Canadian Cancer Trials Group | Groupe canadien des essais sur le cancer

2024 NCITC Workshop:

Sample Size Determination- Methodology and Philosophy

Chris O'Callaghan *DVM MSc PhD*

Learning Objectives

- Identify the key statistical components that drive sample sizes
- Discuss practical limitations and how to incorporate them in study design
- Discuss the ‘sample size tango’ for creating a successful sample size calculation

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-β)

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

Experimental Errors

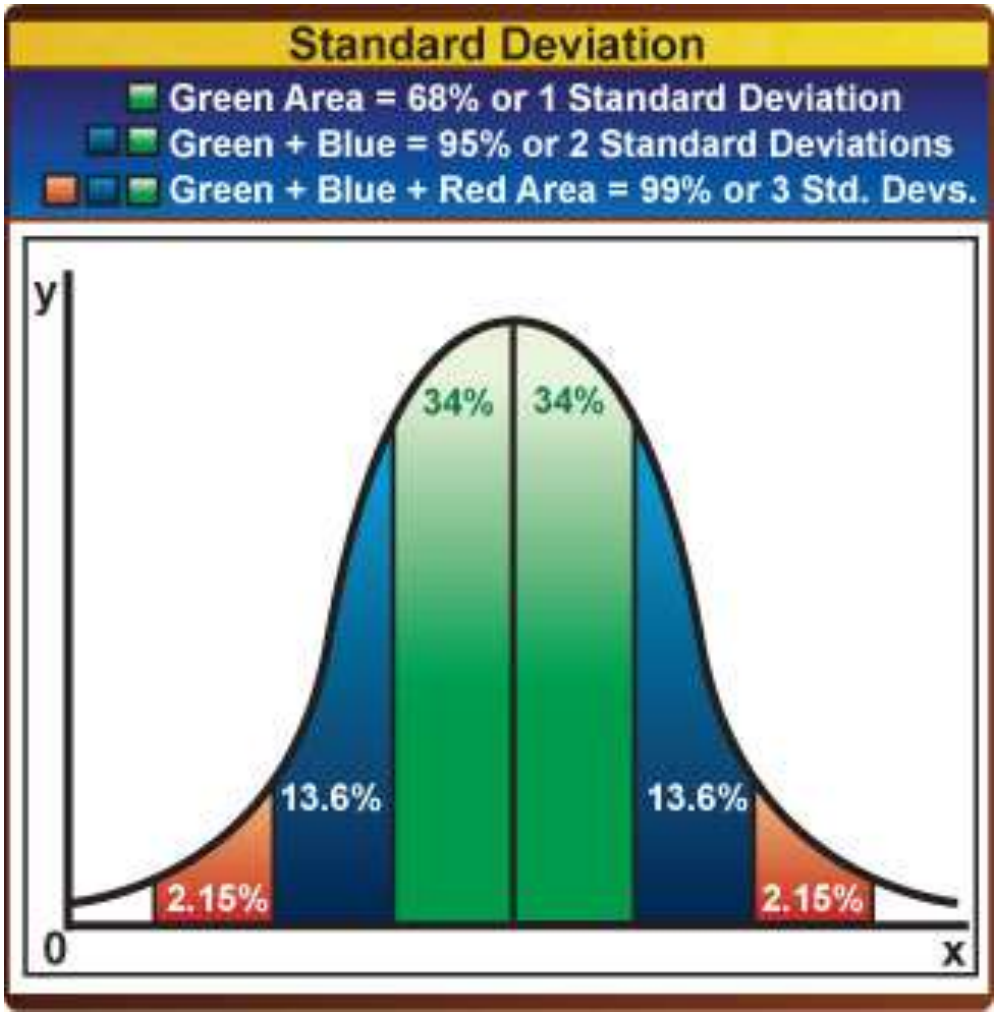
Sample

Results of Statistical Analysis

Results of Statistical Analysis	State of Nature (Reality) **UNKNOWN**	
	No Effect	Effect
No Effect	✓ 'Accept' null hypothesis when it is true	Type II (β) error 'Accept' null hypothesis when it is false
Effect	Type I (α, p) error Reject null hypothesis when it is true	✓ Reject null hypothesis when it is false

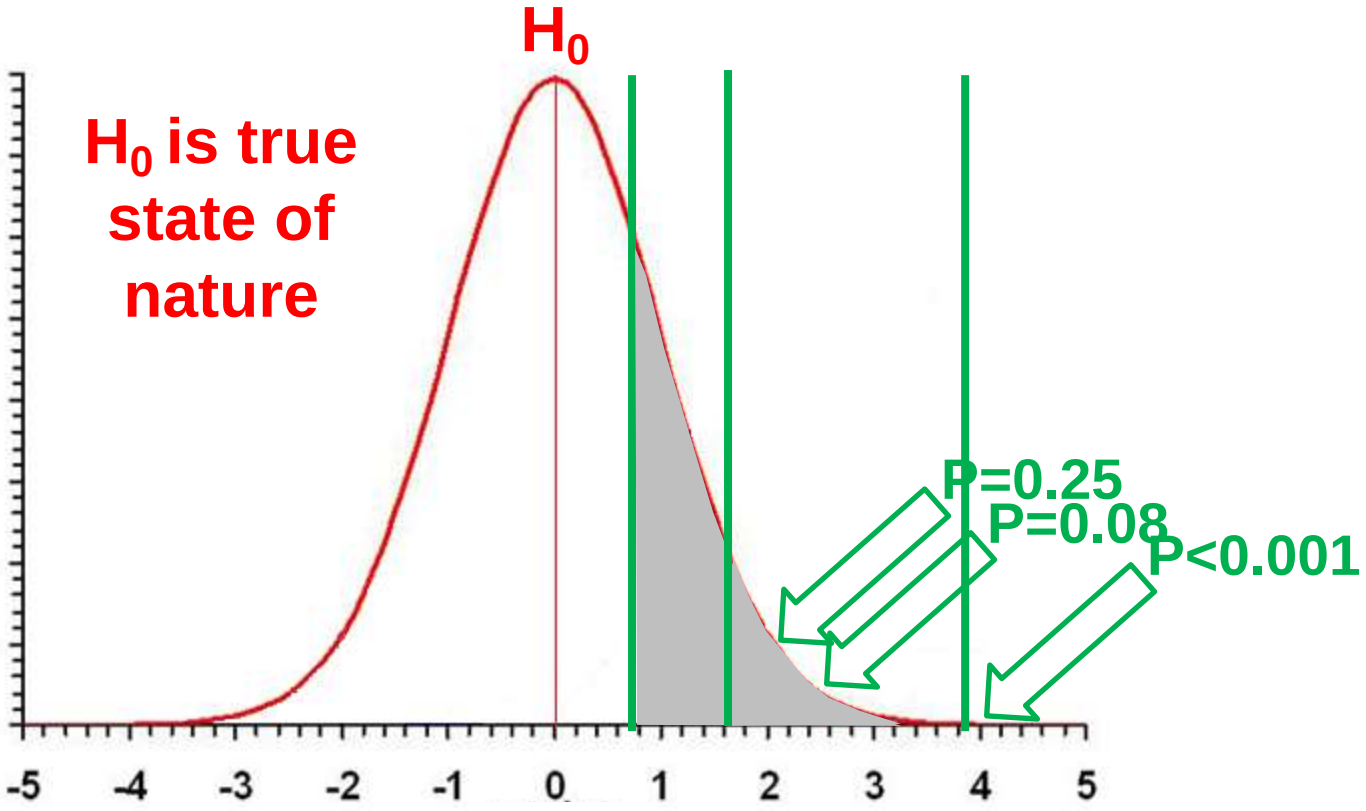
Aside: Sampling Distribution

- 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"*

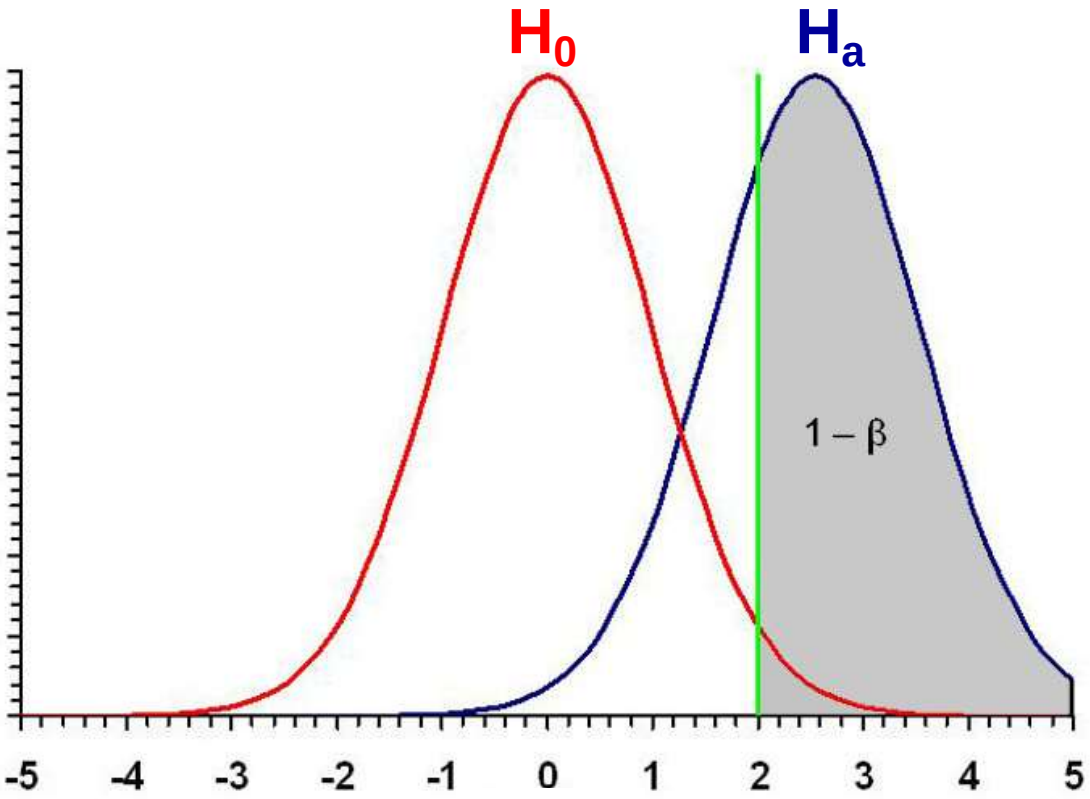


Significance Level

- 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.

Power (1-β)

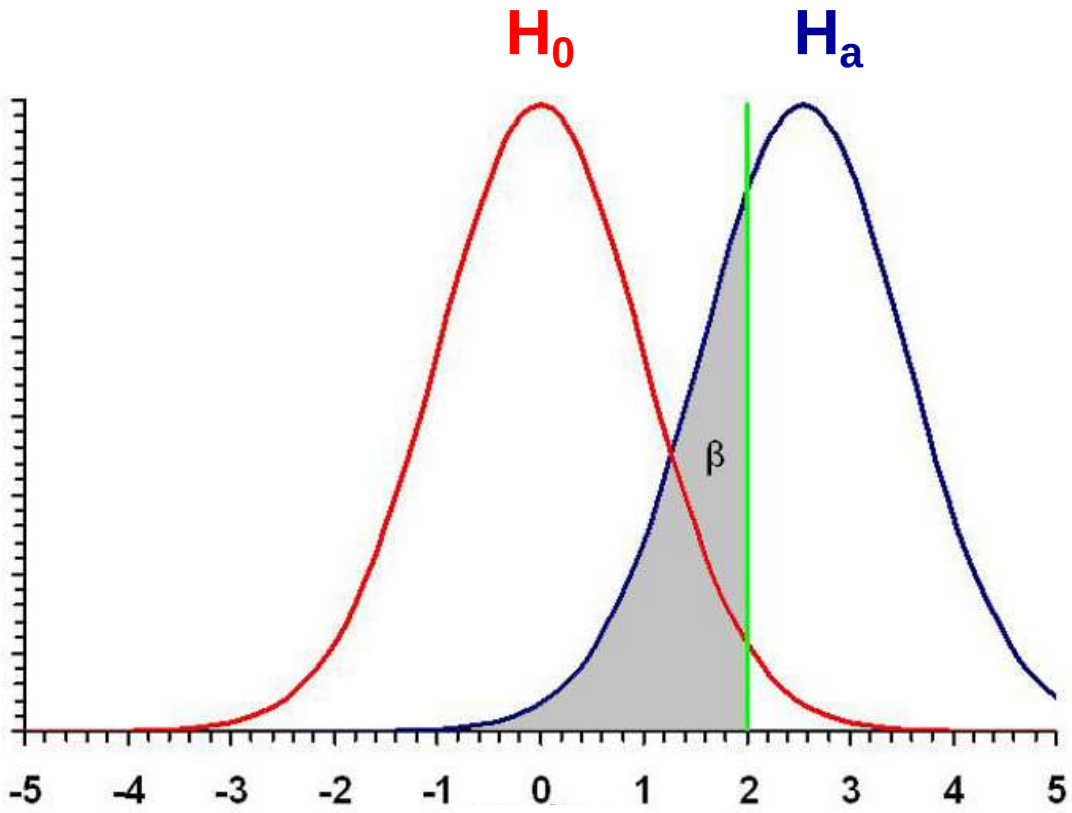
- Probability of correctly reject H_0 (probability of rejecting the H_0 given that H_a is true)
- Power=1-type II error



H_a is true state of nature

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, Type II error (β)

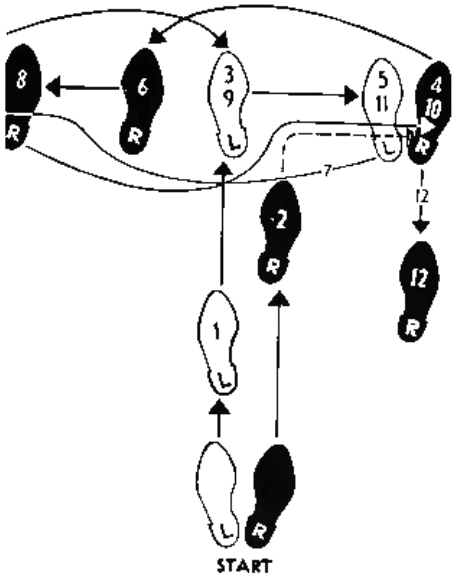
- Traditionally, power is fixed *a priori*, usually at 0.80 (1-β) with the chance of a Type II error (β) at 0.20
- Few studies are powered greater than 90% but **MANY** have lower power
- Affects the credibility of "negative" studies
  - Medical *versus* Ecological implications
- Be suspicious of small studies and/or those where apriori power is not explicitly reported.

“The Tango”

α= 5%
Type I error
Statistical Significance

Sample Size

N



-

-

Δ

β

-

Difference / Efficacy

Type II error
1- Power

Canadian Cancer Trials Group
Groupe canadien des essais sur le cancer

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, what difference in that measure the trial will be designed to detect, and how this can be done in a timely fashion

Reference

Practical help for specifying the target difference in sample size calculations for RCTs: the DELTA five-stage study, including a workshop

JA Cook et al, Health Technology Assessment, 23(60): October 2019

BOX 3 The DELTA² recommendations for undertaking a sample size calculation and choosing the target difference

The following are recommendations for specifying the target difference in a RCT's sample size calculation when the conventional approach to the sample size calculation is used. Recommendations on the use (or not) of individual methods are made. More detailed advice on the application of the individual methods can be found elsewhere.[15]

Recommendations

- Begin by searching for relevant literature to inform the specification of the target difference. Relevant literature can:
  - relate to a candidate primary outcome and/or the comparison of interest
  - inform what is an important and/or realistic difference for that outcome, comparison and population (estimand of interest).
- Candidate primary outcomes should be considered in turn and the corresponding sample size explored. When multiple candidate outcomes are considered, the choice of primary outcome and target difference should be based on consideration of the views of relevant stakeholders groups (e.g. patients), as well as the practicality of undertaking such a study and the required sample size. The choice should not be based solely on which yields the minimum sample size. Ideally, the final sample size will be sufficient for all key outcomes, although this is not always practical.
- The importance of observing a particular magnitude of a difference in an outcome, with the exception of mortality and other serious adverse events, cannot be presumed to be self-evident. Therefore, the target difference for all other outcomes requires additional justification to infer importance to a stakeholder group.
- The target difference for a definitive (e.g. Phase III) trial should be one considered to be important to at least one key stakeholder group.
- The target difference does not necessarily have to be the minimum value that would be considered important if a larger difference is considered a realistic possibility or would be necessary to alter practice.
- When additional research is needed to inform what would be an important difference, the anchor and opinion-seeking methods are to be favoured. The distribution should not be used. Specifying the target difference based solely on a SES approach should be considered a last resort, although it may be helpful as a secondary approach.
- When additional research is needed to inform what would be a realistic difference, the opinion-seeking and review of the evidence-based methods are recommended. Pilot studies are typically too small to inform what would be a realistic difference and primarily address other aspects of trial design and conduct.
- Use existing studies to inform the value of key 'nuisance' parameters that are part of the sample size calculation. For example, a pilot trial can be used to inform the choice of SD value for a continuous outcome or the control group proportion for a binary outcome, along with other relevant inputs, such as the number of missing outcome data.
- Sensitivity analyses that consider the impact of uncertainty around key inputs (e.g. the target difference and the control group proportion for a binary outcome) used in the sample size calculation should be carried out.
- Specification of the sample size calculation, including the target difference, should be reported in accordance with the recommendations for reporting items (see *Chapter 4, Figure 1*) when preparing key trial documents (grant applications, protocols and result manuscripts).

SD, standard deviation; SES, standardised effect size.

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 required:

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

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

Statistical Significance

Power

Difference/Effect

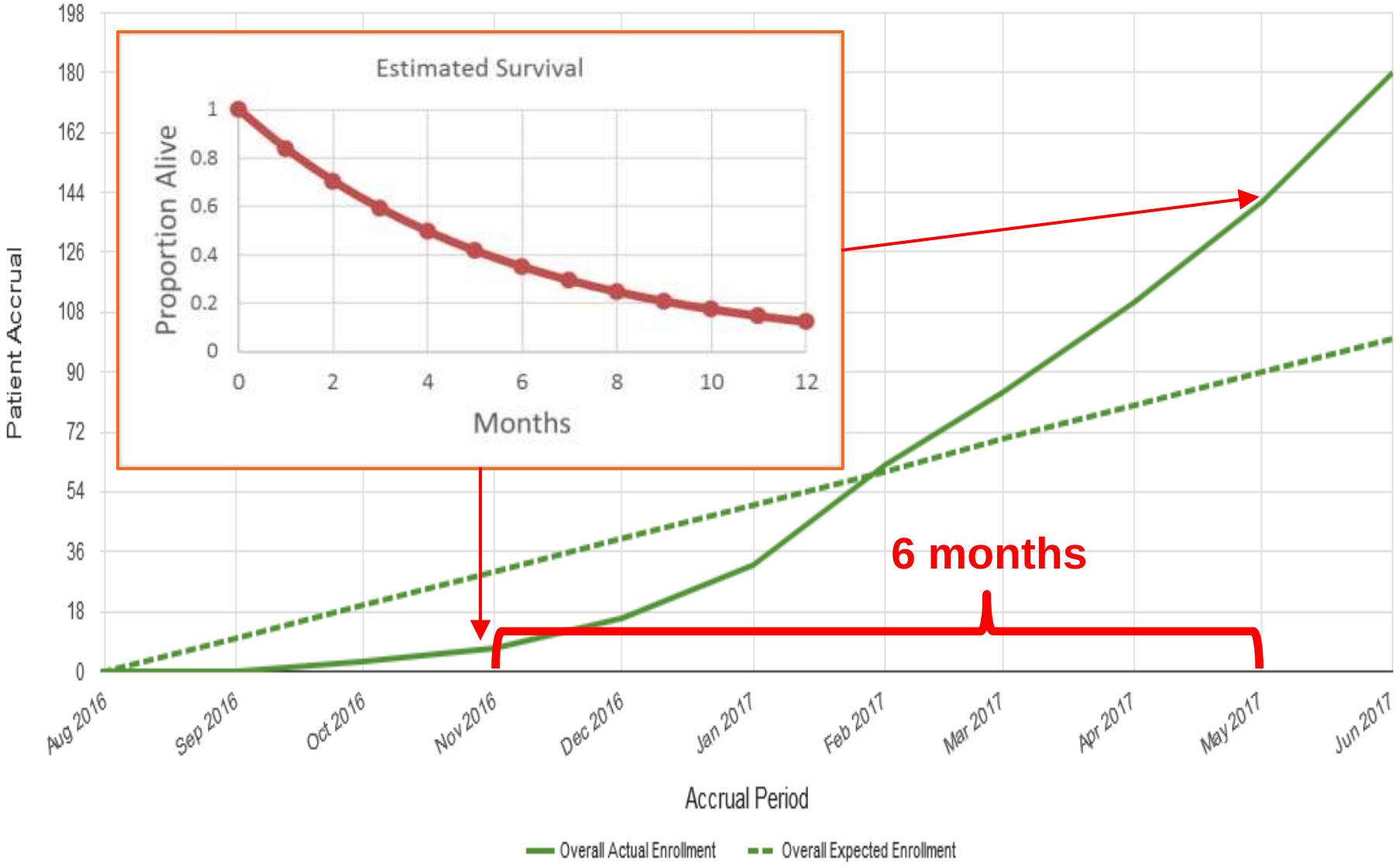
Example – Number of Events

- $H_0: S_e(t) = S_c(t)$ 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 (HR)	# Events
α=0.05, 1-β=0.8			
1.5	1	1.5	191
2.0	1	2.0	65
1.25	1	1.25	631
3.0	2	1.5	191
4.0	2	2.0	65

- 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

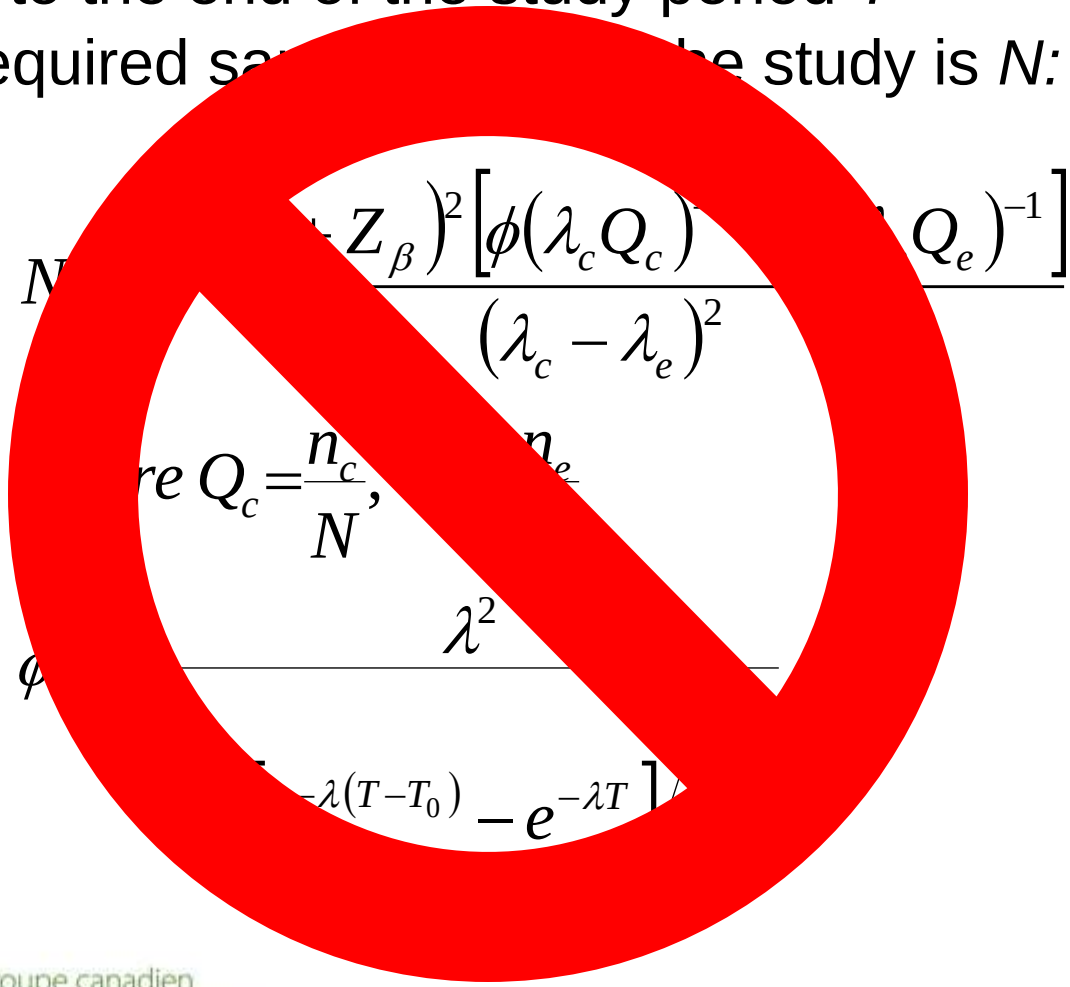
Example: CO.26



Canadian Cancer Trials Group | Groupe canadien des essais sur le cancer

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 sa… …e study is *N:*



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

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

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

Help is at hand!

CRAB Cancer Research And Biostatistics

SWOG Cancer Research Network

STATISTICAL TOOLS | DESIGN | ANALYSIS | PROBABILITIES | OTHER TOOLS | ABOUT US

Two Arm Survival

Two Arm Survival is a program to calculate either estimates accrual or power for differences in survival times between two groups. The program allows for unequal sample size allocation between the two groups. The survival time estimates also allow for multiple strata or risk groups.

For further details, view the Help Document.

User Input	Program Output

Select Parameters

Type Calculation	Type Input	Sided
◉ Sample Size ○ Power	○ Hazard Ratio ○ Survival Proportions ◉ Medians	○ 1 Sided ◉ 2 Sided

Number Strata	Proportion in Standard Group	Alpha
1	0.5	0.05
Years of Accrual	**Years of Follow-up**	**Power**

Stratum	Proportion	Hazard Rate, Std.	Hazard Rate, Exp.	Hazard Ratio	Proportion Surviving, Std.	Survival Time, Std.	Proportion Surviving, Exp.	Survival Time, Exp.
1	1				0.5		0.5	

Accrual Rate	Total Accrual

Calculate

Help Document

Canadian Cancer Trials Group | Groupe canadien des essais sur le cancer

Two Arm Survival (crab.org)

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 = Survival
- 1:1 Randomization
- Alpha = 0.05, 2-sided
- Power = 90%
- Median Survival Control = 6 months
- Hazard Ratio to Detect = 1.25 (0.80)
- 6 months – 7.5 months
- 845 events required
- Accrual Rate = 100 / year
- Duration of Follow-up = 6 months

= 890

- Accrued over ~ 9 years
- Total duration ~ 9.5 years

- Primary Outcome = Survival
- 1:1 Randomization
- Alpha = 0.05, **1-sided ↓**
- Power = **80% ↓**
- Median Survival Control = 6 months
- Hazard Ratio to Detect = **1.50(0.67)↑**
- 6 months – **9 months ↑**
- 151 events required
- Accrual Rate = 100 / year
- Duration of Follow-up = **18 months ↑**

= 166

- Accrued over ~ 1.67 years
- Total duration ~ 3.33 years

Canadian Cancer Trials Group / Groupe canadien des essais sur le cancer

Another Example of "the Tango"…

- Adjuvant trial in resected biliary cancer evaluating capecitabine vs capecitabine + gemcitabine
- Primary endpoint Relapse-Free Survival (RFS)
- 1:1 randomization
- Alpha = 5%, 2-sided (Type I error)
- Power = 80% (Type II error = 20%)
- Median RFS with capecitabine = 24 months
- Hazard Ratio = 1.4 (/0.714 or 28.6% reduction in risk of relapse)
  - Median RFS with combination = 33.6 months
  - Absolute improvement in median of 9.6 months

278 "Events" Required

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

How do we get 278 events?

Need to know accrual RATE!

Accrue at 18 patients per month (~216 per year):

a) Accrue for 2 years to enroll 422 patients then follow for an additional 2.75 years = Total Duration of 4.75 years (66%)

b) Accrue for 1.5 years to enroll 320 patients then follow for an additional 6.25 years = Total Duration of 7.75 years (87%*)

c) Accrue for 3 years to enroll 659 patients then follow for an additional 0.5 years = Total Duration of 3.5 years (42%)

* CAUTION – Assumes constant risk and therefore exponential distribution

Fill in the Blanks!

CRAB Cancer Research And Biostatistics

SWOG Cancer Research Network

STATISTICAL TOOLS | DESIGN | ANALYSIS | PROBABILITIES | OTHER TOOLS | ABOUT US

Two Arm Survival

Two Arm Survival is a program to calculate either estimates accrual or power for differences in survival times between two groups. The program allows for unequal sample size allocation between the two groups. The survival time estimates also allow for multiple strata or risk groups.

For further details, view the Help Document.

User Input | Program Output

Select Parameters

Type Calculation	Type Input	Sided
(•) Sample Size ( ) Power	( ) Hazard Ratio ( ) Survival Proportions (•) Medians	( ) 1 Sided (•) 2 Sided

Number Strata	Proportion in Standard Group	Alpha
1	0.5	0.05
Years of Accrual	**Years of Follow-up**	**Power**
2	**2.75**	**0.80**

Stratum	Proportion	Hazard Rate, Std.	Hazard Rate, Exp.	Hazard Ratio	Proportion Surviving, Std.	Survival Time, Std.	Proportion Surviving, Exp.	Survival Time, Exp.
1	1	0.347	0.248	1.4	0.5	**2**	0.5	**2.8**

Accrual Rate	Total Accrual
211.48	422

Calculate

Help Document

Canadian Cancer Trials Group | Groupe canadien des essais sur le cancer

Two Arm Survival (crab.org)

“Too optimistic…”

- Adjuvant trial in resected biliary cancer evaluating capecitabine vs capecitabine + gemcitabine
- Primary endpoint Relapse-Free Survival (RFS)
- 1:1 randomization
- Alpha = 5%, 2-sided (Type I error)
- Power = 80% (Type II error = 20%)
- Median RFS with capecitabine = 24 months
- Hazard Ratio = **1.3** (/0.769 or 23.1% reduction in risk of relapse)
- Median RFS with combination of **31.2** months
- Absolute improvement in median of **7.2** months

457 “Events” Required

29% to 23% risk reduction = 278 to 457 Events

How do we get 457 events?

Accrue at 18 patients per month (~216 per year):

a) Accrue for 3 years to enroll 640 patients then follow for an additional 2.75 years = Total Duration of 5.75 years (71%)

b) Accrue for 2.5 years to enroll 534 patients then follow for an additional 5.25 years = Total Duration of 7.75 years (86%*)

c) Accrue for 4 years to enroll 850 patients then follow for an additional 0.75 years = Total Duration of 4.75 years (54%)

"Too optimistic…"

Accrue at **10 patients per month** (120 per year):

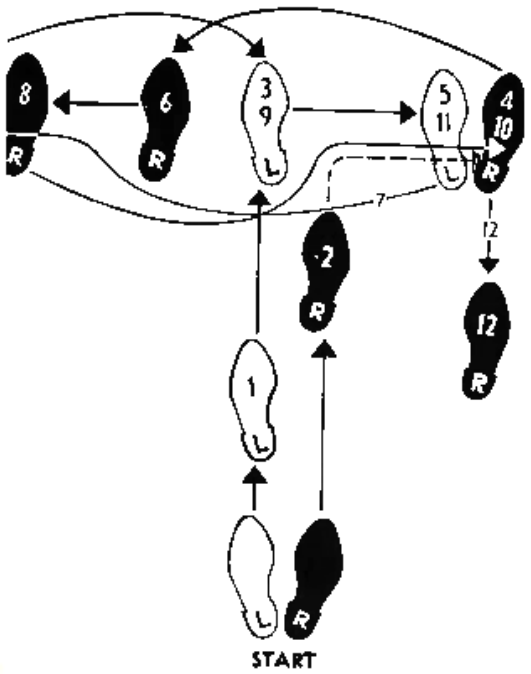
a) Accrue for 4.5 years to enroll 538 patients then follow for an additional 4.25 years = Total Duration of 8.75 years (85%*)

b) Accrue for 4 years to enroll 482 patients then follow for an additional 8 years = Total Duration of 12 years (95%*)

c) Accrue for 6 years to enroll 693 patients then follow for an additional 1 year = Total Duration of 7 years (66%)

d) Accrue for 5 years to enroll 589 patients then follow for an additional 2.75 years = Total Duration of 7.75 years (78%)

The Dance Continues!

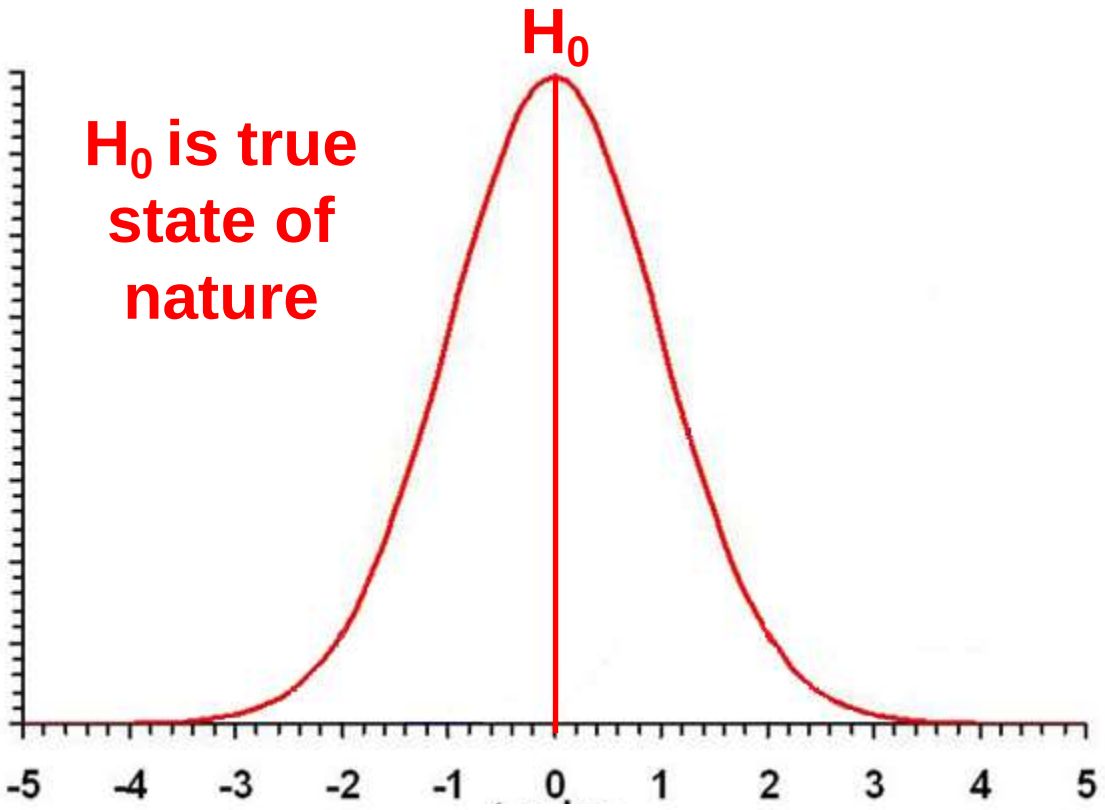


"To call in the statistician after the experiment is done may be no more than asking him to perform a postmortem examination: he may be able to say what the experiment died of."

Sir R.A Fisher

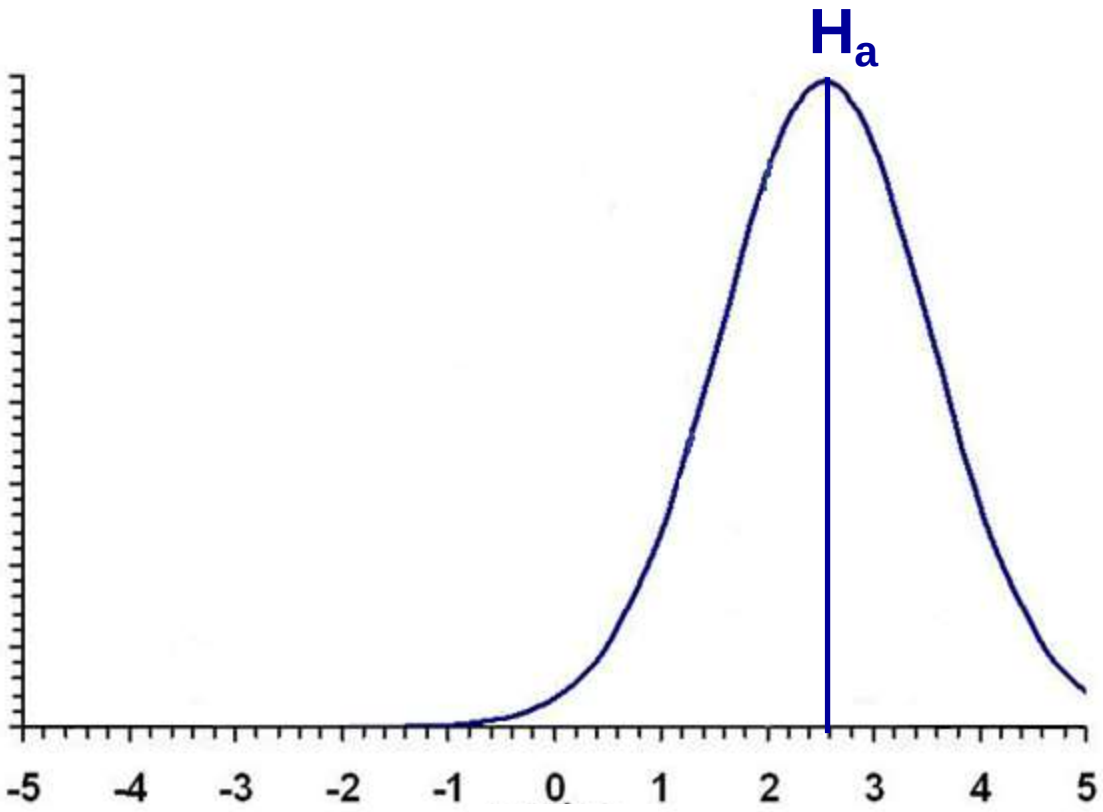
H_0 Sampling Distribution

- Suppose H_0 is true – difference between treatments = “0”
- Repeat trial over and over and over keeping track of results of each in a frequency distribution…



H_a Sampling Distribution

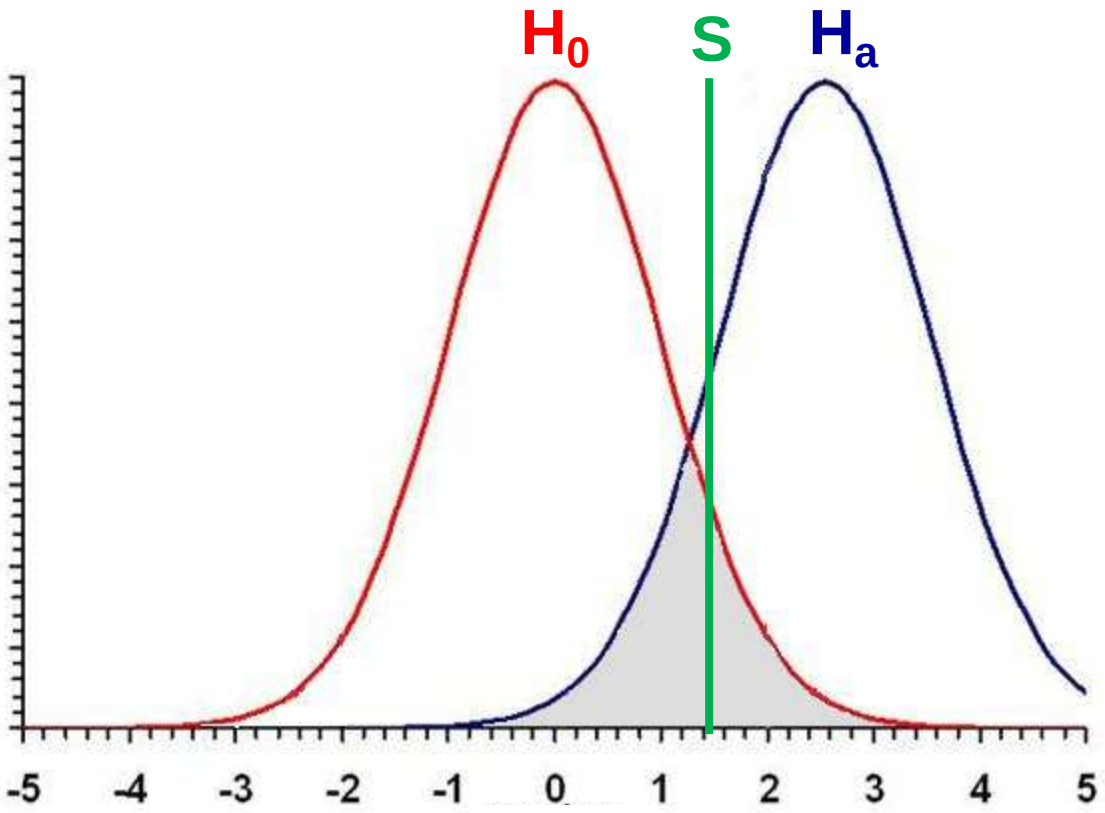
- Suppose H_a is true – difference between treatments = “2.6”
- Repeat trial over and over and over keeping track of results of each in a frequency distribution…



H_a is true state of nature

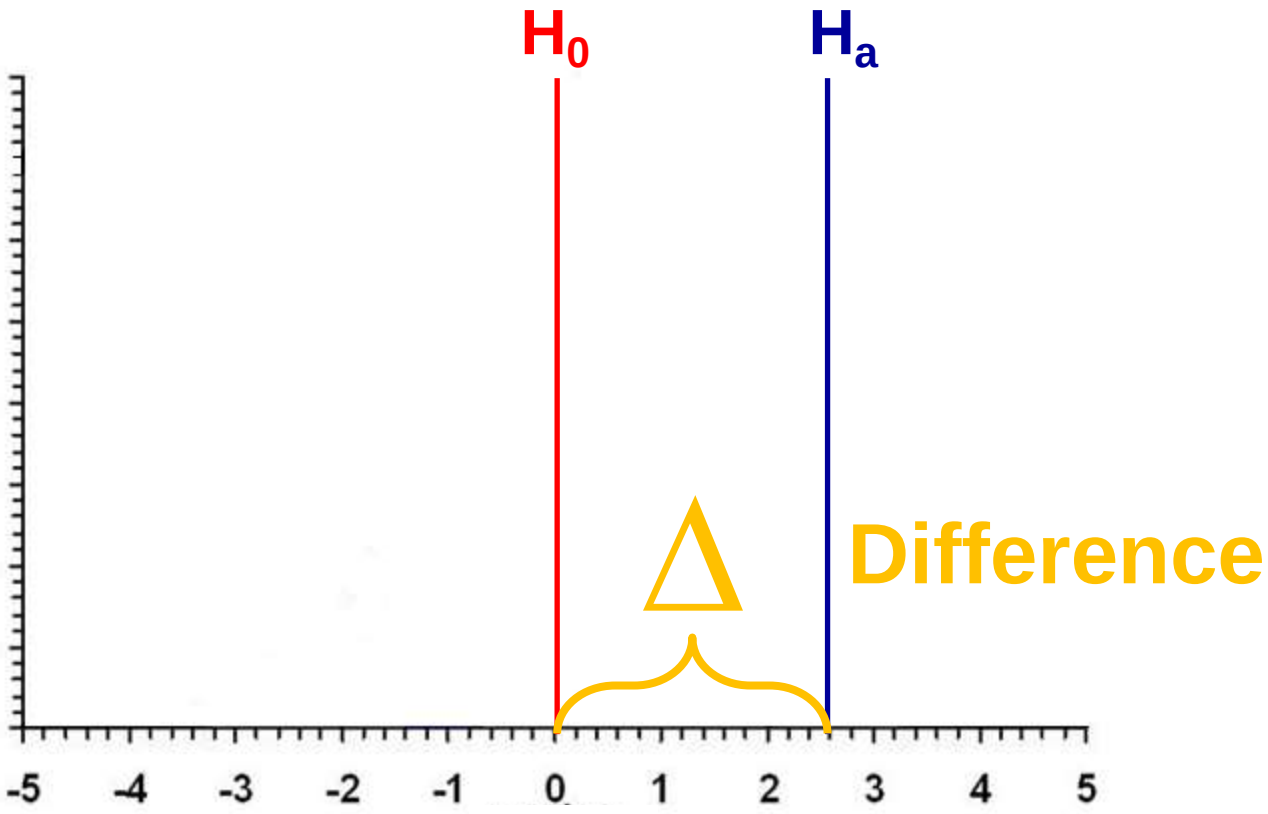
Sampling Distribution Overlap

- One is right…. and one is wrong
- But we only “see” one single result.



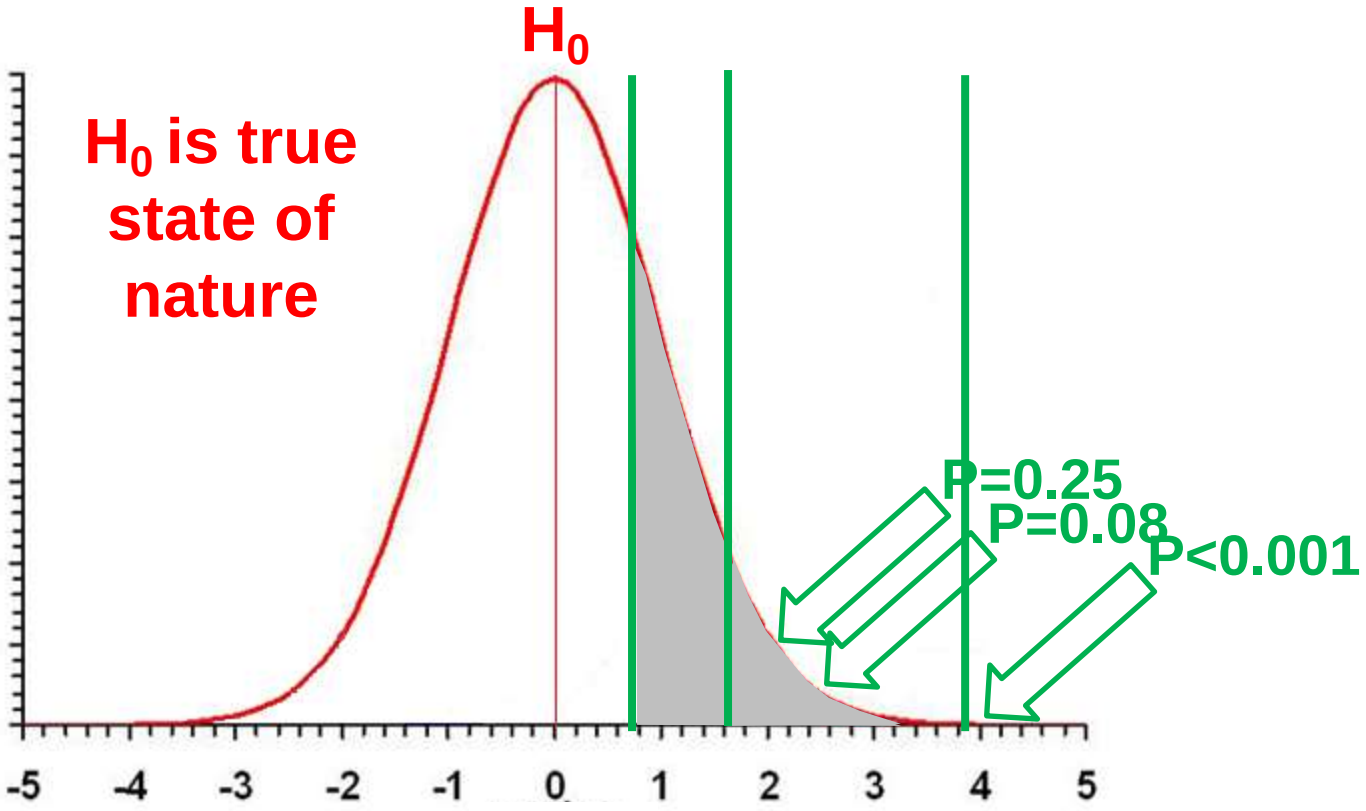
Difference to Detect

- The difference between H_0 and H_a
- … increasing difference will decrease overlap…



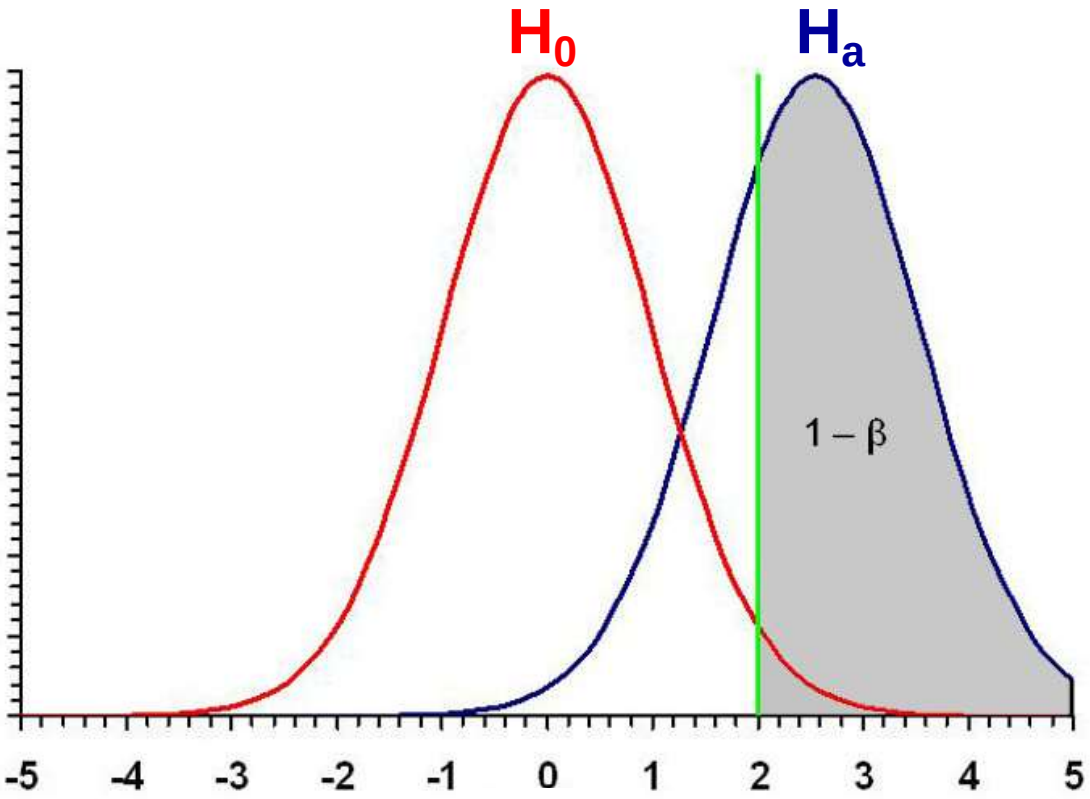
Type I error (α)

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



Power (1-β)

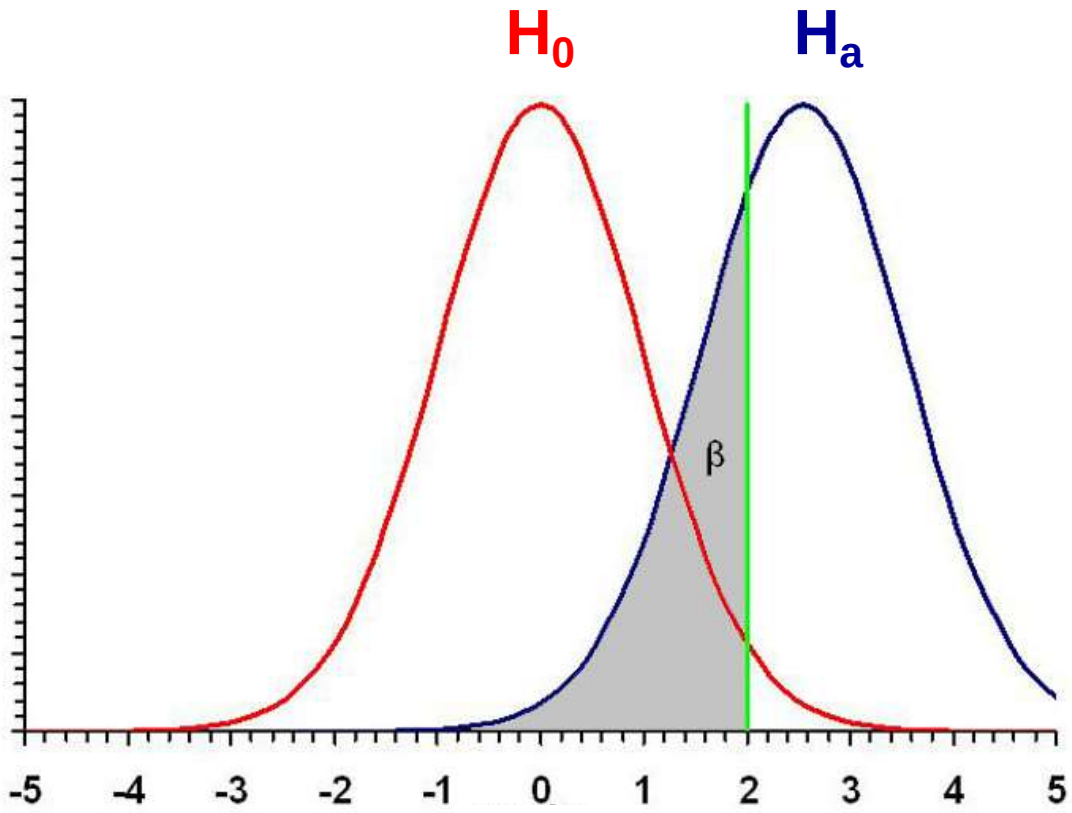
- Probability of correctly reject H_0 (probability of rejecting the H_0 given that H_a is true)
- Power=1-type II error



H_a is true state of nature

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



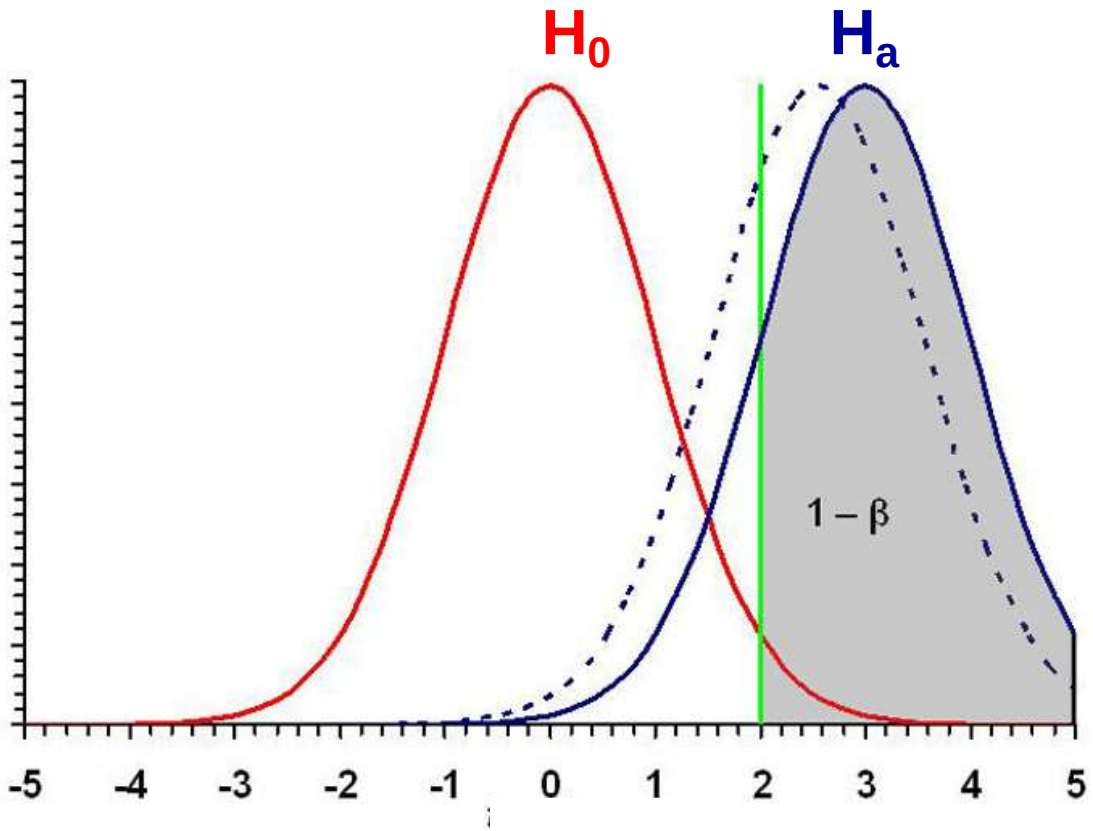
H_a is true state of nature

Power, Type II error (β)

- Traditionally, power is fixed *a priori*, usually at 0.80 (1-β) 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
- Be suspicious of small studies and/or those where apriori power is not explicitly reported.

Power (1-β)

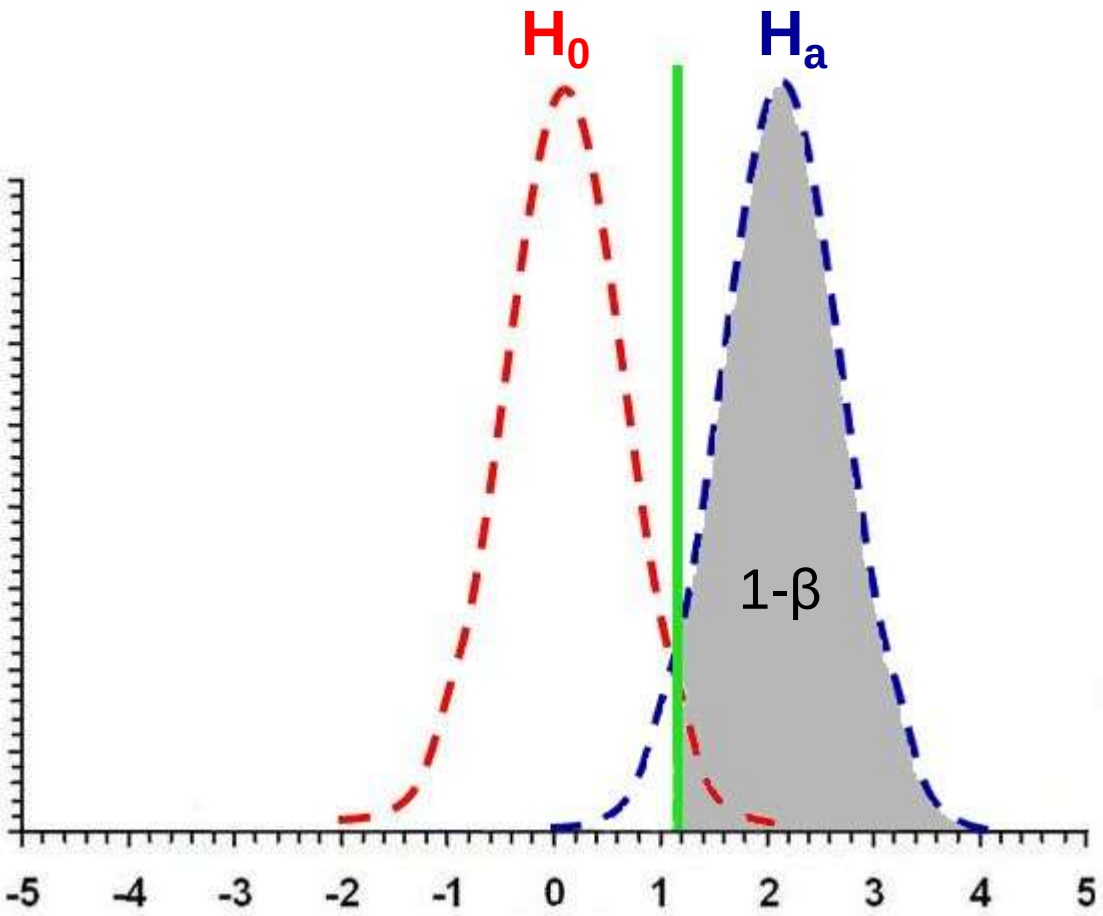
- How to increase power?
- Increase minimum detectable difference– shift H_a and reduce overlap



H_a is true state of nature

Power (1-β)

- How to increase power?
  - Increase N – narrow shape of distributions



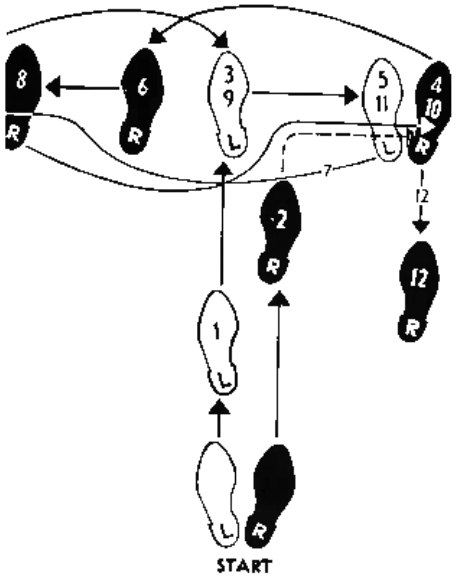
H_a is true state of nature

“The Tango”

α= 5%
Type I error
Statistical Significance

Sample Size

N



-

-

Δ

β

-

Difference / Efficacy

Type II error
1- Power

Canadian Cancer Trials Group
Groupe canadien des essais sur le cancer