httpswww.vox.comfuture-perfect21504366science-replication-crisis-peer-review-statisticsfbclid=IwAR3IlJXfXBVwFWaE5aw4RXHKY

A QUICK REFRESHER ON NULL HYPOTHESIS SIGNIFICANCE TESTING (NHST) AND CONFIDENCE INTERVAL (CI)

Ph.D Programme in Psychology, Linguistics and Cognitive Neurosciences

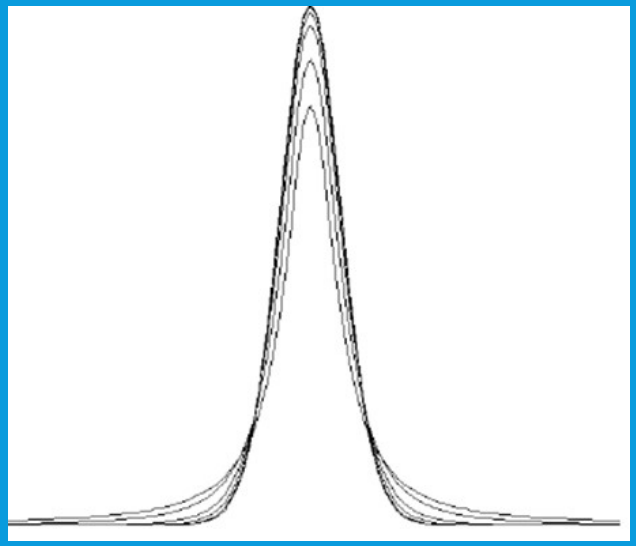
NULL HYPOTHESIS SIGNIFICANCE TESTING (NHST)

- NHST is a method of statistical inference by which an experimental factor is tested against a hypothesis of no effect or no relationship based on a given observation.
- The method has the hypothesis of no effect as the reference, i.e. as if the null hypothesis is true. Hence the denomination Null Hypothesis Testing
- It is recommended to set a *level of significance* (a theoretical p-value) that acts as a reference point to identify significant results
- The approach proposed is of 'proof by contradiction', the null model is the reference and the researcher tests if data conform to it.

A STEP BACK: THE (ESTIMATOR) SAMPLE MEAN

- Our NHST is based on sampling. The reference inferential problem is on the difference between the means of two distinct population, so we test the difference bwtween two independent samples.
- Our sampling method is simple random sample with reimmission. To guess the true value of the means in the populations, we compute the sample means, i.e. the means of the sample . These are sample statistics with optimal properties (called estimators). The sample mean, computed for all possible samples, is a random variable . Sample mean estimator: $\bar{X}$ with capital letter, single value *small*.
- Whatever the distribution of thevariable in the population : $E(\bar{X}) = \mu$ and $\sigma(\bar{X})$= $\sigma/\sqrt{n}$ where E stands for expected value (mean of all sample mean), μ and σ are the true value of the mean and the standard deviation in the population. The standard deviation of the sample mean is called the standard error. Why? Because our estimate becomes more precise when we have a bigger sample, because we divide the standard error

Bigger n, smaller $\sigma/\sqrt{n}$.
The curve tails become smaller and smaller, the curtosis (hight) increases.

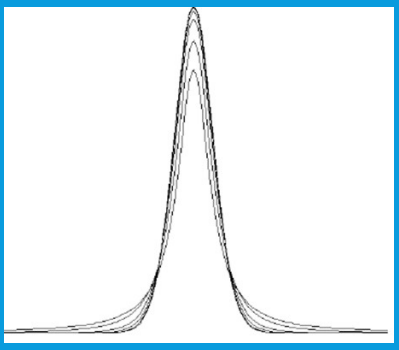


HOW DO WE USE THE SAMPLE MEAN IN THE NHST?

- We move from an idea on the value of the two means, i.e. that they are equal:
- We standardise our test statistics under this null hypothesis
- The denominator depends upon the sample size. The bigger the sample size, the smaller the deviance of our estimator, i.e. the standard error.
- Our precision increases.

Ho: $\mu_{01} - \mu_{02}$=0

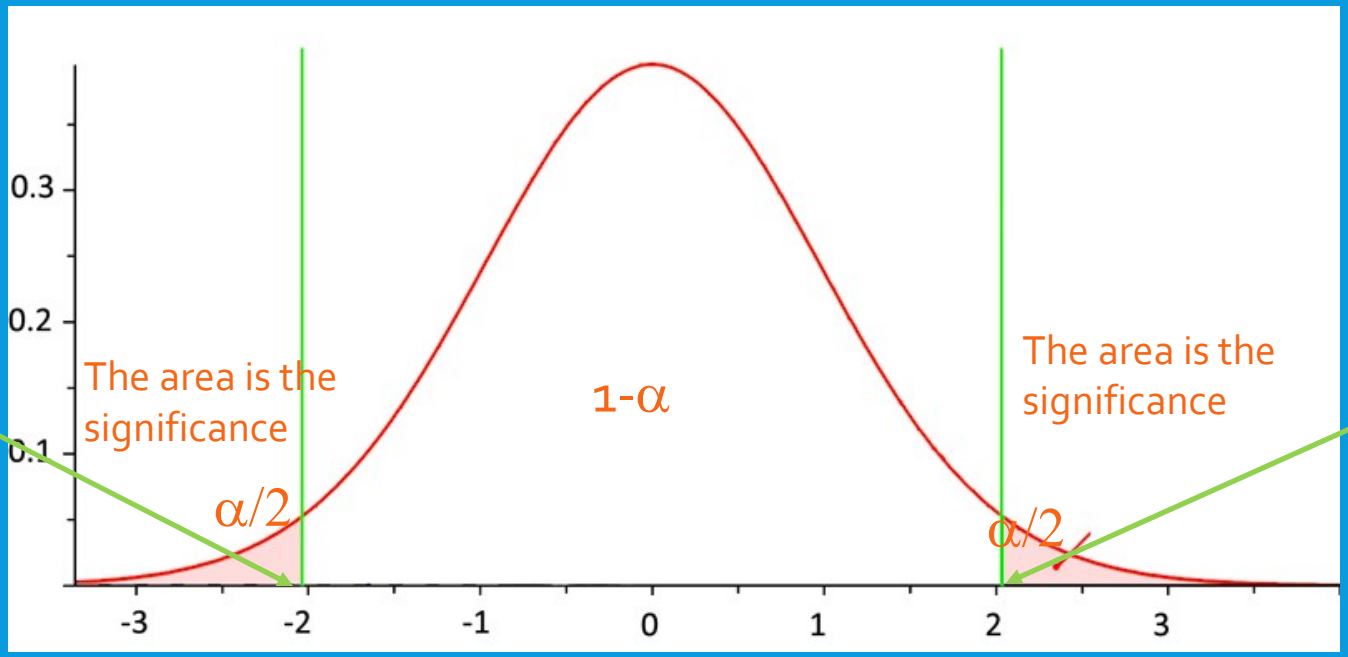
$$\frac{\bar{X}_1 - \bar{X}_2 - \overbrace{(\mu_{01} - \mu_{02})}^{\emptyset}}{\sigma_{\bar{X}_1 - \bar{X}_2}}$$



NULL HYPOTHESIS: SIGNIFICANCE VALUE AND CRITICAL VALUE

- Let us set the level of significance (a theoretical p-value) and the derive the critical value(s)

Criterion (-1.96). The observed value must fall to the right of this point to be significant

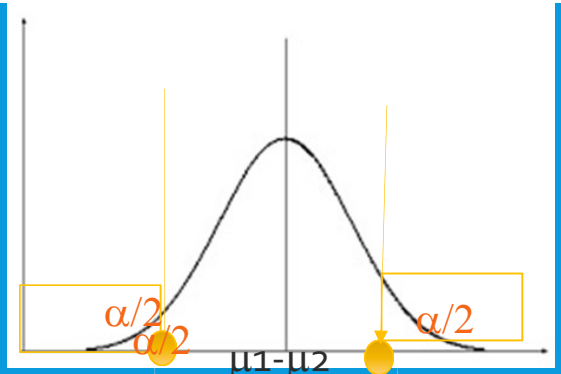


Criterion (+1.96). The observed value must fall to the right of this point to be significant

-1,96

+1,96

BILATERAL OR MONOLATERAL NULL HYPOTHESIS?



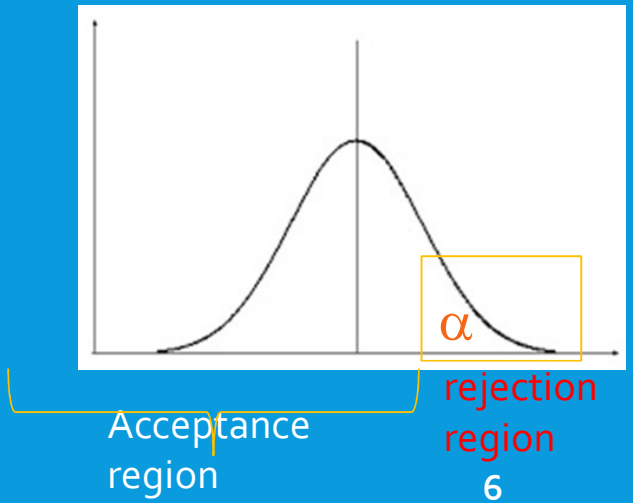
The null hypothesis states NO effect. The opposite can be whatever effect, i.e. $\mu 1 > \mu 2$ or $\mu 1 < \mu 2$. The alternative hypothesis is : Ho: $\mu_1 \neq \mu_2 = 0$

The significance level is split in two equal parts , the two tails in the figure on the left. The rejection region is for values on the abscissa on the two extremes.

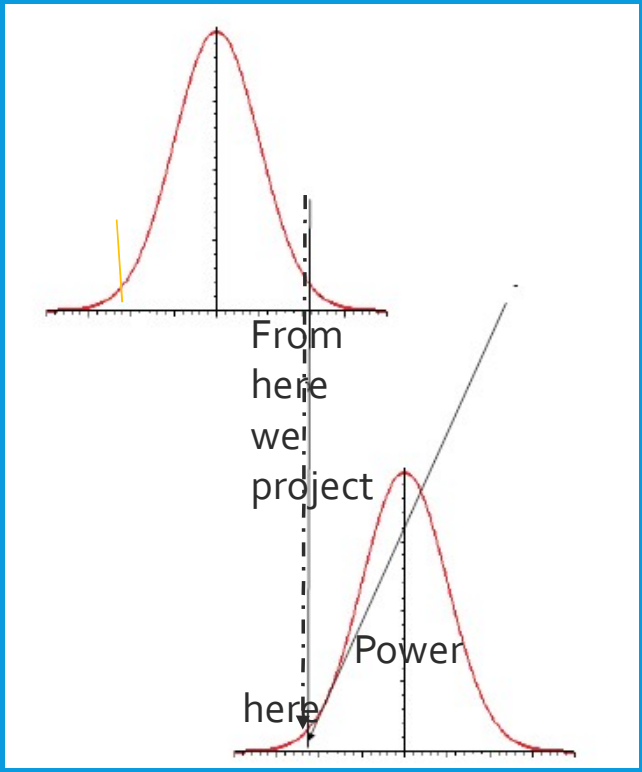
The null hypothesis states NO effect. If we have a precise alternative, our interest can be only for $\mu 1 > \mu 2$ (or only on $\mu 1 < \mu 2$) . The alternative hypothesis is :

Ho: $\mu_1 > \mu_2$

The significance level is on a single tail, the rejection region on the abscissa on the corresponding extreme, as in the figure on the right.



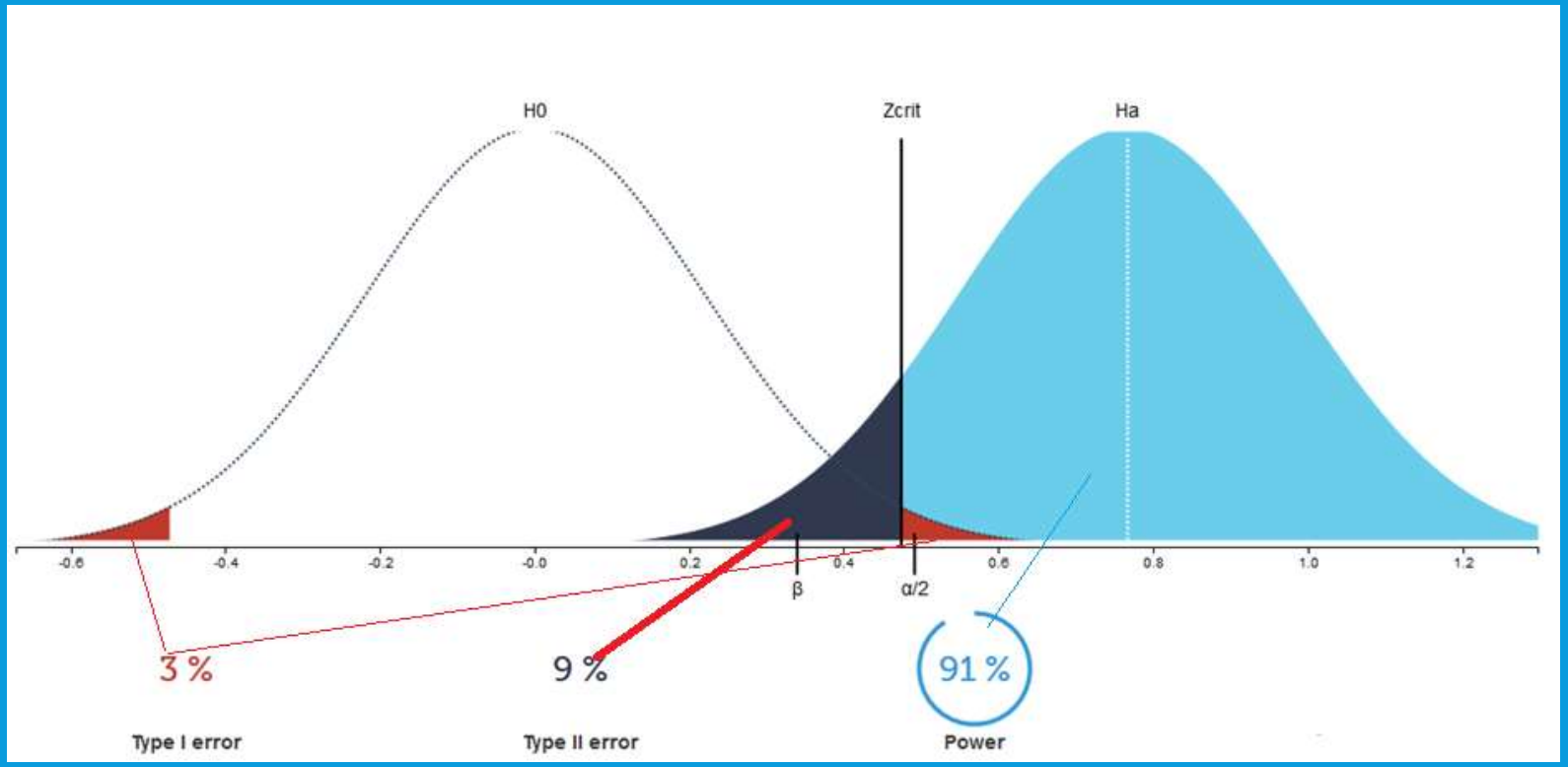
FROM NHST SIGNIFICANCE/CRITICAL VALUE TO POWER



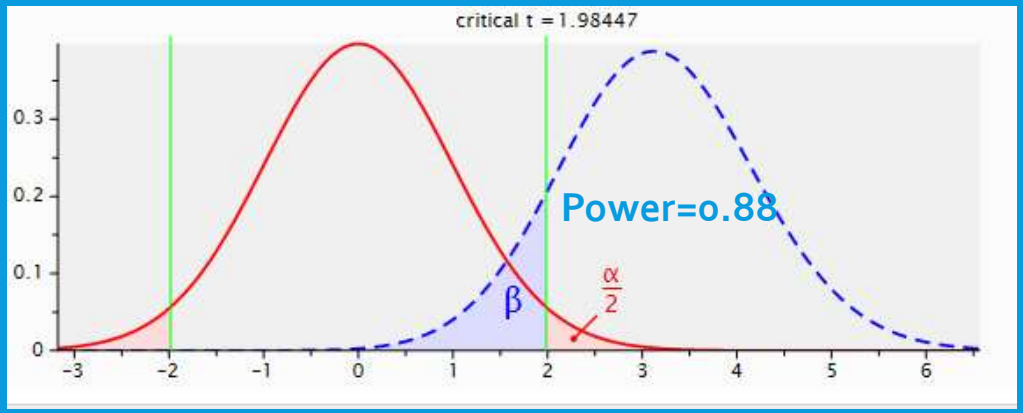
The upper curve is the null hypothesis, the lower curve the alternative one (we drew one of the many possible alternatives).

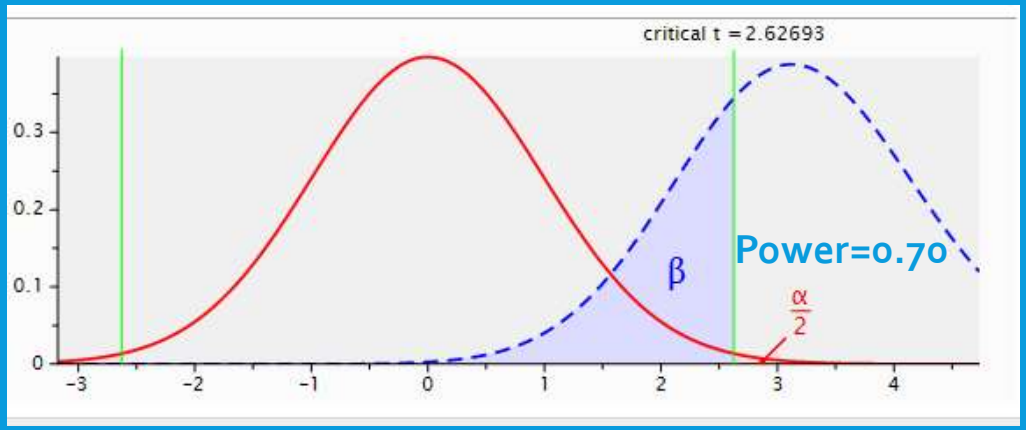
One we fix the significance value, with that specific alternative, the type 2 error and the power (1.type 2 error) is determined.

Here we see null and alternative hypothesis on the same axes, as we do in G*Power



Different significance levels and related power (n=100, ES=0.3, obtained in G*Power)





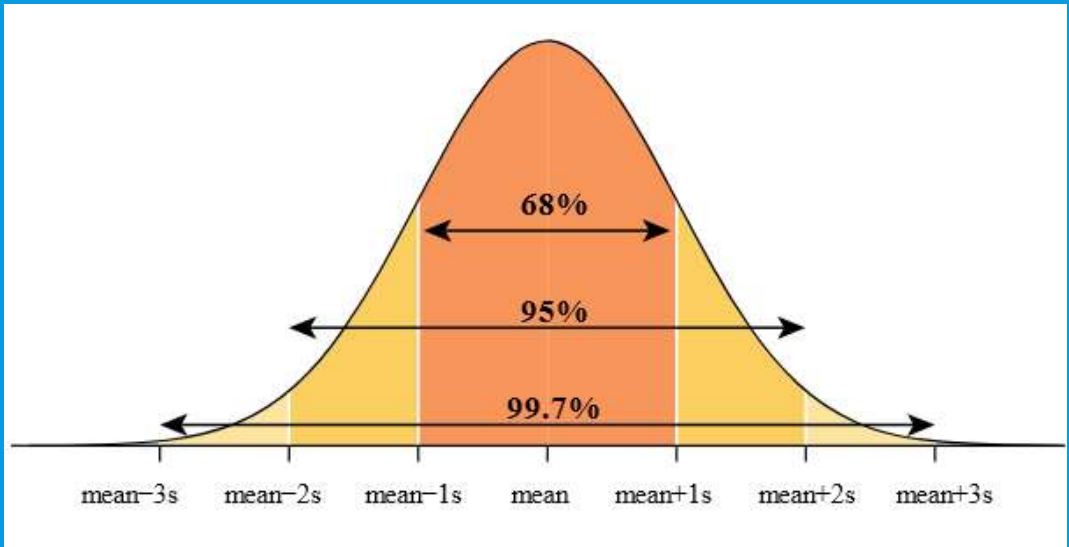
In the first test, significance is 0.05, in the second 0.01.

Holding everyting else constant, a decrease in the significance level implies a decrease in the power of the test

LET'S GET IT RIGHT

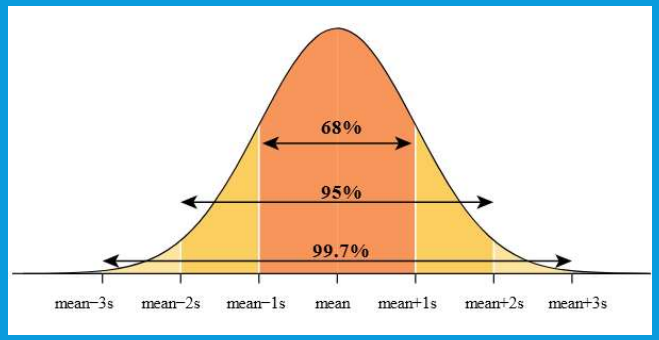
- The *p*-value is not an indication of the strength of an effect.
- 1-*p* is not the probability to replicate an effect
- In the single experiment, either we draw the right conclution or we do not.

CONFIDENCE INTERVALS



- When we compute a confidence interval, we have no a priori idea of the true value of the mean in the population.
- The logic is to capture the mean in an interval that, with probability $1-\alpha$, contains the true value of the mean.

CONFIDENCE INTERVALS/2



- How do we compute a CI? Standardising the sample mean(s), our estimator(s). With CI we do not set any possible value of the mean(s)

$$-t\alpha/2 < \frac{\bar{X}-\mu}{\sigma_{\bar{X}}} + t\alpha/2$$

$$\bar{X} - \sigma_{\bar{X}}\, t\alpha/2 < \mu < \bar{X} + \sigma_{\bar{X}}\, t\alpha/2$$

- We can compute the limits of the CI.
- If the confidence is 95%, it mean that, in the long run, 95% of confidence interval contain the true value of the mean of the population, 5% don't.
- BEWARE: a single CI either contains the true value or it does not contain it.

CI AND REPLICABILITY