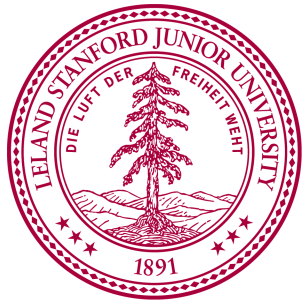
Sailing Through Data: Discoveries and Mirages

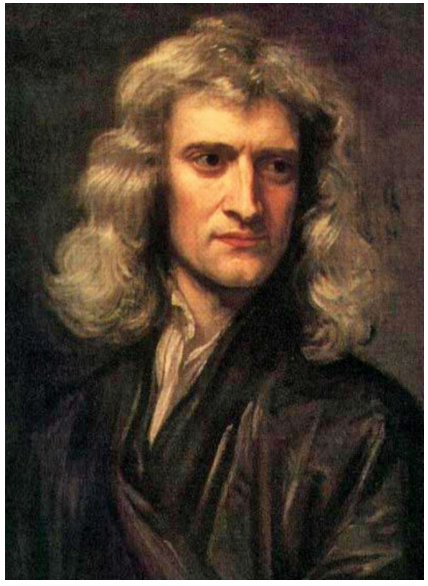
Emmanuel Candès, *Stanford University*



2018 Machine Learning Summer School, Buenos Aires, June 2018

Mathematics and Nature

Galileo



Newton

Euler

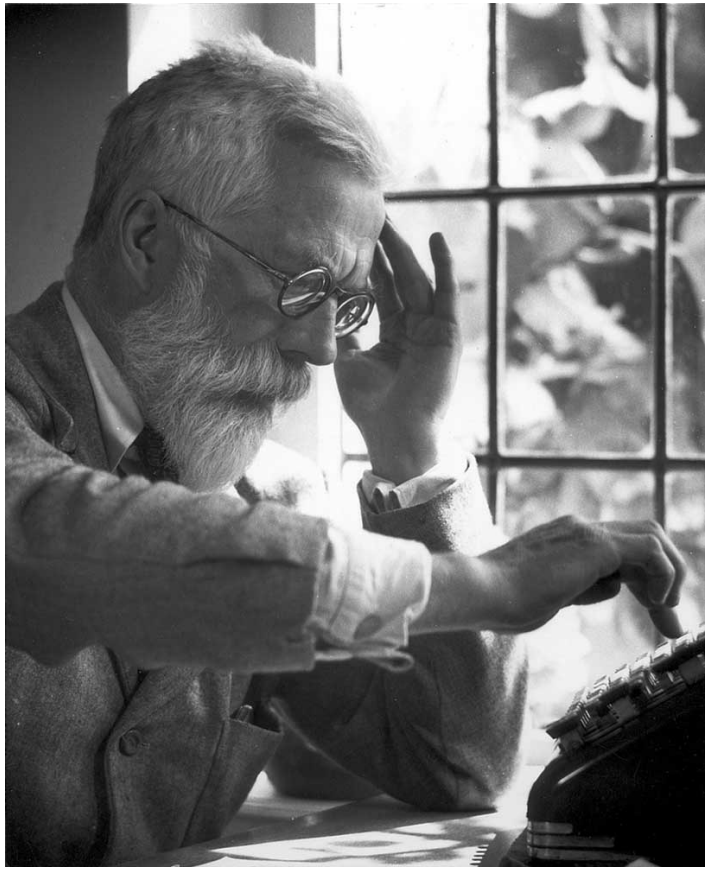
The laws of Nature are written in the language of mathematics ... the symbols are triangles, circles and other geometrical figures, without whose help it is impossible to comprehend a single word (Galileo)

‘The statistical century’ (B. Efron)

The scientific method

- Formulate hypothesis
- Collect data to test predictions
- Falsify or corroborate

Operationalization: R. A. Fisher



Ronald Fisher (1890-1962)

Statistical thinking in action: the Salk vaccine trial

Group	Size	Rate per 100,000
Treatment	200,000	28
Control	200,000	71

Source: Thomas Francis, Jr., *American Journal of Public Health (1955)*

NEW YORK, WEDNESDAY, APRIL 13, 1955. Times Square, New York 36, N. Y. Telephone Lackawanna 4-1000 FIVE CENTS

GH COURT HEARS
)UTH WILL DEFY
UICK END TO BIAS

dual Approaches Urged
Integration of Schools—
egro Lawyers Opposed

y LUTHER A. HUSTON
Special to The New York Times.

ASHINGTON, April 12—
tesmen for South Carolina
Virginia told the Supreme
t today that their people
ld not obey a decree ordering
mmediate end to racial seg-
tion in the public schools.
hen Chief Justice Earl War-
asked S. E. Rogers, repre-
ing Clarendon County, S. C.,
s were willing to say that
'honest attempt" would be
e to conform to whatever
se the court might issue, Mr.
irs said:
et's get that word 'honest'
of there. It would depend
the kind of decree. The

SALK POLIO VACCINE PROVES SUCCESS; MILLIONS WILL BE IMMUNIZED SOON; CITY SCHOOLS BEGIN SHOTS APRIL 25

TRIAL DATA GIVEN

Efficacy of 80 to 90% Shown—Salk Sees Further Advance

Abstract of report, summary of data on tests, Page 22.

By WILLIAM L. LAURENCE
Special to The New York Times.

ANN ARBOR, Mich., April 12—The world learned today that its hopes for finding an effective weapon against paralytic polio had been realized.

The Economist

OCTOBER 19TH–25TH 2013 Economist.com

Washington's lawyer surplus

How to do a nuclear deal with Iran

Investment tips from Nobel economists

Junk bonds are back

The meaning of Sachin Tendulkar

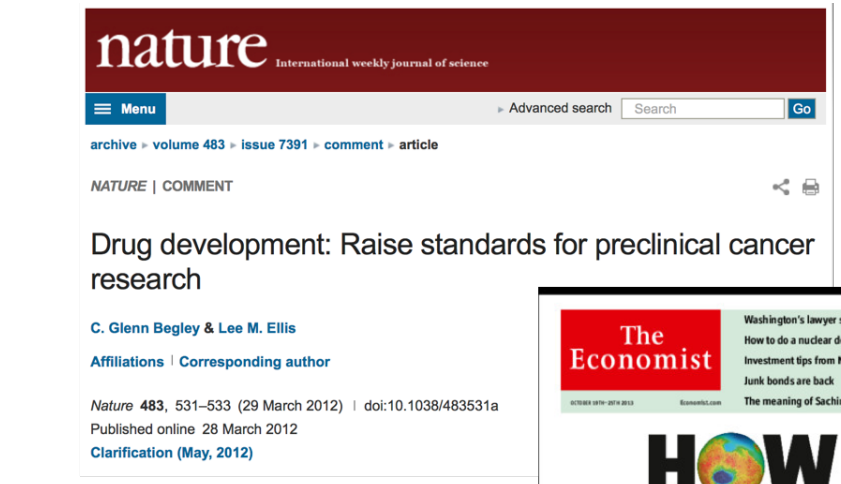
HOW SCIENCE GOES WRONG.

99
Einsteinium

The replicability crisis

Begley and Ellis, *Nature* (2012)

- Amgen could only replicate 6 of 53 studies they considered landmarks in basic cancer science
- HealthCare could only replicate about 25% of 67 seminal studies
- Systematic attempts to replicate widely cited priming experiments have failed



Media coverage...

SECTIONS | The New York Times | SUBSCRIBE NOW | LOG IN

A College Town Gets Ready for Its Moment Under No Sun

TRILOBITES
This Beautiful Parasitic Bird Could Soon Turn Up in Your Yard

TRILOBITES
Drumming Cockatoos and the Rhythms of Love

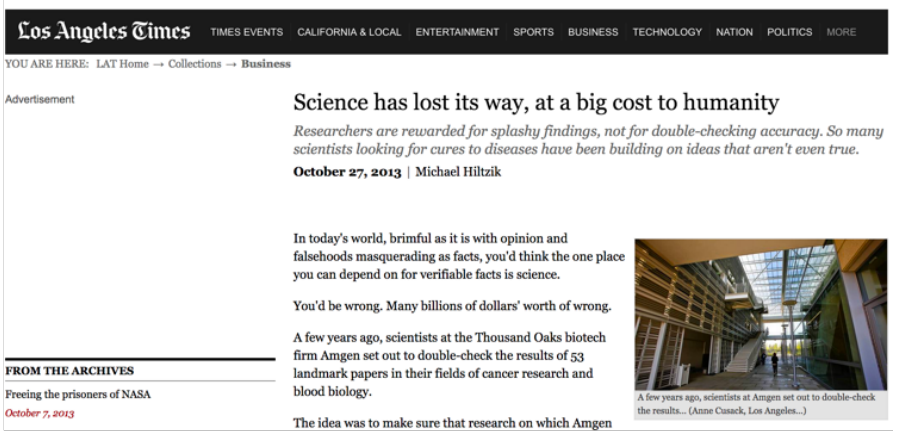
SCIENCE

New Truths That Only One Can See

George Johnson
RAW DATA JAN. 20, 2014

Carl Wiens

Since 1955, The Journal of Irreproducible Results has offered "spoofs, parodies, whimsies, burlesques, lampoons and satires" about life in the laboratory. Among its greatest hits: "Acoustic Oscillations in Jell-O, With and Without Fruit, Subjected to Varying Levels of Stress" and "Utilizing Infinite Loops to Compute an Approximate Value of Infinity." The good-natured jibes are a backhanded celebration of science. What really goes on in the lab is, by implication, of a loftier, more serious nature.



Los Angeles Times | TIMES EVENTS | CALIFORNIA & LOCAL | ENTERTAINMENT | SPORTS | BUSINESS | TECHNOLOGY | NATION | POLITICS | MORE

YOU ARE HERE: LAT Home → Collections → Business

Advertisement

Science has lost its way, at a big cost to humanity

Researchers are rewarded for splashy findings, not for double-checking accuracy. So many scientists looking for cures to diseases have been building on ideas that aren't even true.

October 27, 2013 | Michael Hiltzik

In today's world, brimful as it is with opinion and falsehoods masquerading as facts, you'd think the one place you can depend on for verifiable facts is science.

You'd be wrong. Many billions of dollars' worth of wrong.

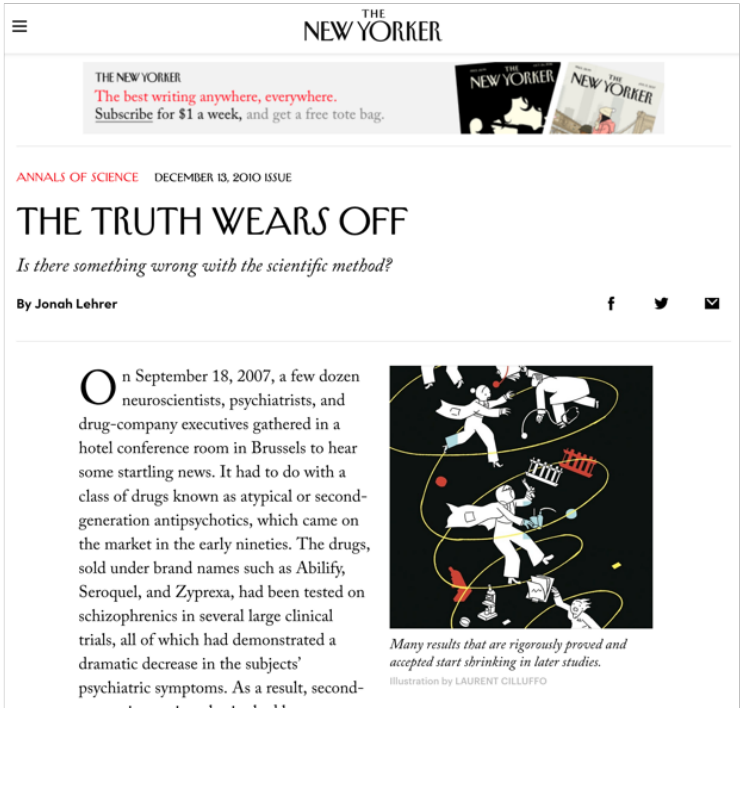
A few years ago, scientists at the Thousand Oaks biotech firm Amgen set out to double-check the results of 53 landmark papers in their fields of cancer research and blood biology.

The idea was to make sure that research on which Amgen

A few years ago, scientists at Amgen set out to double-check the results... (Anne Cusack, Los Angeles...)

FROM THE ARCHIVES

Freeing the prisoners of NASA
October 7, 2013



THE NEW YORKER

THE NEW YORKER
The best writing anywhere, everywhere.
Subscribe for $1 a week, and get a free tote bag.

ANNALS OF SCIENCE DECEMBER 13, 2010 ISSUE

THE TRUTH WEARS OFF

Is there something wrong with the scientific method?

By Jonah Lehrer

On September 18, 2007, a few dozen neuroscientists, psychiatrists, and drug-company executives gathered in a hotel conference room in Brussels to hear some startling news. It had to do with a class of drugs known as atypical or second-generation antipsychotics, which came on the market in the early nineties. The drugs, sold under brand names such as Abilify, Seroquel, and Zyprexa, had been tested on schizophrenics in several large clinical trials, all of which had demonstrated a dramatic decrease in the subjects' psychiatric symptoms. As a result, second-

Many results that are rigorously proved and accepted start shrinking in later studies.

Illustration by LAURENT CILLUFFO

Personal and societal concern

Snippets from media

"Significance chasing"

"Publication bias"

"Selective reporting"

Why most published research findings are false (Ioannidis, '05)

Personal and societal concern

Snippets from media

"Significance chasing"

"Publication bias"

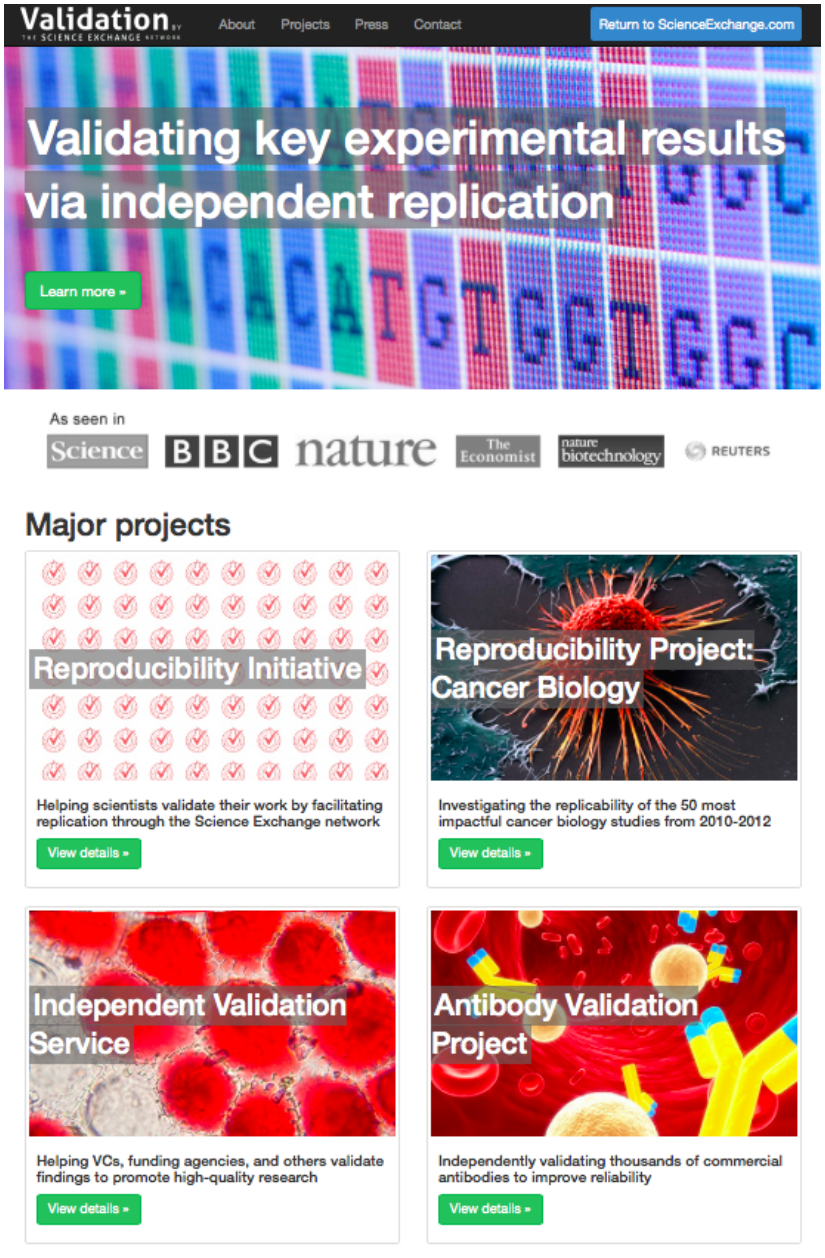
"Selective reporting"

Why most published research findings are false (Ioannidis, '05)

Great danger in seeing erosion of public confidence in science

Scientific community is responding

Response: reproducibility intiatives



Reproducibility Initiative

`http://validation.scienceexchange.com/`

Response: editorial policies

ANNOUNCEMENT

Reducing our irreproducibility

Over the past year, *Nature* has published a string of articles that highlight failures in the reliability and reproducibility of published research (collected and freely available at go.nature.com/huhbyr). The problems arise in laboratories, but journals such as this one compound them when they fail to exert sufficient scrutiny over the results that they publish, and when they do not publish enough information for other researchers to assess results properly.

From next month, *Nature* and the Nature research journals will introduce editorial measures to address the problem by improving the consistency and quality of reporting in life-sciences articles. To ease the interpretation and improve the reliability of published results we will more systematically ensure that key methodological details are reported, and we will give more space to methods sections. We will examine statistics more closely and encourage authors to be transparent, for example by including their raw data.

Central to this initiative is a checklist intended to prompt authors to disclose technical and statistical information in their submissions, and to encourage referees to consider aspects important for research reproducibility (go.nature.com/oloeip). It was developed after discussions with researchers on the problems that lead to irreproducibility, including workshops organized last year by US National Institutes of Health (NIH) institutes. It also draws on published concerns about reporting standards (or the lack of them) and the collective experience of editors at Nature journals.

The checklist is not exhaustive. It focuses on a few experimental and analytical design elements that are crucial for the interpretation of research results but are often reported incompletely. For example, authors will need to describe methodological parameters that can introduce bias or influence robustness, and provide precise characterization of key reagents that may be subject to biological variability, such as cell lines and antibodies. The checklist also consolidates existing policies about data deposition and presentation.

We will also demand more precise descriptions of statistics, and we will commission statisticians as consultants on certain papers, at the editor's discretion and at the referees' suggestion.

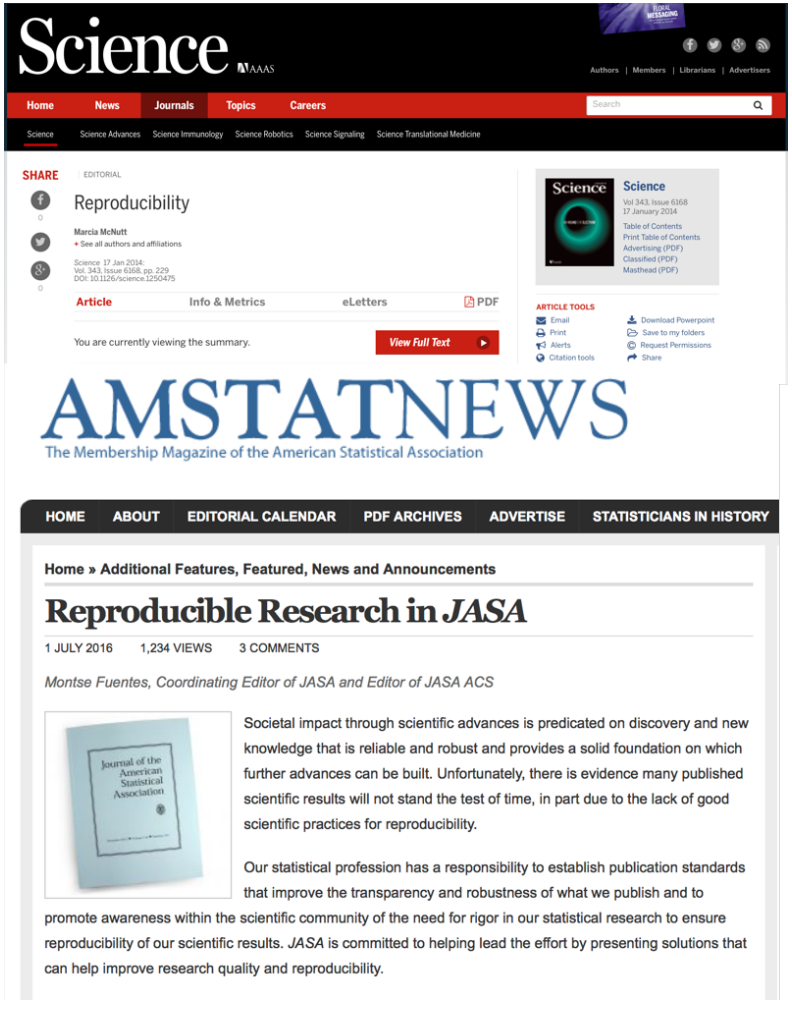
We recognize that there is no single way to conduct an experimental study. Exploratory investigations cannot be done with the same level of statistical rigour as hypothesis-testing studies. Few academic laboratories have the means to perform the level of validation required, for example, to translate a finding from the laboratory to the clinic. However, that should not stand in the way of a full report of how a study was designed, conducted and analysed that will allow reviewers and readers to adequately interpret and build on the results.

To allow authors to describe their experimental design and methods in as much detail as necessary, the participating journals, including *Nature*, will abolish space restrictions on the methods section.

To further increase transparency, we will encourage authors to provide tables of the data behind graphs and figures. This builds on our established data-deposition policy for specific experiments and large data sets. The source data will be made available directly from the figure legend, for easy access. We continue to encourage authors to share detailed methods and reagent descriptions by depositing protocols in Protocol Exchange (www.nature.com/protocolexchange), an open resource linked from the primary paper.

Renewed attention to reporting and transparency is a small step. Much bigger underlying issues contribute to the problem, and are beyond the reach of journals alone. Too few biologists receive adequate training in statistics and other quantitative aspects of their subject. Mentoring of young scientists on matters of rigour and transparency is inconsistent at best. In academia, the ever increasing pressures to publish and chase funds provide little incentive to pursue studies and publish results that contradict or confirm previous papers. Those who document the validity or irreproducibility of a published piece of work seldom get a welcome from journals and funders, even as money and effort are wasted on false assumptions.

Tackling these issues is a long-term endeavour that will require the commitment of funders, institutions, researchers and publishers. It is encouraging that NIH institutes have led community discussions on this topic and are considering their own recommendations. We urge others to take note of these and of our initiatives, and do whatever they can to improve research reproducibility. ■



Response: best practices

Research Reproducibility, Replicability, Reliability
A Speech by Ralph J. Cicerone, President
National Academy of Sciences
Presented at the Academy's 152nd Annual Meeting
April 27, 2015

President's Council of Advisors on Science and Technology (PCAST)
Public Meeting Agenda
January 31, 2014

National Academy of Sciences (NAS)
2101 Constitution Avenue, NW
Washington, DC

Lecture Room

9:00 am **Welcome from PCAST Co-Chairs**
John Holdren, Assistant to the President for Science and Technology; Director, Office of Science and Technology Policy (OSTP); Co-Chair, PCAST
Eric Lander, Co-Chair, PCAST

9:05 am **Improving Scientific Reproducibility in an Age of International Competition and Big Data I: Researchers**
Glenn Begley, Chief Scientific Officer and Senior Vice-President R&D, TetraLogic Pharmaceuticals
Donald Berry, Professor, Department of Biostatistics, University of Texas MD Anderson Cancer Center
Daniel MacArthur, Assistant Professor, Harvard Medical School and Massachusetts General Hospital and Associate Member, Broad Institute of Harvard and MIT

10:00 am **Improving Scientific Reproducibility in an Age of International Competition and Big Data II: Editors**
Marcia McNutt, Editor-In-Chief, *Science*
Philip Campbell, Editor-In-Chief, *Nature* and Nature Publishing Group
Véronique Kiermer, Executive Editor and Head of Researchers Services, Nature Publishing Group

Statistical Challenges in Assessing and Fostering the Reproducibility of Scientific Results

A Workshop
of the Committee on Applied and Theoretical Statistics

NATIONAL RESEARCH COUNCIL
OF THE NATIONAL ACADEMIES

The replicability issue

Many different components

1. Publishing culture
2. Granting agencies culture
3. Computational reproducibility
4. Statistics: how to choose a finding?
   Statistical methodology enhancing replicability

Can only do 3 and 4
1 and 2 above pay grade

Why is this happening? A new scientific paradigm

Collect data first $\Longrightarrow$ Ask questions later

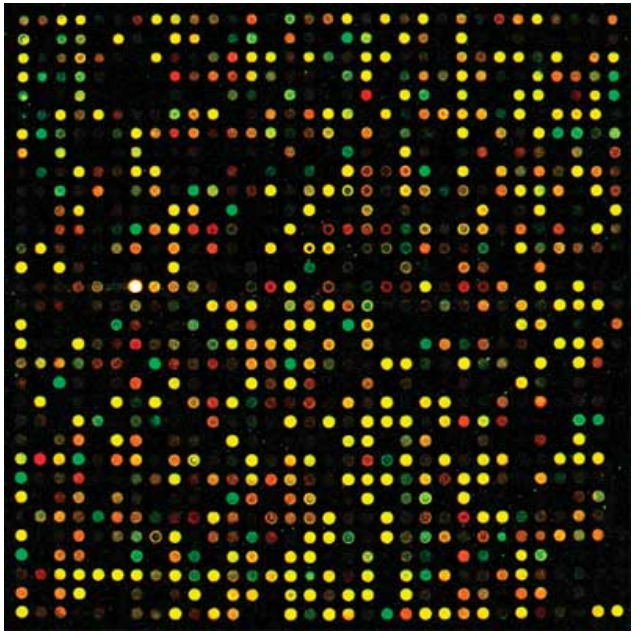
- Large data sets available prior to formulation of scientific hypotheses/theories
- Very different from hypothesis-driven research

Example from genomics

Historically, molecular biology was hypothesis-driven research

"Sometime in the 90's"

- Eruption of high-throughput technologies
- Enabled thousands of genes to be tested simultaneously for differential expression
  - Small # of samples
  - High # of variables
- Researchers begin to look everywhere



gene expression microarray

Complete revolution: from hypothesis- to data-driven research

Warnings: Sorić ('89) & Ioannidis ('05)

Statistical "Discoveries" and Effect-Size Estimation

BRANKO SORIĆ*

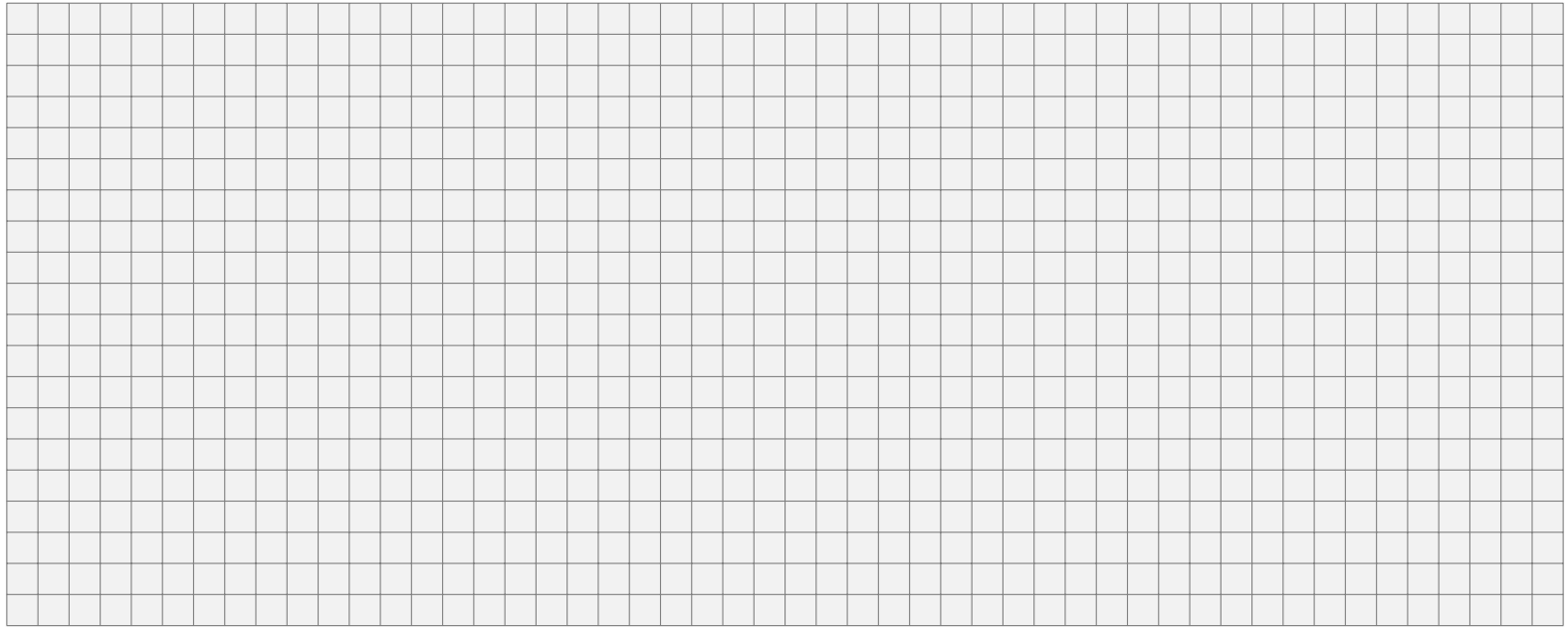
Essay

Why Most Published Research Findings Are False

John P. A. Ioannidis

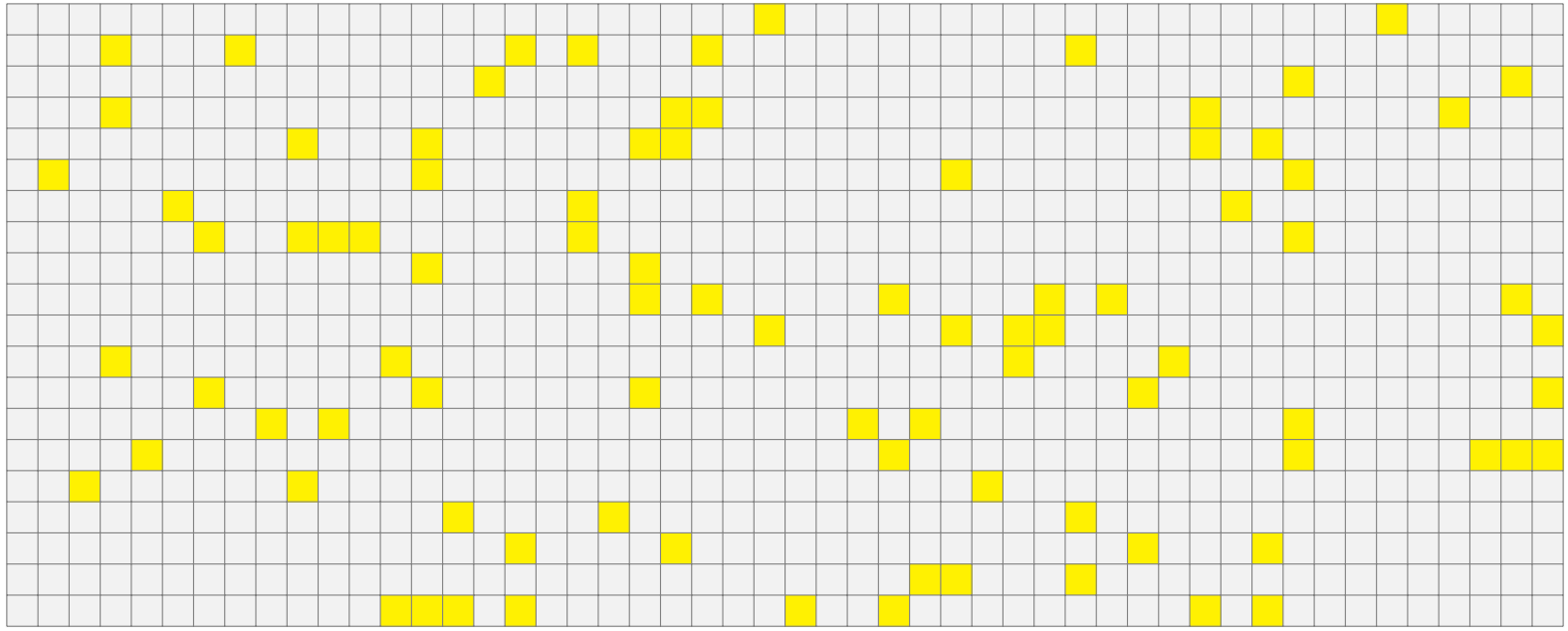
Ioannidis: "It can be proven that most claimed research findings are false"

False discovery rate (FDR), Benjamini-Hochberg '95



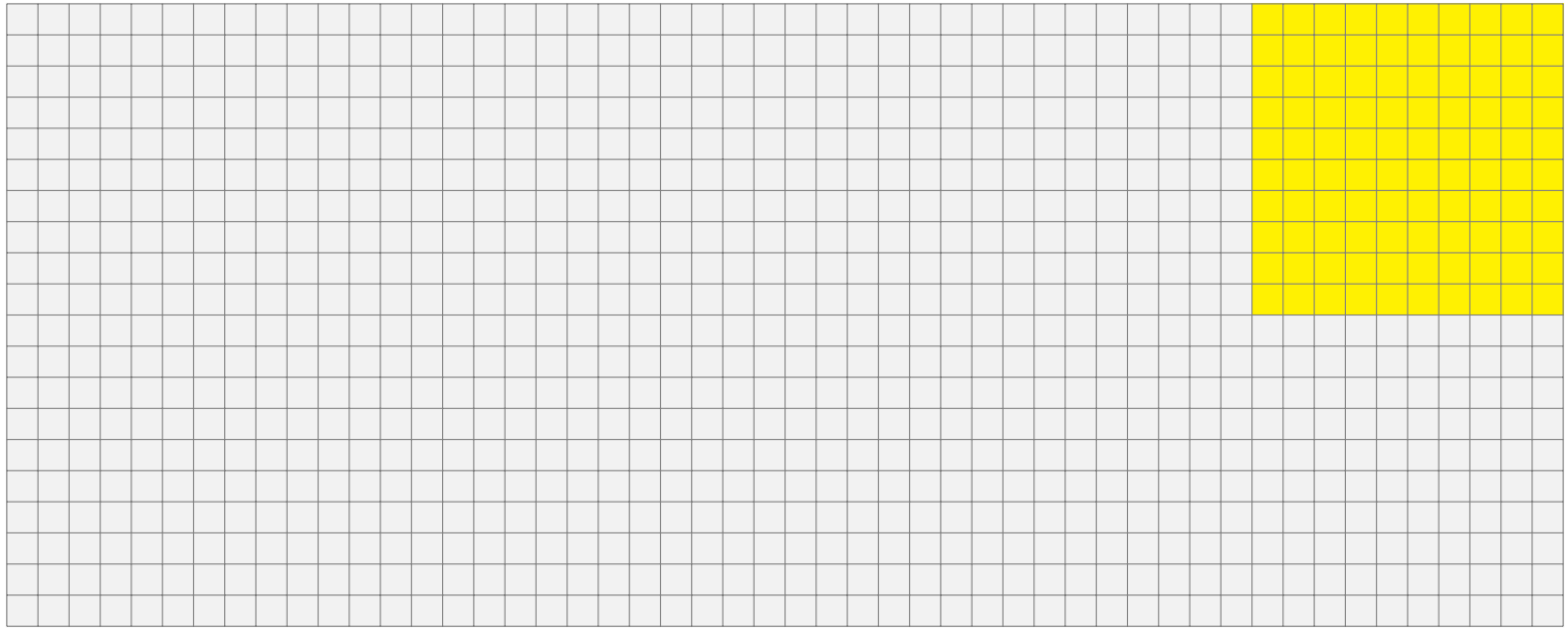
1000 hypotheses to test

False discovery rate (FDR), Benjamini-Hochberg '95



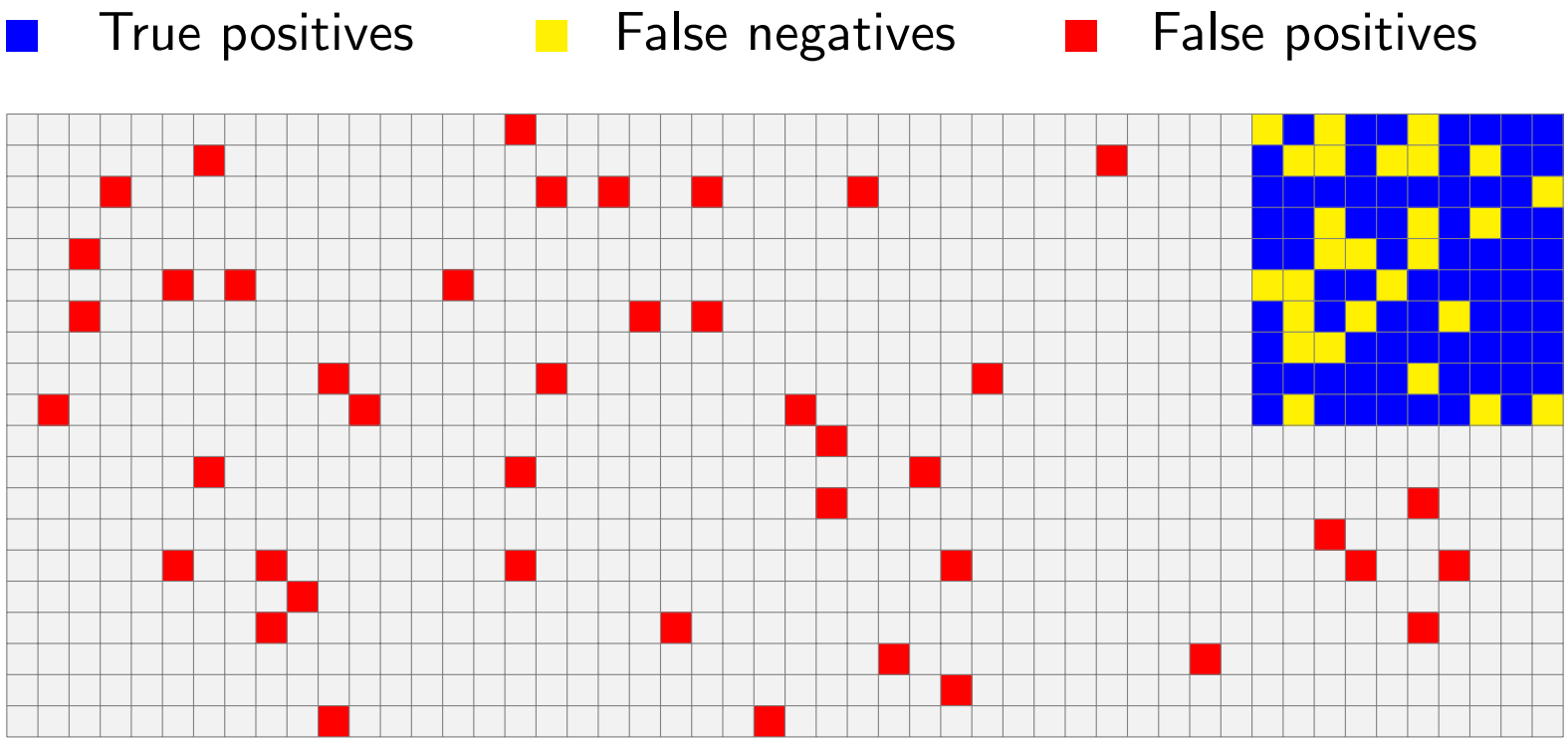
1000 hypotheses, 100 potential discoveries

False discovery rate (FDR), Benjamini-Hochberg '95



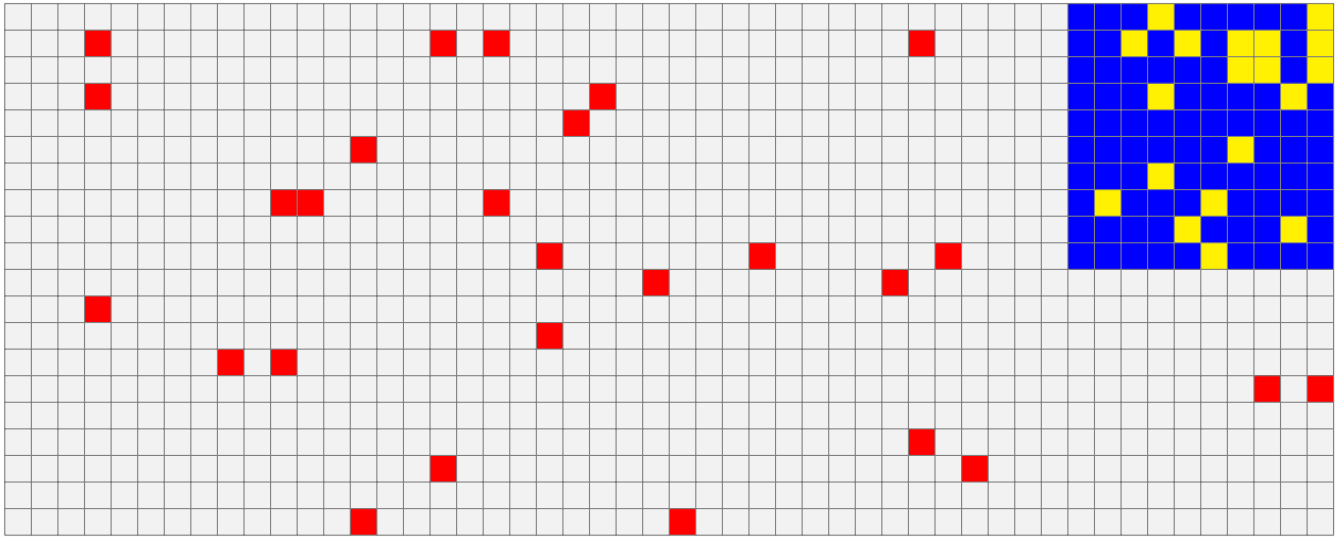
1000 hypotheses, 100 potential discoveries

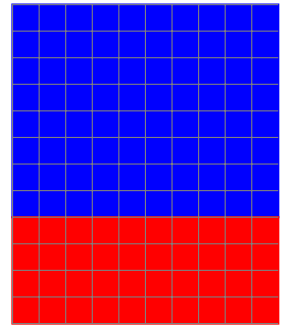
False discovery rate (FDR), Benjamini-Hochberg '95



False discovery rate (FDR), Benjamini-Hochberg '95

True positives False negatives False positives



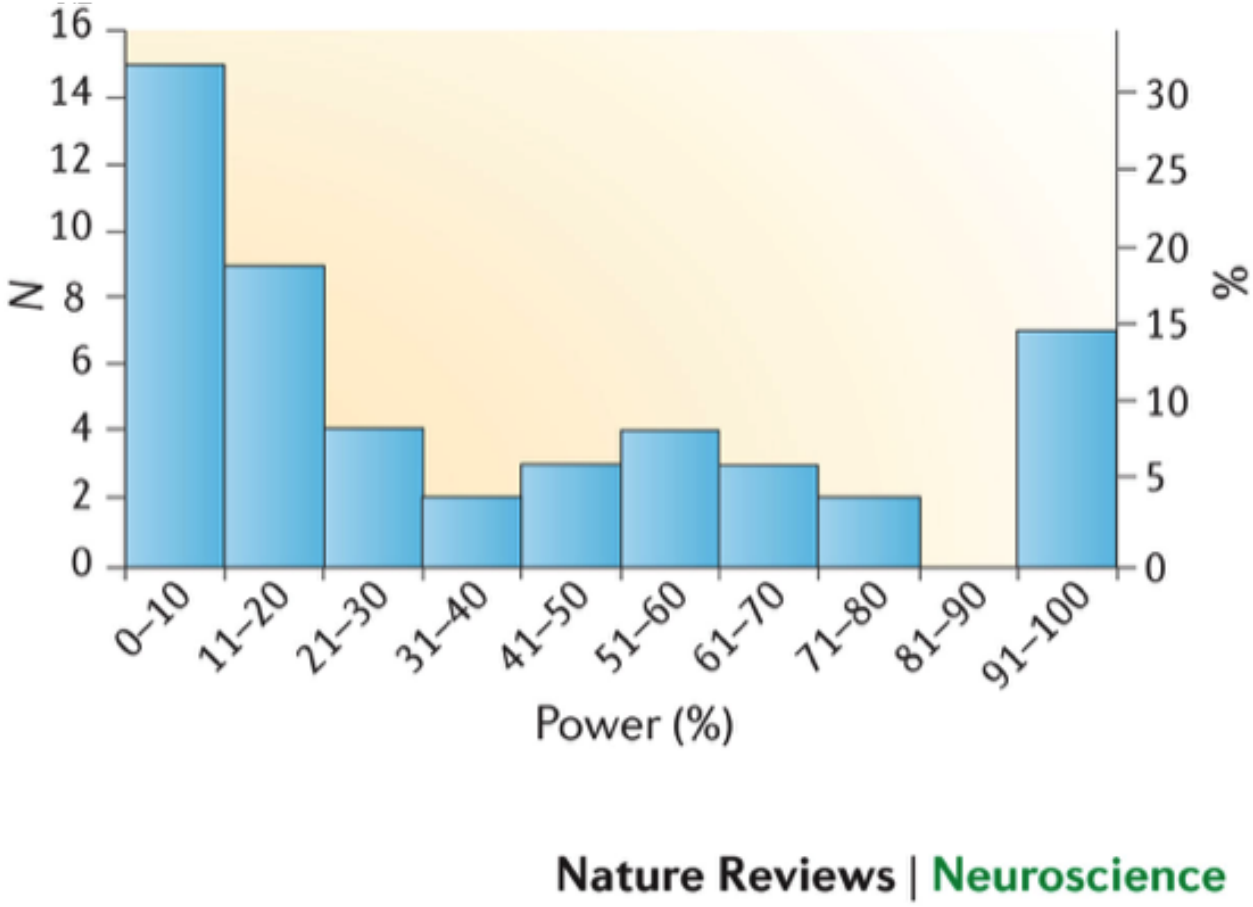


Selected

$$\text{FDR} = \mathbb{E}\left[\frac{\#\text{false positives}}{\#\text{selections}}\right]$$

Example: meta-analysis in neuroscience

Button et al. (2013) *Power failure: why small sample size undermines the reliability of neuroscience*



False Discovery Rate (FDR): Benjamini & Hochberg ('95)

$H_1, \dots H_n$ hypotheses subject to some testing procedure

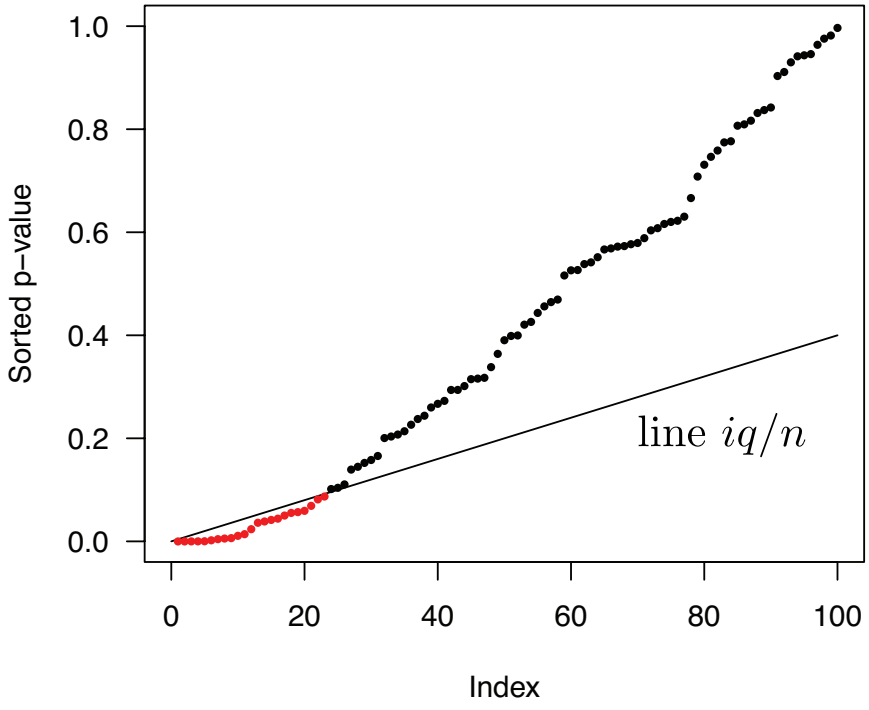
$$\text{FDR} = \mathbb{E}\left[\frac{\#\text{false discoveries}}{\#\text{discoveries}}\right] \qquad \text{'}0/0 = 0\text{'}$$

- Natural type I error
- Under independence (and PRDS) simple rules control FDR (BHq)
- Widely used $\longrightarrow$ enormous influence on medical research

FDR control with BHq (under independence)

FDR: expected proportion of false discoveries

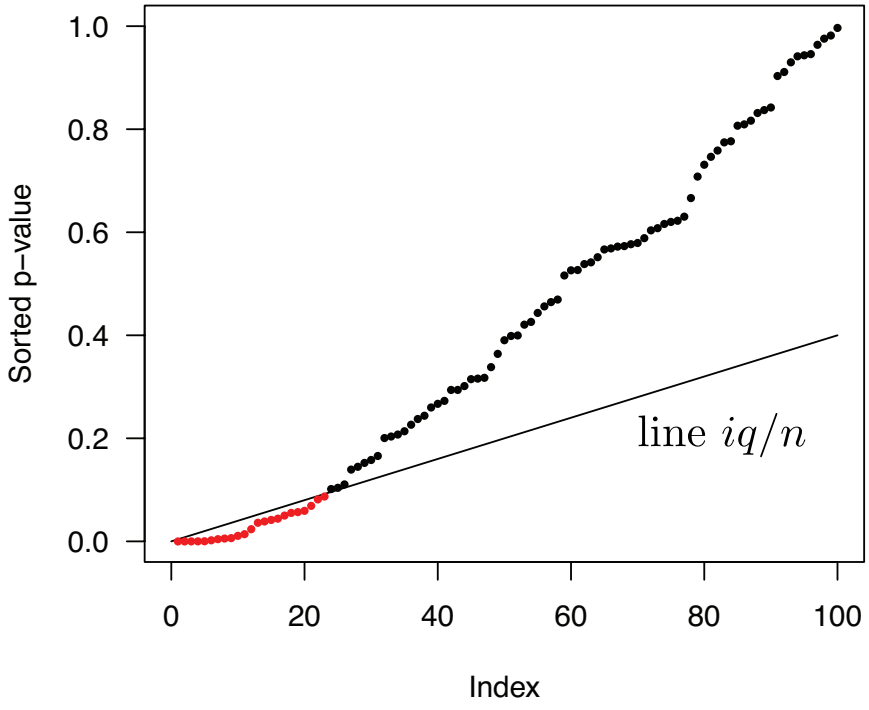
- Sorted p-values: $p_{(1)} \leq p_{(2)} \leq \ldots \leq p_{(n)}$ (from most to least significant)
- Target FDR q

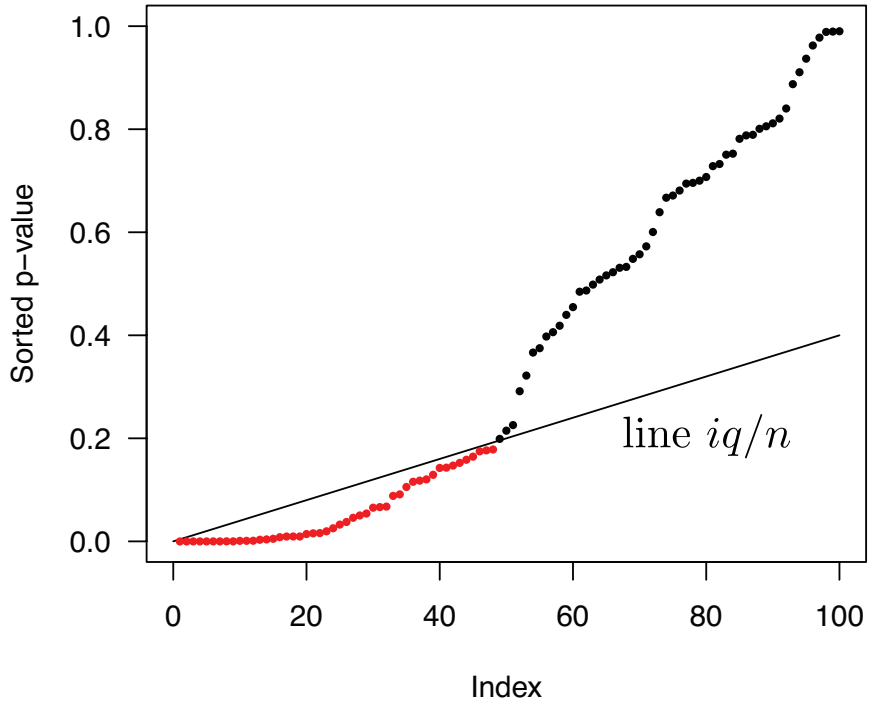


FDR control with BHq (under independence)

FDR: expected proportion of false discoveries

- Sorted p-values: $p_{(1)} \leq p_{(2)} \leq \ldots \leq p_{(n)}$ (from most to least significant)
- Target FDR q





The cut-off is adaptive to number of non-nulls

Not just about testing: confidence intervals

- Multiple parameters $\theta_1, \ldots, \theta_n$
- Regular 95% CI

$$\mathbb{E}\, 1(\text{CI}_i \text{ covers } \theta_i) \geq 0.95 \quad \Longrightarrow \quad \mathbb{E}\, \frac{\sum_{i=1}^{n} 1(\text{CI}_i \text{ covers } \theta_i)}{n} \geq 0.95$$

- Expected proportion is all right

Sorić's warning: after selection...

"In a large number of 95% confidence intervals, 95% of them contain the population parameter [...] but it would be wrong to imagine that the same rule also applies to a large number of 95% interesting confidence intervals"

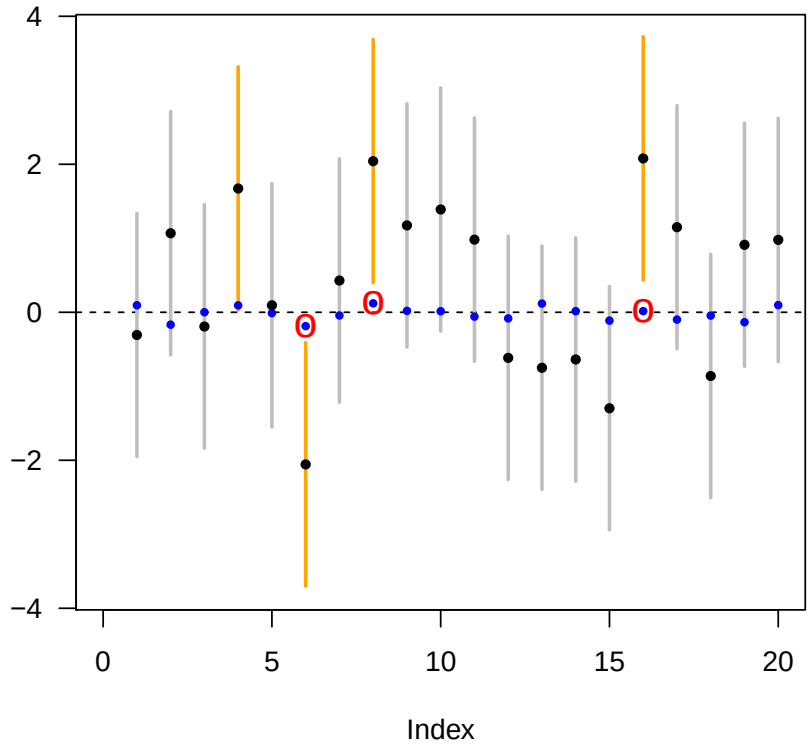
Sorić's warning: after selection...

> *"In a large number of 95% confidence intervals, 95% of them contain the population parameter [...] but it would be wrong to imagine that the same rule also applies to a large number of 95% interesting confidence intervals"*

90% CI

- 17/20 over all cover
- 1/4 over selected cover!

Selection bias
$\Longrightarrow$ will tend to fail when replicated



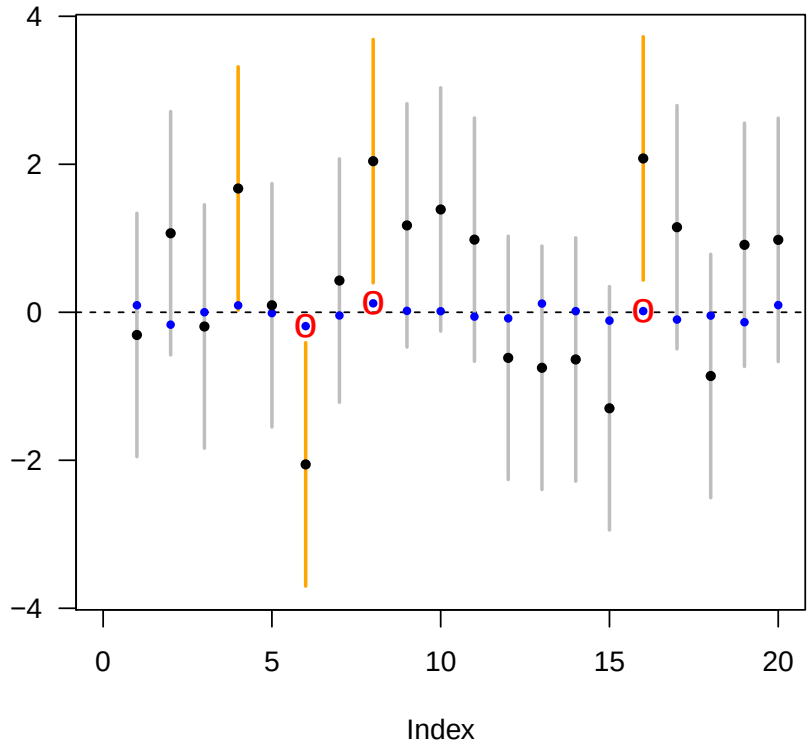
Sorić's warning: after selection...

> *"In a large number of 95% confidence intervals, 95% of them contain the population parameter [...] but it would be wrong to imagine that the same rule also applies to a large number of 95% interesting confidence intervals"*

90% CI

- 17/20 over all cover
- 1/4 over selected cover!

Selection bias
$\Longrightarrow$ will tend to fail when replicated



Statisticians's response

False coverage rate (Benjamini & Yekutieli, '05)

Historical success story: clinical trials

Testing of new drugs rigorous and safe thanks to

- pre-registered protocols
- pre-specified endpoints

... and thanks to sophisticated statistical methods

- sequential testing
- repeated hypothesis testing (alpha spending function)

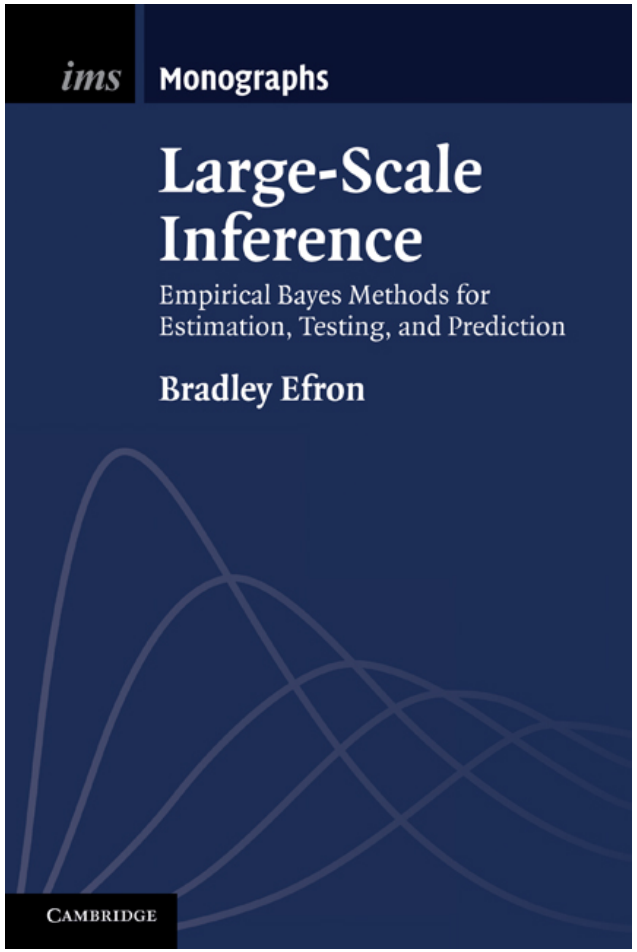
Positive outlook from gene expression studies

- Recognized multiplicity problem early
- Worked hand-in-hand with statisticians
- Beginning to rewrite statistical theory

Statistical Science
2003, Vol. 18, No. 1, 71–103

Multiple Hypothesis Testing in Microarray Experiments

Sandrine Dudoit, Juliet Popper Shaffer and Jennifer C. Boldrick



Summary: a new scientific paradigm

Textbook practice

(1) Select hypotheses/model/question
(2) Collect data
(3) Perform inference

Modern practice

(1) Collect data
(2) Select hypotheses/model/questions
(3) Perform inference

Elementary stats textbooks and researchers often ignore selection issues
$\rightsquigarrow$ inference completely wrong and misleading

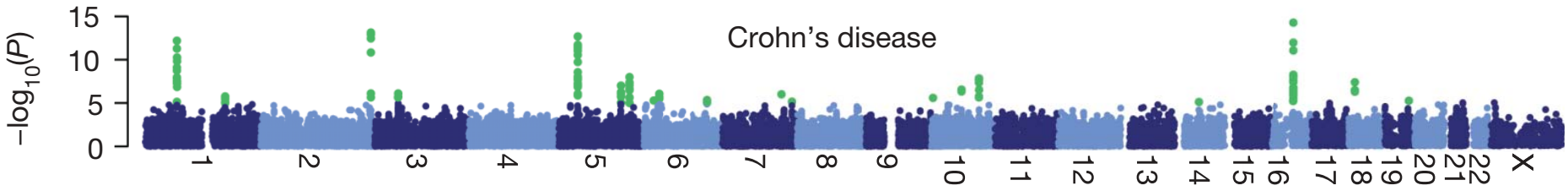
Knockoffs: Tools for Replicable Selections

Joint with R. Barber
and Y. Fan, L. Janson and J. Lv

Some data-driven scientific problems

- One response Y: phenotype; e.g. Crohn's disease status, cholesterol level
- Hundreds of thousands of variables X: genotype information

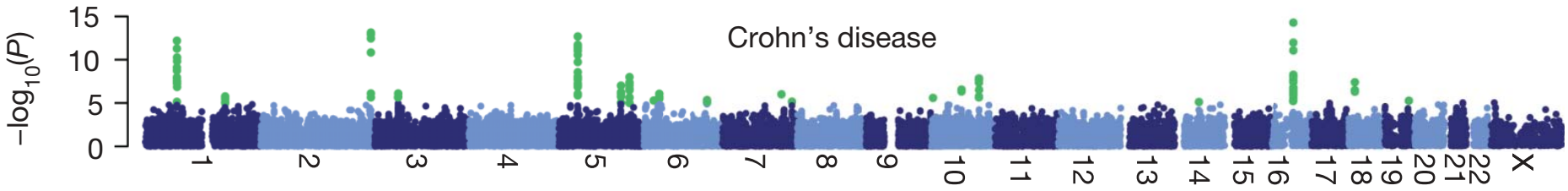
Ex. 1: which genetic variations affect traits, e.g. the risk of a disease?



Some data-driven scientific problems

- One response Y: phenotype; e.g. Crohn's disease status, cholesterol level
- Hundreds of thousands of variables X: genotype information

Ex. 1: which genetic variations affect traits, e.g. the risk of a disease?



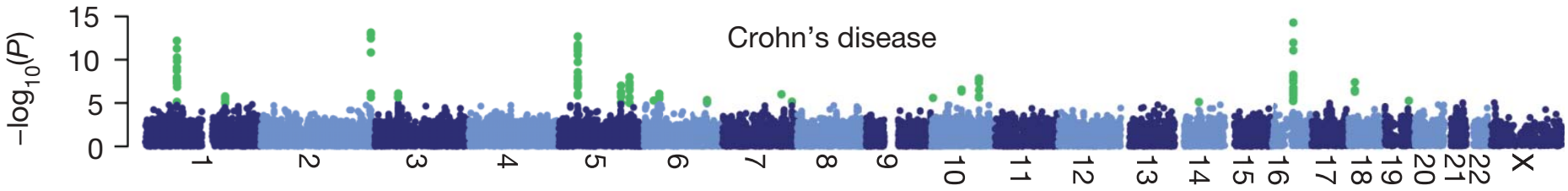
Ex. 2: which gene expression profiles help determine severity of a tumor?

Ex. 3: which factors/variables help determine whether a loan will be repaid?

Some data-driven scientific problems

- One response Y: phenotype; e.g. Crohn's disease status, cholesterol level
- Hundreds of thousands of variables X: genotype information

Ex. 1: which genetic variations affect traits, e.g. the risk of a disease?



Ex. 2: which gene expression profiles help determine severity of a tumor?

Ex. 3: which factors/variables help determine whether a loan will be repaid?

How can we select variables without too many false positives?
$\rightsquigarrow$ do not run into problem of irreproducibility

Formalizing the selection problem

Each with their X and Y

variables response

Thousands/millions of variables X

Which ones are important?

Distribution of $Y \mid X$ depends on X through which variables?

Formalizing the selection problem

Each with their X and Y

variables response

Thousands/millions of variables X

Which ones are important?

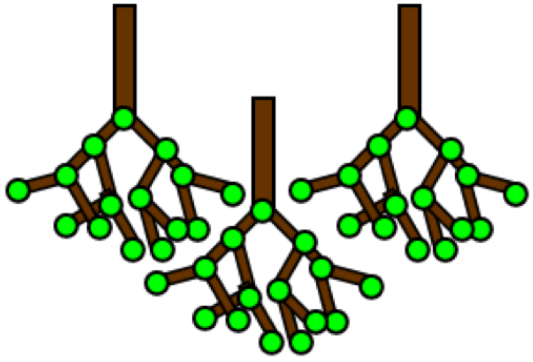
Distribution of $Y \mid X$ depends on X through which variables?

Variable is a discovery if

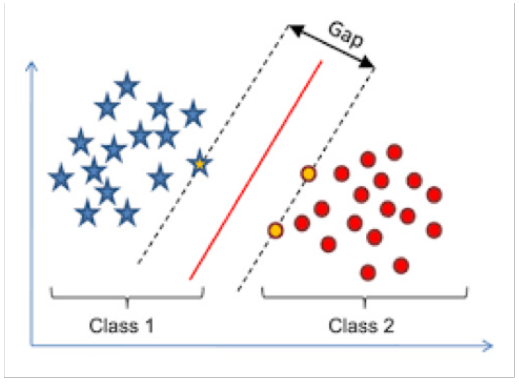
$$p(\text{response} \mid \text{variable}, \text{others}) \neq p(\text{response} \mid \text{others})$$

Selection in the computer age

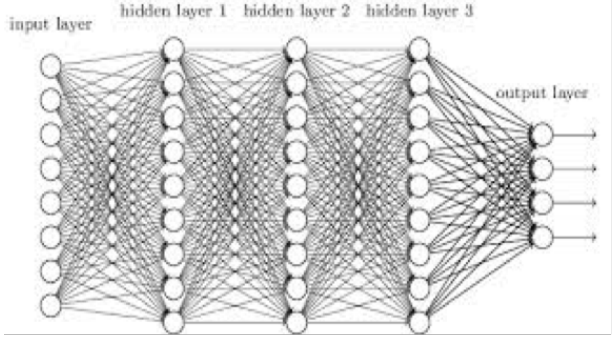
Many sophisticated tools to measure strength of dependence



Random forests

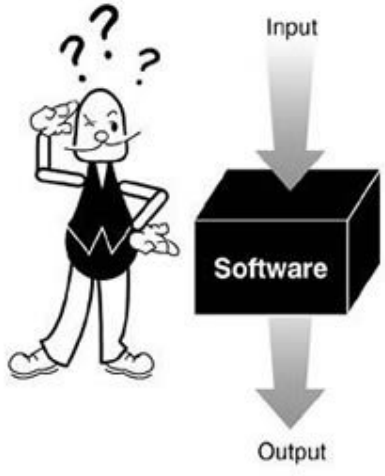


SVM



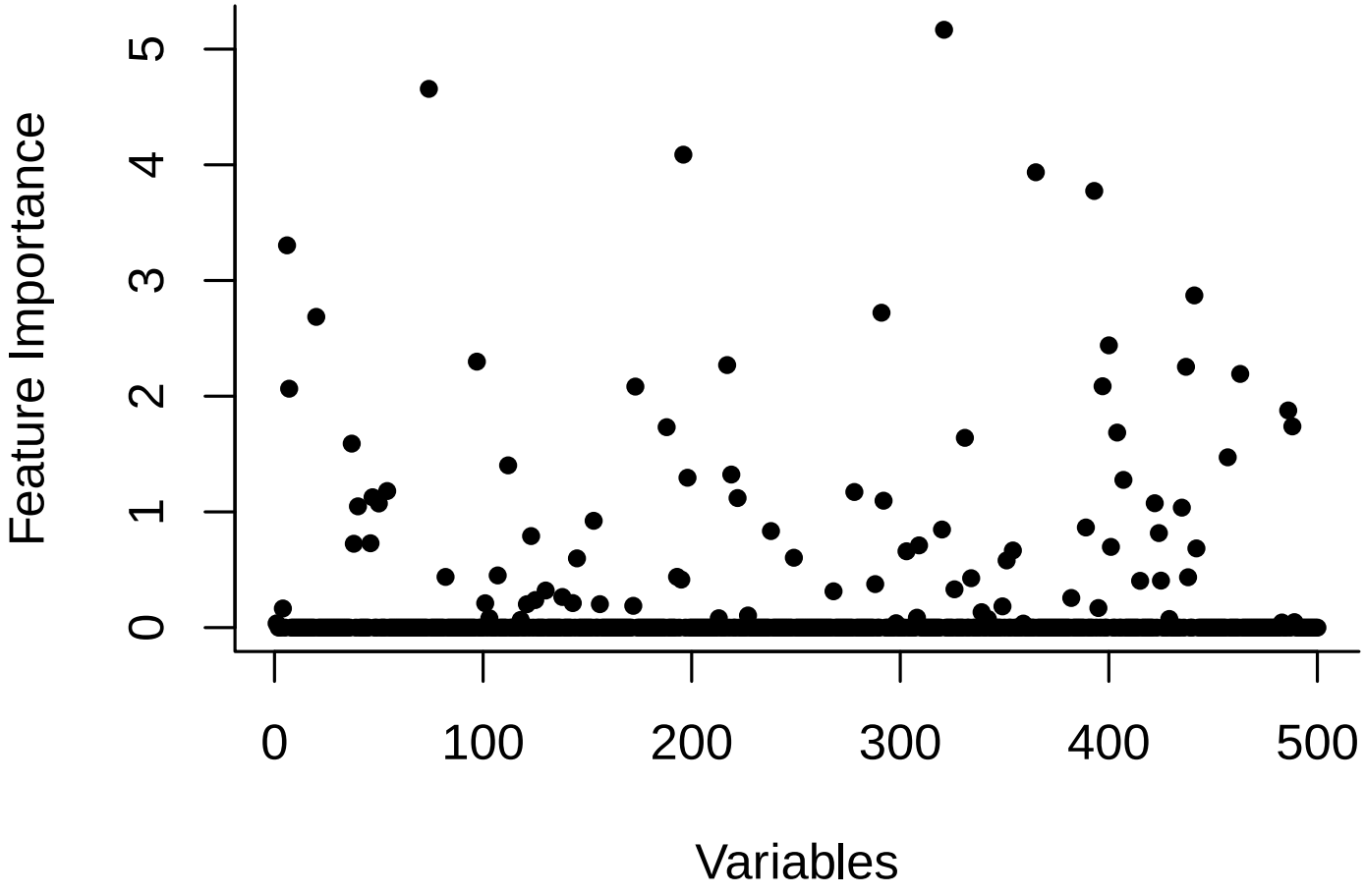
Deep nets

Bayes posteriors (MCMC)



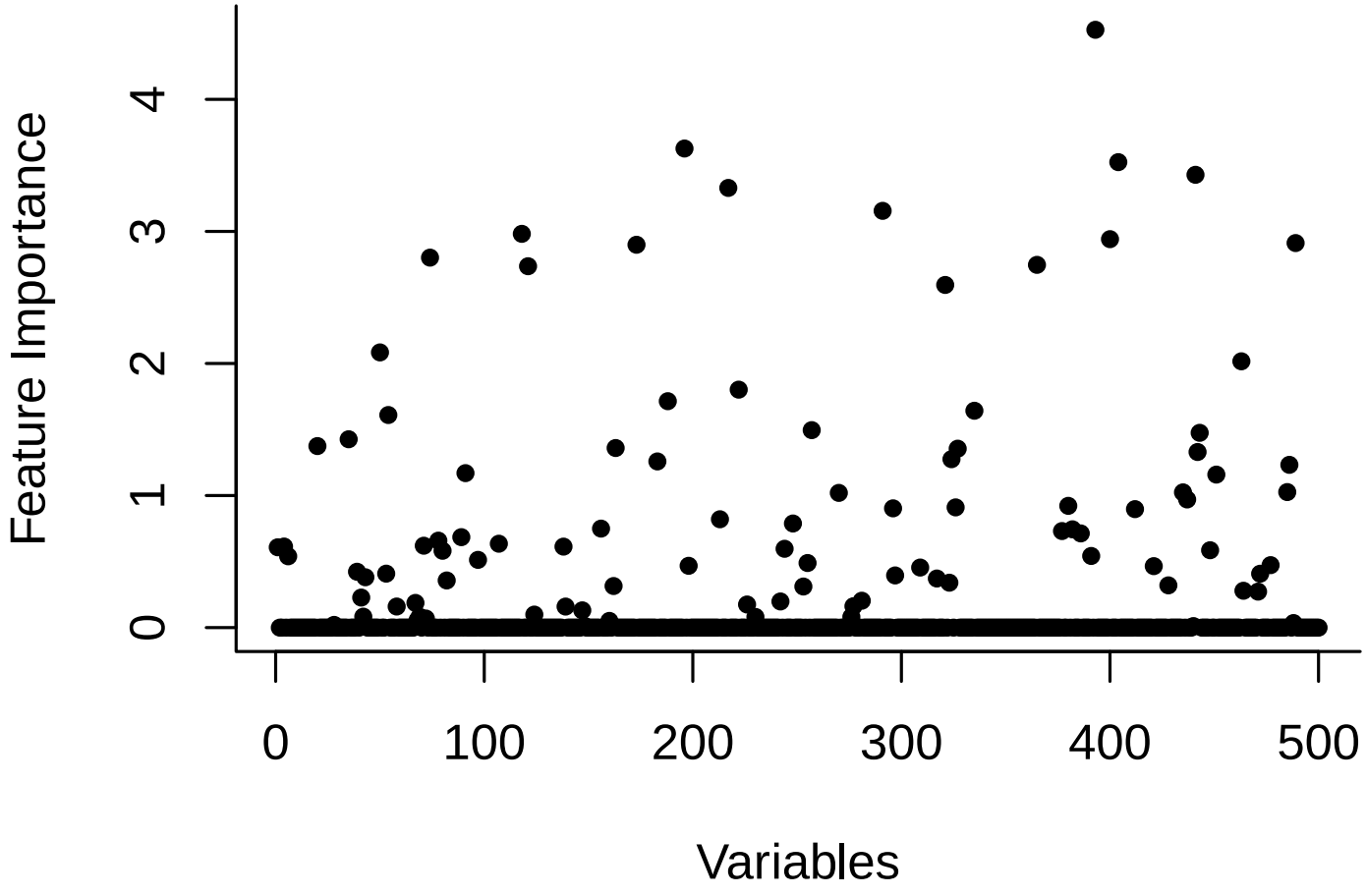
Black-box algorithm

‘Black box’ produces measures of strength of dependence



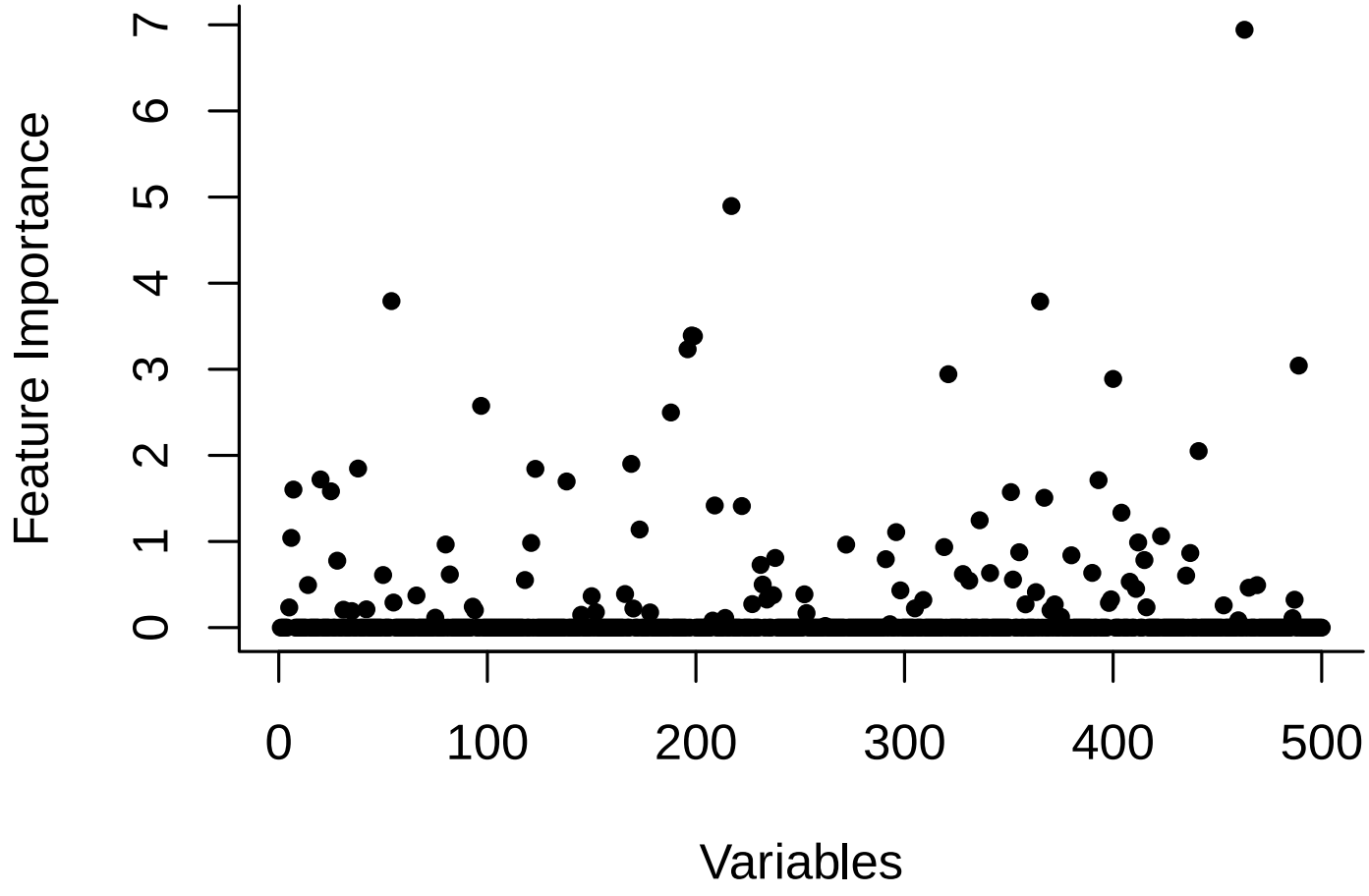
How do we make reliable decisions in the face of unknown statistical variability?

‘Black box’ produces measures of strength of dependence



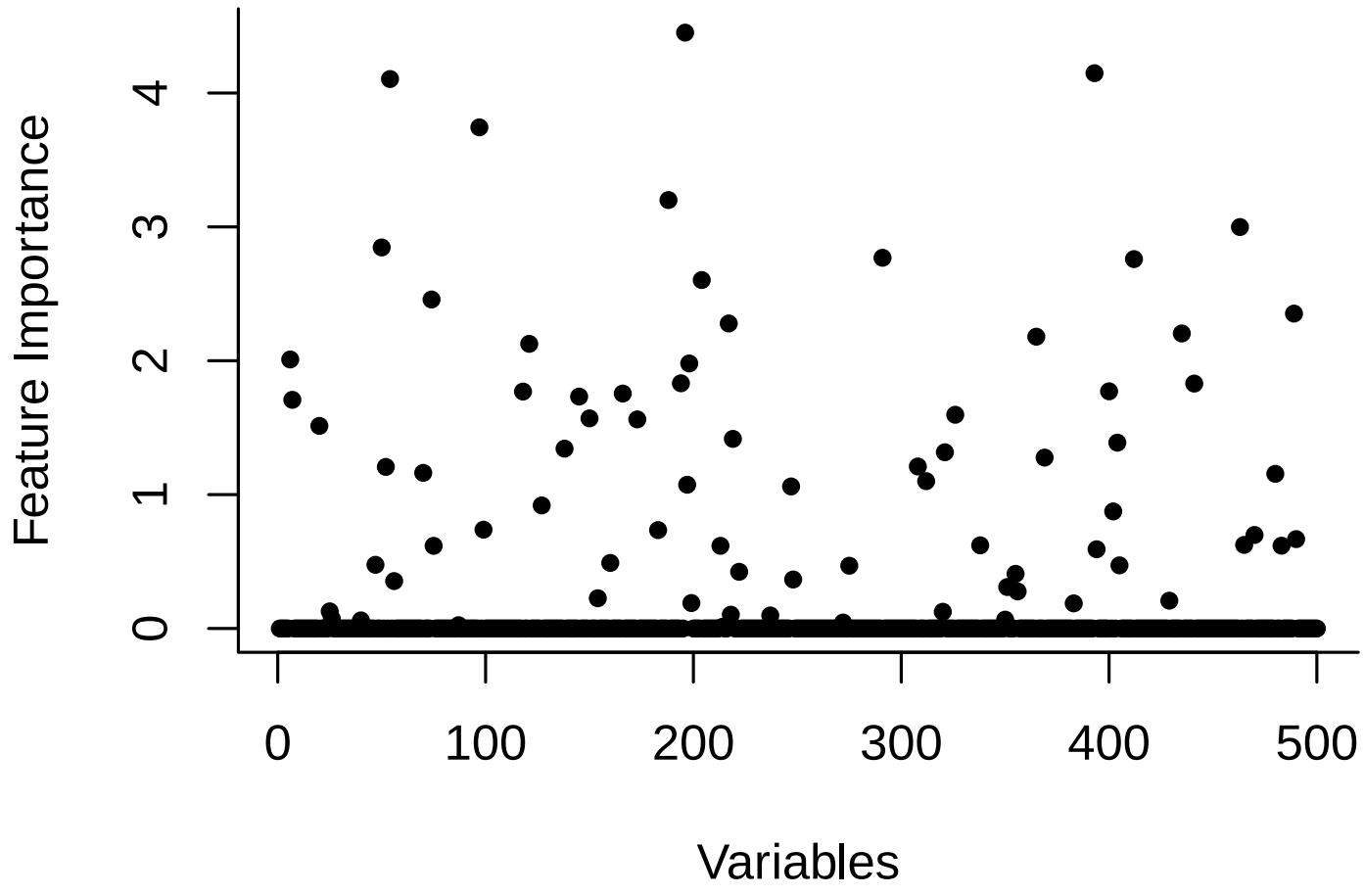
How do we make reliable decisions in the face of unknown statistical variability?

‘Black box’ produces measures of strength of dependence



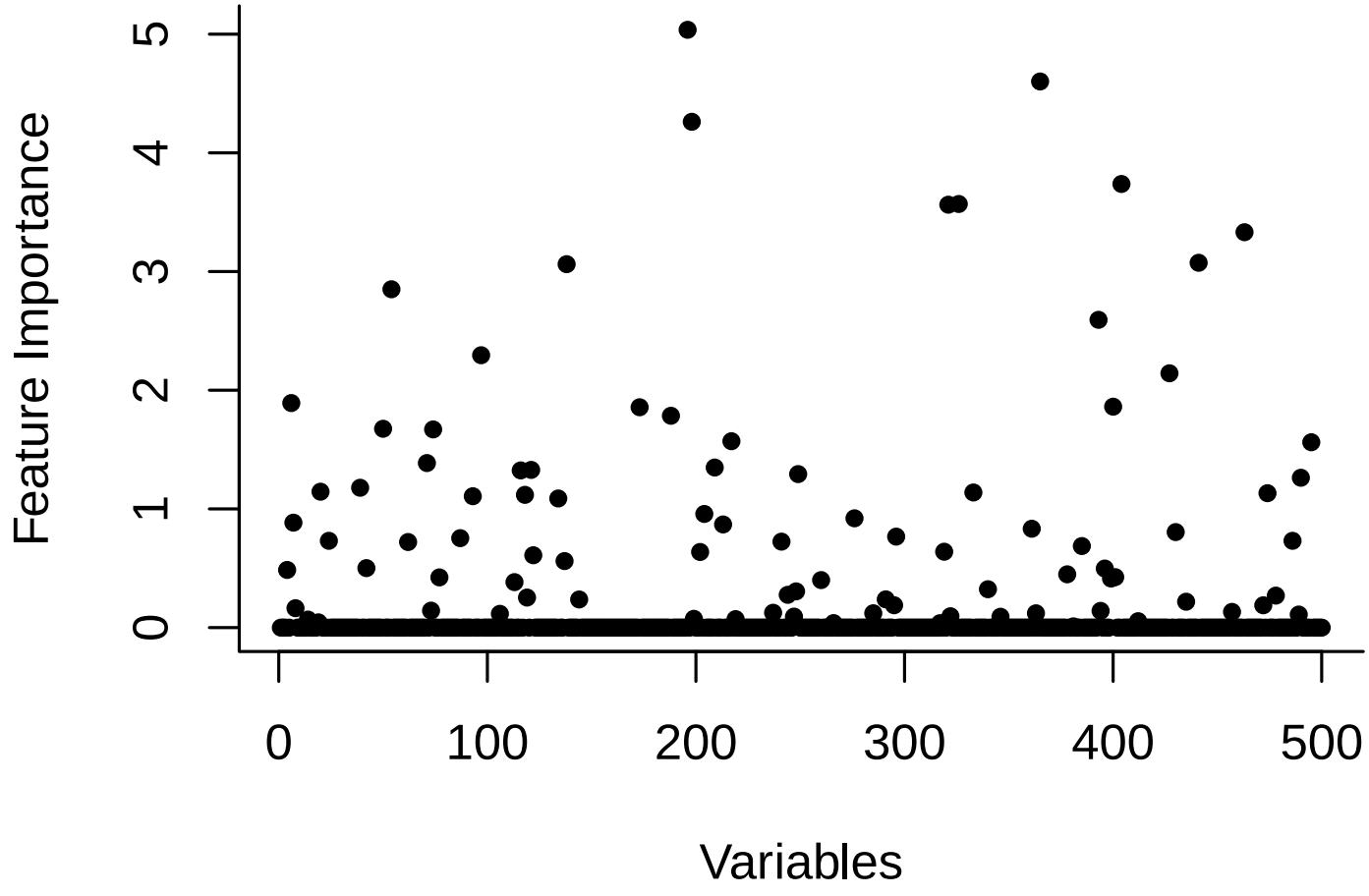
How do we make reliable decisions in the face of unknown statistical variability?

‘Black box’ produces measures of strength of dependence



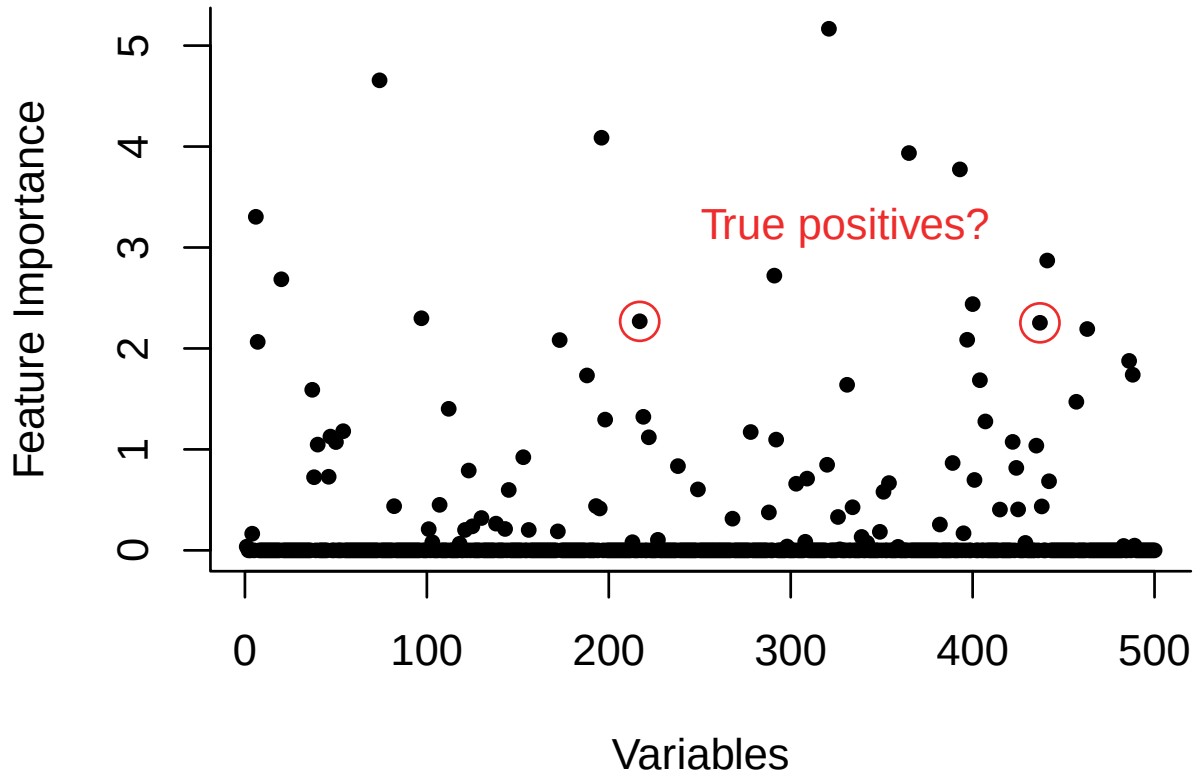
How do we make reliable decisions in the face of unknown statistical variability?

‘Black box’ produces measures of strength of dependence



How do we make reliable decisions in the face of unknown statistical variability?

Only one dataset: what should we report?

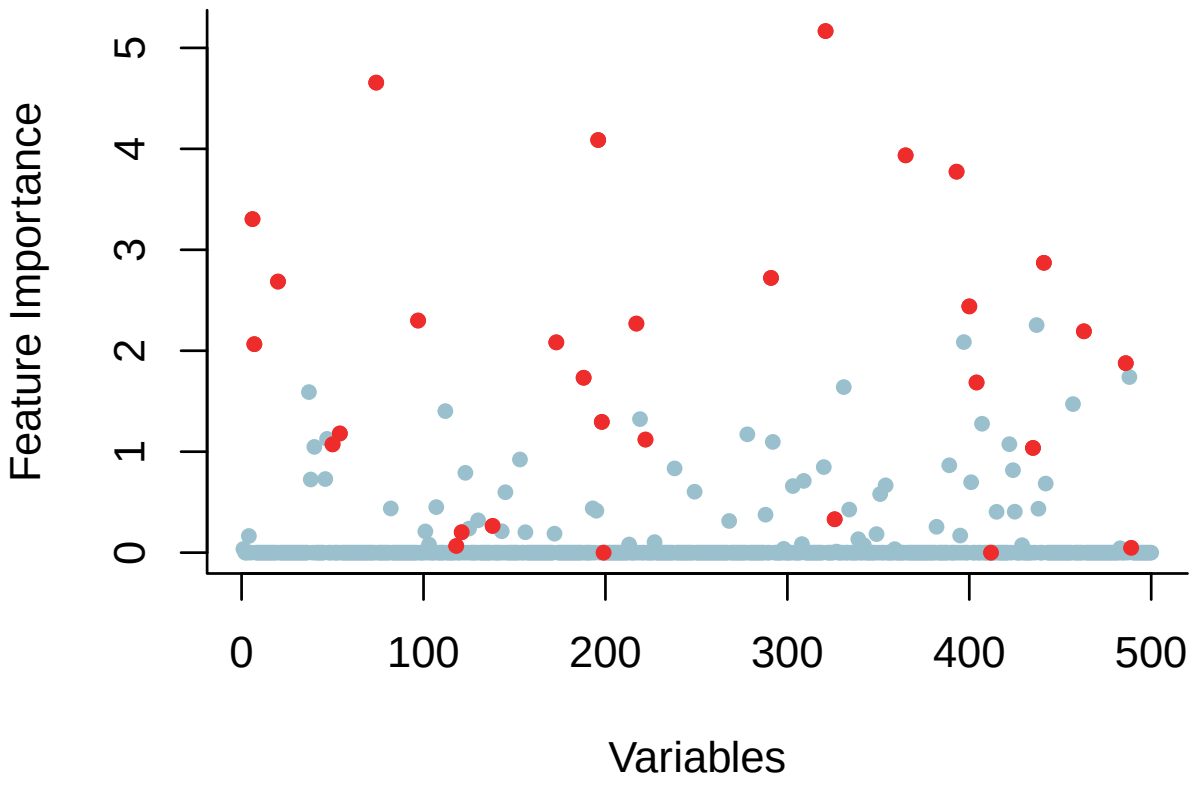


Modern science faces the problem of selection of promising findings from the noisy estimates of many.

Y. Benjamini and Y. Hechtlinger

Selecting promising findings from noisy estimates of many?

Feature importance statistic $Z_j = |\hat{\beta}_j(\lambda = 3)|$

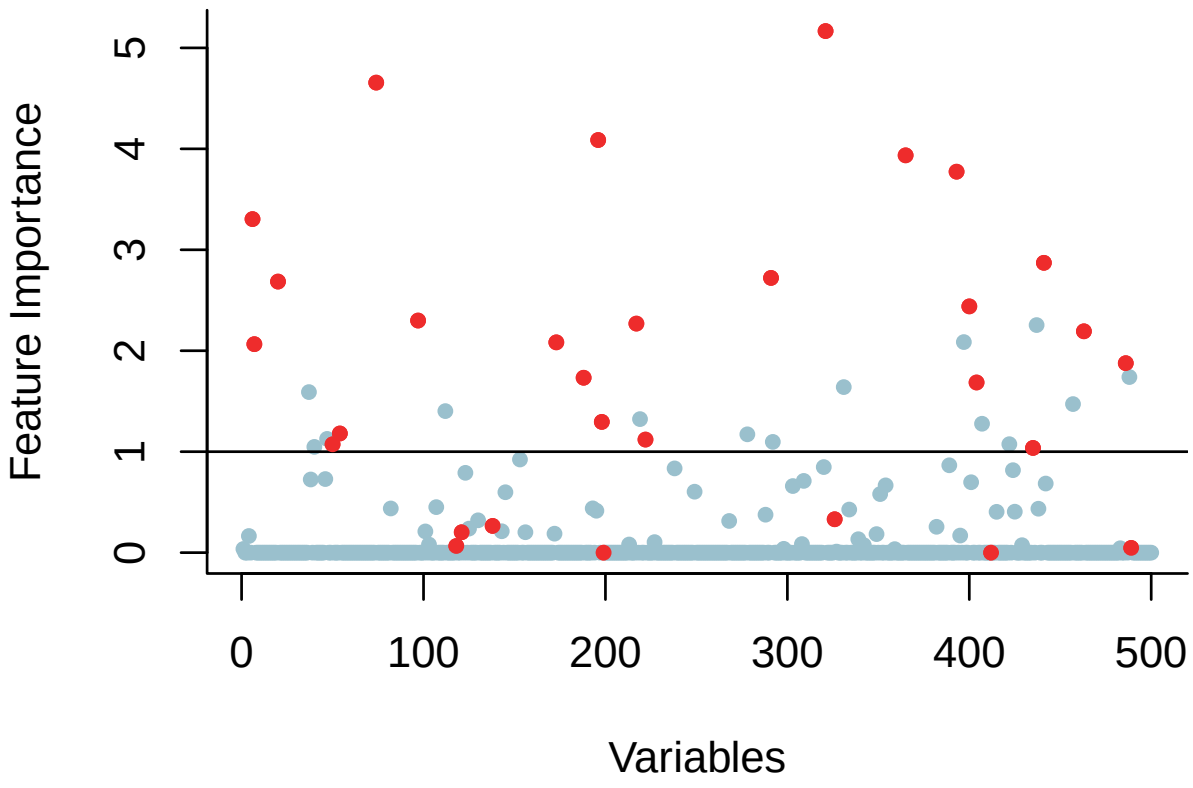


Select $Z_j > 0$

$$\text{FDP} = \frac{62}{90} \approx 69\%$$

Selecting promising findings from noisy estimates of many?

Feature importance statistic $Z_j = |\hat{\beta}_j(\lambda = 3)|$



Select $Z_j \geq 1$

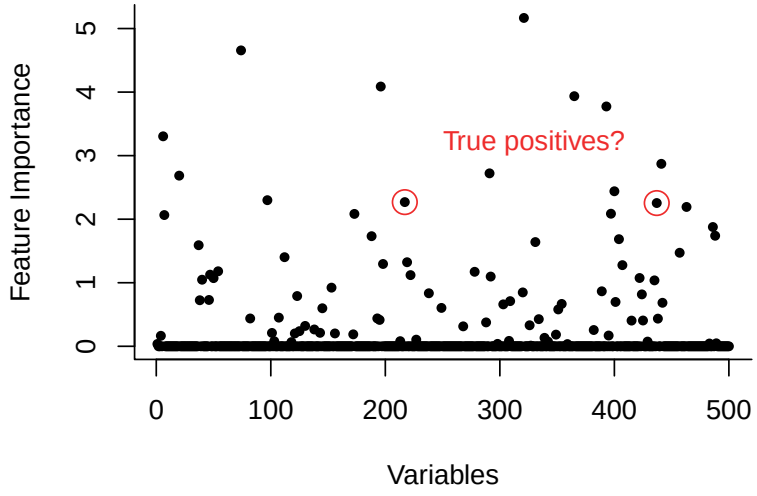
$$\text{FDP} = \frac{14}{37} \approx 38\%$$

Myriad of delicate inference problems

Does dist. of $Y \,|\, X$ depend on X_j?

$$H_j : Y \perp\!\!\!\perp X_j \,|\, X_{-j}$$

Not interested in marginal testing (wrong question)

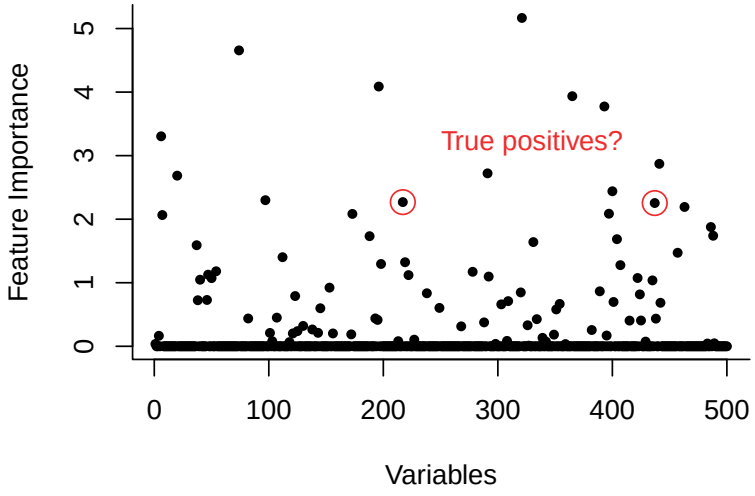


Myriad of delicate inference problems

Does dist. of $Y \,|\, X$ depend on X_j?

$$H_j : Y \perp\!\!\!\perp X_j \,|\, X_{-j}$$

Not interested in marginal testing (wrong question)



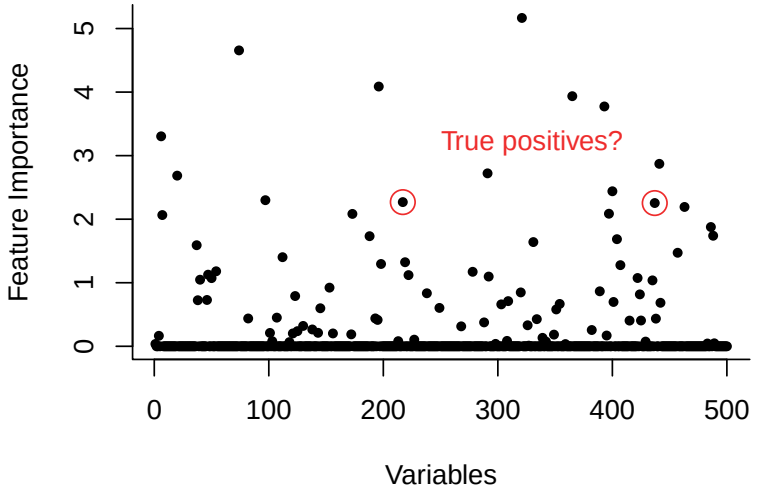
- High dimensionality $p \approx n$ or $p \gg n$
- No idea how to compute sampling distribution of statistical estimates
- No idea how to compute p-values in general
- Test statistics are not independent
- ...

Myriad of delicate inference problems

Does dist. of $Y \,|\, X$ depend on X_j?

$$H_j : Y \perp\!\!\!\perp X_j \,|\, X_{-j}$$

Not interested in marginal testing (wrong question)

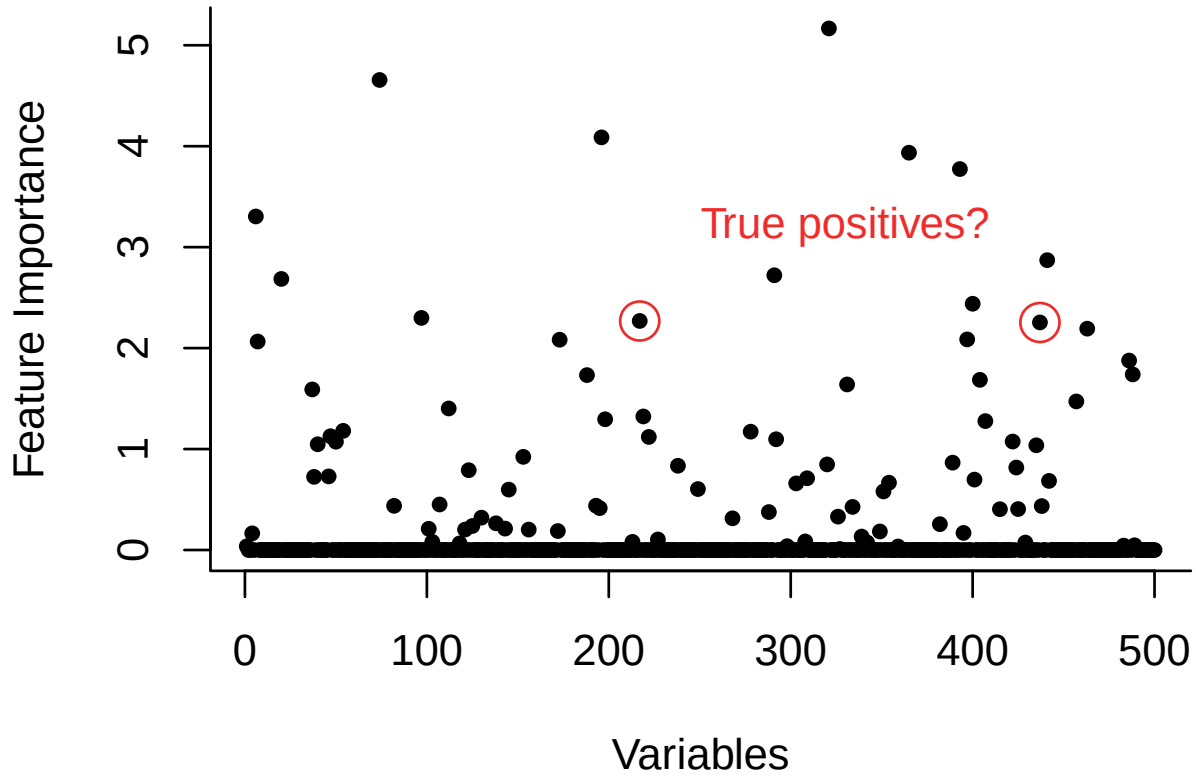


- High dimensionality $p \approx n$ or $p \gg n$
- No idea how to compute sampling distribution of statistical estimates
- No idea how to compute p-values in general
- Test statistics are not independent
- ...

Our goal

Provide some solutions to these important problems

How do the nulls look like?

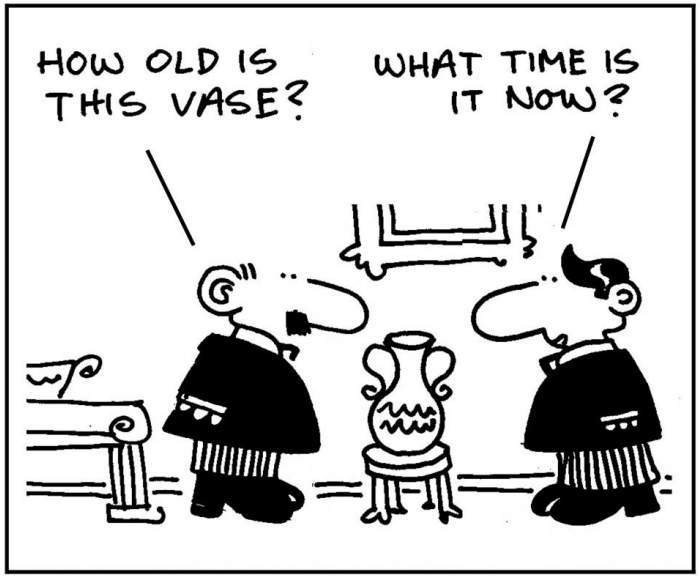


Would like comparison points (controls) for the nulls. How?

Knockoffs

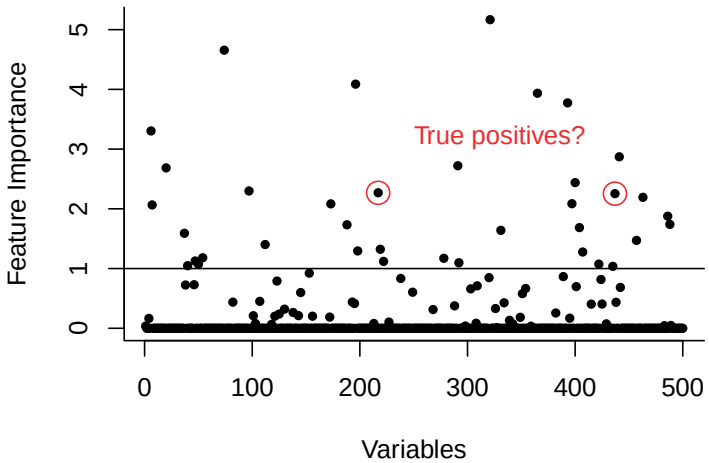
It's not a name brand bag
just a cheap knockoff

Thesaurize.com

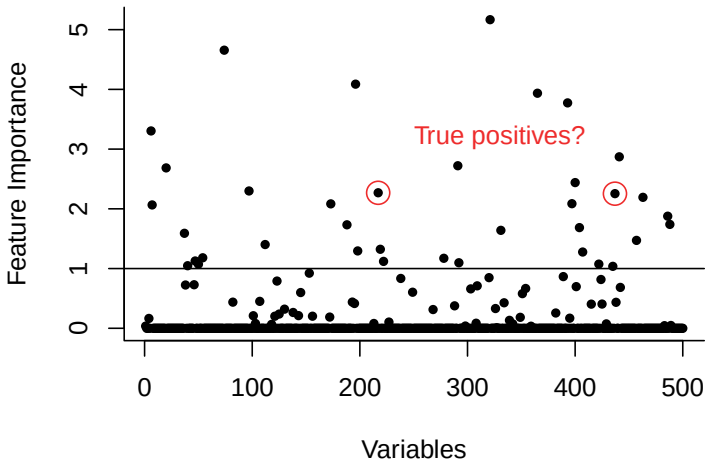


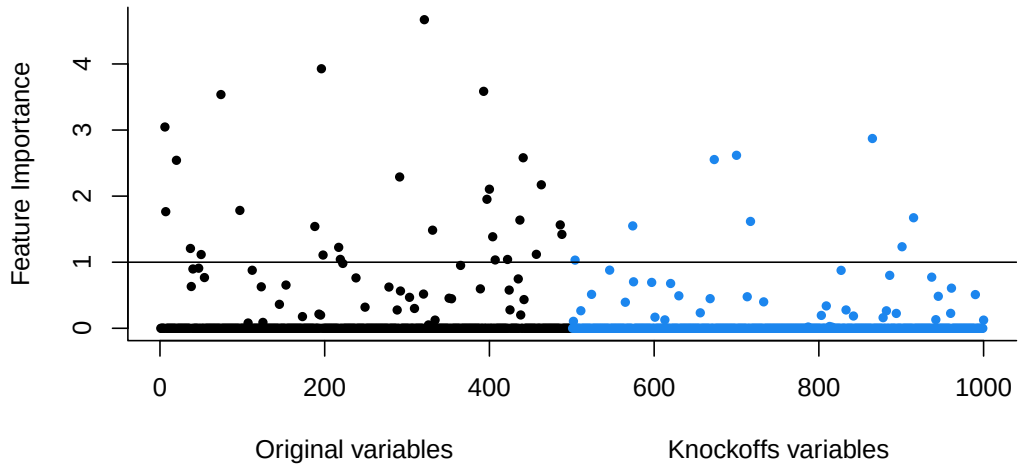
For each feature X_j, construct a knockoff version $\tilde{X}_j$
Knockoffs serve as "control group" $\Longrightarrow$ can estimate FDP & control FDR

Knockoffs framework (Barber and Candès, 2015)



Knockoffs framework (Barber and Candès, 2015)





Run same procedure on original and knockoff features 'serving as controls'

Not just dummy variables...

For linear models, Miller ('84, '02) creates 'dummy' variables with entries drawn i.i.d. at random

- Forward selection procedure is applied to augmented list of variables
- Stop when selects a dummy variable for the first time

Pseudovariables (permuted rows and variants): Wu, Boos and Stefanski ('07, '09)

Not just dummy variables...

For linear models, Miller ('84, '02) creates 'dummy' variables with entries drawn i.i.d. at random

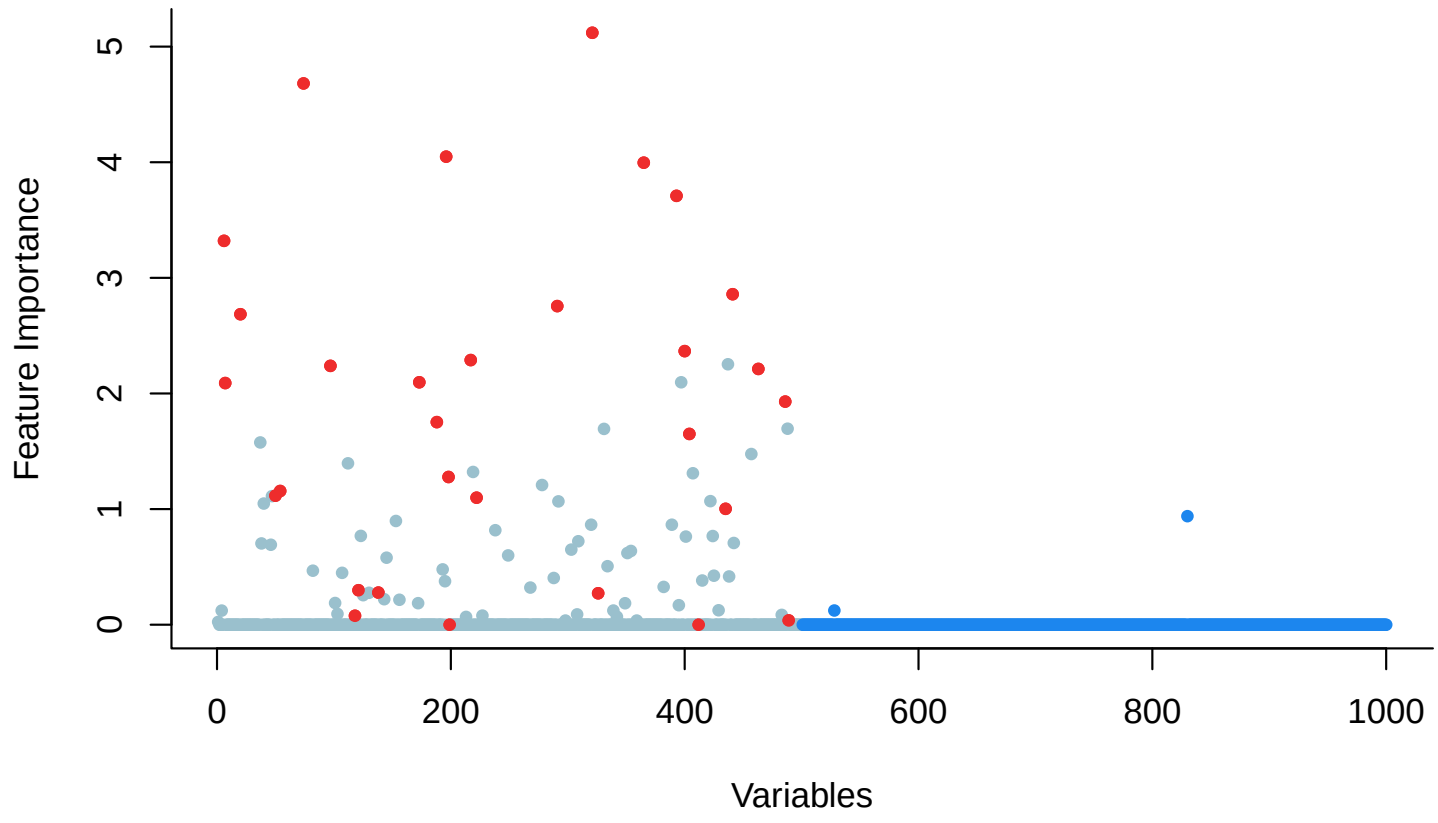
- Forward selection procedure is applied to augmented list of variables
- Stop when selects a dummy variable for the first time

Pseudovariables (permuted rows and variants): Wu, Boos and Stefanski ('07, '09)

Dummies	Structure preserved
Independent Gaussian variables	Mean and marginal variance
Permuted rows `X[sample(n),]`	(Joint) distribution
Knockoffs	More...

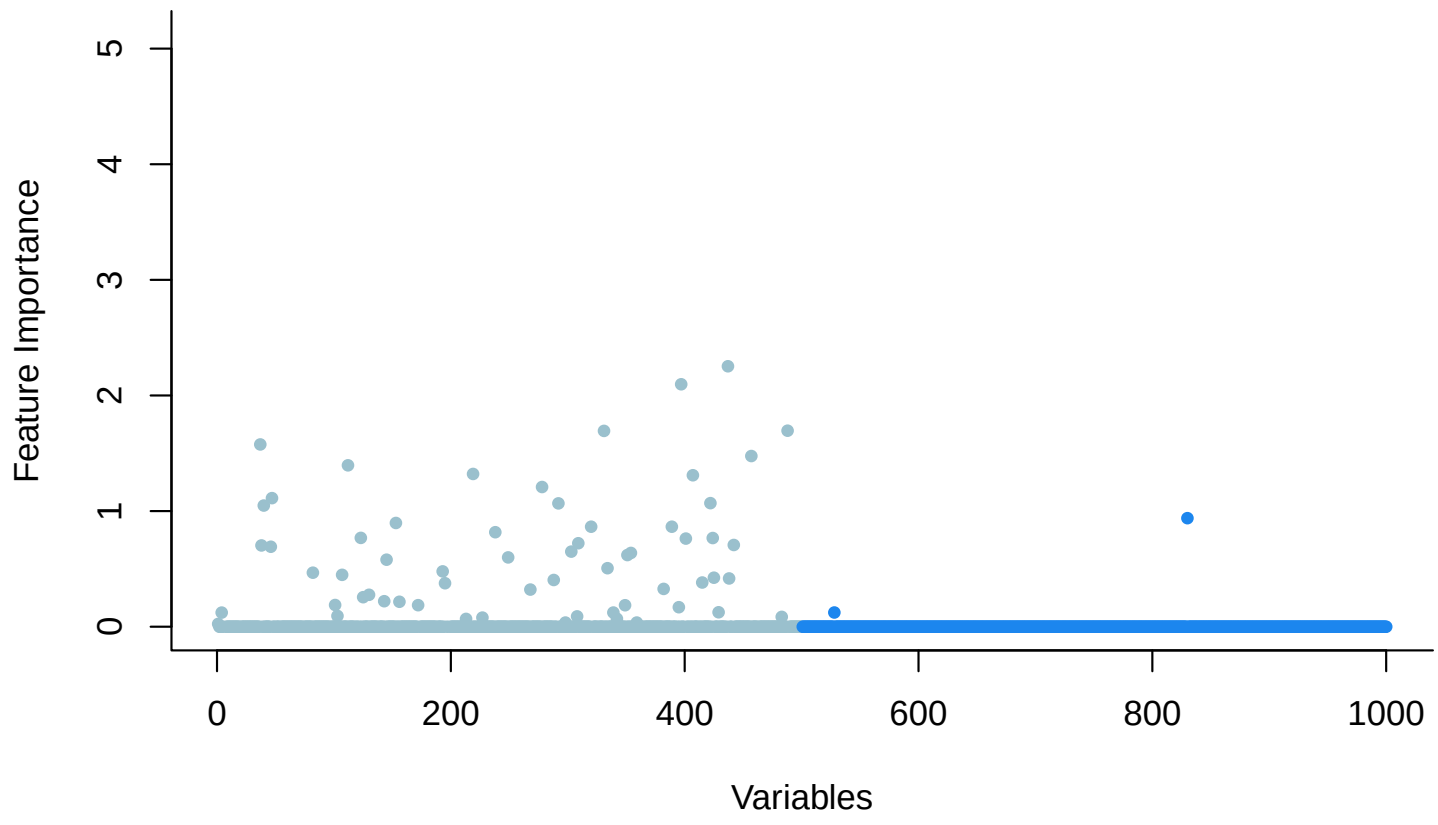
Gaussian dummies

Feature importance statistic $Z_j = |\hat{\beta}_j(\lambda = 3)|$

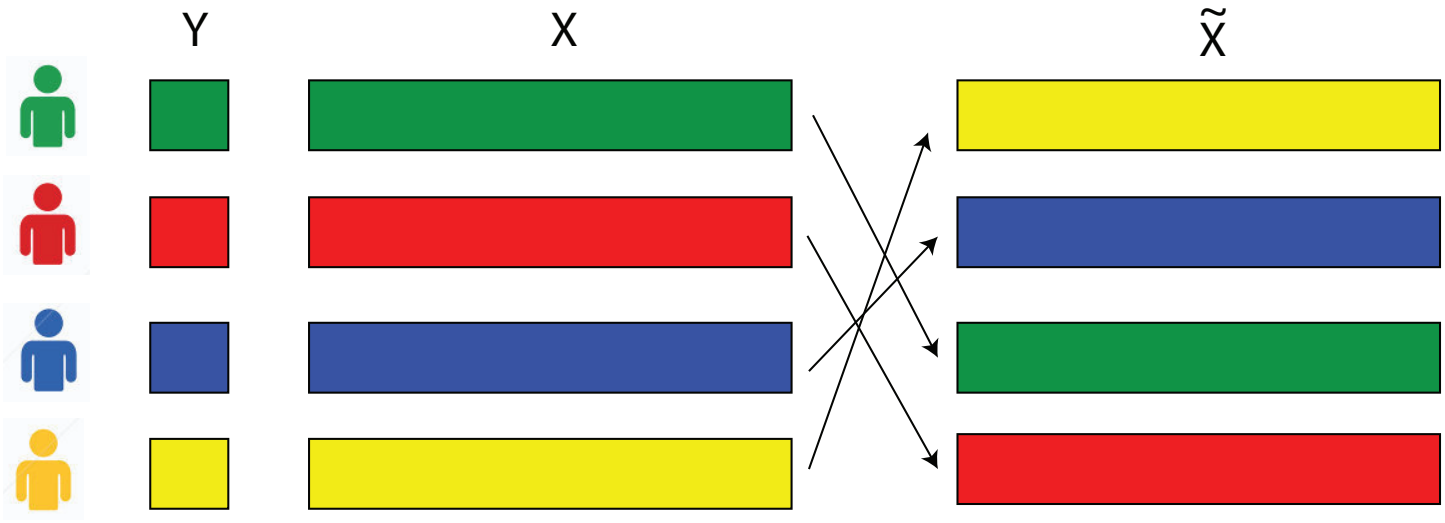


Gaussian dummies

Feature importance statistic $Z_j = |\hat{\beta}_j(\lambda = 3)|$

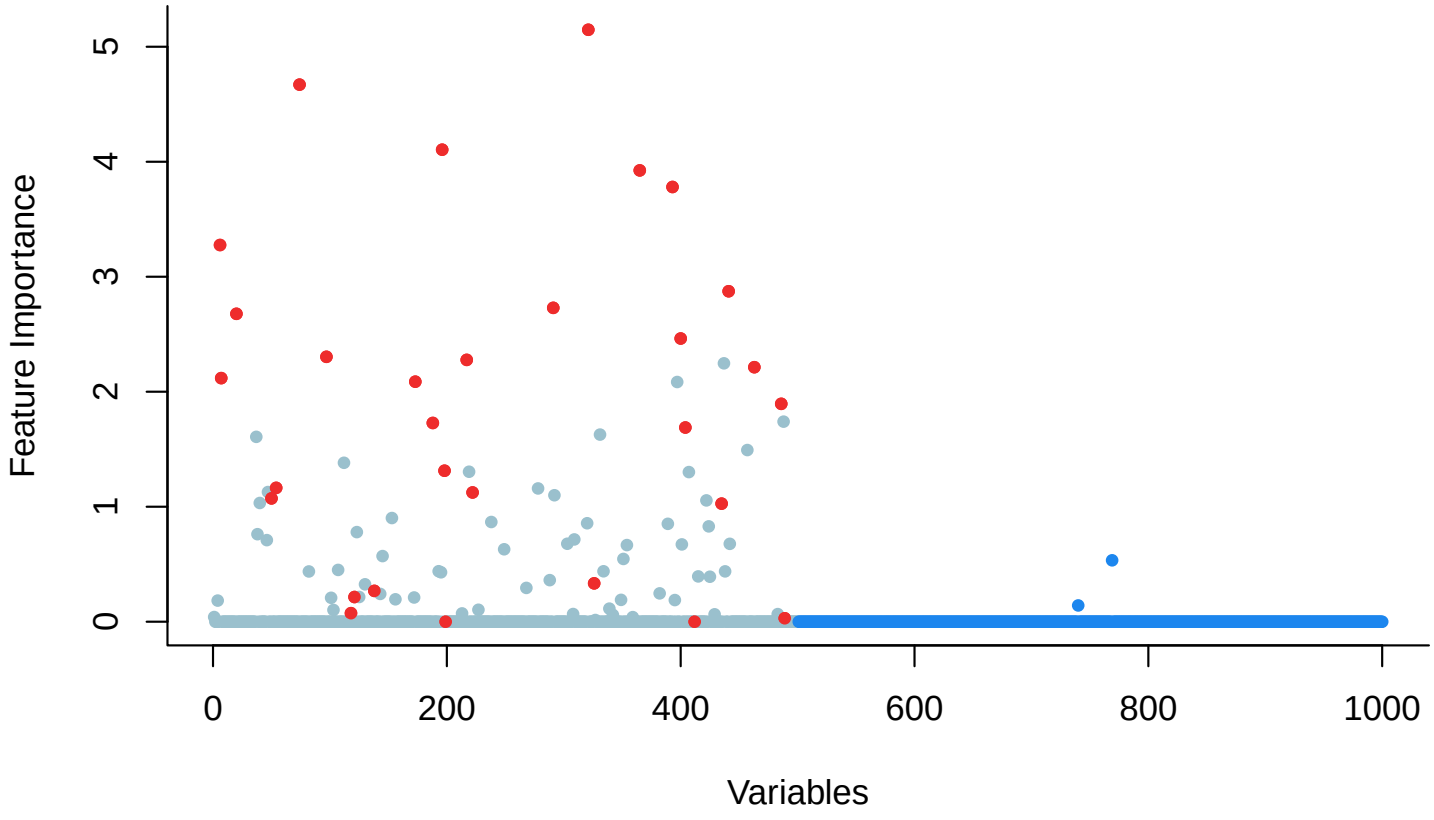


How? By permutation?



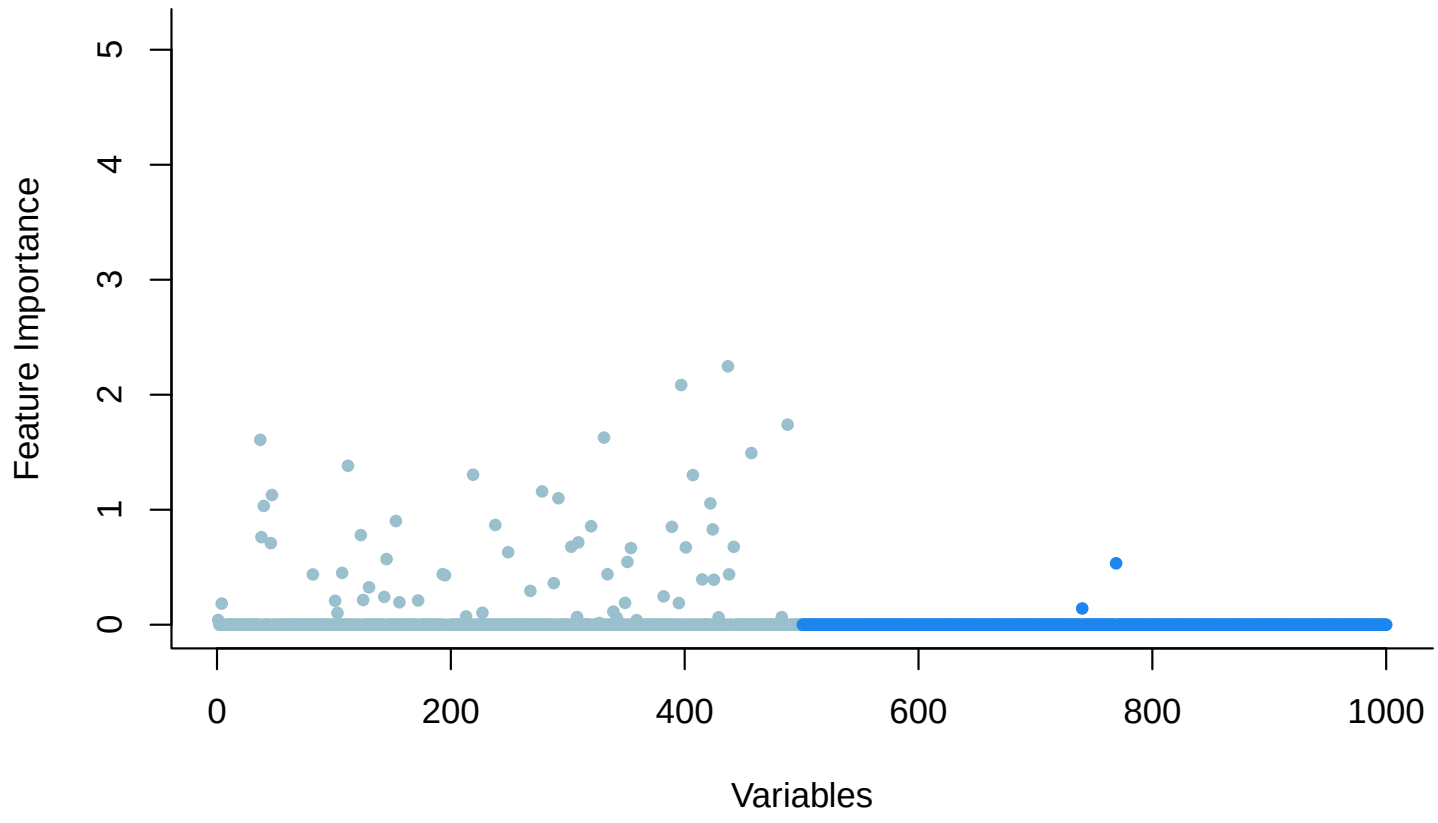
Permuted dummies

Feature importance $Z_j = |\hat{\beta}_j(\lambda = 3)|$



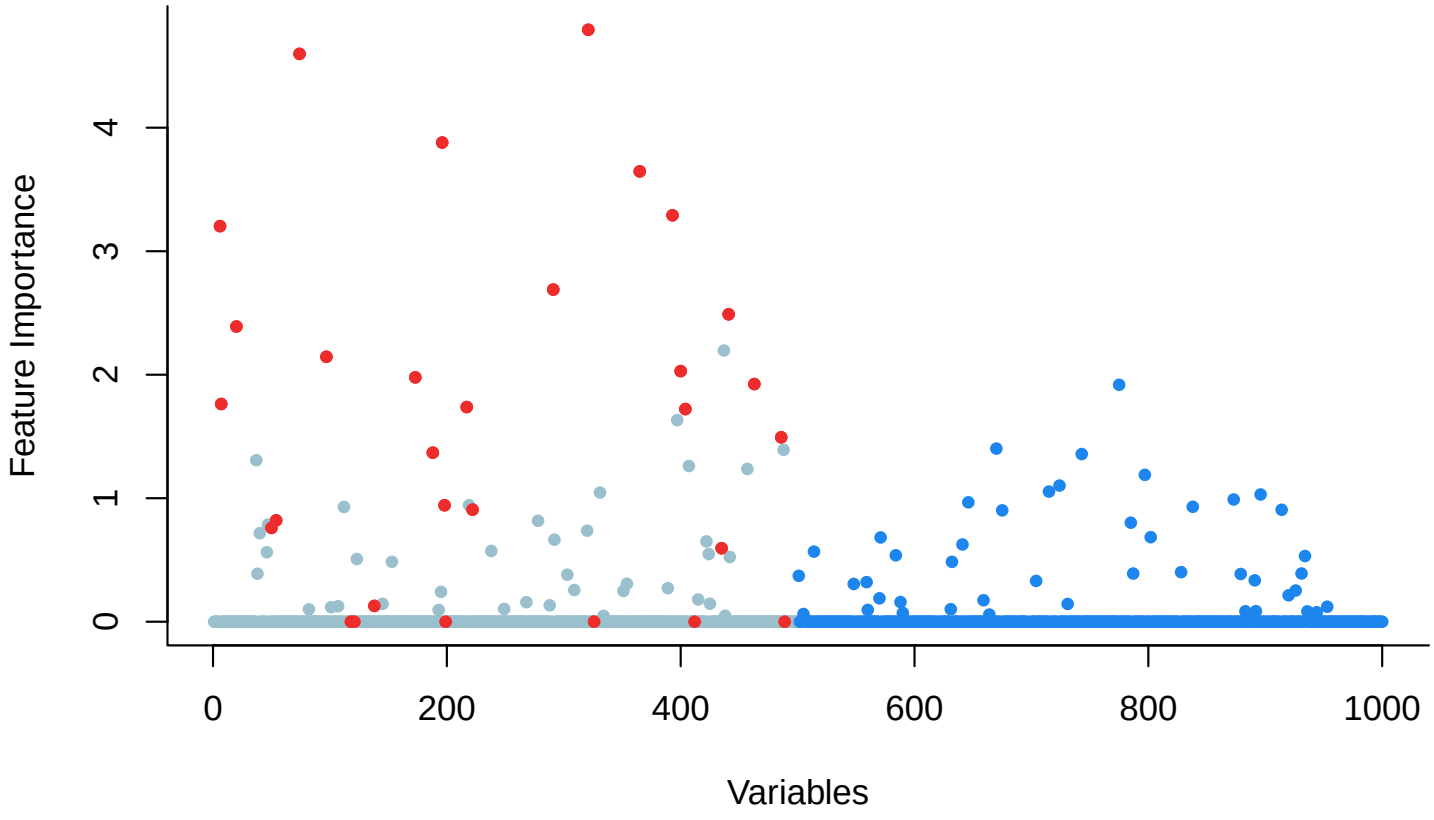
Permuted dummies

Feature importance $Z_j = |\hat{\beta}_j(\lambda = 3)|$



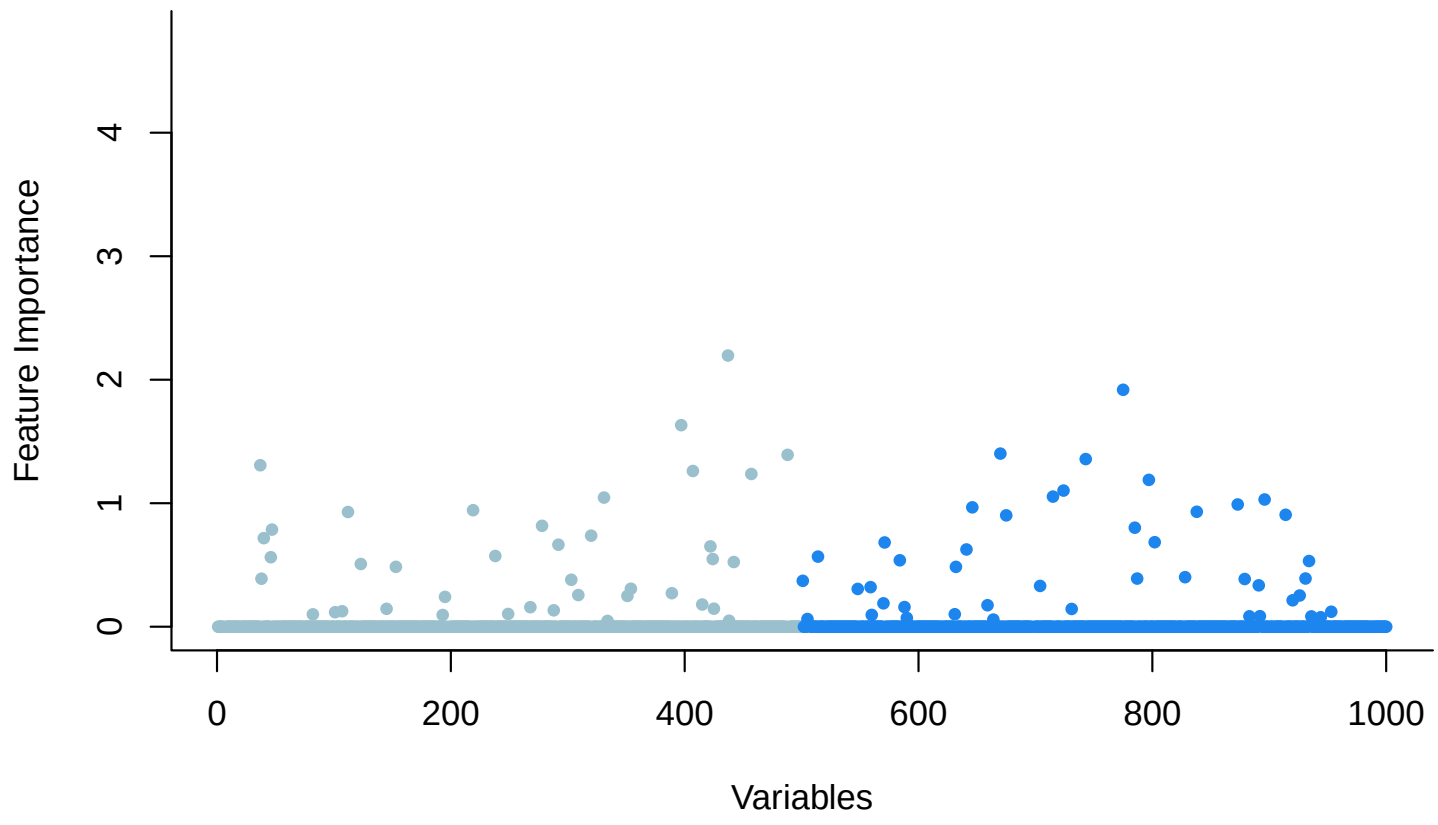
Knockoff dummies

Feature importance $Z_j = |\hat{\beta}_j(\lambda = 3)|$



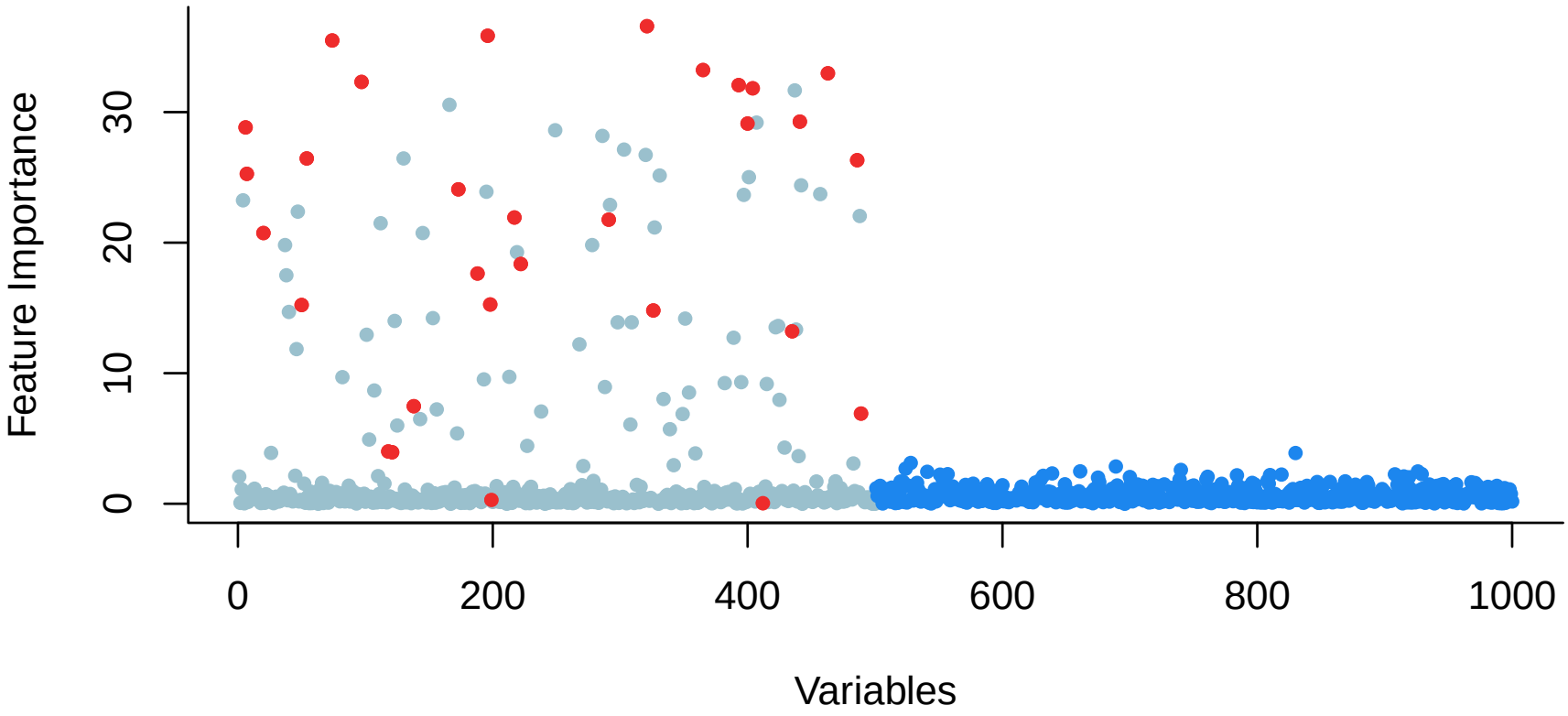
Knockoff dummies

Feature importance $Z_j = |\hat{\beta}_j(\lambda = 3)|$



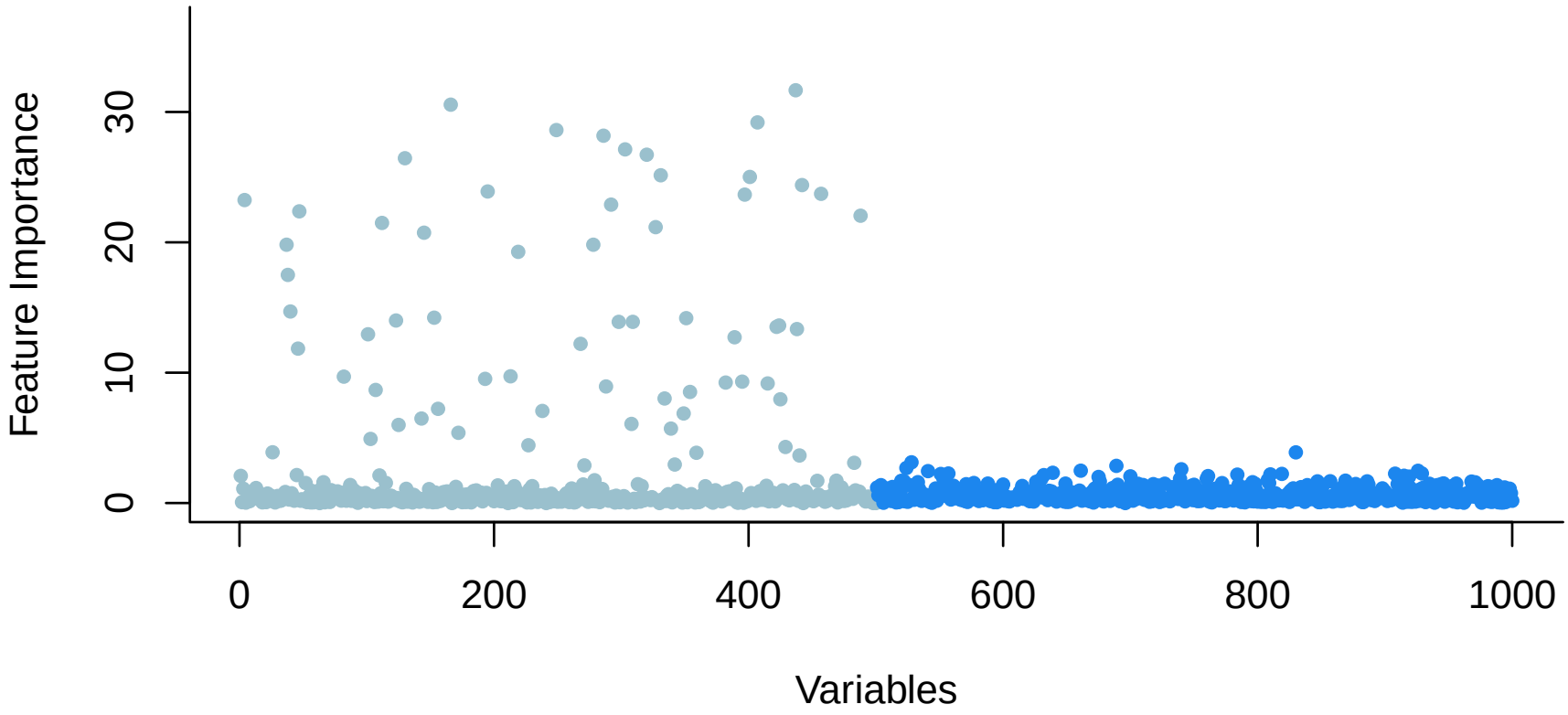
Gaussian dummies: different statistic Z_j

Feature importance $Z_j = \sup\{\lambda : |\hat{\beta}_j(\lambda)| \neq 0\}$



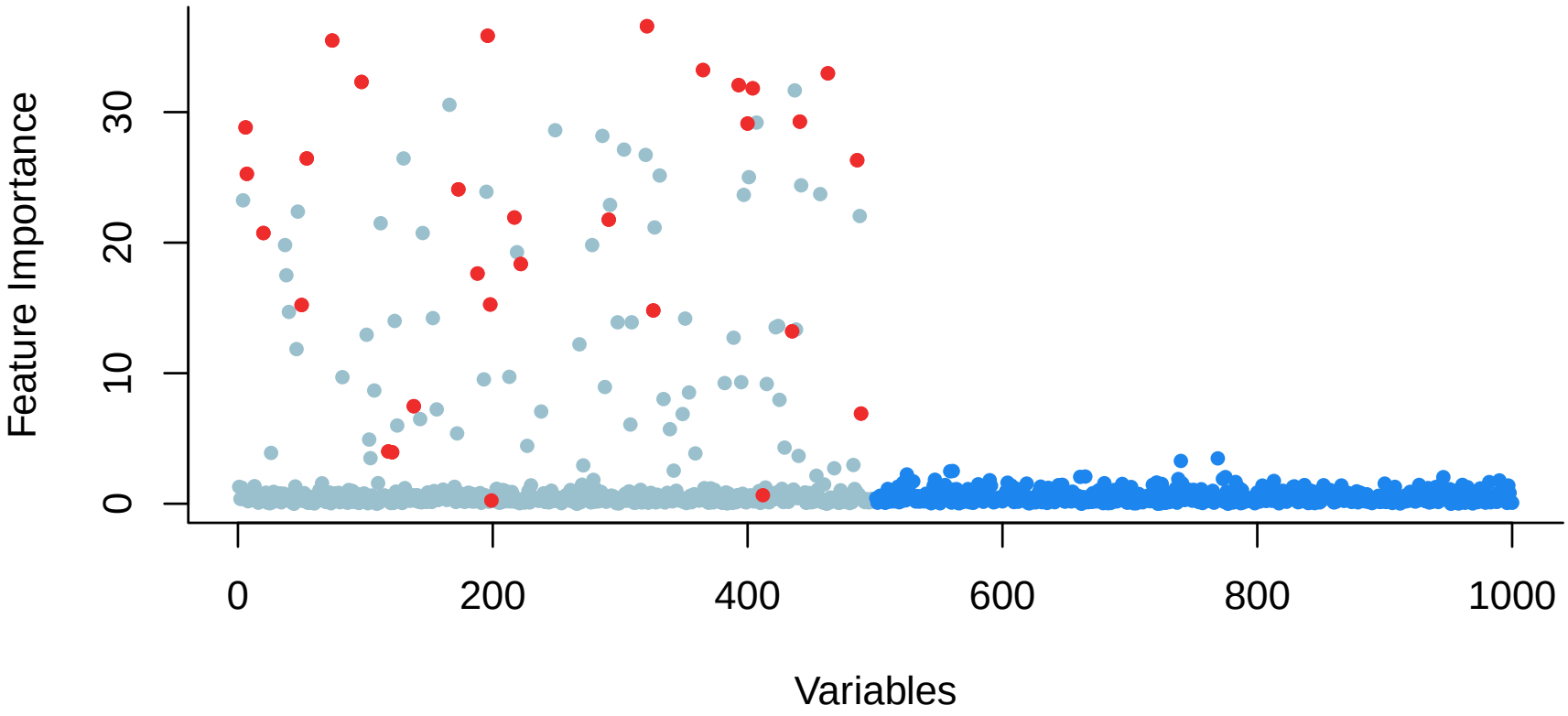
Gaussian dummies: different statistic Z_j

Feature importance $Z_j = \sup\{\lambda : |\hat{\beta}_j(\lambda)| \neq 0\}$



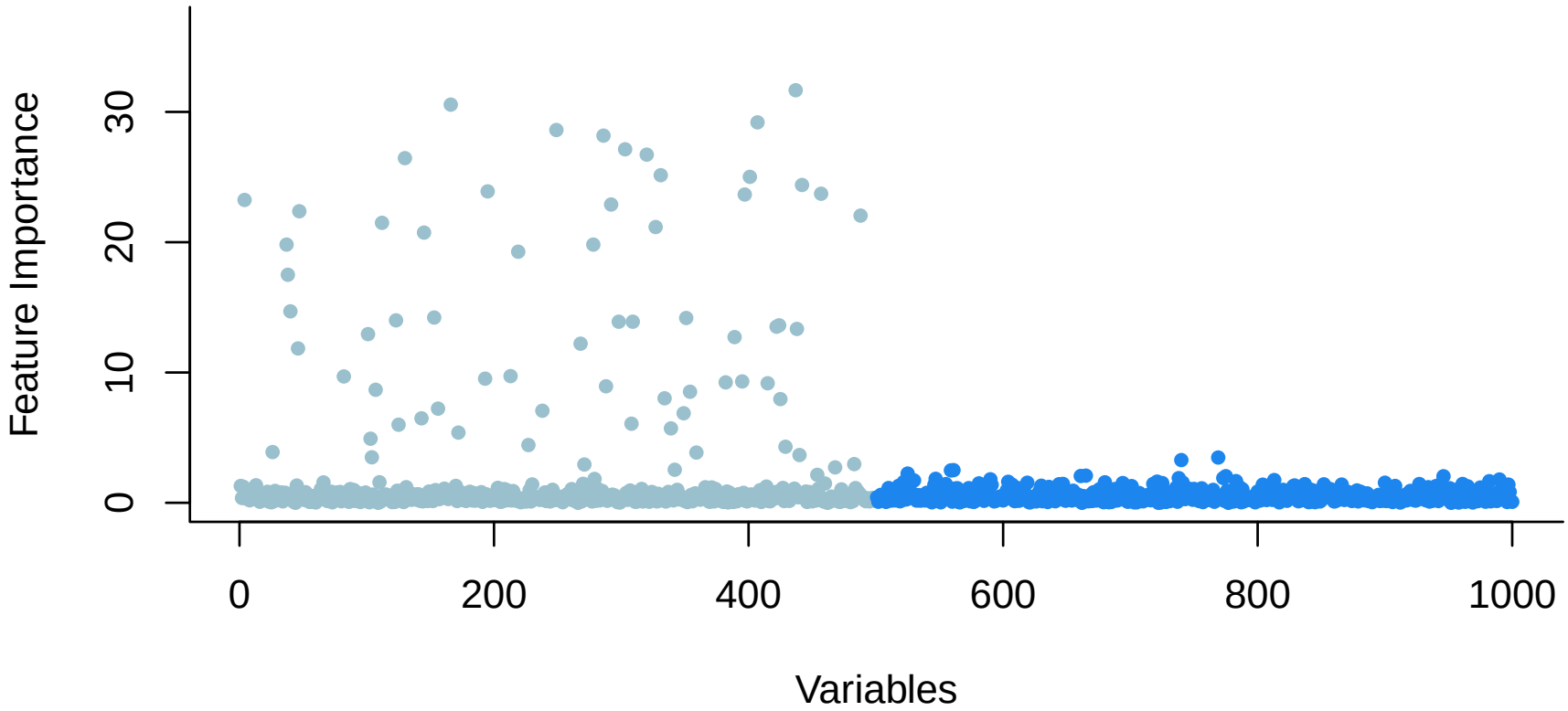
Permuted dummies: different statistic Z_j

Feature importance $Z_j = \sup\{\lambda : |\hat{\beta}_j(\lambda)| \neq 0\}$



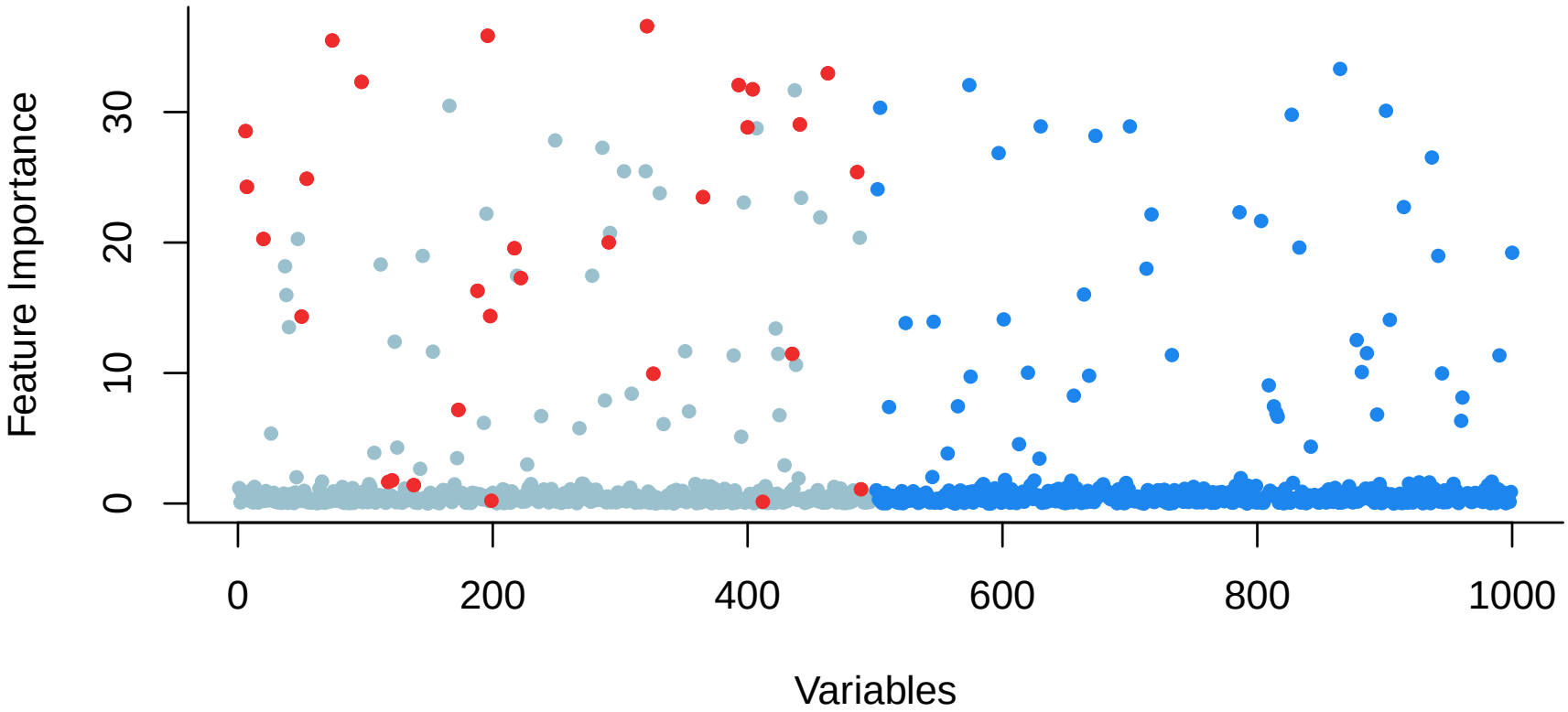
Permuted dummies: different statistic Z_j

Feature importance $Z_j = \sup\{\lambda : |\hat{\beta}_j(\lambda)| \neq 0\}$



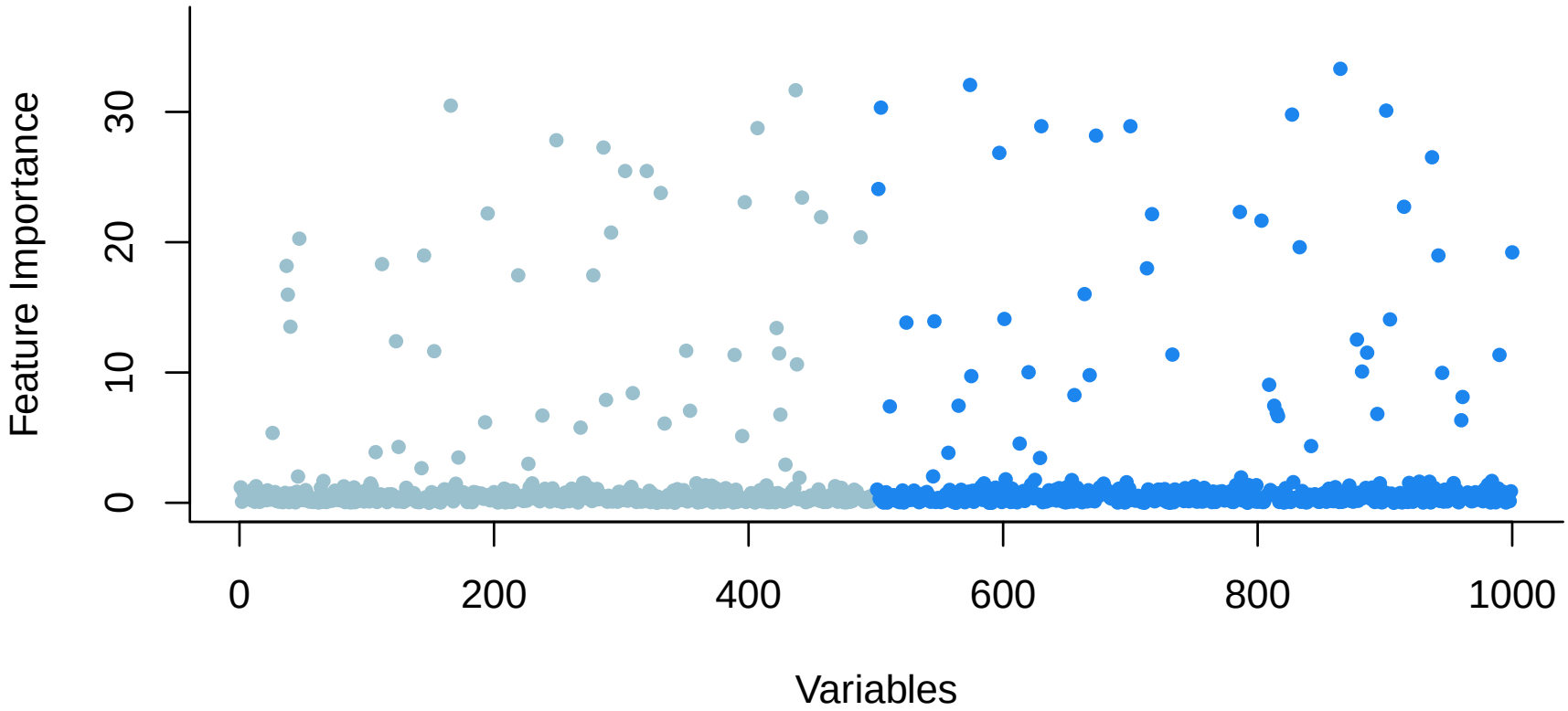
Knockoff dummies: different statistic Z_j

Feature importance $Z_j = \sup\{\lambda : |\hat{\beta}_j(\lambda)| \neq 0\}$

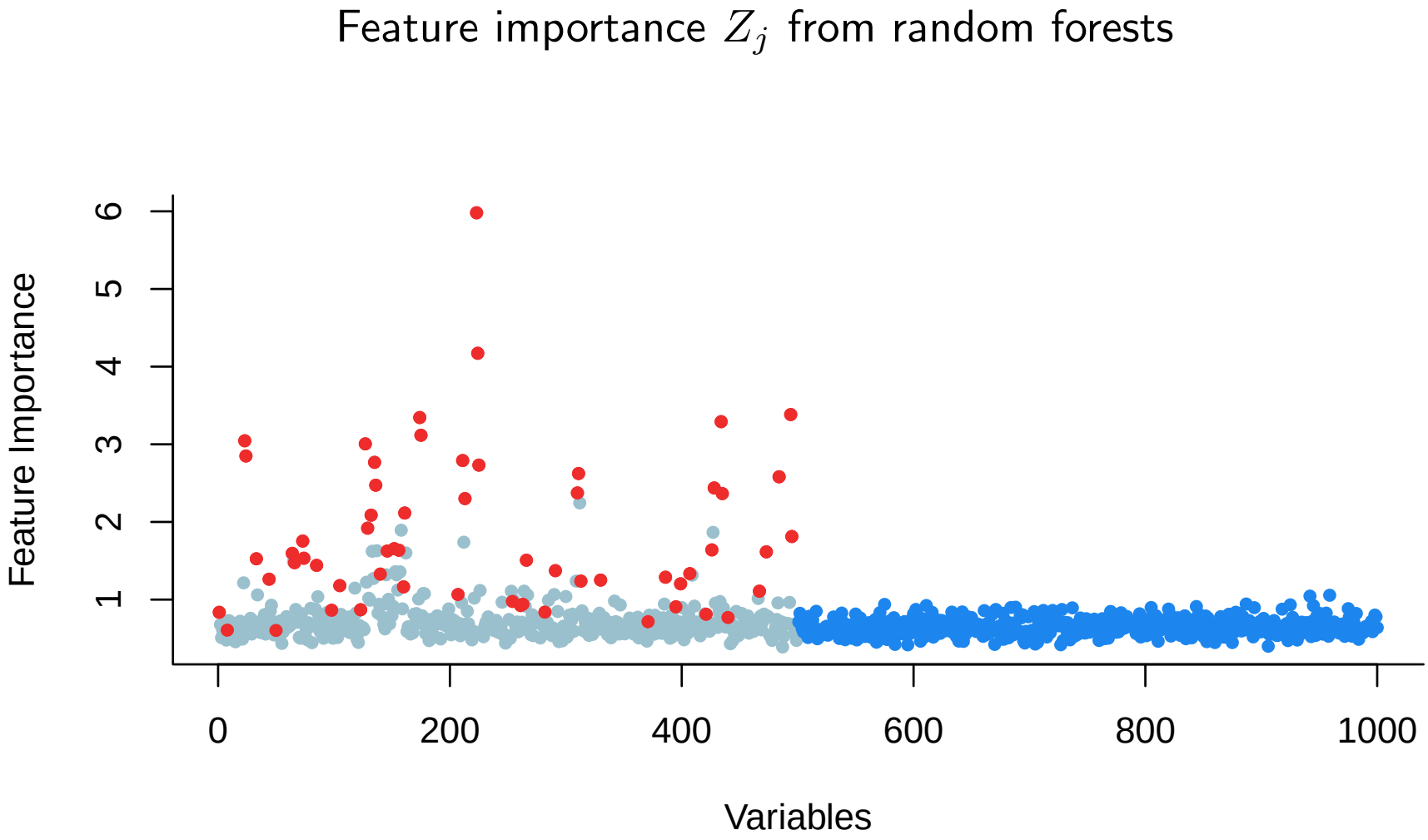


Knockoff dummies: different statistic Z_j

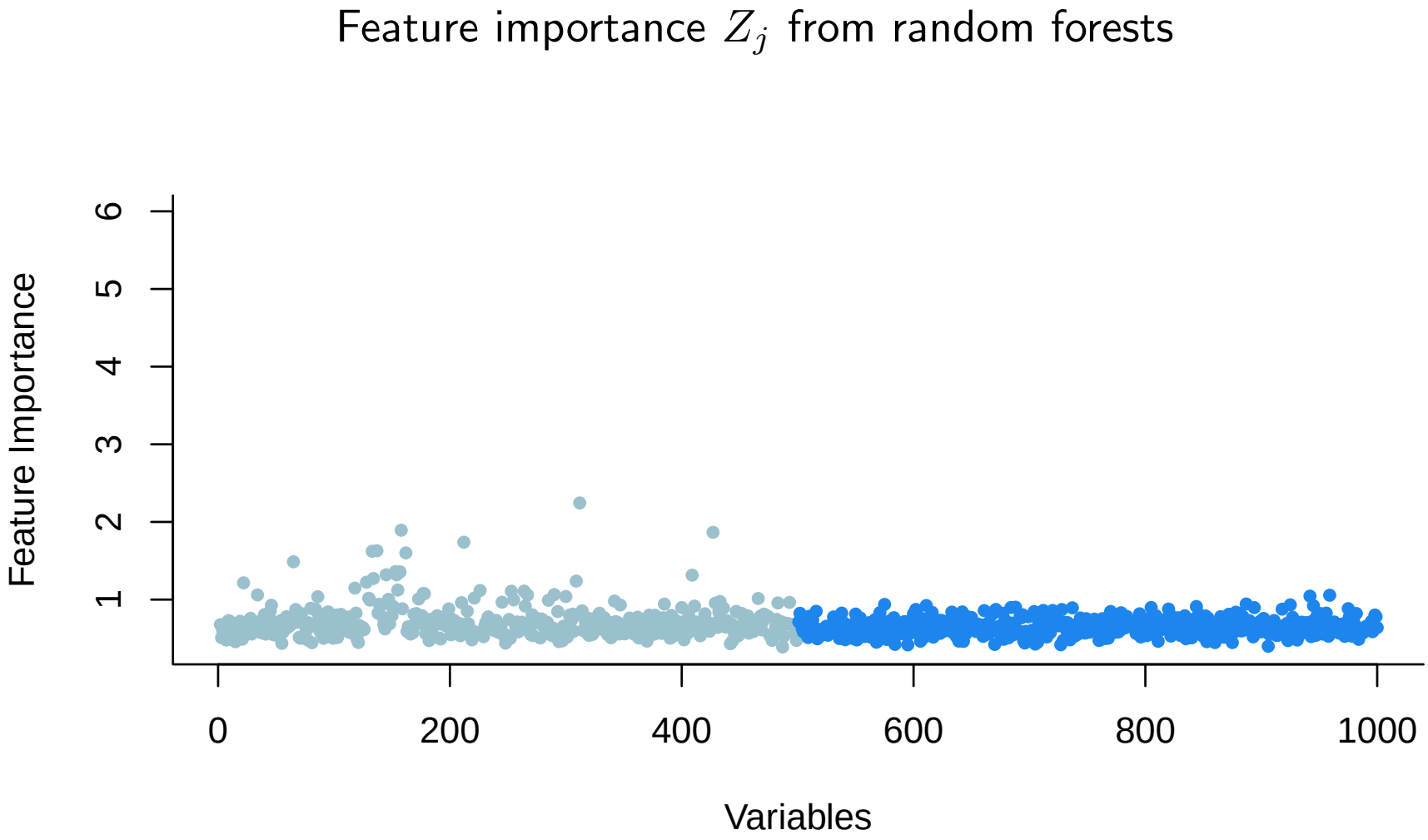
Feature importance $Z_j = \sup\{\lambda : |\hat{\beta}_j(\lambda)| \neq 0\}$



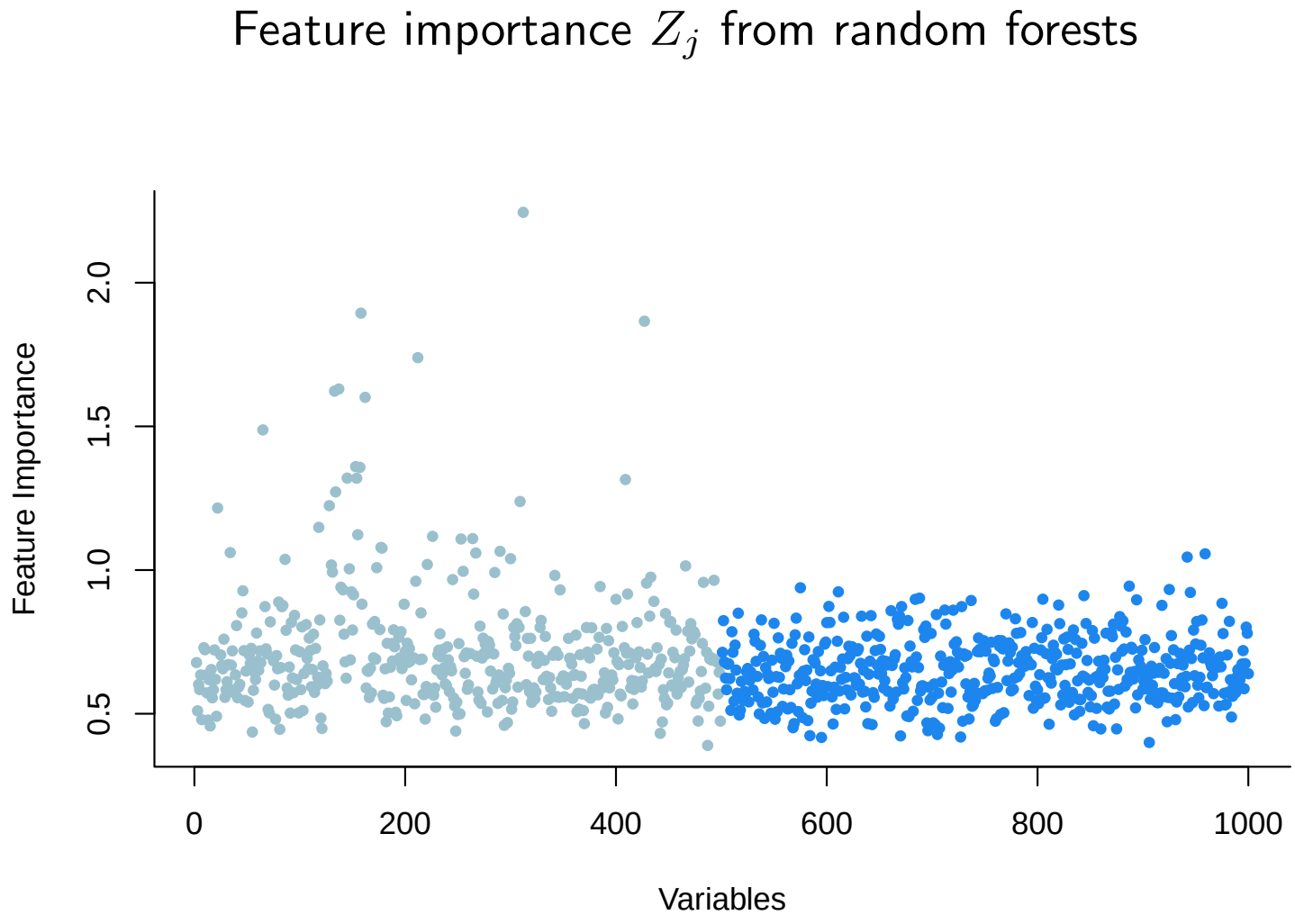
Gaussian dummies with binary response



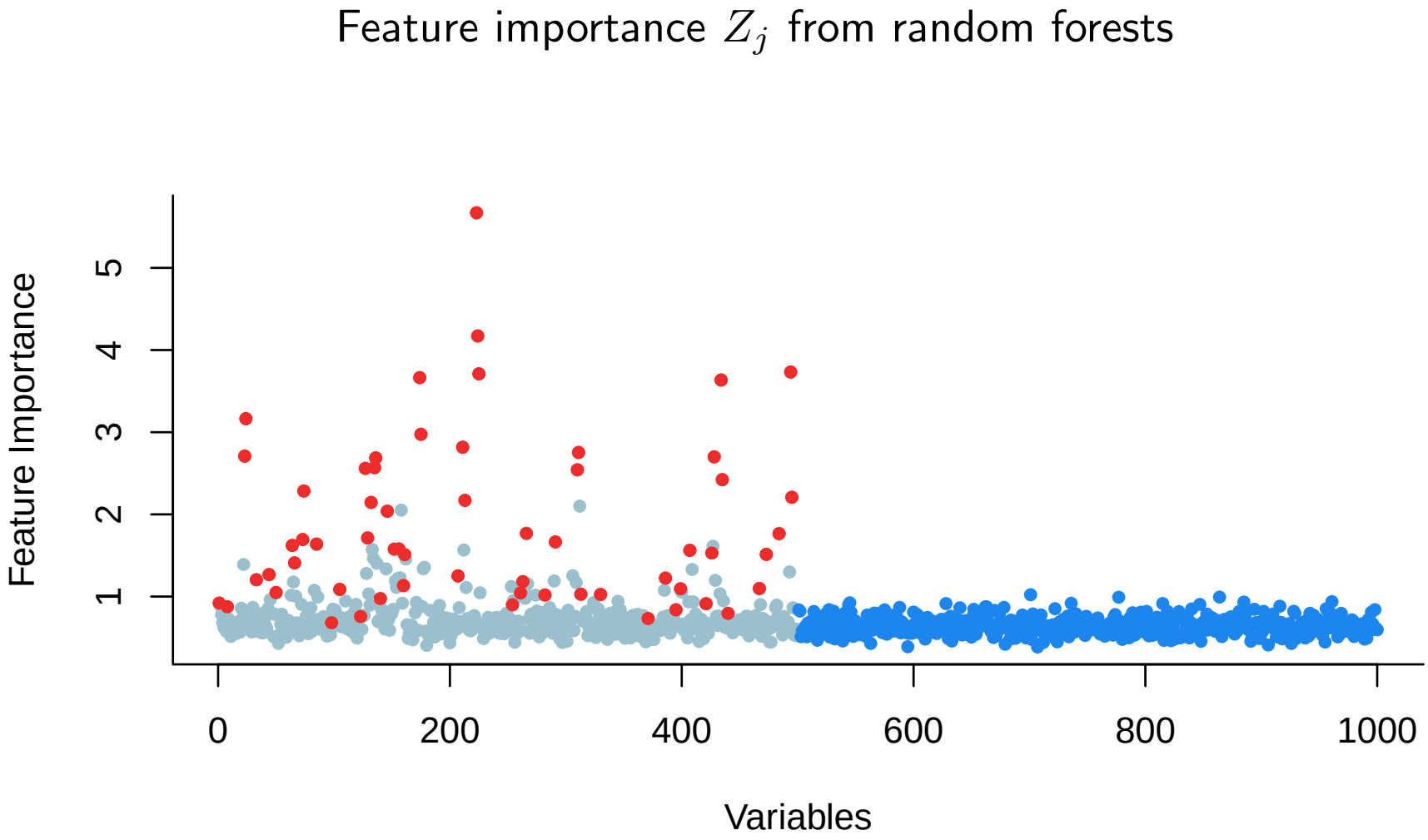
Gaussian dummies with binary response



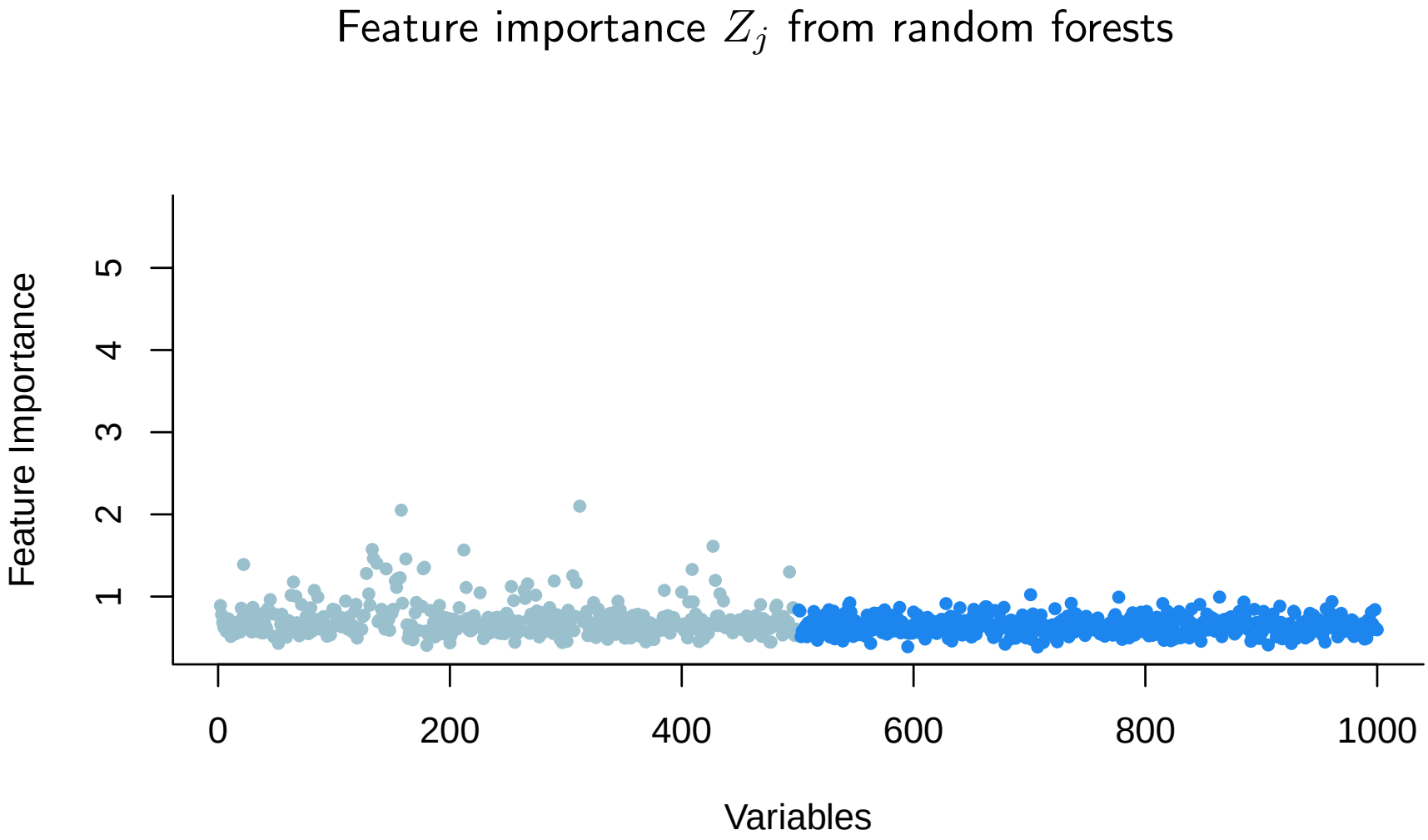
Gaussian dummies with binary response



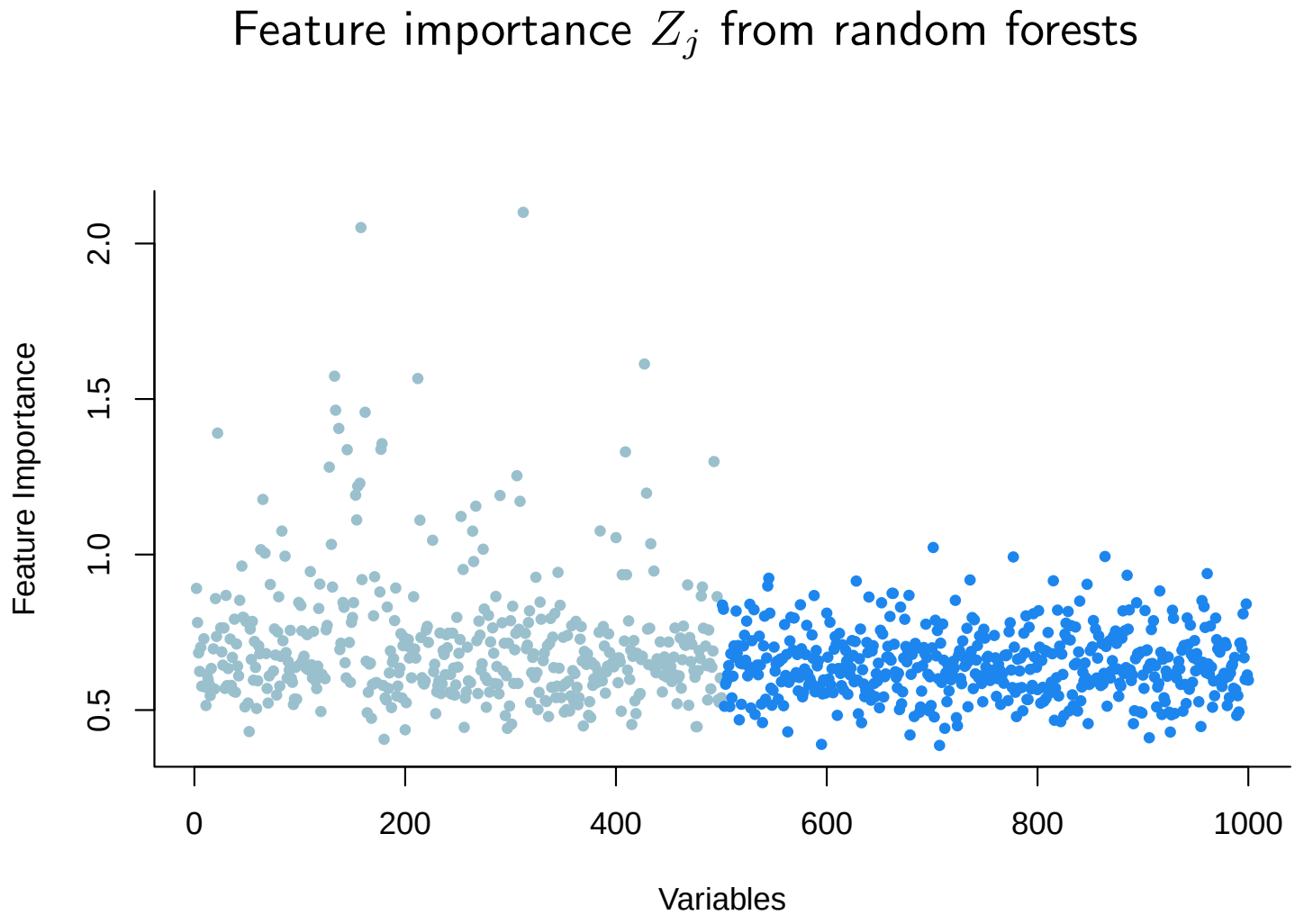
Permuted dummies with binary response



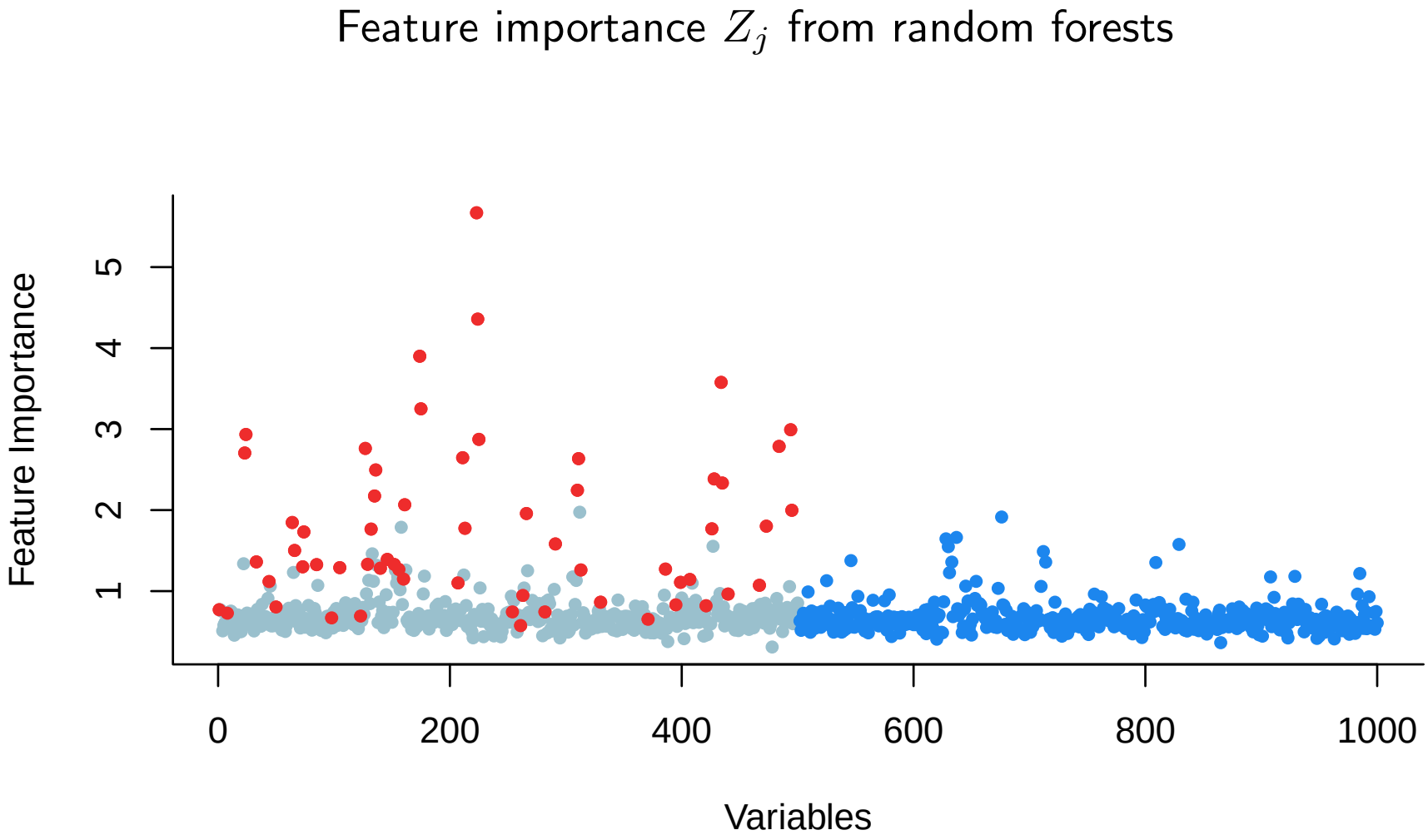
Permuted dummies with binary response



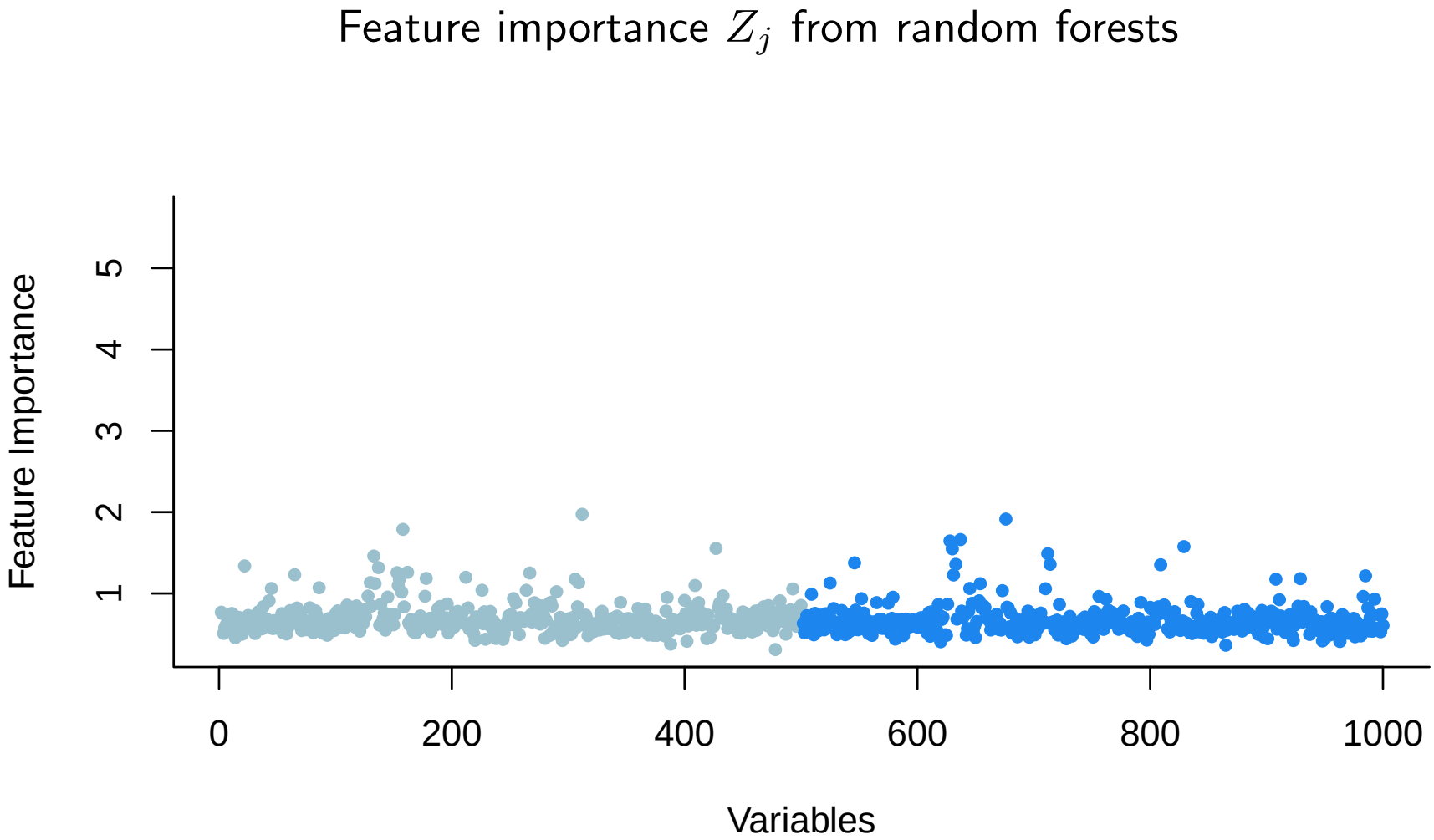
Permuted dummies with binary response



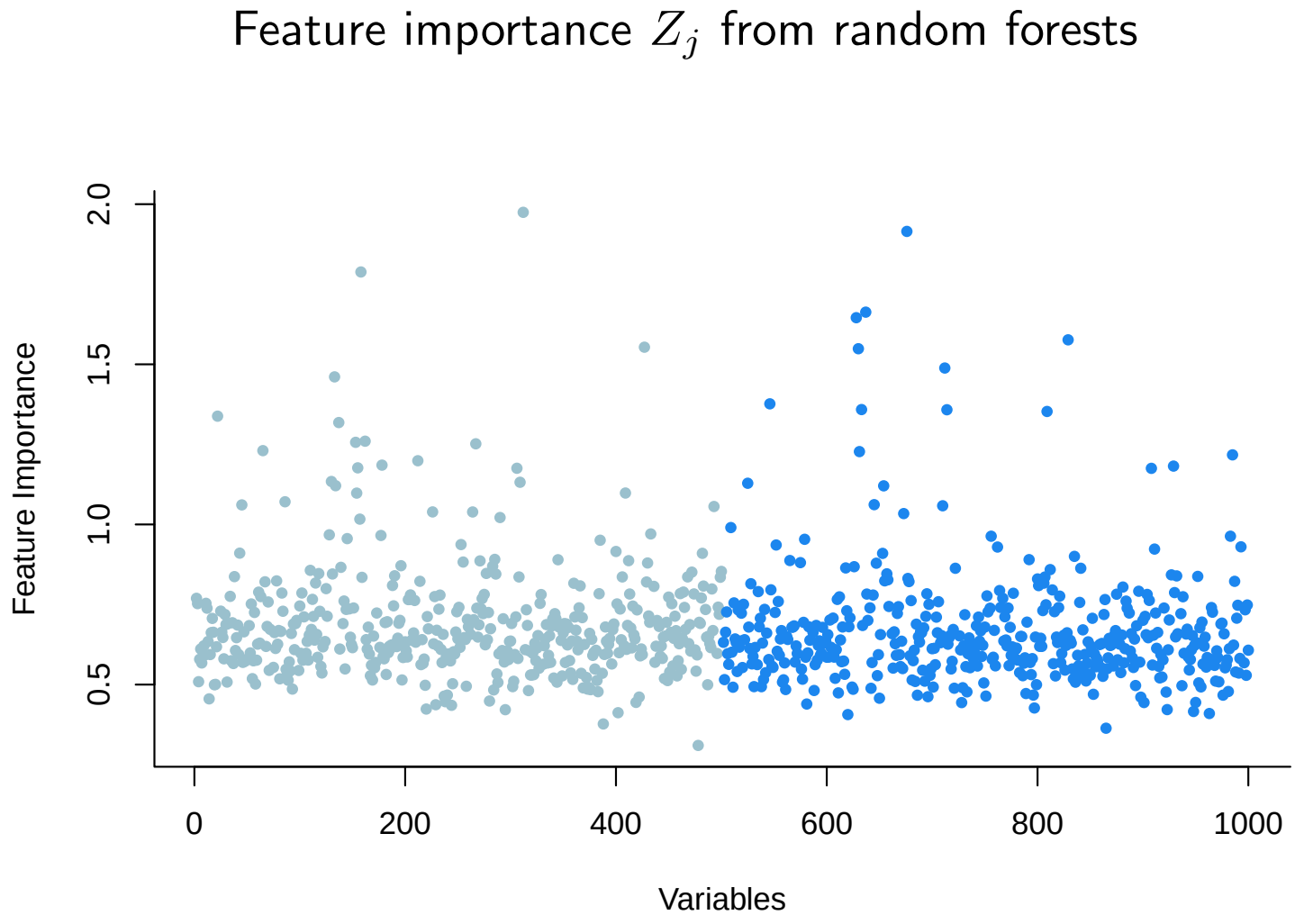
Knockoff dummies with binary response



Knockoff dummies with binary response



Knockoff dummies with binary response



Lessons

- Permuted dummies cannot serve as controls
- Having same distribution as original variables is not sufficient

Why permuted dummies cannot serve as controls?

Because there may be a relationship between X_1 and X_2 which makes the test statistic for a null variable larger than that of its permuted version

Why permuted dummies cannot serve as controls?

Because there may be a relationship between X_1 and X_2 which makes the test statistic for a null variable larger than that of its permuted version

- Linear model with two variables X_1 and X_2

$$Y = X_1 + \epsilon$$

Why permuted dummies cannot serve as controls?

Because there may be a relationship between X_1 and X_2 which makes the test statistic for a null variable larger than that of its permuted version

- Linear model with two variables X_1 and X_2

$$Y = X_1 + \epsilon$$

- Standardized variables (mean zero & variance one) with

$$\text{cor}(X_1, X_2) = 0.5$$

Why permuted dummies cannot serve as controls?

Because there may be a relationship between X_1 and X_2 which makes the test statistic for a null variable larger than that of its permuted version

- Linear model with two variables X_1 and X_2

$$Y = X_1 + \epsilon$$

- Standardized variables (mean zero & variance one) with

$$\text{cor}(X_1, X_2) = 0.5$$

- Marginal correlations

$$\mathbb{E}(YX_1) = 1 \qquad \mathbb{E}(YX_2) = \mathbb{E}(X_1 + \epsilon)X_2 = 0.5$$

Why permuted dummies cannot serve as controls?

Because there may be a relationship between X_1 and X_2 which makes the test statistic for a null variable larger than that of its permuted version

- Linear model with two variables X_1 and X_2

$$Y = X_1 + \epsilon$$

- Standardized variables (mean zero & variance one) with

$$\text{cor}(X_1, X_2) = 0.5$$

- Marginal correlations

$$\mathbb{E}(YX_1) = 1 \qquad \mathbb{E}(YX_2) = \mathbb{E}(X_1 + \epsilon)X_2 = 0.5$$

- Marginal correlations with permuted features

$$\mathbb{E}(X_{2,\pi}Y) = 0$$

Why permuted dummies cannot serve as controls?

Because there may be a relationship between X_1 and X_2 which makes the test statistic for a null variable larger than that of its permuted version

- Linear model with two variables X_1 and X_2

$$Y = X_1 + \epsilon$$

- Standardized variables (mean zero & variance one) with

$$\text{cor}(X_1, X_2) = 0.5$$

- Marginal correlations

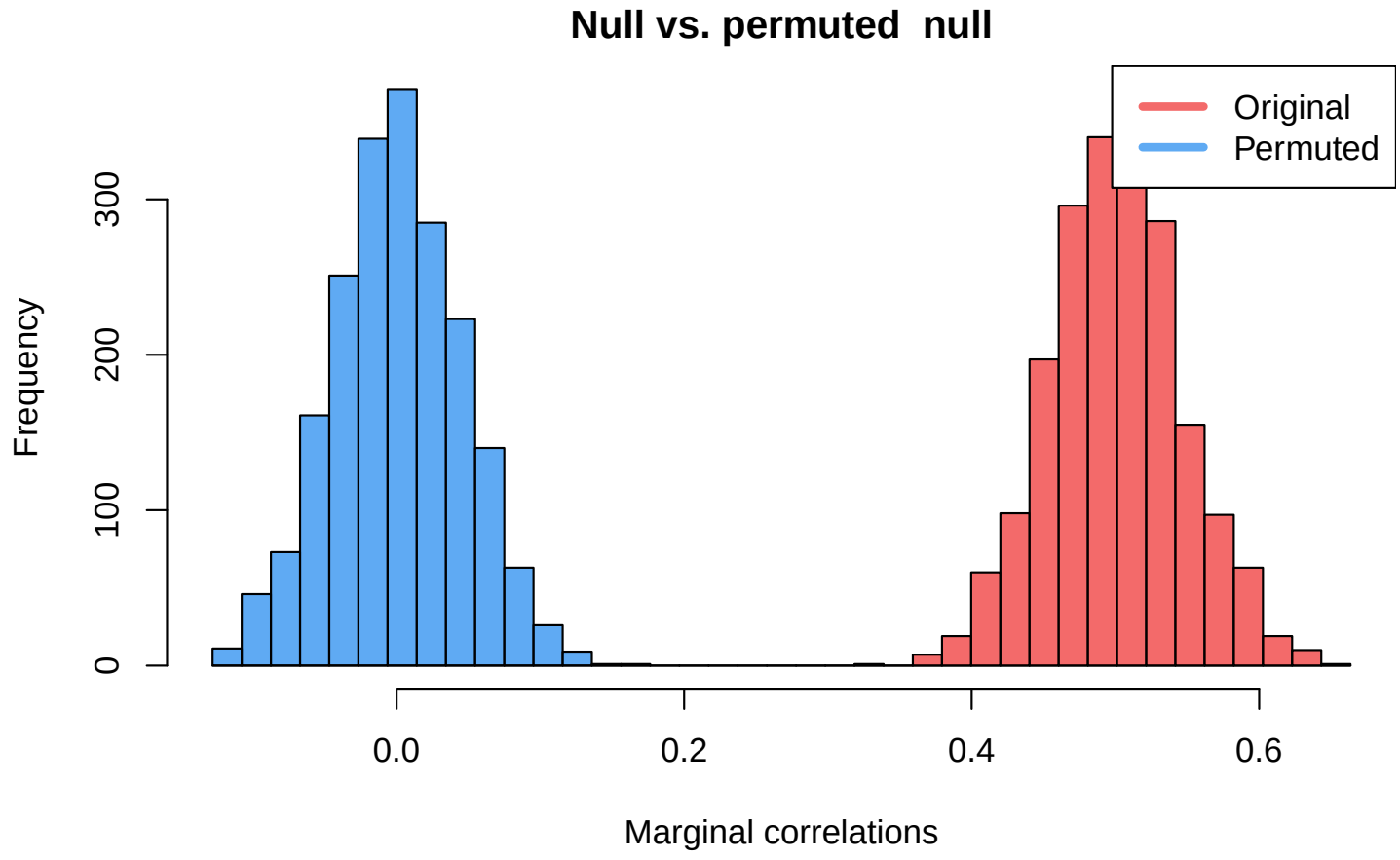
$$\mathbb{E}(YX_1) = 1 \qquad \mathbb{E}(YX_2) = \mathbb{E}(X_1 + \epsilon)X_2 = 0.5$$

- Marginal correlations with permuted features

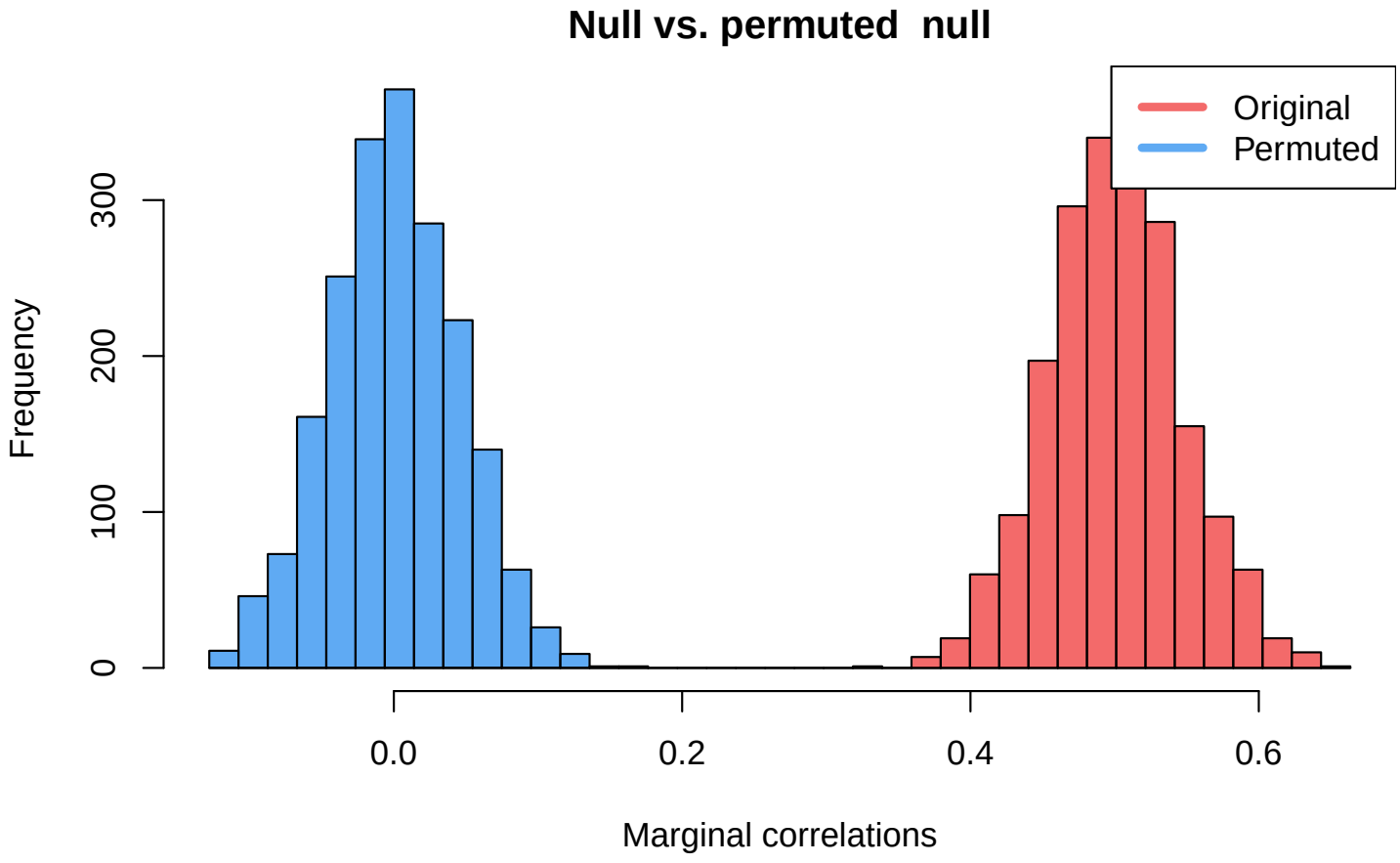
$$\mathbb{E}(X_{2,\pi}Y) = 0$$

Statistics $X_2^\top y$ and $X_{2,\pi}^\top y$ cannot be the same!

Why permuted dummies cannot serve as controls?



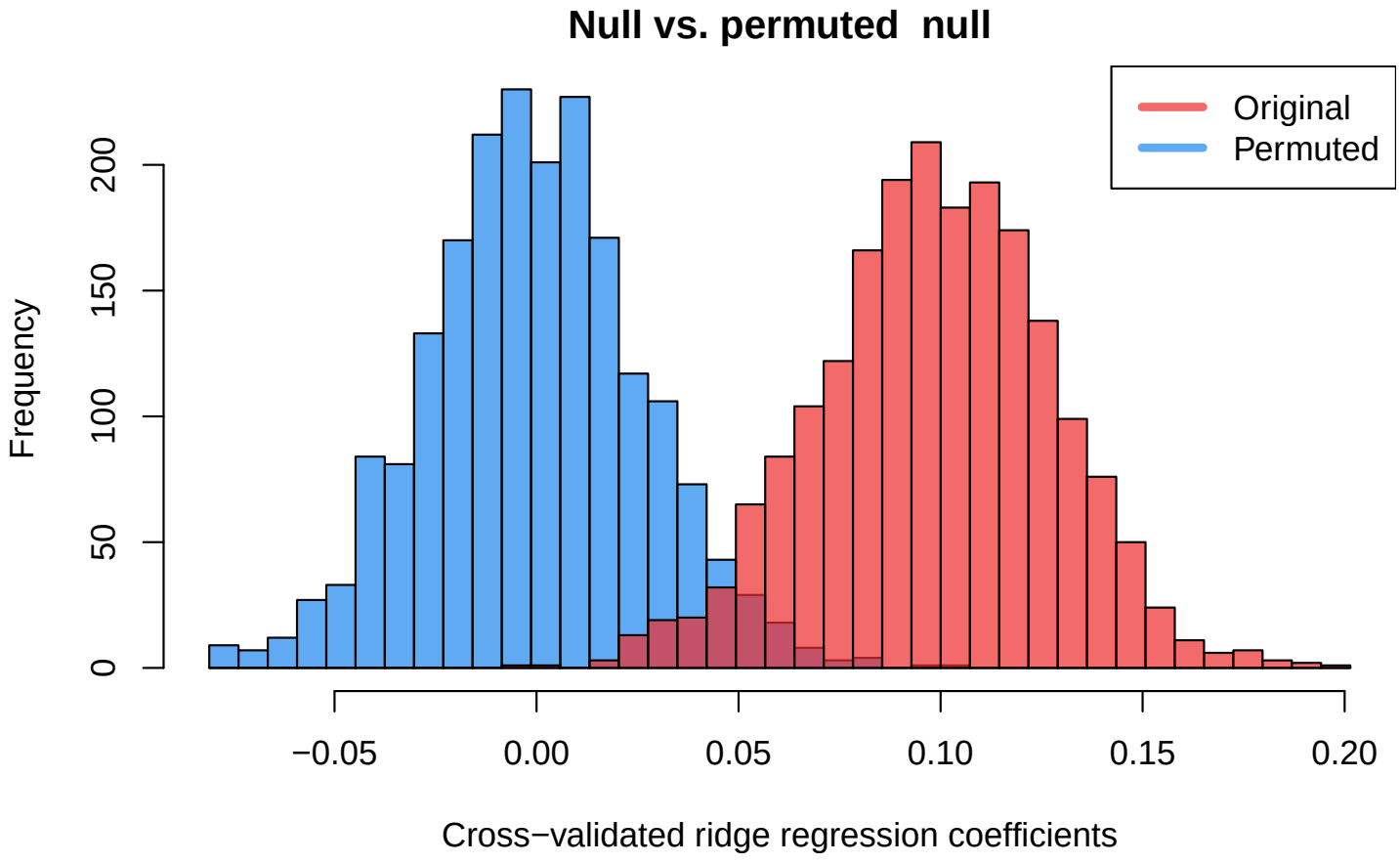
Why permuted dummies cannot serve as controls?



How could a permuted feature statistic possibly serve as control?

$$(Z_2, Z_{2,\pi}) = t(X_1^\top y, X_2^\top y, X_{1,\pi}^\top y, X_{2,\pi}^\top y)$$

Why permuted dummies cannot serve as controls?



How could a permuted feature statistic possibly serve as control?

$$(Z_2, Z_{2,\pi}) = t(X_1^\top y, X_2^\top y, X_{1,\pi}^\top y, X_{2,\pi}^\top y)$$

Next lecture

Build knockoff features $(\tilde{X}_1, \tilde{X}_2)$ such that if X_2 is null

$$\tilde{X}_2^\top y \stackrel{d}{=} X_2^\top y$$

X_2 'correlated' with X_1 $\quad\Longrightarrow\quad$ $\tilde{X}_2$ 'correlated' with X_1 (and $\tilde{X}_1$) in same way

$\rightsquigarrow$ different from permuted features

$\quad\Longrightarrow\quad$ $X_{2,\pi}$ 'correlated' with $X_{1,\pi}$ in same way
$X_{2,\pi}$ uncorrelated with X_1

Next lecture

Build knockoff features $(\tilde{X}_1, \tilde{X}_2)$ such that if X_2 is null

$$\tilde{X}_2^\top y \stackrel{d}{=} X_2^\top y$$

X_2 'correlated' with X_1 $\implies$ $\tilde{X}_2$ 'correlated' with X_1 (and $\tilde{X}_1$) in same way

$\rightsquigarrow$ different from permuted features

$\implies$ $X_{2,\pi}$ 'correlated' with $X_{1,\pi}$ in same way
$X_{2,\pi}$ uncorrelated with X_1

If X_j is null, then we would like that our black box is equally likely to pick X_j or $\tilde{X}_j$ regardless of other covariates selected

The Knockoff Inference Machine

Originally proposed with Barber
Re-interpreted with Fan, Janson & Lv

Hypotheses and test statistics

Examples of questions of interest

- (Parametric) linear model $Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p + \epsilon$

$$H_j : \beta_j = 0$$

- Nonparametric model

$$H_j : Y \perp\!\!\!\perp X_j \,|\, X_{-j}$$

Hypotheses and test statistics

Examples of questions of interest

- (Parametric) linear model $Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p + \epsilon$

$$H_j : \beta_j = 0$$

- Nonparametric model

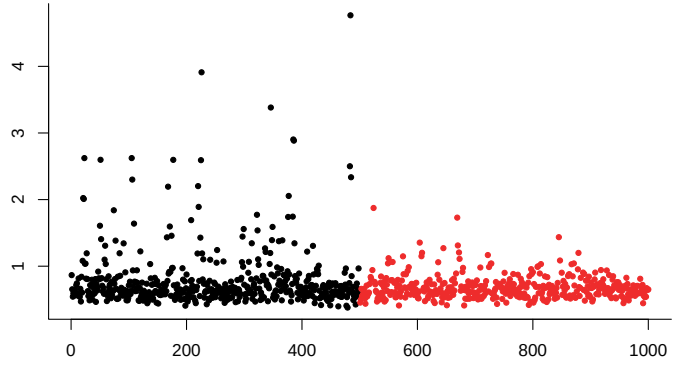
$$H_j : Y \perp\!\!\!\perp X_j \,|\, X_{-j}$$

Examples of test statistics Z_j

- Some random forest feature importance statistic
- Value of square-root lasso coefficient
- Posterior probability calculated according to some Bayesian model
- Your favorite deep learning feature
- ...

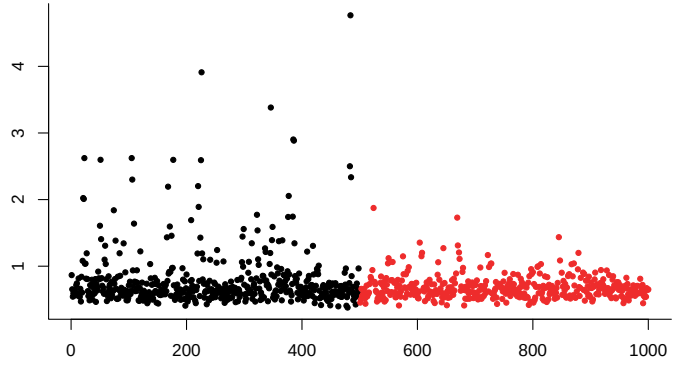
Exchangeability of feature importance statistics

$$(\underbrace{Z_1, \ldots, Z_p}_{\text{originals}}, \underbrace{\tilde{Z}_1, \ldots, \tilde{Z}_p}_{\text{knockoffs}}) = z([\boldsymbol{X}, \tilde{\boldsymbol{X}}], \boldsymbol{y})$$



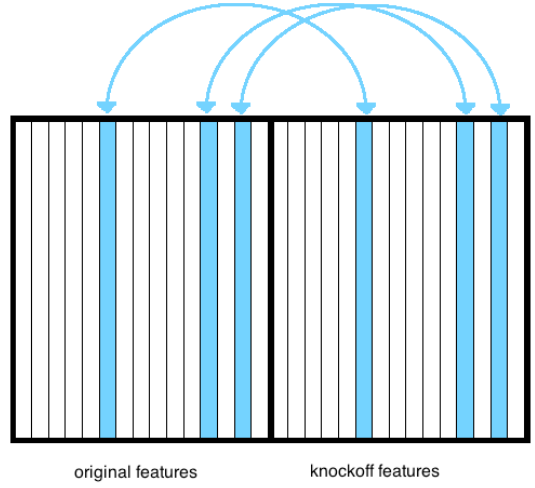
Exchangeability of feature importance statistics

$$(\underbrace{Z_1, \ldots, Z_p}_{\text{originals}}, \underbrace{\tilde{Z}_1, \ldots, \tilde{Z}_p}_{\text{knockoffs}}) = z([\boldsymbol{X}, \tilde{\boldsymbol{X}}], \boldsymbol{y})$$



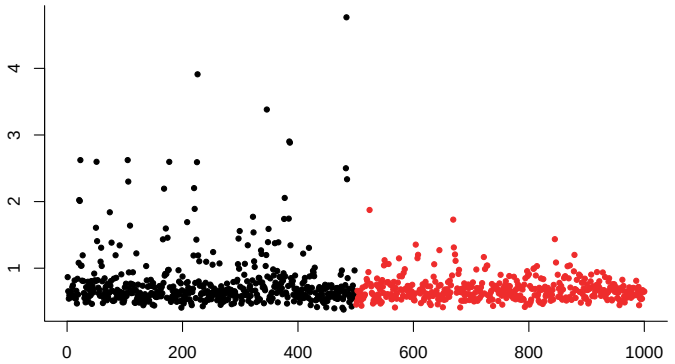
Swapping originals and knockoffs swaps the Z's

$$\overbrace{(Z_1, \tilde{Z}_2, \tilde{Z}_3, \tilde{Z}_1, Z_2, Z_3)}^{(Z,\, \tilde{Z})_{\text{swap}\{2,3\}}} = z([\boldsymbol{X}, \tilde{\boldsymbol{X}}]_{\text{swap}\{2,3\}}, \boldsymbol{y})$$



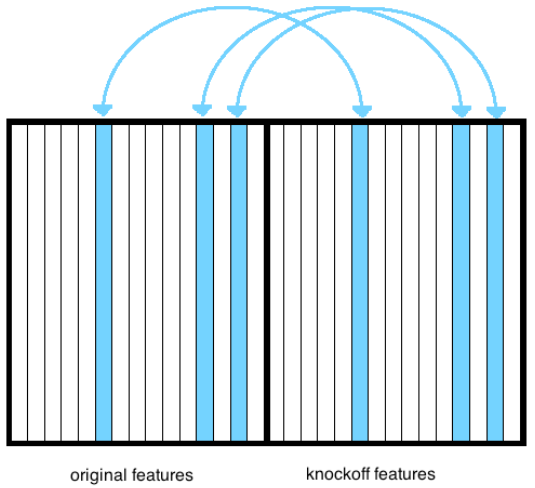
Exchangeability of feature importance statistics

$$(\underbrace{Z_1, \ldots, Z_p}_{\text{originals}}, \underbrace{\tilde{Z}_1, \ldots, \tilde{Z}_p}_{\text{knockoffs}}) = z([\boldsymbol{X}, \tilde{\boldsymbol{X}}], \boldsymbol{y})$$



Swapping originals and knockoffs swaps the Z's

$$\overbrace{(Z_1, \tilde{Z}_2, \tilde{Z}_3, \tilde{Z}_1, Z_2, Z_3)}^{(Z,\, \tilde{Z})_{\text{swap}\{2,3\}}} = z([\boldsymbol{X}, \tilde{\boldsymbol{X}}]_{\text{swap}\{2,3\}}, \boldsymbol{y})$$



Next Lecture

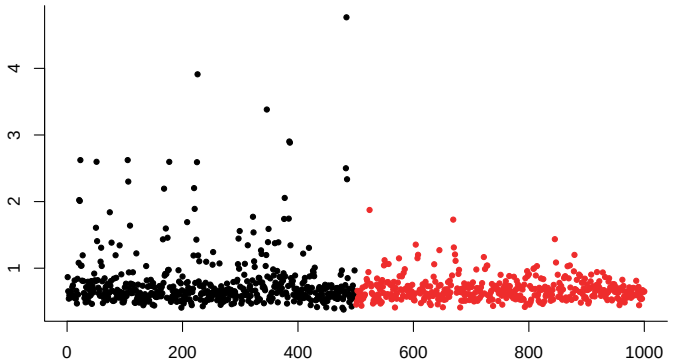
Can construct knockoff features such that

$$j \in \mathcal{H}_0 \implies (Z_j, \tilde{Z}_j) \overset{d}{=} (\tilde{Z}_j, Z_j)$$

more generally $\mathcal{T} \subset \mathcal{H}_0 \implies (Z, \tilde{Z})_{\text{swap}(\mathcal{T})} \overset{d}{=} (Z, \tilde{Z})$

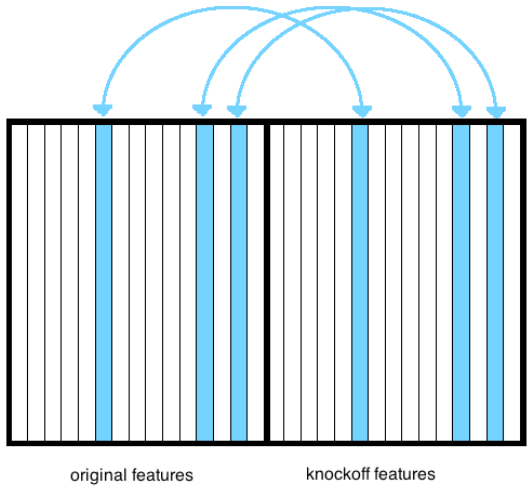
Exchangeability of feature importance statistics

$$(\underbrace{Z_1, \ldots, Z_p}_{\text{originals}}, \underbrace{\tilde{Z}_1, \ldots, \tilde{Z}_p}_{\text{knockoffs}}) = z([\boldsymbol{X}, \tilde{\boldsymbol{X}}], \boldsymbol{y})$$



Swapping originals and knockoffs swaps the Z's

$$\overbrace{(Z_1, \tilde{Z}_2, \tilde{Z}_3, \tilde{Z}_1, Z_2, Z_3)}^{(Z,\, \tilde{Z})_{\text{swap}\{2,3\}}} = z([\boldsymbol{X}, \tilde{\boldsymbol{X}}]_{\text{swap}\{2,3\}}, \boldsymbol{y})$$



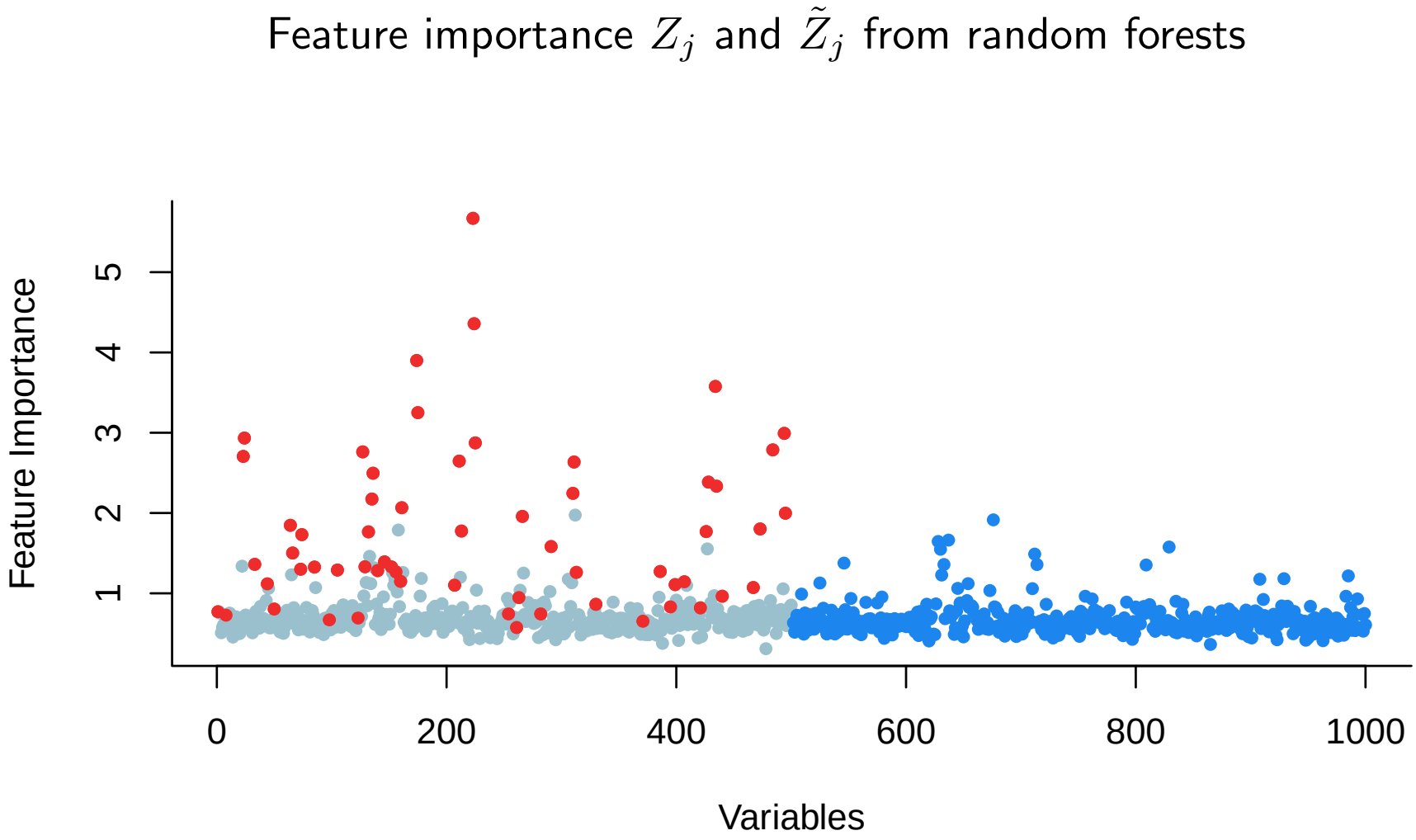
Next Lecture

Can construct knockoff features such that

$$j \in \mathcal{H}_0 \implies (Z_j, \tilde{Z}_j) \overset{d}{=} (\tilde{Z}_j, Z_j)$$

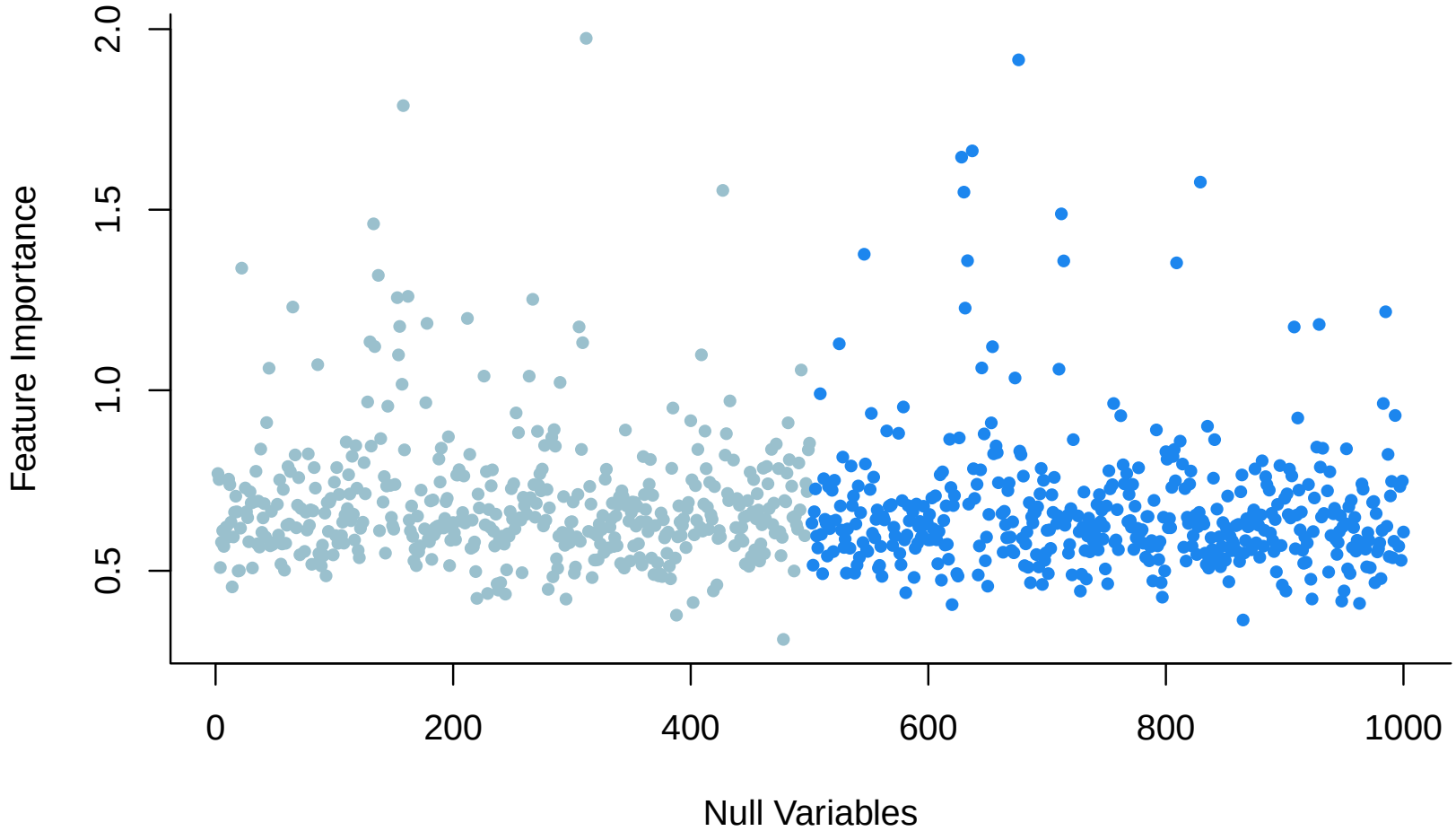
$$\text{e.g. } \{2,3\} \subset \mathcal{H}_0 \implies (Z_1, \tilde{Z}_2, \tilde{Z}_3, \tilde{Z}_1, Z_2, Z_3) \overset{d}{=} (Z_1, Z_2, Z_3, \tilde{Z}_1, \tilde{Z}_2, \tilde{Z}_3)$$

Knockoffs with binary response



Knockoffs with binary response

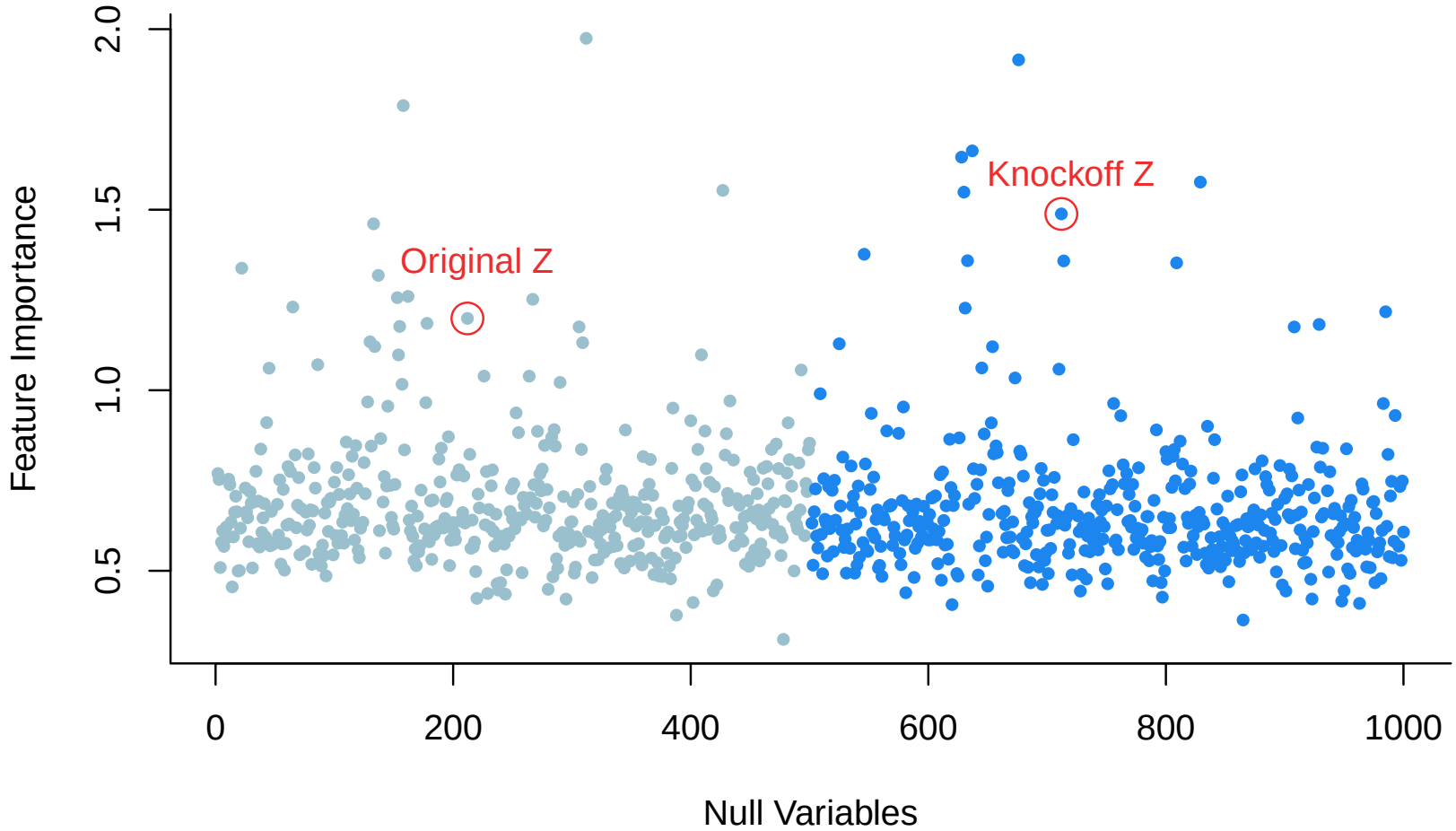
Feature importance Z_j and $\tilde{Z}_j$ from random forests



$$(Z_j, \tilde{Z}_j) \stackrel{d}{=} (\tilde{Z}_j, Z_j)$$

Knockoffs with binary response

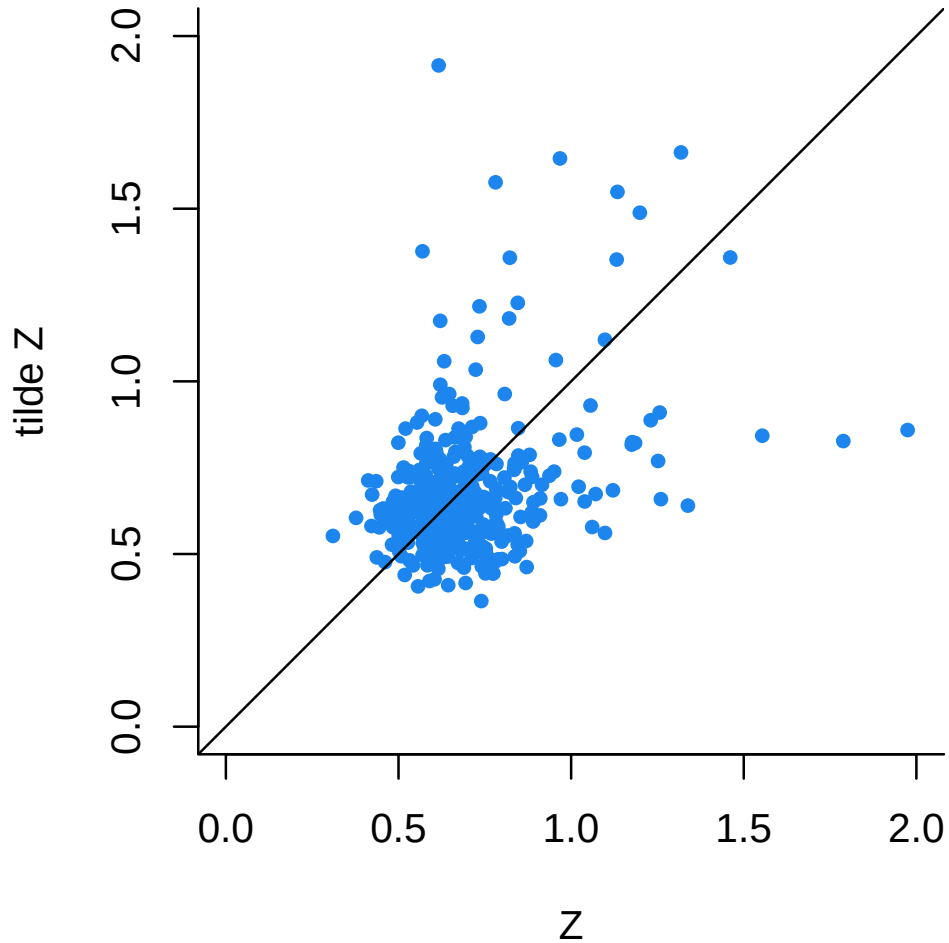
Feature importance Z_j and $\tilde{Z}_j$ from random forests



$$(Z_j, \tilde{Z}_j) \stackrel{d}{=} (\tilde{Z}_j, Z_j)$$

Knockoffs with binary response

Feature importance Z_j and $\tilde{Z}_j$ from random forests



$$(Z_j, \tilde{Z}_j) \stackrel{d}{=} (\tilde{Z}_j, Z_j)$$

Knockoffs-adjusted scores

Adjusted scores W_j with flip-sign property

Combine Z_j and $\tilde{Z}_j$ into single (knockoff) score W_j

$$W_j = w_j(Z_j, \tilde{Z}_j) \qquad w_j(\tilde{Z}_j, Z_j) = -w_j(Z_j, \tilde{Z}_j)$$

Knockoffs-adjusted scores

Adjusted scores W_j with flip-sign property

Combine Z_j and $\tilde{Z}_j$ into single (knockoff) score W_j

$$W_j = w_j(Z_j, \tilde{Z}_j) \qquad w_j(\tilde{Z}_j, Z_j) = -w_j(Z_j, \tilde{Z}_j)$$

e.g. $\quad W_j = Z_j - \tilde{Z}_j$

Knockoffs-adjusted scores

Adjusted scores W_j with flip-sign property

Combine Z_j and $\tilde{Z}_j$ into single (knockoff) score W_j

$$W_j = w_j(Z_j, \tilde{Z}_j) \qquad w_j(\tilde{Z}_j, Z_j) = -w_j(Z_j, \tilde{Z}_j)$$

e.g. $\qquad W_j = Z_j - \tilde{Z}_j \qquad W_j = Z_j \vee \tilde{Z}_j \cdot \begin{cases} +1 & Z_j > \tilde{Z}_j \\ -1 & \text{else} \end{cases}$

Knockoffs-adjusted scores

Adjusted scores W_j with flip-sign property

Combine Z_j and $\tilde{Z}_j$ into single (knockoff) score W_j

$$W_j = w_j(Z_j, \tilde{Z}_j) \qquad w_j(\tilde{Z}_j, Z_j) = -w_j(Z_j, \tilde{Z}_j)$$

e.g. $\quad W_j = Z_j - \tilde{Z}_j \qquad W_j = Z_j \vee \tilde{Z}_j \cdot \begin{cases} +1 & Z_j > \tilde{Z}_j \\ -1 & \text{else} \end{cases}$

- Null W_j's are symmetrically distributed

Knockoffs-adjusted scores

Adjusted scores W_j with flip-sign property

Combine Z_j and $\tilde{Z}_j$ into single (knockoff) score W_j

$$W_j = w_j(Z_j, \tilde{Z}_j) \qquad w_j(\tilde{Z}_j, Z_j) = -w_j(Z_j, \tilde{Z}_j)$$

e.g. $\quad W_j = Z_j - \tilde{Z}_j \qquad W_j = Z_j \vee \tilde{Z}_j \cdot \begin{cases} +1 & Z_j > \tilde{Z}_j \\ -1 & \text{else} \end{cases}$

- Null W_j's are symmetrically distributed
- Conditional on $|W|$, signs of null W_j's are i.i.d. coin flips

Knockoffs-adjusted scores

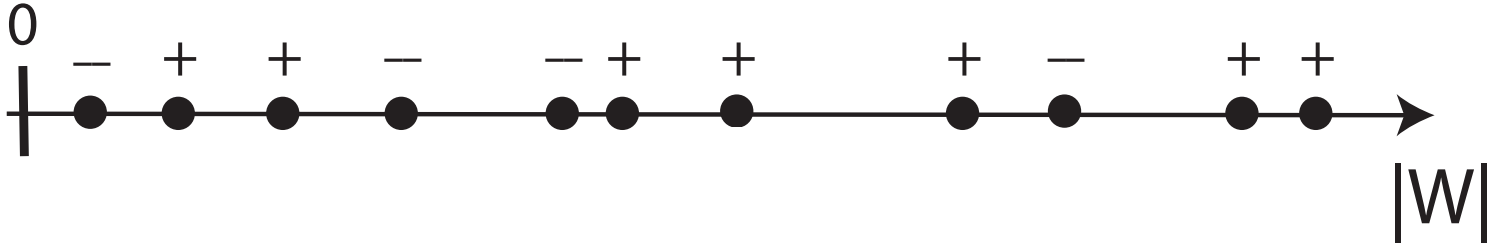
Adjusted scores W_j with flip-sign property

Combine Z_j and $\tilde{Z}_j$ into single (knockoff) score W_j

$$W_j = w_j(Z_j, \tilde{Z}_j) \qquad w_j(\tilde{Z}_j, Z_j) = -w_j(Z_j, \tilde{Z}_j)$$

e.g. $\quad W_j = Z_j - \tilde{Z}_j \qquad W_j = Z_j \vee \tilde{Z}_j \cdot \begin{cases} +1 & Z_j > \tilde{Z}_j \\ -1 & \text{else} \end{cases}$

- Null W_j's are symmetrically distributed
- Conditional on $|W|$, signs of null W_j's are i.i.d. coin flips



Knockoffs-adjusted scores

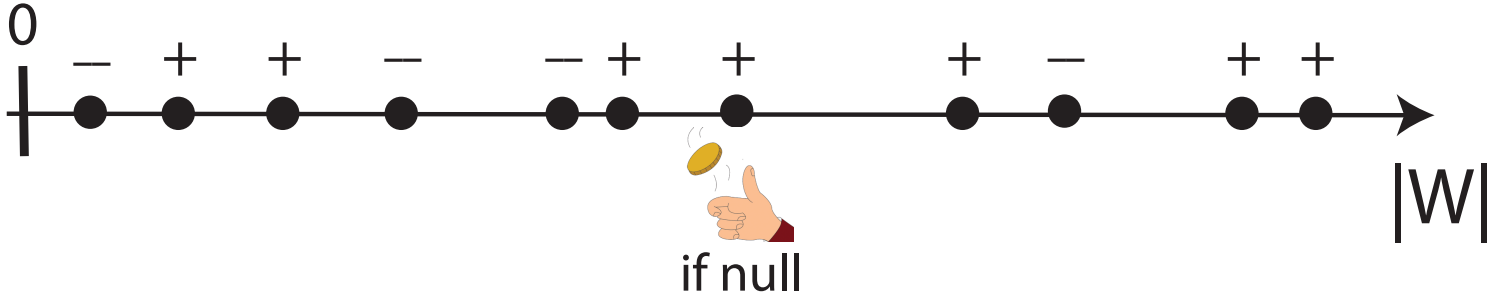
Adjusted scores W_j with flip-sign property

Combine Z_j and $\tilde{Z}_j$ into single (knockoff) score W_j

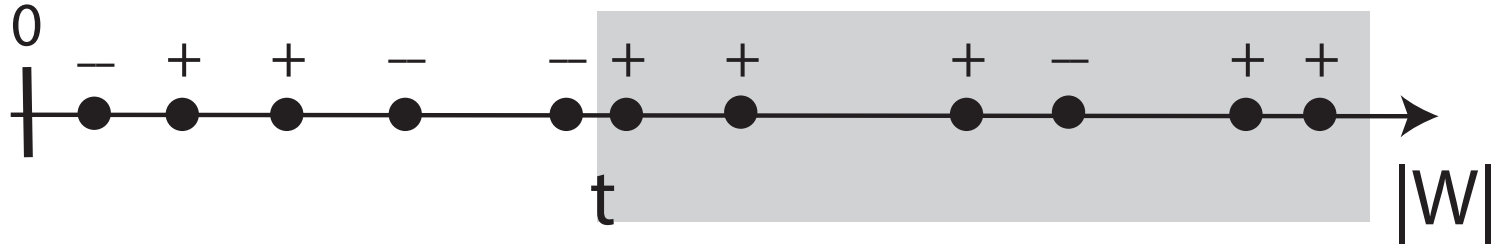
$$W_j = w_j(Z_j, \tilde{Z}_j) \qquad w_j(\tilde{Z}_j, Z_j) = -w_j(Z_j, \tilde{Z}_j)$$

e.g. $\quad W_j = Z_j - \tilde{Z}_j \qquad W_j = Z_j \vee \tilde{Z}_j \cdot \begin{cases} +1 & Z_j > \tilde{Z}_j \\ -1 & \text{else} \end{cases}$

- Null W_j's are symmetrically distributed
- Conditional on $|W|$, signs of null W_j's are i.i.d. coin flips

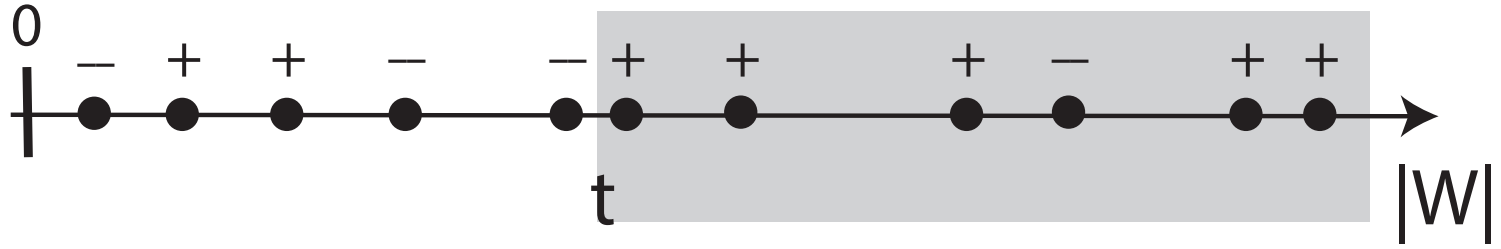


Knockoff estimate of FDR



Interested in selecting $\{j : W_j \geq t\}$

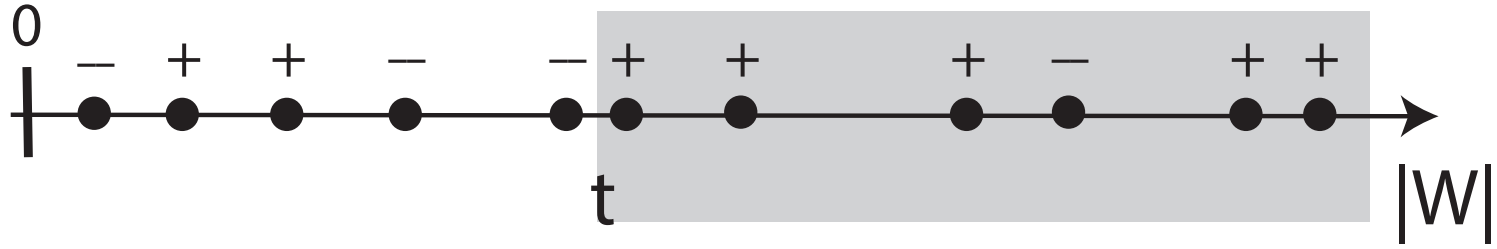
Knockoff estimate of FDR



Interested in selecting $\{j : W_j \geq t\}$

$$\text{FDP}(t) = \frac{\#\{j \text{ null } : W_j \geq t\}}{\#\{j : W_j \geq t\} \vee 1}$$

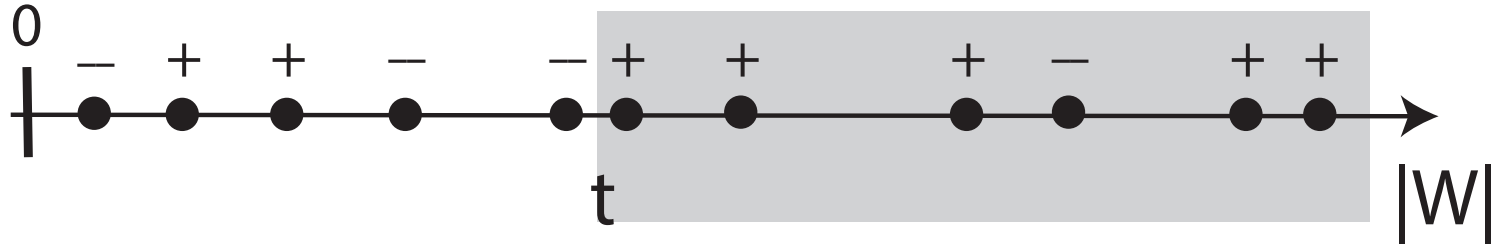
Knockoff estimate of FDR



Interested in selecting $\{j : W_j \geq t\}$

$$\text{FDP}(t) = \frac{\#\{j \text{ null } : W_j \geq t\}}{\#\{j : W_j \geq t\} \vee 1} \approx \frac{\#\{j \text{ null } : W_j \leq -t\}}{\#\{j : W_j \geq t\} \vee 1}$$

Knockoff estimate of FDR

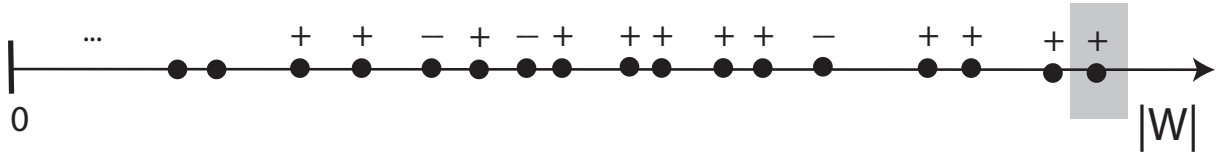


Interested in selecting $\{j : W_j \geq t\}$

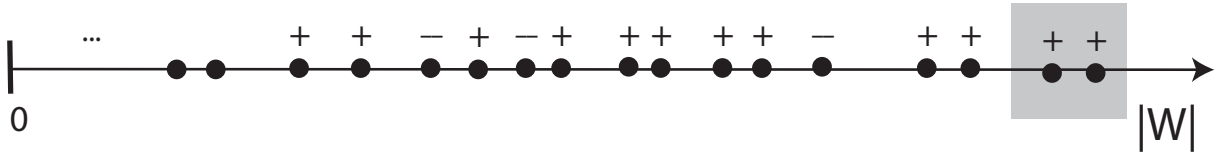
$$\mathsf{FDP}(t) = \frac{\#\{j \text{ null } : W_j \geq t\}}{\#\{j : W_j \geq t\} \vee 1} \approx \frac{\#\{j \text{ null } : W_j \leq -t\}}{\#\{j : W_j \geq t\} \vee 1}$$

$$\leq \frac{\#\{j : W_j \leq -t\}}{\#\{j : W_j \geq t\} \vee 1} := \widehat{\mathsf{FDP}}(t)$$

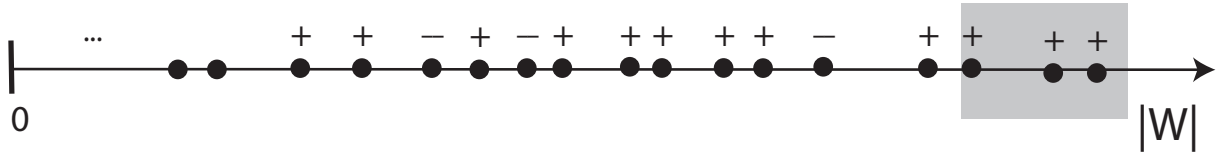
Selection (via sequential testing)



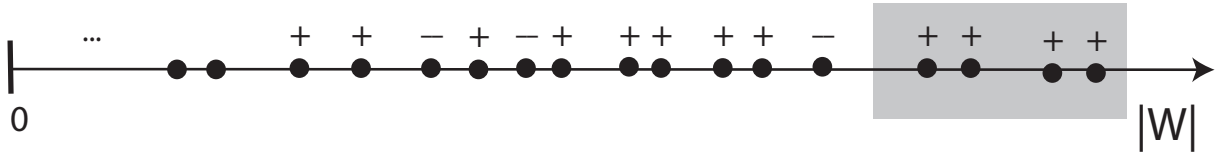
Selection (via sequential testing)



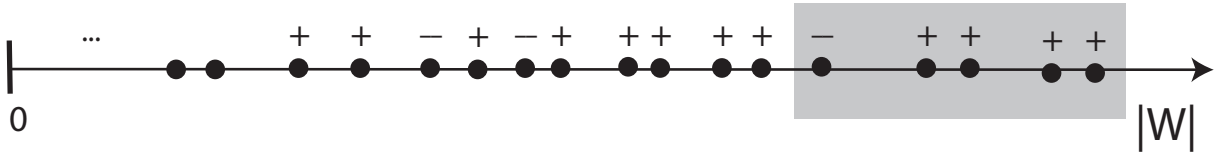
Selection (via sequential testing)



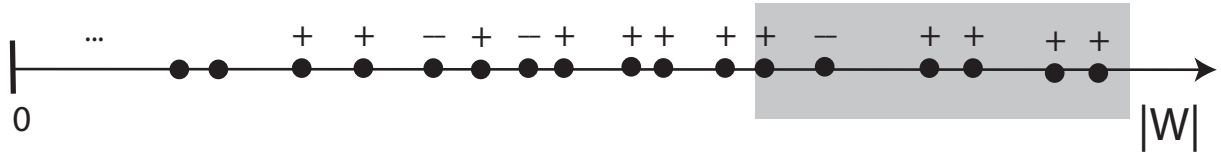
Selection (via sequential testing)



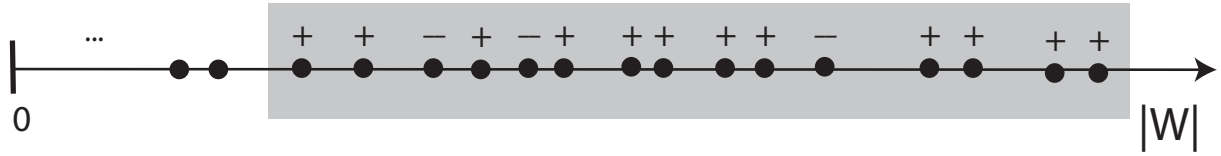
Selection (via sequential testing)



Selection (via sequential testing)

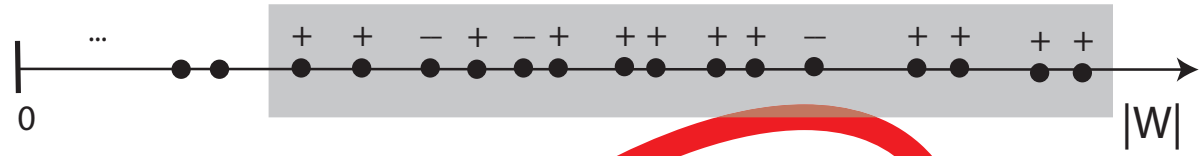


Selection (via sequential testing)

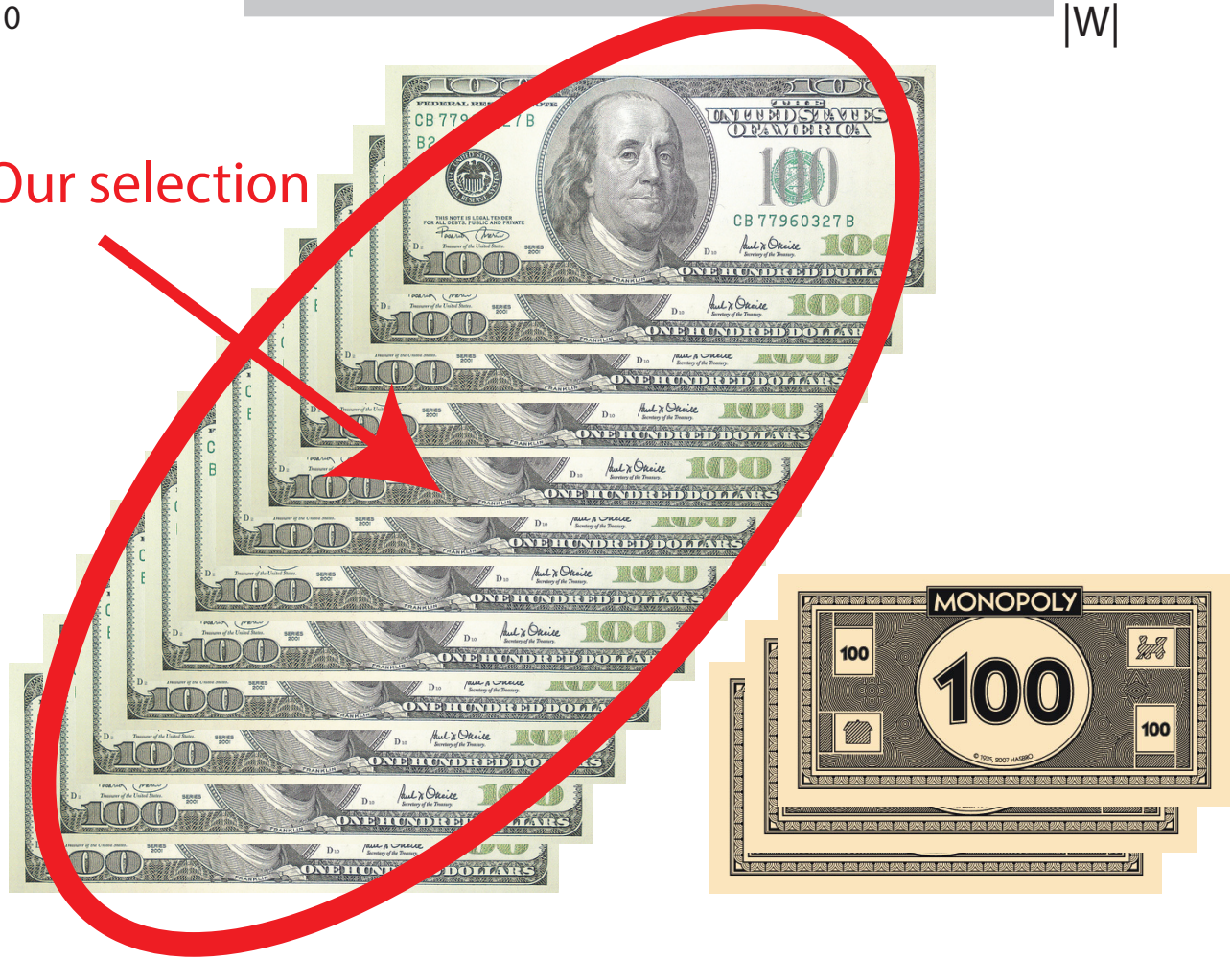


Step-up rule: stop last time ratio between '-' and '+' below target FDR level

Selection (via sequential testing)



Our selection



Select '+'s

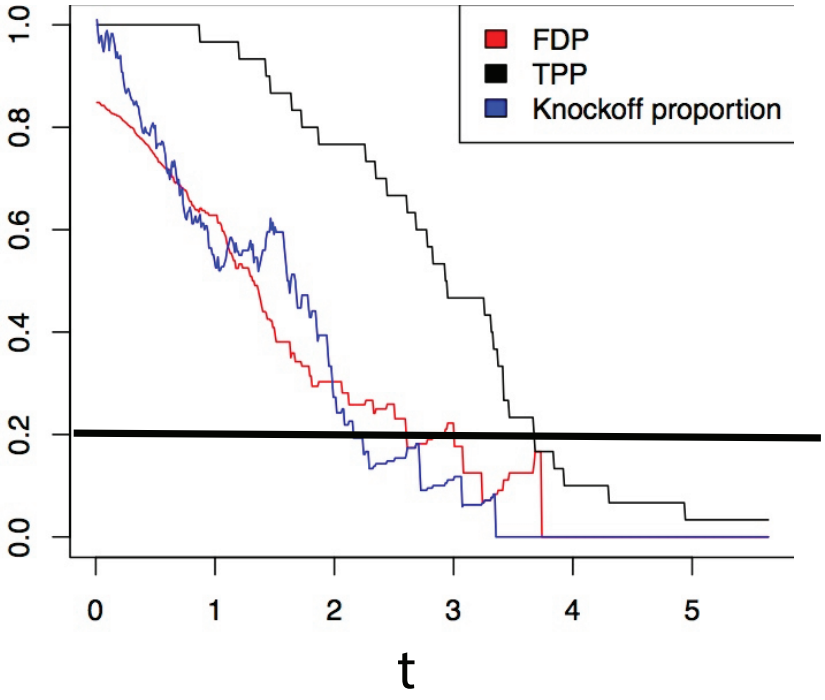
Step-up rule: stop last time ratio between '-' and '+' below target FDR level

FDR control

$$\mathcal{S}^{\pm}(t) = \{j : |W_j| \geq t \text{ and } \operatorname{sgn}(W_j) = \pm\}$$

$$\tau = \min\left\{t : \widehat{\mathsf{FDP}}(t) = \frac{1+|\mathcal{S}^{-}(t)|}{1 \vee |\mathcal{S}^{+}(t)|} \leq q\right\}$$

$$\hat{\mathcal{S}} = \{W_j \geq \tau\}$$

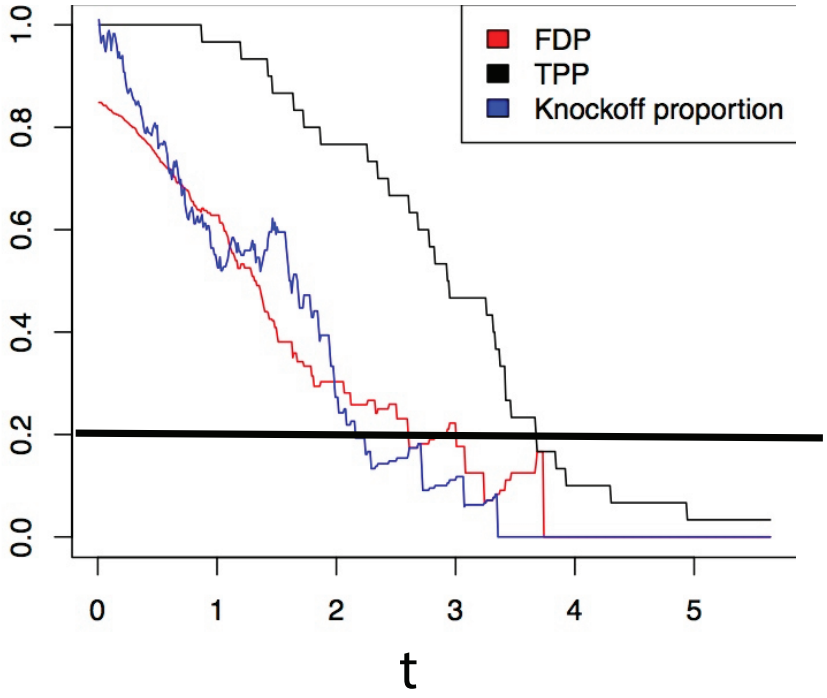


FDR control

$$\mathcal{S}^{\pm}(t) = \{j : |W_j| \geq t \text{ and } \operatorname{sgn}(W_j) = \pm\}$$

$$\tau = \min\left\{t : \widehat{\mathsf{FDP}}(t) = \frac{1+|\mathcal{S}^{-}(t)|}{1 \vee |\mathcal{S}^{+}(t)|} \leq q\right\}$$

$$\hat{\mathcal{S}} = \{W_j \geq \tau\}$$



Theorem (Barber and C. ('15))

- *Knockoff*

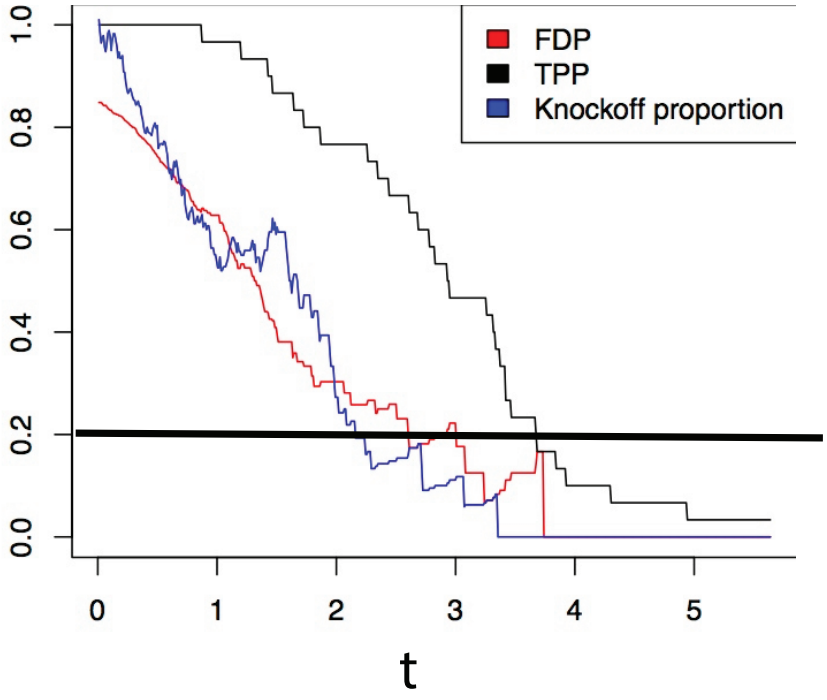
$$\mathbb{E}\left[\frac{\#\text{ false positives}}{\#\text{ selections} + q^{-1}}\right] \leq q$$

FDR control

$$\mathcal{S}^{\pm}(t) = \{j : |W_j| \geq t \text{ and } \operatorname{sgn}(W_j) = \pm\}$$

$$\tau = \min\left\{t : \widehat{\mathrm{FDP}}(t) = \frac{1+|\mathcal{S}^{-}(t)|}{1 \vee |\mathcal{S}^{+}(t)|} \leq q\right\}$$

$$\hat{\mathcal{S}} = \{W_j \geq \tau\}$$



Theorem (Barber and C. ('15))

- *Knockoff*

$$\mathbb{E}\left[\frac{\#\text{ false positives}}{\#\text{ selections} + q^{-1}}\right] \leq q$$

- *Knockoff+*

$$\mathbb{E}\left[\frac{\#\text{ false positives}}{\#\text{ selections}}\right] \leq q$$

Analysis of Genetic Data

Joint with Fan, Janson & Lv and with Sabatti & Sesia

Data

WTCCC data

- $n = 4,913$ subjects ($1,917$ CD patients and $2,996$ healthy controls)
- $p = 377,749$ SNPs
- Previously analyzed in WTCCC (2007)

NFBC data

- $4,700$ HDL subjects and $4,682$ LDL subjects
- $p = 328,934$ SNPs
- Previously analyzed in Sabatti et al. (2009)

Peek at the results from Wald Lecture II

Knockoffs with nominal FDR level of 10%

Peek at the results from Wald Lecture II

Knockoffs with nominal FDR level of 10%

- Power is much higher:

Dataset	Number of discoveries	
	Original study	Knockoffs (average)
CD	9	22.8
HDL	5	8
LDL	6	9.8

Peek at the results from Wald Lecture II

Knockoffs with nominal FDR level of 10%

- Power is much higher:

Dataset	Number of discoveries	
	Original study	Knockoffs (average)
CD	9	22.8
HDL	5	8
LDL	6	9.8

- Quite a few of the discoveries made by knockoffs were confirmed by larger GWAS (Franke et al., '10, Willer et al., '13)

Peek at the results from Wald Lecture II

Knockoffs with nominal FDR level of 10%

- Power is much higher:

Dataset	Number of discoveries	
	Original study	Knockoffs (average)
CD	9	22.8
HDL	5	8
LDL	6	9.8

- Quite a few of the discoveries made by knockoffs were confirmed by larger GWAS (Franke et al., '10, Willer et al., '13)

- Knockoffs made a number of new discoveries

Peek at the results from Wald Lecture II

Knockoffs with nominal FDR level of 10%

- Power is much higher:

Dataset	Number of discoveries	
	Original study	Knockoffs (average)
CD	9	22.8
HDL	5	8
LDL	6	9.8

- Quite a few of the discoveries made by knockoffs were confirmed by larger GWAS (Franke et al., '10, Willer et al., '13)

- Knockoffs made a number of new discoveries
  - Expect some (roughly 10%) of these to be false discoveries

Peek at the results from Wald Lecture II

Knockoffs with nominal FDR level of 10%

- Power is much higher:

Dataset	Number of discoveries	
	Original study	Knockoffs (average)
CD	9	22.8
HDL	5	8
LDL	6	9.8

- Quite a few of the discoveries made by knockoffs were confirmed by larger GWAS (Franke et al., '10, Willer et al., '13)
- Knockoffs made a number of new discoveries
  - Expect some (roughly 10%) of these to be false discoveries
  - It is likely that many of these correspond to true discoveries

Peek at the results from Wald Lecture II

Knockoffs with nominal FDR level of 10%

- Power is much higher:

Dataset	Number of discoveries	
	Original study	Knockoffs (average)
CD	9	22.8
HDL	5	8
LDL	6	9.8

- Quite a few of the discoveries made by knockoffs were confirmed by larger GWAS (Franke et al., '10, Willer et al., '13)
- Knockoffs made a number of new discoveries
  - Expect some (roughly 10%) of these to be false discoveries
  - It is likely that many of these correspond to true discoveries
  - Evidence from independent studies about adjacent genes shows some of the top unconfirmed hits to be promising candidates

Selection frequency	SNP (cluster size)	Chr.	Position range (Mb)	Franke et al. '10	WTCCC '07
100%	rs11209026 (2)	1	67.31–67.42	yes	yes
99%	rs6431654 (20)	2	233.94–234.11	yes	yes
98%	rs6688532 (33)	1	169.4–169.65		yes
97%	rs17234657 (1)	5	40.44–40.44	yes	yes
95%	rs11805303 (16)	1	67.31–67.46	yes	yes
91%	rs7095491 (18)	10	101.26–101.32	yes	yes
91%	rs3135503 (16)	16	49.28–49.36	yes	yes
81%	rs7768538 (1145)	6	25.19–32.91	yes	yes
80%	rs6601764 (1)	10	3.85–3.85		yes
75%	rs7655059 (5)	4	89.5–89.53		
73%	rs6500315 (4)	16	49.03–49.07	yes	yes
72%	rs2738758 (5)	20	61.71–61.82	yes	
70%	rs7726744 (46)	5	40.35–40.71	yes	yes
68%	rs11627513 (7)	14	96.61–96.63		
66%	rs4246045 (46)	5	150.07–150.41	yes	yes
62%	rs9783122 (234)	10	106.43–107.61		
61%	rs6825958 (3)	4	55.73–55.77		

Table: SNP clusters found to be important for CD over 100 repetitions of knockoffs.

Selection frequency	SNP (cluster size)	Chr.	Position range (Mb)	Confirmed in Willer et al. '13	Found in Sabatti et al. '09
100%	rs1532085 (4)	15	58.68–58.7	yes	yes
100%	rs7499892 (1)	16	57.01–57.01	yes	yes
100%	rs1800961 (1)	20	43.04–43.04	yes	
99%	rs1532624 (2)	16	56.99–57.01	yes	yes
95%	rs255049 (142)	16	66.41–69.41	yes	yes

Table: SNP clusters found to be important for HDL over 100 repetitions of knockoffs.

Selection frequency	SNP (cluster size)	Chr.	Position range (Mb)	Confirmed in Willer et al. '13	Found in Sabatti et al. '09
99%	rs4844614 (34)	1	207.3–207.88		yes
97%	rs646776 (5)	1	109.8–109.82	yes	yes
97%	rs2228671 (2)	19	11.2–11.21	yes	yes
94%	rs157580 (4)	19	45.4–45.41	yes	yes
92%	rs557435 (21)	1	55.52–55.72	yes	
80%	rs10198175 (1)	2	21.13–21.13	yes	yes
76%	rs10953541 (58)	7	106.48–107.3		
62%	rs6575501 (1)	14	95.64–95.64		

Table: SNP clusters found to be important for LDL over 100 repetitions of knockoffs.

Summary: very precise inference machine

You design the statistics, knockoffs takes care of inference

Addresses the reproducibility issue (at least partially)

Summary: very precise inference machine

You design the statistics, knockoffs takes care of inference

Addresses the reproducibility issue (at least partially)

Summary: very precise inference machine

You design the statistics, knockoffs takes care of inference

Addresses the reproducibility issue (at least partially)

Summary: very precise inference machine

You design the statistics, knockoffs takes care of inference

Addresses the reproducibility issue (at least partially)

Summary: very precise inference machine

You design the statistics, knockoffs takes care of inference

Addresses the reproducibility issue (at least partially)

Summary: very precise inference machine

You design the statistics, knockoffs takes care of inference

Addresses the reproducibility issue (at least partially)

Thank You!
Let's Make Knockoffs!

A better reasoning

Conditional randomization test: C. , Fan, Janson, Lv ('16)

A better reasoning

Conditional randomization test: C. , Fan, Janson, Lv ('16)

Want to understand the effect of X_2 while controlling for X_1

A better reasoning

Conditional randomization test: C. , Fan, Janson, Lv ('16)

Want to understand the effect of X_2 while controlling for X_1

- Assume X is random and resample many copies X_2' from $X_2 \,|\, X_1$ ($\perp\!\!\!\perp Y$)

A better reasoning

Conditional randomization test: C. , Fan, Janson, Lv ('16)

Want to understand the effect of X_2 while controlling for X_1

- Assume X is random and resample many copies X_2' from $X_2 \,|\, X_1$ ($\perp\!\!\!\perp Y$)
- Compare obs. value of $t(X_1, X_2, Y)$ with (resampling) distribution of $t(X_1, X_2', Y)$

A better reasoning

Conditional randomization test: C. , Fan, Janson, Lv ('16)

Want to understand the effect of X_2 while controlling for X_1

- Assume X is random and resample many copies X_2' from $X_2 \,|\, X_1$ ($\perp\!\!\!\perp Y$)
- Compare obs. value of $t(X_1, X_2, Y)$ with (resampling) distribution of $t(X_1, X_2', Y)$

Valid procedure $\longrightarrow$ valid p-values

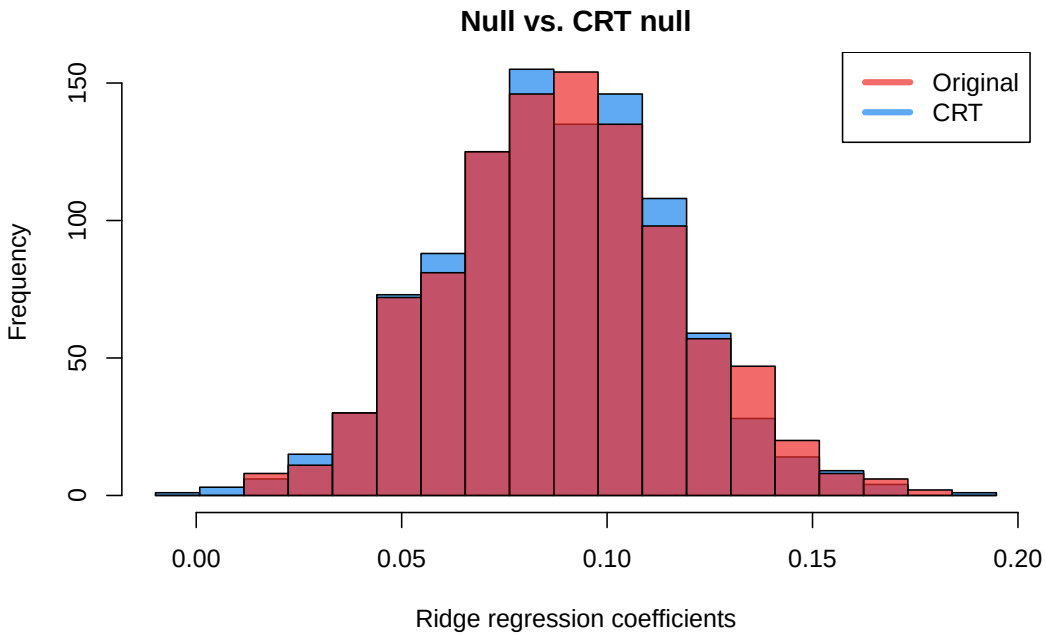
A better reasoning

Conditional randomization test: C. , Fan, Janson, Lv ('16)

Want to understand the effect of X_2 while controlling for X_1

- Assume X is random and resample many copies X_2' from $X_2 \,|\, X_1$ ($\perp\!\!\!\perp Y$)
- Compare obs. value of $t(X_1, X_2, Y)$ with (resampling) distribution of $t(X_1, X_2', Y)$

Valid procedure $\longrightarrow$ valid p-values



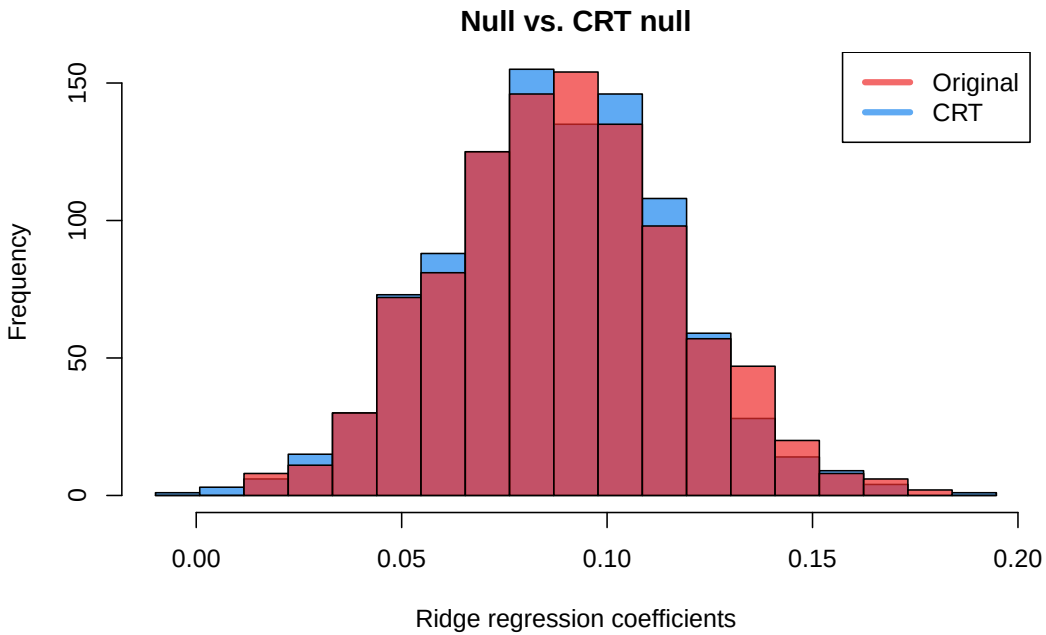
A better reasoning

Conditional randomization test: C. , Fan, Janson, Lv ('16)

Want to understand the effect of X_2 while controlling for X_1

- Assume X is random and resample many copies X_2' from $X_2 \,|\, X_1$ ($\perp\!\!\!\perp Y$)
- Compare obs. value of $t(X_1, X_2, Y)$ with (resampling) distribution of $t(X_1, X_2', Y)$

Valid procedure $\longrightarrow$ valid p-values



Dummies obey

$$\mathbb{E}\, X_2' X_1 = \mathbb{E}\, X_2 X_1$$