

Multiple Testen

– mit medizinischen Anwendungen

F. Bretz*, L.A. Hothorn
Universität Hannover

Heidelberg, 22.11.2000

AG Statistische Methodik in der klinischen Forschung

*email: bretz@ifgb.uni-hannover.de

Outline

- Motivation, Beispiele
- Bonferroni–, Holm–Methode
- Ausnutzung logischer und stochastischer Abhängigkeiten
- Multiple Kontrasttests (Tukey, Dunnett, ...)
- Abschlußtest
- *A-priori* geordnete Hypothesen
- Implementierungen in SAS
- Zusammenfassung
- Literatur

MOTIVATION

Erster Zugang

Multiplizitätsprobleme

treten dann auf, wenn

*mehrere (simultane) Tests
am gleichen Datenmaterial (Studie)*

durchgeführt werden.

Anwendungen in klinischen Studien

- Mehrere Behandlungs-/Dosisgruppen
- Mehrere Endpunkte
- Subgruppenanalyse
- Zwischenauswertungen
- ...

Unterschied zu multivariaten Verfahren

Multiple Verfahren (Tukey-Test, Dunnett-Test, ...):

simultan, lokal

Multivariate Verfahren (F -Test, T^2 -Test, ...):

global, zusammenfassend

Situation

System von Hypothesen

$$\{H_i, i = 1, \dots, n\}$$

mit n Entscheidungsfunktionen

$$\{\varphi_i(x), i = 1, \dots, n\}$$

wobei

$$\varphi_i(x) = \begin{cases} 1, & H_i \text{ ablehnen} \\ 0, & H_i \text{ nicht ablehnen} \end{cases}$$

Globalhypothese:

$$H = \bigcap_{i=1}^n H_i$$

Beispiel

All-Paar-Vergleich von $k = 4$ Behandlungen

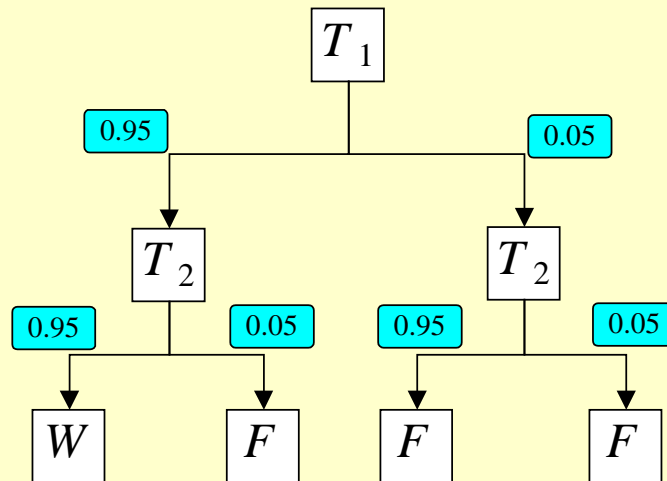
System von Einzelhypothesen ($n = 6$):

- $H_1: \mu_1 = \mu_2,$
- $H_2: \mu_1 = \mu_3,$
- $H_3: \mu_1 = \mu_4,$
- $H_4: \mu_2 = \mu_3,$
- $H_5: \mu_2 = \mu_4,$
- $H_6: \mu_3 = \mu_4.$

Globalhypothese: $H: \mu_1 = \mu_2 = \mu_3 = \mu_4.$

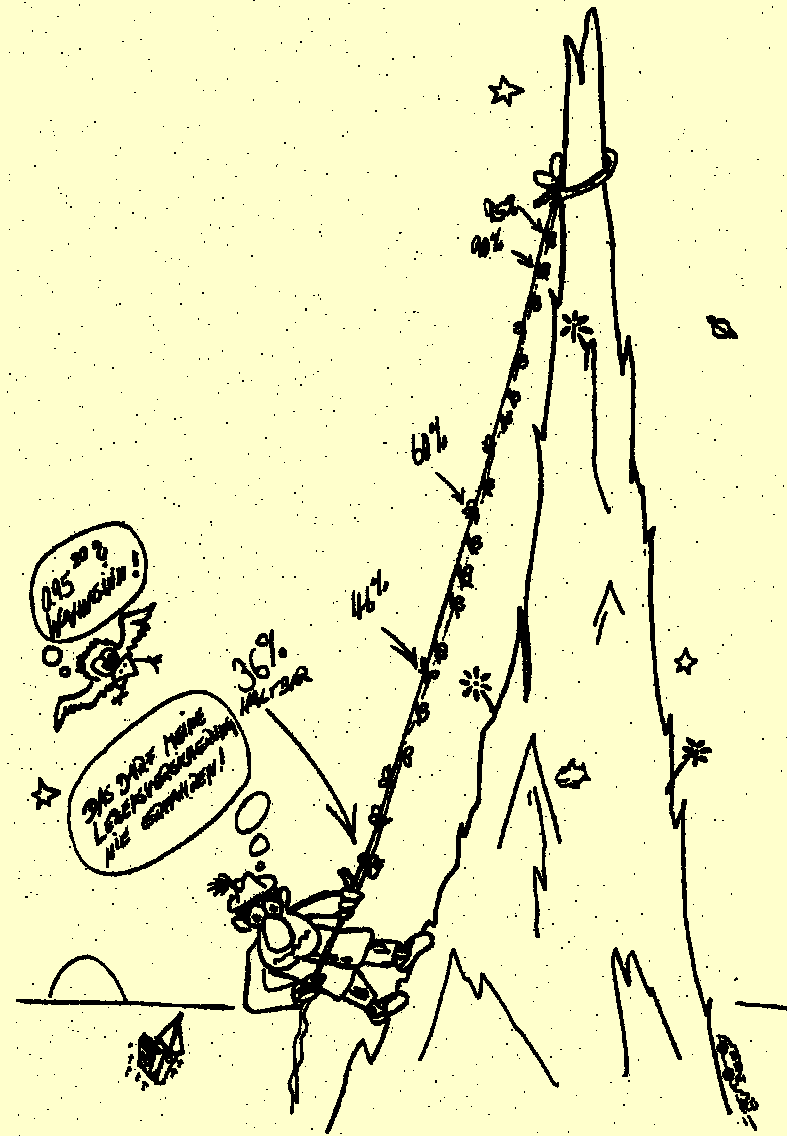
Problem der Multiplizität

Kombination der Aussagen zweier Tests zu je 5%:



Damit ist

$$\begin{aligned} P(F) &= P(\text{mindestens eine } H_i \text{ fälschlicherweise abzulehnen}) \\ &= 2 * 0.95 * 0.05 + 0.05 * 0.05 = 0.0975 > 0.05 \end{aligned}$$



(aus Beck-Bornholdt und Dubben, 1999)

Notation

$\alpha_{\text{versuchsbezogen}} = P(\text{mindestens eine } H_i \text{ fälschlicherweise abzulehnen}), \text{ d.h. der Fehler 1. Art einer Gesamtaussage}$

$\alpha_{\text{lokal}} = P(\text{falsche Ablehnung einer einzelnen } H_i),$
d.h. der *Fehler 1. Art eines einzelnen Tests*

Motivation

Zentrales Problem der multiplen Testtheorie:
Die Kontrolle des **versuchsbezogenen** Fehlers 1. Art

Beispiel (1)

Vergleich mehrerer Behandlungen

Toxikologische Studie bei Jungmäusen zum Vergleich von 6 chemischen Substanzen (Endpunkt: Gewichtsveränderung zum Baselinewert).

Frage: Welche der 6 Substanzen unterscheiden sich im Vergleich zur Kontrolle (bzw. untereinander)?

	Control	1	2	3	4	5	6
MEAN	105.38	95.9	80.48	72.14	91.88	84.68	74.24
STD	13.443	23.8904	12.6830	8.4111	9.4397	18.3549	7.8065
N	4	4	4	4	4	4	4

Beispiel (2)

Vgl. zweier Dosierungen vs. Placebo für mehrerer Endpunkte

Dosis	Endpunkte									
	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10
0	0	0	1	0	0	1	0	0	0	0
0	0	0	0	0	0	0	0	0	0	1
0	0	0	0	0	0	0	0	0	1	0
0	0	0	0	0	0	0	0	0	0	0
0	0	1	0	0	0	0	0	0	0	0
1	1	0	0	1	0	1	0	0	1	0
1	0	0	0	1	1	0	0	1	0	1
1	0	1	0	0	0	0	1	0	0	0
1	0	0	1	0	0	0	0	0	0	1
1	1	0	1	0	0	0	0	1	0	0
2	0	1	1	1	0	1	1	1	0	1
2	1	1	1	1	1	1	0	1	1	0
2	1	0	0	1	0	1	1	0	1	0
2	0	1	1	1	1	0	1	1	1	1
2	1	0	1	0	1	1	1	0	0	1

LÖSUNGSANSÄTZE

Bonferroni–Methode (1)

Hypothesen $\{H_i, i = 1, \dots, n\}$

p –Werte $\{p_i, i = 1, \dots, n\}$

→ adjustierte p –Werte: $\tilde{p}_i = np_i < \alpha_{\text{versuchsbezogen}}$

→ äquivalent hierzu: $p_i < \alpha_{\text{lokal}} / n$

Bonferroni–Methode (2)

Vorteile:

- einfach
- generell
- korrekt

Nachteile:

- konservativ

Bonferroni–Holm–Methode (1)

Geordnete p -Werte $p_{(1)} \leq \dots \leq p_{(n)}$

→ adjustierte p -Werte: $\tilde{p}_{(i)} = (n - i + 1)p_{(i)} < \alpha_{\text{versuchbezogen}}$

→ äquivalent hierzu: $p_{(i)} < \alpha_{\text{lokal}} / (n - i + 1)$

Schrittweise Prozedur (setze $i = 1$ am Anfang):

IF $\tilde{p}_{(i)} < \alpha$ **THEN** $i = i + 1$

ELSE STOP

$$p_{(1)} \stackrel{?}{<} \alpha/n \xrightarrow{\text{Abl.}} p_{(2)} \stackrel{?}{<} \alpha/(n-1) \xrightarrow{\text{Abl.}} \dots \xrightarrow{\text{Abl.}} p_{(n)} \stackrel{?}{<} \alpha$$

Bonferroni–Holm–Methode (2)

Beispiele (hier: $\alpha = \alpha_{\text{lokal}}$)

Holm:	n			Schranken		
	2	$\alpha/2$	α			
	3	$\alpha/3$	$\alpha/2$	α		
	4	$\alpha/4$	$\alpha/3$	$\alpha/2$	α	
	5	$\alpha/5$	$\alpha/4$	$\alpha/3$	$\alpha/2$	α

Bonferroni: jeweils zu α/n

Bonferroni–Holm–Methode (3)

Vorteile:

- einfach
- generell
- korrekt
- gleichmäßig besser als Bonferroni

Nachteile:

- logische und stochastische Abhängigkeiten werden nicht genutzt.

Ausnutzung logischer Abhängigkeiten

- Shaffer (1986)

Wichtiger Spezialfall: alle Paarvergleiche

$k = 3$ Behandlungen: $\alpha/3, \alpha, \alpha$

$$\left. \begin{array}{l} H_1: \mu_1 = \mu_2 \\ H_2: \mu_2 = \mu_3 \end{array} \right\} \Rightarrow H_3: \mu_1 = \mu_3$$

$k = 4$: $\alpha/6, \alpha/3, \alpha/3, \alpha/3, \alpha/2, \alpha$

$k = 5$: $\alpha/10, \alpha/6, \alpha/6, \alpha/6, \alpha/6, \alpha/4, \alpha/4, \alpha/3, \alpha/2, \alpha$

- Hommel und Bernhard (1999)
- Westfall (1997)

Ausnutzung stochastischer Abhängigkeiten

Abhängigkeiten (d.h. Korrelationen) zwischen den Teststatistiken führen zu simultanen Prüfverteilungen.

Zur Vorbereitung:

Klassisches Einweg-ANOVA-Modell

$$X_{jl} = \mu_j + \varepsilon_{jl}$$

$j = 1, \dots, k$ (Behandlungen)

$l = 1, \dots, m_j$ (Wiederholungen)

$$\varepsilon_{jl} \sim N(0, \sigma^2)$$

Kontrasttests

Definition

Einzelkontrast:

$$T = \frac{\sum_{j=1}^k c_j \bar{X}_j}{s \sqrt{\sum_{j=1}^k c_j^2 / m_j}} \quad \text{mit} \quad \sum_{j=1}^k c_j = 0.$$

Multipler Kontrasttest: $T^{MC} = \max\{T_1, \dots, T_q\}$

➔ Gemeinsame Verteilung von T_1, \dots, T_q ist multivariat t und numerisch verfügbar (Genz und Bretz, 1999).

Dunnett-Test

Vergleich mehrerer Behandlungen gegen eine Kontrolle

(many-to-one Situation)

Kontrastmatrix ($k = 4, q = 3$):

$$\begin{pmatrix} -1 & 1 & 0 & 0 \\ -1 & 0 & 1 & 0 \\ -1 & 0 & 0 & 1 \end{pmatrix}$$

Lokale Entscheidungen und Konfidenzintervalle verfügbar.

Tukey-Test

Vergleich mehrer Behandlungen gegeneinander

(All-Paar Vergleiche)

Kontrastmatrix $\left(k = 4, q = \binom{k}{2} = 6 \right)$:

$$\begin{pmatrix} -1 & 1 & 0 & 0 \\ -1 & 0 & 1 & 0 \\ -1 & 0 & 0 & 1 \\ 0 & -1 & 1 & 0 \\ 0 & -1 & 0 & 1 \\ 0 & 0 & -1 & 1 \end{pmatrix}$$

Abschlusstest (1)

Formale Definition

Sei $I = \{1, \dots, n\}$,

$\mathcal{H} = \{H_i: i \in I\}$ System der Elementarhypothesen,

$\bar{\mathcal{H}} = \{H_J: \emptyset \neq J \subseteq I\}$ dessen Abschluss mit $H_J = \bigcap_{j \in J} H_j$ und
zugehörigem Niveau- α -Test φ_J .

Dann (Marcus *et al.*, 1976) bildet

$$\Psi_{\mathcal{H}} = \{\Psi_J: \emptyset \neq J \subseteq I\}$$

multiplen Niveau- α -Test für $\bar{\mathcal{H}}$, wobei

$$\Psi_J = \min_{J \subseteq M \subseteq I} \varphi_M$$

Insbesondere bilden die Elemente $\Psi_{\{i\}}$ gerade einen multiplen
Niveau- α -Test für \mathcal{H} .

Abschlusstest (2)

Durchführung

- Gegeben seien die n Elementarhypothesen H_i von Interesse.
- Bilde alle möglichen Durchschnitte dieser Hypothesen (→ durchschnittsabgeschlossenes System mit $m > n$ Hypothesen.)
- Finde für jede der Hypothesen einen Niveau- α -Test (Test zum lokalen Niveau α_{lokal})
- H_i wird nur dann abgelehnt, wenn alle Schnitt-hypothesen, die H_i implizieren, auch abgelehnt werden (Kohärenz).

Abschlusstest (3)

Vorteile:

- generell
- hohe Güte
- logische Abhängigkeiten werden genutzt
- gleichmäßig besser als Einschrittverfahren
- Güteberechnungen möglich

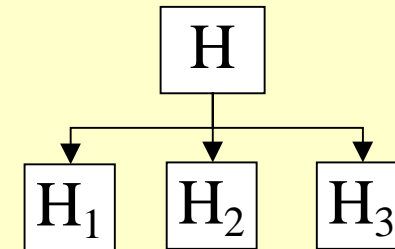
Nachteile:

- häufig keine Konfidenzintervalle
- u.U. sehr umfangreiches Hypothesensystem
- Aussagen zu Fehlern dritter Art?

Allpaarvergleiche, $k = 3$

$$n = \binom{3}{2} = 3 \Rightarrow H_1: \mu_1 = \mu_2, H_2: \mu_1 = \mu_3, H_3: \mu_2 = \mu_3,$$

$$H = \bigcap_{i=1}^3 H_i: \mu_1 = \mu_2 = \mu_3$$



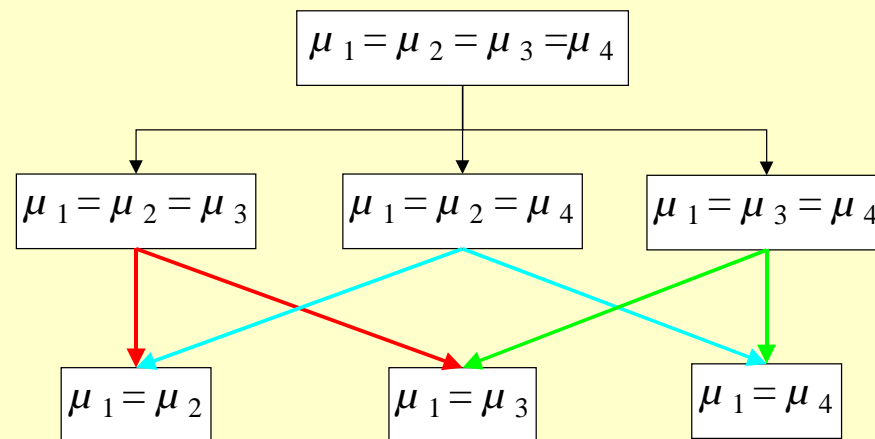
Mögliche Tests für H :

a) F -Test, b) Spannweiten-Test, c) Bonferroni-Globaltest

Strategie: Wird H nicht abgelehnt, stoppt das Verfahren, keine Einzelhypothese wird abgelehnt. Sonst erfolgt das Testen von H_1 , H_2 und H_3 zum Niveau α_{lokal} .

a) führt zu Fisher's LSD, c) zur Shaffer-Prozedur $\alpha/3$, α , α .
Prozeduren b) und c) sind konsonant.

Many-to-one Vergleiche, $k = 4$



→ Dunnett-Tests auf jeder Stufe (= t -Test für Elementarhyp.);
im Falle von Bonferroni-Einzeltests entsteht Holm-Prozedur
 $\alpha/3, \alpha/2, \alpha$

Beachte: Alle Durchschnitte der Elementarhypothesen sind
voneinander verschieden (→ keine logischen Abhängigkeiten;
“free combinations”)

A-priori geordnete Hypothesen (1)

Reihenfolge der $H_{(1)}, \dots, H_{(n)}$ fest inhaltlich vorgegeben.

Schrittweise Prozedur (setze $i = 1$ am Anfang):

IF $p_{(i)} < \alpha$ **THEN** $i = i + 1$

ELSE STOP

Vorteile:

- einfach
- generell
- *alle* Tests zum Niveau α_{lokal}
- kein aufgeblähtes Hyp.-system wie beim AT

Nachteile:

- vorgegebene Reihenfolge

***A-priori* geordnete Hypothesen (2)**

Anwendung: Multiple Endpunkte

Placebo-kontrollierte Studie mit *Gingko biloba* Extrakt mit multiplen Endpunkten **unterschiedlicher Priorität** (Kieser *et al.*, 1995). Endpunkte bzw. Testhierarchie:

- (1) Unterschied der schmerzfreien Wegstrecke bei Therapieende (24 Wochen).
- (2) Falls (1) gezeigt, dann die Frage: Effekt schon nach 16 Wochen sichtbar?
- (3) u.s.w., d.h. weitere Zeitpunkte o.ä.

A-priori geordnete Hypothesen (3)

Anwendung: Dosis-Wirkungsanalyse

$k = 4$: Vergleich einer Placebogruppe mit drei Dosierungen

$$H_{(1)}: \mu_1 = \mu_2 = \mu_3 = \mu_4$$

$$H_{(2)}: \mu_1 = \mu_2 = \mu_3$$

$$H_{(3)}: \mu_1 = \mu_2$$

- ➔ S. Hsu und Berger (1999) für schrittweise Konfidenzintervalle paarweiser Kontraste
- ➔ Für weitere Anwendungen und Verallgemeinerungen s. Maurer *et al.* (1995).

IMPLEMENTIERUNGEN IN SAS

Multiple Vergleiche mit PROC GLM (1)

```
data waste;
  do temp = 1 to 3;
    do envir = 1 to 5;
      do rep=1 to 2;
        input waste @@;
        output;
      end;
    end;
  end;
datalines;
7.09 5.90 7.94 9.15 9.23 9.85 5.43 7.73 9.43 6.90
7.01 5.82 6.18 7.19 7.86 6.33 8.49 8.67 9.62 9.07
7.78 7.73 10.39 8.78 9.27 8.90 12.17 10.95 13.07 9.76
;
run;

proc glm data=waste;
  class envir temp;
  model waste = envir temp envir*temp;
  lsmeans temp envir/pdiff cl adjust=tukey;
run;
```

Multiple Vergleiche mit PROC GLM (2)

```
data detergent;
  do detergent=1 to 5;
    do block =1 to 10;
      input plates @@;
      output;
    end;
  end;
datalines;
27 28 30 31 29 30 . . . .
26 26 29 . . . 30 21 26 .
30 . . 34 32 . 34 31 . 33
. 29 . 33 . 34 31 . 33 31
. . 26 . 24 25 . 23 24 26
;

proc glm data=detergent;
  class block detergent;
  model plates = block detergent;
  lsmeans detergent/pdiff cl adjust=simulate (acc=.001
report seed=121211);
run;
```

Multiple Vergleiche mit PROC MIXED (1)

```
proc mixed data=waste;  
  class envir temp;  
  model waste = temp/ddfm=satterth;  
  random envir envir*temp;  
  lsmeans temp/cl adjust=tukey;  
  ods output diffs=diffs;  
run;
```

```
proc print data=diffs noobs;  
  title "Multiple Comparisons in Random Block Model with  
  Interaction";  
  var temp _temp Estimate StdErr df AdjLow AdjUpp;  
run;
```

Multiple Vergleiche mit PROC MIXED (2)

```
proc mixed data=detergent;  
  class block detergent;  
  model plates = detergent/ddfm=satterth;  
  random block;  
  lsmeans detergent/cl adjust=simulate(seed=121211);  
  ods output diffs=diffs;  
run;  
  
proc print data=diffs noobs;  
  title "Multiple Comparisons in Random Block Model -  
        Incomplete Blocks";  
  var detergent _detergent Estimate StdErr df AdjLow AdjUpp;  
run;
```

Multiple Vergleiche mit PROC MIXED (3)

```
data Halothane;
  do Dog =1 to 19;
    do Treatment = 'HA', 'LA', 'HP', 'LP';
      input Rate @@;      output;
    end;
  end;
datalines;
426 609 556 600      253 236 392 395      359 433 349 357
432 431 522 600      405 426 513 513      324 438 507 539
310 312 410 456      326 326 350 504      375 447 547 548
286 286 403 422      349 382 473 497      429 410 488 547
348 377 447 514      412 473 472 446      347 326 455 468
434 458 637 524      364 367 432 469      420 395 508 531
397 556 645 625
;

proc mixed data=Halothane;
  class Dog Treatment;
  model Rate = Treatment / ddfm=satterth;
  repeated / type=un subject=Dog;
  lsmeans Treatment / adjust=simulate(nsamp=200000 seed=121211) cl pdiff;
run;
```

Multiple Endpunkte m. PROC MULTTEST

```
data a;  input group$ s1-s10;
datalines;
0MG 0 0 1 0 0 1 0 0 0 0
0MG 0 0 0 0 0 0 0 0 0 0 1
0MG 0 0 0 0 0 0 0 0 0 1 0
0MG 0 0 0 0 0 0 0 0 0 0 0
0MG 0 1 0 0 0 0 0 0 0 0 0
1MG 1 0 0 1 0 1 0 0 1 0 0
1MG 0 0 0 1 1 0 0 0 1 0 1
1MG 0 1 0 0 0 0 1 0 0 0 0
1MG 0 0 1 0 0 0 0 0 0 0 1
1MG 1 0 1 0 0 0 0 0 1 0 0
2MG 0 1 1 1 0 1 1 1 1 0 1
2MG 1 1 1 1 1 1 0 1 1 1 0
2MG 1 0 0 1 0 1 1 0 1 1 0
2MG 0 1 1 1 1 0 1 1 1 1 1
2MG 1 0 1 0 1 1 1 0 0 0 1
;
proc multtest bootstrap nsample=1000 seed=41287 order=data notables pvals;
  class group;
  test ca(s1-s10);
  contrast 'Trend' 0 1 2;
run;
```

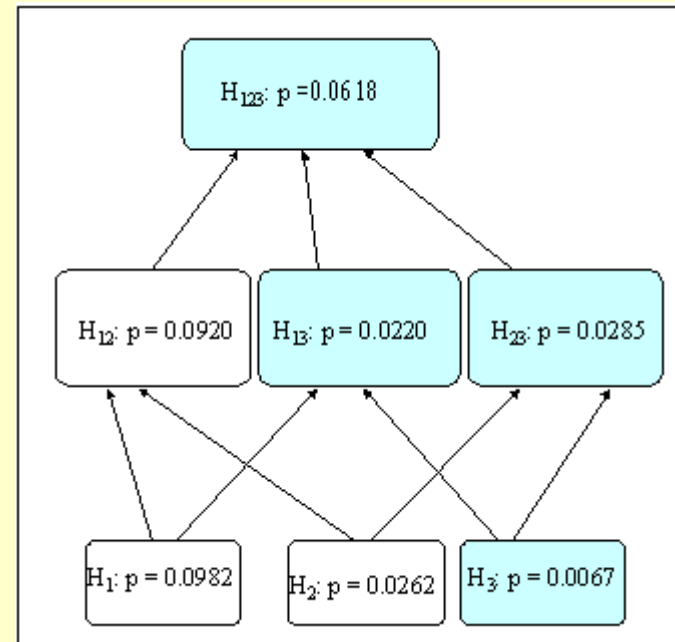

Abschlußtest in SAS (1)

```
data mult;  
    input G Y1 Y2 Y3;  
    datalines;  
0 14.4 7.00 4.30  
0 14.6 7.09 3.88  
0 13.8 7.06 5.34  
0 10.1 4.26 4.26  
0 11.1 5.49 4.52  
0 12.4 6.13 5.69  
0 12.7 6.69 4.45  
1 11.8 5.44 3.94  
1 18.3 1.28 0.67  
1 18.0 1.50 0.67  
1 20.8 1.51 0.72  
1 18.3 1.14 0.67  
1 14.8 2.74 0.67  
1 13.8 7.08 3.43  
1 11.5 6.37 5.64  
1 10.9 6.26 3.47  
;
```

Abschlußtest in SAS (2)

Hotelling's T^2 :

```
proc reg data=mult;  
  model Y1 Y2 Y3 = G;  
  H1: mtest Y1;  
  H2: mtest Y2;  
  H3: mtest Y3;  
  H12: mtest Y1, Y2;  
  H13: mtest Y1, Y3;  
  H23: mtest Y2, Y3;  
  H123: mtest Y1, Y2, Y3;  
run;
```



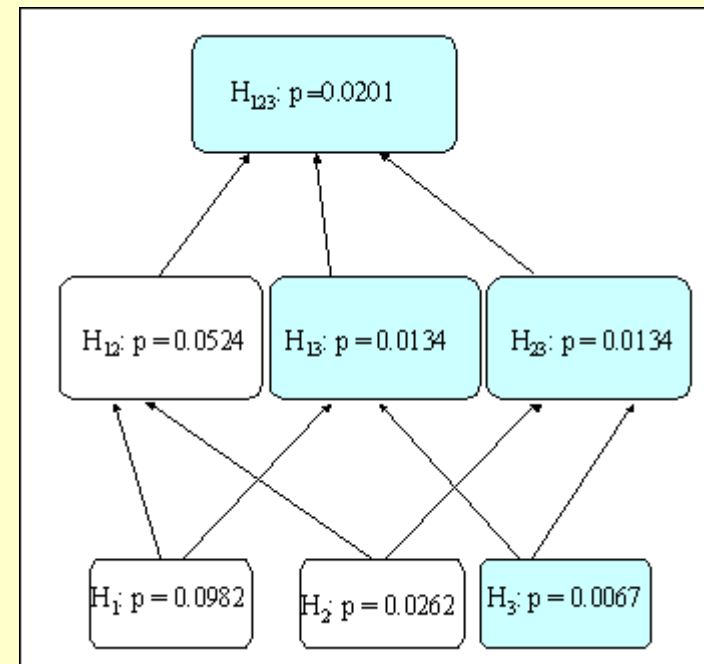
Abschlußtest in SAS (3)

Bonferroni-Holm:

```
proc multtest data=mult holm pvals;  
  class g;  
  test mean(Y1 Y2 Y3);  
  contrast "0 vs 1" -1 1;  
run;
```

```
data pvals;  
  input test$ raw_p @@;  
  datalines;  
  Y1 .0982 Y2 .0262 Y3 .0067  
  ;
```

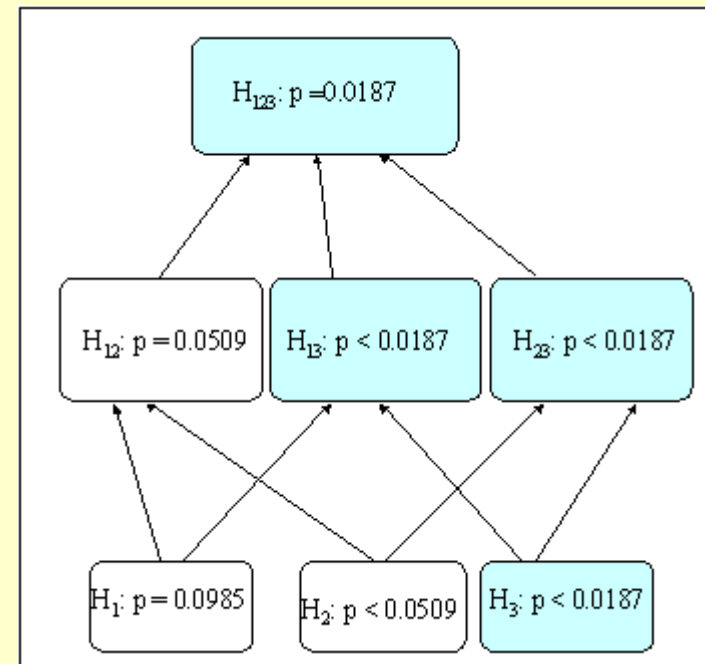
```
proc multtest pdata=pvals holm out=results; run;  
proc print data=results; run;
```



Abschlußtest in SAS (4)

Bootstrap-Holm minP:

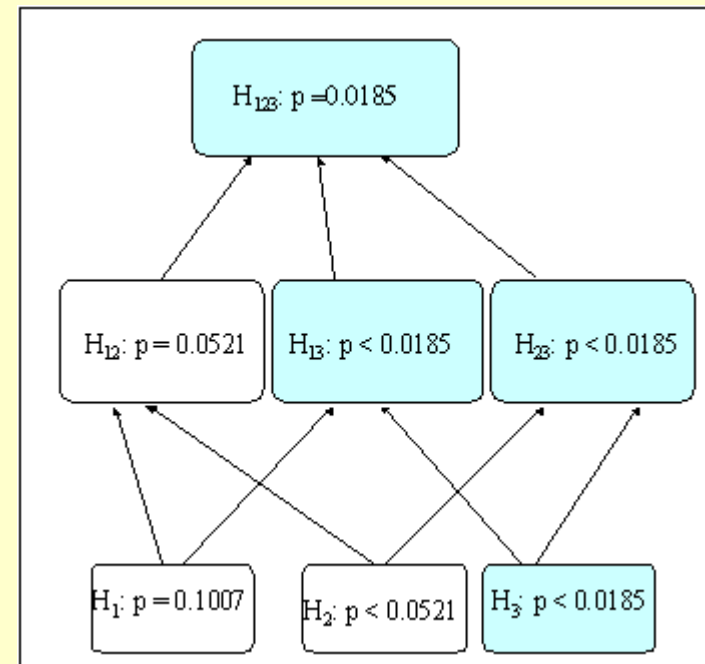
```
proc multtest data=mult holm stepboot pvals n=1000000;  
  class g;  
  test mean(Y1 Y2 Y3);  
  contrast "0 vs 1" -1 1;  
run;
```



Abschlußtest in SAS (5)

Permutation-Holm minP:

```
proc multtest data=mult holm stepperm pvals n=1000000;  
  class g;  
  test mean(Y1 Y2 Y3);  
  contrast "0 vs 1" -1 1;  
run;
```



Ausblick

- Motivation multiplen Testens
- Einführung verschiedener Verfahren und Strategien mit unterschiedlichen Vor- und Nachteilen
- Anwendungen in der Interims- und Subgruppenanalyse, ...
- Weitere multiple Testprozeduren (Simes, Hommel, ...)
- Verallgemeinerungen auf verteilungsfreie Verfahren, andere Datensituationen, ...
- Problematik der Konfidenzintervalle und Fehler 3. Art?

Literatur (1)

- **Miller, R.G.** (1981) *Simultaneous statistical inference*. Springer, New York.
- **Hochberg, Y. und Tamhane, A.C.** (1987) *Multiple comparison procedures*. Wiley, New York.
- **Bauer, P., Hommel, G. and Sonnemann, E.** (1988) *Multiple hypotheses testing*. Springer, Berlin.
- **Westfall, P.H. und Young, S.S.** (1993) *Resampling-based multiple testing*. Wiley, New York.
- **Horn, M. und Vollandt, R.** (1995) *Multiple Tests und Auswahlverfahren*. Fischer, Stuttgart.
- **Hsu, J.C.** (1996) *Multiple comparisons*. Chapman and Hall, London.
- **Westfall, P.H., Tobias, R.D., Rom, D., Wolfinger, R.D. und Hochberg, Y.** (1999) *Multiple comparisons and multiple tests using the SAS system*. Cary, NC.
- **Westfall, P.H. and Tobias, R.D.** (2000) *Multiple comparisons and multiple tests using the SAS system workbook*. Cary, NC.

Literatur (2)

- **Dunnett, C.W.** (1955) A multiple comparison procedure for comparing several treatments with a control. *Journal of the American Statistical Association*, **50**, 1096-1121.
- **Genz, A. and Bretz, F.** (1999) Numerical computation of multivariate t -probabilities with application to power calculation of multiple contrasts. *Journal of Statistical Computation and Simulation*, **63**, 361-378.
- **Hommel, G.** (1988) A stagewise rejective multiple test procedure based on modified Bonferroni test. *Biometrika*, **75**, 383-386.
- **Hommel, G. and Bernhard, G.** (1999) Bonferroni procedures for logically related hypotheses. *Journal of Statistical Planning and Inference*, **82**, 119-128.
- **Hsu, J.C. and Berger, R.L.** (1999) Stepwise confidence intervals without multiplicity adjustment for dose-response and toxicity studies. *Journal of the American Statistical Association*, **94**, 468-482.
- **Kieser, M., Reitmeir, P. and Wassmer, G.** (1995) Test procedures for clinical trials with multiple endpoints. *Biometrie in der chemisch-pharmazeutischen Industrie, Volume 6*, Vollmar, J. (ed.), Gustav Fischer Verlag, Stuttgart.
- **Marcus, R., Peritz, E. and Gabriel, K.B.** (1976) On closed testing procedures with special reference to ordered analysis of variance. *Biometrika*, **63**, 655-660.

Literatur (3)

- **Maurer, W., Hothorn, L. and Lehmacher, W.** (1995) Multiple comparisons in drug clinical trials and preclinical assays: a-priori ordered hypotheses. *Biometrie in der chemisch-pharmazeutischen Industrie, Volume 6*, Vollmar, J. (ed.), Gustav Fischer Verlag, Stuttgart.
- **Pigeot, I.** (2000) Basic concepts of multiple tests - a survey. *Statistical Papers*, **41**, 3-36.
- **Shaffer, J.** (1986) Modified sequentially rejective multiple test procedures. *Journal of the American Statistical Association*, **81**, 826-831.
- **Simes, R.J.** (1986) An improved Bonferroni procedure for multiple tests of significance. *Biometrika*, **73**, 751-754.
- **Tukey, J.W.** (1953) The problem of multiple comparisons. Unpublished manuscript reprinted in: *The Collected Works of John W. Tukey, Volume 8 (1994)*, Braun, H.I. (ed.), Chapman and Hall, New York.
- **Wassmer, G.** (2000) Basic concepts of group sequential and adaptive group sequential test procedures. *Statistical Papers*, **41**, 253-279.
- **Westfall, P.H.** (1997) Multiple testing of general contrasts using logical constraints and correlations. *Journal of the American Statistical Association*, **92**, 299-306.
- **Westfall, P.H. and Wolfinger, R.D.** (2000) Closed Multiple Testing Procedures and PROC MULTTEST. <http://www.sas.com/service/library/periodicals/obs/obswww23/>