# The Surprising Conditional Adventures of the Bootstrap

G. Alastair Young

Department of Mathematics
Imperial College London

Inaugural Lecture, 13 March 2006

# Acknowledgements


- Early influences: Eric Renshaw, David Kendall, David Cox, Bernard Silverman, Henry Daniels, David Williams.

# Acknowledgements


- Early influences: Eric Renshaw, David Kendall, David Cox, Bernard Silverman, Henry Daniels, David Williams.
- Co-authors (21, last count), especially ($\geq 5$ joint publications) Daniela De Angelis, Tom DiCiccio, Peter Hall, Stephen Lee, Michael Martin.

# Acknowledgements


- Early influences: Eric Renshaw, David Kendall, David Cox, Bernard Silverman, Henry Daniels, David Williams.
- Co-authors (21, last count), especially ($\geq 5$ joint publications) Daniela De Angelis, Tom DiCiccio, Peter Hall, Stephen Lee, Michael Martin.
- Daniela De Angelis, again.

# Acknowledgements


- Early influences: Eric Renshaw, David Kendall, David Cox, Bernard Silverman, Henry Daniels, David Williams.
- Co-authors (21, last count), especially ($\geq 5$ joint publications) Daniela De Angelis, Tom DiCiccio, Peter Hall, Stephen Lee, Michael Martin.
- Daniela De Angelis, again.
- Tom DiCiccio, Yvonne Ho, Russell Zaretzki, contributions to this lecture.

# Acknowledgements


- Early influences: Eric Renshaw, David Kendall, David Cox, Bernard Silverman, Henry Daniels, David Williams.
- Co-authors (21, last count), especially ($\geq 5$ joint publications) Daniela De Angelis, Tom DiCiccio, Peter Hall, Stephen Lee, Michael Martin.
- Daniela De Angelis, again.
- Tom DiCiccio, Yvonne Ho, Russell Zaretzki, contributions to this lecture.
- Rudolf Raspe, publication in 1785 of *The Travels and Surprising Adventures of Baron Munchausen*.

# The rogue Raspe

# The Adventures of Baron Munchausen

- A voyage to an island composed entirely of cheese, and judged larger than Europe.

# The Adventures of Baron Munchausen

- A voyage to an island composed entirely of cheese, and judged larger than Europe.
- Two voyages to the moon (one accidental).

# The Adventures of Baron Munchausen

- A voyage to an island composed entirely of cheese, and judged larger than Europe.
- Two voyages to the moon (one accidental).
- Salvation from death by drowning in a swamp of quicksand by lifting himself (and his horse) up by his hair, in later versions pulling himself up by his bootstraps.

# Bootstrap methods of inference

Sample data $D$ plays an interventionist role in determining the analysis applied to itself. 'Random data corruption' methods, among others.

Small-scale problems, small $D$.

Different answers from different approaches. Resolve conflict?

# The bootstrap

Formalised, 'the bootstrap', by Efron (1979): estimates of statistical variability, by using empirical sampling models constructed from $D$ together with simulation from the empirical model.

Replace analytic calculations and approximations required by conventional approaches.

Pull ourselves up by bootstraps, by pretending empirical model is true (unknown) probability distribution from which $D$ has come.

# Parametric statistical inference

Data $D = \{X_1, \ldots, X_n\}$, assumed to be a random sample from (infinite) population, the randomness expressed through a probability density function, which expresses relative plausibility of different values.

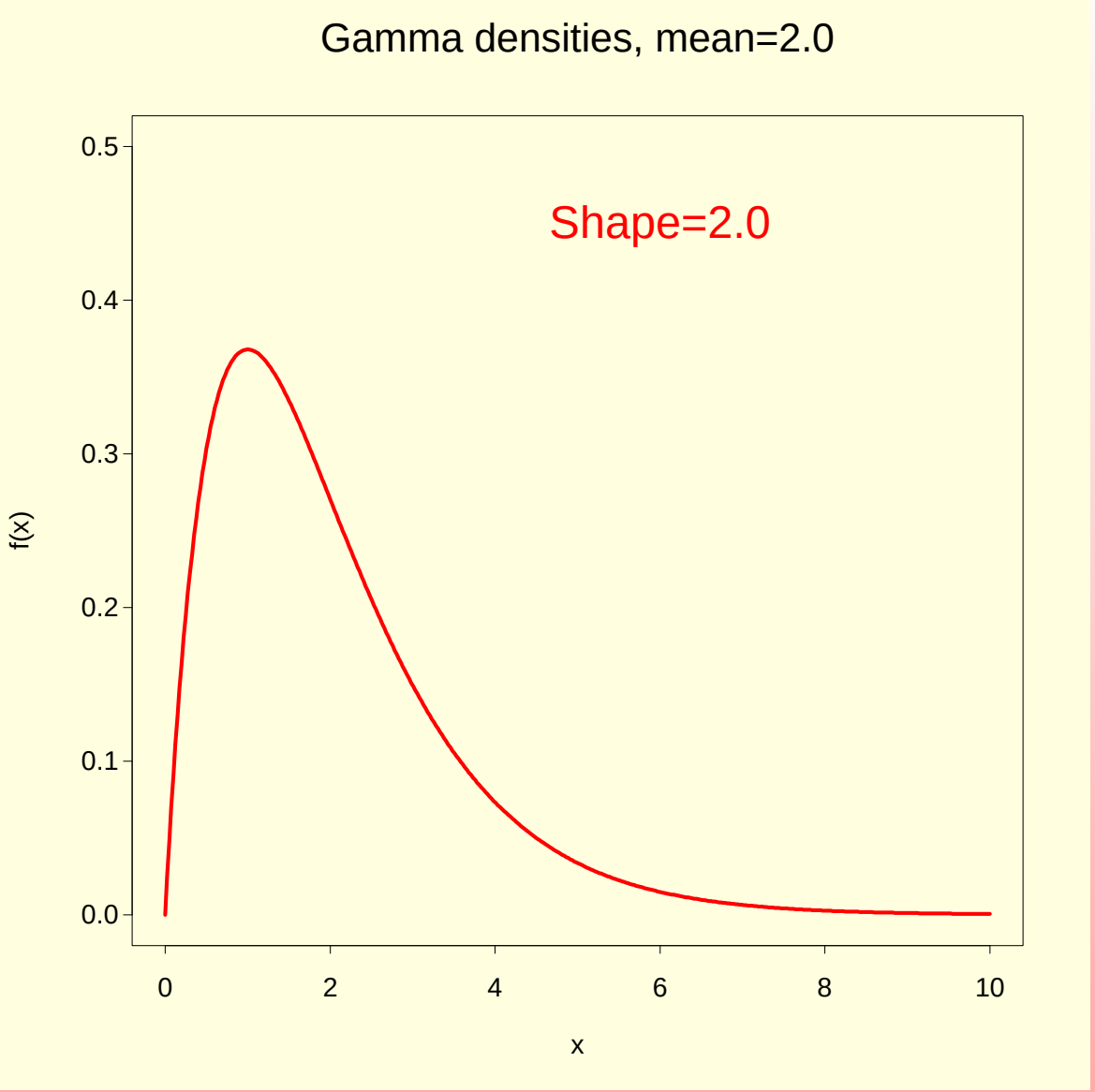
Density function $f(X; \theta)$ of specified functional form, but depending on parameter $\theta = (\mu, \nu)$, value unspecified.

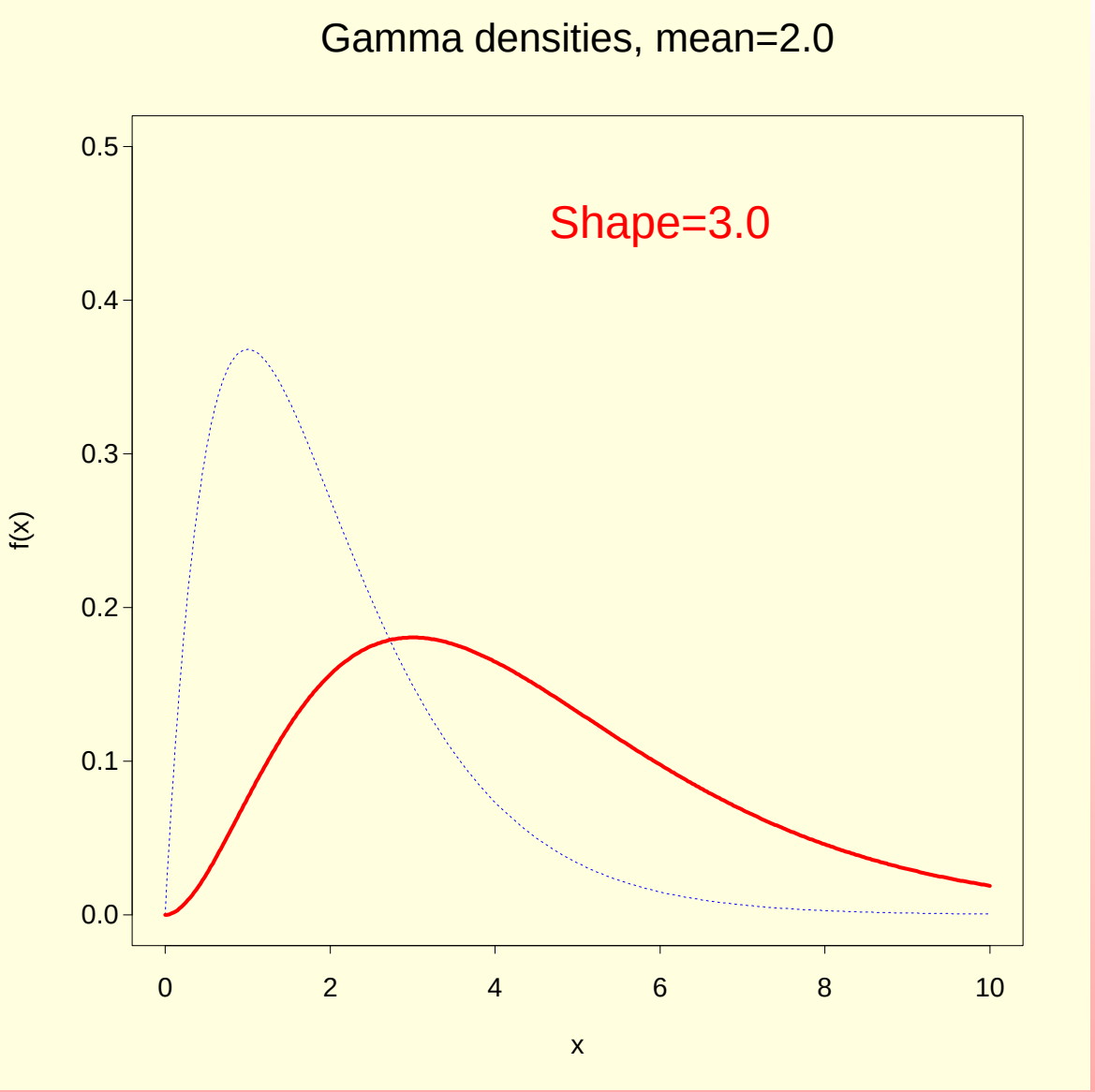
# An example

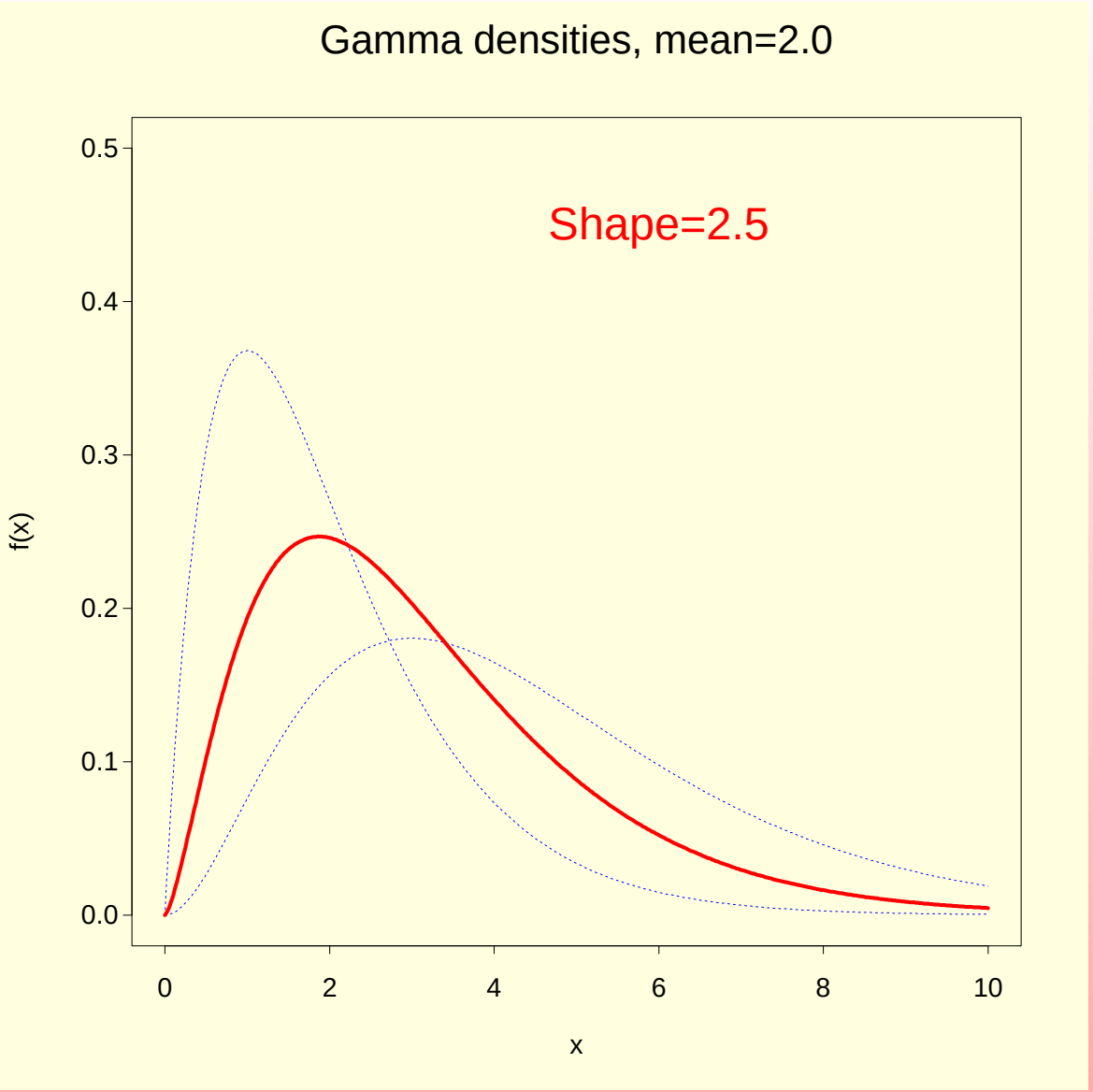
A Gamma density, of mean $\mu$ and shape parameter $\nu$ has density of functional form:

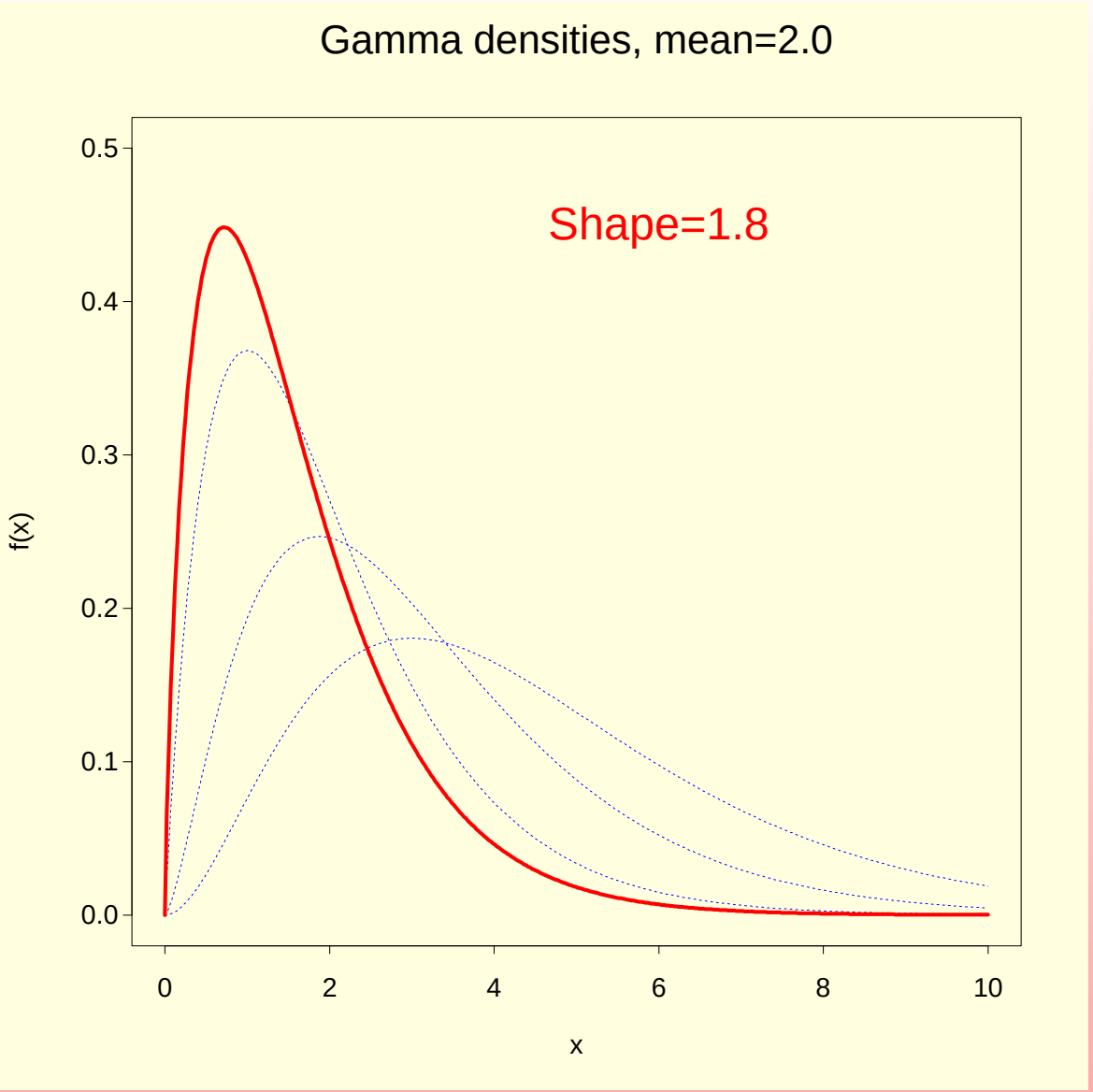
$$f(X;\theta) = \frac{\nu^{\nu}}{\Gamma(\nu)} \exp[-\nu\{\frac{X}{\mu} - \log(\frac{X}{\nu})\}]\frac{1}{X}.$$

By allowing $\mu$ (location) and $\nu$ (concentration) to vary, generate a flexible class of probability distributions.


Gamma densities, mean=2.0
Shape=2.0
f(x)
0.0
0.1
0.2
0.3
0.4
0.5
0
2
4
6
8
10
x


Gamma densities, mean=2.0
Shape=3.0
f(x)
0.0
0.1
0.2
0.3
0.4
0.5
0
2
4
6
8
10
x


Gamma densities, mean=2.0
Shape=2.5
f(x)
0.0
0.1
0.2
0.3
0.4
0.5
0
2
4
6
8
10
x


Gamma densities, mean=2.0
Shape=1.8
f(x)
x
0.0
0.1
0.2
0.3
0.4
0.5
0
2
4
6
8
10

# The inference problem

Typically, $\mu$ is an interest parameter, and $\nu$ is a nuisance parameter.

Objective of data analysis: test consistency of $D$ with the hypothesis ('null hypothesis, $H_0$') that $\mu$ takes a particular value, $\mu = \mu_0$.

Let

$$l(\theta) \equiv l(\mu, \nu) = \Sigma_{i=1}^{n} \log f(X_i; \mu, \nu)$$

be the log-likelihood for $(\mu, \nu)$, given $D$.

Let $\hat{\theta} = (\hat{\mu}, \hat{\nu})$ be the global maximum likelihood estimator (MLE), which maximises $l(\theta)$.

Let $\hat{\theta}_0 = (\mu_0, \hat{\nu}_0)$ be the constrained MLE , which maximises $l(\theta)$, subject to the constraint $\mu = \mu_0$.

Inference is based on the test statistic

$$r(\mu_0) = \mathrm{sgn}(\hat{\mu} - \mu_0)\sqrt{2\{l(\hat{\theta}) - l(\hat{\theta}_0)\}}.$$

Denote value of $r(\mu_0)$ for $D$ by $r_D$. Frequentist approach: compare $r_D$ with the distribution of values of $r(\mu_0)$ for (hypothetical) datasets from the population, if $H_0$ is true. Reject $H_0$ if $r_D$ is 'extreme' for this distribution.

Specifically, calculate the p-value, $\mathrm{prob}\{r(\mu_0) \geq r_D \mid H_0\}$, reject $H_0$ if this is small.

But the sampling distribution of $r(\mu_0)$ under $H_0$ is not known, as $\nu$ remains unspecified, even under $H_0$. Can't calculate the p-value.

But the sampling distribution of $r(\mu_0)$ under $H_0$ is not known, as $\nu$ remains unspecified, even under $H_0$. Can't calculate the p-value.

Either: Approximate distribution analytically. Asymptotically ($n \to \infty$) the sampling distribution if $H_0$ is true is 'standard Gaussian', $N(0, 1)$. More sophisticated approximations possible.

But the sampling distribution of $r(\mu_0)$ under $H_0$ is not known, as $\nu$ remains unspecified, even under $H_0$. Can't calculate the p-value.

Either: Approximate distribution analytically. Asymptotically ($n \to \infty$) the sampling distribution if $H_0$ is true is 'standard Gaussian', $N(0, 1)$. More sophisticated approximations possible.

Or: bootstrap.

# A Gamma dataset

Dataset $D$ of size $n = 20$ on survival times (in some units) of 20 mice exposed to 240 rads of gamma radiation: 152, 115, 152, 109, 137, 88, 94, 77, 160, 165, 125, 40, 128, 123, 136, 101, 62, 153, 83, 69.

Reasonable to model survival time using Gamma distribution. Consider testing $H_0 : \mu = \mu_0$, mean survival time is $\mu_0$.

# The bootstrap calculation, to test $H_0$

- From $D$, obtain $\hat{\mu}, \hat{\nu}, \hat{\nu}_0$ and calculate $r_D$.

# The bootstrap calculation, to test $H_0$

- From $D$, obtain $\hat{\mu}, \hat{\nu}, \hat{\nu}_0$ and calculate $r_D$.
- Simulate $B$ (actual) datasets, $D_1^*, \ldots, D_B^*$, say, each of size $n = 20$, from the Gamma density $f(X; \mu_0, \hat{\nu}_0)$. [Easy: big $B$, millions, feasible].

# The bootstrap calculation, to test $H_0$

- From $D$, obtain $\hat{\mu}, \hat{\nu}, \hat{\nu}_0$ and calculate $r_D$.
- Simulate $B$ (actual) datasets, $D_1^*, \ldots, D_B^*$, say, each of size $n = 20$, from the Gamma density $f(X; \mu_0, \hat{\nu}_0)$. [Easy: big $B$, millions, feasible].
- By repeating for each $D_i^*$ the calculations involved in determining $r(\mu_0)$, obtain associated values $r_{D_1^*}, \ldots, r_{D_B^*}$ of the test statistic, representing $H_0$.

# The bootstrap calculation, to test $H_0$

- From $D$, obtain $\hat{\mu}, \hat{\nu}, \hat{\nu}_0$ and calculate $r_D$.
- Simulate $B$ (actual) datasets, $D_1^*, \ldots, D_B^*$, say, each of size $n = 20$, from the Gamma density $f(X; \mu_0, \hat{\nu}_0)$. [Easy: big $B$, millions, feasible].
- By repeating for each $D_i^*$ the calculations involved in determining $r(\mu_0)$, obtain associated values $r_{D_1^*}, \ldots, r_{D_B^*}$ of the test statistic, representing $H_0$.
- The bootstrap p-value is the proportion of the $B$ simulated values $\geq r_D$.

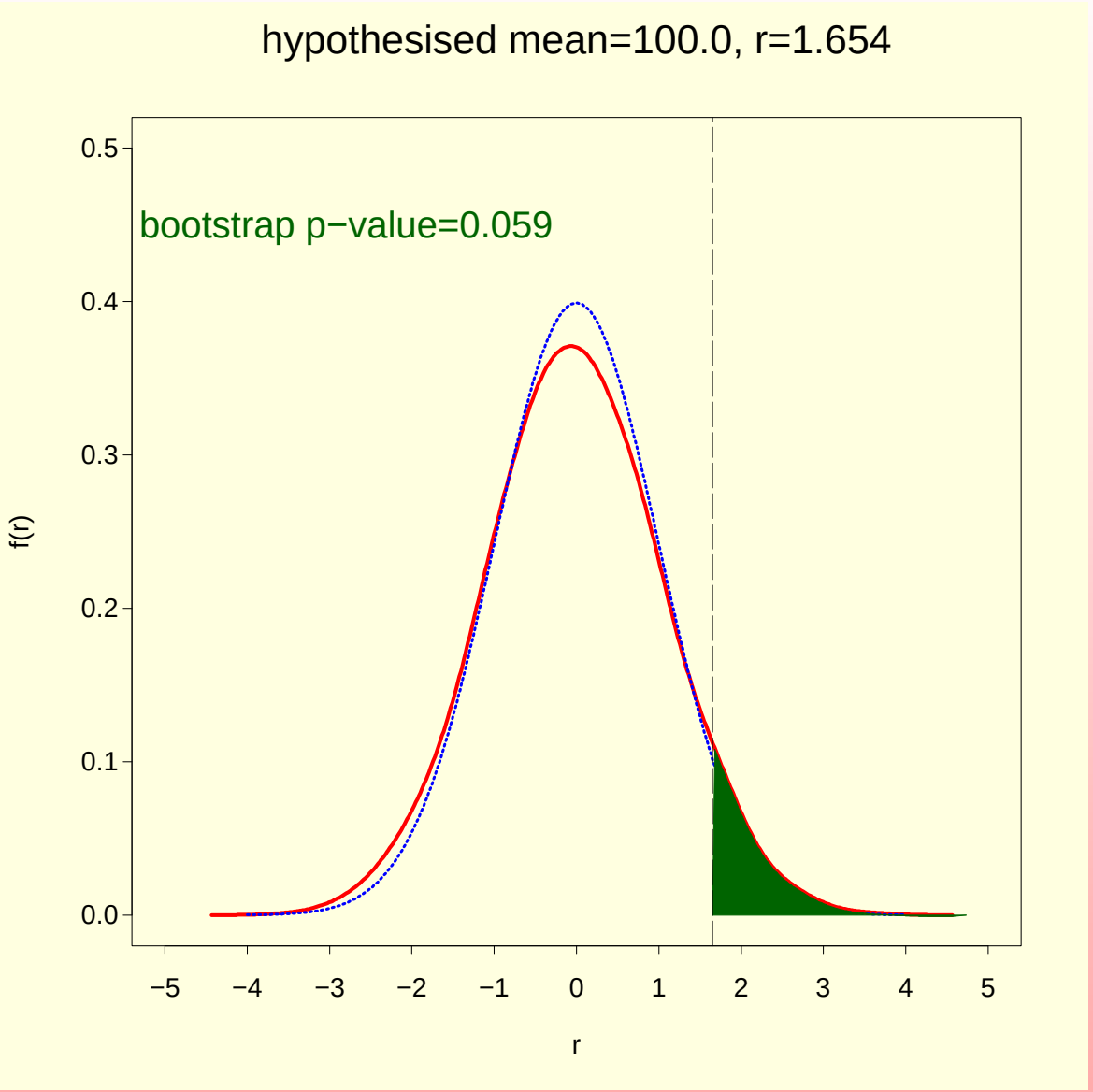
# Illustration: testing $H_0 : \mu = 100$

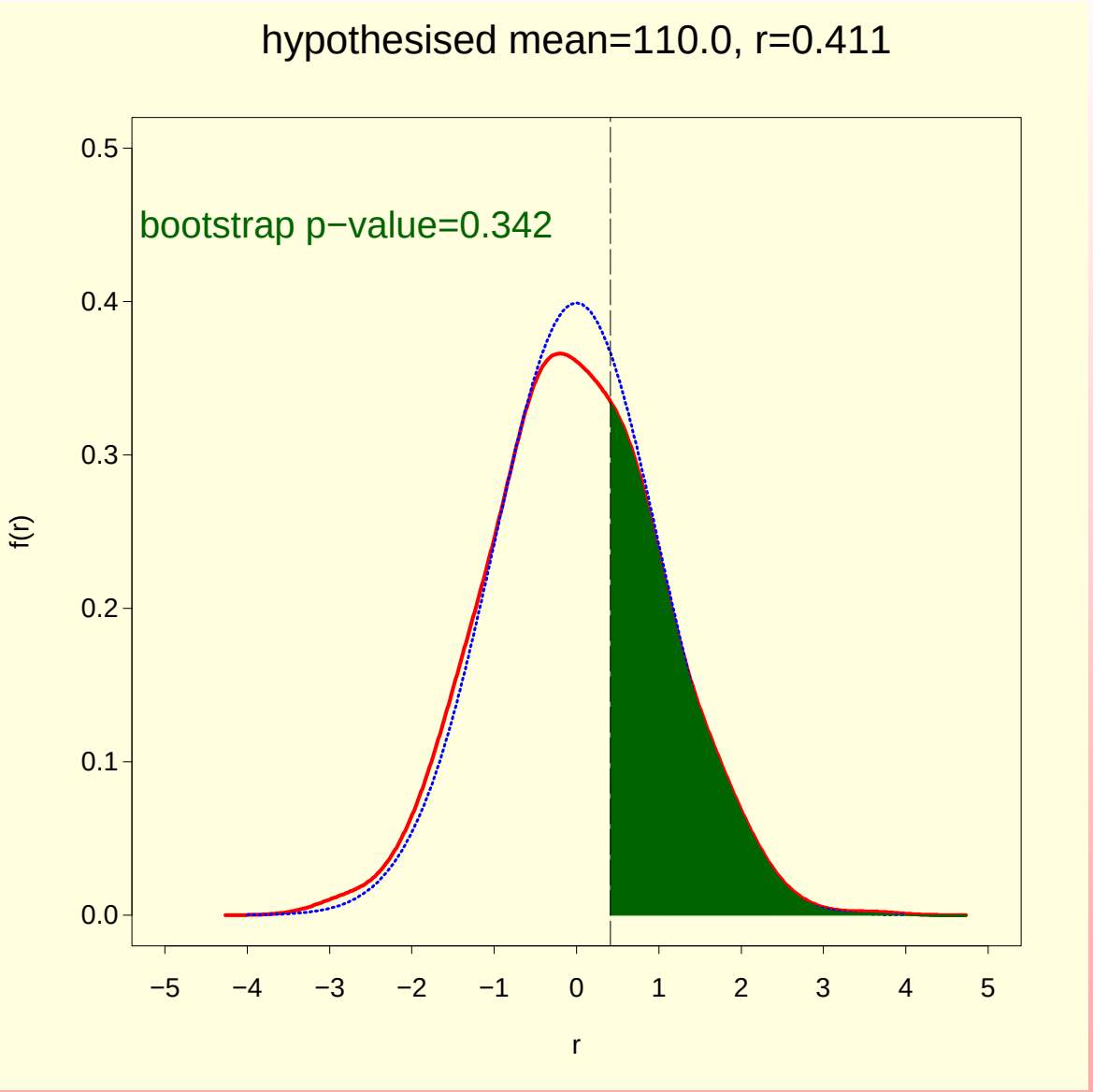
With $\mu_0 = 100$, we find $\hat{\nu}_0 = 7.715$, $r_D = 1.654$.

Simulate $B = 5,000,000$ data samples of size n=20 from Gamma density $f(X; 100, 7.715)$.

For each simulated dataset, compute $r(\mu_0)$, observe that the proportion giving value larger than $r_D$ is 5.9%.

This bootstrap p-value is large enough that we would conclude there is no strong evidence against $H_0$.


hypothesised mean=100.0, r=1.654
bootstrap p−value=0.059
f(r)
r
0.0
0.1
0.2
0.3
0.4
0.5
−5
−4
−3
−2
−1
0
1
2
3
4
5


hypothesised mean=110.0, r=0.411
bootstrap p−value=0.342
f(r)
0.0
0.1
0.2
0.3
0.4
0.5
−5
−4
−3
−2
−1
0
1
2
3
4
5
r

# Is the bootstrap any good?

Repeated sampling perspective: yes.

If $H_0$ is true (population really has $\mu = \mu_0$), under repeated sampling of $D$, bootstrap p-value is distributed as uniform on $(0, 1)$, to error of order $n^{-3/2}$. Correct answers would correspond to distribution being exactly uniform.

# Conditional inference

Repeated sampling perspective is narrow. A more sophisticated analysis should take into account demands of conditional inference.

Controversial, murky area. Basic idea: need to make inference relevant to $D$ by conditioning on features of $D$ which, in themselves, say nothing about quantity of interest, $\mu$, but which control propensity for extreme values of test statistic to occur for artificial reasons.

Restrict frequentist inference to involve (hypothetical) datasets which have same value as $D$ of some 'ancillary statistic'. 'Conditionality Principle'.

# Motivating example

Physical quantity $\theta$ can be measured by two machines, both giving (Gaussian) measurements $X$ which have mean $\theta$. First machine is precise, measurement error is low, Gaussian distribution has low variance, but second machine gives measurements of high variability about $\theta$.

Precise machine is often busy, second machine will be used only if first is unavailable: through repeated observation we know that each machine is equally likely to be used.

We are given an observation, and told it has come from the first machine.

Should our inference on $\theta$ take into account that the second machine might have been used, but in the event wasn't?

We are given an observation, and told it has come from the first machine.

Should our inference on $\theta$ take into account that the second machine might have been used, but in the event wasn't?

Silly to take into account (as in frequentist approach) that second machine might have been used, when we know that it wasn't.
Draw inference from distribution of $X$ for machine actually used: 'machine used' is ancillary statistic.

# The swamp of conditional inference

Some of the formal difficulties with conditional inference:

# The swamp of conditional inference

Some of the formal difficulties with conditional inference:

- Conflict between conditioning and 'power', ability to correctly identify that $H_0$ is false. 'Cut off right hand to save left'.

# The swamp of conditional inference

Some of the formal difficulties with conditional inference:

- Conflict between conditioning and 'power', ability to correctly identify that $H_0$ is false. 'Cut off right hand to save left'.
- Typically, arbitrariness of what to condition on, ancillary statistics are not unique. 'Dithering'.

# The swamp of conditional inference

Some of the formal difficulties with conditional inference:

- Conflict between conditioning and 'power', ability to correctly identify that $H_0$ is false. 'Cut off right hand to save left'.
- Typically, arbitrariness of what to condition on, ancillary statistics are not unique. 'Dithering'.
- Mathematical contradiction. Formally, acceptance of (totally uncontroversial) 'sufficiency principle' together with conditionality principle requires acceptance of 'likelihood principle'. The likelihood principle is incompatible with common methods of inference such as calculation of p-values. 'Sawing off the branch of the tree that you are sitting on'.

A schizophrenic attitude is quite common.

# Lifting ourselves out of the swamp

# Lifting ourselves out of the swamp

Be Bayesian. 'Out of the frying pan into the fire'.

# Lifting ourselves out of the swamp

Be Bayesian. 'Out of the frying pan into the fire'.

Or, construct new approach to inference e.g. Fisher's fiducial theory. "Fisher's biggest blunder".

# Lifting ourselves out of the swamp

Be Bayesian. 'Out of the frying pan into the fire'.

Or, construct new approach to inference e.g. Fisher's fiducial theory. "Fisher's biggest blunder".

Or, more modestly, use forms of frequentist inference which deliver the same solution, whether applied unconditionally or conditionally on any relevant ancillary.

Identify unconditional procedures which yield same inference as we would obtain from conditional inference, were we to agree on it.

# Efron's bootstrap lifts us..

Bootstrap calculations as described (no conditioning specified) yield inference which respects conditioning to astonishing degree.

# Two contexts for conditioning

- 'exponential family models', where $l$ has particular form: conditioning has effect of eliminating nuisance parameter (uncontroversial);

# Two contexts for conditioning

- 'exponential family models', where $l$ has particular form: conditioning has effect of eliminating nuisance parameter (uncontroversial);
- where there exists ancillary statistic $a$. [Statistic with distribution not depending on $\theta$ which, together with MLE, is 'minimal sufficient']. Inference should consider only (hypothetical) samples with same value of $a$ as $D$ (very controversial).

# Bootstrap and conditional inference

Exponential family context: (unconditional) bootstrap calculations approximate exact conditional inference to third-order, $O(n^{-3/2})$, as good as sophisticated analytic methods.

# Bootstrap and conditional inference

Exponential family context: (unconditional) bootstrap calculations approximate exact conditional inference to third-order, $O(n^{-3/2})$, as good as sophisticated analytic methods.

Ancillary statistic context: bootstrap calculations approximate exact conditional inference to second-order, $O(n^{-1})$. Good enough?

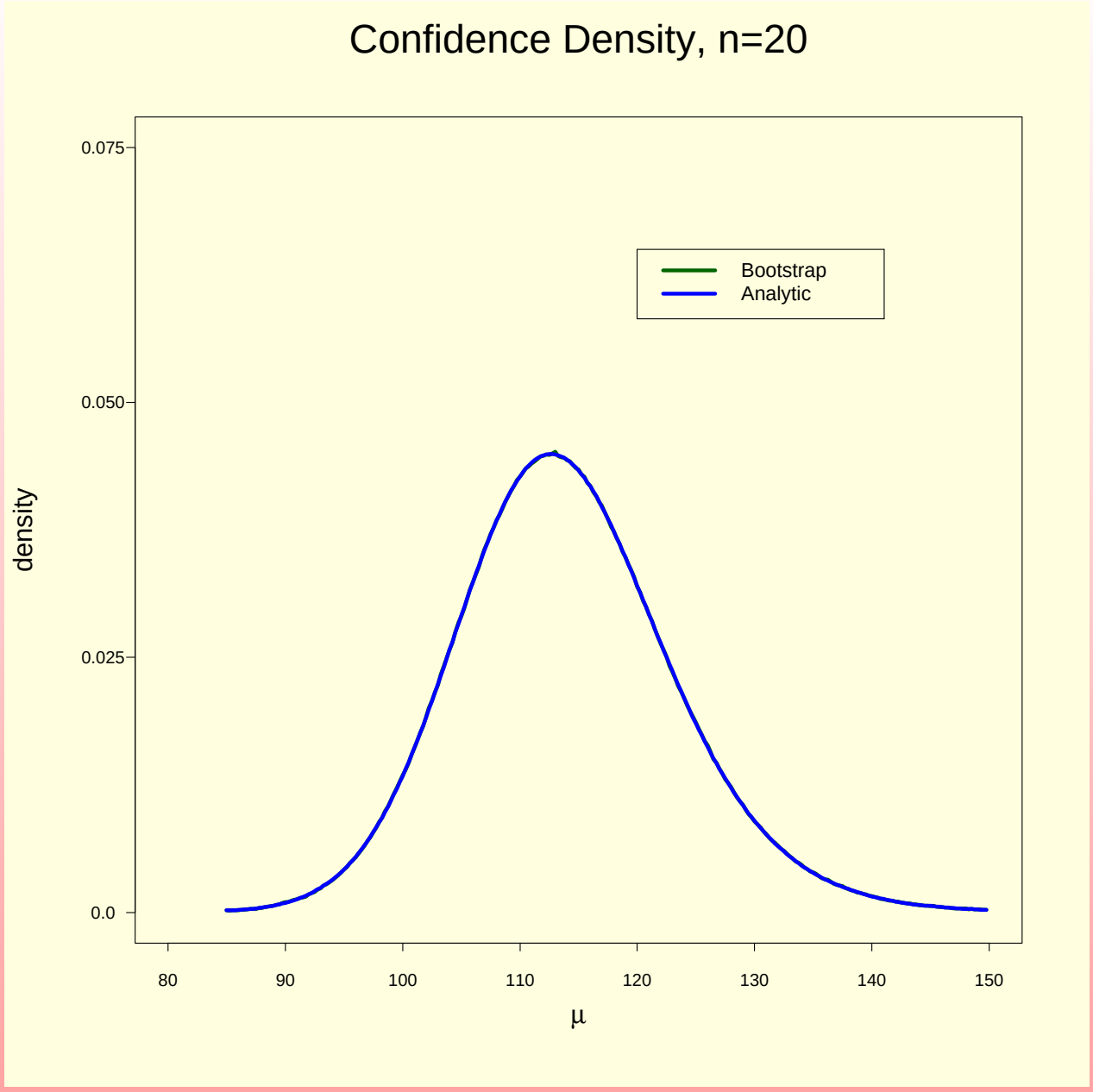
# Bootstrap and conditional inference

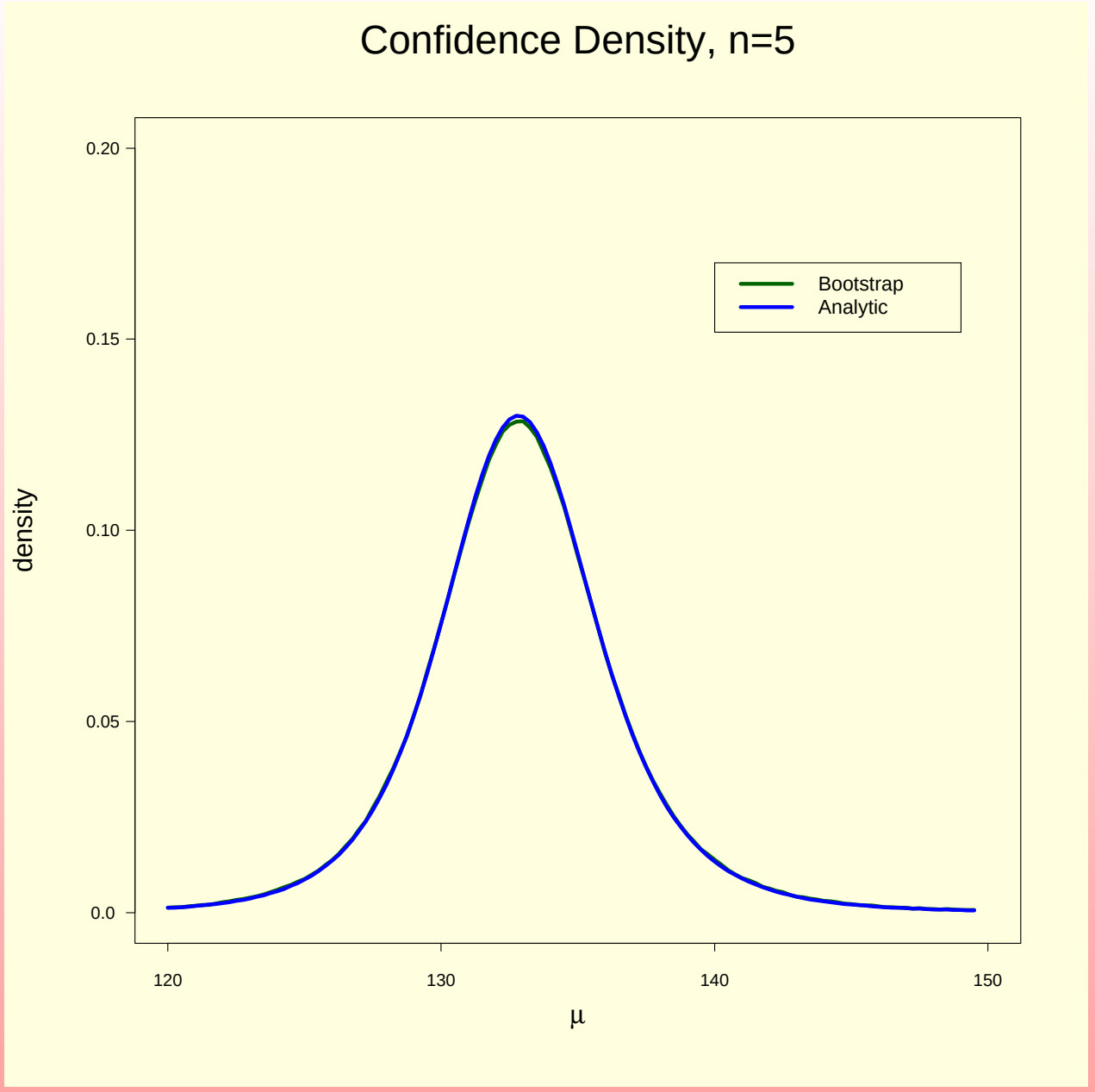
Exponential family context: (unconditional) bootstrap calculations approximate exact conditional inference to third-order, $O(n^{-3/2})$, as good as sophisticated analytic methods.

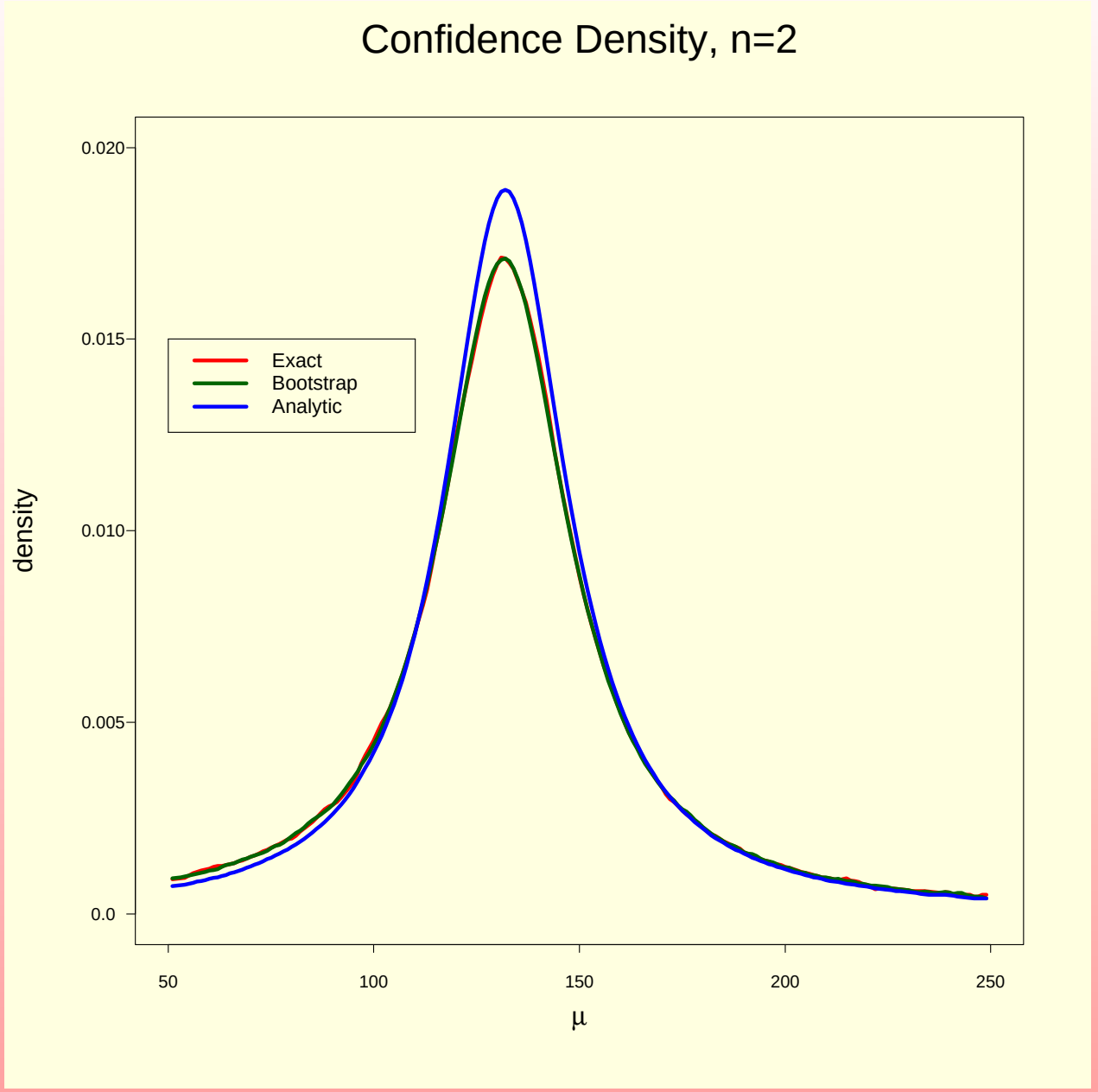
Ancillary statistic context: bootstrap calculations approximate exact conditional inference to second-order, $O(n^{-1})$. Good enough? Yes, insisting on greater conditional accuracy is unwarranted.

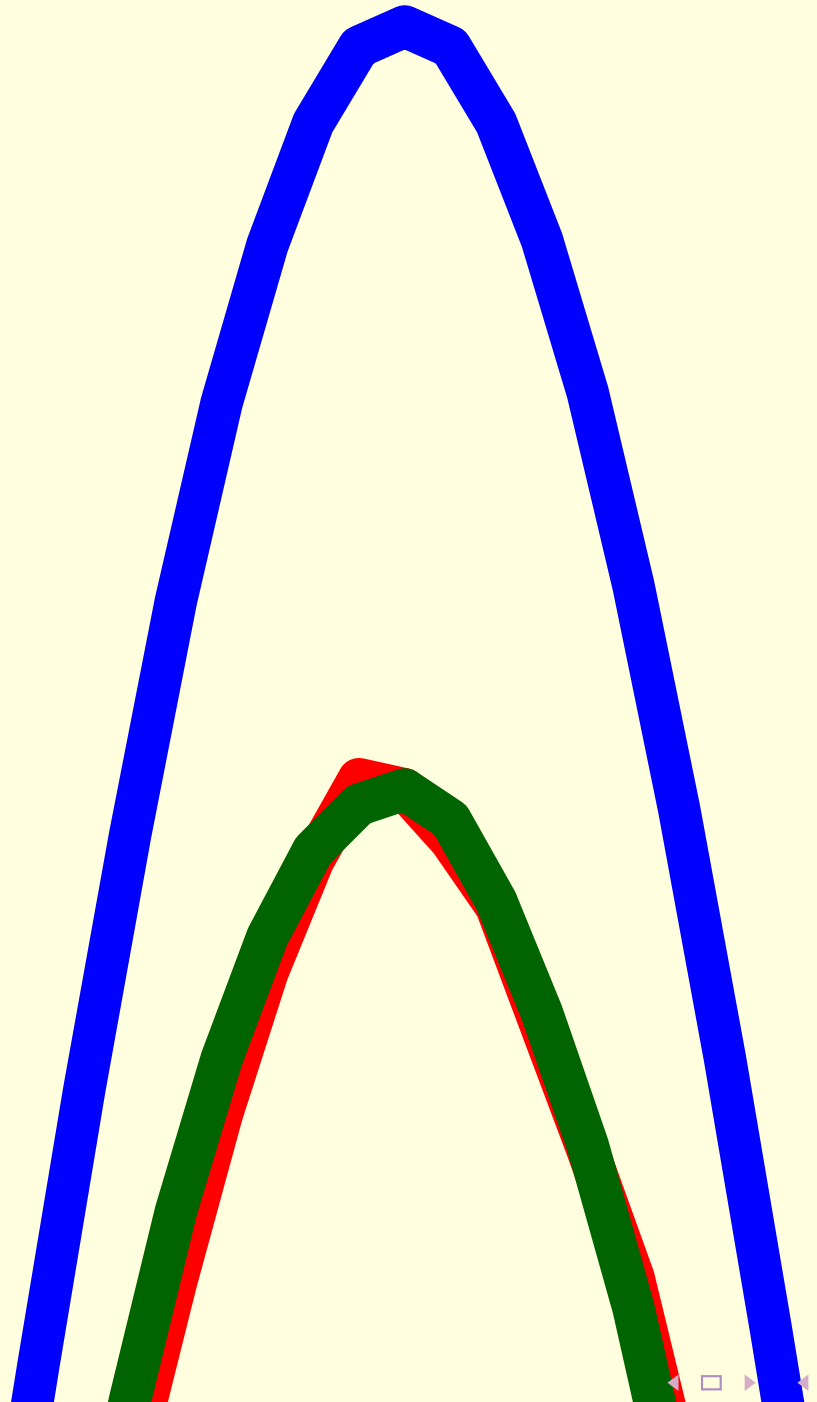
# The Gamma mean problem

Exponential family. Exact conditional test is analytically intractable, except for $n = 2$ or 3. Compare bootstrap with analytic procedures, specifically designed to approximate the exact inference to third-order.


Confidence Density, n=20
0.075
0.050
0.025
0.0
density
Bootstrap
Analytic
80
90
100
110
120
130
140
150
μ


Confidence Density, n=5
density
0.20
0.15
0.10
0.05
0.0
120
130
140
150
μ
Bootstrap
Analytic


Confidence Density, n=2
density
0.020
0.015
0.010
0.005
0.0
Exact
Bootstrap
Analytic
50
100
150
200
250
μ



# Exponential regression

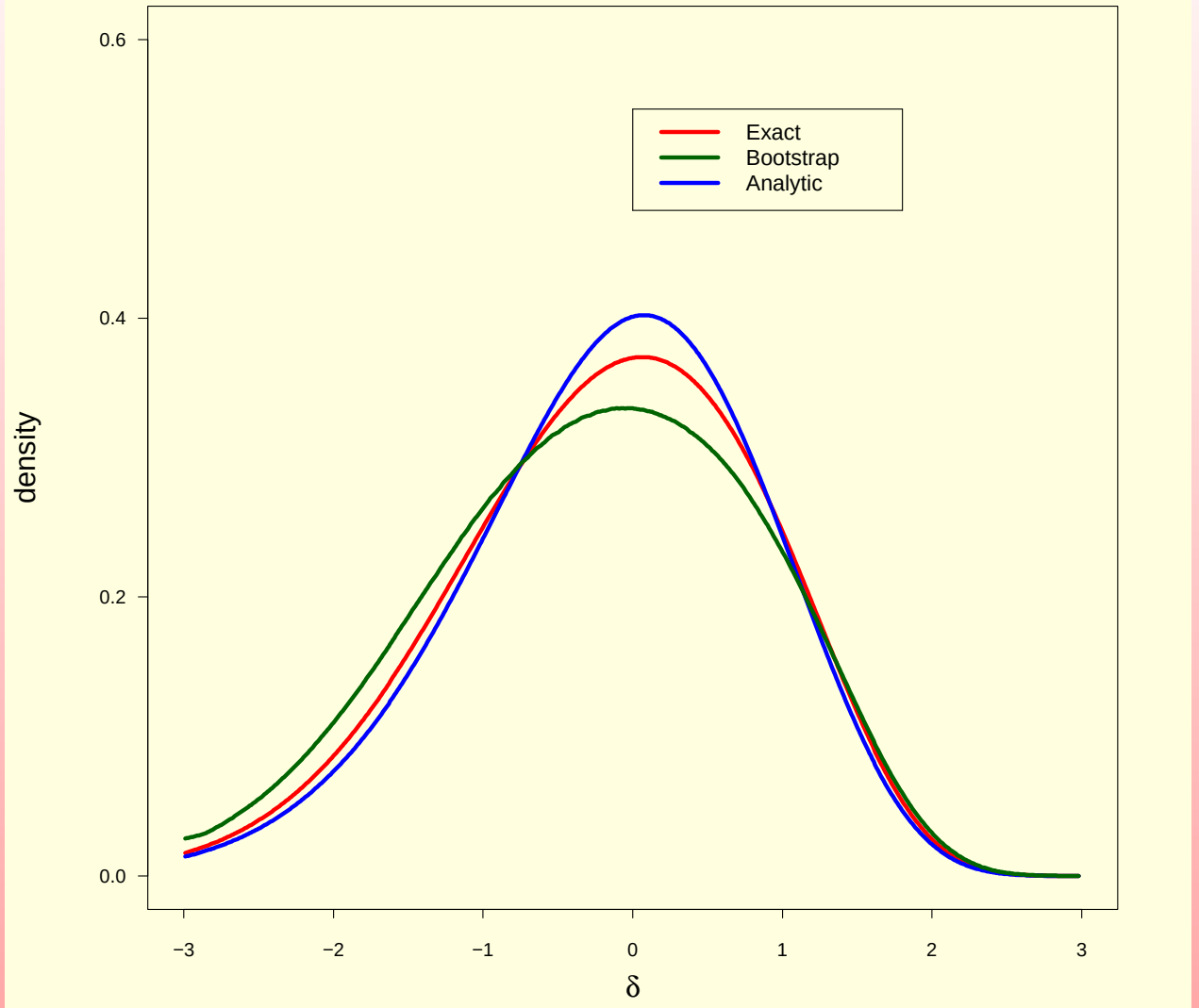
Ancillary statistic model. Exact conditional inference feasible, awkward.
Have $D = \{X_1, \ldots, X_n\}$ independent survival times, $f(X_i; \mu_i)$ exponential,

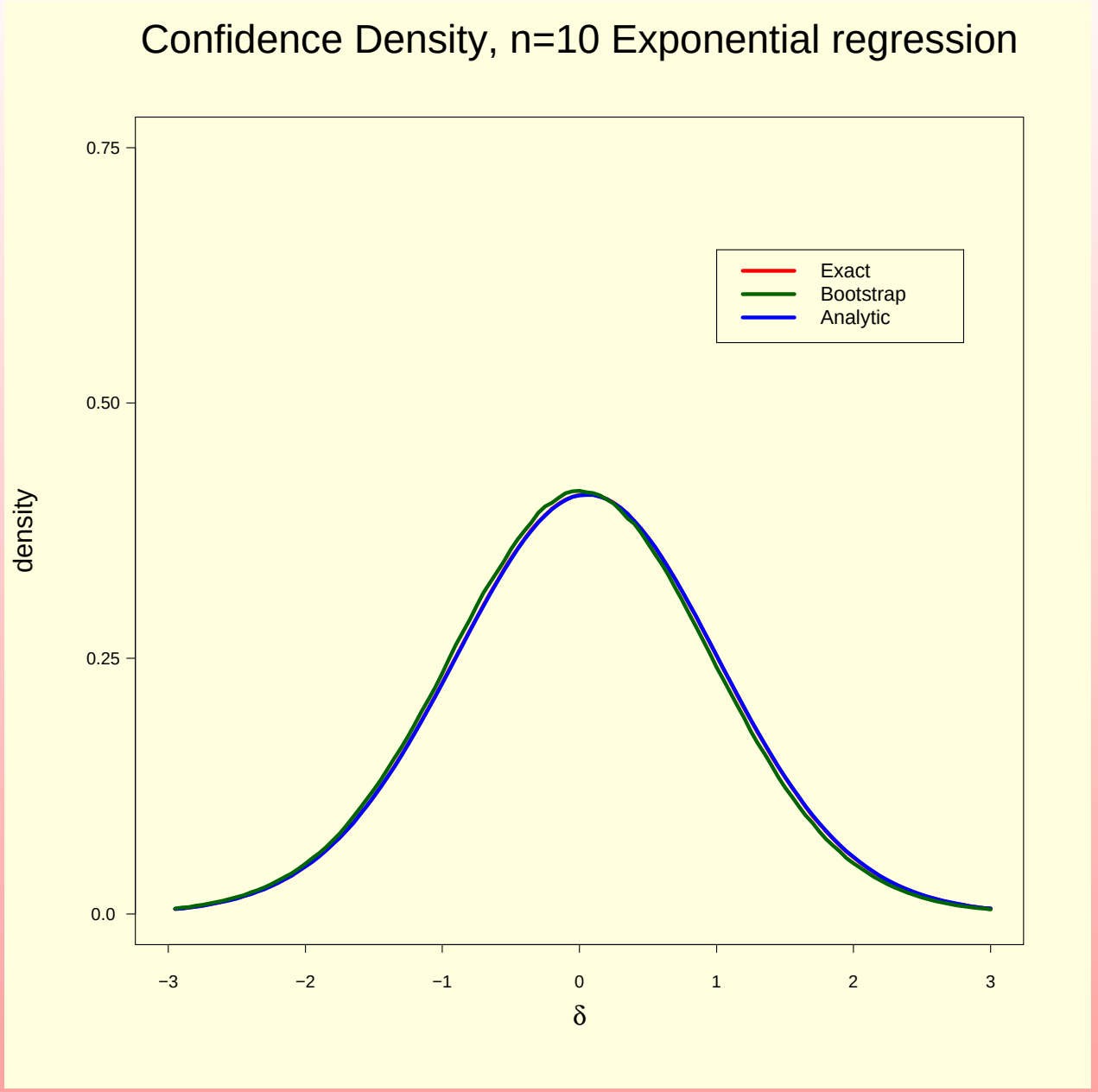
$$f(X_i; \mu_i) = \frac{1}{\mu_i} \exp(-X_i/\mu_i),$$

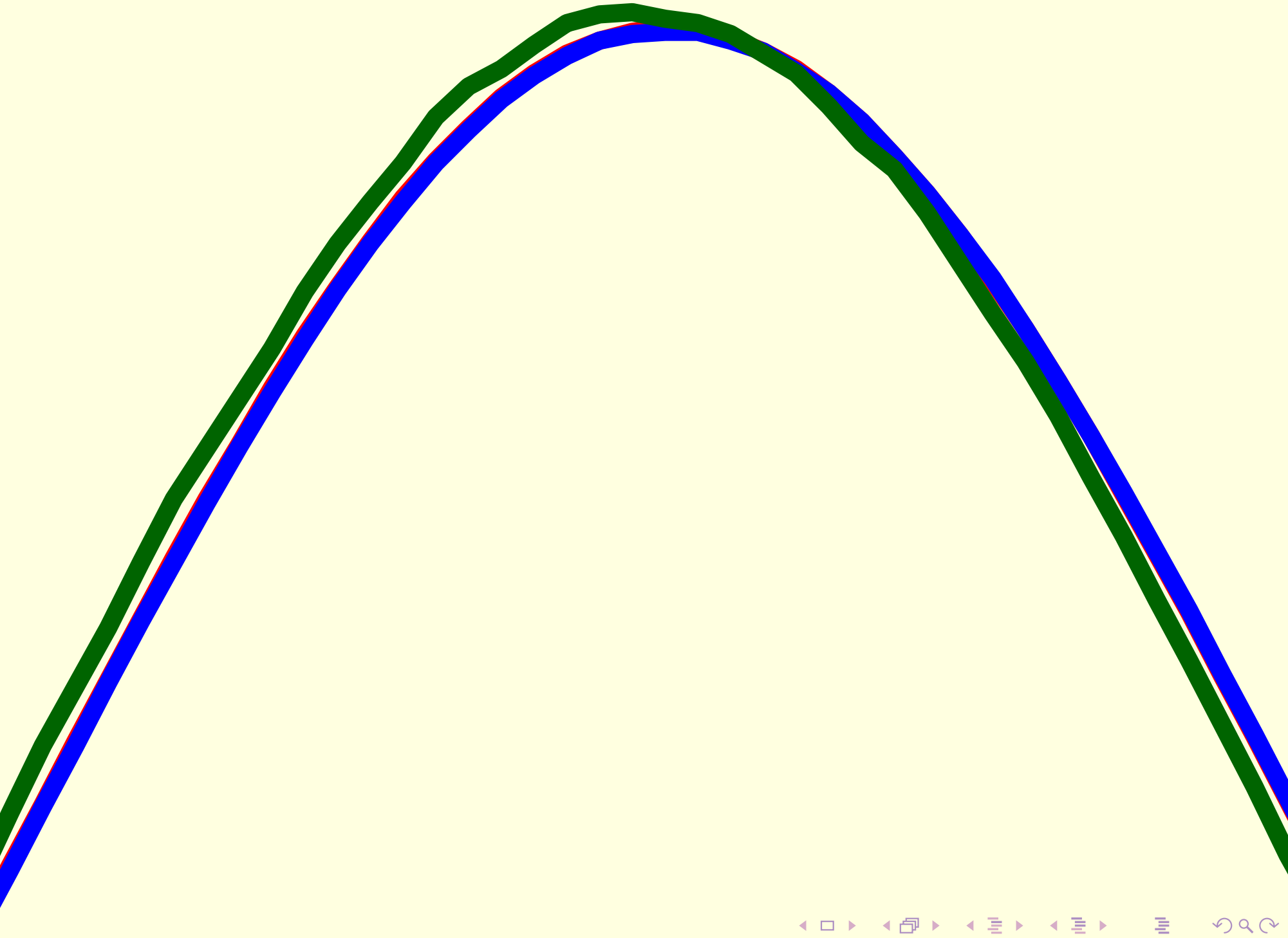
with mean $\mu_i$ depending on known covariate value $z_i$.
Interested in mean survival $\mu$ for covariate $z_0$, in presence of a nuisance parameter. Test $H_0 : \mu = \hat{\mu} + \delta$.

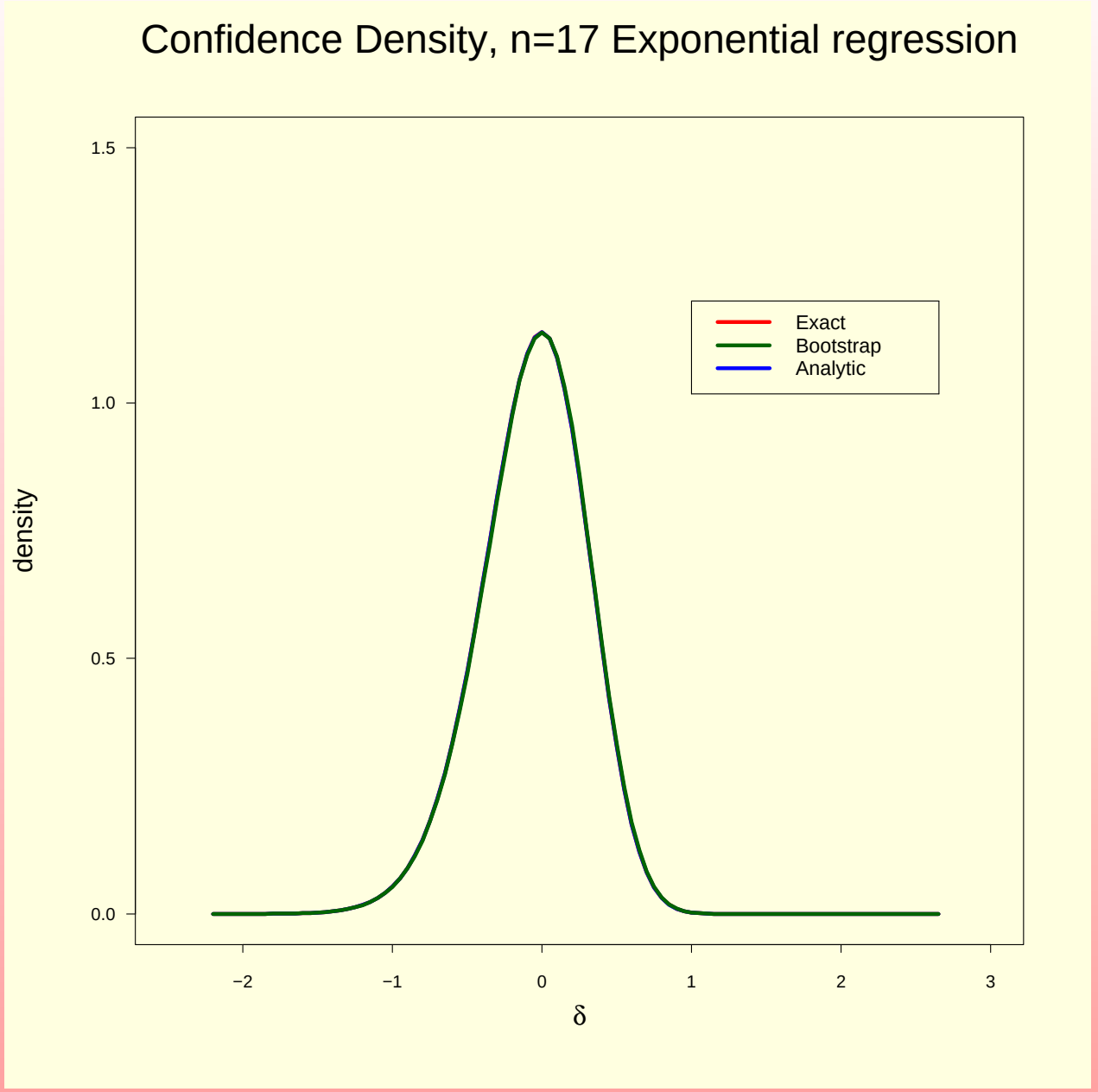
The $n = 5$ responses $X_i$ are 156, 108, 143, 65 and 1, survival times (in weeks) of leukaemia patients, covariate values $z_i$ are base-10 logarithms of initial blood cell count: 2.88, 4.02, 3.85, 5.0, 5.0. Take $z_0 = \log_{10}(50000) \approx 4.7$.

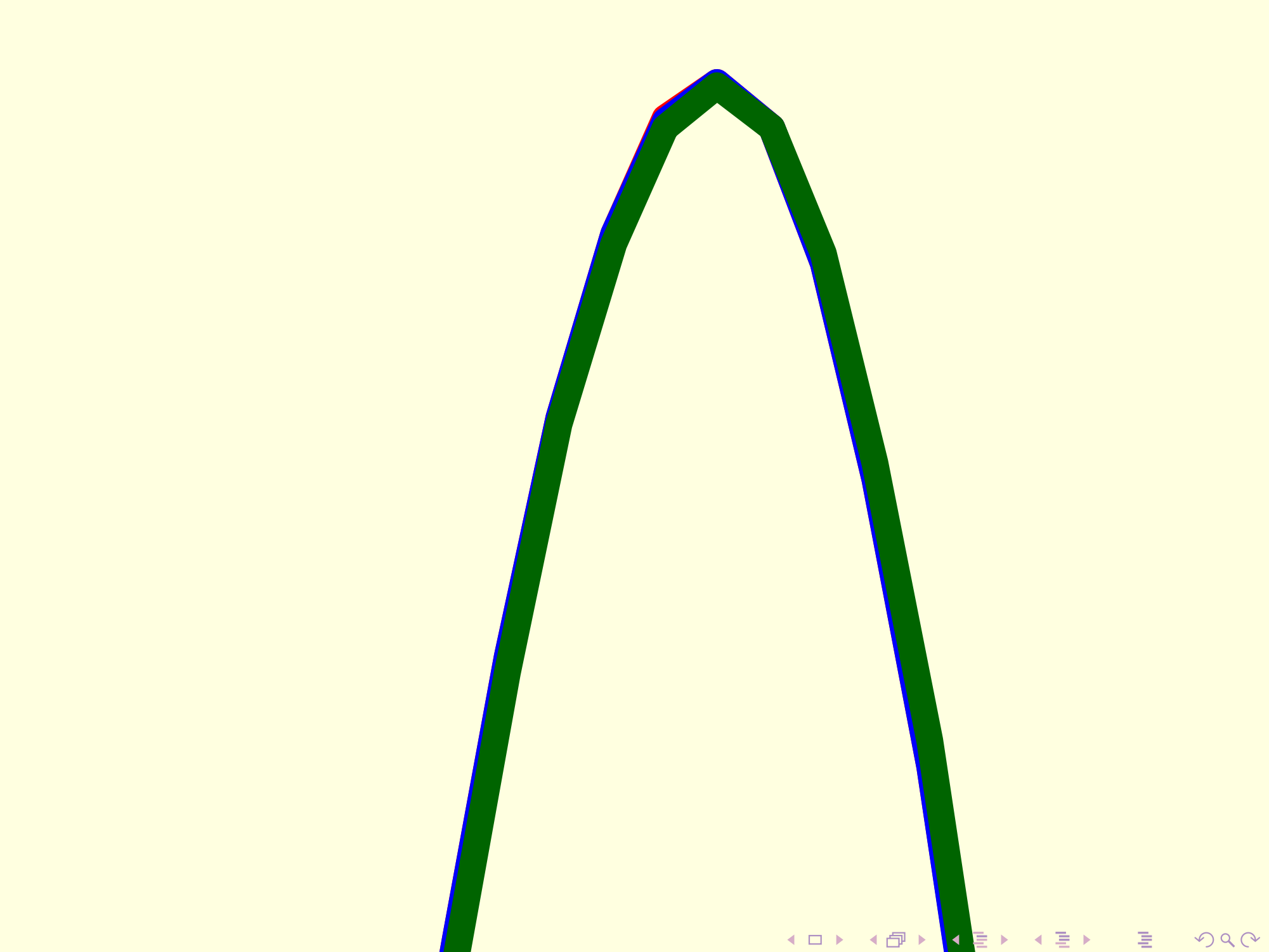
# Confidence Density, n=5 Exponential regression




Confidence Density, n=10 Exponential regression
density
0.75
0.50
0.25
0.0
−3
−2
−1
0
1
2
3
δ
Exact
Bootstrap
Analytic




Confidence Density, n=17 Exponential regression
density
δ
Exact
Bootstrap
Analytic
1.5
1.0
0.5
0.0
−2
−1
0
1
2
3



# Bootstrap is more subtle.....

Bootstrap methods allow accurate estimation of sampling characteristics of inferentially important statistics. Work well from repeated sampling perspective.

Also automatically encapsulate sophisticated statistical thought. Provide pragmatic solution to debate on conditional inference.

# Last word to the Baron...