Principled Statistical Inference in Data Science

Todd A. Kuffner
Washington University in St. Louis
kuffner@math.wustl.edu

G. Alastair Young
Imperial College London
alastair.young@imperial.ac.uk

August 5, 2017

Abstract

We discuss the challenges of principled statistical inference in modern data science. Conditionality principles are argued as key to achieving valid statistical inference, in particular when this is performed after selecting a model from sample data itself.

Keywords and phrases: Statistical inference; principles; data science; conditioning; post-selection inference; validity.

1 Introduction

In recent times, even prominent figures in statistics have come to doubt the importance of foundational principles for data analysis.

> "If a statistical analysis is clearly shown to be effective at answering the questions of interest, it gains nothing from being described as principled." (Speed, 2016)

The above statement was made by Terry Speed in the September 2016 *IMS Bulletin*. It is our primary purpose in this article to refute Professor Speed's assertion! We argue that a principled approach to inference in the data science context is essential, to avoid erroneous conclusions, in particular invalid statements about significance.

We will be concerned here with statistical inference, specifically calculation and interpretation of $p-$values and construction of confidence intervals. While the greater part of the data science literature is concerned with prediction rather than inference, we believe that our focus is justified for two solid reasons. In many circumstances, such, say, as microarray studies, we are interested in identifying significant 'features', such as genes linked to particular forms of cancer, as well as the identity and strength of evidence. Further, the current

reproducibility crisis in science demands attention be paid to the formal repeated sampling properties of inferential methods.

2 Key principles

The key notions which should drive consideration of methods of statistical inference are: validity, whether a claimed criterion or assumption is satisfied, regardless of the true unknown state of nature; and, relevance, whether the analysis performed is actually relevant to the particular data sample under study.

It is most appropriate to consider the notion of validity in the context of procedures motivated by the principle of error control. Then, a valid statistical procedure is one for which the probability is small that the procedure has a higher error rate than stated. For example, the random set $\mathcal{C}_{1-\alpha}$ is an (approximately) valid $(1-\alpha)$ confidence set for a parameter θ if $\Pr(\theta \notin \mathcal{C}_{1-\alpha}) = \alpha + \epsilon$ for some very small (negligible) ϵ, whatever the true value of θ.

Relevance is achieved by adherence to what we term the 'Fisherian proposition' (Fisher 1925, 1934). This advocates appropriate conditioning of the hypothetical data samples that are the basis of non-Bayesian statistics. Specifically, the Conditionality Principle, formally described below, would maintain that to ensure relevance to the actual data under study the hypothetical repetitions should be conditioned on certain features of the available data sample.

It is useful to frame our discussion as done by Cox & Mayo (2010). Suppose that for testing a specified null hypothesis $H_0 : \psi = \psi_0$ on an interest parameter ψ we calculate the observed value t_{obs} of a test statistic T and the associated $p-$value $p = P(T \geq t_{obs}; \psi = \psi_0)$. Then, if p is very low, e.g. 0.001, t_{obs} is argued as grounds to reject H_0 or infer discordance with H_0 in the direction of the specified alternative, at level 0.001.

This is not strictly valid, since it amounts to choosing the decision rule based on the observed data (Kuffner & Walker, 2017). A valid statistical test requires that the decision rule be specified in advance. However, there are two rationales for the interpretation of the $p-$value described in the preceding paragraph.

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

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

The evidential construal in (2) is only accomplished to the extent that it can be assured that the small observed $p-$value is due to the actual data-generating process being discrepant from that described by H_0. As noted by Cox & Mayo (2010), once the requirements of (2) are satisfied, the low error-rate rationale (1) follows.

The key to principled inference which provides the required interpretation is to ensure relevancy of the sampling distribution on which $p-$values are based. This is achieved through the Conditionality Principle, which may formally be stated as follows.

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

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

In practice, the requirement that A be distribution constant is often relaxed. It is (see, for instance, Barndorff-Nielsen & Cox, 1994) well-established in statistical theory that to condition on the observed data value of a random variable whose distribution does depend on θ might, under some circumstances, be convenient and meaningful, though this would in some sense sacrifice information on θ.

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

Suppose that the minimal sufficient statistic can again be partitioned as $S = (T, A)$, where the distribution of T given $A = a$ depends only on ψ. We may extend the Conditionality Principle to advocate that inference on ψ should be based on this latter conditional distribution, under appropriate conditions on the distribution of A. We note that the case where the distribution of A depends on λ but not on ψ is just one rather special instance.

A simple illustration of conditioning on an exactly distribution constant statistic is given by Barndorff-Nielsen & Cox (1994, Example 2.20). Suppose Y_1, Y_2 are independent Poisson variables with means $(1 - \psi)l, \psi l$, where l is a known constant. There is no reduction by sufficiency, but the random variable $A = Y_1 + Y_2$ has a known distribution, Poisson of mean l, not depending on ψ. Inference would, say, be based on the conditional distribution of Y_2, given $A = a$, which is binomial with index a and parameter ψ.

Justifications for many standard procedures of applied statistics, such as analysis of 2×2 contingency tables, derive from the Conditionality Principle, even when A has a distribution that depends on both ψ and λ, but when observation of A alone would make inference on ψ imprecise. The contingency table example concerns inference on the log-odds ratio when comparing two binomial variables: see Barndorff-Nielsen & Cox (1994, Example 2.22). Here Y_1, Y_2 are independent binomial random variables corresponding to the number of successes in (m_1, m_2) independent trials, with success probabilities (θ_1, θ_2). The interest parameter is $\psi = \log\{\theta_2/(1 - \theta_2)\} - \log\{\theta_1/(1 - \theta_1)\}$. Inference on ψ would, following the Conditionality Principle, be based on the conditional distribution of Y_2 given $A = a$, where $A = Y_1 + Y_2$ has a marginal distribution depending in a complicated way on *both* ψ *and* whatever nuisance parameter λ is defined to complete the parametric specification.

Central to our discussion, therefore, is recognition that conditioning an inference on the observed data value of a statistic which is, to some degree, informative about the parameter of interest is an established part of statistical theory. Conditioning is supported as a means of controlling the Type 1 error rate, while ensuring relevance to the data sample under test.

Of course, generally, conditioning will run counter to the objective of maximising power (minimising Type 2 error rate), which is a fundamental principle of much of statistical theory. However, loss of power due to adoption of a conditional approach to inference may be very slight, as demonstrated by the following example.

Suppose Y is normally distributed as $N(\theta, 1)$ or $N(\theta, 4)$, depending on whether the outcome δ of tossing a fair coin is heads ($\delta = 1$) or tails ($\delta = 2$). It is desired to test the null hypothesis $H_0 : \theta = -1$ against the alternative $H_1 : \theta = 1$, controlling the Type 1 error rate at level $\alpha = 0.05$. The most powerful unconditional test, as given by Neyman-Pearson optimality theory, has rejection region given by $Y \geq 0.598$ if $\delta = 1$ and $Y \geq 2.392$ if $\delta = 2$. The Conditionality Principle advocates that instead we should condition on the outcome of the coin toss, δ. Then, given $\delta = 1$, the most powerful test of the required Type 1 error rate rejects H_0 if $Y \geq 0.645$, while, given $\delta = 2$ the rejection region is $Y \geq 2.290$. The power of the unconditional test is 0.4497, while the power of the more intuitive conditional test is 0.4488, only marginally less.

Further support for conditioning, to eliminate dependence of the inference on unknown nuisance parameters, is provided by the Neyman-Pearson theory of optimal frequentist inference (see, for example, Young & Smith, 2005).

A key context where this theory applies is when the parameter of interest is a component of the canonical parameter in a multiparameter exponential family model. Suppose Y has a density of the form

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

Then (T_1, T_2) is minimal sufficient and the conditional distribution of $T_1(Y)$, given $T_2(Y) = t_2$, say, depends only on ψ. The distribution of $T_2(Y)$ may, in special cases, depend only on λ, but will, in general, depend in a complicated way on both ψ and λ. The extended form of the Conditionality Principle argues that inference should be based on the distribution of $T_1(Y)$, given $T_2(Y) = t_2$. But, in Neyman-Pearson theory this same conditioning is justified by a requirement of full elimination of dependence on the nuisance parameter λ, achieved in the light of completeness of the minimal sufficient statistic only by this conditioning. The resulting conditional inference is actually optimal, in terms of furnishing a uniformly most power unbiased test on the interest parameter ψ: see Young & Smith (2005, Chapter 7).

Our central thesis is that the *same* Fisherian principles of conditioning are necessary to steer appropriate statistical inference in a data science era, when models and the associated inferential questions are arrived at after examination of data:

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

Our assertion is that appropriate conditioning is needed to ensure validity of the inferential methods used. Importantly, however, the justifications used for conditioning are not new, but mirror the arguments used in established statistical theory.

3 Classical and 'post-selection' inference

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

$$P(\text{reject } H_0|M, H_0) \leq \alpha.$$

The appropriate paradigm for data science is, in our view, the structure for inference that is known as 'post-selection Inference', as described, for example, by Lee et al. (2016) and Fithian, Sun & Taylor (2014).

Now it is recognised that inference is performed after having arrived at a statistical model adaptively, through examination of the observed data.

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

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

That is, we require that we control the Type 1 error rate of the test given that it was actually performed. The thinking leading to this principle is really just a 21st century re-expression of Fisherian thought.

A simple example, the 'File Drawer Effect', serves to illustrate the central ideas, and is a template (Fithian, Sun & Taylor, 2014) for how statistical inference is performed in data science. Suppose data consists of a set of n independent observations Y_i distributed as $N(\mu_i, 1)$. We choose, however, to focus attention only on the apparently large effects, selecting for formal inference only those indices i for which $|Y_i| > 1$, $\hat{I} = \{i : |Y_i| > 1\}$. We wish, for each $i \in \hat{I}$, to test $H_{0,i} : \mu_i = 0$, each individual test to be performed at significance level $\alpha = 0.05$.

A test which rejects $H_{0,i}$ when $|Y_i| > 1.96$ is invalidated by the selection of the tests to be performed. Though the probability of falsely rejecting a given $H_{0,i}$ is certainly α, since most of the time that hypothesis is not actually tested, the error rate among the hypotheses that are actually selected for testing is much higher than α.

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

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

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

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

In a formal framework of post-selection inference, we assume that our data Y lies in some measurable space with unknown sampling distribution $Y \sim F$. The task is to pose, on the basis of Y itself, a reasonable probability model $\hat{M}$, then carry out inference, using the same data Y.

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

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

In terms of our discussion above, the selection event S will typically be informative about the quantity θ of interest, and conditioning will therefore discard information. But, to ignore the selection event loses control over the (Type 1) error rate, potentially badly. Principled inference requires conditioning on the selection event, and therefore drawing inferences from leftover information in Y, given S.

4 Example: File Drawer Effect

Consider the File Drawer Effect example, but now take the selection event as $\{Y > 1\}$. We compare 'nominal' confidence intervals, not accounting for selection, and selective confidence intervals, of coverage 95%.

Figure 1 compares the selective and non-selective confidence intervals, as a function of the observed value Y. If Y is much larger than 1, there is hardly any selection bias, so no adjustment for selection is really required. When Y is close to 1, the need to properly account for selection is stark.

Figure 2 compares the lengths of the selective and non-selective confidence intervals: the non-selective interval is $Y \pm 1.96$, and therefore has length 3.92, whatever the true mean μ or data value Y. Figure 3 illustrates the coverage of the invalid non-selective confidence interval: this generally exceeds 95%, but if the true mean μ is much less than 1, undercoverage is substantial. If the true mean is exactly 1, but only in this case, the non-selective interval has coverage exactly 95%.

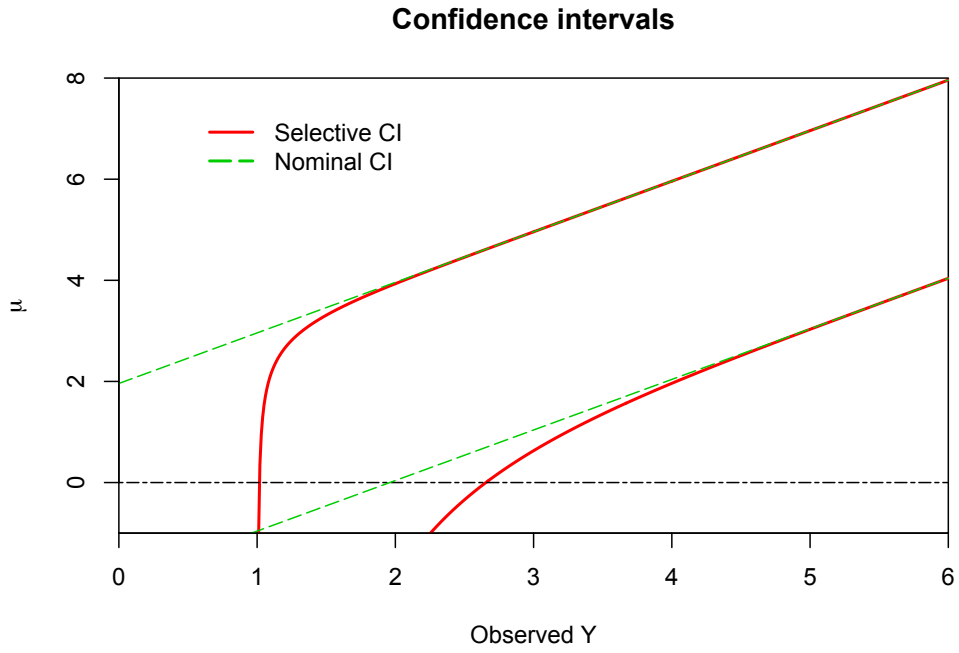

Figure 1: File Drawer Effect, comparison of selective and non-selective confidence intervals.

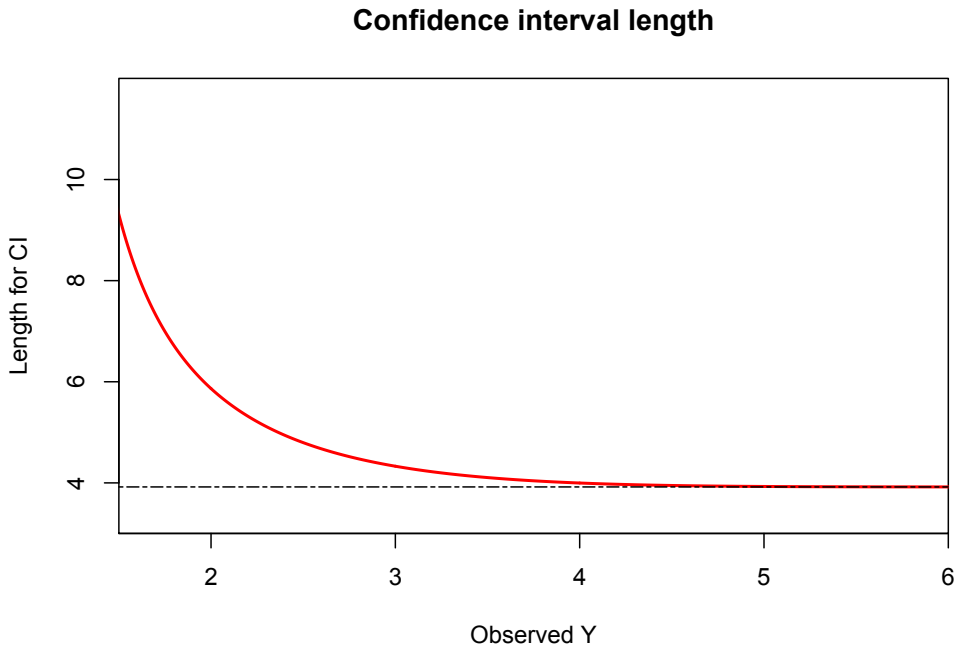

Figure 2: File Drawer Effect, length of selective confidence interval compared to fixed length of non-selective interval.

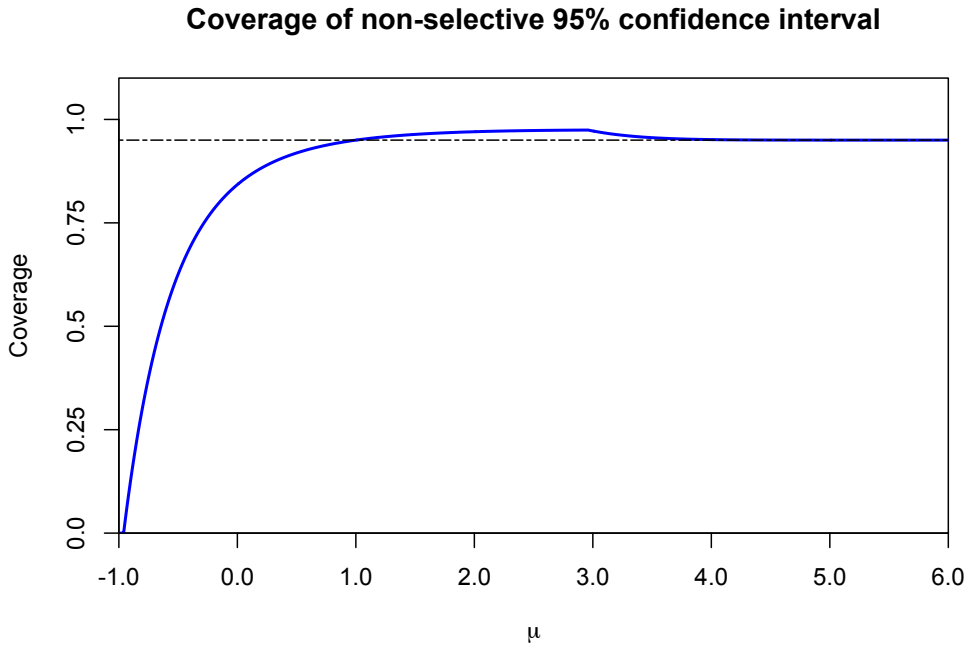

Figure 3: File Drawer Effect, coverage of non-selective confidence interval.

5 Borrowing from classical theory

Conditioning the inference performed on the selection event is especially convenient if Y is assumed to have an exponential family distribution. Then the distribution of Y conditional on a measurable selection event $S(Y)$ is also an exponential family distribution, allowing support for the techniques of selective inference to be drawn from the established classical theory for inference in exponential families: see Fithian, Sun & Taylor (2014).

As further illustration, we consider a simple normal linear regression model. Suppose that $Y \sim N_n(\mu, \sigma^2 I_n)$, with $\mu \equiv X\beta$, β a vector of unknown parameters, and X a matrix of p predictors with columns $X_1, \ldots, X_p \in \mathbb{R}^n$. We suppose, for simplicity, that σ^2 is known.

Suppose that some variable selection procedure is utilised to select a model $M \subset \{1, \ldots, p\}$ consisting of a subset of the p predictors. Under the selected model, $\mu = X_M \beta^M$, where X_M is $n \times |M|$, with columns $(X_M)_1, \ldots, (X_M)_{|M|}$, say: we assume that X_M is of full rank, so that $\beta^M = (\beta^M_1, \ldots, \beta^M_{|M|})$ is well-defined.

Conventional principles of inference in exponential family distributions, adapted to this selective inference context, indicate that inference on β^M_j should be based on the conditional distribution of $(X_M)^T_j Y$, given the observed values of $(X_M)^T_k Y$, $k = 1, \ldots, |M|, k \neq j$, and the selection event that model M is chosen. Use of this sampling distribution is termed (Fithian, Sun & Taylor, 2014) inference under the 'selected model'.

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

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

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

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

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

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

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

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

The issue then arises of whether to perform inference under the selected or saturated models. Do we assume $P_M^\perp \mu = 0$, or treat it as an unknown nuisance parameter, to be eliminated by further conditioning?

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

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

If we condition on V when, in fact, $P_M^\perp \mu = 0$, we might expect to lose power, while inferential procedures may badly lose their control of (Type 1) error rate if this quantity is large, so that the selected model is actually false. We contend that such further conditioning is, however, necessary to ensure validity of the conclusions drawn from the specific data set under analysis.

6 Example: Bivariate Regression

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

We choose (using Least Angle Regression, lasso, or some such procedure) a 'one-sparse model', that is X_M is specified to have just one column. The selection procedure chooses $M = \{1\}$ if $|Y_1| > |Y_2|$ and $M = \{2\}$ otherwise.

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

The selected model has Y distributed as $N_2((\mu_1, 0), I_2)$. Inference on μ_1 would base a test of $H_0 : \mu_1 = 0$ against $H_1 : \mu_1 > 0$ on rejection for large values of Y_1, $Y_1 > c$, say. This test may be expressed as $H_0 : \eta^T \mu = 0$, with $\eta = (1, 0)^T$. In the test of nominal Type 1 error α based on the selected model, c is fixed by requiring $P_{H_0}(Y_1 > c \,|M, |Y_1| > |Y_2|) = \alpha$, explicitly assuming that $\mu_2 = 0$. Notice that, in terms of the discussion of the previous Section, there is no U in this example, since X_M has only one column. The issue is whether to condition only on the selection event, or also on $V = P_M^{\perp} Y \equiv P_{\eta}^{\perp} Y = Y_2$.

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

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

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

Figure 4 compares the power functions of the tests in the selected and saturated models. If the selected model is true, $\mu_2 = 0$, the test under the selected model is generally more powerful than the test derived from the saturated model, though we note the latter is actually marginally more powerful for small values of μ_1. However, if the selected model is false (the Figure illustrates the case $\mu_2 = 2$), control of Type 1 error at the nominal 5% level is lost: the test of $\mu_1 = 0$ has Type 1 error rate exceeding 10% when the selected model is false and μ_2 is actually equal to 2.

Figures 5 and 6 examine the distributions of Y_2 and Y_1 respectively, conditional on the selection event $|Y_1| > |Y_2|$. Figure 5 demonstrates that the conditional distribution of Y_2 varies little with μ_1, the interest parameter, so that Y_2 is rather uninformative about μ_1. By contrast, the conditional distribution of Y_1, shown in Figure 6, depends strongly on μ_1. Conditioning on the observed value of Y_2 is justified on the grounds that conditional on the selection event this value is, relative to Y_1, uninformative about μ_1, while this further conditioning ensures exact control of Type 1 error.

What do we conclude from this analysis? The operational difference between the saturated and selected model perspectives may (Fithian, Sun & Taylor, 2014) be important in key practical contexts, such as early steps of sequential model-selection procedures. However, the case being made is that a principled approach to inference is forced to give central consideration to the saturated model in contexts such as those discussed here, where valid interpretation of significance is key. The Fisherian proposition requires conditioning on the selection event, as it is necessary (Young, 1986) to condition the inference on features of the data sample which control the propensity for extreme value of the test statistic to occur for spurious reasons. Precise control of the Type 1 error rate then demands elimination of nuisance parameter effects, achieved only by further conditioning on $P_{\eta}^{\perp} Y$: this leads to inference from the saturated model perspective.

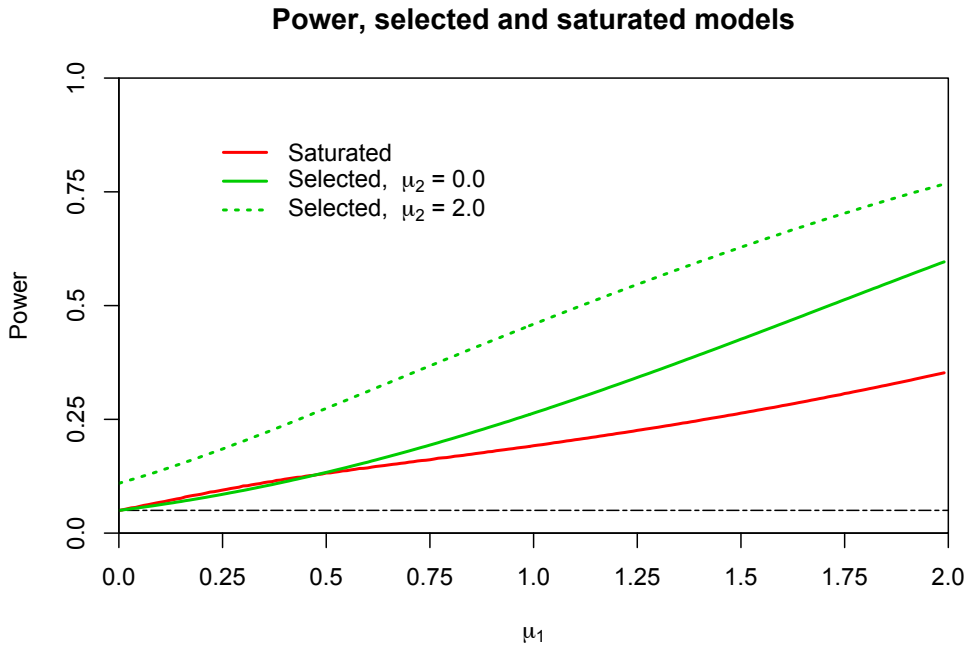

Figure 4: Bivariate regression, power functions under selected and saturated models.

7 Some other points

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

In some generality, the selection event can be expressed as a polyhedron $S(Y) = \{AY \leq b\}$, for A, b not depending on Y. This is true for forward stepwise regression, the lasso with fixed penalty parameter λ, Least Angle Regression and other procedures. If inference is required for $\eta^T \mu$, then further conditioning on $P_\eta^\perp Y$ yields the conditional distribution required for the inference to be a truncated Gaussian with explicitly available endpoints, allowing a simple analytic solution.

Notice that here conditioning on $P_\eta^\perp Y$ is generally promoted (Lee et al., 2016) as a means of obtaining an analytically simple distribution for the inference. We have argued, however, that this conditioning is necessary to eliminate dependence on the nuisance parameter and provide control over Type 1 error. Marginally, that is ignoring the selection event, $\eta^T Y$ is independent of $P_\eta^\perp Y$, so the conditioning is justified by ancillarity, but this is not true conditional on the selection event: justification stronger than analytic convenience is provided by necessary elimination of the nuisance parameter.

(ii) In the non-Gaussian setting and in general under the selective model, Monte Carlo procedures, such as MCMC and acceptance/rejection methods, will be necessary to determine the necessary conditional distribution of Y, but such computational demands are unlikely

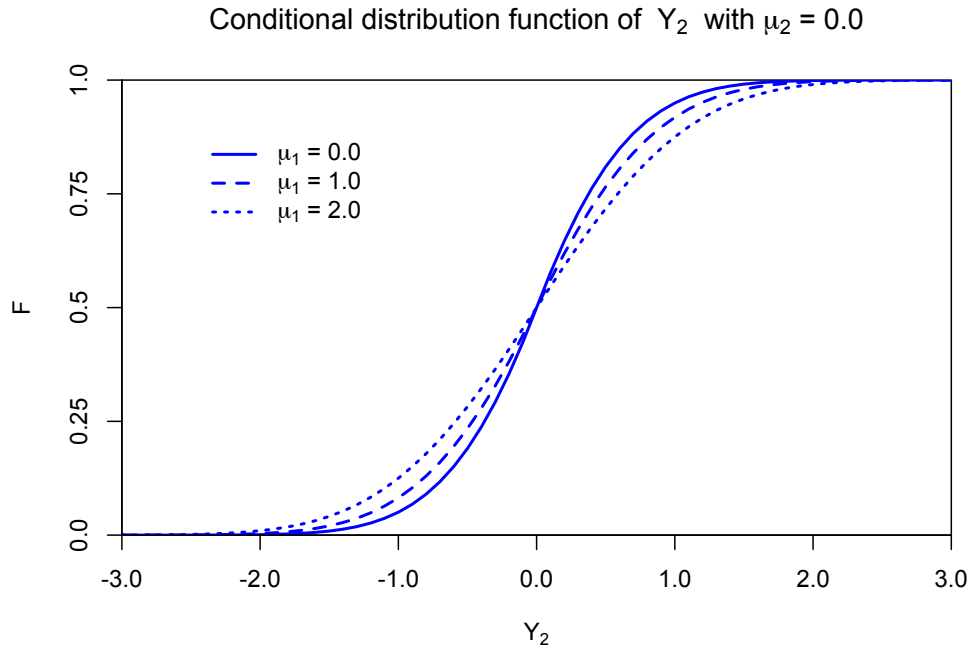

Figure 5: Bivariate regrssion, conditional distributions of Y_2.

to prove an obstacle to principled inference (Young & DiCiccio, 2010).
(iii) Tibshirani et al. (2015) offer a different perspective on selective inference, potentially relevant to data science.

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

The notion of validity now is that, for the selected target, which is not fixed but varies according to Y, it is required that under repeated sampling of Y, a specified proportion $1-\alpha$ of the time, the inference on the selected target should be correct. Implicitly, perhaps, this is what is sought in much of data science. However, this perspective abandons the requirement that we have argued is central to principled inference, of ensuring validity and relevance to the actual data sample.

8 Conclusions

We have argued that a principled approach to inference in data science is necessary to provide the rationale by which claimed error-rate properties of inferential procedures are justified. The appropriate conceptual framework for valid inference is that discussed in the statistical

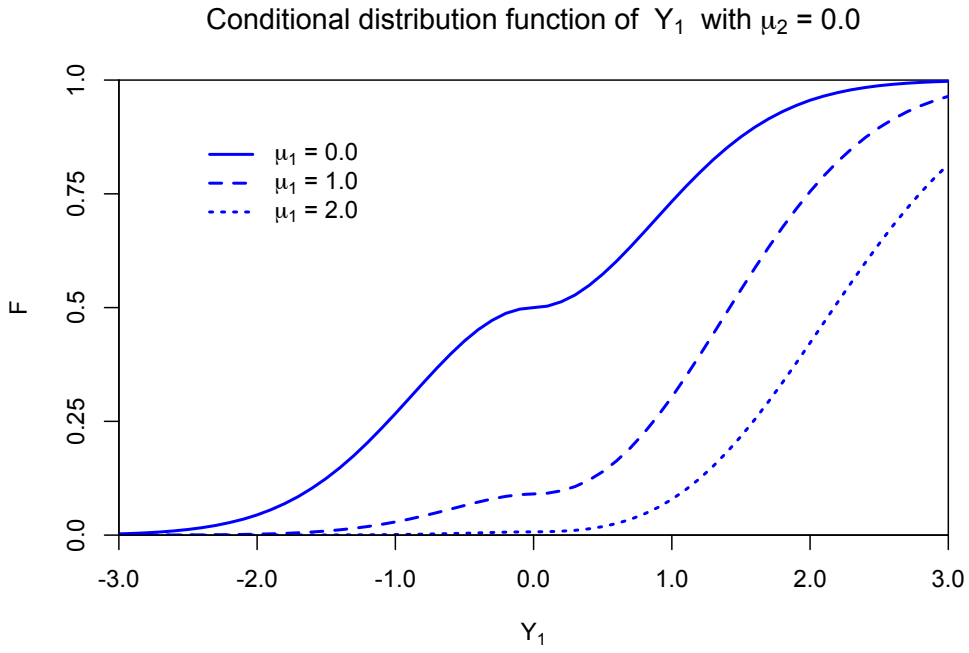

Figure 6: Bivariate regression, conditional distributions of Y_1.

literature as 'post-selection inference', which is based on ensuring relevance of sampling distributions used for inference to the particular data sample. These are classical, Fisherian ideas: no new paradigm for inference in data science is involved.

Specifically, inference after adaptive model determination ('data snooping') requires conditioning on the selection event and control of the error rate of the inference given it was actually performed. As commented by Fithian, Sun & Taylor (2014) 'the answer must be valid, given that the question was asked.'

Care is however required, as the selected model for inference may be wrong, and can lead to substantially distorted error rates. The primary cause is the assumption that nuisance parameters effects are known: elimination by the classical device of (further) conditioning ensures precise control of error rates. What we have described as the saturated model framework is the appropriate basis for inference. The potential loss of accuracy (power) through the necessary conditioning is undesirable, but may not be practically consequential: possible overconditioning is a worthwhile price to be paid for validity.

References

Barndorff-Nielsen, O.E. & Cox, D.R. *Inference and Asymptotics.* Chapman & Hall, 2004.

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

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

Fisher, R.A. Theory of statistical estimation. *Proc. Camb. Phil. Soc.*, **22**, 700–725, 1925.

Fisher, R.A. Two new properties of mathematical likelihood. *Proc. R. Soc. Lond. A*, **144**, 285–307.

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

Kuffner, T.A. & Walker, S.G. Why are p-values controversial? *American Statistician*, to appear, 2017.

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

Speed, T. Terence's stuff: Principles. *IMS Bulletin*, **45**, 17.

Tibshirani, R., Rinaldo, A., Tibshirani, R. & Wasserman, L. Uniform asymptotic inference and the bootstrap after model selection. *arXiv preprint arXiv:1506.06266*, 2015.

Young, G.A. Conditioned data-based simulations: some examples from geometrical statistics. *Int. Stat. Rev.*, **54**, 1–13.

Young, G.A. & DiCiccio, T.J. Computer-intensive conditional inference. In Mantovan, P. & Secchi, P. (Eds.) *Complex Data Modeling and Computationally Intensive Statistical Methods*, Springer-Verlag Italia, 138–150, 2010.

Young, G.A. & Smith, R.L. *Essentials of Statistical Inference*. Cambridge University Press, 2005.