How to get it right: why you should think twice before planning your next study

Part II

Marco Perugini

Rovereto, 20/05/2016

The problem

- Assume that, as scientists, we all want to get it right
- What can we do to increase our chances?

a) Get it right ≠ I am right

b) Get it right ≠ Get it published

Why now?

- The last few years have probably seen more developments in research methodology than in the previous decades
- Rapid changes in standards for publishing
- Until 2011: Methodology = Boooring!
- From 2012: Methodology = Cool!
- Why?

(see also Perugini 2014, GIP)

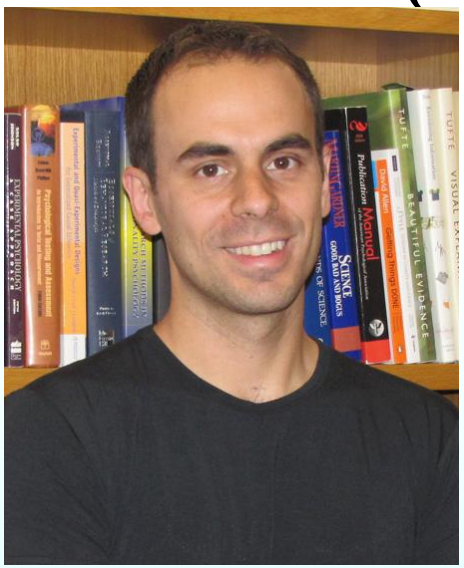
A year to remember

- The year 2011 has been an *annus horribilis* for Psychology
- Three main events:

Bem

Stapel

Simmons (et al)



Bem (JPSP, 2011)

- Nine experiments showing ESP
- Strong reactions
- Initial article failing to replicate the results was refused by JPSP
- Hard questions on the modal way of analyzing data and on “cherry-picking” results
- Galak et al (2012): 7 failed replication attempts (n=3289)

Stapel (September 2011)

- Resigned from Dean at Tilburg University (NL)
- Faked data: 53 retracted papers and 10 PhD thesis with invented or dubious data
- Levelt report (2012): proofs beyond doubts of faked data and strong criticisms to the scientific community
- Huge media impact

Simmons et al (PS, 2011)

False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant

Joseph P. Simmons[1], Leif D. Nelson[2], and Uri Simonsohn[1]
[1]The Wharton School, University of Pennsylvania, and [2]Haas School of Business, University of California, Berkeley

- They show that commonly used questionable research practices can allow to provide empirical evidence even for null effects (false positive)
- Huge scientific impact (740 citations, 2nd most cited paper from 2011 in Psychology)

And many other events…

- Bargh, Kanheman, Sanna, Smeesters, Simonsohn, Francis, and so on
- Until one buzzword came out

Replicability

- and Psychology was born again …

Replicability

- If a result is not replicated, it is not valid
- To be replicated, it needs to be replicable

Replicability

- A key concept in Science
- Almost forgotten in Psychology
- Now in the forefront
- What does it mean?

European Journal of Personality, *Eur. J. Pers.* **27**: 108–119 (2013)
Published online in Wiley Online Library (wileyonlinelibrary.com) **DOI**: 10.1002/per.1919

Recommendations for Increasing Replicability in Psychology†

JENS B. ASENDORPF[1]*, MARK CONNER[2], FILIP DE FRUYT[3], JAN DE HOUWER[4], JAAP J. A. DENISSEN[5], KLAUS FIEDLER[6], SUSANN FIEDLER[7], DAVID C. FUNDER[8], REINHOLD KLIEGL[9], BRIAN A. NOSEK[10], MARCO PERUGINI[11], BRENT W. ROBERTS[12], MANFRED SCHMITT[13], MARCEL A. G. VANAKEN[14], HANNELORE WEBER[15] and JELTE M. WICHERTS[5]

[1]*Department of Psychology, Humboldt University Berlin, Berlin, Germany*
[2]*Institute of Psychological Sciences, University of Leeds, Leeds, UK*
[3]*Department of Developmental, Personality and Social Psychology, Ghent University, Ghent, Belgium*
[4]*Department of Experimental Clinical and Health Psychology, Ghent University, Ghent, Belgium*
[5]*School of Social and Behavioral Sciences, Tilburg University, Tilburg, The Netherlands*
[6]*Department of Psychology, University of Heidelberg, Heidelberg, Germany*
[7]*Max Planck Institute for Research on Collective Goods, Bonn, Germany*
[8]*Department of Psychology, University of California at Riverside, Riverside, CA USA*
[9]*Department of Psychology, University of Potsdam, Potsdam, Germany*
[10]*Department of Psychology, University of Virginia, Charlottesville, VA USA*
[11]*Department of Psychology, University of Milano-Bicocca, Milan, Italy*
[12]*Department of Psychology, University of Illinois, Chicago, IL USA*
[13]*Department of Psychology, University of Koblenz–Landau, Landau, Germany*
[14]*Department of Psychology, Utrecht University, Utrecht, The Netherlands*
[15]*Department of Psychology, University of Greifswald, Greifswald, Germany*

Conditions for Replicability

- The study should be described in a way such that everyone qualified can replicate it
- This implies a very detailed method section, including information that often is not disclosed
- Also, the data should be publicly available (at very least, upon request) to replicate the results using appropriate analyses (internal replication)
- Transparency in research

Back to basics (and beyond…)

- As scientists, we all want to get something right
- If we get it right, it is replicable and will be replicated
- But what does it mean “to get it right”?
- What can we do to increase our chances?
- We need first to go back to some basic concepts

The real problem

- The world is uncertain
- Knowledge is imperfect
- We deal with "samples" rather than "population" (no matter whether you use a Bayesian or a Frequentist approach)
- We try to make inferences from them
- Bertrand Russell's inductivist turkey

"Essentially, all models are wrong, but some are useful" (Box & Draper, 1987)

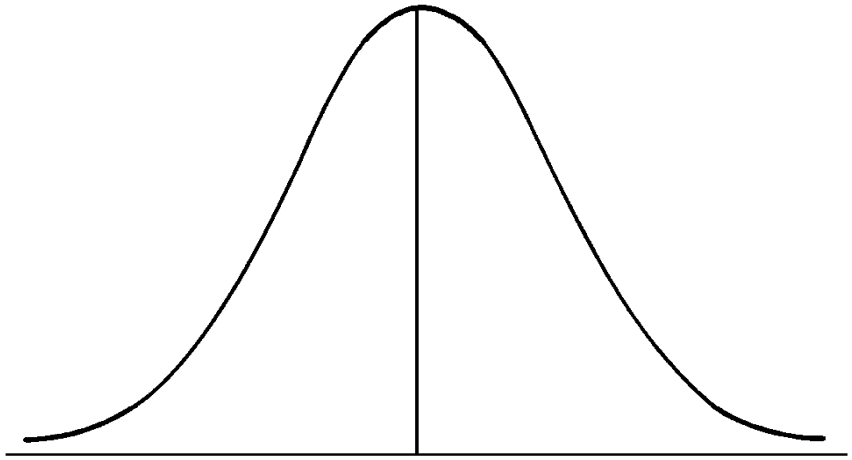
What comes next

- Mean, Standard deviation, Standard error
- Confidence intervals
- Effect size
- Errors of statistical inference
- Power analysis
- Replicability of psychological research

Mean

- A single value that reflects the central point of a distribution
- If the distribution is normal, it is also the best simple way to summarize it

$$\overline{X} = \frac{\sum X_i}{N}$$

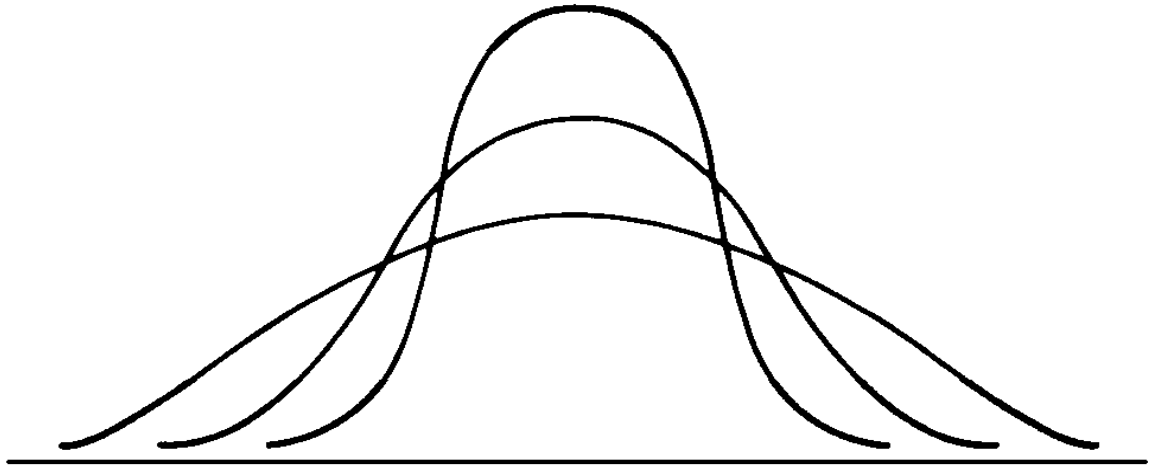


Variance and standard deviation

- Reflects the dispersion (variability) around the mean

$$s^2 = \frac{\sum (X_i - \overline{X})^2}{N} = \frac{\sum X^2}{N} - \overline{X}^2$$

$$s = \sqrt{s^2}$$



Standard error

- When we measure something, the more the observed data, the less the measurement error

- For example, exit polls are more accurate (less error) the more the sampled voters or polling stations

- The point is that we have a sample but would like to say something about the underlying population (or anyway something that generalizes beyond that sample)

Standard error and variance

- Standard error does not depend only from how big is a sample size but also from the variability (variance) of the study object

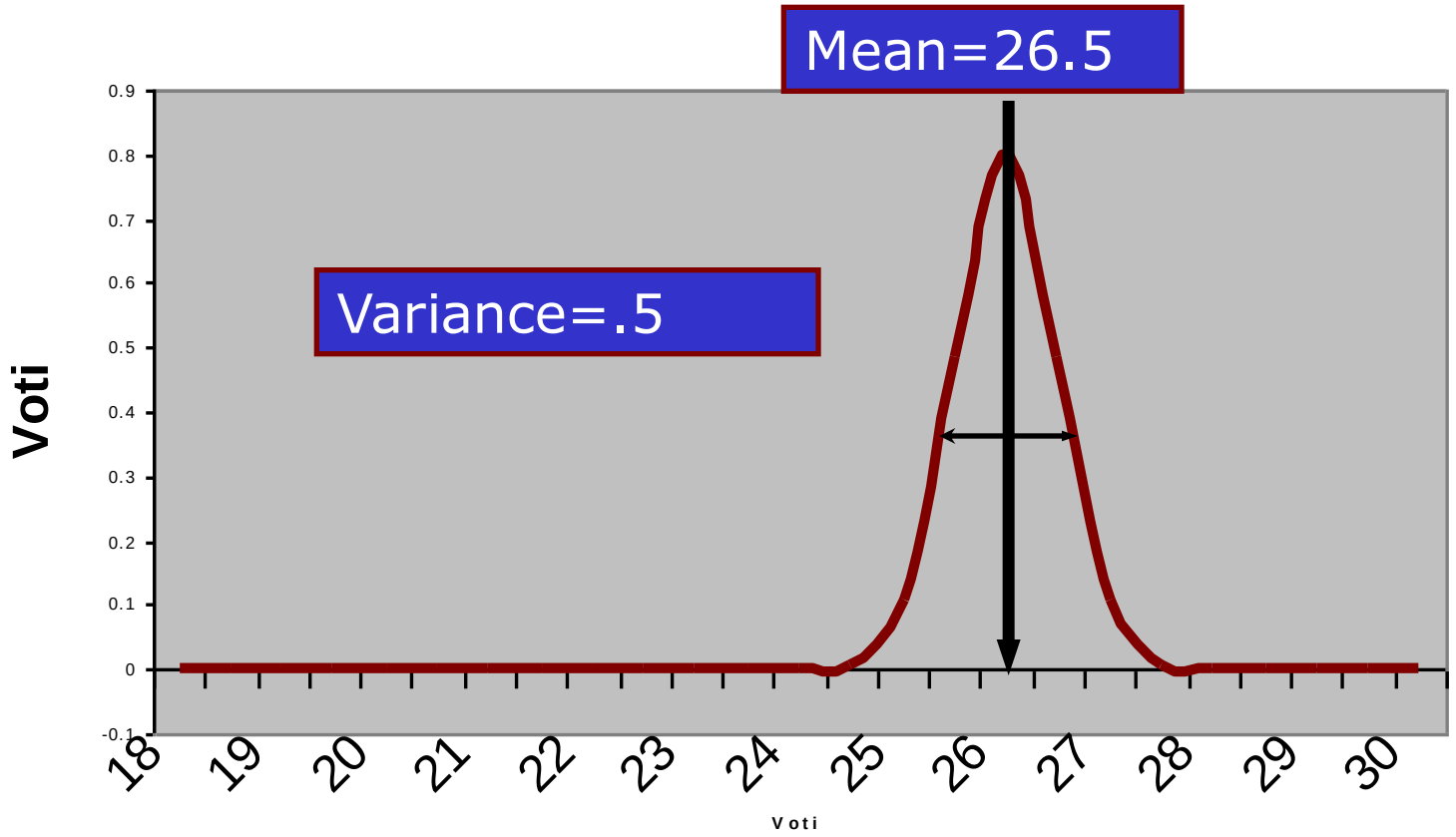
- If everyone answers in the same way, one needs to ask to only one person…

- If people have very different opinions, one need many of them to be able to say something about «what they think»…

- Standard error provides a link between sample and population

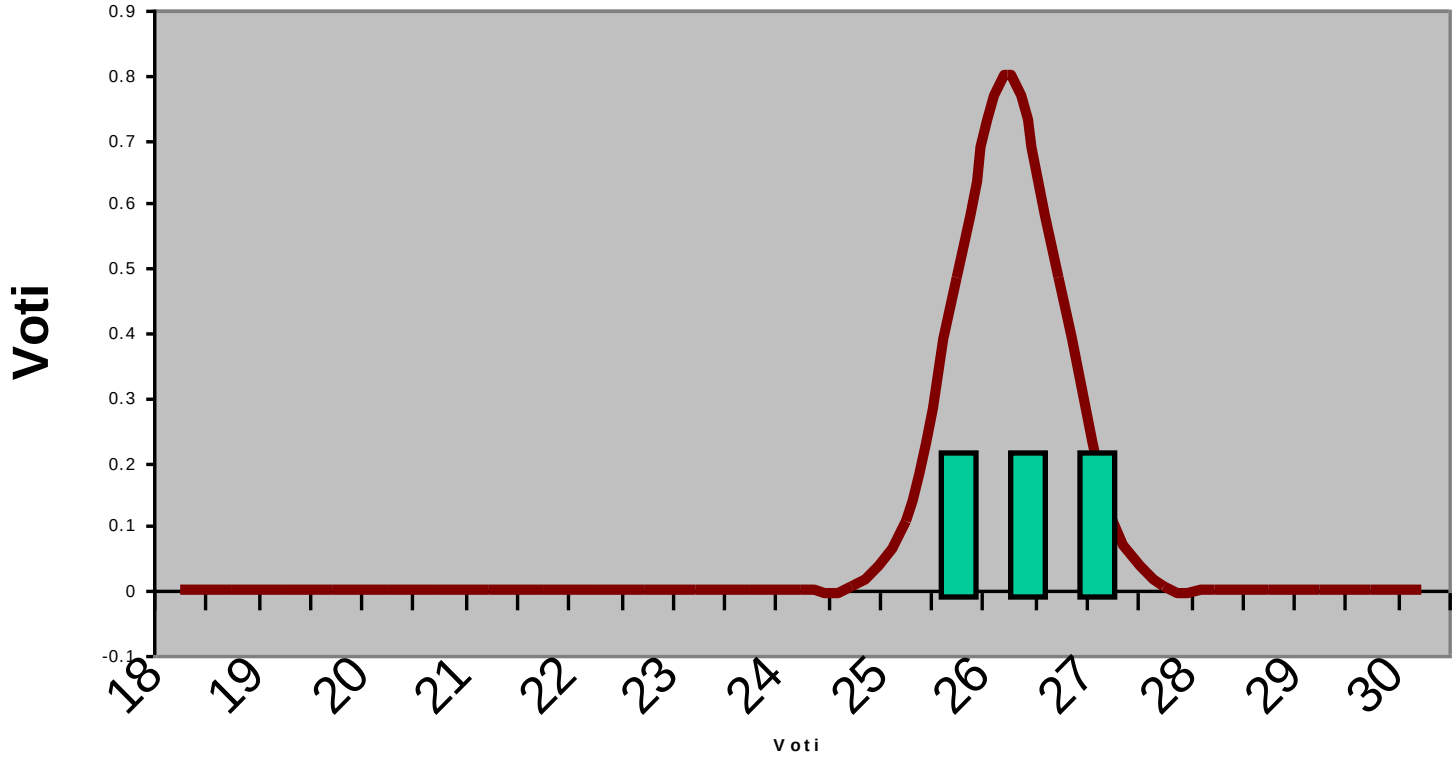
Error and variability

- Suppose that most students have a mark between 26 and 27



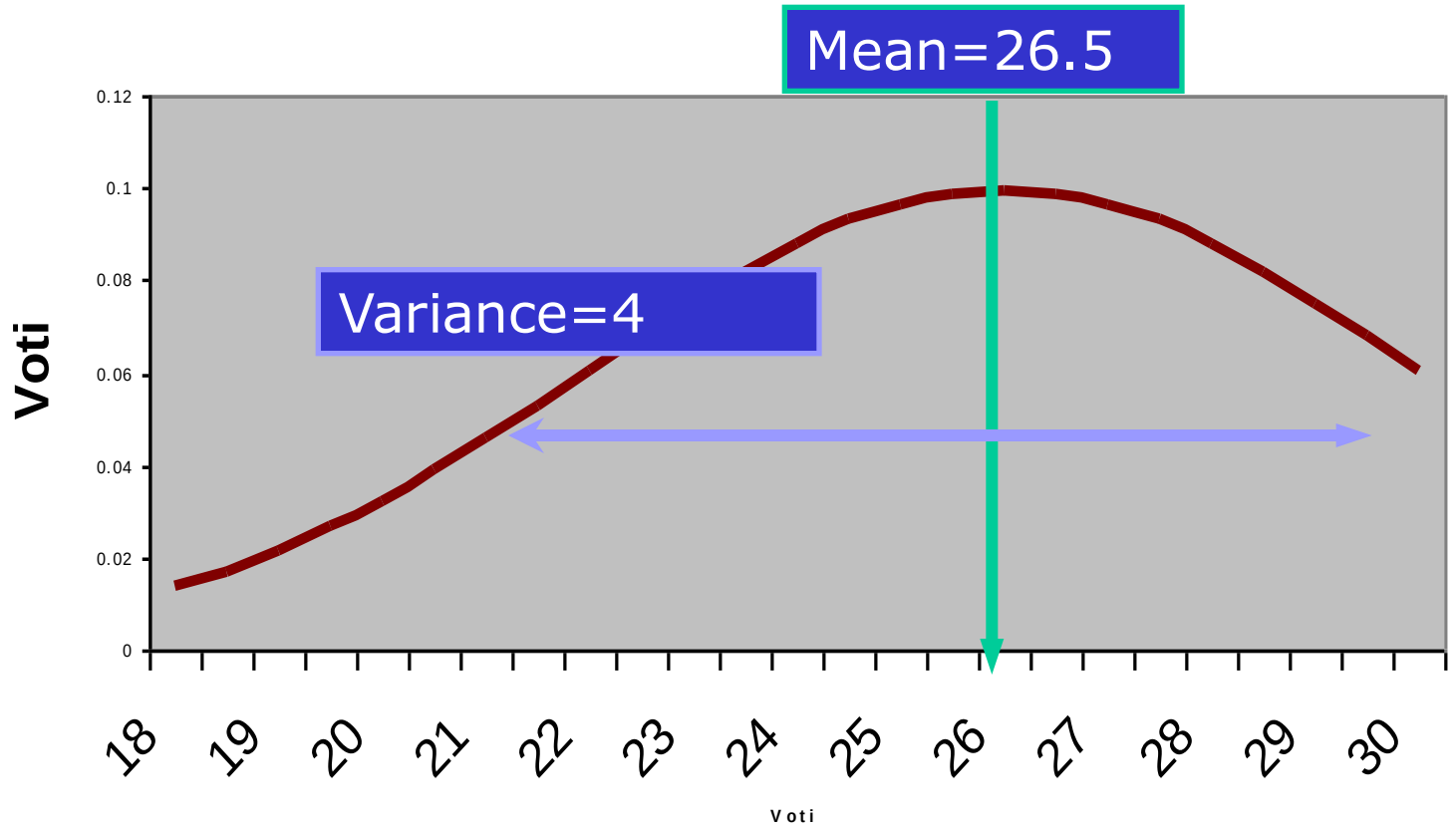
Error and variability

- Almost all possible samples, even if small, will have values near to the population mean of 26.5



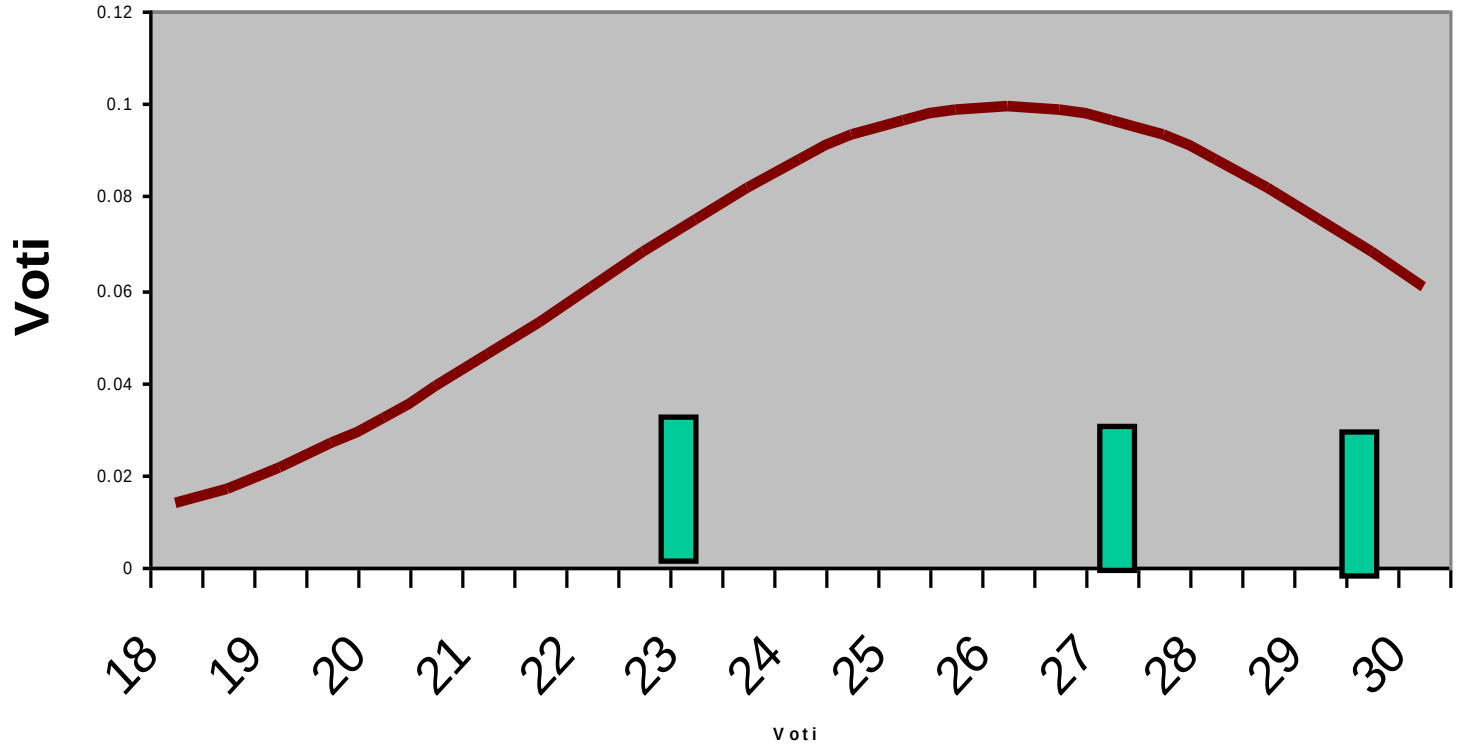
Error and variability

- Suppose now that there is a lot of variability in the marks



Error and variability

- The samples, especially if small, can be very different from each other and from the mean population value



Standard error

- When we estimate a parameter (e.g., mean) in a sample we will make an *estimation error* of the population parameter

- The size of this error is given by the *standard error*

Standard error

$$SE = \sqrt{\frac{S^2}{n}}$$

Variance

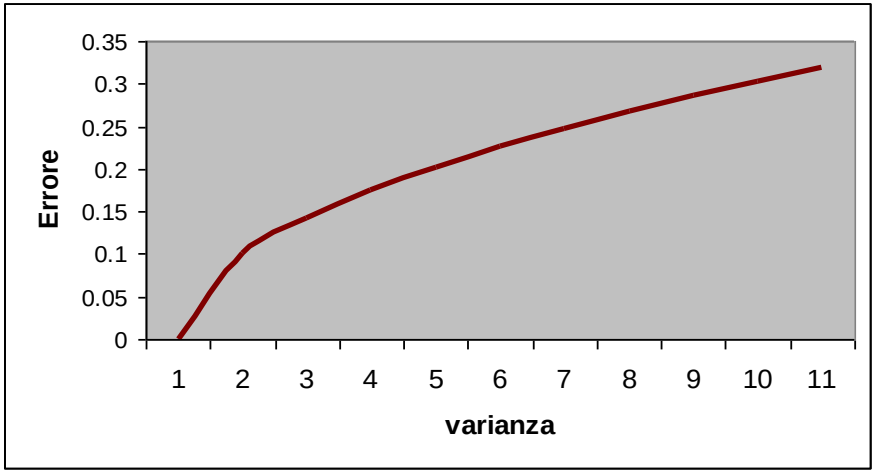
Sample size

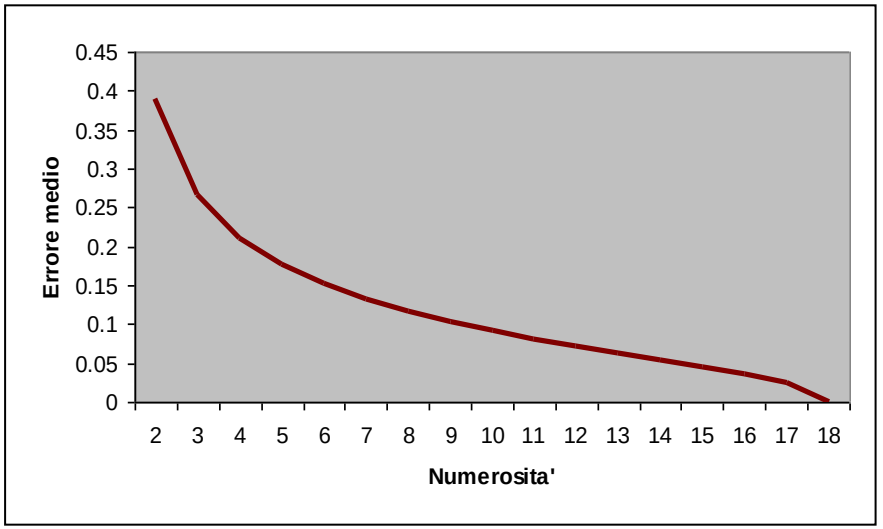
Standard error

Goes up with increasing variance

$$SE = \sqrt{\frac{S^2}{n}}$$

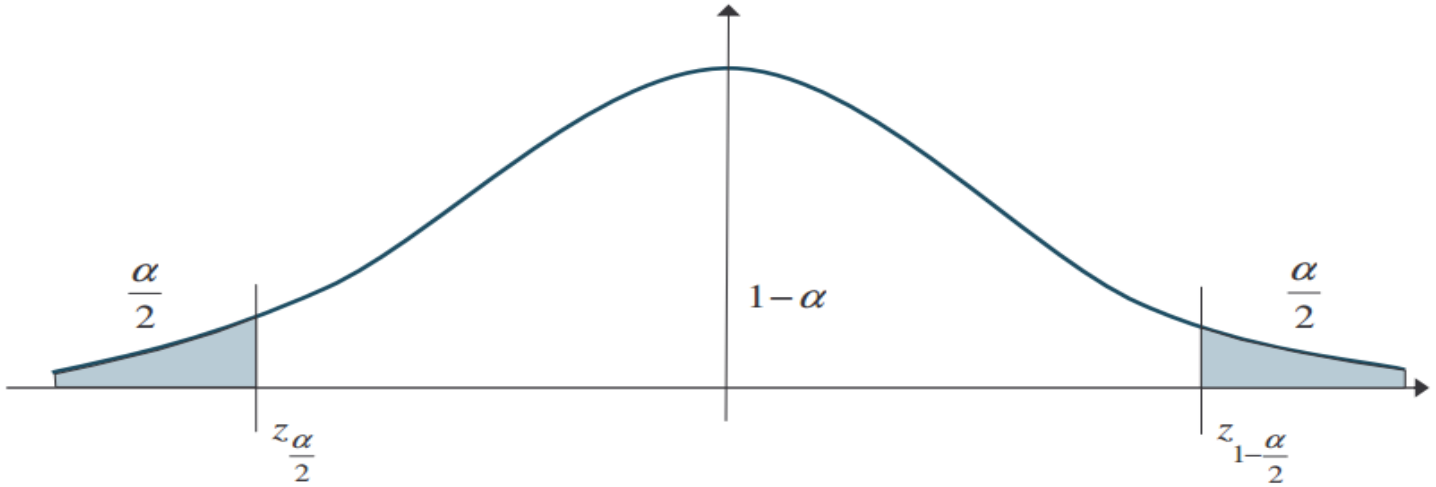
Goes down with increasing sample size





Confidence Interval (CI)

The sample estimate does not necessarily correspond to the population value
It is possible to estimate a Confidence Interval that provides a range of values that contain the population value with a certain likelihood



$$\mu \in \left(\overline{X}_n \pm t^{(n-1)}_{1-\frac{\alpha}{2}} \sqrt{\frac{\overline{s}_n^2}{n}} \right)$$

Standard Error

Confidence Interval (IC)

The typical CI is 95%, meaning that there is a 95% likelihood that if we were to repeat the study 100 times, 95% of the times we would obtain one of the values within the interval including the sample estimate of the parameter (e.g., mean)

To simplify, CI 95% is roughly equal to the sample mean +/- 2 SE.

For example

M = 5; DS = 4 N=100

$$SE = \sqrt{\frac{4^2}{100}} \text{ o } \frac{4}{\sqrt{100}} = 0.4 \quad \text{Range: } 2 \times SE = 0.8$$

95% CI = [4.2, 5.8]

Effect size

- A significant effect says little about the size of the effect
- The p-value depends on sample size and not only effect size
- There are a number of ways to measure effect size
- Most common: Cohen's ***d***

$$\boldsymbol{d = \frac{M_1 - M_2}{pooled\ SD}}\ ,\ \text{with } pooled\ SD = \sqrt{\frac{(n_1-1)SD_1^{\,2}+(n_2-1)SD_2^{\,2}}{n_1+n_2-2}},$$

If the two groups have the same sample size,

$$\boldsymbol{pooled\ SD = \sqrt{\frac{SD_1^{\,2}+SD_2^{\,2}}{2}}}$$

Effect size

- But also **r (correlation coefficient)**

- $r = \sqrt{\frac{t^2}{t^2 + df}}$ $r = \sqrt{\frac{F(1,-)}{F(1,-) + df_R}}$ $r = \sqrt{\frac{\chi^2(1)}{N}}$

- From **d** to **r** and from **r** to **d** **(e.g.** $d = \frac{2r}{\sqrt{1 - r^2}}$

Example of Cohen's d

$$pooled\,SD = \sqrt{\frac{(n_1-1)SD_1{}^2+(n_2-1)SD_2{}^2}{n_1+n_2-2}} = \sqrt{\frac{(246-1)1.107^2+(219-1)1.197^2}{246+219-2}} = 1.150$$

$$\boldsymbol{d = \frac{M_1 - M_2}{pooled\,SD} = \frac{7.79 - 7.57}{1.150} = 0.19}$$

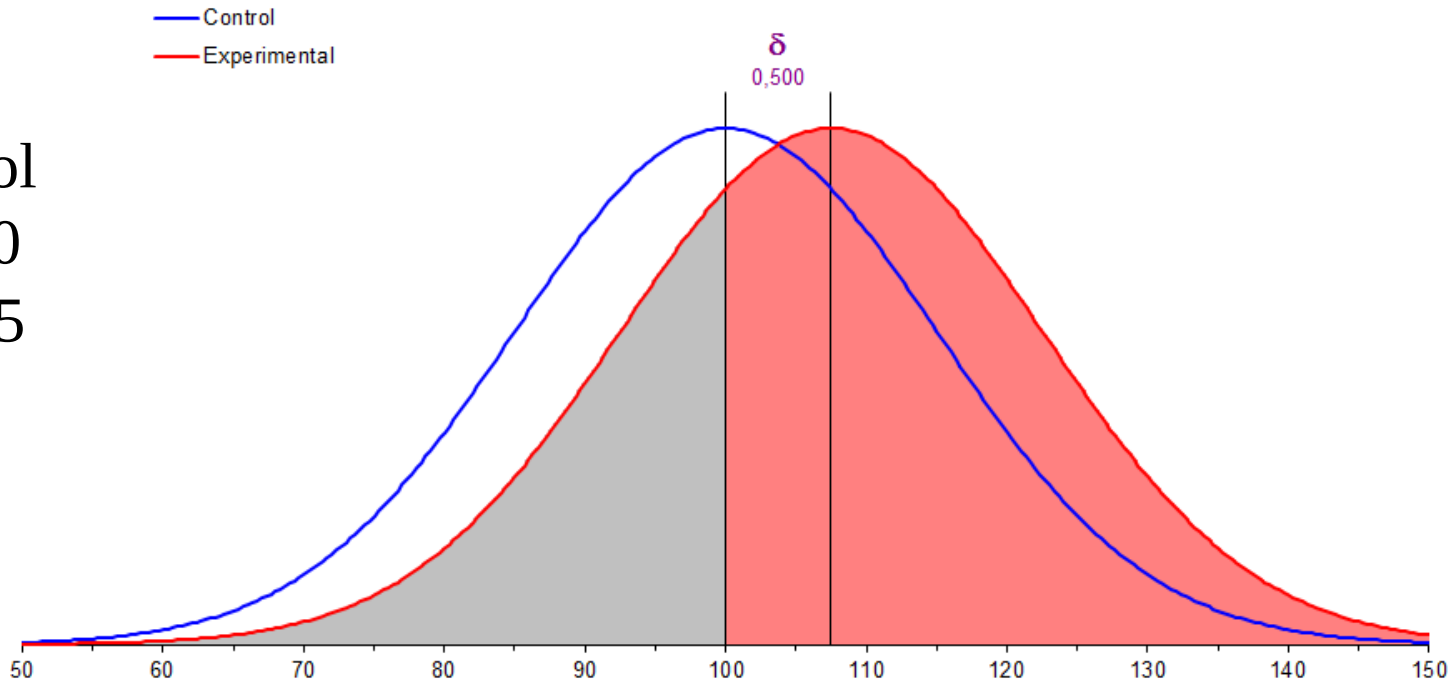
Statistiche di gruppo

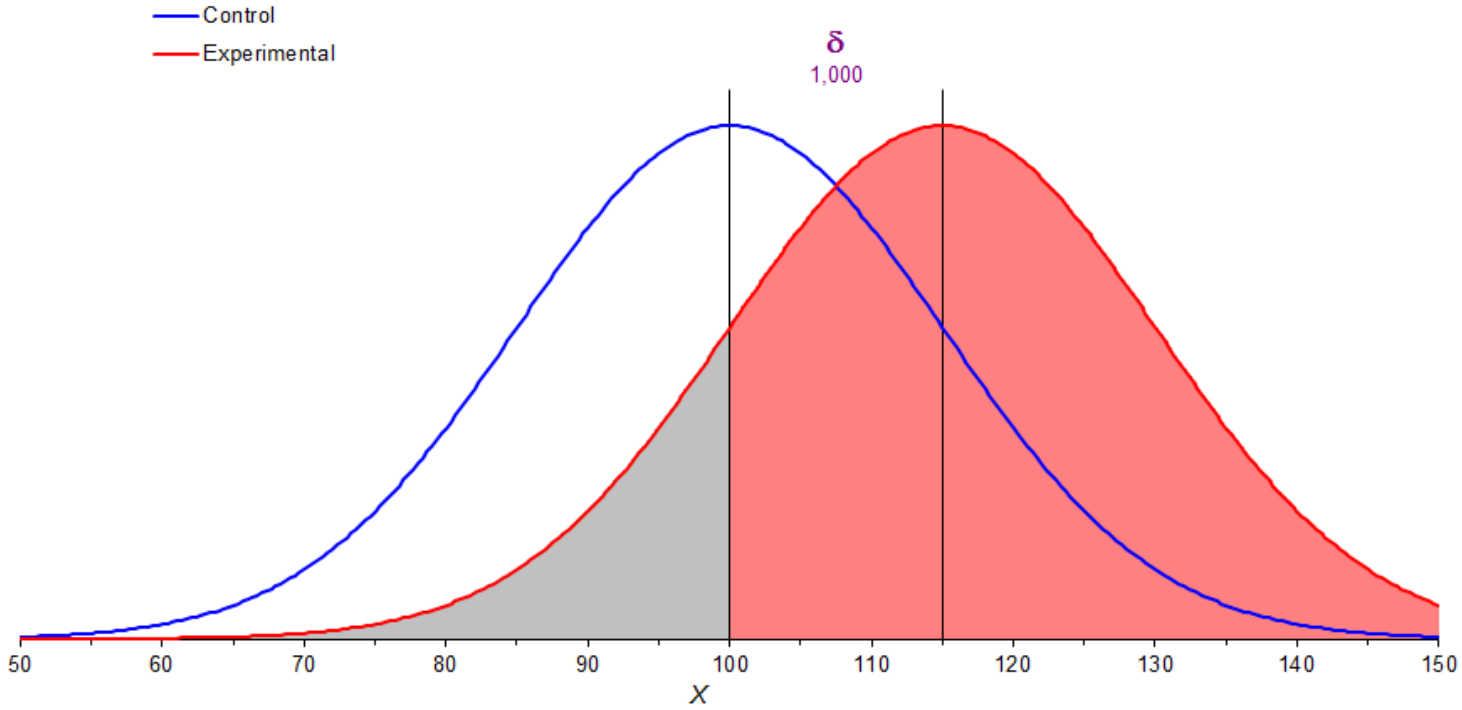
	forma Forma.email	N	Media	Deviazione std.	Errore std. Media
persuasività Giudizio. Persuasività	1 Tu	246	7,57	1,107	,071
	2 Lei	219	7,79	1,197	,081

Conventional values:

0.2 small
0.5 medium
0.8 large

Control
X=100
SD=15





Other Effect Size indicators

- Common in factorial designs:

$$\hat{\eta}^2 = \frac{SS_{\text{Effect}}}{SS_{\text{T}}}, \qquad \hat{\eta}_p^2 = \frac{SS_{\text{Effect}}}{SS_{\text{Effect}} + SS_{\text{s/Cells}}}, \qquad \hat{\omega}_p^2 = \frac{SS_{\text{Effect}} - df_{\text{Effect}} MS_{\text{s/Cells}}}{SS_{\text{Effect}} + (N - df_{\text{Effect}})\, MS_{\text{s/Cells}}}$$

- Some bibliografic references:

- Fritz, C. O., Morris, P. E., & Richler, J. J. (2012). Effect size estimates: Current use, calculations, and interpretation. Journal of Experimental Psychology: General, 141, 2–18.
- Olejinik & Algina (2003). Generalized eta and omega squared..., Psychological Methods
- Cohen (1992). A power primer, Psychological Bulletin
- Cohen (1994). The earth is round (ρ < .05). American Psychologist
- Cohen (1988). Statistical power analysis for the behavioral sciences

Errors of inference

- Frequentist approach
- There are three types of errors
- NHST*: Type I error (False positives)
  Type II error (False negatives)
- CI: Estimate error (imprecision)

NHST= Null Hypothesis Significance Testing (what you have been most likely taught as a student)
H0 vs. H1

Errors of inference in NHST

Real World (**POPULATION**)

Conclusion of the significance test (**SAMPLE**)	Null is true (H0 is correct)	Null is false (H1 is correct)
Null is true	Correct decision	Type II error
Null is false	Type I error	Correct decision

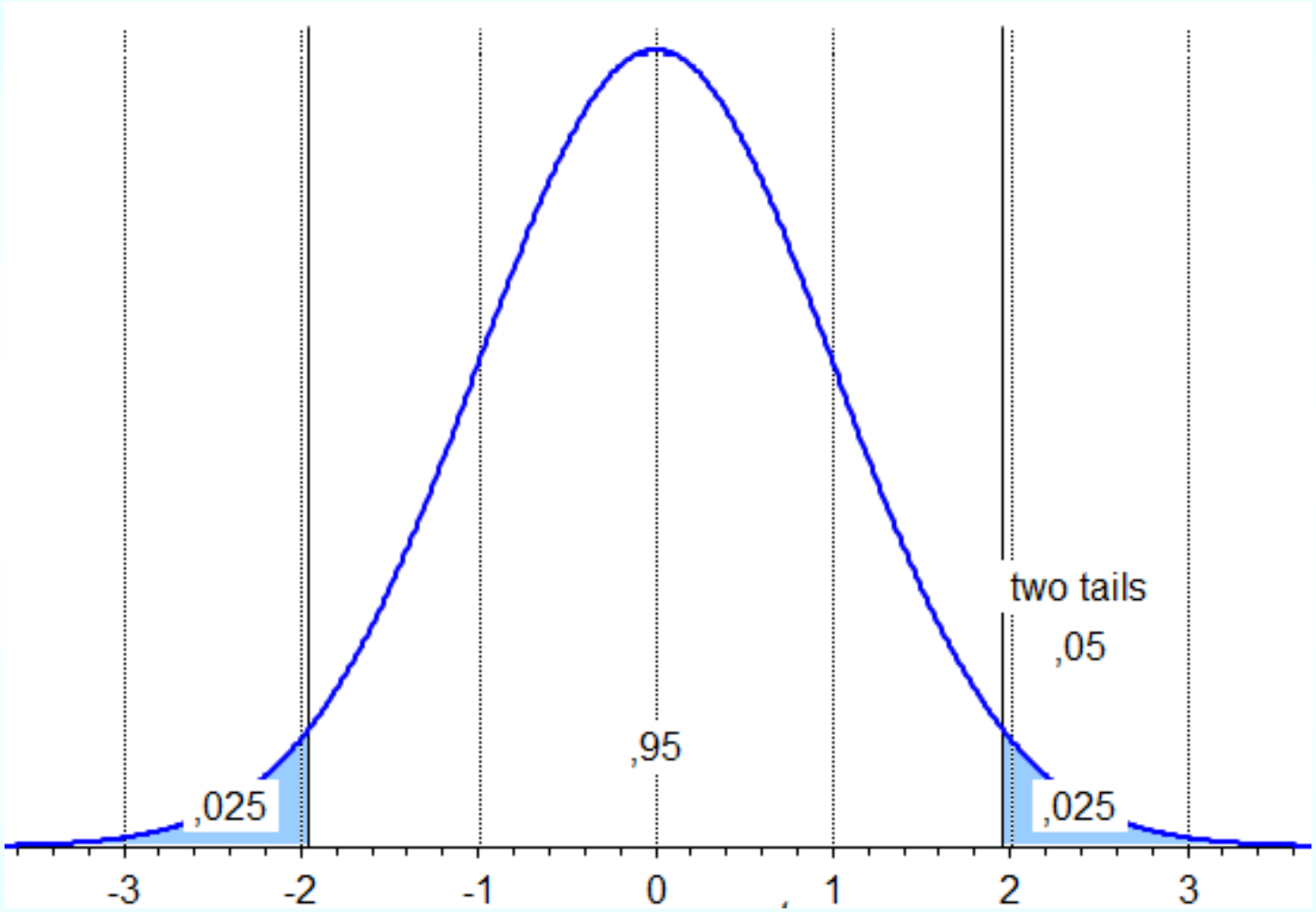
Errors of inference in NHST

- **Type I error**: ***Erroneously rejecting the null hypothesis (<u>False positive</u>)***.
  The result in the sample is significant ($p < .05$), so the null hypothesis is rejected, but the null hypothesis is actually true in the population.

- **Type II error**: ***Erroneously accepting the null hypothesis (<u>False negative</u>)***. The result in the sample is not significant ($p > .05$), so the null hypothesis is not rejected, but it is actually false in the population.

How to control Type I errors?

- The Type I error rate (***False positive***) is controlled by the researcher.
- It is called the **alpha rate**, and corresponds to the probability cut-off that one uses in a significance test.
- Conventionally, researchers use an alpha rate (α) of .05. This means that the null hypothesis is rejected when a value such as the one found is likely to occur 5% of the time or less when the null hypothesis is true.
- The test can be two-tailed (more common) or one-tailed (directional)


two tails
,05
,95
,025
,025
-3
-2
-1
0
1
2
3

How to control Type II errors?

- The Type II error (***False negative***) can also be controlled by the experimenter.
- The Type II error rate is called **beta** (β) as a complement to alpha.
- How can the beta rate be controlled? The easiest way to control Type II errors is by increase the **statistical power** of a test.
- **Statistical power**= probability of finding an effect, if it exists
- **Power** = **1 – β**
- Conventionally a power of at least .80 (**β=.20**) is considered as acceptable

What affects power?

- Power goes up with larger effect sizes and sample sizes, given a certain decision criterion (e.g., α=.05)

- When effect sizes become larger? When the portion of variability (difference) ascribed to the effect of interest grows more than the general (non specific) variability

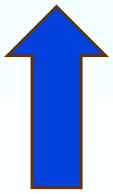
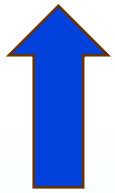
$$d = \frac{M_1 - M_2}{pooled\ SD}$$

$$\hat{\eta}^2 = \frac{SS_{\text{Effect}}}{SS_{\text{T}}}$$

$$r(v,x) = \frac{\text{cov}(v,x)}{sd(v) * sd(x)}$$

How to increase power?

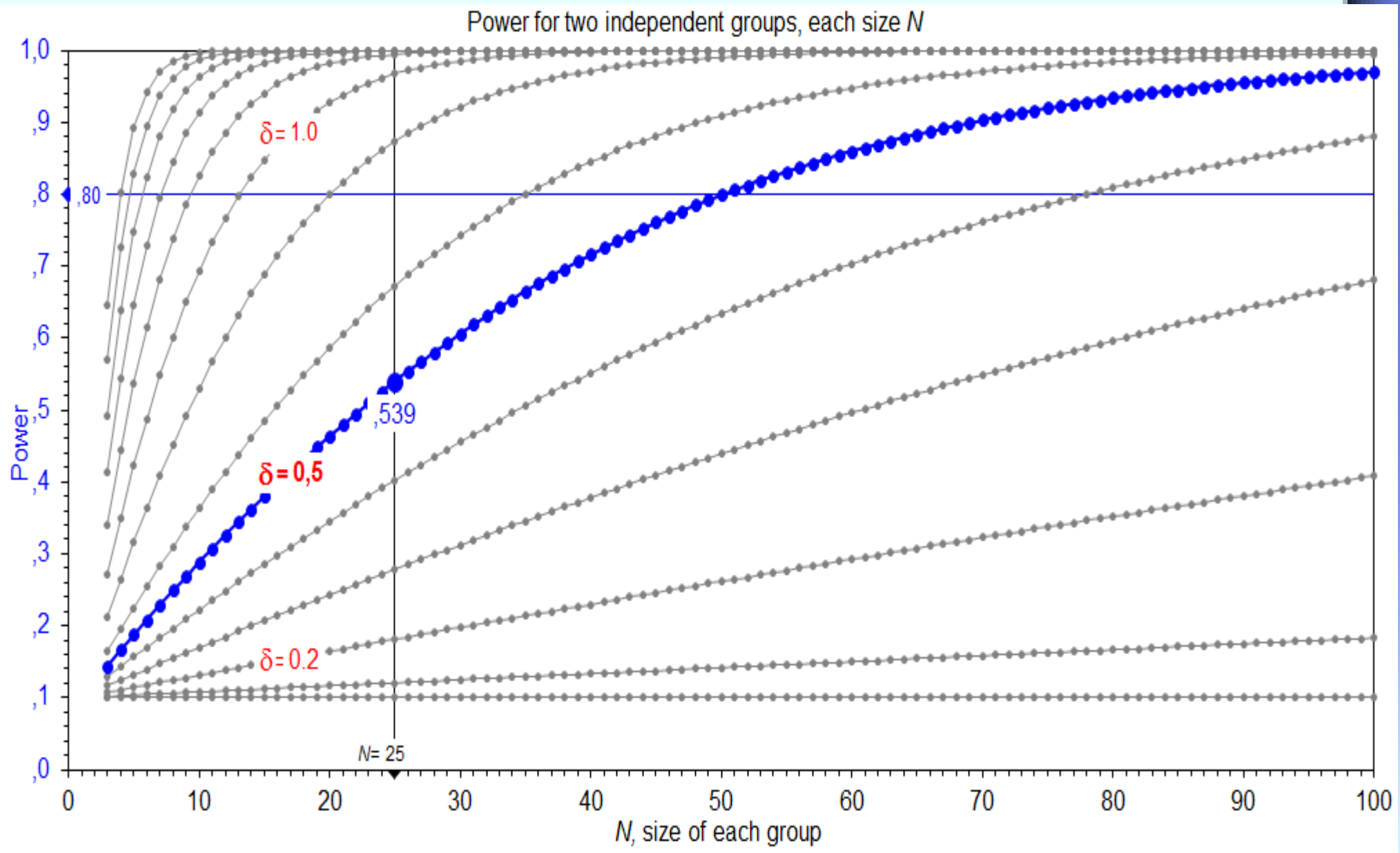
Power is affected by

- ***Sample size*** 
- ***Construct-related (i.e., signal) variance*** 
- ***Construct-unrelated (i.e., noise) variance*** 

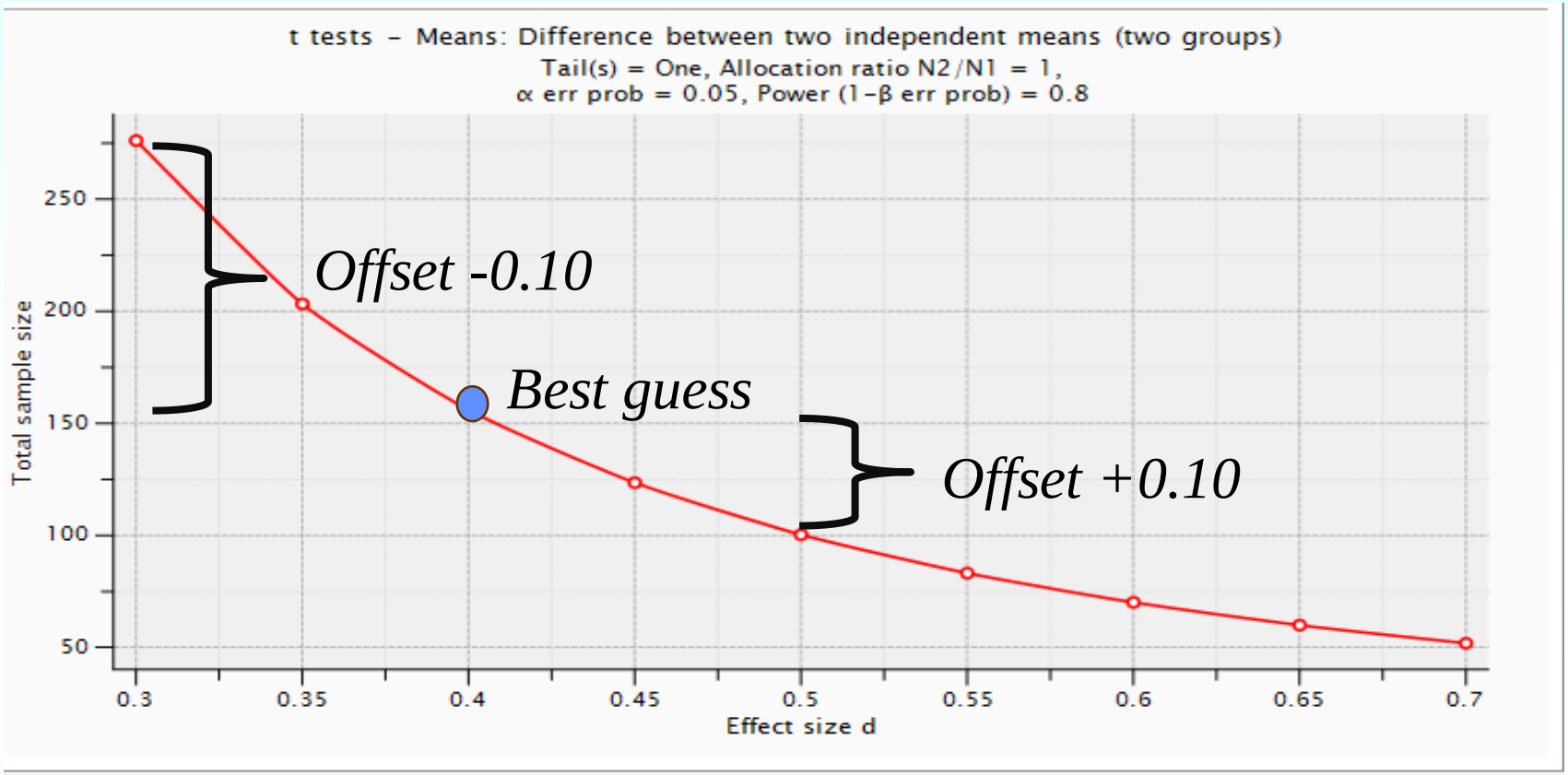
How to increase power?

- **Increase sample size**
- Administer stronger treatments (e.g., experimental manipulation) BUT be wary of possible reduced ecological validity
- Avoid restrictions of range for dependent variables
- Standardize experimental procedures
- Increase reliability of measures
- Use more homogenous subject samples BUT increased risks to generalizability of results
- Use blocking or repeated measures (within) design BUT sometimes can be inappropriate
- Meta-analytical mindset

Power Between Ss



Power Between Ss



Plot Parameters

Plot (on y axis) Total sample size ☑ with markers ☐ and displaying the values in the plot

s a function of Effect size d from 0.3 in steps of 0.05 through to 0.7

lot 1 graph(s) interpolating points

with Power (1−β err prob) at 0.80

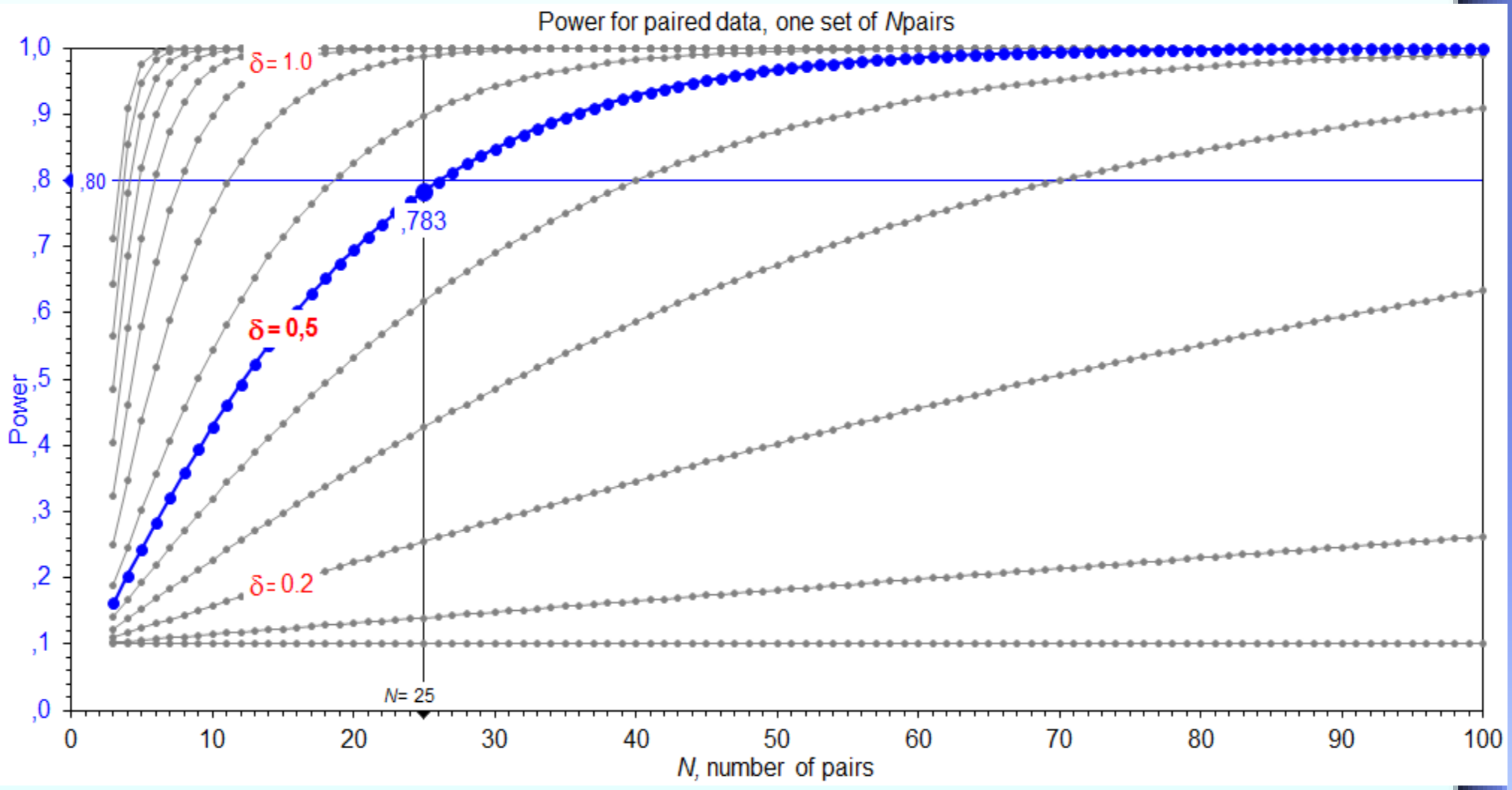
and α err prob at 0.05

Draw plot

Gpower, http://www.gpower.hhu.de/en.html

Power Within Ss

- Power for within Ss studies is greater (*ceteris paribus*) but depends on r (e.g., r = .50) between DVs



Bottom line

- Either you expect a large effect size or you need a substantial sample size
- The average ES in Psychology is at best **d=0.5**
- You can have a large sample (e.g., N=274) to reliably detect a small effect size (e.g., d=0.30)
- You can have a small sample (e.g., N=26) to reliably detect a very large effect size (e.g., d=1.0)
- But you **cannot** have a small sample size to detect a subtle effect (small effect size)
- Within Ss studies can be more powerful

Errors of inference

- Frequentist approach
- There are three types of errors
- NHST: Type I error (False positives)
  Type II error (False negatives)
- CI: Estimate error (imprecision)
  (e.g., AIPE, Maxwell, ARP 2008)

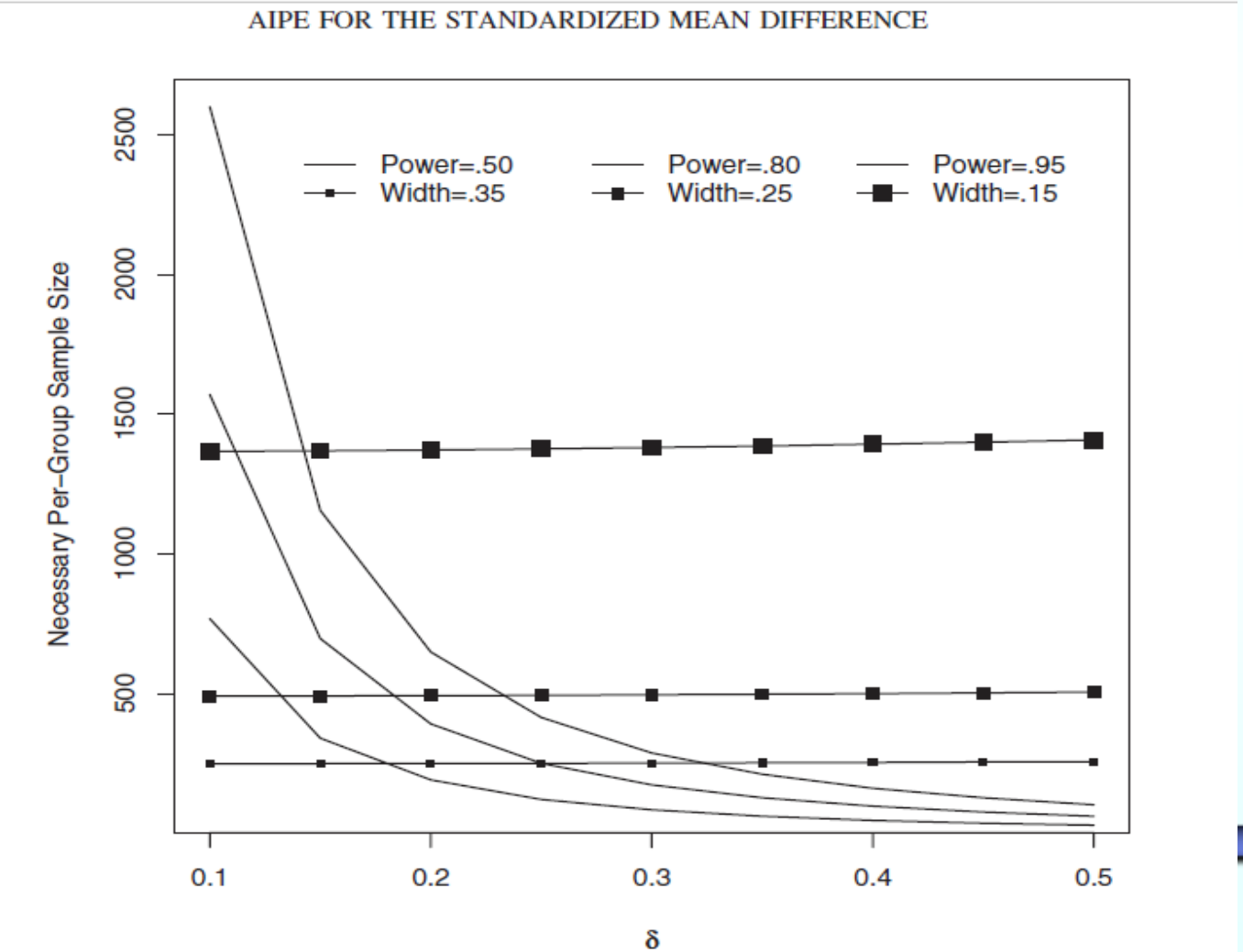
CI error

- In a CI approach, one thing matters a lot: sample size, the bigger, the better (*ceteris paribus*)
- The point is not whether some effect exists (or not) but how precise is our estimate of it
- All effects exist given an infinite sample size (Cohen)

- *If you want to get it right, increase sample size*

A practical problem

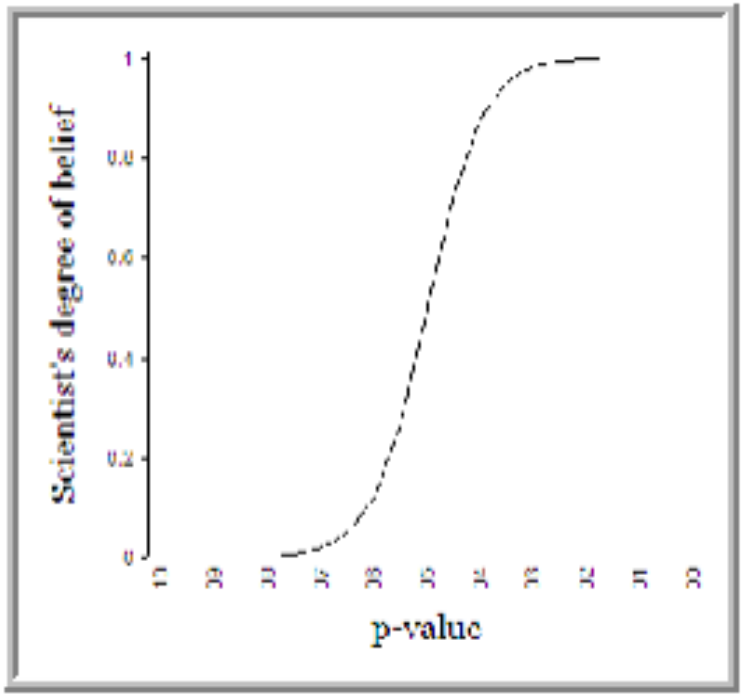
- **Big** sample sizes are needed for precise estimates no matter the effect size



Why NHST can be problematic

- But NHST dominates
- Publications, theories, discussions, fights and careers gravitate around p<.05

"We are not interested in the logic itself, nor will we argue for replacing the .05 alpha with another level of alpha, but at this point in our discussion we only wish to emphasize that dichotomous significance testing has no ontological basis. That is, we want to underscore that, surely, God loves the .06 nearly as much as the .05. Can there be any doubt that God views the strength of evidence for or against the null as a fairly continuous function of the magnitude of p?"



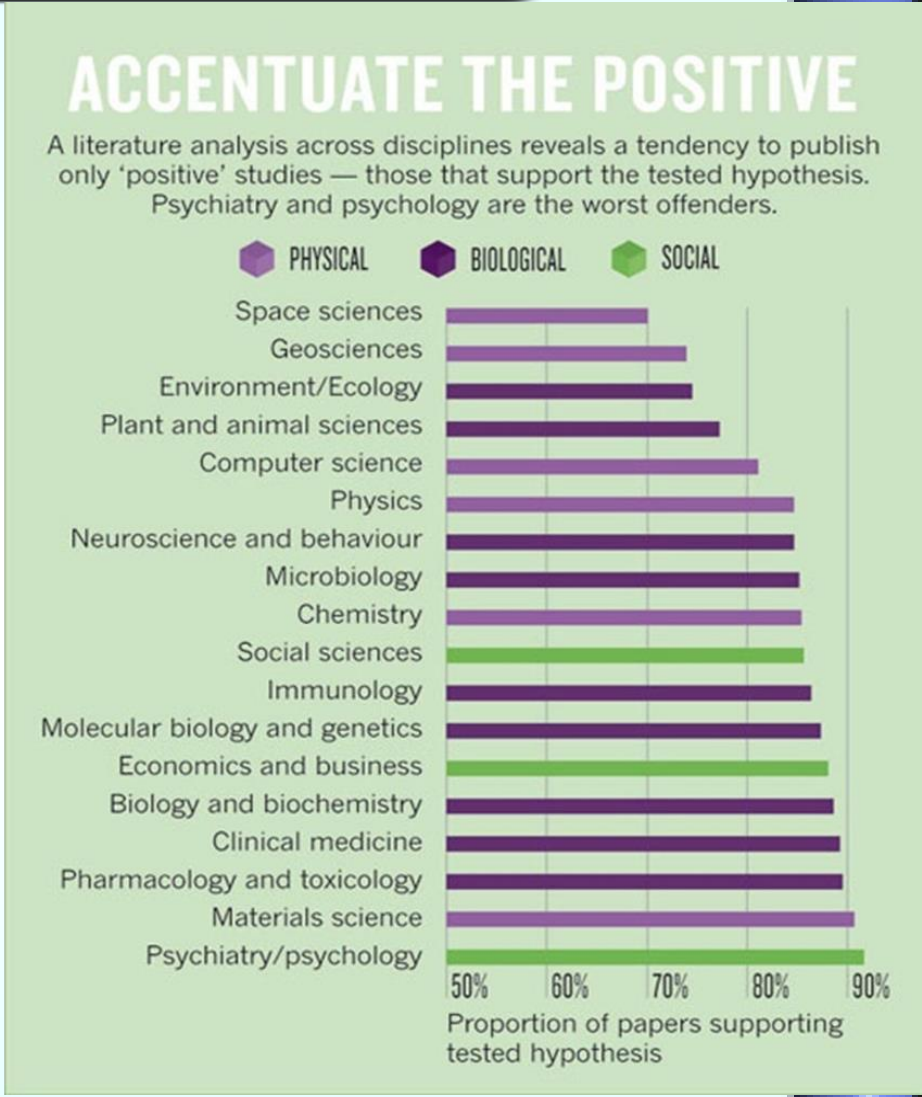
Rosnow, R.L. & Rosenthal, R. (1989). Statistical procedures and the justification of knowledge in psychological science. American Psychologist, 44, 1276-1284.

Back to the present

- Psychology is now at the forefront of Science in addressing the replicability issue
- Methodological innovations
- Massive effort to estimate replicability
- The Reproducibility Project (OSC)
- 270 volunteers, 64 universities, 11 countries, 100 replicated studies published in three main journals (JPSP, PS, JEP:LMC) in 2008

Why is it needed?

- Average ES: Cohen's d=0.50
- Average Sample size: n= 40
- Typical power: (1-β)=.35
- This means around 1/3 chance of positive findings
- The literature should be full of non-significant findings
- Really?



Bakker, van Dijk, & Wicherts, 2012; Asendorpf et al., 2013

Is the problem unique to Psychology?

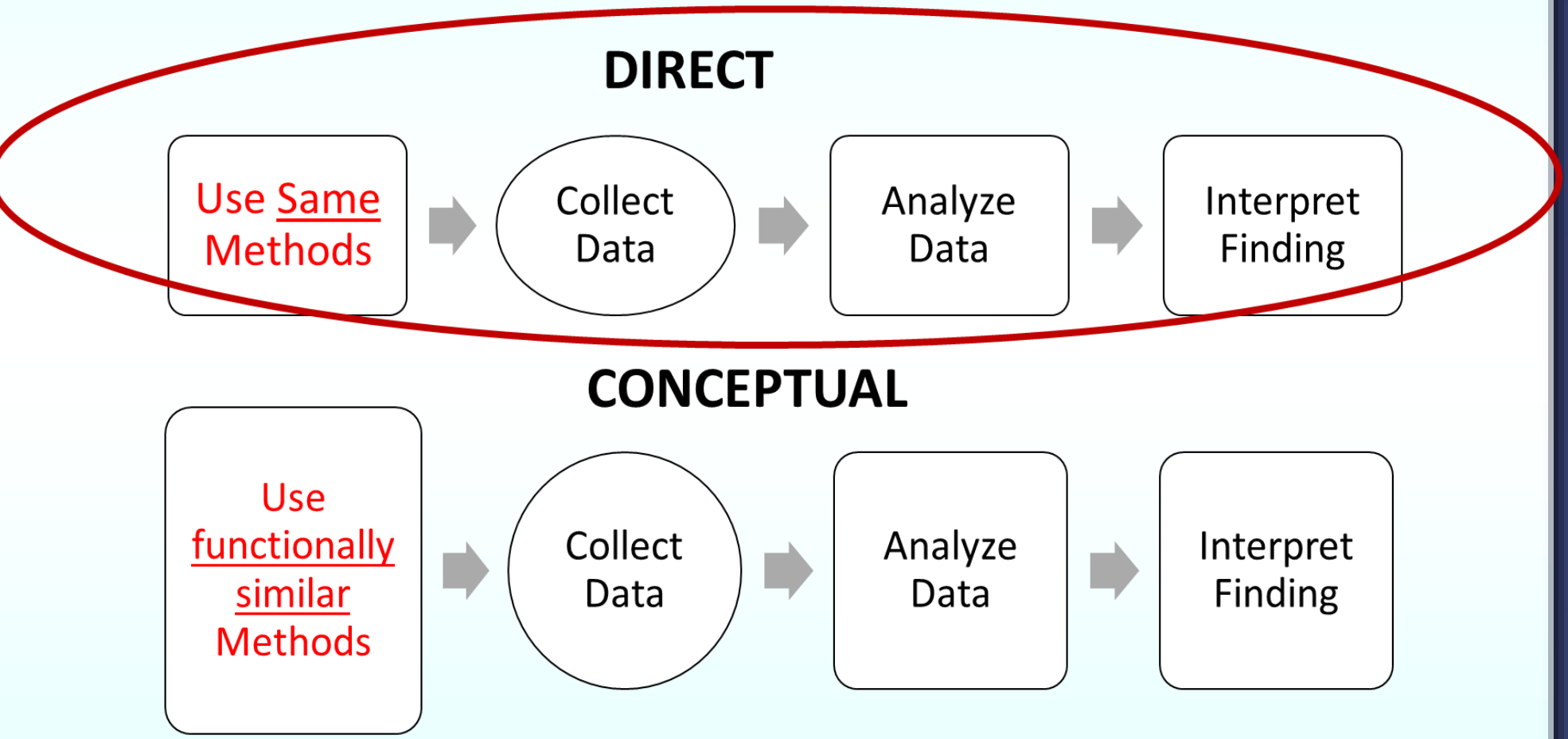
- **NO !!** (Ioannidis, 2005)

- Average power in **Neuroscience**: .21 (Button et al., 2013)
  This means around 1/5 chance of positive findings

- **Cancer Biology**: Replication rate of main results from pre-clinical trails (Begley & Ellis, 2012): from 11% (Amgen, 2011) to 25% (Prinz et al. – Bayer HC team, 2011)

The Reproducibility Project

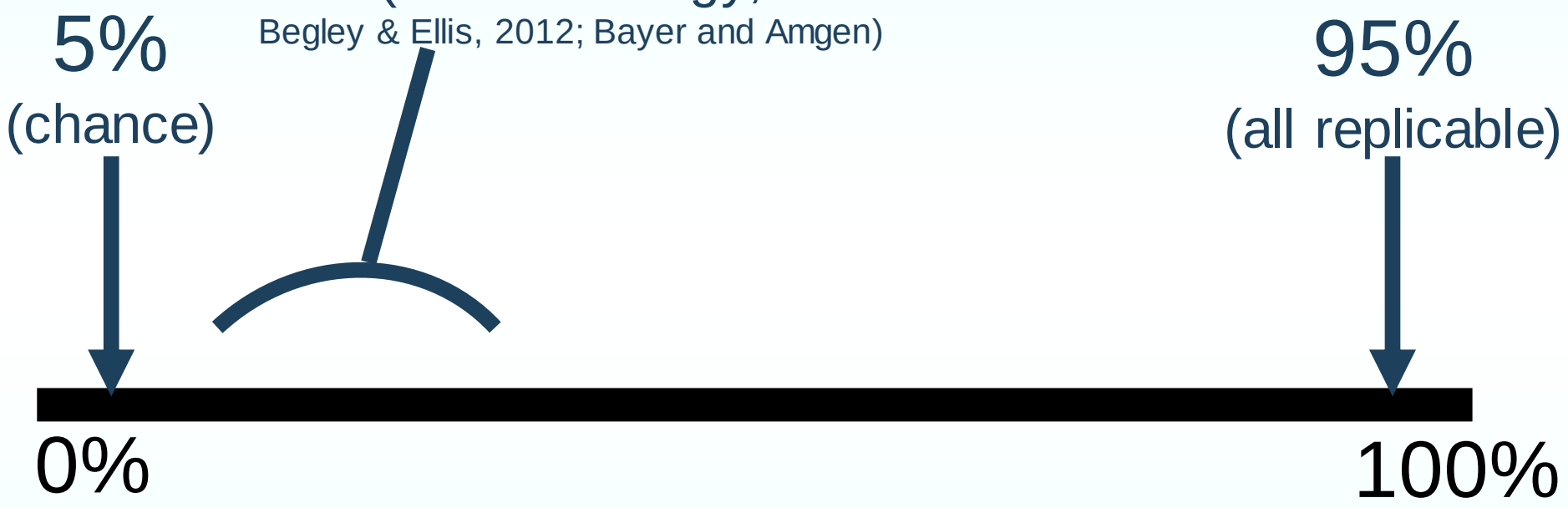
Direct vs. Conceptual Replication



What one can expect?

11-25% (89-75% did not replicate)

(cancer biology;
Begley & Ellis, 2012; Bayer and Amgen)



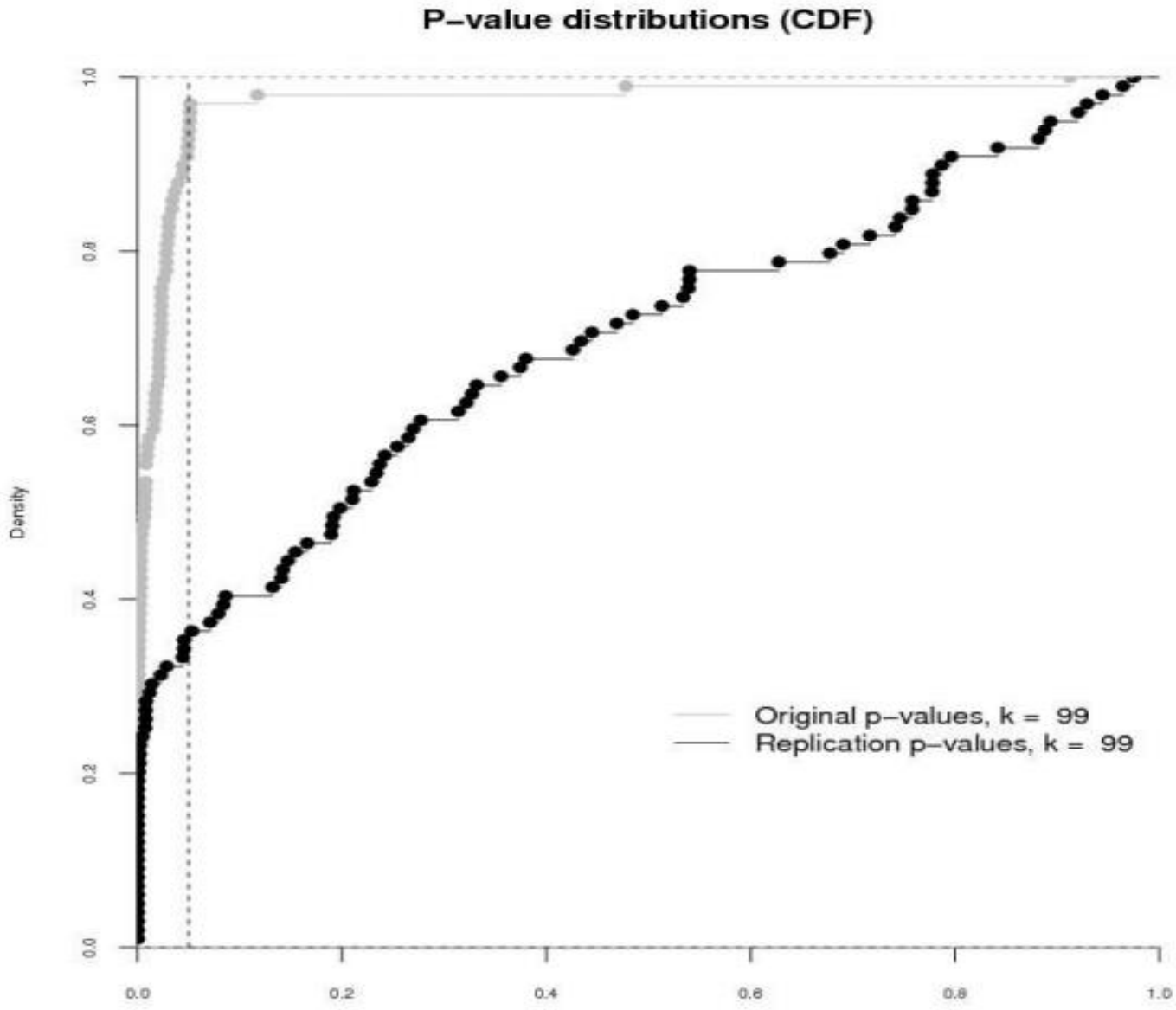
What counts as a replication?

1. Effect itself against null (p-values)
2. Effect size comparison
3. Meta-analytic precision estimate

Open Science Collaboration (2015).
Estimating the reproducibility of psychological science.
Science, 349(6251). DOI: 10.1126/science.aac4716

Effects against null I

Figure S1: Cumulative *P* value distributions of original and replication studies.



Effects against null II

Table 1. Summary of reproducibility rates and effect sizes for original and replication studies overall and by journal/discipline. *df/N* refers to the information on which the test of the effect was based (for example, *df* of *t* test, denominator *df* of *F* test, sample size −3 of correlation, and sample size for *z* and χ^2). Four original results had *P* values slightly higher than 0.05 but were considered positive results in the original article and are treated that way here. Exclusions (explanation provided in supplementary materials, A3) are "replications $P < 0.05$" (3 original nulls excluded; $n = 97$ studies); "mean original and replication effect sizes" (3 excluded; $n = 97$ studies); "meta-analytic mean estimates" (27 excluded; $n = 73$ studies); "percent meta-analytic ($P < 0.05$)" (25 excluded; $n = 75$ studies); and, "percent original effect size within replication 95% CI" (5 excluded, $n = 95$ studies).

			Effect size comparison					Original and replication combined			
	Replications $P < 0.05$ in original direction	Percent	Mean (SD) original effect size	Median original *df/N*	Mean (SD) replication effect size	Median replication *df/N*	Average replication power	Meta-analytic mean (SD) estimate	Percent meta-analytic ($P < 0.05$)	Percent original effect size within replication 95% CI	Percent subjective "yes" to "Did it replicate?"
Overall	35/97	36	0.403 (0.188)	54	0.197 (0.257)	68	0.92	0.309 (0.223)	68	47	39
JPSP, social	7/31	23	0.29 (0.10)	73	0.07 (0.11)	120	0.91	0.138 (0.087)	43	34	25
JEP:LMC, cognitive	13/27	48	0.47 (0.18)	36.5	0.27 (0.24)	43	0.93	0.393 (0.209)	86	62	54
PSCI, social	7/24	29	0.39 (0.20)	76	0.21 (0.30)	122	0.92	0.286 (0.228)	58	40	32
PSCI, cognitive	8/15	53	0.53 (0.2)	23	0.29 (0.35)	21	0.94	0.464 (0.221)	92	60	53

Effect size comparison

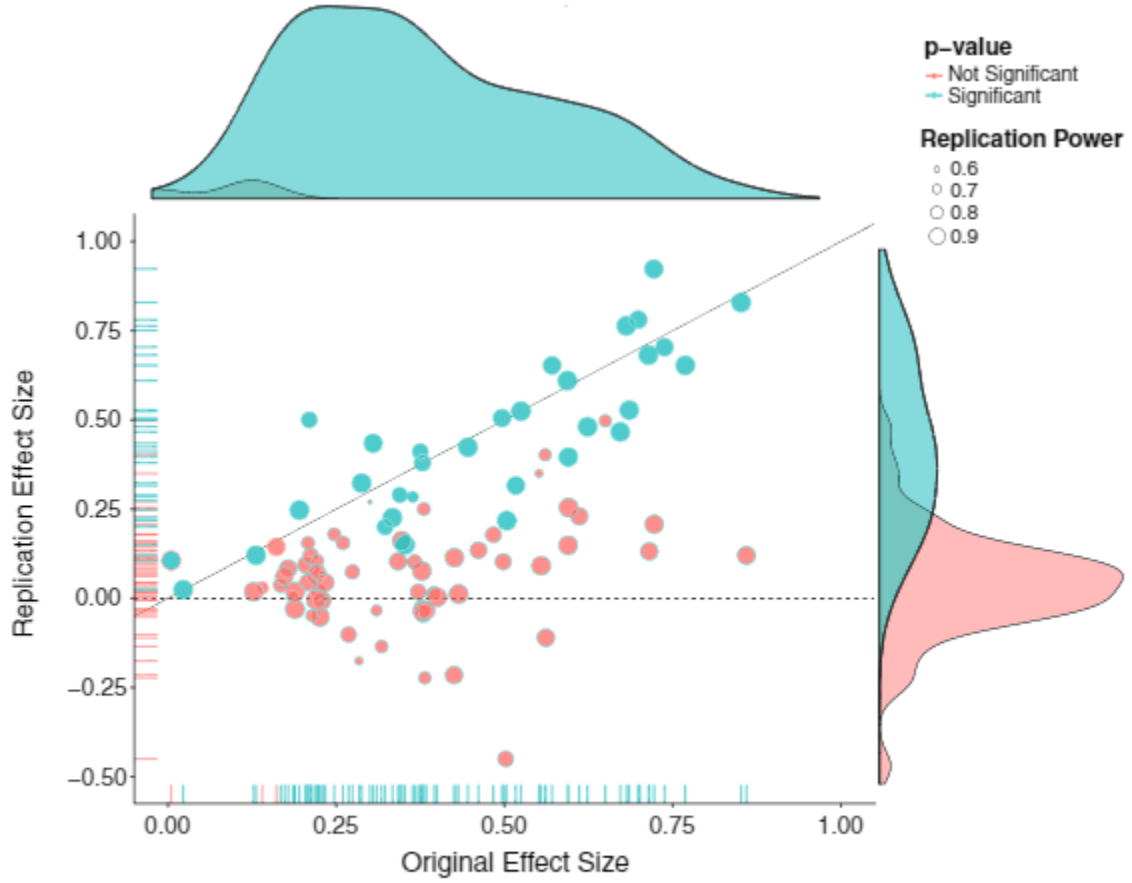


Fig. 3. Original study effect size versus replication effect size (correlation coefficients). Diagonal line represents replication effect size equal to original effect size. Dotted line represents replication effect size of 0. Points below the dotted line were effects in the opposite direction of the original. Density plots are separated by significant (blue) and nonsignificant (red) effects.

r(spearman) = 0.51

Most replications smaller effect size than original

Meta-analytic comparison

Table 1. Summary of reproducibility rates and effect sizes for original and replication studies overall and by journal/discipline. *df/N* refers to the information on which the test of the effect was based (for example, *df* of *t* test, denominator *df* of *F* test, sample size −3 of correlation, and sample size for *z* and χ^2). Four original results had *P* values slightly higher than 0.05 but were considered positive results in the original article and are treated that way here. Exclusions (explanation provided in supplementary materials, A3) are "replications $P < 0.05$" (3 original nulls excluded; $n = 97$ studies); "mean original and replication effect sizes" (3 excluded; $n = 97$ studies); "meta-analytic mean estimates" (27 excluded; $n = 73$ studies); "percent meta-analytic ($P < 0.05$)" (25 excluded; $n = 75$ studies); and, "percent original effect size within replication 95% CI" (5 excluded, $n = 95$ studies).

			Effect size comparison					Original and replication combined			
	Replications $P < 0.05$ in original direction	Percent	Mean (SD) original effect size	Median original *df/N*	Mean (SD) replication effect size	Median replication *df/N*	Average replication power	Meta-analytic mean (SD) estimate	Percent meta-analytic ($P < 0.05$)	Percent original effect size within replication 95% CI	Percent subjective "yes" to "Did it replicate?"
Overall	35/97	36	0.403 (0.188)	54	0.197 (0.257)	68	0.92	0.309 (0.223)	68	47	39
JPSP, social	7/31	23	0.29 (0.10)	73	0.07 (0.11)	120	0.91	0.138 (0.087)	43	34	25
JEP:LMC, cognitive	13/27	48	0.47 (0.18)	36.5	0.27 (0.24)	43	0.93	0.393 (0.209)	86	62	54
PSCI, social	7/24	29	0.39 (0.20)	76	0.21 (0.30)	122	0.92	0.286 (0.228)	58	40	32
PSCI, cognitive	8/15	53	0.53 (0.2)	23	0.29 (0.35)	21	0.94	0.464 (0.221)	92	60	53

Summing up RP main results

- 36% replicate at p<.05 (simple answer)
- Effect size are half (publication bias, file drawer effect)
- Less likely to replicate if:
  a) weaker evidence in original study (p<.05 worse than p<.001)
  b) results considered to be "surprising" (*sexy but unreliable findings*)
- Milestone achievement of Psychology
- RP in Cancer Biology has just started

Back to the future

- As scientists, we all want to get something right
- If we get it right, it is replicable and will be replicated
- But what does it mean “to get it right”?
- So, what can we do to increase our chances?

First pointer

- Design your study with adequate power
- It is ok to run initial exploratory/pilot studies (N needs not be too small) to identify an effect and to have a rough estimate of its effect size (**don't trust it too much…**)
- If you find something, then you need to plan a confirmatory study with adequate power to confirm your effect
- Be careful on overestimations of effect size (*Winner's curse*)

Winner's curse I

- Inflated effect size from initial study are likely and can affect later confirmatory studies (Ioannides, 2008).
- More likely with asymmetries in publication standards (publish only significant results) and underpowered studies (difficult to establish a priori)
- Under these common conditions, true effect size can be much lower than published effect size
- d=0.80 (n=50), CI [0.22, 1.37]
- N for power 80% can be from 516 to 16, with n=42 for d=0.80
  (cf. Safeguard power analysis – PGC, 2014)

Winner's curse II

- This is true also for your own pilot studies, because you will be likely to pursue significant findings and abandon not significant ones (your own "publication bias")

Contents lists available at ScienceDirect

Journal of Experimental Social Psychology

journal homepage: www.elsevier.com/locate/jesp

Exploring Small, Confirming Big: An alternative system to The New Statistics for advancing cumulative and replicable psychological research☆

John Kitchener Sakaluk *

Department of Psychology, University of Toronto Mississauga, Canada

- It is not good advice
- Better advice "explore reliably, confirm more reliably"

aps ASSOCIATION FOR PSYCHOLOGICAL SCIENCE

Safeguard Power as a Protection Against Imprecise Power Estimates

Perspectives on Psychological Science
2014, Vol. 9(3) 319–332

DOI: 10.1177/1745691614528519
pps.sagepub.com
SAGE

Marco Perugini, Marcello Gallucci, and Giulio Costantini
University of Milan-Bicocca, Italy

1. Calculate the effect size in the metric available or preferred for the study (e.g., d_o)
2. Calculate the two-tail confidence interval (e.g., 60%) of this effect size measure.
3. Consider the lower boundary of the confidence interval (e.g., 20th percentile)[5].
4. Use this as the estimate of the true effect size (d_s).
5. Calculate the needed sample size at a chosen power level (e.g., $P = .80$) for a given decision level (e.g., $\alpha = .05$).

Table 3. Application of Safeguard Power Analysis to Some Hypothetical Studies

Study	d_0	N_0	d_{s80}	N_p	N_{s80}	$N_{2.5}$	SSR_{80}
A	0.70	50	0.45	52	124	125	2.48
B	0.70	200	0.58	52	76	500	0.38
C	1.0	30	0.66	28	58	75	1.93
D	1.5	20	1.05	14	24	50	1.20
E	0.4	80	0.21	156	570	200	7.12

Note: d_0 = effect size of the original study; N_0 = sample size of the original study; d_{s80} = effect size considered for safeguard power analysis at an 80% level of protection; N_p = sample size computed using standard power analysis; N_{s80} = sample size considered for safeguard power analysis at an 80% level of protection; $N_{2.5}$ = sample size considered for the 2.5 rule (i.e., sample times 2.5); SSR_{80} = safeguard sample ratio for safeguard power analysis at an 80% level of protection.

Second pointer

- Distinguish between exploratory and confirmatory studies
- If you find an “unexpected effect”, confirm it with another well powered study before building on it
- Results can be significant simply out of random sampling

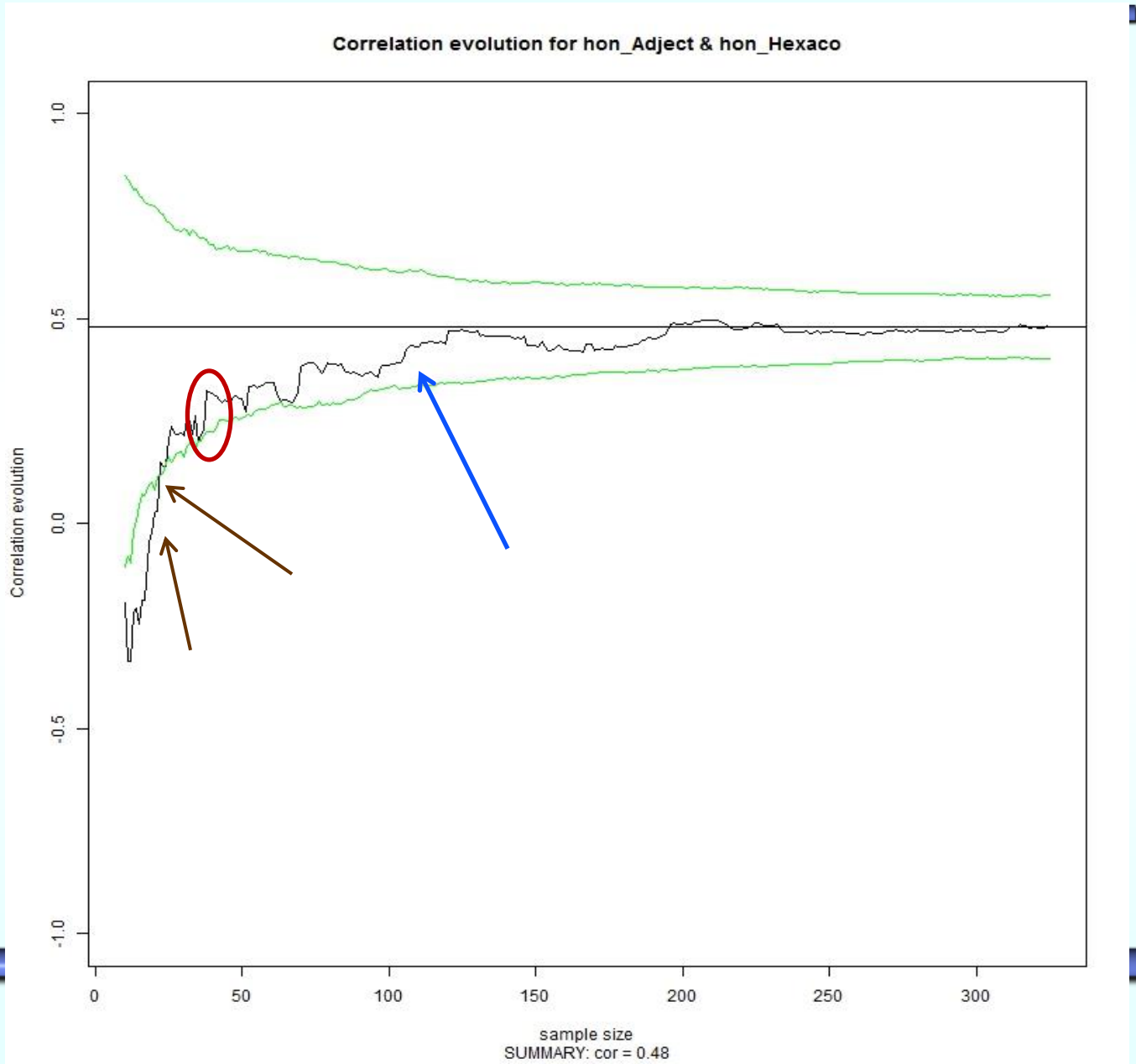
Third pointer

- Results stabilize with bigger sample sizes
- Try to have a decent sample size
- Sometimes results can be significant in opposite directions given small sample sizes
- For example, stability of correlation coefficients (cf. Schonbrodt & Perugini, 2013)

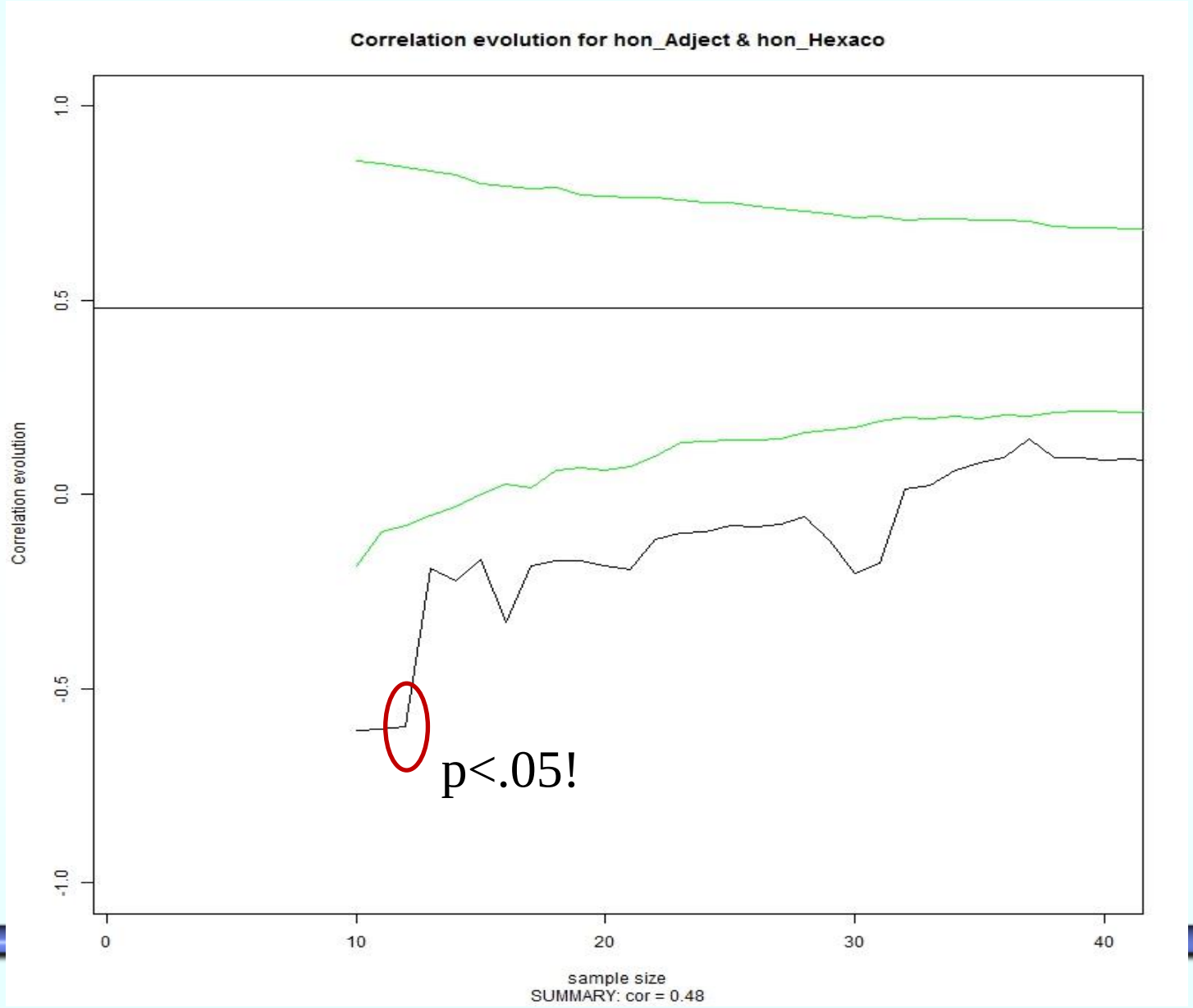
Stability of correlations

- Examples from my own research
- S1: one effect that exists in the full sample (r between H/H_{quest} and H/H_{adjec}=.48) and one that does not (r between H/H_{quest} and Ext_{quest}=-.05)
- Correlations calculated adding Ss at each step starting from N=10 to full sample (evolution of r)
- Real Ss order
- Boostrapped (s=1000) CI 95%

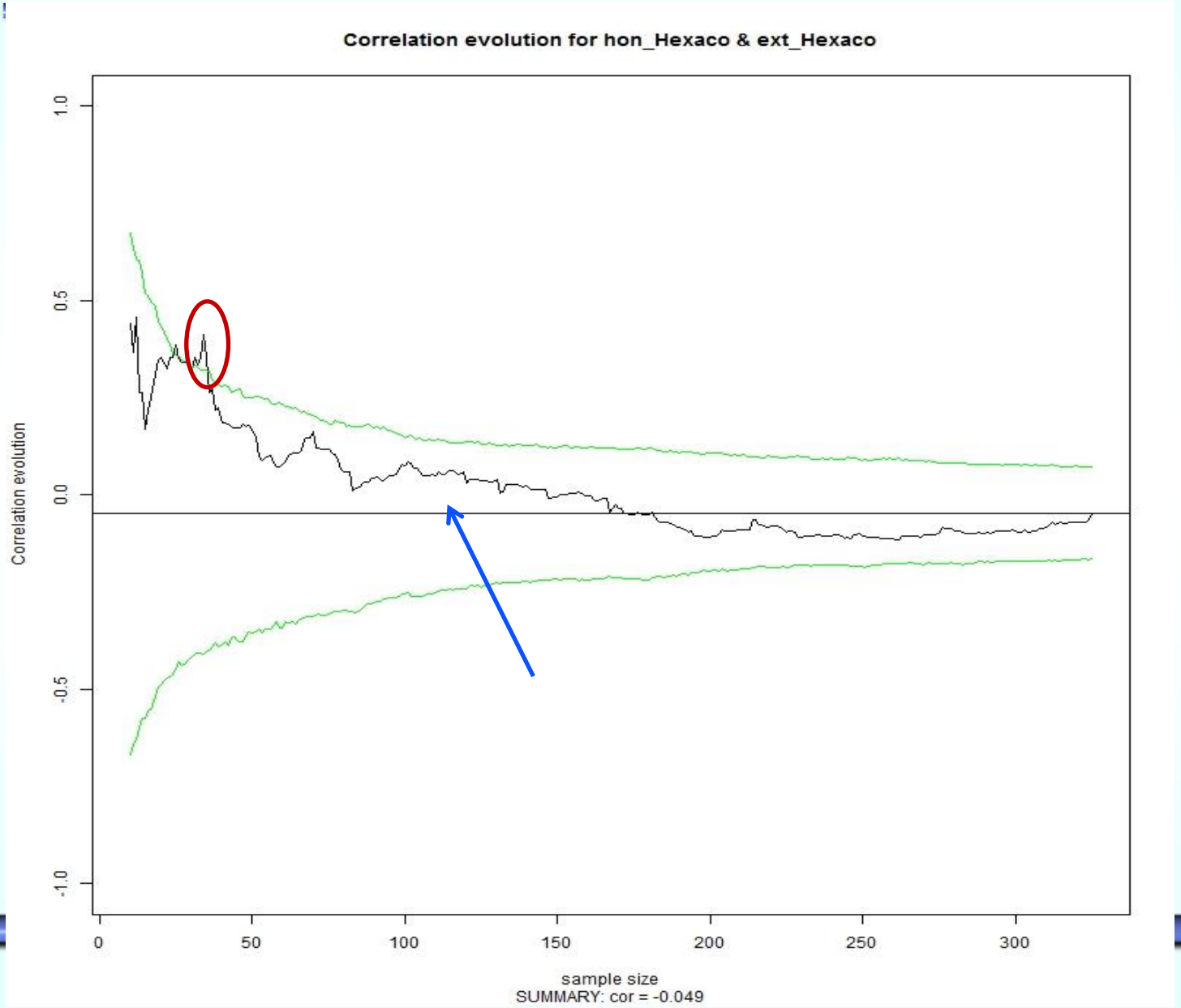
H vs. H


Correlation evolution for hon_Adject & hon_Hexaco
Correlation evolution
1.0
0.5
0.0
-0.5
-1.0
0
50
100
150
200
250
300
sample size
SUMMARY: cor = 0.48

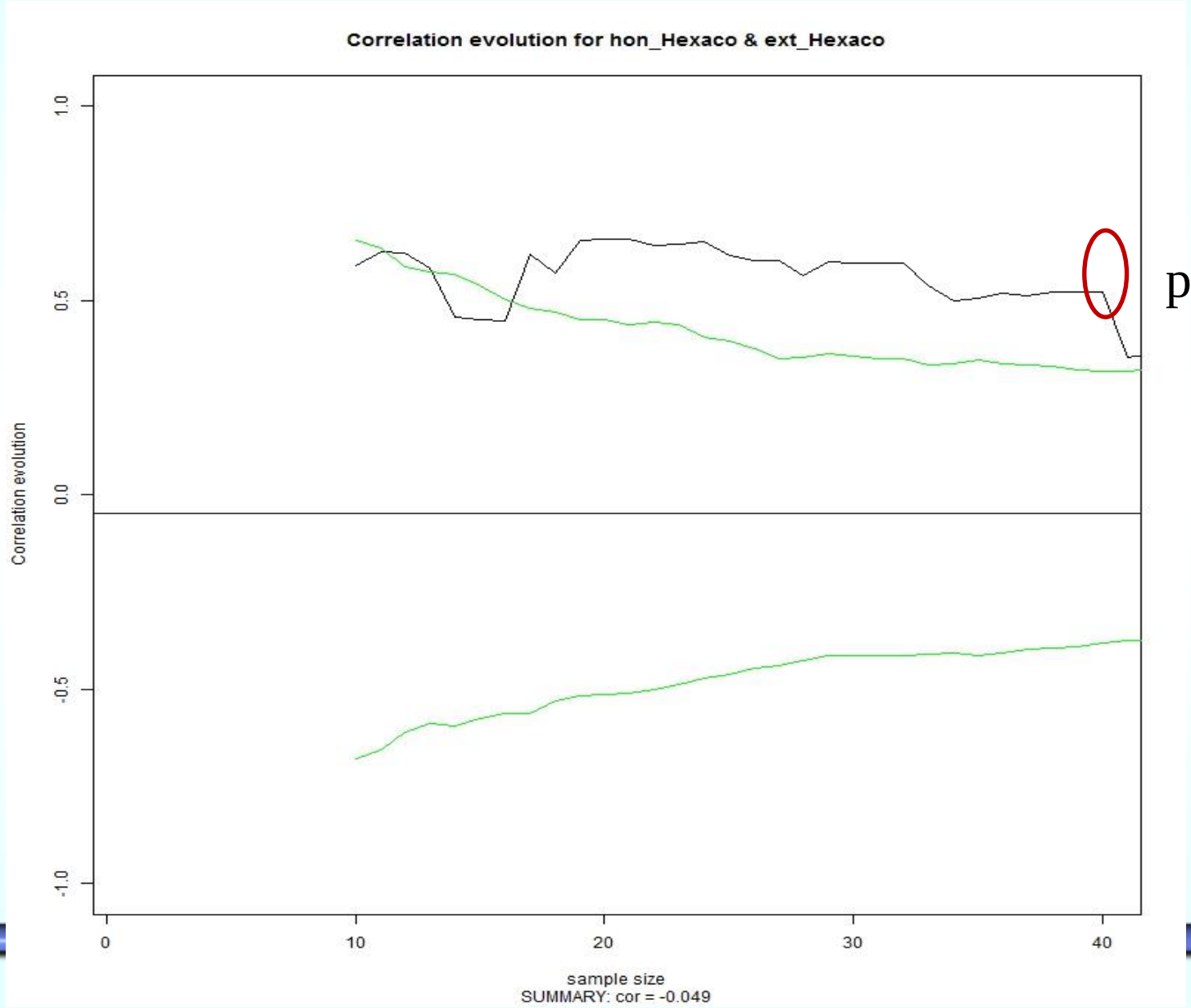
H vs. H


Correlation evolution for hon_Adject & hon_Hexaco
Correlation evolution
1.0
0.5
0.0
-0.5
-1.0
0
10
20
30
40
sample size
SUMMARY: cor = 0.48
p<.05!

The results: H vs. E



H vs. E


Correlation evolution for hon_Hexaco & ext_Hexaco
Correlation evolution
1.0
0.5
0.0
-0.5
-1.0
0
10
20
30
40
sample size
SUMMARY: cor = -0.049

p<.001!

Contents lists available at SciVerse ScienceDirect

Journal of Research in Personality

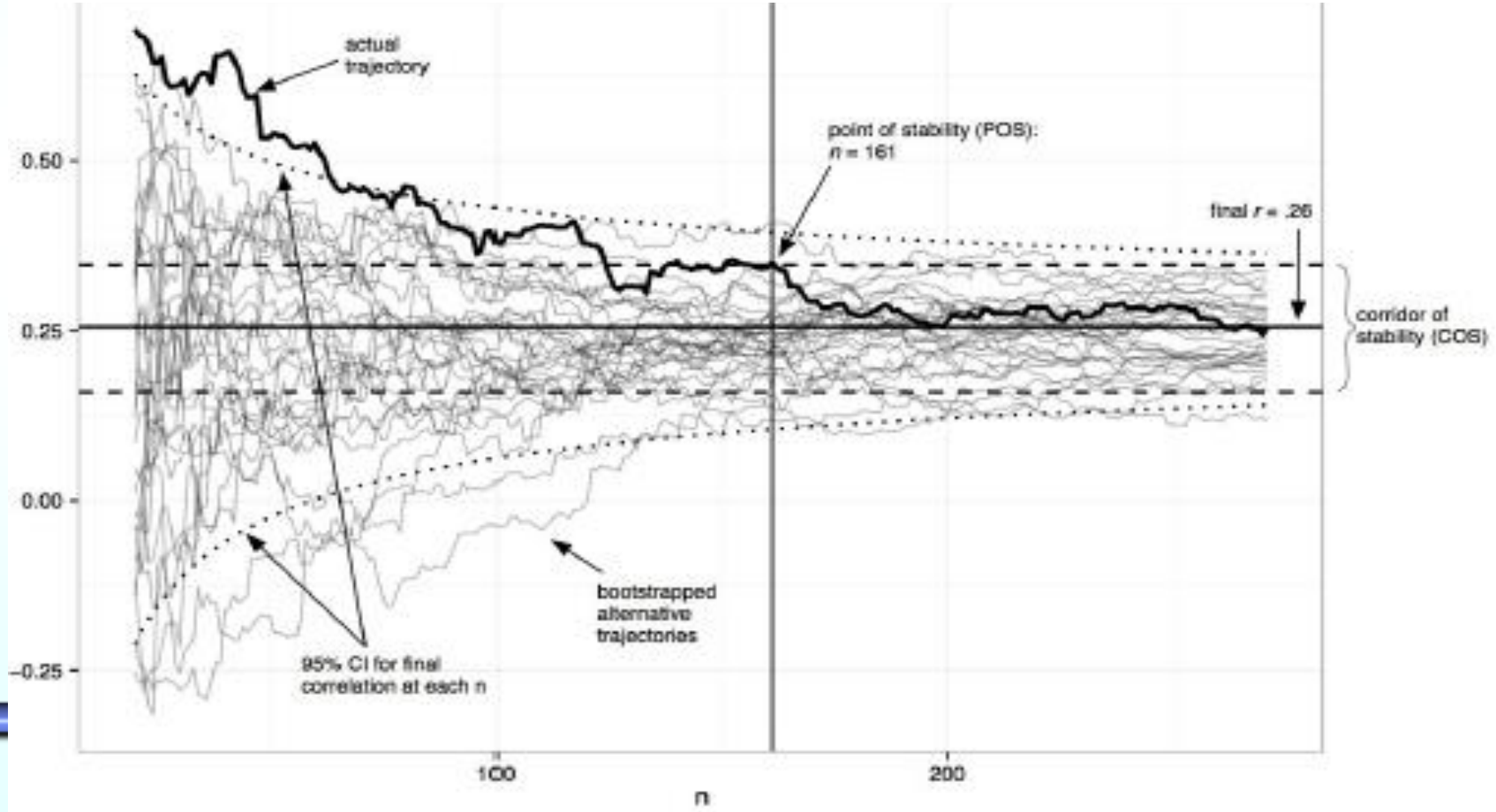
journal homepage: www.elsevier.com/locate/jrp

Brief Report

At what sample size do correlations stabilize?

Felix D. Schönbrodt [a,*], Marco Perugini [b]

[a] Department of Psychology, Ludwig-Maximilians-Universität, Leopoldstr. 13, 80802 München, Germany
[b] Department of Psychology, University of Milan, Bicocca, Piazza dell'Ateneo Nuovo 1 (U6), 20126 Milan, Italy



Implications

- Sequential effects can be devastating for small samples ($N \leq 60$)

- Estimates start to stabilize for $N \geq 100$
  (but it strictly depends on the expected correlation;
  e.g., $N \approx 180$ for $r = .4$)

- Small samples ($N \leq 60$) can give many false positives/negatives, especially for small effects (but see Sequential testing with Bayes Factors, *Schönbrodt, Wagenmakers, Zehetleitner, & Perugini, in press, PM*)

Practical implications

$$d = \frac{2r}{\sqrt{1-r^2}}$$

Lakens, D., & Evers, E. R. K. (2014). Sailing from the seas of chaos into the corridor of stability: Practical recommendations to increase the informational value of studies. *Perspectives in Psychological Science, 9*, 278–292. doi:10.1177/1745691614528520

In columns 3 and 4 of Table 1 on p. 280, the recommended sample sizes are for the total sample, not for the sample size per condition. The corrected table appears below.

Table 1. Recommended Sample Size Per Condition When Comparing Two Independent Groups Based on Different Effect Sizes (r and Cohen's d_{pop}) to Achieve the Point of Stability (POS) With 80% Confidence and Corridor Widths of .2 and .1 (See Part 1), to Achieve 80% or 90% Power to Observe the Effect With an Alpha of .05, and to Achieve a v Statistic Higher Than .5 (See Part 3)

r	d_{pop}	POS, 80% w = .2	POS, 80% w = .1	80% power	90% power	v > .5
.1	0.20	31	126	394	527	404
.2	0.41	29	119	95	126	99
.3	0.63	26	106	41	54	43
.4	0.87	22	91	22	29	23
.5	1.15	17	72	13	17	14
.6	1.50	13	52	9	11	9
.7	1.96	10	33	6	7	6

Fourth pointer

- Results stabilize with smaller standard errors
- Standard errors depend on N and SD
- Smaller SD means smaller SE
- The SD can be reduced (*ceteris paribus*) with more reliable measures, more precise experimental design, less within Ss variability
- Plan your design as simple and as clean as possible

Fifth pointer

- Be very careful with Questionable Research Practices (QRP)
- Do not cherry-pick DVs among many that you have
- Do not exclude cases as is
- Do not make multiple interim analyses to decide whether to collect additional Ss
- Read Simmons et al. (2011): not every recommendation is perfect, but they do give many good ones

Sixth pointer

- Some of this stuff is already implemented in top level journals
- For example, Psychological Science

DISCLOSURE QUESTIONS:

For all studies in your recently published article titled [publication title], please endorse the following statements: (please type an X to indicate your answer)

We reported the total number of observations which were excluded (if any) and the criterion for doing so. (If no observations excluded, please indicate Yes)
Yes: ___ No: ___
If no, please report this information here (e.g., data from 3 participants in Study 2 excluded due to computer malfunction; 4 participants in Study 1 excluded for not following instructions):

We reported all tested experimental conditions, including failed manipulations. Yes: ___ No: ___
If no, please provide brief explanation for not reporting this information (e.g., critical software implementation error; editorial request):

We reported all administered measures/items. Yes: ___ No: ___
If no, please provide brief explanation for not reporting this information (e.g., measures not related to research question; scores from unreported measure insufficiently reliable):

We reported (a) how we determined our sample size **and** (b) our data collection stopping rule. Yes: ___ No: ___
If no, please describe (a) the basis for the sample sizes used and (b) how you decided to stop collecting data (e.g., decided ahead of time to collect data until minimum sample size achieved and this was followed; sample size determined by power analysis but did not achieve it by the end of term):

Current scenario

a) Average effect size in published studies are small to medium ($d \approx 0.50$)

b) Average sample sizes are small ($n \approx 40$)

c) Typical power is low ($1-\beta \approx .35$)

d) In neuroscience even lower ($1-\beta \approx .21$)

This means that, assuming that all effects are true, there should be approx. 35% of positive findings but they actually are **over 90%** and above all other sciences (Fanelli, 2012)

- How this is possible?

UNIVERSITÀ DEGLI STUDI DI MILANO BICOCCA

How To Win Lotto

8.7 Times Out Of Every 10 Times?

Click Here

For The Most Respected Lottery System On The Internet
100% Guaranteed

A few final considerations

- **Increase sample size if you want to get it right**
- Decrease "noise" in the study
- Dichotomous thinking does not help: everything happens under some circumstances
- Ask what, when, how much, how something happens
- Get it right ≠ I am right
- To get it right means to reduce False positives (Type I error), False negatives (Type II error) and to have reasonably precise estimates

Tversky & Kanheman (1971)
(Belief in the law of small numbers, PB, 76, 105-110

We submit that people view a sample randomly drawn from a population as highly representative, that is, similar to the population in all essential characteristics. Consequently, they expect any two samples drawn from a particular population to be more similar to one another and to the population than sampling theory predicts, at least for small samples.

In review, we have seen that the believer in the law of small numbers practices science as follows:

1. He gambles his research hypotheses on small samples without realizing that the odds against him are unreasonably high. He overestimates power.
2. He has undue confidence in early trends (e.g., the data of the first few subjects) and in the stability of observed patterns (e.g., the number and identity of significant results). He overestimates significance.
3. In evaluating replications, his or others', he has unreasonably high expectations about the replicability of significant results. He underestimates the breadth of confidence intervals.
4. He rarely attributes a deviation of results from expectations to sampling variability, because he finds a causal "explanation" for any discrepancy. Thus, he has little opportunity to recognize sampling variation in action. His belief in the law of small numbers, therefore, will forever remain intact.