



How To Test Hypotheses If You Must

Andy Grieve
VP Innovation Centre

PSI Journal Club, Sponsored by Wiley
16th September 2015

BASIC AND APPLIED SOCIAL PSYCHOLOGY, 37:1–2, 2015
Copyright © Taylor & Francis Group, LLC
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

“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”

Robert Matthews

Sunday Telegraph, 13 September 1998

2012 Ecology Papers on Significance Testing

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.

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.

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

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)

“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 α 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.”

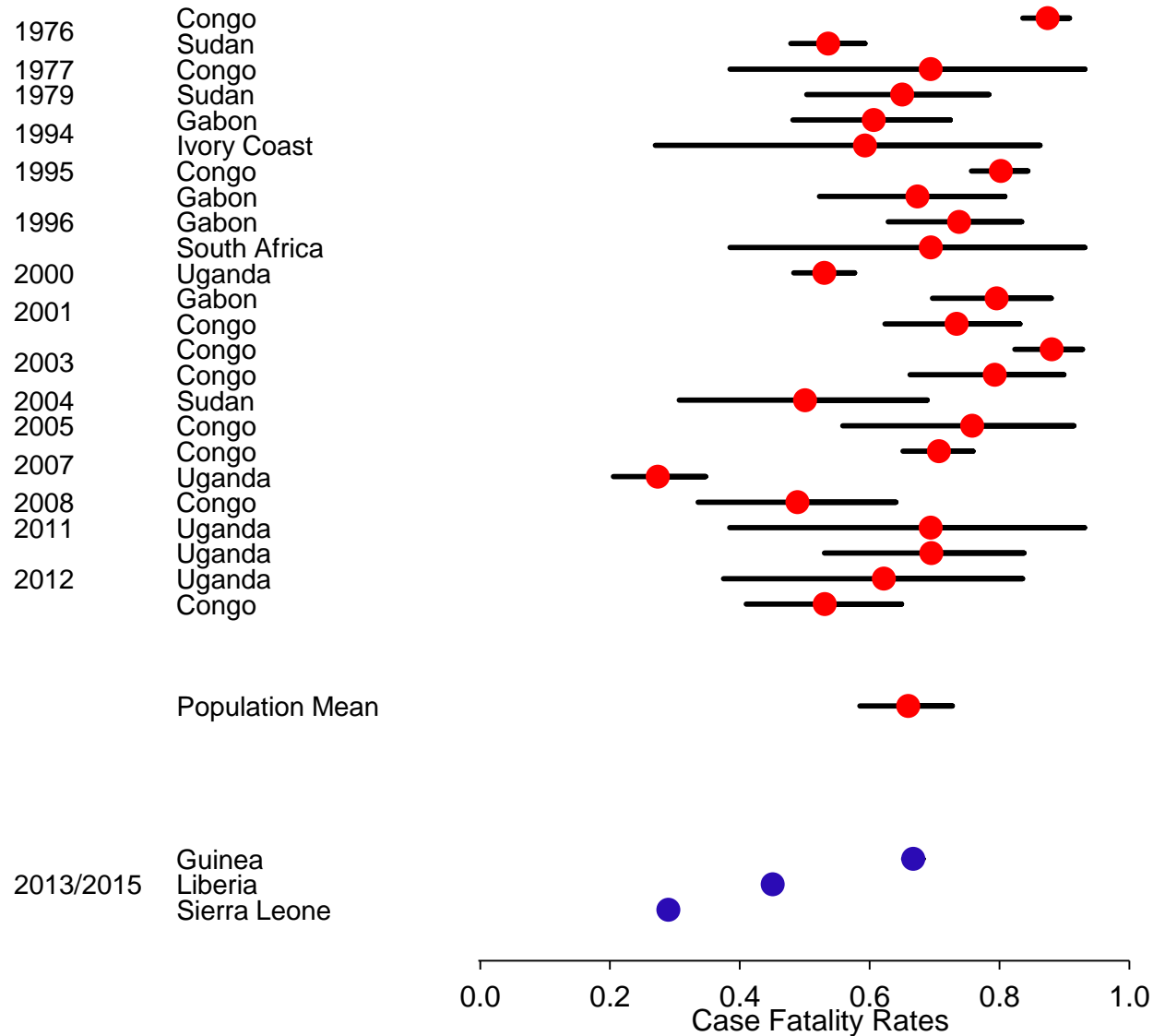
ICH E9 (1998) - Statistical Principles for Clinical Trials

. . Two sources of error can rarely be eliminated completely; in some cases it will be more important to avoid the first, in others the second. Is it more serious to convict an innocent man or to acquit a guilty? That will depend upon the consequences of the error; is the punishment death or fine; what is the danger to the community of released criminals; what are the current ethical views on punishment? From the point of view of mathematical theory all that we can do is to show how the risk of the errors may be controlled and minimised. The use of these statistical tools in any given case, in determining just how the balance should be struck, must be left to the investigator.

“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)

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



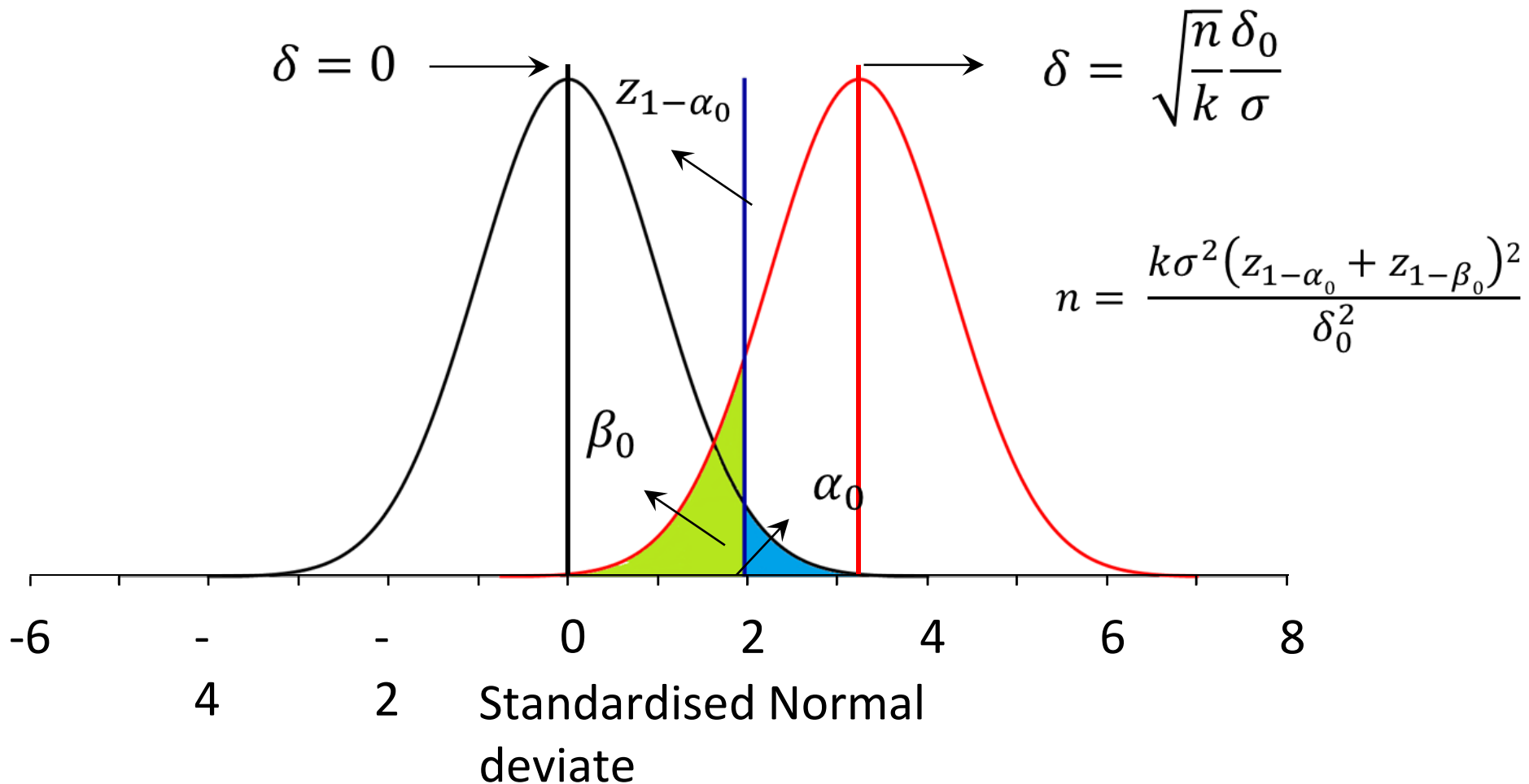
● These suggestions correspond to :

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



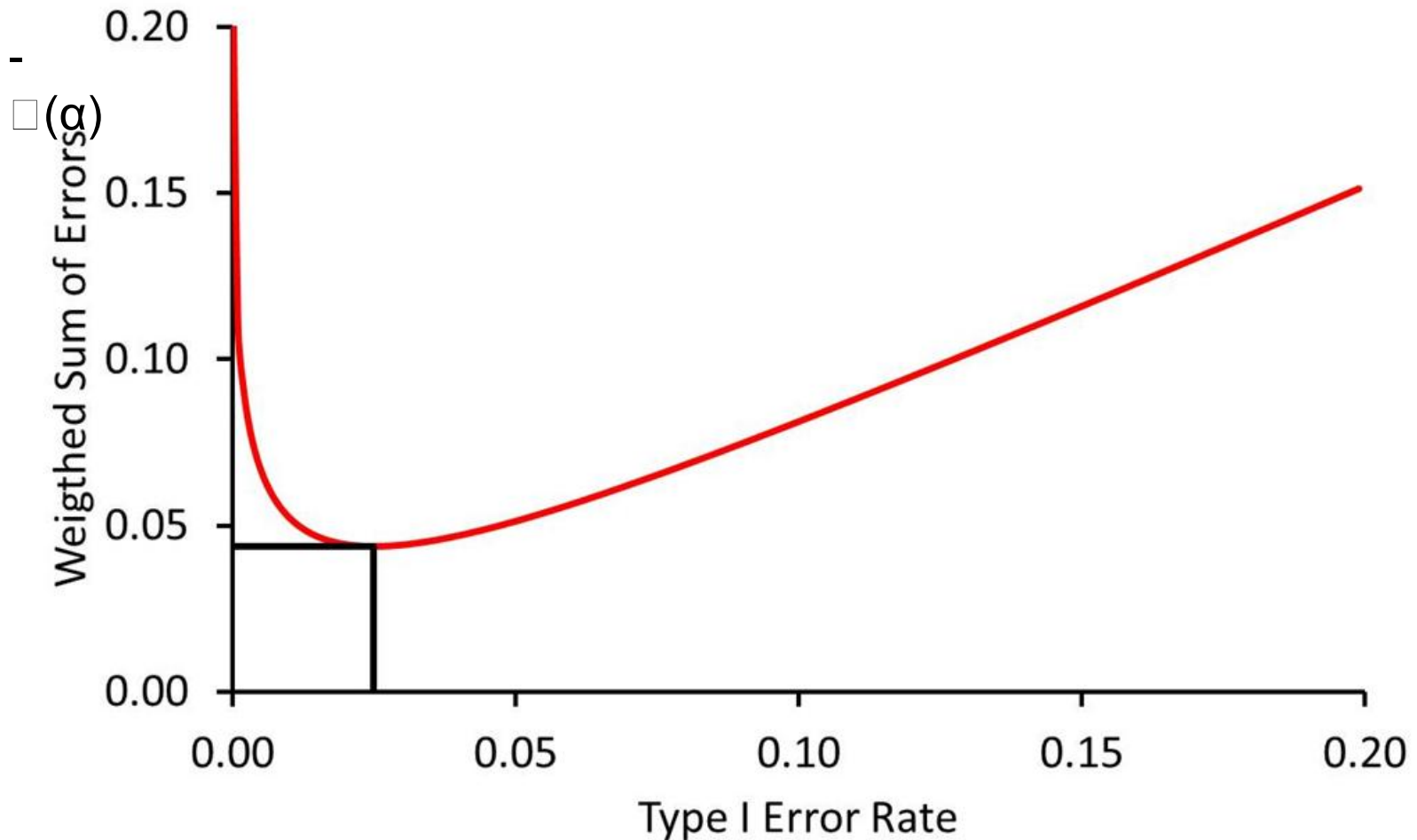
- For a given n , k , σ , α and δ_0 the probability of a type II error for testing $H_0: \mu = \mu_0$ vs $H_1: \mu = \mu_0 + \delta_0$ is given by

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

- 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 ($k=1$, $\sigma=1$, $\delta_0=\sqrt{2}$, $n=21$, $\omega=3$)



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

then given a value of ω the optimal value of α is given by

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

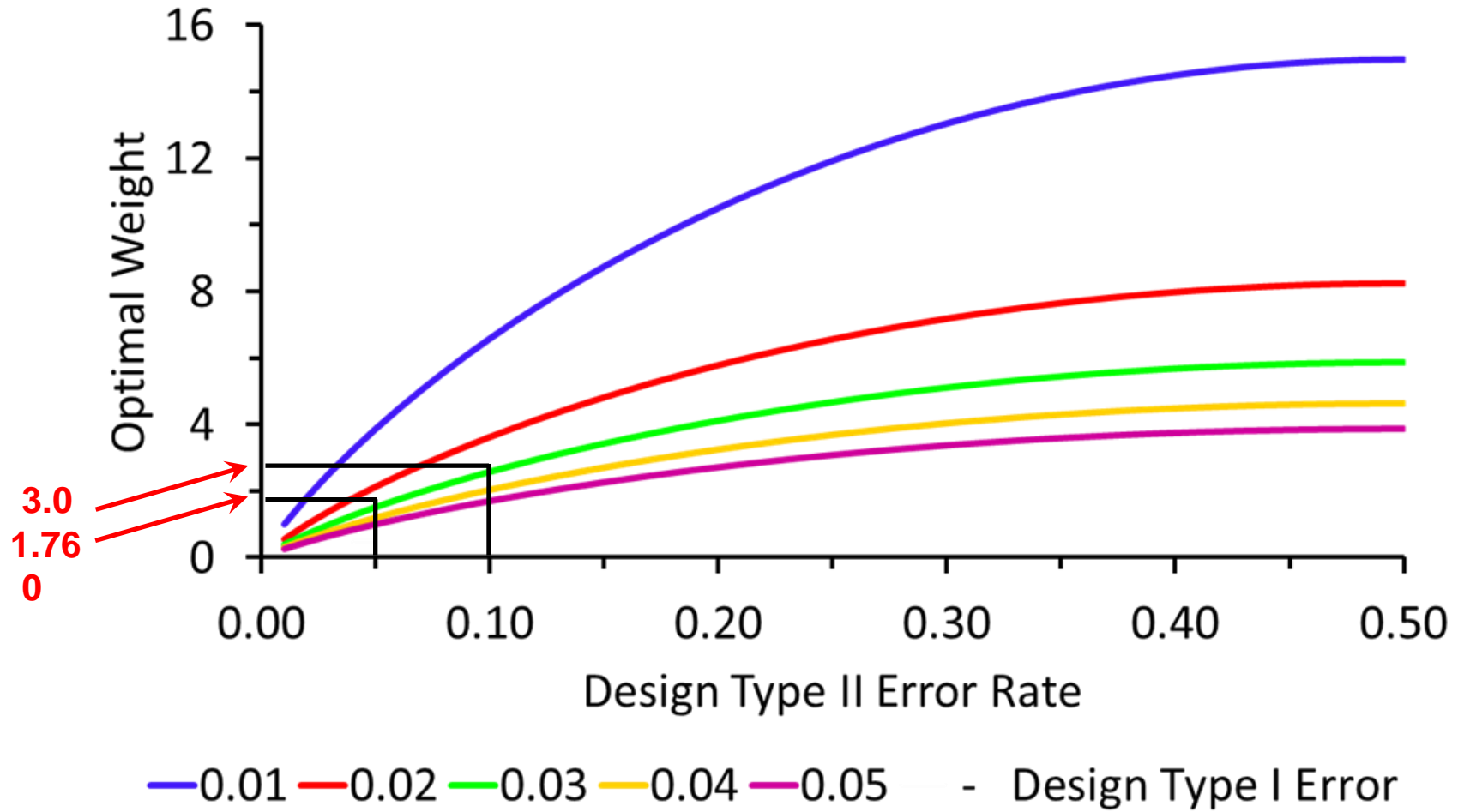
- For what value of ω does $\alpha = \alpha_0$?

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

and since

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

Optimal Weights Giving Standard Type I and Type II Error Rates



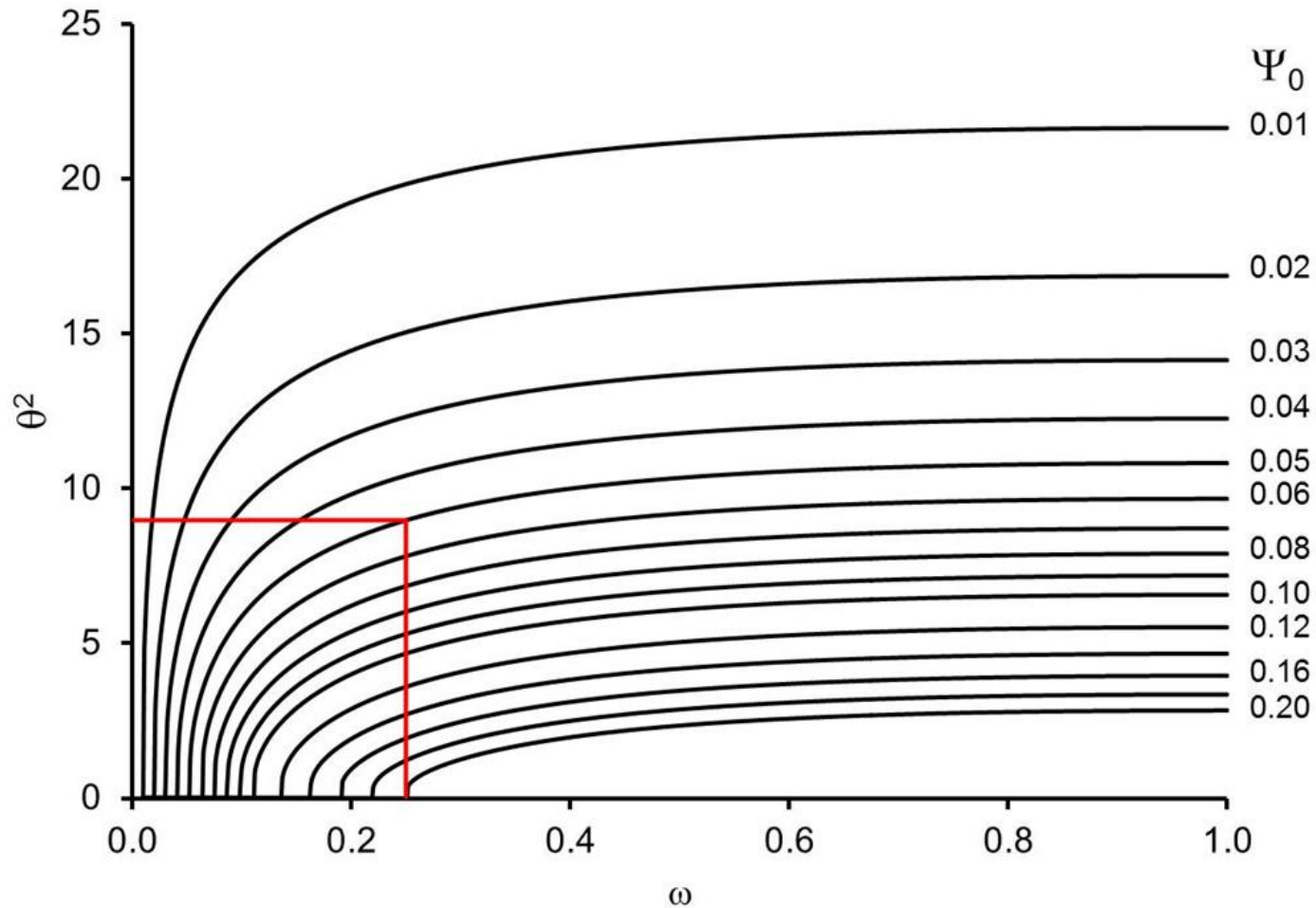
- **Mudge et al (2012)**: “Alpha–beta optimization can also allow sample sizes to be estimated for a desired average probability or cost of error”

- 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 $n = k(z_{1-\alpha_0} + z_{1-\beta_0})^2\sigma^2/\delta_0^2$
- Must be solved numerically.

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




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

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

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

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

likelihood ratio 

Simplest Case - One-Armed Study

Normal mean (k=1), known variance

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

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

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

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

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

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

- 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.

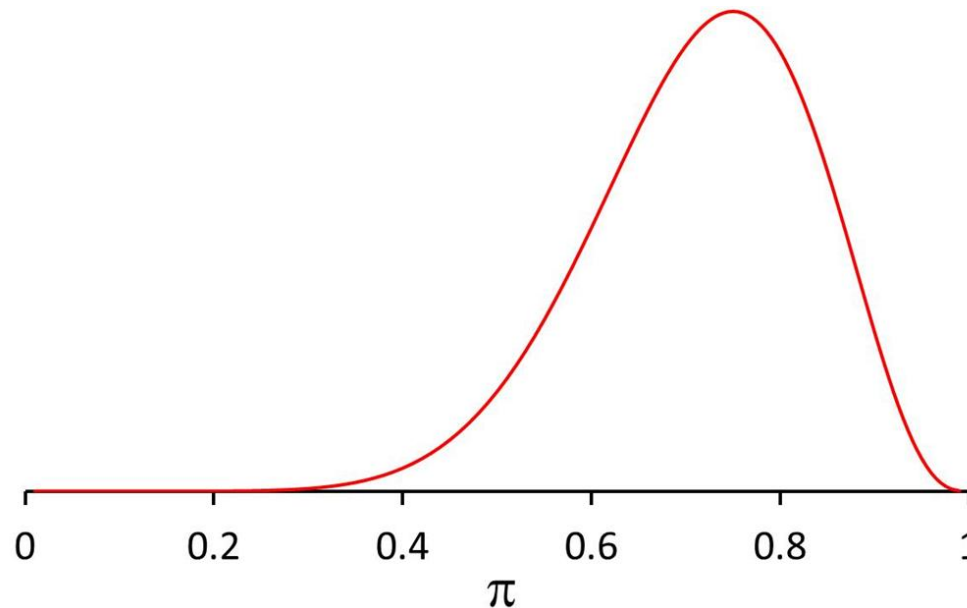
Scenario

1. Fixed sample study 12 patients are treated; of these 9 respond successfully. $H_0: \pi=0.5$.
2. Patients to be treated until 3 treatment failures. The 3rd failure occurs when 12 patients have been treated. $H_0: \pi=0.5$.
3. Patients to be recruited for 2 weeks at which 12 patients treated with 9 successes.
4. Plan to recruit 50 patients but funding runs out after 12 patients treated with 9 successes.

p-value

1. $\sum_{k=9}^{12} \binom{12}{k} 0.5^{12} = \mathbf{0.073}$
2. $\sum_{k=9}^{\infty} \binom{k+3-1}{k} 0.5^{k+3} = \mathbf{0.033}$
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, $p=\mathbf{0.180}$; if mean=20, $p=\mathbf{0.018}$
4. No idea

- 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: $\pi^9(1-\pi)^3$



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

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

- $P(H_0|x) < 0.5 \Rightarrow \frac{p(x|H_1)}{p(x|H_0)} > \frac{\pi_0}{\pi_1}$

(Pericchi and Pereira, 2012, Unpublished)

- 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.”