Perzentilenkurven sind ein Werkzeug für den Arzt.

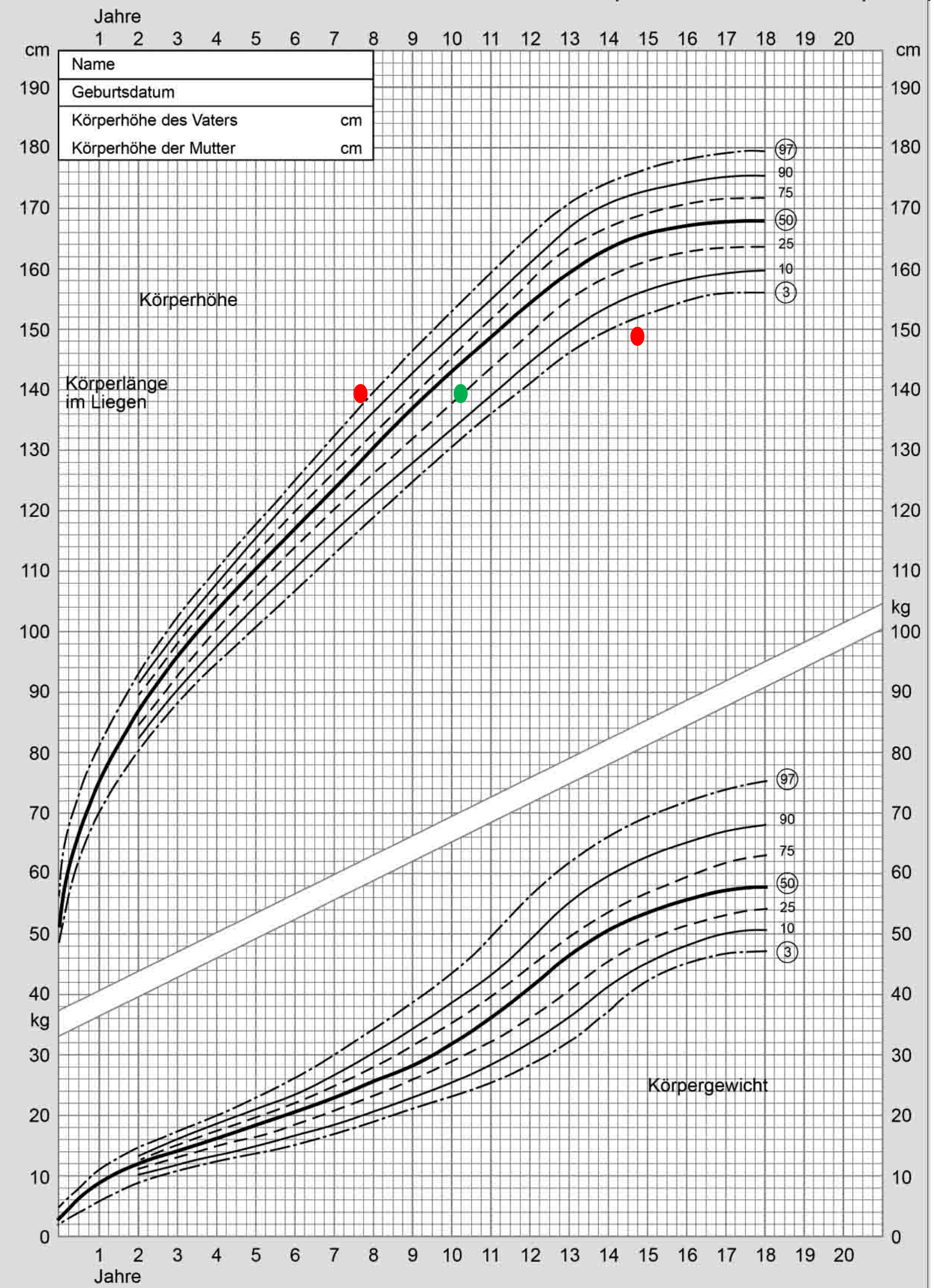
Die Fälle am Rahmen sind einfach sichtbar.

z.B.

Wachstums- und Gewichtskurven für Mädchen

=percentile(...)

=Quantil(...)



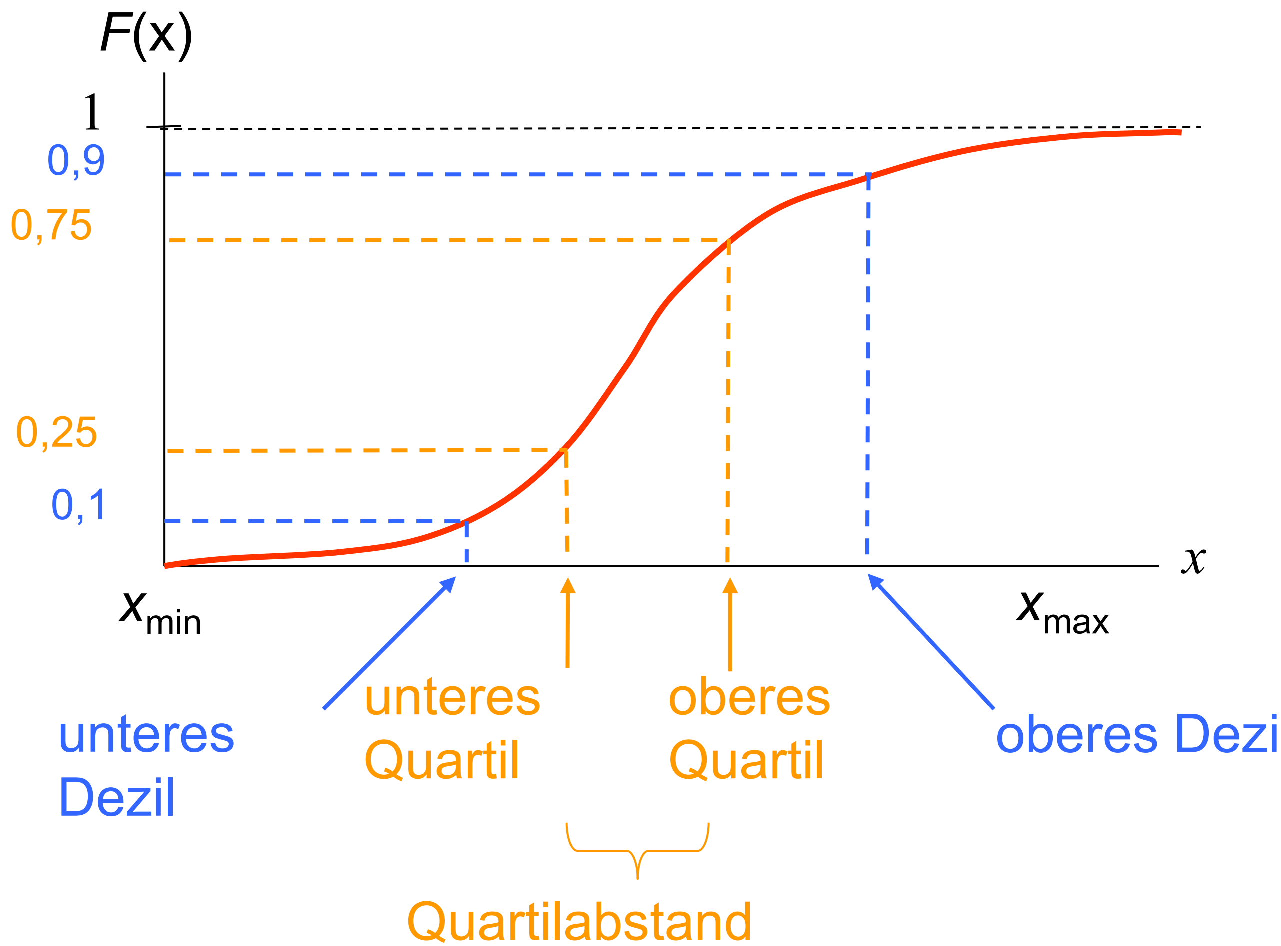
Vorsicht mit Skalentypen!

Besonders mit zahlen: die originelle Skala kann gut „nur” nominal, oder ordinal sein (z.B. Noten)

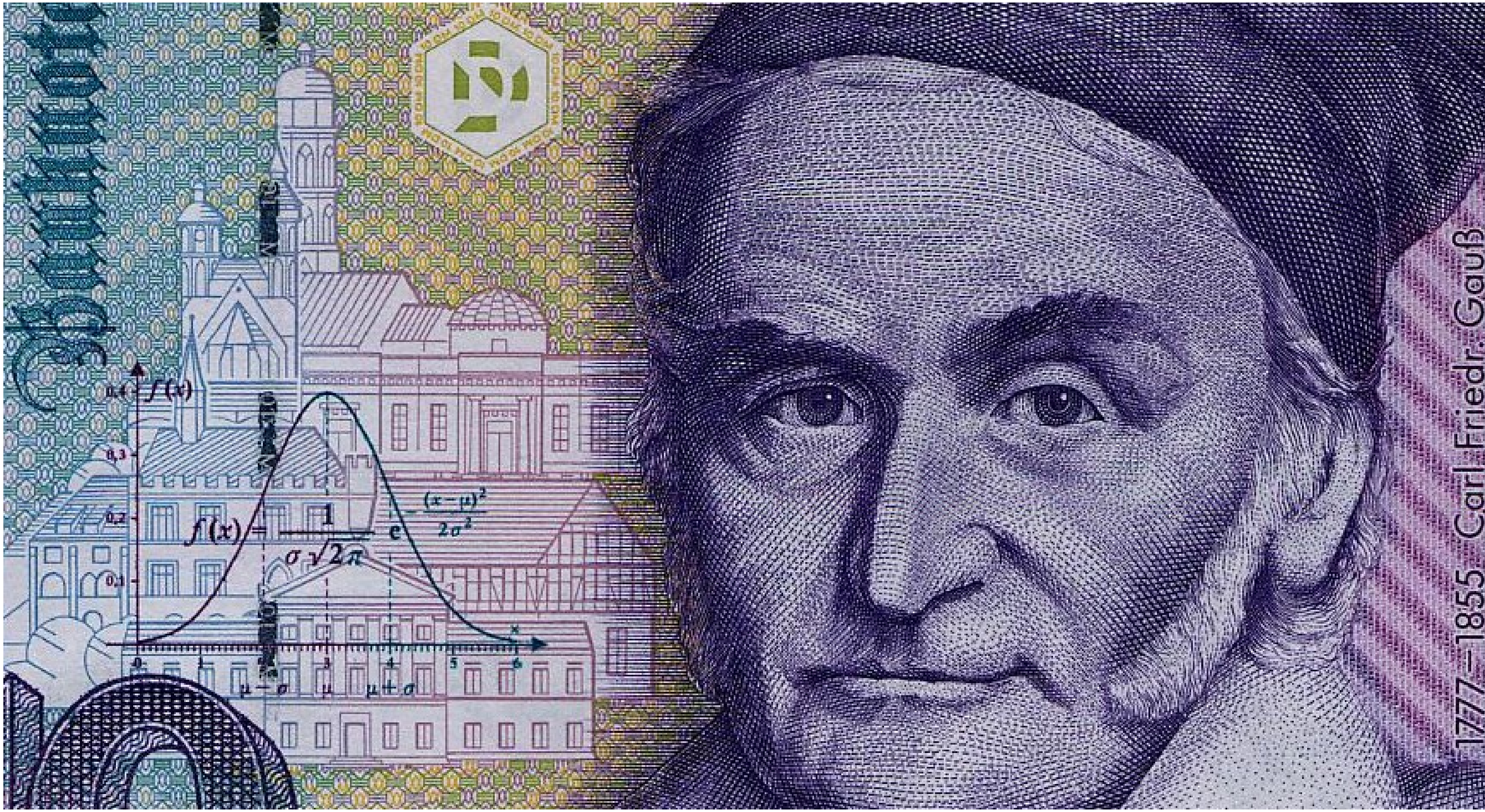
Skalentypen	zulässige Lage-Parameter	zulässige Streuungs-Parameter
Nominalskala	Modus	–
Ordinalskala	Modus, Median	–
numerische Skalen	Modus, Median, Durchschnittswert	Spannweite, Quartilabstand, Standardabweichung

Quantile und die relative Summenhäufigkeits-verteilung

F(x): wie viel % der Werte sind kleiner als x



Verteilungen und Schätzungen
Grundlagen der Wahrscheinlichkeitslehre



Wahrscheinlichkeit

Bernoulli (1654-1705), Laplace (1749-1827)
(**klassische Wahrscheinlichkeit**)

Bei einem Zufallsexperiment, was endlich viele Ausgänge hat, die (zB. wegen Symmetriegründen) **gleichwahrscheinlich** sind, die Wahrscheinlichkeit eines Ereignisses (*E*) ist:

$$\boxed{p(E) = \frac{\text{Anzahl der für } E \text{ günstigen Elementarereignisse}}{\text{Anzahl aller gleichmöglichen Elementarereignisse}}}$$

Dabei denken wir, das alle interessante Ereignisse eigentlich aus Kombinationen verschiedener Elementarereignisse aufbaubar sind, der Anzahl wovon kann auch sehr gross sein (wie im Lego-Spiel).

p=probability, Probabilität

Beispiele

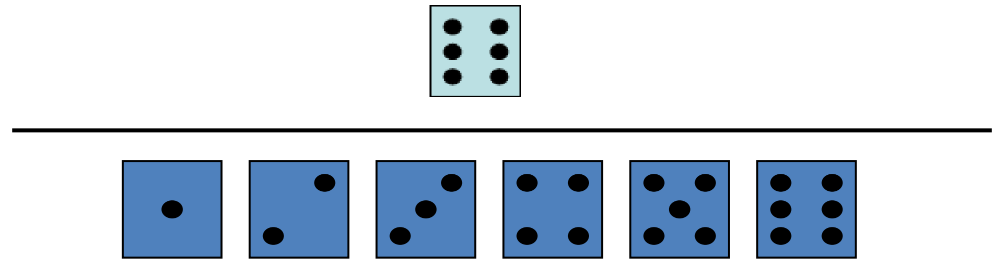
$$\boxed{p(E) = \frac{g}{m}}$$

günstig

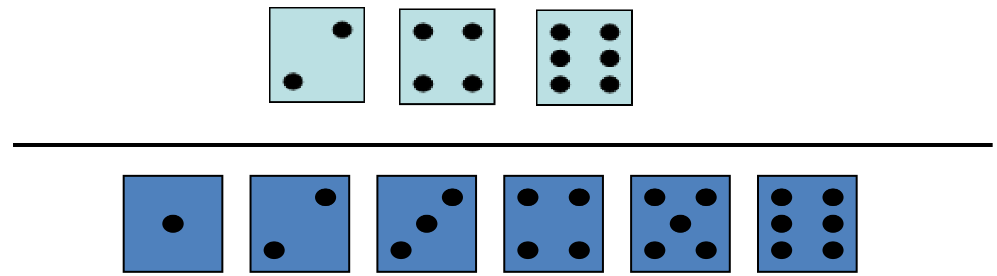
alle

Würfelexperiment:

$$p(6) = \frac{1}{6}$$



$$p(gerade\,Zahl) = \frac{3}{6}$$



Münzenexperiment:

$$p(Kopf) = \frac{1}{2}$$

Lageparameter der Verteilung

Es sei X eine diskrete Zufallsgröße mit Werten x_1, x_2... dann heisst

$$\boxed{\mu = \sum_i x_i p(x_i)}$$

Erwartungswert von *X.*

Der Erwartungswert gibt denjenigen Wert an, den man als Mittelwert (durchschnittlichen Wert) über viele Versuchswiederholungen "erwarten" kann.

Dabei ist es durchaus möglich, dass der Erwartungswert bei keinem einzigen Versuch realisiert wird oder sogar überhaupt nicht vorkommen kann.

Normalverteilung

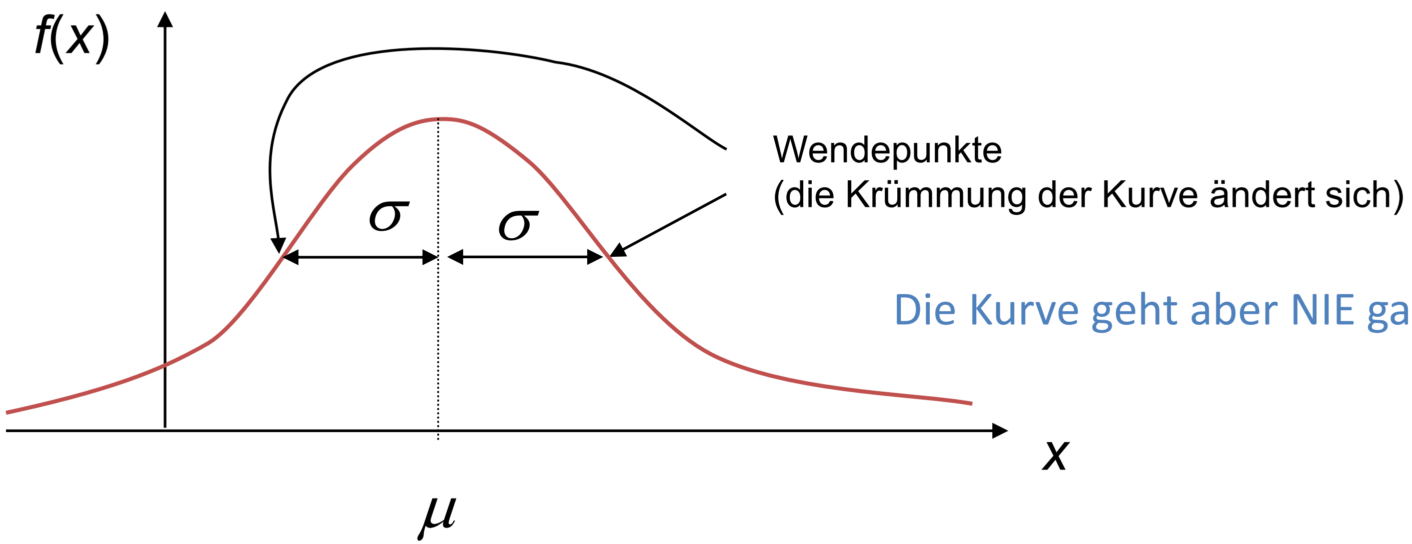
Verteilungsdichtefunktion:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Parameter der Normalverteilung:

Erwartungswert: μ
Streuung: σ

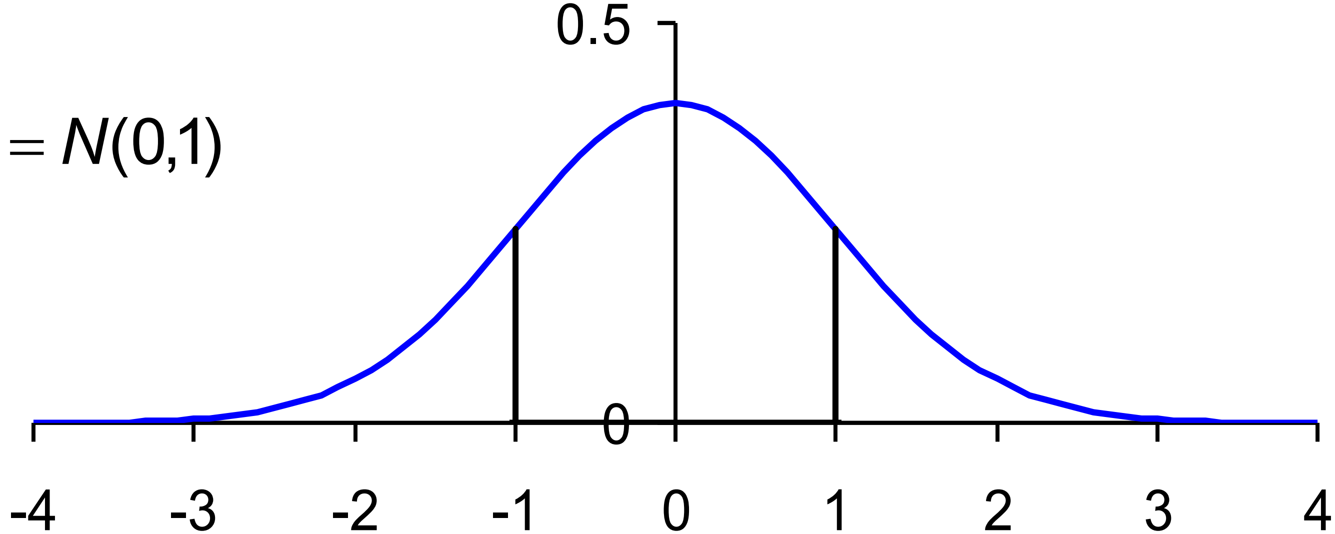
Oberfläche unter der Kurve = 1.
(gilt für alle verteilungsdichtefuntionen!)



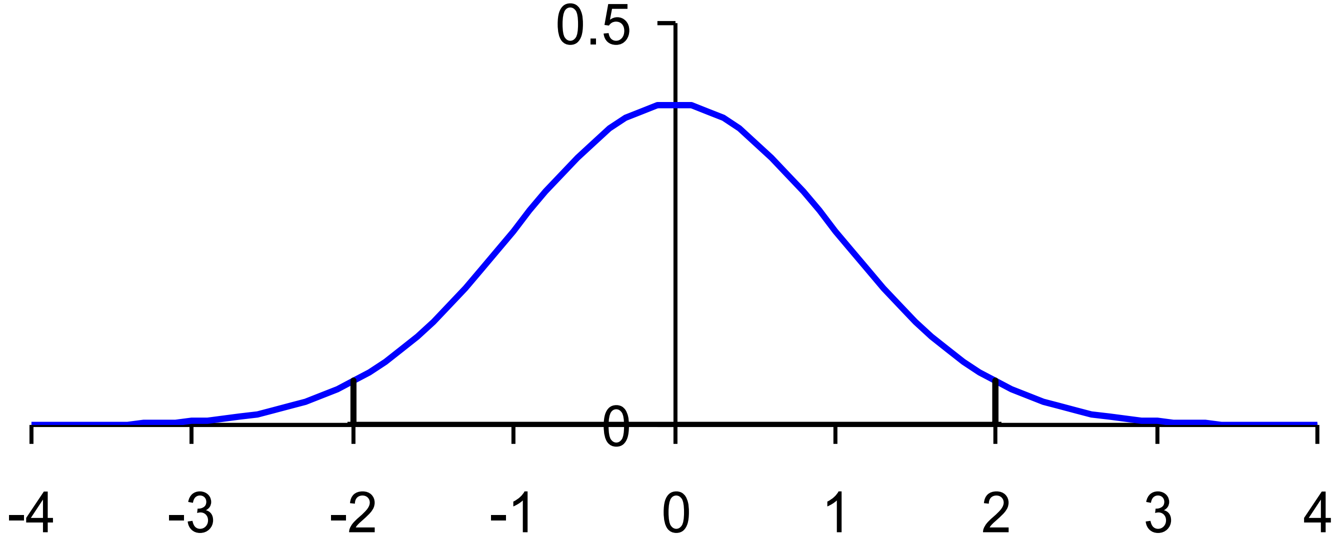
Die Kurve geht aber NIE ganz zu Null!

Standard - Normalverteilung

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} = N(0,1)$$



$$\boxed{\begin{array}{l} \mu = 0 \\ \sigma = 1 \end{array}}$$



Analytische Statistik

Population

Stichprobe

N = „unendlich”

n = endlich

Theoretische Verteilung
Erwartungswert
Theoretische Streuung

?

Häufigkeitsverteilung
Durchschnitt
Standardabweichung

Aufgabe der Schätztheorie

Aus einer Stichprobe Schätzwerte für

- Wahrscheinlichkeiten
- Erwartungswert
- Streuung
- oder andere Parametern

einer Verteilung zu ermitteln.

Typen der Schätzungen:

- ***Punktschätzung***
- ***Intervallschätzung***

Punktschätzungen

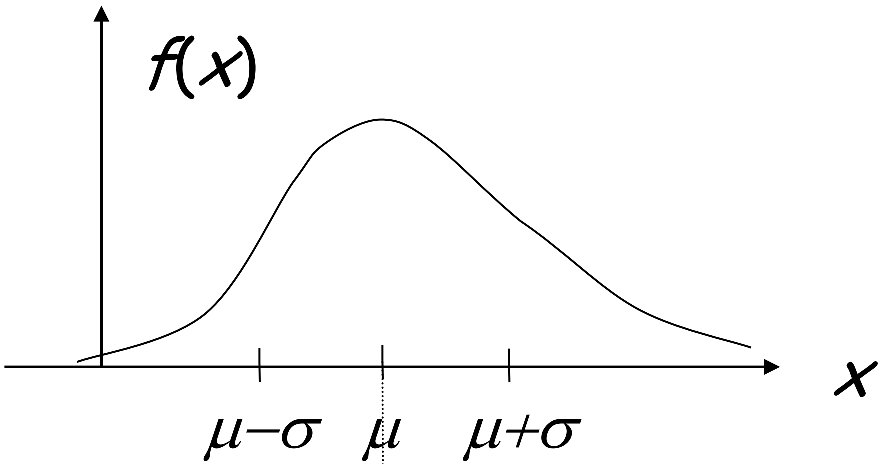
- Der unbekannte Parameter wird mit *einem Wert* geschätzt.
- Relative Häufigkeit
  ist ein Schätzwert für die Wahrscheinlichkeit
- Durchschnitt
  ist ein Schätzwert für den Erwartungswert
- Standardabweichung
  ist ein Schätzwert für die Streuung
- Punktschätzungen sagen
  nichts über die Genauigkeit bzw. Sicherheit der Schätzung!

Intervallschätzungen

- Intervallschätzung oder Konfidenzschätzung gibt zu einer vorgewählten Sicherheitswahrscheinlichkeit γ, (Konfidenzniveau) ein Intervall (c_1,c_2) an, in dem der unbekannte Parameter (zB. μ oder σ) mit einer Wahrscheinlichkeit von mindestens γ liegt.

Zb.: Erwartungswert der Pulszahl ist bei 95% Konfidenzniveau: $74 \pm 6\ ^{1}/_{\mathrm{Min}}$

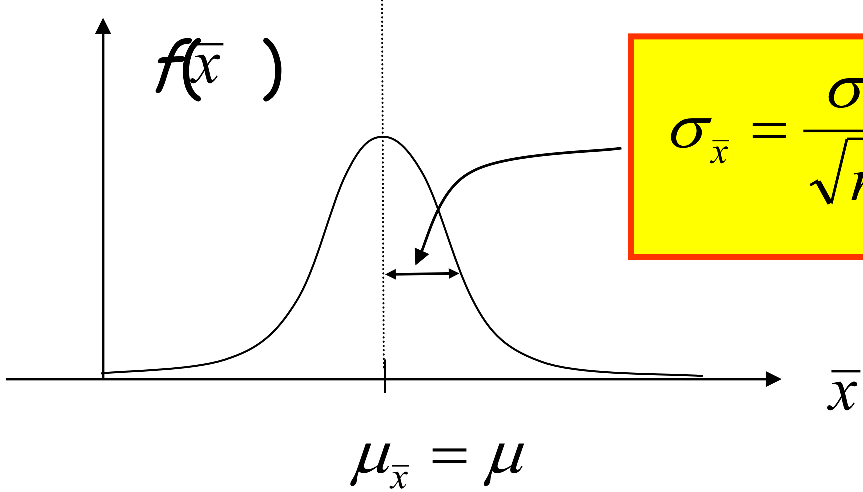
Konfidenzintervall für den Erwartungswert



zB:Körperhöhe

Standardfehler

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \approx s_{\bar{x}}$$



zB: durchschnittliche Körperhöhe in einem Studentengruppe von *n* Studenten

Zusammenfassung der Schätzungen

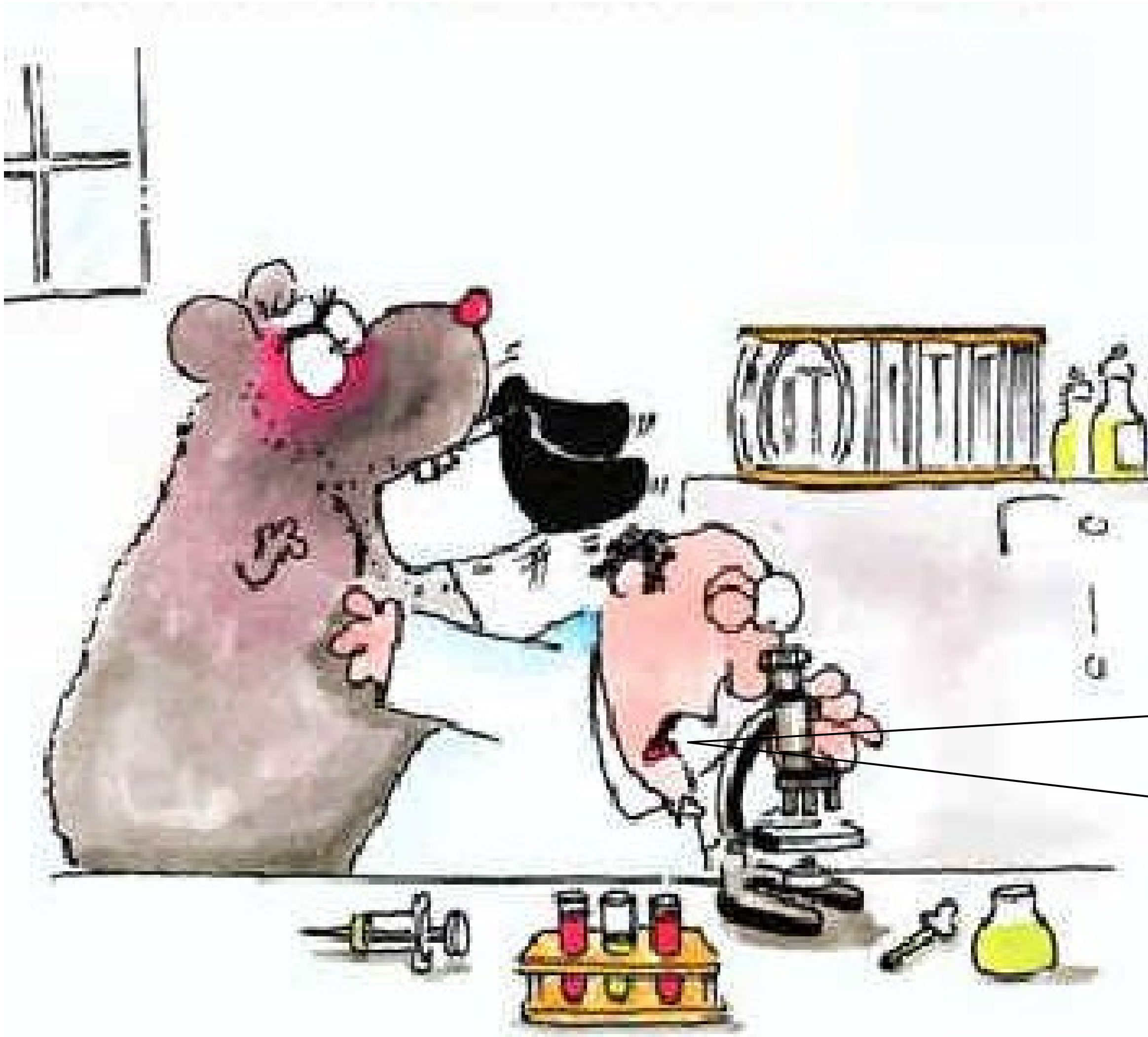
•Punktsätzungen:

Stich-probe	Grund-gesamtheit
$\bar{x} \longrightarrow$	μ
$s \longrightarrow$	σ
$n \longrightarrow$	∞

Intervallschätzung
für den Erwartunswert:

$\boxed{\bar{x} \pm 2 s_{\bar{x}} \quad 95\%}$

Hypothesenprüfungen. *t*-Tests



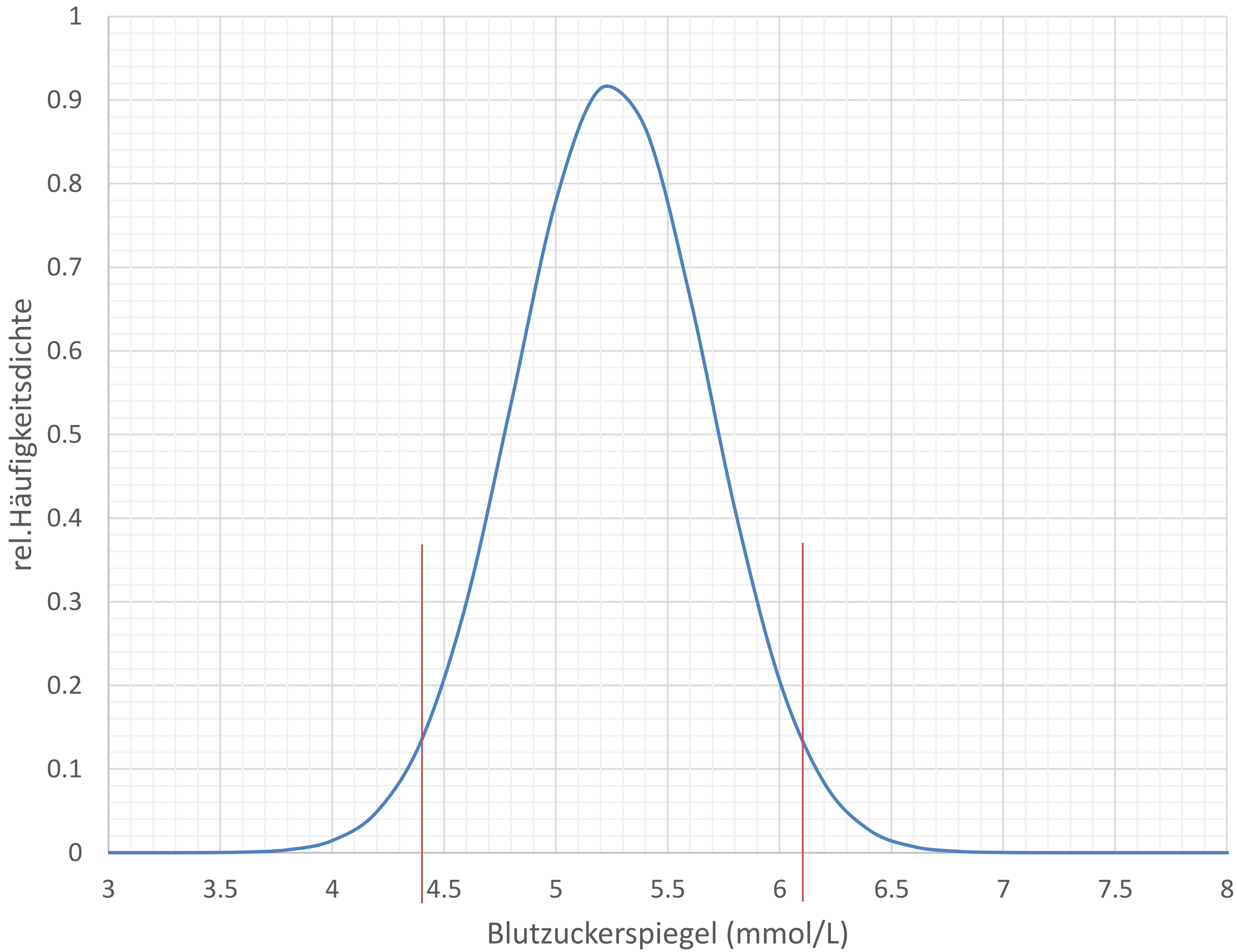
Warum?

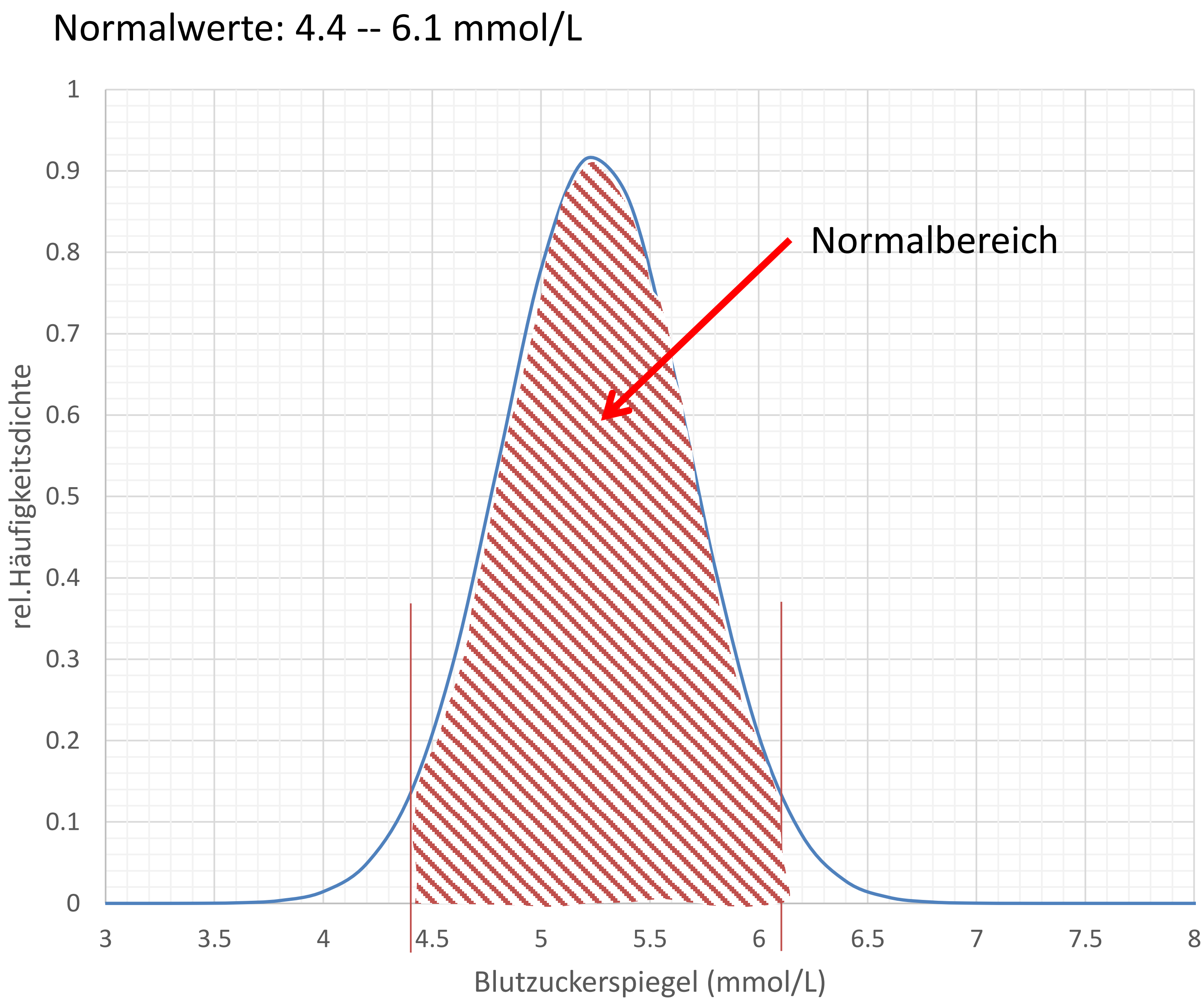
Wir wollen eine medizinisch relevante **Frage beantworten**.

z.B.
Blutzuckerwert (Glukosespiegel)

Wir messen bei jemanden den Wert 6.3, ist unser Patient jetzt krank?

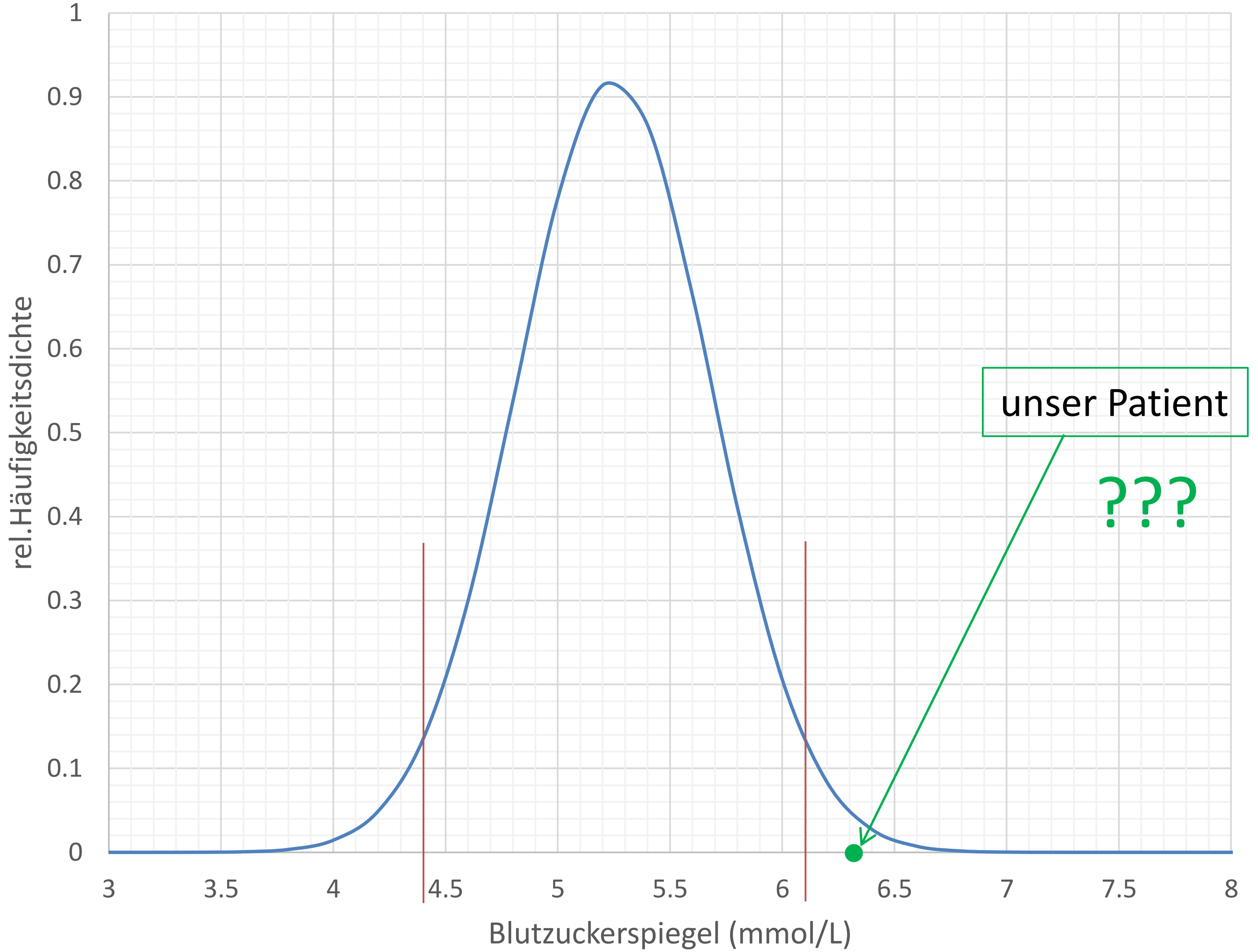
Normalwerte: 4.4 -- 6.1 mmol/L



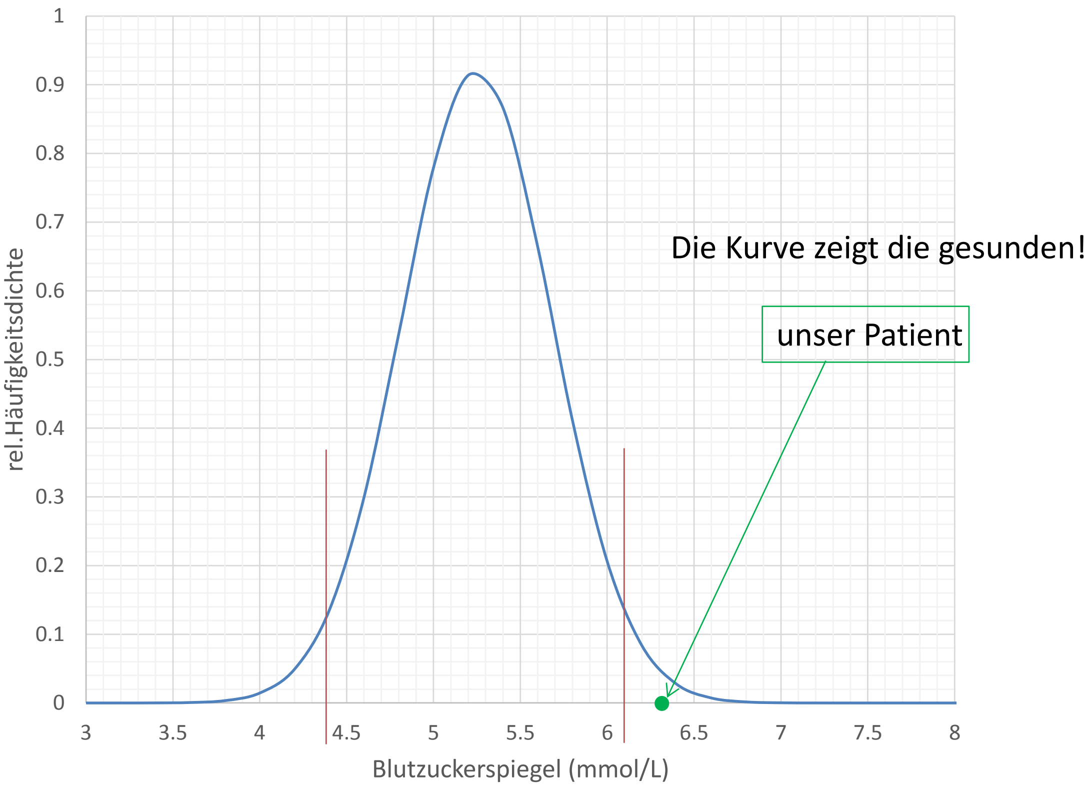


Normalwerte: 4.4 -- 6.1 mmol/L
Normalbereich
rel.Häufigkeitsdichte
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
3
3.5
4
4.5
5
5.5
6
6.5
7
7.5
8
Blutzuckerspiegel (mmol/L)

Normalwerte: 4.4 -- 6.1 mmol/L



Normalwerte: 4.4 -- 6.1 mmol/L

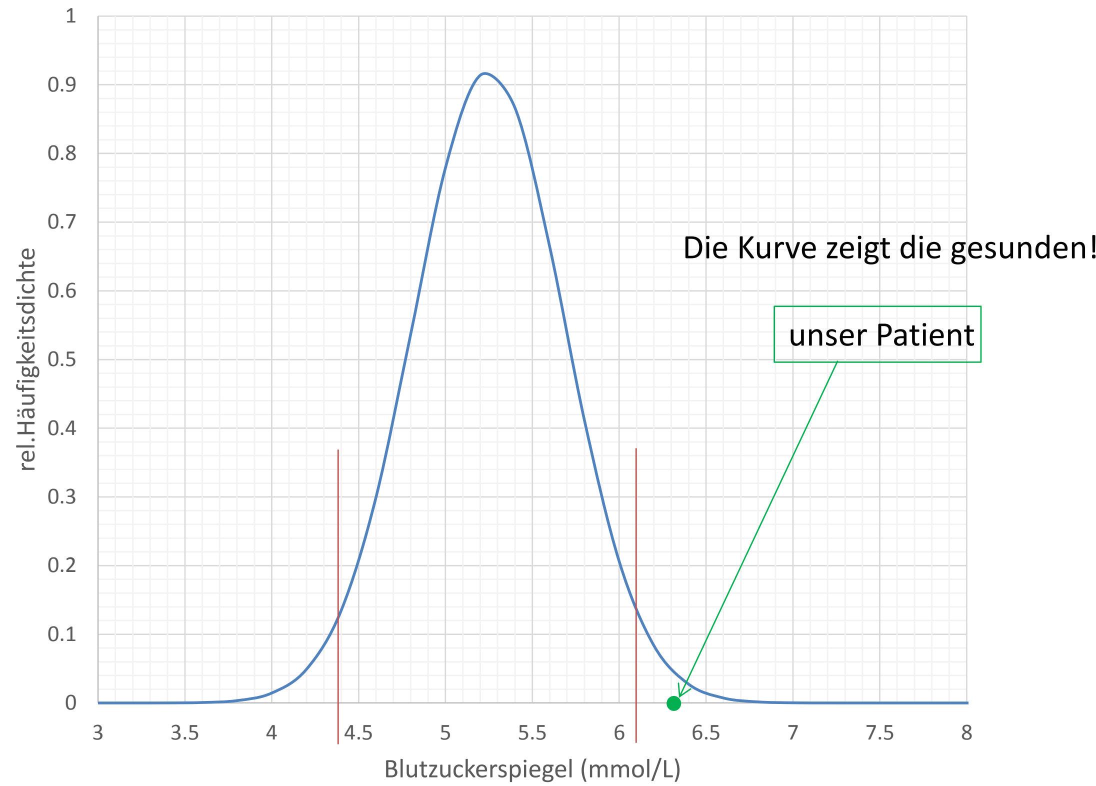


Wir wollen ein Zahl:

Eine **Wahrscheinlichkeit**

P(krank) = ?

Normalwerte: 4.4 -- 6.1 mmol/L



Wir wollen ein Zahl:

Eine **Wahrscheinlichkeit**

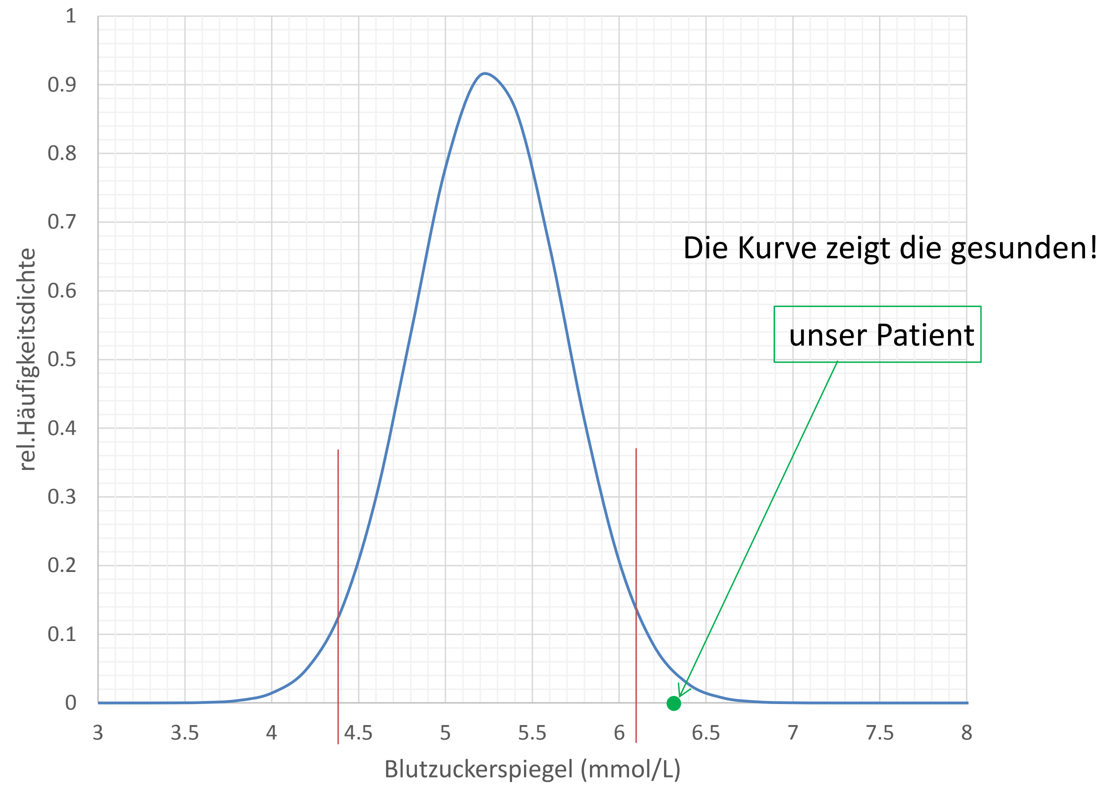
P(krank) = ?

Aber! (krank) hat keine Wahrscheinlichkeit!

ist nicht etwas wozu es die relative Häufigkeit gäbe

(ist nicht wiederholbar: entweder krank oder nicht, und nochdazu jetzt. Wir können den Zustand nicht wiederholen, ist kein Ereigniss)

Normalwerte: 4.4 -- 6.1 mmol/L



Wir wollen ein Zahl:

Eine **Wahrscheinlichkeit**

P(krank) = ?

Ansatz = „Nullhypothese”

Das hier aber ist ausrechenbar:

P(Konz=6.3mmol/L |)

Normalwerte: 4.4 -- 6.1 mmol/L

Wahrscheinlichkeit ist
Oberfläche unter der Kurve

P(c=6.3mmol/L | gesund) ~ (+)

Wir können also bedingte Wahrscheinlichkeiten ausrechnen.
(zumindest etwas dazu direkt proportional)

Wenn diese Wahrscheinlichkeit ist sehr klein, dann passen die gemessene Werte nicht zu dem Ansatz.

Wenn aber die Messwerte in Ordnung sind, dann kann nur unser Ansatz falsch sein!
Also:
Wenn P, was wir ausrechnen fällt unter einen **Schwellenwert**, dann werden wir unseren Ansatz (H0, Nullhypothese) ablehnen, sonst aber behalten.

Den Schwellenwert („Signifikanz”) legen wir am Anfang fest, abhängig davon welche mögliche Entscheidungsfehler ist ungewünschter.

Obwohl wir mit hilfe der Statistik meistens die richtige Entscheidung treffen werden, *wegen die eingebaute unsicherheit der Natur*, manchmal werden wir (erst später!) erfahren das wir in der Zeitpunkt der Entscheidung doch ein Fehler gemacht haben.

die Wahrheit (erst immer zu spät bekannt)

	Ansatz richtig	**Ansatz falsch**
Entscheidung: annehmen	Gut gemacht ☺	Typ II. Fehler β
Entscheidung: ablehnen	Typ I. Fehler α	Gut gemacht ☺

Allgemeine Schritte

1.: Frage -> Entscheidungsfrage (Ja/Nein)
H0: Nullhypothese oder Ansatz.
H0: Was wir als Daten bekommen werden,
gehört zu einer BEKANNTEn Verteilung.
2.: Festlegung des Signifikanzniveaus (Psign, Pkritisch,Signifikanz)
(Fürchten wir mehr falsche Annahme, oder ablehnen?)
3.: Experiment -> Daten   Representativ, unverzerrt!
4.: Rechnen wir ein wenig ☺
Daten -> „ξ" : Parameter des Tests (ein Zahl)
„ ξ" -> P, die Wahrscheinlichkeit.
P ~ P(Daten | H0)
5.: Entscheidung
wenn P<Pkrit , H0 ablehnen (Fehler Typ I möglich)
wenn P>=Pkrit, H0 behalten (Fehler Typ II möglich)

Einstichproben *t*-Test

Variable	x	$w = x - \mu$	$z = \dfrac{x - \mu}{\sigma}$
Verteilung	$N(\mu,\sigma)$,	$N(0,\sigma)$,	$N(0,1)$

Wenn die originale Variable x zu einer **Normalverteilung** mit Parameter μ und σ gehört,
dann gehört die transformierte Variable z zu der Standardnormalverteilung.

Wenn H_0 richtig ist, kennen wir den Wert von μ,
aber σ nicht.

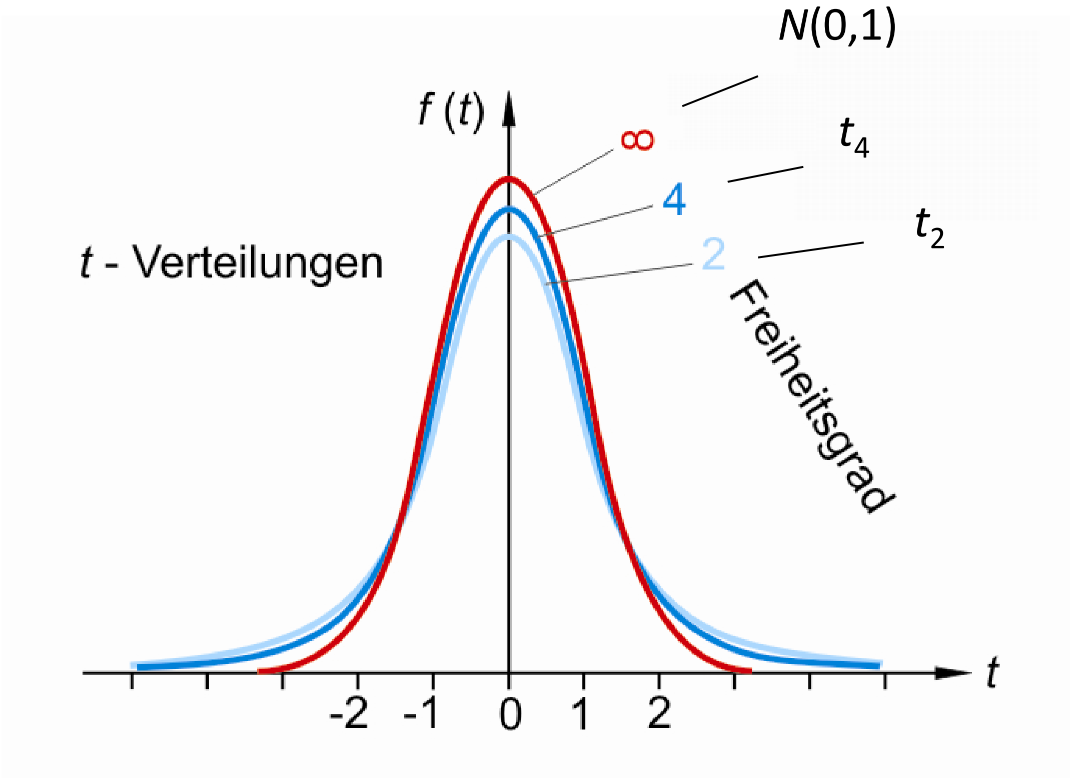
Die durchgeführte Transformation:

$$\boxed{t = \frac{\bar{x} - \mu}{s_{\bar{x}}}}$$

unserer „ ξ”-Wert, was wir aus der Daten bekommen, ist jetzt ein t-Wert.

$$t = \frac{\bar{x} - \mu}{s_{\bar{x}}}$$

t_{n-1}



Pr.Buch Abb. 21

Einstichproben *t*-Test

$$t_{\text{n-1}} = \frac{\bar{x} - \mu_0}{s_{\bar{x}}} = \frac{\bar{x} - \mu_0}{s}\sqrt{n} \quad \text{, wo} \qquad s = \sqrt{\frac{Q}{n-1}}$$

Zweistichproben *t*-Test

$$t_{n_1+n_2-2} = \frac{\bar{x}_1 - \bar{x}_2}{s^*} \cdot \sqrt{\frac{n_1 n_2}{n_1 + n_2}} \quad \text{, wo} \qquad s^* = \sqrt{\frac{Q_1 + Q_2}{n_1 + n_2 - 2}}$$

Vergleichen wir die Formeln!

$$t_{\text{n-1}} = \frac{\bar{x}}{s\sqrt{\frac{1}{n}}} \qquad\qquad t_{n_1-1+n_2-1} = \frac{\bar{x}_1 - \bar{x}_2}{s^*\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

Einstichproben

Zweistichproben

Mit Excel ist alles viel einfacher...

2-Seitiger Test.

T.TEST (daten1, daten2, **2**, Typ)

	n1	n2	dn		
m	36	37	1	1.: gibt es ein Effekt?	
				H0: n1 und n2 sind identisch	
m	33	35	2	Sei Pkrit 3%	
m	34	37	3	**P=T.TEST(n1,n2,2,1)**	**0.00027**
m	33	37	4	P<3% also H0 wird abgelehnt.	
m	34	38	4	(Fehler Typ I max P gross)	
m	37	41	4		
m	37	46	9	2.: sind m und w unterschiedlich?	
w	30	29	-1	H0: ja, die sind indentisch	
w	45	49	4	Sei Pkrit 3%	
w	50	55	5	**P=T.TEST(dn w,dn m,2,2)**	**0.23118**
w	35	40	5	p>3% also H0 wird behalten	
w	28	37	9	(Fehler Typ II =?)	
w	39	48	9		
w	49	63	14		

Typ : 1: gepaart
2: 2-Stichproben gleicher Var.
3: 2-Stichproben ungleicher Var.

Übersicht der Testmethode

Verteilung / *Stichproben*	*normalverteilte Daten*	*die Verteilung der Daten ist unbekannt*
eine Stichprobe	Einstichproben t-Test	Wilcoxon Test
zwei Stichproben	Zweistichproben t-Test	Mann-Whittney U-Test
mehrere Stichproben	ANOVA (Varianzanalyse)	Kruskal-Wallis Test

Wie gut passen die Messpunkte an die Regressionsgerade?

Korrelationsrechnung beschreibt die lineare Beziehung zwischen zwei oder mehr statistischen Variablen

es beschreibt die Stärke der Korrelation
es gibt starke und schwache Korrelation

Korrelationskoeffizient (Pearson)

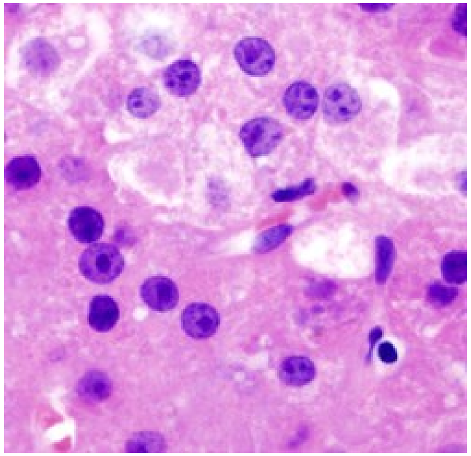
$$\boxed{r = \frac{Q_{xy}}{\sqrt{Q_{xx} \cdot Q_{yy}}} = \frac{s_{xy}^2}{s_x s_y}}$$

der Zähler ist gleich dem Zähler der Steigung der Regressionsgerade (der Nenner ist im beiden Fall positiv)

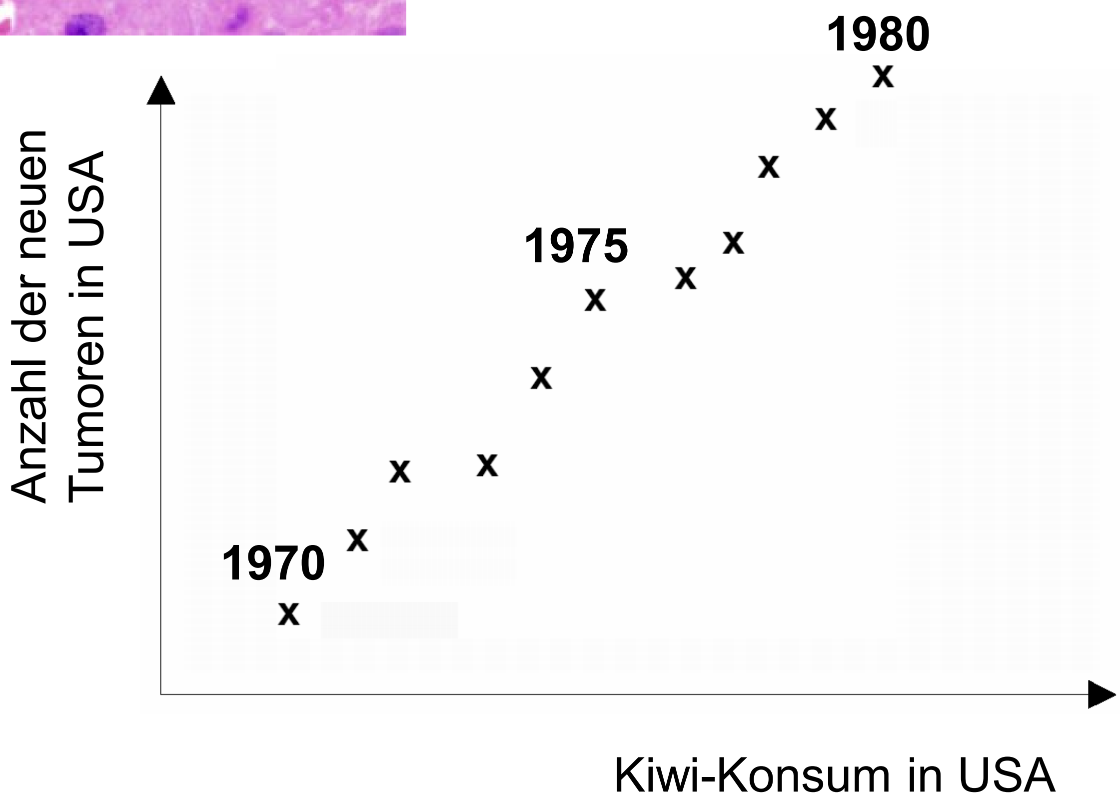
$$a^* = \frac{Q_{xy}}{Q_{xx}} \Rightarrow$$

positive Steigung: *r* ist positive Zahl
negative Steigung: *r* ist negative Zahl

$$-1 \leq r \leq 1$$



Korreliert heisst nicht notwendigerweise kausal verknüpft(!)

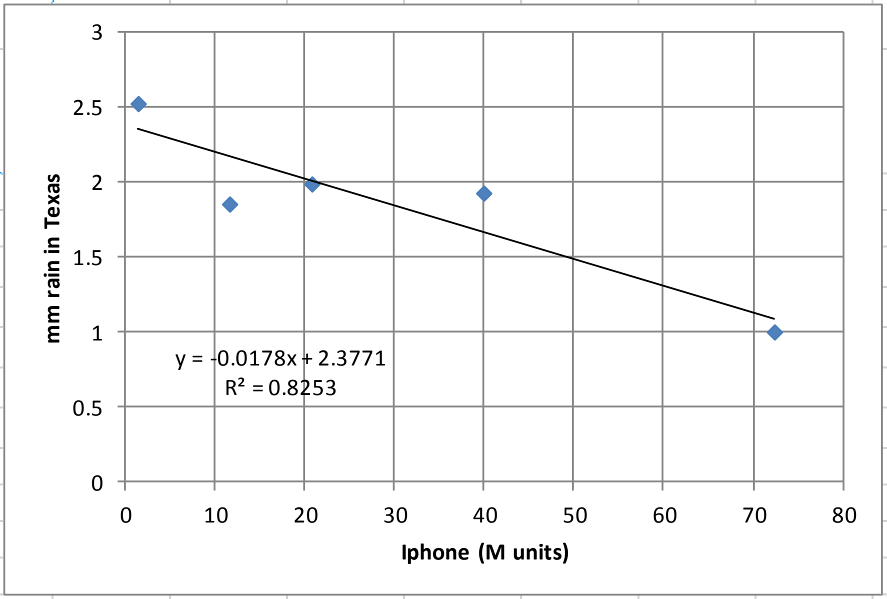


Korrelation heisst noch lange nicht Uhrsache!!!

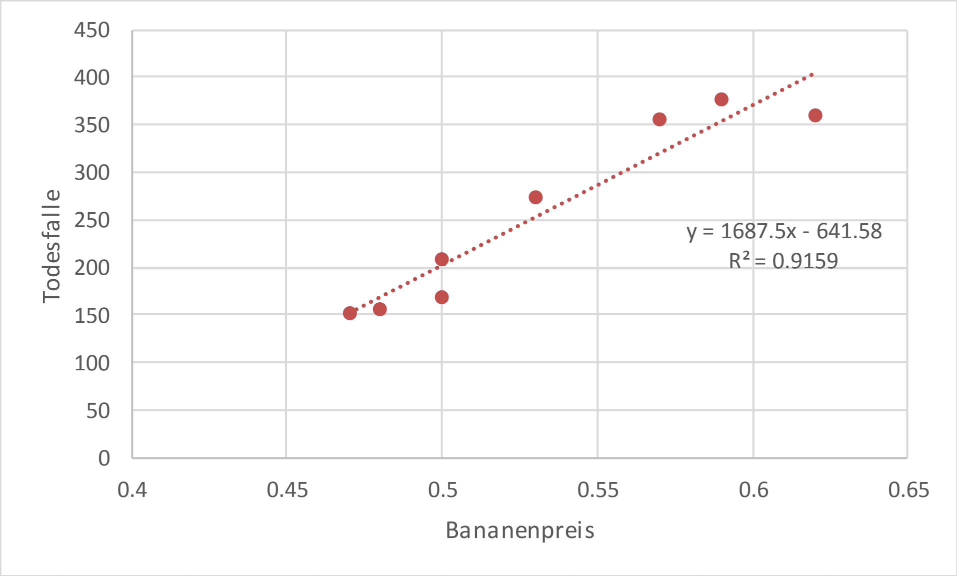
	2007	2008	2009	2010	2011
Apple iPhone sales Millions of units ()	1.39	11.63	20.73	39.99	72.29
Precipitation in Texas *Avg Daily Precipitation (mm) (CDC)*	2.52	1.85	1.99	1.93	1

1.39	2.52
11.63	1.85
20.73	1.99
39.99	1.93
72.29	1

R	-0.90846
n	5
t	-3.76471
P	0.032784
	p<0.05



	1999	2000	2001	2002	2003	2004	2005	2006
Cost of bananas (unadjusted) *Dollars per pound (Bureau of Labor)*	0.5	0.47	0.48	0.5	0.53	0.62	0.57	0.59
People who died by falling out of their wheelchair *Deaths (US) (CDC)*	169	154	157	209	274	360	356	377



http://tylervigen.com/discover