Hypothesis testing on two samples: Sample size for the comparison of two means

Paola Rebora

Example

A randomized trial **aims to evaluate a new (N) blood pressure lowering drug with one already in use (V).** 240 subjects with high blood pressure are recruited and are randomized to the two treatments.

The sample size n = 120 (for each group) was calculated to ensure that a **minimal clinically relevant difference δ = 5 mmHg** could be highlighted

with a **prob. type II error** (do not reject false H_0) $\beta = 0.10$

$1 - \beta = 0.90$ is the prob. to reject H_0 when it is false

1 - β is the power of the test

Given

- variability of both groups: $\sigma = 10$ mmHg
- a probability of type I error (reject true H_0) of 0.01

Type I error risk (α)

Probability of reject H_0 when is true H_0

ex. We conclude that N is better(or worse) than V when it is not (efficacy of treatments N and V is the same).

Usually ≤ 5%

Power (1-β):

Probability of reject H_0 when is true a specific H_1

ex. We conclude that N is better(or worse) than V when it is (efficacy of treatments N and V is different)..

Usually ≥ 80%

Sample size for the comparison of two means

Two sample of 120 subjects guarantee that:

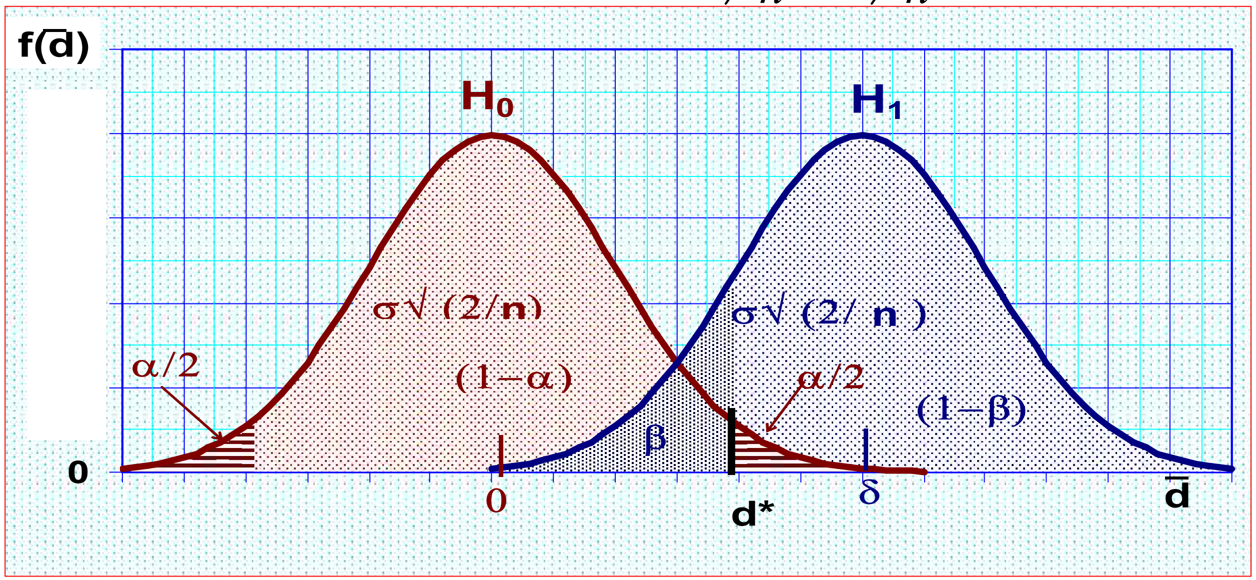
- I will not recognize differences in efficacy between V and N drugs if $\mu_V = \mu_N$ with a probability of 99%.
- I will recognize differences in efficacy **equal to or greater than the lowest clinically relevant value δ** with a probability of 90%.

Sample size for the comparison of two means

δ is in the original scale, so we consider the distribution of the difference $(\bar{x}_N - \bar{x}_V)$ not commensurate with the standard error that (for 2 samples of size n) is Gaussian with

- mean δ and variance $\sigma^2/n + \sigma^2/n = \sigma^2(2/n)$ under H_1
- mean 0 e variance $\sigma^2/n + \sigma^2/n = \sigma^2(2/n)$ under H_0

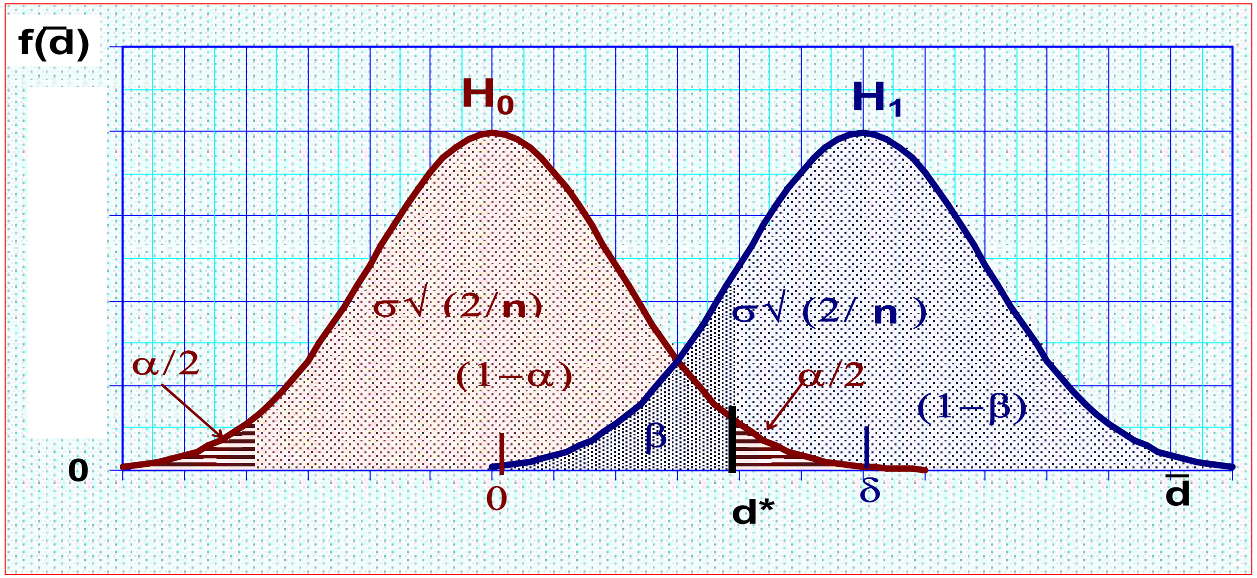
$(\bar{x}_N - \bar{x}_V)$



d* is the threshold of the rejection zone in the original scale

Sample size for the comparison of two means

$$(\bar{x}_N - \bar{x}_V)$$



Under H_0: $$z_{\alpha/2} = \frac{d^* - 0}{\sigma\sqrt{2/n}} \qquad d^* = 0 + z_{\alpha/2} \cdot \sigma\sqrt{2/n}$$

Under H_1: $$-z_{\beta} = \frac{d^* - \delta}{\sigma\sqrt{2/n}} \qquad d^* = \delta - z_{\beta} \cdot \sigma\sqrt{2/n}$$

By equating the two expressions, the required size can be obtained:

$$n = 2(z_{\alpha/2} + z_{\beta})^2 \cdot \frac{\sigma^2}{\delta^2}$$

Sample size calculation

When planning a study we have to power it in order to be able to get an answer for it, that is we have to be sure that we are able to see a difference (δ), if that difference exists.

$$n = 2(z_{\alpha/2} + z_{\beta})^2 \frac{\sigma^2}{\delta^2}$$

α: first type error
1-β: power
σ: standard deviation of the outcome variable in each of the two groups
δ : clinically relevant difference

n: sample size **for each group**

Sample size for the comparison of two means

In the example:

- Given a variability of both groups: σ = 10 mmHg
- a probability of type I error α = 0.01

To highlight

a minimal clinically relevant difference δ = 5 mmHg

with a power 1 - β =0.90

we obtain the following sample size for each arm:

$$n = 2 \cdot (z_{\alpha/2} + z_{\beta})^2 \cdot (\sigma / \delta)^2 = 2 \cdot (2.58 + 1.28)^2 \cdot (10/5)^2 = 119.2$$

Standard error (ES):

In the planning of the study illustrated in our example, we proposed to follow a total of 240 subjects (120 with V and 120 with N): this split of the subjects into the two groups is the most efficient, in the sense that the standard error obtained (for the difference between N and V) is the minimum possible :

$$E.S.(\bar{x}_N - \bar{x}_V) = \sqrt{\sigma^2\left(\frac{1}{n_N} + \frac{1}{n_V}\right)} = \sqrt{10^2\left(\frac{1}{120} + \frac{1}{120}\right)} = 1.29$$

If 60 subjects had been assigned to drug N and 180 to V, the same amount of work would have been done, but a greater standard error would have been obtained:

$$E.S.(\bar{x}_N - \bar{x}_V) = \sqrt{\sigma^2\left(\frac{1}{n_N} + \frac{1}{n_V}\right)} = \sqrt{10^2\left(\frac{1}{60} + \frac{1}{180}\right)} = 1.49$$

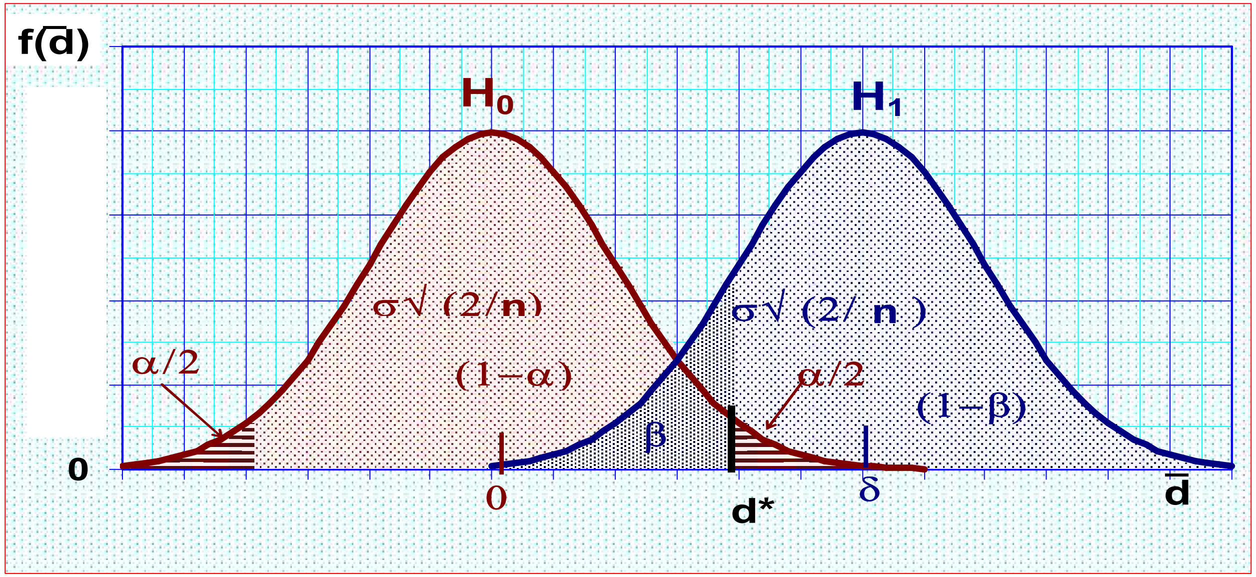
Let us guess how the Power changes

Reducing δ

Increasing σ

Reducing the sample size n

Increasing the α



Exercises

For two analytical methods for the determination of uricemia, one already in use (V) and the other new (N), are known:

- the form of the error distribution (Gaussian)
- the extent of the imprecision (σ = 0.3 mg / dl)

One wonders if "on average" the two methods tend to provide the same value and therefore have the same "accuracy“.

1) Fixing α=0.01 and β=0.1, to highlight a minimum difference of 0.45 mg/dl how many measurments should I perform to test the difference among the two methods?

$$\begin{cases} H_0 : \mu_N = \mu_V \\ H_1 : \mu_N \neq \mu_V \end{cases}$$

1) Given an imprecision of both methods: σ = 0.30 mg / dl and type I error risk set at 0.01 to highlight a minimum technically relevant difference δ = 0.45 mg / dl with a power 1 - β = 0.90 we obtain a single sample size equal to

$$n = 2 \cdot (z_{\alpha/2} + z_{\beta})^2 \cdot (\sigma/\delta)^2 = 2 \cdot (2.58 + 1.28)^2 \cdot (0.30/0.45)^2 \cong 14$$

Thus I need to do 14 measuments with the standard method (V) and 14 with the new one (N) for a total of 28 measurements

Sample size calculation: STATA

power two means 0 0.20, sd(1) power(0.90)

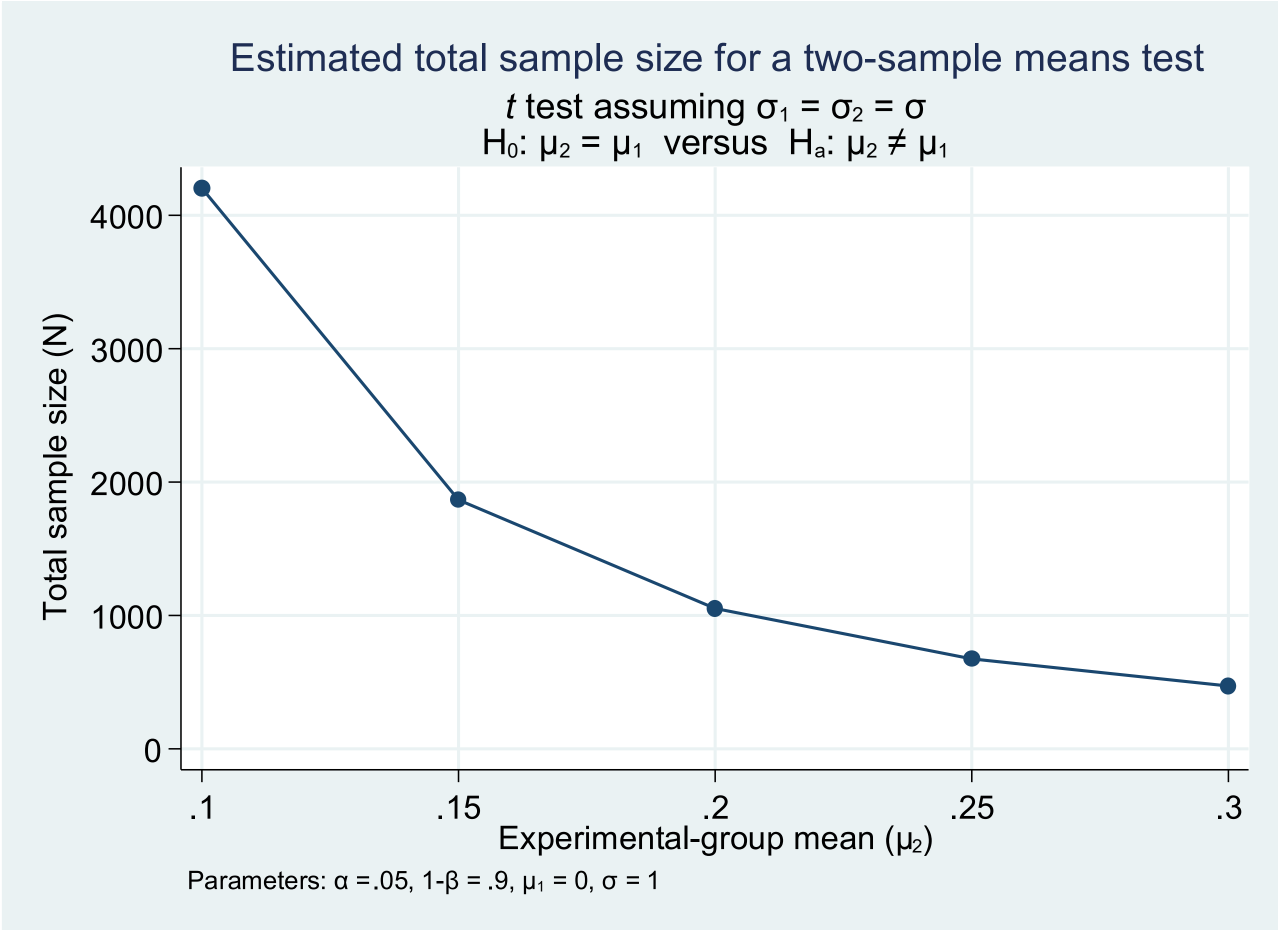
```
Estimated sample sizes for a two-sample means test
t test assuming sd1 = sd2 = sd
Ho: m2 = m1  versus  Ha: m2 != m1

Study parameters:

        alpha =    0.0500
        power =    0.9000
        delta =    0.2000
           m1 =    0.0000
           m2 =    0.2000
           sd =    1.0000

Estimated sample sizes:

            N =      1054
  N per group =       527
```

Sample size calculation

```
power two means 0 (0.10 (0.05) 0.30), sd(1) power(0.90) graph
```

Parameters

- Type I error 5%
- Power 90%
- Mean of control group 0 (μ_1)
- Standard deviation of Y_1 and Y_2 1 (σ)



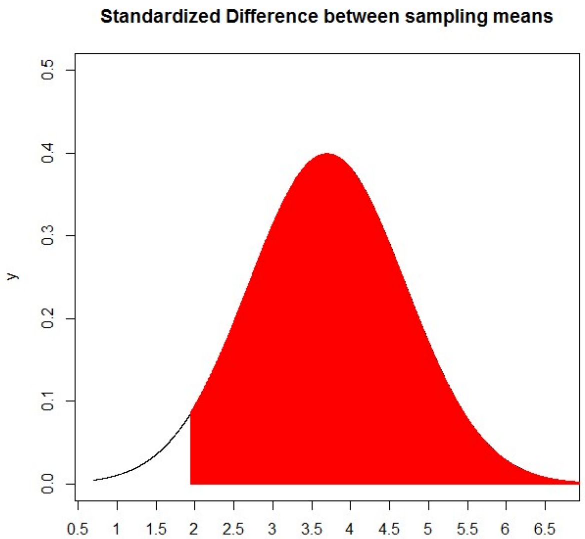
μ_2-μ_1	0.1	0.15	0.2	0.25	0.3

Power assessment

1- β = P(Deciding for H_1 | given that H_1 is true $\mu_E - \mu_{NE} \neq 0$)

Let us assume that $H_1 : \mu_E - \mu_{NE} = \Delta = 0.30$ is true, this implies

$$T = \frac{\bar{Y}_{NE} - \bar{Y}_{NE}}{\sigma * \sqrt{\frac{1}{n_E} + \frac{1}{n_{NE}}}} \sim N\left(\frac{\Delta}{\sigma * \sqrt{\frac{1}{n_E} + \frac{1}{n_{NE}}}} ; 1\right)$$

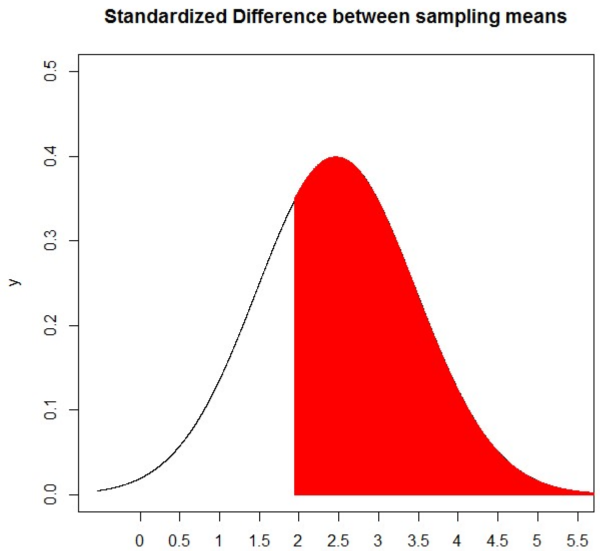


The red area represents the chance (90%) of rejecting H_0 if H_1 is true

Power assessment

Let us change the reference Δ=0.20 for H1

$$T = \frac{\bar{Y}_{NE} - \bar{Y}_{NE}}{\sigma * \sqrt{\frac{1}{n_E} + \frac{1}{n_{NE}}}} \sim N\left(\frac{0.20}{\sigma * \sqrt{\frac{1}{n_E} + \frac{1}{n_{NE}}}}; 1\right)$$

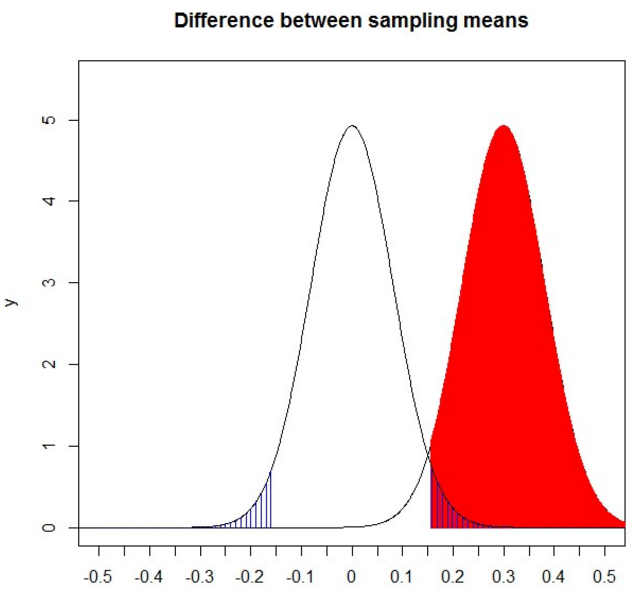


The red area represents the chance of rejecting H_0 if H_1 is true.
The chance is reduced assuming a lower Δ!

Power assessment

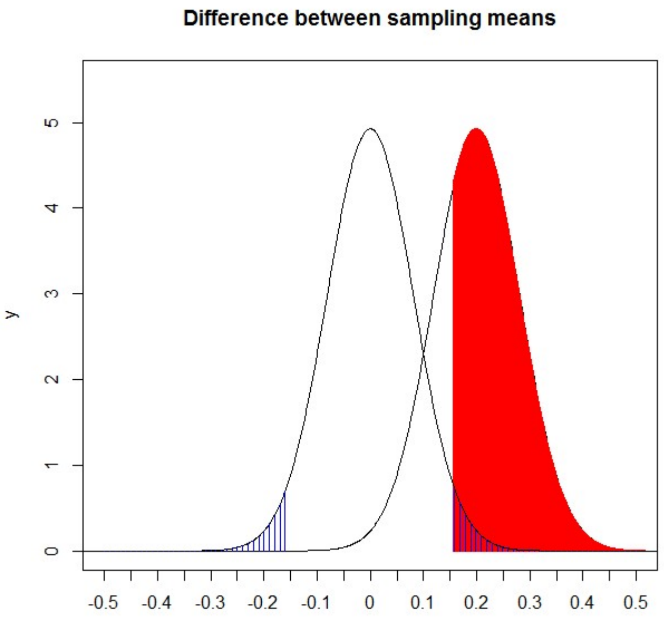
Δ=0.30

$$\bar{Y}_E - \bar{Y}_{NE} \sim N\left(0.30;\sigma\sqrt{\frac{1}{n_E}+\frac{1}{n_{NE}}}\right)$$



Δ=0.20

$$\bar{Y}_E - \bar{Y}_{NE} \sim N\left(0.20;\sigma\sqrt{\frac{1}{n_E}+\frac{1}{n_{NE}}}\right)$$



The red area depends on:

Δ Value for H_1, σ biological variability, sample size n_E and n_{NE}, Blue area

Let us guess the Power changes

$$\bar{Y}_E - \bar{Y}_{NE} \sim N\left(\Delta;\sigma\sqrt{\frac{1}{n_E}+\frac{1}{n_{NE}}}\right)$$

Reducing Δ

Increasing σ

Reducing the sample size n_E and n_{NE}

Increasing the Blue area