Madrid 2011
Statistics and Scientific Method I
The Controversy about Hypothesis Testing

Wolfgang Rolke
University of Puerto Rico – Mayaguez
Consultant with CMS Statistics Committee

Outline of Talk

What the conference was about

What I thought the conference was going to be about

And was meant to be about by the conveners

Jose Bernardo

Hypothesis Testing from a Decision Theory Viewpoint: A General Objective Bayesian Approach

- Problem with Bayesian Hypothesis testing: If there is a sharp null H0: $\theta=\theta 0$ prior needs have $P(\theta=\theta 0)=\alpha>0$
- Always subjective!
- Different prior from estimation problem
- Generally not invariant

Bernardo cont.

Bernardo's idea: use decision theory but instead of the usual 0-1 loss function use a "smooth" one

Intrinsic Discrepancy Loss Function:

$\delta(p1,p2)= \min [\ K(p1|p2), K(p2|p1)]$

$K(p1|p2)=\int p1 \log(p1/p2)$

is the Kullback-Leibler divergence

Bernardo cont.

(In) Famous Example: ESP

Jahn, Dunne and Nelson (1987) with RNG:

104,490,000 trials, 52,263,471 successes

Estimated probability 0.5001768

H0: $p=0.5$ vs Ha: $p \neq 0$

Frequentist test(s): p_value= 0.0003

Art De Vos and Marc Francke (Free University Amsterdam) No More Null Hypotheses, Just Decisions

Premise: the main objective for hypothesis testing is decision making

Bayesians know how to do this, but it's hard work

Frequentist hypothesis testing is easy:

reject H0 if $p<\alpha$

But it is easy because the costs of wrong decisions are

Art De Vos and Marc Francke cont.

- Idea: find α using Bayesian decision theory:

Idea: find α using Bayesian decision theory:

Let S be some test statistic, and BF(S) it's Bayes factor

Let S be some test statistic, and BF(S) it's Bayes factor

$$BF(H_0)=\frac{\pi(H_0|x)}{1-\pi(H_0|x)}\Big/\frac{\pi(H_0)}{1-\pi(H_0)}$$

BF(H0)=/

Let К=[π(H_0)L(1,0)]/[π(H_1)L(0,1)]

Let К=[π(H0)L(1;0)]/[π(H1)L(0,1)]

Then α=P(BF(S)> К|H_0)

Then α=P(BF(S)> К|H0)

Similar to Bernardos work in that it makes use of decision theory, but leads to subjective choices of α

Similar to Bernardos work in that it makes use of decision theory, but leads to subjective choices of α

Zeynep Baskurt and Michael Evans

University of Toronto

Hypothesis Assessment via Bayes Factors and Relative Belief Ratios

- Bayes factor: $BF(H_0) = \frac{\pi(H_0|x)}{1-\pi(H_0|x)} / \frac{\pi(H_0)}{1-\pi(H_0)}$

relative believe ratio: $RB(H_0) = \frac{\pi(H_0|x)}{\pi(H_0)}$

what if $\pi(H_0)=0$?

Zeynep Baskurt, Michael Evans cont.

- usual solution: =γ>0

usual solution: $\pi(H_0)$=γ>0

their solution: if $H_0: \theta = \theta_0$ define a transformation $\psi = \Psi(\theta)$ and $H_0 = \Psi^{-1}\{\psi_0\}$

their solution: if 0 define a transformation

ψ=Ψ() and H0= Ψ-1{ψ0}

"embed" in larger set $\psi_0 \in C_\varepsilon(\psi_0)$ that "shrinks" to ψ_0 as $\varepsilon \downarrow 0$

"embed" in larger set

that "shrinks" to ψ0 as ε

choose Ψ =d_{H0} where $d_{H0}(\theta)$ is a measure of the distance from θ to H_0. Then

$$BF(0) = RB(0) = \frac{\pi_\psi(0|x)}{\pi_\psi(0)}$$

choose Ψ =dH0 where dH0() is a measure of the distance from to H0. Then

Valen Johnson (University of Texas M.D. Anderson Cancer Center)

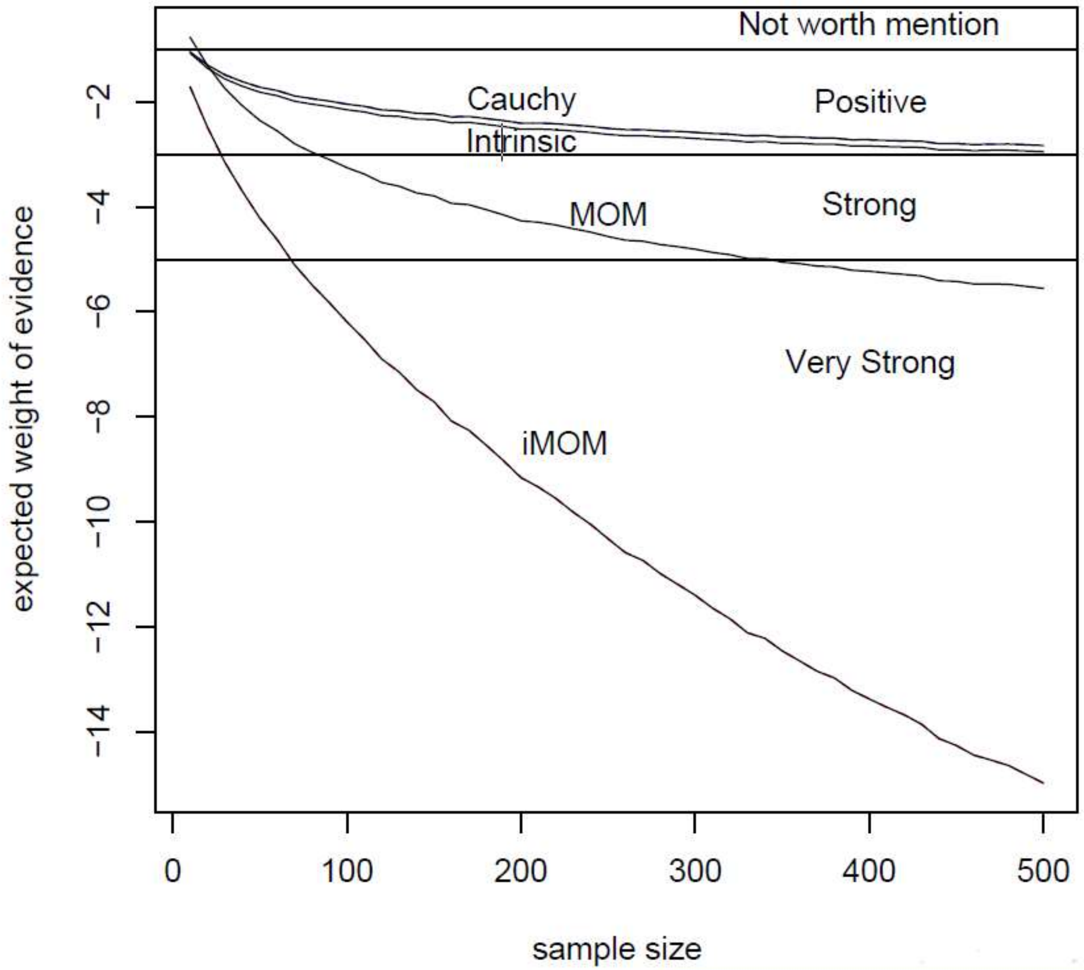
On The Importance of Distinguishing Between Hypotheses: The Role of Non-local Prior Densities in Bayesian Hypothesis Testing and Model Selection

- Johnson defined non-local prior alternative prior densities as prior densities that take the value of 0 for all parameter values consistent with the null hypothesis.
- Essentially all standard Bayesian hypothesis tests of point null hypotheses define alternative hypotheses with priors that take their maximum value at or near the null hypothesis value.

Valen Johnson, cont.

- In many applications, the use of local alternative prior densities (e.g., intrinsic priors, fractional Bayes factors) makes it impossible to obtain strong evidence in favor of a true null hypothesis.
- The use of non-local prior densities in Bayesian hypothesis testing results in much faster accumulation of evidence in favor of true null hypotheses and true alternative hypotheses.

Valen Johnson, cont.



Valen Johnson, cont.

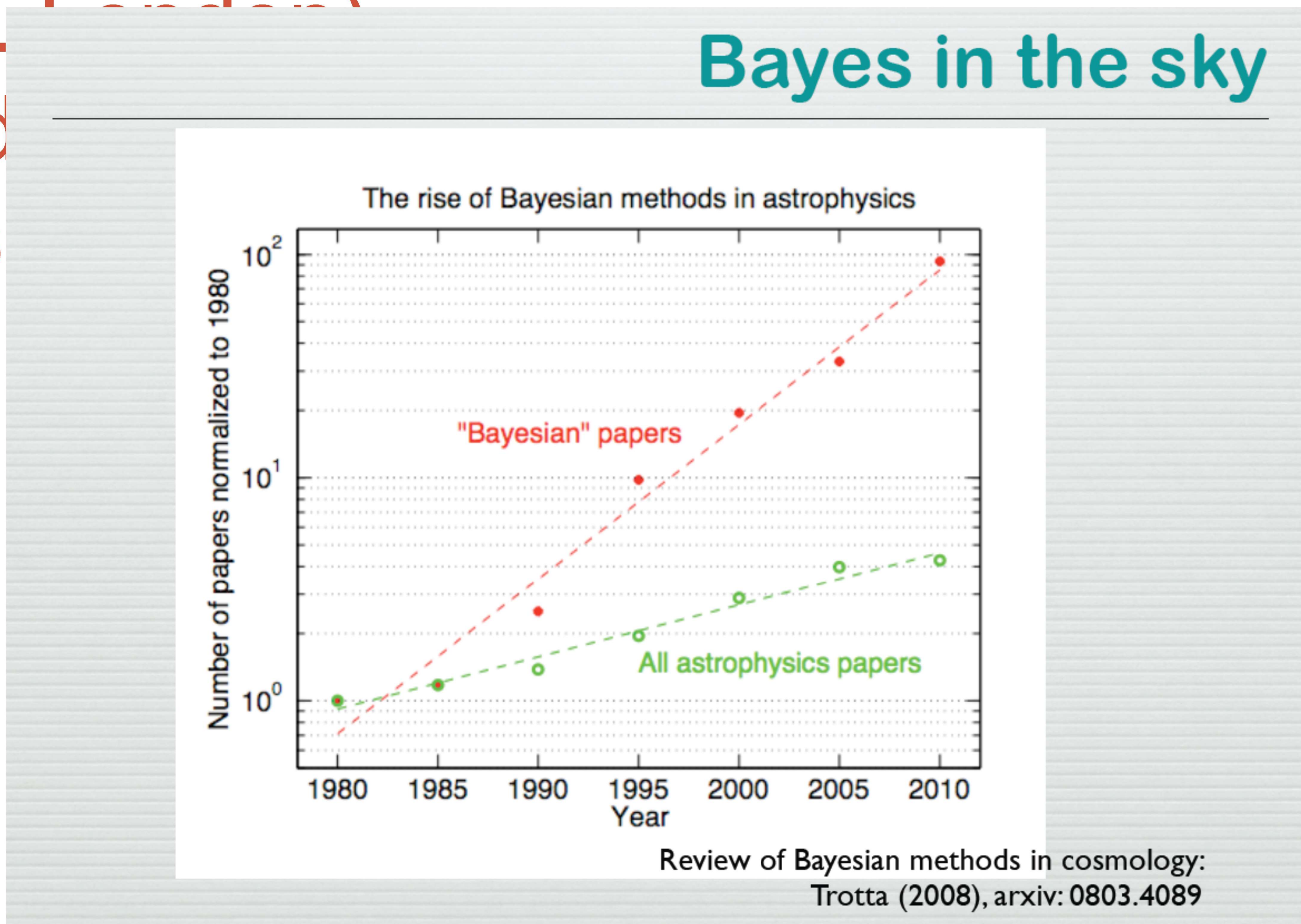
- Results for hypothesis testing using non-local priors available at *http://blades.byu.edu/seminar/valjohnsonJRSSB.pdf*
- Preprint of forthcoming *Journal of American Statistical Association* article describing Bayesian variable selection based on non-local prior densities available at *http://biostats.bepress.com/mdandersonbiostat/paper67/*

Trotta, A. Jaffe, D. Mortlock and D.Van Dyke

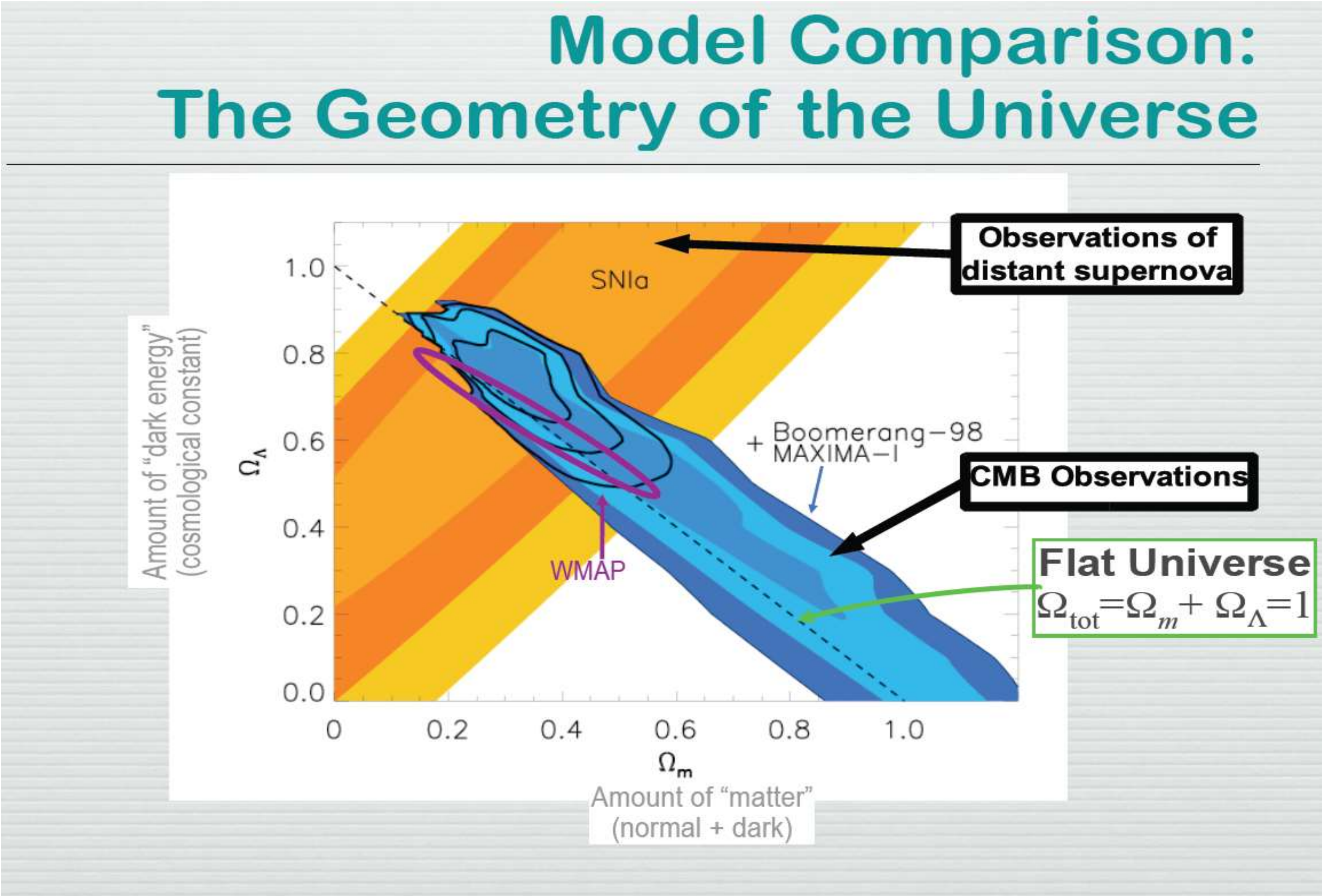
(I.C.

Mod

Cos



A. Jaffe, cont.



Model Comparison:
The Geometry of the Universe
Observations of distant supernova
SNIa
1.0
0.8
0.6
0.4
0.2
0.0
Ω_Λ
Amount of "dark energy" (cosmological constant)
+ Boomerang−98
MAXIMA−I
CMB Observations
WMAP
Flat Universe
$\Omega_{tot}=\Omega_m+\Omega_\Lambda=1$
0
0.2
0.4
0.6
0.8
1.0
Ω_m
Amount of "matter" (normal + dark)

A. Jaffe, cont.

Results: current model comparison

- A positive lnB favours the flat model over curved one

prior = 1/3 prior = 2/3

Data sets and models	$\ln B_{01}$	$\ln B_{0-1}$	$p(\mathcal{M}_0\|d)$	$p(N_U = \infty\|d)$	Notes
				Astronomer's prior (flat in Ω_κ)	
WMAP5+BAO ($w = -1$)	4.1	5.3	0.98	0.98	Moderate evide
WMAP5+BAO+SNIa ($w = -1$)	4.2	5.3	0.98	0.98	Moderate evide
WMAP5+BAO ($w \neq -1$)	1.0	6.1	0.74	0.74	Weak evidence
WMAP5+BAO+SNIa ($w \neq -1$)	3.9	5.3	0.98	0.98	Moderate evide
				Curvature scale prior (flat in o_κ)	
WMAP5+BAO ($w = -1$)	0.4	0.6	0.45	0.69	Inconclusive
WMAP5+BAO+SNIa ($w = -1$)	0.4	0.6	0.45	0.69	Inconclusive
WMAP5+BAO ($w \neq -1$)	−0.8	0.5	0.26	0.42	Inconclusive
WMAP5+BAO+SNIa ($w \neq -1$)	0.3	0.6	0.44	0.67	Inconclusive

posterior probability of flatness

posterior probability of an infinite Universe

Vardanyan, Trotta & Silk (2009)

- http://astro.imperial.ac.uk/

Deborah Mayo (Virginia Tech) Are Frequentist Significance Tests Inconsistent? Breaking through "Birnbaum's Breakthrough"

- *Strong Likelihood Principle: (SLP)* If two experiments result in proportional likelihoods they should yield the same inference
- *Conditionality Principle (CP):* only the experiment that was actually done matters, not any experiment we could have done but didn't.
- *Sufficiency Principle (SP):* a sufficient statistic summarizes the results of an experiment with no loss of information.

Deborah Mayo, cont.

Birnbaum 1962: CP+SP → SLP

L.J. Savage: Without any intent to speak with exaggeration or rhetorically, it seems to me that this is really a historic occasion ...

But not to take the principle (SLP) seriously no longer seems possible ...

I can't know what everyone will do, but I suspect that once the likelihood principle is widely recognized, people will not long stop at the halfway house but will go forward and accept the implications of personalistic probability for statistics.

Deborah Mayo, cont

Proof of Birnbaum's theorem can be found in

- Casella and Berger (2nd Ed) p294
- every other Statistics textbook in the last 50 years
- and yet, Deborah Mayo claims to show that Birnbaums proof is wrong.
- So maybe Frequentist statistics isn't altogether silly.
- Mayo, D. (2010). " An Error in the Argument from Conditionality and Sufficiency tc " in *Error and Inference: Recent Exchanges on Experimental Reasoning, Reliability and the Objectivity and Rationality of Science* (D Mayo and A. Spanos eds.), Cambridge: Cambridge University Press: 305-14.

K.Brewer, G.Hayes and A. Gillison (Australian National University) Using Fisher's p to Measure Significance

4-part paper, using both information criteria (ICs) and Bayesian hypothesis tests.

- Part 1 shows that if the null hypothesis is precise, p can be grossly misinterpreted.
- BIC can be grossly parsimonious, so in need of additional penalty terms.
- new IC is then a simple function of Student's T, thus also a function of the p-value.
- It is also, for practical purposes, always intermediate between the AIC and the BIC.

K.Brewer, G.Hayes and A. Gillison, cont.

- Part 2 develops an approximately and asymptotically Bayesian hypothesis test, using Benford's Law of Numbers to specify a "complete ignorance" prior for the alternative hypothesis.
- This test is also equivalent to the new IC of Part 1.
- Part 3 applies the above test to 1294 regression slopes from a biodiversity data set.
- Part 4 develops a related and fully Bayesian hypothesis test using even fewer assumptions.

Ken.Brewer@anu.edu.au

Kevin Hoover (Duke)
The Role of Hypothesis Testing in the Molding of Econometric Models

Econometrics and Philosophy of Science

Upshot: Economists have a lot of models supposedly derived from theory, but they don't test those models, and their theories are not very good.

Scary, because many of them work for banks, and the banks have our money!

The Controversy about Null Hypothesis Significance Testing (NHST)

- Bakan (1966) *A great deal of mischief has been associated with NHST*
- Carver (1993): *NHST is a corrupt form of the scientific method*
- *Meehl (1968) NHST is a potent but sterile intellectual rake who leaves in his merry path a long train of ravished maidens but no viable scientific offspring*
- In 1996 the American Psychological Association formed a high level task force which considered to recommend banning NHST from any of their journals.
- For once, not a Frequentist vs Bayesian issue
- Mostly discussed in Psychology, Sociology, Education and

The Controversy about NHST

So what is the problem? There are two separate issues:

1) NHST (especially p-values) are badly understood, misused and misinterpreted:

- p-value is the probability that the null hypothesis is true
- $\alpha=0.05$, so if $p=0.045$ reject the null but if $p=0.055$ do not.

p-value is a random variable, with an often surprisingly

The Controversy about NHST

2) NHST is used when it probably should not be

In many fields the null is usually known to be false a priori:

H0: Median Income = \$25000

25000? Not 25000.01?

But if H0 is false test will always reject null as long as sample size is large enough

Not true in our fields: H0: Higgs does not exist

Statistical significance ≠ practical significance

Say a new medication decreases the time until cure from 100 days to 99 days on average. If the study is huge this is stat. sign., but does it really matter?

Again, not really a problem for us (?)

$\alpha=0.05$ is sacrosanct (because Fisher said so)

No consideration of consequences of type I and type I errors.

Definitely an issue for us: $\alpha \sim 5\sigma$

Proposed solution? Don’t test but find interval estimates.

Sounds silly to Statisticians because the two are the “same” anyway

NHST has been around for a long time: Arbuthnot (1710)
H0: God does not exist

Its likely going to be around for a while longer

For a discussion of these issues see paper by David Krantz:
http://www.unt.edu/rss/class/mike/5030/articles/krantznhst.

Thanks!

Supplemental Material

How common are these misunderstandings?

Suppose you have a treatment that you suspect may alter performance on a certain task. You compare the means of your control and experimental groups (say, 20 subjects in each sample). Furthermore, suppose you use a simple independent means *t*-test and your result is significant ($t = 2.7$, $df = 18$, $p = .01$). Please mark each of the statements below as "true" or "false." *False* means that the statement does not follow logically from the above premises. Also note that several or none of the statements may be correct.

(1) You have absolutely disproved the null hypothesis (i.e., there is no difference between the population means).	☐ True	False ☐
(2) You have found the probability of the null hypothesis being true.	☐ True	False ☐
(3) You have absolutely proved your experimental hypothesis (that there is a difference between the population means).	☐ True	False ☐
(4) You can deduce the probability of the experimental hypothesis being true.	☐ True	False ☐
(5) You know, if you decide to reject the null hypothesis, the probability that you are making the wrong decision.	☐ True	False ☐
(6) You have a reliable experimental finding in the sense that if, hypothetically, the experiment were repeated a great number of times, you would obtain a significant result on 99% of occasions.	☐ True	False ☐

Percentages of False Answers (i.e., Statements Marked as True) in the Three Groups of Figure 1

Statement (abbreviated)	*Germany 2000*			*United Kingdom 1986*
	Psychology students	*Professors and lecturers: not teaching statistics*	*Professors and lecturers: teaching statistics*	*Professors and lecturers*
1. H_0 is absolutely disproved	34	15	10	1
2. Probability of H_0 is found	32	26	17	36
3. H_1 is absolutely proved	20	13	10	6
4. Probability of H_1 is found	59	33	33	66
5. Probability of wrong decision	68	67	73	86
6. Probability of replication	41	49	37	60

Note. For comparison, the results of Oakes' (1986) study with academic psychologists in the United Kingdom are shown in the right column.

Talks given at Conference

J. M. Bernardo (U. Valencia) Keynote address:

Hypothesis Testing from a Decision Theory Viewpoint:

A General Objective Bayesian Approach

Art De Vos and Marc Francke (Free University Amsterdam)

No More Null Hypotheses, Just Decisions

Mike Evans and Zeynep Baskurt (University of Toronto)

Hypothesis Assessment via Bayes Factors and Relative Belief Ratios

Cecilia Nardini (University of Milan & SEMM & IEO)

Can Likelihood-based Tests Be Reliable in Sequential Clinical Trials?

Trotta, A. Jaffe, D. Mortlock and D.Van Dyke (I.C. London)

Model Criticism and Model Selection in Cosmology

Valeriano Iranzo (U. Valencia)

Some Remarks on Bayesian Measures of Explanatory Power

Deborah Mayo (Virginia Tech)

K.Brewer, G.Hayes and A. Gillison (Australian National University)

Using Fisher's p to Measure Significance

Kevin Hoover (Duke) Keynote address:

The Role of Hypothesis Testing in the Molding of Econometric Models

Nicholas Longford (SNTL and Universitat Pompeu Fabra)

Statistics Without Hypothesis Testing

Ian Hunt

We posed the question with the six multiple-choice answers to 44 students of psychology, 39 lecturers and professors of psychology, and 30 statistics teachers, who included professors of psychology, lecturers, and teaching assistants. All students had successfully passed one or more statistics courses in which significance testing was taught. Furthermore, each of the teachers confirmed that he or she taught null hypothesis testing. To get a quasi-representative sample, we drew the participants from six German universities (Haller & Krauss, 2002).

Percentages of False Answers (i.e., Statements Marked as True) in the Three Groups of Figure 1

	Germany 2000			*United Kingdom 1986*
Statement (abbreviated)	*Psychology students*	*Professors and lecturers: not teaching statistics*	*Professors and lecturers: teaching statistics*	*Professors and lecturers*
1. H_0 is absolutely disproved	34	15	10	1
2. Probability of H_0 is found	32	26	17	36
3. H_1 is absolutely proved	20	13	10	6
4. Probability of H_1 is found	59	33	33	66
5. Probability of wrong decision	68	67	73	86
6. Probability of replication	41	49	37	60

Note. For comparison, the results of Oakes' (1986) study with academic psychologists in the United Kingdom are shown in the right column.