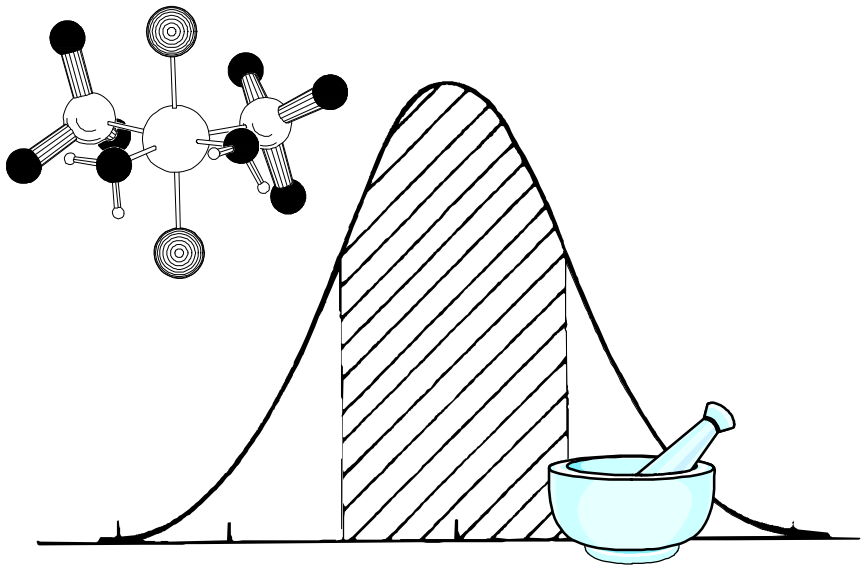
Ralph Puchta

Chemometrie

eine Vorlesung von Professor D. K. Breitinger



Inhaltsverzeichnis

0. Einführung ***3***

1. Eichung ***6***

2. Messung ***7***

3. Statistische Behandlung von sogenannten „Ausreißern“ (outliners) ***10***

3.1 Q-Test **10**

3.2 Grubb’s - Test **10**

4. Vergleich von Datensätzen ***12***

4.1 Student-t-Test **12**

4. 2 Fischers F-Test **13**

4.3 „Faust“-Test **15**

5. Fehlerfortpflanzung ***16***

6. Datenanalyse ***20***

6.1 Univariate Datenanalyse/Modellierung - lineare Regression **20**

6.2 Multivariate Datenanalyse/Modellierung **24**

6.3 Mustererkennung **27**

7. Anhang ***32***

Chemometrie (Chemometrik)

0. Einführung

Unter Chemometrik versteht man die chemische Teildisziplin, die sich mit der Anwendung mathematischer und statistischer Methoden beschäftigt, um in optimaler Weise chemische Verfahren und Experimente zu planen, zu entwickeln, auszuwählen oder so auszuwerten, daß ein Maximum an chemischen Informationen aus experimentellen Meßdaten extrahiert werden kann.

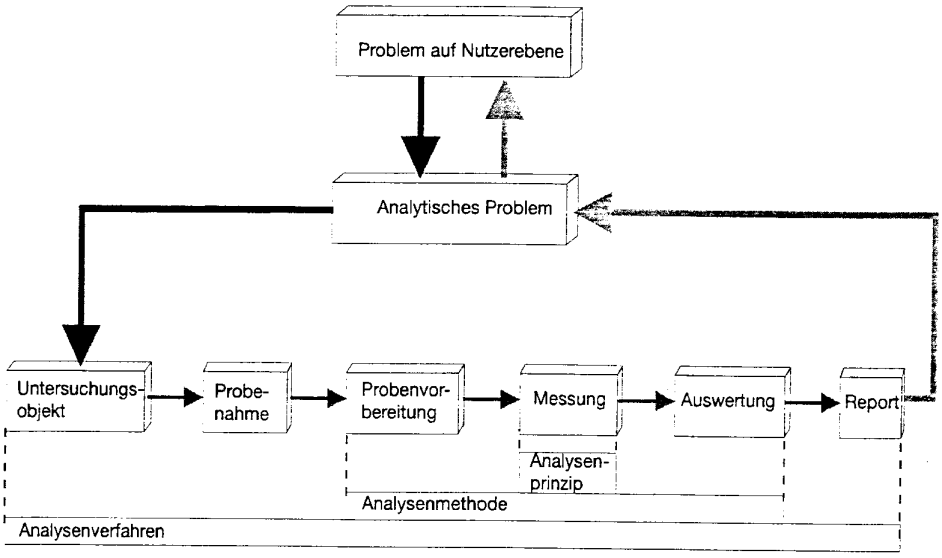
K. Danzer (Jena)
International Chemometrics Society

Aufgaben:
- Optimierung von Prozessen
- kritische Bewertung von Prozessen
- Qualitätssicherung
- Kritik an der Analyse
- etc.

Literatur:

M. Otto:	Analytische Chemie, VCH, Weinheim,1995
G. Schwedt:	Analytische Chemie - Grundlagen, Methoden und Praxis, Thieme Verlag, Stuttgart, 1995
L. Kolditz (Hrsg.):	Anorganikum, 13. Aufl., Bd. 2 Kap. 35.7 J. A. Barth Deutscher Verlag der Wissenschaften, Leipzig, Berlin, 1993
Autorenkollektiv:	Analytikum, 9. Aufl., Kap. 9 Deutscher Verlag der Grundstoffindustrie, Leipzig, Stuttgart, 1994
F. W. Fifield, D. Kealy:	Principles and Practice of Analytical Chemistry, Blackie Academic & Professional, London, 1990
R.C.Graham:	Data Analysis for the Chemical Science - A Guide to Statistical Techniques, VCH, Weinheim, 1993

1 (- 3) Wert(e) - nix wert !



(Graphik entnommen aus M. Otto: Analytische Chemie)

Fragestellungen:

Empfindlichkeit der Methode (sensitivity)

Aus der Kalibrierung ersichtlich, wie verändert sich das Signal bei Veränderung der Konzentration

Nachweisgrenze (detection limit)

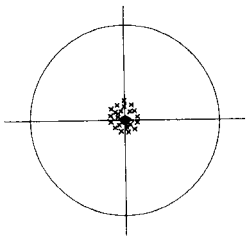
Wann ist ein Signal nur noch Untergrundrauschen?

Genauigkeit (precision)

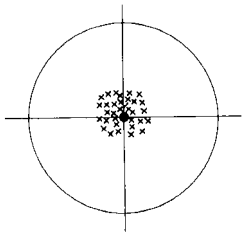
Richtigkeit (accuracy)

Fehlerquellen

Beispiel aus Analytikum: Trefferbilder eines Gewehrschützen

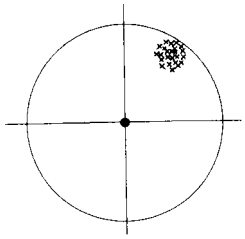


kein systematischer Fehler
(gutes Gewehr)
kleiner Zufallsfehler

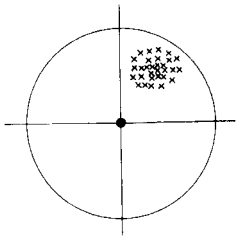


kein systematischer Fehler

großer Zufallsfehler



deutlicher systematischer Fehler (schlechtes G.)
kleiner Zufallsfehler



deutlicher systematischer Fehler
großer Zufallsfehler

Statistische /Zufallsfehler (random / indeterminated error)
systematischer Fehler (determinated error)

Faktoranalyse: Einfluß der Meßparameter auf Analyse => Optimierung

1. Eichung

Eichung: (calibration)

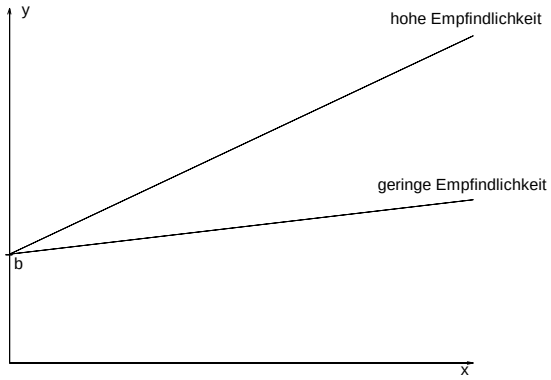
6 - 8 Meßwerte sind minimale Voraussetzung!

Meßsignal y als Funktion von x => lineare Regression (Genauigkeit ist zu beachten)

Kalibrierfunktion:

$$y = ax + b$$

a: Blindwert (c[Analyt] = 0) (background)
b: Empfindlichkeit (sensitivity)



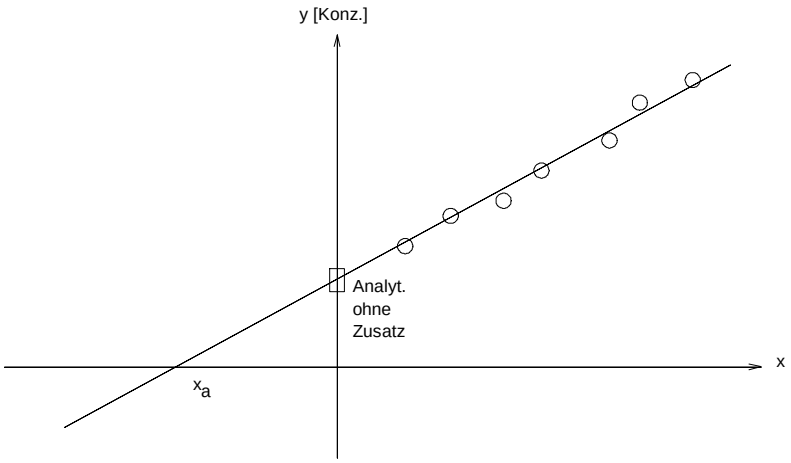
Analysenfunktion = Umkehrfunktion der Kalibrierfunktion: $x = \frac{(y-b)}{a}$

Störeffekte der Matrix (interferences) müssen vermieden werden.

Matrix: alles, außer dem Analyten

Vermeidung der Matrixstöreffekte ist möglich z. B. mit der Standardadditionsmethode

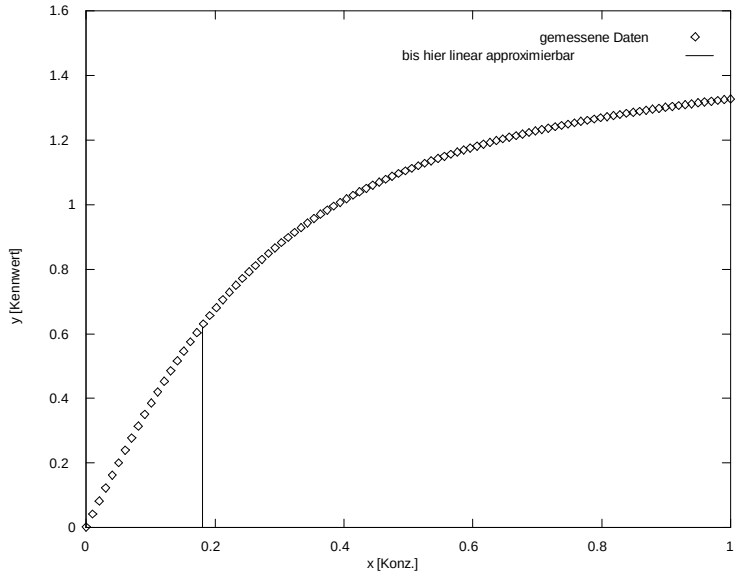
Standardadditionsmethode:



Analyt + Zusatz von einer bekannten Menge zu analysierender Substanz

(nach Messung:
Behandlung wie oben)

Problem: bei höheren Konzentration ist die Funktion oft keine Gerade mehr.

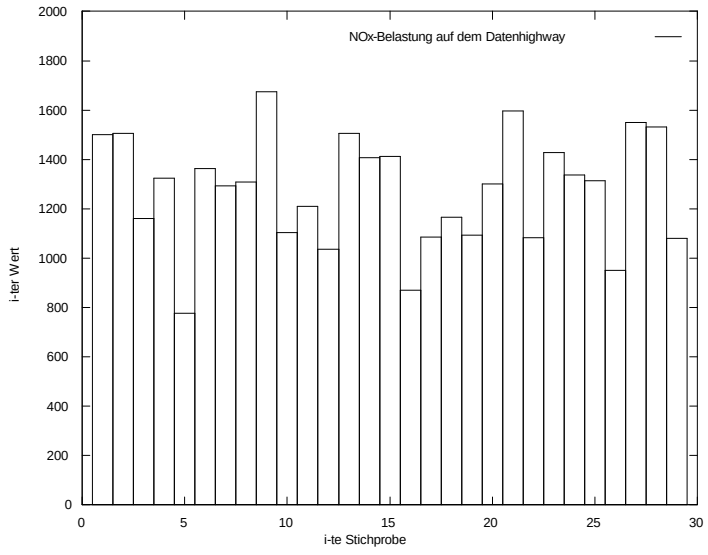


Lösungen:

- quadratische - kubische - ... Anpassung
- Verdünnung, damit die Funktion linear bleibt

2. Messung

Stichproben (endlicher Größe) (sample) → Ergebnisse → Mittelung der Ergebnisse



arithmetisches Mittel (mean):

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

empirische Standardabweichung s:

$$s_y = \sqrt{\left(\frac{\sum_i (y_i - \bar{y})^2}{n-1}\right)}$$

s^2 : Varianz (variance)
n - 1 : Freiheitsgrade

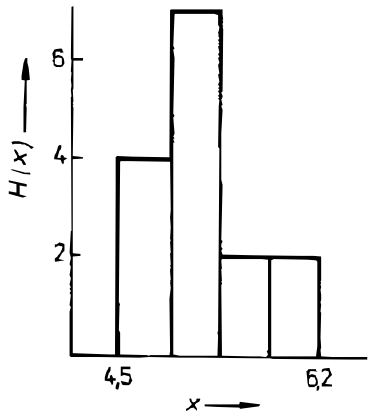
relative Standardabweichung

$s_{yr} = s_y / y$

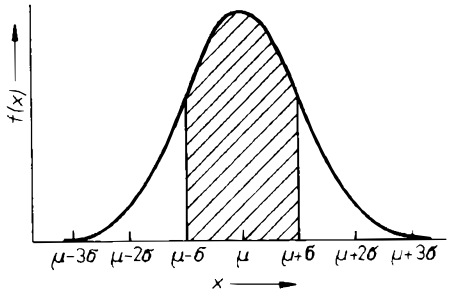
Ergebnisse von 15 Wiederholungsbestimmungen für die Kupferkonzentration in einer Weizenprobe. Die folgenden Spalten bezeichnen die arithmetischen Mittel bzw. empirische Standardabweichungen für jeweils drei aufeinanderfolgende bzw. die ersten i Meßwerte.

Messung Nr.	Meßwert ([Cu]/ppm)	$\bar{x}$ (3)	s (3)	$\bar{x}$ (i)	s (i)
1	5,3			5,30	---
2	5,3	5,20	0,17	5,30	0,00
3	5,0			5,20	0,17
4	5,0			5,15	0,17
5	6,1	5,30	0,70	5,34	0,45
6	4,8			5,25	0,46
7	4,5			5,14	0,51
8	5,7	5,00	0,62	5,21	0,51
9	4,8			5,17	0,49
10	4,9			5,14	0,47
11	5,6	5,20	0,36	5,18	0,47
12	5,1			5,18	0,45
13	6,2			5,25	0,52
14	5,3	5,43	0,71	5,26	0,50
15	4,8			**5,23**	**0,49**

Die Verteilung der Stichproben läßt sich mit sogenannten Häufigkeitsdiagrammen veranschaulichen. Hierzu unterteilt man den Bereich zwischen größtem und kleinstem Meßwert in gleich große Intervalle, deren Zahl per Definition $=\sqrt{n}$ (n = Zahl der Daten) ist, und zeichnet über jedem Intervall einen Streifen ein, dessen Höhe der Anzahl der Meßwerte in diesem Intervall entspricht. In unserem Beispiel werden 4 Balken eingezeichnet. Es deutet sich eine bei Zufallsfehlern typische Verteilungsform mit maximaler Häufigkeit im mittleren Bereich an. Im allgemeinen läßt sie sich als Normal- oder Gaußverteilung beschreiben.



Häufigkeitsdiagramm für die Daten aus der Tabelle



Dichtefunktion einer Normalverteilung

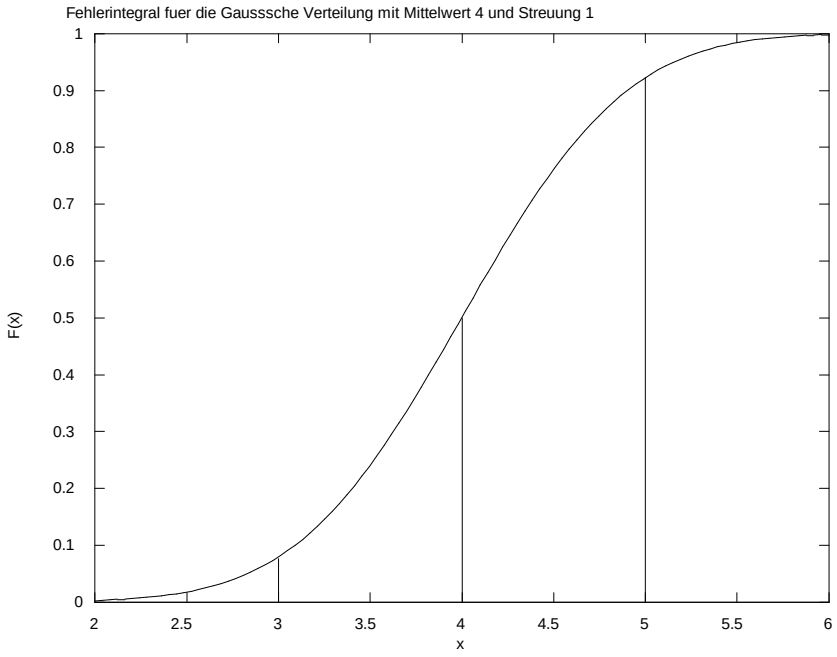
Daten entnommen aus:
L. Kolditz: Anorganikum 2
13. Auflage, 1993

Gauss-Verteilung:

$$f(x) = \frac{1}{s\sqrt{2p}} \cdot e^{-\frac{(x-m)^2}{2s^2}}$$

μ: theoretischer Maximumwert der Verteilung,
oft wird $\bar{x}$ als Näherung herangezogen
σ^2: theoretische Varianz
x= $\pm\sigma$ Steigung ist maximal ($f''(x) = 0$)

Bei sehr großen Stichproben nähert sich die Gestalt der Häufigkeitsdiagramme jener der Wahrscheinlichkeitsdichte f(x) an. Aus der Dichte lassen sich nun Wahrscheinlichkeiten ausrechnen. Die Flächengröße eines Intervalls unter einer Gaußverteilung entspricht der Wahrscheinlichkeit, daß der Meßwert im Intervall liegt. Mathematisch läßt sich die durch Integration der Gaußverteilung leicht bestimmen. Man erhält dann das Gaußsche oder Fehlerintegral F(x)



Gauss-Verteilung:
- kontinuierlich,
- Grenzwert

=> Fehlerintegral F(x)

Grenzen auf der x - Achse x_1	x_2	Teil der Gesamtfläche	Wahrscheinlichkeit P [%] = Teil der Gesamtfläche · 100
μ-1,00σ	μ+1,00σ	0,683	68,3
μ-1,96σ	μ+1,96σ	0,950	95,0
μ-2,00σ	μ+2,00σ	0,954	95,4
μ-2,58σ	μ+2,58σ	0,990	99,0
μ-3,00σ	μ+3,00σ	0,997	99,7
μ-3.29σ	μ+3.29σ	0,999	99,9

Werte, die mehr als 3σ vom Mittelwert μ entfernt sind, treten kaum auf, man spricht auch von der 3σ - Grenze.

Nachweisgrenze (detection limit (d. l.)) = f (Empfindlichkeit = Steigung)

Erfassungsgrenze (practical d. l.): erst ab hier sind vernünftige Aussagen möglich, wenn die Blindwertkurve und Meßkurve nur noch außerhalb der 3σ-Grenze überlagern.

3. Statistische Behandlung von sogenannten „Ausreißern“ (outliners)

3.1 Q-Test

Sortierung der Meßwerte nach Größe

$$Q = \frac{(x_n - x_{n-1})}{(x_n - x_1)}$$

x_n-x_1: größte Differenz
x_n: höchster Wert

Vergleich mit Q-Werten aus Normalverteilung für bestimmtes Vertrauensintervall (confidence level / interval) z. B.: 90%

Anzahl der Ergebnisse	$Q_{krit.}$ (90% Vertrauen)
2	---
3	0,94
4	0,76
5	0,64
6	**0,56**
7	0,51
8	0,47
9	0,44
10	0,41

bei 6 Meßwerten darf Q nur max. 0,56 betragen, sonst liegt der Wert mit 90%-Wahrscheinlichkeit außerhalb.

3.2 Grubb’s - Test

$$r^* = \frac{x^* - \bar{x}}{s\sqrt{\frac{n}{n-1}}}$$

x^*	: Ausreißer
$\bar{x}$	: Mittelwert
s	: empirische Standardabweichung
n	: Meßwert
$\sqrt{\frac{n}{n-1}}$	: Faktor fehlt in manchen Tabellenwerten!!!!!

Vergleich mit r^*-Werten der Normalverteilung

Wahrscheinlichkeit (probability): P
Irrtumswahrscheinlichkeit: α

$$P + \alpha = 1$$

Beispiel:

volumetrische Titration:

25,35
25,80
25,28
25,50
25,45
25,43

Ist 25,80 Ausreißer?

Q-Test: $Q = \frac{(x_n - x_{n-1})}{(x_n - x_1)}$ = (25,80 - 25,50)/(25,80 - 25,28) = 0,577

Vergleich: 0,56 für n = 6 bei 90% Wahrscheinlichkeit
=> für 90% Wert liegt über Schwelle
=> für 99% Wert liegt unter Schwelle (Q = 0,70)

Grubb'sTest:

$$r^* = \frac{x^* - \bar{x}}{s} = 1{,}83$$

$$r^* = \frac{x^* - \bar{x}}{s\sqrt{\frac{n}{n-1}}} = 1{,}67$$

$\bar{x}$: 25,47
s : 0,18

einfache Darstellung : 95% Vertrauen (P): Ausreißer ist ein wirklicher Ausreißer (1,822)
: 99% Vertrauen (P): Ausreißer ist kein wirklicher Ausreißer (1,944)

4. Vergleich von Datensätzen

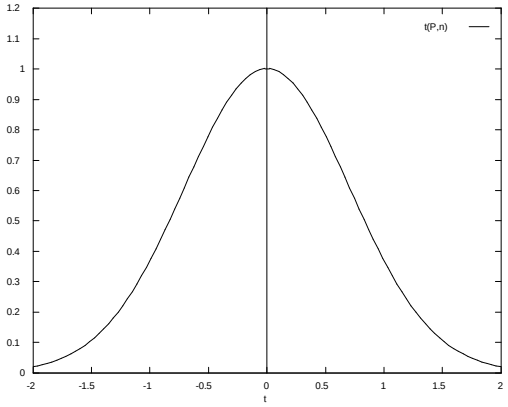
Nullhypothese: H_0: $\bar{x}$ (experimentell) = μ (wahr)
Differenz der Datensätze ist zufällig, kein statistisch signifikanter Unterschied

Alternativhypothese: $H_{i:}$ $\bar{x}$ (experimentell) ≠ μ (wahr)
Differenz der Datensätze ist nicht zufällig, statistisch signifikanter Unterschied

typische Irrtumswahrscheinlichkeiten: $\alpha = 0{,}05$ oder $\alpha = 0{,}01$

4.1 Student-t-Test

(William Gossett, Pseudonym: Student)



Einfacher Student-t-Test:

$$t(P,n) = \frac{|\bar{x} - m(wahr)|}{s_d} \cdot \sqrt{n}$$

n: Meßdaten
s_d: Standardabweichung
P: Wahrscheinlichkeit

Erweiterter Student-t-Test: (Vergleich von verschiedenen Mittelwerten)

$$t(P,n) = \frac{|(\bar{x}_1 - \bar{x}_2)|}{s_d} \sqrt{\left(\frac{n_1 n_2}{n_1 + n_2}\right)}$$

n: Meßdaten
s_d: Standardabweichung
P: Wahrscheinlichkeit
$\bar{x}_1, \bar{x}_2$: Mittelwerte aus verschiedenen Messungen

$$s_d = \sqrt{\left(\frac{(n_1 - 1)s_1^{\,2} + (n_2 - 1)s_2^{\,2}}{n_1 + n_2 - 2}\right)}$$

s_d: Standardabweichung
$(n_1\text{-}1),(n_2\text{-}1)$: Freiheitsgrade (degrees of freedom) (da x schon abgeleiteter Parameter aus Rohdatensatz ist)
$n_1+n_2\text{-}2$: Summe der Freiheitsgrade des Systems
s_1, s_2: Varianz

Aussage des Student-t-Testes:

- errechneter t-Wert liegt unter den tabellierten Werten
  $\Rightarrow$ Nullhypothese H_0 bzw. Aussage wahr
- errechneter t-Wert liegt über den tabellierten Werten
  $\Rightarrow$ Nullhypothese bzw. Aussage ist falsch
  $\Rightarrow$ Alternativhypothese H_i ist wahr

4. 2 Fischers F-Test

$$F = s_1^2/s_2^2$$

F : Prüfwert
s_1, s_2 : Varianz

Aussage des Fischers F-Test:

- errechneter F-Wert liegt unter den tabellierten Werten
  $\Rightarrow$ Nullhypothese H_0 bzw. Aussage wahr
- errechneter F-Wert liegt über den tabellierten Werten
  $\Rightarrow$ Nullhypothese bzw. Aussage ist falsch
  $\Rightarrow$ Alternativhypothese H_i ist wahr

t-Test erst anwenden, nach dem F-Test erfolgreich war!

Vertrauensbereich/Vertrauensintervall (confidence interval):

$$\text{Vertrauensbereich} = \bar{x} \pm t \cdot \frac{s}{\sqrt{n}}$$

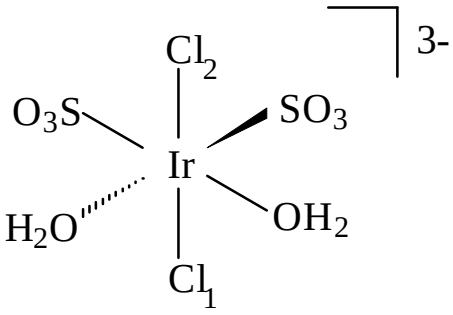
s : Standardabweichung
n : Zahl der Meßdaten
$\bar{x}$: Mittelwert

besondere Form der Standardabweichung = $\frac{s}{\sqrt{n}}$

relative Standardabweichung = $\frac{s}{\bar{x}} \cdot 100\%$

Beispiel: Untersuchung der Bindungslängen eines Iridiumkomplexes (AK Prof. Breitinger)

Sind die beiden Ir-Cl Bindungen in $Na_3[IrCl_2(SO_3)_2(OH_2)_2]\bullet 7H_2O$ gleich?



Messung: Ir - Cl_2: 237,9(2) pm

(2) = Varianz der letzten Stelle

Raumgruppe: Pnma

Messung: Ir - Cl_1: 233,0(2) pm

$$t(P,n) = \frac{|(\bar{x}_1 - \bar{x}_2)|}{s_d}\sqrt{\left(\frac{n_1 n_2}{n_1 + n_2}\right)} = \frac{|(233{,}9 - 237{,}0)|}{s_d}\sqrt{\left(\frac{n^2}{2n_2}\right)}$$

Die Werte für n_1 und n_2 sowie s_1 und s_2 sind gleich, da sie aus dem gleichen Datensatz der gleichen Meßung mit der gleichen Anzahl von Meßwerten stammen und die Bindungen in der gleichen Spiegelebene liegen.

$$s_d = \sqrt{\left(\frac{(n_1 - 1)s_1^{\,2} + (n_2 - 1)s_2^{\,2}}{n_1 + n_2 - 2}\right)} = \sqrt{\left(\frac{2(n-1)s^2}{2(n-1)}\right)} = s = 0{,}2 \text{ pm}$$

$$t(P,n) = \frac{|(233{,}9 - 237{,}0)|}{0{,}2}\sqrt{\left(\frac{n}{2}\right)}$$

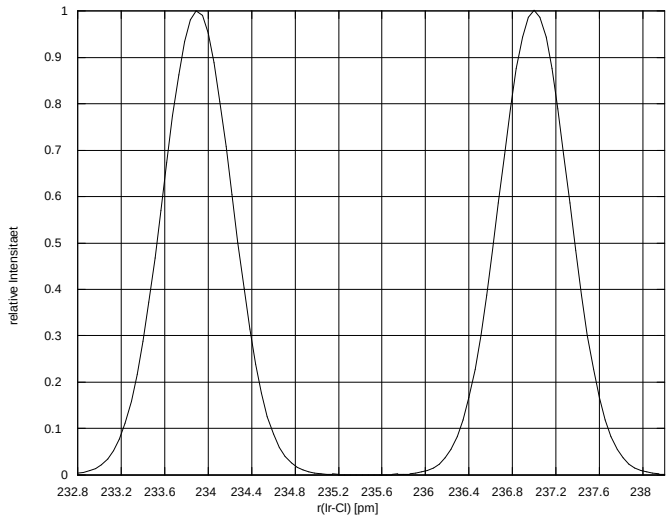
da n größer 2, muß n/2 größer 1 sein, und damit auch die Wurzel.

t = größer als 15,5 $\Rightarrow$ zu mind. 95% sind die beiden Bindungen verschieden!

Aufgrund der unterschiedlichen Koordinationssphäre (Kationen!) der Cl-Atome sind die Bindungen wirklich unterschiedlich.

4.3 „Faust"-Test - in der Praxis sehr hilfreich

Annahme: Ir-Cl_1 233,9(2) pm , Ir-Cl_2 237,0(2) pm sind Mittelwerte von Normalverteilungen.



n: 1 = 68%, n: 2 = 95%,
n: 3 = 99,7% -Wahrscheinlichkeit

=> Berechnung der 3σ-Schranke
→ klare Lücke -
Werte sind eindeutig unterschiedlich

→ Überlappung oder Berührung:
genauere Betrachtung nötig

$$\mu \pm n \cdot \sigma$$

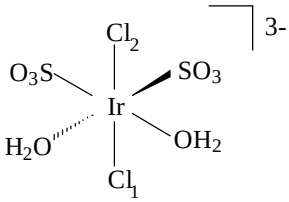
Beispiel:

Ir-Cl_1: 233,9(2) pm — + $3 \cdot \sigma$ → 234,5 pm
Ir-Cl_2: 237,0(2) pm — - $3 \cdot \sigma$ → 236,4 pm

=> klare Lücke → unterschiedlicher Wert

5. Fehlerfortpflanzung

Einführung am Beispiel der Kristallstrukturbestimmung von $Na_3[IrCl_2(SO_3)_2(OH_2)_2]\cdot 7H_2O$



meßtechnische Daten der experimentellen Messung:

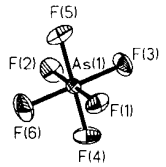
3408 Reflexe wurden gemessen

1750 Reflexe sind unabhängig (Mittelung über Friedelpaare hkl, $\bar{h}\bar{k}\bar{l}$)

115 Parameter sind zu bestimmen (Symmetriegründe, z.B.:
Cl_1 - Ir - Cl_2 liegen auf einer Spiegelebene => 1 Lageparameter fest => nur noch 2 Lageparameter frei)

Bei der Bestimmung einer Kristallstruktur werden bei anisotroper Berechnung für jedes Atom 9 Parameter bestimmt. 3 Lageparameter, 3 Hauptachsen und 3 Orientierungsparameter (=> Darstellung der Atome als Ellipsoide)
Für die Protonen werden nur 3 Lagekoordinaten und ein Temperaturparameter berechnet unter der Annahme, daß sich das Proton in einer Kugel „bewegt".

Einschub zur Darstellung der Atome als Ellipsoide:
Die Ellipsoide sind ein Maß für die temperaturbedingten Schwingungen der Atome um ihre Ruhelage. Schwerere Atome z. B.: Arsen zeigen kleinere Ellipsoide oder nähern sich einer Kugel an. Leichtere Atome z.B. Fluor zeigen im Gegensatz dazu größere Ellipsoide.



R. Mews et. al., Angew. Chem., **106** (1994) 1724.

Oft sind Einschränkungen für die Messung durch die Symmetrie möglich:
z.B.: ein Ortsparameter ist durch die Spiegelebene festgelegt
=> nach allen Einschränkungen: für den Beispielkomplex: 115

Qualität einer röntgenographischen Untersuchung:

$$\frac{\text{unabhängige Reflexe}}{\text{Parameter}} = \frac{1750}{115} = 15{,}3$$ (jeder Wert ist 15,3 mal überbestimmt)

Beurteilung: bis 8 sehr problematisch, bis 10 in Ordnung, 15 - 20 sehr gut

Braggsche Gleichung: $n\cdot\lambda = 2\cdot d_{hkl}\cdot\sin\vartheta$

ϑ: Beugungswinkel
n: enthält hkl mit
d_{hkl}: Gitterkonstante

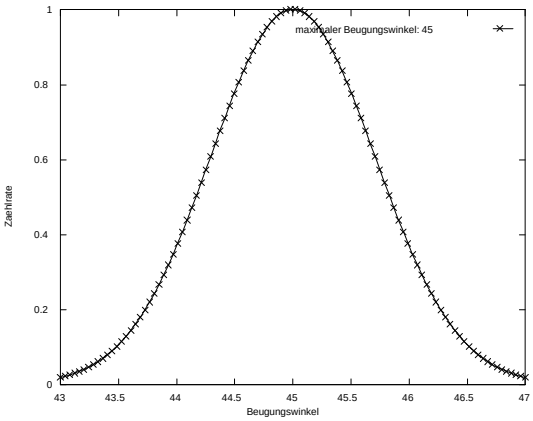
$$(1/d_{hkl})^2 = h^2a^{*2}+k^2b^{*2}+l^2c^{*2}+2hka^*b^*\cos(\gamma)+2klb^*c^*\cos(\alpha)+2hla^*c^*\cos(\beta)$$

a^*,b^*,c^* = reziproke Gitterkonstante

kubisches System: $\alpha = \beta = \gamma = 90^o \Rightarrow \cos(90^o) = 0$; $(1/d_{hkl})^2 = h^2a^{*2}+k^2b^{*2}+l^2c^{*2}$
monoklines System: $\cos(\beta)$ bleibt erhalten (Konvention!)

Meßgrößen: λ: ist sehr genau vorgegeben => Fehler wird ignoriert, da vernachlässigbar
ϑ: Beugungswinkel - guter Vertrauensbereich
Beugungswinkel liefert Gitterkonstante

Reflexe : I : Reflexe - geringe Reflexe: - geringer Vertrauensbereich
- starke Reflexe: - guter Vertrauensbereich
Reflexe liefern relative Lageparameter



ϑ-Bestimmung durch öftere Messung:

Maximum = $\vartheta \pm s(\vartheta)$

s: Standardabweichung -
wird in guter Analyse angegeben

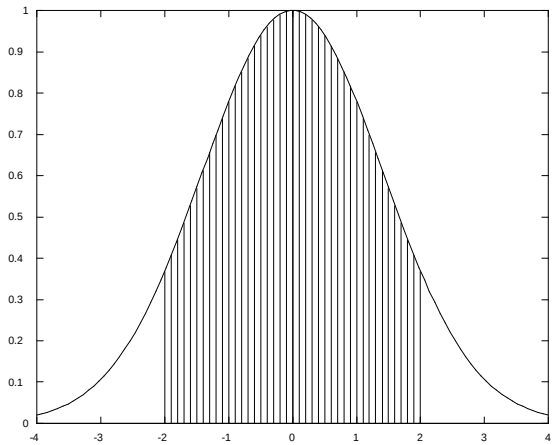
in Braggsche Gleichung eingesetzt:

$\sin(\vartheta) \rightarrow \partial(\sin(\vartheta)) = \cos(\vartheta) \cdot \partial(\vartheta)$

$\partial(\vartheta)$ wird als $s(\vartheta)$ genähert betrachtet; Meßwerte ist diskret, die Formel aber kontinuierlich

Wie wirkt sich der Fehler aus, wenn die Braggsche Gleichung nach d aufgelöst wird?

Relativlageparameter:



weiß: Untergrund, liniert: Meßwert

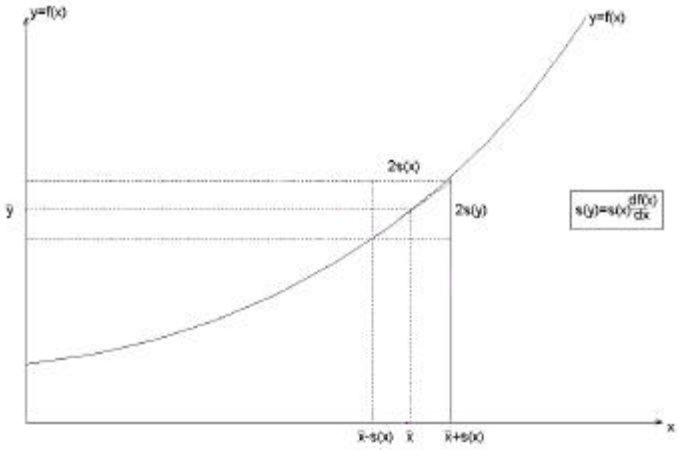
Bestimmung des Untergrunds → Meßwert - Untergrund

$$I_{hkl} \sim F_{hkl}^2$$

I : Intensität
(mit statistischen Fehlern behaftet)
F: Strukturfaktor

$$I_{hkl} = \Sigma_j f_j \bullet \exp(2\pi i (h_{xj} + k_{yj} + l_{zj}))$$

$h_x, k_y, l_z,$: relative Ortskoordinaten
i : Einheit der imaginären Zahl

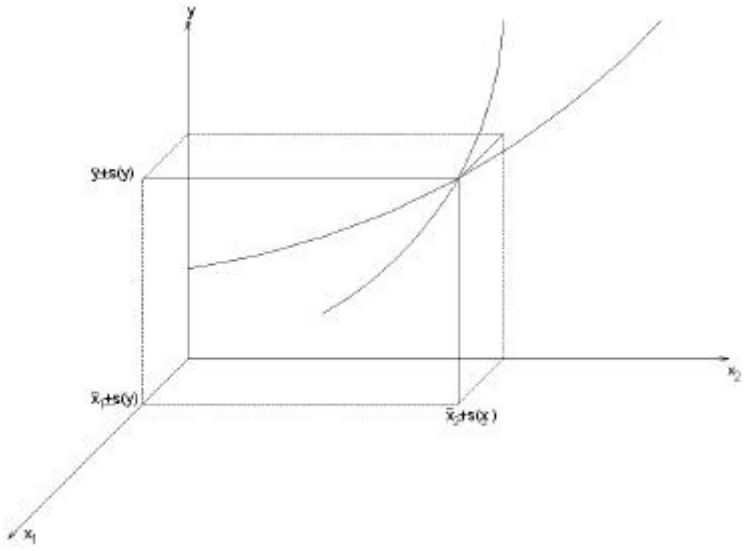


Annahme: im Bereich 2s(x) ist die Funktion linear => Ableitung $\frac{\partial f(\bar{x})}{\partial(x)}$

Fehlerbalken (error bar)

Fehlerfortpflanzung im 2-dimensionalen Fall:

$$d(y) = \frac{\partial f(x_1, x_2)}{\partial x_1} dx_1 + \frac{\partial f(x_1, x_2)}{\partial x_2} dx_2 = \sum_{i=1}^{2} \left(\frac{\partial f(x)}{\partial(x_i)} \right) dx_i$$



Fehlerfortpflanzung im n-dimensionalen Fall:

$$d(y) = \left(\frac{\partial f}{\partial (x_1)}\right) dx_1 + \left(\frac{\partial f}{\partial (x_2)}\right) dx_2 + \ldots + \left(\frac{\partial f}{\partial (x_n)}\right) dx_n = \sum_i \left(\frac{\partial f}{\partial x_i}\right) dx_i$$

Eine Ableitung kann positiv oder negativ sein => es könnte die $\Sigma = 0$ sein, dies wäre wenig sinnvoll, zur Vermeidung dieses Problems quadriert man die Beiträge und erhält somit als statistische Größe die Varianz.

$$\boxed{s^2(y) = \sum_i \left(\frac{\partial f(\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_n)}{\partial x_i}\right)^2 s^2(x_i)}$$

$s^2(y)$ ist somit die Varianz für die abgeleitete Größe y.

Beispiele:

Beispiel 1:

$y = x_1 + x_2$; $\left(\frac{\partial y}{\partial x_1}\right) = 1; \left(\frac{\partial y}{\partial x_2}\right) = 1;$ (Vorzeichen fallen durch das Quadrieren weg!)

$y = \Sigma_i x_i$

$s^2(y) = \Sigma_i s^2(x_i)$

Beispiel 2:

$y = a_1x_1 + a_2x_2$; $\left(\frac{\partial y}{\partial x_1}\right) = a_1 \left(\frac{\partial y}{\partial x_2}\right) = a_2;$

$s^2(y) = \Sigma_i a_i^2 s^2(x_i)$

Dieses Beispiel wird wichtig, wenn z. B. Mittelwerte gewichtetet werden müssen. Dieses Vorgehen ist z. B. in der Kristallstrukturanalyse üblich, wenn starke Reflexe, die relativ genau sind, mehr gewichtet werden sollen als schwache, mit großen Fehlern behaftete Reflexe.

Übersicht:

¦(y)	s(y)
1/x	$s(x)/x^2$
x^2	$2x \cdot s(x)$
x^n	$nx^{(n-1)} \cdot s(x)$

6. Datenanalyse

Für dieses Kapitel ist als Sekundärliteratur zu empfehlen:
M. Otto, Analytische Chemie, VCH, Weinheim, 1995, Kapitel 6.3 Multivariate Methoden

6.1 Univariate Datenanalyse/Modellierung - lineare Regression

Voraussetzung: eine unabhängige und eine abhängige Variable mit linearem Zusammenhang

$$y = b_0 + b_1x_1 (+ b_2x_2^2.....)$$

Erweiterung im Bedarfsfall zur Polynomdarstellung.

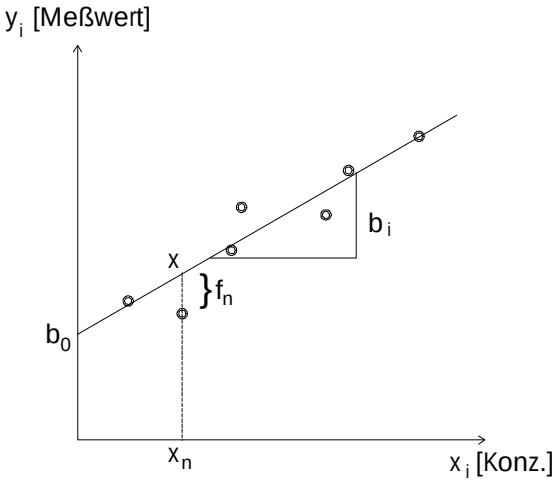
Vorprobe: graphische Auftragung der Daten
→ Ausreißer ? → vgl. Statistische Behandlung von sogenannten „Ausreißern“
→ graphische Regression - Gerade in Werte legen

Hintergrund: Minimierung der Summe der Fehlerquadrate

$f(b_0,b_1) = b_0 + b_1x_i - y_i$ x_i, y_i: Meßwerte
b_0, b_1: Variable

$f(b_0,b_1) = \sum_i(b_0 + b_1x_i - y_i)^2$ soll minimiert werden.

Modellvorstellung:



Achtung: Terminologiefalle!

Graph ist die Zeichnung (Graphik); es wird x und f(x) = y dargestellt.
Funktion ist die algebraische Zuordnung: x → f(x).

Schematische Ableitung der Regressionsformeln aus $f(b_0,b_1) = \sum_i (b_0 + b_1 x_i - y_i)^2$:

Extrema - Bestimmung

$$\left(\frac{\partial f(b_0,b_1)}{\partial b_0}\right) = 2\sum_{i=1}^{n} b_o + b_1 x_i - y_i = 0 \qquad \left(\frac{\partial f(b_0,b_1)}{\partial b_1}\right) = 2\sum_{i=1}^{n} x_i\left(b_o + b_1 x_i - y_i\right) = 0$$

Aus diesen Bedingungen wird ein lineares Gleichungssystem für b_o und b_1 aufgestellt:

$$b_0 n + b_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i$$

$$b_0 \sum_{i=1}^{n} x_i + b_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i$$

Da die Koeffizientendeterminante D > 0 ist,

$$D = n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2 = n\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2 > 0$$

kann unter Anwendung der Cramerschen Regel das lineare inhomogene Gleichungssystem gelöst werden mit dem Ergebnis:

$$b_1 = \frac{n\sum_i x_i y_i - \sum_i x_i \sum_i y_i}{n\sum_i x_i^2 - \left(\sum_i x_i\right)^2}$$

x_i, y_i: Wertepaare aus Messung

$$b_0 = \bar{y} - b_1 \bar{x}$$

$\bar{x}, \bar{y}$: Mittelwerte der x_i und y_i

Fehlerfortpflanzung

Mittelbarer Meßfehler steckt auch mit in Kalibrierfunktion!

Varianz der Steigung:

$$s^2(b_1) = \frac{s^2(y)}{\sum_i (x_i - \bar{x})^2}$$

$\sum_i (x_i - \bar{x})^2$: Summe der Fehlerquadrate

Varianz des Achsenabschnitts:

$$s^2(b_0) = \frac{s^2(y)\sum_i x_i^2}{n\sum_i (x_i - \bar{x})^2}$$

Varianz des y-Werts:

$$s^2(y) = \frac{\sum_i (y_i - \hat{y})^2}{n-2}$$

$\sum_i (y_i - \hat{y})^2$: Summe der Abweichungsquadrate

n - 2 : Zahl der Freiheitsgrade (Reduktion um 2, da für y-Wert und x-Wert Mittelwerte gebildet wurden)

y_i: Meßwert

$\hat{y}$: Erwartungswert, durch Kalibrierfunktion ermittelt

Zur linearen Regression stehen inzwischen gute Computerprogramme zur Verfügung.
Ein Beispiel in IBM-Basic ist im Anhang zu finden.
In der Praxis ist oft der umgekehrte Weg gewünscht, daß zu einem experimentell bestimmten y das zugehörige x gesucht wird; x kann hierbei z.B. eine Konzentration o. ä. sein.

$$x_0 = \frac{y_0 - b_0}{b_1}$$

Varianz des x-Werts:

$$s^2(x_0) = \frac{s^2(y)}{b_1^2}\left(\frac{1}{P} + \frac{1}{n} + \frac{\bar{y}_0 - \bar{y}}{b_1^2 \sum_i (x_i - \bar{x})^2}\right)$$

P: Zahl der Parallelmessungen

Je mehr Daten aus Messungen zur Verfügung stehen, desto besser ist das Ergebnis der Regression, da so die Zahl der Ausreißer minimiert werden.

Korrelationskoeffizient
(Maß für die Qualität der Funktion)

$$r = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\left[\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2\right]}}$$

Der Korrelationskoeffizient r schwankt zwischen ± 0 und ± 1. Je näher der Wert an ± 1 liegt, desto besser ist die Korrelation zwischen gemessenen Daten und der Funktion. Liegt der Wert näher an ± 0, so ist die errechnete Funktion schlecht an die Meßwerte angepaßt.

Trotz aller algebraischen Überprüfungen ist ein optischer Test der aufgetragenen Funktion und der Meßwerte unerläßlich.

Zur Verallgemeinerung des Regressionsproblems wird die Matrixschreibweise eingeführt. Das Modell für die Geradengleichung lautet dann:

$$\begin{pmatrix} y_1 \\ y_2 \\ . \\ . \\ . \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ . & . \\ . & . \\ . & . \\ 1 & x_n \end{pmatrix} \begin{pmatrix} b_0 \\ b_1 \end{pmatrix}$$

komprimierte Form: $\mathbf{y} = \mathbf{Xb}$

Bei Polynomen wächst die $n \times 2$ -Matrix und die Spaltenmatrix.

Warum wird die Matrizengleichung $\mathbf{y = Xb}$ nicht schon nach $\mathbf{b= yX^{-1}}$ umgeformt?

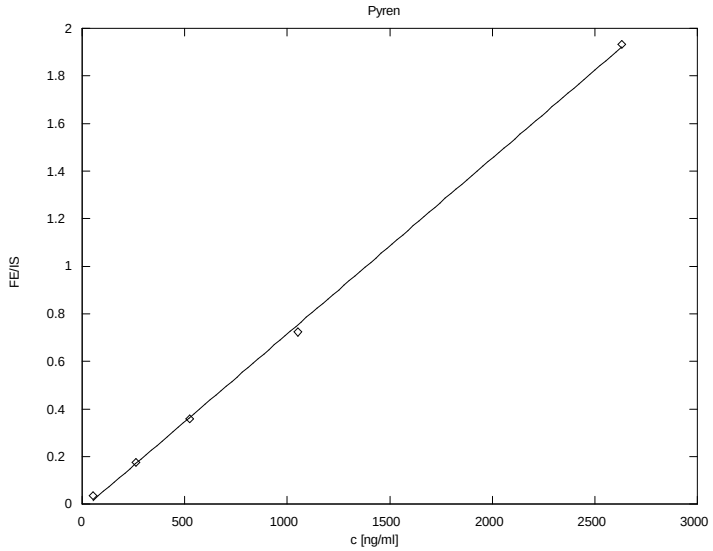
X ist nicht quadratisch und nur reguläre quadratische Matrizen haben eine Inverse => Umweg über Multiplikation der Matrix mit ihrer Transponierten (macht Matrix quadratisch).

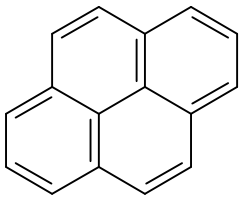
$$b = (X^T X)^{-1} X^T y$$

Beispiel:

Zur Bestimmung von Pyren mittels GC-MS in Luftanalysen soll eine Eichgerade erstellt werden. Da aus meßtechnischen Gründen kein genaues Volumen der am Ende zu vermessenden Lösung bekannt ist, wird ein interner Standard zugesetzt und als Meßgröße die Fläche der gemessenen Peaks (FE) pro Fläche des interne Standardpeaks (IS) angegeben.

x (ng/ml)	52.600	263.000	526.000	1052.000	2630.000
y (FE/IS)	0.033	0.174	0.356	0.721	1.931





Pyren - ein potentiell krebserregender polyzyklischer Kohlenwasserstoff

Als Geradengleichung wurde ermittelt: $y = -0.025685 + 0.000739x$

Als Korrelationskoeffizient wurde für die erhaltene Gleichung $r(x,y) = 0.999661$ ermittelt.

Ein Zusammenhang zwischen x und y ist nur gesichert, wenn r > r(P,f) bei f = n - 2 Freiheitsgraden. Bei einem n (Anzahl der Messungen) von 5 ergibt sich folgendes Bild:

Grenzwerte r(P,f) zur Prüfung des Korrelationskoeffizienten:
(aus K. Doerffel: Analytikum, 1994)

f	P = 0,95	P = 0,99	f	P = 0,95	P = 0,99
1	1,00	1,00	6	0,71	0,83
2	0,95	0,99	7	0,67	0,80
3	0,88	0,96	8	0,63	0,77
4	0,81	0,92	9	0,60	0,74
5	0,75	0,87	10	0,58	0,71

α: Irrtumswahrscheinlichkeit $P + \alpha = 1$
P: Wahrscheinlichkeit

Wir können also annehmen, daß mit einer Irrtumswahrscheinlichkeit $\alpha = 0.01$ % von r(P,f) = 0,96 bei einem ermittelten Wert r = 0,999661 der lineare Zusammenhang gesichert ist.

6.2 Multivariate Datenanalyse/Modellierung

Voraussetzung: mehrere unabhängige und abhängige Variable

Beispiel: Spektroskopische Mehrkomponentenanalyse
(Voraussetzung: Für alle Komponenten gilt das Lambert-Beersche Gesetz; d.h. es ist ein linearer Zusammenhang zwischen Absorbanz und Konzentration gegeben.)

$$A = \varepsilon(\lambda) \cdot c \cdot l$$

A: Absorbanz
$\varepsilon(\lambda)$: molarer Extinktionskoeffizient bei einer definierten Wellenlänge
c: Konzentration
l: Länge/Schichtdicke der Küvette für alle Messungen gleich

Aus der Gleichung des Lambert-Beerschen Gesetzes läßt sich das Gleichungssystem für die Messungen der Absorbanzen A_i bei verschiedenen Wellenlängen λ_i erstellen.
Es ist darauf zu achten, daß immer mindestens so viele Wellenlängen λ gemessen werden, wie Komponenten im System vertreten sind.

$$A_1 = e_1(\lambda_1)c_1 l + e_2(\lambda_1)c_2 l + \ldots = l\sum_{k=1}^{m} e_k(\lambda_1)c_k$$

$$A_2 = l\sum_{k=1}^{m} e_k(\lambda_2)c_k$$

$$A_3 = l\sum_{k=1}^{m} e_k(\lambda_3)c_k$$

.
.
.

$$A_n = l\sum_{k=1}^{m} e_k(\lambda_n)c_k \qquad k = n$$

In eine Matrix umformuliert erhält man:

$$\mathbf{A} = l \cdot e \cdot \mathbf{c}$$

Da die $\varepsilon_k(\lambda_n)$ bekannt sind erhält man für die Analysenfunktion wenn k = n ist:

$$c_0 = \mathbf{A_0}^T(le)^{-1}$$

Ist die Gleichung überbestimmt k < n (Zahl der Messungen größer als die Zahl der Komponenten), so erhält man für die Analysenfunktion:

$$c_0 = \mathbf{A_0}((le)^T(le))^{-1}$$

Zur Lösung dieser Matrizengleichung bietet sich für kleinere System die Cramersche Regel an, bei großen Systemen muß auf Verfahren wie z. B. die Gauß-Eliminierung zurückgegriffen werden.

Grundbegriffe in der multivariten Datenanalyse:

Objekte: Proben, Untersuchungsgegenstände

Merkmale: Variable
Meßgrößen (univariate Eigenschaften)
Meßwerte - metrisch skalierte Merkmale (B.: Banden im IR-Spektrum)
- nominal skalierte Merkmale (B.: Etwas ist rot.)
- binär skalierte Merkmale (B.: Etwas ist oder ist nicht.)

Beispiel: Nach einem Autounfall soll kriminalistisch ermittelt werden:
Sind Lacksplitter am beschädigten Auto zu finden - binär skaliertes Merkmal.
Die Lacksplitter sind rot und zeigen unter dem Mikroskop Metallpigmente - nominal skaliertes Merkmal.
Mittels IR-Spektroskopie können Bindemittel identifiziert werden - metrisch skalierte Merkmal.

Klassen: Gruppen zusammengehöriger (d. h. ähnlicher) oder als zusammengehörig zu betrachtender Objekte - z. B. aufgrund von Herstellungsbedingungen, Herkunft, Entnahmeort oder auch von Gütekriterien.

Ähnlichkeit: Multivariate Eigenschaft von Objekten - begründet auf einer Übereinstimmung in allen wesentlichen Merkmalen. Grundlage: Datenstruktur im n-dimensionalen Raum, die sich nach einer mathematischen Dimensionsreduzierung in Form von Bildern veranschaulichen läßt, auch mehrdimensional.
Beispiel: Sternenhimmel - räumliche Anhäufung von Sternen werden z. B. zu Galaxien zusammengefaßt.

Datenmatrix: Darstellung der Merkmale (Meßwerte) einer Reihe von Objekten in Form einer Matrix (Matrixalgebra)

Pattern: Typische Merkmalskombination für ein Objekte oder eine Klasse von Objekten bzw. auch für ein Merkmal (eine Variable)

Pattern Recognition: Mustererkennung, Ermittlung typischer Objektmuster z. B. Klassen, in denen bestimmte Objekte zusammengefaßt werden können. Im engeren Sinne: Musterwiedererkennung, d. h. Klassifikationsverfahren, womit nach bestimmten erlernten Klassifikationsregeln Objekte in a priori bekannte oder gesetzte Gruppen eingeteilt werden.

Supervised Learning: „Lernen“ der Klassenzugehörigkeitsfunktion unter Anleitung (mit Vorinformation über existierende Klassen)

Unsupervised Learning: Mustererkennungsverfahren für Objekte von denen a priori keine Klassenzugehörigkeit bekannt ist (Suche nach charakteristischen Datengruppen = Clustern)

Datenmatrix für Pattern Recognition-Methode

Pattern Vektor eines Objektes i(x_{1i} + x_{2i} ... x_{m1})
n : Anzahl der verschiedenen Objekte (→)
m : Anzahl der unterschiedlichen Analysenergebnisse (↓)

$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & x_{13} & \cdots & x_{1m} \\ x_{21} & x_{22} & x_{23} & \cdots & x_{2m} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{n1} & x_{n2} & x_{n3} & \cdots & x_{nm} \end{pmatrix}$$

Die Datenmatrix muß vollständig sein, um sinnvolle Aussagen treffen zu können.

Euklidischer Abstand als multivariates Ähnlichkeitsmaß in der Clusteranayse

$$d_{ij} = \sqrt{\sum_{k=1}^{p} (x_{ik} - x_{jk})^2}$$

p: Variable (Analysenergebnis) zwischen allen Objektpaaren i und j

6.3 Mustererkennung

Beispiel: Haaranalytik

In der Haaranalytik können Schadstoffe festgestellt werden. Wird längs des Wachstumsrichtung der Haare gemessen, so können sogar zeitliche Veränderungen festgestellt werden. z. B.: Drogenkonsum, Schwermetallbelastung o. ä.

Elementgehalte von Haaren verschiedener Personen in ppm.

Haar Nr.	Kupfer	Mangan	Chlor	Brom	Jod
1	9,2	0,30	1730	12,0	3,6
2	12,4	0,39	930	50,0	2,3
3	7,2	0,32	2750	65,3	3,4
4	10,2	0,36	1500	3,4	5,3
5	10,1	0,50	1040	39,2	1,9
6	6,5	0,20	2490	90,0	4,6
7	5,6	0,29	2940	88,0	5,6
8	11,8	0,42	867	43,1	1,5
9	8,5	0,25	1620	5,2	6,2

oberflächliche augenscheinliche Betrachtung der Werte für Chlor:
=> 3 Gruppen: (1,4,9), (2,5,8) und (3,6,7)

oberflächliche augenscheinliche Betrachtung der Werte für Brom:
=> 3 Gruppen: (1,4,9), (2,5,8) und (3,6,7)

Standardisierung von Merkmalsdaten

Merkmal j: z. B. Chlor-Gehalt

s_j : Streuung $s_j = \sqrt{\left(\frac{1}{n-1}\sum x_{ij} - \bar{x}_j\right)}$

$\bar{x}_j$: Mittelwert $\frac{1}{n}\sum x_{ij}$

x_{ij} : Meßwert

$$\frac{x_{ij} - \bar{x}_j}{s_j} = x_{ij}'$$

Durch die Division mit s_j wird der Abstand normiert.

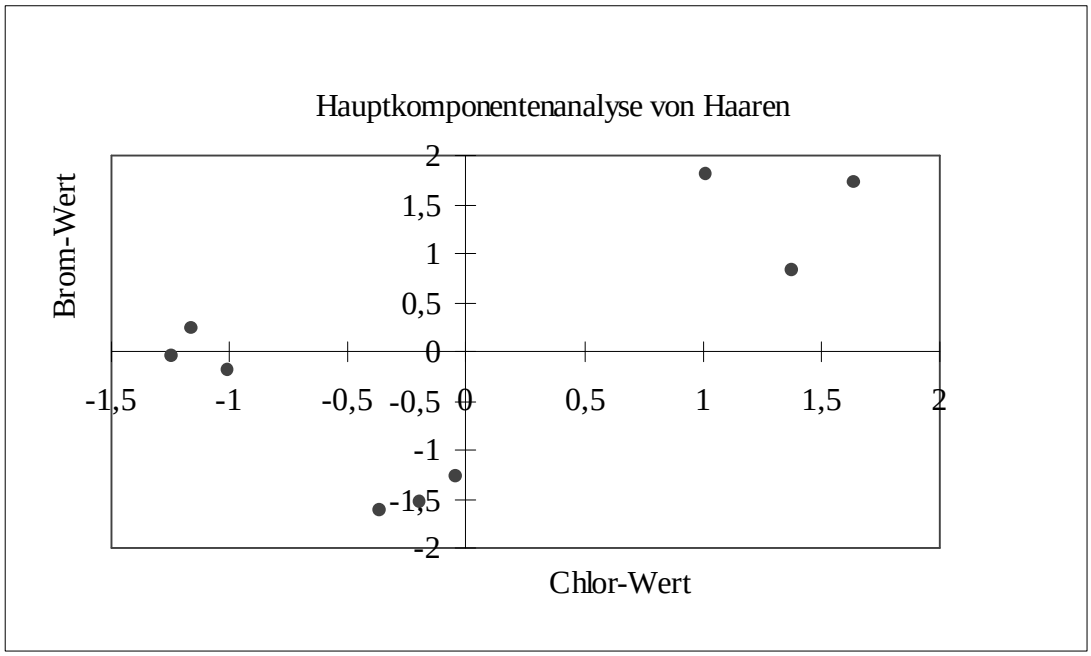
Unüberwachtes Lernen

Projektionsmethode

Eine der wichtigsten Projektionsmethoden ist die Hauptkomponentenanalyse.

Beispiel: Anwendung der Standardisierung von Meßdaten auf die Haaranalyse:

Haar-Nr.	x_{ij}´ Chlor-Wert	x_{ij}´ Brom-Wert
1	0,36	-1,27
2	-0,81	0,24
3	-1,79	0,84
4	0,02	-1,61
5	-0,66	-0,19
6	1,47	1,82
7	2,13	1,74
8	-0,91	-0,04
9	0,19	-1,53



Klassifizierung:
Die Ähnlichkeiten der Werte (2,5,8) und (1,4,9) sind stark ausgeprägt, für die verbleibenden Werte (3,6,7) ist das Bild nicht ganz so deutlich.

Clusteranalyse

Bei der Clusteranalyse faßt man die Objekte an Hand der Ähnlichkeit ihrer Merkmale schrittweise zusammen.
Zur Bewertung der Ähnlichkeit dient häufig der Euklidische Abstand. Er berechnet sich nach

$$d_{ij} = \sqrt{\sum_{k=1}^{p} (x_{ik} - x_{jk})^2}$$

p: Variable (Analysenergebnis) zwischen allen Objektpaaren i und j

Ist der Abstand zwischen zwei Objekten klein, so ist die Ähnlichkeit sehr groß.
=> Gruppenbildung
=> Dendrogramm

Statt des Abstandsmaßes, hier dem Euklidischen Abstand, können Ähnlichkeitsmaße S angewendet werden. Unter Berücksichtigung des Euklidischen Abstands erhält man so:

$$S_{ij} = \frac{d_{ij}}{d_{ij}(\max)}$$

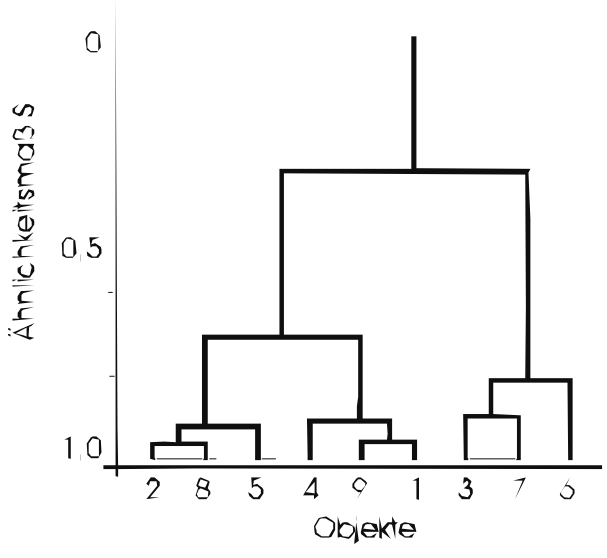
$\frac{d_{ij}}{d_{ij}(\max)}$: Art normierter Euklidischer Abstand

d_{ij} (max): größter in den Daten auftretender Abstand

Wie man sich leicht verdeutlichen kann, bedeutet dabei:

$S_{ij} = 1$: hohes Maß an Ähnlichkeit
$S_{ij} = 0$: keine Ähnlichkeit

Aus dem Ähnlichkeitsmaß S kann ebenso leicht ein Dendrogramm erstellt werden.



Zur Erstellung eines Dendrogrammes können mehrere Wege beschritten werden:
Man startet hierbei immer bei der Berechnung einer Abstandsmatrix.

Objekt	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
1	0								
2	2,405	0							
3	2,250	3,392	0						
4	1,327	2,570	3,017	0					
5	2,713	1,613	3,391	2,799	0				
6	3,062	4,232	1,724	3,766	4,623	0			
7	3,378	4,611	1,666	3,793	4,655	1,337	0		
8	2,562	*0,666*	3,531	2,821	1,189	4,528	4,860	0	
9	1,667	3,603	2,991	1,511	3,987	3,116	3,306	3,910	0

In unserem Fall wird dann daraus noch das Ähnlichkeitsmaß S errechnet. Wir wollen an dieser Stelle jedoch zur Vereinfachung weiterhin den Euklidischen Abstand betrachten.

Den geringsten Abstand zeigen die Objekte 2 und 8, deshalb werden sie zu einem Cluster zusammengefaßt und als erstes in das zu erstellende Dendrogramm eingetragen. Für den nächsten Schritt existieren 3 Möglichkeiten:

- single linkage: z. B. 8 wird als Repräsentant für das neue Objekt (2,8) ausgesucht, anschließend das Objekt mit dem kleinsten Abstand zum Repräsentanten gesucht; das Objekt 2 wird gestrichen.
- complete linkage: nachdem 2 und 8 ein neues Objekt sind wird nach den nächsten beiden Werten gesucht, die den kleinsten Abstand zueinander zeigen, Objekt 2 wird gestrichen.
- average linkage: Berechnung des Mittelpunktes von 2 und 8 und Erstellung einer neuen Matrix.

Die nach einem der obigen Verfahren bestimmten Werte werden in das Dendrogramm eingetragen, und man verfährt danach in gleicher Art weiter.

In unserem Fall erhalten wir die drei Cluster, die wir schon mit der Projektionsmethode vermutet haben. Oft sind die Anzahl der Cluster aber schon aus den Versuchsbedingungen bekannt, z. B. stammen in unserem Beispiel die Haare von drei Tatverdächtigen.

Überwachtes Lernen

Ist ein Mustererkennungsproblem durch Methoden des unüberwachten Lernens gelöst oder bereits anderweitig bekannt, so kann ein Klassifizierungsmodell aufgestellt werden, mit dem dann andere bzw. neue Daten bearbeitet werden können.

In unserem Fall wurden die Haare wie folgt zugeordnet:

Person	*Haare*
A	2, 5, 8
B	1, 4, 9
C	3, 6, 7,

Eine wichtige Klassifizierungsmethode ist die lineare Diskriminanzanalyse (LDA). Sie kann angewendet werden, wenn normalverteilte Daten und unterschiedliche Klassenschwerpunkte vorliegen . Bei der linearen Diskriminanzanalyse wird an Hand der Kenntnis der Zugehörigkeit der Objekte zu einer bestimmten Klasse (z. B.: Person A: Haare 2,5,8) eine Eigenwertanalyse durchgeführt. Der zum größten Eigenwert gehörende Eigenvektor liefert die erste lineare Diskriminanzfunktion s_1; der zum nächstgrößten Eigenwert gehörige Eigenvektor liefert die zweite Diskriminanzfunktion s_2, usw. Dieses Verfahren wird solange betrieben, bis das Problem als gelöst betrachtet werden kann.

Nach dem Lernvorgang erfolgt dann im nächsten Schritt die Anwendung: in unserem Fall die Zuordnung des Haares vom Tatort.

Kupfer	Mangan	Chlor	Brom	Jod
9,2	0,27	2200	9,8	4,7

Die Zuordnung des Haares erfolgt durch Einsetzung der Meßwerte in die Diskriminanzfunktionen. Das Haar gilt dann als zugehörig zu der Klasse, für die der Euklidische Abstand zwischen dem Objekt und dem Schwerpunkt der Klasse am geringsten ist. In unserem Fall konnte das Haar dem Tatverdächtigen B zugeordnet werden.

7. Anhang

Dieses IBM-Basic Programm führt eine gewichtete lineare Regression aus. Es stammt aus:
K. Ebert, H. Ederer, T. L. Isenhour: Computer Applications in Chemistry,
VCH 1989, Weinheim

```
0 REM "REGLINW "         EBERT/EDERER           880506
1 REM ******************************************
2 REM ***       Weighted Linear Regression        ***
3 REM *** The linear least squares fit to a       ***
4 REM *** straight line is calculated with eighted***
5 REM *** data points. In addition the errors     ***
6 REM *** in the parameters a and b are obtained. ***
9 REM ******************************************
1000  DIM X(100),Y(100),W(100)
1010  A$="                                  "
1100  INPUT "How many data points (y,x) ";N
1120  FOR I=1 TO N : W(I)=1 : NEXT I : PRINT
1140  PRINT "Do you want to weight your data points ;
1150  INPUT B$
1200  FOR I=1 TO N
1300    PRINT "  y(";I;") = ";
1310    INPUT Y(I)  : LOCATE CSRLIN-1,1
1330    PRINT A$;A$ : LOCATE CSRLIN-1,1
1350    PRINT "  y(";I;") =";Y(I);
1400    PRINT "  x(";I;") = ";
1410    INPUT X(I)  : LOCATE CSRLIN-1,1
1450    IF B$<>"yes" THEN GOTO 1615
1490    PRINT A$;A$ : LOCATE CSRLIN-1,1
1500    PRINT "  y(";I;") =";Y(I);
1600    PRINT "  x(";I;") =";X(I);
1610    PRINT "  w(";I;") = ";
1613    INPUT W(I)  : LOCATE CSRLIN-1,1
1615    PRINT A$;A$ : LOCATE CSRLIN-1,1
1620    PRINT "  y(";I;") =";Y(I);
1630    PRINT "  x(";I;") =";X(I);
1640    PRINT "  w(";I;") =";W(I)
1700  NEXT I
2000  S1=0
2100  S2=0 : S3=0 : S4=0 : S5=0
2200  FOR I=1 TO N
2250    S1=S1+W(I)
2300    S2=S2+X(I)*W(I)
2400    S3=S3+Y(I)*W(I)
2500    S4=S4+X(I)*X(I)*W(I)
2600    S5=S5+Y(I)*X(I)*W(I)
2700  NEXT I
```

```
3000  D1=S1*S4-S2*S2
3100  D2=S3*S4-S5*S2
3200  D3=S1*S5-S2*S3
4000  A=D2/D1
4100  B=D3/D1
5000  PRINT "The regression line for the data is"
5100  PRINT " y  = ";A;"  +  ";B;"* x"
6000  S=0
6100  FOR I=1 TO N
6200    S=S+W(I)*(Y(I)-(A+B*X(I)))^2
6300  NEXT I
6500  PRINT:PRINT "Sum of squared deviations =";S
6600  D=SQR(S/(N-2))
6605  PRINT:PRINT "Standard deviation  =";D
6610  DA=D*SQR(S4/D1) : DB=D*SQR(S1/D1)
6620  PRINT:PRINT "Number of data points =";N : PRINT
6630  PRINT "a = ";A;" +/- ";DA:PRINT
6640  PRINT "b = ";B;" +/- ";DB:PRINT
7000  PRINT:PRINT "Do you want to compare the input "
7100  PRINT "values with the regression values ";
7200  INPUT A$
7300  IF A$="yes" THEN GOTO 8000
7500  GOTO 8400
8000  PRINT
8050  PRINT " x          y(input)   y(regression) "
8100  FOR I=1 TO N
8200    PRINT X(I);TAB(12);Y(I);TAB(23);A+B*X(I)
8300  NEXT I
8400  PRINT:PRINT "Do you want to interpolate data"
8500  PRINT "using the regression line ";
8600  INPUT A$
8700  IF A$="yes" THEN GOTO 9000
8800  GOTO 9999
9000  PRINT "x-value";:INPUT X
9050    PRINT A$;A$;: LOCATE CSRLIN-1,1
9100  PRINT "y(";X;") = ";A+B*X
9200  GOTO 9000
9999  END
```