Statistical power, importance of effect sizes, and statistical analysis

Thomas Steckler

The views expressed in this presentation are solely those of the individual authors, and do not necessarily reflect the views of their employers.

Information Meeting for Grant Applicants on the Call for Proposals for Confirmatory Preclinical Studies and Systematic Reviews - Quality in Health Research, Feb 19, 2019

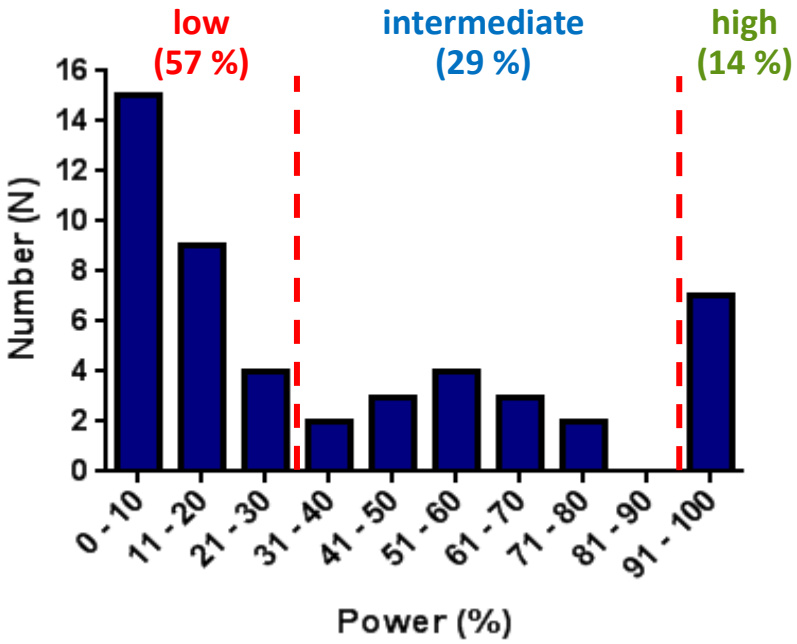
janssen | PHARMACEUTICAL COMPANIES OF Johnson&Johnson

"Although conceptionally simple, the notion of statistical power is rarely considered by most scientists"

Phillippe Vandenbroeck et al. (2006)
J Biopharm Statistics 16, 61-75

Underpowered Studies – A Common Observation

Median Power of Studies Included in Neuroscience Meta-Analyses



Studies with low statistical power have:

- Reduced chance of detecting a true effect
- Low likelihood that a statistically significant result reflects a true effect
- Overestimated effect sizes
- Low reproducibility

Button et al., Nature Rev Neurosci, 2013

Included in the analysis were articles published in 2011 that described at least one meta-analysis of previously published studies in neuroscience with a summary effect estimate (mean difference or odds/risk ratio) as well as study level data on group sample size and, for odds/risk ratios, the number of events in the control group.

Another Consequences of Underpowered Studies: Violation of the 3Rs

Experiments that use only a small number of animals are common, but might not give meaningful results.

MEDICAL RESEARCH

UK funders demand strong statistics for animal studies

Move addresses concerns that some experiments are not using enough animals.

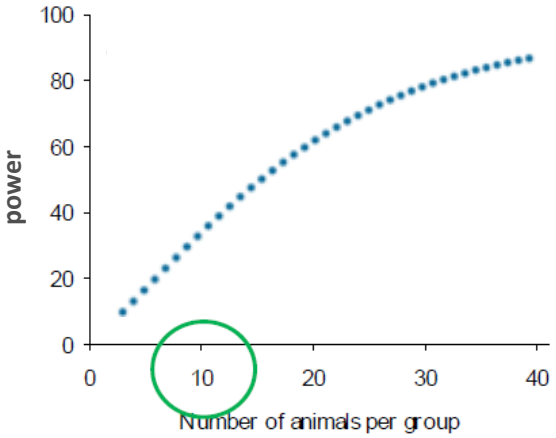
16 APRIL 2015 | VOL 520 | NATURE | 271

Numbers matter

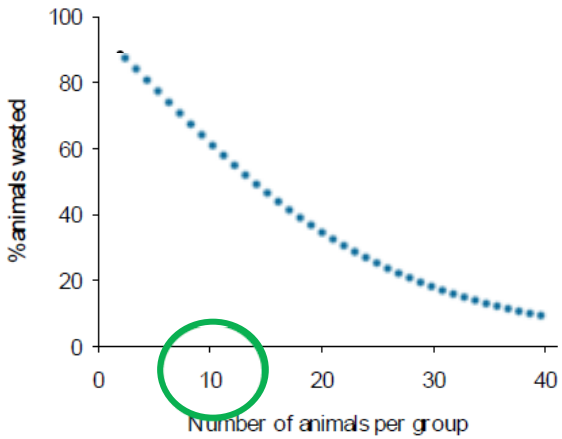
Researchers need help in making the statistical power of animal experiments clear.

16 APRIL 2015 | VOL 520 | NATURE | 263

Power as function of animal number



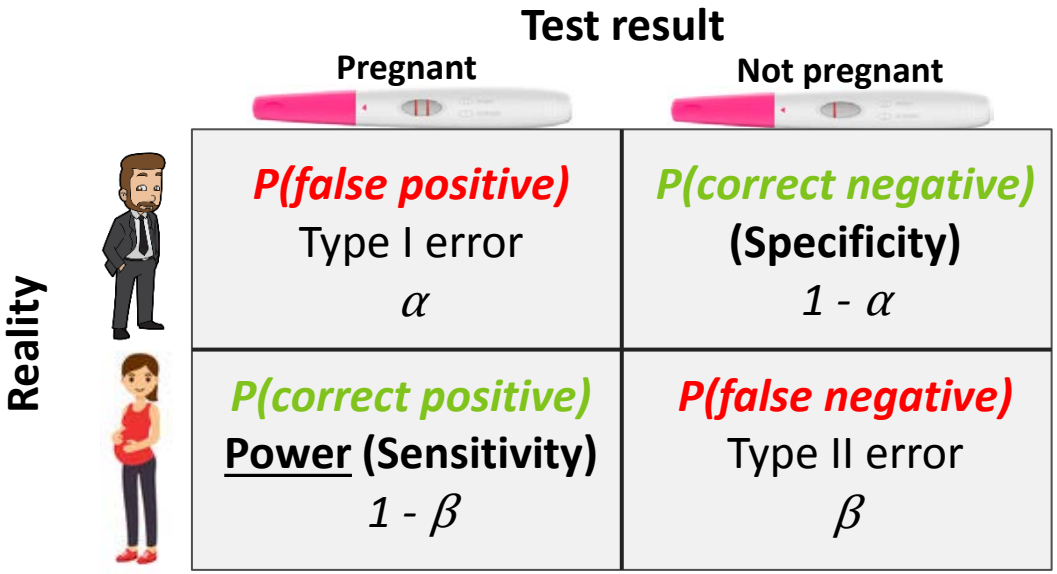
Chances of wasting an animal



Adapted from M Macleod

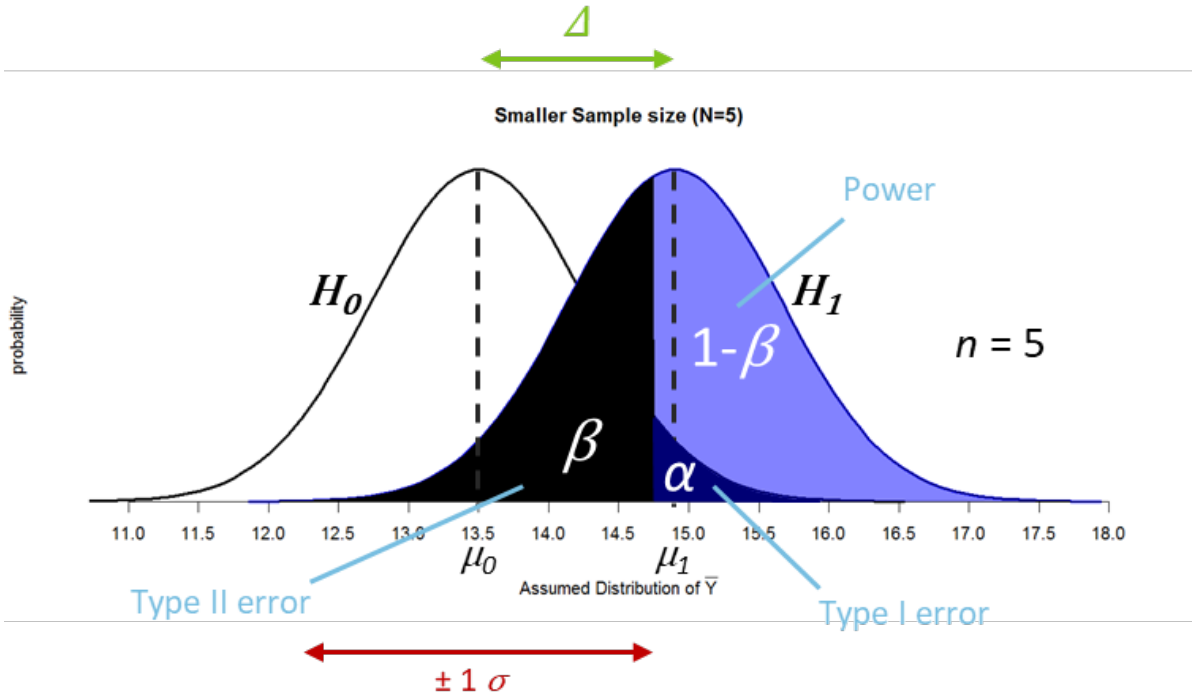
What is Power?

- $1 - \beta$
- Probability to correctly reject H_0
- Probability to avoid a Type II error or a false negative
- Chance to correctly detect a difference (Sensitivity)



Statistical Hypothesis Testing

Normal Distribution of H_0 and H_1



H_0: Null Hypothesis

H_1: Alternative Hypothesis

α: P(false positive)

β: P(false negative)

$1\text{-}\beta$: P(correct positive)

μ: Population mean

σ: Population variance

Δ: Effect size

Statistical Hypothesis Testing

- What is the *probability* of your data showing the observed difference when in reality there is no difference (i.e., $\mu_1 - \mu_0 = 0$)?
  - p **- values** describe these probabilities
    - A formal decision-making process that ensures false positives (or Type I error, α) occur only at a predefined rate
    - The smaller the p-value, the lower the probability that there is a false positive
    - **NOT** the probability that an observed finding is true
    - **NOT** implying a biologically meaningful effect
  - Statistical significance if $p < 0.05$
    - Means that if there is a false positive rate below $\alpha = 5\%$, we are willing to reject H_0
    - Is an **arbitrary convention**, not a fundamental principle

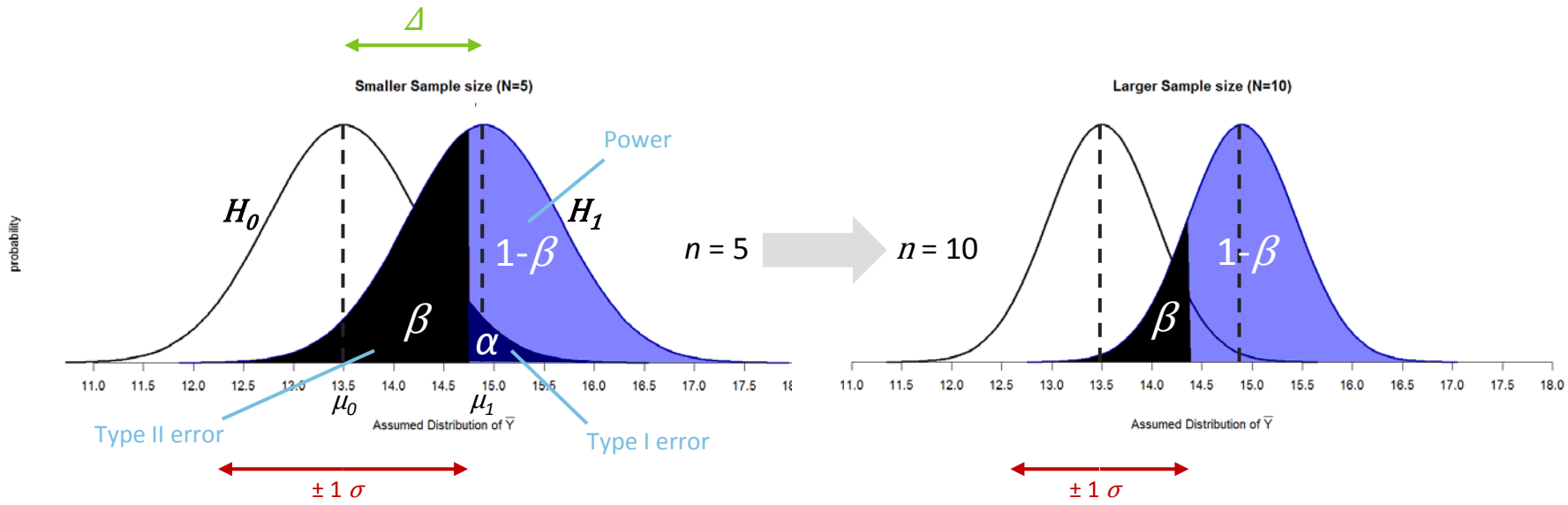
What Makes a P-Value?

- p is a value computed by a statistical test (and may differ according to the test used)

- There is a mathematic relationship in any data set between:
  - Accepted significance level (α), and hence p-value
  - Observed/desired effect size (Δ)
  - Variability within tested sample (standard deviation; σ)
  - Sample size (n)

Changing one factor will affect the others

Relevance of Sample Size for Power



- Type 2 error (β) will decrease as a function of sample size
- **Power** (1-β) will increase as a function of sample size
- Sampling mean variability will decrease as a function of sample size

Sample Size is Where You Randomize!

- **Sample size** = number of experimental units per group
- **Experimental unit**: entity subjected to an intervention independent of all other units

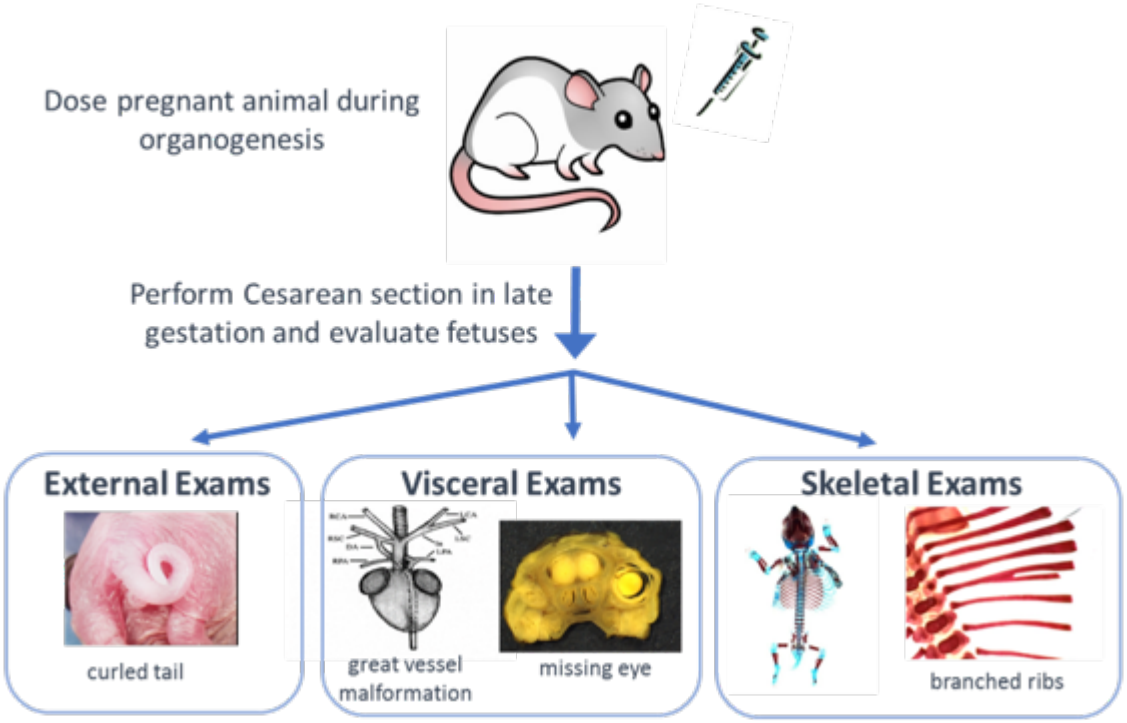
- **Observational unit:** entity on which measures are taken (primary) outcome measure

- **Biological unit:** entity about which inferences are made

Example Teratogenicity Study

- Regulatory requirement as part of drug development program
- Q: Does treatment X have embryotoxic or teratogenic effects?

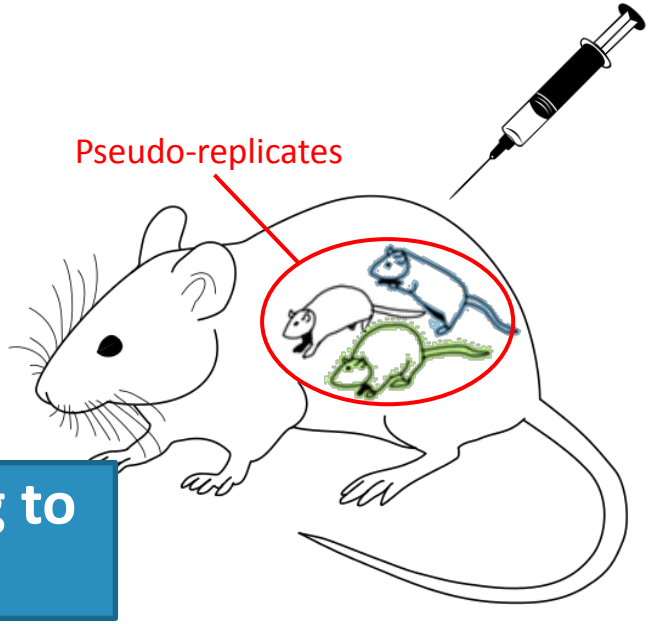
- Study design:



Susan B. Laffan, Teratology Primer, 3rd Edition

Example Teratogenicity Study

- **Biological unit:**
  - The offspring (do fetuses show abnormalities?)
- **Observational unit:**
  - Observations made in the offspring during external, visceral and skeletal examination
  - E.g., fetuses with branched ribs
- **Experimental unit:**
  - The litter, not the individual fetuses!
  - (entity subjected to an intervention independent of all other units)

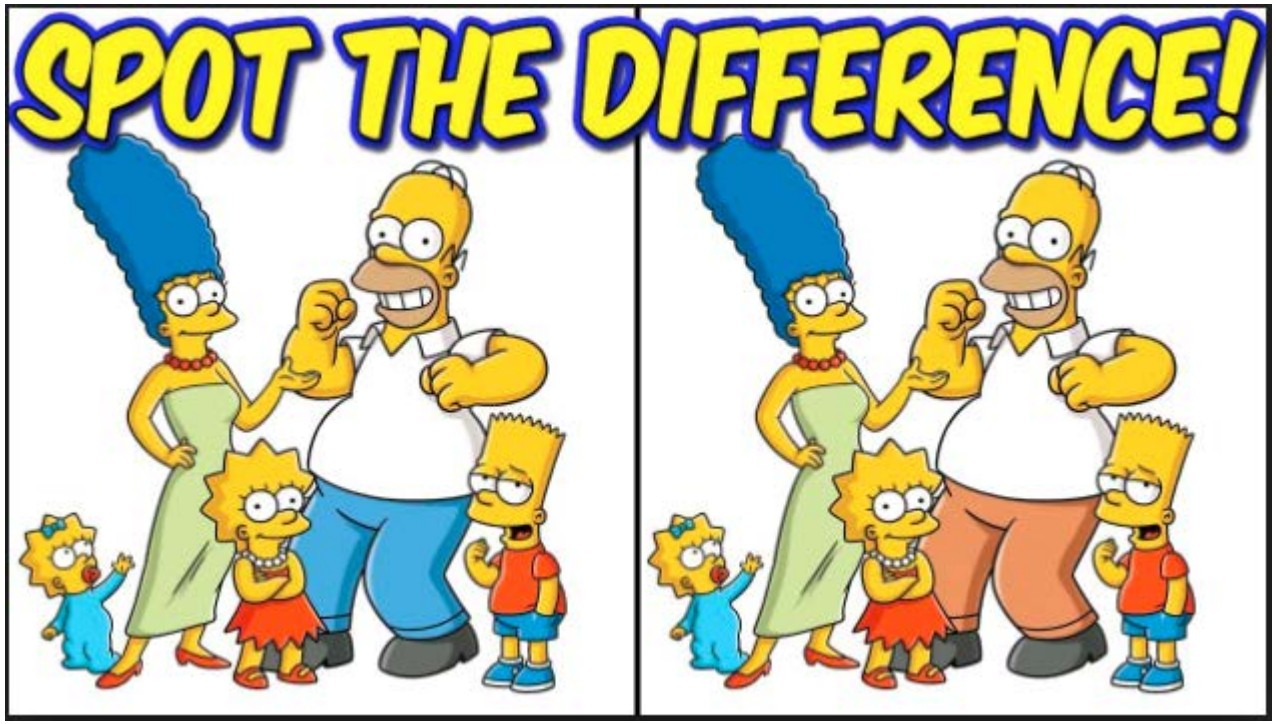


OECD and ICH do not allow offspring to be used as independent samples!

Why is Power Important?

A study may not show a difference between groups because:

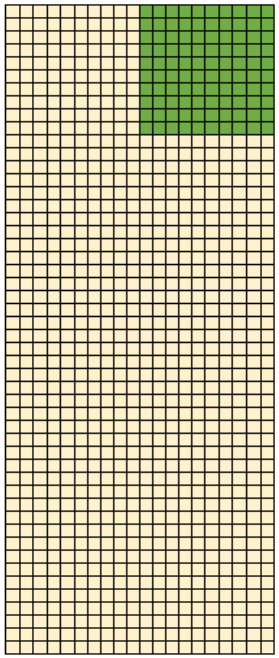
- There is no difference (true negative)
- The study failed to detect that difference (false negative)

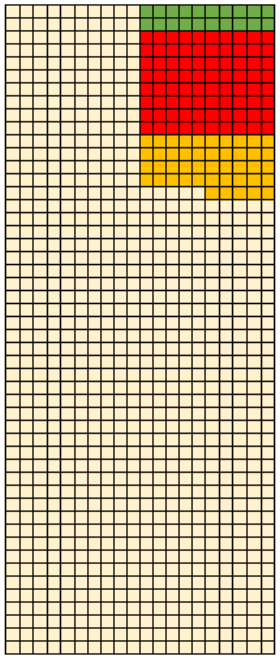


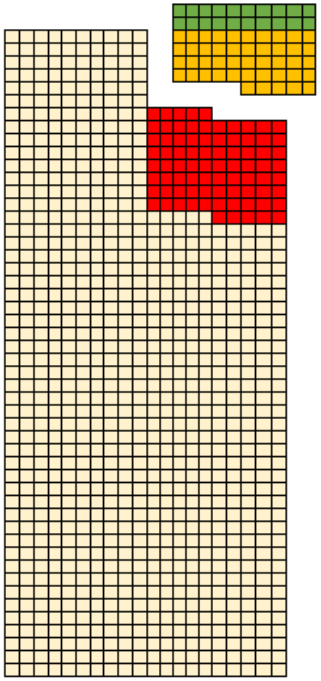
https://www.youtube.com/watch?v=SRyP2BPGUgg

Issues with Low Power

Potentiation by Bad Practice







1000 hypotheses
10% true → n = 100

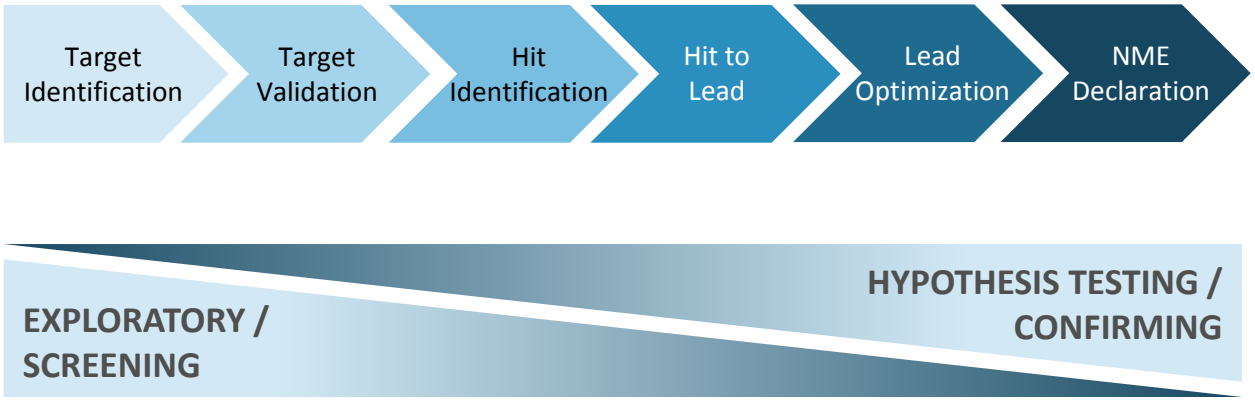
False positive rate α = 0.05
5% → *n* = 45
Power = true positive rate = 0.2
20% → *n* = 20
False negatives
80% → n = 80

65 experiments published
69% false positives

Power 0.8:
125 experiments published
36% false positives

Adapted from The Economist, 2013

Strategies Used Along the Discovery Chain



- Absence of a scientifically plausible hypothesis
- Methods can be adapted (to some degree)
- p-values cannot be interpreted at face value

- A scientifically plausible hypothesis exists
- Requires pre-specification of H_0, experimental methods incl. sample size and analytical methods
- Leads to a statement of significance

Elements can be combined
- Confirmatory for primary (and key secondary) endpoint
- Exploratory for other endpoints

Strategies Used Along the Discovery Chain

- Look for **strong effects**
- Importance to discover potential early on
- Discovery of false positives less serious
- Primary risk with the company

- Look for **true effects**
- Avoid exposure of patients to non-active therapy
- Only invest in development of efficacious therapy
- Risk of the company and of authorities

Statistical power, importance of effect sizes, and statistical analysis

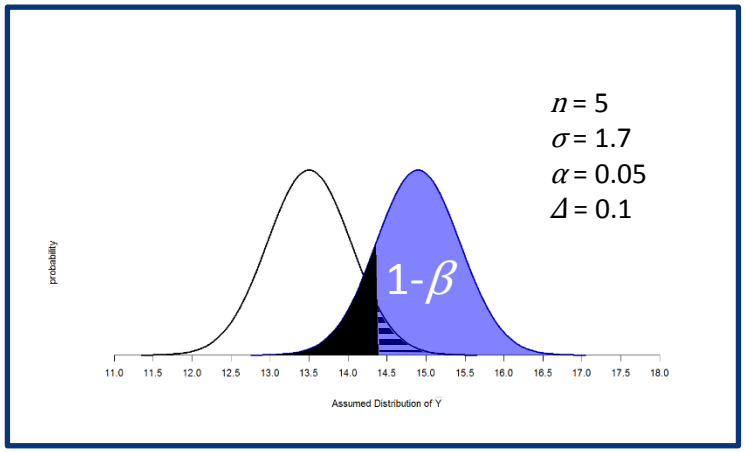
Thomas Steckler

The views expressed in this presentation are solely those of the individual authors, and do not necessarily reflect the views of their employers.

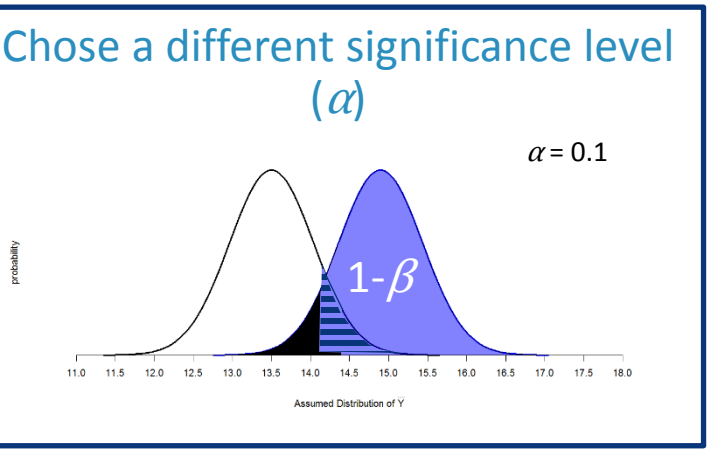
Information Meeting for Grant Applicants on the Call for Proposals for Confirmatory Preclinical Studies and Systematic Reviews - Quality in Health Research, Feb 19, 2019

janssen | PHARMACEUTICAL COMPANIES OF Johnson&Johnson

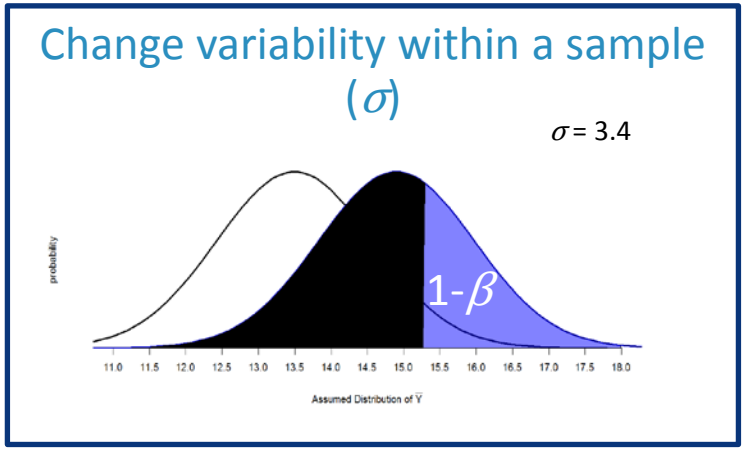
Other Ways to Affect Power



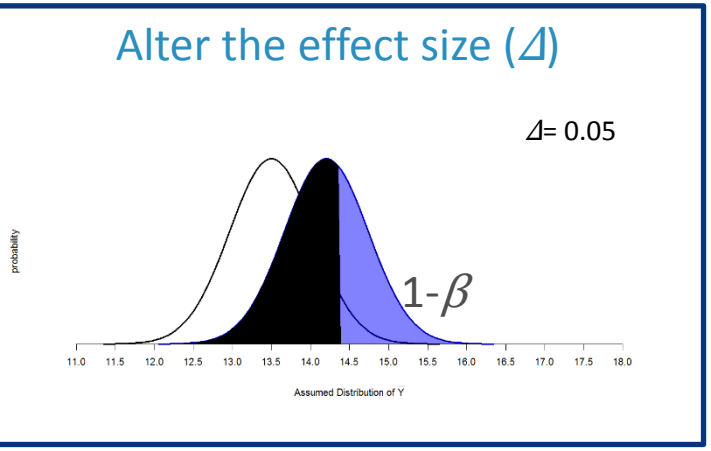
Chose a different significance level (α)



Change variability within a sample (σ)



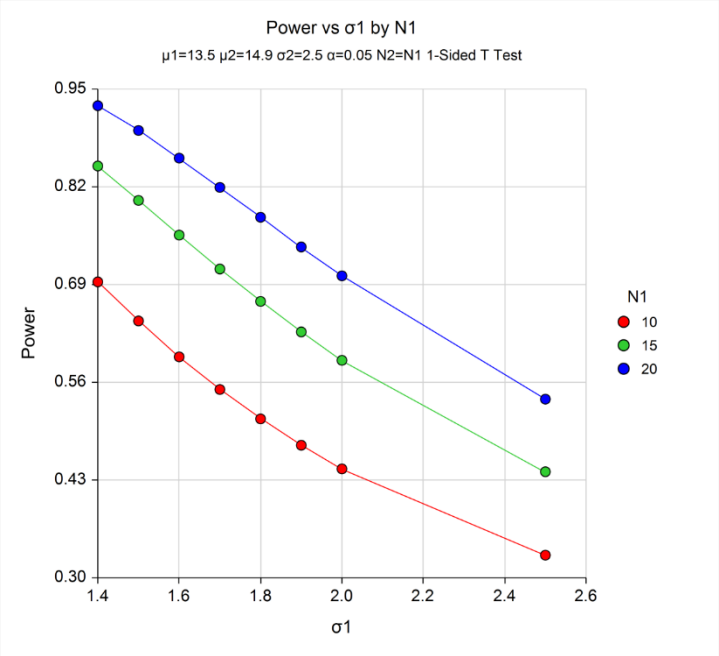
Alter the effect size (Δ)



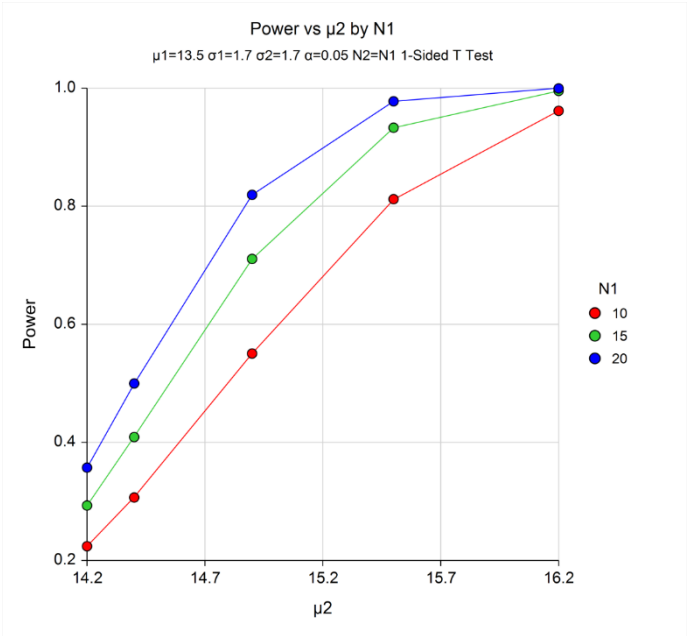
Adapted from F Tekle

The Different Factors Interact

Measures with large variability require larger sample sizes



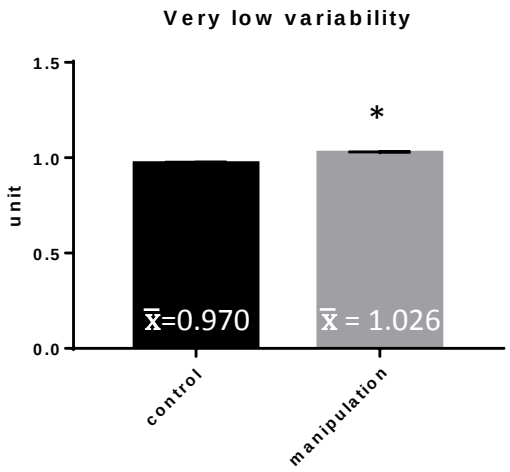
Smaller effect sizes require larger sample sizes



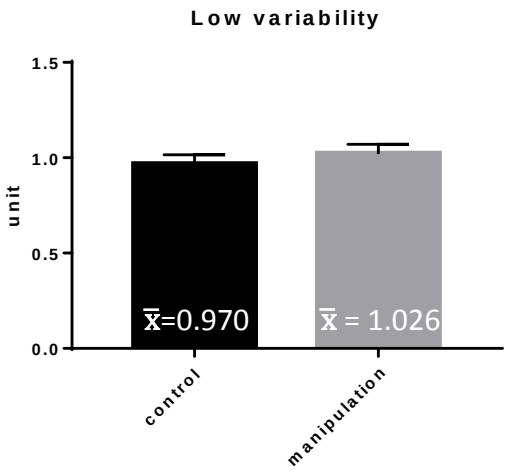
Adapted from F Tekle

Increasing Sample Size Can Make a Result Significant

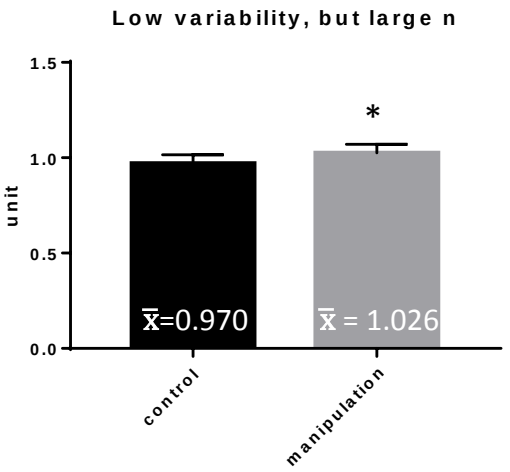
Hypothetical Data Set



n = **5**
σ = ±0.007 vs. ±0.005
p = **0.002** in t-test



n = **5**
σ = ±0.046 vs. ±0.045
p = **0.411** in t-test



n = **25**
σ = ±0.046 vs. ±0.045
p = **0.039** in t-test

But is it meaningful?

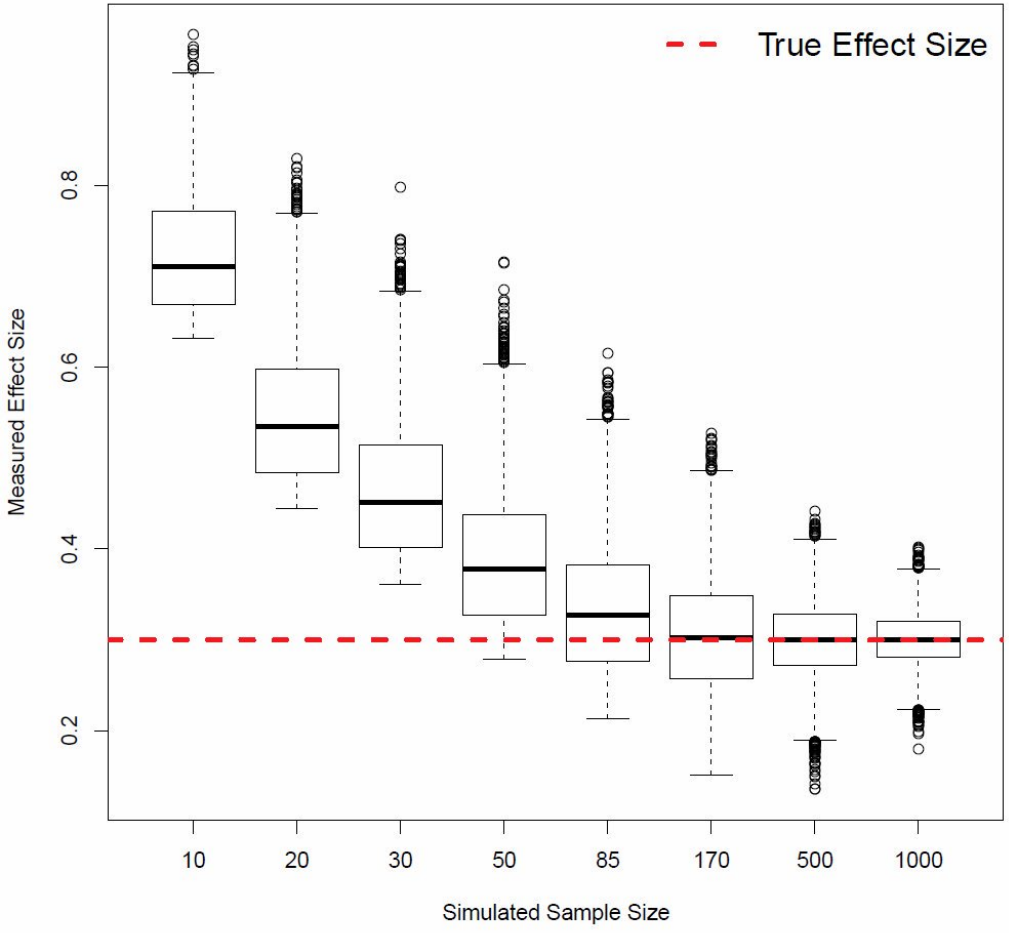
Adapted from M. Michel, 2017

What is a Meaningful Effect?

- If sample size is too high and desired effect size is not defined, there is an increased risk
  - of making a false positive conclusion (Literature is full of it!)
  - that statistically significant effects are declared that are of no practical relevance
  - that resources are wasted
  - of ethical issues (e.g., in case of animal studies)

- **A desired effect**
  - should be **relevant** (scientifically, clinically, economically,…)
  - should be **determined upfront** (which effect must be detectable to make the study relevant?)
  - should be **documented** (to avoid bias)
  - is a **choice by the researcher**
  - based on estimation or experience, not statistically determined
  - **but will affect statistics!**

Low Sample Size Inflates Measured Effect Sizes



- Exploration of sample size sets of n = 10, 20, 30, 50, 85, 170, 500, and 1000, drawn from a multivariate normal distribution
- 1,000 studies simulated per n
- True effect size arbitrarily set to $\Delta = 0.3$
- To obtain 80% power for Δ=0.3 requires 85 samples

Jim Grange (2017)
https://jimgrange.wordpress.com/2017/03/06/low-power-effect-sizes/

False Discovery Rate vs. Positive Predictive Value

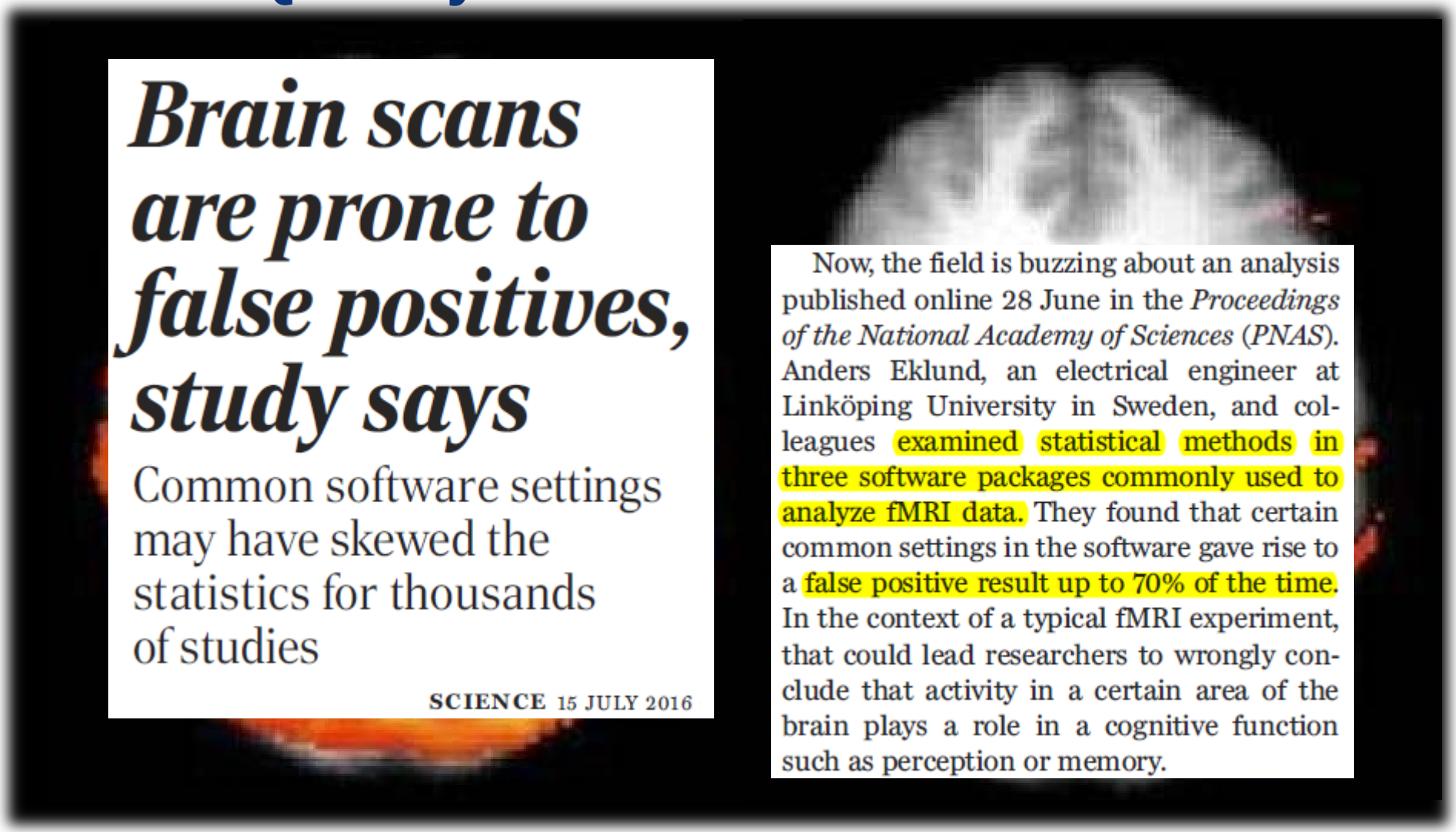
- **Positive predictive Value (PPV)**
  - Probability that a real effect exists if a "significant" result has been obtained
- **False Discovery Rate (FDR)**
  - Probability that a real effect does NOT exist if a "significant" result has been obtained
- PPV and FDR are flipsides of the same coin (FDR = 100 – PPV)

A research finding is less likely to be true in case of

- **Small studies**
- **Small effect sizes**

→low power and high FDR

Unintended Flawed Procedures Impact Data Quality



Brain scans are prone to false positives, study says

Common software settings may have skewed the statistics for thousands of studies

SCIENCE 15 JULY 2016

Now, the field is buzzing about an analysis published online 28 June in the *Proceedings of the National Academy of Sciences* (*PNAS*). Anders Eklund, an electrical engineer at Linköping University in Sweden, and colleagues examined statistical methods in three software packages commonly used to analyze fMRI data. They found that certain common settings in the software gave rise to a false positive result up to 70% of the time. In the context of a typical fMRI experiment, that could lead researchers to wrongly conclude that activity in a certain area of the brain plays a role in a cognitive function such as perception or memory.

Statistical power, importance of effect sizes, and statistical analysis

Thomas Steckler

The views expressed in this presentation are solely those of the individual authors, and do not necessarily reflect the views of their employers.

Information Meeting for Grant Applicants on the Call for Proposals for Confirmatory Preclinical Studies and Systematic Reviews - Quality in Health Research, Feb 19, 2019

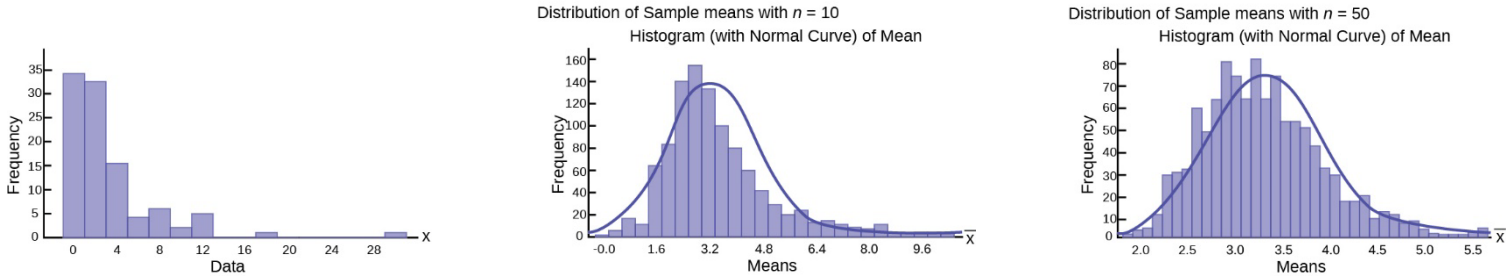
janssen | PHARMACEUTICAL COMPANIES OF Johnson & Johnson

Sample Size Affects Sampling Distribution
Hence Statistical Approach

Central Limit Theorem

- For a random sampling with a large n, the sampling distribution is approximately normal, regardless of the underlying distribution

Simulated data set of a skewed distribution:



1,000 randomly drawn samples for sample sizes n=10 and n=50
As the sample size increases, and the number of samples taken remains constant, the distribution of the 1,000 sample means becomes closer to the smooth line that represents the normal distribution

https://archive.cnx.org/contents/323cb760-5f99-4899-a96b-f3919f982a6a@10/using-the-central-limit-theorem#eip-id1169478792440

- Note also tests for normality will struggle if the sample size is small

What is the Right Power?

Cohen (1988) proposed as a convention that, when the investigator has no other basis for setting the desired power value, the value **.80** be used

- **Rationale:**
  1. **α / β Interrelationship:**
     - if 1-β increasing → β decreasing → α increasing
       (effect size and sample size being unchanged)

  2. **Relative error seriousness:** Scientists usually consider false positive claims more serious than false negative claims
     - if α = .05 and β = .20 → relative seriousness = .20/.05
       (i.e., seriousness of Type I errors = 4 x seriousness of Type II errors)

  3. **Feasibility:** Power >.90 would demand very large sample sizes
     - if 1-β increasing and α constant → needs increasing effect size (may not be feasible)
       or
       increasing sample size

There may be reasons to divert from the convention

Cohen J (1988) Statistical Power Analysis for the Behavioral Sciences, 2nd ed., Lawrence Erlbaum Associates

How to Get to the Right Sample Size?

- Based on the given **variability** and the **expected difference** between the effect of manipulation X and the effect of the control groups a minimum number of samples is required
- **Variability:**
  - Should be provided in standard deviation units (σ)
  - Can be estimated from pilot studies or data reported in the literature
  - Can be calculated from SEM and n: $\sigma = \mathrm{SEM} * \sqrt{n}$
- **Effect size (Cohen's *d*):**
  - Potential mean difference between any two groups in terms of standard deviation units
  - Signal (effect size of scientific interest)/noise (variability)
  - Can be estimated from pilot studies or data reported in the literature
  - Convention: small (Δ = 0.2), medium (Δ= 0.5), large (Δ = 0.8)*
  - If in doubt, hypothesize a Δ = 0.5

*based on benchmarks by Cohen (1988)
may be larger in case of animal studies: small (d = 0.5), medium (d = 1.0), large (d = 1.5)

Power Depends on Experimental Design and Statistical Analysis

Example: Two independent samples, comparing two means

- **Analysis:** unpaired T-test
  - T-test can be one- or two-tailed (two-tailed preferred)
    can be paired (dependent samples) or unpaired (independent samples)
- **Assumption:** comparison of two independent samples of equal n
  - Equal-sized samples are desirable, since it is demonstrable that with a given number of cases available, equal division yields greater power than does unequal division

Sample size calculation (one-tailed):

$$n = (Z_{\alpha} + Z_{1\text{-}\beta})^2 * 2\sigma^2 / \Delta^2$$

with Z_{α} = constant according to accepted α level [100(α) percentile of the standard normal distribution], depending on whether test is one- or two-tailed (in the latter case would be $Z_{\alpha/2}$)
$Z_{1\text{-}\beta}$ = constant according to power of the study [100(1-β) percentile of the standard normal distribution]
σ = common population variance
Δ = estimated effect size

Common population variance:

$$\sigma = \sqrt{(\sigma_1^{\,2} + \sigma_2^{\,2}) / 2}$$

with σ_1 = standard deviation control group
σ_2 = standard deviation experimental group

Cohen's *d*:

$$\Delta = (\mu_1 + \mu_2) / \sigma$$

with μ_1 = mean of control group
μ_2 = mean of experimental group

Z-values for T-test:

Z_α & $Z_{\alpha/2}$

	α	0.01	0.05	0.1
one-tailed	Z_α	2.326	1.645	1.282
two-tailed	$Z_{\alpha/2}$	2.576	1.960	1.645

$Z_{1-\beta}$

$1-\beta$	0.8	0.90	0.95
$Z_{1-\beta}$	0.842	1.282	1.645

Finally: Software for Sample Size Calculation

- G*Power
  - http://www.gpower.hhu.de
  - Stand-alone program
  - Windows, Mac
  - Free
- PASS
  - https://www.ncss.com/software/pass
  - Stand-alone program
  - Windows
  - Not freely available
- Various R packages
  - Windows, Mac, Linux
  - Generally free
- SAS, SPSS, Minitab, Microsoft Excel packages
- NC3Rs Experimental Design Assistant (EDA)
  - https://eda.nc3rs.org.uk/experimental-design-group
  - Free

If in doubt, ask your STATISTICIAN !

Questions?

Know Your Research Question

Research Question:

Does **Intervention X (I_X)** *alter* the **Primary Outcome Measure Y (POM_Y)** ?

- increase
- decrease
- increase or decrease

Null Hypothesis (H_0):

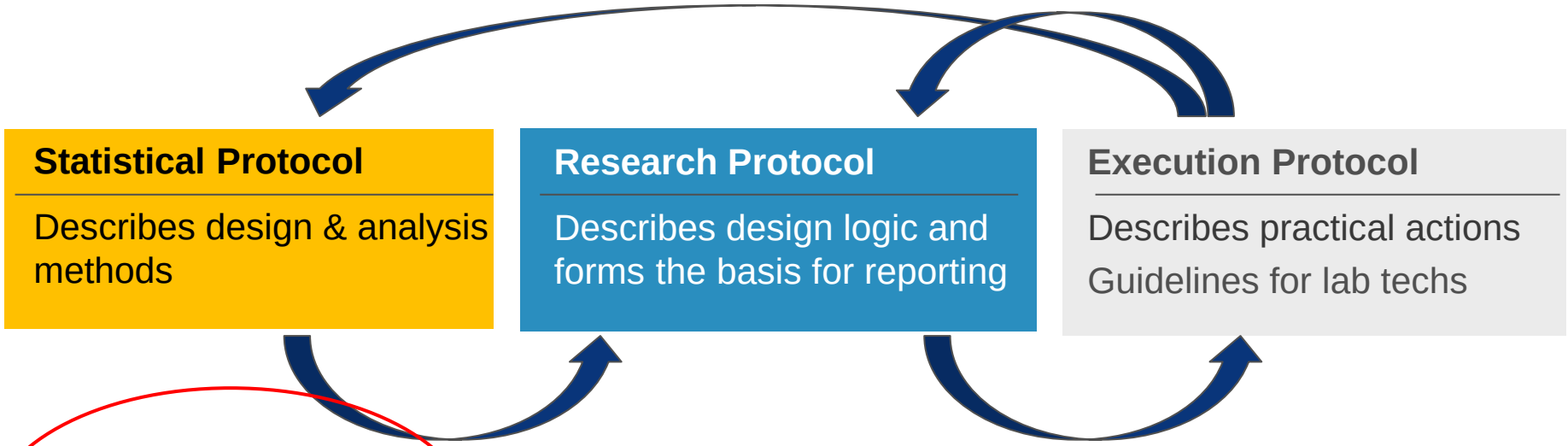
- I_X has **NO** effect on POM_Y

Alternative Hypothesis (H_1):

- I_X has an effect on POM_Y
  - directional (one-sided)
  - non-directional (two sided)

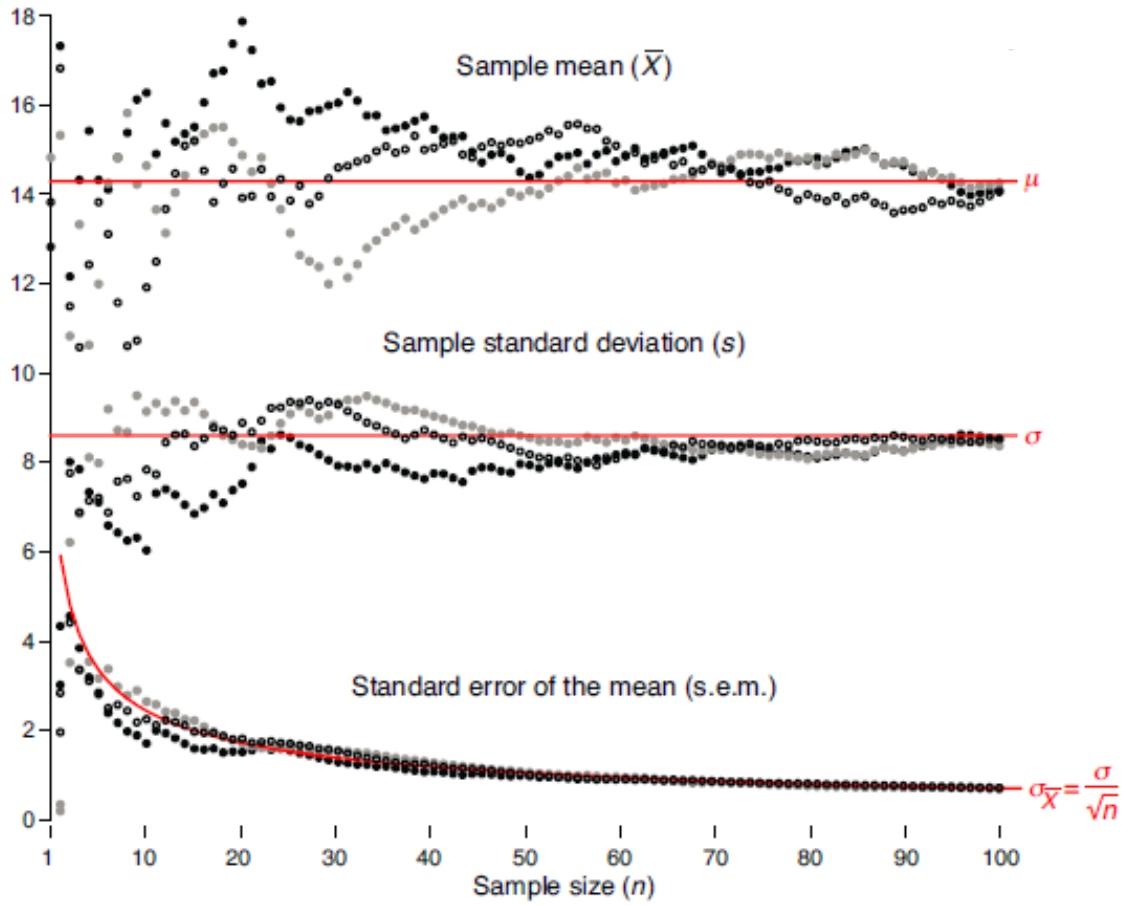
The Protocol

Reflects the various layers in the design and execution of the experiment



Adapted from L. Bijnens, 2017

Relevance of Sample Size



Mean, S_D and SEM for three samples increasing size n=1 to n=100

Adapted from F Tekle

Power Analysis

A-priori

- What do I need for my experiment and is it feasible?
  - What sample size is needed to detect a difference at predefined effect size, anticipated variability and with a certain probability, given that this difference indeed exists?
  - What sample size is needed to detect a difference at predefined effect size, anticipated variability and with a certain probability, given that this difference indeed exists?
  - Will I be able to run this experiment?
    - Can I get enough samples?
    - Can I test all the samples?
    - Can I afford the experiment?
  - If not, does it still make sense to run the experiment?

Post-hoc

- How reliable are the data (reported)?
  - What was the power of a reported study, given the observed effect size, variability and number of samples reported?
  - Can I “trust” the data?

Underpowered Studies – Still an Issue 5 Years Later

Efficacy of Analgetics

- Effect of drug interventions in models of chemotherapy-induced peripheral neuropathy
- Systematic search, inclusion of 341 publications by Nov 2015

Outcome measure	Model (examples)	Number comparisons	Post-hoc power
Evoked limb withdrawal to mechanical stimuli	e.g., electronic "von Frey", mechanical monofilament, pin prick, pinch test	648	0.06
Evoked limb withdrawal or vocalisation to pressure	e.g., Randall-Selitto paw pressure	235	0.07
Evoked limb withdrawal to cold	e.g., acetone/ethylchloride/menthol, cold plate, cold water	251	0.06
Evoked limb withdrawal to heat	e.g., radiant heat, hot plate, paw immersion	140	0.07
Evoked tail withdrawal to cold	e.g., tail immersion	50	0.06
Evoked tail withdrawal to heat	e.g., tail flick, tail immersion	38	0.19
Complex behaviour, pain-related	e.g., TRPA1 agonist- or capsaicin-evoked nocifensive behavior, burrowing activity, CPP, thermal place preference	12	0.10

unpublished data, based on Currie et al., bioRxiv, 2018, http://dx.doi.org/10.1101/293480, courtesy of Ezgi Tanriver-Ayder, Gillian L. Currie, Emily Sena

More Likely for a Research Claim to be False Than True

PLoS Medicine | August 2005 | Volume 2 | Issue 8 | e124

Open access, freely available online

Essay

Why Most Published Research Findings Are False

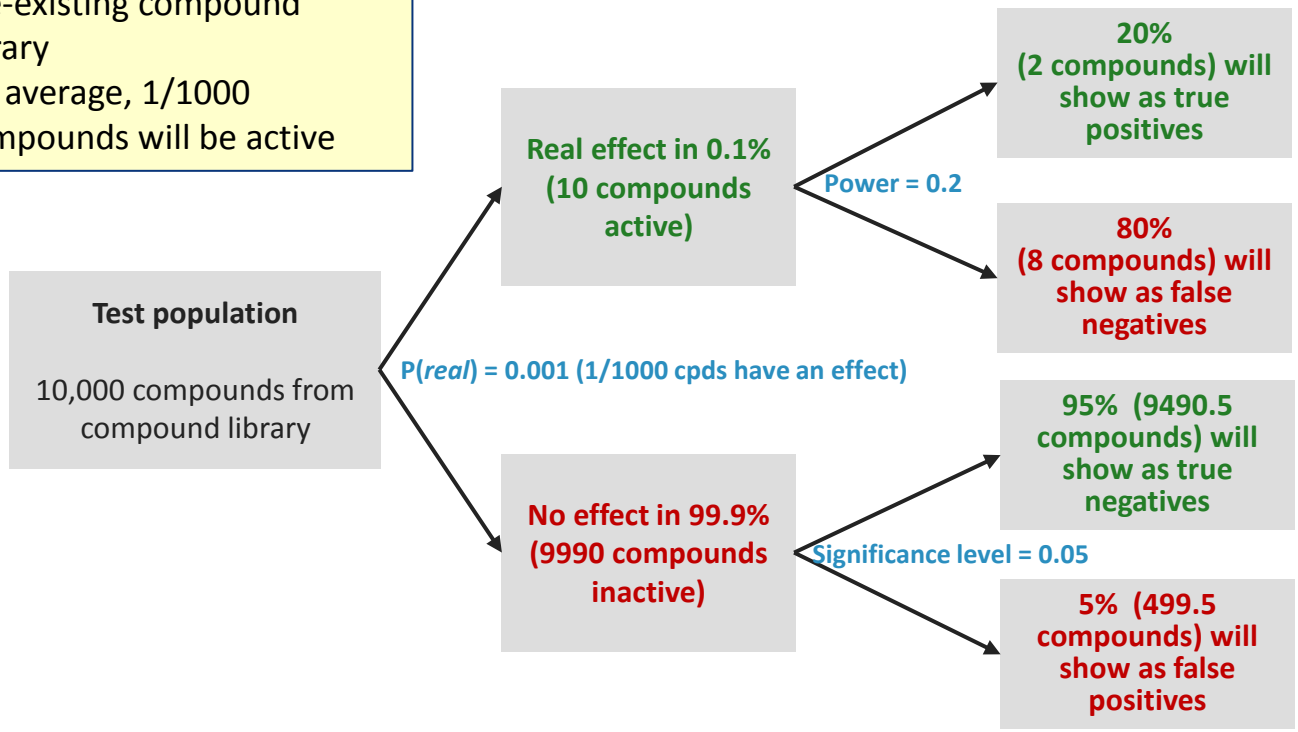
John P. A. Ioannidis

- A research finding is less likely to be true in case of
  - Small studies
  - Small effect sizes
  - Many statistical comparisons, less pre-defined criteria
  - Greater flexibility in designs, definitions, outcomes, and analytical modes

Example: Phenotypic Screen

Assumptions:

- Pre-existing compound library
- On average, 1/1000 compounds will be active



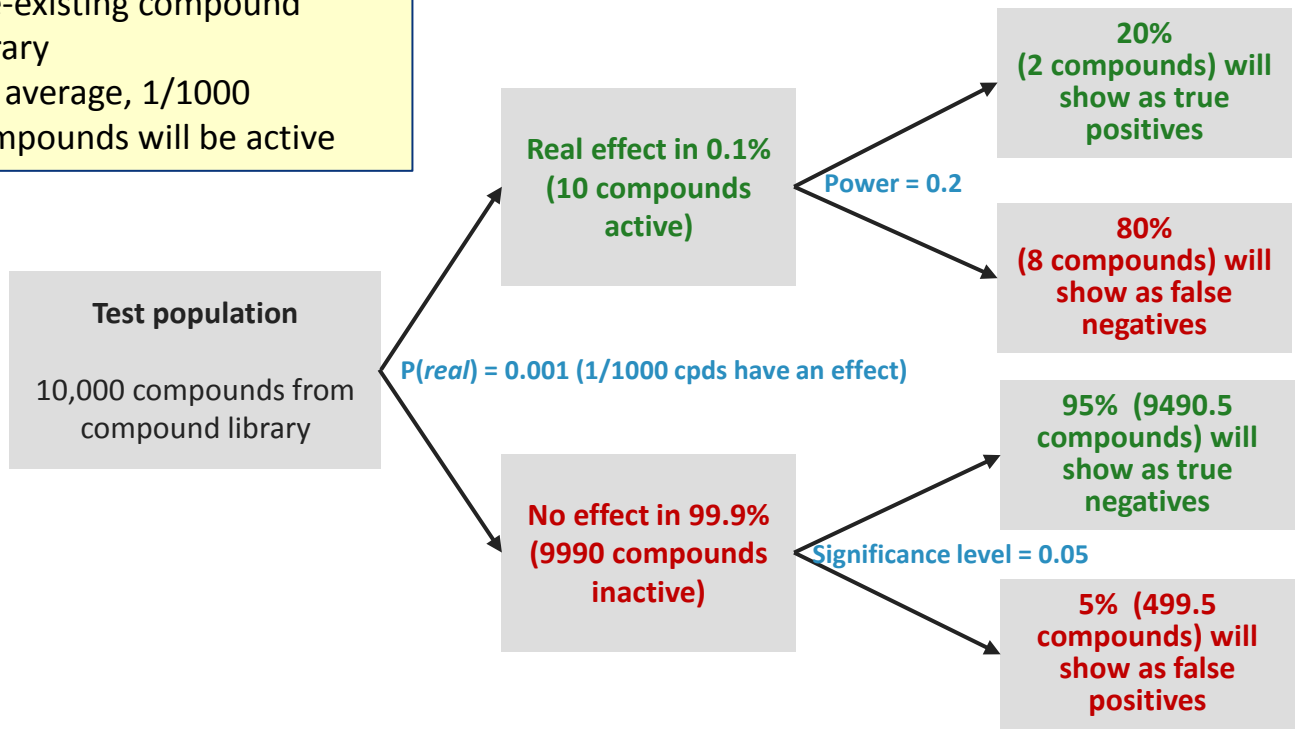
What is the probability that there is no efficacy even if a "significant" result has been obtained?

FDR = False Positives ( )

Example: Phenotypic Screen

Assumptions:

- Pre-existing compound library
- On average, 1/1000 compounds will be active



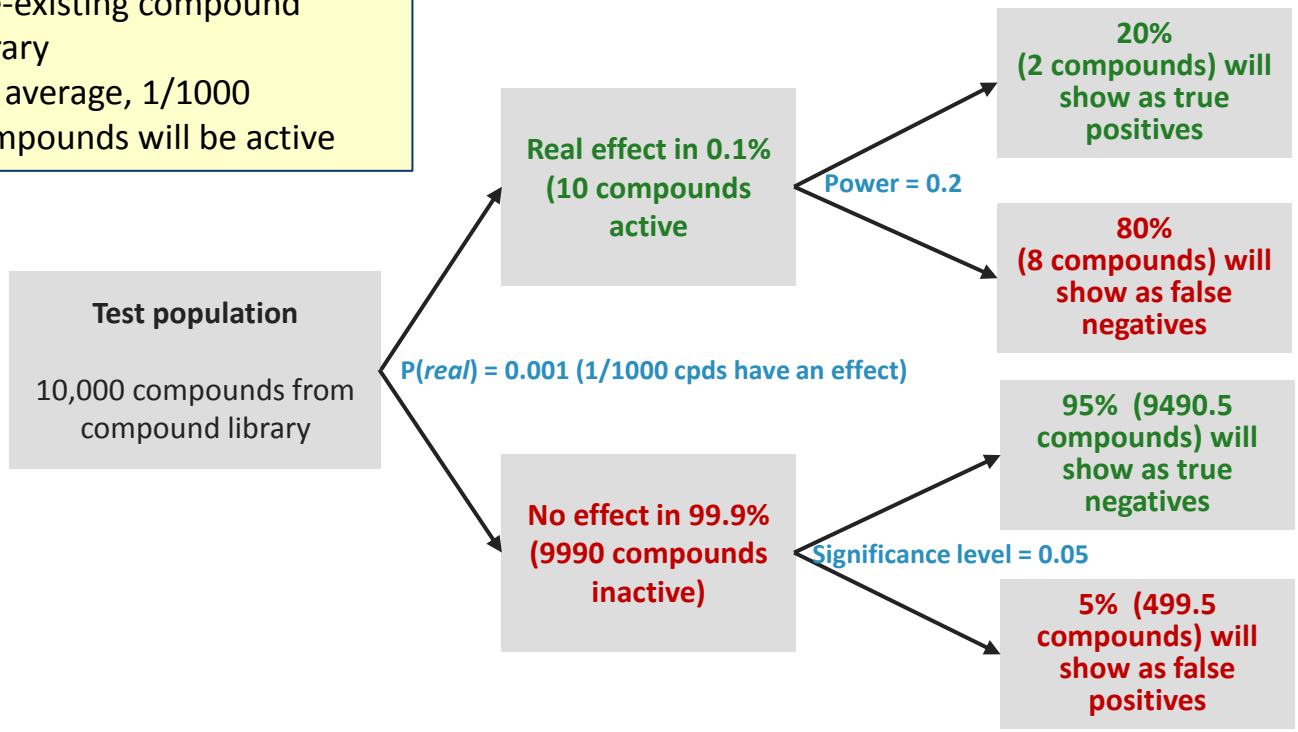
What is the probability that there is no efficacy even if a "significant" result has been obtained?

FDR = False Positives (499.5) / All Positives ( )

Example: Phenotypic Screen

Assumptions:

- Pre-existing compound library
- On average, 1/1000 compounds will be active



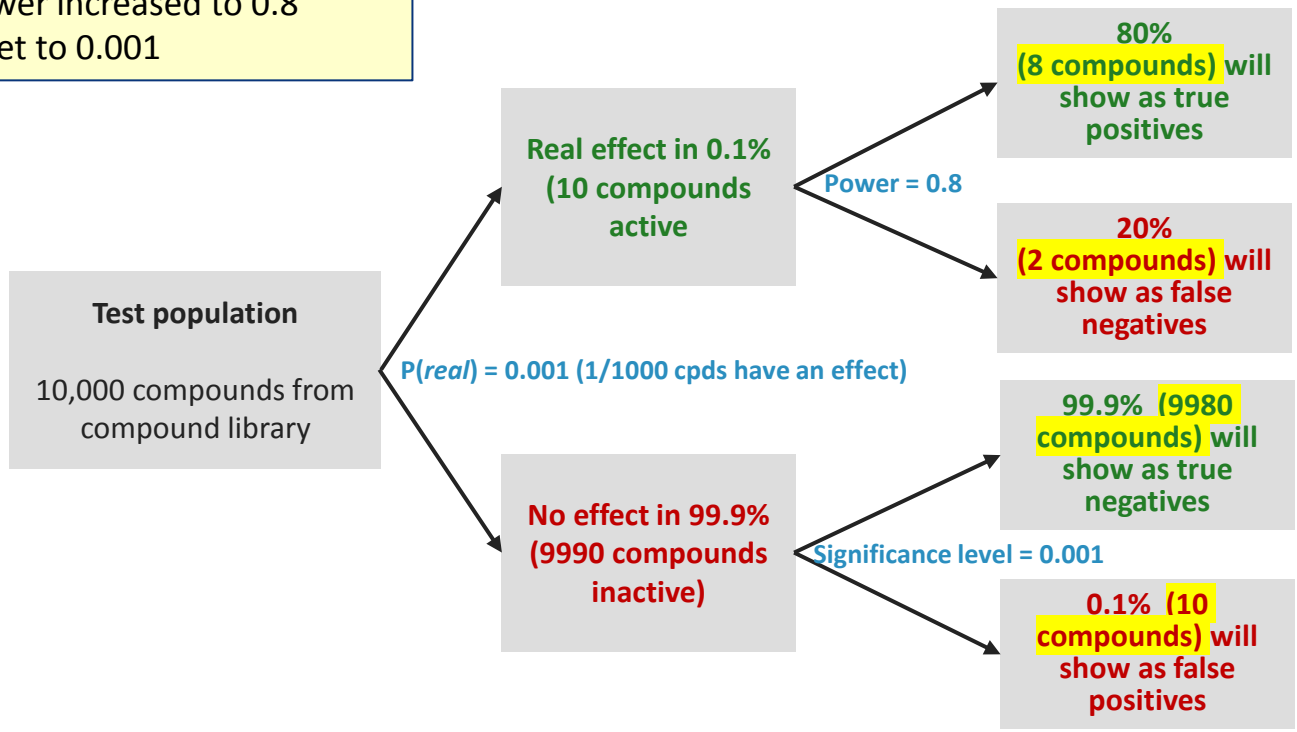
What is the probability that there is no efficacy even if a "significant" result has been obtained?

FDR = False Positives (499.5) / All Positives (499.5+2) = 0.99.6 (99.6%)

→ > 99% of compounds are falsely detected as "active"

Making Statistics More Robust

Assumptions:
- Power increased to 0.8
- α set to 0.001



What is the probability that there is no efficacy even if a "significant" result has been obtained?

FDR = False Positives (10) / All Positives (10 + 8) = 0.55 (55%)

→ 55% of compounds are falsely detected as "active"

Example Cell Culture

- Cells suspended and pipetted into wells of microtiter plate, treatment randomized to wells, replicates on multiple days

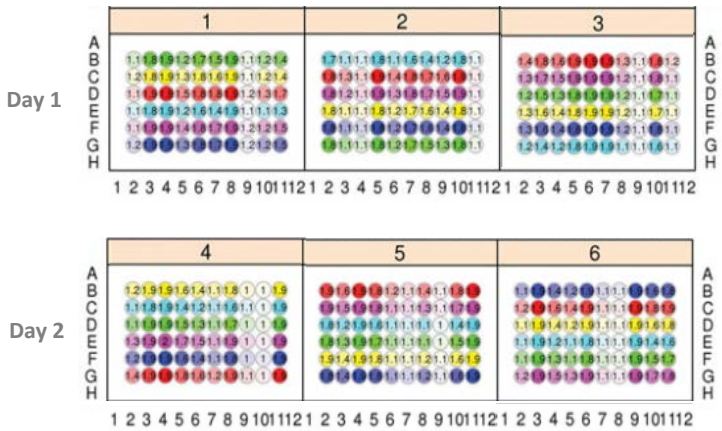
  → EU: the well, not the individual cells (possible)

  → the replicates, not the wells (better!)

  could be well plates run on one day would be subsamples

  improves robustness / consistency / generalizability

  becomes a scientific judgement!



- Experimental material (cells) are artificially homogenous
- Experimental conditions very narrowly defined

Multiple Interventions 1

- **Example 1: New drug against vehicle and multiple active comparators**
  Q: do drugs differ from vehicle and when compared to each other?

- Multiple (m) interventions, independent samples, multiple comparisons (comparing more then 2 means)

- Stats: curve fitting if possible (e.g., to calculate and compare EC_{50}'s)
  often not possible →ANOVA (overall effect), *post-hoc* test (pairwise comparisons of all conditions with each other)

- Sample size calculation similar to unpaired T-test
  - Take desired effect size
  - Calculate population variance
  - Use Z-values for T-test

Balanced design, comparisons estimated best if same number of samples are allocated per group

Multiple Interventions 2

- **Example 2: Dose-response study with planned comparisons**
  Q: is drug X active when compared to vehicle?

- Multiple (m) interventions , independent samples, 1 control, pairwise comparisons

- Stats: ANOVA (overall effect), planned comparisons (individual drug doses vs. vehicle)

- Sample size calculation similar to unpaired T-test

- **But:** Number of samples in control group should be $\sqrt{m}$ – times the number of samples in intervention groups

(Bate & Karp, PLOS One, 2014)

$$N_c = \sqrt{m} * n_m$$

with N_c = adjusted control group size
m = number of interventions (e.g. drug doses tested)
n_m = calculated sample size

Unbalanced design, gains sensitivity at cost of multiple comparisons

Multiple Interventions 3

- **Example 3: Comparison of 2 drugs, all samples receive all treatments**
  Q: is one drug better than the other drug?
- Multiple (m) interventions , dependent samples
- Stats: Cross-over design, repeated measures ANOVA
- Sample size calculation uses same formula as for unpaired T-test
- **But:** Variability used is the within sample standard deviation σ_w

$\sigma_w = \sqrt{(WMSE)}$, with WMSE = within mean square error from the ANOVA table

ANOVA

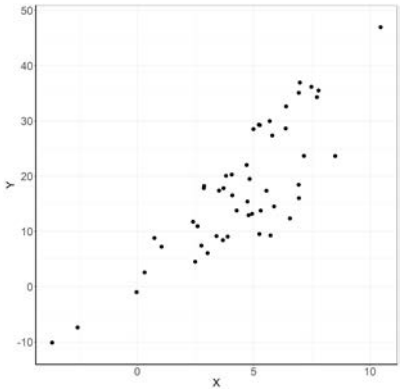
	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	31.444	2	15.722	7.447	.006
Within Groups	31.667	15	2.111		
Total	63.111	17			

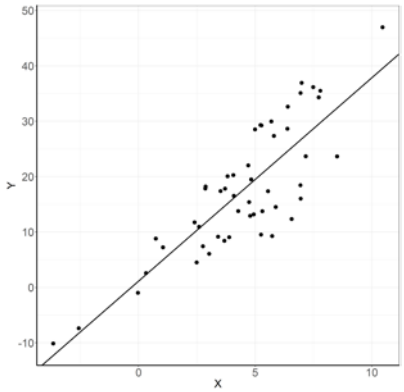
Special Cases and Additional Requirements

- Large sample size due to high variability
  - → consider covariates
  - → consider blocking factors
- Anticipated dropouts
  - → add samples
- Studies designed to show lack of effect (zero-data)?
  - → need high power and variability (95% CI) covers effect size too small to be considered relevant
- Other statistical approaches
  - → **ask your statistician!**

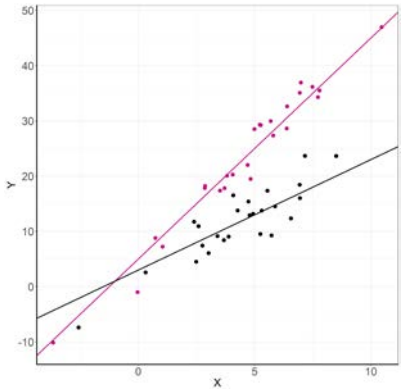
How to Deal with…

- High variability?
  - Consider using **covariates**





$Y = \beta X_1 + \varepsilon$

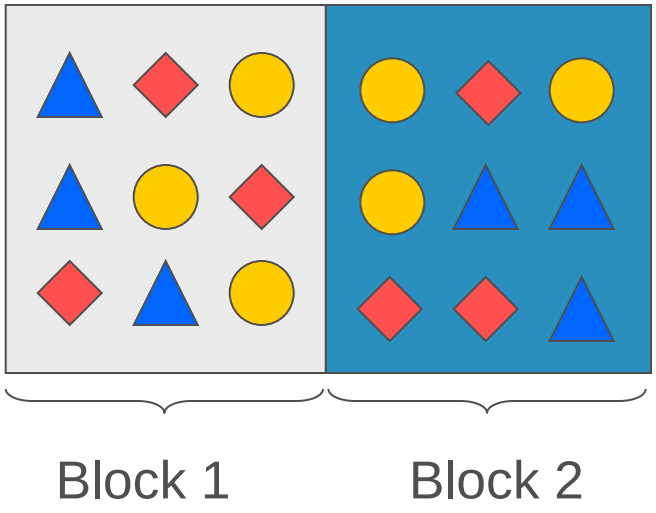


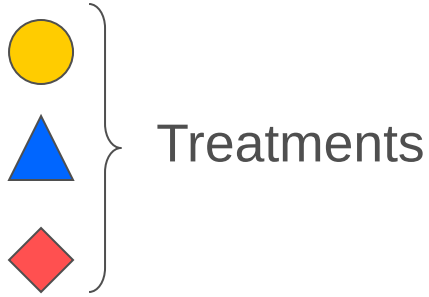
$Y = \beta_0 + \beta X_1 + \beta X_2 + \ldots + \varepsilon$

with Y= measurement, X1 = explanatory variable and ε = error term

How to Deal with…

- High variability?
  - Consider **blocking** factors
    - A variable that has an effect on an experimental outcome, but is itself of no interest
    - Age, gender, experimenter, time of day, batches…
  - Use block designs to reduce unexplained variability
  - Key concept: variability within each block is less than the variability of the entire sample → increasing efficiency to estimate an effect





Experimental units in one block are more 'alike' than others

How to Deal with…

- **Anticipated dropouts?**
- Consider need for additional samples!
  - E.g. x% reduced viability of offspring from mutant mouse line

$$N_1 = n/(1\text{-}\ (z/100))$$

with N_1 = adjusted sample size

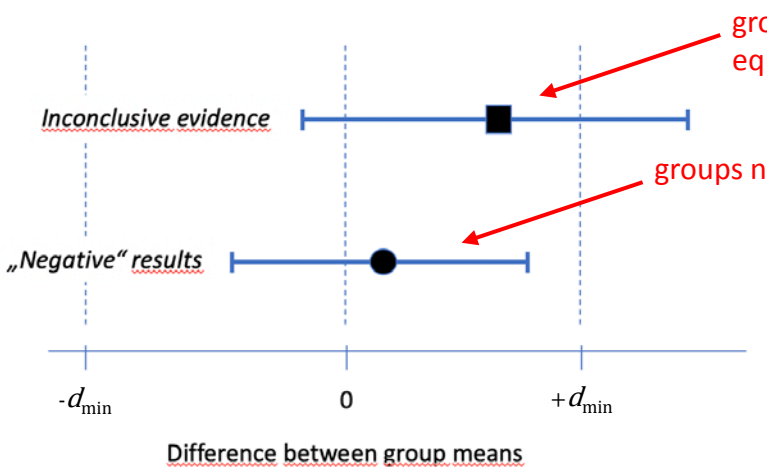
n = calculated sample size

z = anticipated dropout rate (%)

How to Deal with…

- Studies designed to show lack of effect (zero-data)?
  - Example: Attempt to confirm published work
- Minimize false negatives! → Power >> .80!
- Demonstrate result is conclusive! → Show effect size is less than originally reported

A non-significant result may or may not be conclusive



if 95% confidence interval overlaps with important effect we cannot exclude possibility that there is an important impact

Modified from: Bespalov et al., unpublished

What is convincingly negative?

High research rigor

- Robust and unbiased experimental design, methodology, analysis, interpretation, and reporting
- Authors of original paper consulted (if possible)

Properly validated methods

No evidence for technical failure

Converging evidence on multiple readouts (if possible)

Ideally multi-lab collaborative effort

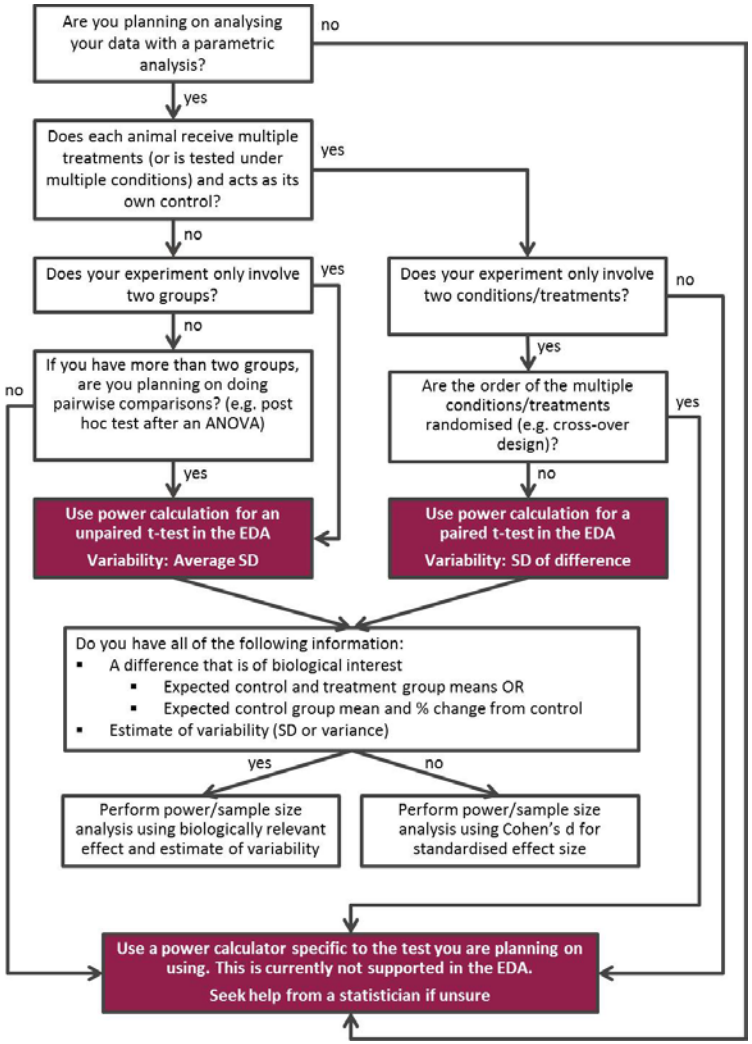
Adequate data analysis methods establish the observed results as statistically negative

- No experiment can provide an absolute proof of absence of an effect, i.e., $p > 0.05$ does not proof negative results
- Often more rigorous design and stronger statistical power required than in the original report (power of 0.2 is not sufficient!)
- Confidence intervals should be narrow and cover only those effect sizes that aren't considered scientifically relevant

Full access to methods and raw data are provided

Results likely to have a significant impact if published

Choosing the Appropriate Power Calculation



https://eda.nc3rs.org.uk/experimental-design-group

Additional Strategies to Increase Power

- Use as few treatment groups as possible
- Investigate only main effects rather then interactions
- Use direct rather than indirect dependent variables
- Use sensitive measures
- Use reliable measures
- Use covariates and/or blocking variables
- Use cross-over designs