P-values: The good, the bad, and the (really) ugly

WoW

December 10, 2004

Steven Goodman, MD, PhD
sgoodman@jhmi.edu

or, P-values: A search for meaning

WoW

December 10, 2004

Steven Goodman, MD, PhD
sgoodman@jhmi.edu

Central Problem of Inference

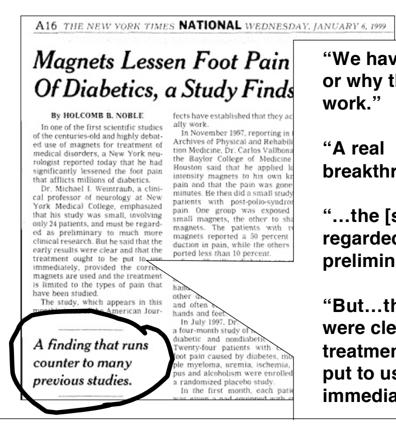
What is the chance that what we say about nature is true?

Things identified as cancer risks

(Altman and Simon, JNCI, 1992)

- Electric Razors
- Broken Arms (in women)
- Fluorescent lights
- Allergies
- Breeding Reindeer
- Being a waiter
- Owning a pet bird
- Hot dogs
- Being short
- Being tall

Having a *refrigerator*

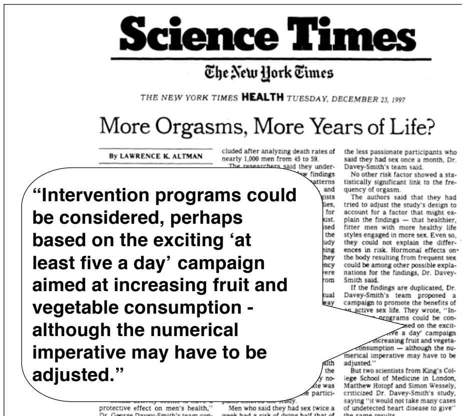


"We have no idea how or why the magnets work."

"A real breakthrough…"

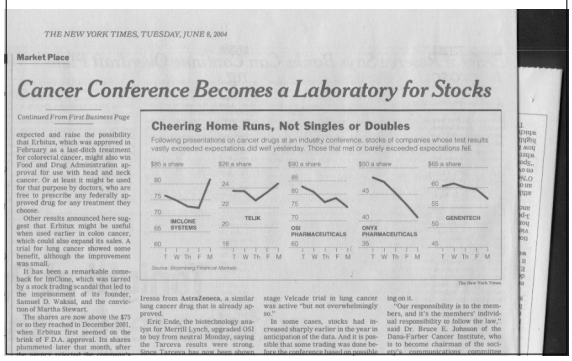
"…the [study] must be regarded as preliminary…."

"But…the early results were clear and... the treatment ought to be put to use immediately."



"Intervention programs could be considered, perhaps based on the exciting 'at least five a day' campaign aimed at increasing fruit and vegetable consumption - although the numerical imperative may have to be adjusted."

Cancer statistics, 2004



THE NEW YORK TIMES, TUESDAY, JUNE 8, 2004

Market Place

Cancer Conference Becomes a Laboratory for Stocks

Continued From First Business Page

Cheering Home Runs, Not Singles or Doubles

Following presentations on cancer drugs at an industry conference, stocks of companies whose test results vastly exceeded expectations did well yesterday. Those that met or barely exceeded expectations fell.

$85 a share — IMCLONE SYSTEMS — $26 a share — TELIK — $90 a share — OSI PHARMACEUTICALS — $50 a share — ONYX PHARMACEUTICALS — $69 a share — GENENTECH

T W Th F M

Source: Bloomberg Financial Markets

The New York Times

expected, and raise the possibility that Erbitux, which was approved in February as a last-ditch treatment for colorectal cancer, might also win Food and Drug Administration approval for use with head and neck cancer. Or at least it might be used for that purpose by doctors, who are free to prescribe any federally approved drug for any treatment they choose.

Other results announced here suggest that Erbitux might be useful when used earlier in colon cancer, which could also expand its sales. A trial for lung cancer showed some benefit, although the improvement was small.

It has been a remarkable comeback for ImClone, which was tarred by a stock trading scandal that led to the imprisonment of its founder, Samuel D. Waksal, and the conviction of Martha Stewart.

The shares are now above the $75 or so they reached in December 2001, when Erbitux first seemed on the brink of F.D.A. approval. Its shares plummeted later that month, after

Iressa from AstraZeneca, a similar lung cancer drug that is already approved.

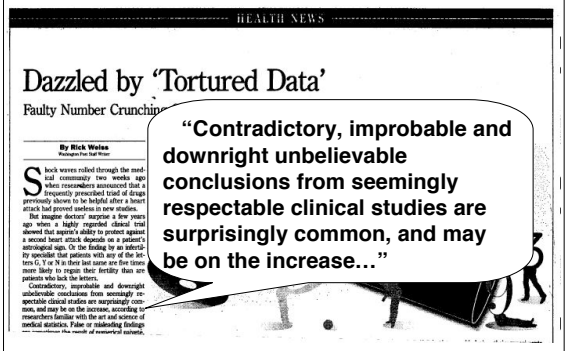
Eric Ende, the biotechnology analyst for Merrill Lynch, upgraded OSI to buy from neutral Monday, saying the Tarceva results were strong. Since Tarceva has now been shown

stage Velcade trial in lung cancer was active "but not overwhelmingly so."

In some cases, stocks had increased sharply earlier in the year in anticipation of the data. And it is possible that some trading was done before the conference based on possible

ing on it.

"Our responsibility is to the members, and it's the members' individual responsibility to follow the law," said Dr. Bruce E. Johnson of the Dana-Farber Cancer Institute, who is to become chairman of the society's communications committee



HEALTH NEWS

Dazzled by 'Tortured Data'

Faulty Number Crunchin

By Rick Weiss
Washington Post Staff Writer

Shock waves rolled through the medical community two weeks ago when researchers announced that a frequently prescribed triad of drugs previously shown to be helpful after a heart attack had proved useless in new studies.

But imagine doctors' surprise a few years ago when a highly regarded clinical trial showed that aspirin's ability to protect against a second heart attack depends on a patient's astrological sign. Or the finding by an infertility specialist that patients with any of the letters G, Y or N in their last name are five times more likely to regain their fertility than are patients who lack the letters.

Contradictory, improbable and downright unbelievable conclusions from seemingly respectable clinical studies are surprisingly common, and may be on the increase, according to researchers familiar with the art and science of medical statistics. False or misleading findings

"Contradictory, improbable and downright unbelievable conclusions from seemingly respectable clinical studies are surprisingly common, and may be on the increase…"

A short research quiz

A study is done on risk factors for childhood leukemia in a suburban community, and the authors state that a surprising association has turned up (i.e., one that they thought had less than a 30% chance of being true before the experiment) p=0.05, OR=2. The probability that this association is real is:

a.) < 75%

b.) 75% to 94.99%

c.) ≥ 95%

How do we represent that question?

- Hypothesis ("Ha"): There is a **SOME effect** of the exposure on leukemia risk.
- Null hypothesis (Ho): There is **NO effect** of the exposure on leukemia risk
- Data (x): OR=2.0, CI 1-4, p=0.05.
- The question was "What is the probability that this association is real?":

$$Pr(Ha \mid x) = ?$$
$$= 1 - Pr(Ho \mid x)$$

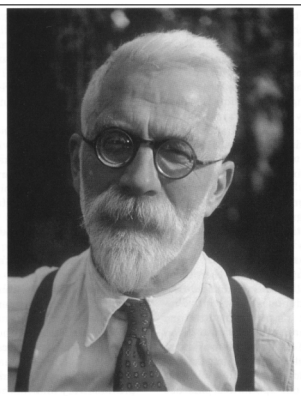


FIG. 2. *R. A. Fisher in 1953, photographed by G. S. Watson on a picnic near Melbourne.*

In search of "p"

…from the world's most definitive statistical sources.

In search of "p"

Armitage P-value definition

"The dividing line between "likely" and "unlikely" classes [of results, under the null hypothesis] is clearly arbitrary, but is usually defined in terms of a probability, *P,* which is referred to as the significance level. Thus, a result would be declared *significant at the 5% level* if the sample were in the class containing those samples most removed from the null hypothesis in the direction of the relevant alternatives, and that class contained samples with a total probability of no more than 5% on the null hypothesis."

"Statistics Made Clear" P-value definition

- A p-value is the probability of obtaining a result as extreme or more extreme than the value of the test statistic, given that the null hypothesis is not rejected, if the dissimilarity is entirely due to chance alone."
- "The p-value is an estimate of the degree to which the result is representative of the population. Commonly selected p-values are arbitrary choices based on general research experience."

"Intuitive Biostatistics" P-value definition

"Assuming the null hypothesis is true, calculate the likelihood of observing various results. Determine the fraction of those possible results in which the difference…is as large or larger than what you observed. The answer…is called the *P value."*

"Intuitive Biostatistics" P-value definition, cont.

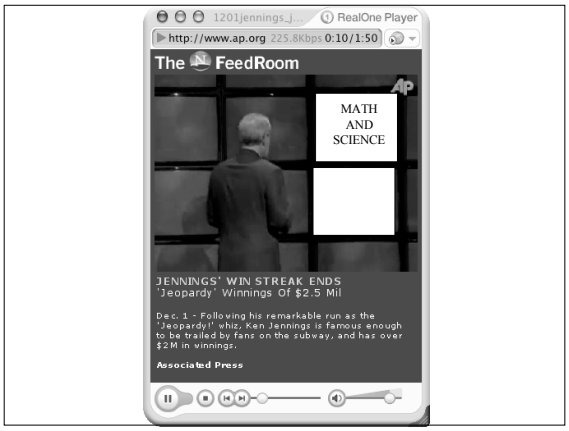
"Thinking about P values seems quite counterintuitive at first, as you must use backwards, awkward logic. Unless you are a lawyer or a Talmudic scholar…you will probably find this sort of reasoning uncomfortable."

<u>After calculating the p-value:</u>

"What conclusions should you reach? That's up to you."

In search of "p"

…from the world's smartest person.



In search of "p"

…from the school's most successful person.

Message from the Mount

In search of "p"

…from the world's wisest person.

The Final Quest

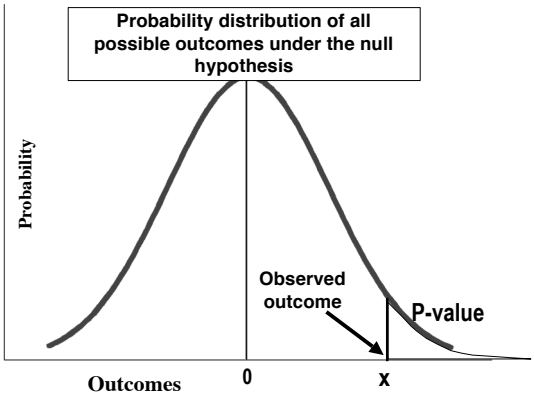
The P-value is ….

- …not almost anything intuitive that you can think of.
- … a rough guide to the strength of statistical evidence for the null hypothesis versus the hypothesis that you happen to have observed the exact truth.

The P-value is….

- The probability of getting a result as or more extreme than the observed result, if the null hypothesis (of chance) were true.

$$\text{P-value} = \Pr(X \geq x \mid \text{Ho})$$



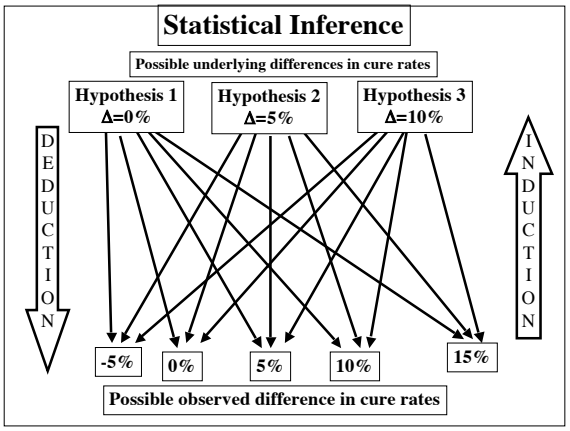
What the P-value is not….

P-value = Pr(X ≥ *x* I Ho)	
The probability of the null hypothesis, given the data.	Pr(Ho I *x*)
The probability of the data under Ho (i.e. if only chance were operating).	Pr(*x* I Ho)
The probability that the data were observed by chance.	Pr(Ho I *x*)
The probability that a non-null association is "real", given the data	Pr(Ha I *x*) =1-Pr(Ho I *x*)

How do we calculate Pr(H|D), the probability of the truth of our claims?

Bayes Theorem

$$\underbrace{\frac{\Pr(H_0 \mid \text{Data})}{\Pr(H_1 \mid \text{Data})}}_{\text{Post–test Odds}} = \underbrace{\frac{\Pr(H_0)}{\Pr(H_1)}}_{\text{Pre–test Odds}} \times \underbrace{\frac{\Pr(\text{Data} \mid H_0)}{\Pr(\text{Data} \mid H_1)}}_{\text{Likelihood Ratio}}$$



Rarity is not enough: Evidence is relative

Someone wins the "pick 5" lottery……. $p=10^{-5}$
The winner is the son of the person who picked the balls.

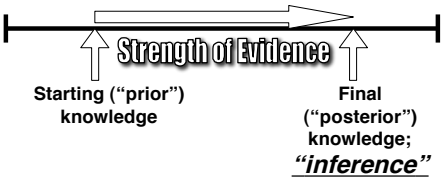
A roulette wheel comes up 3, 14, 6 and 27…. $p=6 \times 10^{-7}$
You notice the numbers are adjacent.

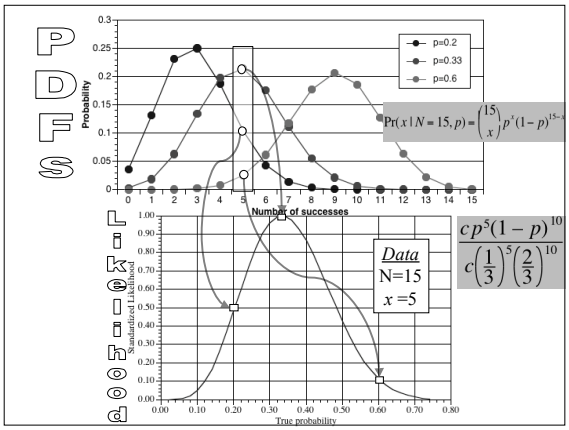
A previously unsuspected and implausible association shows up in a study…… p=0.01
A reviewer suggests a biologic explanation for the finding .

Bayes Theorem

$$\underbrace{\frac{\Pr(H_0 \mid \text{Data})}{\Pr(H_1 \mid \text{Data})}}_{\text{Post–test Odds}} = \underbrace{\frac{\Pr(H_0)}{\Pr(H_1)}}_{\text{Pre–test Odds}} \times \underbrace{\frac{\Pr(\text{Data} \mid H_0)}{\Pr(\text{Data} \mid H_1)}}_{\text{Likelihood Ratio}}$$

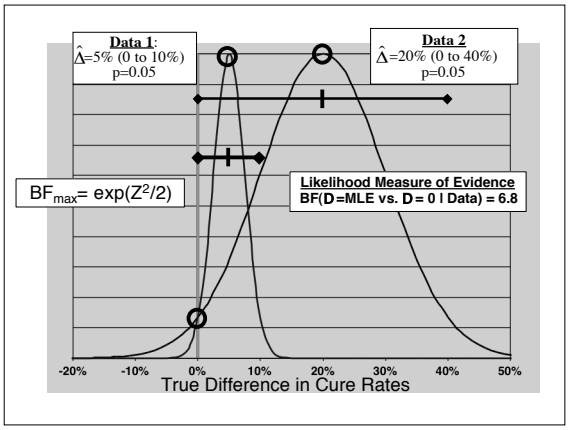
Bayes Theorem





LR/Bayes factor vs. P-value

P-value	Bayes factor
Non-comparative	Comparative
Observed + hypothetical data	Only observed data
Alternative hypothesis implicit, partly data-defined	Alternative hypothesis explicit, pre-defined
Evidence only negative	Evidence negative or positive
Sensitive to stopping rules	Insensitive to stopping rules
No formal justification or interpretation	Formal justification and interpretation



P-values <-> Bayes factors

- For any outcome that has an (approximately) Gaussian distribution, the maximum Bayes Factor (or likelihood ratio) associated with a given Z-score, is:

Max LR(Ha vs. Ho I Z) = exp($Z^2/2$)

P-value : Bayes factor : Inference

Strength of			*Maximum* final probability of Ha when prior probability is:		
Evidence	P	LR	25%	50%	75%
Weak	0.1	4	56	79	92
Mod	0.05	7	69	87	95
Mod	0.03	11	78	91	97
Mod/Strong	0.01	28	90	96	99
Very Strong	0.001	203	98.5	99.5	99.8

A short research quiz

A study is done on risk factors for childhood leukemia in a suburban community, and the authors state that a surprising association has turned up (i.e., one that they thought had less than a 30% chance of being true before the experiment) p=0.05. The probability that this association is real is:

a.) < 75%

b.) 75% to 94.99%

c.) ≥ 95%

Inferential calculations

Prior probability = 30%

What is the probability that relationship is real after p=0.05?

Prior odds = 0.3/0.7 = 0.43

Max. LR(+) = exp(1.96^2/2) = 6.8 ***[the evidence]***

Odds of disease (+) = LR(+) x Prior Odds
= 6.8 x 0.43 = 2.9

The inference:
Max probability of Ha = 2.9/3.9 = 74.3%

The Good

The New England
Journal of Medicine

Volume 332 APRIL 13, 1995 Number 15

DIETARY INTAKE OF MARINE n−3 FATTY ACIDS, FISH INTAKE, AND THE RISK OF CORONARY DISEASE AMONG MEN

ALBERTO ASCHERIO, M.D., ERIC B. RIMM, SC.D., MEIR J. STAMPFER, M.D., EDWARD L. GIOVANNUCCI, M.D., AND WALTER C. WILLETT, M.D.

Abstract *Background.* It has been hypothesized that a diet containing n−3 fatty acids from fish reduces the risk of coronary heart disease, but few large epidemiologic studies have examined this question.

Methods. In 1986, 44,895 male health professionals, 40 to 75 years of age, who were free of known cardiovascular disease completed detailed and validated dietary questionnaires as part of the Health Professionals Follow-up Study. During six years of follow-up, we documented 1543 coronary events in this group: 264 deaths from coronary disease, 547 nonfatal myocardial infarctions, and 732 coronary-artery bypass or angioplasty procedures.

Results. After controlling for age and several coronary risk factors, we observed no significant associations between dietary intake of n−3 fatty acids or fish intake and the risk of coronary disease. For men in the top fifth of the group in terms of intake of n−3 fatty acids (median, 0.58 g per day), the multivariate relative risk of coronary heart disease was 1.12 (95 percent confidence interval, 0.96 to 1.31), as compared with the men in the bottom fifth (median, 0.07 g per day). For men who consumed six or more servings of fish per week, as compared with those who consumed one serving per month or less, the multivariate relative risk of coronary disease was 1.14 (95 percent confidence interval, 0.86 to 1.51). The risk of death due to coronary disease among men who ate any amount of fish, as compared with those who ate no fish, was 0.74 (95 percent confidence interval, 0.44 to 1.23), but the risk did not decrease as fish consumption increased.

Conclusions. Although the possibility of residual confounding by unmeasured factors cannot be entirely excluded, these data suggest that increasing fish intake from one to two servings per week to five to six servings per week does not substantially reduce the risk of coronary heart disease among men who are initially free of cardiovascular disease. (N Engl J Med 1995;332:977-82.)

Honest conclusions

We have no convincing explanation for the suggestion of an increased frequency of coronary-artery bypass surgery among men with higher fish intake in this study.

We have no convincing explanation for the suggestion of an increased frequency of coronary-artery bypass surgery among men with higher fish intake in this study. Perhaps men with higher fish intake are more health-conscious and more willing to undergo angiography and elective coronary surgery. We also cannot exclude the possibility that coronary surgery is less likely to be performed in geographic areas where fish may be less available. However, relative-risk estimates for coronary-artery bypass grafting or myocardial infarction did not change materially after adjustment for the region of residence (data not shown).

Our results suggest that increasing fish intake from one to two servings per week to five to six servings per week is unlikely to reduce the risk of coronary disease substantially among men without preexisting cardiovascular disease. However, an effect of fish or fish oil at lower or higher levels of intake, or among persons with dietary habits or other risk factors that are markedly different from those of the men in our cohort, cannot be excluded by these data.

We are indebted to the participants in the Health Professionals Follow-up Study for their continued cooperation and participation; to

The Bad

The Bad: Confusion of evidence and inference

"This is the first study to demonstrate a therapeutic benefit of corticosteroids in chronic fatigue syndrome (p=0.06)….."
(JAMA, 1998)

- Mechanism not shown
- Inconsistent with prior studies
- Other endpoints inconsistent

ORIGINAL INVESTIGATION

A Randomized, Controlled Trial of the Effects of [redacted] on Outcomes in Patients Admitted to the Coronary Care Unit

William S. Harris, PhD; Manohar Gowda, MD; Jerry W. Kolb, MDiv; Christopher P. Strychacz, PhD; James L. Vacek, MD; Philip G. Jones, MS; Alan Forker, MD; James H. O'Keefe, MD; Ben D. McCallister, MD

Context: [redacted] has been a common response to sickness [redacted] but it has received little scientific attention. The positive findings of a previous controlled trial of [redacted] have yet to be replicated.

Objective: To determine whether [redacted] or hospitalized, cardiac patients will reduce overall adverse events and length of stay.

Design: Randomized, controlled, double-blind, prospective, parallel-group trial.

Setting: Private, university-associated hospital.

Patients: Nine hundred ninety consecutive patients who were newly admitted to the coronary care unit (CCU).

Intervention: At the time of admission, patients were randomized to receive [redacted] group) or not (usual care group). [redacted]

Main Outcome Measures: The medical course from CCU admission to hospital discharge was summarized in a CCU course score derived from blinded, retrospective chart review.

Results: Compared with the usual care group (n = 524), the [redacted] group (n = 466) had lower mean ± SEM weighted (6.35 ± 0.26 vs 7.13 ± 0.27; *P* = .04) and unweighted (2.7 ± 0.1 vs 3.0 ± 0.1; *P* = .04) CCU course scores. Lengths of CCU and hospital stays were not different.

Conclusions: [redacted] was associated with lower CCU course scores. This result suggests that [redacted] may be an effective adjunct to standard medical care.

Arch Intern Med. 1999;159:2273-2278

SPECIAL ARTICLE

A Psychometric Experiment in Causal Inference to Estimate Evidential Weights Used by Epidemiologists

C. D'Arcy J. Holman, Diane E. Arnold-Reed, Nicholas de Klerk, Christine McComb, and Dallas R. English

A psychometric experiment in causal inference was performed on 159 Australian and New Zealand epidemiologists. Subjects each decided whether to attribute causality to 12 summaries of evidence concerning a disease and a chemical exposure. The 1,748 unique summaries embodied predetermined distributions of 19 characteristics generated by computerized evidence simulation. Effects of characteristics of evidence on causal attribution were estimated from logistic regression, and interactions were identified from a regression tree analysis. Factors with the strongest influence on the odds of causal attribution were statistical significance (odds ratio = 4.5 if $0.001 \le P < 0.05$ and 7.2 if $P < 0.001$, vs $P \ge 0.05$); refutation of alternative explanations (odds ratio = 8.1 for no known confounder vs none adjusted); strength of association (odds ratio = 2.0 if $1.5 <$ relative risk ≤ 2.0 and 3.6 if relative risk > 2.0, vs relative risk ≤ 1.5); and adjunct information concerning biological, factual, and theoretical coherence. The refutation of confounding reduced the cutpoint in the regression tree for decision-making based on strength of association. The effect of the number of supportive studies reached saturation after it exceeded 12 studies. There was evidence of flawed logic in the responses concerning specificity of effects of exposure and a tendency to discount evidence if the *P*-value was a "near miss" ($0.050 < P < 0.065$). Evidential weights based on regression coefficients for causal criteria can be applied to actual scientific evidence. (Epidemiology 2001;12:246–255)

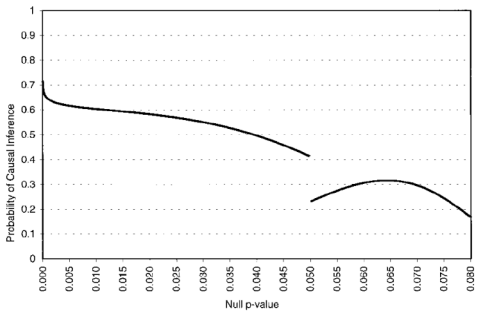


FIGURE 3. Estimated probability of causal attribution according to the null *P*-value, modeled using fractional polynomials with a cutpoint at $P = 0.05$.

Commentary

"The other depressing result is the 20% gap in the authors' Figure 3 between the proportions of epidemiologists who declared causality when confronted with *P*-values abutting the shopworn 0.05 benchmark. Every epidemiologist has enough innate common sense to know that there is no meaningful difference between $P = 0.04$ and $P = 0.06$, if only we were not brainwashed into believing otherwise... Those who persist in teaching null hypothesis testing uncritically to epidemiology students should have Figure 3 tattooed onto their foreheads in reverse image, to remind them with each glance into a mirror of the pox they continue to spread upon our field."

Poole C, "Causal Values," Epidemiology, 12:139-141, 2001

The [Pretty] Ugly

Uncontrolled error

- **Table 16.1**: Physical characteristics of 22 patients (mean ± sem) pre-operatively and after post-operative weight reduction

Characteristic	Pre-operative	Post-operative	Significance
Weight	146 ± 4	87.9 ± 2.9	***
Height	170 ± 1.6	170 ± 1.6	NS
Age	34.7 ± 1.9	36.6 ± 1.9	***

Paired t-tests, ***P<0.001. NS = Not significant
From Acta Med Scand 1979; 205:367

FDA Discussion

(Fisher, CCT, 20:16-39,1999)

L. Moyé, MD, PhD

"What we have to wrestle with is how to interpret p-values for secondary endpoints in a trial which frankly was negative for the primary. …In a trial with a positive endpoint…you haven't spent all of the alpha on that primary endpoint, and so you have some alpha to spend on secondary endpoints….In a trial with a negative finding for the primary endpoint, you have no more alpha to spend for the secondary endpoints."

FDA Discussion, cont.

(Fisher, CCT, 20:16-39,1999)

Dr. Lipicky: What are the p-values needed for the secondary endpoints? …Certainly we're not talking 0.05 anymore. …You're out of this 0.05 stuff and I would have like to have seen what you thought was significant and at what level…

What p-value tells you that it's there study after study?

Dr. Konstam: …what kind of statistical correction would you have to do that survival data given the fact that it's not a specified endpoint? I have no idea how to do that from a mathematical viewpoint.

Confusion of evidence and inference

"The results were insignificant because of small sample size."

Instead of:

"The evidence for the effect was modest, but we believe the relationship exists because of…"

- Prior studies with similar results
- Consistency with known mechanism
- Coherence of multiple outcomes within study

Confusion of evidence and inference

"Of the 40 variables examined, only liver cancer was caused by transfusions (p=0.01)."

Confusion of evidence and inference

Instead of:

"There was moderate evidence (LR=25) for the relationship between liver cancer and transfusions, but this was not strong enough to make the association highly likely because of:

- Prior studies with different results
- No excess of liver cancer in populations with frequent transfusions
- No accepted mechanism...

Take-to-happy-hour messages

- There are no "negative" or "positive" studies - only ones that supply weak and strong evidence, for various hypotheses.
- No formula based on the data alone can tell us how sure we should be about a conclusion, which is based on combining the statistical evidence with biologic or mechanistic understanding.
- I'd tell you to forget all about "testing", but I've run out of time, so just keep doing it.

RA Fisher on statistics education

"I am quite sure it is only personal contact with ... the natural sciences that is capable to keep straight the thought of mathematically-minded people...I think it is worse in this country [the USA] than in most, though I may be wrong. Certainly there is grave confusion of thought. We are quite in danger of sending highly trained and intelligent young men out into the world with tables of erroneous numbers under their arms, and with a dense fog in the place where their brains ought to be. In this century, of course, they will be working on guided missiles and advising the medical profession on the control of disease, and there is no limit to the extent to which they could impede every sort of national effort." 1958

Final thoughts

"What used to be called judgment is now called prejudice, and what used to be called prejudice is now called the null hypothesis....it is dangerous nonsense (dressed up as 'the scientific method') and will cause much trouble before it is widely appreciated as such."

A.W.F. Edwards (1972)

LET'S HOPE LIFE ON MARS