Freedman's paradox

Céline Cunen

Department of Mathematics, University of Oslo

07/12/2018

The (replication) crisis in Science

advancedsciencestudies.wordpress.com

There is growing concern on the validity of scientific findings. Specifically, there are indications that many (most?) published results are false discoveries, i.e. spurious associations.

How can this be explained?

- fraud?
- publishing practices, institutional incentives, the file-drawer problem?
- flawed statistical tools?

The (replication) crisis in Science

The phenomenon sometimes referred to as Freedman's paradox was described in Freedman (1983) and fits within this picture because it constitutes

- an explanation for how (reasonably) standard use of statistical methods can lead to false discoveries;
- a warning to statisticians and practitioners.

The (replication) crisis in Science

The phenomenon sometimes referred to as Freedman's paradox was described in Freedman (1983) and fits within this picture because it constitutes

- an explanation for how (reasonably) standard use of statistical methods can lead to false discoveries;
- a warning to statisticians and practitioners.

Plan:

- Freedman (1983): empirical and theoretical results.
- Paradox?
- Solutions to the R^2 problem.
- Model selection and post-selection inference.
- Solutions to post-selection problems.

Freedman (1983)

A linear regression setting with n observations of some response variable Y and explanatory variables $X_1, X_2, \ldots, X_p$,

$$Y = X\beta + \epsilon$$

with $\epsilon_i \sim N(0, \sigma^2)$. Here $p < n$, and we will be interested in:

- the coefficient of determination $R^2 = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \hat{y}_i)^2 + \sum_{i=1}^n (\hat{y}_i - \bar{y})^2}$ with $\widehat{y} = X\widehat{\beta}$;
- the test H_0: $\beta = 0$ with test statistic $F = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2 / p}{\sum_{i=1}^n (y_i - \hat{y}_i)^2 / (n - p - 1)}$;
- the tests H_0: $\beta_j = 0$ with test statistics $T_j = \widehat{\beta}_j / s_j$, with $s_j^2 = \widehat{\sigma}^2 \{(X^{\mathrm{t}} X)^{-1}\}_{j,j}$, and p-values p_j.

Freedman (1983)

Freedman considers the situation where $\beta = 0$, i.e. there really is no association between Y and X!

First, Freedman studies the behaviour of R^2, F and the p-values in a simple simulation study. He draws a number of datasets with both n and p reasonably large, say $n = 100$ and $p = 50$.

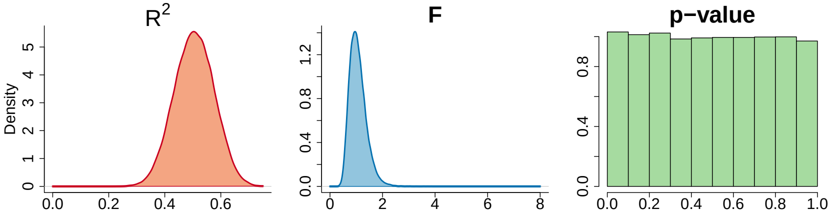
For each dataset he performs two rounds of regressions:

1. with all p variables;
2. with only q_α variables, where the selected variables are the ones with $p_j < \alpha$ in the first regression.

In the next slides we see the results of a large number of such simulations.

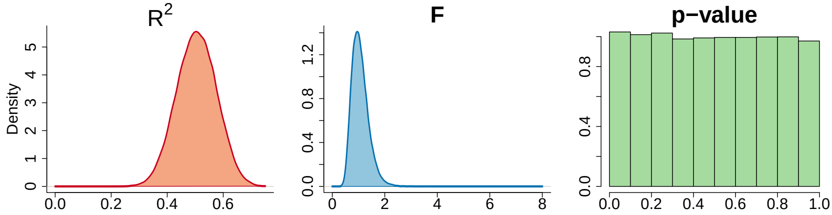
Freedman (1983) – Empirical results

In the first regression we observe,

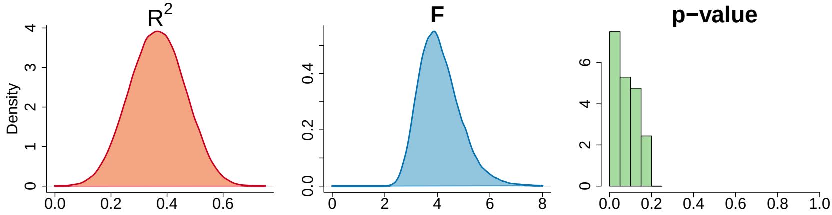


Freedman (1983) – Empirical results

In the first regression we observe,

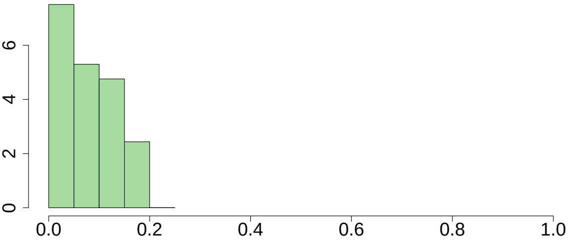


Then, redo the regression keeping only the variables with $p_j < 0.25$:



Freedman (1983) – Empirical results

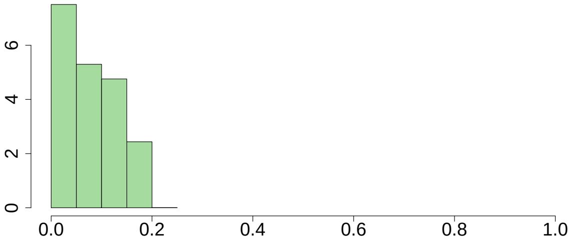
What happens with the p-values in the second regression?



- Many variables seem highly significant (say we use $\alpha_2 = 0.05$) and give the indication of an association between X and y.

Freedman (1983) – Empirical results

What happens with the p-values in the second regression?



- Many variables seem highly significant (say we use $\alpha_2 = 0.05$) and give the indication of an association between X and y.
- The distribution of p_j is no longer uniform.
- The probability of false rejections (type 1 error) is much higher than 0.05.
- Confidence intervals for β_j are no longer valid, i.e. $\Pr(\beta_j \in \mathrm{CI}_{0.95}) < 0.95$.

Freedman (1983) – Theoretical results (1)

We still assume $\beta = 0$. Suppose $n \to \infty$, $p \to \infty$ and $p/n \to \rho$ with $0 < \rho < 1$. Also, we assume that $rank(X) = p$ (no collinearity).

In the first regression, we have $R^2 \xrightarrow{pr} \rho$ and $F \xrightarrow{pr} 1$.

These results follow straightforwardly from the definitions of R^2 and F.

Freedman (1983) – Theoretical results (2)

Suppose $n \to \infty$, $p \to \infty$ and $p/n \to \rho$ with $0 < \rho < 1$. Also, we assume that all the explanatory variables are orthonormal. After the second regression, we have

$$R_\alpha^2 \xrightarrow{pr} \rho g(\lambda),$$
$$F_\alpha \xrightarrow{pr} \frac{g(\lambda)(1-\alpha\rho)}{\alpha(1-g(\lambda)\rho)},$$
$$T_{\alpha,j} \xrightarrow{d} Z_\lambda \sqrt{\frac{1-\alpha\rho}{1-g(\lambda)\rho}},$$

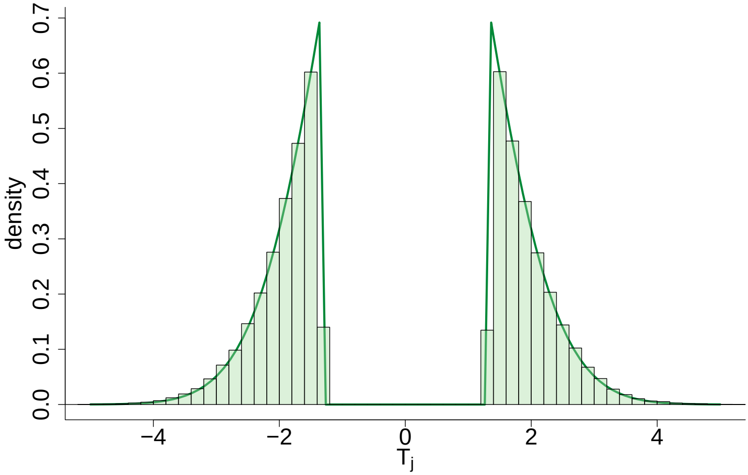
where $\Pr(|Z| > \lambda) = \alpha$, $g(\lambda) = \alpha + \sqrt{2/\pi}\lambda \exp(-\lambda^2/2)$, and $Z_\lambda \stackrel{d}{=} (Z \mid |Z| > \lambda)$. These result follow from considering q_α, the number of variables which are kept after the first regression,

$$q_\alpha/n \xrightarrow{pr} \alpha\rho$$

and studying the distribution of $\widehat{\beta}_j$ given that the first test was passed.

Freedman (1983) – The distribution of $T_{\alpha,j}$

The asymptotic $T_{\alpha,j}$ distribution along with a histogram of $T_{\alpha,j}$ values from the simulations (again with $n = 100$ and $p = 50$). The simulations fit well with the theory.



Paradox?

- Freedman himself did not describe his findings as paradoxical, but wrote "The existence of this effect is well known, but its magnitude may come as a surprise, even to a hardened statistician."
- Parts of the literature have later used the term paradox:
  - Raftery, Madigan and Hoeting (1993)
  - Anderson and Burnham (2002)
  - Lukacs, Burnham and Anderson (2009)

The surprise?

1. R^2 is high in the first regression.
2. After variable selection, F_α and many $T_{\alpha,j}$ can look highly significant.

R^2

The inflation of R^2 in the first regression is a typical case of overfitting: within a given sample, any pattern in y may be explained by a sufficiently large number of explanatory variables.

Solutions?

- R^2 adjusted: $R^2_{\text{adj}} = 1 - (1 - R^2)\frac{n}{n-p}$, we get $R^2_{\text{adj}} \xrightarrow{pr} 0$.
- Predicted R^2.
- Construct a confidence distribution for r^2, the population R^2 (see for instance Helland, 1987). In this setting, the confidence distribution will typically have a point-mass in 0.

Post-selection inference

2. After variable selection, F_α and many $T_{\alpha,j}$ can look highly significant.

The second part of Freedman's "paradox" is an illustration of the problems with post-selection inference, i.e. statistical inference after model selection.

Model selection methods are data-driven tools for choosing a model $\widehat{M}$ among several candidates. Variable selection based on p-values, like in Freedman, is a special case. There are a great number of different criteria and frameworks: AIC, BIC, FIC, Lasso, forward selection, backward elimination, ...

The estimators after model selection, $\widehat{\beta}_{\widehat{M}}$, will have unusual distributional properties. Similarly for the test statistics. The ordinary tests and confidence intervals are therefore no longer valid.

Post-selection inference

- Intuition:
  - The model should be specified before the data are analysed: "Using the data twice".
  - There is randomness in the choice of model, i.e. more uncertainty in the final inference.
  - We let data decide which questions to focus on, then proceed as if these were decided on beforehand.
- The "naive" use of ordinary inference methods after model selection is extremely common:
  - The practice is often taught in basic courses.
  - The phenomenon arises in all types of model selection, and in all kinds of models (not limited to regression!).

Why do scientists want to do model selection?

There are a number of reasons for why scientists use model selection methods. Typically, the need will depend on the purpose of the investigations and the extent of prior knowledge.

- To find a "good" model in a prediction setting.
  - Explain vs predict. The problems of post-selection inference are typically more acute in an explanatory setting (because predictions are "almost always" validated on test sets).

Why do scientists want to do model selection?

There are a number of reasons for why scientists use model selection methods. Typically, the need will depend on the purpose of the investigations and the extent of prior knowledge.

- To find the "true" model.
- To generate interesting hypotheses.
- To choose a between a set of equally likely models, differing in their secondary features.
- To obtain a smaller model.

Why do scientists want to do model selection?

There are a number of reasons for why scientists use model selection methods. Typically, the need will depend on the purpose of the investigations and the extent of prior knowledge.

- To find the "true" model. Exploratory
- To generate interesting hypotheses. Exploratory
- To choose a between a set of equally likely models, differing in their secondary features. Confirmatory
- To obtain a smaller model. Confirmatory

Solutions

- Ignore the problem.

Solutions

- Ignore the problem.
- Avoid model selection (in confirmatory analyses).
  - Simple, but not always possible.
  - Very difficult to avoid any kind of informal "model selection" in practice.

Solutions

- Ignore the problem.
- Avoid model selection (in confirmatory analyses).
  - Simple, but not always possible.
  - Very difficult to avoid any kind of informal “model selection” in practice.
- Do model averaging instead.
  - Advocated by for instance Raftery, Madigan and Hoeting (1993), and Lukacs, Burnham and Anderson (2009).
  - Model selection criteria are used to weight the candidate models and then constructs an estimator for the parameter of interested which is a weighted sum of estimators from the different models.
  - Some issues with interpretability.

Solutions

- Ignore the problem.
- Avoid model selection (in confirmatory analyses).
  - Simple, but not always possible.
  - Very difficult to avoid any kind of informal “model selection” in practice.
- Do model averaging instead.
  - Advocated by for instance Raftery, Madigan and Hoeting (1993), and Lukacs, Burnham and Anderson (2009).
  - Model selection criteria are used to weight the candidate models and then constructs an estimator for the parameter of interested which is a weighted sum of estimators from the different models.
  - Some issues with interpretability.
- Split the data: one part for model selection, one part for inference.

Solutions

- Ignore the problem.
- Avoid model selection (in confirmatory analyses).
  - Simple, but not always possible.
  - Very difficult to avoid any kind of informal "model selection" in practice.
- Do model averaging instead.
  - Advocated by for instance Raftery, Madigan and Hoeting (1993), and Lukacs, Burnham and Anderson (2009).
  - Model selection criteria are used to weight the candidate models and then constructs an estimator for the parameter of interested which is a weighted sum of estimators from the different models.
  - Some issues with interpretability.
- Split the data: one part for model selection, one part for inference.
- Attempt to correct for the model selection step.

Correcting for model selection

If we can understand the distributional properties of the post-selection estimators, we can hope to make corrected intervals and tests (i.e. which have the right coverage properties).

Simple example: Say we want to test H_0: $\beta_j = 0$ in the second regression. The "naive" p-value $p_j = \Pr(|T_{\alpha,j}| > |\hat{\beta}_j|/s_j)$ $\approx \Pr(|Z| > |\hat{\beta}_j|/s_j)$ will reject the null hypothesis far too often (as we have already seen). The following result from Freedman

$$T_{\alpha,j} \xrightarrow{d} Z_\lambda \sqrt{\frac{1-\alpha\rho}{1-g(\lambda)\rho}},$$

with $Z_\lambda \overset{d}{=} (Z \,|\, |Z| > \lambda)$, allows us to compute an adjusted p-value with the correct frequentist properties:

$$p_{\text{adj,f},j} \approx \Pr\left(|Z_\lambda| > |\hat{\beta}_j|/s_j \sqrt{\frac{1-g(\lambda)\rho}{1-\alpha\rho}}\right).$$

Correcting for model selection

There is a huge literature on similar attempts, but in much more complicated and general situations. Typically, these procedures have to be worked out separately for different model selection criteria and frameworks.

- See for example Kabaila (1998); Hjort & Claeskens (2003); Claeskens & Hjort (2008); Berk, Brown, Buja, Zhang & Zhao (2013); Efron (2014); Bachoc, Leeb & Pötscher (2015); Charki & Claeskens (2018).
- I will take a particular look at a specific framework, selective inference, see Lee and Taylor (2014); Taylor and Tibshirani (2015); Lee, Sun, Sun and Taylor (2016); Taylor, Lockhart, Tibshirani & Tibshirani (2016).

Selective inference

The whole idea relies on being able to express the model selection event (i.e. the event that $\widehat{M}$ was selected) as

$$\widehat{M} \Longleftrightarrow \{y \colon Ay \leq b\}$$

with some matrix A and a vector b, which will be different for different model selection procedures.
Then, the conditional distribution of $\widehat{\beta}$ given the model that was selected follows a certain truncated normal distribution

$$\widehat{\beta}_j \sim \mathrm{TN}^{\mathcal{V}_j^-, \mathcal{V}_j^+}(\beta_j, \sigma^2\{(X^{\mathrm{t}}X)^{-1}\}_{j,j})$$

with $\mathcal{V}_j^-, \mathcal{V}_j^+$ some functions of A, b and $\widehat{\beta}$. With this result, we can carry out valid hypothesis of H_0: $\beta_j = c$ and construct valid confidence intervals.
If σ is unknown, plug in $\widehat{\sigma}$.

Comparisons with Freedman

It turns out that the selective inference framework takes a particularly simple form in the setting with orthonormal X columns and variable selection based on $p_j < \alpha$. Then we have

$$A = \begin{bmatrix} -\mathrm{sgn}_S X_S^{\mathrm{t}} \\ \mathrm{sgn}_N X_N^{\mathrm{t}} \end{bmatrix} \qquad b = \begin{bmatrix} -t_{\alpha,n-p}\widehat{\sigma}_1/\sqrt{n}\ \mathbf{1}_S \\ t_{\alpha,n-p}\widehat{\sigma}_1/\sqrt{n}\ \mathbf{1}_N \end{bmatrix}$$

and we get

$$\mathcal{V}_j^- = t_{\alpha,n-p}\widehat{\sigma}_1/\sqrt{n}, \quad \mathcal{V}_j^+ = +\infty \quad \text{if } \widehat{\beta}_j > 0, \text{ and}$$
$$\mathcal{V}_j^- = -\infty, \quad \mathcal{V}_j^+ = -t_{\alpha,n-p}\widehat{\sigma}_1/\sqrt{n} \quad \text{if } \widehat{\beta}_j < 0.$$

This gives the following expression for the adjusted p-value for H_0: $\beta_j = 0$ (in the case of $\widehat{\beta}_j < 0$):

$$p_{\mathrm{adj,si},j} = \frac{\Phi(\widehat{\beta}_j/(\widehat{\sigma}_1/\sqrt{n}))}{\Phi(-t_{\alpha,n-p})},$$

which we can compare with the one from Freedman:

$$p_{\mathrm{adj,f},j} = \frac{\Phi\left(\widehat{\beta}_j/(\widehat{\sigma}_2/\sqrt{n})\sqrt{\frac{1-g(\lambda)\rho}{1-\alpha\rho}}\right)}{\alpha/2}.$$

Conclusions

- We have learnt
  - to be careful R^2 when p is of the same order as n;
  - to be careful with inference after model selection.
- The problems associated with post-selection inference can be amended in various ways, but the most important message is know what you are doing.
  - Know what the goal of the analysis is.
  - Know what hypotheses you are trying to confirm (if any).
  - Know the assumptions you are making.

It is easy to lie with statistics, but a whole lot easier without them.
(Fred Mosteller)

Some more references

- Berk, Brown, Buja, Zhang & Zhao. Valid Post-Selection Inference. The Annals of Statistics (2013).
- Berk, Brown & Zhao. Statistical Inference After Model Selection. Journal of Quantitative Criminology (2010).
- Charkhi & Claeskens. Asymptotic Post-Selection Inference for the Akaike Information Criterion. Biometrika (2018).
- Freedman. A Note on Screening Regression Equations. The American Statistician (1983).
- Freedman, L. & Pee. Return to a Note on Screening Regression Equations. The American Statistician (1989).
- Helland. On the Interpretation and Use of R^2 in Regression Analysis. Biometrics (1987).
- Holmes. Statistical Proof? The Problem of Irreproducibility. Bulletin of the American Mathematical Society (2017).
- Lee & Taylor. Exact Post Model Selection Inference for Marginal Screening. In Advances in Neural Information Processing Systems (2014).
- Leeb, Pötscher & Ewald. On Various Confidence Intervals Post-Model-Selection. Statistical Science (2015).
- Liu, Markovic & Tibshirani. More Powerful Post-Selection Inference, with Application to the Lasso. ArXiv (2018).
- Taylor & Tibshirani. Statistical Learning and Selective Inference. Proceedings of the National Academy of Sciences (2015).