A General Framework for Combining Information & A Frequentist Approach to Incorporate Expert Opinions

Minge Xie
Department of Statistics
Rutgers, The State University of New Jersey
Supported in part by grants from NSF, NSA and OMJSA

Outline
- Motivating example
  - Clinical trial on Migraine Pain (Johnson & Johnson Inc.)
  - Potential problems of Bayesian solutions
- Statistical inference using confidence distributions (CDs)
  - CD concept: a “distribution estimator” of a parameter
  - General framework for combining information
- Incorporating expert opinions to clinical trials
  - Modeling of expert opinions
  - Combining Information via confidence distributions

Motivating Example: Migraine Pain Clinical Trial (J&J)
- Binomial Clinical trial on migraine headache
  - Randomized trial of two treatment groups:
    - n₁ subjects in A group; n₂ subjects in B group (control)
  - Binary responses:
    - \( X_{1i} \sim \text{Bernoulli}(p_1) \) & \( X_{0j} \sim \text{Bernoulli}(p_0) \)
- Expert opinions on the improvement (\( \delta = p_1 - p_2 \))
  - Being solicited from 11 experts, prior to the clinical trial.
  - Following the design of Parmar et al. (1994), Spiegelhalter et al. (1994)
- Goal: Incorporate expert opinions with information in the clinical data
  - Generating hypothesis testing questions for future studies

Univariate Bayesian Approach
Naïve Bayesian: directly apply Spiegelhalter et al. (1994)’s approach for normal clinical trials
- Normal Prior (fit a normal curve to histogram)
  \( \pi(\hat{\theta}) = N(\bar{y}_{data}, \sigma_{data}^2) \)
- Likelihood (estimated)
  \( \hat{L}(\theta|\hat{\delta}) = \frac{1}{2} \log(2\pi \sigma_{data}^2) - \frac{1}{2} (\delta - \bar{y})^2 / \sigma_{data}^2 \)
  Where \( \delta = X_1 - X_2 \) and \( \bar{y} = X_1(1 - X_2)/n_1 + X_2(1 - X_1)/n_2 \)
- Posterior (explicit)
  \( \pi(\theta|data) \propto \pi(\hat{\theta}|data) \pi(\hat{\delta}|\theta) = N(\frac{\hat{\delta}/\sigma_{data}^2 + \bar{y}/\sigma_{data}^2}{1/\sigma_{data}^2 + 1/\sigma_{data}^2}, \frac{1}{1/\sigma_{data}^2 + 1/\sigma_{data}^2}) \)
Issues in the Naïve Bayesian Approach

- **Theoretical Issue:** Strictly speaking, Bayesian theory does not apply in the univariate Bayesian approach.
  - Need to know \( f(d) \) to apply the Bayesian formula, which is NOT possible in the binomial trial!
  - Bayesian method is not good at “division of labor” (Efron 1986, Wasserman 2007)
- Normal prior could also be an issue
  - Bounded range of \( \delta = p_1 - p_0 \)
  - Possible skewed distribution of expert opinions
- **Full Bayesian solution:** use “a comprehensive Bayesian model” to jointly model \( p_0 \) and \( p_1 \).

Bi-variate Bayesian Approach

- **Bi-Beta prior** (Olkin and Liu, 2003)
  - Marginals are beta-distributed; Correlation ranges from 0 to 1.
- Additional information (or assumption) on \( p_0 \) is required
- Likelihood function
  - Two independent binomial distributions (or normal approximations)
  - Posterior
  - No explicit formula and need to use an MCMC algorithm

Frequentist Alternative

- Feature:
  - Directly model the parameter of interest \( \delta \), instead of both \( p_0 \) & \( p_1 \)
  - This is the so called “the edge of division of labor” of a frequentist approach (Efron, 1986, Wasserman, 2007).
  - No prior information on \( p_0 \) and \( p_1 \) is collected or available in the J&J project

- Problem:
  - Only experts opinion, no actual data, are available on prior experiments
  - Not possible in conventional frequentist inference

Solution: A Frequentist Bayesian compromise approach utilizing confidence distributions (CDs)
**Background on Confidence Distribution (CD)**
- Parameter Estimations
- Point Estimation
- Confidence Intervals
- Confidence Distributions (CDs)
  - Loaded with information (comparable to Bayesian posterior distributions)
  - Useful device for constructing all types of frequentist's statistical inferences, e.g., point estimator, confidence interval, p-value, etc.
- Long history but little attention on CDs in the past
  - Historic connection to fiducial distribution, which is largely considered as “Fisher’s biggest blunder” (see, Efron, 1998)
- Not yet seriously looked at the possible utilities of CDs especially in statistical applications

**Renewed interest in CDs**
- The renewed interest starts from Efron (1998)
  - Bootstrap distributions are “distribution estimators” and CDs.
  - Efron (1998) has “a prediction” that “something like fiducial inference may will ‘become a big hit in the 21st century’”.
- Some recent articles:
- Highlights: New definition (pure frequentist/non-fiducial); Likelihood connection; CD combination; Predictive distribution; HCDR, etc.

**Classical Notion of CD**
- Classical CD function (Efron 1993):
  - Let \( -\infty < \xi_n(\alpha) < \infty \), be a 100\(\alpha\)% lower-side confidence interval for \(\theta\), at every level \(\alpha \in (0,1)\).
  - Assume the \(\xi_n(\alpha) = \xi_n(X_n, \alpha)\) is continuous and increasing in \(\alpha\).
  - The inverse function \(H_n(\cdot) = \xi^{-1}_n(\cdot)\) is a CD in the usual Fisherian sense.

**A Formal Definition**
A function \(H_n(\theta) = H_n(X_n, \theta)\) on \(\mathcal{X} \times \Theta \to (0,1)\) is called a confidence distribution (CD) for a parameter \(\theta\), if

(i): For each given sample set \(X_n \in \mathcal{X}\), \(H_n(\cdot)\) is a continuous cumulative distribution function in the parameter space \(\Theta\);

(ii): At the true value \(\theta = \theta_0\), \(H_n(\theta_0) = H_n(X_n, \theta_0)\), as a function of the sample set \(X_n\), has the uniform distribution \(U(0,1)\).

(Adopted from Singh et al., 2005)

**A Formal Definition (cont.)**
The function \(H_n(\cdot)\) is called an asymptotic confidence distribution (aCD), if (ii) is replaced by

(ii)\(\): At \(\theta = \theta_0\), \(H_n(\theta_0) \to U(0,1)\), as \(n \to \infty\), and the continuity requirement on \(H_n(\cdot)\) is dropped.

**Connection to Classical Notion**
In the classical example where \(H_n(\theta) = \xi^{-1}_n(\theta)\) and \(\xi_n(\alpha)\) being the upper end of lower-side confidence interval:

For any \(\alpha \in (0,1)\) and at \(\theta = \theta_0\),

\[
P[H_n(\theta_0) \leq \alpha] = P[\theta \leq \xi_n(\alpha)] = \alpha.
\]

\(\Rightarrow H_n(\theta_0) = H_n(X_n, \theta_0)\), as a function of the sample set \(X_n\) is \(U(0,1)\) distributed!
Depiction of the 2nd Requirement

Right Amount of Information

\[ H(\theta) \leq 1 - H(\theta) \]

when \( \theta < \theta_0 \)

Wrong Amount of Information

\[ H(\theta) \geq 1 - H(\theta) \]

when \( \theta > \theta_0 \)

CD Subsumes a Broad Range of Examples

- Regular parametric cases
  - Including classical examples of fiducial distributions & plus more
- Bootstrap distributions
- Likelihood function (normalized as function of parameters)
  - Including likelihood function, profile likelihood, implied likelihood, etc. (e.g., Singh, Xie, Strawderman, 2005, 2007, Efron, 1993, Schweder and Hjort, 2008)
- The significance (p-value) functions (Fraser, 1991)
- Predictive distribution (Lawless and Fredette 2005)
- Some Bayesian priors and posteriors
- Among others

CD Example 1: Normal Mean

Assume \( X_1, \ldots, X_n \) is from \( N(\mu, \sigma^2) \).

- A CD for \( \mu \) (\( \sigma \) known):
  \[ H_n(\theta) = \Phi\left( \theta - X_\bar{n}/\sigma\right) \]

  \( \Rightarrow \mu \) is estimated by \( \bar{X} \).

- A CD for \( \mu \) (\( \sigma \) unknown):
  \[ H_n(\theta) = F_{t_{n-1}}\left( \theta - X_\bar{n}/\sigma_0^2\right) \]

  where \( F_{t_{n-1}} \) is the cumulative distribution function of the \( t_{n-1} \) distribution.

CD Example 2: Bootstrap Distribution

Let \( \hat{\theta} \) be a consistent estimator of \( \theta \)

\( \hat{\theta}_B \) be the corresponding bootstrap estimator

- An aCD for \( \theta \) is
  \[ H_n(\theta) = 1 - P(\hat{\theta}_B - \theta \leq 0 | \text{data}) = P(\hat{\theta}_B \geq 2 \theta - \theta | \text{data}) \]

- As \( n \to 1 \), the limiting distribution of centered \( \hat{\theta} \) is often symmetric. In this case, the raw bootstrap distribution
  \[ H_n(\theta) = 1 - P(\hat{\theta} - \theta \leq 0 | \text{data}) = P(\hat{\theta} \geq 0 | \text{data}) \]
  is also an aCD for \( \theta \)

CD Example 3: Informative Bayesian Prior

Informative prior distribution can be viewed as a CD

- An informative prior distribution of mean parameter \( \theta \)

  \[ \pi(\theta) \sim N(\mu_0, \sigma_0)^2 \]

- Past experiment: summary statistic \( X_0 \) with variance \( \text{var}(X_0) = \sigma_0^2 \) and realization (observation) \( X_0 = \mu_0 \).

- Denote by \( \theta \) the sample space of the past experiment

- The parameter space of \( \theta \)

  \[ H_0(\theta) = \Phi\left( \theta - X_0/\sigma_0\right) \]

  \( \Rightarrow \) \( H_n(\theta) = \Phi\left( \theta - X_\bar{n}/\sigma_0\right) \) is a CD function on \( \theta \)

A General Framework of Combining Information

Assume \( H_1(\theta), \ldots, H_k(\theta) \) are CDs from \( k \) independent studies.

- A continuous & monotonic function (in each coordinate)

  \[ g_r(u_1, \ldots, u_k) : [0, 1]^k \to R \]

  Define

  \[ G_r(t) = P(g_r(U_1, \ldots, U_k) \leq t) \]

  where \( U_1, \ldots, U_k \) are i.i.d. U(0,1) random variables

- Combined CD function

  \[ H_{\text{comb}}(\theta) = G_r\{g_r(H_1(\theta), \ldots, H_k(\theta))\} \]

  is a CD function for the parameter \( \theta \)
**Convenient Choices**

- A convenient choice
  
  \[ g_c(u_1, \ldots, u_k) = F_1^{-1}(u_1) + \ldots + F_k^{-1}(u_k), \]
  
  for any continuous cumulative distribution function \( F \).

  \[ H_c^*(\theta) - G_c[F_1^{-1}(H_1(\theta)) + \ldots + F_k^{-1}(H_k(\theta))], \]
  
  where \( c \in \{ F_1 \} \times \ldots \times \{ F_k \} \), and \( \ast \) means convolution.

**A Special Case \( F_0 = \Phi \)**

- No weights combination
  
  \[ H_c^*(\theta) = \Phi \left[ \sum_{i=1}^{k} w_i \Phi^{-1}(H_i(\theta)) \right] \]
  
  \( \Rightarrow \) Classical meta-analysis of p-value combination using normal distribution function

- Weighted combination
  
  \[ H_c^*(\theta) = \Phi \left[ \sum_{i=1}^{k} w_i \Phi^{-1}(H_i(\theta)) / \sqrt{\sum_{i=1}^{k} w_i^2} \right], \]
  
  \( \Rightarrow \) Normal model based meta-analysis methods

**Error Bounds on Combining**

**Lemma** (Xie et al 2008). Let

\[ P(H_i(h_i) \leq t) - t \leq \epsilon_i \]

for all \( t \in [0,1], i = 1, 2 \ldots \). Then

\[ P(H_c^*(\theta) \leq t) - t \leq \epsilon_1 + \ldots + \epsilon_k \]

- **Theoretical:** If individual CD \( H_i \) is accurate up to \( a^{-1/2} \), the combined CD function \( H^* \) is accurate up to \( n^{-1/2} \), where \( n = \#(H_1, \ldots, H_k) \).
- **Practical:** The CD combination approach still works, even if the model assumptions might be slightly off or approximations of CD functions are used.

**Unification of Meta-analysis approaches**

<table>
<thead>
<tr>
<th>Classical p-value combination approaches from Marden (1991)</th>
<th>Can be derived under the CD combination framework?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Classical p-value combination</td>
<td>Yes</td>
</tr>
<tr>
<td>Fisher method</td>
<td>Yes</td>
</tr>
<tr>
<td>Normal method</td>
<td>Yes</td>
</tr>
<tr>
<td>Tippett method</td>
<td>Yes</td>
</tr>
<tr>
<td>Normal method</td>
<td>Yes</td>
</tr>
<tr>
<td>Sum method</td>
<td>Yes</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Model based meta-analysis approaches from Table IV of Normand (1999)</th>
<th>Can be derived under the CD combination framework?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fixed effects model: MLE method</td>
<td>Yes</td>
</tr>
<tr>
<td>Fixed effects model: Bayesian method</td>
<td>Yes</td>
</tr>
<tr>
<td>Random effects model: Method of moment</td>
<td>Yes</td>
</tr>
<tr>
<td>Random effects model: REML method</td>
<td>Yes</td>
</tr>
<tr>
<td>Random effects model: Bayesian method (normal prior and fixed ( t ))</td>
<td>Yes</td>
</tr>
</tbody>
</table>

From Xie et al. (2008)

**Summary on CDs**

- **Two levels of unifications:**
  - CD, as a pure frequentist concept, covers a wide range of examples
  - Useful device for statistical inference
  - The proposed CD combination methods subsume important existing information combination approaches
  - Flexible, efficient and broad
- **Offers new methods beyond conventional approaches**
  - Frequentist’s Bayesian compromise to incorporate expert opinions
  - Robust Meta-analysis approach
CD to Summarize Expert Opinions

**Question:** How to use a CD to summarize the expert opinions?

**Answer:** The distribution underlying the histogram is the CD!

---

**Model/CD of Past Information**
- Let \( Y_0 \) be a “sufficient statistic” (sample realization) of past experiments or knowledge.
- **(Assumption)** There is pre-existing knowledge on the improvement \( \delta = p_1 - p_0 \), to which the experts are exposed.
- These experiments could be some informal, similar, or even virtual experiments, but we assume no actual data are available.

**Model/CD of Expert Opinion**
- Let \( Y_0 \) be the summary statistic of the past experiments or knowledge to which the \( i \)-th expert has been exposed.
- **(Error model)** Experts are subject to errors: Assume: \( \epsilon_i \sim N(0, \sigma^2) \).
- A CD of \( \delta \) is

---

CD to Summarize Clinical Trial Data
- An aCD from its MLE estimator

\[
H_\delta = \hat{\delta}^* - \delta
\]

with

\[
\hat{\delta}^* = \frac{\tilde{X}_1}{n_1} + \frac{\tilde{X}_0}{n_0}
\]

A CD from density/profile likelihood function

\[
H_\delta = \int_\delta^{\infty} h_\delta(\gamma)d\gamma, \quad h_\delta(\gamma) = \frac{e^{\gamma^2/2}}{\sqrt{2\pi}}
\]

Denote by \( \ell_\delta(\delta) \) the log profile likelihood function of \( \delta \).

\[
\ell_\delta(\delta) = \ell_\text{MLE}(\delta) - \ell_\text{MLE}(\delta_0)
\]

\[
H_\delta = \int_\delta^{\infty} e^{\gamma^2/2} d\gamma
\]

**CD Combination**

- $H_0(\delta)$: a CD for expert opinions (histogram or fitted curve)
- $H_F(\delta)$: a CD from clinical data.

- Combined CD using normal function (explicit):
  \[
  H^{(c)}(\delta) = \Phi \left( \frac{w_1 \Phi^{-1}(H_0(\delta)) + w_2 \Phi^{-1}(H_F(\delta))}{(w_1^2 + w_2^2)^{1/2}} \right)
  \]

An example: weights $w_1 = 1/\delta_c, w_2 = 1/\delta_d$

- When we use the normal approximation and fit a normal curve to the histogram, it matches with the naive posterior

---

**Summary & Discussion**

**Empirical results:**
- When histogram is close to bell shaped, all agree with each other
- When histogram is skewed
  - Naive Bayesian approach cannot catch the skewness
  - Full Bayesian approach produces a counter-intuitive solution

**Computational effort and implementation:**
- The CD approach is explicit and computationally simple
- No need of simulation or use of an expensive algorithm
- The histogram of expert opinions can be directly used in the CD approach
- No need to assume the form of the prior distribution or extra computing

**Theoretical issues:**
- A fundamental flaw exists in the naive Bayesian approach
  - Not possible to find a conditional density function $f(\theta|\delta)$
  - A full Bayesian solution is to jointly model $p_0, p_1$
  - Requires additional prior information on $p_0$ or $p_1$
  - (Bigger problem) Could generate a “paradox” in the skewed case.

---

**Relevant papers**

Thank You!