
Statistische Tests?

Dr. Dirk Hasenclever
University of Leipzig

[imise.](#)

Traditionell gibt es zwei Basisverfahren statistischer Inferenz

■ Schätzen von unbekannten Modellparametern

- **Punktschätzer**
 - Was ist ein guter Schätzwert?
- **Konfidenzintervalle** zum Punktschätzer
 - Wie genau wissen wir es eigentlich?

■ Gute biometrische Praxis:

- **Kein Schätzwert ohne Konfidenzintervall!**

■ Beurteilen von statistischen Hypothesen

- **Nullhypothese**
 - Behauptung / Theorie über den unbekannten Parameter
- **Alternativhypothese**
 - Hypothese über Art der Abweichung von der Nullhypothese.
- **Statistischer Test**
 - Sind die **Daten verträglich mit der Nullhypothese?**
 - *Ergebnis eines statistischen Tests ist ein **Argument!***

Meine Meinung: Eigentlich...

- **Eigentlich...** braucht man nur gute Punktschätzer mit guten zugehörigen $(1-\alpha)$ -Konfidenzintervall-Verfahren!
 - Guter Punktschätzer:
 - Konsistent
 - möglichst Erwartungstreu und
 - minimaler Mean Square Error
 - Gutes $(1-\alpha)$ -Konfidenzintervall-Verfahren:
 - Einhaltung der $(1-\alpha)$ -Überdeckungseigenschaft
 - bei minimaler Breite
- Aus historischen Gründen spielen statistische Tests eine zu große Rolle und leider werden oft missinterpretiert.

Was sollen statistische Tests leisten?

- **Rationalisiere wissenschaftlicher Diskurs über sogenannte statistische Hypothesen:**
- **Statistische Hypothesen** = Behauptungen über latente Eigenschaften identisch wiederholbarer Zufallsexperimente,
 - d.h. über ihre Verteilungsfunktion
 - bzw. Größen, die man aus dieser theoretisch ableiten kann.
- Beispiele für statistische Hypothesen mit Hämoglobinwerten
 - sind normalverteilt,
 - Erwartungswert = 15 g/dl bei gesunden Männern
 - Frauen haben im Erwartungswert niedrigere Hb-Werte als Männer

Argumentieren mit Konfidenzintervallen

- Theorie/Nullhypothese: Jemand behauptet
 - H_0 : Die Erwartungswerte des Hb von Männern und Frauen sind gleich.
- Überprüfung/ Widerlegung dieser Theorie:
 - Einigung auf Messverfahren und Skala [g/dl]
 - Wähle *hinreichend große* Stichprobe von Probanden
 - Achtung: Auswahl möglichst ohne differentielle Selektionseffekte.
 - Messe die Hb-Werte bei N_m Männern und N_f Frauen
 - Schätze die Differenz: $E(\text{Hb}|\text{Mann}) - E(\text{Hb}|\text{Frau})$ mit Konfidenzintervall(en)

Wie schätzen wir eine Differenz im Erwartungswert?

- 2 Zufallsvariable: HbM, HbF approximativ normalverteilt
- Schätze $E(\text{HbM})$ durch $\overline{\text{HbM}}_{N_m}$ (arithmetischer Mittelwert)
- und $\widehat{\text{se}}(\overline{\text{HbM}}_{N_m}) = \overline{\text{SD}}(\text{HbM}) / \text{Wurzel}(N_m)$ (mit unserem Varianzschätzer)
- und analog für Frauen.
- Schätze $E(\text{HbM}) - E(\text{HbF})$ durch $\overline{\text{HbM}}_{N_m} - \overline{\text{HbF}}_{N_f}$
- Standardfehler: $\widehat{\text{se}}(\text{Differenzschätzung}) =$

$$\text{Wurzel}(\widehat{\text{se}}(\overline{\text{HbM}}_{N_m})^2 + \widehat{\text{se}}(\overline{\text{HbF}}_{N_f})^2)$$

weil $\text{var}(X-Y) = \text{var}(X) + \text{var}(Y)$ bei Unabhängigkeit

Wie schätzen wir eine Differenz im Erwartungswert? II

- Zweiseitiges $(1-\alpha)$ – Konfidenzintervall:

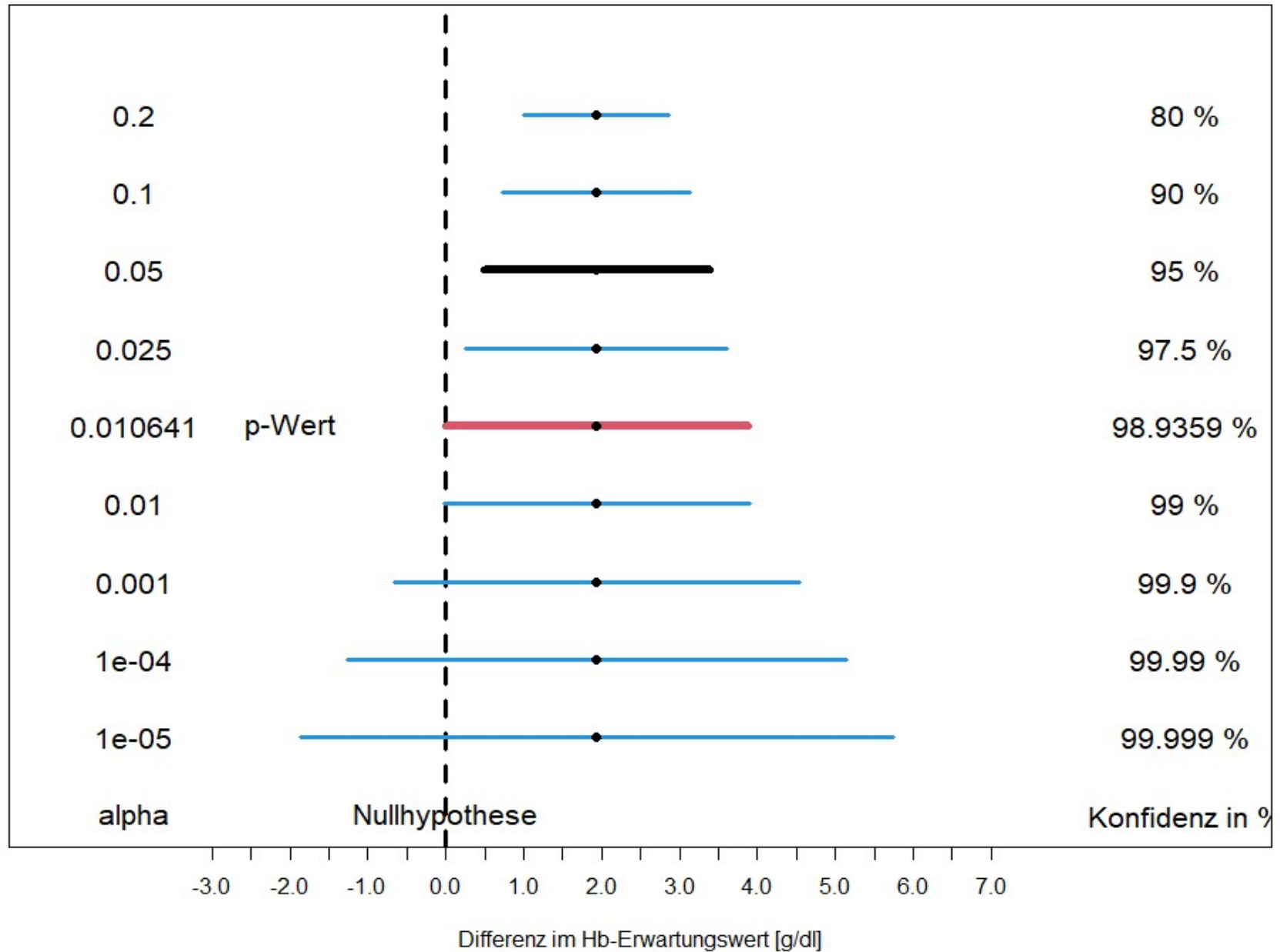
- $\overline{\text{HbM}}_{\text{Nm}} - \overline{\text{HbF}}_{\text{Nf}} \pm$
 $qt(1-\alpha/2, df = \text{Nm} + \text{Nf} - 2) * \hat{s}(\text{Diffschätzung})$

↑
Quantil der t Verteilung mit geeigneter Zahl von Freiheitsgraden (df)

Wie stark ist die Evidenz gegen die Nullhypothese?

- p-Wert = Supremum aller α , so dass die Nullhypothese außerhalb des $(1-\alpha)$ Konfidentintervalls liegt, gegeben die vorliegenden Daten.
- Beantwortet die Frage: Wie gut ist die Nullhypothese verträglich mit den gegebenen Daten.
- Alternative (nur bei normalverteilten Schätzern):
 - Abstand Nullhypothese zum Schätzwert in Einheiten des Standardfehlers.

Bestimmung des p-Werts



Zwei-Stichproben (ungepaarter) t-Test I

$$X_i \sim N(\mu_X, \sigma_X^2) \quad i = 1..n$$

$$Y_k \sim N(\mu_Y, \sigma_Y^2) \quad k = 1..m$$

$$\bar{X}_n \sim N(\mu_X, \frac{\sigma_X^2}{n}) \quad \text{und} \quad \bar{Y}_m \sim N(\mu_Y, \frac{\sigma_Y^2}{m})$$

$$(\bar{X}_n - \bar{Y}_m) \sim N(\mu_X - \mu_Y, \frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m})$$

$$\frac{\bar{X}_n - \bar{Y}_m}{\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}} \sim N\left(\frac{\mu_X - \mu_Y}{\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}}, 1\right)$$

Zwei-Stichproben (ungepaarter) t-Test II

Unter $H_0 : \mu_X = \mu_Y$ und falls σ_X und σ_Y bekannt sind, gilt :

$$\frac{\bar{X}_n - \bar{Y}_m}{\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}} \sim N(0,1)$$

Wenn die Varianzen aus den Daten geschätzt werden müssen, gilt approximativ :

$$\frac{\bar{X}_n - \bar{Y}_m}{\sqrt{\frac{\hat{\sigma}_X^2}{n} + \frac{\hat{\sigma}_Y^2}{m}}} \sim t_{df=\min(n,m)-1}$$

bzw. eine noch genauere Approximation nach Welch welche gebrochene Freiheitsgrade (df) liefert.

Zwei Gruppen t-test

```
# Vergleich mit t.test  
> t.test(HbM, HbF)
```

Welch Two Sample t-test

```
data:  HbM and HbF  
t = 2.7407, df = 23.916, p-value = 0.01141  
alternative hypothesis: true difference in means is  
not equal to 0  
95 percent confidence interval:  
  0.4210228  2.9905772  
sample estimates:  
  mean of x mean of y  
14.26539   12.55959
```

Grundidee des statistischen Test

- Ein **Proponent** behauptet eine statistische Hypothese = **Nullhypothese**.
- Um diese zu überprüfen / zu widerlegen spezifiziert der **Opponent** eine **Teststatistik T** (Stichprobenfunktion, also Zufallsvariable), von der er meint, dass sie die vermutete Abweichung von der Nullhypothese gut darstellt.
- Der **Proponent** berechnet basierend auf seiner Nullhypothese die **erwartete Verteilung F_T der Teststatistik T unter der Nullhypothese = die Nullverteilung**.

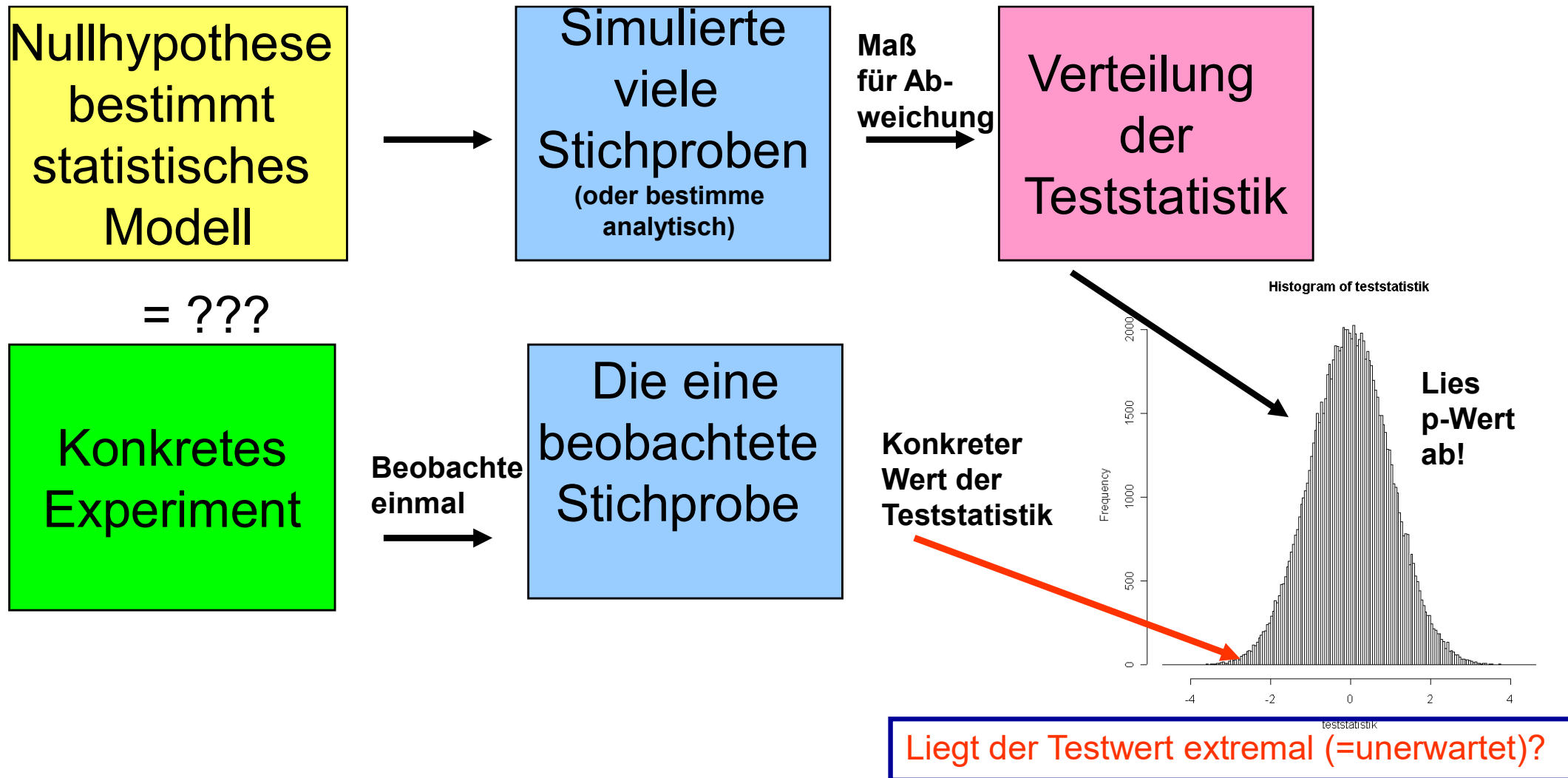
Grundidee des statistischen Test III

- Das Experiment wird mit einer geeigneten Fallzahl durchgeführt und liefert einen **konkreten Wert der Teststatistik $T = t$** .
- Nun wird anhand der Nullverteilung **bewertet wie extrem $T=t$ bzgl. der Erwartung unter der Nullhypothese ist**:
- $T=t$ kann unerwartet klein oder unerwartet groß sein:
Berechne den zweiseitigen p-Wert als
$$p = 2 * \min (F_T^{-1}(t); 1 - F_T^{-1}(t))$$
- F_T^{-1} ist dabei die inverse kumulative Verteilungsfunktion von F_T

Grundidee des statistischen Test III

- p ist die Wahrscheinlichkeit unter der Nullhypothese $T=t$ oder nach beiden Seiten noch extremere T zu beobachten.
- Falls p groß ist, liegt $T=t$ im Bereich des unter der Nullhypothese erwarteten.
- Falls p klein ist, liegt $T=t$ am Rande dessen, was der Proponent erwartet hat:
- Der **Proponent** muss dann
 - entweder behaupten, **es sei etwas nach seiner Erwartung sehr Unwahrscheinliches passiert** ist,
 - oder zugeben, **dass die Nullhypothese falsch** ist.

Statistischer Test als p-Wert Prozedur



Verteilung des P Werts unter der Nullhypothese

Die Teststatistik T ist eine Zufallsvariable mit Verteilung $F_T(t)$ und Dichte $f_T(t)$ dt

Der einseitige p-Wert ist eine Zufallsvariable $P = F_T(T)$, die unter der Nullhypothese gleichverteilt auf $[0,1]$ ist:

Nach Regel für die Dichtefunktion einer transformierten Zufallsvariable gilt:

$$f_P(p)dp = f_T(F_T^{-1}(p)) \cdot \frac{d}{dp}(F_T^{-1}(p)) = 1 dp$$

da die Ableitung der Umkehrfunktion das Inverse der Ableitung ist.

Schätzen Sie p und das 95%CI!

Therapie A				Therapie B	
E1=	13			E2=	10
N1=	20			N2=	20
P1=	0.65			P2=	0.5

Erfolgsraten 13/20 versus 10/20

2-sample test for equality of proportions
with continuity correction

data: c(13, 10) out of c(20, 20)

X-squared = 0.4092, df = 1, p-value = 0.5224

alternative hypothesis: two.sided

95 percent confidence interval:

-0.2028446 0.5028446

sample estimates:

prop 1 prop 2

0.65 0.50

Schätzen Sie p und das 95%CI!

Therapie A				Therapie B	
E1=	130			E2=	100
N1=	200			N2=	200
P1=	0,65			P2=	0,5

Erfolgsraten 130/200 versus 100/200

2-sample test for equality of proportions
with continuity correction

data: c(130, 100) out of c(200, 200)

X-squared = 8.6036, df = 1, **p-value = 0.003355**

alternative hypothesis: two.sided

95 percent confidence interval:

0.04923214 0.25076786

sample estimates:

prop 1 prop 2

0.65 0.50

Beispiel: Gauss-Test I

Die mathematische Stichprobe $X_1, \dots, X_N$ sei unabhängig und identisch normalverteilt $N(\mu, \sigma^2)$ mit bekannter Varianz σ^2 .

Nullhypothese: $\mu = \mu_0$

Alternativhypothese: $\mu \neq \mu_0$

Wähle als Teststatistik

$$G = \frac{(\bar{X}_N - \mu_0)}{\left(\frac{\sigma}{\sqrt{N}} \right)}$$

G ist eine Zufallsvariable und unter der Nullhypothese standardnormalverteilt $G \sim N(0,1)$ (Nullverteilung).

(Nachrechnen mit Eigenschaften der Normalverteilung!)

Beispiel: Gauss-Test II

Das Experiment liefert eine konkret realisierte Stichprobe $x_1, \dots, x_N$.

Bei $\sigma=1$ und $\mu_0 = 1$, sei z.B. $N = 16$, $\bar{x} = 1.75$

Dies ergibt für die Teststatistik G :
$$g = \frac{(1.75-1)}{\left(\frac{1}{\sqrt{16}}\right)} = 3$$

Sei $\Phi(x)$ die Verteilungsfunktion der Standardnormalverteilung.

$$p\text{-Wert} = pr(G \leq -3) + pr(G \geq 3) = \Phi(-3) + (1 - \Phi(3)) < 0.0027$$

(In R: `pnorm(-3)+(1-pnorm(3))`)

Resultat: Die Nullhypothese H_0 ist kaum zu halten. ("hochsignifikant")

p-Wert Prozedur gegen Neyman-Pearson Entscheidungsprozedur

■ p-Wert Prozedure:

- Überprüfung einer wissenschaftlichen Theorie (Nullhypothese)
- Theorie mit Daten verträglich?
- p-Wert = Wahrscheinlichkeit unter H_0 einen so extremen oder extremeren Testwert zu beobachten
- p-Wert als **Argument** gegen die Nullhypothese.

■ Entscheidungsprozedur:

- Proponent und Opponent vereinbaren hartes **Entscheidungsverfahren** zwischen H_0 und H_A aufgrund noch zu erhebender Daten!
- „**Experimentum crucis**“
 - Z.B. Zulassungsverfahren eines neuen Medikaments
 - Qualitätstest vor Abnahme einer Warenlieferung

Test als Entscheidungsverfahren I

Entscheidungsregel:

Bestimme p-Wert:

Falls $p \leq \alpha$ H_0 wird abgelehnt (Opponent gewinnt)

Falls $p > \alpha$ H_0 wird beibehalten (Proponent gewinnt)

Proponent wählt Signifikanzniveau α . Z.B. $\alpha = 0.05$ oder $\alpha = 0.01$

Damit kontrolliert er vorab seinen

$$\alpha = \text{Fehler erster Art} = pr(\text{Test gegen } H_0 \mid H_0 \text{ richtig})$$

Test als Entscheidungsverfahren II

Opponent wählt eine spezifische Alternativhypothese H_A^*
(eine Punkthypothese in Ω_A , damit er die Verteilung
der Teststatistik unter seiner Erwartung H_A^* berechnen kann)

Darauf basierend bestimmt er die nötige Fallzahl.

Damit kontrolliert er vorab (bevor das Experiment durchgeführt wird)
seinen

$$\beta = \text{Fehler zweiter Art} = \text{pr}(\text{Test behält } H_0 \text{ bei} \mid H_A^* \text{ richtig})$$

bzw. seine

$$\text{Power} = 1 - \beta = \text{pr}(\text{Test gegen } H_0 \mid H_A^* \text{ richtig})$$

Diagnostische Tests und Statistische Tests

... in der Medizin

		Patient	
		krank	gesund
T e s t	+	😊 <i>Sensitivität</i>	falsch pos. ☹️
	–	falsch neg. ☹️	😊 <i>Spezifität</i>

... in der Statistik

		Nullhypothese („null Unterschied“)	
		falsch	richtig
T e s t	+	😊 <i>Power</i>	F. 1. Art <i>α</i> ☹️
	–	F. 2. Art <i>β</i> ☹️	😊 <i>Konfidenz- niveau</i>

Beispiel: Geschmackstest (vgl. Übung) I

Experiment: Proband soll aus $N=8$ Paaren von Geschmacksproben

einen bestimmten Geschmack herausfinden.

Sei p seine unbekannte Trefferwahrscheinlichkeit

Nullhypothese: $H_0 : p=0.5$ (Proband rät bloß!)

Spezifische Alternative: $H_A : p=0.9$ (Proband ist sich ziemlich sicher)

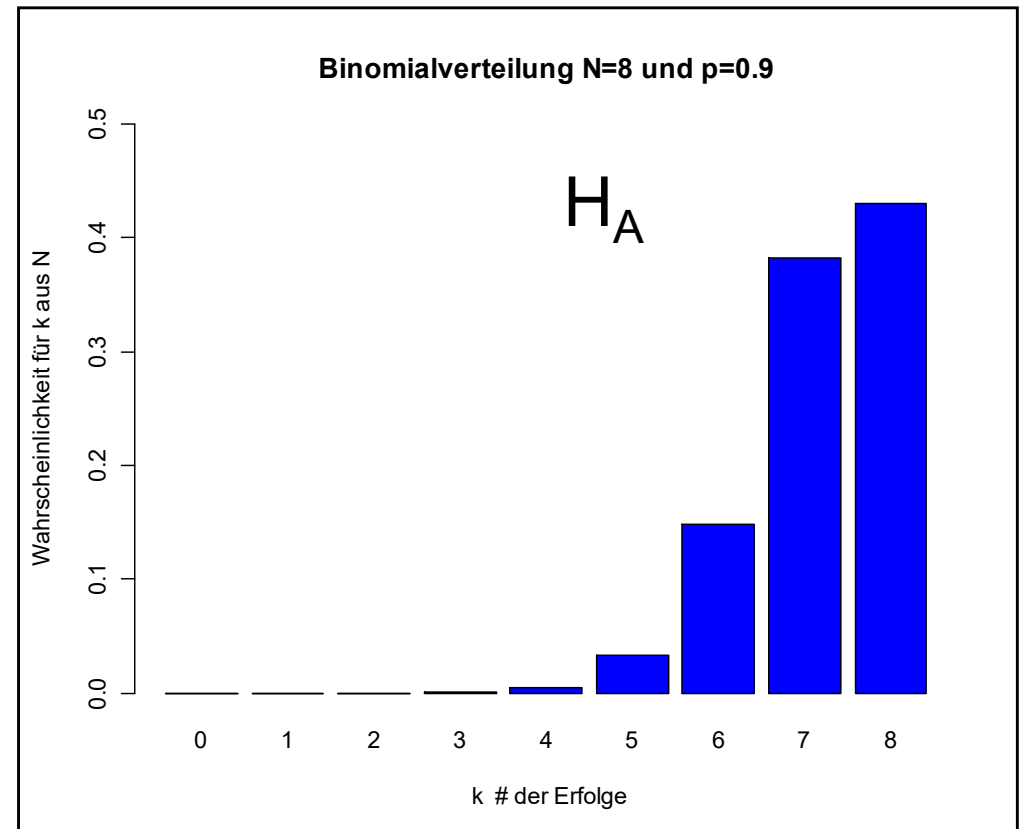
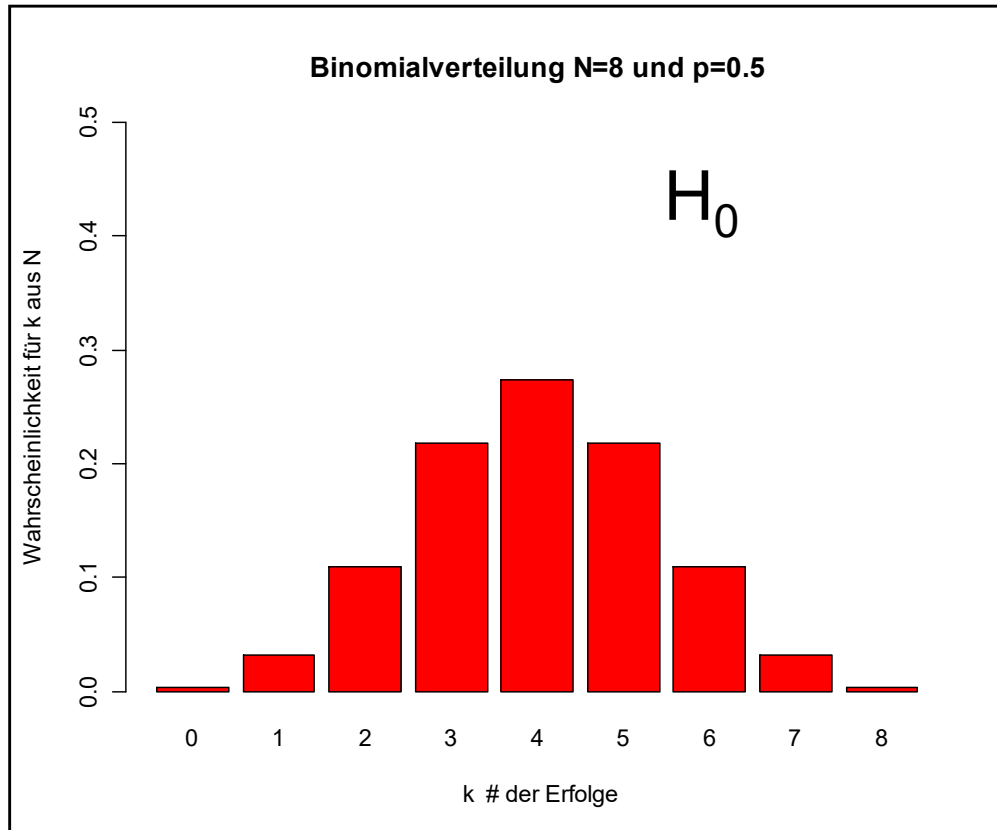
Entscheidungsregel: Sei k die Anzahl der Treffer

Lehne H_0 ab wenn $k=7$ oder 8 , sonst behalte H_0 bei.

Binomial-Verteilung

Wahrscheinlichkeit unter N Versuchen
k Erfolge zu beobachten, wenn die
Erfolgswahrscheinlichkeit p zugrunde liegt.

$$\text{pr}(X=k) = \binom{N}{k} p^k (1-p)^{N-k}$$



Beispiel: Geschmackstest (vgl. Übung) II

Der Versuch ist binomial mit $N=8$.

Fehler 1.Art:

$$\begin{aligned} \text{pr}(k \geq 7 | H_0 : p=0.5) &= \text{dbinom}(7, 8, 0.5) + \text{dbinom}(8, 8, 0.5) \approx \\ &0.03125 + 0.00390625 \approx 0.03515625 < 0.0352 \end{aligned}$$

Fehler 2.Art:

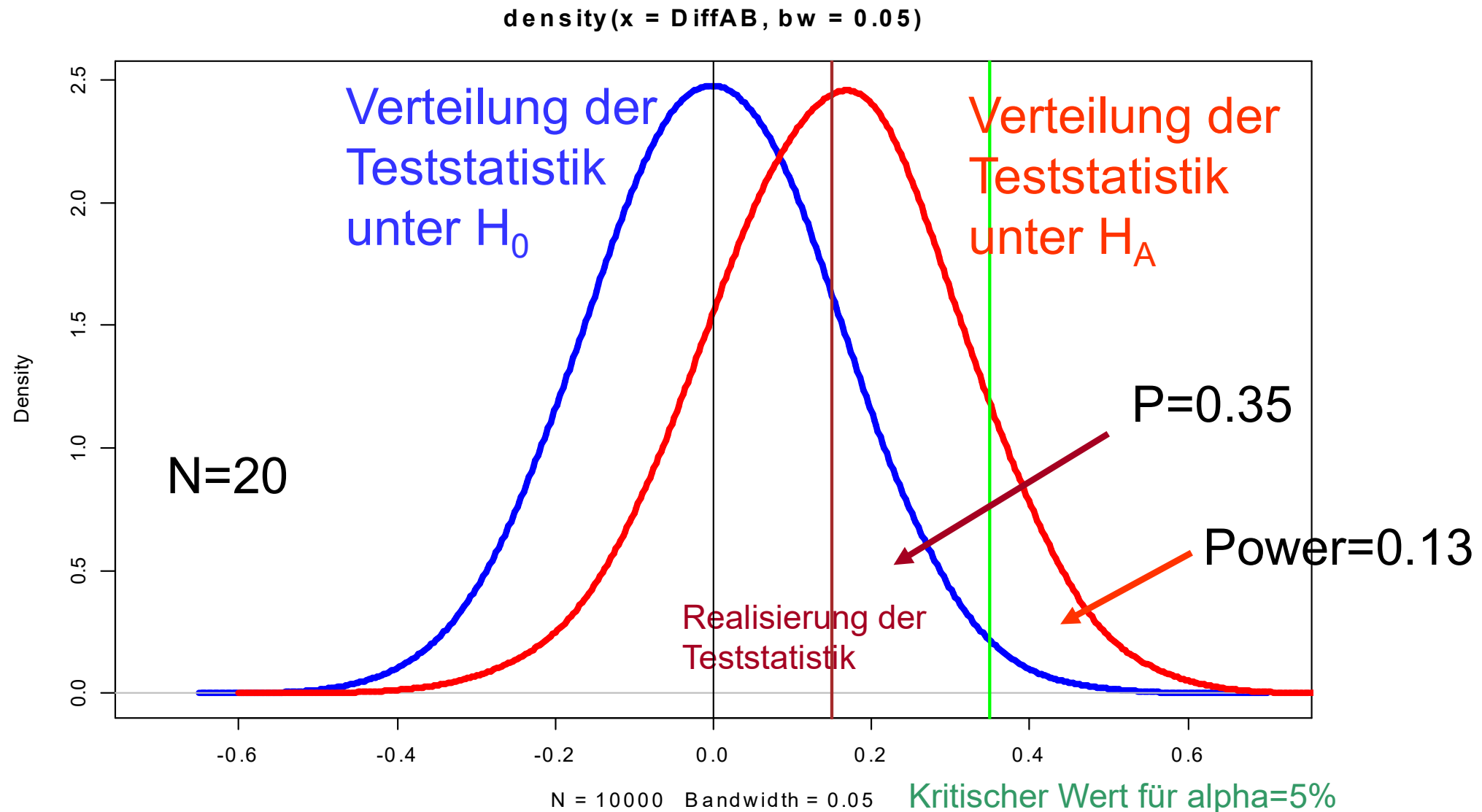
$$\text{pr}(k < 7 | H_A^* : p=0.9) = \text{sum}(\text{dbinom}(0:6, 8, 0.9)) < 0.187$$

Also Power $\approx 81.3\%$

Proponent und Opponent können sich auf den Versuch einlassen.

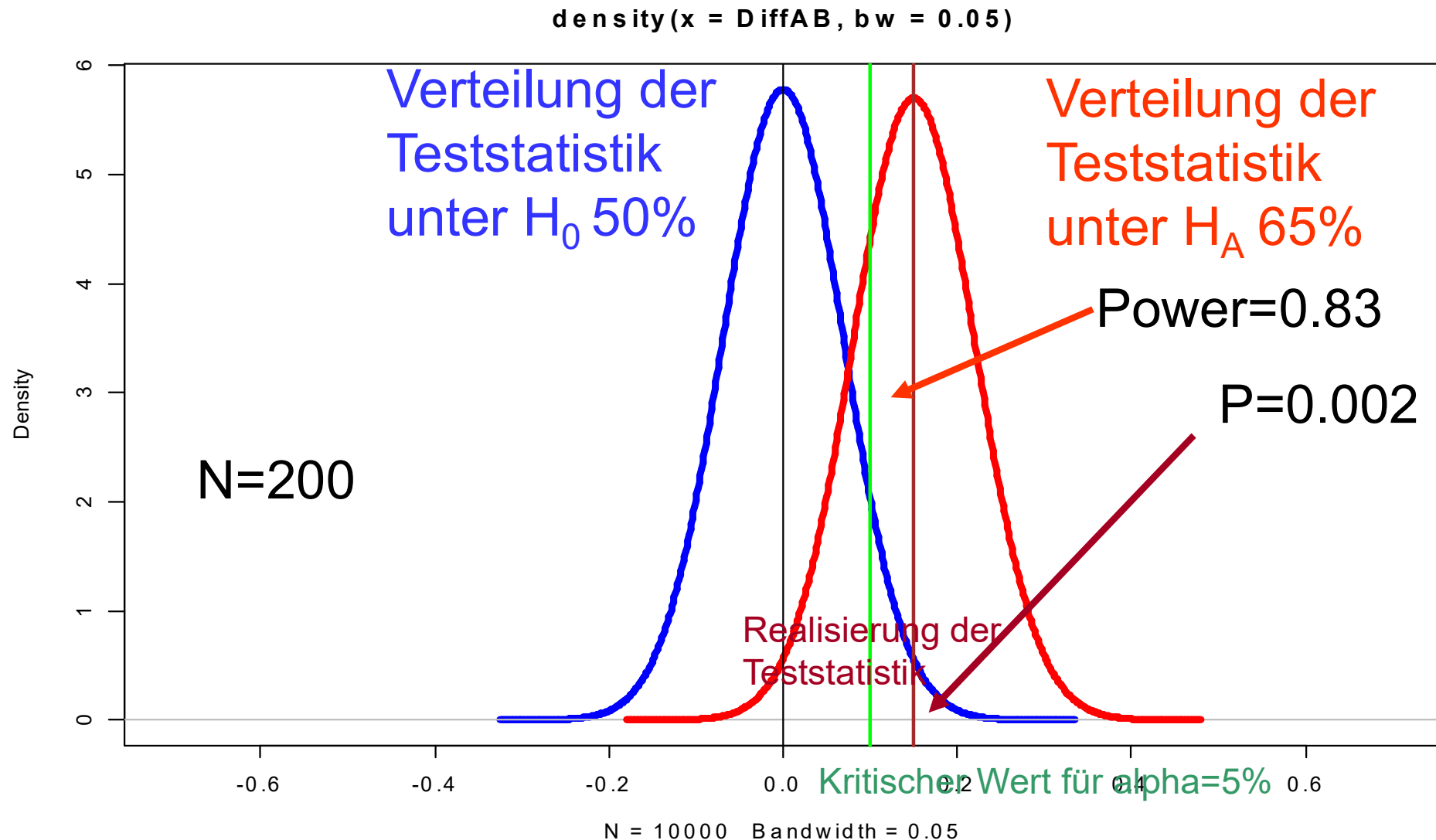
Verteilung der Teststatistik unter Null- und Alternativhypothese

Beispiel: 13/20 vs 10/20



Verteilung der Teststatistik unter Null- und Alternativhypothese

Beispiel: 130/200 vs 100/200



VORSICHT

- Ein **Test als Entscheidungsverfahren setzt voraus, dass Proponent und Opponent des Versuch gemeinsam geplant haben** (bzw. mit der Kontrolle der Fehler 1. und 2. Art zufrieden sein können).
- Ein Test als p-Wert Prozedure beurteilt nur, **ob die Evidenz aus den Daten ausreicht, die Nullhypothese in Frage zu stellen**. Eine Entscheidung fällt nicht automatisch.

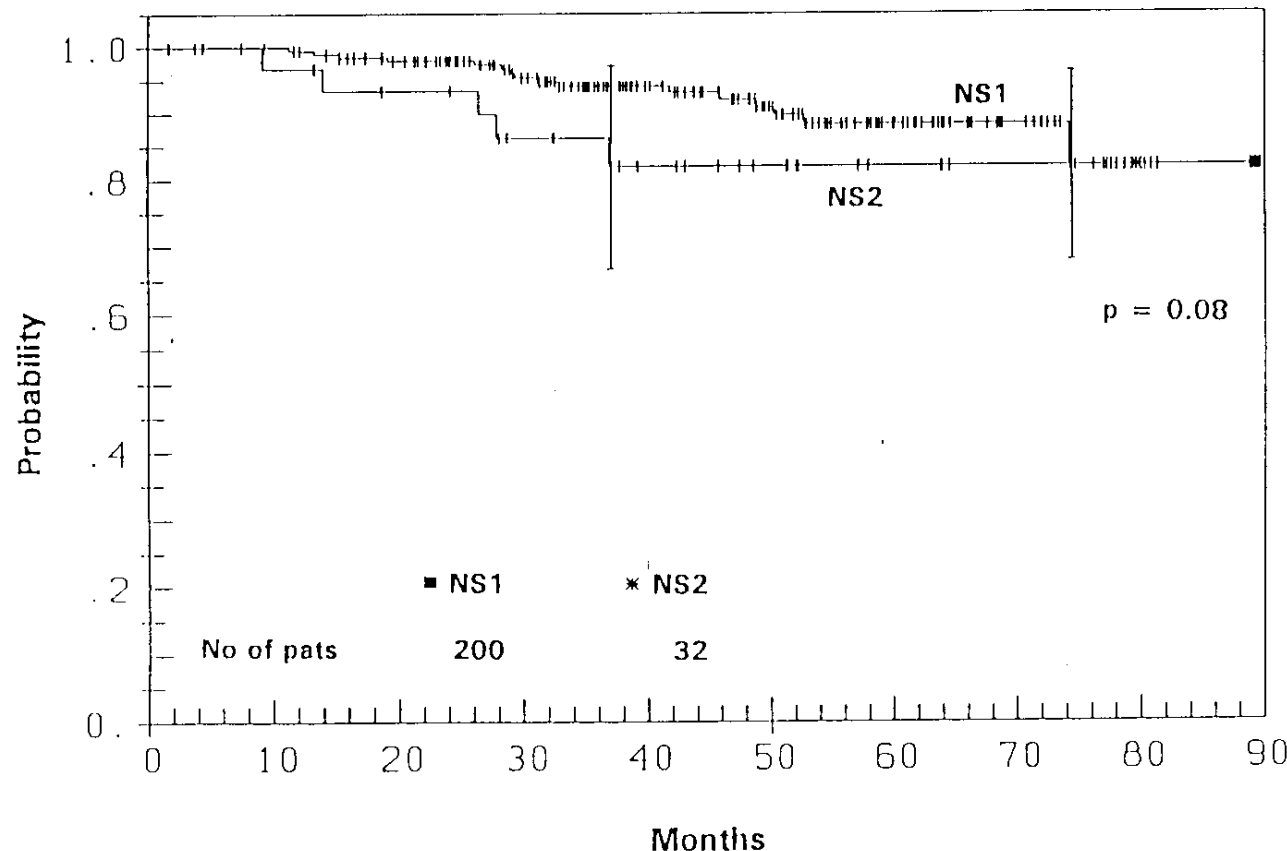
Logik des statistischen Tests

**„Unterschied nicht signifikant“
heißt nicht:
„Es gibt keinen Unterschied!“**

- Wenn Sie einen Zufallsbefund nicht ausschließen können, d.h. einen Skeptiker (noch) nicht überzeugen können, bedeutet dies nicht, dass die Nullhypothese wahr ist, oder Sie an sie glauben sollten.

HD1-3 trials of the German Hodgkin Study Group

Survival of patients with grade I and grade II nodular sclerosis subtype in HD1-3 trials of the German Hodgkin's Lymphoma Study Group



**Ziel: Publizierte Daten
BNLI (N > 3000)
zu reproduzieren.**

**Effekt geht in
erwartete Richtung,
ist aber „nicht signifikant“!**

Power ca. 43%!!

**2 Jahre später mit mehr
Patienten und Follow-up
„signifikant“.**

Logik des statistischen Tests II

- Wenn Sie einen quantitativ relevanten, aber nicht signifikanten Unterschied beobachten, wissen Sie nur, dass Ihre Fallzahl (!) nicht ausreicht, um bei der sich andeutenden Effektgröße einen Zufallsbefund auszuschließen.
- Denken Sie quantitativ, geben Sie einen Schätzwert für die Differenz mit einem Konfidenzintervall an.
- Bestimmen Sie die nötige Fallzahl bzw. analysieren Sie die statistische Power Ihrer Daten.

Logik des statistischen Tests III

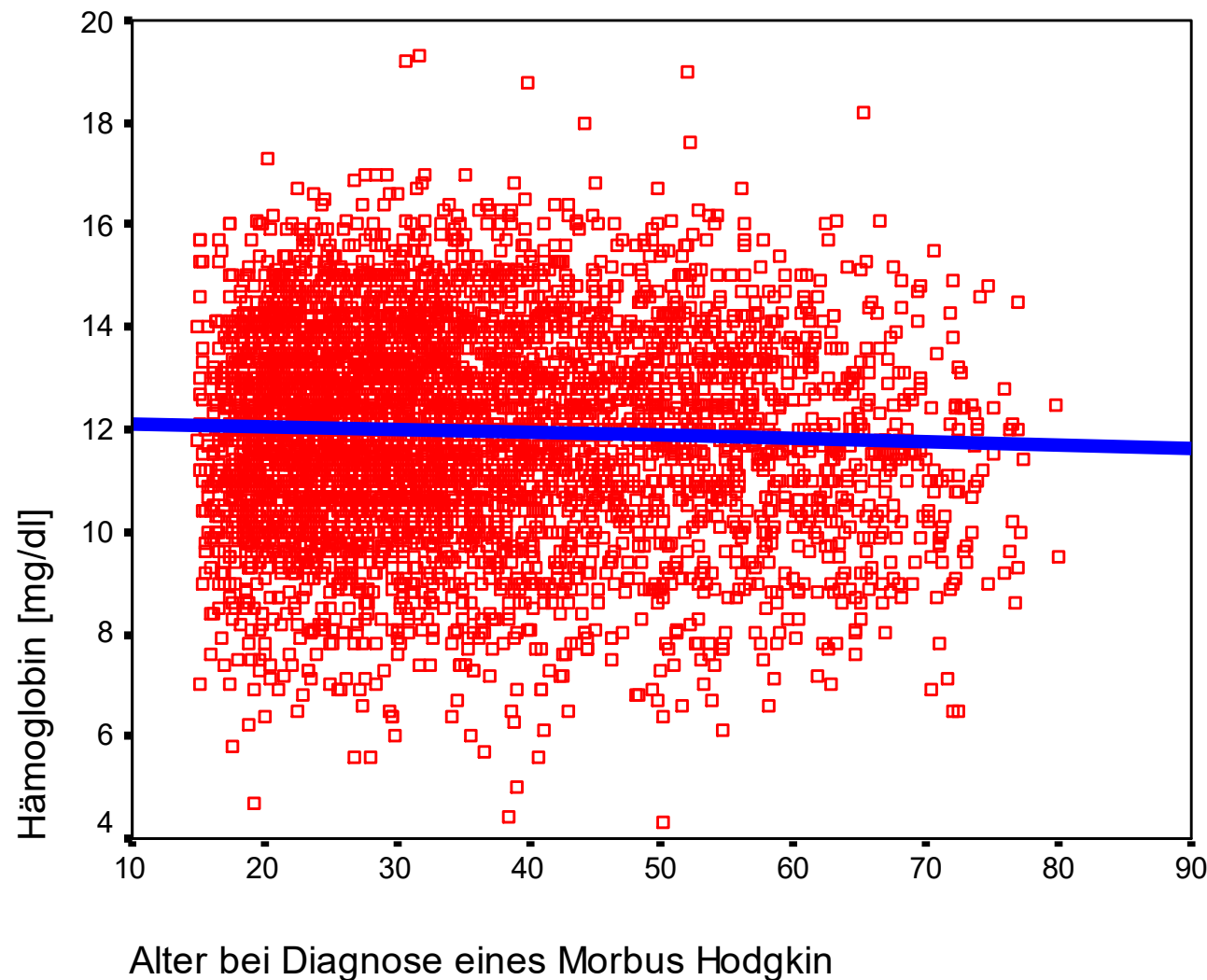
„Unterschied signifikant“

heißt nicht:

„Es gibt einen relevanten Unterschied!“

- p Werte sind kein Maß für die Effektgröße, da sie stark von der Fallzahl abhängen.
- Bei großen Fallzahlen können auch irrelevant kleine Unterschiede „signifikant“ sein.

Statistisch signifikant, aber quantitativ irrelevant



N=4606

Korrelations-
Koeffizient
 $R = -0.043$

$P = 0.004$
„hoch signifikanter
linearer
Zusammenhang..!“

Wenn man Schätzen kann, braucht man eigentlich kein Test

Kann man ein $(1-\alpha)$ Konfidenzintervall für den unbekannten Parameter θ_x berechnen, so gilt:

Liegen diejenigen Werte von θ , welche mit der Nullhypothese verträglich sind (Ω_0), alle außerhalb des $(1-\alpha)$ Konfidenzintervall, so entspricht dies einem Testresultat $p < \alpha$.

Zugleich hat man die quantitativen Gegebenheiten vor Augen.

Genauer besteht eine Dualität von Testen und Schätzen:

Dualität: Signifikanztests – Konfidenzintervalle I

Gegeben eine parametrische Familie von Verteilungen ($\theta \in \Omega$).

X eine mathematische Stichprobe dazu. $X=x$ eine Realisierung.

(1) Sei $\Psi(\theta_0)$ eine Familie $\theta_0 \in \Omega$ von Signifikanz-Tests zum Niveau α der Hypothesen $H(\theta_0): \theta = \theta_0$

$$\Psi(\theta_0)(X) = \begin{cases} 1: & \text{Ablehnen} \\ 0: & \text{Beibehalten} \end{cases}$$

Dann definiert $CI(x) = \{\theta \in \Omega \mid \Psi(\theta)(x) = 0\}$ eine $1-\alpha$ Konfidenzmenge.

Beweis: $\text{pr}(\theta \in CI(x)) = \text{pr}(\Psi(\theta)(x) = 0) = 1-\alpha$

Dies kann man benutzen, um Konfidenzintervallverfahren zu konstruieren!

Dualität von Testen und Konfidenzintervallschätzung

Tests von
Punkthypothesen
lehnt H_0 ab

Tests von
Punkthypothesen
lehnt H_0 nicht ab

Tests von
Punkthypothesen
lehnt H_0 ab

Konfidenzintervall
besteht aus allen
 θ in Ω , die durch
ihren Test nicht
abgelehnt werden
und also mit den
Daten verträglich
sind.

θ in Ω

Dualität: Signifikanztests – Konfidenzintervalle II

(2) Sei CI ein Verfahren zur Bestimmung von $1-\alpha$ Konfidenzmengen, d.h.

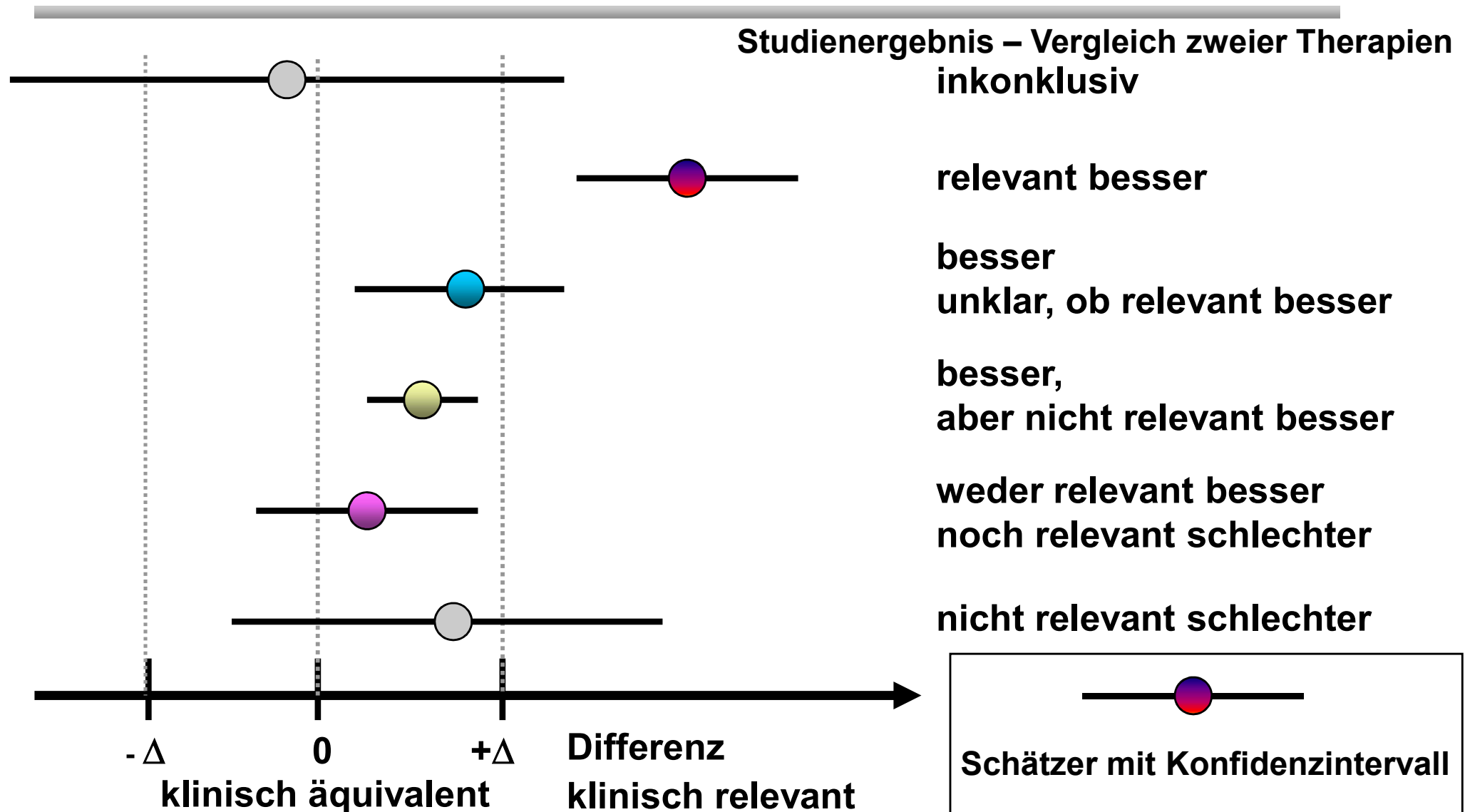
$$\text{pr}(\theta \in CI(x)) = 1-\alpha$$

Dann definiert

$$\Psi(\theta)(x) = \begin{cases} 1 & \text{falls } \theta \notin CI(x) \\ 0 & \text{falls } \theta \in CI(x) \end{cases}$$

eine Familie von Niveau α Signifikanztests.

Punktschätzer & Konfidenzintervall ≈ statistischer Test + quantitative Information



Bonusmaterial

Test als Entscheidungsprozedur

Entscheidungstheoretische Kritik

- Mit dem Test als Entscheidungsprozedur wird ein allgemeines statistisches Entscheidungsverfahren als „rational“ propagiert.
- Die Rationalität der Entscheidung hängt bei konkreten Entscheidungen von der Bewertung der Handlungsfolgen (Verlustfunktion) ab.
- Bei geplanten Versuchen können α und β nach Analyse der Handlungsfolgen sachgemäß auch eventuell unter Berücksichtigung der Kosten der Fehler gewählt und dann die Fallzahl bestimmt werden.
- Werden Tests jedoch schematisch als Entscheidungsprozeduren angewandt, **wie dies in der Medizin leider endemisch ist**, ist die Rationalität nicht gesichert.

Test als Entscheidungsprozedur

Evidenztheoretische Kritik

Betrachte einen Binomialen Versuch $\text{Binom}(k, n, \theta)$

Teste $H_0: \theta = 1/4$ gegen $H_1: \theta = 3/4$ (θ bezeichnet die "wahre" zugrundeliegende Wahrscheinlichkeit)

Wir beobachten k erfolge aus n und berechnen die Likelihoodratio,

d.h. in diesem Fall das Verhältnis der Wahrscheinlichkeit der Beobachtung unter den jeweiligen Hypothesen.

$$\text{LR}(k) = \frac{\text{pr}(k|H_1)}{\text{pr}(k|H_0)} = \frac{\binom{n}{k} * 0.75^k * 0.25^{n-k}}{\binom{n}{k} * 0.25^k * 0.75^{n-k}} = 3^{2k-n}$$

!

Sei $n = 30$.

Da $\sum_{k=12}^{30} \binom{30}{k} * 0.25^k * 0.75^{30-k} = 0.05$ lehnt der N-P-Test H_0 ab und wählt H_1 , wenn $k \geq 12$

Aber $\text{LR}(12) = 1/729$ zugunsten von H_0 !!!

erst $\text{LR}(15) = 1$ ist neutral zwischen den Hypothesen und $\text{LR}(16) = 9$ spricht moderat für H_1 .

Bei ungenügender Fallzahl kann wegen der Asymmetrie von Fehler erste und zweiter Art eine Hypothese angenommen werden, welche die Daten dramatisch unwahrscheinlicher erscheinen lässt als die verworfene Gegenhypothese.

Dieses Phänomen tritt nicht auf, wenn der Parameterraum Ω adäquat gewählt wird.

(Hier: Statt $\Omega = \{0.25, 0.75\}$ nehme $\Omega = [0, 1]$ und teste $\Omega_0 = [0, 0.25]$ gegen $\Omega_A = (0.25, 1]$.)

Test als Entscheidungsverfahren

Bayesianische Kritik

Tests als Entscheidungsverfahren kontrollieren
die Fehler erster und zweiter Art.

Dies suggeriert, es werde die Wahrscheinlichkeit, einen Fehler zu machen, kontrolliert.
Kontrolliert werden aber nur bedingte Wahrscheinlichkeiten.

Typische Vorgabewerte bei klinischen Studien sind $\alpha = 0.05$ und $\beta = 0.2$.

Angenommen wir halten beide Hypothesen für gleichwahrscheinlich,
d.h. unser Prior ist $\text{pr}(H_0) = 0.5$ und $\text{pr}(H_1) = 0.5$

Dann gilt zum Beispiel:

$$\begin{aligned}\text{pr}(\text{Fehler}) &= \text{pr}(H_0 \text{ ablehnen} \mid H_0) \cdot \text{pr}(H_0) + \text{pr}(H_1 \text{ ablehnen} \mid H_1) \cdot \text{pr}(H_1) \\ &= \alpha * 0.5 + \beta * 0.5 = 0.125\end{aligned}$$

Die eigentlich interessierende Fehlerwahrscheinlichkeit $\text{pr}(\text{Fehler})$ hängt
vom Prior ab und liegt $\max(\alpha, \beta) = \text{pr}(\text{Fehler}) = \min(\alpha, \beta)$

Test als Entscheidungsverfahren

Kritik der „Konditionalisten“

EXAMPLE 13. Suppose X is 1, 2, or 3 and θ is 0 or 1, with X having the following probability density in each case:

	x		
	1	2	3
$f(x 0)$	0.005	0.005	0.99
$f(x 1)$	0.0051	0.9849	0.01

The classical most powerful test of $H_0: \theta = 0$ versus $H_1: \theta = 1$, at level $\alpha = 0.01$, concludes H_1 when $X = 1$ or 2, and this test also has a Type II error probability of 0.01. Hence, a standard frequentist, upon observing $x = 1$, would report that the decision is H_1 and that the test had error probabilities of 0.01. This certainly *gives the impression* that one can place a great deal of confidence in the conclusion, but is this the case? Conditional reasoning shows that the answer is sometimes no! When $x = 1$ is observed, the likelihood ratio between $\theta = 0$ and $\theta = 1$ is $(0.005)/(0.0051)$ which is very close to one. To a conditionalist (and to most other statisticians also), a likelihood ratio near one means that the data does very little to distinguish between $\theta = 0$ and $\theta = 1$. Hence the conditional “confidence” in the decision to conclude H_1 , when $x = 1$ is observed, would be only about 50%. (Of course $x = 1$ is unlikely to occur, but, when it does, should not a sensible answer be given?)

- Regeln, die sich an Verfahrenseigenschaften orientieren, können irreführen, wenn zufällig eine unwahrscheinliche, aber bzgl. ihres Informationswerts extreme Stichprobe beobachtet wird.

Berger 1985