Calibration of Bayesian analyses including historical data: The perspective matters

Christian Röver, Tim Friede

Department of Medical Statistics, University Medical Center Göttingen, Göttingen, Germany

IBS-DR Bayes methods working group webinar
September 16, 2020
Overview

- Introduction
- A simple (1-stage) example
  - “Bayesian” vs. “frequentist” calibration
- A 2-stage example
  - more calibration perspectives
- Discussion
use of **historical data** (extrapolation, bridging, ...) encouraged to better utilize existing evidence (also in regulatory guidances\(^1\))

- often: via (informative) *priors*

- implementations often include investigations of **operating characteristics**

- common: provide an exemplary “historical data” scenario, then investigate resulting operating characteristics

- question is, how such investigations (esp.: calibration, coverage) may / should be approached

- differences w.r.t. what to *condition* on, what to *marginalize* over

---

\(^1\) e.g., EMA/199678/2016, FDA-2016-D-2153, CHMP/EWP/83561/2005, FDA-2015-D-1376
Introduction
Calibration

- statistical analyses yield **probabilistic statements** (credible / confidence intervals, …)
- to be meaningful, probabilities need to be **calibrated**
- sometimes given **by construction** (e.g.: assumptions are met)
- sometimes need to **verify** (e.g.: asymptotics are used, assumptions are violated)

**How can we tell?**
- usually via simulation: may be checked by matching **probabilities** with (long-run) **frequencies** in replicated samples
- what constitutes a **proper replication** depends on the context (esp.: Bayesian vs. frequentist approaches)
here: consider **coverage probabilities** of **95% intervals** (credible / confidence intervals)

calibration may be checked in several ways; besides calibration (ignored here): **sharpness** is important

---

Bayesian and frequentist methods have differing requirements regarding calibration:

Bayesian

intervals provide coverage on average over the prior distribution: a “marginal” property

(marginalisation over $p(\theta, y)$)

frequentist

intervals provide coverage for any point in parameter space: a “(uniform) conditional” property

(marginalisation over $p(y|\theta)$)
A simple example

1-stage example: setup

- *nature* provides a parameter value:
  \[ \theta \sim \text{Normal}(\mu_\Omega = 0, \sigma^2_\Omega = 1) \]

- *experimenter* gathers an i.i.d. sample \((y_1, \ldots, y_m)\) of size \(m\):
  \[ \bar{y} | \theta \sim \text{Normal}(\theta, s^2_{\bar{y}} = \frac{1}{m}) \]
A simple example

1-stage example: analysis

- prior:

\[ \theta \sim \text{Normal}(\mu_p = 0, \sigma_p^2 = 1) \]

(assumptions \((\mu_p, \sigma_p)\) match setting \((\mu_\Omega, \sigma_\Omega)\))

- likelihood:

\[ \bar{y} | \theta \sim \text{Normal}(\theta, s_\bar{y}^2 = \frac{1}{m}) \]

- posterior:

\[ \theta | \bar{y} \sim \text{Normal} \left( \frac{1}{\sigma_p^2} \mu_p + \frac{1}{s_\bar{y}^2} \bar{y}, \frac{1}{\sigma_p^2 + \frac{1}{s_\bar{y}^2}} \right) \]
A simple example

1-stage example: illustration

- a sample of size $n = 9$ yields $\bar{y} = 1$
- posterior:

$$\theta | \bar{y} = 1 \sim \text{Normal}(0.90, 0.32^2)$$
Checking (marginal) calibration

- a Bayesian analysis (with proper prior) is **calibrated by construction** (prior expected coverage probability; averaged over $p(\theta, \bar{y})$)
- may be assessed via **simulation**
  (e.g. in order to check implementation)$^3$
For $j = 1, \ldots, N$:
  1. generate parameter value $\theta_j$ from the prior $p(\theta)$
  2. generate an $\bar{y}_j$ value (the data) from $p(\bar{y}_j | \theta_j)$
  3. derive the posterior distribution $p(\theta | \bar{y}_j)$
  4. check whether true value $\theta_j$ is within 95% interval
- true value should be within 95% interval in $\approx 95\%$ of cases

(here: coverage probabilities via numerical integration)

---

Checking (marginal) calibration
1-stage example

- may vary prior parameters \((\mu_p, \sigma_p)\) used in analysis (i.e., assumptions are violated!)

<table>
<thead>
<tr>
<th>prior</th>
<th>(\mu_p)</th>
<th>(\sigma_p)</th>
<th>coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0</td>
<td>1.0</td>
<td></td>
<td>95.0%</td>
</tr>
<tr>
<td>0.0</td>
<td>2.0</td>
<td></td>
<td>95.2%</td>
</tr>
<tr>
<td>0.0</td>
<td>0.5</td>
<td></td>
<td>84.2%</td>
</tr>
<tr>
<td>1.0</td>
<td>1.0</td>
<td></td>
<td>93.8%</td>
</tr>
<tr>
<td>1.0</td>
<td>2.0</td>
<td></td>
<td>95.1%</td>
</tr>
</tbody>
</table>

- 95% intervals are calibrated for matching prior \((\mu_p = \mu_\Omega, \sigma_p = \sigma_\Omega)\)
- not calibrated otherwise
Checking “conditional” calibration

1-stage example

- a Bayesian analysis in general **does not** provide frequentist coverage  
  *(conditional coverage probability; averaged over $p(\bar{y}|\theta)$)*

- may again be assessed via simulation
  *(for given parameter value $\theta^*$)*

For $j = 1, \ldots, N$:

1. generate an $\bar{y}_j$ value (the data) from $p(\bar{y}_j|\theta^*)$
2. derive the posterior distribution $p(\theta|\bar{y}_j)$
3. check whether true value $\theta^*$ is within 95% interval

- (here: coverage probabilities again via numerical integration)
may vary (fixed) true parameter ($\theta^*$) used in analysis

<table>
<thead>
<tr>
<th>true $\theta^*$</th>
<th>coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0</td>
<td>96.1%</td>
</tr>
<tr>
<td>1.0</td>
<td>95.0%</td>
</tr>
<tr>
<td>2.0</td>
<td>91.6%</td>
</tr>
<tr>
<td>3.0</td>
<td>85.6%</td>
</tr>
</tbody>
</table>
Checking “conditional” calibration

1-stage example

- may vary (fixed) true parameter \((\theta^*)\) used in analysis
  (marginalizing over \(p(\bar{y} | \theta^*)\))
Two-stage sampling

2-stage example

- variation of first example with sequential sampling in two stages ("historical" / "current")

2-stage example: setup

- parameter value:

  \[ \theta \sim \text{Normal}(\mu_\Omega = 0, \sigma^2_\Omega = 1) \]

- first sample (size \( n \)):

  \[ \bar{x} | \theta \sim \text{Normal}(\theta, s^2_\bar{x} = \frac{1}{n}) \]

- second sample (size \( m \)):

  \[ \bar{y} | \theta \sim \text{Normal}(\theta, s^2_\bar{y} = \frac{1}{m}) \]
Two-stage example

2-stage example: analysis (i)

- **prior:**
  \[ \theta \sim \text{Normal}(\mu_p = 0, \sigma_p^2 = 1) \]

- **likelihood:**
  \[ \bar{x} | \theta \sim \text{Normal}(\theta, s_{\bar{x}}^2 = \frac{1}{n}) \]
  \[ \bar{y} | \theta \sim \text{Normal}(\theta, s_{\bar{y}}^2 = \frac{1}{m}) \]
Two-stage example

2-stage example: analysis (ii)

- posterior may be expressed as:

\[
p(\theta \mid \bar{x}, \bar{y}) \propto p(\bar{y} \mid \theta) \, p(\bar{x} \mid \theta) \, p(\theta)
\]

\[
\propto \left( p(\bar{y} \mid \theta) \, p(\bar{x} \mid \theta) \right) \, p(\theta) \propto \underbrace{p(\bar{x}, \bar{y} \mid \theta)}_{\text{joint likelihood}} \underbrace{p(\theta)}_{\text{prior}}
\]

\[
\propto p(\bar{y} \mid \theta) \left( p(\bar{x} \mid \theta) \, p(\theta) \right) \propto \underbrace{p(\bar{y} \mid \theta)}_{\bar{y} \text{ likelihood}} \underbrace{p(\theta \mid \bar{x})}_{\text{historical prior}}
\]

- historical-data (\(\bar{x}\)) posterior as (informative) prior for new data (\(\bar{y}\))

- “technically” little has changed: calibration unaffected
  (may also think of data as larger sample of size \(n + m\))
Two-stage example

2-stage example: illustration

- a **first** sample of size $n = 9$ yields $\bar{x} = 1$
- analysis of **second** sample ($\bar{y}$, of size $m = 9$) now proceeds with Normal($0.90$, $0.32^2$) prior

- NB: two-stage analysis is **calibrated** from **two viewpoints**:
  - drawing $\theta_j$ from $p(\theta)$ and proceeding with 1st- and 2nd-stage analyses (marginalizing over $p(\theta, \bar{x}, \bar{y})$)
  - drawing $\theta_j$ from $p(\theta | \bar{x})$ and proceeding with 2nd-stage analysis (marginalizing over $p(\theta, \bar{y} | \bar{x})$)

- may again consider **conditional** calibration...
Two-stage example

“conditional” calibration

- may again investigate “conditional” coverage probabilities (fixing $\bar{x}$, marginalizing over $p(\bar{y} | \theta^*)$)

![Diagram showing a prior density and conditional coverage probability plot.]

(dotted lines from previous figure)
“Conditional” calibration issues
in the 2-stage example

why not to worry

- proper calibration still guaranteed
- posterior distribution (informative prior \( p(\theta | \bar{x} = 1) \)) reflects relevant range of \( \theta \) values
  (e.g., little reason to worry about, say, \( \theta^* = 0 \) when \( \bar{x} = 1 \) was observed)

why still to worry

- distrust in relevance of posterior \( p(\theta | \bar{x} = 1) \),
  i.e., disagreement with (data-) model assumptions
- should be addressed by adapting model (e.g., robustification)
“Conditional” calibration issues
What to condition on?

- NB: in previous figure, we were **conditioning on both** $\bar{x}$ and $\theta$, i.e.
  1. fixing (1st-stage) data and
  2. varying the parameter

  (ignoring $p(\bar{x} | \theta)$ or $p(\theta | \bar{x})$)

- **how sensible** is it to **marginalize over** $p(\bar{y} | \bar{x}, \theta)$?

- shouldn’t the 2-stage procedure be judged **as a whole**?

- shouldn’t **both** 1st-stage and 2nd-stage data be treated as **data**?

- might instead
  - condition on data $\bar{x}$ only
    (2nd stage is calibrated w.r.t. informative prior $p(\theta, y | \bar{x})$ by construction)
  - condition on parameter $\theta$ only
    (marginalize over both $\bar{x}$ and $\bar{y}$: $p(\bar{x}, \bar{y} | \theta)$)
“Conditional” calibration issues

What to condition on?

instead of *conditioning* on some 1st-stage data...

...may also *marginalize* over both 1st and 2nd stages
“Conditional” calibration issues
What to condition on?

- **marginalizing over** \( p(\bar{x}, \bar{y} | \theta) \) instead
  - (seemingly) “pathological” appearance is reduced when considering 1st- and 2nd-stage data jointly
    - *(not conditioning on 1st-stage data)*
  - figure reflects that inference improves upon “stage-1-only” or “stage-2-only” analyses
    - possibly the more sensible “conditional” figure to consider?

- or should one look at “marginal” calibration right away? (marginalize over \( p(\theta, \bar{x}, \bar{y}) \)?)
Conclusions
Calibration perspectives

several calibration “perspectives”
depending on what is conditioned upon:

<table>
<thead>
<tr>
<th>Conditioning on:</th>
<th>conditioning on:</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.) $p(\theta, \bar{x}, \bar{y})$</td>
<td>(prior) marginal</td>
</tr>
<tr>
<td>2.) $p(\bar{x}, \bar{y}</td>
<td>\theta)$</td>
</tr>
<tr>
<td>3.) $p(\bar{y}, \theta</td>
<td>\bar{x})$</td>
</tr>
<tr>
<td>4.) $p(\bar{y}</td>
<td>\theta, \bar{x})$</td>
</tr>
</tbody>
</table>

- our suggestion: **prefer 1st or 3rd**
- or possibly 2nd (relevant in regulatory context?)
Conclusions

Discussion

- importance of what to **condition** upon
- when considering “**conditional**” calibration, clarify
  - what properties you **require** and **may expect**
  - whether it is the right figure to look at
- in particular, **uniform coverage probability** seems impossible for an informative prior; marginalisation over 1st-stage data may be the more sensible approach
- Bayes model calibrated by construction
  (w.r.t. **prior** \( p(\theta, \bar{x}, \bar{y}) \) as well as **posterior** \( p(\theta, \bar{y} | \bar{x}) \)); important to convincingly **motivate assumptions, build in robustness**.
- implications for **related/similar problems**
  (evidence synthesis, adaptive designs,...)