„Datenmanagement – Teil II: Datenscreening und -transformation“

Dipl.-Psych. W. Igl & Dipl.-Psych. A. Reusch
(Methodenberatung)

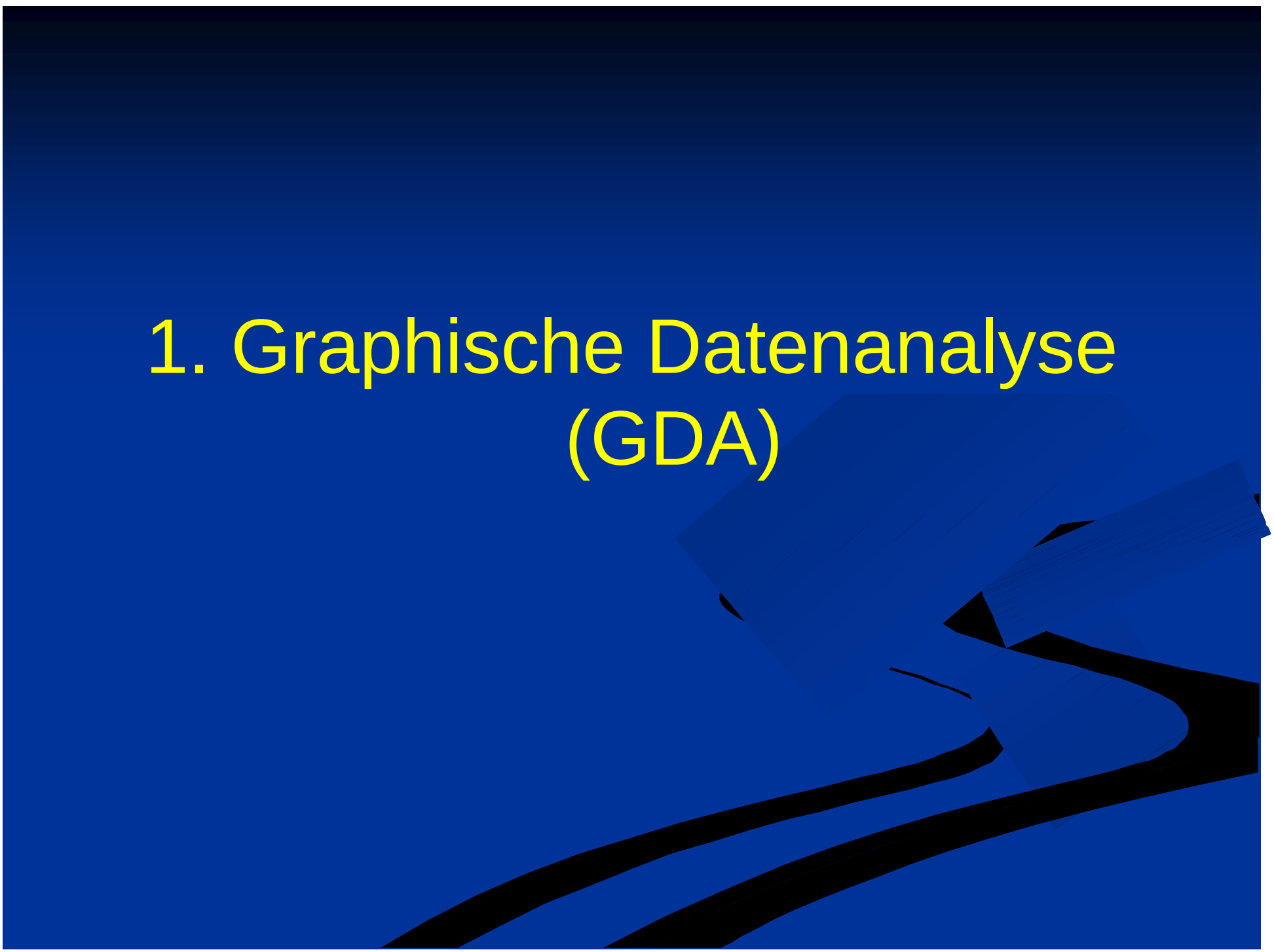
Rehabilitationswissenschaftlicher
Forschungsverbund
Bayern (RFB)

Referent: Dipl.-Psych. Wilmar Igl

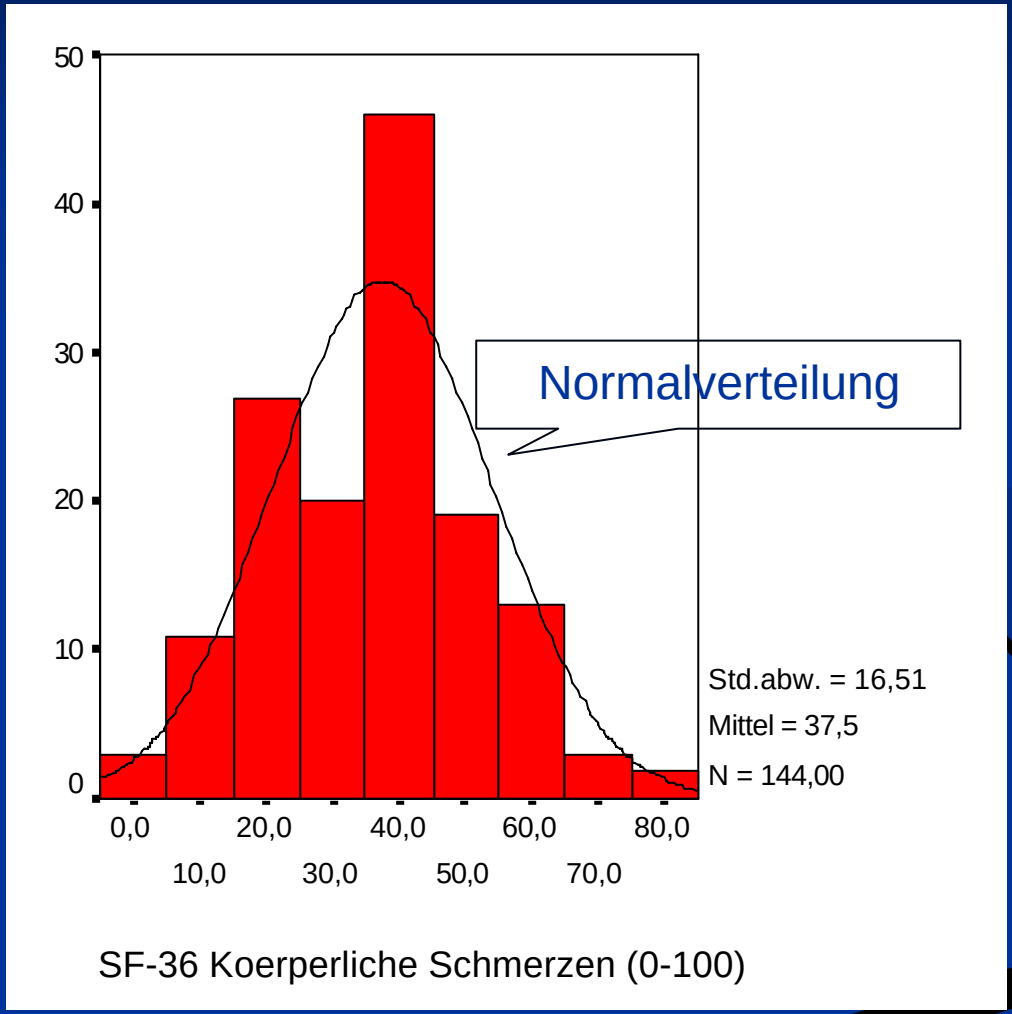
Einleitung

- Zusammenhang von Qualität der Daten und Aussagekraft der statistischen Ergebnisse („garbage in, garbage out phenomenon“)
- Data Screening („Datensichtung“):
  Untersuchung der erhobenen Daten auf mögliche Verzerrungen und (vorsichtige) Behebung dieser Verzerrungen zur Steigerung der Aussagekraft der Daten
- Wichtige (sich wiederholende) Schritte:
  1. Graphische Datenanalyse
  2. Analyse und Behandlung von Ausreißern
  3. Analyse und Behandlung von Fehlwerten (missing data = MD)
  4. Datentransformationen zur Korrektur von Verletzungen der statistischen Voraussetzungen

1. Graphische Datenanalyse (GDA)



GDA - Histogramm



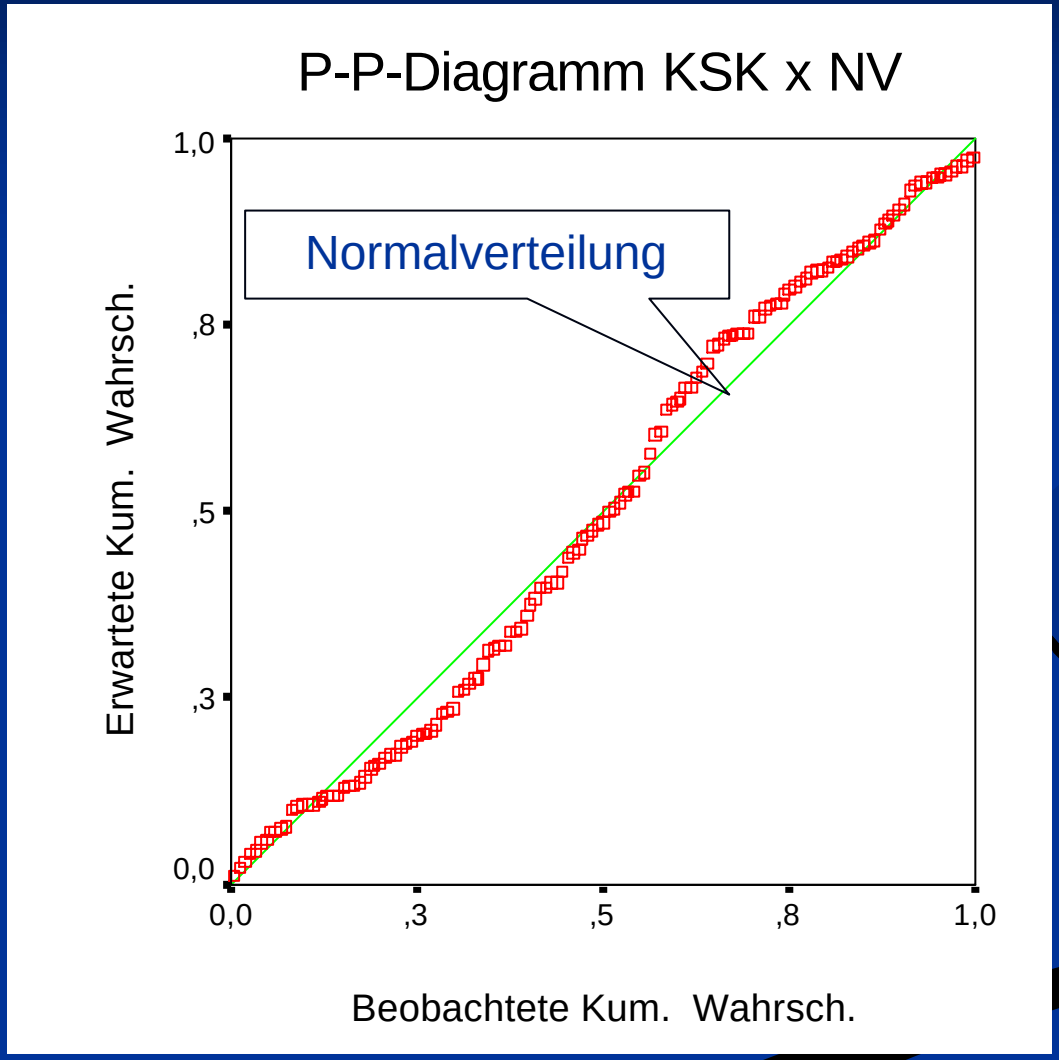
SF-36 Koerperliche Schmerzen (0-100)

SPSS-MENÜ
> Grafiken > Histogramm
(+ NV anzeigen)

Anwendung:

- Beurteilung der (Normal)Verteilung
- Erkennen von Ausreißern
- Hinweise auf geeignete Datentransformationen
- Verzerrung durch Breite der Intervalle möglich

GDA – P-P-Diagramm



SPSS-MENÜ

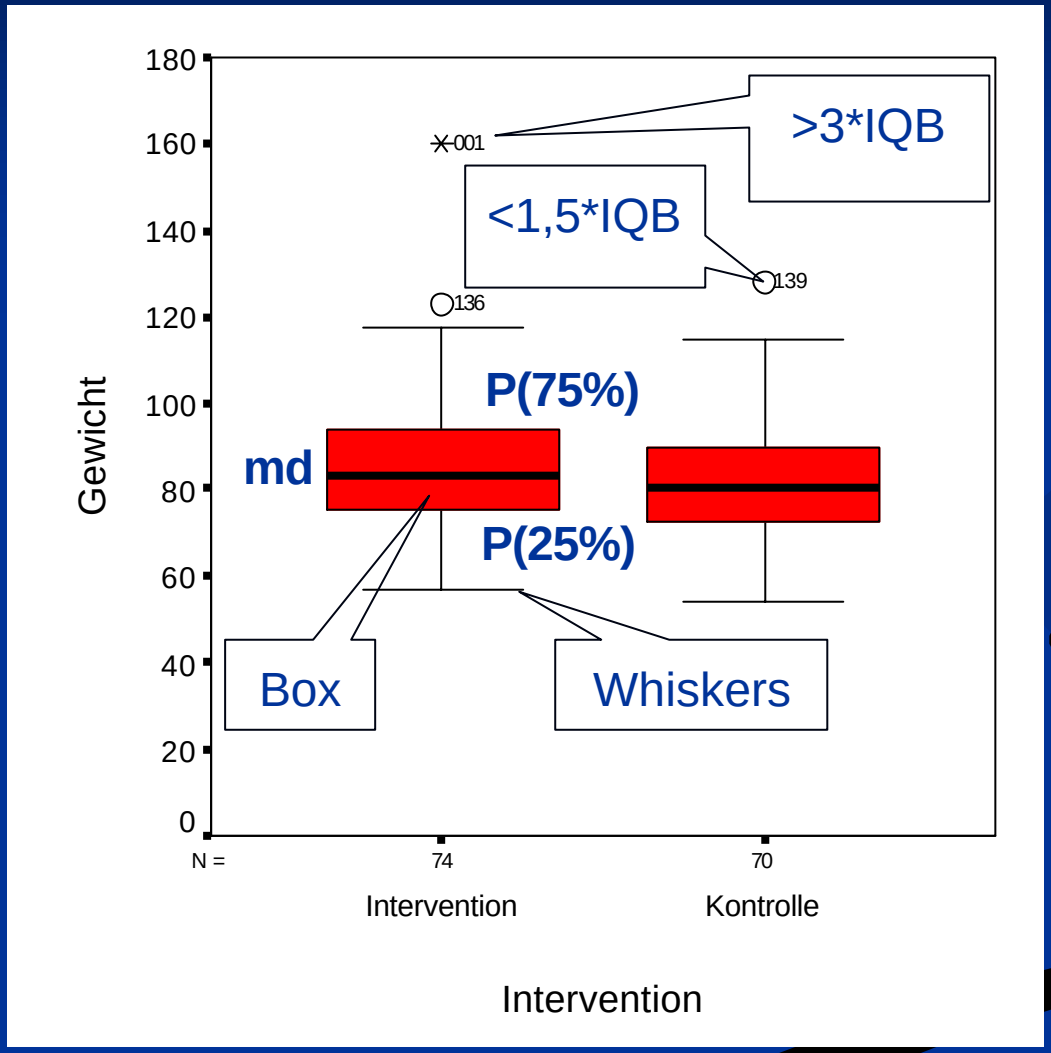
> Grafiken > P-P-...

Anwendung:

Beurteilung der (Normal-) Verteilung möglich bzgl.

- Kurtosis („Gipfligkeit“)
- Schiefe

GDA – Box-and-Whisker-Plot



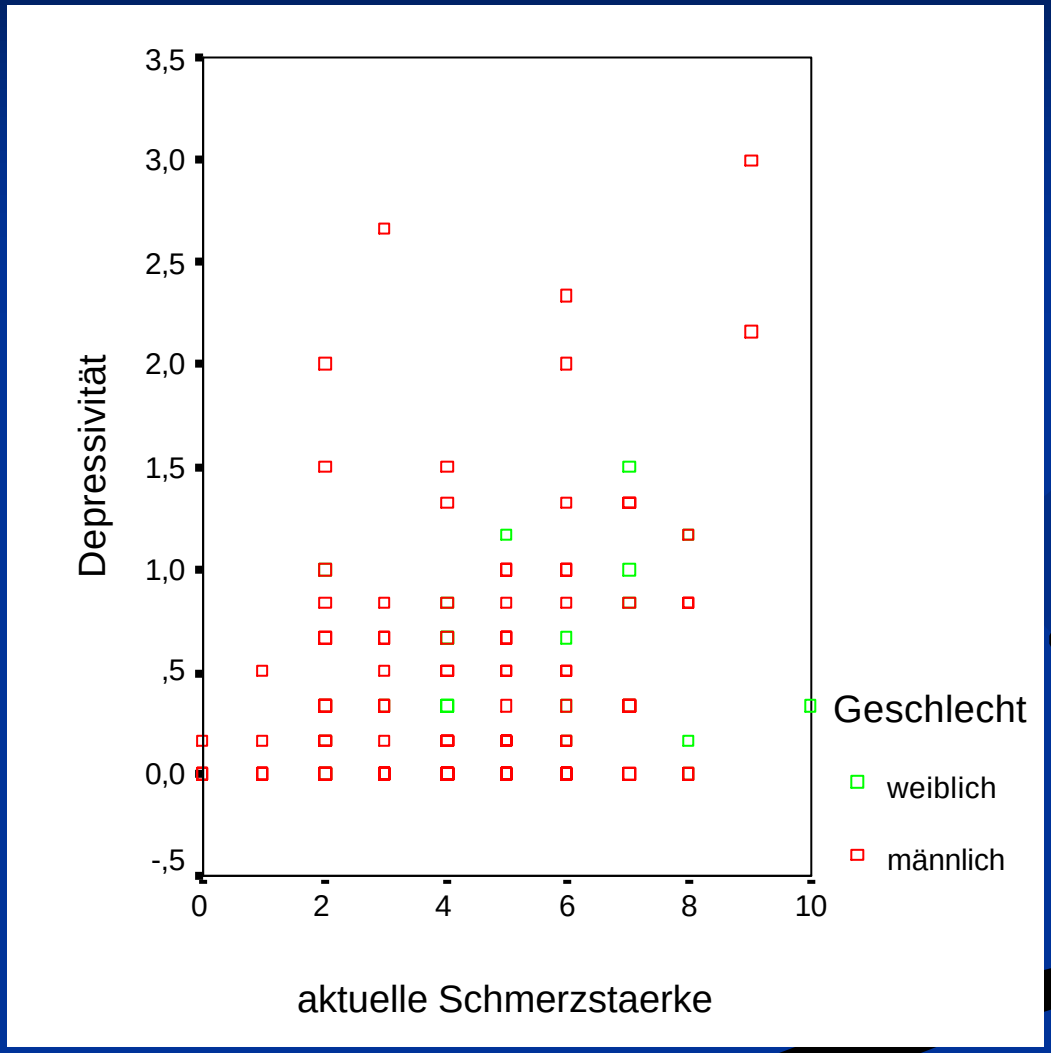
SPSS-MENÜ

> Grafiken > Box-Plot

Anwendung:

- Erkennen von Ausreißern
- Beurteilen von Lageunterschieden
- Beurteilen der Verteilung

GDA – Scatterplots (Streudiagramm)



SPSS-MENÜ

> Grafiken

> Streudiagramm

Anwendung:

- Analyse von Zusammenhängen
- Beurteilung von Linearität
- Erkennen von einflussreichen Werten (influential values)
- Erkennen von Ausreißern

2. Ausreißer

Ausreißer (1)

- Def. „Ausreißer“:

  „... observations with a unique combination of characteristics identifiable as distinctly different from the other observations.“
  (Hair, 1998)

- Ausreißer als ...
  - Fehler

    => Aufblähung der Fehlervarianz, Verzerrung der Ergebnisse
  - bedeutsames Ereignis

    => Generalisierbarkeit, Hinweise auf Wechselwirkungen, bedeutsamer Indikator (z.B. erhöhte Selbstmordrate als Hinweis auf erhöhte Depressivität einer Population)

Ausreißer (2)

- Faustregeln (univariat):
  - n <= 80: Werte größer als +-2.5 * sd
  - n > 80: Werte größer als +-3 bis +-4 * sd

- Behandlung:
  - Wenn Ausreißer repräsentativ/ valide sind für die Stichprobe, dann Behalten
  - Wenn Ausreißer nicht repräsentativ sind für die Stichprobe, dann Löschen
  - alternativ: stabilere Statistiken verwenden (z.B. Median statt Mittelwert, Kendall‘s Tau statt Spearman‘s Rho, non-parametrische statt parametrische Methoden)

3. missing data

missing data

- Definition: „...missing data liegt vor, wenn Werte wider Erwarten in der Datenmatrix fehlen.“ (Müller, 2002)

- Verzerrungen der Ergebnisse und
  Verringerung der Effizienz von statistischen Verfahren möglich

- Behandlung von MD abhängig von der Art des
  missing-data-Prozesses (Systematik?)
  - Non-Random Missing (NRM)
  - Missing At Random (MAR)
  - Missing Completely At Random (MCAR)

MAR - Missing At Random

- Statistische Bedeutung:
  Fehlwerte der Variable Y hängen nicht von den Werten der Variable Y ab, sondern von einer anderen Variable X.
- Veranschaulichung:
  Datenmatrix als Leinentuch, das von Schrotkugeln durchsiebt wird
  => Die Löcher bilden Linien oder Rechtecke
  - waagrechte Linie: Hinweise patientenbezogene Ursachen
  - senkrechte Linie: Hinweis auf itembezogene Ursachen
  - Rechtecke: Hinweise auf klinikmitarbeiterbezogene Ursachen
- Beispiel:
  Angaben zum „Einkommen“ (Y) fehlen unabhängig von der Höhe des Einkommens (Y) , aber hängen trotzdem vom „Geschlecht“ (X) ab

MCAR – Missing Completely At Random

- Statistische Bedeutung:
  Fehlwerte von Y weisen keinen Zusammenhang mit einer anderen Variable auf
- Veranschaulichung:
  Datenmatrix als Leinentuch, das aus Fäden unterschiedlicher Dicke besteht; Fäden unterschiedliche Dicke werden gleich häufig getroffen
- Beispiel:
  Angaben zum „Einkommen“ (Y) fehlen unabhängig von der Höhe des Einkommens (Y) oder anderen Merkmalen wie „Geschlecht“ (X1), Alter (X2),... ab

Diagnose des missing-data-Prozesses (1)

- Screening der Datenmatrix (Fälle X Variablen):
  Häufige missings bei Fällen oder Variablen oder Kombinationen von Fällen und Variablen?

- Auswertung basaler Statistiken:
  - gültige Werte: [f], [%], m, sd
  - fehlende Werte: [f], [%], m, sd

Diagnose des missing-data-Prozesses (2)

- Erstellung einer Indikatormatrix:
  Codierung von gültigen Werten mit 1 und von fehlenden Werten mit 0 zur Bildung von Gruppen

- Analyse von Gruppenunterschieden in anderen Variablen

- Analyse von Korrelationen in der Indikatormatrix

- Bei signifikanten Unterschieden/ Zusammenhängen kann man nicht mehr von MCAR ausgehen.

=> Welche Maßnahmen können ergriffen werden?

Maßnahmen - Listenweiser Fallausschluss

- Vorgehen: Ausschluss aller unvollständiger Fälle/ Variablen („complete information approach“)

- Anwendung bei :
  - MCAR
  - große Stichprobe
  - starke Effekte

- Nachteile: Reduktion der Stichprobe bis zur Unbrauchbarkeit möglich

Maßnahmen - Imputationsverfahren

- Definition:
  Verfahren, durch das Fehlwerte geschätzt und ersetzt werden

- Vorgehen: Schätzen von fehlenden Werten basierend auf den validen Werten von anderen Variablen /Fällen in der Stichprobe.

- Anwendung bei:
  - MCAR
  - intervallskalierten/metrischen Variablen

Imputation – Paarweiser Fallausschluss

- Vorgehen: Alle gültigen Fälle, der in die Berechnung eingehenden Variablen, werden ausgewertet. Übernehmen der Verteilungs-charakteristika der gültigen Werte (“all available approach”)
- Anwendung bei:
  - MCAR
  - Berechnung von Korrelationen, Mittelwerten, Standardabw.
- Nachteile:
  - Statistiken können auf unterschiedlichen Stichproben von Beobachtungen basieren (unterschiedliches N !)
  - mathematische Inkonsistenzen möglich (z.B. zwischen Korrelationen zweier Variablen X, Y und deren Partialkorrelationen mit Z)

Imputation - Regression

- Verfahren: Schätzen und Ersetzen von Fehlwerten durch (multiple) Regression unter Anwendung bekannter Beziehungen zwischen Variablen
- Anwendung bei:
  - Vorliegen substantieller Zusammenhänge mit anderen Variablen
  - mäßiger Grad von weit verstreuten missing data
- Nachteile:
  - Unterschätzung der Varianz
  - Verstärkung (Verzerrung) bestehender Zusammenhänge
  - Werte ausserhalb des Wertebereichs möglich
  - geringere Generalisierbarkeit

Weitere Imputationsverfahren

- Mittelwertersetzung: Ersetzen des fehlenden Wertes durch Mittelwert der gültigen Werte
  - Vorteile: einfach, vollständiger Datensatz
  - Nachteile: Verzerrung der wahren Verteilung, Unterschätzung der wahren Varianz, Unterschätzung der wahren Zusammenhänge

- Ersetzung aus externem Datensatz: Ersetzen des fehlenden Wertes durch Werte einer externen Quelle/ frühere Forschung, die valider ist als die untersuchte Stichprobe

- Fallersetzung: Ersetzen des fehlenden Wertes durch andere, neue (ähnliche) Beobachtung (Neurekrutierung)

missing data - Fazit

1. Diagnose des vorliegenden missing-data-Prozesses
2. Rationale, auf theoretischen Überlegungen und empirischen Fakten beruhende Auswahl eines Verfahrens
3. Vergleich der Effekte anderer Verfahren
4. (ggf. begründete Auswahl eines anderen Verfahrens)

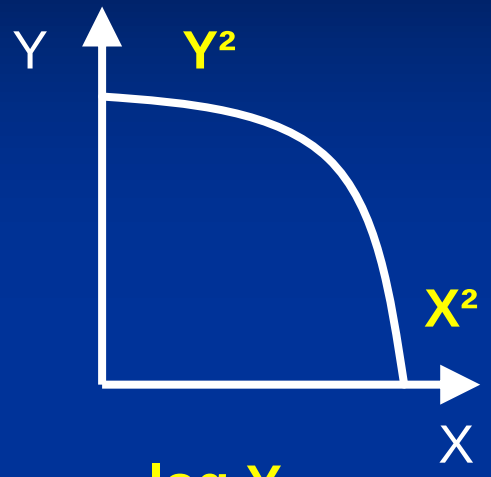
=>„begründete Entscheidung“

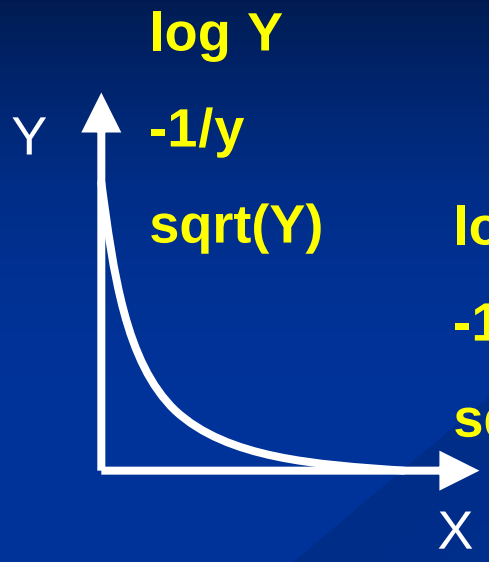
4. Datentransformationen

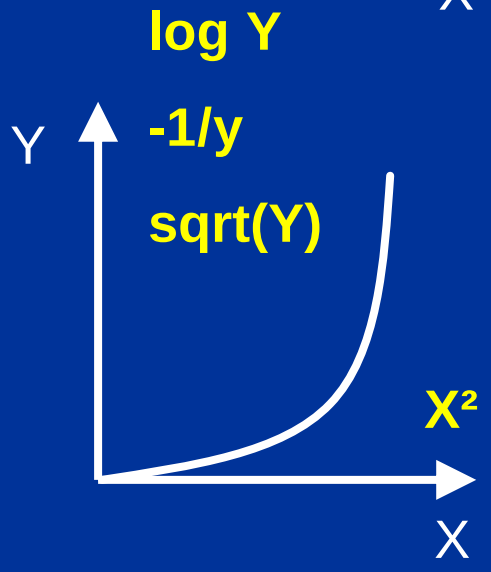
Normalität u. Heteroskedaszität - Transformationen

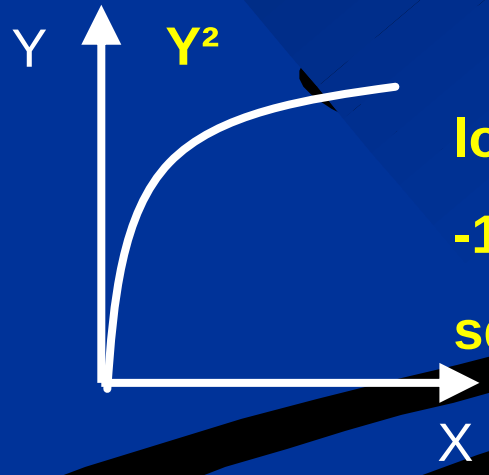
- „Flache“ Verteilungen: $y = 1/x$
- „Schiefe“ Verteilungen
  - $y = \sqrt{x}$ (bei rechtssteilen Verteilungen geeignet)
  - $y = \log(x)$ (bei linkssteilen Verteilungen geeignet)
  - $y = 1/x$
- falls noch Heteroskedaszität vorliegt: $y = 2 \arcsin \sqrt{x}$
- Auswahl der Transformation nach bestem Ergebnis

Linearität – Transformationen









nach Hair et al. (1998)

Leitlinien für Datentransformationen

1. $m(x)/sd(x) < 4$

2. Bei Auswahl zwischen zwei Variablen, wähle die mit dem kleinsten Quotienten aus 1)

3. Transformationen sollten nur auf unabhängige Variablen angewendet werden.

4. Heteroskedaszität kann nur durch Transformation verringert werden.

5. Die Interpretation transformierter Variablen kann sich ändern.

Literatur

- Bland, M. (2000). An Introduction to Medical Statistics (3rd edition). Oxford: University Press
- Bühl, A. & Zöfel, P. (1998). SPSS für Windows Version 7.5. Bonn: Addison-Wesley
- Diehl, J. M. & Staufenbiehl, T. (2001). Statistik mit SPSS Version 10.0 (1. Auflage). Eschborn: Klotz
- Hair, J. F., Anderson, R. E., Tatham, R. L., Black, W. C. (1998). Multivariate data analysis. 5. Auflage. New Jersey: Prentice Hall.
- Reusch, A., Zwingmann, Ch., Faller, H. (Hrsg.) (2002). Empfehlungen zum Umgang mit Daten in der Rehabilitationsforschung. Regensburg: Roderer Wilkinson, L. & The Task Force on Statistical Inference (1999). Statistical Methods in Psychology Journals – Guidelines and Explanations. American Psychologist, Vol. 54,594-604
- Wirtz, M. Umgang mit fehlenden Werten (Vortrag). Methodenzentrum des Rehabilitationswissenschaftlichen Forschungsverbundes Freiburg/Bad Säckingen

Vielen Dank für

Ihre Aufmerksamkeit!

Kontakt: wilmar.igl@mail.uni-wuerzburg.de