The following slides were presented at the 83rd Annual Meeting of the Transportation Research Board, Washington, DC, January 2004, in a session titled "How useful is it to test hypotheses in highway safety?". The other session speakers were Gary A. Davis (U. Minnesota), Ezra Hauer (U. Toronto) and Simon P. Washington (U. Arizona), and the presiding officer was Michael S. Griffith (Federal Highway Administration).

Admittedly, the slide content is abbreviated and at a later date, I plan to add explanatory notes. For now, however, I hope the slides alone will be sufficient illustration for the point I tried to make in my presentation (i.e., that claims of "no difference" should be supported at $p<.05$ just as claims of "difference" are).

Slide 1

Significance Testing Cuts Both Ways (or should):

Claims of "No Real Difference" Should be Supported at p<.05

Richard A. Harshman, Psychology Dept.
University of Western Ontario, Canada

TRB Annual Meeting, January 2004, Wash. DC

Slide 2

Hauer is right: the error *is* widespread.
Here's an example from Psychology:

Article:

"Women's advantage on verbal memory is not restricted to concrete words"

in *Perceptual and Motor Skills*, 2003

Slide 3

Typical ambiguous wording:

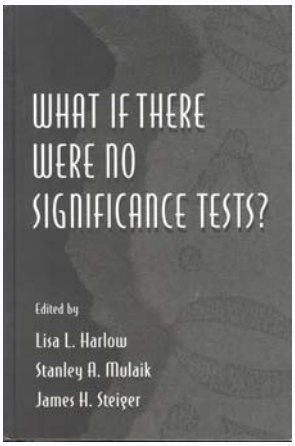
"There was no significant interaction of sex with test … indicating that women were equally advantaged on all 3 tests."

"There was no sex x word-type interaction… indicating that women were equally advantaged on the two kinds of words"

Slide 4

Concern Spawned Radical Proposal

- Some urged **abolishing** sig tests (eg, Schmidt & Hunter)
- A heated controversy ensued & book published
- American Psych. Assoc. set up methodology committee
- Conclusion: didn't abolish sig tests, but urged reporting confidence intervals, etc.



Why keep significance testing?

- "Yes-No" choices must sometimes be made, e.g.
  whether to outlaw right turns on red light
  whether to accept article for publication
- Sig tests with p-value conventions provide objective, agreed, yardstick for evidence
- Size of p-values (.03 vs .002) further informs us about *strength of evidence against null*

Slide 6

Also, *proving* “no difference” has become more important in recent years:

- FDA regulations require *proof* of equivalent “bioavailability” of generic drug.
- Substituting a different therapy requires first proving it to have 1-sided equivalence (be as good as the expensive one).
- Structural models, goodness of fit tests, etc

Proposed alternative to abolition:

We can keep sig testing as one of several useful tools (along with confidence intervals, etc.), if we require any "proof of no effect" to be *valid:*
i.e., demand the same support for a conclusion of NO DIFFERENCE as for one of a DIFFERENCE: i.e. *evidence* signif at p<.05 (or .01)

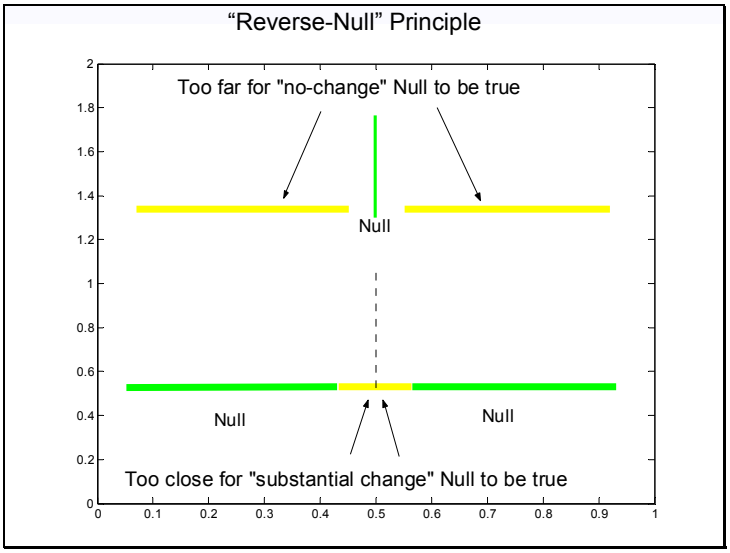
Slide 8

How to do it: The Reverse-null Principle:

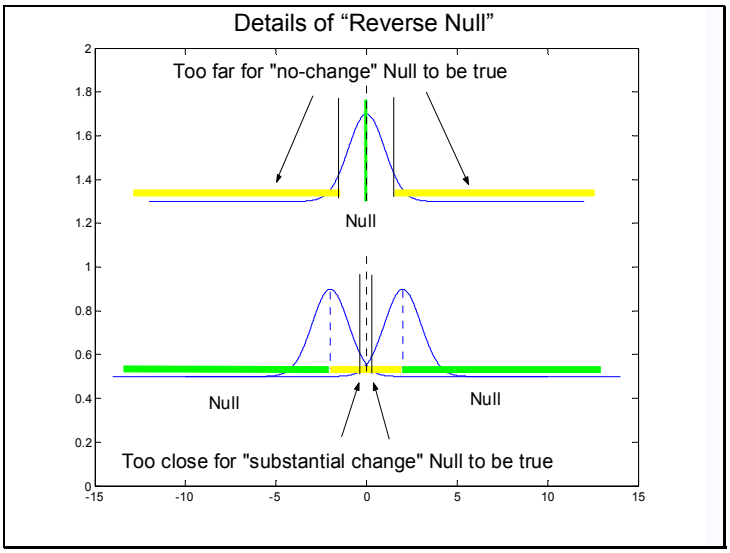
To conclude there is a substantial difference you must observe a diff big enough to reject null that it's due to chance.

To conclude there's no substantial difference you must observe a diff small enough to reject null that closeness is due to chance (i.e., a chance deviation from a bigger 'true' population diff.)

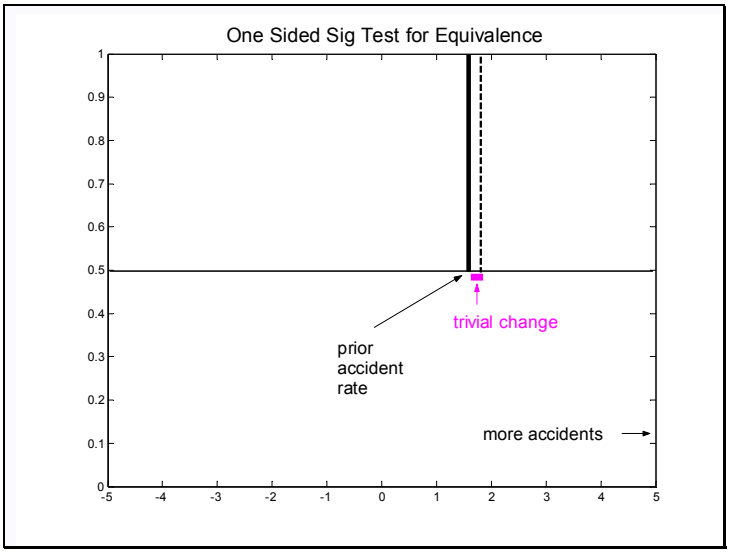
Slide 9


“Reverse-Null” Principle
Too far for "no-change" Null to be true
Null
Null
Null
Too close for "substantial change" Null to be true
2
1.8
1.6
1.4
1.2
1
0.8
0.6
0.4
0.2
0
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

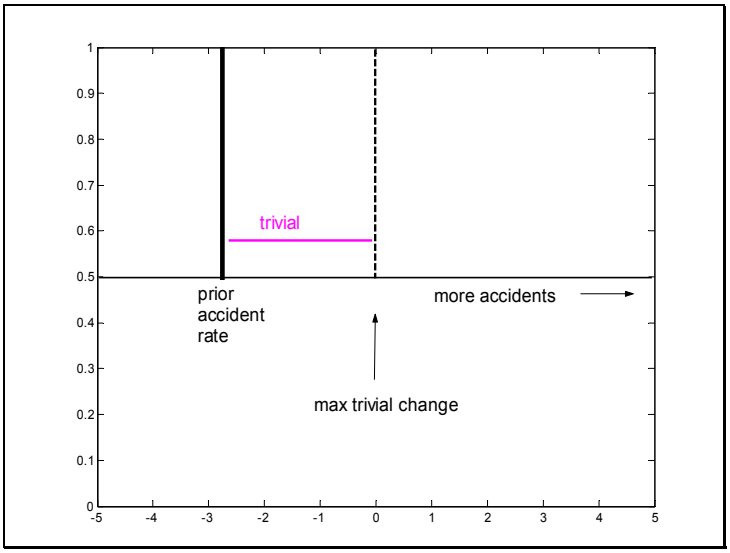
Slide 10


Details of “Reverse Null”
Too far for "no-change" Null to be true
Null
Null
Null
Too close for "substantial change" Null to be true
2
1.8
1.6
1.4
1.2
1
0.8
0.6
0.4
0.2
0
-15
-10
-5
0
5
10
15

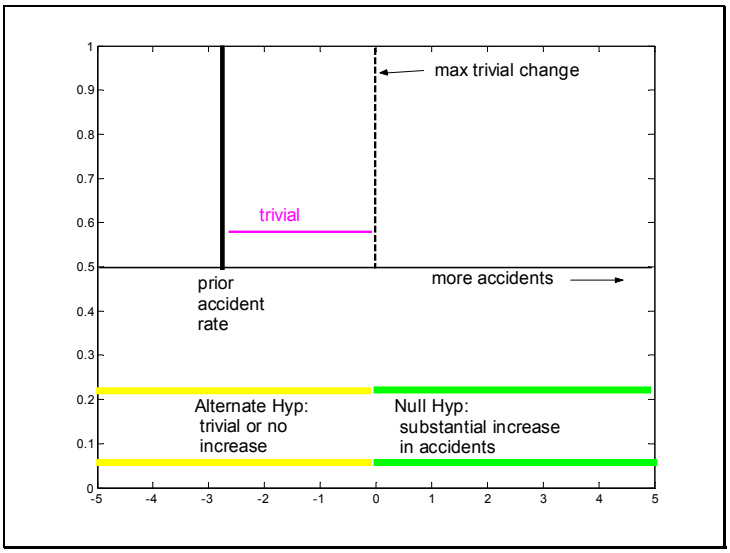
Slide 11


One Sided Sig Test for Equivalence
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
-5
-4
-3
-2
-1
0
1
2
3
4
5
trivial change
prior
accident
rate
more accidents

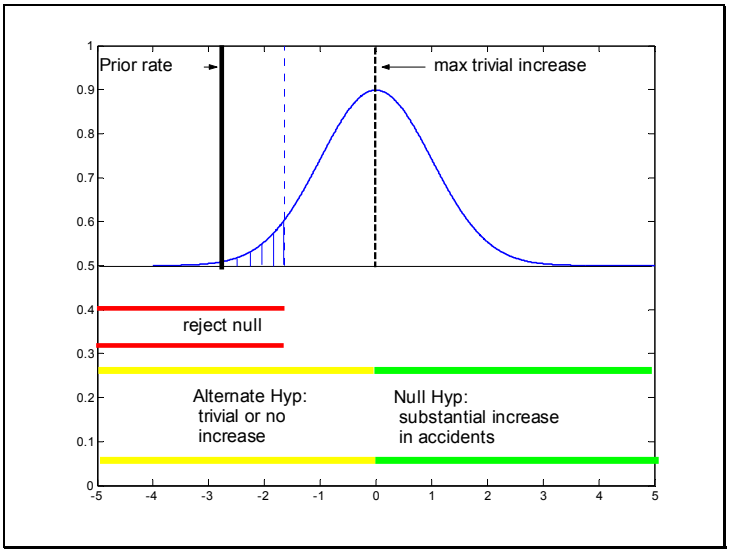
Slide 12


1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
-5
-4
-3
-2
-1
0
1
2
3
4
5
trivial
prior
accident
rate
more accidents
max trivial change

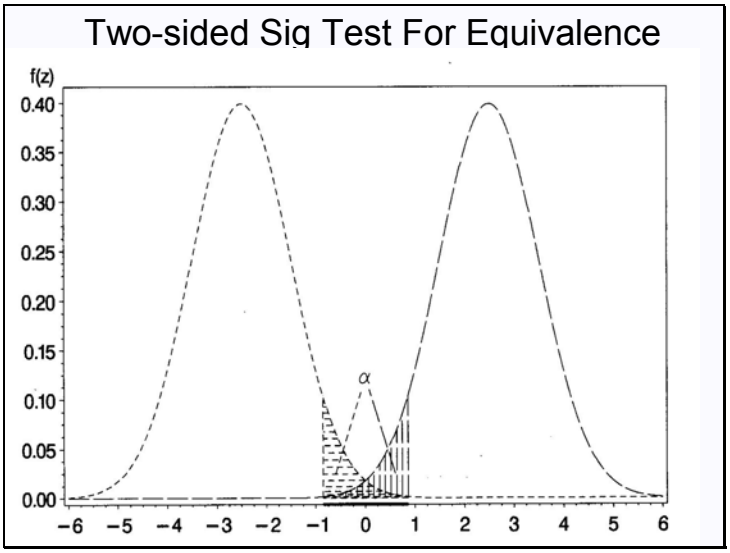
Slide 13


max trivial change
trivial
prior
accident
rate
more accidents
Alternate Hyp:
trivial or no
increase
Null Hyp:
substantial increase
in accidents
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
-5
-4
-3
-2
-1
0
1
2
3
4
5

Slide 14

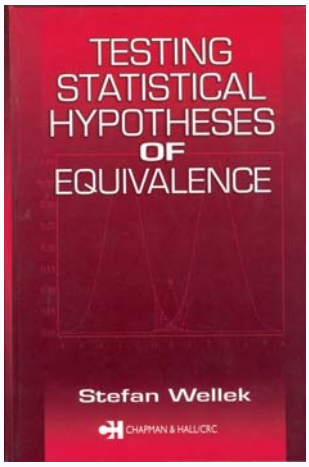

Prior rate
max trivial increase
reject null
Alternate Hyp:
trivial or no
increase
Null Hyp:
substantial increase
in accidents
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
-5
-4
-3
-2
-1
0
1
2
3
4
5

Slide 15


Two-sided Sig Test For Equivalence
f(z)
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
α
−6 −5 −4 −3 −2 −1 0 1 2 3 4 5 6

Slide 16

Resource:



Gives both theory and use of software that is downloadable from the internet.

Slide 17

Equivalence Testing Software on Internet:

- Exact Fisher Test: power, sample sizes, p
- Bayesian tests
- Hazard rate tests
- One and Two sample t-tests
- Signed Rank, Mann-Whitney, etc.
- Equivalence of Variances
- True test of 'Goodness of Fit' of observed to specified multinomial distributions
- Etc.

Caveats

- Larger samples will be required, hence more work to conduct a study
- As with other rules, must discourage automatic unthoughtful use
- Valid sig testing must be used *in conjunction with* other important statistical methods, particularly reporting of confidence intervals, awareness of power, effect sizes, etc.

Slide 19

Advantages of new approach

Transparently Reasonable:

- All conclusions *should* be based on hard evidence

Good chance of success:

- Simple rule to understand and follow
- Hard to argue with (new application of familiar requirement)
- Already shown to work for FDA regulators
- May engender less psychological resistance--people will feel virtuous when reporting "p<.05"

"It's the right thing to do"

- Improves scientific integrity, prevents "puffery"
- Reduces practical harm

Slide 20