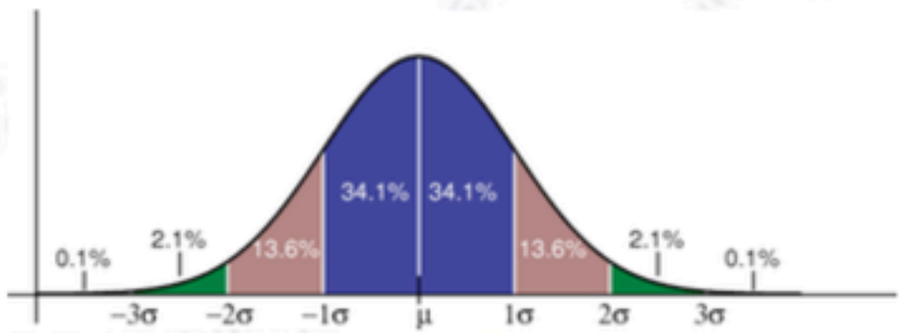
Applied Statistics

Systematic Uncertainties






Troels C. Petersen and Mathias Luidor Heltberg (NBI)

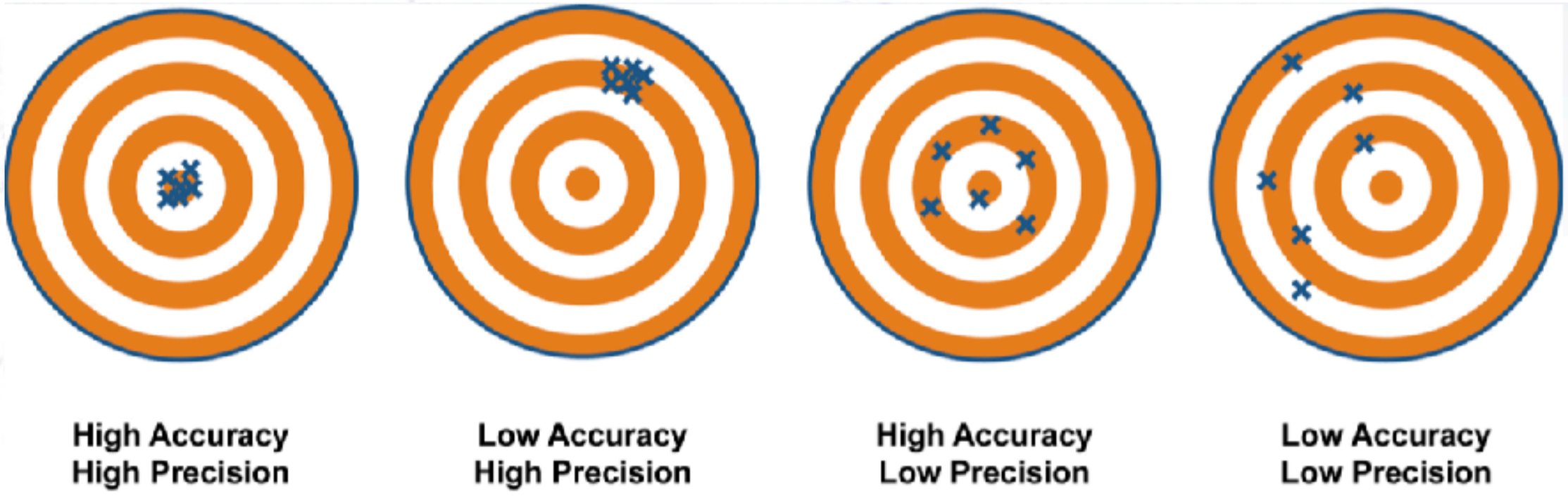


"Statistics is merely a quantisation of common sense"

What is a systematic uncertainty?

Concept and definitions of ‘systematic uncertainties’ originates from physics, not from fundamental statistical methodology.

A common definition is: “Systematic uncertainties are all uncertainties that are not directly due to the statistics of the data”

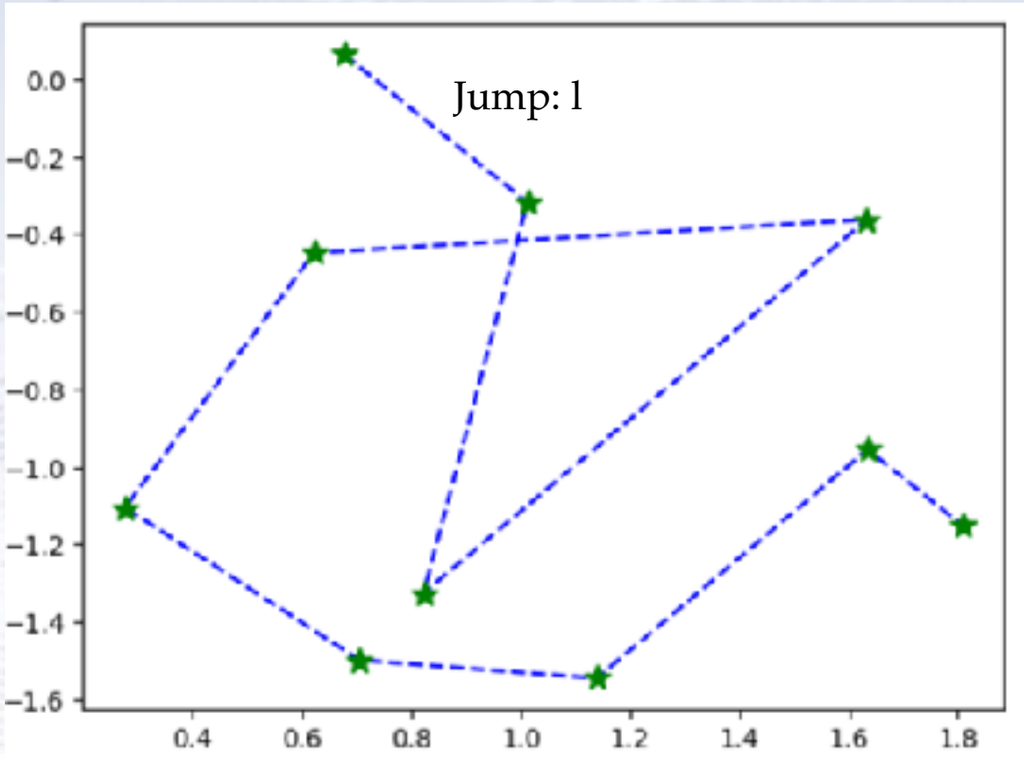


Example

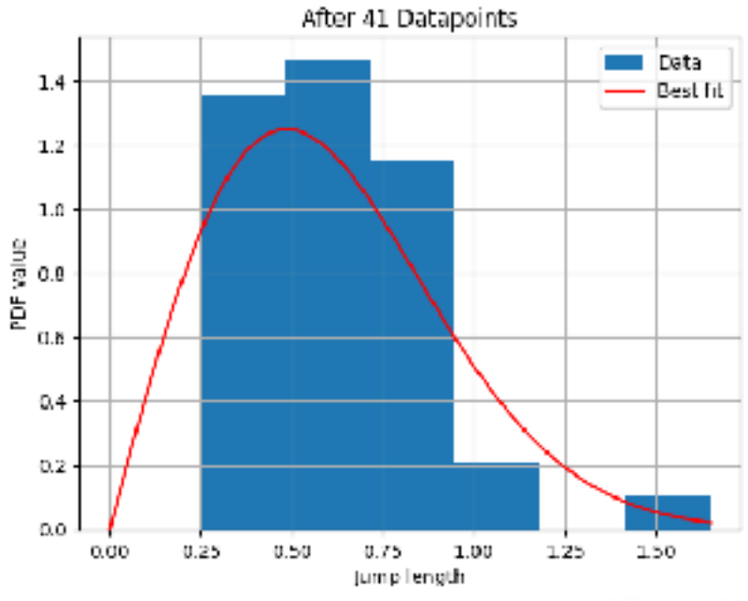
The ideal experiment

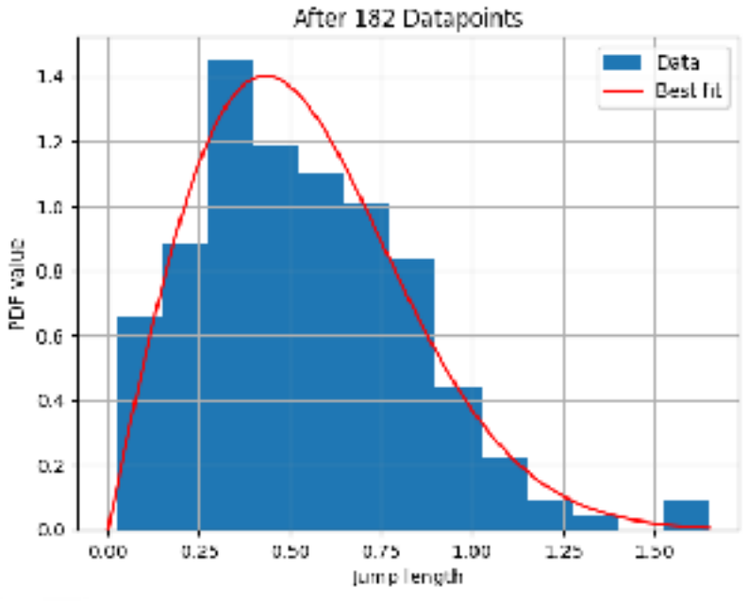
Lets consider an ideal/trivial experiment.

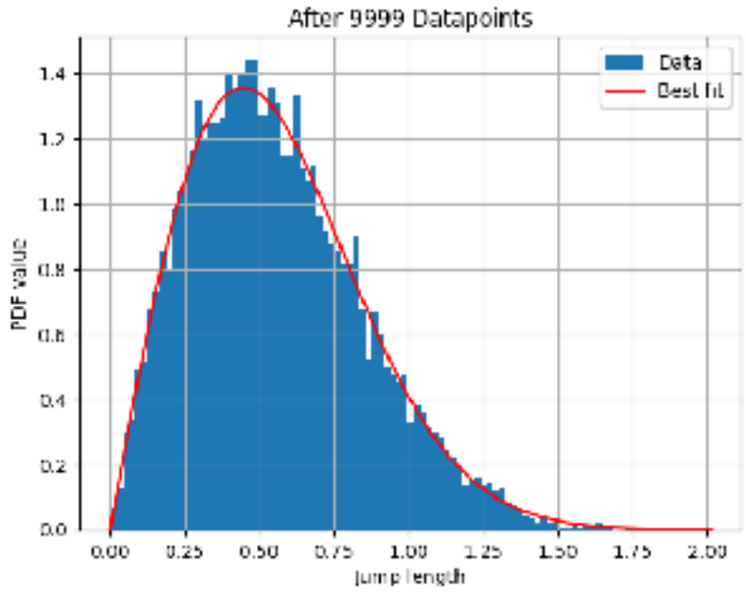
We are measuring the diffusion of a particle (green).



The ML gives: $D^* = \frac{\langle l^2 \rangle}{4\tau}$







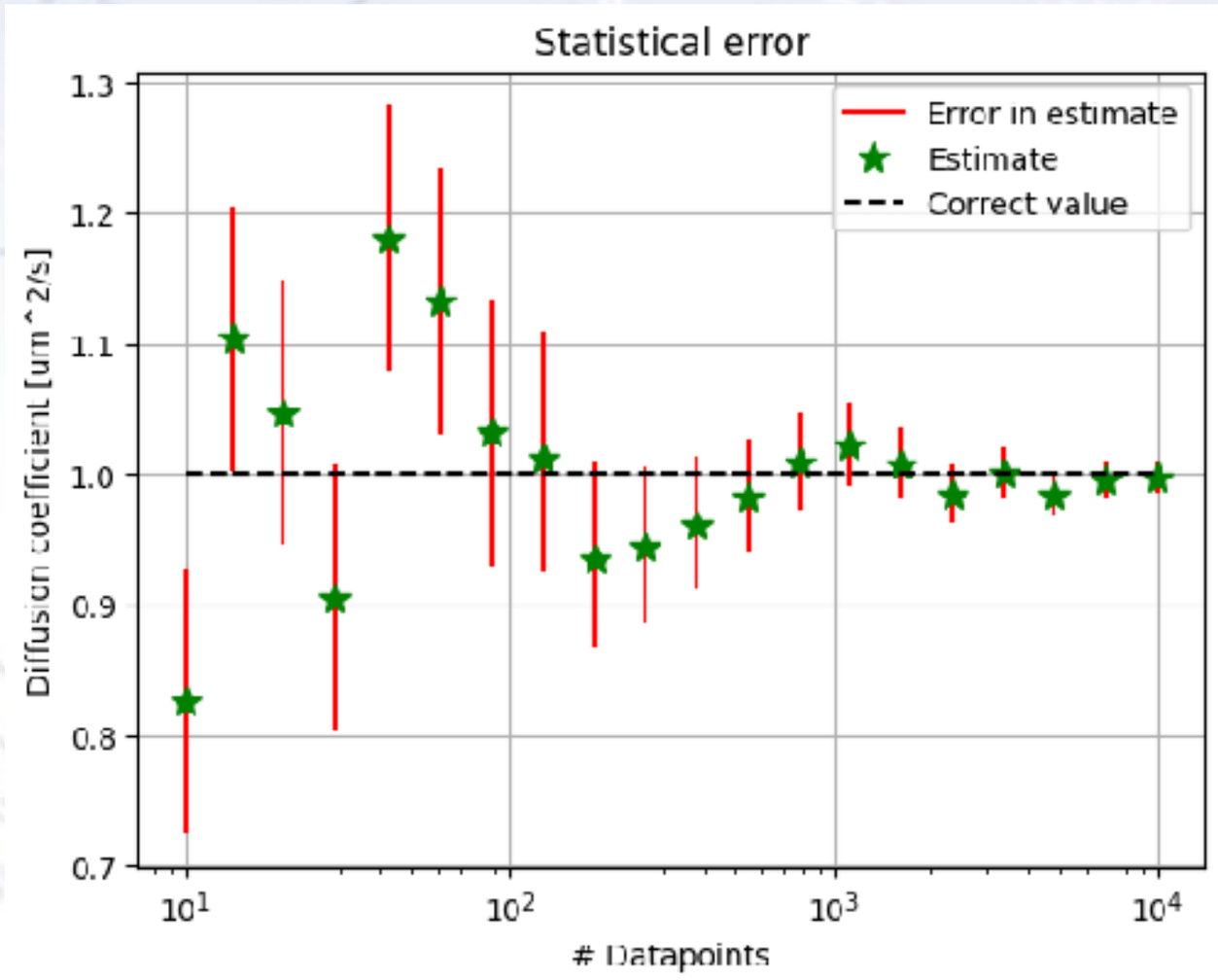
The ideal experiment

With *statistical* uncertainty:

As number of datapoints increase

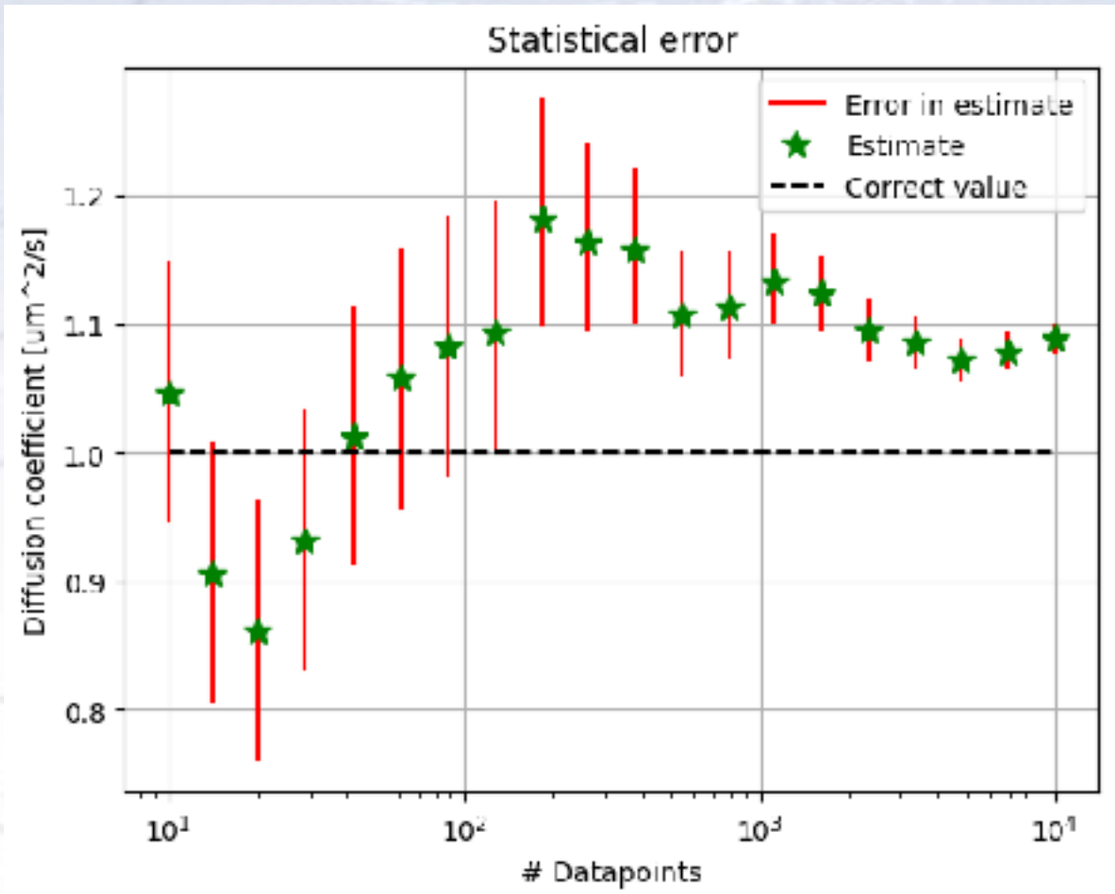
- > the error decrease

-> the result converge to correct value



The non-ideal experiment

Suppose I have now done another experiment: the result does not converge to the correct value.



I have 3 main suspects:

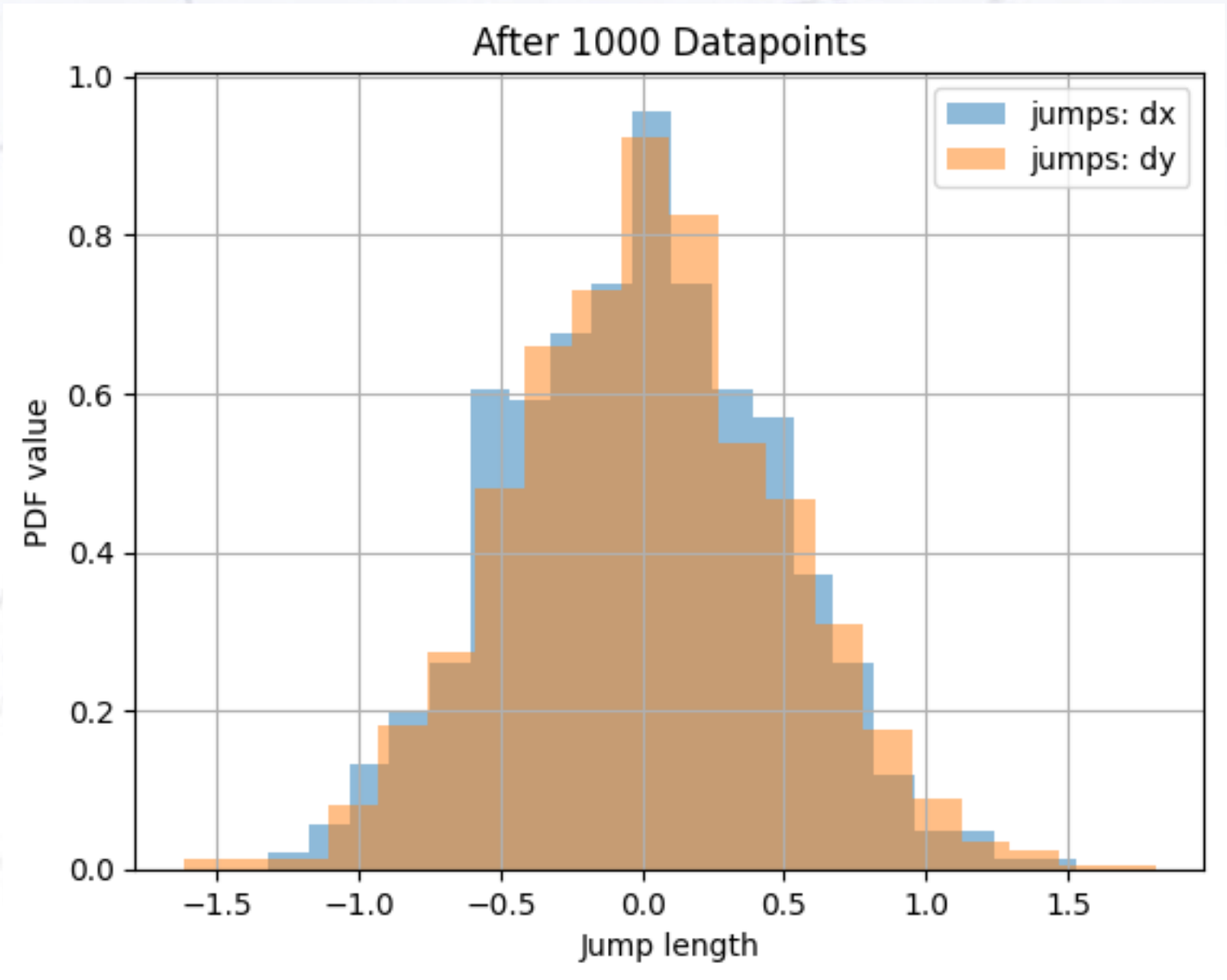
1) A flow from the left might have pushed the particles additionally

2) Another (faster) molecule might have been measured

3) Experimental noise might have influenced results

1) A flow from the left might have pushed the particles additionally I could compare the jumps only in x and only in y.

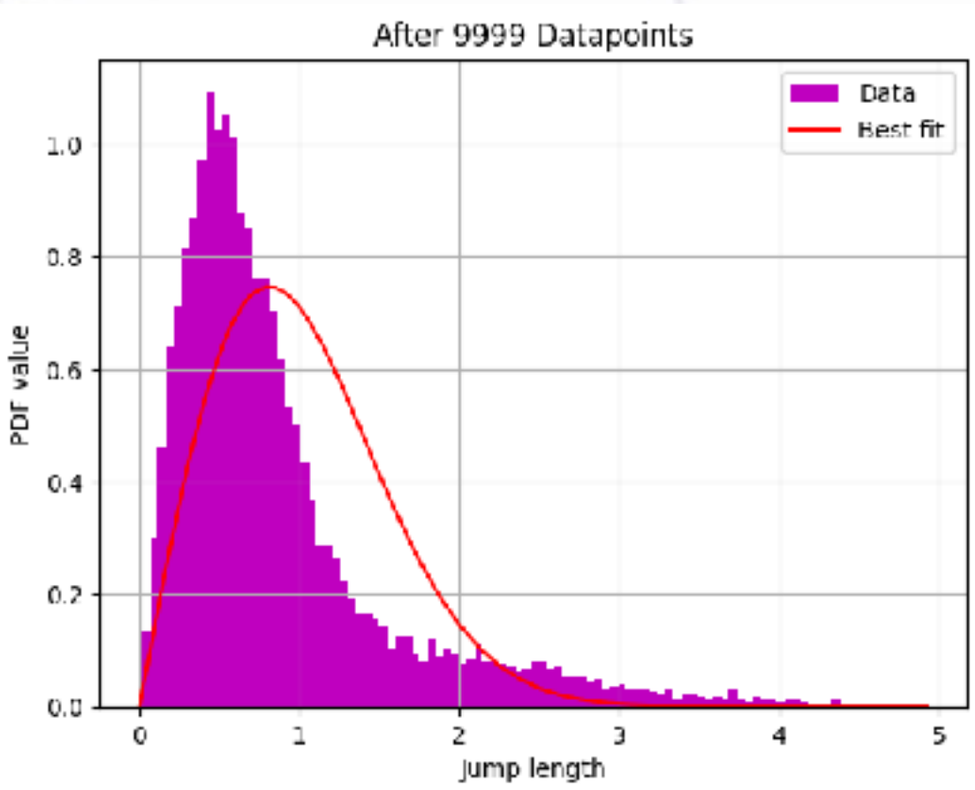
I find that the distributions are statistically the same - thereby rejecting this hypothesis

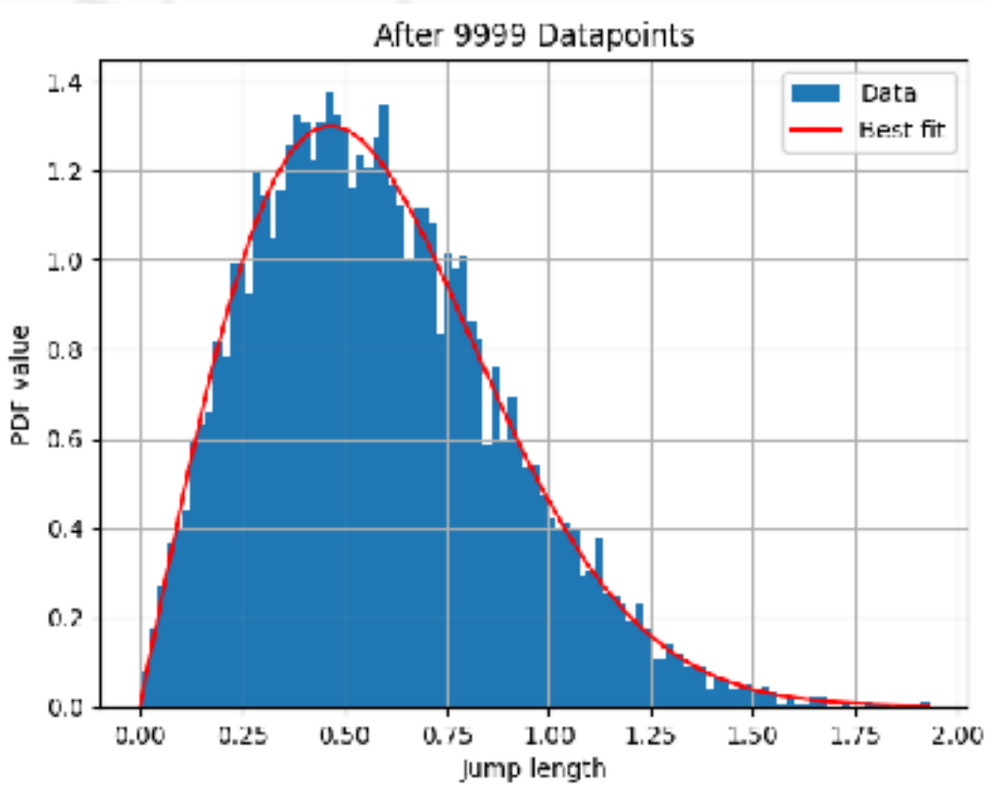


2) Another (faster) molecule might have been measured

I simulate the data for 10% of the data, having a higher diffusion coefficient. Here I find that the best fit is a terrible fit.

For my actual data I get a very good fit - thereby rejecting this hypothesis





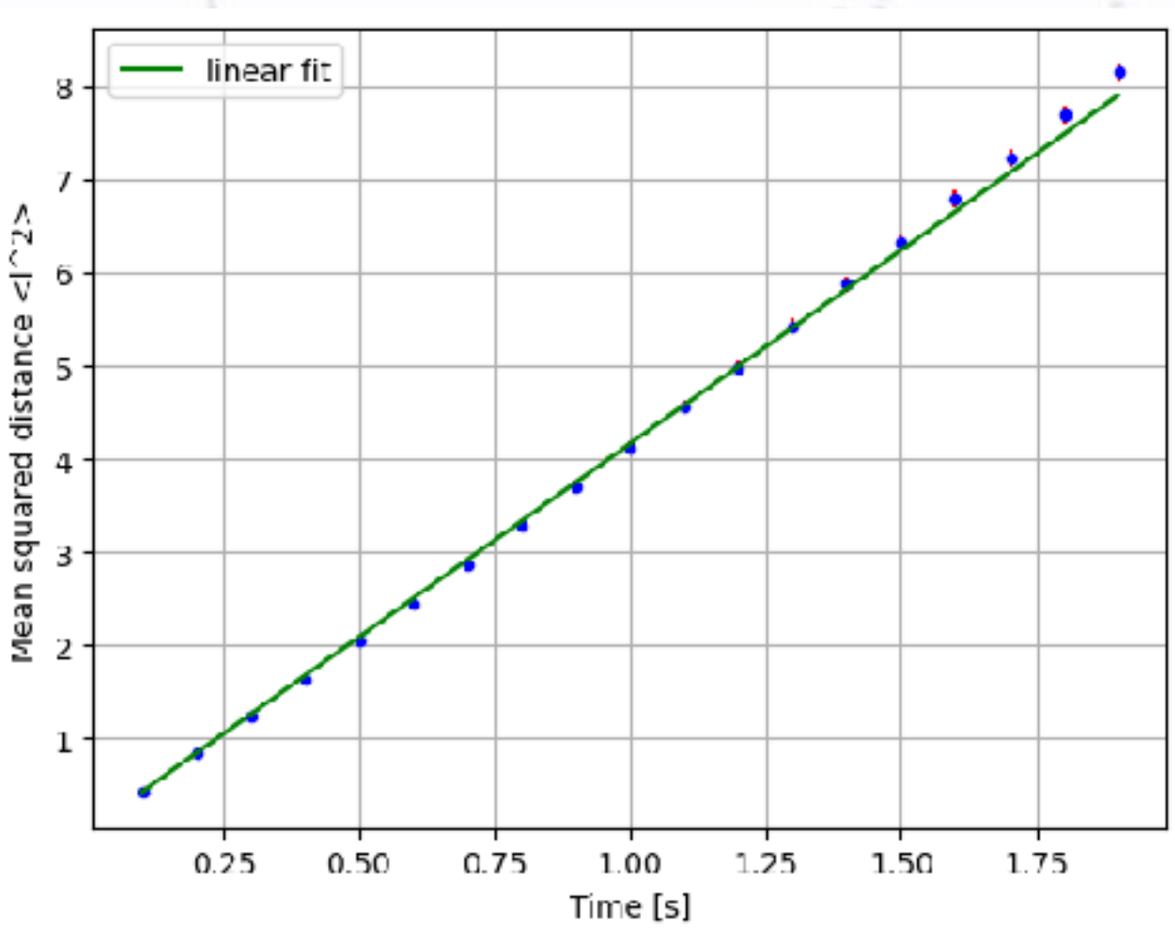
3) Experimental noise might have influenced results

How can we separate experimental noise from stochastic fluctuations?
Careful investigation yields:

$$dX = \mathcal{N}(0, \sqrt{2D\tau}) + \mathcal{N}(0, \sigma) \quad = \mathcal{N}\left(0, \sqrt{2(D + \frac{\sigma^2}{2\tau})\tau}\right)$$

Now I now how to get the basic estimate of the diffusion coefficient -

$$\langle l^2 \rangle = \langle l_x^2 \rangle + \langle l_y^2 \rangle = 4\tau D + 2\sigma^2$$



D estimate from slope:
1.02 +/- 0.02 m^2/s

Conclusion

Understanding systematic errors require a detailed understanding of the data

When investigating the effect of a hypothesized source of systematic error: calculate/simulate the effect!

Sources of Systematic Errors

We can broadly divide the source of systematic errors in 3 categories

1. Instrumental (e.g., miscalibration, measurement errors).

2. Environmental (e.g., temperature fluctuations, electromagnetic interference).

3. Procedural (e.g., incorrect assumptions, bias).

If systematic errors are *smaller* than statistical errors - our way of improving is to get more data.

If systematic errors are *larger* than statistical errors - our way of improving to enhance our understanding of the experiment

Instrumental errors

Example case

Example *Backgrounds*
You are measuring the efficiencies of several Geiger counters using a Pb^{210} source, labelled 10 μCi. The efficiency is given by

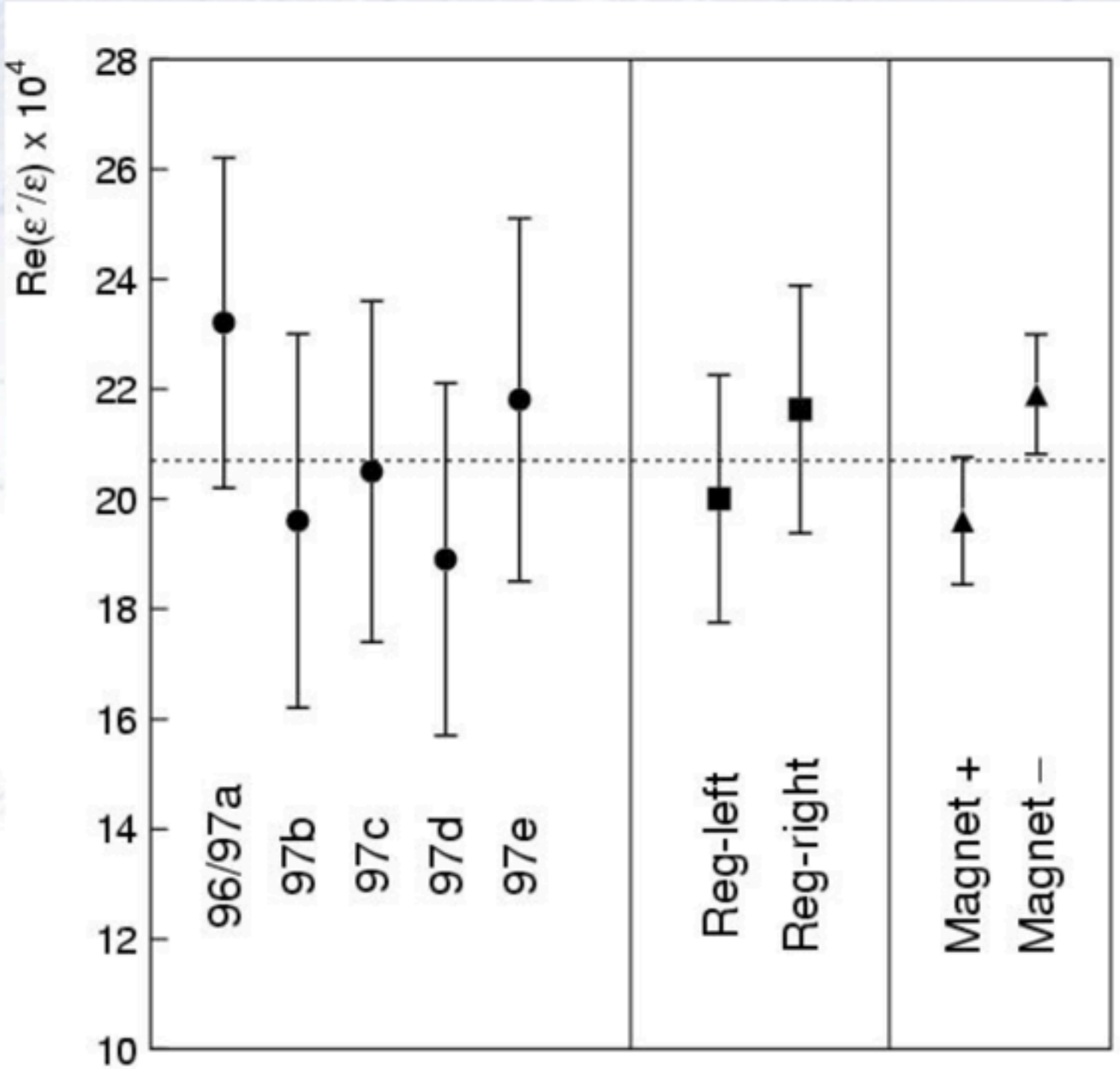
$$\eta = \frac{\text{observed rate}}{\text{true rate}}.$$

Now, 10 μCi is 370 000 disintegrations per second—but unfortunately you do not know when it was measured, and the half-life of Pb^{210} is only 21 years, so the true rate may really be less than the number on the label.

You decide, by inspecting the condition of the source container and from what you know about the way things work in your lab, that it is most unlikely for the source to be more than 5 years old, and in this worst case the rate is $370\,000 \times 2^{-5/21} =$ 313 710 counts per second. You take the most likely value as being midway between the two, 341 855, and appeal to the fact that the variance of a uniform distribution is 1/12 its width to give the error as $(370\,000 - 313\,710)/\sqrt{12} = 16\,250$, and the error is thus 5%. You are not entirely happy with this, but it is the best you can do with the data available, and you comfort yourself with the thought that it is probably an overestimate.

escape your notice.

Cross check of data



Classic check of systematic errors, by dividing the data according to:

- Period of data taking
- Direction of regulator
- Direction of B-field

If any of these showed an inconsistency between the subsamples, one would know that this had an impact on the result.

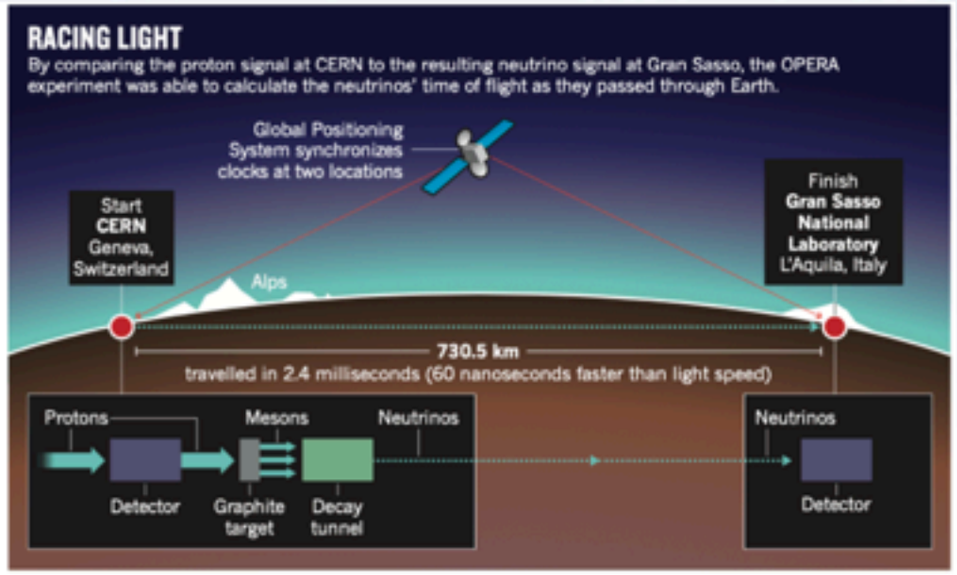
This type of cross checks is at the heart of data analysis.

(Another) Example of systematic error

One of the best "recent" examples is the case of physicists measuring neutrinos to travel faster than speed of light.

This would (if true) turn the foundations of physics in ruins…
After 6 months of intense studies, the researchers found two possible systematic errors:

- A link from a GPS receiver to the OPERA master clock was loose, which increased the delay through the fiber.
- A clock on an electronic board ticked faster than its expected 10 MHz frequency, lengthening the reported flight-time of neutrinos, thereby somewhat reducing the seeming faster-than-light effect.



PARTICLE PHYSICS

Speedy neutrinos challenge physicists

Experiment under scrutiny as teams prepare to test claim that particles can beat light speed.

BY EUGENIE SAMUEL REICH

The joke begins with the barman saying: "I'm sorry, we don't serve neutrinos." Then the punch line: a neutrino walks into a bar.

Such causality-bending humour has been rife on the Internet in the past week, following the news that an experiment at the Gran Sasso

worth of physics upended, starting with Albert Einstein's special theory of relativity. This sets the velocity of light as the inviolable and unattainable limit for matter in motion, and links it to deeper aspects of reality, such as causality.

Physicists, for the most part, suspect that an unknown systematic error lies behind OPERA's startling result. But nothing obvious has emerged, and many see the experiment as

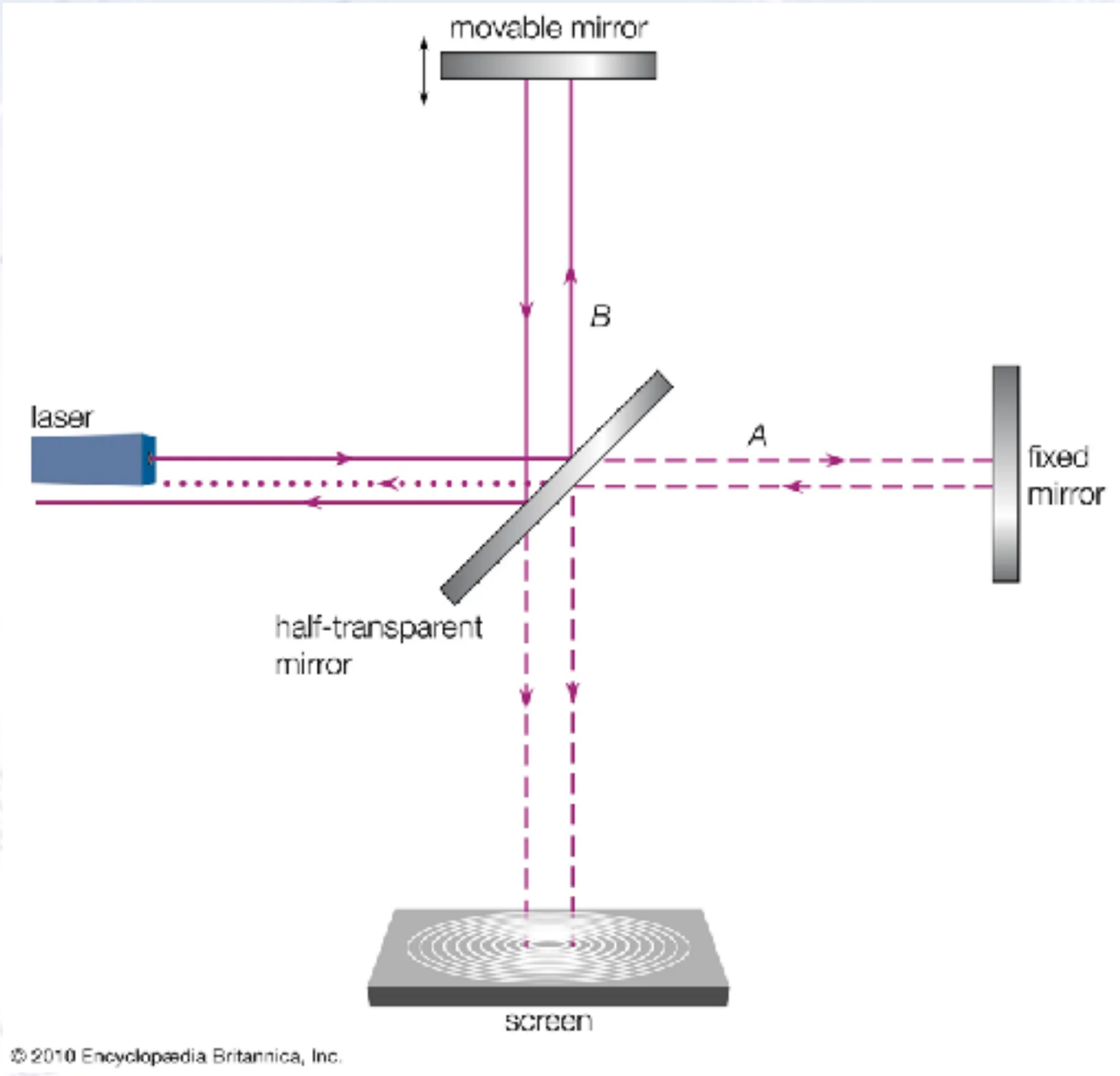
Environmental errors

Classical examples

Experiment: The Michelson-Morley experiment aimed to detect the "aether wind" by measuring the speed of light in different directions.

Systematic Error: Thermal expansion of the apparatus caused small but measurable shifts in the interference pattern, complicating data interpretation.

Resolution: Improved materials and careful temperature control helped isolate the null result that challenged the existence of the aether.



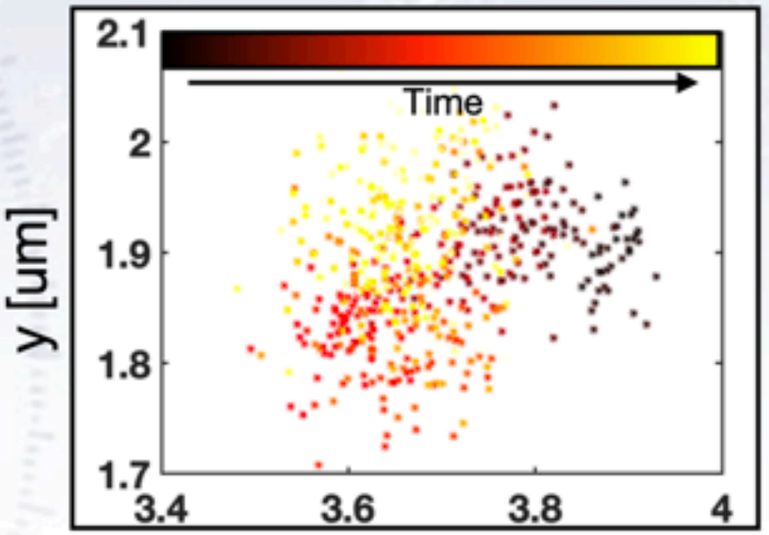
(Another) Example of systematic error

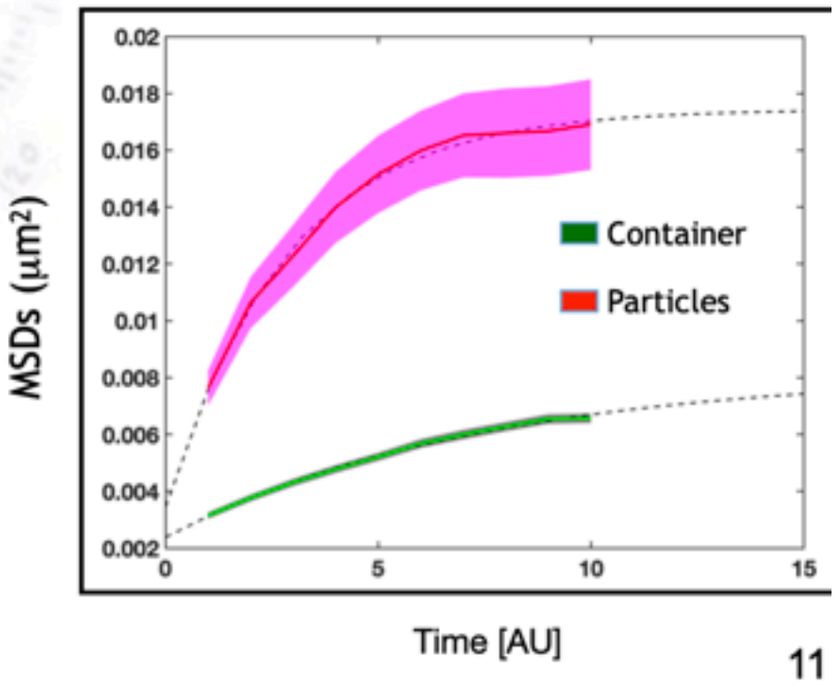
Imagine you have a set of measurements (trapped particles) , and you want to measure the size of the container.

You look at them and think you can just measure their circumference and use that as an estimator.

Next you realize that the container is not constant in time! This leads to a serious overestimation of that container.

In order to resolve this, you need to come up with new ways on analysing the data with methods that do not assume constant position of that container.





Procedural errors

The problem for every scientist

Typically in science we have some kind of theory that we is testing a hypothesis of.

Clearly some measurements represent something completely different - for instance the background in an experiment.

Ideally we want to understand everything but as time and ressources are limited we might treat something as systematic uncertainties.

However - we really don’t want to be the scientists that threw out their Nobel Prize of interesting data because we treated it as systematic uncertainties.

Biased measurements

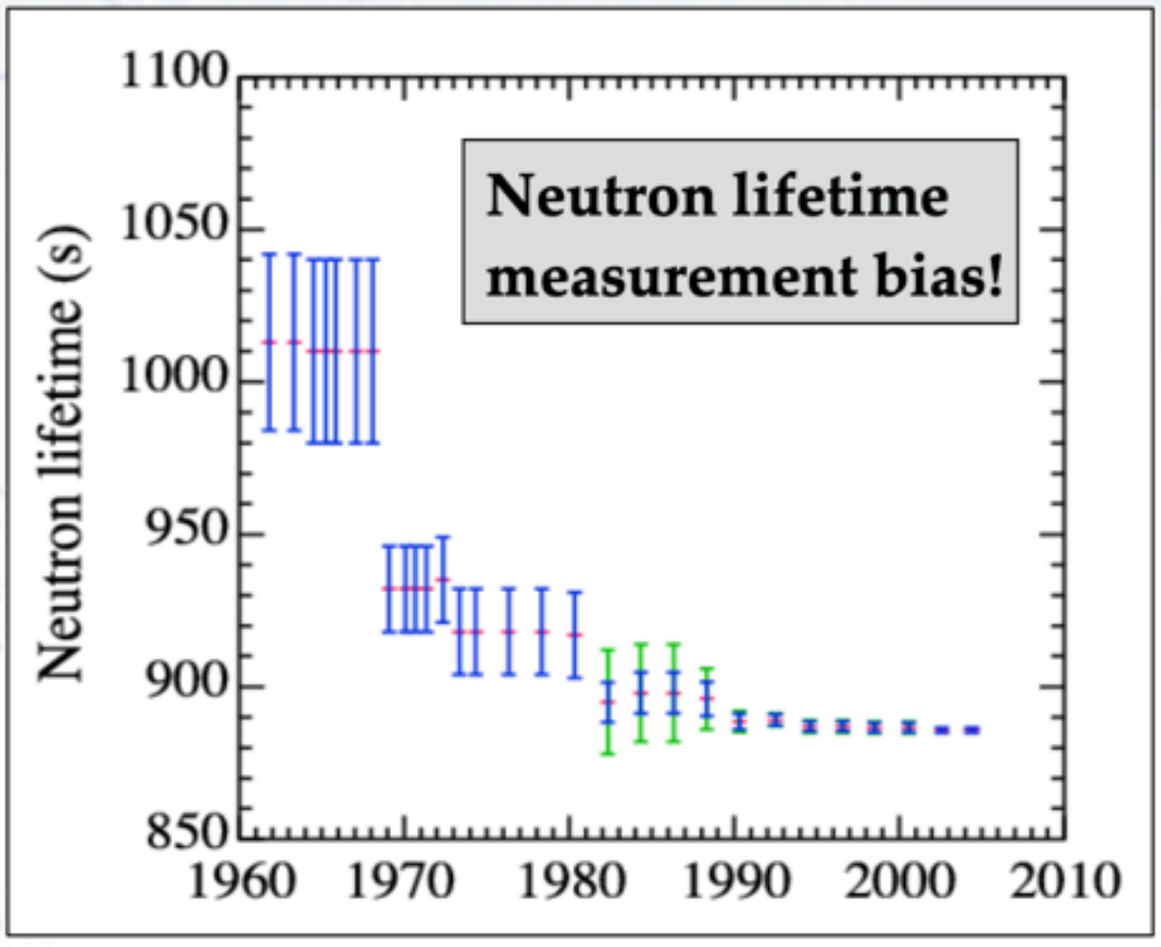
Why does my experiment find a lower value than others?

It is questions like these, that makes you start looking for effects that could yield a higher value, leading to…

Biases!

When measuring a parameter for which there are already expectations/predictions, the result can be biased. Examples:

- Millikan's oil-drop experiment.
- Epsilon prime (CERN vs. FNAL).
- Most politically influenced decisions!



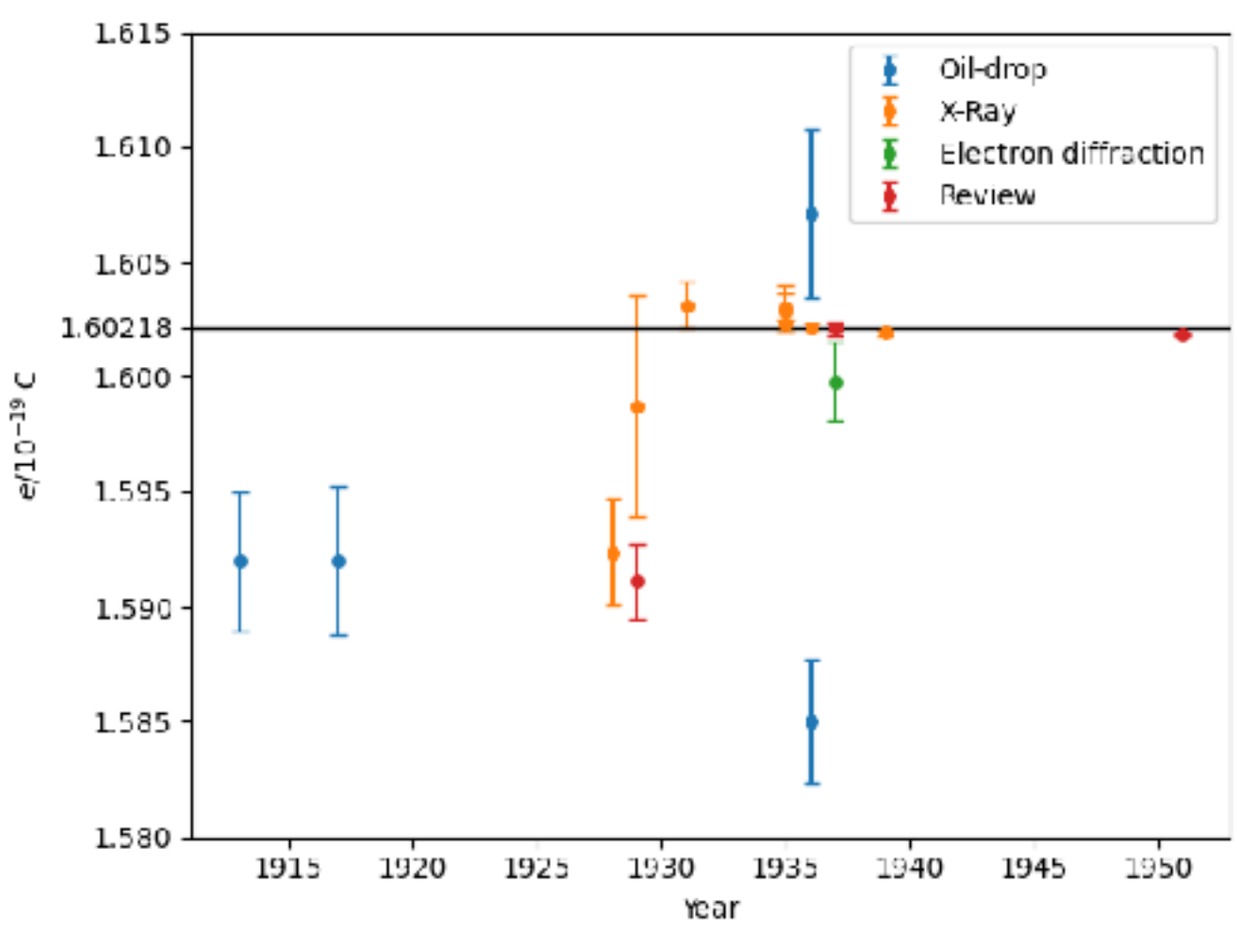
Those who forget good and evil and seek only the facts are more likely to achieve good, than those who view the world through the distorting medium of their own desires. [Bertrand Russell]

The charge of an electron

We have learned a lot from experience about how to handle some of the ways we fool ourselves.

…Millikan measured the charge on an electron by an experiment with falling oil drops, and got an answer which we now know not to be quite right.

…it's apparent that people did things like this: When they got a number that was too high above Millikan's, they thought something must be wrong—and they would look for and find a reason why something might be wrong. When they got a number close to Millikan's value they didn't look so hard…



Blinding of results

To avoid experimenters biases, **blinding** has been introduced.

This means that the computer adds a random number to the result, which is not removed before the analysis has been thoroughly checked.

Example:

```
> ./FitSin2beta
Result is: sin(2beta) = x.xx +- 0.37
Do you wish to unblind (y/n)?
```

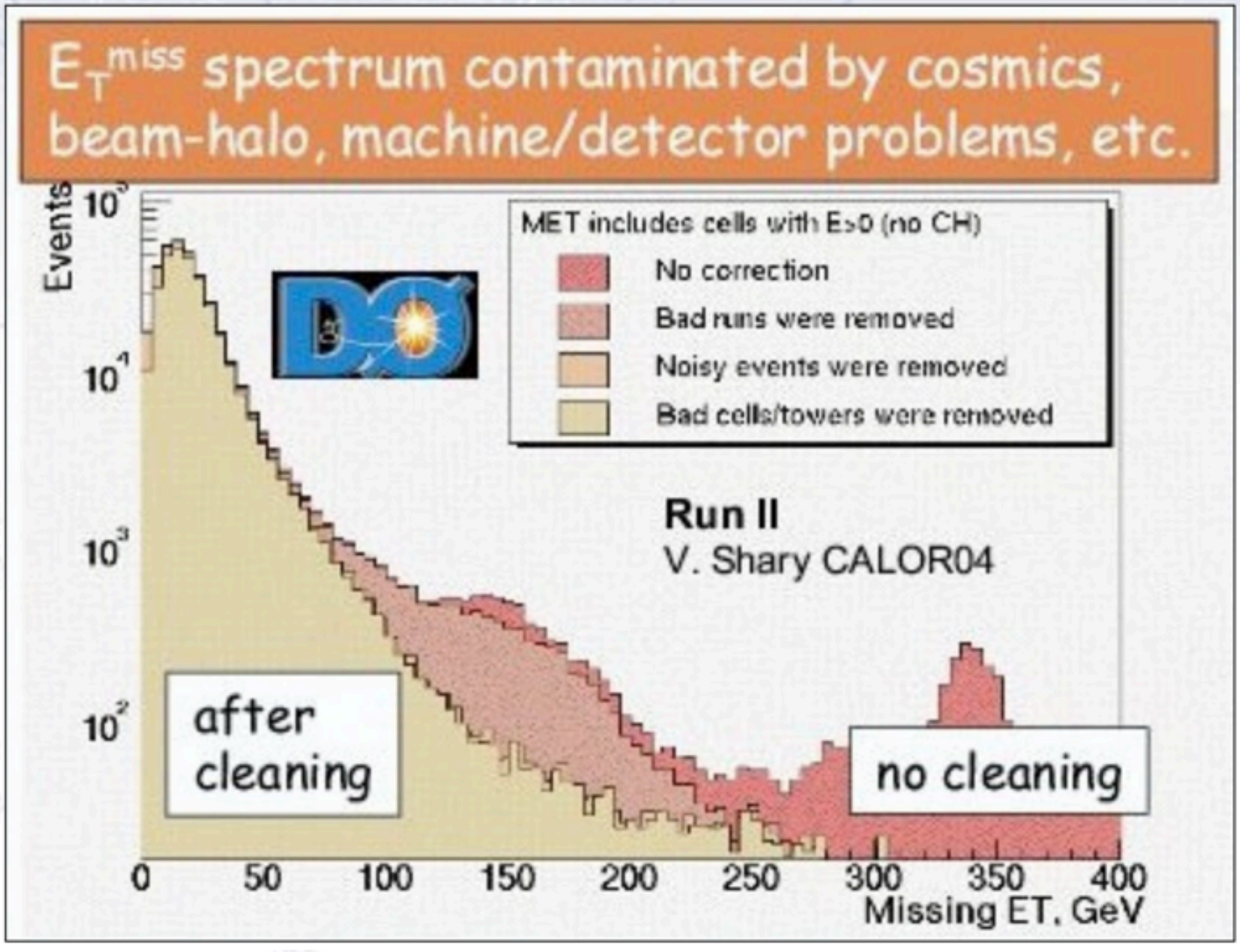
Emblem used by the BaBar experiment to label blinded analysis

This was first introduced by the French Academy of Science (1784), and has since become standard procedure in most science and medical experiments.

In this way experimenters bias is removed, and the results become truly independent and unaffected by wishful thinking and "common belief".

Cleaning data

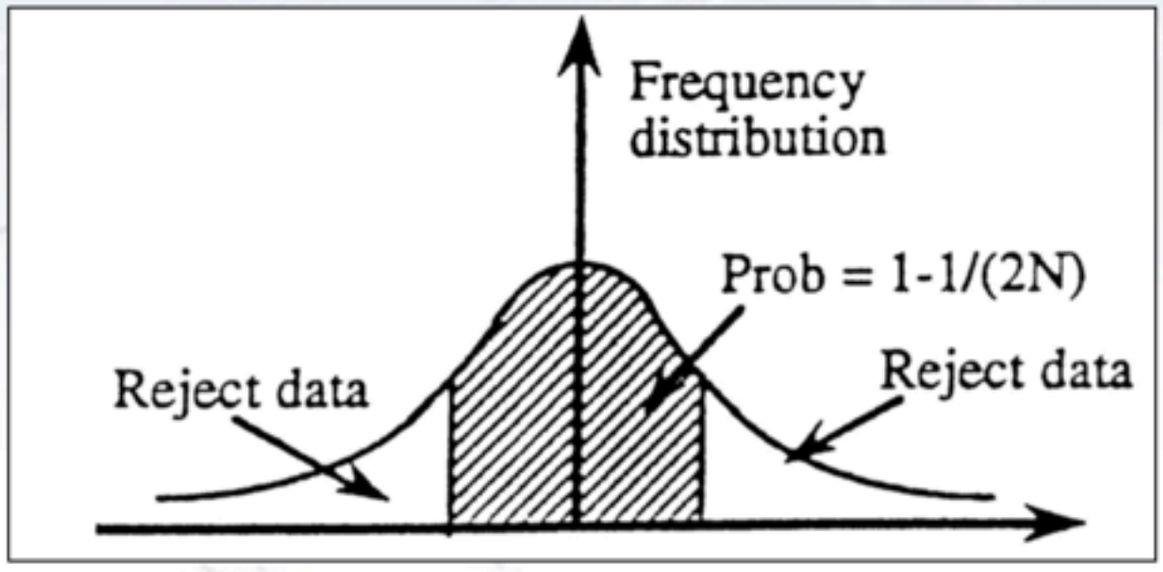
Example of experimental error, which would be a disaster if not corrected for.



Removing data points

One should always be careful about removing data points, yet at the same to be willing to do so, if very good arguments can be found:

- It is an error measurement.
- Measurement is improbable.



Removing improbable data points is formalised in **Chauvenet's Criterion,** though many other methods exists (Pierce, Grubbs, etc.)

The idea is to assume that the distribution is Gaussian, and ask what the probability of the farthest point is. If it is below some value (which is preferably to be determined ahead of applying the criterion), then the point is removed, and the criterion is reapplied until no more points should be removed.

However, **ALWAYS keep a record of your original data**, as it may contain more effects than you originally thought.

Example

Assume we have 10 measurements:

$$46,\ 48,\ 44,\ 38,\ 45,\ 47,\ 58,\ 44,\ 45,\ 43\,.$$

Calculate mean and std:

$$\bar{x} = 45.8 \quad \text{and} \quad \sigma_x = 5.1.$$

Now 58 looks suspicious - how bad is it?

$$t_{\text{sus}} = \frac{x_{\text{sus}} - \bar{x}}{\sigma_x} = \frac{58 - 45.8}{5.1} = 2.4.$$

And the probability is:

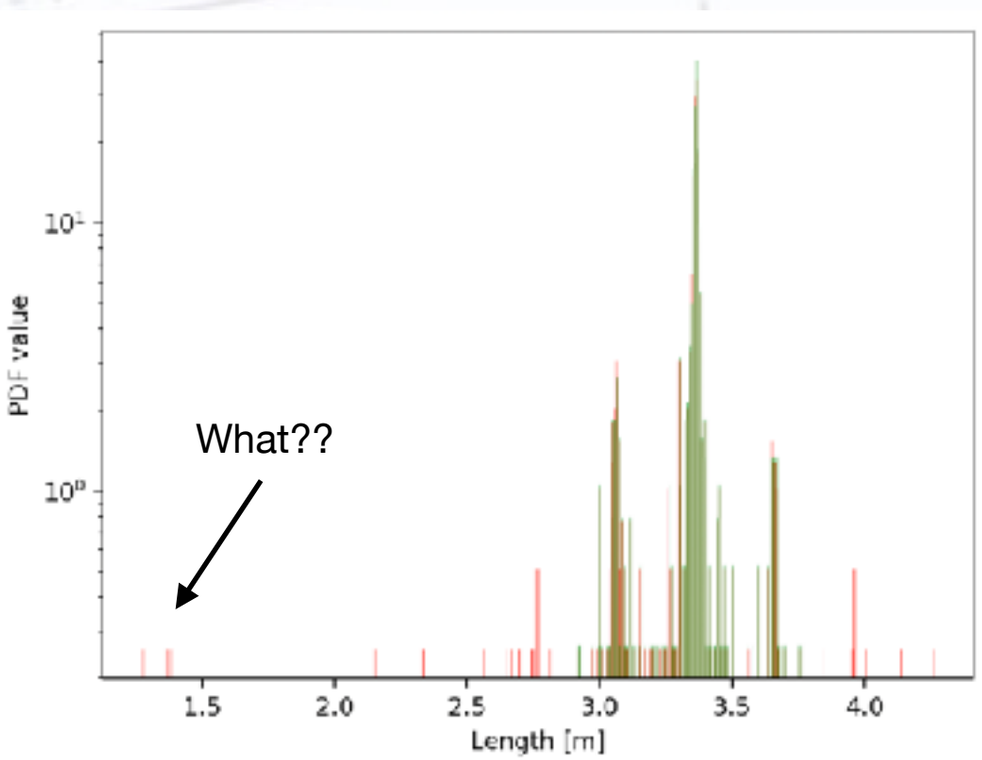
$$\begin{aligned} \textit{Prob}(\text{outside } 2.4\sigma) &= 1 - \textit{Prob}(\text{within } 2.4\sigma) \\ &= 1 - 0.984 \\ &= 0.016. \end{aligned}$$

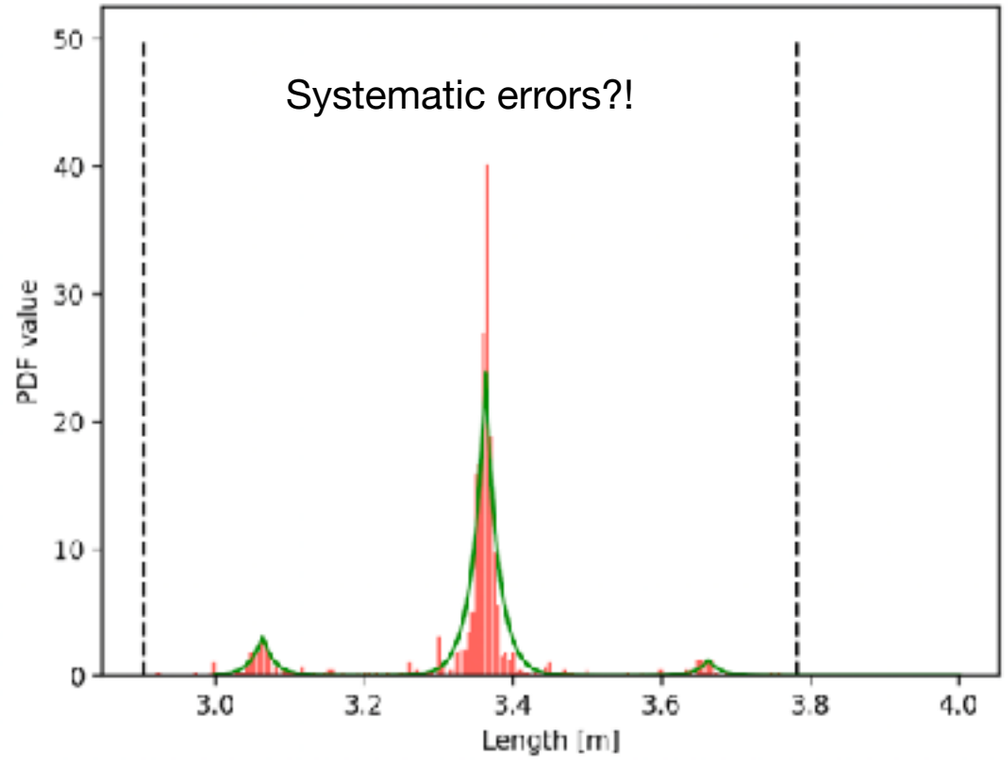
And according to Chauvenets criteria we reject - and recalculate mean and std.

Removing data points and systematic errors

A (in this sense messy and for training) very good data set is the one containing all the table measurements.

Here we have some clear outliers that should be removed - but also some systematic errors.





Example of bad data handling

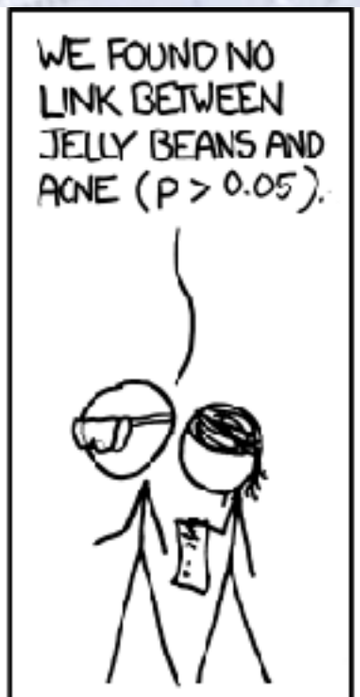
The result of our panel GLS model predicting the stringency of Covid-19 policies in OECD countries is shown in Table S3. We carefully test each independent variable by inserting them hierarchically to avoid collinearity.[6] The coefficients of Table S3 shows that the epidemiological baseline indicator (*death rate*) now strongly predicts the stringency of countries' policy adoptions (*stringency index*), as does population density. A stronger electoral democracy, on the other hand, weakens the stringency of measures adopted. Looking across the analysis of timing of adoption and stringency of adoption, we see that population density predicts NPI adoption that is both rapid and stringent. A higher score on the democracy index, on the other hand, predicts slower as well as less stringent adoption of NPIs.[7]

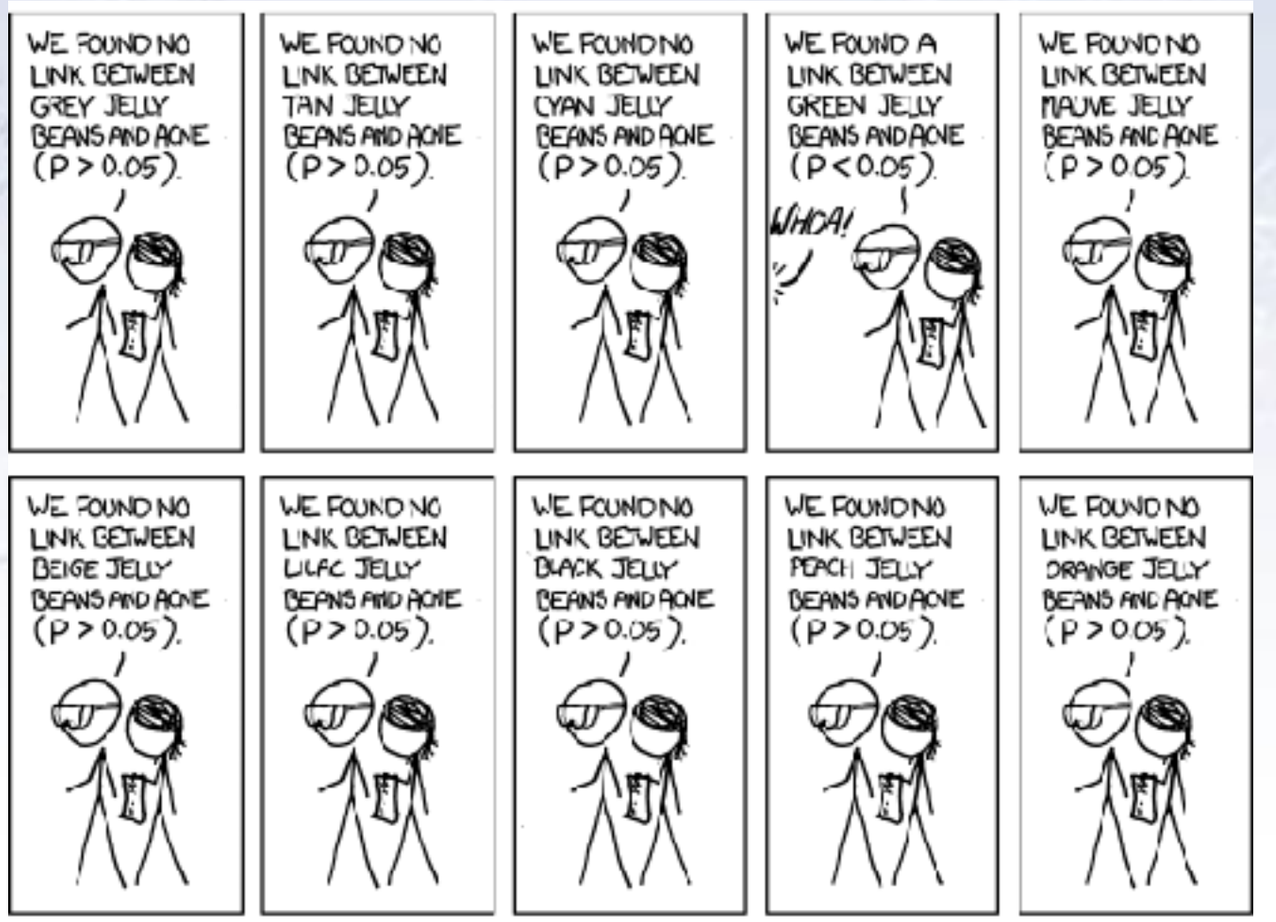
Table S3. Predicting the stringency of Covid-19 policies in OECD countries.

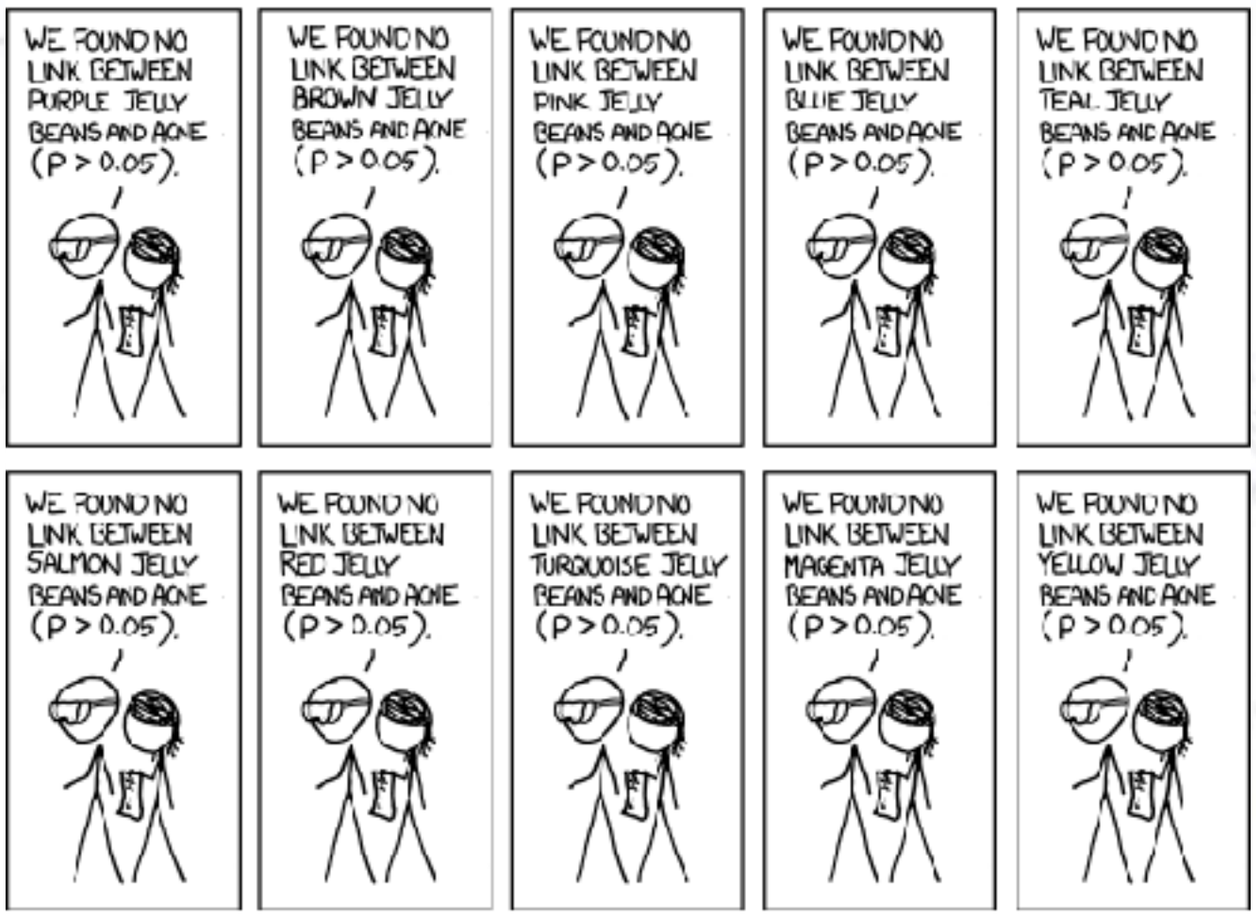
	Model 1	Model 2	Model 3
GDP per capita (log)	-3.715	4.445	3.457
	(7.339)	(6.759)	(7.239)
Tax revenue (% of GDP)	-0.041	-0.162	-0.081
	(0.580)	(0.541)	(0.553)
GINI index (income)	-0.776	-0.815	-0.814
	(0.672)	(0.601)	(0.618)
Hospital beds (/1,000 people)	0.527	-0.424	0.091
	(1.386)	(1.314)	(1.310)
Population age ≥65 (%)	-1.741*	-0.711	-0.821
	(0.773)	(0.754)	(0.788)
Urban population (%)	0.022	0.048	0.009
	(0.274)	(0.253)	(0.245)
Population density (log)	5.372**	5.905***	4.719*
	(2.055)	(1.792)	(1.844)
Death rate (/100,000)	11.701**		11.706**
	(3.615)		(3.610)
Electoral democracy		-0.709***	-0.748***
		(0.193)	(0.200)
Constant	98.528	58.126	75.504
	(87.690)	(76.766)	(83.036)
Observations	1,356	1,356	1,356
Number of countries	36	36	36
Within R-square	0.150	0	0.150
Between R-square	0.276	0.420	0.400
Overall R-square	0.208	0.123	0.240
Model chi-square	41.97	50.34	60.96
Root MSE	21.91	23.75	21.91

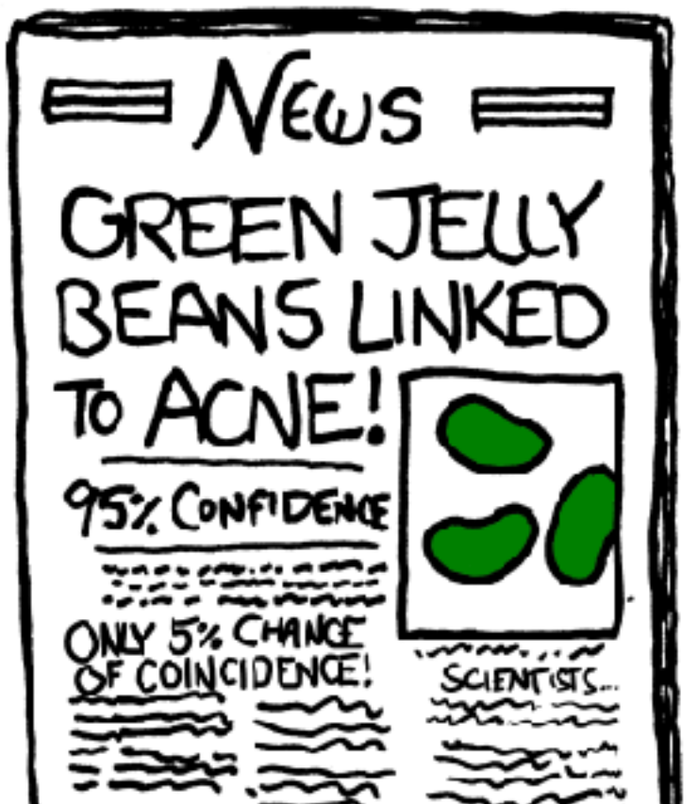
Note: Models specified as CLS Panel models with random effects. Stringency of policy adoption and death rate updated daily, all other variables constant. Measured daily January 15–March 30 2020. Standard errors in parentheses are clustered at the country level. *** p<0.001, ** p<0.01, * p<0.05, + p<0.10

XKCD on statistics

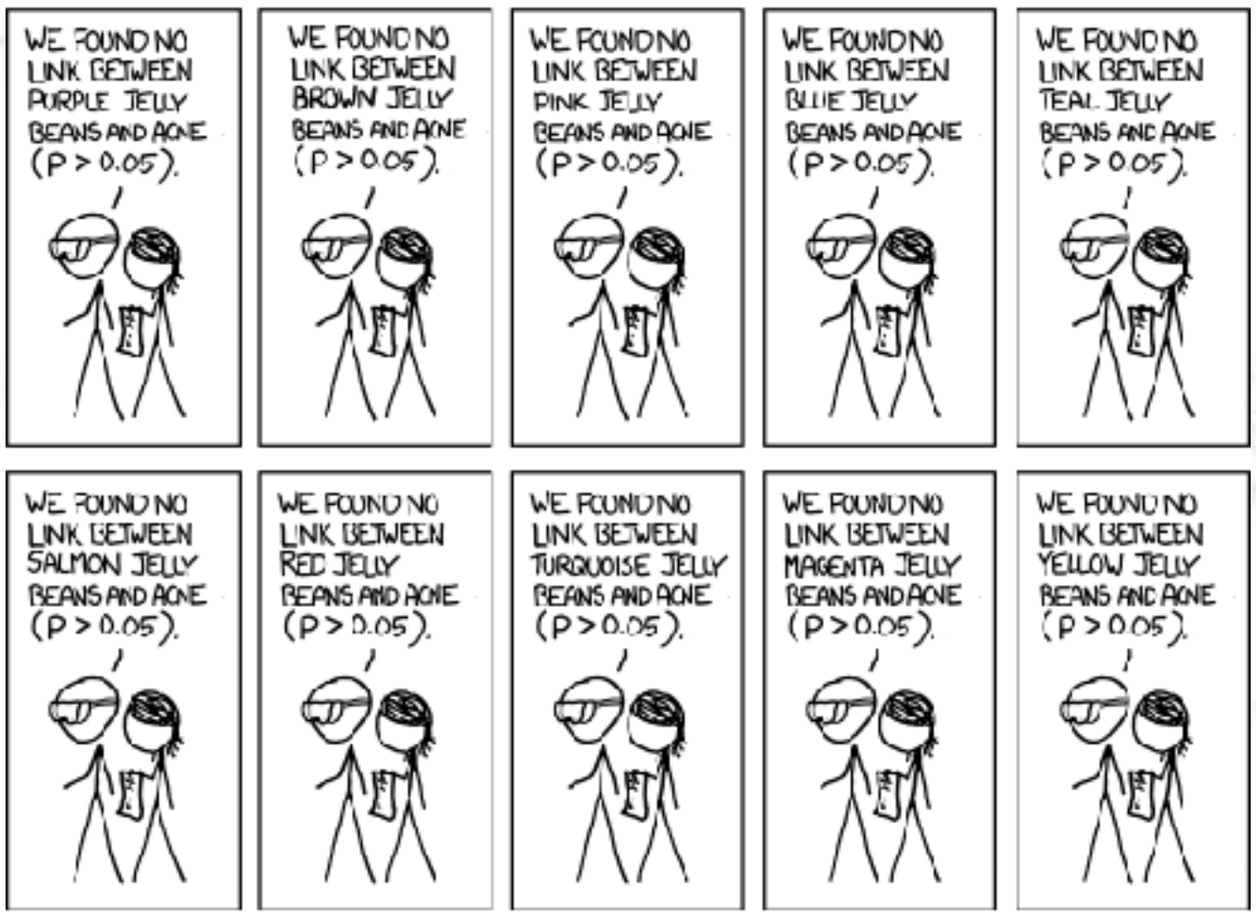


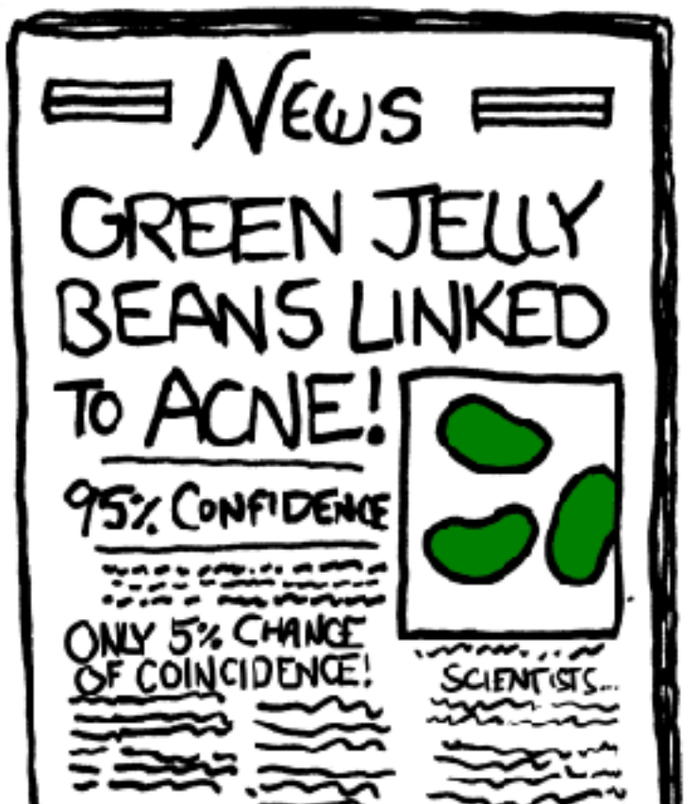




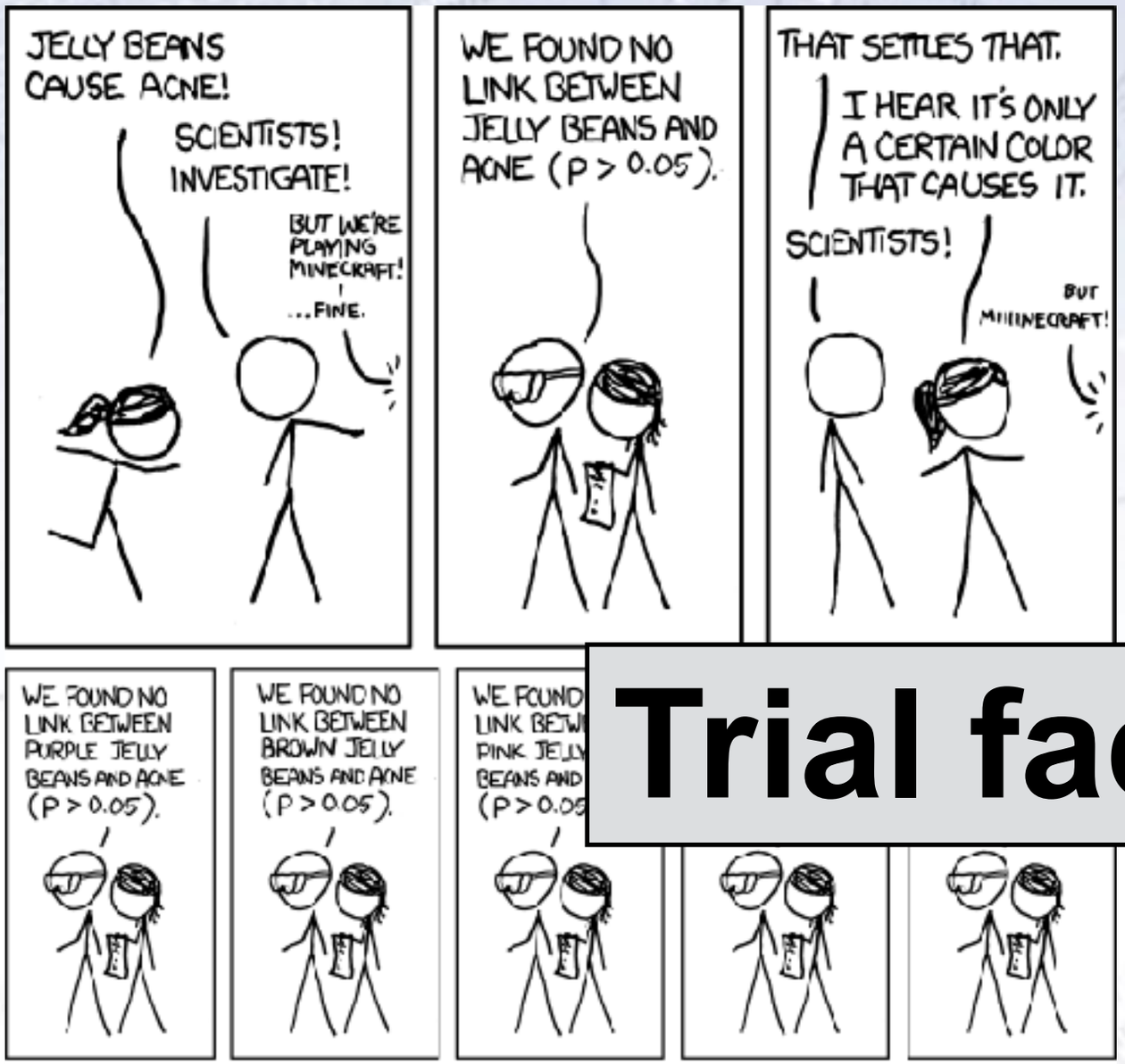


XKCD on statistics

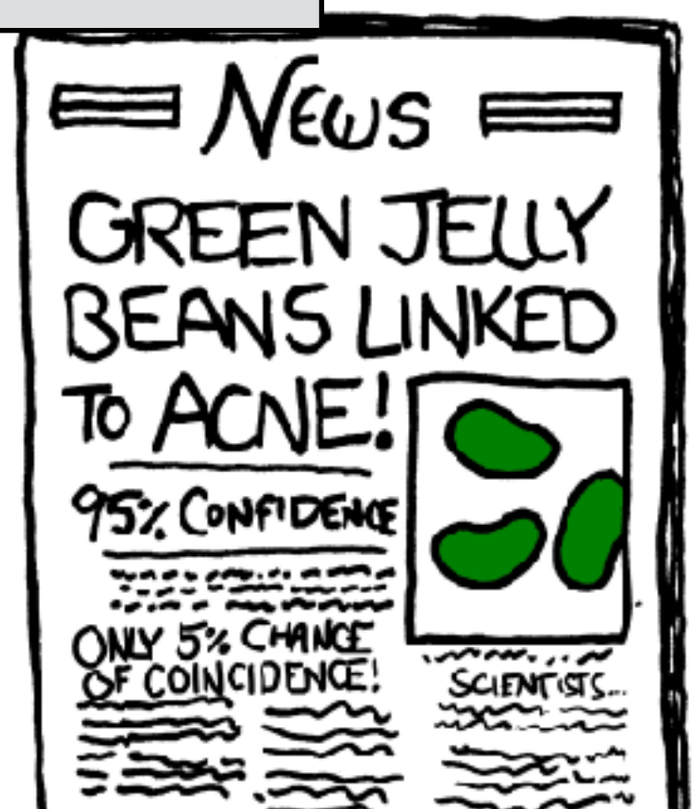




XKCD on statistics



Trial factor: 20



Trial factor / Look-Elsewhere Effect

"If you look enough times or places, you will find something unlikely"

The "Look-Elsewhere Effect" refers to observing an **apparent** statistically significant observation, which has arisen from searching a large parameter space (i.e. many places).

To account for this, one uses a **trial factor**, which is the ratio between the probability of observing a possible excess at some fixed point, to the probability of observing it anywhere in the range.

The significance of the (fitted) amplitude tells you the **local significance**. As you might be searching in many places, this reduces your certainty to the **global significance**:

$$p_{global} = 1 - (1 - p_{local})^N \simeq N \text{ local}$$

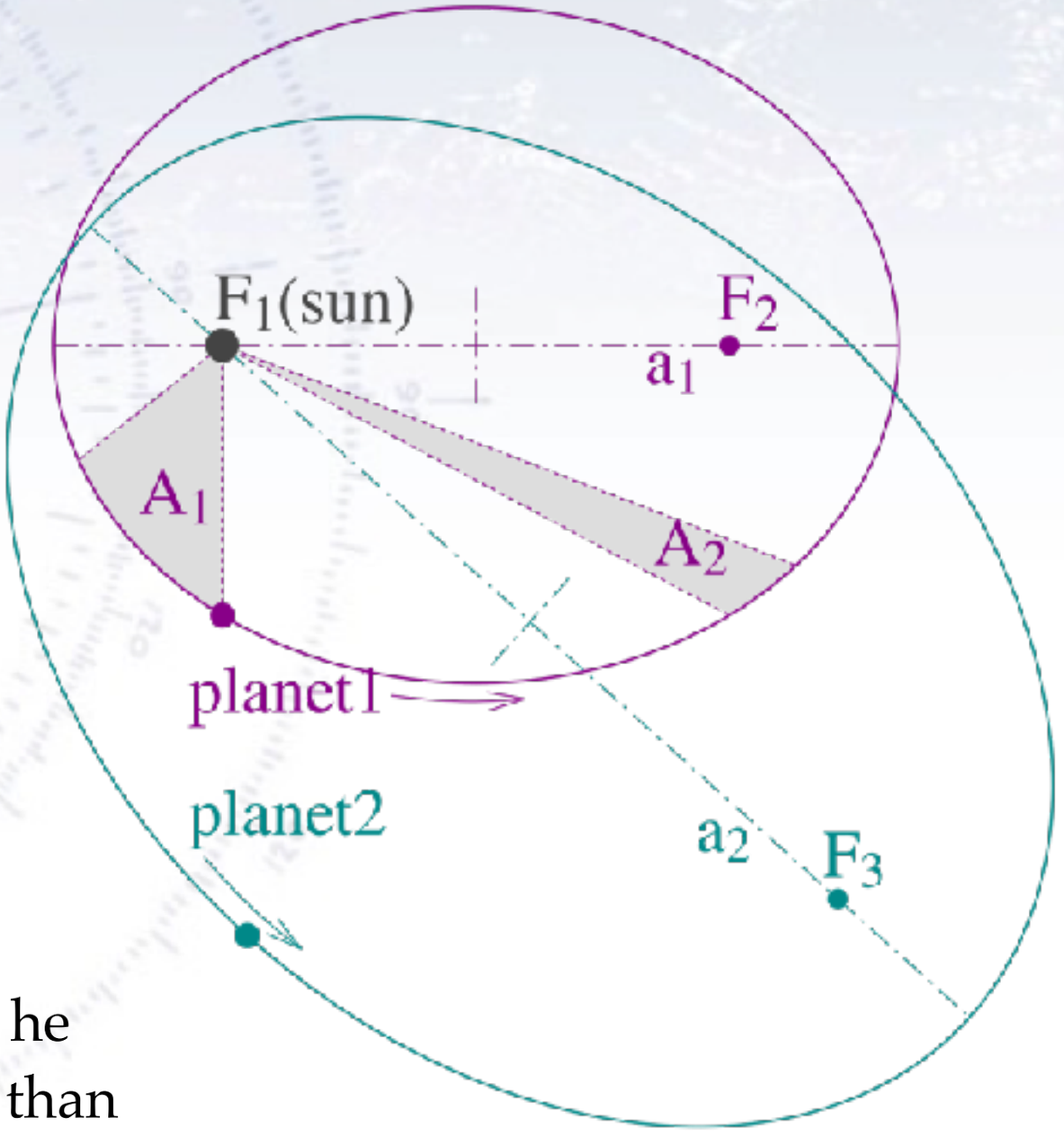
Thus, the global significance is (roughly) reduced by the trial factor.

A good paper with discussion of statistical treatment: https://arxiv.org/abs/1005.1891

When systematic errors are actually new discoveries…

Kepler tested the circular orbit model against Brahe's data and found discrepancies of up to 8 arcminutes

Brahe's methods were reliable enough to rule out measurement biases or calibration errors



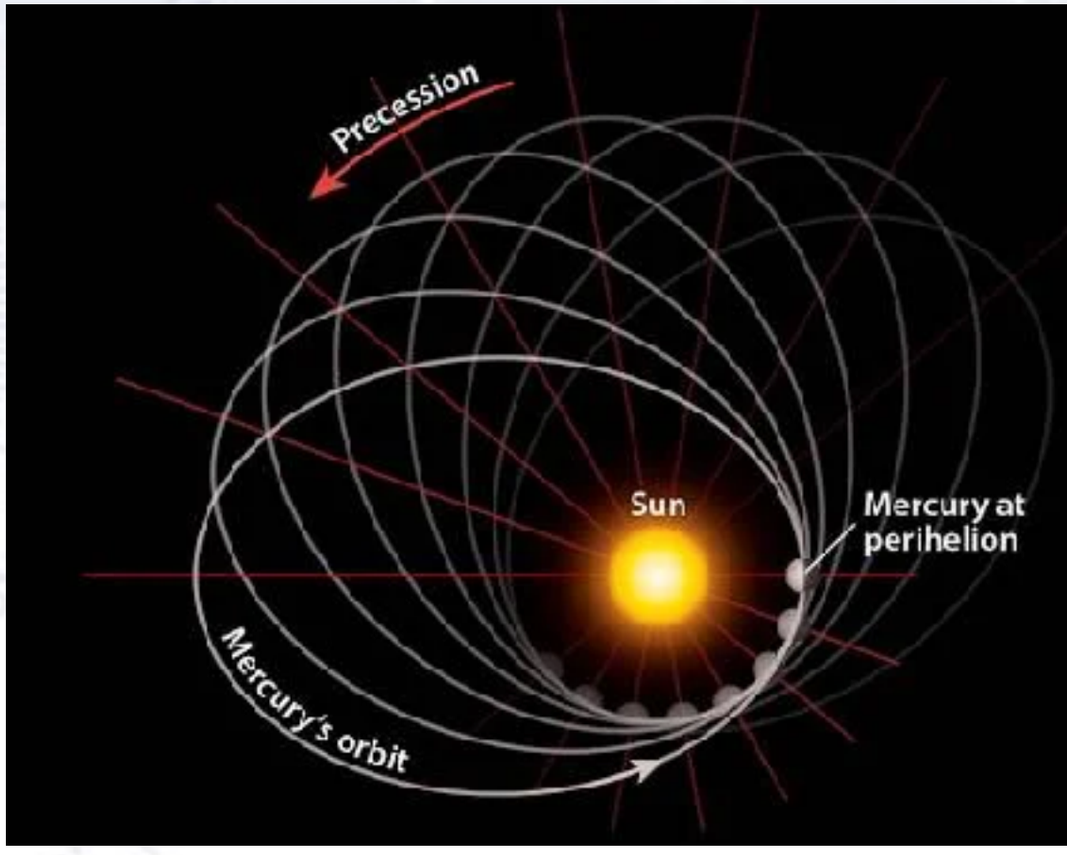
Eventually, through exhaustive calculations, he realized the orbit had to be an **ellipse** rather than a circle, with the Sun at one focus

When systematic errors are actually new discoveries…

Mercury's orbit showed a precession that could not be fully explained by Newtonian mechanics or known perturbations from other planets.

The discrepancy was attributed to systematic errors in observations or incomplete knowledge of celestial mechanics.

This "error" was later explained by Einstein's General Theory of Relativity which showed that spacetime curvature near the Sun caused the observed precession.



To work with systematic errors

Example of systematic error

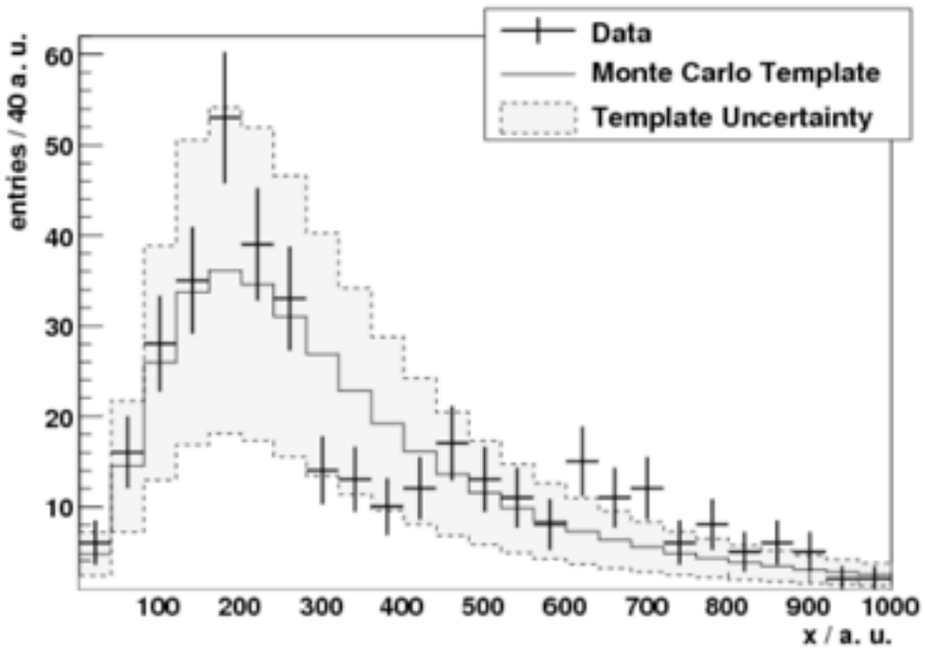
Measurements are taken with a steel ruler, the ruler was calibrated at 15°C, the measurements done at 22°C. This is a systematic **bias** and not only a systematic **uncertainty**! To neglect such an effect is a systematic **mistake**.

Effects can be corrected for! If the temperature coefficient and lab temperature is known (exactly), then there is no systematic uncertainty.

If we correct for effect, but corrections are not known exactly, then we have to introduce a systematic uncertainty (error propagation!).

A sign of a systematic error (or bug), is that one can see in data, that "something" strange is going on.

One should of course work hard to understand the effect, but occasionally one must give up, and suffer a large systematic uncertainty.



Evaluating systematic errors

Known sources:

- Error on factors in the analysis, energy calibration, efficiencies, corrections, ...
- Error on external input: theory error, error on temperature, masses, ...

Evaluate from varying conditions, and compute result for each. Error is RMSE.

Unsuspected sources:

Repeating the analysis in different form helps to find such systematic effects.

- Use subset of data, or change selection of data used in analysis.
- Change histogram binning, change parameterisations, change fit techniques.
- Look for impossibilities.

If you do not a priori expect a systematic effect and if the deviation is not significant, then do not add this in the systematic error.

If there is a deviation, try to understand, where the mistake is and fix it!

Only as a last resort include non-understood discrepancy as systematic error.

Take-home messages

Systematic errors present a serious challenge - Cross checks and tedious investigations is your best ally

Bias of results is a typical human mistake - blinding of results solves this issue

Removing data points is a delicate issue - Chauvenets criterion presents one way to deal with this (but it is not a law of nature!)

End of presentation

Example of bad data handling

KOMMENTARER

Nedlukninger har kun reddet fem danskere fra døden

Effekten af nedlukningerne på antal døde er yderst beskeden. Uanset, hvor travlt Mette Frederiksen har haft med at redde danskernes liv.

Forskere kritiserer: Nej, ny analyse kan ikke konkludere, at nedlukninger kun har reddet fem danske liv

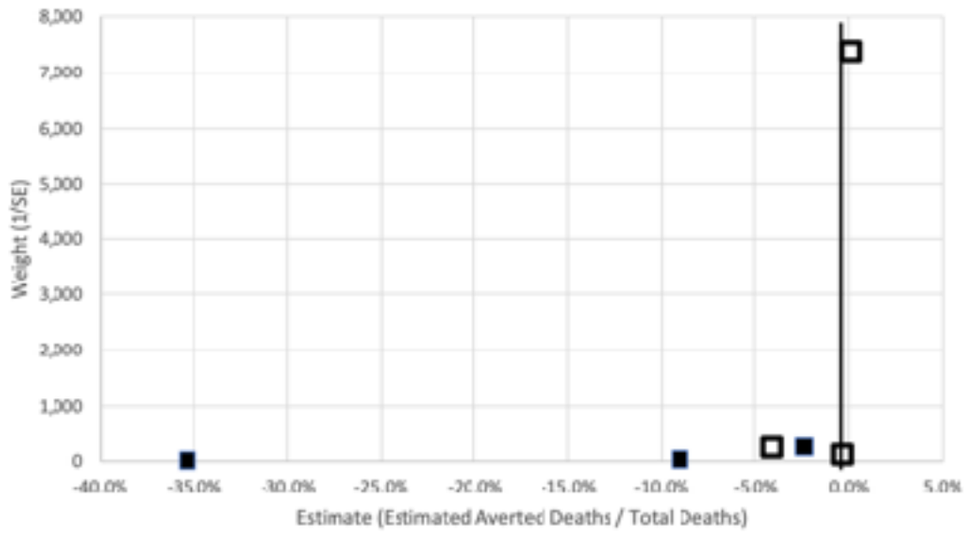
»Jeg ville dumpe mine studerende, hvis de havde lavet den,« lyder det fra en forsker om analysen, som blandt andet CEPOS har stået bag.

Example of bad data handling

Table 3: Overview of common estimates from studies based on stringency indexes

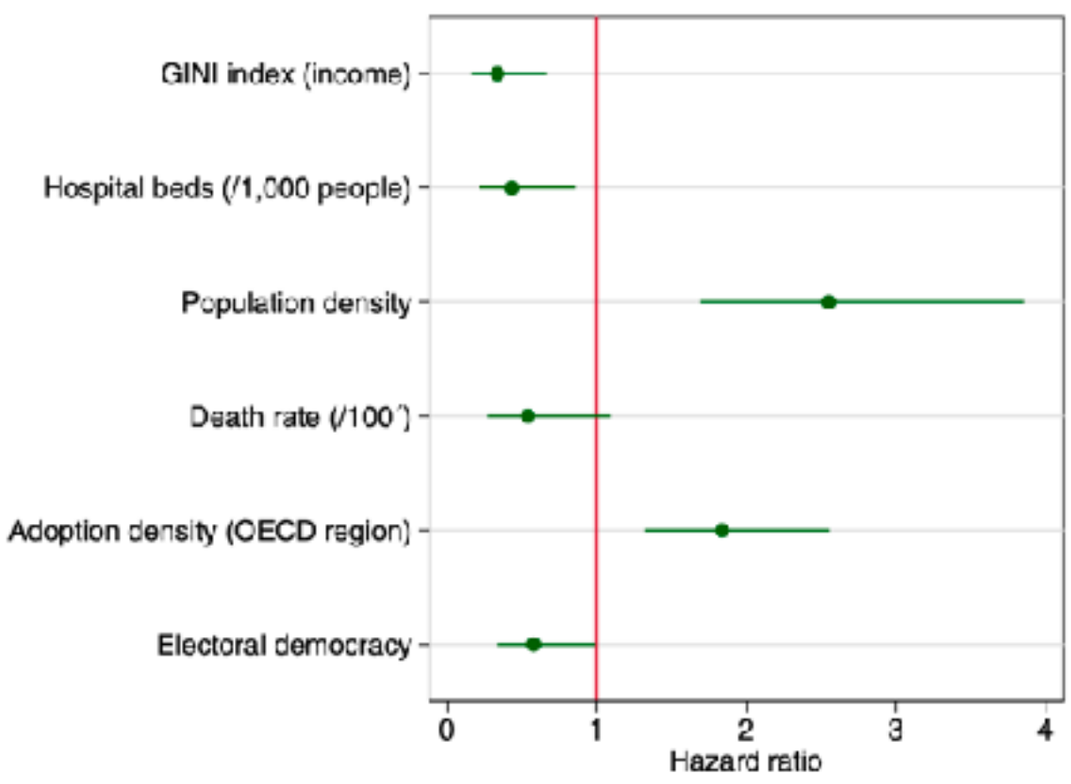
Effect on COVID-19 mortality	Estimate (Estimated Averted Deaths / Total Deaths)	Standard error	Weight (1/SE)	Quality dimensions
Bjørnskov (2021)	-0.3%	0.8%	119	3
Shiva and Molana (2021)	-4.1%	0.4%	248	4
Stockenhuber (2020)*	0.0%	n/a	n/a	3
Chisadza et al. (2021)	0.1%	0.0%	7,390	4
Goldstein et al. (2021)	-9.0%	3.8%	26	2
Fuller et al. (2021)	-35.3%	9.1%	11	2
Ashraf (2020)	-2.4%	0.4%	256	2
Precision-weighted average (arithmetic average / median)	**-0.2% (-7.3%/-2.4%)**			

Figure 5: Funnel plot for estimates from studies based on stringency indexes



Example of bad data handli

Several studies explicitly claim that they estimate the actual causal relationship between lockdowns and COVID-19 mortality. Some studies use instrumental variables to justify the causality associated with their analysis, while others make causality probable using anecdotal evidence.[25] But, Sebhatu et al. (2020) show that government policies are strongly driven by the policies initiated in neighboring countries rather than by the severity of the pandemic in their own countries. In short, it is not the severity of the pandemic that drives the adoption of lockdowns, but rather the propensity to copy policies initiated by neighboring countries. The Sebhatu et al. conclusion throws into doubt the notion of a causal relationship between lockdowns and COVID-19 mortality.



Example of bad data handli

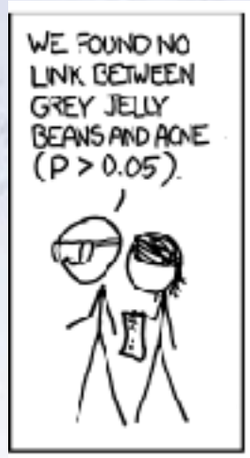
The result of our panel GLS model predicting the stringency of Covid-19 policies in OECD countries is shown in Table S3. We carefully test each independent variable by inserting them hierarchically to avoid collinearity.[6] The coefficients of Table S3 shows that the epidemiological baseline indicator (*death rate*) now strongly predicts the stringency of countries' policy adoptions (*stringency index*), as does population density. A stronger electoral democracy, on the other hand, weakens the stringency of measures adopted. Looking across the analysis of timing of adoption and stringency of adoption, we see that population density predicts NPI adoption that is both rapid and stringent. A higher score on the democracy index, on the other hand, predicts slower as well as less stringent adoption of NPIs.[7]

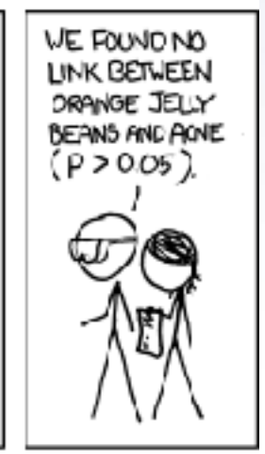
Table S3. Predicting the stringency of Covid-19 policies in OECD countries.

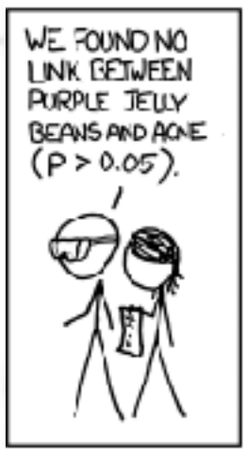
	Model 1	Model 2	Model 3
GDP per capita (log)	-3.715	4.445	3.457
	(7.339)	(6.759)	(7.239)
Tax revenue (% of GDP)	-0.041	-0.162	-0.081
	(0.580)	(0.541)	(0.553)
GINI index (income)	-0.776	-0.815	-0.814
	(0.672)	(0.601)	(0.618)
Hospital beds (/1,000 people)	0.527	-0.424	0.091
	(1.386)	(1.314)	(1.310)
Population age ≥65 (%)	-1.741*	-0.711	-0.821
	(0.773)	(0.754)	(0.788)
Urban population (%)	0.022	0.048	0.009
	(0.274)	(0.253)	(0.245)
Population density (log)	5.372**	5.905***	4.719*
	(2.055)	(1.792)	(1.844)
Death rate (/100,000)	11.701**		11.706**
	(3.615)		(3.610)
Electoral democracy		-0.709***	-0.748***
		(0.193)	(0.200)
Constant	98.528	58.126	75.504
	(87.690)	(76.766)	(83.036)
Observations	1,356	1,356	1,356
Number of countries	36	36	36
Within R-square	0.150	0	0.150
Between R-square	0.276	0.420	0.400
Overall R-square	0.208	0.123	0.240
Model chi-square	41.97	50.34	60.96
Root MSE	21.91	23.75	21.91

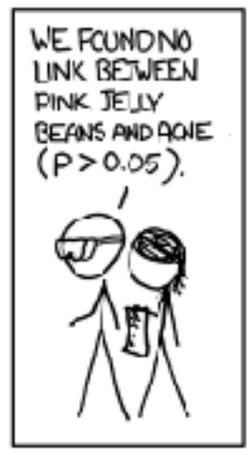
Note: Models specified as CLS Panel models with random effects. Stringency of policy adoption and death rate updated daily, all other variables constant. Measured daily January 15–March 30 2020. Standard errors in parentheses are clustered at the country level. *** p<0.001, ** p<0.01, * p<0.05, + p<0.10

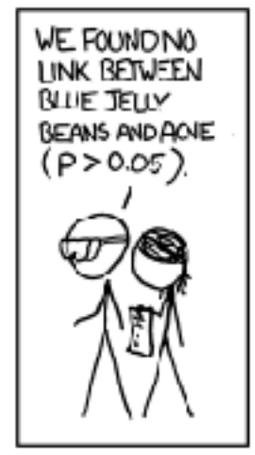
XKCD on statistics

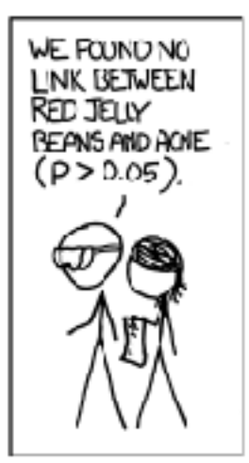


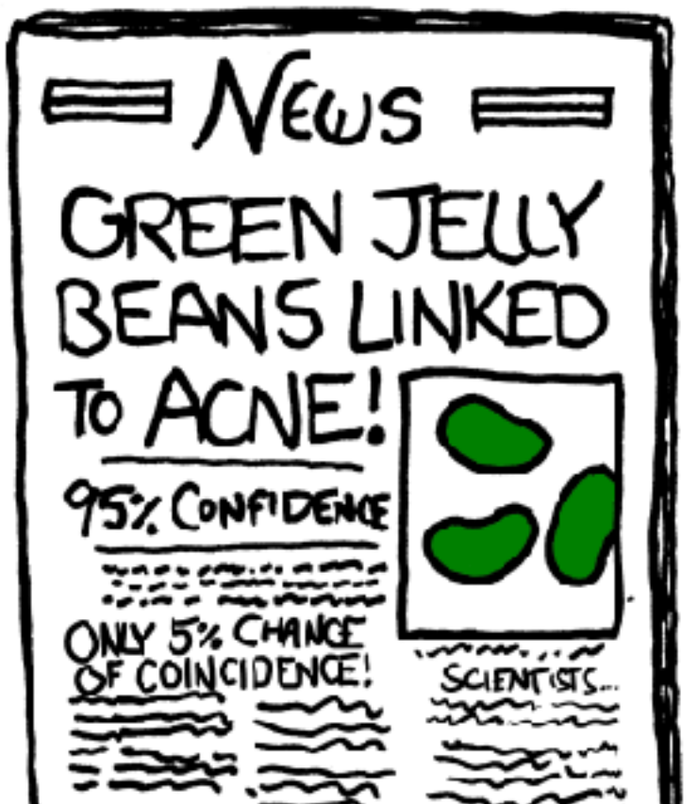




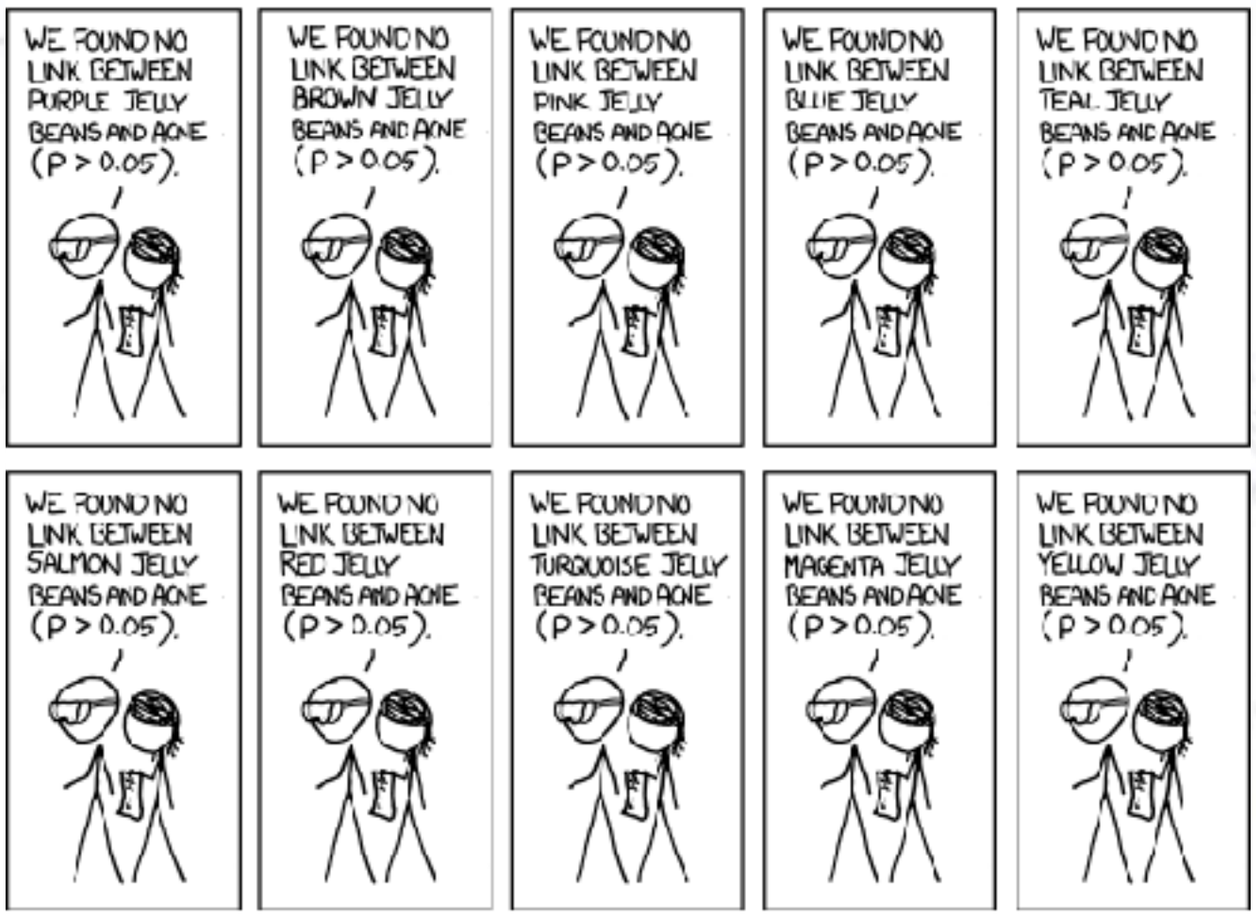


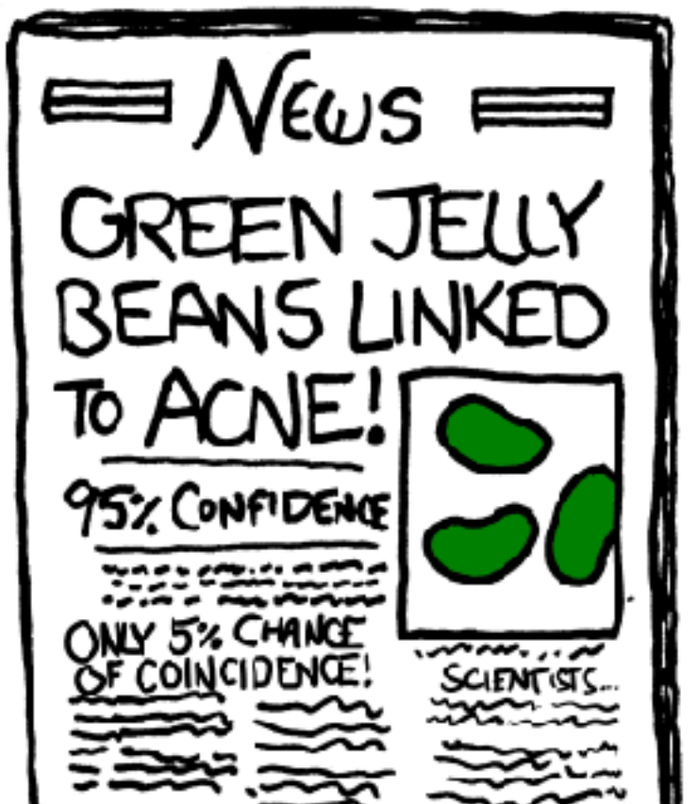




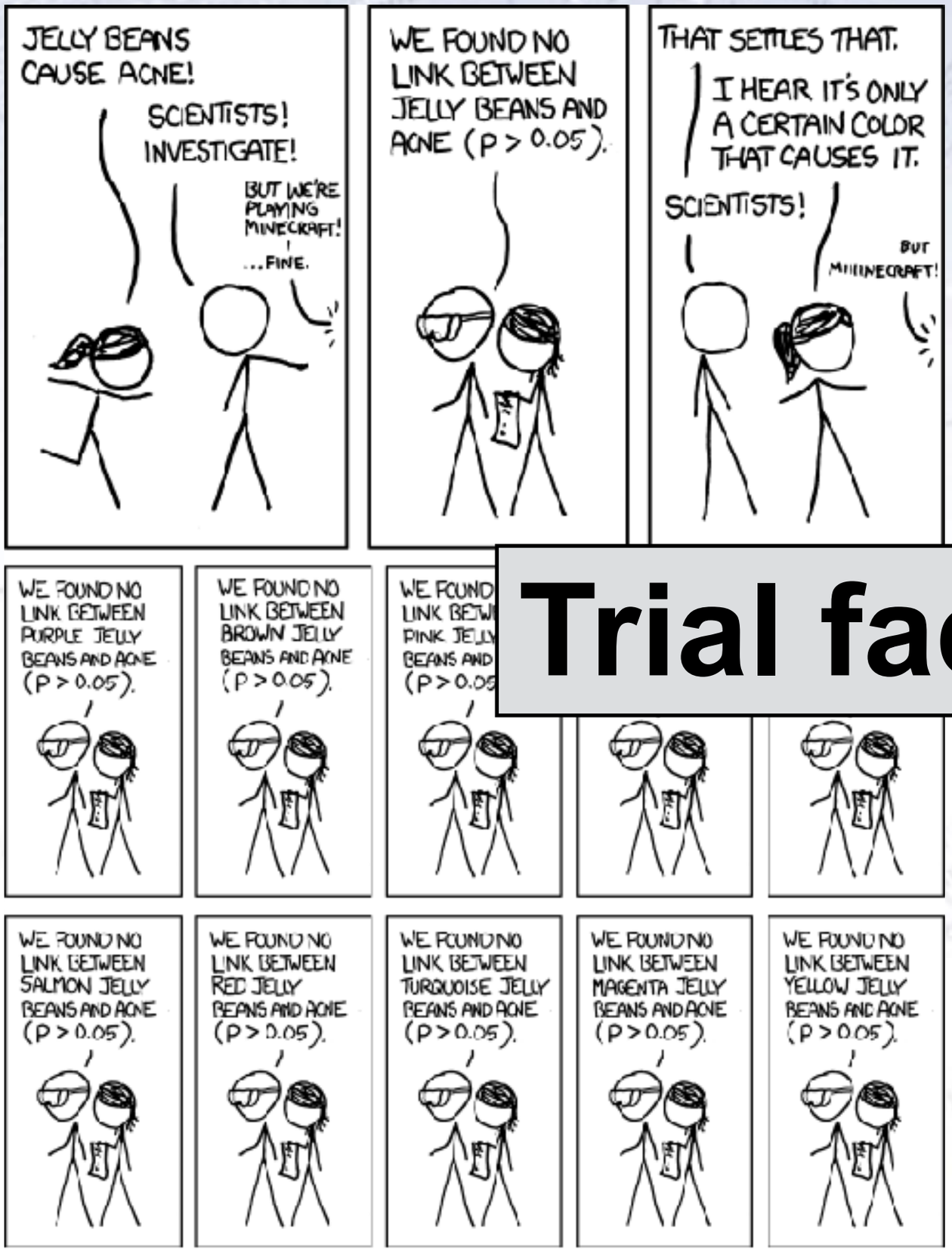


XKCD on statistics

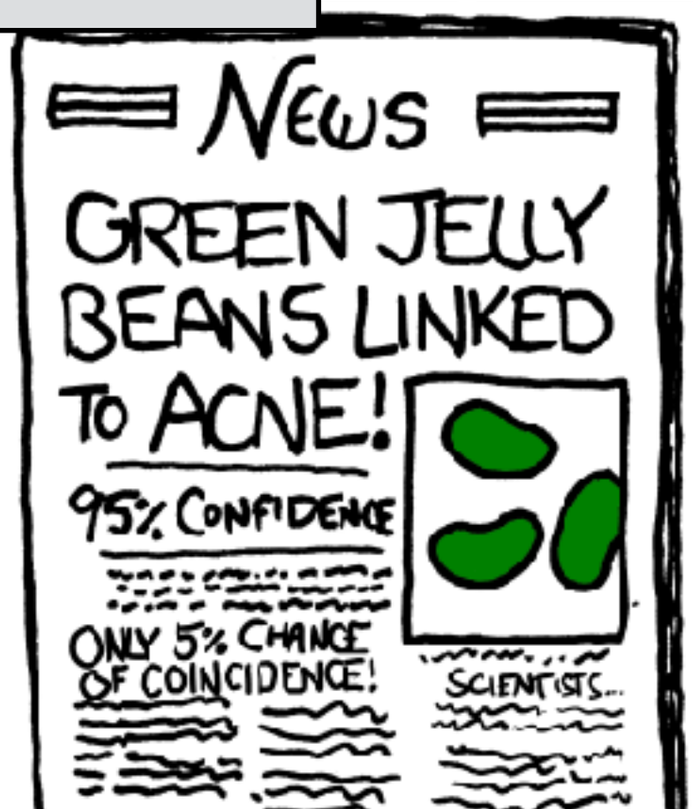




XKCD on statistics



Trial factor: 20



Trial factor / Look-Elsewhere Effect

"If you look enough times or places, you will find something unlikely"

The "Look-Elsewhere Effect" refers to observing an **apparent** statistically significant observation, which has arisen from searching a large parameter space (i.e. many places).

To account for this, one uses a **trial factor**, which is the ratio between the probability of observing a possible excess at some fixed point, to the probability of observing it anywhere in the range.

The significance of the (fitted) amplitude tells you the **local significance**. As you might be searching in many places, this reduces your certainty to the **global significance**:

$$p_{global} = 1 - (1 - p_{local})^N \simeq N \text{ local}$$

Thus, the global significance is (roughly) reduced by the trial factor.

A good paper with discussion of statistical treatment: https://arxiv.org/abs/1005.1891

Take-home messages

Systematic errors present a serious challenge - Cross checks and tedious investigations is your best ally

Bias of results is a typical human mistake - blinding of results solves this issue

Removing data points is a delicate issue - Chauvenets criterion presents one way to deal with this (but it is not a law of nature!)

Look-elsewhere effect is a typical mistake that should be avoided by scaling over the number of cases you are looking at

Example case

A Swedish study in 1992 tried to determine whether or not power lines caused some kind of poor health effects. The researchers surveyed everyone living within 300 m of high-voltage power lines over a 25-year period and looked for statistically significant increases in rates of **over 800 ailments**.

The study found that the incidence of childhood leukemia was four times higher among those that lived closest to the power lines, and it spurred calls to action by the Swedish government.

Example case

A Swedish study in 1992 tried to determine whether or not power lines caused some kind of poor health effects. The researchers surveyed everyone living within 300 m of high-voltage power lines over a 25-year period and looked for statistically significant increases in rates of **over 800 ailments**.

The study found that the incidence of childhood leukemia was four times higher among those that lived closest to the power lines, and it spurred calls to action by the Swedish government.

The problem with the conclusion, however, was that they failed to compensate for the **look-elsewhere effect**; in any collection of 800 random samples, it is likely that at least one will be at least 3 standard deviations above the expected value, **by chance alone**. Subsequent studies failed to show any links between power lines and childhood leukemia, neither in causation nor even in correlation.

[Jon Palfreman, "Currents of fear" (1995-06-13), Frontline, PBS,]

De-trending algorithms

A typical example is in time series analyses where would like a process to be stationary. This can be done by applying different kinds of filters.

Of particular importance is the Polynomial filter

$$\hat{\vec{\theta}} = (\tilde{\mathbf{C}}\mathbf{V}^{-1}\mathbf{C})^{-1}\tilde{\mathbf{C}}\mathbf{V}^{-1}\mathbf{y}$$

$$\mathbf{C} = \begin{pmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ \vdots & \vdots & \vdots \\ 1 & x_n & x_n^2 \end{pmatrix} \qquad \mathbf{V} = \begin{pmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_n^2 \end{pmatrix}$$