Statistik

Literatur:

Glantz, S.A. (2002). *Primer of Biostatistics.* New York: McGraw-Hill.

Maxwell, S.E. & Delaney, H.D. (2000). *Designing Experiments and Analyzing Data.* Mahwah, NJ: Erlbaum.

Das Grundproblem der Statistik: der Zufall.

• Jede empirische Messung kann nur mit einer bestimmten Präzision erfolgen.

• Im allgemeinen treten von Messung zu Messung mehr oder weniger starke Schwankungen auf, die durch „Zufall“ zustande kommen.

• Die entscheidende Frage:
Spiegelt ein interessantes Datenmuster tatsächlich die Wirklichkeit wieder, oder könnte es durch reinen Zufall zustande gekommen sein?

• Die Statistik dient dazu, Wahrscheinlichkeiten dafür abzuschätzen, daß ein gemessener Effekt durch Zufallsschwankungen zustande gekommen sein könnte.

Wozu braucht man Statistik?

- Entscheiden, wie extrem eine Beobachtung innerhalb einer normierten Population ist.

- Zeigen, daß die beobachteten Effekte stärker sind als die Zufallseinflüsse in der Studie.

- Qualitätssicherung: statistisch signifikante Ergebnisse zeigen, daß die Daten mit hinreichender Präzision gemessen wurden.

- „Statistik ist ein organisiertes Argument“ (Maxwell & Delaney, 2000): Die statistische Analyse ist genauso organisiert wie die logische Argumentation.

Stichprobe und Grundgesamtheit

Grundgesamtheit:

Die Menge aller *möglichen* Beobachtungen einer festgelegten Kategorie

die IQ-Ergebnisse aller amerikanischen Busfahrer
die BSE-Werte aller Rinder in Deutschland

Stichprobe:

eine **zufällig** und **unabhängig** gezogene Teilmenge davon

Grundlegende statistische Maße („Statistiken“)

Mittelwert (*mean*):
Der Durchschnitt aller Werte

Varianz (*variance*):
Die durchschnittliche quadrierte Abweichung aller Werte vom Mittelwert

Standardabweichung (*standard deviation, SD*, Streuung):
Die Quadratwurzel aus der Varianz

Freiheitsgrade

Die Freiheitsgrade (*df*, degrees of freedom) sind die Zahl der Elemente einer Stichprobe, die frei variieren können.

In einer (theoretischen) Population entspricht die Zahl der Freiheitsgrade einfach der Zahl der Elemente: $df = n$.

Wenn man weiß, daß die Elemente bestimmte Randbedingungen erfüllen müssen, gehen Freiheitsgrade verloren: ein Freiheitsgrad pro einzuhaltende Randbedingung.

> Beispiel: In der Formel für die empirische Varianz steht $(n-1)$ statt n, weil die Varianz ja relativ zum Mittelwert berechnet werden muß – einer der Meßwerte ist also durch die anderen genau festgelegt, wenn alle zusammen den Mittelwert ergeben sollen.

Viele Wahrscheinlichkeitsverteilungen hängen von der Zahl der Freiheitsgrade ab!

Deskriptive Statistik

Busfahrer:	*IQ*
1	119
2	100
3	90
4	110
5	108
6	135

Mittelwert:	**110.33**
Standardabweichung:	**15.55**
Varianz:	**241.87**
Standardmeßfehler:	**6.35**

Der Standardmeßfehler

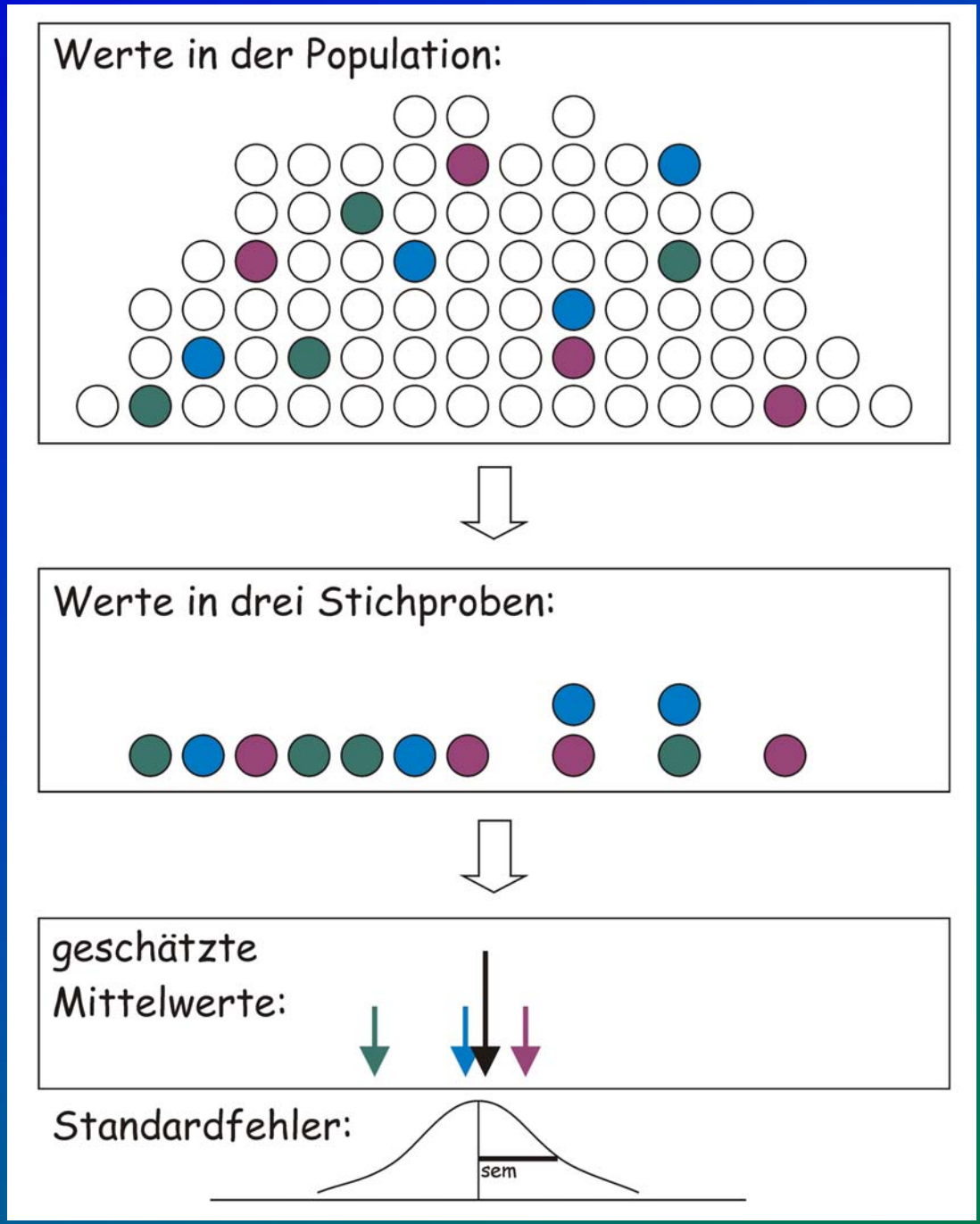
• Eine Stichprobe kann die Population nur unvollkommen widerspiegeln.

• alle geschätzten Statistiken (Mittelwert, Varianz usw.) unterscheiden sich etwas von einer Stichprobe zur anderen.

• Die Standardabweichung dieser leicht verschiedenen Werte nennt man **Standardmeßfehler.**

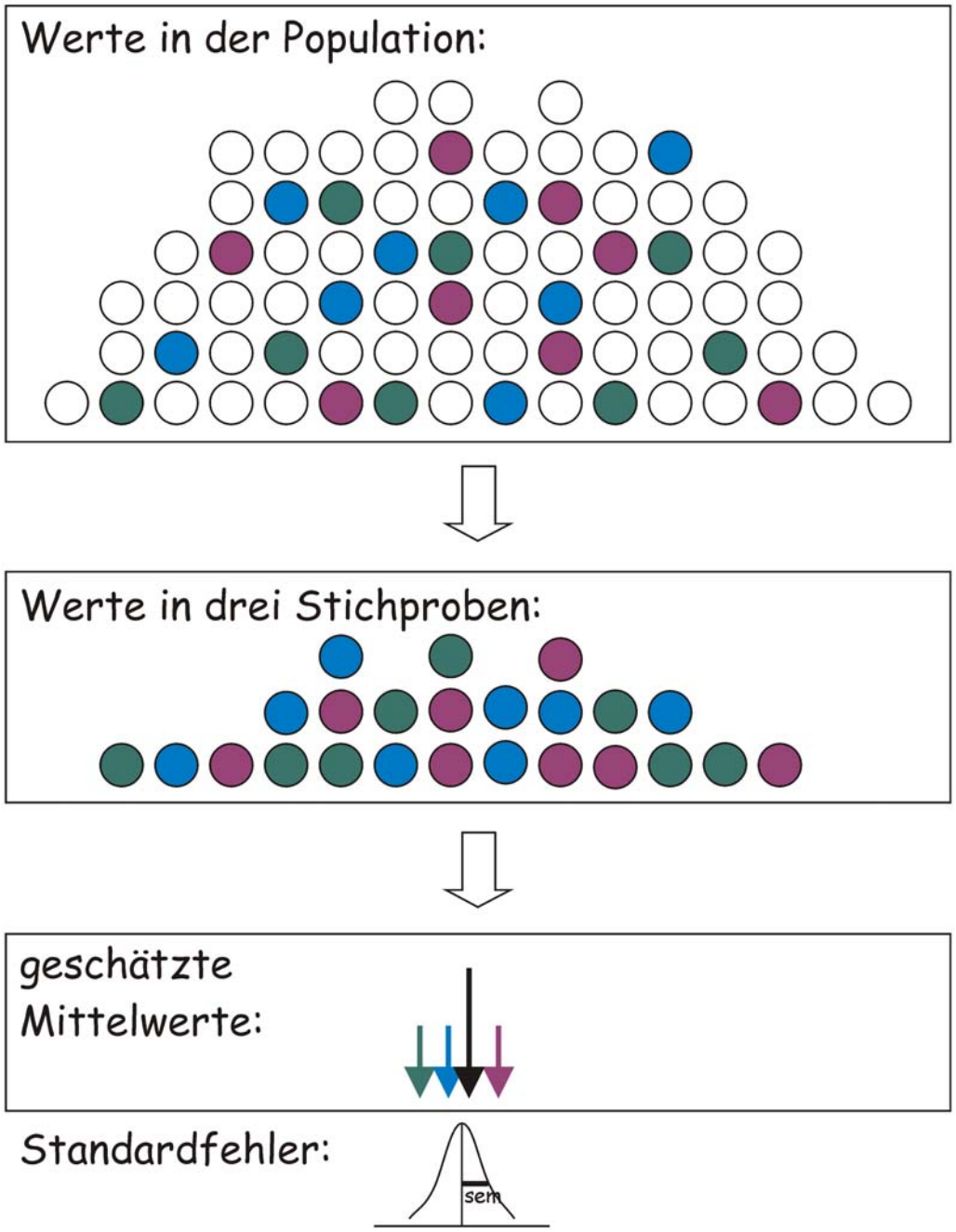
• Je größer die Stichprobe ist, desto genauer werden die Statistiken geschätzt, und umso kleiner ist der Standardmeßfehler.

• z.B. **Standardfehler des Mittelwertes:**
Standardabweichung geteilt durch Wurzel N.



Kleine Stichproben: geschätzte Mittelwerte können durch Zufallseinflüsse weit vom „wahren“ Mittelwert entfernt liegen

⇒ die theoretische Verteilung der geschätzten Mittelwerte ist breit, d.h. der Standardmeß-fehler ist groß.



Größere Stichproben: Zufallseinflüsse spielen eine geringere Rolle, die Mittelwerte können genauer geschätzt werden

⇒ die theoretische Verteilung der geschätzten Mittelwerte ist enger, d.h. der Standardmeß-fehler ist klein.

Konfidenzintervalle

Mit Hilfe des Standardmeßfehlers kann man ein Konfidenzintervall erzeugen:

z.B. Schätzer des Mittelwertes:

mit einer Wahrscheinlichkeit von 0.95 liegt der Populationsmittelwert innerhalb eines Bereichs von ca. 2 Standardfehlern um den Stichproben-mittelwert.

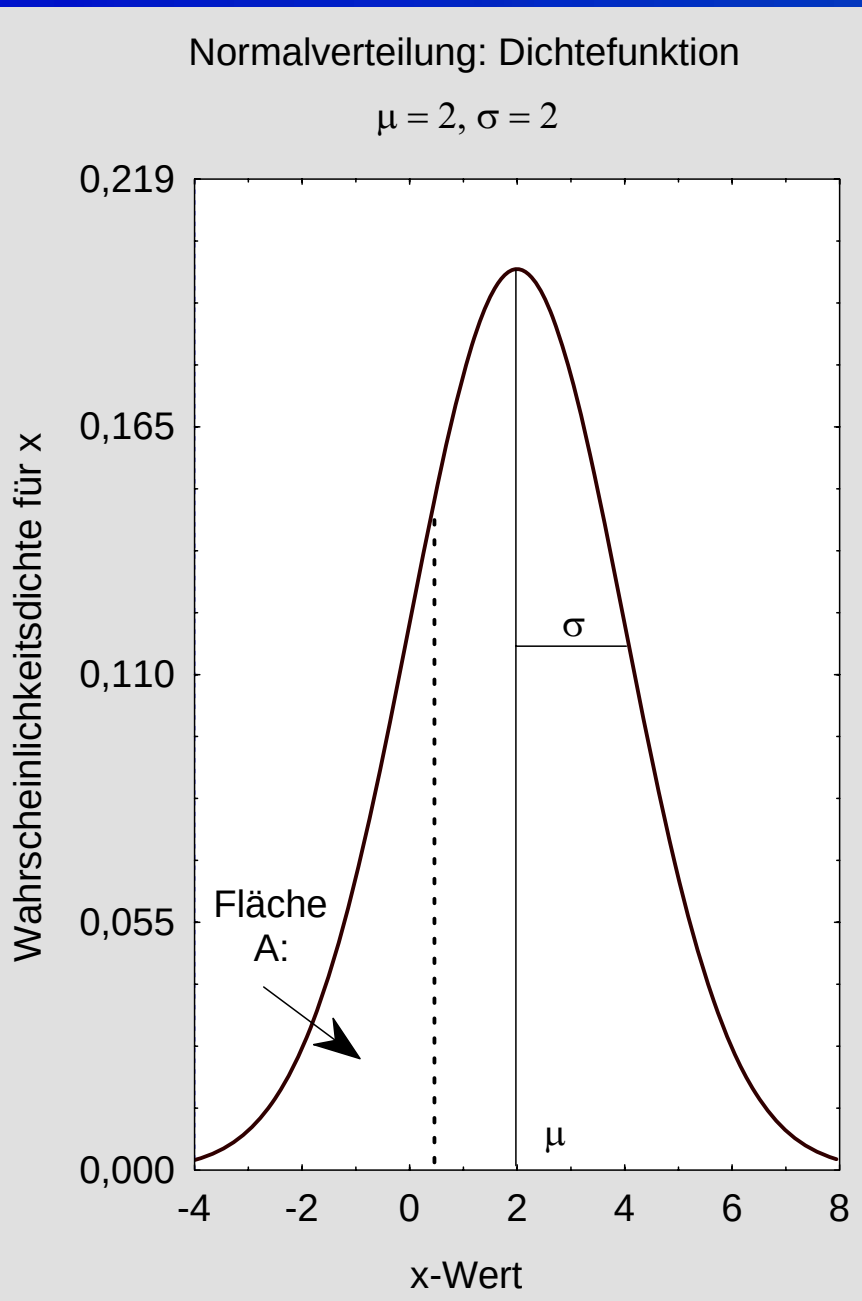
Busfahrer:	*IQ*
1	119
2	100
3	90
4	110
5	108
6	135

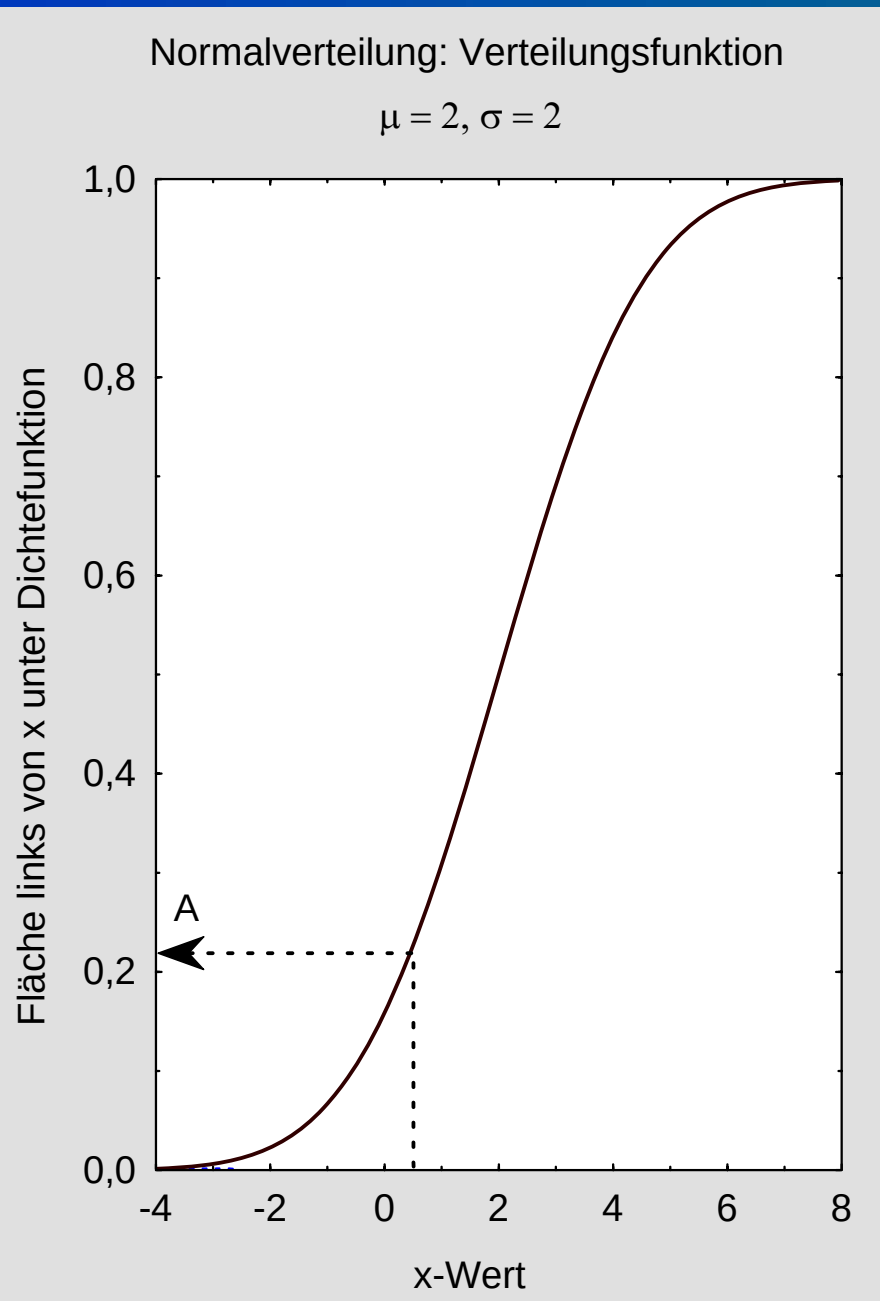
Mittelwert:	**110.33**
Standardmeßfehler:	**6.35**
95%-Konfidenzintervall:	**97.63 - 123.03**

Zufallsvariablen und Wahrscheinlichkeitsverteilungen

Eine Zufallsvariable ist eine Größe, die mit bestimmten Wahrscheinlichkeiten bestimmte Werte annehmen kann.

Eine Wahrscheinlichkeitsverteilung ist eine mathematische Funktion, die angibt, welcher Wert mit welcher Wahrscheinlichkeit vorkommen kann.

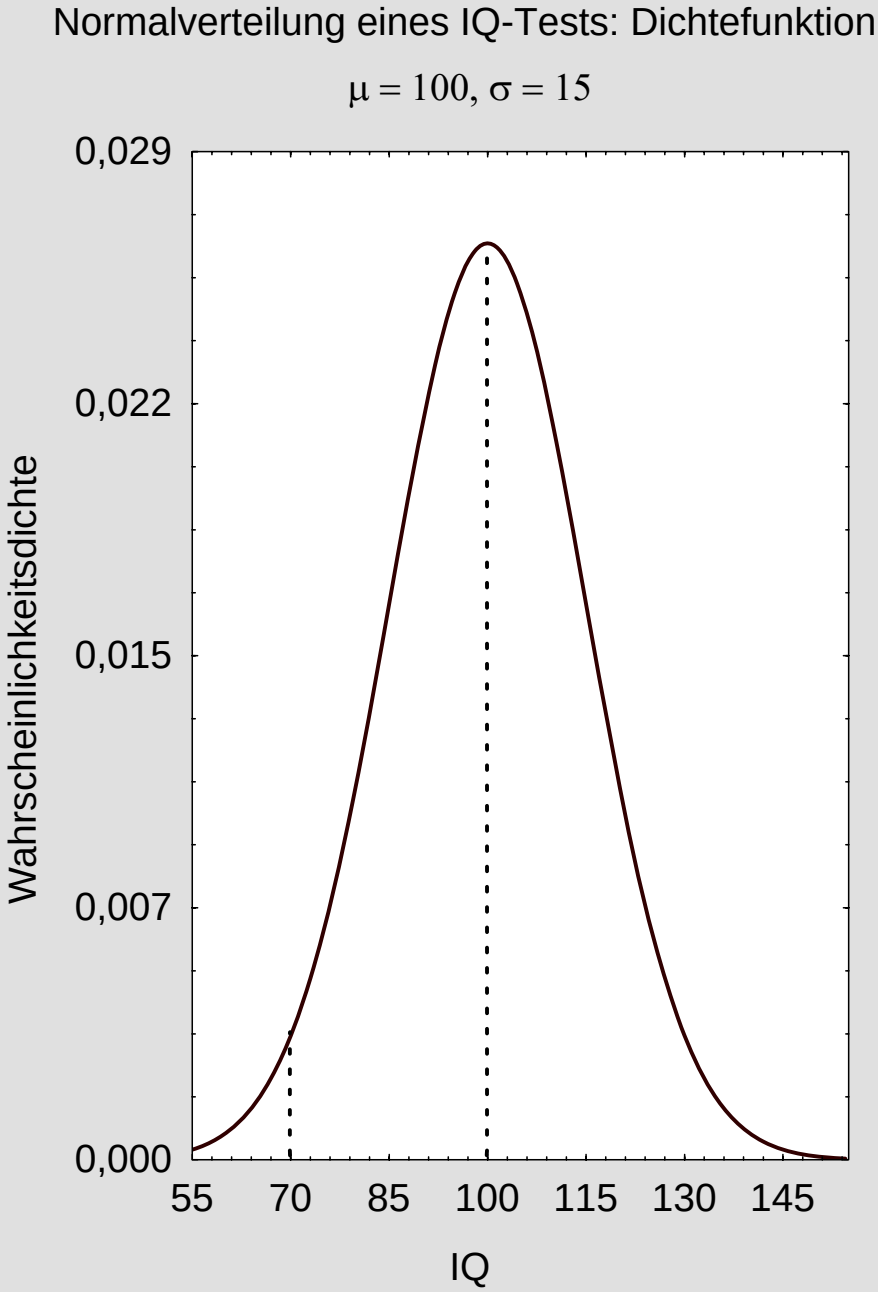

Normalverteilung: Dichtefunktion
μ = 2, σ = 2
Wahrscheinlichkeitsdichte für x
0,219
0,165
0,110
0,055
0,000
Fläche
A:
σ
μ
-4
-2
0
2
4
6
8
x-Wert

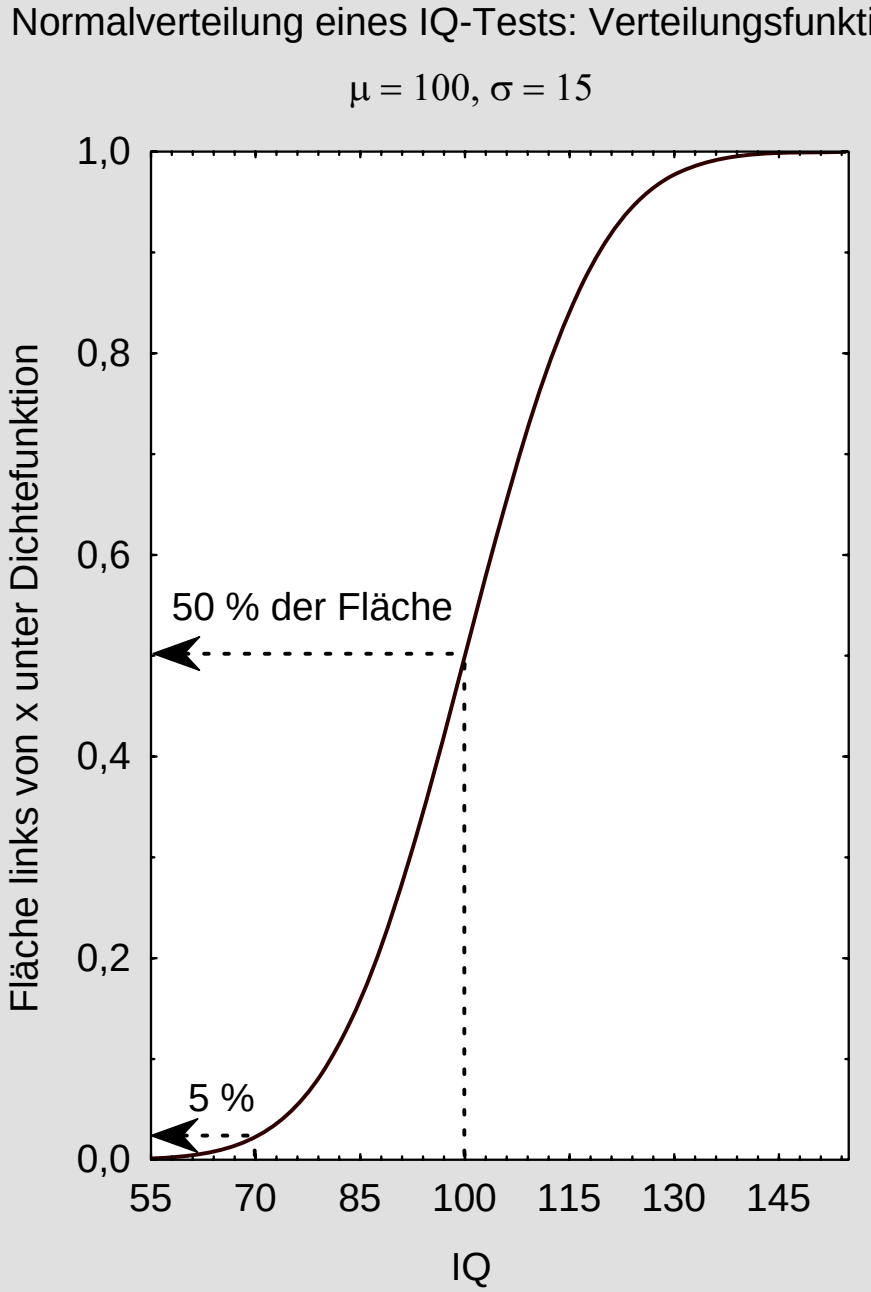

Normalverteilung: Verteilungsfunktion
μ = 2, σ = 2
Fläche links von x unter Dichtefunktion
1,0
0,8
0,6
0,4
0,2
0,0
A
-4
-2
0
2
4
6
8
x-Wert

Beispiel: Intelligenztest

Daten:
IQs von einer sehr großen Stichprobe von zufällig ausgewählten Personen

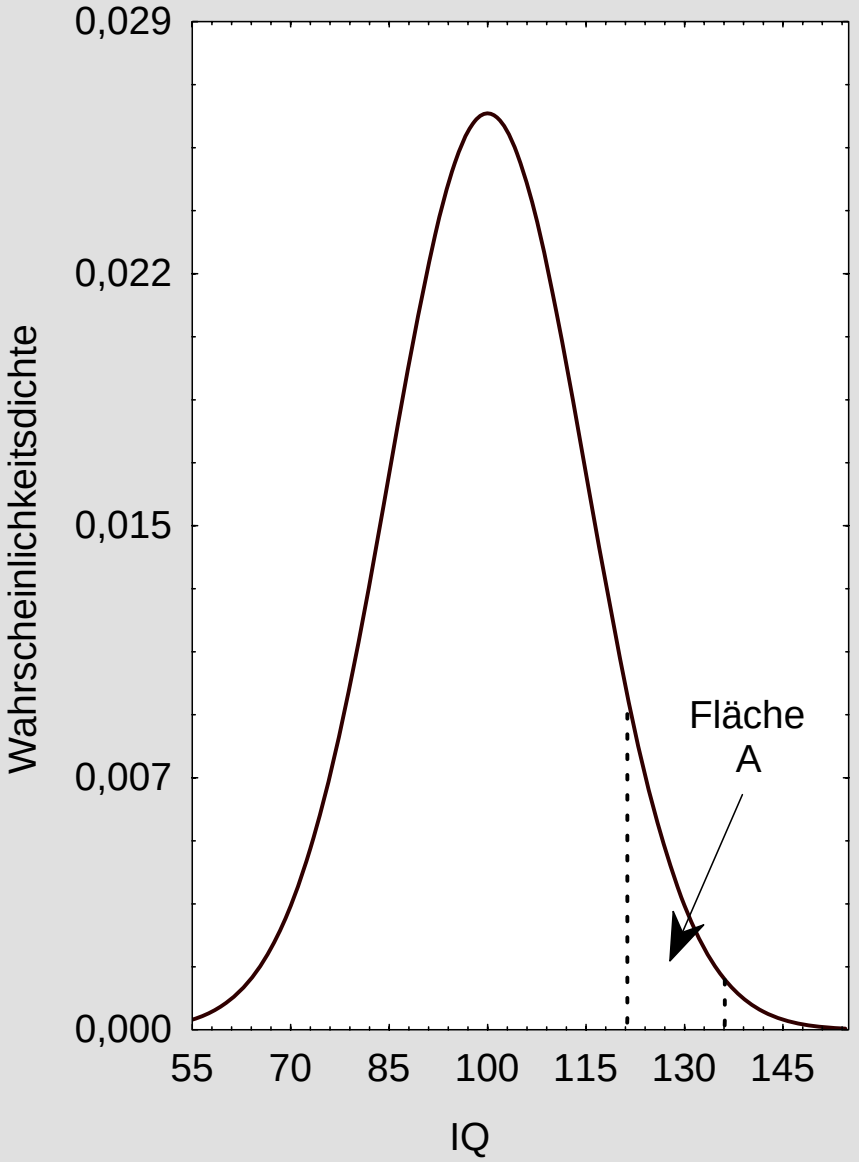
Annahme:
Das Merkmal "Intelligenz" ist in der Population normalverteilt


Normalverteilung eines IQ-Tests: Dichtefunktion
μ = 100, σ = 15
Wahrscheinlichkeitsdichte
0,029
0,022
0,015
0,007
0,000
55
70
85
100
115
130
145
IQ


Normalverteilung eines IQ-Tests: Verteilungsfunktion
μ = 100, σ = 15
Fläche links von x unter Dichtefunktion
1,0
0,8
0,6
0,4
0,2
0,0
50 % der Fläche
5 %
55
70
85
100
115
130
145
IQ

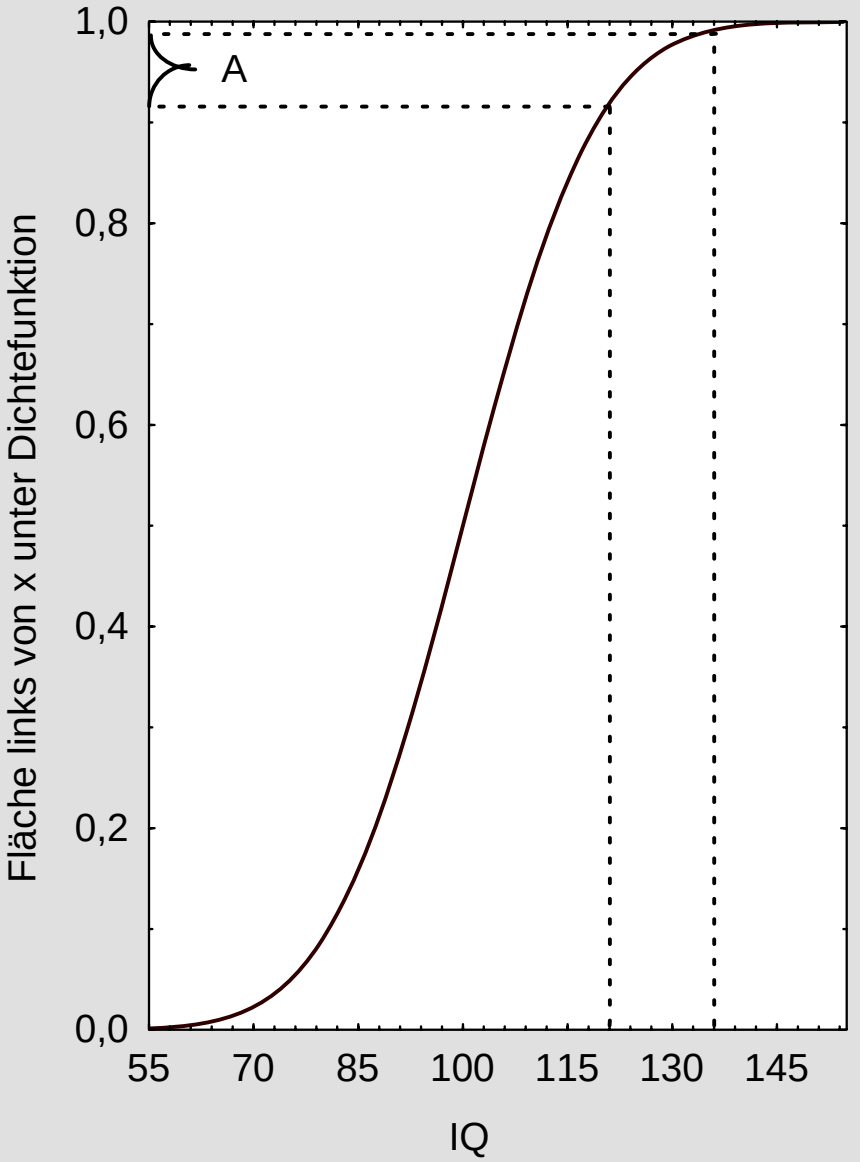
Normalverteilung eines IQ-Tests: Dichtefunktion

$\mu = 100, \sigma = 15$



Normalverteilung eines IQ-Tests: Verteilungsfunktion

$\mu = 100, \sigma = 15$



Wenn man weiß, wie eine Statistik verteilt ist, kann man entscheiden, ob ein bestimmter Wert "extrem" ist oder nicht:

- ein IQ von 130 ist hoch, weil nur wenige Personen einen höheren IQ haben

- ein IQ von 70 weist dagegen auf ein Intelligenzdefizit hin

Testen von Hypothesen

(Achtung: das Folgende gilt zunächst nur für sog. *ungerichtete Tests*!)

In jedem statistischen Test gibt es eine Nullhypothese H_0:

- „Es gibt keinen Unterschied zwischen Experimental- und Kontrollgruppe.“
- „Die Korrelation zwischen zwei Variablen ist 0.“
- „Die Regressionsgerade hat eine Steigung von *genau* 0.5.“

H_0 ist immer eine **Punkthypothese!**

Die Alternativhypothese H_1 ist das logische Komplement der H_0:

- „Es gibt einen Unterschied zwischen Experimental- und Kontrollgruppe.“
- „Die Korrelation zwischen zwei Variablen ist nicht 0.“
- „Die Regressionsgerade hat eine andere Steigung als 0.5.“

H_1 ist immer eine **Bereichshypothese!**

Sind die Daten mit der Nullhypothese vereinbar?

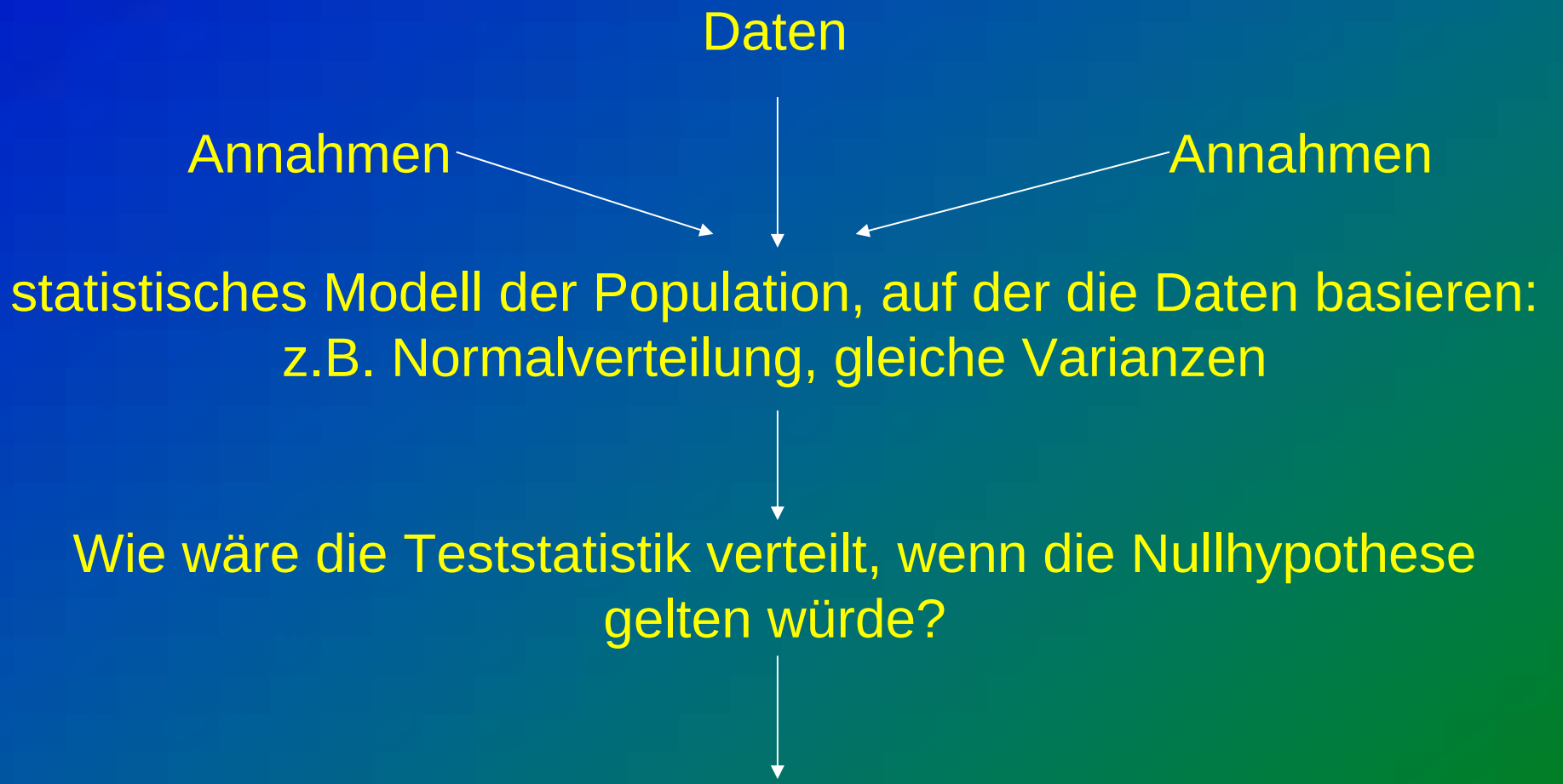
Wenn die Daten zu stark von der H_0 abweichen, wird sie verworfen.

Vorsicht: Das heißt noch nicht, daß die Alternativhypothese **wahr** ist!

Wenn die Daten nicht stark genug von der H_0 abweichen, wird sie beibehalten.

Vorsicht: Das heißt noch nicht, daß die Alternativhypothese **falsch** ist!

Alle statistischen Prozeduren funktionieren auf die gleiche Weise:



Teststatistiken

Eine Teststatistik beschreibt die Größe des Effekts, den man prüfen möchte, und zwar relativ zum Ausmaß der Zufallseinflüsse.

Eine gute Teststatistik sollte dann große Werte annehmen, wenn

- die Daten stark von der Nullhypothese abweichen;
- das Zufallsrauschen gering ist;
- die Stichprobe groß ist.

Unter bestimmten Annahmen kann man ein mathematisches Modell für die Wahrscheinlichkeits-verteilung der Teststatistik entwickeln – so kann man abschätzen, ob ein bestimmter Wert der Teststatistik „extrem“ ist oder nicht.

Alle statistischen Verfahren folgen derselben Logik, aber verwenden unterschiedliche Wahrscheinlichkeitsverteilungen:

Normalverteilung:

- z.B. psychometrische Tests mit bekannten Mittelwerten und Streuungen

t-Verteilung:

- z.B. Differenzen zwischen Mittelwerten,
- Korrelationskoeffizienten,
- Parameter einer Regressionsgleichung

F-Verteilung:

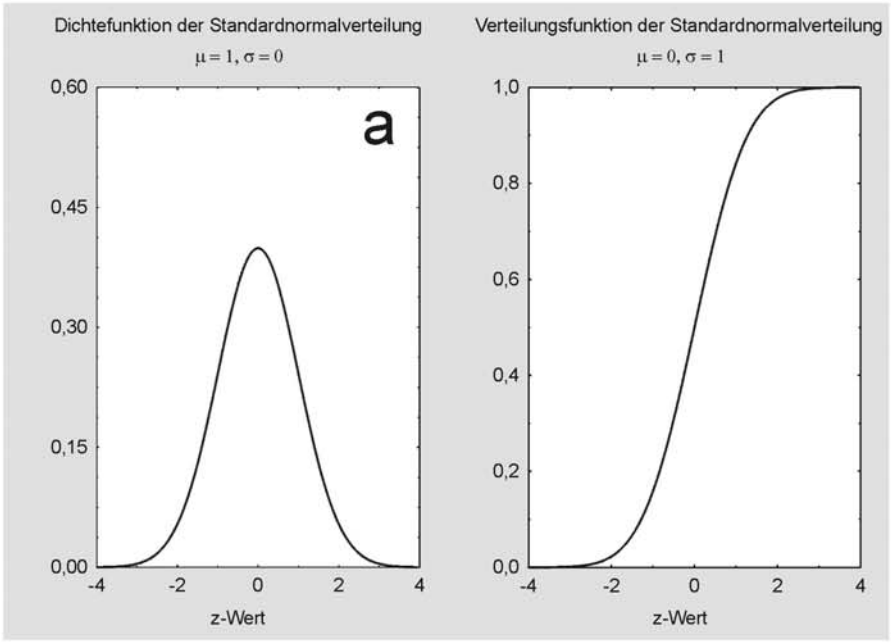
- Verhältnis zweier Varianzen, z.B. in der Varianzanalyse

Chi-Quadrat-Verteilung:

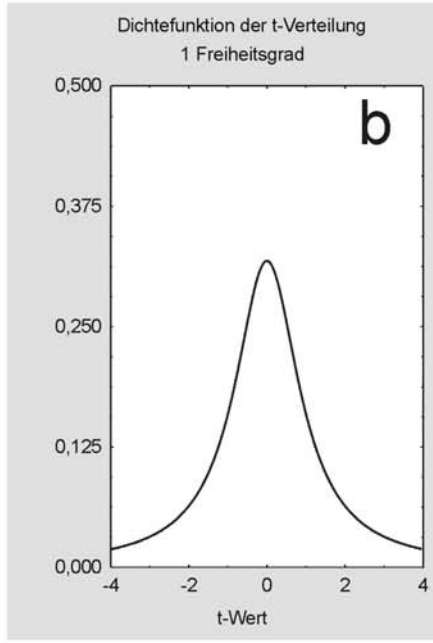
- Abweichungen beobachteter von erwarteten Häufigkeiten

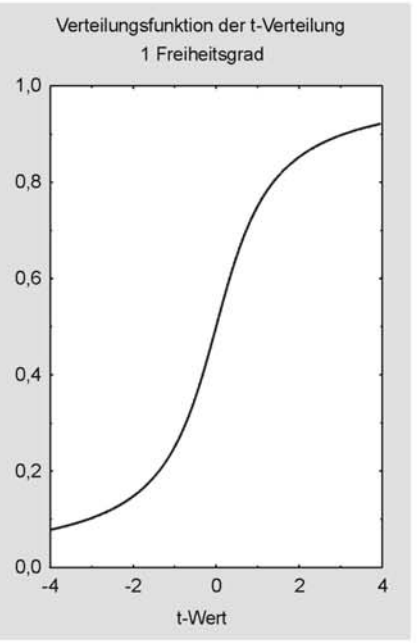
aber:
in allen Fällen geht es darum, ob eine berechnete Statistik einen "extremen" Platz in ihrer Verteilung einnimmt

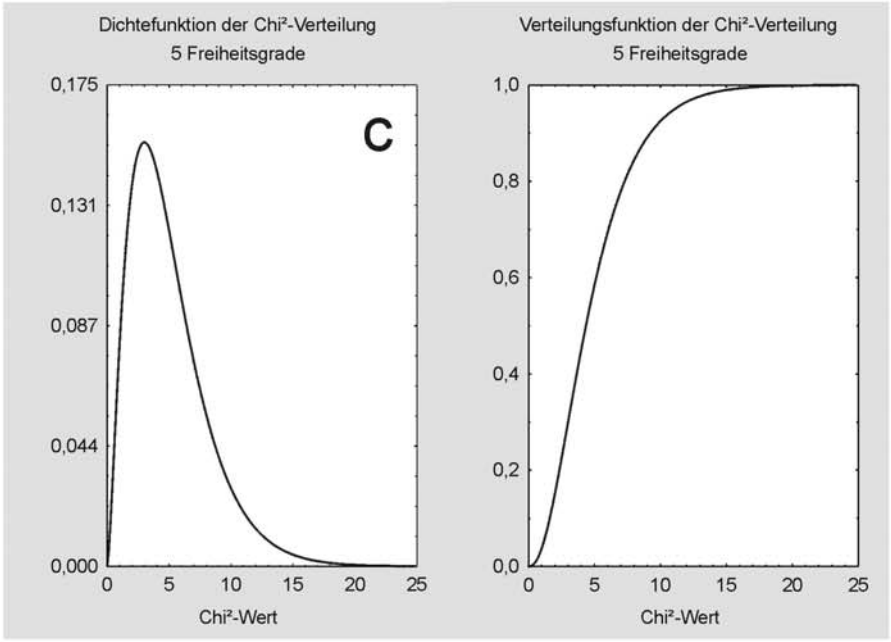
wichtigstes Kriterium: "Signifikanz" des Ergebnisses: sind weniger als 5% der Werte noch extremer als der beobachtete Wert?


Dichtefunktion der Standardnormalverteilung
$\mu = 1, \sigma = 0$
a
0,60
0,45
0,30
0,15
0,00
-4
-2
0
2
4
z-Wert

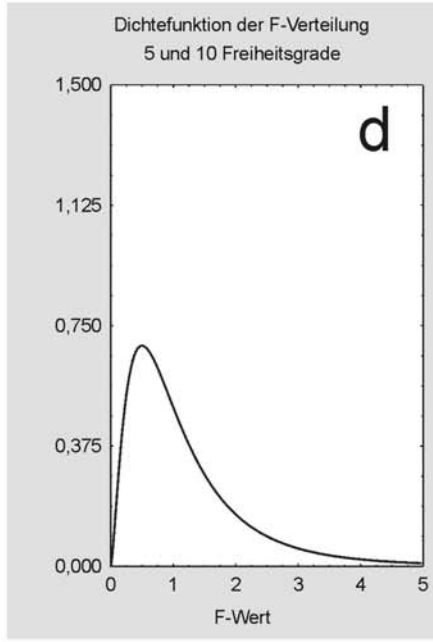
Verteilungsfunktion der Standardnormalverteilung
$\mu = 0, \sigma = 1$
1,0
0,8
0,6
0,4
0,2
0,0
-4
-2
0
2
4
z-Wert

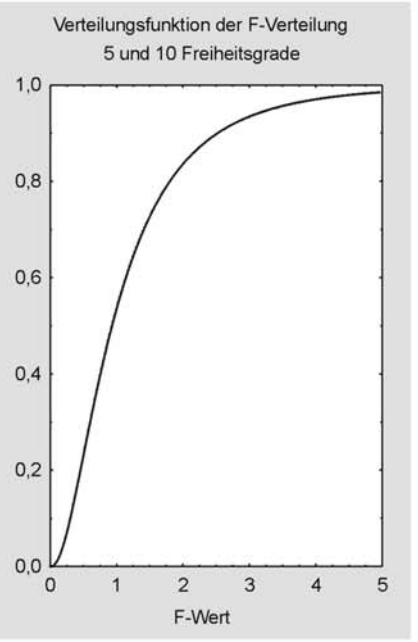

Dichtefunktion der t-Verteilung
1 Freiheitsgrad
b
0,500
0,375
0,250
0,125
0,000
-4
-2
0
2
4
t-Wert


Verteilungsfunktion der t-Verteilung
1 Freiheitsgrad
1,0
0,8
0,6
0,4
0,2
0,0
-4
-2
0
2
4
t-Wert


Dichtefunktion der Chi²-Verteilung
5 Freiheitsgrade
c
0,175
0,131
0,087
0,044
0,000
0
5
10
15
20
25
Chi²-Wert

Verteilungsfunktion der Chi²-Verteilung
5 Freiheitsgrade
1,0
0,8
0,6
0,4
0,2
0,0
0
5
10
15
20
25
Chi²-Wert


Dichtefunktion der F-Verteilung
5 und 10 Freiheitsgrade
d
1,500
1,125
0,750
0,375
0,000
0
1
2
3
4
5
F-Wert


Verteilungsfunktion der F-Verteilung
5 und 10 Freiheitsgrade
1,0
0,8
0,6
0,4
0,2
0,0
0
1
2
3
4
5
F-Wert

Beispiel: Vergleich von zwei Mittelwerten

Auswirkung von verschiedenen Arten von Musik auf die Leistung in einem Intelligenztest:

Gruppe A bearbeitet den Test zu Musik von Mozart

Gruppe B bearbeitet den Test zu Musik von jemand **ganz anderem!**

Nullhypothese: Die beiden Gruppen sind gleich gut.

Alternativhypothese: Eine der beiden Gruppen ist besser.

Der t-Test

Zweck:
Vergleich zweier Mittelwerte

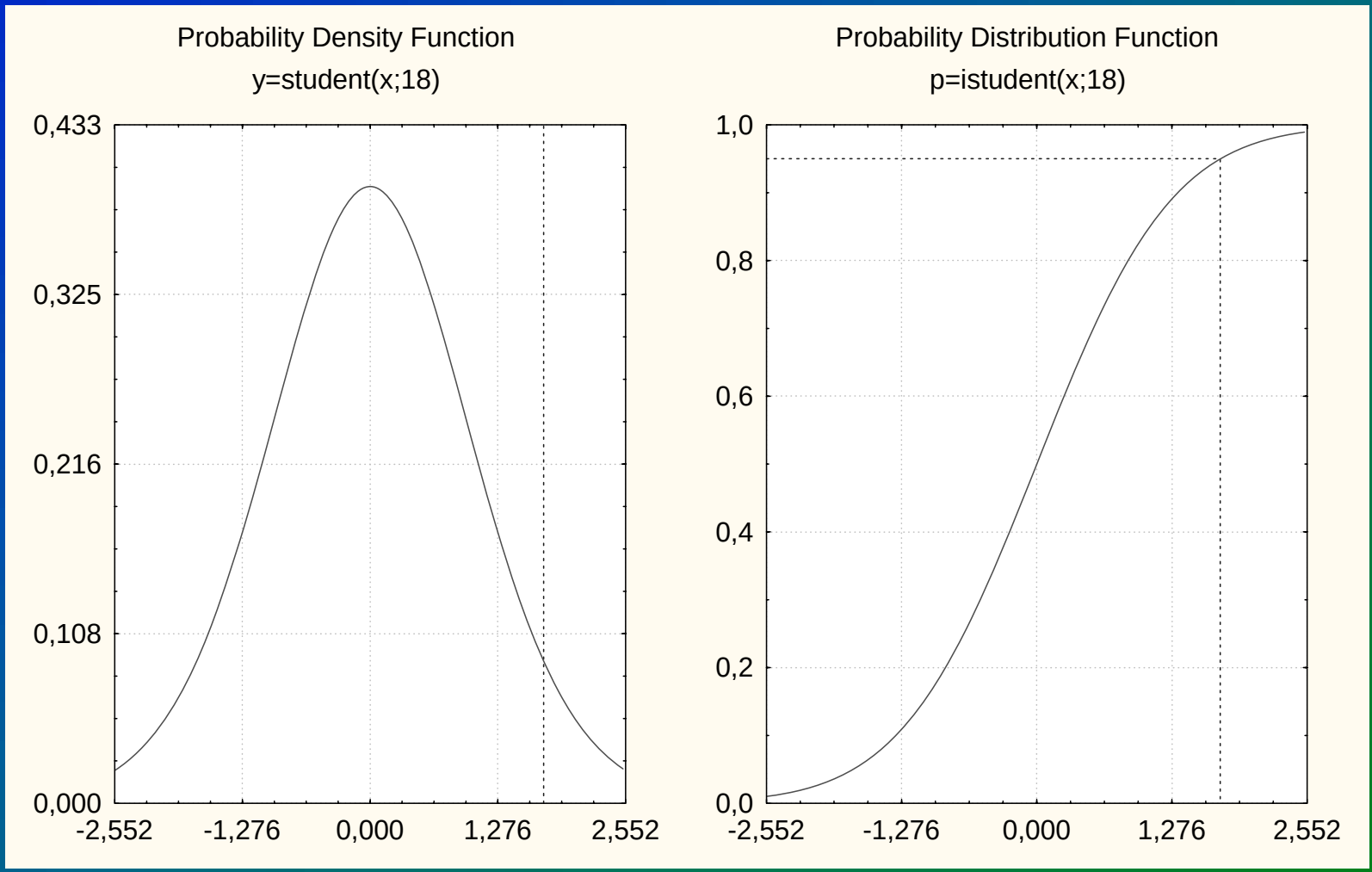
Daten:
Werte von zwei Gruppen von Vpn oder von einer Gruppe in zwei Bedingungen

Modellannahmen:
- Werte sind in den Teilpopulationen normalverteilt und unabhängig
- Verteilungen haben dieselbe Varianz

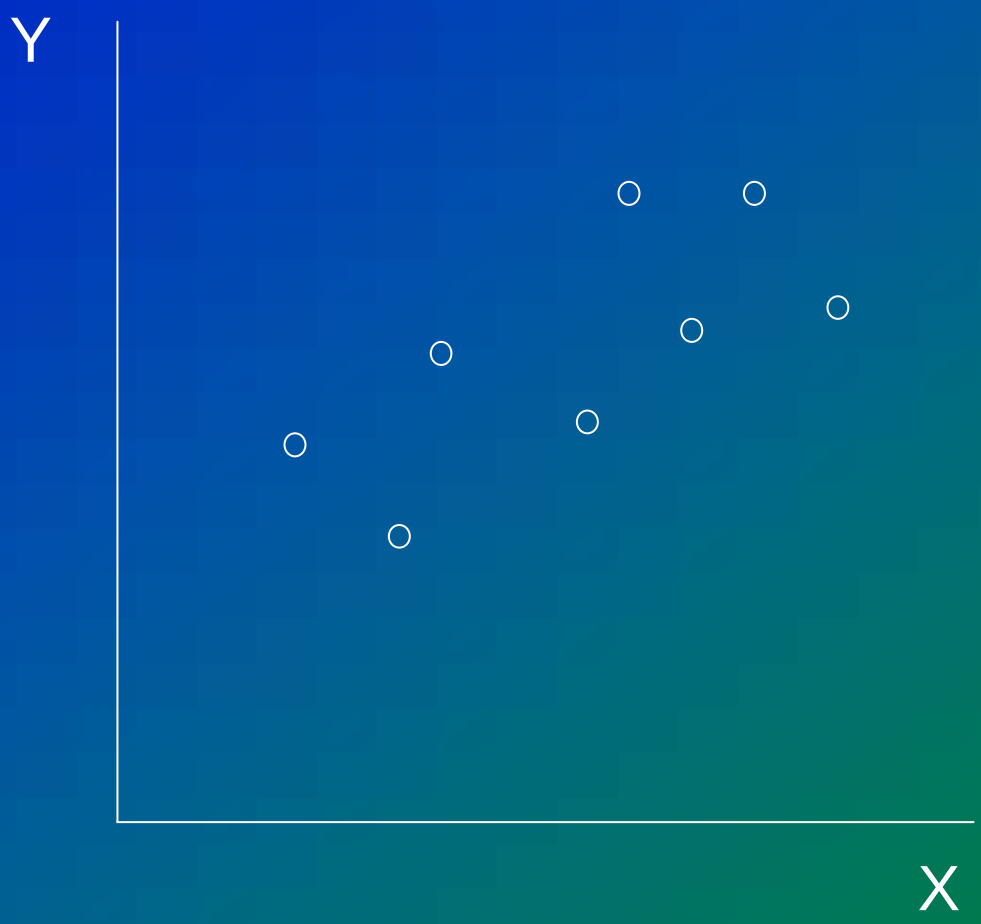
Teststatistik (genaue Formel hängt vom Design ab):
t-Wert, Differenz der Mittelwerte ist t-verteilt mit einer bestimmten Zahl von Freiheitsgraden

Beispiel:

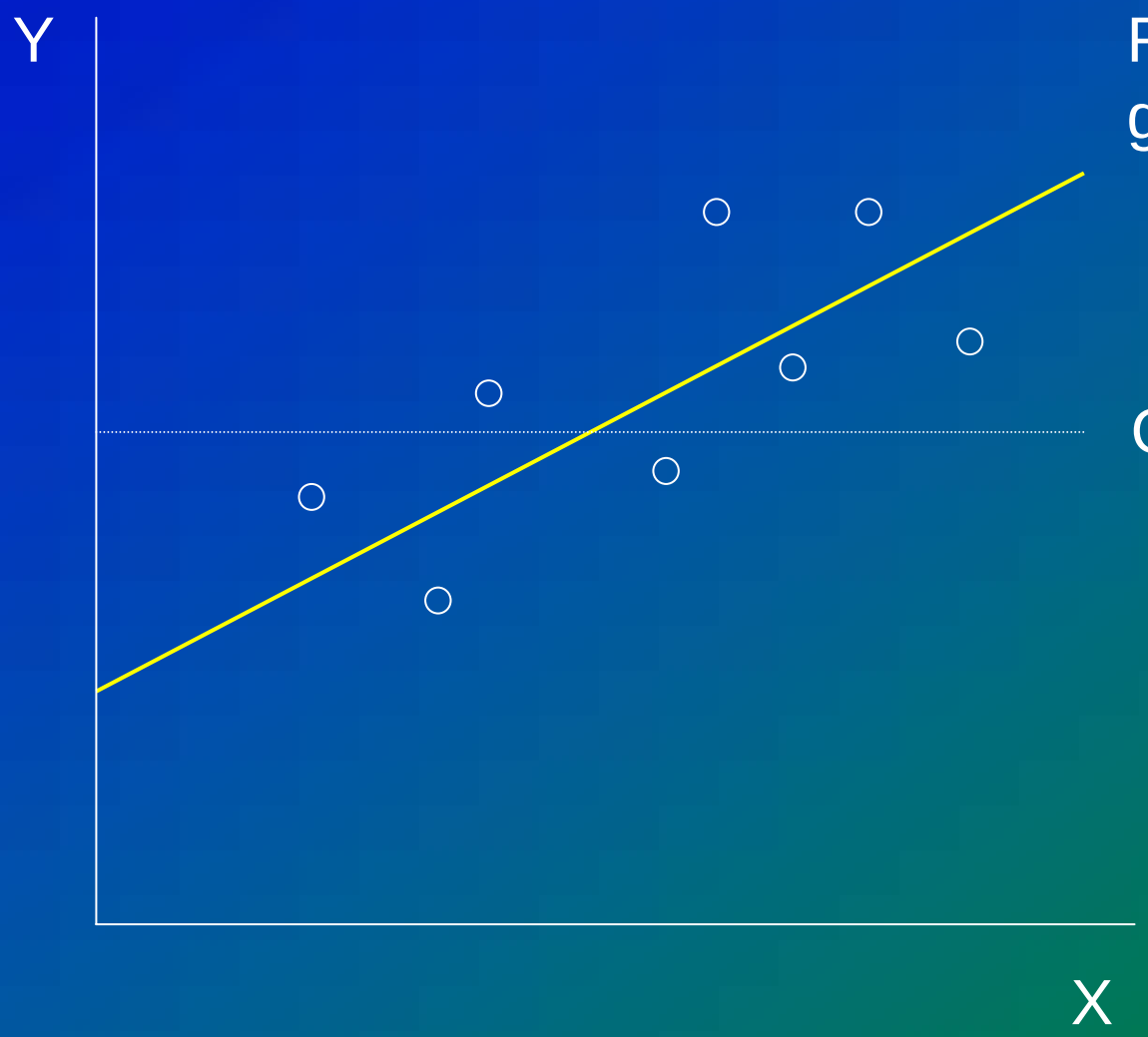
Gerichteter t-Test für unabhängige Stichproben
zwei Gruppen, je 10 Vpn, normalverteilt, gleiche Varianz



Lineare Regression



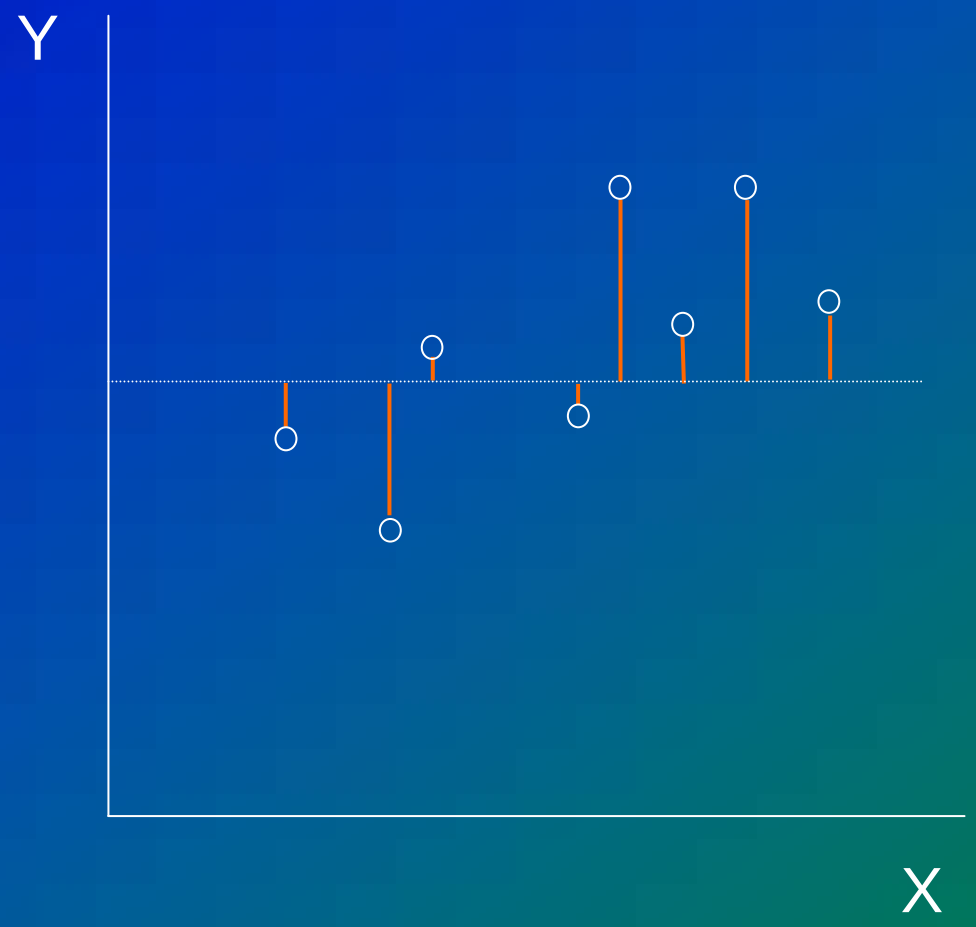
Können wir eine Gerade finden, die die Daten möglichst gut beschreibt?



Regressions-
gerade

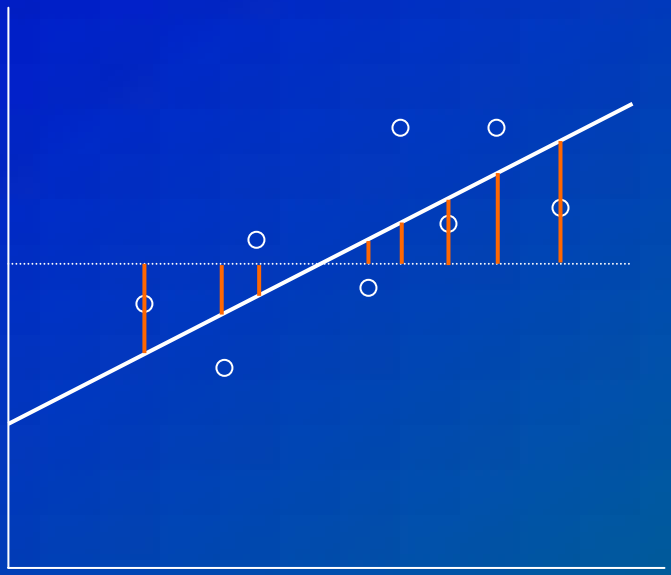
Gesamtmittelwert

Wie gut paßt die Gerade auf die Daten?

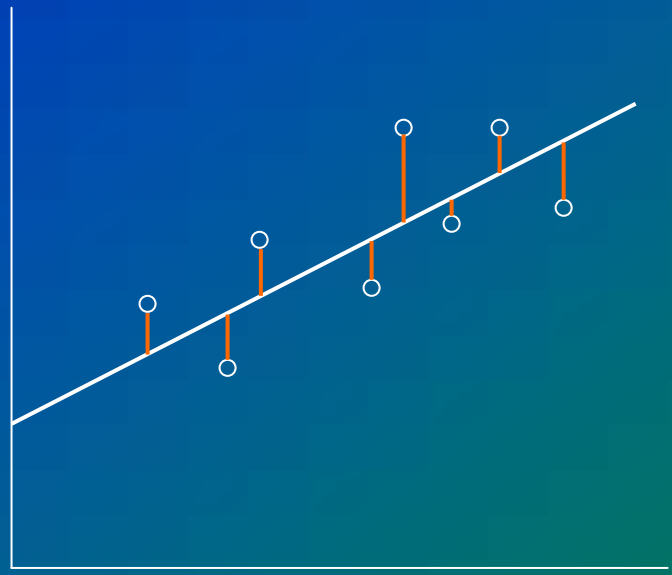


Abstände der Daten vom Gesamtmittelwert quadrieren und aufsummieren: $\mathbf{SS_{tot}}$

"Quadratsummen-Zerlegung":



Abstände zwischen Regression und Gesamtmittelwert quadrieren und aufsummieren: $\mathbf{SS_{reg}}$



Regressionsresiduen quadrieren und aufsummieren: $\mathbf{SS_{res}}$

$$\mathbf{SS_{tot} = SS_{reg} + SS_{res}}$$

$$\mathbf{R^2 = SS_{reg} / SS_{tot}}$$

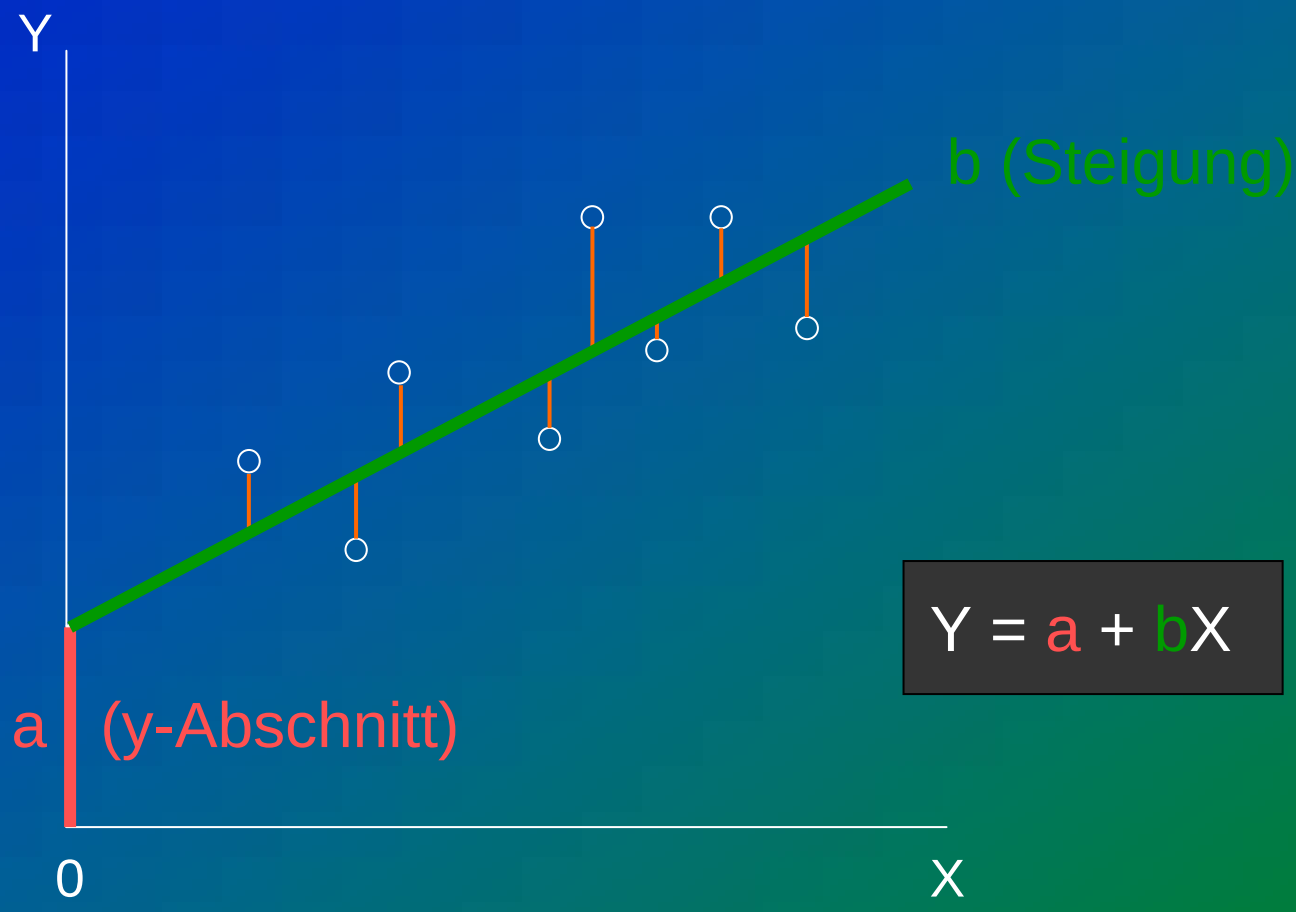
R^2 ist ein Maß dafür, wie gut das Regressionsmodell die Daten beschreibt: eine ***Goodness-of-Fit-Statistik***

R^2 = Prozentsatz der "erklärten Varianz"

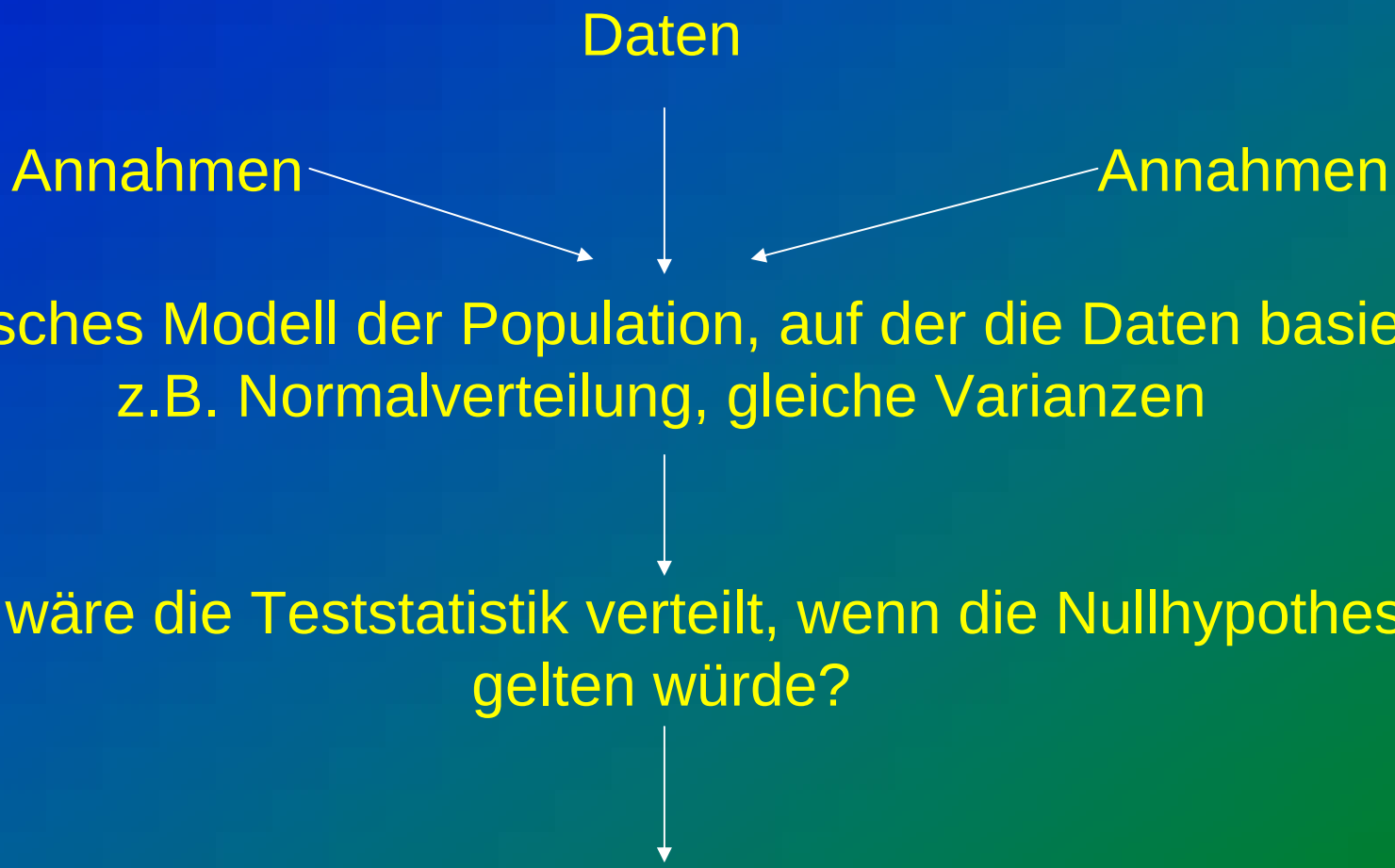
R^2 = 1: Regression erklärt die gesamte Varianz im Datensatz, alle Residuen sind 0

R^2 = 0: Regression erklärt überhaupt keine Varianz

Parameter für die Steigung und den y-Achsen-Abschnitt werden per t-Test gegen 0 getestet:



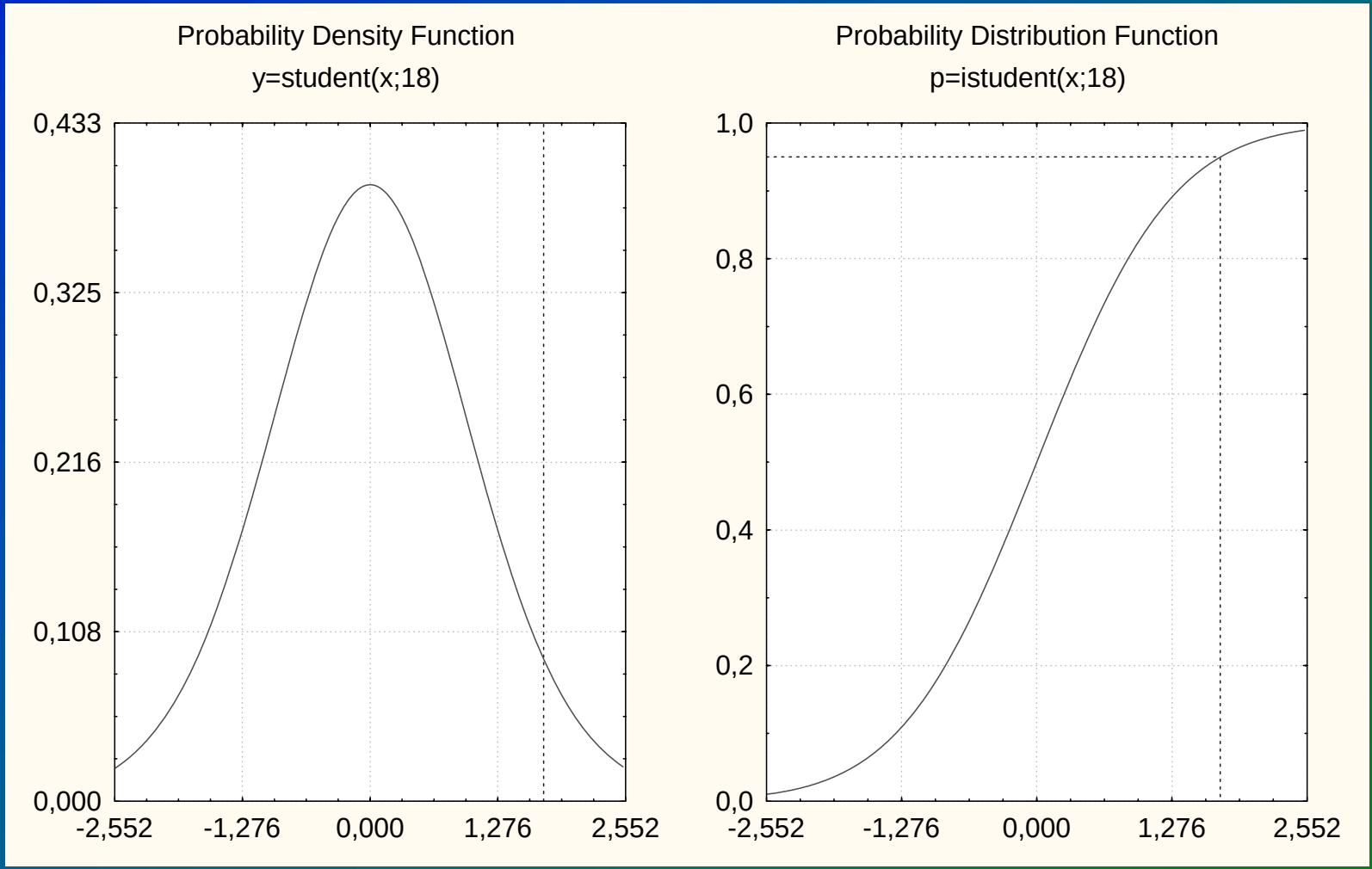
Alle statistischen Prozeduren funktionieren auf die gleiche Weise:

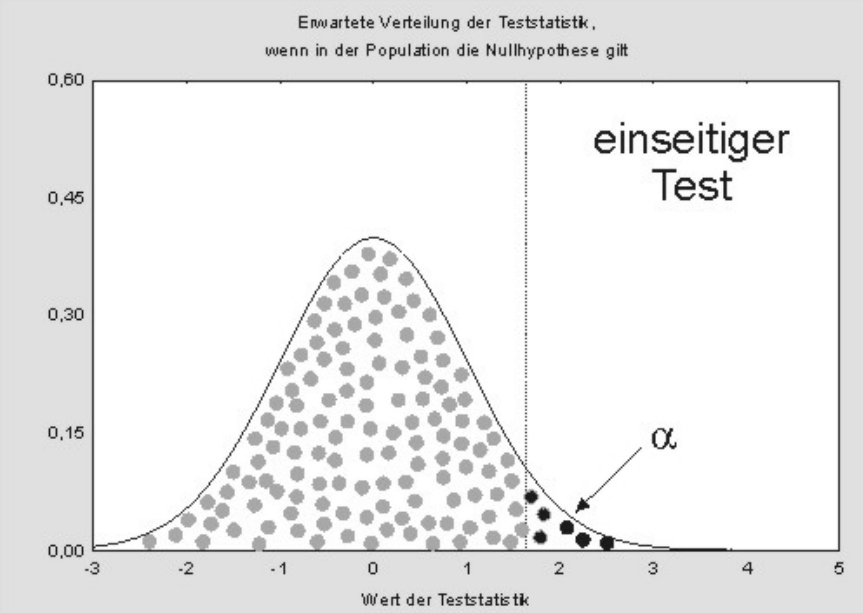


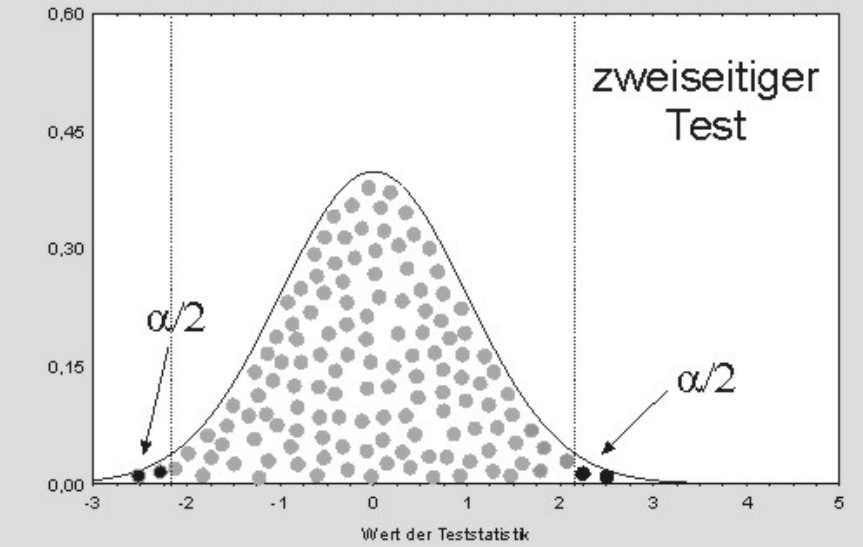
Beispiel:

einseitiger t-Test für unabhängige Stichproben

zwei Gruppen, je 10 Vpn, normalverteilt, gleiche Varianz

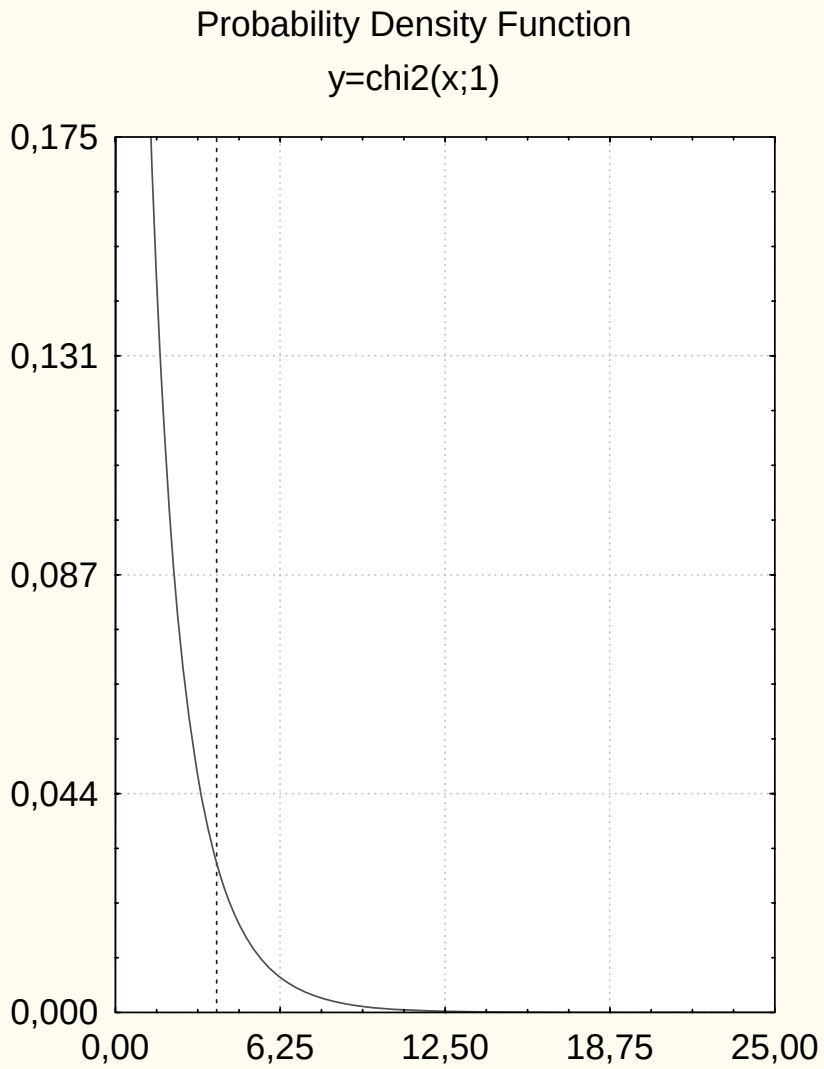


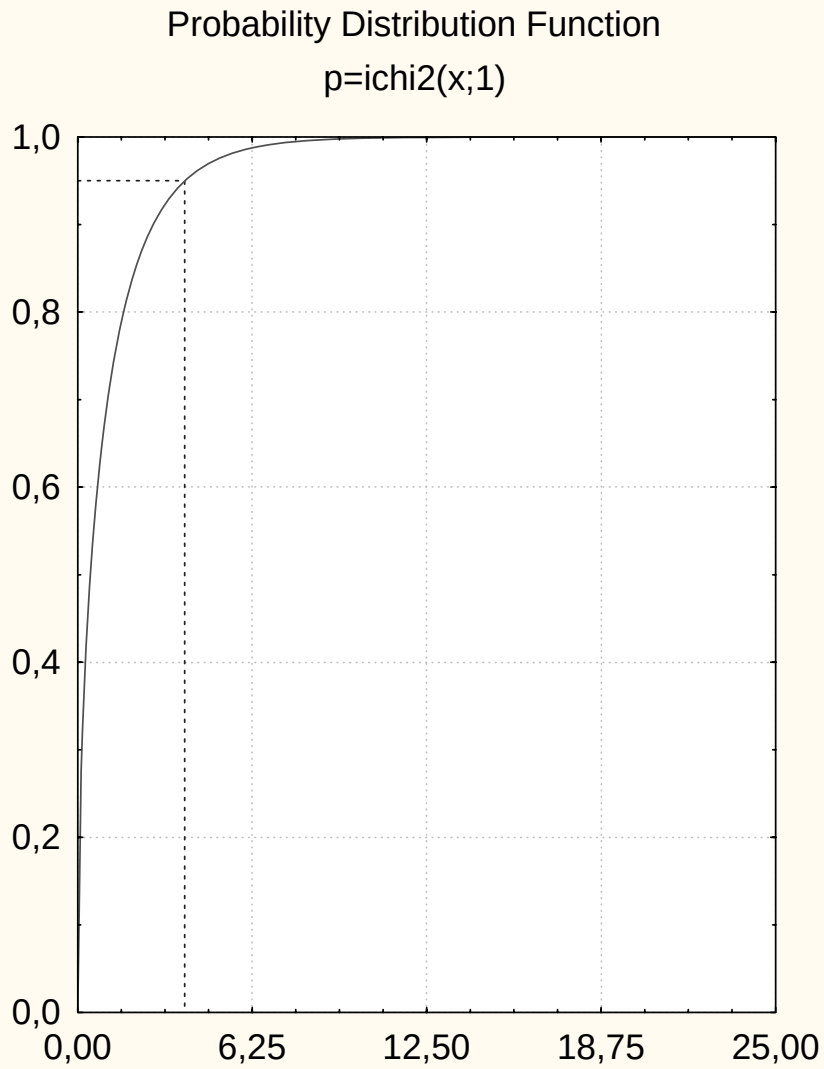




Einseitige Tests und zweiseitige Tests unterscheiden sich nur in ihren Kriterien zur Ablehnung der Nullhypothese, nicht in der Form der Verteilungen!

Beispiel: 4-Felder-Chi-Quadrat-Test

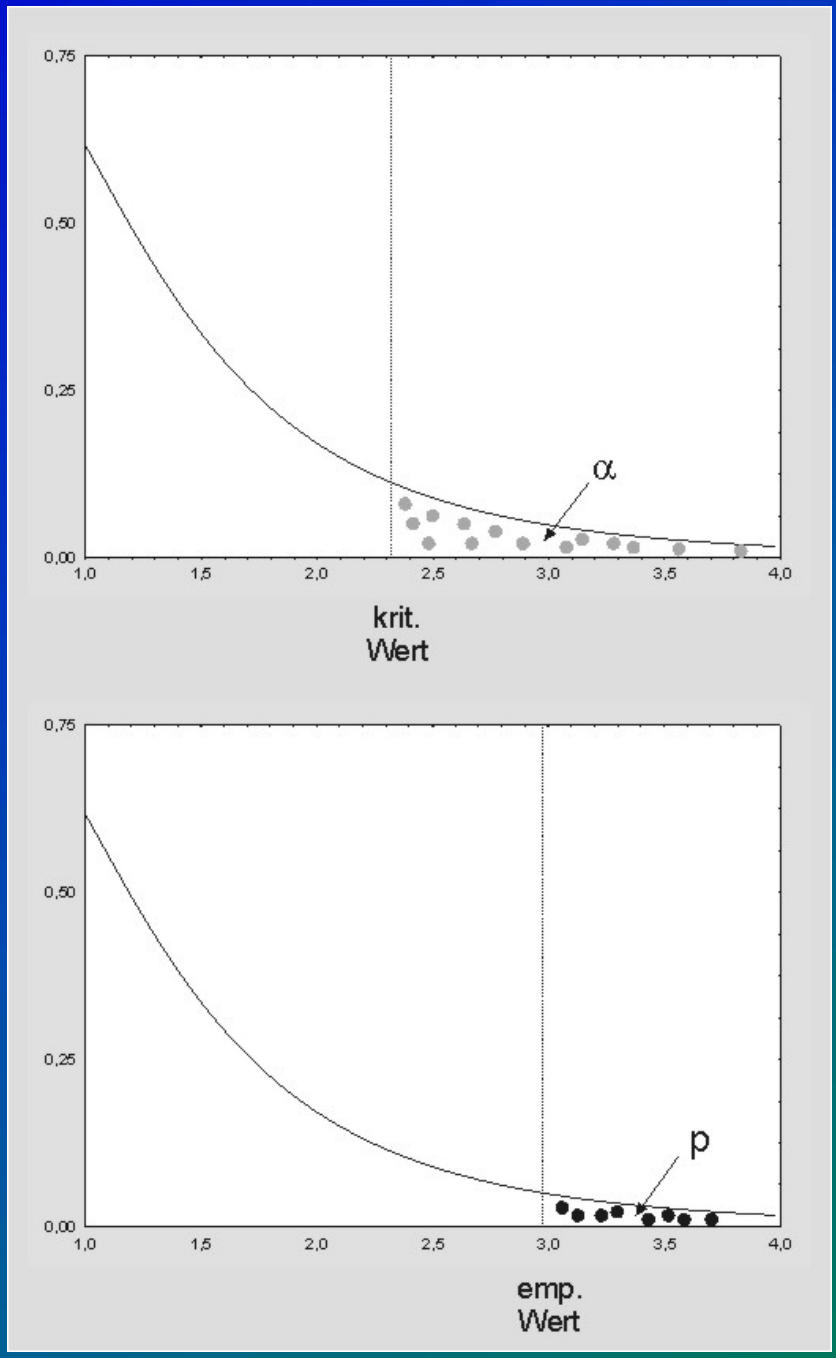




Was bedeutet der p-Wert?

- Stichprobe ist erhoben worden
- Teststatistik (z.B. t-Wert) ist berechnet worden
- die Dichtefunktion der Teststatistik (z.B. t-Verteilung) beschreibt, wie sich die t-Werte verteilen sollten, wenn die Nullhypothese gilt
- p-Wert gibt den Prozentsatz der möglichen t-Werte an, die *noch extremer* sind als der empirisch gefundene Wert

"p" ist die Wahrscheinlichkeit, daß die Daten aus einer Population stammen, in der die Nullhypothese gilt - *vorausgesetzt*, das statistische Modell ist korrekt und alle seine Bedingungen sind erfüllt!



Der kritische Wert liegt dort, wo er die Fläche α von der Verteilung abschneidet.

Der p-Wert ist der Wert rechts vom tatsächlich beobachteten Wert der Teststatistik.

Signifikanz:
(emp. Wert $\geq$ krit. Wert) ?
($p \leq \alpha$) ?

Je kleiner der p-Wert, desto deutlicher widersprechen die Daten der Nullhypothese!

Wann setzt man welches Verfahren ein?

Vorberg, D. & Blankenberger, S. (1999). Die Auswahl statistischer Tests und Maße. *Psychologische Rundschau, 50,* 157-164.

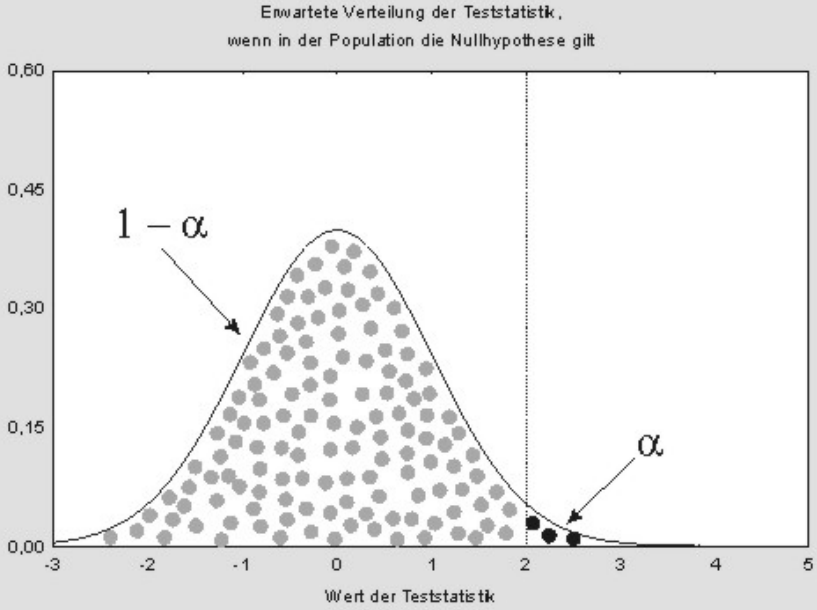
Fehlentscheidungen beim statistischen Testen

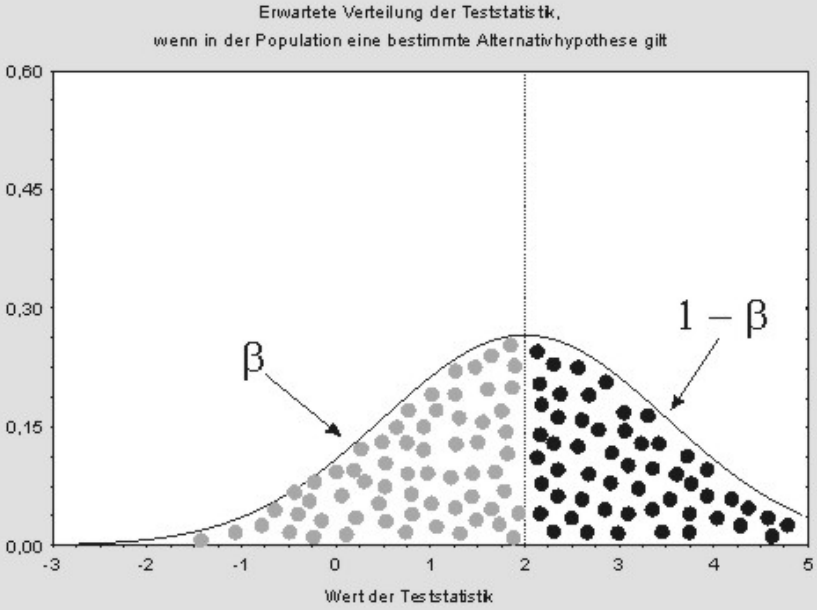
		in der Population gilt:	
		H_0	H_1
der Test entscheidet:	H_0 gilt	Korrekte Beibehaltung	Verpasser (Beta-Fehler)
	H_1 gilt	Falscher Alarm (Alpha-Fehler)	Korrekte Ablehnung

Statistische Power

Beta ist die Wahrscheinlichkeit, daß man einen "echten" Effekt übersieht

(1 - Beta) ist die Wahrscheinlichkeit, daß man einen Effekt **nicht** übersieht: Teststärke oder **Power**


Erwartete Verteilung der Teststatistik,
wenn in der Population die Nullhypothese gilt
0,60
0,45
0,30
0,15
0,00
1 – α
α
-3
-2
-1
0
1
2
3
4
5
Wert der Teststatistik


Erwartete Verteilung der Teststatistik,
wenn in der Population eine bestimmte Alternativhypothese gilt
0,60
0,45
0,30
0,15
0,00
β
1 – β
-3
-2
-1
0
1
2
3
4
5
Wert der Teststatistik

Frage: Was ist schlimmer - Alpha- oder Betafehler?

Antwort: Kommt auf den Schaden an, den die beiden Fehlerarten anrichten können!

Beispiel: Fehldiagnose einer Krankheit

Alpha-Fehler: Risiko, eine Krankheit zu behandeln, die gar nicht da ist

Beta-Fehler: Risiko, eine Krankheit zu übersehen

Teststärke und Stichprobengröße

mangelnde Teststärke ist das Resultat unpräziser Messungen:

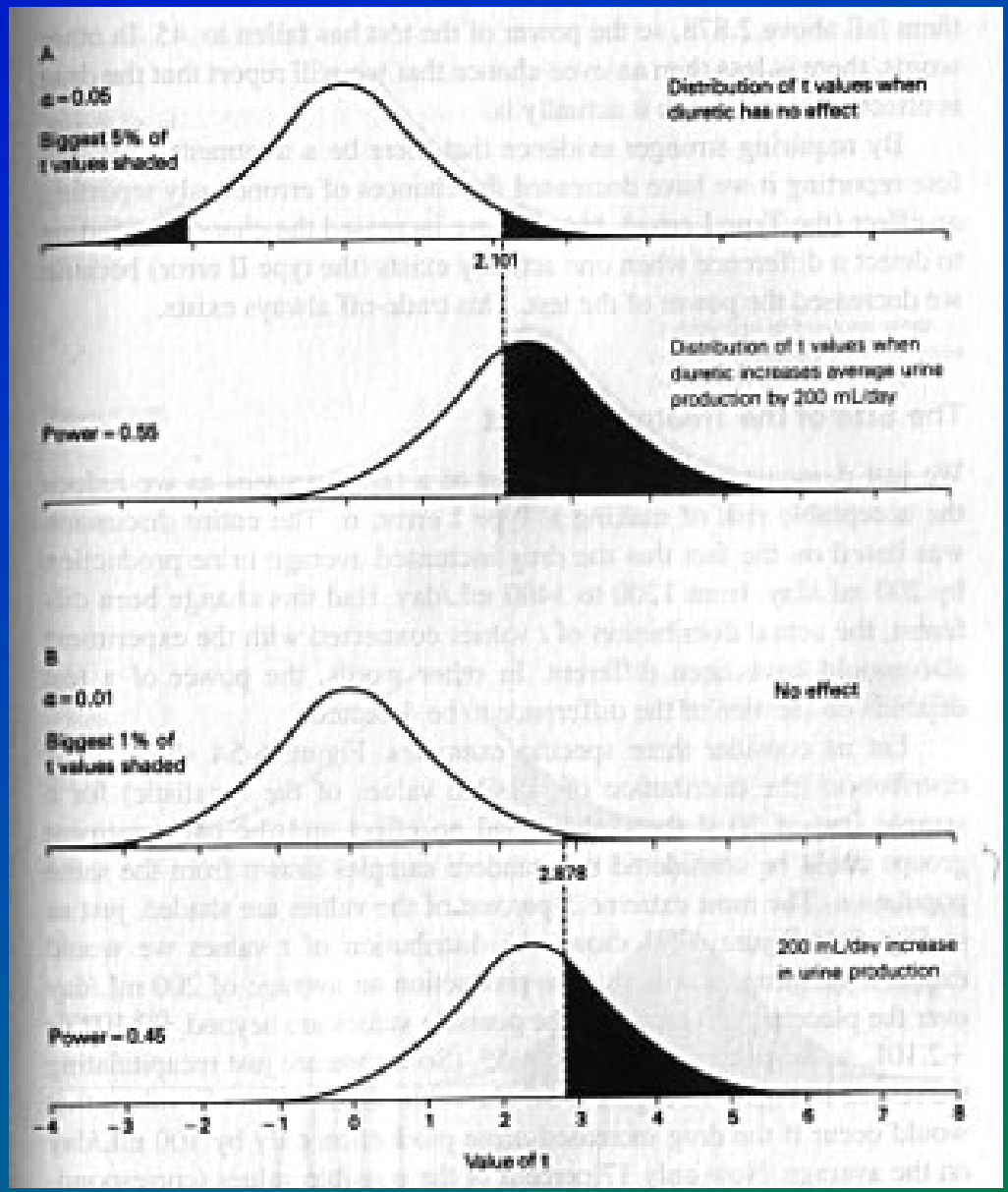
- zu kleine Stichproben
- zu ungünstige Meßbedingungen

Das führt zu großen Problemen:

- in der Population vorhandene Effekte werden nicht signifikant
- Nulleffekte können nicht interpretiert werden

Im Extremfall sind ganze Studien umsonst gewesen!

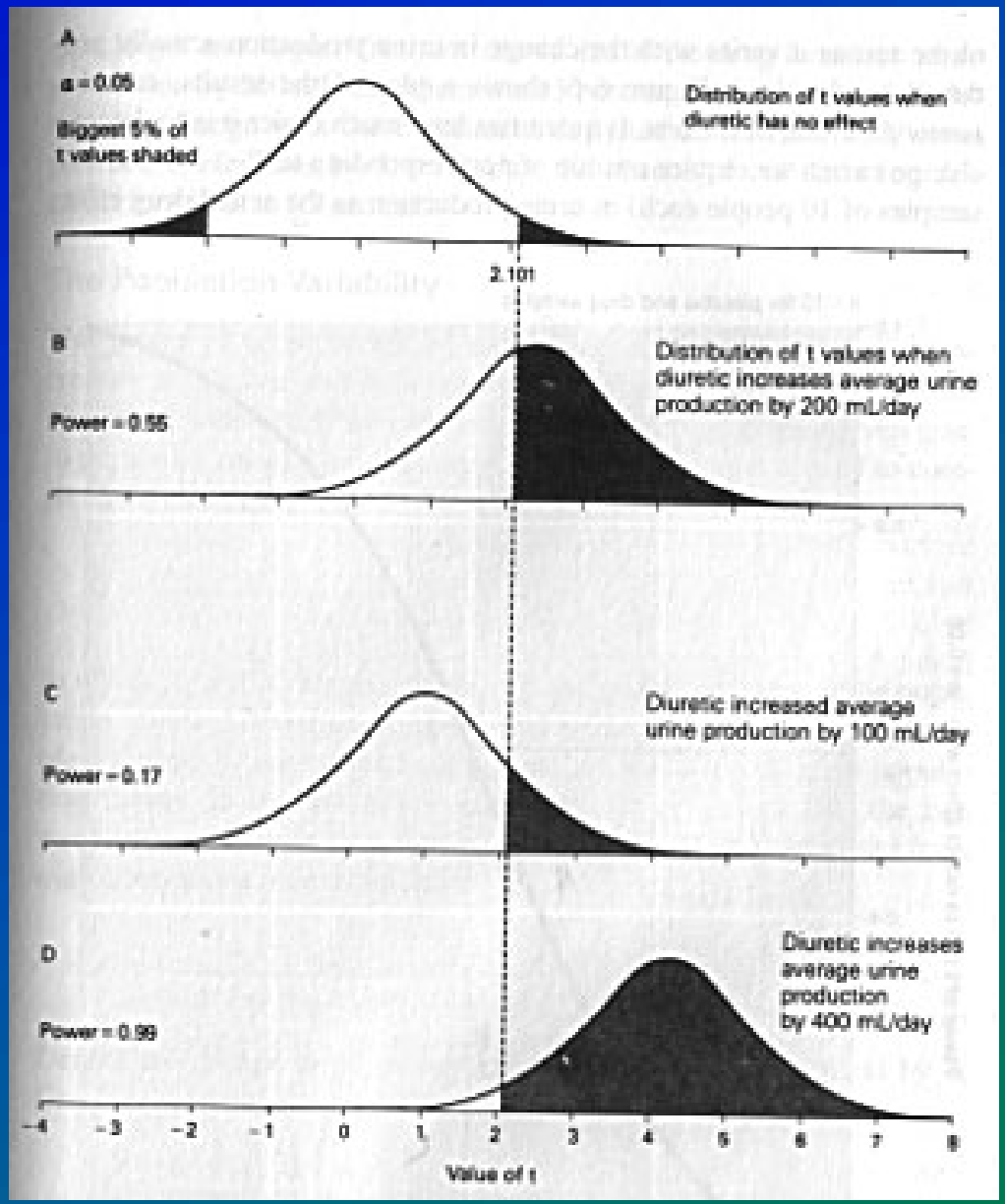
Statistische Power



Strengeres α-Niveau:

- Weniger Fehler 1. Art
- Mehr Fehler 2. Art
- Weniger Power

Statistische Power



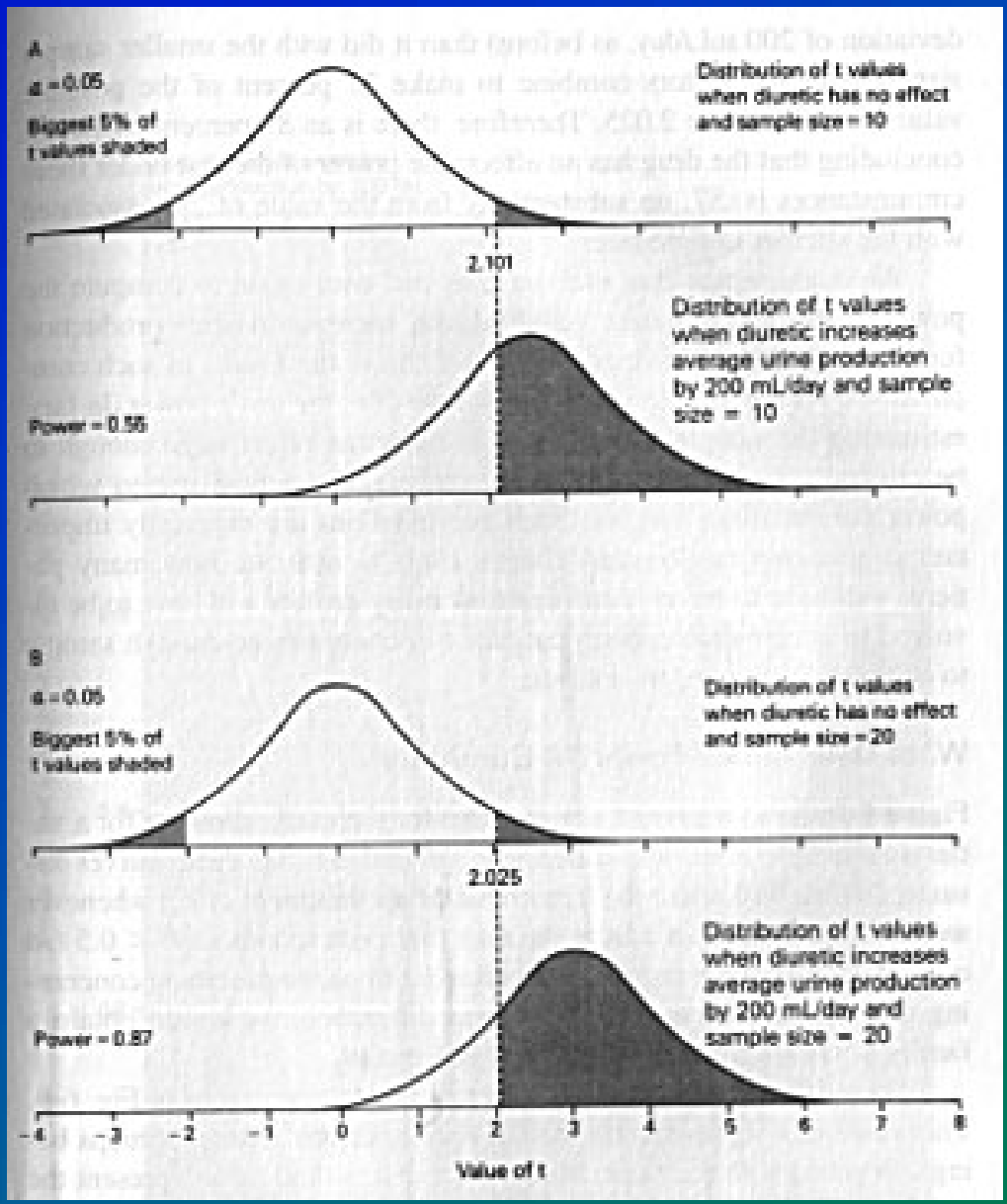
Kleinerer Effekt:

- Mehr Fehler 1. Art
- Mehr Fehler 2. Art
- Weniger Power

Größerer Effekt:

- Weniger Fehler 1. Art
- Weniger Fehler 2. Art
- Mehr Power

Statistische Power



Größere Stichprobe (= mehr Freiheitsgrade):

- kritischer t-Wert sinkt
- rechter Teil der H_0-Verteilung wird flacher
- empirischer t-Wert wird größer

- weniger Fehler 2. Art
- mehr Power

Achtung: der Prozeß verläuft asymptotisch!

Kritik an der klassischen Inferenzstatistik

„Die Nullhypothese ist ohnehin falsch, man kann sie gar nicht akzeptieren.“ (Murphy & Myors, 2000)

„Signifikanz sagt nichts darüber, ob die beobachteten Effekte wichtig oder unwichtig sind.“

„Mit ausreichend großen Stichproben bekommt man jeden Effekt signifikant.“

„Signifikanztests sagen nichts über die Wahrscheinlichkeit der Alternativhypothese.“

„Tests sind nur ein Ritual; man braucht statistisches Denken. Signifikanztests fördern die Verwendung von ungenauen Hypothesen.“ (Gigerenzer, 1998)

„Anwender kennen nur wenige Techniken, verwenden sie inkonsistent, verstehen die Ergebnisse nicht.“

„Anwender sind besessen vom Alpha-Fehler und ignorieren den Beta-Fehler; daher sind die meisten Studien von niedriger Qualität“.

Sind Signifikanztests ein geeignetes Mittel der Qualitätskontrolle?