Repetitorium statistischer Basics:
Konfidenzintervalle, Tests und Fallzahlplanung

Dr. Dirk Hasenclever

Institut für Medizinische Informatik, Statistik & Epidemiologie (IMISE),
and
Zentrum für klinische Studien (ZKS),
University Leipzig

dirk.hasenclever@imise.uni-leipzig.de

SaxoCell Clinics Workshop - Klinische Studien mit ATMPs

2023-03-16 Leipzig

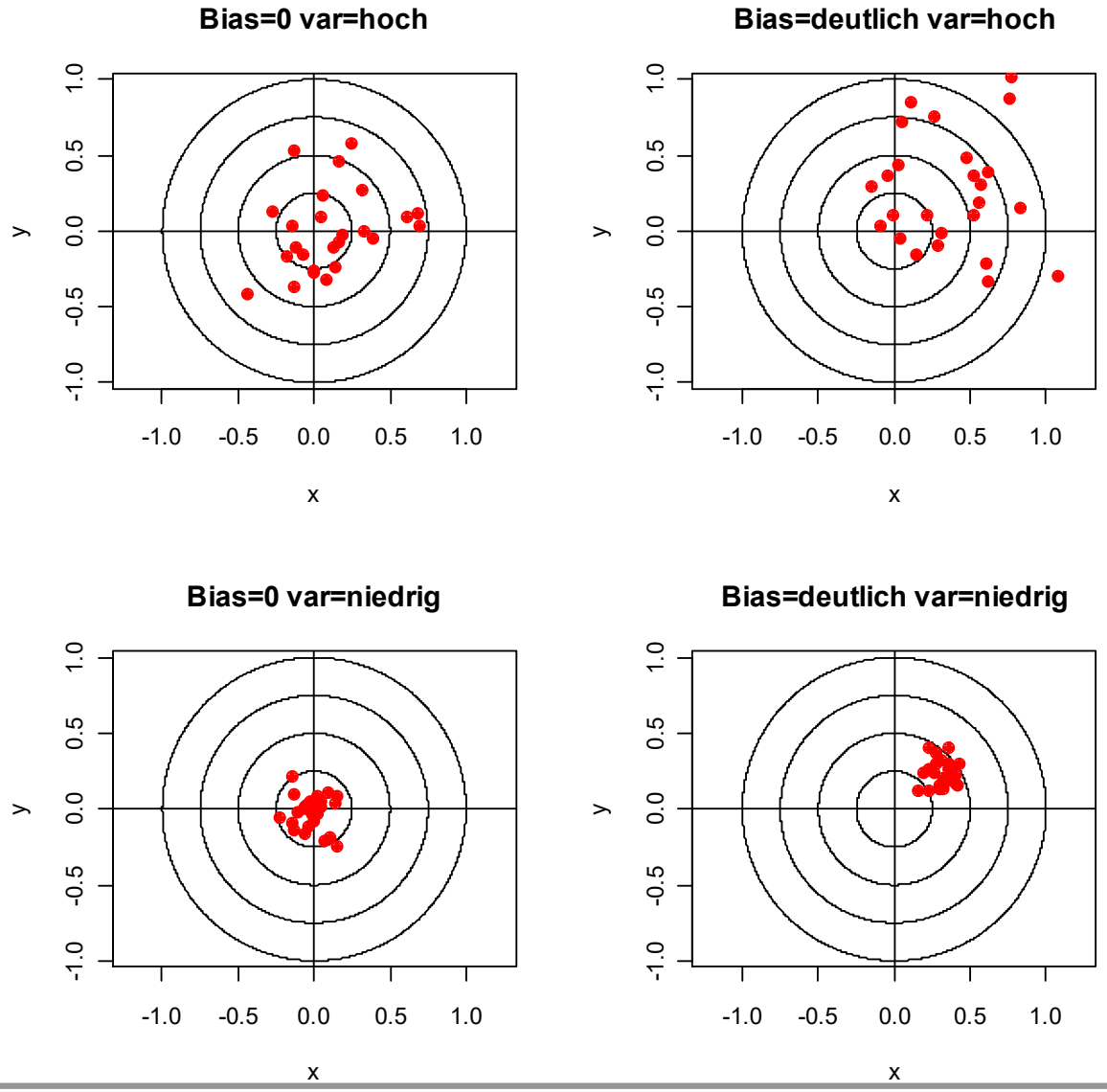
Welche Rolle hat Biometrie und Statistik in klinischen Studien?

Brauchen wir das überhaupt?

- vor allem in frühen Studien?

Antwort: Fehler und Irreführung durch Daten vermeiden!

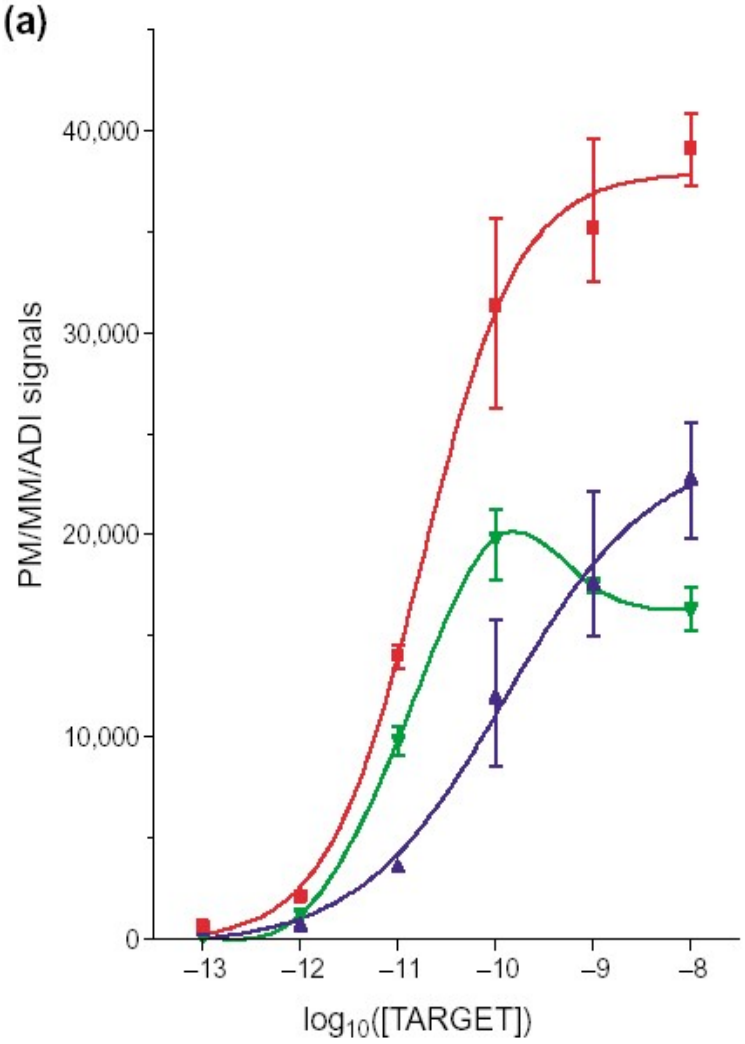
Fehler sind zusammengesetzt aus Verzerrung (Bias) und Zufalls-Streuung



Wann sind Daten interpretierbar und glaubwürdig?

- Keine systematischer Verzerrungen
  - im Messprozess für Einzeldaten und
  - im Prozess der Stichprobenerhebung
- Keine Zufallsbefunde
- (Kein wissenschaftlicher Betrug)

Systematische Verzerrung im Messverfahren



Spezifisches Signal

Hintergrund-Störsignal

Differenz “Hintergrundskorrektur”

Genexpressions-Chips

Chudin 2001

PM red, AvDiff green, MM blue (gemittelt über Probe-Paare); ranges

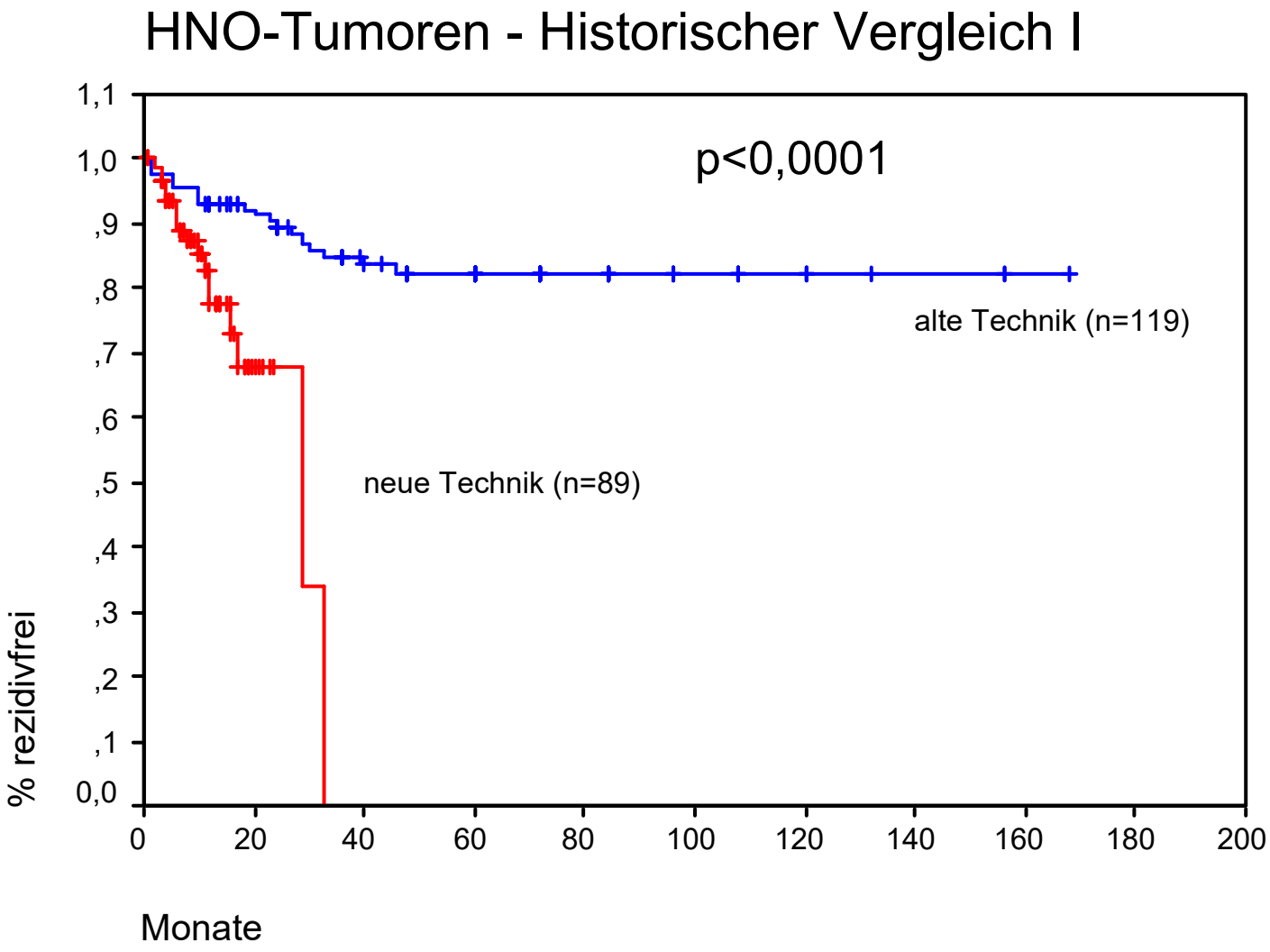
Systematische Verzerrung im Prozess der Stichprobenerhebung

Die Komplikationsrate von Hausgeburten ist signifikant niedriger als die Komplikationsrate von Geburten im Krankenhaus!

(?!)

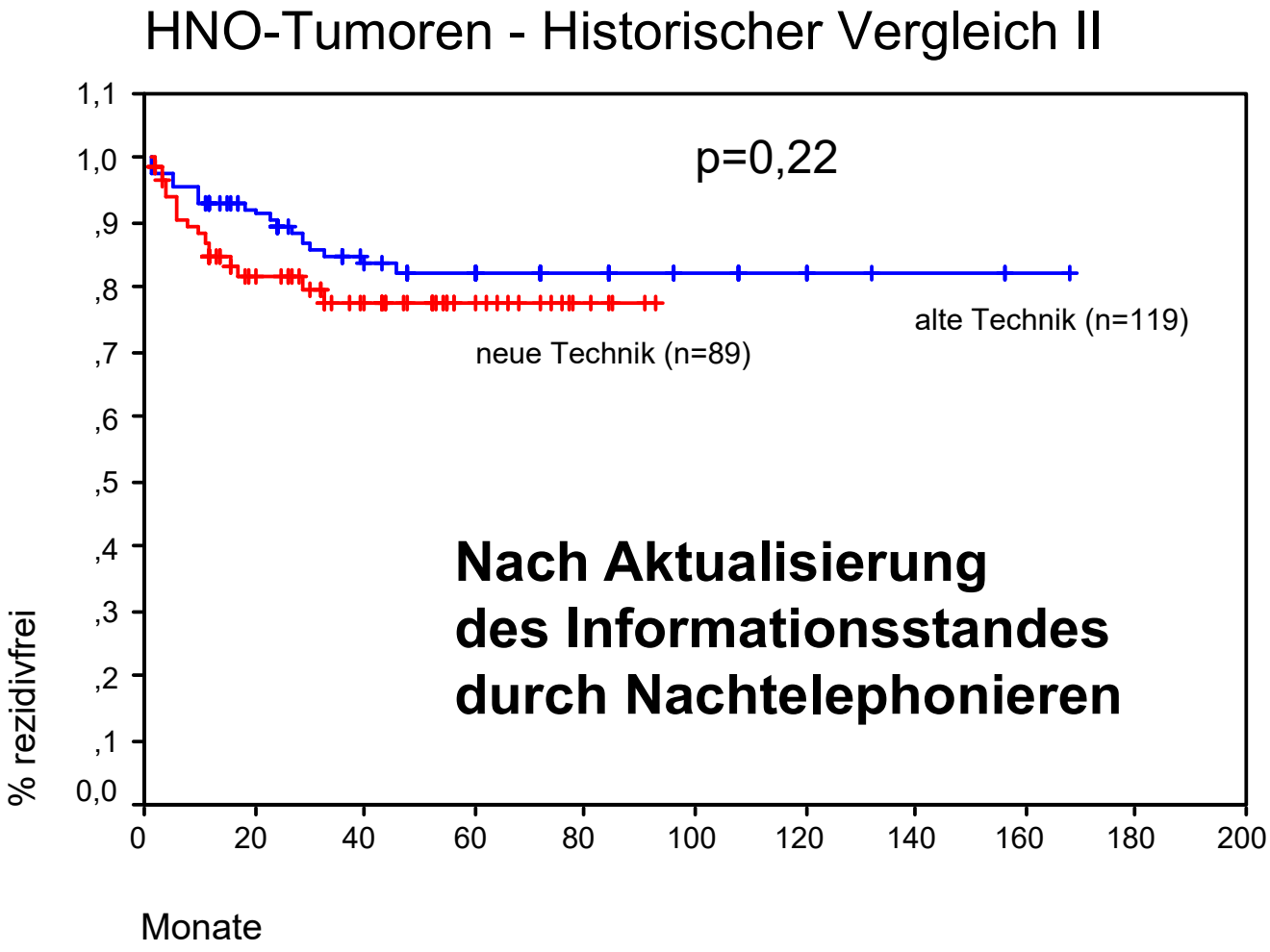
LVZ Panoramaseite ~Frühjahr 1997

Systematische Verzerrung im Informationsfluss



Vergleich der neuen Bestrahlungs-Technik bei HNO-Tumoren mit der alten Technik bzgl. rezidivfreien Überlebens nach Krankenaktenlage

Systematische Verzerrung im Informationsfluss informative Zensur



Kann das Zufall sein?

Dioxin verhindert die Zeugung von Jungen
(*Yahoo! Schlagzeilen: Montag 29. Mai 2000 10.00h*)

Eine hohe Dioxin-Belastung verringert die Chance für Männer, einen Jungen zu zeugen.

Darauf weist das Wissenschaftsmagazin „MorgenWelt“ mit Blick auf eine italienische Studie hin.

Die Forscher hatten die Langzeitfolgen des Dioxin-Unfalls von Seveso im Jahre 1976 untersucht....

Insgesamt hatten die betroffenen Männer seit dem Unfall mit unbelasteten Frauen 88 Jungen und 103 Mädchen gezeugt. Normalerweise liegt die Geburtenrate dagegen bei 106 Jungen zu 100 Mädchen...

Kann das Zufall sein? JA!

Dioxin verhindert die Zeugung von Jungen

(*Yahoo! Schlagzeilen: Montag 29. Mai 2000 10.00h*)

Erwartet: p_0=106/206 = 51,5%

Beobachtet: p = 88/191 = 46,1%

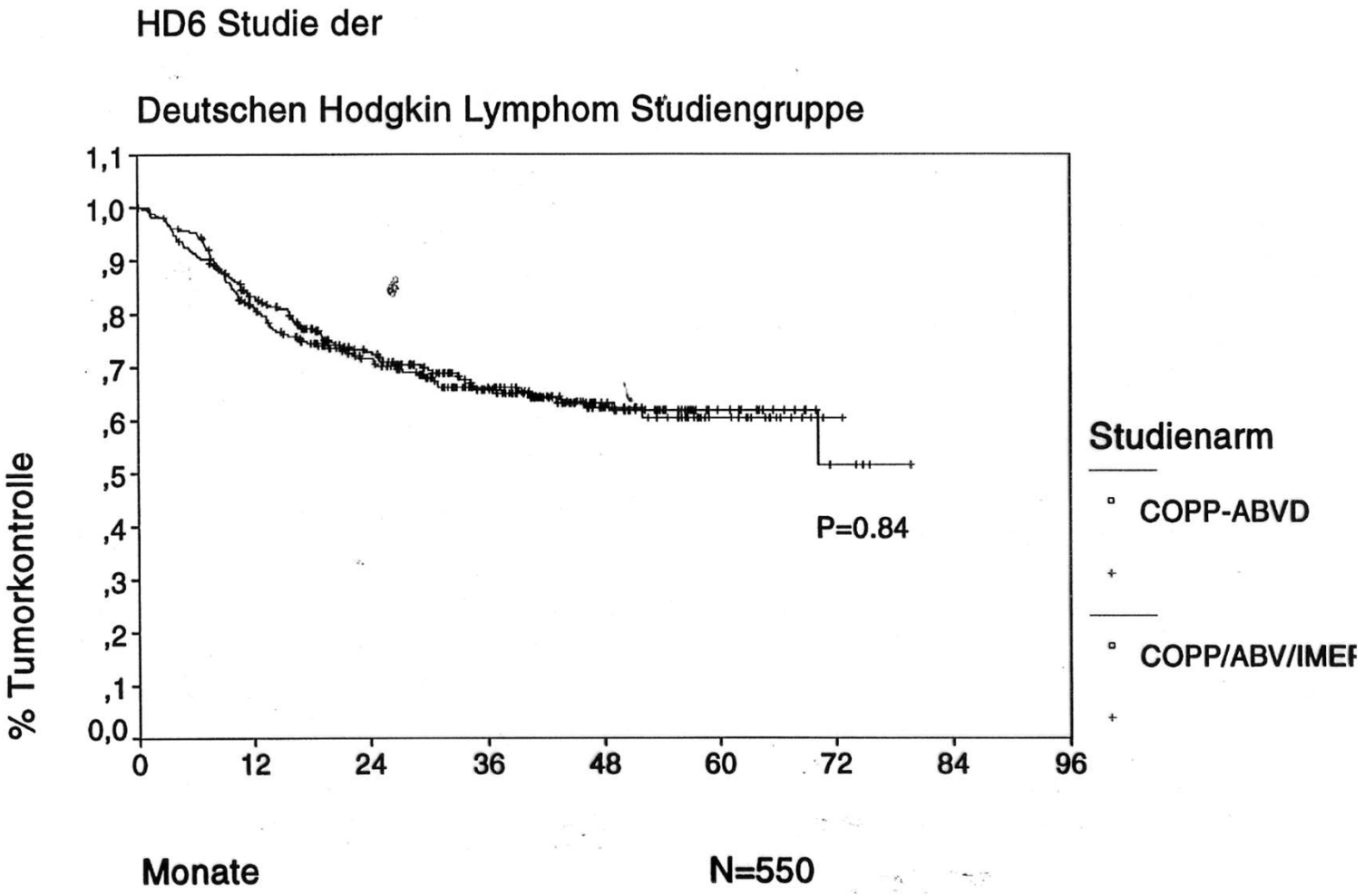
```
> binom.test (88,191, p = 106/206)

        Exact binomial test
data:  88 and 191
number of successes = 88, number of trials = 191, p-value = 0.148
alternative hypothesis: true probability of success is not equal to 0.515
95 percent confidence interval:
 0.389 0.534
sample estimates:
probability of success:           0.460
```

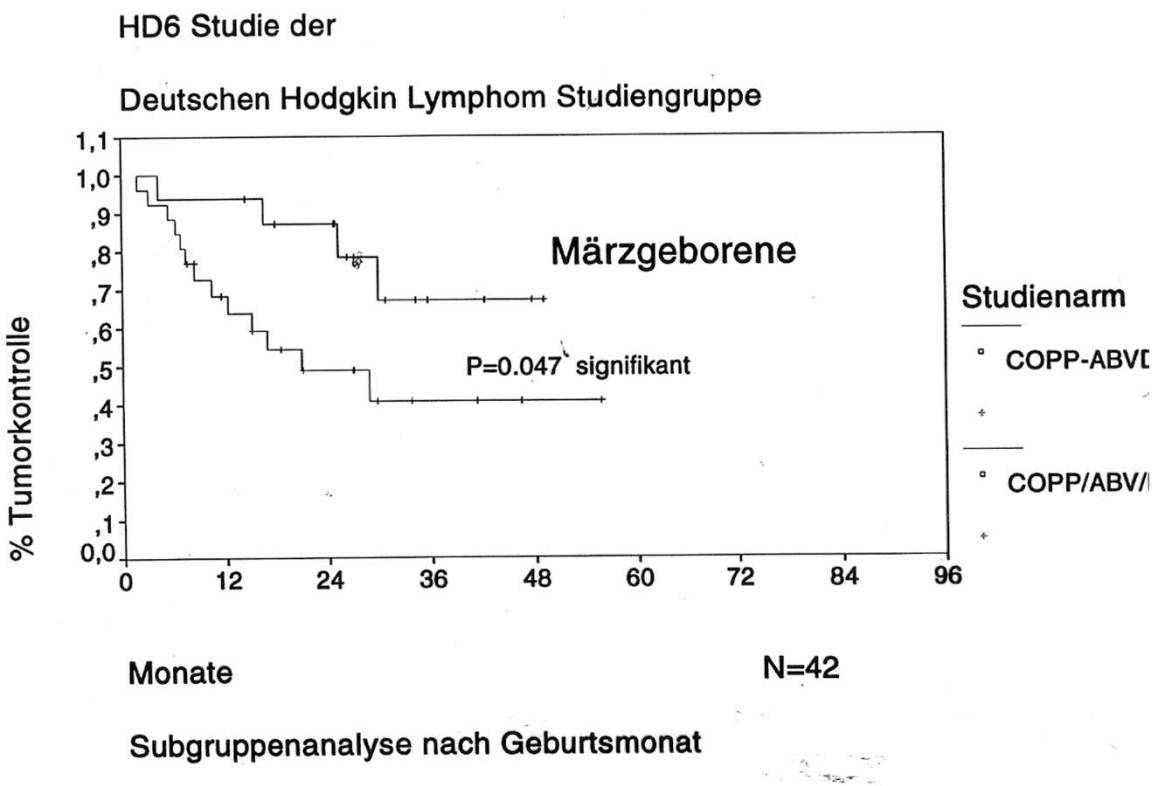
Fazit: **Beobachtete Abweichung ist durchaus mit einem Zufallsbefund verträglich.**

Kann das Zufall sein? Subgruppenanalyse HD6

Randomisierter Chemotherapie Vergleich in fortgeschrittenen Stadien eines Hodgkin Lymphoms.
Keinerlei Unterschied!



Kann das Zufall sein? Subgruppenanalyse HD6



CAVE: Multiples Testen

Test des Therapieeffekts p=0.84

Subgruppenanalyse nach Geburtsmonat:

Januar	p=0.87
Februar	p=0.56
März	**p=0.047** formal signifikant
April	p=0.95
Mai	p=0.28
Juni	p=0.21
Juli	p=0.76
August	p=0.075 Trend in entgegengesetzte Richtung..
September	p=0.94
Oktober	p=0.91
November	p=0.77
Dezember	p=0.22

Irreführende Studienergebnisse

- **Schlechte Datenqualität**:
  Falsche, gefälschte, verfälschte, unvollständige Daten
- kontrolliert durch **Professionelle Qualitätssicherung**,
  Datenmanagement und -monitoring

- **Systematische Verzerrungen**
  - Nicht valide Messmethode
  - Selektive Auswahl der Beobachtungen

  kontrolliert durch **Biometrie**:
  Validierte Messinstrumente
  Biasdiagnostik & Versuchsplanung

- **Zufallsbefunde**

  kontrolliert durch
  Statistik

Der Ansatz der Statistik

Was kann man über Zufallsexperimente lernen?

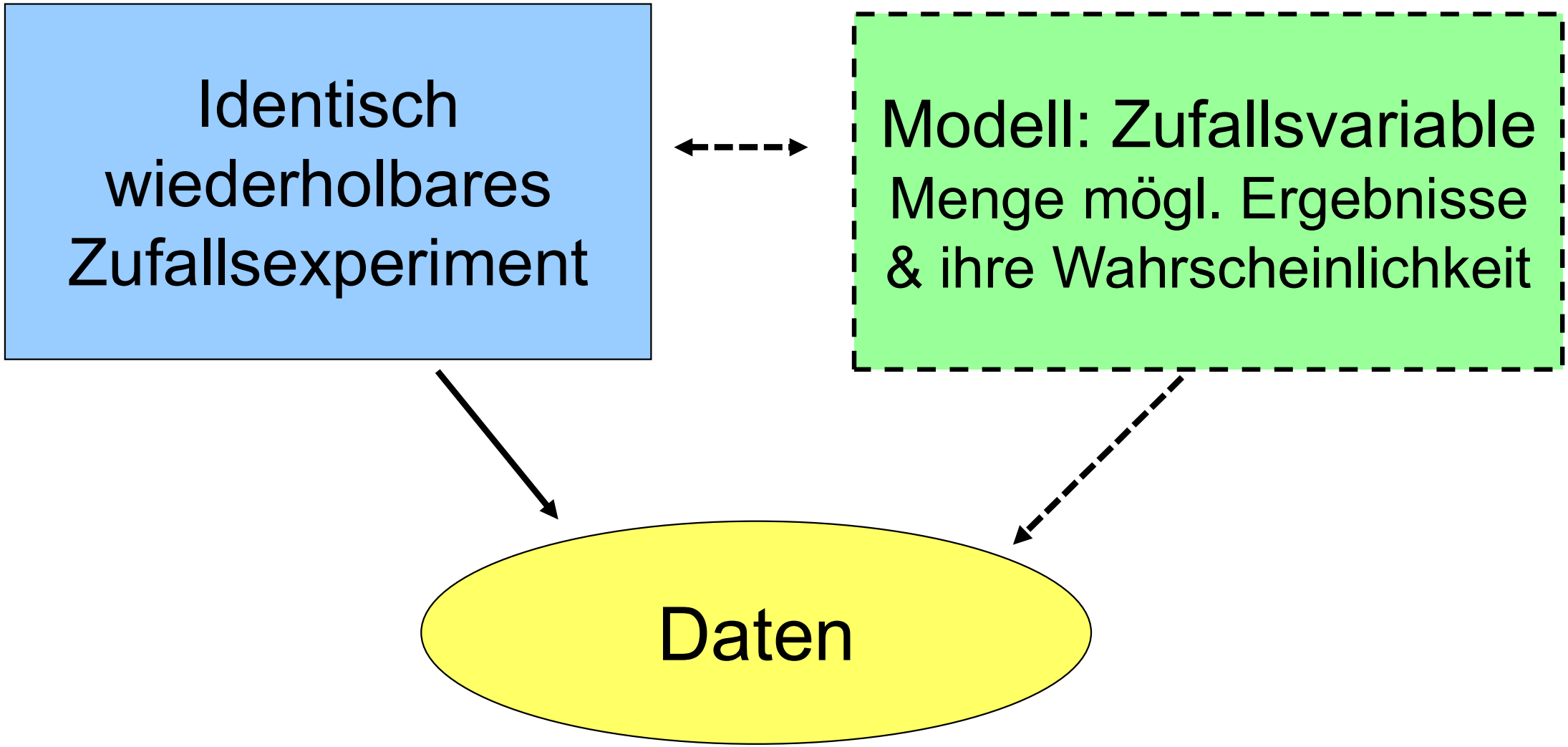
Grundaufgabe der Statistik
Rechenschaft über evidentiale Unsicherheit

- Klinische Studien untersuchen **zufällige, endliche Stichproben**

- Endliche Stichprobe ➔ **Allgemeine Aussage über Versuchsprinzip**
  z.B. Therapieverfahren
- **Evidentiale UNSICHERHEIT** dabei muss **ehrlich offengelegt werden**.

- **Statistik** kontrolliert diese Unsicherheit und schützt vor **irreführenden Zufallsbefunden**.

Statistischer Ansatz



Was ist Wahrscheinlichkeit?

Definition 1.4: *Sei $(\Omega, \mathfrak{S})$ ein meßbarer Raum und sei P eine Abbildung von $\mathfrak{S}$ in $\mathbb{R}$. P heißt ein Wahrscheinlichkeitsmaß oder kurz eine Wahrscheinlichkeit, wenn gilt:*

(I) $A \in \mathfrak{S} \Rightarrow P(A) \geqq 0.$

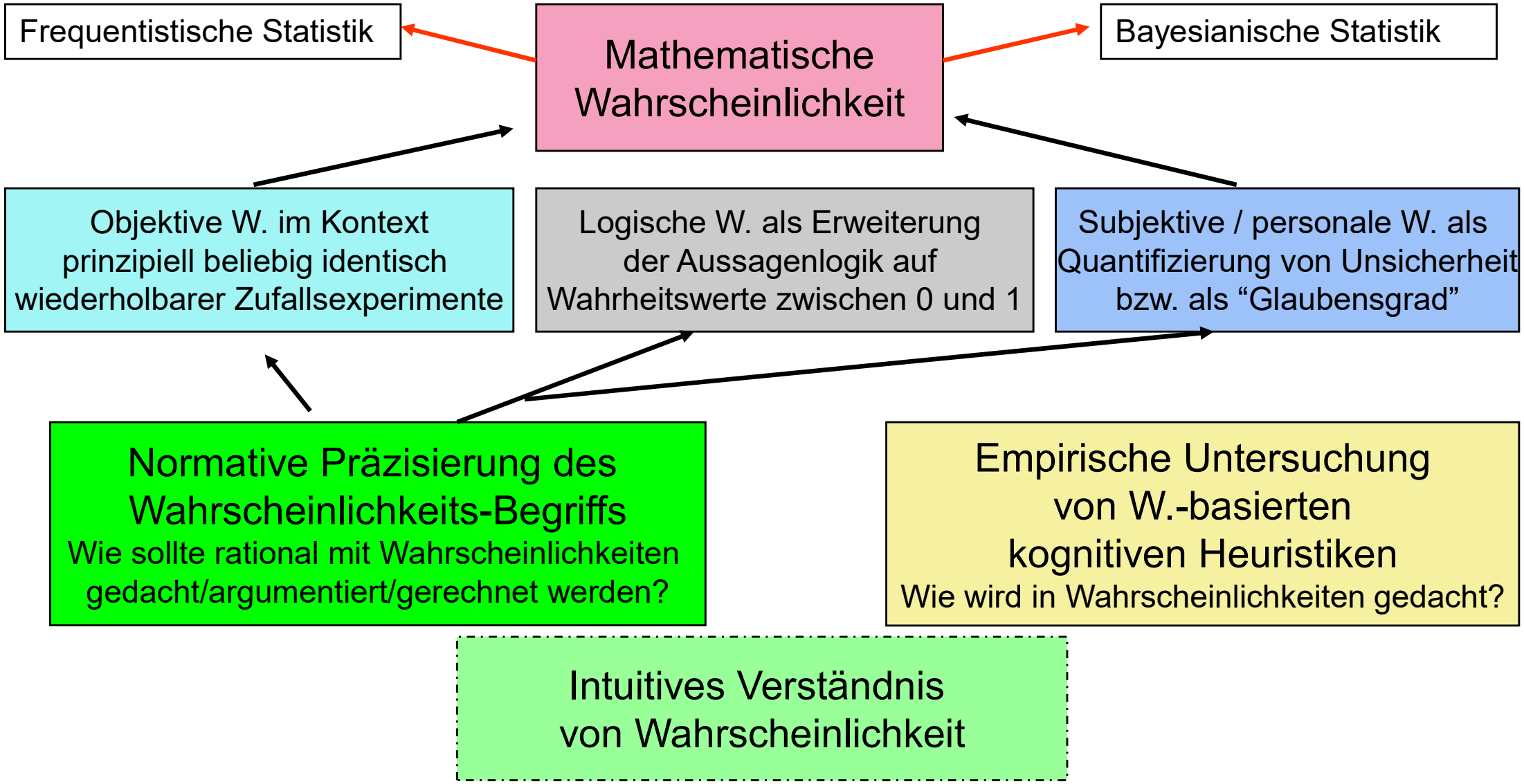
(II) $(A_n \in \mathfrak{S}$ *für alle* $n \in \mathbb{N}$ *und* $A_i \cap A_k = \emptyset$ *für* $i \neq k) \Rightarrow$

$$P\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} P(A_n).$$

(III) $P(\Omega) = 1.$

Der mathematische Begriff der Wahrscheinlichkeit bedarf im Kontext der Erkenntnisgewinnung einer Realinterpretation!

Wahrscheinlichkeiten



Zwei intuitive Quellen für Wahrscheinlichkeit

- Stabile Häufigkeiten bei wiederholbaren Versuchen
  (besonders bei Symmetrie: Würfel, Urne etc.)
  → **frequentistischer W-Begriff**

- Rationale Einschätzungen von Eventualitäten beim Handeln unter Unsicherheit
  Objektiviert sich als „fairer Wettquotient“
  → **personalistischer (Bayes) W-Begriff**

- Mathematische Axiome gelten für beide Realbegriffe.

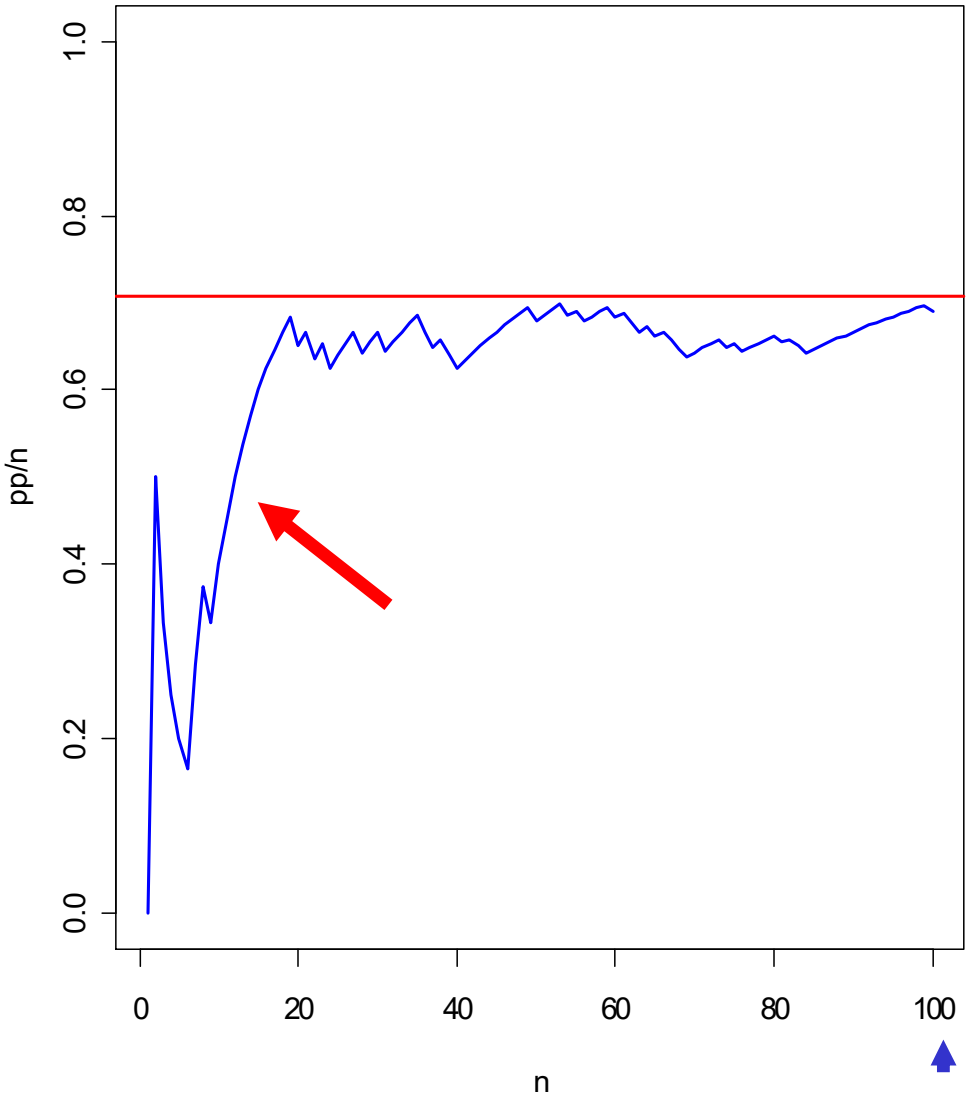
Frequentistischer W-Begriff

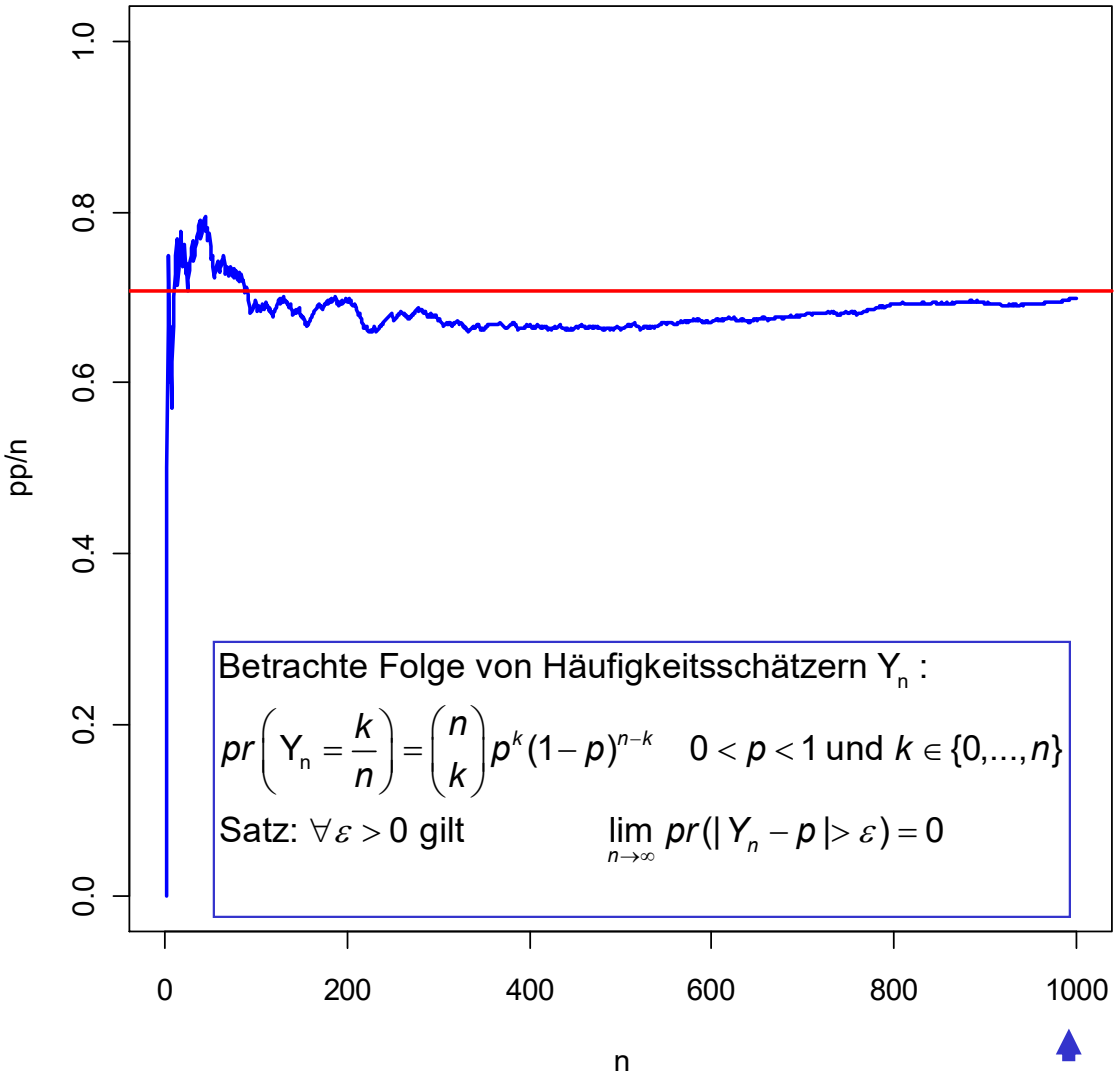
- W. ist eine objektive, aber latente **Eigenschaft (identisch) wiederholbarer Zufallsexperimente**.

- Die W. eines bestimmten Ergebnisses eines wiederholbaren Zufallsexperiments lässt sich durch die **empirische Häufigkeit** in einer hinreichend langen Serie von Experimenten approximativ schätzen.

- W. von singulären Ereignissen (z.B. Ergebnis eines bestimmen Fußballspiels) ist undefiniert.

Gesetz der Großen Zahl: Häufigkeit → Wahrscheinlichkeit





Betrachte Folge von Häufigkeitsschätzern Y_n :

$$pr\left(Y_n = \frac{k}{n}\right) = \binom{n}{k} p^k (1-p)^{n-k} \quad 0 < p < 1 \text{ und } k \in \{0,\ldots,n\}$$

Satz: $\forall \varepsilon > 0$ gilt $\lim_{n\to\infty} pr(|Y_n - p| > \varepsilon) = 0$

Bayesianischer Wahrscheinlichkeitsbegriff I

- W. = **rationaler Grad von Glauben** oder
  rationale Quantifizierung von Unsicherheit.
- W. stellt subjektive Ungewissheit dar.
- W. wird formalisiert als
  - **Erweiterung der Logik auf Wahrheitswerte zwischen 0 und 1 (Jeffreys, Jaynes)**
  - Rationale Wettquotienten
    (Rational= Vermeide Systeme von Wetten mit sicherem Verlust) (de Finetti)

Sind Sie Frequentist oder Bayesianer?

- Kurzbeschreibungen von 100 realen Personen:

  70 Rechtanwälte und 30 Ingenieure.

> *Jack ist 45 Jahre alt, verheiratet, 4 Kinder.*
>
> *Jack ist konservativ, sorgfältig und ehrgeizig.*
>
> *Jack zeigt kein Interesse an politischen oder sozialen Themen und verbringt den größten Teil seiner Freizeit mit seinen vielfältigen Hobbies, darunter Heimwerkerei, Segeln und mathematische Rätsel*

- Die Wahrscheinlichkeit, dass Jack einer der 30 Ingenieure unter den 100 Personen ist, beträgt:

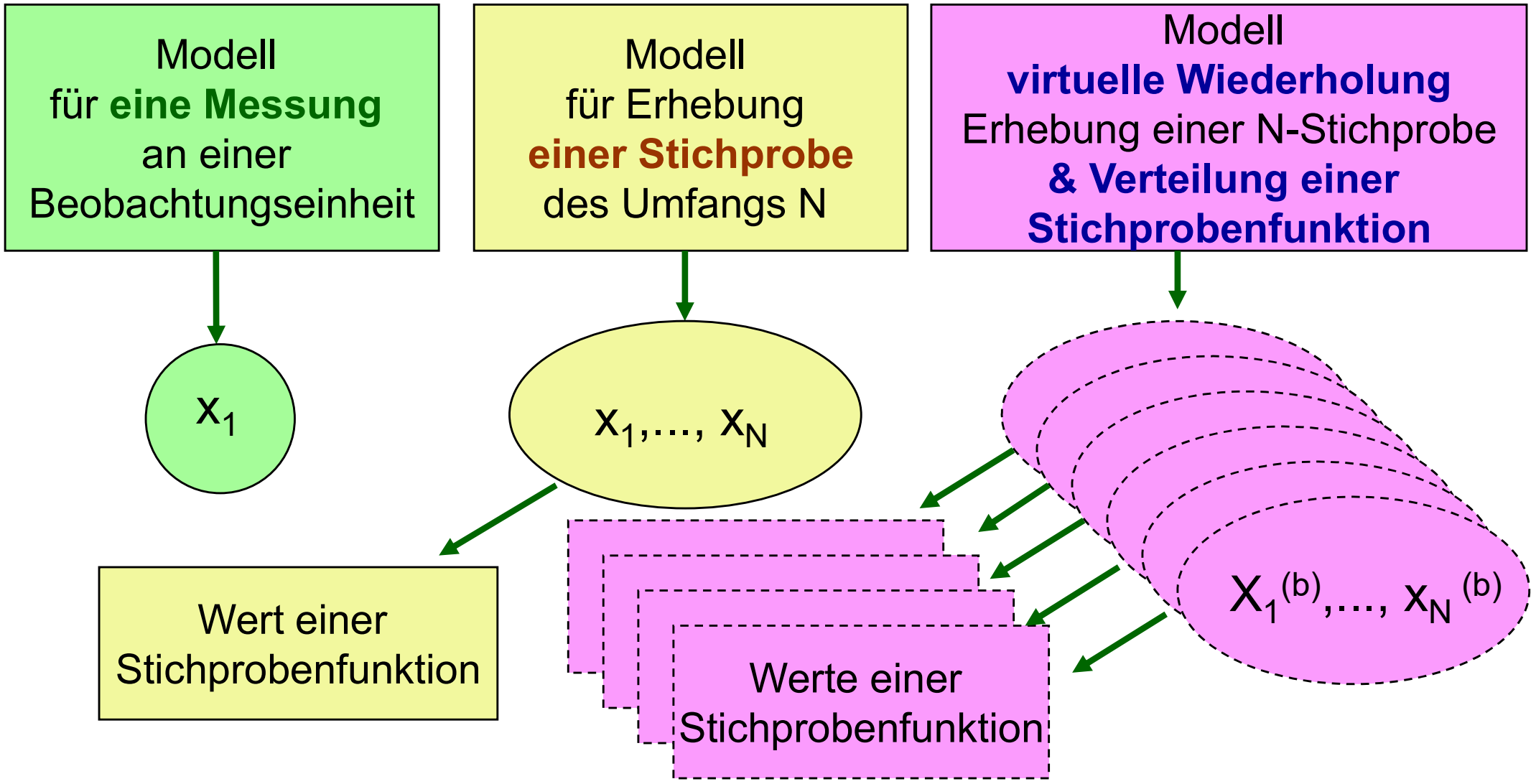
(A) deutlich unter 50% (B) ca. 50% (C) deutlich über 50%

Nach Kahnemann & Tversky (1982)

Mathematische Beschreibung von Zufallsexperimenten

Worüber reden Statistiker?

Modellierung von Zufallsexperiemente



Mathematisches Jenseits

Statistisches Diesseits

X Zufallsvariable → x Beobachtung

Mathematische Stichprobe → $x_1, x_2, \ldots, x_N$ empirische Stichprobe

Wahrscheinlichkeit → Häufigkeit

$F(x) = pr(X \leq x)$ → $\hat{F}(x) = \#\{ x_i \mid x_i \leq x \} / N$

theoretische Verteilungsfunktion → empirische Verteilungsfunktion

Theoretischer Erwartungswert → Empirischer arithmetischer Mittelwert:

$$\bar{x} = \frac{1}{N} \sum x_i$$

Theoretische Varianz → Emprischer Varianz:

$$\hat{\sigma} = \frac{1}{N-1} \sum (x_i - \bar{x})^2$$

Déformation professionnelle?

- **Statistiker reden über Zufallsvariable** als Modellrepräsentationen allgemeiner Verfahren.

- **Statistiker** reden eigentlich **nie über zufällige Daten,**
- sondern mittels Daten über
- **latente Modell-Parameter / Funktionen,**
- **die die Natur in dem fraglichen Zufallsexperiment (der Studie) „versteckt“ hat.**

- **NUR diese Parameter** kann man aus einer Studie **verallgemeinerbar** lernen…
  (z.B. eine Erfolgswahrscheinlichkeit)

Der Grundaufgaben der Statistik

Was tun die eigentlich?

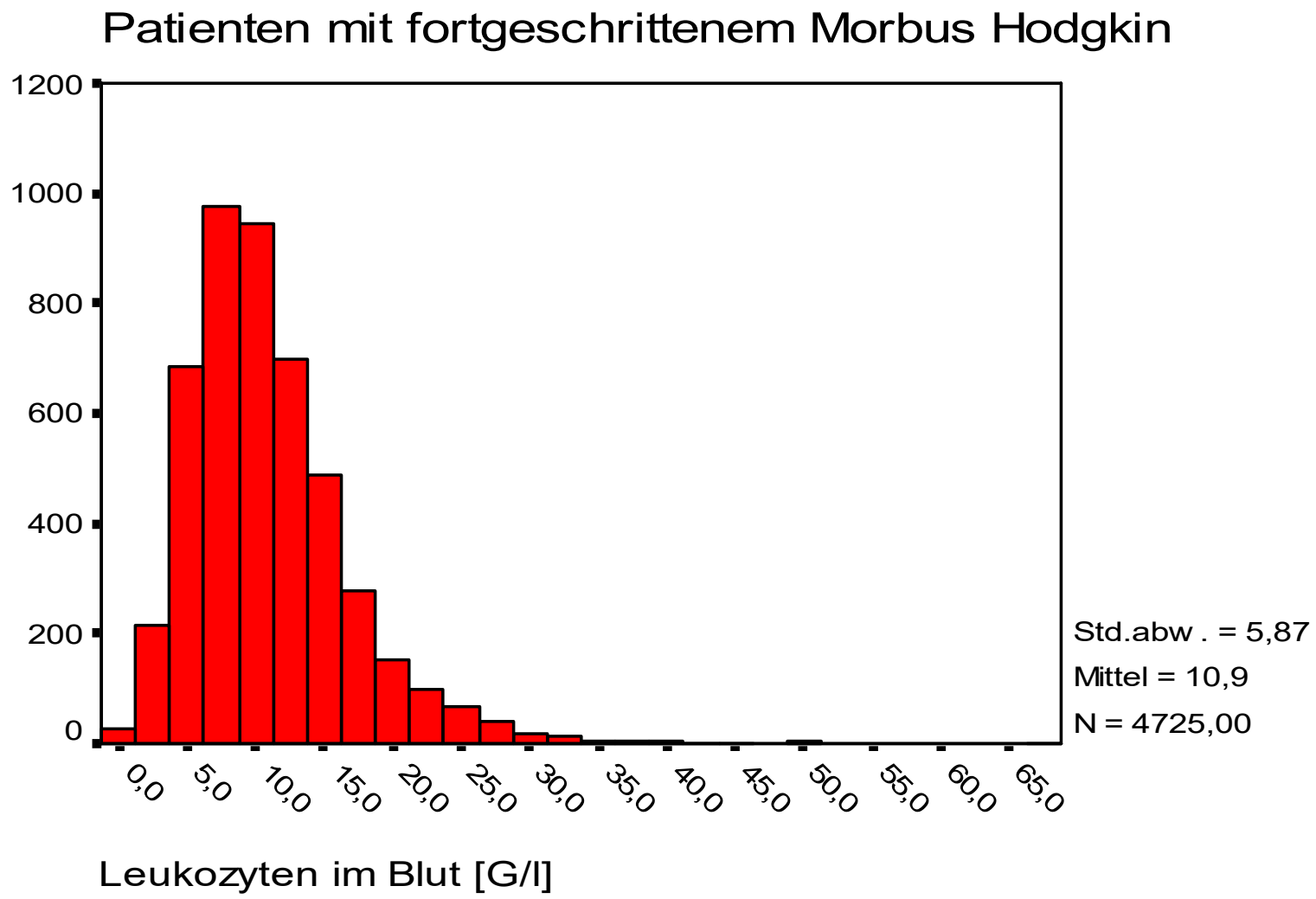
Basisaufgaben der Statisik

- Schätzen von **Verteilungsfunktionen**
  $$F(x) = pr(X \leq x)$$

- **Punktschätzer** für Kenngrößen mit **Konfidenzintervallen** zur Angabe der Schätzgenauigkeit

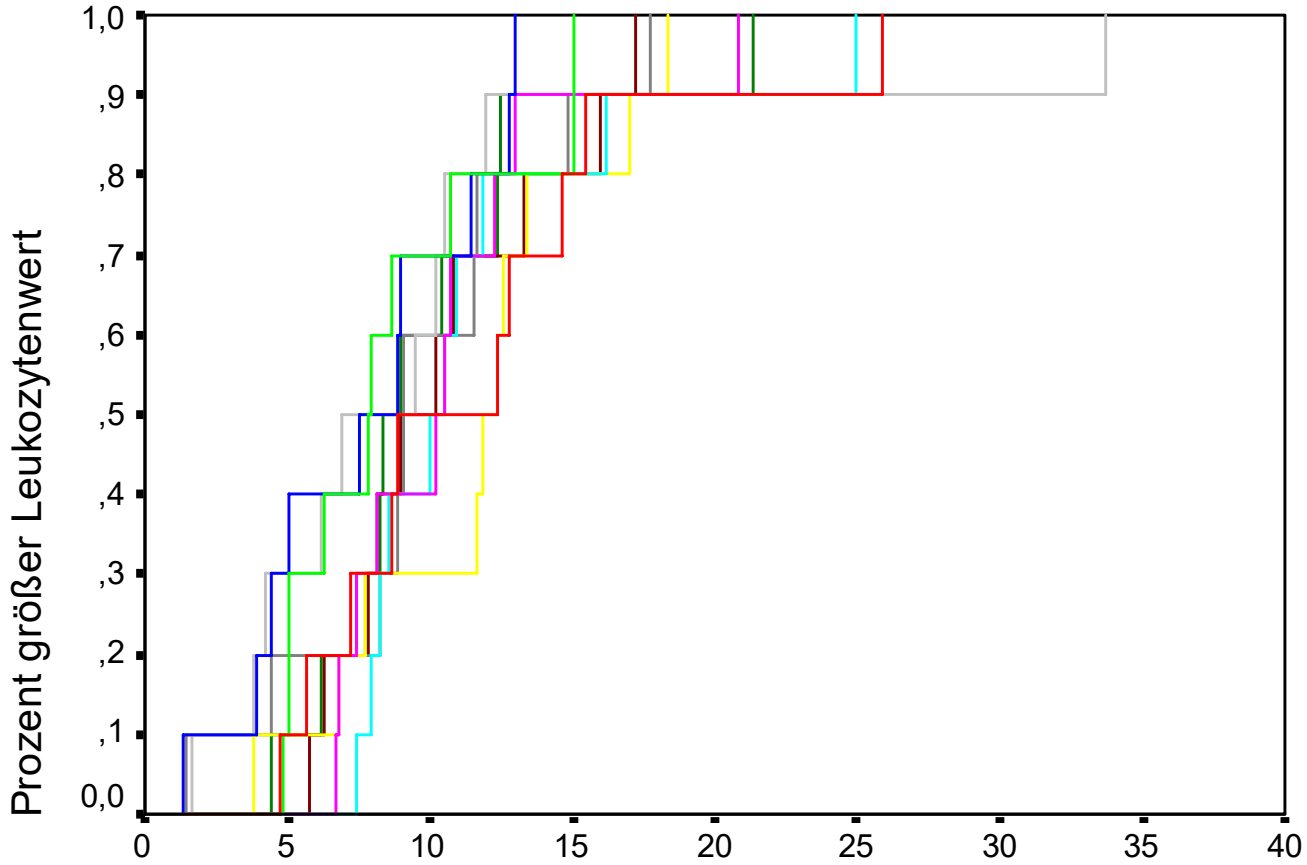
- **Bewertung statistischer Hypothesen** (= Aussagen über latente Parameter) mittels **statistischer Tests**

Histogramm (empirische Verteilungsdichte)



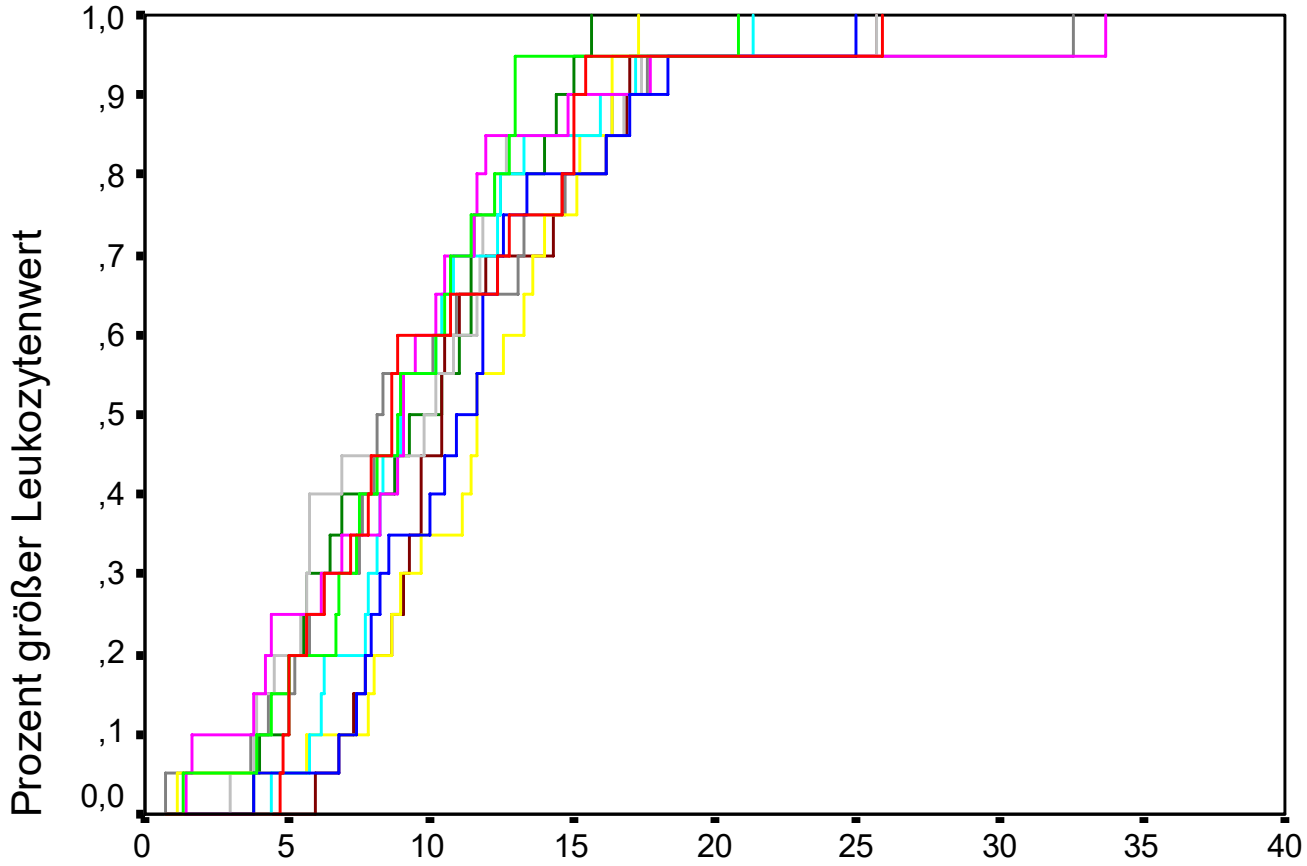
Schätzung einer Verteilung – Effekt der Fallzahl

9 Stichproben mit 10 Patienten



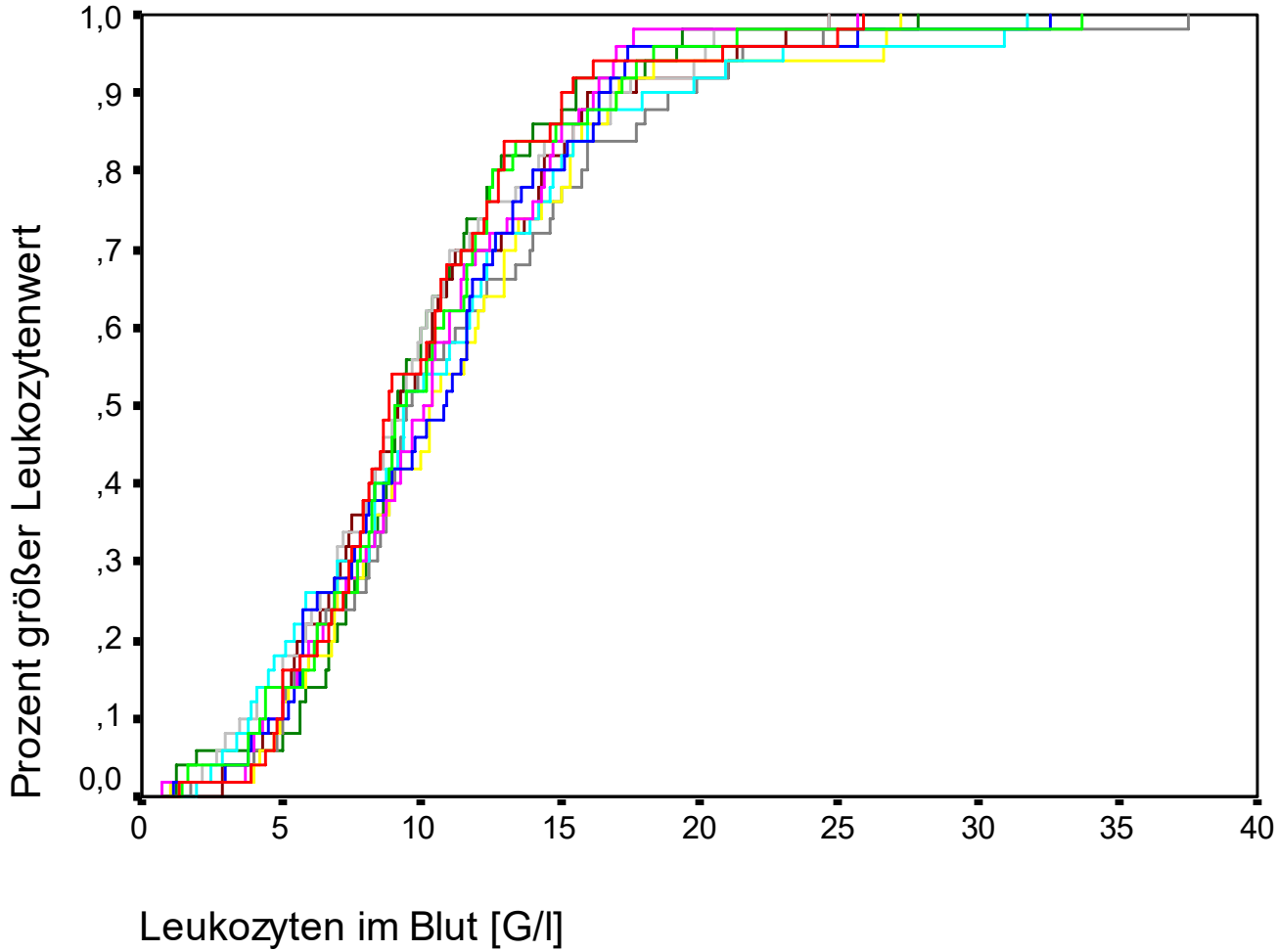
Schätzung einer Verteilung – Effekt der Fallzahl

9 Stichproben mit 20 Patienten



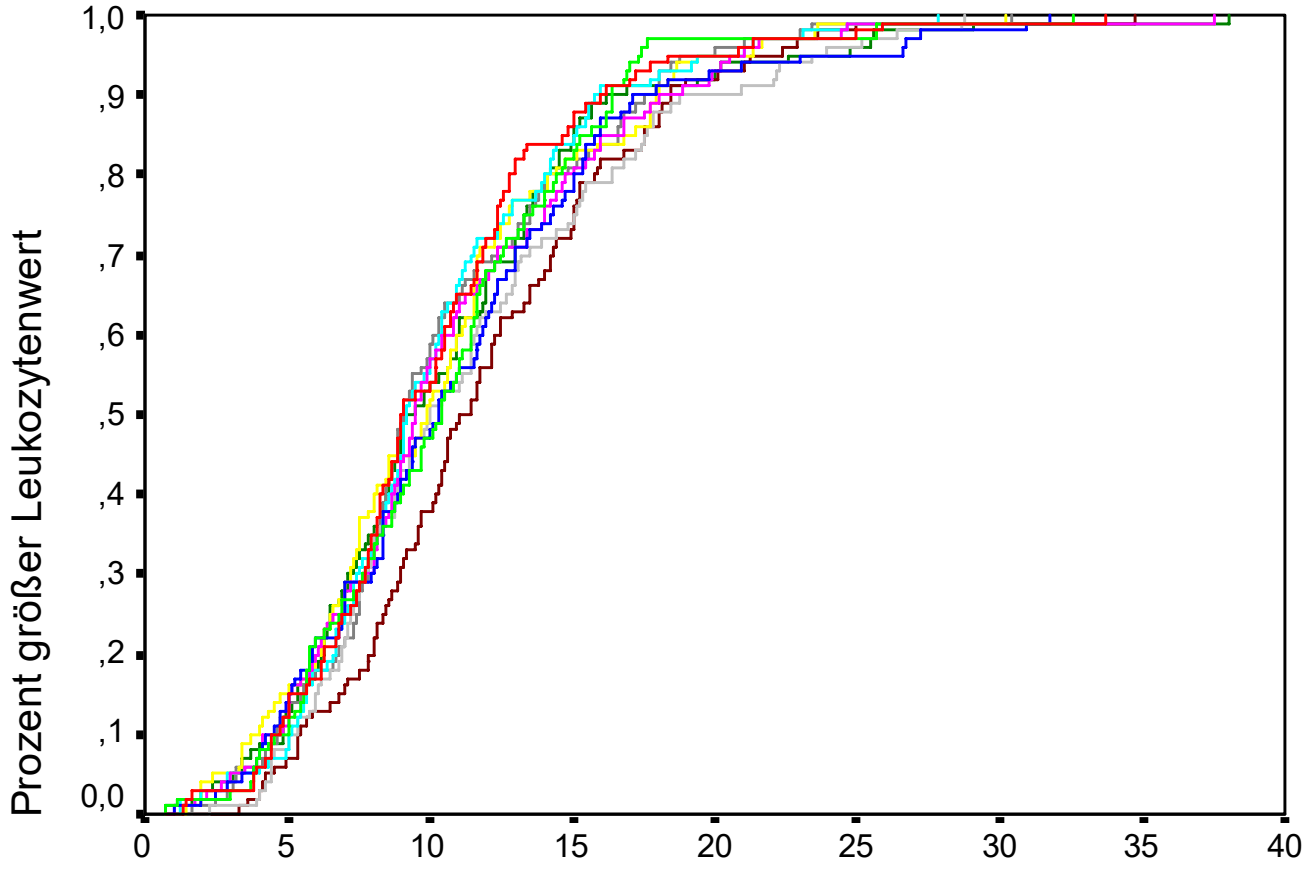
Schätzung einer Verteilung – Effekt der Fallzahl

9 Stichproben von 50 Patienten



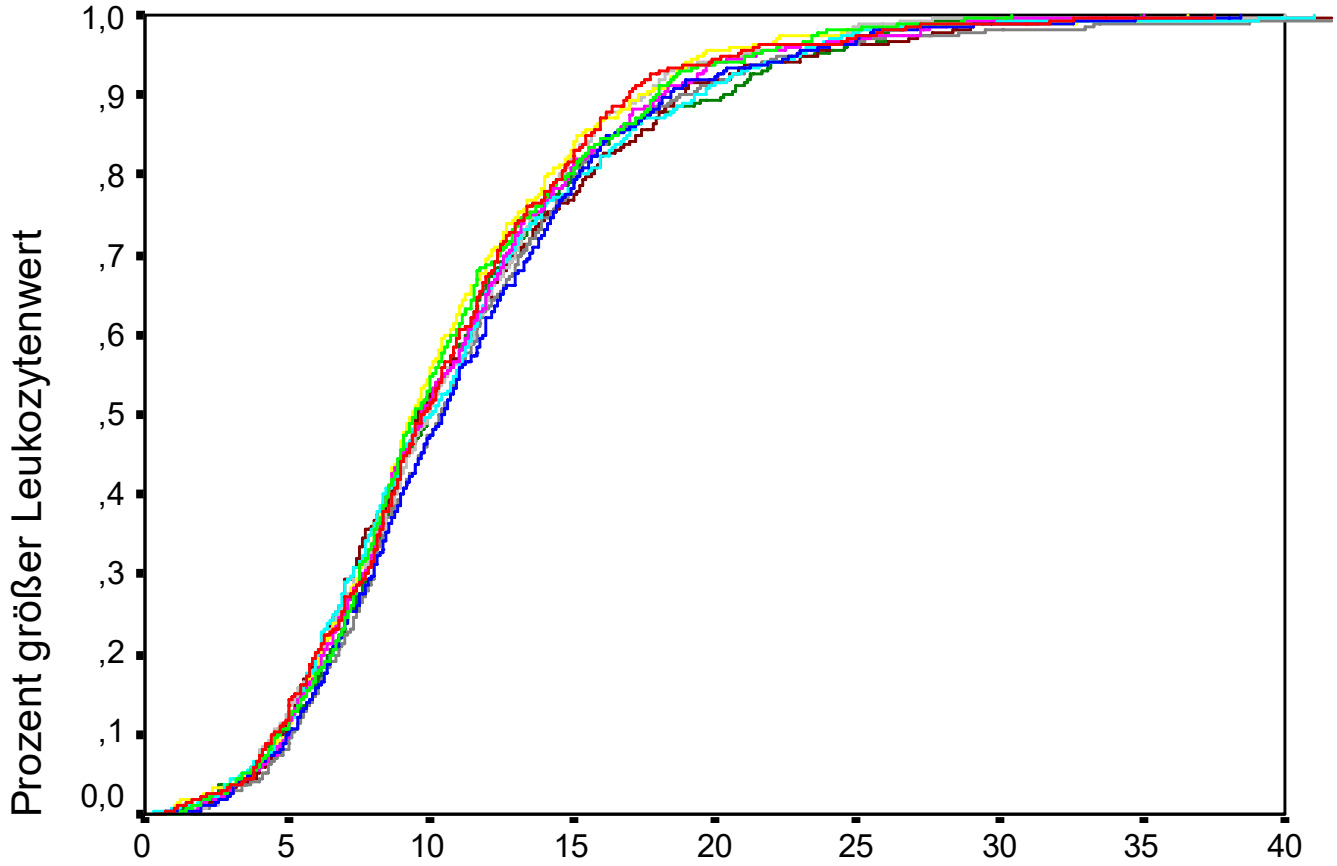
Schätzung einer Verteilung – Effekt der Fallzahl

9 Stichproben mit 100 Patienten

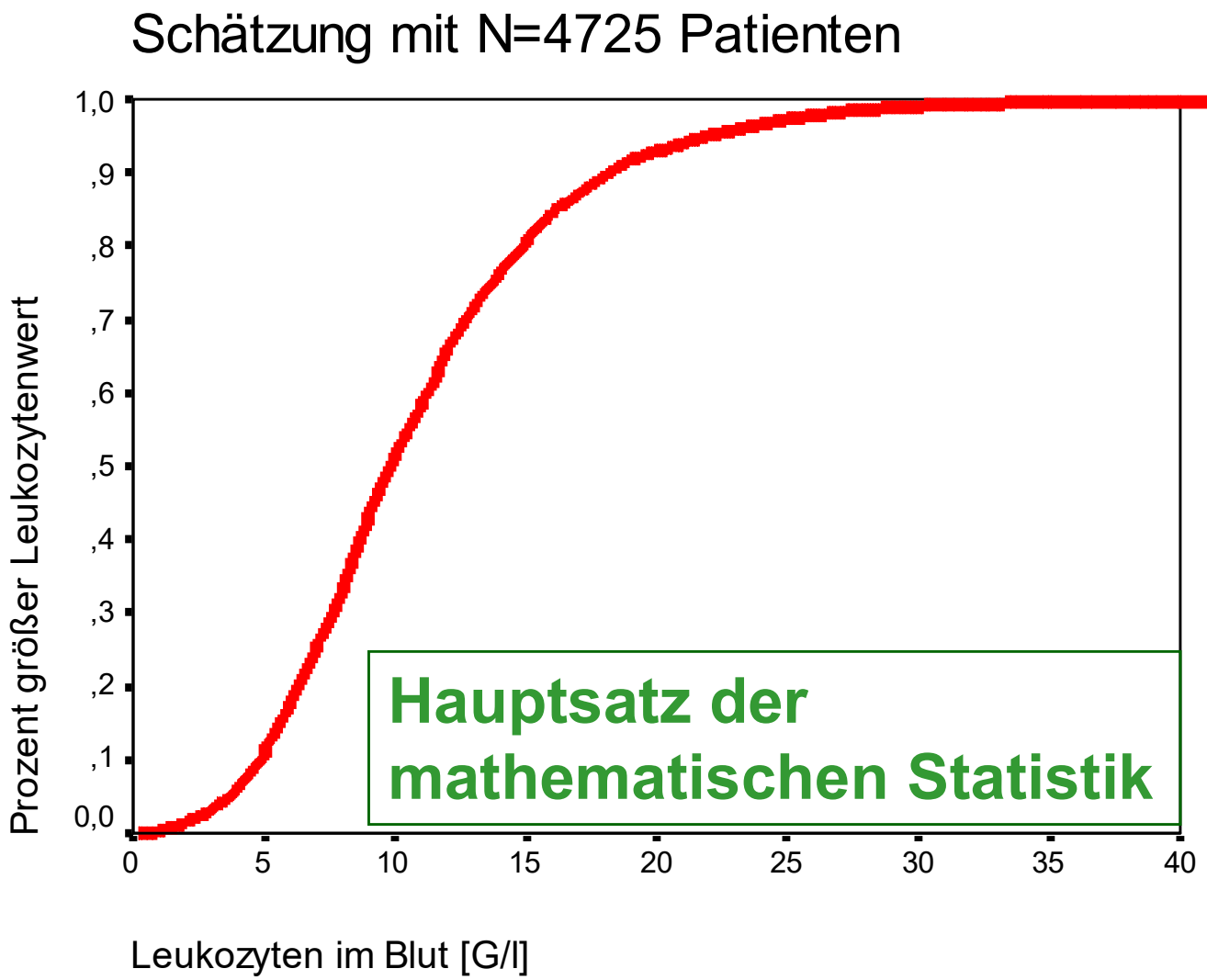


Schätzung einer Verteilung – Effekt der Fallzahl

9 Stichproben mit 400 Patienten



Schätzung einer Verteilung – Effekt der Fallzahl



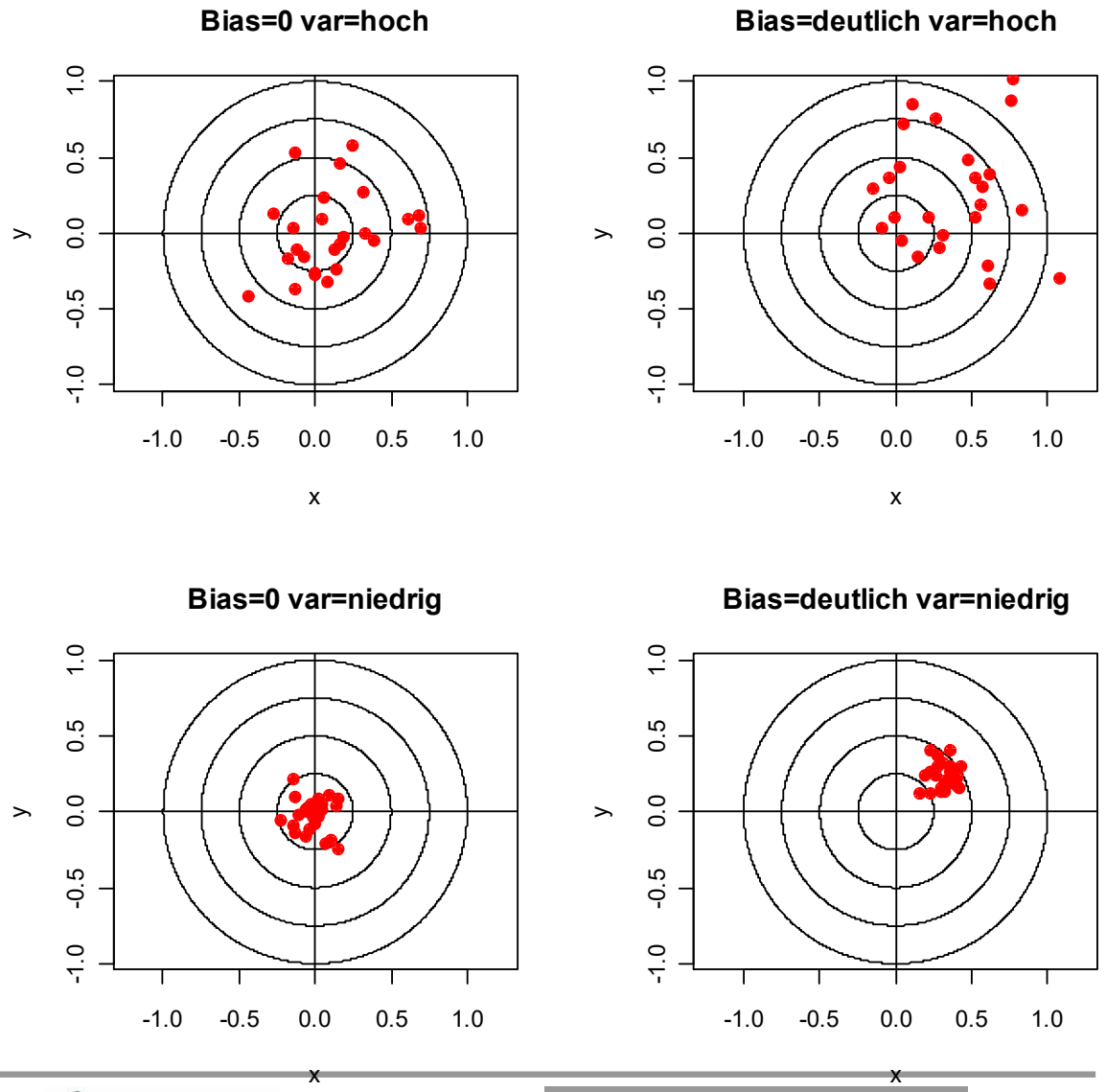
Aber: Konvergenz sehr langsam!

Schätze nicht Verteilungen, sondern Kenngrößen!

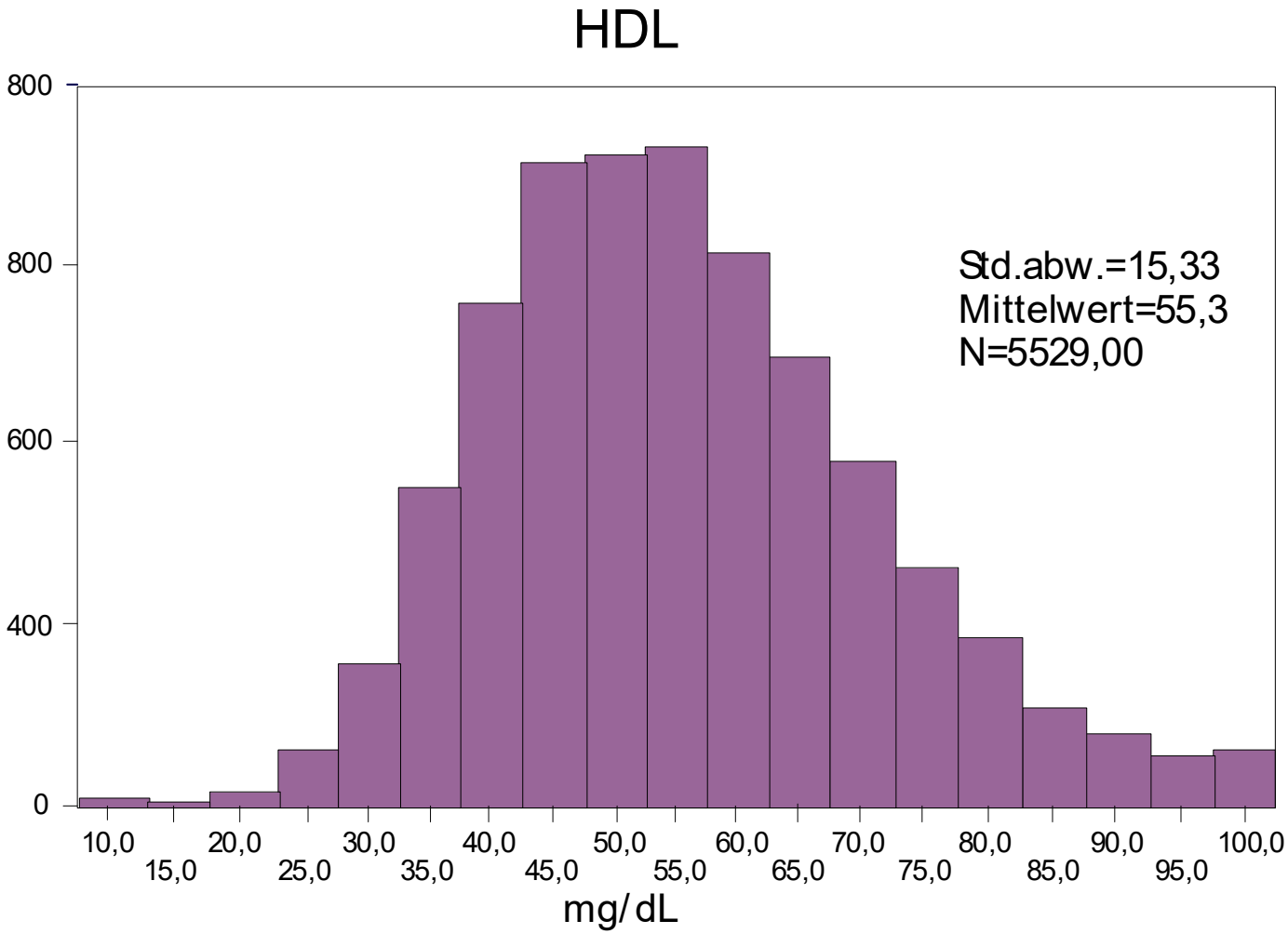
Punktschätzer für Kenngrößen einer Verteilung: konsistent, niedrige Varianz, erwartungstreu

Punktschätzer =
Verfahren aus
einer Stichprobe
einen Schätzwert
für einen latenten
Parameter
zu berechnen.

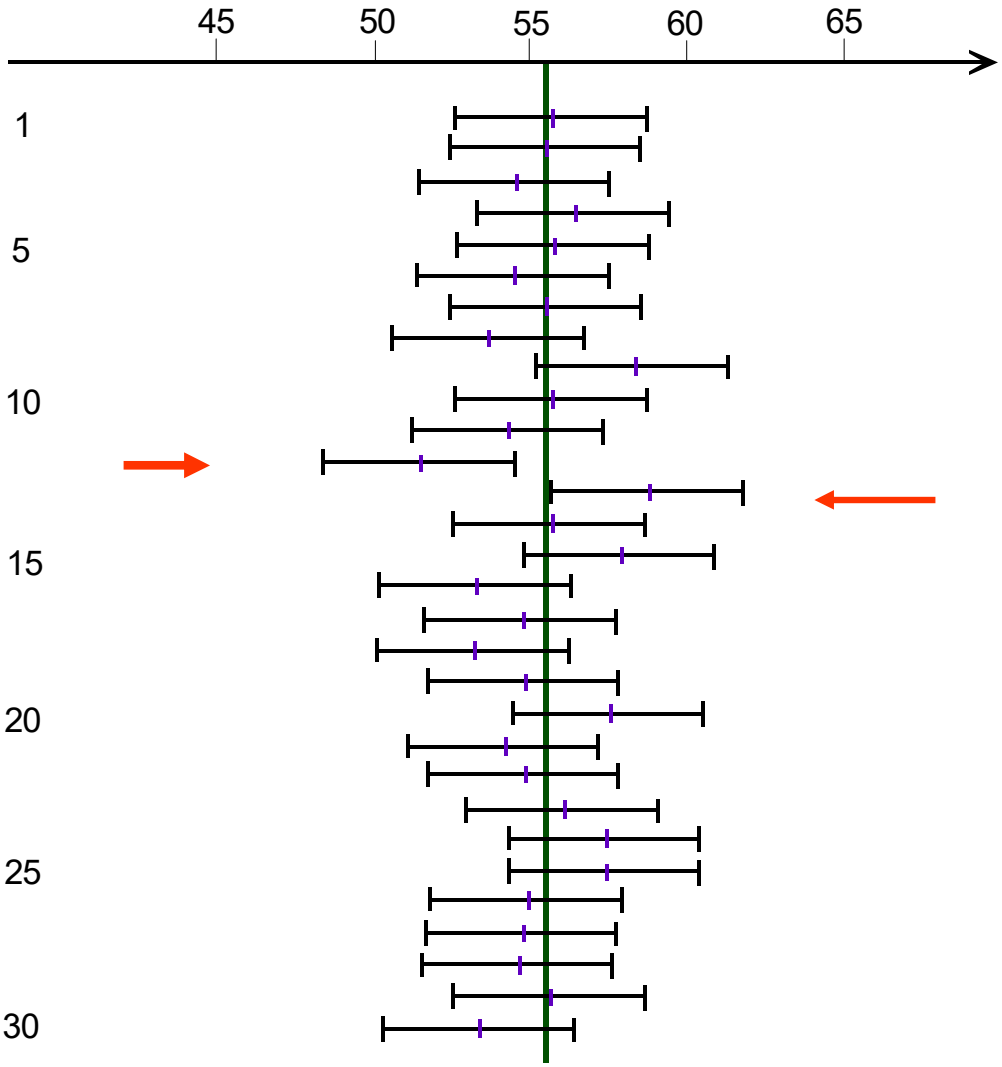
Z.B.
Arithmetische Mittelwert



Kleine Stichproben aus einem großen Datensatz
Lipid-Screening



95% Konfidenzintervalle: Überdecken wahren Wert mit 95% Verfahrenswahrscheinlichkeit



Lipid Screening Daten

30 Stichproben a 100 Probanden

Bedeutung: 95%- Konfidenzintervall

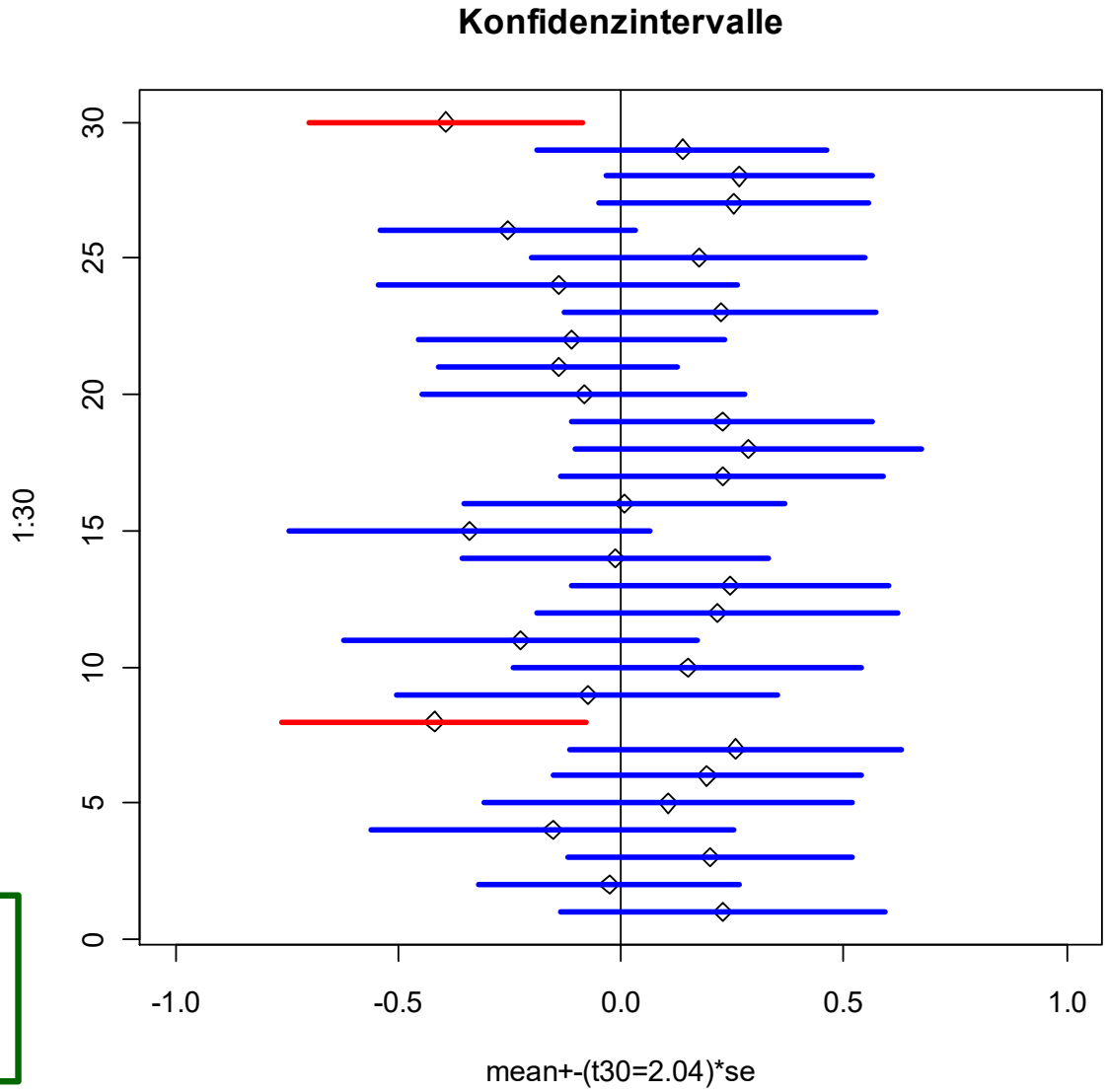
- Bei **wiederholter unabhängiger Anwendung des Konstruktionsverfahrens enthält das (zufällige) Konfidenzintervall in 95% aller Fälle den „wahren“ Wert**.

- Im gegebenen Fall liegt jeweils der wahre Wert
  im Intervall oder auch nicht !?!

- Wahrscheinlichkeitsaussagen darüber, ob der wahre Wert in einem gegebenen Intervall liegt, können nur Bayesianer formulieren...

Simulation Konfidenzintervalle mittels t-Verteilung für N=30 Stichproben aus N(0,1)

5% irreführende 95% Konfidenzintervalle sind nach Konstruktion garantiert!



Breite des Intervalls
~ 1/√N

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**!*

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 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

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 = 0.52
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.203  0.503
sample estimates:
prop 1 prop 2
  0.65   0.50
```

Schätzen Sie p !

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 = 0.0034
alternative hypothesis: two.sided
95 percent confidence interval:
 0.049  0.251
sample estimates:
prop 1 prop 2
  0.65   0.50
```

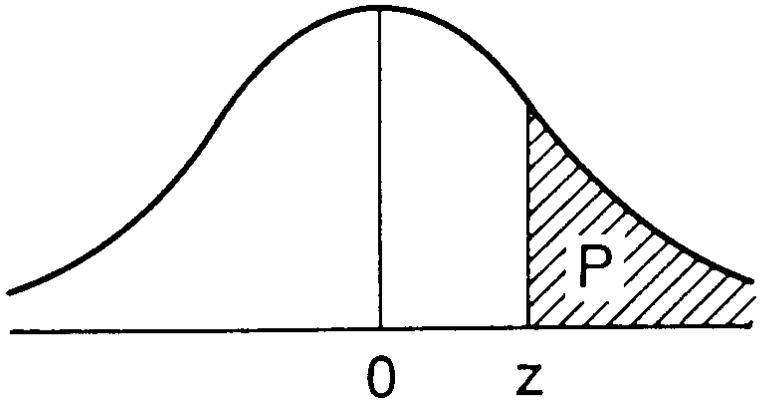
Logik des statistischen Tests III

Konstruktion eines Tests:

- Wähle geeignetes Differenzmaß - **Teststatistik**
- Bestimme deren **Zufallsverteilung unter der Nullhypothese**
    - Was erwarten wir wenn die Nullhypothese wahr wäre?
- Lege fest, was extrem heißen soll (z.b. einseitig vs. zweiseitig)
- Mit beobachtetem Wert der Teststatistik Z
  lese p-Wert ab.

Passt beobachtete Teststatistik

zur Erwartung unter der Nullhypothese?



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.

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.

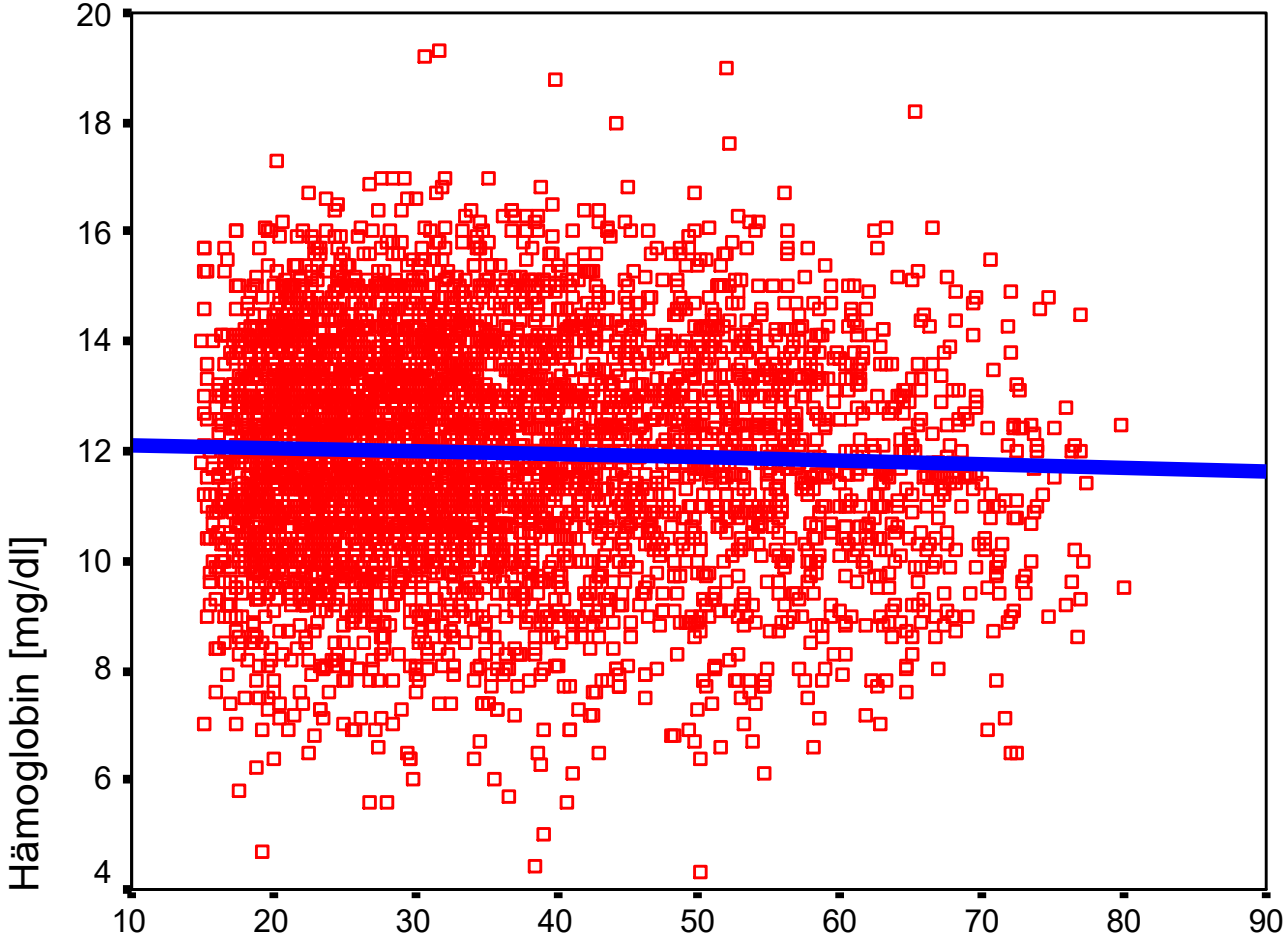
Logik des statistischen Tests V

„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..!“

Evidenzforderungen
Kontrolle Fehler 1. und 2. Art

Was bestimmt die Fallzahl?

p-Wert Prozedur gegen Neyman-Pearson Entscheidungsprozedur

- **p-Wert Prozedur:**
- **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

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

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^* berechen kann)

Darauf basierend bestimmt er die nötige Fallzahl.

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

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

bzw. seine

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

Beispiel: Geschmackstest 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)

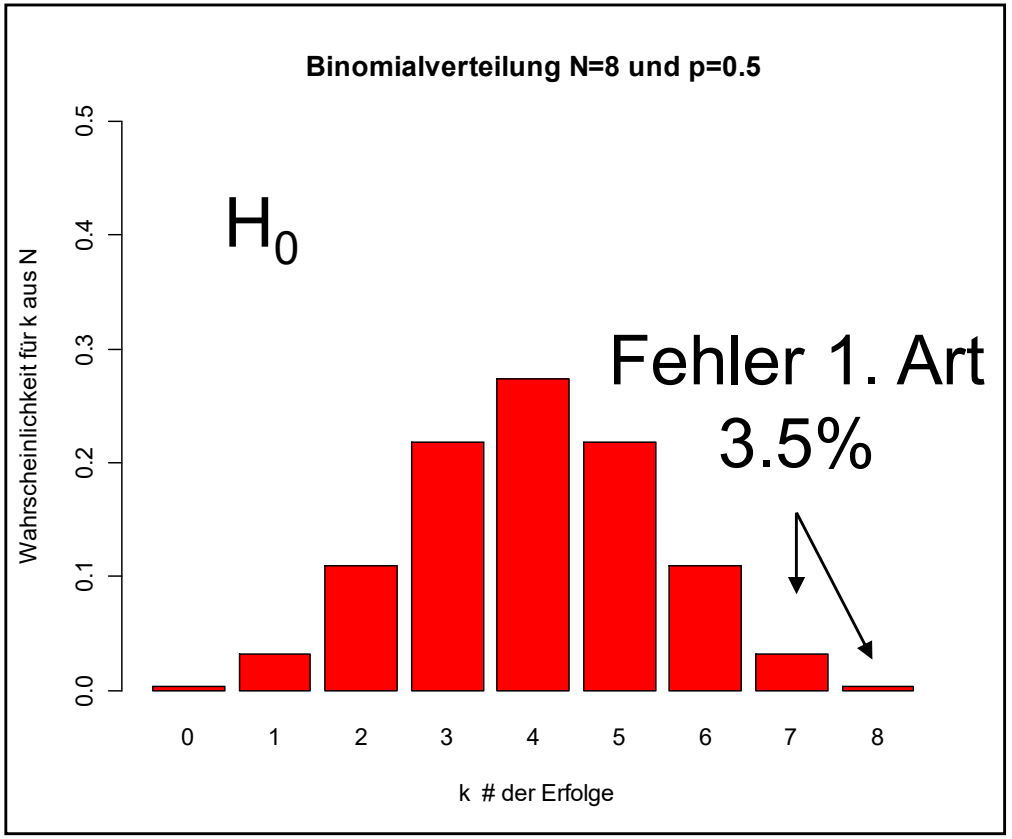
Entscheidungsregel: Sei k die Anzahl der Treffer bei 8 Versuchen

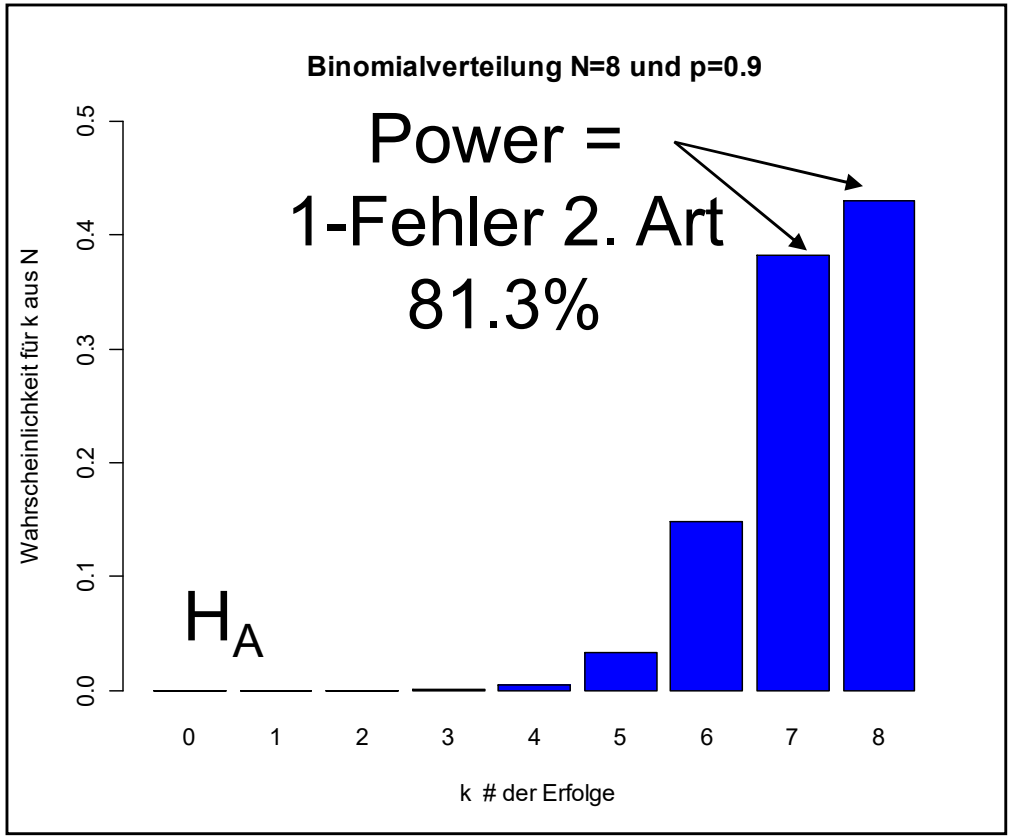
Lehne H_0 ab wenn k= 7 oder 8, sonst behalte H_0 bei.

Binomial-Verteilung

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

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





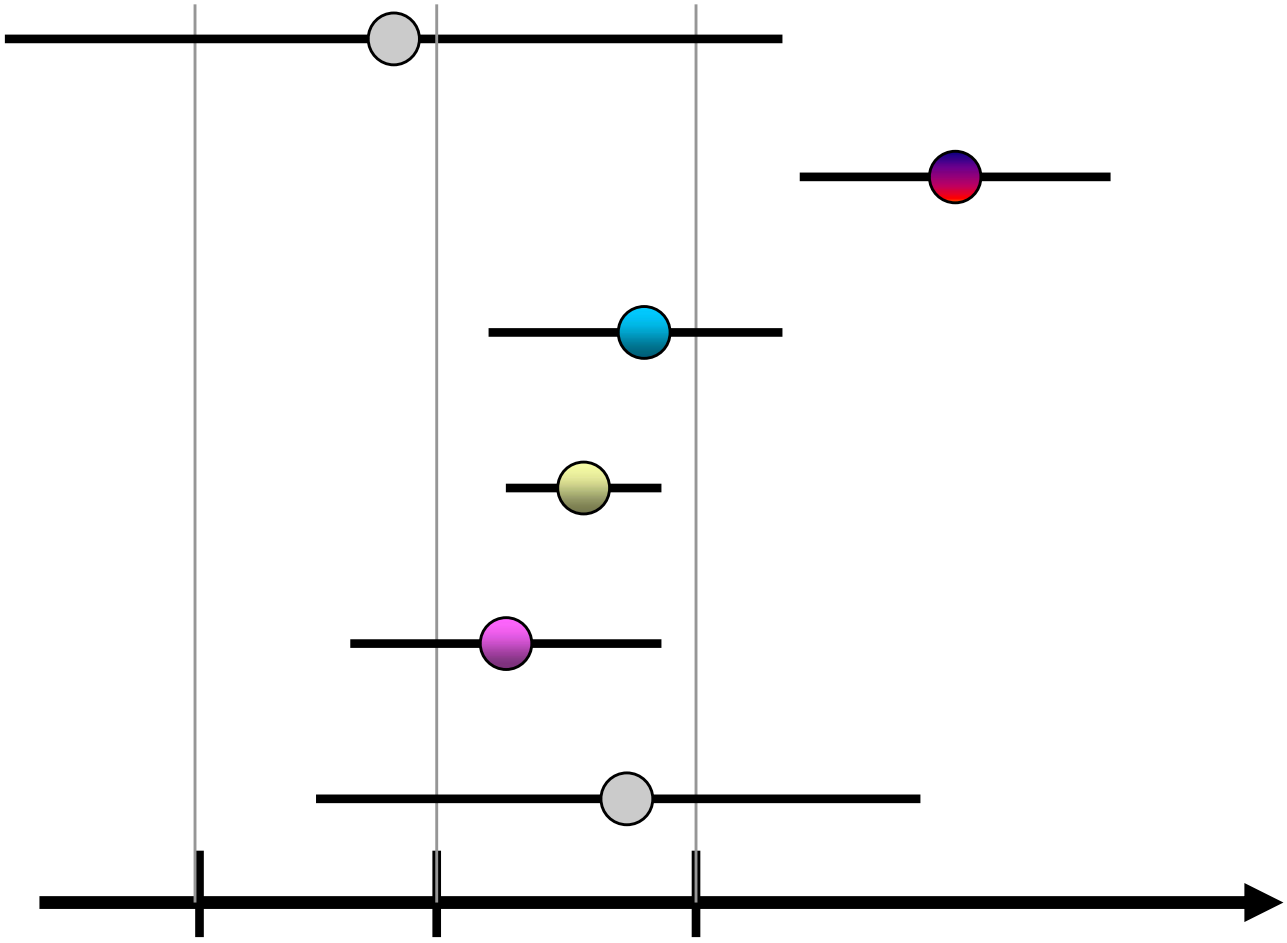
Was haben Konfidenzintervalle und Test miteinander zu tun?

Eigentlich…

Meine Meinung: Eigentlich…

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

Punktschätzer & Konfidenzintervall ≈ statistischer Test + quantitative Information



inkonklusiv

relevant besser

besser
unklar, ob relevant besser

besser,
aber nicht relevant besser

weder relevant besser
noch relevant schlechter

nicht relevant schlechter

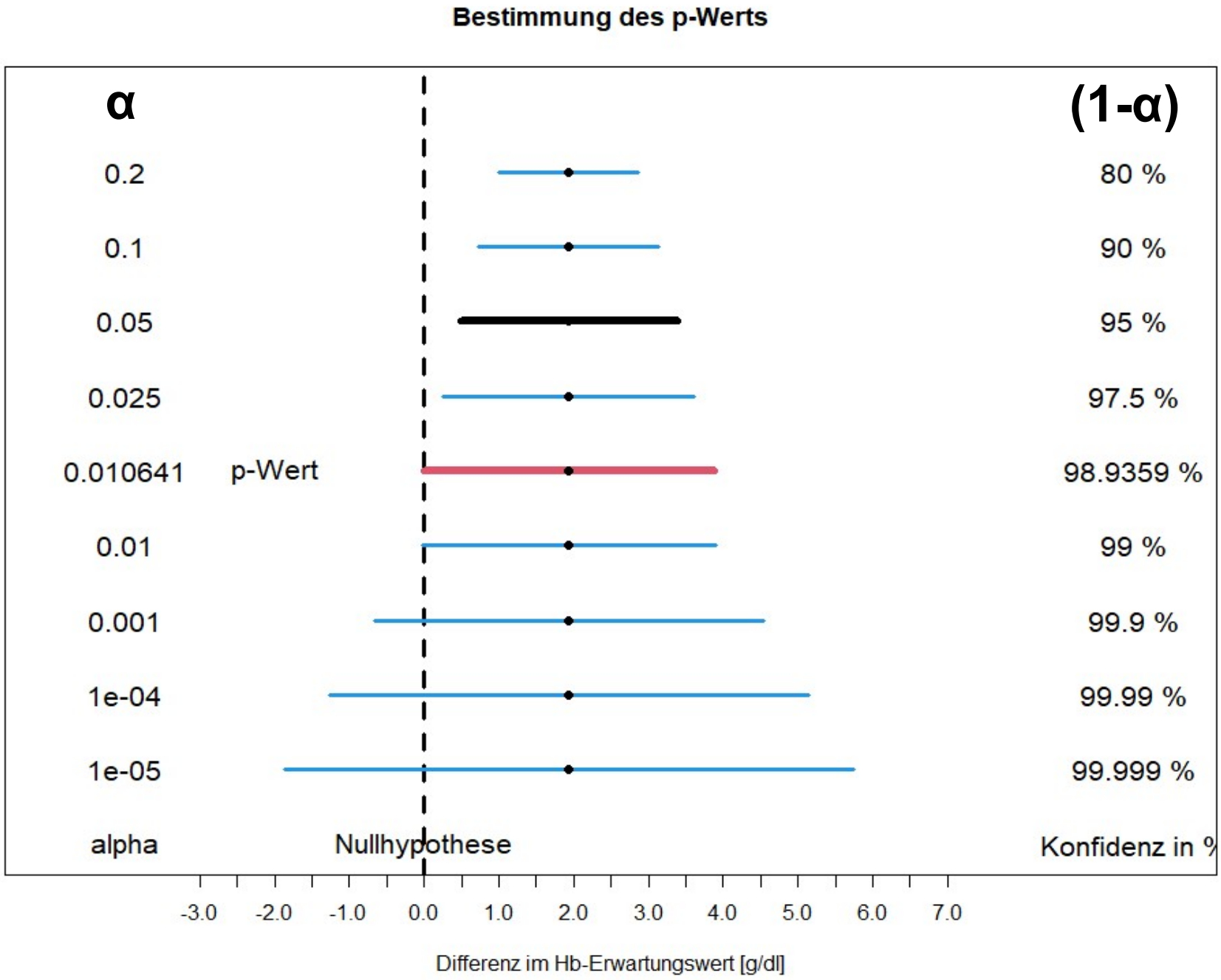
Schätzer mit Konfidenzintervall

p-Wert zu einem Konfidenzintervallverfahren

- Definiere interpretierbare Skala für die Effektsize

- Verwende **(1-α)-**Konfidenzintervall-Verfahren
  - Berichte standardmäßig zweiseitige 95% CIs

- **Braucht man einen p-Wert:**
  - **p-Wert = das α, so dass die Nullhypothese auf dem Rand des (1-α) Konfidentintervalls liegt.**


Bestimmung des p-Werts
α
(1-α)
0.2
80 %
0.1
90 %
0.05
95 %
0.025
97.5 %
0.010641
p-Wert
98.9359 %
0.01
99 %
0.001
99.9 %
1e-04
99.99 %
1e-05
99.999 %
alpha
Nullhypothese
Konfidenz in %
-3.0
-2.0
-1.0
0.0
1.0
2.0
3.0
4.0
5.0
6.0
7.0
Differenz im Hb-Erwartungswert [g/dl]

Fallzahlen

Welche Evidenz wird angestrebt?

- **Wie überzeugend sollen die Daten sein?**
  - Erster Hinweis – Rechtfertigung Folgestudie (Phase I/II)
  - Definitive praxisändernde Klärung einer klinischen Frage (Phase III)

- **Schätzproblem oder Testproblem?** **(eigentlich dasselbe…)**
  - Schätzen eines Interventions-Effekts mit 95%-Konfidenzintervall
  - Testen einer Statistischen Hypothese über den zugrundeliegenden Interventions-Effekt

Evidenzforderungen Schätzproblem

- Spezifiziere **Konfidenzniveau (1-α)**
- Spezifiziere **Präzision = angestrebte Breite** des (1-α) Konfidenzintervals
- Spezifiziere ggfls. angestrebte Wahrscheinlichkeit, dass das realisierte Konfidenzintervall diese Präzision erfüllt.

Evidenzforderungen Tests

- Spezifiziere **Nullhypothese** θ_0
- Spezifiziere **Signifikanzniveau α**
- Spezifiziere **spezifische Alternativehypothese** θ_A und
  weitere nötige Details ζ des Planungsszenario
- Spezifiziere **angestrebte Power = 1-β** d.h.
  Die Wahrscheinlichkeit, dass Test signifikant zu α ausfällt falls Planungsszenario zutrifft

Good conceptual practice I

Empfehlung: Testen und Schätzen konsistent und dual

- Beschreibe **Interventions-Effekt** auf einer **interpretierbaren Skala**
  - Granularität: minimaler klinisch bedeutsamer Unterschied

- **Statistische Hypothesen** entsprechen spezifischen Werten auf dieser Skala

- **Wähle Differenzskala so, dass Effekt-Schätzer (asymptotisch) normalverteilt.**
- **Dann einfache Formeln…**

Good conceptual practice II

Empfehlung: Testen und Schätzen konsistent und dual

- **Leite Test von zweiseitigem Konfidenzintervallverfahren ab!**

- Zweiseitiger Test
  - signifikant zum Niveau α
  - *genau dann, wenn*
  - **Nullhypothese nicht im zweiseitigen (1- α) CI**

Generische Fallzahl - Monsterformel

γ	$z_{1-\gamma}$
0,0005	3,291
0,001	3,090
0,005	2,576
0,01	2,326
0,025	**1,960**
0,05	1,645
0,10	**1,282**
0,20	**0,842**

N = Gesamtfallzahl

$$N = \frac{\left(Z_{1-\alpha/2} + Z_{1-\beta}\right)^2}{\left(\frac{(\theta_A - \theta_0)}{F(\zeta)}\right)^2}$$

$\theta_A - \theta_0$ aufzudeckeder Unterschied auf Differenzskala

$se(\hat{\theta}_N) = \frac{F(\zeta)}{\sqrt{N}}$ für das Planungsszenario ζ

Hier wird der erwartete Standardfehler *se* zu Planung angesetzt.

Monsterformel qualitativ interpretiert – Bessere Evidenz kostet!

- Bessere Kontrolle des Fehlers erster Art, d.h.

  Niedrigeres Signifikanzniveau α → erhöhte Fallzahl

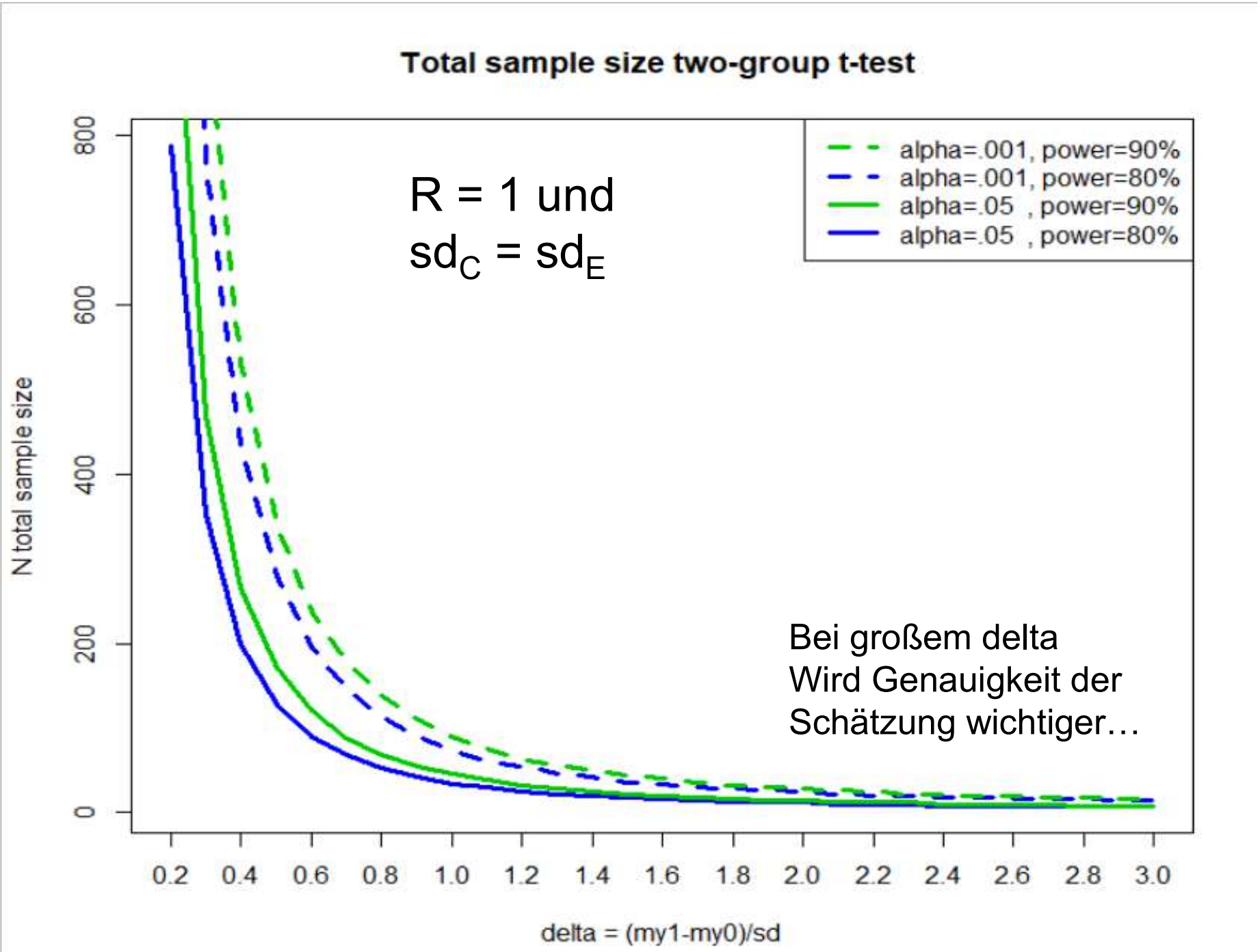
  α = 5% → α = 2.5% 21% mehr Fälle!

- **Höhere Powerforderung (1-β) → erhöhte Fallzahl**

  power = 80% → power = 90% 34% mehr Fälle!

- **Halbierung des aufzudeckenden Unterschieds**

  → Vervierfachung der Fallzahl


Total sample size two-group t-test
alpha=.001, power=90%
alpha=.001, power=80%
alpha=.05 , power=90%
alpha=.05 , power=80%
R = 1 und
$sd_C = sd_E$
Bei großem delta
Wird Genauigkeit der
Schätzung wichtiger…
N total sample size
0
200
400
600
800
0.2
0.4
0.6
0.8
1.0
1.2
1.4
1.6
1.8
2.0
2.2
2.4
2.6
2.8
3.0
delta = (my1-my0)/sd

Fragen?

Jetzt haben wir eine Pause verdient…