Bayesian model selection for informative hypotheses: A comparison of data-based versus default encompassing priors

Thomas J. Faulkenberry & Bryanna Scheuler

Tarleton State University

MAA 2019

Consider the test scores from students in three different treatment conditions:

- Treatment 1 - read and reread
- Treatment 2 - read, then answer prepared questions
- Treatment 3 - read, then create and answer questions

Treatment 1	Treatment 2	Treatment 3
2	5	8
3	9	6
8	10	12
6	13	11
5	8	11
6	9	12
$M = 5$	$M = 9$	$M = 10$

Typical question – are there differences among these condition means?

Standard approach - analysis of variance (ANOVA)

- model $Y_{ij} = \mu + \alpha_j + \varepsilon_{ij}$, where $\varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$
- assume "null hypothesis" $\mathcal{H}_0 : \alpha_j = 0$
- compute probability of observing data Y_{ij} under $\mathcal{H}_0$
- if data is $rare$ under $\mathcal{H}_0$, reject $\mathcal{H}_0$

ANOVA computations

source	SS	df	MS	F
between treatments				
within treatments				
total				

ANOVA computations

source	SS	df	MS	F
between treatments				
within treatments				
total	172			

$$
\begin{aligned}
SS_{\text{total}} &= \sum Y^2 - \frac{(\sum Y)^2}{N} \\
&= 1324 - \frac{144^2}{18} \\
&= 172
\end{aligned}
$$

ANOVA computations

source	SS	df	MS	F
between treatments	84			
within treatments				
total	172			

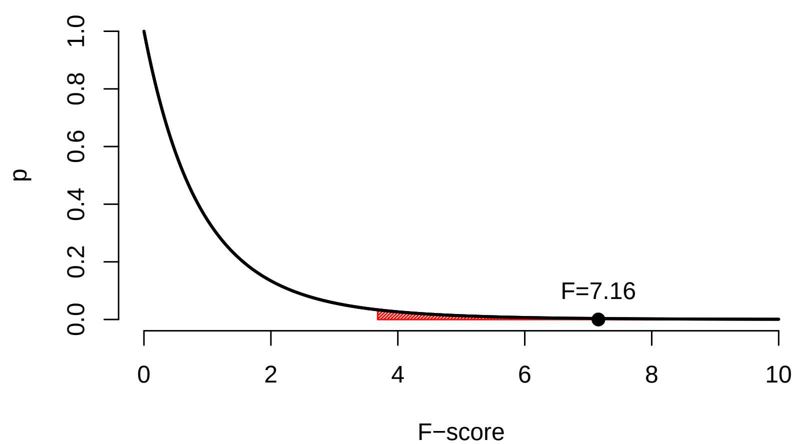
$$
\begin{aligned}
SS_{\text{bet tmts}} &= n\sum_{j=1}^{3}(\overline{Y}_j - \overline{Y})^2 \\
&= 6\Big[(5-8)^2 + (9-8)^2 + (10-8)^2\Big] \\
&= 84
\end{aligned}
$$

ANOVA computations

source	SS	df	MS	F
between treatments	84	2	42	7.16
within treatments	88	15	5.87	
total	172	17		

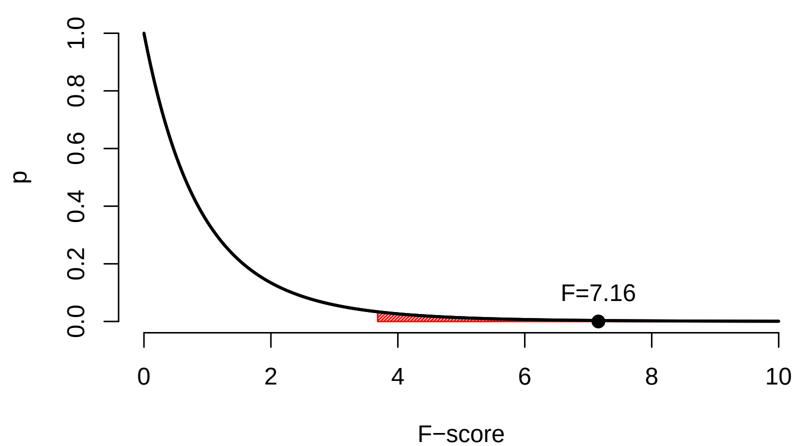
ANOVA computations

source	SS	df	MS	F
between treatments	84	2	42	7.16
within treatments	88	15	5.87	
total	172	17		



ANOVA computations

source	SS	df	MS	F
between treatments	84	2	42	7.16
within treatments	88	15	5.87	
total	172	17		



Since our data Y_{ij} is rare under $\mathcal{H}_0$, we reject $\mathcal{H}_0$ as an implausible model restriction.

What does this tell us?

If we reject $\mathcal{H}_0 : \alpha_j = 0$, this tells us that $\alpha_j \neq 0$ for some j.

- which values of j?
- are they positive / negative?
- the alternative is rather **uninformative**

Informative hypotheses

Consider instead defining competing $informative$ models:

- $\mathcal{M}_1 : \mu_1 < \mu_2 < \mu_3$
- $\mathcal{M}_2 : \mu_2 < \mu_1 < \mu_3$
- $\mathcal{M}_3 : \mu_1 < \mu_3 < \mu_2$

Informative hypotheses

Consider instead defining competing *informative* models:

- $\mathcal{M}_1 : \mu_1 < \mu_2 < \mu_3$
- $\mathcal{M}_2 : \mu_2 < \mu_1 < \mu_3$
- $\mathcal{M}_3 : \mu_1 < \mu_3 < \mu_2$

Note:

1. each model tells a different story about effective study methods
2. typical ANOVA cannot differentiate between $\mathcal{M}_1$, $\mathcal{M}_2$, $\mathcal{M}_3$

Goal - evaluate relative evidence for each model $\mathcal{M}_j$, in light of observed data $\boldsymbol{y}$

Bayes Theorem:

$$\underbrace{p(\mathcal{M}_j \mid \boldsymbol{y})}_{\substack{\text{Posterior belief}\\ \text{about model}}} = \underbrace{p(\mathcal{M}_j)}_{\substack{\text{Prior belief}\\ \text{about model}}} \times \underbrace{\frac{p(\boldsymbol{y} \mid \mathcal{M}_j)}{p(\boldsymbol{y})}}_{\text{predictive updating factor}}$$

Goal - evaluate relative evidence for each model $\mathcal{M}_j$, in light of observed data $\boldsymbol{y}$

taking quotients:

$$\underbrace{\frac{p(\mathcal{M}_j \mid \boldsymbol{y})}{p(\mathcal{M}_k \mid \boldsymbol{y})}}_{\text{posterior odds}} = \underbrace{\frac{p(\mathcal{M}_j)}{p(\mathcal{M}_k)}}_{\text{prior odds}} \times \underbrace{\frac{p(\boldsymbol{y} \mid \mathcal{M}_j)}{p(\boldsymbol{y} \mid \mathcal{M}_k)}}_{\text{predictive updating factor}}$$

The predictive updating factor, or **Bayes factor**,

$$B_{jk} = \frac{p(\boldsymbol{y} \mid \mathcal{M}_j)}{p(\boldsymbol{y} \mid \mathcal{M}_k)}$$

tells us how much better $\mathcal{M}_j$ predicts our observed data compared to $\mathcal{M}_k$.

The predictive updating factor, or **Bayes factor**,

$$B_{jk} = \frac{p(\boldsymbol{y} \mid \mathcal{M}_j)}{p(\boldsymbol{y} \mid \mathcal{M}_k)}$$

tells us how much better $\mathcal{M}_j$ predicts our observed data compared to $\mathcal{M}_k$.

Example: suppose $B_{12} = 5$.

Interpretation: the observed data are 5 times more likely under $\mathcal{M}_1$ than $\mathcal{M}_2$.

Computing Bayes factors for informative hypotheses

Klugkist et al. (2005) proved the following:

Theorem 1. *Consider two models* $\mathcal{M}_1$ *and* $\mathcal{M}_e$*, where* $\mathcal{M}_1$ *is nested within an encompassing model* $\mathcal{M}_e$ *via an inequality constraint on some parameter* δ *that is unconstrained under* $\mathcal{M}_e$*. Then*

$$B_{1e} = \frac{F}{C}$$

where F *and* C *represent the proportions of the posterior and prior of the encompassing model, respectively, that are in agreement with the inequality constraint imposed by the nested model* $\mathcal{M}_1$*.*

Sample from prior:

	Prior					
Iteration	μ_1	μ_2	μ_3	$\mathcal{M}_1$	$\mathcal{M}_2$	$\mathcal{M}_3$
1	6.54	10.15	-1.78	0	0	0
2	22.60	-0.28	8.03	0	0	0
3	3.37	3.01	-0.63	0	0	0
4	-6.13	11.54	12.33	1	0	0
5	13.68	-0.61	1.50	0	0	0
6	27.83	7.43	6.79	0	0	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮
5000	11.00	13.07	23.91	1	0	0
Sum				847	876	807
Proportion (C)				0.169	0.175	0.161

Sample from posterior:

Iteration	Posterior μ_1	μ_2	μ_3	$\mathcal{M}_1$	$\mathcal{M}_2$	$\mathcal{M}_3$
1	5.59	9.90	12.83	1	0	0
2	2.90	8.86	8.35	0	0	1
3	2.63	10.43	10.44	1	0	0
4	5.55	10.17	9.61	0	0	1
5	4.61	7.24	10.24	1	0	0
6	4.72	8.95	9.78	1	0	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮
5000	5.61	8.72	9.99	1	0	0
Sum				3674	29	1286
Proportion (F)				0.735	0.006	0.257
Proportion (C)				0.169	0.175	0.161
B_{je}				4.43	0.03	1.59

From these Bayes factors, we can compute *posterior model probabilities* (PMPs).

First, assume $\mathcal{M}_1$, $\mathcal{M}_2$, and $\mathcal{M}_3$ are equally likely, *a priori*.

Then we have

$$
\begin{aligned}
B_{11} + B_{21} + B_{31} &= \frac{p(\mathcal{M}_1 \mid \boldsymbol{y})}{p(\mathcal{M}_1 \mid \boldsymbol{y})} + \frac{p(\mathcal{M}_2 \mid \boldsymbol{y})}{p(\mathcal{M}_1 \mid \boldsymbol{y})} + \frac{p(\mathcal{M}_3 \mid \boldsymbol{y})}{p(\mathcal{M}_1 \mid \boldsymbol{y})} \\
&= \frac{p(\mathcal{M}_1 \mid \boldsymbol{y}) + p(\mathcal{M}_2 \mid \boldsymbol{y}) + p(\mathcal{M}_3 \mid \boldsymbol{y})}{p(\mathcal{M}_1 \mid \boldsymbol{y})} \\
&= \frac{1}{p(\mathcal{M}_1 \mid \boldsymbol{y})}
\end{aligned}
$$

So we have

$$p(\mathcal{M}_1 \mid \boldsymbol{y}) = \frac{1}{B_{11} + B_{21} + B_{31}}$$

Multiplying top and bottom by B_{1e}, we have

$$\begin{aligned} p(\mathcal{M}_1 \mid \boldsymbol{y}) &= \frac{B_{1e}}{B_{11} \cdot B_{1e} + B_{21} \cdot B_{1e} + B_{31} \cdot B_{1e}} \\ &= \frac{B_{1e}}{B_{1e} + B_{2e} + B_{3e}} \end{aligned}$$

Thus, we can compute our posterior model probabilities for each $\mathcal{M}_j$:

Model	F	C	B_{je}	PMP
$\mathcal{M}_1 : \mu_1 < \mu_2 < \mu_3$	0.735	0.169	4.43	0.731
$\mathcal{M}_2 : \mu_2 < \mu_1 < \mu_3$	0.006	0.175	0.03	0.005
$\mathcal{M}_3 : \mu_1 < \mu_3 < \mu_2$	0.257	0.161	1.59	0.263

Sensitivity to prior?

- in the Klugkist et al. (2005) formulation, priors are **data-based**

$$\pi(\boldsymbol{\mu}, \sigma^2 \mid \mathcal{M}_e) = \pi(\boldsymbol{\mu} \mid \mathcal{M}_e)\pi(\sigma^2)$$

with

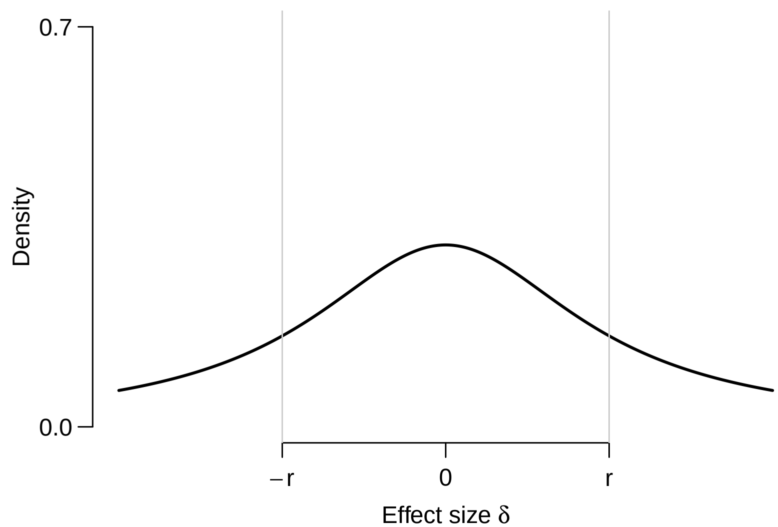
$$\boldsymbol{\mu} \sim \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{T}_0)$$

and

$$\sigma^2 \sim \text{Inv-}\chi^2(v_0, \sigma_0^2)$$

Rouder et al. (2012) instead model such problems as "effects"-driven

- parameter of interest is "effect size" $\delta = \dfrac{\mu_1 - \mu_2}{\sigma}$
- assume $\delta \sim \text{Cauchy}(r)$



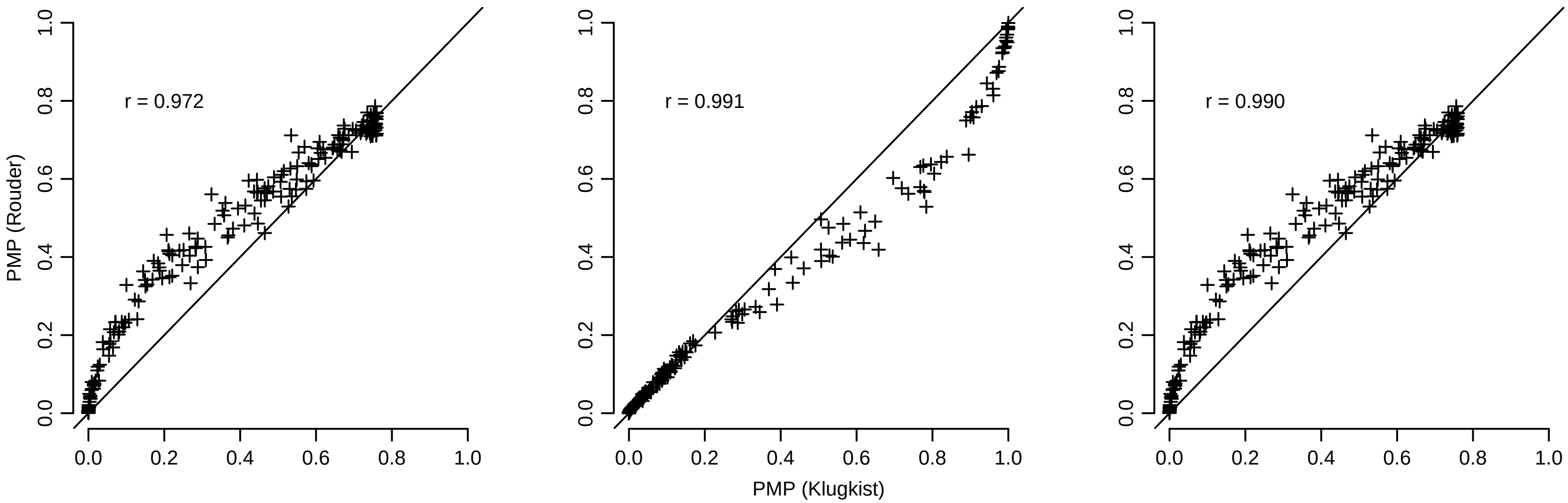
Modeling results are similar:

Model	F	C	B_{je}	PMP (Rouder)	PMP (Klugkist)
$\mathcal{M}_1 : \mu_1 < \mu_2 < \mu_3$	0.732	0.165	4.44	0.739	0.731
$\mathcal{M}_2 : \mu_2 < \mu_1 < \mu_3$	0.009	0.167	0.05	0.009	0.005
$\mathcal{M}_3 : \mu_1 < \mu_3 < \mu_2$	0.256	0.169	1.52	0.252	0.263

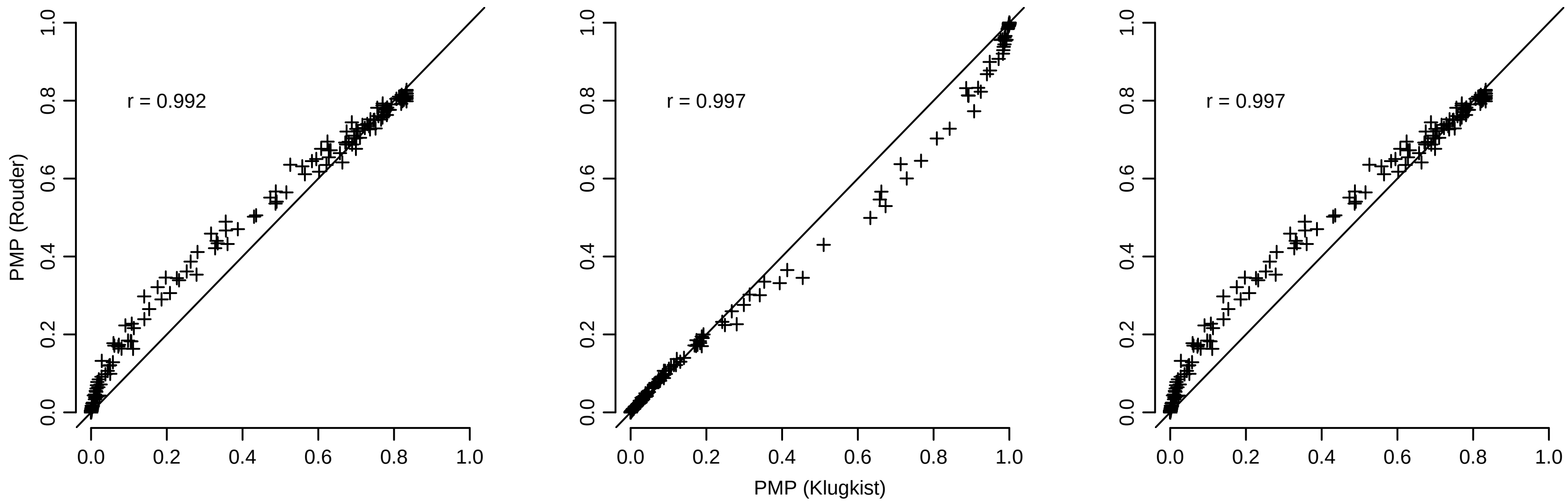
Simulation:

- two populations: $\mathcal{N}(0,1)$ and $\mathcal{N}(\alpha,1)$, where $\alpha \sim$ Uniform$(-1,1)$
- random samples of size $N = 20, 50, 80$
- Three competing models:
  - $\mathcal{M}_1 : \mu_1 \approx \mu_2$
  - $\mathcal{M}_2 : \mu_1 < \mu_2$
  - $\mathcal{M}_3 : \mu_1 > \mu_2$
- computed PMPs using (1) Klugkist method and (2) Rouder method

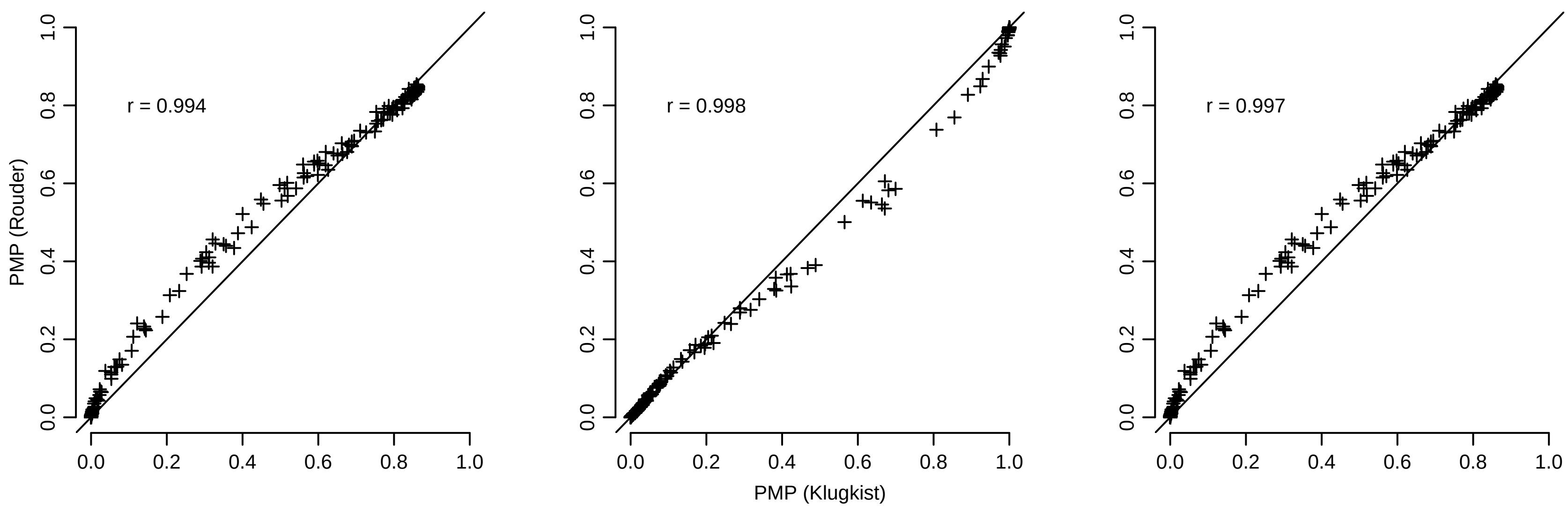
Results: $N = 20$



Results: $N = 50$



Results: $N = 80$



Thank you!

- Thanks to Tarleton Office of Research and Innovation for funding!

- slides available at `github.com/tomfaulkenberry/talks`

- more details in Faulkenberry, T. J. (2019). A tutorial on generalizing the default Bayesian t-test via posterior sampling and encompassing priors. *Communications for Statistical Applications and Methods, 26*(2), 1-22.

- Twitter: `@tomfaulkenberry`

- Email: `faulkenberry@tarleton.edu`