Breaking Bad: Perverse Incentives and Inappropriate Statistics are Ruining Science

Tomáš Fürst and Halina Šimková
(and Honza Strojil and Ondra Vencálek, too)

4B|N: Center for Bayesian Inference
www.4bin.org

October 2019

Understanding Nature

- Three ways to know
  1. Speculative reasoning (Karl Marx)
  2. Expert knowledge (George Soros)
  3. Reasoning from empirical evidence (Michael Bloomberg)
- Reformation
- Archie Cochrane revolution of EBM
- Industry 4.0
- Science?

The Scientific Method

- From observation (or otherwise), propose a Theory

  Theory: Men are heavier than women

- Pose the Null Hypothesis in a testable way

  H_0 : Men and women have the same average mass

- Perform an Experiment to Falsify the Hypothesis

  Round up hundred men and hundred women and weigh them

- Compute the P-value of the t-test, it will be small, e.g. $p \sim 10^{-6}$

  $p = prob(\text{I observe what I observe (or worse!)} | H_0 \text{ is true})$

The Scientific Method

- If the p-value is small, it casts doubt on the null hypothesis
- If $p < 0.05$, I reject the null hypothesis
  (and stick to the Theory)
- If $p > 0.05$, I do not reject the null hypothesis.

The Scientific Method

- If the p-value is small, it casts doubt on the null hypothesis
- If $p < 0.05$, I reject the null hypothesis
  (and stick to the Theory)
- If $p > 0.05$, I do not reject the null hypothesis.

In our case . . .

. . . $p \sim 10^{-6}$ and therefore men are heavier than women.

The Scientific Method

The Scientific Method Revisited

- From observation (or otherwise), propose a Theory

  Theory: Little redhead girls have magical powers

- Pose the Null Hypothesis in a testable way

  H_0 : LRG do not have magical powers

- Perform an Experiment to Falsify the Hypothesis

  Get 5 LRGs to roll a dice: All of them get a six

- Compute the P-value of a test, it will be small, e.g. $p \sim 0.0003$

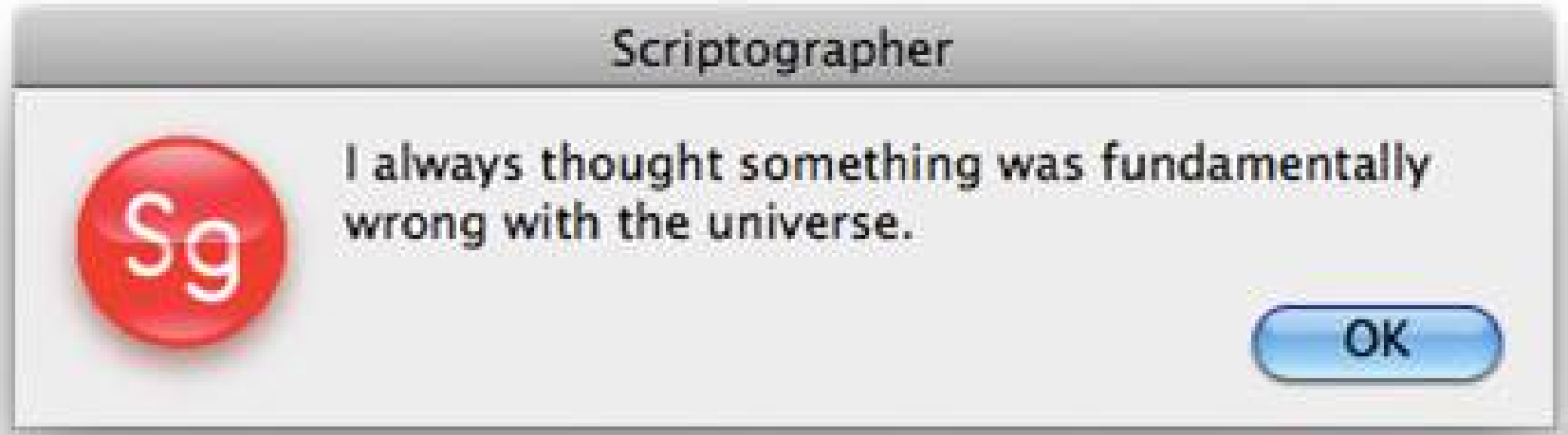
  $p = prob(\text{I observe what I observe (or worse!)} | H_0 \text{ is true})$

The curse of p-values

AND THEREFORE

Little redhead girls have magical powers!

The curse of p-values



The curse of p-values

- The p-value i really small, e.g. $p \sim 10^{-6}$

$$p = prob(\text{I observe what I observe (or worse!)} | H_0 \text{ is true})$$

- A small p-value means that the data are improbable given the hypothesis is true
- That tells me that the probability of the hypothesis being true increases e.g. from 10^{-18} to 10^{-17}.

HOWEVER

Although the p-value is really small, it is still much more probable that the theory is **wrong**. That is, the hypothesis should not be rejected!

Interludium: All statistical tests are equal

- Under the assumptions of the null hypothesis ...

  e.g. two samples of equal size and equal variance

- ...some function (not a function of the data though!) ...

$$t = \frac{\sqrt{N}(\overline{X_1} - \overline{X_2})}{\sqrt{2}S_p}$$

- ...has (asymptotically) a known distribution.

  $t \sim$ Student's t-distribution with $2N - 2$ degrees of freedom

- Compute the value of this function for the observed data ...
- ... and assess how likely it is, under the null hypothesis, that the data is at least this extreme.

Interludium: All statistical tests are equal

. . . , t-test, F-test, chi-square test, Wilcoxon, Mann-Whitney, Kolmogoroff-Smirnoff, Fisher, ANOVA, Kruskal-Wallis, Friedman, Cox PH, log-rank, . . .

Interludium: All statistical tests are equal

. . . , t-test, F-test, chi-square test, Wilcoxon, Mann-Whitney, Kolmogoroff-Smirnoff, Fisher, ANOVA, Kruskal-Wallis, Friedman, Cox PH, log-rank, . . .

. . . , M-1, M-2, A bombs, H bombs, P bombs, Q bombs, Chinese checks, Hindus, Bindus, Italianos, Polacks Germans, Youse, Jews, Ups and downs, Vietnam, Johnson, high school, sex, coffee, books, food, scissors, magazines, news, cigarettes, . . .

Interludium: All statistical tests are equal

. . . , t-test, F-test, chi-square test, Wilcoxon, Mann-Whitney, Kolmogoroff-Smirnoff, Fisher, ANOVA, Kruskal-Wallis, Friedman, Cox PH, log-rank, . . .

To a single well-posed question . . .

classical statistician offers to you 20 different tests and 4 criteria which one is better.

Something's wrong

- Amgen, CA (biotech firm) tried to replicate 53 landmark studies in cancer treatment. In 6 cases they obtained confirming results.

  Begley: Drug development: Raise standards for preclinical cancer research, Nature 2012

- Bayer HealthCare tried to replicate 67 landmark studies. About 80 % failed.

  Prinz: Believe it or not: how much can we rely on published data on potential drug targets?, Nature Drug Discovery, 2011

- More than 100 potential drugs for ALS tested in an established mouse model. Many of these drugs had been reported to slow down disease in that same mouse model. None was found to be beneficial in our experiments.

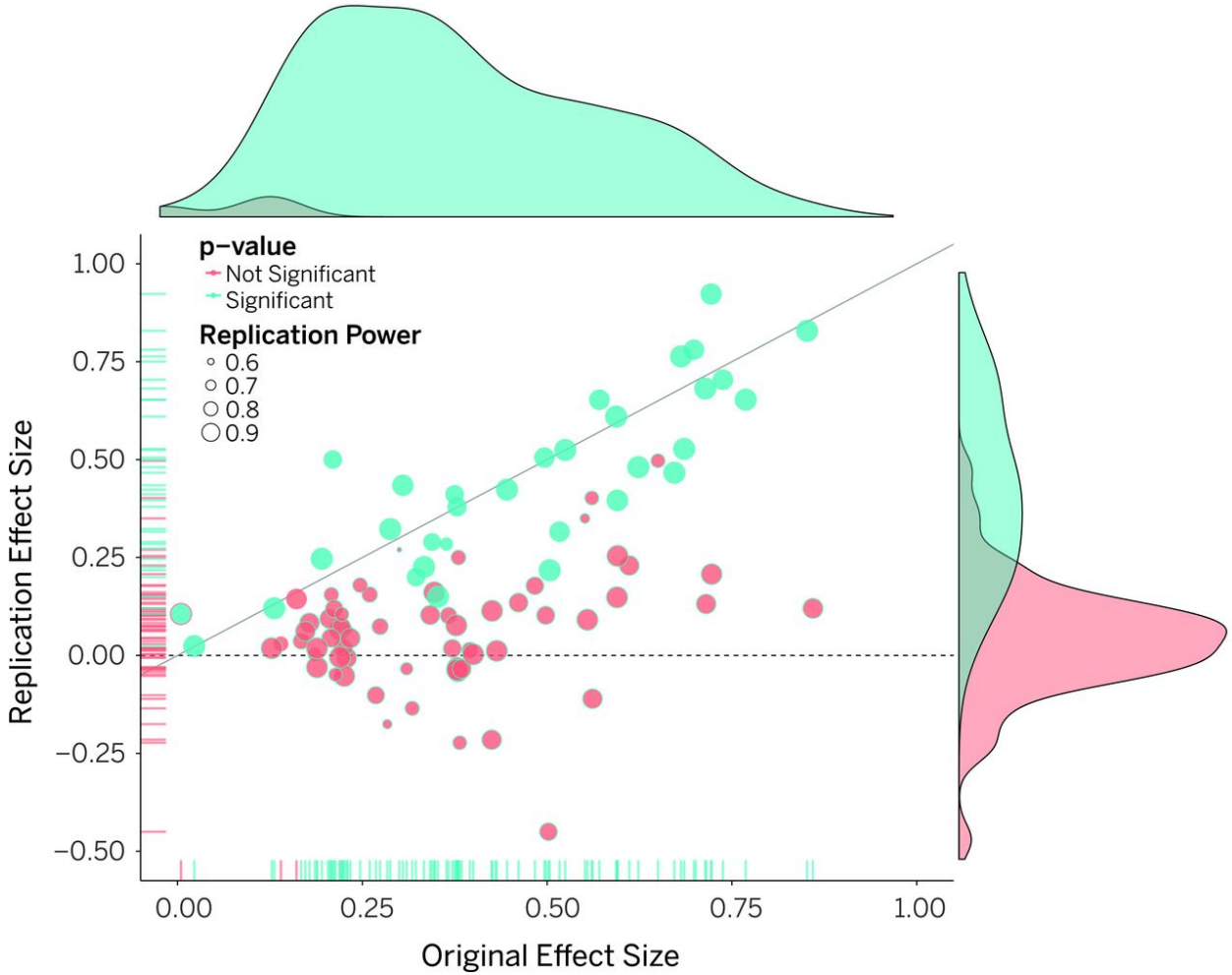
  Perrin: Preclinical research: Make mouse studies work, Nature, 2014

Evidence That Raises Concern

- fMRI is 25 years old, more than 60 000 papers based on it
- Resting-state fMRI data from 499 healthy controls to conduct 3 million task group analyses.
- In theory, we should find 5 % false positives.
- Instead we found that the most common software packages have false-positive rates of up to 70 %.

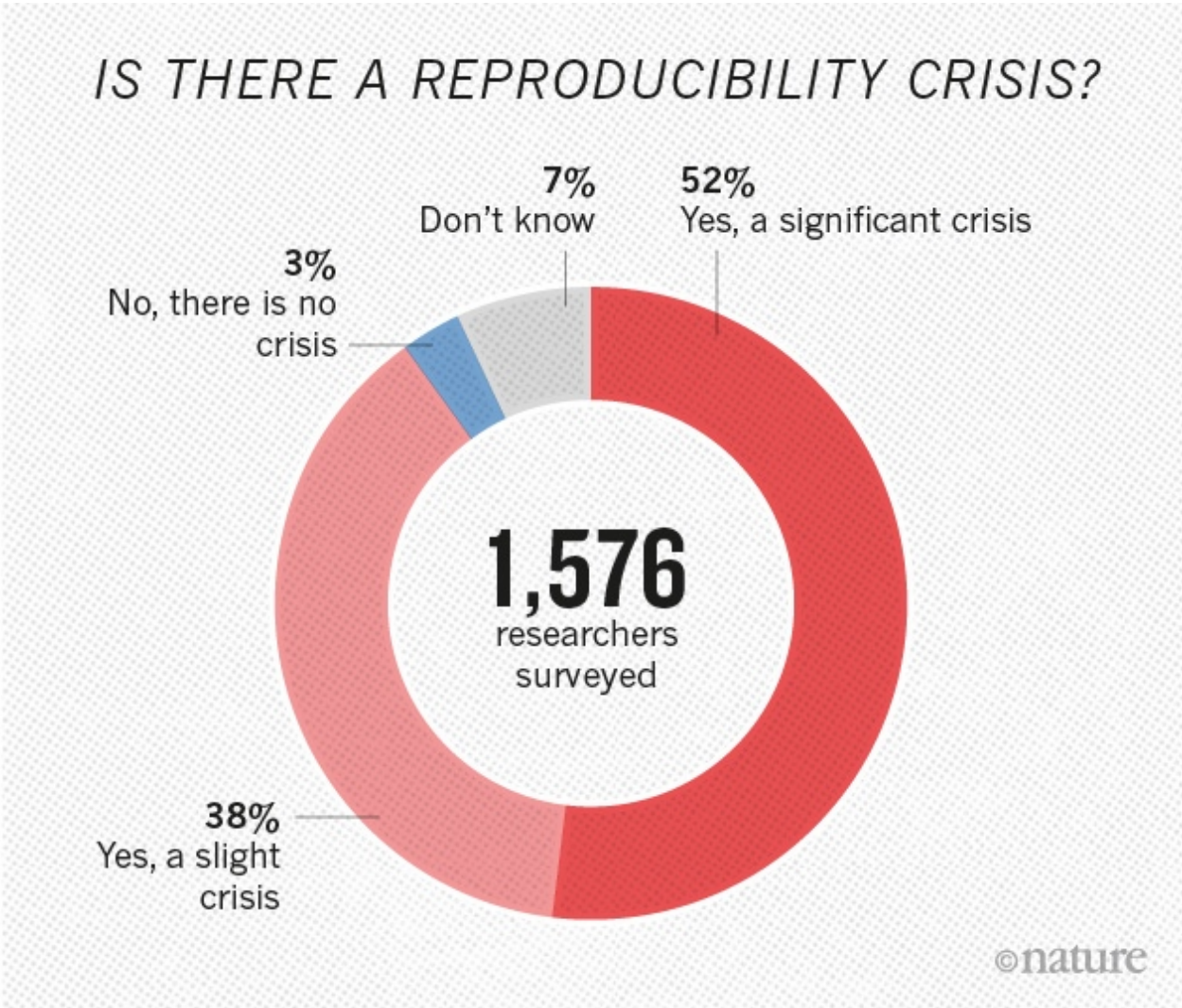
  Eklund: Cluster failure: Why fMRI inferences for spatial extent have inflated false-positive rates, PNAS, 2016

Open Science Project



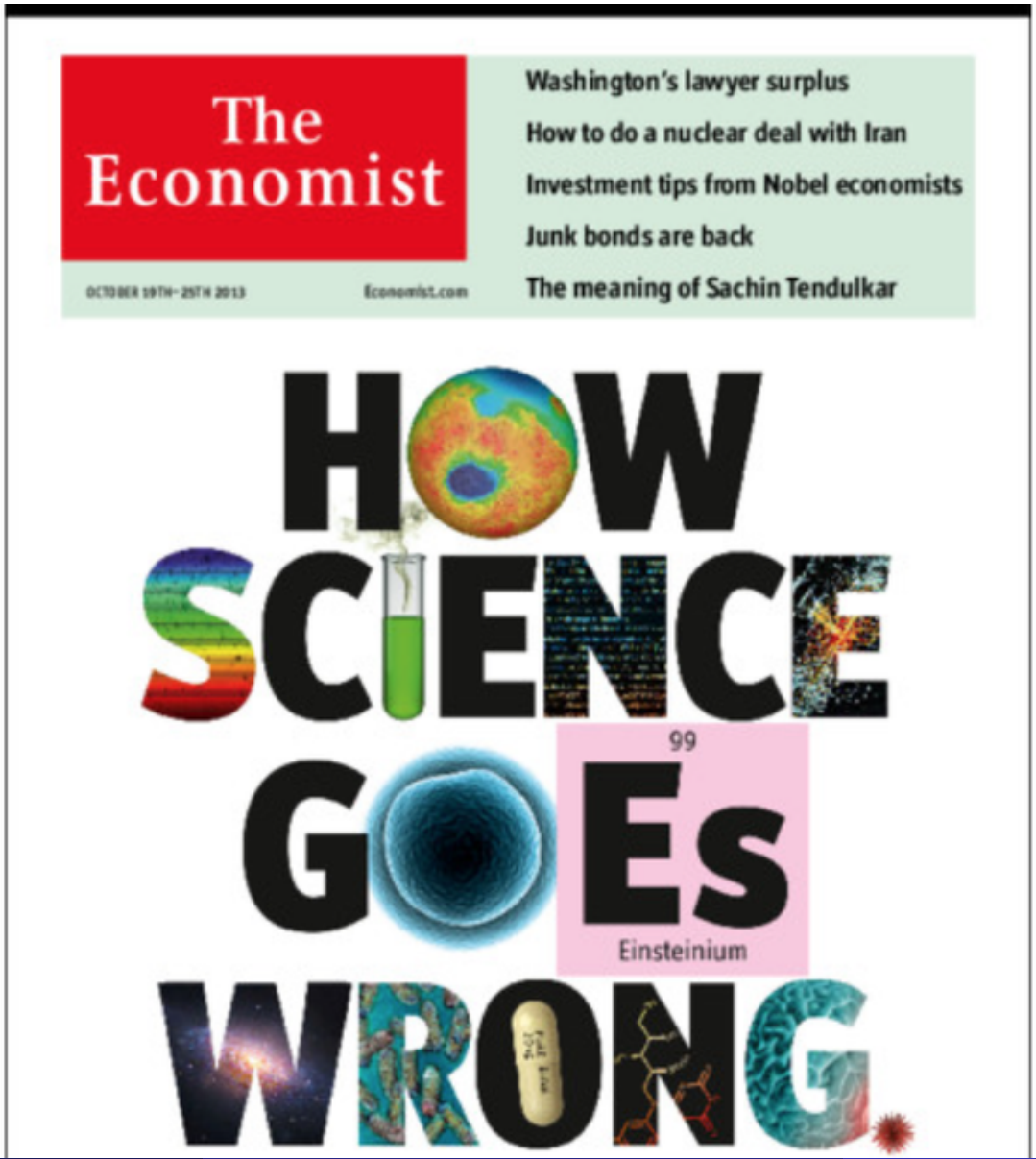
Science: Estimating the reproducibility of psychological science

The most influential research journal



Nature: 1,500 scientists lift the lid on reproducibility, May 2016

The most influential journal



Wrong Incentives in Science

1. Career depending on number of articles published.
2. Public finance (grants, scholarships, institutional money, ...) available according to the number of articles published.
3. Peer review system failure.
4. Reward for getting things wrong, cost of not getting them published.
5. The pressure to publish original results (the more striking the better) rather than repeat experiments
6. Preference for positive findings over negative ones.

Wrong Incentives in Science

There are ...

... more than enough wrong incentives in Science that weren't in place some 40 years ago.

Wrong Incentives in Science

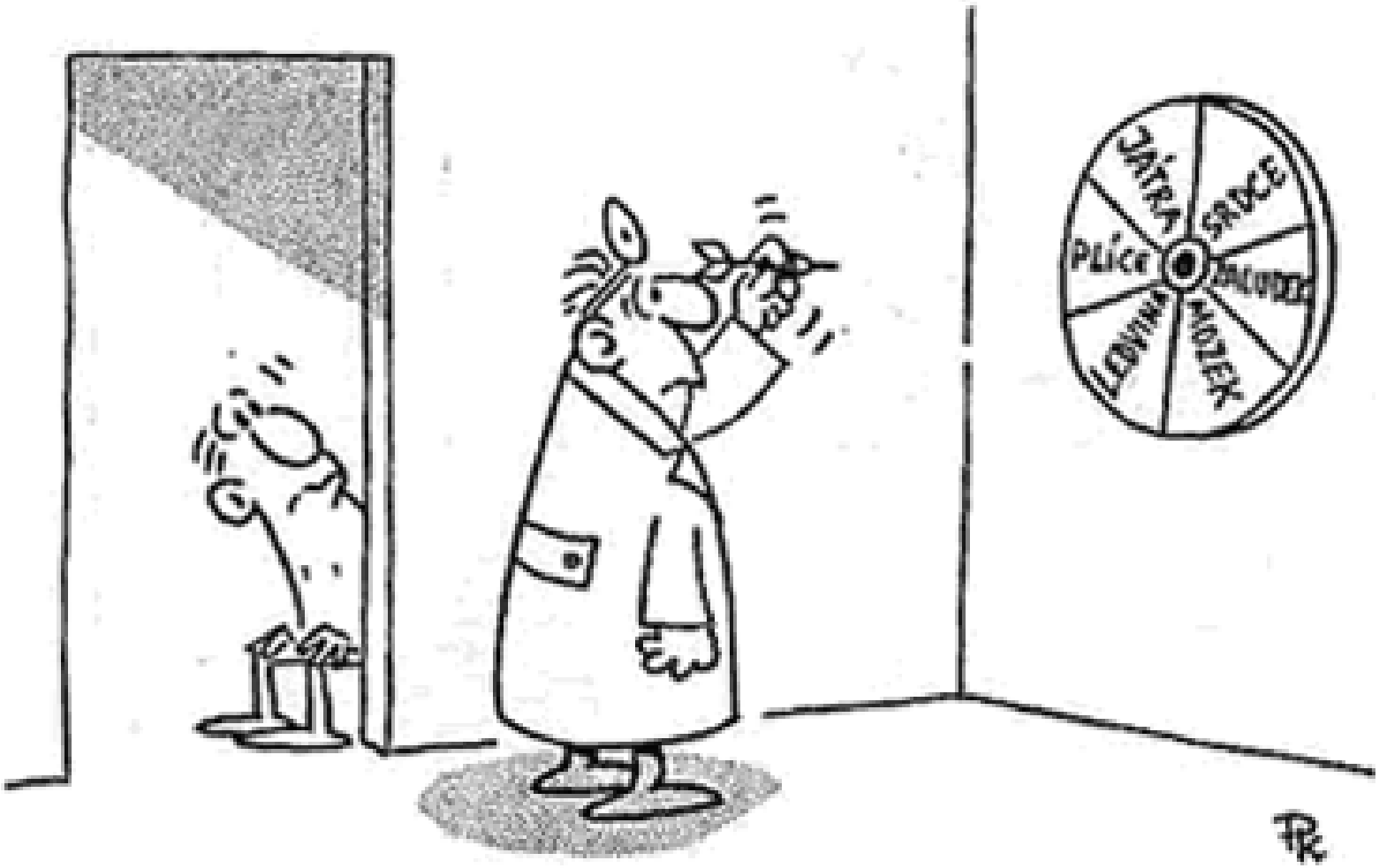
There are ...

... more than enough wrong incentives in Science that weren't in place some 40 years ago.

Moreover ...

... these wrong incentives have found a uniquely unfortunate companion in the old statistical mindset of hypothesis testing.

Pavel Kantorek



Towards correct, i.e. Bayesian inference

- Bayes' Theorem: The single most efficient crutch of the common sense
- Test T for a disease D
- Sensitivity: $Sens = p(T+|D+)$, assume $Sens = 0.99$
  Specificity: $Spec = p(T-|D-)$, assume $Spec = 0.99$
- Prevalence

$$Prev = p(\text{randomly selected person has disease } D)$$

Assume $Prev = 0.001$

Towards correct, i.e. Bayesian inference

- Bayes' Theorem: The single most efficient crutch of the common sense
- Test T for a disease D
- Sensitivity: $Sens = p(T+|D+)$, assume $Sens = 0.99$
  Specificity: $Spec = p(T-|D-)$, assume $Spec = 0.99$
- Prevalence

$$Prev = p(\text{randomly selected person has disease } D)$$

Assume $Prev = 0.001$

The Question

In a randomly selected person, the test comes out positive. What is the probability that he/she really has the disease?

Towards correct, i.e. Bayesian inference

The Question

In a randomly selected person, the test comes out positive.
What is the probability that he/she really has the disease?

Warning

Use the hypothesis testing mindset and face the consequences!

Path to the Bayes' Rule

1. Conditional probability

$$p(A|B) = \frac{p(A \text{ and } B)}{p(B)}$$

so that

$$p(A \text{ and } B) = p(A|B)p(B) = p(B|A)p(A)$$

Path to the Bayes' Rule

1. Conditional probability

$$p(A|B) = \frac{p(A \text{ and } B)}{p(B)}$$

so that

$$p(A \text{ and } B) = p(A|B)p(B) = p(B|A)p(A)$$

2. Common sense observation

$$p(B) = p(B|A)p(A) + p(B|nonA)p(nonA)$$

Path to the Bayes' Rule

1. Conditional probability

$$p(A|B) = \frac{p(A \text{ and } B)}{p(B)}$$

so that

$$p(A \text{ and } B) = p(A|B)p(B) = p(B|A)p(A)$$

2. Common sense observation

$$p(B) = p(B|A)p(A) + p(B|nonA)p(nonA)$$

3. Put it together

$$p(A|B) = \frac{p(A \text{ and } B)}{p(B)} = \frac{p(B|A)p(A)}{p(B)}$$

$$= \frac{p(B|A)p(A)}{p(B|A)p(A) + p(B|nonA)p(nonA)}$$

The Bayes Rule in Action

The Bayes Rule

$$p(A|B) = \frac{p(B|A)p(A)}{p(B|A)p(A) + p(B|nonA)p(nonA)}$$

The Bayes Rule in Action

The Bayes Rule

$$p(A|B) = \frac{p(B|A)p(A)}{p(B|A)p(A) + p(B|nonA)p(nonA)}$$

In the context of the disease D and the test T :

$$p(D+|T+) = \frac{p(T+|D+)p(D+)}{p(T+|D+)p(D+) + p(T+|D-)p(D-)}$$

The Bayes Rule in Action

The Bayes Rule

$$p(A|B) = \frac{p(B|A)p(A)}{p(B|A)p(A) + p(B|nonA)p(nonA)}$$

In the context of the disease D and the test T :

$$\begin{aligned} p(D+|T+) &= \frac{p(T+|D+)p(D+)}{p(T+|D+)p(D+) + p(T+|D-)p(D-)} \\ &= \frac{Sens \times Prev}{Sens \times Prev + (1 - Spec) \times (1 - Prev))} \end{aligned}$$

The Bayes Rule in Action

The Bayes Rule

$$p(A|B) = \frac{p(B|A)p(A)}{p(B|A)p(A) + p(B|nonA)p(nonA)}$$

In the context of the disease D and the test T :

$$\begin{aligned} p(D+|T+) &= \frac{p(T+|D+)p(D+)}{p(T+|D+)p(D+) + p(T+|D-)p(D-)} \\ &= \frac{Sens \times Prev}{Sens \times Prev + (1 - Spec) \times (1 - Prev))} \\ &= \frac{0.99 \times 0.001}{0.99 \times 0.001 + 0.01 \times 0.999} \\ &= \frac{0.00099}{0.00099 + 0.00999} \sim \frac{0.001}{0.001 + 0.01} = 0.09 \end{aligned}$$

How to Understand the Result

- In a population of a million, the error due to Sensitivity

  1 % of false negatives

  applies for only a thousand diseased people.
- However, the error due to Specificity

  1 % of false positives

  applies for almost a million healthy people.
- So you should expect about 990 true positive, and 9990 false positives.
- Almost all $T+$ cases are false positives due to the low prior.
- The Most Downloaded Paper in PLoS Medicine in 2005:
  John Ioannides: Why most published research findings are wrong

An Example

- Suppose 1000 theories out of which 100 are true.

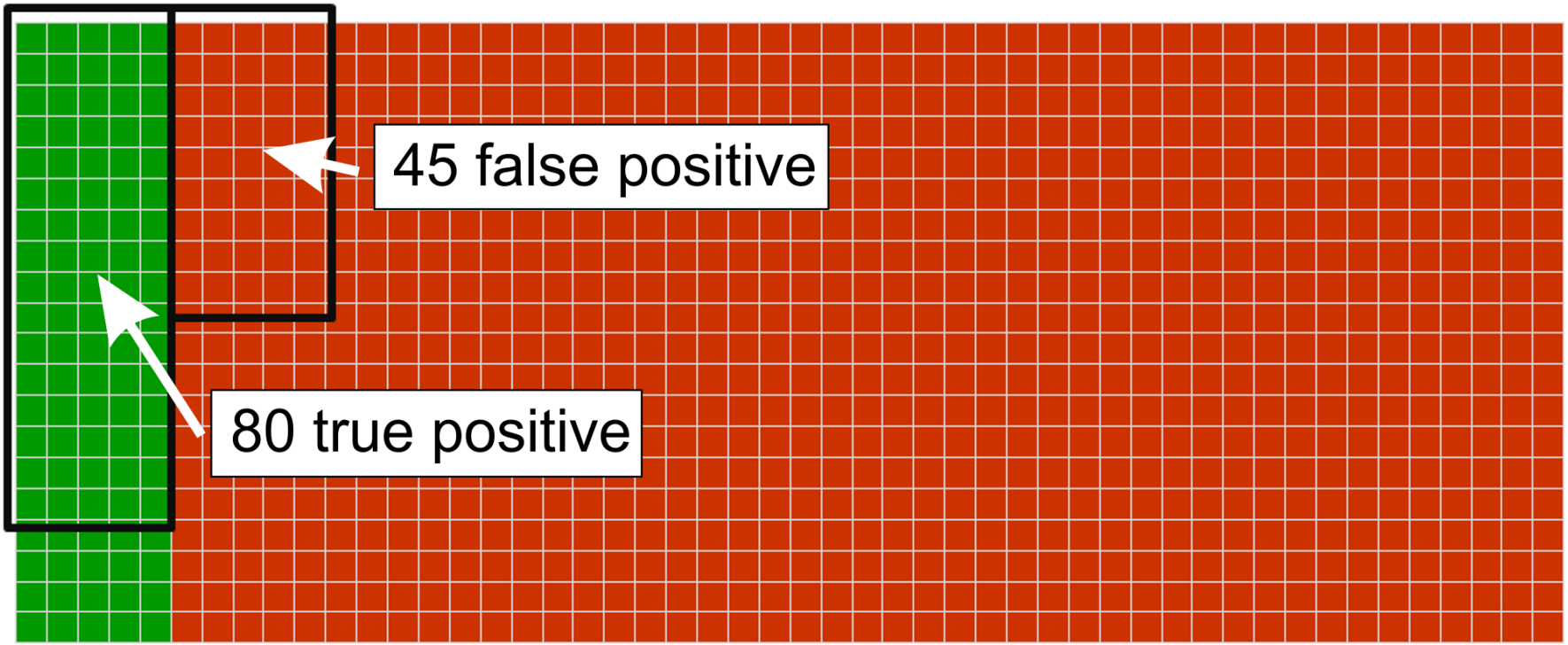
  Gene $X_1 \ldots X_{1000}$ is associated with cancer

- That poses 1000 null hypotheses out of which 100 should be rejected
- A study with a power of 0.8 will find 80 of the 100 truly associated, and miss 20 due to false negatives.
- Of the 900 wrong theories (null hypotheses that should not be rejected), 45 (that is 5%) will look right (and will be rejected) due to false positives.
- That makes $80 + 45 = 125$ rejected hypotheses (positive results), more than third are false positives.
- That also makes 875 not rejected hypotheses (negative results), only 20 of which are false negatives, 97% accuracy.

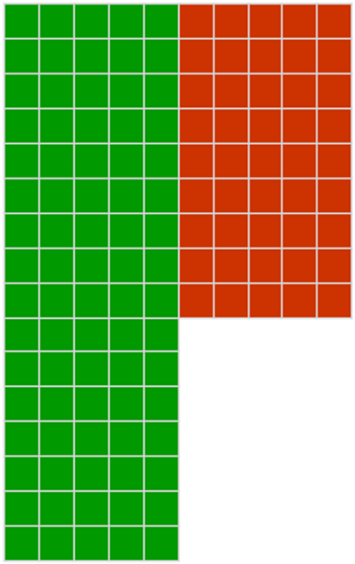
A Scarier Example

- Let's drop the prior to 10 correct theories out of 1000.
- Keep the power at 0.8 and the level of significance at 0.05.
- We get $8 + 49 = 57$ rejected null hypotheses (positive results = papers published), more than 85% are false positives...
- ... and 943 not rejected hypotheses (negative results = no papers) only two of which are false negatives, 99.8% accuracy!

Example Illustrated

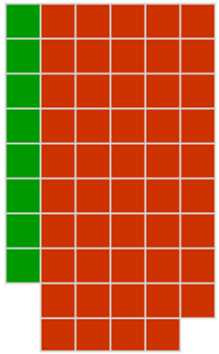


A Scarier Example Illustrated



80 true positive

45 false positive



8 true positive

49 false positive

Important Consequences

1. Negative results (which tend not to get published) are much more likely to be true.

Important Consequences

1. Negative results (which tend not to get published) are much more likely to be true.
2. Smaller studies have smaller power.

   In large randomized studies research findings are more likely to be true.

Important Consequences

1. Negative results (which tend not to get published) are much more likely to be true.
2. Smaller studies have smaller power.

   In large randomized studies research findings are more likely to be true.
3. Small effect sizes decrease the power.

   In scientific fields where the effects are big research findings are more likely to be true.

Important Consequences

1. Negative results (which tend not to get published) are much more likely to be true.
2. Smaller studies have smaller power.

   In large randomized studies research findings are more likely to be true.
3. Small effect sizes decrease the power.

   In scientific fields where the effects are big research findings are more likely to be true.
4. Small priors dramatically decrease the probability of research findings to be true.

   In particular in fields where many hypotheses are tested (microarrays, high throughput discovery oriented research).

Important Consequences

5 The "hotter" the scientific field, the less likely the research findings to be true.

If many teams are testing the same hypothesis,
all of them fail to reject it (and thus do not publish)
but one team rejects it (and publishes) the probability
of the hypothesis being true decreases even further.

Other Issues of Classical Statistics

1. The curse of hypothesis testing.
2. Overfitting: Classical regression models tend to work on test data only, fail to generalize.
3. Interactions: Accounting for interactions by taking products of predictors is wildly inadequate.
4. Estimator construction: How to to that? Usually many answers and many criteria.
5. Failing to show a clearly existing effect.
6. Dependence of p-values on irrelevant information.
7. Weird nature of confidence intervals.

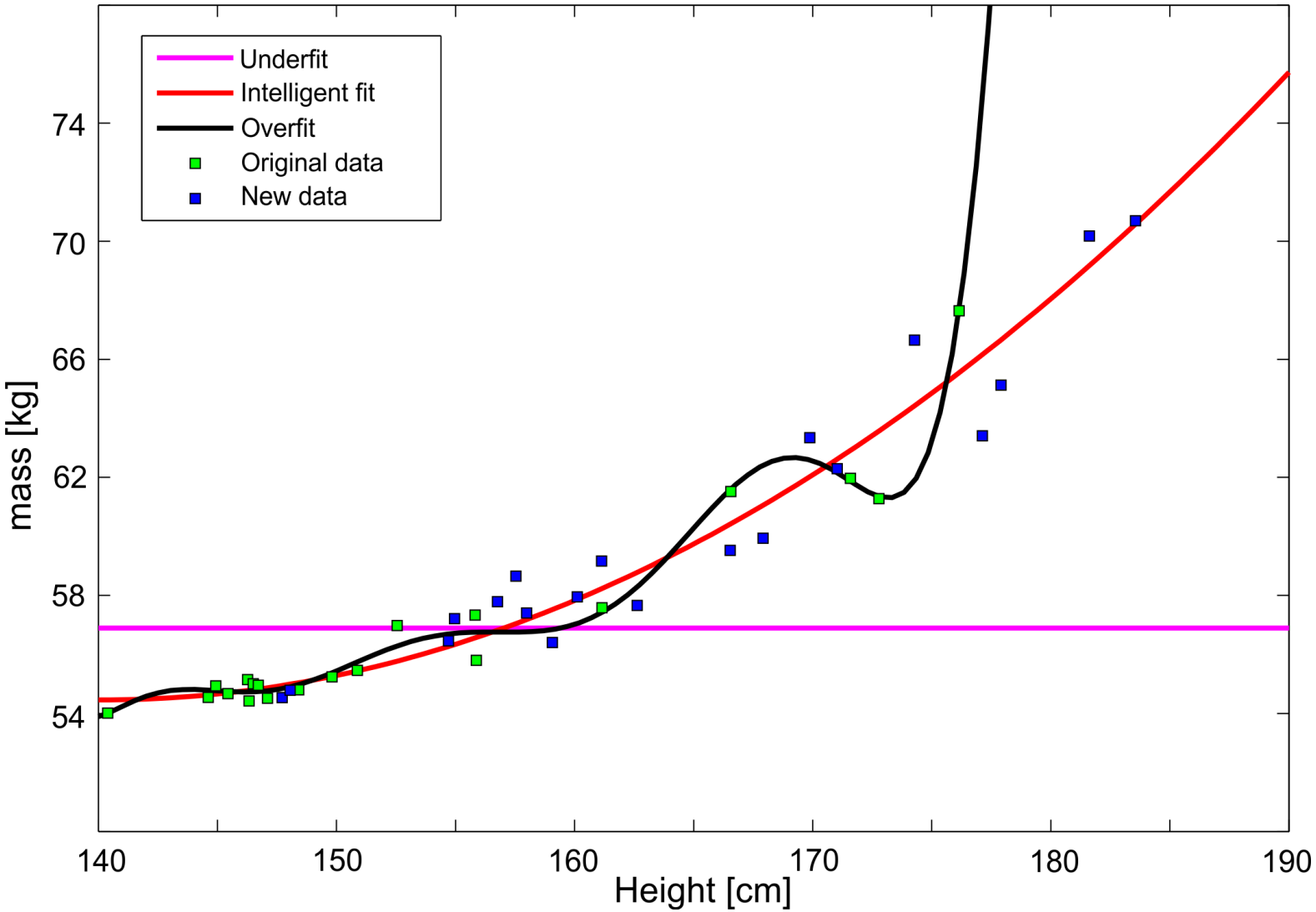
Problems of classical hypothesis testing

Problem 2: Overfitting

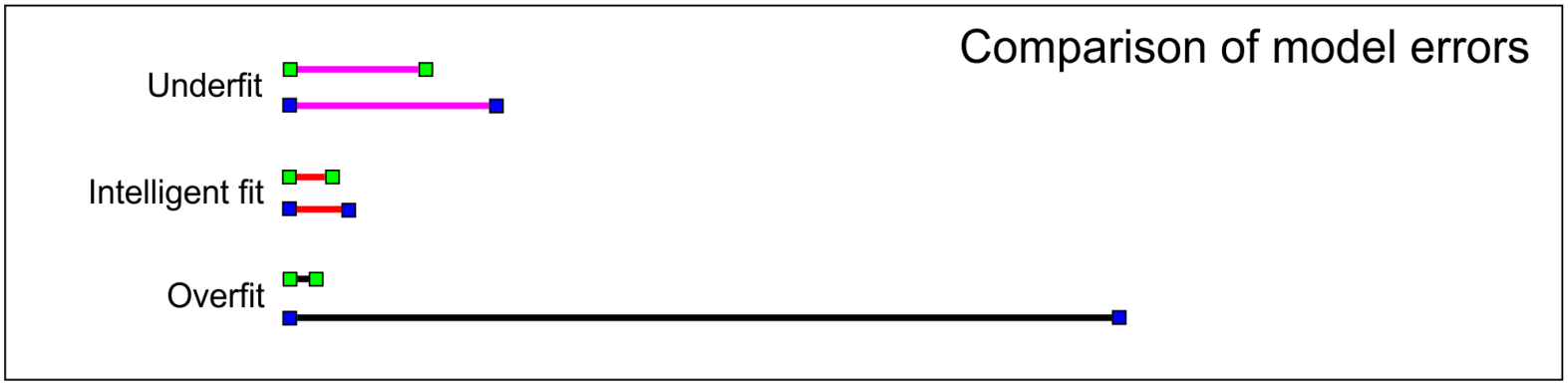
Classical regression models tend to work on the training data only

- A curse of medical research
- Classical statistics lures into a swamp by testing the significance of individual predictors on the training set.
- The problem of building multivariate regression models
- A model can be overfitted even if all the regressors are significant!
- At certain time, around 75% ML articles: plain overfitting
- ML avoid overfitting by validation, Bayesian models have intrinsically built in overfitting prevention!
- BMs allow for natural model comparison.
- BMs do not mislead by testing the significance.

Overfitting



Overfitting



Overfitted model may be worse than prediction by a constant!

Problems with classical statistics

Problem 3: Interactions

Classical regression models tend to ignore interactions among predictors.

How to sweeten your coffee?

1. Stir it (won't work – stirring does not make coffee sweet)
2. Put sugar in it (won't work – sugar sinks to the bottom)
3. Put sugar in and then stir it!

"This week I'd like you to try some cream and sugar.
See how that goes."

Problems with classical statistics

- Classical approach: Products of regressors
- The case of BMI
  1. We know the height and mass and want to predict BMI
  2. We build a model with interactions.
  3. Although all the information for an *exact* prediction are known, our model will be pretty lousy, and totally misleading.
- Products of regressors work only for binary predictors, even there lead to severe overfitting.
- ML solution: ANN look for relevant interactions themselves, SVM switch to the topology of the data.

The problem of interaction

Problems with classical statistics

Problem 4: How to construct estimators

Classical statistician comes up with 10 estimators and 20 criteria which one is the best. Bayesian has a single correct answer for a single well-posed question.

- Particles leave a source and decay at point x with prob

$$p(x) = \frac{1}{\lambda} \exp\left(-\frac{x}{\lambda}\right)$$

- I can observe only a window $x \in (1, 20)$!
- How to estimate λ from the data?
- No classical estimator.
- But the problem trivial and very practical!

Problems with classical statistics

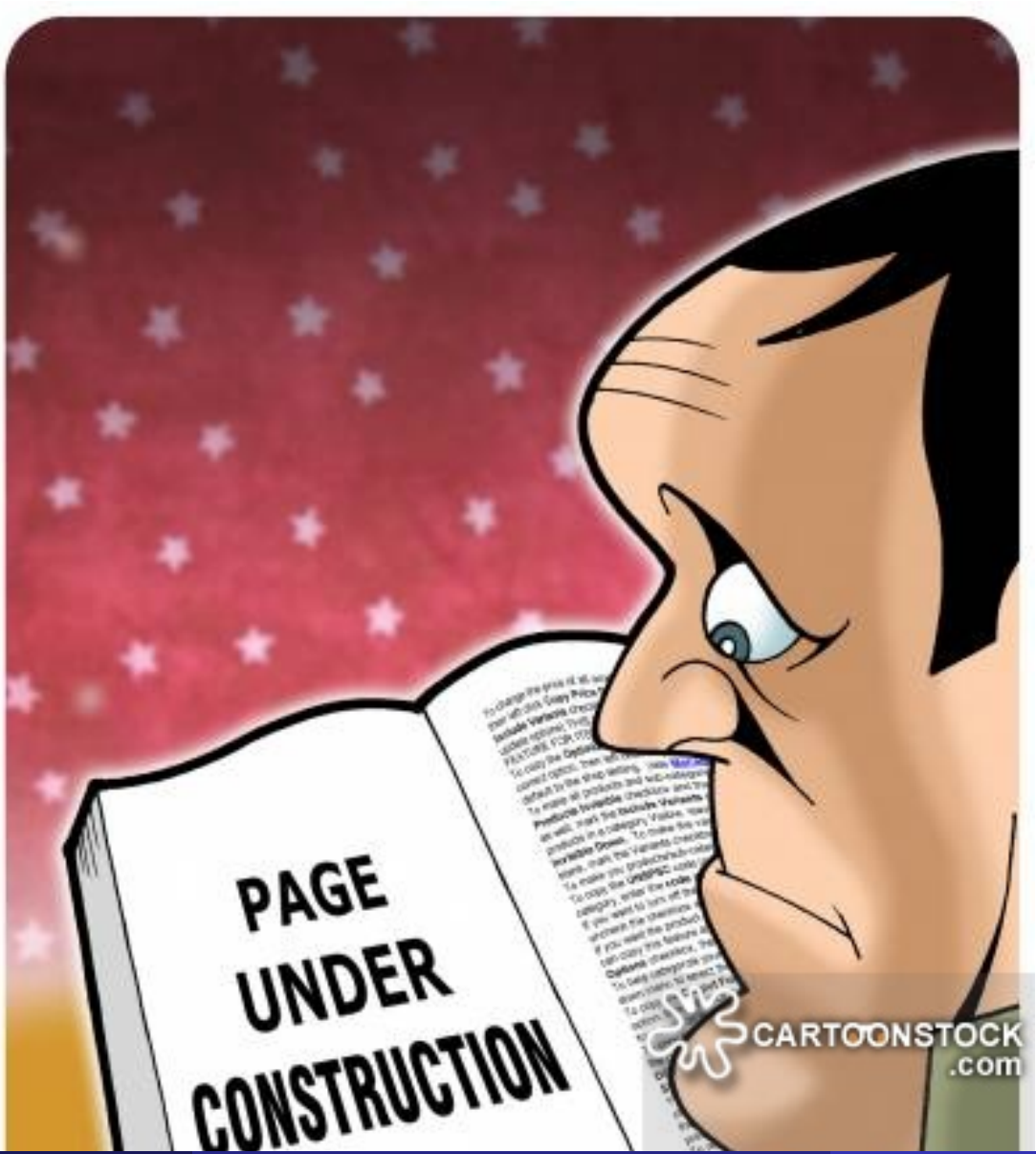
Problem 4: How to construct estimators

Classical statistician comes up with 10 estimators and 20 criteria which one is the best. Bayesian has a single correct answer for a single well-posed question.

- Bayesian solution:

$$\begin{aligned} p(\lambda|x_1,\dots x_n) &= \frac{p(x_1,\dots x_n|\lambda)p(\lambda)}{p(x_1,\dots x_n)} \\ &\sim p(x_1|\lambda)\dots p(x_n|\lambda)p(\lambda) \end{aligned}$$

Estimator under construction



Problems with classical statistics

Problem 5: Dependence of p-values on irrelevant information

The p-values of classical tests are dependent on irrelevant information!

- 12 times toss a coin and obtain

$$aaabaaaabaab \qquad \text{or} \qquad 3 \times b \text{ and } 9 \times a$$

- Is the coin fair?
- Classical statistics tests $H_0 : p = 0.5$ against the rest of the Universe.

$$prob(r \leq 3 | p = 0.5) = \sum_{r=0}^{3} \binom{12}{r} 0.5^{12} = 0.07$$

- The RV is a but the number of tosses is fixed (non-random).

Problems with classical statistics

- The experimenter says: No, I was waiting for three b's to appear and then I stopped tossing.
- Suddenly, we have a different hypothesis test:

$$prob(n \geq 12 | r = 3) = \sum_{n=12}^{\infty} \binom{n-1}{r-1} 0.5^n = 0.03$$

- What if someone had watched through the window and does not know why the experiment stopped? What inference should he make?
- The Bayesian answer is simple, clear, and independent of the stopping rule.

Irrelevance

A very practical example

We try to reduce incidence of a disease called *frequentism*.

A practical example

- We try to reduce incidence of a disease called *frequentism*.
- Two vaccinations A (active) and B (placebo) are tested.
- 30 people get A, 10 get B, patients randomized.

A practical example

- We try to reduce incidence of a disease called *frequentism*.
- Two vaccinations A (active) and B (placebo) are tested.
- 30 people get A, 10 get B, patients randomized.
- After one year, 1 gets *frequentism* in group A and 3 get it in group B.
- Question 1: Is A better than B?
- Question 2: How likely is it that A is at least 10 times better than B?

The frequentist "solution"

- Null hypothesis: A and B have the same effect. Compare to the alternative.
- This is not what you want: You want to compare "A is better than B" to "A is worse than B".
- Construct a test statistic (why exactly this one?)

$$\chi^2 = \frac{(A^+ - EA^+)^2}{EA^+} + \frac{(A^- - EA^-)^2}{EA^-} + \frac{(B^+ - EB^+)^2}{EB^+} + \frac{(B^- - EB^-)^2}{EB^-}$$

- where e.g. EA^+ is the expected number under the null hypothesis

The frequentist "solution"

- The Yates's correction! What is it? Should you use it or not?
- Degrees of freedom (What the hell is that?)
- Estimate EA^+ from the data by

$$\frac{A^+ + B^+}{N} A$$

- Plug it into the statistic, get $p < 0.05$ and reject H_0.
- Or use the Yates's correction, get $p > 0.05$ and accept H_0.
- Why 0.05 exactly?
- But χ^2 asymptotic approximation is all right only if all observed frequencies are greater than 5. Why 5? What does asymptotic approximation mean?

The Bayesian solution

- p_{A+} is the prob that within one year from vaccination by A one will get *frequentism*. p_{B+} analogously.
- The Bayes' Theorem:

$$p(p_{A+}, p_{B+}|Data) = \frac{p(Data|p_{A+}, p_{B+})p(p_{A+}, p_{B+})}{p(Data)}$$

The Bayesian solution

- p_{A+} is the prob that within one year from vaccination by A one will get *frequentism*. p_{B+} analogously.
- The Bayes' Theorem:

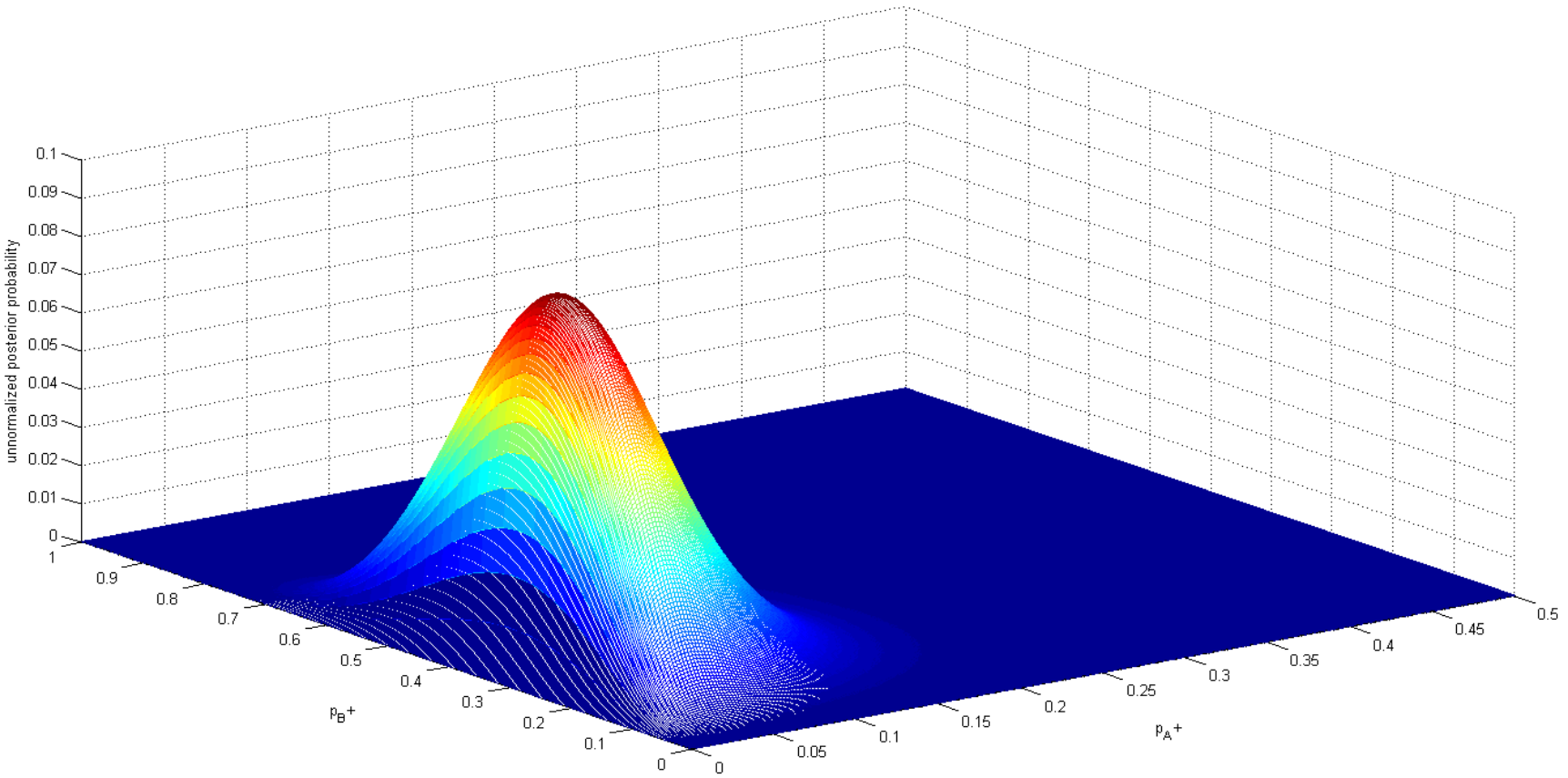
$$p(p_{A+}, p_{B+}|Data) = \frac{p(Data|p_{A+}, p_{B+})p(p_{A+}, p_{B+})}{p(Data)}$$

- A high school exercise:

$$p(Data|p_{A+}, p_{B+}) = \binom{30}{1}(p_{A+})^1(p_{A-})^{29}\binom{10}{3}(p_{B+})^3(p_{B-})^7$$

- Use vanilla prior $p(p_{A+}, p_{B+}) = 1$
- Or use whatever prior knowledge you might have!

MatLab simulation



Famous quotes

Winston Churchill (actuallyJoseph Goebbels)

I only believe in statistics that I doctored myself.

Michael Bloomberg

In God we trust. Everyone else, bring data!

What Can Be Done

1. Encourage replication! Our PhD initiative.

What Can Be Done

1. Encourage replication! Our PhD initiative.
2. Raise peer-review standards! ArXiv, PeerJ, PLoS ONE

What Can Be Done

1. Encourage replication! Our PhD initiative.
2. Raise peer-review standards! ArXiv, PeerJ, PLoS ONE
3. Encourage publication of negative results.

What Can Be Done

1. Encourage replication! Our PhD initiative.
2. Raise peer-review standards! ArXiv, PeerJ, PLoS ONE
3. Encourage publication of negative results.
4. Register all studies beforehand.

What Can Be Done

1. Encourage replication! Our PhD initiative.
2. Raise peer-review standards! ArXiv, PeerJ, PLoS ONE
3. Encourage publication of negative results.
4. Register all studies beforehand.
5. Remove the bad incentives to publish excessively.

What Can Be Done

1. Encourage replication! Our PhD initiative.
2. Raise peer-review standards! ArXiv, PeerJ, PLoS ONE
3. Encourage publication of negative results.
4. Register all studies beforehand.
5. Remove the bad incentives to publish excessively.
6. Abolish financing based on number of publications.

What Can Be Done

1. Encourage replication! Our PhD initiative.
2. Raise peer-review standards! ArXiv, PeerJ, PLoS ONE
3. Encourage publication of negative results.
4. Register all studies beforehand.
5. Remove the bad incentives to publish excessively.
6. Abolish financing based on number of publications.
7. Abolish career decisions based on number of publications

What Can Be Done

1. Encourage replication! Our PhD initiative.
2. Raise peer-review standards! ArXiv, PeerJ, PLoS ONE
3. Encourage publication of negative results.
4. Register all studies beforehand.
5. Remove the bad incentives to publish excessively.
6. Abolish financing based on number of publications.
7. Abolish career decisions based on number of publications
8. Abolish the CE system of academic titles!

What Can Be Done

1. Encourage replication! Our PhD initiative.
2. Raise peer-review standards! ArXiv, PeerJ, PLoS ONE
3. Encourage publication of negative results.
4. Register all studies beforehand.
5. Remove the bad incentives to publish excessively.
6. Abolish financing based on number of publications.
7. Abolish career decisions based on number of publications
8. Abolish the CE system of academic titles!
9. Go Bayes!

First Success: The ASA Statement on p-values

1. P-values can indicate how incompatible the data are with a specified statistical model.
2. P-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone.
3. Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold.
4. Proper inference requires full reporting and transparency (p-hacking)
5. A p-value, or statistical significance, does not measure the size of an effect or the importance of a result.
6. By itself, a p-value does not provide a good measure of evidence regarding a model or hypothesis.

Wasserstein: The ASA's Statement on p-Values,
The American Statistician, 2016

It's So Simple!

Go Bayes!