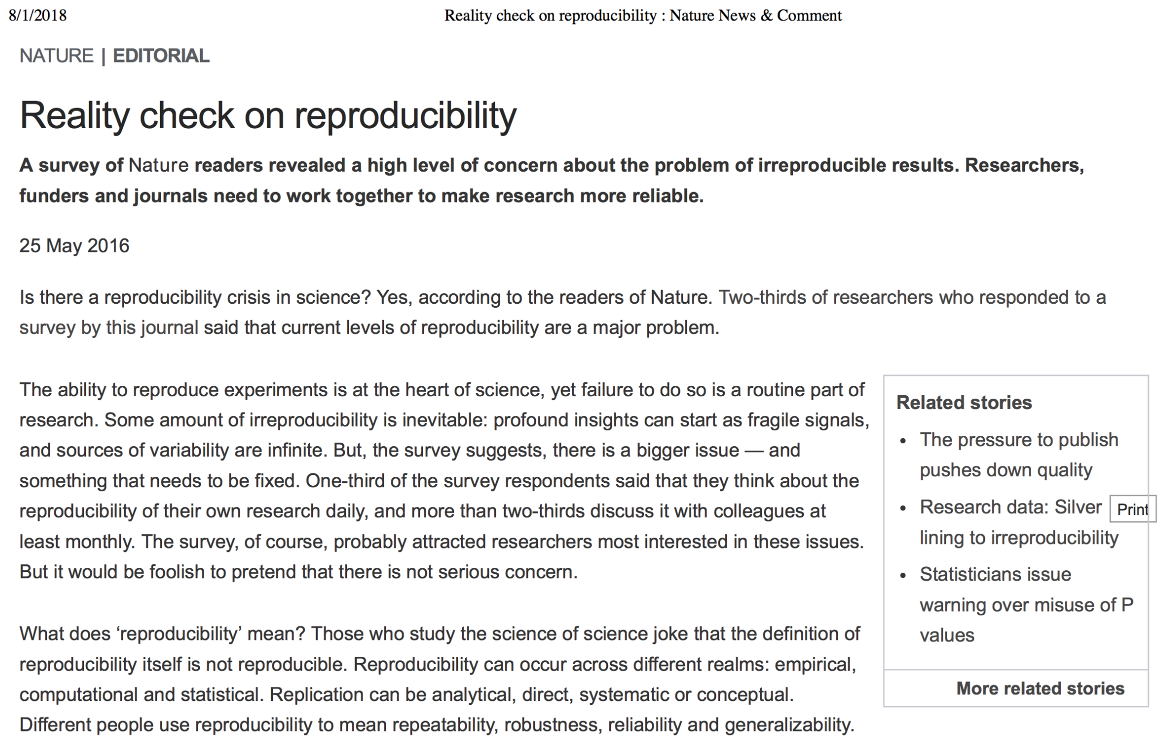
A selective survey of selective inference

J. Taylor (Stanford)

ICM 2018

Loading [MathJax]/jax/output/HTML-CSS/jax.js

Replicability crisis in science



8/1/2018 Reality check on reproducibility : Nature News & Comment

NATURE | **EDITORIAL**

Reality check on reproducibility

A survey of Nature **readers revealed a high level of concern about the problem of irreproducible results. Researchers, funders and journals need to work together to make research more reliable.**

25 May 2016

Is there a reproducibility crisis in science? Yes, according to the readers of Nature. Two-thirds of researchers who responded to a survey by this journal said that current levels of reproducibility are a major problem.

The ability to reproduce experiments is at the heart of science, yet failure to do so is a routine part of research. Some amount of irreproducibility is inevitable: profound insights can start as fragile signals, and sources of variability are infinite. But, the survey suggests, there is a bigger issue — and something that needs to be fixed. One-third of the survey respondents said that they think about the reproducibility of their own research daily, and more than two-thirds discuss it with colleagues at least monthly. The survey, of course, probably attracted researchers most interested in these issues. But it would be foolish to pretend that there is not serious concern.

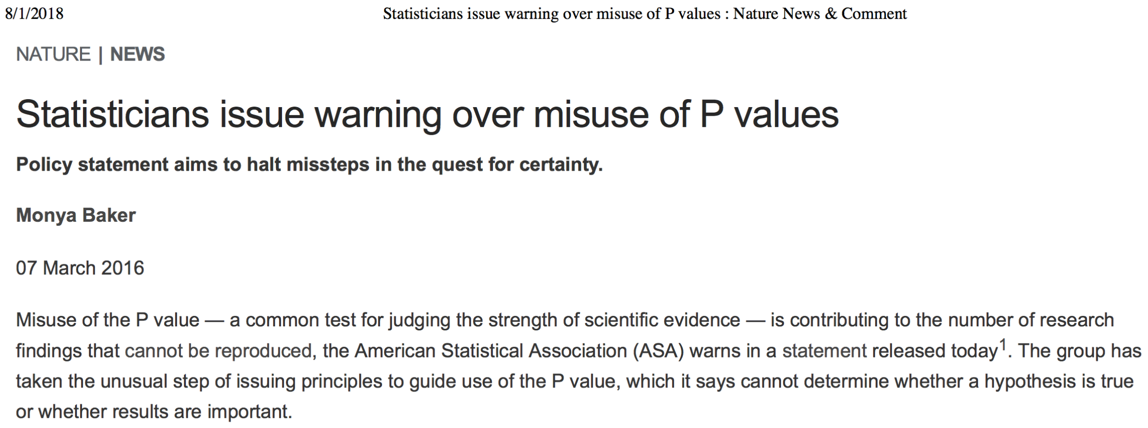
What does ‘reproducibility’ mean? Those who study the science of science joke that the definition of reproducibility itself is not reproducible. Reproducibility can occur across different realms: empirical, computational and statistical. Replication can be analytical, direct, systematic or conceptual. Different people use reproducibility to mean repeatability, robustness, reliability and generalizability.

Related stories

- The pressure to publish pushes down quality
- Research data: Silver lining to irreproducibility
- Statisticians issue warning over misuse of P values

Print

More related stories



8/1/2018 Statisticians issue warning over misuse of P values : Nature News & Comment

NATURE | **NEWS**

Statisticians issue warning over misuse of P values

Policy statement aims to halt missteps in the quest for certainty.

Monya Baker

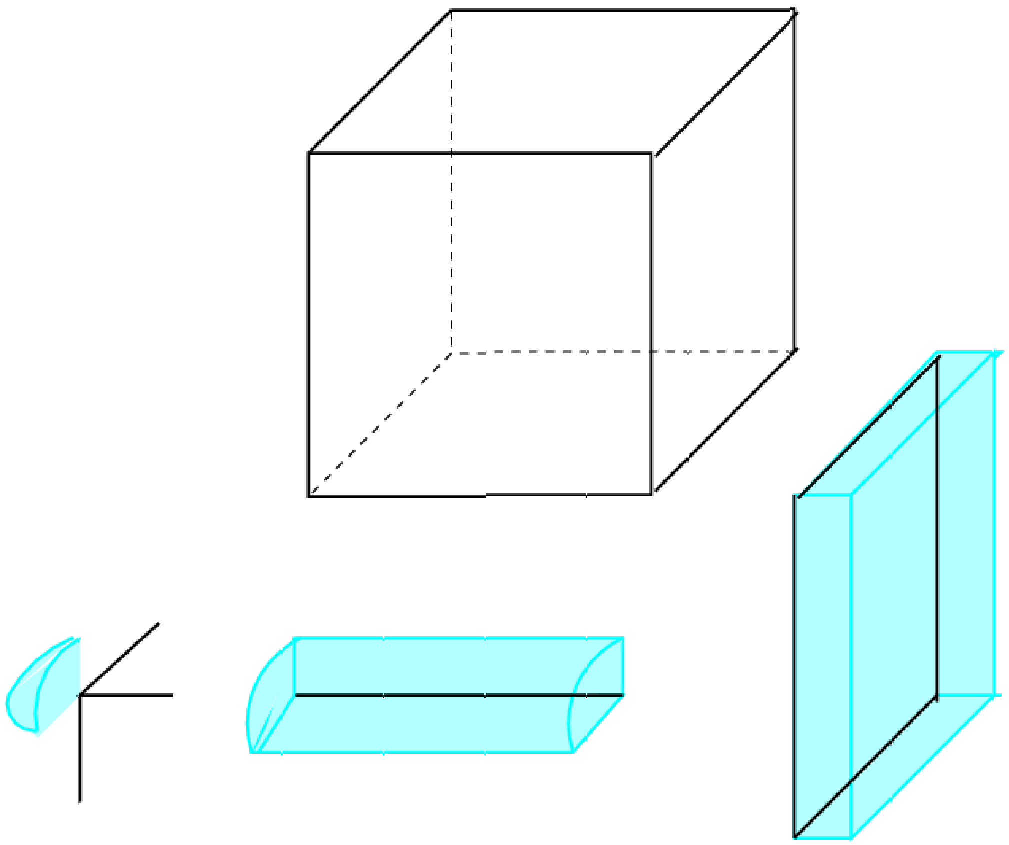
07 March 2016

Misuse of the P value — a common test for judging the strength of scientific evidence — is contributing to the number of research findings that cannot be reproduced, the American Statistical Association (ASA) warns in a statement released today[1]. The group has taken the unusual step of issuing principles to guide use of the P value, which it says cannot determine whether a hypothesis is true or whether results are important.

- **Caveat:** embarassingly untechnical talk
- Begin with focus on statisticians’ efforts addressing replicability in science.

Common mathematical theme

Normal cycle $N(K)$



$$N(K) = \{(u,\beta) : u \in K, \beta \in N_u K\}$$

Replicability crisis in science

Scientists collect data first and ask questions later. (Candes)

The idea of a scientist, struck, as if by lightning with a question, is far from the truth. (Tukey)

Replicability crisis in science

Exploratory data analysis

- Tukey: scientists have always used data to form new questions – this is classical!

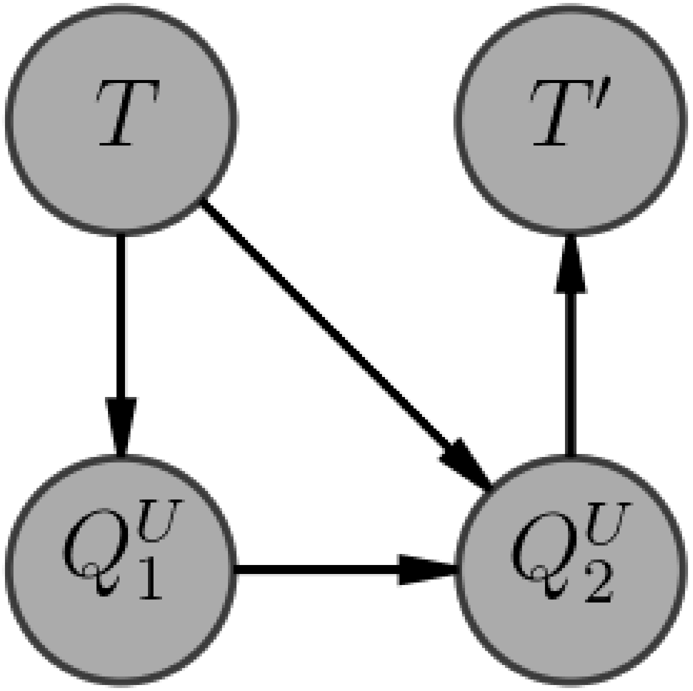
Confirmatory data analysis

- The standards of science require some confirmation (often statistical).

Conflict identified by Candes (and certainly others)

- Misleading to (naively) use the same data for exploration and confirmation.

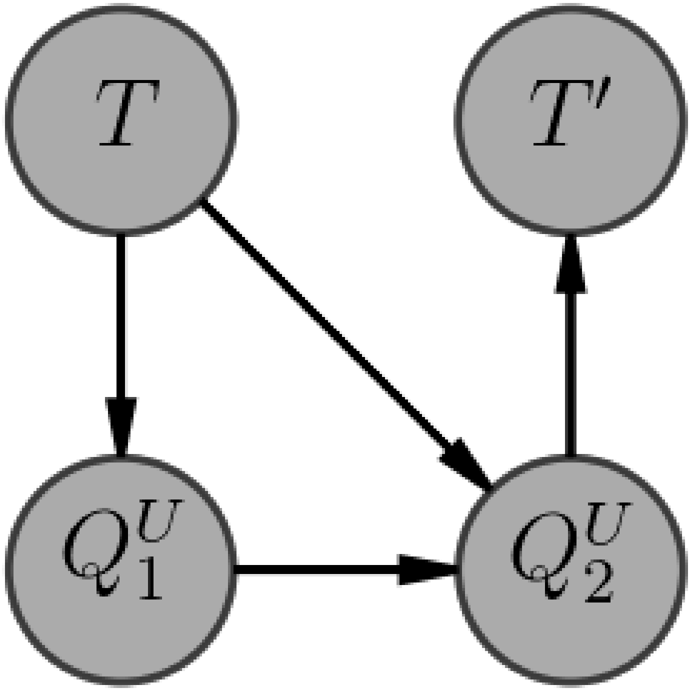
Modern science



A (typical?) data scientist U's workflow…

- Query Q_1^U might be choice of a tuning parameter
- Query Q_2^U might be a feature selection step
- Data T' might be validation data

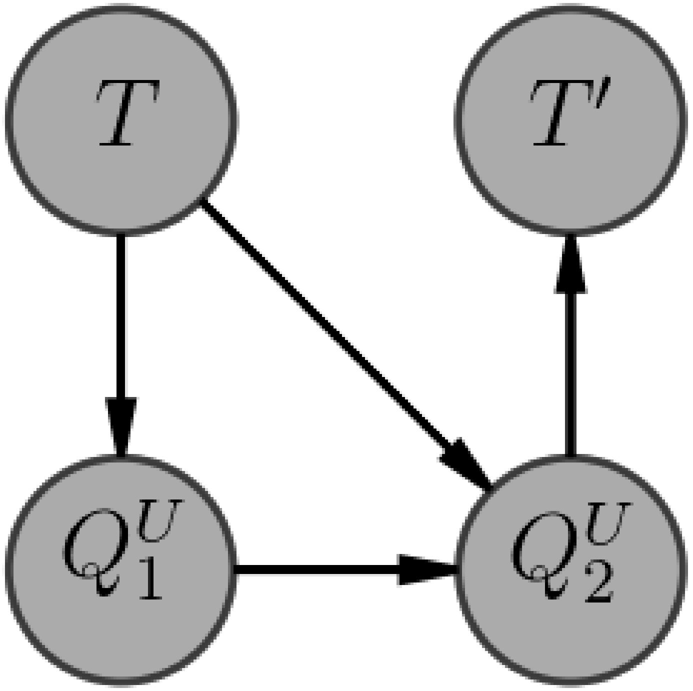
Modern science



A (typical?) data scientist U's workflow…

- Could have more data nodes…
- Could have more queries…

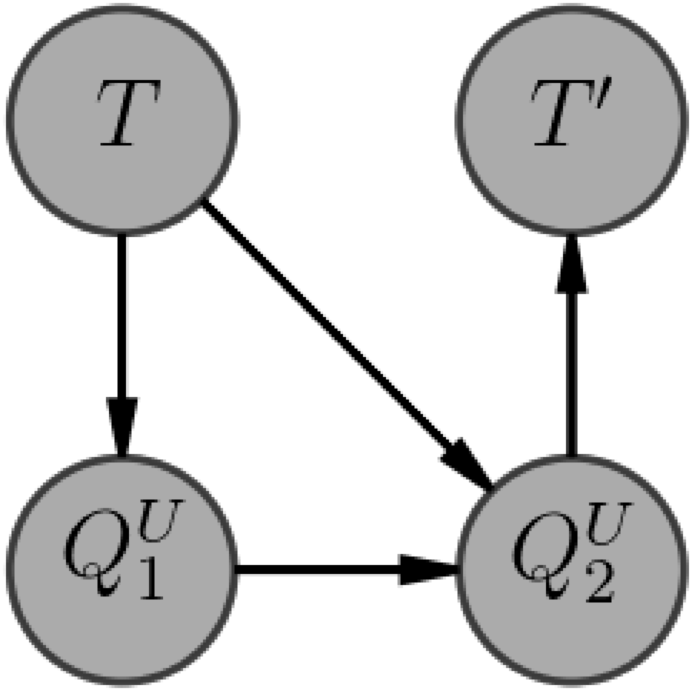
Modern science



Example: predicting drug resistance

- T denotes mutation patterns of HIV viruses and in vitro response to 3TC
- Q_1^U asks for important main effects, Q_2^U asks for important interaction; T' is empty – no new data.

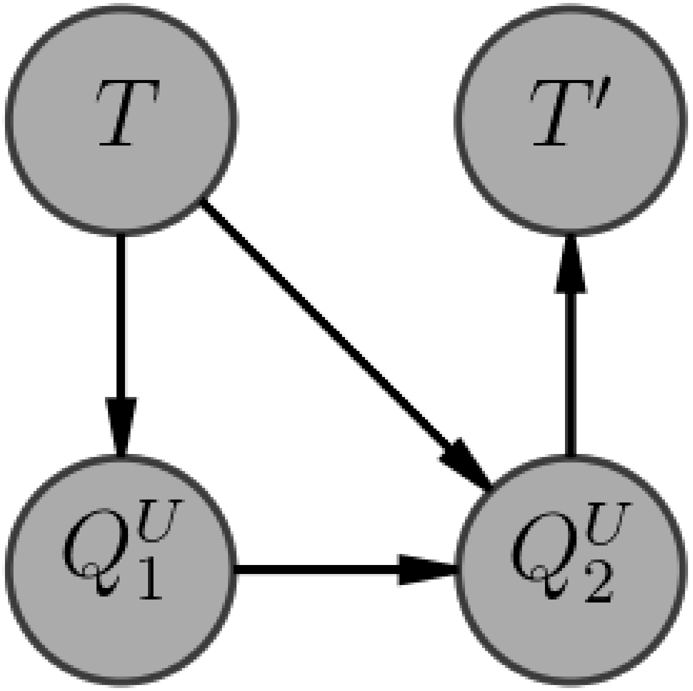
Modern science



Simple example: Drop the losers (Sampson and Sill)

- T denotes K different treatments in a clinical trial.
- Q_1^U asks which treatment (apparently) works best.

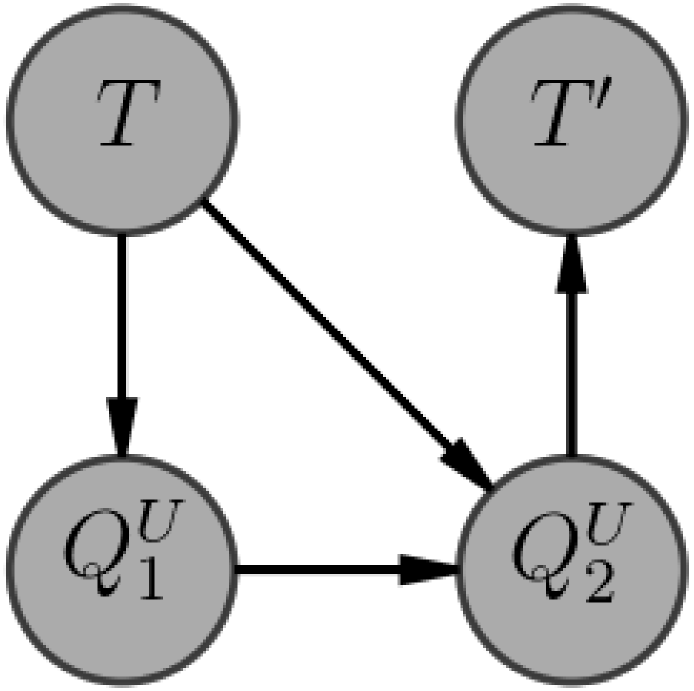
Modern science



Simple example: Drop the losers (Sampson and Sill)

- Q_2^U asks which is second best (variant of Sampson and Sill).
- Data T' is confirmatory sample: **gold standard** reports an estimate of best treatment effect based on T' alone.

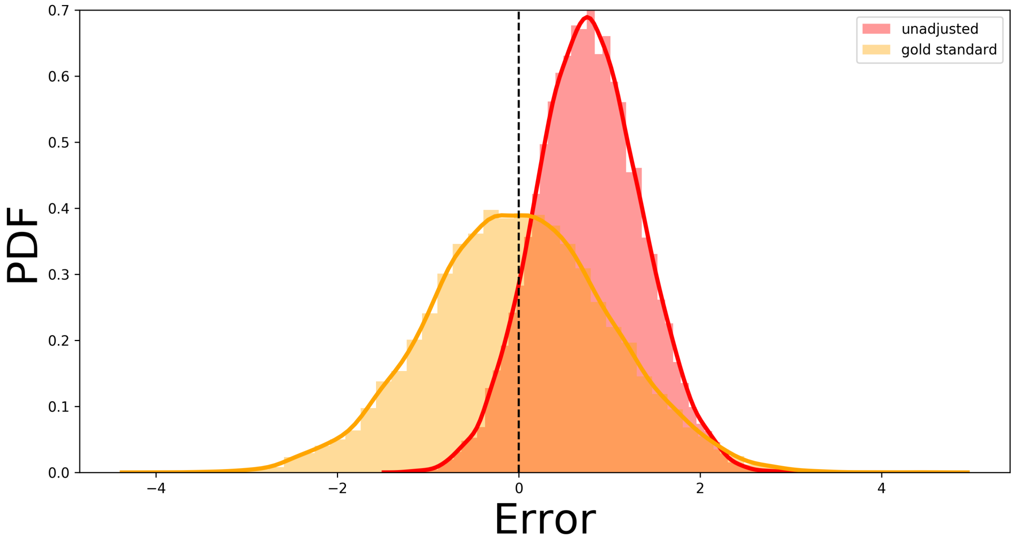
Modern science



Simple example: Drop the losers (Sampson and Sill)

- Wasteful to only use T for selection?
- Can we reuse earlier data?

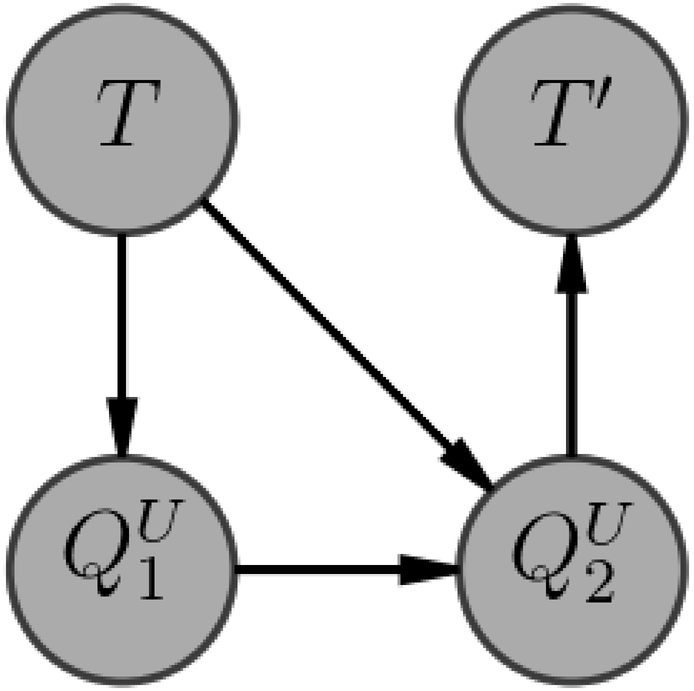
Modern science



Conflict between confirmatory and exploratory

- Unadjusted estimator $(T_k + T'_k)/2$ is biased, gold standard estimator T'_k is unbiased but more variable.
- Simple manifestation of Candes' observation, *researcher degrees of freedom*.
- Difference can easily be bigger with larger K, different sample sizes, etc.

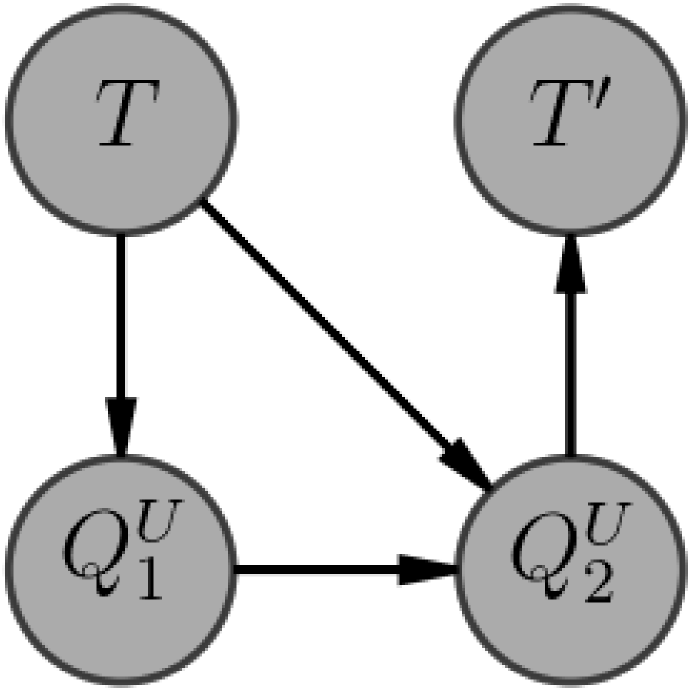
Modern science



Reproducibility and replicability

- Great efforts have been made to make U's results **reproducible**.
- For **replicability** we need to (statistically) understand this collection of random variables.

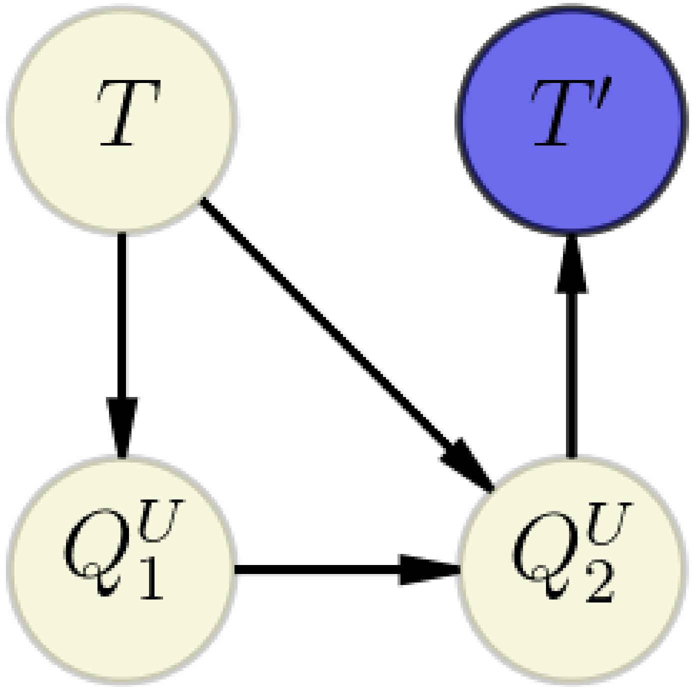
Modern science



Selective inference

- Valid inference in the presence of *selection effects* determined by Q_1^U, Q_2^U.

Modern science



Classical inference

- U's exploratory interaction with the data is limited to T.
- Earlier data is “wasted”, confirmatory focus on T'.
- No selection effect.

Modern science

Challenge for selective inference

- A scientist's question does not always translate easily into statistical objects, a necessary step to model U's workflow.

Statistical objects

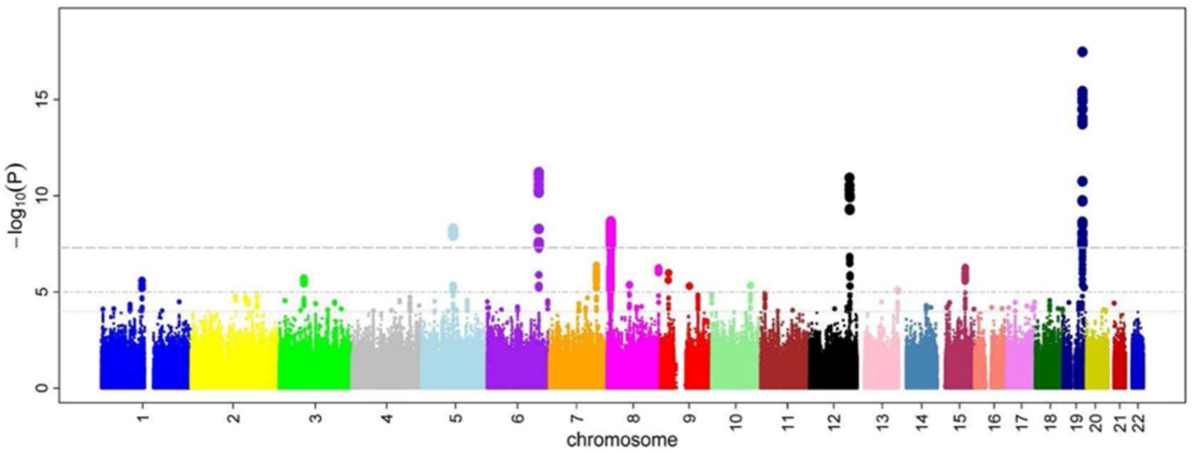
- Statistical model: e.g. for gold standard $T'|T \sim F \in \mathcal{M}$
- Parameter: $\theta : \mathcal{M} \to \mathbb{R}$, e.g. treament effect for "best" treatment.
- Language of statistics: for parameter θ we have

1. point estimators
2. confidence intervals
3. posterior distributions

Scientists, when struck by anything, are not struck with null hypotheses…

Modern science

A prototype: simultaneous (large scale) inference



(Wikipedia)

- Measure disease status D and a genomic signature for each of N markers, $(M_i)_{1 \leq i \leq N}$.

- Natural choice of parameters: θ_i be the *association* of marker M_i with disease D.

- Data: $T' = (D, (M_i)_{1 \leq i \leq N})$.

Modern science

Large scale inference and multiple comparisons

- Within each marker, estimate association $\hat{\theta}_i$ and consider testing no association between D and marker M_i

$$H_i : \theta_i(F) = 0, \text{ i.e. } F \ni \{G \in \mathcal{M} : \theta_i(G) = 0\}?$$

False Discovery Rate (FDR)

- Benjamini & Hochberg (1995) **hugely influential** in multiple comparisons over past 20 years, particularly in *large scale inference*.

Modern science

Family Wise Error Rate (FWER)

- Tests based on maximum association

$$\max_{1 \leq i \leq N} \left| \frac{\hat{\theta}_i - \theta_i}{SD(\hat{\theta}_i)} \right| = \max_{1 \leq i \leq N} |Z_i|$$

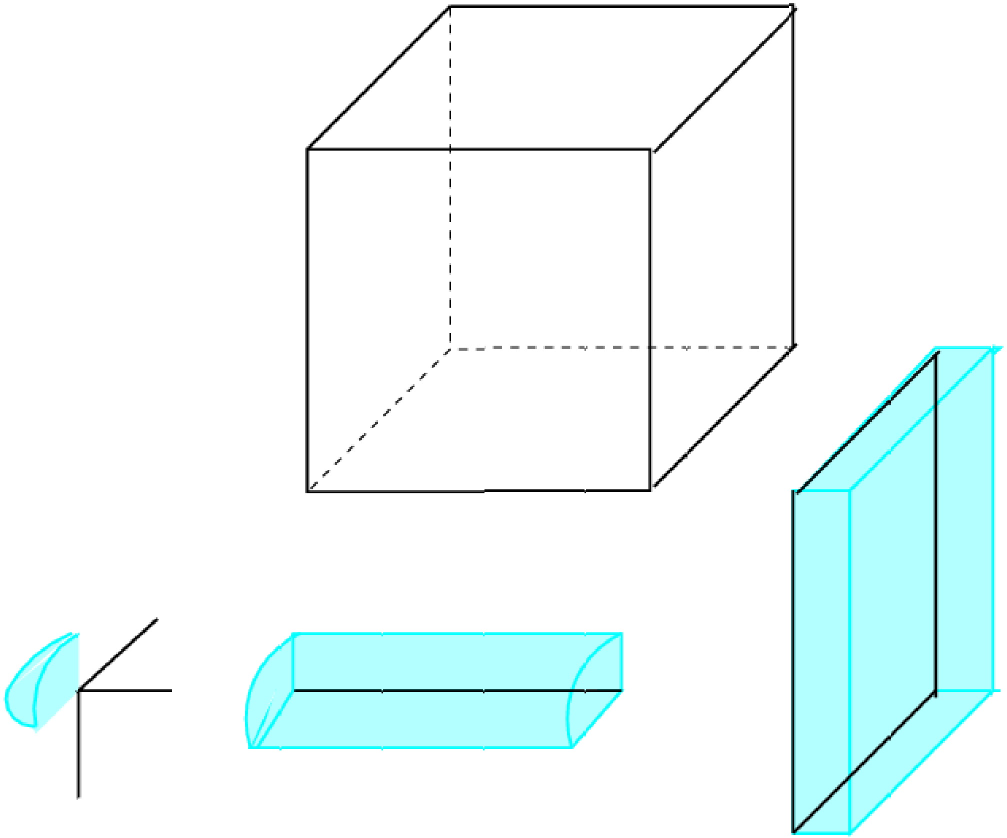
Bonferroni and volume of tubes

- Embedding sampling of genome (or other measurements) into some continuous space

$$P\left(\sup_{x \in M} |Z_x| \geq u \right) \approx \sum_j \mathcal{L}_j(M) \rho_j(u)$$

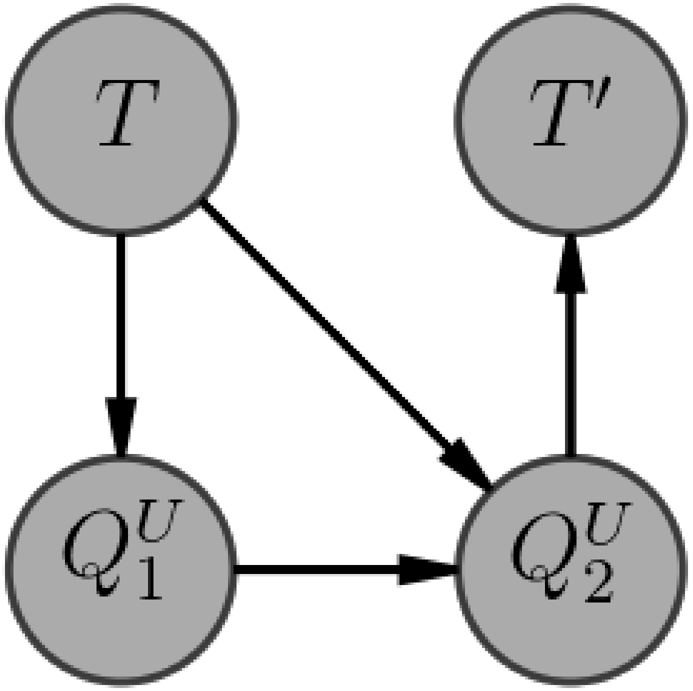
Modern science

Volume of tubes



$$\lambda\,(\mathrm{Tube}(M,r)) = \int_{N(M)} J(u,\beta)\mathcal{H}(du\ d\beta)$$

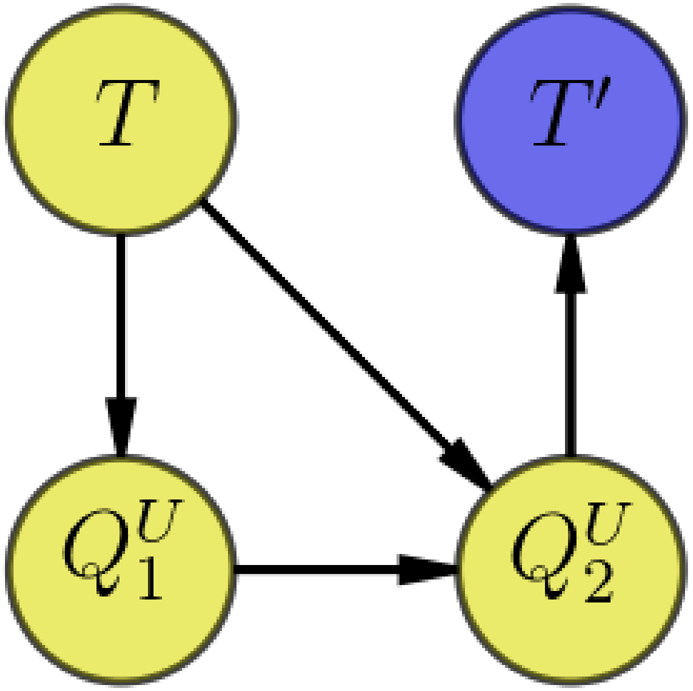
Modern science



Pause: does large scale inference address U's workflow?

- Arguments for: Bonferroni can be used for a confidence interval in *drop the losers*.
- Arguments against: questions are determined entirely by structure of T'.

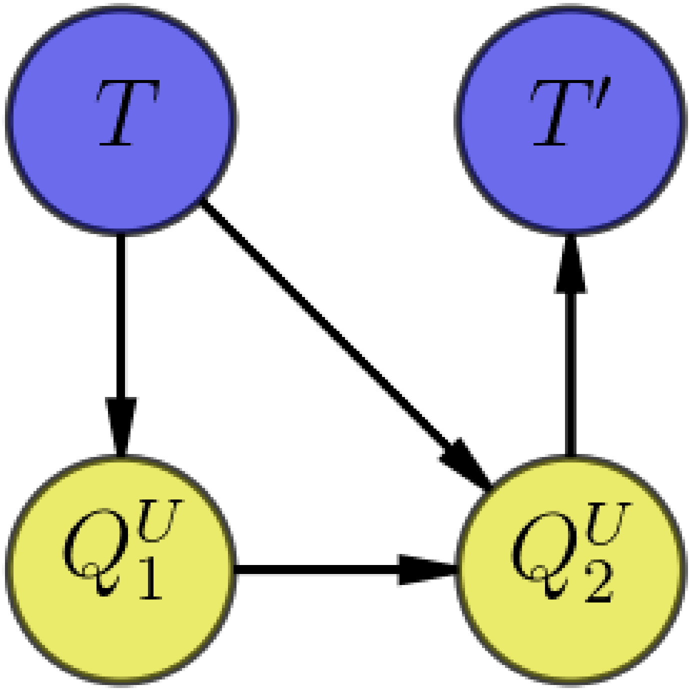
Conditional inference



Classical inference

- Required to collect data T'.
- Throwing away T is *conditioning* on T.

Conditional inference



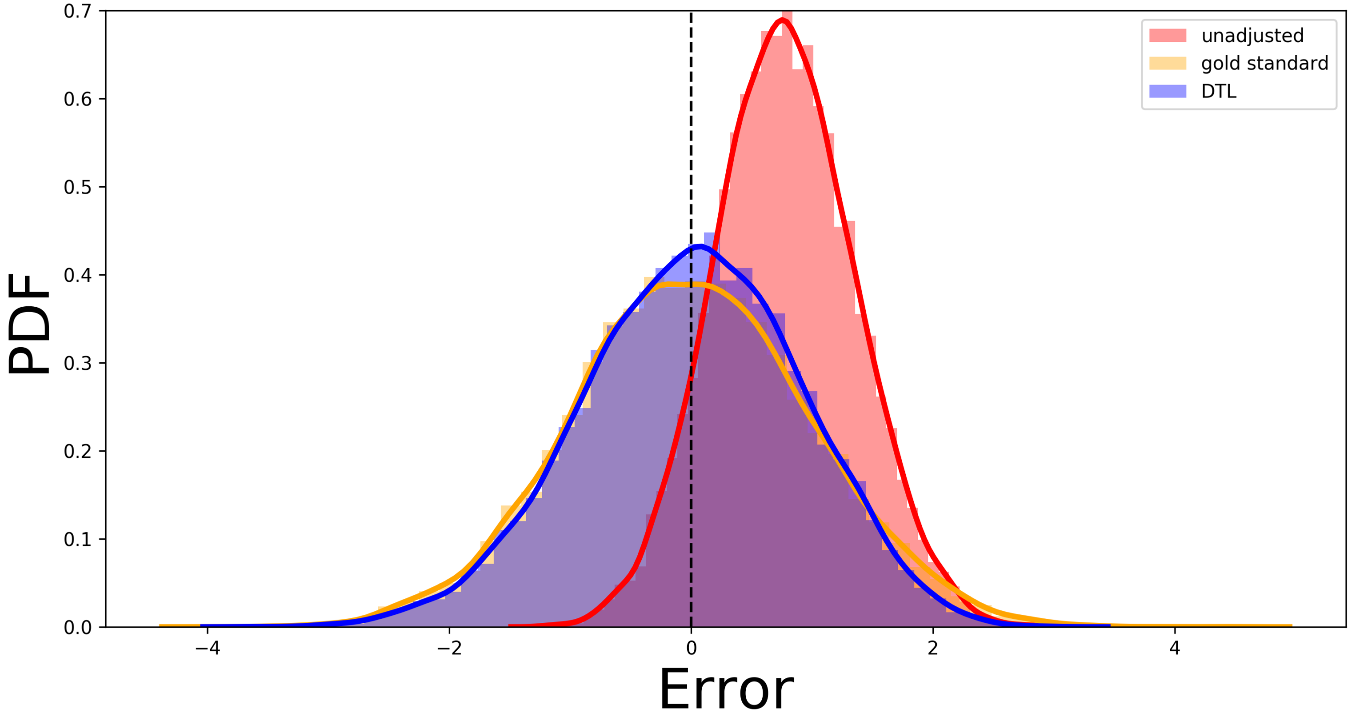
Drop the losers

- Instead of throwing out all of T, condition only on which treatment is apparently best:
- Rao-Blackwell (Cohen + Sacrowicz)

$$\hat{\theta}_{\hat{K}} = E[T_k' | (T' + T)_k, (T_j)_{j \neq k}, \hat{K} = k]$$

Conditional inference

The (classical) scientific method is inadmissible!

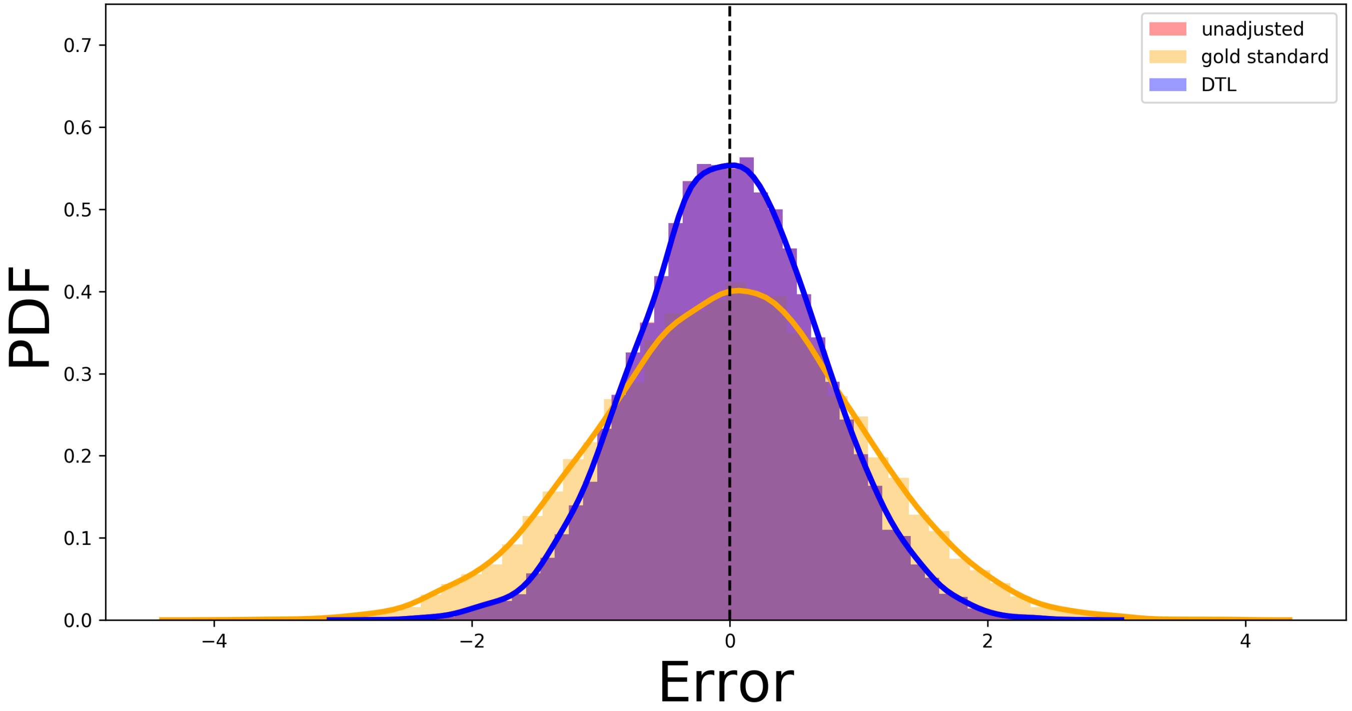


- Tests and confidence intervals also available.

- Similar technique can be used when looking at best 2 treatments, rather than just single best treatment.

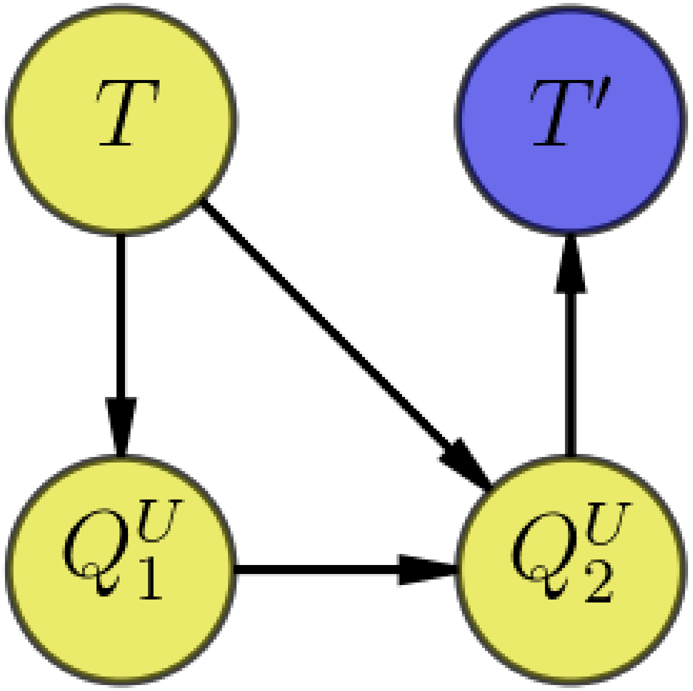
Conditional inference

The (classical) scientific method is inadmissible!



- Tests and confidence intervals also available.
- Similar technique can be used when looking at best 2 treatments, rather than just single best treatment.

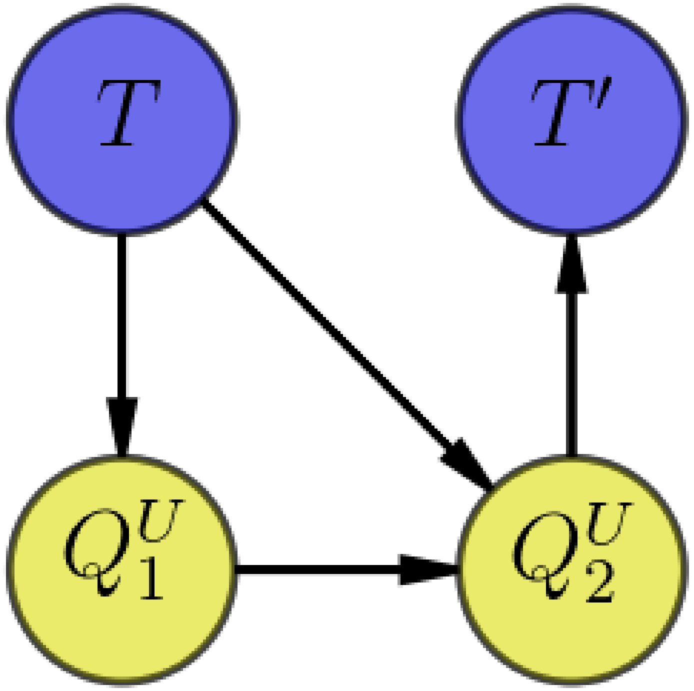
Conditional inference



Classical scientific method

- Specifically allows U's intervention – even model $\mathcal{M}$ is chosen **after observing** (Q_1^U, Q_2^U)**!**
- U specifies model $\mathcal{M}$ for the law $T'|T$.

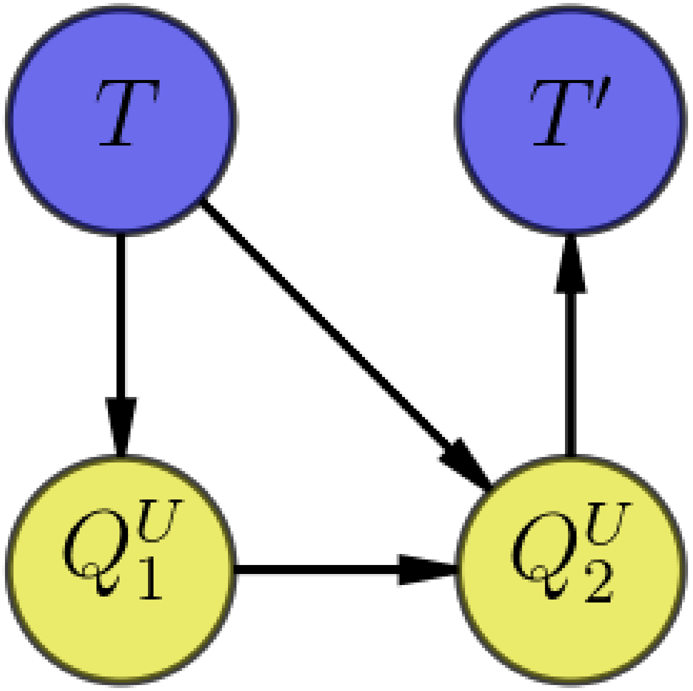
Conditional inference



General approach

- U specifies model $\mathcal{M}$ for the law (T', T).
- *Do statistics* on all data (T, T').

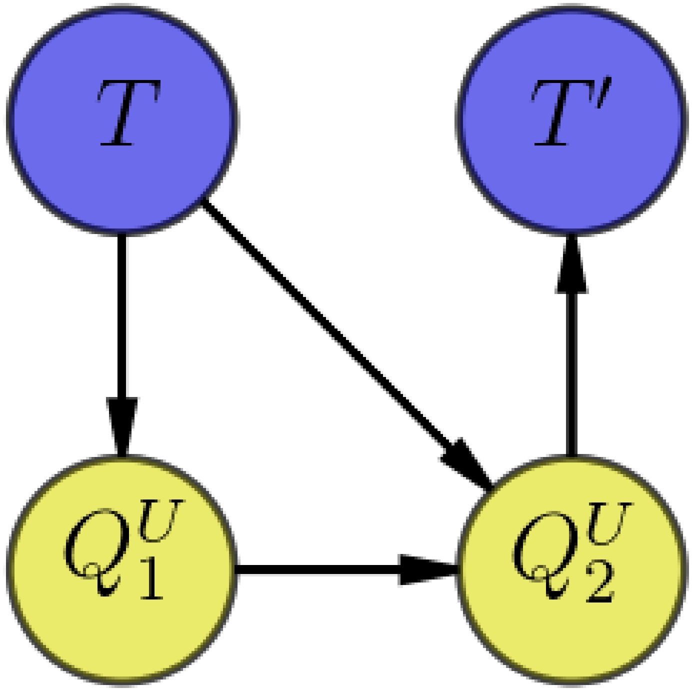
Conditional inference



General approach

- Is this improvement limited to *drop the losers*? **No.**
- Do we need confirmatory sample T'? **No. We can even have $T = T'$.**
- Can we allow arbitrary queries? **Probably not.**

Conditional inference



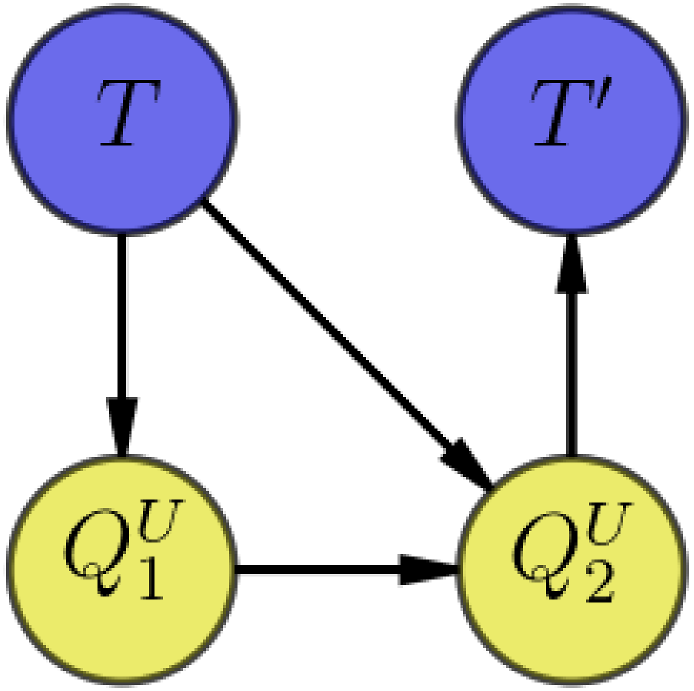
General approach

- Inference is carried out in **selective model**

$$\mathcal{M}^* = \left\{ F^* : \frac{dF^*}{dF} \propto \zeta^* \right\}$$

- The function ζ^* can be "read off" the dependency graph knowing the observed values of Q_1^U and Q_2^U.

Conditional inference



General approach

- Theoretical crux becomes transferring what we know about $\mathcal{M}$ to $\mathcal{M}^*$ (i.e. consistency, CLT, etc.)
- Computational crux becomes describing ζ^* in silico.

Conditional inference

Randomized convex programs

- In drop the losers, for Q_1^U we solve

$$\text{maximize}_{\alpha\in\mathcal{S}_\mathcal{K}}\langle\alpha, T\rangle$$

with

$$\mathcal{S}_K = \left\{ \alpha \in \mathbb{R}^K : \mathbb{R}^K : \alpha_i \geq 0 \sum_{i=1}^{K} \alpha_i = 1 \right\}.$$

- With $\omega = T' - T$, this is (essentially) equivalent to

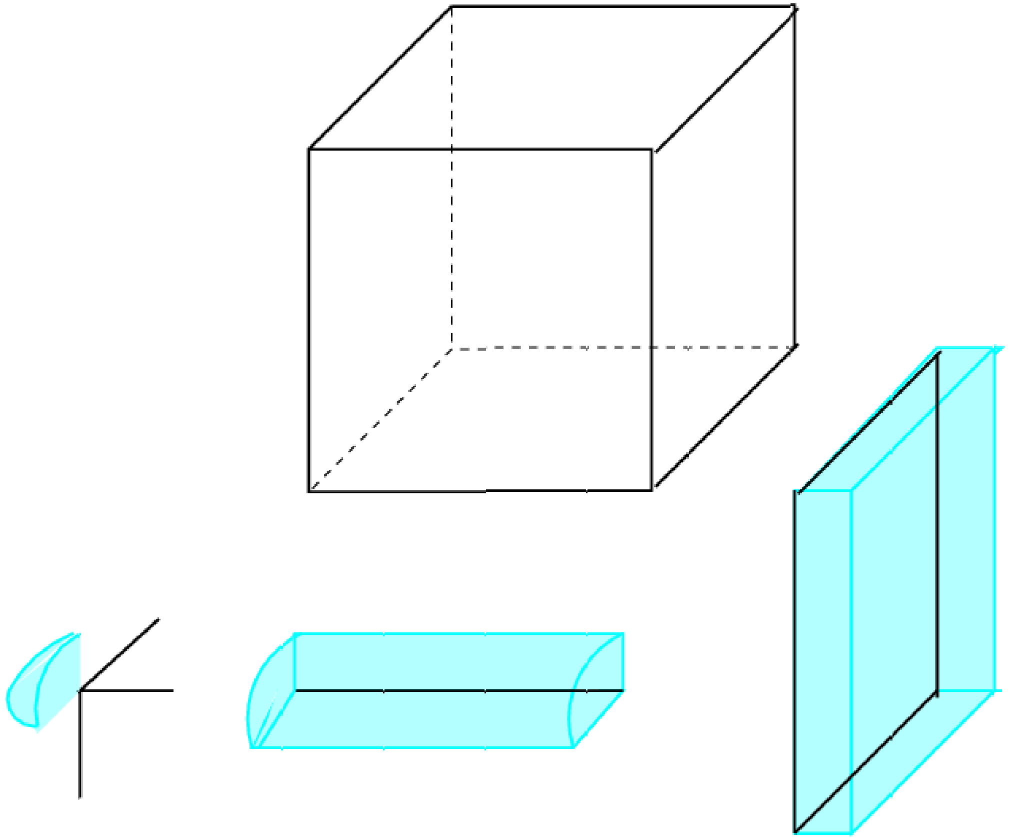
$$\text{maximize}_{\alpha\in\mathcal{S}_\mathcal{K}}\langle\alpha, T + T'\rangle - \langle\alpha, \omega\rangle$$

- A perturbed version of

$$\text{maximize}_{\alpha\in\mathcal{S}_\mathcal{K}}\langle\alpha, T + T'\rangle$$

Conditional inference

Randomized convex programs



$$\zeta^*(T+T') = \int_{N_k \mathcal{S}_K} g_\omega(T+T'-\eta) \cdot J(T+T',\eta)\; d\eta$$

Conditional inference

Randomized convex programs

$$\text{minimize}_\beta \ell(\beta; T) + \mathcal{P}(\beta) - \omega^T \beta, \qquad \omega \sim G$$

Structure inducing penalties

$$\mathcal{P}(\beta) = \sup_{u \in K} \langle u, \beta \rangle, \qquad \text{e.g. } \mathcal{P}(\beta) = \lambda \|\beta\|_1$$

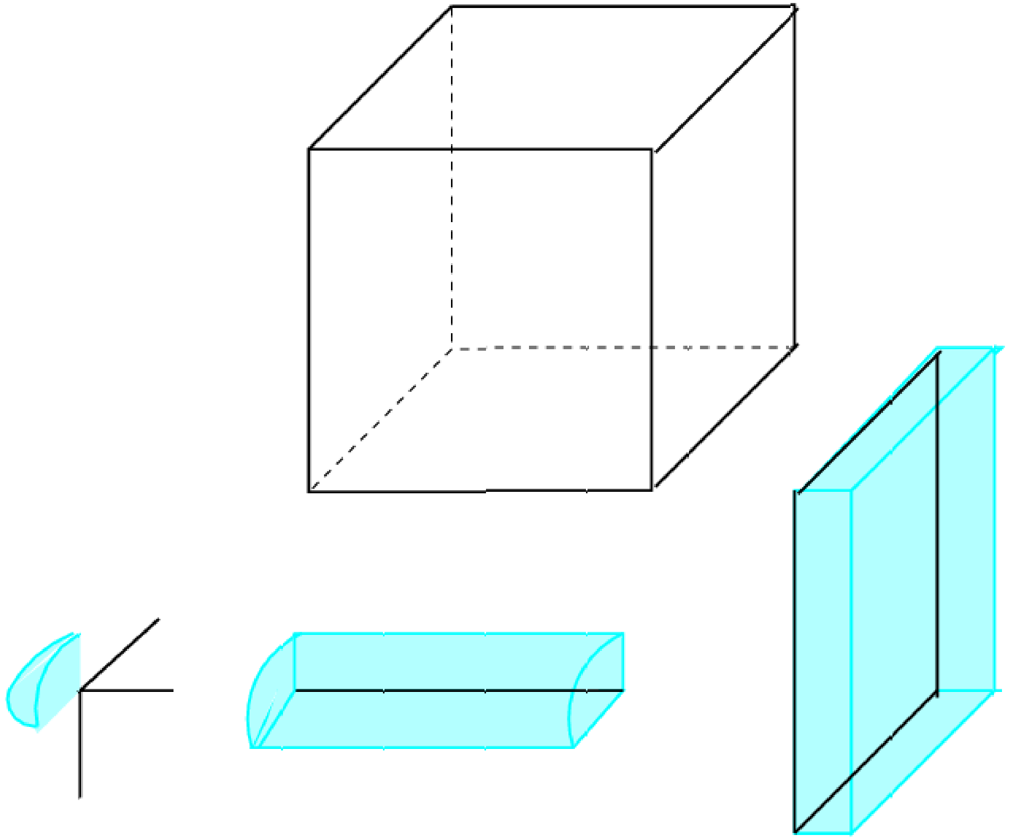
Subgradient equations

$$\hat{u} = -\nabla \ell(\hat{\beta}; T), \qquad \hat{u} \in \partial \mathcal{P}(\hat{\beta})$$

A model for queries: $Q^U = Q^U(\hat{\beta}, \hat{u})$

Conditional inference

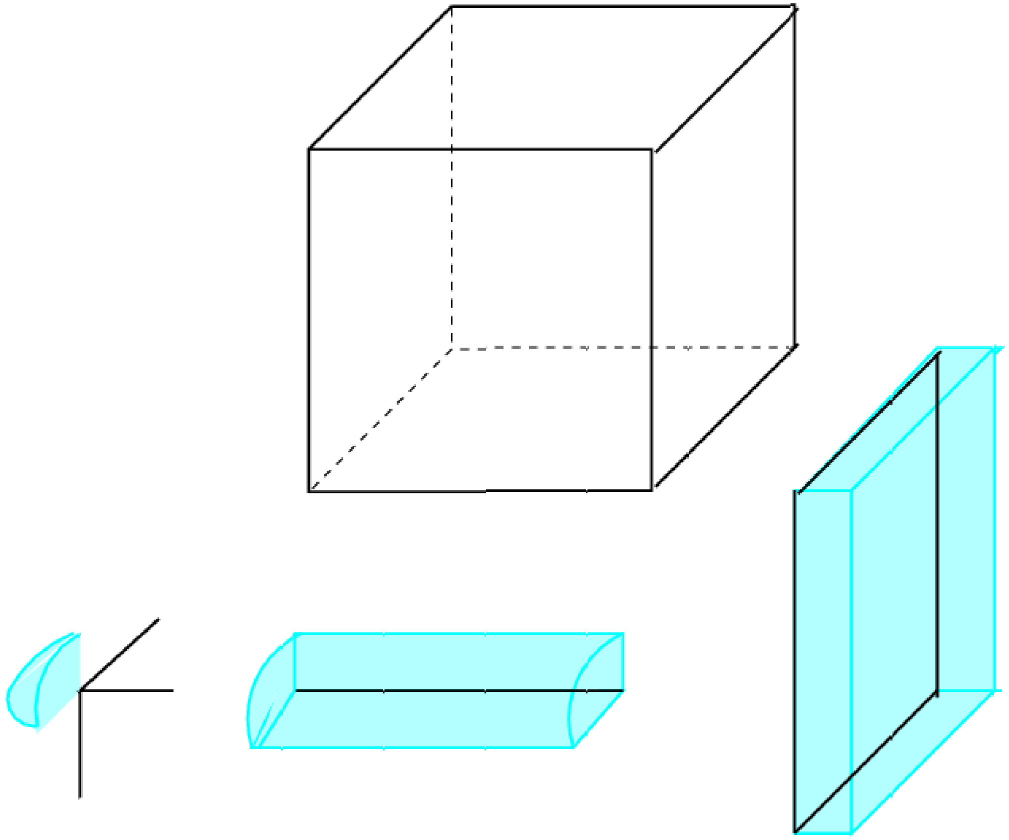
Structure inducing penalties



$$u \in \partial \mathcal{P}(\beta) \iff \beta \in N_u K.$$

Conditional inference

Adjustment factor



$$\zeta^*(T) = \int_{\{(u,\beta):Q^U(\beta,u)=q\}} g_\omega\left(\nabla\ell(\beta;T)+u\right)J(T,\beta,u)\mathcal{H}(du\ d\beta)$$

Conditional inference

What is the payoff?

- Many structure-detection algorithms in modern applied statistics can be cast as convex problems.

Canonical example

- LASSO (noisy version of compressed sensing)

$$\text{minimize}_{\beta} \ell(\beta; T) + \lambda \|\beta\|_1$$

- Solution is sparse for large values of λ.
- Also **hugely influential** over last 20 years in statistics.

Conditional inference

Randomized LASSO (Tian-Harris et al. 2016, arxiv/1609.05609)

$$\text{minimize}_{\beta} \ell(\beta; T) + \lambda \|\beta\|_1 - \omega^T \beta + \frac{\epsilon}{2} \|\beta\|_2^2$$

- Query:

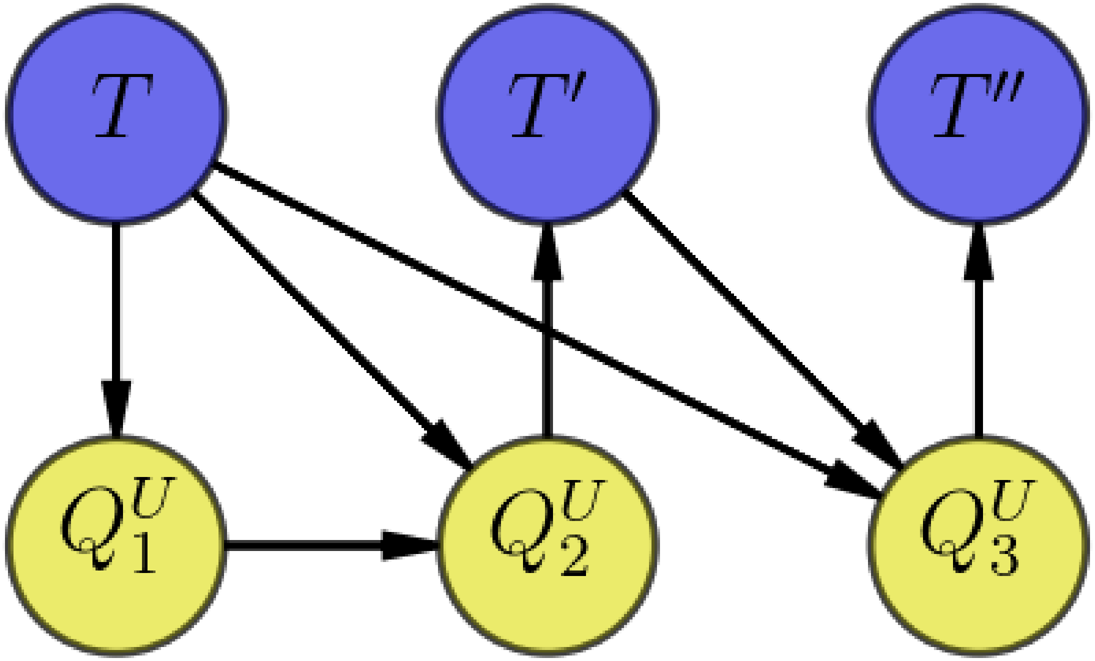
$$Q^U(T, \omega) = \hat{u}(T, \omega) = u_{\text{obs}}$$

- Adjustment:

$$\zeta^*(T) = \int_{N_u[-\lambda,\lambda]^p} g_\omega \left(\nabla \ell(T; \beta) + \epsilon \beta + u_{\text{obs}} \right) \ \mathcal{H}(d\beta)$$

Conditional inference

What is the payoff?

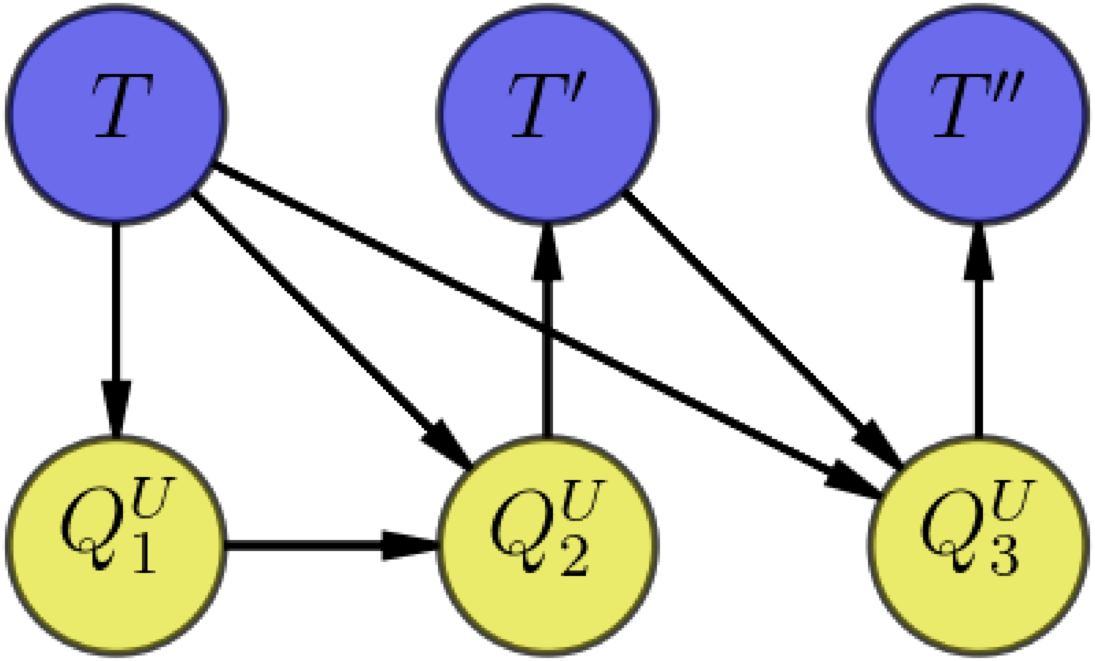


- Many queries can be combined.

- Fairly flexible set of analysis pipelines can be subsumed in this model.

Conditional inference

What is the payoff?



$$\zeta^*(T, T', T'') = \zeta^*_{1,q_1}(T) \cdot \zeta^*_{2,(q_1,q_2)}(T) \cdot \zeta^*_{3,(q_1,q_2,q_3)}(T, T')$$

Conditional inference

What is the payoff?

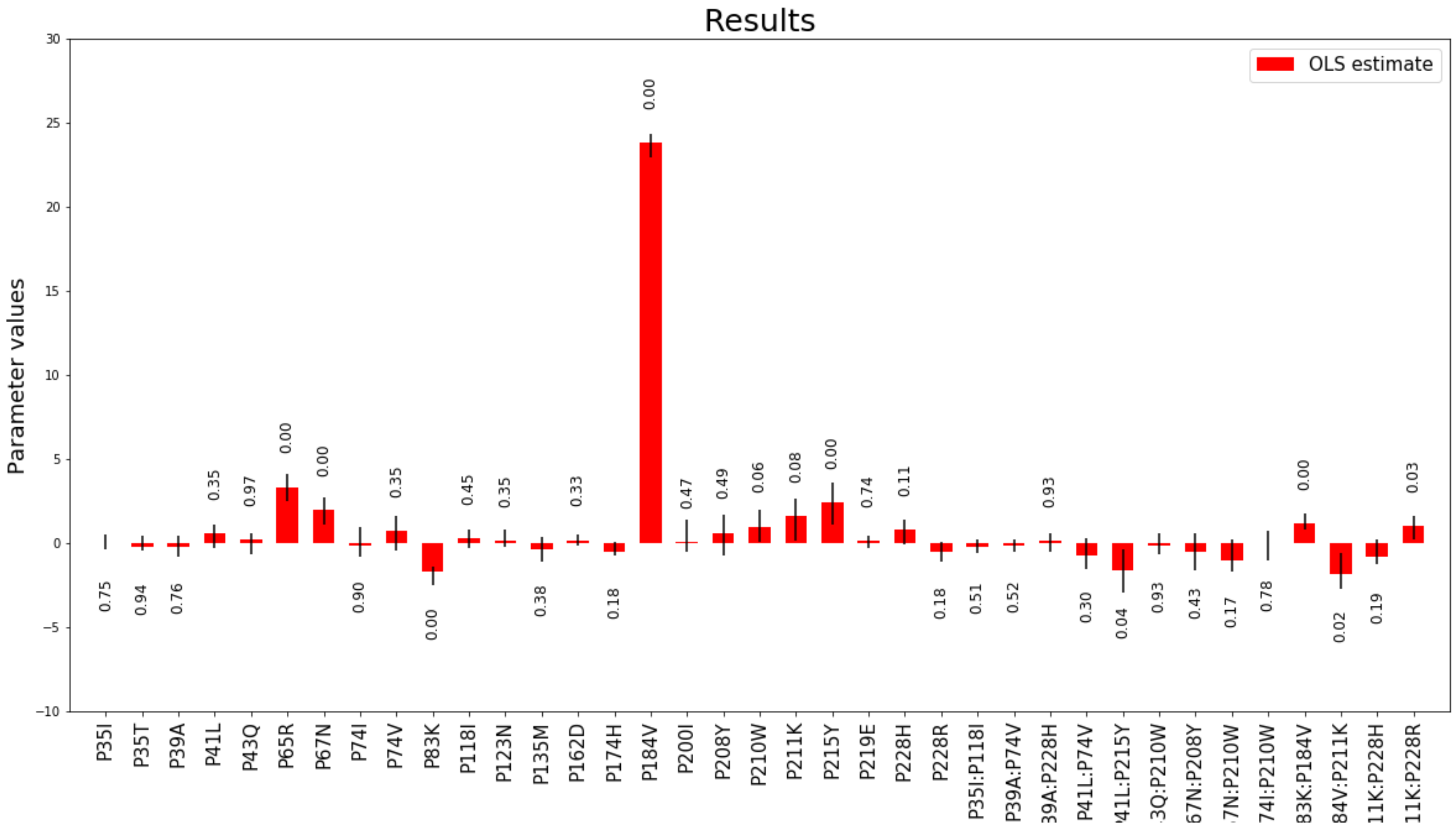
- Revisit our HIV resistance data: $n, p = 633, 91$.
- Goal: predict in-vitro resistance from mutation pattern.

Workflow

1. Search for important main effects using (randomized) marginal screening at some threshold.
2. U decides that even though mutation K65R was not discovered 1., it should be included.
3. Interaction effects for these first stage mutations are discovered using a (randomized) LASSO.
4. Report desired p-values, point estimates, intervals.

Conditional inference

What is the payoff?



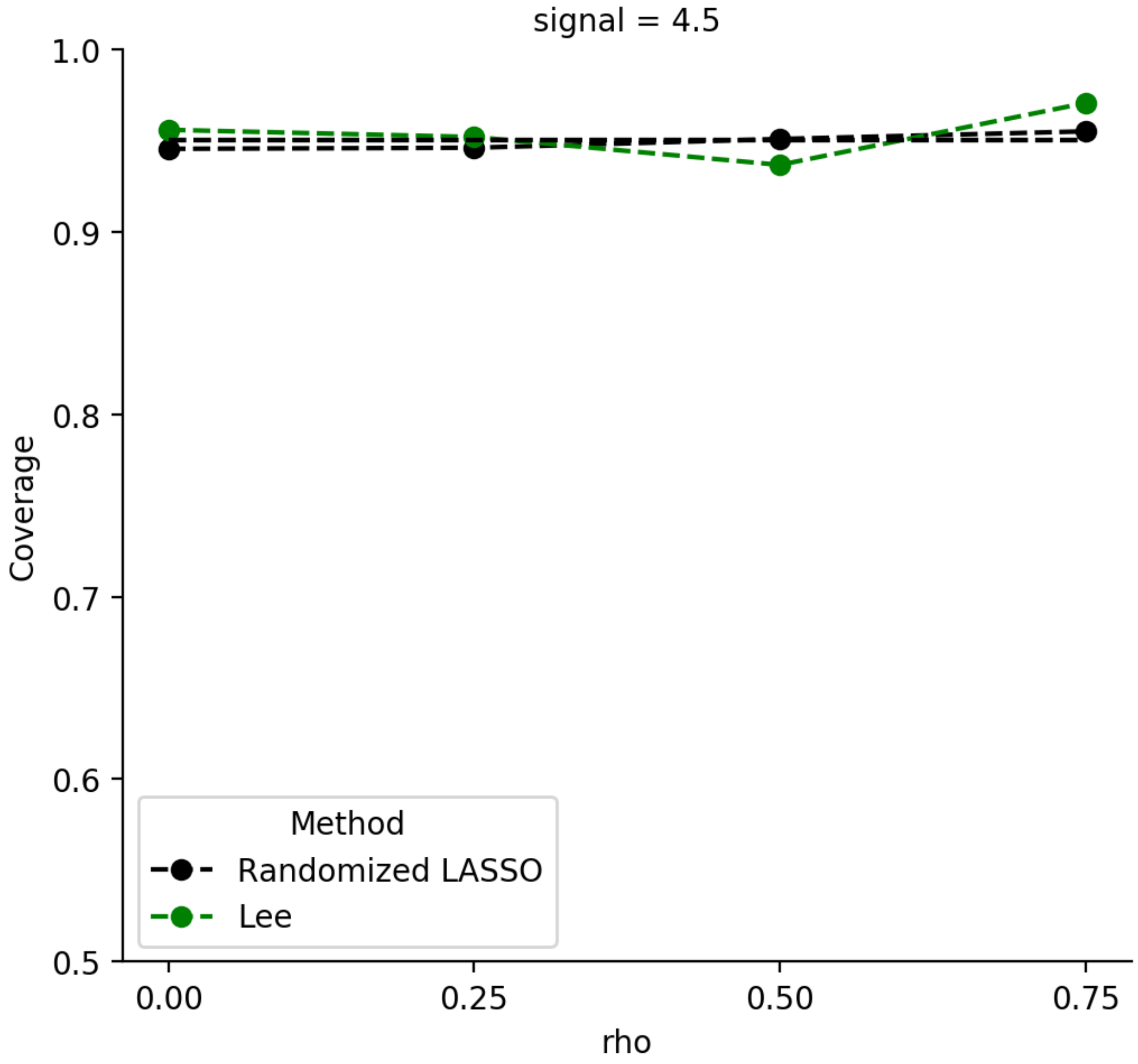
- Not clear how to do valid inference in other ways besides data splitting (or collecting new data).

Conditional inference

Cost of selection

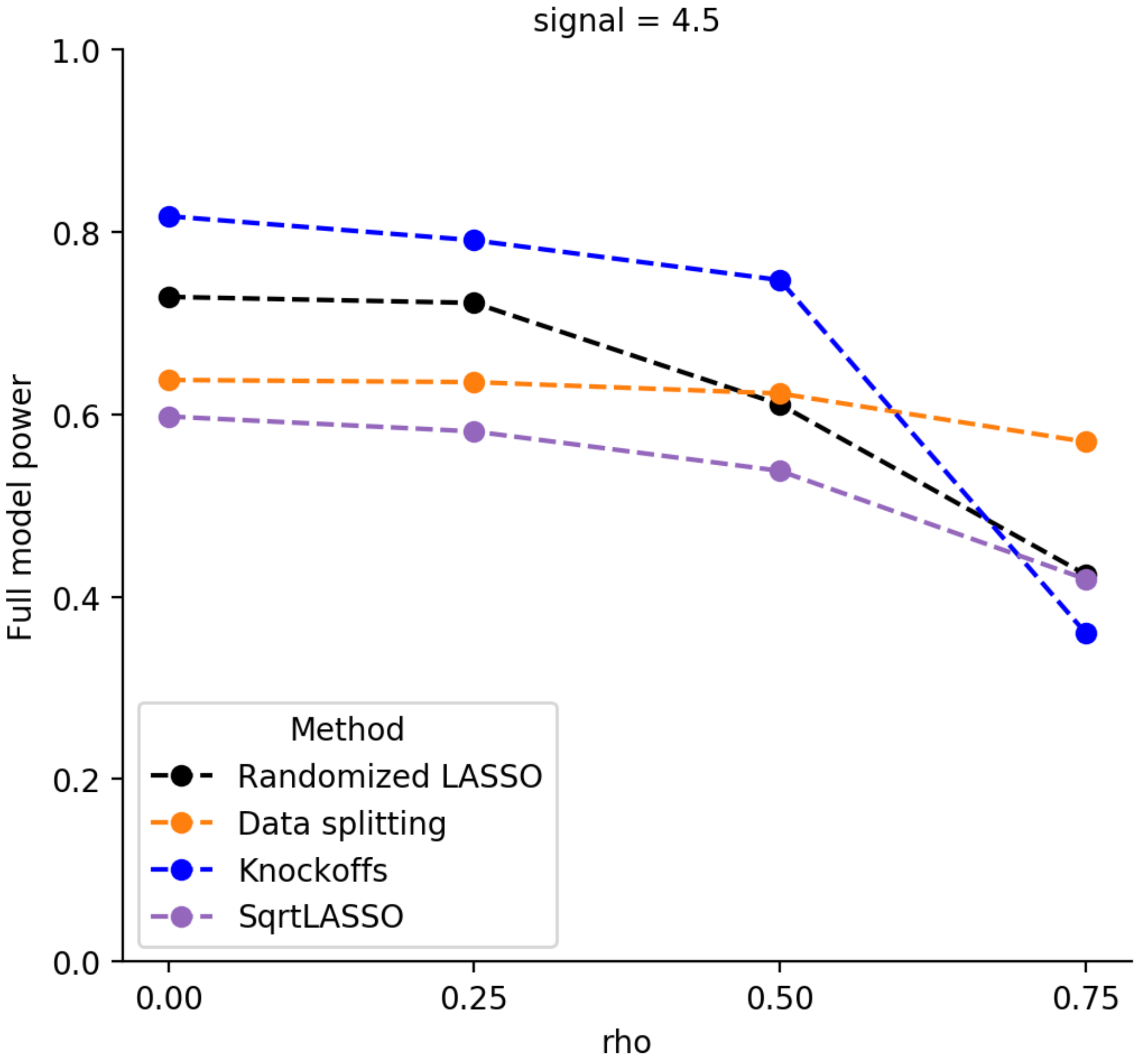
Conditional inference

Does it work?



Conditional inference

Does it work?



Conditional inference

Challenges

- Practical concerns:

  1. Tradeoff between selection quality and inferential power.

- Theoretical properties (some preliminary results e.g. Tian and Taylor (2018)):

  1. consistency
  2. CLT
  3. High dimensions

- Computational properties:

  1. Evaluation of ζ^*
  2. Quality of MCMC (also a theoretical question)

Conclusion

Takeaways

- Modern science requires statistics to adjust to how data is used.
- Simultaneous and conditional inference: we need both!
- Interesting practical, theoretical and computation questions.

Image credits

- Wikipedia for Tukey and Manhattan plot.