
Beurteilung von Hypothesen in drei statistischen Paradigmata I

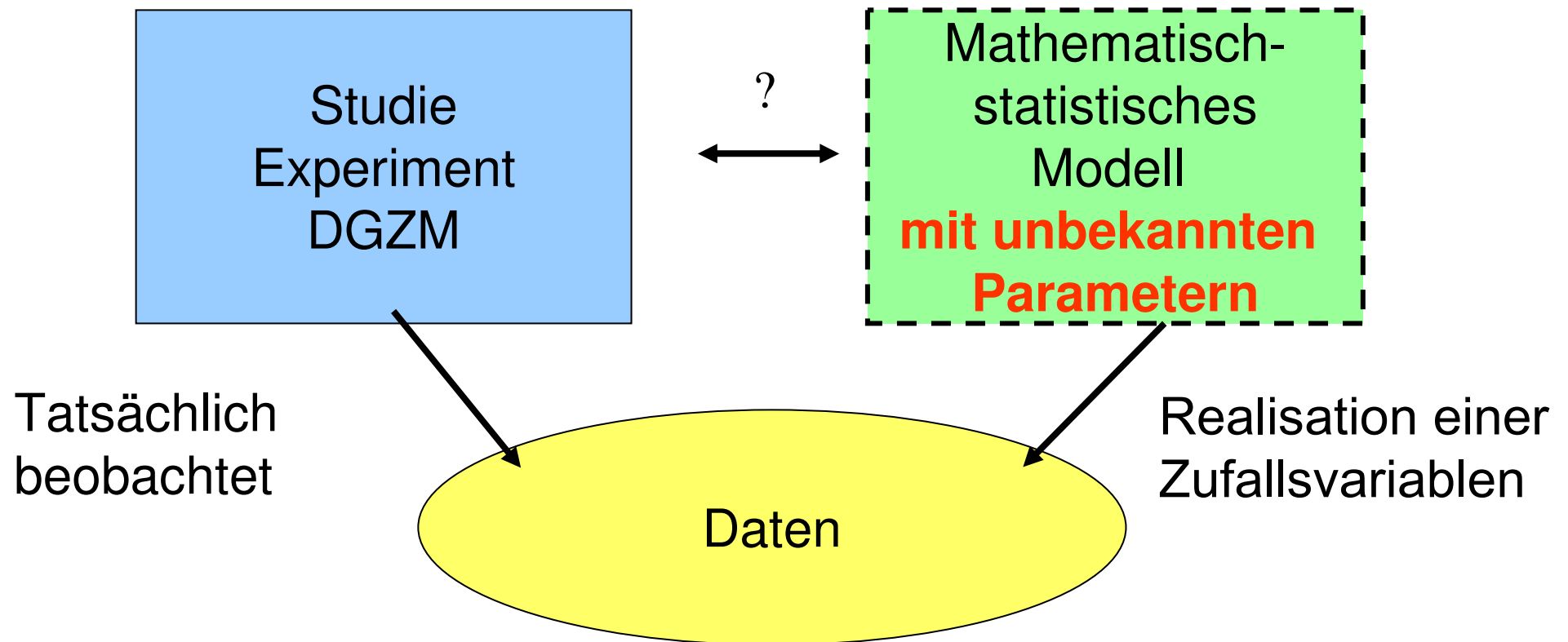
Dr. Dirk Hasenclever
IMISE, Leipzig

Hasenclever@IMISE.uni-Leipzig.de

Zielsetzung

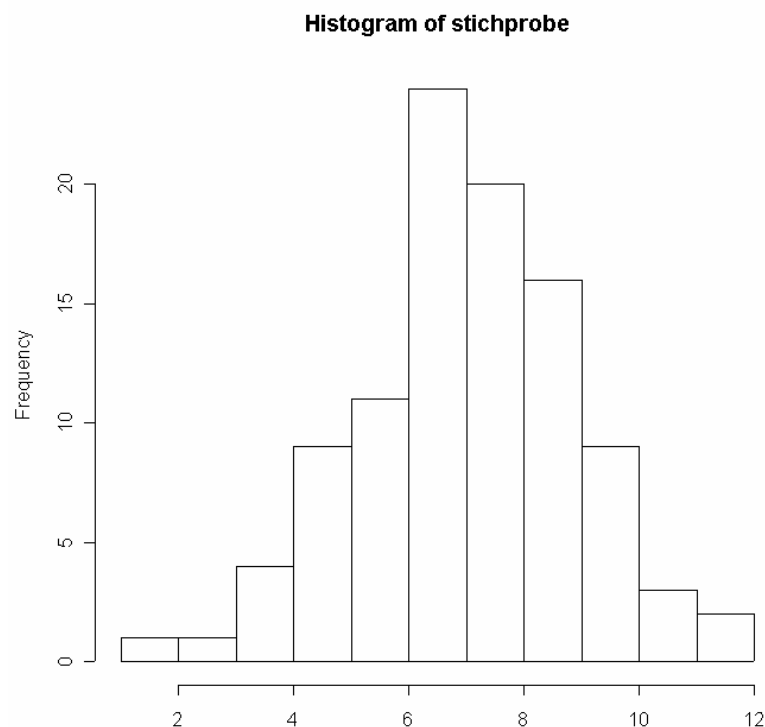
- Grundprobleme der Statistik:
 - Schätzen
 - Beurteilen von statistischen Hypothesen
- Paradigmatische Unterschiede im Vorgehen zwischen:
 - Frequentisten
 - Likelihoodians
 - Bayesianern

Grundprobleme der Statistik



- (1) **Schätze** unbekannte Parameter und beurteile die **Genauigkeit**
- (2) **Beurteile Hypothesen** über unbekannte Parameter

Beispiel: Stichprobe aus Normalverteilung



```
>stichprobe<-rnorm(100,7,2)
>mean(stichprobe)
[1] 7.013944
> sqrt(var(stichprobe))
[1] 1.891915
```

Stichprobe $N=100$ aus $N(\mu, \sigma^2)$
 μ, σ unbekannte Parameter

Schätzer:

$$\hat{\mu} = \frac{1}{N} \sum_{i=1}^N x_i; \quad \hat{\sigma}^2 = \frac{1}{N-1} \sum_{i=1}^N (x_i - \hat{\mu})^2$$

Beurteile : $H_0 : \mu=5$

Vergleiche bzgl. Evidenz

$H_0 : \mu=5$ mit $H_1 : \mu=10$

Paradigmatische Leitfragen

- **Was soll ich glauben**, angesichts der Daten?
(-> Vorwissen berücksichtigen!) „**Bayesianer**“
 - **Wie entscheide ich** zwischen Handlungsoptionen,
aufgrund der Daten? (-> Entscheidungsregel.
Berücksichtige Verlust-Funktion über mögliche
Handlungsfolgen!) „**Frequentisten**“
 - **Was sagen die Daten** bzgl. Hypothesen A und B
über die unbekannten Parameter des Modells?
(-> relative Evidenz, argumentative Kraft)
„**Likelihoodians**“
- (Diagnostikbeispiel: Antworten weitgehend unabhängig!!)**

Frequentisten A: Neyman-Pearson-Test I

Die Verteilung P_θ der Zufallsvariablen X liege in einer Klasse $\{P_\theta \mid \theta \in \Omega\}$ von Verteilungen.
Betrachte eine disjunkte Zerlegung $\Omega = \Omega_0 \dot{\cup} \Omega_1$ des Parameterraums.

Aufgabe: Entscheidungsregel zwischen den Hypothesen $H_0 : \theta \in \Omega_0$ und $H_1 : \theta \in \Omega_1$
Zum Beispiel bei Schraubenproduktion: H_0 : Schrauben im Toleranzbereich. H_1 : Ausschuss.

Allgemeine Form einer Entscheidungsregel:

Sei $S_x = \{x \mid x \text{ mögliche Realisierung von } X\}$ der Samplespace der Zufallsvariable X .

Im Beispiel: Mögliche Realisierungen von Stichproben der Größe $N = 100$.

Bilde disjunkte Zerlegung: $S_x = S_0 \dot{\cup} S_1$ und entscheide:

Falls $x \in S_0 \rightarrow H_0$ annehmen! Falls $x \in S_1 \rightarrow H_1$ annehmen!

Frequentisten A: Neyman-Pearson-Test II

Viele Aufteilungen des Samplespace sind möglich. Welche sind gut?

Grundgedanke frequentistischer Verfahren :

Kontrolliere die Güte einer Entscheidungsregel durch Verfahrenseigenschaften
bei wiederholter Anwendung!

Entscheidungsregel - Vorschlag von Neyman - Pearson :

Zeichne eine der Hypothesen aus und kontrolliere die Wahrscheinlichkeit
eines "Fehlers erster Art" bei wiederholter Anwendung des Verfahrens :

$$\text{pr}(x \in S_1 \mid \theta \in \Omega_0) \leq \alpha \text{ für alle } \theta \in \Omega_0$$

Unter den Entscheidungsregeln, die dieser Bedingung genügen, wähle diejenige(n),
welche - bei gegebener Fallzahl -

die Power $\text{pr}(x \in S_1 \mid \theta \in \Omega_1) = 1 - \beta$ möglichst für alle $\theta \in \Omega_1$ gleichmäßig maximiert.

Frequentisten A: Neyman-Pearson-Test III

Neyman-Pearson-Testtheorie:

Finde solche optimalen Entscheidungsregeln (=Tests) für diverse Situationen.

Nicht immer existiert eindeutig bestes Verfahren. Jedoch gilt:

Lemma von Neyman-Pearson:

Bester Ablehnungsbereich beim Test von **Punkthypothesen** wird durch

$$\text{Likelihoodratio} = \frac{\text{Likelihood der Daten unter } H_0}{\text{Likelihood der Daten unter } H_1} > C_\alpha$$

definiert.

(Maximale Power bei gegebener „Irrtumswahrscheinlichkeit“)

Einfaches Beispiel: Gauss-Test

Annahme : $x_1, \dots, x_N$ realisierte Stichprobe zu Zufallsvariablen $X_1, \dots, X_N$ unabhängig und identisch normalverteilt $N(\mu, 1)$ mit bekannter Varianz 1.

Nullhypothese : $\mu = 0$

Alternativhypothese : $\mu \neq 0$

Wähle Signifikanzniveau $\alpha = 0.05$

Die Teststatistik : $G = \frac{\bar{x}}{\sqrt{N}}$ ist eine Zufallsvariable

und unter der Nullhypothese standardnormalverteilt $N(0, 1)$.

(Nachrechnen mit Rechenregel für Normalverteilungen!)

Wähle Ablehnungsbereich: $|G| > 1.96$

Entscheidungsregel:

Berechne Wert der Teststatistik g aus der Stichprobe.

Lehne Nullhypothese ab, falls $|g| > 1.96$

Frequentisten A: Neyman-Pearson-Test IV

	Entscheidung H0	Entscheidung H1
Wahrheit H0	OK $> 1 - \alpha$	Fehler erster Art $< \alpha$ qua Verfahren
Wahrheit H1	Fehler zweiter Art β	OK (Power) $1 - \beta$ Fallzahlabhängig

Bei der Versuchplanung kann nach Wahl des Testverfahrens das Signifikanzniveau α und die gewünschte Power $1 - \beta$ (bzgl. einer konkreten Alternative) vorgegeben und daraus die nötige Fallzahl errechnet werden.
„experimentum crucis“

Frequentisten A: Neyman-Pearson-Test V

Entscheidungstheoretische Kritik

- Mit diesem Newman-Pearson Ansatz 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 gewählt und dann die Fallzahl bestimmt werden.
- Werden N-P-Tests jedoch schematisch angewandt, **wie dies in der Medizin leider endemisch ist**, ist die Rationalität nicht gesichert.

Frequentisten A: Neyman-Pearson-Test VII

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.**

Frequentisten A: Neyman-Pearson-Test VI

Bayesianische Kritik

- Newman-Pearson-Test 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 :

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

- Die eigentlich interessierende Fehlerwahrscheinlichkeit f hängt vom Prior ab und liegt $\max(\alpha, \beta) \geq f \geq \min(\alpha, \beta)$

Frequentisten A: Neyman-Pearson-Test VIII

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:

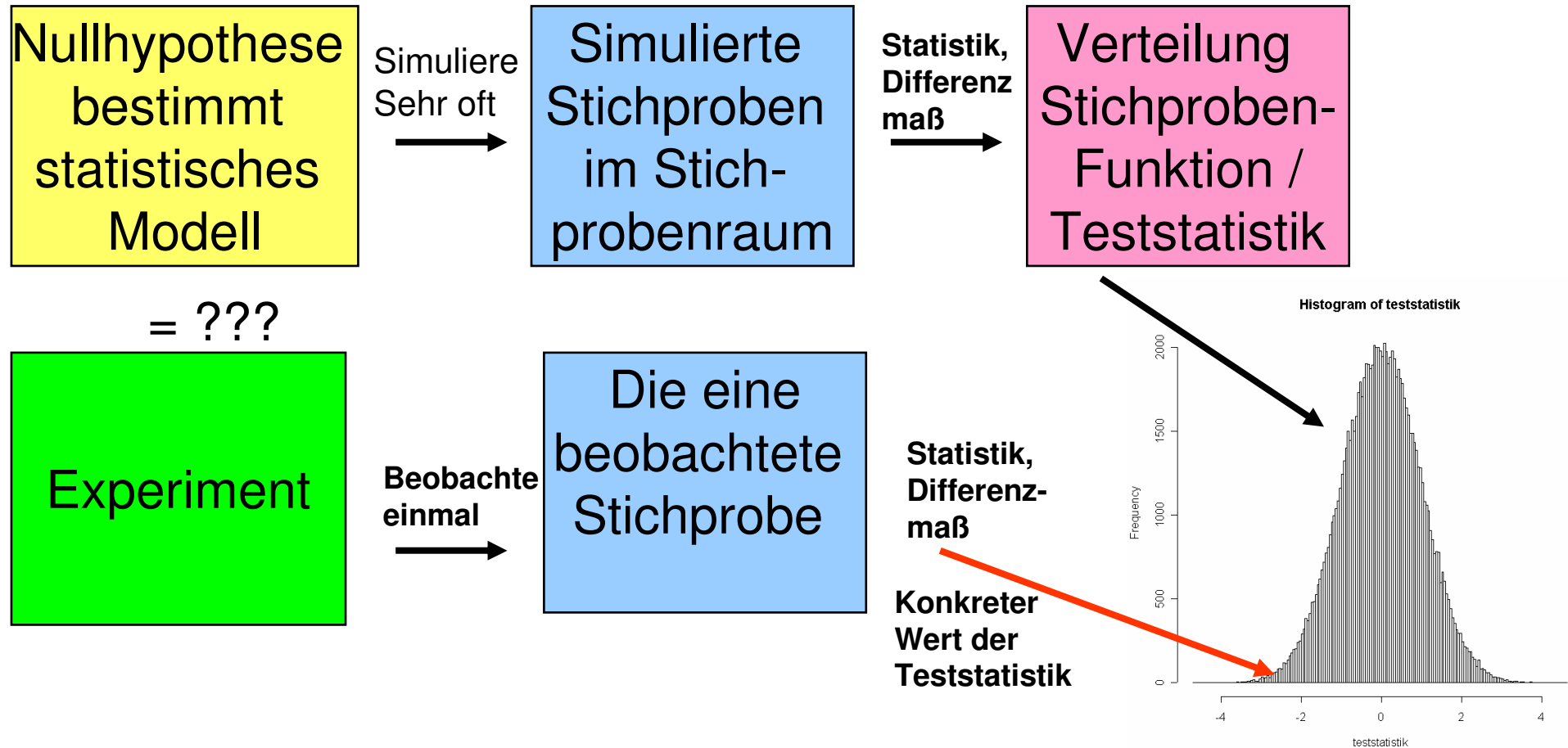
	x		
	1	2	3
$f(x 0)$	0.005	0.005	0.99
$f(x 1)$	0.0051	0.9849	0.01

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

Frequentisten B: P-Wert Prozedur I



Liegt der Testwert extremal?

Neyman-Pearson gegen P-Wert Prozedur

- Entscheidungsverfahren zwischen zwei Hypothesen
- Entscheide aufgrund der Daten!
- Vorgegebene Verfahrensparameter
 - Signifikanzniveau
 - Power
- Kontrolle Fehler erster Art
- Überprüfung einer wissenschaftlichen Theorie (Nullhypothese)
- Daten mit Theorie verträglich?
- p-Wert = Wahrscheinlichkeit unter H_0 einen so extremen oder extremeren Testwert zu beobachten
- p-Wert als Argument gegen die Nullhypothese.

Logik des statistischen Tests I

Beobachtet: Unterschied zwischen zwei Therapien $A > B$

Skeptiker: **Bloßer Zufallsbefund!**

Angenommen: Der Skeptiker habe Recht
Kein realer Unterschied (Nullhypothese)

Frage: Wie gut lässt sich die Beobachtung $A > B$
als Zufallsbefund wegerklären?

Logik des statistischen Tests II

Maß: **Wahrscheinlichkeit p , einen solchen Unterschied oder einen noch extremeren rein zufällig** (d.h. bei Gültigkeit der Nullhypothese) **zu beobachten.**

Statistischer Schluss:

Falls p klein ist

- **entweder etwas sehr unwahrscheinliches passiert**
- **oder die Nullhypothese ist falsch.**

Je kleiner p , desto unplausibler der Einwand des Skeptikers

Konvention: **$p < 0.05$ („signifikant“)**
-> Verwerfe Einwand des Skeptikers

Schätzen Sie p !

Therapie A				Therapie B	
E1=	13			E2=	10
N1=	20			N2=	20
P1=	0,65			P2=	0,5

Auflösung: Schätzen Sie p !

Therapie A				Therapie B	
E1=	13			E2=	10
N1=	20			N2=	20
P1=	0,65			P2=	0,5
se1=	0,10665			se2=	0,1118
		(p1-p2)=	0,15		
		se12=	0,1545		
		95%CI:	-0,153	0,4529	
		Z=	0,9708		
		"p"=	0,3317		

Schätzen Sie p !

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

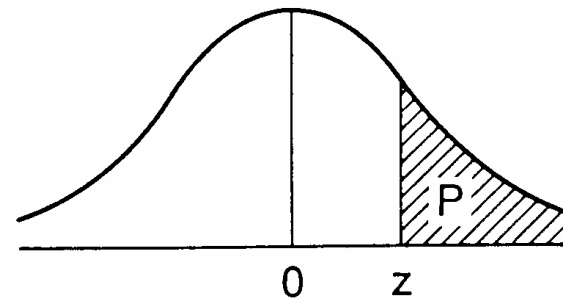
Auflösung: Schätzen Sie p !

Therapie A				Therapie B	
E1=	130			E2=	100
N1=	200			N2=	200
P1=	0,65			P2=	0,5
se1=	0,03373			se2=	0,0354
		(p1-p2)=	0,15		
		se12=	0,0489		
		95%CI:	0,0542	0,2458	
		Z=	3,0699		
		"p"=	0,0021		

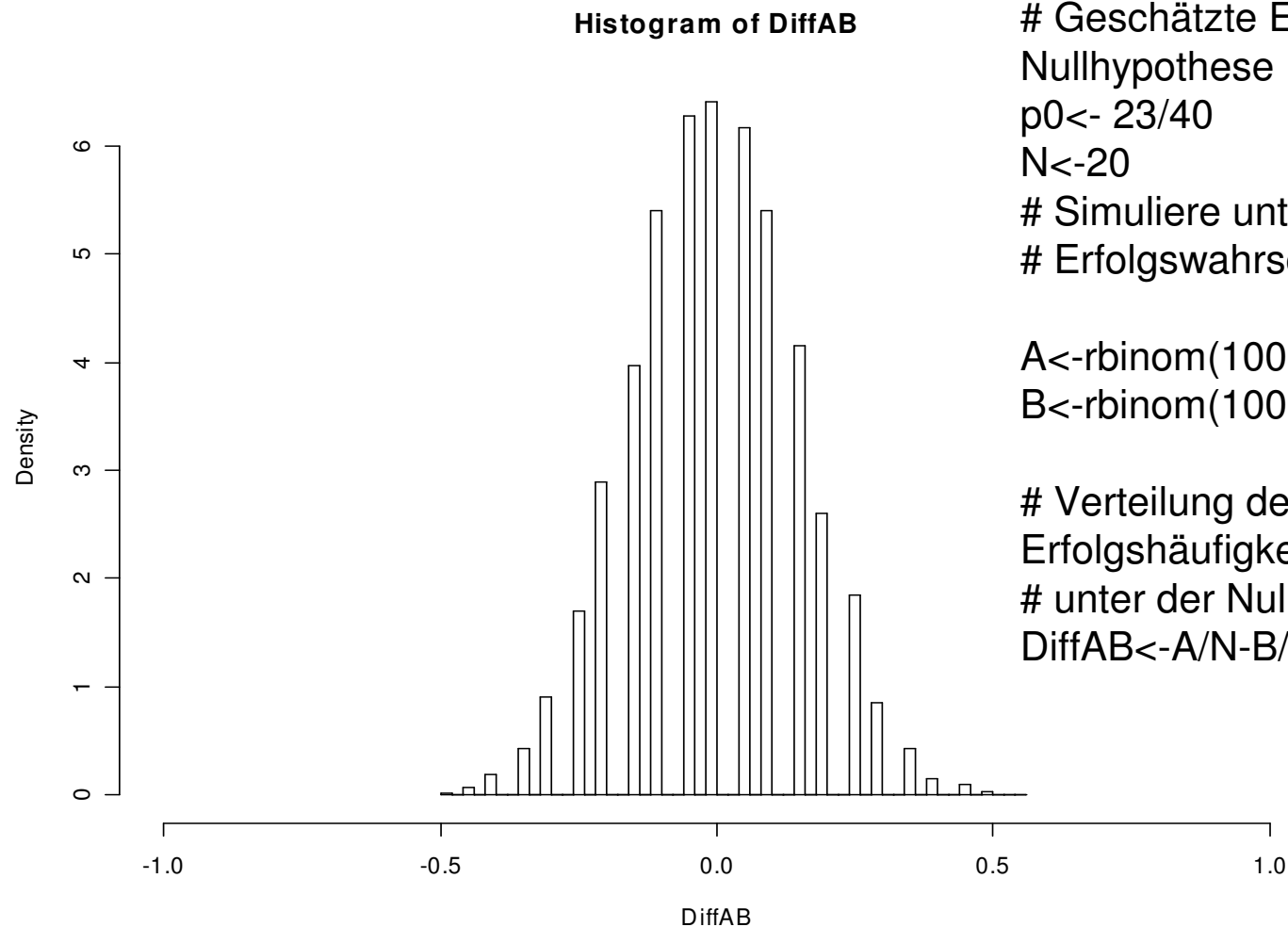
Logik des statistischen Tests III

Konstruktion eines Tests:

- Wähle geeignetes Differenzmaß (Teststatistik)
- Bestimme deren Zufallsverteilung
unter der Nullhypothese
- Lege fest, was extrem heißen soll
(z.b. einseitig vs. zweiseitig)
- Mit beobachtetem Wert Z
der Teststatistik lese p-Wert
an der Verteilung ab.



Simulation der Verteilung der Differenzen unter der Nullhypothese



Geschätzte Erfolgsrate unter der Nullhypothese

$p_0 <- 23/40$

$N <- 20$

Simuliere unter der Nullhypothese p_0

Erfolgswahrscheinlichkeit in A & B

$A <- \text{rbinom}(10000, N, p_0)$

$B <- \text{rbinom}(10000, N, p_0)$

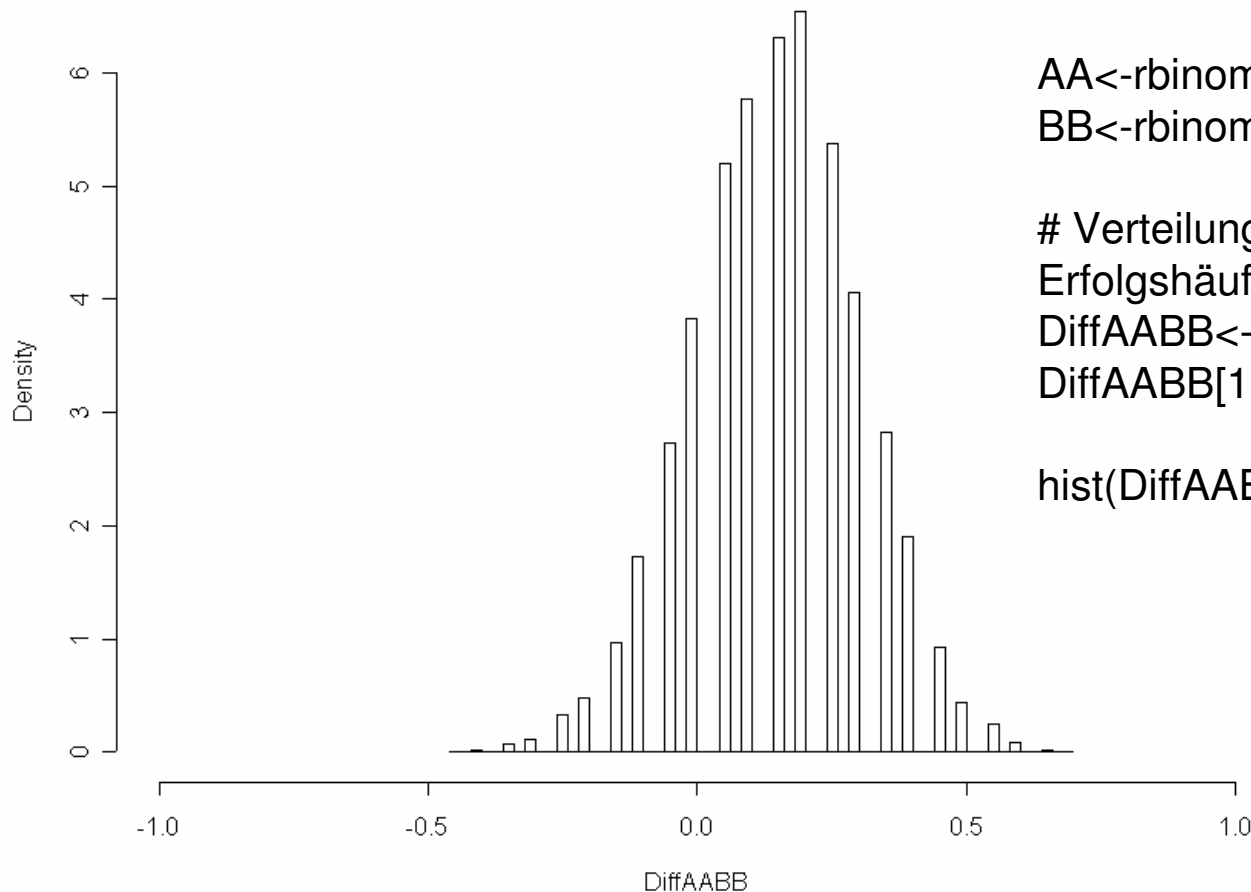
Verteilung der beobachteten Differenzen in Erfolgshäufigkeiten

unter der Nullhypothese

$\text{DiffAB} <- A/N - B/N$

Simulation der Verteilung der Differenzen unter der Alternative: Unterschied= 15%

Histogram of DiffAABB



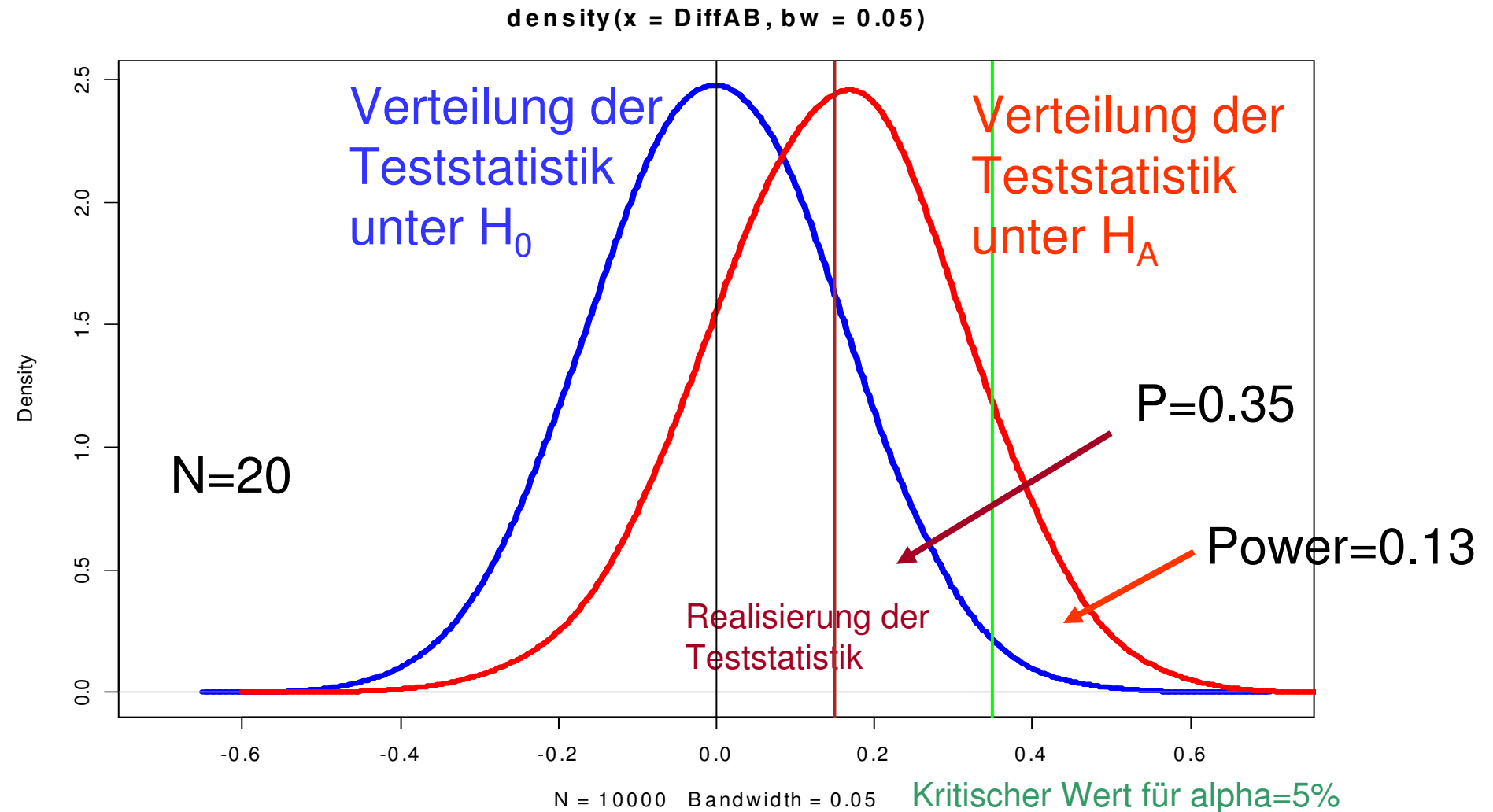
Angenommen Alternativ-Hypothese gilt:
Therapie A 65% Therapie B 50%

```
AA<-rbinom(10000,N,0.65)  
BB<-rbinom(10000,N,0.50)
```

```
# Verteilung der beobachteten Differenzen in  
Erfolgshäufigkeiten  
DiffAABB<-AA/N-BB/N  
DiffAABB[1:100]
```

```
hist(DiffAABB,freq=FALSE,nclass=42,xlim=c(-1,1))
```

Realisierte Teststatistik in Bezug auf die Verteilung der Teststatistik unter Null- und Alternativhypothese



p-Wert Bestimmung in Simulation

Bestimme zweiseitigen p-Wert:

pWert<-sum(abs(DiffAB)>= 0.15)/10000

"P-Wert = Wahrscheinlichkeit unter H0 eine solche"

"oder eine extremere Differenz zu beobachten:"

Bestimmung des kritischen Werts in Simulation (N=20)

```
# Suche Kritischen Wert zu alpha=0.05
> sort(abs(DiffAB),decreasing=TRUE)[1:700]
[1] 0.60 0.55 0.55 0.55 0.50 0.50 0.50 0.50 0.50
...
[298] 0.35 0.35 0.35 0.30 0.30 0.30 0.30 0.30 0.30
[307] 0.30 0.30 0.30 0.30 0.30 0.30 0.30 0.30 0.30
[316] 0.30 0.30 0.30 0.30 0.30 0.30 0.30 0.30 0.30
[325] 0.30 0.30 0.30 0.30 0.30 0.30 0.30 0.30 0.30
[334] 0.30 0.30 0.30 0.30 0.30 0.30 0.30 0.30 0.30
...
[496] 0.30 0.30 0.30 0.30 0.30 0.30 0.30 0.30 0.30
[505] 0.30 0.30 0.30 0.30 0.30 0.30 0.30 0.30 0.30
```

Problem weil diskrete Verteilung: Alpha nicht voll ausschöpfbar

**Pr (Teststatistik $\geq$ kritischer Wert) $\leq$ 0.05 ->
kritischerWert=0.30**

Bestimmung der Power in Simulation (N=20)

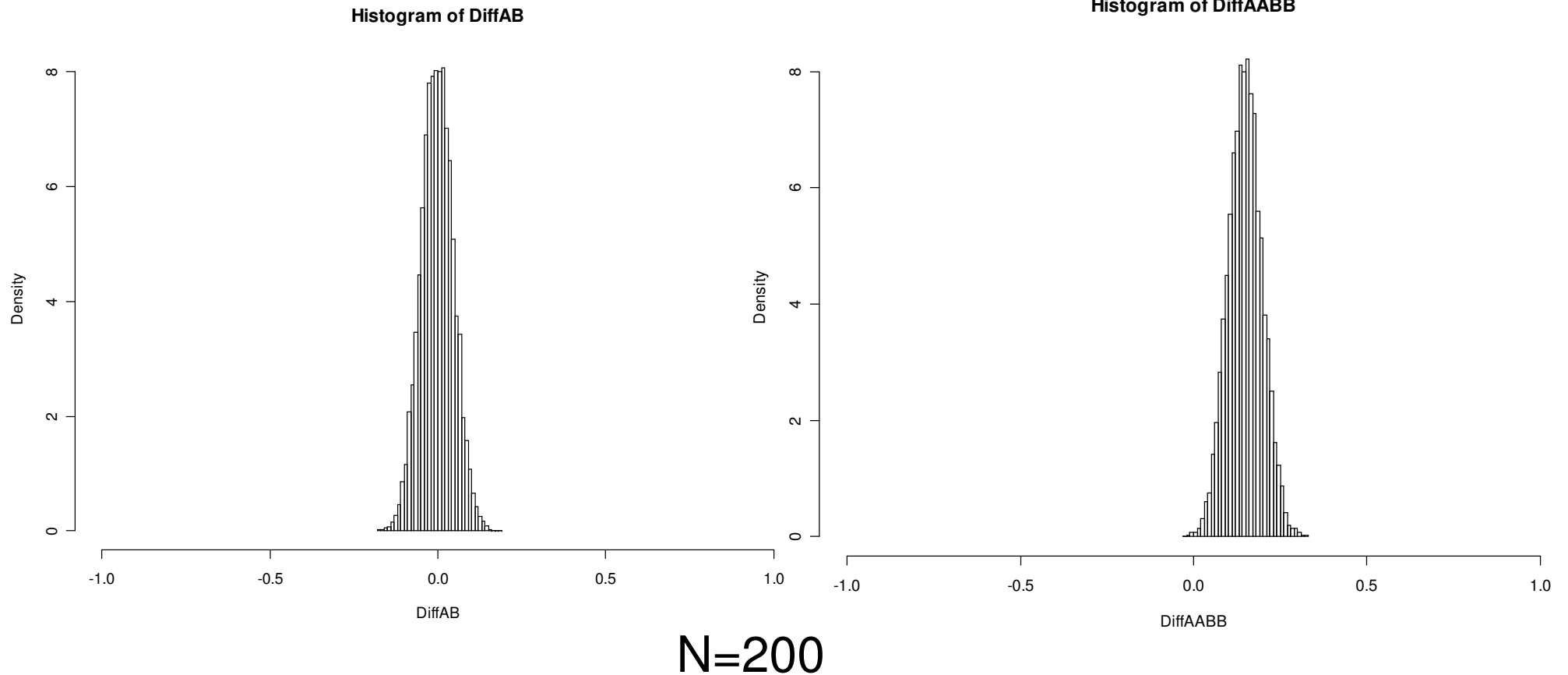
Power<-sum(abs(DiffAABB)>= kritischer Wert)/10000

Power=13% bedeutet:

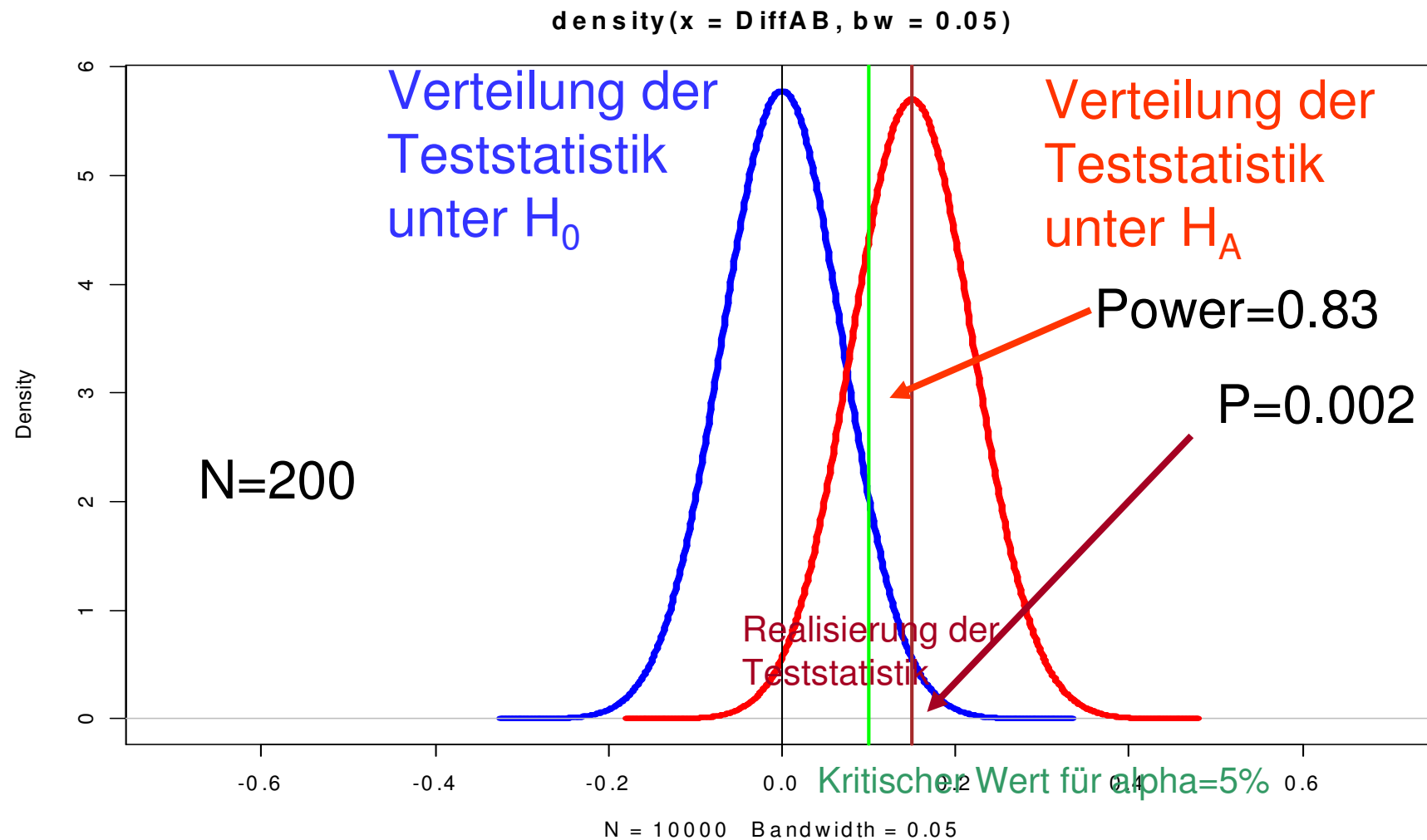
**Wenn der Unterschied tatsächlich 15% ist,
dann haben Sie 87% Wahrscheinlichkeit die
Nullhypothese fälschlicherweise nicht abzulehnen.**

**In einer solchen Situation sollte man sich wegen
unzureichender Information besser nicht entscheiden..**

Histogramm Teststatistik unter der Nullhypothese und der Alternative 15%



Realisierte Teststatistik in Bezug auf die Verteilung der Teststatistik unter Null- und Alternativhypothese



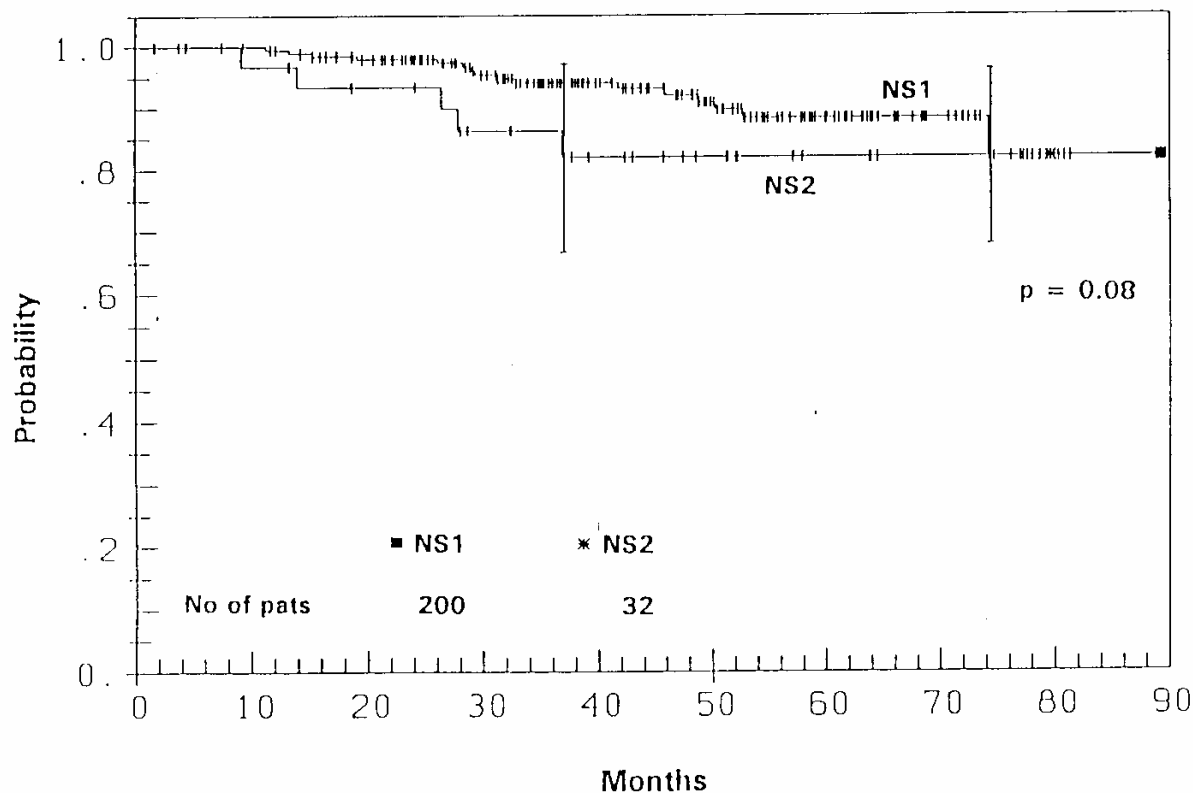
Logik des statistischen Tests IV

**„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 IV

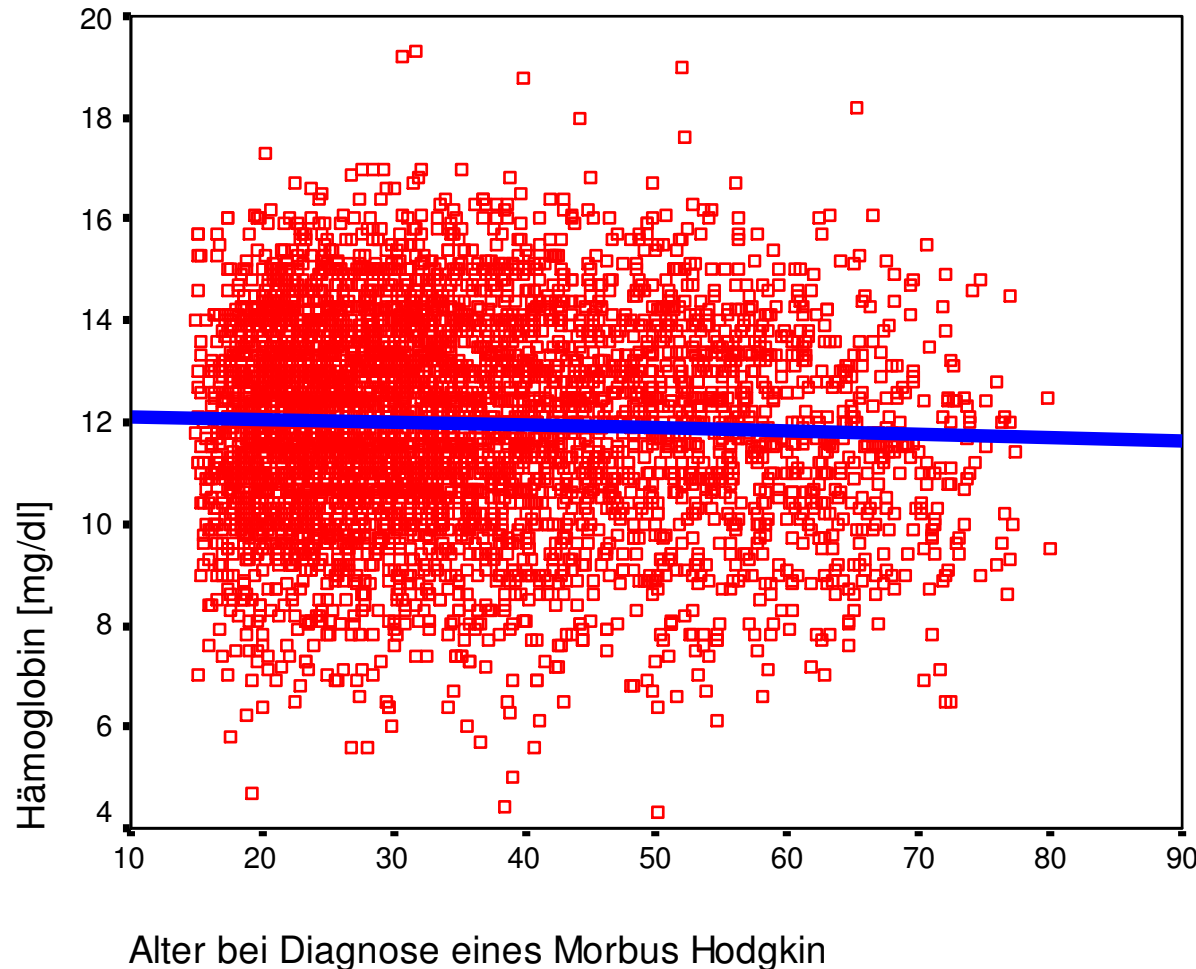
- Wenn Sie einen quantitativ relevanten, aber nicht signifikanten Unterschied beobachten, wissen Sie nur, dass Ihre Fallzahl (!) nicht ausreicht, um bei der sich andeutenden Effektgröße einen Zufallsbefund auszuschließen.
- Denken Sie quantitativ, geben Sie einen Schätzwert für die Differenz mit einem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

Logik des statistischen Tests V

**„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..!“