Comparing Minimum Hellinger Distance Estimation and Hypothesis Testing to Traditional Statistical Analyses – a Simulation Study

Matthew M Hutmacher[1,*], Anand Vidyashankar[2], and Debu Mukherjee[1]
[1]PGRD Ann Arbor Laboratories, Pfizer Inc, Ann Arbor, MI 48105, USA
[2]Department of Statistical Science, Cornell University, Ithaca, NY 14853, USA
[*]Corresponding author: Matt.Hutmacher@Pfizer.com

1. INTRODUCTION

- The Hellinger Distance (HD) is a discrepancy measure between distributions.
- The HD metric is also a well-defined objective function, HD^2, between the empirical density of the data (h_n) and a model density (f_θ).

$$HD^2(\theta) = 2 - 2\int f_\theta^{1/2} h_n^{1/2} dy, \quad \gamma(\theta) = \int f_\theta^{1/2} h_n^{1/2} dy$$

  - ➢HD^2 (or γ) can be optimized to estimate a vector of parameters (θ).
  - ➢It provides an absolute estimate of lack-of-fit
  - ➢Is bounded (HD^2 = 0 when $f \equiv g$ and ranges from 0-2 inclusive).
- Minimizing the Hellinger distance objective function for estimation (MHDE) results in consistent and asymptotically normal estimates, which are theoretically asymptotically efficient and robust.
- This work evaluates MHDE and compares it to traditional ANOVA and ANCOVA analyses via a simulation study. Specific objectives were:
  - ➢Efficiency of the estimates
  - ➢Robustness of estimated treatment effect
  - ➢Robustness of estimated residual variability
- Bias in the estimates as well as type 1 error and power for testing the difference between treatment means were compared.

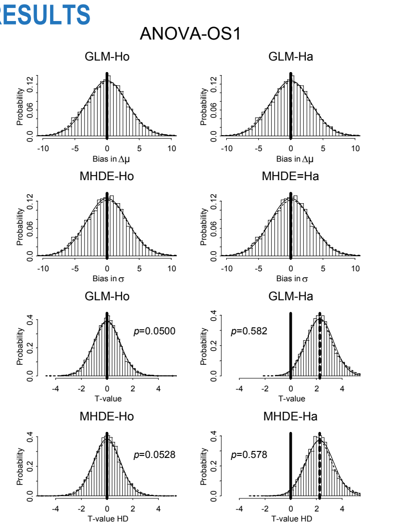
2. METHODS

- Data (y) for two treatment groups (TG) – a 'reference' (R) and a 'test' (T) (n=20/group, 1 observation/subject) – were generated using
  - ➢Ho: μ_R = 100, μ_T = 100, σ = 10; Ha: μ_T = 107
  - ➢For ANCOVA: β (slope) = 0.15 with covariate z ~ N(0,50^2)
- Hypotheses were evaluated for three outlier generation scenarios (OS)
  - ➢OS1: 0% outliers (efficiency)
  - ➢OS2: 10% outliers assigned to the T group systematically (robustness) Generated by sampling y ~ Uniform(140,160)
  - ➢OS3: 10% outliers assigned to the T group randomly (robustness) Generated by E(y|z) + δ, δ ~ Uniform(-80,80)
- The integral in HD^2 was approximated (M=5000):

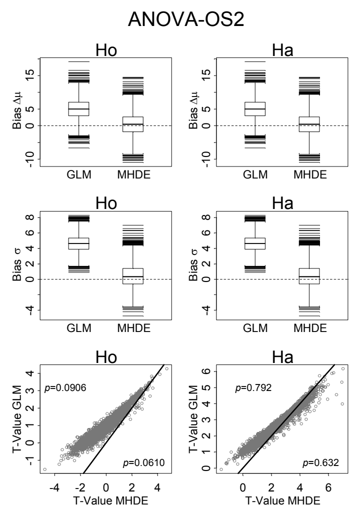
$$\gamma(\theta) = \int f_\theta^{1/2} h_n^{1/2} dy = \int \frac{f_\theta^{1/2}}{h_n^{1/2}} h_n dy = E_y\left(\frac{f_\theta^{1/2}}{h_n^{1/2}}\right)_n \approx \frac{1}{M}\sum_j \left[\frac{f_\theta(y_j^*)}{h_n(y_j^*)}\right]^{\frac{1}{2}}$$

  where $y_j^* \sim h_n$, h_n estimated using the Epanechnikov kernel ($K(w) = \frac{3}{4}(1-w^2)$, $|w| \le 1$) with tuning constant = 1.5 (not optimized).
- 5000 simulations per model/hypothesis/data generation scenario were completed. Estimations performed using SAS 8.2 PROC GLM for the traditional analyses and PROC NLP for MHDE

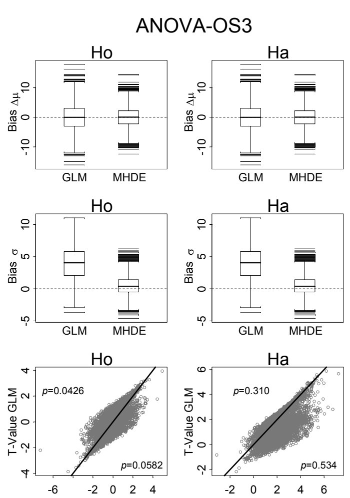
3. RESULTS



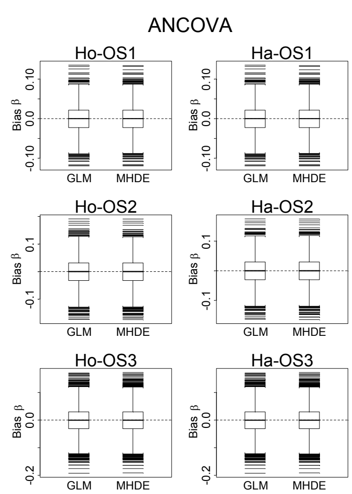
➢ GLM and MHDE demonstrated similar (lack of) biases, distributions of the estimates (normality), and p-values.



➢ Systematic outliers caused biases for GLM. MHDE demonstrated robust estimates. GLM Type 1 error and power inflated – MHDE more in line with OS1.



➢ Random outliers caused bias in σ estimates for GLM – MHDE was robust. GLM power deflated. MHDE modestly decrease in power with modest increase in Type 1 error.



➢ For ANCOVA, GLM and MHDE showed little bias in β estimates. Estimates of $\Delta\mu$ and σ showed similar distributions to ANOVA, and p-values were similar to ANOVA – depending on OS (not shown).

4. DISCUSSION

- MHDE, in some sense, finds estimates indexing the normal distribution closest to the observed data's distribution.
- MHDE T-values calculated by plugging the MHDE estimate of σ into the traditional ANOVA/ANCOVA standard error formula.
- Systematic outliers (OS2) – GLM
  - ➢ Overestimated the treatment differences ($\Delta\mu$).
  - ➢ Overestimated the residual variability (σ).
  - ➢ Resulted in inflated power (0.792) at the cost of inflated type 1 error (0.091) – (nominal power $\cong$0.6 and type 1 = 0.05)
- Systematic outliers (OS2) – MHDE
  - ➢ Provided robust estimates of $\Delta\mu$ and σ.
  - ➢ Type 1 error (0.0610) and power (0.632) marginally inflated – more consistent with no outliers (OS1).
- Random outliers (OS3) – GLM
  - ➢ Overestimated σ.
  - ➢ Increased variability in $\Delta\mu$ distribution
  - ➢ Resulted in deflated power (0.310), yet maintained the type 1 error rate (0.043).
- Random outliers (OS3) – MHDE
  - ➢ Provided robust estimates of $\Delta\mu$ and σ.
  - ➢ Increased marginally the type 1 error (0.058) with a modest power decrease (0.534) – more consistent with no outliers (OS1) than GLM.
- The robustness of MHDE derives from little 'weight' $(\cong \frac{1}{n})$ given to the observations (and outliers).
- Estimates obtained by minimizing the squared error $(y-\mu)^2$ (ANOVA/ANCOVA estimators – GLM) are adversely influenced by the 'usual weight' given to outlying observations.
- Ultimately, GLM sacrifices power to maintain the type 1 error rate when random outliers are present. This is achieved by inflated σ estimates – perhaps making trials 'appear' more variable and inconsistent. Sample size must be spent to overcome this variability. Elimination of outliers is subjective and subject to controversy even if pre-specified criteria are used.
- MHDE demonstrated little bias and influence with respect to the outlying observations yielding more consistent data interpretations across scenarios.

5. REFERENCES

R. Beran. Minimum Hellinger distance estimates for parametric models. *The Annals of Statistics.* 3:445-465 (1977).

A. Cheng and A. N. Vidyashankar. Minimum Hellinger distance estimation for randomized play the winner design (submitted for publication).