Multiples Testen

-Multiple Gruppenvergleiche-

Dr. Martin Scharpenberg

MSc Medical Biometry/Biostatistics

WiSe 2019/2020

Beispiel: Vergleich von zwei zu einer Gruppe

- **Gruppe A:** 50% Stickstoff und 50% Sauerstoffgemisch für 24 Stunden. (Erwartungswert μ_1)
- **Gruppe B:** 50% Stickstoff und 50% Sauerstoffgemisch nur wärend der Operation. (Erwartungswert μ_2)
- **Gruppe C:** kein Stickstoff, 35-50% Sauerstoff für 24 Stunden. (Erwartungswert μ_3)
- Unterscheiden sich A oder B (mit Stickoxid) von C (kein Stickoxid)?
- Wir testen die zwei Hypothesen

$$H_0^1 : \theta_1 = \mu_1 - \mu_3 = 0 \qquad \text{und} \qquad H_0^2 : \theta_2 = \mu_2 - \mu_3 = 0$$

- **ANOVA-Modell:** $Y_{ij} \sim N(\mu_i, \sigma^2)$ unabhängige Beobachtungen.

Bonferroni-Test

- Wir verwerfen $H_0 = H_0^1 \cap H_0^2 = \{\theta = (0,0)\}$ falls

$$\max(|T_1|, |T_2|) \geq t_{\alpha/2}, \qquad t_{\alpha/2} = Q_{N-3}^t(1-\alpha/4)$$

  wobei $Q_{N-3}^t(u)$ das u-Quantil der t-Verteilung mit $N-3$ Freiheitsgraden ist ($N = n_1 + n_2 + n_3$) und

$$|T_i| = |\bar{Y}_i - \bar{Y}_3| / \left(\hat{\sigma}\sqrt{\frac{1}{n_i} + \frac{1}{n_3}}\right) \qquad (i = 1, 2)$$

- Wir wissen, dass für $\theta = \mathbf{0} = (0,0)$ ($\iff$ $\theta \in H_0$)

$$P_{\mathbf{0}}\{\max(|T_1|, |T_2|) \geq t_{\alpha/2}\} < \sum_{i=1,2} P_{\mathbf{0}}\{|T_i| \geq t_{\alpha/2}\} = 2\alpha/2 = \alpha$$

- Können wir diesen Test durch eine kleinere Verwerfungsschranke $d_\alpha < t_{\alpha/2}$ verbessern?

Gemeinsame Verteilung der Teststatistiken

- Wir können schreiben:

$$(T_1, T_2) = \left(\frac{\bar{Y}_1 - \bar{Y}_3}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_3}}}, \frac{\bar{Y}_2 - \bar{Y}_3}{\sigma\sqrt{\frac{1}{n_2} + \frac{1}{n_3}}} \right) \Big/ (\hat{\sigma}/\sigma)$$

- Mit $\rho = 1/\sqrt{(1 + \frac{n_3}{n_1})(1 + \frac{n_3}{n_2})}$ und $\theta_i^* = \theta_i/(\sigma\sqrt{n_i^{-1} + n_3^{-1}})$ gilt:

$$\text{Zähler} = \left(\frac{\bar{y}_1 - \bar{y}_3}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_3}}}, \frac{\bar{y}_2 - \bar{y}_3}{\sigma\sqrt{\frac{1}{n_2} + \frac{1}{n_3}}} \right) \sim N\left(\begin{pmatrix} \theta_1^* \\ \theta_2^* \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right)$$

- Es gilt zudem: $\text{Nenner} = \hat{\sigma}^2/\sigma^2 \sim \chi^2_{N-3}/(N-3)$
- und: Zähler und Nenner sind stochastisch unabhängig
- Man nennt die Verteilung von (T_1, T_2) die *bivariate t-Verteilung*

Vergleich von zwei zu einer Gruppe

- Bivariate Verteilungsfunktion von $(|T_1|, |T_2|)$ durch num. Integration:

$$F_\theta(x_1, x_2) = P_\theta(|T_1| \leq x_1, |T_2| \leq x_2) \ .$$

- Kritische Grenze d_α durch num. Lösen der Gleichung

$$P_{\mathbf{0}}\left(\max(|T_1|, |T_2|) \geq d_\alpha\right) = 1 - F_{\mathbf{0}}(d_\alpha, d_\alpha) = \alpha$$

- **Das multiple Signifikanzniveau wird damit ausgeschöpft!**
- In unserem Beispiel ($n_1 = 8$, $n_2 = 9$, $n_3 = 5$) gilt $d_\alpha = 2.3649$ (berechnet mit `qmvt` des R-Packages `mvtnorm`).
- Da $1 - F_{\mathbf{0}}(t_{\alpha/2}, t_{\alpha/2}) < \alpha$ und $F_{\mathbf{0}}(c, c)$ monoton wachsend in c gilt:

$$d_\alpha < t_{\alpha/2} \qquad (t_{\alpha/2} = Q^t_{19}(0.9875) = 2.4334 \text{ im Bsp.})$$

- **Wir verwerfen also öfter als mit Bonferroni!**

Dunnett-Test für drei Gruppen (2 Exp. + 1 Kontr.)

- Der Test, der H_0 verwirft, wenn

$$\max(|T_1|, |T_2|) \geq d_\alpha$$

heißt (zweiseitiger) **Dunnett-Test**.

- Die kritische Grenze d_α heißt **Dunnett-Grenze.**
- d_α hängt nicht nur von α ab, sondern auch von $df = N - 3$ und

$$\rho = 1/\sqrt{(1 + n_3/n_1)(1 + n_3/n_2)}$$

- Kritischen Grenzen für $df = N - 3 = 19$:

ρ	0	0.2	0.6	0.8	1	$t_{\alpha/2}$	t_α
d_α	2.4208	2.4007	2.3709	2.3143	2.0930	2.4334	2.0930

Zweiseitiger Dunnett-Test für k Gruppen

- Aus ANOVA-Modell folgt:
  - k Gruppenmittelwerte $\bar{Y}_j \sim N(\mu_j, \sigma^2/n_j)$ stoch. unabhängig.
  - Varianzschätzer: $\hat{\sigma}^2/\sigma^2 \sim \chi^2_{N-k}/(N-k)$ $\quad (N = \sum_{i=1}^{k} n_i)$ dessen Verteilung von θ unabhängig ist.
  - $\bar{Y}_1, \ldots, \bar{Y}_k, \hat{\sigma}^2$ sind stoch. unabhängig

- **Many-To-One Comparison (zweiseitig):**

  Wir testen die Hypothesen

$$H_0^1 : \mu_1 = \mu_k \quad , \ldots , \quad H_0^{k-1} : \mu_{k-1} = \mu_k$$

Zweiseitiger Dunnett-Test für k Gruppen

- Betrachten für jedes H_0^j die Teststatistik

$$T_j = (\bar{Y}_j - \bar{Y}_k)/(\hat{\sigma}\sqrt{n_j^{-1} + n_k^{-1}})$$

- Bestimmen $d_{k,\alpha}$, so dass für $\theta = \mathbf{0} = (0, \ldots, 0) \in \mathbb{R}^{k-1}$

$$P_{\mathbf{0}}\left(\max_{j=1}^{k-1} |T_j| \geq d_{k,\alpha}\right) = \alpha$$

- Wir verwerfen H_0^j falls $|T_j| \geq d_{k,\alpha}$.

- **Bem.:** Krit. Grenze $d_{k,\alpha}$ hängt nun ab von α, k, $N - k$ und den Korrelationen

$$\rho_{ij} = \mathrm{Corr}(\bar{Y}_i - \bar{Y}_k, \bar{Y}_j - \bar{Y}_k) = 1/\sqrt{(1 + n_k/n_i)(1 + n_k/n_j)},\ 1 \leq i \leq j \leq k-1$$

Beispiel: Effektivität von Eniporide

Vergleich von 4 Dosen Eniporide zu Placebo bei akutem Herzinfarkt in randomisierter Studie mit ingesamt 430 Patienten

- **Gruppe 1:** Placebo (88 Pat.)
- **Gruppe 2:** 50 mg Eniporide (86 Pat.)
- **Gruppe 3:** 100 mg Eniporide (91 Pat.)
- **Gruppe 4:** 150 mg Eniporide (74 Pat.)
- **Gruppe 5:** 200 mg Eniporide (91 Pat.)

Primärer Endpunkt: α-HBDH AUC (0 bis 72 Stunden)

Dunnett-Methode – Beispiel Eniporide mit SAS

```
proc glm data=grvgl.zeym;
class group;
model HBDH = group;
lsmeans group / adjust=dunnett;
run;
```

The GLM Procedure

Least Squares Means

Adjustment for Multiple Comparisons: Dunnett

group	HBDH LSMEAN	H0:LSMean= Control Pr > |t|
0	44.2000000	
1	45.3000000	0.9960
2	40.2000000	0.6921
3	33.9000000	0.0425
4	34.6000000	0.0477

Dunnett-Methode – Beispiel Eniporide mit R

```
> zeymer2  <- read.table('ZeymerS2.dat',header=T)
> library(multcomp)
> bmod <- aov(HBDH ~ group, data=zeymer2)
> bmod_glht <- glht(bmod, linfct = mcp( group="Dunnet") )
> summary(bmod_glht)

         Simultaneous Tests for General Linear Hypotheses
Multiple Comparisons of Means: Dunnett Contrasts
Fit: aov(formula = HBDH ~ group, data = zeymer2)

Linear Hypotheses:
           Estimate Std. Error t value p value       p Bonf  p raw
1 - 0 == 0    1.100      3.938   0.279  0.9960       1.000   0.780
2 - 0 == 0   -4.000      3.883  -1.030  0.6921       1.000   0.303
3 - 0 == 0  -10.300      4.096  -2.515  0.0425 *     0.049   0.012
4 - 0 == 0   -9.600      3.883  -2.473  0.0477 *     0.054   0.014
--
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Adjusted p values reported - single-step method)
```

Dunnett Methode – Vor- und Nachteile

- **Vorteile:**
  - Man verwirft immer mehr Hypothesen $H_0^{(i,k)} : \mu_i = \mu_k$ als mit dem Bonferronitest.
- **Nachteile:**
  - Kann nur bei homogener Varianz angewendet werden.
  - Erlaubt keine anderen Vergleiche als mit μ_k (oder einem anderen vorab gewählten μ_i wie z.B. μ_1).

Adjustierte p-Werte

Defintion: adjustierter p-Wert

Der adjustierte p-Wert für eine Nullhypothese H_0^i eines multiplen Tests ist das kleinste (multiple) Signifikanzniveau, auf dem H_0^i mit dem multiplen Test verworfen wird.

- Es gilt: Der multiple Test verwirft H_0^i auf dem Niveau $\alpha \iff p_i^{adj} \leq \alpha$
- **Bonferroni-Test:**
  Testen $H_0^1, \ldots, H_0^h$ mit p-Werten $p_1, \ldots, p_h$ und verwerfen H_0^j falls $p_j \leq \alpha/h$.
  Adjustierter p-Wert für H_0^j ist (Erklärung an Tafel)

$$p_j^{adj} = \min(h \cdot p_j, 1)$$

Übung: Bestimme den adjustierten p-Wert des (a) gewichteten Bonferroni-Tests und (b) Bonferroni-Holm-Tests (komplizierter).

Adjustierter p-Wert des Dunnett-Tests

- t_j = beobachter Wert von T_j
- Verwerfen H_0^j auf multiplen Signifikanzniveau u falls $|t_j| \geq d_{k,u}$
- $\Rightarrow$ kleinstes u unter dem H_0^j verworfen werden kann, erfüllt

$$|t_j| = d_{k,u} \quad \Longleftrightarrow \quad u = P_0\left(\max_{l=1}^{k-1} |T_l| \geq |t_j|\right)$$

- Damit ist der adjustierte p-Wert des Dunnett-Test

$$p_j^{adj} = P_0\left(\max_{l=1}^{k-1} |T_l| \geq |t_j|\right)$$

- Es gilt:

$$|t_j| \geq d_{k,\alpha} \quad \Longleftrightarrow \quad p_j^{adj} \leq \alpha$$

Simultane Test-Prozeduren

- Wir testen h Nullhypothesen $H_0^1, \ldots H_0^h$
- Wir benutzen die h Teststatistiken $T_1, \ldots, T_h$
- Wir bestimmen d_α, so dass für $H_0 = H_0^1 \cap \cdots \cap H_0^h$

$$P_{H_0}(\max_{l=1}^{h} T_l \geq d_\alpha) = \alpha$$

- Wir verwerfen H_j, falls $T_j \geq d_\alpha$.
- Adjustierter p-Wert für H_0^j: t_j beobachteter Wert von T_j

$$p_j^{adj} = P_{H_0}(\max_{l=1}^{h} T_l \geq t_j)$$

- Dunnett-Test ist ein Beispiel

Beispieldatensatz 'cholesterol'

- Datensatz `cholesterol` im R-Package `multcomp`:
- Vergleich von 5 Behandlungen zur Cholesterol-Reduktion in randomisierter Studie mit 10 Patienten pro Gruppe (N=50)
  - **Gruppe 1:** Neues Medikament, 1 mal 20 mg pro Tag
  - **Gruppe 2:** Neues Medikament, 2 mal 10 mg pro Tag
  - **Gruppe 3:** Neues Medikament, 4 mal 5 mg pro Tag
  - **Gruppe 4:** Kontrollmedikament D
  - **Gruppe 5:** Kontrollmedikament E.
- Wir interessieren uns nun für alle $h = 5 \cdot 4/2 = 10$ Gruppenvergleiche

Tukey-Test für alle paarweisen Vergleiche

- Wir testen für k Gruppen unter Annahme des ANOVA-Modells die $h = k(k-1)/2$ Nullhypothesen

$$H_0^{ij} : \mu_i = \mu_j, \quad i \neq j, \quad 1 \leq i < j \leq k$$

mit den Teststatistiken

$$T_{ij} = (\bar{Y}_i - \bar{Y}_j)/(\hat{\sigma}\sqrt{n_i^{-1} + n_j^{-1}})$$

- Der **adjustierte p-Wert** für H_0^{ij} des Tukey-Tests ist

$$p_{ij}^{adj} = P_{H_0}(\max_{i<j} |T_{ij}| \geq |t_{ij}|)$$

wobei t_{ij} der beobachtete Wert von T_{ij} ist

- Wir verwerfen H_0^{ij}, wenn $p_{ij}^{adj} \leq \alpha$

Tukey-Methode – Beispiel 'Cholesterol' mit SAS

```
proc glm data=grvgl.cholest;
class trt;
model response = trt;
lsmeans trt / adjust=tukey;
run;
```

Least Squares Means for effect trt
Pr > |t| for H0: LSMean(i)=LSMean(j)
Dependent Variable: response

i/j	1	2	3	4	5
1		0.1381	0.0004	<.0001	<.0001
2	0.1381		0.2050	0.0010	<.0001
3	0.0004	0.2050		0.2512	<.0001
4	<.0001	0.0010	0.2512		0.0031
5	<.0001	<.0001	<.0001	0.0031	

Tukey-Methode – Beispiel 'Cholesterol' mit R

```
> library(multcomp)
> amod <- aov(response ~ trt, data=cholesterol)
> amod_glht <- glht(amod, linfct = mcp(trt = "Tukey"))
> summary(amod_glht)

         Simultaneous Tests for General Linear Hypotheses
Multiple Comparisons of Means: Tukey Contrasts
Fit: aov(formula = response ~ trt, data = cholesterol)

Linear Hypotheses:
                     Estimate Std. Error t value  p value
2times - 1time == 0     3.443      1.443   2.385 0.138134
4times - 1time == 0     6.593      1.443   4.568 0.000354 ***
drugD - 1time == 0      9.579      1.443   6.637  < 1e-04 ***
drugE - 1time == 0     15.166      1.443  10.507  < 1e-04 ***
4times - 2times == 0    3.150      1.443   2.182 0.205086
drugD - 2times == 0     6.136      1.443   4.251 0.000950 ***
drugE - 2times == 0    11.723      1.443   8.122  < 1e-04 ***
drugD - 4times == 0     2.986      1.443   2.069 0.251194
drugE - 4times == 0     8.573      1.443   5.939  < 1e-04 ***
drugE - drugD == 0      5.586      1.443   3.870 0.003041 **
--
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Adjusted p values reported -- single-step method)
```

Tukey Methode – Vor- und Nachteile

- **Vorteile:**
  - Man verwirft immer mehr Hypothesen als mit dem Bonferroni-Test.
- **Nachteile:**
  - Tukey Methode kann nicht auf beliebige statistische Tests angewendet werden, sondern nur auf den Kontrasttest für homogene Varianzen.
  - Man kann nur die k Paarvergleichskontraste testen.

Scheffé Methode

- Bei der Methode von Scheffé testen wir alle (d.h. beliebige) Kontraste

$$H_0^c : c^T \cdot \mu = 0 \quad \text{für } c = (c_1, \ldots, c_k) \text{ mit } \sum_i^k c_i = 0 \ .$$

- Es ist $H_0 = \cap_{c \text{ Konstrast}} H_0^c = \{\mu_1 = \ldots = \mu_k\}$
- Der **adjustierte p-Wert** für die Hypothese H_0^c ist entsprechend

$$p_c^{\text{adj}} = P_{H_0}(\max_{c' \text{Kontrast}} |T_{c'}| \geq |t_c|) = P_{H_0}(\max_{c' \text{Kontrast}} T_{c'}^2 \geq t_c^2)$$

  wobei t_c der für T_c beobachtet Wert ist. T_c ist die letzte Woche definierte Kontrastteststatistik

- Scheffé hat gezeigt, dass unter H_0

$$\max_{c' \text{Kontrast}} T_{c'}^2/(k-1) \quad \sim \quad F_{k-1,N-k}$$

Scheffé Methode - Ableitung des maximalen $T^2_{c'}$

Mit $\bar{y} = \sum_{j=1}^{k} n_j \bar{y}_j / \sum_{j=1}^{k} n_j$ gilt (Cauchy-Schwarzsche Ungleichung):

$$|c^T \hat{\mu}|^2 = \left| \sum_{j=1}^{k} c_j \bar{y}_j \right|^2 = \left| \sum_{j=1}^{k} c_j (\bar{y}_j - \bar{y}) \right|^2 = \left| \sum_{j=1}^{k} \frac{c_j}{\sqrt{n_j}} \sqrt{n_j} (\bar{y}_j - \bar{y}) \right|^2 \leq \left(\sum_{j=1}^{k} \frac{c_j^2}{n_j} \right) \sum_{l=1}^{k} n_l (\bar{y}_l - \bar{y})^2$$

Hieraus ergibt sich die Ungleichung

$$|c^T \hat{\mu}|^2 / \sum_{j=1}^{k} (c_j^2 / n_j) \leq \sum_{l=1}^{k} n_l (\bar{y}_l - \bar{y})^2$$

Diese Ungleichung wird zur Gleichung für

$$c = c_{\max} = (\sqrt{n_1}(\bar{y}_1 - \bar{y}), \ldots, \sqrt{n_k}(\bar{y}_k - \bar{y}))^T$$

Hieraus ergibt sich

$$\frac{\sup_{c'} T^2_{c'}}{k-1} = \frac{|c_{\max}^T \hat{\mu}|^2 / \sum_{j=1}^{k} (c^2_{\max,j} / n_j)}{(k-1)\,\hat{\sigma}^2} = \frac{\sum_{j=1}^{k} n_j (\bar{y}_j - \bar{y})^2}{(k-1)\,\hat{\sigma}^2} = \text{Statistik des F-Tests}$$

Scheffé Methode

- Testen die Kontraste

$$H_0^c : c^T \cdot \mu = 0 \quad \text{für } c = (c_1, \ldots, c_k) \text{ mit } \sum_i^k c_i = 0$$

  mit den adjustierten p-Werten p_c^{adj}.

- D.h. man kann H_0^c verwerfen, wenn $p_c^{\text{adj}} \leq \alpha$.
  - **Vorteil:** Kontraste müssen vorab nicht festgelegt werden.
  - **Nachteil** Adjustierte p-Werte sind oft unnötig gross, insbesondere bei kleiner Zahl an Kontrasten.
- Es gilt immer

$$p_{ij}^{Tukey} < p_{ij}^{Bonferroni} \quad \text{und} \quad p_{ij}^{Tukey} < p_{ij}^{\text{Scheffé}}$$

Scheffé-Methode – Beispiel 'Cholesterol' mit SAS

```
proc glm data=grvgl.cholest;
class trt;
model response = trt;
lsmeans trt / adjust=scheffe;
run;
```

Least Squares Means for effect trt
Pr > |t| for H0: LSMean(i)=LSMean(j)
Dependent Variable: response

i/j	1	2	3	4	5
1		0.2420	0.0015	<.0001	<.0001
2	0.2420		0.3279	0.0037	<.0001
3	0.0015	0.3279		0.3824	<.0001
4	<.0001	0.0037	0.3824		0.0103
5	<.0001	<.0001	<.0001	0.0103	

Scheffé Methode – Beispiel 'Cholesterol' mit R

```
> Scheffe.p.value <- (1-pf(cmw^2/(4*SE^2),df1=4,df2=50-5))
> data.frame(Mean.Diff=cmw,SE=sqrt(gvar/5),Bonf.p.value,
+ Scheffe.p.value)
```

	Mean.Diff	SE	Bonf.p.value	Scheffe.p.value
2times-1time	3.44	1.44	2.13e-01	2.42e-01
4times-1time	6.59	1.44	3.82e-04	1.54e-03
drugD-1time	9.58	1.44	3.53e-07	2.60e-06
drugE-1time	15.17	1.44	1.08e-12	1.32e-11
4times-2times	3.15	1.44	3.44e-01	3.28e-01
drugD-2times	6.14	1.44	1.06e-03	3.74e-03
drugE-2times	11.72	1.44	2.29e-09	2.18e-08
drugD-4times	2.99	1.44	4.43e-01	3.82e-01
drugE-4times	8.57	1.44	3.84e-06	2.40e-05
drugE-drugD	5.59	1.44	3.48e-03	1.03e-02