The End of Statistical Significance?

Jonathan Sterne
Department of Social Medicine,
University of Bristol UK

Acknowledgements: Doug Altman, George Davey Smith, Tim Peters, Mervyn Stone and Kate Tilling

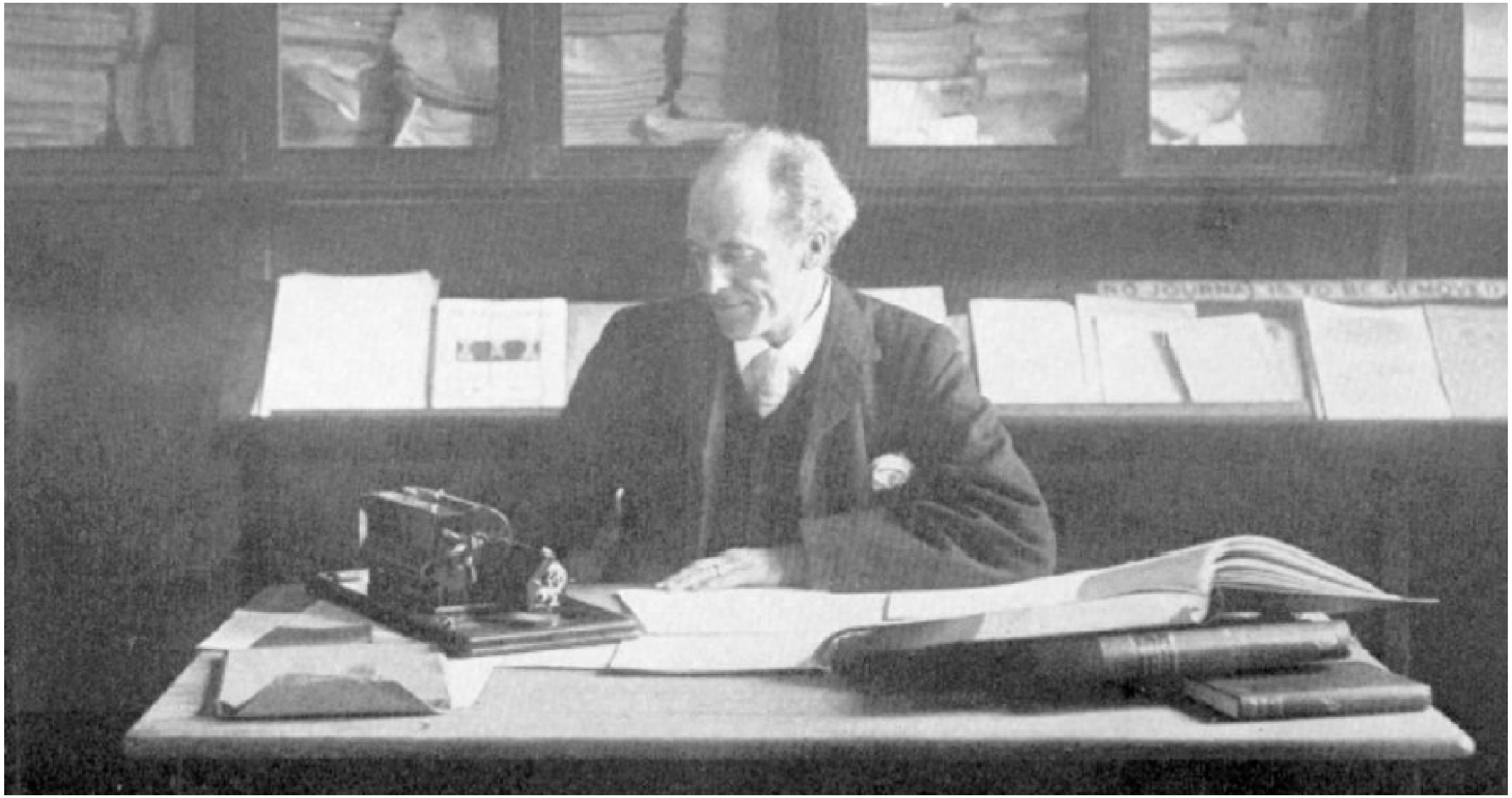
Outline

- P-values (significance levels)
- A brief history
- Using P-values and confidence intervals to interpret statistical analyses
- Interpretation of P-values
- Some recommendations, and a question….

Problems with interpretation of research findings

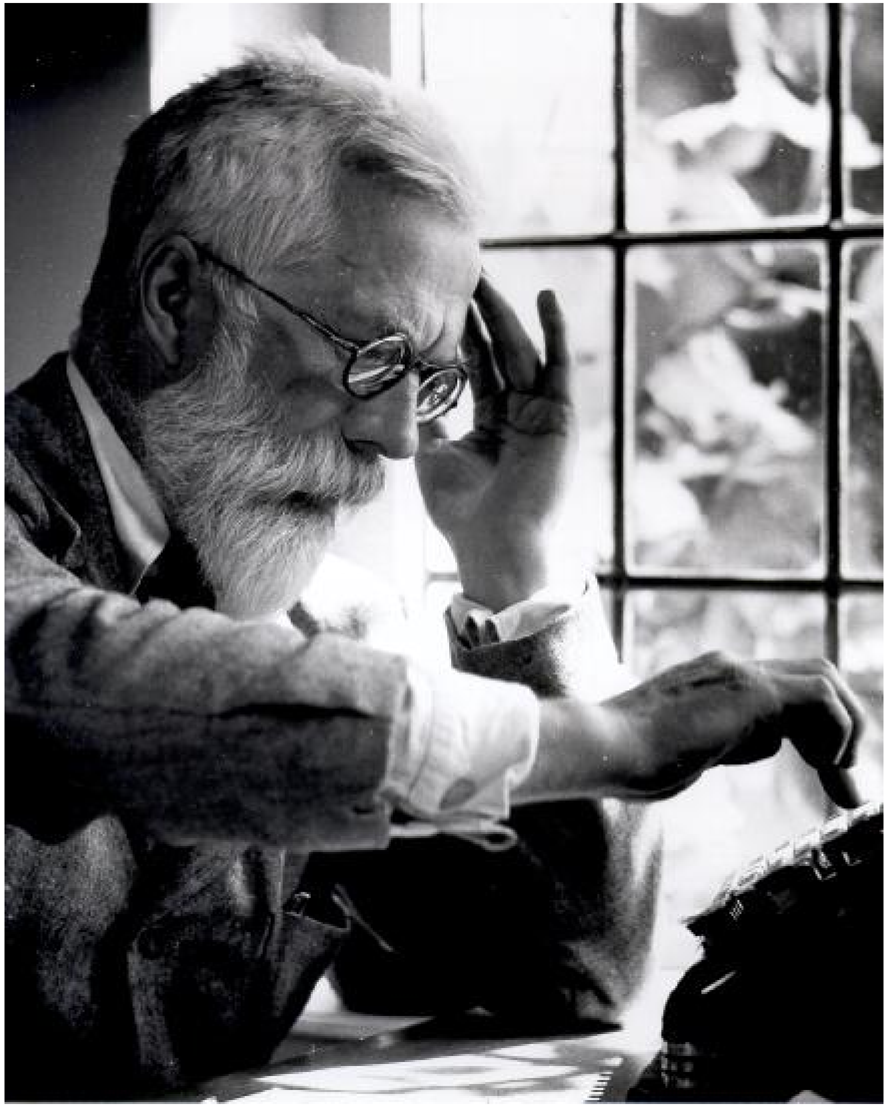
- Confounding
- Bias
- **Misinterpretation of statistical evidence**

Karl Pearson (1857-1936)



- Developed the formula for the correlation coefficient, and introduced the chi-squared (χ^2) test
- Published the first statistical tables, and did the first meta-analysis

R.A. Fisher (1890-1962)



- The father of modern statistical inference (and of statistical genetics)
- Introduced the idea of *significance levels* as a means of examining the discrepancy between the data and a null hypothesis

R.A. Fisher - quotes

“perhaps the most original mathematical scientist of the [twentieth] century”
Bradley Efron *Annals of Statistics* (1976)

“Fisher was a genius who almost single-handedly created the foundations for modern statistical science ….”
Anders Hald *A History of Mathematical Statistics* (1998)

“Sir Ronald Fisher … could be regarded as Darwin’s greatest twentieth-century successor.”
Richard Dawkins *River out of Eden* (1995)

“I occasionally meet geneticists who ask me whether it is true that the great geneticist R. A. Fisher was also an important statistician.”
Leonard J. Savage *Annals of Statistics* (1976)

From Fisher’s obituary

“In character he was, let us admit it, a difficult man. Among his wide circle of admirers he accepted homage with fair grace. With most other people he seemed to feel that his genius entitled him to more social indulgence than they were willing to concede: in short he could be, and not infrequently was, gratuitously rude. In his written work he rarely acknowledged any kind of indebtedness to other statisticians and mentioned them, if at all, only to correct or castigate. In fact, to win written approbation from him in his later work one had to have been dead for some time.”

W.G Kendall, *Biometrika* 1963 **50:** 1-15.

P-values (significance levels)

- We postulate a **null hypothesis**, eg
  - MMR vaccination does not affect a child's subsequent risk of autism
  - Birth weight is not associated with subsequent IQ
  - Living close to power lines does not change children's risk of leukaemia
- Does the data in our sample provide evidence ***against*** the null hypothesis?
- We calculate the **P-value** - the probability of getting a difference at as big as the one observed, if the null hypothesis is true

Example

Lung capacity (FVC) measured in 100 men

Group	Number	Mean FVC	*s.d.*
Non-smokers (0)	n_0=64	$\bar{x}_0$ = 5.0	s_0 = 0.6
Smokers (1)	n_1=36	$\bar{x}_1$ = 4.7	s_1 = 0.6

The **difference** in mean FVC is $\bar{x}_1 - \bar{x}_0$ = -0.3

The **standard error** of the difference in mean FVC is 0.125

Does this study provides evidence against the null hypothesis that, in the population, the difference in mean FVC is zero?

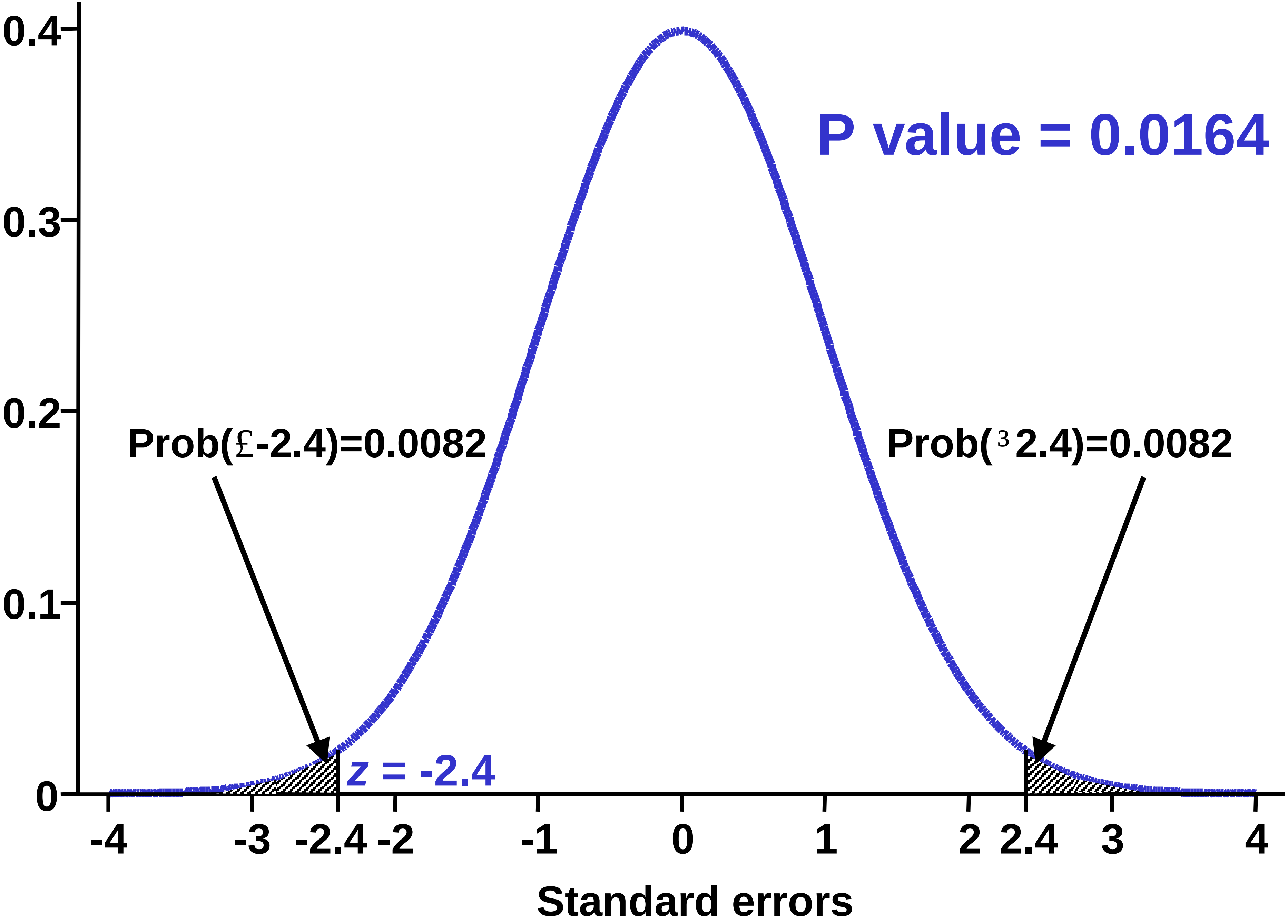
z-tests

The value: $\dfrac{\text{difference in means}}{\text{s.e. of difference}}$ is known as a ***z*-statistic**

This expresses the difference in terms of standard error units from the null value of 0

Here, *z* = -0.3/0.125 = -2.4

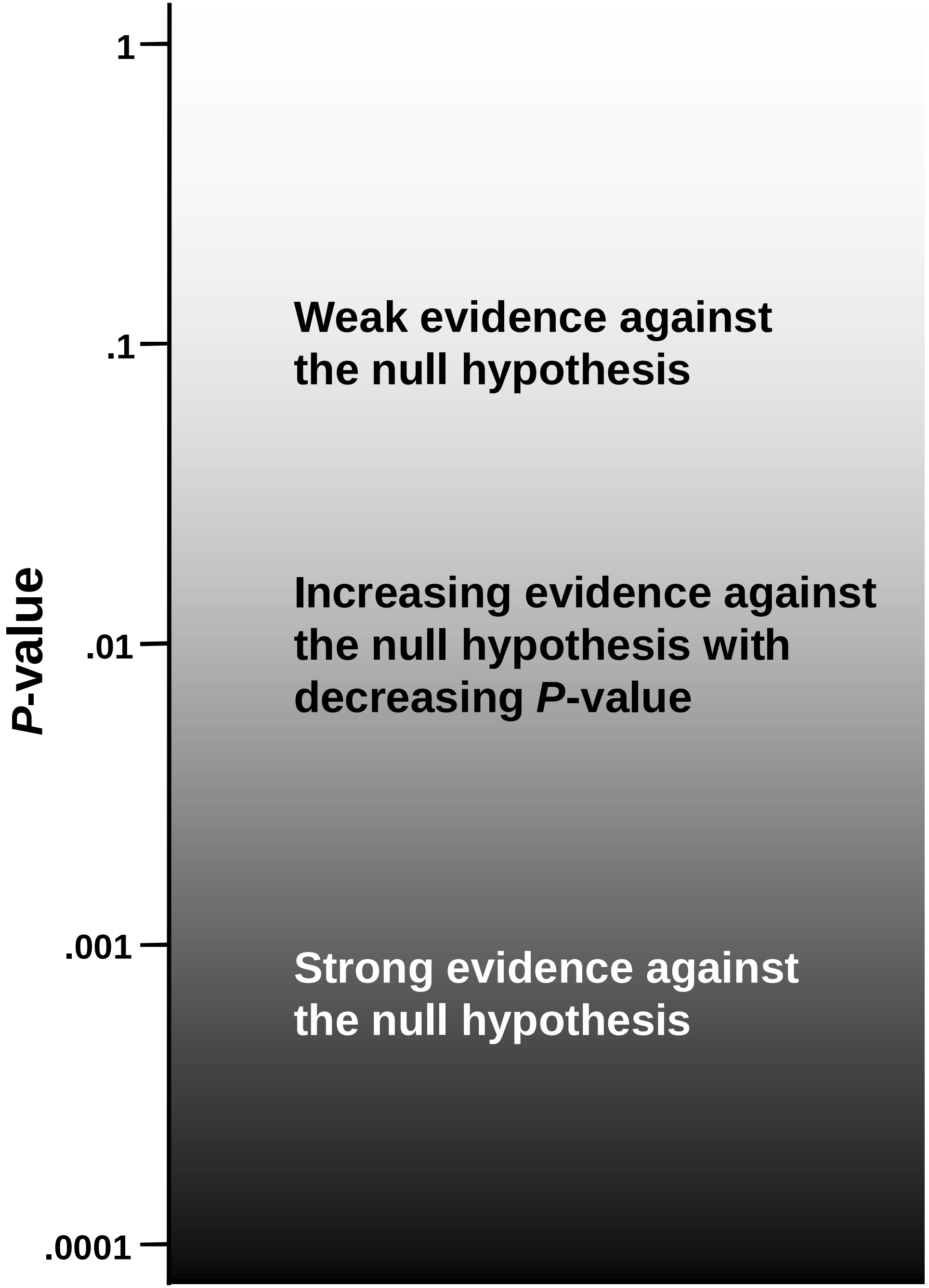
We use this to conduct a ***z*-test**, by deriving a *P*-value – the probability of getting a difference of at least 2.4 (in either direction) if the null hypothesis is true


P value = 0.0164
Prob(£-2.4)=0.0082
Prob(³2.4)=0.0082
z = -2.4
0.4
0.3
0.2
0.1
0
-4
-3
-2.4
-2
-1
0
1
2
2.4
3
4
Standard errors

Interpretation of *P*-values

The smaller the *P*-value, the lower the chance of getting a difference as big as the one observed *if the null hypothesis is true*

Therefore **the smaller the *P*-value, the stronger the evidence against the null hypothesis**


P-value
1
.1
.01
.001
.0001
Weak evidence against the null hypothesis
Increasing evidence against the null hypothesis with decreasing P-value
Strong evidence against the null hypothesis

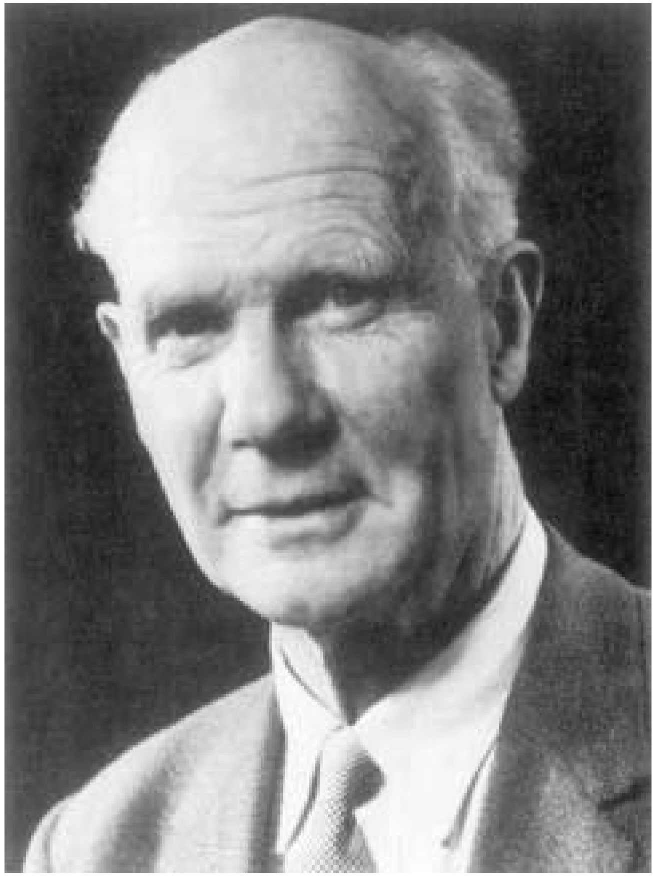
Fisher’s view of significance testing

- Fisher saw the P value as an **informal** index to be used as a measure of discrepancy between the data and the null hypothesis
  - *“The null hypothesis is never proved or established, but is possibly disproved”*
- He advocated 5% significance as a standard level for concluding that there is evidence against the hypothesis tested, though not as an absolute rule
  - *“If P is between .1 and .9 there is certainly no reason to suspect the hypothesis tested. If it is below .02 it is strongly indicated that the hypothesis fails to account for the whole of the facts. We shall not often be astray if we draw a conventional line at .05 …”*

Digression – fiducial probability

- Fisher rejected the frequentist (long run) view of probability
- Instead, he proposed **fiducial inference**, based on so-called pivotal quantities such as the error of observation $e = x - \theta$ in a single observation x of a parameter θ
- Fisher's arguments allowed an objective probability distribution for *e* to be assigned to θ, as its *fiducial distribution*, with *x* fixed at its observed value
  - "Bayesianism without the priors"
  - Never widely accepted

Egon Pearson (1895-1980)



Jerzy Neyman (1894-1981)

Neyman-Pearson hypothesis tests

- Aimed to replace the subjective interpretation inherent in significance testing with an objective, decision-theoretic approach to the results of experiments

	The truth	
Result of experiment	Null hypothesis true	Alternative hypothesis true
Reject null hypothesis	Type I error rate (α)	**Power = 1-** b
Accept null hypothesis		Type II error rate (β)

- By fixing, in advance, the type I (α) and type II (β) error rates, the number of mistakes made over many different experiments would be limited

Neyman-Pearson hypothesis tests

- *"no test based upon a theory of probability can by itself provide any valuable evidence of the truth or falsehood of a hypothesis.*

  But we may look at the purpose of tests from another viewpoint. Without hoping to know whether each separate hypothesis is true or false, we may search for rules to govern our behaviour with regard to them, in following which we insure that, in the long run of experience, we shall not often be wrong."
- In the Neyman-Pearson approach the result of our analysis is simply the rejection or acceptance of the null hypothesis.
  - **We make no attempt to interpret the P value to assess the strength of evidence against the null hypothesis in an individual study**.

Two mutually exclusive approaches

- Decision-theoretic
  - divide our results according to whether or not they are statistically significant
- Subjective
  - the use of a cut-off is simply a guide to the strength of the evidence
- **Transatlantic divide?**
  - the Neyman-Pearson approach suited the industrial view of statistics current in America during the 1930s and 1940s
  - Fisher’s more subjective approach corresponds to a more individualistic view of scientific reasoning

Fisher’s views…..

- “ *… in fact no scientific worker has a fixed level of significance at which from year to year, and in all circumstances , he rejects hypotheses; he rather gives his mind to each particular case in the light of his evidence and his ideas”…..*”
- “*….I am casting no contempt on acceptance procedures, and I am thankful, whenever I travel by air, that the high level of precision and reliability required can really be achieved by such means. But the logical differences between such an operation and the work of scientific discovery by physical or biological experimentation seem to me so wide that the analogy between them is not helpful…..*”

Origins of the 0.05 threshold for statistical significance

- When Fisher was writing Statistical Methods for Research Workers (1925) he applied to Karl Pearson to reproduce a chi-squared table from his *Tables for Statisticians and Biometricians*
  - KP refused, probably because he relied on money from the sales of his tables
- Fisher decided to tabulate quantiles of distributions (0.1, 0.05, 0.02, 0.01) when he produced his own tables
  - Egon Pearson subsequently acknowledged Fisher’s decision to tabulate in this way as one of the key contributions to the development of Neyman-Pearson theory
  - EP also wrote subsequently that the only reason for emphasis on P=0.05 rather than “exact” P-values was the need for manageable tables

What's wrong with Neyman-Pearson in practice?

- Calculation of the Type II error rate requires the specification of a **precise** alternative hypothesis
  - But the use of statistics in medicine became dominated by a division of results into "statistically significant" or "not significant", with little consideration of the type II error rate

	The truth	
Result of experiment	Null hypothesis true	Null hypothesis false
Reject null hypothesis	Type I error rate (α)	**Power = 1-** b
Accept null hypothesis		Type II error rate (β)

What's wrong with Neyman-Pearson in practice?

- Calculation of the Type II error rate requires the specification of a **precise** alternative hypothesis
  - But the use of statistics in medicine became dominated by a division of results into "statistically significant" or "not significant", with little consideration of the type II error rate

Result of experiment	Null hypothesis true
Reject null hypothesis	Type I error rate (α)
Accept null hypothesis	

"Simple laziness! - let a simple technique replace harder thinking"

(Footnote: sample size calculations are of great importance in clinical epidemiology and do require Neyman-Pearson theory)

Confidence intervals (Neyman 1937)

If a distribution is normal *then 95% of observations lie within 1.96 s.d.'s of the mean*

Therefore, in 95% of samples, the interval

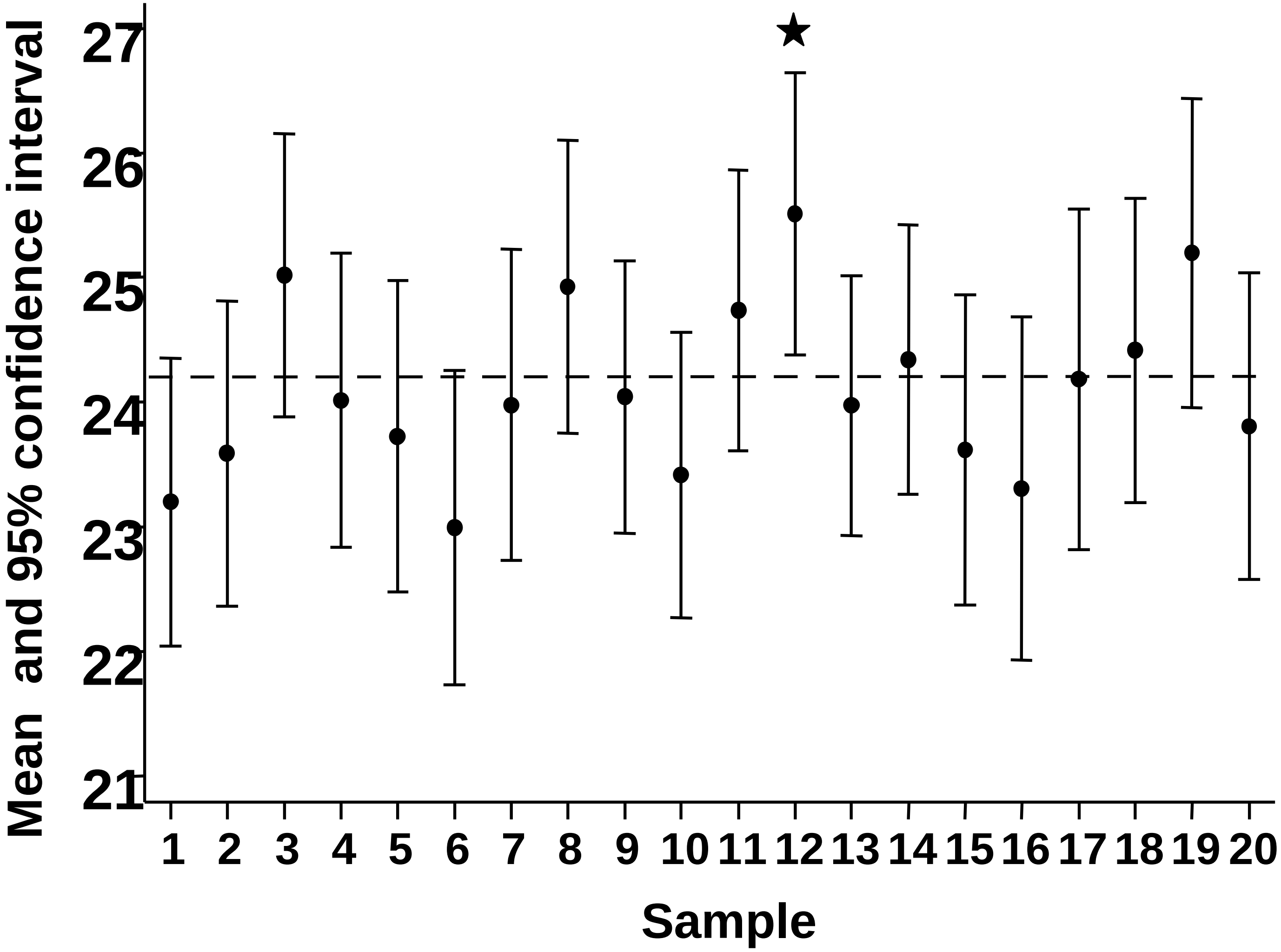
$$(\bar{x}\text{-}1.96\times\text{s.e. to } \bar{x}+1.96\times\text{s.e.})$$

contains the (unknown) population mean

This interval is called a **95% confidence interval**

Understanding confidence intervals

The population mean (*m*) is a fixed unknown number: it is the **confidence interval** that will vary between samples.



20 samples of size 100, from a population with mean 24.2 and s.d. 5.9.

The sample means vary around the population mean *m*

One of the twenty 95% C.I.s does not contain *m*

Promotion of confidence intervals

- During the 1980s, a number of British statisticians tried to improve the presentation of statistical analyses by encouraging authors to present confidence intervals
- Key references:
  - Altman DG, Gore SM, Gardner MJ, Pocock SJ. (1983) Statistical guidelines for contributors to medical journals. *British Medical Journal* **286**:1489-1493
  - Gardner MJ, Altman DG. (1986) Confidence intervals rather than P values: estimation rather than hypothesis testing. *British Medical Journal* **292**:746-750
  - Gardner MJ, Altman DG. (1989) *Statistics with Confidence - Confidence intervals and statistical guidelines*. BMJ Books, London

We need confidence intervals and P values to interpret the results of statistical analyses

Example

- Effect of stroke unit care by stroke subtype (*Stroke* 2002; **33**: 449-455)
  - Randomized trial of treatment in stroke units or general medical wards with stroke team support
  - Separate analyses for 164 patients with large-vessel infarcts and 103 with lacunar infarcts
  - “Stroke units improve the outcome in patients with large-vessel infarcts but not lacunar infarcts”
- Results for mortality and institutionalization:

Large vessel	**Lacunar**
P=0.01	P=0.06
OR 2.8 (95% CI 1.3 to 6.2)	OR 4.9 (95% CI 0.9 to 25.0)

Three common errors

- Potentially clinically important differences observed in small studies, for which P>0.05, are denoted as “non-significant” and ignored
  - **Always examine the confidence interval!**
- Statistically significant (P<0.05) findings are assumed to result from real treatment effects
  - **By definition 1 in 20 comparisons in which the null hypothesis is true will result in P<0.05**
- Statistically significant (P<0.05) findings are assumed to be of clinical importance
  - **Given a large sample size, even a small difference will be statistically significantly different from zero**

Can we get rid of P-values?

- Practical experience says no – why?
  - In the context of RCTs, to justify spending resources on a particular treatment we need, as a minimum, evidence against there being no treatment effect at all
  - In epidemiology, so many factors have been postulated over the years to be associated with a multitude of disease outcomes that some quantification of the possible role of chance in explaining observed results is enduringly useful
  - In choosing a statistical model we have to make decisions about the inclusion or otherwise of different covariates, and different forms of these covariates. This is difficult without some recourse to null hypotheses
- **Easier to interpret a CI without a P-value, than a P-value without a CI**

Interpretation of P-values

	The truth	
Result of experiment	Null hypothesis true	Null hypothesis false
Reject null hypothesis	Type I error rate (α)	Power = 1- β
Accept null hypothesis		Type II error rate (β)
	900	**100**

- Assumptions:
  - **The proportion of false null hypothesis is 10%**
(e.g. by 1985 nearly 300 risk factors for CHD had been identified - it is unlikely that more than a small fraction of these increase the risk CHD)
  - Because studies are often too small, **the average power of studies reported in the medical literature is 50%**
(this is consistent with published surveys of the size of trials)
  - **We conduct 1000 studies**

Interpretation of P-values

	The truth	
Result of experiment	Null hypothesis true	Null hypothesis false
Reject null hypothesis	**45**	Power = 1- β
Accept null hypothesis	**855**	Type II error rate (β)
	900	**100**

- Assumptions:
  - **The proportion of false null hypothesis is 10%**
    (e.g. by 1985 nearly 300 risk factors for CHD had been identified - it is unlikely that more than a small fraction of these increase the risk CHD)
  - Because studies are often too small, **the average power of studies reported in the medical literature is 50%**
    (this is consistent with published surveys of the size of trials)
  - **We conduct 1000 studies**

Interpretation of P-values

	The truth	
Result of experiment	Null hypothesis true	Null hypothesis false
Reject null hypothesis	**45**	**50**
Accept null hypothesis	**855**	**50**
	900	**100**

- Assumptions:
  - **The proportion of false null hypothesis is 10%** (e.g. by 1985 nearly 300 risk factors for CHD had been identified - it is unlikely that more than a small fraction of these increase the risk CHD)
  - Because studies are often too small, **the average power of studies reported in the medical literature is 50%** (this is consistent with published surveys of the size of trials)
  - **We conduct 1000 studies**

Interpretation of P-values

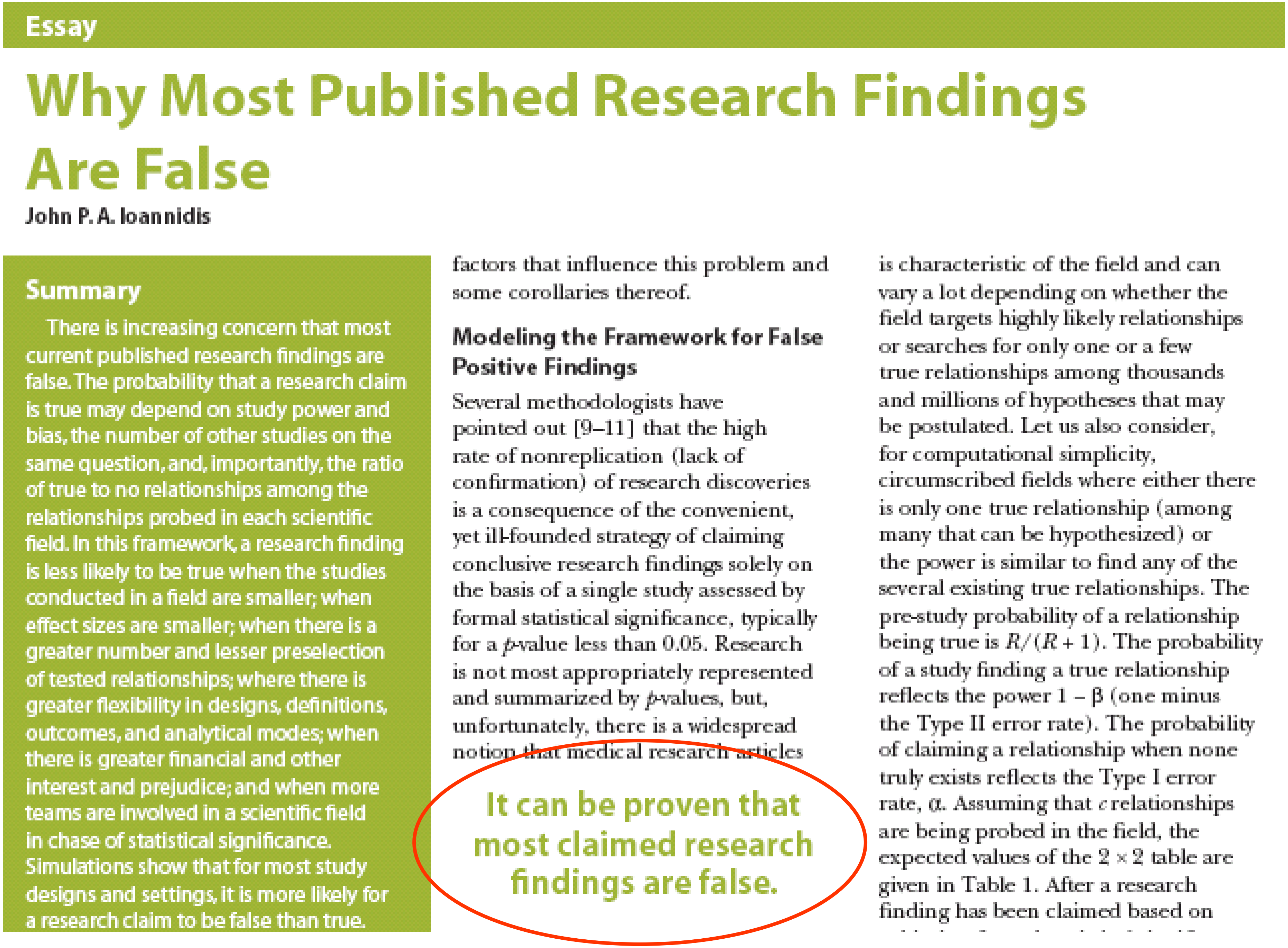
	The truth	
Result of experiment	Null hypothesis true	Null hypothesis false
Reject null hypothesis	**45**	**50**
Accept null hypothesis	**855**	**50**
	900	**100**

- **Of the 95 studies which result in a "statistically significant" (i.e. p<0.05) result, 45 (47%) are true null hypotheses and so are "false alarms"**

(Example adapted from Oakes (1986))

Interpretation of P-values

Proportion of ideas that are correct (null hypothesis false)	Power of study	Percentage of “significant” results that are false-positives		
		P=0.05	P=0.01	P=0.001
80%	20%	5.9	1.2	0.1
	50%	2.4	0.5	0.0
	80%	1.5	0.3	0.0
50%	20%	20.0	4.8	0.5
	50%	9.1	2.0	0.2
	80%	5.9	1.2	0.1
10%	20%	69.2	31.0	4.3
	50%	**47.4**	15.3	1.8
	80%	36.0	10.1	1.1
1%	20%	96.1	83.2	33.1
	50%	90.8	66.4	16.5
	80%	86.1	55.3	11.0



Essay

Why Most Published Research Findings Are False

John P. A. Ioannidis

Summary

There is increasing concern that most current published research findings are false. The probability that a research claim is true may depend on study power and bias, the number of other studies on the same question, and, importantly, the ratio of true to no relationships among the relationships probed in each scientific field. In this framework, a research finding is less likely to be true when the studies conducted in a field are smaller; when effect sizes are smaller; when there is a greater number and lesser preselection of tested relationships; where there is greater flexibility in designs, definitions, outcomes, and analytical modes; when there is greater financial and other interest and prejudice; and when more teams are involved in a scientific field in chase of statistical significance. Simulations show that for most study designs and settings, it is more likely for a research claim to be false than true.

factors that influence this problem and some corollaries thereof.

Modeling the Framework for False Positive Findings

Several methodologists have pointed out [9–11] that the high rate of nonreplication (lack of confirmation) of research discoveries is a consequence of the convenient, yet ill-founded strategy of claiming conclusive research findings solely on the basis of a single study assessed by formal statistical significance, typically for a *p*-value less than 0.05. Research is not most appropriately represented and summarized by *p*-values, but, unfortunately, there is a widespread notion that medical research articles

It can be proven that most claimed research findings are false.

is characteristic of the field and can vary a lot depending on whether the field targets highly likely relationships or searches for only one or a few true relationships among thousands and millions of hypotheses that may be postulated. Let us also consider, for computational simplicity, circumscribed fields where either there is only one true relationship (among many that can be hypothesized) or the power is similar to find any of the several existing true relationships. The pre-study probability of a relationship being true is $R/(R+1)$. The probability of a study finding a true relationship reflects the power $1 - \beta$ (one minus the Type II error rate). The probability of claiming a relationship when none truly exists reflects the Type I error rate, α. Assuming that c relationships are being probed in the field, the expected values of the 2×2 table are given in Table 1. After a research finding has been claimed based on

Ioannidis, *PLOS Medicine*, 2005

Increasing power

- It can be shown, using standard power calculations, that the maximum amount by which a study size would have to be increased is by a factor of:
  - **1.75 if moving from P<0.05 to P<0.01**
  - **2.82 from P<0.05 to P<0.001**
- It is also possible, and generally preferable, to increase power by decreasing measurement error rather than by increasing sample size
- **By doing fewer but more powerful studies it is perfectly possible to stop the discrediting of medical research**

Interpretation of P-values depends on context

- Evidence against the null hypothesis of no treatment effect from a large, high quality, randomized controlled trial, based on the primary outcome specified in the protocol
- Unexpected finding from an epidemiological study
- P-value for interaction, from many such tests

Salvation in Bayes?

- The previous discussion about interpretation of P-values clearly has a Bayesian flavour
- Others are better to discuss Bayesian developments than me…..
- Bayesian methods have recently been used to quantify the size of P-value needed to provide convincing evidence of an association, in genetic association studies (Wacholder et al. *JNCI* 2004)
  - e.g. if the prior odds are 1000:1 against a true association then the posterior odds, after a "significant" finding, are still ~50:1 in favour of the null hypothesis
  - Significance levels between 10^{-4} and 10^{-6} may be needed to provide convincing evidence of association

Suggested guidelines for reporting statistical analyses

(Sterne and Davey Smith, *BMJ* 2001; 322: 226-231)

- The description of differences as “statistically significant” is not acceptable
- Confidence intervals for the main results should always be included:
  - Confidence intervals should not be used as a surrogate means of examining statistical significance at the conventional 5% level
  - Interpretation of confidence intervals should focus on the plausible range of values for the statistic presented

Suggested guidelines for reporting statistical analyses

(Sterne and Davey Smith, *BMJ* 2001; 322: 226-231)

- When there is a meaningful null hypothesis, the corresponding P-value should be reported
- Authors should take a very skeptical view of subgroup analyses – in clinical trials and observational studies
  - The strength of the evidence for interaction - that effects really differ between subgroups - should always be presented
  - Claims made on the basis of subgroup findings should be even more tempered than claims made about main effects
- In observational studies it should be remembered that considerations of confounding and bias are at least as important as the issues discussed here

You don’t need to say “significant”!

- “There was a statistically significant odds ratio of 2.8 (95% CI 1.3 to 6.2) …..”

You don’t need to say “significant”!

- “The odds ratio was 2.8 (95% CI 1.3 to 6.2, P=0.01) …..”

- ~~“P values less than 0.05 were regarded as statistically significant.”~~

What’s your excuse for (ab)using the phrase “statistically significant”?

Common excuses….

- **That's what I was taught!**
  - Fair point. But now you know better….
- **It's all too difficult. I need a simple rule for how to interpret my results**
  - Which of the common errors in interpreting results do you not mind making?
- **I'm using a procedure that doesn't give CIs**
  - Modern statistical methods (e.g. bootstrapping) will solve your problem
- **Journal editors and referees expect me to**
  - Have you tried?
  - Refer to Sterne and Davey Smith (*BMJ* 2001; 322: 226-231)
- **You are wrong. It is useful to use the 0.05 threshold, because…..**