordelen at det er (relativt) enkle regneoperasjoner og enkel statistisk teori som ligger bak. Men alt har sin pris: disse enkle testene kan bare gi svar på enkle problemstillinger, slik som:

- Er det forskjell mellom prøve A og prøve B?
- Er prøve C søtere enn prøve D?
- Hvilken prøve foretrekker du: E eller F?

Siden et sensorisk panel ikke benyttes til å uttale seg om preferanser, er den siste problemstillinga bare av interesse i forbindelse med forbrukertester.

The Lady tasting Tea

Et eksempel av statistisk-historisk interesse er «The Lady tasting Tea», første gang referert til av R A Fisher i hans bok «The Design of Experiments» som utkom i 1935. En dame påstår hun kan kjenne på smaken hvorvidt teen eller melka ble helt opp i koppen først. Fisher diskuterer så hvordan en slik påstand skal kunne testes. Det kan jo tenkes at damen bløffer, så et eksperiment er absolutt på sin plass. I boka gir hele episoden inntrykk av å være oppkonstruert, men i en biografi publisert mye seinere, påstås det at eksemplet har sitt utspring i en virkelig hendelse. Fisher - gentleman som han var - ville være høflig og tilbød i en tepause en kvinnelig medarbeider en kopp te. «Men kjære Dem, Hr. Fisher, denne kan jeg jo ikke drikke, De har jo hatt i melken først!», svarte damen forferdet. Fisher ristet på hodet og påsto at dette kunne jo ikke damen kjenne forskjell på. Myten vil så ha det til at han gikk inn i sitt lønnkammer og skrev ned innledningskapitlet til en av de store klassikerne innen statistisk litteratur.

Damen med teen lever videre – i 2001 ga David Salsburg ut boka «The Lady Tasting Tea: How Statistics Revolutionized Science in the Twentieth Century» – en bok om statistikk og personene som la grunnlaget for statistisk teori. Mange år seinere ble for øvrig damen som ble utsatt for Fisher's lett tilslørte sjekketriks navngitt som Muriel Bristol, forlovet, og seinere gift, med en annen deltaker rundt te-bordet: William Roach. Sjekketrikset var mislykket, men den moderne statistikk var vinneren. Og frøken Roach besto testen: hun kunne virkelig kjenne forskjell!

Et vanlig sensorisk panel består som regel av så få personer at en forskjellstest blir ytterst ustabil: at bare en dommer endrer oppfatning kan snu om på hele konklusjonen. Forskjellstester brukes derfor mer i forbindelse med forbrukertester. Vanligvis vil det delta 100–200 personer i slike tester, og den n som vi seinere skal se inngår i de generelle formlene er lik antall personer som deltar i testen.

Nå kunne det vært fristende å oppnå en høyere verdi av n ved å la panelet gjøre bedømmelsen flere ganger, f.eks. la et panel på 12 personer gjøre en binomisk test 5 ganger for å oppnå n = 60. Denne framgangsmåten frarådes, da det får konsekvenser for den teoretiske vurderingen av testens godhet. Derfor unngår man en del teoretiske komplikasjoner ved å holde seg til prinsippet om at binomiske gjøres enten på et panel ved at hver dommer gjør en bedømmelse (n = antall dommere), eller ved at én person gjør n bedømmelser. I det siste tilfellet er det personen som sådan som vi tester.

Et alternativ kan være at vi sier en dommer har rett hvis vedkommende har identifisert rett prøve i for eksempel minst 3 av 4 gjentak. Da vil n i alle formlene fremdeles være lik antall dommere, men sannsynligheten for å gjette riktig vil avhenge av hvordan vi definerer å ha «rett». For eksempel vil en triangeltest hvor vi forlanger minst 3 av 4 rette gi oss en binomisk test med p= 1/9 =0,1111. Hvis vi i stedet vil forlange minst 4 av 5 rette, blir p= 11/243 = 0,0453.

11.3.1 Duo-trio-test

Denne testen brukes for å gi svar på spørsmålet: – Er det forskjell mellom prøve A og prøve B? (Se kapittel 4.2. Forskjellstester)

Hvis prøvene virkelig er forskjellige, vil vi forvente at mange dommere vil gjenkjenne den prøven som er ulik referansen. Hvis derimot prøvene er omtrent like, vil vi forvente at bare omtrent halvparten gjør det. Er det ingen forskjell mellom prøvene, må dommerne gjette seg til hvilken prøve som er ulik referansen, og sannsynligheten for å gjette riktig er p = ½.

Nullhypotesen er at A og B er like, alternativet er at de er forskjellige, det vil si at det er flere dommere som identifiserer riktig prøve enn det man skulle forvente ved ren gjetning. Følgelig blir:

H_0: p = ½ mot p > ½

Hvis 7 av 12 personer har identifisert korrekt, må vi beregne sannsynligheten for at vi får et slikt – eller mer ekstremt – resultat (punkt 6 i den skjematiske framstillinga av gangen i en hypoteseprøving). Vi må altså beregne sannsynligheten for at minst 7 av 12 personer skal velge A, under forutsetning av at A og B er like. Denne sannsynligheten kan vi finne ved hjelp av den binomiske fordelinga beskrevet i forrige kapittel:

$$\Pr(X \geq 7) = \sum_{i=7}^{12} \binom{12}{i} (\tfrac{1}{2})^i (1-\tfrac{1}{2})^{12-i} = \sum_{i=7}^{12} \binom{12}{i} (\tfrac{1}{2})^{12} = 0{,}3872$$

I praksis vil vi aldri beregne en slik sannsynlighet for hånd, men enten bruke et regneark eller et statistikkprogram. I Excel vil vi få fram denne sannsynligheten ved å skrive

=1-BINOM.FORDELING.N(6;12;0,5;SANN) eller:

=1-BINOM.FORDELING(6;12;0,5;SANN)

Da har vi benyttet oss av at $\Pr(\bar{A})=1-\Pr(\bar{A})$ for en vilkårlig begivenhet A. Eller med andre ord: sannsynligheten for en begivenhet (her 7 eller flere) er lik 1 minus sannsynligheten for den omvendte sannsynligheten (her: færre enn 7). Når begivenheten A betyr X≥7, blir dette: $\Pr(X \geq 7) = 1 - \Pr(X < 7) = 1 - \Pr(X \leq 6)$. Legg merke til at det motsatte av X≥7 altså er X<7, som igjen er lik X≤6 fordi X nødvendigvis er et helt tall.

I statistikkprogrammet R (som er gratis), får vi den samme sannsynligheten ved hjelp av:

sum(dbinom(7:12,12,0.5))

En sannsynlighet på 0,3872 kan definitivt ikke betraktes som «liten»

her. Det betyr at selv om 7 av 12 (altså et simpelt flertall, eller 58%) korrekt identifiserer A til å være lik referansen, så kan ikke det kalles usannsynlig. Det er derfor ingen grunn til å forkaste H_0, og vi har altså ingen grunn til å påstå at det er noen forskjell mellom de to prøvene.

Hvis vi hadde benyttet langt flere enn 12 personer, for eksempel 150, så ville det faktum at minst 58% (dvs minst 87 personer) av disse hadde valgt prøve A, hatt en sannsynlighet på bare 0,030 hvis hypotesen om at de egentlig var like, var sann.

Et alternativ til å gjøre disse beregningene er å slå opp i statistiske tabeller for å finne ut hvor mange korrekte identifikasjoner man må ha for å kunne forkaste H_0 – etter først å ha bestemt seg for nivået for testen. Slike tabeller pleier – selv om statistikkprogrammene stort sett har overtatt alt regnearbeidet – å være tilgjengelige i elementære lærebøker i statistikk.

Visste du at…?

I sannsynlighetsregning er det vanlig å snakke om begivenheter, suksesser og gunstige utfall. For andre enn statistikere kan det virke rart at man for eksempel i forbindelse med dødsfall snakker om «suksesser» for å omtale personer som døde i et medisinsk forsøk. Vi får bare akseptere at «sånn er det».....

11.3.2 Triangeltest

I likhet med duo-trio-testen har vi også her to prøver som skal sammenlignes, og hver av forsøkspersonene får tre testprøver. To av testprøvene er like, mens den tredje er forskjellig fra de to første. Siden det er 3 testprøver som blir presentert, er det nå en sannsynlighet på 1/3 for å gjette riktig.

Nullhypotesen blir:

H_0: p= ½ mot p>½

I likhet med under duo-trio-testen, vet altså forsøkslederen hva som er riktig: fasiten er kjent. For å avgjøre om det er signifikant forskjell

mellom de to prøvene må vi beregne sannsynligheten for at det resultatet vi observerte kan oppnås ved tilfeldigheter. Sagt på en annen måte: hvis f. eks. 23 av 45 personer korrekt identifiserte den testprøven som var forskjellig fra de to andre – hvor stor sannsynlighet er det for et slikt resultat hvis de to prøvene i virkeligheten er like? Denne gang blir p=1/3 i formelen, og vi får:

$$\Pr(X \geq 23) = \sum_{i=23}^{45} \binom{45}{i} \left(\frac{1}{3}\right)^{i} \left(\frac{2}{3}\right)^{45-i}$$

som vi ved hjelp av Excel, R eller et annet regneark- eller statistikkprogram finner er lik 0,0103.

Følgelig er det en liten sannsynlighet for at 23 av 45 personer skal utpeke den ene prøven som forskjellig fra referansen hvis det i virkeligheten ikke er noen forskjell. Siden den sannsynligheten vi fant er liten, er det naturlig å forkaste hypotesen om at prøvene er like.

Samme konklusjon ville vi ha fått ved å konsultere tabellen i vedlegg 16b. Der framgår det at med 45 personer er det tilstrekkelig med 21 korrekte svar for å forkaste hypotesen om at prøvene er like hvis vi ønsker nivå 0,05 for testen.

11.3.3 Partest

Ved en partest får dommeren to prøver som skal sammenlignes, for eksempel hvilken av de to som er søtest, har mest gulrotsmak, eller hvilken av de to prøvene synes du smaker best? I det etterfølgende holder vi oss til eksempel med søtsmak. En naturlig nullhypotese vil være at de to prøvene er like søte.

Partestene kan ha ensidige eller tosidige alternativ. Hvis vi bare er interessert i å vite om det er en forskjell, men ikke har noen forutinntatt mening om hvilken vei forskjellen kan gå, har vi en tosidig test: konklusjonen kan bli at A er mest søt, eller at B er mest søt.

Da sier vi at testen er tosidig, med symboler skriver vi nullhypotesen og alternativet slik:

H_0: A=B mot A≠B

Men i noen situasjoner kan det tenkes at bare det ene alternativet er av interesse. En slik situasjon kan være at vi vurderer å innføre en ny, billigere ingrediens i et produkt. Denne ingrediensen kan tenkes å ha innflytelse på søtheten. Hvis søtsmak er en ønsket egenskap, ønsker vi å avsløre det hvis den nye resepten resulterer i mindre søtsmak enn den gamle. I så fall vil vi ikke introdusere den billigere ingrediensen. Om den nye ingrediensen gir enda søtere smak enn den gamle er egentlig ikke så interessant, det blir nærmest bare en bonus. Det er sannsynlighetsteoretiske grunner til at vi i en slik situasjon velger en ensidig test. Det er nemlig slik at det nivået vi velger for testen, er en form for forsikring mot å gjøre feil. Ved en tosidig test kan vi komme til å påstå at A er søtere enn B når faktisk H_0 er sann, og vi kan komme til å påstå at B er søtere enn A når faktisk H_0 er sann. Siden vi ikke er interessert i alternativet B er søtere enn A, er det heller ingen grunn til å «bruke opp» noe av nivået for å gardere oss mot en slik «feil». Altså er testen ensidig, med symboler skriver vi nullhypotesen og alternativet slik:

H_0: A=B mot A>B

Alternativet er altså at A er søtere enn B. Hvis prøvene virkelig er like, vil dommernes valg være bestemt av rene tilfeldigheter, og vi vil forvente at omtrent like mange peker ut den ene som den andre prøven. Selve den statistiske testen består i å notere hvor mange som har valgt prøve A og hvor mange som har valgt prøve B, og sjekke hvor sannsynlig dette er *under forutsetning av at prøvene faktisk er like*.

Sannsynligheten for å velge den ene eller den andre prøven, forutsatt at nullhypotesen er sann (dvs. at prøvene er like), er lik ½. Hvis 7 av 12 personer har sagt at italiensk salat produsert etter resept A er best, mens 5 har foretrukket resept B, må vi beregne sannsynligheten for at vi får et slikt – eller mer ekstremt – resultat (punkt 6 i den skjematiske framstillinga av gangen i en hypoteseprøving). Vi må altså beregne sannsynligheten for at minst 7 av 12 skal velge A, under forutsetning

av at A og B er like. Denne sannsynligheten finner vi – akkurat på samme måte som for duo-trio-testen ved hjelp av:

$$\Pr(X \geq 7) = \sum_{i=7}^{12} \binom{12}{i} (\tfrac{1}{2})^i (1-\tfrac{1}{2})^{12-i} = \sum_{i=7}^{12} \binom{12}{i} (\tfrac{1}{2})^{12} = 0{,}3872$$

Konklusjonen blir også den samme: det er ingen grunn til å påstå at A er søtere enn B.

Et eksempel på en par-test hvor det er aktuelt med et tosidig alternativ, er følgende: vi har valget mellom å tilsette 2 forskjellige ingredienser og er interessert i hvilken som gir det søteste produktet. Men på forhånd vet vi ingenting om hvilken av de 2 som vil gi det søteste produktet. Det betyr at vi forkaster H_0 om at de er like, hvis *mange* dommere sier at A er søtest, eller at mange sier at B er søtest. Eller sagt på en annen måte: Vi forkaster H_0 hvis mange svarer A, eller få svarer A. Dette følger av at dommerne må svare noe, slik at antall som svarer A pluss antall som svarer B er lik antall dommere.

Anta at en partest med 15 dommere, så har 12 påstått at A er søtest, men 3 har påstått at B er søtest. Hvor sannsynlig er det at et slikt resultat skal oppstå av rene tilfeldigheter, det vil si hvis H_0 faktisk er sann? Før vi regner ut svaret, kan vi selvfølgelig sjekke med tabellen i vedlegg 16b og konkludere med at sannsynligheten for dette er mindre enn 0,05. Vil vi sjekke det, blir formelen:

$$\Pr(X \geq 12) + \Pr(X \leq 2) = \sum_{i=12}^{15} \binom{15}{i} (\tfrac{1}{2})^i (1-\tfrac{1}{2})^{15-i} + \sum_{i=0}^{2} \binom{15}{i} (\tfrac{1}{2})^i (1-\tfrac{1}{2})^{15-i}$$

$$= \sum_{i=12}^{15} \binom{15}{i} (\tfrac{1}{2})^{15} + \sum_{i=0}^{2} \binom{15}{i} (\tfrac{1}{2})^{15} = 0{,}0213$$

I tillegg til at dette bekrefter konklusjonen vi fikk ved å sjekke tabellen, får vi den tilleggsinformasjonen at vi ligger relativt godt under nivået 0,05 – konklusjonen burde derfor være klar.

11.3.4 To-av-5-test

Denne testen er en variant av triangeltesten og har fått en viss utbredelse, trolig fordi sannsynligheten for å gjette riktig hvis det ikke er noen forskjell mellom prøvene, er så liten som 0,1. Igjen er vi interessert i å sammenlikne 2 prøver, men nå får dommerne 2 testprøver av den ene prøven og 3 av den andre. Oppgaven er å finne hvilke 3 som hører sammen.

I denne testen vil vi derfor igjen bruke den binomiske fordelinga, nå med p=0,1 der vi brukte p = 1/3 i triangeltesten. Med 16 dommere og korrekt identifikasjon fra 5 av dem, får vi at

$$\Pr(X \geq 5) = \sum_{i=5}^{16} \binom{16}{i} 0{,}1^{i}(1-0{,}1)^{16-i} = 0{,}0170$$

som må regnes som en «liten» sannsynlighet, altså vil vi forkaste H_0: A = B og påstå at de er forskjellige.

I prinsippet kunne man konstruere massevis av slike n-av-m-tester. Sannsynligheten p for å gjette riktig vil da være gitt ved 1 dividert med den binomiske koeffisienten:

$$p = \frac{1}{\binom{m}{n}} = \frac{1}{\frac{m!}{(m-n)!n!}}$$

En 2-av-4-test (n = 2, m = 4) vil ha p = 1/6 = 0,1667, en 4-av-9-test (n = 4, m = 9) vil ha p = 1/126 = 0,0079.

11.4. Beskrivende tester

I en beskrivende test bedømmer dommerne produktene etter en eller flere egenskaper ved å gi dem karakterer på en skala. Vanlige skalaer er 1-7, 1-9, 0-100, ofte med muligheten til å bruke desimaler.

11.4.1 Enveis variansanalyse – urealistisk modell

En statistisk test som ofte brukes til å analysere data fra beskrivende test er slike data, er variansanalyse, ofte omtalt som ANOVA etter den engelske betegnelsen *Analysis of Variance*. En variansanalyse kan være alt fra en helt enkel og oversiktlig situasjon til en komplisert modell som kan være vanskelig å fortolke og som i praksis ville vært nær umulig å beregne manuelt.

En helt enkel situasjonen kan eksemplifiseres ved det følgende urealistiske datasettet, hvor 16 dommere har vært i aksjon og bedømt søthet: 4 av dem har bedømt sort A, 4 har bedømt sort B, 4 har bedømt sort C og 4 har bedømt sort D. Her kan det ikke understrekes tydelig nok at dette ikke er en realistisk situasjon i sensorikk: der lar vi – nesten uten unntak – alle dommerne bedømme alle sortene. *Men denne modellen lar oss demonstrere gangen i den aller enkleste formen for variansanalyse.*

Lesere som kjenner til begrepet T-test kan her bli fristet til å benytte den flere ganger: først sammenlikne A med B, så A med C, A med D, B med C, B med D og C med D, totalt 6 T-tester. Dette er en framgangsmåte som det sterkt advares mot, fordi vi da mister kontrollen over nivået for hele testsituasjonen. Selv om nivået for hver av testene er for eksempel 0,05, så gjelder dette ikke hvis vi ser alle testene under ett, det vil si å teste H_0: A = B = C = D.

Grunntanken i en variansanalyse er – ikke overraskende – å sammenlikne varianser. Det som derimot er overraskende, er at det vi egentlig uttaler oss om, er middelverdier. Idéen er ikke umiddelbart opplagt og krever en viss matematisk skolering for å kunne forstås helt ut. La oss anta at resultatene ble som i tabell 11.1:

Tabell 11.1: Middelverdier for 4 sorter

SORT A	SORT B	SORT C	SORT D
3	4	5	8
6	1	5	5
4	2	4	8
7	5	6	7

Middelverdiene for de 4 sortene blir 5, 3, 5 og 7. Variansen i datamaterialet kan beregnes på to måter, forutsatt at det ikke er noen systematisk forskjell mellom sortene. En måte å regne på, er å beregne variansen innen hver av de 4 sortene, og så ta gjennomsnittet av de 4 variansene. Avrundet til 2 desimaler blir de 4 variansene 3,33 – 3,33 – 0,67 og 2,00, og gjennomsnittet av dem blir 2,33.

Men hvis det ikke er noen forskjell mellom sortene, så vil variansen for de 4 gjennomsnittene også være et fornuftig uttrykk for variansen i datasettet. Denne verdien blir 2,67. Når variansen for et sett observasjoner er σ^2, så er variansen for middelverdien over n slike observasjoner lik σ^2/n. For at uttrykket ovenfor (2,67) skal være sammenlignbart med den gjennomsnittlige variansen vi fant (2,33), må vi derfor multiplisere den førstnevnte med antall observasjoner som ligger bak hvert middeltall, nemlig 4, og vi får 10,67. (Her har vi brukt de ikke-avrundete delresultatene i beregningene). Siden vi har to tall som begge er et uttrykk for variansen, forutsatt at det ikke er noen forskjell mellom sortene, er det naturlig å sammenlikne dem, for eksempel ved å dividere dem med hverandre:

$$F = \frac{10,67}{2,33} = 4,57$$

Under forutsetning av at H_0 var sann, det vil si at det bare er tilfeldige variasjoner som gjør at ikke alle 4 middelverdiene blir identiske, så ville vi forventet at F var nær 1. Når den er større enn 1, tyder det på at det estimatet for totalvariansen i datasettet som vi beregnet, inneholder noe mer enn bare den tilfeldige variasjonen mellom middelverdiene. Eller med andre ord: de 4 middelverdiene kan ikke betraktes som å være like allikevel. Telleren i F består altså av noe mer enn variasjonen i datasettet og dette «noe» kan direkte tilskrives forskjellene i middelverdier.

Hvor sannsynlig er det at det skal bli så store forskjeller som det har blitt, forutsatt at H_0 er sann? Det kan vi få ut ved hjelp av et av de mange statistikkprogrammene som er tilgjengelige (vi har igjen benyttet Statistix 9):

One-Way AOV for X by Sort

Source	DF	SS	MS	F	P
Sort	3	32.0000	10.6667	4.57	0.0234
Error	12	28.0000	2.3333		
Total	15	60.0000			

Her kjenner vi igjen verdiene 10,67 og 2,33 fra den manuelle beregningen av variansene, samt F-verdien 4,57. Programmet har riktignok skrevet ut variansene med 4 desimaler i stedet for våre 2. Sannsynligheten finner vi under overskriften P: den er 0,0234 og er ikke noe vi kan regne ut selv. For å teste H_0 på nivå 0,05 uten tilgang til program av denne typen måtte vi ha funnet ut at 0,95-fraktilen i F-fordelinga med 3 og 12 frihetsgrader var 3,4903. Siden den F vi beregnet er større enn den verdien vi fant i tabellen, må vi altså forkaste H_0.

Formlene vi har benyttet, skrevet ut i en mer formell drakt, ser slik ut (lesere med formelfobi gis herved tillatelse til å hoppe over dette avsnittet):

Variansen innen gruppe nr. i:

$$\frac{1}{n-1}\sum_{j=1}^{n}(X_{ij}-\bar{X}_{i\bullet})^2$$

Middelverdien over disse variansene:

$$Q_0=\frac{1}{m}\sum_{i=1}^{m}\frac{1}{n-1}\sum_{j=1}^{n}(X_{ij}-\bar{X}_{i\bullet})^2$$

Variansen for gruppemiddeltallene blir:

$$\frac{1}{m-1}\sum_{i=1}^{m}(\bar{X}_{i\bullet}-\bar{X}_{\bullet\bullet})^2$$

Etter multiplikasjon med n:

$$Q_1 = \frac{n}{m-1}\sum_{i=1}^{m}(\bar{X}_{i\cdot} - \bar{X}_{\cdot\cdot})^2$$

$$F = \frac{Q_1}{Q_0}$$

Beviset for at F er Fisher-fordelt med m-1 og m(n-1) frihetsgrader når H_0 er sann, forbigår vi denne sammenhengen, men henviser til lærebøker i generell statistikk.

Selv om en variansanalyse kan ses på som en komplisert prosedyre rent beregnings-teknisk sett, eksisterer det altså en helt mekanisk løsning av problemet:

1. Mat dataene inn i programmet.
2. Trykk på de rette knappene(!).
3. Sjekk om p-verdien er mindre enn det nivået man vil teste på.

11.4.2 Toveis variansanalyse uten gjentak

Vi vil nå se på en mer realistisk sensorisk situasjon, hvor vi lar et panel på 9 dommere bedømme bitterhet i 3 forskjellige kaffesorter, for enkelthets skyld kalt A, B og C. Bedømmelsene ga følgende resultat:

Tabell 11.2: Dommerbedømmelser og middelverdier for 3 sorter

DOMMER	SORT A	SORT B	SORT C
1	4,6	3,8	8,2
2	4,2	7,1	8,1
3	5,9	5,2	8,4
4	6,0	3,2	7,6
5	5,5	3,1	5,1
6	4,7	3,6	7,3
7	5,0	3,9	7,7
8	4,8	4,1	8,0
9	4,3	6,5	7,1
Middel:	5,0	4,5	7,5

Som andre målinger (kjemiske, fysiske, ...) er også disse sensoriske målingene forbundet med en usikkerhet. Selve «måleinstrumentet», nemlig dommerne i panelet, varierer; og det kan de gjøre både på en systematisk og en tilfeldig måte. Videre kan det være variasjon i prøvematerialet: selv om vi tilbereder kaffen på samme måte hver gang, kan det være råvareforskjeller også innen en enkelt sort. Vi kan derfor ikke uten videre konkludere med at kaffesort C er mer bitter enn de to andre. Spørsmålet vi må svare på, er om de forskjellene vi observerer, er signifikante. Sagt på en annen måte: kan hypotesen H_0: A = B = C forkastes på et gitt nivå, f.eks. 0,05?

Selv om tabellen over i sin struktur likner tabellen i det forrige delkapitlet til forveksling, er det en viktig forskjell: i kaffe-eksemplet er det de samme 9 dommerne som har bedømt alle sortene, slik at vi har ikke bare en sortseffekt vi må ta hensyn til i modellen, men også en dommereffekt.

Kortversjonen av hvordan vi skal analysere disse kaffedataene, er som følger:

1. Mat dataene inn i programmet.
2. Trykk på de rette knappene(!).
3. Sjekk om p-verdien er mindre enn det nivået man vil teste på.

Statistix 9 gir oss:

Analysis of Variance Table for SensScore

Source	DF	SS	MS	F	P
Sort	2	46.5000	23.2500	19.14	0.0001
Dommer	8	8.6800	1.0850	0.89	0.5441
Error	16	19.4400	1.2150		
Total	26				

Middelverdiene og standardavvik kan vi også få:

Descriptive Statistics for Sort = A			
Variable	**N**	**Mean**	**SD**
SensScore	9	5.0000	0.6595
Descriptive Statistics for Sort = B			
Variable	**N**	**Mean**	**SD**
SensScore	9	4.5000	1.4457
Descriptive Statistics for Sort = C			
Variable	**N**	**Mean**	**SD**
SensScore	9	7.5000	0.9950

Et hvilket som helst statistikkprogram vil gi oss den samme informasjonen, men den kan være organisert på en annen måte. Et gratisprogram som R, her i versjon 3.1.1 og med modulene («package» i R-terminologi) *mixlm* og *describe* gir følgende utskrift, lettere redigert av plasshensyn:

Analysis of variance (unrestricted model)
Response: SensScore

	Mean Sq	Sum Sq	Df	F value	Pr(>F)
Sort	23.25	46.50	2	19.14	0.0001
Dommer	1.09	8.68	8	0.89	0.5441
Residuals	1.21	19.44	16		

Group A

	n	mean	sd
SensScore	9	5.00	0.66

Group B

	n	mean	sd
SensScore	9	4.50	1.45

Group C

	n	mean	sd
SensScore	9	7.50	0.99

I praksis har altså data fra en beskrivende test minst en toveis-analyse, og hvis de forskjellige prøvene kan deles inn i flere undergrupper eller forsøksledd (for eksempel: sorter, lagringsforhold, dyrkingssted …) kan det være snakk om tre-, fire-, og flerveisanalyser. Selv om dette kompliserer formlene og gjør en manuell utregning nesten uoverkommelig, er prinsippet grovt sett det samme uansett hvor komplisert modellen er: totalvariansen deles inn i delvarianser som kan tilskrives de forskjellige forsøksleddene som inngår i modellen, samt at det alltid vil være igjen en rest – i utskrifter fra dataprogrammene gjerne betegnet Error. En annen betegnelse som benyttes på denne restvariansen, er MSE (Mean Square Error) – et begrep som vil dukke opp seinere under beskrivelsen av programvaren PanelCheck.

11.4.3 Rangeringer: Friedman's test

I stedet for å bruke bedømmelsene fra dommerne direkte, kan vi gjøre dem om til rangeringer. Da er det naturlig å analysere data ved hjelp av Friedman's test. Den tar utgangspunkt i rangereringer, for eksempel ser vi av tabellen i kapittel 11.2 at dommer 5 har rangert prøvene i rekkefølgen A – C – B, hvor A har fått høyest score. Da vil resultatene for dommer 5 for sort A kodes om til 3, sort C til 2 og sort B til 1. Tilsvarende blir gjort for de andre dommerne. Tabellen omgjort til rangeringer blir da:

Tabell 11.3: Rangering av 3 sorter, rangsum og gjennomsnittlig rang

DOMMER	SORT A	SORT B	SORT C
1	2	1	3
2	1	2	3
3	2	1	3
4	2	1	3
5	3	1	2
6	2	1	3
7	2	1	3
8	2	1	3
9	1	2	3
Rangsum	17	11	26
Gj.snittlig rang	1,89	1,22	2,89

Hvis sortene er like, vil vi forvente at rangeringene fordeler seg slik at A, B og C får omtrent samme rangsum (= summen av rangeringene). Spørsmålet i eksemplet blir nå om rangsummene 17, 11, 26 er så forskjellige at vi må forkaste hypotesen om at de er like, eller at forskjellene i rangsummer like gjerne skyldes rene tilfeldigheter. Å regne ut sannsynligheten for å få så store (eller større) forskjeller er ingen enkel oppgave. Vi støtter oss her til statistisk teori, nærmere bestemt Friedman's formel, som sier at T definert under er tilnærmet kji-kvadratfordelt:

$$T = \frac{12}{ds(s+1)} \sum_{i=1}^{s} R_i^2 - 3d(s+1)$$

der d = antall dommere (= 9 i eksemplet over), s = antall sorter (= 3), og Ri (i=1,2, ..., s) er rangsummene (17, 11 og 26 i tabellen over). Da blir:

$$T = \frac{12}{9 \times 3 \times (3+1)} (17^2 + 11^2 + 26^2) - 3 \times 9 \times (3+1) = 12{,}67$$

Siden T under H_0 er tilnærmet kji-kvadratfordelt med s – 1 frihetsgrader (= 2 i dette eksemplet), må vi sammenlikne T = 12,67 med 0,95-fraktilen i denne fordelinga. Av statistiske tabeller (vedlegg 16 b, tabell 3) ser vi at den verdien er lik 5,9915. Siden T > 5,9915, må derfor H_0 forkastes. I praksis vil vi vanligvis la et statistikkprogram stå for selve utregningene, og da kan resultatet bli seende slik ut (her har vi benyttet Statistix 9)

Friedman Two-Way Nonparametric AOV

Variable	Mean Rank	Sample Size
Sort A	1.89	9
Sort B	1.22	9
Sort C	2.89	9

Friedman Statistic	**12.667**
P-value, Chi-Squared Approximation	**0.0018**
Degrees of Freedom	**2**

I stedet for å sammenlikne T (*Friedman Statistic* i utskriften fra programmet) med verdien fra tabellen, sammenlikner vi p-verdien (0,0018) i utskriften med nivået vi tester på (0,05). Siden den beregnede p-verdien i utskriften er mye mindre enn nivået, forkaster vi H_0. Konklusjonen blir altså den samme, men bruk av statistikkprogrammet vil i større grad kunne si oss hvor signifikant forskjellene er: det er forskjell på p = 0,0003 og p = 0,0499 selv om begge sier at det er forskjell på nivå 0,05.

Det programmet som er benyttet her, skriver ut de gjennomsnittlige rangeringene i stedet for rangsummene 17, 11 og 26; dette vil variere fra program til program.

Hvis en dommer har bedømt 2 eller flere sorter likt, deles rangeringene: hvis to prøver har fått 6,0 og en prøve 4,7 så får de to første hver rang 2,5 (= gjennomsnittet av 3 og 2, som de ville fått hvis de i stedet hadde hatt for eksempel verdiene 6,1 og 6,0) og den tredje får rang 1.

Friedman's test kan benyttes enten dommerne har bedømt egenskapene langs en skala, eller de har gjort rangeringene direkte. Testen er et interessant alternativ hvis man ikke er overbevist om at data tilfredsstiller de kravene som en variansanalyse stiller. Hvis dommerne har lite trening i å bruke en skala, for eksempel hvis det dreier seg om forbrukere, kan det være en ide å bruke Friedman's test til analysene selv om data opprinnelig er på en 1-9 skala.

11.4.4 Multiple sammenlikninger

Å forkaste H_0 er vanligvis ikke en fullstendig konklusjon. Hvis vi har flere enn 2 grupper, vil vi gjerne vite hvilke grupper som er forskjellige. I eksemplet fra kapitlet over kan vi stille spørsmål som:

Er alle 3 sortene forskjellige fra hverandre?

Er sort 1 og sort 2 forskjellige fra sort 3, mens vi ikke kan skille sort 1 fra sort 2?

Er sort 2 forskjellig fra sort 3, mens vi ikke kan skille sort 1 fra sort 2 eller sort 1 fra sort 3?

Her er det nærliggende å benytte T-tester for hver av sammenlikningene:

Sort 1 mot sort 2

Sort 1 mot sort 3

Sort 2 mot sort 3

Den teknikken må det advares sterk mot av så vel sannsynlighetsteoretiske som praktiske grunner. Den korrekte måten å besvare slike spørsmål på, er å benytte metoder som går under betegnelsen multiple sammenlikninger. Av dem finnes det flere, og den vanligste er sannsynligvis Tukey's test. Også den vil ligge inne i alle statistikkprogrammer med respekt for seg selv. En grei måte å presentere slike resultater på, er denne, hentet fra Statistix 9:

Tukey HSD All-Pairwise Comparisons Test of SensScore for Sort

Sort	Mean	Homogeneous Groups
Sort C	7.5000	A
Sort A	5.0000	B
Sort B	4.5000	B

Alpha	0.05	Standard Error for Comparison	0.5196
Critical Q Value	3.651	Critical Value for Comparison	1.3413

Error term used: Sort*Dommer, 16 DF
There are 2 groups (A and B) in which the means
are not significantly different from one another.

Hva kan vi så lese ut av dette? Prinsippet er at sorter som ikke er signifikant forskjellige fra hverandre, markeres med samme bokstav i kolonnen Homogeneous Groups. Hovedkonklusjonen er at det er signifikant forskjell mellom A og C og mellom B og C, men ikke mellom A og B. Alpha er nivået for testen, og Critical Value for Comparison er den verdien som middelverdiene sammenliknes med. Forskjellen mellom A og B er 0,5 som er mindre enn 1,3413: altså er det ikke forskjell mellom A og B. Men forskjellene mellom A og C (2,5) og mellom B og

C (3,0) er begge større enn 1,3413 – følgelig er både A og B forskjellige fra C.

I større forsøk kan resultatet av multiple sammenlikninger bli ganske komplisert, som i dette forsøket med 8 prøver:

	Mean	G1	G2	G3	G4
Prøve JK407	4.104762	A			
Prøve JO171	3.861905	A			
Prøve BG891	3.566667	A	B		
Prøve BY100	3.366667	A	B		
Prøve CX301	3.161905	A	B	C	
Prøve EE291	2.119048		B	C	D
Referanse	1.685714			C	D
Prøve LK197	1.461905				D

Skal vi beskrive dette med ord, blir det:

JK407 og JO171 er forskjellige fra EE291, Referanse og LK197
BG891 og BY100 er forskjellige fra Referanse og LK197
CX301 er forskjellig fra LK197

Mellom de øvrige prøvene er det ingen forskjeller. Merk at her er forskjellig fra benyttet i betydningen signifikant forskjellig på nivå 0,05.

I en type situasjoner kan det være aktuelt med en annen test, nemlig hvis vi ikke er interessert i å sammenlikne et visst antall sorter som i utgangspunktet er likeverdige, men ønsker å sammenlikne disse sortene mot en standard eller kontroll. Med 4 sorter A, B, C, D og en kontroll K, er vi altså ikke interessert i å sammenlikne alle 5 sortene innbyrdes, men bare de 4 første mot den siste. I en viss forstand slipper vi å «kaste bort» noe av nivået for testen ved å gardere oss mot feilaktig å påstå at det er forskjeller innen A, B, C og D. Denne testen kalles Dunnett's test og er tilgjengelig i alle standard statistikkprogram. I det ovenstående eksemplet med 8 prøver kunne Dunnett's test vært et alternativ, siden vi har en prøve som kan betraktes som en kontroll eller en referanse.

Også i Friedman's test er det aktuelt å gjøre multiple sammenlikninger hvis vi har forkastet H_0. I motsetning til Tukey's test er dette ikke en standardanalyse, og den må derfor gjennomføres for hånd. Første punkt er å beregne den kritiske verdien som alle parene skal sammenliknes med. En navnløs test beskrevet i Hochberg og Tamhane (1987) gir en kritisk verdi c gitt ved:

$$c = \frac{n}{\sqrt{2}} q_{k,\infty;\alpha} \sqrt{\frac{k(k+1)}{6n}}$$

Her er n lik antall dommere, k = antall grupper, og $q_{k,\infty,\alpha}$ er den kritiske verdien for den Studentiserte variasjonsbredde (Studentized range) for k grupper, uendelig mange frihetsgrader og nivå α for testen. Siden den tabellen ikke er vanlig å gjengi i statistiske lærebøker, er en tabell for k = 3, 4, …, 20 og α = 0,01, 0,05 og 0,10 gjengitt i vedlegget. De gruppene som har rangsummer som avviker med mer enn c, er signifikant forskjellige. Med $q_{k,\infty,\alpha}$ = 3,315 fra tabellen i vedlegg 16 får vi:

$$c = \frac{9}{\sqrt{2}} \times 3{,}315 \times \sqrt{\frac{3 \times 4}{6 \times 9}} = 9{,}945$$

Siden forskjellen mellom A og C og mellom A og B er mindre enn denne kritiske verdien, mens forskjellen mellom B og C er større enn den, kan konklusjonen framstilles slik:

C	26^{a}
A	17^{ab}
B	11^{b}

Altså: det er ikke signifikant forskjell mellom A og C, og heller ikke mellom A og B, mens det er signifikant forskjell mellom B og C. Legg merke til at denne konklusjonen er noe svakere enn den vi kunne trekke i variansanalyseeksemplet: der var både A og B signifikant forskjellige fra C.

11.4.5 Toveis variansanalyse med gjentak

Dette er kanskje den vanligste modellen i sensorisk analyse: vi har et visst antall prøver: A, B, C, osv., og alle dommerne i et sensorisk panel har bedømt alle prøvene flere ganger (som oftest 2 eller 3). Den underliggende metoden for variansanalyse som vi har sett på tidligere, benyttes også her. Et konkret eksempel:

Tabell 11.4: Data fra 3 sorter bedømt av 10 dommere i 2 gjentak

	Sort A	Sort A	Sort B	Sort B	Sort C	Sort C
Dommer	Gjentak 1	Gjentak 2	Gjentak 1	Gjentak 2	Gjentak 1	Gjentak 2
1	3,6	3,9	1,9	2,7	4,0	3,1
2	2,1	1,6	1,0	1,0	2,0	2,8
3	3,9	4,5	1,0	3,7	4,3	4,0
4	1,0	3,0	1,0	1,0	5,0	4,8
5	1,5	6,6	3,1	2,5	6,6	5,1
6	2,5	2,7	2,3	3,7	2,7	3,4
7	3,3	5,6	1,0	1,0	3,6	2,8
8	2,9	2,6	3,6	3,3	2,6	4,1
9	3,8	1,3	1,5	1,1	4,0	1,2
10	3,9	6,1	2,6	1,9	5,0	7,7

I denne modellen oppstår tre nye begreper som vi må ta hensyn til. Det første er relativt greit å forholde seg til: samspillet mellom to faktorer, i dette tilfellet: samspillet mellom sort og dommer. At det er samspill mellom faktorene sort og dommer, betyr i korthet at ikke alle dommerne har bedømt alle sortene likt: dommer 1 kan ha gitt høyere score til prøve A enn til prøve B, mens dommer 2 har gitt høyere score til prøve B enn til prøve A. Aller helst hadde vi jo ønsket at dommerne utviste større enighet enn det. Det er ikke så farlig om de legger seg på forskjellige deler av skalaen, så lenge de er konsekvente. Med konsekvente menes her at de bedømmer prøvene i omtrent samme rekkefølge.

Det andre begrepet er at effekter kan deles inn i de som er tilfeldige (engelsk: random) eller faste (engelsk: fixed). Om en effekt er fast eller tilfeldig, får konsekvenser for beregningen av F-verdien. I de to enkle modellene vi har sett på hittil, får det ingen praktiske konsekvenser om effektene er faste eller tilfeldige: formlene – og dermed alle utregningene – blir akkurat de samme. Men med en gang vi får tilfeldige effekter inn i modellen, endrer det til dels dramatisk noe av det teoretiske grunnlaget for testene.

En fast effekt er en effekt hvor vi er interessert i de nivåene, eller verdiene, som effekten har. I et eksempel med 3 sorter A, B og C er det naturlig at disse 3 sortene er valgt ut fordi det er nettopp disse sortene vi er interessert i. En annen situasjon ville vi fått hvis det var snakk om eplesorter, og vi i stedet for å bestemme oss på forhånd hvilke sorter vi ville undersøke, så gikk vi i en butikk og kjøpte de sortene som tilfeldigvis var til salgs den dagen. En tilfeldig effekt kan bringe tanken hen på at det er nettopp noe tilsvarende vi gjør når en effekt defineres som tilfeldig. Det er imidlertid ikke tilfelle når det gjelder de tilfeldige effekten i herværende modell, nemlig dommereffekten. Når vi definerer den som tilfeldig, betyr det selvfølgelig ikke at det sensoriske panelet består av personer trukket tilfeldig ut fra en befolkningen over 18 år. De aller fleste sensoriske panel er hentet fra personer som er trenet til oppgaver innen sensoriske analyser. Men det er likevel tilfeldig at det er nettopp disse personene som utgjør panelet.

Vi holder oss til den tradisjonen som definerer en fast effekt hvor vi er interessert i de konkrete verdiene (sortene) vi analyserer, og alt annet defineres som tilfeldige effekter.

Det tredje av de nye begrepene er skillet mellom kryssete (engelsk: crossed) og nøstet (engelsk: nested), eller hierarkiske effekter. I den modellen vi ser på her, er dommer og sort krysset: alle dommerne har bedømt alle sortene. Hvis vi hadde benyttet totalt 30 dommere, og 10 dommere hadde bedømt sort A, 10 andre dommere hadde bedømt sort B og de 10 siste dommerne hadde bedømt sort C, så hadde dommereffekten vært hierarkisk under sortseffekten. I andre, og mer kompliserte modeller, hvor vi har 3 slaktemetoder og 3 dyr fra hver metode, ville det være naturlig å modellere dyreeffekten som hierarkisk under slaktemetoden. Ett enkelt dyr kan av opplagte grunner bare bli slaktet etter en metode. I en slik situasjon ville ikke dyr 1 fra metode A hatt noe som helst felles med dyr 1 fra metode B og C, noe som vi må ta hensyn til i utregningene. Fôringsforsøk er en annen type forsøk hvor det er naturlig å benytte en modell hvor dyr er hierarkisk under et fôringsregime.

Et annet eksempel på hvor dette er aktuelt, er følgende situasjon: vi skal sammenlikne sauser som leveres i poser med pulver, 3 poser fra

hver sort. Da kan vi lage 3 sauser av hver sort og la hver pose være et gjentak, eller blande alle 3 posene og lage en stor porsjon som vi serverer 3 gjentak fra til hver dommer. I det første tilfellet vil vi ha en pose-effekt hierarkisk under sortseffekten: pose 1 fra sort D vil ikke ha noe felles (bortsett fra den tilfeldige nummereringen) med pose 1 fra sort C. En slik modell vil gi oss informasjon om variasjonen mellom enkeltposer, mens den andre lager en form for middelverdi ved å blande de 3 posene og dermed miste all informasjon om variasjon mellom enkeltposene.

Gjentakene er vanligvis både en tilfeldig og hierarkisk effekt, i hvert fall i slike situasjoner hvor gjentak 1 og gjentak 2 bare er tilfeldige nummereringer. Noe annet ville det vært hvis gjentak 1 ble foretatt på en viss dato, og gjentak 2 en uke seinere. Da ville gjentakseffekten vært krysset med både sort og dommer, og vi ville havnet utenfor rammene til dette underkapitlet.

En variansanalyse av data fra en modell som beskrevet: en fast sortseffekt krysset med en tilfeldig dommereffekt, samspill mellom sort og dommer, og en hierarkisk og tilfeldig gjentakseffekt, vil gi oss svar som dette, lettere redigert for lesbarhetens skyld (her er benyttet SAS versjon 9.4):

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Sort	2	37.340333	18.670167	11.19	0.0007
Dmr	9	39.661500	4.406833	2.64	0.0380
Error	18	30.033000	1.668500		
Error: MS(Sort*Dmr)					

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Sort*Dmr	18	30.033000	1.668500	1.24	0.2913
Error: MS(Error)	30	40.275000	1.342500		

Her ser vi at det er god grunn til å forkaste H_0, siden p-verdien er så liten som 0,0007. At dommereffekten også er signifikant, er av mindre betydning, det betyr bare at de har benyttet forskjellige deler av

skalaen. At samspillet mellom sort og dommer ikke er signifikant, er også interessant – det betyr at dommerne stort sett har vært enige. Hvis for eksempel en dommer bedømmer søtheten i prøvene A – B – C til henholdsvis 2,3 – 3,1 – 4,0 og en annen dommer bedømmer de samme prøvene til 5,2 – 5,5 – 6,1 så er disse enige: C er søtere enn B som igjen er søtere enn A. Sagt på en annen måte: de to dommerne har rangert prøvene på samme måte med hensyn på søthet.

11.4.6 Flerveis variansanalyse

I prinsippet kan variansanalysemodellen utvides nærmest i det uendelige ved å legge til flere effekter, som både kan være krysset med eller være hierarkisk under en eller flere andre effekter, være faste eller tilfeldige. Med tilgang til statistisk programvare vil det være få problemer med å gjøre selve utregningene som skal til, men det kan fort oppstå problemer med tolkningene. Hvis det for eksempel er signifikante samspill, får dette betydning for tolkningen av hovedeffektene.

En annen type komplikasjon inntreffer når datasettet ikke er balansert. Et ubalansert datasett har vi når for eksempel en eller flere av dommerne ikke har bedømt alle prøvene fordi de har vært fraværende under deler av forsøket. Ubalanse kan også oppstå hvis en produksjon har gått galt, og det ikke går an å kjøre produksjonen på nytt. Sammenlikningene kan bli urettferdige og gale.

I noen situasjoner utelukker man enkelte kombinasjoner av effekter rett og slett fordi et fullstendig design med alle effektene kombinert med alle de andre effektene ville gi et uoverkommelig stort datasett. Såkalte «2-i-n'te» forsøk (2n-forsøk) faller i denne kategorien. Slike forsøk består i å analysere et spesifisert antall effekter (n), hver på 2 nivåer. Da kan antall analyseprøver fort bli uhåndterlig stort: hvis 7 slike effekter inngår, blir det 2^7=128 prøver som skal bedømmes – og da har vi sett helt bort fra eventuelle gjentak. Her er det viktig at man følger et definert forsøksopplegg – i litteraturen finner man detaljene under Design of Experiments eller Experimental Designs.

11.4.7 Manglende verdier

I forsøk som går over flere dager, hender det at dommere er forhindret fra å være tilstede under hele bedømmelsen. Slike «hull» i datamatrisa gjør at teorien bak variansanalysen blir mer komplisert, og noen statistikkprogrammer kan også få problemer med å utføre beregningene hvis antall manglende verdier blir for stort. Hvis det mangler en eller flere verdier i datamatrisa, sier vi at modellen er ubalansert. Fortolkningsmessig er ubalanserte modeller mer kompliserte enn balanserte. Det er derfor en stor fordel å ha balanserte modeller. Det finnes en opplagt måte å gjøre en ubalansert modell balansert på, og det finnes en som er mindre opplagt. Hvis vi i utgangspunktet har relativt mange dommere (mer enn ca. 10), og en av disse bare har vært tilstede på noen av bedømmelsene, kan vi velge å utelate denne dommeren helt i beregningene. Dermed er modellen balansert. Et annet alternativ er å erstatte de manglende verdiene med ett eller annet gjennomsnitt. Denne framgangsmåten kan være akseptabel hvis programmet selv beregner hva disse manglende verdiene skal erstattes med, men er ikke å anbefale hvis de må beregnes mer eller mindre for hånd. Å erstatte manglende verdier med middelverdier innebærer også at man må justere frihetsgradene for restkvadratsummen.

Har man en ubalansert modell er altså valget om man vil kaste bort noen av dataene og dermed få en enkel, balansert og lett tolkbar modell, eller man vil analysere data med en ubalansert modell med de fortolkningsproblemer det medfører.

En annen årsak enn dommere som uteblir til at man får manglende verdier, ser vi ofte i dommerpanel hvor data registreres for hånd. I slike situasjoner er det fort gjort å glemme å fylle inn en eller flere karakterer.

11.5 Flere egenskaper samtidig: PCA

Alle metodene vi har sett på så langt i dette kapitlet har vært såkalte univariate – vi ser på hver egenskap for seg. Noen ganger er det også av interesse å se hvordan egenskapene varierer i sammen, og til det

trenger vi multivariable metoder. En mye brukt metode er prinsipal komponentanalyse, forkortet PCA etter den engelske betegnelsen Principal Component Analysis. En vanlig datamatrise i et sensorisk forsøk vil ikke bare se slik ut som de tabellene vi hittil har presentert, men de vil ha en rad som tilsvarer bedømmelsene for en dommer av en sort i ett gjentak, og så en kolonne for hver egenskap. Et eksempel med 8 dommere, sorter A-H, 2 gjentak og totalt 26 sensoriske egenskaper (luktegenskaper L1-L9, smaksegenskaper S1-S12 og teksturegenskaper T1-T5) kan se slik ut (bare de første og siste radene og de første og siste kolonnene vises):

Tabell 11.5 Eksempel på full datamatrise av en større sensorisk analyse

Dmr	Sort	Gjt	L1	L2	L3	L4	L5	L6	.	S8	S9	S10	S11	S12	T1	T2	T3	T4	T5
1	A	1	6,5	2,9	4,6	2,5	5,3	1,0	.	2,6	2,5	2,0	6,6	5,4	5,6	5,1	4,3	4,3	7,2
1	A	1	7,2	2,9	3,1	4,9	2,6	2,0	.	3,3	3,6	2,8	3,4	4,2	7,9	2,5	1,4	2,5	8,2
3	A	1	6,5	1,0	4,7	1,0	1,0	1,0	.	4,9	1,7	5,2	5,5	2,5	3,9	4,3	1,3	4,2	7,6
4	A	1	7,7	1,0	3,8	7,2	5,3	5,5	.	2,8	2,7	6,7	5,9	6,6	5,0	3,1	2,2	2,3	5,7
5	A	1	7,1	1,0	5,0	3,7	3,2	1,0	.	4,1	2,7	5,7	6,7	6,3	5,5	2,5	1,0	2,7	9,0
6	A	1	6,4	1,9	5,6	5,2	1,0	1,0	.	2,4	2,0	3,5	4,6	5,5	5,9	5,0	1,7	4,1	6,9
7	A	1	7,2	1,9	5,5	5,2	5,8	2,7	.	4,5	1,6	4,2	6,2	2,4	3,3	3,5	1,2	4,3	7,7
8	A	1	8,5	1,1	3,4	3,4	5,7	1,1	.	4,7	1,0	1,0	1,0	9,0	6,2	1,1	1,1	2,0	9,0
1	A	2	5,7	3,6	4,2	4,5	3,3	3,0	.	2,5	2,4	2,1	6,2	1,0	5,4	4,7	3,2	3,6	6,9
2	A	2	6,1	4,1	4,6	2,2	2,0	1,0	.	3,7	3,6	2,7	4,0	3,7	8,3	2,4	1,4	2,6	8,3
3	A	2	6,9	1,0	3,7	1,0	1,0	1,0	.	3,0	1,7	5,2	6,9	2,6	5,1	3,3	1,4	5,6	7,7
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
3	H	2	7,1	1,0	3,8	4,0	3,2	1,0	.	3,1	2,3	2,5	5,6	1,0	3,7	2,6	1,0	5,6	7,5
4	H	2	5,7	3,2	5,2	3,3	3,2	1,0	.	3,0	2,5	3,1	5,8	5,1	4,3	4,2	4,1	4,7	5,5
5	H	2	5,8	1,0	4,8	5,0	3,3	2,8	.	3,7	2,8	5,8	5,9	6,9	5,0	3,2	2,7	3,8	9,0
6	H	2	5,7	4,1	4,0	3,5	2,3	1,0	.	3,3	1,9	3,0	5,8	2,4	5,2	4,3	2,5	4,9	7,4
7	H	2	6,0	1,3	5,1	6,2	5,2	2,1	.	4,2	2,5	4,7	6,6	3,4	5,0	4,9	2,3	4,6	6,2
8	H	2	6,3	2,8	5,0	3,8	5,1	1,1	.	2,9	2,5	1,0	4,5	6,1	3,7	1,1	1,0	5,1	9,0

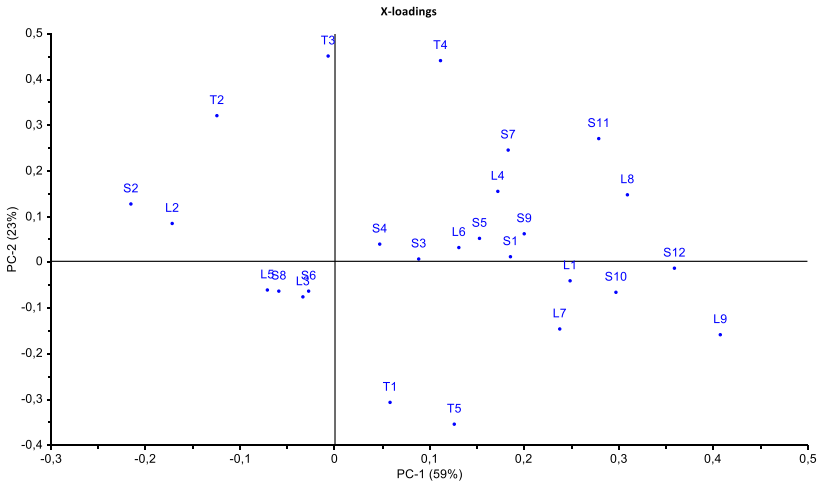
I motsetning til mer klassiske statistiske tester, kan vi ikke her sette opp en enkel formel som vi bare kan sette inn dataene våre i og så få resultatet ut, gjerne i form av ett enkelt tall, for eksempel en F-verdi. I stedet må man benytte en iterativ algoritme som konvergerer etter et visst iterasjoner. Detaljene i algoritmene kan variere fra program til program – dette er en metode som aldri vil gjøres ved hjelp av

manuelle beregninger. Hovedresultatet fra en PCA er skåringsplot (scores) og ladningsplot (loadings). Ofte brukes resultatene fra en PCA som en måte å få et første oversiktsbilde av datasettet på. Bruker vi PCA på dataeksemplet over, vil vi kunne identifisere enkeltbedømmelser som skiller seg ut, såkalte outliere . Hva man i så fall gjør med slike avvikende resultater, er et annet spørsmål! Men hvis det skulle vise seg at alle bedømmelsene på stikkende lukt på prøver fra en bestemt fisk, så kan det være grunn til å undersøke om det under forbehandlingen av denne prøven har skjedd noe spesielt som gjør at den bør utelukkes fra analysene. Er man fornøyd med datasettet slik det er, er det vanlig å beregne gjennomsnitt over både gjentak og dommere når man gjør en PCA. Da vil datasettet redusere seg til:

Tabell 11.6: Eksempel på redusert datamatrise av et større datasett

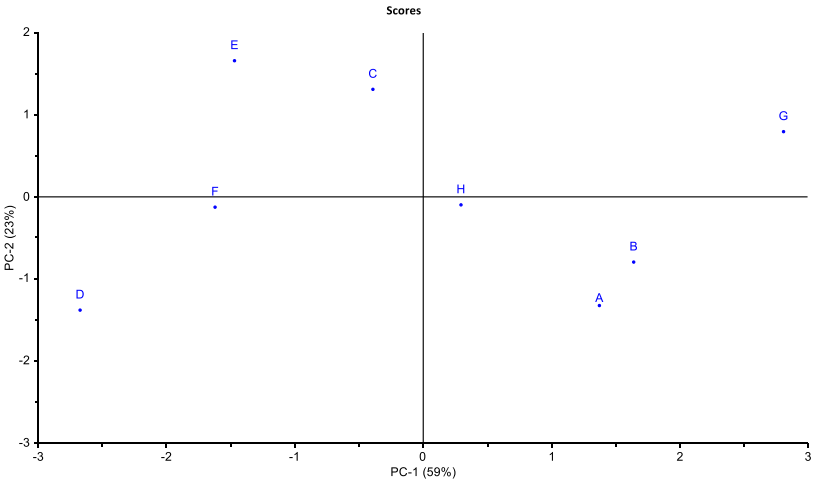
Prøve	L1	L2	L3	L4	L5	L6	L7	.	S8	S9	S10	S11	S12	T1	T2	T3	T4	T5
A	6,55	2,23	4,33	3,91	3,51	1,84	3,19	.	3,69	2,49	3,53	5,16	4,32	5,45	3,46	1,89	3,66	7,73
B	6,76	2,16	4,37	4,48	3,65	1,83	2,86	.	3,57	2,46	3,63	5,58	3,64	5,52	3,08	2,02	4,13	7,63
C	5,74	2,94	4,01	4,13	3,46	1,73	1,86	.	3,41	2,44	2,94	5,51	3,45	4,81	4,03	3,03	4,95	7,02
D	5,63	2,73	4,66	3,33	4,16	1,48	2,21	.	4,14	1,54	2,16	3,81	2,67	5,25	3,72	1,94	3,58	6,96
E	5,99	2,76	4,36	4,14	3,73	1,53	2,13	.	3,83	1,91	2,38	5,15	2,93	4,38	4,66	3,21	4,70	5,99
F	5,94	2,64	4,89	3,35	3,97	1,44	2,85	.	3,76	1,64	2,53	4,86	2,71	4,34	4,43	2,46	3,83	6,63
G	7,03	1,84	4,61	4,38	3,78	2,30	3,68	.	3,75	2,63	3,58	5,90	4,59	4,53	4,07	2,95	5,03	6,59
H	6,23	2,34	4,51	3,89	3,72	1,36	2,64	.	3,61	1,93	3,02	5,68	3,76	4,78	3,73	2,14	4,29	7,13

En PCA kjørt ved hjelp av Unscrambler X ga følgende resultat:



Figur 11.1: PCA plot med markering av produktegenskaper (loadings)

Ladningsplottet viser hvordan egenskapene plasserer seg når vi ser på deres plassering i det 2-dimensjonale planet definert av de to viktigste dimensjonene. Disse representerer 59% + 23% = 82% av all variasjonen i datasettet. Her kan vi røpe at S2 og L2 er syrlig smak og lukt, og L8, L9, S11, S12 har med stram og emmen å gjøre. Den viktigste dimensjonen er altså en akse fra syrlig til emmen.T1 og T3 er hardhet og saftighet, så disse 2 egenskapene er viktige når det gjelder å definere prinsipalkomponent 2.



Figur 11.2: PCA plot med markering av prøveplassering (scores)

I skåringsplot'et ser vi hvordan prøvene plasserer seg langs de samme dimensjonene. D og F ser dermed ut til å være de som har skåret høyest på syrlig smak og lukt, mens G er mest emmen. Litt verre er det å plassere prøvene langs en hardhet-saftighet-akse (dimensjon 2), utover at C ser ut til å være den som scorer høyest på saftighet.

I noen tilfeller kan det også være aktuelt å se på flere dimensjoner og plotte dem mot hverandre i tillegg.

PCA opptrer i litteraturen (særlig den litt «eldre» – før 1980…) også under andre navn: Singular Value Decomposition (SVD) eller egenvektor-dekomposisjon.

11.6 Dypdykking i datamaterialet

En sensoriker vil ofte føle et behov for å gå nærmere inn i datamaterialet enn det en som bare er interessert i å sammenlikne sorter eller prøver er. Kvalitetskontroll av det sensoriske panelet er en viktig del av den daglige driften, og i forbindelse med trening og rekruttering av enkeltdommere, er det også viktig med en form for kvalitetskontroll. Et verktøy som etter hvert har fått en viss status i det sensoriske miljøet er programmet PanelCheck, utviklet i et samarbeid mellom sensorikere og statistikere ved Nofima og Danmarks Tekniske Universitet og med støtte fra forskningsråd og industripartnere i Norge og Danmark. PanelCheck kan lastes ned gratis fra www.panelcheck.com.

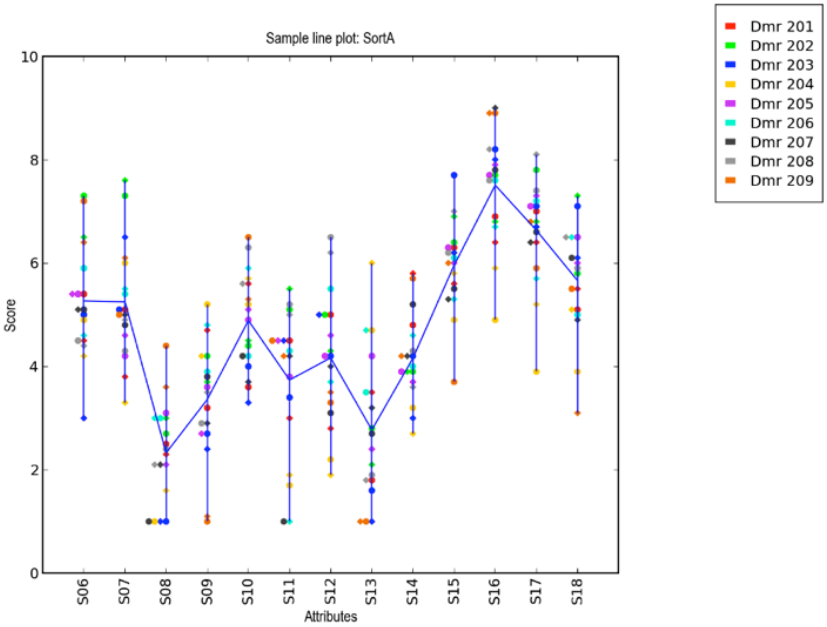
PanelCheck er delt inn i 4 hoveddeler:

- Univariate
- Mutivariate
- Consensus
- Overall

Noen av de ovenstående punktene inngår i forskjellige indekser for enighet mellom enkeltdommere og panelet (AGRPROD), hvordan dommerne er enige med panelt når det gjelder korrelasjoner mellom egenskapene (AGRATT), hvordan dommerne er enige med seg selv over gjentakene (REPPROD), hvordan dommerne er enige med seg selv over gjentak når det gjelder egenskapene (REPATT) og antall egenskaper som dommerne kan skille fra hverandre på nivå 5% (DIS). Disse indeksene er basert på RV2-koeffisienten (Tomic og flere (2013)). Indeksene vil etter hvert bli lagt inn i nye versjoner av PanelCheck.

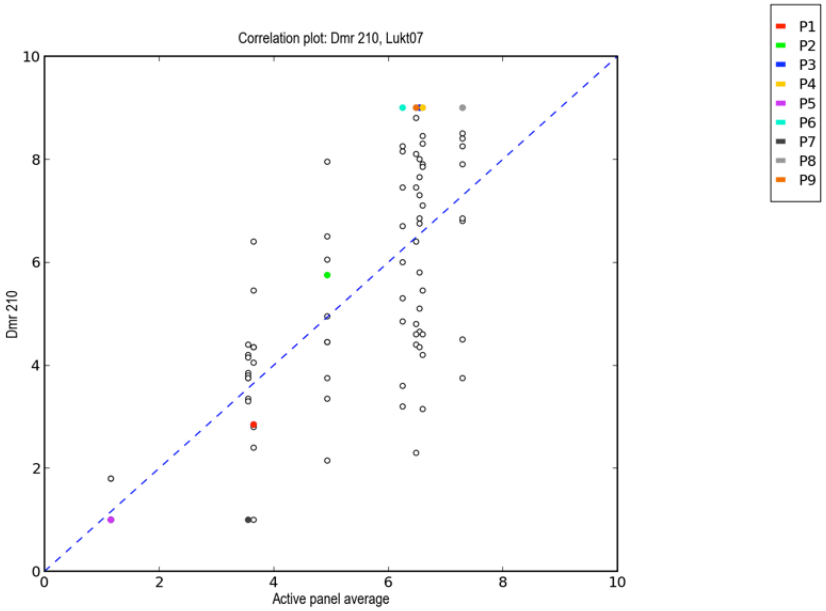
Data kan importeres fra flere standardformater, så som rene tekstfiler (i diverse varianter) og Excel-filer.

Line Plots er nettopp det: enkle figurer som for hver sort (her: Sort 9) viser enkeltgjentakene mot egenskap (her: S1 – S10), pluss panelets gjennomsnitt (blå linje).



Figur 11.3: Line plot som viser alle enkeltbedømmelsene for Sort A for utvalgte egenskaper (S06-S18)

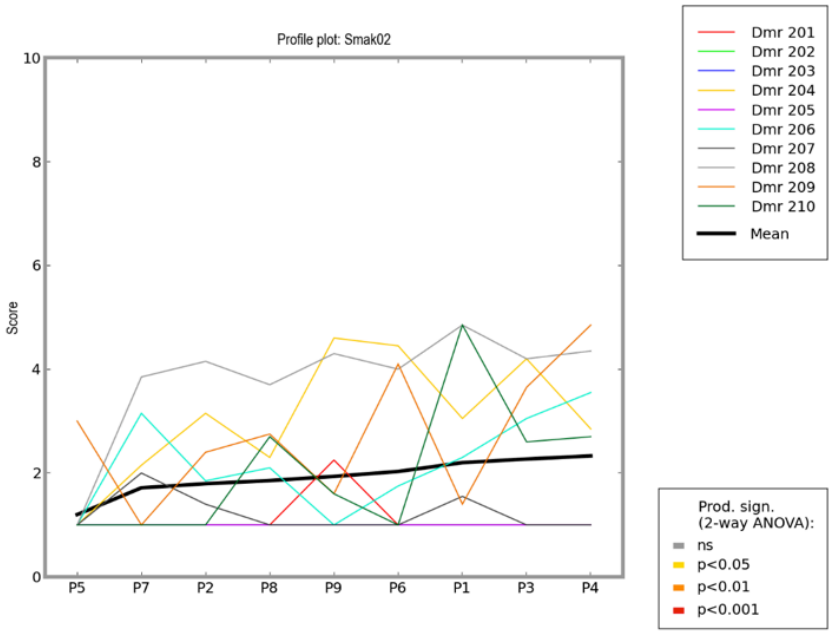
For å se på enkeltdommeres innsats har PanelCheck også andre måter å vise resultatene på: middelverdi over prøver (har dommerne en tendens til å ligge lavt eller høyt på skalaen?), og hvordan ligger en dommer i forhold til de andre dommerne. En enkeltdommers bedømmelser sammenliknet med de andre dommernes bedømmelser vises som et Correlation plot:



Figur 11.4: Correlation Plots viser enkeltdommernes bedømmelser sammenliknet med panelets gjennomsnitt.

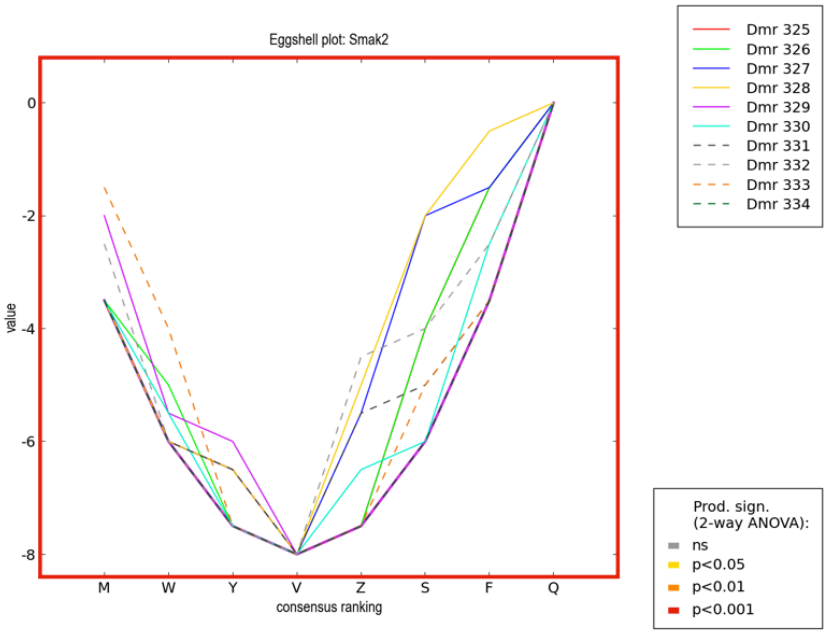
I eksemplet over er det dommer 210 som vises ved de fargede punktene (som i sin tur representerer prøve P1, P2, …, P9 i henhold til kodingene øverst til høyre). En dommer som er 100% enig med panelgjennomsnittet, vil ha alle punktene på den stiplede linja. En dommer som stort sett ligger over linja har en tendens til å gi høyere bedømmelser enn panelet, en som stort sett ligger under har en tendens til å gi lavere bedømmelser enn panelet.

Profile Plots viser enkeltdommernes resultater (midlet over gjentak) sammen med panelets middelverdi. X-aksen utgjøres av de forskjellige prøvene (her: P1, P2, …, P9) ordnet i stigende rekkefølge etter panelmiddelverdiene (andre rekkefølger kan velges). Eksemplet under viser ganske stor spredning mellom dommerne, og sammenliknet med denne er det liten forskjell mellom laveste verdi (P5 = 1,20) og høyeste verdi (P4 = 2,33). At prøvene ikke kan sies å være signifikant forskjellige, sees også ved at rammen rundt plottet er grå.



Figur 11.5: Profilplot for Smak02. Det er liten forskjell mellom prøvene, og panelet har heller ikke funnet signifikante forskjeller. Enkeltdommerens bedømmelser spriker en god del.

Eggshell Plots er en måte å sjekke i hvilken grad dommerne er enige når det gjelder rangering av prøvene. Samspillseffekten Dommer × Prøve i en variansanalyse vil også være ett uttrykk for dette, men samspillet er et lite håndfast begrep for de fleste. Dette vises bedre med figurer:



Figur 11.6: Eggeskallplot for egenskapen Smak2.

Dommerne er stort sett enige om rangeringen mellom prøvene, og panelet har funnet signifikante forskjeller mellom prøvene. For en egenskap hvor det er mer uenighet, vil de foskjellige kurvene sprike mye mer, og ofte vil heller ikke panelet ha funnet signifikante forskjeller

Ideelt sett skal alle dommerne falle sammen med den nederste kurven, men at alle dommerne skulle 100% enige i rangeringene av alle prøvene i et datasett, er urealistisk.

Når dommernes bedømmelser spriker ganske mye, så kan det ha flere årsaker. Det kan godt tenkes at prøvene skiller seg veldig lite fra hverandre, og da er det heller ikke noe poeng i at dommerne ikke er enige i rangeringene av dem. Eggeskallplottet tar ikke hensyn til om prøvene ligger i området 1,3 til 8,7 eller 5,6 til 5,9. At dommernes bedømmelser spriker mye, kan i seg selv bidra til at forskjellene ikke blir signifikante: nevneren i F-testen blir for stor.

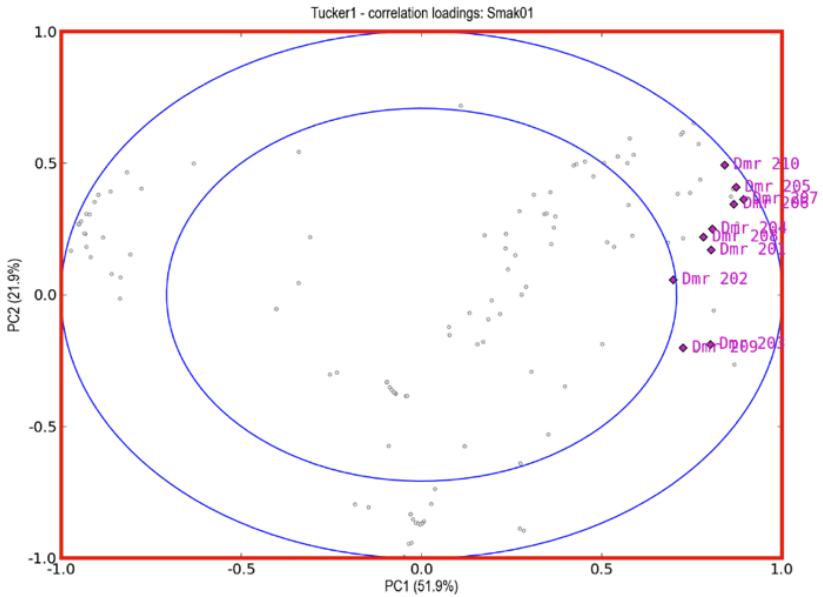
Om en dommer skulle avvike fra resten av panelet, er det også viktig å være klar over at dette kan skyldes at akkurat denne dommeren er spesielt sensitiv overfor denne egenskapen, og at det faktisk er denne som har «rett», mens alle de andre har «feil». Den ideelle bruken av et eggeskall-plot er situasjoner hvor fasiten er kjent, for eksempel prøver tilsatt kjente mengder sukker som skal bedømmes for søthet. Da kan eggeskall-plot benyttes for å vise dommernes prestasjoner.

F & p Plots og MSE Plots lager oversikter over F-, p- og MSE-verdiene fra variansanalysen. Denne typen figurer er godt egnet til å få en rask oversikt om det er enkelte egenskaper som stiller seg ut i en eller annen retning.

p-MSE Plots er et annet plot som kan hjelpe til å fortelle hvordan dommeren har prestert. Det tar utgangspunkt i at det gjøres en variansanalyse for hver dommer separat. MSE er et uttrykk for hvor godt dommeren har gjentatt seg selv, altså om vedkommende har bedømt forskjellige gjentak fra samme prøve omtrent likt. Samtidig er det av interesse å få vite om dommeren har kjent forskjell mellom prøvene, p er resultatet fra F-testen. Hvis p er liten, har dommeren kjent forskjell. Også disse plot'ene må tolkes med forsiktighet: en variansanalyse basert på et veldig lite datasett, som er vanlig når vi analyserer dommerne hver for seg, er i sin natur ustabilt. Men det gir allikevel en viss pekepinn på hvordan dommeren har prestert.

Tucker-1 Plots

Datasettet «brettes ut» ved at matrisene for enkeltdommernes bedømmelser (antall rader=antall prøver, antall kolonner=egenskaper) settes ved siden av hverandre og analyseres ved hjelp av PCA. Et Correlation loading plot avslører i hvilken grad dommerne tolker egenskapene forskjellig.



Figur 11.7: Tucker1-plot for Smak01

Her ligger dommerne godt samlet og stort sett nær den ytterste ellipsen; et tegn på at panelet fungerer godt for denne egenskapen. Hvis dommerne hadde vært mer uenige, ville de ligget spredd mer eller mindre ut over hele ellipsen. Med få prøver er Tucker-1 mindre egnet, siden dommerne ville havne nær den ytre ellipsen uansett. Derfor anbefales minst 7 prøver hvis denne analysen skal tillegges noe vekt.

Manhattan Plots

Dette er et annet alternativ for se på forskjeller mellom dommerne: her ser vi på forklart varians for hver dommer/egenskap-kombinasjon. Mange lyse felter i figuren forteller oss at dommerne har forklart mesteparten av variansen allerede etter de første prinsipalkomponentene. Forklarer dommerne en mindre del av variansen blir det desto flere mørke partier.

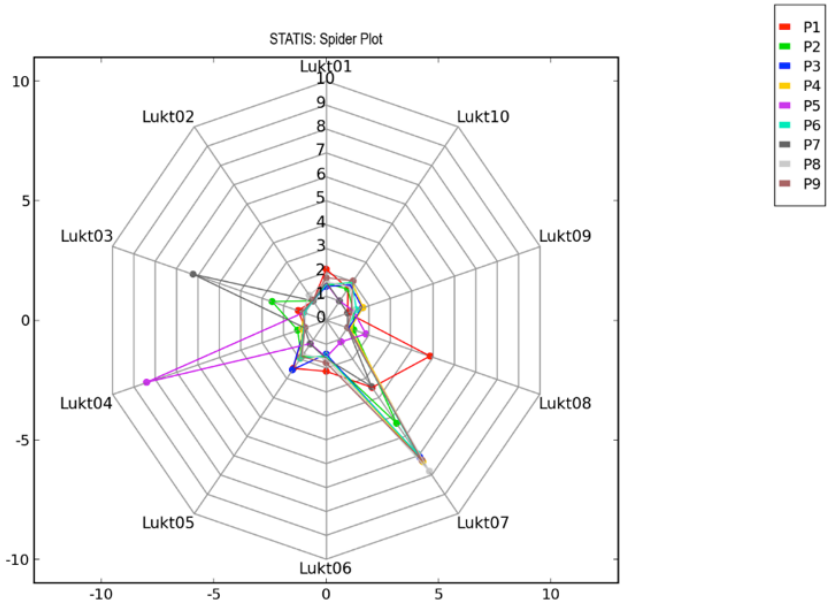
Consensus – Original – Standardized – STATIS

Disse 3 variantene gjør stort sett det samme: PCA enten på originaldata eller på standardiserte data (alle verdiene for en egenskap divideres med standardavviket over alle verdiene), dermed sikrer man at alle egenskapene får samme standardavvik. I sensorikk er det ikke så vanlig at dataene standardiseres, siden de i utgangspunktet er målt på samme skala. Standardisering benyttes heller i situasjoner hvor dataene består av målinger langs forskjellige skalaer, for eksempel hvis de sensoriske dataene skal inngå i en matrise sammen med vekt i gram, vannprosent og liknende. I tillegg til PCA kan STATIS vise enkeltdommernes bidrag til vektingen som STATIS benytter i de øvrige beregningene og et spider-plot, som er en form for linjeplot for middelverdiene av egenskapene.

Under fanen Overall kan PanelCheck utføre variansanalyser i noen utvalgte enkle, men ofte forekommende, modeller. I versjonen tilgjengelig per april 2015, er for eksempel følgende 3 ANOVA-modeller tilgjengelige:

- 2-way ANOVA (1 rep): Toveis variansanalyse uten gjentak i henhold til terminologien i denne boka
- 2-way ANOVA: Toveis variansanalyse med gjentak i henhold til terminologien i denne boka
- 3-way ANOVA: To faste effekter og dommer-effekt, alle faktorer krysset, ingen gjentak.

PanelCheck gir de samme resultatene som tradisjonelle statistikkprogram gir, men utskriften egner seg lite som vedlegg i rapporter.



Figur 11.8: Spider plot fra STATIS-delen

Referanser:

Daniel M Ennis, Benoît Rousseau, Technical Report: Identifying and removing sources of bias in product tests and surveys. The Institute for Perception, Newsletter Spring 2015.

David Hirst, Tormod Næs, A graphical technique for assessing differences among a set of rankings. J. Chemometrics 1994, 8, pp 81-93

Yosef Hochberg, Ajit C Tamhane, Multiple Comparison Procedures. New York: John Wiley & Sons, 1987. ISBN 0-471-82222-1

Daniel Kahneman, Thinking, fast and slow.
New York: Farrar, Strauss and Giroux, 2011. ISBN 978-0374275631
Norsk utgave: Tenke, fort og langsomt. Oslo: Pax forlag, 2012.
ISBN 978-82-530-3552-9

Per Lea, Marit Rødbotten, Tormod Næs, Measuring validity in sensory analysis. Food Quality and Preference 1995, 6, pp 321-326

Oliver Tomic, Ciaran Forde, Conor Delahunty, Tormod Næs, Performance indices in descriptive sensory analysis – A complimentary screening tool for assessor and panel performance. Food Quality and Preferences 2013, 28, pp 122-133

www.panelcheck.com.