

# Vergleich von $k$ unabhängigen Gruppen

(einfaktorielle, einfache Varianzanalyse)

A: Faktor (Gruppenvariable) mit  $k$  Stufen (Faktorstufen)

## Modell

$$Y_{ij} = \mu + \alpha_i + \epsilon_{ij}, \quad i = 1 \dots k, j = 1 \dots n_i$$

- $\mu$ : Gesamterwartungswert
- $\alpha_i$ : Effekt der  $i$ -ten Stufe von A
- $\epsilon_{ij}$ : Fehler,  $\epsilon_{ij} \sim (0, \sigma^2)$
- $Y_{ij}$ :  $j$ -te Beobachtung der  $i$ -ten Faktorstufe
- $\sum_{i=1}^k \alpha_i = 0$  Parametrisierungsbedingung

# Einfache Varianzanalyse

$$H_0 : \alpha_1 = \alpha_2 = \dots = \alpha_k$$

$$H_1 : \alpha_i \neq \alpha_l \text{ (für ein } i \neq l \text{)}$$

Im Fall  $k = 2$  führt dieses Testproblem auf das Zweistichprobenproblem ( $\rightarrow$  t-Test).

Output der Maschinen gleich?

Klausurergebnisse unterschiedlich?

Mageninhalt der Eidechsen gleich?

Cortisolgehalt unterschiedlich?

ANOVA\_Maschinen

Varianzanalyse\_Modelle\PI12erg

GLM\_Eidechsen

GLM\_Cortisol

# Varianzanalyse

Varianzanalyse macht eine Streuungszerlegung:

Gesamt- varianz	=	Varianz zwischen den Faktorstufen	+	Varianz innerhalb der Faktorstufen
SST (Total)	=	SSB (Between)	+	SSW   (SSE) (Within) (Error)

$$N = \sum_{i=1}^k n_i$$

$$\bar{Y}_i = \frac{1}{n_i} \cdot \sum_{j=1}^{n_i} Y_{ij}, \quad \bar{Y} = \frac{1}{N} \sum_{i,j} Y_{i,j}$$

# Einfache Varianzanalyse

**Satz:** Es gilt

$$SSB + SSW = SST$$

wobei

$$SSB = \sum_{i=1}^k n_i (\bar{Y}_i - \bar{Y})^2 \quad (\underline{\text{Between}})$$

$$SSW = \sum_{i=1}^k \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_i)^2 \quad (\underline{\text{Within}})$$

$$SST = \sum_{i=1}^k \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y})^2. \quad (\underline{\text{Total}})$$

# Einfache Varianzanalyse

**Satz:**  $SSB + SSW = SST$

Beweis:

$$SSB = \sum_i n_i \bar{Y}_i^2 - 2 \cdot N \cdot \bar{Y}^2 + \bar{Y}^2 \cdot N$$

$$SSW = \sum_{i,j} Y_{ij}^2 - 2 \cdot \sum_i n_i \bar{Y}_i^2 + \sum_i n_i \bar{Y}_i^2$$

$SSB + SSW =$

$$= \sum_{i,j} Y_{ij}^2 + \sum_i n_i \bar{Y}_i^2 - N \cdot \bar{Y}^2 - \sum_i n_i \bar{Y}_i^2$$

$$= \sum_{i,j} Y_{ij}^2 - N \cdot \bar{Y}^2 = \sum_j \sum_i (Y_{ij} - \bar{Y})^2 = SST \blacksquare$$

# Varianzanalyse

## PROC ANOVA;

```
CLASS A; /*A: Faktor */
```

```
MODEL var=A;
```

```
MEANS A /
```

```
HOVTEST=Levene (TYPE=ABS); oder
```

```
HOVTEST=BF;
```

```
MEANS OUT=SAS-Ausgabedatei;
```

## RUN;

oder: PROC GLM anstelle von PROC ANOVA;

und weiter wie oben.

ANOVA: schneller

GLM: zusätzliche Auswertungen möglich, z.B. Ausgabe der Residuen

HOVTEST: Test auf Varianzhomogenität

# Einfache Varianzanalyse

Dependent Variable: Y

Source	DF	Sum of Squares	Mean Squares	F-value	$Pr > F$
MODEL	$k-1$	SSB(M)	MSB	$\frac{MSB}{MSE}$	p-Wert
ERROR	$N-k$	SSW(E)	MSE		
Total	$N-1$	SST			

$$MSB = \frac{SSB}{k-1},$$

$$MSE = \frac{SSW}{N-k}$$

$$H_0 : \alpha_1 = \dots = \alpha_k$$

$$H_1 : \exists(i, j) : \alpha_i \neq \alpha_j$$

# Einfache Varianzanalyse

$H_0$  wird getestet mit

$$F = \frac{MSB}{MSE} = \frac{\text{Mittlere Var. zwischen d. Gruppen}}{\text{Mittlere Var. innerhalb d. Gruppen}}$$

$$= \frac{N - k}{k - 1} \quad \frac{SSB}{SSW} = \frac{N - k}{k - 1} \quad \frac{SST - SSW}{SSW}$$

$F$  groß,  $F > F_{1-\alpha, k-1, N-k} \Leftrightarrow H_0$  abgelehnt

Bestimmtheitsmaß

$$R^2 := \frac{SSB}{SST} = \frac{SST - SSW}{SST} = 1 - \frac{SSW}{SST}$$

Der Anteil der Varianz, der durch das Modell bestimmt wird, heißt Bestimmtheitsmaß

# Einfache Varianzanalyse

Offenbar:  $0 \leq R^2 \leq 1$ .

$$F = \frac{MSB}{MSE} = \frac{N - k}{k - 1} \cdot \frac{SSB}{SST} \cdot \frac{SST}{SSW} = \frac{N - k}{k - 1} \cdot \frac{R^2}{1 - R^2}$$

$$R^2 \rightarrow 0 \implies F \rightarrow 0$$

$$R^2 \rightarrow 1 \implies F \rightarrow \infty.$$

Schätzung der Modellstandardabweichung  $\sigma$

$$RootMSE = \sqrt{MSE} = \sqrt{\frac{1}{N-k} SSE}$$

Variationskoeffizient

$$CV = \frac{100 \cdot RootMSE}{\bar{Y}}$$

# Einfache Varianzanalyse

## Anmerkungen

- Der  $F$ -Test in der Varianzanalyse ist (einigermaßen) robust gegenüber Abweichungen von der Normalverteilungsannahme
- Wenn man die Prozedur GLM verwendet, dann kann man die sogen. Residuen

$$\hat{\epsilon}_{ij} = Y_{ij} - \hat{\alpha}_i - \hat{\mu}$$

abspeichern (Option RESIDUAL im OUTPUT-Statement)  
und auf Normalität testen.  
(PROC UNIVARIATE NORMAL)

# Varianzanalyse

- ● *F*-Test verlangt auch Varianzhomogenität  
Daten balanziert (gleiche Stichprobenumfänge)  
→ Abweichungen nicht so schwerwiegend.
- ● Wenn die Varianzen verschieden sind, kann die Welch-Modifikation verwendet werden:  
MEANS Var/WELCH;

# Einfache Varianzanalyse

## Test auf Varianzhomogenität

$$H_0 : \sigma_1^2 = \sigma_2^2 = \dots = \sigma_k^2$$

$$H_1 : \exists(i, l) : \sigma_i^2 \neq \sigma_l^2$$

### Levene Test (1960)

HOVTEST= LEVENE im MEANS-Statement

$$Z_{ij}^* = |Y_{ij} - \bar{Y}_{i.}|$$

### Brown-Forsythe-Test (1974)

HOVTEST = BF

$$Z_{ij}^* = |Y_{ij} - \text{med}Y_i|$$

# Einfache Varianzanalyse

## Test auf Varianzhomogenität (2)

Mit diesen neuen ZV wird eine Varianzanalyse durchgeführt.

$$W = \frac{\frac{1}{k-1} \sum n_i (\bar{Z}_{i.}^* - \bar{Z}^*)^2}{\frac{1}{N-k} \sum_{i,j} (Z_{ij}^* - \bar{Z}_{i.}^*)^2} \sim F_{k-1, N-k}.$$

GLM\_Cortisol

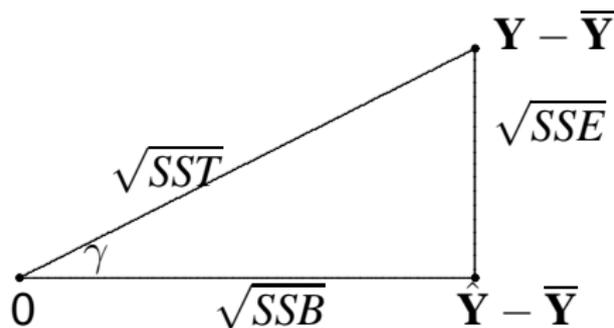
# Geometrische Veranschaulichung

zur Varianzanalyse

$$\mathbf{Y} = (Y_{11}, \dots, Y_{kn_k}) \quad \text{Dimension } N$$

$$\hat{\mathbf{Y}} = (\underbrace{\bar{Y}_1, \dots, \bar{Y}_1}_{n_1 \text{ mal}}, \dots, \underbrace{\bar{Y}_k, \dots, \bar{Y}_k}_{n_2 \text{ mal}})$$

$$\bar{\mathbf{Y}} = (\underbrace{\bar{Y}, \dots, \bar{Y}}_{n_1 \text{ mal}}), \quad \bar{Y} = \frac{1}{N} \sum_{i,j} Y_{ij}$$



$$SSB + SSW = SST$$

$$R^2 = \cos^2 \gamma$$

$$\|\hat{\mathbf{Y}} - \bar{\mathbf{Y}}\|^2 + \|\mathbf{Y} - \hat{\mathbf{Y}}\|^2 = \|\mathbf{Y} - \bar{\mathbf{Y}}\|^2$$

# Multiple Vergleiche

Problemstellung:  $H_0$  abgelehnt, aber zwischen welchen Faktorstufen liegt der Unterschied?

- Idee: Alle Paarvergleiche machen.
- Problem: Wenn wir das Signifikanzniveau  $\alpha (= 0.05)$  so lassen, wird das Testniveau nicht eingehalten!
- Veranschaulichung: Bei 20 gleichzeitigen Tests können wir  $20 \cdot \alpha = 1$  Ablehnung erwarten, auch wenn  $H_0$  richtig ist.

# Multiple Vergleiche

## Lösungsmöglichkeiten

Option BON im MEANS Statement

Signifikanzniveau für die gleichzeitigen Tests herabsetzen auf

$$\frac{\alpha_{nom}}{\binom{k}{2}},$$

bei  $k = 4$  wäre das etwa  $\frac{\alpha_{nom}}{\binom{4}{2}} = \frac{0.05}{6}$ .

Begründung: Bonferroni-Ungleichung.

Option TUKEY im MEANS Statement

Bilden die  $\bar{Y}_j$  und die Spannweite dazu

$$w = \max_{i,j} |\bar{Y}_i - \bar{Y}_j|$$

Dazu kommt noch die empirische Standardabweichung  $s$ .

# Multiple Vergleiche

## Lösungsmöglichkeiten

### Option TUKEY im MEANS Statement

$$t_{\max} = \frac{w}{s}$$

die sogenannte *studentisierte Spannweite*.

Diese hat (wenn die  $Y_i \sim \mathcal{N}$ ) eine (dem SAS-Programmierer) wohlbekanntere Verteilung, und entsprechende Quantile und kritische Werte.

Damit erhalten wir simultane Konfidenzintervalle für alle Paardifferenzen  $\mu_i - \mu_j$ . Liegt 0 nicht darin, so wird  $H_{0,ij} : \mu_i = \mu_j$  abgelehnt zugunsten von  $H_{A,ij} : \mu_i \neq \mu_j$ .

**Bem.:** Es gibt eine Fülle weiterer Varianten.