A Symbol of Excellence

Bayesian Methods in Clinical Research

Andy Grieve
VP Innovation Centre

18th March 2015

Reviews of Bayesian Statistics
Medical / Pharmaceutical

ICON
A Symbol of Excellence

2006

STATISTICS IN MEDICINE
Statist. Med. (in press)
Published online in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/sim.2672

WILEY InterScience® DISCOVER SOMETHING GREAT

25 Years Paper Celebrating the 25th Anniversary of *Statistics in Medicine*

Bayesian statistics in medicine: A 25 year review‡

Deborah Ashby*,†

2007

PHARMACEUTICAL STATISTICS
Pharmaceut. Statist. 2007; **6**: 261–281
Published online 22 October 2007 in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/pst.315

WILEY InterScience® DISCOVER SOMETHING GREAT

25 years of Bayesian methods in the pharmaceutical industry: a personal, statistical bummel

30TH ANNIVERSARY PSI STATISTICIANS IN THE PHARMACEUTICAL INDUSTRY LIMITED 1977 – 2007

Andrew P. Grieve*,†,‡
King's College London, Department of Public Health Sciences, London, UK

Personal Use of Bayesian Methods in Pharmaceutical R&D

A Symbol of Excellence

- Pharmaceutical Development
  - Bayesian Hierarchical model for drug stability studies
  - Assessment of Bioequivalence – single and two-stage Bayesian designs
- Pre-Clinical Toxicology
  - Bayesian models incorporating historical control information in carcinogenicity studies
  - Bayesian hierarchical model accounting for litter effects in teratology studies
  - Acute toxicity studies – estimation of LD50
- Clinical Development
  - Model Uncertainty in Crossover Designs
  - Bayesian Adaptive Designs – Phase I and Phase IIb
  - Assessment of Clinical equivalence
  - Bayesian models incorporating historical control information
- Production
  - Acceptance Sampling for Rare Defects – Utilising historical data

Unifying Themes

- Probability of belonging to regions of parameter space
  - "probabilities computed from the Bayesian approach provide more relevant information to decision makers and are easier to interpret"

    (Harrell F and Shih YC. *International Journal of Health Technology Assessment,*2001)

- Model Uncertainties
- Prediction
- Parametrisation
- Priors

Topics that Currently Interest Me

Model-based approaches

Calibration of Bayesian Procedures

Adaptive Designs

Use of Historical Data

Use of Electronic patient records

Banishment of p-values

Prediction

Topics that Currently Interest Me

Model-based approaches

Calibration of Bayesian Procedures

Adaptive Designs

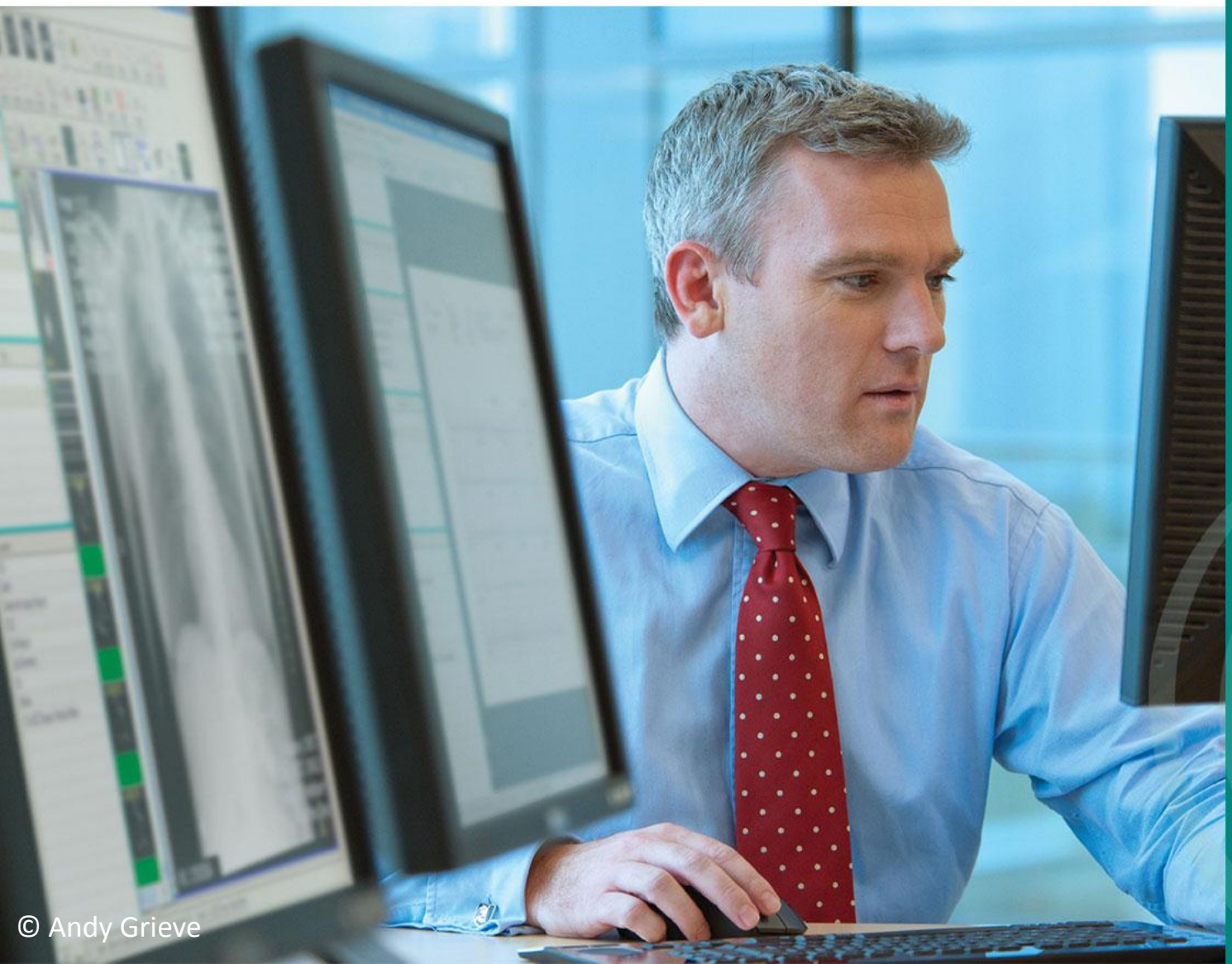
Use of Electronic patient records

Use of Historical Data

Prediction

Banishment of p-values

Calibration of Bayesian Procedures



Guidance for the Use of Bayesian Statistics in Medical Device Clinical Trials

- Requires simulations to assess Bayesian approaches.
- If type I error to large
  - change success criterion (posterior probability)
  - reduce number of interim analyses
  - discount prior information
  - increase sample size
  - altering calculation of type I error
- "the degree to which we might relax the type I error control is a case-by-case decision that depends …. Primarily on the confidence we have in prior information"

Guidelines for Reporting Bayesian Analyses

ICON
A Symbol of Excellence

ROBUST

- Prior Distribution
  - Specified
  - Justified
  - Sensitivity analysis
- Analysis
  - Statistical model
  - Analytical technique
- Results
  - Central tendency
  - SD or Credible Interval

What's Missing?

BAYESWATCH

- Introduction
  - Intervention described
  - Objectives of study
- Methods
  - Design of Study
  - Statistical model
  - Prior / Loss function?
    - When constructed
    - Prior described
    - Loss function described
  - Use of Software – MCMC , starting values, run-in, length of runs, convergence, diagnostics
- Results
- Interpretation
  - Posterior distribution summarized
  - Sensitivity analysis if alternativepriors used

BASIS

- Research question
- Statistical model
  - Likelihood, structure, prior & rationale
- Computation
  - Software - convergence if MCMC, validation, methods for generating posterior summaries
- Model checks, sensitivity analysis
- Posterior Distribution
  - Summaries used: i). Mean, std, quintiles ii) shape of posterior, (iii) joint posterior for multiple comparisons, (iv) Bayes factors
- Results of model checks and sensitivity analyses
- Interpretation of Results
- Limitation of Analysis

Guidance for the Use of Bayesian Statistics in Medical Device Clinical Trials – FDA/CDRH 2010

- "Because of the inherent flexibility in the design of a Bayesian clinical trial, a thorough evaluation of the operating characteristics should be part of the trial design. This includes evaluation of:
  - probability of erroneously approving an ineffective or unsafe device (type I error)
  - probability of erroneously disapproving a safe and effective device (type II error)
  - power (the converse of type II error: the probability of appropriately approving a safe and effective device)
  - sample size distribution (and expected sample size)
  - prior probability of claims for the device
  - if applicable, probability of stopping at each interim look. "

Bayesian Monitoring of Clinical Trials

- Development follows Grossman et al (SIM, 1994)
  - All data & priors are normal (known variance σ^2)
  - A maximum of n patients in each of 2 groups (trts:A and B)
  - T interim analyses after tn/T (t=1,..T) patients per group
  - Of interest is $\delta=\mu_A-\mu_B$
  - The observed difference between groups of the tth cohort is d_t with variance $T\sigma_\delta^2/n$ (where $\sigma_\delta^2=2\sigma^2$)
  - Prior information for δ is available: corresponding to fn patients per group centred at δ_0

- Bayes Theorem implies that at the tth interim the posterior for δ is:

$$p\left(\delta \mid D_t = \sum_{i=1}^{t} d_i\right) \sim N\left(\frac{\frac{n}{T}D_t + fn\delta_0}{\frac{tn}{T}+fn}, \frac{\sigma^2}{\frac{tn}{T}+fn}\right)$$

Bayesian Monitoring of Clinical Trials

- Stopping rule: $\mathrm{Prob}(\delta > \delta_C \mid D_t) > 1 - \psi_t$

equivalent to: $$\Psi_t > \Phi\left[\frac{\delta_C - (D_t + Tf\delta_0)/(t + fT)}{\sigma/(tn/T + fn)^{1/2}}\right]$$

requiring $$D_t > -\frac{T^{1/2}(t + fT)^{1/2} Z_{\Psi_t}\sigma}{n^{1/2}} + \delta_C(t + fT) - Tf\delta_0$$

This is the general case and there are a number of "tuning" parameters: ψ_t, f, δ_C and δ_0

Bayesian Monitoring of Clinical Trials
Special Case 1: T=1, δ_C=0

- Stopping rule requires: $D > -\frac{(1+f)^{1/2} Z_\Psi \sigma}{n^{1/2}} - f\delta_0$

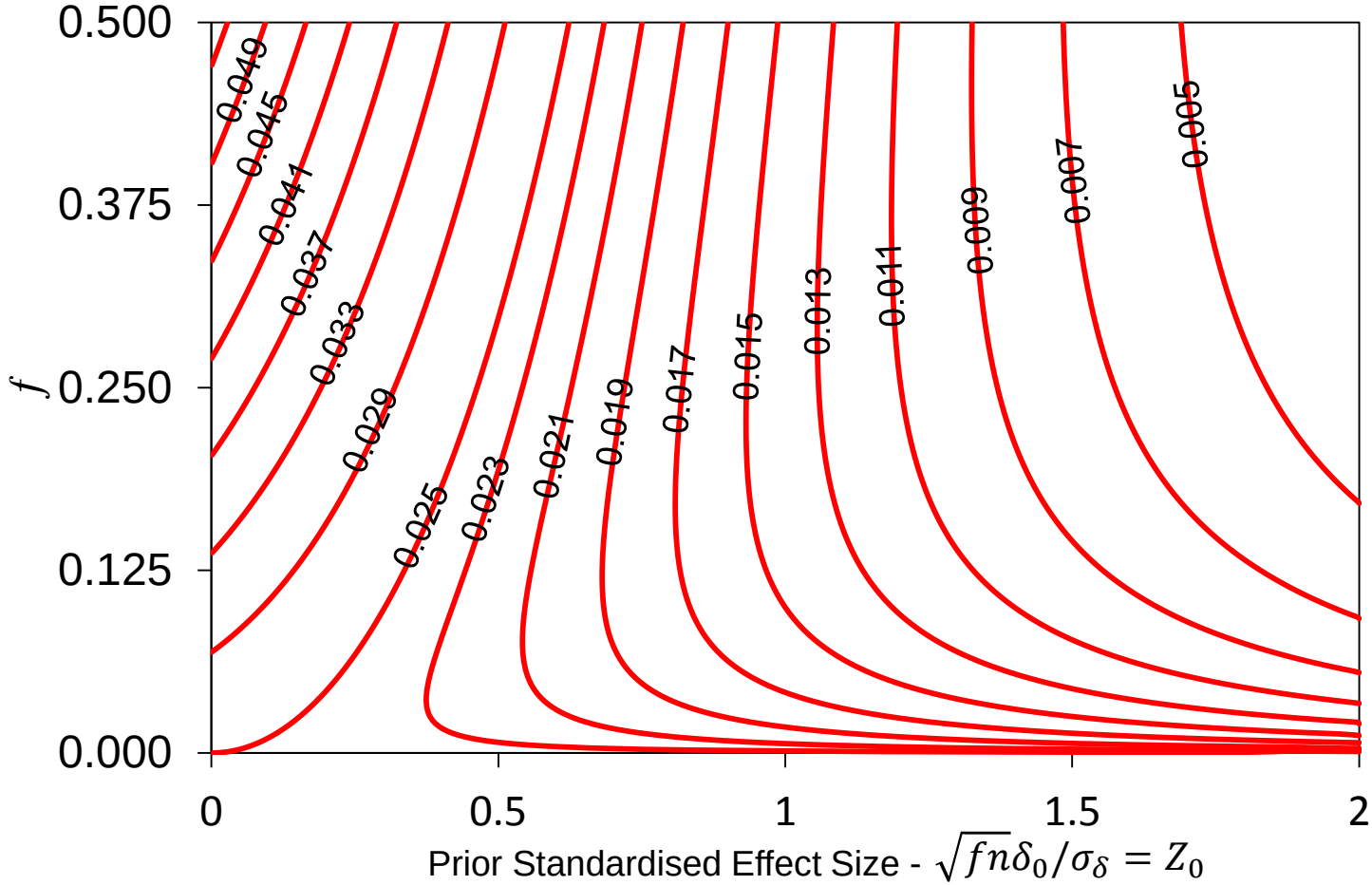
- What are the frequency properties of this rule?
- Under the Null Hypothesis: $\delta \sim N(0,\sigma^2_\delta/n)$

$$\Rightarrow P\left[D > \frac{-\sigma_\delta z_\psi (1+f)^{1/2}}{n^{1/2}} - f\delta_0\right] = 1 - \Phi\left(-z_\psi (1+f)^{1/2} - \frac{f^{1/2}(nf)^{1/2}\delta_0}{\sigma_\delta}\right)$$

- To control this at the 2.5% level we need

$$z_{1-\psi} = \frac{z_{0.975} + f^{1/2} z_0}{(1+f)^{1/2}}$$

Contours of Bayesian Decision Rule (ψ) To give a One-sided Type I Error 0f 2.5%



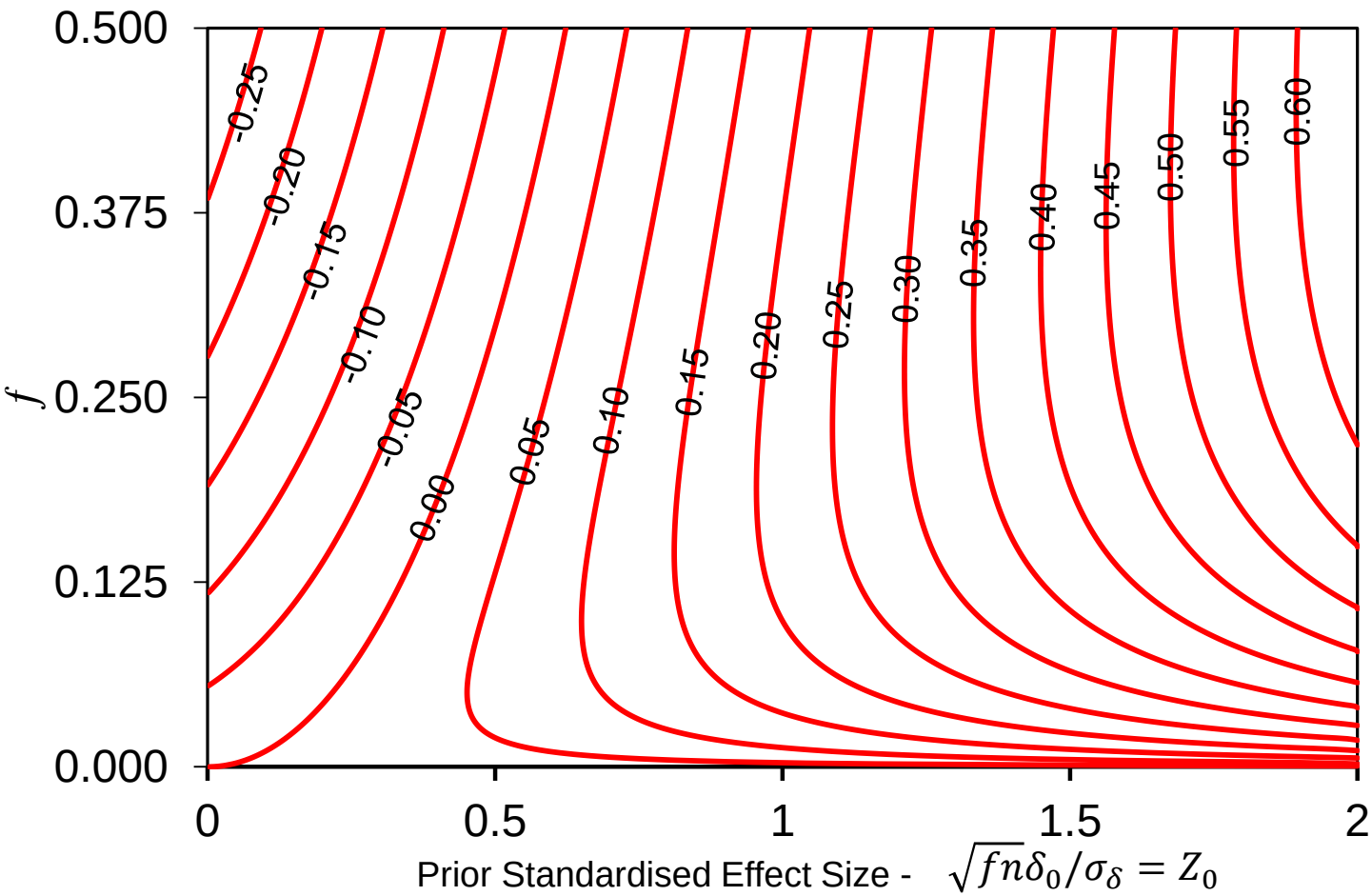
Bayesian Monitoring of Clinical Trials
Special Case 2: T=1, ψ=0.025

- In this case

$$\mathrm{Prob}(\delta > \delta_c \mid D) = 0.975 = 1 - \Phi\left(\frac{n^{1/2}\left[(1+f)\delta_c - (D + f\delta_0)\right]}{\sigma_\delta (1+f)^{1/2}}\right)$$

- giving a condition for D which can be used to find a value for δ_c to give the appropriate type I error.

Contours of Bayesian Decision Rule ($\delta_C\sigma_\delta/n^{1/2}$) To give a One-sided Type I Error 0f 2.5%



Bayesian Monitoring of Clinical Trials
Special Case 1 & 2

- Whichever approach is used is turns out that using this approach is effectively discounting the prior information.

- To see this substitute $z_\psi = \dfrac{z_{0.025} - f^{1/2} z_0}{(1+f)^{1/2}}$ into

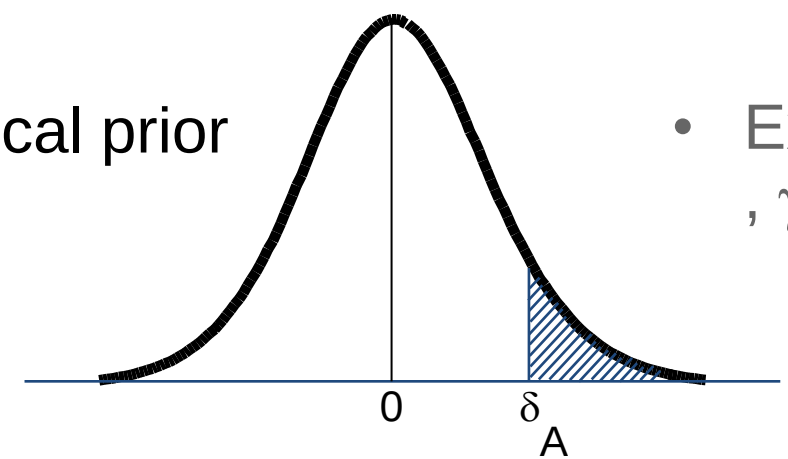
$$D > -\frac{(1+f)^{1/2} Z_\Psi \sigma}{n^{1/2}} - f\delta_0 \quad \text{giving} \quad D > \frac{\sigma_\delta Z_{0.975}}{n^{1/2}}$$

which is the standard, frequentist decision criteria – in other words 100% discounting

Bayesian Monitoring of Clinical Trials
Special Case 3: ψ_t=0.025, δ_C=0, δ_0=0 (Sceptical Prior)

- A sceptical prior can be set up formally.
- Prior centred around 0, with a small probability γ of achieving the alternative δ_A - $p(\delta>\delta_A) = \gamma$
- From which: $\delta_A = -\dfrac{\sigma_\delta z_{1-\gamma}}{(fn)^{1/2}}$

sceptical prior



- Now suppose the trial has been designed with size α and power 1-β to detect the alternative hypothesis δ_A.
- So that: $n = \dfrac{\sigma_\delta^2 \left(z_{1-\alpha/2} + z_{1-\beta}\right)^2}{\delta_A^2}$
- From which: $f = \left(\dfrac{z_\gamma}{z_{1-\alpha/2} + z_{1-\beta}}\right)^2$
- Example: α=0.05, 1- β=0.90 , γ=0.05 => f ~ 1/4

Bayesian Monitoring of Clinical Trials
Special Case 3: ψ_t=0.025, δ_C=0, δ_0=0

- In this case: $$\mathrm{Prob}(\delta > \delta_C \mid D_t) = 1 - \Phi\left[\frac{-D_t/(t+fT)}{\sigma_\delta/(t/T+f)^{1/2}}\right] > 0.975$$
$$= \Phi\left[\frac{T^{1/2}D_t}{\sigma_\delta n^{1/2}(t+fT)^{1/2}}\right] > 0.975$$
$$= \Phi\left[\frac{T^{1/2}D_t}{\sigma_\delta (nt)^{1/2}}\frac{t^{1/2}}{(t+fT)^{1/2}}\right] > 0.975$$

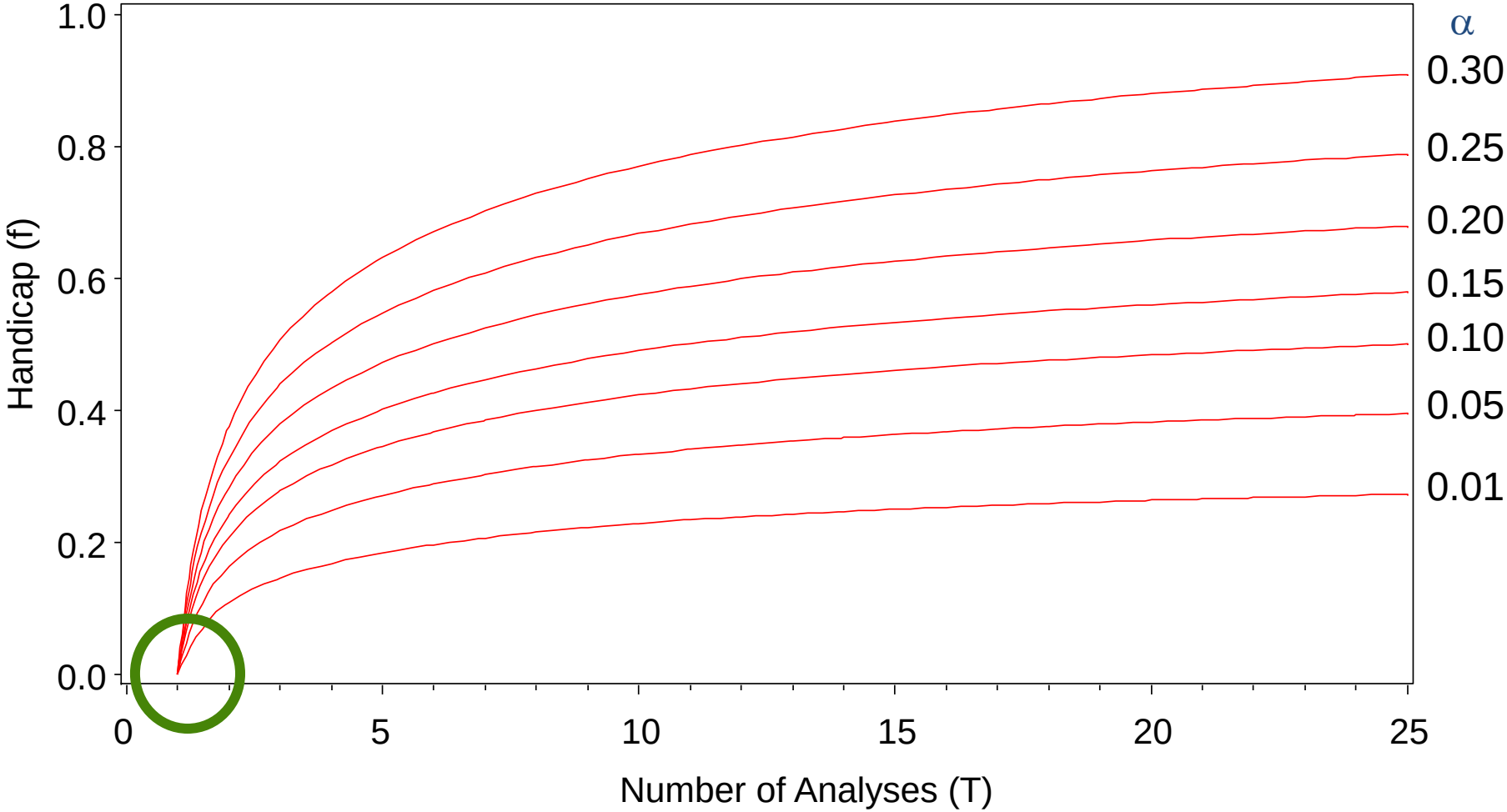
- which is equivalent to increasing the critical region by a factor $$\sqrt{\frac{t+fT}{t}}$$

Grossman et al(1994) call f the "handicap"

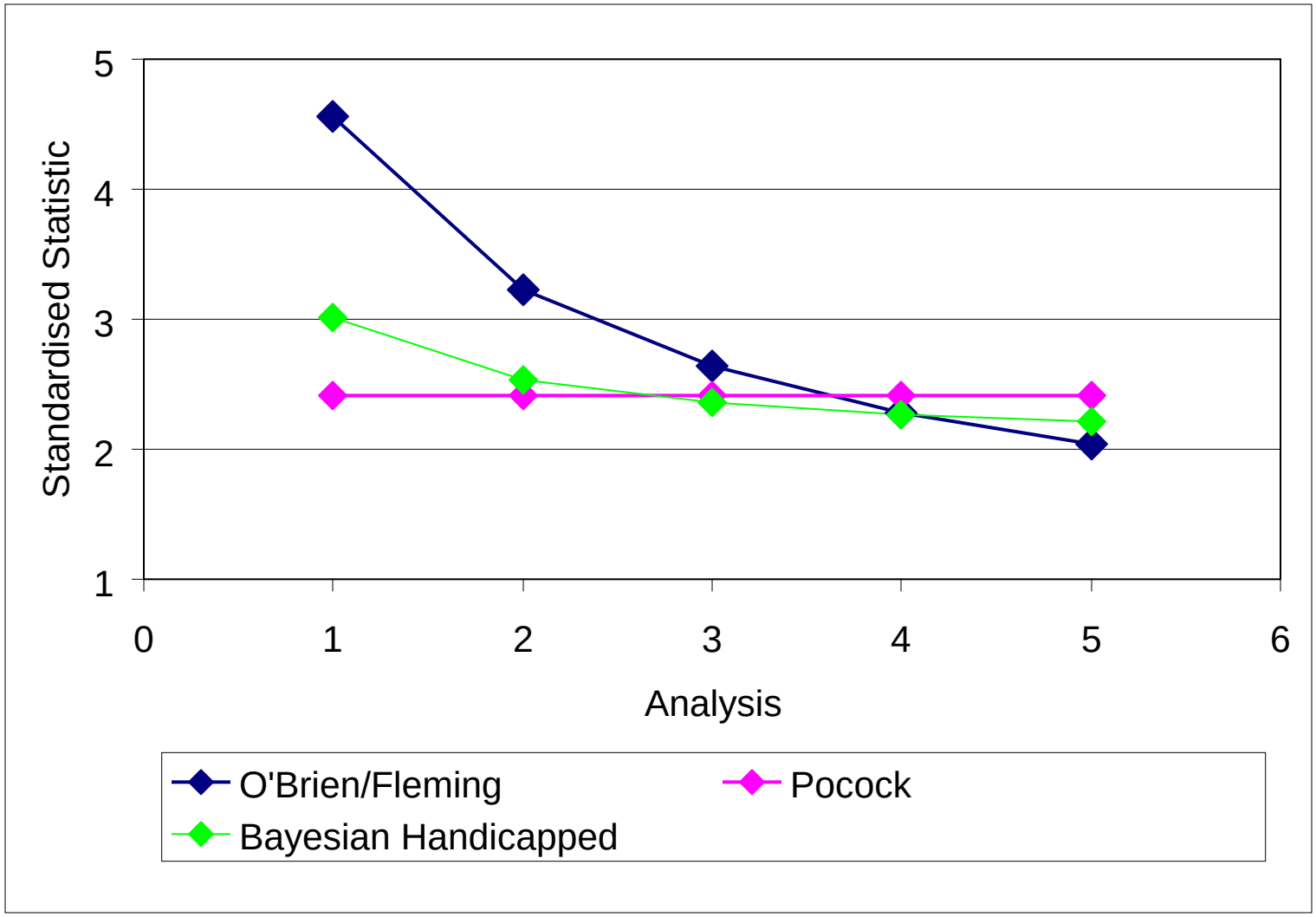
Bayesian Monitoring of Clinical Trials
Special Case 3: ψ_t=0.025, δ_C=0, δ_0=0

- The frequentist properties of this handicapping are not so easy to derive.
- For T=2 – a single interim – the frequentist type-I error can be calculated using a bivariate normal probability function, e.g. the SAS function PROBNRM.
- For T > 3 Grossman et al (1994) use simulation to determine the handicap f that controls the two-sided type I error at 5% and 1% (20,000,000 trials)
- Alternatively use can be made of the algorithm derived by Armitage, MacPherson and Rowe (JRSSA, 1969) – used a SAS implementation of FORTRAN program by Reboussin, DeMets, Kim and Lan or SEQ, SEQSCALE & SEQSHIFT (PROC IML)

Handicaps(f) To Control the Two-sided α for Upto 25 Analyses



Comparison of Critical Values
O'Brien/Fleming, Pocock & Handicapped Bayes



Handicapped Bayes versus Optimal Designs (Pocock, 1982)

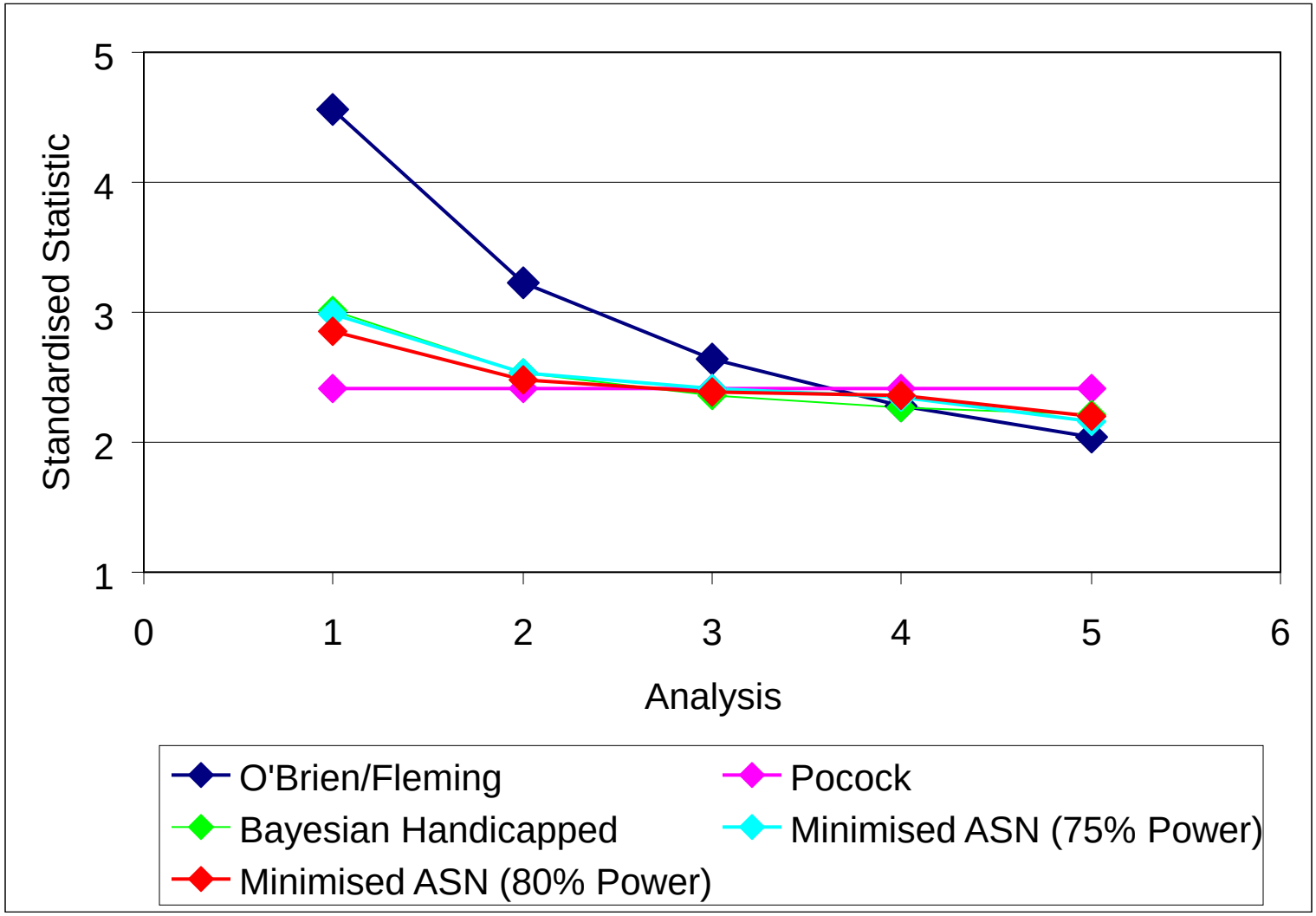
- Investigated properties of group sequential designs, in particular the Average Sample Number (ASN)

Maximum number of groups, K	Nominal significance level, α'	Required number* of patients per group 2n	Maximum number* of patients 2nN	Average number of patients until stopped under H_A (ASN)
1	0.05	51.98	52.0	52.0
2	0.0294	28.39	56.8	37.2
3	0.0221	19.73	59.2	33.7
4	0.0182	15.19	60.8	32.3
5	0.0158	12.38	61.9	31.3
10	0.0106	6.50	65.0	29.8
20	0.0075	3.38	67.6	29.5

- **Multiply by σ^2/δ^2**

23

Comparison of Critical Values
Optimal ASN (75/80% Power) & Handicapped Bayes

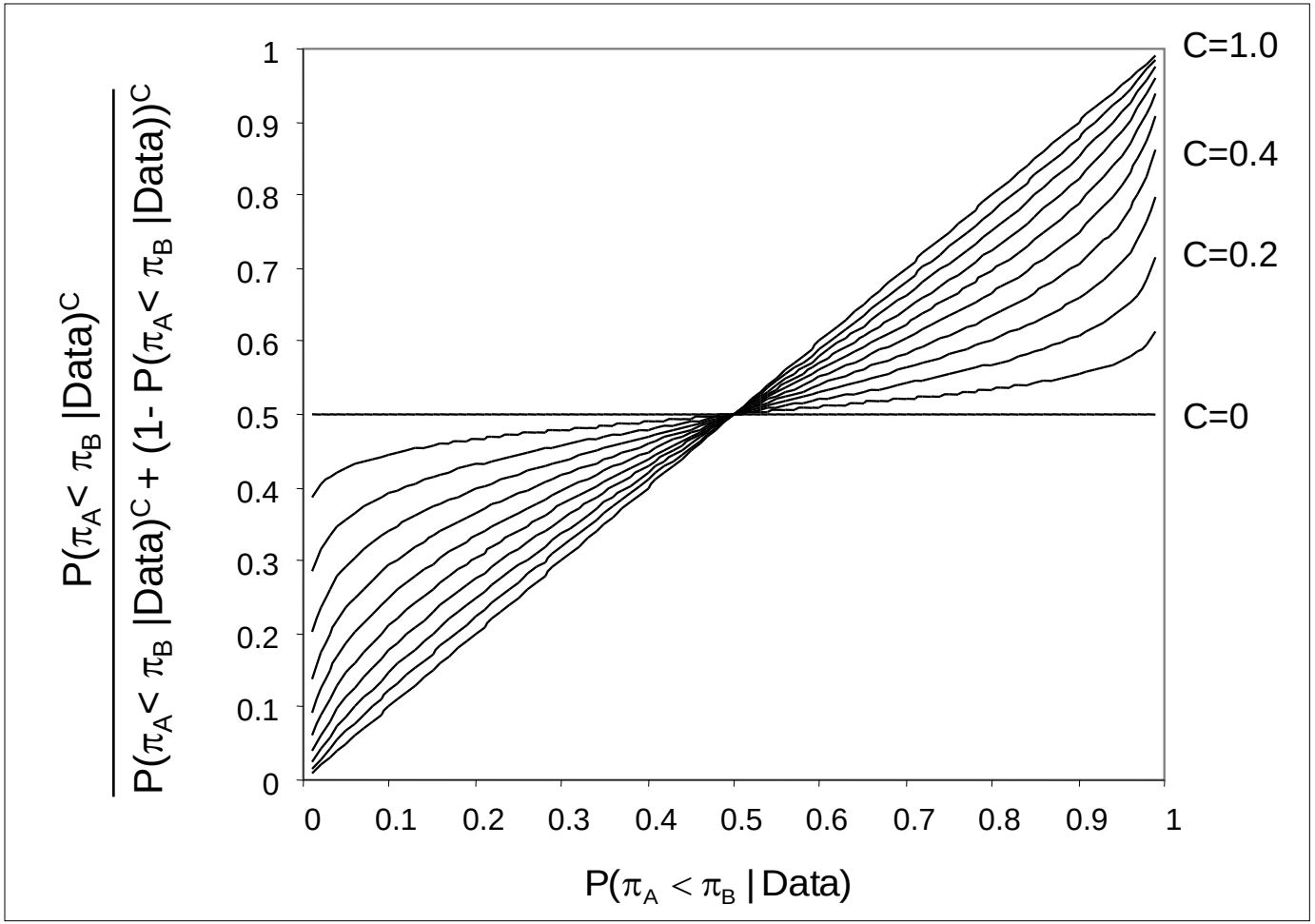


Bayesian Adaptive Randomisation
Thall and Wathen (Eur J Cancer, 2007)

- Back to an idea of Thompson (Biometrika, 1933)
- Similar to RPW – binary outcome
- Randomisation to treatment B on the basis of a function of $P(\pi_A < \pi_B|\text{Data})$ although in practice Thompson used $P(\pi_A < \pi_B|\text{Data})$.
- Unstable
- Thall and Wathen (European J Cancer, 2007)

$$\frac{P(\pi_A < \pi_B \mid \text{Data})^c}{P(\pi_A < \pi_B \mid \text{Data})^c + \left[1 - P(\pi_A < \pi_B \mid \text{Data})\right]^c}$$

Bayesian Adaptive Randomisation
Impact of Choice of C

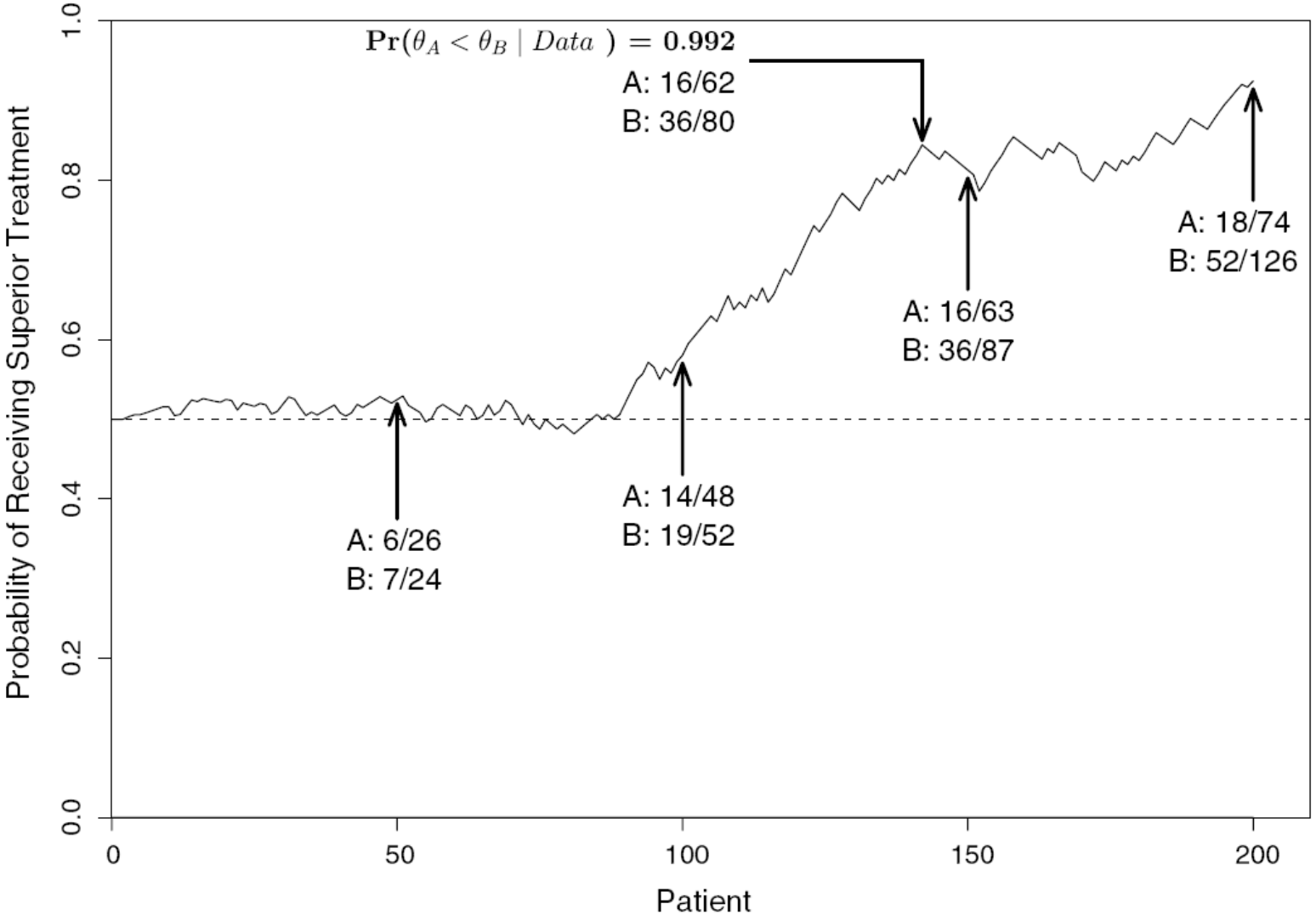


Bayesian Adaptive Randomisation
Impact of Choice of C

- Thall and Whalen recommend C= n/(2N)
  - n=current sample size
  - N=study's maximum sample size
- Begins with C=0, ends with C=1/2
- C=1/2 "works well in many applications"

- Giles et al (J Clin Oncology, 2003)
  - Similar idea – but now with 3 arms (2 experimental, 1 control) using functions of $P(m_1<m_0|data)$, $P(m_2<m_0|data)$, and $P(m_1<m_2|data)$, - m_2, m_1, and m_0 are the median survival times

An Example



A Major Issue

- As sample size increases posterior probability increases
- Even if treatments are similar
- This is in contrast to RPW based on success rates
- Maybe appropriate if the new treatment is much safer than the standard

2 x 2 Contingency Table Data Structure

	Response	No Response
Treatment A	$r_1\ (\pi_A)$	$n_1\text{-}r_1(1\text{-}\pi_A)$
Treatment B	$r_2\ (\pi_B)$	$n_2\text{-}r_2\ (1\text{-}\pi_B)$

Likelihood $$\propto \pi_A^{r_1}\left(1-\pi_{1A}\right)^{n_1-r_1}\pi_B^{r_2}\left(1-\pi_B\right)^{n_2-r_2}$$

Prior $$\propto \pi_A^{\alpha_1-1}\left(1-\pi_A\right)^{\beta_1-1}\pi_B^{\alpha_2-1}\left(1-\pi_B\right)^{\beta_2-1}$$

Posterior $$\propto \pi_A^{r_1+\alpha_1-1}\left(1-\pi_A\right)^{n_1-r_1+\beta_1-1}\pi_B^{r_2+\alpha_2-1}\left(1-\pi_B\right)^{n_2-r_2+\beta_2-1}$$

2x2 Contingency Table - Posterior Inference "Uninformative Priors" : $\alpha_A = \beta_A = \alpha_B = \beta_B = 1$

- The probability of interest is

$$\mathrm{Prob}(\pi_A < \pi_B \mid \mathrm{Data}) = \sum_{k=0}^{n_1 - r_1} \frac{\binom{n_1 + n_2 - r_1 - r_2 - k}{n_2 - r_2}\binom{r_1 + r_2 + 1 + k}{r_2}}{\binom{n_1 + n_2 + 1}{n_1 + 1}}$$

based on the cumulative hypergeometric function as is Fisher's exact test (Altham JRSSB1969; Raiffa & Schlaifer, Applied Statistical Decision Theory, 1960))

Historical Aside

- Thompson(1935) proved the identity:

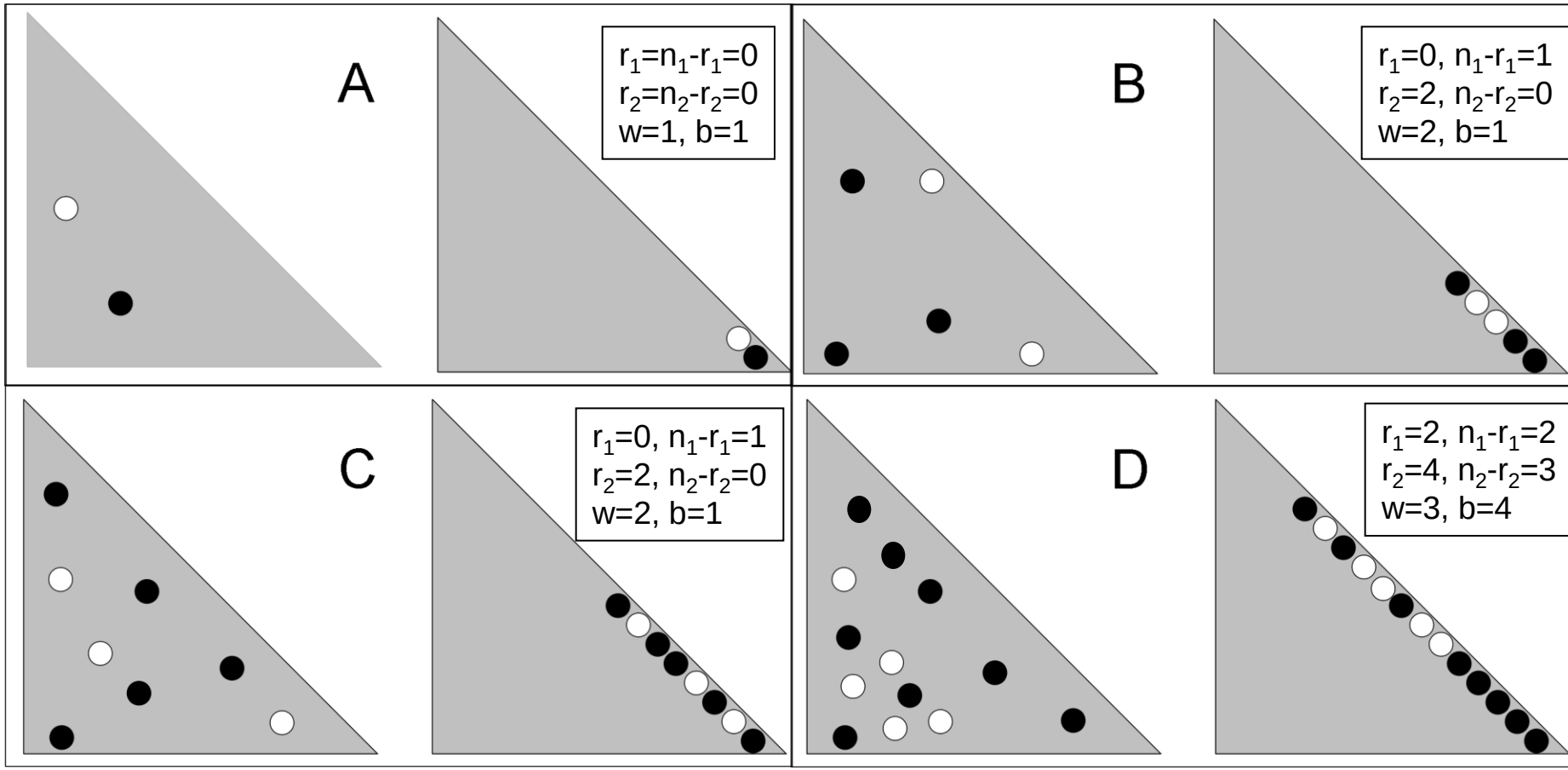
$$\sum_{k=0}^{n_1-r_1} \frac{\binom{n_1+n_2-r_1-r_2-k}{n_2-r_2}\binom{r_1+r_2+1+k}{r_2}}{\binom{n_1+n_2+1}{n_1+1}} = \sum_{k=0}^{\min(b-1,W-w)} \frac{\binom{W}{w+\alpha}\binom{B}{b-1-\alpha}}{\binom{W+B}{w+b-1}}$$

where: $W=n_1+1$, $B=n_2+1$, $w=n_1-r_1$ and $b=n_2-r_2$

- This second term is the probability under sampling without replacement from a mixture of W white balls and B black balls that we will get w white balls before b black balls

Thompson(1935)
Mechanical Randomisation & Simulation

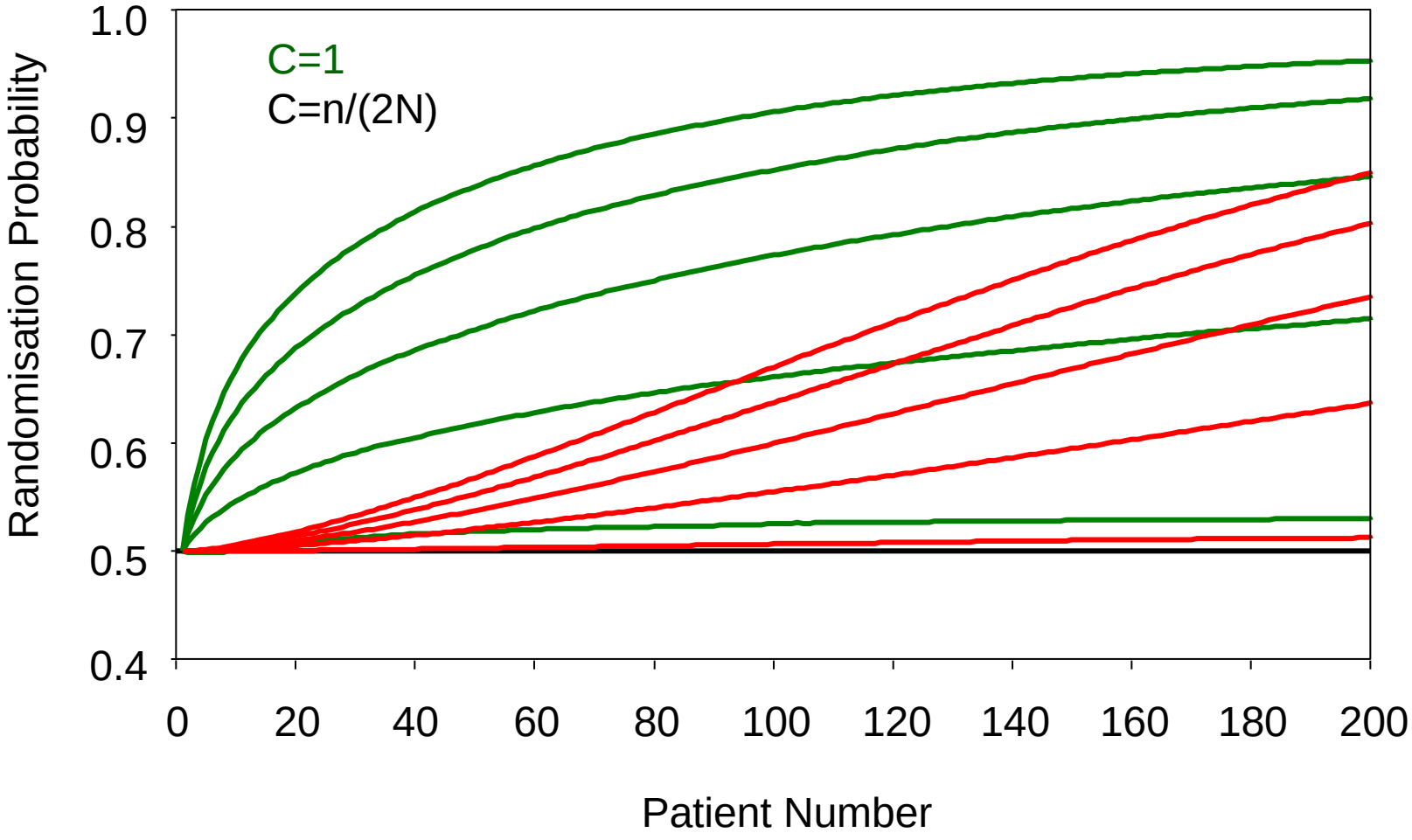
- For $W=n_1+1$, $B=n_2+1$: choose A if $w=n_1-r_1+1$ white balls occur before $b=n_2-r_2+1$ black balls



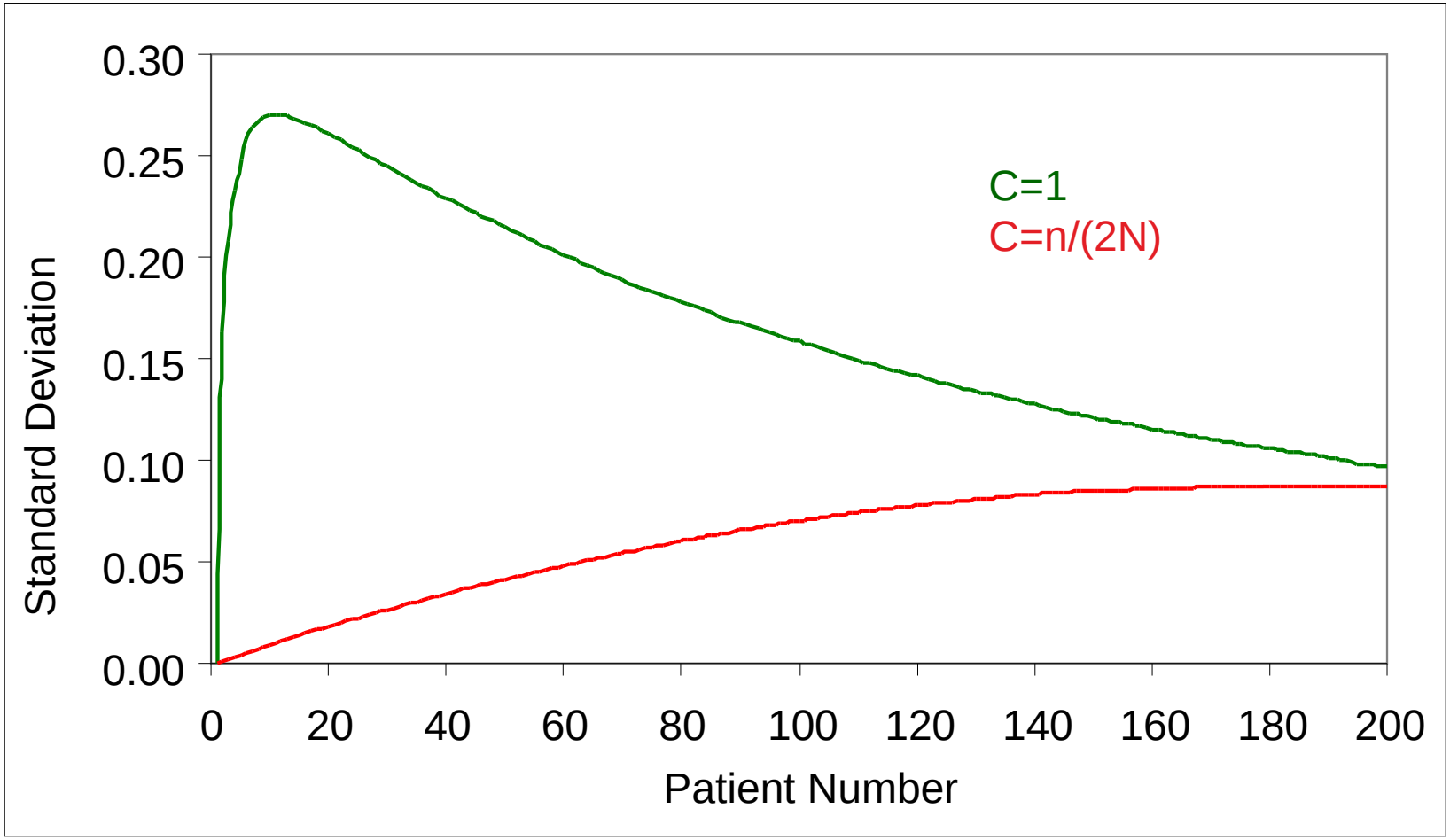
Bayesian AD – Thall & Wathen(EJC,2007)
Type-I Error Based on T&W Criterion

- Thall & Wathen illustration is based on:
  - N = 200
  - Stopping Rules
    - If $P(\pi_A<\pi_B|Data) > 0.99$ stop and "choose" B
    - If $P(\pi_A<\pi_B|Data) < 0.01$ stop and "choose" A (futility)
- What does the type I error look like ?
- A complication is that the control rate - π_A - is a nuisance parameter

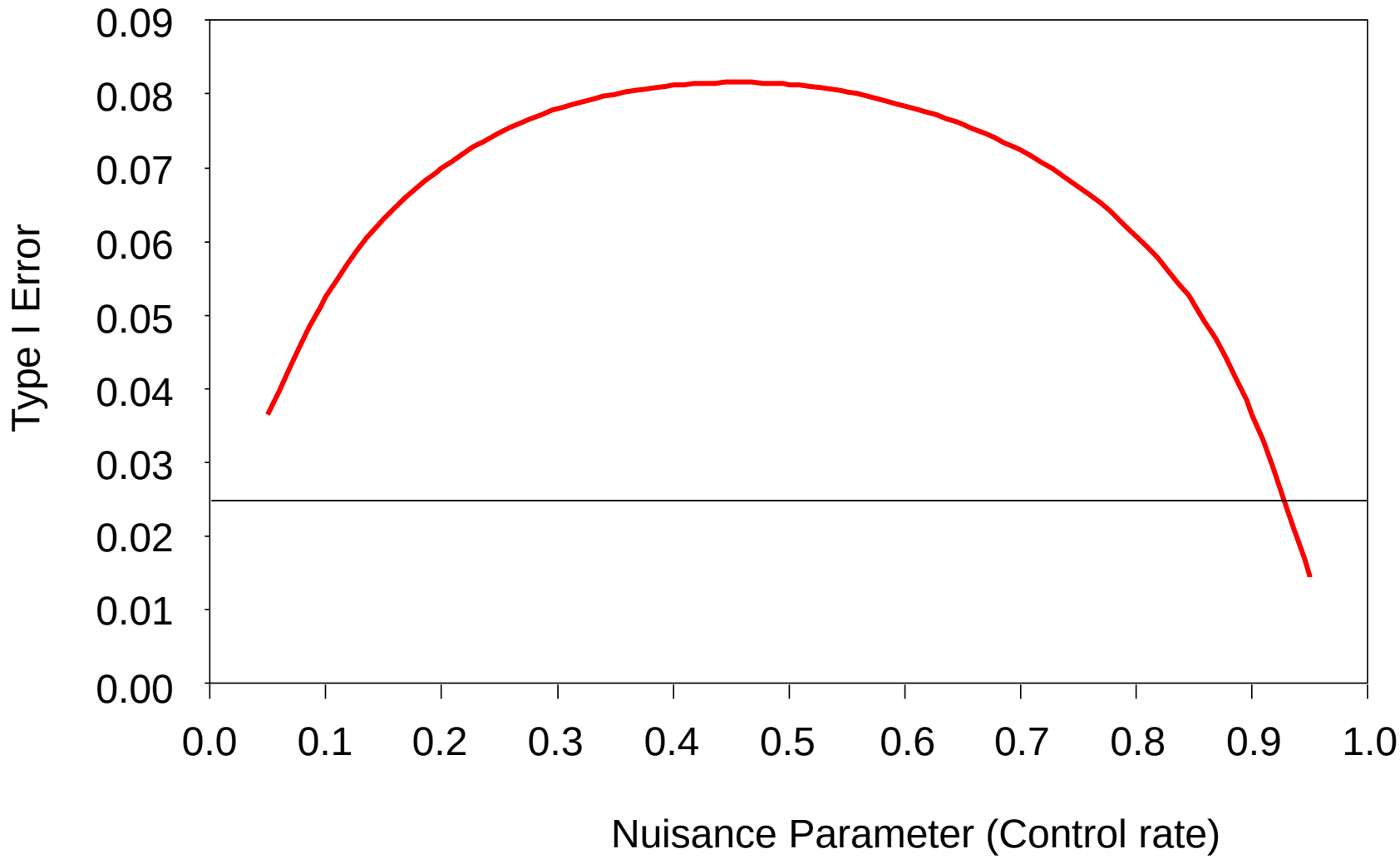
Bayesian AD – Thall & Wathen(EJC,2007) N=200 Randomisation Probabilities (10^5 simulations) π_A=0.25 , π_B=0.25(0.05)0.45



Bayesian AD – Thall & Wathen(EJC,2007) N=200
SD of Randomisation Probabilities (10^5 simulations)
π_A=0.25 , π_B=0.45



Bayesian AD – Thall & Wathen(EJC,2007) N=200
Type-I Error Based on $P(\pi_A > \pi_B | Data) > 0.99$
10^6 Simulations / control rate



Bayesian AD – Thall & Wathen(EJC, 2007) N=200
Control One-Sided Type-I Error

- The issue is the number of tests being conducted

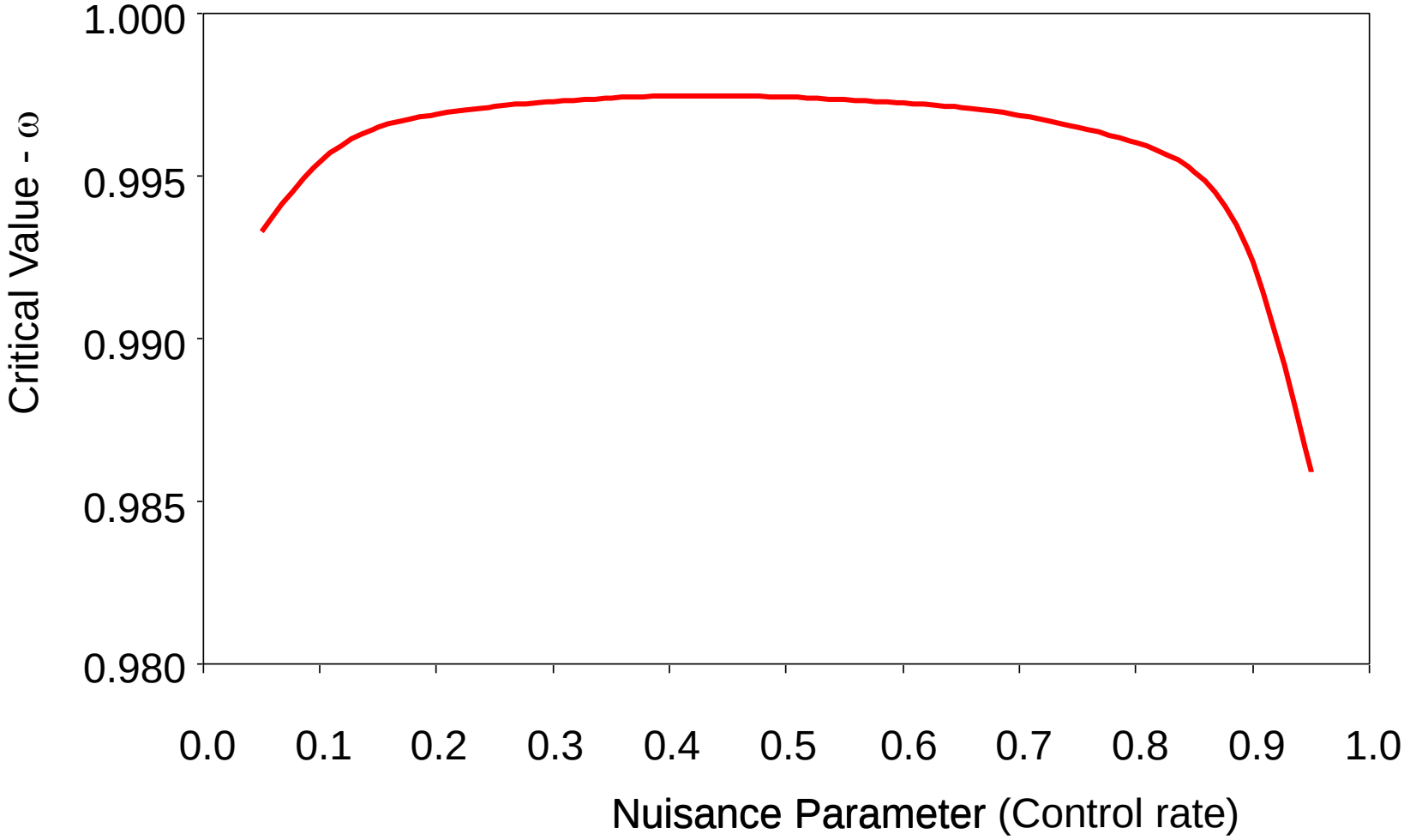
1. Reduce the problem using cohorts (20, 50 or ?)

2. Or choose decision criterion

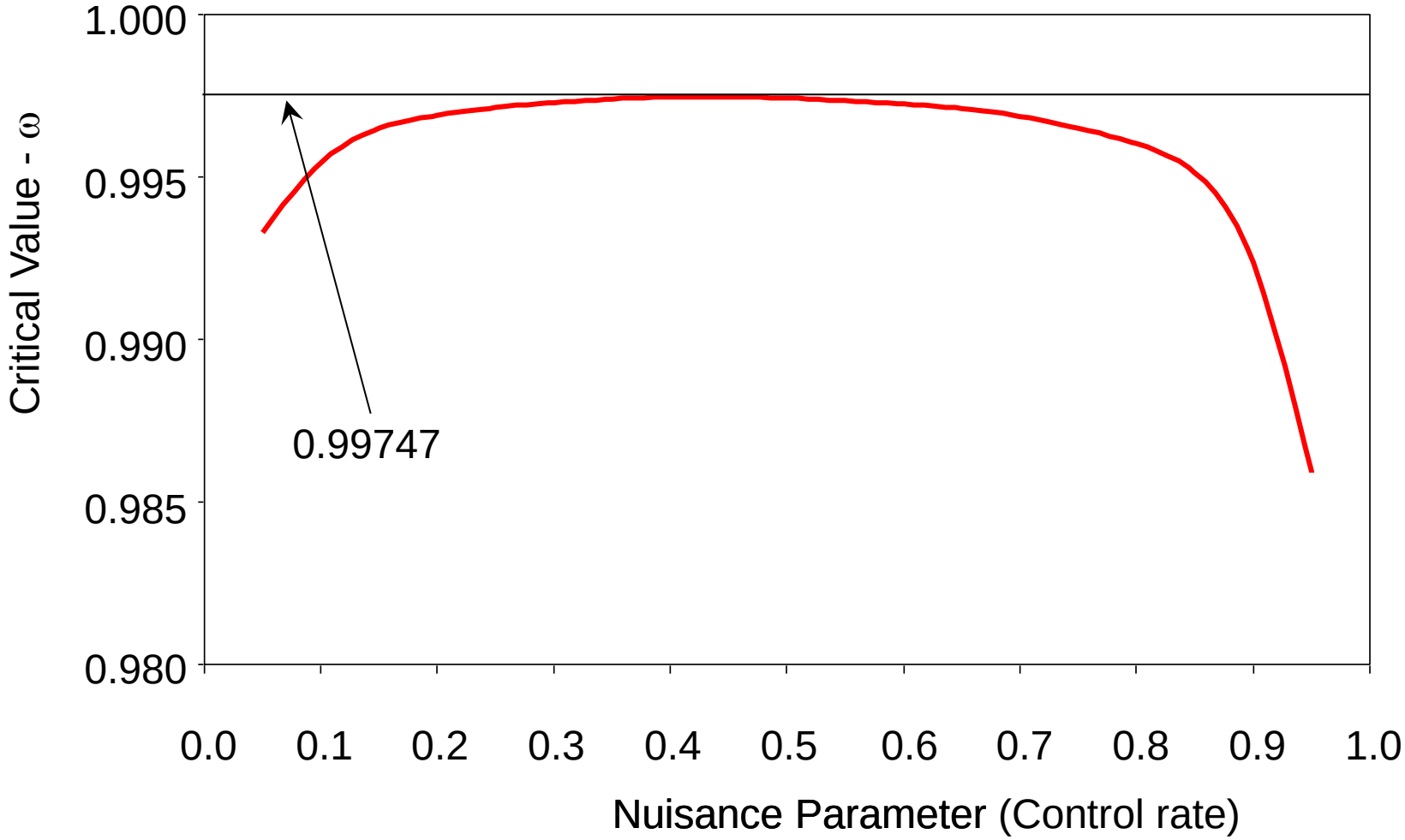
$$P(\pi_A<\pi_B|\text{Data})> \omega$$

to control type-I error

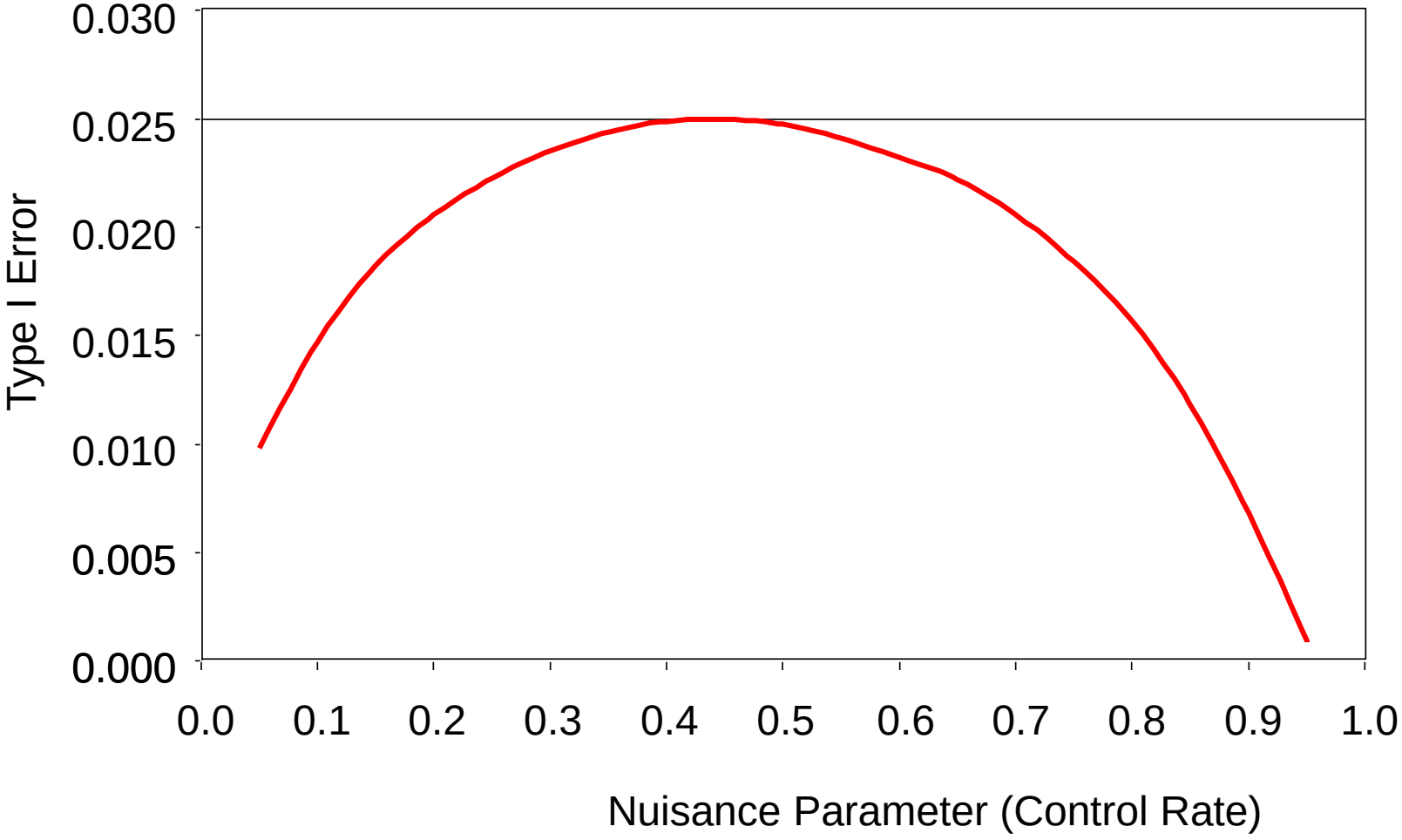
Bayesian AD – Thall & Wathen(EJC, 2007) N=200
Critical Value to Control One-Sided Type-I Error
10^6 Simulations / control rate



Bayesian AD – Thall & Wathen(EJC, 2007) N=200
Critical Value to Control One-Sided Type-I Error
10^6 Simulations / control rate



Bayesian AD – Thall & Wathen(EJC, 2007) N=200
Type-I Error Based on $P(\pi_A<\pi_B|Data)>.99747$
10^6 Simulations / control rate



Criticism of This Approach

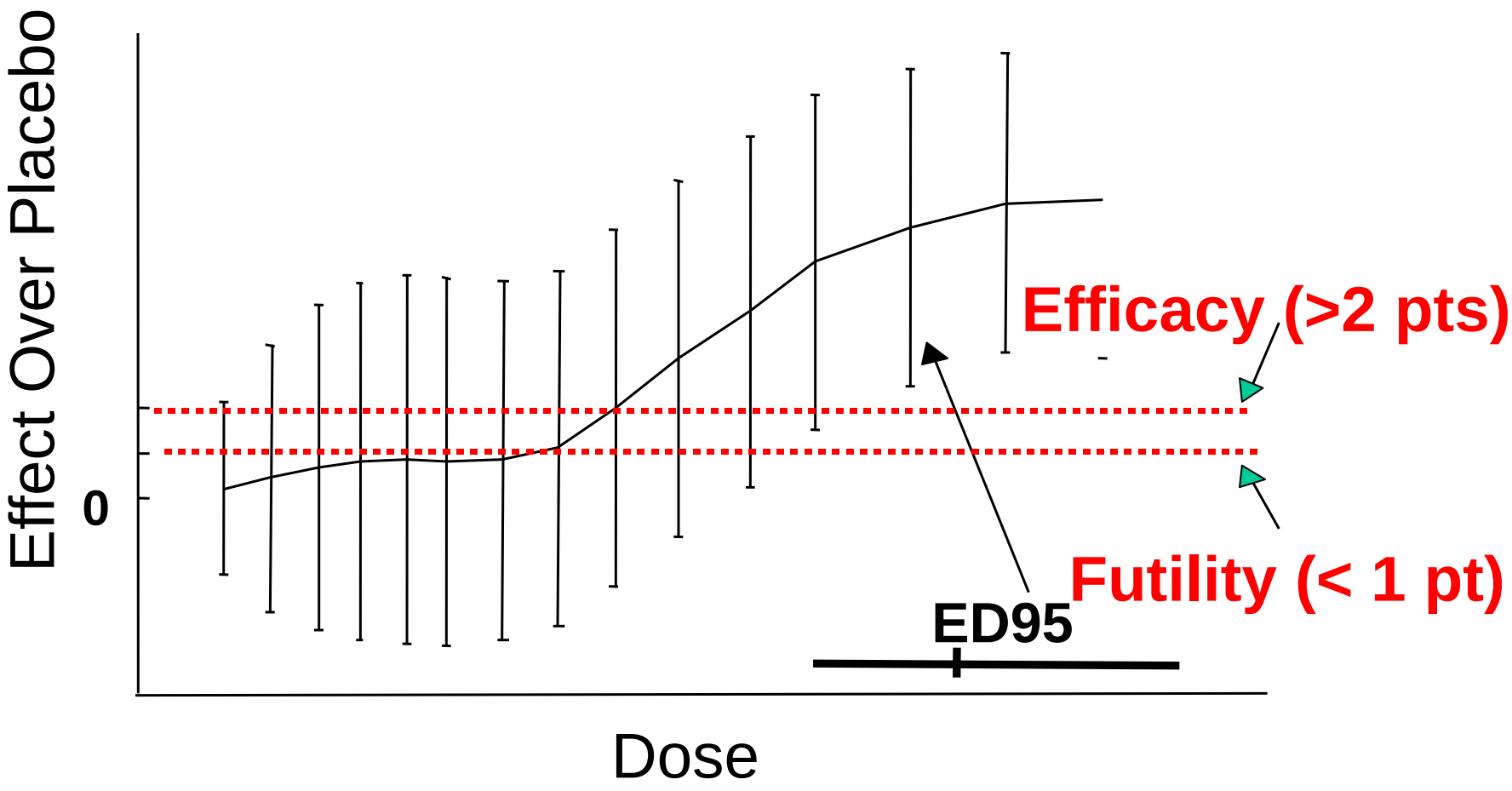
- Korn and Freidlin (J Clin Oncol, 2011)
- Their simulations “show”:
  - Thall & Wathen AD inferior to1:1 randomisation in terms of information, benefits to patients in trial

- True
- I agree with Don Berry (J Clin Ocol 2011) that the greatest benefits are likely to accrue for trials with more than 2 arms
- Rather as in the case of T=1 in the group sequential case greater complexity gives more scope for Bayesian designs

Conclusions
Determining Decision Criteria

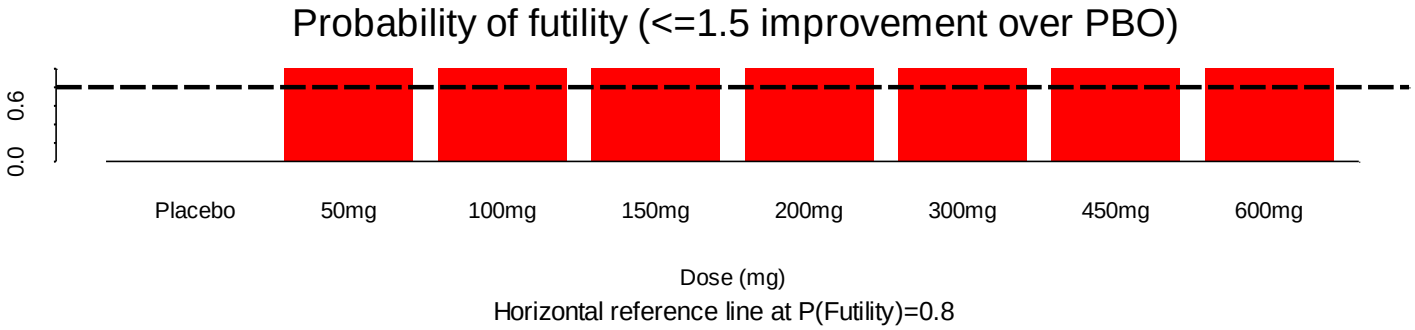
- Appropriate approach:
  - Choose decision rule based on clinical or commercial criteria.

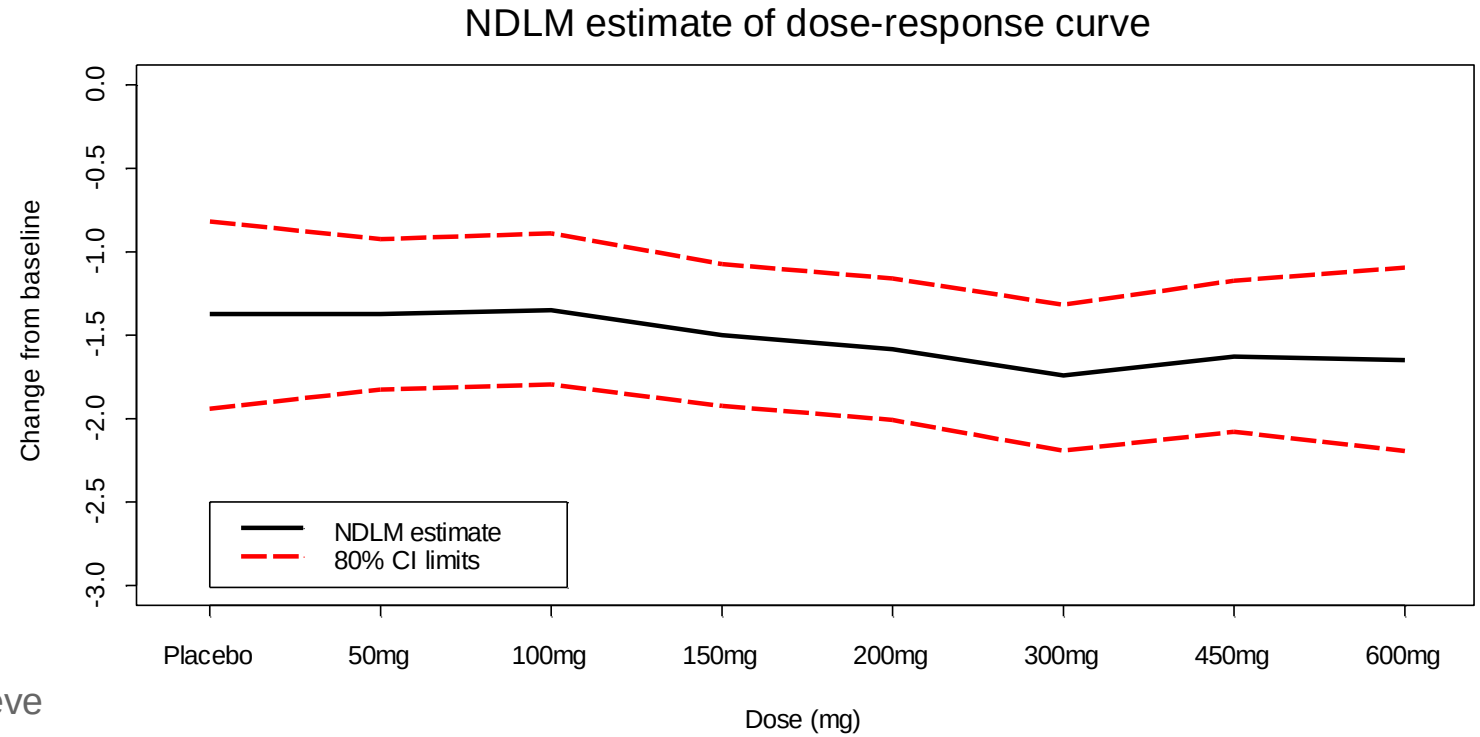
ASTIN Trial – Acute Stroke: Dose Effect Curve (Grieve and Krams, Clinical Trials, 2005)



POC Study in Neuropathic Pain
Smith et al (Pharmaceutical Statistics, 2006)

Probability of futility and dose-response curve. Change from baseline in mean pain score

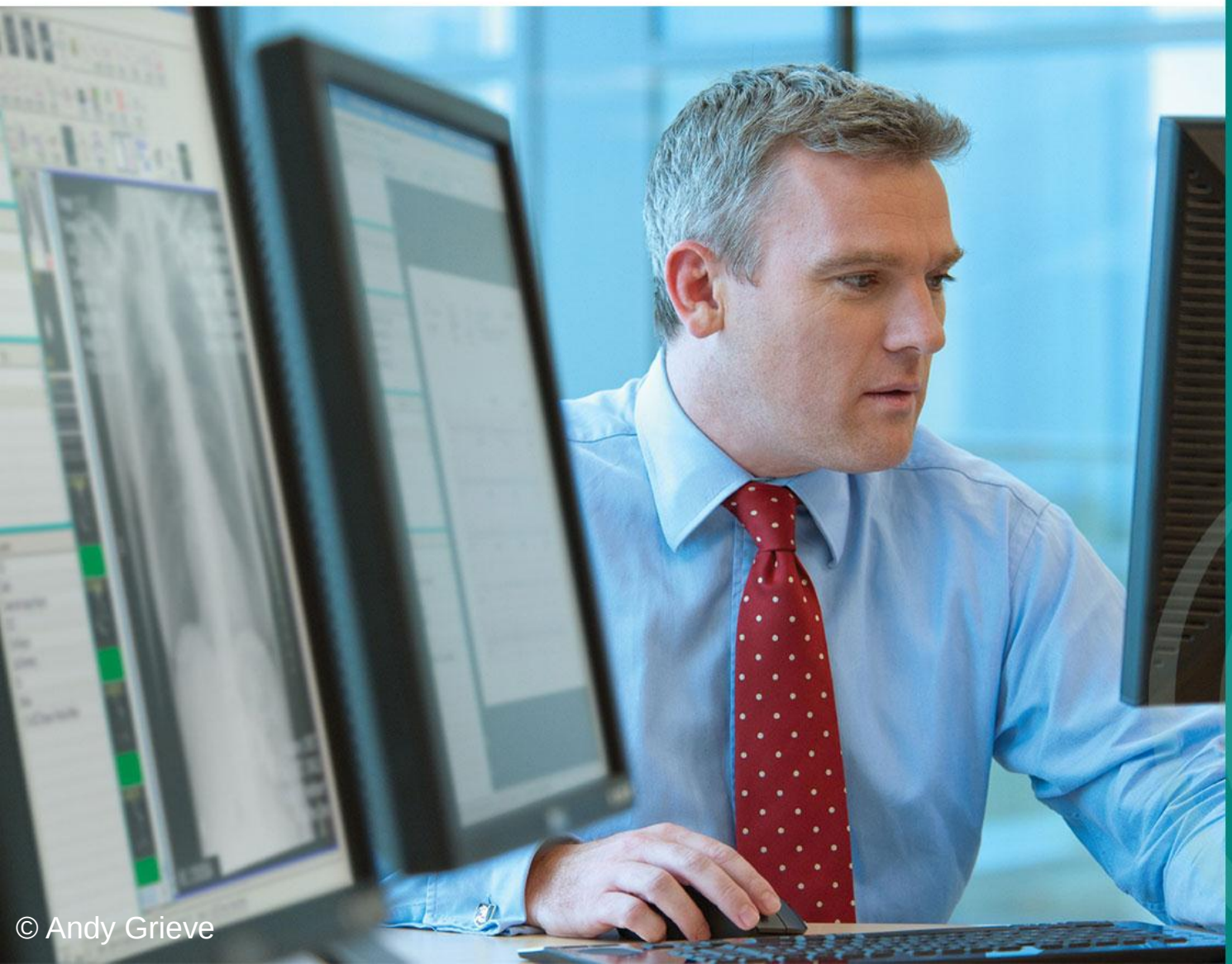




Conclusions
Determining Decision Criteria

- Appropriate approach:
  - Choose decision rule based on clinical or commercial criteria.
  - Investigate operating characteristics
  - If they are unacceptable e.g., type I error > 20% then look to change them
  - BUT don’t strive to get exact control

Banishment of p-values



© Andy Grieve

Recent Editorial

BASIC AND APPLIED SOCIAL PSYCHOLOGY, 37:1–2, 2015

ISSN: 0197-3533 print/1532-4834 online
DOI: 10.1080/01973533.2015.1012991

Editorial

David Trafimow and Michael Marks
New Mexico State University

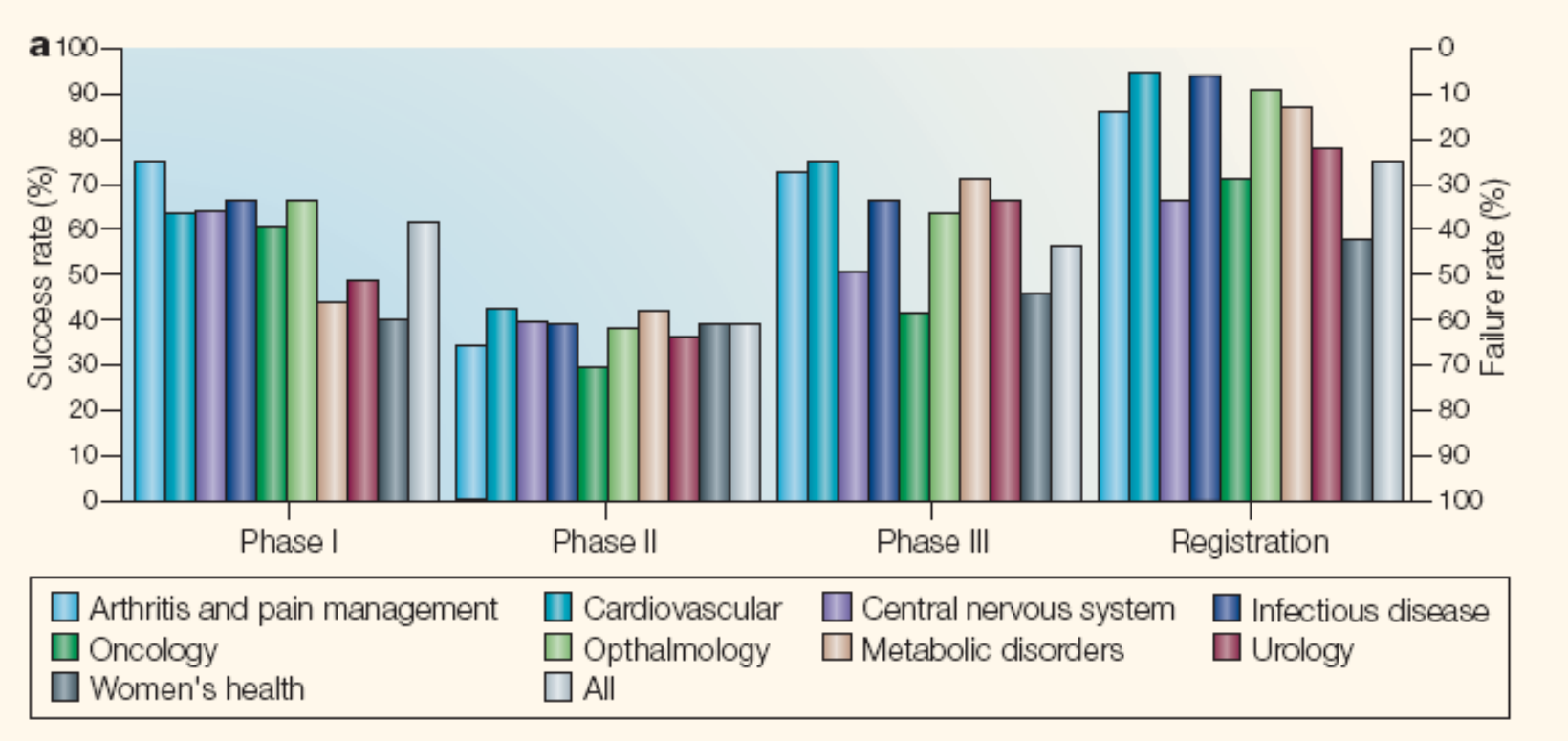
- "from now on BASP is banning NHSTP (null hypothesis significance testing procedure"
- NO MORE p-values
- Unthinking use of statistics

Robert Matthews
Sunday Telegraph, 13 September 1998

"The plain fact is that 70 years ago Ronald Fisher gave scientists a mathematical machine for turning baloney into breakthroughs, and flukes into funding. It is time to pull the plug"

Choosing Type I and Type II Errors

Can the pharmaceutical industry reduce attrition rates?



Kola & Landis (2004) NATURE REVIEWS | DRUG DISCOVERY

2012 Ecology Papers on Significance Testing

ICON
A Symbol of Excellence

Environmental Science & Technology
Policy Analysis
pubs.acs.org/est

Negative Consequences of Using $\alpha = 0.05$ for Environmental Monitoring Decisions: A Case Study from a Decade of Canada's Environmental Effects Monitoring Program

Joseph F. Mudge,*[,†] Timothy J. Barrett,[†] Kelly R. Munkittrick,*[,†,‡] and Jeff E. Houlahan[†]

Environ. Sci. Technol., 46, 9249-9255, 2012.

IF ALL OF YOUR FRIENDS USED $\alpha = 0.05$, WOULD YOU DO IT TOO?

Joseph F Mudge,* Christopher B Edge, Leanne F Baker, and Jeff E Houlahan
University of New Brunswick, Saint John, New Brunswick, Canada
*joe.mudge@unb.ca
DOI: 10.1002/ieam.1313

A NEW APPROACH TO SETTING α LEVELS

Integ. Environ. Ass. Man. 8, 563-369, 2012

Making statistical significance more significant

We routinely set significance levels at 0.05, giving us one chance in 20 of a false positive result if the null hypothesis were true. Why? Why not instead choose values that minimise the combined chances of both false positives *and* false negatives? It is easy, say **Leanne F. Baker** and **Joseph F. Mudge**, so why not do it?

Significance, June 2012, 29-30.

Setting an Optimal α That Minimizes Errors in Null Hypothesis Significance Tests

Joseph F. Mudge*, Leanne F. Baker, Christopher B. Edge, Jeff E. Houlahan

PLoS ONE, 7, e32734, 2012.

Should Type I Error be Fixed in Drug Development?

"If XXX during the 1st week is kept as the primary endpoint, it has at least to be supported by a convincing positive trend for clinically relevant long-term effects like XXX at a time-point of at least six months. It is recommended that XXX is considered as a key secondary endpoint, even if statistical significance at the usual level of 5% two-sided might not be necessary."

EMA Scientific Advice Response – 2012

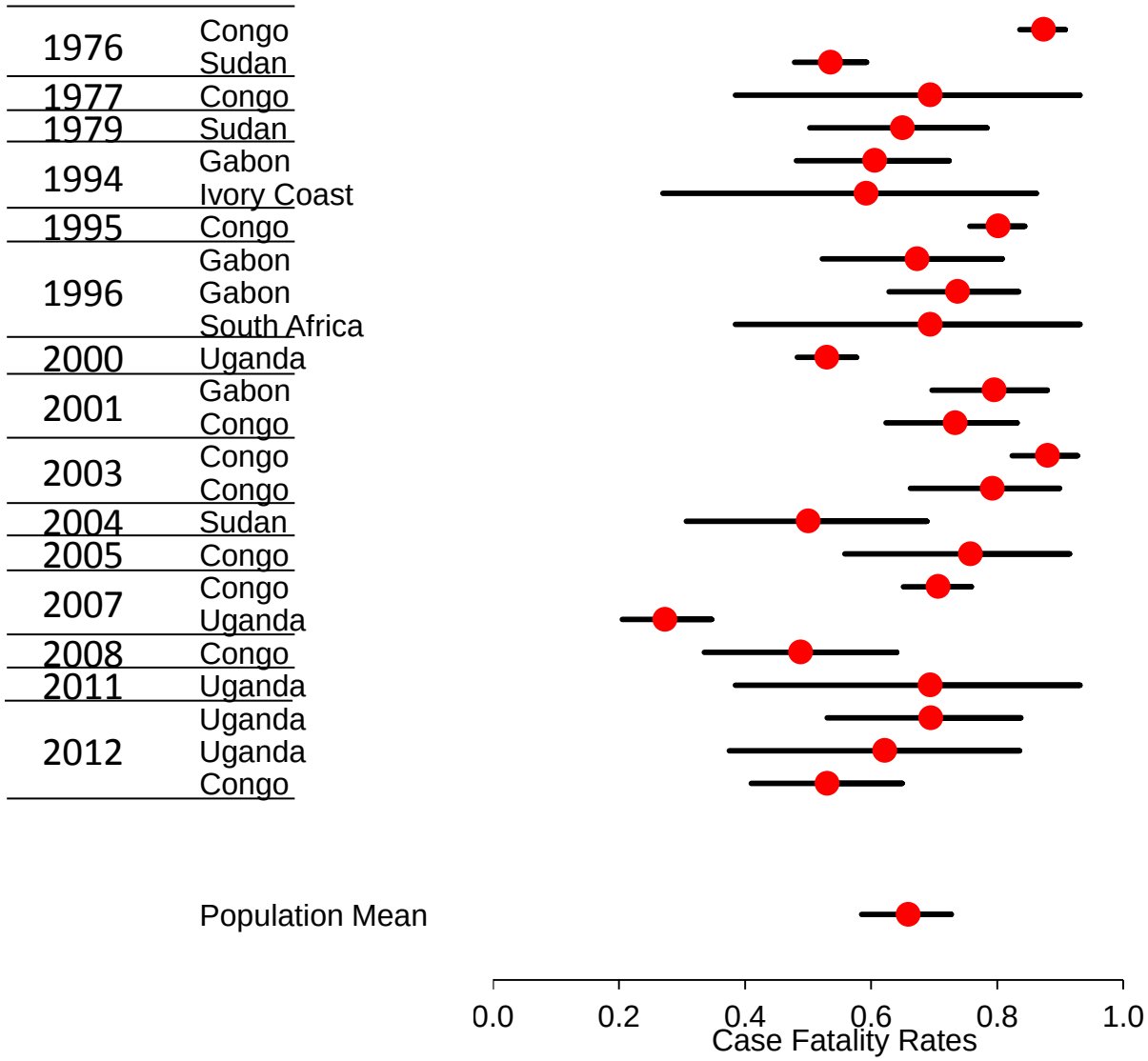
"no scientific worker has a fixed level of significance at which, from year to year, and in all circumstances, he rejects hypotheses; he rather gives his mind to each particular case in the light of his evidence and his ideas"

Fisher (Statistical Methods and Scientific Inference, 1956)

"We and others propose that a one-sided test of the null hypothesis that the true primary outcome is no different between treatment and control with a false-positive rate of 0.20 (type I error) is appropriate."

Ratain and Sargent (Eur. J Cancer, 2009)

Bayesian Hierarchical Meta-analysis of Case Fatality Rate Data (source: www.who,int)



Alternatives to Maximizing Power for Fixed Type I Error

"The extent to which scientific caution need be exercised and the importance of discovery of an effect (alternatively the cost of making type 1 and type 2 errors) will vary from situation to situation. This would imply that conventional significance levels should be abandoned and that with any particular piece of research a should be set with regard to the costs in hand"

Statistical Inference: A Commentary for the Social & Behavioural Sciences – M Oakes, 1986

"Conventionally the probability of type I error is set at 5% or less or as dictated by any adjustments made necessary for multiplicity considerations; the precise choice may be influenced by the prior plausibility of the hypothesis under test and the desired impact of the results."

ICH E9 (1998) - Statistical Principles for Clinical Trials

Alternative to Maximizing Power for Fixed Type-I Error

Journal of Biopharmaceutical Statistics, 22: 596–607, 2012

ISSN: 1054-3406 print/1520-5711 online
DOI: 10.1080/10543406.2011.564340

Taylor & Francis
Taylor & Francis Group

A PORTFOLIO-BASED APPROACH TO OPTIMIZE PROOF-OF-CONCEPT CLINICAL TRIALS

Craig Mallinckrodt[1], Geert Molenberghs[2], Charles Persinger[1], Stephen Ruberg[1], Andreas Sashegyi[1], and Stacy Lindborg[1]

- Choose α and β to minimise

$$C(\alpha, \beta) = \alpha \cdot [1 - p(E)] \cdot C_\alpha + \beta \cdot p(E) \cdot C_\beta$$

Alternative to Maximizing Power for Fixed Type-I Error

"goal of statistical testing is to aid us in making conclusions that limit the probabilities of making mistakes, whether Type I or II errors. We think a strong case can be made that in most studies … α should be set with the objective of either minimising the combined probabilities of making Type I or Type II errors at a critical effect size, or minimizing the overall cost associated with Type I and Type II errors given their respective probabilities"

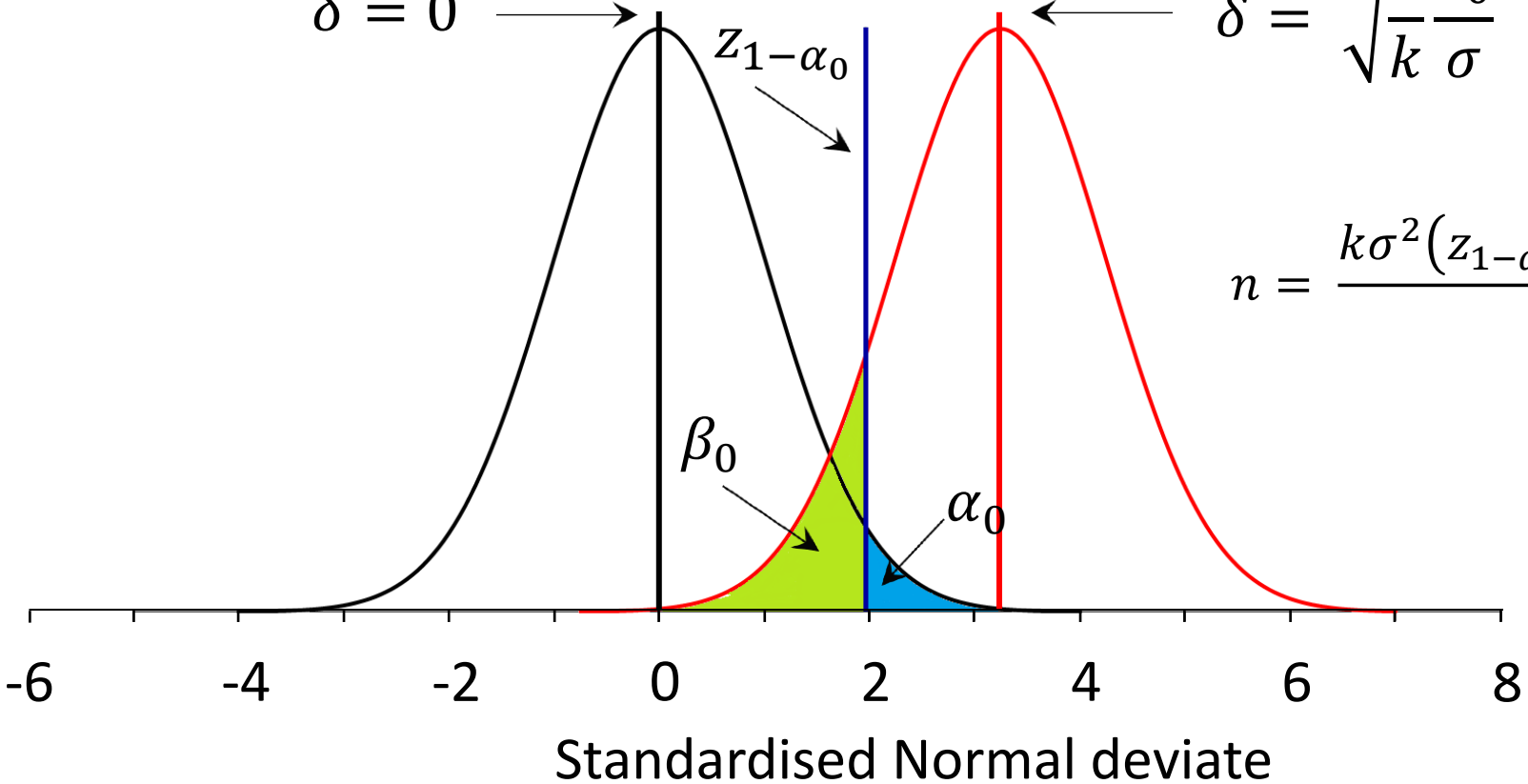
Mudge et al (PLoS, 2012)

Alternative to Maximizing Power for Fixed Type-I Error

- These suggestions correspond to :
  1. Minimise $\Psi = \frac{\alpha+\beta}{2}$
  2. Minimise $\Psi = \frac{\omega_0\alpha+\omega_1\beta}{\omega_0+\omega_1} = \frac{\omega\alpha+\beta}{\omega+1}$, where $\omega = \omega_0/\omega_1$ is the ratio of the costs of making the corresponding error.

(Mudge et al also consider the case where ω_0 and ω_1 are the prior probabilities associated with the null and alternative hypothesis.)

Determination of Sample Size



$$\delta = \sqrt{\frac{n}{k}}\frac{\delta_0}{\sigma}$$

$$n = \frac{k\sigma^2\left(z_{1-\alpha_0} + z_{1-\beta_0}\right)^2}{\delta_0^2}$$

Determining The Optimal α (1)

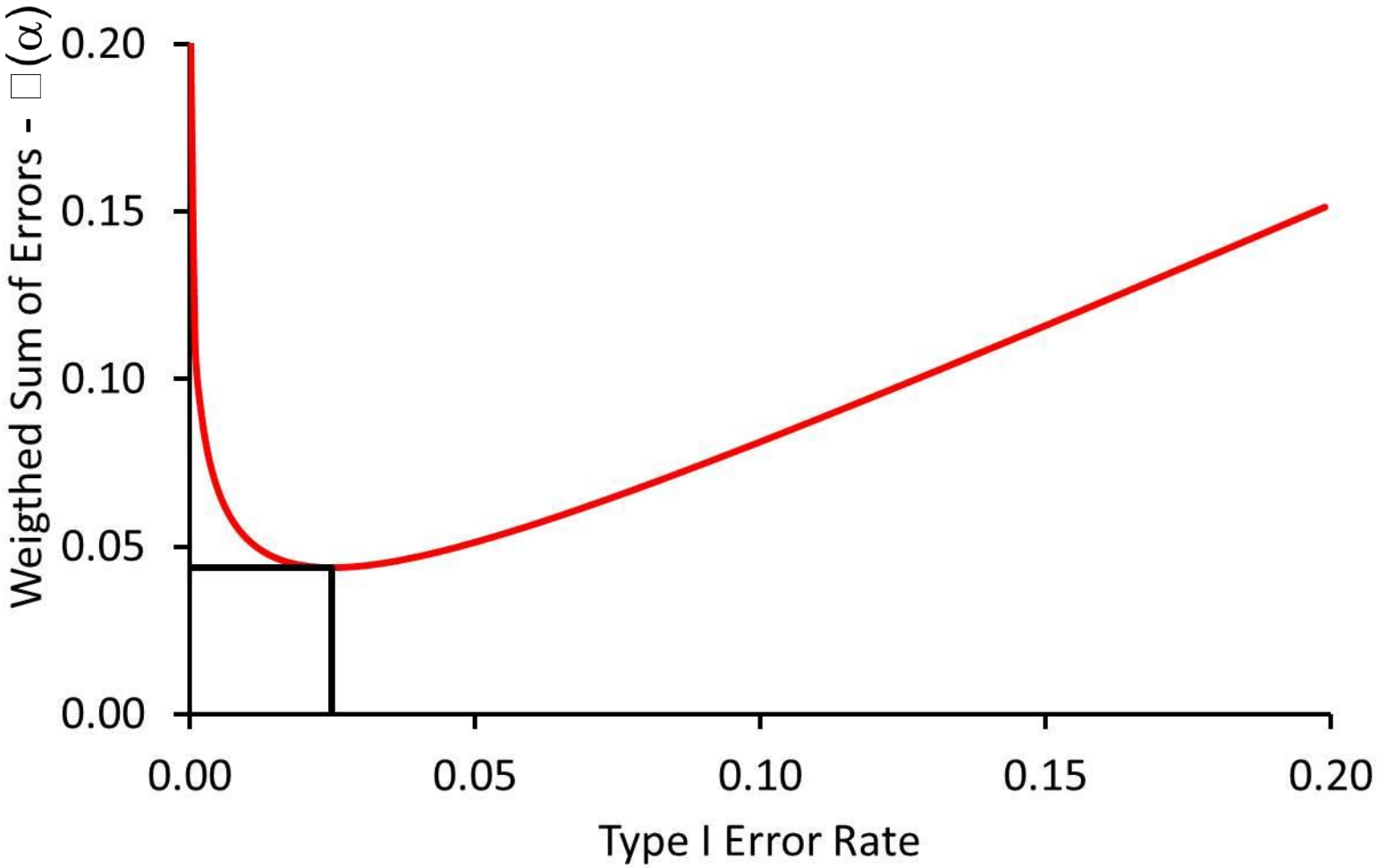
- For a given n, k, σ, α and δ_0 the probability of a type II error for testing H_0: $\mu = \mu_0$ vs H_1: $\mu = \mu_0+\delta_0$ is given by

$$\beta = 1 - \Phi\left(\sqrt{\frac{n}{k}}\frac{\delta_0}{\sigma} + Z_\alpha\right) = \Phi(\theta + Z_\alpha) \quad \left[\theta = \sqrt{\frac{n}{k}}\frac{\delta_0}{\sigma}\right]$$

- For a given weight ω – relative prior probabilities or ratio of costs – the weighted sum of the type I and type II error is

$$\Psi(\alpha) = \frac{\omega\alpha + 1 - \Phi(\theta + Z_\alpha)}{\omega + 1}$$

Weighted Sum of Error Rates as Function of α (k=1, σ=1, δ_0= $\sqrt{2}$, n=21, ω=3)



Determining The Optimal α (2)

- The minimum of this function occurs when

$$\alpha = \Phi\left(-\frac{\ln(\omega)}{\theta} - \frac{\theta}{2}\right) \quad \text{and} \quad \beta = 1 - \Phi\left(-\frac{\ln(\omega)}{\theta} + \frac{\theta}{2}\right)$$

- Minimum value

$$\Psi = \frac{\omega\Phi\left(-\frac{ln(\omega)}{\theta} - \frac{\theta}{2}\right) + \Phi\left(\frac{\ln(\omega)}{\theta} - \frac{\theta}{2}\right)}{\omega + 1}$$

($\omega = 1 \Rightarrow \alpha = \beta$)

Typical Values for Type I and Type II Rates and Implications for the Relative Costs of These Errors

- If n has been chosen on the basis of $n = \frac{k\sigma^2\left(z_{1-\alpha_0} + z_{1-\beta_0}\right)^2}{\delta_0^2}$ then given a value of ω the optimal value of α is given by

$$\alpha = \Phi\left(-\frac{\ln(\omega)}{\theta} - \frac{\theta}{2}\right)$$

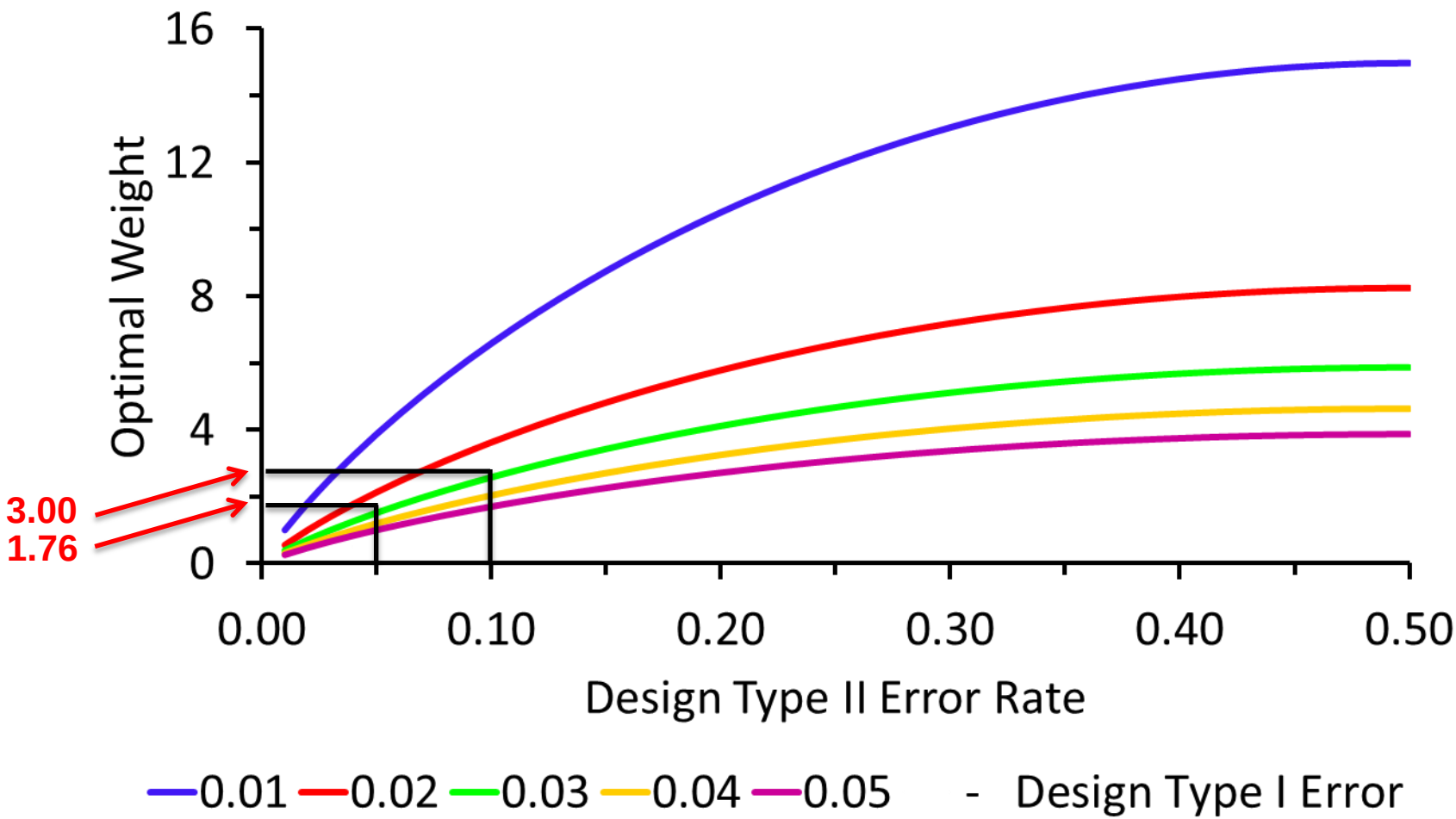
- For what value of ω does $\alpha=\alpha_0$?

$$n = \frac{k\sigma^2\left(z_{1-\alpha_0} + z_{1-\beta_0}\right)^2}{\delta_0^2} \Rightarrow \frac{\sqrt{n}\delta_0}{\sigma} = \theta = Z_{1-\alpha_0} + Z_{1-\beta_0}$$

and since

$$\omega = \frac{\phi(\theta + Z_\alpha)}{\phi(Z_\alpha)} \Rightarrow \omega = \frac{\phi\left(Z_{1-\beta_0}\right)}{\phi\left(Z_{\alpha_0}\right)}$$

Optimal Weights Giving Standard Type I and Type II Error Rates



Sample Sizing Based on Weighted Errors

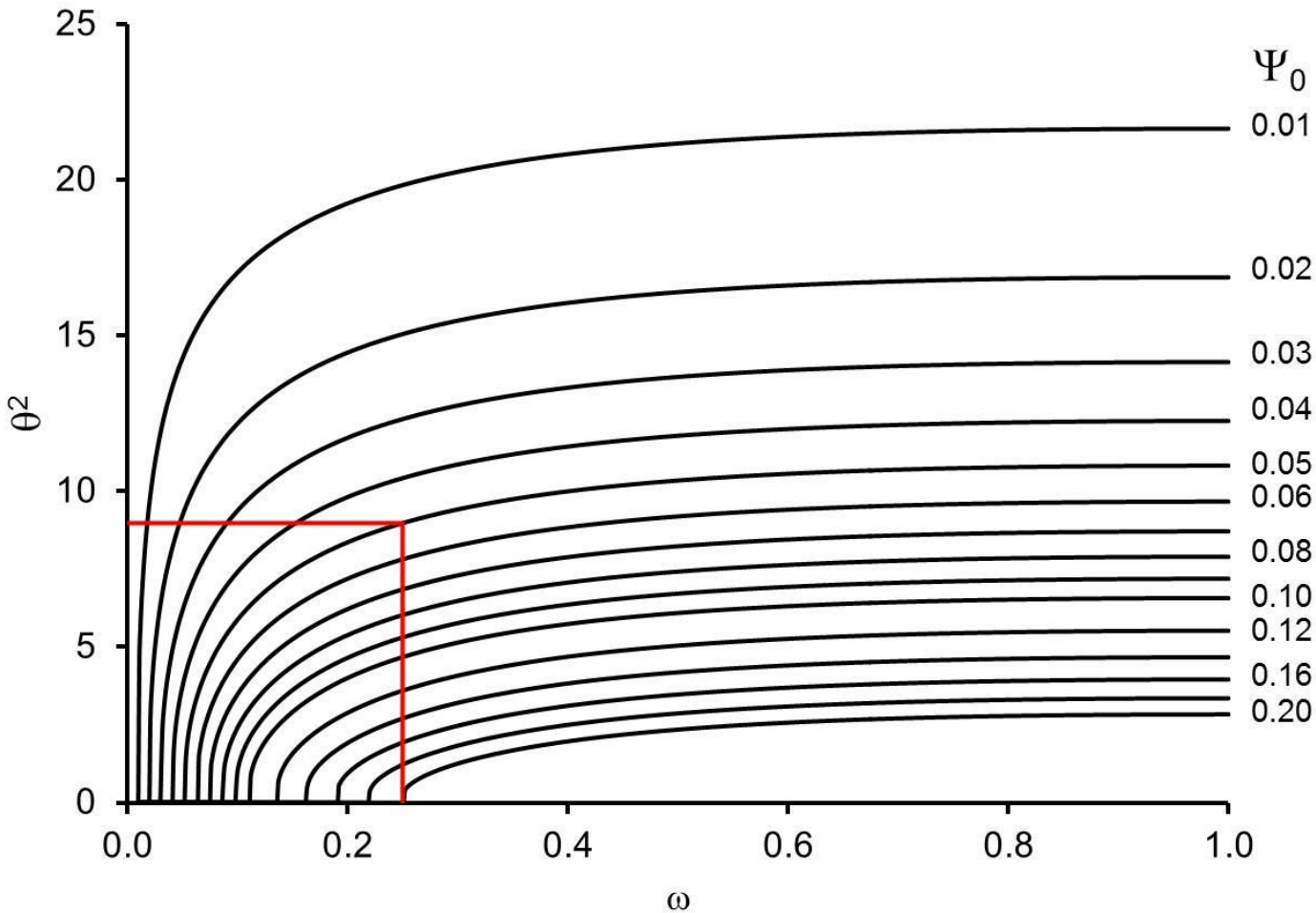
- Mudge et al (2012): "Alpha–beta optimization can also allow sample sizes to be estimated for a desired average probability or cost of error"
- How ?

$$\Psi = \frac{\omega\Phi\left(-\frac{ln(\omega)}{\theta} - \frac{\theta}{2}\right) + \Phi\left(\frac{\ln(\omega)}{\theta} - \frac{\theta}{2}\right)}{\omega + 1}$$

is a function ω and θ.

- If Ψ_0 is the maximum value of $\Psi(\omega,\theta)$ - solve $\Psi_0 = \Psi(\omega,\theta)$ in terms of θ^2
- The appropriate sample size is $n = k\theta^2\sigma^2/\delta_0^2$ which has the standard form for sample sizing $n = k\left(z_{1-\alpha_0} + z_{1-\beta_0}\right)^2\sigma^2/\delta_0^2$
- Must be solved numerically.

Sample Size Factor to Control the Weighted (ω or ω^{-1}) Sum of Error Rates to be $\leq \Psi_0$



Alternative Form of Neyman-Pearson Approach

- Neyman Pearson Lemma (1933) sought a critical region R(x) maximised the power 1-β.
- Suppose now we seek a critical region to minimise the weighted average of α and β – weights w_0 and w_1.

$$\Psi = \omega_0 \text{Prob(Type I error)} + \omega_1 \text{Prob(Type II error)}$$

$$= \omega_1 - \int_{R(x)} [\omega_1 p(x|H_1) - \omega_0 p(x|H_0)] dx$$

$$\Rightarrow R(x) = \{x: \omega_1 p(x|H_1) > \omega_0 p(x|H_0)\} \Rightarrow \frac{p(x|H_1)}{p(x|H_0)} > \frac{\omega_0}{\omega_1} = \omega$$

likelihood ratio

Simplest Case - One-Armed Study
Normal mean (k=1), known variance

- Null hypothesis - H_0: $\mu = \mu_0$

  Alternative hypothesis - H_1: $\mu = \mu_0 + \delta_0$

$$p(x; H_0) = (\sigma^2)^{-n/2} exp\left[-\frac{1}{2\sigma^2}\{(n-1)s^2 + n(\bar{x} - \mu_0)^2\}\right]$$

$$p(x; H_1) = (\sigma^2)^{-n/2} exp\left[-\frac{1}{2\sigma^2}\{(n-1)s^2 + n(\bar{x} - \mu_0 - \delta_0)^2\}\right]$$

$$\frac{p(x; H_1)}{p(x; H_0)} = exp\left\{-\frac{n}{2\sigma^2}[-2(\bar{x} - \mu_0)\delta_0 + \delta_0^2]\right\} > \omega$$

$$\Rightarrow \frac{\sqrt{n}(\bar{x} - \mu_0)}{\sigma} > \sqrt{n}\frac{\delta_0}{2\sigma} + \frac{\sigma}{\sqrt{n}\delta_0} ln(\omega) = \frac{\theta}{2} + \frac{ln(\omega)}{\theta}$$

Likelihood Principle

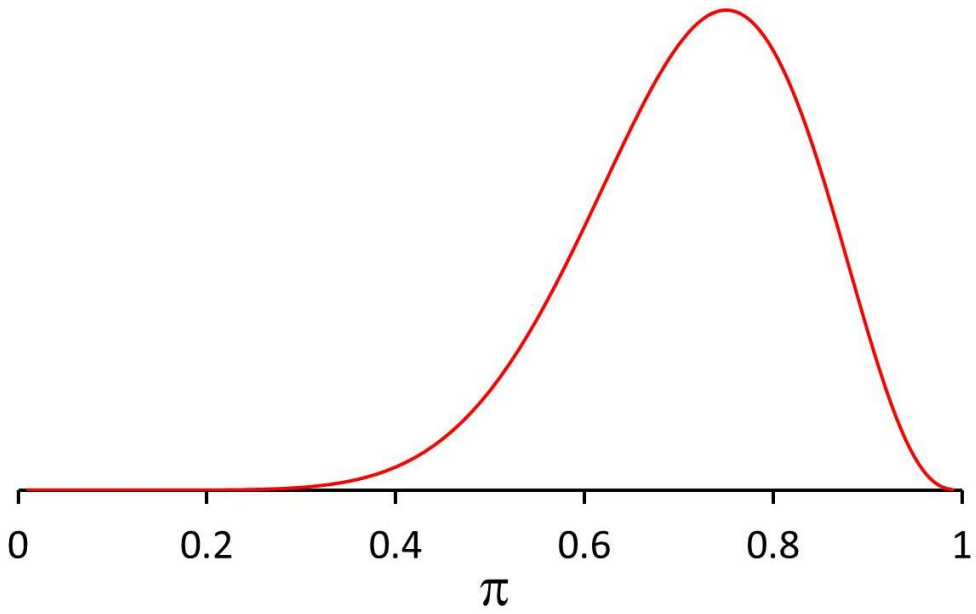
- The likelihood principle says that how the data are arrived at is irrelevant to the inferences that are to be drawn.
- e.g. a single arm, open-label, clinical trial is run and the outcome is binary, success or failure – perhaps a phase II oncology study.

Likelihood Principle – 4 Scenarios

Scenario	p-value
1. Fixed sample study 12 patients are treated; of these 9 respond successfully. H_0: π=0.5.	1. $\sum_{k=9}^{12} \binom{12}{k} 0.5^{12}$ = **0.073**
2. Patients to be treated until 3 treatment failures. The 3rd failure occurs when 12 patients have been treated. H_0: π=0.5.	2. $\sum_{k=9}^{\infty} \binom{k+3-1}{k} 0.5^{k+3}$ = **0.033**
3. Patients to be recruited for 2 weeks at which 12 patients treated with 9 successes.	3. What is basis for a p-value? Assume number of patients recruited is Poisson with mean 10. What are more extreme cases: 8/10 & 13/15? If so, p-value is **0.079**. If mean is 5, p=**0.180**; if mean=20, p=**0.018**
4. Plan to recruit 50 patients but funding runs out after 12 patients treated with 9 successes.	4. No idea

Likelihood Function

- For some scenarios the calculation of the p-value was simple, for others more complicated and for Scenario 4. perhaps impossible. Despite these difficulties the likelihood function for the unknown success proportion π is the same for each scenario: $\pi^9(1-\pi)^3$



Link to Bayesian Inference

- Priors
  - Null - $P(H_0\colon \mu = \mu_0) = \pi_0$
  - Alternative - $P(H_1\colon \mu = \mu_0 + \delta_0) = \pi_1$

- Bayes theorem :

$$P(H_0|x) = \frac{\pi_0 p(x|H_0)}{\pi_0 p(x|H_0) + \pi_1 p(x|H_1)}$$

- $P(H_0|x) < 0.5$ =>

$$\frac{p(x|H_1)}{p(x|H_0)} > \frac{\pi_0}{\pi_1}$$

(Pericchi and Pereira, 2012, Unpublished)

Discussion

- This is not new - Savage & Lindley, Cornfield (1960s), DeGroot (1970s), Bernardo & Smith (1990s), Perrichi & Pereira (2012, 2013) -> solves Lindley ‘s paradox.
- Cornfield(1966) showed that minimising the weighted errors is also appropriate in sequential (adaptive) trials.
- Spieglehalter, Abrams & Myles (2004) quote Cornfield “the entire basis for sequential analysis depends upon nothing more profound than a preference for minimizing β for given α rather than minimizing their linear combination. Rarely has so mighty a structure and one so surprising to scientific common sense, rested on so frail a distinction and so delicate a preference.”