# Class 2: One Health, BioStatistics I: The Power and Crisis

## 1 The Crisis

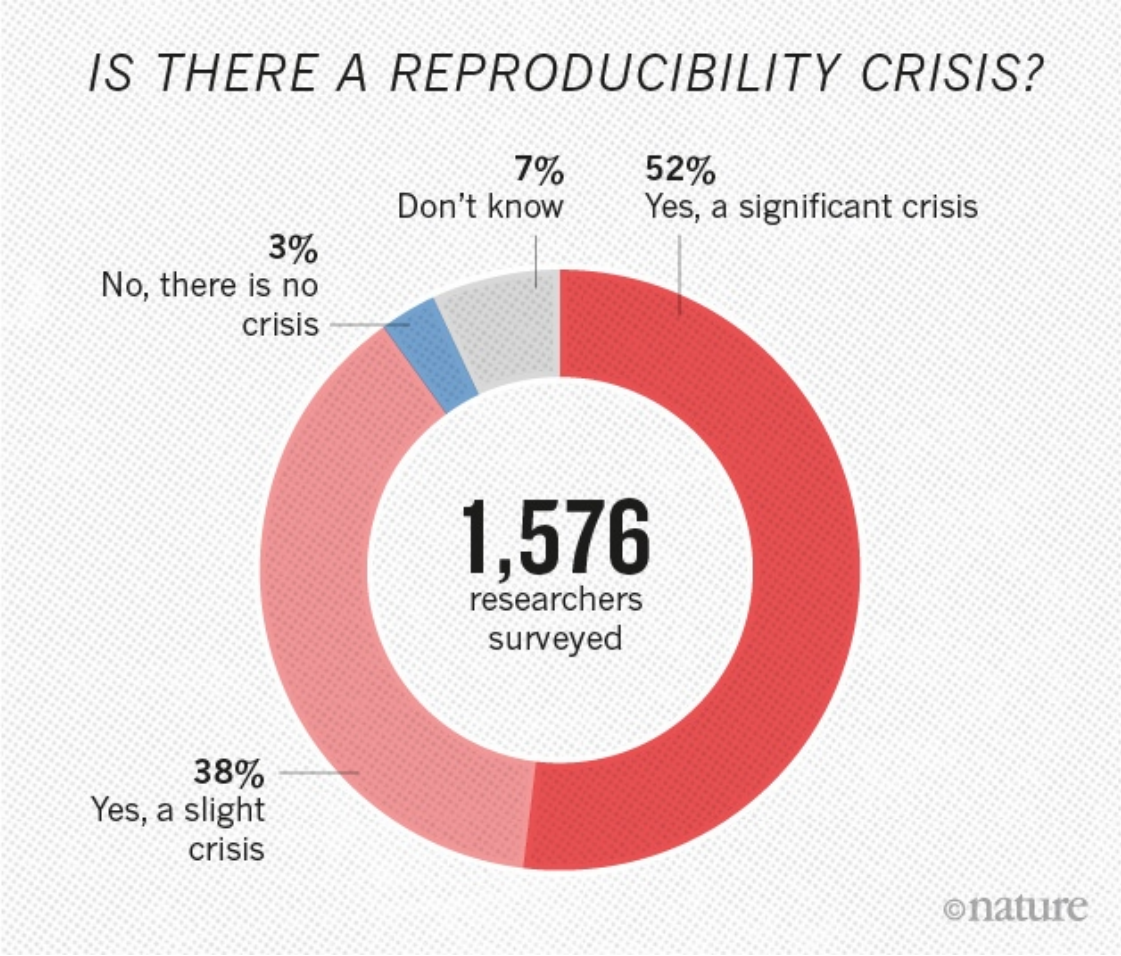
### 1.1 The Crisis


### 1.2 Objectives & Agenda

- Critical Thinking & Objectivity
- Bias in Knowledge & Believes
- Customs & Best Practice
- Societal Pressures

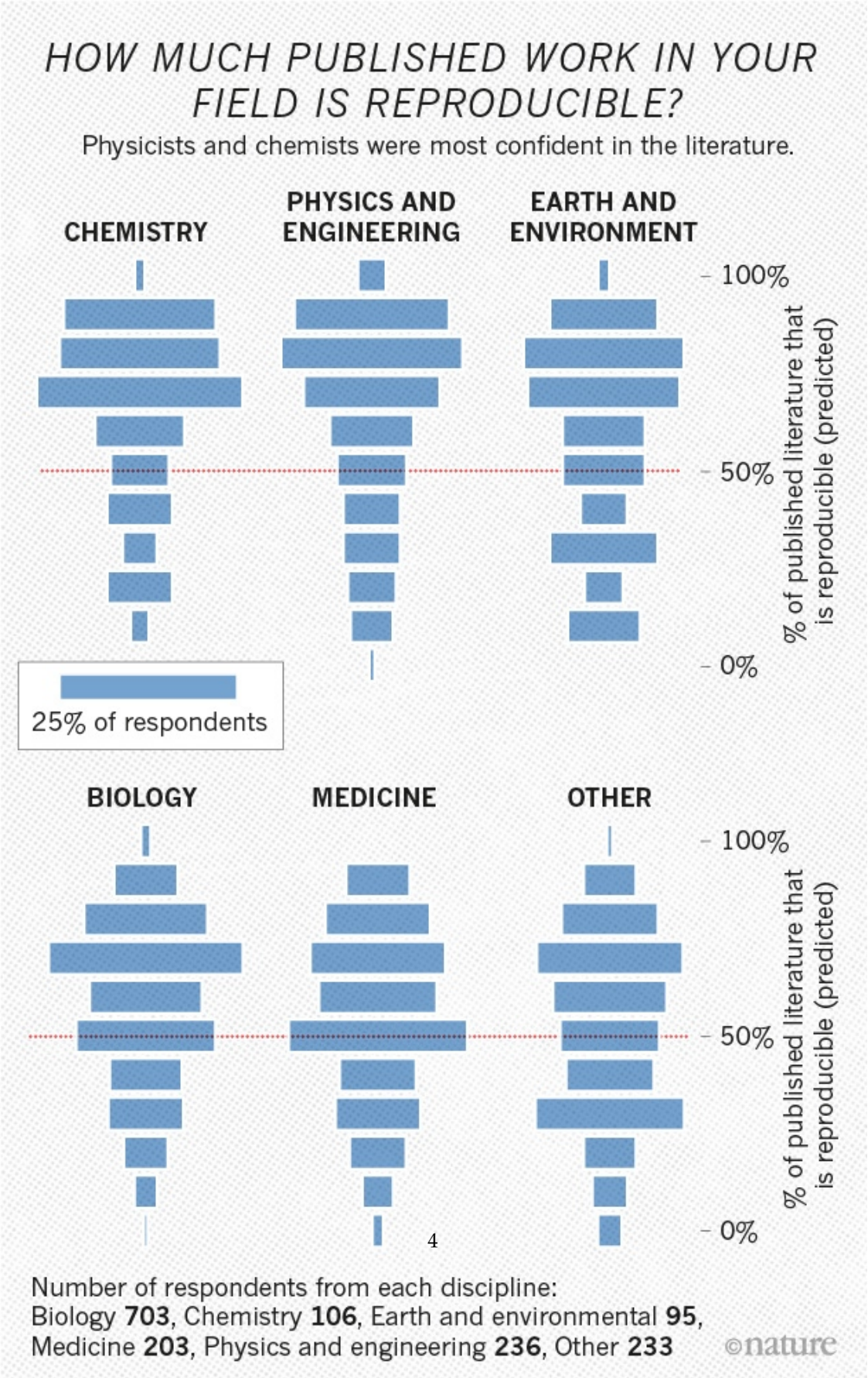
## 1.3 Where we are

### 1.3.1 Ask yourself



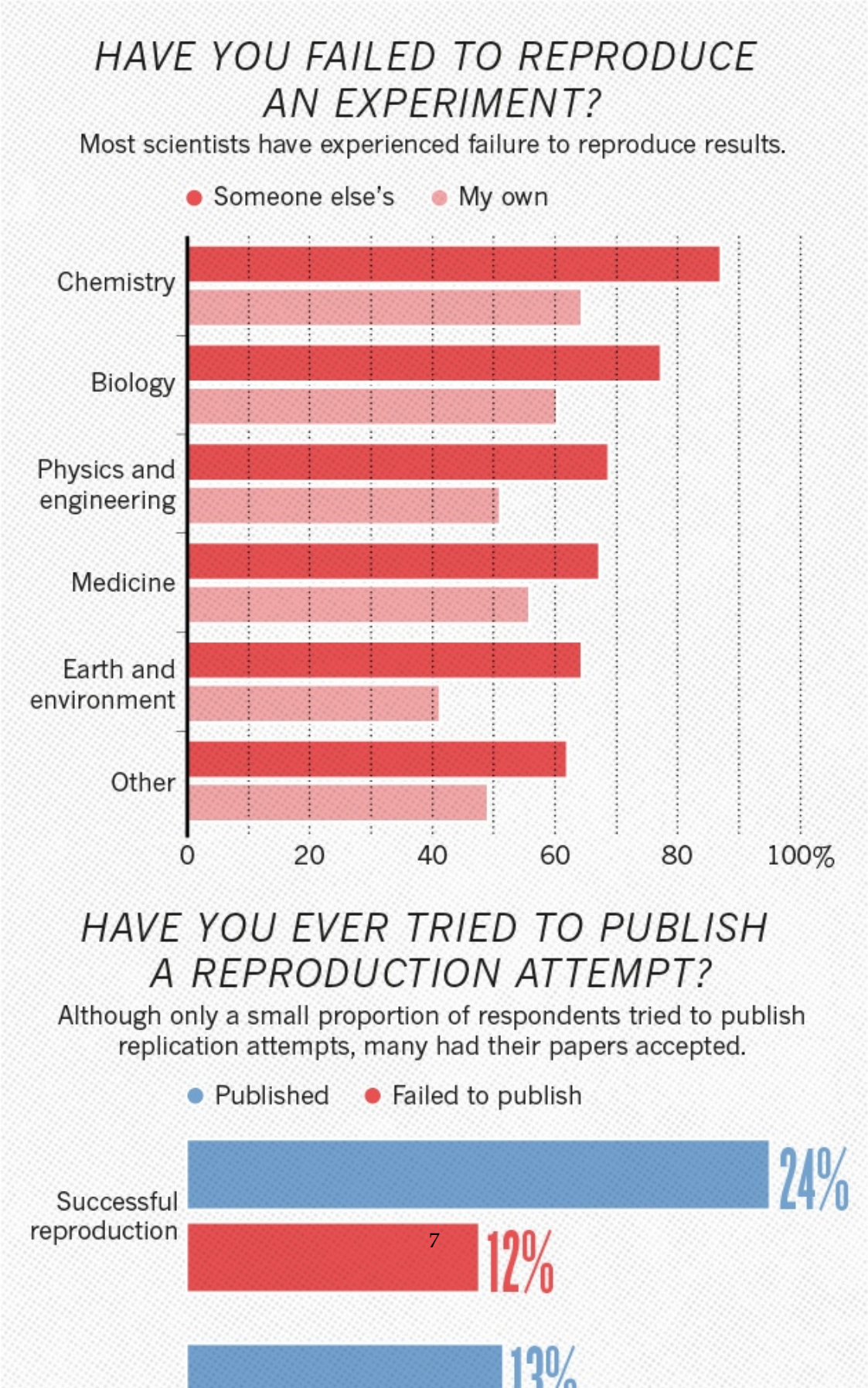

- 90% Recognize a **Crisis on REPRODUCIBILITY**

### 1.3.2 Trust you field?



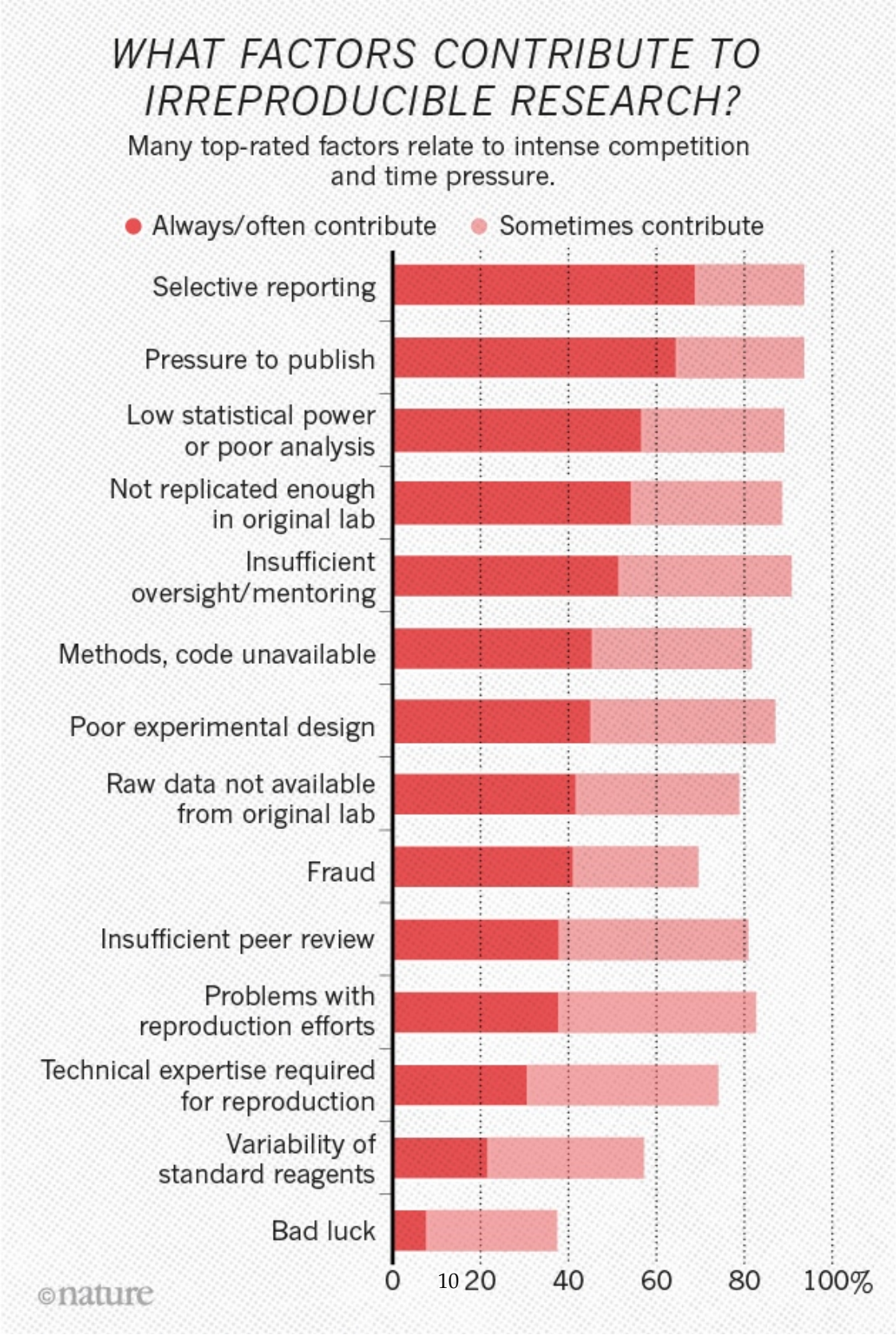
- Quantification makes a difference?.
- Physicist & chemists more confident

### 1.3.3 Have you?


HAVE YOU FAILED TO REPRODUCE AN EXPERIMENT?
Most scientists have experienced failure to reproduce results.
Someone else's
My own
Chemistry
Biology
Physics and engineering
Medicine
Earth and environment
Other
0
20
40
60
80
100%
HAVE YOU EVER TRIED TO PUBLISH A REPRODUCTION ATTEMPT?
Although only a small proportion of respondents tried to publish replication attempts, many had their papers accepted.
Published
Failed to publish
Successful reproduction
24%
12%
13%

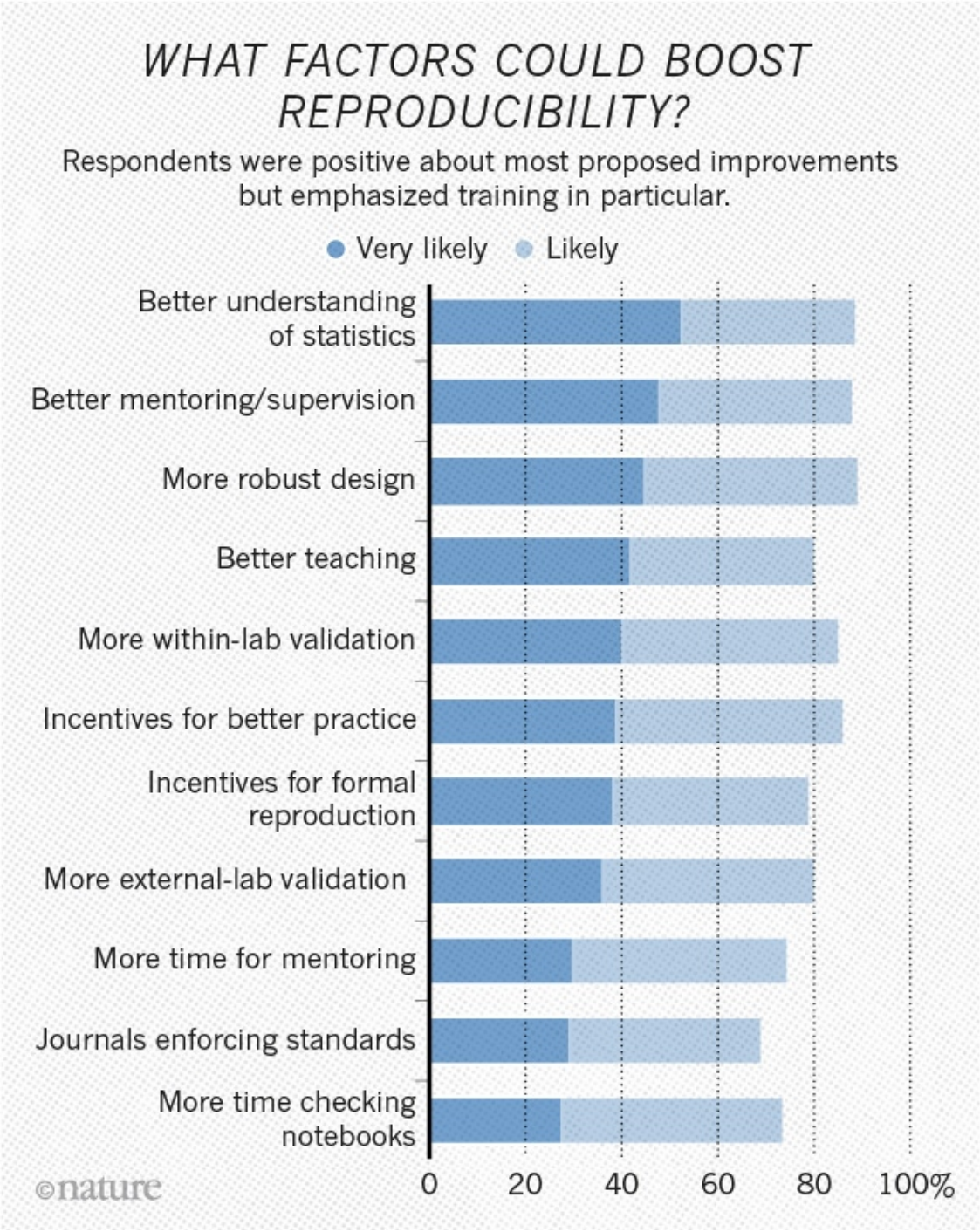
- Fail to reproduce results:
  - Someone else 60-80%
  - My own 40-60%
- Publishing difficulty
  - of failing reproduction 13%
  - vs successful reproduction 24%

### 1.3.4 Why?


WHAT FACTORS CONTRIBUTE TO IRREPRODUCIBLE RESEARCH?
Many top-rated factors relate to intense competition and time pressure.
Always/often contribute
Sometimes contribute
Selective reporting
Pressure to publish
Low statistical power or poor analysis
Not replicated enough in original lab
Insufficient oversight/mentoring
Methods, code unavailable
Poor experimental design
Raw data not available from original lab
Fraud
Insufficient peer review
Problems with reproduction efforts
Technical expertise required for reproduction
Variability of standard reagents
Bad luck
0
10
20
40
60
80
100%
©nature

- ~70% fraud
- >80% poor design
- Selective reporting & pressure ~90%

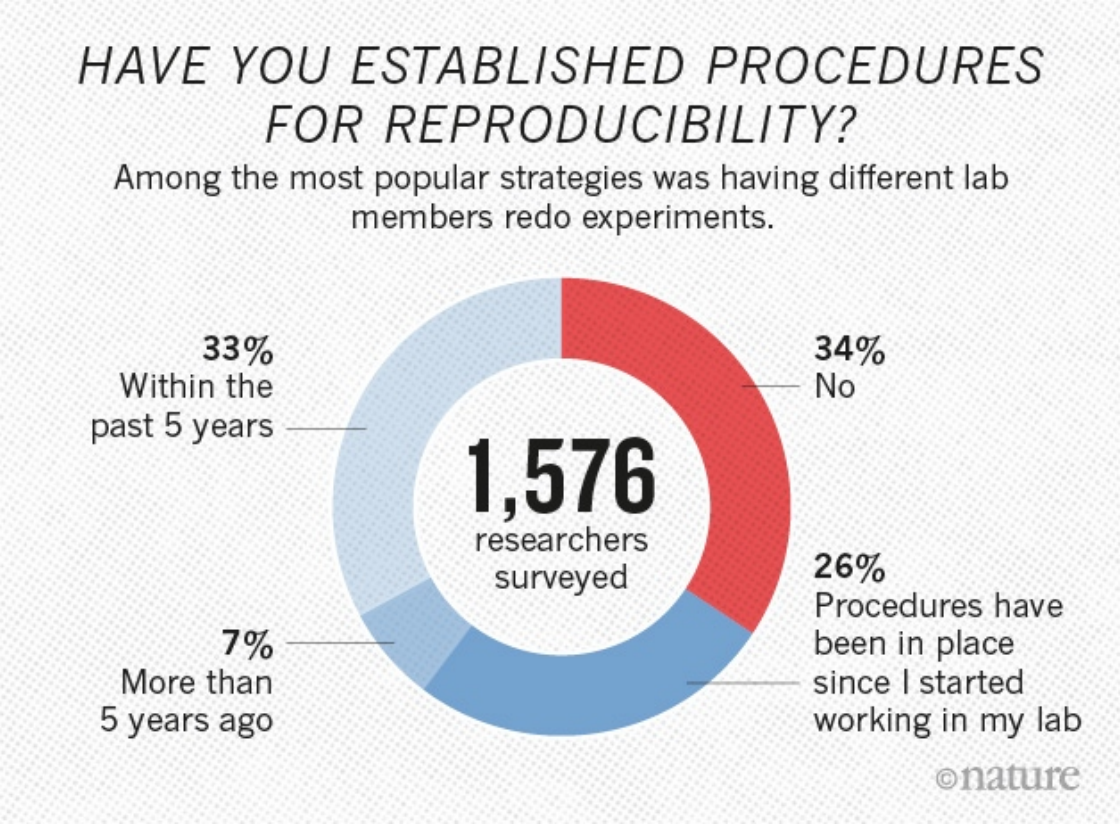
### 1.3.5 What to Change?




- ~90% Better Statistical undersanding
- Robust design

- Mentoring
- Better practices

#### 1.3.6 Did you?




- 34% did take actions
- 33% last 5 yrs
- 7% more than 5 yrs
- 26% From the beginning

### 1.4 Replication Studies

- **Replicability crisis** is a serious issue in which many scientific studies are difficult to reproduce or replicate.
- **Cancer research, only about 10–25%** of published studies could be validated or reproduced.
- In **psychology only about 36%** were reproduced.

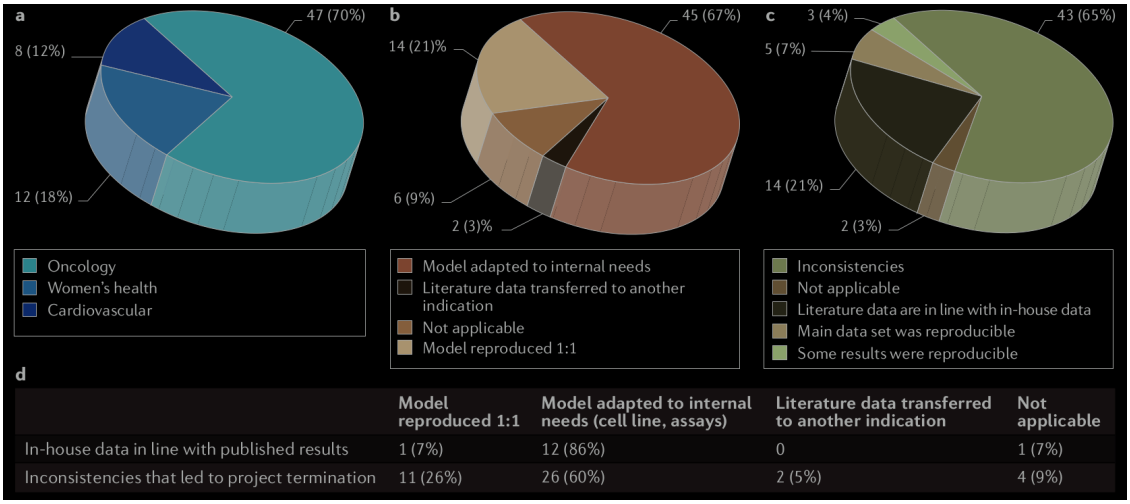
- Other
  - Medicine.
  - Genetics.
  - Economics.
  - **Neuroscience**.

### 1.4.1 Reasons

- Inappropriate practices of science,
  - HARKing (Hypothesizing After the Results are Known)
  - p-hacking.
  - Selective reporting of positive results.
  - Poor research design.
  - Lack of raw data.

## 1.5 Pharma

### 1.5.1 Bayer




| | Model reproduced 1:1 | Model adapted to internal needs (cell line, assays) | Literature data transferred to another indication | Not applicable |
|---|---|---|---|---|
| In-house data in line with published results | 1 (7%) | 12 (86%) | 0 | 1 (7%) |
| Inconsistencies that led to project termination | 11 (26%) | 26 (60%) | 2 (5%) | 4 (9%) |

- Oncology, woman health, cardiovascular.
- 65 % where not reproducible.

### 1.5.2 Amgen

**REPRODUCIBILITY OF RESEARCH FINDINGS**

Preclinical research generates many secondary publications, even when results cannot be reproduced.

| Journal impact factor | Number of articles | Mean number of citations of non-reproduced articles* | Mean number of citations of reproduced articles |
|---|---|---|---|
| >20 | 21 | 248 (range 3–800) | 231 (range 82–519) |
| 5–19 | 32 | 169 (range 6–1,909) | 13 (range 3–24) |

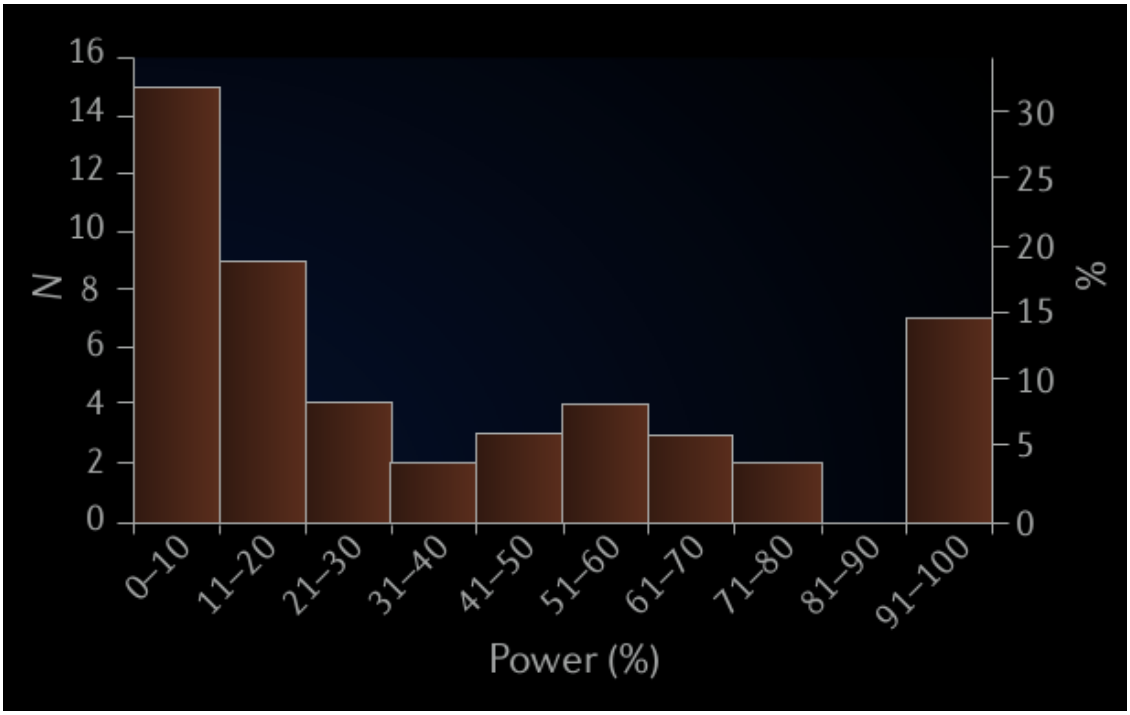
Results from ten-year retrospective analysis of experiments performed prospectively. The term 'non-reproduced' was assigned on the basis of findings not being sufficiently robust to drive a drug-development programme.
*Source of citations: Google Scholar, May 2011.

- Oncology and hematology
- From 53 works, only 6 (11%) where confirmed.

## 1.6 Give me the Power

### 1.6.1 Power Failure

- Median statistical power is 18-21%
- Neuroimaging studies 8%.
- Animal models 18-31%.



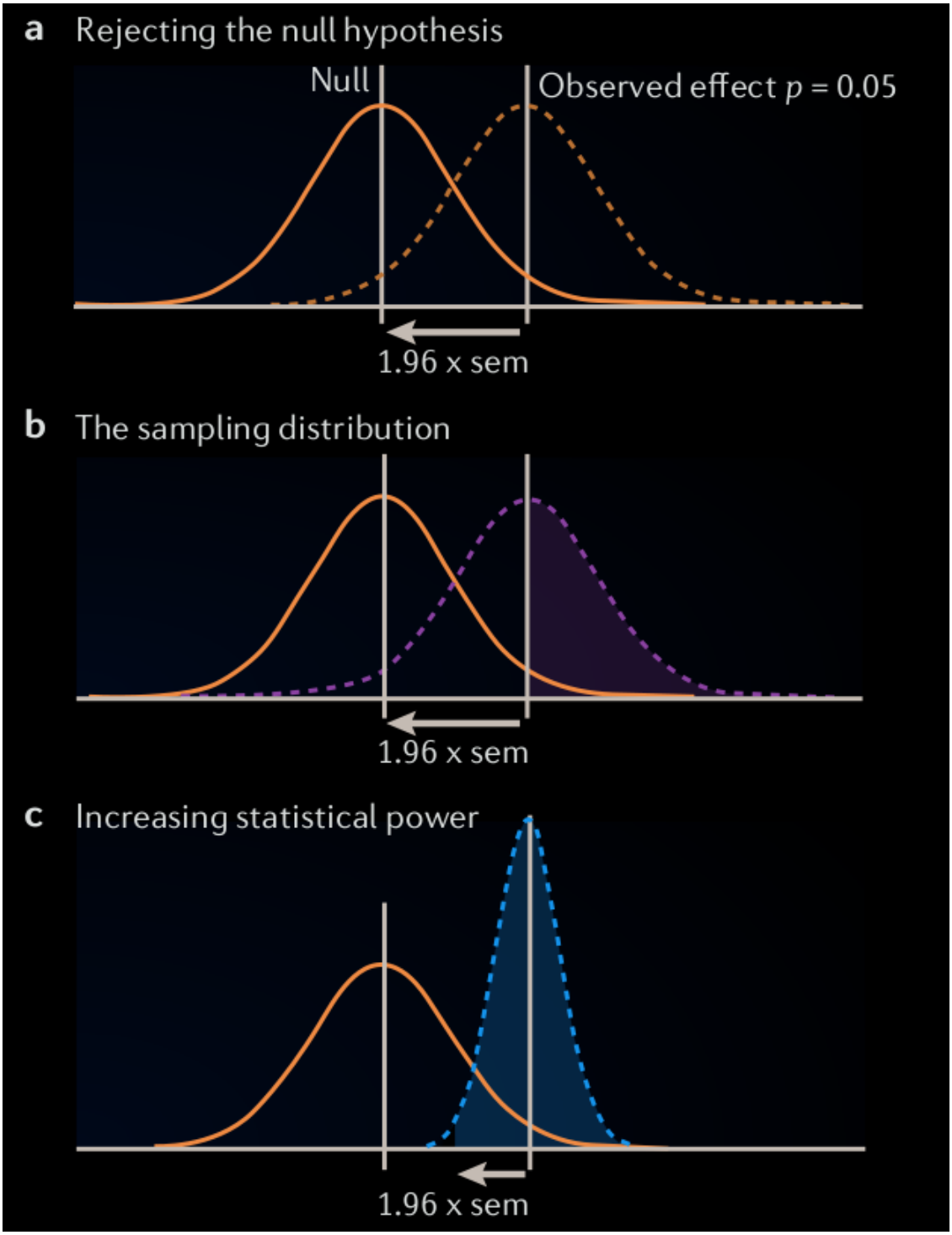
### 1.6.2 Sample Size

Table 2 | Sample size required to detect sex differences in water maze and radial maze performance

| | Total animals used | Required *N* per study | | Typical *N* per study | | Detectable effect for typical *N* | |
|---|---|---|---|---|---|---|---|
| | | 80% power | 95% power | Mean | Median | 80% power | 95% power |
| Water maze | 420 | 134 | 220 | 22 | 20 | $d = 1.26$ | $d = 1.62$ |
| Radial maze | 514 | 68 | 112 | 24 | 20 | $d = 1.20$ | $d = 1.54$ |

Meta-analysis indicated an effect size of Cohen's $d = 0.49$ for water maze studies and $d = 0.69$ for radial maze studies.

### 1.6.3 Power Effects




- **Low Power**

  - Discovering effects that are genuinely true is low.
  - Produce more false negatives than high-powered studies.

- **Low PPV**
  - Positive Predictable Value
  - PPV = ([1 – ] × R)/([1 ] × R + )

- **Effect inflation**
  - Effect inflation **occur** whenever claims of **discovery** are based **on thresholds** of **statistical significance**
    * for example, $p < 0.05$, or other selection filters.

#### 1.6.4 Low power and other biases

- **Low-powered** studies are **more likely** to provide a **wide range of estimates** of the magnitude of an **effect**.
- **Publication bias**, **selective data analysis and selective reporting** are **more likely to affect low-powered studies**.
- **Small studies** may be of **lower quality in other aspects of their design as well**.

#### 1.6.5 More Power

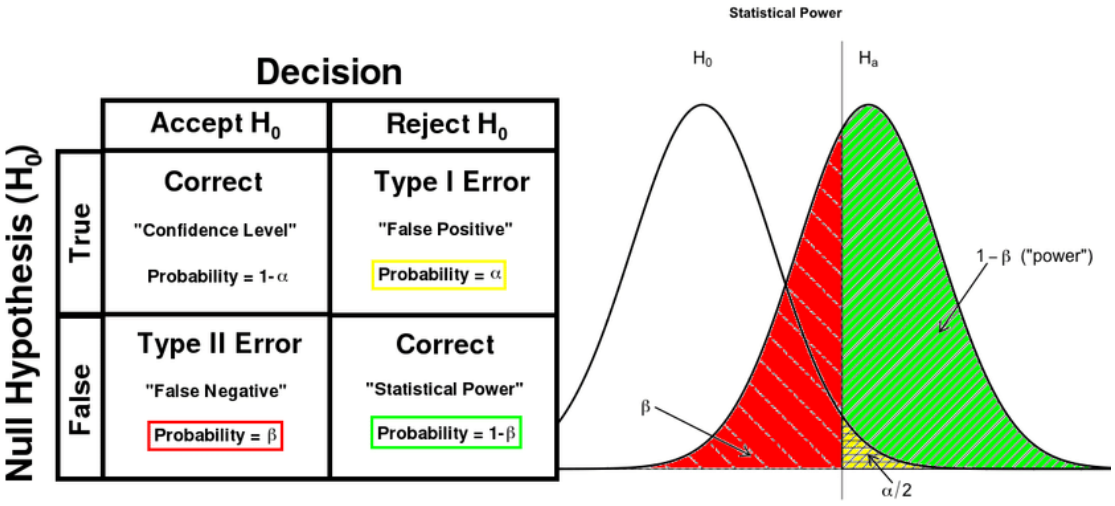
The probability that a research **finding** is indeed true **depends on:**

- the **prior probability** of it being true (before doing the study),
- the **statistical power** of the study,
- and the **level of statistical significance**.

#### 1.6.6 PPV

- After a research finding has been claimed based on achieving formal statistical significance, the **post-study probability that it is true** is the **positive predictive value**, PPV.

### 1.6.7 Graphical Assessment




### 1.6.8 Power & Bias

| Finding | True Relationship | | |
|---|---|---|---|
| | Yes | No | Total |
| Yes | c(1-$\beta$)$R/(R+1)$ | c$\alpha/(R+1)$ | c(R+$\alpha-\beta R)/(R+1)$ |
| No | c$\beta R/(R+1)$ | c(1-$\alpha$)$/(R+1)$ | c(1-$\alpha+\beta R)/(R+1)$ |
| Total | cR/(R+1) | c/(R+1) | c |

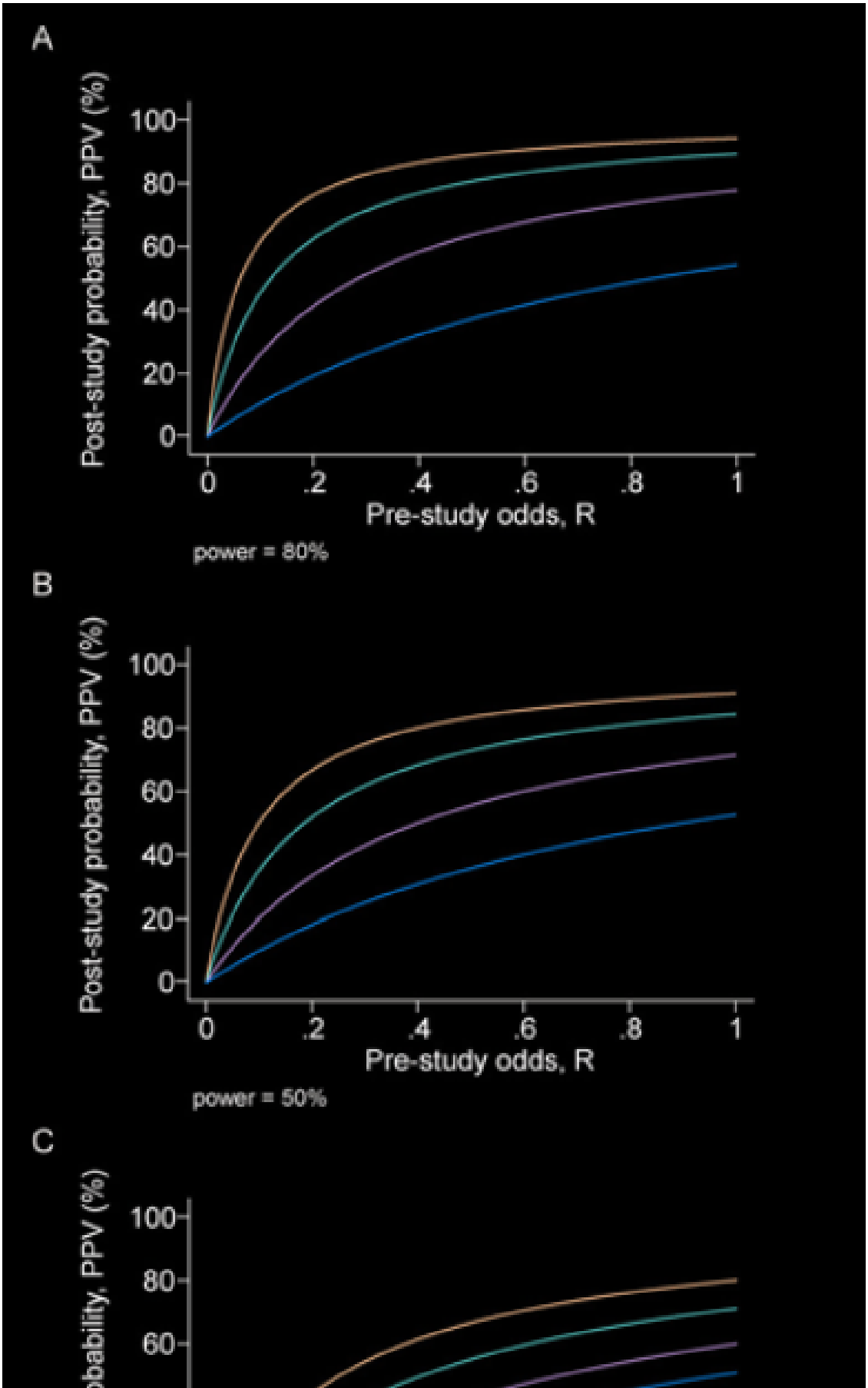
- (c=relationships are being probed in the field)

### 1.6.9 Bias

A combination of various factors that tend to produce research findings when they should not be produced including:

- Design
- Data
- Analysis
- Presentation factors

### 1.6.10 Corollaries



1. The **smaller the studies** conducted in a scientific field, the **less likely** the research **findings are to be true**.
2. The **smaller the effect sizes** in a scientific field, the **less likely** the research **findings are to be true**.
3. The **greater the number and the lesser the selection of tested relationships** in a scientific field, the **less likely** the research **findings are to be true**.
4. The **greater the flexibility in designs**, definitions, outcomes, and analytical modes in a scientific field, the **less likely** the research **findings are to be true**.
5. The **greater the financial and other interests** and prejudices in a scientific field, the **less likely** the research **findings are to be true**.
6. The **hotter a scientific field** (with more scientific teams involved), the **less likely** the research **findings are to be true**.

#### 1.6.11 Some Estimates

| $1-\beta$ | R | Bias, u | Example | PPV |
|---|---|---|---|---|
| 0.80 | 1:1 | 0.10 | Powered RCT with little bias a 1:1 pre-study odds | 0.85 |
| 0.95 | 2:1 | 0.30 | Confirmatory meta-analysis of good quality RCTs | 0.85 |
| 0.80 | 1:10 | 0.30 | Adequately powered exploratory epidemiological study | 0.20 |
| 0.20 | 1:1000 | 0.80 | Discovery oriented exploratory research with massive testing limited bias | 0.0015 |

- Randomized control Trial

## 1.7 Erroneous Interactions

#### 1.7.1 Even Best Families Top-ranking journals

- Behavioural, Systems Neuroscience.
  - ~50% correct comparison procedures for two experimental effects.
  - 2/3 of erroneous cases it may have had serious consequences.

**Table 1 Outcome of the main literature analysis**

| | *Nature* | *Science* | *Nature Neuroscience* | *Neuron* | *Journal of Neuroscience* | Summed |
|---|---|---|---|---|---|---|
| Total reviewed | 34 | 45 | 117 | 106 | 211 | 513 |
| Correct count | 3 | 9 | 17 | 13 | 36 | 78 |
| Error count | 7 | 11 | 16 | 15 | 30 | 79 |

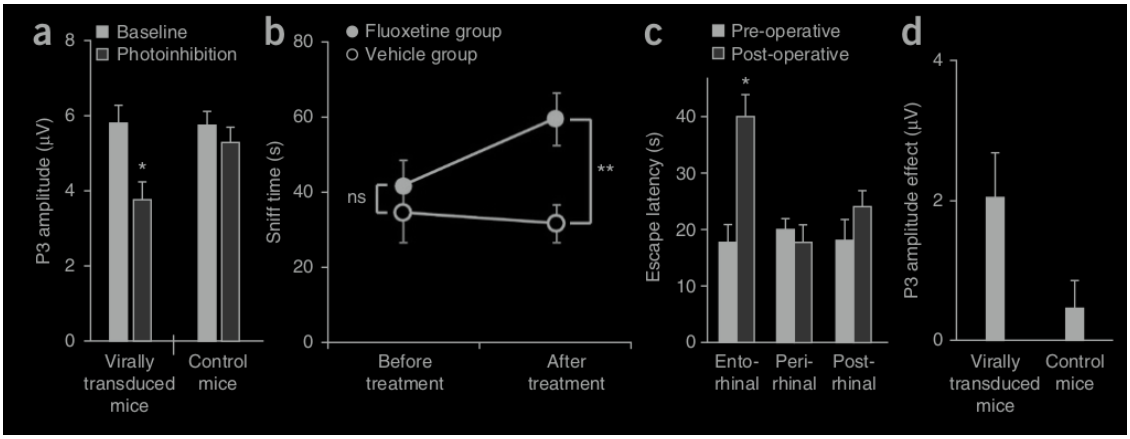
For this analysis, we included every article of which the abstract referred to behavior, cognitive function or brain imaging.

- Cellular and molecular neuroscience.

– From 120 additional articles in, none uses correct statistical procedure to compare effect sizes.

– 25 used incorrect procedures to compared significance levels.

### 1.7.2 Comparison Errors




- Tree situation where effect size comparison are incorrectly made.

## 1.8 DATA sets

### 1.8.1 Raw Data Withdraw

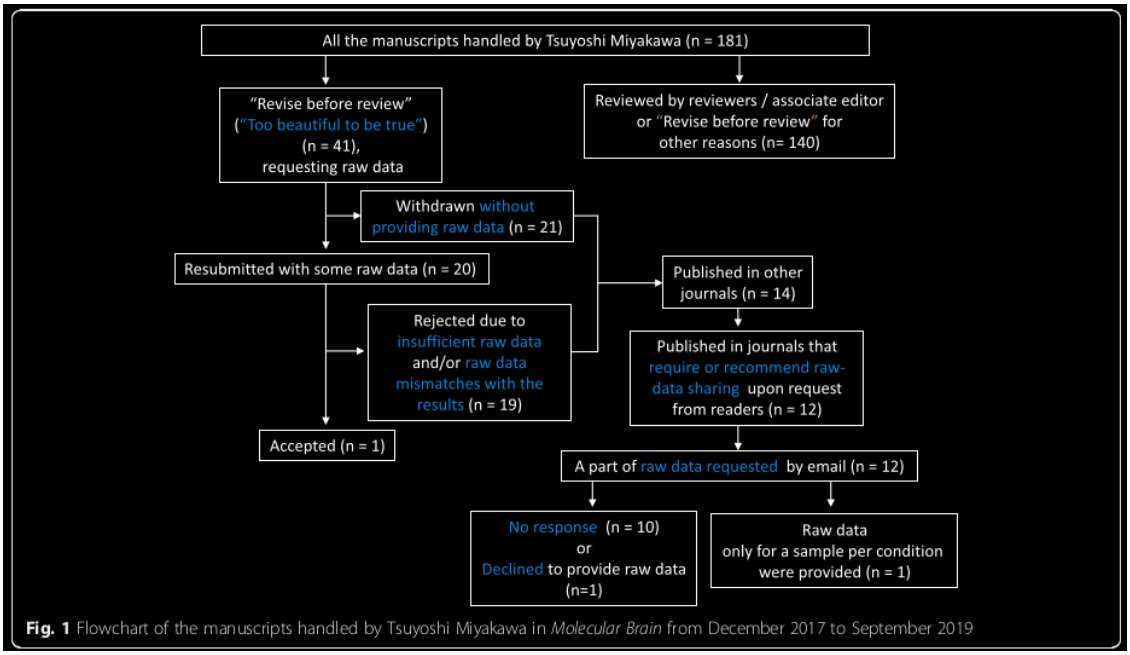



**Fig. 1** Flowchart of the manuscripts handled by Tsuyoshi Miyakawa in *Molecular Brain* from December 2017 to September 2019

#### 1.8.2 Absence of Raw Data means absence of science

#### 1.8.3 Open Science Open Data

# 2 Recalling the Power

## 2.1 Objectives

- Probability & Statistics
- Descriptive/Exploratory
- Inference
- Hypothesis Testing
- Some Recommendations for Biology

## 2.2 Intro

### 2.2.1 Basic Definition

- Statistical inference is the process of **drawing formal conclusions from data**.
- Statistical inference occurs where one wants to infer facts about a **population** using noisy statistical data **where uncertainty** must taken into account.
- Statistical inference **requires assessment of assumptions and tools** and **thinking how to draw conclusions** from data.

### 2.2.2 Some Inference Goals

- **Benchmarking**
  - Effectiveness of a treatment
- **Quantify**
  - Proportion of voting
- **Relationship**
  - Slope of Hoocke's law
- **Impact**
  - Confinements
- **Probability**
  - Raining tomorrow

### 2.2.3 Some tools in Inference

- ***Randomization***.
  - Unobserved variables may confound inferences of interest.
- ***Random sampling***.
  - Data representative of a population.
- ***Sampling models***.
  - Creating a model for the sampling process.
  - Independent Identically Distributed (i.i.d).
- ***Hypothesis testing***.
  - Decision making under uncertainty.
- ***Confidence intervals***.
  - Quantify uncertainty in estimation.
- ***Probability Models***.
  - Formal connection between the data and population of interest.
- ***Study Design***.
  - Experiment to minimize biases and variability.
- ***Nonparametric bootstrapping***.
  - Using data to create inference with minimal probability model assumptions.
- ***Permutation***.
  - Randomization and exchangeability testing to perform inferences.

### 2.2.4 Schools Styles

- Frequentist **Probability & Inference**
  - **Long run proportion** of times an event occurs in **independent, identically distributed repetitions**.
  - Interpretations of **probabilities to control error rates**.
  - **Given my data** controlling the **long run proportion** of mistakes I make **at a tolerable level**.
- Bayesian **Probability & Inference**

- Estimate or **calculate of beliefs**, which **follow certain rules**.
- **Inference** is performed by **Bayesian probability representation of beliefs**.
- **Subjective beliefs** and the **objective information** from the data **to infer**.

### 2.2.5 Probability Definition

Given a random variable (experiment; say rolling a die) a **probability measure** is a population quantity that *summarizes the randomness*.

- **number between 0 and 1**.
- **probability that something occurs is 1** (the die must be rolled) and
- The probability of the union of **any two sets of outcomes that have nothing in common** (mutually exclusive) is the **sum of their respective probabilities**.

### 2.2.6 Rules probability must follow

The Russian mathematician Andrey Nikolaevich Kolmogorov formalized these rules.

- The probability that **nothing occurs is 0**
- The probability that **something occurs is 1**
- The **probability of something is** 1 minus the probability that the opposite occurs
- The **probability of at least one of or more things** that can not simultaneously occur, **mutually exclusive**, is the **sum of their respective probabilities**.
- **More interestingly**
  - If an event **"A" implies the occurrence of event "B"**, then the probability of **"A" occurring is less than the probability that "B" occurs**.
  - For **any two events the probability that at least one occurs** is the **sum of their probabilities minus their intersection**.

### 2.2.7 Simple Example

- Event/Condition **X** with incidence of **3%** in the population
- Whereas **10%** of the population with Event/Condition **Y**.
- Does this imply that 13% of people will have at least one these Event/Condition?
  - Answer: **If the events can simultaneously occur**; they are not mutually exclusive so **NO**.

**lets:**

$$\begin{aligned} A_1 &= \{\text{Event X}\} \\ A_2 &= \{\text{Event Y}\} \end{aligned}$$

**Then**

$$\begin{aligned} P(A_1 \cup A_2) &= P(A_1) + P(A_2) - P(A_1 \cap A_2) \\ &= 0.13 - \text{Probability of having both} \end{aligned}$$

Likely, some fraction of the population has both.

### 2.2.8 Random variables

- A **random variable** is a numerical outcome of an experiment.
- The random variables come in **two varieties**, **discrete** or **continuous**.
  - **Discrete** random variable take on only a **countable number of possibilities**; the probability takes specific values.
  - **Continuous** random variable can take **any value on the real line**, or some subset; the probability they take within some range.

### 2.2.9 Quantiles

- **Famous sample quantiles**.
  - The 95th percentile on an exam, 95% of people scored worse than 5% scored better.
- Population analogs.
- **Definition**
  - The $\alpha^{th}$ **quantile** of a distribution with distribution function $F$ is the point $x_\alpha$ so that

$$F(x_\alpha) = \alpha$$

  - A **percentile** is simply a quantile with $\alpha$ expressed as a percent
  - The **median** is the $50^{th}$ percentile
- **For example**
  - The $75^{th}$ percentile of a distribution is the point so that:

* The probability, that a random variable from the population, **is less is 75%**.
* The probability, that a random variable from the population, **is more is 25%**.

### 2.2.10 Conditional Probability

- **Motivating example**
  - **The probability of getting** a one when rolling a (standard) die is usually assumed to be one sixth.
  - Suppose you were given the **extra information** that the die roll was an odd number (hence 1, 3 or 5).
  - **Conditional on this new information**, the probability of a one is now one third.
- **Definition**
  - Let $B$ be an event so that $P(B) > 0$
    * Then the conditional **probability of an event** $A$ **given that** $B$ **has occurred** is

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

    * Notice that **if** $A$ and $B$ **are independent**, then

$$P(A \mid B) = \frac{P(A)P(B)}{P(B)} = P(A)$$

    *

$$\cap = \text{ intersection}$$

- **Little Example**
  - Consider our die roll example, $P(\text{one given that roll is odd}) = P^*$.
    * $A = \{1\}$ and $B = \{1,3,5\}$.
      · Then

$$\begin{aligned} P^* &= P(A \mid B) \\ &= \frac{P(A \cap B)}{P(B)} \\ &= \frac{P(A)}{P(B)} = \frac{1/6}{3/6} = \frac{1}{3} \end{aligned}$$

### 2.2.11 Bayes' rule

- Baye's rule allows us to **reverse the conditioning** set provided that we know some marginal probabilities.
  - $$P(B \mid A) = \frac{P(A \mid B)P(B)}{P(A \mid B)P(B) + P(A \mid \neg B)P(\neg B)}$$
  - Where
    $$P(\neg B)$$
    is the initial degree of belief in not-B (B is false), and
    $$P(\neg B) = 1 - P(B)$$

- **Diagnostic tests**
  - Let $+$ and $-$ be the events that the result of a diagnostic test is positive or negative respectively.
  - Let $D$ and $D^c$ be the event that the subject of the test has or does not have the disease respectively.
  - The **sensitivity** is the probability that the test is positive given that the subject actually has the disease, $P(+ \mid D)$.
  - The **specificity** is the probability that the test is negative given that the subject does not have the disease, $P(- \mid D^c)$.

- **More definitions**
  - The **positive predictive value** is the probability that the subject has the disease given that the test is positive, $P(D \mid +)$
  - The **negative predictive value** is the probability that the subject does not have the disease given that the test is negative, $P(D^c \mid -)$
  - The **prevalence of the disease** is the marginal probability of disease, $P(D)$

### 2.2.12 Using Bayes' formula

$$\begin{aligned} P(D|+) &= \frac{P(+|D)P(D)}{P(+|D)P(D) + P(+|D^c)P(D^c)} \\ &= \frac{P(+|D)P(D)}{P(+|D)P(D) + \{1 - P(-|D^c)\}\{1 - P(D)\}} \\ &= \frac{.997 \times .001}{.997 \times .001 + .015 \times .999} = 0.062 \end{aligned}$$

- Then,

- **A positive test** result only suggests a **6% probability** that the subject has the **disease**.
- The **positive predictive value is 6%** for this test.

#### 2.2.13 Likelihood ratios, using Bayes rule

$$P(D|+) = \frac{P(+|D)P(D)}{P(+|D)P(D) + P(+|D^c)P(D^c)}$$

$$P(D^c|+) = \frac{P(+|D^c)P(D^c)}{P(+|D)P(D) + P(+|D^c)P(D^c)}$$

- **Therefore**
  - $$\frac{P(D|+)}{P(D^c|+)} = \frac{P(+|D)}{P(+|D^c)} \times \frac{P(D)}{P(D^c)}$$

    ie

    $$\text{post-test odds of } D = DLR_+ \times \text{pre-test odds of } D$$

    * DLR, **Diagnostic Likelihood Ratio** test
    * Similarly, $DLR_-$ relates the decrease in the odds of the disease after a negative test result to the odds of disease prior to the test.

#### 2.2.14 Expected values

- Expected values are useful for **characterizing a distributions**.
- The **mean** is a characterization of **its center**.
- The **variance and standard deviation** are characterizations of how **spread out** it is.
- Our **sample expected values** (the sample mean and variance) will **estimate the population** versions.

#### 2.2.15 The population mean

- The **expected value** or **mean** of a random variable is the center of its distribution
- For discrete random variable $X$ with PMF $p(x)$, it is defined as follows

  $$E[X] = \sum_x xp(x).$$

  where the **sum is taken over the possible values of** $x$
- $E[X]$ represents the center of mass of a collection of locations and weights, $\{x, p(x)\}$

#### 2.2.16 The sample mean

- The ample mean estimates this population mean.
- The **center of mass of the data is the empirical mean**.

$$\bar{X} = \sum_{i=1}^{n} x_i p(x_i)$$

where $p(x_i) = 1/n$

#### 2.2.17 What about a biased coin?

- Suppose that a random variable, $X$, is so that

$P(X = 1) = p$ and $P(X = 0) = (1 - p)$

- (This is a biased coin when $p \neq 0.5$)
- What is its expected value?

$$E[X] = 0 * (1 - p) + 1 * p = p$$

#### 2.2.18 Continuous random variables

- For a continuous random variable, $X$, with density, $f$, the expected value is again exactly the center of mass of the density.

#### 2.2.19 Summary of Expected Values

- **Facts**
  - Expected values are **properties of distributions**.
  - The **average of random variables is** itself **a random variable** and its associated distribution **has an expected value**.
  - The **center** of this distribution is **the same as that of the original distribution**.
  - Therefore, the expected value of the **sample mean** is the **population mean** trying to estimate.
  - When the **expected value of an estimator is what its trying to estimate**, we say that the estimator is **unbiased**

#### 2.2.20 The variance

- The variance of a random variable is a measure of **spread**
- If $X$ is a random variable with mean $\mu$, the variance of $X$ is defined as

$$Var(X) = E[(X-\mu)^2] = E[X^2] - E[X]^2$$

- The expected (squared) distance from the mean
- Densities with a higher variance are more spread out than densities with a lower variance
- The square root of the variance is called the **standard deviation**
- The standard deviation has the same units as $X$

—

#### 2.2.21 Examples Variance

- Example
  - What's the variance from the result of a toss of a die?
    - $E[X] = 3.5$
    - $E[X^2] = 1^2 \times \frac{1}{6} + 2^2 \times \frac{1}{6} + 3^2 \times \frac{1}{6} + 4^2 \times \frac{1}{6} + 5^2 \times \frac{1}{6} + 6^2 \times \frac{1}{6} = 15.17$
  - $Var(X) = E[X^2] - E[X]^2 \approx 2.92$

—

- Example
  - What's the variance from the result of the toss of a coin with probability of heads (1) of $p$?
    - $E[X] = 0 \times (1-p) + 1 \times p = p$
    - $E[X^2] = E[X] = p$

$$Var(X) = E[X^2] - E[X]^2 = p - p^2 = p(1-p)$$

—

#### 2.2.22 The sample variance

- The sample variance is

$$S^2 = \frac{\sum_{i=1}(X_i - \bar{X})^2}{n-1}$$

(almost, but not quite, the average squared deviation from the sample mean)

- It is also a random variable
  - It has an associate population distribution

  - Its expected value is the population variance
  - Its distribution gets more concentrated around the population variance with mroe data
- Its square root is the sample standard deviation

—

#### 2.2.23 Recall the mean

- Recall that the average of random sample from a population is itself a random variable
- We know that this distribution is centered around the population mean, $E[\bar{X}] = \mu$
- We also know what its variance is $Var(\bar{X}) = \sigma^2/n$
- This is very useful, since we don't have repeat sample means to get its variance; now we know how it relates to the population variance
- We call the standard deviation of a statistic a standard error

—

#### 2.2.24 To summarize

- The sample variance, $S^2$, estimates the population variance, $\sigma^2$
- The distribution of the sample variance is centered around $\sigma^2$
- The variance of the sample mean is $\sigma^2/n$
  - Its logical estimate is $s^2/n$
  - The logical estimate of the standard error is $S/\sqrt{n}$
- $S$, the standard deviation, talks about how variable the population is
- $S/\sqrt{n}$, the standard error, talks about how variable averages of random samples of size $n$ from the population are

—

#### 2.2.25 Summarizing what we know about variances

- The sample variance estimates the population variance
- The distribution of the sample variance is centered at what its estimating
- It gets more concentrated around the population variance with larger sample sizes

- The variance of the sample mean is the population variance divided by $n$
  - The square root is the standard error
- It turns out that we can say a lot about the distribution of averages from random samples, even though we only get one to look at in a given data set

#### 2.2.26 Hypothesis testing

- **Hypothesis testing** is concerned with making decisions using data.
- **A null hypothesis** is specified that represents the status quo, usually **labeled** $H_0$.
- The **null hypothesis** is assumed true and statistical **evidence is required to reject it** in favor of a research or alternative hypothesis.

#### 2.2.27 Hypothesis testing decision

- The alternative hypotheses are typically of the form $<$, $>$ or $\neq$
- Note that there are **four possible outcomes** of our statistical decision process

| Truth | Decide | Result |
|---|---|---|
| $H_0$ | $H_0$ | Correctly accept null |
| $H_0$ | $H_a$ | Type I error |
| $H_a$ | $H_a$ | Correctly reject null |
| $H_a$ | $H_0$ | Type II error |

#### 2.2.28 General rules

- The $Z$ test for $H_0 : \mu = \mu_0$, versus
  - $H_1 : \mu < \mu_0$
  - $H_2 : \mu \neq \mu_0$
  - $H_3 : \mu > \mu_0$
  - Test statistic
$$TS = \frac{\bar{X} - \mu_0}{S/\sqrt{n}}$$
  - Reject the null hypothesis when
    - $TS \leq Z_\alpha = -Z_{1-\alpha}$
    - $|TS| \geq Z_{1-\alpha/2}$
    - $TS \geq Z_{1-\alpha}$

#### 2.2.29 Notes

- We:
  - Fix $\alpha$ to be low, so if we reject $H_0$: our model is wrong or there is a **low probability that we have made an error**.
  - **Not fixed the probability of a type II error**, $\beta$; we tend to say *Fail to reject* $H_0$ rather than accepting $H_0$.
  - **Statistical significance** is no the same as **Scientific significance**.

#### 2.2.30 Connections with confidence intervals

- Consider testing $H_0 : \mu = \mu_0$ versus $H_a : \mu \neq \mu_0$.
- Take the set of all possible values for which you fail to reject $H_0$, this set is a $(1 - \alpha)100\%$ confidence interval for $\mu$.
- The same works in reverse; if a $(1 - \alpha)100\%$ interval contains $\mu_0$, then we **fail to** reject $H_0$.

#### 2.2.31 P-values

- Most common measure of statistical significance.
- Their ubiquity, along with concern over their interpretation and use makes them controversial among statisticians.

#### 2.2.32 What is a P-value?

Idea: Suppose nothing is going on - how unusual is it to see the estimate we got? Approach:

1. Define the hypothetical distribution of a data summary (statistic) when "nothing is going on" (null hypothesis)
2. Calculate the summary/statistic with the data we have (test statistic)
3. Compare what we calculated to our hypothetical distribution and see if the value is "extreme" (p-value)

#### 2.2.33 P-values

- The P-value is the probability under the null hypothesis of obtaining evidence as extreme or more extreme than that obtained
- If the P-value is small, then either $H_0$ is true and we have observed a rare event or $H_0$ is false
- Suppos that you get a $T$ statistic of 2.5 for 15 df testing $H_0 : \mu = \mu_0$ versus $H_a : \mu > \mu_0$.

- What's the probability of getting a $T$ statistic as large as 2.5?

"'r pt(2.5, 15, lower.tail = FALSE) "'
"' ## [1] 0.01225 "'

- Therefore, the probability of seeing evidence as extreme or more extreme than that actually obtained under $H_0$ is 0.0123

—

#### 2.2.34 The attained significance level

- Our test statistic was 2 for $H_0 : \mu_0 = 30$ versus $H_a : \mu > 30$.
- Notice that we rejected the one sided test when $\alpha = 0.05$, would we reject if $\alpha = 0.01$, how about 0.001?
- The smallest value for alpha that you still reject the null hypothesis is called the **attained significance level**
- This is equivalent, but philosophically a little different from, the **P-value**

—

#### 2.2.35 Notes

- By reporting a **p-value** the reader can perform the hypothesis test at whatever $\alpha$ **level**.
- If the **p-value** is **less than** $\alpha$ you reject the null hypothesis.
- For **two sided hypothesis test**, double the smaller of the two one sided hypothesis test P-values.

#### 2.2.36 Power

- Power is the probability of rejecting the null hypothesis when it is false
- Ergo, power (as its name would suggest) is a good thing; you want more power
- A type II error (a bad thing, as its name would suggest) is failing to reject the null hypothesis when it's false; the probability of a type II error is usually called $\beta$
- Note Power $= 1 - \beta$

—

### 2.2.37 Notes

- Consider our previous example involving RDI
- $H_0 : \mu = 30$ versus $H_a : \mu > 30$
- Then power is

$$P\left(\frac{\bar{X} - 30}{s/\sqrt{n}} > t_{1-\alpha,n-1} \; ; \; \mu = \mu_a\right)$$

- Note that this is a function that depends on the specific value of $\mu_a$!
- Notice as $\mu_a$ approaches 30 the power approaches $\alpha$

—

### 2.2.38 Calculating power for Gaussian data

- We reject if $\frac{\bar{X}-30}{\sigma/\sqrt{n}} > z_{1-\alpha}$
  - Equivalently if $\bar{X} > 30 + Z_{1-\alpha}\frac{\sigma}{\sqrt{n}}$
- Under $H_0 : \bar{X} \sim N(\mu_0, \sigma^2/n)$
- Under $H_a : \bar{X} \sim N(\mu_a, \sigma^2/n)$
- So we want

“‘r alpha = 0.05 z = qnorm(1 - alpha) pnorm(mu0 + z * sigma / sqrt(n), mean = mua, sd = sigma / sqrt(n), lower.tail = FALSE) “‘

—

### 2.2.39 Example continued

- $\mu_a = 32$, $\mu_0 = 30$, $n = 16$, $\sigma = 4$

“‘r mu0 = 30; mua = 32; sigma = 4; n = 16 z = qnorm(1 - alpha) “‘
“‘ ## Error: object ’alpha’ not found “‘
“‘r pnorm(mu0 + z * sigma / sqrt(n), mean = mu0, sd = sigma / sqrt(n), lower.tail = FALSE) “‘
“‘ ## Error: object ’z’ not found “‘
“‘r pnorm(mu0 + z * sigma / sqrt(n), mean = mua, sd = sigma / sqrt(n), lower.tail = FALSE) “‘
“‘ ## Error: object ’z’ not found “‘

—

### 2.2.40 Question

- When testing $H_a : \mu > \mu_0$, notice if power is $1 - \beta$, then

$$1 - \beta = P\left(\bar{X} > \mu_0 + z_{1-\alpha}\frac{\sigma}{\sqrt{n}}; \mu = \mu_a\right)$$

- where $\bar{X} \sim N(\mu_a, \sigma^2/n)$
- Unknowns: $\mu_a$, $\sigma$, $n$, $\beta$
- Knowns: $\mu_0$, $\alpha$
- Specify any 3 of the unknowns and you can solve for the remainder

—

### 2.2.41 Notes

- The calculation for $H_a : \mu < \mu_0$ is similar
- For $H_a : \mu \neq \mu_0$ calculate the one sided power using $\alpha/2$ (this is only approximately right, it excludes the probability of getting a large TS in the opposite direction of the truth)
- Power goes up as $\alpha$ gets larger
- Power of a one sided test is greater than the power of the associated two sided test
- Power goes up as $\mu_1$ gets further away from $\mu_0$
- Power goes up as $n$ goes up
- Power doesn't need $\mu_a$, $\sigma$ and $n$, instead only $\frac{\sqrt{n}(\mu_a - \mu_0)}{\sigma}$
  - The quantity $\frac{\mu_a - \mu_0}{\sigma}$ is called the effect size, the difference in the means in standard deviation units.
  - Being unit free, it has some hope of interpretability across settings

—

### 2.2.42 T-test power

- Consider calculating power for a Gossett's $T$ test for our example
- The power is

$$P\left(\frac{\bar{X} - \mu_0}{S/\sqrt{n}} > t_{1-\alpha,n-1}\ ;\ \mu = \mu_a\right)$$

- Calcuting this requires the non-central t distribution.

- 'power.t.test' does this very well
  - Omit one of the arguments and it solves for it

—

#### 2.2.43 Key ideas

- Hypothesis testing/significance analysis is commonly overused
- Correcting for multiple testing avoids false positives or discoveries
- Two key components
  - Error measure
  - Correction

—

#### 2.2.44 Three eras of statistics

The age of Quetelet and his successors, in which huge census-level data sets were brought to bear on simple but important questions: Are there more male than female births? Is the rate of insanity rising?

The classical period of Pearson, Fisher, Neyman, Hotelling, and their successors, intellectual giants who developed a theory of optimal inference capable of wringing every drop of information out of a scientific experiment. The questions dealt with still tended to be simple Is treatment A better than treatment B?

The era of scientific mass production, in which new technologies typified by the microarray allow a single team of scientists to produce data sets of a size Quetelet would envy. But now the flood of data is accompanied by a deluge of questions, perhaps thousands of estimates or hypothesis tests that the statistician is charged with answering together; not at all what the classical masters had in mind. Which variables matter among the thousands measured? How do you relate unrelated information?

[http://www-stat.stanford.edu/~ckirby/brad/papers/2010LSIexcerpt.pdf](http://www-stat.stanford.edu/~ckirby/brad/papers/2010LSIexcerpt.pdf)

—

#### 2.2.45 Types of errors

Suppose you are testing a hypothesis that a parameter $\beta$ equals zero versus the alternative that it does not equal zero. These are the possible outcomes.

| | $\beta = 0$ | $\beta \neq 0$ | Hypotheses |
|---|---|---|---|
| Claim $\beta = 0$ | $U$ | $T$ | $m - R$ |
| Claim $\beta \neq 0$ | $V$ | $S$ | $R$ |
| Claims | $m_0$ | $m - m_0$ | $m$ |

Type I error or false positive ($V$) Say that the parameter does not equal zero when it does
Type II error or false negative ($T$) Say that the parameter equals zero when it doesn't
—

#### 2.2.46 Error rates

False positive rate

- The rate at which false results ($\beta = 0$) are called significant:

$$E\left[\frac{V}{m_0}\right]$$

Family wise error rate (FWER)

- The probability of at least one false positive $\Pr(V \geq 1)$

False discovery rate (FDR)

- The rate at which claims of significance are false $E\left[\frac{V}{R}\right]$
- The false positive rate is closely related to the type I error rate [`http://en.wikipedia.org/wiki/False_positive_rate`](`http://en.wikipedia.org/wiki/False_positive_rate`)

—

#### 2.2.47 Controlling the false positive rate

If P-values are correctly calculated calling all $P < \alpha$ significant will control the false positive rate at level $\alpha$ on average.
Problem: Suppose that you perform 10,000 tests and $\beta = 0$ for all of them.
Suppose that you call all $P < 0.05$ significant.
The expected number of false positives is: $10,000 \times 0.05 = 500$ false positives.
How do we avoid so many false positives?
—

#### 2.2.48 Controlling family-wise error rate (FWER)

The Bonferroni correction is the oldest multiple testing correction.
Basic idea:

- Suppose you do $m$ tests
- You want to control FWER at level $\alpha$ so $Pr(V \geq 1) < \alpha$
- Calculate P-values normally

- Set $\alpha_{fwer} = \alpha/m$
- Call all $P$-values less than $\alpha_{fwer}$ significant

Pros: Easy to calculate, conservative Cons: May be very conservative

—

#### 2.2.49 Controlling false discovery rate (FDR)

This is the most popular correction when performing lots of tests say in genomics, imaging, astronomy, or other signal-processing disciplines.

Basic idea:

- Suppose you do $m$ tests
- You want to control FDR at level $\alpha$ so $E\left[\frac{V}{R}\right]$
- Calculate P-values normally
- Order the P-values from smallest to largest $P_{(1)}, ..., P_{(m)}$
- Call any $P_{(i)} \leq \alpha \times \frac{i}{m}$ significant

Pros: Still pretty easy to calculate, less conservative (maybe much less)
Cons: Allows for more false positives, may behave strangely under dependence

—

#### 2.2.50 Adjusted P-values

- One approach is to adjust the threshold $\alpha$
- A different approach is to calculate "adjusted p-values"
- They are not p-values anymore
- But they can be used directly without adjusting $\alpha$

Example:

- Suppose P-values are $P_1, \ldots, P_m$
- You could adjust them by taking $P_i^{fwer} = \max m \times P_i, 1$ for each P-value.
- Then if you call all $P_i^{fwer} < \alpha$ significant you will control the FWER.

—

#### 2.2.51 Case Study

- Case study I: no true positives

  “‘r set.seed(1010093) pValues <- rep(NA, 1000) for (i in 1:1000) { y <- rnorm(20) x <- rnorm(20) pValues[i] <- summary(lm(y ~ x))$coeff[2, 4] }

  sum(pValues < 0.05) “‘

  “‘ ## [1] 51 “‘

  —

- Case study I: no true positives

  “‘r

  sum(p.adjust(pValues, method = “bonferroni”) < 0.05) “‘

  “‘ ## [1] 0 “‘

  “‘r

  sum(p.adjust(pValues, method = “BH”) < 0.05) “‘

  “‘ ## [1] 0 “‘

  —

- Case study II: 50% true positives

  “‘r set.seed(1010093) pValues <- rep(NA, 1000) for (i in 1:1000) { x <- rnorm(20)

  if (i <= 500) { y <- rnorm(20) } else { y <- rnorm(20, mean = 2 * x) } pValues[i] <- summary(lm(y ~ x))$coeff[2, 4] } trueStatus <- rep(c(“zero”, “not zero”), each = 500) table(pValues < 0.05, trueStatus) “‘

  “‘ ## trueStatus ## not zero zero ## FALSE 0 476 ## TRUE 500 24 “‘

  —

- Case study II: 50% true positives

  “‘r

  table(p.adjust(pValues, method = “bonferroni”) < 0.05, trueStatus) “‘

  “‘ ## trueStatus ## not zero zero ## FALSE 23 500 ## TRUE 477 0 “‘

  “‘r

  table(p.adjust(pValues, method = “BH”) < 0.05, trueStatus) “‘

  “‘ ## trueStatus ## not zero zero ## FALSE 0 487 ## TRUE 500 13 “‘

  —

- Case study II: 50% true positives

  _P-values versus adjusted P-values__

“‘r par(mfrow = c(1, 2)) plot(pValues, p.adjust(pValues, method = “bonferroni”), pch = 19) plot(pValues, p.adjust(pValues, method = “BH”), pch = 19) “‘

![plot of chunk unnamed-chunk-3](assets/fig/unnamed-chunk-3.png)

—

#### 2.2.52 Notes and resources

Notes: Multiple testing is an entire subfield A basic Bonferroni/BH correction is usually enough If there is strong dependence between tests there may be problems

- Consider method=“BY”

Further resources: [Multiple testing procedures with applications to genomics](`http://www.amazon.com/Multiple-Procedures-Applications-Genomics-Statistics/dp/0387493166/ref=sr_1_2/102-3292576-129059?ie=UTF8&s=books&qid=1187394873&sr=1-2`) [Statistical significance for genome-wide studies](`http://www.pnas.org/content/100/16/9440.full`) [Introduction to multiple testing](`http://ies.ed.gov/ncee/pubs/20084018/app_b.asp`)

# 3 StatsBiol


Figure 1: (Network and Science Compexity)

## 3.1 Descriptive Statistics

| Term | Meaning | Common Uses |
|---|---|---|
| Standard deviation | The typical difference between each value and the mean value. | Describing how broadly the sample values are distributed. $$s.d. = \sqrt{\sum (X - \bar{X})^2/(N-1)}$$ |
| Standard error of the mean (s.e.m) | An estimate how variable the means will be if the experiment is repeated multiple times. | Inferring where the population mean is likely to lie or whether set of samples are likely to come from the sample population. $$s.e.m. = s.d./\sqrt{N}$$ |
| Confidence Interval (CI:95%) | with 95% confidence, the population mean will lie in this interval. | Top interfere where the population mean lies, and to compare two populations $$CI = mean \pm s.e.m. \times t_{(N-1)}$$ |
| Independent Data | Values from separate of the same type that are not linked | Testing hypothesis about population. |
| Replicate data | Values from experiment where everything is linked as much as possible. | Serves as an internal check on performance of an experiment. |
| Sampling error | Variation caused by sampling part of a population rather than measuring the whole population. | Can reveal bias in the data or problems with conduct of experiment. In binomial distributions the expected is $$\sqrt{Np(1-p)}$$ ; in Poisson the expected s.d. is $$\sqrt{mean}$$ |
| | | |

## 3.2 Statistical Hypothesis Testing

- Null hypothesis
  - **Pearson's correlation test** is that there is no relationship between two variables.
  - The null hypothesis for the **Student's t test** is that there is no difference between the means of two populations.

## 3.3 p-value (p)

- A **p-value**, which is the **probability of observing the result** given that the null hypothesis is true.
  - not the reverse, as is often the case with misinterpretations.
- $p <= \alpha$: **reject** $H_0$, different distribution.
- $p > \alpha$: **fail to reject** $H_0$, same distribution.

## 3.4 Errors

- There are **two types of errors**:
- **Type I Error**. Reject the null hypothesis when there is in fact no significant effect - false positive.
  - The p-value is optimistically small.
- **Type II Error**. Not reject the null hypothesis when there is a significant effect -false negative.
  - The p-value is pessimistically large.

## 3.5 What Is Statistical Power?

- Statistical power, or the power of a hypothesis test is the **probability that the test correctly rejects the null hypothesis**.
  - Power = 1 - Type II Error
  - Pr(True Positive) = 1 - Pr(False Negative)

More intuitively, the **statistical power** can be thought of as the **probability of accepting an alternative hypothesis, when the alternative hypothesis is true**.

- **Low Statistical Power**:
  - Large risk of committing Type II errors.
- **High Statistical Power**:
  - Small risk of committing Type II errors.

## 3.6 Statistical Power

- The **statistical power of a hypothesis test** is the probability of detecting an effect, if there is a true effect present to detect.

## 3.7 Power Analysis

- **Effect Size**.
  - The quantified magnitude of a result present in the population.
  - Effect size is calculated using a specific statistical measure, such as Pearson's correlation coefficient for the relationship between variables.
- **Sample Size**.
  - The number of observations in the sample.
- **Significance**.
  - The significance level used in the statistical test, e.g. alpha. Often set to 5% or 0.05.
- **Statistical Power**.
  - The probability of accepting the alternative hypothesis if it is true.