Enveis variansanalyse (One-way ANOVA, fixed effects model)
(Notat til Kap. 12 i Rosner)

Vi rekapitulerer først t-testen for to uavhengige utvalg. Situasjonen var at vi hadde to grupper, f.eks. G_1 og G_2 og et sett uavhengige og identisk normalfordelte observasjoner fra begge.

Observasjoner fra G_1: $Y_{1j} \sim N(\mu_1, \sigma^2)$, $j=1,2,...,n_1$

Observasjoner fra G_2: $Y_{2j} \sim N(\mu_2, \sigma^2)$, $j=1,2,...,n_2$

Estimatorene for forventningene er henholdsvis $\hat{\mu}_1 = \frac{1}{n_1}\sum_{j=1}^{n_1} Y_{1j} = \overline{Y}_1$ og $\hat{\mu}_2 = \frac{1}{n_2}\sum_{j=1}^{n_2} Y_{2j} = \overline{Y}_2$.

Oppgaven er nå å teste hypotesen $H_0 : \mu_1 = \mu_2$ mot $H_1 : \mu_1 \neq \mu_2$. Forutsatt at variansen er den samme i begge grupper, får vi under H_0 testobservatoren

$$T = \frac{\hat{\mu}_1 - \hat{\mu}_2}{s\sqrt{1/n_1 + 1/n_2}}$$

der s er estimert felles standardavvik, og vi forkaster H_0 dersom

$$|T| > t_{n_1+n_1-2,1-\alpha/2}$$

I tilfelle H_0 forkastes, er konklusjonen enten at $\mu_1 > \mu_2$ eller at $\mu_1 < \mu_2$. Hvilken at de to alternativene er bestemt av størrelsen på de tilhørende estimatene $\hat{\mu}_1$ og $\hat{\mu}_2$. Vi ender således opp med en entydig konklusjon.

Nå tenker vi oss at vi har flere enn to grupper der vi skal sammenlikne forventningene. Dette blir på sett og vis en generalisering av to-utvalgstesten ovenfor. Anta at vi har k uavhengige grupper, hver med n_i uavhengige og normalfordelte observasjoner med varians σ_{ij}^2, $i=1,2,...,k$ og $j=1,2,...n_i$.
Vi får da:

Observasjoner fra G_1: $Y_{1j} \sim N(\mu_1, \sigma_{1j}^2)$, $j=1,2,...,n_1$

Observasjoner fra G_2: $Y_{2j} \sim N(\mu_2, \sigma_{2j}^2)$, $j=1,2,...,n_2$

.....

Observasjoner fra G_k: $Y_{kj} \sim N(\mu_k, \sigma_{kj}^2)$, $j=1,2,...,n_k$

Vi skal nå teste $H_0 : \mu_1 = \mu_2 = ... = \mu_k$. Alternativet til H_0 rommer en rekke muligheter. Dersom k=3, kan vi f.eks. ha at enten er alle forventningene forskjellige, eller så er to like og den tredje forskjellig fra de to. Generelt kan H_1 formuleres som $H_1 : \mu_i \neq \mu_{i'}$ for minst ett

par $\mu_i, \mu_{i'}$.
Vi skal m.a.o. teste

$$H_0 : \mu_1 = \mu_2 = ... = \mu_k$$

mot

$$H_1 : \mu_i \neq \mu_{i'}, i \neq i' \text{ for minst ett par } \mu_i, \mu_{i'}.$$

Dersom resultatet av testen er forkastning av H_0, blir det neste å finne ut hva avviket fra H_0 består av. Dette skal vi komme tilbake til senere.

Vi har følgende modell:

$$Y_{ij} = \mu + \alpha_i + e_{ij}$$

Y_{ij} er j-te observasjon i gruppe i
μ er forventningen til alle Y_{ij}-ene samlet sett, "grand mean".
α_i er avviket i i-te gruppes forventning fra "grand mean" μ.

Vi har mao at hver enkelt observasjon består av en konstant μ , et gruppespesifikt tillegg (eller fradrag) α_i og et stokastisk tillegg (fradrag) e_{ij} .
Vi forutsetter at e_{ij}-ene er normalfordelte med forventning null og fra nå av *samme* ukjente varians σ^2 , slik at

$$e_{ij} \sim N(0, \sigma^2) \text{ for alle } i \text{ og } j$$

Dessuten forutsettes at alle e_{ij}-ene er uavhengige. Modellen gir at $E(Y_{ij}) = \mu_i = \mu + \alpha_i$. Dette medfører at $H_0 : \mu_1 = \mu_2 = ... = \mu_k$ er ekvivalent med $H_0 : \alpha_1 = \alpha_2 = ... = \alpha_k = 0$. Alternativene blir $H_1 : \alpha_i \neq 0$ for minst én i.

Vi definerer noen notasjoner til senere bruk:

$\bar{Y}_i = \frac{1}{n_j} \sum_{j=1}^{n_i} Y_{ij}$, dvs gjennomsnittet av alle observasjonene i gruppe i

$Y_{..} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} Y_{ij}$, dvs summen av alle observasjonene

$N = \sum_{i=1}^{k} n_i$, dvs antall observasjoner totalt

$\bar{\bar{Y}} = \frac{Y_{..}}{N}$, dvs gjennomsnittet av alle observasjonene

Dersom $n_i = n$ for alle i (likt antall i alle grupper), har vi *balansert design*, noe som bør tilstrebes fordi det gir optimalitet i testing.

For å klargjøre notasjonen samler vi alle observasjonene i en tabell:

Gruppe 1	*Gruppe 2*	...	*Gruppe k*
Y_{11}	Y_{21}	...	Y_{k1}
Y_{12}	Y_{22}	...	Y_{k2}
...	...	...	...
...	...	...	...
Y_{1n_1}	Y_{2n_2}	...	$\overline{Y}_{kn_k}$
$\overline{Y}_1$	$\overline{Y}_2$	...	$\overline{Y}_k$

Merk notasjonen. For eksempel vil Y_{34} bety observasjon nr. 4 i gruppe 3, osv.

Det kan vises at $\hat{\mu}_i = \overline{Y}_i$ er en forventningsrett estimator for $\mu_i = \mu + \alpha_i$ og at $Var(\hat{\mu}_i) = \frac{\sigma^2}{n_i}$.

Variansen σ^2 kan i prinsipp estimeres separat innen hver gruppe, men det er mer effektivt å estimere den ut fra samtlige data, noe vi skal komme tilbake til.

Som i regresjonsanalysen betrakter vi nå k*vadratsummer* basert på variasjon i dataene. Den totale kvadratsummen

$$\text{Total SS} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij} - \overline{\overline{Y}})^2$$

er et uttrykk for dataens (responsvariabelens) variasjon rundt det store gjennomsnittet. Den del av Total SS som forklares av modellen, uttrykkes ved de enkelte gruppers variasjon rundt det store gjennomsnittet uttrykt ved kvadratsummen

$$\text{Between SS} = \sum_{i=1}^{k} n_i(\overline{Y}_i - \overline{\overline{Y}})^2$$

Det som nå gjenstår av den totale variasjonen, er den som skyldes variasjonen innen de enkelte grupper, noe som vi tilskriver de tilfeldige avvik og som senere danner grunnlag for estimering av variansen σ^2. Kvadratsummene dette representerer, uttrykkes som

$$\text{Within SS} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y}_i)^2$$

Det kan vises at

$$\text{Total SS=Between SS + Within SS}$$

Vi ser at jo mer ett eller flere av gruppegjennomsnittene $\overline{Y}_i$ avviker fra det totale gjennomsnitt $\overline{\overline{Y}}$, desto større blir andelen av modellforklart variasjon i forhold til total variasjon, noe som trekker i retning av at nullhypotesen må forkastes. Men for å kunne ta en formell beslutning, må vi konstruere en testobservator med kjent fordeling når H_0 gjelder.

Som i regresjonsanalysen samler vi resultatene i en ANOVA-tabell:

Kilde til variasjon	Kvadratsum (*SS*)	*df*	$MS = \frac{SS}{df}$	F	p
Between (grupper imellom)	$\text{Between SS} = \sum_{i=1}^{k} n_i(\bar{Y}_i - \bar{\bar{Y}})^2$	k-1	$\text{Between MS} = \frac{\text{Between SS}}{k-1}$	$F_0 = \frac{\text{Between MS}}{\text{Within MS}}$	
Within (innen grupper, residual)	$\text{Within SS} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_i)^2$	N-k	$\text{Within MS} = \frac{\text{Within SS}}{N-k}$		
Total	$\text{Total SS} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij} - \bar{\bar{Y}})^2$	N-1			

Det kan vises at Within MS er en konsistent estimator for σ^2, variansen til støyleddene og dermed til Y. En av forutsetningen i ANOVA-modellen var at variansen var lik i alle grupper. De fleste programpakker har test for dette.

Under H_0 har vi at

$$F_0 = \frac{\text{Between MS}}{\text{Within MS}} \sim F(k-1, N-k)$$

dvs F-fordelt med *(k-1, N-k)* frihetsgrader. Dersom observert $F_0 > F_{k-1,N-k,1-\alpha}$, forkastes H_0, og vi konkluderer at minst ett par μ_i, $\mu_{i'}$ er forskjellige, eller ekvivalent, minst én $\alpha_i \neq 0$. Noe utover det kan vi foreløpig ikke uttale oss om.

Etter ANOVA

ANOVA-kalkulasjonene tester nullhypotesen at alle observasjonene fra alle gruppene stammer fra fordelinger med samme forventning. Å teste denne hypotesen alene er sjelden formålet med en studie. Hensikten er som regel å vurdere gruppene seg imellom, for eksempel å sammenlikne grupper parvis eller gruppesammensetning mot en annen gruppesammensetning. *Multiple sammenlikninger* er en fellesbetegnelse for dette.

Dersom *F*-testen ovenfor ikke gir signifikant resultat, er saken avgjort: De observerte data gir ikke grunn til å hevde at gruppeforventningene er forskjellige, og vi stopper der. Slår derimot testen signifikant ut, må vi gå videre og finne ut hva avviket fra H_0 består i. Men her fins ingen automatikk. Anta at vi har 5 grupper betegnet A, B, C, D og F. Prinsipielt kan vi tenke oss 4 situasjoner:

1. Alle mulige sammenlikninger, også gjennomsnitt eller evt. veide gjennomsnitt av grupper. For eksempel kan det være interessant å sammenlikne gjennomsnittet av gruppe A og B mot gjennomsnittet av gruppe C, D og E. Eller sammenlikne A mot gjennomsnittet av B, C og D osv. Til dette formålet en Scheffé's test den rette.
2. Alle mulige parvise sammenlikninger. Har vi *g* grupper, utgjør dette i alt

$g(g-1)/2$ enkelttester. Til dette formålet benyttes Tukey eller Newman-Keuls test.

3. Alle mot én kontroll. Dersom gruppe A er kontroll, skal det testes mot gruppe B, C, D, E og F. Her benytter Dunnett's test.
4. Bare noen få sammenlikninger bestemt på forhånd ut fra vitenskapelige kriterier. For eksempel gruppe A mot B og C, og det et alt. Hen benyttes Bonferroni's test.

Multiple sammenlikninger betyr at det utføres flere sammenlikninger samtidig slik som i alle 4 situasjoner ovenfor.

Post hoc- test brukes i situasjoner der man *etter* å ha studert de aktuelle data avgjør hvilke tester som skal utføres. Man trenger ikke å planlegge på forhånd. Situasjon 1 ovenfor er åpenbart post hoc-testing. Situasjon 2 og 3 krever at det gjøres noen beslutninger basert på konkrete problemstillinger. Hvis man for eksempel sammenlikner alle par av grupper, har man implisitt besluttet ikke å utføre alle de generelle sammenlikninger nevnt i punkt 1. I situasjon 3 har man besluttet ikke å sammenlikne alle grupper seg imellom, bare å sammenlikne mot én kontroll. Fordi bruken av disse testene krever beslutninger foretatt før dataene observeres, er situasjon 2 og 3 strengt tatt ikke post hoc-tester.

Planlagte sammenlikninger forutsetter at man konsentrerer seg om et fåtall vitenskapelig baserte problemstillinger. Det er ikke tillatt å la dataene tale for seg og på det grunnlag velge ut hva som skal sammenliknes. Hva som skal testes, må være bestemt på forhånd som ledd i design av en studie eller et eksperiment i vid forstand. Situasjon 4 er et typisk eksempel på dette.

Terminologien ovenfor er ikke alltid brukt konsistent. Tester for multiple sammenlikninger brukes ofte synonymt med post hoc-tester. Strengt tatt utgjør sistnevnte tester en undergruppe av multiple sammenlikninger og benyttes i de tilfeller der man etter å ha observert dataene bestemmer seg for hvilke sammenlikninger som skal gjøres. I slike tifeller stilles det høyere krav til testnivået for ikke å begå feil av Type I.

Alle multiple sammenlikningstester er forskjellige fra *ordinære* t-tester. Selv ved sammenlikning av to og to grupper, der en standard toutvalgs t-test ville være nærliggende, bruker testen et variansestimat basert på data fra alle gruppene, også fra de gruppene som ikke er direkte involverte i hypotesen. Anta at vi har g grupper og skal teste $H_0: \mu_1 = \mu_2$ mot $H_0: \mu_1 \neq \mu_2$. Vi har da at

$$\overline{Y}_1 \sim N(\mu_1, \frac{\sigma^2}{n_1})$$

og

$$\overline{Y}_2 \sim N(\mu_{2'}, \frac{\sigma^2}{n_2})$$

Vi forkaster H_0 dersom

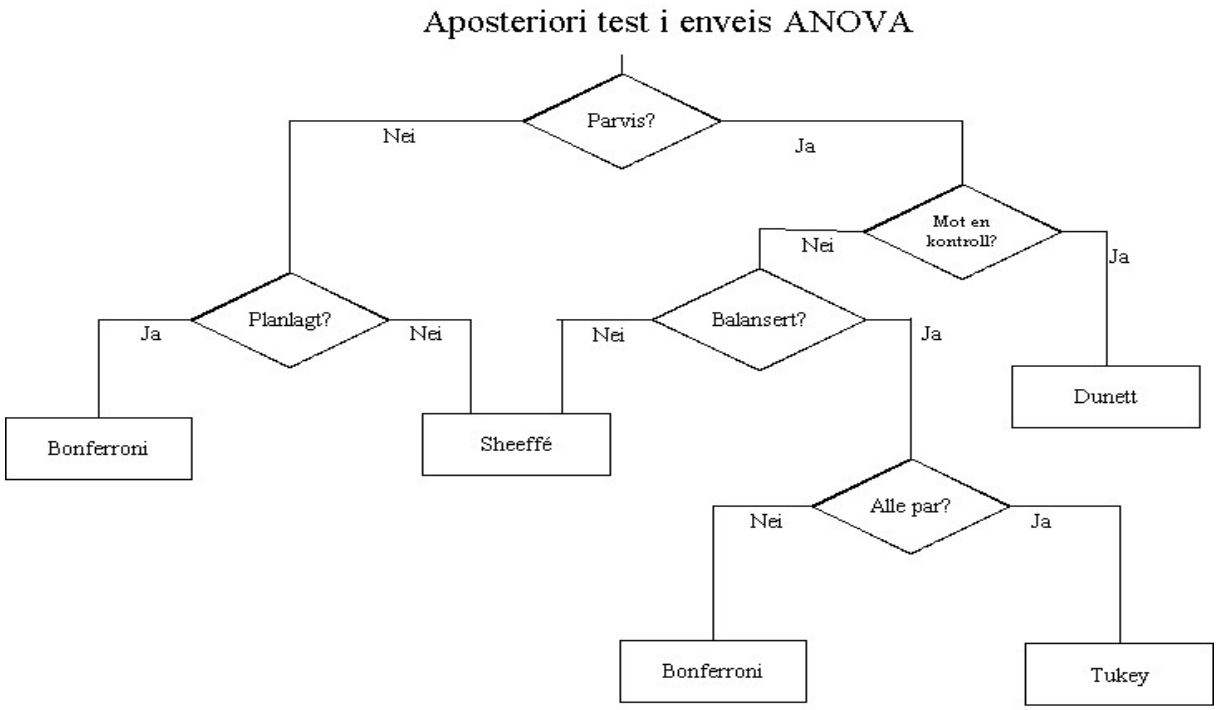
$$|\overline{Y}_1 - \overline{Y}_2| > t_{N-g,1-\alpha/2}\sqrt{\text{Within MS}\left(1/n_1 + 1/n_2\right)}$$

Merk at variansen σ^2 estimeres med *Within MS* fra ANOVA-tabellen. Dette gir flere frihetsgrader og derved et presisere estimat enn kun å bruke dataene fra de to gruppene vi sammenlikner. Det gir mindre signifikanssannsynlighet (lavere p-verdi) og større teststyrke. Generelt gjelder det i statistikk at en bør tilstrebe å bruke mest mulig tilgjengelig informasjon. I eksemplet ovenfor er det benyttet informasjon fra grupper som selv ikke er aktuelle for sammenlikning, men som bidrar med informasjon om variansen.

Det andre aspektet ved multiple sammenlikninger er forsøket på å opprettholde signifikansnivået for *hele familien* av sammenlikninger – snarere enn nivået til den enkelte sammenlikningstest. Det betyr at dersom alle gruppene i virkeligheten hadde den samme forventningsverdi, er sannsynligheten α for at en hvilken som helst eller flere av sammenlikningene ville få en statistisk signifikant konklusjon ved ren tilfeldighet. For å oppnå dette må metodene for multiple sammenlikninger benytte en mer strikt definisjon av signifikans. En enkel korreksjon for slike multiple t-tester er å stramme inn nivået for den enkelte t-test til α / m, der α er nivået i *F*-testen og *m* er antall parsammenlikninger. Dette kalles *Bonferroni*-korreksjon. Metoden er imidlertid konservativ i den forstand at det i mange tilfeller stilles unødig strenge krav til nivået til den enkelte test.

Enkelte statistikere anbefaler ikke å korrigere for multiple sammenlikninger når man bare utfører et fåtall, opptil *g-1*, planlagte sammenlikninger. Begrunnelsen er at man skal få en slags bonus som belønning for å ha planlagt en målrettet studie. Argumentet er dette: Anta at det på forhånd planlegges tre sammenlikninger i ett og samme eksperiment. Alternativet er å utføre tre separate eksperimenter. I så tilfelle trengte man ikke å justere for multiple sammenlikninger, men hvorfor gjøre det dersom man planla på forhånd og utførte sammenlikningene som del av ett og samme eksperiment designet til det bestemte formålet? Andre statistikere mener at man alltid skal justere for multiple sammenlikninger. Ettersom saken er kontroversiell, er det nødvendig at man beskriver sin problemstilling og metode detaljert.

Harald Johnsen, februar 2008



Utfyllende kommentarer

Dersom alle parvise sammenlikninger er av interesse, er Tukeys metode bedre enn Bonferronis metode ved at den gir smalere konfidensintervall. Dersom ikke alle parsammenlikninger er aktuelle, er Bonferronis metode vanligvis bedre.

Bonferronis metode er bedre enn Scheffés metode når antall kontraster som skal estimeres, er om lag det samme som antall faktornivå (antall grupper), eller mindre. Antall utsagn må overstige antall faktornivå betraktelig før Scheffés metode blir bedre.

Uansett kan man ved et gitt problem benytte både Tukeys og Bonferronis metode og så velge den som gir "best" resultat. Dette er tillatt fordi ingen av metodene avhenger at de observerte data.

Bonferronis metode støtter ikke "data-snooping" med mindre man på forhånd kan spesifisere familien av utsagn som er av interesse, og at denne familien ikke er for stor. Både Tukeys og Scheffés metode støtter data-snooping.

HJ okt. 2003