Fallzahlplanung

Dr. Dirk Hasenclever

Institut für Medizinische Informatik, Statistik & Epidemiologie (IMISE),
&
Zentrum für Klinische Studien Leipzig, Universität Leipzig

dirk.hasenclever@imise.uni-leipzig.de

Welche Evidenz wird angestrebt?

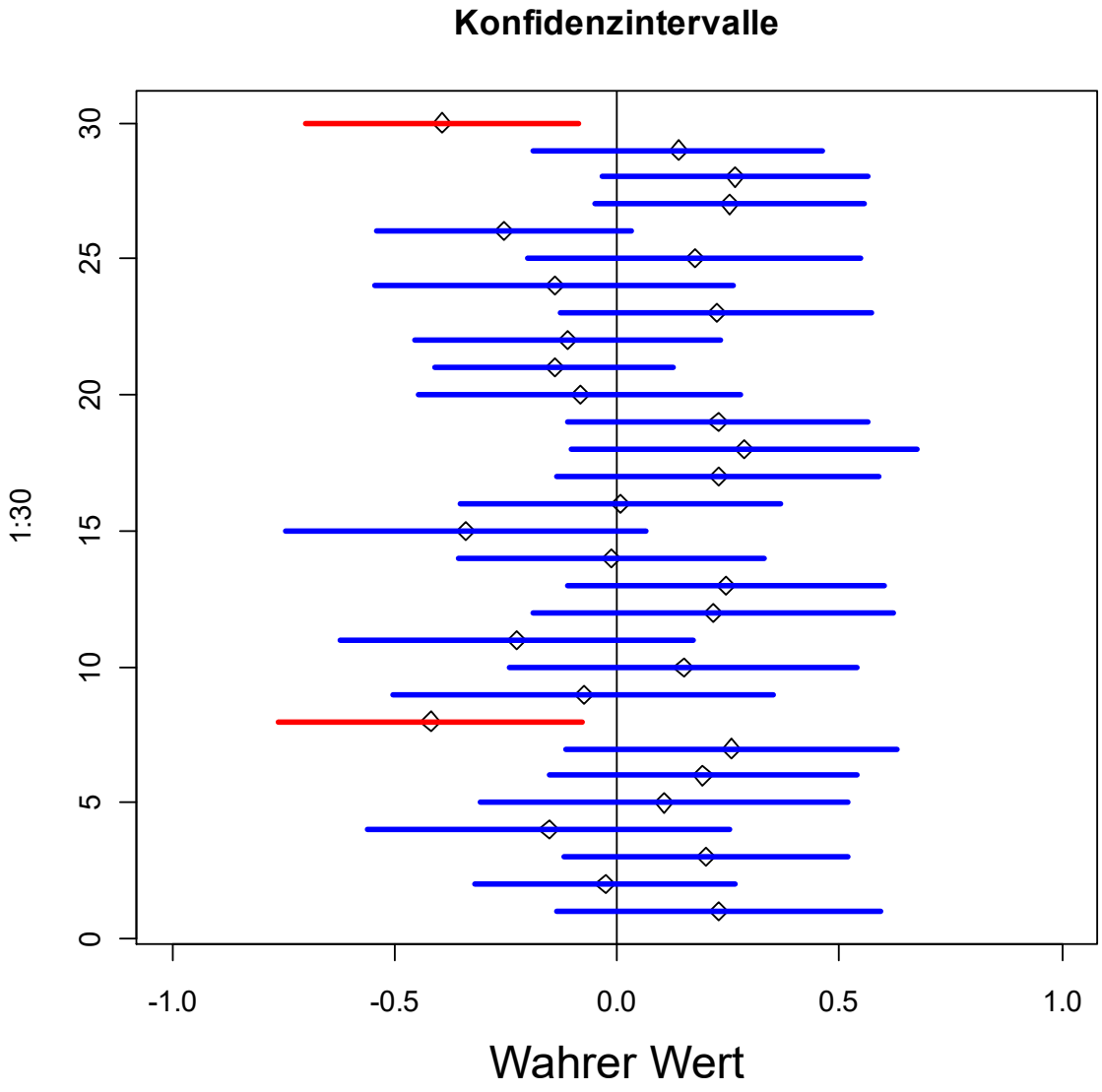
- Voraussetzung: Klare Formulierung des Studienziels:
  - Was genau soll gezeigt werden?
- **Wie überzeugend sollen die Daten sein?**
  - Erster Hinweis – Rechtfertigung Folgestudie (Phase I/II)
  - Definitive praxisändernde Klärung einer klinischen Frage (Phase III)
- **Schätzproblem oder Testproblem?** **(eigentlich dasselbe…)**
  - Schätzen eines Interventions-Effekts mit 95%-Konfidenzintervall
  - Testen einer Statistischen Hypothese über den zugrundeliegenden Interventions-Effekt

Bedeutung: 95%- Konfidenzintervall

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

5% irreführende 95% Konfidenzintervalle sind nach Konstruktion garantiert!

Breite des Intervalls ~ $1/\sqrt{N}$



Dualität von Testen und Konfidenzintervallschätzung

Tests von Punkthypothesen lehnt H0 ab

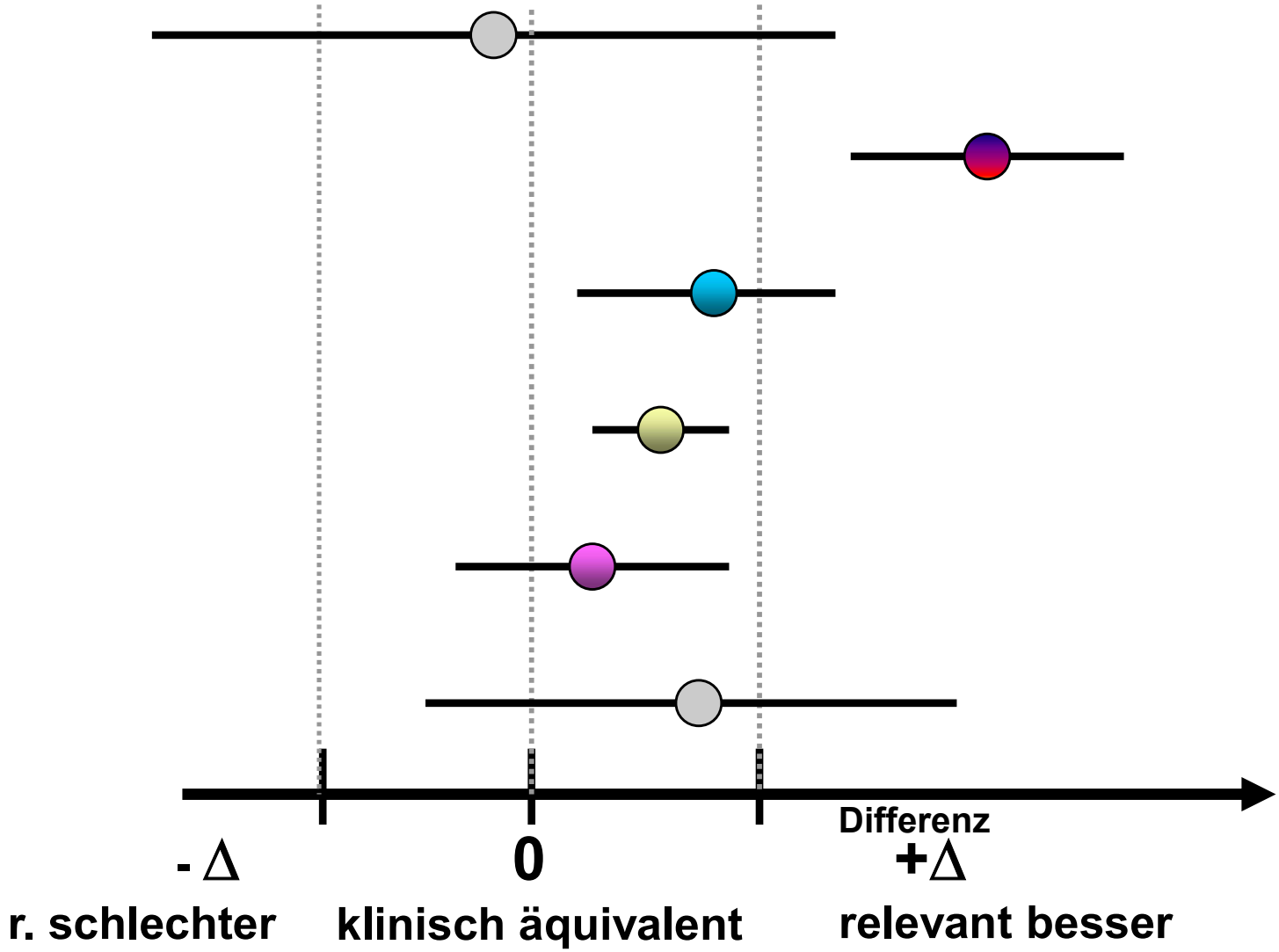
Tests von Punkthypothesen lehnt H0 nicht ab

Tests von Punkthypothesen lehnt H0 ab

θ in Ω

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

Effektschätzer auf Differenzkala mit Konfidentintervallen



inkonklusiv

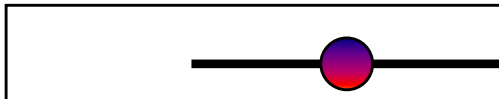
qualifiziert besser

besser
unklar, ob relevant besser

besser,
aber nicht relevant besser

weder relevant besser
noch relevant schlechter

nicht relevant schlechter



Schätzer mit Konfidenzintervall

Good conceptual practice I

Empfehlung: Testen und Schätzen konsistent und dual

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

- Statistische Hypothesen entsprechen spezifischen Werten auf dieser Skala, z.B.
  - Studie auf Differenz: $H0: \theta = 0$
  - Nicht Unterlegenheit: $H0: \theta \leq -\Delta$
    (Δ ist vorspezifizierte Toleranz)
  - Qualifizierte Überlegenheit: $H0: \theta \leq +\Delta$
    (Δ ist vorspezifizierte Erfolgsschwelle)

Good conceptual practice II

Empfehlung: Testen und Schätzen konsistent und dual

- Leite Test aus zweiseitigem Konfidenzintervallverfahren ab
  - Zweiseitiger Test
    - signifikant zu α *genau dann, wenn*
      Nullhypothese liegt nicht im zweiseitigen (1- α) -Konfidenzinterval
  - Einseitiger Test
    - signifikant zu α/2 *genau dann, wenn*
      Nullhypothese **auf der erwarteten Seite** außerhalb des zweiseitigen (1- α) -Konfidenzinterval liegt
    - Standard: „Gerne einseitiger Test, aber zu α/2 …“

Good conceptual practice II

Empfehlung: Testen und Schätzen konsistent und dual

Wähle Differenzskala so, dass Effekt-Schätzer (asymptotisch) normalverteilt

$$\hat{\theta}_N \sim \mathrm{N}\left(\theta\,;\ se\left(\hat{\theta}_N\right)^2\right)$$

Zweiseitiges (1-α) - Konfidenzinterval

$$\left[\hat{\theta}_N - Z_{1-\alpha/2} \cdot \widehat{se}\left(\hat{\theta}_N\right);\ \ \hat{\theta}_N + Z_{1-\alpha/2} \cdot \widehat{se}\left(\hat{\theta}_N\right)\right]$$

Duale Teststatistik (evtl. Korrektur mit t-Verteilung)

$$Z = \frac{\left(\hat{\theta}_N - \theta_0\right)}{\widehat{se}\left(\hat{\theta}_N\right)} \ \sim^{H_0} \ \mathrm{N}(0\,;1)$$

Evidenzforderungen

- Schätzproblem:
  - Spezifiziere Konfidenzniveau 1-α
  - Spezifiziere angestrebte Breite des 1-α Konfidenzinterval
  - Spezifiziere angestrebte Wahrscheinlichkeit, dass realisiertes Konfidenzintervall diese Bedingung erfüllt.
- Testproblem
  - Spezifiziere Nullhypothese θ_0 und Signifikanzniveau α
  - Spezifiziere spezifische Alternative θ_A und weitere nötige Details ζ des Planungsszenario
  - Spezifiziere angestrebte Power = 1-β d.h. Wahrscheinlichkeit, dass Test signifikant zu α ausfällt falls Planungsszenario zutrifft

Generische Fallzahlformel

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.

Herleitung*

$$Z = \frac{(\hat{\theta}_N - \theta_0)}{\widehat{se}(\hat{\theta}_N)} \sim^{H_0} \mathrm{N}(0\,;1) \qquad Z = \frac{(\hat{\theta}_N - \theta_0)}{\widehat{se}(\hat{\theta}_N)} \sim^{H_A} \mathrm{N}(\theta_A - \theta_0; 1)$$

$$\text{Power} = 1-\beta := pr\left(\frac{(\hat{\theta}_N-\theta_0)}{\widehat{se}(\hat{\theta}_N)} > Z_{1-\alpha/2}\middle| H_A\right) =$$

$$pr\left(\frac{(\hat{\theta}_N - \theta_A)}{\widehat{se}(\hat{\theta}_N)} > Z_{1-\alpha/2} - \frac{(\theta_A-\theta_0)}{\widehat{se}(\hat{\theta}_N)}\middle| H_A\right) =$$

$$\Phi\left(\frac{(\theta_A-\theta_0)}{\widehat{se}(\hat{\theta}_N)} - Z_{1-\alpha/2}\right)$$

Φ = Verteilungsfunktion der Standardnormalverteilung

$$Z_{1-\beta} = \frac{(\theta_A-\theta_0)}{\widehat{se}(\hat{\theta}_N)} - Z_{1-\alpha/2} = \sqrt{N}\frac{(\theta_A-\theta_0)}{F(\zeta)} - Z_{1-\alpha/2}$$

Qualitativ

- Bessere Kontrolle des Fehlers erster Art
  - Niedrigeres Signifikanzniveau α → erhöhte Fallzahl
- Höhere Powerforderung (1-β) → erhöhte Fallzahl
- Halbierung des aufzudeckenden Unterschieds
  → Vervierfachung der Fallzahl

Was ist F?

- Die Funktion F spezifiziert den erwarteten Standardfehler

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

  - hängt **nicht** mehr von N ab,
  - sondern in der Regel von θ_0 und θ_A,
  - **und zusätzlichen Annahmen** des **Planungsszenarios ζ** und
  - der gewählten Differenzskala

Differenz Erwartungswerte
(Zweigruppenvergleich – metrische Skala)

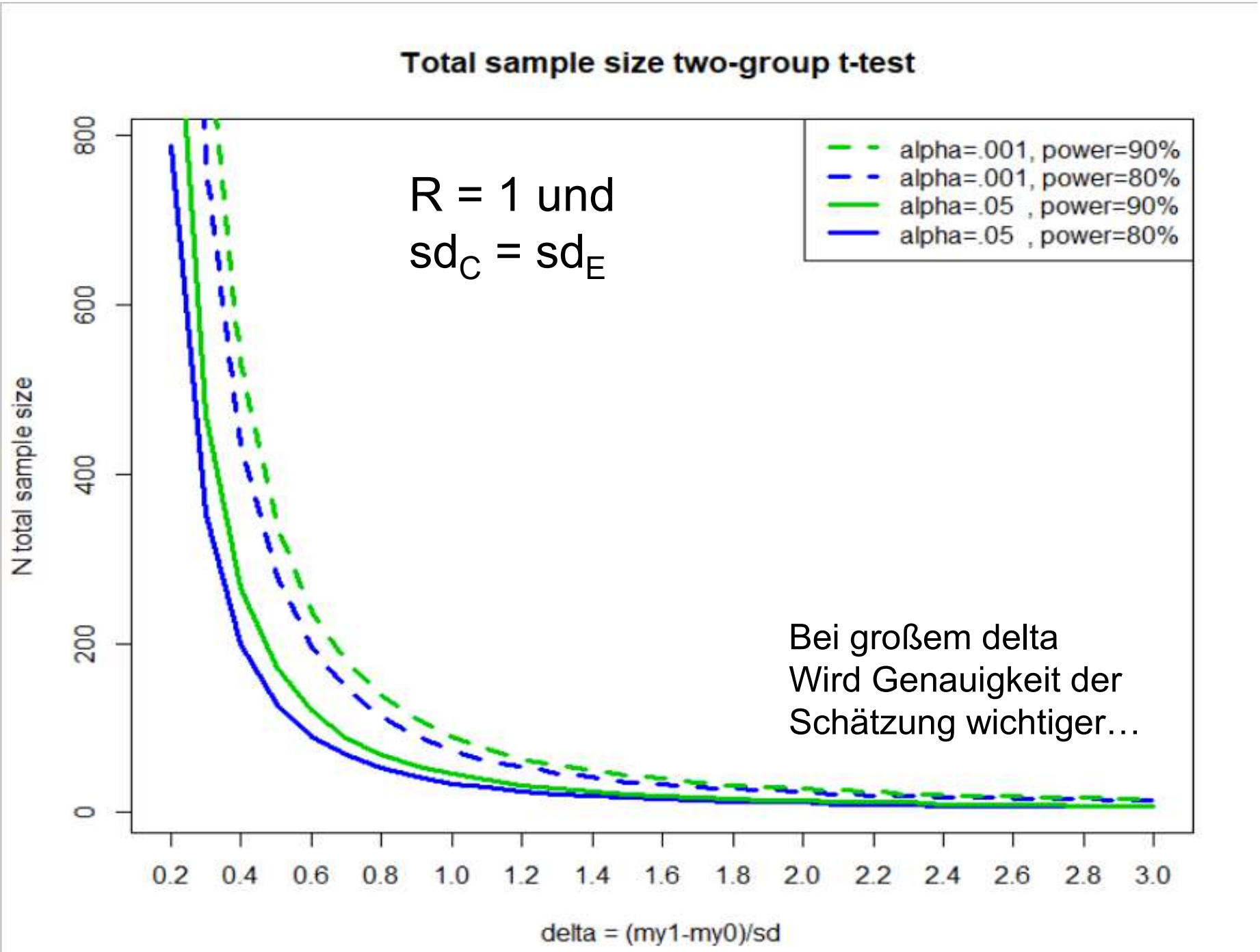
$R := \frac{N_E}{N_C}$ Randomisationsverhältnis

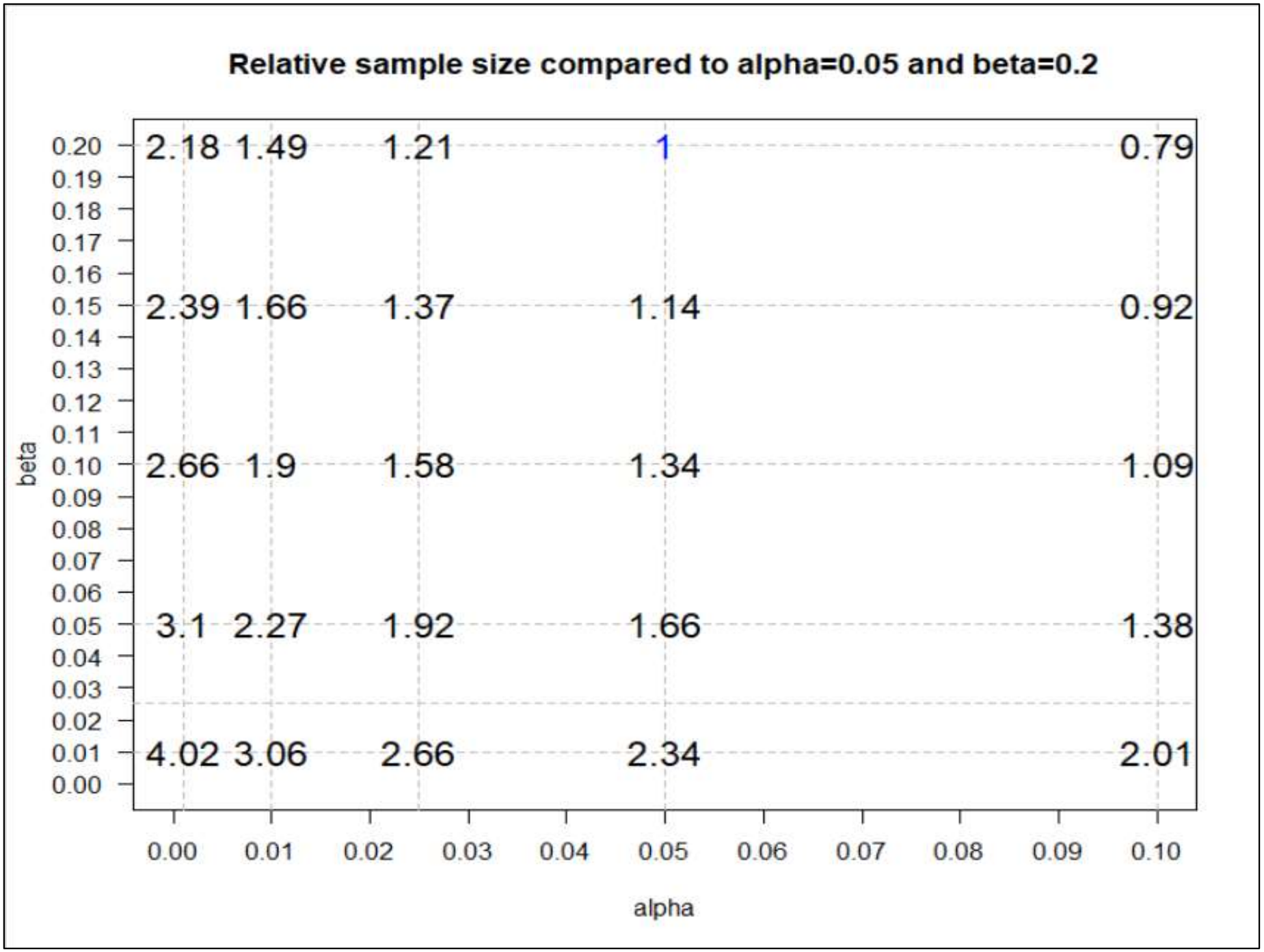
$$\theta = E(M_E) - E(M_C)$$

$$\hat{\theta} = \overline{M_E} - \overline{M_C}$$

$$F(\zeta) = \sqrt{\frac{sd_C^2}{1/(1+R)} + \frac{sd_E^2}{R/(1+R)}}$$

Nötig ζ : Standardabweichung in Vergleichsgruppen
Adjustierung: Lineare Regression


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


Relative sample size compared to alpha=0.05 and beta=0.2
beta
alpha
0.20
0.19
0.18
0.17
0.16
0.15
0.14
0.13
0.12
0.11
0.10
0.09
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0.00
2.18 1.49 1.21 1 0.79
2.39 1.66 1.37 1.14 0.92
2.66 1.9 1.58 1.34 1.09
3.1 2.27 1.92 1.66 1.38
4.02 3.06 2.66 2.34 2.01
0.00 0.01 0.02 0.03 0.04 0.05 0.06 0.07 0.08 0.09 0.10

Differenz Erfolgswahrscheinlichkeit
(Zweigruppenvergleich - Wahrscheinlichkeitsskala)

$R := \frac{N_E}{N_C}$ Randomisationsverhältnis

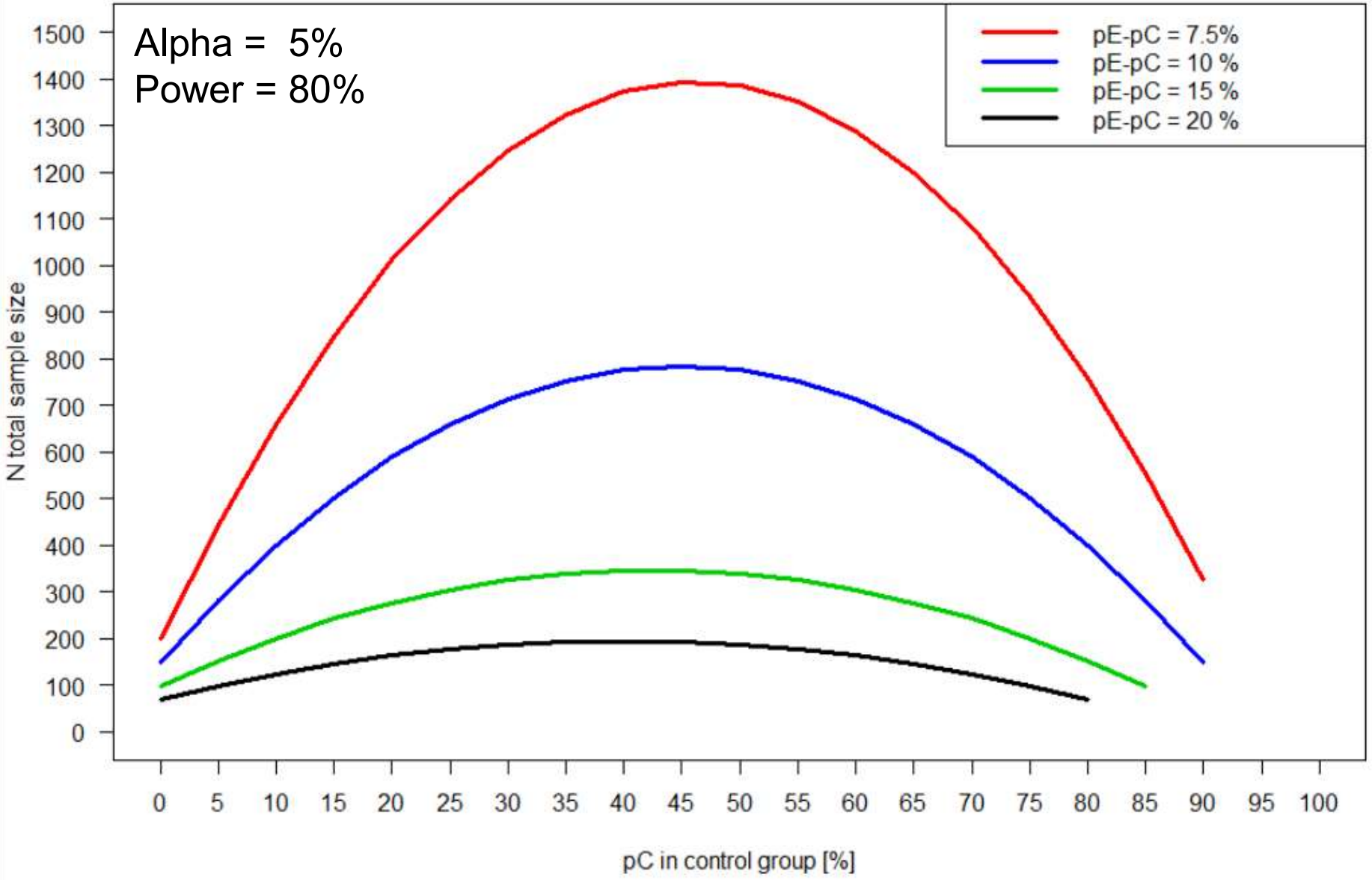
$$\theta = E(S_E) - E(S_C) = p_E - p_C$$

$$\hat{\theta} = \widehat{p_E} - \widehat{p_C}$$

$$F(\zeta) = \sqrt{\frac{p_C \cdot (1 - p_C)}{1/(1+R)} + \frac{p_E \cdot (1 - p_E)}{R/(1+R)}}$$

Nötig ζ : Absolute Erfolgswahrscheinlichkeiten nicht nur Differenz
Adjustierung: Nur indirekt über logistische Regression

Total sample size Chi2-test



logOddsRato Erfolgswahrscheinlichkeiten
(Zweigruppenvergleich – logit Skala)

$R := \frac{N_E}{N_C}$ Randomisationsverhältnis Skala: $logit(p) := \log({}^{p}/_{(1-p)})$

$$\theta = \log(OR) = \log\left(\frac{p_E \cdot (1 - p_C)}{p_C \cdot (1 - p_E)}\right) = logit(p_E) - logit(p_C)$$

$$\hat{\theta} = \log\left(\frac{\widehat{p_E} \cdot (1 - \widehat{p_C})}{\widehat{p_C} \cdot (1 - \widehat{p_E})}\right)$$

$$F(\zeta) = \sqrt{\frac{1}{{p_C}/{(1+R)}} + \frac{1}{{(1-p_C)}/{(1+R)}} + \frac{1}{{p_E \cdot R}/{(1+R)}} + \frac{1}{{(1-p_E) \cdot R}/{(1+R)}}}$$

Nötig ζ : Absolute Erfolgswahrscheinlichkeiten nicht nur Differenz
Adjustierung: Logistische Regression

logRelativeRisk Erfolgswahrscheinlichkeiten
(Zweigruppenvergleich – log Skala)

$R := \frac{N_E}{N_C}$ Randomisationsverhältnis Skala: $\log(p)$

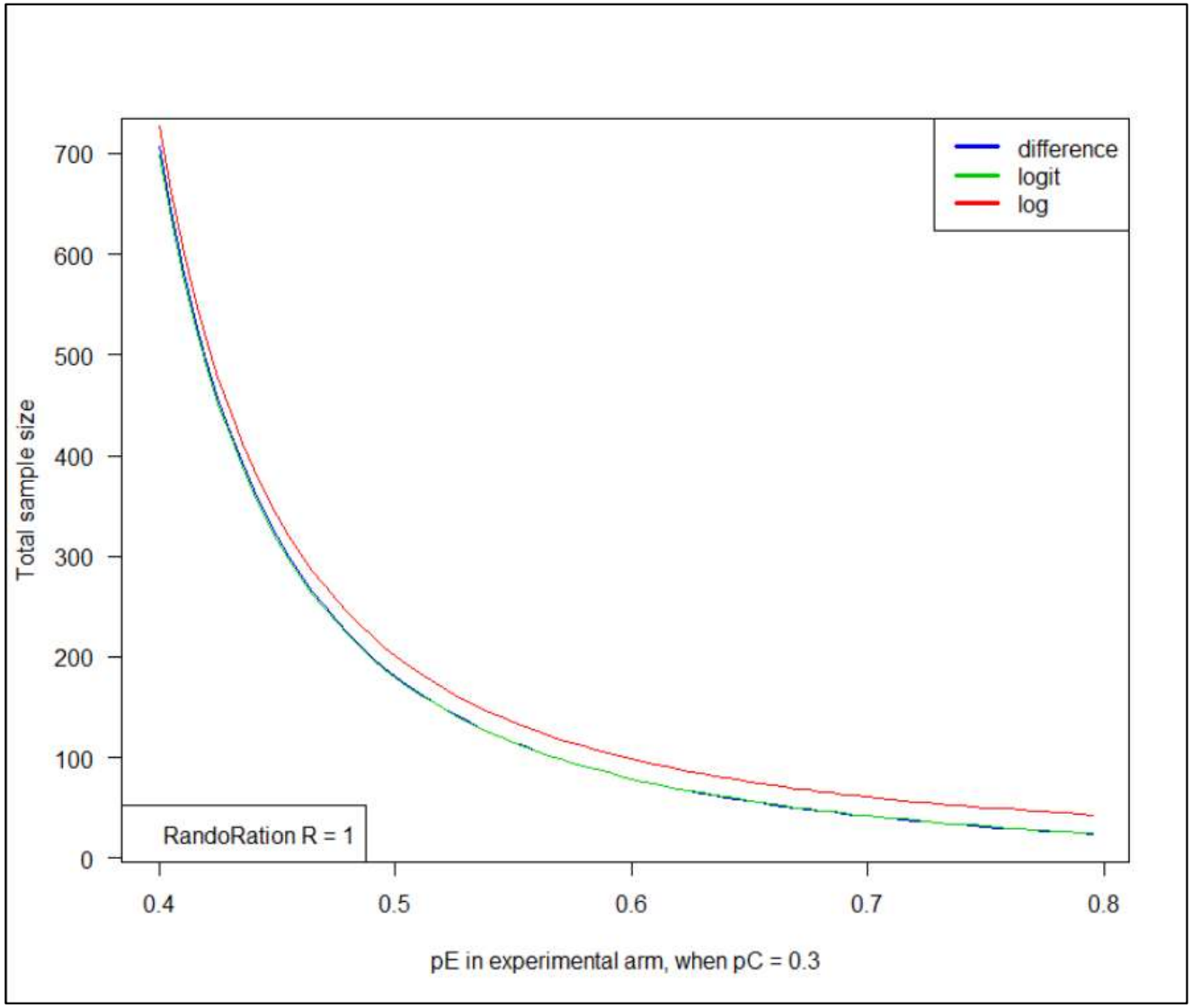
$$\theta = \log(RR) = \log\left(\frac{p_E}{p_C}\right) = log(p_E) - log(p_C)$$

$$\hat{\theta} = \log\left(\frac{\widehat{p_E}}{\widehat{p_C}}\right)$$

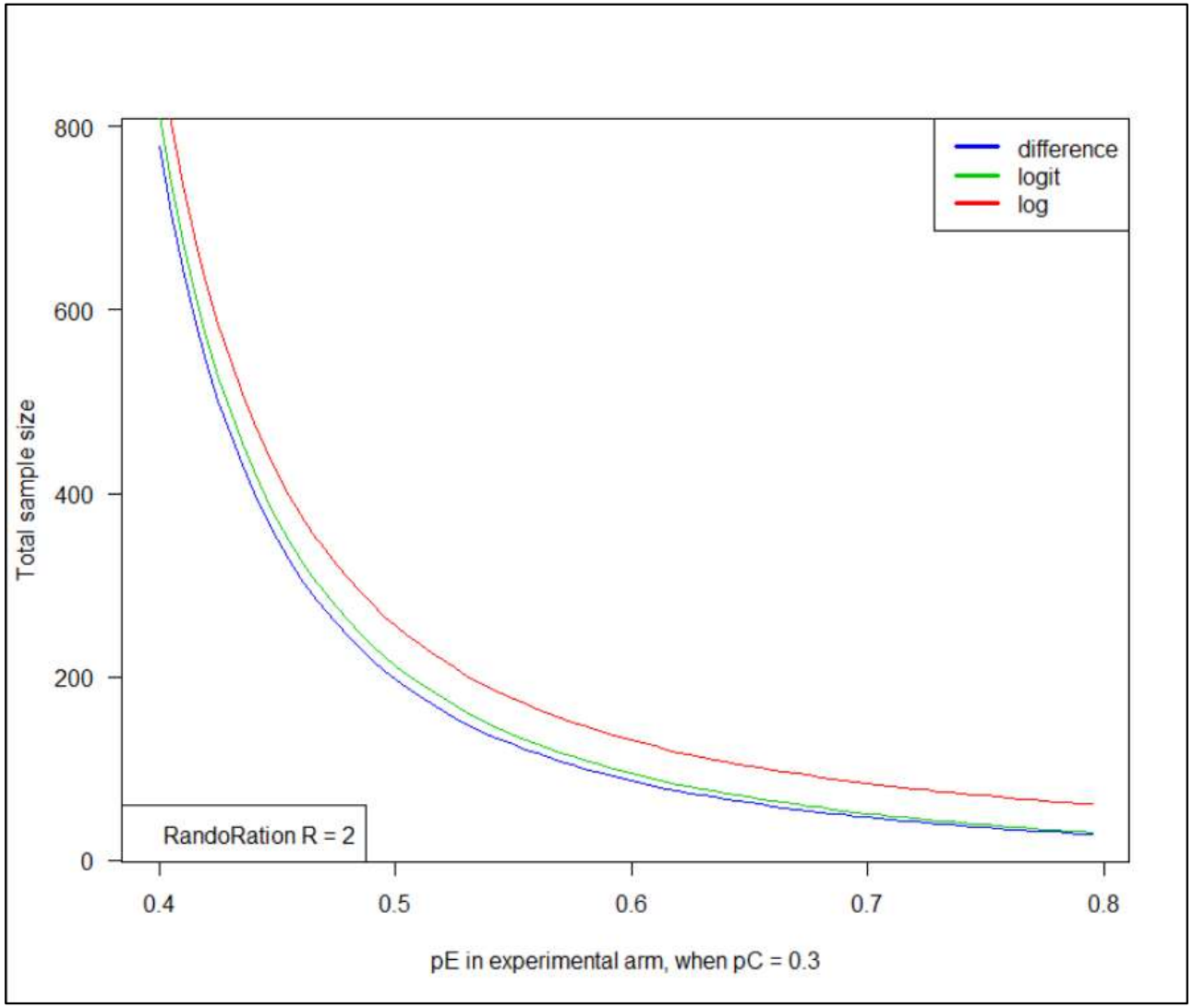
$$F(\zeta) = \sqrt{\frac{1}{p_C/(1+R)} - \frac{1}{1/(1+R)} + \frac{1}{p_E \cdot R/(1+R)} - \frac{1}{R/(1+R)}}$$

Nötig ζ : Absolute Erfolgswahrscheinlichkeiten nicht nur Differenz
Adjustierung: Loglineare Regression

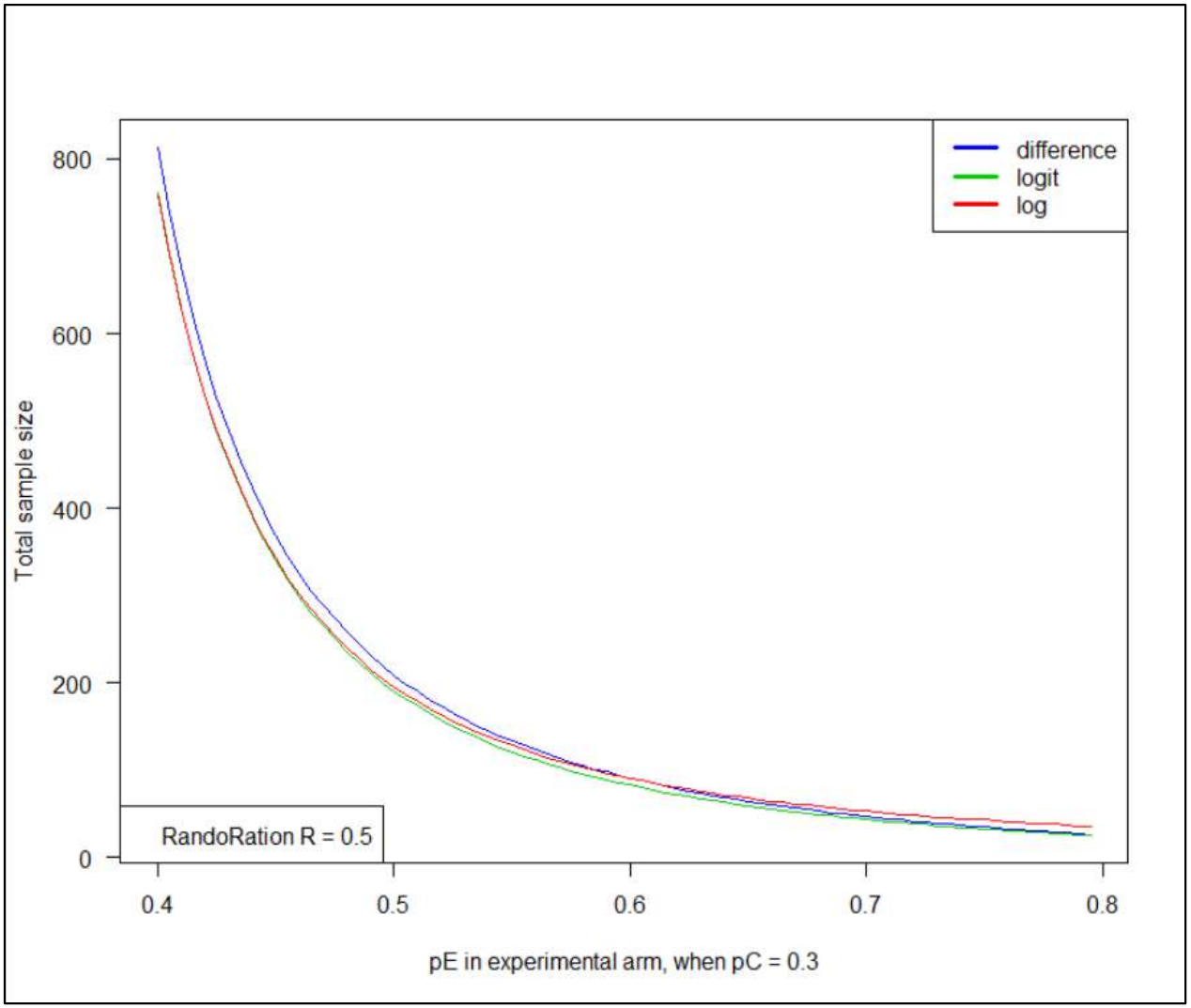
Vergleich der Skalen – Fallzahl R = 1



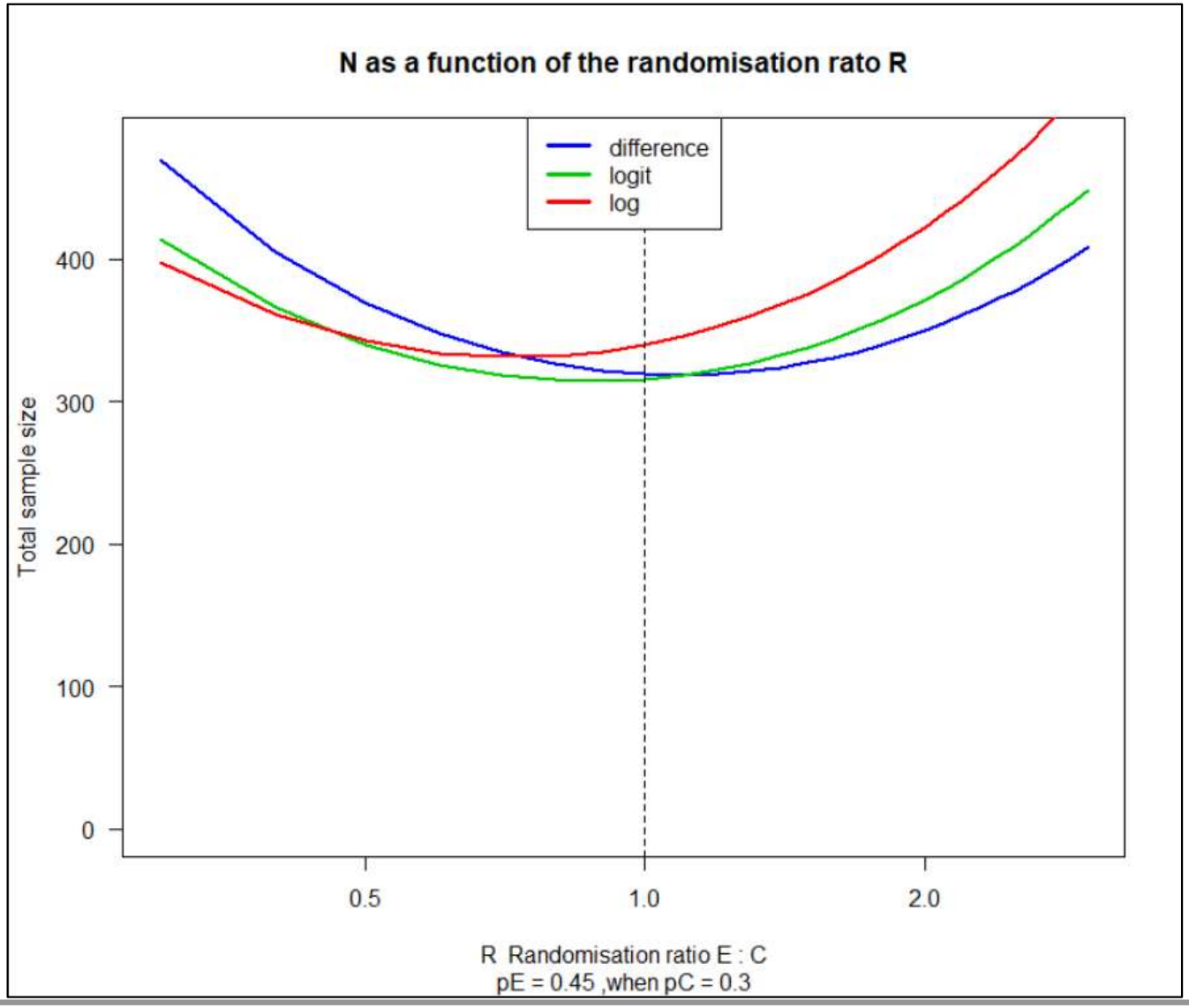
Vergleich der Skalen – Fallzahl R = 2



Vergleich der Skalen – Fallzahl R = 1/2



Vergleich der Skalen – Abhängigkeit von R



Zweigruppen Vergleich bei Time-to-Event Endpunkten

- Problem „zensierte Beobachtungen“´, i.e. Patienten, die während ihrer Beobachtung in der Studie das Ereignis nicht haben.
  - VORSICHT: Informative Zensur

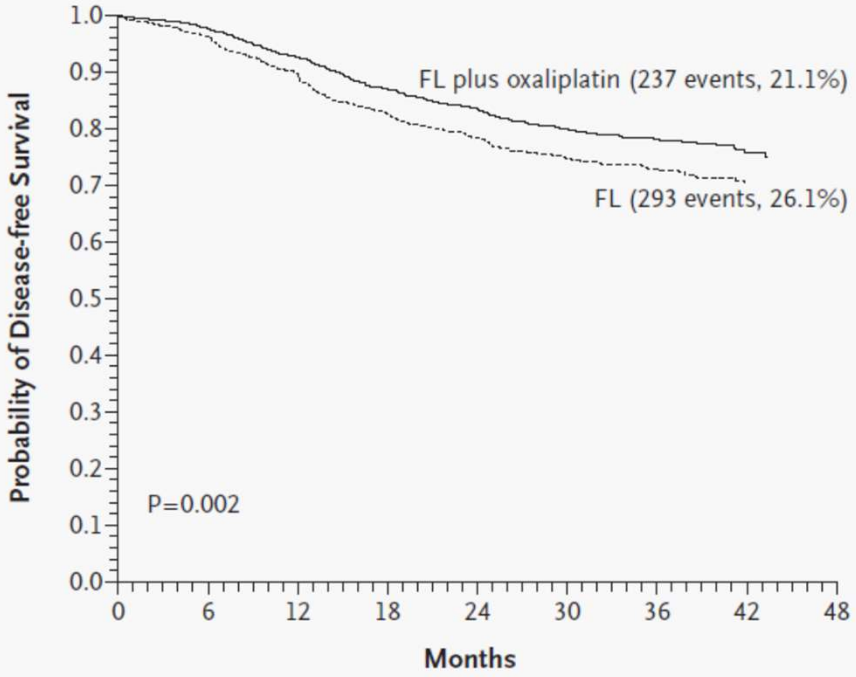
- Information hängt im wesentlichen
  von der Anzahl E der beobachteten Ereignisse
  und nur indirekt von N ab!!!
  - Wenig Information trotz großer Fallzahl bei seltenen Ereignissen, z.B. hohen Heilungsraten.

Zweigruppen Vergleich bei Time-to-Event Endpunkten II

- Wie viele **E** Ereignisse beobachtet werden hängt ab
  - der Form der Time-to-Event Kurven
  - der Dynamik der Rekrutierung
  - Der Qualität und Vollständigkeit des Follow-Ups
- Empfehlung: Simuliere den erwarteten Informationsfluss explizit. Große Gefahr sich Studiendesigns schön zu rechnen
- Für Überschlagsrechnungen:
  - Rate einen „event-treuen“ Zeitpunkt t, sodass im Planungsszenario

$$E \approx \left(1 - S_C(t)\right) \cdot \frac{N}{1+R} + \left(1 - S_E(t)\right) \cdot \frac{R \cdot N}{1+R}$$

Zweigruppen Vergleich bei Time-to-Event Endpunkten III



No. at Risk

FL+oxaliplatin	1123	1086	1023	959	888	663	395
FL	1123	1066	981	903	817	619	356

Figure 1. Kaplan–Meier Estimates of Disease-free Survival in the Group Given Fluorouracil and Leucovorin (FL) and the Group Given FL plus Oxaliplatin, According to the Intention to Treat.

The hazard ratio for recurrence in the group given FL plus oxaliplatin, as compared with the FL group, was 0.77 (95 percent confidence interval, 0.65 to 0.91; P=0.002).

André T et al: N Engl J Med 2004;350:2343-51

- Proportional Hazard Modell
  - Differenzskala ist die log Hazard rato

$$S_{Exp}(t) = S_{Control}(t)^{HR}$$

Für alle t

$$HR = \frac{\log(S_{Exp}(t))}{\log(S_{Control}(t))}$$

Für alle t

$$HR = \frac{\log(0.783)}{\log(0.729)} = 0.778$$

Zweigruppen Vergleich bei Time-to-Event Endpunkten IV

$R := \frac{N_E}{N_C}$ Randomisationsverhältnis Skala: $\log(Hazardratio)$

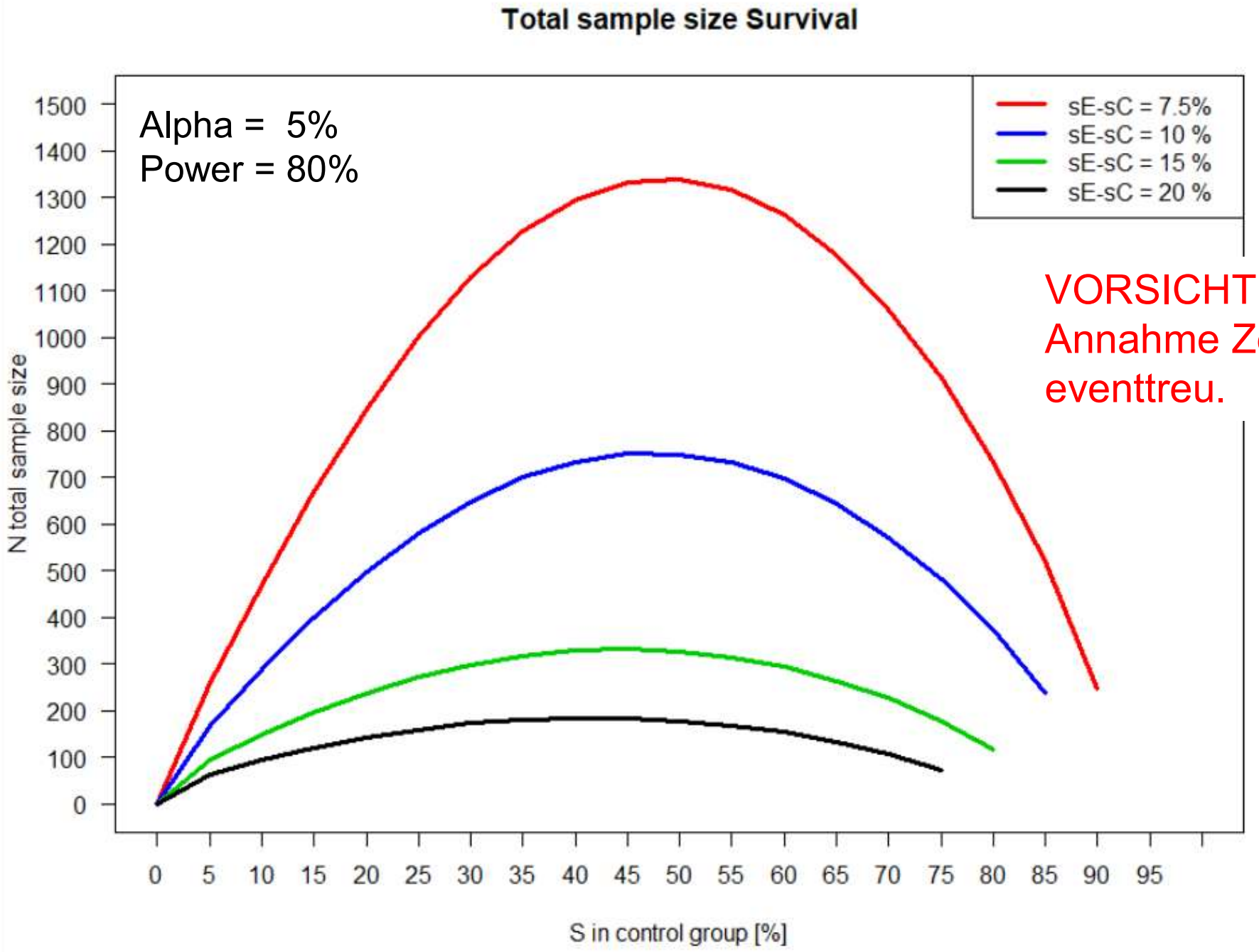
$$\theta = \log(HR)$$

$\hat{\theta} = \log(\widehat{HR})$ aus Proportional Hazard COX Regression

$$F(\zeta) \approx \sqrt{\frac{\frac{(1+R)^2}{R}}{\left(1 - S_C(t)\right) * \frac{1}{1+R} + \left(1 - S_E(t)\right) * \frac{R}{1+R}}}$$

t ist der geratene Event-treue Zeitpunkt

Nötig ζ : Abschätzung der zu beobachtenden Ereignisse
Adjustierung: COX Regression



VORSICHT
Annahme Zeitpunkt eventtreu.