Reproducibility and Statistical Methodology

Anthony Almudevar

Department of Biostatistics and Computational Biology, University of Rochester

Center for Biomedical Informatics Seminar Series, October 24, 2017

Table of Contents

From: Open Science Collaboration (2015) Estimating the reproducibility of psychological science, Science [Brian Nosek, Corresponding Author]

- We conducted replications of 100 experimental and correlational studies published in three psychology journals using high-powered designs and original materials when available;
- 97% of original studies had significant results ($P < .05$);
- 36% of replications had significant results;
- 47% of original effect sizes were in the 95% confidence interval of the replication effect size;
- 39% of effects were subjectively rated to have replicated the original result.

What does the Open Science Consortium conclude?

- A large portion of replications produced weaker evidence for the original findings ...
- ... [V]ariation in the strength of initial evidence (such as original P value) was more predictive of replication success than variation in the characteristics of the teams conducting the research (such as experience and expertise).
- Reproducibility is not well understood because the incentives for individual scientists prioritize novelty over replication.
- Replication can increase certainty when findings are reproduced and promote innovation when they are not.

What might a researcher conclude?

- Reproducibility rates attained by the Reproducility Project are presented as being self-evidently low.
- At the same time, the original article presents quite sensible (and sometimes benign) reasons why reproducibility rates would not be near perfect.
- If there were no uncertainty regarding the experimental outcome, we wouldn't need to do the experiment.
- No systematic flaws in experimental or statistical methodology are discerned.
- A problem with the incentive system in the scientific community is conjectured.
- So it is worth asking, **"What should we expect the reproducibility rate to be?"**

The issue of reproducibility is among the most significant ones in science.

- Collins & Tabak (2014) *Policy: NIH plans to enhance reproducibility*. Nature Comment
- Baker Monya (2015) *US societies push back against NIH reproducibility guidelines.* Nature News

Attempts to characterize, measure and explain reproducibility consistently appear in the literature.

- Button et al (2013) *Power failure: why small sample size undermines the reliability of neuroscience.* Nature Reviews Neuroscience.
- Hoppe C (2013) *A test is not a test*. Nature Reviews Neuroscience Correspondence.
- Button et al (2013) *Empirical evidence for low reproducibility indicates low pre-study odds*. Nature Reviews Neuroscience Correspondence.

Not all efforts to directly measure reproducibility yield pessimistic conclusions ...

Etz & Vandekerckhove (2016) *A Bayesian perspective on the reproducibility project: Psychology*. PLoS One:

> We revisit the results of the recent Reproducibility Project ... Overall, 75% of studies gave qualitatively similar results in terms of the amount of evidence provided ... We conclude that **the apparent failure of the Reproducibility Project to replicate many target effects can be adequately explained by overestimation of effect sizes**.

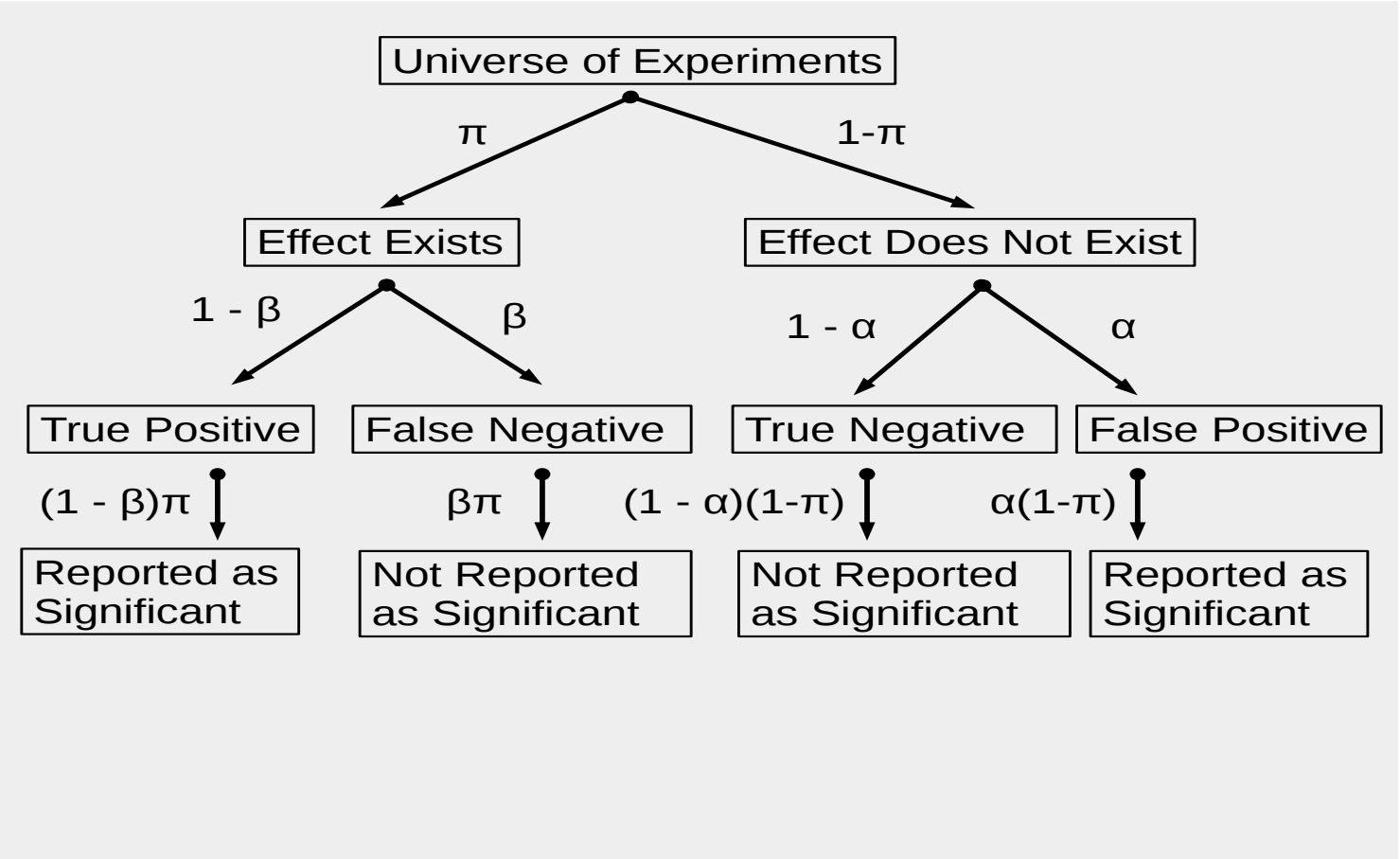
Klein et al (2014) *Investigating variation in replicability: The many labs replication project*. Social Psychology:

- "This research tested variation in the replicability of 13 classic and contemporary effects across 36 independent samples totaling 6,344 participants."
- $11/13 \approx 85\%$ successfully replicated (1 weakly, 10 strongly).
- "... [I]f MLP (Many Labs Project) had used OSCs method, they would have reported an unsettling replication rate of 34% rather than the heartening 85% they actually reported". Gilbert et al. (2016) *Comment on 'Estimating the reproducibility of psychological science'*. Science, with response by the authors.

Baker & Dolgin (2017) *Cancer reproducibility project releases first results*. Nature News:

> An open-science effort to replicate dozens of cancer-biology studies is off to a confusing start.

The process of reporting scientific results is typically represented as a decision tree.



Defining, Measuring and Interpreting Reproducibility

- There is a universe $\mathcal{U}$ of hypothesis tests. For a proportion π of these, the alternative hypothesis is true.
- This can be understood as a significant finding of scientific interest.
- The intention is to report these in a peer-reviewed journal.

	$P > \alpha$	$P \leq \alpha$	
H_0	True Negative	False Positive	$1 - \pi$
H_a	False Negative	True Positive	π

We refer to π as an effect prevalence, or alternatively as a prior effect probability.

We wish to know if A is true, based on evidence E.

$$\begin{aligned} P(E \mid A) &= sens \\ P(E^c \mid A^c) &= spec \\ P(A \mid E) &= PPV \\ P(A^c \mid E^c) &= NPV \\ P(A) &= prev. \end{aligned}$$

Applying Bayes Rule:

$$PPV = \frac{sens \times prev}{sens \times prev + (1 - spec) \times (1 - prev)}$$

In terms of *odds*, where $Odds(p) = p/(1-p)$:

$$Odds(PPV) = \left(\frac{sens}{1 - spec} \right) \times Odds(prev)$$

This allows us to relate $P(A \mid E)$ and $P(E \mid A)$, transposing the conditional.

The process of scientific reporting is analogous to diagnostic testing:

$$\begin{aligned} E &= \{ \text{ Patient tests positive } \} = \{P \leq 0.05\} \\ A &= \{ \text{ Patient has infection } \} = \{H_a \text{ is true } \} \end{aligned}$$

Then Type I error α and Type II error β are

$$\begin{aligned} \alpha &= 1 - \mathit{spec} \\ \beta &= 1 - \mathit{sens} \end{aligned}$$

Then ***PPV* is the proportion of findings reported as significant that really are significant**, and is related to π by

$$\mathit{Odds}(\mathit{PPV}) = \left(\frac{1-\beta}{\alpha}\right) \times \mathit{Odds}(\pi).$$

So, the odds that a positive is a true positive is about 18 times the odds that the investigated effect exists, where $\beta = 0.1$ and $\alpha = 0.05$.

While *PPV* drives reproducibility, the observed reproducibility among replication studies is influenced by the Type I, II errors of the replication protocol.

Suppose a replication study adopts Type I,II errors α^*, β^*. These need not be the same as the orignal study. We can define

$$\begin{aligned} PPV_{obs} &= \text{The proportion of positive replication studies} \\ &\approx PPV(1-\beta^*) + (1-PPV)\alpha^*, \end{aligned}$$

equivalently,

$$PPV \approx \frac{PPV_{obs} - \alpha^*}{1-\alpha^*-\beta^*}.$$

Therefore, the distinction between PPV and PPV_{obs} is dependent only on the replication protocol.

Open Science Collaboration (2015) Estimating the reproducibility of psychological science, Science:

$$\begin{aligned} \alpha^* &= 0.05 \\ 1-\beta^* &= 0.92 \text{ nominal average.} \end{aligned}$$

Klein et al (2014) *Investigating variation in replicability: The many labs replication project*. Social Psychology:

- The number of study replications was 36. As the number of replications increase, α^* and β^* approach zero, so that $PPV \approx PPV_{obs}$.

Since the discrepancy between PPV_{obs} and PPV is entirely attributable to replication protocol, the goal should be to control for α^* and β^*, in order to estimate PPV.

This leaves,

$$Odds(PPV) = \left(\frac{1-\beta}{\alpha}\right) \times Odds(\pi),$$

which can be decomposed in a natural way.

α: Possibly, P-values are being underestimated (no mutiple testing control, improper cross-validation, etc). The effect of this is to increase α.

β: If a study is underpowered $(1-\beta)$ decreases (small sample sizes dilute the pool of published true positives).

π: We will discuss this next ...

What do we expect π to be?

From OSC (2015) *Estimating the reproducibility of psychological science*. Science:

> On the basis of only the average replication power of the 97 original, significant effects [M = 0.92, median (Mdn) = 0.95], we would expect approximately 89 positive results in the replications **if all original effects were true and accurately estimated**; however, there were just 35 [36.1%; 95% CI = (26.6%, 46.2%)], a significant reduction [McNemar test, $\chi^2(1) = 59.1$, $P < 0.001$].

- In other words, the authors use as a benchmark $PPV = 1$.
- **This also means $\pi = 1$.**
- Of course, this can't be the case. If it were, there would be no need to do any investigation at all.

How much difference does this make? Assume $\alpha^* = 0.05$, $\beta^* = 0.1$.

$$\begin{aligned} \pi &= 1 \\ PPV &= 1 \\ PPV_{obs} &\approx 0.9. \end{aligned}$$

$$\begin{aligned} \pi &= 0.5 \\ PPV &\approx 0.947 \\ PPV_{obs} &\approx 0.808. \end{aligned}$$

$$\begin{aligned} \pi &= 0.25 \\ PPV &\approx 0.857 \\ PPV_{obs} &\approx 0.776. \end{aligned}$$

$$\begin{aligned} \pi &= 0.05 \\ PPV &\approx 0.486 \\ PPV_{obs} &\approx 0.439. \end{aligned}$$

So, the reported reproducibility rate is consistent with $\pi \approx 0.05$, assuming the error rates of both the original and follow-up studies are correctly calculated.

- What would be a realistic value for π? First of all, we need to define our universe of hypothesis tests $\mathcal{U}$. It will be useful to make the following distinction:
- **Primary Analysis:** Resolves a hypothesis of scientific consequence. Power analysis is used to ensure, at least, $\beta \leq 0.2$, usually smaller.
- **Secondary Analysis:** A finding that enhances or qualifies a primary analysis. Type II error probability need not be controlled.
- A primary analysis concerns a hypothesis that is usually supported by prior evidence, both statistical and mechanistic. This is comparable to a *Specific Aim* of a research proposal.
- **A reasonable conjecture is that π would be smaller for secondary or exploratory analyses.**

In principle, π for clinical trials can be estimated from reported success rates.

From *The Positive Value Of Negative Drug Trials*, Forbes Magazine, Paul Hsieh, Aug 30, 2015

> Under current US law, drug trials are supposed to be registered on a government website, `ClinicalTrials.gov`. After the study has been completed, researchers are required to post their results within one year positive or negative.

The Forbes article cites Anderson et al (2015) *Compliance with Results Reporting at* `ClinicalTrials.gov`, NEJM:

> At 12 months, results had been reported for 17.0% of trials that were funded by industry, 8.1% of trials funded by the NIH, and 5.7% of trials funded by other government or academic institutions. At 5 years, results had been reported by 41.5% of trials funded by industry, 38.9% of those funded by the NIH, and 27.7% of those funded by other government or academic institutions.

From FDA website www.fda.gov

In general, the success rate of clinical trials is discussed in the literature, and reported to some degree by research institutes and agencies.

- **Phase 1:** Testing of drug on healthy volunteers for dose-ranging.
  Approximately 70% of drugs move to the next phase.
- **Phase 2:** Testing of drug on patients to assess efficacy and side effects.
  Approximately 33% of drugs move to the next phase.
- **Phase 3:** Testing of drug on patients to assess efficacy, effectiveness and safety.
  Approximately 25-30% of drugs move to the next phase.

This suggests that even for significant primary analyses, we can expect $\pi < 1/2$.

SWOG (formerly the Southwest Oncology Group) is a National Cancer Institute (NCI) supported organization that conducts clinical trials in adult cancers

From *In defense of the negative result*, March 10, 2016, Rachel Tompa, Fred Hutch News Service.

> The researchers [Unger et al (2015)] delved into data from a series of large cancer clinical trials conducted in the U.S. between 1985 and 2014 by SWOG, a nationwide cancer clinical trial consortium ... **He found that SWOGs clinical trials showed similar trends to what others had seen: The rate of positive trials was about 30%**

Unger et al (2015) *The Scientific Impact of Positive and Negative Phase 3 Cancer Clinical Trials*, JAMA Oncology.

Phase II trial success rates

From Prinz et al (2011) *Believe it or not: how much can we rely on published data on potential drug targets?* Nature Reviews Drug Discovery:

> A recent report by Arrowsmith [2011] noted that the **success rates for new development projects in Phase II trials have fallen from 28% to 18% in recent years**, with insufficient efficacy being the most frequent reason for failure ... This indicates the limitations of the predictivity of disease models and also that the validity of the targets being investigated is frequently questionable, which is a crucial issue to address if success rates in clinical trials are to be improved.

Arrowsmith (2011) *Phase II failures: 2008–2010*. Nature Reviews Drug Discovery

Retzios D (2009) *Why Do So Many Phase 3 Clinical Trials Fail?* Bay Clinical R & D Services:

> The outlined approached should limit failures in pivotal studies if the Phase 2 program is well implemented. Unfortunately, this is not the case. **The rate of failure in pivotal studies is still quite substantial, standing recently at about 45% [1].**
>
> In certain key areas and with more novel compounds, the failure rate has been substantially higher. For example, **for biopharmaceuticals that entered clinical trials in oncology throughout the 90s, the success rate was a very low 13% [2]. Recent estimates by the FDA have lowered this estimate to approximately 8% [3].**

[1] Kola & Landis (2004) *Can the pharmaceutical industry reduce attrition rates*. Nat Rev

[2] Pavlou & and Reichert (2004) *Recombinant Protein Therapeutics - Success rates, market trends and values to 2010*. Nat Biotechnol

[3] FDA (2004) *Challenges and Opportunities Report*

What value of π do we expect for the Reproducibility Project?

From OSC (2015) *Estimating the reproducibility of psychological science*. Science:

> **By default, the last experiment reported in each article was the subject of replication**. This decision established an objective standard for study selection within an article and was based on the intuition that the first study in a multiple-study article (the obvious alternative selection strategy) was more frequently a preliminary demonstration.

Thus, if the studies used in the Reproducibility Project are largely secondary analyses, then we must consider the possibility that π for their study universe $\mathcal{U}$ would be well below 1/2.

The problem of sample size estimation for a reproducibility study introduces technical issues not normally encountered in conventional experimental design.

To fix ideas, consider a one-sided test for a normal observation of known variance σ^2. We are given an *iid* sample of size n from $N(\mu, \sigma^2)$ to test

$$H_o : \mu = 0 \text{ against } H_a : \mu > 0.$$

Equivalently, we are interested in *standardized effect size* and *noncentrality* parameter

$$\delta = \frac{\mu}{\sigma}, \quad \eta = \sqrt{n}\frac{\mu}{\sigma} = \sqrt{n}\delta.$$

If the observed sample mean is $\bar{X}_{obs}$, estimates of effect size and noncentrality parameter are

$$\hat{\delta} = \frac{\bar{X}_{obs}}{\sigma}, \quad \hat{\eta} = z_p = \frac{\bar{X}_{obs}}{\sigma/\sqrt{n}}.$$

Textbook Power Analysis

The standard formula for a sample size required for power $1 - \beta$ is

$$n^* = \left(\frac{\sigma(z_\alpha + z_\beta)}{\mu} \right)^2.$$

Using approximation $\bar{X}_{obs} \approx \mu$ for the true alternative mean, we have

$$n^* = n \left(\frac{z_\alpha + z_\beta}{z_p} \right)^2.$$

However, since z_p is random, so is n^* and β, which can be shown to have distribution:

$$\hat{\beta} = \Phi \left(z_\alpha - \frac{z_\alpha + z_\beta}{\frac{Z}{\eta} + 1} \right), \quad Z \sim N(0, 1).$$

Interestingly, this depends on the original study parameters only by noncentrality parameter η.

However, we cannot avoid a selection effect, which truncates the distribution of n^* and $\hat{\beta}$

- As $z_p = Z + \eta$ approach 0 from above, n^* is unbounded, and neither n^* or $\hat{\beta}$ are interpretable when $z_p < 0$.
- We therefore consider two decision rules. In each we choose some threshold $t > 0$. The decision then depends on the event $z_p \geq t$.
- **Unconditional Decision**: Commit to a subsequent study for any value of z_p, but bound sample size at n_t^* calculated assuming $z_p = t$.
- **Conditional Decision**: We do not undertake a subsequent study unless $z_p \geq t$.

By how much does $E[\hat{\beta}[t]]$ differ from a nominal value $\beta = 0.2$

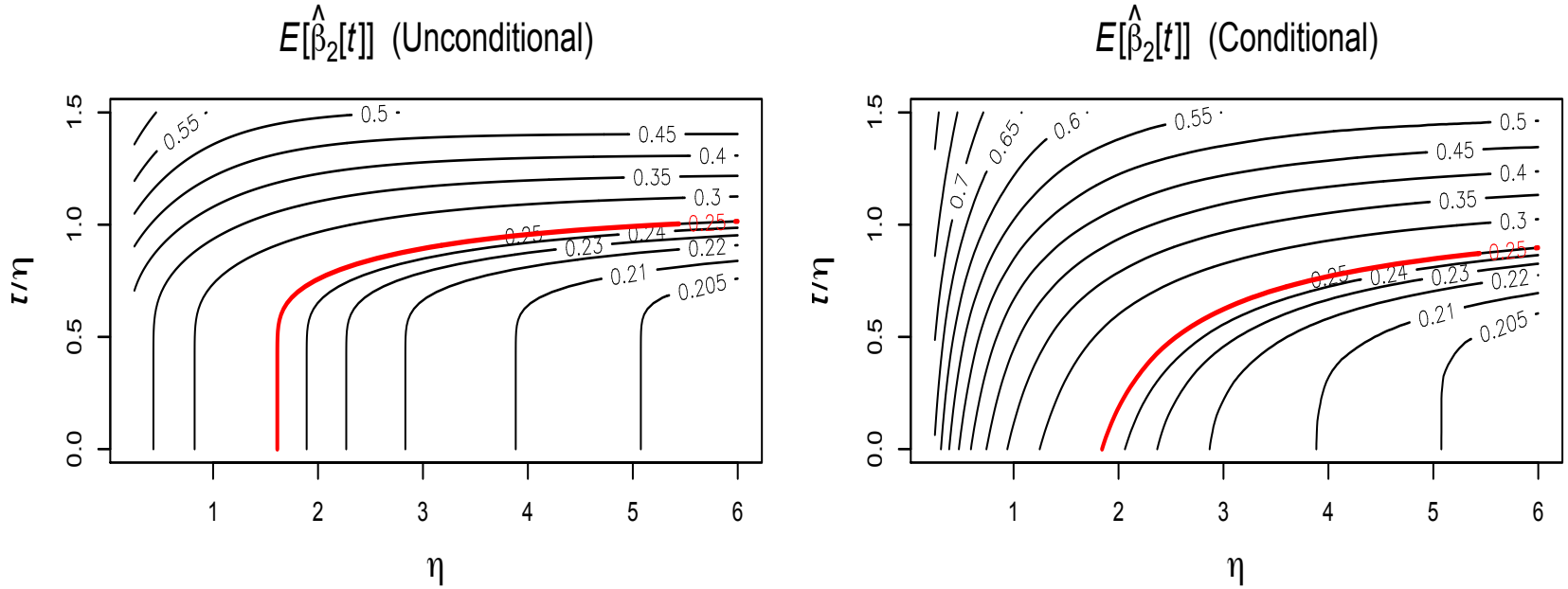


Figure: Estimated marginal power $E[\hat{\beta}[t]]$. The contour $E[\hat{\beta}[t]] = 0.25$ is superimposed on each plot. Note that the distribution of $\hat{\beta}$ now depends on threshold t.

How did the Reproducibility Project handle power analysis?

Each replication study was repowered. From OSC (2015) *Estimating the reproducibility of psychological science*. Science, Supplementary Materials:

> **Power analysis.** After identifying the key effect, power analyses estimated the sample sizes needed to achieve 80%, 90%, and 95% power **to detect the originally reported effect size** ... Post-hoc calculations showed an average of 92% power **to detect an effect size equivalent to the original studies**
>
> ... **Note that these power estimates do not account for the possibility that the published effect sizes are overestimated because of publication biases. Indeed, this is one of the potential challenges for reproducibility**.

What should a power analysis claim?

- "The sample size suffices to detect an effect size of δ^* or more with power of 90%". The truncation effect described above need not affect this.
- From the previous analysis, powering for an alternative $\eta = 4$ seems safe, if we also set threshold $t = 1$ ($t/\eta = 0.25$).
- Then suppose preliminary data yields $z_p \geq 1$ for sample size n. If these numbers are used to calculate new sample size n^* as described above, then $\beta \leq 0.2$ for alternatives

$$\mu/\sigma \geq 4/\sqrt{n},$$

  with only minimal bias in the estimate of β.
- That is, we don't know η, but we know the smallest η for which we are adequately powered.

What does the Reproducibility Project claim?

- Any $\eta > 0$ is a true effect.
- In most power analyses, β is relevant to some hypothetical effect size. The object is to report the smallest effect size we can confidently detect.
- In contrast, the Reproducibility Project claims they are powered for the true values of η.
- The η values enter the Reproducibility Project according to some unknown distribution of outcomes.
- Therefore, there is no way to rule out the possibility that a significant proportion of those values are in the interval, say, $\eta \in (0, 2]$.
- If that were true, the actual power of the Reproducibility Project replication studies would be significantly lower than the nominal average power ≈ 0.92 reported.

Data from the Reproducibility Project suggests that effect sizes from studies which did not reproduce significance are the upper tail of a distribution.

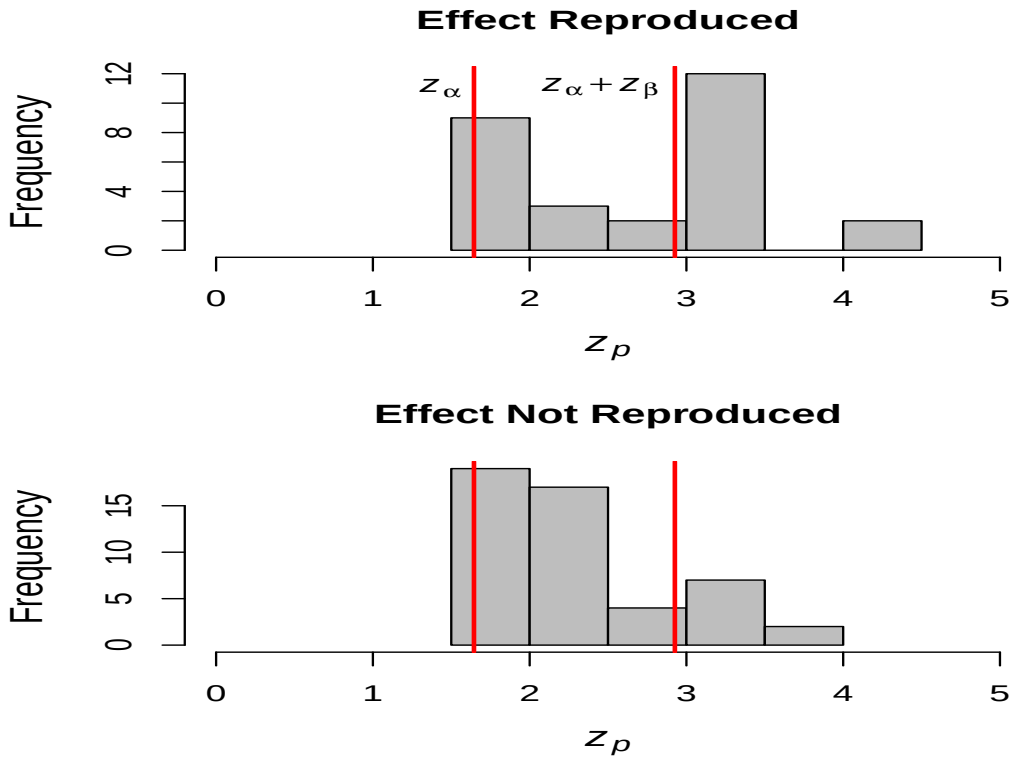


Figure: For the sake of demonstration, we can convert p-values from the Reproducibility Project, and convert them to z-scores z_p (whatever the original test used). Here, $\alpha = 0.05$, $\beta = 0.1$.

The purpose of a clinical trial is to reduce uncertainty.

From Djulbegovic et al. (2013) *Medical research: Trial unpredictability yields predictable therapy gains*. Nature:

> Here we provide empirical evidence that **the system's success rate is optimal**.
>
> ... **We find that just over half the time, RCTs [randomized controlled trials] show that new treatments are better than existing ones**.
>
> ... This success rate is incremental, but maintains a system that has served us well.
>
> ... On ethical as well as scientific grounds, RCTs should be done only when there are genuine uncertainties about the relative merits of alternative treatments.
>
> ... Progress in therapeutics has occurred precisely because science and ethics require that the results of individual RCTs are not predictable.

Replication and validation have already been accepted as an essential part of science, precisely because we anticipate that reproducibility rates will be significantly less than 100%

The Logic of Scientific Discovery (1934), Karl Popper

> Only when certain events recur in accordance with rules or regularities, as is the case with repeatable experiments, can our observations be tested - in principle - by anyone. We do not take even our own observations quite seriously, or accept them as scientific observations, until we have repeated and tested them.
>
> Only by such repetitions can we convince ourselves that we are not dealing with a mere isolated coincidence, but with events which, on account of their regularity and reproducibility, are in principle inter-subjectively testable.

Statistical Methods for Research Workers (1925), Ronald A. Fisher

> The salutary habit of repeating important experiments, or of carrying out original observations in replicate, shows a tacit appreciation of the fact that the object of our study is not the individual result, but the population of possibilities of which we do our best to make our experiments representative.
>
> The calculation of means and probable errors shows a deliberate attempt to find out something about that population.

Conclusion - False Positives and False Negative Rates are a Ubiquitous Trade-off

- High reproducibilty rates decrease false positives at the expense of increasing false negatives, which often cost more.
- We accept $PPV = 0.3$ for a diagnostic test, because we can always repeat the test. But NPV less than 100% represents undiagnosed illness.
- In experimental science, false positives can be flagged by validation. False negatives are contributions lost to science.
- The problem is to balance false positive and false negative rates.
- Therefore, reproducibility rates can be too high, as well as too low.

Conclusion - The Goal of Experimental Science is to Reduce Uncertainty

Recall the relationship: $Odds(PPV) = \left(\frac{1-\beta}{\alpha}\right) \times Odds(\pi)$.

- Reproducibility is dependent on both methodology parameters α, β and effect prevalence π.
- Optimal entropy reduction occurs for $\pi = 1/2$.
- We should also anticipate $\pi < 1/2$, particularly for secondary or exploratory analyses.
- Thus, if the reproducibility rate really has fallen in recent decades, this may be explained as much by increased exploratory or data-driven research, as by any deterioration of methodological standards.
- If so, this wouldn't imply that such research is unsound, but it would imply that validation needs to play a larger role.