Prof. Dr. Bernd Heilig ***für IB***

Statistik *- incl. Digitale Datenanalyse -*

DHBW Mergentheim IB 22 A/B (30 Ustd) Oktober 2022

Intro Piratenschatz

Das statistische **Modell**

DATEN / Digitalisierung

Messen von Daten

Beschreiben: Deskriptive Statistik
- Häufigkeiten
- Mittelwerte
- Streuung

- Clustern (Klumpen, Klassen, Gruppen)
- Zusammenhang *Ähnlichkeits*-Analyse (Korrelation)
- Einflussnahme *Wirkungs*-Analyse (Regression)

Vorhersagen: Induktive Statistik
- Wahrscheinlichkeit / ODDs
- Bedingte Wahrscheinlichkeiten
- EXKURS: Klassifizierung mit dem Naiven Bayes Algorithmus

Zufallsprozesse
- Erwartungswert
- Analyse des Zufalls
- Qualitätskontrollen (Stichproben) /
  - Hypothesentest / Alphafehler

VERTIEFUNG: Präsentation von Daten

LAST: Delphi–Methode für funktionierende Kurz-Befragungen

Verwendete Literatur:

Ariely, D. ***Denken hilft zwar, nützt aber nichts*** Droemer 2008
Bayer, H.C. von ***Das informative Universum*** Beck 2005
Bari, A. u.a. ***Die analytische Glaskugel*** Wiley 2014
Bohley,P.: ***Statistik*** Oldenbourg 2000
Bowers, D.: ***Statistics for Economics and Business*** Mcmillan Press 1997
. Fahrmeir, L. u.a ***Statistik Der Weg zur Datenanalyse*** Springer 2016
Fasel,D. Meier,H. ***Big Data*** Springer 2016
Foster/Stine/Watermann: ***Basic business statistics*** Springer 1998
Fricke, W. ***Statistik in der Arbeitsorganisation*** Hanser 2004
Herger, Mario ***Wenn Affen von Affen lernen*** Plassen 2020
Hesse,C. ***Warum Mathematik glücklich macht*** Beck 2014
Krämer,W.: ***So lügt man mit Statistik*** Campus 1991
Neubauer, G: ***Statistische Methoden*** Vahlen 2002
Provost/Fawcett ***Data Science für Unternehmen*** mitp 2017
Reil, H. ***Predictive Analytics*** Genios 2015
Runkler, T.A. ***Data Mining*** Springer 2015
Schallmo, R.A. ***Digitale Transformation von Geschäftsmodellen*** Springer 2017
Scharnbacher, K.: ***Statistik im Betrieb*** Gabler 2004
Schneider, T. ***Digitalisierung und künstliche Intelligenz*** Springer Gabler 2022
Schwarze,J.: ***Grundlagen der Statistik*** NWB 2005
Seeberg, P ***Wie KI unser Leben verändert*** Hanser 2021
Spiegelhalter, D.: ***Die Kunst der Statistik*** Redline 2020
Taleb, N.N.: ***Der schwarze Schwan*** dtv 2011
Wennker, P ***Künstliche Intelligenz in der Praxis*** Springer Gabler 2020
Wewel, M.C. ***Statistik im bachelor Studium*** Pearson 2006
Wong, D.M. ***Die perfekte Infografik*** redline 2011
Zelasny, G.: ***Wie aus Zahlen Bilder werden*** Gabler 2005
Zöfel,P.: ***Statistik verstehen*** Addison-Wesley 2002

ANHANG

1 FS Zoo
2 FS werk
3 FS Bar
4 Lösung: Werte eines Würfels
5 a) FS Disco b) Lösung
6 Lungenkrebs
7 Statistische Tabellen – Normalverteilung

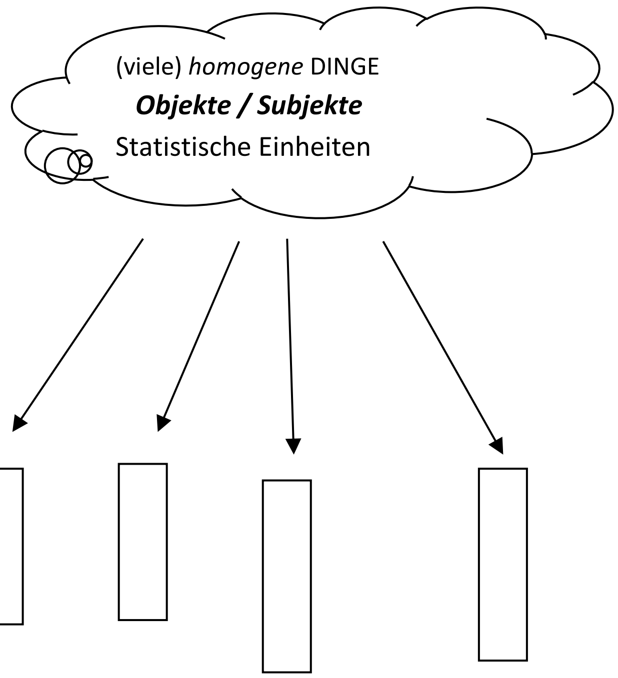
FS Fallstudien (mini)
case studies: Lernaufgaben

Formelsammlung *ist Teil der Klausur*

Intro ***Piratenschatz***

1 Das statistische Modell (Beschreibung der Realität / Dinge)

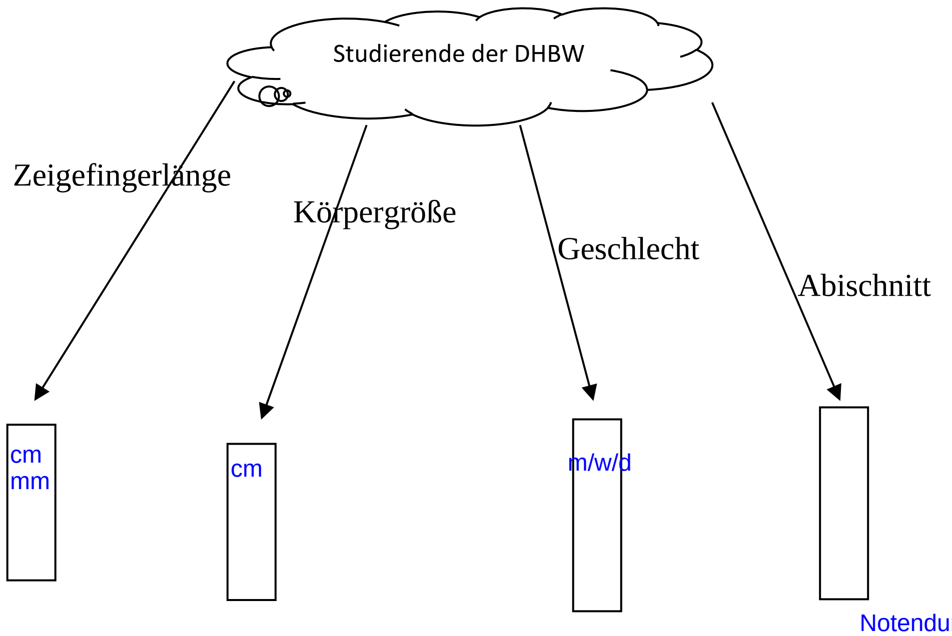
WELT (Universum, Realität)
Grundgesamtheit GG
Masse, (population)



Merkmale (variables)
X, Y, ...

Erfassung, Beobachtung,
Messung, Merkmalsausprägung
Skalenwerte (scale)
X_i, Y_i, ...
(***DATEN***)
genormte Zeichen

FS Aus dem Leben



cm ; 0 bis 20cm

0 bis 200cm

?
männlich/weiblich/divers

Notendurchschnitt; Zahl mit Nachkommastelle; 0,8 bis 6,0

- möchte immer auf ganze Zahlen kommen, aber manchmal auch Kommazahlen in Ordnung
- je genauer, desto kleiner muss die Skala gewählt werden
- Skala auch verschiedene Optionen (z.B. m/w/d)

<u>2 Die Statistik befasst sich mit DATEN / Digitalisierung / Information</u>

sie hilft vor allem, wenn es „*unübersichtlich viele sind*“

Definitionen

Informationen

- ugs. (semantisch): Alle Formen von zweckdienlichen ***Nachrichten*** (nicht: ***Wissen***).
- technisch (operational): ***Symbole***, mit denen Nachrichtet übermittelt werden

*Shannon: Der **Informationsgehalt** einer Nachricht ist die Anzahl der bits, die gebraucht werden. (keine Bedeutung!)*

Daten: „genormte Zeichen, die Information enthalten“
„sind digitale Darstellung von Information“
„sind Informationen die binär digital umgewandelt wurden

Arten: alphanumerische (alphabetische, numerische, sonstige „Sonder“zeichen)

Qualität: Sachdaten, Ordnungsdaten

FS Sachdaten

Das Taschengeld von vier Kinder in drei Monaten wurde ermittelt

Die Sachdaten sind die rot unterlegten 12 Zahlen, alles andere sind Ordnungsdaten

in Euro (€)	Februar	März	April	*Summen*
Egon	14	17	22	53
Fritz	21	21	21	63
Eugen	34	45	18	97
Ludwig	5	7	6	18
Summen	74	90	67	

Statistische Methode

sie erfasst, analysiert und präsentiert die Daten (***messen, beschreiben, darstellen***)

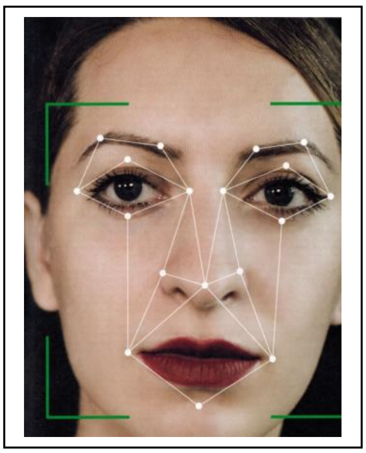
sie ist deshalb eine **Methodenlehre** (Hilfswissenschaft).

Digitale Revolution (Digitalisierung in der Wirtschaft)

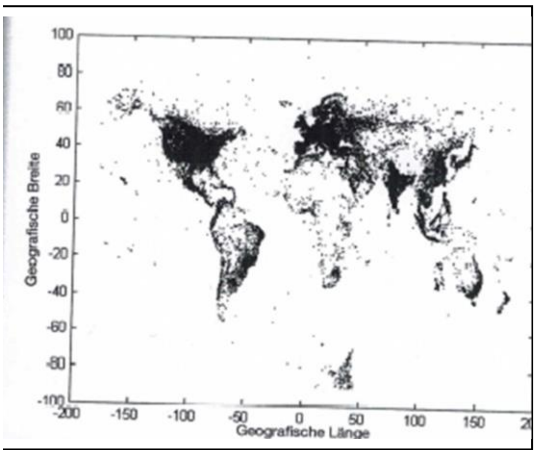
Digitalisierung bedeutet die Übertragung der Wirklichkeit in die Welt der Zahlen.
(EDV: in die beiden „Ziffern“ :Null und Eins).

Transformierung der Realität in ein formales Gerüst (Fiktion)
Darstellung des Wirklichen in einer Zeichen/Zahlen/Datenfolge (binär)
Beispiele Digitalisierung:

REAL ⟶ Digital ⟶ „REAL“



Zeile	Spalte



3 Messen (Erhebung, Erfassung) von Daten

3.1 Skalenarten

Diskrete Mess-Skala
Einzelne Werte, keine Werte dazwischen

Stetige Mess-Skala
Es gibt alle Werte, durchgehende Linie

In der Praxis oft: ***quasi-stetige*** Skala: man kann die Skala „beliebig“ kleiner machen (Geldeinheiten, Gewichte, Zeit, Längen ……)

kardinale Skala: “echte” Zahlen / metrisch (messbar) — ***quantitative*** Merkmale

ordinale Skala: ordenbare Zeichen
Rangordnung ist möglich (kategorial***)***

nominale Skala: “nur” Namen möglich

qualitative Merkmale

TEAM

FS Hemden-Wühltisch

In einem Kaufhaus liegen als Sonderangebot eine Menge Hemden (n = 178)

MERKMAL	Größe	Farbe	Länge	Muster
SKALA “dimension“	M,L,XL,S	Farbspektrum	cm	gestreift, getigert einfarbig, gezackt
Skalenart *nom, kar, ord*	ordinal	nominal	kardinal	nominal

3.2 Erfassung von Daten

nur zwei Methoden:

Urliste
Die Werte werden der Reihe nach aufgeschrieben

Strichliste
In einer vorgefertigten Tabellen werden durch Abfrage die Anzahl festgestellt.

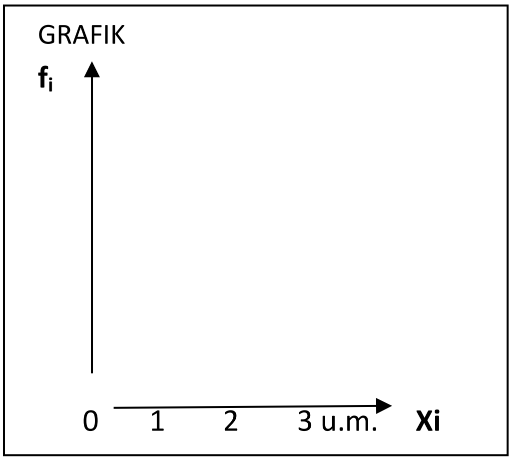
FS Alter der Kursteilnehmer (Urliste)

FS Anzahl Geschwister (Strichliste)

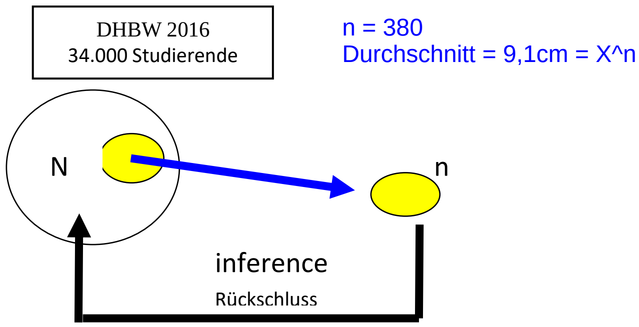
Häufigkeits-tabelle

X_i Anzahl	**Striche**	**absolute Häufigkeit f_i**
0		
1		
2		
3 u. mehr		
Summen		

Häufigkeits-bild



3.3 Vollerhebung, Teilerhebung (Stichprobe)



Grundgesamtheit; Teilmenge (**Stichprobe**)
Rückschluss (inference)

Güte des „Rückschlusses"
(Gesetz der großen Zahl; repräsentativ)

3.4 Ausreißerwerte

Definition: „weit entfernt vom Durchschnitt", weit von den anderen Werten

„***Normale*** Ausreißer" (Körpergröße, -gewicht, Lebensalter, IQ…),
die einzelne Werte sind nicht gewichtig, nicht wesentlich für den Durchschnitt.
„***Extreme*** Ausreißer" (Vermögen, Buchverkäufe, ...)
einzelne Werte sind sehr gewichtig, prägend, beeinflussen den Durchschnitt.

Wie behandelt man **Ausreißer**-Werte:
Nach der Erfassung nach Grund forschen, dann ***begründet*** von der Analyse fernhalten.
Immer am Anfang der Analyse entscheiden.

FS Anzahl Münzen (Urliste)
n=Anzahl; min/max: kleinster / größter Skalenwert
[R: Spannweite (Range)] Ausreißerwerte ???

4 Deskriptive (beschreibende) Statistik

4.1 Häufigkeiten (absolute, relative, kumulierte)

die Statistik behandelt Phänomene ***unabhängig*** von der Anzahl der Werte.

.. es ist egal, wie viele man beobachtet, man braucht nur die Häufigkeiten.

Absolute **Häufigkeiten**

absolute Häufigkeiten

absolut kumulierte Häufigkeiten - Summenhäufigkeiten – $\boxed{f_i}$

Relative **Häufigkeiten**

relative Häufigkeiten

Normierung auf "1"

Normierung auf "100" (in v.H.) ("*prozentuale Anteile*")

relativ kumulierte Häufigkeiten

$$h_i = \frac{f_i}{n}$$

FS zurück zur FS Geschwister

Anzahl	Striche	absolute Häufigkeit	absolut kumuliert (bis zu)	relativ (~~%~~)	relativ kumuliert (%)
0		1	1	0,043 = 4,3%	4,3
1		13	14	0,565= 56,6%	60,9
2		8	22	0,348 = 34,8	95,7
3 u. mehr		1	23	0,043 = 4,3%	100
Summen	23	23	X	1 = 100%	

beim verfälschen immer größten wert nehmen

4.2 Grafische Darstellung (Diagramme)

Die grafische Darstellung der Verteilung der Häufigkeiten auf die Skalenwerte ergibt das ***Histogramm*** (Verteilungsfunktion)

Diskussion von "Verteilungen"

Hüllkurven (smooth-curve) zur optischen Gestaltung

❖ **Gipfel:** eingipflig (unimodal), zwei- oder mehrgipflige Verteilungen

Sonderfall: ***Gleich****verteilung*

❖ **Symmetrie:** Symmetrische, links/rechts-steile Verteilung, **-Schiefe-**

Sonderfall: ***Normal****verteilung*

FS Bild

In Deutschland verdienten 2010 berufstätige Männer durchschnittlich 3.500 €, Frauen hingegen nur 2.800 €. Stellen Sie die ***Daten*** in einer statistischen Grafik dar. Balkendiagramm; X = mann/frau; f = 2800/3500

kein mathematisches Koordinatensystem (kein Nullpunkt); nur Präsentation

FS Welches Bild ?

** Sie würfeln mit einem Würfel (n=570) X = 1 bis 6, da es 6 verschiedene "Ereignisse" gibt; f = 0 bis 590, da man 590x würfelt

** Verteilung der Schuhgrößen aller DHBW-Studierenden (n=34.000)

Eventuell Vertiefung in Kapitel 7

4.3 Mittelwerte (mode, median, mean)

TEAM

B 31 Fünf Dörfer liegen an einer Strasse.
Es soll ein zentraler Kindergarten gebaut werden.
Wo ist der ***optimale Standort*** ?

Dorf	km
a	**5**
b	**15**
d	**30**
f	**75**
g	**95**

= Median; stelle wo die Kilometerzahl zu allen Ortschaften am minimalsten ist

❖ **Modus** (Häufigster Wert, ***mode***):
am **häufigsten** vorkommender Wert einer Reihe (Modalwert)

Es ist bewiesen, dass das Feiern von Geburtstagen gesund ist. Statistiken belegen, dass Menschen, die die meisten Geburtstage feiern, am ältesten werden. *(Hesse S. 123)*

❖ **Median** (Zentralwert, ***median***):
mittlerer Wert einer ***geordneten*** Reihe von Werten

❖ **Mean** = **Arithmetischer Mittelwert** (**Durchschnitts**wert)
nur für *additive* (**kardinale**) Skalen

$$X^M = \frac{1}{n}\sum x_i$$ = ***Summe*** der Werte dividiert durch die Anzahl der Werte

Eigenschaften:

★ Der ***arithmetische Mittelwert*** hat die ***Schwerpunkteigenschaft***, er bildet die Mitte einer Zahlenreihe (Wippe) weil die Summe der Abweichungen links und rechts gleich sind.

Bsp. Zahlenreihe	5	7	11	21	28	33	42
Abweichungen	*-16*	*-14*	*-10*	*0*	*+7*	*+12*	*+21*

Summe der Abweichungen $\Sigma = -30$ $\Sigma = +30$

Daraus folgt: Die Summe der Abweichungen (linke negative und rechte positive) sind gleich,

also $\Sigma(X_i - X^M) = 0$ / $\Sigma X_i - nX^M = 0$ / $\Sigma X_i = n X^M$ / $\Sigma X_i / n = X^M$!

★ Der ***Median*** hat die Eigenschaft, dass er die Wertereihe in gleiche Teile gliedert, d.h. links und rechts sind gleichviele Werte (Anzahl).

★ Der ***Modus*** zeigt den schwersten Wert, der der am häufigsten vorkommt.

FS Restposten „Herrenhemden in einem Regal" ***„Größe" cm***

Größe	cm	f_i	f_i kum
S	36	6	6
M	38	9	15
L	40	11	26
XL	42	3	29
		29	

n	29		
min	S	36	
max	XL	42	
modus	L	40	
median	M	38	n=29, also gesucht 15.Stelle mit kummulierter Nummer
mean	etwas mehr als M; aber nicht addierbar, also eigentlich kein ergebnis	38,76, also 38	

__**Analysieren**__ Sie die Daten (***Datenanalyse***).

FS project

Die IT-Abteilung besteht aus 15 Projektteams mit unterschiedlicher Mitarbeiterzahl (MA)

team	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
MA	**4**	**9**	**4**	**4**	**9**	**8**	**7**	**8**	**7**	**2**	**4**	**8**	**9**	**4**	**3**

Ermitteln Sie die "***mittlere" Mitarbeiterzahl pro Team*** (modus, median, mean).

FS Augenfarben Neun Damen haben unterschiedliche Augenfarben

	1	2	3	4	5	6	7	8	9
	blau	**schwarz**	**blau**	**schwarz**	**braun**	**blau**	**grün**	**blau**	**braun**

Ermitteln Sie die "***mittlere" Augenfarbe*** (modus, median, mean).

modus = blau
median, mean nicht berechenbar, da nominal (aka nicht ordnenbar)

LernZielKontrollFrage

Wo passt was ?			**kardinale Skala**	**ordinale Skala**	**nominale Skala**
Mean	Arithmetischer Mittelwert	***mean***	X	muss umgewandelt werden	
Median	Zentralwert	***median***	X	X	
Modus	Häufigster Wert	***mode***	X	X	X
			echte Zahlen	*ordenbar*	*nur Namen*

FS Familie

Familie Oberberg	Körpergröße -cm-	T-Shirtgröße	Augenfarbe
mam	168	XL	blau
dad	178	L	braun
oskar	102	S	blau
marlene	126	S	blau
susi	84	XS	schwarz

Beschreiben Sie das durchschnittliche Familienmitglied.

Der durchschnittliche Oberberger ist …

4.4 Streuung (Abweichung, Varianz)

Durch die Berechnung eines Mittelwertes ergeben sich Abweichungen der einzelnen Werte von diesem Mittelwert („Varianz“).

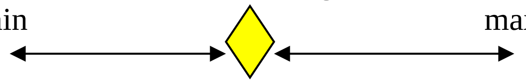
Eine **Datenreihe** geordneter Werte X_i

min max **-ohne** Ausreißerwerte-

R (***range***, Spannweite

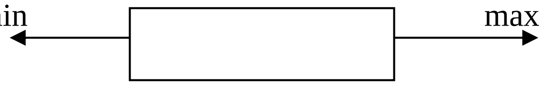
Die **Mitte** und damit **Abweichungen** der Werte von der Mitte in beide Seiten

min max



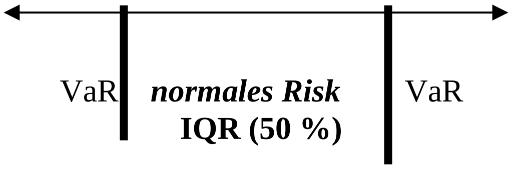
Trennung von normalen zu außergewöhnlichen Abweichungen (**Varianzanalyse**)

min max



Bezeichnung, engl.

Gesamte Abweichung ***R i s k (R)*** ***Range***



Value at Risk (Schwellenwerte)
(Inter Quantil Range); Toleranz

4.4.1 Quantilsmethode (Streuungsanalyse für den Median) –für ordinale Skalen

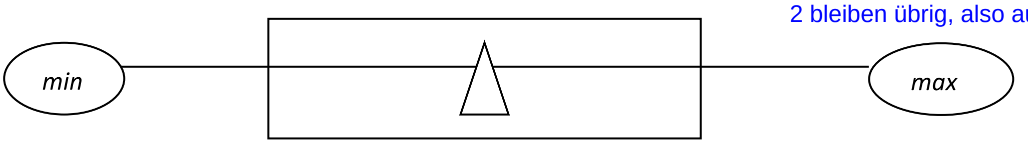
FS Körpergrößen

gemessen wurden folgende Körpergößen (cm)
(geordnet) keine Ausreißer, n = 8

165	**168**	**173**	**176**	**180**	**182**	**182**	**185**

Wie groß ist die gesamte Spannweite (R) (vom MIN zum MAX)

Wie groß ist die 50% - Streuung um den Median (IQR 50%)?
Ermitteln Sie den Boxplot mit min/max Median und IQR(75%). = eine von 1 Zahlen als außergewöhnlich große Zahl; da 75% = 3/4 = 6/8; 2 bleiben übrig, also auf jeder Seite bleibt 1 übrig



(normale, außergewöhnliche Abweichung um die Mitte) ***Boxplot***

TEAM

FS SCRUM ***Wie viele Teilnehmer kommen morgen*** ?

Für die täglichen stattfindenden Scrum-Sitzungen soll morgen bei einem Caterer Mittagessen bestellt werden. In der Vergangenheit kamen jeweils folgenden Projektmitarbeiter (MA). Ermitteln Sie mit Hilfe einer **Querschnittsanalyse** eine **Punktschätzung** (PLANWERT) und eine **Intervallschätzung** (Szenario)

MA pro Sitzung	absolute Häufigkeit	kumulierte Häufigkeit
19	**3**	3
20	**10**	13
21	**7**	20
22	**15**	35
23	**18**	53
24	**37**	90
25	**13**	103
	103	

103 Werktage = 20 Wochen, also 5 Monate

Wie viele Daten sind es (n=?) n = 103

Wie lange ist diese Vergangenheit ? 103 Werktage

Wie ist die Spannweite (R)? 19 bis 25; R = 6

Ermitteln Sie den Median. *(Punktschätzung)* = 23, da Wert 52 in Mitte und 23 bei 52

Bestimmen Sie den IQR 50%

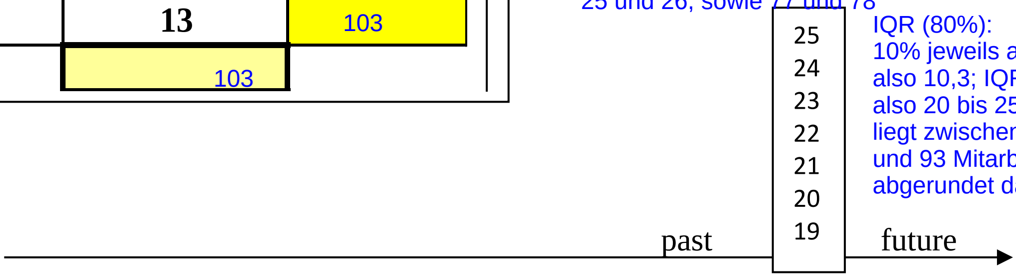
Bestimmen Sie den IQR 80%

(Intervallschätzungen)

(Value at risk – Werte VaR)

IQR (50%): 25,5 werte liegen jeweils außerhalb; IQR von 26 bis 77; also von 22is 24 Mitarbeiter; eigentlich zwischen wert 25 und 26, sowie 77 und 78

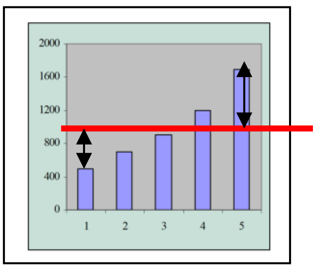
IQR (80%):
10% jeweils außerhalb
also 10,3; IQR von wert 11 bis 93
also 20 bis 25 Mitarbeiter.
liegt zwischen 10 und 11 und 92 und 93 Mitarbeiter, aber abgerundet da mehr sinn



4.4.2 Standardabweichung /Volatilität (Streuungsanalyse für den *Mean*) – nur für kardinale Skalen

FS Eine **Aldi-Filiale in Mannheim** hat abends folgende Kassenbestände

	X_i (Euro)	AbwX $(X_i - X^M)$	absolute Werte aka Betrag	(Xi-Xm)^2
Kasse 1	**500**	-500	500	250000
Kasse 2	**700**	-300	300	90.000
Kasse 3	**900**	- 100	100	10.000
Kasse 4	**1200**	200	200	40.000
Kasse 5	**1700**	700	700	490.000
S U M M E	5000	0	1800	880.000
durch n 5	1000		360	176.000



Der arithmetische Mittelwert ist 1000 € (durchschnittlich hat ***jede*** Kasse 1000 € Bestand)
Wie ist die **durchschnittliche Abweichung** (Streuung) von der Mitte ?

➡ **mittlere absolute Abweichung** (Mean Absolute Deviation: MAD)
MAD = $1/n \ \sum [X_i - X^M]$ Summe der absoluten Abweichungen / n

➡ **Varianz**: mittlere quadratische Abweichung
VAR = $1/n \ \sum (X_i - X^M)^2$ Summe der quadrierten Abweichungen / n
für Stichproben genauer: / (n-1)

➡ **Standardabweichung**: "mittlere Abweichung"
s = $\sqrt{VAR}$ Quadratwurzel der Varianz [X^M +/- s]

FS Grizmek
Eine Analyse in Afrika und Südamerika ergab folgende Daten:

Merkmal: Rüssellänge (cm)	Arithm. Mittelwert X^M	Stand.abweichung s	Vergleich ????
Elefanten	375	52,7	14,1% = (52,7/375)x100
Kolibris	8,5	2,4	28,2% = (2,4/8,5) x 100

durchschn. Standardabweichung ist bei Kolibris 4x so groß
d.h. bei Kolibris gibt es größere Unterschiede im Durchschnitt als bei Elefanten

Vergleichen Sie die Abweichungen in beiden Tierarten.

➡ **Variationskoeffizient**:: **relative** mittlere Abweichung in v.H. (Variationszahl v)
v = s / X^M (*100) relatives Streuungsmaß [X^M +/- v (in%)]
("***Vola***": ***Volatilität*** als Maß der Schwankungen (Abweichungen, Streuung)

FS Firmenwahl
Sie haben zwei Zusagen der Firmen Alpha / Beta, bei beiden gibt es einen durchschnittlichen Einstiegsgehalt von 3.300 / 3.450 Euro, die Vola beträgt 25% / 8%.
Wo gehen Sie hin, wenn der Verdienst ausschlaggebend ist ?

a) v = 3.300 x 0,25 = 825
b) v = 3.450 x 0,08 = 276

Team FS Fingerlänge / (und Körpergröße)
Messen und bestimmen Sie X^M, s, und v(%) der Länge des linken Zeigefingers (mm ganzzahlig).

Tschebyscheff Ungleichung für beliebige Verteilungen (!)
(aus: Griffith, D. Statistik von Kopf bis Fuß O´Reilly 2009 S. 645
Unabhängig von der Häufigkeitsverteilung gilt:
+/- 1,25 s > 36 % der Werte
+/- 1,5 s > 55 % der Werte
+/- 2 s > 75 % der Werte
+/- 3 s > 89 % der Werte
+/- 4 s > 94 % der Werte

4.4.3 Covarianz (Kovarianz) - ohne Sinn

gemeinsame durchschnittliche Abweichung; für zwei Merkmale, Variablen (X und Y):

VAR = $1/n \sum (X_i - X^M)^2$

COV = $1/n \sum (X_i - X^M)(Y_i - Y^M)$

TEAM

FS Kovarianz:

X_i Fingerlänge - cm -	Y_i Körpergröße - dm -	Xi-Xm	Yi-Ym	COV(XY)	
10	**18**	1	0	0	
8	**18**	-2	0	0	
6	**17**	-3	-1	3	
12	**19**	3	2	6	6/4=1,5
9	18				
			COV	1,5	

4.4.4 Normalverteilung -Konfidenz/Toleranz- (Streungsanalyse für typischen Modus)

Intro **Sandkörner**

gestern n: 12.550 **min: 50** **max: 316** **Mitte ???**

Wie viel sind es ***morgen*** ? Median: zwischen 6275 u. 6276 (s=+/-43)

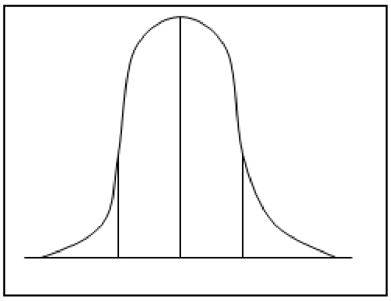
Theorie

Normalverteilung (Gauss): eingipflig, symmetrisch, assymptotisch

Mitte m "μ" : mode = mean = median (X^M; μ)

Streuung s "σ" : Varianz (s^2, σ^2)

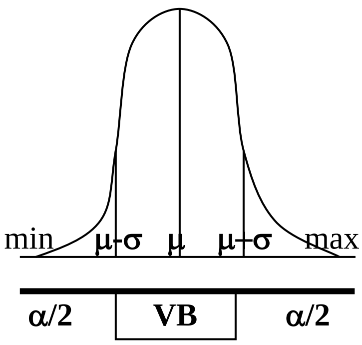
Standardabweichung (s, σ)



Eigenschaften:
Eingipflig –unimodal- (Ballung)
Symmetrisch
Knickpunkte (gleich-hoch)
Alle Mittelwerte in der Mitte
Abstand Mitte-Knick: Standardabweichung
Assymptotisch (schneidet nicht die X-Achse)

***Konfidenz*intervall** (Vertrauensbereich **VB**)
*Normale Abweichung (**Toleranz**)*

***Signifikanz*bereich** („Irrtumsbereich“ α)
Außergewöhnliche Abweichung



Praxis:

typische VB-Bereiche (Konfidenz-, Toleranzbereich):

Fläche in dem σ **-Bereich**	beträgt	0,6826	(68,3 %)	***("rule 66")***
2 σ Bereich		0,9544	(95,4 %)	(95%-ige Sicherheit)
3 σ -Bereich		0,9974	(99,7 %)	(99%-ige Sicherheit)

typische α-Bereiche (Signifikanzbereiche):

10%, **5%**, 2,5%, 1%, 0,5%, ….
links und rechts jeweils 2,5% : 95 Vertrauensbereich

Morgen *war* heute

FS Audi 8 ¼ (audi.xls)

Das Modell hat getestet Verbrauchsdaten: μ = 8,4 σ = 0,45 l/100km.

Für das Modell A 8 ¼ versprechen wir, dass 95,44% der Fahrzeuge einen Durchschnittsverbrauch zwischen 7,5 l und 9,3 l haben werden.

wenn abweichung "zufällig" sind
= nur 4,66 % aller produzierten Audis verbrauchen mehr als 9,3 l oder weniger als 7,5l auf 100km

Team	***Gummibärchen, 5-Cent-Münzen, Schraubenlänge, .Fingerlänge...)***

vor klausur mal machen

4.5 Clustern (Klumpen, Klassen, Gruppen, „zusammenfassen“)

4.5.1 Möglichkeiten der Gruppierung (Cluster)

Betriebliche (ökonomische und technische Zahlen-(Werte) haben eine Mengen- und/oder Wertdefinition; NEU: BIG DATA (Viel, vielfältig, schnell)

Bsp. Kunden, Produkte, Absatz, Gewinn, Teile, Kosten, Umsatz

Einteilung einer Datenreihe (***Cluster; Gruppen, Segmente, Klumpen***):

Zwei (Wichtig; Unwichtig)
Drei (sehr wichtig; weniger wichtig; nicht wichtig)
Vier (sehr gut gut schlecht sehr schlecht
Fünf oder mehr (.++ + neutral - --.)

Über die Klassenbildung (Anzahl):

Faustregel: mindestens 5 und höchstens 20; möglichst Klassen mit gleicher Breite
REFA-Vorschlag: Klassenanzahl zwischen $n^{1/2}$ und $n^{1/3}$ passend.
Papula-Vorschlag: Anzahl der Klassen $= n^{1/2}$
Sturgess-Regel $k = 1 + \log(2)\, n$
Pareto-Regel (80/20-Regel: 80% des Wertes werden von 20% der Menge generiert)

TEAM ***hotel***

Aufgabe: Für die Seminarteilnehmer sollen Hotels zum Selbstbuchen angeboten werden.					
	Hotelpreise MA eine Übernachtung ohne Frühstück				
		Euros			
DATEN	*goldene Gans*	**98**	<100	mittel	ansonsten auch teuer
	Sheraton	**115**	>100	"mittel"	
	Central	**75**	<100	<90	
	Steigenberger	**245**	>100	teuer	
	ibis	**83**	<100	<90	
	Holländer Hof	**90**	<100	mittel	
	Prinzenpark	**85**	<100	<90	

FS Zoo (ANHANG 1)

4.5.2 Maschinenlernen („künstliche Intelligenz“ KI AI)

(Lernen bedeutet Verhalten ändern !)

DATEN

Suchen nach **Strukturen**	Einteilung in **Klassen**
- ohne Vorgaben –	***- nach Vorgaben („labels“) -***
Gruppen, Cluster	Klassen
unüberwachtes Lernen	***überwachtes*** Lernen
uninformiertes Verfahren	informierte Verfahren
untrainierte Methoden	trainierte Methoden
unsupervised learning	supervised learning
ZIEL: ähnliche Daten zusammenfassen	ZIEL: mit Klassen arbeiten und vorhersagen

Clustern: unüberwachtes Gruppieren
Klassifizieren: überwachtes Gruppieren

4.5.2.1 Clustern: k-means – Algorithmus (eindimensional)

„unüberwachtes Maschinenlernen“; *Gruppen diskriminieren*

automatisches Clustern

FS kmeans.xls Anzahl der Gruppen

(1) Ordne die Datenreihe (hier: kardinal)
(2) Bilde k ***zufällige*** Center (Gruppenmitten)
(3) Ordne die Werte dem nächsten Center zu (nearest neighbour) = k Cluster
(4) Bilde die means (k Stück) der einzelnen Cluster = arith.Mittelwert wird neuer Center
(5) Wiederhole (3)
(6) Solange wiederholen, bis die Cluster sich nicht mehr ändern.
= Ergebnisse

Auch mehrdimensional möglich. --> unüberwachtes Maschinenlernprogramm

4.5.2.2 Klassifizieren: k-nearest-Neighbour-Algorithmus KNN (zwei-dimensional)

„überwachtes Maschinenlernen“ (Vorgabe der beiden Center); *Gruppenmuster erkennen*

Zuordnung der Werte nach den **k**-nearest-Neighbours

(auch für Vektorräume verwendbar)

nach Vorgaben Klassen bilden

FS KNN.xls

gibt zwei verschiedene Kriterien (X1 & X2)

nearest neighbor : nächster benachbarter Punkt

k1 = nur 1.nächster nachbar
k2 = zwei nächste nachbarn
kX = Anzahl X der nächsten Nachbarn

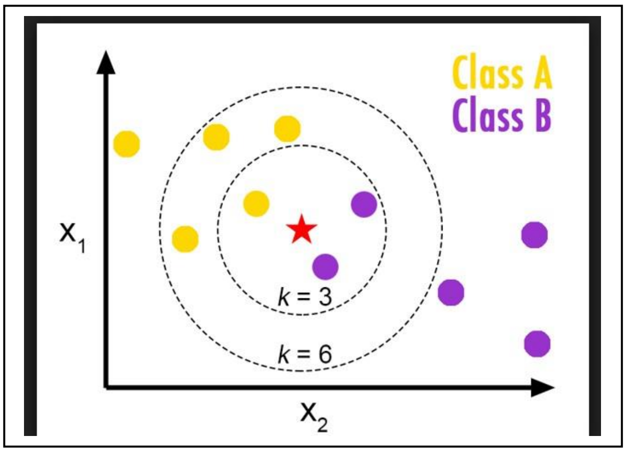
in Praxis nur ungerade k nehmen, damit es keine Uneindeutigkeiten gibt

Algorithmus

-) Vorgabe Anzahl und Bereich der Klassen (Klassifizierung / labels)

-) Trainingsphase

-) danach: Jeder neue Datenpunkt wird automatisch einer der Klassen zugeordnet
(k: Anzahl der Neighbour)



bei 3 nächsten nachbarn ist mehrheit lila, daher lila

bei 6 nächsten Nachbarn Mehrheit der Nachbarn gelb, daher wird Punkt dann gelb
--> k hat Auswirkung auf die Zuordnung zu einer Farbe

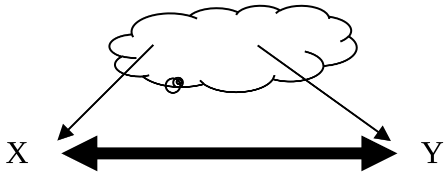
Distanzmaße

- euklidischer Abstand: Satz des Pythagoras
  D = wurzel $((X_1abw)^2 + (X_2abw)^2)$

- Manhatten- (Mannheimer-) Abstand;
  city-block-Distanz
  D = $abs(X_1abw)+abs(X_2abw)$

Auch mehrdimensional möglich !!!! (Pythagoras, Manhatten)

4.6 Zusammenhang (Beziehung); Ähnlichkeit zwischen zwei Merkmalen (X, Y)

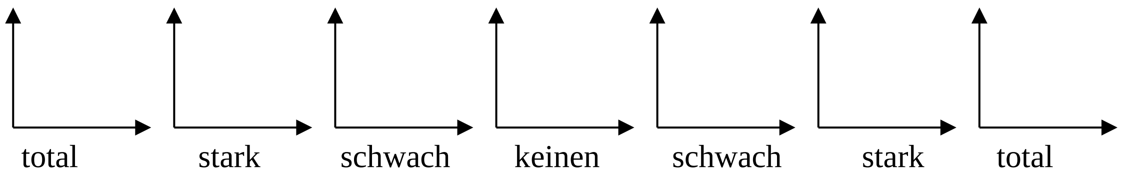


4.6.1 Ähnlichkeitsanalyse (Korrelation)

4.6.1.1 Korrelationskoefizient –Bravais/Pearson– (für kardinale Skalen)

FS Fingerlänge (mm, ganzzahlig) / Körperlänge (cm, ganzzahlig)
Erfassung mit Urliste, Analyse: hängt das zusammen ???????????

n Beobachtungen ergeben n Punkte (scatter-Diagramm)



r	**-1**	negativ	Null	positiv	**+1**

$$\textbf{Korrelationskoeffizient } r = \frac{COV_{XY}}{s_x * s_y}$$

durchschnittliche gemeinsame Abweichung (vgl. 4.4.3)
Produkt der Standard-Abweichungen

$COV = 1/n \sum (Xi - X^M)(Yi - Y^M)$

Zahlenraum -1 =< r <= +1 Vorzeichen: *Richtung* Wert: *Stärke*

FS Beispiel BIG Data (Partnervermittlung)

Zur Stärke ......

"Faustzahlen" *(nach Zöfel* Statistik verstehen *Addison 2002)*

Wert des Koeffizienten	
[r] <= 0,2	sehr geringe Korrelation
0,2 < [r] <= 0,5	geringe Korrelation
0,5 < [r] <= 0,7	*mittlere* Korrelation
0,7 < [r] <= 0,9	hohe Korrelation
0,9 < [r] <= 1,0	sehr hohe Korrelation

EXKURS: Kritikpunkte an den Ähnlichkeitsmaßen

KRITIK bei der Verwendung des Korrelationskoeffizienten

- ➔ **Größe** unsicher (zufall), Faustregel: vor allem bei kleiner Masse: mindestens 0,7
- ➔ nicht geeignet für **nichtlineare** Zusammenhänge (dracula)
- ➔ Achtung vor **Unsinns-** (Nonsens-, Schein-) Korrelation (dow_jones)
- ➔ Stärke abhängig von **Ausreißer** (kingkong)
- ➔ Abhängigkeit von einer **dritten Größe** (glatze)
- ➔ nur möglich bei zwei Variablen: ***Korrelationsmatrix*** (korrmat.xls)
- ➔ Zahlenergebnis schützt nicht vor genauer Analyse (simpson_paradoxon)
  Korreliert heißt nicht kausal !!!!

4.6.1.2 Rang-Korrelations-Koeffizient (Spearman) für ordinale Skalen

FS rang

Differenz (Rd - Re)

	X	Y			Abweichung	
Schüler	Deutsch	Englisch	R_D	R_E	D_i	D_i^2
Anton	**2**	**3,5**	1	3	-2	4
Berta	**3**	**2**	2	1	1	1
Curdi	**3,5**	**3**	3	2	1	1
Doris	**4,5**	**4**	4	4	0	0
Erwin	**5**	**6**	5	5	0	0
Summen			15	15	immer 0	6

Summe Abweichung = 0, da Anzahl gleich

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

mit Di: Differenz der Rangziffern

A. Bravais (1811 – 1863)
K. Pearson (1857 – 1936)
C.E. Spearman (1863 – 1945)

- ◆ gleicher Zahlenraum wie r; sachlogische Erklärung der obigen Formel
- ◆ Rule of Calculation for the same ranks "Die Gesamt-Summe muss erhalten bleiben",

FS Ranggleichheit Ermittle die Rangfolge

Student	#	Note	Rang
Meier	**1**	**1**	2
Alwer	**2**	**1**	2
Kosch	**3**	**1**	2
Gurku	**4**	**4**	4,5
Bose	**5**	**4**	4,5
Summe	**15**		**15**

kein Zusammenhang

Zur Übung: z.B. Deutsch z.B.: Englisch

Student		M1	M2	R1	R2	R1-R2 Di	Di^2
A	1	**2**	**4**	1,5	4,5	-3	9
B	2	**2**	**2**	1,5	1,5	0	0
C	3	**3**	**3**	3	3	0	0
D	4	**4**	**2**	4,5	1,5	3	9
E	5	**4**	**4**	4,5	4,5	0	0
	15			**15**	**15**		**18**

FS Klausur Ermittle den Rangkorrelationskoeffizient

rs= 1- 6x18/5^3 - 5 = 0,1 --> Kein Zusammenhang

Student	Note Mathe	Note Statistik	R_M	R_S	D_i	$D_i^{\,2}$
1024	**1**	**3**	1	3	-2	4
1126 2	**2**	**5**	2,5	6,5	-4	16
1287 3	**2**	**2**	2,5	1	1,5	2,25
1876	**3**	**3**	4	3	1	1
1433 5	**4**	**4**	6	5	1	1
1156 6	**4**	**3**	6	3	3	9
1543 7	**4**	**5**	6	6,5	-0,5	0,25
Summe 28			28	28	0	33,5

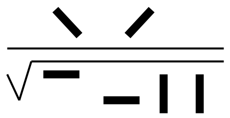
rs = 1 - 6 x 33,5 / 7^3 - 7 = 1 - 0,5982 = 0,4018
Die Note in Mathe und Statistik hängen nur mittel miteinander zusammen

4.6.1.3 Kontingenz-Koeffizient (Chi-Quadrat) für nominale Skalen

meist **nominale „dichitome" Vierfeldertabelle** (zwei Zeilen/zwei Spalten SACHDATEN)

Vierfelderkorrelationskoeffizient rk

a	**b**
c	**d**

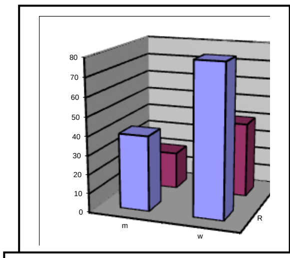


$$r_k = \frac{ad - bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}}$$

r_k Werte zwischen -1 und +1
(Vorzeichen spielt meist keine Rolle)

FS Kontingenz I („nominaler Zusammenhang“)
In einem Unikurs wurden 180 Studierenden nach zwei Merkmalen befragt.
Gibt es einen Zusammenhang zwischen den beiden (dichotomen, "zweiwertige") Variablen ?

absolut	männlich	weiblich	
Raucher	40	80	120
Nichtraucher	20	40	60
	60	120	180

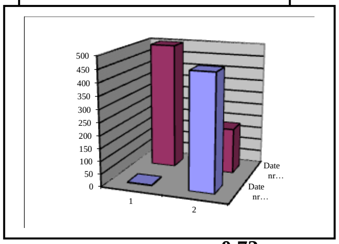


$\mathbf{r_k = 0}$ (!!!!) bei dieser Stichprobe kein Zusammenhang !

FS Kontingenz II
In den Abiklasssen einer Schule wurde ebenfalls gefragt.

absolut	männlich	weiblich		
Raucher	**3**	**455**	458	
Nichtraucher	**500**	**180**	680	
	503	635	1138	-0,72

$\mathbf{r_k}$ **= ??????**



= -0,72

ähnlich: ChiQuadrat ($\chi 2$)
Zähler und Nenner im Quadrat (nur positive Vorzeichen), ***nicht*** von 0 bis 1 !!!!!!
Analog auch für mehrdimensionale Zusammenhänge

<u>4.6.1.4 Kendalls Tau</u> (für ordinale Daten)

Für zwei ordinal gemessene Merkmale X und Y.
Als erstes wir die erste Variable geordnet (nach dem Merkmal X).Dann werden die Rangziffern gebildet. Dann werden die zweite Werte (Rangziffern Y) nach folgendem Muster beurteilt:

$\mathbf{P_i}$ Anzahl der nachfolgenden größeren Y-Rangzahlen
$\mathbf{Q_i}$ Anzahl der nachfolgenden kleineren Y-Rangzahlen

Dann werden die Zahlen addiert zu **P** und **Q**.

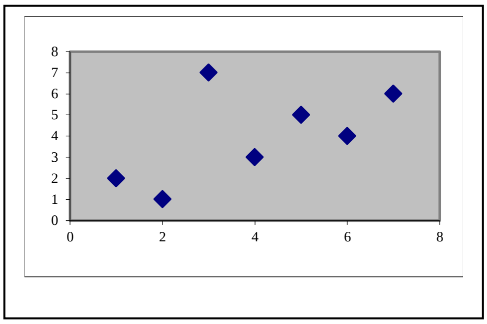
τ (tau) = (P-Q) / (P+Q) Zahlenraum, wie gehabt -1 …. +1

Gibt es einen Zusammenhang zwischen den Schulabschlüssen von Kinder und Eltern ?

Schulabschluss							
Student	X	Abi	Haupt	ohne	Real		
Vater	Y	Real	ohne	Haupt	Abi		
Ordungszahlen	X	1	3	4	2		
	Y	2	4	3	1		
geordnet nach X	X	1	2	3	4		
	Y	2	1	4	3		
folgend größer	P_i (Y)	2	2	0	0	P = S	4
folgend kleiner	Q_i (Y)	1	0	1	0	Q = S	2
tau = (P-Q) / (P+Q)		2	6		**0,33**		

FS Wein

Sieben Weinsorten wurden nach
Geschmack (X) und Bekömmlichkeit (Y) getestet.
Die Siebener-Likertskala ist eine ordinale Skala.



Weinsorte	**a**	**b**	**c**	**d**	**e**	**f**	**g**
Geschmack (X)	**4**	**3**	**6**	**2**	**7**	**1**	**5**
Geruch (Y)	**3**	**7**	**4**	**1**	**6**	**2**	**5**

Bilde die Rangziffern der Bewertungen:

Weinsorte							
Geschmack (X)							
Geruch (Y)							
Pi							
Qi							

FS rang (S.14), ermitteln Sie den Kendall tau

4.6.1.5 Sonstige Ähnlichkeitsmaße

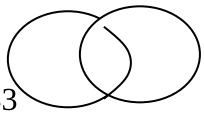
Jaccard-Koeffizient für mengentheoretische Zusammenhänge

Datenfolge A, Datenfolge B

Anzahl Daten ($A \cap B$) / Anzahl Daten ($A \cup B$) (A und B) bezogen auf –durch- (A oder B)

Zahlenraum ?????

FS Zeichen-ähnlichkeiten



A: 11 5 6 8 B: 6 7 11 22 33

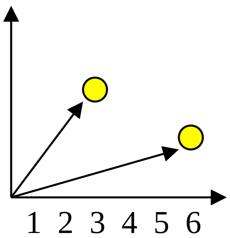
A: Klausur - B: Urlaub

A und B : a l r u
A oder B : K l a u s r U b

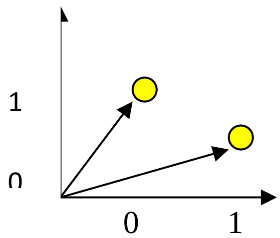
4/8 : 0,5 --> relativ mittlere Übereinstimmung
wenn egal ob groß oder Klein : 4/7 = 0,57

Cosinus-Ähnlichkeit Vektor-Winkel berechnen von Null bis Eins (bis 90°)

X	Y
6	2
3	4



X	Y
0	1
1	0



Zahlenraum (cosinus (Winkel 0 bis 90°) = von Null bis eins)

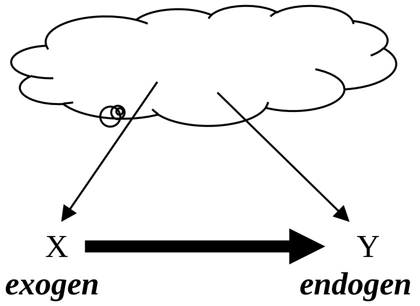
je enger der winkel desto besser --> desto näher ist der cosinus an 1 : 1= starke korrelation

4.6.2 Wirkungs-Analyse (Regression) (zwei Merkmale) nur bei kardinaler Skala

- techn.: Ausgleichsgerade –

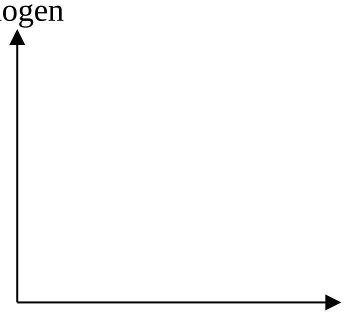
Modell **:** ***endogen* = *Funktion (exogen)***

$Y = f(X)$



FS Fingerlänge, Körperlänge

endogen



exogen

oder ***FS werk (ANHANG 2)***

	X_i	Y_i
Tag	Prod.menge	Prod.kosten
MO	**180**	**5**
DI	**185**	**8**
MI	**170**	**4**
DO	**175**	**6**
FR	**190**	**7**

4.6.2.1 Lineare Regression (OLS: Ordinary Least Squares)

Schätzung der Gerade mit Hilfe der „Kleinst-Quadrat-Methode“)

Regressionsgerade

$$\mathbf{Y = f\ (X)} \qquad \mathbf{(Y = a + b\,X)}$$

Berechnung:

$$\mathbf{b} = \frac{\mathbf{COV}}{\mathbf{VAR_X}}$$

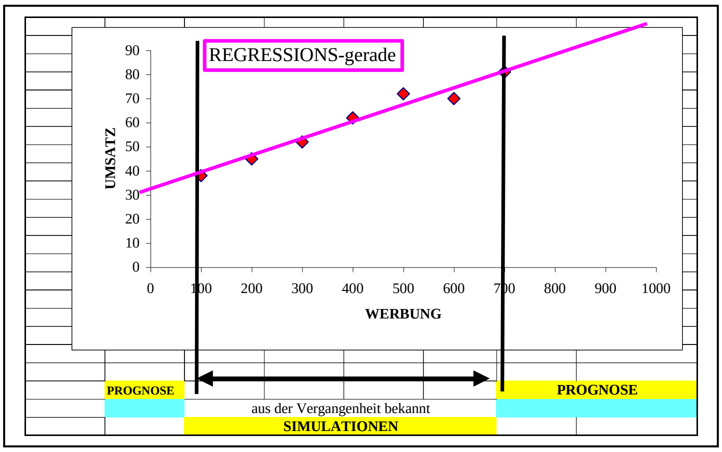
Steigung b = Covarianz durch die Varianz der ***exogenen*** Größe

$$\mathbf{a} = Y^M - b\,X^M$$

danach der **Achsenabschnitt a** nach der Formel

a = Durchschnitt Y - Steigung x Durchschnitt X

FS Werbung



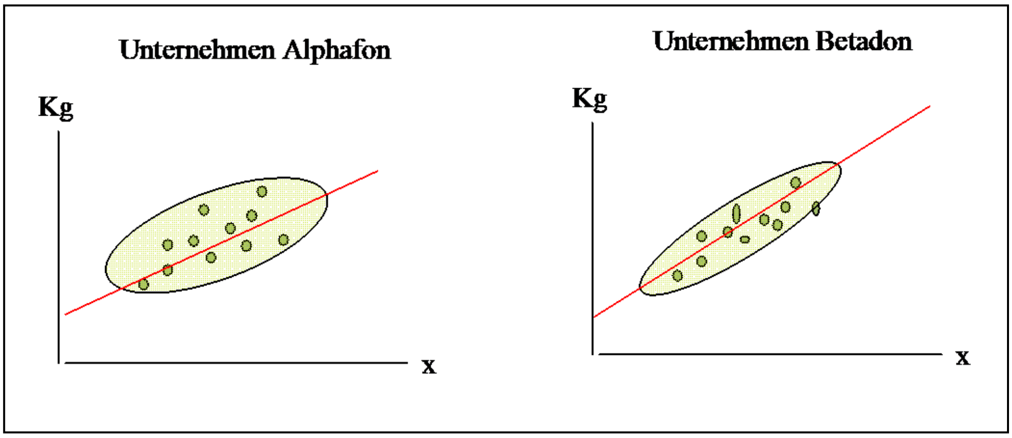
Jahr (X)	Werbung (Tsd Euro)	Umsatz (Mill Euro)
2012	100	38
2013	200	45
2014	300	52
2015	400	62
2016	500	72
2017	600	70
2018	700	81
2019	xxx	???

EXCEL :
$y = 0{,}0711x + 31{,}571$
$R^2 = 0{,}9674$

☞ **Interpolation** Simulationsmodelle
☞ **Extrapolation** Prognoserechnungen

Zur ***Güte*** der Regression

FS Vergleich



Determinationskoeffizient (**R^2**; Bestimmtheitsmaß) $0 <= R^2 <= 1$
Ein Muß für die Regressionsrechnung wie die Streuung zum Mittelwert

(im linearen Fall ist der Det.koeffizient gleich dem quadrierte Korrelationskoeffizient r^2)

4.6.2.2 Nichtlineare Regression

FS Bremsweg (Papula, S. 736)

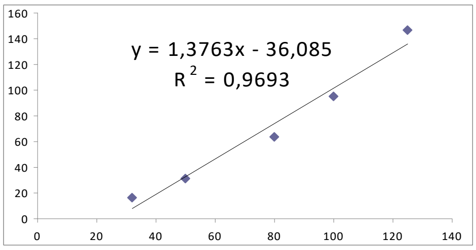
Auf einer Teststrecke ergab sich für ein Auto bei unterschiedlichen Geschwindigkeiten (v) verschiedene Bremswege (s) folgende Werte bei fünf Testläufe.

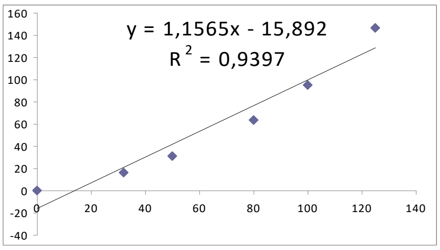
	1	2	3	4	5
v (km/h)	32	50	80	100	125
s (in Meter)	16,2	31,0	63,5	95,0	146,5

	1	2	3	4	5
v (km/h)	32	50	80	100	125
s (in Meter)	16,2	31	63,5	95	146,5

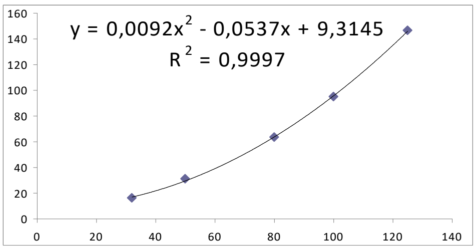
mit dem "technischen" Null/Null-Punkt

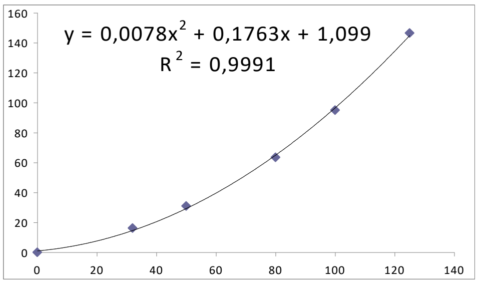
0	32	50	80	100	125
0	16,2	31	63,5	95	146,5





LINEAR





NICHT-LINEAR

4.6.2.3 Multiple Regression (mehrere exogene Faktoren) ; mehrdimensional

$$Y^{endogen} = f(X_1, X_2, X_3, \ldots)^{exogen}$$

Bsp. Körpergewicht

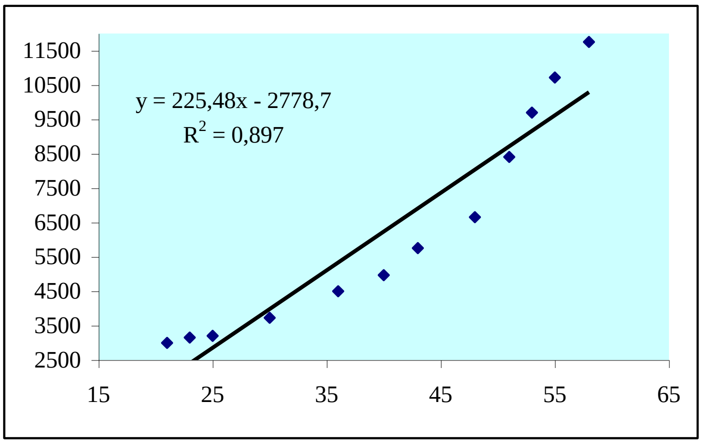
Gewicht = f (Größe, Ernährung, Sportaktivität, Gewicht(Mutter), Gewicht(Vater), ……)

X1 und x2 und etc werden normalerweise addiert

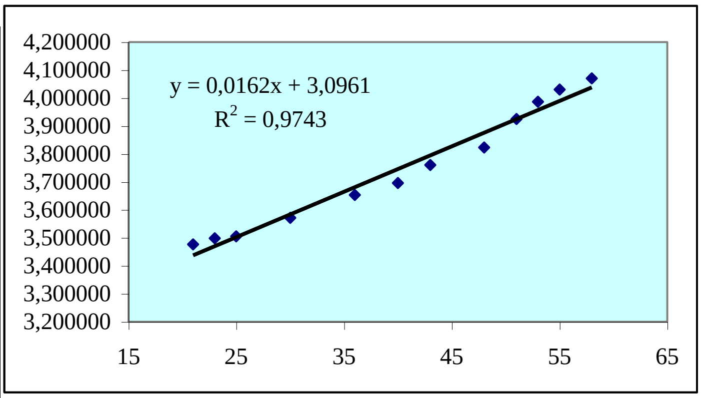
4.6.2.4 Skalentransformation

Der Zusammenhang bleibt, wenn man die Skalenwert „transformiert“, nur die „Art“ des Zusammenhanges ändert sich (nichtlinear → linear)

	X	Y
	exogen	endogen
	Alter	Einkünfte/Monat
Personen	Jahre	€
1	23	3150
2	43	5760
3	55	10720
4	30	3730
5	58	11760
6	36	4500
7	51	8410
8	40	4970
9	25	3200
10	21	3000
11	48	6660
12	53	9700



	X	log (Y)
	exogen	endogen
	Alter	Einkünfte/Monat
Personen	Jahre	€
1	23	3,498311
2	43	3,760422
3	55	4,030195
4	30	3,571709
5	58	4,070407
6	36	3,653213
7	51	3,924796
8	40	3,696356
9	25	3,505150
10	21	3,477121
11	48	3,823474
12	53	3,986772

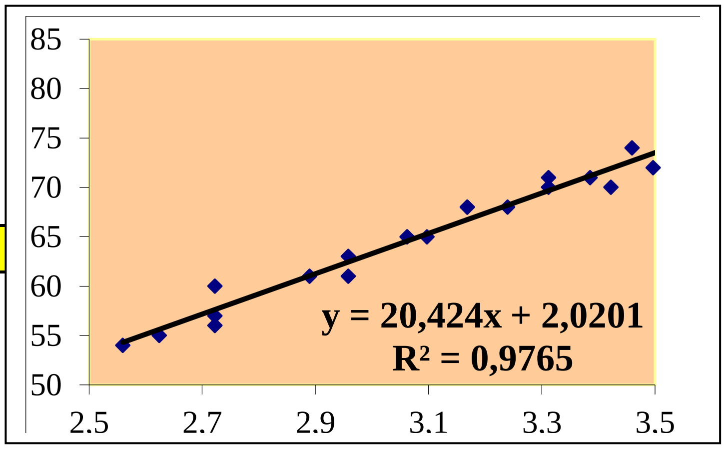


EXKURS Fingerlänge/Körpergröße

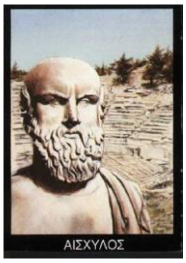
Transformieren Sie die exogenen Merkmalswerte

FS BMI (Spieldaten)

Gewicht = 20,4 * Größe^2



5 Induktive Statistik



intro Schildkröte
Tod des Aischylos (456 v. Chr.):
Zufall oder nicht ?

5.1 Wahrscheinlichkeit / ODDs

DEF: -Die **Wahrscheinlichkeit** p (lat. probabilitas) ist ein Maß, welches das Verhältnis der günstigen Möglichkeiten (Ereignisse) zu allen Möglichkeiten (Ereignissen) angibt.

Die Wahrscheinlichkeit ist die formale mathematische Ausdrucksform der Unsicherheit.

Ein fairer Würfel wird geworfen, mit welcher Wahrscheinlichkeit fällt die vier ?
Zahlenraum 0 (unmöglich) bis 1 (sicher)

Ein unfairer Würfel wird geworfen, mit welcher Wahrscheinlichkeit fällt die vier ?
relative Häufigkeit als Schätzung von p (Gesetz der großen Zahl)

Arten von Wahrscheinlichkeiten:
objektive Wahrscheinlichkeit
a priori (im vorhinein, logisch, deduktiv, mathematisch) ex ante
a posteriori (im nachhinein, statistisch, induktiv, empirisch) ex post
relative Häufigkeit als Schätzung; Gesetz der großen Zahl
subjektive Wahrscheinlichkeit (aus dem Bauch heraus)

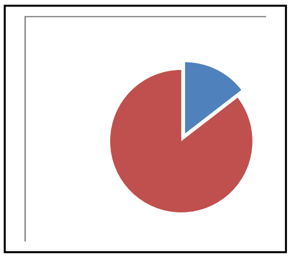
FS Zocker Spezial one
Wie hoch ist die Wahrscheinlichkeit mit einem Würfel eine **gerade Zahl** zu würfeln 3/6 = 0,5 = 50%

FS Zocker Spezial two
Wie hoch ist die Wahrscheinlichkeit mit zwei Würfel die **Summe „7“** zu erzielen ? 6/36 = 1/6 = 0,16
7 = 1 +6; 2 + 5; 3+4; 4+3; 5+2; 6+1

36 Möglichkeiten, da 6 Möglichkeiten pro Zahl (1 bis 6)

DEF: **Odds** (Chance) ist das Verhältnis zwischen der **Wahrscheinlichkeit**, dass ein Ereignis eintritt zu der Wahrscheinlichkeit, dass es nicht eintritt.
formal: p(A) / (1 – p(A))

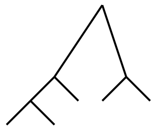
FS Vegan
Wenn von 100 Menschen 16 Veganer sind,
ist die p = 16/100 **(= 0,16)**
Die odds (Chance) für einen Veganer betragen „16 zu 84“;
16/n zu 84/n = 0,16/0,84 = „16/84“ **(= 0, 19)**



FS Lotterie
In einer Lotterie gibt es 500 Lose mit 120 Gewinne.
Die Wahrscheinlichkeit (p) eines Gewinnes ist 0,24 (=120/500)
Die Chance (odds) eines Gewinnes ist 0,32 (=120/380)

5.2 Wahrscheinlichkeitsrechnen, Entscheidungsbaum
5.2.1 Grundlegende Axiome

FS Karten (32 Karten, vier Asse)



Es werden Karten aus einem Skatblatt gezogen.
Wie groß ist die Wahrscheinlichkeit x Asse zu ziehen ?
Darstellung mit dem **Entscheidungsbaum** (hierarchische Struktur)

PROBLEM: unabhängige / abhängige Experimente *(mit oder ohne Zurücklegen)*
- ☞ **unabhängige** (disjunkte) Ereignisse
- ☞ **abhängige** (nichtdisjunkte) Ereignisse) - ***bedingte Wahrscheinlichkeiten*** -

Logisches **ODER** : **Addition** der Wahrscheinlichkeiten
Logisches **UND** : **Multiplikation** der Wahrscheinlichkeiten

FS Würfeln

Wie groß ist die Wahrscheinlichkeit erst beim zweiten Mal eine „6" zu würfeln ?
Wie groß ist die Wahrscheinlichkeit mit drei Würfen die Reihe „1 2 3" oder einen SechserPasch (6 6 6) zu würfeln.

FS Roulett 1/37 = 0,027

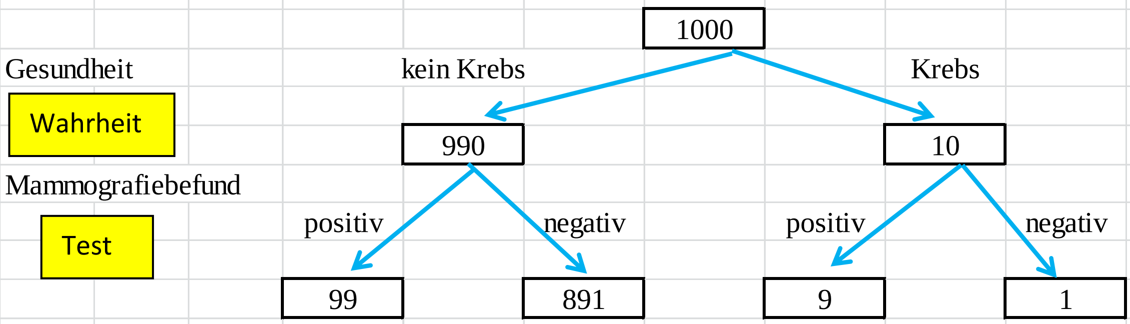
Wie groß ist die Wahrscheinlichkeit für die 17 ? (Zahlen 0 bis 36)
Wie hoch, dass sie dreimal hintereinander kommt ? 1/37 x 1/37 x 1/37=0,0000197
Wie hoch, dass sie dann ein viertes Mal noch mal kommt ? 1/37
Wie hoch, dass sie viermal hintereinander kommt ?
1/37 x 1/37 x 1/37 x 1/37 = 0,00000053

Am 18.5.1913 kam in Monte Carlo 17 mal schwarz hintereinander. Immer mehr Spieler kamen hinzu.. Erst beim 27. Mal fiel die Kugel auf rot.

Objektive Theorie und subjektives Wahrnehmen unterscheiden sich manchmal:

FS Mammografie (Achtung: Kopf-, Bauchgefühl)

1000 Frauen wurden auf Brustkrebs durch eine Mammografie untersucht (didaktisches Beispiel)		
Annahmen:	Ein Prozent aller Frauen haben Krebs	
	Bei Frauen mit Krebs liefert die Mammografie zu 90% einen positiven Befund	***Sensitivität***
	Bei Frauen ohne Krebs liefert die Mammografie zu 90% einen negativen Befund	***Spezifität***



ACHTUNG Es sind insgesamt 108 positive Befunde davon aber nur 9 korrekte !!!!

falsch positiv (10%) ***falsch negativ (10%)***
Spezifität (90%) ***Sensitivät (90%)***

5.2.2 Predictive Analytics (PA) als Bsp. Für KI

Übergangswahrscheinlichkeiten, Markoff´sche Ketten

KI: künstliche Intelligenz (*engl. AI*)
besser: Maschinenlernen
– Lernen heißt Verhalten ändern –

intro Goldburger McFritz

Für eine *Werbekampagne*: beim Eintritt eines Kunden brauchen wir KI,
wenn der neue Kunde evtl. einen Goldburger kaufen will, soll er mit einem Gutschein motiviert werden (potentieller Käufer: Gutschein; sonst kein Gutschein).
Vorgabe: vollautomatisiert.

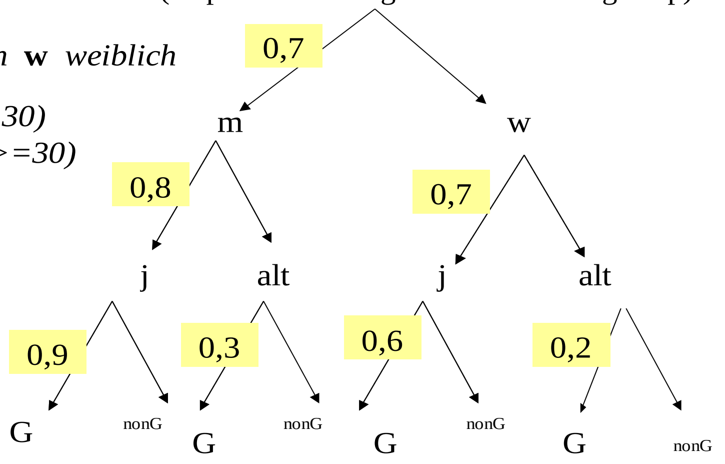
Daten der Vergangenheit:

Erfassung	**Kunde**	**Geschlecht**	**Jung/älter**	**Goldburger**
	Nr	w / m	j / alt	G / nonG
	1	w	j	nonG
	2	m	alt	G
	…..			

Aus der Kundenstatistik: (empirische Häufigkeit als Schätzung für p)

m *männlich* **w** *weiblich*

j *jung (< 30)*
alt *älter (>=30)*



G: kauftGoldburger

Folgerungen für die KI

Jemand betritt den Laden: mit welcher Wahrscheinlichkeit ist er
männlich, jünger und kauft einen Goldburger ?
Jemand betritt den Laden: mit welcher Wahrscheinlichkeit ist er
weiblich, älter und kauft keinen Goldburger ?
Wenn jemand kommt, mit welcher Wahrscheinlichkeit kauft er einen Goldburger ?

Prognose ***lernt*** selbständig:
rel. Häufigkeit aus der Kundenstatistik ist geschätzte Wahrscheinlichkeit;
wird sich mit jedem Kunden ändern – ***fortlaufendes Lernen***)

5.3 Bedingte Wahrscheinlichkeiten (Bayes)
5.3.1 Einführung

FS Bar Bar.xls (vgl. ANHANG 3)
In einer Bar sind 100 ***Personen***. Darunter befinden sich 60 Männer, 40 Raucher und 20 ***männliche Raucher***.
Sachdaten sind rot unterlegt !!

		m	*non m*	Summen
		B	non B	
R	A	20		
non R	non A			
	Summen			100

- Eine Person wird zufällig ausgewählt, um ein Freigetränk zu bekommen:
  Mit welcher Wahrscheinlichkeit ist die Person ein männlicher Raucher ?
  p (R ∩ m) = ??? INFO: Anzahl Personen (alle) = 100 **p = 20/100 = 0,20**

- Die ausgewählte Person ist männlich (Annahme; BEDINGUNG; EVIDENZ)
  mit welcher Wahrscheinlichkeit raucht sie ?
  p (R | m) = ??? ZUSATZINFO: Anzahl Männer = 60 **p = 20/60 = 0,33**

		m	*non m*	Summen
		B	non B	
R	A	20		
non R	Non A			
	Summen	**60**		100

- Die Person ist Raucher, mit welcher Wahrscheinlichkeit ist sie männlich ?
  p (m | R) = ??? ZUSATZINFO: Anzahl Raucher = 40 **p = 20/40 = 0,5**

		m	*non m*	Summen
		B	non B	
R	A	20		**40**
non R	Non A			
	Summen			100

Grafische Zusammenfassung: formale Definition:

Anteile (in Prozent)

	B (m)		
A (R)	20		40
	60		100

p (R$\cap$m) bezogen auf ***alle***
R$\cap$m / n = 20/100 = 0,2
p (R$\cap$m) bezogen auf Evidenz :***männlich***
R$\cap$m / 60 = 20/60 = 0,33
p (R$\cap$m) bezogen auf Evidenz: ***Raucher***
R$\cap$m / 40 = 20/40 = 0,5

5.3.2 Bayes Formel für bedingte Wahrscheinlichkeiten (Evidence)

Bayes Formel (für bedingte p) p(A|B) = p(A$\cap$B) / p(B)
P (A) unter der Evidenz –Annahme–, dass B vorliegt.

Geht auch mit Wahrscheinlichkeiten statt den absoluten Werten (Bayes)

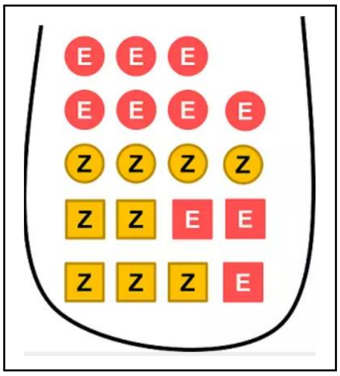
	B (m)		
A (R)	0,20		0,40
	0,60		1

p (A) = 0,40 p(B) = 0,60

p(A|B) bezogen auf Evidenz B:***männlich***
p(A|B) = p(A$\cap$B) / p(B) = 0,20 / 0,60 = 0,33
p(B|A) bezogen auf Evidenz: ***Raucher***
p(B|A) = p(A$\cap$B) / p(A) = 0,20 / 0,40 = 0,5

FS Bayes-Bonbons

In einer Bonbontüte befinden sich 19 Bonbons, 11 Bonbons sind rund und 8 quadratisch.
Von den runden Bonbons haben 7 den Geschmack Erdbeere und 4 nach Zitrone.
Von den quadratischen Bonbons haben 3 Erdbeer-Geschmack und 5 Zitrone.



Es wird eines der 19 Bonbons blind gezogen:

a) Wie groß ist p(rund)
b) Wie groß ist p(Erdbeere)

c) Wie groß ist p(Nicht-Zitrone)
d) Wie groß ist p(quadratisch und Zitrone)
e) Wie groß ist p(Erdbeere und rund)

f) Es wird ein Erdbeer-Bonbon gezogen, Wie groß ist die Wahrscheinlichkeit, dass dieses Bonbon rund ist p(rund|Erdbeere) ?
g) Es wird ein quadratisches Bonbon gezogen, wie groß ist die Wahrscheinlichkeit, dass dieses Bonbon Erdbeergeschmack hat p(Zitrone|quadratisch) ?

Lö a) 11/19 b) 10/19 c) 10/19 d) 5/19 e) 7/19 / 10/19 = 10/19 f) 5/19 / 8/19 = 5/8 ohne Gewähr

Übung FS Disco (ANHANG 5a; 5b)

5.3.3 Bayes Satz

Intro Frage

An der Uni Transsylvanien sind in dem Fach "Kosmologie" nur 8% Frauen, im zweiten Semester gibt es die Angstklausur "Theoretische Physik", die nur 50% aller Studierenden bestehen.
Wenn von den Frauen nur 25 % die Klausur "Physik" bestehen, wie groß ist dann die Wahrscheinlichkeit, dass eine Frau besteht ?

Bayes Formel

p(A|B) = p(AundB) / p(B)
p(B|A) = p(AundB) / p(A)

Satz von Bayes, Herleitung

bedingt eins	p(A|B) = p(AundB) / p(B)
umformen	p(AundB) = p(A|B) * p(B)
bedingt zwei	p(B|A) = p(BundA) / p(A)
umformen	p(BundA) = p(B|A) * p(A)
gleichsetzen	p(A|B) * p(B) = p(B|A) * p(A)
Satz Bayes	**p (B|A) = p (A|B) * (p(B) / p(A))**

Zur Intro Frage

		A		
	in %	Frau	Mann	Summe
B	bestanden	x		50
	nicht b.			
	Summe	8		100

Wenn von den Frauen nur 25 % die Klausur "Physik" bestehen
p(A|B) =p(best|Frau)= 0,25
wie groß ist dann die Wahrscheinlichkeit, dass eine Frau besteht ?
p(B|A) = p(best|Frau) = ?

Lösung Tabelle

		A		
		Frau	Mann	Summe
B	bestanden	2		50
	nicht b.			
	Summe	8		100

x = 8 *0,25 = 2
2/50 = 2/50 = 0,04 (4%)

Es sind nur 4% !!!!

		A			
	p	Frau	Mann		Summe
B	bestanden	0,20		0,50	50
	nicht b.				
	Summe	0,08		1	100

p (B|A) = p (A|B) * p(B) / p(A)

p(B|A) = 0,25 * 0,08 / 0,5 = 0,04

Die p ist nur 0,04 !!!!

Übung Lungenkrebs (Anhang 6)

Auch **mehrdimensional** gültig:

Evidenz/Annahme läßt sich teilen A: A_1, A_2, ….A_n

$P(B|A_i) = [\ p(A_1|B) * p(A_2|B) * \ldots\ldots\ p(A_n|B)\ ] * p(B)\ /\ p(A)$

[Annahme: die Unterteilung ist disjunkt, die einzelnen Evidenzen sind unabhängig]

5.4 EXKURS Klassifizierung mit dem Naiven Bayes Algorithmus (naive baysian classifier)

Klassifizierung: Eine Gruppe von Elemente wird in Klassen (class, labels) eingeteilt.
Diese Einteilung kann „automatisch“ mit Hilfe von KI erfolgen.
Dabei muss aber der **Algorithmus** zuerst angelernt (supervised, antrainiert) werden.
Erst nach dieser Testphase arbeitet der Algorithmus selbständig.

Modell								
	Eine Menge (Masse) von Elementen (Objekten: Menschen oder Dinge) haben							
		verschiedene Merkmale (Evidenzen; Offensichtlichkeiten)						
	Aufgrund dieser Evidenstrukturen sollen die Datensätze (Objekte) klassifiziert werden.							
	Grundgesamtheit			***objects***				
					Klassifizierung	C1		
					Kategorisierung			
		MERKMALE			***labels***	C2		
	e1	e2	e3	***evidences***				
Beispiel	SPAM-FILTER							
	Struktur der	Vorkommen	Anrede ohne		Klassifizierung	C1	SPAM	*ja*
	email-Adresse	Geld	Fehler		***labels***	C2	kein SPAM	*nein*
	Aufgrund historischer Daten (Trainings-, Übungs-, Anlernphase)							
		soll ein Algorithmus automatisch eine Zuordnung zu einer Klasse vornehmen.						
	Algorithmus	naive Bayes classification			$p(\text{Ereignis} \mid \text{Evidenz})$			
		Satz Bayes	p (B/A)	$= p(A \mid B) * (p(B) / p(A))$				
			posterior	= likelihood * (prior / evidence)				

„naiv“: wegen der Annahme, dass die Merkmale/Bedingungen (evidences) unabhängig sind.

THEORIE

Wenn A in **unabhängige** Teilmengen darstellbar ist (A_i), dann gilt allgemein

Satz Bayes

$$p(B|A) = \frac{p(A|B) * (p(B)}{p(A)} \qquad p(B|A_i) = \frac{\Pi\, p(A_i|B) * (p(B)}{p(A)}$$

a posteriori $\dfrac{\textit{Produkt der apriori} * p(B)}{p(\textit{Evidenz})}$

$$p(C1|A_i) = \frac{\Pi\, p(A_i|C1) * (p(C1)}{p(A)}$$

Wahrscheinlichkeit der Klasse C_i bei gegebenen Evidenzen
Wenn es nur zwei Klassen gibt (C1 und C2)

$$p(C2|A_i) = \frac{\Pi\, p(A_i|C2) * (p(C2)}{p(A)}$$

Datensatz Zuordnung in die Klasse mit der höheren p; vom Nenner unabhängig !!!!

$$\boxed{\Pi\, p(A_i|C1) * p(C1) \;\; >??< \;\; \Pi\, p(A_i|C2) * p(C2)}$$

Zuordnung in die wahrscheinlichere (offensichtliche) Klasse

FS Bierklassen

Evidenzen: Geschlecht (m: männlich, w: weiblich; Führerschein(F; kF: kein Führerschein)
Klassen: Biertrinker (ja / nein) Testphase: neun Personen.

Erste algorithmitische Zuordnung welche Klasse (weiblich und Führerschein) ?

Trainingsphase
Anlernphase

	Evidenz E		Klasse
	e1	e2	C
1	m	F	ja
2	w	kF	nein
3	w	F	ja
4	m	F	nein
5	m	kF	nein
6	w	F	nein
7	w	kF	ja
8	m	F	ja
9	w	kF	nein
10	**w**	**F**	**???**

nach der Frage geordnet

Evidenz E		Klasse
e1	e2	C
w	F	ja
w	kF	ja
m	F	ja
m	F	ja
w	F	nein
w	kF	nein
w	kF	nein
m	F	nein
m	kF	nein
w	F	???

p(ja | w;F) = p(w|ja)*p(F|ja)*p(ja)

2/4 3/4 4/9

0,50 *0,75* *0,44* **0,17**

p(nein | w;F) = p(w|nein)*p(F|nein)*p(nein)

3/5 2/5 5/9

0,6 0,4 0,556 **0,13**

Bayes ohne Nenner !
keine Wahrscheinlichkeiten !

Einordnung in die Klasse der Biertrinker.

FS Studierende Besteht die Nr. 13 die Klausur oder nicht ? (Klasse/label)

Studierende in die Klassen ja: C1 (bestehen die Klausur)
nein: C2 (bestehen die Klausur nicht)

Geschlecht
m: männlich
w: weiblich

Anwesenheit in Vorlesung
i: immer
o: oft
s: selten

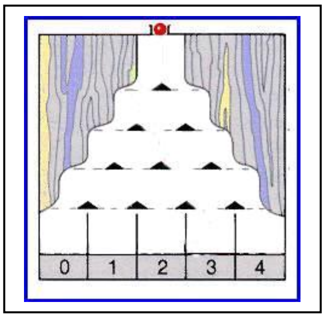
Vorbereiten/ Übungen
k: keine
a: ab und zu
v: viel

geordnet nach C

	e1	e2	e3	C (Klasse)
1	w	i	v	ja
2	w	o	v	ja
3	w	o	a	ja
4	w	i	v	ja
5	w	i	k	ja
6	m	o	a	ja
7	w	s	a	ja
8	w	i	k	ja
9	m	s	a	nein
10	w	o	k	nein
11	m	s	k	nein
12	m	o	a	nein
13	**w**	**o**	**a**	**?????**

6 Zufallsprozesse
6.1 Theorie des Zufall

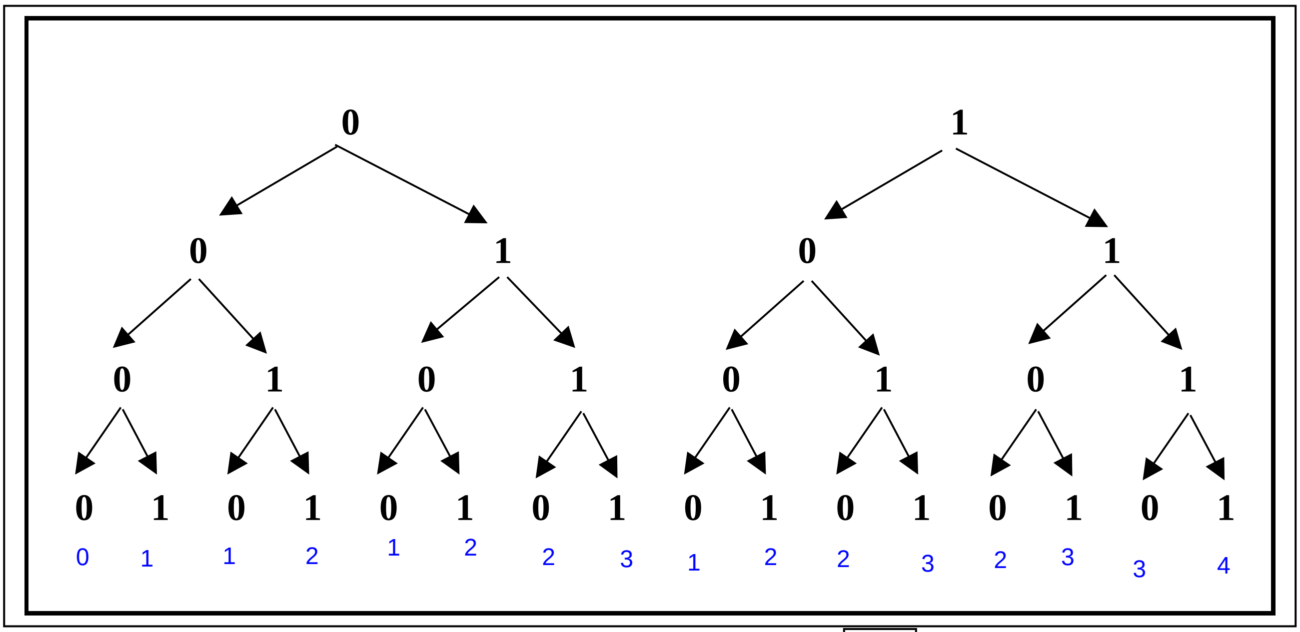
FS Galton
Galton´sches Brett



F. Galton (1822 – 1912)
Auch Begründer der Regression,

ZUFALL ist die Gleichzeitigkeit des **kausal** Nichtzusammenhängenden (A. Schopenhauer)

FS pascal (digital: null, eins)
Merkmal Binäre Zahlen: Stellen; Ziffern („0“, „1“) Messgröße: Summe der Ziffern



Bei einem Lauf (0, 1)

Summe	0	1
Häufigkeit (n)	**1**	**1**

einfach

Bei zwei Läufen (00, 01, 10, 11)

Summe	0	1	2
Häufigkeit (n)	**1**	**2**	**1**

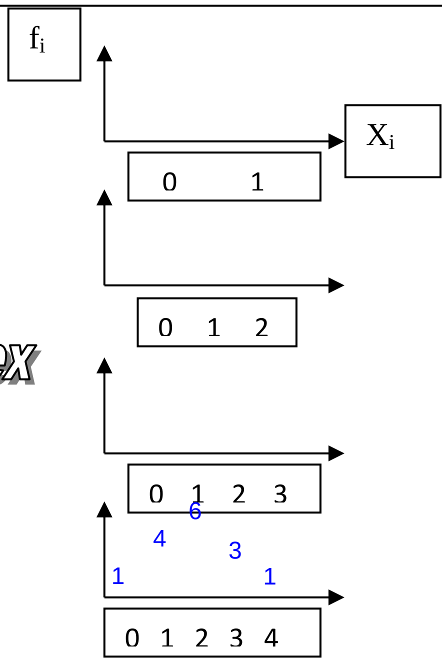
Bei drei Läufen (000, ... , 111)

Summe	0	1	2	3
Häufigkeit (n)	**1**	**3**	**3**	**1**

komplex

Bei vier Läufen (0000, ..., 1111)

Summe	0	1	2	3	4
Häufigkeit (n)	1	4	6	4	1



.....“Ballung um die MITTE“, der Zufall arbeitet zentralisiert, nicht gleich !!

6.2 Erwartungswert (expected value) esperance : „mathematische Hoffnung"

Vorab:
Wenn man alle Ergebnisse eines Experimentes kennt und alle Wahrscheinlichkeiten, dann muss die Summe der Wahrscheinlichkeiten *Eins* sein !!!!!!!!

MERKE wenn alle Möglichkeiten und p_i bekannt sind, dann ist die Summe der $p_i = 1$ ($\Sigma p_i = 1$)

THEORIE

FS würfel I Wurf mit einem Würfel, Gewinn = Augenzahl in Euro .
Frage: welchen Gewinn haben Sie bei einem Wurf ***durchschnittlich*** zu erwarten ?

X_i (€)	p_i	
1	1/6	1/6
2	1/6	2/6
3	1/6	3/6
4	1/6	4/6
5	1/6	5/6
6	1/6	6/6
	E (X)	3,5

$(X_i - E(X)^2$

Erwartungswert E(X) $= \Sigma\, X_i\, p_i$ wenn gilt: $\Sigma\, p_i = 1$

= zukünftige durchschnittlich zu erwartende Wert (Zukunfts-***Mittelwert***)

Erwartungswert E(X) ist der "theoretische" Mittelwert, der a priori bestimmt werden kann. (der Wert der "durchschnittlich" zu erwarten ist)

Berechnen Sie: Erwartungswert E(X); Standardabweichung s (X);Variationskoeffizient v(X)

$\mathrm{VAR}(X) = 1/n\, \Sigma (Xi - E(X))^2$ $\quad s(X) = \sqrt{VAR}$ $\quad v(X) = s / X^M$ (ANHANG 4)

eigentlich ist der Erwartungswert nicht zu erwarten

FS gezinkter Würfel

Berechnen Sie den Erwartungswert eines gezinkten Würfel.

Wie in der Schule gelernt

X_i (€)	p_i	
1	0,5	0,5
2	0,1	0,2
3	0,1	0,3
4	0,1	0,4
5	0,1	0,5
6	0,1	0,6
	E (X)	2,5

FS Roulett (37 Zahlen)

Wenn ein Spieler 100 Euro auf eine Zahl setzt, wie hoch ist dann der Erwartungswert der Spielbank ? Jedes Spiel hat zwei Gesichter; „faires Spiel“

	X_i (€)	p_i	
Spieler verliert	+100	36/37 0,972	
Spieler gewinnt	-3500, da 3600 auszahlung - 100€ einsatz	1/37 0,027	

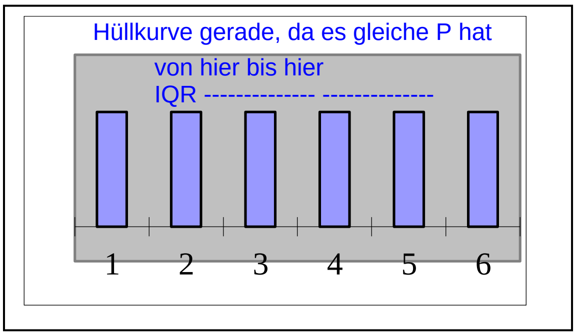
6.3 Analyse des Zufalls

6.3.1 Theorie (Idee)

FS würfel I

WERTE (x_i)	$p(x_i)$
1	0,167
2	0,167
3	0,167
4	0,167
5	0,167
6	0,167
Summe	1,0

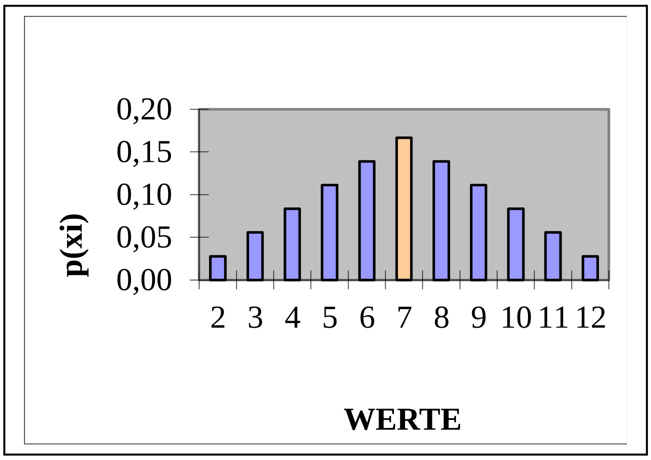
E(X)	**3,50**
VAR(X)	**2,92**
STD(X)	**1,71**
Varkoeff	**48,9 %**



FS würfel II

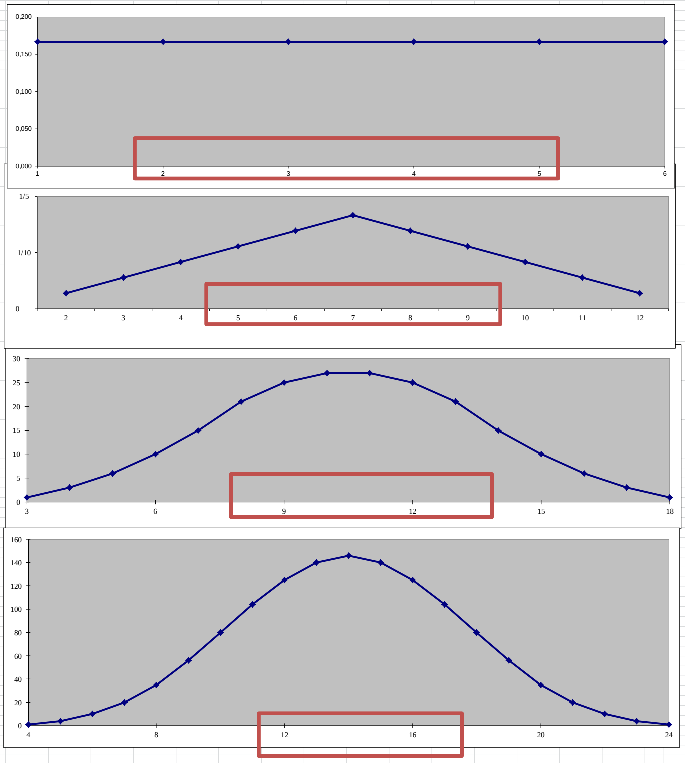
	1	2	3	4	5	6
1	2	3	4	5	6	7
2	3	4	5	6	7	8
3	4	5	6	7	8	9
4	5	6	7	8	9	10
5	6	7	8	9	10	11
6	7	8	9	10	11	12

Und wie sieht das bei zwei Würfeln aus ?



E(X) = 7
STD(X) = 2,42
v(X) = 34,5 %

Ein bis vier Würfel: Übergang von der Gleich- zu der Normalverteilung



Würfel	1	2	3	4
min	1	2	3	4
max	6	12	18	24
E(X)	3,5	7	10,5	14
s	1,71	2,42	2,96	3,42
v (v.H.)	48,9	34,6	28,2	24,4

Sigmabereich

6.3.2 EMPIRIE (Praxis)

FS TEAM würfel III

Jedes Team würfelt mit drei Würfel (Summe der Augenzahlen). Notieren der Ergebnisse.

um 1€ gespielt

FS schere stein papier Spieltheorie

Gegner

S I E

(Ich)

Euro		Schere	Stein	Papier	Brunnen	**E(X)**
		z_0	z_1	z_2	z3	
Schere	a_0	0	-1	+1	-1	- 0,25
Stein	a_1	+1	0	-1	-1	- 0,25
Papier	a_2	-1	+1	0	+1	1/4
Brunnen		+1	+1	-1	0	1/4

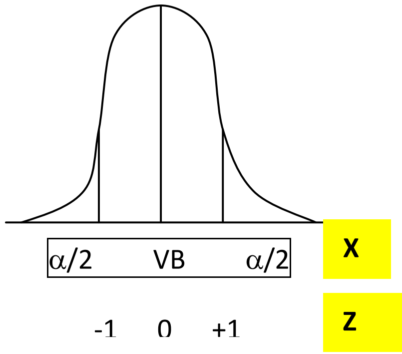
= Auszahlungsmatrix

6.4 Praktisches Arbeiten mit Hilfe von statistischen Tabellen (vgl. ANHANG 7)

Z-Werte: TRANSFORMATION der X – Werte

$$Z = \frac{X - \mu}{\sigma}$$

(am.) standard score: **z-score.** Der Abstand eines Wertes von dem arithmetischen Mittelwert bezogen auf die Streuung.



Übung: Würfelwurf mit drei, Merkmal Augenzahl: min=3; max=18; (μ = 10,5 σ = 3)

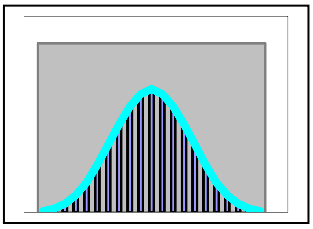
$$Z = \boxed{\frac{X - 10{,}5}{3}}$$

a) X_i : 17 21 8 6 Berechnen Sie die Z – Werte (zweistellig)
b) Lesen Sie bei gegebenen Z-werten die Wahrscheinlichkeiten ab (Anhang 7)

FS Auto

Die Lebensdauer eines Pkws eines bestimmten Typs sei normalverteilt mit den Parametern Mittelwert (μ) = 10 Jahre und der Standardabweichung (σ)= 2 Jahre.
Wenn man einen solchen PKW kauft,
wie groß ist die Wahrscheinlichkeit, dass er eine Lebensdauer

- bis zu 10 Jahren hat
- bis zu 12 Jahren hat



- zwischen 10 und 12 Jahren hat
- zwischen 8 und 12 Jahre hat
- von mindestens 13,5 Jahren hat

- von höchstens 7 Jahren hat
- zwischen 9 und 13 Jahre hat
- zwischen 4 und 10 Jahre hat
- zwischen 6 und 14 Jahre hat

- mindestens 15 Jahre hat
- zwischen 7 2/3 und 13 1/3 Jahre dauern wird.
- zwischen 12 und 13 Jahre dauern wird.

$\boxed{\mu +/- Z\,\sigma}$

- genau 13 Jahre alt wird (!!!: Punktwahrscheinlichkeit)

Betriebliche Praxis:

VORGABE σ:- Lebensdauer im 1, 1,5-..3-σ- Bereich Sigma-bereiche
VORGABE α: - mit einem α = 5% , 1%, 0,05% Signifikanzniveau
VORGABE VB: - Vertrauensbereich von 90%; 95% … Konfidenzintervall

6.5 Qualitätskontrollen (Stichproben) / Hypothesentest / Alphafehler

Innerhalb der betrieblichen Abläufe finden oft sogenannte Qualitätskontrollen statt, im Einkauf, Fertigung und Verkauf – eigentlich überall.
Qualität: ist eine vordefinierte Eigenschaft eines Produktes/Dienstleistung
Qualitäskontrolle: ist eine meist in Form von Stichproben vorgenommene Überprüfung der zugesagten definierten Eigenschaft.
Modell: Die Eigenschaft in der Stichprobe wird ermittelt und dann:

Annahme (true) die definierte Qualität ist gegeben (Null-Hypothese)
Annahme (false) die definierte Qualität ist nicht gegeben (Alternativ-Hypothese)

FS Schrauben

Entwerfen Sie ein Modell der Qualitätsüberprüfung, wenn die Fertigungsqualität verspricht:

Schrauben mit durchschnittlich 5,5 cm bei einer Varianz von 1,5 %
Achtung: ***Alphafehler***

		die WIRKLICHKEIT Grundgesamtheit U M W E L T	
	unser Test „unsere Meinung“	Null-Hypothese (H_0) ist wahr	Alternativ- Hypothese ist richtig
Stich-pro-be	H_0 akzeptieren (Abweichungen nur zufällig)	**O. K.**	**β - Fehler** Fehler 2. Art
	H_0 ablehnen H_a annehmen (Abweichung signifikant)	**α - Fehler** Fehler 1. Art	**O. K.**

FS jura

die WIRKLICHKEIT

unser Test „Indizienurteil“	***ist Täter***	***unschuldig***
verurteilen	***ok***	**β-Fehler**
frei sprechen	**α-Fehler**	***ok***

	richtig	falsch
positiv		falsch/positiv
negativ	richtig/negativ	

liberales Recht: **β-Fehler möglichst klein,** dass heißt aber große α-Fehler tolerieren.

FS heiraten

die WIRKLICHKEIT

unser Test „unsere Meinung“	***ist der Richtige***	***ist der Falsche***
verloben	***ok***	**β-Fehler**
nicht verloben	**α-Fehler**	***ok***

7 Vertiefung: Präsentation von Daten (Listen,Tabellen / Grafiken)

Formaler Aufbau einer Präsentation

Statistik befaßt sich mit SACH-DATEN (=Zeichen)
Visualisieren = ein Bild sagt mehr als 1000 Worte

Daten

Sach daten	Ordnungs- daten

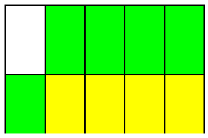
FS Poker

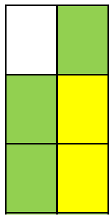
Die fünf Ehepaare Bose, Emal, Faller, Odera und Santa spielen Poker
Die Schuhgrößen der weiblichen Teilnehmer
Anke, Carmen, Friederike, Helga und Marlene sind
39 41 43 36 42 (dt. Schuhgröße)
Die Körpergrößen der männlichen Teilnehmer
Erwin, Hermann, Jürgen, Klaus und Thorsten sind
170 180 195 165 175 (cm)

Eine **LISTE** hat eine Zeile oder eine Spalte

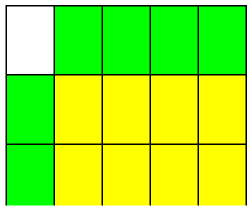
Eine **TABELLE** hat mindestens zwei Zeilen und mindestens zwei Spalten

I Erstellen Sie zwei Listen
Erstellen Sie eine Tabelle
II und Diagramme der beiden Listen
und ein Diagramm der Tabelle





.... Punktwolke (scatter diagramm)



7.1 Liste / Tabelle

DEF: Eine statistische Tabelle (Liste) ist eine möglich genaue, anschauliche, übersichtliche und eventuell aggregierte Zusammenstellung von statistischem Datenmaterial.

Die drei Arten von Nichts (DIN 55301)
wirklich nichts, "-"
Wert zu gering (Rundungsnull) "0"
fehlender Wert "*" bzw. "."

tabelle

C O G A AG & Co., KG

Monatliche Umsatzzahlen (in Mill Euro)

unserer vier Filialen im Jahr 2012

	Januar	Februar	März	April	Mai	summe
Mosbach	12,0	15,3	23,6	0	45,6	
Leimen	4,5[1]	8,2	12,2	154,8[2]	13,5	
Mannheim	-	3,2	*	56,3	89,2	
Paris	125,3	235,6	189,6	156,3	222,2	
summe						

Fußnoten: - kein Umsatz vorhanden
* fehlender Wert
[1] wegen Umbau nur 10 Umsatztage
[2] signifikant abweichend (Großauftrag)
Quelle: Eigene Erhebung (Abt. kolei/zen)

Sachdaten sind natürlich gleich formatiert

7.2 Statistische Diagramme

Visualisieren von statistischen Daten.

1854 brach in London eine Cholera-epidemie aus, an der viele starben. Der Arzt Dr. Stone zeichnete die Todesfälle in einen Stadtplan ein, und erkannte als Ursache eine Wasserleitung mit einer vergifteten Quelle.

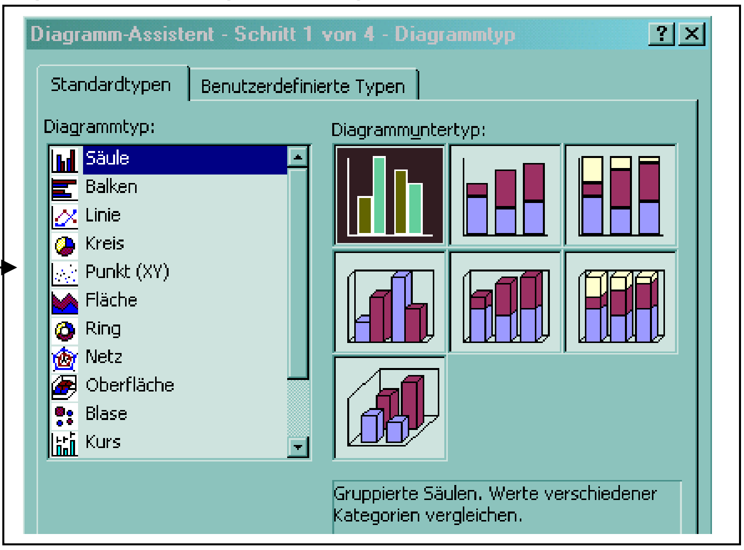
DEF: Eine statistische Grafik soll das Datenmaterial übersichtlich,
schnell lesbar, knapp und einprägsam darstellen.

!!! Das richtige Diagramm ist so wichtig wie das richtige Wort !!!

Arten von statistischen Diagrammen

☞ Punkt-, Stab-, Säulen-, Balken- Diagramme (diskret) - einfach, multipel
☞ Linien-, Kurven- Diagramme (quasi-stetig) - einfach, multipel /halb-logarithmisch
☞ Kreis- Kuchen- Torten- Diagramme (Nur bei Darstellung der Verteilung eines Ganzen)
☞ Bilder- Flächen-, Körper- Diagramme; Kartogramme (landscapes)
☞ Portfolio-Diagramme
☞ Statistische Diagramme: Histogramm, Ogive, Box-Plot

Mathe



FS Umsatz

Die AAA AG Bank hatte folgende Jahresumsätze:

	2019	2020	2021
Umsatz (Mrd €)	41,3	49,8	58,3

Stellen Sie die Umsätze der drei Jahren grafisch dar.

Stellen Sie die Umsatzänderungen grafisch dar.

8 Delphi – Methode

Definition: eine (mindestens) zweistufige Expertenbefragung (subjektive Einschätzung)

Entwickelt von der RAND Corp. für das amerikanischen Militär.

Vorteile: schriftlich, schnell, leicht durchführbar, schneller Rücklauf

FS Durchschnittsalter

Stufe eins: Wie hoch schätzen Sie das Durchschnittsalter Ihrer Gruppe (geheim),
Alter mit einer Nachkommastelle angeben

Stufe zwei: Noch mal schätzen mit Kenntnis des ersten Durchschnittes,
noch mal mit einer Kommastelle (+ eigenes Alter ganzzahlig)

ANHANG 1
FS Zoo

Controlling Für ein Unternehmen mit sieben Zoos sollen die Futterkosten eines Zoos analysiert werden.

Tierart	**Futterkosten €/Tag**
Esel	**77**
Pfau	**25**
Elefant	**275**
Kaninchen	**22**
Zebra	**63**
Papagei	**3**
Tiger	**200**
Enten	**5**
Erdmännchen	**15**
Antilopen	**120**
Flamingo	**33**
Geier	**28**

Unterteilen Sie die Futterkosten

zwei	wichtig/weniger wichtig
	teure / billige
drei	ABC-Analyse
	teure / mittlere / billige
	sehr wichtig / wichtig / weniger wichtig

ANHANG 2
FS werk_lösung

Tag	X_i Prod.menge	Y_i Prod.kosten	**Abweichung** $(X_i - X^M)$	**Abweichung** $(Y_i - Y^M)$	**COV** $_{XY}$	**VAR** $_X$
MO	**180**	**5**	**0**	**-1**	**0**	**0**
DI	**185**	**8**	**5**	**2**	**10**	**25**
MI	**170**	**4**	**-10**	**-2**	**20**	**100**
DO	**175**	**6**	**-5**	**0**	**0**	**25**
FR	**190**	**7**	**10**	**1**	**10**	**100**
n = 5	900	30	0	0	40	250
	X^M	**Y^M**			**COV**	**VAR**
	180	**6**			**8**	**50**

Regressionsgleichung

$$Y^{reg} = -22{,}8 + 0{,}16 * X_i$$

Steigung **0,16** $\frac{\text{COV}}{\text{VARx}}$

Achsenabschnitt **-22,8** $Y^M - b\,X^M$

Eingabe Menge

Ausgabe Kosten

ANHANG 3
FS Bar

bar

In einer Bar sind 100 ***Personen*** . Darunter befinden sich 20 ***männliche Raucher*** .

n	**100**
R ∩ m	**20**

Anteil
relative Häufigkeit
Wahrscheinlichkeit

- Eine Person wird zufällig ausgewählt, um ein Freigetränk zu bekommen:
Mit welcher Wahrscheinlichkeit ist die Person ein männlicher Raucher ?
p (R ∩ m) =

(R∩m)/n	**20/100**	**0,2**

- Die ausgewählte Person ist männlich, mit welcher Wahrscheinlichkeit raucht sie ?
p (R | m) =
Zusatzinfo es gibt 60 männliche

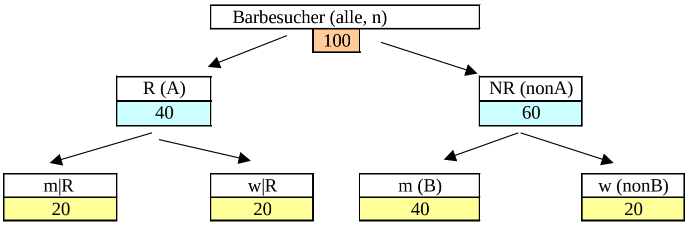
p(R∩m)/m	**20/???**	
p(R∩m)/m	**20/60**	**0,33**

- Die Person ist Raucher, mit welcher Wahrscheinlichkeit ist sie männlich ?
p (m | R) =
Zusatzinfo es gibt 40 Raucher

p(R∩m)/m	**20/???**	
p(R∩m)/m	**20/40**	**0,5**

Strukturanalyse Eine Grundgesamtheit hat zwei Merkmale

		B männlich m	nonB weiblich w	
A	Raucher R	20	20	40
nonA	Nichtraucher NR	40	20	60
		60	40	100



ANHANG 4
Lösung ein Würfel

			Abweichungen	Quadrat	
E(X)	3,50	Xi (€)	X_i - E(X)	$(Xi - E(X)^2$	
		1	-2,50	6,25	
		2	-1,50	2,25	
		3	-0,50	0,25	
		4	0,50	0,25	
		5	1,50	2,25	
		6	2,50	6,25	
		Summe	*0,00*	*17,50*	
			Varianz	***2,917***	**durch n**
			s	**1,71**	**Wurzel**
			v	**48,9%**	**s / E(X) * 100**

ANHANG 5a
Übung Disco

120 Besucher einer Disco werden nach zwei nominalen Merkmalen erfasst.

A	Raucher
non A	Nichtraucher
B	weiblich
non B	nicht weiblich

DATEN

	12	nicht		
	weiblich	weiblich		
absolut	B	non B	Summe	Außenstruktur
A			18	
non A			102	
Summe	48	72	**120**	n

		nicht			
		weiblich	weiblich		
	absolut	B	non B	Summe	
Raucher	A	12	6	18	absolute Werte
Nichtraucher	non A	36	66	102	
	Summe	48	72	**120**	n

relativ (%)	B	non B	Summe	
A	10	5	15	Anteile (in v.H.)
non A	30	55	85	
Summe	40	60	**100**	alle

p	B	non B	Summe	
A	0,10	0,05	0,15	Wahrscheinlichkeiten
non A	0,30	0,55	0,85	
Summe	0,40	0,60	**1**	

Globale Wahrscheinlichkeiten -Analyse der **Außenstruktur** (acht Möglichkeiten)-
Analyse der

p(A) =
p(nonA) =
p(B) =
p(nonB) =

Analyse der **Binnenwerte** in Bezug auf **alle (n)**

P(A∩B) =
P(A∩nonB) =
P(nonA∩B) =
P(nonA∩nonB) =

Bedingte Wahrscheinlichkeiten -Analyse der **Binnenstruktur** (acht Möglichkeiten)-

	absolut	p	**Bayes Satz**		p
p(A|B)	12 / 48	0,25	p(A|B) = p(AundB) / p(B)	0,10 / 0,40	0,25
p(B|A)	12 / 18	0,67	p(B|A) = p(AundB) / p(A)	0,10 / 0,15	0,67
p(A|nonB)			p(A|nonB) = p(AundnonB) / p(nonB)		
p(nonB|A)			p(nonB|A)= p(AundnonB) / p(A)		
P(nonA|B)			p(nonA|B) = p(nonAundB) / p(B)		
p(B|nonA)			p(B|nonA)= p(nonAundB) / p(nonA)		
p(nonA|nonB)			p(nonA|nonB)= p(nonAundnonB)/p(nonB)		
p(nonB|nonA)			p(nonB|nonA)= p(nonAundnonB)/p(nonA)		

Lösung ANHANG 5b

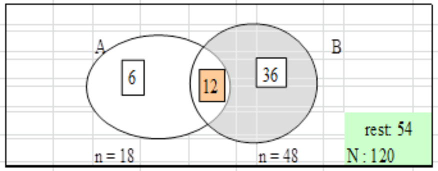
ANHANG 5b
Lösung FS Disco

I Ausgangslage Studienergebnisse

A	Raucher
non A	Nichtraucher
B	weiblich
non B	männlich

absolut	B	non B	Summe	
A	12	6	18	
non A	36	66	102	
Summe	48	72	**120**	N

p	B	non B	Summe	
A	0,10	0,05	0,15	
non A	0,30	0,55	0,85	
Summe	0,40	0,60	1	alle



II Aussenstruktur, global (bezogen auf N)

p(A)	0,15	*15%*
p(non A)	0,85	*85%*
p(B)	0,40	*40%*
p(non B)	0,60	*60%*

III Binnenstruktur, global (bezogen auf N)

P(A∩B)	0,10
P(A∩nonB)	0,05
P(nonA∩B)	0,30
P(nonA∩nonB	0,55

IV Auswertung Binnenstruktur (bedingte Wahrscheinlichkeiten)

		absolut			**Bayes Formel**	
Ausgang	p(A∣B)	12 / 48	0,25	p(A∣B) = p(AundB) / p(B)	=0,10/0,40	0,25
	p(B∣A)	12 / 18	0,67	p(B∣A) = p(AundB) / p(A)	=0,10/0,15	0,67
	p(A∣nonB)	6/72	0,08	p(A∣nonB) = p(AundnonB) / p(nonB)	=0,05/0,60	0,08
	p(nonB∣A)	6/18	0,33	p(nonB∣A)= p(AundnonB) / p(A)	=0,05/0,15	0,33
	p(nonA∣B)	36/48	0,75	p(nonA∣B) = p(nonAundB) / p(B)	=0,30/0,40	0,75
	p(B∣nonA)	36/102	0,35	p(B∣nonA)= p(nonAundB) / p(nonA)	=0,30/0,85	0,35
	p(nonA∣nonB	66/72	0,92	p(nonA∣nonB)=p (nonAundnonB)/p(nonB)	=0,55/0,60	0,92
	p(nonB∣nonA	66/102	0,65	p(nonB∣nonA)= p(nonAundnonB)/p(nonA)	=0,55/0,85	0,65

ANHANG 6
Übung Lungenkrebs
Aufgabe

Die deutsche Gesundheits-Statistik kennt folgende Wahrscheinlichkeiten

	p(Tod durch Lungenkrebs -A-)	**p(A)**	**0,15**
	p(Raucher -B-)	**p(B)**	**0,40**

Wenn die Wahrscheinlichkeit, dass ein Raucher an Lungenkrebs stirbt

	p(Lungenkrebstod|Raucher)	**p (A|B)**	**0,25**

wie groß ist dann die Wahrscheinlichkeit, dass ein Lungenkrebstod auf einen Raucher schließen lässt

	p(Raucher|Lungenkrebstod)	**p (B|A)**	**???**

		Raucher	Nichtraucher	
	relativ	B	non B	Summe
Lungenkrebs	A	0,25 ?		**0,15**
kein L.krebs	non A	0,15		
	Summe	**0,40**		**1,00**

Lösen Sie mithilfe der Tabelle, und dann mit
Satz Bayes **p (B/A) = p (A|B) * (p(B) / p(A))**

Lösung

gegeben	**p(A)**	**0,15**		
	p(B)	**0,40**		
	p (A|B)	**0,25**		

Lösen Sie mithilfe der Tabelle **p (B|A)**

p (A|B) = p(AundB) / p(B) 0,25=p/0,4 = 0,1 !!!!

		Raucher	Nichtraucher	
	relativ	B	non B	Summe
Lungenkrebs	A	**0,10**		**0,15**
kein L.krebs	non A			
	Summe	**0,40**		**1,00**

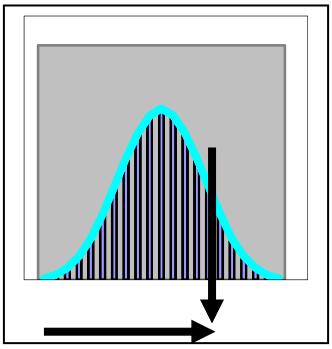
		Raucher	Nichtraucher	
	relativ	B	non B	Summe
	A	0,10	0,05	0,15
	non A	0,30	0,55	0,85
	Summe	0,40	0,60	1,00

ist gleich 0,10 / 0,15 **0,667**

Lösung mithilfe **p (B|A)**
Satz Bayes **p (B/A) = p (A|B) * (p(B) / p(A))**
= 0,25 * (0,4/0,15)
0,667

ANHANG 7
Tabelle: Normalverteilung

Standardnormalverteilung (Fläche von links bis zum positiven Z-Wert)										
Z	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,5000	0,5040	0,5080	0,5120	0,5160	0,5199	0,5239	0,5279	0,5319	0,5359
0,1	0,5398	0,5438	0,5478	0,5517	0,5557	0,5596	0,5636	0,5675	0,5714	0,5753
0,2	0,5793	0,5832	0,5871	0,5910	0,5948	0,5987	0,6026	0,6064	0,6103	0,6141
0,3	0,6179	0,6217	0,6255	0,6293	0,6331	0,6368	0,6406	0,6443	0,6480	0,6517
0,4	0,6554	0,6591	0,6628	0,6664	0,6700	0,6736	0,6772	0,6808	0,6844	0,6879
0,5	0,6915	0,6950	0,6985	0,7019	0,7054	0,7088	0,7123	0,7157	0,7190	0,7224
0,6	0,7257	0,7291	0,7324	0,7357	0,7389	0,7422	0,7454	0,7486	0,7517	0,7549
0,7	0,7580	0,7611	0,7642	0,7673	0,7704	0,7734	0,7764	0,7794	0,7823	0,7852
0,8	0,7881	0,7910	0,7939	0,7967	0,7995	0,8023	0,8051	0,8078	0,8106	0,8133
0,9	0,8159	0,8186	0,8212	0,8238	0,8264	0,8289	0,8315	0,8340	0,8365	0,8389
1,0	0,8413	0,8438	0,8461	0,8485	0,8508	0,8531	0,8554	0,8577	0,8599	0,8621
1,1	0,8643	0,8665	0,8686	0,8708	0,8729	0,8749	0,8770	0,8790	0,8810	0,8830
1,2	0,8849	0,8869	0,8888	0,8907	0,8925	0,8944	0,8962	0,8980	0,8997	0,9015
1,3	0,9032	0,9049	0,9066	0,9082	0,9099	0,9115	0,9131	0,9147	0,9162	0,9177
1,4	0,9192	0,9207	0,9222	0,9236	0,9251	0,9265	0,9279	0,9292	0,9306	0,9319
1,5	0,9332	0,9345	0,9357	0,9370	0,9382	0,9394	0,9406	0,9418	0,9429	0,9441
1,6	0,9452	0,9463	0,9474	0,9484	0,9495	0,9505	0,9515	0,9525	0,9535	0,9545
1,7	0,9554	0,9564	0,9573	0,9582	0,9591	0,9599	0,9608	0,9616	0,9625	0,9633
1,8	0,9641	0,9649	0,9656	0,9664	0,9671	0,9678	0,9686	0,9693	0,9699	0,9706
1,9	0,9713	0,9719	0,9726	0,9732	0,9738	0,9744	0,9750	0,9756	0,9761	0,9767
2,0	0,9772	0,9778	0,9783	0,9788	0,9793	0,9798	0,9803	0,9808	0,9812	0,9817
2,1	0,9821	0,9826	0,9830	0,9834	0,9838	0,9842	0,9846	0,9850	0,9854	0,9857
2,2	0,9861	0,9864	0,9868	0,9871	0,9875	0,9878	0,9881	0,9884	0,9887	0,9890
2,3	0,9893	0,9896	0,9898	0,9901	0,9904	0,9906	0,9909	0,9911	0,9913	0,9916
2,4	0,9918	0,9920	0,9922	0,9925	0,9927	0,9929	0,9931	0,9932	0,9934	0,9936
2,5	0,9938	0,9940	0,9941	0,9943	0,9945	0,9946	0,9948	0,9949	0,9951	0,9952
2,6	0,9953	0,9955	0,9956	0,9957	0,9959	0,9960	0,9961	0,9962	0,9963	0,9964
2,7	0,9965	0,9966	0,9967	0,9968	0,9969	0,9970	0,9971	0,9972	0,9973	0,9974
2,8	0,9974	0,9975	0,9976	0,9977	0,9977	0,9978	0,9979	0,9979	0,9980	0,9981
2,9	0,9981	0,9982	0,9982	0,9983	0,9984	0,9984	0,9985	0,9985	0,9986	0,9986
3,0	0,9987	0,9987	0,9987	0,9988	0,9988	0,9989	0,9989	0,9989	0,9990	0,9990



FORMEL – SAMMLUNG / KLAUSUR

Mean: arithmetischer Mittelwert $X^M = 1/n\ \Sigma X_i$
Modus: häufigster Wert
Median: mittlerer Wert einer geordneten Datenreihe

mittlere absolute Abweichung (Mean Absolute Deviation) $MAD = 1/n\ \Sigma| X_i – X^M|$
Varianz: mittlere quadratische Abweichung $VAR = 1/n\ \Sigma(X_i – X^M)^2$
Standardabweichung: "mittlere Abweichung" $s = \sqrt{VAR}$
Variationskoeffizient:: relative Abweichung in v.H. $v = s / X^M * 100$
Kovarianz:: gemeinsame Abweichung = $1/n\ \Sigma(X_i – X^M) * (Y_i – Y^M)$

Fläche in dem	σ-Bereich beträgt	0,6826	(68,3 %)
	2 σ -Bereich	0,9544	(95,4 %)
	3 σ -Bereich	0,9974	(99,7 %)

Korrelationskoeffizient $r = COV_{XY} / (s_x * s_y)$

Rangkorrelation $\mathbf{r_s} = 1 - \dfrac{6\sum D_i^2}{n^3 - n}$ mit D_i: Differenz der Rangziffern

Kendalls tau = (P – Q) / (P + Q)
$P = \Sigma P_i$: größere Y-Rangziffern; $Q = \Sigma Q_i$: kleinere Y-Rangziffern des Merkmals Y

Vierfelder-Korrelationskoeffizient r_k

$$r_k = \frac{ad - bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}}$$

a	**b**
c	**d**

Jaccard-Koeffizient
Anzahl Daten ($A \cap B$) / Anzahl Daten ($A \cup B$)

Lineares Regressionsmodell (X ist exogen) $Y^{reg} = a + b X_i$
b (Steigung) = COV / Var(X)
a (Achsenabschnitt) = $Y^M - b X^M$

Bayes Formel (für bedingte p) p(A|B) = p(AundB) / p(B)

Satz Bayes p (B|A) = p (A|B) * [p(B) / p(A)]

Erwartungswert $E(X) = \Sigma X_i p_i$, wenn gilt: $\Sigma p_i = 1$