

Workshop 1:

Sample Size Determination- Methodology and Philosophy

NCIC Clinical Trials Group
NCIC Groupe des essais cliniques



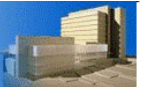
Dr CJ O'Callaghan

NCIC Clinical Trials Group
NCIC Groupe des essais cliniques



or

*“How many do
I need ?....”*



Objectives

- Not a statistics or programming course!
- Enough information to enable you to:
 - understand (\pm critique) what you read in the medical literature.

“In order to have 90% power to detect a hazards ratio of 1.33 between the two treatment arms (an improvement of median survival from 6 to 8 months), using a two-sided 5% level test, a minimum of 520 deaths will be needed before the final analysis.”

- think clearly about your own research **before**, during and after data collection and identify some common pitfalls.
- know what your input should be when seeking additional statistical assistance for study design / sample size.



Statisticians are like drunks leaning against the lamp post - they are there for support, not illumination.



Sample Size in Medical Trials

"How many subjects are needed to assure a given probability of detecting a statistically significant effect, of a given magnitude, if one truly exists?"

What is the...

- smallest effect worth detecting?
 - Clinical relevance
- acceptable risk of seeing it, if it doesn't exist?
 - Statistical significance level α , Type I error
- acceptable risk of missing it, if it exists?
 - Power β , Type II error ($1-\beta$)



Experimental Errors

State of Nature (Reality)

Results of
Statistical Analysis

	No Effect	Effect
No Effect	<p>No Effect</p> <p>✓</p> <p>'Accept' null hypothesis when it is true</p>	<p>Type II</p> <p>(β) error</p> <p>'Accept' null hypothesis when it is false</p>
Effect	<p>Type I</p> <p>(α, p) error</p> <p>Reject null hypothesis when it is true</p>	<p>✓</p> <p>Reject null hypothesis when it is false</p>



Sample Size Calculations

- Define null and alternative hypotheses
 - determine minimum difference to be detected or of interest
- Specify type I error (significance level)
- Specify type II error (power)
 - specify sample size and determine power...



Statistical Hypotheses

- An experiment or set of observations never **proved** anything.
- The purpose of statistical tests, is to determine if the obtained results provide a reason to reject the ***hypothesis*** that they are merely a product of chance factors.
 - Null Hypothesis: H_0
 - Alternate Hypothesis: H_A



Induction and Deduction

- White Swans

"No matter how many instances of white swans we may have observed, this does not justify the conclusion that all swans are white"

Sir Karl Popper



- A black one may be lurking just around the corner?



Define H_0 and H_a

- H_0 is the reverse of what we hope/believe; it is put forward to allow the data to contradict it and is typically a hypothesis of no difference or no effect.
- Examples (different endpoints/objectives):
 - Comparing two means
 - $H_0: M_1 = M_2$ vs $H_a: M_1 > M_2$ (or $M_1 \neq M_2$)
 - Comparing two proportions
 - $H_0: P_1 = P_2$ vs $H_a: P_1 > P_2$ (or $P_1 \neq P_2$)
 - Comparing two survival functions
 - $H_0: S_1 = S_2$ for all t vs $H_a: S_1 > S_2$ (or $S_1 \neq S_2$) for some t



Clinical “Significance”

- a.k.a. a clinically meaningful difference
- statistical significance is necessary but not sufficient for clinical significance
- depends on implications of detected difference (e.g. 1 week improvement in median overall survival**)
- *“given a large enough sample size, you will likely detect a statistically significant difference”*



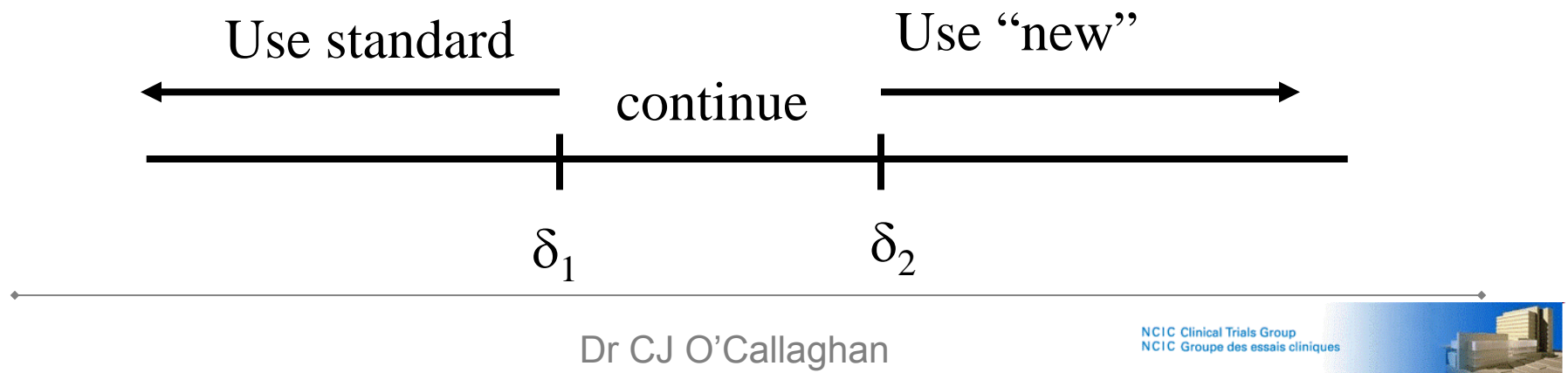
Minimum Difference to be Detected

- This difference can be the difference that:
 - is likely to be present
 - would make a difference to clinical practices
- Determine minimum clinically important difference
 - Previous results
 - Pre-clinical or pilot studies
 - Clinical experiences and judgments



Freedman *et al.* Formulation

- What is the minimum improvement from the new treatment which would lead you to adopt the “new” treatment as routine (δ_2)?
- What is the maximum improvement from the new treatment which would lead to your retention of the standard treatment as routine (δ_1)?



Significance Level

- In hypothesis testing, the significance level is the criterion used for rejecting the null hypothesis.
- The significance level is used in hypothesis testing as follows:
 - The difference between the results of the trial and H_0 is determined.
 - Assuming H_0 is true, the probability of a difference that large or larger is computed .
 - This probability (p) is compared to the significance level (α). If $p \leq \alpha$, then H_0 is rejected and the outcome is said to be statistically significant.



Significance Level

- Traditionally, either the 0.05 level (sometimes called the 5% level) or the 0.01 level (1% level) have been used, although the choice of levels is largely subjective.
- The lower the significance level, the more the data must diverge from the null hypothesis to be significant. Therefore, the 0.01 level is more conservative than the 0.05 level... but not a linear relationship.



An Aside: Probability Value

p-value versus α

- In hypothesis testing, the probability value (sometimes called the p value) is the **probability** of obtaining a statistic as different from or more different from the parameter specified in H_0 as the statistic obtained in the experiment.
- The significance level (α) is an arbitrary **threshold** for comparison / decision



$$p < 0.05$$



p = 0.049



p = 0.059



$p < 0.0000000000000001$



“Statistics are like a bikini.



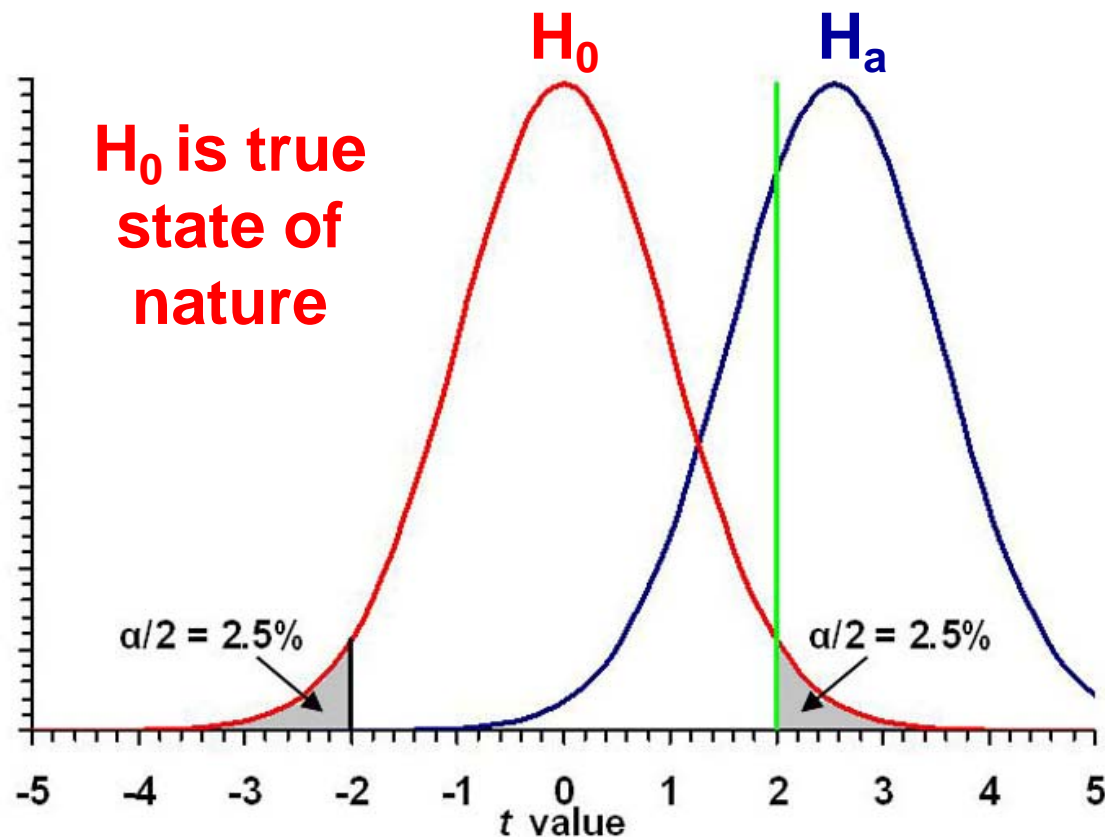
*What they reveal is suggestive, but
what they conceal is vital”*

Aaron Levenstein



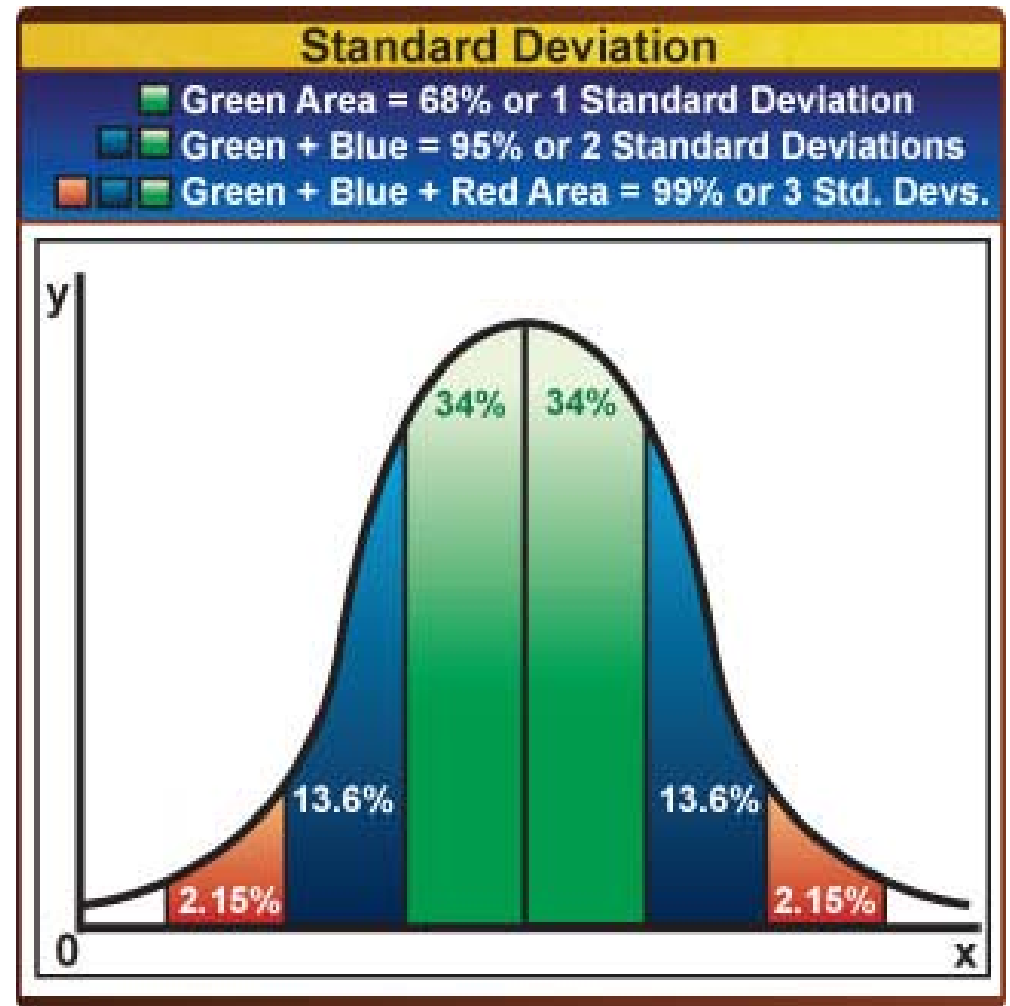
Type I error (α)

- Probability of falsely rejecting H_0 (probability of rejecting the null when null is true)
- Consumer's or Regulatory risk, "*False Discovery Rate*"



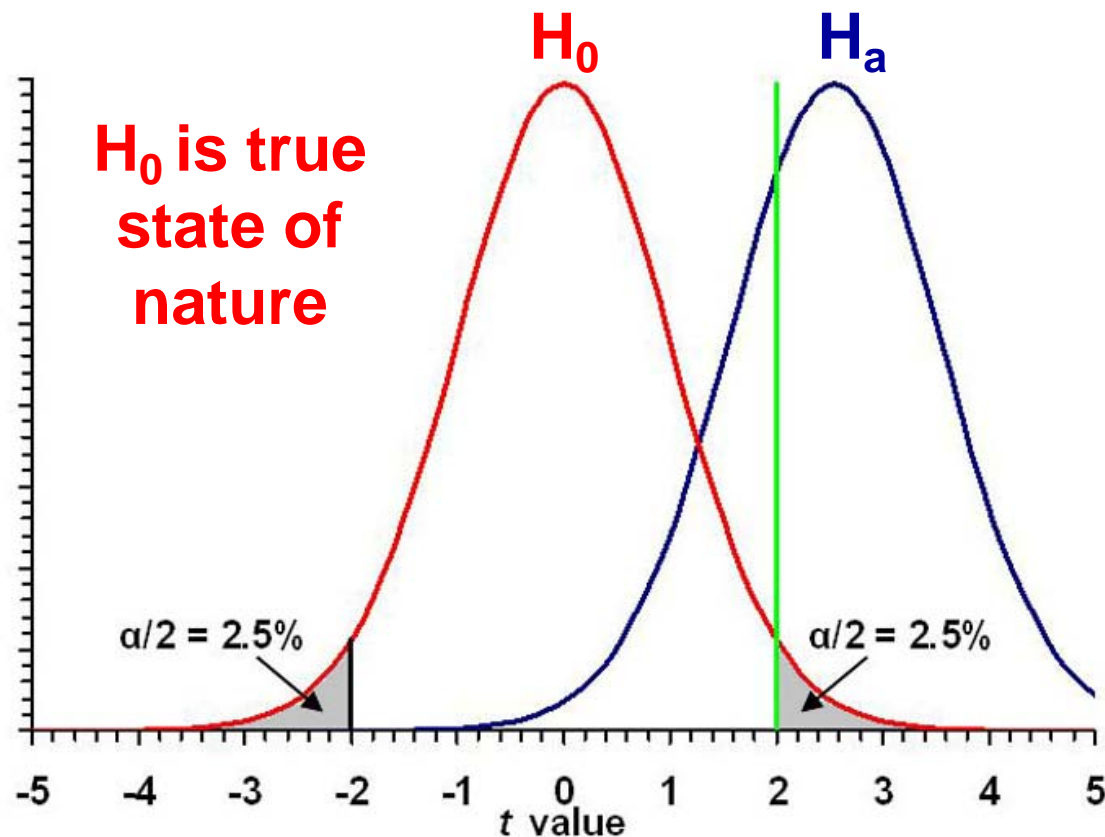
Aside: Sampling Distribution

- a sampling distribution is the **probability distribution** of a given statistic based on a random sample of certain size n .
- It may be considered as the distribution of the statistic for all possible samples of a given size.
- The sampling distribution depends on the underlying distribution of the population, the statistic being considered, and the sample size used.



Type I error (α)

- Probability of falsely rejecting H_0 (probability of rejecting the null when null is true)
- Consumer's or Regulatory risk, "*False Discovery Rate*"



1-sided vs 2-sided Alternatives

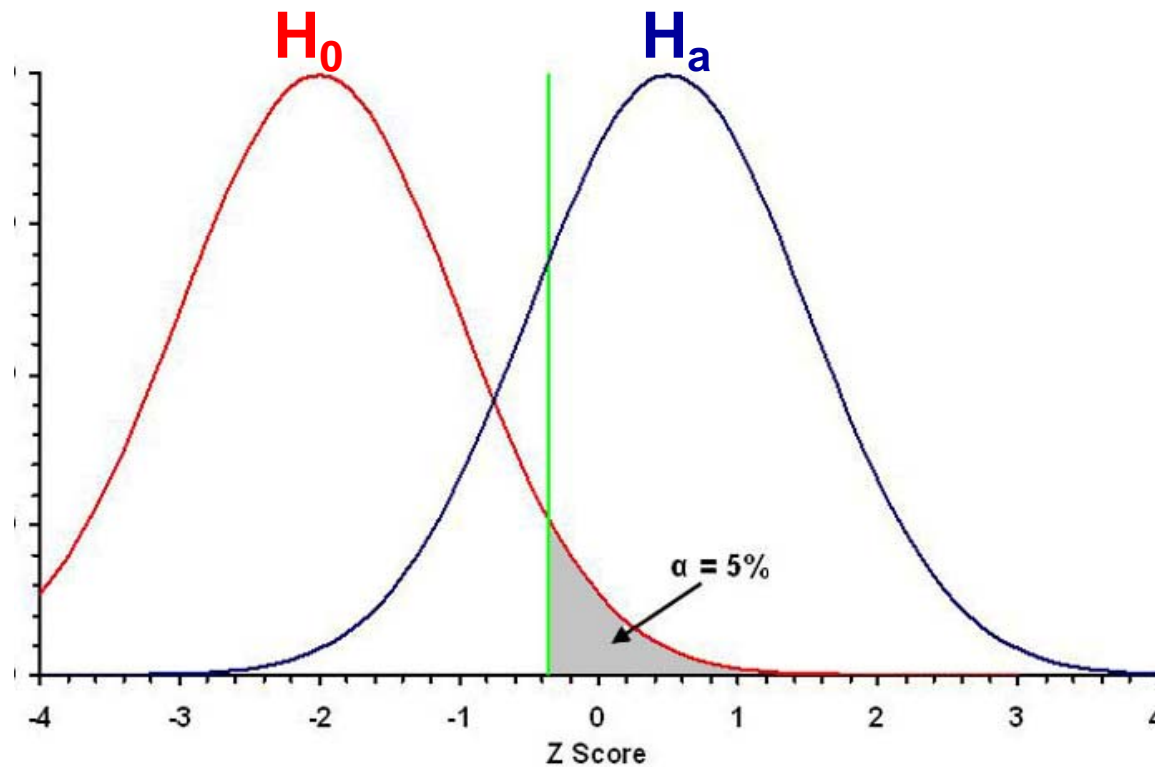
- Use one-sided test if you “know” the experimental arm is better than the standard arm (then why do you need a RCT?) or you are only interested in this type of question
- ...but if the null hypothesis is not rejected, it cannot tell whether experiment arm is worse than the standard arm
- FDA’s position is two-sided alternatives for almost all studies
- For hypothesis generation, a two-sided test should be used



One-sided α

- “Lowers the bar” for the same apparent degree of risk
- Implies knowledge which may not necessarily be assumed
- Cheating? – a one-sided test could make ‘significant’ a non-significant two-sided test

**H_0 is true
state of
nature**



Corollary: Accepting the Null Hypothesis?

- A null hypothesis is not accepted just because it is not rejected.
- Data not sufficient to show convincingly that a difference between arms of a trial is not zero do not prove that the difference is zero.
- Such data may even suggest that the null hypothesis is false but not be strong enough to make a convincing case, for example if the probability value were $p=0.08$
- H_0 may or may not be true, there just is not strong enough evidence to reject it
- so called “*trending toward significance*”, a.k.a. “*pilot study*”



Minimum difference to be detected

- A negative result (i.e., when the null hypothesis is not rejected by the data) does not indicate the two arms are the same
- It only means that the actual difference is less than what we intended to detect and our sample size is not large enough to detect this difference
- A study should have enough **power** to detect a minimum difference which is clinically important



Power, Type II error (β)

- Traditionally, power is fixed *a priori*, usually at 0.80 ($1-\beta$) with the chance of a Type II error (β) at 0.20
- Few studies are powered greater than 90% but **MANY** have lower power
- Affects the credibility of “negative” studies
 - Medical *versus* Ecological implications





On Power:

- Here are the results of our drug testing study on rabbits:
- 1/3 of the sample died
- 1/3 of the sample survived
- the other one got away



“To call in the statistician after the experiment is done may be no more than asking him to perform a postmortem examination: he may be able to say what the experiment died of.”

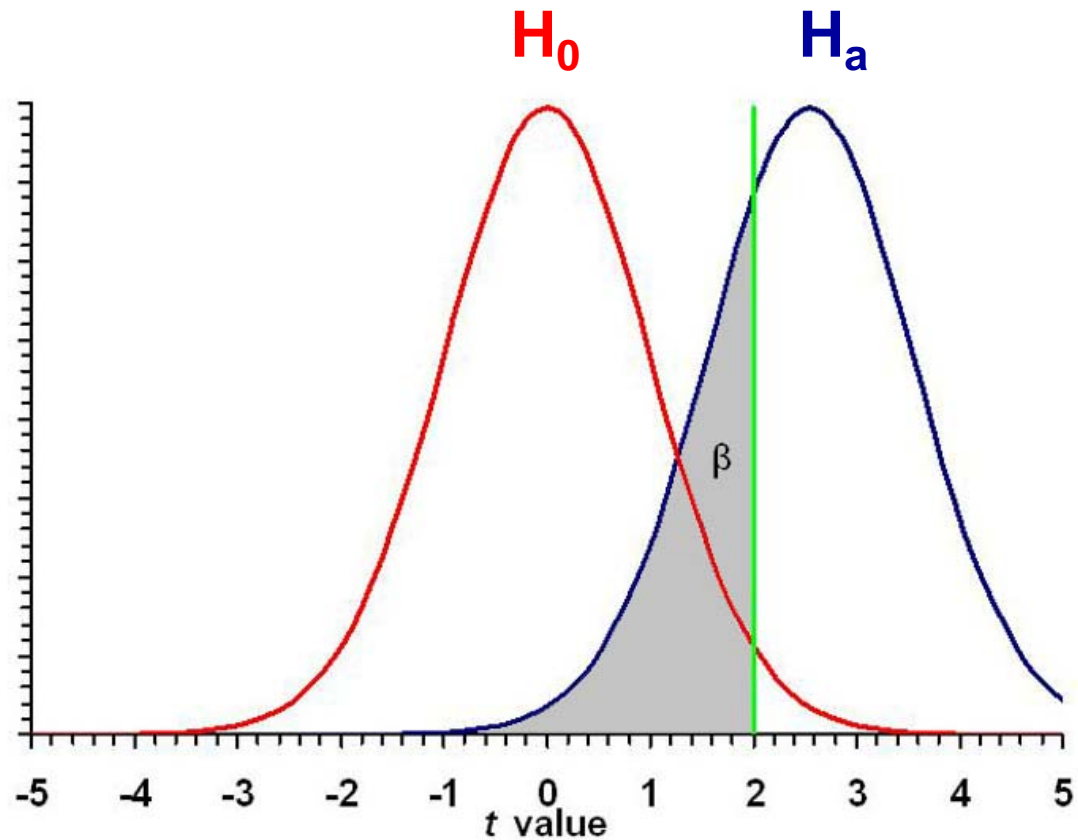


Sir R.A Fisher



Type II error (β)

- Probability of falsely accepting H_0 (probability of failing to reject H_0 given that H_a is true)
- Sponsor's or investigator's risk

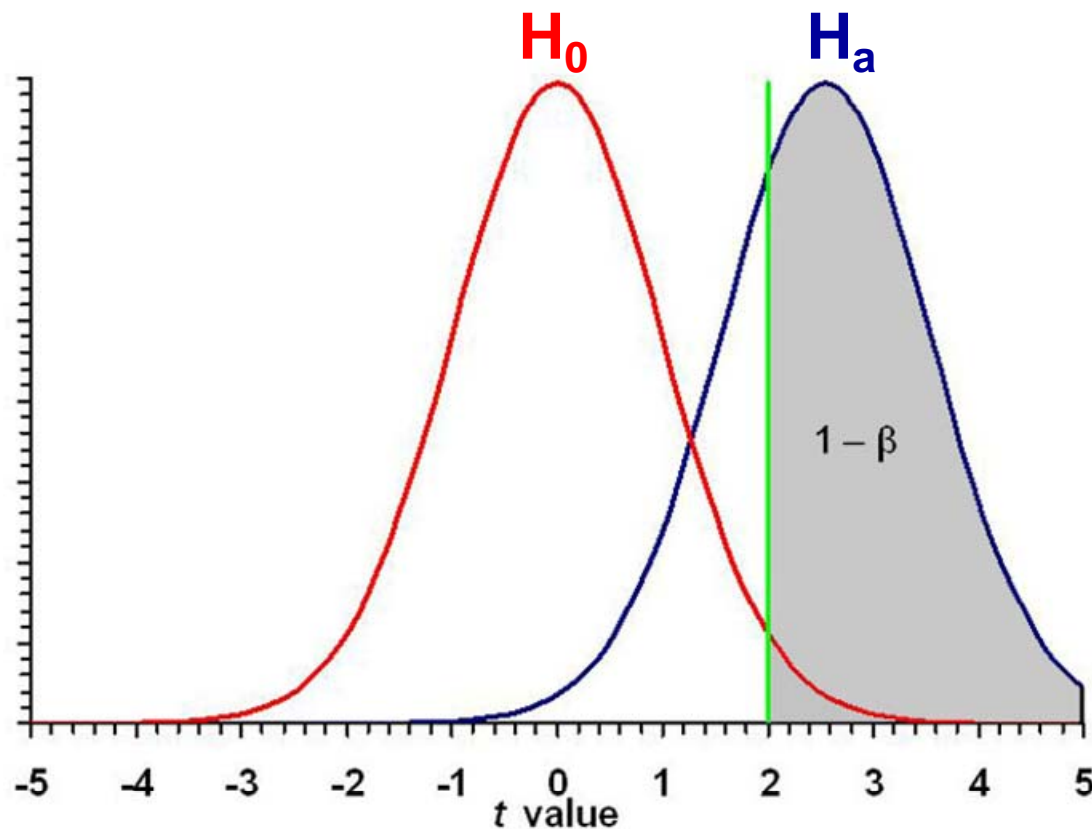


**H_a is true
state of
nature**



Power ($1-\beta$)

- Probability of correctly reject H_0 (probability of rejecting the H_0 given that H_a is true)
- Power=1-type II error

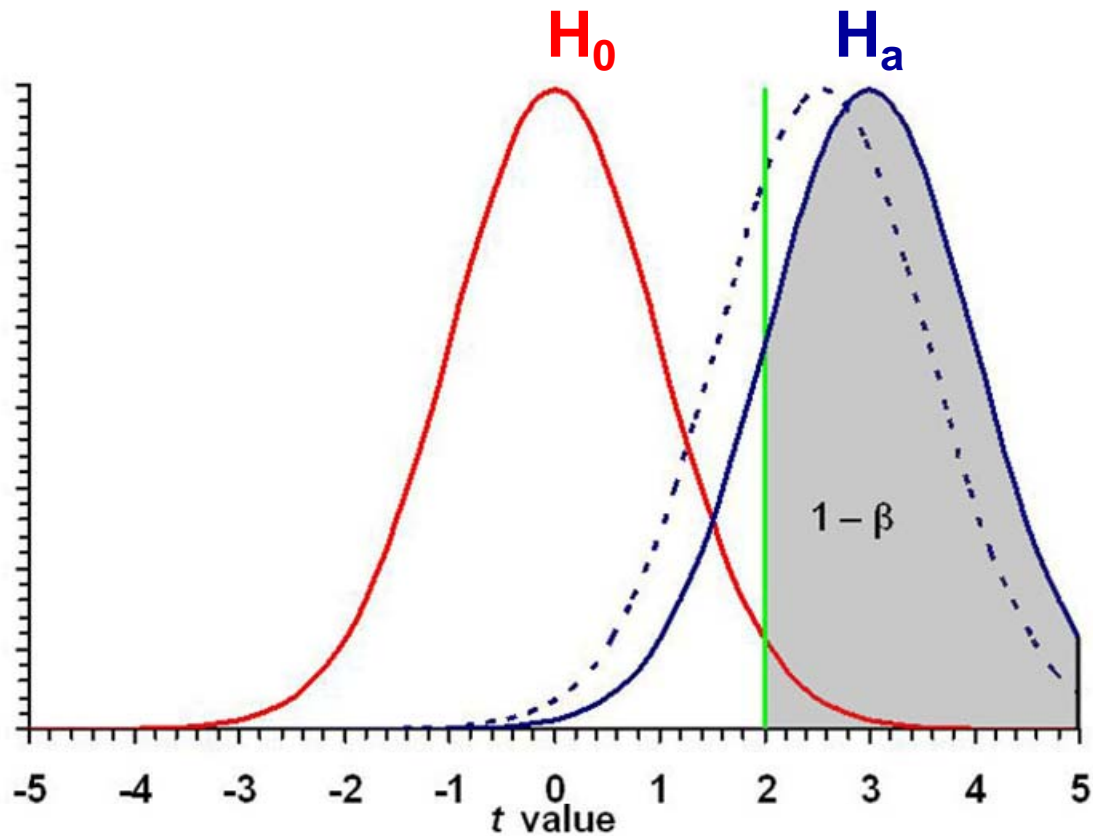


**H_a is true
state of
nature**



Power ($1-\beta$)

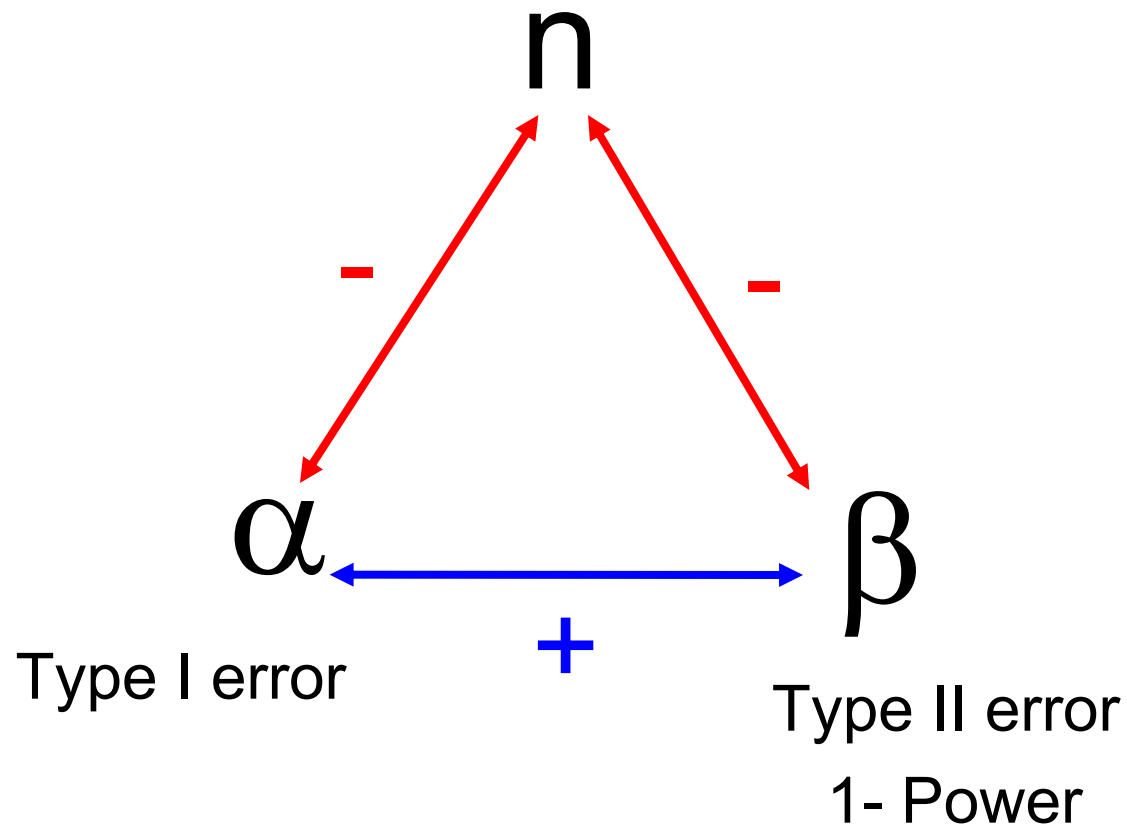
- How to increase power?
 - Increase N
 - Increase minimum detectable difference



**H_a is true
state of
nature**



Sample Size (n), α , β



Calculating a Sample Size

- The most difficult - and important - aspect of “sizing” a study is not the mathematics of sample size calculation
- - it’s deciding what the really relevant outcome measure is and what difference in that measure the trial will be designed to detect.



Sample Size Description for a Difference in Times to Events

In order to have 80% power to detect a hazard ratio of 1.28 (i.e. an improvement of 4% disease-free survival from 80% at 4 years) using a two sided 5% level test, the maximum number of recurrences we would need to observe is 523. Assuming we could enter 2380 patients in 2 years, we would need to follow all patients for about 4 years before the final analysis. The maximum total duration of the trial would be 6 years. If the risk of relapse for the control group is much lower, with 2380 patients entered in two years followed for an additional four years, we would have 80% power to detect a hazard ratio of 1.5 (i.e. an improvement of 2.6% disease free survival from 92% at four years).



Sample size for time to an event outcome

- Assume independent and exponential life times with hazard rates λ_c and λ_e for control and experimental groups respectively
- $H_0: S_e(t) = S_c(t)$ vs $H_a: S_e(t) \neq S_c(t)$
- Since exponential times have constant hazard rates, the above hypotheses can be written as hypotheses for the hazards ratio of $\Delta = \lambda_c / \lambda_e$.



Number of events (d) required

- Assume all the patients will have an event at the time of final analysis. We can determine number of events (per group) required:

$$H_0 : \Delta = 1 \text{ vs } H_a : \Delta = \frac{\lambda_c}{\lambda_e} \neq 1$$

$$d = \frac{2(z_{\alpha/2} + z_{1-\beta})^2}{(\ln \Delta)^2}$$

- Since there will be patients censored at the time of final analysis, we have to enter more patients and follow them for some time in order to observe the given number of events



Total Size & Duration

- Patients are recruited over an interval 0 to T_0 and then follow to the end of the study period T
- The required sample size for the study is N :


$$N = \frac{(Z_{\alpha/2} + Z_{\beta})^2 [\phi(\lambda_c Q_c)^{-1} + \phi(\lambda_e Q_e)^{-1}]}{(\lambda_c - \lambda_e)^2}$$

$$\text{where } Q_c = \frac{n_c}{N}, \quad Q_e = \frac{n_e}{N}$$

$$\phi(\lambda) = \frac{\lambda^2}{1 - [e^{-\lambda(T-T_0)} - e^{-\lambda T}] / \lambda T_0}$$



Help is at hand!



Survival Program to Calculate Power or Sample Size

User input Program output Not applicable

Select type of calculation Select type of input

Power Sample Size Hazard rates Survival prop

strata 1 Change types or strata

test type

1-sided 2-sided

years of accrual years follow-up prop. in std. group 0.5 alpha 0.05

accrual rate 0 Total accrual 0 power 0

stratum	proportion	hazard rate, standard	hazard ratio	hazard rate, expt	proportion surviving @	survival time
1	1					

Calculate

[Help Document](#) [Help Tips](#) [Return to Tools menu](#)

<http://www.swogstat.org/statoolsout.html>



Example

- $H_0: S_e(t) = S_c(t)$ vs $H_a: S_e(t) \neq S_c(t)$
- M_e and M_c are median survivals of the experimental and control arms respectively

M_e	M_c	Δ	$D=2*d$	$T = T_0(N)$
$\alpha=0.05, 1-\beta=0.8$				
1.5	1	1.5	191	3.0 (300)
2.0	1	2.0	65	1.5 (143)
2.5	2	1.25	631	8.0 (910)
3.0	2	1.5	191	4.0 (370)
4.0	2	2.0	65	2.5 (160)



Practicum & Examples

Dr CJ O'Callaghan

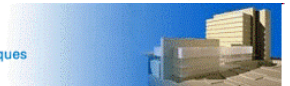
NCIC Clinical Trials Group
NCIC Groupe des essais cliniques



A PHASE III RANDOMIZED STUDY OF YTTRIUM-90 GLASS MICROSPHERES PLUS BEST SUPPORTIVE CARE VERSUS BEST SUPPORTIVE CARE ALONE IN PATIENTS WITH PRETREATED LIVER-DOMINANT METASTATIC COLORECTAL CARCINOMA

- Primary Outcome = Overall Survival
- 1:1 Randomization (2:1?)
- Alpha = 0.05, 2-sided (1-sided?)
- Power = 90% (80%?)
- Accrual Rate = 100 patients per year (150?)
- Duration of Follow-up = 12 months (6 months?)
- Median Survival in Control Arm = 4.6 months (6 months?)
- Hazard Ratio to Detect = 1.43 (1.36?)
 - (6.6 months to 4.6 months ... or ... 8.6 months to 6 months)
 - (6.3 months to 4.6 months ... or ... 7.7 months to 6 months)

Sample Size?



- Primary Outcome = OS
- 1:1 Randomization
- Alpha = 0.05, 2-sided
- Power = 90%
- Accrual Rate = 100 / year
- Duration of Follow-up = 12 months
- Median BSC Survival = 4.6 months
- Hazard Ratio to Detect = 1.43

= 356

- Accrued over 3.6 years
- Total duration = 4.6 years

- Primary Outcome = OS
- 2:1 Randomization ↑
- Alpha = 0.05, 1-sided ↓
- Power = 80% ↓
- Accrual Rate = 150 / year ↓
- Duration of Follow-up = 6 months ↑
- Median BSC Survival = 6 months ↓
- Hazard Ratio to Detect = 1.36 ↑

= 350

- Accrued over 2.3 years
- Total duration = 2.8 years



© Mike Baldwin / Cornered

Baldwin



“Do a double-blind test. Give the new drug to rich patients and a placebo to the poor. No sense getting their hopes up. They couldn’t afford it even if it works.”

