Comparing Objective and Subjective Bayes Factors for the Two-Sample Comparison

Wesley Johnson
UC Irvine

UCI 2017

Bayesian Model Selection: Two Competing Models

- Two models for data:

$$y \mid M_0 \sim p(\cdot \mid, M_0, \theta_0) \qquad y \mid M_1 \sim p(\cdot \mid M_1, \theta_1)$$

- Compare

  Log-Normal model to Weibull model for non-negative data
  Logistic Regression model to Probit regression model for Binomial data
  Models in the same family with parameters $\theta_0 \subset \theta_1$, or not

- Elicit prior distributions: $p_0(\theta_0)$ and $p_1(\theta_1)$
- Specify prior probabilities

$$q_0 = Pr(H_0) \qquad q_1 = 1 - q_0 = Pr(H_1)$$

- Calculate $Pr(M_i \mid y)$. Choose M_1 if

$$Pr(M_1 \mid y) > 0.99 \Leftrightarrow Pr(M_1 \mid y)/Pr(M_0 \mid y) > 99$$

Bayesian Model Selection: Two Competing Models

- We require the marginal predictive densities for the observed data:

$$p(y \mid H_i) \equiv \int p(y \mid H_i, \theta_i) p_i(\theta_i) d\theta_i \qquad i = 0, 1.$$

- We regard this as the (marginal) "plausibility" of the observed data under model M_i
- Also termed the marginal $Lik(M_i \mid y)$
- With improper priors the marginal pdf often doesn't exist
- Applying Bayes Theorem:

$$Pr(M_1 \mid y) = \frac{q_1 p(y \mid M_1)}{q_0 p(y \mid M_0) + q_1 p(y \mid M_1)}.$$

Bayesian Model Selection: Two Competing Models

- Decision theoretic criteria make it "easy" to decide on a cutoff for deciding in favor of either model
- Since $Pr(M_1 \mid y) > k \Leftrightarrow Pr(M_1 \mid y)/[1 - Pr(M_1 \mid y)] > k/(1-k)$, the posterior odds are often used to make a decision
- We have

$$\begin{aligned} \frac{Pr(M_1 \mid y)}{Pr(M_0 \mid y)} &= \frac{q_1 p(y \mid M_1)}{q_0 p(y \mid M_0)} \\ &\equiv \frac{q_1}{q_0} BF, \end{aligned}$$

- BF is the ratio of marginal likelihoods of the models M_1 and M_0

Bayesian Model Selection: Two Competing Models

- Thus, posterior Odds equals the prior odds q_1/q_0 times the Bayes Factor
- If we favor both models equally, we set the prior odds to one, so posterior odds reduced to the BF
- With $k = 0.5$, we select model M_1 if $BF > 1$
- With $k = 0.95$, we select model M_1 if

$$BF > 0.95/0.05 = 19 \Leftrightarrow p(y \mid M_1) > 19\, p(y \mid M_0),$$

which is considerably more stringent

Bayesian Model Selection: Two Competing Models

- The *BF* is often preferred to $Pr(M_1 \mid y)$ as a criterion for making a decision because it may be difficult or controversial to pick a value for the prior probability $Pr(M_1)$

  Suppose M_0 corresponds to the hypothesis that a particular cancer drug will have no effect in terms of prolonging life
  Suppose M_1 corresponds to a hypothesis that a defendant is guilty of murder

- Using the *BF* eliminates this issue at the expense of having to decide how large does *BF* need to make a decision
- But recall that the marginal likelihoods

$$p(y \mid H_i) \equiv \int p(y \mid H_i, \theta_i) p_i(\theta_i) d\theta_i \qquad i = 0, 1.$$

still depend on the marginal prior specifications

The Two-Sample Comparison

- Assume two independent normal samples with different means, (μ_1, μ_2), and common variance σ^2
- Define

$$t = \frac{\bar{y}_1 - \bar{y}_2 - 0}{s_p\sqrt{1/n_1 + 1/n_2}} \equiv \frac{\sqrt{n_\delta}(\bar{y}_1 - \bar{y}_2)}{s_p}$$

- Define the effect size (and nuisance parameter)

$$\delta = \frac{\mu_1 - \mu_2}{\sigma}, \qquad \gamma = \frac{\mu_1 + \mu_2}{2}$$

- Let $p(\gamma, \sigma^2) \propto 1/\sigma^2$, and (independently) $\delta \sim \mathrm{N}(\lambda, \sigma_\delta^2)$
- Then the *BF* comparing

$$M_0 : \delta = 0 \quad \text{to} \quad M_1 : \delta \neq 0$$

is an analytical function of t^2 cf. Gönen et al. (2005)

Subjective Prior for PSI Study (Study of Paranormal Phenomena)

- GBF specify prior for δ based on subjective information
- A PSI study was performed to ascertain whether a particular type of psychic phenomenon exists based on the frequentist analysis of nine experiments (Bem, 2011)
- Bem et al. (2011) re-analyzed the data using GBF
- $\hat{\delta}$ s in PSI studies typically range from 0.2 to 0.3; previous meta-analysis of 56 psi experiments had est median $\hat{\delta}$ across studies of 0.18 (Utts et al., 2010); meta-analysis of 38 studies with ave $\hat{\delta} = 0.28$ (Mossbridge, Tressoldi, and Utts, 2011)
- They asserted *no reasonable observer would ever expect* $\delta > 0.8$ *in laboratory psi experiments*
- Selected prior for δ was $N(0, (0.5/1.645)^2)$; 95th percentile is 0.5; $Pr(-0.5 < \delta < 0.5) = 0.9$, and $Pr(\delta < 0) = 0.5$

Effect Sizes

- It is common in medical and psychology literature focus on the effect size, $\delta = (\mu_1 - \mu_2)/\sigma$
- So if $\delta = 2$, we know that the difference in population means is two population standard deviations above zero
- This is is an extraordinary difference and exemplifies a difference that one could literally see
- For example, the effect size for the difference between adult male and female heights in the US is 2 (cf. Utts and Heckard, p. 541)
- If one can actually see the difference, there may be no real need for conducting an expensive experiment to a forgone conclusion
- In some literature, effect sizes between 0.2 and 0.5 are common and effect sizes larger than 0.8 are considered quite large

A Controversy between Objective and Subjective Points of View

- Wang and Liu (2015) presented an objective Bayes factor (BF) as an alternative to a subjective one presented by Gönen et al. (2005). Their *BF* is also a function of the data only through t^2
- They intended to show superiority of WBF to GBF based on *undesirable behavior* of GBF
- An evident premise in WBF is that Objectivity is good and Subjectivity not so much
- Wonderful Bayesian feature is that we get to *lay all cards on the table*
- Major distinguishing feature of various BFs is the choice of priors (cards)

Properties for BFs

- Objective BFs typically involve priors chosen to have *nice* frequentist properties
- Often/mostly don't incorporate subjective scientific input
- Objective priors are often termed as *diffuse*, or *non-informative*
- But so-called *non-informative* priors may be *dis-informative* eg. a uniform prior on the unknown prevalence of HIV infection among blood donors; more to come
- We are all concerned about the sensitivity of inferences to the choice of prior distribution

Properties for BFs

- Desiderata of BFs have been discussed in the literature, including Bayarri et al. (2012), Rossell and V. Johnson (2011) and V. Johnson (2013AB)
- We give a list of essentially objective properties, not necessarily *good* ones in our opinion
- We then offer new desiderata that we argue may be *scientifically* more relevant

Properties for BFs: Objective Point of View

- 1: **Consistency**: As the sample sizes grow , $BF \to \infty$ if M_1 is true, and $BF \to 0$ M_0 is true
- 2: **Finite Sample Consistency** (FSC): FSC means that for fixed sample sizes, as $|t| \to \infty$, $BF \to \infty$
- 3: **Robustness to Prior**: Standard motivation for using BFs rather than $Pr(M_i \mid \text{data})$ is that BF is free of $Pr(M_i)$. BFs should also be reasonably insensitive to the within-model priors, $p(\theta_i|M_i)$
- 4: **Compatibility with Frequentist Testing**: If $|t|$ is large, leading to frequentist rejection of M_0, the BF should not favor M_0 eg. no Bartlett paradox (Bartlett, 1957)

Properties for BFs: Objective Point of View

- 5: **Ability to Accumulate Evidence Favoring** M_0: As the sample size grows, the evidence favoring either M_0 or M_1 should be able to grow at the same rate

- 6: **High Power**: Under M_1, the (frequentist) probability (for a particular prior specification) that the BF exceeds a given threshold should be large relative to BFs computed using other priors

  Researchers typically propose prior distributions to get BFs that satisfy these desiderata, rather than to incorporate scientifically relevant information (SRI)

Properties for BFs

- In followup paper to Gönen et. al. (2005), Gönen et. al. (2017)
  - (i) Discuss these criteria
  - (ii) Lay subjective and objective cards on the table
  - (iii) Encourage use of minimizing TPM to compare BFs
- They argue that objective priors for δ can be silly/absurd from a subjective viewpoint eg. priors that allow large prior probability that $\delta > 2$ when only relatively small values of δ can be reasonably anticipated
- They argue in favor of selecting a classification rule based on minimizing the total probability of mis-classification; Fisher's linear discriminant function for classifying multivariate normals into groups can be derived as a Bayes Rule that minimizes the TPM

- More generally, the maximum posterior probability rule minimizes TPM when TI and TII errors are comparable

Additional Property

- 7 **Incorporate Scientifically Relevant Information**:
  - Our focus is on taking account of SRI when specifying λ and σ_δ in our normal prior for δ
  - In our experience in working with scientists in collaborative settings, their attitude as been of the nature "why wouldn't I want to take account of my own expertise and knowledge" in the analysis of my data
  - Moreover we argue that, when attempting to specify an objective prior, it may be dangerous to not take account of any unintended consequences that are implied by its specification
  - Consequently, we argue for the necessity of inspecting objective priors to highlight any inconsistencies/incompatibility with known SRI
  - Berger and Delampady (1987) argued that there was no such thing as an objective prior for this problem

PSI Study with Cauchy Prior on δ: Jeffreys BF

- If $\delta \sim$ Cauchy (Wagenmakers et al. 2011), we get prior probabilities

  $Pr(|\delta| > 0.8) = 0.57, Pr(|\delta| > 2) = 0.30, Pr(|\delta| > 5) = 0.16$

  Corresponding BF is a function of t^2
- If effect sizes were really that large, there would be no debate about the reality of PSI
- Recall that the effect size for measuring the effect of the difference of male and female heights in the US is about 2
- We consider this to be a wildly unrealistic prior for the particular PSI problem under study
- The combined BF for the 9 experiments using the Cauchy prior is 0.632 while the corresponding GBF = 13,699
- WOW!

Cauchy Prior on δ

- Rouder et. al. (2009) discuss the situations: $\delta \sim \text{N}(0, 1)$, $\delta \sim \text{Cauchy}(0, r^2)$ (Jeffreys was first with $r = 1$; then Zellner-Siow, (1980))
- Rouder et al argue, as we do, that not too much prior weight should be attached to "large" values of δ
- They argue in favor of Subj normal priors on δ, when information is available, and the "Obj" Cauchy when not
- They point out that, in a single sample problem with $n = 500$, the t value corresponding to $BF = 10$ in favor of M_0 is 1.44, and the corresponding t value corresp to $BF = 10$ in favor of M_1 is 3.38; quite different from the usual cutoff in latter case; Zellner and Siow (1980) made the same kind of point
- The dramatic difference in BFs in the PSI example makes the point that the choice of prior really can matter
- They note $\lambda = 0$, $1/\sigma_\delta^2 \sim \chi_1^2 \Rightarrow$

$$\delta \sim \text{Cauchy}(0, 1)$$

WBF

- Wang and Liu (2015) used $\sigma_\delta^2 \sim$ Pearson VI(a, b); BF is a function of t^2
- In order to satisfy finite sample consistency, they selected (a, b) such that

$$b = (n_1 + n_2 - 1)/2 - a - 5/2$$

  and $a \in (-1, -1/2]$
- They argue that WBF will be robust if $a = -0.75$ (to other choices admissible a)
- With $n_1 = n_2 = 10$, the WL prior has median and 90th percentiles for σ_δ of 6.2 and 158 respectively, while the prior mode is only 1.11
- Induced prior on δ has $Pr(|\delta| > k) = (0.8, 0.64, 0.44, 0.10)$ for $k = (1, 2, 5, 10)$; subjectively "worse" than the Cauchy

Criticisms of GBF

- A main criticism of GBF by WL is that it may suffer from Bartletts paradox
- It is true that if you let me pick σ_δ to be as large as I like, then with t values in the usual rejection range and larger, I can probably make the BF in favor of M_0 grow indefinitely
- But why on earth would I want to let σ_g be much larger than 1, much less 10^6?
- When any prior places a huge amount of mass on very large values of δ, and when the data are suggesting that δ is actually somewhat small, eg. 0.2 say, is it surprising that the model believes that δ is so small relative to the anticipated huge size, as to be effectively zero?

Criticisms of GBF

- Let $Pr(M_i) = 0.5$, $\sigma_\delta = 10$ and $n_1 = n_2 = 500$, Then with
  - $t = 2, 2.5, 4, 5$
  - $Pr(M_0 \mid \text{data}) = (0.96, 0.87, 0.053, 0.0007)$
  - $(\hat{\delta} = 0.13, 0.16, 0.26, 0.32)$
- This is all reasonable: With a prior that anticipates huge effect sizes, it will take a larger t to conclude M_1. It is a really stupid prior
- With a small estimated effect size < 0.5, and with large enough σ_δ, we were able to make the BF in favor of M_0 quite small for substantial values of t
- We view this as perfectly reasonable behavior of the GBF, an argument in favor of the Bartlett effect being a good effect, and an argument for carefully choosing ones prior

Lindley Paradox Revisited:

- Specify $\delta \sim \mathrm{N}(0, 1/9)$. This results in $Pr(\delta > 1) = 0.003$
- Plot $Pr(M_0 \mid n_\delta, t = 3)$ versus n_δ
- Corresponding effect sizes are

$$\hat{\delta} = (3.0, 0.95, 0.3, 0.095, 0.03, 0.0095, 0.003, 0.00095, 0.0003)$$

- Recall that $t = \sqrt{n_\delta}\hat{\delta}$ so $\hat{\delta} = t/\sqrt{n_\delta}$
- It's not a paradox; it corrects for sample size

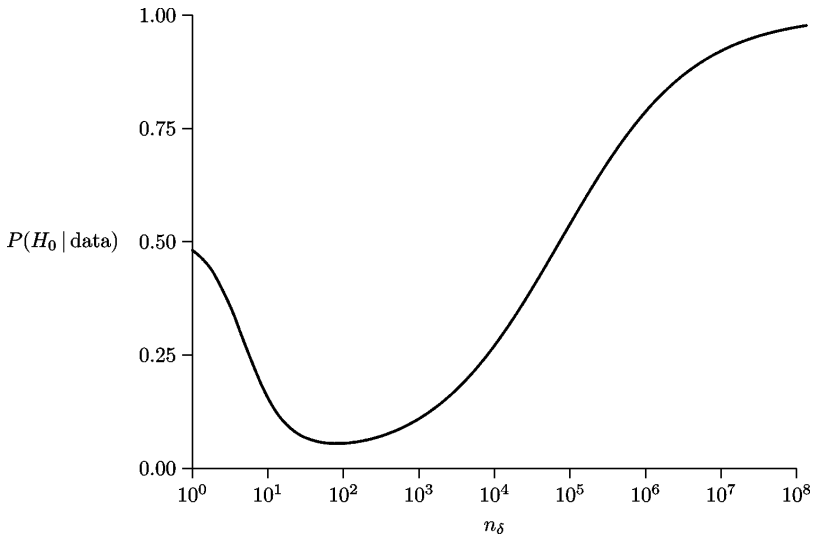


Figure 3. Posterior probability of H_0 as a function of n_δ when π_0 = .5, (λ; σ_δ) = (0, 1/3), and t = 3.00, illustrating Lindley's paradox.

As a concluding note, it is simple to discuss "Lindley's Para-dox" (Lindley 1957) using (2). Lindley noticed that data from large sample sizes that are "highly significant" from a frequentist standpoint can support H_0 better than H_1. Imagine, in the case above, that $t = 3$ .00, highly significant by any measure.

Finite Sample Consistency

- Wang and Liu (2015) note that the GBF is not finite sample consistent since

$$GBF \to (1 + n_\delta \sigma_\delta^2)^{(n_1+n_2-2)/2} < \infty\,, \quad |t| \to \infty$$

- Let $\sigma_\delta^2 = 1$, $Pr(M_i) = 0.5$, and $n_1 = n_2 = 10$. Then as $t \to \infty$

$$Pr(M_1 \mid \text{data}) = GBF/(1 + GBF) \to 0.9999999$$

- If the observed t is 3, 5 or 7, the corresponding posterior probabilities are 0.90, 0.995, and 0.9997, so the limiting bound plays no important role under these circumstances
- Tempest in a teapot

Bayes Factors that Don't Depend on t

- IBF: Intrinsic BF developed by Berger and Pericchi, avoids proper prior specification
- RJBF: Rossell and Johnson (2011) argued against traditional objective BFs involving local priors having mode at 0 under the alternative. They propose using symmetric non-local priors, which require low probability mass near the null, under the alternative specification
- VJBF: Johnson (2013A, 2013B) proposes a BF using an objective prior that maximizes the probability of exceeding a given evidence threshold for all possible alternative prior distributions. The resulting analysis has a close correspondence with frequentist fixed α test procedures

Additional Property

- 8: **Correct Classification:** Select a decision rule that classifies models correctly as much as possible, eg that minimizes the true/oracle TPM

$$Pr(M_0)Pr(\text{Decide } M_1 \mid M_0) + Pr(M_1)Pr(\text{Decide } M_0 \mid M_1)$$

  - With equal error costs and with $q_0 = q_1 = 0.5$, the Bayes decision rule is

$$\text{choose } M_1 \text{ if } BF > 1 \qquad \text{choose } M_0 \text{ otherwise}$$

  - This rule only minimizes the oracle TPM if $q_0 = Pr(M_0)$

TPM

Minimizing TPM is:

- Objective and Subjective
- Easily verified by simulation eg. simulate under what is believed to be the true conditions and use the decision rule that accords with that
- Clearly shows the effects (positive or negative) of assuming particular priors

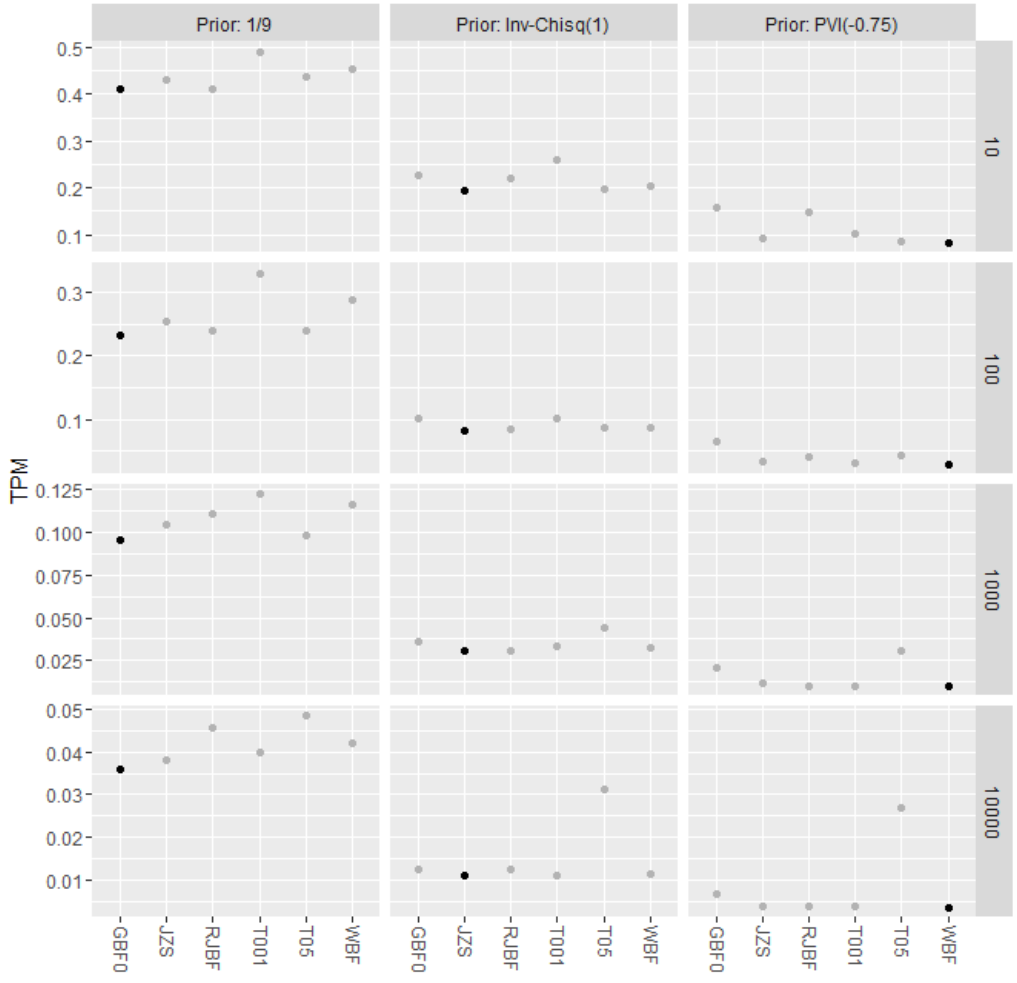
Properties for BFs

- We thus would like the rule to classify the data as coming from the correct model according to the actual frequency of occurrence of these models
- There is no right answer to the question, *how often might the different models occur? for the current experiment*
- However, because the answer to this question affects classification rates substantially, researchers should carefully consider this question when choosing a *BF*
- We subsequently address this question using historical data to provide partial answers

Simulation Study to Compare Two-Sample Comparisons BFs Objectively:

- Step 1: Randomly generate σ_δ^2 from an assumed prior distribution or assign a pre-specified positive value. Fix the value of λ according to the selected method
- Step 2: Randomly generate $\delta \sim \mathrm{N}(\lambda, \sigma_\delta^2)$, or set $\delta = 0$, each with probability 0.5
- Step 3: Randomly generate the $t \sim \mathrm{N}(n_\delta\,\delta, 1)/\sqrt{\chi^2/\nu}$
- Step 4: Calculate BFs that depend on t using the simulated t value from Step 3
- Step 5: A misclassification error occurs if $BF > 1$ and $\delta = 0$ or if $BF < 1$ and $\delta \neq 0$
- Step 6: Repeat Steps 1-5 NSIM times to estimate the TPM

Simulations of TPM for Various Models: TPM as function of n_δ



Historical Studies

- A BF considered in Gönen et al. (2005) used the prior

$$\delta \sim \mathrm{N}(2.8\sqrt{1/n_\delta}, [2.19\sqrt{1/n_\delta}]^2)$$

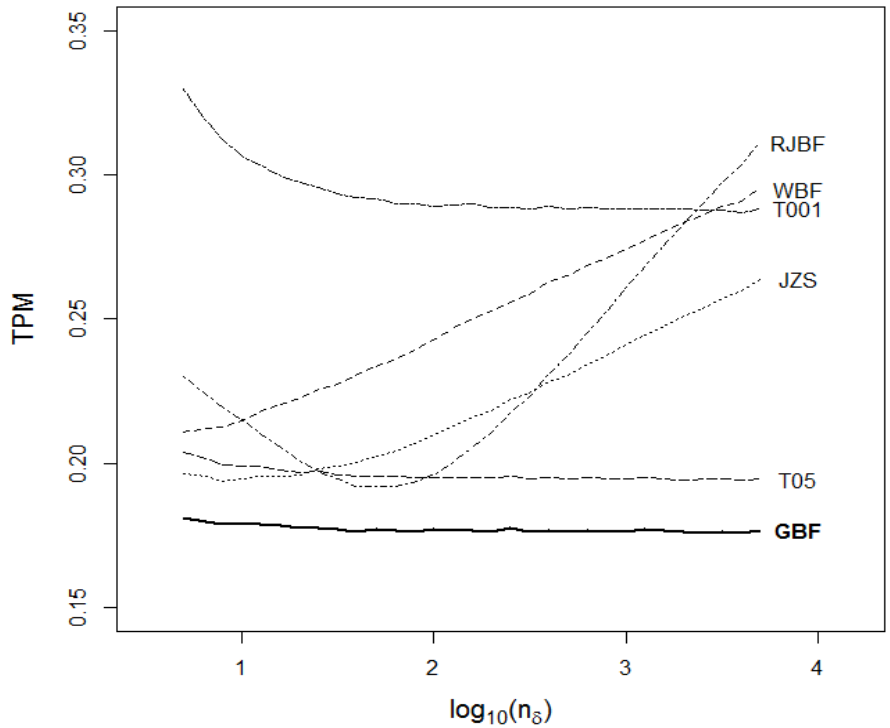
  which was designed to be consistent with subjective prior information that is commonly used in power analysis

- Researchers often choose a large sample size to accommodate *a priori* information that the effect size is small

- This prior makes predictions that are reasonably consistent with published oncology studies

- Not recommended for general use, but researchers do have prior information that can be used to construct their prior distribution (See Gönen et al., 2005 for careful illustration)
- Performed simulations using this model for the generation of data and compared different BFs

Plots of TPM versus $log_{10}(n_\delta)$ under $\delta \sim N(2.8/\sqrt{n_\delta}, [2.19/\sqrt{n_\delta}]^2)$



Wetzels

- Wetzels et al. (2011) report results on 855 t tests from publications in psychology journals, of which 166 were for two-sample comparisons
- The next Figure shows plot of pairs of $|\hat{\delta}|$ versus $1/\sqrt{n_\delta}$ for these studies
- Absolute values used since not clear whether negative estimates were in the anticipated directions
- Least squares fit is $\lambda = 0.20 + 2.75/\sqrt{n_\delta}$
- Comfortably agrees with the Gönen et al. suggestion of $\lambda = 0.00 + 2.8/\sqrt{n_\delta}$
- Additional studies of studies discussed in Gönen et al. (2017) also suggest the value of using this type of information in constructing subjective priors

Plot of Wetzels et al. estimated effect sizes versus $1/\sqrt{n_\delta}$