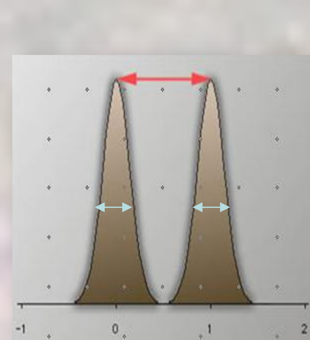
How much does effect size matter?

Aguilar Folgado, A.*; Contreras Ruiz, A**; Marín Montaner, C. ***

Introduction

- Null hypothesis significance testing (NHST) is the dominant statistical approach in biology, although it has many, frequently unappreciated, problems (Jennions, 2001). Most importantly, NHST does not provide us with two crucial pieces of information:
  - *The magnitude of an effect of interest.*
  - *The precision of the estimate of the magnitude of that effect.*

What do you think when you listen "effect size" ?



- Effect size (ES) is the magnitude of an outcome seen in a research as it would be in a population. It represents how different are the results we obtain in a survey. This ES is standardized so we can compare it across different studies.

Why can´t I just judge my result by looking at the *p*-value?

- *P*-value is used to determine if the means of our study are equal or different (statistical significance), but it doesn't give us the result's importance to make future decisions.
- To make this kind of decisions, we need to use the ES because it shows us the importance of the result's comparison.

In this poster, we want to illustrate some effect size's calculation utilities with an experience that investigates the behavior of some measurements made on human hands.

Computing and Interpreting effect size

- Cohen (1969) has been credited with popularizing the term "effect size" because of his famous benchmarks. Table 1 of ***T-shirt effect sizes*** ("small", "medium", and "large")

Table 1

Test	ES. Index	Effect Size: Small	Medium	Large
Comparison of independent Means, m_A, m_B.	$d=\frac{m_A-m_B}{\sigma}$	0.20	0.50	0.80
Correlation.	r	0.10	0.30	0.50
Comparision of two correlations.	$q=Z_A-Z_B$ z= Fisher's z	0.10	0.30	0.50
P=0.5 and the sign test..	g= P-0.5	0.05	0.15	0.25
Difference between proportions.	$h=\phi_A-\phi_B$; $\phi=$ arcsine transformation	0.20	0.50	0.80
Crosstabulation, chi-square for goodness of fit.	$w=\sqrt{\sum_{i=1}^{k}\frac{(P_{1i}-P_{0i})^2}{P_{0i}}}$	0.10	0.30	0.50
ANOVA, one-way.	$f=\frac{\sigma_m}{\sigma}$	0.10	0.25	0.40
Multiple regresion	$f^2=\frac{R^2}{1-R^2}$	0.02	0.15	0.35

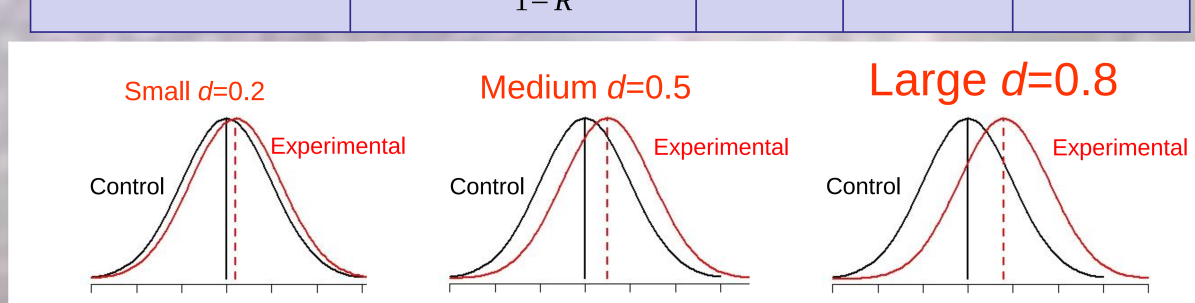


Table 2

Effect Size $d=\frac{m_A-m_B}{\sigma}$	Percentage of control group who would be below average person in experimental group	Rank of person in a control group of 25 who would be equivalent to the average person in experimental group	Probability that you could guess which group a person was in from knowledge of their 'score'.	Equivalent correlation, r	Probability that person from experimental group will be higher than person from control, if both chosen at random
0.0	50%	13th	0.50	0.00	0.50
0.1	54%	12th	0.52	0.05	0.53
Small 0.2	58%	11th	0.54	0.10	0.56
0.3	62%	10th	0.56	0.15	0.58
0.4	66%	9th	0.58	0.20	0.61
Medium 0.5	69%	8th	0.60	0.24	0.64
0.6	73%	7th	0.62	0.29	0.66
0.7	76%	6th	0.64	0.33	0.69
Large 0.8	79%	6th	0.66	0.37	0.71
0.9	82%	5th	0.67	0.41	0.74
1.0	84%	4th	0.69	0.45	0.76
1.2	88%	3rd	0.73	0.51	0.80
1.8	96%	1st	0.82	0.67	0.90
2.5	99%	1st out of 160	0.89	0.78	0.96
3.0	99.9%	1st out of 740	0.93	0.83	0.98

- For the interpretation of Cohen's *d* we have the Table 2.

Material and Methods

- We have made measurements of the area and perimeter (of both hands, left and right) and height of 40 adult people, randomly selected **(21♂ , 20♀).**
- We draw the outline of both hands that we scan later (image resolution: 96 ppp).

- Analyzing the selected hands, Adobe® Photoshop® CS6 extended ,calculated their perimeter and closed area in pixels, that we transform to cm.
- After comparing different measures with paired, independents t-tests and linear models, we calculate useful effect size indicator using formulas in Table 3:

Table 3

Standardized difference		Percentage variance (PV)
$d=t_{ind}\sqrt{\frac{n_1+n_2}{n_1\cdot n_2}}$	$d=t_{Pair}\sqrt{\frac{2\cdot(1-r)}{n}}$	$\eta^2=\frac{f^2}{1+f^2}=\frac{SS_B}{SS_B+SS_W}=\frac{F\cdot df_B}{df_W\cdot F+df_W}$

- We calculate the confidence interval of the *ES* with asymptotic or with bootstrap, using R routines (Nakagawa and Cuthill, 2007) .
- Power analysis allow us to plan future studies using estimations from our pilot experience.
- Finally, we will make hypothesis contrast with non punctual null hypothesis.
- For processing data we used the following statistical software: SPSS v21 and R v2.15.2. GPower 3.1.6 OneStop3

Results

Unpaired comparison

- We found very large effect in differences between sex in hand's area and perimeter.
- We show the biggest difference found in right hand areas: Figure 1 & Table 4

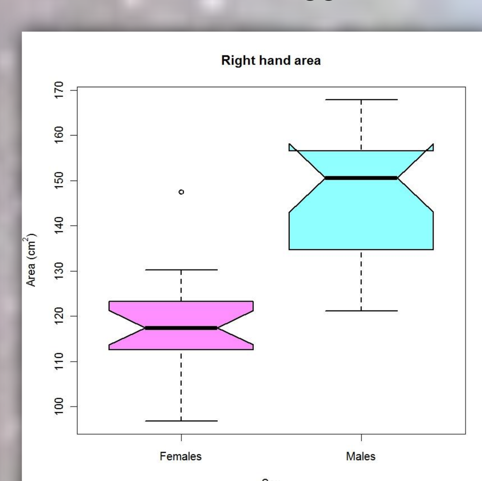


Figure 1. *Right Hand area by gender*.

Table 4

Mean ±SD	Difference	*d*	*PV*
m_{Female}= 118.22 cm² ±10.88 cm² m_{Male}= 146.71 cm² ±12.95 cm²	$m_{Male-Female}$= 28.50 cm² *t*=7.606 P-val=3.2E-9	*d*=2.37 $CI_{95\%}$= [1.53;3.22]	η^2=0.58 $CI_{95\%}$= [0.37;0.72]

Paired comparison

- We don't found any difference between perimeters' hand (*p*=0.96 for Female, *p*=0.85 for Male). In area we found differences statistical significances but the effect size is small. Figure 2.

$m_{difference}$ = 4.65 *t*=3.57 *p*=0.002
Effect size: *d*=0.38 $IC_{95\%}$=[0.18 ; 0.64]

$m_{difference}$ = 2.02 *t*=2.2 *p*=0.04
Effect size: *d*=0.17 $IC_{95\%}$=[0.02 ; 0.43]

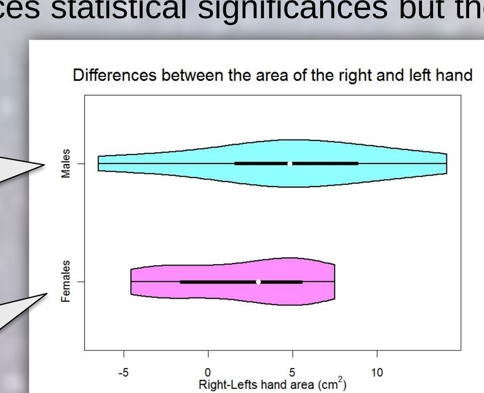
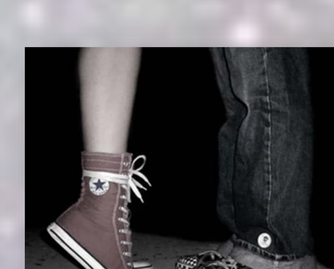


Figure 2. *Difference between right and left hand areas..*



Comparing slops and intercepts in linear model

- Studding the relation between height and hand's area with a linear model, we find small difference in slop by sex, and assuming equal slop, a medium difference in area for the same height. Fig. 3 and 4.

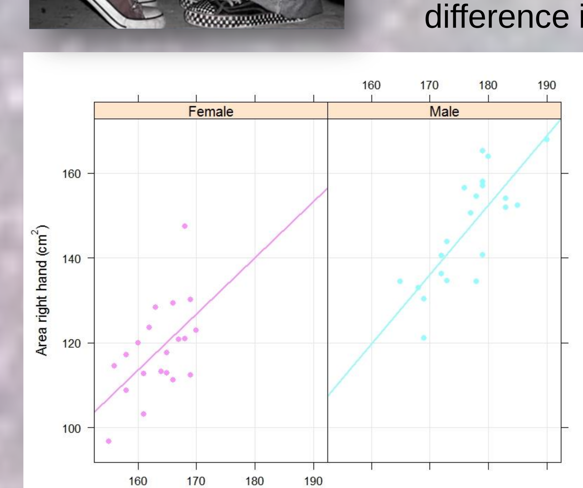


Figure 3 *Model:*
Area~Cte+Sex+βHeight+β_{Sex}Height

β_{Sex} =0,311 *F*=0,322 *p*=0,574
Effect size: η^2=PV=0.009 *d*=0.18
$CI_{95\%}$(PV)=[0.00 ; 0.17]

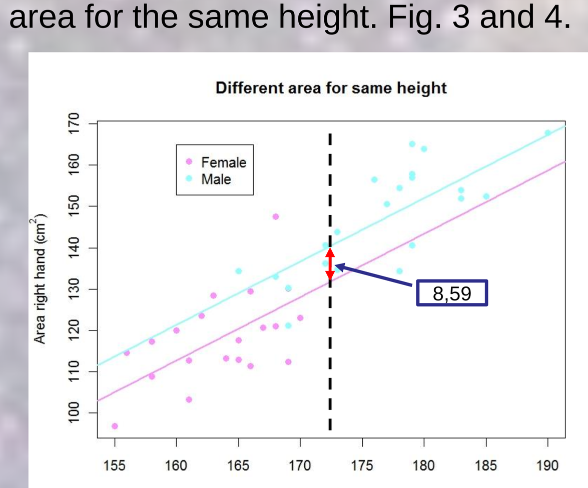
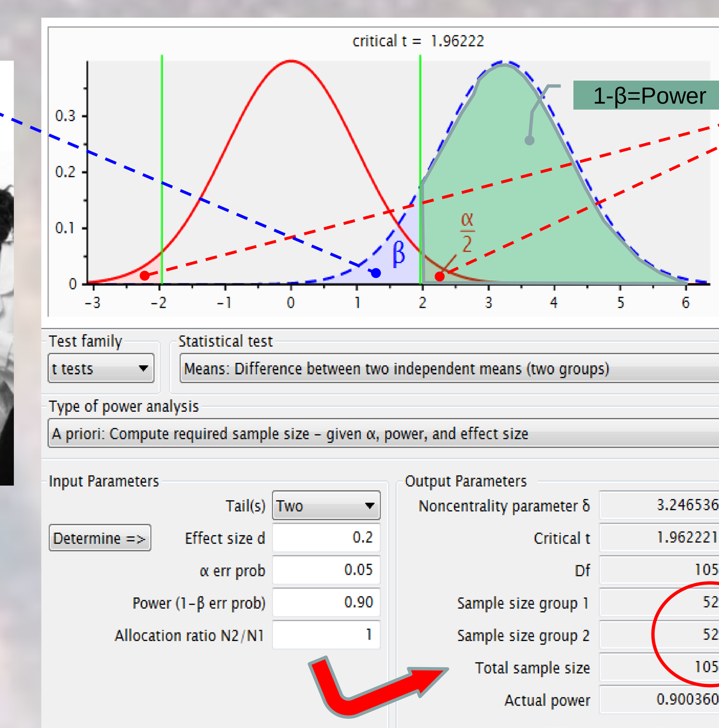


Figure 4 *Model:*
Area~Cte+Sex+βHeight

Sex_{Male} =8.59 *F*=3.97 *p*=0.053
Effect size: η^2=PV=0.10 *d*=0.65
$CI_{95\%}$(PV)=[0.002 ; 0.34]

Use effect size in power analysis

- In our experience we have a mean of ratio Perimeter/Height of 0.61 equal for both sexes. So if there are a small difference (say d=0.2) ¿How big must be the sample for detect such difference significative (α=0.05) with high probability (Power=0.9)?. Results of GPower3 in Figures 5 & 6.



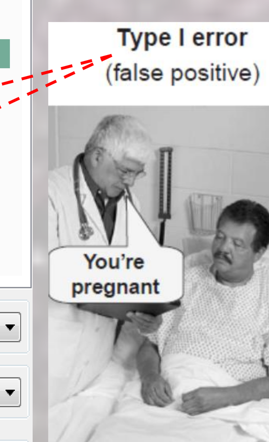


Figure 5. *Power analysisi with GPower3.*

- As we can see in Figure 6, we need about 500 cases in each group to detect a small effect! In Fig.6 the graph shows total sample size for others effect size.

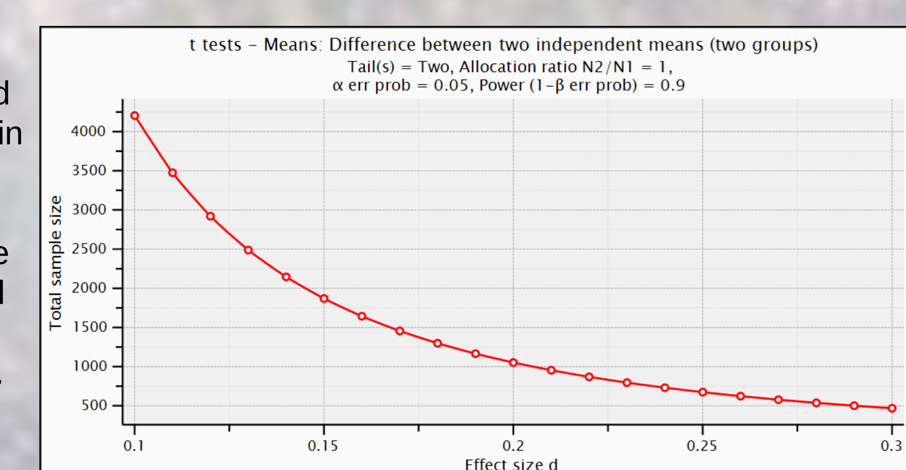


Figure 6. *Sample size versus EF with: d=0.2, .α=0,05, Power=0,9*

Effect size for testing minimum-effect hypothesis

- Rather than testing the hypothesis that treatment have no effect, we might want to test the minimum-effect hypothesis that treatment effect is less than a small PV%(say 1% or 5%).

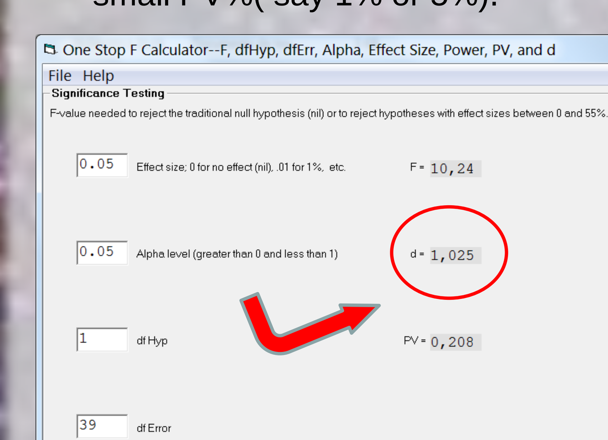


Figure 7. *EF needed to reject minimum-effect hypothesis.*

- Repeating measurements on the same hand, we have *sd*=3 for perimeter hand. What give us a negligible *PV*= 0.05 (*d*=0.5) in the difference between ♂ and ♀ perimeters.
- OneStop *F* (Murphy, 2005), calculate a minimum *d*=1.025 to reject minimum-effect, see Figure 7.
- We observe *d*=1.63 in our study, so we reject minimum-effect hypothesis *d*=0.5.

Conclusions

- With a big sample size virtually any study can be made to show significant results. Alternatively, we can calculate ES and its confidence interval.
- The fact that ES is dimensionless facilitates its comparison through different studies, specially at meta-analysis. Versus this, it complicates its interpretation.
- There can be problems in the standardized ES' interpretation when a sample does not come from a Normal distribution.
- Finally, don't forget that to know the right value a study has, we must take into account the biological importance of the effect.

"If people interpreted effect sizes (using fixed benchmarks) with the same rigidity that α = .05 has been used in statistical testing, we would merely be being stupid in another metric"

Thompson, 2001

References

Coe, R (2002). It´s the Effect Size, Stupid: What effect size is and why it is important. *Education-line.*

Cohen, J. (1969).*Statistical Power Analysis for the Behavioral Sciences,* 1st Edition, Lawrence Erlbaum Associates, Hillsdale (2nd Edition, 1988)

Cohen, J (1992). A power primer. Psychological Bulletin **112** (1): 155-159.

Jennions, MD; Müler, AP. A survey of the statistical power of research in behavioral ecology and animal behavior. *Behavioral ecology* **14** (3): 438-445.

Murphy, KR; Myors, B (2005). *Statistical power analysis: A Simple and General Model for Traditional and Modern Hypothesis Tests,* 2nd edition, Lawrence Eralbum Associates.

Nakagawa, S; Cuthill, IC (2007). Effect size, confidence interval and statistical significance: a practical guide for biologist. *Biological reviews* **82**: 591-605.

Acknowledgments

- This work has been supported by: Servei de Formació Permanent i Innovació Educativa and
- We thank to the numerous classmates and friends who took part in this experience.

anafol@alumni.uv.es; **alcon2@alumni.uv.es; ***montane2@alumni.uv.es** - ***Grade in Biology 1st B 2013