
Parvise sammenlikninger mellom tre grupper

MEDISIN OG TALL

STIAN LYDERSEN

stian.lydersen@ntnu.no

Stian Lydersen er dr.ing. og professor i medisinsk statistikk ved Regionalt kunnskapssenter for barn og unge – psykisk helse og barnevern (RKBU Midt-Norge) ved Institutt for psykisk helse, NTNU. Forfatteren har fylt ut ICMJE-skjemaet og oppgir ingen interessekonflikter.

I noen studier vil man sammenlikne tre eller flere grupper. Dette kan for eksempel være en randomisert kontrollert studie med flere behandlinger. Da er det relevant å gjøre parvise sammenlikninger mellom gruppene.

Hvis studien omfatter tre grupper – A, B og C – kan man gjøre inntil tre parvise sammenlikninger i form av tre hypotesetester. Og hvis studien omfatter fire grupper – A, B, C og D – er det mulig å gjøre inntil seks parvise sammenlikninger: A-B, A-C, A-D, B-C, B-D og C-D. Når man har flere hypoteser, er det aktuelt å kontrollere familievis feilrisiko (*familywise error rate*, FWER), slik at sannsynligheten for å feilaktig påstå at det er forskjell mellom minst ett par grupper ikke overstiger et overordnet signifikansnivå, vanligvis 5 %. I prinsippet kan man gjøre dette ved å regne ut en p -verdi for hver av de parvise sammenlikningene, og deretter justere p -verdiene med en av metodene tidligere beskrevet i denne spalten ([1](#)). Men når man gjør parvise sammenlikninger, finnes det metoder som tar hensyn til den parvise strukturen i hypotesene, og som har vesentlig høyere statistisk styrke.

Ulike metoder

Det finnes mange ulike metoder for å gjøre parvise sammenlikninger. Valg av metode er avhengig av flere forutsetninger: Skal man gjøre parvise sammenlikninger mellom alle gruppene, eller for eksempel bare mot en kontrollgruppe? Er variansen lik eller ulik? Og er gruppestørrelsene like eller ulike? En oversikt i [\(2\)](#) lister opp totalt 16 forskjellige anbefalte metoder under forskjellige forutsetninger. Hvis dataene er normalfordelt, anbefales Tukeys test for alle parvise sammenlikninger, eller Dunnetts test ved sammenlikninger bare mot en kontrollgruppe. Men denne anbefalingen gjelder strengt tatt bare ved tilnærmet like store grupper og lik varians i gruppene. Det kan være vanskelig å velge en egnet metode, selv når dataene er normalfordelt.

Tre grupper

Hvis studien omfatter bare tre grupper, hvilket ofte er tilfelle, finnes en vesentlig enklere fremgangsmåte, som ikke engang krever forutsetninger om fordeling eller gruppestørrelse: Først beregner man den globale p -verdien for nullhypotesen om at alle de tre gruppene er like. Deretter beregner man den ujusterte p -verdien enkeltvis for hver av de tre parvise sammenlikningene. Til slutt justerer man hver av disse tre p -verdiene ved å erstatte dem med den globale p -verdien, dersom den globale p -verdien er størst. Dette er illustrert i et eksempel nedenfor. Denne prosedyren kontrollerer alltid for familievis feilrisiko [\(3\)](#), noe mange forskere ikke synes å være klar over. Selv om dataene er normalfordelt, og man kunne brukt Tukeys test, vil denne enkle metoden gi minst like høy statistisk styrke ved tre grupper [\(4\)](#).

Hvis dataene er normalfordelt, kan man regne ut den globale p -verdien fra en enveis variansanalyse, og deretter sammenlikne parvis med t -tester. Hvis man bruker ikke-parametriske metoder, kan man først gjøre en global Kruskal-Wallis-test, og deretter parvise Wilcoxon-Mann-Whitney-tester. Og hvis dataene er kategoriske, kan man først gjennomføre Pearsons khikvadrattest for tre grupper og deretter Pearsons khikvadrattest for hver av de tre parvise sammenlikningene.

La oss illustrere dette med et eksempel: Weider og medarbeidere sammenliknet kognitiv funksjon hos tre grupper personer, nemlig 41 med anoreksi, 40 med bulimi og 40 friske kontrollpersoner (5, tabell 3). Wechslers intelligensskala [\(5\)](#) viste en gjennomsnittsscore (standardavvik) på henholdsvis 10,51 (3,26), 10,00 (2,42) og 11,85 (2,83) i de tre gruppene. Den globale p -verdien fra enveis variansanalyse var 0,014. P -verdier for parvise sammenlikninger for noen alternative metoder er vist i tabell 1. Ved å bruke denne metoden ser vi at både anoreksigruppen og bulimigruppen fremstår som signifikant forskjellige fra kontrollgruppen ved signifikansnivå 5 %. Hvis man brukte Tukeys eller Dunnetts test, ville bare forskjellen mellom bulimigruppen og kontrollgruppen bli signifikant.

Tabell 1

Parvise sammenlikninger for Wechslers intelligensskala mellom personer med anoreksi (A), bulimi (B) og friske kontrollpersoner (K) (basert på data fra (5), tabell 3). Global p -verdi fra enveis variansanalyse var 0,014. Ujusterte p -verdier ble beregnet ved LSD (*least significant difference*), som er en generalisering av t -testen.

Par	Ujustert p -verdi	Justert p -verdi		
	LSD	Tukey	Dunnett	Maksimum av global og ujustert
A-B	0,422	0,701		0,422
A-K	0,038	0,094	0,069	0,038
B-K	0,005	0,013	0,009	0,014

Kun ved tre grupper

Det må presiseres at den beskrevne metoden bare kontrollerer familievis feilrisiko ved tre grupper. Hvis man for eksempel sammenlikner tre forskjellige behandlinger mot en kontrollgruppe, er fire grupper involvert, og en slik prosedyre vil ikke kontrollere for familievis feilrisiko. For øvrig er metoden enkel å gjennomføre, har høy statistisk styrke, og kan alltid anbefales ved parvise sammenlikninger mellom tre grupper.

LITTERATUR

1. Lydersen S. Justering av p -verdier ved multiple hypoteser. Tidsskr Nor Legeforen 2021; 141. doi: 10.4045/tidsskr.21.0357. [CrossRef]
2. Kirk RE. Experimental design. Procedures for the behavioral sciences. 4. utg. Thousand Oaks: Sage Publications, 2013.
3. Levin JR, Serlin RC, Seaman MAA. Controlled, Powerful Multiple-Comparison Strategy for Several Situations. Psychol Bull 1994; 115: 153–9. [CrossRef]
4. Seaman MA, Levin JR, Serlin RC. New Developments in Pairwise Multiple Comparisons: Some Powerful and Practicable Procedures. Psychol Bull 1991; 110: 577–86. [CrossRef]
5. Weider S, Indredavik MS, Lydersen S et al. Intellectual function in patients with anorexia nervosa and bulimia nervosa. Eur Eat Disord Rev 2014; 22: 15–24. [PubMed][CrossRef]