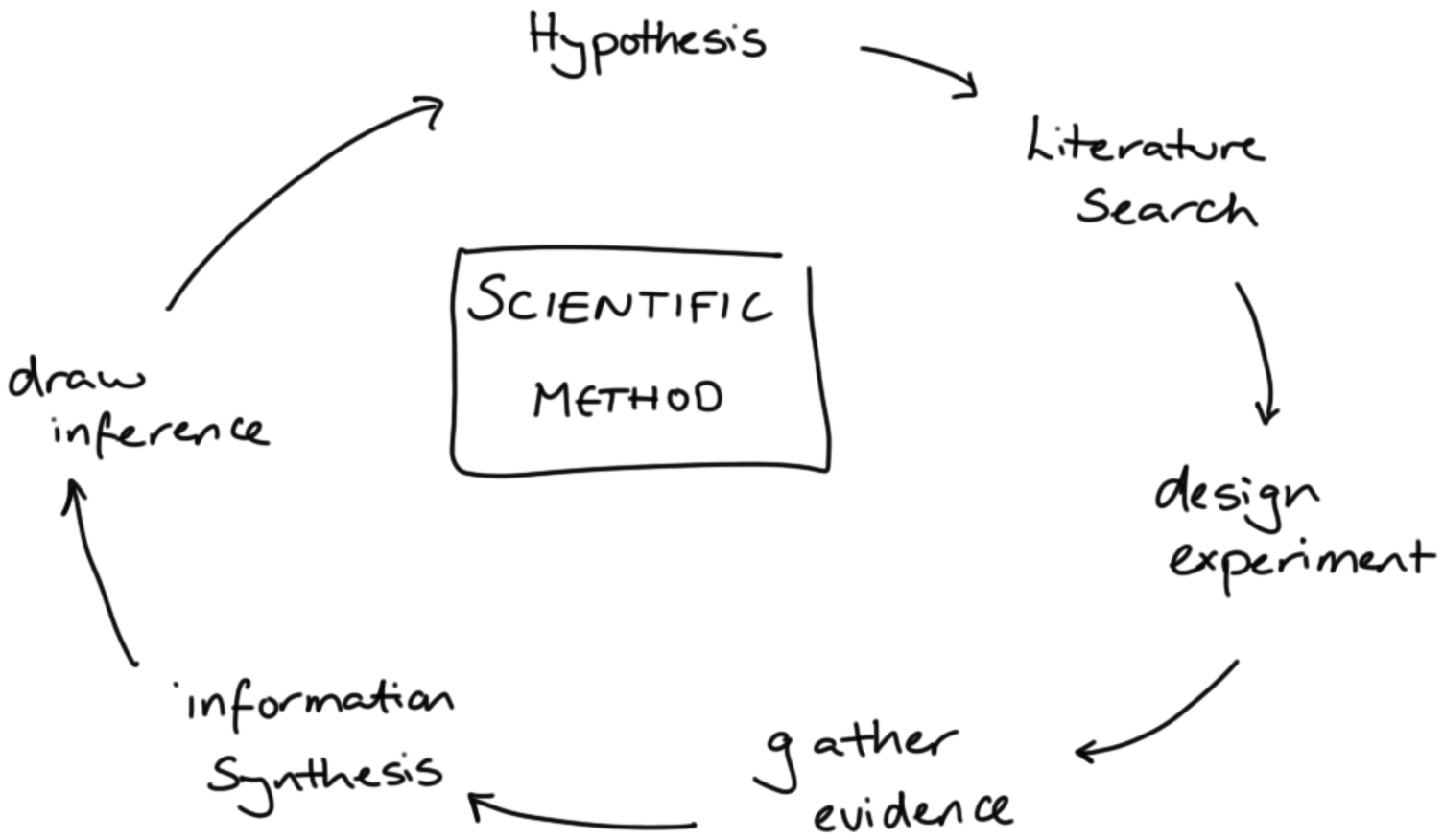
Reproduction and Replication

Preparing for Production

Dr Zak Varty

Putting the Science in Data Science

The Scientific Method



Issue: Multiple, Dependent Tests

- Projects are usually not a single hypothesis test
- Sequence of dependent decisions
- e.g. Model development
- Can fool ourselves by looking lots of times or ignoring sequential and dependent structure.

The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the research hypothesis was posited ahead of time*

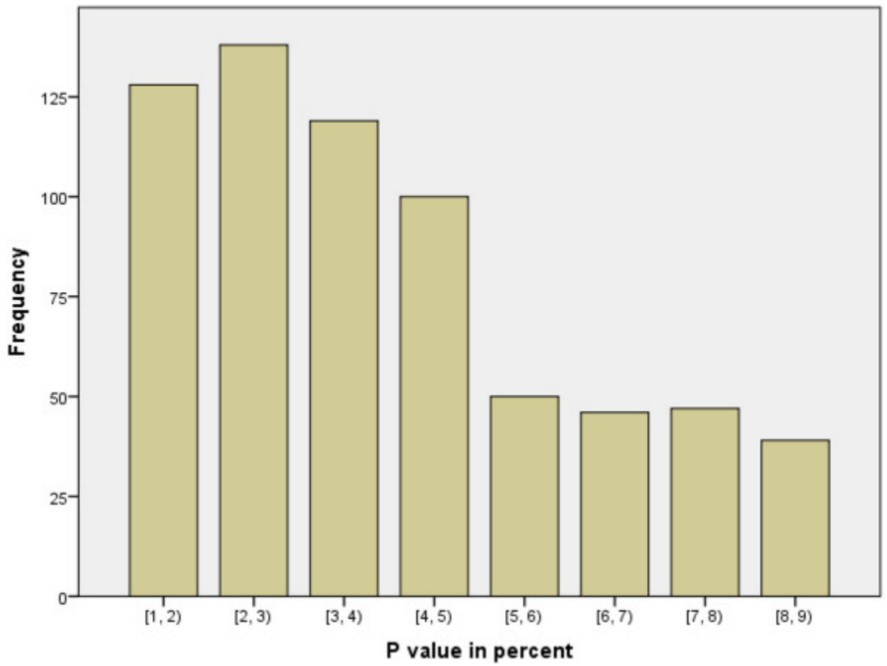
Andrew Gelman† and Eric Loken‡
14 Nov 2013

"I thought of a labyrinth of labyrinths, of one sinuous spreading labyrinth that would encompass the past and the future . . . I felt myself to be, for an unknown period of time, an abstract perceiver of the world." — Borges (1941)

Abstract

Researcher degrees of freedom can lead to a multiple comparisons problem, even in settings where researchers perform only a single analysis on their data. The problem is there can be a large number of *potential* comparisons when the details of data analysis are highly contingent on data, without the researcher having to perform any conscious procedure of fishing or examining multiple p-values. We discuss in the context of several examples of published papers where data-analysis decisions were theoretically-motivated based on previous literature, but where the details of data selection and analysis were not pre-specified and, as a result, were contingent on data.

Issues: p-hacking and Publiction Bias



Distribution of p-values in medical publications. From Perneger and Combescure (2017)

A stack of unpublished work.

Image credit: Sear Greyson

Reproducucibility, Replicability and Going into Production

Reproducibility

> **Reproducibility:** given the original raw data and code, can you get all of the results again?

- Reproducible != Correct
- “Code available on request” is the new “Data available on request”
- Reproducible data analysis requires effort, time and skill.

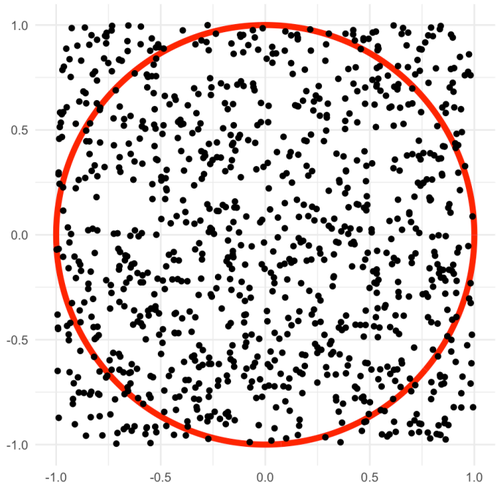
Replicability

> **Replicable:** if the experiment were repeated by an independent investigator, you would get slightly different data but would the substative conclusions be the same?

- In the specific sense, this is the core worry for a statistician!
- Also used more generally: are results stable to perturbations in population / study design / modelling / analysis?
- Only real test is to try it. Control risk with shadow and parallel deployment.

Reproduction, Replication and Statistical Data Science

Monte Carlo Methods



Monte Carlo methods are used extensively in data science.

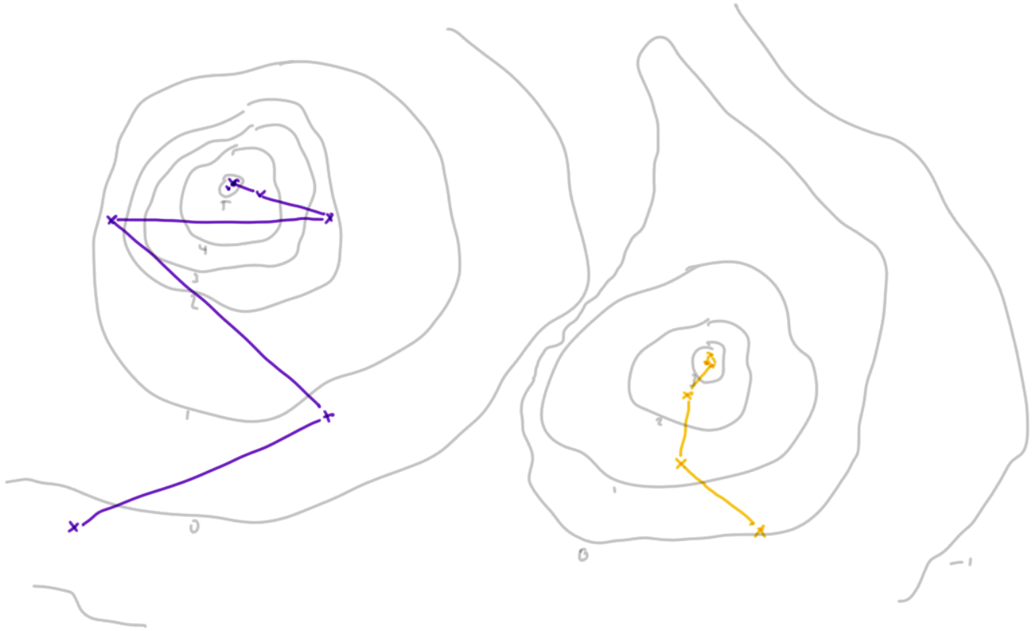
Usually solve a difficult problem (integration) or help to ensure results replicate (partitioning, sampling variability, posterior samples).

Almost always make reproduction of results more difficult. (seeding and LLN)

Optimisation

Is the optimum you find stable over:

- realisations?
- starting points?
- step size / learning rate?
- realisations of the data?



A poorly drawn contour plot. Local modes make this optimiation unstable to the choice of starting point.

Pseudo-random Number Generators

- Computers are deterministic, randomness is hard.
- **Pesudo**-random number generation.
- Set starting point with `set.seed()`.
- Beware: parallel programming and language interfacing.

```r
# different values
rnorm(n = 4)
```

```
[1] -1.1546783 -0.6895971 -1.3753532 -0.3621124
```

```r
rnorm(n = 4)
```

```
[1] -0.3592553 -0.9052359  0.2669273 -0.8162782
```

```r
# the same value
set.seed(1234)
rnorm(n = 4)
```

```
[1] -1.2070657  0.2774292  1.0844412 -2.3456977
```

```r
set.seed(1234)
rnorm(n = 4)
```

```
[1] -1.2070657  0.2774292  1.0844412 -2.3456977
```

Wrapping up

- **Reproducible:** can recreate the same results from the same code and data
- **Replicable:** core results remain valid when using different data

- Stochasticity causes problems: make use of LLN and set.seed()
- Be very careful with you need to be both efficient and replicable.