Inference After Adaptive Design

Nancy Flournoy
University of Missouri

flournoyn@missouri.edu

September 2019
4th Summer School on Data Science, Split

Outline

Classes of Inference Methods

- Parametric: Maximum Likelihood and Bayes
- Semi-parmemtric: Centered Isotonic Regression
- Non-parametric (not discussed here)

Maximum LIkelihood Estimation

The likelihood is the density function, interpreted as a function of the parameters with the data is considered fixed: $p(y|\theta) = \mathcal{L}(\theta|y)$.

Maximum Likelihood Estimates

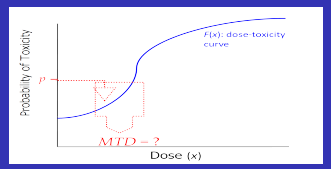
MLEs are the parameter values that maximize the likelihood function, MLEs are the most likely parameter values given the data.

$$\widehat{\theta} = \arg\max_{\theta \in \Theta} \mathcal{L}(\theta|y).$$

Example: Let Y be Bernouli$(1, p)$, $Y \in (0, 1)$:
Consider $\prod_{i=1}^{n} p_i^y (1-p)^{1-y_i}$ as a function of p versus y.

1. take logs: $\sum_{i=1}^{n}(1-y_i)\log(1-p) + \sum_{i=1}^{n} y_i \log p$.
2. set derivate to zero: $-\sum_{i=1}^{n} \frac{(1-y_i)}{(1-p)} + \frac{y_i}{p} = 0$.
3. solve to get $\widehat{p} = \frac{\sum_{i=1}^{n} y_i}{n}$.

Centered Isotonic Regression



$$N_m = \text{ Number of subjects allocated to dose } d_m;$$
$$T_m = \text{ Number of observed toxicities at dose } d_m;$$
$$R_m = T_m/N_m \text{ Observed toxicity rate at dose } d_m.$$

1. Compute the isotonic toxicity rate estimators;
   - $\widetilde{R}_m = R_m$ if the observed dose-specific toxicity rates $\{R_m\}$ are monotonic increasing.
   - If not, take sample-size weighted averages of violators.
2. Move estimators along the x-axis to sample-size weighted point between study doses.
3. Piece-wise linear interpolation between re-located estimators.

Oron, AP, Flournoy, N. (2017). Centered isotonic regression: Point and interval estimation for dose-response studies. *Statistics in Biopharmaceutical Research*

What are Adaptive Designs?

Experiments in which

1. the way data are generated and/or sampled (observed) changes with time and
2. the changes are determined, in whole or in part, by the prior data.

Examples of Adaptive Designs

- Changing treatment allocation probabilities in a randomized experiment
  - Balancing the number of subjects on each treatment
  - Increasing the proportion of subject to the better treatment
- Keeping a sequence of treatments on target, e.g., satellite tracking
- Sampling at or in the neighborhood of a target, e.g., greedy sampling, dose-finding designs, adaptive testing, on-line advertising.
- During a study,
  - Making decisions to stop the study
  - Deciding to change the number of subject to be studies
  - Deciding to add/drop treatment arms

Bias Induced by Adapting

Induced Bias Illustrated in the Context of Dose-Finding Designs

1. Bias Defined
2. Types of Dose-Finding Designs
3. Model-based versus Isotonic Regression Procedures
4. Bias in Observed Toxicity Rates.
5. Numerical Illustration
6. Mitigating the Bias.
7. Impact of Bias and Shrinkage on Estimators of Model Parameters and Target Dose
   - MLEs
   - Isotonic Regression Interval Estimates

What is Bias?

If the expected value of a statistic equals a parameter, it is said to be unbiased for that parameter.

Unbiased Example:
If $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$, $E\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n} E[X_i] = \mu$

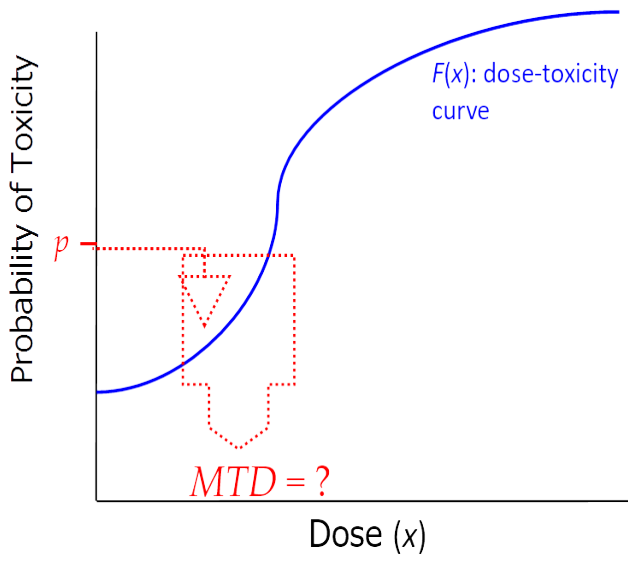
Bias Example: If the X-values are not observed directly, but some piece of equipment is out of calibration and $X_1 + \delta, \ldots, X_n + \delta$ are observed instead, then the observed random variables and their mean are biased.

In general

If θ is a parameter of interest and it is estimated by a statistic T, then

$$\text{bias} = E(T) - \theta$$

What is a Dose-Finding Design?



Context:
Binary Regression Following Sequential Informative Selection of Doses.

Use inverse estimation of a target quantile (e.g., Tsutakawa, 1980).
Alternatively,
consider dose-selection.

Dose-finding designs are risk adverse;
they want to avoid treating at high probabilities of toxicity.

Examples assume $P\{toxicity\}$ increases with dose.

Models for Dose-finding

1. Models of toxicity functions are either semi-parametric or parametric.
2. Even when designs are driven by semi-parametric models, parametric model may be employed for analysis at the end of a study.
   - in simulations to study the behavior of a design;
   - at the end of the study for inference;
3. When a parametric model of the response function assumes increase monotonicity, it is common to assume $\mathrm{P}(\mathrm{toxicity})$ follows a cumulative distribution function.
   - Range is from zero to one.
   - Common are location-scale models: $F[x_i - \alpha)/\beta]$
   - Examples: logistic, normal (probit)

Some Adaptive Dose-Allocation Procedures

Short Memory

- Classical Up-and-Down Design (Markov chain)
- Biased Coin Design (BCD, Markov chain)
- K-in-a-row (Geometric) (Krow, Markov chain)

Long Memory

- Continual Reassessment Method (CRM, Bayesian)
- EWOC (Bayesian design)
- Interval Designs (CCD Frequentist & mTPI Bayesian)
- Adaptive Optimal Design (AO, Frequentist & Bayesian)

Fundamental Challenges to Inference Following Dose-finding Designs Were Recognized Long Ago by Robbins (1954), Wetherill (1963) & others

1. Inverse Estimation Using Parametric Regression – MLE, LSE or Bayes

$\Gamma =$ the target toxicity rate
$\mu =$ the dose to be estimated, aka, the *target dose*.

A location-scale model $F[(x_i - \alpha)/\beta]$

$\rightarrow$ the estimated target is $\widehat{\mu} = \widehat{\beta} F^{-1}(\Gamma) + \widehat{\alpha}$.

1. $\min \mathrm{Var}[\widehat{\mu}] \Rightarrow$ allocate to the target
   but $\widehat{\alpha}$ and $\widehat{\beta}$ are both needed to estimate μ
2. $\min$ confidence ellipsoid around (α, β)
   $\Rightarrow$ allocate near the 20th and 80th quantiles

$\Rightarrow$ Concentrated allocations yield poor estimates of $\widehat{\beta}$.

- In small studies,
  - Bayes estimates will depend heavily on priors;
  - MLEs frequently do not exist.

2. Inverse Estimation Using Centered Isotonic Regression

CIREs=Observed toxicity rates if they are monotonic

Otherwise adjust using Pooled Adjacent Values Algorithm and interpolate between doses
(Oron & Flournoy, 2017).

Advantages

1. Eliminates need to estimate slope parameter;
2. Quality of estimate does not depend on sample sizes at doses far from the target.

3. Last Dose is the Estimated Target

Starts with Stochastic Approximation (Robbins, 1954); Tracking algorithms, . . . , Continual Reassessment Method.

Used with Allocation Methods that assign subjects to the model-estimated target $\widehat{\mu}$.

1. Such allocation sequences cannot be guaranteed to converge to the target μ (Azriel, et al., 2011) on the space of monotone response functions.
2. Tracking algorithms (with continuous responses) incorporate some randomness (with larger sample sizes) to insure convergence.
3. If a sequence happens to converge quickly to μ, one can't estimate $F'(\mu)$ or β.

   For this reason, Wetherill, I and others did not consider such methods further.

Some Notation To Talk About Bias:

n	Number of subjects in the study
$d_1 < \cdots < d_M$	Doses in the treatment space
$X(j)$	Dose for the jth subject, $j = 1, \ldots, n$
$\delta_m(j)$	$= \begin{cases} 1 & \text{if } X(j) = d_m \\ 0 & \text{if else.} \end{cases}$
$N_m = \sum_{j=1}^{n} \delta_m(j)$	Frequency of allocations to dose d_m;
$Y(j)$	$= \begin{cases} 1 & \text{if } j\text{th subject has toxicity;} \\ 0 & \text{if else.} \end{cases}$
F_m	$= P\{Y(j) = 1 \mid \delta_m(j) = 1\} \quad \forall n.$
$T_m = \sum_{j=1}^{n} Y(j)\delta_m(j)$	Frequency of toxicities at dose d_m;

Observed Toxicity Rates Are Fundamental Summary Statistics

Isotonic regression methods

Isotonic regression methods use observed toxicity rates directly.

Likelihood-based methods

Standard *likelihood-based methods* use observed toxicity rates indirectly: $\mathcal{L} = \prod_{m=1}^{M} F_m^{T_m}(1 - F_m)^{N_m - T_m}$.

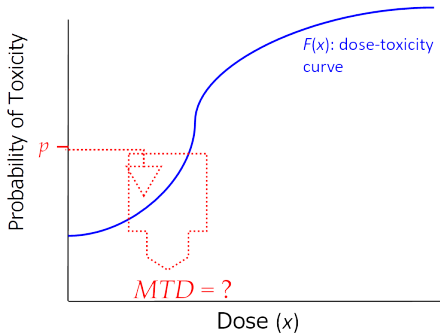
- MLE of F_m is a function of $\{T_m/N_m\}$;
- MLE of F_m is T_m/N_m when F_m is not a function of additional parameters.

Observed Toxicity Rates

$R_m = T_m/N_m, \qquad N_m \geq 1; m = 1, \ldots, M$

An erroneous presumption is widespread:

$$T_m|N_m \propto \text{Binomial}(N_m, F_m)$$



- Obtaining the Binomial distribution requires conditioning on the exact observed value of N_m (N_m is informative⇒loss of information).
- Probabilities under this conditioning are not the same as under the conditional distribution of $R_m|N_m$.

$\mathbf{E}[T_m|N_m] \neq F_m$

except in special circumstances, as we will show.

Dose-Specific Toxicity Rate: $F_m = \mathrm{E}[T_m]/\ \mathrm{E}[N_m] \neq \mathrm{E}[T_m/N_m]$

$N_m = \sum_{j=1}^{N} \delta_m(j)$ Number of allocations to dose d_m;

$T_m = \sum_{j=1}^{N} Y(j)\delta_m(j)$ Number of toxicities at dose d_m.

$$\mathrm{E}[T_m] \overset{definition}{=} \sum_{j=1}^{N} \mathrm{E}[Y(j)\delta_m(j)]$$

$$= \sum_{j=1}^{N} P\{Y(j)|\delta_m(j) = 1\}P\{\delta_m(j) = 1\}$$

$$\overset{definition}{=} F_m \sum_{j=1}^{N} P\{\delta_m(j) = 1\} = F_m\, \mathrm{E}[N_m].$$

$$\Rightarrow F_m = \mathrm{E}[T_m]/\ \mathrm{E}[N_m].$$

Amazing Insight Comes Straight from the Definition of Covariance

$R_m \equiv T_m / N_m$.

$$\begin{aligned}
\mathrm{Cov}[R_m, N_m] &\overset{\text{definition}}{=} \mathrm{E}[(R_m - \mathrm{E}[R_m])(N_m - \mathrm{E}[N_m])] \\
&= \mathrm{E}[R_m N_m] - \mathrm{E}[R_m]\,\mathrm{E}[N_m] \\
&\overset{\text{definition}}{=} \mathrm{E}\left[\frac{T_m}{N_m} N_m\right] - \mathrm{E}[R_m]\,\mathrm{E}[N_m] \\
&= \mathrm{E}[T_m] - \mathrm{E}[R_m]\,\mathrm{E}[N_m].
\end{aligned}$$

$$\begin{aligned}
\Rightarrow \mathrm{E}[R_m] &= \frac{\mathrm{E}[T_m]}{\mathrm{E}[N_m]} - \frac{\mathrm{Cov}[R_m, N_m]}{\mathrm{E}[N_m]} \\
&\overset{\text{last slide}}{=} F_m - \frac{\mathrm{Cov}[R_m, N_m]}{\mathrm{E}[N_m]}.
\end{aligned}$$

Having allocations depend on outcomes induces bias in using observed toxicity rate to estimate probability of toxicity.

Bias: $\mathrm{E}[R_m] - F_m = \frac{\mathrm{Cov}[R_m, N_m]}{\mathrm{E}[N_m]}$

1. $\mathrm{E}[R_m] - F_m \in [-1, 1]$.
2. Because correlations are in [-1,1],

$$|\mathrm{E}[R_m] - F_m| \leq \frac{\sqrt{\mathrm{Var}\,[R_m]}\sqrt{\mathrm{Var}\,[N_m]}}{\mathrm{E}[N_m]} \leq \frac{1}{2}\mathrm{CV}[N_m],$$

3. $$\mathrm{Cov}[R_m, N_m] \begin{cases} > 0 & \text{iff } \mathrm{E}[R_m] - F_m < 0; \\ = 0 & \text{iff } \mathrm{E}[R_m] - F_m = 0; \\ < 0 & \text{iff } \mathrm{E}[R_m] - F_m > 0. \end{cases}$$

4. If $\{N_m\}$ are unimodal around a dose $\mu = F^{-1}(\Gamma)$, and $\{R_m\}$ are increasing with values that bound Γ,

$$\mathrm{Cov}[R_m, N_m] \begin{cases} > 0 & \text{for doses } < \mu; \\ < 0 & \text{for doses } > \mu. \end{cases}$$

Illustration

Consider Logistic Dose-Response Function

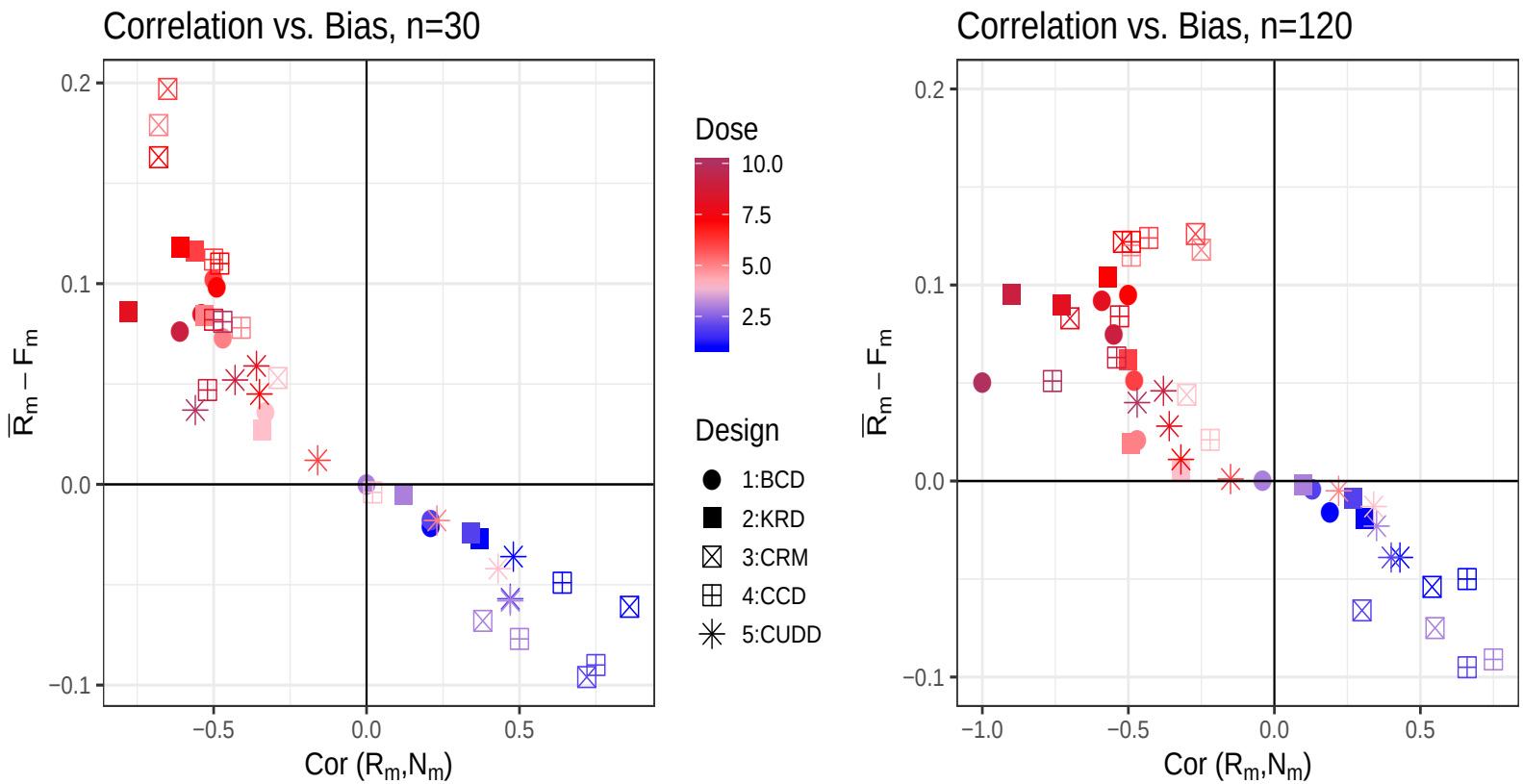
$logit[F(x)] = (x - 5.6)/2$.

- $\mu = F^{-1}(0.3) = LD30 = 3.9$
- $\alpha = F^{-1}(0.5) = LD50 = 5.6$
- "slope" $= 0.5$

Design Space

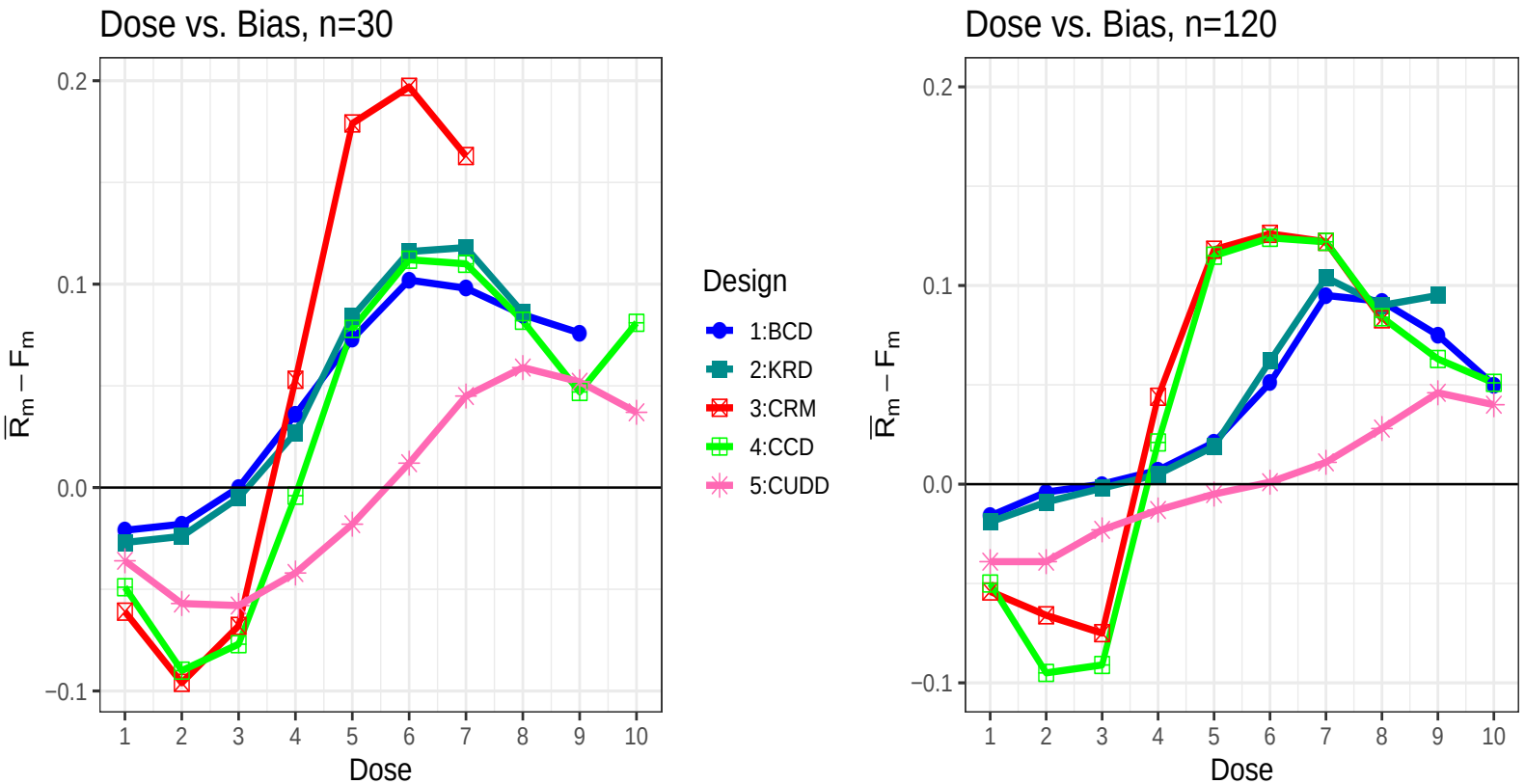
- $x \in \{1, \ldots, 10\}$

Toxicity Rate Bias $(\bar{R}_m - F_m)$ by $\rho(R_m, N_m)$



Data is averaged across doses.

Bias of Observed Toxicity Rate ($\bar{R}_m - F_m$) by Dose. LD30=3.9; LD50=5.6.



Shrinking the Bias in Observed Toxicity Rates – Inspired by Firth (1993).

Make use of our observation that the bias "flares out" away from target by slightly shrinking all R_m's towards target Γ:

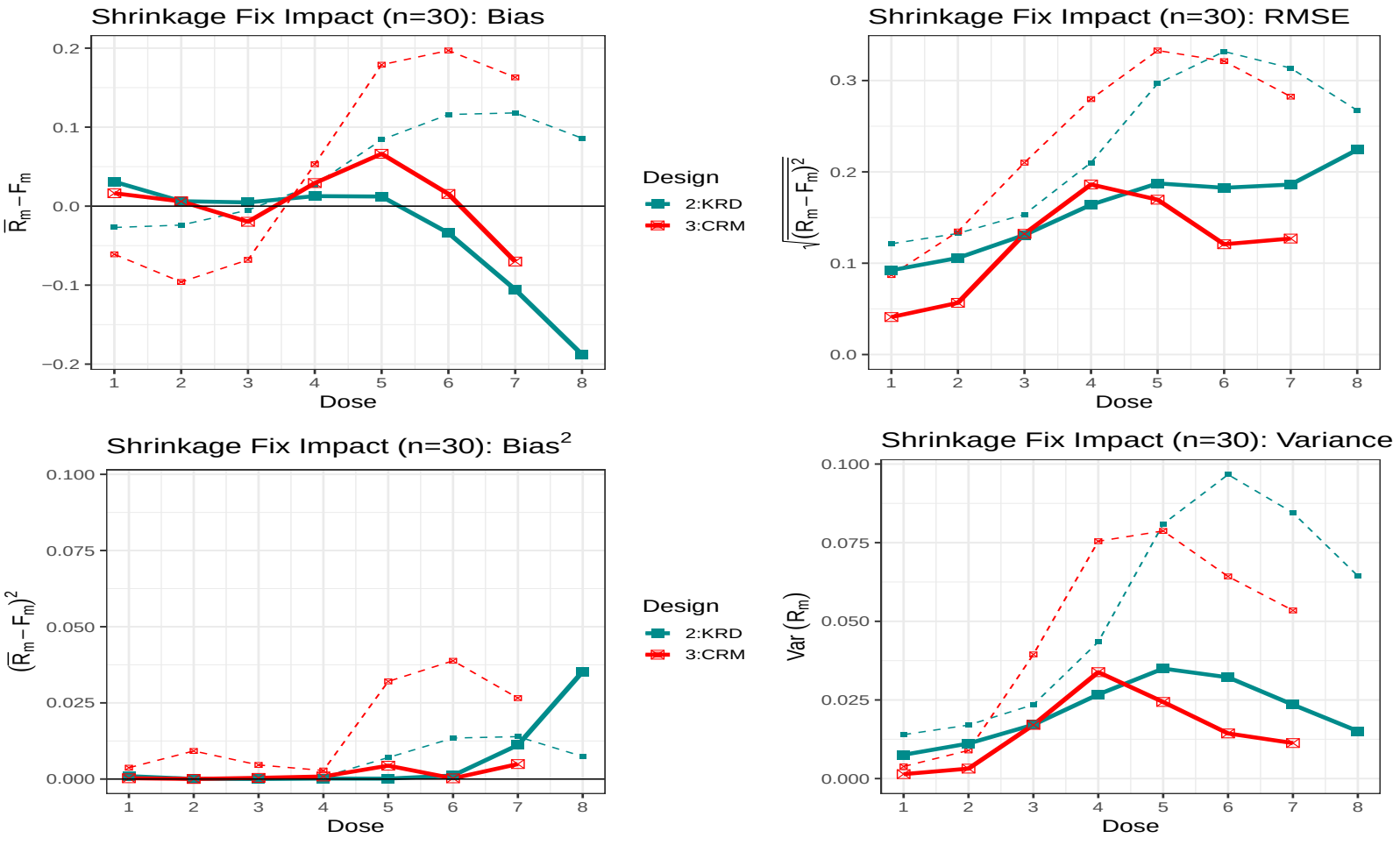
$$\widetilde{R}_m = \frac{T_m + \Gamma}{N_m + 1} = \frac{N_m R_m + \Gamma}{N_m + 1}.$$

The magnitude of shrinkage is approximately inverse to N_m, in agreement with the bias expression.

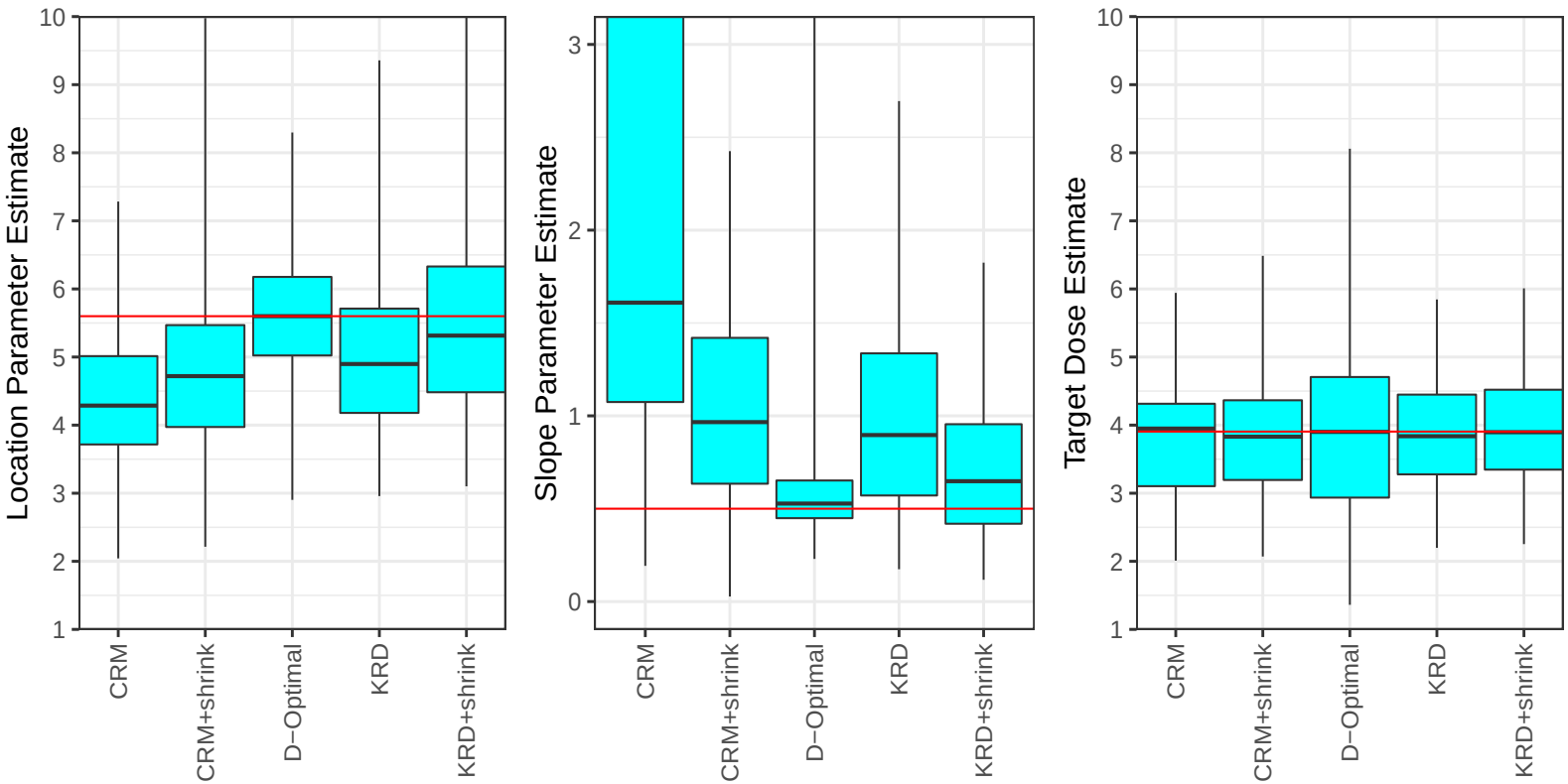
When $\Gamma = 0.5$

This formula is identical to the commonly used correction for calculating the empirical logit in the presence of zero cell counts [Woolf (1955), Anscombe (1956)].

Impact of Shrinking the Bias of Observed Toxicity Rates by Dose. (Target = LD30=3.9).



Impact of Bias and Shrinkage on Logistic Parametric Estimates.



Bias with Independent Groups

Adaptive randomization

Increase assignments to treatments showing less toxicity: negative correlation between sample sizes and toxicity
$\Rightarrow$ positive bias

With Continuous RVs,

Re-define R_m = observed group-specific sample mean and T_m = observed group-specific sum.
Increase P(sampling from high performing groups)
$\Rightarrow$ negative bias
e.g.,

- Greedy sampling,
- Thomson Sampling,
- Randomly Reinforced Urns

Warning

Don't put faith in estimates of F except in a very close neighborhood of the target.

Current methods for dropping doses with high observed toxicity rates are more aggressive than they are thought to be.

Introduction to SSR: What Does Sample Size Calculations Do?

A sample is a subset of a population.
The sample size is the number of participants in a study.

- An appropriate sample size is required for quantifying the validity of an estimate or a **hypothesis test**.
- If the sample size it too small, the study has a low probability of finding existing effects.
- The results from a study with a small sample size will be questionable.
- A sample size that is too large will result in wasting money and time.
- If the subjects are humans or animals, it is unethical to choose too large a sample size.

The Language of Basic Hypothesis Testing

- The **Null Hypothesis**, denoted by H_0, is a status quo or prevailing viewpoint about a population.
- The **Alternative Hypothesis**, denoted by H_A, is an alternative to the null hypothesis - the change in the population that researchers hope is true.

The Associated Press (September 7, 1995)

Taking an aspirin every other day for 20 years can cut your risk of colon cancer nearly in half, a study suggests.
However, the benefits may not kick in until at least a decade of use.
The lifetime risk of developing cancer is 1 in 16 according to the American Cancer Society.

H_0 : Aspirin will not change the 1 in 16 risk of getting colon cancer.

H_A : Aspirin will reduce the risk to less than 1 in 16.

Errors that Can Be Made

The **test statistic** is the summary of the data that will be used to decide between H_0 and H_A.

Your Decision Based on the Data	The Truth is that H_0 is true	The Truth is that H_A is true
H_0 is supported	No error	Type II Error
H_A is supported	Type I Error	No Error

α=**level of significance**=the accepted chance of a Type I error occurring (usually 5%).

β=the accepted chance of a Type II error occurring
= the chance of accepting H_0 when H_A is true.

Power $= 1 - \beta$ = the chance of deciding H_A is true when it is (usually desired to be at least 80%).

Decision Rules

- A **Decision Rule** is a formal statement of when, based on the data obtained, to reject H_0.

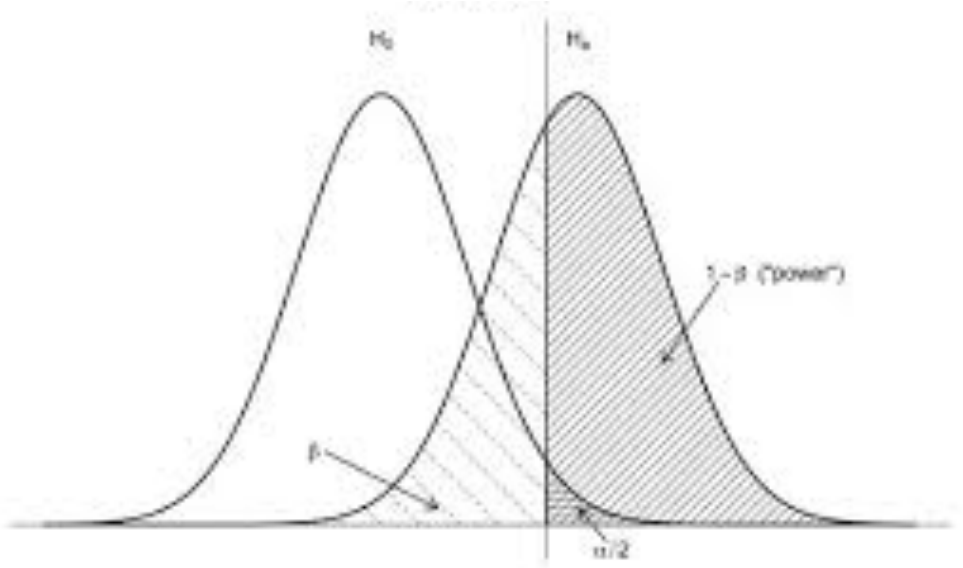
  It generally specifies a set of values, based on the data to be collected, which are contradictory to H_0 and favor the alternative hypothesis H_A.
- The **Direction of the Extreme** corresponds to the position of the values which are more likely under H_A than under H_0.

  If the larger values are more likely under H_A then the direction of the extreme is said to be **to the right**.
- The ***p*-value** is the chance, computed under the assumption that H_0 is true, of getting the observed test statistic plus the chance of getting any more extreme values (which show even less support for H_A).

Decision Rule Terminology

- A **rejection region** is a set of values for which you would reject H_0.
  They are contradictory to H_0 and favor H_A.
- An **acceptance region** is the set of values for which you would accept H_0.
- The **cut-off value** or **critical value** is the value that marks the start of the rejection region.
- The **Minimal Relevant Effect** is the minimum distance of the parameter of interest under H_0 and H_A that has power guaranteed at a specific level (usually 80%.)

The Critical Value and Power for Normal Data



Example: Finding the Critical Value with Normal Data

1. Let $\{Y_1, \ldots, Y_n\} \sim \mathcal{N}(\mu, \sigma^2) \Rightarrow \bar{Y} \sim \mathcal{N}(\mu, \sigma^2/\sqrt{n})$.
2. Test statistic $Z = \frac{(\bar{y}-\mu_0)}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1)$ if $H_0 : \mu = \mu_0$ true.
3. Test statistic $Z = \frac{(\bar{y}-\mu_A)}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1)$ if $H_A : \mu = \mu_A > 0$ true.
   The rejection region is **one-sided** and the direction of the extreme is to the right.
4. $\alpha = P(\bar{Y} \geq c|H_0) = 1 - P(\bar{Y} \leq c|H_0)$
   $= 1 - P\left(\frac{\bar{Y}-\mu_0}{\sigma/\sqrt{n}} \leq \frac{c-\mu_0}{\sigma/\sqrt{n}}|H_0\right) = 1 - \left(\frac{c-\mu_0}{\sigma/\sqrt{n}}\right)$th percentile of the standard normal distribution: $z_{1-\alpha/2} = \frac{c-\mu_0}{\sigma/\sqrt{n}}$
   $\Rightarrow c = \mu_0 + z_{1-\alpha/2}\sigma$

Normal Example cont.: Power

$$
\begin{aligned}
\textit{Power} &= P(\bar{Y} \geq c | H_A) = P(\bar{Y} \geq \mu_0 + z_{1-\alpha/2}\sigma/\sqrt{n} | H_A) \\
&= 1 - P(\bar{Y} \leq \mu_0 + z_{1-\alpha/2}\sigma/\sqrt{n} | H_A) \\
&= 1 - P\left(\frac{\bar{Y} - \mu_A}{\sigma/\sqrt{n}} \leq \frac{\mu_0 + z_{1-\alpha/2}\sigma/\sqrt{n} - \mu_A}{\sigma/\sqrt{n}} | H_A \right) \\
&= 1 - P\left(\frac{\bar{Y} - \mu_A}{\sigma/\sqrt{n}} \leq \frac{\mu_0 - \mu_A}{\sigma/\sqrt{n}} + z_{1-\alpha/2} | H_A \right) \\
&= \left(\frac{\mu_A - \mu_0}{\sigma/\sqrt{n}} - z_{1-\alpha/2} \right) \text{ th percentile of } \mathcal{N}(0,1).
\end{aligned}
$$

Set percentile equal to $z_{1-\beta}$ and solve for n:

$$
n = \frac{\sigma^2 (z_{1-\alpha/2} + z_{1-\beta})^2}{(\mu_0 - \mu_A)^2} = \frac{\sigma^2 c^2}{(\mu_0 - \mu_A)^2}
$$

The Problem with Estimating the Variance

Sample size formula depends on σ^2 which is unknown.

If an *assumed* value of σ is

smaller than the true value,

the calculated sample size will be smaller than it should be which results in under-powered studies.

higher than the true value,

the calculated sample size will be higher than it should be, which results in over-powered studies that are more expensive than needed.

Possible help comes from **sample size re-calculation (SSR)** in the middle of the study.

Complications in Re-estimating the Sample Size

When statistics are *asymptotically normal*,
the variance of the parameter of interest may or may not depend on its expected value,
in which case the distribution of the parameter estimate is not longer purely normal.

Some want to update estimate of minimal effect size
$|\mu_0 - \mu_A|$
as well as the variance of the parameter of interest σ_θ.

Options for Sample Size Re-estimation

Let $\widehat{\theta}_1$ denote an interim estimate of θ at the time of the interim analysis.

Recall $\text{Var}(\bar{Y}) = \frac{\sigma^2}{n}$. Analogously $\widehat{\text{Var}}\left(\widehat{\theta}_1\right) = \frac{\widehat{\sigma}^2}{n}$.

Options

SSR N0: variance known

SSR N1: estimate the variance: $\widehat{\sigma}^2_\theta := n\,\widehat{\text{Var}}\left(\widehat{\theta}_1\right)$.

SSR N2: estimate both the variance and the effect size: $\widehat{\sigma}^2_\theta := n\,\widehat{\text{Var}}\left(\widehat{\theta}_1\right)$ and $\widehat{\mu}_A - \widehat{\mu}_0$.

Hypotheses involve only the mean.

Variance is called a nuisance parameter.

SSR N1 leaves estimates of the mean and variance independent.

Consider Study Statistics Based on Asymptotic Normal Approximations

Study design and analyses frequently rely on assumption

$$\sqrt{n}\left(\widehat{\theta}-\theta\right) \xrightarrow{d} N(0,\sigma_\theta) \qquad \text{as } n \to \infty.$$

Now one wants to use this approximation at an interim time point to re-estimate the sample size.

This implies that $n = N_0$ is the finite number of subjects at the time of the interim analysis.

But n is going to infinity.

This presents a **conundrum** for sample size re-estimation

Sample Size Re-estimation When Using Asymptotic Approximations

How does the n in the asymptotic approximation relate to the sample size N_0 required to obtain 80% power?

The sample size required to test $H_0 : \theta = 0$ at the α % significance level with $100(1-\beta)$ Power

$$N_0 = \frac{\sigma_\theta^2 c^2}{\theta_A^2},$$

$c^2 = \left(z_{1-\alpha/2} + z_{1-\beta}\right)^2$
$z_q = q$-th quantile of the $\mathcal{N}(0,1)$ distribution.

- If $n = N_0 < \infty$, there is 80% power to detect H_A.
- But if $n \to \infty$, power (typically) $\to 1$
  $\Rightarrow$ If $n \to \infty$ drags N_0 along with it, all studies would stop at interim analysis.

This leads to testing at *local alternatives.*

Local and Fixed Alternative Hypotheses

Fixed alternatives: $H_0 : \{\theta = 0$; $H_A : \theta = \theta_A \neq 0$; θ_A is pre-determined.

A *finite* sample size (which cannot go to infinity) is needed to secure α and $1 - \beta$:

$$N_0 = \frac{\sigma_\theta^2 c^2}{\theta_A^2}.$$

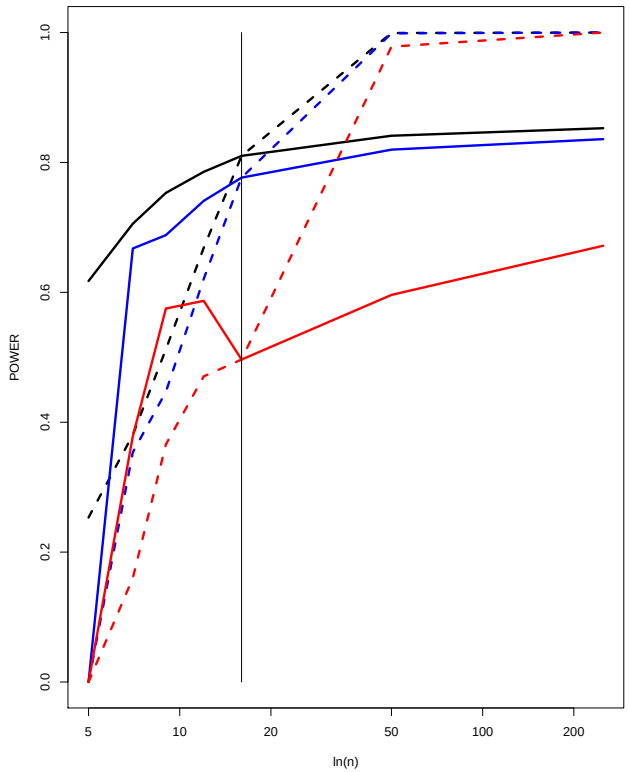
Local alternatives: $H_0 : \{\theta = 0$; $H_A : \theta = h/\sqrt{n} \neq 0$; h is pre-determined.

$$N_0 = n\frac{\sigma_\theta^2 c^2}{h^2}.$$

Note $\sigma_\theta^2 := n\text{Var}\left(\widehat{\theta}_1\right)$ depends on the sequence $n \to \infty$.

Power by $\ln(n)$

$H_0 : N(0, 1)$
H_A fixed alternative: $N(3/4, 1)$ [dotted]
H_A local alternatives: $N(3/\sqrt{n}, 1)$ [solid] .



Power by Log(n)

Black: One sample t-tests
Blue: Wilcoxon signed tests
Red: Sign tests
(non-monotonicity due to discreteness).

Power curves for each test cross at
$n = 16$ ($\theta = 3/4$).

We consider local alternatives (aka Pitman drift).

SSR N_1: Re-estimate the Variance without Changing the Prior Effect Size

The reassessed sample size

$$N_1 := \left[N_0 = n\frac{\sigma_\theta^2 c^2}{h^2} \right]_{\sigma_\theta^2 = n\,\widehat{\mathrm{Var}}(\hat{\theta}_1)}$$

$$= n\,\widehat{\mathrm{Var}}\left(\hat{\theta}_1\right)\frac{c^2}{\theta^2} = n^2\,\widehat{\mathrm{Var}}\left(\hat{\theta}_1\right)\left(\frac{c}{h}\right)^2$$

With local alternatives, $\frac{N_1}{n}$ has a constant limit which is the same as when the variance is known:

$$\frac{N_1}{n} \to \frac{N_0}{n} = \sigma_\theta^2\left(\frac{c}{h}\right)^2.$$

SSR N_1: Only Variance is Re-estimated

Summary : If SSR depends only on nuisance parameters, (local) asymptotic properties are not affected.
See P-values from Shapiro-Wilks test for normality.

n	Average N_1	θ $h/\sqrt{n}$	Ave $\bar{X}$	Power	SW P-value
10	83	0.316	0.317	0.741	$<$0.001
100	769	0.100	0.100	0.793	0.672
10	36	0	0.000	0.054	$<$0.001
100	1003	0	0.001	0.054	0.251

Monte-Carlo Simulated $\mathcal{N}(h/\sqrt{n},\ 1)$ Observations with SSR N_1; $h = 1$.

SSR N_2: Re-estimate Variance and Effect Size Relying on Asymptotic Normality

$$N_2 = \left[N_0 = n\frac{\sigma_\theta^2 c^2}{h^2} \right]_{\sigma_\theta^2 = n\,\widehat{\mathrm{Var}}(\hat{\theta}_1),\ \theta = \hat{\theta}_1}$$

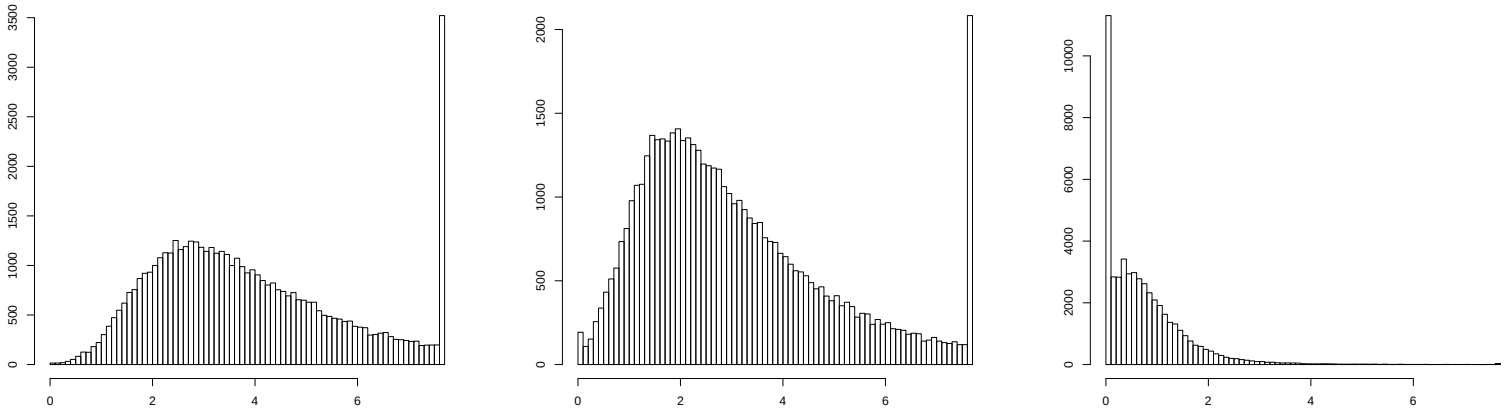
$$= n\,\widehat{\mathrm{Var}}\left(\hat{\theta}_1\right)\left(\frac{c^2}{\hat{\theta}_1^2}\right) = n^2\,\widehat{\mathrm{Var}}\left(\hat{\theta}_1\right)\left(\frac{c}{\hat{h}}\right)^2 = n^2\,\widehat{CV}^2 c^2.$$

CV $:= \sigma_\theta/h$ is a *local coefficient of variation*.
The global version, $\sigma_\theta/\theta = \sqrt{n}\,\mathrm{CV}$, diverges to infinity.

$$\frac{N_2}{n} = \frac{\widehat{\mathrm{Var}}(\hat{\theta}_1)}{(\hat{\theta}_1)^2}\,c^2 = \frac{(n-1)\widehat{\mathrm{Var}}(\hat{\theta}_1)/\sigma_\theta^2}{(n-1)(\hat{\theta}_1)^2/\sigma_\theta^2}\,c^2 \xrightarrow{d} \frac{c^2}{\chi_1^2(h^2/\sigma_\theta^2)},$$

where $\chi_1^2(h^2/\sigma_\theta^2)$ is a χ^2 **random variable** with 1 degree of freedom and non-centrality parameter $h^2/\sigma_\theta^2 = CV^{-2}$.

Histograms: $N_2 \longleftarrow \min\left[\max\left(n, N_2\right), n_{max}\right]$
$ln(N_2/n) =$ LN[Total to Interim SS Ratio]



(a) Under $H_0 : h = 0$ (b) Under $H_A; h = 1$ (c) If true $h = 3$

N_2 calculated to have $\alpha = 0.05$ at $H_0 : h = 0$;
and power $= 0.80$ at $H_A : h = 1$;
$n = 100$, $n_{max} = 2000$, $\sigma_\theta = 1$.

SSR N_2 depends on θ and σ_θ^2 Being Re-estimated after n Observations

Summary : If SSR depends on a re-estimated treatment effect, (local) asymptotic properties DO change.

n	Average N_2	θ $h/\sqrt{n}$	Ave $\bar{X}$	Power	SW P-value
10	35	0.316	0.393	0.785	<0.001
100	414	0.100	0.122	0.809	<0.001
1000	16944	0.032	0.039	0.810	<0.001
10	89	0	0.001	0.092	<0.001
100	2879	0	0.000	0.085	<0.001
1000	112573	0	0.000	0.084	<0.001

Table: Monte-Carlo Simulated $\mathcal{N}(h/\sqrt{n},\ 1)$ Observations with SSR N_2; $h = 1$.

$N \longleftarrow \min\left[\max\left(n, N\right), n_{max}\right]$
$n =$ sample size in asymptotic approx.

Let N be **any** SSR that has

$$\tau_n := \frac{N}{(N-n)} \xrightarrow{d} \tau \quad \text{as } n \to \infty,$$

where τ a random variable on $(-\infty, \infty)$.

τ_n (through N) may depend on both interim statistics $\widehat{\theta}_1$ and $\widehat{\sigma}_\theta$;

$\widetilde{\theta} :=$ full data MLE.

$V := \sqrt{N}\left(\frac{\widetilde{\theta}-\theta}{\sigma_\theta}\right) = \sqrt{N}\left(\frac{\widetilde{\theta}-h/\sqrt{n}}{\sigma_\theta}\right)$

Theorem

If $\tau_n = N/(N-n) \xrightarrow{d} \tau$, *then under* $\theta = h/\sqrt{n}$, *the asymptotic distribution of the standardized MLE does not depend on n:*

$$Pr\,(V < v) \to Pr\left(\frac{N}{n} \leq 1\right)\ \Phi\,(v|\,\tau < 1)$$

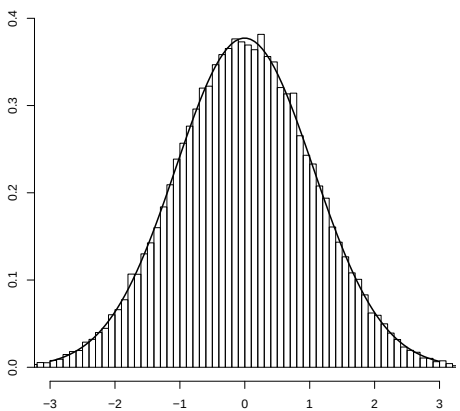
$$+\,Pr\left(\frac{N}{n} > 1\right) \times \int_{\tau \geq 1} \Phi\left(\sqrt{\tau}v - \sqrt{\tau-1}y \middle| \tau \geq 1\right)\phi\,(y)\,dy,$$

where $\Phi\,(v|\,A) = \int_{-\infty}^{v} I(y \in A)\phi\,(y)\,dy/\Phi(A)$ and $\tau = \tau(y)$.

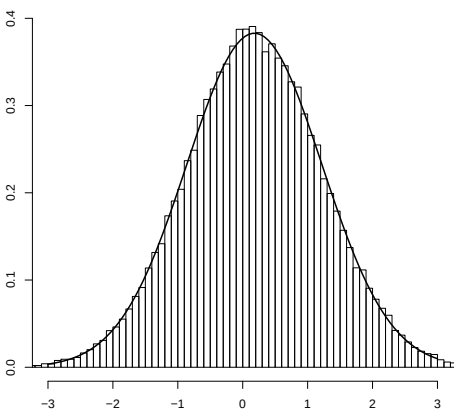
$\Phi(\cdot)$ is standard normal cdf.

Distribution of standardized MLE V.

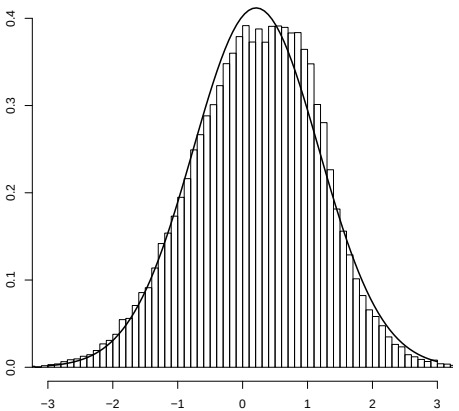
Using SSR N_2 formula assuming
$H_0 : h = 0$, $\alpha = 0.05$, $\beta = 0.80$ at $H_A : h = 1$;
$n = 100$, $n_{max} = 2000$, $\sigma_\theta = 1$.



(d) Under $H_0 : h = 0$



(e) Under $H_A : h = 1$



(f) If true $h = 3$

With General Sample Size Formula, Standardized MLE Distribution is Free of Model Parameters

CV $:= \sigma_\theta / h$ is a *local coefficient of variation*.

Corollary

Under $H_0 : \theta = h = 0$,
when a sample size formulae can be expressed as
$N = n\, g(\widehat{CV}^2)$ *with* $g(\cdot)$ *a positive nonrandom function,*
the distribution of V *does not depend on model parameters.*

Example:

$$\frac{N_2}{n} \overset{d}{\approx} \frac{c^2}{\chi^2_1(h^2/\sigma^2_\theta)} \overset{d}{\approx} \frac{c^2}{\chi^2_1\left(CV^{-2}\right)}.$$

$$\text{Under } H_0, \quad \frac{N_2}{n} \overset{d}{=} \frac{c^2}{\chi^2_1}.$$

Pocock's One-Sided Two-Group Sequential Z-test

Pocock's design is a simple two-stage study design for testing $H_0 : \theta = 0$ versus $H_0 : \theta = \theta_1$.

1. Assume $X_i \sim \mathcal{N}(\theta, 1)$.
2. n_1 =number of subjects before interim analysis;
   n_2 =number of additional subjects if study continues.
   We take $n_1 = n_2 = 100$.
3. Chose a critical value c_1 for continuing or stopping the study.
   We use 2.18 to secure overall type 1 error rate $\alpha = 0.025$.
4. After n observations, calculate $Z_1 = \sqrt{n_1}(\bar{X} - 0)/1$.
5. If $Z_1 > c_1$, stop with n_1 observations.
   If $Z_1 \leq c_1$, n_2 more are observed.

Some Notation

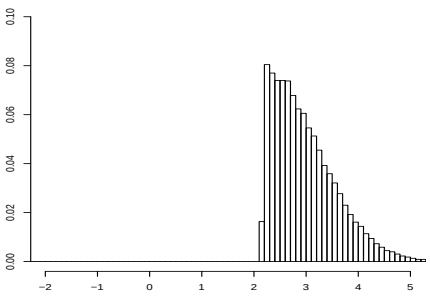
Starting from joint density for whole experiment

- $\widehat{\theta}_{(j)}$ = MLE from the from cumulative data through stage j.
  - $\widehat{\theta}_{(1)}$ = MLE based on data available if study stops at interim analysis.
  - $\widehat{\theta}_{(2)}$ = MLE based on data available if study continues to the end.

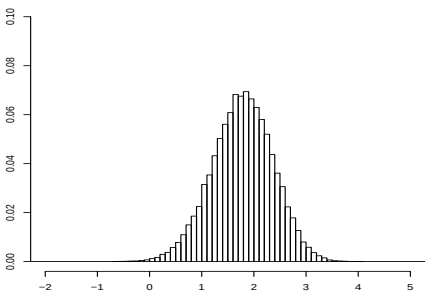
Starting from density conditional on stopping at stage j

- $\widehat{\theta}^{c}_{(j)}$ = MLE from using cumulative data through stage j.
  - $\widehat{\theta}^{c}_{(1)}$ = MLE based on data available if study stops at interim analysis.
  - $\widehat{\theta}^{c}_{(2)}$ = MLE based on data available if study continues to the end.

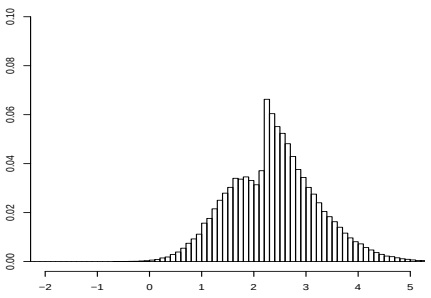
Histograms of MLE pivots from Pocock's Design Example



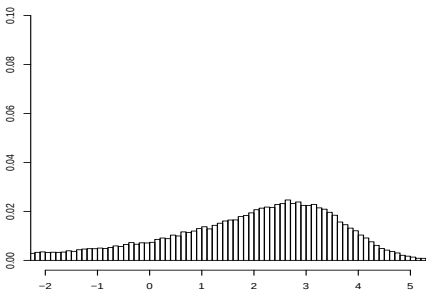
(g) $\hat{\theta}_{(1)}$



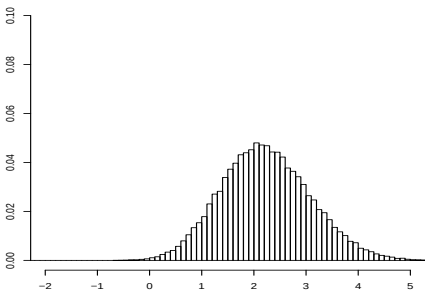
(h) $\hat{\theta}_{(2)}$



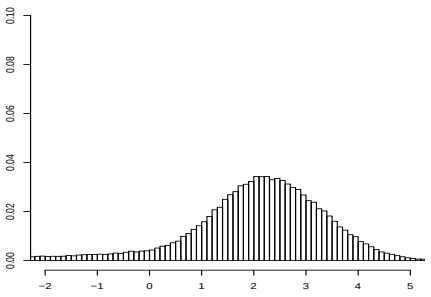
(i) $\hat{\theta}_{(1)}$ & $\hat{\theta}_{(2)}$ Mixture



(j) $\hat{\theta}^c_{(1)}$



(k) $\hat{\theta}^c_{(2)}$



(l) $\hat{\theta}^c_{(1)}$ & $\hat{\theta}^c_{(2)}$ Mixture

Sample Size Summary

- Interim sample size recalculations based on unknown nuisance parameters do not change MLE distribution under local asymptotics.
- Interim sample size recalculations based on re-estimation of the treatment effect do change MLE distributions under local asymptotics.
- Local asymptotics is a more adequate way to analyze large sample properties for hypothesis testing problems.
- Prior to study, one cannot condition on the stopping stage.
  After the study, Conditioning matters
- The impact of using local asymptotics and conditioning on stopping is illustrated with Pocock's two-stage design.

Conclusions on Inference for Adaptive Designs

Beware when making data-dependent changes to your sampling procedure, you sample size or anything else, that your analysis will likely need adjustments to be valid.

Hvala!