

# **p-VALUES, STATISTICAL EVIDENCE AND THE SAFETY OF GENETICALLY MODIFIED FOODS**

W.H. Ross  
Cunye Qiao

Food Directorate  
Health Canada

Statistical Society of Ottawa  
November 21, 2008

## Example: Possible toxicity of GM maize

- August, 2002: Monsanto application to EU for import of GM maize (MON863), including a 90 day rat feeding study
- April, 2004: EFSA publishes favourable review of safety
  - Some statistically significant, but not biologically meaningful effects.
  - *Le Monde* reports on safety concerns raised by some EU organizations
- *June, 2005*: Greenpeace publishes feeding study
- January, 2006: EU Commission authorizes use of MON863 in food
  - *Greenpeace mounts an internet campaign*
- *January 2007*: G.-E. Sérelini et. al. publish “re-analysis” of feeding study
  - *Critical of original results, additional statistical analysis*
- *June 2007*: EFSA review of statistical analyses. (146 pages)

## GM FOOD AND SUBSTANTIAL EQUIVALENCE

- Testing must assure that *“foods produced from these new products are **as safe as** food produced from conventionally bred crops. There must be reasonable certainty that no harm will result from intended uses under the anticipated conditions of consumption”*.
  - WHO, European Food Safety Authority (EFSA), OECD, UNFAO, Codex Alimentarius
- What is the appropriate type of statistical evidence?
  - “Show me” (hypothesis testing)
    - Strict test for evidential support
  - Relative weight
  - Bayes

## A SHORT DETOUR ABOUT NULL HYPOTHESES

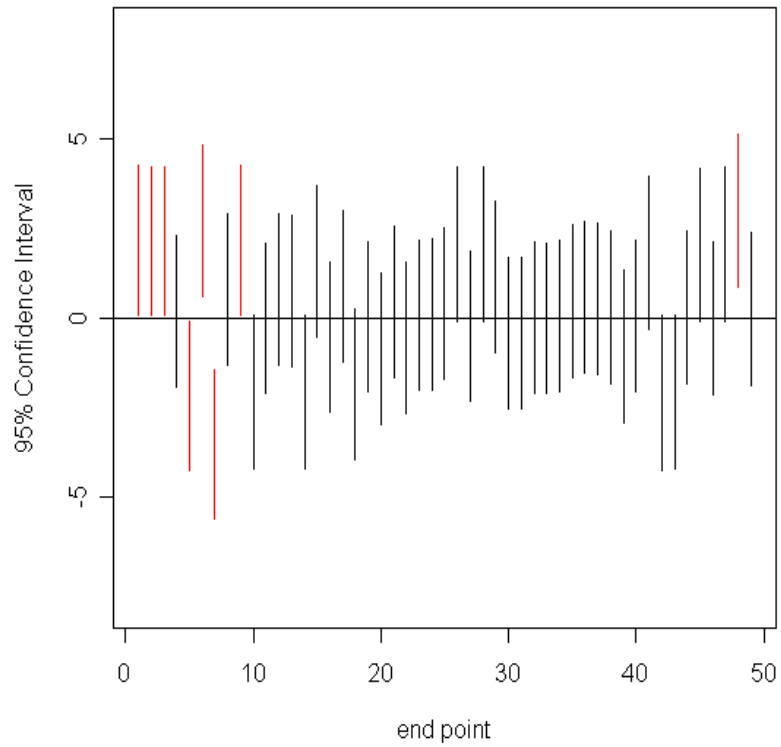
- We only need to think about two types
  - “Effect” is a quantitative proxy for an important outcome
  
- 1. There is a small effect
  - Prefers affirming no large effect when there is one (type I error)
    - “No effect” is the most common and most stringent
  - Scientific ‘integrity’, efficacy, ...
  - Prefers affirming no effect when there is one
  
- 2. There is a large effect versus there is a small effect
  - Incorporates acceptable variability in effects
    - Equivalence testing
  - Prefers affirming a large effect when there isn’t one (type I error)
    - Safety testing (precaution)

# MONSANTO RAT FEEDING STUDY RESULTS (VIA EFSA)

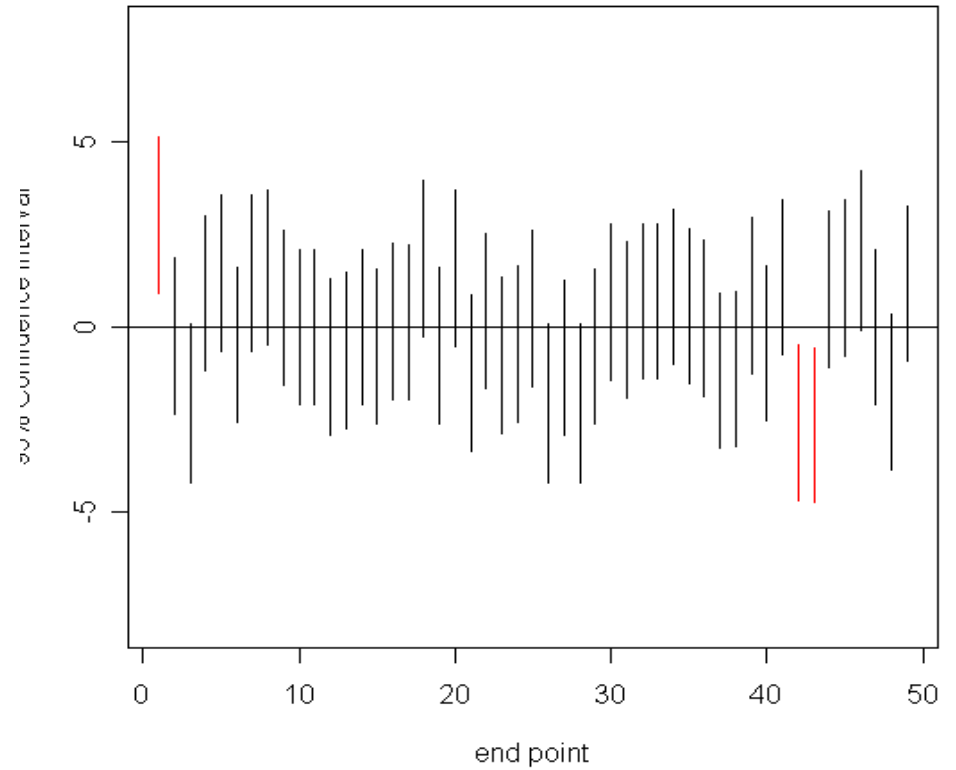
- Four factors
  - Gender (M & F)
  - Genotype (MON863, non-transgenic reference)
  - Dose (11% or 33% level maize in diet – 11% supplemented with 22% non-transgenic maize)
  - Time (total of 14 weeks)
- Measured Hematology, clinical-chemistry and urinalysis parameters, histopathology and organ weights (10/group), and body weight over time(20/gp).
  - 479 endpoints (excluding body weight)
  - EFSA report stresses difference between statistical significance and biological significance
    - General agreement on the distinction in literature and debate.



**95% Confidence interval-Females 11% Week 14**



**95% Confidence interval-Females 33% Week 14**



## AVERAGE EQUIVALENCE: STATISTICAL MODEL

$$X_i \sim N(\mu_T, \sigma^2) \quad i = 1, \dots, n: \text{Test sample}$$

$$Y_j \sim N(\mu_R, \sigma^2) \quad j = 1, \dots, n: \text{Reference sample}$$

### Likelihood Parameterization

$$\theta = \frac{n\mu_T + m\mu_R}{m+n} \quad \psi = \frac{\sqrt{mn}(\mu_T - \mu_R)}{m+n}$$

$$l = -\frac{m+n}{2} \left\{ \ln(\sigma^2) + \frac{\hat{\sigma}^2 + (\theta - \hat{\theta})^2 + (\psi - \hat{\psi})^2}{\sigma^2} \right\}$$

# HYPOTHESIS TESTING: AVERAGE EQUIVALENCE

## Current Test

$$H_0 : \frac{|\psi|}{\sigma} = 0 \quad \text{vs} \quad H_1 : \frac{|\psi|}{\sigma} \neq 0$$

## Critical Regions

$$\{T : T^2 > f_{1, m+n-2, 1-\alpha_1}(0)\}$$

## Equivalence Test

$$\frac{|\mu_T - \mu_R|}{\sigma} \leq \varepsilon \quad \text{Defines equivalence region}$$

$$\{T : T^2 \leq f_{1, m+n-2, \alpha_2}(\delta^2)\}$$

$$H_{\varepsilon 0} : \frac{|\psi|}{\sigma} \geq \frac{\sqrt{mn}}{m+n} \varepsilon \quad \text{vs} \quad H_{\varepsilon 1} : \frac{|\psi|}{\sigma} \leq \frac{\sqrt{mn}}{m+n} \varepsilon$$

$$\delta^2 = \frac{mn\varepsilon^2}{m+n}$$

# USING THE WRONG TEST

Two tests agree when

$$f_{1,m+n-2,1-\alpha_1}(0) = f_{1,m+n-2,\alpha_2}(\delta^2)$$

Two cases

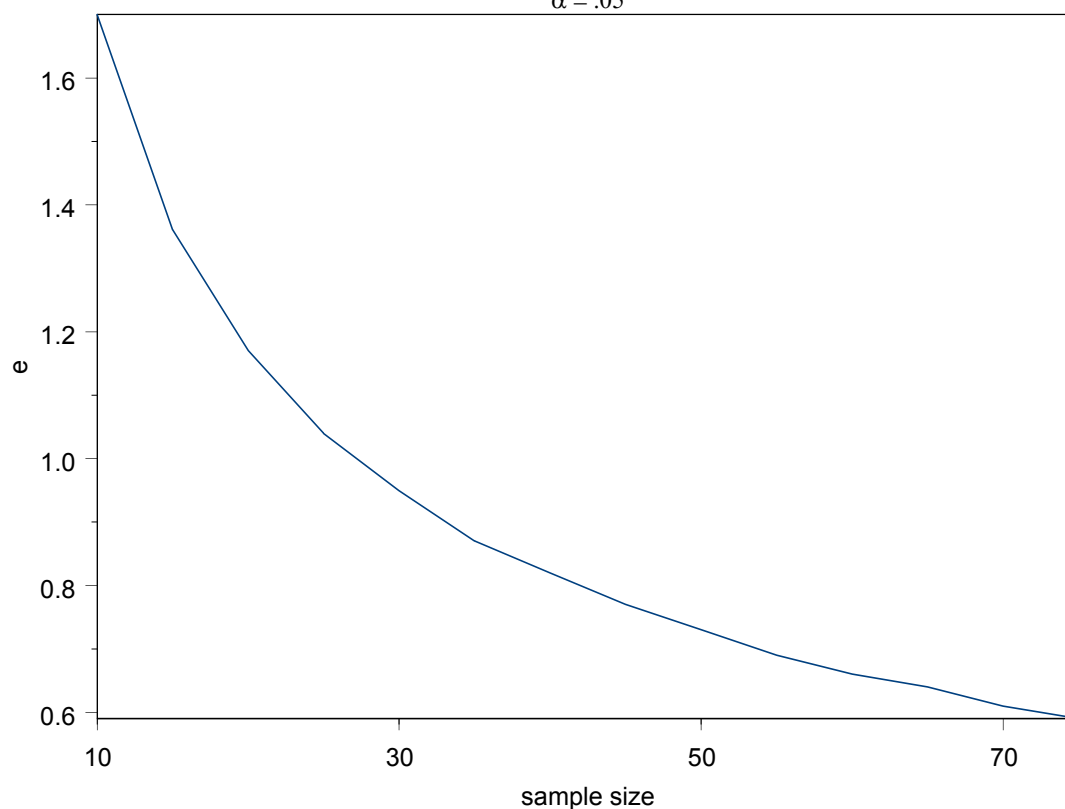
$$\alpha = \alpha_1 = \alpha_2$$

$$\alpha_1 = \alpha_1(\delta^2)$$

# ADJUSTING THE EQUIVALENCE CRITERION

Equivalence Criterion for Equality of Tests

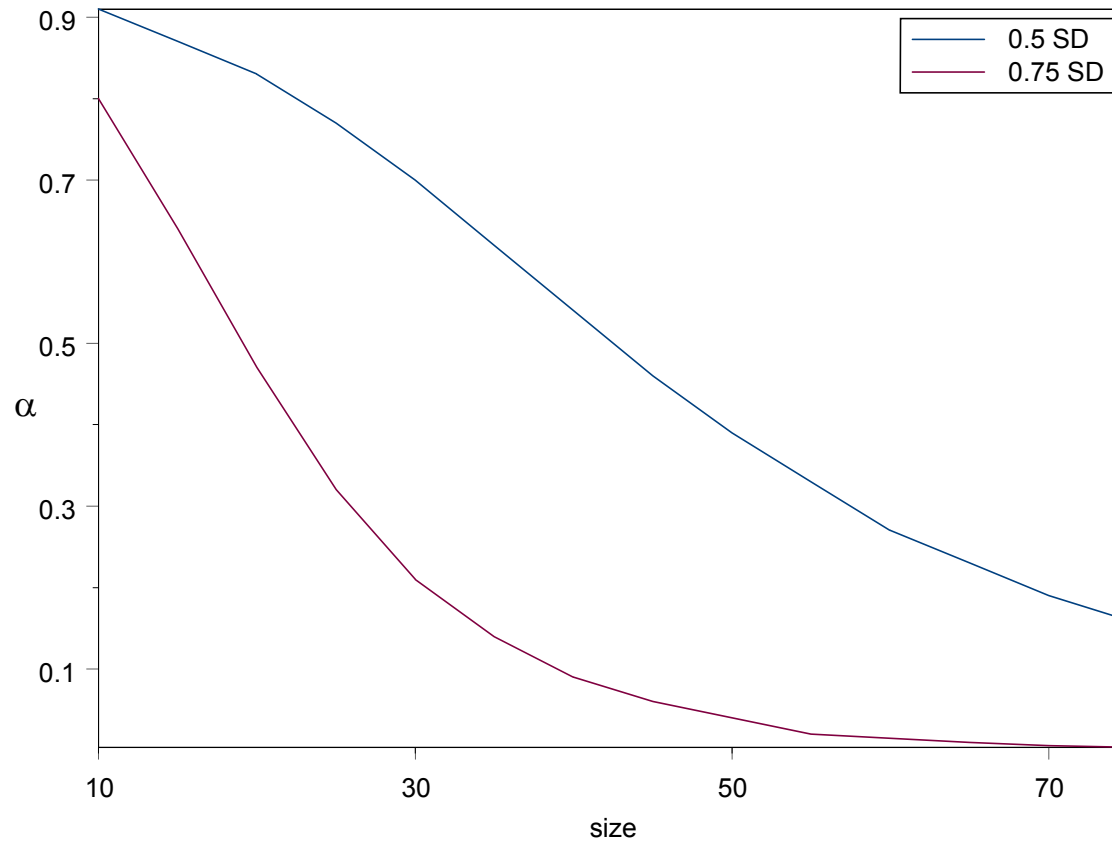
$\alpha = .05$





# ADJUSTING THE SIGNIFICANCE OF A TEST

Adjusting the significance of the wrong test



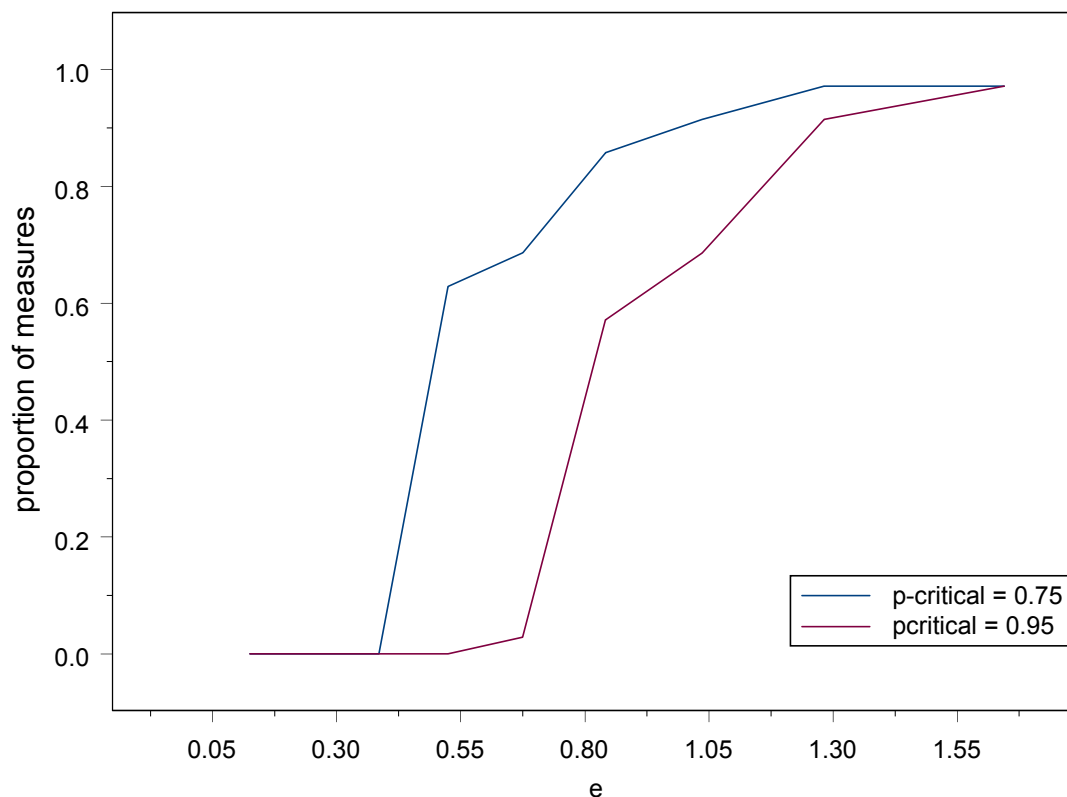
## Msc: FALSE DISCOVERY etc.

- Null hypothesis is composite
  - Not a pure significance test
    - One sided with small values rejecting
- Using boundary of equivalence
  - No acceptable hypotheses (i.e. rejected) unless boundary exceeds 1 SD.
- Within subject correlation masked by lack of power.
  - EFSA studies showed Expected Error Rate about right for the wrong hypothesis.

# BAYESIAN EVIDENCE

Prior on  $\psi$  symmetric, with  $\Pr\{\text{equivalence}\} = 1/2$

Posterior Probability of Equivalence



# FINAL COMMENT

- Published critiques (EFSA, Greenpeace, Peer Reviewed articles)
  - Used wrong null hypothesis (efficacy)
  - Most measures showed “no statistically significant effect”
    - “effect” is a shift in the measure average
- Safety Test
  - Is the correct null hypothesis if p-values are the correct measure of evidence
  - Most measures showed:
    - “no statistically significant effect” unless equivalence criterion large
    - “effect” is substantial equivalence
    - Large equivalence criterion suggests that hypothesis testing is not appropriate as a measure of evidence.