Orléans Graduate Schools, Ethics Day 2020

How to get into trouble with statistics

Nils Berglund

Institut Denis Poisson, University of Orléans, France

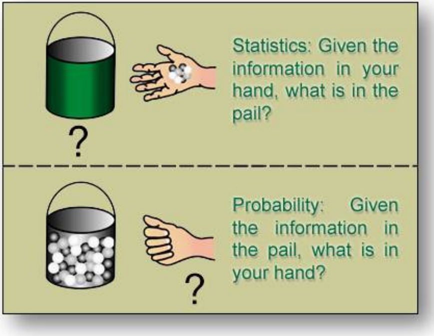
INSTITUT ORLÉANS • TOURS Mathématiques & Physique Théorique DENIS POISSON

October 2020

Disclaimer

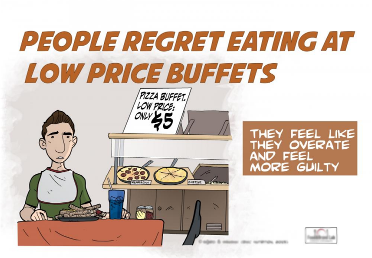
I am probabilist, not statistician

I thus understand the mathematical background of (simple) statistical methods. However, I will not be able to help you with your statistical analyses.



The case of Prof B.[1]

- Professor of Marketing in a prestigious American University
- Studies on nutrition, for example
  - People eat more soup from a "bottomless bowl"
  - When given a choice between a cookie and an apple for desert at lunch, children at an elementary school were more like to pick the apple when it had a sticker of Elmo (from Sesame Street)



- Many articles, highly quoted, policy making in some cases

[1] This is a true story. Out of respect of the victims, the names have been changed. Out of respect of Statistics, the rest will be told exactly as it occurred.

Trouble ahead...

- Blog entry in 2016 comparing a postdoc and a PhD student:
  - The foreign (unpaid) visiting PhD student was given a data set from a failed study. The student tried all kinds of statistical tests, and finally managed to extract significant results, and write 3 paper.
  - At the same time, the (paid) postdoc was given another data set. The postdoc declined to spend time with it.
- Statisticians objected and started analysing several of B.'s papers. They found inconsistencies that B. could not explain.
- At least 14 retractions, and many more corrections. B. resigned from his post in 2018.

Statistical re-analysis of one of B.'s papers

Abstract

We present the initial results of a reanalysis of four articles from the XXX Lab based on data collected from diners at an Italian restaurant buffet. On a first glance at these articles, we immediately noticed a number of apparent inconsistencies in the summary statistics. A thorough reading of the articles and careful reanalysis of the results revealed additional problems. The sample sizes for the number of diners in each condition are incongruous both within and between the four articles. In some cases, the degrees of freedom of between-participant test statistics are larger than the sample size, which is impossible. Many of the computed F and t statistics are inconsistent with the reported means and standard deviations. In some cases, the number of possible inconsistencies for a single statistic was such that we were unable to determine which of the components of that statistic were incorrect [...] The attached Appendix reports approximately 150 inconsistencies in these four articles, which we were able to identify from the reported statistics alone [...]

What were B.'s deadly sins?

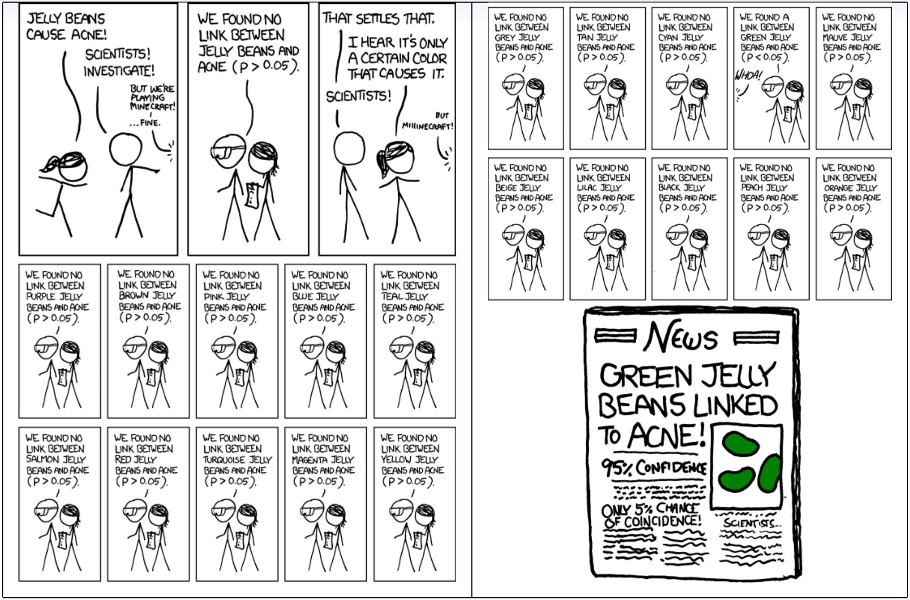
"If you don't reveal some insights soon, I'm going to be forced to slice, dice, and drill!"

Source: atozmarkets

Source: georgiapoliticalreview

- ▷ *p*-**hacking**
- ▷ **HARKing** (**H**ypothesizing **A**fter **R**esults are **K**nown)

p-hacking and HARKing: a first explanation



Source: `https://xkcd.com/882/`

Confirmation bias

Clever Hans, the horse that was able to perform arithmetic

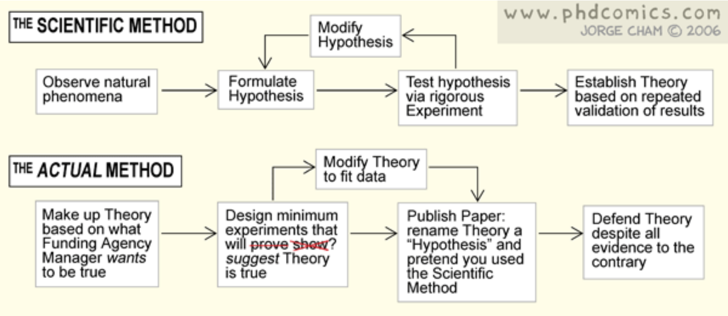
Wilhelm von Osten and Clever Hans

"After a formal investigation in 1907, psychologist Oskar Pfungst demonstrated that the horse was not actually performing these mental tasks, but was watching the reactions of his trainer."

https://en.wikipedia.org/wiki/Clever_Hans

How to avoid confirmation bias

- Formulate a hypothesis
- Design an experiment
- Perform the experiment
- Determine whether the results are (likely to be) compatible with the hypothesis



Hypothesis testing

Example: does hydroxychloroquine (HCQ) help curing Covid19?

Form two groups of patients:

- ▷ One group gets HCQ
- ▷ Control group gets placebo ("sugar pills")

	Cured	Not cured	TOTAL
HCQ group	45	15	60
Control group	25	15	40
TOTAL	70	30	100

- ▷ 75% of HCQ group are cured
- ▷ 62.5% of control group are cured

Hypothesis testing

Null hypothesis H_0: Taking HCQ and being cured are independent

Expected number of cured HCQ patients under H_0:

$$(\% \text{ of HCQ}) \cdot (\% \text{ of Cured}) \cdot 100 = \frac{60}{100} \cdot \frac{70}{100} \cdot 100 = \frac{60 \cdot 70}{100} = 42$$

Theoretical table under H_0:

Under H_0	Cured	Not cured	TOTAL
HCQ group	42	18	60
Control group	28	12	40
TOTAL	70	30	100

- ▷ 70% of HCQ group are cured
- ▷ 70% of control group are cured

Hypothesis testing

Actual	Cured	Not cured	TOTAL
HCQ group	45	15	60
Control group	25	15	40
TOTAL	70	30	100

Under H_0	Cured	Not cured	TOTAL
HCQ group	42	18	60
Control group	28	12	40
TOTAL	70	30	100

Chi-square distance: $d^2_{\chi^2} = \dfrac{(45-42)^2}{42} + \dfrac{(15-18)^2}{18} + \cdots = 1.7857$

Theorem [Pearson]:

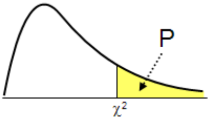
Under H_0, $d^2_{\chi^2}$ follows (approx) a chi-squared law with 1 degree of freedom.

Hypothesis testing

Chi-squared test, "classical" version:

- Fix a level of significance α, say $\alpha = 0.05$
  (α is the probability of getting a false positive, i.e. to wrongly reject H_0 if it is true, also called type I error)
- Look up in a table: $\mathbb{P}\{\chi_1^2 > x\} = 0.05 \Rightarrow x = 3.841$
- Since $d_{\chi^2}^2 = 1.7857 < x$, we cannot reject null hypothesis H_0:
  Can't rule out that being cured is independent of having taken HCQ

Chi-squared test, "modern" version:



- Fix a level of significance α, say $\alpha = 0.05$.
- Compute p-value: $\mathbb{P}\{\chi_1^2 > 1.7857\} = 0.1814$.
- Since p-value is larger than $\alpha = 0.05$, we cannot reject null hypothesis H_0

Euphemisms for lack of significance

The page
https://mchankins.wordpress.com/2013/04/21/still-not-significant-2/
contains a list of over 400 euphemisms for failed tests, such as

- a certain trend toward significance ($p = 0.08$)
- a margin at the edge of significance ($p = 0.0608$)
- a moderate trend toward significance ($p = 0.068$)
- a nonsignificant trend toward significance ($p = 0.1$)
- almost attained significance ($p < 0.06$)
- an apparent trend ($p = 0.286$)
- an evident trend ($p = 0.13$)
- approached acceptable levels of statistical significance ($p = 0.054$)
- arguably significant ($p = 0.07$)
- at the verge of significance ($p = 0.058$)
- ...

p-hacking and HARKing

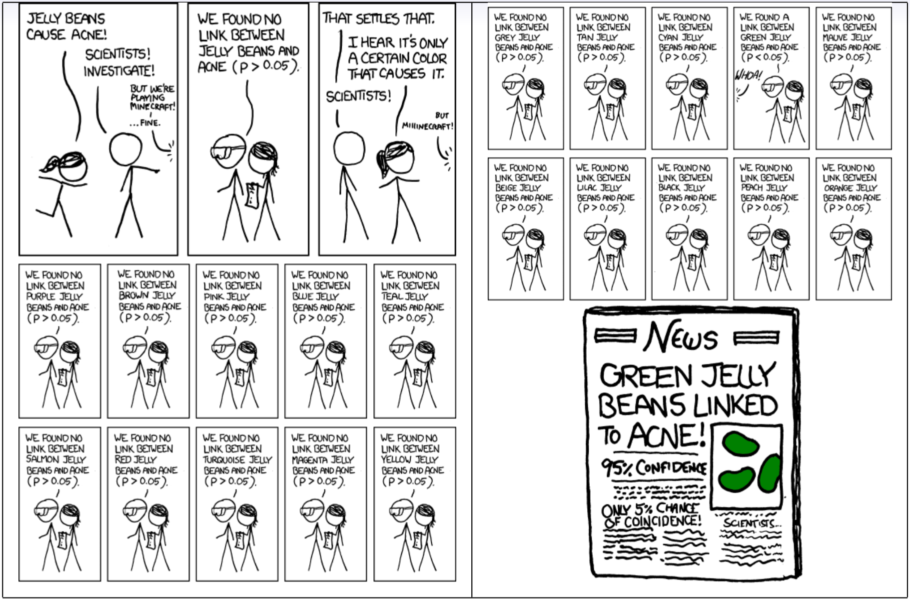
From http://en.wikipedia.org/wiki/Data_dredging:

Data dredging (also data fishing, data snooping, data butchery, and p-hacking) is the misuse of data analysis to find patterns in data that can be presented as statistically significant, thus dramatically increasing and understating the risk of false positives. [It] involves testing multiple hypotheses using a single data set by exhaustively searching – perhaps for combinations of variables that might show a correlation [...]

From https://en.wikipedia.org/wiki/HARKing:

The term HARKing [...] refers to the questionable research practice of Hypothesizing After the Results are Known. Kerr (1998) defined HARKing as "presenting a post hoc hypothesis in the introduction of a research report as if it were an a priori hypothesis". HARKing may also occur when a researcher tests an a priori hypothesis but then omits that hypothesis from their research report after they find out the results of their test.

A second look at XKCD



Analysis: what is the mistake?

- Null hypothesis H_0: having acne is independent of eating jelly beans
- First experiment: Observed chi squared distance d
  $\mathbb{P}\{\chi^2 > d^2 | H_0 \text{ is true}\} > 0.05$
  We cannot reject H_0
- 20 experiments: $x = 3.841$ such that
  $\mathbb{P}\{\text{one false positive}\} = \mathbb{P}\{d^2_{\text{observed}} > x | H_0 \text{ is true}\} = 0.05$

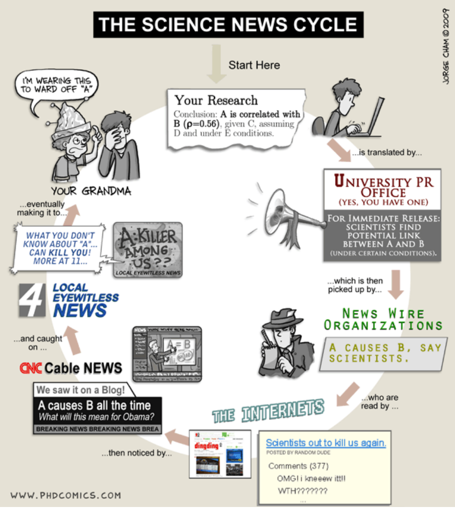
  $\mathbb{P}\{\text{no false positive in 20 tests}\} = (1 - 0.05)^{20} = (0.95)^{20}$

  $\mathbb{P}\{\text{at least one false positive in 20 tests}\} = 1 - (0.95)^{20} = 0.64$
  The significance level has changed from 0.05 to 0.64!
- Possible cure: change α such that $(1-\alpha)^{20} = 0.95 \quad \Rightarrow \quad \alpha = 0.00256$
  i.e. p-value of at least one experiment must be smaller than 0.00256
  (reject H_0 only if $d^2_{\text{observed}} > 9.14$ instead of $d^2_{\text{observed}} > 3.841$)

  Remark: this is very close to dividing α by 20 (Bonferroni correction)

Conclusions

- It is okay to do surveys, experiments . . . first, and then to formulate hypotheses based on results (this is what we do)
- What is dangerous is to reuse a dataset from a single experiment, making statistical tests until one finds something significant
- Also, beware of attention from the media. . .



http://phdcomics.com/comics.php?n=1174

Further reading at
https://simplifaster.com/articles/p-hacking-harking-scientific-replication/