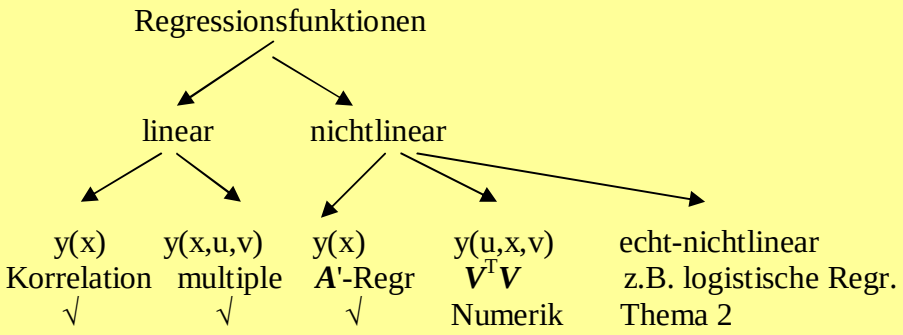
INHALT

a) linear, 2 Variable
Regressionsgerade y(x) → b) linear, > 2 Variable
multiple Regression y (u, v, w, x)

c) nichtlinear, 2 Variable
linearisierbar $y = a\ ^3(x) + b$ → d) nichtlinear, > 2 Variable multiple
multiple nichtlineare Regression y (u, v, w, x)
z.B. $y(x, u, v) = a \ln x + b\, e^u + c\, v + d$)
siehe Numerik Abschnitt 4.23



Regression, Korrelation, FISHER-Prüfmaß, t-Prüfmaß, DW_1-Prüfmaß

Stufe 1: lineare Einfachregression => r, r^2, F-Prüfmaß $xF_{empir} > xF_{crit}$??

Stufe 2: lineare multiple Regression: Datenausgabe interpretieren, Abschn. 1.13
$r^2_{adjustiert}$, xF_{empir} , $t_{empir}(b_i)$, $r_{ij} < 0{,}5$?, DW | { $d_{unten} < 2 < d_{oben}$ }
Regressionsfunktion y(u,v,w), Schätzwert für Szenario

Stufe 3: nichtlineare Einfachregression => r^2 als Varianzenquotient, Abschnitt 1.17
typisch $y(x) = a \ln x + b$
A'-Regression $\hat{y} = a\, \varphi(x) + b$

Stufe 4: multiple Regressionsanlysen, auch nicht-linear, Abschnitt 1.20
typisch $y(u, x) = b_0 + b_1 \ln u + b_2 x^2$
Regressionsgleichung über $\boldsymbol{V.V^T}$
V^T - Regression in Numerik 2012

Stufe 5: echte nichtlineare Regression,
nicht analog zu $y = m x + b$ lösbar, sondern mit logarithmieren
im Thema 2
z.B. logistische Funktionen: $y = \dfrac{10}{1 + e^{-1{,}5x+5}}$

1.1 ORGANISATION

a) Gliederung

Thema 1	Korrelation	sta1korr.doc	Zusammenhänge
Thema 2	Zeitreihen	sta2zeit.doc	Zusammenhänge
Thema 3	Häufigkeit	sta3haeuf.doc	Theorie
Thema 4	Wahrscheinlichkeit	sta4wahr.doc	Theorie
Thema 5	Verteilungen	sta5verteil.doc	Unterschiede | Abweichungen
Thema 6	Stichprobentests	sta6tests.doc	Unterschiede | Abweichungen

b) Dateien

Zum Download von www.NEFFF.de
sta*.doc , sta*.pdf Skript zu den 6 Themen
sta0ex.xls für Übungen während der Vorlesung, regelmäßig aktualisiert
sta9uebung.doc Übungsaufgaben, sta9loesung.xls Lösungen dazu

c) Literatur (Kursive Angabe zum *Zitieren*)

Puhani, Josef, Statistik, 11.Auflage 2008, ISBN 3-89694-433-9, 20 €
Bleymüller, Josef, u.a., Statistik für Wirtschaftswiss., 14.Auflage, ISBN 3-8006-31156, 16,50 €
Bamberg, u.a. Statistik, 12. Auflage 2002, ISBN 3-486-27218-7, 19,80 €
Rinne, Horst Taschenbuch der Statistik, 4. Aufl. 2008, ISBN 978-3-8171-1827-4, >1000 S. 40 €
Elser, Thomas, Statistik für die Praxis, 2004, ISBN 3-527-50097-9, 39,90 €
Fahrmeir, Ludwig, u.a., 5. Auflage 2004, ISBN 3-540-21232-9, 29,95 €
Lambacher-Schweizer, Stochastik, Leistungskurs, Klett-Verlag, beliebige Auflage
oder andere "Schulbücher"

d) Für die Klausur zugelassene Hilfsmittel

Taschenrechner, Eingeführte Formelsammlung.

1.2 BEHAUPTUNGEN

Täglich werden Vermutungen geäußert und Behauptungen aufgestellt bei unsicherer Datenlage.
Statistische Methoden erlauben uns, vernünftige Entscheidungen unter Ungewissheit zu treffen.

Auszug aus der Medizin– Zeitschrift "Jama" 8.2.2006 **Sind ältere Frauen, die auf fetthaltige Kost verzichten, besser vor Herzinfarkten, Schlaganfällen, Brust- oder Darmkrebs geschützt ?** **Teilstudie Brustkrebs** **Es wurden 48 835 übergewichtige bis fettsüchtige Frauen im Alter zwischen 50 und 79 Jahren über 8 Jahre beobachtet. 60% blieben bei ihrer gewohnten Kost. 40% senkten den Fettanteil ihrer Kost auf maximal 20% nach Anleitung von Ernährungsberatern…** **Was Brustkrebs angeht, so senkte sich die Quote von 45 auf 42 Krebsfälle je 10 000 Personen…** **Das ist kein Hinweis auf einen schützenden Effekt fettarmer Diät, das ist statistisch nicht signifikant, das könnte auch Zufall sein**… *Unterschied*	**28.2.2009** **Triumphierende Meldung der Bundesfamilienministerin:** **In Deutschland werden wieder mehr Kinder geboren. Nach einer Schätzung des Statistischen Bundesamtes kamen im vergangenen Jahr 680.000 bis 690.000 Kinder zur Welt, das sind 10.000 bis 20.000 mehr als im Jahr 2006.** **Die endgültige Zahl wird das Amt erst im Frühsommer verkünden, wenn alle Landesämter ihre Daten vollständig ausgewertet haben. Die Zahlen der ersten drei Quartale von 2007 wurden auf das ganze Jahr 2007 hochgerechnet.** **Gehen wir von 685.000 Kinder aus und einem Zuwachs von 15.000, dann sind das 2,2 %.** **Liegt das noch im Zufallsbereich, oder ist der "Trend" statistisch gesichert?** **Auswirkungen der Elternzeiten??**
Metastudie zum Verhalten von männlichen Jugendlichen zwischen 12 und 16 Jahren in den USA, Kanada, Mexiko und Australien: (2005) **Zwischen dem mittleren täglichen Fernsehkonsum und dem BMI besteht ein statistisch gesicherter Zusammenhang der Form:** **BMI = 1,87 . TVzeit[h/d] + 18,83 [kg/m^2]** Zusammenhang	**Evidenzorientierte Wissenschaft** **evidence (Nachweis) n Evidenz (Offensichtlichkeit)** **S P R A B B** **1. Verum-Gruppe und Placebo-Gruppe** (z.B. Pille, die genauso aussieht wie die zu untersuchende Substanz) **placebokontrolliert** **2. randomisiert** d.h. Zufallsauswahl unter den Patienten **3. doppelt-blind** d.h. weder die Patienten noch die Untersucher wissen, wer Placebo eingenommen hat **4. möglichst große Stichprobe, Signifikanz** **5. anonymisert** Untersucher ≠ Auswerter S P R A B B

Grundsätzlich geht man von der Nullhypothese H_0 aus:
diese lautet:
der vermutete Zusammenhang besteht nicht,
die beobachtete Abweichung ist zufällig.

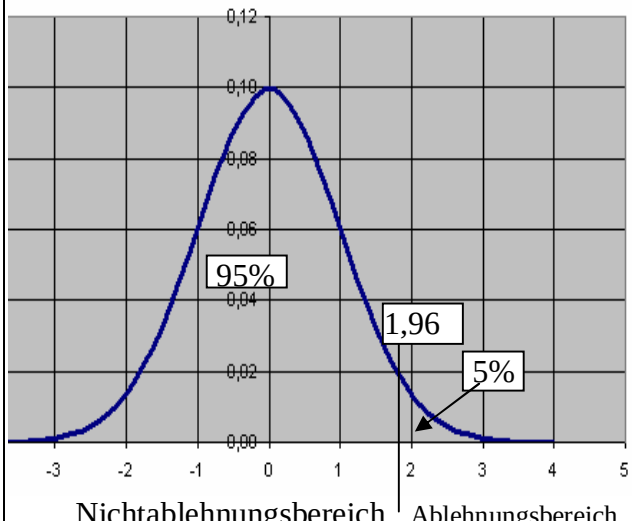
Mit Hilfe statistischer Methoden muss man dann prüfen, ob der vermutete Zusammenhang statistisch gesichert ist, d.h. ob er überzufällig ist (ob er signifikant ist);
erst dann kann man die Nullhypothese H_0 ablehnen und
der Alternativ-Hypothese H_1 zustimmen,
erst dann kann man darauf vertrauen, dass es diesen Zusammenhang / diese Abweichung gibt.

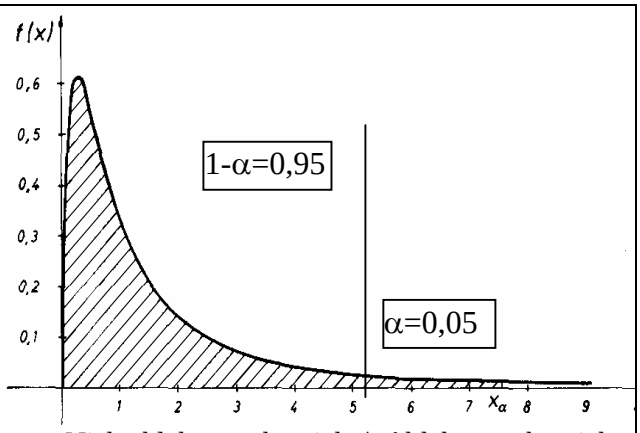
Man formuliert dann:
es besteht ein signifikanter Zusammenhang
es besteht eine signifikante Abweichung (ein signifikanter Unterschied).

Das sind die beiden Haupt-Blickrichtungen der Statistik: Zusammenhänge und Unterschiede.

1.3 PRÜFMAßE – AUSBLICK

Für die Prüfung auf "statistisch gesichert" benutzt man Prüfmaße.





Hinweis zu Flächen, die ins Unendliche ragen: Seite 1.23 1. *Neyman-Niveau, 1940, London*

Die Funktionsgraphen zeigen Wahrscheinlichkeitsdichten *f(x)* für Ereignisse *x* an.

Die Flächeninhalte zwischen den Funktionsgraphen zu *f(x)* und der x-Achse sind 100% = 1.

Die Flächeninhalte sind die Wahrscheinlichkeiten für das Eintreffen der Ereignisse $-\infty < x < x_{krit}$.

Bei der Wahrscheinlichkeit von mindestens 95% setzt man das Prüfmaß $x_{kritisch}$.

Solange der empirische (d.h. bei der Untersuchung gemessene Wert) kleiner ist, als dieses Prüfmaß, muss man von einem zufälligen Zusammenhang bzw. einer zufälligen Abweichung ausgehen.

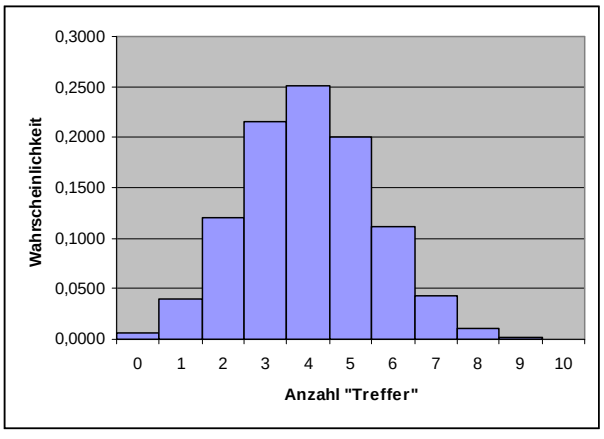
Nullhypothese heißt: der Zusammenhang bzw. die Abweichung ist zufällig.

Nullhypothese nicht ablehnen heißt: der Zusammenhang bzw. die Abweichung ist zufällig.

Nullhypothese ablehnen heißt: der Zusammenhang bzw. die Abweichung ist nicht mehr zufällig, oder der Zusammenhang bzw. die Abweichung ist statistisch gesichert.

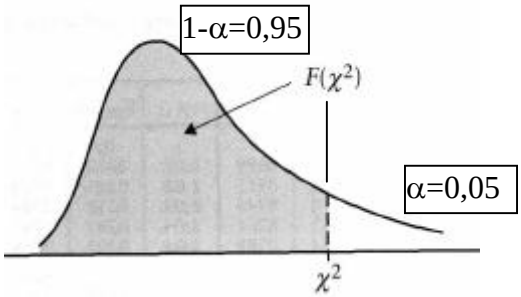
Wenn für den tatsächlich gemessene Wert $x_{empirisch}$ gilt $x_{empirisch} > x_{kritisch}$

dann ist der vermutete Zusammenhang bzw. die vermutete Abweichung statistisch gesichert (überzufällig, signifikant)



Eine diskrete Wahrscheinlichkeitsfunktion (Einzelwerte) … signifikanter Unterschied, wenn 7 Personen oder mehr die betreffende Eigenschaft aufweisen

$$f_{10;0,4}(x) = \binom{10}{x} 0,4^{x} \cdot 0,6^{10-x}$$



Nichtablehnungsbereich | Ablehnungsbereich

Der Zusammenhang zwischen den Merkmalen X und Y ist statistisch gesichert, wenn gilt

$\chi^2_{berechnet} > \chi^2_{kritisch}$.

d.h. wenn $\chi^2_{empirisch}$ im Ablehnungsbereich für die 95% Aussagesicherheit liegt.

… nächstes Standard-Niveau bei 99% …

Hier geht es um die Chi²-Verteilung.

1.4 VERTEILUNGSFUNKTIONEN

In der komplexen Welt des Lebens, der Technik und der Wirtschaft kann man das Geschehen nicht mit der Strenge physikalischer Gesetze verfolgen, sondern man kalkuliert zufällige Abweichungen und Ungenauigkeiten ein. Ein Ergebnis *e* heißt zufällig, wenn es nicht vorhersehbar ist.

3 Arten Variable: x, $\mathbf{x}$, X

Für Merkmalsträger (Personen, Unternehmen, Produkte usw.) wird das Merkmal X untersucht.
Diese Größe X nennt man Zufallsvariable.
Für ein Merkmal X (z.B. Alter von Personen) werden die Beobachtungswerte x_i gemessen.
Das Merkmal X nimmt den Wert (Ausprägung, Beobachtungswert, Ergebnis, Zustand) x_i an,
dafür schreibt man $X = x_i$ und
$f(X = x_i)$ ist die "Wahrscheinlichkeit" dafür, dass das Merkmal X genau den Wert x_i annimmt.
$F(X \le x_i)$ ist die Wahrscheinlichk. dafür, dass das Merkmal X höchstens den Wert x_i annimmt.
vgl. in der Analysis: f(3) f(x = 3) hier: f(X = 3)

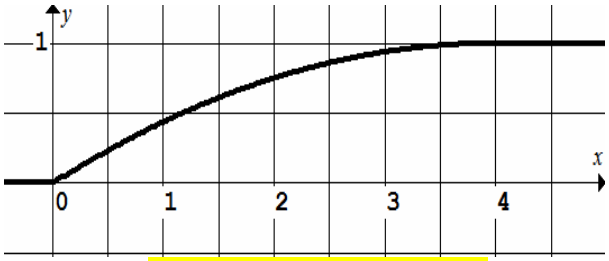
Beispiel 1.1

Die Verspätung der S-Bahn an einer bestimmten Haltestelle ist eine stetige Zufallsvariable X.

Für die Funktion f sei gegeben: $f(x) = \begin{cases} 0,5 - 0,125\,x & \text{für } 0 \le x \le 4 \\ 0 & \text{für alle übrigen } x \end{cases}$ [Minuten]

Man nennt f Dichtefunktion, $f(x)$ sind die Dichten. (vgl. den Begriff "Randfunktion")

Dichtefunktion f(x)



Verteilungsfunktion $F(x_2)$

Für die Dichten f(x) gilt: $f(x) \ge 0$ und $\int_{-\infty}^{+\infty} f(x)dx = 1 = 100\%$

Die Verteilungsfunktion $F(x_2)$ ist die Flächeninhaltsfunktion, sie ordnet jeder oberen Grenze x_2 einen Flächeninhalt zu: $F(x_2) = \int_{-\infty}^{x_2} f(x)dx$

hier: $\int (0,5 - 0,125x)\,dx = 0,5x - 0,0625x^2 \Rightarrow F(x) = \begin{cases} 0 & \text{für } x < 0 \\ 0,5x - 0,0625x^2 & \text{für } 0 \le x \le 4 \\ 1 & \text{für } x > 4 \end{cases}$

Die Flächeninhalte werden als Wahrscheinlichkeiten interpretiert:
Die Wahrscheinlichkeit dafür, dass X im Intervall [a; b] liegt, ist

$$W(a \le X \le b) = \int_a^b f(x)dx = F(b) - F(a) = [F(x_2)]_a^b$$

□ Wahrscheinlichkeit von 2: $W(X = 2) = \int_2^2 f(x)dx = F(2) - F(2) = 0$

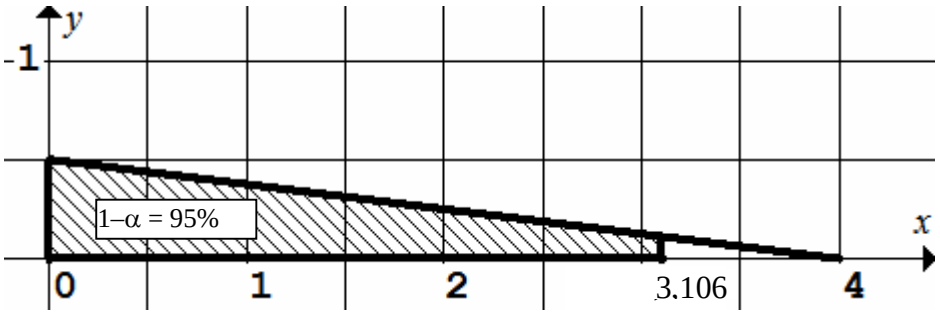
$f(x)$ gibt also keine Wahrscheinlichkeiten an, Wahrscheinlichkeiten gibt es nur für Intervalle.

□ Wahrscheinlichkeit dafür, dass die S-Bahn zwischen 1,6 und 2,4 Minuten Verspätung hat

$$W(1,6 \le X \le 2,4) = \int_{1,6}^{2,4} \left(\frac{1}{2} - \frac{1}{8}x\right)dx = \left[\frac{1}{2}x - \frac{1}{16}x^2\right]_{1,6}^{2,4} = 1,2 - 0,36 - (0,8 - 0,16) = 0,2 = 20\%$$

1.5 SICHERHEIT

Beispiel 1.1 Fortsetzung



Wie viel Minuten Verspätung ist mit 95-prozentiger Sicherheit zu erwarten?
Dazu muss $x_2 = x_{kritisch} = x_c$ für $F(x_2) = 0{,}95$ berechnet werden:

$$0{,}95 = \int_0^{x_c}\left(\frac{1}{2}-\frac{1}{8}x\right)dx = \left[\frac{1}{2}x-\frac{1}{16}x^2\right]_0^{x_c} \Rightarrow \frac{1}{2}x_c-\frac{1}{16}x_c^2-0=0{,}95 \quad \|\cdot(-16)$$

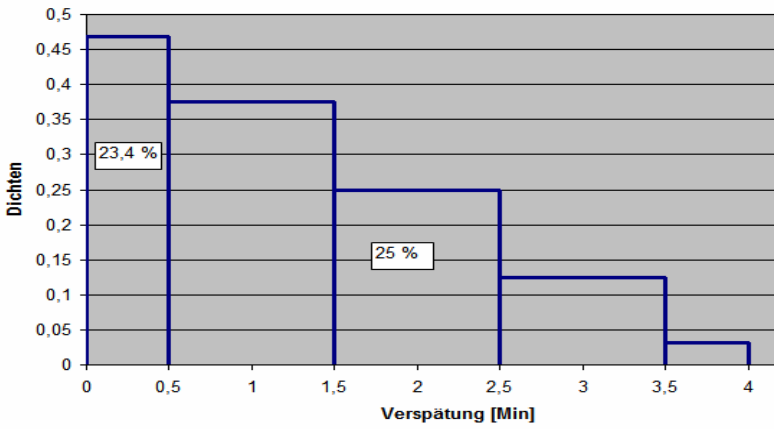
$$x_c^2-8x_c+15{,}2=0 \Rightarrow x=4\pm\sqrt{16-15{,}2}=4\pm0{,}894 \Rightarrow x_c=3{,}106$$

Der Zufallsbereich für 1–α = 0,95 reicht also bis 3,1 Minuten Verspätung.
Ist die Verspätung größer als 3,106 Minuten, dann ist die S-Bahn überzufällig spät.
(*Ein einzelnes Ereignis oder eine Stichprobe von n Ereignissen? → Thema 4)*
Die Wahrscheinlichkeit α nennt man Restunsicherheit oder Irrtums-Wahrscheinlichkeit oder Signifikanzniveau. Die Wahrscheinlichkeit 1–α nennt man Aussagensicherheit.
Für statistische Schlussfolgerungen sollte 1–α mindestens 0,95 sein und α < 0,05.

Die Grenzen x_c und die Flächeninhalte bzw. Wahrscheinlichkeiten F(x) müssen nicht jeweils mit Hilfe der Integralrechnung bestimmt werden, für die (sieben) Standardverteilungen gibt es entsprechende Tabellen, vgl. Formelsammlung.

Die Zufallsvariable X ist stetig. In der Praxis wird man aber diskrete Zeitpunkte messen und diese passenden Intervallen zuordnen. Man könnte z.B. die Merkmalsausprägungen x_i = 0, 1, 2, 3, 4 [Minuten] messen und dazu die Wahrscheinlichkeiten angeben:

x_i	Merkmals-klasse [a , b]	$F_{[a,b]}$
	[0 ; 0,5[	0,234
1	[0,5 ; 1,5[	0,375
2	[1,5 ; 2,5[	0,250
3	[2,5 ; 3,5[	0,125
4	[3,5 ; 4]	0,016
	Summe	1,000



Die Dichtefunktion $f(x_i)$ einer diskreten Zufallsvariable wird als Histogramm dargestellt.
Dabei entsprechen die Flächeninhalten genau den Wahrscheinlichkeiten.
"Dichten", weil sich die y-Werte (hier: Wahrscheinlichkeiten) auf das ganze Intervall beziehen.

Für die Verteilungsfunktionen gilt dann $F(x_i) = \sum_{i=0}^{n} f(x_i)$ *Dichte = m / V*

Wir unterscheiden stetige, diskrete und nominale Zufallsvariable.
Nominale Zufallsvariable sind nicht-quantitativ z.B. X = {kath., evang., islam., sonst}

1.6 MAßZAHLEN

1. Stichproben

Die Untersuchung aller Daten einer Grundgesamtheit N (z.B. alle Wahlberechtigte) ist viel zu aufwendig, man beschränkt sich auf Stichproben vom Umfang n und schließt mit geeigneten Verfahren von den Ergebnissen der Stichprobe auf die der Grundgesamtheit. Um Gesetze über Wahrscheinlichkeiten herzuleiten, bildet man Modelle mit passenden Ereignismengen $\{x_i\}$, z.B. "Zwei Würfe mit einem idealen Würfel", Zufallsvariable könnte die Augensumme X sein.

2. Einzelwerte

In den Themen 1 und 2 benutzen wir Datenreihen aus einzelnen Werten x_i.

3. Mittelwert (Arithmetischer Mittelwert)

Mittelwert in der Grundgesamtheit

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$$ erwarteter Mittelwert

Mittelwert in der Stichprobe:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$ gemessen/berechnet

4. Streuungen misst man allgemein als Summe von Abweichungsquadraten $A_x = \sum_{i=0}^{n}\left(x_i - \bar{x}\right)^2$.

Durch Ausmultiplizieren erhält man die numerisch stabilere 2.Formel $^{II}A_x = \sum x_i^2 - \frac{1}{n}\left(\sum x_i\right)^2$

Für viele weitere Formeln existieren jeweils zwei Formen: Differenzform und Produkte-Form.
vgl. 1.7 Umformungen.

5. Varianz ist der Mittelwert der Abweichungsquadrate $\frac{1}{n}\sum_{i=0}^{n}\left(x_i - \bar{x}\right)^2$

Erwartete Varianz in der Grundgesamtheit

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}\left(x_i - \mu\right)^2 = \boxed{\sigma_n}^2$$

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N} x_i^2 - \mu^2$$

Varianz in der Stichprobe

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2 = \boxed{\sigma_{n-1}}^2$$

$$s^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - n\,\bar{x}^2\right)$$

6. Standardabweichung

Varianzen haben unanschauliche Benennungen wie Personen^2, kg^2, Stunden^2.

Deshalb arbeitet man mit der Standardabweichung $\sigma = \sqrt{\sigma^2}$ bzw. $s = \sqrt{s^2}$

meistens mit $$s = \sqrt{\frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - n\,\bar{x}^2\right)}$$ Taste: [¾n-1]

7. Varianzverhältnisse

Setzt man Varianzen aus derselben Stichprobe ins Verhältnis zueinander, dann heben sich die Nenner auf, z.B. beim Bestimmtheitsmaß

$$r^2 = \frac{s_{\hat{y}}^2}{s_y^2} = \frac{\frac{\sum_{i=1}^{n}\left(\hat{y}_i - \bar{y}\right)^2}{n-1}}{\frac{\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2}{n-1}} = \frac{\sum_{i=1}^{n}\left(\hat{y}_i - \bar{y}\right)^2}{\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2}$$

1.7 UMFORMUNGEN

a) Konstanten $\sum_{i=1}^{n} c = n \cdot c = c_1 + c_2 + ... + c_n$ n-mal

b) Summenregel: $\sum_{i=1}^{n} x_i \pm \sum_{i=1}^{n} y_i = \sum_{i=1}^{n} (x_i \pm y_i)$

c) Faktorregel: $\sum_{i=1}^{n} k \cdot x_i = k \cdot \sum_{i=1}^{n} x_i$

d) Aber Achtung ! $\sum_{i=1}^{n} x_i \cdot \sum_{i=1}^{n} y_i \neq \sum_{i=1}^{n} (x_i \cdot y_i)$ und speziell: $\left(\sum_{i=1}^{n} x_i \right)^2 \neq \sum_{i=1}^{n} (x_i)^2$

Produkt zweier Summen ≠ Summe der Produkte, Quadrat einer Summe ≠ Summe der Quadrate

$$(x_1 + x_2 + ... + x_n) \cdot (y_1 + y_2 + ... + y_n) = (x_1 y_1) + (x_1 y_2) + ... \neq (x_1 y_1) + (x_2 y_2) + ... + (x_n y_n)$$

und für Quadrate: $\left(\sum_{i=1}^{2} x_i \right)^2 = (x_1 + x_2)^2 = x_1^2 + 2x_1 \cdot x_2 + x_2^2 \quad \neq \quad x_1^2 + x_2^2 = \sum_{i=1}^{2} (x_i)^2$

e) Umrechnung $\sum (x_i - \bar{x})^2$ [1. Formel] $\sum x_i^2 - n\,\bar{x}^2$ [2. Formel]

Aus $\bar{x} = \frac{1}{n} \sum x_i$ ergibt sich $n\bar{x} = \sum x_i$, damit lässt sich $\sum (x_i - \bar{x})^2$ umrechnen:

$$\sum (x_i - \bar{x})^2 = \sum (x_i^2 - 2\bar{x}x_i + \bar{x}^2) = \sum x_i^2 - \sum 2\bar{x}\, x_i + \sum \bar{x}^2 = \sum x_i^2 - 2\bar{x} \sum x_i + n\bar{x}^2 =$$

$$= \sum x_i^2 - 2\bar{x}\, n\, \bar{x} + 1n\bar{x}^2 = \sum x_i^2 - 2n\,\bar{x}^2 + n\bar{x}^2 = \sum x_i^2 - 1n\,\bar{x}^2$$

$$\frac{1}{n}\left(\sum x_i^2 - n\,\bar{x}^2 \right) = \frac{1}{n} \sum x_i^2 - \bar{x}^2$$

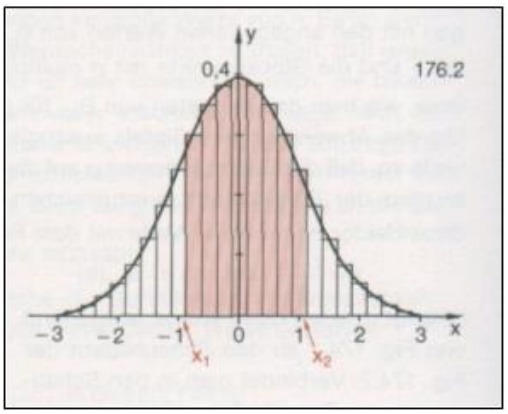
f) Bei der Varianz einer Stichprobe wird durch n–1 dividiert $s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$

Die Freiheitsgrade df (degrees of freedom) sind die Anzahl der unabhängigen Einzelwerte, die in die statistischen Berechnungen einbezogen werden können, es ist die Anzahl der frei wählbaren Einzelwerte. In obiger Rechnung liegt $\bar{x}$ bereits fest, es gibt also noch n–1 Freiheitsgrade. Bei großen Stichproben ist $n - 1 \approx n$.

- □ Eine Datenreihe 2, 4, 5, 6, 9. $\bar{x}$ = 26/5 = 5,2 => nur noch n–1 = 4 frei wählbare Daten.
- □ Bei der Bestimmung der Grenze x_F des FISHER-Prüfmaßes liegt sowohl $\bar{x}$ als auch $\bar{y}$ bereits fest. Der Freiheitsgrad ist dann n–2.

g) Bei stetigen Daten benutzt man die analogen Formeln für Integrale, die man ja vereinfacht als Summen der Flächen aus
"unendlich vielen Rechtecken, mit den Längen f(x)" und den im Verschwinden begriffenen Rechteckbreiten dx begreifen kann.

$$\lim_{\substack{n \to \infty \\ \Delta x \to 0}} \sum f(x) \cdot \Delta x = \int f(x)\,dx$$



1.8 REGRESSIONSGERADE

Beisp. 1.2 Eine Messreihe, zur Erinnerung: Abschnitt 1.22

Bei einem chemischen Prozess wird ein Zusammenhang zwischen der Bestrahlung mit UV-Licht x [Min] und dem Härtegrad y eines Polymers vermutet.
X sei die Einflussgröße und Y die beeinflusste Größe. (Bivariate Untersuchung)
Entsprechenden Tests ergaben folgende Wertepaare (x_i | y_i).
Die Wertepaare (x_i | y_i) lassen sich als Punktwolke darstellen. → *Excel / Korrelation*

Im einfachsten Fall könnte man einen linearen Zusammenhang der Form $y = m\,x + b$ vermuten.
Nur eine Einflussgröße, es handelt sich um eine lineare Einfach-Regression.
Die Regressionsgerade $\hat{y} = m\,x + b$ ist die optimal passende Gerade durch die Punktwolke.
Das Symbol für Schätzwerte der Variablen y ist $\hat{\mathbf{y}}$.

1. Jede mögliche Gerade wird durch die beiden Parameter m und b festgelegt.
Wir haben aber 10 Messwerte. Die Berechnung von m, b ist also überbestimmt.
Zu jedem Datenpunkt (x_i | y_i) gibt es eine Abweichung vom Geradenpunkt $(\hat{x} \mid \hat{y})$.
Die Regressionsgerade mit der Funktionsgleichung $\hat{y} = mx + b$ wird so gewählt, dass die Summe der Abweichungsquadrate minimal wird ("Methode der kleinsten Quadrate").
Einzelne Abweichungen: $e_i = y_i - \hat{y}_i = y_i - (m\,x_i + b) = y_i - m\,x_i - b$
Einzelne Abweichungsquadrate: $(y_i - m\,x_i - b)^2$
Summe der Abw.-Quadrate: $A = \sum_{i=1}^{n} (y_i - mx_i - b)^2$ $\quad A_{\text{Error}} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} e_i^2$
(e_i errors, Residuen, Reste) m und b sind die gesuchten unbekannten Variablen

2. Die Funktion $A(m,b)$ liefert für jede Kombination m,b einen Wert A.
Das Minimum der Funktion $A(m,b)$ erhält man, wenn man die 1. Ableitung null setzt.
Hinweis: Seite 1.24

3. Die Funktion $A = \sum_{i=1}^{n} (y_i - mx_i - b)^2$ leitet man partiell nach m und nach b ab:

$$\begin{cases} A'(m) = \dfrac{\partial A}{\partial m} = \sum 2 \cdot (y_i - mx_i - b)^1 \cdot (-x_i) = 0 \\ A'(b) = \dfrac{\partial A}{\partial b} = \sum 2 \cdot (y_i - mx_i - b)^1 \cdot (-1) = 0 \end{cases} \qquad \frac{\partial y}{\partial x} \text{ sind partielle Ableitungen}$$

(-2) vor die Summe, ausmultiplizieren:
$$\begin{cases} 0 = -2\sum (x_i\,y_i - mx_i^2 - b\,x_i) \\ 0 = -2\sum (y_i - mx_i - b) \end{cases}$$

:(-2), Einzelsummen schreiben:
$$\begin{cases} 0 = \sum x_i y_i - m\sum x_i^2 - b\sum x_i \\ 0 = \sum y_i - m\sum x_i - nb \end{cases}$$

Dieses lineare Gleichungssystem lässt sich in Matrizen-Schreibweise formulieren und nach m und b auflösen.

$$\begin{cases} m\sum x_i^2 + b\sum x_i = \sum y_i x_i \\ m\sum x_i + nb = \sum y_i \end{cases} \Rightarrow \quad \begin{pmatrix} \sum x_i^2 & \sum x_i \\ \sum x_i & n \end{pmatrix} \cdot \begin{pmatrix} m \\ b \end{pmatrix} = \begin{pmatrix} \sum x_i y_i \\ \sum y_i \end{pmatrix}$$

Diese nennt man Normalgleichungen bzw. Normalgleichungs-System.
Achtung: $\Sigma x\,\Sigma y \neq \Sigma\, xy \qquad \Sigma\, x^2 \neq (\Sigma\, x)^2$

1.9 REGRESSIONSKOEFFIZIENTEN

Statt die Regressionskoeffizienten m und b mit dem obigen linearen Gleichungssystem zu bestimmen, werden in der Praxis die Regressionskoeffizienten oft direkt angegeben:

$$\begin{cases} 0 = \sum y_i x_i - m\sum x_i^2 - b\sum x_i \\ 0 = \sum y_i - m\sum x_i - nb \end{cases}$$

2.Zeile nach b auflösen: $b = \frac{1}{n}\sum y_i - \frac{m}{n}\sum x_i$

b in 1.Zeile einsetzen: $0 = \sum x_i y_i - m\sum x_i^2 - \left(\frac{1}{n}\sum y_i - \frac{m}{n}\sum x_i\right)\cdot\sum x_i$

ausmultiplizieren: $0 = \sum x_i y_i - m\sum x_i^2 - \frac{1}{n}\sum y_i \cdot \sum x_i + \frac{m}{n}\left(\sum x_i\right)^2$

m isolieren: $m\sum x_i^2 - \frac{m}{n}\left(\sum x_i\right)^2 = \sum x_i y_i - \frac{1}{n}\sum y_i \cdot \sum x_i$

m ausklammern: $m\left(\sum x_i^2 - \frac{1}{n}\left(\sum x_i\right)^2\right) = \sum x_i y_i - \frac{1}{n}\sum y_i \cdot \sum x_i$

: Klammer $m = \dfrac{\sum x_i y_i - \frac{1}{n}\sum y_i \cdot \sum x_i}{\sum x_i^2 - \frac{1}{n}\left(\sum x_i\right)^2}$

mit $\frac{n}{n}$ erweitern: $m = \dfrac{n\sum x_i y_i - \sum x_i \cdot \sum y_i}{n\sum x_i^2 - \left(\sum x_i\right)^2}$ $\qquad b = \frac{1}{n}\sum y_i - \frac{m}{n}\sum x_i$

Achtung: $\Sigma x\, \Sigma y \neq \Sigma\, xy$ $\qquad \Sigma\, x^2 \neq (\Sigma\, x)^2$

Fortsetzung Beisp. 1.2 Eine Messreihe → *Excel / A'–Regression*

I) Lösung über das System der Normalgleichungen:

$$\begin{pmatrix} \sum x_i^2 & \sum x_i \\ \sum x_i & n \end{pmatrix}\cdot\begin{pmatrix} m \\ b \end{pmatrix} = \begin{pmatrix} \sum x_i y_i \\ \sum y_i \end{pmatrix} \quad \Rightarrow \begin{pmatrix} 298,25 & 49,5 \\ 49,5 & 10 \end{pmatrix}\cdot\begin{pmatrix} m \\ b \end{pmatrix} = \begin{pmatrix} 105,75 \\ 26,5 \end{pmatrix}$$

$$\left.\begin{array}{l} 298,25m + 49,5b = 105,75 \\ 49,5m + 10b = 26,5 \end{array}\right| \begin{array}{l} (-) \\ \cdot 6,025 \end{array} \quad \Rightarrow 10,75b = 53,9125 \Rightarrow b = 5,015 \Rightarrow m = -0,478$$

II) Lösung über Formeln für Regressionskoeffizienten:

$$m = \frac{n\sum x_i y_i - \sum x_i \cdot \sum y_i}{n\sum x_i^2 - \left(\sum x_i\right)^2} = \frac{10\cdot 105,75 - 49,5\cdot 26,5}{10\cdot 298,25 - 49,5^2} = \frac{-254,25}{532,25} = -0,478$$

$$b = \frac{1}{n}\sum y_i - \frac{m}{n}\sum x_i = \frac{1}{10}\cdot 26,5 - \frac{-0,478}{10}\cdot 49,5 = 2,65 + 2,366 = 5,016$$

Es ergibt sich $\hat{y} = -0,478\,x + 5,015$ als beste Approximation (Näherung) für den vermuteten Zusammenhang. ist das ein signifikanter Zusammenhang?

Mit der Funktionsgleichung der Regressionsfunktion kann man Interpolationen durchführen:
Für die gegebene Bestrahlungszeit von $x = 5$ Minuten
schätzt man den Härtegrad auf $\hat{y}(5) = -0,478\cdot 5 + 5,015 = 2,626$.

1.10 KORRELATION

1. Die Regressionsgerade läuft durch die Punkte $(0 \mid b)$ und $(\bar{x} \mid \bar{y})$.
   Die Regressionskoeffizienten kann man auch mit den Varianzen herleiten:

$$\bar{x} = \frac{1}{n}\sum x_i,\ \bar{y} = \frac{1}{n}\sum y_i \ \Rightarrow\quad m = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \frac{s_{xy}}{s_x^2} = \frac{A_{xy}}{A_x} \quad \text{und} \quad b = \bar{y} - m\,\bar{x}$$

2. Varianzzerlegung.
   Sind die beobachteten Werte y_i, deren Mittelwert $\bar{y}$ und die Schätzwerte $\hat{y}_i$ bekannt, dann lassen sich drei Summen der Abweichungsquadrate bzw. Varianzen unterscheiden:
   (1) die gesamte Varianz der Beobachtungswerte: $\sum (y_i - \bar{y})^2$
   (2) die durch die Regression erklärte Varianz: $\sum (\hat{y}_i - \bar{y})^2$
   (3) die nicht durch die Regression erklärte Varianz: $\sum (y_i - \hat{y}_i)^2$
   Es lässt sich zeigen, dass gilt: $\sum (y_i - \bar{y})^2 = \sum (y_i - \hat{y}_i)^2 + \sum (\hat{y}_i - \bar{y})^2$

3. Das Bestimmtheitsmaß r^2 ist ein Maß für die Stärke des Zusammenhangs.
   Es gibt an, wie stark das Merkmal Y durch das Merkmal X bestimmt wird.

$$r^2 = \frac{\sum (\hat{y}_i - \bar{y})^2}{\sum (y_i - \bar{y})^2} = \frac{s_{\hat{y}}^2}{s_y^2} = \frac{\text{der durch die Regression erklärte Anteil der Varianz}}{\text{gesamte Varianz der Beobachtungswerte } y_i}$$

   □ $r^2 = 0{,}898 = 89{,}8\ \%$ bedeutet, dass 89,8 % der Varianz beim Härtegrad durch die Dauer der Bestrahlung erklärt werden kann.
   Man folgert (etwas unscharf), dass 89,8 % der Senkung des Härtegrads Y auf die Erhöhung der Bestrahlungsdauer X zurückzuführen ist (oder umgekehrt: X könnte auch von Y beeinflusst sein). 10,2 % ($= 1 - r^2$) der Ursachen bleiben unerklärt.

   Für r^2 gilt: $0 \le r^2 \le 1$.

4. Der Korrelationskoeffizient r ist ein weiteres Maß für die Stärke des Zusammenhangs.
   Speziell für lineare Regressionen verwendet man die Quadratwurzel aus r^2.

$$r = \pm\sqrt{r^2} = \pm\sqrt{\frac{\sum (\hat{y}_i - \bar{y})^2}{\sum (y_i - \bar{y})^2}}$$

   Mit Korrelationskoeffizienten vergleicht man Richtung und Stärke linearer Zusammenhänge.
   [PEARSON, KARL, ca. 1896, London]

   □ $r = -0{,}948 = -94{,}8\ \%$ heißt, dass zwischen der UV-Bestrahlung X und dem Härtegrad Y ein entgegengerichteter 94,8-prozentiger Zusammenhang besteht.

   Für r gilt: $-1 \le r \le 1$. Für r wählt man das Vorzeichen der Geradensteigung m.
   $r < 0$ => gegengerichtete Korrelation, wenn X zunimmt, nimmt Y ab.
   Man sagt auch "indirekte, umgekehrte, negative" Korrelation.
   $r > 0$ => positive Korrelation.

5. Funktionen als Grenzfälle
   Eine Funktion der Form $y = m\,x + b$ wäre eine 100-prozentige Korrelation.
   Wir erwarten hier nur einen statistischen Zusammenhang, also eine Korrelation von $r < 100\%$.
   Funktionen sind Grenzfälle statistischer Zusammenhänge, es sind Korrelationen mit $r = 1$.
   Natürlich gilt dann auch für das Bestimmtheitsmaß $r^2 = 1$.

→ *Excel / Korrelation*

1.10A INVERSE REGRESSIONSGERADEN

Die Berechnung von r bzw. r^2 lässt sich für **lineare** Zusammenhänge einfacher durchführen.

a) Die Zufallsvariablen X und Y sind eigentlich "gleichberechtigt". Mathematisch lässt sich nicht entscheiden, ob X von Y abhängt oder Y von X. Wir gehen prinzipiell von einer Interdependenz, also von einer gegenseitigen Abhängigkeit aus. Neben der Regressionsgeraden $\hat{y}(x)$ existiert demnach auch die inverse Regressionsgerade $\hat{x}(y) = m_{invers} \cdot y + b_{invers}$

b) Bei einem funktionellen Zusammenhang (100-prozentiger Korrelation) stimmen die beiden inversen Regressionsgeraden überein, die zugehörigen Geraden sind identisch.
In diesem Fall gilt: $y = m \cdot x + b$, nach x aufgelöst: $y - b = mx$ und
$\frac{y}{m} - \frac{b}{m} = x$ *und* $x = \frac{1}{m} y - \frac{b}{m}$; $\frac{1}{m}$ ist die inverse Steigung und $-\frac{b}{m}$ der x-Achsen-Abschnitt.

Für die Steigung einer inversen Geraden m_{inv} gilt also: $m_{inv} = \frac{1}{m}$ *oder* $m_{inv} \cdot m = 1$

c) Bei einem funktionellen Zusammenhang (100-prozentiger Korrelation) gilt für die Steigungen der beiden Regressionsgeraden: $m \cdot m_{invers} = 1 = 100\%$.
Das Produkt $m_{inv} \cdot m$ ist das oben dargestellte Bestimmtheitsmaß r^2 für lineare Regression.

d) In den Formeln für die Koeffizienten m_{inv} und b_{inv} der inversen Regressionsgeraden sind die Variablen x und y vertauscht:

$$m_{invers} = \frac{n\sum x_i y_i - \sum x_i \cdot \sum y_i}{n\sum y_i^2 - \left(\sum y_i\right)^2} \qquad b_{invers} = \frac{1}{n}\sum x_i - \frac{m_{invers}}{n}\sum y_i$$

Für das Bestimmtheitsmaß r^2 ergibt sich also

$$r^2 = m \cdot m_{invers} = \frac{n\sum x_i y_i - \sum x_i \cdot \sum y_i}{n\sum x_i^2 - \left(\sum x_i\right)^2} \cdot \frac{n\sum x_i y_i - \sum x_i \cdot \sum y_i}{n\sum y_i^2 - \left(\sum y_i\right)^2}$$

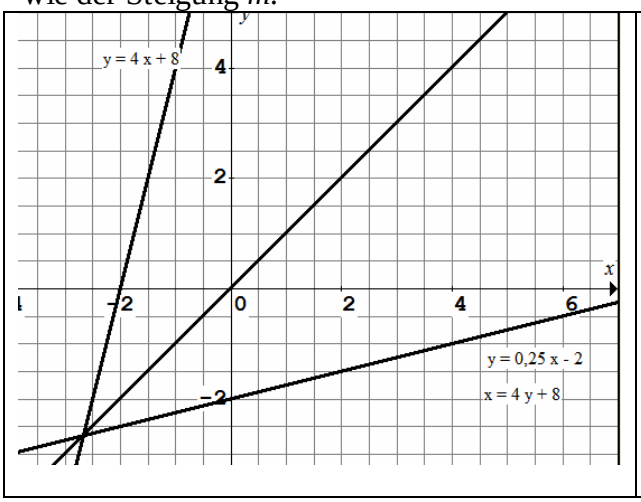
$$= \frac{\left(n\sum x_i y_i - \sum x_i \cdot \sum y_i\right)^2}{\left(n\sum x_i^2 - \left(\sum x_i\right)^2\right) \cdot \left(n\sum y_i^2 - \left(\sum y_i\right)^2\right)}$$

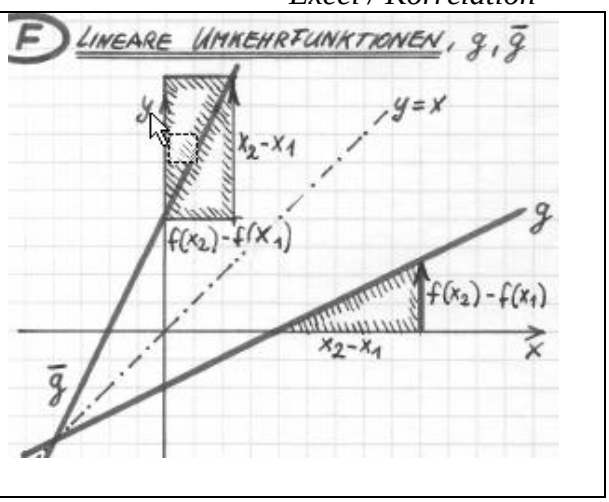
e) Der Korrelationskoeffizient ergibt sich demnach

$$r = \frac{n\sum x_i y_i - \sum x_i \cdot \sum y_i}{\sqrt{\left(n\sum x_i^2 - \left(\sum x_i\right)^2\right) \cdot \left(n\sum y_i^2 - \left(\sum y_i\right)^2\right)}}$$

Das richtige Vorzeichen ergibt sich automatisch, weil im Zähler derselbe Ausdruck steht wie der Steigung m.

→ *Excel / Korrelation*





1.11 FISHER-PRÜFMAß

1. Nullhypothese H_0

Die Nullhypothese lautet: es besteht kein Zusammenhang zwischen X und Y.
Wenn man nichts über den Umfang der Stichprobe weiß, dann verwirft man vereinfachend die Null-Hypothese, wenn $|r| > 0{,}5$, d.h. man bejaht den Zusammenhang zwischen X und Y.
Man akzeptiert also einen Zufallsbereich $-0{,}5 \le r \le +0{,}5$.
Die Grenze ist jedoch abhängig vom Stichprobenumfang.

2. Die Nullhypothese H_0 kann man nur ablehnen, wenn r^2 "groß genug" ist, nur dann ist der Zusammenhang statistisch gesichert.
r^2 ist groß genug, wenn der empirische Wert $xF_{empirisch}$ größer ist als der kritische Wert $xF_{critical}$ aus der Tabelle der F-Verteilung (*Tabelle F.3*).

Der Zusammenhang ist statistisch gesichert, wenn $xF_{empirisch} > xF_{critical}$.
xF_{crit} in der Tabelle ablesen, xF_{empir} berechnen

3. Berechnung von $xF_{empirisch}$ $x_{F,empir.}$

$$xF_{empirisch} = \frac{\sum(\hat{y}-\bar{y})^2}{\frac{\sum(\hat{y}-y_i)^2}{n-2}} = \frac{\text{Summe der erklärten Abweichungsquadrate}}{\frac{\text{Summe der nicht erklärten Abw.Quadrate}}{\text{Freiheitsgrade}}} = \frac{r^2}{1-r^2}\cdot(n-2)$$

a) Die viel einfachere zweite Formulierung ergibt sich aus der Varianzzerlegung (*siehe 1.12*)

$$\sum(y_i-\bar{y})^2 = \sum(y_i-\hat{y}_i)^2 + \sum(\hat{y}_i-\bar{y})^2 \quad \text{und} \quad \sum(y_i-\hat{y}_i)^2 = (1-r^2)\cdot\sum(y_i-\bar{y})^2$$

b) Nur n-2 Freiheitsgrade, weil $\bar{x}$ und $\bar{y}$ vorgegeben sind.

c) Das Prüfmaß aus der FISHER-Verteilung verwendet zwei Freiheitsgrade:
df1 = p = Anzahl der Einflussvariablen, hier p = 1, weil nur eine Einflussvariable.
df2 = ν = n – p – 1, hier ν = 10 – 1 – 1 = 8
Das Prüfmaß $xF_{crit}(0{,}05; 1; 8)$ findet man mit der Excel-Funktion FINV(0,05; 1; 8)
Die kritischen Werte $xF_{crit}(\alpha,p,\nu) = xF_{\alpha|p|\nu}$ sind hinsichtlich dieser Parameter tabelliert.

4. Die FISHER-Verteilung

Die Nullhypothese wird abgelehnt, wenn

$xF_{emp} > xF_{crit}$

mit Formel berechnen *aus Tabelle ablesen*

mit $xF_{empirisch} = \frac{r^2}{1-r^2}\cdot(n-2)$

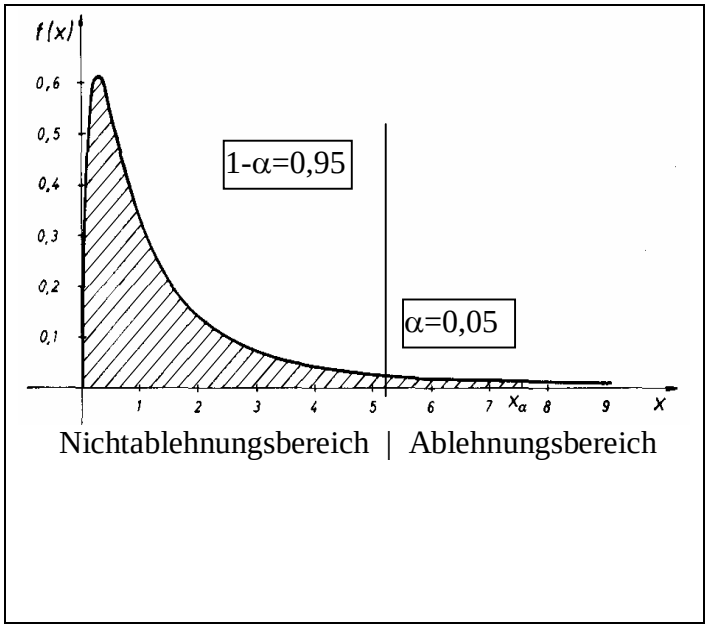
im Ablehnungsbereich für 95% oder 99% Aussagesicherheit liegt.

"empirisch" heißt: aus den Beobachtungswerten berechnet.
"kritisch" heißt die obere Integrationsgrenze für den 95% - bzw. 99%-Flächeninhalt oder die untere Grenze für die Irrtumswahrscheinlichkeit von $\alpha = 0{,}05$ (bzw. $\alpha=0{,}01$).

[FISHER, RONALD, Rothamsted, GB, 1918]

→ *Excel / Korrelation*

→ *Bleymüller S. 152-153,* → *Tabelle 7.3*



Nichtablehnungsbereich | Ablehnungsbereich

1.12 KORRELATIONSANALYSE

Aufgabe Korrelation (Bivariate Korrelationsanalyse)

gegeben: Beobachtungsdaten für die Zufallsvariablen X und Y, einen zusätzlichen x-Wert, Arbeitstabelle mit lückenhaften oder überflüssigen Spalten

gesucht:
a) Arbeitstabelle vervollständigen
b) Regressionskoeffizienten und Funktionsgleichung der Regressionsgeraden berechnen, die Zwischenergebnisse sind anzugeben.
Evtl. Regressionsgerade in gegebene Koordinatenebene einzeichnen.
c) Bestimmtheitsmaß und Korrelationskoeffizient berechnen und interpretieren.
d) Mit Hilfe des Prüfmaßes xF aus der FISHER-Verteilung prüfen, ob der Zusammenhang statistisch gesichert ist.
e) Eine Interpolation durchführen

Schritte:
a) Man benötigt $\sum x_i, \sum y_i, \sum x_i y_i, \sum x_i^2, \sum y_i^2$
b) *m, b, y(x)* mit den Formeln berechnen, *m,b* evtl. mit den Normalgleichungen Zeichnen durch die beiden Punkte $(0 \mid b)$ und $(\bar{x} \mid \bar{y})$.
c) r^2, r mit den Formeln berechnen, beide interpretieren
d) xF_{empir} berechnen, kritische Grenze $xF_{crit}(\alpha,1,\nu)$ in der Tabelle ablesen.
e) den zusätzlich gegebenen x-Wert in die Funktionsgleichung *y(x)* einsetzen, sinnvoll runden und gegebene Einheit (Benennung) angeben.

Umformungen für das FISHER-PRÜFMASS

$$xF_{empirisch} = \frac{\frac{\sum(\hat{y}-\bar{y})^2}{\sum(\hat{y}-y_i)^2}}{n-2} = \frac{\text{Summe der erklärten Abweichungsquadrate}}{\text{Summe der nicht erklärten Abw.Quadrate}} = \frac{r^2}{1-r^2}\cdot(n-2)$$

Freiheitsgrade

Ohne Berücksichtigung der Freiheitsgrade:

$$\frac{\sum(\hat{y}-\bar{y})^2}{\sum(\hat{y}-y_i)^2} = \frac{r^2}{1-r^2}$$ dies ergibt sich aus der Varianzzerlegung:

$$\sum(y_i-\bar{y})^2 = \sum(y_i-\hat{y}_i)^2 + \sum(\hat{y}_i-\bar{y})^2$$

$$\Rightarrow \sum(y_i-\hat{y}_i)^2 = \sum(y_i-\bar{y})^2 - \sum(\hat{y}_i-\bar{y})^2$$

a) Bestimmtheitsmaß $r^2 = \frac{\sum(\hat{y}_i-\bar{y})^2}{\sum(y_i-\bar{y})^2}$ $\Rightarrow r^2\cdot\sum(y_i-\bar{y})^2 = \sum(\hat{y}_i-\bar{y})^2$

b) Ausdruck einsetzen:

$$\Rightarrow \sum(y_i-\hat{y}_i)^2 = \sum(y_i-\bar{y})^2 - r^2\cdot\sum(y_i-\bar{y})^2 = (1-r^2)\cdot\sum(y_i-\bar{y})^2$$

$$\Rightarrow 1-r^2 = \frac{\sum(y_i-\hat{y}_i)^2}{\sum(y_i-\bar{y})^2}$$

c) Definition des FISHER-Prüfmaßes:

$$\frac{\sum(\hat{y}-\bar{y})^2}{\sum(\hat{y}-y_i)^2} = \frac{r^2}{1-r^2} = \frac{\frac{\sum(\hat{y}_i-\bar{y})^2}{\sum(y_i-\bar{y})^2}}{\frac{\sum(y_i-\hat{y}_i)^2}{\sum(y_i-\bar{y})^2}}$$

1.13 MULTIPLE REGRESSION

1. Multiple lineare Regression

Bisher haben wir Zusammenhänge zwischen zwei Variablen untersucht, zwischen dem Einflussfaktor x und der beeinflussten Variablen y. Wenn man mehr als eine Einflussvariable in die Regressionsanalyse einbezieht, arbeitet man mit der multiplen Regressionsanalyse:

Wir haben *p* Einfluss-Parameter (Einflussfaktoren) x_k mit k = 1; 2; … ; p

Ist eine der Einflussvariablen die Zeit *x*, dann handelt es sich um eine multiple Trendanalyse.

Wir betrachten nur linearen multiple Zusammenhänge.

Die Regressionsgleichung hat also die Form:

$y = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3 + ...$ $\quad y = b_0 + b_1 u + b_2 v + b_3 w + ...$

mit den p Datenreihen (Vektoren) ***u, v, w, …***, den Regressionskoeffizienten b_i.

Die Regressionskoeffizienten b_i berechnet man mit Hilfe der Matrizenrechnung (siehe unten).

2. Regressionsanalysen führt man mit Software-Programmen durch

Weit verbreitet sind die Programme von SPSS und SAS.

Analysefunktionen sind Hauptbestandteile der Business-Intelligence-Software mit den Marktführern: Hyperion (Oracle), Cognos, SPSS (IBM), Business Objects (SAP), SAS.

Eine einfache multiple Regressionsanalyse liefern die Analysefunktionen von Excel (seit 1995).

Add-In "Analyse-Funktionen" aktivieren: /Extras /Add-In-Manager /Analyse-Funktionen

→ Excel / Korrelation

Beispiel 1.3

In einer Studie soll untersucht werden, wie die drei Einflussfaktoren:
zeitliche Entwicklung ***x*** [Quartale], Schulung der Mitarbeiter im Vertrieb ***u*** [Stunden/Monat] und Aufwand zur Bindung der Kunden (CRM) **v** [€/Kunde] die Umsatzrendite ***y*** [%] beeinflussen.
Für 12 Quartale (3 Jahre) sind die Datenreihen zusammengestellt. *→ Excel / Multiple*

	A	B	C	D
1	Zeitraum	Schulung	CRM-Aufwand	Umsatz-rendite
2	x	u	v	y
3	Quartale	Std/Monat	€ / Kunde	%
4	1	16	650	4,3
5	2	13	700	4,5
6	3	14	690	4,3
7	4	15	720	5,3
8	5	13	780	6,2
9	6	16	740	5,3
10	7	18	650	4,5
11	8	14	760	4,8
12	9	17	780	6,1
13	10	14	660	4,6
14	11	15	730	5,2
15	12	15	720	5,0

AUSGABE: ZUSAMMENFASSUNG

Regressions-Statistik	
Multipler Korrela	0,870231416
Bestimmtheitsm	0,757302717
Adjustiertes Bes	0,666291235
Standardfehler	0,371495355
Beobachtungen	12

ANOVA

	Freiheitsgrade (a	dratsummen (	Quadratsumn	Prüfgröße (F)	F krit
Regression	3	3,445096275	1,14836542	8,32095803	0,00766226
Residue	8	1,104070392	0,1380088		
Gesamt	11	4,549166667			

	Koeffizienten	Standardfehler	t-Statistik	P-Wert	Untere 95%	Obere 95%
Schnittpunkt	-5,057320751	2,385792073	-2,11976593	0,06684954	-10,5589707	0,44432919
X Variable 1	0,00425843	0,033311948	0,12783492	0,90143464	-0,07255911	0,08107597
X Variable 2	0,085901754	0,077135867	1,11364217	0,29777085	-0,09197399	0,2637775
X Variable 3	0,01223699	0,002606995	4,69390569	0,00155366	0,00622524	0,01824874

AUSGABE: RESIDUENPLOT

Beobachtung	Schätzung für Y	Residuen	e(i)²	(e(i)-e(i-1))²
1	4,275409359	0,024590641	0,0006047	
2	4,633812037	-0,13381204	0,01790566	0,02509141
3	4,601602319	-0,30160232	0,09096396	0,02815358
4	5,058872209	0,241127791	0,05814261	0,29455597
5	5,625546542	0,574453458	0,32999678	0,111106
6	5,398030627	-0,09803063	0,00961	0,45223484
7	4,472763448	0,027236552	0,00074183	0,01569187
8	5,479483783	-0,67948378	0,46169821	0,49945363
9	5,986187278	0,113812722	0,01295334	0,62931934
10	4,264301625	0,335698375	0,1126934	0,04923324
11	5,211051122	-0,01105112	0,00012213	0,12023521
12	5,09293965	-0,09293965	0,00863778	0,00670573
			1,10407039	2,23178083

Interkorrelation				
	Zeitraum	Schulung	CRM-Aufw.	Ums.Rendite
Zeitraum	1			
Schulung	0,18040061	1		
CRM-Aufw.	0,2661668	-0,2165649	1	
Ums.Rendite	0,29601452	0,01838982	0,84512878	1

$$DW_1 = \frac{\sum_{i=2}^{n}(e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2}$$

1.14 ANALYSE-AUSSAGEN

Wir beschränken uns auf die gelb unterlegten Werte.

1. Freiheitsgrade:

Anzahl der Beobachtungen	n = 12	
Varianz-Freiheitsgrade	$\nu = n - 1 = 11$	
Anzahl der Einflussvariablen	p = 3	(df1)
(FISHER)-Freiheitsgrade	$\nu_{FISHER} = n - p - 1 = 8$	(df2)

2. Bestimmtheitsmaß $r^2 = 0{,}757$

Das adjustierte Bestimmtheitsmaß berücksichtigt den Stichprobenumfang *n* und die Anzahl der Einflussvariablen, es ist deshalb erwartungstreuer als das gewöhnliche Bestimmtheitsmaß.

66,6% der Veränderung des Umsatzrendite ist auf die drei Einflussfaktoren *x, v, w* zurückzuführen; 33,4% sind unerklärte Ursachen.

$$r_{adj}^{\ 2} = 1 - \frac{MS_{Residuen}}{MS_{gesamt}} = 0{,}666$$ (MS = Mittlere Summe der Abweichungsquadrate) siehe Excel / Multiple

Der Multiple Korrelationskoeffizient ist einfach nur $\sqrt{r^2}$, er bleibt außer Betracht.

3. Regressionskoeffizienten und Regressionsfunktion

$b_0 = -5{,}057$ $b_1 = 0{,}004$ $b_2 = 0{,}086$ $b_3 = 0{,}012$

$\hat{y} = -5{,}057 + 0{,}004\,x + 0{,}086\,u + 0{,}012\,v$

Es sind auch Grenzen angegeben, in denen die Koeffizienten mit 95% Sicherheit liegen.
Z.B. liegt der Koeffizient für die "Schulung" im Intervall: $-0{,}092 < b_2 < 0{,}264$

4. Interpolationen und Prognosen

Für gegebene Szenarien lassen sich die Werte $\hat{y}$ schätzen:

Bestimmen Sie den besten Schätzwert für die Umsatzrendite im Quartal 17, bei einem Schulungsaufwand von 13 Std/Monat und einem CRM-Aufwand von 700 €/Kunde.

$\hat{y}(17;13;700) = -5{,}057 + 0{,}004 \cdot 17 + 0{,}086 \cdot 13 + 0{,}012 \cdot 700 = 4{,}698\ \% \approx 4{,}7\%$

5. FISHER-Testgröße *xF* zeigt, ob die gesamte Regression statistisch gesichert ist.

Die Regression ist statistisch gesichert, wenn $xF_{empir} > xF_{crit} = xF_{Tabellenwert}$

□ "Prüfgröße F" = $xF_{empir} = 8{,}321$ $xF_{crit}(\alpha=0{,}05;\ df1=3,\ df2=8) = xF_{0{,}05\,|\,3\,|\,8} = 4{,}066$

Die Regression ist statistisch gesichert auf dem 95%-Sicherheitsniveau, weil 8,321 > 4,066

Sie ist sogar auf dem 99%-Niveau gesichert, weil 8,321 > 7,591 (hochsignifikant).

Berechnung von xF_{empir} siehe Excel / Multiple

6. Die STUDENT-t-Testgröße zeigt, welche Einflussvariablen einen signifikanten Beitrag für den untersuchten Zusammenhang liefern. Die empirischen Testgrößen $t_{b(k)}$ werden angezeigt.

$$\textit{für } b_1\text{:}\quad t_{empir|b_k} = \frac{|b_k|}{SE_{b_k}} \quad \text{mit} \quad SE_{b_k} = \sqrt{\frac{\sum e^2}{n\text{-}p\text{-}1} \cdot \beta_{kk}}$$

β_{kk} sind die Elemente der Matrix $(\boldsymbol{V}^T\boldsymbol{V})^{-1}$ siehe unter $\boldsymbol{V}^T$-Regression, Standardfehler SE außer Betracht.
Vgl. Bleymüller S.168

Der Einflussfaktor x_i liefert einen signifikanten Beitrag für die Regression

wenn gilt: $|t_{empirisch}| > t_{critical} = t_{Tabellenwert}$ *Tabelle 7.5a* für $\alpha = 0{,}05$, $\nu = n - p - 1$

□ hier: $t_{0{,}05\,|\,8}$ = 4,694 > 1,86 => der CRM-Aufwand liefert einen signifikanten Beitrag.
1,114 < 1,86 => der Schulungsaufwand liefert keine sign. Beitrag.
0,128 < 1,86 => die Zeitentwicklung liefert keinen sign. Beitrag.

[GOSSET, WILLIAM, "STUDENT", Dublin, Irland, 1908]

1.15 INTER- UND AUTO-KORRELATION

7. Interkorrelation:

Wenn zwei Einflussvariablen x_j , x_k selbst miteinander korrelieren, werden die Ergebnisse der Regressionsanalyse verzerrt.

Interkorrelation liegt vor, wenn für den Korrelationskoeffizient gilt $|r_{jk}| > 0{,}5$.

Die Korrelationsmatrix zeigt paarweise die Korrelationskoeffizienten r_{jk}.

□ hier: es besteht keine signifikante Interkorrelation, weil alle $r_{jk} < 0{,}5$.

Zusätzlich zeigen die Korrelationskoeffizienten zur beeinflussten Variablen y in ähnlicher Weise wie die STUDENT-t-Testgrößen, welche Einflussvariablen einen vergleichsweise höheren Beitrag für den untersuchten Zusammenhang liefern.

8. Autokorrelation: "selbst" Suche nach Zyklen (Astronomie, Wirtschaft, Biochemie)

Man spricht von Autokorrelation, wenn die Residuen e_i , die aufeinander folgen, miteinander korrelieren. $e_i = y_i - \hat{y}_i$

Im Extremfall A wechseln die Residuen e_i einander ab, zeigen also ein zeitliches Muster der Form + – + – + –. Im Extremfall B folgen positiven bzw. negative Residuen regelmäßig aufeinander, etwa in der Form + + + + – – – – – + + + + + .

Je höher diese Korrelation zwischen aufeinander folgender Beobachtungswerten ist, desto größer sind die Schätzfehler aus der Regressionsanalyse; man sollte dann andere Methoden benutzen, die diese auto-regressive Eigenschaft der Datenreihen berücksichtigen.(ARIMA)

Das DURBIN-WATSON-Maß DW_1 ist ein Maß für die Stärke des Zusammenhangs aufeinander folgender Residuen, d.h. für die Stärke der Autokorrelation.

$$DW_1 = \frac{\sum_{i=2}^{n} (e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2} \qquad DW_k = \frac{\sum_{i=k+1}^{n} (e_i - e_{i-k})^2}{\sum_{i=1}^{n} e_i^2}$$

Die Summen $\sum_{i=1}^{n} e_i^2$ und $\sum_{i=2}^{n} (e_i - e_{i-1})^2$ sind in der Ausgabe Residuenplot.

Liegt das DW_1-Maß nahe bei der Zahl 2, dann besteht keine Autokorrelation.

Die DW_1-Tabelle zeigt Intervalle um den Wert 2 für "keine wesentliche Autokorrelation". Liegt das DW_1-Maß oberhalb dieses Intervalls, dann besteht Autokorrelation vom Typ A, liegt das DW_1-Maß unterhalb dieses Intervalls, dann besteht Autokorrelation vom Typ B.

Wenn man untersuchen möchte, ob sich Beobachtungswerte nach jeweils 3 Perioden signifikant auswirken, dann benutzt man DW_3, d.h. dann arbeitet man in der o.g. Formel mit den Abweichungen $e_i - e_{i-3}$.

[DURBIN, JAMES, Cambridge, GB, 1950 WATSON, GEOFFREY, Princeton, USA, 1950]

□ *hier: 2,02 ∈ [1,32 ; 2,68] für n = 12.*

1.16 AUFGABE MULTIPLE

Aufgabe Multiple (Multiple lineare Regressionsanalyse)

gegeben: Ein Analyseblatt wie auf Seite 1.13, jedoch keine gelben Markierungen,
ein Datensatz für ein Szenario, Sicherheitsgrad bei Nr.5 und 6 beachten!

gesucht / *Schritte*:

1. Adjustiertes Bestimmtheitsmaß, Interpretation, *Mit welchem Prozentsatz wird die Größe y durch die Einflussvariablen erklärt?*
   Ablesen, interpretieren
2. Gleichung der Regressionsfunktion
   Koeffizienten ablesen, als Funktionsgleichung schreiben
3. Interpolation oder Prognose für das gegebene Szenario
   Die gegebenen Werte (den Datensatz) in Funktionsgleichung einsetzen
4. Ist das Regressionsmodell bei α = … statistisch gesichert ? (oder bei $1-\alpha$ = …)
   Kann man die Nullhypothese "H_0 Zusammenhang" ablehnen?
   F-Prüfgröße ablesen: vergleichen $xF_{empirisch} > xF_{crit}$ *? Folgerung.*
5. Liefert die Einflussvariable x_k einen signifikanten Beitrag für den behaupteten Zusammenhang? Ist der Einfluss von Faktor x_k signifikant ?
   Schließen Sie mit einem Sicherheitsgrad von $1-\alpha$ =… [%] oder bei α = …
   t-Prüfgrößen ablesen und vergleichen mit t_{crit} *. Folgerung. Tabelle 7.5a*
6. Prüfen auf Interkorrelation zwischen den Einflussfaktoren.
   Wie stark hängen die Einflussfaktoren gegenseitig voneinander ab?
   In der Korrelationsmatrix die Koeffizienten r_{jk} ablesen und *interpretieren.*
7. Prüfen auf Autoregression hintereinander folgender Beobachtungswerte.

   Die Summen $\sum_{i=1}^{n} e_i^2$ *und* $\sum_{i=2}^{n} (e_i - e_{i-1})^2$ *ablesen, dividieren => DW₁.*

   Mit der Tabelle prüfen, ob DW-Maß im erforderlichen Intervall liegt.

1.17 A' - Regression

Regressionsfunktionen der Form $\hat{y} = a\,\varphi(x) + b$ sind Regressionsfunktionen, die man wie einfache Regressionsgeraden entwickeln kann: wenn $\varphi(x) = x$ dann ist $\hat{y} = a\,x + b$.
Bevor man eine Regressionsanalyse durchführt, wählt man eine passende Ansatzfunktion $\varphi(x)$, also beispielsweise $\varphi(x) = e^x$, $\varphi(x) = \frac{1}{x^2}$, $\varphi(x) = \sqrt[3]{x}$, $\varphi(x) = \ln x$

Die Regressionsfunktion mit der Funktionsgleichung $\hat{y} = a\,\varphi(x) + b$ wird so gewählt, dass die Summe der Abweichungsquadrate minimal wird ("Methode der kleinsten Quadrate").
Wir nennen dieses Verfahren "A'-Regression";
üblich ist auch: Regression mit linearisierbaren Modellfunktionen.
Ganz analog der Herleitung in Abschnitt 1.7 folgt:

Einzelne Abweichungen: $e_i = y_i - \hat{y}_i = y_i - (a\,\varphi(x_i) + b) = y_i - a\,\varphi(x)_i - b$

Einzelne Abweichungsquadrate: $(y_i - a\,\varphi(x)_i - b)^2$

Summe der Abw.-Quadrate: $A = \sum_{i=1}^{n} (y_i - a\,\varphi(x_i) - b)^2$

a und b sind die gesuchten unbekannten Koeffizienten

Die Ableitungen sind

$$\begin{cases} A'(a) = \dfrac{\partial A}{\partial a} = \sum 2 \cdot (y_i - a\,\varphi(x_i) - b) \cdot (-\varphi(x_i)) = 0 \\ A'(b) = \dfrac{\partial A}{\partial b} = \sum 2 \cdot (y_i - a\,\varphi(x_i) - b) \cdot (-1) = 0 \end{cases}$$

$$\begin{cases} 0 = \sum y_i \cdot \varphi(x_i) - a \sum (\varphi(x_i))^2 - b \sum \varphi(x_i) \\ 0 = \sum y_i - a \sum \varphi(x_i) - nb \end{cases}$$

Auch in den Normalgleichungen werden die Ausdrücke x durch $\varphi(x_i)$ ersetzt und es entsteht das lineare Gleichungssystem:

$$\begin{cases} a \sum (\varphi(x_i))^2 + b \sum \varphi(x_i) = \sum y_i\,\varphi(x_i) \\ a \sum \varphi(x_i) + nb = \sum y_i \end{cases}$$

Für die Regressionskoeffizienten kann man dann schreiben:

$$a = \frac{n \sum y_i \cdot \varphi(x_i) - \sum y_i \cdot \sum \varphi(x_i)}{n \sum (\varphi(x_i))^2 - \left(\sum \varphi(x_i)\right)^2} \qquad b = \frac{1}{n} \sum y_i - \frac{a}{n} \sum \varphi(x_i)$$

Zum Beispiel gilt für die Regressionskoeffizienten a und b
mit der Ansatzfunktion $\hat{y} = \frac{a}{\sqrt{x}} + b = a \cdot \frac{1}{\sqrt{x}} + b$

$$a = \frac{n \sum y_i \cdot \frac{1}{\sqrt{x_i}} - \sum y_i \cdot \sum \frac{1}{\sqrt{x_i}}}{n \sum \left(\frac{1}{\sqrt{x_i}}\right)^2 - \left(\sum \frac{1}{\sqrt{x_i}}\right)^2} \qquad b = \frac{1}{n} \sum y_i - \frac{a}{n} \sum \frac{1}{\sqrt{x_i}}$$

1.18 REGRESSIONSANALYSE

Beisp. 1.4 Regressionsanalyse mit der Ansatzfunktion $\hat{y} = a \ln x + b$

Ein Unternehmen zeichnet über die Zeiträume x_i (12 Monate) die Absatzmengen y_i eines bestimmten Produktes auf. Das Unternehmen erwartet eine gewisse Marktsättigung und benutzt daher für die Bestimmung des Trends und für Prognosen die Ansatzfunktion $\hat{y} = a \ln x + b$.
Zu bestimmen sind:

a) Die Ableitungen $\frac{\partial A}{\partial a}$ und $\frac{\partial A}{\partial b}$ für die Summe der Abweichungsquadrate A.
b) Die Normalgleichungen und das lineare Gleichungssystem mit den gegebenen Zahlenwerten
c) Die Formeln für die Regressionskoeffizienten (Herleitung nicht erforderlich).
d) Die Funktionsgleichung der Regressionsfunktion, das ist eine Trendfunktion.
e) Das Bestimmtheitsmaß und dessen Interpretation.
f) Die Prognosen für den 13. und 14. Monat.

a) $A = \sum_{i=1}^{n} (y_i - a \ln x_i - b)^2$ $\quad \begin{cases} A'(a) = \frac{\partial A}{\partial a} = \sum 2 \cdot (y_i - a \ln x_i - b) \cdot (-\ln x_i) = 0 \\ A'(b) = \frac{\partial A}{\partial b} = \sum 2 \cdot (y_i - a \ln x_i - b) \cdot (-1) = 0 \end{cases}$

→ *Excel / A'–Regression*

b) $\begin{cases} a \sum (\ln x_i)^2 + b \sum \ln x_i = \sum y_i \ln x_i \\ a \sum \ln x_i + nb = \sum y_i \end{cases}$ $\quad \begin{cases} 39,575\,a + 19,987\,b = 65,66 \\ 19,987\,a + 12\,b = 35 \end{cases}$

c) $a = \frac{n \sum y_i \cdot \ln x_i - \sum y_i \cdot \sum \ln x_i}{n \sum (\ln x_i)^2 - (\sum \ln x_i)^2}$ $\qquad b = \frac{1}{n} \sum y_i - \frac{a}{n} \sum \ln x_i$

d) $\boxed{\hat{y} = 1,172 \ln x + 0,965}$

e) Bestimmtheitsmaß $r^2 = \frac{s^2_{erklärt}}{s^2_{gesamt}} = \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} = \frac{8,630}{9,917} = 0,870$ **Achtung nicht-linear**

87% der Änderung der Absatzmenge ist auf die Zeitentwicklung zurückzuführen.
13% der Ursachen werden durch die Regressionsanalyse nicht erklärt. Formeln 1.3

f) 13.Monat 3971 kg, 14.Monat 4057 kg

Aufgabe A'–Regression mit $\hat{y} = a\,\varphi(x) + b$

Gegeben: Wertetabelle $(x_i \mid y_i)$, n = 5..12, Teile der Arbeitstabelle, Ansatzfunktion $\varphi(x)$
zusätzlicher Wert für eine Interpolation

Gesucht:

a) Die Ableitungen $\frac{\partial A}{\partial a}$ und $\frac{\partial A}{\partial b}$ für die Summe der Abweichungsquadrate A.
b) Die Normalgleichungen, das lineare Gleichungssystem mit den gegebenen Werten.
c) Die Formeln für die Regressionskoeffizienten (Herleitung nicht erforderlich).
d) Die Funktionsgleichung der Regressionsfunktion.
e) Das Bestimmtheitsmaß und dessen Interpretation. Formeln 1.3
f) Ein Prognose- oder Interpolationswert.

Schritte: genau wie in Beispiel 1.4 gezeigt

1.19 TRANSPONIERTE VANDERMONDE-MATRIX

Beisp. 1.5

x_i	y_i
1	6
2	6,8
3	10
4	10,5

Vier Messungen ergaben nebenstehende Wertetabelle.
Die Ansatzfunktion sei eine Gerade, $\hat{y} = a\,x + b$ oder $\hat{y} = a_0 + a_1\,x$.
Wir bestimmen die Regressionskoeffizienten mit dem analytischen Ansatz über die Ableitungen der Summe der Abweichungsquadrate:

→ *Excel / V^T-Regression*

Als Normalgleichungssystem ergibt sich:

$$\begin{pmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{pmatrix} \cdot \begin{pmatrix} b \\ m \end{pmatrix} = \begin{pmatrix} \sum y_i \\ \sum x_i y_i \end{pmatrix} \Rightarrow \begin{pmatrix} 4 & 10 \\ 10 & 30 \end{pmatrix} \cdot \begin{pmatrix} a_0 \\ a_1 \end{pmatrix} = \begin{pmatrix} 33,3 \\ 91,6 \end{pmatrix}$$

Mit der Lösung $a_1 = m = 1{,}67$. $a_0 = b = 4{,}15$. $\hat{y} = 1{,}67\,x + 4{,}15$
(Spalten getauscht, wegen der Index-Reihenfolge in der VANDERMONDE-Matrix)
(Zeilen getauscht, spielt für die Lösung keine Rolle)

Mit der Vandermonde-Matrix könnte man aus 2 Wertepaaren die Geradengleichung bestimmen.
Kennt man 2 Stützpunkte $(x_i \mid y_i)$ dann lassen sich die Koeffizienten a_0, a_1 bestimmen:

$$\begin{cases} a_0 + a_1 x_0 = y_0 \\ a_0 + a_1 x_1 = y_1 \end{cases} \Leftrightarrow \begin{pmatrix} 1 & x_0 \\ 1 & x_1 \end{pmatrix} \cdot \begin{pmatrix} a_0 \\ a_1 \end{pmatrix} = \begin{pmatrix} y_0 \\ y_1 \end{pmatrix} \Leftrightarrow \boldsymbol{V} \cdot \boldsymbol{a} = \boldsymbol{y}$$

Das geht problemlos, wenn $\boldsymbol{V}$ quadratisch ist. Wir haben aber eine Matrix $\boldsymbol{V}$ mit 4 Zeilen.
Unser lineares Gleichungssystem ist überbestimmt:

$$\begin{cases} a_0 + a_1 x_0 = y_0 \\ a_0 + a_1 x_1 = y_1 \\ a_0 + a_1 x_2 = y_2 \\ a_0 + a_1 x_3 = y_3 \end{cases} \Leftrightarrow \begin{pmatrix} 1 & x_0 \\ 1 & x_1 \\ 1 & x_2 \\ 1 & x_3 \end{pmatrix} \cdot \begin{pmatrix} a_0 \\ a_1 \end{pmatrix} = \begin{pmatrix} y_0 \\ y_1 \\ y_2 \\ y_3 \end{pmatrix}$$

Das führt in aller Regel zu einem Widerspruch.
Man kann das lineare Gleichungssystem aber mit einem minimalen Fehler A_{min} lösen.

Durch Multiplikation der Gleichung $\boldsymbol{V} \cdot \boldsymbol{a} = \boldsymbol{y}$ mit der transponierten Matrix $\boldsymbol{V}^T$ entsteht die Gleichung $\boldsymbol{V}^T\boldsymbol{V}\,\boldsymbol{a} = \boldsymbol{V}^T\,\boldsymbol{y}$.
Die linke Seite $\boldsymbol{V}^T\boldsymbol{V}$ ist eine quadratische, sogar symmetrische Matrix.
Mit der Matrizengleichung $\boldsymbol{V}^T\boldsymbol{V}\,\boldsymbol{a} = \boldsymbol{V}^T\,\boldsymbol{y}$ kann man also die Koeffizienten $\boldsymbol{a}$ berechnen.

→ *Excel / V^T-Regression*

				$\boldsymbol{V}$			$\boldsymbol{y}$
				1	1		6
				1	2		6,8
				1	3		10
$\boldsymbol{V}^T$				1	4		10,5
1	1	1	1				
1	2	3	4				
				$\boldsymbol{V}^T\boldsymbol{V}$			$\boldsymbol{V}^T\boldsymbol{y}$

⇒

				1	1		6
				1	2		6,8
				1	3		10
$\boldsymbol{V}^T$				1	4		10,5
1	1	1	1	4	10		33,3
1	2	3	4	10	30		91,6

Es ergeben sich dieselben Normalgleichungen,
man erkennt wie bei der Multiplikation mit $\boldsymbol{V}^T$ nacheinander die Summen $n, \sum x_i, \sum x_i^2, \sum y_i, \sum x_i y_i$ entstehen.

Die Regressionsanalyse mit Hilfe der Gleichung $\boldsymbol{V}^T\boldsymbol{V}\,\boldsymbol{a} = \boldsymbol{V}^T\,\boldsymbol{y}$ nennen wir V^T-Regression, sie ist viel weitreichender als die A'–Regression über die Ableitungen *A'(m,b)*.

1.20 V^T-REGRESSION

Übersicht

1. **A'**–Regression
   1.1 Lineare Regression im engeren Sinne.Berechnung einer Regressionsgeraden $\hat{y}(x) = m\,x + b$
   Man erstellt eine Arbeitstabelle für $x_i y_i$, x_i^2,bestimmt die Summen und löst $\boldsymbol{A\,a = y}$
   1.2 Regressionsmodelle der Form $\hat{y}(x) = a\varphi(x) + b = a_0 + a_1\,\varphi(x)$
   Man erstellt eine Arbeitstabelle für $\varphi(x_i)$, $\varphi(x_i)\cdot y_i$, $\varphi(x_i)^2$,bestimmt die Summen und löst $\boldsymbol{A\,a = y}$. Man erhält eine Regressionsfunktion.
2. $\mathbf{V^T}$-Regression
   Damit können fast alle Regressionsmodelle bearbeitet werden.
   Regressionsmodelle i.w.S. haben die Form $\hat{y}(x) = a_0 + a_1\varphi_1(x) + a_2\,\varphi_2(x) + \ldots + a_k\,\varphi_k(x)$
   mit fast beliebigen Ansatzfunktionen $\varphi(x)$.
   Man entwickelt eine VANDERMONDE-Matrix $\boldsymbol{V_\varphi}$ und löst die Gleichung $\boldsymbol{V^T V\,a = V^T\,y}$ nach den Regressionskoeffizienten $\boldsymbol{a}$ auf.
   2.1 Regressionsfunktionen mit $k > 1$, z.B. $\hat{y}(x)=a_0 + a_1\,x + a_2\,x^2$
   2.2 Multiple Regressionen, das sind Regressionsmodelle mit mehr als eine Einflussvariable.
   z.B. $\hat{y}(x)= a_0 + a_1\cdot u + a_2\cdot v + a_3\cdot x$
3. Echte nichtlineare Regression
   z.B. $\hat{y} = a\cdot e^{bx}$, $\hat{y} = a\cdot\sin(bx)$, $\hat{y} = a\cdot x^b$ verwenden wir in Thema 2 "Logistischer Trend".

Beisp. 1.6

Fünf Punkte sind gegeben (0 | 3) (2 | 5) (3 | 5) (5 | –3) (6 | 0)
Das Ausgleichpolynom 2. Grades ist zu bestimmen, eine Regressionsparabel,
eine Funktion nach dem Regressionsmodell $\hat{y}(x)=a_0 + a_1\,x + a_2\,x^2$ ist zu bestimmen.

→ *Excel / V^T-Regression*

$$\boldsymbol{V} = \begin{pmatrix} 1 & x_0 & x_0^2 \\ 1 & x_1 & x_1^2 \\ & \cdots & \\ 1 & x_k & x_k^2 \end{pmatrix} \qquad \boldsymbol{V\cdot a = y} \quad \Rightarrow \quad \hat{y} = 3{,}550 + 1{,}017x - 0{,}325\,x^2$$

Mit Hilfe der Regressionsfunktion kann man die beste Schätzung für $x = 4$ bestimmen:
$\hat{y}\,(4) = 3{,}550 + 1{,}017\cdot 4 - 0{,}325\cdot 4^2 = 2{,}418$
In Excel: Punkt aus der Punktwolke markieren, dann Rechtsklick, "Trendlinie einfügen"

Beisp. 1.7

Der Einfluss der beiden Variablen U und X auf die Größe Y soll untersucht werden.
Es liegt dazu die nebenstehende Wertetabelle vor.

u_i	x_i	y_i
100	3	142
95	2	138
102	4	167
128	6	182
125	8	191
102	9	179
124	9	190
107	10	178
119	11	194

Gesucht ist ein Regressionsfunktion nach dem multiplen Modell $\hat{y} = a_0 + a_1\,u + a_2\,x$.

$$\boldsymbol{V} = \begin{pmatrix} 1 & u_0 & x_0 \\ 1 & u_1 & x_1 \\ & \cdots & \\ 1 & u_k & x_k \end{pmatrix} \cdot \boldsymbol{a = y} \quad \Rightarrow \quad \hat{y} = 60{,}802 + 0{,}767\,u + 3{,}961\,x$$

Die beste Schätzung für das Szenario u = 110, x = 7 ist
$\hat{y}(110\,;\,7) = 60{,}802 + 0{,}767\cdot 110 + 3{,}961\cdot 7 = 172{,}86$

1.21 AUFGABE V^T-REGRESSION

Aufgabe $\mathbf{V^T}$-Regression

Gegeben: Wertetabelle $(x_i \mid y_i)$, $(u_i \mid v_i \mid x_i \mid y_i)$ $n = 5..6$, Regressionsmodell
zusätzlich Wert(e) für eine Interpolation

Gesucht:

a) VANDERMONDE-Matrix für das angegebene Regressionsmodell
$\hat{y}(x) = a_0 + a_1\varphi_1(x) + a_2\,\varphi_2(x) + \ldots + a_k\,\varphi_k(x)$

b) Gleichung der Regressionsfunktion

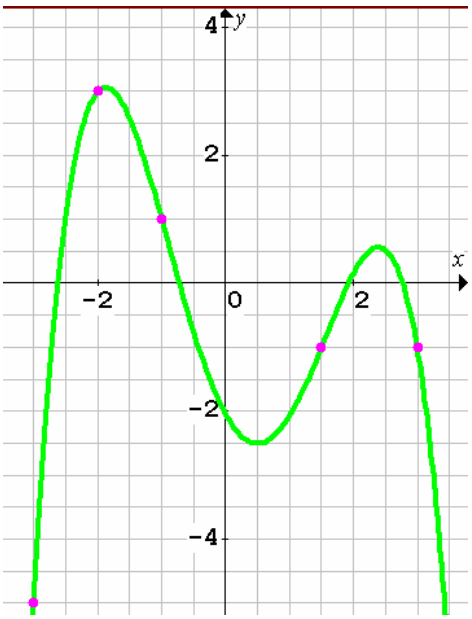
c) Interpolationswert $\hat{y}$

Schritte:

1. VANDERMONDE-Matrix für das gegebene Zahlenmaterial formulieren
2. Schema für die Lösung der Gleichung $\boldsymbol{V^T V\,a = V^T\,y}$ erstellen, evtl. Vordruck benutzen
3. $\boldsymbol{V^T V\,a = V^T\,y}$ nach $\boldsymbol{a}$ auflösen, bei $\boldsymbol{V^T V}$ die Symmetrie ausnutzen.
   Das lineare Gleichungssystem mit GAUß-JORDAN-Verfahren lösen.
4. Regressionsgleichung formulieren
5. Gegebene Wert(e) für die Interpolation in die Regressionsgleichung einsetzen.

1.22 Interpolation, zur Erinnerung

❶ $P_0(-3|-5)$ $P_1(-1|1)$ $P_2(3|-1)$ $P_3(-2|3)$ $P_4(1,5|-1)$
5 Stützstellen

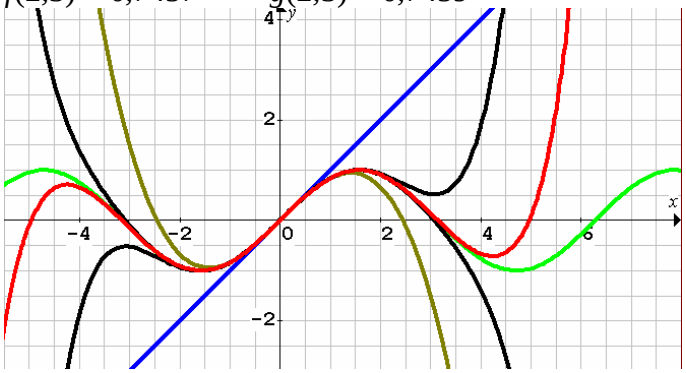


❷ Taylor-Reihe für $f(x) = \sin x$
entwickelt für $x = 0$.

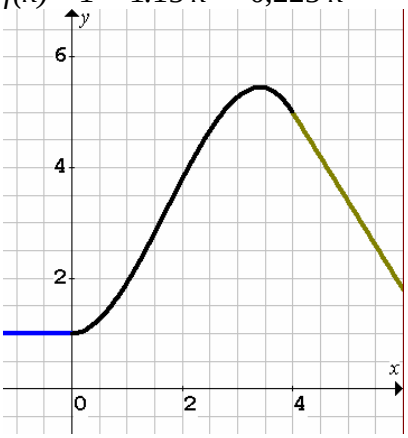
$$g(x) = x - \frac{x^3}{6} + \frac{x^5}{120} - \frac{x^7}{5040} + \frac{x^9}{362880} =$$

$$= \left(\left(\left(\left(\frac{1}{362880}xx - \frac{1}{5040}\right)xx + \frac{1}{120}\right)xx - \frac{1}{6}\right)xx + 1\right)x$$

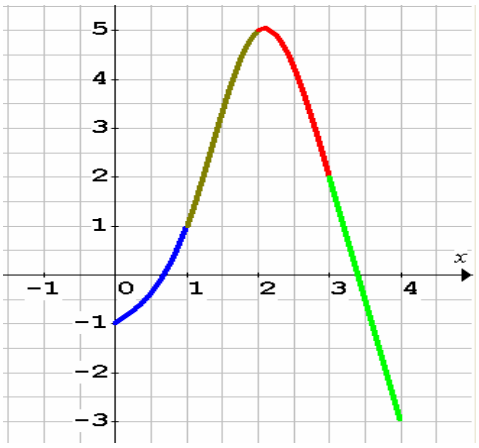
$f(2,3) = 0,7457$ $g(2,3) = 0,7459$



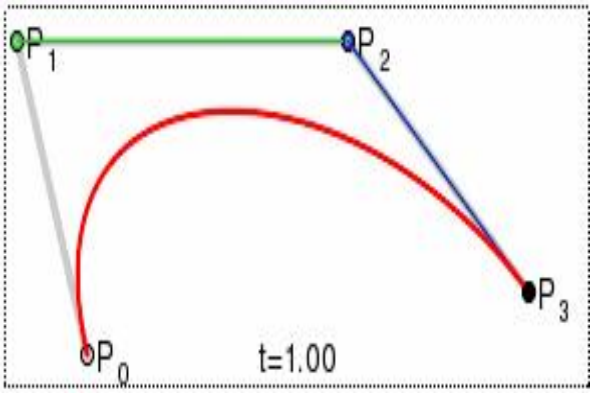
❸ Steigungswinkel 0° bei P_0 (0|1)
Steigungswinkel 122° bei P_1 (4|5)
$f(x) = 1 + 1.15\,x^2 - 0,225\,x^3$



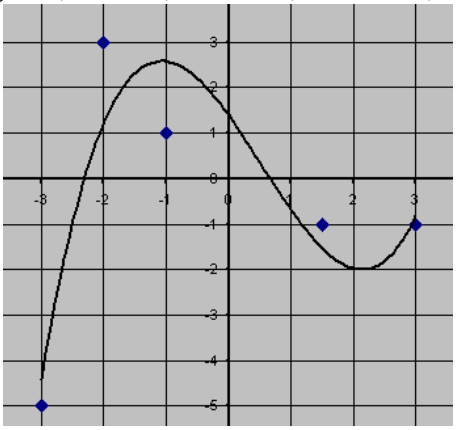
❹ kubische Splinefunktion, 5 Stützstellen P_0 … P_4
4 Teilpolynome $s_0(x)$, $s_1(x)$, $s_2(x)$, $s_3(x)$,



❺
Bézier-Polynom zum Polygon P_0, P_1, P_2, P_3



❻ Ausgleichspolynom
$f(x) = 1,392 - 1,909\ x - 0,446\,x^2 + 0,279\,x^3$

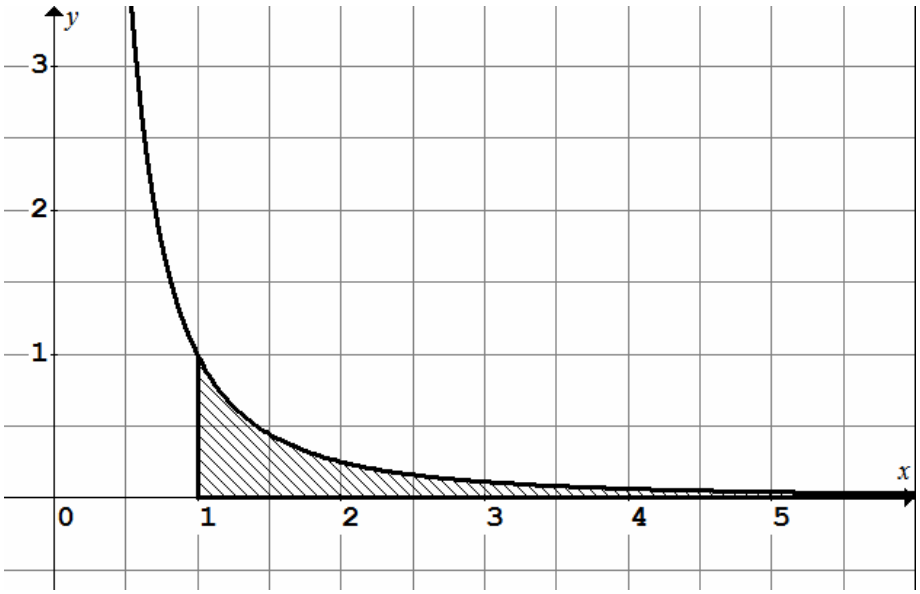


1.23 Uneigentliche Integrale

Flächen zwischen Randfunktionen f(x) und ihrer Asymptoten a(x) ragen ins Unendliche.
Mit Hilfe der Grenzwertrechnung kann man die Inhalte solcher Flächen bestimmen.
In bestimmten Fällen haben diese Flächen einen endlichen Inhalt.

Die Ausdrücke für solche Flächen nennt man uneigentliche Integrale.
Viele Integraltafeln enthalten ein Liste aller uneigentlichen Integrale.

□ $f(x)=\frac{1}{x^2}$ mit der x-Achse als Asymptote $a(x)=0.$ $f(x)=\frac{1}{x^1}$



$$F(x_2)=\int_1^{\infty} x^{-2}\,dx=\left[-x^{-1}\right]_1^{x_2}=\left[-\frac{1}{x}\right]_1^{x_2}=-\frac{1}{x_2}-\left(-\frac{1}{1}\right)=-\frac{1}{x_2}+1$$

$$F_{\infty}=\lim_{x_2\to\infty}\left(-\frac{1}{x_2}+1\right)=1\ [\mathrm{cm}^2]$$

1.24 ABWEICHUNGSQUADRATE

a) Beim Ableiten von $A=\sum_{i=1}^{n}(y_i-mx_i-b)^2$ ergeben sich 3 Probleme:

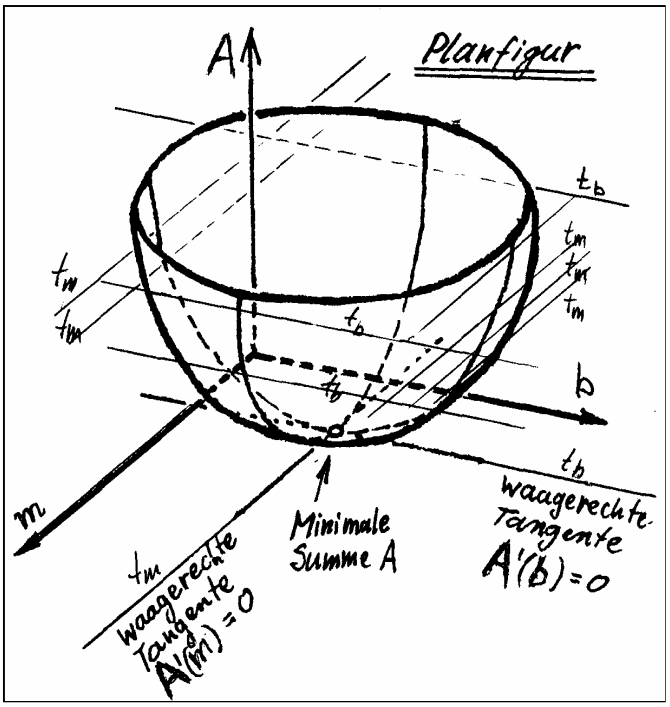
(1) Die Funktionsvariable ("Unbekannte") sind hier nicht x und y, sondern m und b. Die optimalen Werte für m (Geradensteigung) und b (y-Achsen-Abschnitt) sind gesucht. x_i und y_i sind die gemessenen und damit bekannten Tabellenwerte.

(2) Die Funktion enthält ein Summensymbol. Es gilt:

$$\left(f(x)+g(x)\right)' = f'(x)+g'(x),$$

das ist die Summenregel. Man kann also die Summe als Ganzes ableiten.

(3) $(y_i-mx_i-b)^2$ ist ein verketteter Ausdruck. Die innere Funktion ist $u(x) = y_i - mx_i - b$ und die äußere heißt $v(u) = (u(x))^2$. Man benutzt die Kettenregel $f(u(x))' = v'(u) \cdot u'(x)$.



b) Der Funktionsgraph zu A(m,b) ist eine "Mulde" im Koordinatenraum mit den drei Achsen A, m und b. Für jede Kombination (m,b) kann man die Summe der Abweichungsquadrate A(m,b) berechnen.

c) Die Tangenten t_m sind die "waagerechten" Tangenten in Richtung der m-Achse, für ihre Steigungen gilt A'(m) = 0. Eine dieser Tangenten t_m verläuft durch den Tiefpunkt.
Die Tangenten t_b sind die "waagerechten" Tangenten in Richtung der b-Achse, für ihre Steigungen gilt A'(b) = 0. Eine dieser Tangenten t_b verläuft durch den tiefsten Punkt.
Am tiefsten Punkt der "Mulde" schneiden sich die Tangenten, es gilt A'(m) = 0 und gleichzeitig A'(b) = 0. Man erhält also m und b durch Lösen des Gleichungssystems

$$\begin{cases} A'(m)=\dfrac{\partial A}{\partial m}=\sum 2\cdot(y_i-mx_i-b)^1\cdot(-x_i)=0 \\ A'(b)=\dfrac{\partial A}{\partial b}=\sum 2\cdot(y_i-mx_i-b)^1\cdot(-1)=0 \end{cases}$$

$\dfrac{\partial y}{\partial x}$ sind partielle Ableitungen