NTNU
Kunnskap for en bedre verden

Multiple hypoteser eller multiple sammenlikninger

av
Stian Lydersen, professor i medisinsk statistikk, RKBU Midt-Norge

RBUP, Oslo, 9 oktober 2015

Slides revidert 13 oktober 2015
http://folk.ntnu.no/slyderse/medstat/Multcomp9okt2015.pdf

Multiple hypoteser eller multiple sammenlikninger.

Hva bør du gjøre hvis du skal undersøke flere hypoteser samtidig? Dette er tilfelle hvis du har mer enn én avhengig variabel, dvs flere endepunkt, eller hvis mer enn to grupper (feks eksponeringsgrupper eller behandlingsgrupper) skal sammenliknes. Dersom hver av hypotesene testes ved signifikansnivå α, vil sannsynligheten for å gjøre minst en type I feil (FWER, familywise error rate) bli betydelig større enn α.

I denne forelesningen vil jeg beskrive metoder for å kontrollere FWER, og forklare i hvilke situasjoner de er anbefalt brukt.

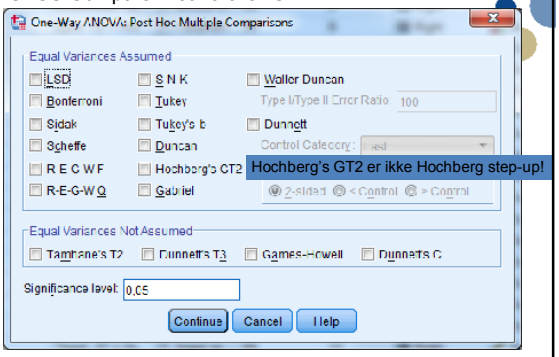
Dette inkluderer generelle metoder som Bonferroni, Šidák, og Hochberg, samt metoder knyttet til enveis ANOVA som Dunnett, Dunn, Tukey og Scheffé. En spesiell situasjon er når man skal sammenlikne bare tre grupper. Her er det faktisk ikke nødvendig å justere (ved Bonferroni e.l.) for å bevare FWER, såfremt man gjør parvise sammenlikninger bare hvis omnibus-hypotesen (alle tre gruppene er like) forkastes.

Innhold

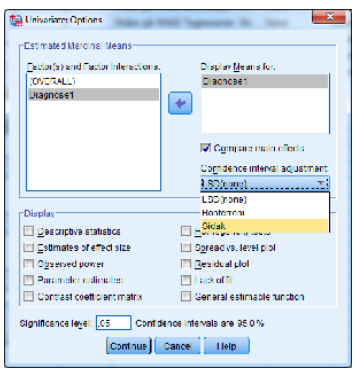
- Når bør vi justere for multiple hypoteser?
- I så fall, hva vil vi kontrollere
  - Familywise error rate (FWER): Sannsynligheten for minst ett falskt positivt funn
  - False discovery rate (FDR): Forventet andel falske positive funn blant våre positive funn
- P-verdibaserte metoder:
  - Bonferroni, Šidák, Holm step-down, Hochberg step-up
- Sekvensielle metoder basert på «the closed testing principle»
- Parametriske metoder for parvise sammenlikninger i enveis ANOVA:
  - Dunnett, Tukey, Scheffé, …
- Parvise sammenlikninger, tre grupper, generelt: Tilstrekkelig å kombinere en global test med ujusterte(!) lokale tester

3

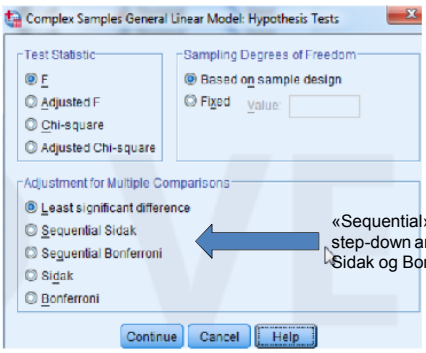
SPSS Compare means eller GLM



SPSS GLM med minst en kovariat



SPSS Complex samples GLM



Eksempel
Kaasboll, J., Lydersen, S., & Indredavik, M. S. 2012, "Psychological symptoms in children of parents with chronic pain-the HUNT study", Pain, vol. 153, no. 5, pp. 1054-1062.

4 grupper ungdommer: Foreldre med kroniske smerter:
•Ingen,
•bare mor (M),
•bare far (F),
•både mor og far (MF)

2 avhengige variable:
•Angst/depresjon (SCL-5)
•Atferdsproblemer

Separate analyser for jenter og gutter

Kaasbøll J, Lydersen S, Indredavik M: (Pain, 2012)
"Psychological symptoms in children of parents with chronic pain – the HUNT study"

Results adjusted for age – boys:

Parents with chronic pain	Number of children	Risk for conduct problems: Odds ratio (OR)		
		estimate	Conf. int.	P-value
None	801	1 (ref.)		
Only mother	289	1.30	1.02 to 1.67	0.036
Only father	216	0.99	0.74 to 1.31	0.93
Both parents	117	1.36	0.96 to 1.93	0.087

Tre(?) problemstillinger:

a) Flere avhengige variable
b) Sammenlikning mellom flere enn 2 grupper
c) Flere subgrupper

9

Flere avhengige variable
("Uavhengige" eller "urelaterte" hypoteser)
Alternativer:

- Rothman, K. J. "No adjustments are needed for multiple comparisons" (1990). "Six persistent research misconceptions" (2014).
  - Justere når flere hypoteser i samme artikkel?
  - Splitte opp i flere artikler?
  - Justere for alle hypoteser jeg tester i min karriere?
  - Justere for alle hypoteser menneskeheten tester?
- Ett primært endepunkt, legge "mindre vekt" på signifikante funn på sekundære endepunkt
- Pragmatisk forslag (Peter Fayers): Two-sided p-values < 0.05 are taken to indicate statistical significance. Due to multiple hypotheses, p-values between 0.01 and 0.05 should be interpreted with caution.
- Kontrollere FWER (familywise error rate)
- Kontrollere FDR (false discovery rate)

10

Eksempel:
Senn, S. 2007, Statistical issues in drug development, 2nd ed, John Wiley. Chapter 10: Multiplicity
RCT, medikamenter for lungefunksjon hos astmatikere:

Mulige avhengige variable:
•FEV
•PEV
•RAW
•…
•Bivirkninger

"Multiplicity continues to be a field of much research and controvercy."

	Akseptert	Forkastet	Totalt
Sanne nullhypoteser	U	V	m_0
Usanne nullhypoteser	T	S	m-m_0
	m-R	R	m

Per comparison error rate PCER: $E[V/m]$
Hvis hver nullhypotese testes med nivå α :
Hvis alle nullhypotesene er sanne: $E[V/m] = \alpha$
Hvis noen nullhypoteser er usanne: $E[V/m] < \alpha$

Family-wise error rate FWER: $P(V \geq 1)$
(Also called experimentwise error rate)

False discovery rate FDR: $E[V/R]$
er forventet andel sanne nullhypoteser blant de som forkastes.

Weak control of FWER:
FWER ≤ α given that all null hypotheses are true (m_0=m)

Strong control of FWER:
FWER ≤ α for any combination of true and false null hypotheses

Strong control of FWER is mandated by regulators in the USA in all confirmatory clinical trials (Phase III) trials.

In the rest of this presentation, control of FWER means in the strong sense.

Problem:

Ved å bruke signifikansnivå α på hver av hypotesetestene, blir sannsynligheten for minst ett falskt positivt funn, dvs FWER, (betydelig) større enn α.

Control of FDR is especially relevant with a large number of null hypotheses, such as in genomics.

Goeman, J.J. & Solari, A. 2014. Tutorial in biostatistics: Multiple hypothesis testing in genomics. Statistics in Medicine, 33, (11) 1946-1978

Distributional assumptions:

- P-value based procedures: Based on univariate p-values. No assumptions about joint distribution of test statistics. F.ex. Bonferroni, Šidák
- Semiparametric p-value based procedures. F.ex. Šidák, Hochberg, Hommel assuming independent or positively correlated p-values.
- Parametric procedures assuming a specific multivariate distribution. F.ex. Dunnett, Dunn, Tukey, Scheffé in one-way ANOVA (without additional covariates) assume normal distribution.
- Resampling-based procedures (bootstrap or permuation methods). Control FWER only when sample size approaches infinity (Dmitrienko and D'Agostino, 2013, page 5181)

P-value based procedures tend to perform poorly compared to parametric and resampling based procedures, when there are many hypotheses or strongly correlated test statistics. (Dmitrienko et al, 2010, page 49 - 50)

16

m hypoteser, p-verdier $p_{(1)} \le p_{(2)} \le \ldots \le p_{(m)}$.

Bonferronikorreksjon:
Forkast bare hypoteser med $mp_{(j)} \le \alpha$ (ekvivalent: $p_{(j)} \le \tilde{\alpha} = \alpha / m$)
Kontrollerer alltid $FWER \le \alpha$

Bonferroniulikheten gjelder alltid:
$P(F_1 \cup F_2 \cup \ldots \cup F_m) \le P(F_1) + P(F_2) + \ldots + P(F_m) = m\tilde{\alpha}$

«Union» betyr «(og)/eller»

Šidák korreksjon:

Forkast bare hypoteser med $p_{(j)} \le \tilde{\alpha} = 1-(1-\alpha)^{1/m}$

Kontrollerer $FWER \le \alpha$ ved
uavhengige p-verdier,
eller uavhengige testobservatorer,
eller multinormalfordelte testobservatorer.

Šidák-ulikheten gjelder med = for uavhengige og < for positivt korrelerte p-verdier:

$P(F_1 \cup F_2 \cup \ldots \cup F_m)$
$\le 1-[1-P(F_1)][1-P(F_2)]\ldots[1-(F_m)] = 1-(1-\tilde{\alpha})^m$

Eksempler:

	α	m	$\tilde{\alpha}$
Bonferroni	0.05	3	0.01667
Šidák	0.05	3	0.01695
Bonferroni	0.05	10	0.00500
Šidák	0.05	10	0.00512

«The Šidák procedure provides a fairly small improvement over the Bonferroni procedure … For this reason, the Šidák procedure and related multiple testing procedures have not found many applications in a clinical setting …» (Dmitrienko and D'Agostino 2013, page 5188)

Kirk (2013, Tabell 5.1-1 side 169) anbefaler Šidák-korreksjon fremfor Bonferronikorreksjon for Dunn's test.

Holm step-down correction (også kalt Bonferroni step-down correction)
Hvis $p_{(1)} \le \alpha_{(1)} = \alpha / m$ forkast hypotese (1)
Hvis hypotese (1) ble forkastet:
Hvis $p_{(2)} \le \alpha_{(2)} = \alpha / (m-1)$ forkast hypotese (2)
Hvis hypotese (2) ble forkastet:
Hvis $p_{(3)} \le \alpha_{(3)} = \alpha / (m-2)$ forkast hypotese (3)
…
Hvis hypotese (m-1) ble forkastet:
Hvis $p_{(m)} \le \alpha_{(m)} = \alpha / 1 = \alpha$ forkast hypotese (m)

Dvs:
I steg j, bruk Bonferronikorreksjon på de $m - j + 1$ siste p-verdiene.

Kontrollerer alltid $FWER \le \alpha$.
Uniformt mindre konservativ (høyere styrke) enn Bonferroni

Hochberg step-up correction

Hvis $p_{(m)} \le \alpha_{(m)} = \alpha$ forkast alle m hypotesene.
Hvis $p_{(m-1)} \le \alpha_{(m-1)} = \alpha / 2$ forkast hypotese (m-1), (m-2), …, 1
Hvis $p_{(m-2)} \le \alpha_{(m-2)} = \alpha / 3$ forkast hypotese (m-2), …,1
…
Hvis $p_{(1)} \le \alpha_{(1)} = \alpha / m$ forkast hypotese (1)
Kontrollerer også $FWER \le \alpha$ under forholdsvis generelle betingelser (multivariat normalfordelte testobservatorer med ikke-negative korrelasjonskoeffisienter)
Mindre konservativ (høyere styrke) enn Holm (og Bonferroni)

NB! Hochberg som kontrollerer FWER må ikke forveksles med Benjamini-Hochberg som bare kontrollerer FDR!

Holm step-down correction | Hochberg step-up correction

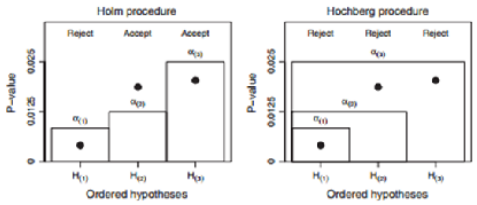


Figure 8. Adjusted significance levels used in the Holm procedure (left panel) and Hochberg procedure (right panel) in a multiplicity problem with three null hypotheses H_1, H_2, and H_3 (ordered one-sided p-values, $p_{(1)} = 0.0041$, $p_{(2)} = 0.0187$, and $p_{(3)} = 0.0204$ are represented by dots).

Dmitrienko and D'Agostino 2013

Hommel's (1988) procedure is more powerful than Hochberg's but is more difficult to understand and apply. Let j be the largest integer for which $p_{(m-j+k)} < k\alpha/j$ for all $k = 1,\ldots, j$. If no such j exists, reject all hypotheses; otherwise, reject all H_i with $p_i \le \alpha/j$. Both i and j, by the way, go from 1 to m.

(Shaffer, 1995, page 571, adapted to our notation with *m* instead of *n*)

Hommel's procedure is explained in a slightly different way in the Appendix of Dmitrienko and d'Agostino (2013). See also Wright (1992).

Benjamini-Hochberg step-up

Hvis $p_{(m)} \le \alpha_{(m)} = m\alpha / m = \alpha$ forkast alle m hypotesene.
Hvis $p_{(m-1)} \le \alpha_{(m-1)} = (m-1)\alpha / m$ forkast hypotese (m-1), (m-2), …, 1
Hvis $p_{(m-2)} \le \alpha_{(m-2)} = (m-2)\alpha / m$ forkast hypotese (m-2), …,1
…
Hvis $p_{(1)} \le \alpha_{(1)} = \alpha / m$ forkast hypotese (1)

Kontrollerer ikke $FWER \le \alpha$
"The Benjamini & Hochberg procedure has valid FDR control if the test statistics underlying the p-values are positively correlated for one-sided tests, under a different condition that allows some negative correlations for two-sided tests." (Goeman and Solari 2014)

m hypoteser, p-verdier $p_{(1)} \le p_{(2)} \le \ldots \le p_{(m)}$.

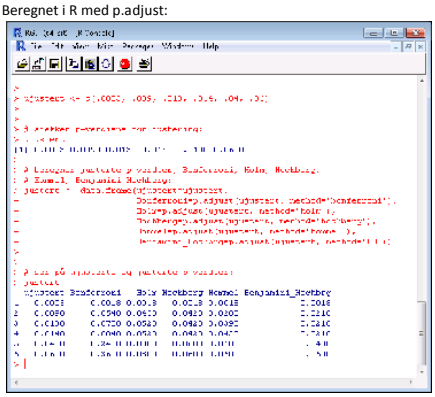
p-value unad-justed	Bon-fer-roni	Holm step-down	Hochberg step-up	Benjamini-Hochberg step-up
$p_{(1)}$	$mp_{(1)}$	$mp_{(1)}$	$\min\left(mp_{(1)},\ldots,\frac{mp_{(m)}}{m}\right)$	$\min\left(\frac{mp_{(1)}}{1},\ldots,\frac{mp_{(1)}}{m}\right)$
$p_{(2)}$	$mp_{(2)}$	$\max\left(mp_{(1)},(m-1)p_{(2)}\right)$	$\min\left((m-1)p_{(2)},\ldots,\frac{mp_{(m)}}{m}\right)$	$\min\left(\frac{mp_{(2)}}{2},\ldots,\frac{mp_{(1)}}{m}\right)$
$p_{(m-1)}$	$mp_{(m-1)}$	$\max\left(mp_{(1)},\ldots,2p_{(m-1)}\right)$	$\min\left(2p_{(m-1)},\frac{mp_{(m)}}{m}\right)$	$\min\left(\frac{mp_{(m-1)}}{m-1},\frac{mp_{(1)}}{m}\right)$
$p_{(m)}$	$mp_{(m)}$	$\max\left(mp_{(1)},\ldots,p_{(m)}\right)$	$\frac{mp_{(m)}}{m}$	$\frac{mp_{(m)}}{m}$

Eksempel, m=6 tester

Beregnet i Excel:

p_(j)	(j)	(m-j+1)p_(j)	mp_(j)/j	Bonferroni	Holm step-down	Hochberg step-up	Benjamini-Hochberg step-up
0,0003	1	0,0018	0,0018	0,0018	0,0018	0,0018	0,0018
0,009	2	0,0450	0,0270	0,0540	0,0450	0,0420	0,0210
0,013	3	0,0520	0,0260	0,0780	0,0520	0,0420	0,0210
0,014	4	0,0420	0,0210	0,0840	0,0520	0,0420	0,0210
0,04	5	0,0800	0,0480	0,2400	0,0800	0,0600	0,0480
0,06	6	0,0600	0,0600	0,3600	0,0800	0,0600	0,0600

Eksempel, m=6 tester.
Beregnet i R med p.adjust:



5.4. Summary

The section considered the class of multiplicity problems where no logical restrictions are imposed on the null hypotheses of interest. The four p-value-based MTPs used in this setting are easily arranged in terms of increasing power.

$$\text{Bonferroni} < \text{Holm} < \text{Hochberg} < \text{Hommel}.$$

Therefore, the Hochberg and Hommel procedures are preferred over the other two procedures in any multiplicity problem without hypothesis ordering. However, there is an important caveat that needs to be noted here. As explained in Section 5.3, the Bonferroni and Holm procedures protect the FWER strongly in any multiplicity problem, but the Hochberg and Hommel procedures require additional distributional assumptions to protect the FWER. The distributional assumptions are not restrictive and are met in many multiplicity problems arising in clinical trials.

Dmitrienko and d'Agostino 2013 page 5191

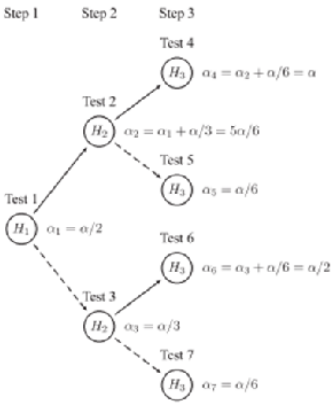
Fallback procedure in the analysis of multiple dose-placebo comparisons (Dmitrienko and D'Agostino 2013, page 5183-5184)

RCT with 3 dose-placebo comparisons:
H_1: Dose H vs placebo
H_2: Dose M vs placebo
H_3: Dose L vs placebo

The sponsor introduces unequal hypothesis weights to increase statistical power for the higher doses: $w_1 = 1/2$, $w_2 = 1/3$, $w_3 = 1/6$

The weights are (of course!) pre-specified and $\sum_j w_j = 1$.

Fallback procedure with α-propagation ("use it or lose it" principle): Test the hypotheses sequentially (in a pre-specified order) with initial local significance levels $\alpha/2$, $\alpha/3$, $\alpha/6$. See Fig 6 in the ref.



A solid (dashed) arrow shows the decision path after a hypotheses has been rejected (accepted).

«Loose it or use it» principle for α-propgagation.

Dmitrienko and D'Agostino (2013, Fig 6).

STATISTICS IN MEDICINE
Statist. Med. 2009; **28**:586–604
Published online 2 December 2008 in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/sim.3495

A graphical approach to sequentially rejective multiple test procedures

Frank Bretz[1,2], Willi Maurer[1], Werner Brannath[3] and Martin Posch[3,*,†]

[1]*Novartis Pharma AG, Lichtstrasse 35, 4002 Basel, Switzerland*
[2]*Department of Biometry, Medical University of Hannover, 30625 Hannover, Germany*
[3]*Section of Medical Statistics, Medical University of Vienna, Spitalgasse 23, 1090 Wien, Austria*

Example:
RCT with three treatment arms. Drug high dose, drug low dose, placebo.

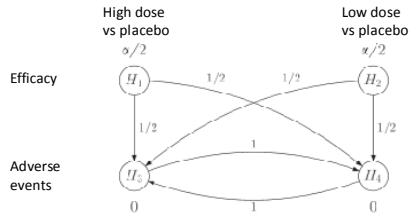


Figure 2. Graphical illustration of the gatekeeping procedure from [7] with four hypotheses

THE LANCET

Comprehensive geriatric care for patients with hip fractures: a prospective, randomised, controlled trial

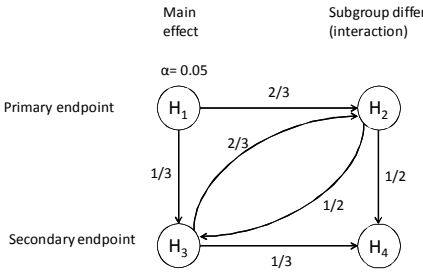
RCT:
Comprehensive geriatric care versus usual orthopaedic care

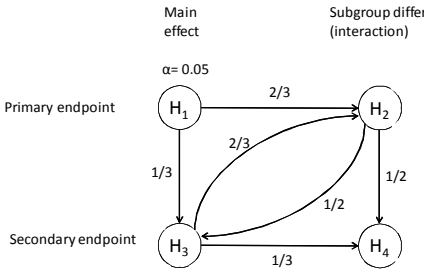
Primary outcome: Mobility (SPPB) 4 months after surgery
Secondary outcomes: Minimental State Examination (MMSE), Barthel Index, …

Subgroup analyses in a forthcoming publication

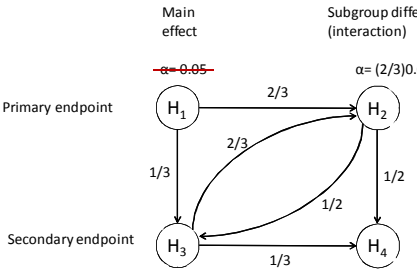
Outcomes and subgroup analyses were pre-specified in the statistical analysis plan.

A possible approach to preserve FWER:



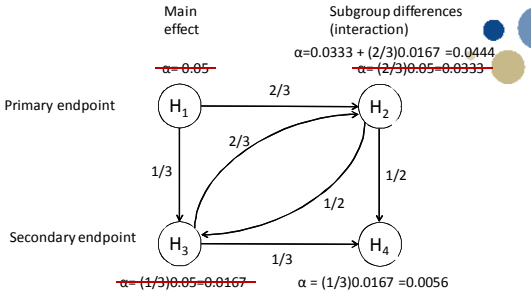


	Treatment group Main effect	Treatment * fracture type (interaction)
SPPB	P=0.010	P=0.07
Barthel Index	P=0.013	P=0.037



α= (1/3)0.05=0.0133

	Treatment group Main effect	Treatment * fracture type (interaction)
SPPB	P=0.010	P=0.07
Barthel Index	P=0.013	P=0.037



	Treatment group Main effect	Treatment * fracture type (interaction)
SPPB	P=0.010	P=0.07
Barthel Index	P=0.013	P=0.037

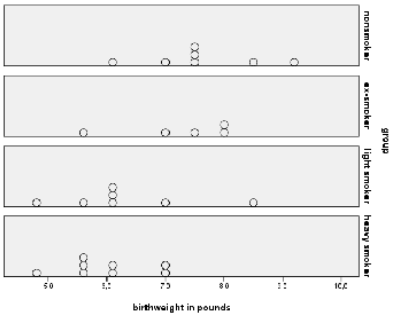
Tre eller flere grupper (behandlings- eller eksponeringsgrupper):
Parvise sammenlikninger mellom noen av eller alle parene.

- 4 eller flere grupper
- 3 grupper

38

Eksempel: Rosner (2006), oppg 12.11

Eksempel: Opp 12.14 i Rosner 6th ed, 2006



Descriptives

birthweight in pounds

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean	
					Lower Bound	Upper Bound
nonsmoker	7	7,586	,9516	,3638	6,695	8,475
ex-smoker	5	7,240	,9127	,4082	6,107	8,373
light smoker	7	6,329	1,1398	,4308	5,274	7,383
heavy smoker	8	6,013	,7200	,2546	5,411	6,614
Total	27	6,730	1,1090	,2134	6,291	7,169

ANOVA

birthweight in pounds

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	11,673	3	3,891	4,400	,014
Within Groups	20,304	23	,883		
Total	31,976	26			

Eksempel: 4 grupper.

Global hypotese:

$$\mu_1 = \mu_2 = \mu_3 = \mu_4$$

Totalt $\binom{4}{2} = \frac{4!}{(4-2)!2!} = 6$ parvise sammenlininger:

$\mu_1 = \mu_2$
$\mu_1 = \mu_3$
$\mu_1 = \mu_4$
$\mu_2 = \mu_3$
$\mu_2 = \mu_4$
$\mu_3 = \mu_4$

Bare teste mot en kontroll:

$\mu_1 = \mu_2$
$\mu_1 = \mu_3$
$\mu_1 = \mu_4$

Lineære kontraster :

$\sum_{i=1}^{k} c_i \mu_i = 0$ hvor $\sum_{i=1}^{k} c_i = 0$

For eksempel:

$\mu_1 - \mu_2 = 0$

$\mu_3 - \mu_4 = 0$

$\frac{1}{2}(\mu_1 + \mu_2) - \frac{1}{2}(\mu_3 + \mu_4) = 0$

Lineære kontraster er ortogonale hvis

$\sum_{j=1}^{k} c_{ij} c_{i'j} = 0$.

De tre lineære kontrastene i eksempelet ovenfor er ortogonale. Parvise kontraster som inneholder samme forventningsverdi, som for eksempel alle parvise kontraster eller teste mot en kontroll, er ikke ortogonale.

Metoder spesielt for ortogonale kontraster se for eksempel Kirk (2013).

De fleste prosedyrer for multiple sammenlikninger bruker testobservatorer (test statistics) av typen

$$t = \frac{\bar{Y}_{\bullet 1} - \bar{Y}_{\bullet 2}}{\sqrt{MS_{error}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

som vanligvis er fordelt som Student's t eller Studentized range.

Kirk (2013) «Experimental design. Procedures for the behavioural sciences» 4th ed, Chapter 5 «Multiple Comparison tests» (side 154-208). (Første utgave kom i 1969)

Fra omtalen på bokomslaget:
«Up-to-date coverage of multiple comparison procedures».

Table 5.1_1:
http://folk.ntnu.no/slyderse/medstat/Kirk_table5.1_1.pdf

Anbefalte prosedyrer ved p (≥4) grupper, enveis ANOVA:

Test hva?	Likt antall (balansert) eller lik varians i gruppene?	Anbefalt test
p-1 grupper mot en kontrollgruppe	Balansert og lik varians	Dunnett
	Ubalansert eller ulik varians	Dunnett med modifikasjoner
C planlagte kontraster	Balansert og lik varians	Dunn(-Bonferroni) Dunn- Šidák (Dunn-) Holm Dunn-Hochberg?
	Ulik varians	Som over med Welch frihetsgrader
Alle parvise kontraster	Balansert og lik varians	Tukey Fisher-Hayter REGW* F, FQ og Q
	Ubalansert	Tukey-Kramer Fisher-Hayter
	Ulik varians	Brown-Forsythe
Planlagte og uplanlagte kontraster	Balansert eller ubalansert	Scheffé
	Ulik varians	Dunnett's T3 Dunnett's C Games-Howell

I stor grad basert på Kirk (2013, Table 5.1-1).
Der flere alternativer er listet, er det i økende statistisk styrke
* REGW står for Ryan, Einot, Gabriel, Welsch

Sammenlikne grupper justert for minst en kovariat (for eksempel alder eller kjønn). (« ANCOVA»)

Ved 4 eller flere grupper, i økende statistisk styrke:
Bonferroni < Holm < Hochberg < Hommel
Šidák er såvidt sterkere enn Bonferroni

Ved 3 grupper:
Kombiner global test med lokal test ved parvise sammenlikninger

Closed testing procedure for m hypotheses $H_1, H_2, \ldots, H_m$:

Define the closed family of hypotheses. For each non-empty set of indices $I \subseteq \{1, 2, \ldots, m\}$, consider the intersection hypothesis $\bigcap_{i \in I} H_i$.

«Intersection» («snitt») betyr «(og)/eller»

Define local (that is, unadjusted) α-level tests for the intersection hypotheses.

Reject H_i at level α if and only if all intersection hypotheses containing H_i are rejected by local α-level tests.

(Marcus et al. 1976) showed that this closed testing procedure for the hypotheses $H_1, H_2, \ldots, H_m$ controls the FWER in the strong sense at the α-level.

Spesialtilfelle:
Parvise sammenlikninger av 3 parametre, feks forventningsverdi i 3 grupper:

$H_1 : \mu_2 = \mu_3$, $H_2 : \mu_1 = \mu_3$, $H_3 : \mu_1 = \mu_2$

Global hypotese:
$H_{Global} : \mu_1 = \mu_2 = \mu_3$

I følge «The closed testing procedure» forkaster vi H_1 på nivå α hvis vi forkaster alle disse på lokalt nivå α:

H_1
$H_1 \cap H_2$
$H_1 \cap H_3$
$H_1 \cap H_2 \cap H_3 \left(= H_{Global}\right)$

Men
$H_1 \cap H_2 = \left\{\left(\mu_2 = \mu_3\right) \cap \left(\mu_1 = \mu_3\right)\right\} = \left\{\mu_1 = \mu_2 = \mu_3\right\} = H_{Global}$
og
$H_1 \cap H_3 = \left\{\left(\mu_2 = \mu_3\right) \cap \left(\mu_1 = \mu_2\right)\right\} = \left\{\mu_1 = \mu_2 = \mu_3\right\} = H_{Global}$

så vi forkaster H_1 på nivå α hvis vi forkaster H_1 på lokalt nivå α og forkaster H_{Global} på nivå α.

Hvis vi forkaster $\mu_i = \mu_j$ bare når både lokal p-verdi og p-verdi for global test er under α så bevarer vi FWER! Dvs vi utfører testen for $\mu_i = \mu_j$ (ujustert) bare hvis vi forkaster H_{Global} på nivå α.

Ekvivalent:
p-verdien for den parvise testen settes lik den største av den lokale og globale *p*-verdien: $p_{j,adjusted} = \max(p_0, p_j)$.

Dette er lite kjent.
Bender and Lange (2001).

NB!
Dette gjelder bare for 3 grupper. Det gjelder ikke for eksempel for 4 grupper, selv når du bare har 3 hypoteser (som 3 alternativer mot en kontroll).

"In the frequent case of three groups the principle of closed testing leads to the following simple procedure that keeps the multiple level α. At first, test the global null hypothesis that all three groups are equal by a suitable level α test (e.g., and F test or the Kruskal–Wallis test). If the global null hypothesis is rejected proceed with level α tests for the three pairwise comparisons (e.g., t tests or Wilcoxon rank sum tests)."

(Bender and Lange 2001, Section 5.1)

Hence:

- In studies with more than one hypothesis, some adjustment is needed to control the probability of falsely rejecting at least one true hypothesis (Familywise error rate, FWER).
- If only three quantities are involved, such as mean outcome in three groups, you can reject the equality between two groups if the local test and the global test show statistical significance. This follows from the so-called principle of closed testing.
- Few researchers are aware that no additional adjustment is necessary to control FWER when this procedure is followed for three quantities.

Eksempel fra Weider et al (2014), Tabell 3:

Descriptives

Skåre på WAIS Informasjon- Skalert skåre

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean	
					Lower Bound	Upper Bound
AN	41	10,51	3,264	,510	9,48	11,54
BN	40	10,00	2,418	,382	9,23	10,77
Kontroll	40	11,35	2,833	,448	10,44	12,26
Total	121	10,76	2,916	,265	10,25	11,32

ANOVA

Skåre på WAIS Informasjon- Skalert skåre

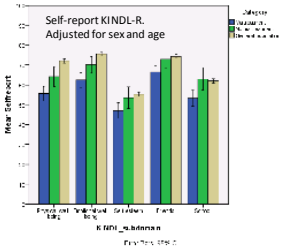
	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	73,009	2	36,505	4,457	,014
Within Groups	967,344	118	8,198		
Total	1040,413	120			

Global p = 0.0136

	p-verdi				
Par	Šidák	Tukey HSD	Dunnett	LSD (ujustert)	Max (ujustert, global)
AN vs BN	0.807	0.701	-	0.422	0.422
AN vs HC	0.109	0.094	0.069	0.038	0.038
BN vs HC	0.0137	0.013	0.009	0.005	0.0136

Eksempel fra Greger et al, submitted 2015:
Child Maltreatment and Quality of Life: A Study of Adolescents in Residential Care

Residencial youth care institutions:
115 adolescents with history of maltreatment
55 adolescents without history of maltreatment
General population:
1017 adolescents



KINDL-R Self-report, school
Linear model, adjusted for sex and age

Global p = 0.067

	p-value			
Pair	Bonferroni	Šidák	LSD (unadjusted)	Max (unadjusted, global)
Maltrt vs no maltrt	0.062	0.060	0.021	0.067
Maltrt vs normpop	1.000	0.735	0.358	0.358
No maltrt vs normpop	0.214	0.199	0.071	0.067

Testing, P-verdier og konfidensintervall:

- Metodene tar utgangspunkt i testing (aksepter eller forkast ved gitt signifikansnivå)
- Forholdsvis rett frem å beregne p-verdier (minste signifikansnivå som ville medført forkastning)
- Konfidensintervall kan beregnes for noen metoder (som Bonferroni, Šidák, Dunnett, Tukey, Scheffé), men ikke for flerstegs metoder (Som Holm step-down, Hochberg step-up, Benjamini-Hochberg step-up).
- "In practice, if a stepwise multiple testing procedures is applied to deal with multiplicity in a clinical trial, the trial's sponsor typically has to resort to presenting unadjusted/marginal CIs for the treatment parameters with the understanding that the joint coverage probability of these intervals is not controlled." (Dmitrienko & D'Agostino 2013, page 5205)

60

Control FWER or FDR?

- Control of FDR instead of FWER results in higher statistical power at the cost of increased type I error rate.
- Traditionally, FDR is used in studies with large numbers of tests
- Glickman & al (2014) advocate use of FDR also in studies with small to moderate numbers of simultaneous tests

61

"Most applications of false discovery rate control have been in situations where tens of thousands of tests (or more) are performed, but the procedures work reliably in smaller numbers of tests. Simulation analyses (Benjamini and Hochberg, 1995), Verhoeven & al, 2005) and (Williams & al, 1999) have indicated that false discovery rate control has uniformly better power than other competitor methods (including FWER control), and the fraction of false positives is about what would be expected, even in small to moderate numbers of simultaneous tests. Thus, false discovery rate control has application in smaller studies, though the advantages are more pronounced with larger numbers of tests."

(Glickman & al, 2014)

How small studies?

Williams, & al (1999) "Controlling error in multiple comparisons, with examples from state-to-state differences in educational achievement. " Journal of Educational and Behavioral Statistics, 24, (1) 42-69:
Examples with 45 or more comparisons

Verhoeven & al (2005). Implementing false discovery rate control: increasing your power. Oikos, 108, (3) 643-647
Example with 50 tests.

Benjamini, Y. & Hochberg, Y. (1995). Controlling the False Discovery Rate - A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society Series B-Methodological.
See next slide …

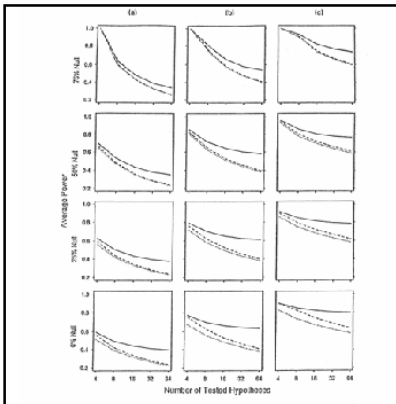


Fig. 1. Simulation-based estimates of the average power (the proportion of the false null hypotheses which are correctly rejected) for two FWER controlling methods, the Bonferroni (·····) and Hochberg's (1988) (——) methods, and the FDR controlling procedure (——): (a) decreasing; (b) equally spaced; (c) increasing

FWER:
Hochberg step-up is more powerful than Bonferroni especially with few (0-25%) true null hypotheses

FDR versus FWER:
FDR (Benjamini-Hochberg) is substantially more powerful with many (say, at least 8 or 16) hypotheses.

Figure 1, Benjamini & Hochberg (1995)

Control FDR or FWER?

- With many hypotheses, control of FDR can be sensible.
- But with few (less than 8 or 16?) hypotheses, the power gain of Benjamini-Hochberg FDR is not large compared to Hochberg step-up
- With few hypotheses, under certain conditions (3 groups or oneway ANOVA) there even exist methods to control FWER, which are more simpler and/or more powerful than Hochberg step-up (see next slide).

65

Choice of method for control of FWER:
Simple advice for comparison between groups:

- 3 groups: Combine a global test with (unadjusted) local tests
- ≥ 4 groups, oneway ANOVA without adjusting for covariates (see also table based on Kirk 2013):
  - Dunnett (only versus control group)
  - Tukey (all pairwise comparisons, when balanced and equal variances)
  - Scheffé (all planned or unplanned contrasts, also unbalanced)
- ≥ 4 groups else:
  - Bonferroni < Šidák < Holm < Hochberg step-up < Hommel

66

References

Bender, R. & Lange, S. 2001. Adjusting for multiple testing--when and how? J.Clin.Epidemiol., 54, (4) 343-349

Benjamini, Y. & Hochberg, Y. 1995. Controlling the False Discovery Rate - A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society Series B-Methodological*, 57, (1) 289-300

Bretz, F., Maurer, W., Brannath, W., & Posch, M. 2009. A graphical approach to sequentially rejective multiple test procedures. *Stat.Med.*, 28, (4) 586-604

Dmitrienko, A., D'Agostino, R.B., & Huque, M.F. 2013. Key multiplicity issues in clinical drug development. Statistics in Medicine, 32, (7) 1079-1111

Dmitrienko, A. & D'Agostino, R. 2013. Tutorial in Biostatistics: Traditional multiplicity adjustment methods in clinical trials. Statistics in Medicine, 32, 5172-5218

Dmitrienko, A., Tamhane, A.C., & Bretz, F. 2010. Multiple testing problems in pharmaceutical statistics Boca Raton, FL, Chapman & Hall/CRC.

Glickman, M.E., Rao, S.R., & Schultz, M.R. 2014. False discovery rate control is a recommended alternative to Bonferroni-type adjustments in health studies. *Journal of Clinical Epidemiology*, 67, (8) 850-857

References (continued)

Goeman, J.J. & Solari, A. 2014. Tutorial in biostatistics: Multiple hypothesis testing in genomics. Statistics in Medicine, 33, (11) 1946-1978

Greger, H, K., Myhre, A. K., Lydersen, S., & Jozefiak, T.: Child Maltreatment and Quality of Life: a Study of Adolescents in Residential Care. Submitted, 2015.

Kaasboll, J., Lydersen, S., & Indredavik, M.S. 2012. Psychological symptoms in children of parents with chronic pain-the HUNT study. Pain, 153, (5) 1054-1062

Kirk, R.E. 2013. Experimental design. Procedures for the behavioral sciences, 4th ed. Thousand Oaks, Sage Publications.

Marcus, R., Peritz, E., & Gabriel, K.R. 1976. Closed Testing Procedures with Special Reference to Ordered Analysis of Variance. Biometrika, 63, (3) 655-660

Prestmo, A., Hagen, G., Sletvold, O., Helbostad, J.L., Thingstad, P., Taraldsen, K., Lydersen, S., Halsteinli, V., Saltnes, T., Lamb, S.E., Johnsen, L.G., & Saltvedt, I. 2015. Comprehensive geriatric care for patients with hip fractures: a prospective, randomised, controlled trial. *Lancet*, 385, (9978) 1623-1633

References (continued)

Rosner, B. 2006. Fundamentals of biostatistics, 6th ed ed. Belmont, CA, Thomson-Brooks/Cole.

Rothman, K.J. 1990. No adjustments are needed for multiple comparisons. Epidemiology, 1, (1) 43-46

Rothman, K.J. 2014. Six persistent research misconceptions. J.Gen.Intern.Med., 29, (7) 1060-1064

Senn, S. 2007. Statistical issues in drug development, 2nd ed. Chichester, England, John Wiley & Sons.

Shaffer, J.P. 1995. Multiple Hypothesis Testing. Annual Reviews Psychology, 46, 561-584

Weider, S., Indredavik, M.S., Lydersen, S., & Hestad, K. 2014. Intellectual Function in Patients with Anorexia Nervosa and Bulimia Nervosa. Eur.Eat.Disord.Rev., 22, (1) 15-22

Wright, S.P. 1992. Adjusted P-Values for Simultaneous Inference. *Biometrics*, 48, (4) 1005-1013