Overarching Issues in Pharmaceutical Risk

Bob Obenchain

Commercial Information Sciences

Outcomes Research, US Medical Division

Lilly

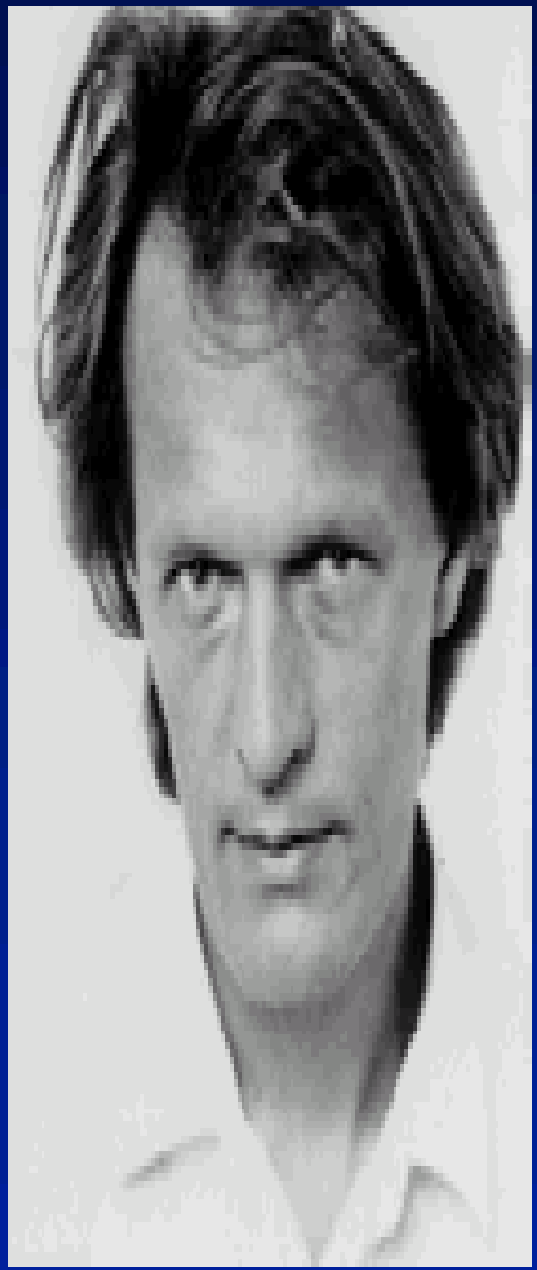
Answers That Matter.

Overarching Issues

…Head-to-Head Comparison of 2 Treatments

1. **Analysis of data from non-randomized human studies**
2. **Modern terminology for even-handed testing of composite hypotheses**
3. **Postponing decisions when cost-effectiveness data are insufficient**

…Statistical methods can become more-and-more realistic as data accumulate



"Investing in outcomes research is worse than just destroying the money because it gives the illusion of information."

- Richard Peto, *Science*, February 1994.

Cochran WG, Cox G.
Experimental Designs 1957

- **Blocking**
- Randomization
- **Replication**

Forms of Local Control for Human Studies

- Epidemiology (case-control & cohort) studies
- Post-stratification and re-weighting in surveys
- Stratified, dynamic randomization to improve balance on predictors of outcome
- Matching using estimated propensity scores
- Econometric simultaneous equations (IV)
- Marginal structural models (IPTW)
- Unsupervised propensity scoring: treatment within cluster (nested) models

Local Control:
A Subgrouping / Sensitivity approach to Robustness

- **Replace covariate adjustment based upon a global model with inference based upon local clustering (sub-grouping) of patients in X-covariate space.**
- **Explore sensitivity by increasing the number of clusters, intentionally over-shooting, then recombining.**
- **Also vary distance metric and clustering method while employing computationally intensive algorithms and interactive graphical displays.**

Notation

y = observed outcome variable(s)

t = observed treatment assignment (usually non-random)

x = observed baseline covariate(s)

z = unobserved explanatory variable(s)

The Propensity Score
of a Patient with Baseline Characteristic Vector *x*

$$PS = Pr(t \mid x)$$

is a vector of conditional probabilities that sum to 1.

The length of the PS vector is the total (finite) number of different treatments.

Propensity Scores for only 2 Treatments:

t = 1 (new) or 0 (standard)

p = Propensity for New Treatment
= Pr(t = 1 | X)
= E(t | X) = a scalar valued function of X

X = vector of baseline covariate values for patient

PS = (p, 1–p)

Fundamental PS Theorem

Joint distribution of ***x*** and ***t*** given **p**:

$$\Pr(\boldsymbol{x}, \boldsymbol{t} \mid \mathbf{p}) = \Pr(\boldsymbol{x} \mid \mathbf{p}) \Pr(\boldsymbol{t} \mid \boldsymbol{x}, \mathbf{p})$$
$$= \Pr(\boldsymbol{x} \mid \mathbf{p}) \mathbf{Pr(\boldsymbol{t} \mid \boldsymbol{x})}$$
$$= \Pr(\boldsymbol{x} \mid \mathbf{p}) \text{ times } \mathbf{p} \text{ or } \mathbf{(1–p)}$$
$$= \Pr(\boldsymbol{x} \mid \mathbf{p}) \Pr(\boldsymbol{t} \mid \mathbf{p})$$

...i.e ***x*** and ***t*** are **conditionally independent** given the propensity for new, $\mathbf{p} = \Pr(\boldsymbol{t} = 1 \mid \boldsymbol{x})$.

The unknown true propensity score (in non-randomized human studies) is the "most coarse" possible balancing score.

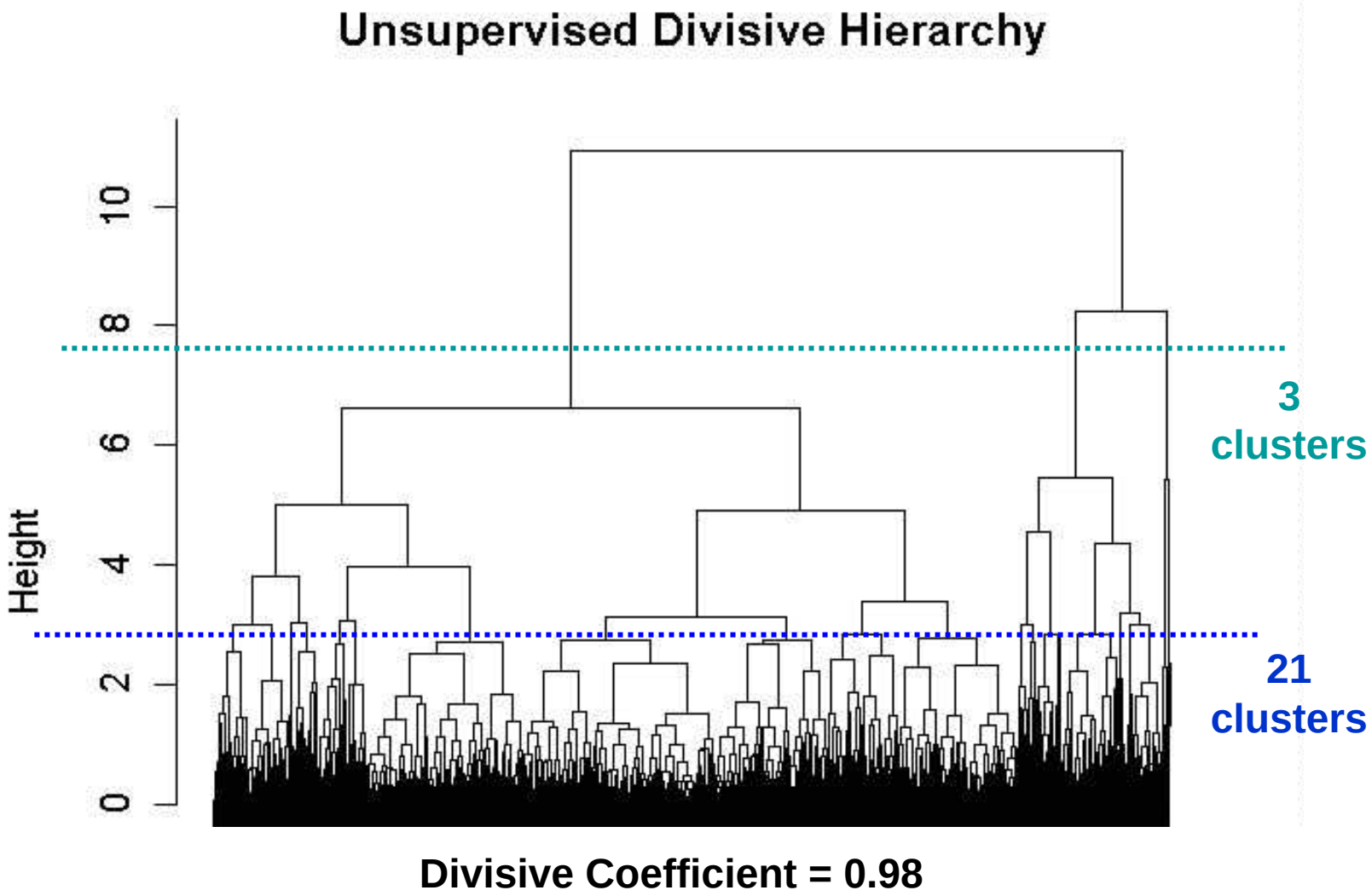
The known X-vector itself is the "most detailed" balancing score…

$$\Pr(x, t) = \Pr(x) \Pr(t \mid x)$$

Conditioning upon Cluster Membership is intuitively somewhere between these two extremes in the limit as individual clusters become numerous, small and compact…

$$\Pr(x, t \mid C) = \Pr(x \mid C) \Pr(t \mid x, C)$$
$$\approx \Pr(x \mid C) \Pr(t \mid C)$$

Start by Clustering Patients in X-Space



Nested ANOVA

Source	Degrees-of-Freedom	Interpretation
Clusters (Subgroups)	C = Number of Clusters	Local Average Treatment Effects (LATEs) are **Cluster Means**
Treatment within Cluster	Number of "Informative" Clusters ≤ C	Local Treatment Differences (LTDs)
Error	≥ Number of Patients – 2C	Uncertainty

Although a NESTED model can be (technically) WRONG, it is <u>sufficiently versatile</u> to almost always be USEFUL as the number of "clusters" increases.

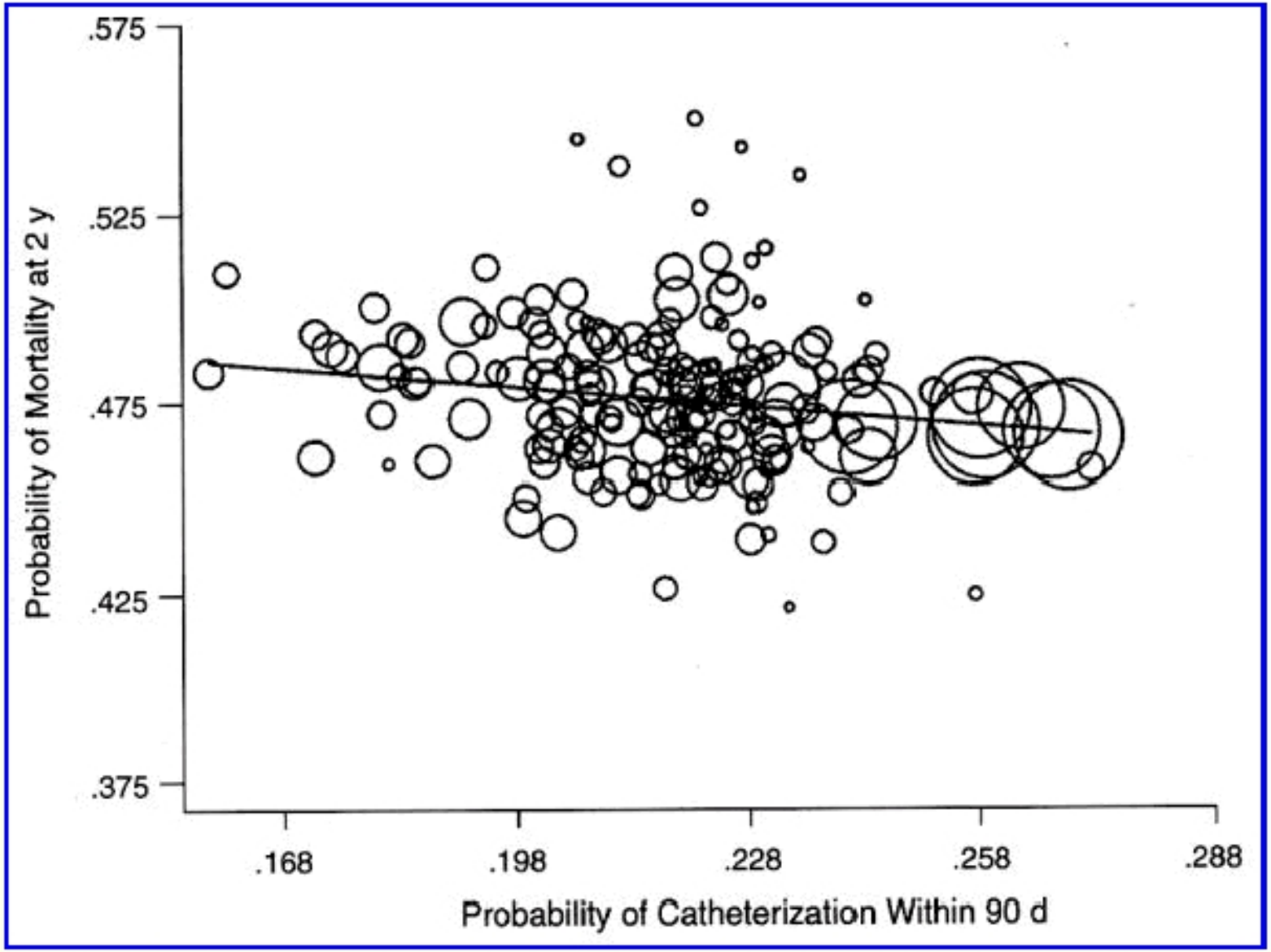
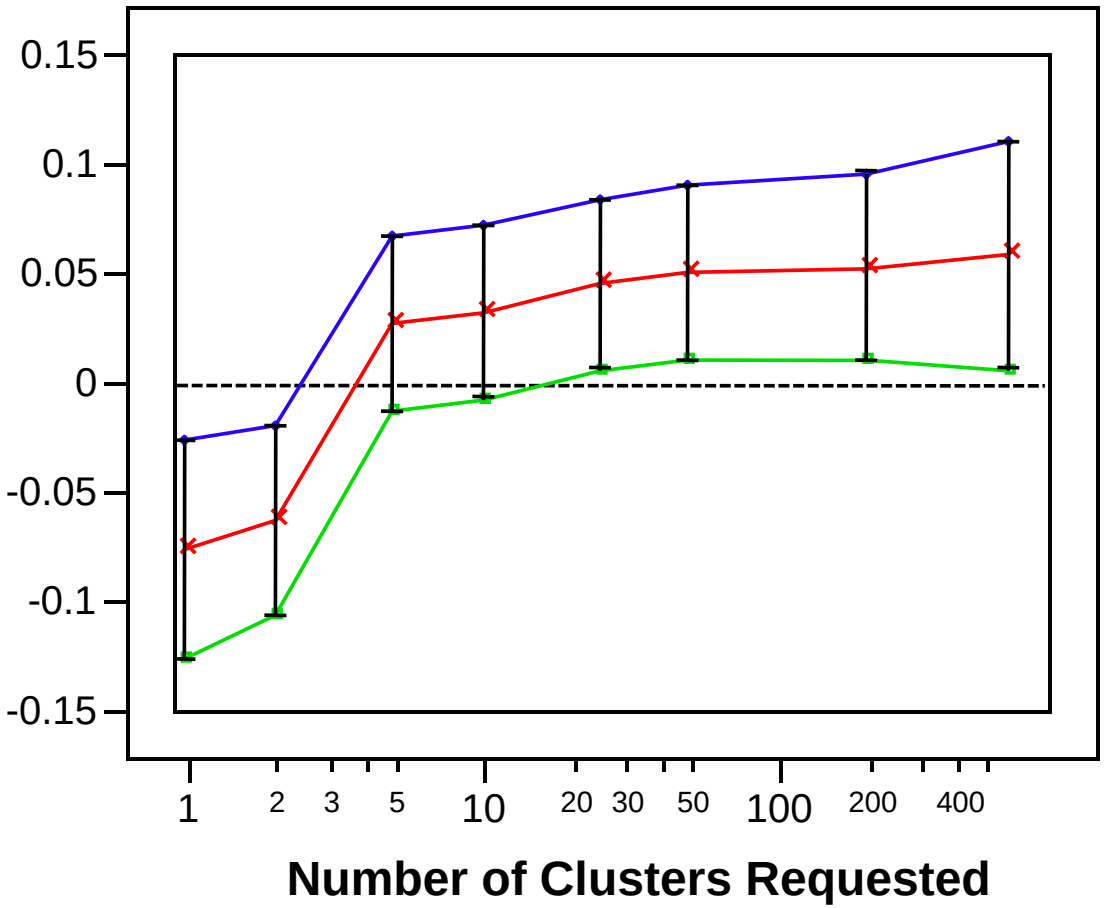


Figure 1. Graphic illustration of instrumental variables (IV) approach to estimating the incremental effect of invasive procedure use on acute myocardial infarction mortality at 2 years. Each circle represents a group of patients in a particular differential-distance group with a particular set of age, gender, and race characteristics. The model is thus fully adjusted for effects of observable patient demographic characteristics, and remaining differences in probabilities of catheterization across groups are strictly the result of different IV values for patients who are identical in terms of observable characteristics. The area of each circle is proportional to the number of patients in the group. The slope of the line fitted to the cells estimates the weighted-average effect of invasive procedure use

20 Year Mortality of 1314 Whickham Women: Smokers minus Nonsmokers



$LTD + 2\hat{\sigma}$

LTD

$LTD - 2\hat{\sigma}$

X = Age in 1974

Appleton DR, French JM, Vanderpump MPJ. "Ignoring a Covariate: An Example of Simpson's Paradox" ***Amer. Statist.*** **1996; 50: 340-341.**

What is a Treatment Effect?

Global / Marginal Inference…

Difference of Overall Averages

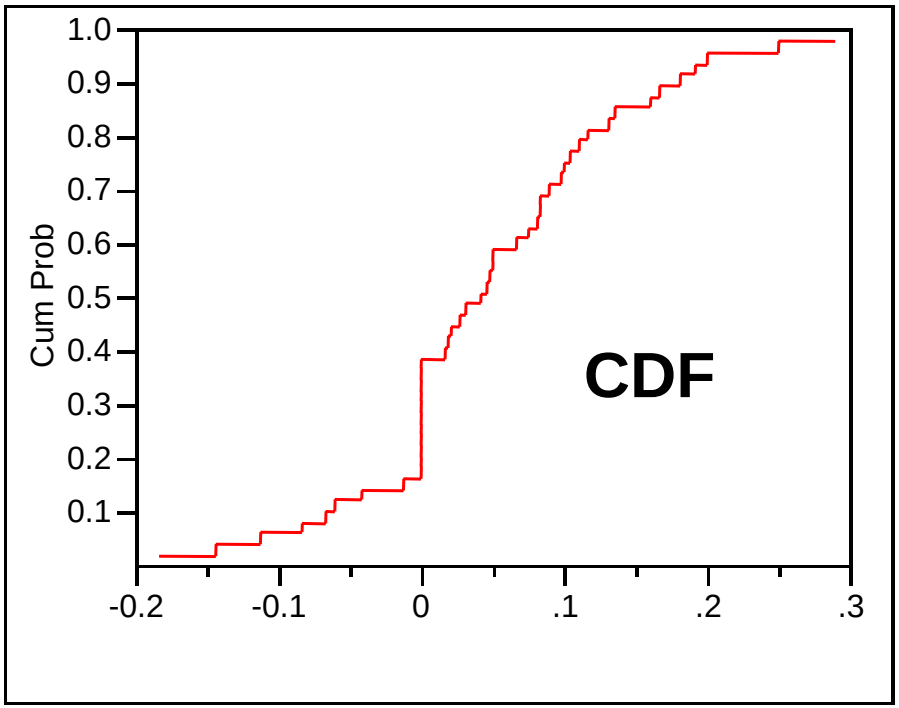
…one average for each treatment group or a simple “contrast” (single degree-of-freedom)

Local / Conditional Inference…

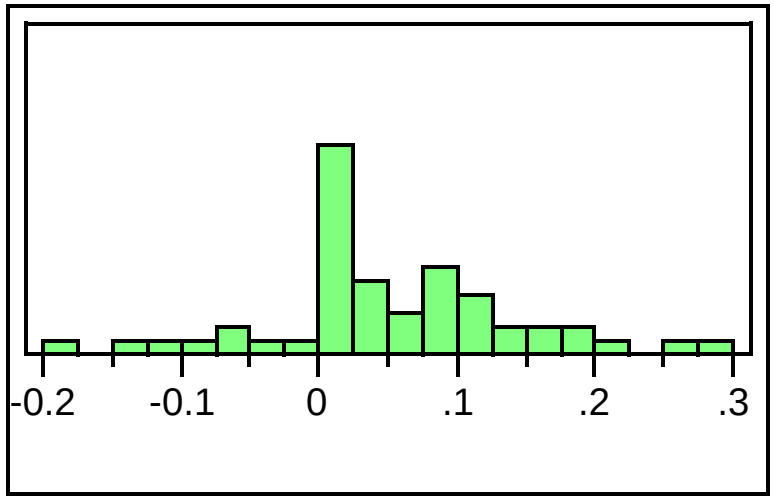
Distribution of Local Differences

…one average treatment difference within each subgroup of well-matched patients

LTD Distribution of Heteroskedastic Estimates…



49 Informative Clusters for 20 Year Mortality of 1314 Wickham Women: Smokers minus Nonsmokers



KEY QUESTIONS:

- **Is this distribution mostly just noise around some central value?**
- **How many local modes might this distribution really have?**
- **Do the Xs predict the most likely LTD for some (or all) patients?**

Ways for non-randomized studies to lose credibility…

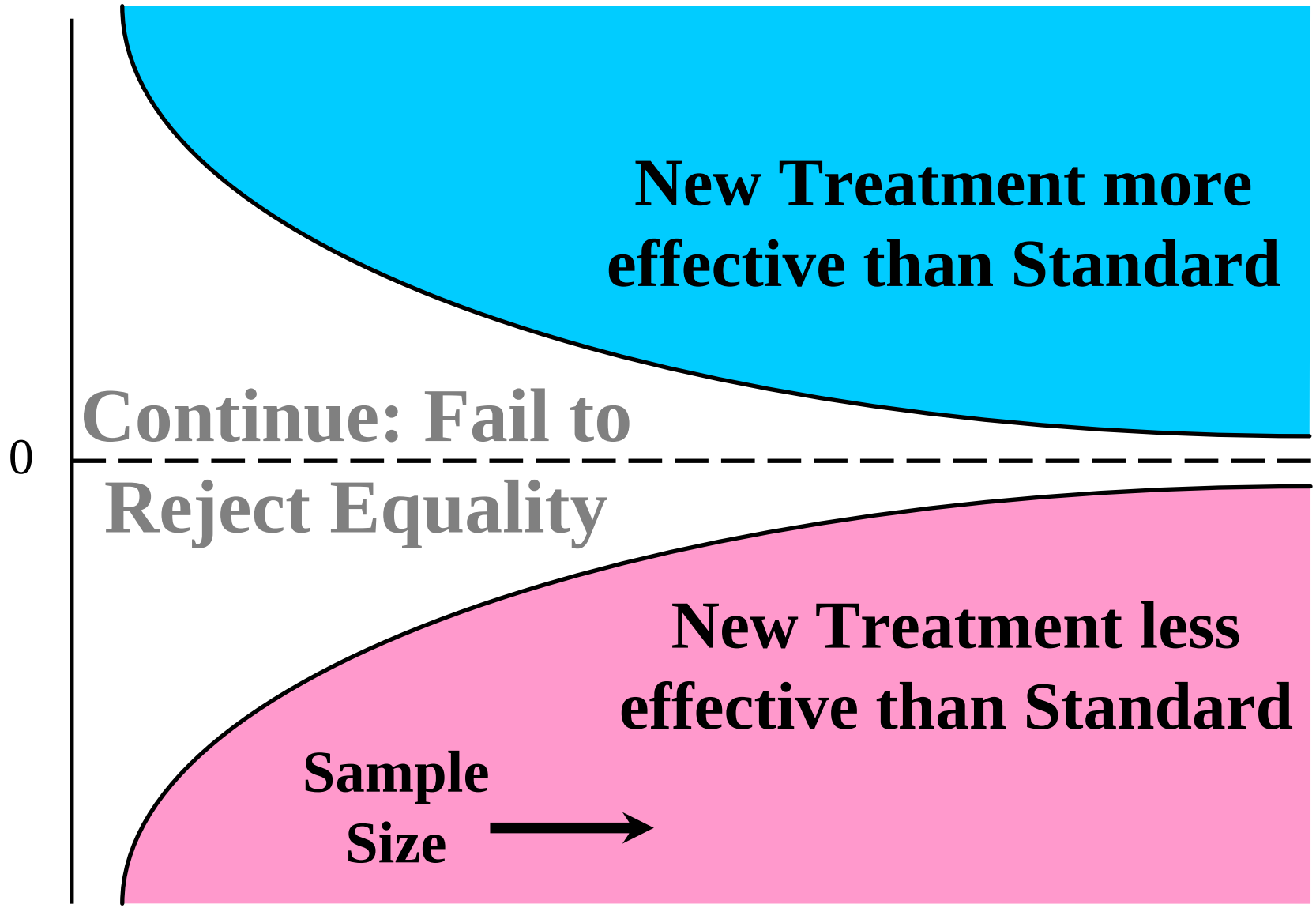
- Select patients by matching using covariates then analyze the data as two unrelated samples
- Use retrospective data to match patients in one FIXED ratio when this ratio actually varies (and should vary)
- Fail to adjust for treatment selection bias (see slide 5)
- Claim that your study results PROVE something
- Use data that nobody else has access to and refuse to share that data with other responsible researchers
- Reach unrealistic, unwarranted conclusions from data that responsible researchers do have access to
- Fail to perform sensitivity analyses (mult. imputation)
- Fail to point out obvious shortcomings in your data

Why do drug companies want to do good non-randomized studies…

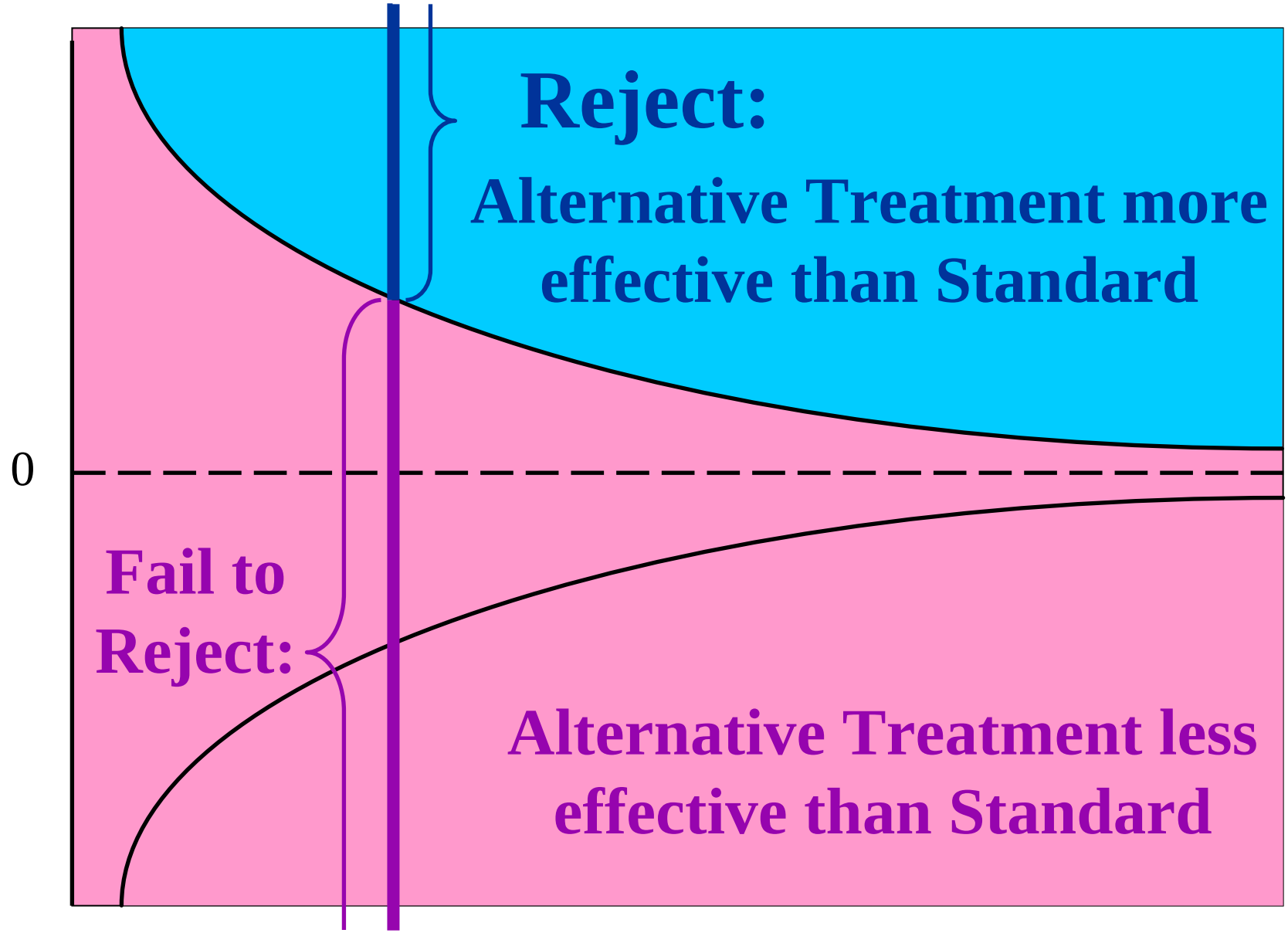
- Analyze realistic data without biasing outcomes via study participation incentives and restrictions
- Estimate actual costs (not protocol driven maximums)
- Provide disease state (burden of illness) information
- Save both TIME and $$$
- Drug companies employ gifted, professional scientists
- Drug companies use SCIENCE to mount high risk efforts to help patients and reward their stockholders
- Address situations where randomization is illegal, impossible or unethical and consent is difficult to get
- Study community-based interventions
- Study long-term treatment effects and safety

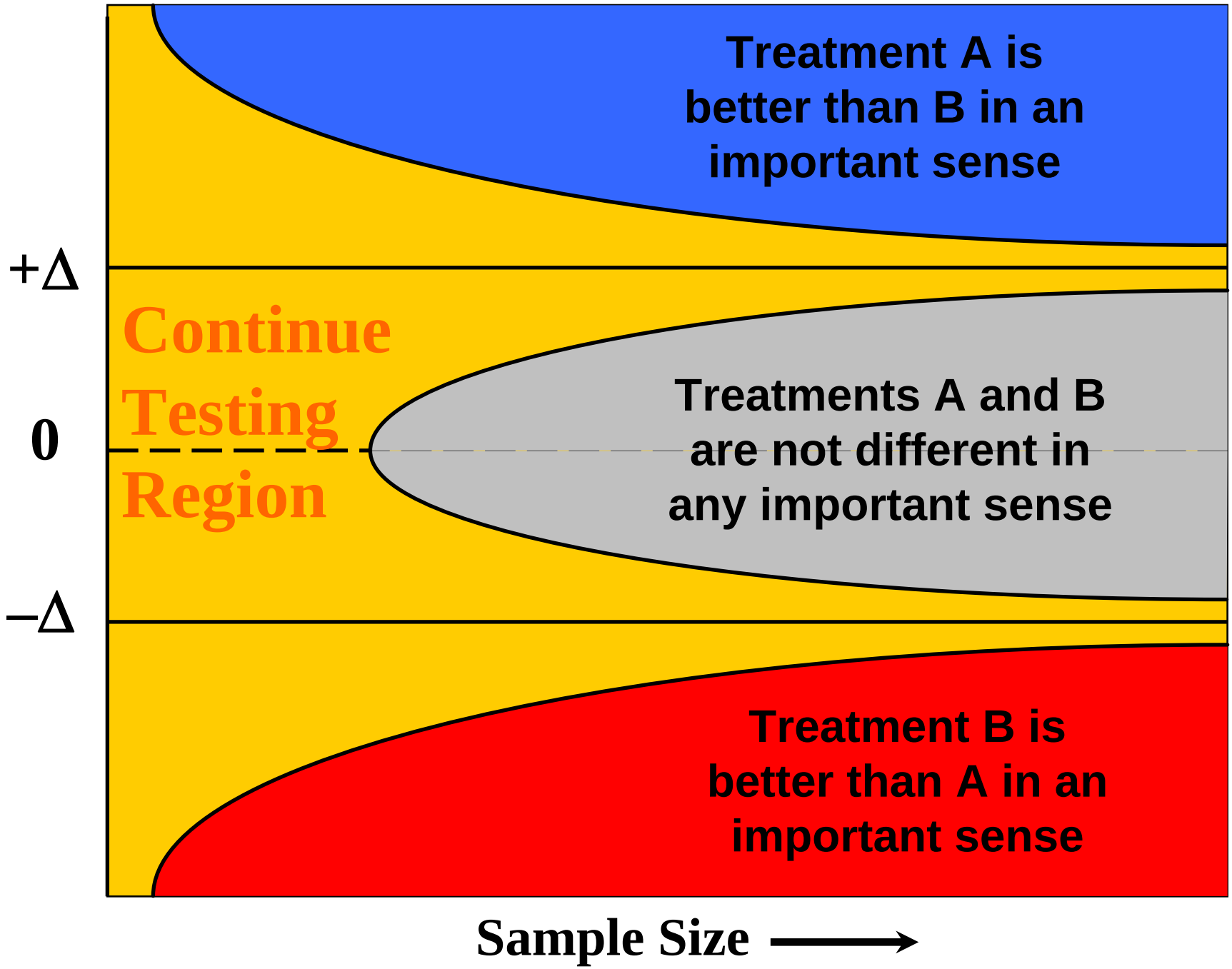
Even-Handed Testing of Composite Hypotheses

Sequential Testing of "Equality"



Fixed Sample Size: **“Sharp” NULL Hypothesis**



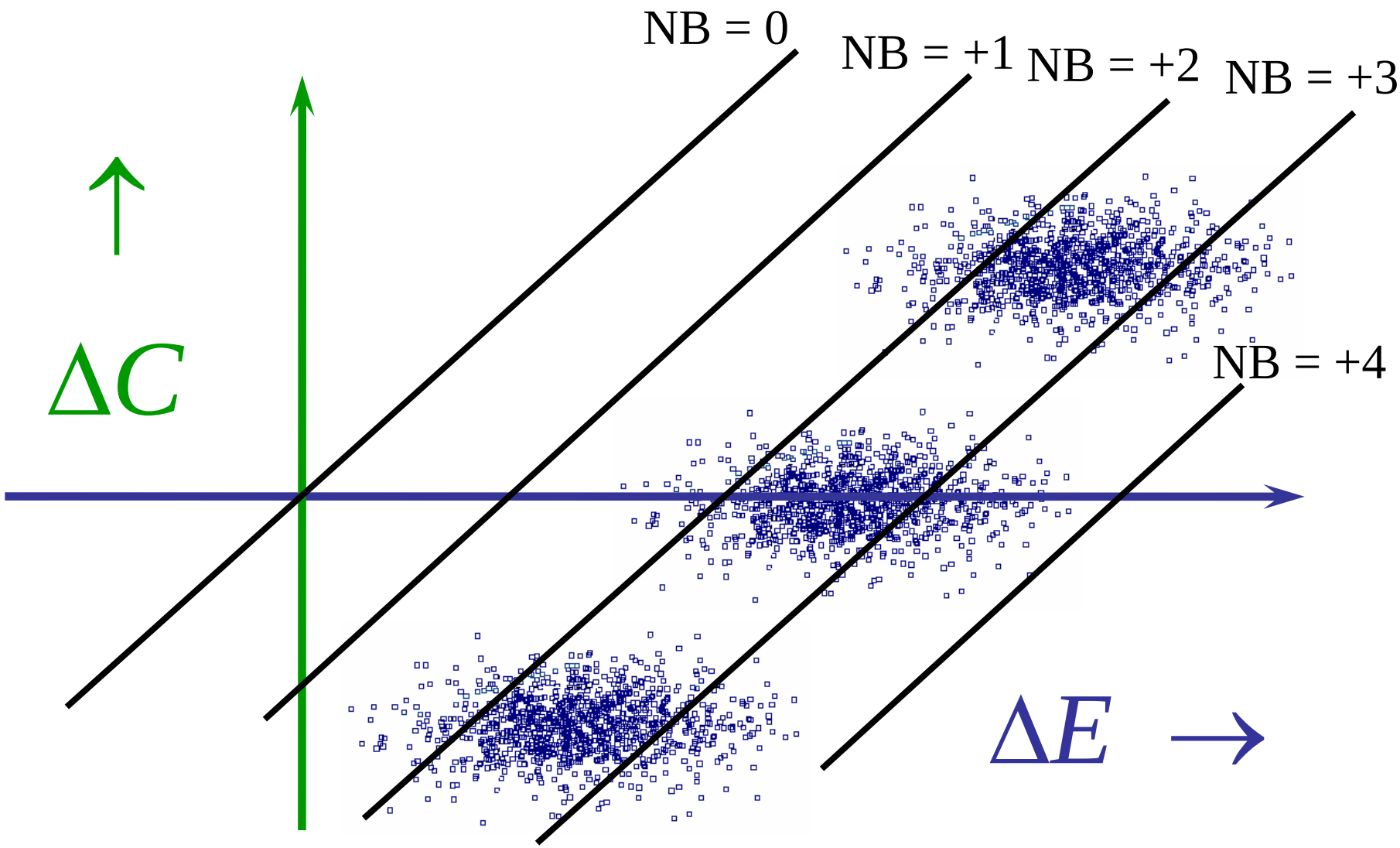

Treatment A is better than B in an important sense
+Δ
Continue Testing Region
0
Treatments A and B are not different in any important sense
–Δ
Treatment B is better than A in an important sense
Sample Size

- **It's NOT Possible to have "Too Much" Data!**
- **Here "∆" is the Admin / Policy Parameter**

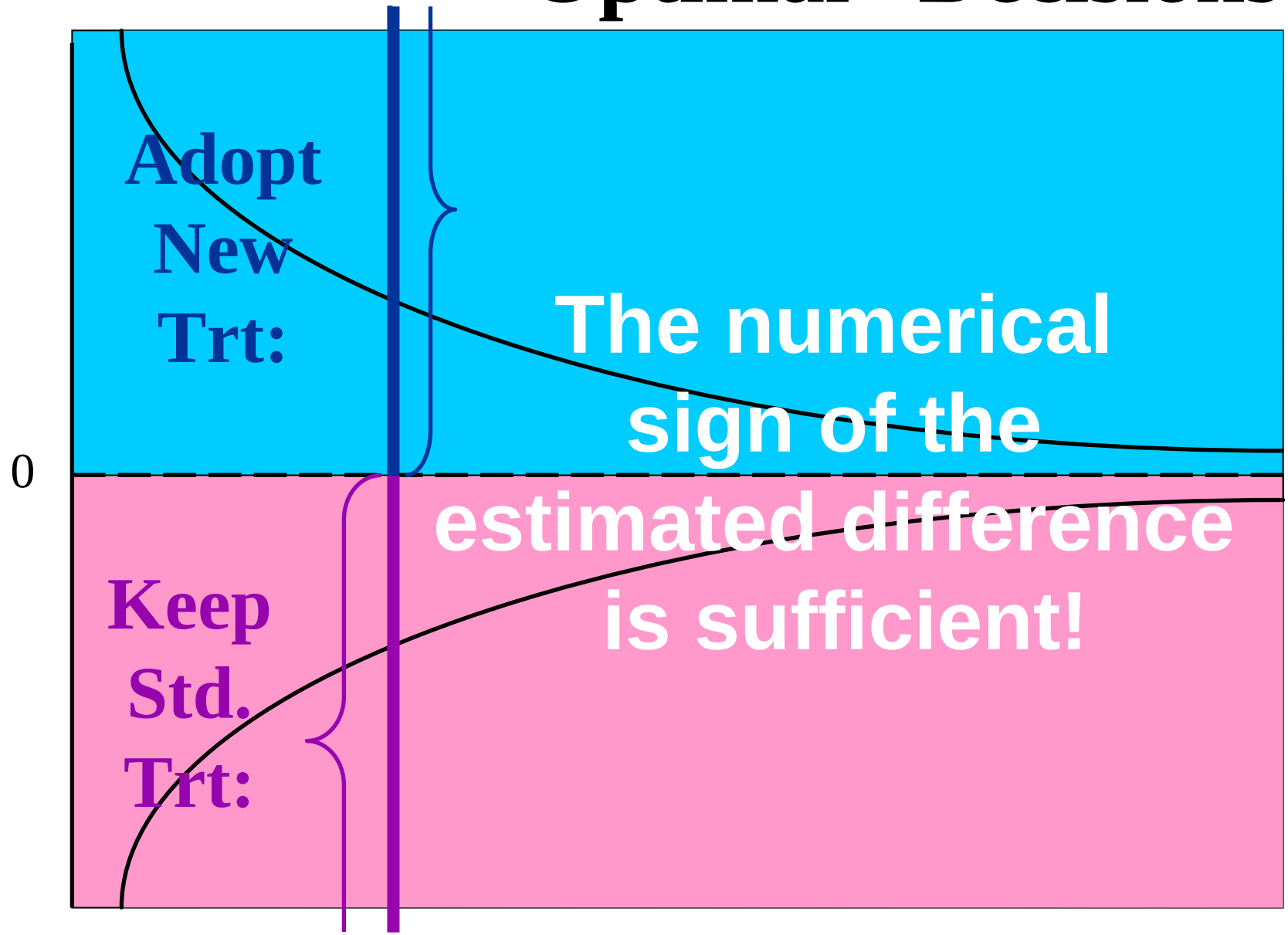
Cost-Effectiveness:

- Bivariate Inference !
- “Optimal” Decisions ?

Constant “Net Benefit” Lines of Slope = λ



Fixed Sample Size: "Optimal" Decisions



- **Inference is NOT Irrelevant!**
- **Poorly informed decisions CANNOT be truly "optimal"**
- **Here "λ" is the Admin / Policy Parameter ...not something to be varied to determine sensitivity**

Future “Needs”

- Continuing evolution in methodologies for and attitudes about analyses of non-randomized human studies
- Even-handed “testing terminology” compatible with voluminous data
- Postpone decisions whenever available data are insufficient to provide high confidence
- Statistical methods that get better-and-better (realistic, robust) as data become dense

Current Guideline Initiatives…

- **Randomized Clinical Trials**
  - **CONSORT: www.consort-statement.org**

- **Observational & Non-randomized Studies**
  - **STROBE: www.strobe-statement.org**
  - **TREND: www.trend-statement.org**

References

Hansen BB. "Full matching in an observational study of coaching for the SAT." *JASA* 2004; 99: 609-618.

Fraley C, Raftery AE. "Model based clustering, discriminant analysis and density estimation." *JASA* 2002; 97: 611-631.

Imbens GW, Angrist JD. "Identification and Estimation of Local Average Treatment Effects." *Econometrica* 1994; 62: 467-475.

McClellan M, McNeil BJ, Newhouse JP. "Does More Intensive Treatment of Myocardial Infarction in the Elderly Reduce Mortality?: Analysis Using Instrumental Variables." *JAMA* 1994; 272: 859-866.

More References

McEntegart D. "The Pursuit of Balance Using Stratified and Dynamic Randomization Techniques: An Overview." *Drug Information Journal* 2003; 37: 293-308.

Obenchain RL. "Unsupervised and Supervised Propensity Scoring in R: the USPS package" March 2005. http://www.math.iupui.edu/~indyasa/download.htm.

Obenchain RL. "Unsupervised Propensity Scoring: NN and IV Plots." *2004 Proceedings of the ASA*.

Rosenbaum PR, Rubin RB. "The Central Role of the Propensity Score in Observational Studies for Causal Effects." *Biometrika* 1983; 70: 41-55.

Rubin DB. "Bias Reduction using Mahalanobis Metric Matching." *Biometrics* 1980; 36: 293-298.

Testing References

Anderson, S. and Hauck, W. W. (1983). "A new Procedure for Testing Equivalence in Comparative Bioavailability and other Clinical Trials." Comm. Stat. A 12, 2663-2692.

Hauck, W. W. and Anderson, S. (1984). "A New Statistical Procedure for Testing Equivalence in Two-Group Comparative Bioavailability Trials." Journal of Pharmacokinetics and Biopharmaceutics 12, 83-91.

Schuirmann, D. J. (1987.) "A Comparison of Two One-Sided Test Procedure and the Power Approach for Assessing the Equivalence of Average Bioavailability." Journal of Pharmacokinetics and Biopharmaceutics, 15, 657-679.

Westlake, W. J. (1976). "Symmetric Confidence Intervals for Bioequivalence Trials." Biometrics 32, 741-744.

ICE References

Stinnett AA, Mullahy J. Net health benefits: a new framework for the analysis of uncertainty in cost-effectiveness analysis. *Med Decis Making (Special Issue on Pharmacoeconomics)* 1998; 18: S68-S80.

Claxton K. The irrelevance of inference: a decision-making approach to the stochastic evaluation of health care technologies. *J Health Econ* 1999; 18: 341-364.

Laska EM, Meisner M, Siegel C, Stinnett AA. Ratio-based and net benefit-based approaches to health care resource allocation: proofs of optimality and equivalence. *Health Econ* 1999; 8: 171-174.

O'Brien BJ, Sculpher MJ. Building uncertainty into cost-effectiveness rankings: portfolio risk-return tradeoffs and implications for decision rules. *Med Care* 2000; 38: 460–468.

Supplementary Slides

Model for the Conditional Expected Value of the Outcome Variable given a vector of Baseline Covariate values:

$$E(y_i \mid \vec{x}_i) = f[\vec{x}_i' \vec{\beta}_i]$$

…where f is a known function and β denotes a vector of parameters to be estimated.

- **Subscript "i" on β could imply a local, conditional model that might be quite general and flexible.**
- **No subscript on β implies no variation from patient-to-patient. This is a smooth, GLOBAL model making strong (and possibly unrealistic) assumptions.**

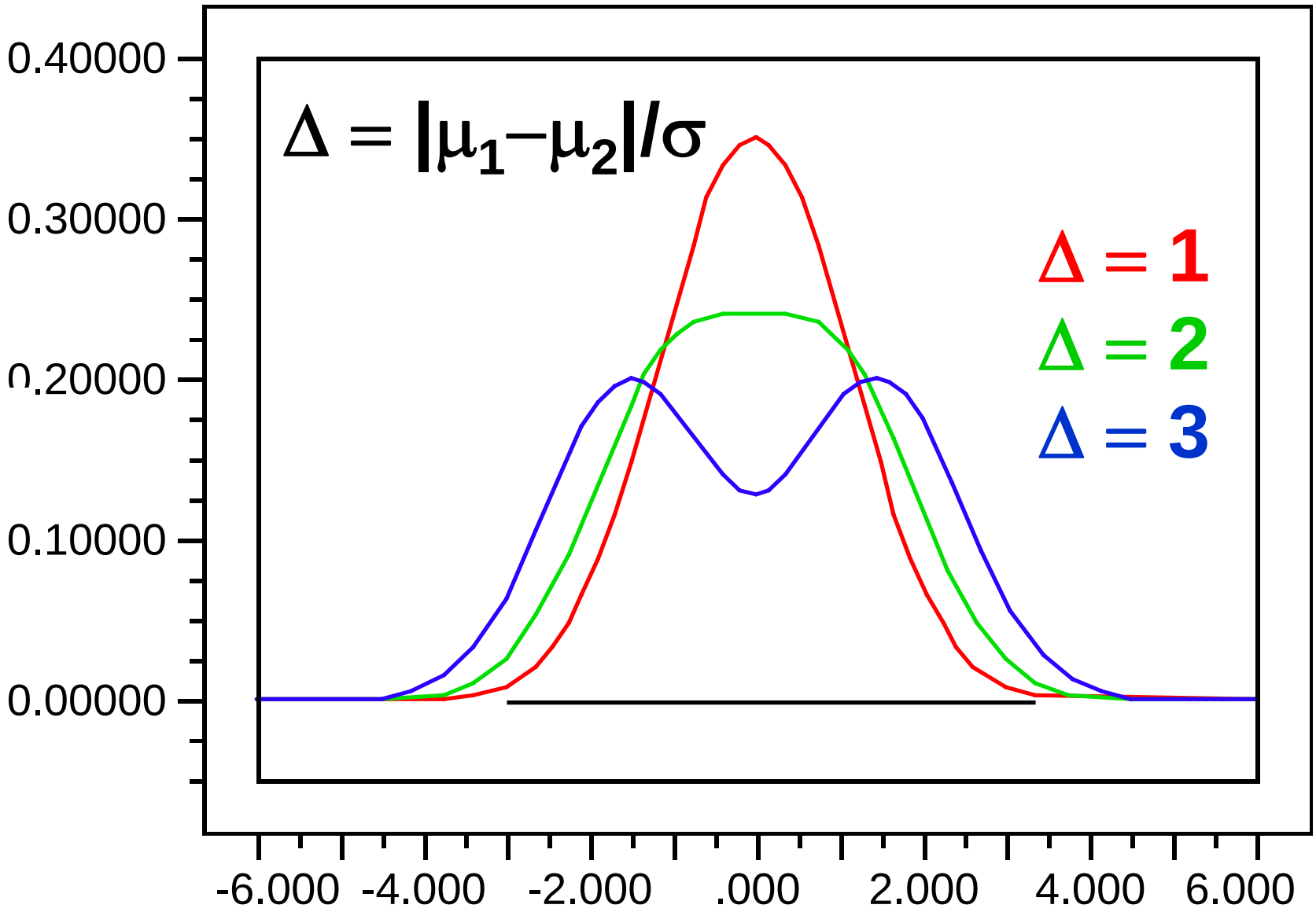
Global CA Model:

$$E(y_i \mid \vec{x}_i) = f[\vec{x}_i' \vec{\beta}]$$

…where f is a known function and β is an unknown vector of parameters common to all patients.

- **Primary purpose is NOT conditional (local) inference.**
- **This is really a smooth, GLOBAL model making strong (and possibly unrealistic) assumptions.**
- **Parameter estimates from this model might be of interest primarily for their MARGINAL implications.**

Mixture Joint Density:



Survey of 1314 Whickham Women

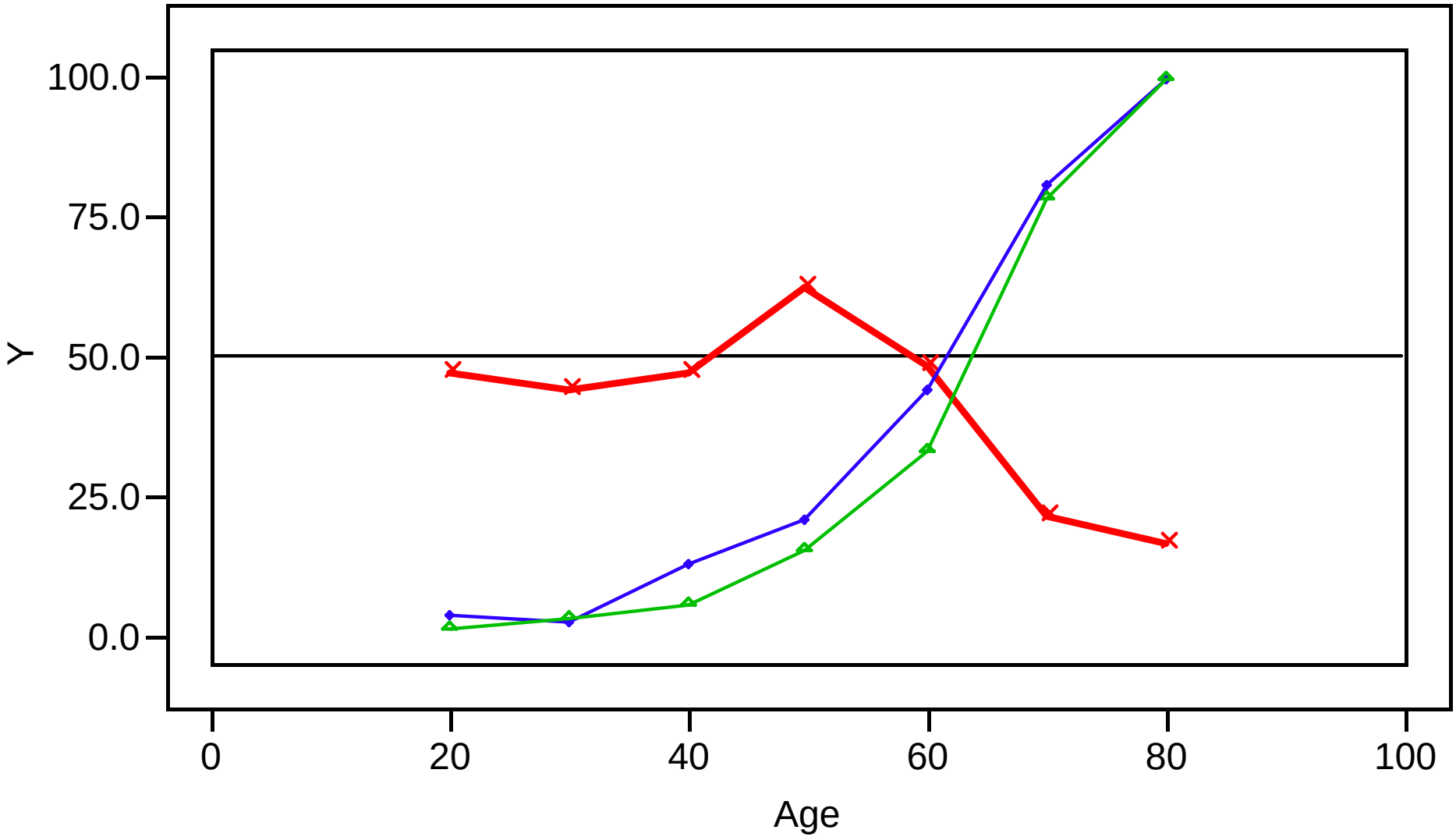
y = 20 year mortality (yes or no) in 1995 follow-up study of a survey made in 1972-1974

t = smoker or non-smoker at the time of the initial survey

x = age decade (20, 30, 40, 50, 60, 70 or 80) at the time of the initial survey

Appleton DR, French JM, Vanderpump MPJ. "Ignoring a Covariate: An Example of Simpson's Paradox" ***Amer. Statist.*** **1996; 50: 340-341.**

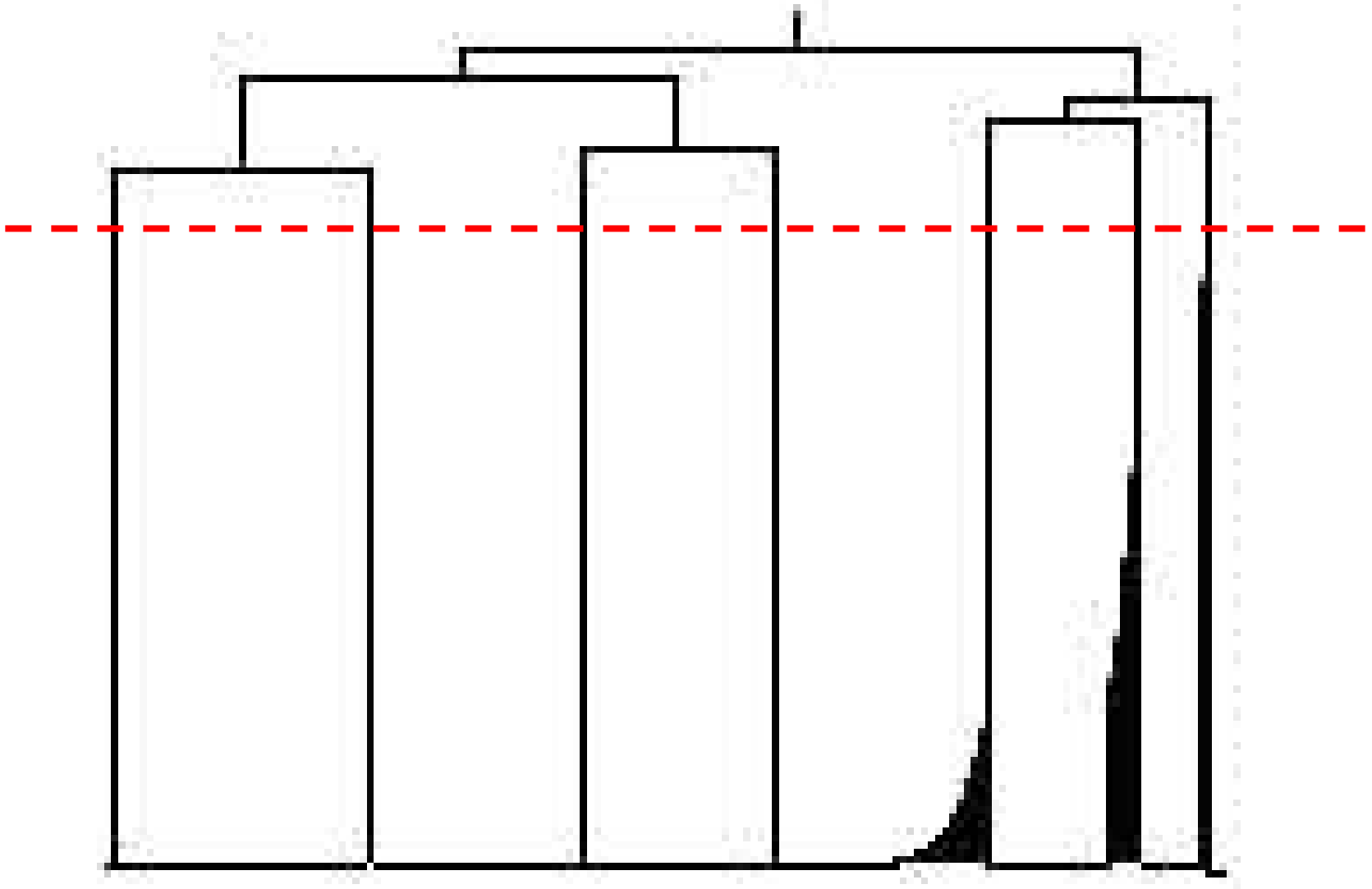
Simpson's Paradox At Work:



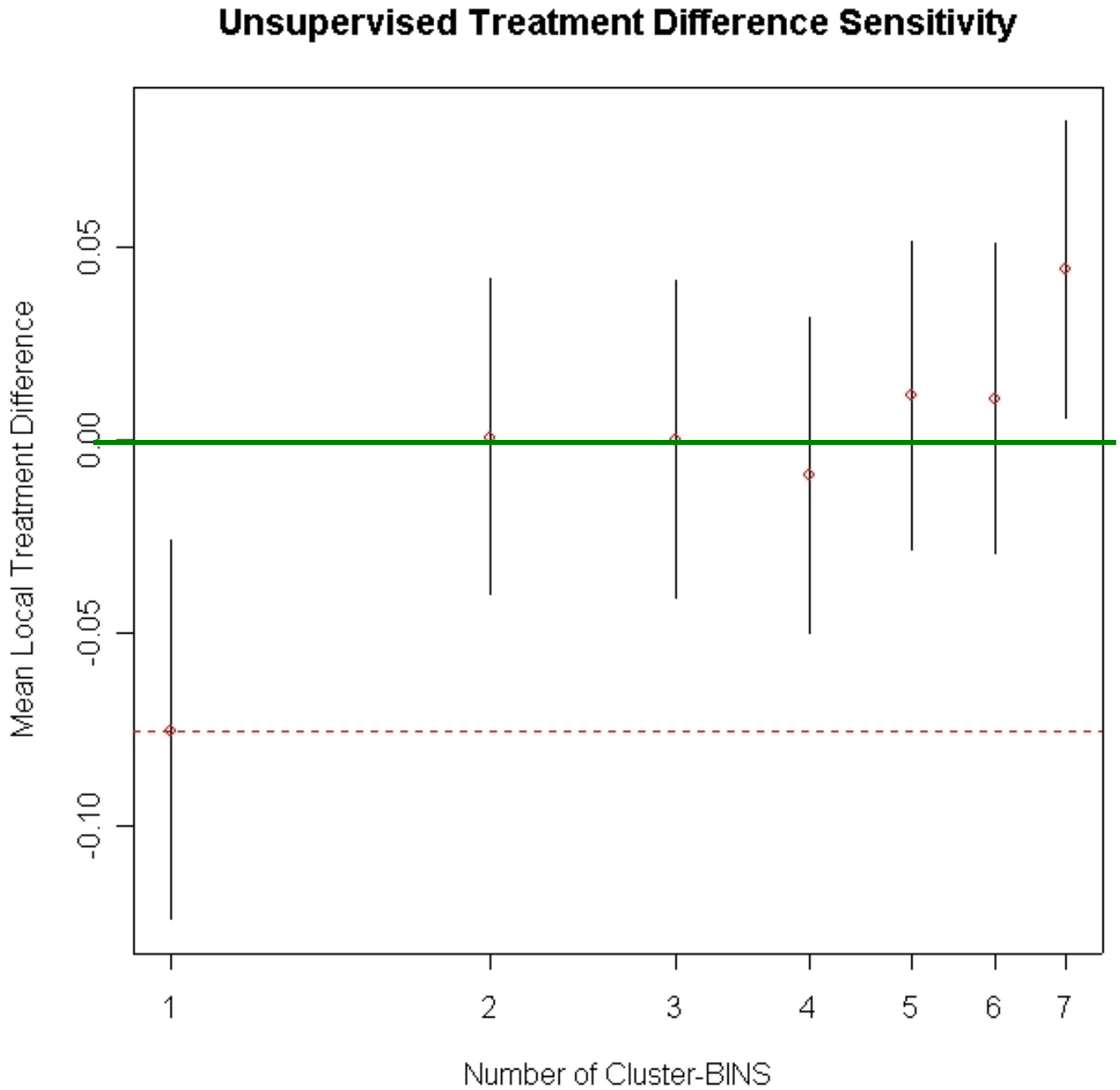
Percentage of Smokers by Age Decade

…and 20 Year Mortality Percentages for Non-Smokers and Smokers

Hierarchical Clustering of Women by Age:



…more than 7 “age decades” can only be formed in arbitrary ways.

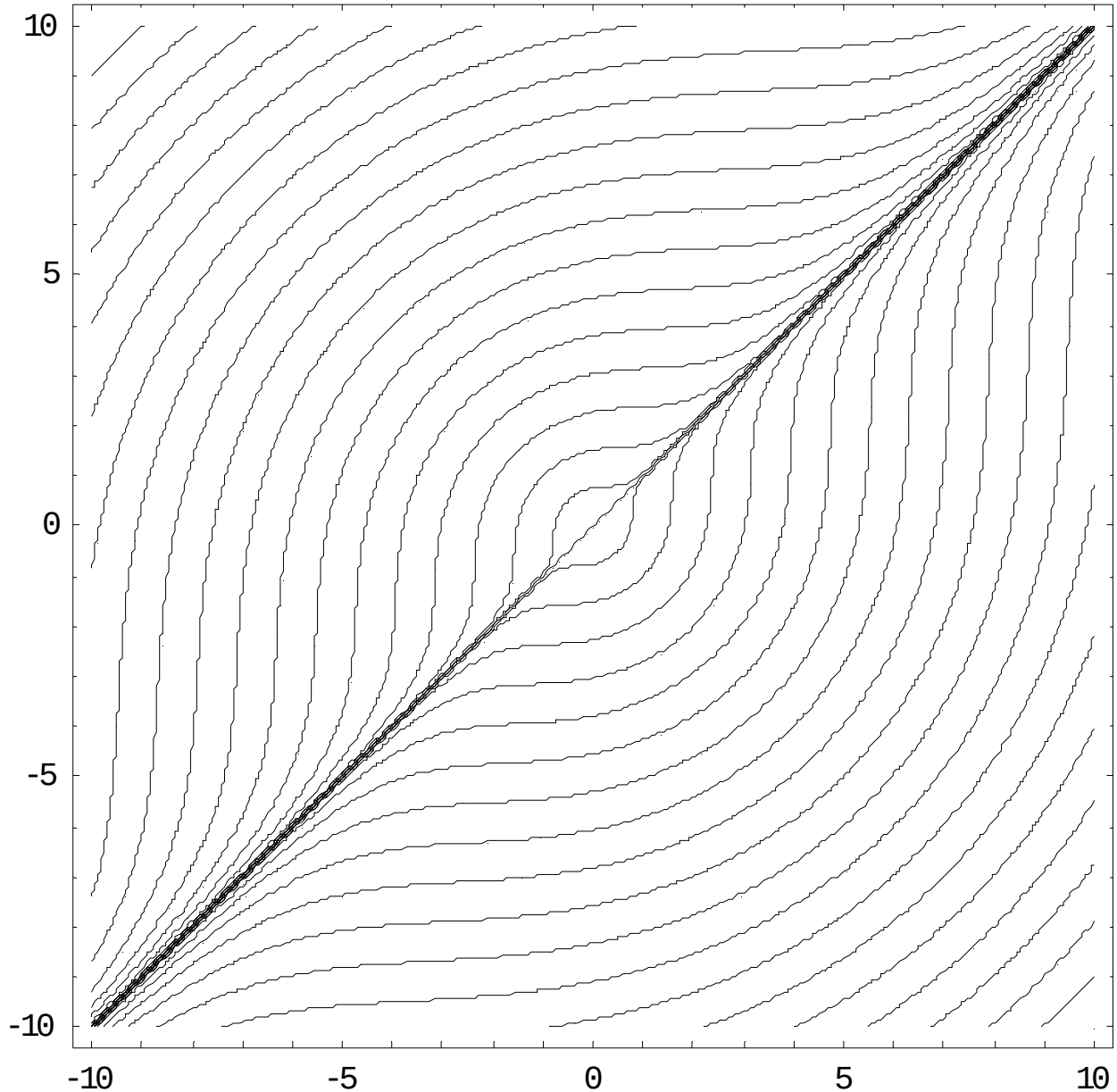


20 Year Mortality Rate Difference

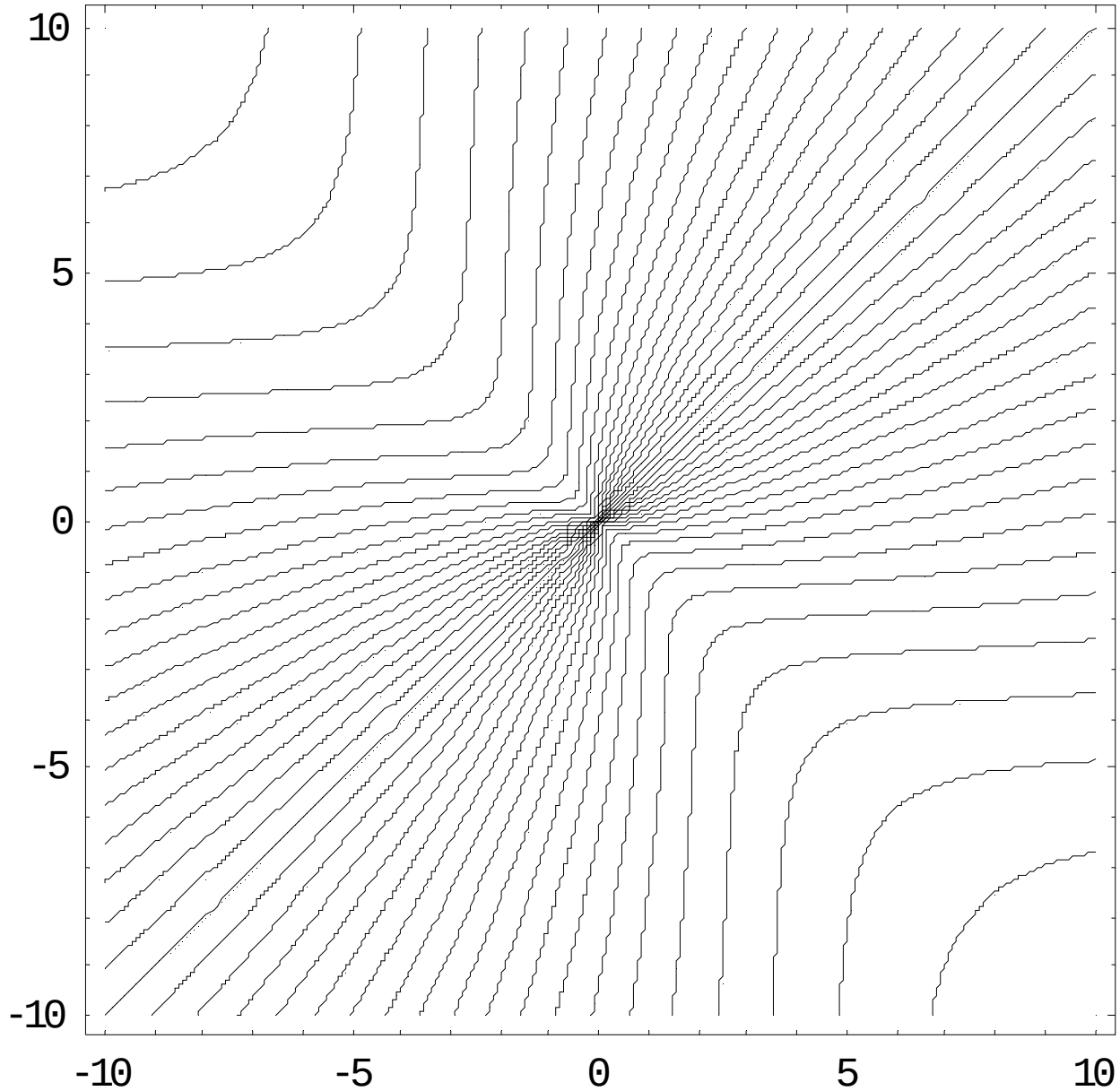
Overall Mean and +/-Two Sigma Limits for the Distribution of LATE Differences:

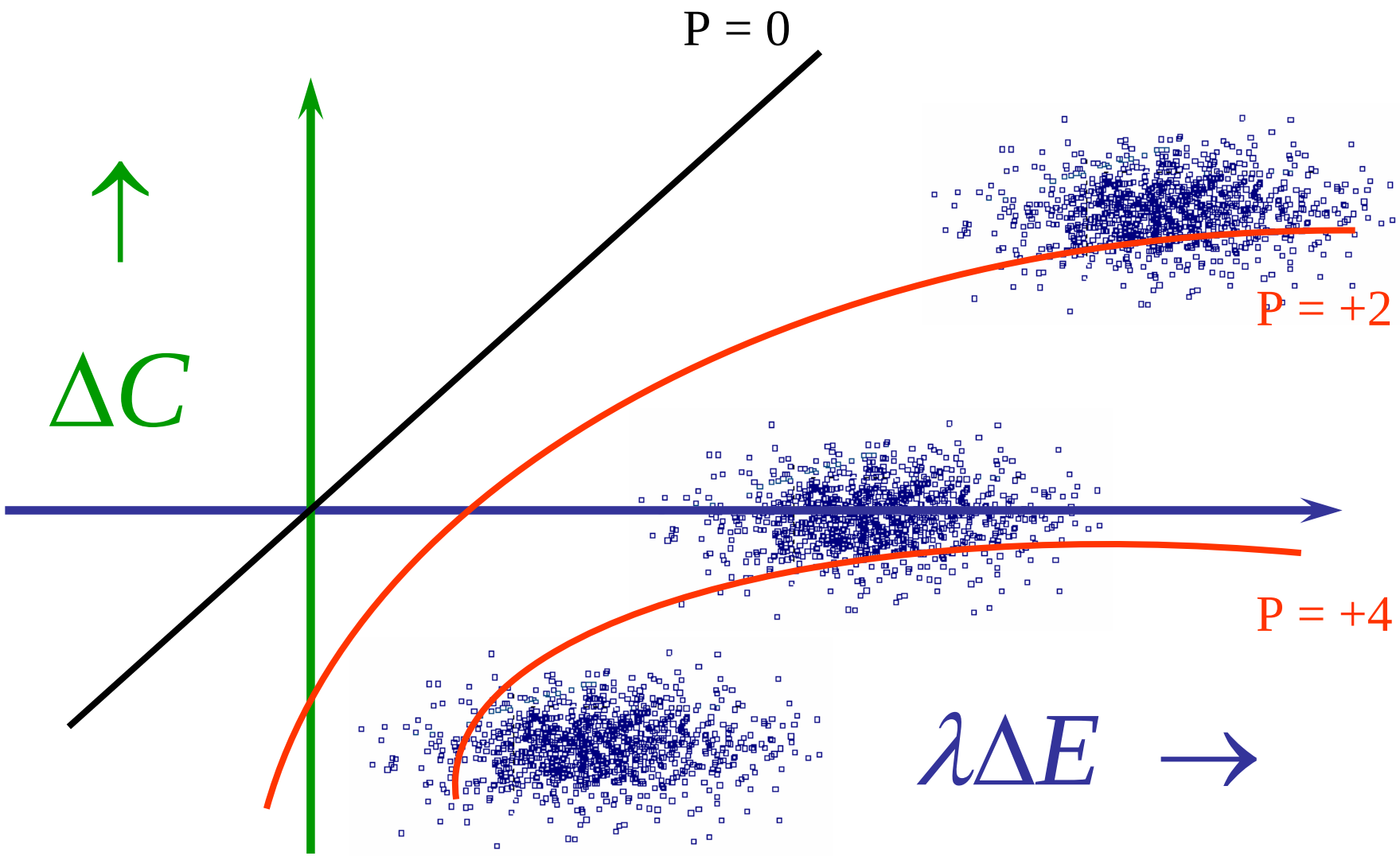
Mortality Rate of Smokers minus that of Nonsmokers.

Non-Linear Preference Map: $\beta = 1, \gamma = 0.2$



Non-Linear Preference Map: $\beta = 0.2, \gamma = 1$





Here we see the 3 cases are not equivalent!

"Wiper Blade" Confidence Regions:

These 3 cases are quite different!