Lecture 9 - Model Criticism

Susie Bayarri (U Valencia), with

Jim Berger , Maria-Eugenia Castellanos and Javier Morales

Duke U, U Rey Juan Carlos, U Miguel Hernandez

CBMS-MUM

UC Santa Cruz July 23-27, 2012

The Problem

We have worked hard and have come with one model for the data that we are pretty happy about:

$$\mathcal{M} : \boldsymbol{X} = \{X_1, \ldots, X_n\} \mid \boldsymbol{\theta} \sim f(\boldsymbol{x} \mid \boldsymbol{\theta})$$

BUT what if I am wrong? The question:

is model O.K.? $\leftrightarrow$ *is observed data* $\boldsymbol{x}_{obs}$ *compatible with model?*

is a very old question in statistics. Can Bayesians provide an answer?

Model criticism vs model comparison. We want:

- Model check (no comparison) $\rightsquigarrow$ no alternatives
- Objective Bayes $\rightsquigarrow$ no subjective priors

Why no alternatives?

- Model comparison *is* the Bayesian way: If one is uncertain about model $\mathcal{M}$, one should select a believable set of models $\mathcal{M}_i$ and do model choice or BMA (or others, depending on the utility function)
- Model criticism only applies when "$\mathcal{M}$ is our model"; one thinks that MS and MA is likely to be too hard and offer little improvement
- Having really no alternatives $\leadsto$ can't reject $\mathcal{M}$

  If data compatible $\leadsto$ pat yourself in the back and continue the analysis

  If data incompatible $\leadsto$ do the hard work!
- Checking as a exploratory tool $\leadsto$ look for alternatives only if needed

Why objective Bayes?

- Most natural at exploratory stage
- Prior assessment might be (way!) too hard (and the effort wasted if the model is not good)
- Most importantly, with a subjective, informative prior, model checking can only check the combination of prior and model:

  Subjective Bayes model criticism can not (and maybe does not want to) separate inadequacy of model from inadequacy of prior

- In 'model criticism' the general goal is to check the adequacy of the data generating model $f(\boldsymbol{x} \mid \boldsymbol{\theta})$

With no alternative models ...

do data $\boldsymbol{x}_{obs}$ looks like it should? are we "surprised" to see this $\boldsymbol{x}_{obs}$?

To investigate this question, choose:

1. a diagnostic statistic $T = t(\boldsymbol{X}\ )$ to investigate incompatibility of data with assumed (null) model. Compute $t_{obs} = t(\boldsymbol{x}_{obs})$
2. a (specified) distribution $f(t)$ of T under the assumed model
3. a way to measure conflict between t_{obs} and $f(t)$

- Different choices of 1,2,3 $\rightsquigarrow$ different model checks
- Concentrate on the optimal choice of $f(t)$ for ANY choice of the statistic T and ANY choice of measuring incompatibility (whether formal measures of surprise or informal 'checks')

… two words about T

We will not be concerned about choice of T in this talk, but

- choice of T *is* important
- Choice is often made on casual, intuitive manner, specially for complex models,
- Often kind of 'surrogate' for alternatives; so if clear alternative(s) in mind we recommend formal Bayesian analysis
- if choosing T too hard $\rightsquigarrow$ devote the effort to formulate the alternative models

NOTE: if T is ancillary (or nearly so) $\rightsquigarrow$ it doesn't matter how we get rid of θ (or matters less)

If distribution of T depends on θ (complex models, T chosen casually) $\rightsquigarrow$ which distribution $f(t)$ is used becomes *crucial*

... two words about measuring conflict

To measure compatibility between observed t_{obs} and 'null' $f(t)$:

- Likelihood-based measures, like the relative height of the density $f(t)$ at t_{obs} (*Relative Predictive Surprise* in Berger 1985) [a]:

$$RPS = \frac{f(t_{obs})}{\sup_t f(t)}$$

(we do not treat these in this talk)

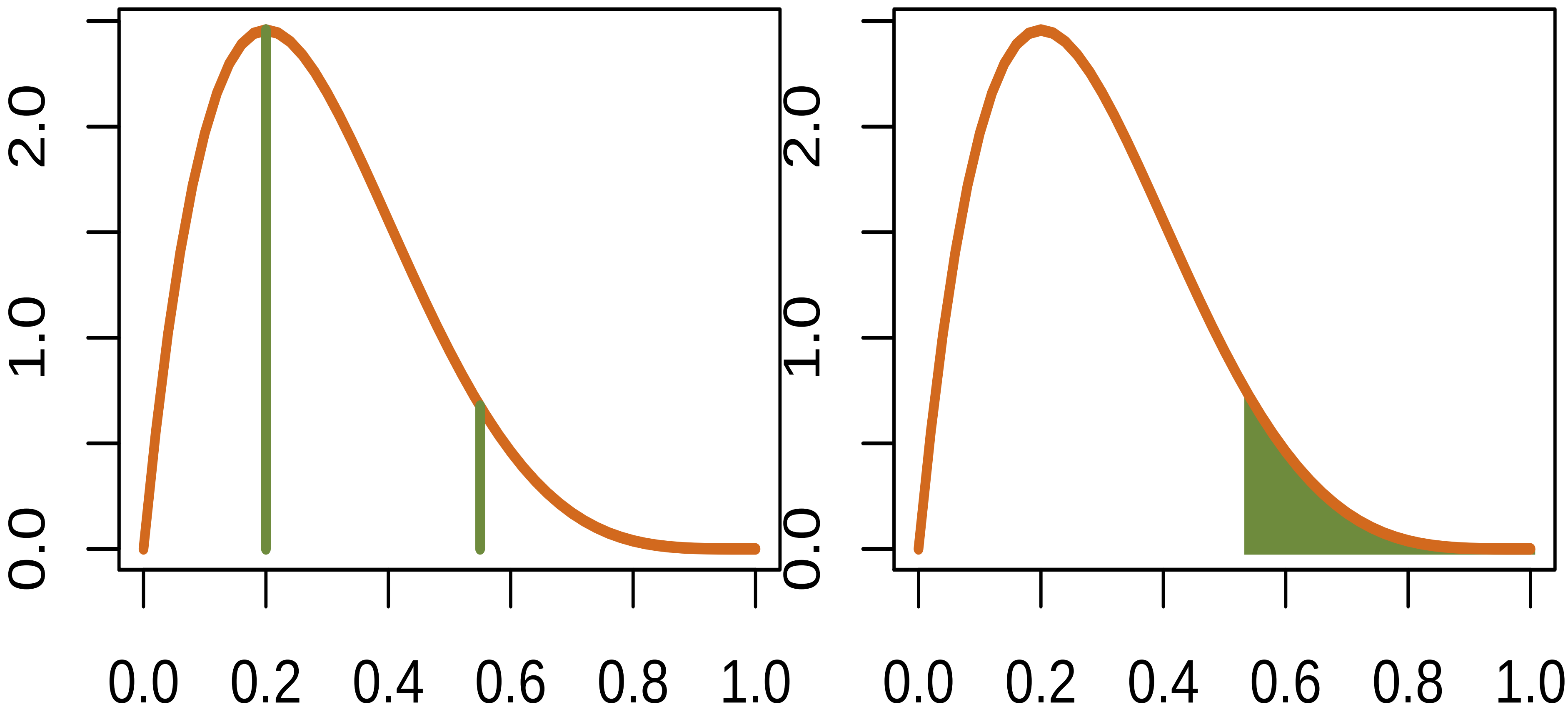
- Tail-areas based measures, like the most popular p-values

$$p = Pr^{f(t)}(t(\boldsymbol{X}) \geq t(\boldsymbol{x}_{obs}))$$

which are the ones we will be considering

[a] For other proposals of surprise indices see Weaver, 48; Good, 56, 83, 88; Berger, 85; Evans, 97, 06; Bayarri and Berger, 97

relative height and p-values



Whaaat???? p-values????

- yeap, we know ... we have been advising you again and again not to use p-values ...
- relative height has a more Bayesian (and likelihood) flavour
- as ugly as they are, p-values have some advantages:
  - easier to compute (and to MCMC)
  - invariant under 1-1 transformations
  - everyone is used to them
- so we stick to them, however:
  - we have explored both (B&B, B&C, B&M)
  - we know how to calibrate for proper interpretation

Note 1. Remember: luckily we can recalibrate for easy interpretation: when $p < e^{-1}$ compute

- $B(p) = -e\,p\log(p)$: interpret as the odds (or Bayes factor) of H_0 to (unspecified) H_1
- $\alpha(p) = (1 + [-e\,p\log(p)]^{-1})^{-1}$: interpret as (conditional) frequentist Type I error probability

Note 2. big problem with p-values $\rightsquigarrow$ they exaggerate the evidence against the null.

However, here this only means more, maybe unneeded work: look for alternative models when maybe the original model was the best of all entertained models $\rightsquigarrow$ not a serious mistake.

Note 3. the opposite, that is a procedure which fails to detect seriously wrong models IS **a serious, worrisome mistake** in model checking.

recap: $T = t(\boldsymbol{X})$ is a test statistics; Assume that large values of T indicate incompatibility with $\mathcal{M}$

$T \mid \boldsymbol{\theta} \sim f(t \mid \boldsymbol{\theta})$ with $\boldsymbol{\theta}$ unknown $\rightsquigarrow$ need to "get rid" of $\boldsymbol{\theta}$ to compute p-values, relative heights, ...

in this talk: Compute $p = Pr^{f(t)}\{T \geq t_{obs}\}$ with $t_{obs} = t(\boldsymbol{x}_{obs})$ Model 'under suspicion' if p small.

several possibilities to get to a completely specified distribution $f(t)$ (under $\mathcal{M}$) to compute 'measures of surprise'

focus: compare some few such ways through their respective p-values, but message applies also to other measures of surprise.

important point in this talk is not so much p-values versus 'likelihood-ratio' type measures, but the distribution used (more so with casually chosen T, like with informal graphical checks)

finding $f(t)$ free of θ

- want to 'eliminate' $\boldsymbol{\theta}$ from $f(t \mid \boldsymbol{\theta})$ to produce a known $f(t)$ for computing the p-value p
- several ways to 'eliminate' the unknown $\boldsymbol{\theta}$
  - plug-in p-value (p_{plug})
  - similar p-value (p_{sim})
  - prior predictive p-value (p_{prior})
  - posterior predictive p-value (p_{post})
  - partial posterior predictive p-value (p_{ppost})
  - conditional predictive p-value (p_{cpred})

Normal example

- under the null, $X_i \sim N(0, \sigma^2)$
  call $s^2 = \sum(x_i - \overline{x})^2/n$
- discrepancy statistic $t(\mathbf{X}) = |\overline{X}|$ (mean)
  $\overline{X} \sim N(0, \sigma^2/n)$
- various p-values are

$$p = \mathsf{P}r\{|\overline{X}| > |\overline{x}_{obs}|\}$$

- usual non-informative prior for σ^2: $\pi(\sigma^2) \propto 1/\sigma^2$

plug-in p-value:

replace θ by some estimate $\hat{\theta}$, such as the MLE:

$$\mathbf{p_{plug}} = \mathsf{P}r^{f(\cdot;\hat{\theta})}(t(\mathbf{X}) \geq t(\mathbf{x}_{obs}))$$

- strengths
  - simplicity
  - intuitive appeal
- weakness
  - failure to account for uncertainty in the estimation of θ
  - double use of the data

Note: distinction between the plug-ing $f(t;\hat{\theta}) = f(t \mid \theta = \hat{\theta}(\boldsymbol{x}_{obs}))$ and the conditional distribution $f(t \mid \hat{\theta}, \theta)$ which can depend on θ

Normal example (cont.)

- $\mathbf{p_{plug}}$
  - MLE $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 = s^2 + \overline{x}^2$ and

$$p_{plug} = 2\left[1 - \Phi\left(\frac{\sqrt{n}\ |\overline{x}_{obs}|}{\sqrt{s_{obs}^2 + \overline{x}_{obs}^2}}\right)\right]$$

  - but $p_{plug} \longrightarrow 2[1 - \Phi(\sqrt{n})]$ (positive constant) as $|\overline{x}_{obs}|/s_{obs} \longrightarrow \infty$
  - p-value will not go to zero , no matter how strong the evidence ! !

similar p-value:

condition on a sufficient statistic U, for θ, so that, by definition $f(\boldsymbol{x} \mid u, \theta) = f(\boldsymbol{x} \mid u)$ is free of θ

$$\mathbf{p_{sim}} = \mathsf{P}r^{f(\cdot|u_{obs})}(t(\mathbf{X}) \geq t(\mathbf{x}_{obs}))$$

- strength
  - based on a proper probability computation (desirable properties)
- weaknesses
  - suitable sufficient U typically does not exist
  - choice of T is then typically forced (and might have poor power)

Normal example (cont.)

- $\mathbf{p_{sim}}$
  - sufficient statistic for $\sigma^2 \leadsto V = \sum_{i=1}^{n} X_i^2 = ||\mathbf{X}||^2$.
  - distribution of $\mathbf{X}$ given $v_{obs} = ||\mathbf{x}_{obs}||^2$ is uniform on $\{\mathbf{x} : \; ||\mathbf{x}||^2 = ||\mathbf{x}_{obs}||^2\}$, and

$$p_{sim} = Pr\left(\frac{|\overline{X}|}{||\mathbf{x}_{obs}||} \; > \; \frac{|\overline{x}_{obs}|}{||\mathbf{x}_{obs}||}\right) = Pr\left(|\overline{Z}| > \frac{|\overline{x}_{obs}|}{||\mathbf{x}_{obs}||}\right)$$

    where $\mathbf{Z} \sim$ uniform on $\{\mathbf{z} : \; ||\mathbf{z}||^2 = 1\}$

  - later $\leadsto p_{sim} = p_{ppost} = p_{cpred}$

prior predictive p-value [Box, 1980]

integrate θ out w.r.t. the (proper) prior $\pi(\theta)$:

$$
\begin{aligned}
f(\boldsymbol{x}) \quad &\equiv \quad m(\boldsymbol{x}) = \int f(\boldsymbol{x};\theta)\pi(\theta)d\theta\,, \\
\mathbf{p_{prior}} \quad &= \quad \mathsf{P}r^{m(\cdot)}(t(\mathbf{X}) \geq t(\mathbf{x}_{obs}))
\end{aligned}
$$

- strengths
  - based on a proper probability computation
  - suggests a natural and simple T: $t(\mathbf{x}) = 1/m(\mathbf{x})$
- weaknesses
  - confounded by compatibility of data with prior
  - improper objective priors cannot be used

posterior predictive p-value: [Guttman, 67, Rubin, 84]

integrate θ out w.r.t. the posterior distribution

$$\pi(\theta \mid \boldsymbol{x}_{obs}) \propto f(\boldsymbol{x}_{obs};\theta)\pi(\theta)$$

leading to

$$\begin{aligned} m_{post}(\boldsymbol{x} \mid \boldsymbol{x}_{obs}) &= \int f(\boldsymbol{x};\theta)\pi(\theta \mid \boldsymbol{x}_{obs})d\theta\,, \\ \mathbf{p_{post}} &= \mathsf{P}r^{m_{post}(\cdot|\mathbf{x}_{obs})}(t(\mathbf{X}) \geq t(\mathbf{x}_{obs})) \end{aligned}$$

(generalizations in Meng 94; Gelman, Carlin, Stern and Rubin 95; Gelman, Meng and Stern 96)

- strengths
  - improper noninformative priors can be used
  - $m_{post}(\boldsymbol{x} \mid \boldsymbol{x}_{obs})$ more influenced by the model than by the prior; for large n, $\pi(\theta \mid \boldsymbol{x}_{obs})$ is concentrated at $\hat{\theta}$ so $p_{post} \approx p_{plug}$
  - easy to compute from MCMC outputs (which has make it very popular)
- weaknesses
  - "double use" of the data (which results in an unnatural behavior)
    * (1) to 'train' the improper $\pi(\theta)$ into $\pi(\theta \mid \boldsymbol{x}_{obs})$
    * (2) to compute the tail area corresponding to $t_{obs} = t(\boldsymbol{x}_{obs})$ in resulting $m(t \mid \boldsymbol{x}_{obs})$
  - lacks a pure Bayesian interpretation

Normal example (cont.)

- $\mathbf{p_{prior}}$ cannot be computed (prior improper)
- $\mathbf{p_{post}}$
  - posterior distribution
    $\pi(\sigma^2|\mathbf{x}_{obs}) = Ga^{-1}(\sigma^2 \mid n/2, n(s^2+\overline{x}^2)/2)$
  - posterior predictive of $\overline{X}$
    $m_{post}(\overline{x}|\mathbf{x}_{obs}) = t_n(\overline{x} \mid 0,\ \frac{1}{n}(s^2_{obs}+\overline{x}^2_{obs}))$
  - posterior predictive p-value

$$p_{post} = 2\left[1 - \Upsilon_n\left(\frac{\sqrt{n}\ \overline{x}_{obs}}{\sqrt{s^2_{obs}+\overline{x}^2_{obs}}}\right)\right] \approx p_{plug}$$

  - similarly to p_{plug}, $p_{post} \longrightarrow 2[1-\Upsilon_n(\sqrt{n})]$, a positive constant, as $|\overline{x}_{obs}|/s_{obs} \longrightarrow \infty$

– when $n = 4$, $p_{post} > 0.12$ no matter how many standard deviations $\overline{x}_{obs}$ is from zero

– inadequacy of p_{post} (and p_{plug}) directly traced to the double use of the data

– the problem with p_{plug} is less severe: $p_{plug} > 0.046$ when $n = 4$

partial posterior predictive p-value

idea: use information in $\boldsymbol{x}_{obs}$ NOT in t_{obs} to 'train' the, possibly improper, $\pi(\theta)$

- integrate θ w.r.t. *partial posterior* $\pi(\theta \mid \boldsymbol{x}_{obs} \setminus t_{obs})$

$$m(t \mid \boldsymbol{x}_{obs} \setminus t_{obs}) = \int f(t \mid \theta)\pi(\theta \mid \boldsymbol{x}_{obs} \setminus t_{obs})\, d\theta$$

$$\pi(\theta \mid \boldsymbol{x}_{obs} \setminus t_{obs}) \propto f(\boldsymbol{x}_{obs} \mid t_{obs}, \theta)\pi(\theta) \propto \frac{f(\boldsymbol{x}_{obs} \mid \theta)}{f(t_{obs} \mid \theta)}\, \pi(\theta)$$

  to produce (our proposal)

$$\mathbf{p_{ppost}} = \Pr{}^{m(\cdot \mid x_{obs} \setminus t_{obs})}(t(\boldsymbol{X}) \geq t(\boldsymbol{x}_{obs}))$$

- Has strengths of p_{post} with no double use of data
  also nice Bayesian justification (in terms of $(m(t \mid u))$

conditional predictive p-values

idea: for model checking with improper priors, use 'slices' of $m(\boldsymbol{x})$

- For some conditioning statistic $U = u(\boldsymbol{X})$, compute conditional predictive p-value as follows:
  - Integrate θ out with respect to the (assumed proper) conditional posterior distribution

$$\pi(\theta \mid u) \propto f(u;\theta)\pi(\theta)$$

    to get the u-conditional predictive distribution

$$m(t \mid u) = \int f(t \mid u;\theta)\pi(\theta \mid u)d\theta,$$

  - Compute the corresponding u-conditional predictive p-values

$$\mathbf{p_{cpred(u)}} = \mathsf{P}r^{m(\cdot|u_{obs})}(T \geq t_{obs})$$

- **the conditional predictive p-value p_{cpred}**
  - is a particular case and our proposal
  - choose the conditioning statistic U to be the conditional MLE of θ in $f(\boldsymbol{x} \mid t, \theta)$

$$\hat{\theta}_{cMLE}(\boldsymbol{x}) = \arg\max f(\boldsymbol{x} \mid t, \theta) = \arg\max \frac{f(\boldsymbol{x}; \theta)}{f(t; \theta)}$$

    or a one-to-one transformation; $m(t \mid u)$ invariant to such
  - so that $\mathbf{p_{cpred}} = p_{cpred(\hat{\theta}_{cMLE})}$

- RESULT: *when T is conditionally independent of $\hat{\theta}_{cMLE}$ and $(T, \hat{\theta}_{cMLE})$ are jointly sufficient, then*

$$\mathbf{p_{ppost}} = \mathbf{p_{cpred}}$$

Normal example (cont.)

- $\mathbf{p_{cpred}}$:
  - conditional m.l.e.

$$f(\boldsymbol{x} \mid t; \sigma^2) \propto \frac{f(\boldsymbol{x}; \sigma^2)}{f(t; \sigma^2)} \propto (\sigma^2)^{-\frac{n-1}{2}} \exp\{-\frac{ns^2}{2\sigma^2}\}$$

    maximized at $\hat{\sigma}^2_{cMLE} = ns^2/(n-1) \rightsquigarrow U = S^2$

  - conditional posterior

$$\pi(\sigma^2 \mid s^2) = Ga^{-1}(\sigma^2 \,|\, (n-1)/2, ns^2/2)$$

  - conditional predictive distribution

$$m(\overline{x} \mid s^2_{obs}) = t_{n-1}(\overline{x} \mid 0, \frac{1}{n-1} s^2_{obs})$$

– conditional predictive p-value

$$p_{cpred} = 2\left[1 - \Upsilon_{n-1}\left(\frac{\sqrt{n-1}\,\overline{x}_{obs}}{s_{obs}}\right)\right]$$

– perfectly satisfactory

– equals usual classical p-value $\rightsquigarrow$ true frequentist p-value

- **$\mathbf{p_{ppost}}$**:
  – $T = \overline{X}$ independent of $U = \hat{\sigma}^2_{cMLE} \propto S^2$

  – (T, U) jointly sufficient

  – partial posterior predictive p-value equals the conditional predictive p-value,

$$p_{ppost} = p_{cpred} = p_{sim} = p_{classic}$$

What do we want in a p-value?

- usual frequentist requirement $\rightsquigarrow p = p(\boldsymbol{X})$ to be $U[0,1]$ under the null, $f(\boldsymbol{x};\theta)$, for all θ

  if not $\rightsquigarrow$ no common interpretation across models $\rightsquigarrow$ not very useful

  'defining' property of a p-value

  [Meng, 94; Rubin, 96; Thompson, 97; Robins, 99; Robins, van der Vaart, and Ventura, 99; De la Horra and Rodríguez, 97]

- exact uniformity is often impossible $\rightsquigarrow$ p-value should be $U[0,1]$ under the null asymptotically (RVV, 99)

- For Bayesians with subjectively chosen priors $\rightsquigarrow$ maybe more natural $U[0,1]$ under $m(\boldsymbol{x}) \rightsquigarrow U[0,1]$ on average over θ (prior predictive p-value) (Meng, 94)

- BUT preliminary model checking $\rightsquigarrow$ objective, usually improper priors $\rightsquigarrow$ no average possible

- if p-value uniform under the null in the frequentist sense $\rightsquigarrow$ marginally $U[0,1]$ under *any* proper prior distribution (strong Bayesian property !!)

- if p-value *always* either conservative or anti-conservative in a frequentist sense (RVV 1999) $\rightsquigarrow$ guaranteed to be conservative or anti-conservative in a Bayesian sense, no matter what the prior (not too good)
- Also, Bayesians $\rightsquigarrow$ reasonable conditional performance not just unconditional uniformity (only few examples, no general results)
- other methods : power comparisons; decision-theoretic evaluations of p-values (with alternatives)

  (Schaafsma, Tolboom and Van Der Meulen 89; Blyth and Staudte 95; Hwang, Casella, Robert, Wells and Farrell 92; Hwang and Pemantle 97; Hwang and Yang 97; Thompson 97)

A toy outliers example

- checking for outliers $\rightsquigarrow T = Y_{(1)} = \min\{Y_1, \ldots, Y_n\}$ (lower tail) or $T = Y_{(n)} = \max\{Y_1, \ldots, Y_n\}$ (upper tail)
- data: 10 observations generated from $N(0, \mathbf{1})$
  - example 1: the $\min$ changed to a $-\mathbf{8}$, $T = Y_{(1)}$:
    -8, -1.27, -1.059, -0.986, -0.874, -0.204, 0.315, 0.42, 0.49, 2.457
  - example 2: the $\max$ changed to a $\mathbf{8}$, $T = Y_{(n)}$:
    -1.28, -1.27, -1.059, -0.986, -0.874, -0.204, 0.315, 0.42, 0.49, 8
- compute plug-in, posterior and partial posterior p-values

	example 1	example 2
ppp	1.59×10^{-3}	5.9×10^{-5}
post	0.133	0.104
plug-in	0.030	0.018

remember: the outlier was **8** S.D. from the rest of the data

Normal linear model example

- $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_n)^t$ response
  $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_k)^t$ regression coefficients
  $\mathbf{V}$ covariables (full rank), $\boldsymbol{\epsilon}$ errors

$$\mathbf{Y} = \mathbf{V}\boldsymbol{\theta} + \boldsymbol{\epsilon} \qquad\qquad \boldsymbol{\epsilon} \sim N_n(\mathbf{0}, \boldsymbol{\sigma^2}\mathbf{I}) \quad \sigma^2 \text{ known.}$$

- departure statistic $T = \mathbf{w}^{\mathbf{t}}\mathbf{Y}$, with given $\mathbf{w} = (w_1, w_2, \ldots, w_n)^t$
- $\pi(\boldsymbol{\theta}) = 1$ and $\pi(\boldsymbol{\theta} \mid \mathbf{y}) = N_k(\boldsymbol{\theta} \mid \hat{\boldsymbol{\theta}}, \sigma^2(\mathbf{V}^{\mathbf{t}}\mathbf{V})^{-1})$
  where $\hat{\boldsymbol{\theta}} = (\mathbf{V}^{\mathbf{t}}\mathbf{V})^{-1}\mathbf{V}^{\mathbf{t}}\mathbf{y}$ usual least squares estimate

- **Plug-in p-value**
  - $p_{plug} = \mathsf{P}r^{f(t;\hat{\theta})}(T > t_{obs}) = 1 - \Phi\left(\frac{t_{obs} - \mathbf{w^t V}\hat{\boldsymbol{\theta}}}{\sigma\sqrt{||\mathbf{w}||^2}}\right)$
  - random $p_{plug}(\mathbf{Y}) = 1 - \Phi\left(\sqrt{\frac{\mathbf{w^t B w}}{||\mathbf{w}||^2}}\ Z\right)$

    where $\mathbf{B} = \mathbf{I} - \mathbf{V}(\mathbf{V^t V})^{-1}\mathbf{V^t}$ and $Z \sim N(0,1)$
  - $p_{plug}(\mathbf{Y}) \sim U[0,1]$ distribution only if $\mathbf{V^t w} = 0$ (i.e., T is a linear function of residuals)
  - $\mathbf{w^t B w}/||\mathbf{w}||^2 < 1$, so p_{plug} is always conservative (i.e., larger than it should be - bad for model checking)

- **Posterior predictive p-value**
  - $p_{post} = \mathsf{P}r^{m_{post}(t|\mathbf{x}_{obs})}(T > t_{obs}) = 1 - \Phi\left(\frac{t_{obs} - \mathbf{w^t V}\hat{\boldsymbol{\theta}}}{\sigma\sqrt{\mathbf{w^t C w}}}\right)$
  - random $p_{post}(\mathbf{Y}) = 1 - \Phi\left(\sqrt{\frac{\mathbf{w^t B w}}{\mathbf{w^t C w}}}\ Z\right)$

    where $Z \sim N(0,1)$
  - $p_{post}(\mathbf{Y}) \sim U[0,1]$ only if $\mathbf{V^t w} = \mathbf{0}$
  - $\mathbf{w^t C w} > ||\mathbf{w}||^2$, so p_{post} is more conservative than p_{plug}

- **Partial posterior predictive p-value**
  - $\pi(\boldsymbol{\theta} \mid \mathbf{x}_{obs} \backslash t_{obs}) = N_k\left(\boldsymbol{\theta} \mid \mathbf{u}_{obs}, \sigma^2 \boldsymbol{\Sigma}\right)$ where

    $\mathbf{U} = \left(\mathbf{V^t H V}\right)^{-1} \mathbf{V^t H Y}$, $\boldsymbol{\Sigma} = \left(\mathbf{V^t H V}\right)^{-1}$, $\mathbf{H} = \left[\mathbf{I} - \mathbf{w}\mathbf{w}^t \ / \ ||\mathbf{w}||^2\right]$

    $$p_{ppost} = 1 - \Phi\left(\frac{t_{obs} - \mathbf{w^t V u}_{obs}}{\sigma\sqrt{\mathbf{w^t}\left[\mathbf{I} + \mathbf{V \Sigma V^t}\right]\mathbf{w}}}\right)$$

  - as a random p-value , $p_{ppost}(\mathbf{Y}) = 1 - \Phi(Z)$ where $Z \sim N(0,1)$ $\rightsquigarrow$ p_{ppost} is a 'valid' p-value

- **Conditional predictive p-value**
  - U maximizing $f(\mathbf{y} \mid t_{obs}; \boldsymbol{\theta})$ the one given before

    $\mathbf{U} = \left(\mathbf{V^t H V}\right)^{-1} \mathbf{V^t H Y}$
  - $\mathbf{Cov}(\mathbf{T}, \mathbf{U}) = \mathbf{0} \rightsquigarrow T$ and U independent $\rightsquigarrow p_{cpred} = p_{ppost}$

Bayesian Motivations

- U-conditional posterior predictive p-values $\leadsto$ positive features of both p_{prior} and p_{post}
  - based on $m(\mathbf{x}) \leadsto$ natural Bayesian meaning; if $\pi(\theta)$ proper $\leadsto m(t \mid u)$ conditional distribution
  - with appropriate $U \leadsto$ reflect surprise in the model
  - noninformative priors can be used, with $\pi(\theta \mid u)$ proper
  - no double use of the data $\leadsto u_{obs}$ to produce the posterior, t_{obs} to compute tail area (in the appropriate distribution)

- key $\rightsquigarrow$ suitable choice of conditioning statistic U

  Different possible choices of U in Bayarri and Berger, 97

  (Related possibility: Evans, 97; also cross-validation as in Gelfand, Dey and Chang, 92)

  – want U to contain as much information about θ as possible but not involve T

    in the example, $\sum x_i^2/n \rightsquigarrow$ all information but involves $t(\mathbf{x}) = |\overline{x}|$. Take $u(\mathbf{x}) = s^2 = \sum(x_i - \overline{x})^2/n \rightsquigarrow$ information about σ^2 independent of $t(\mathbf{X})$

  – also $u(\mathbf{x})$ same dimension as θ

  – achieve all $\rightsquigarrow$ define U as conditional m.l.e. of θ, given $t(\mathbf{x}) = t$

- partial posterior predictive p-value
  - conditional predictive p-value appealing but maybe difficult to compute
  - directly use $c\, f(\mathbf{x} \mid t; \theta)\, \pi(\theta)$ to integrate out $\theta \rightsquigarrow p_{post}$
  - partial predictive p-value very similar to conditional predictive p-value. As a matter of fact, p_{cpred} and p_{ppost} asymptotically equivalent (RVV, 99)

Frequentist motivations

- nice property $\leadsto$ asymptotic distribution of p_{cpred} and p_{ppost} is $U[0,1]$ for all θ (RVV, 99) ...and for small samples ?
- THEOREM *Let $p(\mathbf{X})$ be any U-conditional predictive p-value. If the distribution of $p(\mathbf{X})$ does not depend on θ, then $p(\mathbf{X})$ is a frequentist p-value for all θ (extra conditions for improper $\pi(\theta)$)*
- Obvious application $\leadsto$ U sufficient $\leadsto$ $m(t|u) = f(t|u)$ and U-conditional predictive p-value $=$ frequentist similar p-value.
- Robert and Rousseau (2002) and Fraser and Rousseau (2008) studied u-conditional p-values for $U =$ MLE, including asymptotic properties, higher order asymptotic and equivalence with ancillary and (repeated) bootstrap p-values

Exponential example

- $X_1, X_2, \ldots, X_n$ i.i.d. $Ex(\lambda)$, with $S = \sum_{i=1}^{n} X_i$
- $T = X_{(1)}$ (lower tail)
- usual noninformative prior $\pi(\lambda) = 1/\lambda$
- **$\mathbf{p_{plug}}$**
  - m.l.e. $\hat{\lambda} = n/S$ and $T \sim Ex(n\lambda)$, so that

$$p_{plug} = e^{-n^2 t_{obs}/s_{obs}}$$

  - conditionally unsatisfactory : for $nt_{obs}/s_{obs} \to 1$ model is clearly contraindicated yet $p_{plug} \to e^{-n}$

– for $\alpha > e^{-n}$,

$$\mathsf{P}r(p_{plug}(\mathbf{X}) \leq \alpha) = \left(1 + \frac{\log \alpha}{n}\right)^{n-1}$$

so $p_{plug}(\mathbf{X})$ is not a frequentist p-value

– but it can be shown to be asymptotically

- $\mathbf{p_{sim}}$

  S is sufficient, $\mathbf{X}|s \sim$ uniform on $\{\mathbf{X} : \sum_{i=1}^n X_i = s\}$

$$p_{sim} = \mathsf{P}r(T > t_{obs}|s_{obs}) = \left(1 - \frac{nt_{obs}}{s_{obs}}\right)^{(n-1)}$$

- **p_{post}**
  - posterior distribution of λ is $Ga(n, s_{obs})$
  - posterior predictive density of T is $\frac{n^2}{s_{obs}} \left(\frac{s_{obs}}{nt+s_{obs}} \right)^{n+1}$
  - posterior predictive p-value

$$p_{post} = \Pr^{m_{post}(t|\mathbf{x}_{obs})}(T > t_{obs}) = \left(1 + \frac{nt_{obs}}{s_{obs}}\right)^{-n}$$

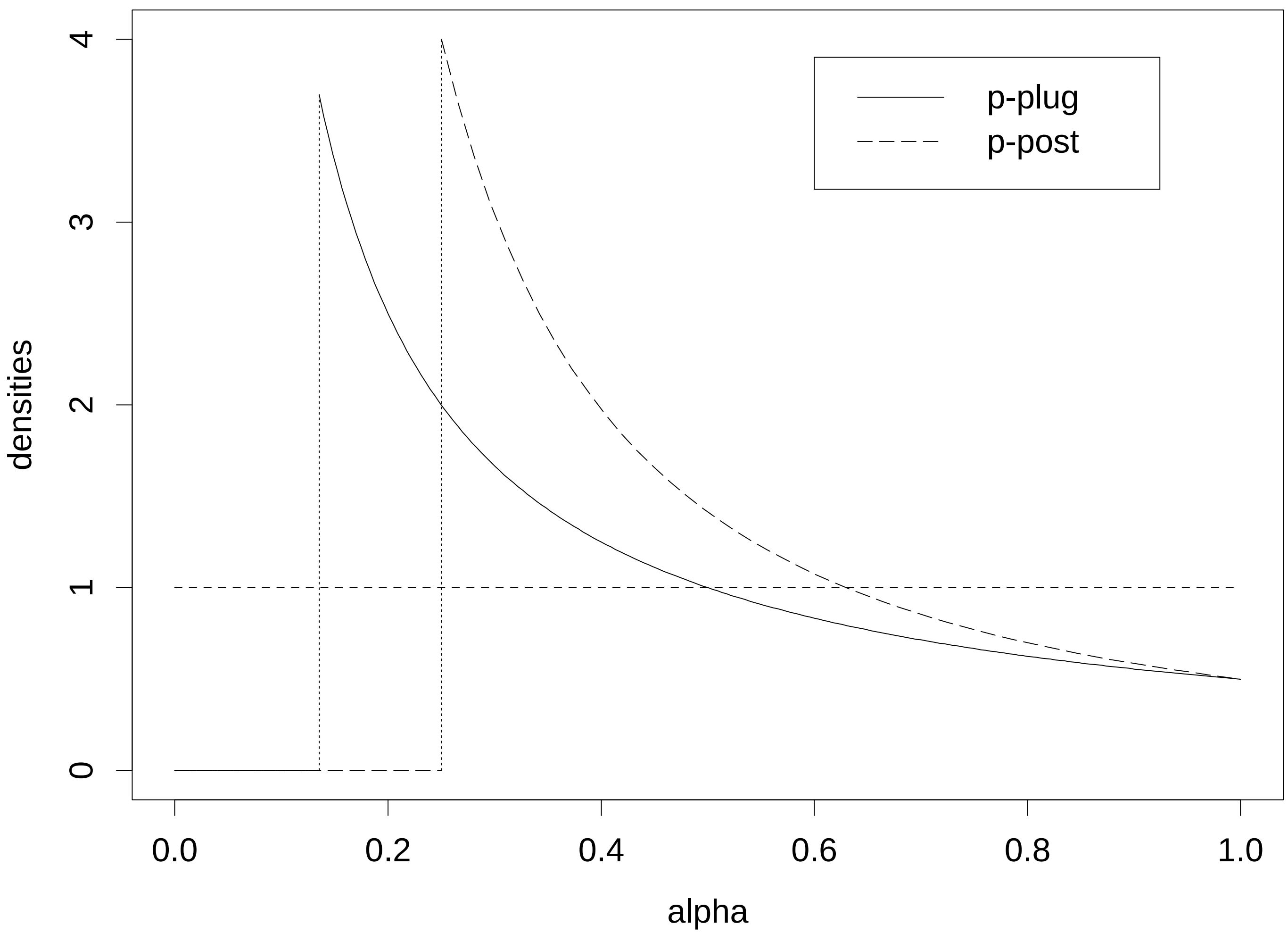
  - conditional behavior not appropriate

$$p_{post} \to 2^{-n} > 0 \ \text{ as } \ nt_{obs}/s_{obs} \to 1$$

  - distribution of p_{post} not $U[0,1]$. For $\alpha > 2^{-n}$,

$$\Pr(p_{post}(\mathbf{X}) \le \alpha) = \left(2 - \alpha^{-1/n}\right)^{n-1}$$

even further from uniformity than p_{plug} ! ! (can be shown to be asymptotically $U[0,1]$)



p-plug
p-post
densities
0
1
2
3
4
0.0
0.2
0.4
0.6
0.8
1.0
alpha

- **p_{ppost}**
  - $f(\mathbf{x} \mid t; \lambda) \propto \lambda^{n-1} \exp\{-\lambda \left(\sum x_i - nt\right)\}$
  - partial posterior for λ

$$\pi(\lambda \mid \mathbf{x}_{obs} \backslash t_{obs}) = \frac{\lambda^{n-2} e^{-\lambda(s_{obs} - nt_{obs})}}{\Gamma(n-1)(s_{obs} - nt_{obs})^{-(n-1)}}$$

  - partial posterior predictive density is

$$m(t \mid \mathbf{x}_{obs} \backslash t_{obs}) = \frac{n(n-1)(s_{obs} - nt_{obs})^{n-1}}{(nt + s_{obs} - nt_{obs})^{n}}$$

  - partial posterior p-value

$$p_{ppost} = \mathsf{P}r^{m(t|\mathbf{x}_{obs} \backslash t_{obs})}(T > t_{obs}) = \left(1 - \frac{nt_{obs}}{s_{obs}}\right)^{n-1}$$

identical to the similar p-value

- It can be shown that $p_{ppost} \to 0$ as $nt_{obs}/s_{obs} \to 1$
- also p_{ppost} is a frequentist p-value for all n

- **$\mathbf{p_{cpred}}$**
  - conditional m.l.e. $\hat{\lambda}_{cMLE} \propto \sum_{i=1}^{n} X_i - nX_1 = S - nT$
  - $\hat{\lambda}_{cMLE}$ is independent of $T \leadsto p_{cpred} = p_{ppost}$
  - derivation of p_{ppost} simpler than that of p_{cpred}
  - P$r(p_{ppost}(\mathbf{X}) \leq \alpha)$ does not depend on λ (Theorem 1) $\leadsto$ p_{cpred} (and p_{ppost} and p_{sim}) frequentist p-value

a curious coincidence

- in examples $p_{sim} = p_{cpred} = p_{ppost}$, even though distributions on completely different (conditional) spaces
- quite useful $\leadsto p_{ppost}$ easier to derive
- THEOREM *If $f(\mathbf{x};\theta)$ (continuous) scale exponential, $S = T + U$ sufficient*

$$f(t, u; \theta) = k\ \theta^{\alpha} t^{\gamma} u^{\alpha-\gamma-2} \exp\{-\theta(t+u)\}$$

with usual noninformative prior, $\pi(\theta) = 1/\theta$

$$p_{cpred} = p_{ppost} = p_{sim}$$

more results in Fraser and Rousseau (2008)

A word about computations

- In general p_{plug} the easiest, then p_{post} then p_{ppost} then p_{cpred}.
- Computation of $\hat{\theta}$ and simulations from posterior predictive $\leadsto$ standard.
- To simulate from $f^*(t) = \int f(t \mid \theta)\, \pi^*(\theta)\, d\theta$:
  - simulate θ from $\pi^*(\theta)$
  - simulate $\boldsymbol{x}$ from $f(\boldsymbol{x} \mid \theta)$ and compute $t = t(\boldsymbol{x})$ (or the p-value)

  where $\pi^*(\theta)$ is the ppost or cpred posterior
- To simulate from $\pi^*(\theta) \leadsto$ M-H (or M-H within Gibbs)

partial posterior p-values

To simulate from $\pi(\theta \mid \mathbf{x}_{obs} \setminus t_{obs}) \propto \dfrac{\pi(\theta \mid \mathbf{x}_{obs})}{f(t_{obs} \mid \theta)}$

- easiest proposal is posterior $\pi(\theta \mid \mathbf{x}_{obs}) \leadsto$ often works, but not when model and data are very incompatible (posterior and partial posterior very distant)

- 'move' (and mix) posterior: If $\theta^* \sim \pi(\theta \mid \mathbf{x}_{obs})$, compute
$$\widetilde{\theta}^* = \theta^* + (\hat{\theta}_{cMLE} - \hat{\theta}_{MLE})$$
$\hat{\theta}_{cMLE} = arg\,max\, f(\boldsymbol{x} \mid t, \theta)$ is conditional MLE
sometimes $\leadsto$ 'mix' with a $U \sim U(0,1)$
when convenient $\leadsto$ log-scale

- moving some factors of $1/f(t_{obs} \mid \theta)$ into $\pi(\theta \mid \mathbf{x}_{obs})$ and renormalizing also works very well when feasible (instead of previous displacement)

- resulting algorithm : `Given` $\widetilde{\theta}^{(t)}$ `at time t,`

  1. `generate` $\theta^* \sim \pi(\theta \mid \boldsymbol{x}_{obs})$
  2. `move` θ^* `to` $\widetilde{\theta}^*$
  3. `acceptance probability:`

$$\alpha = \texttt{min}\left\{1, \frac{\pi(\widetilde{\theta}^* \mid \boldsymbol{x}_{obs}) \;\; f(t_{obs} \mid \widetilde{\theta}^{(t)}) \;\; \pi(\theta^{(t)} \mid \boldsymbol{x}_{obs})}{\pi(\widetilde{\theta}^{(t)} \mid \boldsymbol{x}_{obs}) \;\; f(t_{obs} \mid \widetilde{\theta}^*) \;\; \pi(\theta^* \mid \boldsymbol{x}_{obs})}\right\}$$

- added complication when $f(t \mid \theta)$ not close-form.

u-conditional predictive p-values

For any conditioning statistic U (and in particular for our proposal, $U =$ conditional MLE), $f(\boldsymbol{x} \mid u, \theta)$ is often not available in closed form. General strategy:

- instead of generating from the required $m(\boldsymbol{x} \mid u_{obs})$ we generate from $m(\boldsymbol{x} \mid |u - u_{obs}| < \delta)$
- For small δ this is an approximation to generating from $m(\boldsymbol{x} \mid u_{obs})$ (now called ABC)
- for not so small δ, it can be regarded as a 'less restrictive' conditioning

- Again, for a MH algorithm to simulate from the conditional posterior, the easiest proposal is the usual posterior $\pi(\theta \mid \mathbf{x}_{obs})$, appropriately weighted and re-scaled (if possible) and proposals 'translated' as with *pppp*

- another possibility that works well is a Gibbs-type algorithm:

  `If at time` t `we have the simulations` $(\boldsymbol{x}^{(t)}, \theta^{(t)})$`,`

  1. `Generate` $\theta^{(t+1)} \sim \pi(\theta \mid \boldsymbol{x}^{(t)})$
  2. `Generate` $\boldsymbol{x}^{(t+1)} \sim f(\boldsymbol{x} \mid \theta^{(t)}) 1_{\{|u-u_{obs}|<\delta\}}$ `(that is, simulate repeatedly till` $|u - u_{obs}| < \delta$

Discrete sample spaces

- common analysis is to condition on U for which $f(x|u;\theta)$ does not depend on θ (Fisher Exact Test)
- difficulties
  - conditioning on U yields a severely constrained sample space and serious conservatism of p-values in small or moderate samples
  - choice of T is essentially 'forced' on the user
  - 'conditional issues' in extreme cases

p_{ppost} substantially overcomes these difficulties

2 x 2 contingency tables

	A_1	A_2	Totals
B_1	X_{11}	X_{12}	X_{1+}
B_2	X_{21}	X_{22}	X_{2+}
Totals	X_{+1}	X_{+2}	n

- *Case 1.* One margin $X_{+1} = n_1$, $X_{+2} = n_2$ fixed $\rightsquigarrow$ null model of homogeneity: the two binomial distributions have same success probability θ

- *Case 2.* n fixed; null model is that classification by A and B is independent

Test of homogeneity

- null model: X_{11} and X_{12} are two independent binomial r.v.'s with the same success probability θ

$$f(x_{11}, x_{12}\,; \theta) = \binom{n_1}{x_{11}} \binom{n_2}{x_{12}} \theta^{x_{11}+x_{12}} (1-\theta)^{n-x_{11}-x_{12}}$$

- Fisher exact test $\rightsquigarrow$ conditions on X_{1+} and uses $T = X_{11}$ (textbook choice: essentially forced) resulting in the p-value:

$$p_{fet} = \sum f(t \mid x^o_{1+}) = \sum_{j=t_{obs}}^{\min\{x^o_{1+}, n_1\}} \binom{n_1}{j} \binom{n_2}{x^o_{1+}-j} \Big/ \binom{n}{x^o_{1+}}$$

- for $p_{ppost} \rightsquigarrow$ use the same T as in FET: $T = X_{11}$; this is only for comparison and to judge the power of the methodology ($T = \frac{1}{n_1}X_{11} - \frac{1}{n_2}X_{22}$ would be more sensible unconditionally)

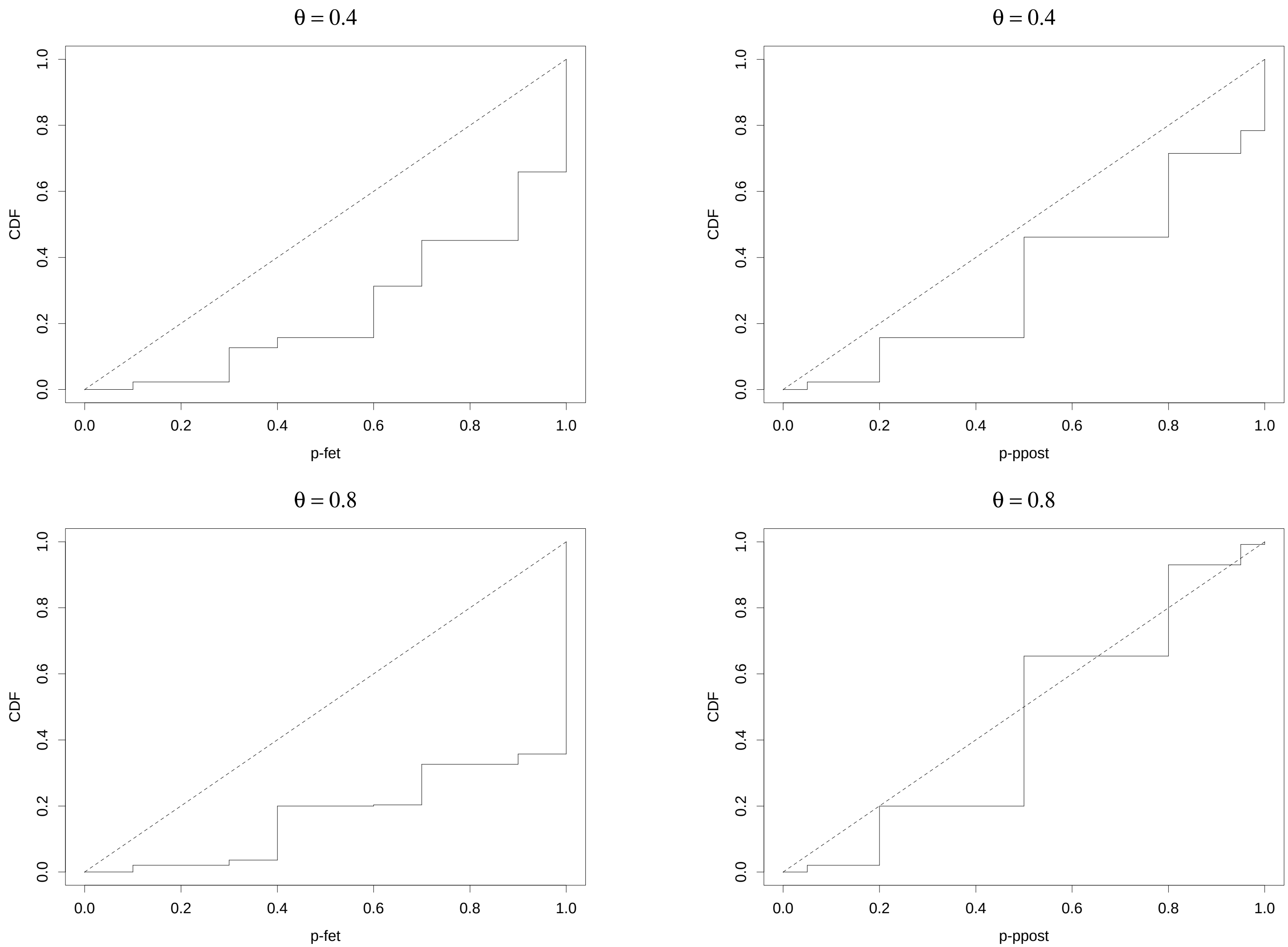
  $\pi(\theta) = 1 \rightsquigarrow$ partial posterior $Beta(x_{12}^o + 1, n_2 - x_{12}^o + 1)$

  $$p_{ppost} = \sum_{j=t_{obs}}^{n_1} \frac{n_2+1}{n_1+1} \left(\begin{array}{c} n_1 \\ j \end{array} \right) \left(\begin{array}{c} n_2 \\ x_{12}^o \end{array} \right) \Big/ \left(\begin{array}{c} n \\ x_{12}^o + j \end{array} \right) .$$

  (here $p_{cpred} = p_{ppost}$)

- specific example $n_1 = 3$ and $n_2 = 2$ (quite extreme case) $\rightsquigarrow$ conditioning on x_{1+} can result in dramatic reduction in the sample space of T (which can have as little as 1 point, or as much as 3); for p_{ppost} this sample space is always $\{0, 1, 2, 3\}$

Distribution functions of p-values p_{fet} (left) and p_{ppost} (right)



Test of independence

- with θ =Pr(A_1) and ξ =Pr(B_1), the null model is

$$f(\mathbf{x};\theta,\xi) = \left(\frac{n!}{x_{11}!x_{12}!x_{21}!x_{22}!}\right)\theta^{x_{+1}}(1-\theta)^{x_{+2}}\xi^{x_{1+}}(1-\xi)^{x_{2+}}$$

- Fisher exact test conditions on both marginals (U) and uses $T = X_{11}$ (forced), with conditional density

$$f(t \mid n, x^o_{1+}, x^o_{+1}) = \begin{pmatrix} x^o_{1+} \\ t \end{pmatrix}\begin{pmatrix} n - x^o_{1+} \\ x^o_{+1} - t \end{pmatrix} \Big/ \begin{pmatrix} n \\ x^o_{+1} \end{pmatrix}$$

leading to the p-value (same as previously)

$$p_{fet} = \sum_{j=t_{obs}}^{\min\{x^o_{1+},n_1\}} \begin{pmatrix} n_1 \\ j \end{pmatrix}\begin{pmatrix} n_2 \\ x^o_{1+} - j \end{pmatrix} \Big/ \begin{pmatrix} n \\ x^o_{1+} \end{pmatrix}$$

- p_{ppost} with same (non optimal) T as in FET
  - with uniform independent priors for θ, ξ

$$p_{ppost} = \int_0^1 \int_0^1 \pi(\theta, \xi \mid \mathbf{x}_{obs} \backslash t_{obs}) \sum_{t=t_{obs}}^{n} Bi(t \mid n, \theta\xi)\, d\theta\, d\xi$$

where the partial posterior is

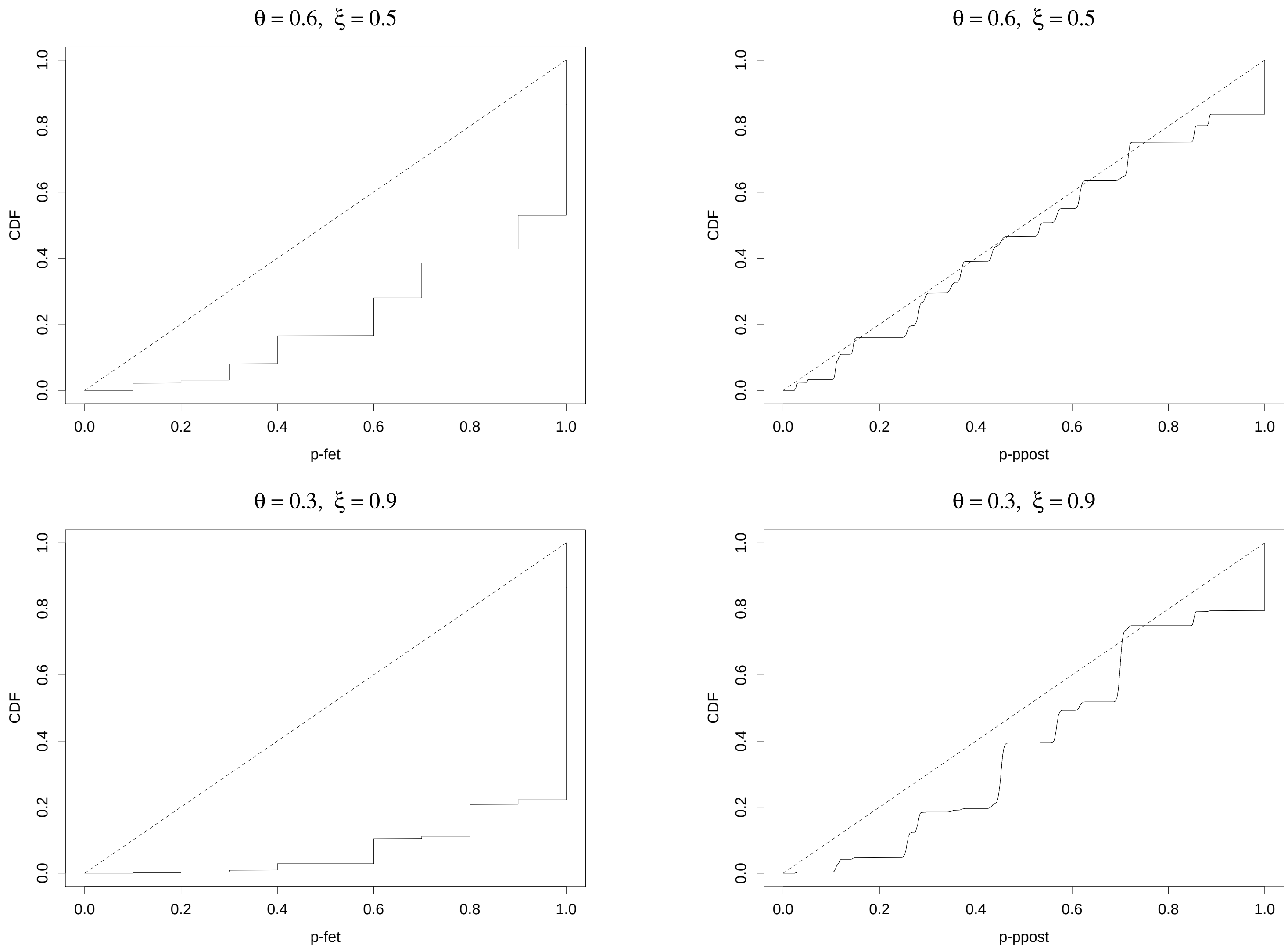
$$\pi(\theta, \xi \mid \mathbf{x}_{obs} \backslash t_{obs}) \propto \theta^{x^o_{21}} (1-\theta)^{x^o_{+2}} \xi^{x^o_{12}} (1-\xi)^{x^o_{2+}} (1-\theta\xi)^{-(n-t_{obs})}$$

  - computation via importance sampling w.r.t.

$$\frac{1}{2}\, Un(\theta \mid 0,1)\, Be(\xi \mid x^o_{12}+1, x^o_{22}+1) + \frac{1}{2}\, Be(\theta \mid x^o_{21}+1, x^o_{22}+1)\, Un(\xi \mid 0,1)$$

easy generation and highly efficient computationally

- particular example
  - $n = 5$
  - support of $p_{fet}(\mathbf{X})$ is $\{0.1, 0.2, 0.3, 0.4, 0.6, 0.7, 0.8, 0.9\}$, support of $p_{ppost}(\mathbf{X})$ noticeably richer
  - next figure gives cdfs of $p_{fet}(X)$ and $p_{ppost}(X)$; if uniform, these would be $F(p) = p$.


θ = 0.6, ξ = 0.5
CDF
p-fet
θ = 0.6, ξ = 0.5
CDF
p-ppost
θ = 0.3, ξ = 0.9
CDF
p-fet
θ = 0.3, ξ = 0.9
CDF
p-ppost
0.0
0.2
0.4
0.6
0.8
1.0

- how large does n need to be for the p-values to be approximately uniform?
  - sample size needed for cdf of a p-value at 0.05 to be within 20% of 0.05
    * when $(\theta, \xi) = (0.6, 0.5)$,
      · $p_{fet}(\mathbf{X}) \approx U[0,1]$ when $n \approx 500$;
      · $p_{ppost}(\mathbf{X}) \approx U[0,1]$ when $n \approx 10$
    * when $(\theta, \xi) = (0.3, 0.9)$,
      · $p_{fet}(\mathbf{X}) \approx U[0,1]$ when $n \approx 1200$,
      · $p_{ppost}(\mathbf{X}) \approx U[0,1]$ when $n \approx 110$

a bad choice: T $\approx$ sufficient

- apparent breakdown of both p_{fet} and p_{ppost} for large values of (θ, ξ)
- $p_{fet} \rightsquigarrow$ hopelessly conservative $\rightsquigarrow$ never stating that data incompatible with model
- $p_{ppost} \rightsquigarrow$ markedly anti-conservative
- At a purely intuitive level, the behavior of p_{ppost} is quite sensible
  - we declare that large values of T means evidence against the null model
  - when (θ, ξ) both large, $T = X_{11}$ is typically very large (leading to rejection)
  - p_{ppost} exhibits exactly this behavior

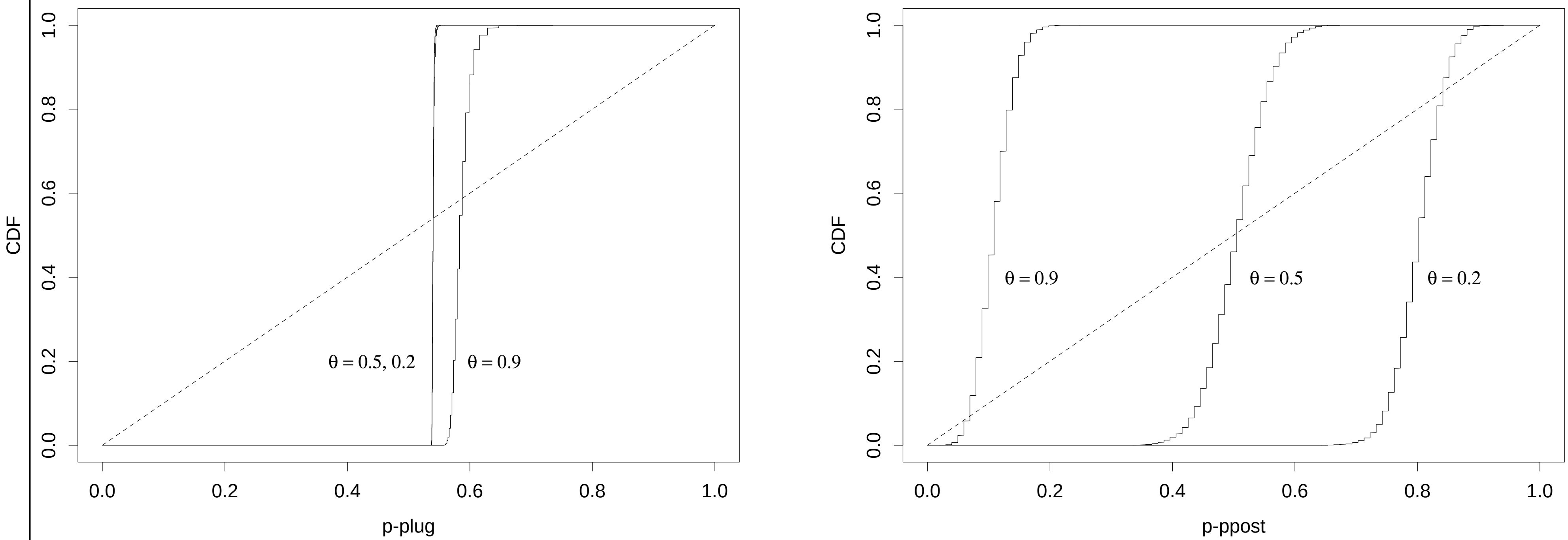
- anti-conservative behavior of p_{ppost} arises because a very large T provides a great deal of information about the parameters, but little information about deviance from the model
- Most extreme example arises when T sufficient, a choice that is nearly useless for model checking
- intuitively, choice of a sufficient statistic for T allocates all the information from the data to learn about the unknown parameters, leaving none to judge model inadequacy
- even in this extremely bad scenario, p_{ppost} seems to convey some information, whereas p_{plug} and p_{post} are useless

example with T sufficient

- $X_i \sim Ber(\theta)$, $T = \sum X_i$, a sufficient statistic
- $p_{ppost} = 1 - t_{obs}/(n+1)$
  - for large n, distribution of p_{ppost} tightly concentrates around $1 - \theta$
  - entirely natural behavior: large $T \approx$ large values of θ and declared to be 'surprising'
- p_{plug} and p_{post}
  - distributions of both concentrate tightly about $1/2$ when n is large for all θ
  - provide completely useless answers here

- the natural requirement for Bayesians $\rightsquigarrow$ require a p-value to be uniform under the prior predictive distribution
  - p_{ppost} is a p-value for a Bayesian $\rightsquigarrow$ 'average' of all the distribution functions of p_{ppost} is uniform
  - no Bayesian averages of the distribution functions of p_{plug} (or p_{post}) can be uniform

CDF's of p-values for sufficient T



What about U $\approx$ sufficient in $p_{cpred(u)}$?

- Remember: for $p_{cpred(u)}$ the 'distribution of reference' was $f(t \mid u)$, with T measuring departure from the entertained model
- It was suggested that optimal choice of U for a given T would be to have $(T, U) \approx$ sufficient, with U 'overlapping' as little as possible with T
- Our proposal was to use $U = \hat{\theta}_c$ the conditional MLE (that is, the MLE of θ from $f(\boldsymbol{x} \mid T = t_{obs}, \theta)$ and the resulting p-value is p_{cpred}
- Robert and Rousseau (03) and Fraser and Rousseau (08) suggest use of $U = \hat{\theta}$, the MLE of θ from $f(\boldsymbol{x} \mid \theta)$
- When $\hat{\theta}$ is sufficient (or nearly so), this makes any T ancillary in the conditional distribution $f(\boldsymbol{x} \mid u, \theta)$ (or nearly so), and hence

p-values (for any T) are approximately uniform

- Nothing is wrong with this except that Bayesian analysis is not really required, in that the 'recentering' of T is done through conditioning on a sufficient statistic, that is, by computing the (frequentist) p_{sim}
- We suspect (work in progress) that, for small n and when $\hat{\theta}$ is not sufficient
  - this choice might be too much conditioning (the discrete sample space gives a hint),
  - power might be an issue,
  - p-values are further from Uniformity than those from our original definition of p_{cpred}

Checking a Gamma distribution

- Entertained model: $X_1, \ldots, X_n \sim Ga(\alpha, \beta)$

$$f(x \mid \alpha, \beta) \propto x^{\alpha-1} e^{-x/\beta}$$

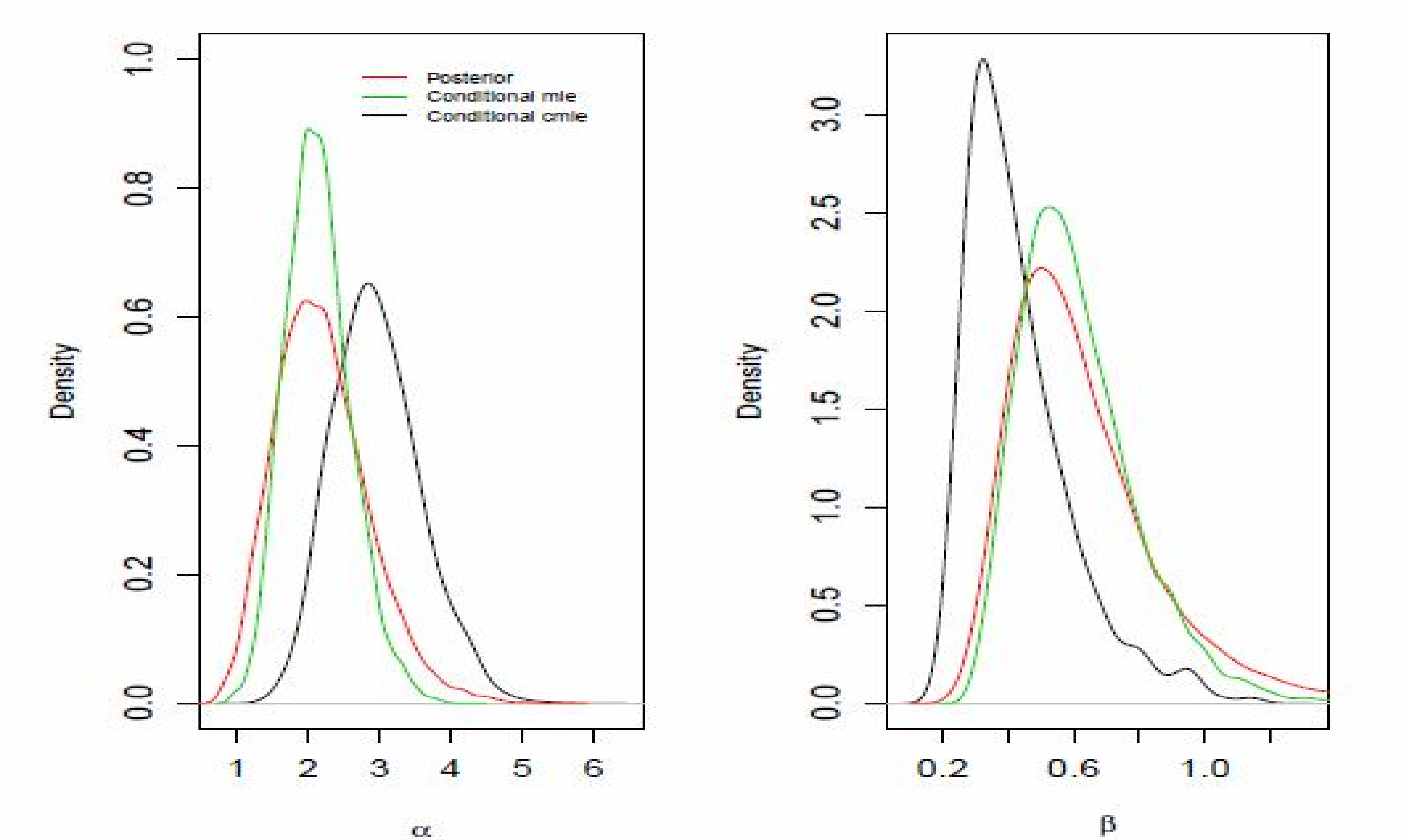
  with α shape and β scale parameters

- Use Jeffrey's prior for $\boldsymbol{\theta} = (\alpha, \beta)$
- Let the departure statistic be $T = \max(X_1, \ldots, X_n)$
- Compare model checks carried in the following distributions:
  - The posterior predictive
  - The conditional predictive (U-conditional with $U =$ the conditional MLE of $\boldsymbol{\theta}$)
  - The U-conditional, with $U =$ the unconditional MLE of $\boldsymbol{\theta}$

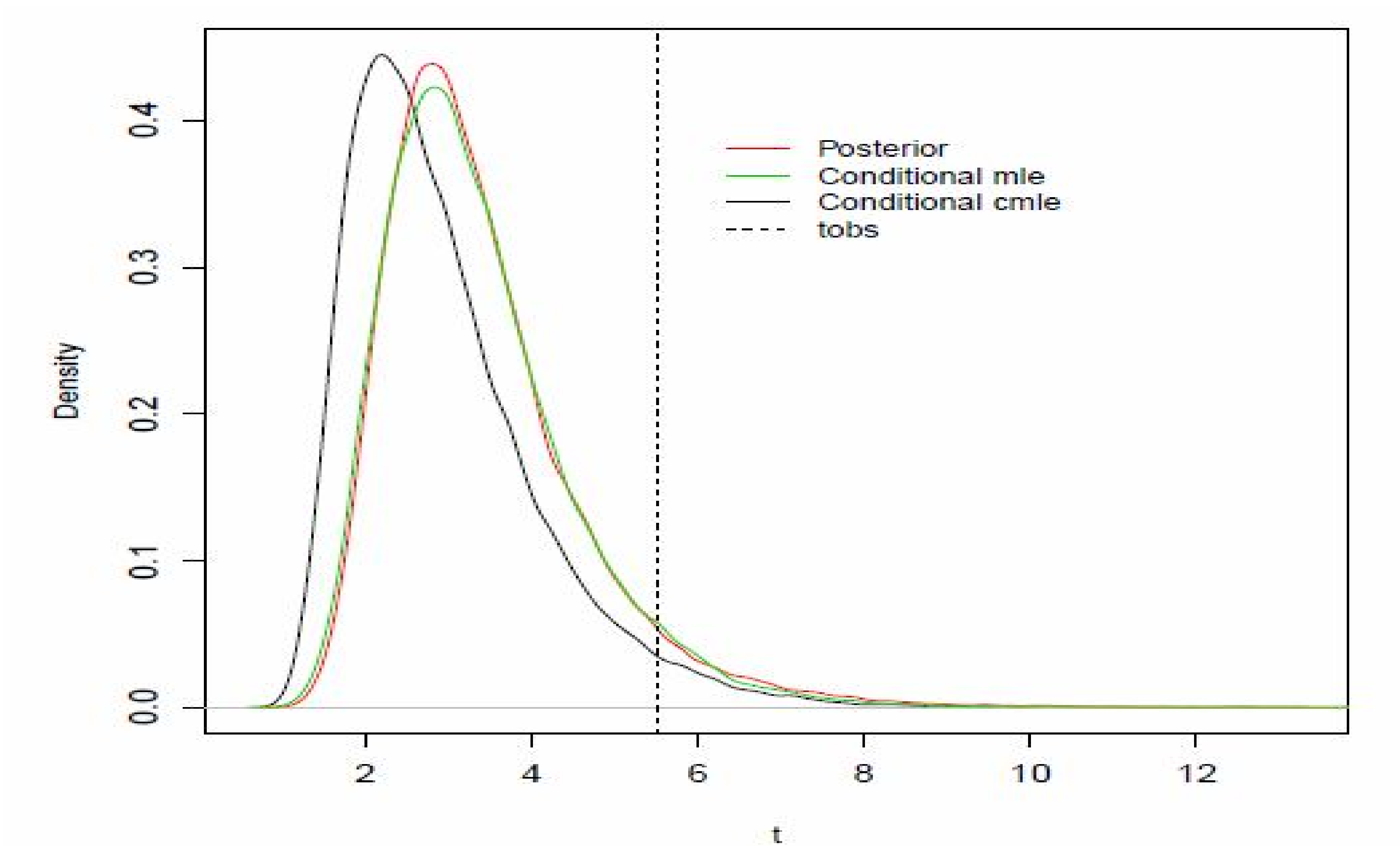
- a little simulated example
  - generate 19 observations from a $Ga(3,3)$ and then add a very extreme observation equal to 5. Ordered data is:

    0.36, 0.37, 0.42, 0.55, 0.56, 0.62, 0.69, 0.74, 0.94, 0.95,1.28, 1.29, 1.39, 1.44, 1.52, 1.58, 1.85, 1.87, 1.87, 5
  - simulate behavior under the null with 500 replicates for $n = 50$

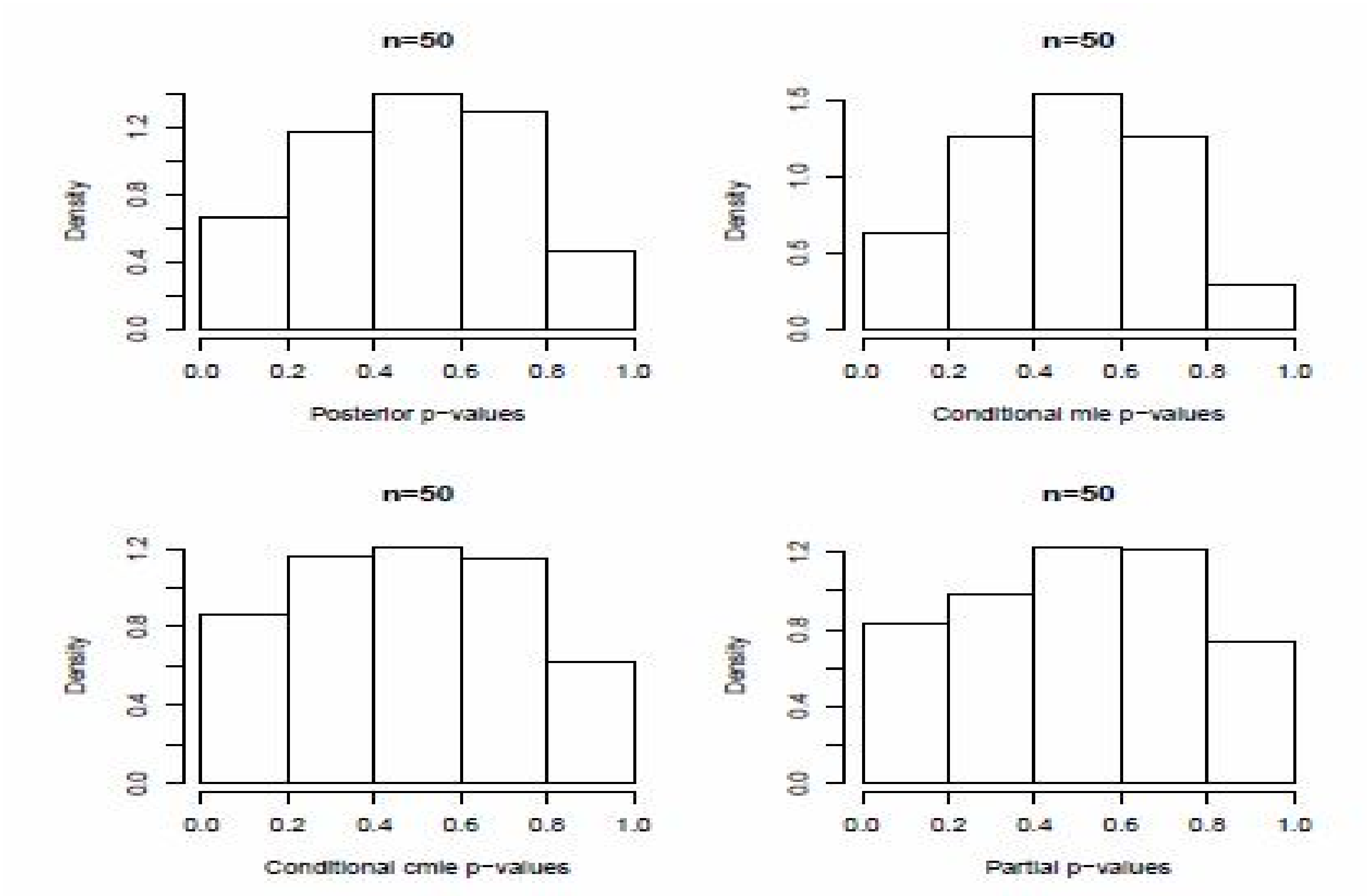
posterior and u-conditional posterior distributions



posterior and u-conditional predictive distributions



distribution of p-values under the null



Normal hierarchical models

Rest of talk: model is usual normal-normal hierarchical model with k groups:

$$\begin{array}{lcll} X_{ij} \,|\, \mu_i & \overset{i}{\sim} & N(\mu_i, \sigma_i^2) & \text{for } i = 1, \ldots, k, \quad j = 1, \ldots, n_i \\ \mu_i \,|\, \nu, \tau & \overset{i}{\sim} & N(\nu, \tau^2) & \text{for } i = 1, \ldots, k\ . \end{array}$$

- use previous ways to get rid of (hyper)parameters (the prior for the means is ‘agreed upon’, and hence part of the ‘model’). Variances σ_i^2 assumed known sometimes
- Investigate different ‘nulls’

Checking the 'hypermean'

To fix ideas, begin with an easy one: testing a specified value for the "great mean" (and sometimes it is even of interest)

- recall $\quad X_{ij} \mid \mu_i \overset{i}{\sim} N(\mu_i, \sigma^2), \quad \mu_i \mid \nu, \tau \overset{i}{\sim} N(\nu, \tau^2)$
  assume k groups, n observations per group, same σ^2 (known)
- to test $H_0 : \nu = \nu_0$
- an intuitive T : $T = \frac{\sum_{i=1}^{k} \overline{X}_{i\cdot}}{k}$
- p-value: $p = Pr^{f(t)}\{\, |T - \nu_0| \geq |t_{obs} - \nu_0|\,\}$
- distribution of T : $f(t \mid \boldsymbol{\mu}\,) = N(t \mid \frac{\sum_{i=1}^{k} \mu_i}{k},\ \frac{\sigma^2}{k\,n})$

integrate $\boldsymbol{\mu}$ out (random effects) w.r.t. several distributions

Empirical Bayes (plug-in)

let $\hat{\tau}$ the MLE from $f(\boldsymbol{x} \mid \tau^2) = \int f(\boldsymbol{x} \mid \boldsymbol{\mu})\pi(\boldsymbol{\mu} \mid \tau^2)d\boldsymbol{\mu}$

Consider *two* EB distributions for $\boldsymbol{\mu}$:

- $\pi^{EB}(\boldsymbol{\mu}) = \pi(\boldsymbol{\mu} \mid \hat{\tau}^2) = \pi(\boldsymbol{\mu} \mid \tau^2 = \hat{\tau}^2)$
  producing $m^{EB}_{prior}(t) = \int f(t \mid \boldsymbol{\mu})\pi^{EB}(\boldsymbol{\mu})d\boldsymbol{\mu}$
- $\pi^{EB}(\boldsymbol{\mu} \mid \boldsymbol{x}_{obs}) \propto f(\boldsymbol{x}_{obs} \mid \boldsymbol{\mu})\pi^{EB}(\boldsymbol{\mu})$
  producing $m^{EB}_{post}(t) = \int f(t \mid \boldsymbol{\mu})\pi^{EB}(\boldsymbol{\mu} \mid \boldsymbol{x}_{obs})d\boldsymbol{\mu}$

Note use of $\pi^{EB}(\boldsymbol{\mu} \mid \boldsymbol{x}_{obs})$ is clearly inappropriate, making an obvious double use of the data. We'll see that it exhibits identical behavior to posterior predictive checks.

Comparing both EB predictive $m(t)$

Prior is $N\left(\nu_0,\ \frac{1}{k}\left(\frac{\sigma^2}{n} + \hat{\tau}^2\right)\right)$

Posterior is $N\left((1-\alpha)t_{obs} + \alpha\,\nu_0,\ \alpha\,\frac{1}{k}\left(\frac{\sigma^2}{n} + 2\,\hat{\tau}^2\right)\right)$

with $\alpha \to 0$ as $n \to \infty$ (or as $\hat{\tau}^2 \to \infty$)

assume now that $t_{obs} \to \infty$ (model very wrong)

$$m^{EB}_{prior}(t) \longrightarrow N(\nu_0, \infty)$$

$$m^{EB}_{post}(t) \longrightarrow N(t_{obs}, \frac{2\sigma^2}{k\,n})$$

inadequacy of $m^{EB}_{post}(t)$ for model checking is obvious, and hence the p-value (or graphical checks, or whatever) will also be seriously inadequate.

posterior and partial posterior distributions

With prior $\pi(\tau^2) \propto 1/\tau$, we use Gibbs to simulate from both $\pi_{post}(\boldsymbol{\mu}, \tau^2 \mid \boldsymbol{x}_{obs})$ and $\pi_{ppp}(\boldsymbol{\mu}\ , \tau^2 \mid \boldsymbol{x}_{obs} \setminus t_{obs})$

- full conditional of τ^2 is common (n.c. χ^2)
- full conditionals of μ_i are N
  - means

    post $\rightsquigarrow (1-\alpha)\,\bar{x}_i + \alpha\,\nu_0 \quad$ (independent)

    ppp $\rightsquigarrow (1-\alpha^*)\left[\bar{x}_i + \bar{\mu}_{rest} - \bar{x}_{rest}\right] + \alpha^*\,\nu_0$
  - 1/variances

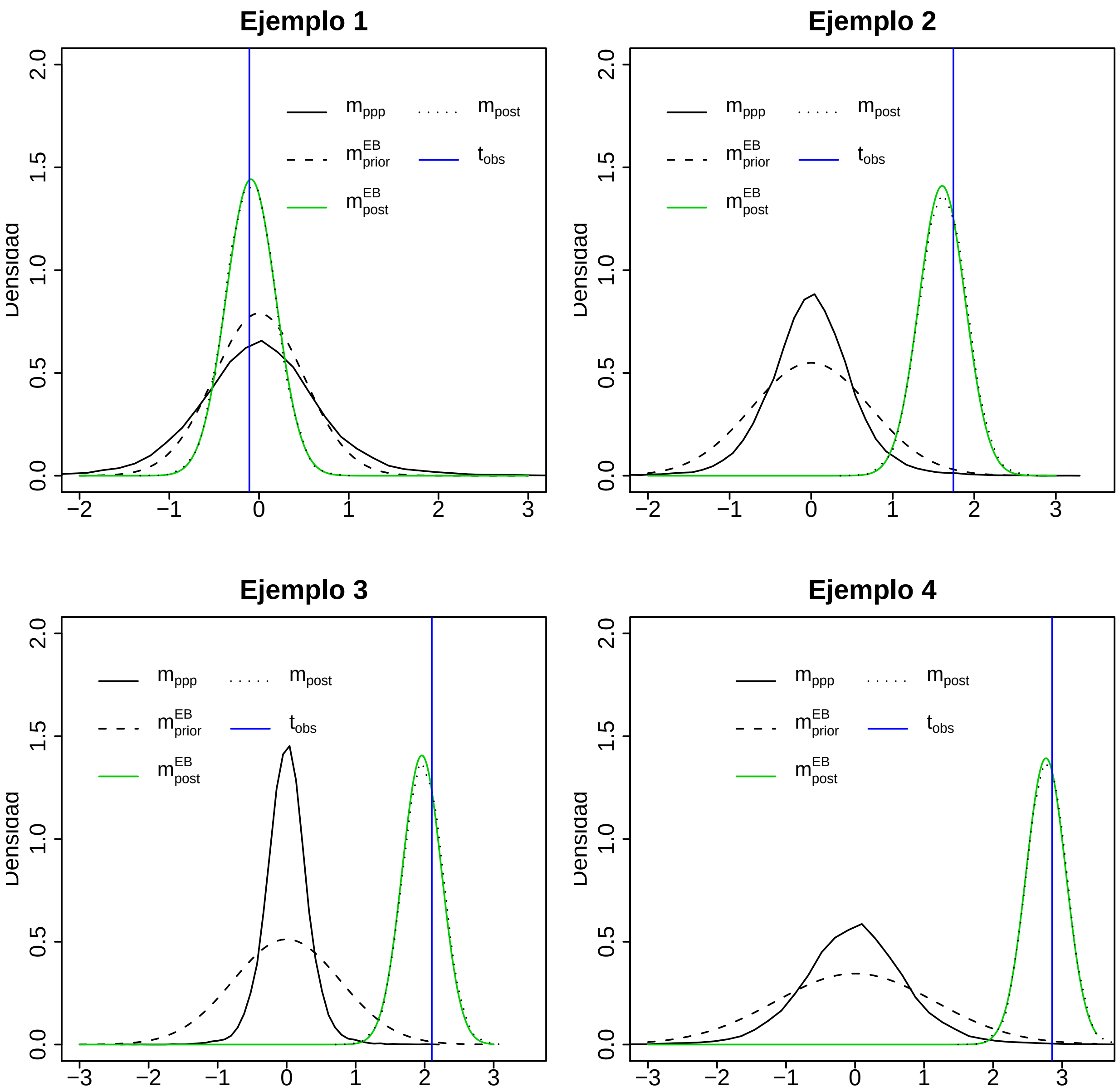
    post $\rightsquigarrow \frac{1}{\sigma^2} + \frac{1}{\tau^2}$

    ppp $\rightsquigarrow \frac{k-1}{k}\ \frac{1}{\sigma^2} + \frac{1}{\tau^2}$

Examples

- 4 simulated examples, $k = 8$ groups, $n = 12$ observations per group
- in all of them, test $H_0 : \nu = 0$ ($\nu =$ mean of μ_i's)
- $X_{ij} \sim N(\mu_i, 4)$
  - $\mu_i \sim N(0, 1)$ in Example 1 (H_0 true)
  - $\mu_i \sim N(1.5, 1)$ in Example 2 (H_0 not true)
  - $\mu_i \sim N(2.5, 1)$ in Example 3 (H_0 not true)
  - $\mu_i \sim N(2.5, 3)$ in Example 4 (H_0 not true)

	Ex. 1	**Ex. 2**	**Ex. 3**	**Ex. 4**
ppp	0.859	0.008	0.000	0.005
EB prior	0.831	0.016	0.007	0.013
EB post	0.711	0.313	0.305	0.378
post	0.712	0.333	0.325	0.392


Ejemplo 1
Ejemplo 2
Ejemplo 3
Ejemplo 4
Densidad
m_{ppp}
m_{post}
m^{EB}_{prior}
t_{obs}
m^{EB}_{post}

Checking the second level

- recall

$$\begin{aligned} X_{ij} \,|\, \mu_i &\overset{i}{\sim} N(\mu_i, \sigma^2) \\ \mu_i \,|\, \nu, \tau &\overset{i}{\sim} N(\nu, \tau^2) \end{aligned}$$

  k groups, n observations per group, same σ^2

- to test the second level of the hierarchy

- intuitive, easy to work with $T = \max\{\bar{X}_1, \ldots, \bar{X}_k\}$

- p-value: $p = Pr^{f(\bullet)}\{\,T \geq t_{obs}\}$

- priors (prior for σ^2 when unknown)

$$\begin{aligned}\pi(\sigma^2) &\propto \frac{1}{\sigma^2}\\ \pi(\nu \mid \tau^2) &\propto 1\\ \pi(\tau^2) &\propto \frac{1}{\tau}\end{aligned}$$

- all distributions and p-values require MC or MCMC

Example

Assume a simulated example with 5 groups, 8 observations per group and

$$
\begin{aligned}
X_{ij} \mid \mu_i &\sim N(\mu_i, 4) \quad \text{for } i = 1, \ldots, 5 \qquad j = 1, \ldots, 8 \\
\mu_i &\sim N(1, 1) \quad \text{for } i = 1, \ldots, 4 \\
\mu_5 &\sim N(5, 1)
\end{aligned}
$$

sample means: $1.56,\ 0.64,\ 1.98,\ 0.01,\ 6.96$

(The mean of the 5th group is 6.65 SD away from the others)

	p_{ppp}	p^{EB}_{prior}	p^{EB}_{post}	p_{post}
σ^2 known	.010	.130	.347	.409
σ^2 unknown	.015	.195	.371	.405

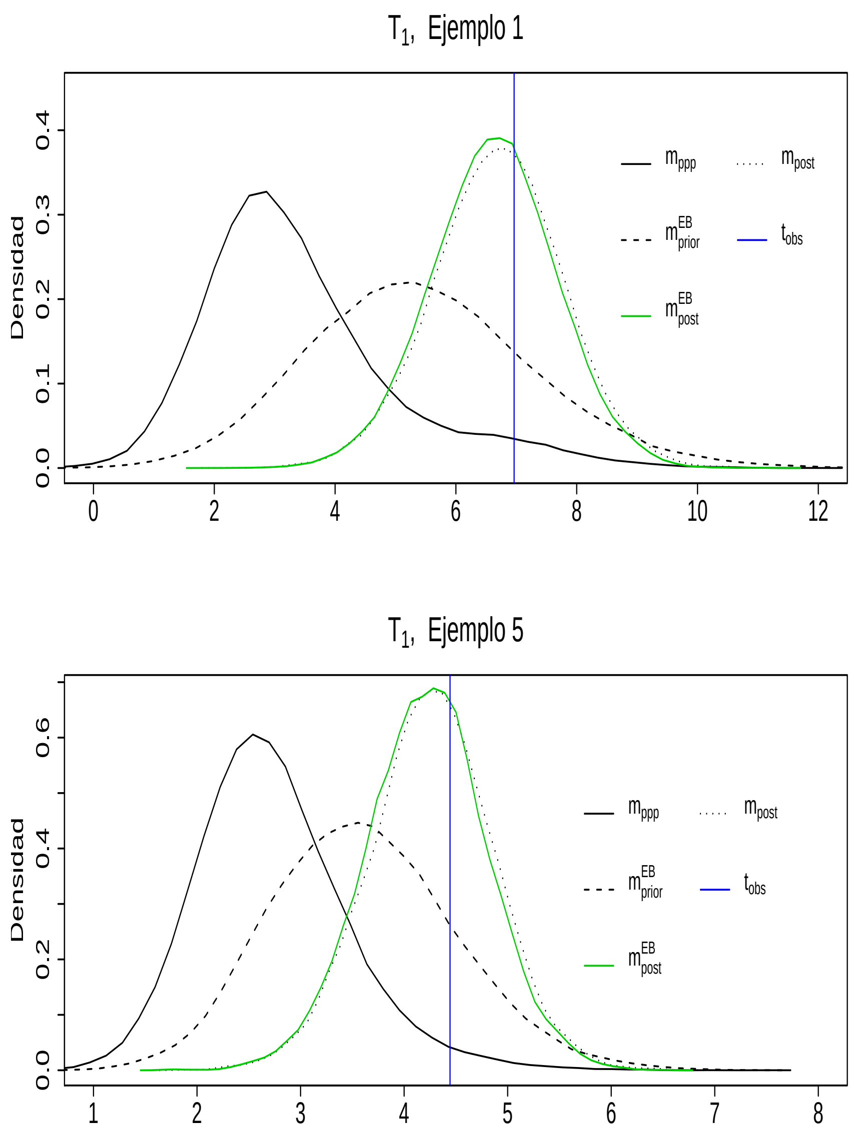


Figure 1: σ^2 unknown

behavior under the null

- Assume $X_1, X_2, \ldots, X_n$ i.i.d. $f(x \mid \theta) \rightsquigarrow T \sim f(t \mid \theta)$
- for known θ (or ancillary T)

$$p = p(\boldsymbol{X}) \sim U(0,1)$$

  pretty convenient $\rightsquigarrow$ same meaning across problems

  also $\rightsquigarrow$ defining property of a p-value

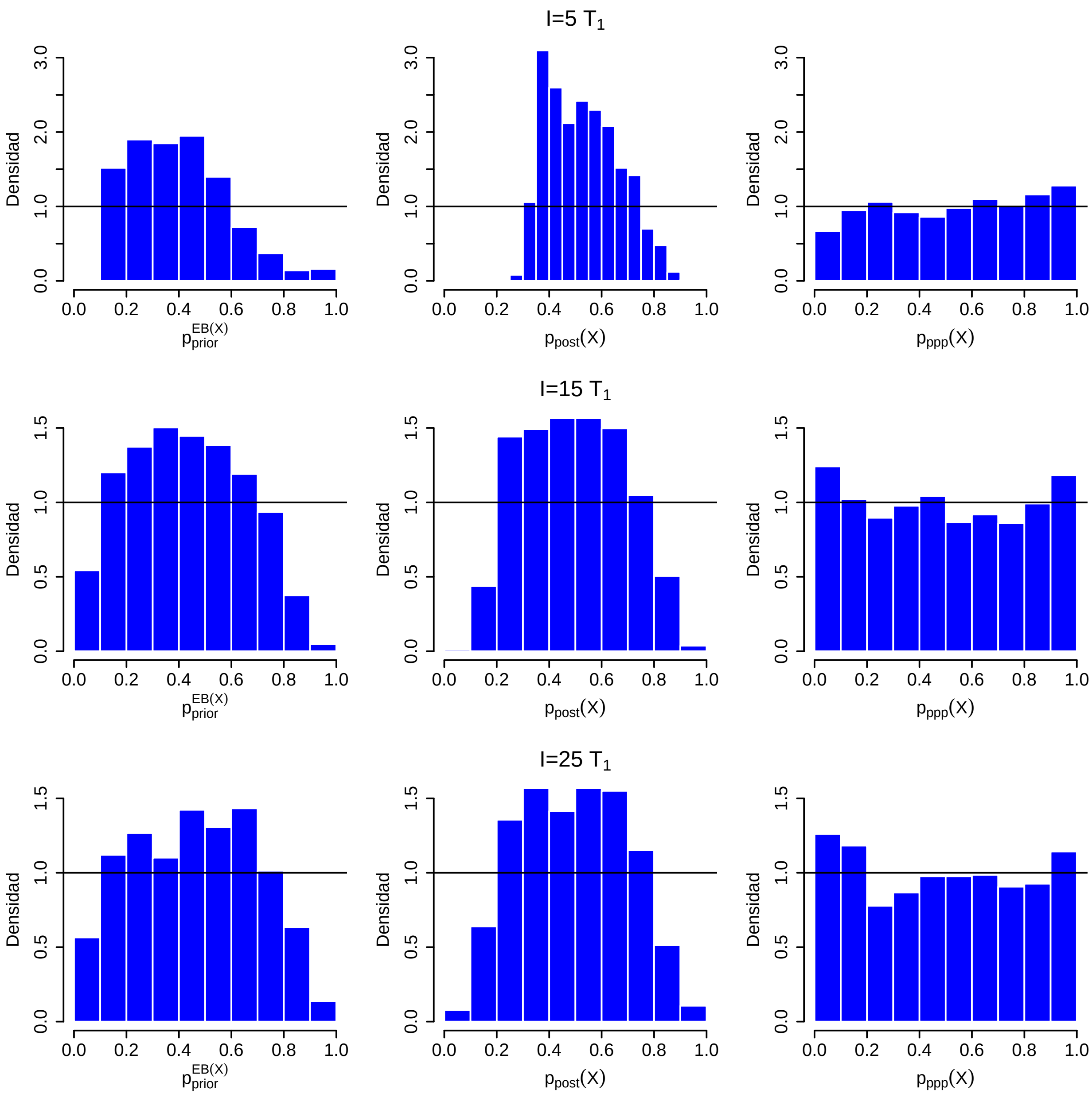
- for unknown $\theta \rightsquigarrow p(\boldsymbol{X}) \sim U(0,1)$ for all θ usually not possible $\rightsquigarrow$ require $p(\boldsymbol{X}) \sim U(0,1)$ asymptotically (RVV, 2000), and approximately so for finite n

- RESULT: for asymptotic normal T, the *only* p-value which is asymptotically $Un(0,1)$ is p_{ppp} (RVV, 00). Also, it is most powerful against Pittman's alternatives. Also, p_{plug} and p_{post} are conservative.

- here T not asympt. N, and also want to exemplify behavior for small/moderate $n \rightsquigarrow$ simulation

- pictures $\rightsquigarrow$ consider $p_{plug}(\boldsymbol{X}\ ), p_{post}(\boldsymbol{X}\ ), p_{ppp}(\boldsymbol{X}\ )$ as R.V. $\rightsquigarrow$ simulate $\boldsymbol{X}$ under the null model, represent density of the p-values $\rightsquigarrow$ should be $\approx U(0,1)$

- null: $X_{ij} \mid \mu_i \sim N(\mu_i, 4), \quad \mu_i \sim N(0,1)$
  $k = 5, 15, 25$ groups, 8 observations per group.

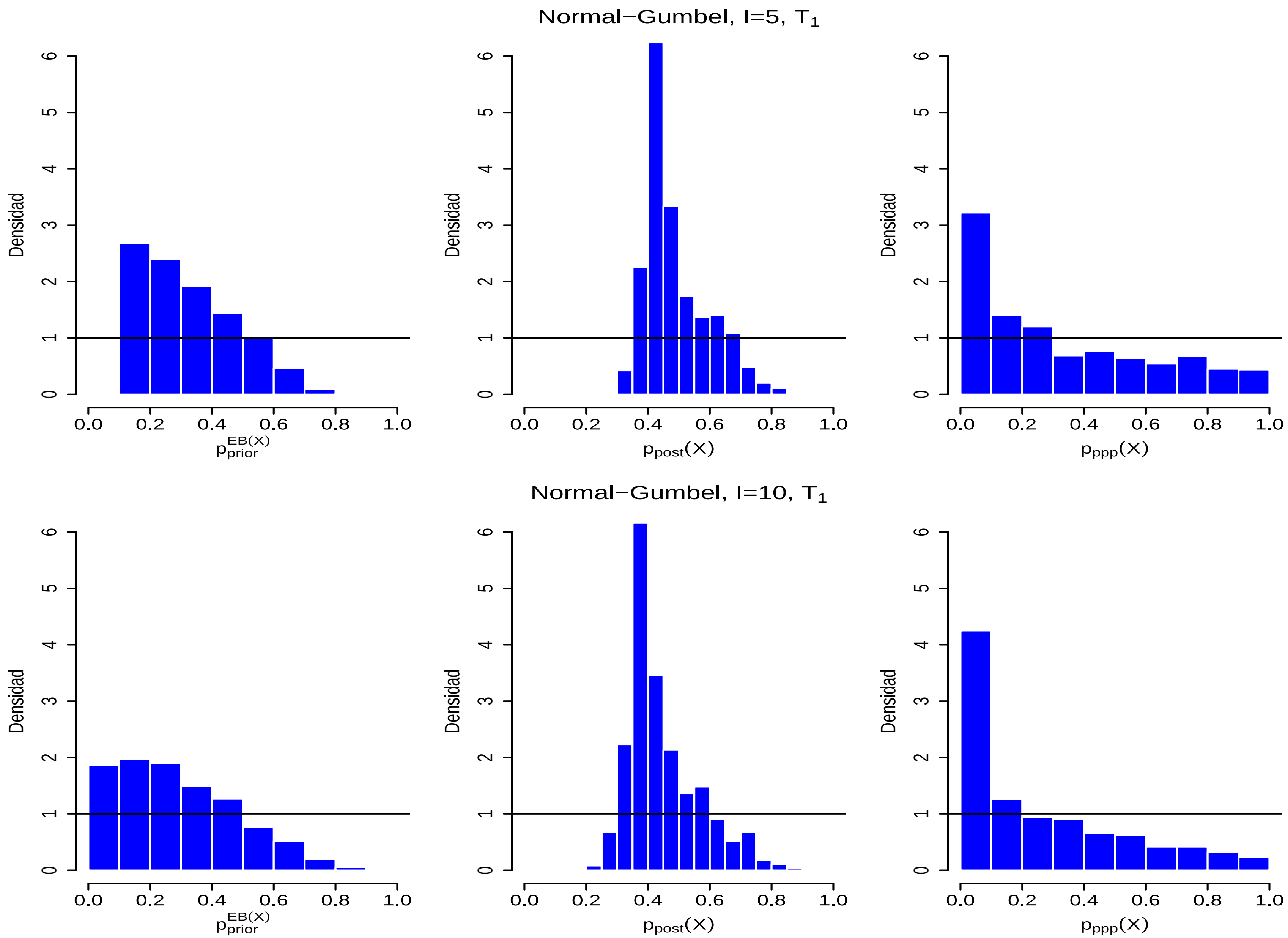

I=5 T_1
I=15 T_1
I=25 T_1
Densidad
$p^{EB(X)}_{prior}$
$p_{post}(X)$
$p_{ppp}(X)$

behavior under alternatives

- “null model”: $X_{ij}\,|\,\mu_i \stackrel{i}{\sim} N(\mu_i,\sigma^2), \quad \mu_i\,|\,\nu,\tau \stackrel{i}{\sim} N(\nu,\tau^2)$
- To explore behavior of $p_{plug}(\boldsymbol{X}\ ), p_{post}(\boldsymbol{X}\ ), p_{ppp}(\boldsymbol{X}\ )$ when “null model” not true $\rightsquigarrow$ POWER
- concentrate in ‘wrong’ second level: simulate X_{ij} from normal and μ_i from non-normal
- First level: $X_{ij} \mid \mu_i \sim N(\mu_i, 4)$
  $n = 8$ observations per group, $k = 5, 10$ groups
- second level: $\mu_i \sim Gumbel(0, 2)$ (similar results with exponential and log-normal, B&C)

$\mathsf{Pr}(p - value \leq \alpha)$

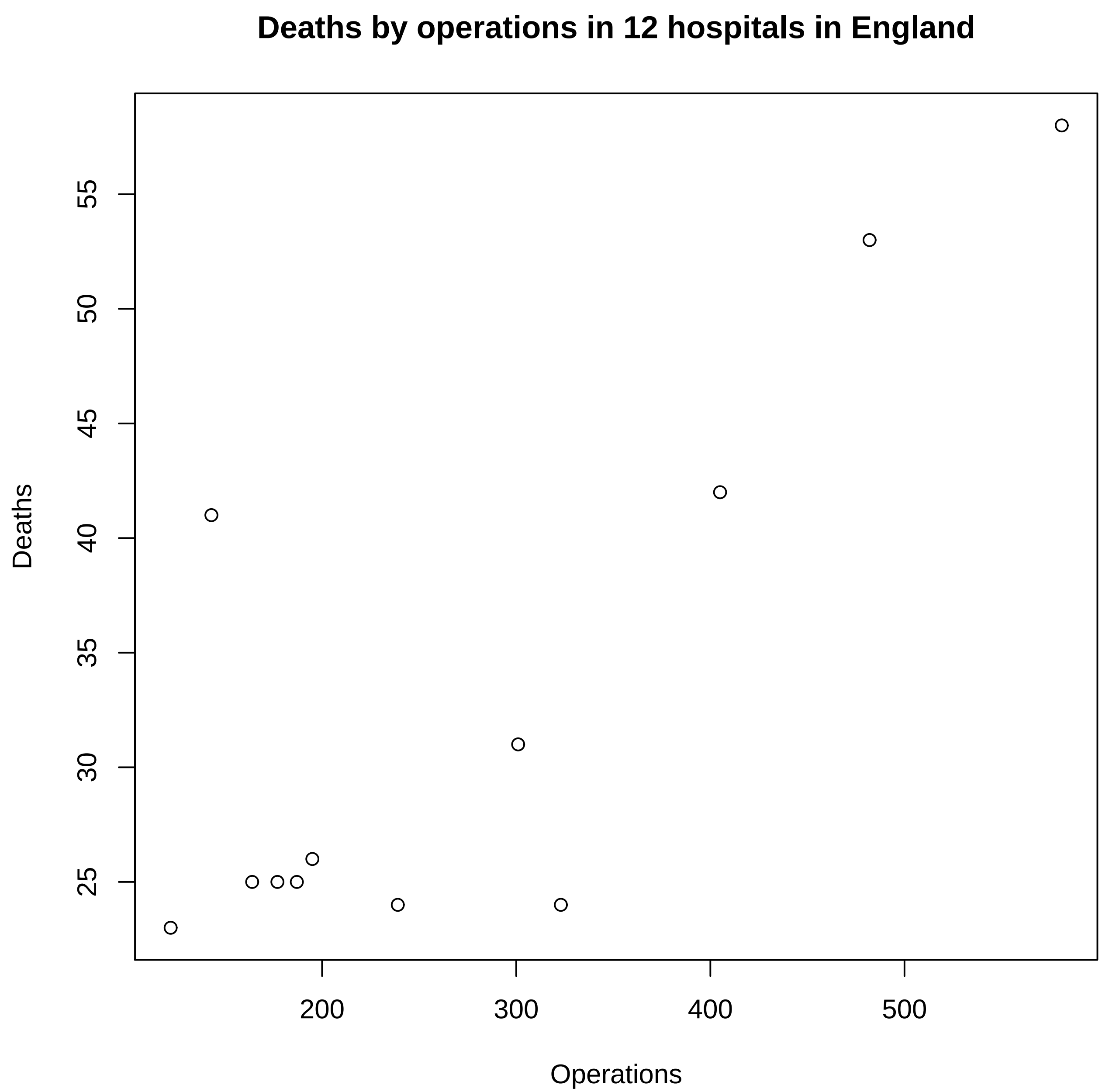
α	0.02	0.05	0.1	0.2
Normal-Gumbel				
k=5				
p_{ppp}	0.124	0.219	0.322	0.462
p_{post}	0.000	0.000	0.000	0.000
p_{prior}^{EB}	0.000	0.000	0.000	0.268
k=10				
p_{ppp}	0.208	0.314	0.425	0.550
p_{post}	0.000	0.000	0.000	0.003
p_{prior}^{EB}	0.001	0.067	0.187	0.383


Normal–Gumbel, I=5, T1
Normal–Gumbel, I=10, T1
Densidad
0 1 2 3 4 5 6
0.0 0.2 0.4 0.6 0.8 1.0
p EB(X) prior
ppost(X)
pppp(X)

Binomial-Beta model example: Bristol Royal Infirmary Inquiry data

Real example: number n_i of open-heart operations and the corresponding number Y_i of deaths of children under 1 year in 12 hospitals in England, (Spiegelhalter et al. 2002).

$$
\begin{aligned}
Y_i \mid \theta_i &\overset{i}{\sim} \mathsf{Bin}(\theta_i, n_i), \quad i = 1, \ldots, I, \\
\pi(\boldsymbol{\theta} \mid \alpha, \beta) &= \textstyle\prod_{i=1}^{I} \mathsf{Beta}(\theta_i \mid \alpha, \beta)\,, \\
\pi(\alpha, \beta) &\propto \text{Jeffreys prior} \qquad \text{(Yang and Berger 87)}
\end{aligned}
$$

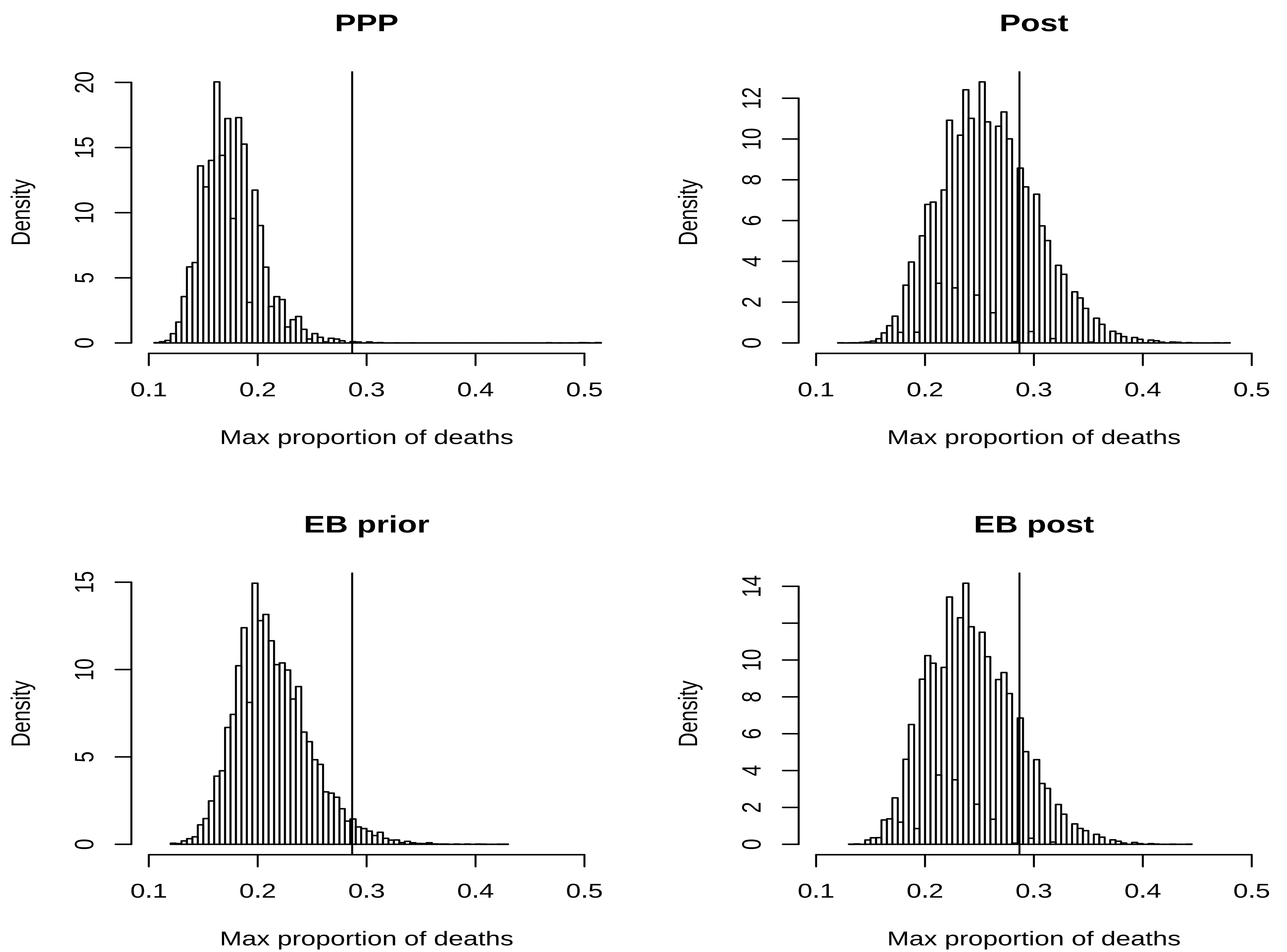

Deaths by operations in 12 hospitals in England
Deaths
55
50
45
40
35
30
25
200
300
400
500
Operations

- As departure statistics we use:

$$\text{Max}\left\{\frac{y_i}{n_i}\right\} \text{ and Min}\left\{\frac{y_i}{n_i}\right\}$$

- To approximate the ppp distribution we use the normal approximation to the binomial.

	p^{EB}_{prior}	p^{EB}_{post}	p_{post}	p_{ppp}
Maximum	0.03	0.16	0.23	0.00
Minimum	0.67	0.56	0.62	0.64


PPP
Post
EB prior
EB post
Density
Max proportion of deaths
0.1 0.2 0.3 0.4 0.5
0 5 10 15 20
0 2 4 6 8 10 12
0 5 10 15
0 2 4 6 8 10 14

Other methods are reviewed and discussed in B&C

- *Simulation-based model checking* proposed by Dey, Gelfand, Swartz and Vlachos, 98, as a computationally intense method for model checking. It seems to work well in detecting the incompatibility between model and the data, but it requires proper priors.

- *O'Hagan method* (O'Hagan, 2003) is highly sensitive to the prior chosen, and in fact it seems to be conservative with non-informative priors.

- *Marshall and Spiegelhalter's conflict p-values* (Marshall and Spiegelhalter, 2003) seems to work well, produce as many p-values as number of groups and multiplicity might be an issue.

- Proposals of Johnson, 2006; Evans and Moshonov, 06.

… in conclussion

- Bayesian checks are better than plug-ing checks
- Posterior predictive check are extremely dangerous, unless T is nearly ancillary. But in this case, plug-ing is recommended because it is easier
- Posterior predictive checks are defended on grounds of simple computations; plug-in checks are simpler and often better
- because of its familiarity, p-values, when *calibrated*, are useful for model checking (but the message is the same for other, formal or informal, checks, like graphical checks)

- if a true p-value ($U[0,1]$) is desired with uncentered T
  - p_{ppost} $(and p_{ppred})$ are best in asymptotic and studied small sample situations; they *automatically* centers T
  - p_{plug} is superior to p_{post}
- computationally,
  - p_{plug} and p_{post} are usually simplest
  - p_{ppost} is easy to compute if $f(t|\theta)$ is available.
  - p_{cpred} is available with ABC techniques

– in discrete settings, p_{ppost} offers dramatic gains and avoids excessive conservatism

.

THANKS !! . . .