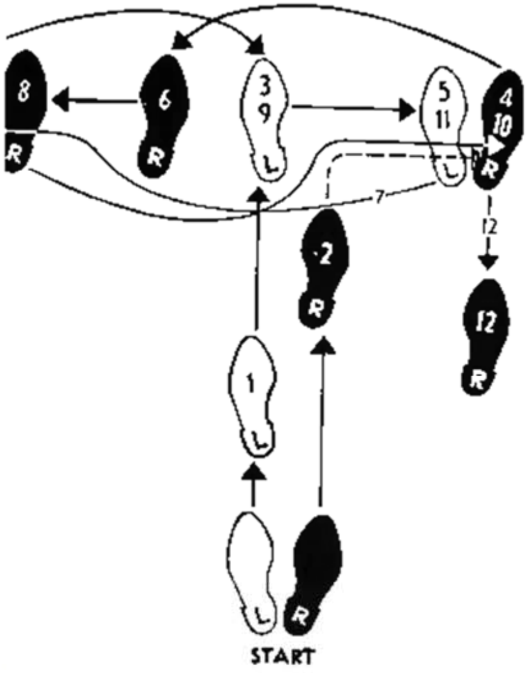
Canadian Cancer Trials Group
Groupe canadien des essais sur le cancer

Workshop 1:

Sample Size Determination-Methodology and Philosophy

C.J. O'Callaghan DVM, MSc, PhD

Population, P-value and Power: Dancing the *“sample-size Tango”* of statistical inference in clinical research



Canadian Cancer
Trials Group

Groupe canadien
des essais sur le cancer

Objectives

- Not a statistics or programming course!
- Enough information to enable you to:
    - understand (± 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.

“How many do I need ?....”

Canadian Cancer Trials Group Groupe canadien des essais sur le cancer

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 ($1-\beta$)

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

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

“Pre-Trial Motions”

- 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…]

Significance Level

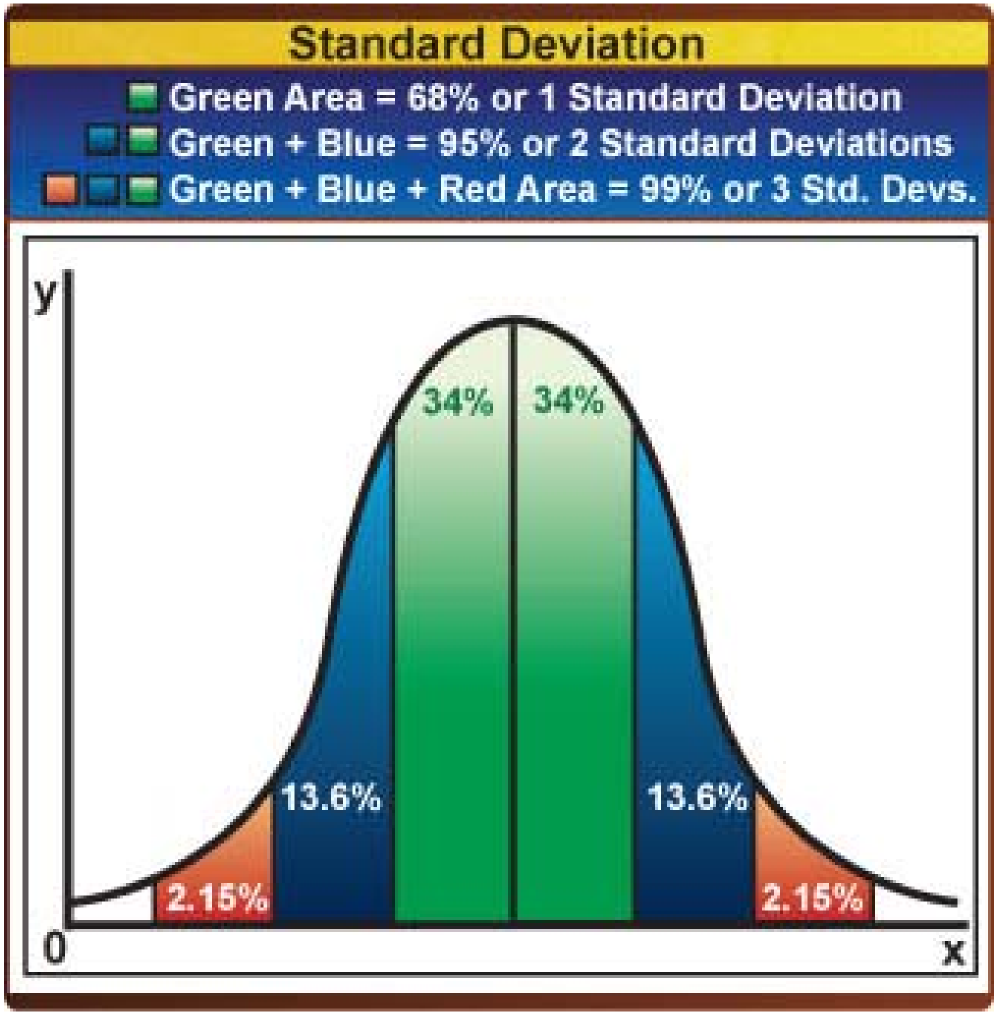
- In hypothesis testing, the significance level is the criterion used for rejecting the null hypothesis.
- The significance level is used in hypothesis testing as follows:
  - The difference between the results of the trial (“the sample”) and H_0 is determined.
  - **Assuming H_0 is true…** the probability (p) 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.

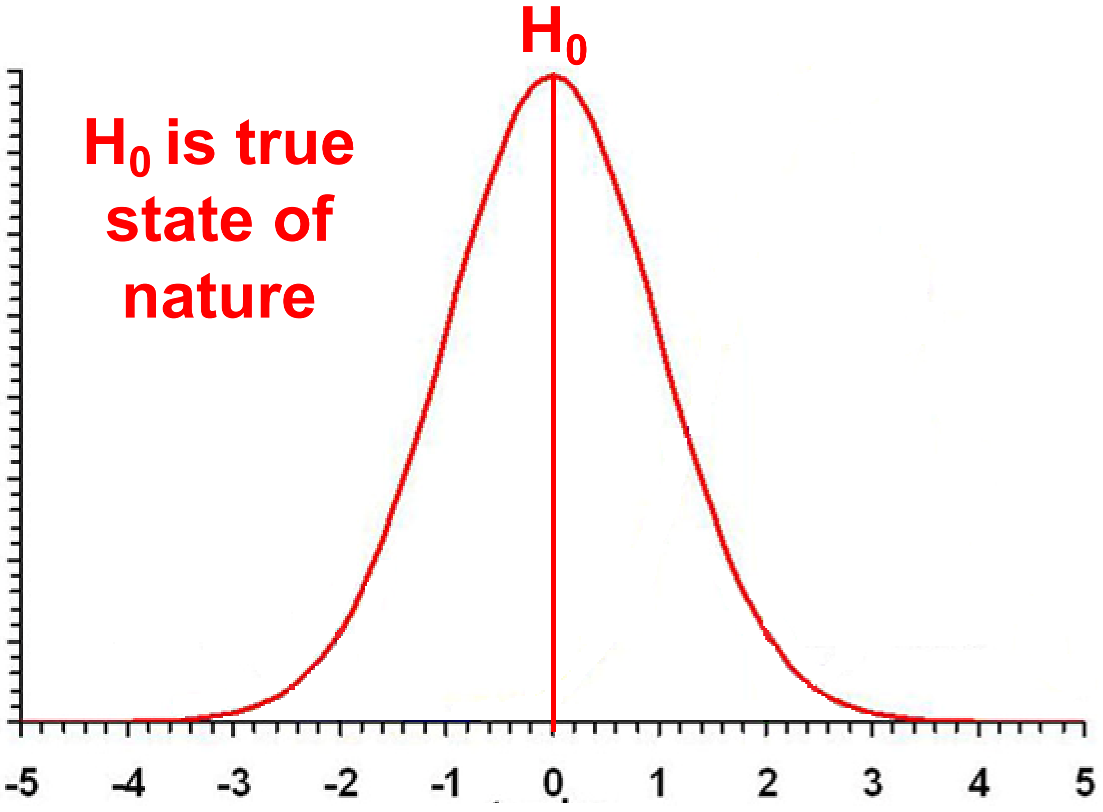
Aside: Sampling Distribution

- 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.



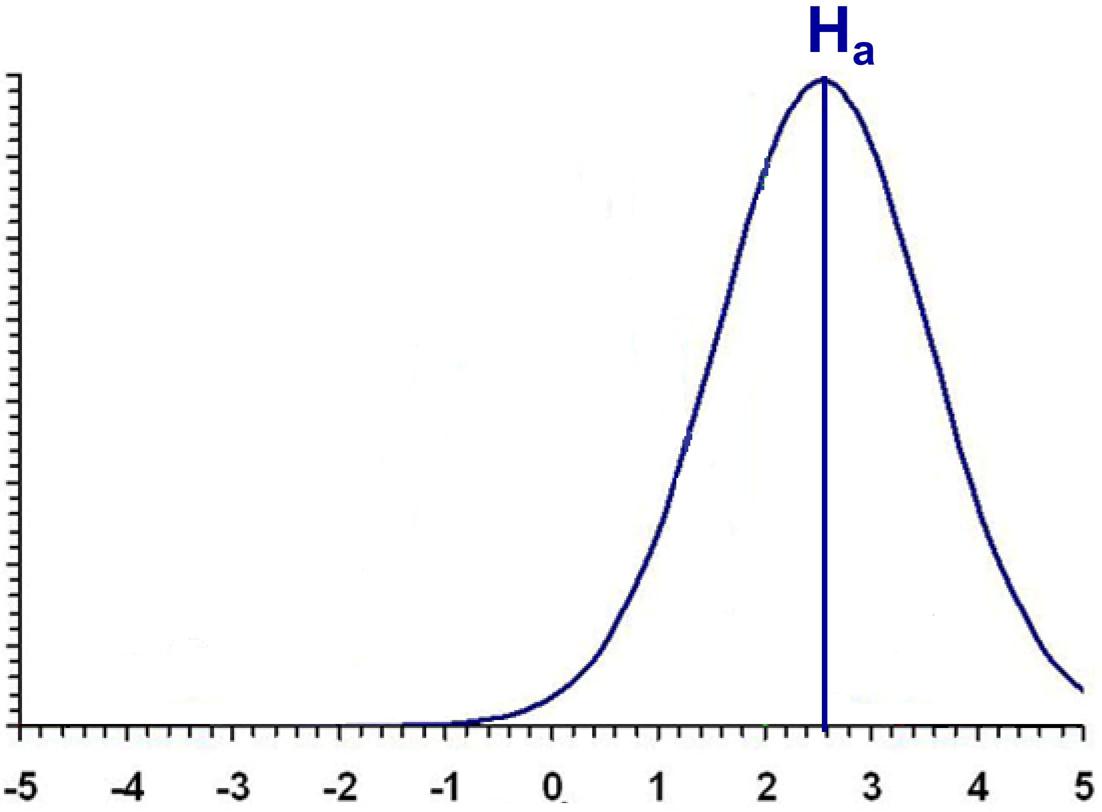
H_0 Sampling Distribution

- 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

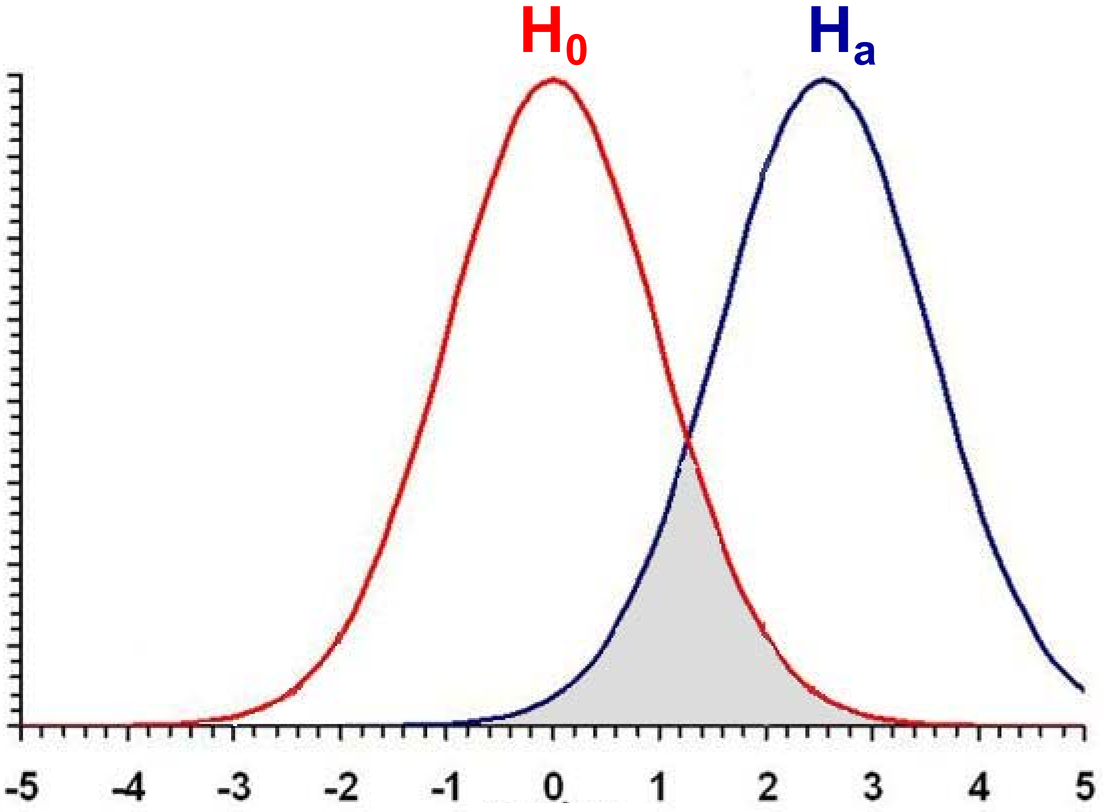
- Suppose H_a is true – difference between treatments = “2.6”
- Repeat trial over and over and over keeping track of results of each in a frequency distribution…



H_a is true state of nature

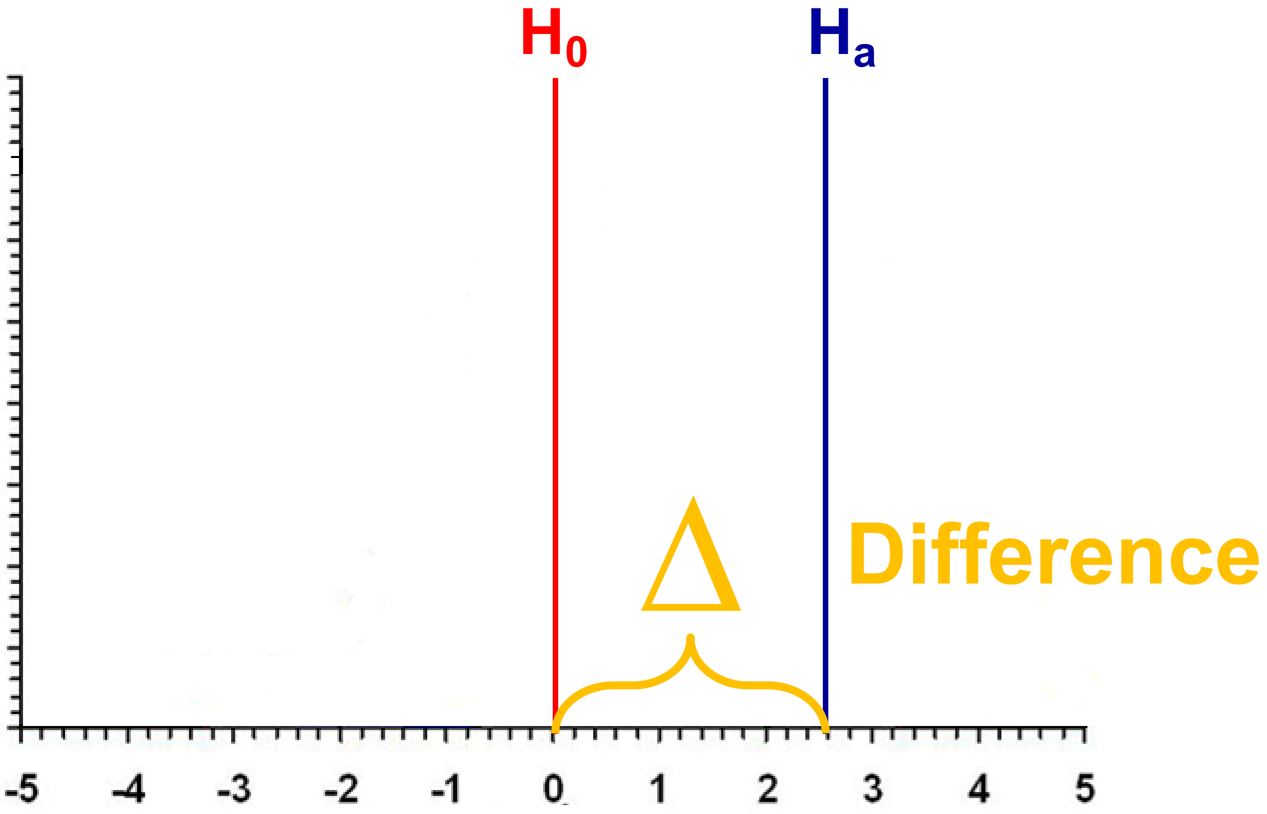
Sampling Distribution Overlap

- 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…

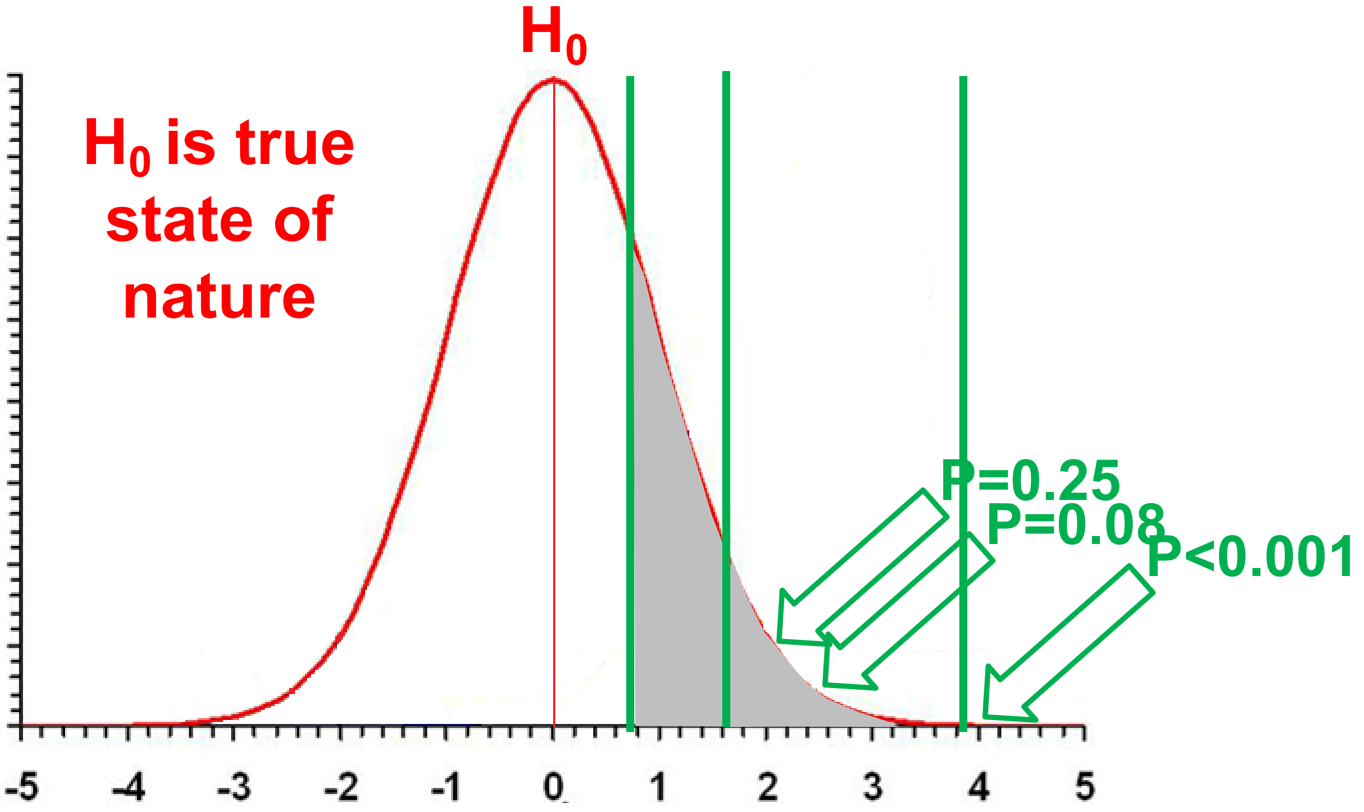


Minimum Difference to be Detected

- 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

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"*



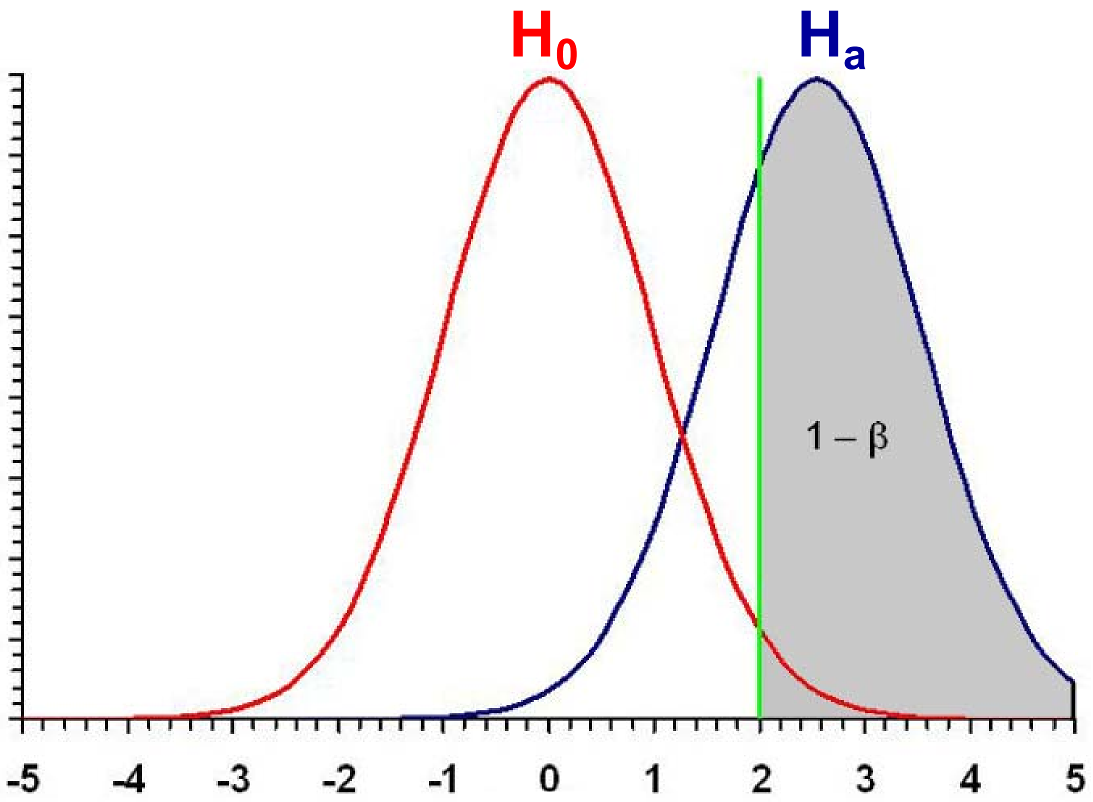
Canadian Cancer Trials Group
Groupe canadien des essais sur le cancer

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/or 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 (1-β)

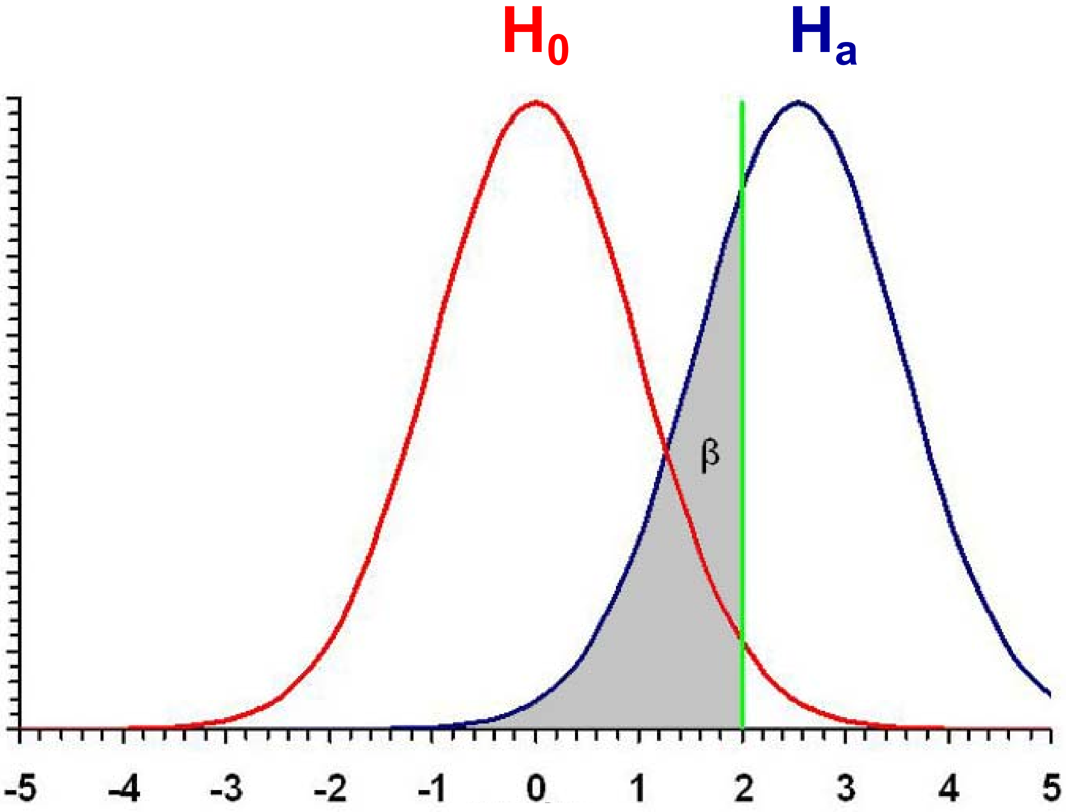
- Probability of correctly reject H_0 (probability of rejecting the H_0 given that H_a is true)
- Power=1-type II error



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



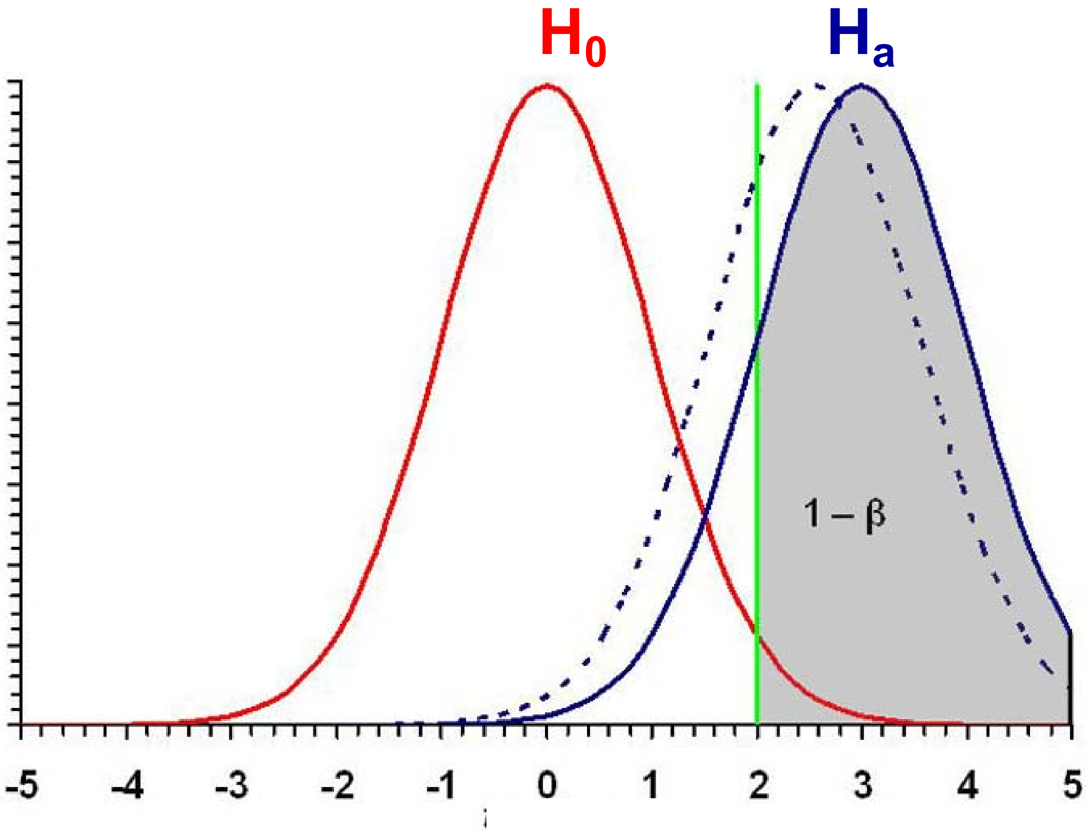
H_a is true state of nature

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)

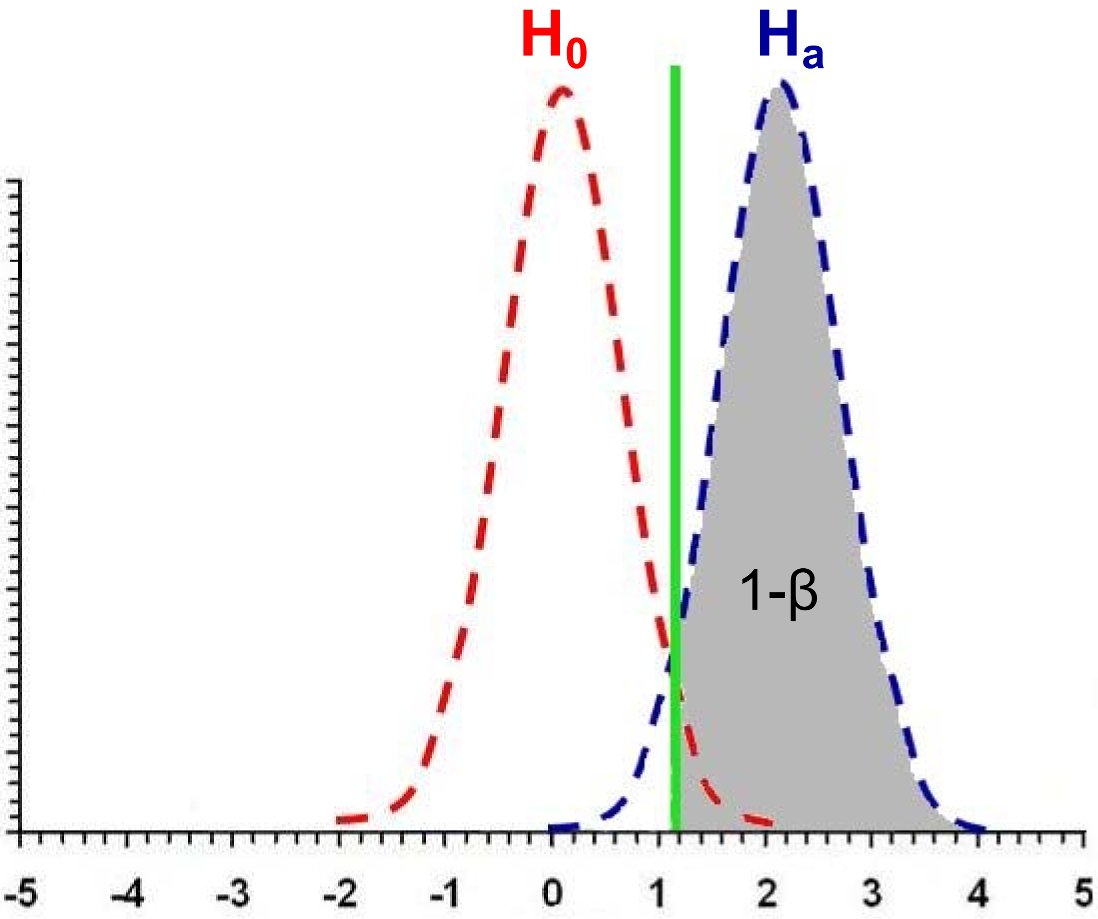
- How to increase power?
- Increase minimum detectable difference– shift H_a and reduce overlap



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

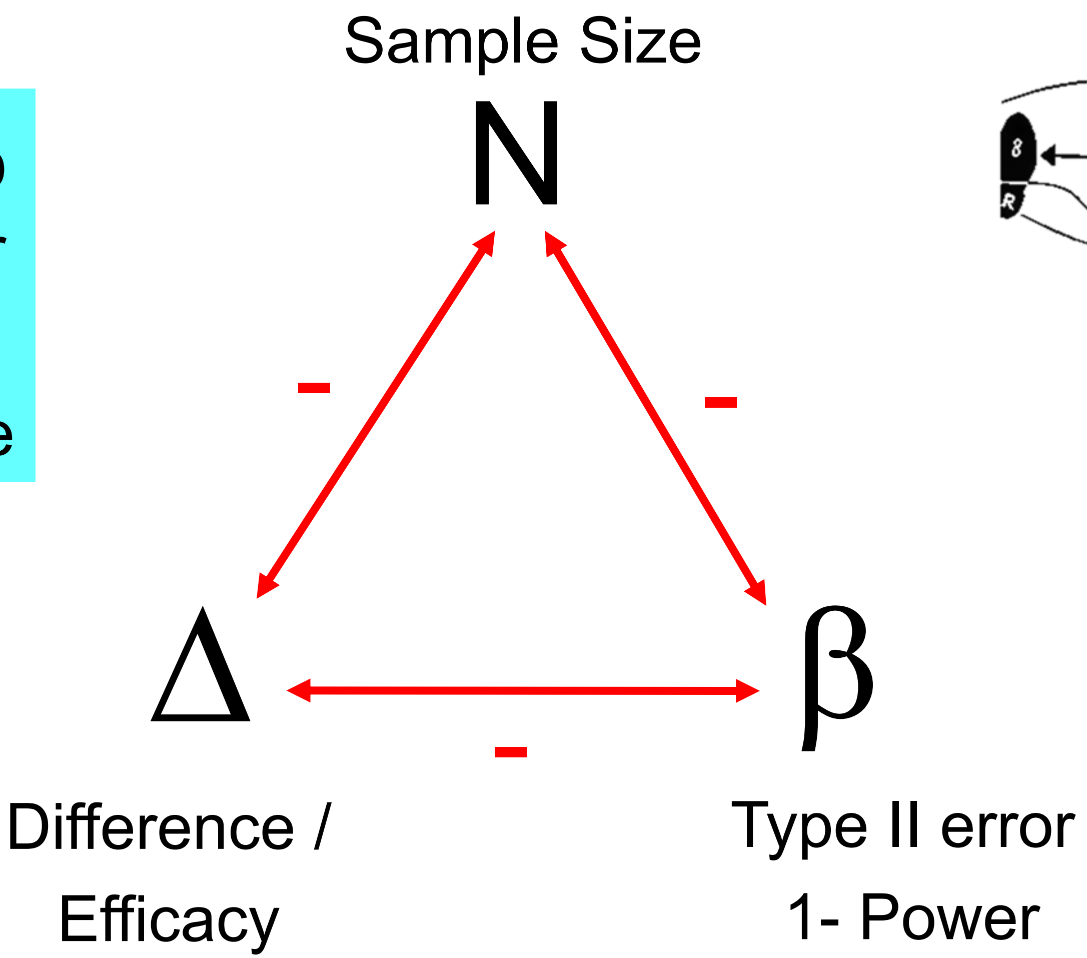
“The Tango”

α= 5%

Type I error

Statistical

Significance



Sample Size

N

-

-

Δ

β

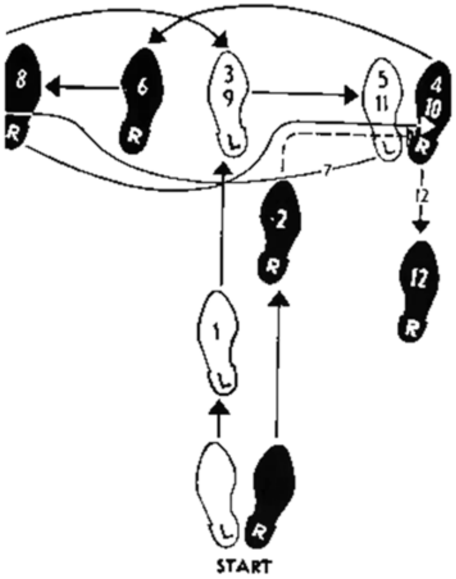
-

Difference /

Efficacy

Type II error

1- Power



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

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:

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

Statistical Significance

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

Power

Difference/Effect

- 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

- 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)***	***D=2*d***
α=0.05, 1-β=0.8			
1.5	1	1.5	191
2.0	1	2.0	65
2.5	2	1.25	631
3.0	2	1.5	191
4.0	2	2.0	65

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 he study is *N:*

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

$$\overline{\qquad (\lambda_c - \lambda_e)^2 \qquad}$$

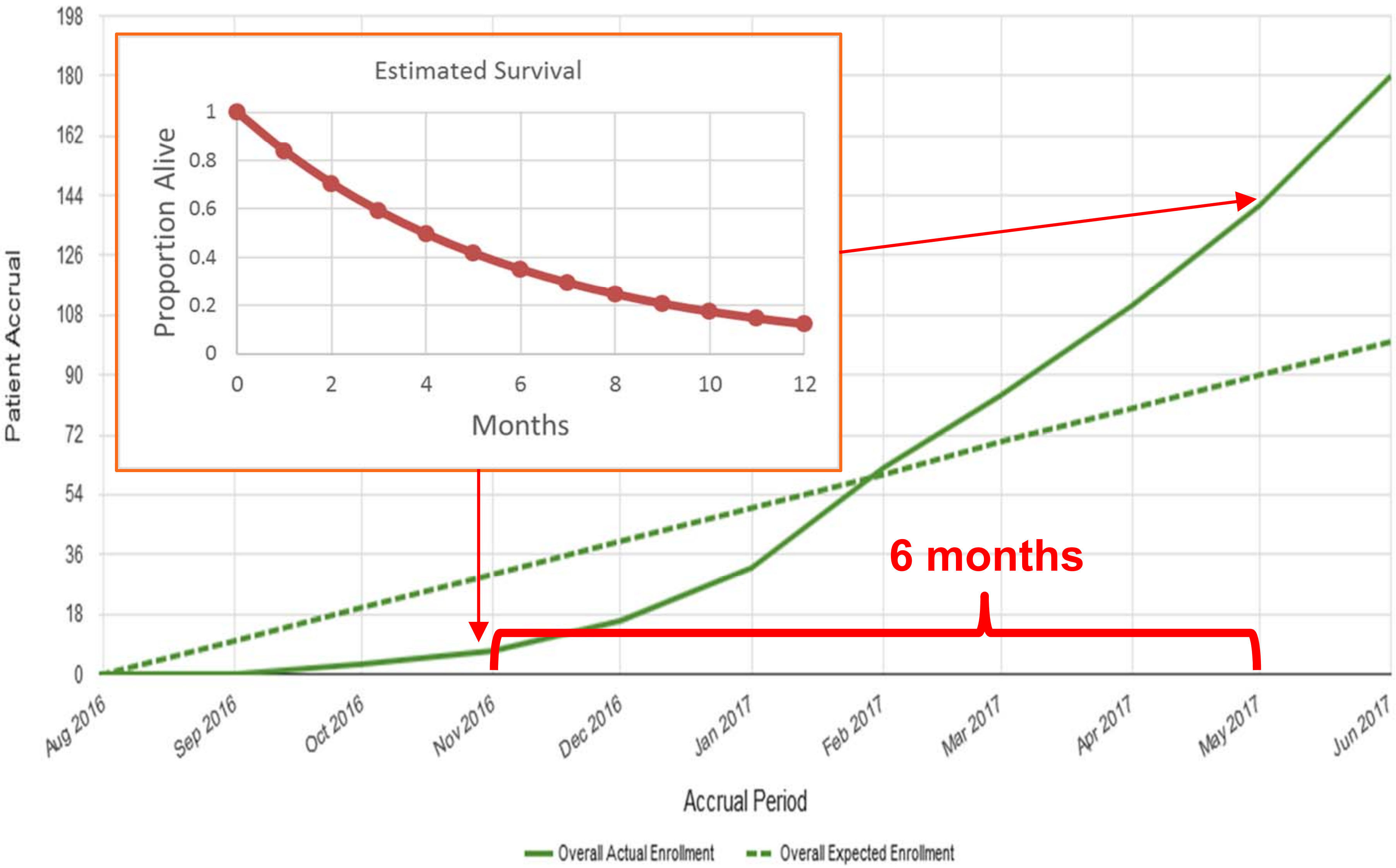
$$re\ Q_c = \frac{n_c}{N}, \quad n_e$$

$$\varphi \quad \frac{\lambda^2}{\qquad}$$

$$-\lambda(T-T_0)) - e^{-\lambda T}\,]/$$

Canadian Cancer Trials Group
Groupe canadien des essais sur le cancer

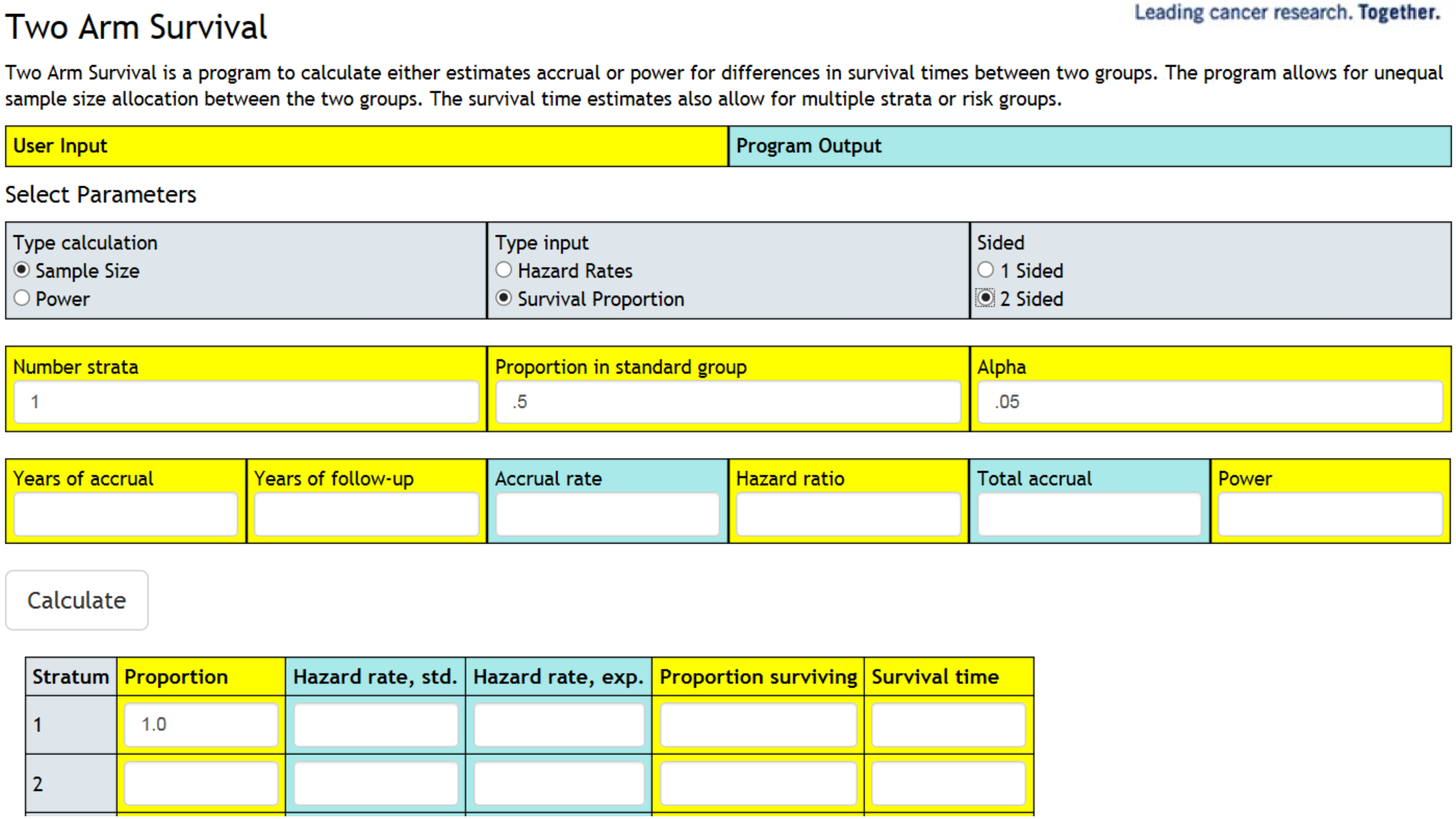
Example: CO.26 Accrual



Canadian Cancer Trials Group
Groupe canadien des essais sur le cancer

Help is at hand!

SWOG
Leading cancer research. **Together.**



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.

User Input	Program Output

Select Parameters

Type calculation	Type input	Sided
● Sample Size ○ Power	○ Hazard Rates ● Survival Proportion	○ 1 Sided ● 2 Sided

Number strata	Proportion in standard group	Alpha
1	.5	.05

Years of accrual	Years of follow-up	Accrual rate	Hazard ratio	Total accrual	Power

Calculate

Stratum	Proportion	Hazard rate, std.	Hazard rate, exp.	Proportion surviving	Survival time
1	1.0				
2					

https://stattools.crab.org/Calculators/twoArmSurvivalColored.htm

Canadian Cancer Trials Group
Groupe canadien des essais sur le cancer

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%
- Accrual Rate = 100 / year
- Duration of Follow-up = 6 months
- Median Survival Control = 6 months
- Hazard Ratio to Detect = 1.25 (0.80)
- 6 months – 7.5 months

= 890

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

- Primary Outcome = Survival
- 1:1 Randomization
- Alpha = 0.05, **1-sided** ↓
- Power = **80%** ↓
- Accrual Rate = 100 / year
- Duration of Follow-up = **18 months** ↑
- Median Survival Control = 6 months
- Hazard Ratio to Detect = **1.50(0.67)**↑
- 6 months – **9 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 (/0.71 or 29% reduction in risk of relapse)
- Median RFS with combination of 35 months
- Absolute improvement in median of 10 months

278 “Events” Required

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 432 patients then follow for an additional 2.75 years = Total Duration of 4.75 years (64%)

b) Accrue for 1.5 years to enroll 324 patients then follow for an additional 6.25 years = Total Duration of 7.75 years (86%)

c) Accrue for 3 years to enroll 648 patients then follow for an additional 0.62 years = Total Duration of 3.62 years (43%)

“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** (/0.77 or 23% reduction in risk of relapse)
- Median RFS with combination of **32.5** months
- Absolute improvement in median of **7.5** 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 648 patients then follow for an additional 2.75 years = Total Duration of 5.75 years (71%)

b) Accrue for 2.5 years to enroll 540 patients then follow for an additional 5.25 years = Total Duration of 7.75 years (85%)

c) Accrue for 4 years to enroll 864 patients then follow for an additional 0.75 years = Total Duration of 4.75 years (43%)

“Too optimistic…”

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

a) Accrue for 4.5 years to enroll 540 patients then follow for an additional 4.25 years = Total Duration of 8.75 years (85%)

b) Accrue for 4 years to enroll 480 patients then follow for an additional 8 years = Total Duration of 12 years (95%)

c) Accrue for 6 years to enroll 720 patients then follow for an additional 1 year = Total Duration of 7 years (63%)

d) Accrue for 5 years to enroll 600 patients then follow for an additional 2.75 years = Total Duration of 7.75 years (76%)