Priors from a differential viewpoint:
How Bayes can deliver 2nd order Accuracy!

D A S Fraser
Statistical Sciences
Univ Toronto

Western University
2014 April 10

www.utstat.toronto.edu/dfraser/documents/UWO2014.pdf

some references as: .... ~/documents/xxx.pdf where xxx =

Priors from a differential viewpoint:
How Bayes can deliver 2nd order Accuracy!

With a long history

great collaborators: Nancy

A Wong	York
M Bédard	U de Montréal
W Lin	Toronto
A M Fraser	UBC
M J Fraser	Toronto

(Preliminary report)

Background:

2nd Order Bayes ?

Background:

1) Science 2014 — Reproducibility & Statistics

2a) Science 2011 — "Data" & the 'dust-up'

2b) but... — Retractions and ...... Reproducibility

3a) Science 2013 — Efron on...... Reproducibilty

3b) Science 2013 — Laplace had confidence.... Reproducibility

2nd Order Bayes ?

Background:

1) Science 2014 — Reproducibility & Statistics

2a) Science 2011 — "Data" & the 'dust-up'

2b) but... — Retractions and ...... Reproducibility

3a) Science 2013 — Efron on...... Reproducibilty

3b) Science 2013 — Laplace had confidence.... Reproducibility

2nd Order Bayes ?

1 Scalar parameter: Welch-Peers 1963 B + 200
Example

2 Scalar linear interest parameter
Example

3 Scalar rotating interest:
Example

4 Scalar curved interest:
Example

Background:

1) Science 2014 — Reproducibility & Statistics

2a) Science 2011 — "Data" & the 'dust-up'

2b) but... — Retractions and ...... Reproducibility

3a) Science 2013 — Efron on...... Reproducibilty

3b) Science 2013 — Laplace had confidence.... Reproducibility

2nd Order Bayes ?

1 Scalar parameter: Welch-Peers 1963 — B + 200
  Example

2 Scalar linear interest parameter
  Example

3 Scalar rotating interest:
  Example

4 Scalar curved interest:
  Example

Discussion

Science 2014

(1) View from Science

Editorial: Marcia McNutt Editor-in-chief
"Science" 17 January 2014

(1)

Editorial by Marcia McNutt Editor-in-chief
"Science" 17 January 2014

Marcia McNutt is Editor-in-Chief of *Science*.

Reproducibility

SCIENCE ADVANCES ON A FOUNDATION OF TRUSTED DISCOVERIES. REPRODUCING AN EXPERIMENT is one important approach that scientists use to gain confidence in their conclusions. Recently, the scientific community was shaken by reports that a troubling proportion of peer-reviewed preclinical studies are not reproducible. Because confidence in results is of paramount importance to the broad scientific community, we are announcing new initiatives to increase confidence in the studies published in *Science*. For preclinical studies (one of the targets of recent concern), we will be adopting recommendations of the U.S. National Institute of Neurological Disorders and Stroke (NINDS) for increasing transparency.* Authors will indicate whether there was a pre-experimental plan for data handling (such as how to deal with outliers), whether they conducted a sample size estimation to ensure a sufficient signal-to-noise ratio, whether samples were treated randomly, and whether the experimenter was blind to the conduct of the experiment. These criteria will be included in our author guidelines.

There are a number of reasons why peer-reviewed preclinical studies may not be reproducible. The system under investigation may be more complex than previously thought, so that the experimenter is not actually controlling all independent variables. Authors may not have divulged all of the details of a complicated experiment, making it irreproducible by another lab. It is also expected that through random chance, a certain number of studies will produce false positives. If researchers are not alert to this possibility and have not set appropriately stringent significance tests for their results, the outcome is a study with irreproducible results. Although there is always the possibility that an occasional study is fraudulent, the number of preclinical studies that cannot be reproduced is inconsistent with the idea that all irreproducibility results from misconduct in such research.

It is unlikely that the issues with irreproducibility are confined to preclinical studies (social science has been equally noted, for example). Unfortunately, there are no equivalents to the NINDS recommendations for other disciplines that provide a basis for requiring transparency across all fields. For the next 6 months, we will be asking reviewers and editors to identify papers submitted to *Science* that demonstrate excellence in transparency and instill confidence in the results. This will inform the next steps in implementing reproducibility guidelines. *Science Translational Medicine*, a sister journal of *Science*, already enforces the NINDS guidelines for preclinical studies. Both journals also are open to improving on the NINDS recommendations for preclinical studies.

There is also a wide range of sophistication in the application of statistics displayed in research analysis, ranging from practically no statistics, to the routine use of generic software packages, to the application of advanced methods that extract subtle signals from noise. Because reviewers who are chosen for their expertise in the subject matter of a study may not be authorities in statistics as well, statistical errors in manuscripts may slip through undetected. For that reason, with the advice of the American Statistical Association and others, we are adding new members to our Board of Reviewing Editors from the statistics community to ensure that manuscripts receive appropriate scrutiny in their methods of data analysis.

Science's standards have always been high, and these measures add to steps we have already taken to increase transparency, such as requiring data accessibility. Nevertheless, journals can only do so much to assure readers of the validity of the studies they publish. The ultimate responsibility lies with authors to be completely open with their methods, all of their findings, and the possible pitfalls that could invalidate their conclusions.

– **Marcia McNutt**

10.1126/science.1250475

*S. C. Landis *et al.*, *Nature* **490**, 187 (2012).

(1)

Editorial by Marcia McNutt Editor-in-chief
"Science" 17 January 2014

Emphasis: Reproducibity!

EDITORIAL

Reproducibility

Marcia McNutt is Editor-in-Chief of *Science*.

SCIENCE ADVANCES ON A FOUNDATION OF TRUSTED DISCOVERIES. REPRODUCING AN EXPERIMENT is one important approach that scientists use to gain confidence in their conclusions. Recently, the scientific community was shaken by reports that a troubling proportion of peer-reviewed preclinical studies are not reproducible. Because confidence in results is of paramount importance to the broad scientific community, we are announcing new initiatives to increase confidence in the studies published in *Science*. For preclinical studies (one of the targets of recent concern), we will be adopting recommendations of the U.S. National Institute of Neurological Disorders and Stroke (NINDS) for increasing transparency.* Authors will indicate whether there was a pre-experimental plan for data handling (such as how to deal with outliers), whether they conducted a sample size estimation to ensure a sufficient signal-to-noise ratio, whether samples were treated randomly, and whether the experimenter was blind to the conduct of the experiment. These criteria will be included in our author guidelines.

There are a number of reasons why peer-reviewed preclinical studies may not be reproducible. The system under investigation may be more complex than previously thought, so that the experimenter is not actually controlling all independent variables. Authors may not have divulged all of the details of a complicated experiment, making it irreproducible by another lab. It is also expected that through random chance, a certain number of studies will produce false positives. If researchers are not alert to this possibility and have not set appropriately stringent significance tests for their results, the outcome is a study with irreproducible results. Although there is always the possibility that an occasional study is fraudulent, the number of preclinical studies that cannot be reproduced is inconsistent with the idea that all irreproducibility results from misconduct in such research.

It is unlikely that the issues with irreproducibility are confined to preclinical studies (social science has been equally noted, for example). Unfortunately, there are no equivalents to the NINDS recommendations for other disciplines that provide a basis for requiring transparency across all fields. For the next 6 months, we will be asking reviewers and editors to identify papers submitted to *Science* that demonstrate excellence in transparency and instill confidence in the results. This will inform the next steps in implementing reproducibility guidelines. *Science Translational Medicine*, a sister journal of *Science*, already enforces the NINDS guidelines for preclinical studies. Both journals also are open to improving on the NINDS recommendations for preclinical studies.

There is also a wide range of sophistication in the application of statistics displayed in research analysis, ranging from practically no statistics, to the routine use of generic software packages, to the application of advanced methods that extract subtle signals from noise. Because reviewers who are chosen for their expertise in the subject matter of a study may not be authorities in statistics as well, statistical errors in manuscripts may slip through undetected. For that reason, with the advice of the American Statistical Association and others, we are adding new members to our Board of Reviewing Editors from the statistics community to ensure that manuscripts receive appropriate scrutiny in their methods of data analysis.

Science's standards have always been high, and these measures add to steps we have already taken to increase transparency, such as requiring data accessibility. Nevertheless, journals can only do so much to assure readers of the validity of the studies they publish. The ultimate responsibility lies with authors to be completely open with their methods, all of their findings, and the possible pitfalls that could invalidate their conclusions.

– Marcia McNutt

10.1126/science.1250475

*S. C. Landis *et al.*, *Nature* **490**, 187 (2012).

(1)

Editorial by Marcia McNutt Editor-in-chief
"Science" 17 January 2014

Emphasis: Reproducibity!

Also: Statistics more generally

EDITORIAL

Reproducibility

Marcia McNutt is Editor-in-Chief of *Science*.

SCIENCE ADVANCES ON A FOUNDATION OF TRUSTED DISCOVERIES. REPRODUCING AN EXPERIMENT is one important approach that scientists use to gain confidence in their conclusions. Recently, the scientific community was shaken by reports that a troubling proportion of peer-reviewed preclinical studies are not reproducible. Because confidence in results is of paramount importance to the broad scientific community, we are announcing new initiatives to increase confidence in the studies published in *Science*. For preclinical studies (one of the targets of recent concern), we will be adopting recommendations of the U.S. National Institute of Neurological Disorders and Stroke (NINDS) for increasing transparency.* Authors will indicate whether there was a pre-experimental plan for data handling (such as how to deal with outliers), whether they conducted a sample size estimation to ensure a sufficient signal-to-noise ratio, whether samples were treated randomly, and whether the experimenter was blind to the conduct of the experiment. These criteria will be included in our author guidelines.

There are a number of reasons why peer-reviewed preclinical studies may not be reproducible. The system under investigation may be more complex than previously thought, so that the experimenter is not actually controlling all independent variables. Authors may not have divulged all of the details of a complicated experiment, making it irreproducible by another lab. It is also expected that through random chance, a certain number of studies will produce false positives. If researchers are not alert to this possibility and have not set appropriately stringent significance tests for their results, the outcome is a study with irreproducible results. Although there is always the possibility that an occasional study is fraudulent, the number of preclinical studies that cannot be reproduced is inconsistent with the idea that all irreproducibility results from misconduct in such research.

It is unlikely that the issues with irreproducibility are confined to preclinical studies (social science has been equally noted, for example). Unfortunately, there are no equivalents to the NINDS recommendations for other disciplines that provide a basis for requiring transparency across all fields. For the next 6 months, we will be asking reviewers and editors to identify papers submitted to *Science* that demonstrate excellence in transparency and instill confidence in the results. This will inform the next steps in implementing reproducibility guidelines. *Science Translational Medicine*, a sister journal of *Science*, already enforces the NINDS guidelines for preclinical studies. Both journals also are open to improving on the NINDS recommendations for preclinical studies.

There is also a wide range of sophistication in the application of statistics displayed in research analysis, ranging from practically no statistics, to the routine use of generic software packages, to the application of advanced methods that extract subtle signals from noise. Because reviewers who are chosen for their expertise in the subject matter of a study may not be authorities in statistics as well, statistical errors in manuscripts may slip through undetected. For that reason, with the advice of the American Statistical Association and others, we are adding new members to our Board of Reviewing Editors from the statistics community to ensure that manuscripts receive appropriate scrutiny in their methods of data analysis.

Science's standards have always been high, and these measures add to steps we have already taken to increase transparency, such as requiring data accessibility. Nevertheless, journals can only do so much to assure readers of the validity of the studies they publish. The ultimate responsibility lies with authors to be completely open with their methods, all of their findings, and the possible pitfalls that could invalidate their conclusions.

– Marcia McNutt

10.1126/science.1250475

*S. C. Landis *et al.*, *Nature* 490, 187 (2012).

1

Editorial by Marcia McNutt Editor-in-chief
"Science" 17 January 2014

Emphasis: Reproducibity!

Also: Statistics more generally

Low key ...

EDITORIAL

Reproducibility

Marcia McNutt is Editor-in-Chief of *Science*.

SCIENCE ADVANCES ON A FOUNDATION OF TRUSTED DISCOVERIES. REPRODUCING AN EXPERIMENT is one important approach that scientists use to gain confidence in their conclusions. Recently, the scientific community was shaken by reports that a troubling proportion of peer-reviewed preclinical studies are not reproducible. Because confidence in results is of paramount importance to the broad scientific community, we are announcing new initiatives to increase confidence in the studies published in *Science*. For preclinical studies (one of the targets of recent concern), we will be adopting recommendations of the U.S. National Institute of Neurological Disorders and Stroke (NINDS) for increasing transparency.* Authors will indicate whether there was a pre-experimental plan for data handling (such as how to deal with outliers), whether they conducted a sample size estimation to ensure a sufficient signal-to-noise ratio, whether samples were treated randomly, and whether the experimenter was blind to the conduct of the experiment. These criteria will be included in our author guidelines.

There are a number of reasons why peer-reviewed preclinical studies may not be reproducible. The system under investigation may be more complex than previously thought, so that the experimenter is not actually controlling all independent variables. Authors may not have divulged all of the details of a complicated experiment, making it irreproducible by another lab. It is also expected that through random chance, a certain number of studies will produce false positives. If researchers are not alert to this possibility and have not set appropriately stringent significance tests for their results, the outcome is a study with irreproducible results. Although there is always the possibility that an occasional study is fraudulent, the number of preclinical studies that cannot be reproduced is inconsistent with the idea that all irreproducibility results from misconduct in such research.

It is unlikely that the issues with irreproducibility are confined to preclinical studies (social science has been equally noted, for example). Unfortunately, there are no equivalents to the NINDS recommendations for other disciplines that provide a basis for requiring transparency across all fields. For the next 6 months, we will be asking reviewers and editors to identify papers submitted to *Science* that demonstrate excellence in transparency and instill confidence in the results. This will inform the next steps in implementing reproducibility guidelines. *Science Translational Medicine*, a sister journal of *Science*, already enforces the NINDS guidelines for preclinical studies. Both journals also are open to improving on the NINDS recommendations for preclinical studies.

There is also a wide range of sophistication in the application of statistics displayed in research analysis, ranging from practically no statistics, to the routine use of generic software packages, to the application of advanced methods that extract subtle signals from noise. Because reviewers who are chosen for their expertise in the subject matter of a study may not be authorities in statistics as well, statistical errors in manuscripts may slip through undetected. For that reason, with the advice of the American Statistical Association and others, we are adding new members to our Board of Reviewing Editors from the statistics community to ensure that manuscripts receive appropriate scrutiny in their methods of data analysis.

Science's standards have always been high, and these measures add to steps we have already taken to increase transparency, such as requiring data accessibility. Nevertheless, journals can only do so much to assure readers of the validity of the studies they publish. The ultimate responsibility lies with authors to be completely open with their methods, all of their findings, and the possible pitfalls that could invalidate their conclusions.

– Marcia McNutt

10.1126/science.1250475

*S. C. Landis *et al.*, *Nature* 490, 187 (2012).

1

Editorial by Marcia McNutt Editor-in-chief
"Science" 17 January 2014

Emphasis: Reproducibity!

Also: Statistics more generally

Low key ...

but from Statistics ... Ho Hum!
Indifference!
Just ... Experimental design

EDITORIAL

Reproducibility

Marcia McNutt is Editor-in-Chief of *Science*.

SCIENCE ADVANCES ON A FOUNDATION OF TRUSTED DISCOVERIES. REPRODUCING AN EXPERIMENT is one important approach that scientists use to gain confidence in their conclusions. Recently, the scientific community was shaken by reports that a troubling proportion of peer-reviewed preclinical studies are not reproducible. Because confidence in results is of paramount importance to the broad scientific community, we are announcing new initiatives to increase confidence in the studies published in *Science*. For preclinical studies (one of the targets of recent concern), we will be adopting recommendations of the U.S. National Institute of Neurological Disorders and Stroke (NINDS) for increasing transparency.* Authors will indicate whether there was a pre-experimental plan for data handling (such as how to deal with outliers), whether they conducted a sample size estimation to ensure a sufficient signal-to-noise ratio, whether samples were treated randomly, and whether the experimenter was blind to the conduct of the experiment. These criteria will be included in our author guidelines.

There are a number of reasons why peer-reviewed preclinical studies may not be reproducible. The system under investigation may be more complex than previously thought, so that the experimenter is not actually controlling all independent variables. Authors may not have divulged all of the details of a complicated experiment, making it irreproducible by another lab. It is also expected that through random chance, a certain number of studies will produce false positives. If researchers are not alert to this possibility and have not set appropriately stringent significance tests for their results, the outcome is a study with irreproducible results. Although there is always the possibility that an occasional study is fraudulent, the number of preclinical studies that cannot be reproduced is inconsistent with the idea that all irreproducibility results from misconduct in such research.

It is unlikely that the issues with irreproducibility are confined to preclinical studies (social science has been equally noted, for example). Unfortunately, there are no equivalents to the NINDS recommendations for other disciplines that provide a basis for requiring transparency across all fields. For the next 6 months, we will be asking reviewers and editors to identify papers submitted to *Science* that demonstrate excellence in transparency and instill confidence in the results. This will inform the next steps in implementing reproducibility guidelines. *Science Translational Medicine*, a sister journal of *Science*, already enforces the NINDS guidelines for preclinical studies. Both journals also are open to improving on the NINDS recommendations for preclinical studies.

There is also a wide range of sophistication in the application of statistics displayed in research analysis, ranging from practically no statistics, to the routine use of generic software packages, to the application of advanced methods that extract subtle signals from noise. Because reviewers who are chosen for their expertise in the subject matter of a study may not be authorities in statistics as well, statistical errors in manuscripts may slip through undetected. For that reason, with the advice of the American Statistical Association and others, we are adding new members to our Board of Reviewing Editors from the statistics community to ensure that manuscripts receive appropriate scrutiny in their methods of data analysis.

Science's standards have always been high, and these measures add to steps we have already taken to increase transparency, such as requiring data accessibility. Nevertheless, journals can only do so much to assure readers of the validity of the studies they publish. The ultimate responsibility lies with authors to be completely open with their methods, all of their findings, and the possible pitfalls that could invalidate their conclusions.

– **Marcia McNutt**

10.1126/science.1250475

*S. C. Landis *et al.*, *Nature* **490**, 187 (2012).

(1)

Editorial by Marcia McNutt Editor-in-chief
"Science" 17 January 2014

Emphasis: Reproducibity!

Also: Statistics more generally

Low key ...

but from Statistics ... Ho Hum!
Indifference ...
Just Experimental design

from Science : LHC $p < 3 \cdot 10^{-7}$ 5σ

EDITORIAL

Reproducibility

Marcia McNutt is Editor-in-Chief of *Science*.

SCIENCE ADVANCES ON A FOUNDATION OF TRUSTED DISCOVERIES. REPRODUCING AN EXPERIMENT is one important approach that scientists use to gain confidence in their conclusions. Recently, the scientific community was shaken by reports that a troubling proportion of peer-reviewed preclinical studies are not reproducible. Because confidence in results is of paramount importance to the broad scientific community, we are announcing new initiatives to increase confidence in the studies published in *Science*. For preclinical studies (one of the targets of recent concern), we will be adopting recommendations of the U.S. National Institute of Neurological Disorders and Stroke (NINDS) for increasing transparency.* Authors will indicate whether there was a pre-experimental plan for data handling (such as how to deal with outliers), whether they conducted a sample size estimation to ensure a sufficient signal-to-noise ratio, whether samples were treated randomly, and whether the experimenter was blind to the conduct of the experiment. These criteria will be included in our author guidelines.

There are a number of reasons why peer-reviewed preclinical studies may not be reproducible. The system under investigation may be more complex than previously thought, so that the experimenter is not actually controlling all independent variables. Authors may not have divulged all of the details of a complicated experiment, making it irreproducible by another lab. It is also expected that through random chance, a certain number of studies will produce false positives. If researchers are not alert to this possibility and have not set appropriately stringent significance tests for their results, the outcome is a study with irreproducible results. Although there is always the possibility that an occasional study is fraudulent, the number of preclinical studies that cannot be reproduced is inconsistent with the idea that all irreproducibility results from misconduct in such research.

It is unlikely that the issues with irreproducibility are confined to preclinical studies (social science has been equally noted, for example). Unfortunately, there are no equivalents to the NINDS recommendations for other disciplines that provide a basis for requiring transparency across all fields. For the next 6 months, we will be asking reviewers and editors to identify papers submitted to *Science* that demonstrate excellence in transparency and instill confidence in the results. This will inform the next steps in implementing reproducibility guidelines. *Science Translational Medicine*, a sister journal of *Science*, already enforces the NINDS guidelines for preclinical studies. Both journals also are open to improving on the NINDS recommendations for preclinical studies.

There is also a wide range of sophistication in the application of statistics displayed in research analysis, ranging from practically no statistics, to the routine use of generic software packages, to the application of advanced methods that extract subtle signals from noise. Because reviewers who are chosen for their expertise in the subject matter of a study may not be authorities in statistics as well, statistical errors in manuscripts may slip through undetected. For that reason, with the advice of the American Statistical Association and others, we are adding new members to our Board of Reviewing Editors from the statistics community to ensure that manuscripts receive appropriate scrutiny in their methods of data analysis.

Science's standards have always been high, and these measures add to steps we have already taken to increase transparency, such as requiring data accessibility. Nevertheless, journals can only do so much to assure readers of the validity of the studies they publish. The ultimate responsibility lies with authors to be completely open with their methods, all of their findings, and the possible pitfalls that could invalidate their conclusions.

– Marcia McNutt

10.1126/science.1250475

*S. C. Landis *et al.*, *Nature* **490**, 187 (2012).

(1)

Editorial by Marcia McNutt Editor-in-chief
"Science" 17 January 2014

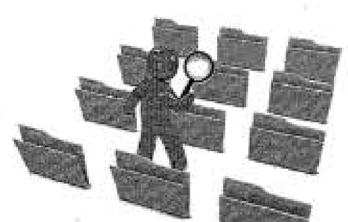
EDITORIAL

Reproducibility

Marcia McNutt is Editor-in-Chief of *Science*.

SCIENCE ADVANCES ON A FOUNDATION OF TRUSTED DISCOVERIES. REPRODUCING AN EXPERIMENT is one important approach that scientists use to gain confidence in their conclusions. Recently, the scientific community was shaken by reports that a troubling proportion of peer-reviewed preclinical studies are not reproducible. Because confidence in results is of paramount importance to the broad scientific community, we are announcing new initiatives to increase confidence in the studies published in *Science*. For preclinical studies (one of the targets of recent concern), we will be adopting recommendations of the U.S. National Institute of Neurological Disorders and Stroke (NINDS) for increasing transparency.* Authors will indicate whether there was a pre-experimental plan for data handling (such as how to deal with outliers), whether they conducted a sample size estimation to ensure a sufficient signal-to-noise ratio, whether samples were treated randomly, and whether the experimenter was blind to the conduct of the experiment. These criteria will be included in our author guidelines.

There are a number of reasons why peer-reviewed preclinical studies may not be reproducible. The system under investigation may be more complex than previously thought, so that the experimenter is not actually controlling all independent variables. Authors may not have divulged all of the details of a complicated experiment, making it irreproducible by another lab. It is also expected that through random chance, a certain number of studies will produce false positives. If researchers are not alert to this possibility and have not set appropriately stringent significance tests for their results, the outcome is a study with irreproducible results. Although there is always the possibility that an occasional study is fraudulent, the number of preclinical studies that cannot be reproduced is inconsistent with the idea that all irreproducibility results from misconduct in such research.



It is unlikely that the issues with irreproducibility are confined to preclinical studies (social science has been equally noted, for example). Unfortunately, there are no equivalents to the NINDS recommendations for other disciplines that provide a basis for requiring transparency across all fields. For the next 6 months, we will be asking reviewers and editors to identify papers submitted to *Science* that demonstrate excellence in transparency and instill confidence in the results. This will inform the next steps in implementing reproducibility guidelines. *Science Translational Medicine*, a sister journal of *Science*, already enforces the NINDS guidelines for preclinical studies. Both journals also are open to improving on the NINDS recommendations for preclinical studies.

There is also a wide range of sophistication in the application of statistics displayed in research analysis, ranging from practically no statistics, to the routine use of generic software packages, to the application of advanced methods that extract subtle signals from noise. Because reviewers who are chosen for their expertise in the subject matter of a study may not be authorities in statistics as well, statistical errors in manuscripts may slip through undetected. For that reason, with the advice of the American Statistical Association and others, we are adding new members to our Board of Reviewing Editors from the statistics community to ensure that manuscripts receive appropriate scrutiny in their methods of data analysis.

Science's standards have always been high, and these measures add to steps we have already taken to increase transparency, such as requiring data accessibility. Nevertheless, journals can only do so much to assure readers of the validity of the studies they publish. The ultimate responsibility lies with authors to be completely open with their methods, all of their findings, and the possible pitfalls that could invalidate their conclusions.

– Marcia McNutt

10.1126/science.1250475

*S. C. Landis *et al.*, *Nature* **490**, 187 (2012).

Emphasis: Reproducibity!

Also: Statistics more generally

Low key ...

but from Statistics ... Ho Hum!
Indifference ...
Just Experimental design

from Science: LHC $p < 3 \cdot 10^{-7}$ 5σ

But there is a lot of background ...
a) re "Data"
b) re "Retraction"
So, not as innocent or high principled as it seems ...

2a

The "Data" Dust-up

Science again:
11 February 2011

(2a)

The "Data" Dust-up

Science again:
11 February 2011

Full issue on Data
Early for 'Big Data'

(2a)

The "Data" Dust-up

Science again:
11 February 2011

Full issue on Data

Early for 'Big Data'

But look at word cloud

"Statistics" doesn't appear!

(2a)

The "Data" Dust-up

Science again:
11 February 2011

Full issue on Data
Early for 'Big Data'

But look at word cloud
"Statistics" doesn't appear!
(or does it? Some one did find it)

(2a)

The "Data" Dust-up

Science again:
11 February 2011

Full issue on Data

Early for 'Big Data'

But look at word cloud

"Statistics" doesn't appear!

(or does it? Some one did find it)

Two senior statisticians complained to Science/AAAS

2a

The "Data" Dust-up

Science again:
11 February 2011

Full issue on Data
Early for 'Big Data'

But look at word cloud
"Statistics" doesn't appear!
(or does it? Some one did find it)

Two senior statisticians
complained to Science/AAAS

A dismissive responce:
"You statisticians have an
Image problem!

2b "Retraction"

But the article on $\boxed{\text{Reproducibity}}$

(2b) "Retraction"

but the article on <u>Reproducibity</u>

also mentioned <u>Retraction</u>

No Journal likes to <u>retract</u> papers

2b "Retraction"

but the article on Reproducibity

also mentioned Retraction

No Journal likes to retract papers

and Science is no exception

(2b) "Retraction"

but the article on Reproducibity

also mentioned Retraction

No Journal likes to retract papers

and Science is no exception

So: they were rethinking their "dismissal of statistics"?

... they had their own Image problem!

(2b) "Retraction"

but the article on <u>Reproducibity</u>

also mentioned <u>Retraction</u>

No Journal likes to <u>retract</u> papers

and Science is no exception

So: they were rethinking their "dismissal of statistics"?

... they had their own <u>Image problem</u>!

... they were back-tracking fast!

(2b) "Retraction"

but the article on Reproducibity

also mentioned Retraction

No Journal likes to retract papers

and Science is no exception

So: they were rethinking their "dismissal of statistics"?

... they had their own Image problem!

... they were back-tracking fast!

.... on high principle!

(2b) "Retraction"

but the article on Reproducibity

also mentioned Retraction

No Journal likes to retract papers

and Science is no exception

So: they were rethinking their "dismissal of statistics"?

... they had their own Image problem!

... they were back-tracking fast!

.... on high principle!

But...

3a and in Statistics

... Statistics isn't immune to all of this!

(3a) and in Statistics

... Statistics isn't immune to all of this!

Statistics has <u>two</u> theories (Bayes / freq.) and they are <u>contradictory</u> | xxx = 266

(3a) and in Statistics

... Statistics isn't immune to all of this!

Statistics has <u>two</u> theories $\binom{\text{Bayes}}{\text{freq.}}$ and they are <u>contradicting</u> | xxx = 266

and no one cares ...

(3a) and in Statistics

... Statistics isn't immune to all this...

Statistics has <u>two</u> theories (Bayes / freq.) and they are <u>contradicting</u> | xxx = 266

and no one cares ...

"we are just exploring ..." John Doyle, G&M, April 1

(3a) and in Statistics

... Statistics isn't immune to all this...

Statistics has <u>two</u> theories (Bayes / freq.) and they are <u>contradicting</u> | xxx = 266

and no one cares ...

"we are just exploring ..."

but sometimes it's for "real"   LHC ; L'Aquila ; Vioxx

(3a) and in Statistics

... Statistics isn't immune to all this...

Statistics has <u>two</u> theories (Bayes / freq.) and they are <u>contradictory</u> | xxx = 266

and no one cares ...

"we are just exploring..."

but sometimes it's for "real" LHC ; <u>L'Aquila</u> ; <u>Vioxx</u>

Should results "mean what they say ?" Courts (legal) ?

3a and in Statistics

... Statistics isn't immune to all this...

Statistics has two theories (Bayes / freq.) and they are contradictory | xxx = 266

and no one cares ...

"we are just exploring..."

but sometimes it's for "real" LHC; L'Aquila; Vioxx

Should results "mean what they say?" Courts (legal)?

Efron Science 7 June 2013

MATHEMATICS

Bayes' Theorem in the 21st Century

Bayes' theorem plays an increasingly prominent role in statistical applications but remains controversial among statisticians.

Bradley Efron

The term "controversial theorem" sounds like an oxymoron, but Bayes' theorem has played this part for two-and-a-half centuries. Twice it has soared to scientific celebrity, twice it has crashed, and it is currently enjoying another boom. The theorem itself is a landmark of logical reasoning and the first serious triumph of statistical inference, yet is still treated with suspicion by most statisticians. There are reasons to believe in the staying power of its current popularity, but also some signs of trouble ahead.

Here is a simple but genuine example of Bayes' rule in action (see sidebar) (1). A physicist couple I know learned, from sonograms, that they were due to be parents of twin boys. They wondered what the probability was that their twins would be identical rather than fraternal. There are two pieces of relevant evidence. One-third of twins are identical; on the other hand, identical twins are twice as likely to yield twin boy sonograms, because they are always same-sex, whereas the likelihood of fraternal twins being same-sex is 50:50. Putting this together, Bayes' rule correctly concludes that the two pieces balance out, and that the odds of the twins being identical are even. (The twins were fraternal.)

Bayes' theorem is thus an algorithm for combining prior experience (one-third of twins are identicals) with current evidence (the sonogram). Followers of Nate Silver's FiveThirtyEight Web blog got to see the rule in spectacular form during the 2012 U.S. presidential campaign: The algorithm updated prior poll results with new data on a daily basis, correctly predicting the actual vote in all 50 states. "Statisticians beat pundits" was the verdict in the press (2).

Bayes' 1763 paper was an impeccable exercise in probability theory. The trouble and the subsequent busts came from overenthusiastic application of the theorem in the absence of genuine prior information, with Pierre-Simon Laplace as a prime violator. Suppose that in the twins example we lacked the prior knowledge that one-third of twins are identical. Laplace would have assumed a uniform distribution between zero and one for the unknown prior probability of identical twins, yielding 2/3 rather than 1/2 as the answer to the physicists' question. In modern parlance, Laplace would be trying to assign an "uninformative prior" or "objective prior" (2), one having only neutral effects on the output of Bayes' rule (3). Whether or not this

Department of Statistics, Stanford University, Stanford, CA 94305, USA. E-mail: brad@stat.stanford.edu

③a and in Statistics

... Statistics isn't immune to all this...

Statistics has two theories (Bayes / freq.) and they are contradicting | xxx = 266

and no one cares ...
"we are just exploring..."

but sometimes it's for "real" LHC ; L'Aquila ; Vioxx

Should results "mean what they say?" Courts (legal)?

Efron Science 7 June 2013

Classify priors $\pi(\theta)$

1) frequency empirical

2) mathematical "Pierre-Simon Laplace

3) opinion

MATHEMATICS

Bayes' Theorem in the 21st Century

Bayes' theorem plays an increasingly prominent role in statistical applications but remains controversial among statisticians.

Bradley Efron

The term "controversial theorem" sounds like an oxymoron, but Bayes' theorem has played this part for two-and-a-half centuries. Twice it has soared to scientific celebrity, twice it has crashed, and it is currently enjoying another boom. The theorem itself is a landmark of logical reasoning and the first serious triumph of statistical inference, yet is still treated with suspicion by most statisticians. There are reasons to believe in the staying power of its current popularity, but also some signs of trouble ahead.

Here is a simple but genuine example of Bayes' rule in action (see sidebar) (*1*). A physicist couple I know learned, from sonograms, that they were due to be parents of twin boys. They wondered what the probability was that their twins would be identical rather than fraternal. There are two pieces of relevant evidence. One-third of twins are identical; on the other hand, identical twins are twice as likely to yield twin boy sonograms, because they are always same-sex, whereas the likelihood of fraternal twins being same-sex is 50:50. Putting this together, Bayes' rule correctly concludes that the two pieces balance out, and that the odds of the twins being identical are even. (The twins were fraternal.)

Bayes' theorem is thus an algorithm for combining prior experience (one-third of twins are identicals) with current evidence (the sonogram). Followers of Nate Silver's FiveThirtyEight Web blog got to see the rule in spectacular form during the 2012 U.S. presidential campaign: The algorithm updated prior poll results with new data on a daily basis, correctly predicting the actual vote in all 50 states. "Statisticians beat pundits" was the verdict in the press (*2*).

Bayes' 1763 paper was an impeccable exercise in probability theory. The trouble and the subsequent busts came from overenthusiastic application of the theorem in the absence of genuine prior information, with Pierre-Simon Laplace as a prime violator. Suppose that in the twins example we lacked the prior knowledge that one-third of twins are identical. Laplace would have assumed a uniform distribution between zero and one for the unknown prior probability of identical twins, yielding 2/3 rather than 1/2 as the answer to the physicists' question. In modern parlance, Laplace would be trying to assign an "uninformative prior" or "objective prior" (*2*), one having only neutral effects on the output of Bayes' rule (*3*). Whether or not this

Department of Statistics, Stanford University, Stanford, CA 94305, USA. E-mail: brad@stat.stanford.edu

3a and in. Statistics

... Statistics isn't immune to all this...

Statistics has two theories ($\substack{\text{Bayes} \\ \text{freq.}}$) and they are contradicting | xxx = 266

and no one cares ...

"we are just exploring..."

but sometimes it's for "real" LHC ; L'Aquila ; Vioxx

Should results "mean what they say?" Courts (legal)?

MATHEMATICS

Bayes' Theorem in the 21st Century

Bayes' theorem plays an increasingly prominent role in statistical applications but remains controversial among statisticians.

Bradley Efron

The term "controversial theorem" sounds like an oxymoron, but Bayes' theorem has played this part for two-and-a-half centuries. Twice it has soared to scientific celebrity, twice it has crashed, and it is currently enjoying another boom. The theorem itself is a landmark of logical reasoning and the first serious triumph of statistical inference, yet is still treated with suspicion by most statisticians. There are reasons to believe in the staying power of its current popularity, but also some signs of trouble ahead.

Here is a simple but genuine example of Bayes' rule in action (see sidebar) (*1*). A physicist couple I know learned, from sonograms, that they were due to be parents of twin boys. They wondered what the probability was that their twins would be identical rather than fraternal. There are two pieces of relevant evidence. One-third of twins are identical; on the other hand, identical twins are twice as likely to yield twin boy sonograms, because they are always same-sex, whereas the likelihood of fraternal twins being same-sex is 50:50. Putting this together, Bayes' rule correctly concludes that the two pieces balance out, and that the odds of the twins being identical are even. (The twins were fraternal.)

Bayes' theorem is thus an algorithm for combining prior experience (one-third of twins are identicals) with current evidence (the sonogram). Followers of Nate Silver's FiveThirtyEight Web blog got to see the rule in spectacular form during the 2012 U.S. presidential campaign: The algorithm updated prior poll results with new data on a daily basis, correctly predicting the actual vote in all 50 states. "Statisticians beat pundits" was the verdict in the press (*2*).

Bayes' 1763 paper was an impeccable exercise in probability theory. The trouble and the subsequent busts came from overenthusiastic application of the theorem in the absence of genuine prior information, with Pierre-Simon Laplace as a prime violator. Suppose that in the twins example we lacked the prior knowledge that one-third of twins are identical. Laplace would have assumed a uniform distribution between zero and one for the unknown prior probability of identical twins, yielding 2/3 rather than 1/2 as the answer to the physicists' question. In modern parlance, Laplace would be trying to assign an "uninformative prior" or "objective prior" (*2*), one having only neutral effects on the output of Bayes' rule (*3*). Whether or not this

Department of Statistics, Stanford University, Stanford, CA 94305, USA. E-mail: brad@stat.stanford.edu

Efron Science 7 June 2013

Classify priors $\pi(\theta)$ "Values"

1) frequency empirical ← genuine

2) mathematical "Pierre-Simon Laplace ← "trouble"

3) opinion ← "trouble"

3a and in Statistics

... Statistics isn't immune to all this...

Statistics has two theories (Bayes / freq.) and they are contradicting | xxx = 266

and no one cares ...

"we are just exploring..."

but sometimes it's for "real"   LHC; L'Aquila; Vioxx

Should results "mean what they say?"   Courts (legal)?

MATHEMATICS

Bayes' Theorem in the 21st Century

Bayes' theorem plays an increasingly prominent role in statistical applications but remains controversial among statisticians.

Bradley Efron

The term "controversial theorem" sounds like an oxymoron, but Bayes' theorem has played this part for two-and-a-half centuries. Twice it has soared to scientific celebrity, twice it has crashed, and it is currently enjoying another boom. The theorem itself is a landmark of logical reasoning and the first serious triumph of statistical inference, yet is still treated with suspicion by most statisticians. There are reasons to believe in the staying power of its current popularity, but also some signs of trouble ahead.

Here is a simple but genuine example of Bayes' rule in action (see sidebar) (*1*). A physicist couple I know learned, from sonograms, that they were due to be parents of twin boys. They wondered what the probability was that their twins would be identical rather than fraternal. There are two pieces of relevant evidence. One-third of twins are identical; on the other hand, identical twins are twice as likely to yield twin boy sonograms, because they are always same-sex, whereas the likelihood of fraternal twins being same-sex is 50:50. Putting this together, Bayes' rule correctly concludes that the two pieces balance out, and that the odds of the twins being identical are even. (The twins were fraternal.)

Bayes' theorem is thus an algorithm for combining prior experience (one-third of twins are identicals) with current evidence (the sonogram). Followers of Nate Silver's FiveThirtyEight Web blog got to see the rule in spectacular form during the 2012 U.S. presidential campaign: The algorithm updated prior poll results with new data on a daily basis, correctly predicting the actual vote in all 50 states. "Statisticians beat pundits" was the verdict in the press (*2*).

Bayes' 1763 paper was an impeccable exercise in probability theory. The trouble and the subsequent busts came from overenthusiastic application of the theorem in the absence of genuine prior information, with Pierre-Simon Laplace as a prime violator. Suppose that in the twins example we lacked the prior knowledge that one-third of twins are identical. Laplace would have assumed a uniform distribution between zero and one for the unknown prior probability of identical twins, yielding 2/3 rather than 1/2 as the answer to the physicists' question. In modern parlance, Laplace would be trying to assign an "uninformative prior" or "objective prior" (*2*), one having only neutral effects on the output of Bayes' rule (*3*). Whether or not this

Department of Statistics, Stanford University, Stanford, CA 94305, USA. E-mail: brad@stat.stanford.edu

Efron Science 7 June 2013

Classify priors $\pi(\theta)$

1) frequency empirical ← genuine

2) mathematical "Pierre-Simon Laplace"   "trouble"

3) opinion   "trouble"

Support | Genuine priors (1)

Avoid | Laplace & Opinion
(2) (3)

(36) "Laplace had confidence": Science Sept 2013 Fraser

Low Marks for Education Funding Priorities

ANYONE INVOLVED SUBSTANTIVELY IN SCIence education during the past five decades will see the irony in the decision by the Office and Management and Budget (OMB) to trim the federal government's science, technology, engineering, and mathematics (STEM) programs on the grounds that many of them lack evaluation data on efficacy ("An invisible hand behind plan to realign U.S. science education," J. Mervis, News Focus, 26 July, p. 338). Although federal funding often supported formative evaluation (assessment in the pilot phase to improve the program itself) during the development of new curricula, it was virtually impossible to secure funding for summative evaluation (assessment of effectiveness after implementation) because of the costs and time frames involved. At the Biological Sciences Curriculum Study (*1*), where the value of summative evaluation always has been self-evident, we often lamented that the federal government funded a series of 90-meter dashes, supporting development of new instructional materials but not their evaluation. Funding from the Institute for Education Sciences for efficacy trials (*2*) that provide one type of summative evaluation constitutes some progress, but it is not enough.

It is perverse for OMB to blame STEM projects for deficiencies that were inherent in the government's funding priorities. Perhaps an evaluation of those priorities is in order.

JOSEPH D. MCINERNEY

Executive Vice President, American Society of Human Genetics, Bethesda, MD 20814, USA. E-mail: jmcinerney@ashg.org

References

1. Biological Sciences Curriculum Study (www.bscs.org).
2. J. K. Spybrook, S. W. Raudenbush, *Educ. Eval. Pol. Anal.* 31, 298 (2009).

171

Bayes' Confidence

NEITHER THE PERSPECTIVE "BAYES' THEOrem in the 21st century" (B. Efron, 7 June, p. 1177) nor the responding Letter "A statistically significant future for Bayes' rule" (R. van Hulst, 26 July, p. 343) refer to the mystical flavor often associated with Bayes in their discussions of the theorem's popularity.

Bayes had a propensity to use names that suggest something more than what is directly being described. For example, "Bayes' rule" is just conditional probability applied in a specialized context. The "controversial theorem" is nothing more than a formula for conditional probability.

Perhaps more disconcerting in Bayes is the term "objective prior" for the uninformative priors used by Pierre-Simon Laplace. Such priors, of course, are just imagined; they are not in fact objective themselves, but rather aim to produce objective conclusions. Indeed, many of Laplace's calculations of posterior probability using uninformative priors are numerically equal to frequentist calculations of confidence.

van Hulst mentions that the life sciences need a "synthesis of multiple categories of evidence." Certainly Bayes provides a simple and accessible means of combining different data results: Just multiply the likelihoods together. But this option is also available to the frequentist: Just combine the likelihoods and ignore what's left. The typical frequentist, however, realizes that this method would lose information and is unwilling to make this tradeoff for simplicity. Thus, he would choose an exact confidence interval when available.

Treating Bayes as a route to approximate confidence could go a long way toward resolving the presence of two theories in statistical inference.

D. A. S. FRASER

Department of Statistical Sciences, University of Toronto, Toronto, ON M5S3G3, Canada. E-mail: dfraser@utstat.toronto.edu

Research Funders Should Take the Field

WE SUPPORT EFFORTS TOWARD "LEVELING the playing field" (M. McNutt, Editorial, 26 July, p. 317) in science, technology, engineering, and mathematics (STEM) disciplines through organizations such as the Committee on Women in Science, Engineering, and Medicine (CWSEM). Targeting interventions at early career researchers is vital.

The Wellcome Trust's Basic Scientist Career Tracker (*1*) demonstrates the disproportionate number of women exiting academia early in their careers. Although an academic research career brings rewards, it remains a risky long-term career choice (*2*), and as McNutt describes, childbearing years typically coincide with the time when a faculty member needs to build a strong portfolio and gain tenure, thereby securing a less risky future.

Academia needs to attract and retain high-quality, highly trained researchers; research funders such as the Wellcome Trust can play an important role by following these steps: (i) Funders need to ensure that career awareness and mentorship are integral components of their training provision. (ii) Funders must ensure that their eligibility and/or funding guidelines do not discriminate against certain researchers (for example, a bias in funding decisions toward grant applications that include a move between institutions may inadvertently discriminate against those with established local ties). (iii) Funders need to promote and develop opportunities for researchers to use their funding flexibly, including options for career breaks, reentry fellowships, opportunities to work in posts other than as a principal investigator, and part-time schedules. (iv) We need to expand the opportunities for female role models working across academia to tell their story; this should be a core component of training programs.

ELIZABETH ALLEN,* HALINA SUWALOWSKA, DAVID LYNN

Strategic Planning and Policy Unit, Wellcome Trust, London, NW1 2BE, UK.

*Corresponding author. E-mail: l.allen@wellcome.ac.uk

Reference

1. Wellcome Trust, "Wellcome Trust Basic Science Career Tracker: Results of Wave 4 (2012)" (2013); www.wellcome.ac.uk/Funding/Biomedical-science/Career-tracker/Basic-tracker/index.htm
2. Ipsos MORI, "Risks and rewards: How PhD students choose their careers" (Ipsos MORI, London, 2013).

CORRECTIONS AND CLARIFICATIONS

This Week in *Science*: "Pushy black hole" (6 September, p. 1041). The last line should be "possibly limiting star formation and galaxy growth" instead of "possibly contributing to star formation and galaxy growth." The HTML and PDF versions online have been corrected.

Reports: "Pandoraviruses: Amoeba viruses with genomes up to 2.5 Mb reaching that of parasitic eukaryotes" by N. Philippe *et al.* (19 July, p. 281). In the first sentence of the legend to Fig. 1, the "(1)" and "(2)" should not have been italicized, as they refer to panels A1/A2 and B1/B2 and not to references 1 and 2. In the legend to Fig. 1E, the "a" and "b" labels should have been transposed. In addition, a reference to panels B1 and B2 is now included. In the acknowledgments, the GenBank accession numbers were incorrectly listed. They should read KC977571 and KC977570 (not KC977471 and KC977470). Also, the financial support of the Provence-Côte-d'Azur Région was missing. The HTML and PDF versions online have been corrected.

Letters to the Editor

Letters (~300 words) discuss material published in *Science* in the past 3 months or matters of general interest. Letters are not acknowledged upon receipt. Whether published in full or in part, Letters are subject to editing for clarity and space. Letters submitted, published, or posted elsewhere, in print or online, will be disqualified. To submit a Letter, go to www.submit2science.org.

(3b) Laplace had confidence : Science Sept 2013 Fraser

Bayes' Confidence

NEITHER THE PERSPECTIVE "BAYES' THEOREM in the 21st century" (B. Efron, 7 June, p. 1177) nor the responding Letter "A statistically significant future for Bayes' rule" (R. van Hulst, 26 July, p. 343) refer to the mystical flavor often associated with Bayes in their discussions of the theorem's popularity.

Bayes had a propensity to use names that suggest something more than what is directly being described. For example, "Bayes' rule" is just conditional probability applied in a specialized context. The "controversial theorem" is nothing more than a formula for conditional probability.

Perhaps more disconcerting in Bayes is the term "objective prior" for the uninformative priors used by Pierre-Simon Laplace. Such priors, of course, are just imagined; they are not in fact objective themselves, but rather aim to produce objective conclusions. Indeed, many of Laplace's calculations of posterior probability using uninformative priors are numerically equal to frequentist calculations of confidence.

van Hulst mentions that the life sciences need a "synthesis of multiple categories of evidence." Certainly Bayes provides a simple and accessible means of combining different data results: Just multiply the likelihoods together. But this option is also available to the frequentist: Just combine the likelihoods and ignore what's left. The typical frequentist, however, realizes that this method would lose information and is unwilling to make this tradeoff for simplicity. Thus, he would choose an exact confidence interval when available.

Treating Bayes as a route to approximate confidence could go a long way toward resolving the presence of two theories in statistical inference.

D. A. S. FRASER

Department of Statistical Sciences, University of Toronto, Toronto, ON M5S3G3, Canada. E-mail: dfraser@utstat.toronto.edu

Recall:

Efron in Science

Priors $\pi(\theta)$	Efron
1) frequency empirical	Genuine
2) mathematical "Pierre-Simon Laplace"	"trouble"
3) opinion	"trouble"

(3b) Laplace had confidence : Science Sept 2013 Fraser

Bayes' Confidence

NEITHER THE PERSPECTIVE "BAYES' THEOREM in the 21st century" (B. Efron, 7 June, p. 1177) nor the responding Letter "A statistically significant future for Bayes' rule" (R. van Hulst, 26 July, p. 343) refer to the mystical flavor often associated with Bayes in their discussions of the theorem's popularity.

Bayes had a propensity to use names that suggest something more than what is directly being described. For example, "Bayes' rule" is just conditional probability applied in a specialized context. The "controversial theorem" is nothing more than a formula for conditional probability.

Perhaps more disconcerting in Bayes is the term "objective prior" for the uninformative priors used by Pierre-Simon Laplace. Such priors, of course, are just imagined; they are not in fact objective themselves, but rather aim to produce objective conclusions. Indeed, many of Laplace's calculations of posterior probability using uninformative priors are numerically equal to frequentist calculations of confidence.

van Hulst mentions that the life sciences need a "synthesis of multiple categories of evidence." Certainly Bayes provides a simple and accessible means of combining different data results: Just multiply the likelihoods together. But this option is also available to the frequentist: Just combine the likelihoods and ignore what's left. The typical frequentist, however, realizes that this method would lose information and is unwilling to make this tradeoff for simplicity. Thus, he would choose an exact confidence interval when available.

Treating Bayes as a route to approximate confidence could go a long way toward resolving the presence of two theories in statistical inference.

D. A. S. FRASER

Department of Statistical Sciences, University of Toronto, Toronto, ON M5S3G3, Canada. E-mail: dfraser@utstat.toronto.edu

Efron in Science → Efron; Here ↓

Priors $\pi(\theta)$	Efron	Here
1) frequency empirical	Genuine	genuine
2) mathematical "Pierre-Simon Laplace	"trouble"	
3) opinion	"trouble"	"trouble"

(3b) Laplace had confidence : Science Sept 2013 Fraser

Bayes' Confidence

NEITHER THE PERSPECTIVE "BAYES' THEOREM in the 21st century" (B. Efron, 7 June, p. 1177) nor the responding Letter "A statistically significant future for Bayes' rule" (R. van Hulst, 26 July, p. 343) refer to the mystical flavor often associated with Bayes in their discussions of the theorem's popularity.

Bayes had a propensity to use names that suggest something more than what is directly being described. For example, "Bayes' rule" is just conditional probability applied in a specialized context. The "controversial theorem" is nothing more than a formula for conditional probability.

Perhaps more disconcerting in Bayes is the term "objective prior" for the uninformative priors used by Pierre-Simon Laplace. Such priors, of course, are just imagined; they are not in fact objective themselves, but rather aim to produce objective conclusions. Indeed, many of Laplace's calculations of posterior probability using uninformative priors are numerically equal to frequentist calculations of confidence.

van Hulst mentions that the life sciences need a "synthesis of multiple categories of evidence." Certainly Bayes provides a simple and accessible means of combining different data results: Just multiply the likelihoods together. But this option is also available to the frequentist: Just combine the likelihoods and ignore what's left. The typical frequentist, however, realizes that this method would lose information and is unwilling to make this tradeoff for simplicity. Thus, he would choose an exact confidence interval when available.

Treating Bayes as a route to approximate confidence could go a long way toward resolving the presence of two theories in statistical inference.

D. A. S. FRASER

Department of Statistical Sciences, University of Toronto, Toronto, ON M5S3G3, Canada. E-mail: dfraser@utstat.toronto.edu

Efron in Science

Priors $\pi(\theta)$	Efron	Here	
1) frequency empirical	Genuine	genuine	✓
2) mathematical "Pierre-Simon Laplace	"trouble" ⇒	- Laplace (when reproducible) - otherwise "trouble"	✓ ✗
3) opinion	"trouble"	"trouble"	✓

(3b) Laplace had confidence : Science Sept 2013 Fraser

Bayes' Confidence

NEITHER THE PERSPECTIVE "BAYES' THEOREM in the 21st century" (B. Efron, 7 June, p. 1177) nor the responding Letter "A statistically significant future for Bayes' rule" (R. van Hulst, 26 July, p. 343) refer to the mystical flavor often associated with Bayes in their discussions of the theorem's popularity.

Bayes had a propensity to use names that suggest something more than what is directly being described. For example, "Bayes' rule" is just conditional probability applied in a specialized context. The "controversial theorem" is nothing more than a formula for conditional probability.

Perhaps more disconcerting in Bayes is the term "objective prior" for the uninformative priors used by Pierre-Simon Laplace. Such priors, of course, are just imagined; they are not in fact objective themselves, but rather aim to produce objective conclusions. Indeed, many of Laplace's calculations of posterior probability using uninformative priors are numerically equal to frequentist calculations of confidence.

van Hulst mentions that the life sciences need a "synthesis of multiple categories of evidence." Certainly Bayes provides a simple and accessible means of combining different data results: Just multiply the likelihoods together. But this option is also available to the frequentist: Just combine the likelihoods and ignore what's left. The typical frequentist, however, realizes that this method would lose information and is unwilling to make this tradeoff for simplicity. Thus, he would choose an exact confidence interval when available.

Treating Bayes as a route to approximate confidence could go a long way toward resolving the presence of two theories in statistical inference.

D. A. S. FRASER

Department of Statistical Sciences, University of Toronto, Toronto, ON M5S3G3, Canada. E-mail: dfraser@utstat.toronto.edu

Efron in Science

Priors $\pi(\theta)$	Efron	Here	
1) frequency empirical	Genuine	Genuine	✓
2) mathematical "Pierre-Simon Laplace	"trouble" ⇒	- Laplace (when reproducible) - otherwise "trouble"	✓ ✗
3) opinion	"trouble"	"trouble"	✓

Summary:

1) If you have opinion, let's hear it!
   but don't use it to analyze data! — otherwise "misconduct!
   Display it in parallel. Let user see both (opinion / analysis) separately

(3b) Laplace had confidence : Science Sept 2013 Fraser

Bayes' Confidence

NEITHER THE PERSPECTIVE "BAYES' THEOrem in the 21st century" (B. Efron, 7 June, p. 1177) nor the responding Letter "A statistically significant future for Bayes' rule" (R. van Hulst, 26 July, p. 343) refer to the mystical flavor often associated with Bayes in their discussions of the theorem's popularity.

Bayes had a propensity to use names that suggest something more than what is directly being described. For example, "Bayes' rule" is just conditional probability applied in a specialized context. The "controversial theorem" is nothing more than a formula for conditional probability.

Perhaps more disconcerting in Bayes is the term "objective prior" for the uninformative priors used by Pierre-Simon Laplace. Such priors, of course, are just imagined; they are not in fact objective themselves, but rather aim to produce objective conclusions. Indeed, many of Laplace's calculations of posterior probability using uninformative priors are numerically equal to frequentist calculations of confidence.

van Hulst mentions that the life sciences need a "synthesis of multiple categories of evidence." Certainly Bayes provides a simple and accessible means of combining different data results: Just multiply the likelihoods together. But this option is also available to the frequentist: Just combine the likelihoods and ignore what's left. The typical frequentist, however, realizes that this method would lose information and is unwilling to make this tradeoff for simplicity. Thus, he would choose an exact confidence interval when available.

Treating Bayes as a route to approximate confidence could go a long way toward resolving the presence of two theories in statistical inference.

D. A. S. FRASER

Department of Statistical Sciences, University of Toronto, Toronto, ON M5S3G3, Canada. E-mail: dfraser@utstat.toronto.edu

Efron in Science

Priors $\pi(\theta)$	Efron	Here	
1) frequency empirical	Genuine	Genuine	✓
2) mathematical "Pierre-Simon Laplace	"trouble" ⇒	- Laplace (when reproducible) - otherwise "trouble"	✓ ✗
3) opinion	"trouble"	"trouble"	✓

Summary:

1) If you have opinion, let's hear it!
but don't use it to analyze data! otherwise "misconduct!
Display it in parallel. Let user see both (opinion / analysis separately

2) Bayes mathematical "Laplace" | OK if reproducible (Laplace had confidence first)
| o/w o/w "trouble"

(3b) Laplace had confidence :   Science Sept 2013 Fraser

Bayes' Confidence

NEITHER THE PERSPECTIVE "BAYES' THEOrem in the 21st century" (B. Efron, 7 June, p. 1177) nor the responding Letter "A statistically significant future for Bayes' rule" (R. van Hulst, 26 July, p. 343) refer to the mystical flavor often associated with Bayes in their discussions of the theorem's popularity.

Bayes had a propensity to use names that suggest something more than what is directly being described. For example, "Bayes' rule" is just conditional probability applied in a specialized context. The "controversial theorem" is nothing more than a formula for conditional probability.

Perhaps more disconcerting in Bayes is the term "objective prior" for the uninformative priors used by Pierre-Simon Laplace. Such priors, of course, are just imagined; they are not in fact objective themselves, but rather aim to produce objective conclusions. Indeed, many of Laplace's calculations of posterior probability using uninformative priors are numerically equal to frequentist calculations of confidence.

van Hulst mentions that the life sciences need a "synthesis of multiple categories of evidence." Certainly Bayes provides a simple and accessible means of combining different data results: Just multiply the likelihoods together. But this option is also available to the frequentist: Just combine the likelihoods and ignore what's left. The typical frequentist, however, realizes that this method would lose information and is unwilling to make this tradeoff for simplicity. Thus, he would choose an exact confidence interval when available.

Treating Bayes as a route to approximate confidence could go a long way toward resolving the presence of two theories in statistical inference.

D. A. S. FRASER

Department of Statistical Sciences, University of Toronto, Toronto, ON M5S3G3, Canada. E-mail: dfraser@utstat.toronto.edu

Efron in Science

Priors $\pi(\theta)$	Efron	Here	
1) frequency empirical	Genuine	Genuine	✓
2) mathematical "Pierre-Simon Laplace	"trouble" ⇒	- Laplace (when reproducible) - otherwise "trouble"	✓ ✗
3) opinion	"trouble"	"trouble"	✓

Summary:

1) If you have opinion, let's hear it!
but don't use it to analyze data! ← otherwise "misconduct!
Display it in parallel. Let user see both (opinion / analysis separately

2) Bayes mathematical "Laplace" | OK if reproducible (Laplace had confidence first)
| o/w o/w "trouble"

3) Question: How to get reproducible Bayes? (confidence!)

(3b) Laplace had confidence : Science Sept 2013 Fraser

Bayes' Confidence

NEITHER THE PERSPECTIVE "BAYES' THEOrem in the 21st century" (B. Efron, 7 June, p. 1177) nor the responding Letter "A statistically significant future for Bayes' rule" (R. van Hulst, 26 July, p. 343) refer to the mystical flavor often associated with Bayes in their discussions of the theorem's popularity.

Bayes had a propensity to use names that suggest something more than what is directly being described. For example, "Bayes' rule" is just conditional probability applied in a specialized context. The "controversial theorem" is nothing more than a formula for conditional probability.

Perhaps more disconcerting in Bayes is the term "objective prior" for the uninformative priors used by Pierre-Simon Laplace. Such priors, of course, are just imagined; they are not in fact objective themselves, but rather aim to produce objective conclusions. Indeed, many of Laplace's calculations of posterior probability using uninformative priors are numerically equal to frequentist calculations of confidence.

van Hulst mentions that the life sciences need a "synthesis of multiple categories of evidence." Certainly Bayes provides a simple and accessible means of combining different data results: Just multiply the likelihoods together. But this option is also available to the frequentist: Just combine the likelihoods and ignore what's left. The typical frequentist, however, realizes that this method would lose information and is unwilling to make this tradeoff for simplicity. Thus, he would choose an exact confidence interval when available.

Treating Bayes as a route to approximate confidence could go a long way toward resolving the presence of two theories in statistical inference.

D. A. S. FRASER

Department of Statistical Sciences, University of Toronto, Toronto, ON M5S3G3, Canada. E-mail: dfraser@utstat.toronto.edu

Efron in Science	Efron	Here	
Priors $\pi(\theta)$	Efron		
1) frequency empirical	Genuine	Genuine	✓
2) mathematical "Pierre-Simon Laplace"	"trouble" ⇒	- Laplace (when reproducible) - otherwise "trouble"	✓ ✗
3) opinion	"trouble"	"trouble"	✓

Summary:

1) If you have opinion, let's hear it!
but don't use it to analyze data! ← otherwise "misconduct!"
Display it in parallel. Let user see both (opinion / analysis separately)

2) Bayes mathematical "Laplace" | OK if reproducible (Laplace had confidence first) | o/w o/w "trouble"

3) Question: How to get reproducible Bayes? (confidence!)
How to get Objective (truly) Bayes

(3b) Laplace had confidence : Science Sept 2013 Fraser

Bayes' Confidence

NEITHER THE PERSPECTIVE "BAYES' THEOrem in the 21st century" (B. Efron, 7 June, p. 1177) nor the responding Letter "A statistically significant future for Bayes' rule" (R. van Hulst, 26 July, p. 343) refer to the mystical flavor often associated with Bayes in their discussions of the theorem's popularity.

Bayes had a propensity to use names that suggest something more than what is directly being described. For example, "Bayes' rule" is just conditional probability applied in a specialized context. The "controversial theorem" is nothing more than a formula for conditional probability.

Perhaps more disconcerting in Bayes is the term "objective prior" for the uninformative priors used by Pierre-Simon Laplace. Such priors, of course, are just imagined; they are not in fact objective themselves, but rather aim to produce objective conclusions. Indeed, many of Laplace's calculations of posterior probability using uninformative priors are numerically equal to frequentist calculations of confidence.

van Hulst mentions that the life sciences need a "synthesis of multiple categories of evidence." Certainly Bayes provides a simple and accessible means of combining different data results: Just multiply the likelihoods together. But this option is also available to the frequentist: Just combine the likelihoods and ignore what's left. The typical frequentist, however, realizes that this method would lose information and is unwilling to make this tradeoff for simplicity. Thus, he would choose an exact confidence interval when available.

Treating Bayes as a route to approximate confidence could go a long way toward resolving the presence of two theories in statistical inference.

D. A. S. FRASER

Department of Statistical Sciences, University of Toronto, Toronto, ON M5S3G3, Canada. E-mail: dfraser@utstat.toronto.edu

Efron in Science ↓ Here ↓

Priors $\pi(\theta)$	Efron	Here
1) frequency empirical	Genuine	Genuine
2) mathematical "Pierre-Simon Laplace	"trouble" ⇒	- Laplace (when reproducible) - otherwise "trouble"
3) opinion	"trouble"	"trouble"

Summary:

1) If you have opinion, let's hear it!
but don't use it to analyze data! ← otherwise "misconduct!
Display it in parallel. Let user see both (opinion / analysis separately

2) Bayes mathematical "Laplace" | OK if reproducible (Laplace had confidence first) | o/w o/w "trouble"

3) Question: How to get reproducible Bayes? (confidence!)

How to get Objective (truly) Bayes

(Term is already in use "without reproducibility" in B-community)

(4a) Can Bayes/Jeffreys give <u>reproducible</u> inference ?

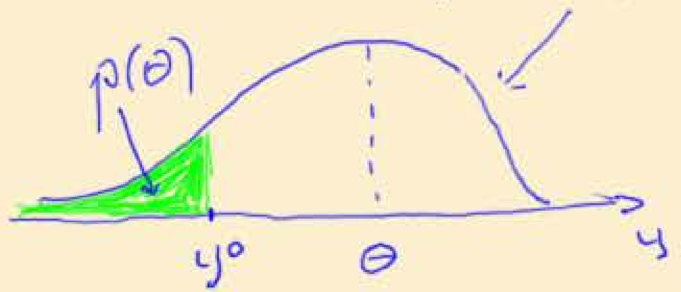
Try: Scalar $f(y-\theta)$ .... very simple case

④a Can Bayes/Jeffreys give reproducible inference ?

Try: Location, Scalar $f(y-\theta)$

1 The Case for Bayes

a) Motivating example $f(y-\theta)$: Data y^0 ; assess θ



Observed p-value $= p^0(\theta) = \int_{-\infty}^{y^0} f(y-\theta)\,dy$

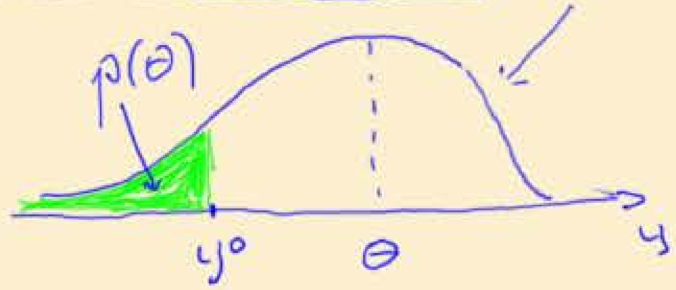
$=$ %age position of y^0 re θ $= F^0(\theta)$

(4a) Can Bayes/Jeffreys give <u>reproducible</u> inference?

Try: Location, Scalar $f(y-\theta)$

1 The Case for Bayes - Location

a) Motivating example $f(y-\theta)$: Data y^0; assess θ

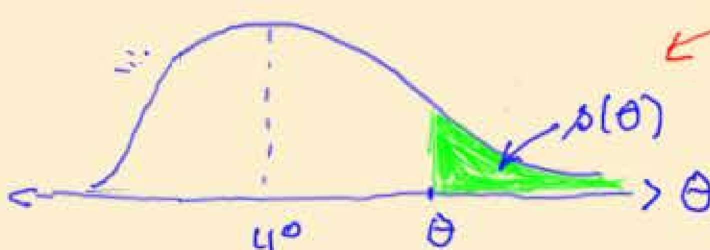


Observed p-value $= p^0(\theta) = \int_{-\infty}^{y^0} f(y-\theta)\,dy$

$=$ % age position of y^0 re θ

b) Bayes (Location/flat prior): pdf at y^0; flipped

$\pi(\theta \mid y^0) = 1 \cdot f(y^0-\theta)$



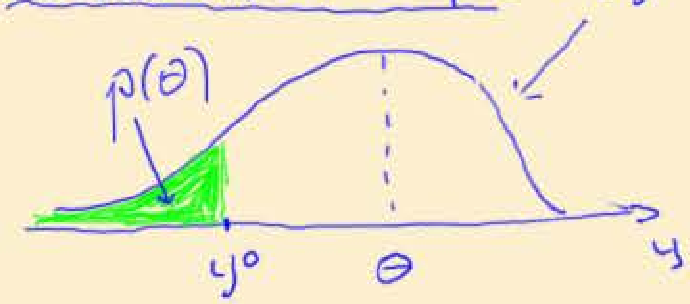
Bayes survivor $= s^0(\theta) = \int_{\theta}^{\infty} 1 \cdot f(y^0-\theta)\,d\theta$

(4a) Can Bayes/Jeffreys give reproducible inference ?

Try: Location, Scalar $f(y-\theta)$ Bayes give confidence

1 The Case for Bayes - Location

a) Motivating example $f(y-\theta)$: Data y^0 ; assess θ

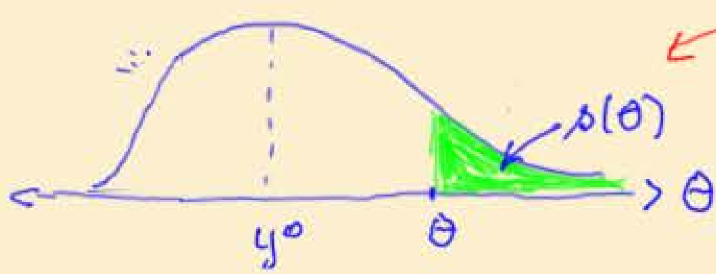


Observed p-value $= p^0(\theta) = \int_{-\infty}^{y^0} f(y-\theta)\,dy$

$=$ %age position of y^0 re θ

b) Bayes (Location/flat prior): pdf at y^0; flipped

$\pi(\theta | y^0) = 1 \cdot f(y^0-\theta)$



Bayes survivor $= s^0(\theta) = \int_{\theta}^{\infty} 1\, f(y^0-\theta)\,d\theta$

c) Thus: $p(\theta) = s(\theta)$ ... "Reflection" Just a "calculus recalculation"

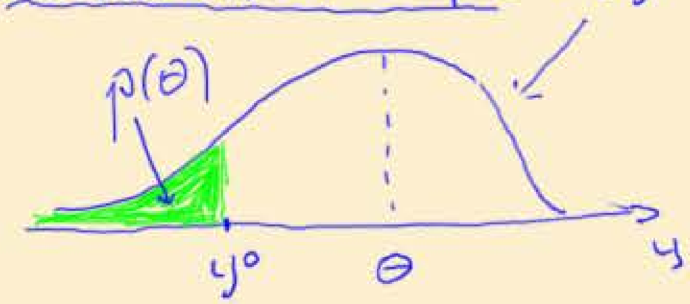
Bayes here gives confidence !

④a Can Bayes/Jeffreys give <u>reproducible</u> inference ?

Try: Location, Scalar $f(y-\theta)$ Bayes give confidence

1 The Case for Bayes - Location

a) Motivating example $f(y-\theta)$: Data y^0 ; assess θ

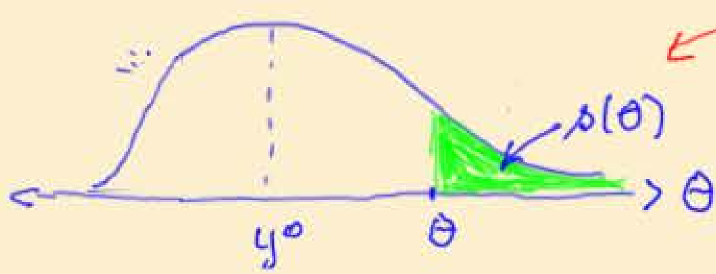


Observed p-value $= p^0(\theta) = \int_{-\infty}^{y^0} f(y-\theta)\,dy$

$=$ %age position of y^0 re θ

b) Bayes (Location/flat prior): pdf at y^0; flipped

$\pi(\theta \mid y^0) = 1 \cdot f(y^0-\theta)$



Bayes survivor $= s^0(\theta) = \int_{\theta}^{\infty} 1\, f(y^0-\theta)\,d\theta$

c) Thus: $p(\theta) = s(\theta)$ ... "Reflection" Just a calculus recalculation

Bayes here gives confidence

Does this generalize ?

4b Scalar parameter regular ... $f(y;\theta)$

(4b) Scalar parameter regular ... $f(y;\theta)$

1) Can always be rewritten as exponential (3rd order)

$$f(y;\theta) = \exp\{\varphi \Delta - k(\varphi)\} h(y)$$

Likelihood, asymptotics $O(n^{-3/2})$

4b Scalar parameter regular ... $f(y;\theta)$

1) Can always be rewritten as exponential (3rd order)

$$f(y;\theta) = \exp\{\varphi s - k(\varphi)\}\, h(y)$$

2) Can always be standardized $O(n^{-1})$

$$f(s;\varphi) = \exp\{\varphi s - \varphi^2/2 \quad - \gamma\, \varphi^3/6n^{1/2}\}\, h(s)$$ Taylor 2nd order

4b) Scalar parameter regular ... $f(y;\theta)$

1) Can always be rewritten as exponential (3rd order)

$$f(y;\theta) = \exp\{\varphi s - k(\varphi)\}\, h(y)$$

2) Can always be standardized

$$f(s;\varphi) = \exp\left\{\varphi s - \varphi^2/2 - \gamma \frac{\varphi^3}{6n^{1/2}}\right\} h(s)$$

$$= \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{(s-\varphi)^2}{2} - \gamma \varphi^3/6n^{1/2} + \underline{\gamma s^3/6n^{1/2}}\right\}(1 - \underline{\gamma s/2n^{1/2}})$$

Expanded

Determined by "pdf"

(4b) Regular, Scalar parameter $f(y;\theta)$

1) Can always be rewritten as exponential (3rd order)

$$f(y;\theta) = \exp\{\varphi s - k(\varphi)\}\, h(y)$$

2) Can always be standardized

$$f(s;\varphi) = \exp\{\varphi s - \varphi^2/2 \quad - \gamma \varphi^3/6n^{1/2}\}\, h(s)$$

$$= \frac{1}{\sqrt{2\pi}} \exp\{- \frac{(s-\varphi)}{2} - \gamma \varphi^3/6n^{1/2} + \gamma s^3/6n^{1/2}\}(1 - \gamma s/2n^{1/2})$$

3) Can use a <u>constant-info</u> parameter $\beta = \varphi + \gamma \varphi^2/2n^{1/2}$

(4b) Regular, Scalar parameter $f(y;\theta)$

1) Can always be rewritten as exponential (3rd order)

$$f(y;\theta) = \exp\{\varphi s - k(\varphi)\}\, h(y)$$

2) Can always be standardized

$$f(s;\varphi) = \exp\{\varphi s - \varphi^2/2 - \gamma\, \varphi^3/6n^{1/2}\}\, h(s)$$

$$= \frac{1}{\sqrt{2\pi}} \exp\{-\frac{(s-\varphi)}{2} - \gamma\, \varphi^3/6n^{1/2} + \gamma\, s^3/6n^{1/2}\}(1 - \gamma s/2n^{1/2})$$

3) Can use a constant-info parameter $\beta = \varphi + \gamma\, \varphi^2/2n^{1/2}$

$$f(\hat\beta;\beta) = \frac{1}{\sqrt{2\pi}} \exp\{\frac{(\hat\beta-\beta)^2}{2} - \gamma\,(\hat\beta-\beta)^3/6n^{1/2}\} \cdot d\hat\beta$$

Rewrite as <u>location</u>

(4b) Regular, Scalar parameter $f(y;\theta)$

1) Can always be rewritten as exponential (3rd order)

$$f(y;\theta) = \exp\{\varphi s - k(\varphi)\}\, h(y)$$

2) Can always be standardized

$$f(s;\varphi) = \exp\{\varphi s - \varphi^2/2 \quad - \gamma\, \varphi^3/6n^{1/2}\}\, h(s)$$

$$= \frac{1}{\sqrt{2\pi}} \exp\left\{- \frac{(s-\varphi)}{2} - \gamma\, \varphi^3/6n^{1/2} + \gamma\, s^3/6n^{1/2}\right\}(1 - \gamma s/2n^{1/2})$$

3) Can use a constant-info parameter $\beta = \varphi + \gamma\varphi^2/2n^{1/2}$

$$f(\hat\beta;\beta) = \frac{1}{\sqrt{2\pi}} \exp\left\{ \frac{(\hat\beta-\beta)^2}{2} - \gamma\,(\hat\beta-\beta)^3/6n^{1/2}\right\} \cdot d\hat\beta$$

4) Jeffreys "automatic" for location model: posterior is

$$f(\beta;\hat\beta) = \frac{1}{\sqrt{2\pi}} \exp\left\{ \frac{(\hat\beta-\beta)^2}{2} - \gamma\,(\hat\beta-\beta)^3/6n^{1/2}\right\} \cdot d\beta$$

(4b) Regular, Scalar parameter $f(y;\theta)$

1) Can always be rewritten as exponential (3rd order)

$$f(y;\theta) = \exp\{\varphi s - k(\varphi)\}\, h(y)$$

2) Can always be standardized

$$f(s;\varphi) = \exp\{\varphi s - \varphi^2/2 - \gamma\, \varphi^3/6n^{1/2}\}\, h(s)$$

$$= \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{(s-\varphi)}{2} - \gamma\varphi^3/6n^{1/2} + \gamma s^3/6n^{1/2}\right\}(1 - \gamma s/2n^{1/2})$$

3) Can use a constant-info parameter $\beta = \varphi + \varphi^2/2n^{1/2}$

$$f(\hat\beta;\beta) = \frac{1}{\sqrt{2\pi}} \exp\left\{\frac{(\hat\beta-\beta)^2}{2} - \gamma(\hat\beta-\beta)^3/6n^{1/2}\right\} \cdot d\hat\beta$$

4) Jeffreys automatic for location model

$$f(\beta;\hat\beta) = \frac{1}{\sqrt{2\pi}} \exp\left\{\frac{(\hat\beta-\beta)^2}{2} - \gamma(\hat\beta-\beta)^3/6n^{1/2}\right\} \cdot d\beta$$

5) Reexpress $d\beta = (1 + \gamma\varphi/n^{1/2})\, d\varphi$

$= \text{Likelihood} \cdot (\text{root info}) \cdot d\varphi$

Jeffreys

- Rewrite posterior differential
- 2nd order

$\Rightarrow$ pure confidence

4b Regular, Scalar parameter $f(y;\theta)$

1) Can always be rewritten as exponential (3rd order)

$$f(y;\theta) = \exp\{\varphi s - k(\varphi)\}\, h(y)$$

2) Can always be standardized $O(n^{-1})$

$$f(s;\varphi) = \exp\{\varphi s - \varphi^2/2 - \gamma \varphi^3/6n^{1/2}\}\, h(s)$$

$$= \frac{1}{\sqrt{2\pi}} \exp\{-\frac{(s-\varphi)}{2} - \gamma\varphi^3/6n^{1/2} + \gamma s^3/6n^{1/2}\}(1-\gamma s/2n^{1/2})$$

3) Can use a constant-info parameter $\beta = \varphi + \varphi^2/2n^{1/2}$

$$f(\hat\beta;\beta) = \frac{1}{\sqrt{2\pi}} \exp\{\frac{(\hat\beta-\beta)^2}{2} - \gamma(\hat\beta-\beta)^3/6n^{1/2}\} \cdot d\hat\beta$$

4) Jeffreys automatic for location model

$$f(\beta;\hat\beta) = \frac{1}{\sqrt{2\pi}} \exp\{\frac{(\hat\beta-\beta)^2}{2} - \gamma(\hat\beta-\beta)^3/6n^{1/2}\} \cdot d\beta$$

5) Reexpress $d\beta = (1 + \gamma\varphi/n^{1/2})\, d\varphi$

$= \text{Likelihood} \cdot (\text{root info}) \cdot d\varphi$

Jeffreys is 2nd order Accurate — Welch Peers 1963

is reproducible — Brown Cai DasGupta 2001 (binomial) | from Bayes to Jeffreys

Scalar parameter model :

Example 1

Setup

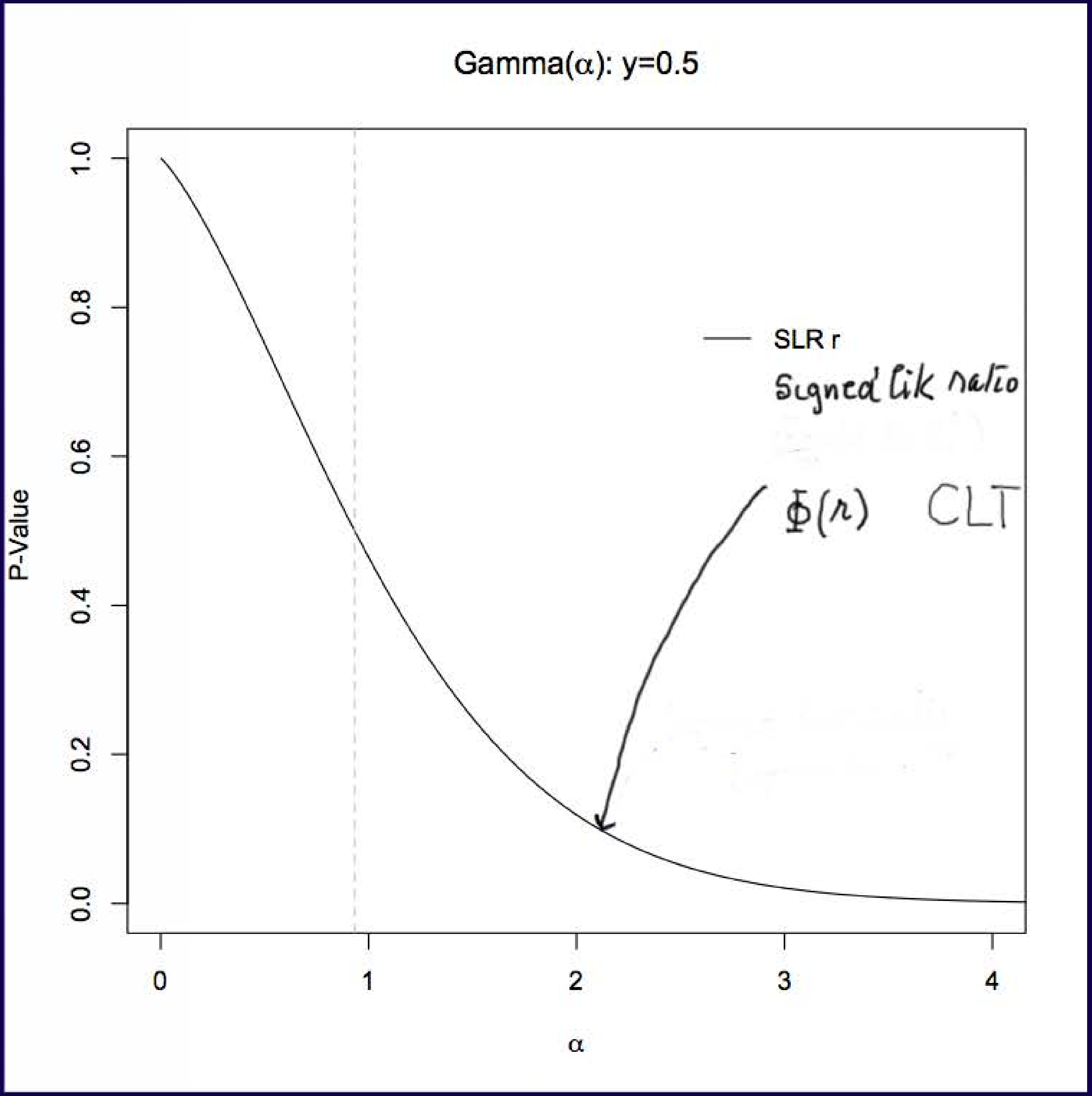
- Model: $Y \sim Gamma(\alpha, \beta)$ where α is shape and β is rate
  pdf is $\frac{\beta^\alpha}{\Gamma(\alpha)} y^{\alpha-1} e^{-\beta y}$
  Let $n = 1$ and data is $y^\circ = .5$

  Parameter of interest is α

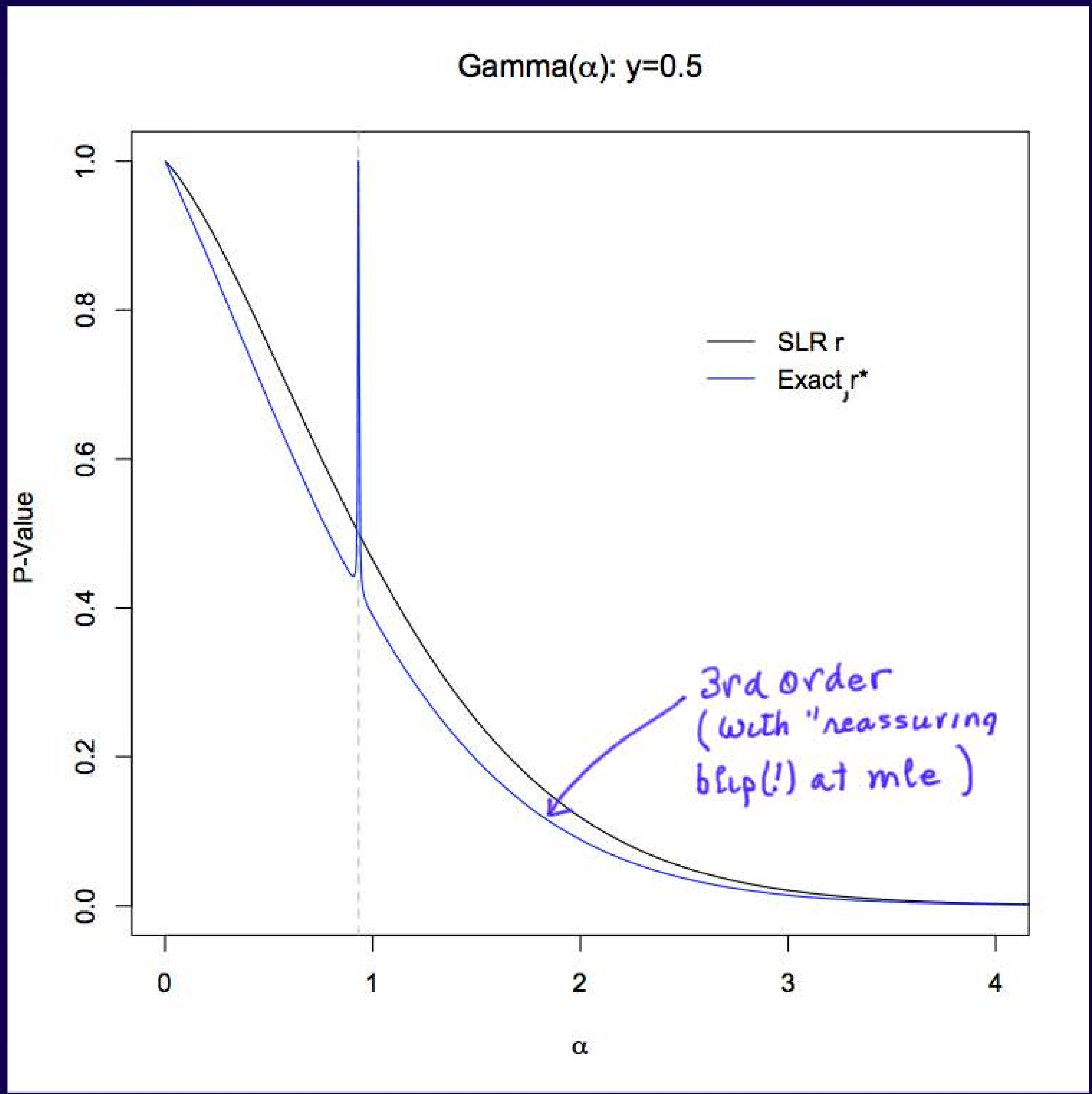
Fix $\beta = 1$

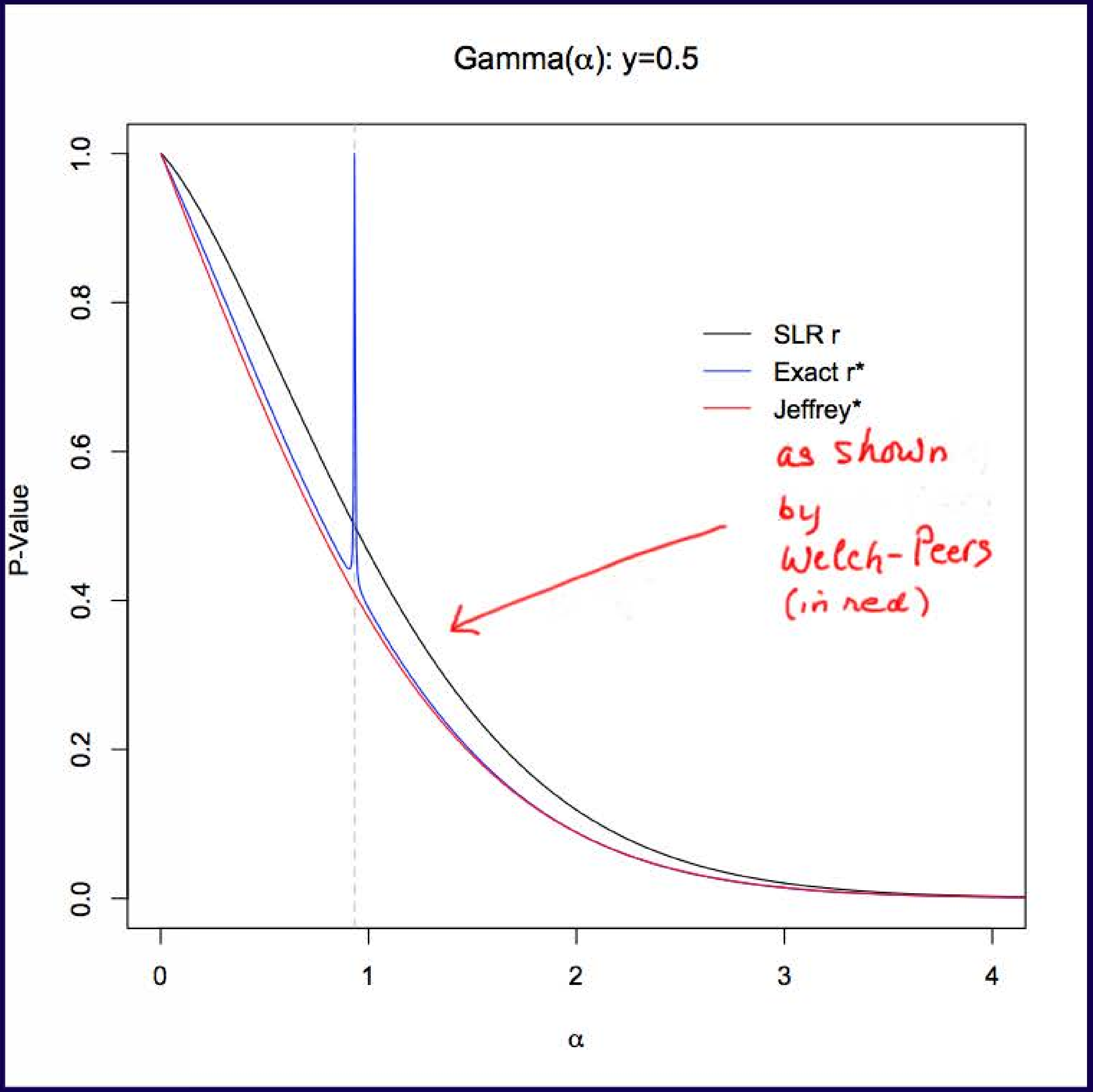
$\Gamma^{-1}(\alpha) y^{\alpha-1} e^{-y} \cdot dy$

$y^\circ = .5$



Gamma(α): y=0.5
P-Value
1.0
0.8
0.6
0.4
0.2
0.0
0
1
2
3
4
α
SLR r
Signed lik ratio
Φ(r) CLT


Gamma(α): y=0.5
P-Value
1.0
0.8
0.6
0.4
0.2
0.0
0
1
2
3
4
α
SLR r
Exact r*
3rd order
(with "reassuring
blip(!) at mle)



reproducible!

④c Regular Statistical model: $f(y;\varphi)$ φ canonical; $\psi(\varphi)$ = scalar interest

1) Can <u>always</u> be <u>re</u>-written as exponential model:

$$f(y;\varphi) = \exp\{\varphi' s - k(\varphi)\}\, h(y)$$

likelihood asymptotics
re-parameterization

2) Can <u>always</u> be separated: $f(s;\psi) \cdot g(t \mid s;\psi,\lambda)$

$$= f(s;\psi) \cdot \frac{|j_{(\lambda\lambda)}(\hat{\varphi}_\psi)|^{-1/2}}{(2\pi)^{\frac{p-1}{2}}} \exp\{\ell - \tilde{\ell}\} \quad \text{re } ds\,dt$$

Saddlepoint analysis

3) Prior to eliminate

2nd factor $\quad (2\pi)^{\frac{p-1}{2}} |j_{(\lambda\lambda)}(\hat{\varphi}_\psi)|^{1/2}$

3rd factor $\quad$ 1 on profile

direct multiplication

or SP integration of (λ) given ψ
re Exp model form for (λ)
over section fixed ψ

4) Prior to calibrate
via W-P 1st factor $\quad |j^{p}_{[\psi\psi]}(\psi)|^{1/2}$

Welch-Peers plus Lik. asy.

5) Combine (1st 2nd) $\quad |j_{\varphi\varphi}(\hat{\varphi}_\psi)|^{1/2} \dfrac{|j_{(\lambda\lambda)}(\hat{\varphi}_\psi)|^{1/2}}{|j_{[\lambda\lambda]}(\hat{\varphi}_\psi)|^{1/2}}$

d into re ψ at $\hat{\varphi}_\psi$
" re χ "

6) 3rd factor eliminates
λ by Laplace
(also by full Bayes)

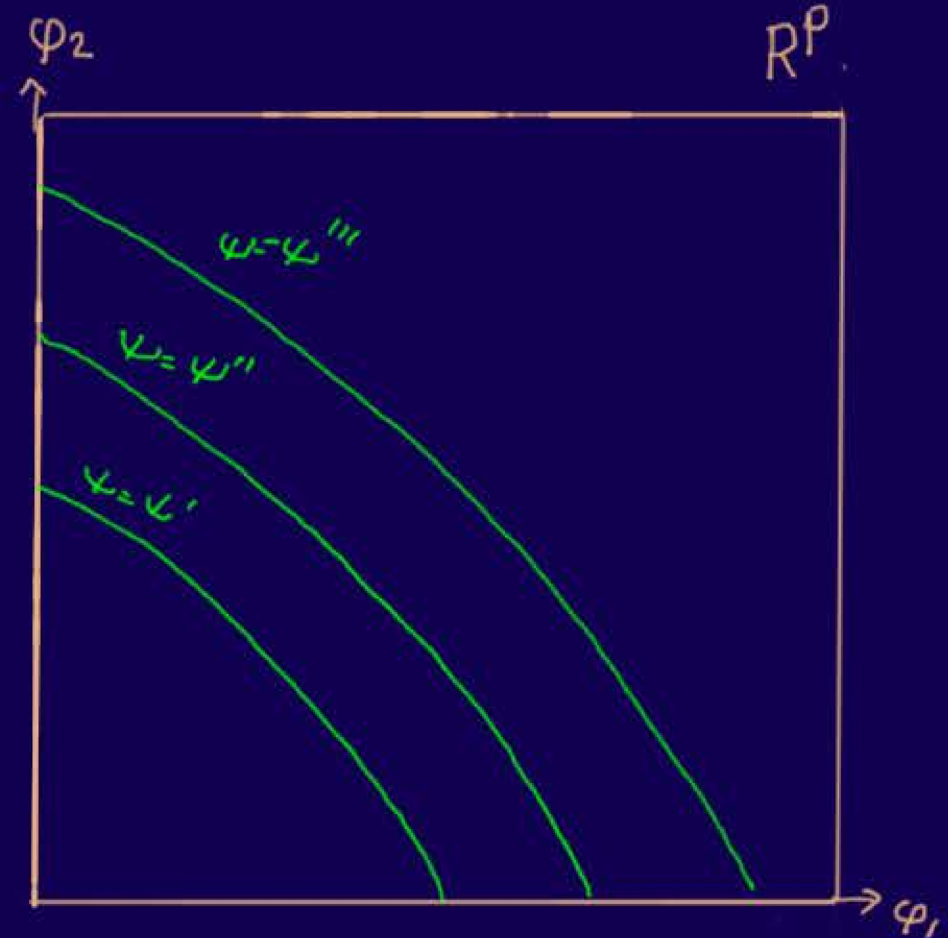
5) Gives <u>full</u> Jeffreys $|j_{\varphi\varphi}(\hat{\varphi}_\psi)|^{1/2}$ on profile (re ψ) + "curvature" correction

Cases: a) Linear ψ b) Rotating ψ c) Curved ψ & calculate $d(\psi)$

The geometry:

1 Parameter space (canonical) Φ

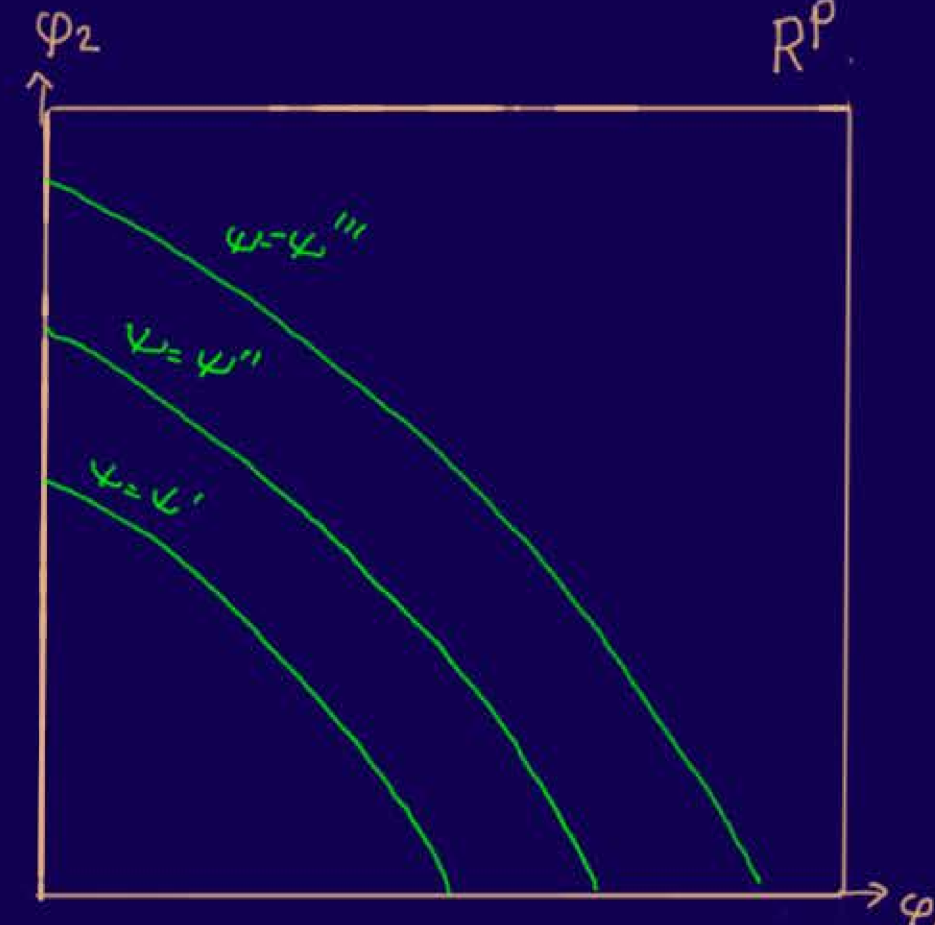
Interest $\psi(\varphi)$



The geometry:

1 Parameter space (canonical) Φ

Interest $\psi(\varphi)$



Jeffreys (usual)

Likelihood = $L^\circ(\varphi)$

J. prior = $\pi(\varphi) = \left| j_{\varphi\varphi}(\varphi) \right|^{1/2}$

Posterior = $L^\circ(\varphi)\,\pi(\varphi)$

& integrate up to contour " "

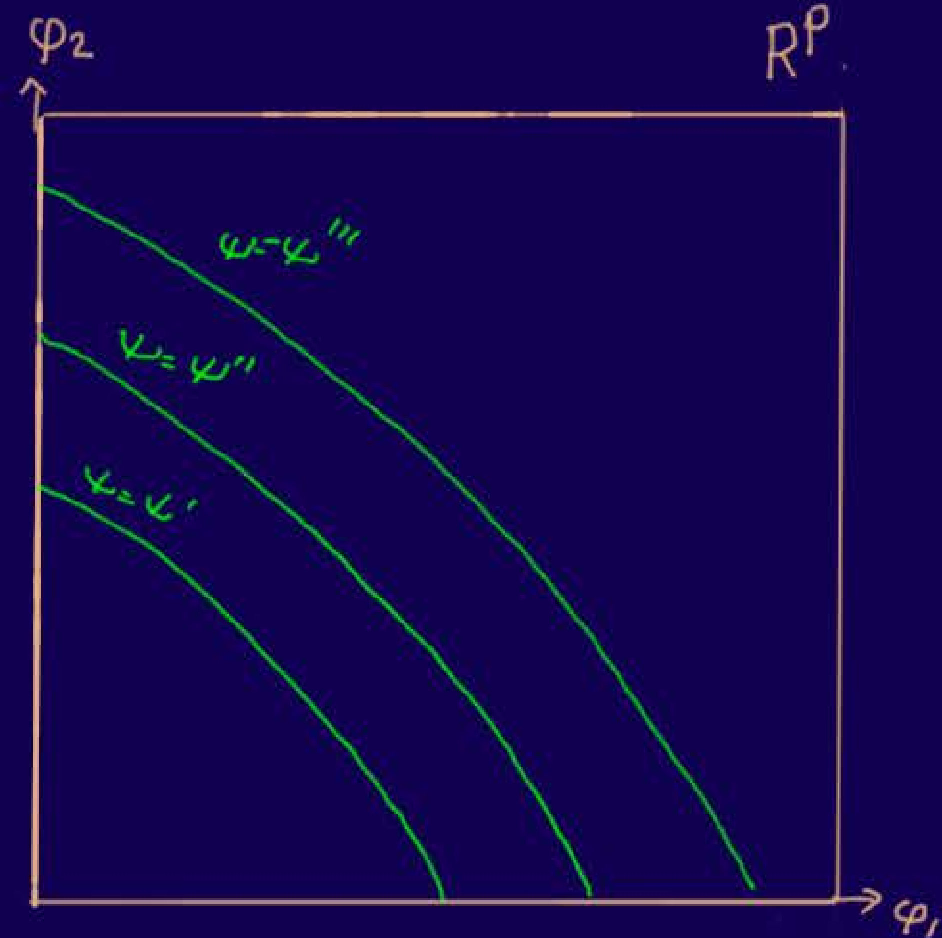
for dist'n of ψ

But...

The geometry:

1 Parameter space (canonical) Φ

Interest $\psi(\varphi)$



Jeffreys (usual)

Likelihood = $L^\circ(\varphi)$

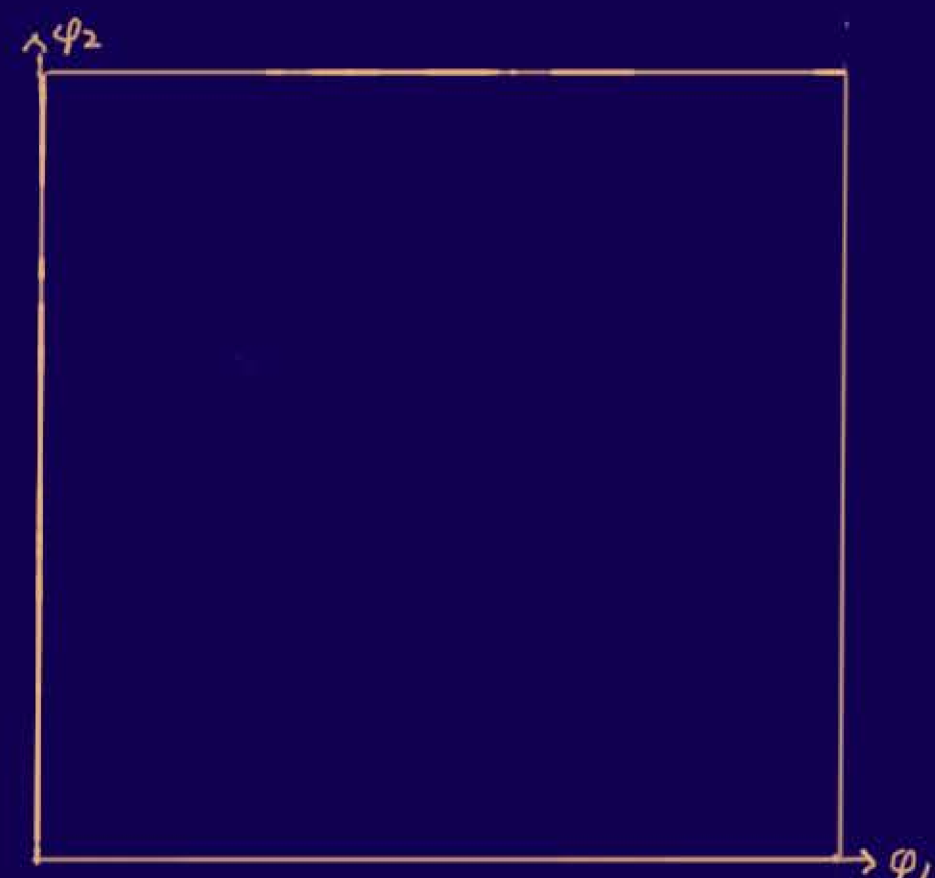
J. prior = $\pi(\varphi) = |j_{\varphi\varphi}(\varphi)|^{1/2}$

Posterior = $L^\circ(\varphi)\,\pi(\varphi)$

& integrate up to contour
for dist'n of ψ

2. Parameter space (Rotnly sym. re y°) $\bar{\Phi}$

Interest $\psi(\varphi)$

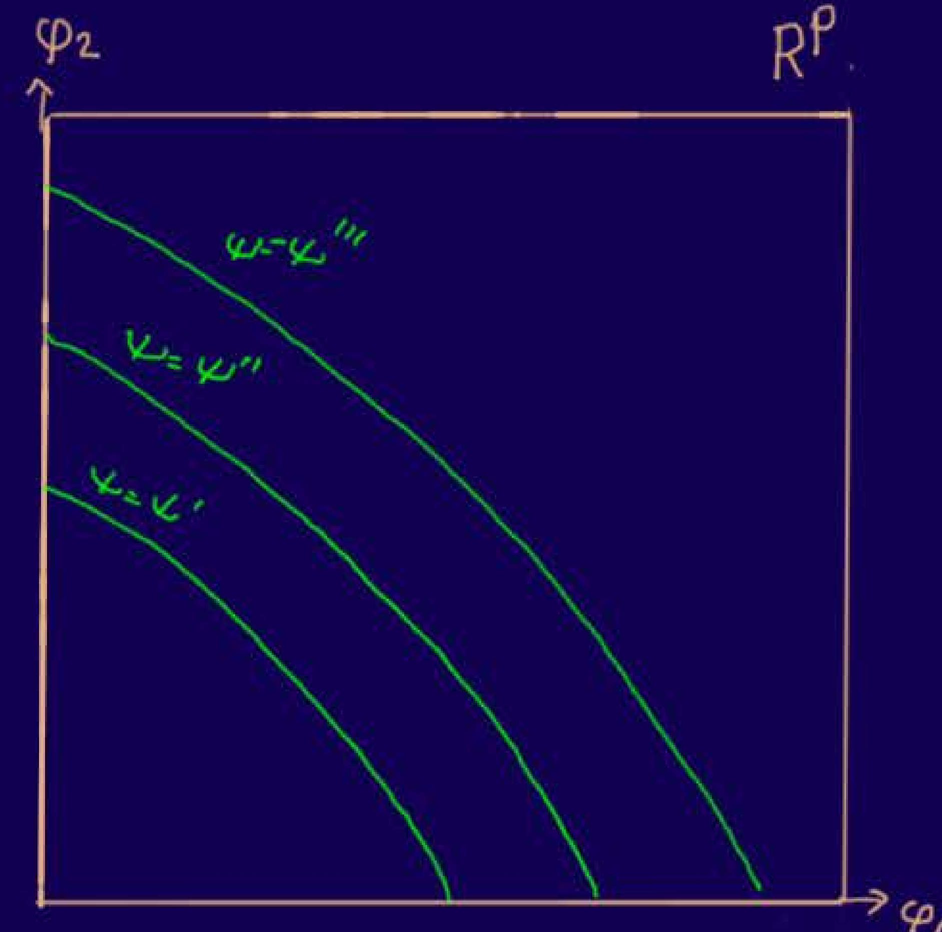


Accelerated Jeffreys: J^*

The geometry:

1 Parameter space (canonical) Φ

Interest $\psi(\varphi)$



<u>Jeffreys</u> (usual).

Likelihood = $L^\circ(\varphi)$

J. prior = $\pi(\varphi) = \left| j_{\varphi\varphi}(\varphi) \right|^{1/2}$

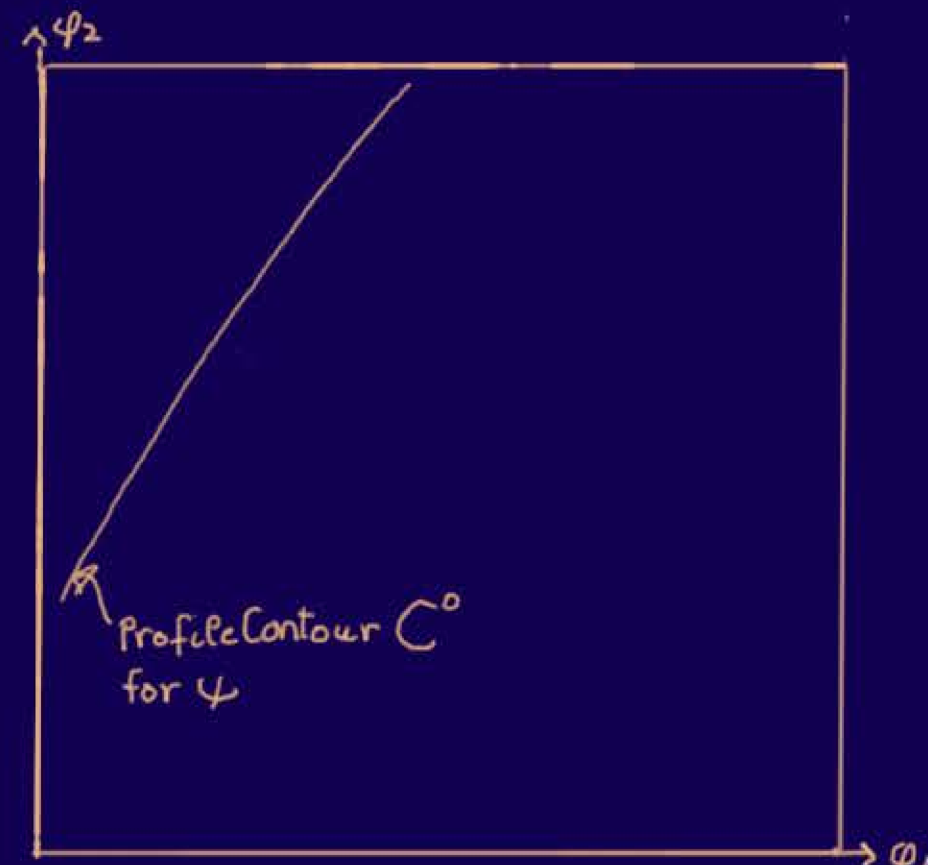
Posterior = $L^\circ(\varphi)\,\pi(\varphi)$

& integrate up to contour "$\setminus$"

for dist'n of ψ

2. Parameter space (Rotnly sym. re ψ°) Φ

Interest $\psi(\varphi)$



<u>Accelerated Jeffreys</u> J^*

Use <u>full</u> Jeffreys

<u>but just</u> on profile curve C° re ψ

Posterior = $L^\circ(\varphi)\,\pi_\psi(\varphi)$ on C°

(one dimensional)

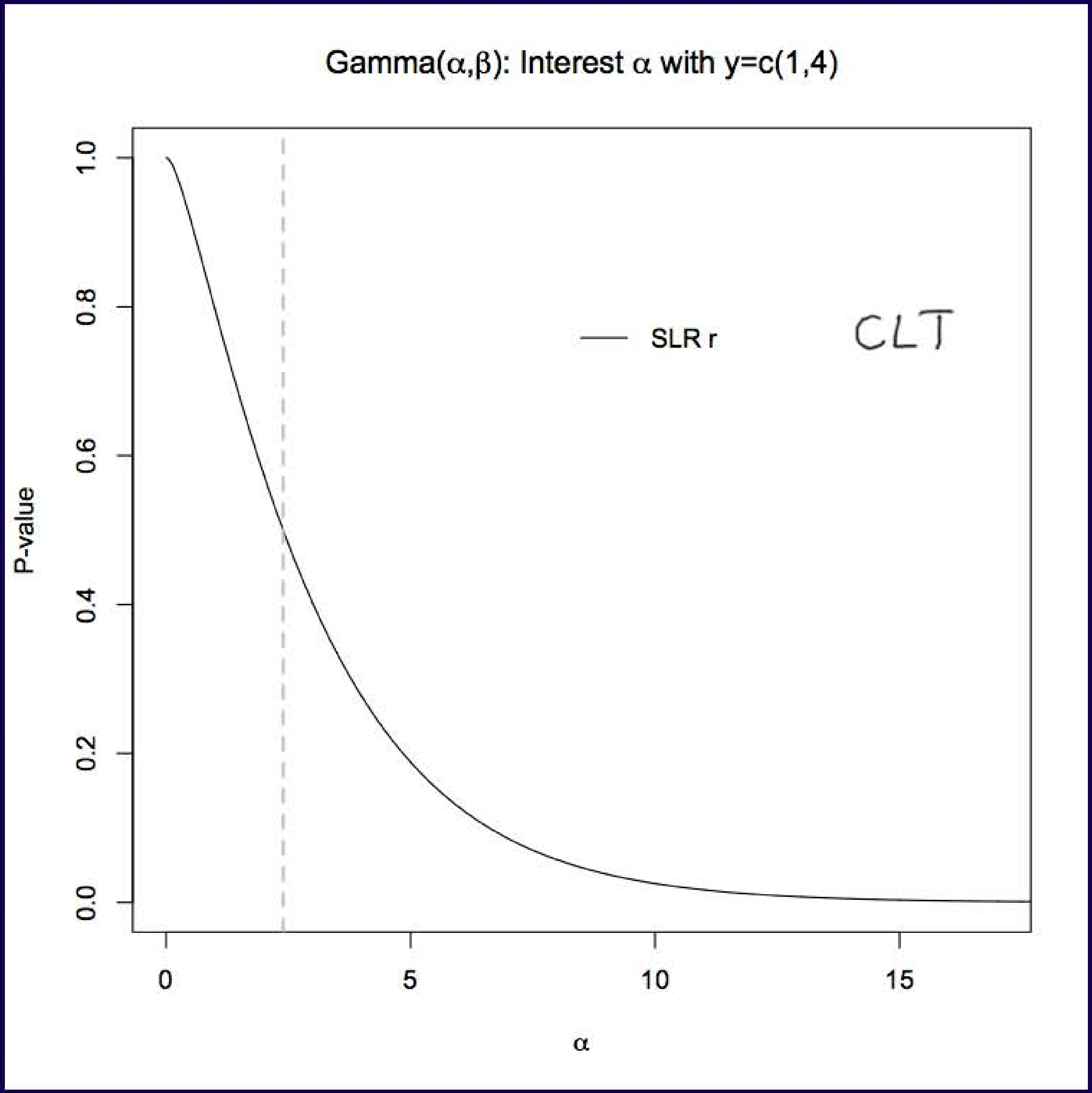
Vector parameter (α, β); Scalar linear interest α

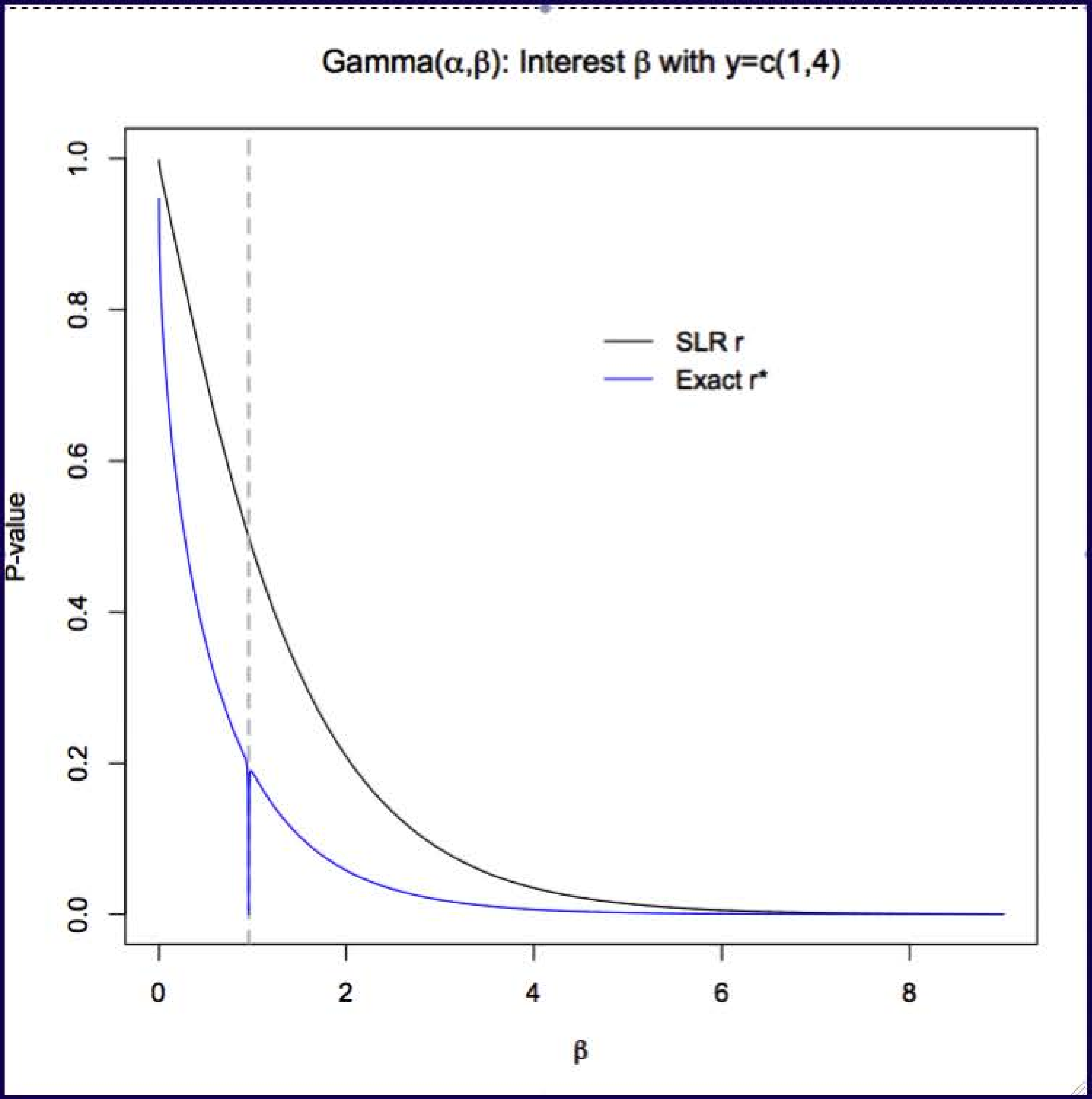
Example 2

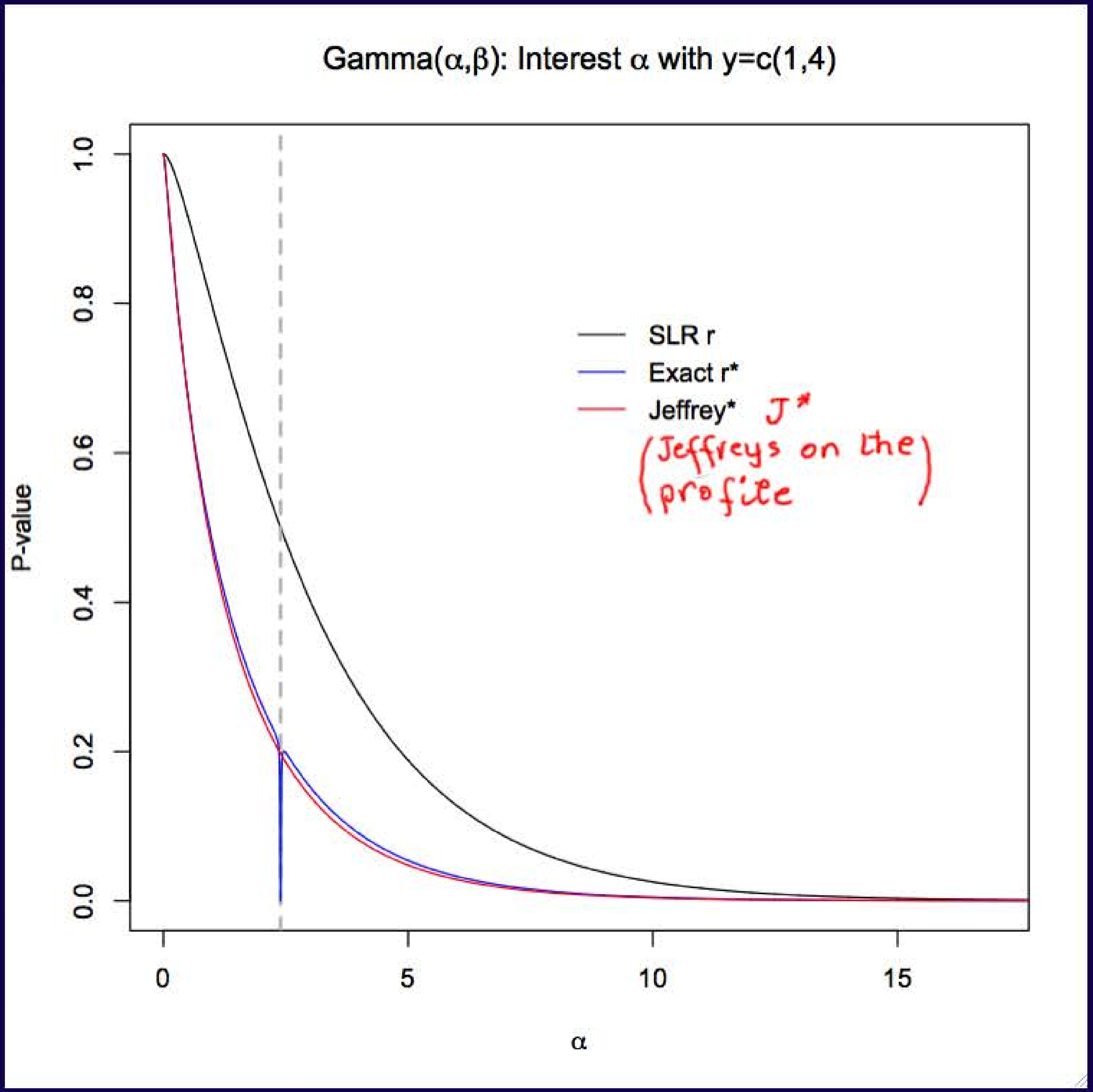
Setup

- Model: $Y \sim Gamma(\alpha, \beta)$ where α is shape and β is rate
  pdf is $\frac{\beta^\alpha}{\Gamma(\alpha)} y^{\alpha-1} e^{-\beta y}$
  Let $n = 2$ and data is $(y_1, y_2) = (1, 4)$

  Parameter of interest is α, β free nuisance


Gamma(α,β): Interest α with y=c(1,4)
P-value
1.0
0.8
0.6
0.4
0.2
0.0
0
5
10
15
α
SLR r
CLT

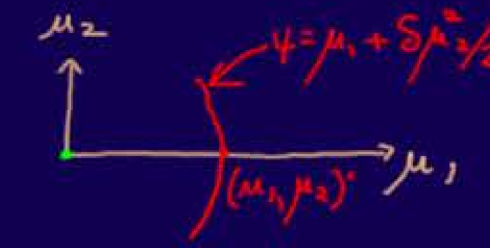

Gamma(α,β): Interest β with y=c(1,4)
P-value
1.0
0.8
0.6
0.4
0.2
0.0
0
2
4
6
8
β
SLR r
Exact r*

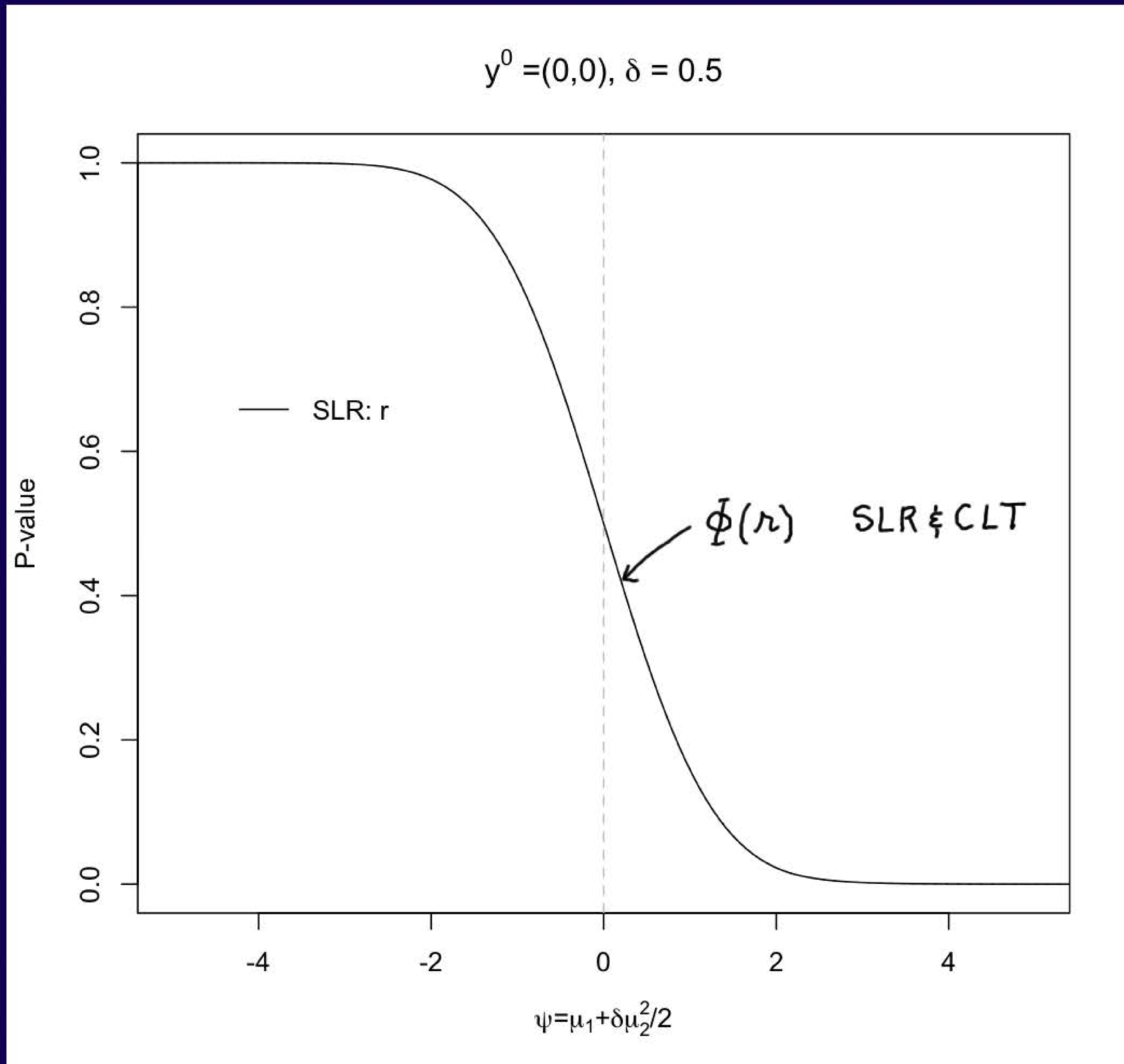


Gamma(α,β): Interest α with y=c(1,4)
P-value
1.0
0.8
0.6
0.4
0.2
0.0
0
5
10
15
α
SLR r
Exact r*
Jeffrey*
J*
(Jeffreys on the profile)

Example 4 $N\left(\begin{pmatrix}\mu_1\\ \mu_2\end{pmatrix}; I\right)$ $\psi = \mu_1 + \delta\mu_2^2/2$

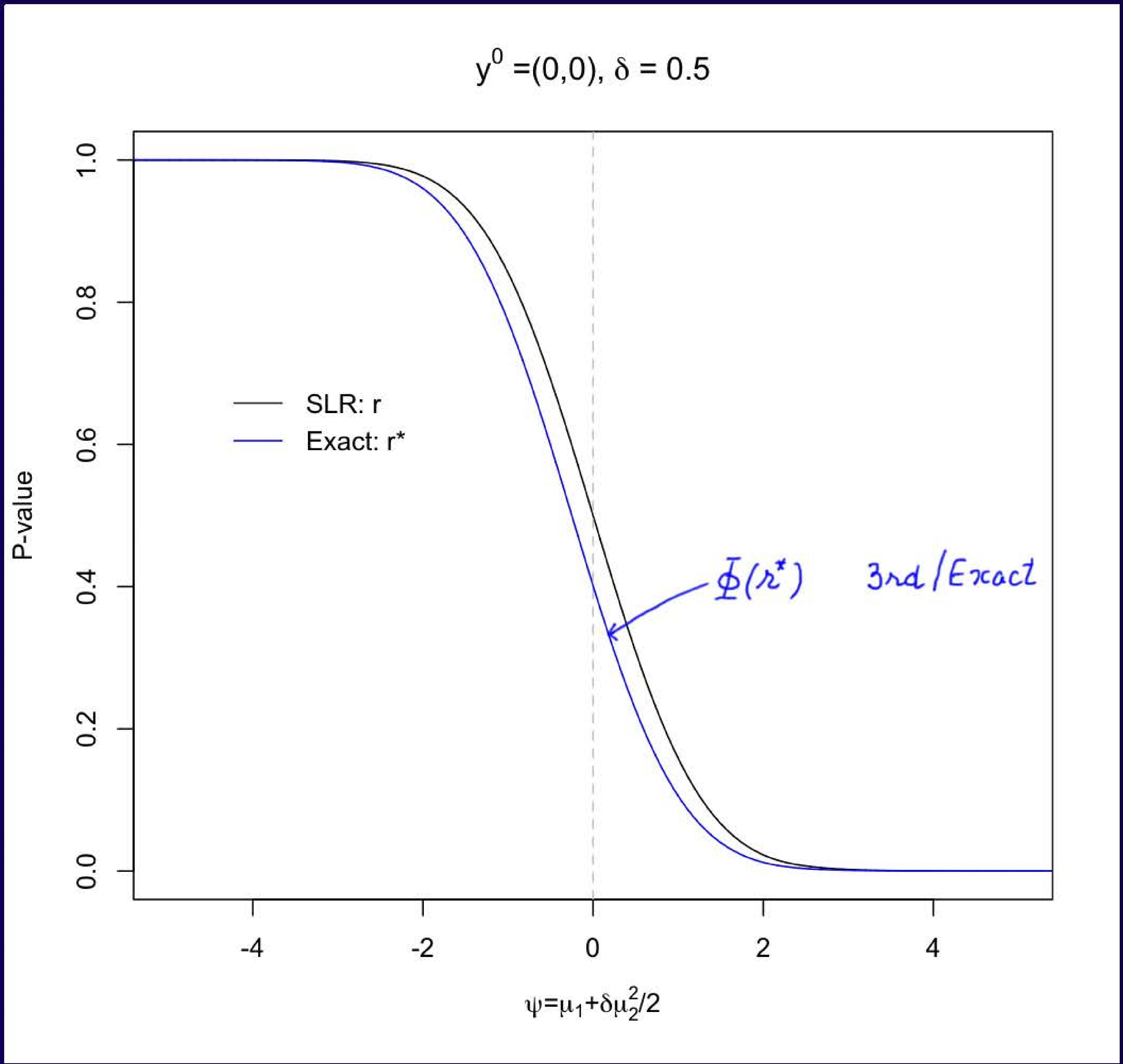
$y^0 = \begin{pmatrix}0\\ 0\end{pmatrix}$ Curved Interest (A story in itself!)





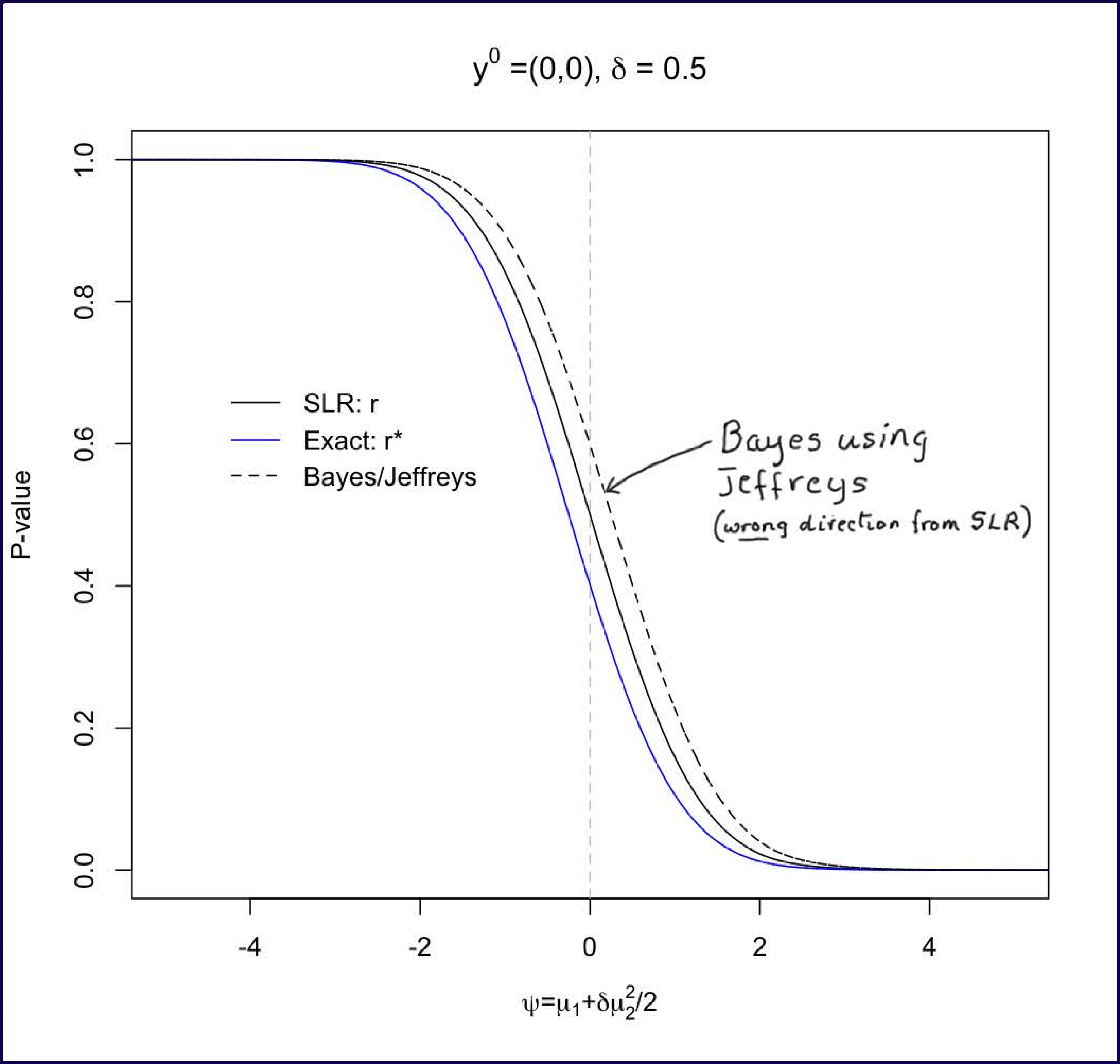
Example 4 $N\left(\begin{smallmatrix}\mu_1\\ \mu_2\end{smallmatrix}; I\right)$ $\psi = \mu_1 + \delta\mu_2^2/2$

$y^0 = \left(\begin{smallmatrix}0\\0\end{smallmatrix}\right)$



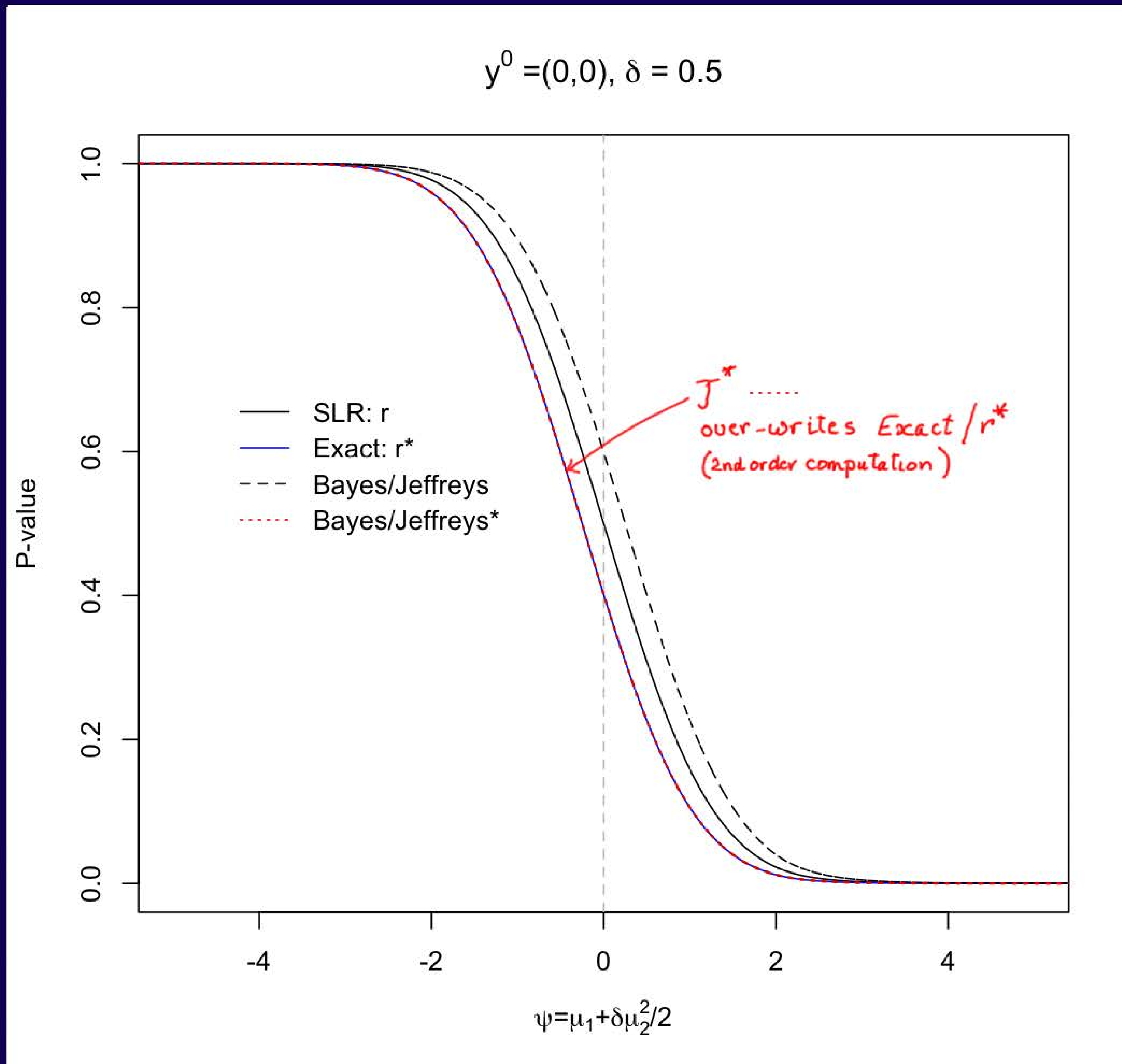
Example 4 $N\left(\begin{smallmatrix}\mu_1\\ \mu_2\end{smallmatrix}; I\right)$ $\psi = \mu_1 + \delta\mu_2^2/2$

$y^0 = \left(\begin{smallmatrix}0\\ 0\end{smallmatrix}\right)$



Example 4 $N\left(\binom{\mu_1}{\mu_2}; I\right)$ $\psi = \mu_1 + \delta \mu_2^2/2$

$y^0 = \binom{0}{0}$



BIG DATA

The Parable of Google Flu: Traps in Big Data Analysis

Large errors in flu prediction were largely avoidable, which offers lessons for the use of big data.

David Lazer,[1,2*] Ryan Kennedy,[1,3,4] Gary King,[3] Alessandro Vespignani[5,6,7]

In February 2013, Google Flu Trends (GFT) made headlines but not for a reason that Google executives or the creators of the flu tracking system would have hoped. *Nature* reported that GFT was predicting more than double the proportion of doctor visits for influenza-like illness (ILI) than the Centers for Disease Control and Prevention (CDC), which bases its estimates on surveillance reports from laboratories across the United States (*1, 2*). This happened despite the fact that GFT was built to predict CDC reports. Given that GFT is often held up as an exemplary use of big data (*3, 4*), what lessons can we draw from this error?

The problems we identify are not limited to GFT. Research on whether search or social media can predict *x* has become commonplace (*5–7*) and is often put in sharp contrast with traditional methods and hypotheses. Although these studies have shown the value of these data, we are far from a place where they can supplant more traditional methods or theories (*8*). We explore two issues that contributed to GFT's mistakes—big data hubris and algorithm dynamics—and offer lessons for moving forward in the big data age.

Big Data Hubris

"Big data hubris" is the often implicit assumption that big data are a substitute for, rather than a supplement to, traditional data collection and analysis. Elsewhere, we have asserted that there are enormous scientific possibilities in big data (*9–11*). However, quantity of data does not mean that one can ignore foundational issues of measurement and construct validity and reliability and dependencies among data (*12*). The core challenge is that most big data that have received popular attention are not the output of instruments designed to produce valid and reliable data amenable for scientific analysis.

The initial version of GFT was a particularly problematic marriage of big and small data. Essentially, the methodology was to find the best matches among 50 million search terms to fit 1152 data points (*13*). The odds of finding search terms that match the propensity of the flu but are structurally unrelated, and so do not predict the future, were quite high. GFT developers, in fact, report weeding out seasonal search terms unrelated to the flu but strongly correlated to the CDC data, such as those regarding high school basketball (*13*). This should have been a warning that the big data were overfitting the small number of cases—a standard concern in data analysis. This ad hoc method of throwing out peculiar search terms failed when GFT completely missed the nonseasonal 2009 influenza A–H1N1 pandemic (*2, 14*). In short, the initial version of GFT was part flu detector, part winter detector. GFT engineers updated the algorithm in 2009, and this model has run ever since, with a few changes announced in October 2013 (*10, 15*).

Although not widely reported until 2013, the new GFT has been persistently overestimating flu prevalence for a much longer time. GFT also missed by a very large margin in the 2011–2012 flu season and has missed high for 100 out of 108 weeks starting with August 2011 (see the graph). These errors are not randomly distributed. For example, last week's errors predict this week's errors (temporal autocorrelation), and the direction and magnitude of error varies with the time of year (seasonality). These patterns mean that GFT overlooks considerable information that could be extracted by traditional statistical methods.

Even after GFT was updated in 2009, the comparative value of the algorithm as a stand-alone flu monitor is questionable. A study in 2010 demonstrated that GFT accuracy was not much better than a fairly simple projection forward using already available (typically on a 2-week lag) CDC data (*4*). The comparison has become even worse since that time, with lagged models significantly outperforming GFT (see the graph). Even 3-week-old CDC data do a better job of projecting current flu prevalence than GFT [see supplementary materials (SM)].

Considering the large number of approaches that provide inference on influenza activity (*16–19*), does this mean that the current version of GFT is not useful? No, greater value can be obtained by combining GFT with other near–real-time health data (*2, 20*). For example, by combining GFT and lagged CDC data, as well as dynamically recalibrating GFT, we can substantially improve on the performance of GFT or the CDC alone (see the chart). This is no substitute for ongoing evaluation and improvement, but, by incorporating this information, GFT could have largely healed itself and would have likely remained out of the headlines.

[1]Lazer Laboratory, Northeastern University, Boston, MA 02115, USA. [2]Harvard Kennedy School, Harvard University, Cambridge, MA 02138, USA. [3]Institute for Quantitative Social Science, Harvard University, Cambridge, MA 02138, USA. [4]University of Houston, Houston, TX 77204, USA. [5]Laboratory for the Modeling of Biological and Sociotechnical Systems, Northeastern University, Boston, MA 02115, USA. [6]Institute for Scientific Interchange Foundation, Turin, Italy. *Corresponding author. E-mail: d.lazer@neu.edu.

Science
2014 Mar 14

Google Flu Trends
vs
CDC

BIG DATA

The Parable of Google Flu: Traps in Big Data Analysis

Large errors in flu prediction were largely avoidable, which offers lessons for the use of big data.

David Lazer,[1,2*] Ryan Kennedy,[1,3,4] Gary King,[3] Alessandro Vespignani[5,6,7]

In February 2013, Google Flu Trends (GFT) made headlines but not for a reason that Google executives or the creators of the flu tracking system would have hoped. *Nature* reported that GFT was predicting more than double the proportion of doctor visits for influenza-like illness (ILI) than the Centers for Disease Control and Prevention (CDC), which bases its estimates on surveillance reports from laboratories across the United States (*1*, *2*). This happened despite the fact that GFT was built to predict CDC reports. Given that GFT is often held up as an exemplary use of big data (*3*, *4*), what lessons can we draw from this error?

The problems we identify are not limited to GFT. Research on whether search or social media can predict *x* has become commonplace (*5*–*7*) and is often put in sharp contrast with traditional methods and hypotheses. Although these studies have shown the value of these data, we are far from a place where they can supplant more traditional methods or theories (*8*). We explore two issues that contributed to GFT's mistakes—big data hubris and algorithm dynamics—and offer lessons for moving forward in the big data age.

CREDIT: ADAPTED FROM AXEL KORES/DESIGN & ART DIRECTION/ISTOCKPHOTO.COM

Big Data Hubris

"Big data hubris" is the often implicit assumption that big data are a substitute for, rather than a supplement to, traditional data collection and analysis. Elsewhere, we have asserted that there are enormous scientific possibilities in big data (*9*–*11*). However, quantity of data does not mean that one can ignore foundational issues of measurement and construct validity and reliability and dependencies among data (*12*). The core challenge is that most big data that have received popular attention are not the output of instruments designed to produce valid and reliable data amenable for scientific analysis.

The initial version of GFT was a particularly problematic marriage of big and small data. Essentially, the methodology was to find the best matches among 50 million search terms to fit 1152 data points (*13*). The odds of finding search terms that match the propensity of the flu but are structurally unrelated, and so do not predict the future, were quite high. GFT developers, in fact, report weeding out seasonal search terms unrelated to the flu but strongly correlated to the CDC data, such as those regarding high school basketball (*13*). This should have been a warning that the big data were overfitting the small number of cases—a standard concern in data analysis. This ad hoc method of throwing out peculiar search terms failed when GFT completely missed the nonseasonal 2009 influenza A–H1N1 pandemic (*2*, *14*). In short, the initial version of GFT was part flu detector, part winter detector. GFT engineers updated the algorithm in 2009, and this model has run ever since, with a few changes announced in October 2013 (*10*, *15*).

Although not widely reported until 2013, the new GFT has been persistently overestimating flu prevalence for a much longer time. GFT also missed by a very large margin in the 2011–2012 flu season and has missed high for 100 out of 108 weeks starting with August 2011 (see the graph). These errors are not randomly distributed. For example, last week's errors predict this week's errors (temporal autocorrelation), and the direction and magnitude of error varies with the time of year (seasonality). These patterns mean that GFT overlooks considerable information that could be extracted by traditional statistical methods.

Even after GFT was updated in 2009, the comparative value of the algorithm as a stand-alone flu monitor is questionable. A study in 2010 demonstrated that GFT accuracy was not much better than a fairly simple projection forward using already available (typically on a 2-week lag) CDC data (*4*). The comparison has become even worse since that time, with lagged models significantly outperforming GFT (see the graph). Even 3-week-old CDC data do a better job of projecting current flu prevalence than GFT [see supplementary materials (SM)].

Considering the large number of approaches that provide inference on influenza activity (*16*–*19*), does this mean that the current version of GFT is not useful? No, greater value can be obtained by combining GFT with other near–real-time health data (*2*, *20*). For example, by combining GFT and lagged CDC data, as well as dynamically recalibrating GFT, we can substantially improve on the performance of GFT or the CDC alone (see the chart). This is no substitute for ongoing evaluation and improvement, but, by incorporating this information, GFT could have largely healed itself and would have likely remained out of the headlines.

[1]Lazer Laboratory, Northeastern University, Boston, MA 02115, USA. [2]Harvard Kennedy School, Harvard University, Cambridge, MA 02138, USA. [3]Institute for Quantitative Social Science, Harvard University, Cambridge, MA 02138, USA. [4]University of Houston, Houston, TX 77204, USA. [5]Laboratory for the Modeling of Biological and Sociotechnical Systems, Northeastern University, Boston, MA 02115, USA. [6]Institute for Scientific Interchange Foundation, Turin, Italy. *Corresponding author. E-mail: d.lazer@neu.edu.

Science
2014 Mar 14

Google Flu Trends
vs
CDC

- GFT predicts more than double that of CDC

BIG DATA

The Parable of Google Flu: Traps in Big Data Analysis

Large errors in flu prediction were largely avoidable, which offers lessons for the use of big data.

David Lazer,[1,2*] Ryan Kennedy,[1,3,4] Gary King,[3] Alessandro Vespignani[5,6,7]

In February 2013, Google Flu Trends (GFT) made headlines but not for a reason that Google executives or the creators of the flu tracking system would have hoped. *Nature* reported that GFT was predicting more than double the proportion of doctor visits for influenza-like illness (ILI) than the Centers for Disease Control and Prevention (CDC), which bases its estimates on surveillance reports from laboratories across the United States (*1, 2*). This happened despite the fact that GFT was built to predict CDC reports. Given that GFT is often held up as an exemplary use of big data (*3, 4*), what lessons can we draw from this error?

The problems we identify are not limited to GFT. Research on whether search or social media can predict *x* has become commonplace (*5–7*) and is often put in sharp contrast with traditional methods and hypotheses. Although these studies have shown the value of these data, we are far from a place where they can supplant more traditional methods or theories (*8*). We explore two issues that contributed to GFT's mistakes—big data hubris and algorithm dynamics—and offer lessons for moving forward in the big data age.

Big Data Hubris

"Big data hubris" is the often implicit assumption that big data are a substitute for, rather than a supplement to, traditional data collection and analysis. Elsewhere, we have asserted that there are enormous scientific possibilities in big data (*9–11*). However, quantity of data does not mean that one can ignore foundational issues of measurement and construct validity and reliability and dependencies among data (*12*). The core challenge is that most big data that have received popular attention are not the output of instruments designed to produce valid and reliable data amenable for scientific analysis.

The initial version of GFT was a particularly problematic marriage of big and small data. Essentially, the methodology was to find the best matches among 50 million search terms to fit 1152 data points (*13*). The odds of finding search terms that match the propensity of the flu but are structurally unrelated, and so do not predict the future, were quite high. GFT developers, in fact, report weeding out seasonal search terms unrelated to the flu but strongly correlated to the CDC data, such as those regarding high school basketball (*13*). This should have been a warning that the big data were overfitting the small number of cases—a standard concern in data analysis. This ad hoc method of throwing out peculiar search terms failed when GFT completely missed the nonseasonal 2009 influenza A–H1N1 pandemic (*2, 14*). In short, the initial version of GFT was part flu detector, part winter detector. GFT engineers updated the algorithm in 2009, and this model has run ever since, with a few changes announced in October 2013 (*10, 15*).

Although not widely reported until 2013, the new GFT has been persistently overestimating flu prevalence for a much longer time. GFT also missed by a very large margin in the 2011–2012 flu season and has missed high for 100 out of 108 weeks starting with August 2011 (see the graph). These errors are not randomly distributed. For example, last week's errors predict this week's errors (temporal autocorrelation), and the direction and magnitude of error varies with the time of year (seasonality). These patterns mean that GFT overlooks considerable information that could be extracted by traditional statistical methods.

Even after GFT was updated in 2009, the comparative value of the algorithm as a stand-alone flu monitor is questionable. A study in 2010 demonstrated that GFT accuracy was not much better than a fairly simple projection forward using already available (typically on a 2-week lag) CDC data (*4*). The comparison has become even worse since that time, with lagged models significantly outperforming GFT (see the graph). Even 3-week-old CDC data do a better job of projecting current flu prevalence than GFT [see supplementary materials (SM)].

Considering the large number of approaches that provide inference on influenza activity (*16–19*), does this mean that the current version of GFT is not useful? No, greater value can be obtained by combining GFT with other near–real-time health data (*2, 20*). For example, by combining GFT and lagged CDC data, as well as dynamically recalibrating GFT, we can substantially improve on the performance of GFT or the CDC alone (see the chart). This is no substitute for ongoing evaluation and improvement, but, by incorporating this information, GFT could have largely healed itself and would have likely remained out of the headlines.

[1]Lazer Laboratory, Northeastern University, Boston, MA 02115, USA. [2]Harvard Kennedy School, Harvard University, Cambridge, MA 02138, USA. [3]Institute for Quantitative Social Science, Harvard University, Cambridge, MA 02138, USA. [4]University of Houston, Houston, TX 77204, USA. [5]Laboratory for the Modeling of Biological and Sociotechnical Systems, Northeastern University, Boston, MA 02115, USA. [6]Institute for Scientific Interchange Foundation, Turin, Italy. *Corresponding author. E-mail: d.lazer@neu.edu.

Science
2014 Mar 14

Google Flu Trends
vs
CDC

GFT predicts more than double that of CDC

- 3-week-old CDC better than GFT

BIG DATA

The Parable of Google Flu: Traps in Big Data Analysis

Large errors in flu prediction were largely avoidable, which offers lessons for the use of big data.

David Lazer,[1,2*] Ryan Kennedy,[1,3,4] Gary King,[3] Alessandro Vespignani[5,6,7]

In February 2013, Google Flu Trends (GFT) made headlines but not for a reason that Google executives or the creators of the flu tracking system would have hoped. *Nature* reported that GFT was predicting more than double the proportion of doctor visits for influenza-like illness (ILI) than the Centers for Disease Control and Prevention (CDC), which bases its estimates on surveillance reports from laboratories across the United States (*1*, *2*). This happened despite the fact that GFT was built to predict CDC reports. Given that GFT is often held up as an exemplary use of big data (*3*, *4*), what lessons can we draw from this error?

The problems we identify are not limited to GFT. Research on whether search or social media can predict x has become commonplace (*5*–*7*) and is often put in sharp contrast with traditional methods and hypotheses. Although these studies have shown the value of these data, we are far from a place where they can supplant more traditional methods or theories (*8*). We explore two issues that contributed to GFT's mistakes—big data hubris and algorithm dynamics—and offer lessons for moving forward in the big data age.

Big Data Hubris

"Big data hubris" is the often implicit assumption that big data are a substitute for, rather than a supplement to, traditional data collection and analysis. Elsewhere, we have asserted that there are enormous scientific possibilities in big data (*9*–*11*). However, quantity of data does not mean that one can ignore foundational issues of measurement and construct validity and reliability and dependencies among data (*12*). The core challenge is that most big data that have received popular attention are not the output of instruments designed to produce valid and reliable data amenable for scientific analysis.

The initial version of GFT was a particularly problematic marriage of big and small data. Essentially, the methodology was to find the best matches among 50 million search terms to fit 1152 data points (*13*). The odds of finding search terms that match the propensity of the flu but are structurally unrelated, and so do not predict the future, were quite high. GFT developers, in fact, report weeding out seasonal search terms unrelated to the flu but strongly correlated to the CDC data, such as those regarding high school basketball (*13*). This should have been a warning that the big data were overfitting the small number of cases—a standard concern in data analysis. This ad hoc method of throwing out peculiar search terms failed when GFT completely missed the nonseasonal 2009 influenza A–H1N1 pandemic (*2*, *14*). In short, the initial version of GFT was part flu detector, part winter detector. GFT engineers updated the algorithm in 2009, and this model has run ever since, with a few changes announced in October 2013 (*10*, *15*).

Although not widely reported until 2013, the new GFT has been persistently overestimating flu prevalence for a much longer time. GFT also missed by a very large margin in the 2011–2012 flu season and has missed high for 100 out of 108 weeks starting with August 2011 (see the graph). These errors are not randomly distributed. For example, last week's errors predict this week's errors (temporal autocorrelation), and the direction and magnitude of error varies with the time of year (seasonality). These patterns mean that GFT overlooks considerable information that could be extracted by traditional statistical methods.

Even after GFT was updated in 2009, the comparative value of the algorithm as a stand-alone flu monitor is questionable. A study in 2010 demonstrated that GFT accuracy was not much better than a fairly simple projection forward using already available (typically on a 2-week lag) CDC data (*4*). The comparison has become even worse since that time, with lagged models significantly outperforming GFT (see the graph). Even 3-week-old CDC data do a better job of projecting current flu prevalence than GFT [see supplementary materials (SM)].

Considering the large number of approaches that provide inference on influenza activity (*16*–*19*), does this mean that the current version of GFT is not useful? No, greater value can be obtained by combining GFT with other near–real-time health data (*2*, *20*). For example, by combining GFT and lagged CDC data, as well as dynamically recalibrating GFT, we can substantially improve on the performance of GFT or the CDC alone (see the chart). This is no substitute for ongoing evaluation and improvement, but, by incorporating this information, GFT could have largely healed itself and would have likely remained out of the headlines.

[1]Lazer Laboratory, Northeastern University, Boston, MA 02115, USA. [2]Harvard Kennedy School, Harvard University, Cambridge, MA 02138, USA. [3]Institute for Quantitative Social Science, Harvard University, Cambridge, MA 02138, USA. [4]University of Houston, Houston, TX 77204, USA. [5]Laboratory for the Modeling of Biological and Sociotechnical Systems, Northeastern University, Boston, MA 02115, USA. [6]Institute for Scientific Interchange Foundation, Turin, Italy. *Corresponding author. E-mail: d.lazer@neu.edu.

Science
2014 Mar 14

Google Flu Trends
vs
CDC

GFT predicts more than double
that of CDC

3-week-old CDC better
than GFT

- careful research article
on BIG DATA

BIG DATA

The Parable of Google Flu: Traps in Big Data Analysis

Large errors in flu prediction were largely avoidable, which offers lessons for the use of big data.

David Lazer,[1,2]* Ryan Kennedy,[1,3,4] Gary King,[3] Alessandro Vespignani[5,6,3]

In February 2013, Google Flu Trends (GFT) made headlines but not for a reason that Google executives or the creators of the flu tracking system would have hoped. *Nature* reported that GFT was predicting more than double the proportion of doctor visits for influenza-like illness (ILI) than the Centers for Disease Control and Prevention (CDC), which bases its estimates on surveillance reports from laboratories across the United States (*1, 2*). This happened despite the fact that GFT was built to predict CDC reports. Given that GFT is often held up as an exemplary use of big data (*3, 4*), what lessons can we draw from this error?

The problems we identify are not limited to GFT. Research on whether search or social media can predict *x* has become commonplace (*5–7*) and is often put in sharp contrast with traditional methods and hypotheses. Although these studies have shown the value of these data, we are far from a place where they can supplant more traditional methods or theories (*8*). We explore two issues that contributed to GFT's mistakes—big data hubris and algorithm dynamics—and offer lessons for moving forward in the big data age.

CREDIT: ADAPTED FROM AXEL KORES/DESIGN & ART DIRECTION/ISTOCKPHOTO.COM

Big Data Hubris

"Big data hubris" is the often implicit assumption that big data are a substitute for, rather than a supplement to, traditional data collection and analysis. Elsewhere, we have asserted that there are enormous scientific possibilities in big data (*9–11*). However, quantity of data does not mean that one can ignore foundational issues of measurement and construct validity and reliability and dependencies among data (*12*). The core challenge is that most big data that have received popular attention are not the output of instruments designed to produce valid and reliable data amenable for scientific analysis.

The initial version of GFT was a particularly problematic marriage of big and small data. Essentially, the methodology was to find the best matches among 50 million search terms to fit 1152 data points (*13*). The odds of finding search terms that match the propensity of the flu but are structurally unrelated, and so do not predict the future, were quite high. GFT developers, in fact, report weeding out seasonal search terms unrelated to the flu but strongly correlated to the CDC data, such as those regarding high school basketball (*13*). This should have been a warning that the big data were overfitting the small number of cases—a standard concern in data analysis. This ad hoc method of throwing out peculiar search terms failed when GFT completely missed the nonseasonal 2009 influenza A–H1N1 pandemic (*2, 14*). In short, the initial version of GFT was part flu detector, part winter detector. GFT engineers updated the algorithm in 2009, and this model has run ever since, with a few changes announced in October 2013 (*10, 15*).

Although not widely reported until 2013, the new GFT has been persistently overestimating flu prevalence for a much longer time. GFT also missed by a very large margin in the 2011–2012 flu season and has missed high for 100 out of 108 weeks starting with August 2011 (see the graph). These errors are not randomly distributed. For example, last week's errors predict this week's errors (temporal autocorrelation), and the direction and magnitude of error varies with the time of year (seasonality). These patterns mean that GFT overlooks considerable information that could be extracted by traditional statistical methods.

Even after GFT was updated in 2009, the comparative value of the algorithm as a stand-alone flu monitor is questionable. A study in 2010 demonstrated that GFT accuracy was not much better than a fairly simple projection forward using already available (typically on a 2-week lag) CDC data (*4*). The comparison has become even worse since that time, with lagged models significantly outperforming GFT (see the graph). Even 3-week-old CDC data do a better job of projecting current flu prevalence than GFT [see supplementary materials (SM)].

Considering the large number of approaches that provide inference on influenza activity (*16–19*), does this mean that the current version of GFT is not useful? No, greater value can be obtained by combining GFT with other near–real-time health data (*2, 20*). For example, by combining GFT and lagged CDC data, as well as dynamically recalibrating GFT, we can substantially improve on the performance of GFT or the CDC alone (see the chart). This is no substitute for ongoing evaluation and improvement, but, by incorporating this information, GFT could have largely healed itself and would have likely remained out of the headlines.

[1]Lazer Laboratory, Northeastern University, Boston, MA 02115, USA. [2]Harvard Kennedy School, Harvard University, Cambridge, MA 02138, USA. [3]Institute for Quantitative Social Science, Harvard University, Cambridge, MA 02138, USA. [4]University of Houston, Houston, TX 77204, USA. [5]Laboratory for the Modeling of Biological and Sociotechnical Systems, Northeastern University, Boston, MA 02115, USA. [6]Institute for Scientific Interchange Foundation, Turin, Italy. *Corresponding author. E-mail: d.lazer@neu.edu.

Science
2014 Mar 14

Google Flu Trends
vs
CDC

GFT predicts more than double that of CDC

3-week-old CDC better than GFT

carefully research article on BIG DATA

⇒ Reproducibility!

A cautionary tale

Discussion:

1. $f(y-\theta)$ with Jeffreys $\Rightarrow$ <u>Exact</u> inference, <u>Reproducibility</u>

Discussion:

1. $f(y-\theta)$ with Jeffreys $\Rightarrow$ Exact inference, Reproducibility

2. $f(y;\theta)$, vector θ, y°,
regular, continuity,
indep. components

Discussion:

1. $f(y-\theta)$ with Jeffreys $\Rightarrow$ Exact inference, Reproducibility

2. $f(y;\theta)$, vector θ, y°, regular, continuity, indep. components $\Rightarrow$ Unique quantile representation $y = y(\theta; z)$

Discussion:

1. $f(y-\theta)$ with Jeffreys $\Rightarrow$ Exact inference, Reproducibility

2. $f(y;\theta)$, vector θ, y°, regular, continuity, indep. components $\Rightarrow$ Unique quantile representation $y = y(\theta; z)$ $\Rightarrow$ Directions $V = (v_1, \ldots, v_p) = \left.\frac{dy}{d\theta}\right|_{\hat{\theta}^\circ}^{y^\circ}$ $\Rightarrow$

where y "measures" θ at y°

Discussion:

1. $f(y-\theta)$ with Jeffreys $\Rightarrow$ Exact inference, Reproducibility

2. $f(y;\theta)$, vector θ, y^0, regular, continuity, indep. components $\Rightarrow$ Unique quantile representation $y = y(\theta; z)$ $\Rightarrow$ Directions $V = (v_1, \dots, v_p) = \left.\frac{dy}{d\theta}\right|_{y^0, \hat{\theta}^0}$ where y "measures" θ at y^0 $\Rightarrow$ Canonical $\varphi(\theta) = \left.\frac{d\,\ell(\theta, y)}{dV}\right|_{y^0}$ of exponential model

Discussion:

1. $f(y-\theta)$ with Jeffreys $\Rightarrow$ Exact inference , Reproducibility

2. $f(y;\theta)$, vector θ, y°, regular, continuity, indep. components $\Rightarrow$ Unique quantile representation $y = y(\theta; z)$ $\Rightarrow$ Directions $V = (v_1, \ldots, v_p) = \left.\frac{dy}{d\theta}\right|_{\hat\theta^\circ}^{y^\circ}$ where y "measures" θ at y° $\Rightarrow$ Canonical $\varphi(\theta) = \left.\frac{d\ell(\theta, y)}{dV}\right|_{y^\circ}$ of exponential model

3. $\ell^\circ(\theta)$, $\varphi(\theta)$, data y°, scalar interest $\psi(\theta)$

Discussion:

1. $f(y-\theta)$ with Jeffreys $\Rightarrow$ Exact inference, Reproducibility

2. $f(y;\theta)$, vector θ, y°, regular, continuity, indep. components $\Rightarrow$ Unique quantile representation $y = y(\theta; z)$ $\Rightarrow$ Directions $V = (v_1, \dots, v_p) = \left.\dfrac{dy}{d\theta}\right|_{\hat{\theta}^\circ}^{y^\circ}$ where y "measures" θ at y° $\Rightarrow$ Canonical $\varphi(\theta) = \left.\dfrac{d\ell(\theta, y)}{dV}\right|_{y^\circ}$ of exponential model

3. $\ell^\circ(\theta)$, $\varphi(\theta)$, data y°, scalar interest $\psi(\theta)$ $\Rightarrow$ Unique $O(n^{-3/2})$ p-value for $\psi(\theta)$

Discussion:

1. $f(y-\theta)$ with Jeffreys $\Rightarrow$ <u>Exact</u> inference, <u>Reproducibility</u>

2. $f(y;\theta)$, vector θ, y°, regular, continuity, indep. components $\Rightarrow$ Unique quantile representation $y = y(\theta; z)$ $\Rightarrow$ Directions $V = (v_1, \ldots, v_p) = \left.\dfrac{dy}{d\theta}\right|_{\hat{\theta}^\circ}^{y^\circ}$ where y "measures" θ at y° $\Rightarrow$ Canonical $\varphi(\theta) = \left.\dfrac{d\,\ell(\theta, y)}{dV}\right|_{y^\circ}$ of exponential model

3. $\ell^\circ(\theta)$, $\varphi(\theta)$, data y°, scalar interest $\psi(\theta)$ $\Rightarrow$ Unique $O(n^{-3/2})$ <u>p-value fn</u> $p(\theta)$ (Suff'cy; Ancillarity not needed/wanted)

Discussion:

1. $f(y-\theta)$ with Jeffreys $\Rightarrow$ Exact inference, Reproducibility

2. $f(y;\theta)$, vector θ, y°, regular, continuity, indep. components $\Rightarrow$ Unique quantile representation $y = y(\theta; z)$ $\Rightarrow$ Directions $V = (v_1, \ldots, v_p) = \left.\frac{dy}{d\theta}\right|_{\hat{\theta}^\circ}^{y^\circ}$ where y "measures" θ at y° $\Rightarrow$ Canonical $\varphi(\theta) = \left.\frac{d\ell(\theta, y)}{dV}\right|_{y^\circ}$ of exponential model

3. $\ell^\circ(\theta)$, $\varphi(\theta)$, data y°, scalar interest $\psi(\theta)$ $\Rightarrow$ Unique $O(n^{-3/2})$ p-value fn $p(\theta)$ (Suff'cy; Ancillarity not needed/wanted) $\Rightarrow$ Unique reproducible inference

Discussion:

1. $f(y-\theta)$ with Jeffreys $\Rightarrow$ Exact inference, Reproducibility

2. $f(y;\theta)$, vector θ, y°, regular, continuity, indep. components $\Rightarrow$ Unique quantile representation $y = y(\theta; z)$ $\Rightarrow$ Directions $V = (v_1, \ldots, v_p) = \left.\frac{dy}{d\theta}\right|_{y^\circ, \hat{\theta}^\circ}$ where y "measures" θ at y° $\Rightarrow$ Canonical $\varphi(\theta) = \left.\frac{d\,\ell(\theta, y)}{dV}\right|_{y^\circ}$ of exponential model

3. $\ell^\circ(\theta)$, $\varphi(\theta)$, data y°, scalar interest $\psi(\theta)$ $\Rightarrow$ Unique $O(n^{-3/2})$ p-value fn $p(\theta)$ (Suff'cy; Ancillarity not needed / wanted) $\Rightarrow$ Unique reproducible inference

4. $\ell^\circ(\theta)$, $\varphi(\theta)$, data y°, scalar interest $\psi(\theta)$

Discussion:

1. $f(y-\theta)$ with Jeffreys $\Rightarrow$ Exact inference, Reproducibility

2. $f(y;\theta)$, vector θ, y°, regular, continuity, indep. components $\Rightarrow$ Unique quantile representation $y = y(\theta; z)$ $\Rightarrow$ Directions $V = (v_1, \ldots, v_p) = \left.\frac{dy}{d\theta}\right|_{\hat{\theta}^\circ}^{y^\circ}$ where y "measures" θ at y° $\Rightarrow$ Canonical $\varphi(\theta) = \left.\frac{d\ell(\theta; y)}{dV}\right|_{y^\circ}$ of exponential model

3. $\ell^\circ(\theta)$, $\varphi(\theta)$, data y°, scalar interest $\psi(\theta)$ $\Rightarrow$ Unique $O(n^{-3/2})$ p-value fn $p(\theta)$ (Suff'cy; Ancillarity not needed / wanted) $\Rightarrow$ Unique reproducible inference

4. $\ell^\circ(\theta)$, $\varphi(\theta)$, data y°, scalar interest $\psi(\theta)$ $\Rightarrow$ Jeff $= \pi(\varphi) = |j_{\varphi\varphi}(\varphi) = -\ell_{\varphi\varphi}(\theta)|^{1/2}$

Rescale so $j_{\varphi\varphi}(\hat{\varphi}^\circ) = I$

Discussion:

1. $f(y-\theta)$ with Jeffreys $\Rightarrow$ Exact inference, Reproducibility

2. $f(y;\theta)$, vector θ, y°, regular, continuity, indep. components $\Rightarrow$ Unique quantile representation $y = y(\theta; z)$ $\Rightarrow$ Directions $V = (v_1, \ldots, v_p) = \left.\frac{dy}{d\theta}\right|_{y^\circ, \hat\theta^\circ}$ where y "measures" θ at y° $\Rightarrow$ Canonical $\varphi(\theta) = \left.\frac{d\ell(\theta, y)}{dV}\right|_{y^\circ}$ of exponential model

3. $\ell^\circ(\theta)$, $\varphi(\theta)$, data y°, scalar interest $\psi(\theta)$ $\Rightarrow$ Unique $O(n^{-3/2})$ p-value fn $p(\theta)$ $\left(\begin{array}{l}\text{Suff'cy; Ancillarity}\\ \text{not needed / wanted}\end{array}\right)$ $\Rightarrow$ Unique reproducible inference

4. $\ell^\circ(\theta)$, $\varphi(\theta)$, data y°, scalar interest $\psi(\theta)$ $\Rightarrow$ Jeff $= \pi(\varphi) = |j_{\varphi\varphi}(\varphi) = -\ell_{\varphi\varphi}(\theta)|^{1/2}$ Rescale so $j_{\varphi\varphi}(\hat\varphi^\circ) = I$ Use $\pi(\varphi)$ only on curve $C_\psi = \{\varphi : \psi(\varphi) = \hat\psi^\circ\}$

Discussion:

1. $f(y-\theta)$ with Jeffreys $\Rightarrow$ Exact inference, Reproducibility

2. $f(y;\theta)$, vector θ, y°, regular, continuity, indep. components $\Rightarrow$ Unique quantile representation $y = y(\theta; z)$ $\Rightarrow$ Directions $V = (v_1, \dots, v_p) = \left.\frac{dy}{d\theta}\right|_{\hat{\theta}^\circ}^{y^\circ}$ where y "measures" θ at y° $\Rightarrow$ Canonical $\varphi(\theta) = \left.\frac{d\ell(\theta, y)}{dV}\right|_{y^\circ}$ of exponential model

3. $\ell^\circ(\theta)$, $\varphi(\theta)$, data y°, scalar interest $\psi(\theta)$ $\Rightarrow$ Unique $O(n^{-3/2})$ p-value fn $p(\theta)$ (Suff'cy; Ancillarity not needed/wanted) $\Rightarrow$ Unique reproducible inference

4. $\ell^\circ(\theta)$, $\varphi(\theta)$, data y°, scalar interest $\psi(\theta)$ $\Rightarrow$ Jeff $= \pi(\varphi) = |j_{\varphi\varphi}(\varphi) = -\ell_{\varphi\varphi}(\theta)|^{1/2}$ Rescale so $j_{\varphi\varphi}(\hat{\varphi}^\circ) = I$. Use $\pi(\varphi)$ only on curve $C_\psi = \{\varphi : \psi(\varphi) = \hat{\psi}^\circ\}$. One dim. integration: $O(n^{-1})$ if $\psi(\varphi)$ linear

Discussion:

1. $f(y-\theta)$ with Jeffreys $\Rightarrow$ Exact inference, Reproducibility

2. $f(y;\theta)$, vector θ, y°, regular, continuity, indep. components $\Rightarrow$ Unique quantile representation $y = y(\theta; z)$ $\Rightarrow$ Directions $V = (v_1, \ldots, v_p) = \left.\frac{dy}{d\theta}\right|_{y^\circ, \hat{\theta}^\circ}$ where y "measures" θ at y° $\Rightarrow$ Canonical $\varphi(\theta) = \left.\frac{d\ell(\theta, y)}{dV}\right|_{y^\circ}$ of exponential model

3. $\ell^\circ(\theta)$, $\varphi(\theta)$, data y°, scalar interest $\psi(\theta)$ $\Rightarrow$ Unique $O(n^{-3/2})$ p-value fn $p(\theta)$ (Suff'cy; Ancillarity not needed/wanted) $\Rightarrow$ Unique reproducible inference

4. $\ell^\circ(\theta)$, $\varphi(\theta)$, data y°, scalar interest $\psi(\theta)$ $\Rightarrow$ Jeff $= \pi(\varphi) = |j_{\varphi\varphi}(\varphi) = -\ell_{\varphi\varphi}(\theta)|^{1/2}$ Rescale so $j_{\varphi\varphi}(\hat{\varphi}^\circ) = I$ Use $\pi(\varphi)$ only on curve $C_\psi = \{\varphi : \psi(\varphi) = \hat{\psi}^\circ\}$ One dim. integration: $O(n^{-1})$ if $\psi(\varphi)$ linear

5. If $\psi(\varphi)$ is curved $\Rightarrow$ simple adjustment to $\pi(\varphi)$

Discussion:

1. $f(y-\theta)$ with Jeffreys $\Rightarrow$ Exact inference, Reproducibility

2. $f(y;\theta)$, vector θ, y°, regular, continuity, indep. components $\Rightarrow$ Unique quantile representation $y = y(\theta; z)$ $\Rightarrow$ Directions $V = (v_1, \ldots, v_p) = \left.\frac{dy}{d\theta}\right|_{\hat\theta^\circ}^{y^\circ}$ where y "measures" θ at y° $\Rightarrow$ Canonical $\varphi(\theta) = \left.\frac{d\ell(\theta, y)}{dV}\right|_{y^\circ}$ of exponential model

3. $\ell^\circ(\theta)$, $\varphi(\theta)$, data y°, scalar interest $\psi(\theta)$ $\Rightarrow$ Unique $O(n^{-3/2})$ p-value fn $p(\theta)$ (Suff'cy; Ancillarity not needed/wanted) $\Rightarrow$ Unique reproducible inference

4. $\ell^\circ(\theta)$, $\varphi(\theta)$, data y°, scalar interest $\psi(\theta)$ $\Rightarrow$ Jeff $= \pi(\varphi) = |j_{\varphi\varphi}(\varphi) = -\ell_{\varphi\varphi}(\theta)|^{1/2}$ Rescale so $j_{\varphi\varphi}(\hat\varphi^\circ) = I$. Use $\pi(\varphi)$ only on curve $C_\psi = \{\varphi : \psi(\varphi) = \hat\psi^\circ\}$. One dim. integration: $O(n^{-1})$ if $\psi(\varphi)$ linear

5. If $\psi(\varphi)$ is curved $\Rightarrow$ simple adjustment to $\pi(\varphi)$

6. Geometry: Exp'l model; can var u; can. par. φ

Discussion:

1. $f(y-\theta)$ with Jeffreys $\Rightarrow$ Exact inference, Reproducibility

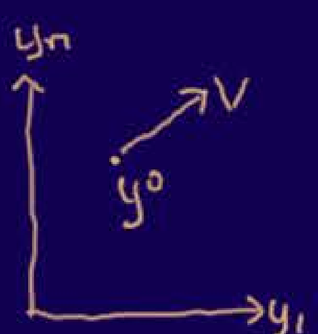
2. $f(y;\theta)$, vector θ, y°, regular, continuity, indep. components $\Rightarrow$ Unique quantile representation $y = y(\theta; z)$ $\Rightarrow$ Directions $V = (v_1, \ldots, v_p) = \left.\dfrac{dy}{d\theta}\right|_{\hat{\theta}^\circ}^{y^\circ}$ where y "measures" θ at y° $\Rightarrow$ Canonical $\varphi(\theta) = \left.\dfrac{d\ell(\theta; y)}{dV}\right|_{y^\circ}$ of exponential model

3. $\ell^\circ(\theta)$, $\varphi(\theta)$, data y°, scalar interest $\psi(\theta)$ $\Rightarrow$ Unique $O(n^{-3/2})$ p-value fn $p(\theta)$ (Suff'cy; Ancillarity not needed / wanted) $\Rightarrow$ Unique reproducible inference

4. $\ell^\circ(\theta)$, $\varphi(\theta)$, data y°, scalar interest $\psi(\theta)$ $\Rightarrow$ Jeff $= \pi(\varphi) = |j_{\varphi\varphi}(\varphi) = -\ell_{\varphi\varphi}(\theta)|^{1/2}$ Rescale so $j_{\varphi\varphi}(\hat{\varphi}^\circ) = I$. Use $\pi(\varphi)$ only on curve $C_\psi = \{\varphi : \psi(\varphi) = \hat{\psi}^\circ\}$. One dim. integration: $O(n^{-1})$ if $\psi(\varphi)$ linear

5. If $\psi(\varphi)$ is curved $\Rightarrow$ simple adjustment to $\pi(\varphi)$

6. Geometry: Exp'l model; can var u; can. par. φ



Data Space

Discussion:

1. $f(y-\theta)$ with Jeffreys $\Rightarrow$ Exact inference, Reproducibility

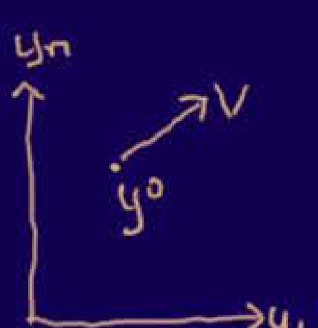
2. $f(y;\theta)$, vector θ, y°, regular, continuity, indep. components $\Rightarrow$ Unique quantile representation $y = y(\theta; z)$ $\Rightarrow$ Directions $V = (v_1, \ldots, v_p) = \left.\dfrac{dy}{d\theta}\right|_{\hat\theta^\circ}^{y^\circ}$ where y "measures" θ at y° $\Rightarrow$ Canonical $\varphi(\theta) = \left.\dfrac{d\ell(\theta, y)}{dV}\right|_{y^\circ}$ of exponential model

3. $\ell^\circ(\theta)$, $\varphi(\theta)$, data y°, scalar interest $\psi(\theta)$ $\Rightarrow$ Unique $O(n^{-3/2})$ p-value fn $p(\theta)$ (Suff'cy; Ancillarity not needed/wanted) $\Rightarrow$ Unique reproducible inference

4. $\ell^\circ(\theta)$, $\varphi(\theta)$, data y°, scalar interest $\psi(\theta)$ $\Rightarrow$ Jeff $= \pi(\varphi) = |j_{\varphi\varphi}(\varphi) = -\ell_{\varphi\varphi}(\theta)|^{1/2}$ Rescale so $j_{\varphi\varphi}(\hat\varphi^\circ) = I$. Use $\pi(\varphi)$ only on curve $C_\psi = \{\varphi : \psi(\varphi) = \hat\psi^\circ\}$. One dim. integration: $O(n^{-1})$ if $\psi(\varphi)$ linear

5. If $\psi(\varphi)$ is curved $\Rightarrow$ simple adjustment to $\pi(\varphi)$

6. Geometry: Exp'l model; can var u; can. par. φ



Data Space

Can. var. space

Discussion:

1. $f(y-\theta)$ with Jeffreys $\Rightarrow$ Exact inference, Reproducibility

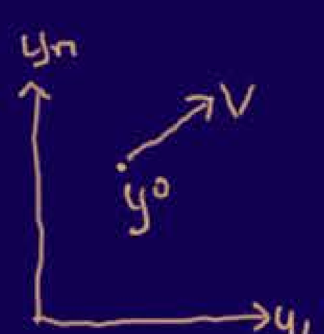
2. $f(y;\theta)$, vector θ, y°, regular, continuity, indep. components $\Rightarrow$ Unique quantile representation $y = y(\theta; z)$ $\Rightarrow$ Directions $V = (v_1, \ldots, v_p) = \left.\dfrac{dy}{d\theta}\right|_{\hat{\theta}^\circ}^{y^\circ}$ where y "measures" θ at y° $\Rightarrow$ Canonical $\varphi(\theta) = \left.\dfrac{d\,\ell(\theta, y)}{dV}\right|_{y^\circ}$ of exponential model

3. $\ell^\circ(\theta)$, $\varphi(\theta)$, data y°, scalar interest $\psi(\theta)$ $\Rightarrow$ Unique $O(n^{-3/2})$ p-value fn $p(\theta)$ (Suff'cy; Ancillarity not needed/wanted) $\Rightarrow$ Unique reproducible inference

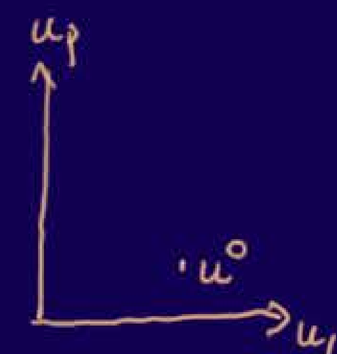
4. $\ell^\circ(\theta)$, $\varphi(\theta)$, data y°, scalar interest $\psi(\theta)$ $\Rightarrow$ Jeff $= \pi(\varphi) = |j_{\varphi\varphi}(\varphi) = -\ell_{\varphi\varphi}(\theta)|^{1/2}$ Rescale so $j_{\varphi\varphi}(\hat{\varphi}^\circ) = I$. Use $\pi(\varphi)$ only on curve $C_\psi = \{\varphi : \psi(\varphi) = \hat{\psi}^\circ\}$. One dim. integration: $O(n^{-1})$ if $\psi(\varphi)$ linear

5. If $\psi(\varphi)$ is curved $\Rightarrow$ simple adjustment to $\pi(\varphi)$

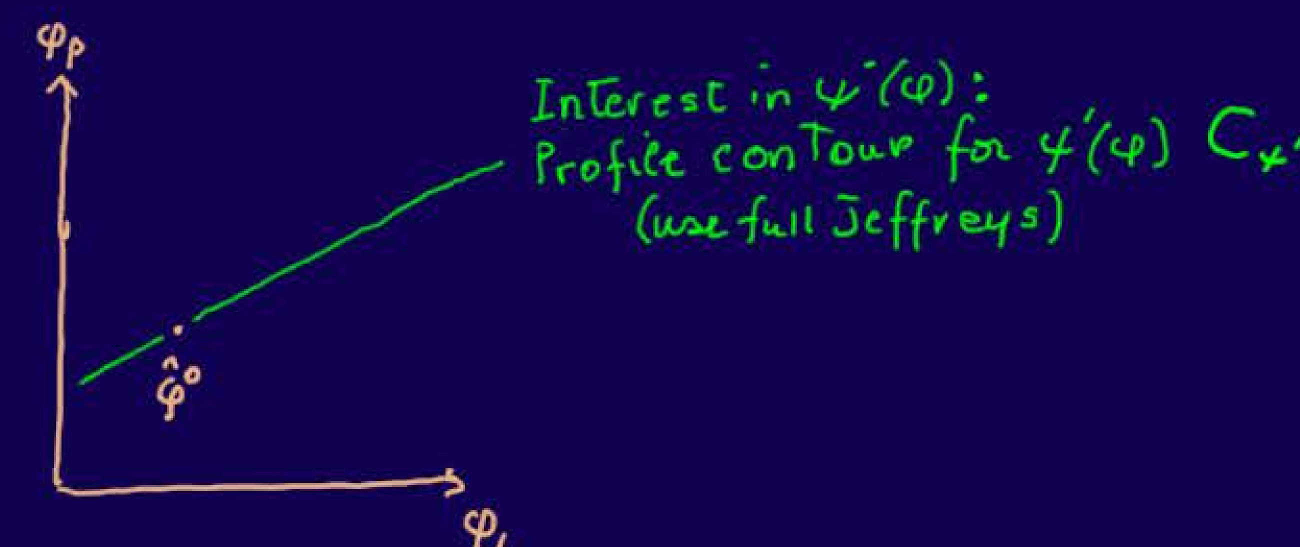
6. Geometry: Exp'l model; can var u; can. par. φ



Data Space



Can. var. space



Can. par. space

Discussion:

1. $f(y-\theta)$ with Jeffreys $\Rightarrow$ Exact inference, Reproducibility

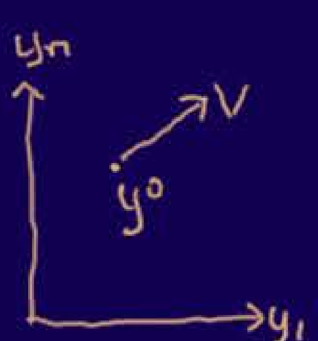
2. $f(y;\theta)$, vector θ, y°, regular, continuity, indep. components $\Rightarrow$ Unique quantile representation $y = y(\theta; z)$ $\Rightarrow$ Directions $V = (v_1, \ldots, v_p) = \left.\dfrac{dy}{d\theta}\right|_{\hat{\theta}^\circ}^{y^\circ}$ where y "measures" θ at y° $\Rightarrow$ Canonical $\varphi(\theta) = \left.\dfrac{d\ell(\theta, y)}{dV}\right|_{y^\circ}$ of exponential model

3. $\ell^\circ(\theta)$, $\varphi(\theta)$, data y°, scalar interest $\psi(\theta)$ $\Rightarrow$ Unique $O(n^{-3/2})$ p-value for $p(\theta)$ (Suff'cy; Ancillarity not needed/wanted) $\Rightarrow$ unique reproducible inference

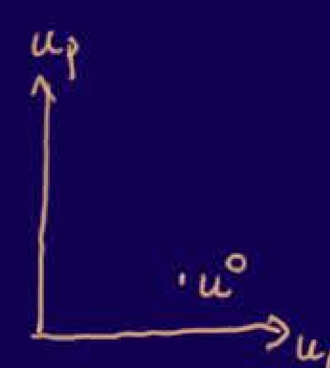
4. $\ell^\circ(\theta)$, $\varphi(\theta)$, data y°, scalar interest $\psi(\theta)$ $\Rightarrow$ Jeff $= \pi(\varphi) = |j_{\varphi\varphi}(\varphi) = -\ell_{\varphi\varphi}(\theta)|^{1/2}$ Rescale so $j_{\varphi\varphi}(\hat{\varphi}^\circ) = I$. Use $\pi(\varphi)$ only on curve $C_\psi = \{\varphi : \psi(\varphi) = \hat{\psi}^\circ\}$. One dim. integration: $O(n^{-1})$ if $\psi(\varphi)$ linear

5. If $\psi(\varphi)$ is curved $\Rightarrow$ simple adjustment to $\pi(\varphi)$

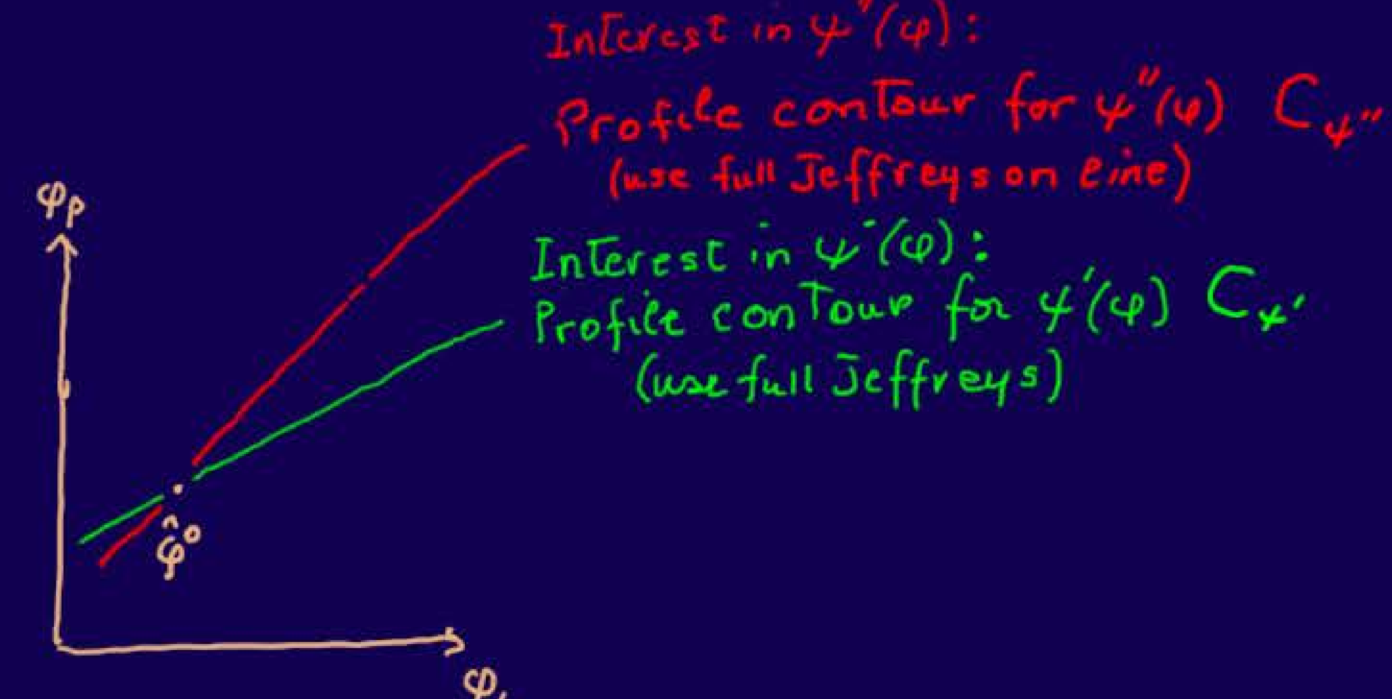
6. Geometry: Exp'l model; canvar u; can. par. φ



Data Space



Can. var. space



Can. par. space

Summary:

1. All info (2nd/3rd) for scalar $\psi(\theta)$ is on profile curve C_ψ (1 dim)

   use full Jeffreys on C_ψ ... <u>not</u> on full space

Summary:

1. All info (2nd/3rd) for scalar $\psi(\theta)$ is on profile curve C_ψ (1 dim)

   Use full Jeffreys on C_ψ ... not on full space

2. Different $\psi''(\theta)$ ... different curve $C_{\psi''}$

Summary:

1. All info (2nd/3rd) for scalar $\psi(\theta)$ is on profile curve C_ψ (1 dim)

   use full Jeffreys on C_ψ ... <u>not</u> on full space

2. <u>Different</u> $\psi''(\theta)$ ... <u>different</u> curve $C_{\psi''}$

3. Gives <u>2nd</u> order inference

Summary:

1. All info (2nd/3rd) for scalar $\psi(\theta)$ is on profile curve C_ψ (1 dim)

   use full Jeffreys on C_ψ ... not on full space

2. Different $\psi''(\theta)$ ... different curve $C_{\psi''}$

3. Gives 2nd order inference

   but if "curved", simple curvature adjustment available for 2nd order

Summary:

1. All info (2nd/3rd) for scalar $\psi(\theta)$ is on profile curve C_ψ (1 dim)
   use full Jeffreys on C_ψ ... not on full space

2. Different $\psi^{*}(\theta)$ ... different curve C_{ψ^*}

3. Gives 2nd order inference
   but if "curved", simple curvature adjustment available for 2nd order

Thank you ....