Fiducial and Objective Bayesian inference

Some comparisons

Leiv Rønneberg

University of Oslo

May 22, 2018

Goals of the talk

- Give some historical context as to why the theories were developed
  - Scientific objectivity
- Fisher's fiducial inference – a short intro
- Objective Bayes methods, quick overview
- Examples
  - Behrens-Fisher problem
  - Paired exponentials

Historical context

- Researchers look to statistics to provide framework for data analysis and inference
- Many (most?) applied fields are frequentist
  - Why?
- Frequentism is perceived as more 'objective'
  - Objective $\sim$ 'free from subjective opinion and bias'
- Bayesian analysis:

$$p(\theta|\text{data}) \propto f(\text{data}|\theta)\ \pi(\theta)$$

- $\pi(\theta)$ feels 'iffy' to some researchers
  - What if my prior knowledge is incorrect?
  - What if there is little available prior information?
  - What if the parameter is hard to interpret?
- In early 20th century, Bayes' theorem was main workhorse of statistical inference
- Can we choose a prior to represent our ignorance, or lack of knowledge?

Historical context cont.

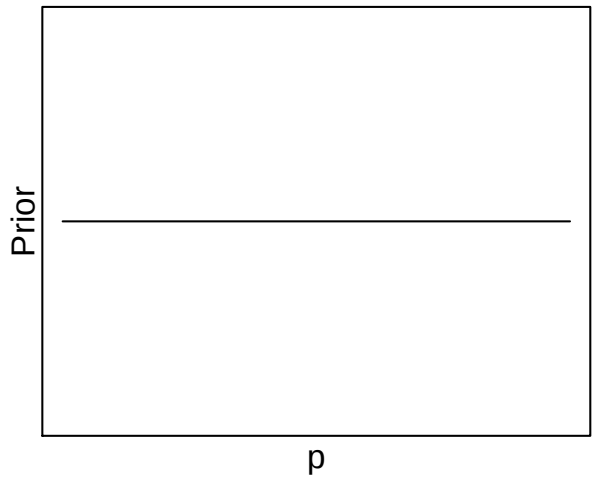
- Traditionally, going back to Laplace, Principle of Insufficient Reason! (PoIR)

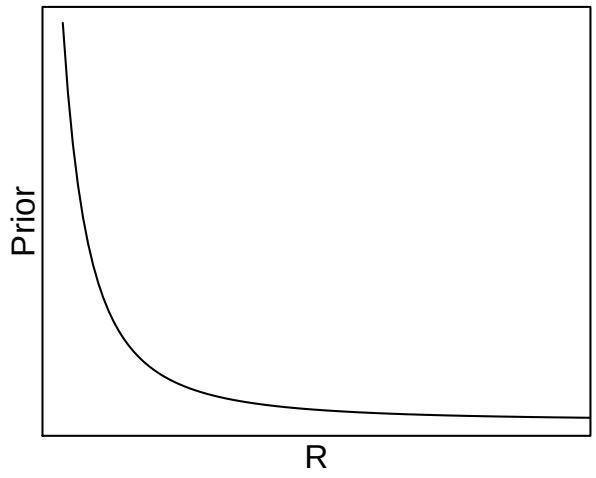
  [...] if there is no known reason for predicating of our subject one rather than another of several alternatives, then relatively to such knowledge the assertions of each of these alternatives have an equal probability. (Keynes 1921)

- Consider a coin toss – has a sample space $\{\text{heads}, \text{tails}\}$
  - If we have no reason to believe that one outcome is more probable than another, PoIR $\implies$ $P(\text{heads}) = P(\text{tails}) = 0.5$
- Seems reasonable, but has troubling consequences

PoIR troubles

- Main issue: PoIR is not invariant to monotone transformations!
  - If I'm ignorant about θ, I should be equally ignorant about any one-to-one transformation of it!
- Example: Consider $X \sim \text{Bin}(n, p)$ for a given n.
  - PoIR $\Longrightarrow$ $\pi(p) = 1$ on $[0, 1]$
  - Consider instead the odds, $R = p/(1-p) \in (0, \infty)$
  - Transformed prior: $\pi(R) = \pi(p(R))|\mathrm{d}p/\mathrm{d}R|$ should be equally non-informative





PoIR troubles cont.

- Clear that PoIR was heavily flawed.

  I know only one case in mathematics of a doctrine which has been accepted and developed by the most eminent men of their time, and is now perhaps accepted by men now living, which at the same time has appeared to a succession of sound writers to be fundamentally false and devoid of foundation. (Fisher 1930)

- Two distinct solutions evolved
  - Fisher's fiducial distribution – prior free!
  - Objective Bayesian method
    - Jeffreys' invariant prior
    - Entropy arguments (Reference priors)

The fiducial distribution

- Fisher's idea – drop the prior altogether!
  - Parameters are fixed, unknown quantities – to be discovered
  - All uncertainty is in the sample space, $\mathcal{X}$
  - But by transferring the uncertainty from $\mathcal{X}$ into the parameter space Θ, one can still obtain a distribution function similar to the Bayesian posterior – without a prior!
- One method:

$$\text{Consider } X_1, \ldots, X_n \overset{\text{iid.}}{\sim} N(\theta, 1). \text{ Then } \sqrt{n}(\bar{x} - \theta) \sim N(0, 1)$$

- Now, consider $\bar{x}$ as a fixed quantity, and let θ vary, the function

$$C(\theta) = 1 - \Phi\left(\sqrt{n}(\bar{x} - \theta)\right)$$

  is a distribution function on Θ!

The fiducial distribution cont.

- An important feature of $C(\theta)$ is that, at true value θ_0, $C(\theta_0) \sim \text{Unif}(0, 1)$
  - $\implies$ quantiles are exact confidence intervals!
  - Key point that ensures inferential validity!
- By taking derivatives of $C(\theta)$ wrt. θ, one finds that

$$\theta \overset{\text{fid.}}{\sim} N(\bar{x}, n^{-1})$$

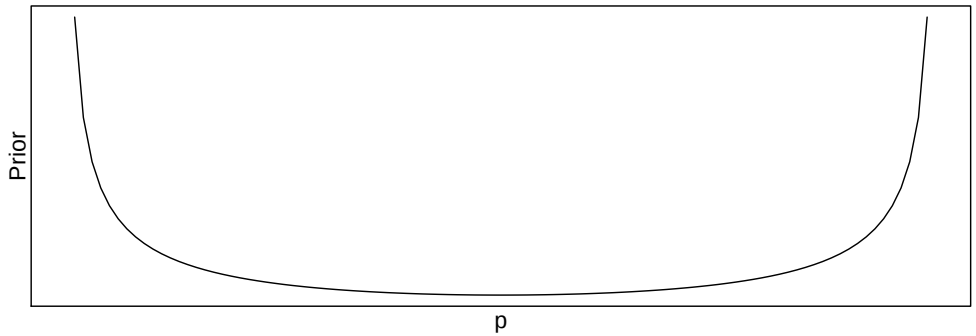
- At this point, some suspicion should arise... θ is a fixed real number..?
- Controversies ensued, especially once Fisher extended the technique to the multiparameter setting
  - To counter the critiques, he gave up the $C(\theta_0) \sim \text{Unif}(0, 1)$ requirement.
  - Interpreting his fiducial distribution as an objective Bayesian posterior distribution, obtained without priors.

Jeffreys prior

- Going back to the world of non-informative priors
  - Can we choose a non-informative prior?
- Jeffreys (1946) suggested using

$$\pi(\boldsymbol{\theta}) \propto \sqrt{\det I(\boldsymbol{\theta})}$$

  where $I(\boldsymbol{\theta})$ is the Fisher information matrix.
- This has the property that it is invariant under monotone transformations!
- But Jeffreys' prior has lost the intuitive explanation given by PoIR!



- ... and, can be heavily inconsistent in the presence of nuisance parameters!

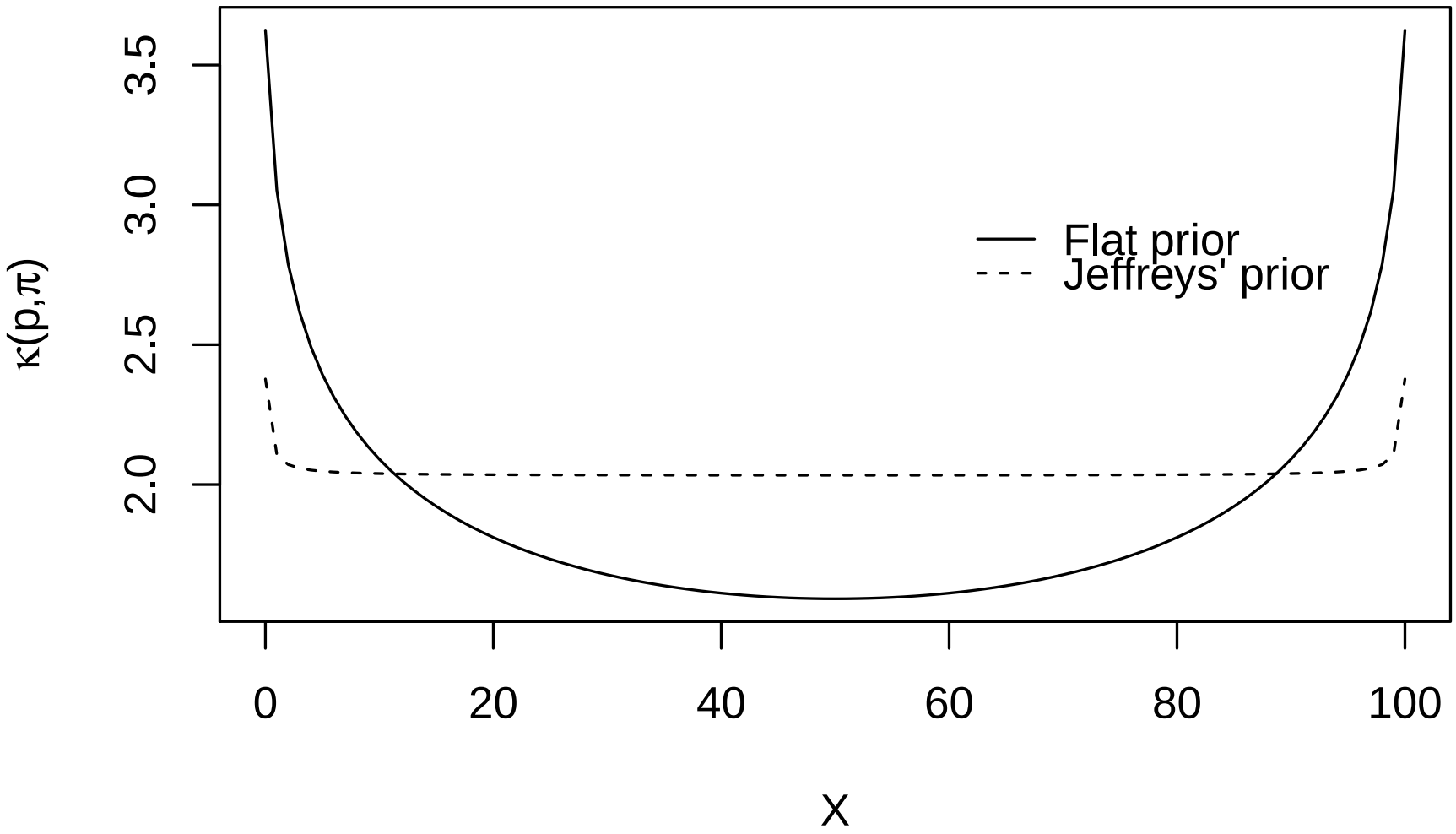
Reference priors

- For dealing with nuisance parameters, the reference prior approach (Bernardo 1979) is a nice extension of Jeffreys' prior.
- Main idea: Prior distributions should be 'as far away as possible' from the posterior distribution.
- One way of measuring this, is the Kullback-Leibler divergence (relative entropy) from prior to posterior.

$$\kappa(p|\pi) = \int_{\Theta} p(\theta|\boldsymbol{x}) \log \frac{p(\theta|\boldsymbol{x})}{\pi(\theta)} \mathrm{d}\theta$$

- The KL-divergence will depend on which sample $\boldsymbol{x}$ is observed,

KL-divergence in Binomial setup



- To remove this dependence, we can take the average over the sample space to obtain the *expected information gain*;

$$\begin{aligned}\mathbb{E}\kappa(p|\pi) &= \int_{\mathcal{X}} p(\boldsymbol{x}|\theta)\kappa(p|\pi)\mathrm{d}\boldsymbol{x} \\ &= \int_{\mathcal{X}}\int_{\Theta} p(\boldsymbol{x}|\theta)\pi(\theta)\log\frac{p(\theta|\boldsymbol{x}}{\pi(\theta)}\mathrm{d}\theta\mathrm{d}\boldsymbol{x}\end{aligned}$$

- *Very* loosely stated; a reference prior, is a prior distribution $\pi_R(\cdot)$ that maximizes the expected information gain while still facilitating a proper posterior distribution.
- Definition is non-constructive (but!) In the one-dimensional case, Jeffreys' Prior==Reference prior
- If nuisance parameters are present, Jeffreys' prior $\neq$ Reference prior

- When nuisance parameters are present, it is derived sequentially, i.e. if we have $(\psi, \lambda_1, \lambda_2)$ in descending order of importance
  - Keep (ψ, λ_1) fixed, and derive $\pi_R(\lambda_2|\psi, \lambda_1)$
  - Then integrate λ_2 out and derive the next one
  - Finally, $\pi_R(\psi, \lambda_1, \lambda_2) = \pi_R(\psi)\pi_R(\lambda_1|\psi)\pi_R(\lambda_2|\psi, \lambda_1)$
- In this case, the reference prior will depend on
  - Ordering and grouping of the parameters by inferential importance
  - Can also depend on the mathematical machinery itself – how a specific limit is taken.
- It does however seem to clear up many of the inconsistencies observed with the multivariate Jeffreys' prior.
- However, reference priors can be difficult to derive.

Objective posterior distributions

- So, now we have two methods claiming to yield somewhat objective posterior distributions
  - Fiducial distribution – no prior, inferential validity ensured by quantiles being CIs
  - Objective Bayesian methods – non-informative priors, inferential validity ensured by ??
- It is of interest to ask when they coincide, if ever?
- Two notable references in this direction is;
  - Lindley (1958): equality can be obtained if parameters are scale or location (one-parameter case)
  - Veronese & Melilli (2017): equality can be obtained in joint models, if the models are members of a subclass of the NEF (cr-NEF).
- What about *focussed* examples?
  - i.e. full model $(\theta_1, \ldots, \theta_p) \mapsto (\psi, \lambda_1, \ldots, \lambda_{p-1})$
  - only really interesting in inference about ψ.

Behrens-Fisher problem

- Notorious problem in statistical inference, of historical and practical importance.

$$\text{Let } X_1, \ldots, X_n \overset{\text{iid.}}{\sim} N(\mu_1, \sigma_1^2) \text{ and } Y_1, \ldots, Y_m \overset{\text{iid.}}{\sim} N(\mu_2, \sigma_2^2).$$

- Variances are unknown, possibly unequal – and inference is sought for the focus parameter $\psi = \mu_2 - \mu_1$, all else are nuisance.
- Let's first consider Fisher's fiducial solution, which came under scrutiny, and led him to change much of his initial argument.

Behrens-Fisher fiducial solution

- Starting from the two familiar pivotal quantities

$$t_1 = \frac{\sqrt{n}(\mu_1 - \bar{x})}{s_1} \text{ and } t_2 = \frac{\sqrt{m}(\mu_2 - \bar{y})}{s_2}$$

  where t_1 and t_2 have the Student's T distribution with $n-1$ and $m-1$ degrees of freedom, respectively.

- By standard fiducial argument, we can rewrite μ_1 and μ_2 as

$$\mu_1 = \bar{x} + \frac{s_1}{\sqrt{n}} T_1 \text{ and } \mu_2 = \bar{y} + \frac{s_2}{\sqrt{m}} T_2$$

  where T_1 and T_2 are random variables with the same distribution as t_1 and t_2 above.

- We can see that $\mu_1 \overset{\text{fid.}}{\sim}$ Non-standard-T distribution with location-scale parameters $(\bar{x}, s_1/\sqrt{n})$, and the same goes for μ_2 with $(\bar{y}, s_2/\sqrt{m})$

- Now we can write

$$\psi = \mu_2 - \mu_1 = \bar{y} - \bar{x} + \frac{s_2}{\sqrt{m}} T_2 - \frac{s_1}{\sqrt{n}} T_1$$

- Which will imply (by switching to polar coordinates)

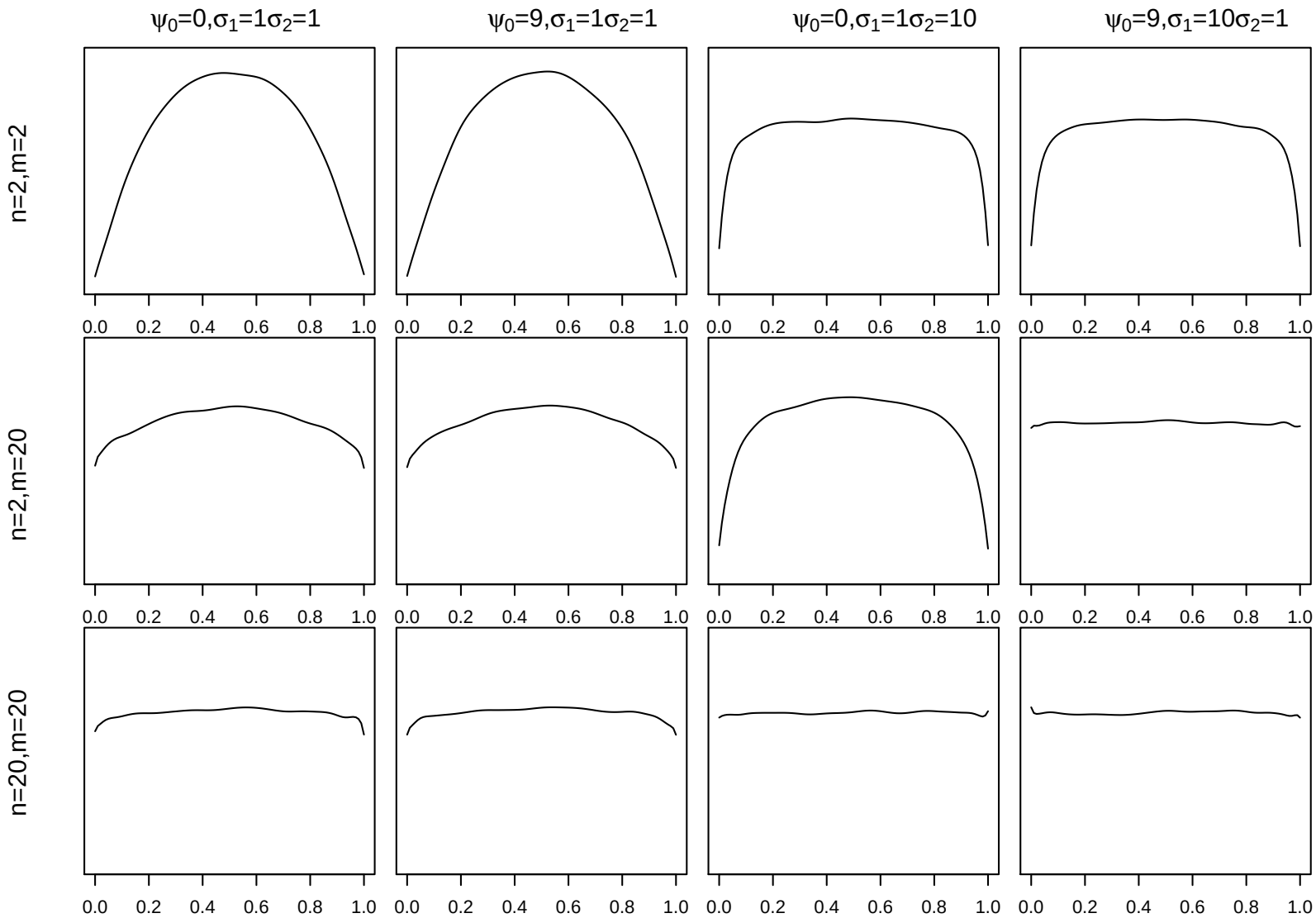
$$\frac{\psi - \hat{\psi}}{\sqrt{s_1^2/n + s_2^2/m}} = T_2 \cos\theta - T_1 \sin\theta,$$

  where $\hat{\psi} = \bar{y} - \bar{x}$ and $\tan\theta = (s_1/\sqrt{n})/(s_2/\sqrt{m})$
- The statistic on the left-hand side is distributed as the convolution of two T-distributions – the Behrens Fisher distribution (BF)!
- BF-distribution doesn't have a nice parametrization, but there are R-packages that will compute critical values.
  - Or, if you're old school, there are books with tables.
- It is also approximately T-distributed with degrees of freedom $\hat{\nu}$ estimated from the data (Welch 1938).

- Now, letting $B_{n,m,\theta}(\cdot)$ denote the cdf of the BF distribution, we have the fiducial distribution of ψ by

$$C(\psi) = B_{n,m,\theta}\left(\frac{\psi - \hat{\psi}}{\sqrt{s_1^2/n + s_2^2/m}}\right)$$

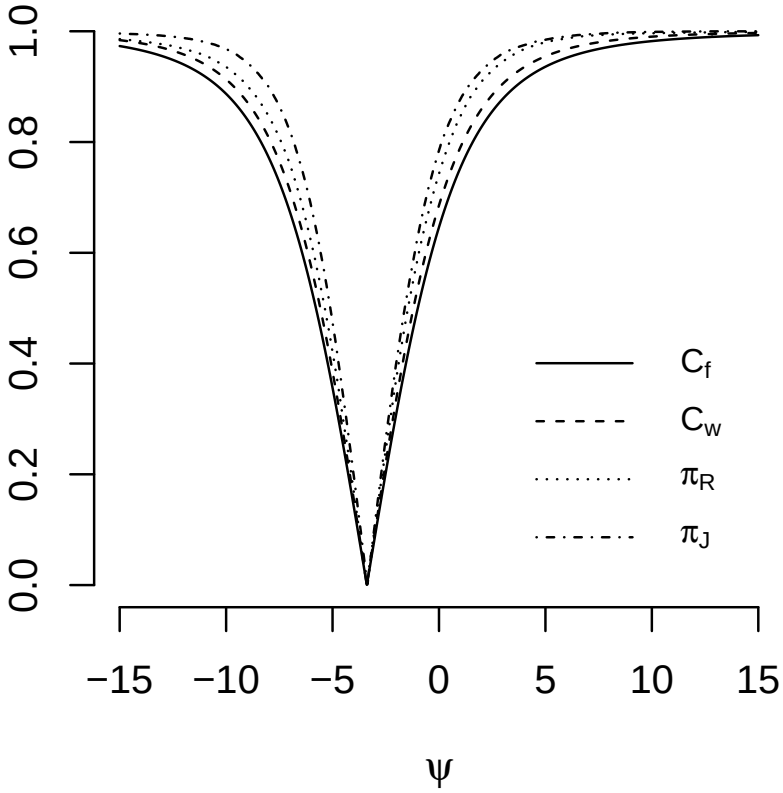
- It was pointed out, by Bartlett (1936), that testing $H_0 : \psi = 0$ leads to a test with wrong level of significance.
- In fact, it turns out that $C(\psi_0) \not\sim \text{Unif}(0, 1)$
  - We were kinda relying on the quantiles being exact CIs for inferential validity.
- Could hope that $C(\psi_0) \approx \text{Unif}(0, 1)$ as sample size increases?

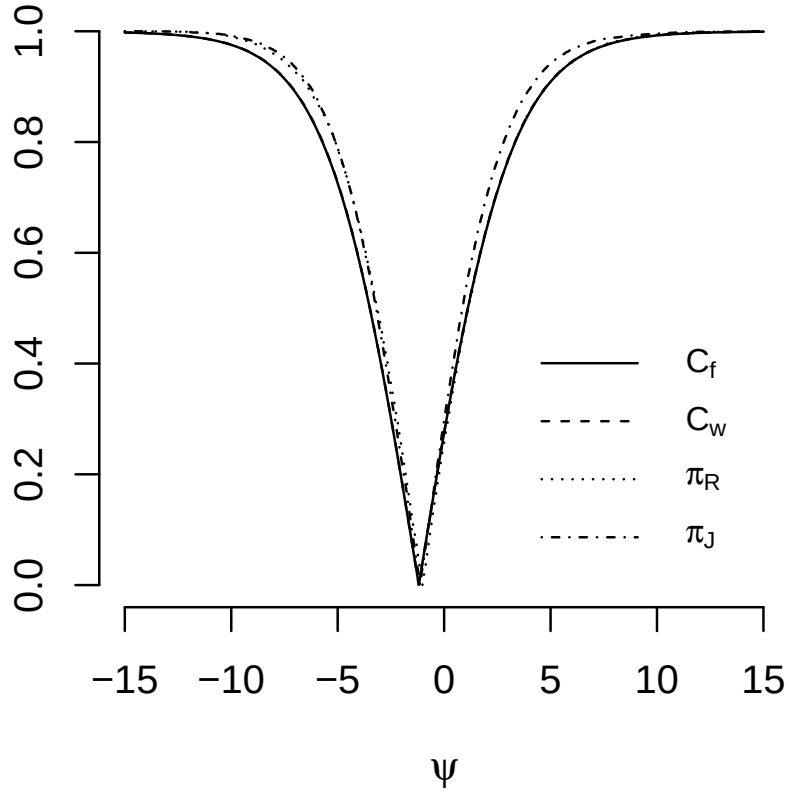

ψ0=0,σ1=1σ2=1
ψ0=9,σ1=1σ2=1
ψ0=0,σ1=1σ2=10
ψ0=9,σ1=10σ2=1
n=2,m=2
n=2,m=20
n=20,m=20
0.0 0.2 0.4 0.6 0.8 1.0

BF – Objective Bayesians

- The objective Bayesian analysis is straight forward.
- Use the reparametrization $(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2) \mapsto (\psi, \lambda, \sigma_1^2, \sigma_2^2)$, where $\psi = \mu_2 - \mu_1$ and $\lambda = \mu_1 + \mu_2$
  - Jeffreys' prior: $\pi_J \propto (\sigma_1 \sigma_2)^{-3}$
  - Reference prior: $\pi_R \propto (\sigma_1 \sigma_2)^{-2}$
  - As a side note, the prior $\pi \propto (\sigma_1 \sigma_2)^{-1}$ leads to the Fiducial solution (Jeffreys 1961)

A numerical comparison





In the left panel, n=m=5, with $(\mu_1, \mu_2, \sigma_1, \sigma_2) = (2, 4, 2, 4)$, while in the right panel n=m=10 samples from each distribution with the same means, but more unbalanced standard deviations $(\sigma_1, \sigma_2) = (1, 9)$.

- Overall, the performance of Fisher's solution isn't that bad!
- Distributions are very similar to the ones from the objective Bayesian scheme,even at low sample sizes!
- Simulations seem to indicate that the fiducial solution is *conservative*, i.e. that an α-level fiducial set, has *at least* α-level coverage.
- Included in the previous graphs are Welch's (1938) solution C_w using effective degrees of freedom – might be easier to teach to new students, but not exact – and it's derivation is a lot harder to grasp than the fiducial argument.

Paired Exponentials

- Consider independent pairs (X_i, Y_i) $i = 1, \ldots, n$, where $X_i \sim \text{Expo}(\theta)$ and $Y_i \sim \text{Expo}(\theta + \psi)$
- Interest is on ψ, while θ is considered a nuisance parameter.
- Schweder & Hjort (2016):

$$C(\psi) = P_\psi \left(\sum_{i=1}^{n} Y_i \leq \sum_{i=1}^{n} y_{i,obs} \middle| \sum_{i=1}^{n} (X_i + Y_i) = \sum_{i=1}^{n} (x_{i,obs} + y_{i,obs}) \right),$$

  is an (optimal!) **confidence distribution**!
- This is easy enough to calculate using some MCMC scheme.

Objective Bayes analysis

- Jeffreys' prior: Fine!

$$\pi_J(\psi, \theta) \propto \frac{1}{\theta(\theta + \psi)}$$

- Reference prior... is another story
  - Step 1: Keep ψ fixed, and derive reference prior for θ – easy! Simply Jeffreys'

$$\pi(\theta|\psi) = \left(\frac{1}{\theta^2} + \frac{1}{(\theta + \psi)^2} \right)^{1/2}$$

- Step 2: Obtain marginal distribution

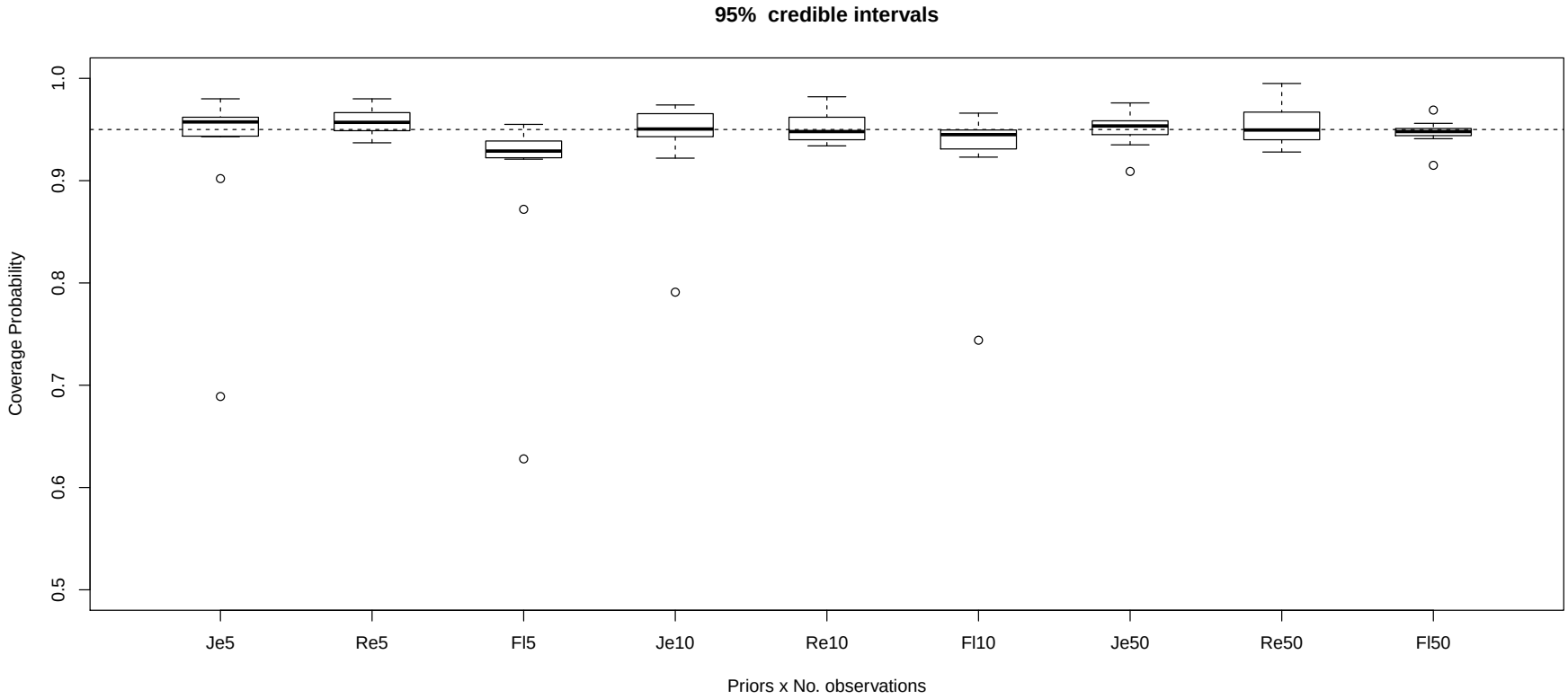
$$p(\boldsymbol{x}, \boldsymbol{y}|\psi) = \int_{\Theta} p(\boldsymbol{x}|\theta, \psi)\pi(\theta|\psi)\mathrm{d}\theta,$$

  difficult, integrals don't really a closed form solution.
- My solution?
  - Enforce $\psi > 0$ *a priori*.
  - Reparametrize $(\psi, \theta) \mapsto (\psi, \lambda)$, where $\lambda = (\theta + \psi)/\theta$.
  - Approximate full parameter space of (Ψ, Λ) by a sequence of expanding compact subsets $\{\Psi_i \times \Lambda_i\}_{i=1}^{\infty}$
  - Resolve all integrals by Taylor expansions
  - Derive reference prior as a limit
- After some time:

$$\pi_R(\psi, \theta) \approx \psi^{-0.47} \left(\frac{1}{\theta^2} + \frac{1}{(\theta + \psi)^2} \right)^{1/2},$$

  - which I still think is incorrect!

- And even after all that work, it doesn't really do any better than Jeffreys' prior in terms of coverage probability.



- This example is interesting as the fiducial (CD) solution is quite easy, while the reference prior approach is incredibly difficult.
- Neither method should have any monopoly on inferential validity
- I hope that the fiducial methods keep delivering interesting solutions to hard problems, and that the statistical community is open for new ideas.
    - Even though they're old.

Thanks for your attention!