

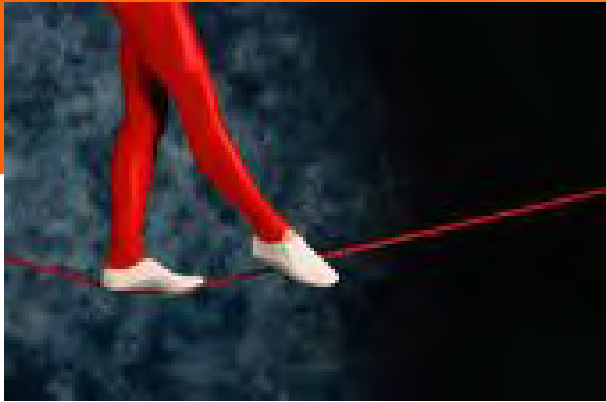
Annex 8

The contrast between....

Statistical significance

and

Biological significance



A significance test used in biology is like a safety net for a tightrope walker

It provides confidence (that unwarranted inferences are not being drawn)



.... but it should *never* be part of the main act.
(Perry, 1986)

Working Group Statistics of GMO Panel

- Panel members:
 - Hans Christer Andersen
 - Salvatore Arpaia
 - Harry Kuiper (formal chair)
 - Joe Perry
 - Willem Seinen
- Ad hoc experts:
 - Marco Acutis
 - Andrew Cockburn
 - Ludwig Hothorn
 - Gijs Kleter
 - John Law
 - Jim McNicol
 - Hilko van der Voet (acting chair)
- EFSA staff
 - Claudia Paoletti

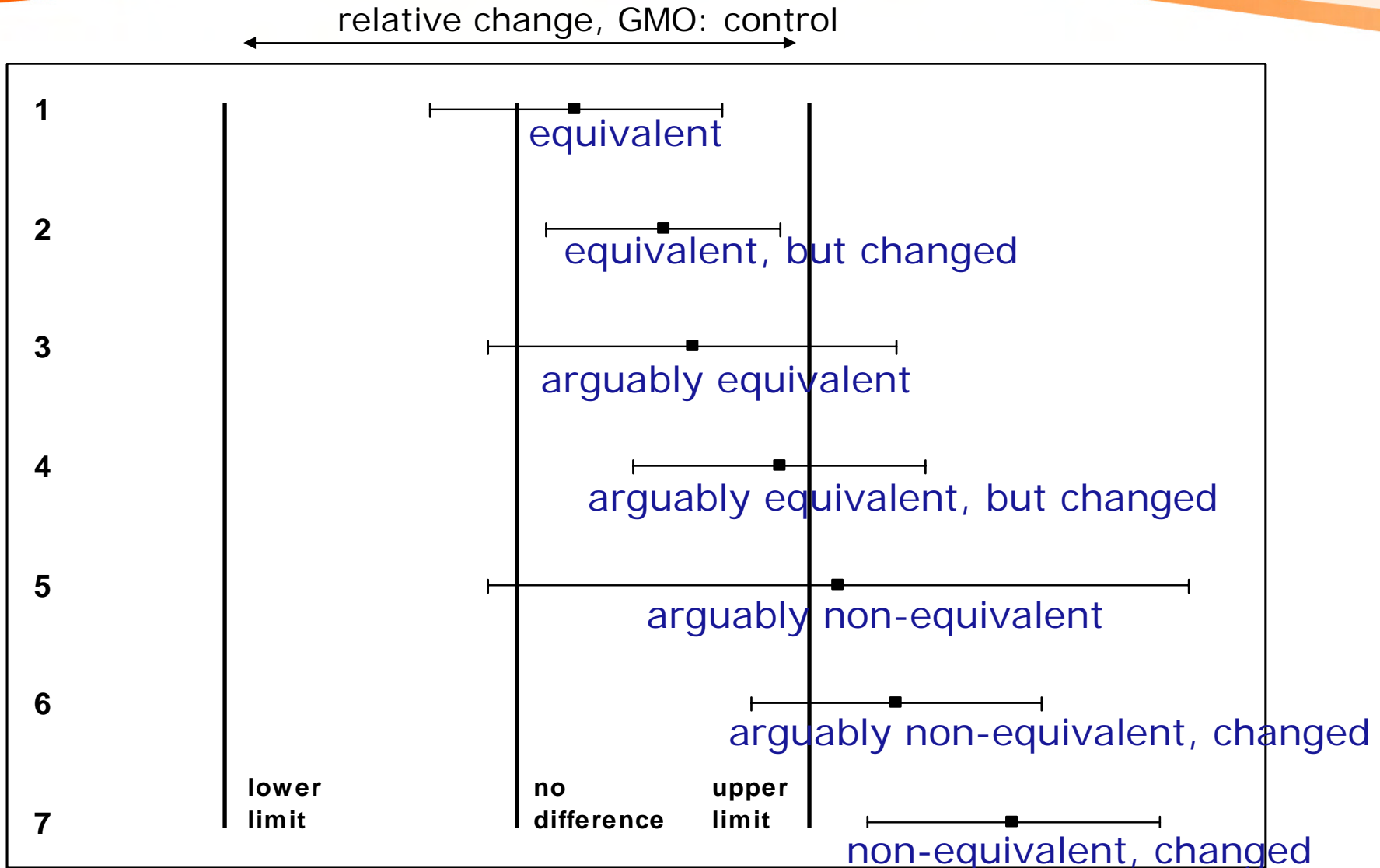
Two types of testing

Test of difference (proof of hazard)		verdict	
		not different	different
truth	H ₀ : the same	OK	Type I error (false positive)
	H ₁ : differ	Type II error (false negative)	OK

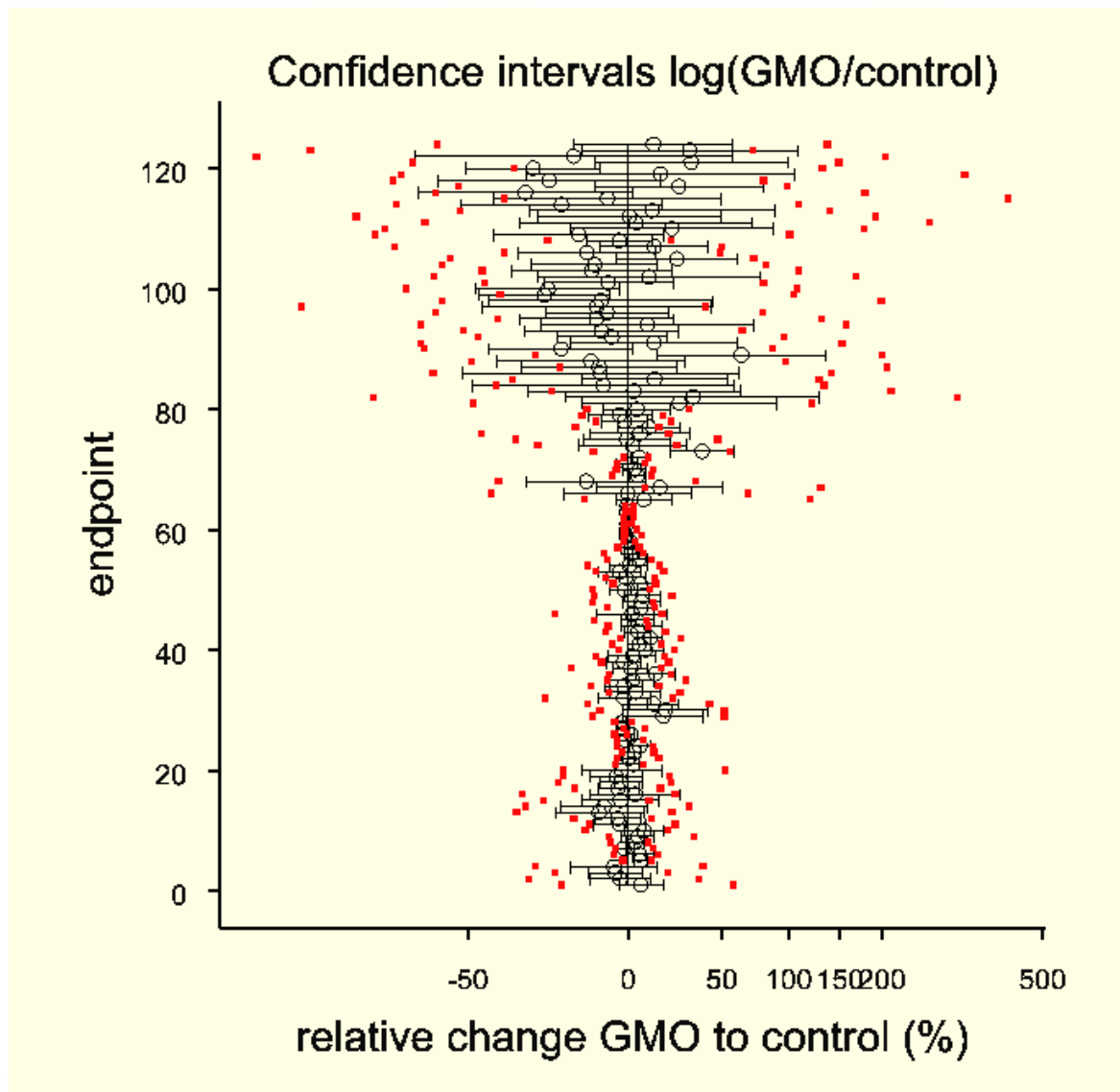
Test of equivalence (proof of safety)		verdict	
		not equivalent	equivalent
truth	H ₀ : non-equivalent	OK	Type I error (false negative)
	H ₁ : equivalent	Type II error (false positive)	OK

- Type I error is set at specified value (e.g. $\alpha = 0.05$)
- More difficult to control Type II error (β)
- Power of the test (= $1 - \beta$) should be high enough
- Note: Consumer risk is
 - Type I error in test of equivalence
 - Type II error in test of difference

Types of possible outcomes



Example, estimated limits of concern



Further work by Statistics Working Group on (i) theory and (ii) data analysis will help elucidate remaining issues of experimental design:

- Experimental design for field trials and animal trials differs
- Many field experiments require replication over years and locations (because there may be site x treatment interactions and year x treatment interactions)
- Power analysis should be done where feasible
- Replication should be large enough to have sufficient degrees of freedom (> 15) for within-site statistical analysis
- Guidelines are not the same as recipes



For a randomized block trial with b blocks and t treatments we have $(b-1) * (t-1)$ degrees of freedom for residual

Say we have $t = 4$ treatments (GM, non-isogenic-control, reference variety 1, reference variety 2) we need $b \geq 6$ (a minimum of six blocks) to obtain 15 residual d.f. .

For $t = 3$ we require a minimum of eight blocks

For $t = 5$ we require a minimum of five blocks, etc.

Further quotes from Perry (1986)

Perry, J.N. (1986) Multiple-comparison procedures : a dissenting view.

Journal of Economic Entomology, **79**, 1149-1155.

The paper has received about 50 citations

Estimation is more important than hypothesis testing

Fisher's early writings placed undue emphasis on significance testing that has persisted until the present day.

The primary interest in agriculture and ecology is to estimate the magnitude of treatment effects.

Typical relevant questions are: "by how much does fecundity increase for each unit increase in food intake?"
rather than: "is there a significant increase in fecundity when nutrition is improved?"

Significance tests have a limited role in biological experiments for at least seven reasons:



Alternative approaches

The role for testing in the analysis of data occurs at the stage of exploring competing models. Concentration should then focus on estimating the parameters of this model, and hence on the magnitude of the biological effects under investigation.

Information regarding the standard errors of the means is of central importance and should always be displayed ... A standard error is much more informative than a significance test.

Significance tests have a limited role in biological experiments, because:

- 1) significance refers merely to plausibility, not to biological importance; significant treatment differences may exist but be biologically trivial
- 2) theories may be proved to be strictly untrue but still of practical use (e.g. Newtonian mechanics of Einsteinian relativity)
- 3) a null hypothesis is often known to be false before experimentation; the resultant significance level, based on an incorrect assumption, has no quantitative meaning, and testing the hypothesis is redundant
- 4) the outcome of a test often depends merely on the size of the experiment (e.g. number of replicates); the more replicates the greater the chance of achieving significance
- 5) in agriculture, ecology and entomology, the really critical, single experiment is rare
- 6) results may indicate merely that a hypothesis is rejected, but not give the magnitude of departures from the hypothesis
- 7) the exact nature of tests is often exaggerated and ignores the fact that all tests are based on assumptions that rarely hold in practise.

Further quotes from Perry (1986) (continued)

Discussion

Most biologists have developed from experience an intuitive understanding of the appropriate size of experiment required to achieve its objectives.

[Too few] text books give sensible advice based on experience in analyzing data, compared with those which rely on 'cook-book' techniques, which are often misused by scientists who do not fully understand them.

While rigid instructions to authors are not desirable, statistical guidelines are valuable; we, therefore, consider each case on its individual merits.

How many sites – how many years?

Guidance document, 7.2(a):

“The comparison ... should cover more than one representative growing season and multiple geographical locations representative of the various environments in which the GM plants will be cultivated.”

This implies ≥ 2 sites and ≥ 2 years

Guidance document, 7.2(b):

A randomised complete block design, for example, could indicate whether the experimental factors (location, year, climatic conditions, plant variety) interact with one another.

This implies an ideal experimental design would include the *same* varieties at the *same* sites in *different* years

Guidance Document, 7.2(b):

Field trial data should be presented separately as well as pooled and should be analysed statistically.

This implies each site should be both analysed separately and pooled together in an overall analysis, and that

Therefore the replication *within* a site should be sufficient to allow a well-founded stand-alone analysis.

(Consensus: that it would not be feasible to try to capture the full range of natural variation across years, but that the number of sites should be as large as possible in each year.

In UK plant variety trials, ≥ 2 years are required for national listing of a new variety but up to 4 years for other purposes. The number of sites per year may exceed 5 in the first year, and at least 5 in subsequent years.)

“Use of statistical techniques such as power analysis shall be utilised prior to finalising experimental design” and “there should be more advice on the application of statistical power of NTO testing”

“The effects of scale and time on the environmental impact of a GMO would always be difficult to predict in environments in constant and unpredictable flux. Upscaling models as discussed above could become useful and so this issue should be under constant review and be studied by the EFSA Statistics Working Group.”

from *Overall Conclusions and Recommendations* of the EFSA Environment Working Group Colloquium at Tabiano, 2007

Statistical power: the probability of detecting a difference between GMO and control, when there is a real difference to detect - often quoted as a percentage.

A power analysis, **executed when the study is being planned and prior to its start**, may be used to estimate power, choose appropriate replication and give confidence that the experiment will detect any significant effect that is present.

A common approach to deal with Type II error, but one of dubious validity, is the **calculation of statistical power from the experimental data obtained** (so-called 'retrospective power analysis'). In this approach an applicant might seek to compensate for a possible lack of power in a relatively poorly replicated experiment by adjustment of the size of the experiment (the Type I error rate), which uniquely determines the retrospective power of the experiment. Exposure of the problems associated with such a strategy was identified by Hoenig and Heisley (2001). Tempelman (2004) pointed out how a poorly executed experiment would be rewarded a greater chance of concluding bioequivalence.

(The approach proposed by the Statistics WG, of specifying explicit limits of concern and then employing two types of hypothesis test, does not suffer from such problems. Notwithstanding the problems of retrospective power analyses, it remains valid to reassess studies for which a prospective power analysis was done, to check model assumptions and parameters estimated *a priori*.)

Implications for regulatory procedures of text in red:



Beware of what you wish for –
you may get it one day

Applicant pre-notifies
with power analysis

*"I am going to do a trial and plan it
to have the following power"*



Trial(s) conducted over two or three year period



Analysis of data –
development of application



Full application –
"How did I do in my trial?"

Farm Scale Evaluations of GMHT Crops

Three experiments

Spring-sown crops:

Beet, Spring oilseed rape (SOSR), Maize
all GMHT

1999 (pilot year for protocols)

1999 power study

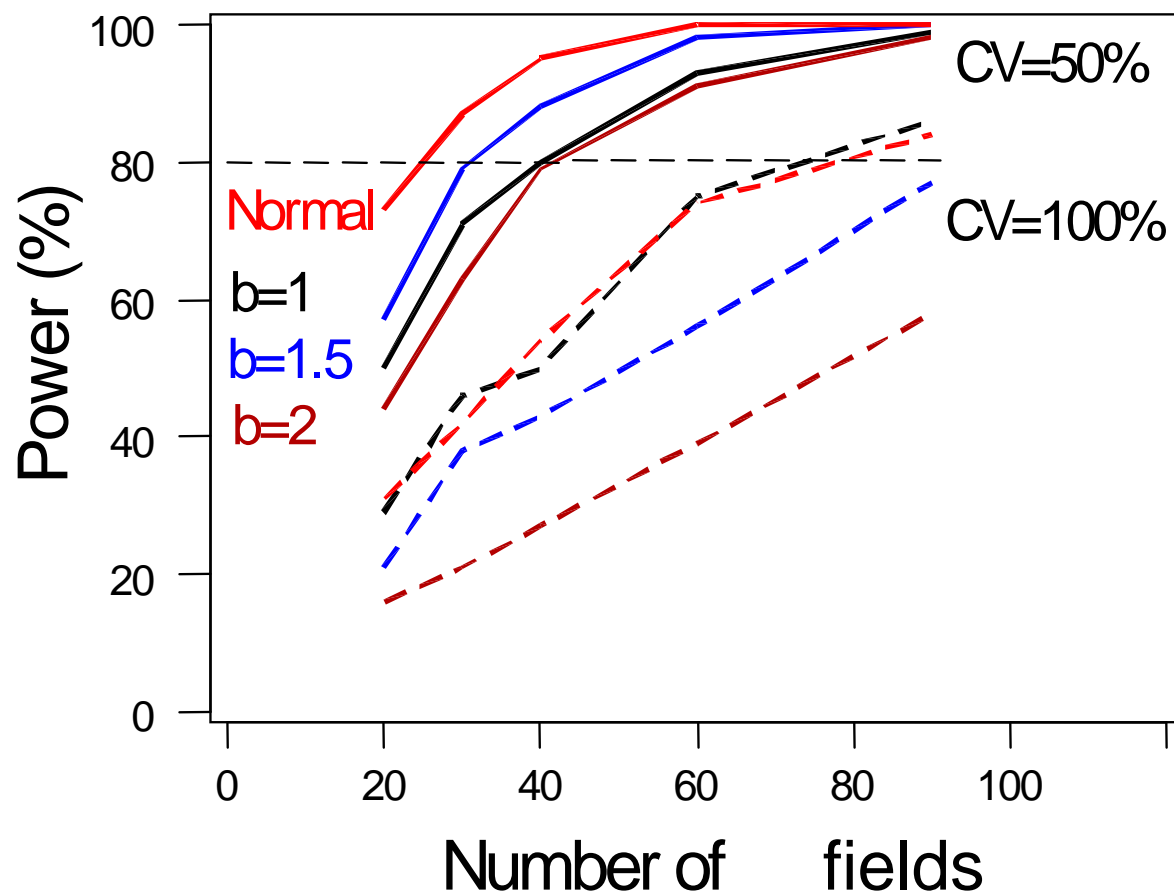
experiments begun in 2000

power study reported February 2003

Results reported October 2003

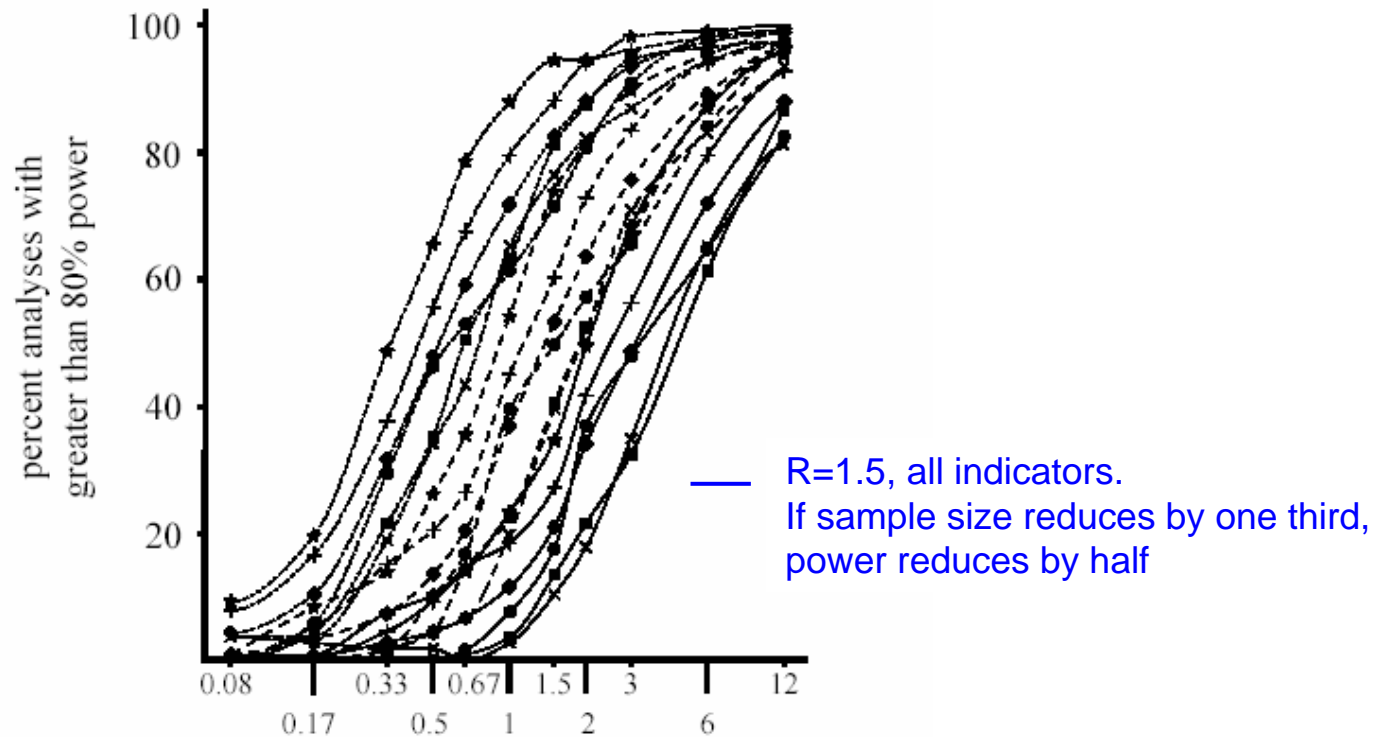
..... power estimated from 500 sets of simulated low-mean count data using a negative binomial distribution, with P values computed from paired Monte Carlo randomization test

and, again, it is vital that biologists must specify size of effect which they seek to detect



Retrospective studies between sites – e.g. effects of change of sample size

$$\text{Probit [Power]} = 0.955 + 0.670\theta - 0.526\theta \sqrt{M} - 1.46\beta + 0.182\theta\beta - 7.67 \sqrt{n} + 3.62\beta \sqrt{n}$$



Fraction of actual achieved sample size, in terms of field replication

Clark, S.J. Rothery, P. & Perry, J.N. (2005) Farm Scale Evaluations of spring-sown genetically modified herbicide-tolerant crops: a statistical assessment. *Proceedings of the Royal Society B.*, **273**, 237 – 243.

Retrospective studies within sites – e.g. effects of change of sampling intensity

Clark, S.J., Rothery, P., Perry, J.N. & Heard, M.S. (2007)

Farm Scale Evaluations of herbicide-tolerant crops: assessment of within-field variation and sampling methodology for arable weeds.

Weed Research, **47**, 157–163.

The amount of within-field sampling could have been reduced from 12 transects (60 quadrats) to 6 transects (30 transects) or even to 3 transects (15 quadrats), with little reduction in the frequency of detection of treatment effects or of power.

Qi, Perry, Pidgeon, Haylock & Brooks (2008)

Cost efficacy in measuring farmland biodiversity –
lessons from the Farm Scale Evaluations of GMHT crops

Annals of Applied Biology (submitted)

The recommended reduction from 12 to 3 transects would have saved £1,356 per site, only 6% of total budget. A minimalist approach using only the single season seedbank protocol would have cost only £3,437 (c. 5850 Euro) per site