Multivariate Statistik, Projektplanung und Versuchsdesign

Vorlesung
Modul M 103 (Vorl.-Nr. 28206)

Dr. G. Lischeid

Gunnar.Lischeid@bayceer.uni-bayreuth.de
http://www.bayceer.uni-bayreuth.de/mod/

Gliederung der Vorlesung

15.04.05	Einführung, Verteilungen	
22.04.05	Datentransformation	▷
29.04.05	(Auto-)Korrelation *(zu verschieben)*	▷
06.05.05	Multiple lineare Regression	▷
13.05.05	Hauptkomponentenanalyse	▷
20.05.05	*(Pfingstwoche)*	
27.05.05	Korrespondenzanalyse	▷
03.06.05	Clusteranalyse	▷
10.06.05	Diskriminanzanalyse	▷
17.06.05	Grundlagen der Versuchsplanung	▷
24.06.05	Mehrfaktorielle Versuche	▷
01.07.05	Parameter-freie Methoden	▷
08.07.05	Nicht-lineare Methoden	
15.07.05	Abschlusskolloquium	

Links und Lehrbücher

http://www.multivariate.de

http://wwwhomes.uni-bielefeld.de/hjawww/glossar/stichwor.htm

Backhaus, K., Erichson, B., Plinke, W., Weiber, R. (2003): Multivariate Analysemethoden. Eine anwendungsorientierte Einführung. 10. Auflage, Springer, ISBN 3-540-00491-2

Bortz, J. (1993): Statistik für Sozialwissenschaftler. 4. Auflage, Springer, ISBN 3-540-56200-1

Wackernagel, H. (1998): Multivariate Geostatistics. 2. Auflage, Springer

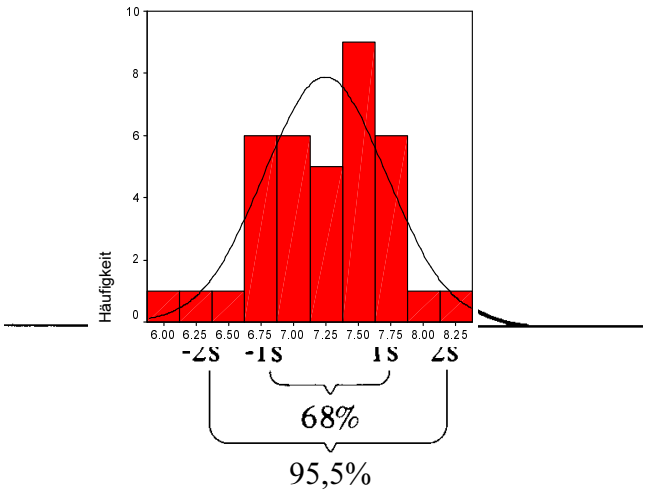
Statistik - Zielrichtungen

1. ***Deskriptiv*** - Beschreibung (der Verteilung) eines Datensatzes
2. ***Konfirmativ*** - Testen von Hypothesen (Zusammenhängen)
3. ***Explorativ*** - Suche nach Strukturen (Zusammenhängen)

Datentypen

1. ***Nominal skaliert*** - Zugehörigkeit zu einer Gruppe (***Name***)
2. ***Ordinal skaliert*** - Rangfolge (***Ordnung***)
3. ***Intervall-skaliert*** - Abstände der Zahlenwerte proportional der Abstände der Merkmalsausprägung
4. ***Metrisch skaliert*** - Zahlenwerte propotional der Merkmalsausprägung (***Maß***)

Dichtefunktion der Normalverteilung



Verteilungen

Normalverteilung: Approximaton der Binomialverteilung für große Stichproben

für eine normalverteilte Zufallsgröße ***z*** *(mit μ = 0 und σ = 1) gilt:*

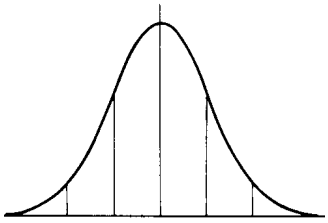
χ2-Verteilung: für $\chi^2 = z^2$, bzw. für ***n*** Freiheitsgrade: $\chi_n^2 = \sum_n z^2$

t-Verteilung: $t_n = \frac{z}{\sqrt{\frac{\chi_n^2}{n}}}$ mit $\sigma = \sqrt{\frac{n}{n-2}}$

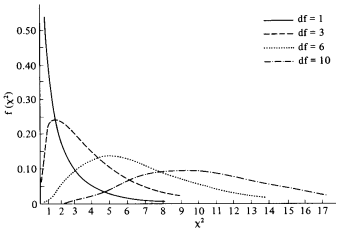
F-Verteilung: $F(n_1, n_2) = \frac{\chi_{n_1}^2}{\chi_{n_2}^2} \cdot \frac{n_2}{n_1}$

Verteilungen

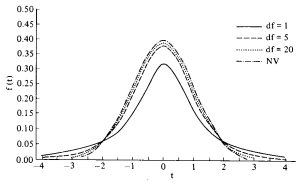
Normalverteilung



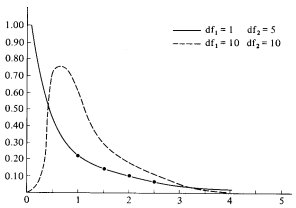
χ2-Verteilung



t-Verteilung

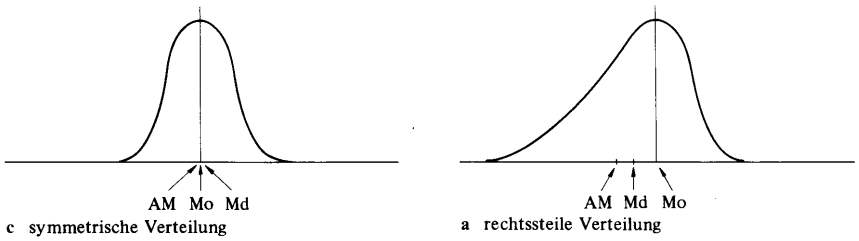


F-Verteilung:



Schiefe, Exzess

AM: Arithmtisches Mittel
Mo: Modalwert
Md: Median



(Bortz 1993)

Momente (I)

k-tes Moment der Variablen ***x***: $\mu_k(x, A) = E[(x - A)^k]$

(= Moment ***k***-ter Ordnung)

gewöhnliches Moment: $A = 0$

Mittelwert: $A = 0; \quad k = 1: \quad \mu = E(x)$

zentrales Moment: $A = E(x)$

Varianz: $A = \mu; \quad k = 2: \quad \sigma^2 = E(x - \mu)^2$

Momente (II)

Normiertes Moment ***k***-ter Ordnung: $\mu_k(x, A) = \frac{E[(x-A)^k]}{\sqrt{\sigma^k}}$

für $z = \frac{x - E(x)}{\sqrt{\sigma^2}}$ (*z-Transformation*):

Schiefe: $\gamma_3 = \frac{\mu_3}{\sqrt{\sigma^3}} = \frac{\sum_{i=1}^{n} z_i^3}{n}$

(Werte <0: negative Schiefe => rechtssteil = linksschief)

Exzess: $\gamma_4 - 3 = \frac{\mu_4}{\sqrt{\sigma^4}} - 3 = \frac{\sum_{i}^{n} z_i^4}{n} - 3$ (= *Kurtosis*, *Wölbung*)

(Werte < 0: breitgipflige Verteilung)

Kovarianz, Korrelation (I)

Varianz: $\mathrm{var}(x) = \frac{\sum_{i=1}^{n}(x-\bar{x})^2}{n} = \frac{\sum_{i=1}^{n}(x-\bar{x})\cdot(x-\bar{x})}{n}$

Kovarianz: $\mathrm{cov}(x, y) = \frac{\sum_{i=1}^{n}(x_i-\bar{x})\cdot(y_i-\bar{y})}{n}$

Korrelation: $r = \frac{\mathrm{cov}(x, y)}{\sqrt{\mathrm{var}(x)}\cdot\sqrt{\mathrm{var}(y)}} = \frac{\mathrm{cov}(x, y)}{s_x \cdot s_y}$

maximal mögliche Kovarianz

(*Produkt-Moment-Korrelation, Pearson-Korrelation*)

Kovarianz, Korrelation (II)

Korrelation:

$$r = \frac{\mathrm{cov}(x,y)}{s_x \cdot s_y}$$

$$= \frac{1}{n} \cdot \frac{\sum_{i=1}^{n} (x_i - \bar{x}) \cdot (y_i - \bar{y})}{s_x \cdot s_y}$$

$$= \frac{1}{n} \cdot \sum_{i=1}^{n} \left[\frac{x_i - \bar{x}}{s_x} \cdot \frac{y_i - \bar{y}}{s_y} \right]$$

z-Transformation

z-Transformation

Unterschiedliche Wertebereiche für unterschiedliche Parameter

=> unterschiedliche Gewichtung in der multivariaten Analyse, die nur vom Wertebereich (Einheit!) abhängig ist

=> Notwendigkeit der Normierung

=> analog zur Korrelation:

1. Normierung auf ***Mittelwert = 0***, d.h.: *Substraktion des Mittelwertes*
2. Normierung auf ***Varianz = 1*** (= Standardabweichung), d.h.: *Division durch die Standardabweichung*

Produkt-Moment-Korrelation

Moment: $$\mu_k(x, A) = \frac{E[(x - A)^k]}{\sqrt{\sigma^k}}$$

1. zentrales Moment: $$\mu_1(x, \bar{x}) = \frac{E[(x - \bar{x})^1]}{\sqrt{\sigma^1}}$$

1. Produkt-Moment zweier Zufallsvariabler:

$$\mu_1(x, A) = \frac{E\left[([x - \bar{x}] \cdot [y - \bar{y}])^1\right]}{\sqrt{\mathrm{var}_x} \cdot \sqrt{\mathrm{var}_y}}$$

Nichtlineare Korrelation

Spearman-Rangkorrelationskoeffizient:

(D_i: Rangplatzdifferenzen)

$$r_R = 1 - \frac{6 \cdot \sum D_i^2}{n^3 - n}$$

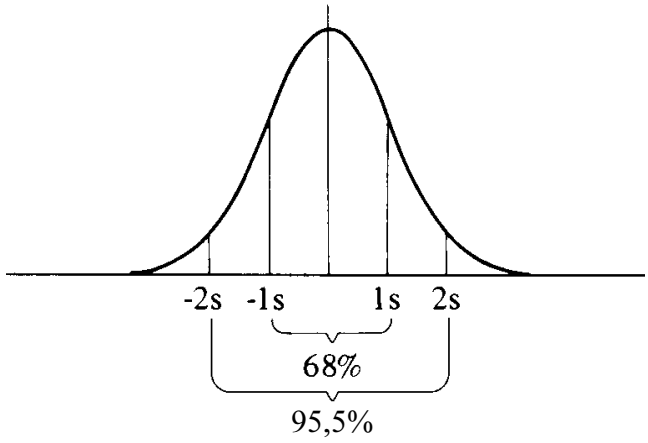
Kendall's τ:

(*Ko*: Konkordanzen = gleichsinnige Änderungen $x_1 \rightarrow x_2$ und $y_1 \rightarrow y_2$,

Di: Diskordanzen)

$$r_K = \frac{2 \cdot (Ko - Di)}{n \cdot (n - 1)}$$

Dichtefunktion der Normalverteilung



Test auf Normalverteilung (I)

X^2-Test:

- Vergleich der beobachteten Häufigkeit f_i von k Klassen (mit $f_i > 10$) mit den erwarteten Häufigkeiten e_i (mit $e_i \geq 1$ und $e_i < 5$ für max. 20% der Klassen)
- Testgröße: $\chi^2 = \sum_{i=1}^{k} \frac{(f_i - e_i)^2}{e_i}$ verteilt nach X^2 mit (k-r-1) Freiheitsgraden (r = Anzahl der geschätzten Parameter der Verteilung)
- Verwerfen der Null-Hypothese "F = Normalverteilung" für $p \leq \alpha$

Fehler 1. und 2. Art
(α-, β-Fehler)

	in Grundgesamtheit gilt	
	H_0	H_1
Entscheidung aufgrund	richtig	**β-Fehler**
der Stichprobe	**α-Fehler**	richtig

Nullhypothese H_0 = Alternative zur eigentlich zu prüfenden Hypothese H_1

Irrtumswahrscheinlichkeit *p*

α*-Fehler**: "Irrtumswahrscheinlichkeit***" ***p*** = Wahrscheinlichkeit, einen bestimmten Wert zu beobachten, wenn tatsächlich die H_0 gilt:

$$p(\textit{beobachteten Wert} \mid H_0 = \textit{wahr})$$

= ***bedingte*** Wahrscheinlichkeit

***β* -Fehler**: $p(\textit{beobachteten Wert} \mid H_1 = \textit{wahr})$

=> quantitativ nur zu bestimmen, wenn die Verteilung der Werte gemäß der H_1-Hypothese ***à priori*** bekannt ist

=> dies ist aber i.d.R. nicht möglich
entsprechend für Test auf Normalverteilung: alternative Verteilung müsste definiert werden

Irrtumswahrscheinlichkeit *p*

Die "***Irrtumswahrscheinlichkeit*** **p**

$\boldsymbol{p}(\textit{beobachteter Wert} \mid H_0 = \textit{wahr})$

Unterscheide:

$\boldsymbol{p}(\textit{beobachteter Wert})$	$\boldsymbol{p}(H_0 = \text{wahr})$
$1 - \boldsymbol{p}(H_0)$	$\boldsymbol{p}(H_0 = \textit{wahr} \mid \textit{beobachteter Wert})$

Fehler 1. und 2. Art
(α-, β-Fehler)

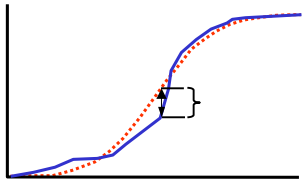
	in Grundgesamtheit gilt	
	H_0	**H_1**
Entscheidung aufgrund	richtig	**β-Fehler**
der Stichprobe	**α-Fehler**	richtig

Nullhypothese H_0 = Alternative zur eigentlich zu prüfenden Hypothese H_1

Test auf Normalverteilung (II)

Kolmogorov-Smirnov mit Lilliefors-Schranken:

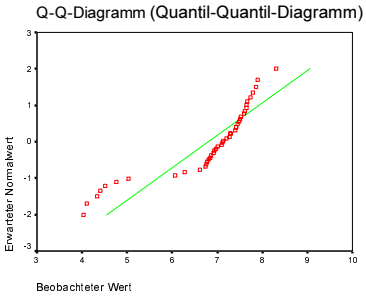
- Testgröße: maximale absolute Abweichung der Ordinatenabstände zwischen der beobachteten und der erwarteten kumulierten Häufigkeitsverteilung
- Verwerfen der Null-Hypothese "F = Normalverteilung" für $p < \alpha$



Test auf Normalverteilung (III)

Anpassungstest nach **Shapiro und Wilk** (für $n \leq 50$):

- Testgröße: Korrelationskoeffizient zwischen beobachteten und erwarteteten Werten der kumulativen Häufigkeitsverteilungen
- Verwerfen der Null-Hypothese "F = Normalverteilung" für $p < \alpha$

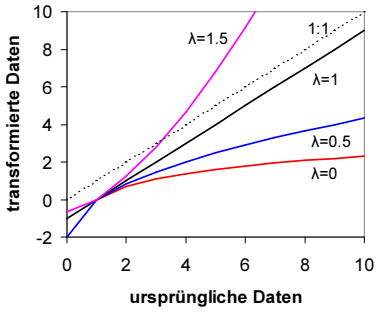


Box-Cox Transformation

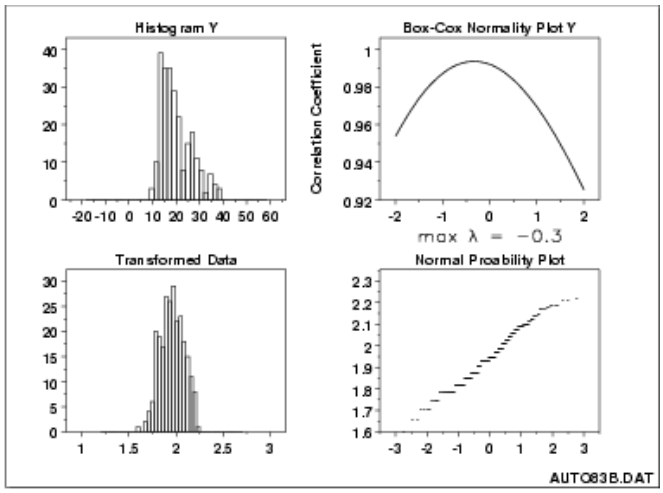
Ziel: Korrektur der Schiefe, so dass die transformierten Daten eine Normalverteilung aufweisen

$$T(x) = \frac{x^{\lambda} - 1}{\lambda} \qquad \textit{für} \quad \lambda > 0$$

$$T(x) = \ln x \qquad \textit{für} \quad \lambda = 0$$



Box-Cox Transformation



(http://www.itl.nist.gov/div898/handbook/eda/section3/eda336.htm)

Aufgabe: Datentransformation

- Ersetzen Sie die Einträge "< *Bestimmungsgrenze*" durch sinnvolle nummerische Werte.
- Überrpüfen Sie die einzelnen Parameter auf Normalverteilung, und führen Sie, falls erforderlich, eine Box-Cox-Transformation durch. Überprüfen Sie anschließend die transformierten Daten auf Normalverteilung.
- Führen Sie anschließend eine z-Transformation für alle Parameter durch.