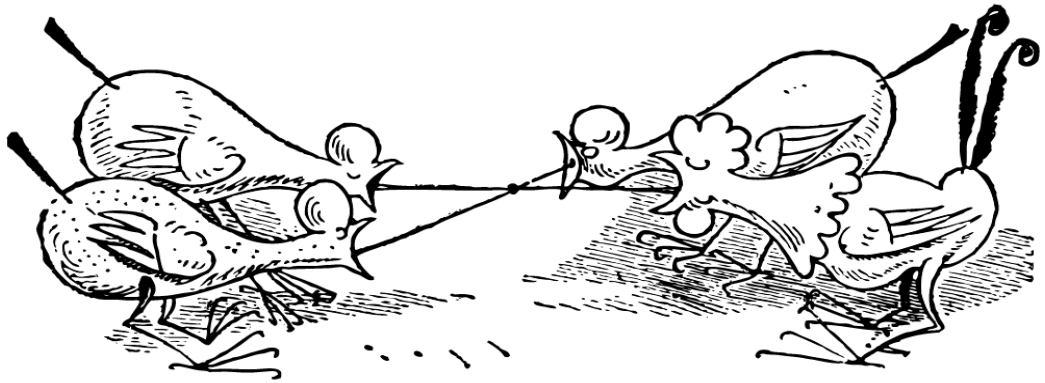
How Many Mice Make Robust Outcomes?

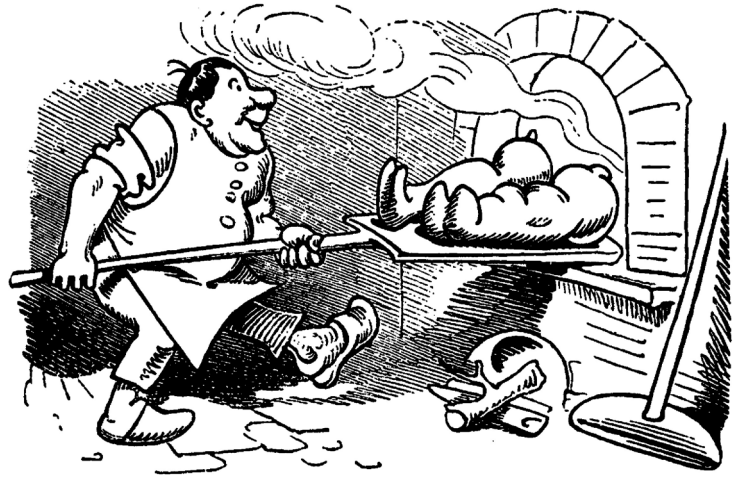
Bernhard Voelkl
Animal Welfare Division, University of Bern

Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

Niemand darf ungerechtfertigt einem Tier Schmerzen, Leiden oder Schäden zufügen, es in Angst versetzen oder in anderer Weise seine Würde missachten. Das Misshandeln, Vernachlässigen oder unnötige Überanstrengen von Tieren ist verboten. (TSchG Art.4, Z.2)

No one shall unjustifiably cause pain, suffering, or harm to an animal, place it in fear, or otherwise disregard its dignity. Abusing, neglecting or unnecessarily overexerting animals is prohibited.





Tierschutzgesetz
(TSchG)

vom 16. Dezember 2005 (Stand am 1. Januar 2022)

Die Bundesversammlung der Schweizerischen Eidgenossenschaft,
gestützt auf die Artikel 80 Absätze 1 und 2 sowie 120 Absatz 2
der Bundesverfassung[1],
nach Einsicht in die Botschaft des Bundesrates vom 9. Dezember 2002[2],
beschliesst:

1. Kapitel: Allgemeines

Art. 1 Zweck

Zweck dieses Gesetzes ist es, die Würde und das Wohlergehen des Tieres zu schützen.

Art. 2 Geltungsbereich

[1] Das Gesetz gilt für Wirbeltiere. Der Bundesrat bestimmt, auf welche wirbellosen Tiere es in welchem Umfang anwendbar ist. Er orientiert sich dabei an den wissenschaftlichen Erkenntnissen über die Empfindungsfähigkeit wirbelloser Tiere.

[2] Vorbehalten bleiben das Jagdgesetz vom 20. Juni 1986[3], das Bundesgesetz vom 1. Juli 1966[4] über den Natur- und Heimatschutz, das Bundesgesetz vom 21. Juni 1991[5] über die Fischerei, das Berufsbildungsgesetz vom 13. Dezember 2002[6] sowie das Tierseuchengesetz vom 1. Juli 1966[7].

Art. 3 Begriffe

In diesem Gesetz bedeuten:

a. *Würde*: Eigenwert des Tieres, der im Umgang mit ihm geachtet werden muss. Die Würde des Tieres wird missachtet, wenn eine Belastung des Tieres nicht durch überwiegende Interessen gerechtfertigt werden kann. Eine Belastung liegt vor, wenn dem Tier insbesondere Schmerzen, Leiden oder Schäden zugefügt werden, es in Angst versetzt oder erniedrigt wird, wenn

Bundesverfassung der Schweizerischen Eidgenossenschaft

vom 18. April 1999 (Stand am 13. Februar 2022)

-
-
-

Art. 18 Sprachenfreiheit

Die Sprachenfreiheit ist gewährleistet.

Art. 19 Anspruch auf Grundschulunterricht

Der Anspruch auf ausreichenden und unentgeltlichen Grundschulunterricht ist gewährleistet.

Art. 20 Wissenschaftsfreiheit

Die Freiheit der wissenschaftlichen Lehre und Forschung ist gewährleistet.

Art. 21 Kunstfreiheit

Die Freiheit der Kunst ist gewährleistet.

Art. 22 Versammlungsfreiheit

[1] Die Versammlungsfreiheit ist gewährleistet.

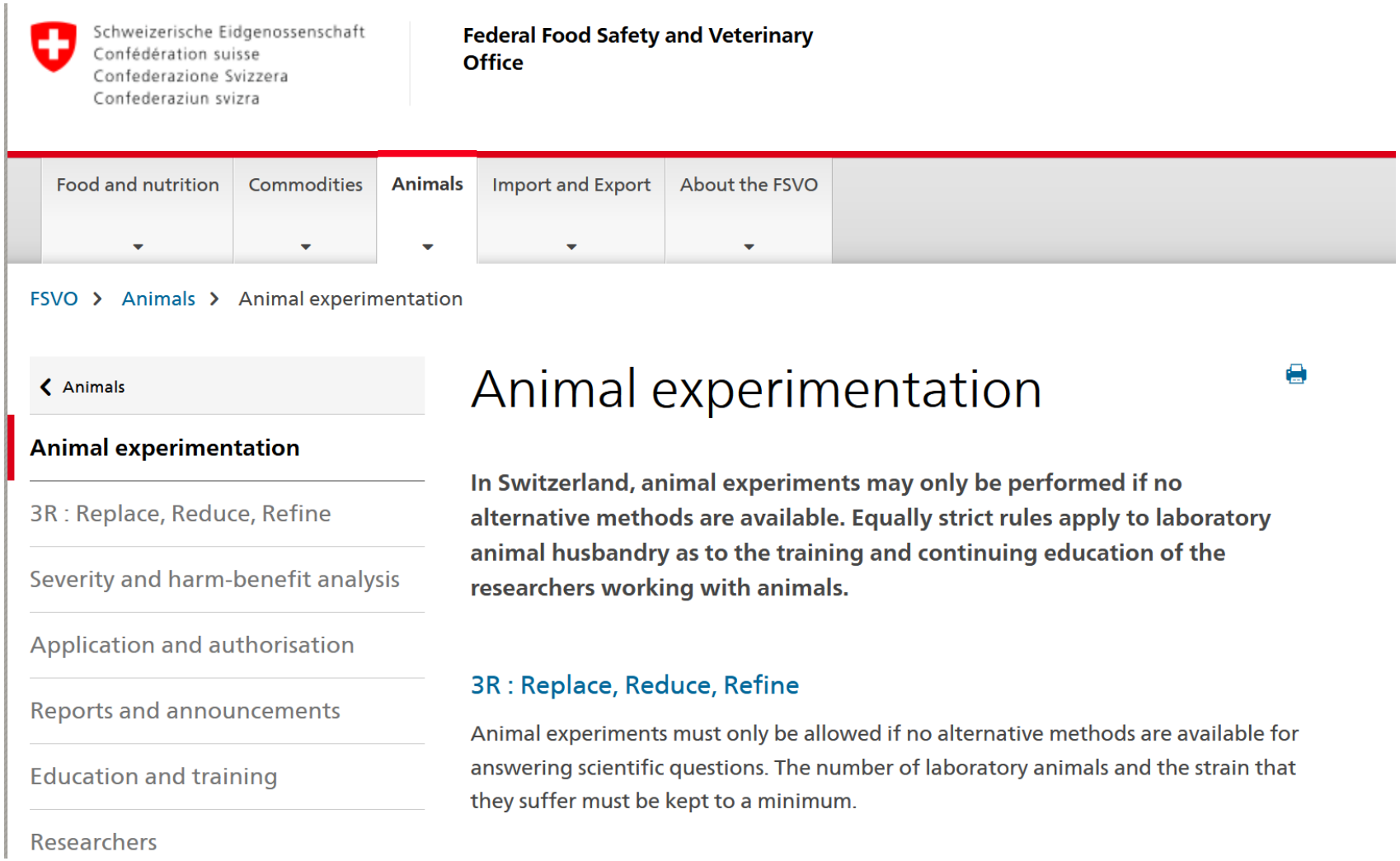
[2] Jede Person hat das Recht, Versammlungen zu organisieren, an Versammlungen teilzunehmen oder Versammlungen fernzubleiben.

Art. 23 Vereinigungsfreiheit

[1] Die Vereinigungsfreiheit ist gewährleistet.

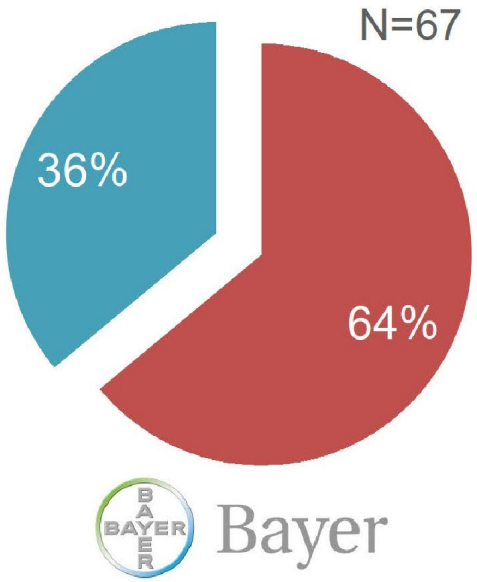
[2] Jede Person hat das Recht, Vereinigungen zu bilden, Vereinigungen beizutreten oder anzugehören und sich an den Tätigkeiten von Vereinigungen zu beteiligen.

[3] Niemand darf gezwungen werden, einer Vereinigung beizutreten oder anzugehören.

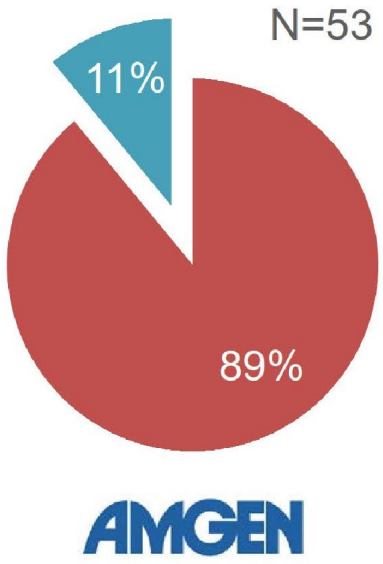


How Many Mice Make Robust Outcomes?

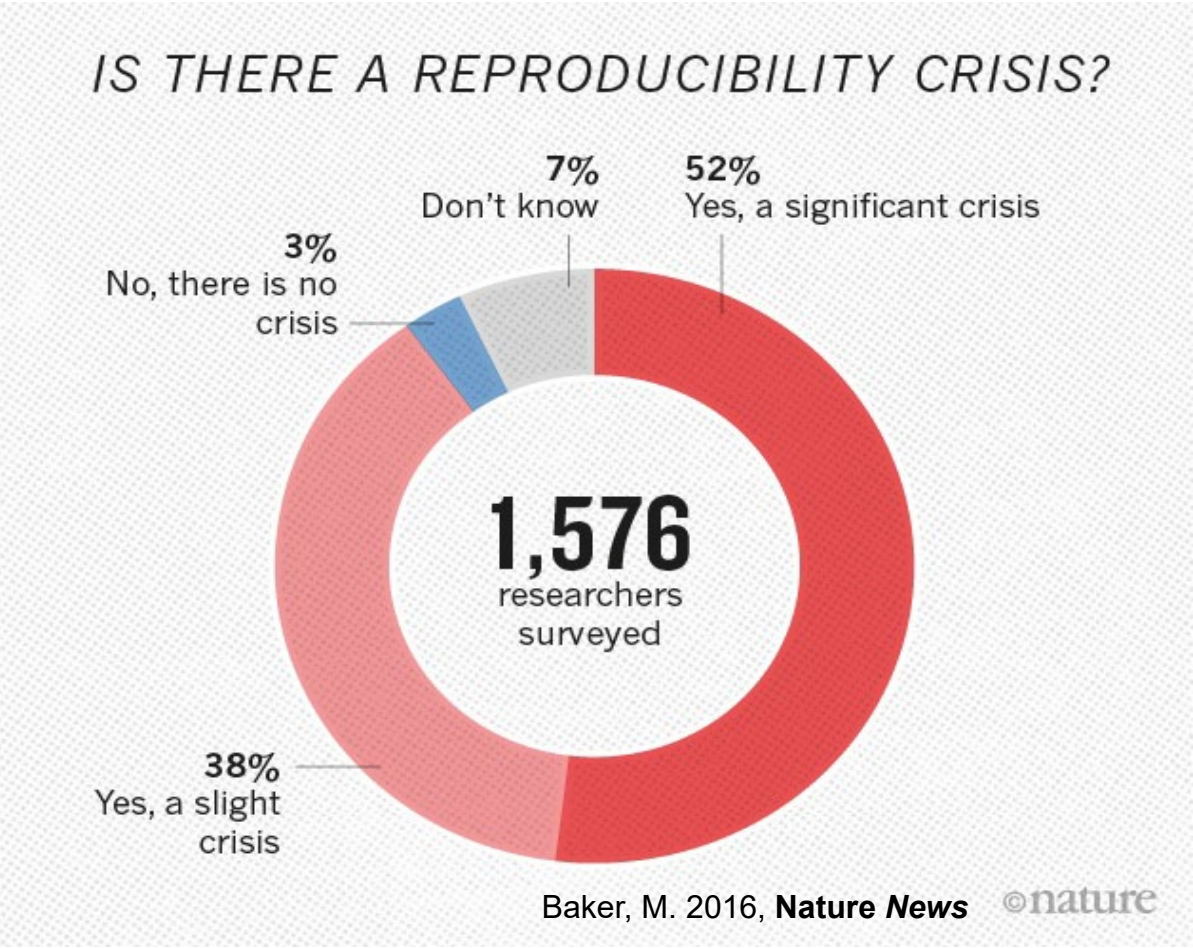
Replication of key studies in biomedical research



Prinz et al. 2011, **Nature Reviews Drug Discover**

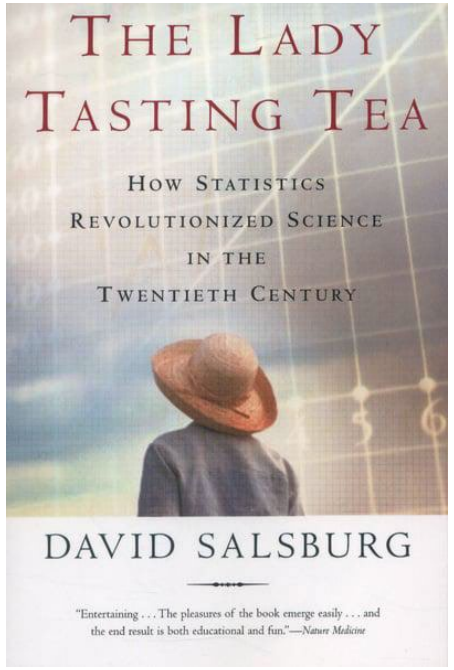


Begley & Ellis 2012, **Nature**



Baker, M. 2016, **Nature *News***

Is our way of hypothesis testing flawed ?

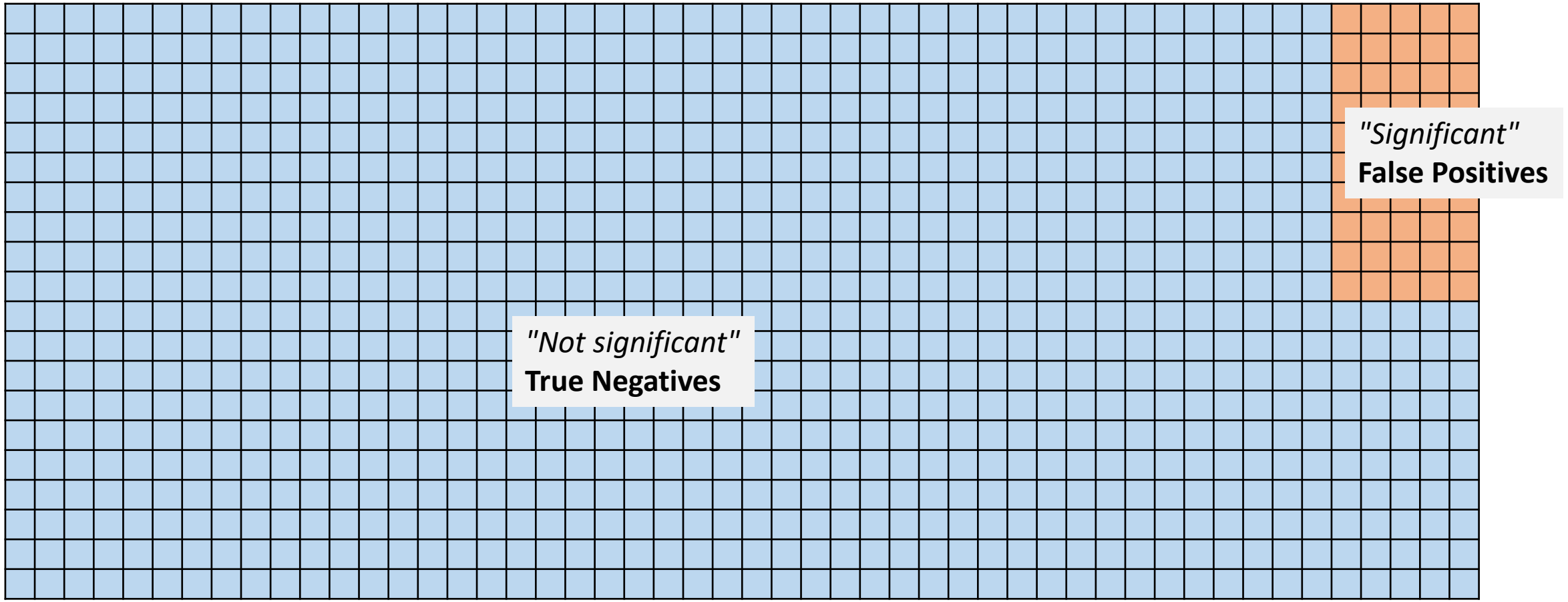


A simplified version of the Neyman-Pearson formulation of hypothesis testing [..] has been accepted by regulatory agencies [..] and is taught in medical schools to future medical researchers [..] When the Neyman-Pearson formulation is taught in this rigid, simplified version of what Neyman developed, it distorts his discoveries by concentrating on the wrong aspects of the formulation.

D. Salsburg

1000 Hypotheses tested: assuming all hypotheses are false (i.e. no effects)

.. setting alpha = 0.05 means: only 5% will be false positives



alpha-error:

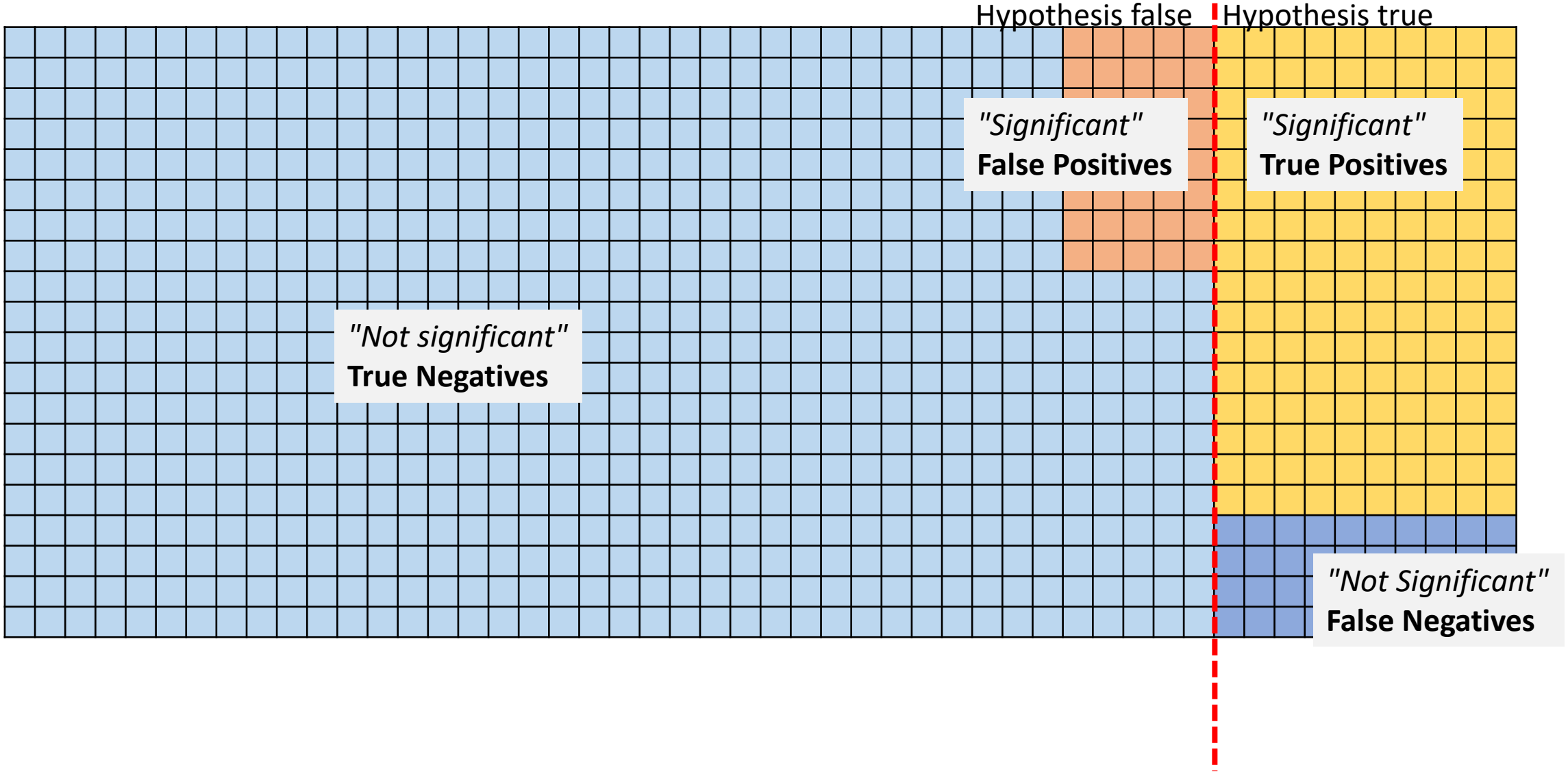
… if all hypothesis are wrong, how often will a test will give a positive result ?

$$P(\text{test positive} \mid \text{hypothesis false})$$

conditional probability

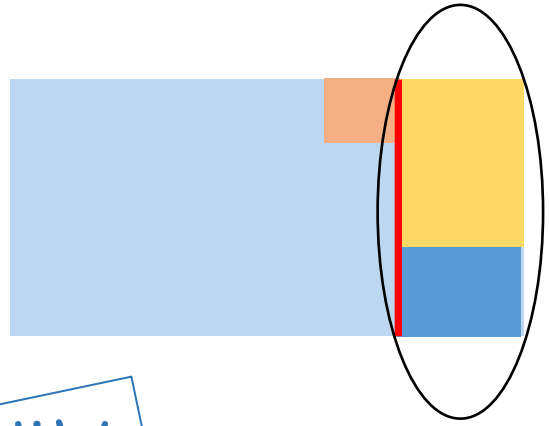
1000 Hypotheses tested: assuming 20 % of hypotheses are true

alpha = 0.05, power = 0.8



Power: = 1 – beta error

...**assuming** all hypotheses are true, how often will we get a positive test result?



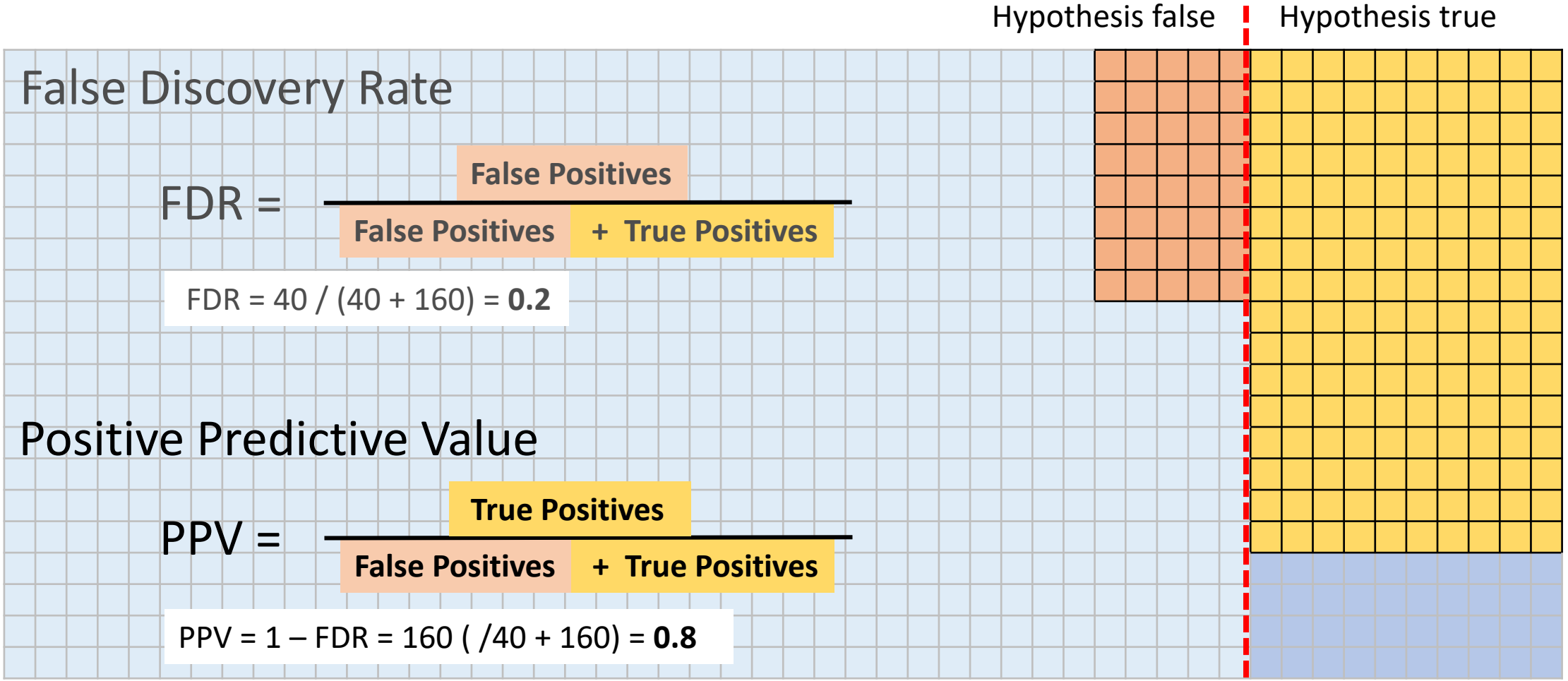
... **if** all hypotheses are true, what is the probability that the test will give a positive result ?

P(test positive) | hypothesis true

conditional probability

1000 Hypotheses tested: assuming 20 % of hypotheses are true

alpha = 0.05, power = 0.8



False Discovery Rate

$$FDR = \frac{\text{False Positives}}{\text{False Positives} + \text{True Positives}}$$

FDR = 40 / (40 + 160) = **0.2**

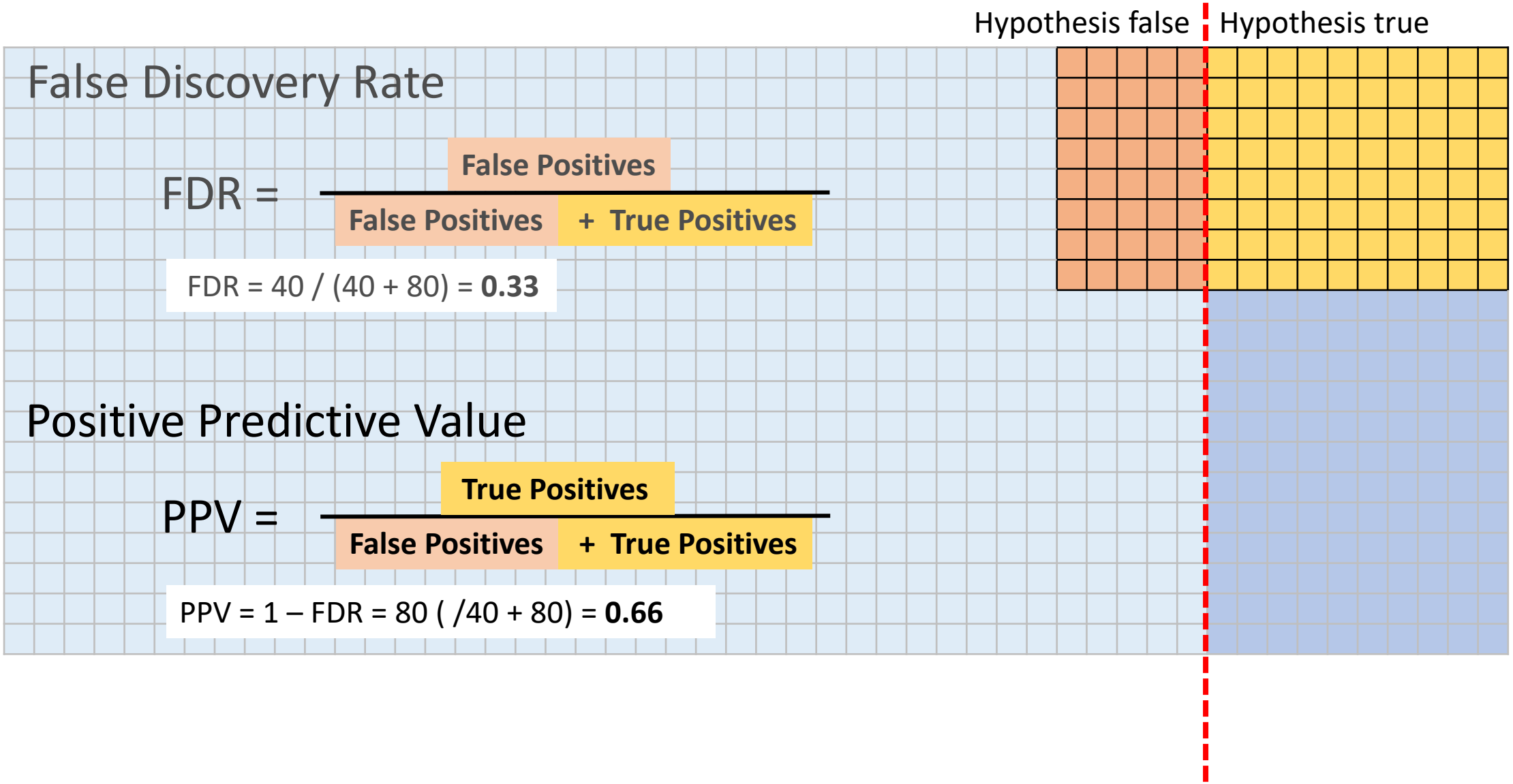
Positive Predictive Value

$$PPV = \frac{\text{True Positives}}{\text{False Positives} + \text{True Positives}}$$

PPV = 1 – FDR = 160 (/40 + 160) = **0.8**

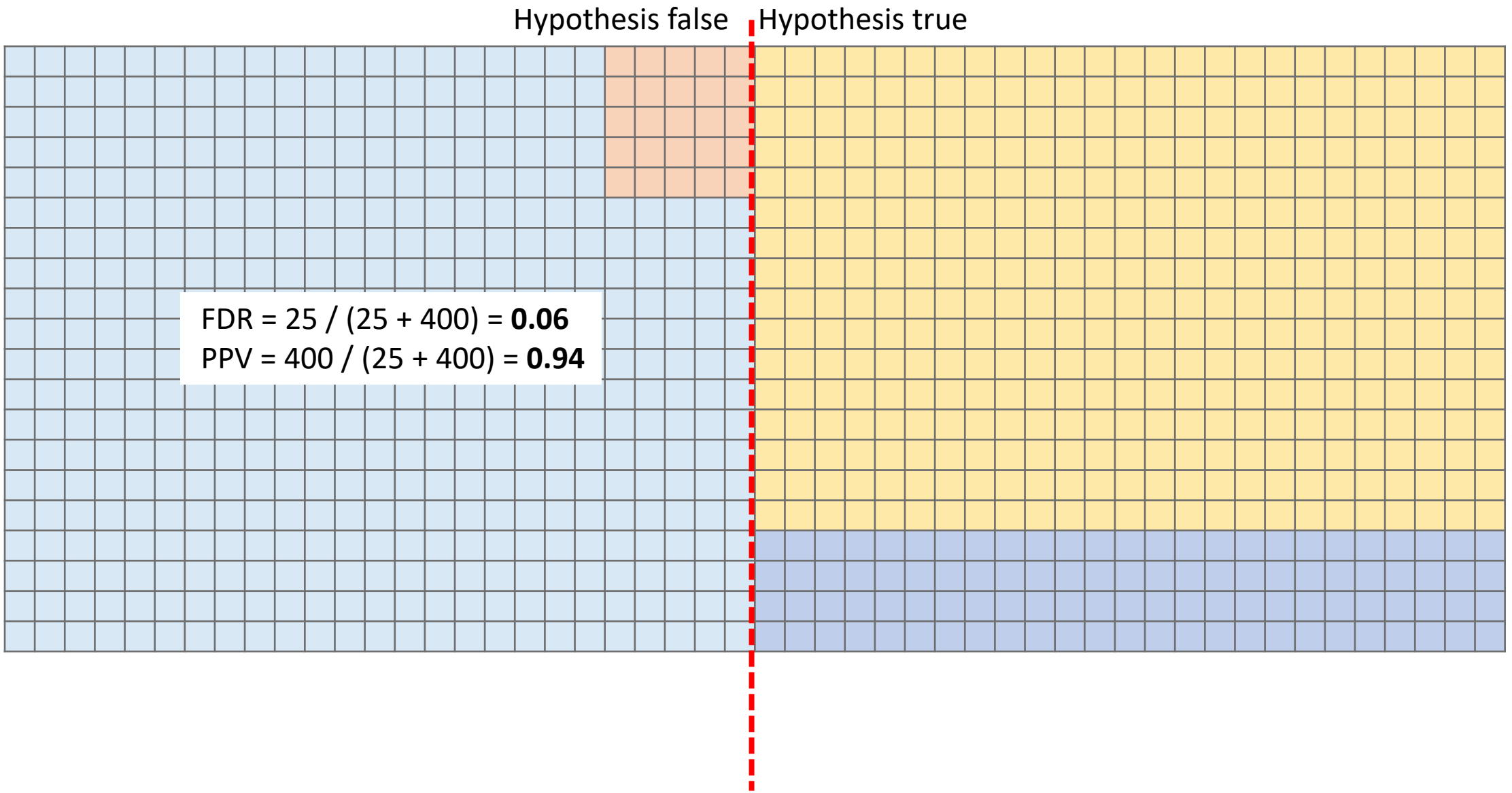
1000 Hypotheses tested: assuming 20 % of hypotheses are true

alpha = 0.05, power = **0.4**



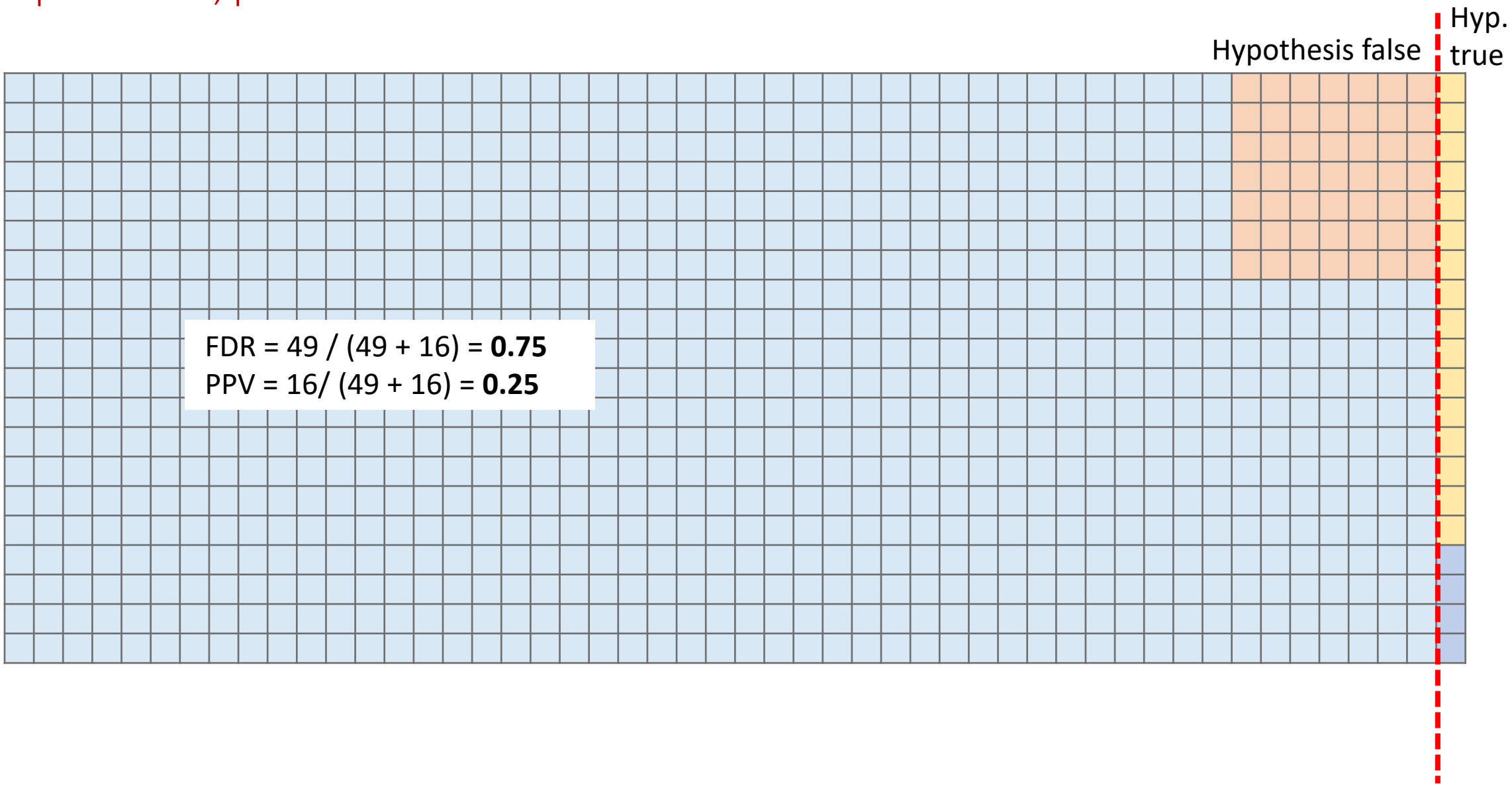
1000 Hypotheses tested: assuming **50 %** of null hypotheses are true

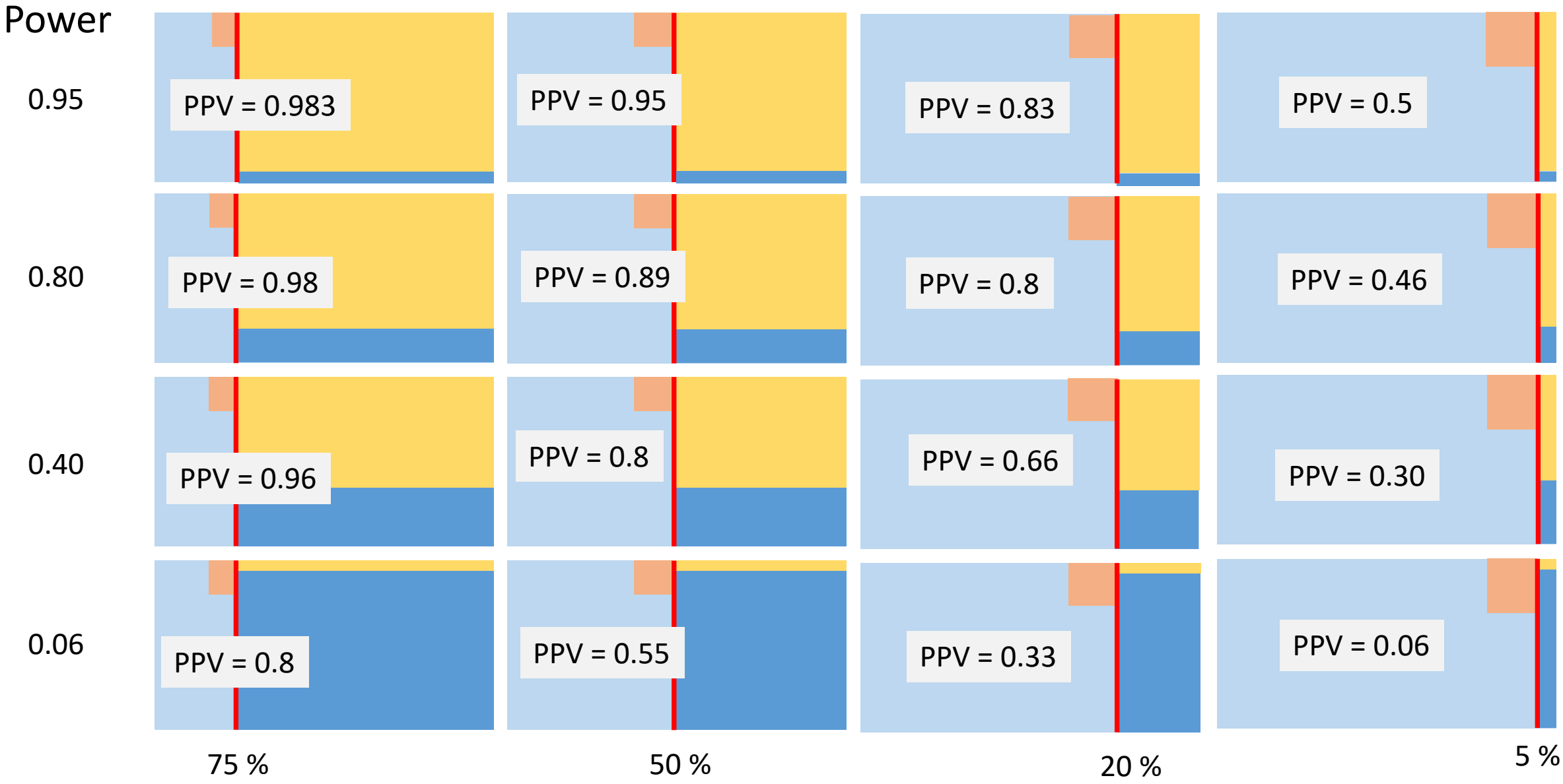
alpha = 0.05, power = 0.8

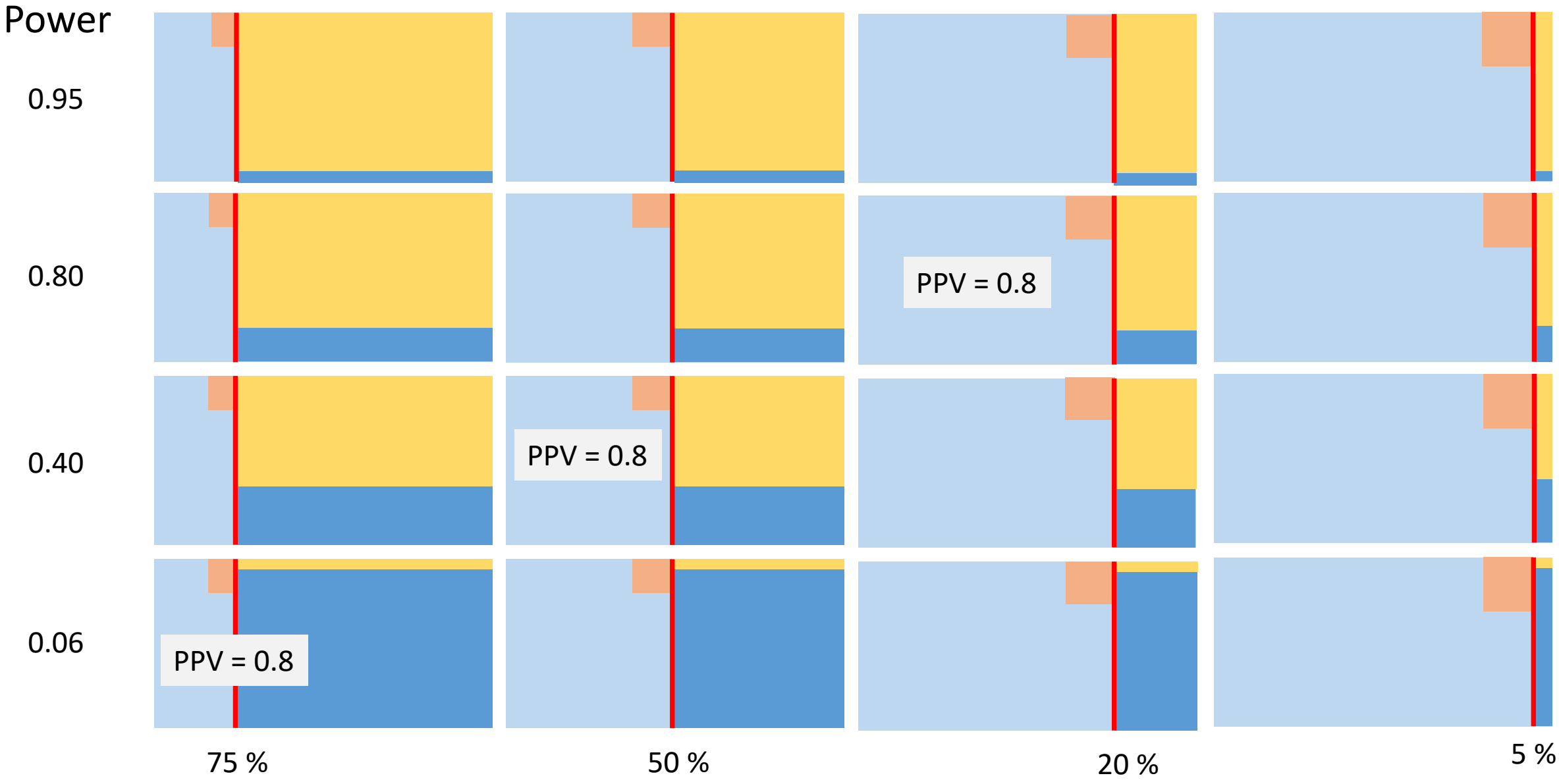


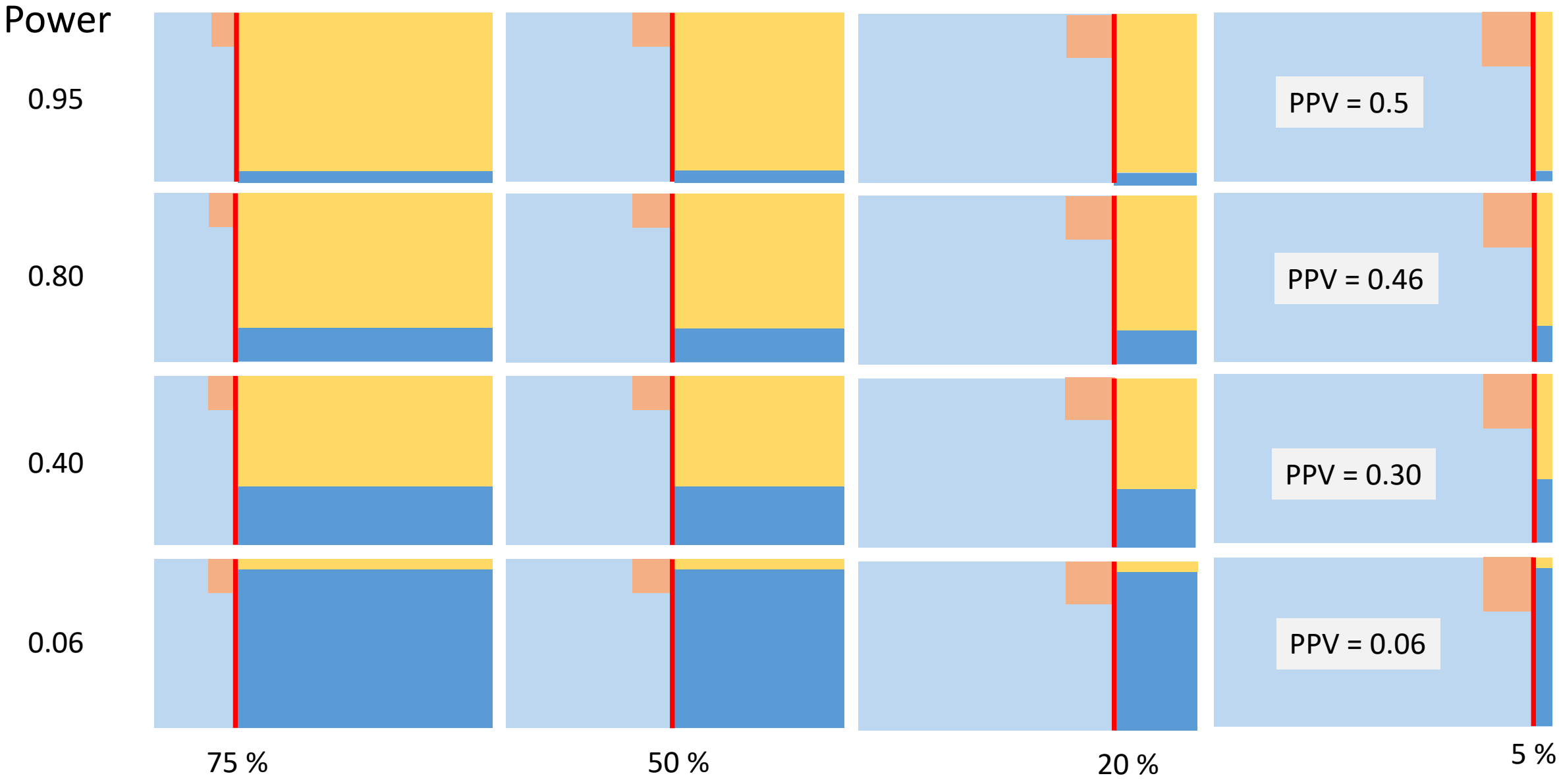
1000 Hypotheses tested: assuming **2 %** of hypotheses are true

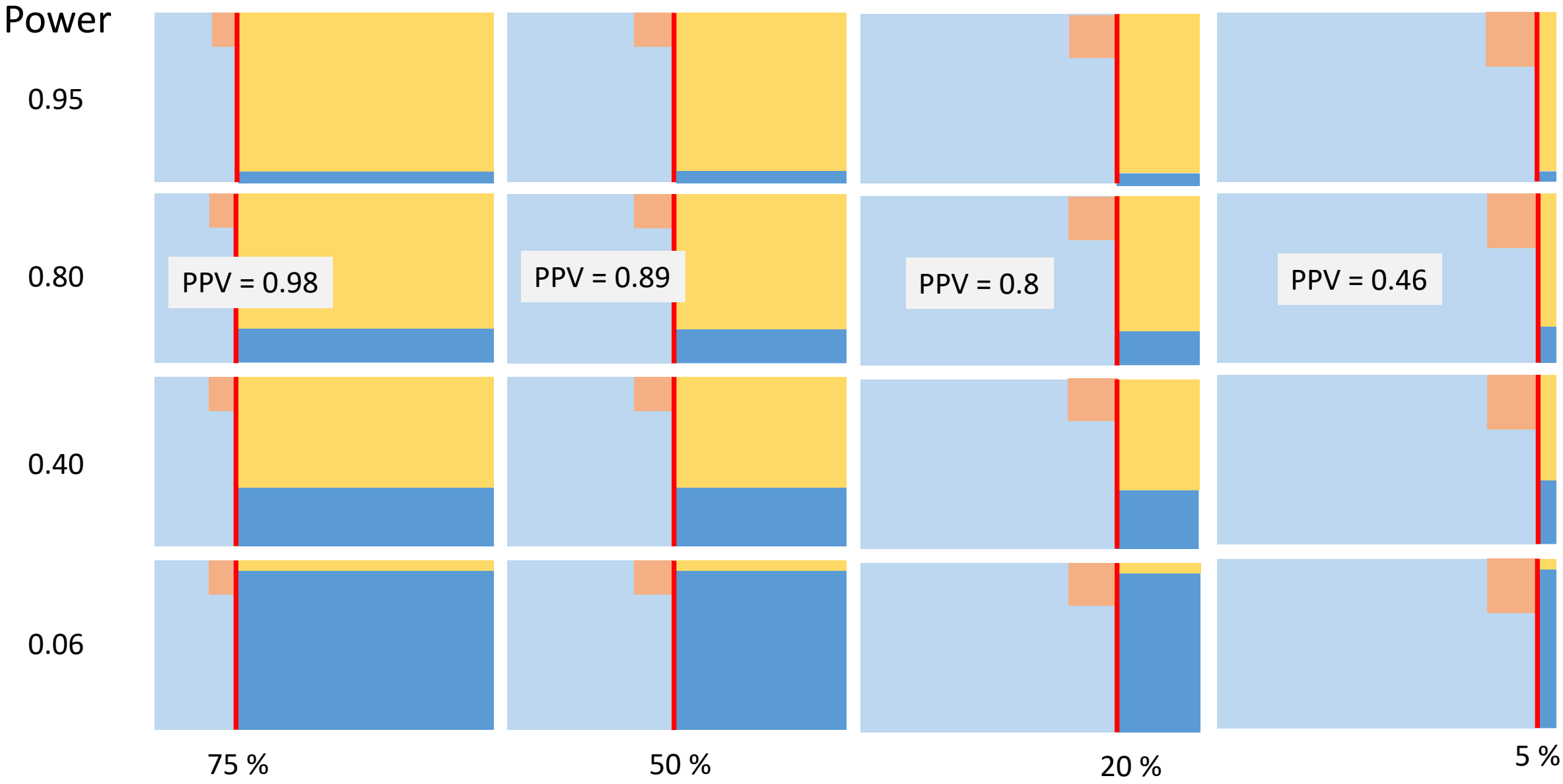
alpha = 0.05, power = 0.8




Power
0.95
0.80
0.40
0.06
PPV = 0.983
PPV = 0.95
PPV = 0.83
PPV = 0.5
PPV = 0.98
PPV = 0.89
PPV = 0.8
PPV = 0.46
PPV = 0.96
PPV = 0.8
PPV = 0.66
PPV = 0.30
PPV = 0.8
PPV = 0.55
PPV = 0.33
PPV = 0.06
75 %
50 %
20 %
5 %

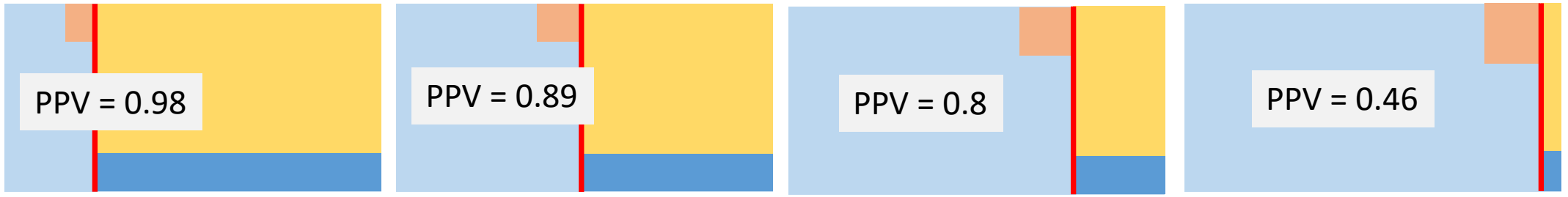

Power
0.95
0.80
0.40
0.06
PPV = 0.8
PPV = 0.8
PPV = 0.8
75 %
50 %
20 %
5 %


Power
0.95
0.80
0.40
0.06
PPV = 0.5
PPV = 0.46
PPV = 0.30
PPV = 0.06
75 %
50 %
20 %
5 %


Power
0.95
0.80
0.40
0.06
PPV = 0.98
PPV = 0.89
PPV = 0.8
PPV = 0.46
75 %
50 %
20 %
5 %

Confirmatory Exploratory

Power
0.80

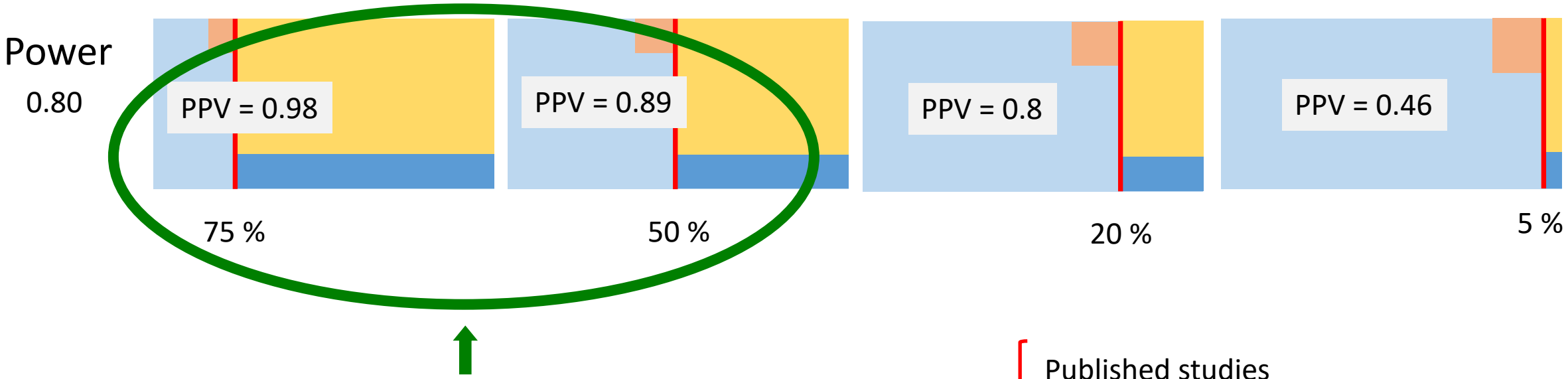


75 % 50 % 20 % 5 %

boring, not really new
almost trivial
we thought that already
as expected

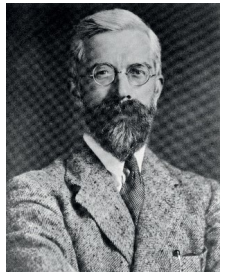
exciting, sexy
big news
surprising discovery
for *Nature*

Confirmatory Exploratory



Hypothesis [/haɪˈpɒθɪsɪs/]: a statement that provides an explanation for why or how something works, **based on facts or some reasonable assumptions,** but that has not yet been specifically confirmed.

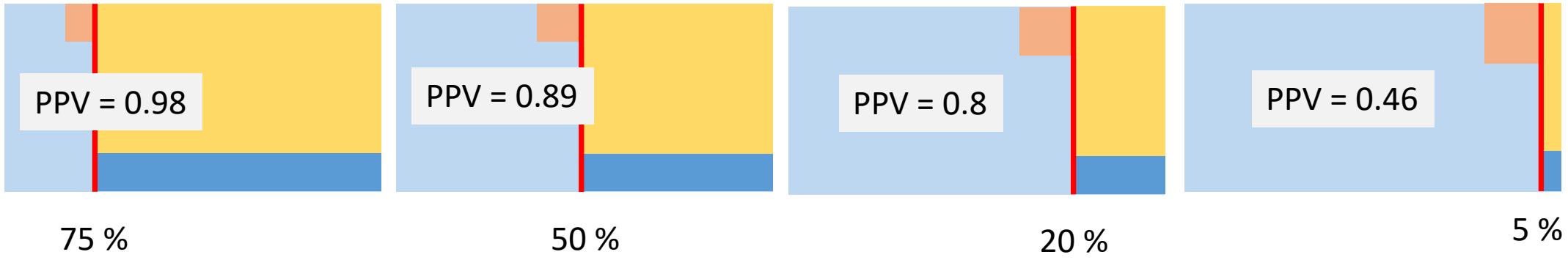
- Published studies
- Theoretical predictions
- Indirect / complementary evidence
- Anecdotal observations
- Exploratory studies



Confirmatory Exploratory

Power

0.80



Take-home message: **Use hypothesis testing only for confirmatory research**

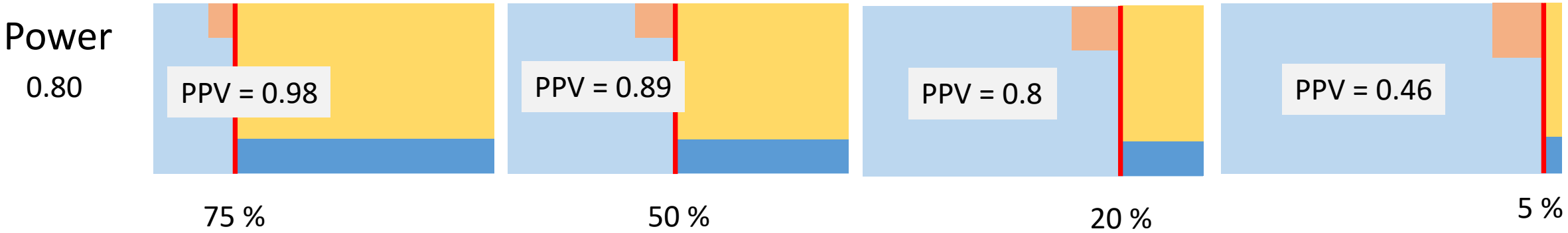
Declare whether your research is confirmatory or exploratory
(Make a statement about the risk you are taking)

Don't calculate p-values when your research is exploratory

Refuse to make a power-analysis when your research is exploratory

.. *but:* How Many Mice Make Robust Outcomes?

$$\text{power} \propto \frac{Es \times \alpha \times \sqrt{n}}{\sigma}$$



How to estimate the required sample size ?

If we have evidence from previous experiments how large the expected effect size (ES) and the expected variance (s.d.) will be, we can calculate the required sample size for a given power level.

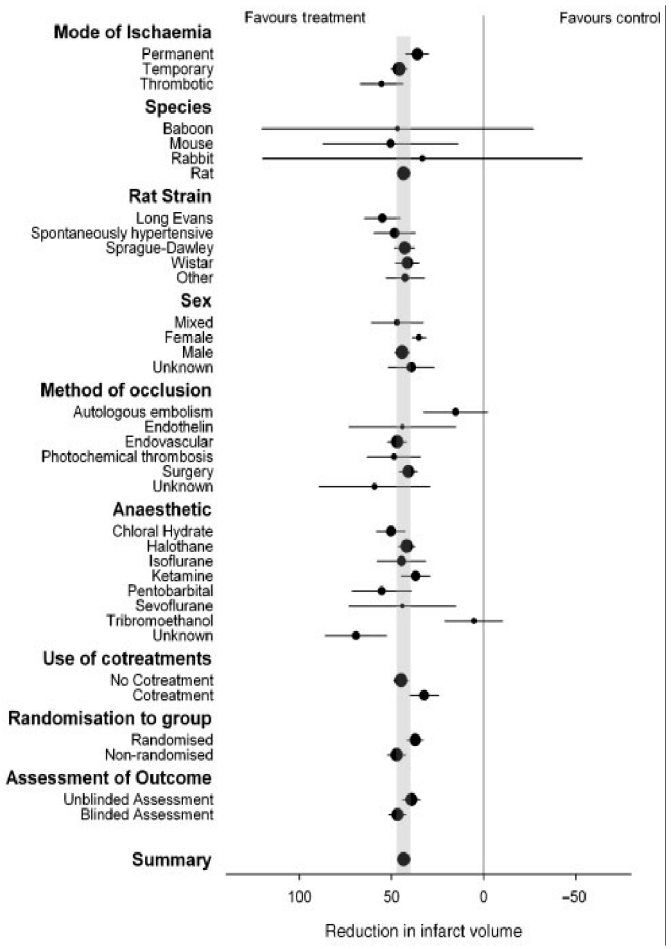
$$\text{power} \propto \frac{Es \times \alpha \times \sqrt{n}}{\sigma}$$

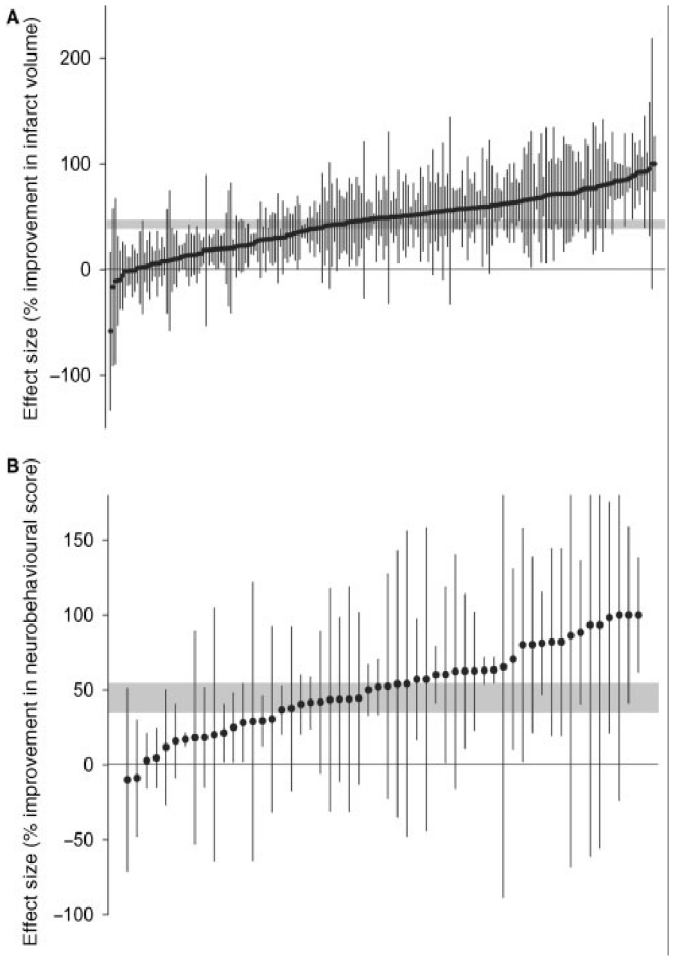
doi:10.1093/brain/awm083 *Brain* (2007), **130**, 3063–3074

REVIEW ARTICLE

Hypothermia in animal models of acute ischaemic stroke: a systematic review and meta-analysis

H. Bart van der Worp,[1] Emily S. Sena,[2] Geoffrey A. Donnan,[3] David W. Howells[3] and Malcolm R. Macleod[2]





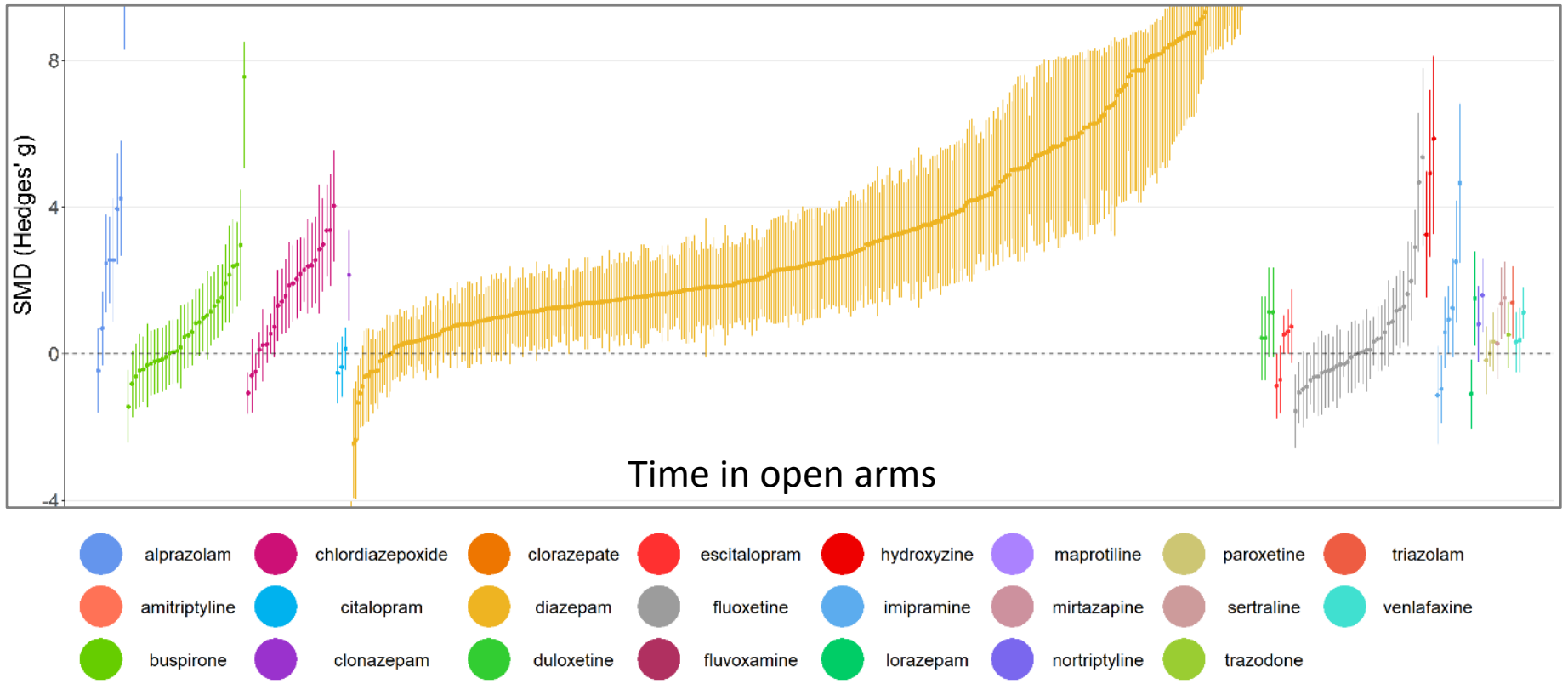
Contents lists available at ScienceDirect

Neuroscience and Biobehavioral Reviews

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

Reliability of common mouse behavioural tests of anxiety: A systematic review and meta-analysis on the effects of anxiolytics

Marianna Rosso *, Robin Wirz, Ariane Vera Loretan, Nicole Alessandra Sutter, Charlène Tatiana Pereira da Cunha, Ivana Jaric, Hanno Würbel, Bernhard Voelkl



How to estimate the required sample size ?

Sample size samba

'*Standard*' effect sizes

based on a pilot study

based on randomly chosen study

based on meta-analysis

Rule of 10

Mead's resource equation

Clinical (biological) relevant effect size

How to estimate the required sample size ?

Sample size samba
(U. Dirnagl)

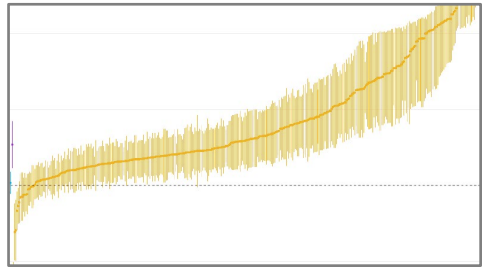
“come up with a number of animals (N) you want to use”

”calculate the ES neded to get power=0.8 given N”

“search for a study where they reported such an ES”

“use this study to make a power analysis”

(not recommended)



How to estimate the required sample size ?

'*Standard*' effect sizes

"take a ES value, where someone claimed that this value is a typical value for small/medium/large effects (in Psychology or Sociology)"

"claim that you expect small/medium/large effects"

"don't worry whether these ES values are sensible for your field"

(not recommended)

How to estimate the required sample size ?

Effect size estimates based on a pilot study

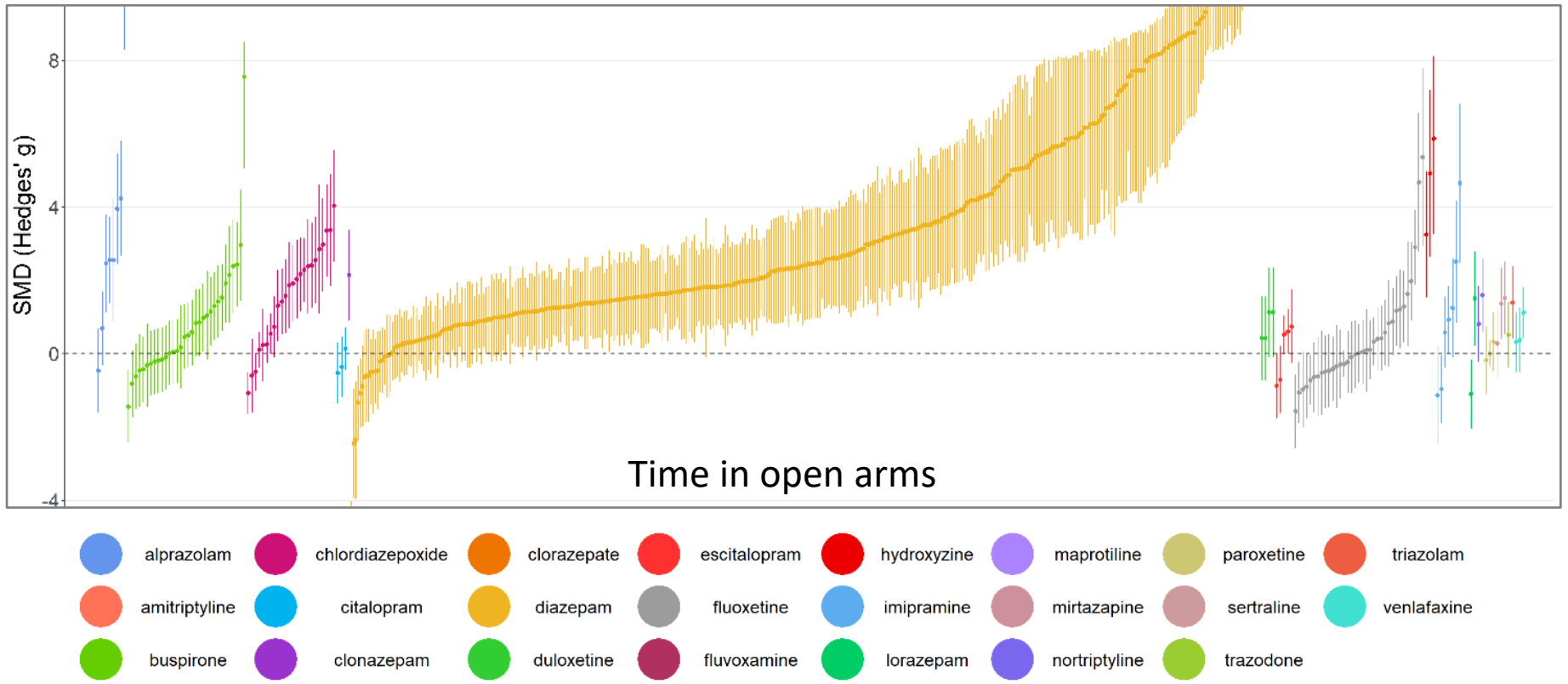
$$\text{power} \propto \frac{Es \times \alpha \times \sqrt{n}}{\sigma}$$

Getting good estimates for the variance requires large samples

Sample experinent:
N(std= 1.0), n=4
observed std:
0.97
0.83
1.90
0.49
0.74

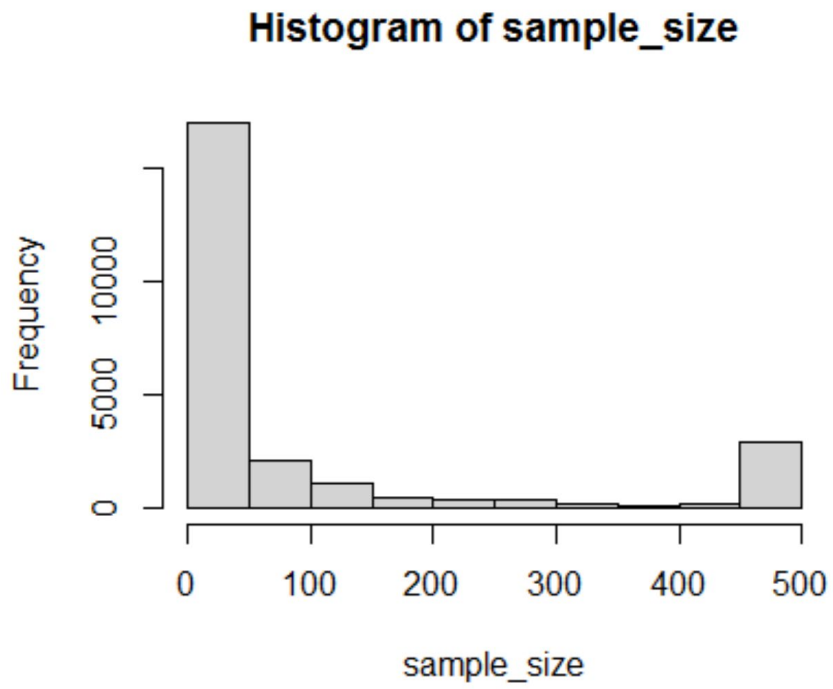
How to estimate the required sample size ?

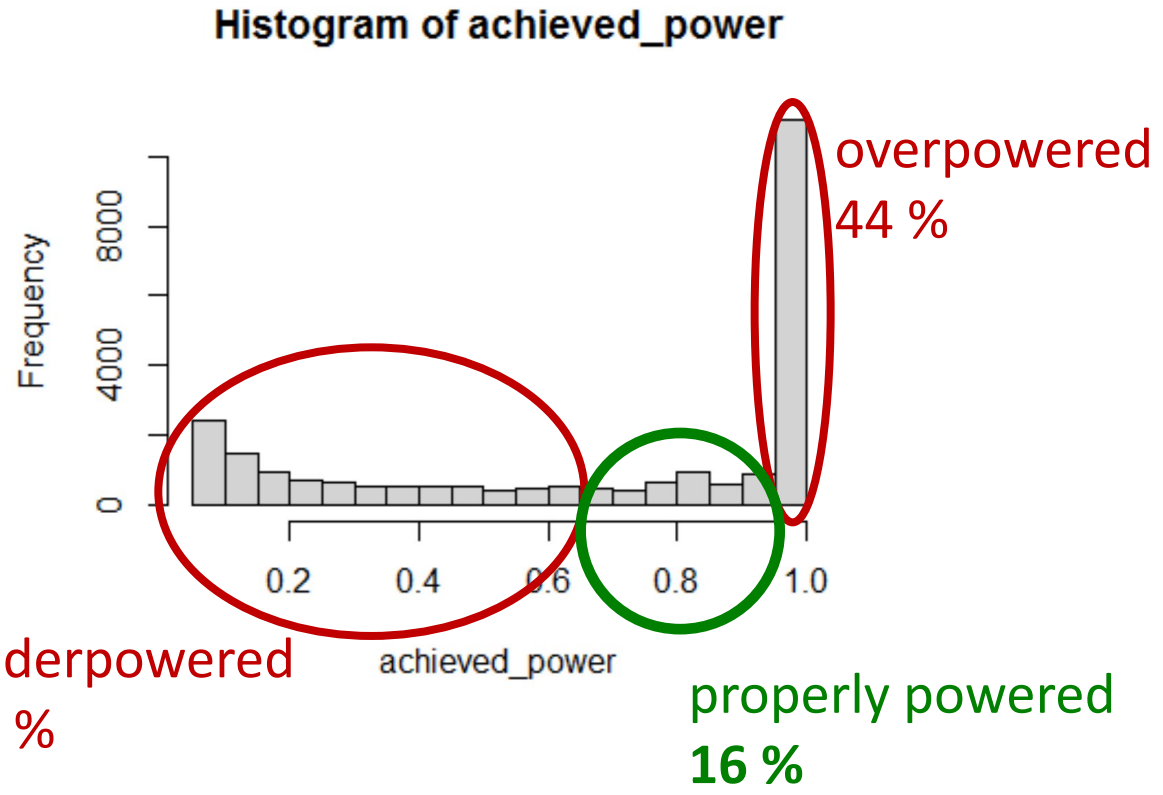
Power based on a single randomly chosen study



How to estimate the required sample size ?

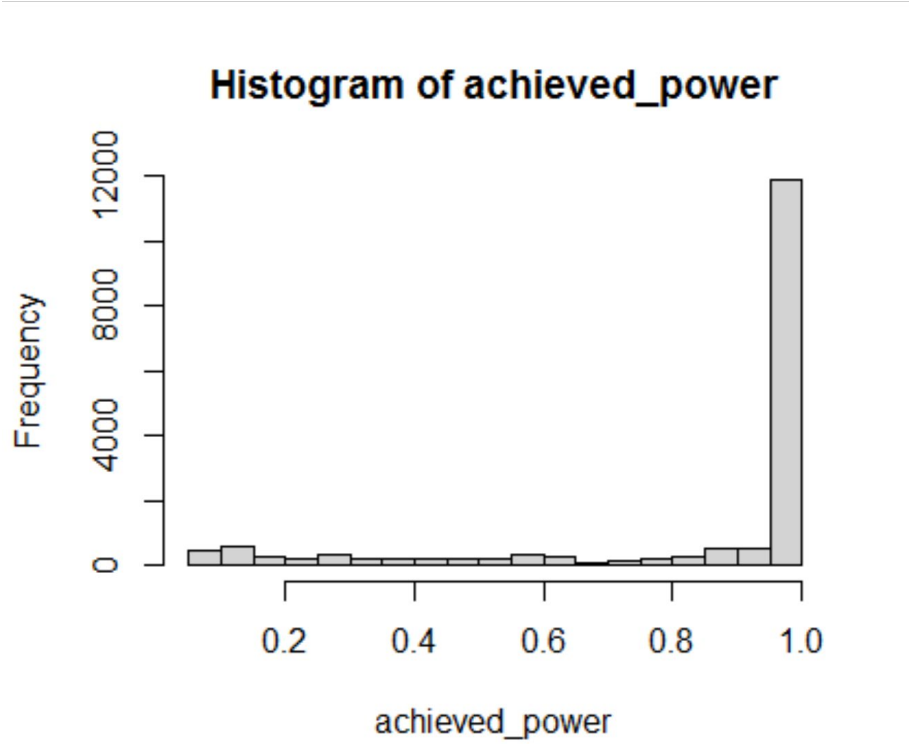
Power based on a single randomly chosen study

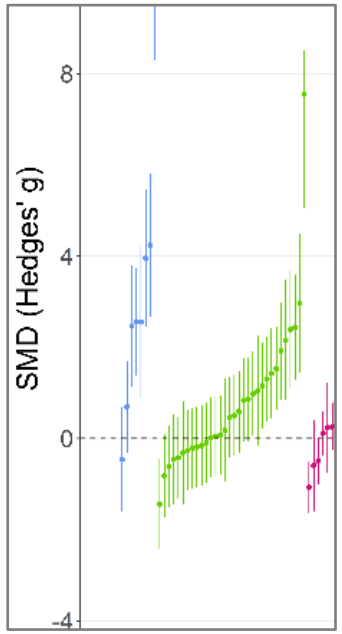




How to estimate the required sample size ?

Power based on a meta analysis





Contents lists available at ScienceDirect

Neuroscience and Biobehavioral Reviews

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

Reliability of common mouse behavioural tests of anxiety: A systematic review and meta-analysis on the effects of anxiolytics

Marianna Rosso *, Robin Wirz, Ariane Vera Loretan, Nicole Alessandra Sutter, Charlène Tatiana Pereira da Cunha, Ivana Jaric, Hanno Würbel, Bernhard Voelkl

Median sample size per group N=60
Median power achieved: 0.999
Overpowered (< 0.95): 70%

How to estimate the required sample size ?

'Rule of 10'

N=10 animals per group (factor combination)
(deemed unscientific)

Mead's resource equation

based on degrees of freedom (df)
E = N – B – T
where:

N is the total number of individuals or units in the study (minus 1)
B is the blocking component, representing environmental effects allowed for in the design (minus 1)
T is the treatment component, corresponding to the number of treatment groups (including control group) being used, or the number of questions being asked (minus 1)
E is the degrees of freedom of the error component and should be somewhere between **10** and **20**.

How to estimate the required sample size ?

Clinical (biological) relevant effect size

explain which effect size would be relevant ..

to justify further reseach

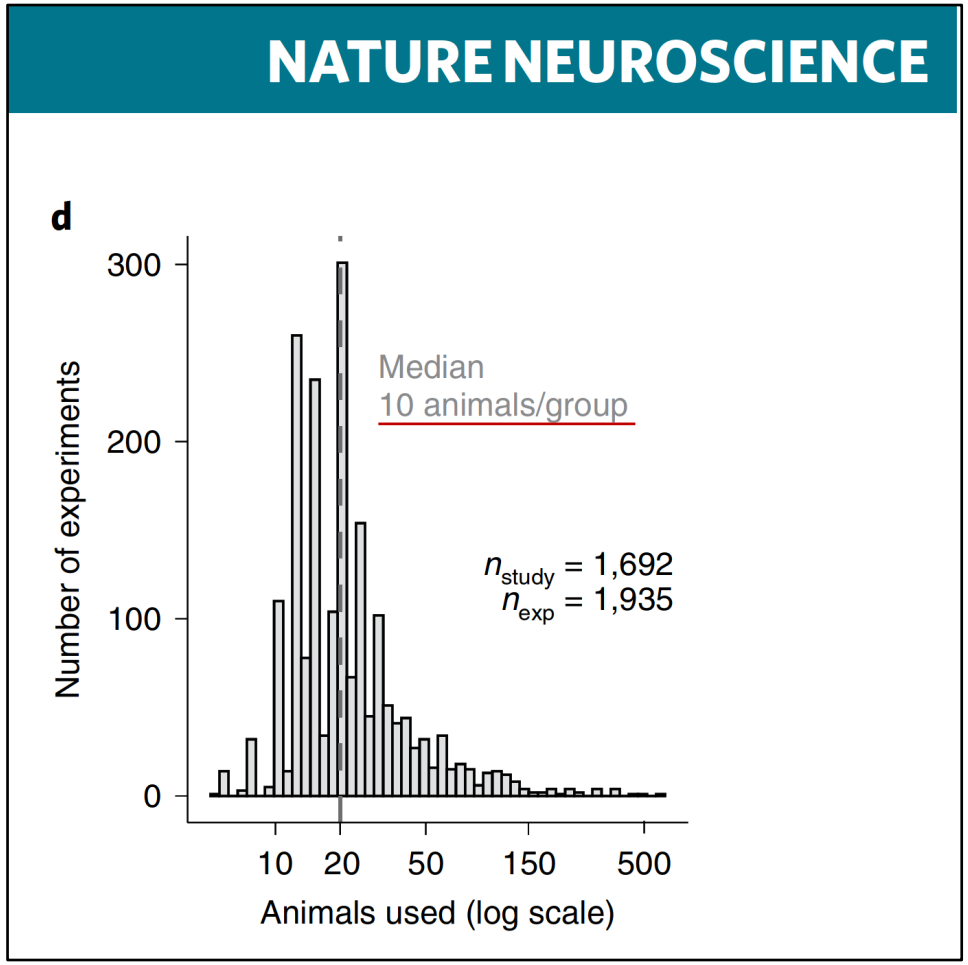
to be able to work with it

to be successful in clinical applications

(recommended by statisticians, difficult, subjective)

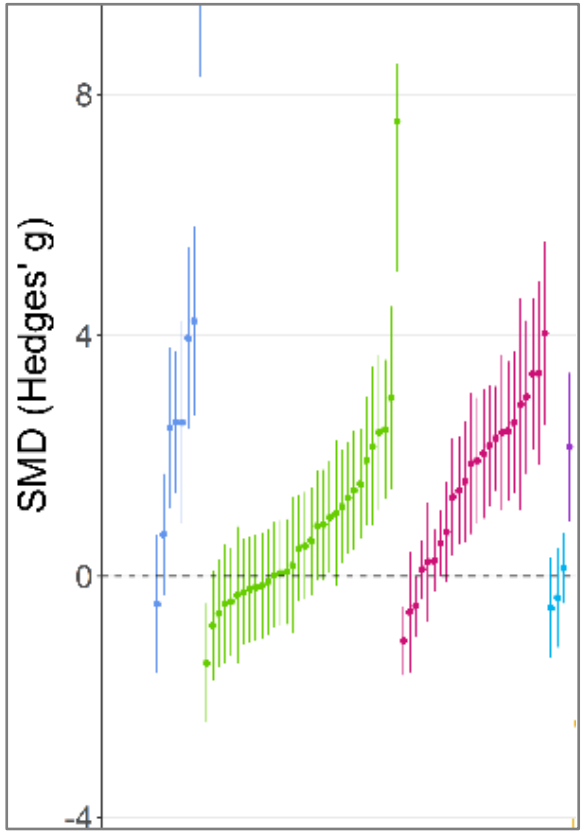
.. so, actually, this means that the question "**how many animals**" is not a statistical question, but a biological question.

How many mice ?



Bonapersona et al. (2021). Nat Neurosci, 24: 470-477.

Will my results be reproducible?



Large between-study vaiation will mean that independent replication studies will frequently deliver different results.

Can I improve reproduciblity with ..?

improving precision and standardization

No. If between-study variability is a major source of variation, then standardization will not improve reproducibility. To the contrary: it can even lead to poorer reproducibility (the standardization fallacy).

increasing sample size

No. If between-study variability is a major source of variation, then increasing the sample size will usually not improve reproducibility (and sometimes even slightly reduce reproducibility).

diversifying the study population

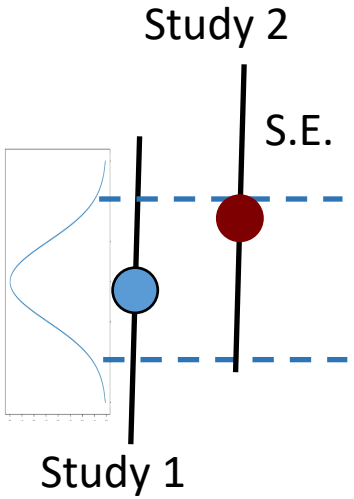
Yes. Diversifying the study population by including both sexes, different strains, outbred animals, different age groups, heterogenized environments etc can help to increase external validity. (Though evidence for success is mixed.)

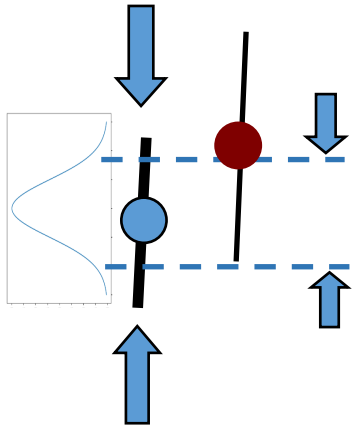
independently replicating experiments

Yes. Independent replication in a different laboratory has shown to be an efficient means to improve external validity and reproducibility of study outcomes.

The standardization fallacy

In the presence of between-lab variation, standardization within a lab reduces reproducibility and external validity.





Standardization within a lab reduces within-lab variation.

The S.E. becomes smaller.

The likelihood that another study falls outside the 95% C.I. increases.

Reproducibility decreases.

Can I improve reproduciblity with ..?

improving precision and standardization

No. If between-study variability is a major source of variation, then standardization will not improve reproducibility. To the contrary: it can even lead to poorer reproducibility (the standardization fallacy).

increasing sample size

No. If between-study variability is a major source of variation, then increasing the sample size will usually not improve reproducibility (and sometimes even slightly reduce reproducibility).

diversifying the study population

Sometimes **Yes.** Diversifying the study population by including both sexes, different strains, outbred animals, different age groups, heterogenized environments etc can help to increase external validity. (Though evidence for success is mixed.)

independently replicating experiments

Yes. Independent replication in a different laboratory has shown to be an efficient means to improve external validity and reproducibility of study outcomes.

Not all those who wander are lost ..