# Forschungsmethodik: Interventions- und Evaluationsforschung theoretisch begründen

*Vorlesung Biostatistik, WS 2026/2027*

*Dr. Marcus Wagner, IMISE*

# Vorlesung VII am 24.11.2026

---

**Inhaltsverzeichnis**

*10. Zweifaktorielle ANOVA mit Meßwiederholung*

*10.5. Kontraste und Querschnittsmodelle*

*11. Tests für Beziehungen zwischen kategorialen Variablen*

*11.1. Wiederholung: Tests bei metrischen Variablen*

*11.2. Kontingenztafel und Indifferenztabelle*

*11.3. Der Chi-Quadrat-Test*

*11.4. Chi-Quadrat-Test: Weitere Beispiele*

*11.5. Fisher's exakter Test*

*11.6. Zusammenfassung*

*12. Exkurs: Fehlwerte in linearen Modellen*

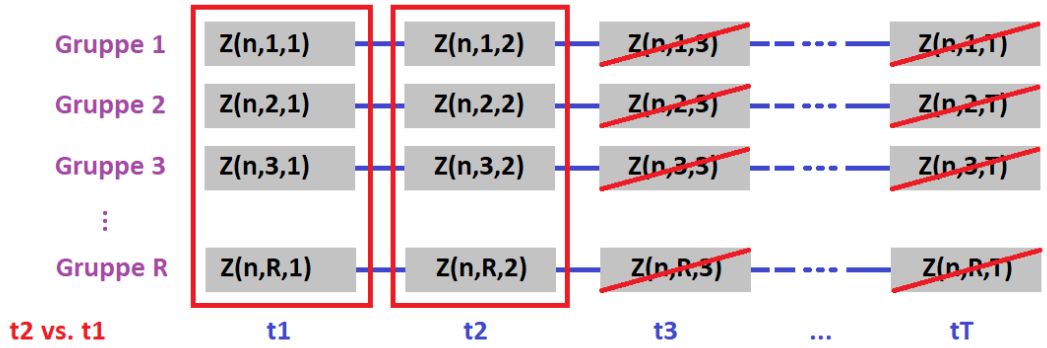
---

# Nachtrag zur zweifaktoriellen ANOVA

**Einfache Kontraste:** *Wir analysieren nur einen Teil des Datensatzes und vergleichen jeweils die Daten zu den Zeitpunkten $t_2$, $t_3$, ... , $t_T$ mit den Daten vom Zeitpunkt $t_1$.*

• **Einfacher Innersubjektkontrast Zeit:** *Verändert sich die Zielvariable im Mittel über alle Gruppen im Vergleich von $t_2$ vs. $t_1$, $t_3$ vs. $t_1$, ... , $t_T$ vs. $t_1$?*

• **Einfacher Innersubjektkontrast Zeit : Gruppe:** *Unterscheiden sich die Gruppen in der Veränderung der Zielvariablen im Vergleich von $t_2$ vs. $t_1$, $t_3$ vs. $t_1$, ... , $t_T$ vs. $t_1$?*



# Nachtrag: Einfache Innersubjektkontraste

• **Beispiel 7.1.:** *Wir fragen, ob sich in Beispiel 6.2. die mittlere Wegstrecke zwischen den Zeitpunkten $t_2$ und $t_1$ unterscheidet (einfacher Innersubjektkontrast Zeit).*

• **Verkürzte Varianz-Zerlegung:** *Berechnung wie in Vorlesung VI, aber darin werden nur die Daten der beiden Zeitpunkte $t_2$ und $t_1$ verwendet, die im Kontrast miteinander verglichen werden. Die Freiheitsgrade werden entsprechend angepaßt. Wir erhalten die modifizierten Zähler $SS'_{Zeit}$, $SS'_{RestInnSubj}$ und die modifizierten empirischen Varianzen $(\widehat{\sigma}')^2_{Zeit}$, $(\widehat{\sigma}')^2_{RestInnSubj}$.*

• **Prüfgröße für den Kontrast:** $\widehat{T}_{Kontrast} = (\widehat{\sigma}')^2_{Zeit} \,/\, (\widehat{\sigma}')^2_{RestInnSubj}$ *ist Fisher-verteilt. Mit $\widehat{T}_{Kontrast}$ kann man einen Post-hoc-Test gegen die Nullhypothese $\mathcal{H}_0$ "$\mu(t_2) = \mu(t_1)$" durchführen, denn aus $\mathcal{H}_0$ folgt $(\widehat{\sigma}')^2_{Zeit} = 0$ und $\widehat{T}_{Kontrast} = 0$.*

• **Bonferroni-Korrektur:** *Bei der zweifaktoriellen ANOVA sollte die Untersuchung der Kontraste vorab geplant werden. Da für jeden Kontrast ein Post-hoc-Test durchgeführt werden muß, ergibt sich die Bonferroni-Korrektur fürs Signifikanzniveau als die Gesamtanzahl aller untersuchten Kontraste.*

# Nachtrag: Einfache Innersubjektkontraste (II)

• **Verkürzte Varianz-Zerlegung:** *mit* $N = 50$, $R = 2$, $T = 2$ *(nur* $t_1$, $t_2$*)*

| *Bezeichnung Zähler* | *Wert Zähler* | *Nenner (df)* | *empirische Varianz* |
|---|---|---|---|
| $SS'_{Gesamt}$ | 251 208.8 | 199 | $(\widehat{\sigma}')^2_{Gesamt} = 1\,262.4$ |
| $SS'_{ZwSubj}$ | 125 710.3 | 99 | $(\widehat{\sigma}')^2_{ZwSubj} = 1\,269.8$ |
| $SS'_{Gruppe}$ | 588.2 | 1 | $(\widehat{\sigma}')^2_{Gruppe} = 588.2$ |
| $SS'_{InnSubj}$ | 125 498.5 | 100 | $(\widehat{\sigma}')^2_{InnSubj} = 125.5$ |
| $SS'_{Zeit}$ | 103 558.0 | 1 | $(\widehat{\sigma}')^2_{Zeit} = 103\,558.0$ |
| $SS'_{Zeit:Gruppe}$ | 1 794.0 | 1 | $(\widehat{\sigma}')^2_{Zeit:Gruppe} = 1\,794.0$ |
| $SS'_{RestInnSubj}$ | 20 146.5 | 98 | $(\widehat{\sigma}')^2_{RestInnSubj} = 205.6$ |

• *Wir testen gegen die Nullhypothese* $\mathcal{H}_0$ *"*$\mu(t_2) = \mu(t_1)$*" zum Bonferroni-korrigierten Signifikanzniveau* $\alpha' = 0.05/12 = 0.0041$. $\widehat{T}_{Kontrast}$ *ist* $F$*-verteilt mit (*$df_1$*,*$df_2$*) = (1,98).*

• *Wir erhalten* $\widehat{T}_{Kontrast} = 503.46$. *Zu* $\alpha' = 0.0041$ *gehört* $Q_{99.59}(F, 1, 98) = 8.64$, *also unterscheiden sich* $\mu(t_1)$ *und* $\mu(t_2)$ *signifikant (*$p < 0.0005$*).*

# Mittelwertvergleiche, parametrische Tests

---

**• Zwei unabhängige Stichproben:**

*— Klassischer $t$-Test: metrische Variable $x$ in $\mathcal{P}'$, $\mathcal{P}''$ normalverteilt, Standardabweichungen in $\mathcal{P}'$, $\mathcal{P}''$ gleich*
*— Welch-Test: metrische Variable $x$ in $\mathcal{P}'$, $\mathcal{P}''$ annähernd normalverteilt, Standardabweichungen in $\mathcal{P}'$, $\mathcal{P}''$ verschieden*

**• Zwei verbundene Stichproben:**

*Typische Situationen: Messungen ein und derselben Größe am gleichen Probanden zu verschiedenen Zeitpunkten / Fallkontrollstudien mit festen Paaren von Probanden*
*— Klassischer $t$-Test für Paardifferenzen: Differenzen der verbundenen Meßwerte annähernd normalverteilt*

**• Modelle mit drei oder mehr unabhängigen Stichproben / Meßwiederholung:**

*— Einfaktorielle ANOVA ("analysis of variance"): F-Globaltest und ggf. Post-hoc-Tests*
*— Zweifaktorielle ANOVA mit Meßwiederholung: F-Globaltests und Post-hoc-Analysen*

---

# Vergleich “zentraler Tendenzen”, nichtparametrische Tests

- **Gemeinsame Voraussetzungen:**

— *Daten metrisch, jedoch kleine Stichproben und / oder*

— *Verteilungen in $\mathcal{P}'$, $\mathcal{P}''$ ganz unbekannt oder stark von Normalverteilung abweichend*

- **Zwei unabhängige Stichproben:**

— *Mann-Whitney-U-Test*

— *Basis- oder Shifthypothesen, Vergleich der Verteilungen im ganzen*

— *Mittelwertvergleich als Folgerung aus den Shifthypothesen*

- **Zwei verbundene Stichproben:**

— *Wilcoxon-Rangsummentest*

- **Drei oder mehr unabhängige Stichproben:**

— *Kruskal-Wallis-Test (nichtparametrisches Äquivalent zur einfaktoriellen ANOVA)*

# Tests bei linearen Modellgleichungen

• **Lineare Regression:**

*— Besteht in $\mathcal{P}$ ein linearer Zusammenhang $Z \sim c_0 + c_1 \cdot K$ zwischen den metrischen Variablen $Z$ und $K$?*
*— Anhand einer Stichprobe $\mathcal{S}$: F-Globaltest gegen die Nullhypothese "$c_1 = 0$"*

• **Allgemeines lineares Modell:**

*— Besteht in $\mathcal{P}$ ein linearer Zusammenhang $Z \sim c_0 + c_1 \cdot K_1 + c_2 \cdot K_2 + c_3 \cdot K_3 + ...$ zwischen der metrischen Zielvariable $Z$ und den Kovariaten $K_1$, $K_2$, $K_3$, ... ?*
*— Anhand der Daten (einer "Stichprobe mit mehreren Variablen"): F-Globaltest gegen die Nullhypothese "$c_1 = c_2 = c_3 = ... = 0$"*

• **Zweifaktorielle ANOVA mit Meßwiederholung:**

*— Mauchly-Test gegen die Nullhypothese, daß die Kovarianzmatrix $C$ in $\mathcal{P}$ die technische Eigenschaft "Sphärizität" besitzt*

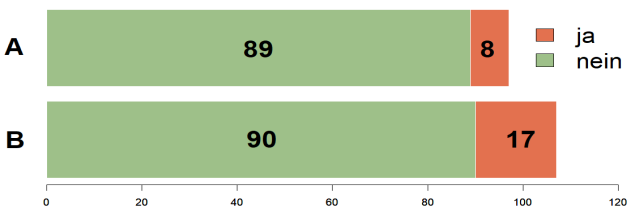
# Vergleich für Verteilungen kategorialer Variablen

• **Beispiel 7.2.:** *Wir betrachten die Verteilung der Personen mit vorzeitigem Studienende ("Dropouts") in den beiden Armen einer randomisierten Studie.*

• **Vierfeldertafel** *...*

| | Vorzeitiges Ende: nein | Vorzeitiges Ende: ja | Zeilen-summe |
|---|---|---|---|
| Arm A | 89 | 8 | 97 |
| Arm B | 90 | 17 | 107 |
| Spalten-summe | 179 | 25 | 204 |

• *... und Balkendiagramm*




• **Frage:** *Unterscheiden sich die Verteilungen der Dropouts in Arm A und Arm B? Wenn ja, so besteht die Gefahr, daß die Prinzipien von Strukturgleichheit und Beobachtungsgleichheit in der Studie verletzt werden ("differentieller Dropout").*

• **Gleiche Verteilung bedeutet:** *Die beobachteten Häufigkeiten stehen im gleichen Verhältnis wie die Zeilen- und Spaltensummen.*

# Identische Verteilungen: Indifferenztabelle

- *Beobachtete Daten:*

| | Vorzeitiges Ende: nein | Vorzeitiges Ende: ja | Zeilen-summe |
|---|---|---|---|
| Arm A | 89 | 8 | 97 |
| Arm B | 90 | 17 | 107 |
| Spalten-summe | 179 | 25 | 204 |

| | Vorzeitiges Ende: nein | Vorzeitiges Ende: ja | Zeilen-summe |
|---|---|---|---|
| Arm A | $N_{1,1}$ | $N_{1,2}$ | $Z_1$ |
| Arm B | $N_{2,1}$ | $N_{2,2}$ | $Z_2$ |
| Spalten-summe | $S_1$ | $S_2$ | $N$ |

- *Erwartete Häufigkeiten bei identischer Verteilung ("Indifferenztabelle"):*

| | Vorzeitiges Ende: nein | Vorzeitiges Ende: ja | Zeilen-summe |
|---|---|---|---|
| Arm A | $Z_1 \cdot S_1 / N$ | $Z_1 \cdot S_2 / N$ | $Z_1$ |
| Arm B | $Z_2 \cdot S_1 / N$ | $Z_2 \cdot S_2 / N$ | $Z_2$ |
| Spalten-summe | $S_1$ | $S_2$ | $N$ |

| | Vorzeitiges Ende: nein | Vorzeitiges Ende: ja | Zeilen-summe |
|---|---|---|---|
| Arm A | 85.1 | 11.9 | 97 |
| Arm B | 93.9 | 13.1 | 107 |
| Spalten-summe | 179 | 25 | 204 |

# Identische Verteilungen: Chi-Quadrat-Statistik

- *Erwartete Häufigkeiten bei identischer Verteilung:* $E_{i,j} = Z_i \cdot S_j / N$ , $1 \leqslant i,\, j \leqslant 2$.

- *Freiheitsgrade in der Vierfeldertafel mit festen Zeilen- und Spaltensummen: genau 1 Wert (zum Beispiel* $N_{1,1}$*) ist frei wählbar, alle anderen sind dadurch festgelegt.*

- *Prüfgröße: quadrierte Differenzen im Zähler, auf* $E_{i,j}$ *normiert*

$$\widehat{T} = \frac{(N_{1,1} - E_{1,1})^2}{E_{1,1}} + \frac{(N_{1,2} - E_{1,2})^2}{E_{1,2}} + \frac{(N_{2,1} - E_{2,1})^2}{E_{2,1}} + \frac{(N_{2,2} - E_{2,2})^2}{E_{2,2}}$$

- **Lehrsatz 7.3.:** *Gegeben sei eine Vierfeldertafel mit* $N \geqslant 40$ *und* $E_{i,j} \geqslant 5$ *für alle erwarteten Häufigkeiten. Bilden wir zufällig Vierfeldertafeln mit denselben Zeilen- und Spaltensummen, so folgt: Wenn die Einträge* $N_{1,1}$ *normalverteilt mit Mittelwert* $\mu = E_{1,1}$ *sind, dann sind alle Differenzen* $(N_{i,j} - E_{i,j})$ *normalverteilt mit* $\mu = 0$*, und die Prüfgröße* $\widehat{T}$ *unterliegt in sehr guter Näherung einer* $\chi^2$*-Verteilung mit einem Freiheitsgrad* $df = 1$*.*

# Der Chi-Quadrat-Test

*Wir formalisieren die Fragestellung aus Beispiel 7.2. als statistischen Hypothesentest.*

• *Populationen $\mathcal{P}'$, $\mathcal{P}''$ — alle denkbaren Studienteilnehmer, die entweder den Prozeduren in Arm A oder in Arm B unterworfen werden.*

• *Stichproben: $\mathcal{S}'$, $\mathcal{S}''$ — die tatsächlichen Studienteilnehmer in Arm A ($N = 97$) bzw. Arm B ($N = 107$)*

• *Nullhypothese $\mathcal{H}_0$: "Die Verteilungen der (potentiellen) Dropouts in $\mathcal{P}'$ und $\mathcal{P}''$ stimmen überein, also sind alle Mittelwerte von $(N_{i,j} - E_{i,j})$ Null."*

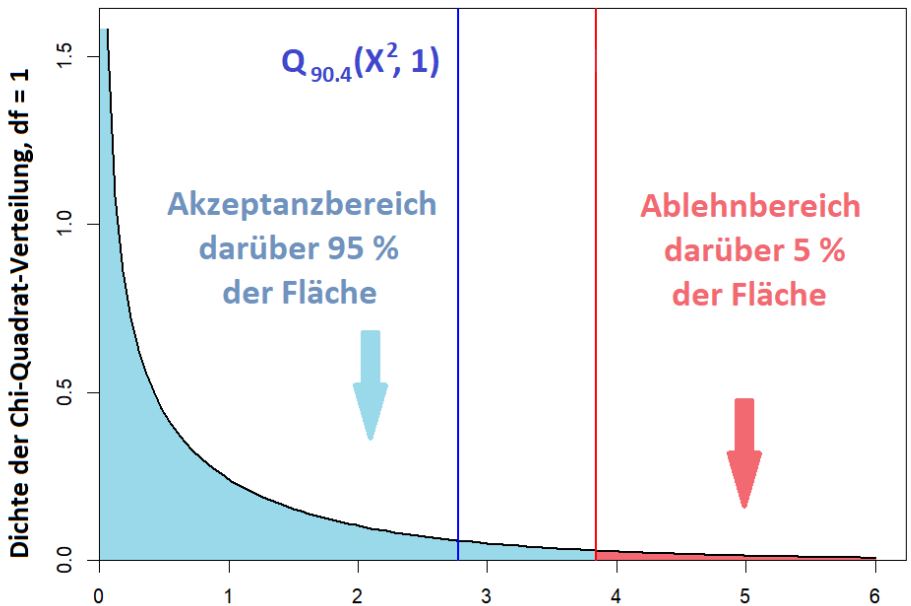
• *Alternativhypothese $\mathcal{H}_A$: "Die Verteilungen der (potentiellen) Dropouts in $\mathcal{P}'$, $\mathcal{P}''$ sind verschieden, also sind nicht alle Mittelwerte von $(N_{i,j} - E_{i,j})$ Null."*

• *Freiheitsgrade berechnen: $k = 2$, $m = 2$, df $= (k-1) \cdot (m-1) = 1$*

• *Signifikanzniveau festlegen: $\alpha = 0.05$, das zugehörige 95-%-Quantil der Chi-Quadrat-Verteilung mit df $= 1$ ist $Q_{95}(\chi^2, 1) = 3.8415$ .*

• *Die Voraussetzungen von Satz 7.3. sind erfüllt: $N = 204 > 40$, alle $E_{i,j} \geqslant 5$.*

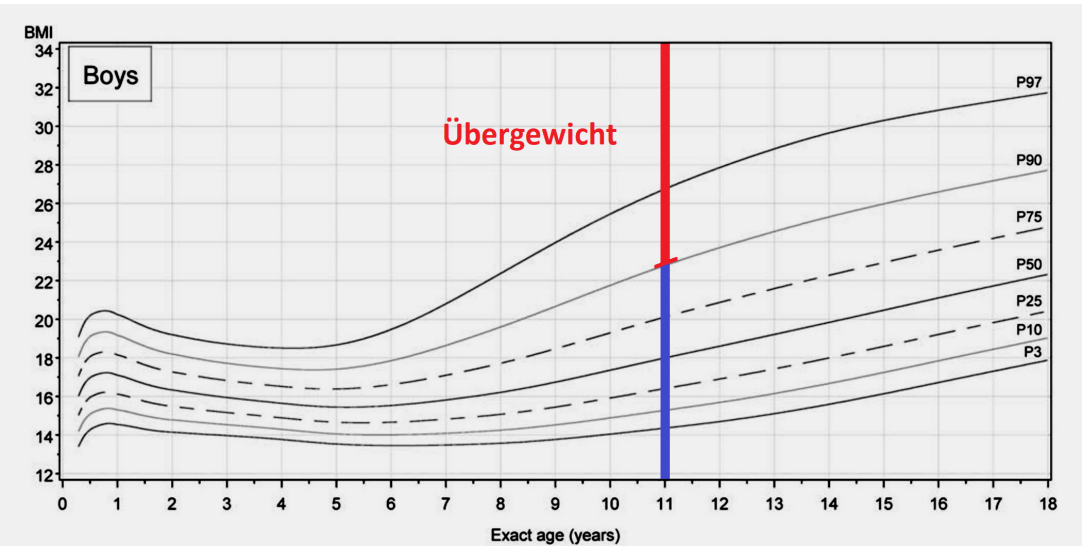
# Der Chi-Quadrat-Test (II)




- **Prüfgröße:** $\widehat{T} = 2.779$. *Dazu gehört das 90.4-%-Quantil* $Q_{90.4}(\chi^2, 1)$ *der* $\chi^2$*-Verteilung mit einem Freiheitsgrad* $df = 1$*, und wir erhalten* $p = 0.096$ *als Restfläche.*
- **Testentscheidung:** *Die Nullhypothese* $\mathcal{H}_0$ *wird beibehalten. "Die Verteilungen der Dropouts in den Studienarmen A und B unterscheiden sich nicht* $(p = 0.096)$*."*

# Naschverhalten und Übergewicht

• **Beispiel 7.4.:** *Gibt es bei etwa 11-jährigen Schulkindern einen Zusammenhang zwischen Naschverhalten und Übergewicht?*




• **Übergewicht bei Kindern:** *Body-Mass-Index muß auf oder über der alters- und geschlechtsabhängigen 90-Prozent-Quantilkurve ("Perzentilkurve") liegen.*

# Naschverhalten und Übergewicht: Hypothesentest

• **Naschverhalten:** *passender Fragebogen mit Scores zwischen 0 und 5; Schwelle für "naschhaft" liegt bei Score $> 2.5$ .*

• **Daten** *aus* [ WALTHER ET AL. 2011 ]

| | Übergewicht: nein | Übergewicht: ja | Zeilen-summe |
|---|---|---|---|
| naschhaft: nein | 81 | 14 | 95 |
| naschhaft: ja | 130 | 7 | 137 |
| Spalten-summe | 211 | 21 | 232 |

• **Indifferenztabelle**

| | Übergewicht: nein | Übergewicht: ja | Zeilen-summe |
|---|---|---|---|
| naschhaft: nein | 86.4 | 8.6 | 95 |
| naschhaft: ja | 124.6 | 12.4 | 137 |
| Spalten-summe | 211 | 21 | 232 |

• *Populationen: $\mathcal{P}'$, $\mathcal{P}''$ — nicht naschhafte bzw. naschhafte Schulkinder im Alter von etwa 11 Jahren; daraus Stichproben: $\mathcal{S}'$, $\mathcal{S}''$ mit $N = 95$ bzw. $N = 137$*

• *Nullhypothese $\mathcal{H}_0$: "Die Verteilungen der übergewichtigen Kinder in $\mathcal{P}'$ und $\mathcal{P}''$ stimmen überein."*

• *Alternativhypothese $\mathcal{H}_A$: "Die Verteilungen der übergewichtigen Kinder in $\mathcal{P}'$ und $\mathcal{P}''$ sind verschieden."*

# Naschverhalten und Übergewicht: Hypothesentest (II)

• *Signifikanzniveau festlegen:* $\alpha = 0.05$*; das zugehörige 95-%-Quantil der Chi-Quadrat-Verteilung mit df = 1 ist wiederum* $Q_{95}(\chi^2, 1) = 3.8415$ .

• *Voraussetzungen für den* $\chi^2$*-Test sind erfüllt:* $N = 232 > 40$*, alle* $E_{i,j} \geqslant 5$.

• *Berechnung der Prüfgröße:* $\widehat{T} = \sum_{i=1}^{2} \sum_{j=1}^{2} (N_{i,j} - E_{i,j})^2 / E_{i,j} = 6.314$ *. Dazu gehört* $Q_{98.8}(\chi^2, 1)$*, und wir erhalten* $p = 0.012$*. Folgt:*

• *Die Nullhypothese* $\mathcal{H}_0$ *wird verworfen und die Alternativhypothese* $\mathcal{H}_A$ *angenommen.*

• **Interpretation:**

— *Voreilig und falsch: "Die Wahrscheinlichkeit von Übergewicht unter Kindern, die gern naschen (*$\mathcal{P}''$*), ist höher als unter solchen, die wenig naschen (*$\mathcal{P}'$*)." Denn:*

— *Die relative Häufigkeit von Übergewicht in* $\mathcal{S}'$ *beträgt 14.7%, in* $\mathcal{S}''$ *nur 5.1%!*

— *Richtig: "Die Wahrscheinlichkeit von Übergewicht unter Kindern, die gern naschen, ist um 9.6% niedriger als unter solchen, die wenig naschen. Dieser Unterschied ist signifikant zum 5-%-Niveau (*$p = 0.012$*)."*

# Naschverhalten und BMI: mehrstufig

• **Beispiel 7.5.:** *Hängt das überraschende Ergebnis von Beispiel 7.4. mit der binären Stufung der Daten zusammen?*
*— Naschverhalten: Score in* $[\,0\,,\,2.5\,]$, $(\,2.5\,,\,3.75\,]$, $(\,3.75\,,\,5.0\,]$ *(dreistufig)*
*— BMI: niedrig bis 33.3-%-Perzentilkurve, mittel zwischen 33.3-%- und 66.6-%-Perzentilkurve, hoch ab 66.6-%-Perzentilkurve (ebenfalls dreistufig)*

• $(3\times3)$-**Kontingenztafel** *statt Vierfeldertabelle für die Daten*

| | BMI niedrig | BMI mittel | BMI hoch | Zeilen-summe |
|---|---|---|---|---|
| [0.00 , 2.50] | 25 | 35 | 35 | 95 |
| (2.50 , 3.75] | 44 | 34 | 36 | 114 |
| (3.75 , 5.00] | 11 | 9 | 3 | 23 |
| Spaltensumme | 80 | 78 | 74 | 232 |

• *Ein Chi-Quadrat-Test kann auch für mehr als zwei Stichproben und mehrstufige Verteilungen durchgeführt werden!*

# Chi-Quadrat-Test für beliebige kategoriale Variablen

- *Beobachtete Daten:* $(k \times m)$*-Kontingenztafel mit Einträgen* $N_{i,j}$*, Zeilensummen* $Z_i$*, Spaltensummen* $S_j$ *und Gesamtsumme* $N$*,* $1 \leqslant i \leqslant k$*,* $1 \leqslant j \leqslant m$
- *Freiheitsgrade in der* $(k \times m)$*-Kontingenztafel mit festen Zeilen- und Spaltensummen: genau* $(k-1) \cdot (m-1)$ *Werte sind frei wählbar, alle anderen sind dadurch festgelegt*
- *Erwartete Häufigkeiten bei unabhängiger Verteilung:* $E_{i,j} = Z_i \cdot S_j / N$
- $(3 \times 3)$-**Indifferenztabelle** *zu Beispiel 7.5.* $(k = 3, m = 3)$

| | BMI niedrig | BMI mittel | BMI hoch | Zeilen-summe |
|---|---|---|---|---|
| [0.00 , 2.50] | 32.8 | 31.9 | 30.3 | 95 |
| (2.50 , 3.75] | 39.3 | 38.3 | 36.4 | 114 |
| (3.75 , 5.00] | 7.9 | 7.8 | 7.3 | 23 |
| Spaltensumme | 80 | 78 | 74 | 232 |

- *Signifikanzniveau festlegen:* $\alpha = 0.05$ *(wie im vorigen Test)*

# Chi-Quadrat-Test für beliebige kategoriale Variablen (II)

• *Nullhypothese $\mathcal{H}_0$: "Bei Schulkindern von etwa 11 Jahren stimmen die Verteilungen der kategorialen Variable 'BMI-Stufe' für alle drei Gruppen, die bezüglich des Naschverhaltens gebildet wurden, überein."*

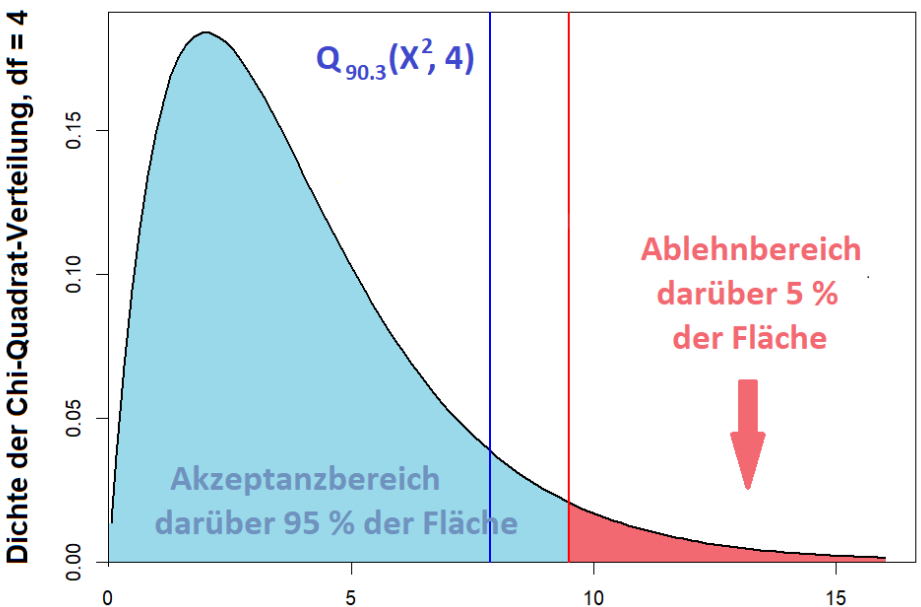
• *Alternativhypothese $\mathcal{H}_A$: "Bei Schulkindern von etwa 11 Jahren unterscheiden sich die Verteilungen der kategorialen Variable 'BMI-Stufe' zwischen mindestens zwei der drei Gruppen, die bezüglich des Naschverhaltens gebildet wurden."*

• *Prüfgröße: quadrierte Differenzen im Zähler, auf $E_{i,j}$ normiert*

$$\widehat{T} = \frac{(N_{1,1} - E_{1,1})^2}{E_{1,1}} + \frac{(N_{1,2} - E_{1,2})^2}{E_{1,2}} + \ \ldots \ + \frac{(N_{k,m} - E_{k,m})^2}{E_{k,m}}$$

• **Lehrsatz 7.6.:** *(Verteilung der Prüfgröße beim allgemeinen $\chi^2$-Test) Gegeben sei eine $(k \times m)$-Kontingenztafel mit $N \geqslant 40$ und $E_{i,j} \geqslant 5$ für mindestens 80% aller erwarteten Häufigkeiten. Bilden wir zufällig $(k \times m)$-Kontingenztafeln mit stets denselben Zeilen- und Spaltensummen, so unterliegt die Prüfgröße $\widehat{T}$ in sehr guter Näherung einer $\chi^2$-Verteilung mit df $= (k-1) \cdot (m-1)$ Freiheitsgraden.*

# Chi-Quadrat-Test für beliebige kategoriale Variablen (III)




- *Voraussetzungen für den $\chi^2$-Test sind erfüllt:*
  $N = 232 > 40$,
  *sogar alle* $E_{i,j} \geqslant 5$
- *Durchführung:*
  $\widehat{T} = 7.8684$,
  *dazu gehört das Quantil* $Q_{90.3}(\chi^2, 4)$
  *und* $p = 0.097$.

- *Wegen $p = 0.097 > 0.05$ wird die Nullhypothese $\mathcal{H}_0$ beibehalten: Die Verteilungen der BMI-Stufen unterscheiden sich nicht signifikant zwischen den drei Gruppen mit verschiedenem Naschverhalten. Das Testergebnis hängt also von der Gruppierung ab!*

# Chi-Quadrat-Test als Unabhängigkeitstest

• **Neue Interpretation:** *Daten in der Vierfelder- bzw. Kontingenztafel werden als gleichzeitige Beobachtungen von zwei kategorialen Variablen (Zeilenvariable, Spaltenvariable) in einer einzigen Stichprobe $\mathcal{S}$ mit Umfang $N$ angesehen.*

• **Neue Nullhypothese:** *"In der Gesamtpopulation $\mathcal{P}$ sind die beiden Variablen statistisch unabhängig." (Dann multiplizieren sich die Wahrscheinlichkeiten für das Auftreten der einzelnen Stufen, und als Häufigkeiten für die gleichzeitigen Beobachtungen entstehen gerade die erwarteten Häufigkeiten aus der Indifferenztabelle!)*

• **Neue Alternativhypothese:** *"In der Gesamtpopulation $\mathcal{P}$ sind die beiden Variablen statistisch abhängig." (Also weichen die Häufigkeiten für die gleichzeitigen Beobachtungen signifikant von der Indifferenztabelle ab!)*

• **Neue Interpretation von Beispiel 7.2.:** *Die Armzuordnung der Studienteilnehmer hat keinen signifikanten Einfluß auf das vorzeitige Ausscheiden.*

• **Neue Interpretation von Beispiel 7.5.:** *Die Verteilung der BMI-Stufen unter etwa 11-jährigen Schulkindern hängt nicht signifikant vom Naschverhalten ab.*

# Zu kleine erwartete Häufigkeiten

• **Beispiel 7.7.:** *Zwei Gruppen von Nichtschwimmern wird Schwimmunterricht nach zwei verschiedenen Methoden erteilt. Am Ende der Kurse wird die "Seepferdchen"-Prüfung abgelegt. Ist eine der beiden Unterrichtsmethoden erfolgreicher?*

• *Beobachtete Daten:*

| | "Seepferdchen" bestanden: ja | "Seepferdchen" bestanden: nein | Zeilen-summe |
|---|---|---|---|
| Unterrichts-methode I | 18 | 6 | 24 |
| Unterrichts-methode II | 22 | 3 | 25 |
| Spalten-summe | 40 | 9 | 49 |

• *Indifferenztabelle:*

| | "Seepferdchen" bestanden: ja | "Seepferdchen" bestanden: nein | Zeilen-summe |
|---|---|---|---|
| Unterrichts-methode I | 19.6 | 4.4 | 24 |
| Unterrichts-methode II | 20.4 | 4.6 | 25 |
| Spalten-summe | 40 | 9 | 49 |

• *Der $\chi^2$-Test ist nicht anwendbar, weil die Voraussetzung über $E_{i,j}$ verletzt ist!*

• *Aber: Unabhängig von der Größe von $N$ und $E_{i,j}$ darf Fisher's exakter Test durchgeführt werden.*

# Fisher's exakter Test

| | | | Zeilen-summe |
|---|---|---|---|
| | $N_{1,1}$ | $N_{1,2}$ | $Z_1$ |
| | $N_{2,1}$ | $N_{2,2}$ | $Z_2$ |
| Spalten-summe | $S_1$ | $S_2$ | $N$ |

• *Die Wahrscheinlichkeit, bei gegebenen Zeilen- und Spaltensummen die Vierfeldertafel links zu erhalten, kann nach Fisher exakt berechnet werden:*

$$Prob = \frac{Z_1! \cdot Z_2! \cdot S_1! \cdot S_2!}{N! \cdot N_{1,1}! \cdot N_{1,2}! \cdot N_{2,1}! \cdot N_{2,2}!}$$

• **Grundidee** *von Fisher's exaktem Test: Berechne zuerst die Wahrscheinlichkeit, die Vierfeldertafel mit den beobachteten Daten zu erhalten, und danach die entsprechenden Wahrscheinlichkeiten für alle "noch extremeren" Tafeln mit denselben Zeilen- und Spaltensummen. Die Summe aller dieser Wahrscheinlichkeiten ergibt die Signifikanz bzw. den $p$-Wert für den Hypothesentest.*

• *"Noch extremere" Tafeln sind solche, die noch stärker als die beobachteten Daten von der Nullhypothese $\mathcal{H}_0$ abweichen.*

# Fisher's exakter Test (II)

- *Beobachtete Daten:*

| | "Seepferdchen" bestanden: ja | "Seepferdchen" bestanden: nein | Zeilen-summe |
|---|---|---|---|
| Unterrichts-methode I | 18 | 6 | 24 |
| Unterrichts-methode II | 22 | 3 | 25 |
| Spalten-summe | 40 | 9 | 49 |

- *Eine "noch extremere" Vierfeldertafel:*

| | "Seepferdchen" bestanden: ja | "Seepferdchen" bestanden: nein | Zeilen-summe |
|---|---|---|---|
| Unterrichts-methode I | 16 | 8 | 24 |
| Unterrichts-methode II | 24 | 1 | 25 |
| Spalten-summe | 40 | 9 | 49 |

$$\textit{Prob} = \frac{24! \cdot 25! \cdot 40! \cdot 9!}{49! \cdot 18! \cdot 6! \cdot 22! \cdot 3!}$$

$$= \frac{40!}{49!} \cdot \frac{25!}{22!} \cdot \frac{24!}{18!} \cdot \frac{9!}{6!} \cdot \frac{1}{3!} = 0.1507$$

$$\textit{Prob} = \frac{24! \cdot 25! \cdot 40! \cdot 9!}{49! \cdot 16! \cdot 8! \cdot 24! \cdot 1!}$$

$$= \frac{40!}{49!} \cdot \frac{25!}{24!} \cdot \frac{24!}{16!} \cdot \frac{9!}{8!} \cdot \frac{1}{1!} = 0.0089$$

- *Als Summe der Wahrscheinlichkeiten in Fisher's exaktem Test erhalten wir $p = 0.289$. Also ist die Nullhypothese $\mathcal{H}_0$ beizubehalten: Die Unterrichtsmethode hat keinen signifikanten Einfluß auf das Bestehen der "Seepferdchen"-Prüfung.*

# Chi-Quadrat- und Fisher's exakter Test: Vergleich

• **Interpretation:** *Beide Tests können sowohl gegen die Homogenität der Verteilungen in allen Gruppen als auch gegen die Unabhängigkeit von zwei kategorialen Variablen durchgeführt werden.*

• **Fehler I. Art:** *Im Unterschied zum $\chi^2$-Test hält Fisher's exakter Test das Niveau für den Fehler I. Art auch für kleines $N$ und kleine $E_{i,j}$ ein — während die Verteilung der Prüfgröße $\widehat{T}$ aus dem $\chi^2$-Test in diesem Fall erheblich von einer $\chi^2$-Verteilung abweicht.*

• **Fehler II. Art:** *Fisher's exakter Test ist "konservativ", d. h. er verschenkt Power (die Wahrscheinlichkeit $\beta$ für ein falsch-negatives Testergebnis ist größer als beim $\chi^2$-Test).*

• **Praktische Umsetzung:** *Fisher's exakter Test ist für Handrechnung ungeeignet. Selbst im Computer entsteht ein erheblicher Aufwand; da sowohl Rechenzeit als auch benötigter Workspace für große Tabellen deutlich anwachsen, bieten Programme wie SPSS Prozeduren an, die den $p$-Wert nicht exakt berechnen, sondern mit Hilfe von Monte-Carlo-Simulationen schätzen.*

# Welcher Test zu welcher Frage?

• **Kategoriale Daten in Vierfeldertafel**

*a)* $N \geqslant 40$ *und alle* $E_{i,j} \geqslant 5$ *— Chi-Quadrat-Test mit einem Freiheitsgrad*

*b)* $N < 40$ *oder mindestens ein* $E_{i,j} < 5$ *— Fisher's exakter Test*

*Nullhypothese: Unabhängigkeit der Variablen / Übereinstimmung der Verteilungen in zwei Populationen*

• **Kategoriale Daten in** $(k \times m)$**-Kontingenztafel** *(*$k > 2$ *oder* $m > 2$*)*

*a)* $N \geqslant 40$ *und* $E_{i,j} \geqslant 5$ *für mindestens 80 % der erwarteten Häufigkeiten*
*— Chi-Quadrat-Test mit df* $= (k-1) \cdot (m-1)$ *Freiheitsgraden*

*b)* $N < 40$ *oder* $E_{i,j} < 5$ *für mehr als 20 % der erwarteten Häufigkeiten*
*— Fisher's exakter Test*

*Nullhypothese: Unabhängigkeit der Variablen / Übereinstimmung der Verteilungen in allen Populationen*

# "Missing values"

• **Problem:** *Im GLM / in der zweifaktoriellen ANOVA mußten wir voraussetzen, daß alle Daten für alle Probanden vollständig sind. In der Praxis entstehen jedoch fast immer Fehlwerte in den Datensätzen.*

• *Der Umgang mit Fehlwerten bei den Kovariaten wird am GLM* $Z \sim c_0 + c_1 \cdot K_1 + ... + c_p \cdot K_p + \varepsilon$ *erläutert. Für die Zielvariable* $Z$ *akzeptieren wir keine Fehlwerte.*

• **Verteilung der Fehlwerte: Fallunterscheidung nach** [ RUBIN 1976 ]

**(MCAR)** *("Missing Completely At Random"): Für jede Kovariate* $K_j$ *ist die Verteilung der Fehlwerte unabhängig von der Zielvariablen und allen Kovariaten (insbesondere unabhängig von* $K_j$ *selbst).*

**(MAR)** *("Missing At Random"): Für jede Kovariate* $K_j$ *hängt die Verteilung der Fehlwerte möglicherweise von der Zielvariable* $Z$ *und allen anderen Kovariaten* $K_m$ *mit* $m \neq j$ *ab, ist aber unabhängig von* $K_j$ *selbst.*

**(MNAR)** *("Missing Not At Random"): Für mindestens eine Kovariate* $K_j$ *führen die Fehlwerte zu einer Verzerrung der tatsächlichen Werteverteilung (die Verteilung der Fehlwerte hängt also von* $K_j$ *selbst ab).*

# Verteilung der Fehlwerte

*An zwei Schulen wird in der Klassenstufe 9 eine Umfrage zu Schul- und Freizeitsport durchgeführt. Erfragt wird u.a. die Teilnote für Bodenturnen im laufenden Halbjahr.*

• **Beispiel 7.8.** *In Schule A fehlt die Note bei 5 von 45 Schülern, in Schule B bei 8 von 63 Schülern. Bewertung: MCAR oder MAR?*

• **Beispiel 7.9.** *In Schule A fehlt die Note bei 7 von 45 Schülern, in Schule B bei 19 von 63 Schülern. Nachfrage: In einer Klasse in Schule B stand die Leistungskontrolle für Bodenturnen noch aus. Bewertung: MCAR oder MAR?*

• **Beispiel 7.10.** *In Schule A fehlt die Note bei 3 von 45 Schülern, in Schule B bei 10 von 63 Schülern, und es gibt keine einzige Note 4 oder 5. Nachfrage: In Schule A wurden die Noten im Fragebogen vom Klassenlehrer ausgefüllt, während in Schule B die Schüler ihre Noten selbst eingetragen haben. Bewertung: starker Verdacht auf MNAR (wurden die schlechten Noten in Schule B einfach ausgelassen?)*

• **Merke:** *Im Rahmen der Statistik kann zwischen MCAR und MAR unterschieden werden, aber nicht zwischen MCAR/MAR und MNAR (wenn überhaupt, dann nur aufgrund genauer Kenntnis der Gesamtsituation).*

# Chi-Quadrat-Test: MCAR oder MAR?

**Beispiel 7.8.:** • *beobachtete Daten*

| | Note vorhanden | Note fehlt | Zeilen-summe |
|---|---|---|---|
| Schule A | 40 | 5 | 45 |
| Schule B | 55 | 8 | 63 |
| Spalten-summe | 95 | 13 | 108 |

• *Indifferenztabelle*

| | Note vorhanden | Note fehlt | Zeilen-summe |
|---|---|---|---|
| Schule A | 39.58 | 5.42 | 45 |
| Schule B | 55.42 | 7.58 | 63 |
| Spalten-summe | 95 | 13 | 108 |

• *Nullhypothese $\mathcal{H}_0$ zum Chi-Quadrat-Test: "Die Verteilungen der Fehlwerte in Schule A und Schule B stimmen überein, also gilt MCAR." / Alternativhypothese $\mathcal{H}_A$: "Die Verteilungen der Fehlwerte in Schule A und Schule B unterscheiden sich, also gilt MAR."*

• *Freiheitsgrade: df $= 1$, Signifikanzniveau: $\alpha = 0.05$, dazu $Q_{95}(\chi^2, 1) = 3.8415$ .*

• *Die Voraussetzungen für den Test (Satz 7.3.) sind erfüllt: $N = 108 > 40$, alle $E_{i,j} \geqslant 5$.*

• *$\widehat{T} = 0.0928$, also ist $\mathcal{H}_0$ beizubehalten, und wir schließen (vorbehaltlich eines Fehlers II. Art) auf MCAR.*

# Strategien zum Umgang mit Fehlwerten

• **Complete Case Analysis bzw. "listwise deletion":** *Im GLM werden nur solche Probanden berücksichtigt, für die alle Werte von $K_1(i)$, ... , $K_p(i)$ vollständig zur Verfügung stehen; alle anderen Zeilen werden gestrichen (Standardverfahren in den meisten Statistik-Softwarepaketen). Möglicher Nachteil: evtl. drastischer Informationsverlust. Voraussetzung: MCAR.*

• **Imputation:** *An allen Fehlstellen $K_j(i)$ werden Werte eingefügt, die nach bestimmten Regeln künstlich erzeugt worden sind. Die Aussagekraft des "vervollständigten" GLM muß abhängig vom Verfahren beurteilt werden. Voraussetzung: MCAR oder MAR.*

• **Mixed linear model:** *Neufassung des GLM, wobei jede Zeile durch eine neue Variable $r_i$ (sog. "random intercept") erweitert wird. Danach darf man alle Einzelterme mit Fehlwerten weglassen. Voraussetzung: MCAR oder MAR.*

**Was tun mit Datenschrott?** *Variablen mit MNAR sind unbrauchbar. Jede statistische Auswertung, in die solche Variablen einbezogen werden, ist wertlos. Das Auftreten von MNAR sollte bereits durch sorgfältige Studienplanung und laufendes Monitoring während der Datenerhebung verhindert werden.*

# Drei Imputationsverfahren

• **Mean Imputation:** *Man setzt für alle Fehlwerte einer metrischen Variablen $K_j$ den Stichproben-Mittelwert $\widehat{\mu}_j$ ein, der aus den vorhandenen Werten gebildet wurde. Bei einer kategorialen Variablen benutzt man den Wert, der unter den vorhandenen Werten am häufigsten vorkommt. Vorteil: einfache Regel / Nachteile: künstlich verringerte Standardabweichung, evtl. Verzerrung linearer Modellgleichungen*

• **Last value carried forward:** *In Datensätzen mit Meßwiederholung ersetzt man jeden Fehlwert durch den letzten vorhandenen Meßwert vom gleichen Probanden. Vorteil: einfache Regel / Nachteil: evtl. Verzerrung linearer Modellgleichungen sogar unter MCAR*

• **Multiple Imputation:** *Man erstellt typischerweise 3 bis 5 Datensätze, in denen alle Fehlwerte nach speziellen Regeln ersetzt werden. Die Ersetzungen in den verschiedenen Datensätzen unterscheiden sich geringfügig. Dann wertet man das GLM für jeden Datensatz separat aus und poolt die Ergebnisse (siehe* [ VAN BUUREN 2018 ]*). Vorteil: statistisch korrektes Verhalten der Ergebnisse / Nachteil: komplizierte Berechnung der künstlichen Werte*

# Zum Vergleich: GLM ...

• **Variablen:** *metrische Zielvariable $Z$ ("abhängige Variable"), Kovariaten $K_1$, ... , $K_p$ ("unabhängige Variablen") und Residuum $\varepsilon$ mit jeweils $N$ Beobachtungen.*

• **Lineare Modellgleichung:** $Z \sim c_0 + c_1 \cdot K_1 + \ ... \ + c_p \cdot K_p + \varepsilon$ .

• *Die Residuen $\varepsilon(i)$ werden als (annähernd) normalverteilt mit $\mu = 0$ und $\sigma > 0$ und paarweise unabhängig vorausgesetzt.*

• **Ermittlung der Koeffizienten** $c_0$, $c_1$, ... , $c_p$ *als diejenigen Lösungen des linearen Gleichungssystems*

$$
\begin{aligned}
Z(1) &= c_0 + c_1 \cdot K_1(1) + \ ... \ + c_p \cdot K_p(1) + \varepsilon(1)\,, \\
Z(2) &= c_0 + c_1 \cdot K_1(2) + \ ... \ + c_p \cdot K_p(2) + \varepsilon(2)\,, \\
Z(3) &= c_0 + c_1 \cdot K_1(3) + \ ... \ + c_p \cdot K_p(3) + \varepsilon(3)\,, \\
&\ \ \vdots \qquad\qquad\qquad\qquad\qquad\quad \vdots
\end{aligned}
$$

*welche die Bedingung $\varepsilon(1)^2 + \varepsilon(2)^2 + \ ... \ + \varepsilon(N)^2 = \mathrm{Min}\,!$ erfüllen. Wegen $N > p+1$ ist das Gleichungssystem überbestimmt.*

# ... und "mixed linear model"

• **Variablen:** *metrische Zielvariable* $Z$, *Kovariaten* $K_1$, ... , $K_p$, *Residuum* $\varepsilon$ *mit jeweils* $N$ *Beobachtungen; dazu in jeder Zeile ein "random intercept"* $r_i$ .

• **Erweiterte lineare Modellgleichung:** $Z \sim c_0 + c_1 \cdot K_1 + \ ... \ + c_p \cdot K_p + \varepsilon + r$ .

• *Die Residuen* $\varepsilon(i)$ *und die "random intercepts"* $r_i$ *seien (annähernd) normalverteilt mit* $\mu = 0$, *positiven Standardabweichungen und jeweils paarweise unabhängig.*

• **Ermittlung der Koeffizienten** $c_0$, $c_1$, ... , $c_p$ *sowie* $r_1$, $r_2$, ... , $r_N$ *als diejenigen Lösungen des linearen Gleichungssystems*

$$
\begin{aligned}
Z(1) &= c_0 + c_1 \cdot K_1(1) + \ ... \ + c_p \cdot K_p(1) + \varepsilon(1) + r_1 \,, \\
Z(2) &= c_0 + c_1 \cdot K_1(2) + \ ... \ + c_p \cdot K_p(2) + \varepsilon(2) \quad + r_2 \,, \\
Z(3) &= c_0 + c_1 \cdot K_1(3) + \ ... \ + c_p \cdot K_p(3) + \varepsilon(3) \quad + r_3 \,, \\
&\ \vdots \qquad\qquad\qquad\qquad\qquad\qquad \vdots
\end{aligned}
$$

*welche die Bedingung* $\varepsilon(1)^2 + \varepsilon(2)^2 + \ ... \ + \varepsilon(N)^2 = \mathrm{Min}\,!$ *erfüllen. Wegen* $N < N + p + 1$ *ist das lineare Gleichungssystem jetzt sogar unterbestimmt, und wir dürfen die Zusatzbedingung* $(r_1)^2 + (r_2)^2 + \ ... \ + (r_N)^2 = \mathrm{Min}\,!$ *für die* $r_i$ *stellen.*