Group Sequential Tests: Beyond Exponential Family Models

Sergey Tarima and Nancy Flournoy

starima@mcw.edu
Institute for Health and Equity
Medical College of Wisconsin

flournoyn@umsystem.edu
Department of Statistics
University of Missouri

MODA 13, Southhampton, UK

July, 2023

Introduction and a Few Summary Comments

- We consider the impact of *(informative) interim adaptations (focus on and Cauchy data)*.
  - *Informative* adaptations use collected data that is non-ancillary to the parameter of interest θ (e.g. location parameter) for making interim decisions.
  - An example of *informative* interim adaptations is a group sequential design.
  - An example of a *non-informative* adaptation is a sample size recalculation based re-estimated a scale parameter σ^2 in the location scale family.
- Informative adaptation leads to (Fisher) information loss: the observed sample and the likelihood function based on the observed sample does not accumulate all sampling evidence.
- Wolfowitz 1947 suggested sequential version of Cramer-Rao lower bound. Simons 1980 shows it did not work even in Gaussian settings.
- We derive new lower bounds for the variance and for the MSE.

Two Samples: No Early Stopping (1)

- Let $\boldsymbol{X}_1 = (X_1, \ldots, X_{n_1})$ and $\boldsymbol{X}_2 = (X_{n_1+1}, \ldots, X_{n_1+n_2})$ be two samples of i.i.d.r.v. with $X_i \sim f_X(x|\theta)$.
- Joint *stage-specific* densities $f_{\boldsymbol{X}_d}(\boldsymbol{x}_d|\theta)$ are used to define Fisher information in $\boldsymbol{X}_d$ $(d = 1, 2)$,

$$\mathcal{I}_{\boldsymbol{X}_d}(\theta) = Var\left[\frac{\partial}{\partial\theta}\log f_{X_d}(\boldsymbol{x}_d|\theta)\right] = -E_{\boldsymbol{X}_d}\left[\frac{\partial^2}{\partial\theta^2}\log f_{\boldsymbol{X}_d}(\boldsymbol{X}_d|\theta)\right]$$

Let $\tilde{\theta}_d$ be an estimator based on $\mathbf{X}_d$ and $\mathrm{E}[\tilde{\theta}_d] = \theta + b_d(\tilde{\theta}|\theta)$, then Cramer-Rao Lower Bound (CRLB) for MSE is

$$\mathrm{E}\left([\tilde{\theta}_d - \theta]^2\right) \geq \frac{\left[1 + \frac{\partial}{\partial\theta} b_d(\tilde{\theta}|\theta)\right]^2}{\mathcal{I}_{\boldsymbol{X}_d}(\theta)} + b_d^2(\tilde{\theta}|\theta).$$

Two Samples: No Early Stopping (2)

Since $\mathcal{I}_{\boldsymbol{X}} = \mathcal{I}_{\boldsymbol{X}_1} + \mathcal{I}_{\boldsymbol{X}_2}$,

$$\mathrm{E}\left([\tilde{\theta} - \theta]^2\right) \geq \frac{\left[1 + \frac{\partial}{\partial\theta} b(\tilde{\theta}|\theta)\right]^2}{\mathcal{I}_{\boldsymbol{X}}(\theta)} + b^2(\tilde{\theta}|\theta),$$

where $\tilde{\theta}$ is based on both samples, $\mathrm{E}[\tilde{\theta}|\theta] = \theta + b(\tilde{\theta}|\theta)$.

The Cramer-Rao Lower Bound holds for all regular estimators.

In one-parameter exponential family with canonical parameterization, the MLE ($\hat{\theta}$) attains the lower bound (for Cauchy, asymptotically).

$$\mathrm{E}\left([\hat{\theta} - \theta]^2\right) = \frac{\left[1 + \frac{\partial}{\partial\theta} b(\hat{\theta}|\theta)\right]^2}{\mathcal{I}_{\boldsymbol{X}}(\theta)} + b^2(\hat{\theta}|\theta),$$

The CRLB depends on the bias and its derivative. A fair application of CRLB for comparing estimators should be conditioned on bias.

Probability Space Changes with an Option to Stop

Suppose $\tilde{\theta}_2$ is only observed if $\tilde{\theta}_1 < c_1$.

1. $\tilde{\theta}_2$ becomes impossible (not just missing) when $\tilde{\theta}_1 \geq c_1$.
   $\Rightarrow$
   $\tilde{\theta} = (\tilde{\theta}_1, \tilde{\theta}_2)$ is non-observable when $\tilde{\theta}_1 \geq c_1$
   $\tilde{\theta}$ is no longer defined on R^2.
2. To define a probability space, we define *observable random variables* and their joint probability measure.
3. The sample space for D (stopping stage) is $\{1, 2\}$ and the sample space for $\tilde{\theta}$ is

$$\Omega = \{\tilde{\theta}_1 \geq c_1\} \quad \cup \quad \{\{\tilde{\theta}_1 < c_1\} \quad \cap \quad \{\tilde{\theta}_2 \in R\}\}$$

4. The distribution of $\tilde{\theta}$ changes with an early stopping rule: information about θ is now in $(D, \tilde{\theta})$.

New Distribution is Very Different from the Old One

- The distribution of D is $f_D(d \mid \theta) = \prod_{d=1}^{2}[\Pr(D = d \mid \theta)]^{I(D=d)}$.
- Then, $\tilde{\theta}$ is a mixture of $\tilde{\theta}_1|D = 1$ and $\tilde{\theta}_{(2)}|D = 2$:

$$f_{\tilde{\theta}}(t) = \Pr(D = 1 \mid \theta) f_{\tilde{\theta}_1|D=1}(t_1|\theta) + \Pr(D = 2 \mid \theta) f_{\tilde{\theta}_{(2)}|D=2}(t|\theta),$$

  where $\tilde{\theta}_{(2)}$ is calculated using both stages given $\tilde{\theta}_1 < c_1$.
- Conditional on $D = 1$, $f_{\tilde{\theta}_1|D=1}(t_1|\theta)$ is a left truncated density.
- Conditional on $D = 2$,

$$f_{\tilde{\theta}_{(2)}|D=2}(t|\theta) = \int_{-\infty}^{c_1} f_{\tilde{\theta}_{(2)}|\tilde{\theta}_1=x}(t(x)|\theta) f_{\tilde{\theta}_1|D=2}(x|\theta)dx$$

  which an integral of a conditional distribution with respect to a truncated density.

Cauchy Example with Early Stopping

Let $X \sim \pi^{-1}\left(1+(x-\theta)^2\right)^{-1}$ which is $Cauchy(\theta, 1)$. The objective is to test $H_0 : \theta = 1$ with 80% power at both $H_1 : \theta = 1.3$ and $H_2 : \theta = 1.6$, while controlling overall type 1 error at 5% with equal rejection probabilities at stages 1 and 2 under H_0.

Consider Cauchy MLE and LRT tests, conditional on $D = d$:

- the MLE $\hat{\theta}_{(d)}$ of θ is found by solving a score equation

$$U\left(\theta|\mathbf{X}_{(d)}\right) = \sum_{i=1}^{n_{(d)}} \frac{2|x_i - \theta|}{1 + (x_i - \theta)^2} = 0.$$

- the LRT (log-likelihood ratio) is

$$l_{(d)}(\mathbf{X}_{(d)}) = -2 \sum_{i=1}^{n_{(d)}} \left\{ \log\left[1 + (x_i - \theta_0)^2\right] - \log\left[1 + (x_i - \theta_1)^2\right] \right\} + c,$$

  where

$$c = -2 \log \Pr_{\theta_0}\left[l_1(\mathbf{X}_1) \geq c_1^{lrt}\right] + 2 \log \Pr_{\theta_1}\left[l_1(\mathbf{X}_1) \geq c_1^{lrt}\right].$$

Densities of Cauchy MLE(LRT) with Early Stopping

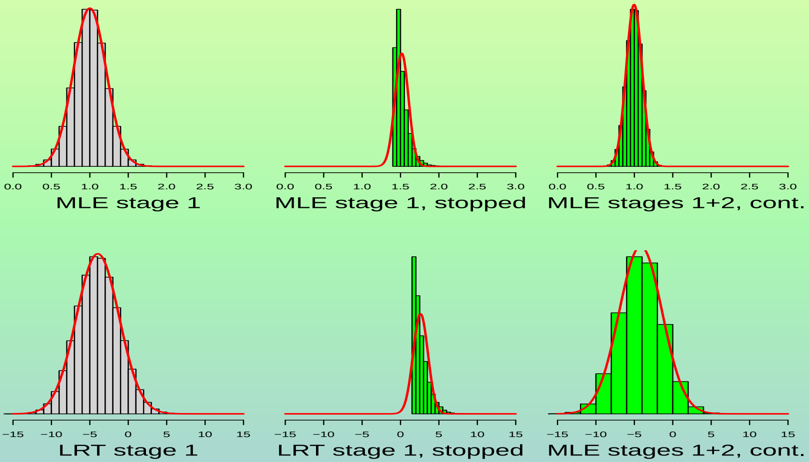


Figure 2: Under $H_0 : \theta = \theta_0 = 1$

Densities of Cauchy MLE(LRT) with Early Stopping

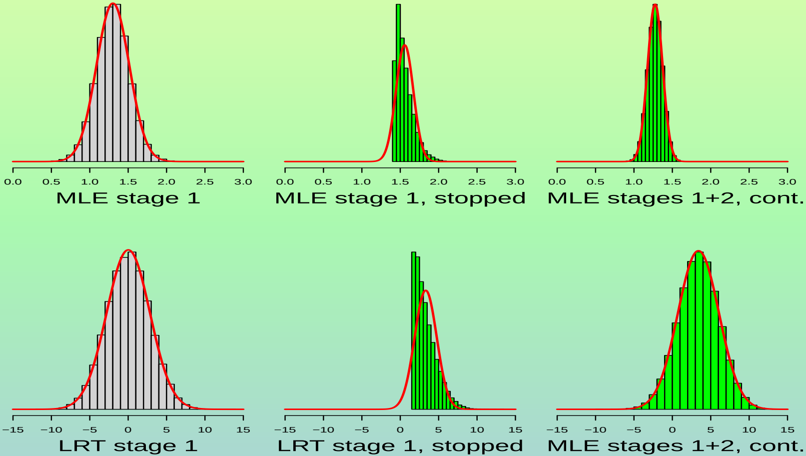


Figure 3: Under $H_1 : \theta = \theta_0 = 1.3$

Densities of Cauchy MLE(LRT) with Early Stopping

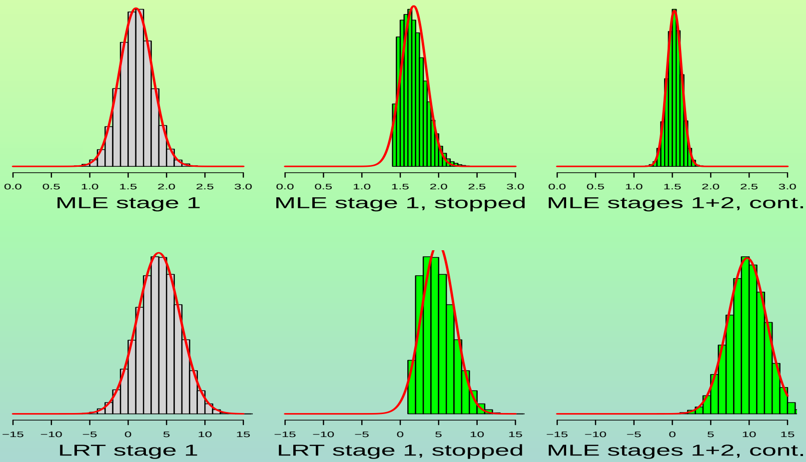


Figure 4: Under $H_2 : \theta = \theta_0 = 1.6$

Operational characteristics (a Cauchy example)

Then, $\alpha_1 = \alpha_2 = 0.0253$ secures the overall Type 1 error of 5%. These error rates can be obtained with critical values $c_1 \approx 1.4236$ and $c_2 \approx 1.3060$ ($c_1^{lrt} \approx 1.5357$ and $c_2^{lrt} \approx 4.2047$) and sample sizes $n_1 = 46$ and $n_2 = 138$.

θ	$\Pr\left(\hat{\theta}_1 > c_1\right)$	$\Pr\left(\hat{\theta}_{(2)} > c_2\right)$	$\Pr\left(l_1(\mathbf{X}_1) > c_1^{lrt}\right)$	$\Pr\left(l_2(\mathbf{X}_{(2)}) > c_2^{lrt}\right)$
1.0	0.0253	**0.0500**	0.0253	**0.0500**
1.3	0.2783	**0.7977**	0.2918	**0.8024**
1.6	**0.7997**	0.9998	**0.8064**	0.9998

Table 1: Cumulative rejection probabilities of the MLE-based and LR tests, $X_i \sim Cauchy(\theta, 1)$, ($n_1 = 46, n_2 = 138$); based on a Monte-Carlo experiments with 10^6 repeats each.

Functional Form of the Likelihood Does Not Change

The log-likelihood function is conditional on observed $\boldsymbol{X}_1 = \boldsymbol{x}_1$ and $\boldsymbol{X}_2 = \boldsymbol{x}_2$, and consequently on the stopping stage d:

$$\log L(\theta|\boldsymbol{X}_1 = \boldsymbol{x}_1, \boldsymbol{X}_2 = \boldsymbol{x}_2, D = d) = \begin{cases} \log L(\theta|\boldsymbol{X}_1 = \boldsymbol{x}_1) & \text{if } D = 1, \\ \log L(\theta|\boldsymbol{X}_1 = \boldsymbol{x}_1, \boldsymbol{X}_2 = \boldsymbol{x}_2) & \text{if } D = 2. \end{cases}$$

- the likelihood does not change
- the score function does not change
- the MLE does not change
- observed Fisher information does not change

Fisher Information Becomes Very Different

(Expected) Fisher Information Decomposition:

$$\mathcal{I}_{\mathbf{X}}(\theta) = \mathcal{I}_{D}(\theta) + \mathcal{I}_{\mathbf{X}|D}(\theta)$$

Information in the Design: $\mathcal{I}_D(\theta)$

- the **cost of informative stopping** is determined by $\mathcal{I}_D(\theta)$. The same *information deficit* was found in Molenberghs et al. 2014 .
- for non-informative stopping $\mathcal{I}_D(\theta) = 0$.
- the higher $\mathcal{I}_D(\theta)$ is the less information is left for estimation.

Information Conditional on the Design: $\mathcal{I}_{\mathbf{X}|D}(\theta)$

$$\begin{aligned}\mathcal{I}_{\mathbf{X}|D}(\theta) = & \Pr(D=1|\theta)\,\mathcal{I}_{\mathbf{X}_1|D=1}(\theta) \\ & + \Pr(D=2|\theta)\left[\mathcal{I}_{\mathbf{X}_1|D=2}(\theta) + \mathcal{I}_{\mathbf{X}_2}(\theta)\right]\end{aligned}$$

Information After Stopping: $\mathcal{I}_{\mathbf{X}|D=d}(\theta)$

- $\mathcal{I}_{\mathbf{X}_1|D=1}(\theta)$ and $\mathcal{I}_{\mathbf{X}_1|D=2}(\theta) + \mathcal{I}_{\mathbf{X}_2}(\theta)$.

Fisher Information in the experiment

$(D, \hat{\theta})$ - solid blue (total)
$\hat{\theta}|D$ - solid black (unconditional)
D - solid red (design)
$\hat{\theta}|D = 1$ - dashed black (conditional)
$\hat{\theta}|D = 2$ - dotted black (conditional)

$c_1 = 1.4236$ is shown by a thin dashed vertical line.

Fisher information in $C(\theta, 1)$ is 0.5.

In $n_1 + n_2 = 46 + 138$ i.i.d. $C(\theta, 1)$ observations, the Fisher information $= (46 + 138)/2 = 92$.

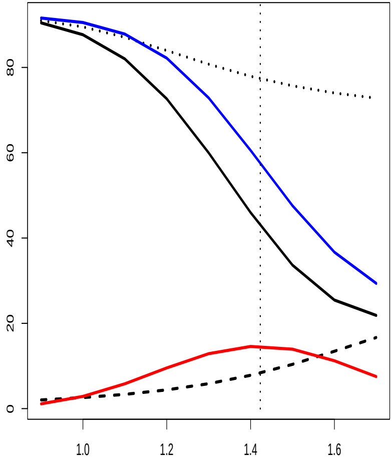


Figure 5: Fisher Information by θ

Some Remarks on Information Conditional on Design

1. The highest reduction of $\mathcal{I}_{\boldsymbol{X}}(\theta)$ happens when $\theta = c_1$.
2. Sample size re-estimation based on ancillary parameters does not reduce $\mathcal{I}_{\boldsymbol{X}|D}(\theta)$
   (e.g., SSR based on a re-estimated variance under a normal model).
3. Group sequential designs (also SPRT) are associated with a reduction in $\mathcal{I}_{\boldsymbol{X}}(\theta)$.

History of Lower Bound for Sequential Experiments

The Cramer Rao inequality goes back to work of Rao (1945) and Cramér 1946. Wolfowitz 1947 suggested a sequential version

$$\mathrm{E}\left([\tilde{\theta}-\theta]^2\right) \geq \frac{\left[1+\frac{\partial}{\partial\theta} b(\tilde{\theta}|\theta)\right]^2}{\mathrm{E}_D \mathcal{I}_{\boldsymbol{X}|\boldsymbol{D}=\boldsymbol{d}}(\theta)} + b^2(\tilde{\theta}|\theta),$$

where expected-over-the-decision-space Fisher information was used.

Simons 1980 showed in normal case that it is possible to have an estimator with a smaller variance than this lower bound claims.

A comprehensive review on sequential versions of CRLB is written by Ghosh and Parkayastha 2010.

Note: $\mathbf{E}_D$ is applied in the denominator!

Lower Bound for MSE in Two-Stage Experiments with Informative Stopping Options

Conditional on stopping stage $D = d$, the CRLB is

$$\mathrm{E}\left([\tilde{\theta}_{(d)} - \theta]^2\right) \geq \frac{\left[1 + \frac{\partial}{\partial\theta} b(\theta)\right]^2}{\mathcal{I}_{\tilde{\theta}|D=d}(\theta)} + b^2_{(d)}(\theta) \qquad (1)$$

and since $\mathrm{E}\left([\tilde{\theta} - \theta]^2\right) = P_1(\theta)\mathrm{E}\left([\tilde{\theta}_{(1)} - \theta]^2\right) + P_2(\theta)\mathrm{E}\left([\tilde{\theta}_{(2)} - \theta]^2\right)$, the lower bound for the MSE for an arbitrary estimator $\tilde{\theta}$ is

$$\mathrm{E}\left([\tilde{\theta} - \theta]^2\right) \geq \sum_{d=1}^{2} P_d(\theta)\left[\frac{\left[1 + \frac{\partial}{\partial\theta} b_{(d)}(\theta)\right]^2}{\mathcal{I}_{\tilde{\theta}|D=d}(\theta)} + b^2_{(d)}(\theta)\right], \qquad (2)$$

where $E\tilde{\theta}_{(d)} = \theta + b_{(d)}(\theta)$ and $P_d(\theta)$ is the stopping probability.

The MLE ($\hat{\theta}$) for a canonical parameter (θ) in an exponential family attains this lower bound (attains asymptotically for Cauchy).

The lower bound also depends on the bias and its derivative. A fair application of estimators should be conditioned on bias.

Two-Stage design: $X_i \sim C(\theta, 1)$, $n_1 = 46$, $n_2 = 138$.

Early stopping rule: Stage 1 Cauchy MLE $> c_1 = 1.4236$.

- **Black line (and circles)** is the Cauchy MLE;
- **Blue line** is the normal asymptotic approximation;
- Green line is the sample median.

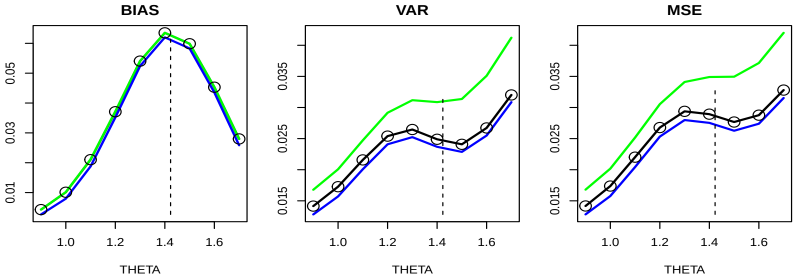
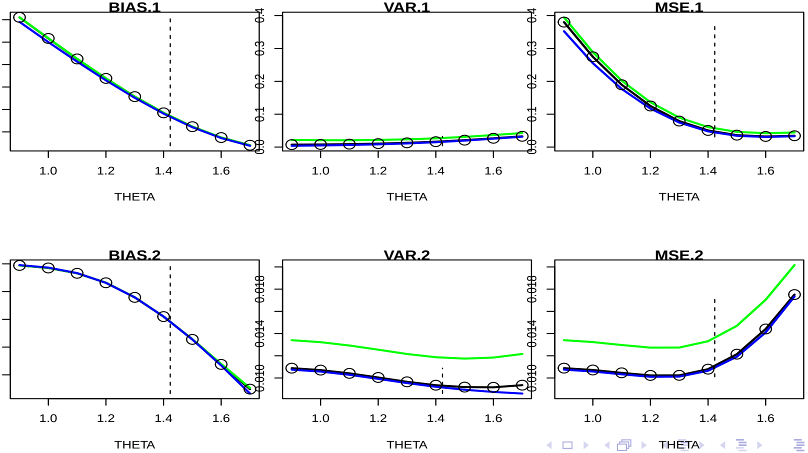


Figure 6: Asymptotic normal approximation of Cauchy MLE bias (1st figure), variance (2nd) and the MSE (3rd). The dashed line is c_1.

Two-Stage design: $X_i \sim C(\theta, 1)$, $n_1 = 46$, $n_2 = 138$.

Figure 7: Asymptotic Normal Approximation of stage specific bias (1st columns), variance (2nd) and mean squared error (3rd). Stage 1 is in the upper row; stage 2 is in the lower.



Summary (Fisher Information)

- Impact of informative interim adaptations on distributions is seen in group sequential designs, informative sample size re-estimation and in enrichment designs. See Flournoy and Tarima 2022; Tarima and Flournoy 2019, 2022 for details.
- Informative adaptations lead to information loss
  - Not all sampling information about θ is absorbed by the likelihood function;
  - From a known likelihood, you cannot reconstruct the design.
- When comparing two sequential procedures, in addition to the traditional averages *sample sizes*, *type 1 and 2 errors*, one may also consider *Fisher information*

Summary (Lower bound for MSE and variance)

- A new lower bound for variance and for the MSE is found
- The lower boundary for the MSE depends on the conditional biases and their derivatives. Thus, the lower bound only applies within such classes of estimators.
- There exist classes of estimators that share the same conditional bias (example - sample mean and sample median). An early stopping rule applied to unbiased estimators induces the same bias across all such estimators.

References

Cramér, Harald (1946). "A contribution to the theory of statistical estimation". In: *Scandinavian Actuarial Journal* 1946.1, pp. 85–94.

Flournoy, Nancy and Sergey Tarima (2022). "Discussion on "Adaptive enrichment designs with a continuous biomarker" by Nigel Stallard". In: *Biometrics*.

Ghosh, JK and Sumitra Parkayastha (2010). "Sequential cramer-rao and bhattachrya bounds: work of GR seth and afterwards". In:

Molenberghs, Geert et al. (2014). "On random sample size, ignorability, ancillarity, completeness, separability, and degeneracy: Sequential trials, random sample sizes, and missing data". In: *Statistical Methods in Medical Research* 23.1, pp. 11–41.

Simons, Gordon (1980). "Sequential estimators and the Cramér-Rao lower bound". In: *Journal of Statistical Planning and Inference* 4.1, pp. 67–74. ISSN: 0378-3758. DOI: https://doi.org/10.1016/0378-3758(80)90034-8. URL: https://www.sciencedirect.com/science/article/pii/0378375880900348.

References (cont.)

Tarima, Sergey and Nancy Flournoy (2019). “Asymptotic properties of maximum likelihood estimators with sample size recalculation”. In: *Statistical Papers* 60 (2), pp. 23–44. DOI: 10.1007/s00362-019-01095-x.

– (2022). “Most powerful test sequences with early stopping options”. In: *Metrika* 85.4, pp. 491–513.

Wolfowitz, Jacob (1947). “The efficiency of sequential estimates and Wald’s equation for sequential processes”. In: *The Annals of Mathematical Statistics* 18.2, pp. 215–230.