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)

Introduction

Calibration: 95% coverages

- here: consider **coverage probabilities** of **95% intervals** (credible / confidence intervals)
- calibration may be checked in several ways;
  besides calibration (ignored here): **sharpness** is important [2]

[2] T. Gneiting, F. Balabdaoui, A. E. Raftery. Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society B*, **69**(2):243–268; 2007.

Bayesian vs. frequentist calibration

"marginal" vs. "conditional"

- 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:

$$\theta \;\sim\; \mathrm{Normal}(\mu_\Omega = 0,\, \sigma^2_\Omega = 1)$$

- *experimenter* gathers an i.i.d. sample $(y_1, \ldots, y_m)$ of size m:

$$\bar{y}|\theta \;\sim\; \mathrm{Normal}(\theta,\, s^2_{\bar{y}} = \tfrac{1}{m})$$

A simple example

1-stage example: analysis

- prior:

$$\theta \;\sim\; \text{Normal}\big(\mu_\text{p} = 0,\; \sigma_\text{p}^2 = 1\big)$$

  (assumptions $(\mu_\text{p}, \sigma_\text{p})$ match setting $(\mu_\Omega, \sigma_\Omega)$)

- likelihood:

$$\bar{y}|\theta \quad \sim \quad \text{Normal}\big(\theta,\; s_{\bar{y}}^2 = \tfrac{1}{m}\big)$$

- posterior:

$$\theta|\bar{y} \;\sim\; \text{Normal}\left(\frac{\frac{1}{\sigma_\text{p}^2}\mu_\text{p} + \frac{1}{s_{\bar{y}}^2}\bar{y}}{\frac{1}{\sigma_\text{p}^2} + \frac{1}{s_{\bar{y}}^2}},\; \frac{1}{\frac{1}{\sigma_\text{p}^2} + \frac{1}{s_{\bar{y}}^2}}\right)$$

A simple example

1-stage example: illustration

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

$$\theta|\bar{y}=1 \;\sim\; \mathrm{Normal}(0.90,\; 0.32^2)$$

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 θ_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 θ_j is within 95% interval
- true value should be within 95% interval in $\approx$ 95% of cases
- (here: coverage probabilities via numerical integration)

[3] S. R. Cook, A. Gelman, D. B. Rubin. Validation of software for Bayesian models using posterior quantiles. *Journal of Computational and Graphical Statistics*, **15**(3):675–692; 2006.

Checking (marginal) calibration

1-stage example

- may **vary prior parameters** (μ_p, σ_p) used in analysis
  (i.e., assumptions are violated!)

prior		
μ_p	σ_p	coverage
0.0	**1.0**	**95.0%**
0.0	2.0	95.2%
0.0	0.5	84.2%
1.0	1.0	93.8%
1.0	2.0	95.1%

- 95% intervals are calibrated for matching prior ($\mu_\mathrm{p} = \mu_\Omega$, $\sigma_\mathrm{p} = \sigma_\Omega$)
- 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 θ^*)
  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 θ^* is within 95% interval
- (here: coverage probabilities again via numerical integration)

Checking "conditional" calibration

1-stage example

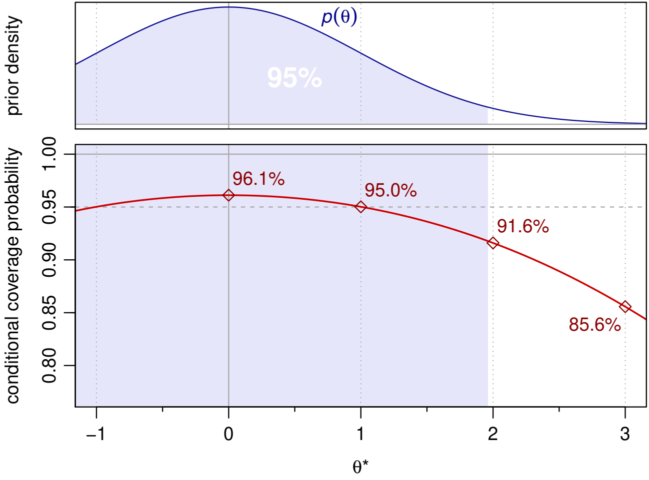
- may **vary (fixed) true parameter** (θ^*) used in analysis

true θ^*	coverage
0.0	96.1%
1.0	95.0%
2.0	91.6%
3.0	85.6%

Checking "conditional" calibration

1-stage example

- may **vary (fixed) true parameter** (θ^*) used in analysis (marginalizing over $p(\bar{y} \mid \theta^*)$)



Two-stage sampling

2-stage example

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

2-stage example: setup

- parameter value:

$$\theta \;\sim\; \mathrm{Normal}\big(\mu_\Omega = 0,\, \sigma_\Omega^2 = 1\big)$$

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

$$\bar{x}|\theta \;\sim\; \mathrm{Normal}\big(\theta,\, s_{\bar{x}}^2 = \tfrac{1}{n}\big)$$

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

$$\bar{y}|\theta \;\sim\; \mathrm{Normal}\big(\theta,\, s_{\bar{y}}^2 = \tfrac{1}{m}\big)$$

Two-stage example

2-stage example: analysis (i)

- prior:

$$\theta \;\sim\; \text{Normal}\big(\mu_{\text{p}} = 0,\, \sigma_{\text{p}}^2 = 1\big)$$

- likelihood:

$$\begin{aligned} \bar{x}|\theta &\sim \text{Normal}\big(\theta,\, s_{\bar{x}}^2 = \tfrac{1}{n}\big) \\ \bar{y}|\theta &\sim \text{Normal}\big(\theta,\, s_{\bar{y}}^2 = \tfrac{1}{m}\big) \end{aligned}$$

Two-stage example

2-stage example: analysis (ii)

- posterior may be expressed as:

$$
\begin{aligned}
p\big(\theta \,\big|\, \bar{x}, \bar{y}\big) &\propto p\big(\bar{y} \,\big|\, \theta\big)\; p\big(\bar{x} \,\big|\, \theta\big)\; p(\theta) \\
&\propto \Big(p\big(\bar{y} \,\big|\, \theta\big)\; p\big(\bar{x} \,\big|\, \theta\big)\Big)\; p(\theta) \;\propto\; \underbrace{p\big(\bar{x}, \bar{y} \,\big|\, \theta\big)}_{\text{joint likelihood}}\; \underbrace{p(\theta)}_{\text{prior}} \\
&\propto p\big(\bar{y} \,\big|\, \theta\big)\; \Big(p\big(\bar{x} \,\big|\, \theta\big)\; p(\theta)\Big) \;\propto\; \underbrace{p\big(\bar{y} \,\big|\, \theta\big)}_{\bar{y}\text{ likelihood}}\; \underbrace{p\big(\theta \,\big|\, \bar{x}\big)}_{\text{historical prior}}
\end{aligned}
$$

- **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

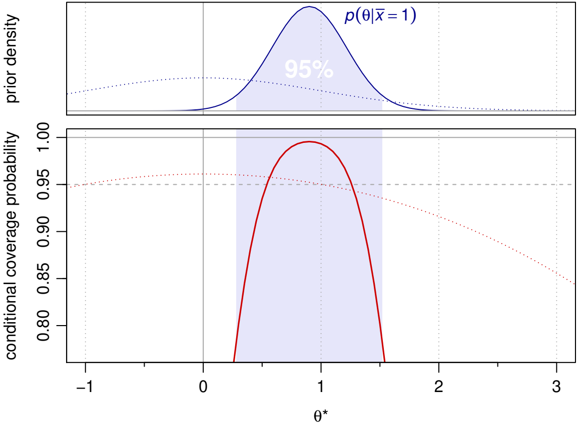
- NB: two-stage analysis is **calibrated** from **two viewpoints**:
  - drawing θ_j from $p(\theta)$ and proceeding with 1st- and 2nd-stage analyses (marginalizing over $p(\theta, \bar{x}, \bar{y})$)
  - drawing θ_j from $p(\theta|\bar{x})$ and proceeding with 2nd-stage analysis (marginalizing over $p(\theta,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

- may again consider *conditional* calibration…

Two-stage example

"conditional" calibration

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



(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 θ 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?

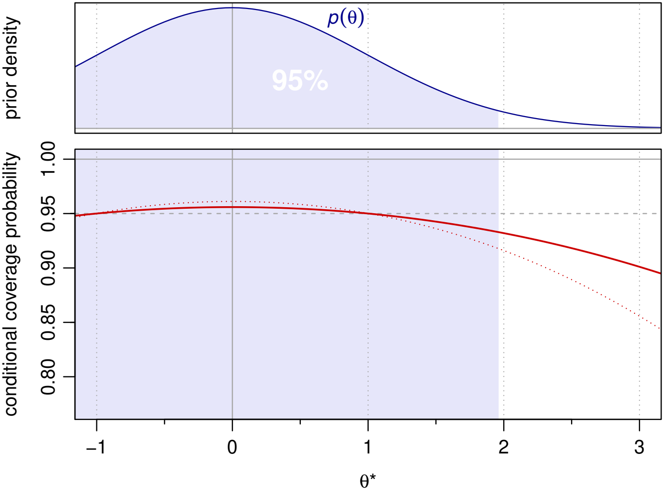
- NB: in previous figure, we were **conditioning** on **both $\bar{x}$ and θ**, i.e.
  (i) fixing (1st-stage) data and
  (ii) varying the parameter

  (ignoring $p(\bar{x}|\theta)$ or $p(\theta|\bar{x})$)
- **how sensible** is it to **marginalize over** $p(\bar{y} \mid \bar{x}, \theta)$?
- 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 $p(\theta, y \mid \bar{x})$ by construction)
  - condition on parameter θ only
    (marginalize over both $\bar{x}$ and $\bar{y}$: $p(\bar{x}, \bar{y} \mid \theta)$)

“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:

			conditioning on:
1.)	$p(\theta, \bar{x}, \bar{y})$	"(prior) marginal"	
2.)	$p(\bar{x}, \bar{y} \mid \theta)$	"conditional on θ^*"	parameter θ
3.)	$p(\bar{y}, \theta \mid \bar{x})$	"($\bar{x}$-posterior) marginal"	data $\bar{x}$
4.)	$p(\bar{y} \mid \theta, \bar{x})$	"conditional on $\bar{x}$ and θ^*"	data $\bar{x}$ and parameter θ

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

Conclusions

Discussion

- importance of what to **condition** upon
- whan 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** $p(\theta, \bar{y} \mid \bar{x})$);
  important to convincingly **motivate assumptions**, **build in robustness**.
- implications for **related/similar problems**
  (evidence synthesis, adaptive designs,. . .)