Principled Statistics for Data Science

Alastair Young

Department of Mathematics, Imperial College London

CUSO, February 2019

Part 1: Preliminaries

"If a statistical analysis is clearly shown to be effective at answering the questions of interest, it gains nothing from being described as principled."

Terry Speed, IMS Bulletin, 2016

Purpose of tutorial: to refute this assertion!

Principled approach to statistical inference, especially in data science context, is essential, to avoid erroneous conclusions, in particular invalid statements about significance.

"The statistician cannot excuse himself from the duty of getting his head clear on the principles of scientific inference, but equally no other thinking man can avoid a like obligation."

R.A. Fisher, *The Design of Experiments*, 1935.

R.A. Fisher, 1953

Statistical Inference: $p-$values and CIs

Inference versus prediction?

Reasons for focus on inference:

- Interested in identifying significant 'features';
- Reproducibility crisis in science demands attention to properties of inferential methods.

Statistical inference

Have data y, assumed to be observed value of random variable Y.

An assumed (family of) models specifies the density of Y to be

$$f_Y(y; \theta),$$

where $\theta \in \Omega_\theta$ is the unknown parameter, about which we wish to draw inductive conclusions.

Two broad approaches: frequentist and Bayesian.

Frequentist

No further probabilistic assumptions: parameter value θ which gave rise to y treated as unknown constant.

Arguments involve probability only via long-run frequency interpretation: repeated sampling principle. Inference from y founded on analysis of variations in data samples seen under repeated (hypothetical) repetitions of experiment giving rise to y ('data sets we might have seen instead of y').

Neyman-Pearson versus Fisher.

Fisherian

Emphasis on statistical inference as a summary of data.

In order to be as relevant as possible to actual data y, must condition on everything that is known and uninformative about θ.

Formalize later as key directing principle of statistics for contemporary applications.

Neyman-Pearson

Emphasis on inferential procedures as decision problems.

Clarity of mathematical formulation, optimum inference procedures should be identified before y is available, optimality defined explicitly in terms of repeated sampling principle (e.g. maximise 'power', probability under repeated sampling that we correctly reject an incorrect hypothesis about θ).

Bayesian

We treat θ as having a probability distribution, both with and without the data y. Consider θ as the realised but unobserved value of a random variable Θ.

The prior distribution, expressed as density $\pi(\theta)$, summarizes information about Θ not arising from y.

Inference is drawn from posterior distribution of Θ, given y. By Bayes' Theorem, the posterior density $\pi(\theta|y) \propto \pi(\theta)f(y|\theta)$, where $f(y|\theta) \equiv f_Y(y;\theta)$ is 'likelihood function'.

A simple (artificial) example

Two observations Y_1 and Y_2 are taken, with

$$Y_i = \begin{cases} \theta + 1, & \text{with probability } 1/2, \\ \theta - 1, & \text{with probability } 1/2. \end{cases}$$

Suppose we are given the following proposal, as a confidence set for the unknown θ:

$$C(Y_1, Y_2) = \begin{cases} \text{the point } \{(Y_1 + Y_2)/2\}, & \text{if } Y_1 \neq Y_2, \\ \text{the point } \{Y_1 - 1\}, & \text{if } Y_1 = Y_2. \end{cases}$$

How do the Frequentist, Fisherian and Bayesian proceed?

Frequentist

The frequentist calculates that this is a 75% confidence set of smallest size for θ:

$$P_\theta(C(Y_1, Y_2) \text{ contains } \theta) = 0.75.$$

End of story.

Fisherian

This is not sensible to report, once the data is at hand.

If $y_1 \neq y_2$ we know for certain that their average is equal to θ, so the confidence set is actually 100% accurate.

If $y_1 = y_2$, we do not know if θ is the (common value $+1$), or the (common value -1): both are equally likely.

So...

To obtain sensible frequentist answers, define the conditioning statistic $S = |Y_1 - Y_2|$.

This is a measure of the strength of evidence in the data: $S = 2$ for data with maximal evidential content, $S = 0$ for data of minimal evidential content.

We define frequentist coverage conditional on the strength of evidence S:

$$P_\theta(C(Y_1, Y_2) \text{ contains } \theta | S = 2) = 1,$$

$$P_\theta(C(Y_1, Y_2) \text{ contains } \theta | S = 0) = \frac{1}{2}.$$

Same unconditional, repeated sampling, property as frequentist's analysis.

Report 100% confidence half the time and 50% confidence half the time, averaging 75% overall.

Bayesian

'Objective' Bayesian approach assigns θ an (uninformative) uniform prior, then calculates the posterior probability of $C(Y_1, Y_2)$ as 1 if $y_1 \neq y_2$, and 0.5 if $y_1 = y_2$.

End of story.

Key desiderata of statistical methods

- Validity: whether a claimed criterion or assumption is satisfied, regardless of the true unknown state of nature.
- Relevance: whether the analysis performed is relevant to the question of interest for the particular case at hand, the actual observed data.

Validity

Most appropriate to consider in the context of procedures motivated by the principle of error control.

A valid statistical procedure is one for which there is negligible probability that the procedure has a higher error rate than stated.

For example, the set $\mathcal{C}_{1-\alpha}$ is a (approximately) valid $(1-\alpha)$ confidence set for parameter θ if $P(\theta \notin \mathcal{C}_{1-\alpha}) = \alpha + \epsilon$ for some very small (negligible) ϵ, whatever the true value of θ.

This inferential correctness to be achieved by having accurate estimates of sampling distributions used in construction of $p-$values and CIs.

Relevance: Fisherian proposition

Appropriate conditioning of the hypothetical data samples that are the basis of non-Bayesian statistics.

The Conditionality Principle would advocate that the hypothetical repetitions should be conditioned on certain features of the available data sample, to ensure relevance to that actual data sample.

Framing: Cox & Mayo, 2010

Suppose for testing a null hypothesis $H_0 : \psi = \psi_0$ on an interest parameter ψ we calculate the observed value t_{obs} of a test statistic T and the associated $p-$value $p = P(T \geq t_{obs}; \psi = \psi_0)$.

If p is very low, e.g. 0.005, t_{obs} is grounds to reject H_0 or infer discordance with H_0 in direction of specified alternative, at level 0.005.

Rationale

- (1) To do so is to follow decision process with low Type 1 error rate, in the long run: if we treat the data as just decisive evidence against H_0, then in hypothetical repetitions, H_0 would be rejected in a proportion p of the cases when it is actually true.

- (2) [What we actually want]. To do so is to follow a rule where the low value of p corresponds to the actual data sample suggesting inconsistency with H_0.

Evidential construal in (2) only accomplished to extent it can be assured that small observed $p-$value is due to actual data-generating process being discrepant from that described by H_0. Once requirements of (2) satisfied, low error-rate rationale (1) follows.

Key

Ensure relevancy of sampling distribution on which $p-$values are based.

A (contrived) example

Suppose Y is distributed as $N(\theta, 1)$, and we have null hypothesis $H_0 : \theta = -3$, to be tested against alternative $H_1 : \theta = 3$.

Since $P(N(-3, 1) \geq 0) = 0.00135$, a rule which says 'Reject H_0 if $Y \geq 0$' has Type 1 (and, indeed, Type 2) error rate, under repeated sampling, $= 0.00135$.

Given actual data value $y = 0$, am I happy to assert that that value is strongly indicative of H_0 being false?

The $p-$value is 0.00135, but this value $y = 0$ is equally plausible under H_0 and H_1.

Sufficiency Principle

If $S \equiv S(Y)$ is a statistic such that the conditional distribution of Y given $S = s$ does not depend on θ for all s, θ, then S is sufficient for θ. S is minimal sufficient if it is a function of every other sufficient statistic.

(Uncontroversial) Sufficiency Principle says that if two data samples y and y' have $S(y) = S(y')$ then identical inferences about θ should be drawn from y and y'.

(Strong) Likelihood Principle

Two different random systems, the first giving observations y corresponding to a random variable Y and the second giving observations z on a random variable Z, the corresponding densities being $f_Y(y;\theta)$ and $f_Z(z;\theta)$, with the same parameter θ and the same parameter space Ω_θ.

The (strong) likelihood principle is that if y and z give proportional likelihood functions, the conclusions drawn from y and z should be identical, assuming adequacy of both models.

If, for all $\theta \in \Omega_\theta$,

$$f_Y(y;\theta) = h(y,z) f_Z(z;\theta),$$

identical conclusions about θ should be drawn from y and z.

Formal statement of CP

Suppose we may partition the minimal sufficient statistic for a model parameter θ of interest as $S = (T, A)$, where T is of the same dimension as θ and the random variable A is distribution constant: the statistic A is said to be ancillary.

Then, the Conditionality Principle says that inference should be based on the conditional distribution of T given $A = a$, the observed value in the actual data sample.

An example

Suppose Y_1, Y_2 are independent Poisson variables with means $(1-\psi)\lambda, \psi\lambda$, where λ is a known constant.

There is no reduction by sufficiency, but the random variable $A = Y_1 + Y_2$ has a known distribution, Poisson of mean λ, not depending on ψ. Inference would, say, be based on the conditional distribution of Y_2, given $A = a$, which is binomial with index a and parameter ψ.

Relaxation

The requirement that A be distribution constant is often relaxed.

Well-established in statistical theory that to condition on the observed data value of a random variable whose distribution does depend on θ might, under some circumstances, be convenient and meaningful, though this would in some sense sacrifice information on θ.

Nuisance parameter context

Extended notion of conditioning is most explicit in problems involving nuisance parameters, where the model parameter θ is partitioned as $\theta = (\psi, \lambda)$, with ψ of interest and λ a nuisance parameter.

Extended CP

Suppose that the minimal sufficient statistic can again be partitioned as $S = (T, A)$, where the distribution of T given $A = a$ depends only on ψ.

Extend the Conditionality Principle to advocate that inference on ψ should be based on this latter conditional distribution, under appropriate conditions on the distribution of A.

The case where the distribution of A depends on λ but not on ψ is just one rather special instance.

Applications

Justifications for many standard procedures of applied statistics, such as analysis of 2×2 contingency tables, derive from the Conditionality Principle, even when A has a distribution that depends on both ψ and λ, but when observation of A alone would make inference on ψ imprecise.

Contingency Table

Inference on the log-odds ratio when comparing two binomial variables.

Have Y_1, Y_2 independent binomial random variables corresponding to the number of successes in (m_1, m_2) independent trials, with success probabilities (θ_1, θ_2). The interest parameter is $\psi = \log\{\theta_2/(1-\theta_2)\} - \log\{\theta_1/(1-\theta_1)\}$. Inference on ψ would, following the Conditionality Principle, be based on the conditional distribution of Y_2 given $A = a$, where $A = Y_1 + Y_2$ has a marginal distribution depending in a complicated way on both ψ *and* whatever nuisance parameter λ is defined to complete the parametric specification.

Remarks

- Conditioning an inference on the observed data value of a statistic which is, to some degree, informative about the parameter of interest is an established part of statistical theory.
- Supported as a means of controlling (Type 1) error rate, while ensuring relevance to the data sample under test.
- Generally, conditioning will run counter to the objective of maximising power (minimising Type 2 error rate), which is a fundamental principle of much of frequentist statistical theory.
- Loss of power due to adoption of a conditional approach to inference may be very slight.

An example of inconsequential power loss

Y is normally distributed as $N(\theta, 1)$ or $N(\theta, 4)$, depending on whether the outcome δ of tossing a fair coin is heads ($\delta = 1$) or tails ($\delta = 2$).

To test the null hypothesis $H_0 : \theta = -1$ against the alternative $H_1 : \theta = 1$, controlling the Type 1 error rate at level $\alpha = 0.05$.

The most powerful unconditional test has rejection region given by $Y \geq 0.598$ if $\delta = 1$ and $Y \geq 2.392$ if $\delta = 2$.

CP advocates that we should condition on the outcome of the coin toss, δ. Then, given $\delta = 1$, the most powerful test of the required Type 1 error rate rejects H_0 if $Y \geq 0.645$, while, given $\delta = 2$ the rejection region is $Y \geq 2.290$.

The power of the unconditional test is 0.4497, while the power of the more natural conditional test is 0.4488, only marginally less.

A problem?

Birnbaum: Sufficiency Principle together with some form of Conditionality Principle implies the Strong Likelihood Principle, essentially incompatible with non-Bayesian statistics.

Much debated, of little consequence for statistical practice.

Neyman-Pearson Theory

Further support for conditioning (to eliminate nuisance parameters) provided by 'Neyman-Pearson theory' of optimal frequentist inference.

Key context

Parameter of interest is a component of the canonical parameter in a multiparameter exponential family model. Suppose Y has a density of the form

$$f(y;\theta) \propto h(y)\exp\{\psi T_1(y) + \lambda T_2(y)\}.$$

Then (T_1, T_2) is minimal sufficient and the conditional distribution of $T_1(Y)$, given $T_2(Y) = t_2$, say, depends only on ψ. The distribution of $T_2(Y)$ may, in special cases, depend only on λ, but will, in general, depend in a complicated way on both ψ and λ.

The extended form of the CP argues that inference should be based on the distribution of $T_1(Y)$, given $T_2(Y) = t_2$.

But, in Neyman-Pearson theory this same conditioning is justified by a requirement of full elimination of dependence on the nuisance parameter λ, achieved in the light of completeness of the minimal sufficient statistic only by this conditioning.

The resulting conditional inference is actually optimal, in terms of furnishing a uniformly most powerful unbiased test on the interest parameter ψ.

Central thesis

Same Fisherian principles of conditioning are necessary to steer appropriate statistics in a Data Science era, when models and the associated inferential questions are typically arrived at after examination of data.

> *"Data science does not exist until there is a dataset".*

Designed experiments typically not relevant. Notion of hypothetical data-generating process, statistical analysis to provide summary of data OK, but shouldn't emphazise repeated sampling properties of inference as in Neyman-Pearson theory.

Conditioning is needed to ensure validity of the methods used. Importantly, the justifications used for conditioning are not new, but mirror the arguments used in established statistical theory.

Classical or selective inference?

Key issue: what is appropriate framework for statistical analysis? Classical or selective?

Classical (frequentist) statistical inference

The analyst specifies the model, as well as the hypothesis to be tested, in advance of examination of the data. A classical $\alpha-$level test for the specified hypothesis H_0 under the specified model M must control the Type 1 error rate

$$P_{H_0}(\text{reject } H_0) \leq \alpha,$$

when model M holds.

Paradigm for (frequentist) statistics in Data Science: 'Post-selection Inference'

Lee et al., 2016; Fithian, Sun & Taylor, 2014.

Inference after having arrived at a statistical model adaptively, through examination of observed data.

Having selected a model $\hat{M}$ based on our observed data y, we wish to test a hypothesis $\hat{H}_0$. The notation here stresses that $\hat{H}_0$ will be random, a function of the selected model and hence of the data y: $\hat{M} \equiv \hat{M}(y), \hat{H}_0 \equiv \hat{H}_0(y)$. The key principle is expressed in terms of selective Type 1 error:

$$P_{\hat{H}_0}(\text{reject } \hat{H}_0 | (\hat{M}, \hat{H}_0)) \leq \alpha.$$

Discussion

We want to control the Type 1 error rate of the test **given it was actually performed.**

The thinking leading to this principle is really just a 21st century re-expression of Fisherian thought.

Example: File Drawer Effect

Suppose data consists of a set of n independent observations Y_i distributed as $N(\mu_i, 1)$. We choose, however, to focus attention only on the apparently large effects, selecting for formal inference only those indices i for which $|Y_i| > 1$, $\hat{I} = \{i : |Y_i| > 1\}$.

We wish, for each $i \in \hat{I}$, to test $H_{0,i} : \mu_i = 0$, each individual test to be performed at significance level $\alpha = 0.05$.

A test which rejects $H_{0,i}$ when $|Y_i| > 1.96$ is invalidated by the selection of the tests to be performed.

Though the probability of falsely rejecting a given $H_{0,i}$ is certainly α, since most of the time that hypothesis is not actually tested, the error rate among the hypotheses that are actually selected for testing is much higher than α.

Letting n_0 be the number of true null effects and suppose that $n_0 \to \infty$ as $n \to \infty$, in the long run, the fraction of errors among the true nulls we test, the ratio of the number of false rejections to the number of true nulls selected for testing, tends to $P_{H_{0,i}}(\text{reject } H_{0,i} | i \in \hat{I}) \approx 0.16$.

Appropriate error control

The probability of a false rejection conditional on selection is the natural and controllable error criterion to consider. We see that

$$P_{H_{0,i}}(|Y_i| > 2.41 | |Y_i| > 1) = 0.05,$$

so that the appropriate test of $H_{0,i}$, given that it is selected for testing, is to reject if $|Y_i| > 2.41$.

Take Bayesian approach?

Conventional wisdom: Bayesian inference should not be altered by selection. Inference is provided conditionally on observed data, any further conditioning on the selection is redundant.

Or is it?

George & Yekutieli (2012), Yekutieli (2012): standpoint that inference doesn't need to be adjusted by selection only justified if selection takes place on parameter space as well as sample space.

If not ('fixed parameter'), Bayesian inference must appropriately account for selection.

The framework

Bayesian model with sampling distribution $f(y|\theta)$ and prior $\pi(\theta)$, with $\theta \in \Omega_\theta \subseteq \mathbb{R}^p$.

Let $\mathcal{E} = \{E_1, E_2, \ldots, E_m\}$ be a partition of sample space.

For each realisation of y we are interested in a parameter $h_i(\theta) \in \{h_1(\theta), h_2(\theta), \ldots, h_m(\theta)\}$ only if $y \in E_i$.

Selective inference: provide valid inference for $h_i(\theta)$, taking into account we are only interested in it because event E_i was observed.

Formalities

Inference about $h_i(\theta)$ should be based on the conditional density

$$f(y|\theta, E_i) = \frac{f(y|\theta)}{P(E_i|\theta)}\mathbb{I}(y \in E_i).$$

Suppose prior $\pi(\theta)$ represents a hypothetical sampling distribution from which values of parameter are generated.

Two cases

- If successive samples of y are generated from $f(y|\theta)$ for a common value of θ, sampled from $\pi(\theta)$, selection adjusted posterior is

$$\pi^{E_i}(\theta|y) \propto \pi(\theta) f(y|\theta, E_i).$$

'Fixed parameter'.

- If each sample y is generated from $f(y|\theta)$ for a different value of θ, sampled from prior, then values of parameter considered in last step are sampled conditionally on E_i, and selection adjusted posterior is

$$\begin{aligned} \pi^{E_i}(\theta|y) &\propto \pi(\theta|E_i) f(y|\theta, E_i) \propto \pi(\theta) P(E_i|\theta) f(y|\theta, E_i) \\ &\propto \pi(\theta) f(y|\theta) \propto \pi(\theta|y), \end{aligned}$$

posterior which ignores selection. 'Random parameter'.

An example

Suppose $\theta = (\theta_1, \theta_2)$ are independent $N(0, 1/2)$, and, conditional on θ, $Y = (Y_1, Y_2)$, with Y_1, Y_2 independent, Y_i distributed as $N(\theta_i, 1/n)$.

In non-selective framework, Bayes estimator of θ_2 is posterior mean of θ_2,

$$\frac{n}{n+2} Y_2,$$

with associated Bayes risk [average squared error over joint distribution of (θ, Y)] equal to posterior variance, $1/(2+n)$.

Suppose selection is applied: inference is only provided for θ_2 if $Y_2 \geq Y_1$.

Random parameter case

Generate, for range of n, 100,000 samples from joint distribution of (θ, Y), retain only those for which $Y_2 \geq Y_1$. Calculate average squared error of estimator $nY_2/(n+2)$ over retained samples: in each case proportion of retained samples very close to 0.5.

n	MSE	$1/(n+2)$
1	0.33346	0.33333
2	0.24970	0.25000
5	0.14251	0.14286
10	0.08300	0.08333
50	0.01918	0.01923
100	0.00978	0.00980

Selection has no effect on inference.

Fixed parameter case

Now:

1. Generate θ from assumed prior;
2. Generate Y from conditional distribution of $Y|\theta$, until $Y_2 \geq Y_1$.

Compare, over 100,000 replications, average squared errors of following estimators of θ_2:

NS Estimator $nY_2/(n+2)$ which ignores selection;

S Posterior mean of θ_2, calculated from selective posterior [constructed using truncated likelihood, and estimated by large MCMC for each replication].

MSEs

n	NS	S
1	0.59114	0.48261
2	0.60349	0.44810
5	0.58507	0.37434
10	0.55719	0.31507
50	0.51474	0.23238
100	0.50252	0.21647

Selection cannot be ignored.

Structure of remainder of course

- Likelihood-based methods of inference, as omnibus procedures in classical setting;
- Higher-order accuracy/validity by analytic methods, bootstrapping;
- Principles/illustrations of frequentist and Bayesian selective statistics.

Part 2: Higher-order Likelihood-based Inference

The classical parametric inference problem

Let $Y = \{Y_1, \ldots, Y_n\}$ be random sample from underlying distribution $F(y; \theta)$, indexed by d-dimensional parameter $\theta = (\theta^1, \ldots, \theta^d) = (\psi, \phi)$, ψ p-dimensional interest parameter, ϕ q-dimensional nuisance parameter, $p + q = d$. May have ϕ high-dimensional.

Wish to test $H_0 : \psi = \psi_0$, or (duality) construct confidence set for ψ.

If $p = 1$, $\psi = \theta^1$, want one-sided inference e.g. test H_0 against (one-sided) alternative $\psi > \psi_0$ or $\psi < \psi_0$, or one-sided confidence limit.

"Break the research question of interest into simple components corresponding to strongly focused and incisive research questions."

(D.R. Cox, 'Principles of Statistical Inference', 2006)

Typically, $p = 1$.

Inference

Let $L(\theta) \equiv L(\theta; Y)$ be log-likelihood, $\hat{\theta} = (\hat{\psi}, \hat{\phi})$ the overall MLE of θ, $\hat{\phi}_\psi$ the constrained MLE of ϕ, for fixed value of ψ. Write $\tilde{\theta} \equiv \tilde{\theta}(\psi) = (\psi, \hat{\phi}_\psi)$.

Profile log-likelihood function for ψ is $M(\psi) = L\{\tilde{\theta}(\psi)\}$.

Likelihood ratio statistic is $W(\psi) = 2\{M(\hat{\psi}) - M(\psi)\}$.

In case of scalar ψ, use signed root likelihood ratio statistic:

$$R(\psi) = \mathrm{sgn}(\hat{\psi} - \psi) W(\psi)^{1/2}.$$

Notation

Differentiation is indicated by subscripts, so $L_r(\theta) = \partial L(\theta)/\partial\theta^r$, $L_{rs}(\theta) = \partial^2 L(\theta)/\partial\theta^r\partial\theta^s$, etc. Then $E\{L_r(\theta)\} = 0$; let $\lambda_{rs} = E\{-L_{rs}(\theta)\}$.

The constants λ_{rs} are assumed to be of order $O(n)$. These assumptions are usually satisfied in situations involving independent observations, structured (e.g. time series) dependent data problems.

Let (λ^{rs}) be the $d \times d$ matrix inverse of (λ_{rs}).

A comment

Calculation of quantities required by methods to be described requires (at most) evaluation of expectations of log-likelihood derivatives.

Other statistics

Consider, for simplicity, scalar case $p = 1$. Variants for $p > 1$ easily defined.

As alternative 'pivots' to $R(\psi)$, could use, for example:

Wald statistic,

$$T_W(\psi) = (\hat{\psi} - \psi)\{\lambda^{11}(\hat{\theta})\}^{-1/2}.$$

Score statistic,

$$T_S(\psi) = L_1\{\tilde{\theta}(\psi)\}\{\lambda^{11}(\hat{\theta})\}^{1/2}.$$

Constructed using expected (inverse) information matrix $[\lambda^{rs}]$, evaluated at global MLE.

Alternatively: use observed (inverse) information matrix $[-L^{rs}]$; evaluate at constrained MLE $\tilde{\theta}(\psi)$,.....

Running Example (RE): Inverse Gaussian distribution

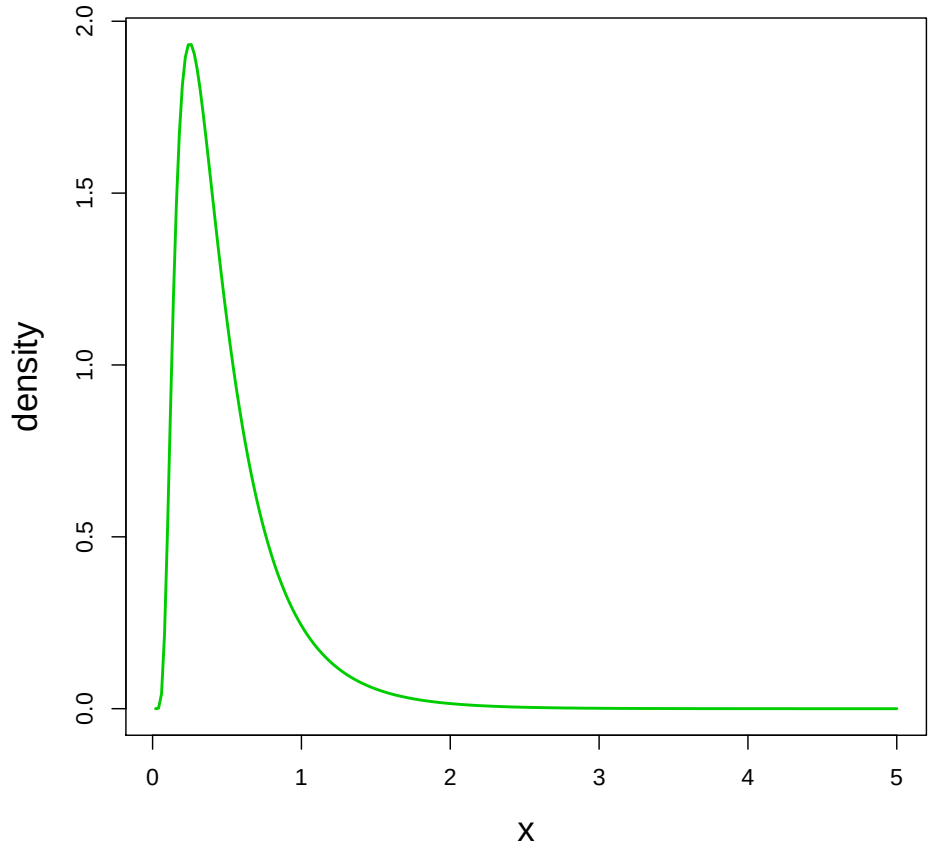
$Y_1, \ldots, Y_n$ IID inverse Gaussian, $IG(\mu, \psi)$, with density

$$f(y; \mu, \psi) = \left(\frac{\psi}{2\pi y^3}\right)^{1/2} \exp\left(-\frac{\psi}{2\mu^2 y}(y-\mu)^2\right), \; y > 0,$$

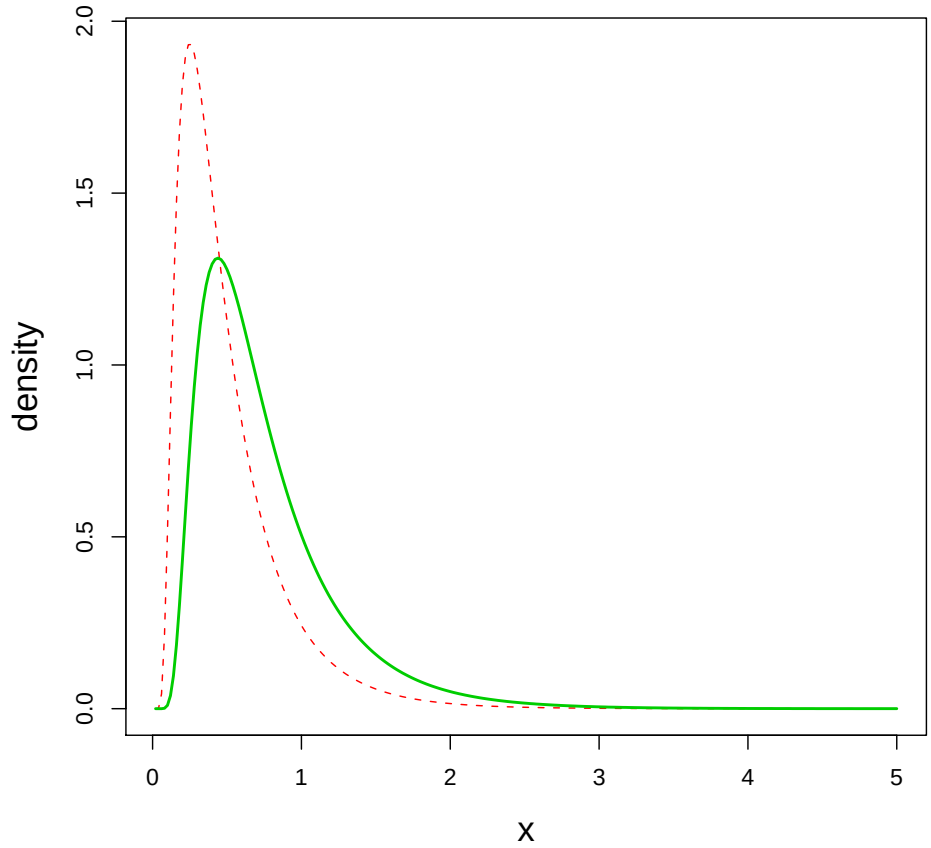
interest parameter is shape $\psi > 0$, mean $\mu > 0$ as nuisance.

First passage time of Brownian motion, widely used to model phenomena in biosciences/reliability/survival/....

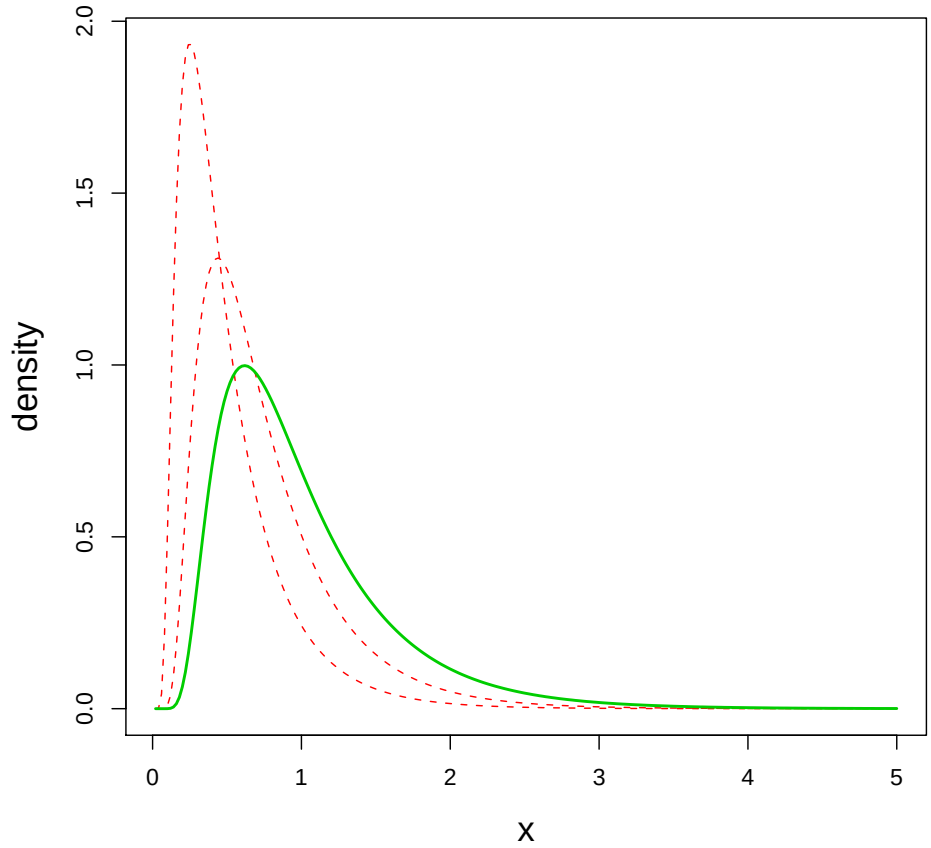
Density: $\psi = 1, \mu = 0.5$



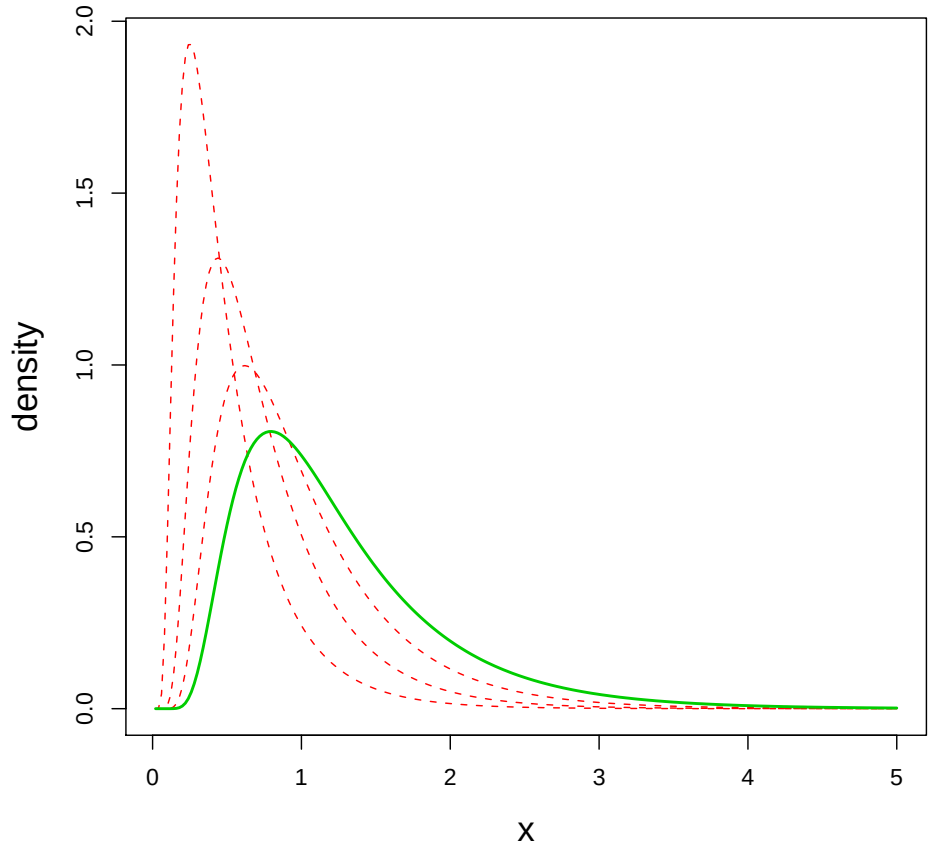
Density: $\psi = 2, \mu = 0.75$



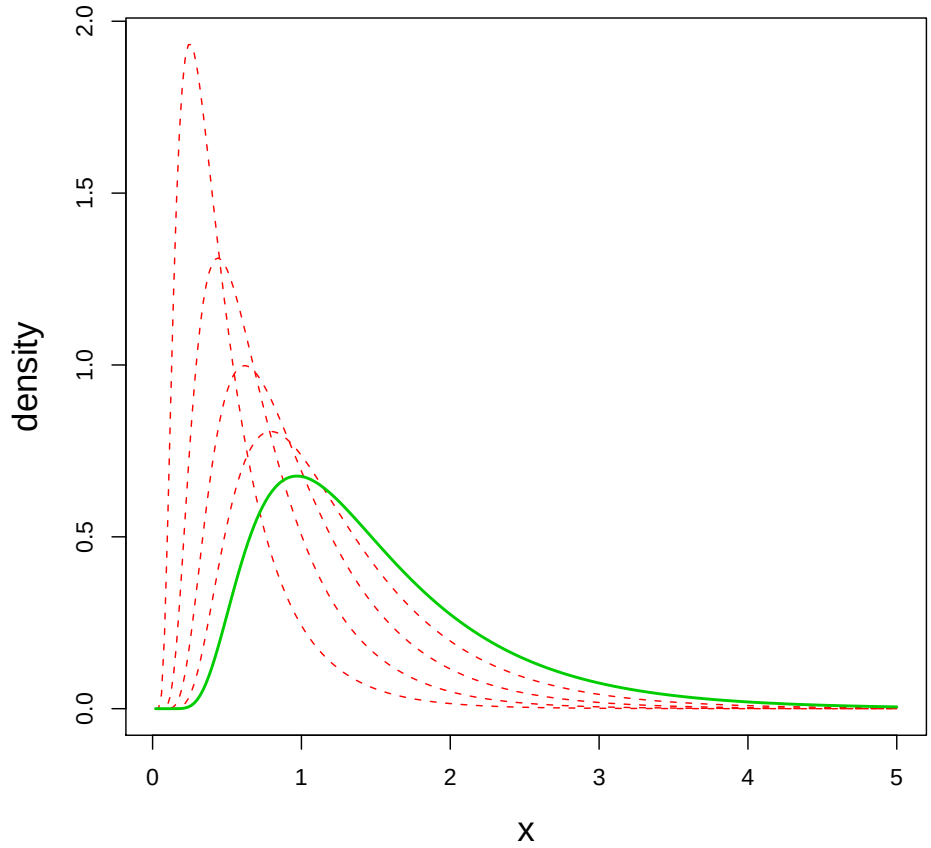
Density: $\psi = 3, \mu = 1.0$



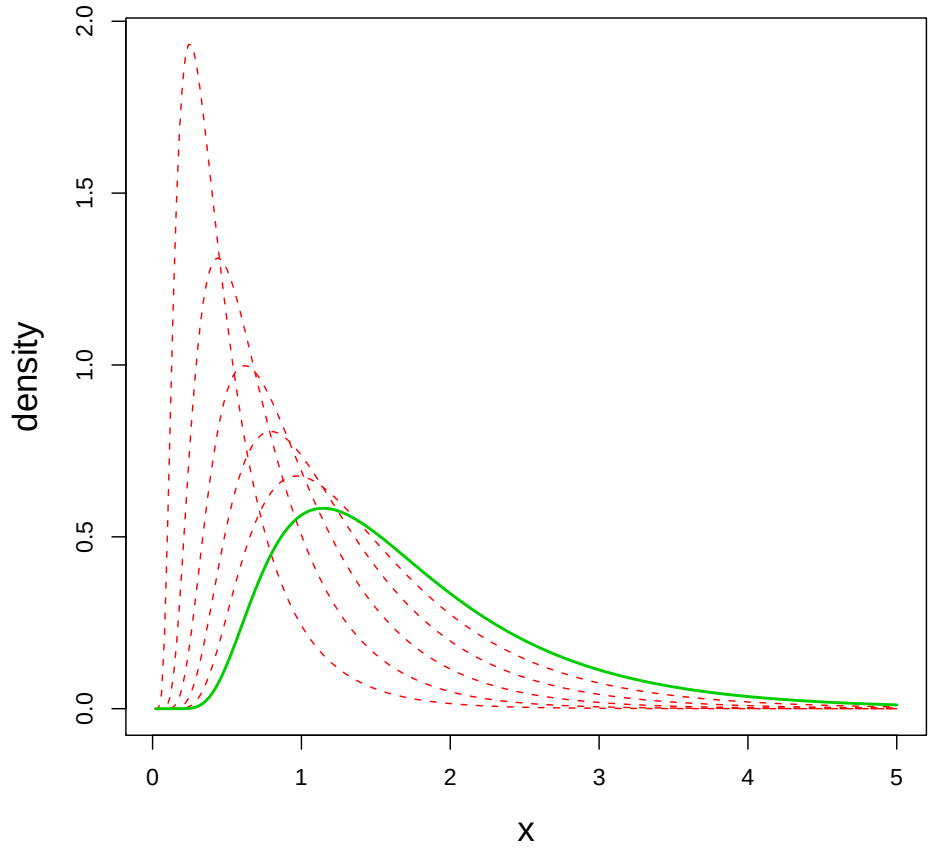
Density: $\psi = 4, \mu = 1.25$



Density: $\psi = 5, \mu = 1.5$



Density: $\psi = 6, \mu = 1.75$



MLES are:

$$\hat{\psi} = n/V,\ \hat{\mu} = \hat{\mu}_\psi = \bar{Y},$$

$$V = \sum_{i=1}^{n}(Y_i^{-1} - \overline{Y}^{-1}), \bar{Y} = n^{-1}\sum_{i=1}^{n} Y_i.$$

Distribution of ψV is χ^2_{n-1}, distribution of $\hat{\mu}$ is $IG(\mu, \psi)$.

$$
\begin{aligned}
R(\psi) &= \operatorname{sgn}(\hat{\psi}-\psi)\{n(\log\hat{\psi}-1-\log\psi+\psi/\hat{\psi})\}^{1/2}, \\
T_W(\psi) &= \sqrt{\frac{n}{2}}(1-\frac{\psi}{\hat{\psi}}), \\
T_S(\psi) &= \sqrt{\frac{n}{2}}(\frac{\hat{\psi}}{\psi}-1)
\end{aligned}
$$

A data sample

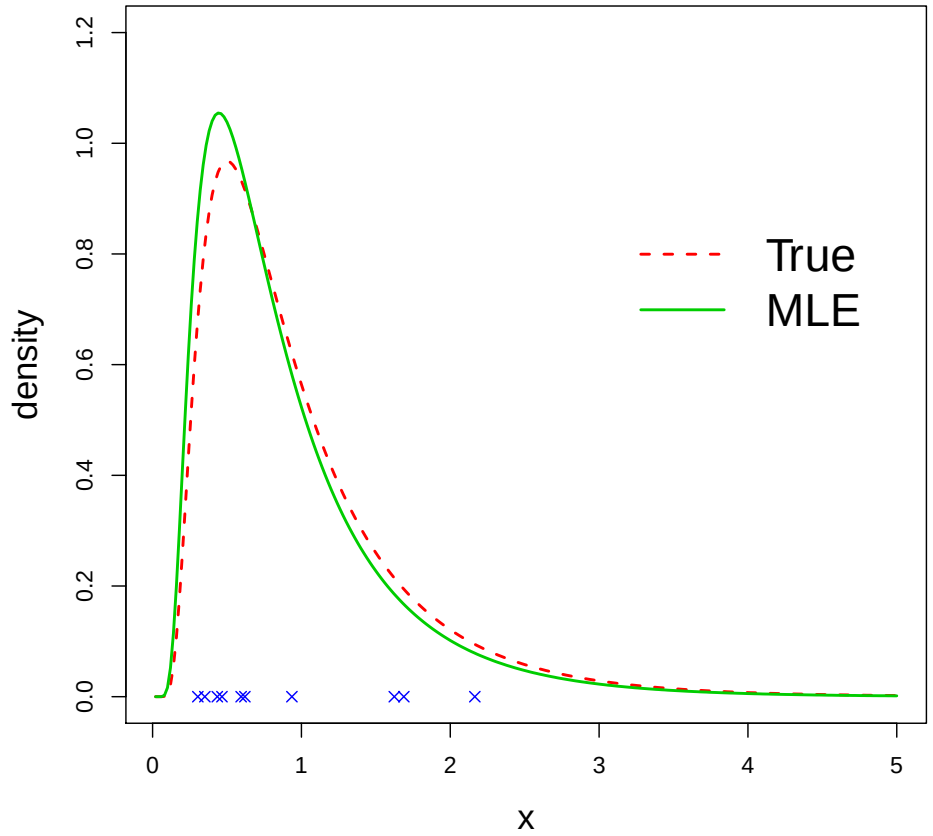
Data sample size $n = 10$, generated with $\mu = 1, \psi = 2$:

```
0.435, 0.466, 1.624, 0.304, 2.165

0.936, 0.620, 0.595, 0.351, 1.688
```

Have: $\hat{\psi} = 1.745, \hat{\mu} = 0.918$.

RE: True and estimated densities



Comment

Concentrate here on inference based on R, W, for simplicity. Most results true also for Wald and score statistics.

Parameterization invariance

Principle of parameterization invariance (PPI) important basis for choosing between different inferential procedures.

If θ and ζ are two alternative parameterizations and $\mathcal{P}(\cdot)$ is an inference procedure, with C_θ and C_ζ the conclusions that $\mathcal{P}(\cdot)$ leads to, expressed in the two parameterizations, then the same conclusion C_ζ should be reached by both application of $\mathcal{P}(\cdot)$ in the ζ parameterization and translation into the ζ parameterization of the conclusion C_θ.

Nuisance parameter

With nuisance parameters, parameterization invariance is restricted to mean invariance under interest respecting reparameterization.

Suppose $\theta = (\psi, \phi)$, with ψ interest parameter and ϕ nuisance parameter. An interest respecting reparameterization is of the form $\upsilon = \upsilon(\theta) = \upsilon(\psi, \phi)$ with $\upsilon = (\varphi, \chi)$, such that

$$\varphi = \varphi(\psi),\ \chi = \chi(\psi, \phi).$$

Reparameterization invariance helpful in practical sense: work on any numerically convenient scale and then transform back to the original one.

Implications of PPI

Inference based on $W(\psi)$ (or $R(\psi)$) does respect PPI.

So does inference based on $T_S(\psi)$, at least if constructed using expected information evaluated at constrained MLE $\tilde{\theta}(\psi)$.

Inference based on $T_W(\psi)$ does not.

First-order theory

Have $W(\psi)$ distributed as χ^2_p, to error of order $O(n^{-1})$.

Also, $R(\psi)$ distributed as $N(0,1)$, to error of order $O(n^{-1/2})$.

Latter true also for $T_W(\psi)$ and $T_S(\psi)$, and variants.

Inference: illustration, $p = 1$

A confidence set of asymptotic coverage $1 - \alpha$ for ψ is

$$\mathcal{I}(Y) \equiv \mathcal{I}_{1-\alpha}(Y) = \{\psi : u(Y, \psi) \leq 1 - \alpha\},$$

with $u(Y, \psi) = \Phi\{R(\psi)\}$, in terms of the $N(0, 1)$ distribution function $\Phi(\cdot)$. Call $u(Y, \psi)$ the 'significance function'.

Equivalently, the confidence set is

$$\mathcal{I}(Y) = \{\psi : R(\psi) \leq \Phi^{-1}(1 - \alpha)\}.$$

The coverage error of the confidence set is $O(n^{-1/2})$: first-order accuracy.

Have that $u(Y, \psi)$ is monotonic in ψ, so confidence set is semi-infinite interval of form $(\hat{\psi}_l(Y), \infty)$. Lower confidence limit.

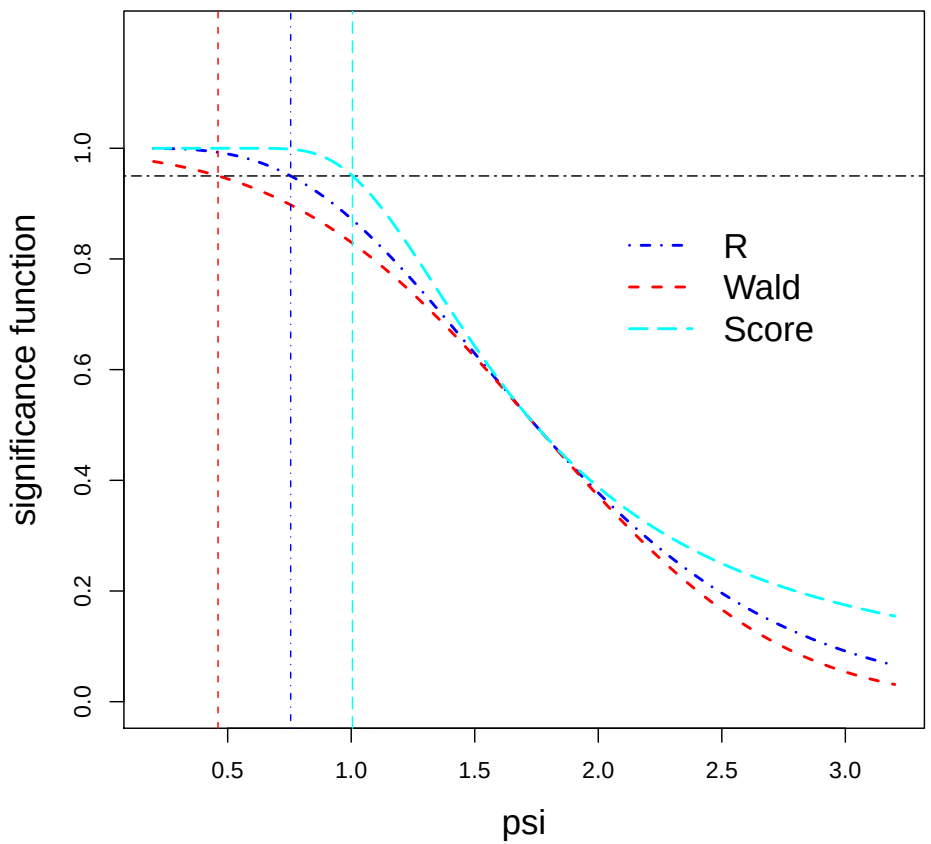
If two-sided inference is required, an equi-tailed two-sided confidence interval $\mathcal{J}(Y)$ of nominal coverage $1 - \alpha$ may be obtained by taking the set difference of two one-sided sets:

$$\mathcal{J}(Y) \equiv \mathcal{J}_{1-\alpha}(Y) = \mathcal{I}_{1-\alpha/2}(Y) \backslash \mathcal{I}_{\alpha/2}(Y).$$

Similar statements about coverage error of confidence sets true for other asymptotically $N(0, 1)$ pivots.

In case $p > 1$, confidence set of coverage error $O(n^{-1})$ (second-order accuracy) from χ^2_p approximation to sampling distribution of $W(\psi)$.

RE, data sample: significance functions



RE, data example: 95% confidence limits

- $R(\psi)$: interval is $(0.755, \infty)$.
- $T_W(\psi)$: interval is $(0.461, \infty)$.
- $T_S(\psi)$: interval is $(1.005, \infty)$.

Motivations for refinements

- To obtain higher-order repeated sampling accuracy.
- To accommodate appropriate conditioning: multi-parameter exponential families (conditioning dictated by theory of optimal tests etc.); ancillary statistic models (relevance, by conditioning on component of minimal sufficient statistic that is approximately distribution constant).

Refinements: approaches

Two most established approaches:

- Analytic procedures, 'small sample asymptotics', saddlepoint, related methods;
- Simulation ('bootstrap') methods.

The third way: objective Bayes

Bayes with prior explicitly specified so (marginal) posterior for ψ yields confidence limits with correct frequentist interpretation, to high-order: 'probability matching prior'.

- conceptually simple;
- typically awkward with high-dimensional nuisance parameter, as need to find marginal posterior of ψ;
- route not always open, higher-order (conditional) accuracy not necessarily obtainable.

Detail

Require prior $\pi(\psi, \phi)$ so that

$$Pr_\theta\{\psi \leq \psi^{(1-\alpha)}(\pi, Y)\} = 1 - \alpha + O(n^{-r/2}),$$

for $r = 2$ or 3, each $0 < \alpha < 1$.

Here:

- n is sample size;
- $\psi^{(1-\alpha)}(\pi, Y)$ is $(1 - \alpha)$ quantile of marginal posterior, given data Y, of ψ, under prior $\pi(\psi, \phi)$;
- Pr_θ denotes frequentist probability, under repeated sampling of Y, when true (fixed) parameter value is $\theta = (\psi, \phi)$.

Probability matching priors

If condition holds with $r = 2$, speak of $\pi(\psi, \phi)$ as first-order probability matching prior.

If condition holds with $r = 3$, speak of $\pi(\psi, \phi)$ as second-order probability matching prior.

Conditional probability matching

Appropriate frequentist inference to match in full exponential family or ancillary statistic context is the conditional one, conditional on the observed value c of some statistic $C(Y)$.

The requirement should be 'conditional probability matching':

$$Pr_\theta\{\psi \leq \psi^{(1-\alpha)}(\pi, Y) \mid C(Y) = c\} = 1 - \alpha + O(n^{-r/2}).$$

Want the posterior $1 - \alpha$ quantile to match the $1 - \alpha$ conditional frequentist confidence limit for ψ.

Analytic methods: the highlights

- Bartlett correction of likelihood ratio statistic $W(\psi)$.
- Analytically modified forms of $R(\psi)$, specifically designed to offer conditional validity, to high (asymptotic) order, in both contexts. 'Barndorff-Nielsen's R^*'.

Bartlett correction

Have

$$E_\theta\{W(\psi)\} = p\left(1 + \frac{b(\theta)}{n} + O(n^{-2})\right),$$

so modify $W(\psi)$ to

$$W_c(\psi) = W(\psi)/\{1 + b(\psi, \hat{\phi}_\psi)/n\},$$

or

$$\bar{W}_c(\psi) = W(\psi)/E_{(\psi,\hat{\phi}_\psi)}\{W(\psi)\}.$$

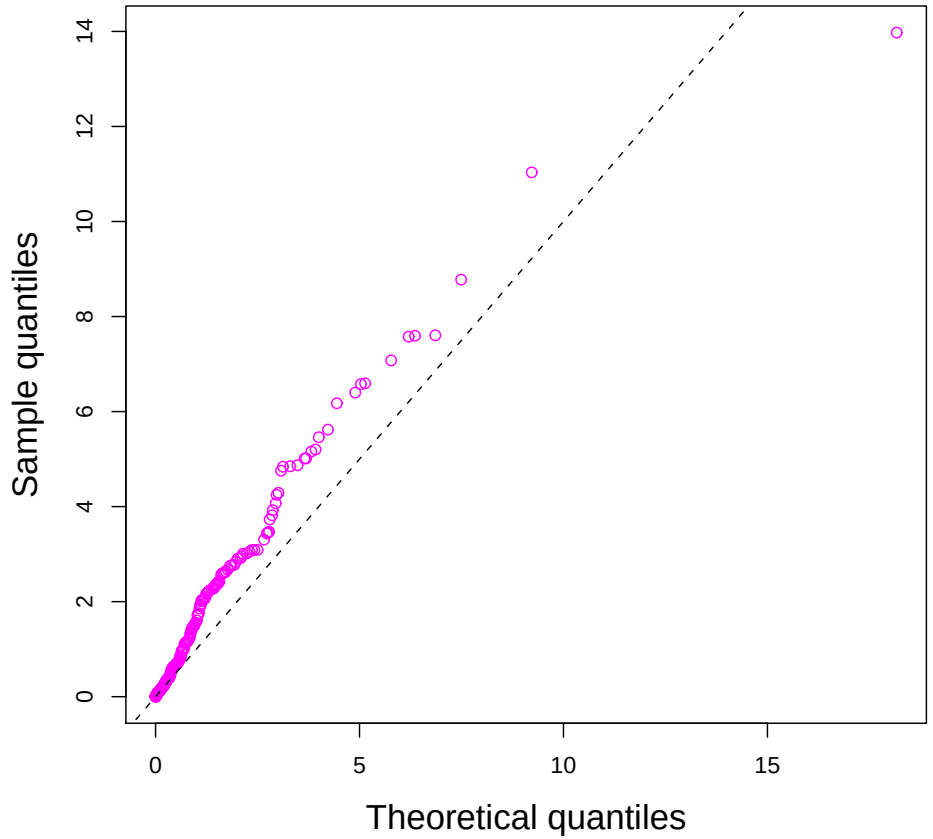
Then $W_c(\psi)$ and $\bar{W}_c(\psi)$ are distributed as χ^2_p, to error of order $O(n^{-2})$. Confidence sets constructed by χ^2_p approximation have coverage error $O(n^{-2})$.

$E_{(\psi,\hat{\phi}_\psi)}\{W(\psi)\}$ constructed by (bootstrap) simulation. Estimation of expectation requires smaller MC simulation that estimation of whole sampling distribution.

Inference by χ^2_p approximation to distribution of $\bar{W}_c(\psi)$: 'Empirical Bartlett correction'.

Could replace χ^2_p approximation to sampling distribution of $W(\psi)$ by bootstrap distribution: sampling distribution under sampling with parameter fixed as $\theta = (\psi, \hat{\phi}_\psi)$. Confidence set will also have coverage error $O(n^{-2})$.

RE: $n = 5, \psi = 2, \mu = 1.0$, χ^2_1 QQ plot, $W(\psi)$

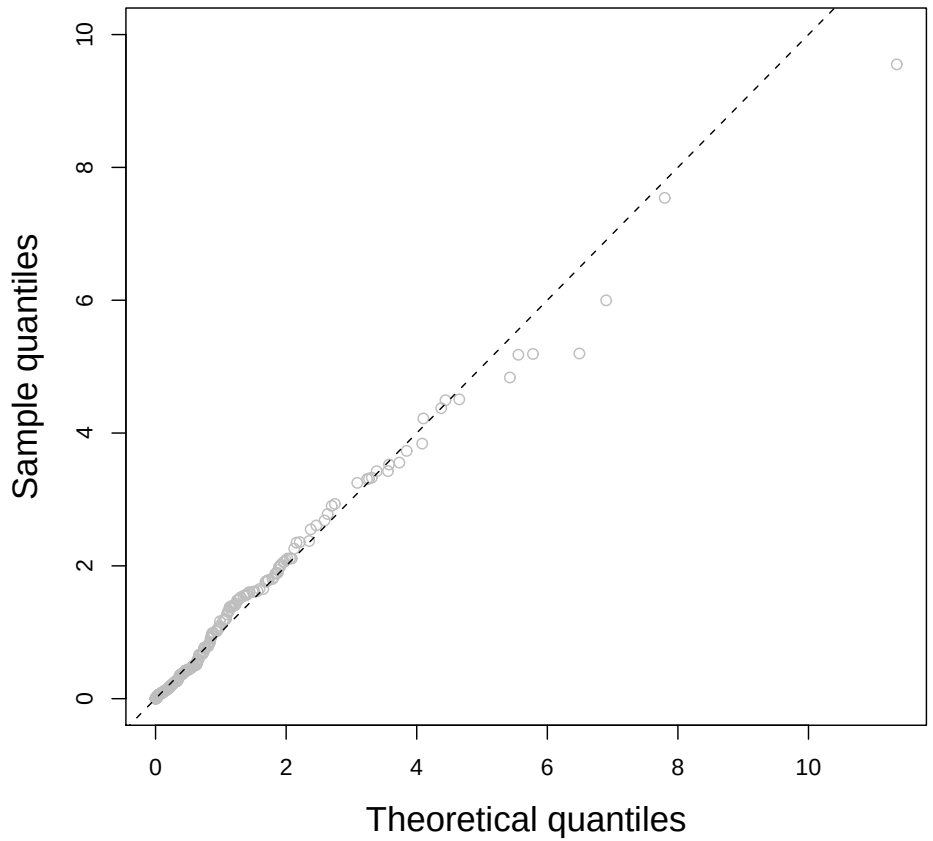


RE: $n = 5, \psi = 2, \mu = 1.0$

In inverse Gaussian example, $E_\theta\{W(\psi)\}$ does not depend on nuisance parameter μ.

Big simulation shows, $E_\theta\{W(\psi)\} = 1.4632$.

RE: $n = 5, \psi = 2, \mu = 1.0$, χ^2_1 QQ plot, $\bar{W}_c(\psi)$



Adjusted signed root statistic R^*

Defined by

$$R^*(\psi) = R(\psi) + \log\{v(\psi)/R(\psi)\}/R(\psi)$$

Here, in formulation considered, adjustment $v(\psi)$ necessitates:

- explicit specification of ancillary A in ancillary statistic (e.g. transformation) context;
- potentially awkward analytic calculations, in both ancillary/exponential family situations.

Adjustment function

Adjustment $v(\psi)$ is given by

$$v(\psi) = \left| \begin{array}{c} L_{;\hat{\theta}}(\hat{\theta}) - L_{;\hat{\theta}}(\tilde{\theta}) \\ L_{\phi;\hat{\theta}}(\tilde{\theta}) \end{array} \right| / \{|j_{\phi\phi}(\tilde{\theta})|^{1/2} |j(\hat{\theta})|^{1/2}\}.$$

Here, the log-likelihood function has been written as $L(\theta; \hat{\theta}, a)$, with $(\hat{\theta}, a)$ minimal sufficient and a ancillary, and

$$L_{;\hat{\theta}}(\theta) \equiv L_{;\hat{\theta}}(\theta; \hat{\theta}, a) = \frac{\partial}{\partial \hat{\theta}} L(\theta; \hat{\theta}, a),$$

$$L_{\phi;\hat{\theta}}(\theta) \equiv L_{\phi;\hat{\theta}}(\theta; \hat{\theta}, a) = \frac{\partial^2}{\partial \phi \partial \hat{\theta}} L(\theta; \hat{\theta}, a).$$

Also, j denotes the observed information matrix, $j(\theta) = (-L_{rs}(\theta))$, with $L_{rs}(\theta) = \partial^2 L(\theta)/\partial\theta^r \partial\theta^s$, and $j_{\phi\phi}$ denotes its (ϕ, ϕ) component.

Other formulations

Other formulations of $v(\psi)$, due to Fraser and co-workers, possible: use of 'tangent exponential model' avoids need to specify transformation $Y \to (\hat{\theta}, A)$.

Still analytically fiddly.

RE: adjustment function

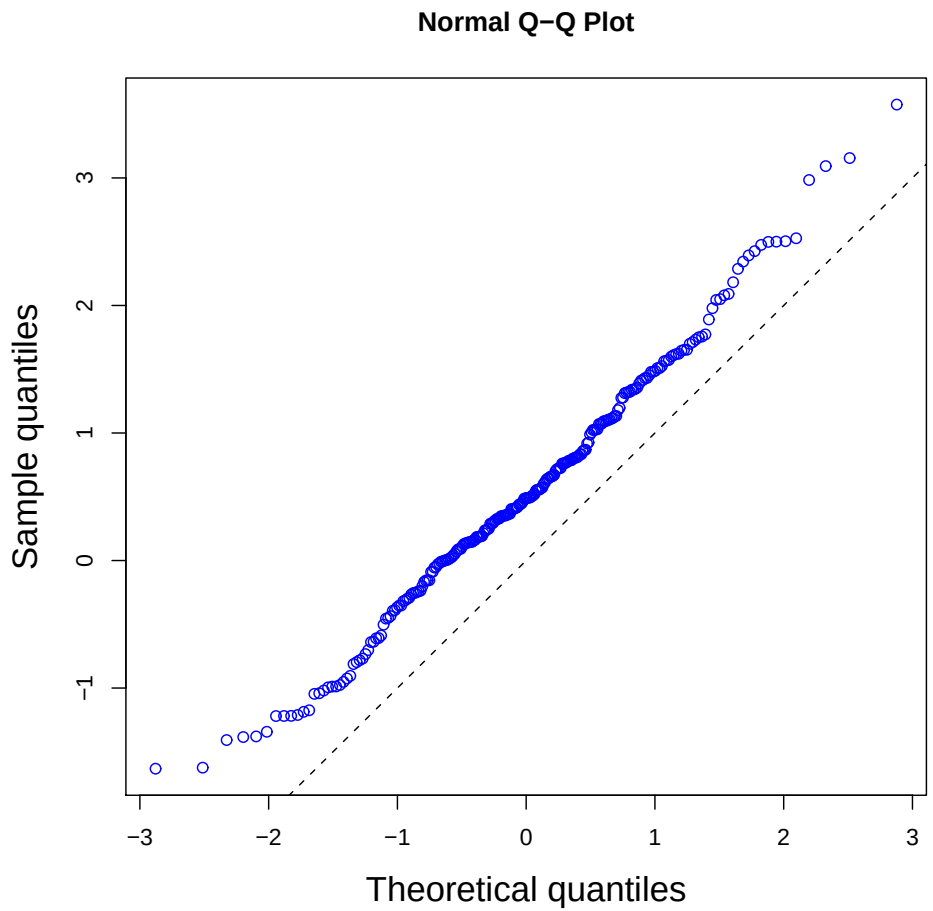
In inverse Gaussian example,

$$v(\psi) = \sqrt{\frac{n\psi}{2\hat{\psi}}}(1 - \frac{\psi}{\hat{\psi}}).$$

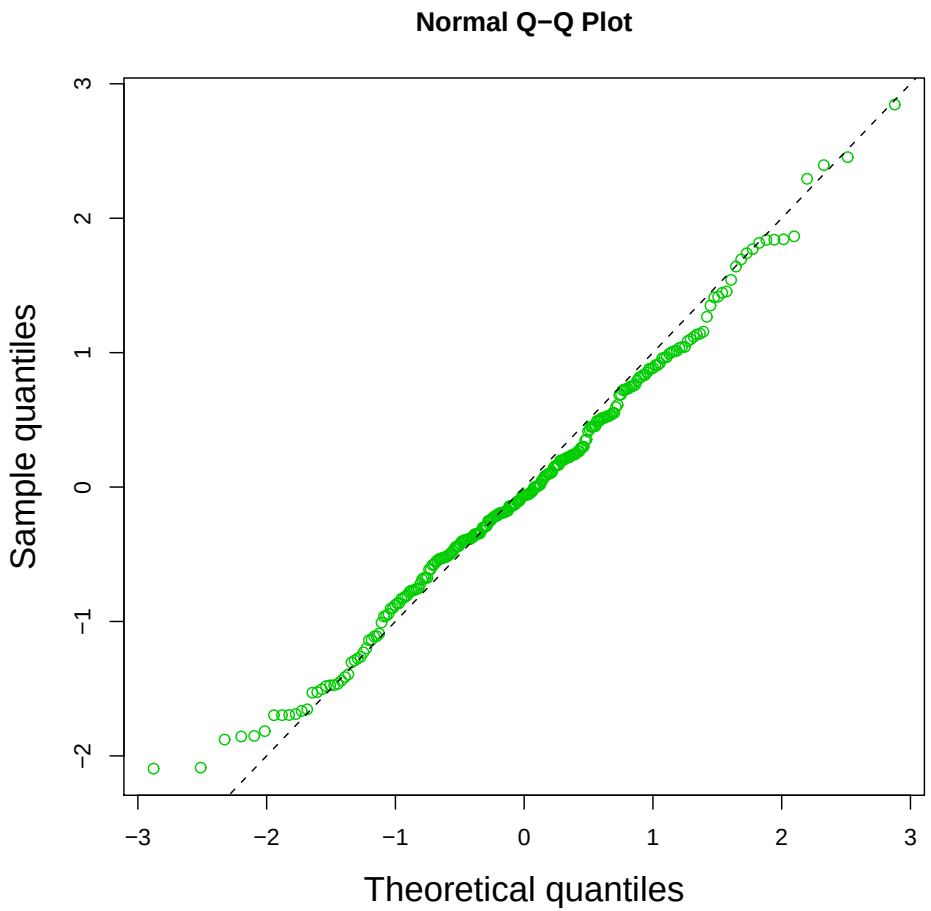
Sampling distribution of $R^*(\psi)$ is $N(0,1)$, to error of order $O(n^{-3/2})$, conditional on ancillary, hence unconditionally. Normal approximation to distribution of $R^*(\psi)$ yields third-order (relative) conditional accuracy in ancillary statistic setting, and confidence sets with third-order repeated sampling coverage accuracy.

Inference which respects that of exact conditional inference in exponential family setting to same third-order.

RE: $n = 5, \psi = 2, \mu = 1.0$, QQ plot, $R(\psi)$



RE: $n = 5, \psi = 2, \mu = 1.0$, QQ plot, $R^*(\psi)$



Some comments on analytic methods

- Often very awkward analytic calculations.
- Successfully packaged (Brazzale et al.) for certain classes of model, e.g. nonlinear heteroscedastic regression models.
- Also, relatively unexplored is idea of using simulation to replace analytic calculations, specifically to calculate Bartlett correction.
- Versions of R^* for vector interest parameters possible, seen as less effective than in case $p = 1$, or than Bartlett correction. But, 'directional approach' (Fraser et al.) which reduces to a one-dimensional integration problem seems to be very effective, even in high-dimensional settings.

(Constrained) Bootstrap

Bootstrap Principle: estimate sampling distribution of interest by that under a fitted model.

Key: appropriate handling of nuisance parameter. Repeated sampling properties of bootstrap are [modulo Monte Carlo error from using finite simulation] entirely determined by nuisance parameter effects.

The key recommendation

Use as basis of bootstrap calculation $F(y; (\psi, \hat{\phi}_\psi))$, fitted model with nuisance parameter taken as constrained MLE, for given value of interest parameter.

Properties: repeated sampling perspective

- 'Essentially exact'.
- Estimate true sampling distribution of $W(\psi)$ to error of order $O(n^{-2})$. Confidence sets constructed from bootstrap distribution of $W(\psi)$ have coverage error of order $O(n^{-2})$.
- Estimate true sampling distribution of $R(\psi)$ to error of order $O(n^{-1})$.
- But, confidence sets constructed from bootstrap distribution of $R(\psi)$ have third-order coverage accuracy: coverage error of order $O(n^{-3/2})$.

Detail

The confidence set is

$$\{\psi : R(\psi) \leq \tilde{G}^{-1}(1-\alpha)\},$$

where $\tilde{G}$ denotes the sampling distribution of $R(\psi)$ under $F(y;\tilde{\theta})$, the distribution with parameter value fixed as $\tilde{\theta} = (\psi, \hat{\phi}_\psi)$.

Corresponds to a significance function $u(Y,\psi) = \tilde{G}(R(\psi))$.

Note: a different bootstrap calculation required for each ψ. The significance function may not be monotonic.

Other schemes, e.g. substituting global MLE of nuisance parameter, less effective, in general. If $\hat{G}$ denotes the distribution of $R(\psi)$ under sampling from $F(y; \hat{\theta})$, the confidence set

$$\{\psi : R(\psi) \leq \hat{G}^{-1}(1-\alpha)\},$$

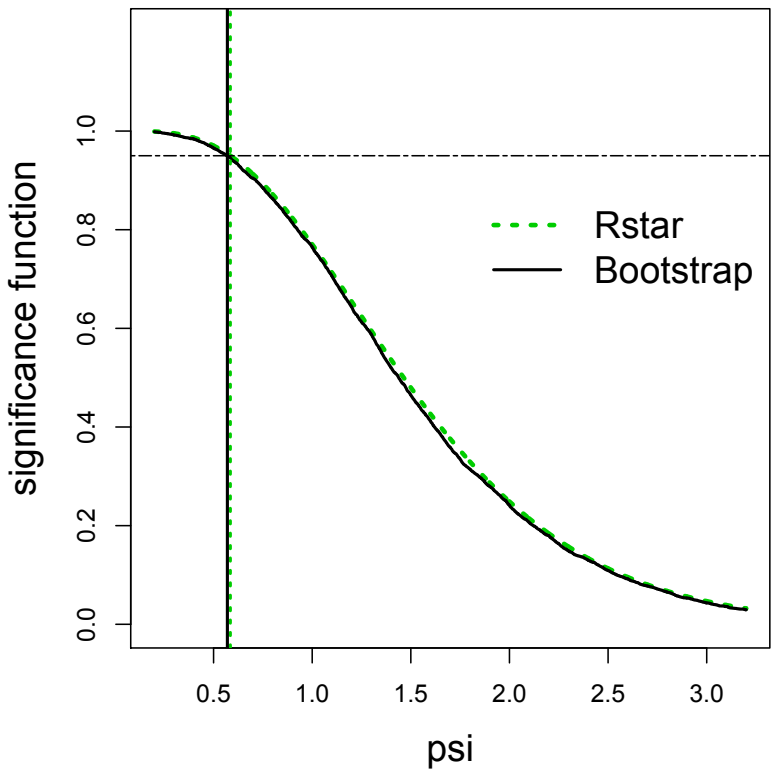
has coverage error of order $O(n^{-1})$.

PPI

Inference based on bootstrapping distribution of $R(\psi)$ respects PPI.

So does making normal approximation to sampling distribution of $R^*(\psi)$.

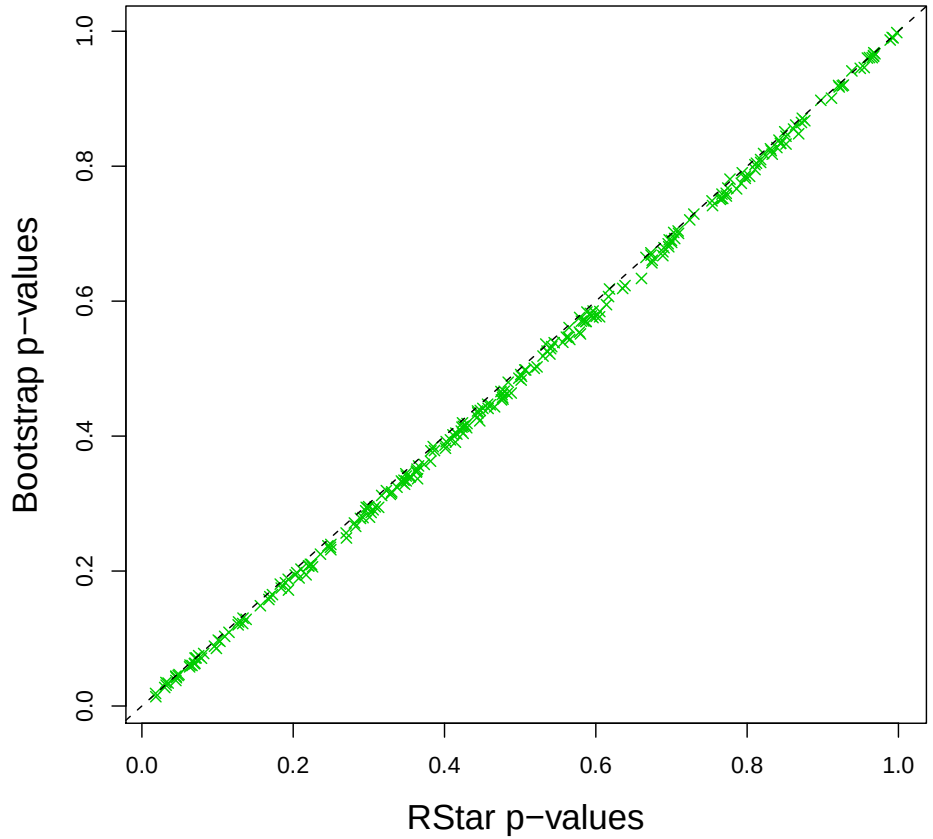
RE, data sample: significance functions



RE, data example: 95% confidence limits

- $R^*(\psi)$: interval is $(0.585, \infty)$.
- Bootstrap $R(\psi)$: interval is $(0.570, \infty)$.

RE: $n = 5$, bootstrap $p-$values vs R^* $p-$values



A practical example: signal detection

LHC: detection of signal in presence of background noise.

Set confidence limits on underlying signal, based on data from observation channel.

Observation is number of times a particular event is observed. Supposed to have Poisson distribution with mean $\psi\gamma + \beta$, where interest parameter ψ represents signal, β and γ represent respectively a background rate at which event occurs and efficiency of the measurement device.

Precise formulation

Available data is y_1, y_2, y_3. Realizations of independent Poisson random variables with means $\psi\gamma + \beta$, βt and γu respectively, where t and u are known and parameters ψ, β, γ are unknown.

In principle, $\psi \geq 0$, and nuisance parameters β, γ are positive.

Consider $y_1 = 1, y_2 = 8, y_3 = 14$, with $t = 27, u = 80$.

Inference

Appropriate inference is test of hypothesis $\psi = 0$ against one-sided alternative $\psi > 0$.

Significance probability is one minus significance function at $\psi = 0$.

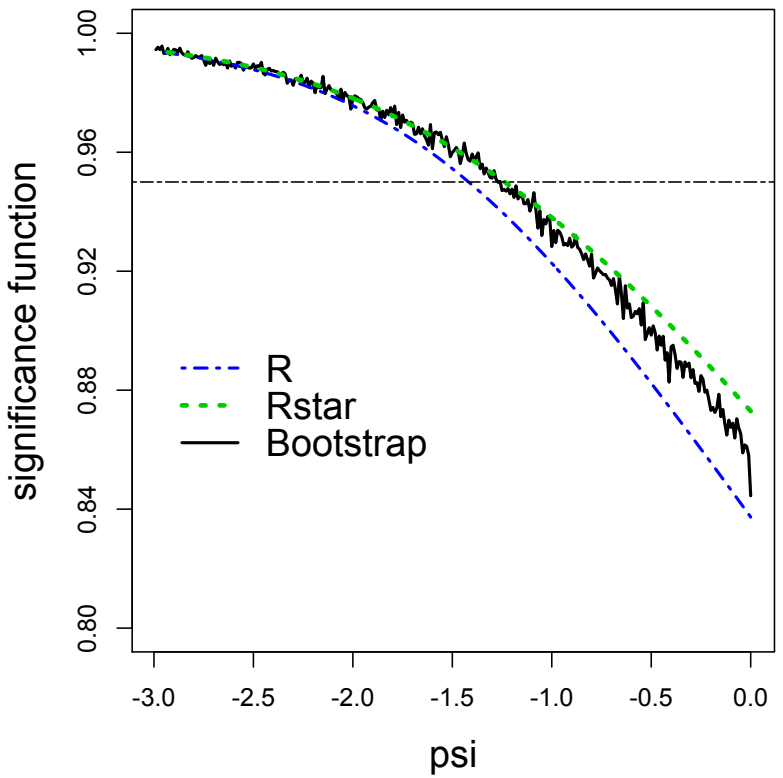
$R(\psi)$: $p-$value is $1 - \Phi\{R(0)\} = 0.163$.

$R^*(\psi)$: $p-$value is $1 - \Phi\{R^*(0)\} = 0.127$.

Bootstrap $R(\psi)$: $p-$value is 0.156 [10,000 bootstrap samples].

Weak evidence of positive signal.

Significance functions



Remarks

- Lower confidence limits are negative. If insist on confidence limits, take lower limit as maximum, $\max\{0, \psi_\alpha\}$, of actual limit ψ_α and lower physically admissible value of zero? All lower confidence limits are zero (coherent with $p-$values for testing for a positive signal). Calculation of $p-$value more appropriate?
- Even though large simulation is carried out, bootstrap significance function here is not smooth. Smoothing required? Discreteness of bootstrap distribution inducing differences with $R^*(\psi)$?
- Discrete distribution. Does not effect essential inferential issues, but introduces (mainly computational) complications. Not all theoretical results about rates of error etc. necessarily apply to such cases. Good practical performance.

Conditional properties of bootstrap, $p = 1$

Recall, bootstrap applied unconditionally.

- Multi-parameter exponential family context: inference agreeing with exact conditional inference to relative error third-order, $O(n^{-3/2})$. Same conditional accuracy as R^*. DiCiccio & Young (2008).
- Same context, automatically reproduces appropriate objective ('conditional second-order probability matching') Bayesian inference to order $O(n^{-3/2})$, in many circumstances.

- Ancillary statistic models: bootstrap inference using $R(\psi)$ agrees with conditional inference to second-order, $O(n^{-1})$;

- Same for other asymptotically $N(0,1)$ pivots, provided these are constructed using observed information. Pivot must be 'stable' to second-order, $O(n^{-1})$: marginal and conditional distributions must agree to that order. Not true, for example, for $T_W(\psi)$ and $T_S(\psi)$.

- Compare with third-order conditional accuracy of R^*.

- Third-order conditional accuracy unwarranted? Ancillary statistics typically not unique, different conditional inferences will typically only agree to second-order.

Vector interest parameter ($p > 1$)

Repeated sampling perspective: simulating the distribution of $W(\psi)$, at either global MLE or constrained MLE, produces $p-$values uniformly distributed under H_0, to error of order $O(n^{-2})$.

Ancillary statistic models: bootstrapping $W(\psi)$ approximates exact conditional inference given $A = a$ to third-order, $O(n^{-3/2})$.

Objective Bayes ($p = 1$)

- Exponential family context: conditional (and hence unconditional, repeated sampling) frequentist inference accurate to $O(n^{-3/2})$ achievable by any prior in a general class, provided a simple model condition holds. DiCiccio & Young (2010).

- Ancillary statistics models: unconditional higher-order probability matching priors give conditional frequentist accuracy to $O(n^{-3/2})$ under some further conditions (DiCiccio, Kuffner & Young, 2012). But now, in key cases exact conditional matching priors exist and are unique. In these cases, objective Bayes might be preferred route to conditional frequentist accuracy?

Methodological Issues

- 'Uniqueness of inference'.
- Computational considerations.
- Relationship between analytic and bootstrap approaches.

When do inferences agree?

In general, $p-$values from different asymptotically $N(0,1)$ pivots will agree only to first-order, $O(n^{-1/2})$.

However, establish simple sufficient conditions, under which $p-$values from two statistics will agree to second-order, $O(n^{-1})$, provided approximations to distributions accurate to $O(n^{-1})$ are employed. Such accurate approximation obtained quite generally by bootstrap.

Consequences

- $T_W(\psi)$ and $T_S(\psi)$ in general do not provide $p-$values that agree with those from $R(\psi)$ to order $O_p(n^{-1})$.
- But, versions of Wald and score statistics constructed using observed information will yield $p-$values agreeing with those from $R(\psi)$ to $O_p(n^{-1})$.
- Etc., etc.

Computational considerations

Use of $W(\psi)$ and $R(\psi)$ requires calculation of both global and constrained MLEs. Potentially unattractive compared to Wald statistic, $T_W(\psi)$ [or multivariate version]. Latter routinely employed in statistical packages etc., but not stable or parameterization invariant.

Bootstrap: must recalculate for a series of B bootstrap samples. General guideline: B of order of few 1000's to reduce Monte Carlo variability to acceptable levels, to 'capture' good theoretical properties. In small samples or with high-dimensional nuisance parameter solution of likelihood equations can be a worry.

$R^*(\psi)$: computationally simple, potentially awkward analytic calculations/coding. (Highly) stable, parameterization invariant.

Relationship between Bootstrap and $R^*(\psi)$

Conceptually related, not distinct methodologies.

Details

Specifically:

- $p-$values calculated from $N(0,1)$ approximation to distribution of $R^*(\psi)$ will quite generally agree with those from bootstrap to order $O_p(n^{-1})$.
- Multi-parameter exponential family models: (unconditional) bootstrap $p-$values agree with those from $R^*(\psi)$ to $O_p(n^{-3/2})$.
- Ancillary statistic models: normal approximation to $R^*(\psi)$ is an $O(n^{-3/2})$ (saddlepoint) approximation to conditional bootstrap [which could use if we could simulate the conditional distribution of $R(\psi)$ given $A = a$].

The bottom line

If likelihood equations can be reliably solved, analytic simplicity indicates bootstrapping of $R(\psi)$ or $W(\psi)$ as a highly effective methodology.

- Competitive in terms of accuracy with analytic alternatives.
- Unlikely to be computationally prohibitive [moderate B adequate to ensure MC variability does not impair good theoretical properties].
- Stable (respects CP to high-order) and parameterization invariant: 'inferentially correctness is OK'.
- Vector ψ: use bootstrap calculation to estimate mean of $W(\psi)$, then base inference on χ^2 approximation to empirically Bartlett-corrected statistic $\bar{W}_c(\psi)$, or use directional tests of Fraser et al.

Part 3: Statistics in Data Science

Formal framework for inference in Data Science

We assume our data Y lies in some measurable space with unknown sampling distribution $Y \sim F$.

The task is to pose, on the basis of Y itself, a reasonable probability model $\hat{M}$, then carry out inference, using the same data Y.

Let $S \equiv S(Y)$ be the selection event. For instance, this might be the event that model $\hat{M}$ is chosen, or, in the context of the File Drawer Effect example, Y distributed as $N(\mu, 1)$, the event $S = \{|Y| > 1\}$.

Central proposal is that to be relevant to the observed data sample and yield precisely interpretable validity, the inference we perform should not be drawn from the original assumed distribution, $Y \sim F$, but by considering the conditional distribution of $Y|S$.

Discussion

This is just the Fisherian proposition, that hypothetical samples used as basis of inference should be conditioned in appropriate way, here on selection event $S(Y)$.

Now the selection event S will typically be informative about the quantity θ of interest, and conditioning discards information. But, to ignore the selection event loses control over the (Type 1) error rate, potentially badly.

Principled inference requires conditioning on the selection event, and therefore drawing inferences from leftover information in Y, given S.

Example: File Drawer Effect, ctd.

Suppose Y is distributed as $N(\mu, 1)$. Take selection event as $\{Y > 1\}$.

Selective inference uses the distribution of Y conditional on $\{Y > 1\}$: this has density

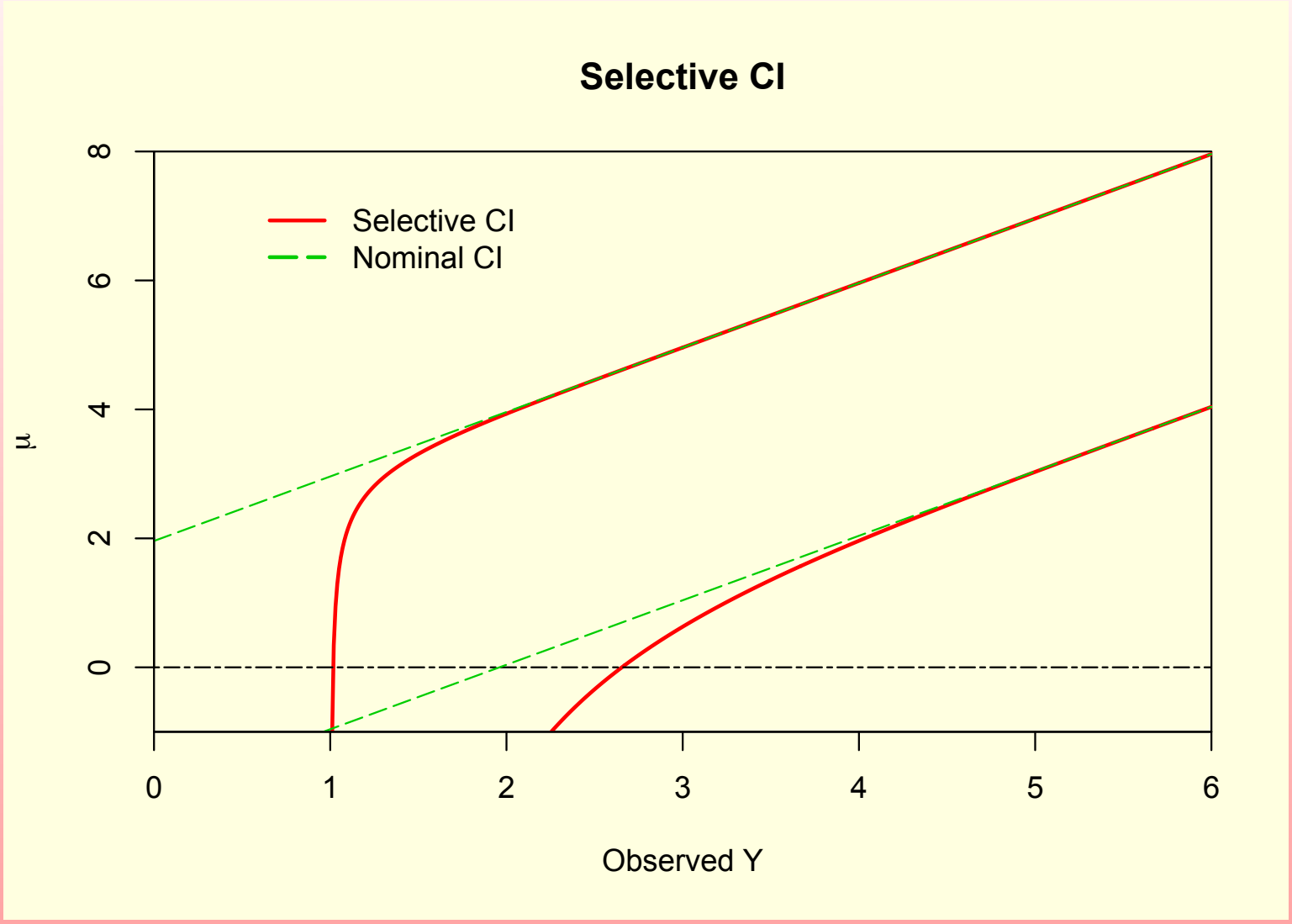
$$f_S(y; \mu) = \frac{\phi((Y - \mu))}{\Phi(\mu - 1)}.$$

Let $F(y; \mu)$ be corresponding distribution function. For given observed y_o we construct the selective CI of coverage $1 - \alpha$ as $\{\mu : \alpha/2 \leq F(y_o; \mu) \leq 1 - \alpha/2\}$.

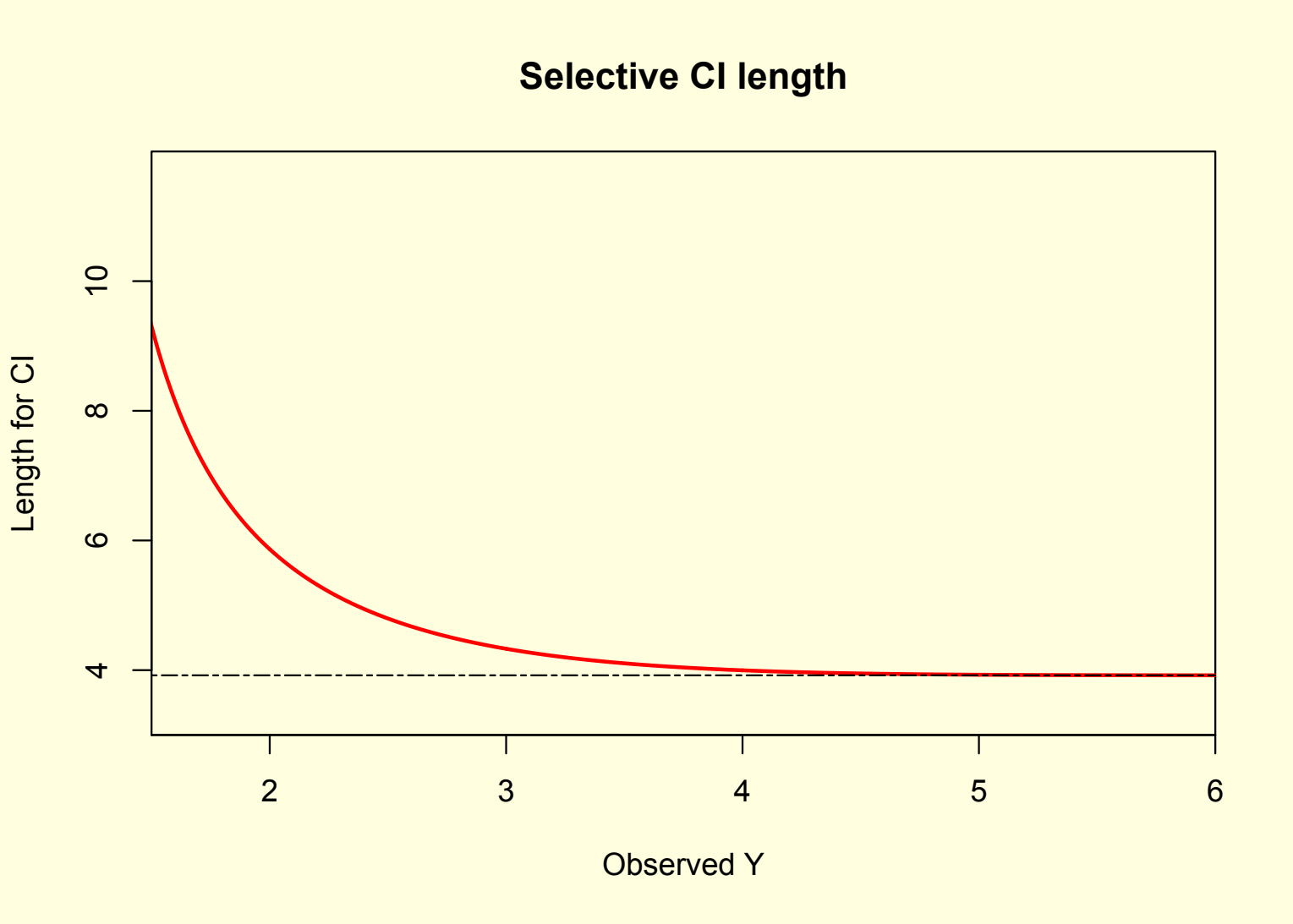
Compare 'nominal' confidence intervals [not accounting for selection] and selective confidence intervals, of coverage 90%.

If Y is much larger than 1, there is hardly any selection bias, no adjustment for selection is really required. When Y is close to 1, need to properly account for selection is stark, and length of the selective CI $\rightarrow \infty$.

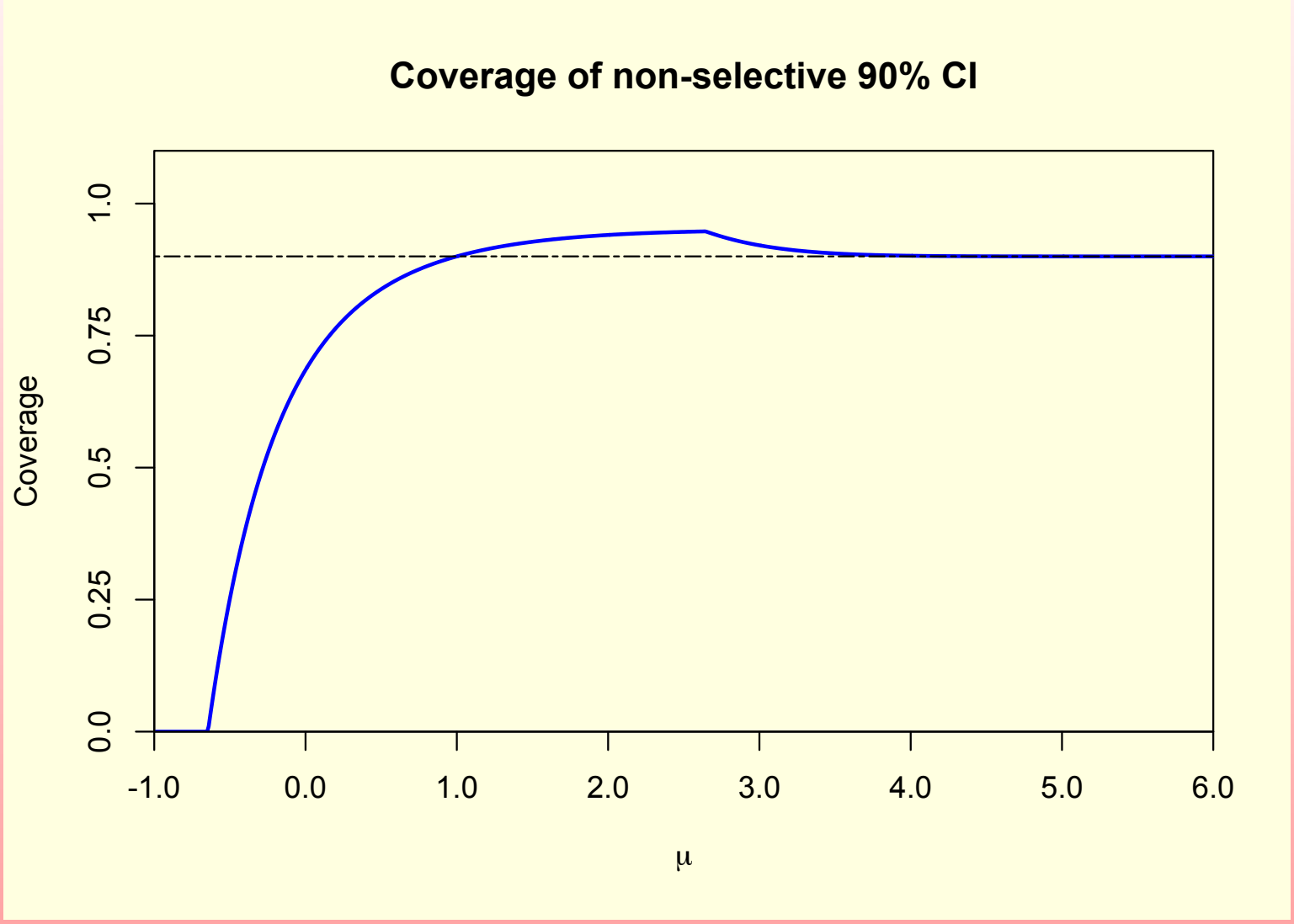
File Drawer Effect: CIs



File Drawer Effect: length of selective CI



File Drawer Effect: coverage of non-selective CI



Borrowing from classical theory

Conditioning the inference performed on the selection event is especially convenient if Y is assumed to have an exponential family distribution.

Then the distribution of Y conditional on a measurable selection event $S(Y)$ is also an exponential family distribution, allowing support for the techniques of selective inference to be drawn from the established classical theory for inference in exponential families.

Key example: variable selection, Normal linear regression

Suppose that $Y \sim N_n(\mu, \sigma^2 I_n)$, with $\mu \equiv X\beta$, β a vector of unknown parameters, and X a matrix of p predictors with columns $X_1, \ldots, X_p \in \mathbb{R}^n$.

Suppose, for simplicity, that σ^2 is known.

Some variable selection procedure (Lasso, LAR, ...) is utilised to select a model $M \subset \{1, \ldots, p\}$ consisting of a subset of the p predictors. Under the selected model, $\mu = X_M \beta^M$, where X_M is $n \times |M|$, say, with columns $(X_M)_1, \ldots, (X_M)_{|M|}$: we assume that X_M is of full rank, so that $\beta^M = (\beta_1^M, \ldots, \beta_{|M|}^M)$ is well-defined.

'Selected model'

Conventional principles of inference in exponential family distributions, adapted to this selective inference context, indicate that inference on β_j^M should be based on the conditional distribution of $(X_M)_j^T Y$, given the observed values of $(X_M)_k^T Y$, $k = 1, \ldots, |M|, k \neq j$, and the selection event that model M is chosen.

'Saturated model'

If we do not take the model M seriously, there is still a well defined linear predictor in the population for design matrix X_M.

Now define target of inference as

$$\beta^M \equiv \arg\min_{b^M} \mathbb{E}\| Y - X_M b^M \|^2 = X_M^+ \mu,$$

$X_M^+ \equiv (X_M^T X_M)^{-1} X_M^T$ is the Moore-Penrose pseudo-inverse of X_M.

This 'saturated model' perspective is convenient as it allows meaningful inference even if, say, our variable selection procedure does a poor job.

Assertion

This point of view can be advocated (see, for example, Berk et al., 2013) as a way of avoiding the need, in the adaptive model determination context typical of Data Science, to consider multiple candidate probabilistic models.

Inference

Under the selected model, β_j^M can be expressed in the form $\beta_j^M = \eta^T \mu$, say, whereas under the saturated model there may not exist any β^M such that $\mu = X_M \beta^M$.

Compared to the selected model, the saturated model has $n - |M|$ additional nuisance parameters, which may be completely eliminated by the classical device of conditioning on the appropriate sufficient statistics: these correspond to $P_M^\perp Y \equiv (I_n - X_M(X_M^T X_M)^{-1} X_M^T) Y$.

Considering the saturated model as an exponential family, again assuming σ^2 is known, and writing the least-squares coefficient β_j^M again in the form $\eta^T \mu$, inference is based on the conditional distribution of $\eta^T Y$, conditional on the observed values of $P_\eta^\perp Y \equiv (I_n - \eta^T(\eta^T \eta)^{-1} \eta^T) Y$, as well as the selection event.

Selected or saturated?

Do we treat $P_M^\perp \mu$ as an unknown nuisance parameter, to be eliminated by further conditioning, or assume $P_M^\perp \mu = 0$?

Denoting by $X_{M\setminus j}$ the matrix obtained from X_M by deleting $(X_M)_j$, and letting $U = X_{M\setminus j}^T Y$ and $V = P_M^\perp Y$, the issue is whether to condition on both U and V, or only on U.

In the classical, non-adaptive, setting this issue does not arise, as $\eta^T Y, U$ and V are mutually independent: they are generally not independent conditional on the selection event.

If we condition on V when, in fact, $P_M^\perp \mu = 0$, we might expect to lose power, while inferential procedures may badly lose their control of (Type 1) error rate if this quantity is large, so that the selected model is actually false.

Our viewpoint

Contend that such conditioning (on V) is necessary to ensure validity of the conclusions drawn from the specific data set under analysis.

Example: bivariate regression

Suppose that Y is distributed as $N_2(\mu, I_2)$, so that $\sigma^2 = 1$ and that the design matrix is $X = I_2$.

We choose a 'one-sparse model', that is X_M is specified to have just one column. The selection procedure chooses $M = \{1\}$ if $|Y_1| > |Y_2|$ and $M = \{2\}$ otherwise.

Suppose data outcome $Y = \{2.9, 2.5\}$, so the chosen model is $M = \{1\}$.

Inference, selected model

The selected model M has Y distributed as $N_2((\mu_1, 0), I_2)$.
Inference on μ_1 would base a test of $H_0 : \mu_1 = 0$ against
$H_1 : \mu_1 > 0$ on rejection for large values of Y_1, $Y_1 > c$, say.

In the test of nominal Type 1 error α based on the selected model, c is fixed by requiring $P_{H_0}(Y_1 > c \,|M, |Y_1| > |Y_2|) = \alpha$, explicitly assuming that $\mu_2 = 0$.

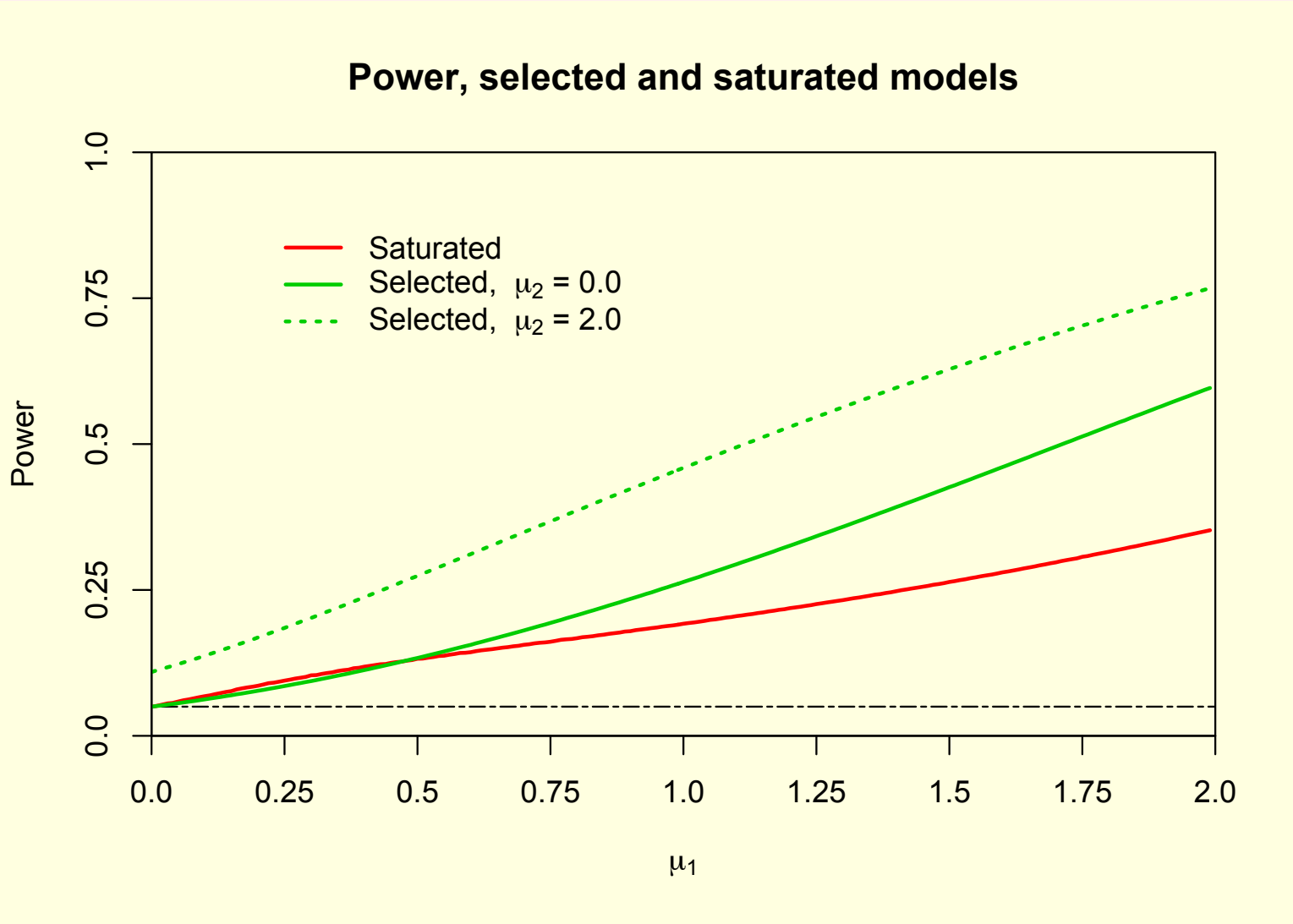
Inference, saturated model

In the saturated model framework, we reject H_0 if $Y_1 > c'$, where c' satisfies

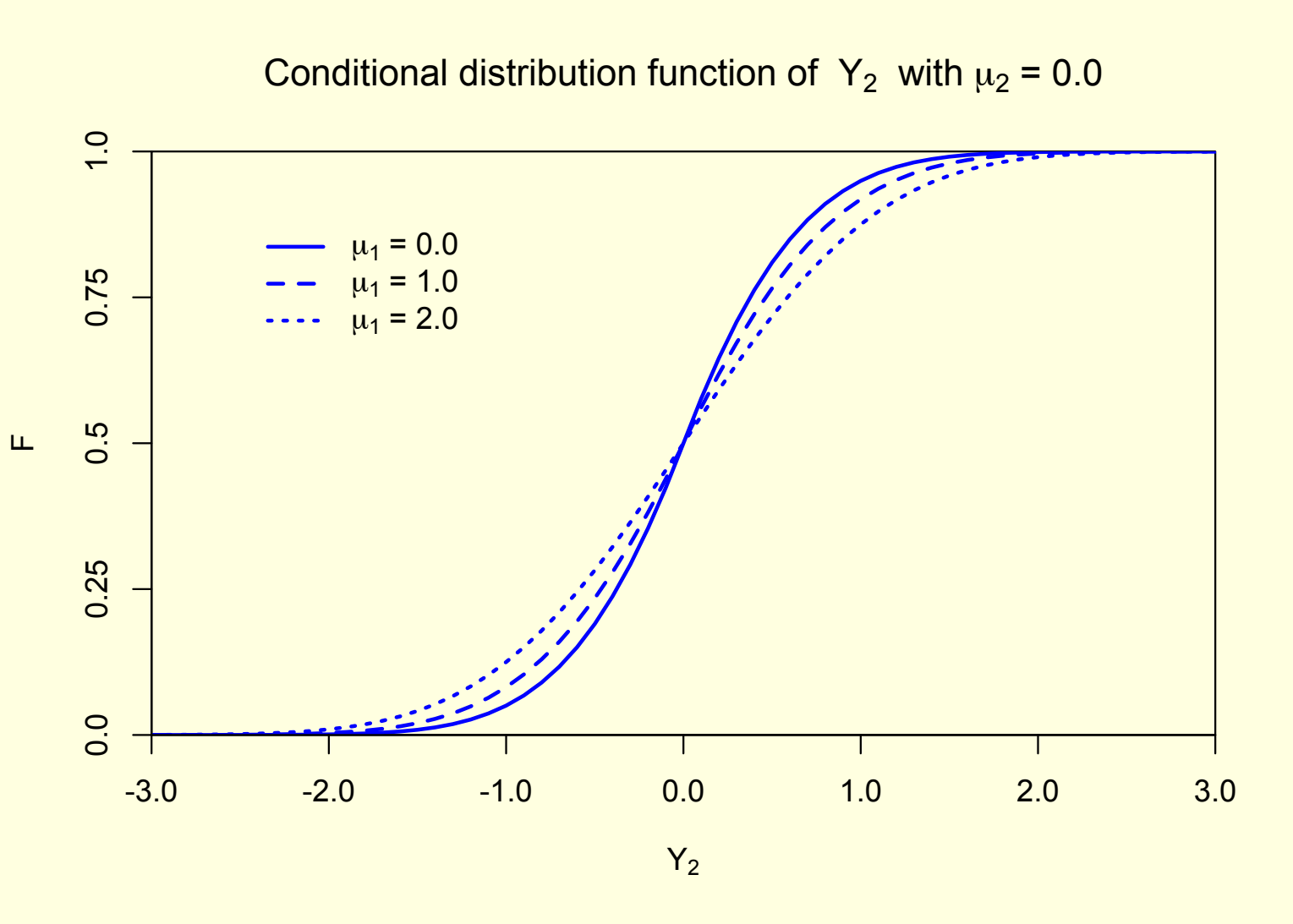
$$P_{H_0}(Y_1 > c' \,|Y_2 = 2.5, |Y_1| > |Y_2|) \equiv P_{H_0}(Y_1 > c' \,| \,|Y_1| > 2.5) = \alpha.$$

Conditioning on the observed value $Y_2 = 2.5$ as well as the selection event eliminates completely dependence of the Type 1 error rate on the value of μ_2. It is immediately established here that $c = 1.95, c' = 3.23$, in tests of nominal Type 1 error rate 0.05.

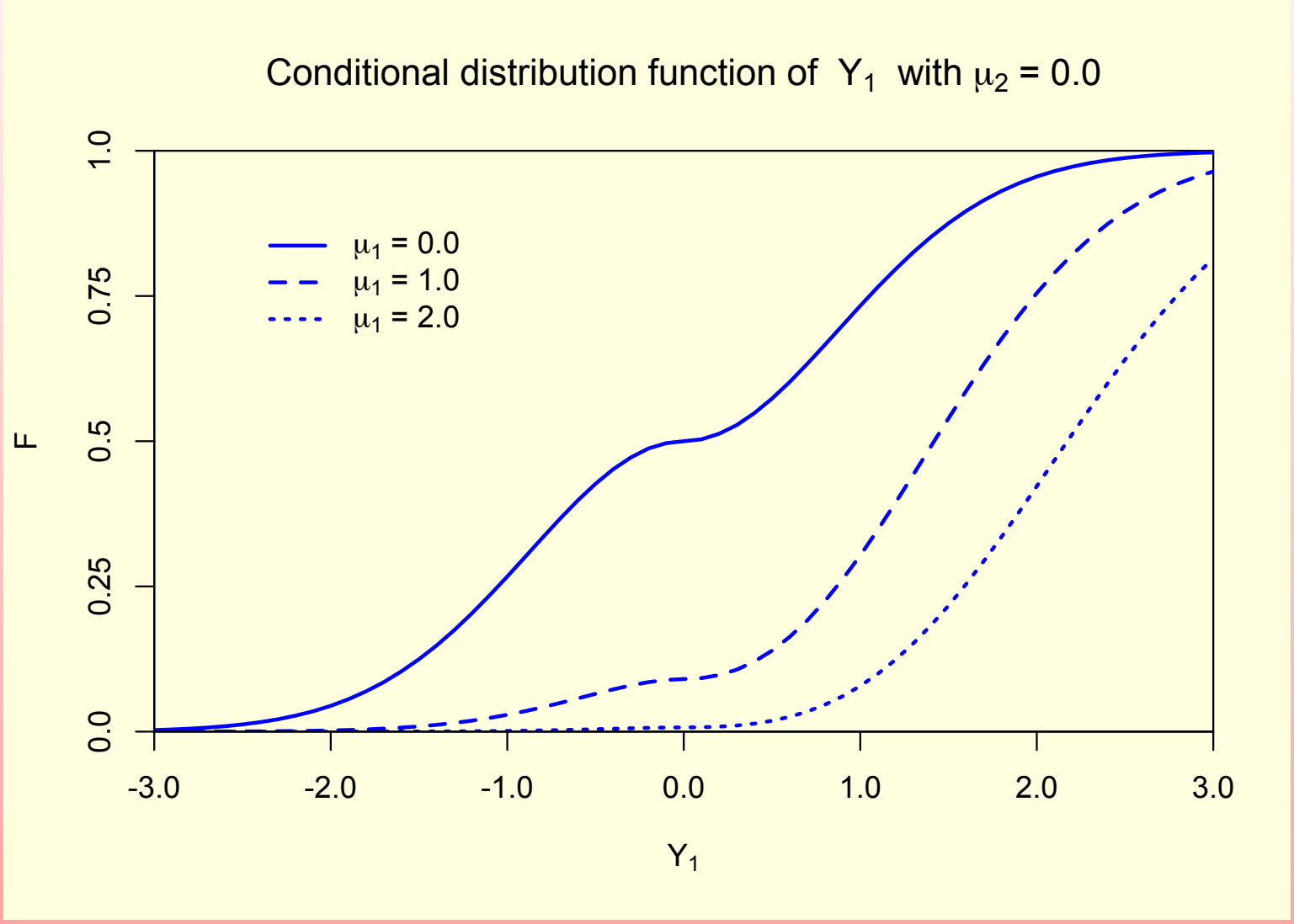
Power functions



Distribution of Y_2 conditional on selection event



Distribution of Y_1 conditional on selection event



What do we conclude?

Operational difference between the saturated and selected model perspectives may be important in key practical contexts, such as early steps of sequential model-selection procedures.

But, the case being made is that a principled approach to inference is forced to give central consideration to the saturated model in contexts such as those discussed here, where valid interpretation of significance is key.

Some other points, (A)

(i) Distribution theory necessary for inference in the saturated model perspective, under the Gaussian assumption, is generally easy.

In some generality, the selection event can be expressed as a polyhedron $S(Y) = \{AY \leq b\}$, for A, b not depending on Y. This is true for forward stepwise regression, the lasso with fixed penalty parameter λ, LAR and other procedures.

If inference is required for $\eta^T \mu$, then further conditioning on $P_\eta^\perp Y$ yields the conditional distribution required for the inference to be a truncated Gaussian with explicitly available endpoints, allowing a simple analytic solution.

Conditioning on $P_\eta^\perp Y$ promoted as a means of obtaining an analytically simple distribution for the inference.

Conditioning is necessary to eliminate dependence on the nuisance parameter and provide control over Type 1 error.

Marginally, $\eta^T Y$ is independent of $P_\eta^\perp Y$, so the conditioning is justified by ancillarity, but this is not true conditional on the selection event: justification stronger than analytic convenience is provided by necessary elimination of the nuisance parameter.

Some other points, (B)

In the non-Gaussian setting and in general under the selective model, Monte Carlo procedures, such as MCMC and acceptance/rejection methods, will be necessary to determine the necessary conditional distribution of Y, but this is unlikely to prove an obstacle to principled inference.

Another perspective on selective inference

Tibshirani et al., 2018, offer a different proposal, potentially relevant to data science.

Consider the multivariate normal model.

Under alternative framework, we recognise that for every possible selected model M, a quantity of interest, $\eta_M^T \mu$, say, is specified. When model $\hat{M}(Y)$ is selected inference is made on the interest parameter $\eta_{\hat{M}(Y)}^T \mu$.

The notion of validity now is that under repeated sampling of Y, a specified proportion $1 - \alpha$ of the time the inference on the selected target, which is not fixed, is correct.

Comment

Implicitly what is sought in much of data science?

Abandons the requirement that we have argued is central to principled inference, of ensuring validity and relevance to the actual data sample.

Concluding remarks, frequentist selective inference

- Principled approach to frequentist inference in data science necessary to provide the rationale by which claimed frequentist error-rate properties are justified.

- Appropriate conceptual framework for valid inference is that discussed in statistical literature as 'Post selection inference', based on ensuring relevance of sampling distributions to particular data sample. Fisherian ideas: no new paradigm involved.

- Inference after adaptive model determination ('data snooping') requires conditioning on selection event, control of error rate of inference given it was performed: 'the answer must be valid, given that the question was asked.'

- Care required, as selected model for inference may be wrong, can lead to substantially distorted error rates. Primary cause is assumption that nuisance parameters effects are known: elimination by classical device of (further) conditioning ensures precise control of error rates. 'Saturated model' framework is the appropriate basis for inference.

- Potential loss of accuracy (power) is undesirable, but may not be practically consequential: possible overconditioning is worthwhile price paid for validity.

Bayesian selective inference

Inference for a model parameter θ provided only if we observe event E: covers situation described before, but also situations in which sampling model (and therefore parameter) is specified after observing E.

Inference about fixed parameter θ is based on selection-adjusted posterior, with density

$$\pi^E(\theta|y) \propto \pi(\theta) f(y|\theta, E).$$

The truncated likelihood is $L^E(\theta; y) \propto f(y|\theta) P(E|\theta)^{-1}$.

An example: one dimensional Gaussian

Let $Y_n|\mu \sim N(\mu, n^{-1})$, where $\mu \in \mathbb{R}$ is parameter of interest. Inference is only provided for μ if we observe $E = \{Y_n > 0\}$.

Truncated likelihood is

$$L^E(\mu; Y_n) \propto \frac{\phi(\sqrt{n}(Y_n - \mu))}{P(N(\mu, n^{-1}) > 0)} = \frac{\phi(\sqrt{n}(Y_n - \mu))}{\Phi(\sqrt{n}\mu)}.$$

Some alerts

For this Gaussian model, well known that as $Y_n \to 0$, the maximizer of the truncated likelihood $L^E(\mu; Y_n)$ tends to $-\infty$.

Finite sample behaviour

Repeated sampling behaviour of, say, mode of selective posterior, as estimator of true mean μ_0 can be awful.

Consider bias and MSE of selective/non-selective posterior modes, under $N(0,1)$ prior for μ, when $Y_n|\mu \sim N(\mu, n^{-1})$, true mean is $\mu_0 = -0.1$, as before $E = \{Y_n > 0\}$. Each figure based on $R = 10,000$ replications of $Y_n|E$.

Bias, MSE of selective (S) and non-selective (NS) posterior modes

n	Bias		MSE	
	S	NS	S	NS
5	-0.0457	0.3678	0.1961	0.1802
10	-0.0904	0.2975	0.1721	0.1136
50	-0.1305	0.1819	0.1178	0.0377
100	-0.1333	0.1520	0.0999	0.0251
500	-0.1369	0.1156	0.0827	0.0136
1000	-0.1370	0.1085	0.0765	0.0118

Asymptotic considerations

Bayesian setting: asymptotic aspect of key concern is consistency.

Sequence of posterior distributions is consistent if it correctly identifies the true value of the parameter asymptotically.

Formalised by requiring that posterior probability of sets bounded away from true (fixed) parameter value decreases exponentially almost surely.

Bernstein-von Mises: for regular IID models, posterior distribution of $\sqrt{n}(\theta - \hat{\theta})$ converges to $N_p(0, i(\theta_0)^{-1})$, under basic conditions, with $\hat{\theta}$ the MLE and $i(\theta_0)$ the Fisher information at the true parameter value θ_0.

Selective regime

Consistency and asymptotic normality need not hold.

One dimensional Gaussian, ctd.

Suppose $Y_n \sim N(\mu_0, n^{-1})$ for all n, selection event $E = (0, \infty)$.

Asymptotic behaviour of truncated likelihood.

- If $\mu_0 > 0$, then $\sqrt{n}(Y_n - \mu_0)|E \xrightarrow{d} N(\mu_0, 1)$.
- If $\mu_0 = 0$, then $\sqrt{n}Y_n|E \xrightarrow{d} N(\mu_0, 1)|E$.
- If $\mu_0 < 0$, then $nY_n|E \xrightarrow{d} \text{Exp}(-\mu_0)$.

Posterior consequences

- If $\mu_0 > 0$, selection region contains true parameter value, posterior inference is asymptotically same as that in IID (non-selective) setting. Unsurprising.
- If μ_0 lies on boundary of E, provided prior gives positive probability to a neighbourhood of μ_0, posterior distributions are consistent (in probability).
- If $\mu_0 < 0$, outside the selection region, truncated likelihood has a non-degenerate limit, given by $\mu \exp\{\mu Z\}, \mu < 0$, where Z is $\text{Exponential}(-\mu_0)$.

Interesting case, $\mu_0 < 0$.

For any fixed set $B \subseteq \mathbb{R}$ with $B \cap (-\infty, 0) \neq \emptyset$, then $P(\mu \in B | Y_n) | \{Y_n \in E\}$ has a non-degenerate asymptotic distribution, where probability is taken with respect to the selective posterior distribution.

Illustration

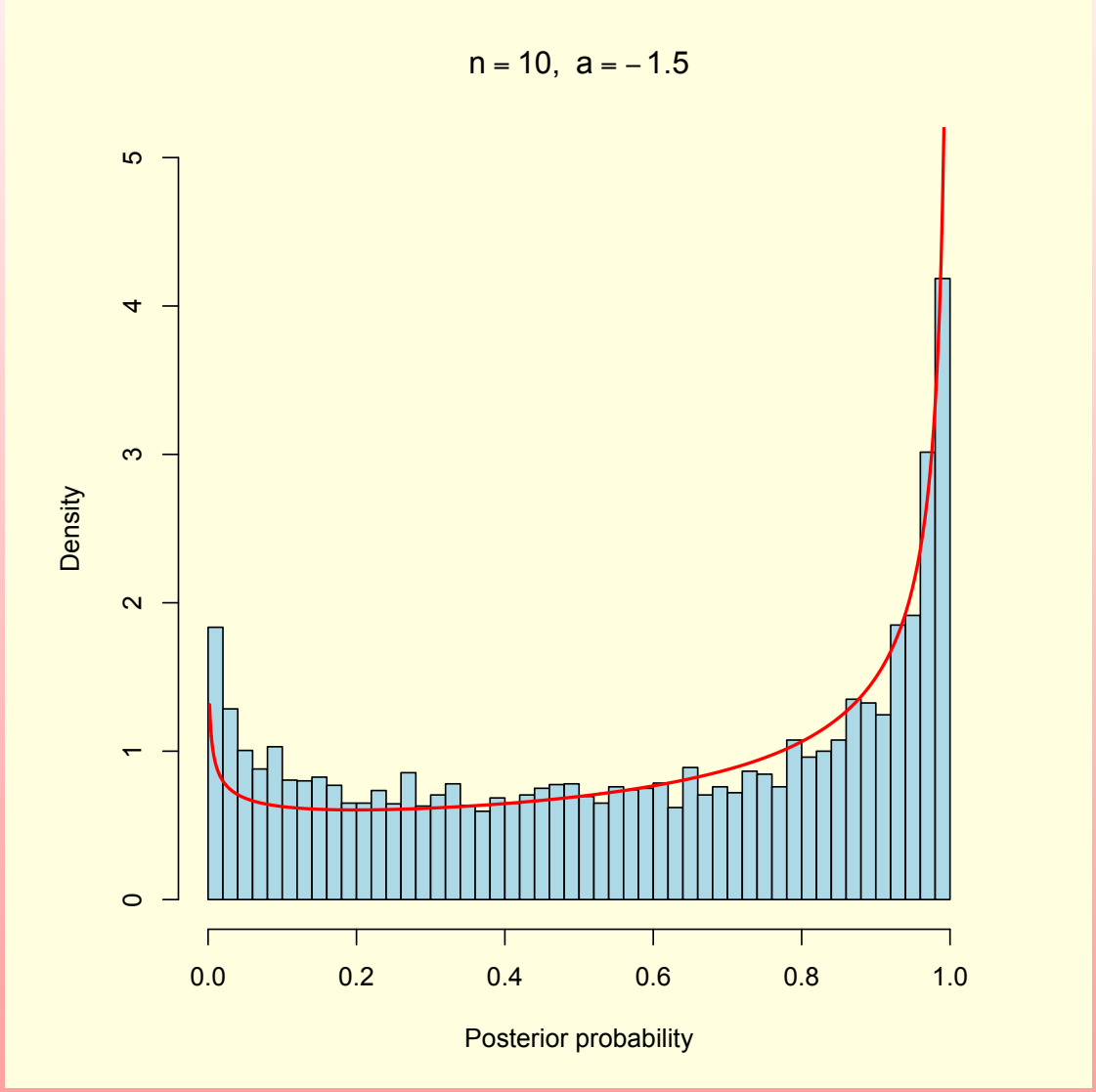
With uniform prior $\pi(\mu) \propto 1$, $B = (-\infty, a), a < 0$, we have

$$P(\mu \in B | Y_n) | \{Y_n \in E\} \xrightarrow{d} (-aZ + 1) \exp(aZ).$$

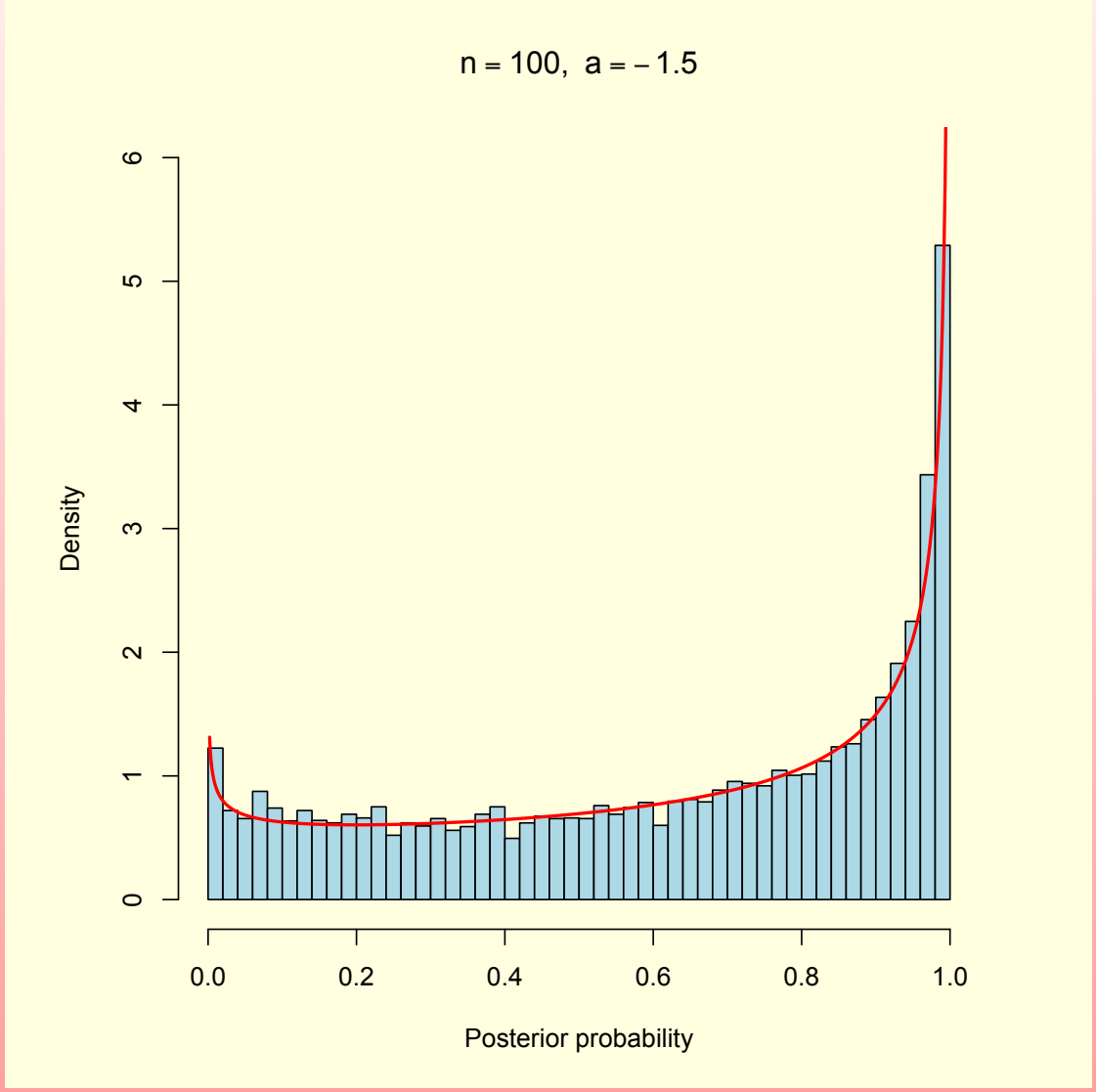
Empirical comparisons, $\mu_0 = -1.0$

Consider case $a = -1.5$. Generate a series of $R = 10,000$ replications of $Y_n|E$, each calculate posterior probability $P(\mu \in B|Y_n)$. Histogram compared with asymptotic limiting distribution.

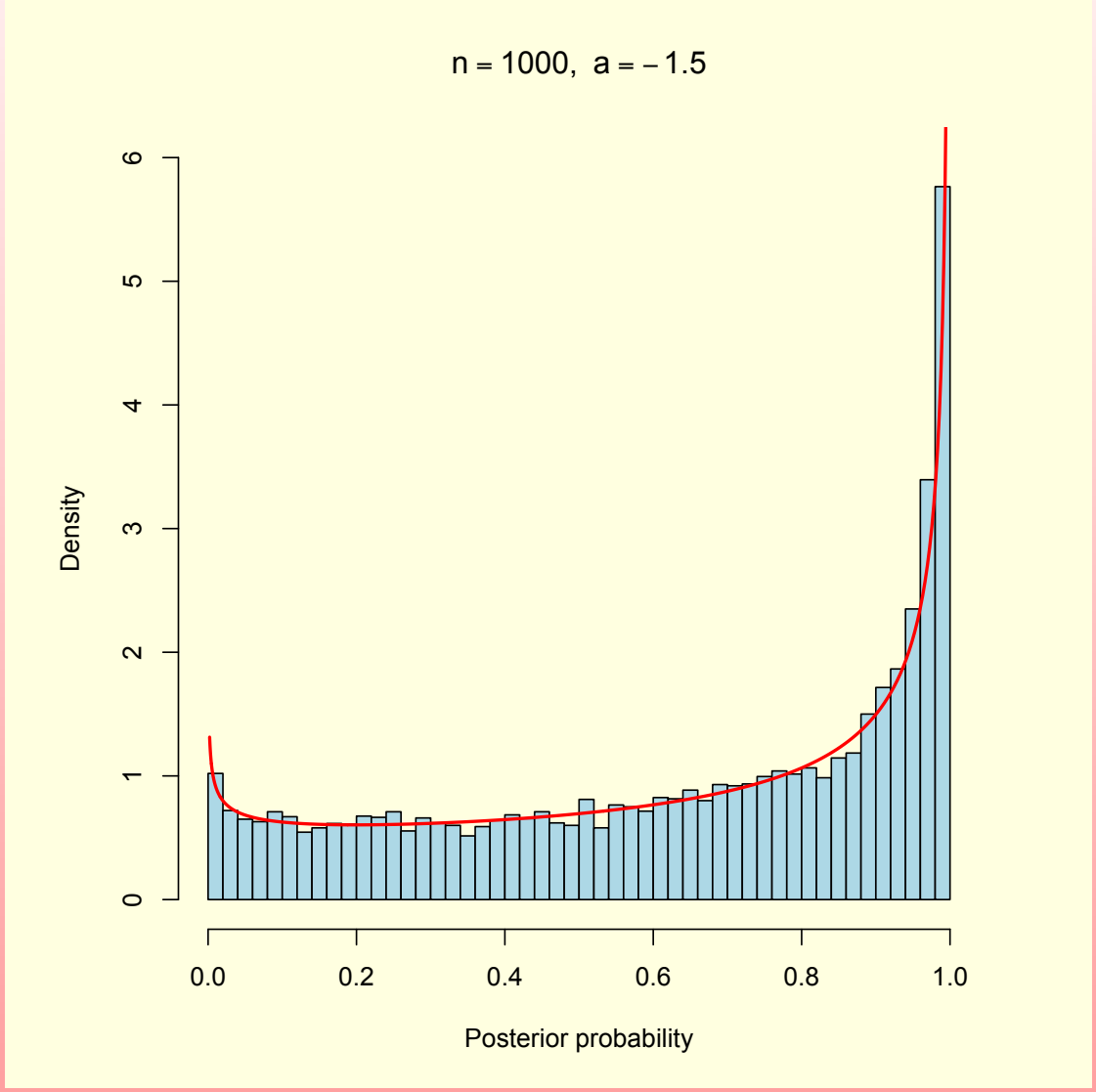
Distribution, posterior probability of $B = (-\infty, -1.5), n = 10$



Distribution, posterior probability of $B = (-\infty, -1.5), n = 100$



Distribution, posterior probability of $B = (-\infty, -1.5), n = 1000$



Multidimensional extension

Challenge: arbitrariness of selection region.

Asymptotic behaviour of truncated likelihood tractable when sampling model is multivariate normal, fixed dimension and known covariance matrix which decreases in norm, and selection region is affine.

Covers, for example, variable selection with Lasso, LAR, marginal screening.

Formulation

Suppose $Y_n|\mu \sim N_p(\mu, n^{-1}\Sigma)$, Σ known, selection region $E = \{y : Ay \leq b\}$, A symmetric, positive definite.

Key results

Let $\hat{y} = \text{argmin}\{||\Sigma^{-1/2}(y - \mu_0)||_2 : y \in E\}$.

Under repeated sampling with true parameter μ_0, there exists a linear reparameterization of μ to (s, t), such that:

- $\dim(s) = ||A\hat{y} - b||_0$.
- π^E is consistent and asymptotically normal for s as $n \to \infty$.
- $\pi^E(t|y)$ has a non-degenerate limit as $n \to \infty$.

Discussion

Essentially, we conclude that non-smoothness of (polytopic) selection regions can have a negative effect on asymptotic consistency of Bayesian selective posterior.

If the closest point to the true parameter within selection region lies in the intersection of m of the hyperplanes defining polytope, posterior distributions are only consistent for a $(p - m)$-dimensional transformation of parameter.

Randomization (Tian and Taylor, 2018), involving applying selection procedure to noisy version of the data, avoiding hard-threshold truncation on sample space, to alleviate problem?

One dimensional Gaussian, ctd.

As before, $Y_n \sim N(\mu_0, n^{-1})$, let $Z_n \sim N(0, n^{-1})$, independently.

Truncation applied to $Y_n + Z_n$: inference on μ only if $Y_n + Z_n \in E = (0, \infty)$.

Illustration

True $\mu_0 = -1$, uniform prior, generate a series of $R = 10,000$ replications, compare average over replications of posterior probability $P(\mu \in (\mu_0 - 0.1, \mu_0 + 0.1)|Y_n)$:

- (a) (nonrandomised) conditional on selection $\{Y_n \in E\}$;
- (b) (randomised) conditional on selection $\{Y_n + Z_n \in E\}$.

$P(\mu \in (\mu_0 - 0.1, \mu_0 + 0.1) | Y_n)$

n	$\lvert\{Y_n \in E\}$	$\lvert\{Y_n + Z_n \in E\}$
10	0.026	0.132
20	0.027	0.183
50	0.029	0.280
100	0.029	0.383
200	0.029	0.519
500	0.030	0.735
1000	0.030	0.887

Concluding remarks, Bayesian selective inference

- Bayesian approach is not a panacea for the selective inference problem: selection has to be taken into account;

- Computational challenges, selective posterior may even be inconsistent, though applying selection on randomized version of data appears to fix things;

- Asymptotic regime reasonable? In example, fixed $\mu_0 = -1$, fixed $E = (0, \infty)$. Have $P(E) \to 0$ rapidly as $n \to \infty$. True also with randomization, though

$$\frac{P(E \text{ with randomization})}{P(E \text{ without randomization})} \to \infty,$$

as $n \to \infty$. Randomization makes rare event E more likely.

So, in summary...

- Statistics, even in data science era, should be directed by well-established, classical principles of inference, especially ideas of appropriate conditioning, at least when a parametric model can reasonably be postulated;
- By this means achieve valid, relevant inference. High levels of accuracy, by computationally intensive or analytic methods;
- Problems posed by selective inference can be satisfactorily addressed by conditioning arguments, at least under assumption that selection event is well-defined.

Work to be done...

- Adapt principles to ad hoc or informal selection procedures;
- Universality perhaps guided [Berk et al.] by consideration of procedures which search for the statistically most significant effect.

Key references

Barndorff-Nielsen, O. E. & Cox, D. R. *Inference and Asymptotics*. Chapman & Hall, 1994.

Berk, R., Brown, L., Buja, A., Zhang, K. & Zhao, L. Valid post-selection inference. Ann. Statist., **41**, 802–837, 2013.

Birnbaum, A. On the foundations of statistical inference (with discussion), J. Amer. Statist. Assoc., **57**, 269–326, 1962.

Brazzale, A.R., Davison, A.C. & Reid, N. *Applied Asymptotics: Case Studies in Small-Sample Statistics*. CUP, 2007.

Cox, D.R. *Principles of Statistical Inference*. CUP, 2006.

Cox, D.R. & Mayo, D.G. Objectivity and conditionality in frequentist inference. In *Error and Inference*, CUP, 276–304, 2010.

Davison, A.C., Fraser, D.A.S., Reid, N. & Sartori, N. Accurate directional inference for vector parameters in linear exponential families. J. Amer. Statist. Assoc., **109**, 302-314, 2014.

DiCiccio, T.J., Kuffner, T.A. & Young, G.A. Objective Bayes, conditional inference and the signed root likelihood ratio statistic. Biometrika, **99**, 675–686, 2012.

DiCiccio, T.J. & Young, G.A. Conditional properties of unconditional parametric bootstrap procedures for inference in exponential families. Biometrika, **95**, 747–758, 2008.

DiCiccio, T.J. & Young, G.A. Objective Bayes and conditional inference in exponential families. Biometrika, **97**, 497–504, 2010.

Efron, B. & Hastie, T. *Computer Age Statistical Inference: Algorithms, Evidence, and Data Science*. CUP, 2016.

Fithian, W., Sun, D. & Taylor, J. Optimal inference after model selection. *arXiv preprint arXiv:1410.2597*, 2014.

Fraser, D.A.S., Reid, N. & Sartori, N. Accurate directional inference for vector parameters. Biometrika, **103**, 625–639, 2016.

George, E.I. & Yekutieli, D. Shrinkage Adjustment for Model Selection. Presentation at JSM, 2012.

Kuffner, T.A. & Young, G.A. Principled statistical inference in data science. In *Statistical Data Science*. N. Adams, E. Cohen and Y.-K. Guo, editors. World Scientific, 21–36, 2018.

Lee, J.D., Sun, D.L., Sun, Y. & Taylor, J.E. Exact post-selection inference, with application to the lasso. Ann. Statist., **44**, 907–927, 2016.

Tian, X. & Taylor, J. Selective inference with a randomized response. Ann. Statist., **46**, 679–710, 2018.

Tibshirani, R.J., Rinaldo, A., Tibshirani, R. & Wasserman, L. Uniform asymptotic inference and the bootstrap after model selection. Ann. Statist., **46**, 1255-1287, 2018.

Yekutieli, D. Adjusted Bayesian inference for selected parameters. J. Roy. Statist. Soc. B, **74**, 515–541, 2012.

Young, G.A. & Smith, R.L. *Essentials of Statistical Inference*. CUP, 2005.