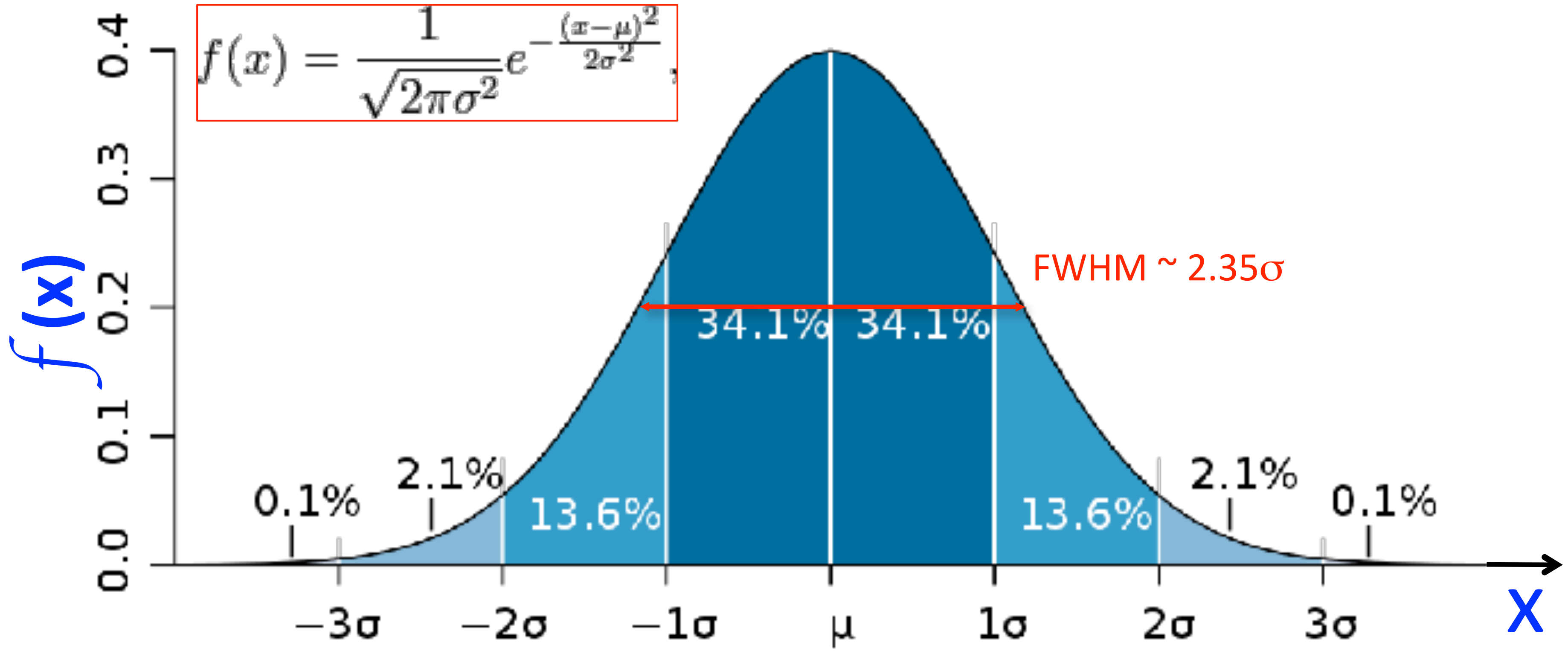
Lecture 2:

Trials and Errors

- Properties of Normal Distributions
- Trials and Tribulations!
- Regression to the Mean
- Correlations
- Uncertainties & Error Propagation



"Two-Sided"

Chance Prob. To Be Within:

- ±1 σ : 0.6827
- ±2 σ : 0.9545
- ±3 σ : 0.9973
- ±4 σ : 0.99994

Chance Prob. To Be Outside:

- ±1 σ : 0.3173
- ±2 σ : 0.0455
- ±3 σ : 0.0027
- ±4 σ : 6×10^{-5}

***e.g.* "Do my data points look ok relative to the model?"**

"One-Sided"

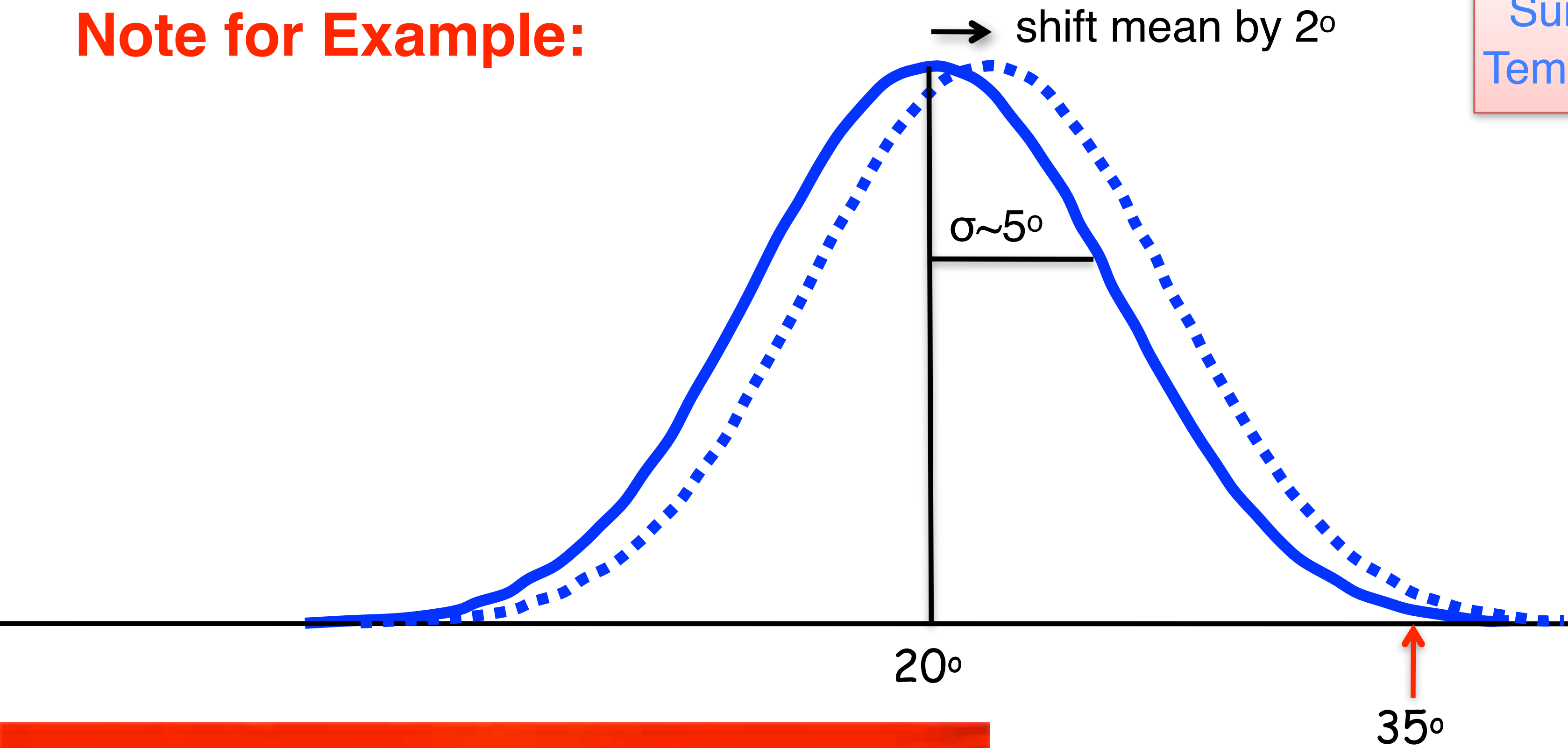
Chance Prob. To Be:

- >1 σ : 0.1587
- >2 σ : 0.0228
- >3 σ : 1.35×10^{-3}
- >4 σ : 3×10^{-5}

***e.g.* "What's the chance of seeing an excess at least this large?"**

Note for Example:

Summer Temperature



$$20 + 15(3\sigma) \rightarrow 22 + 13(2.6\sigma)$$

$$P(> 35^o) : 0.0013 \rightarrow 0.0047$$

(i.e. nearly quadruples, with bigger changes further into the tail)

As a consequence of the Central Limit Theorem, many (but not all!) physical processes often tend towards the Normal Distribution shape.

However, few achieve this exactly !!

Although calculated probabilities are often couched in terms of an ideal Normal Distribution to give a rough intuition of the scaling

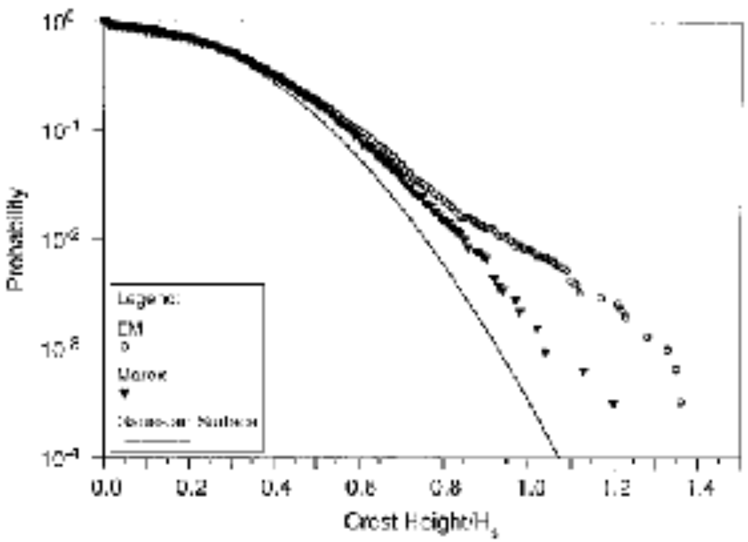
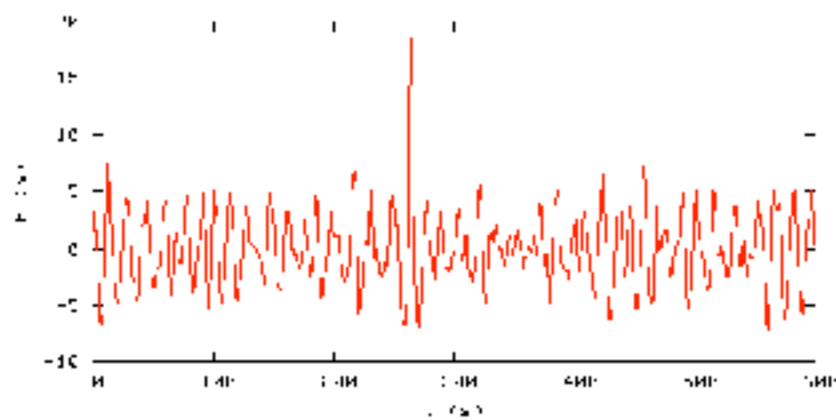


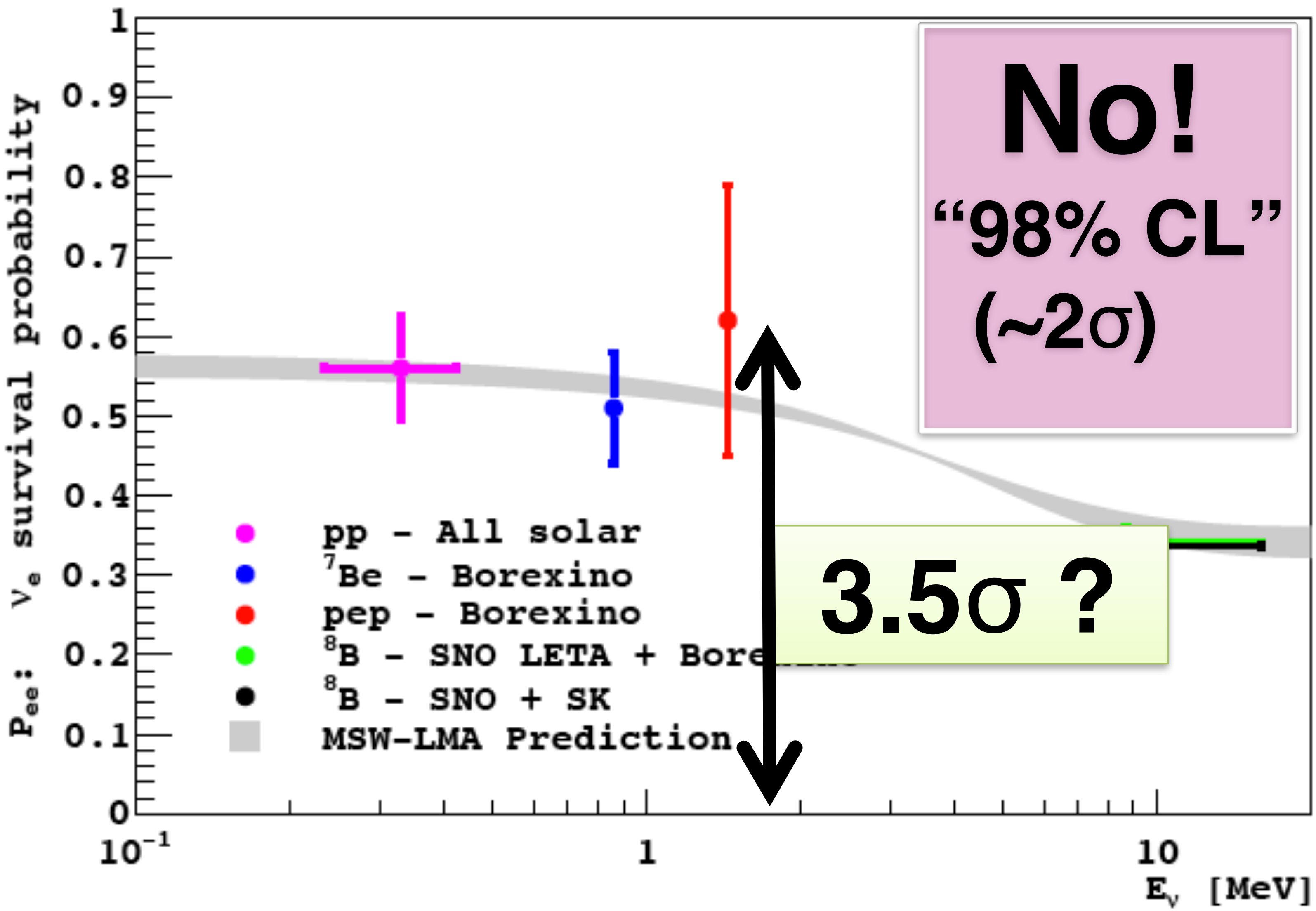
FIG. 4. Probability distribution of normalized crest heights measured at Tern during the storm on 4 Jan 1993. The crest heights are normalized by the significant wave height during each hour of the measurements. Nine hours of measurements with an average significant wave height of about 12 m were combined to produce the observed distribution.



The Draupner wave, a single giant wave measured on New Year's Day 1995, finally confirmed the existence of freak waves, which had previously been considered near-mythical

Beware the Distribution Tails!!

Borexino "1 σ error" on solar pep flux (2011)



Example: Search for Episodic X-Ray Emission

Over the course of a year, 36000 x-rays are observed to come from a particular astrophysical object. However, on one particular day, 130 events are observed. What is the statistical significance of this observed burst?

$$\langle x \rangle = \frac{36000}{365} = 98.6 \qquad \mu \simeq \langle x \rangle \qquad \sigma = \sqrt{\mu}$$

$$s \cong \frac{(130 - 98.6)}{\sqrt{98.6}} = 3.16\sigma$$

odds of getting at least this many events by a chance fluctuation from the average rate of emission

$$P = 8 \times 10^{-4}$$

Is this sufficient to claim the observation of a burst from this object?



20 1 18 4 13 6 10 15 2 17 3 19 7 16 8 11 14 9 12 5

Correct question:

What is the chance of seeing at least one burst with an excess at least as large given the number of independent tests I've done ?

Binomial !!

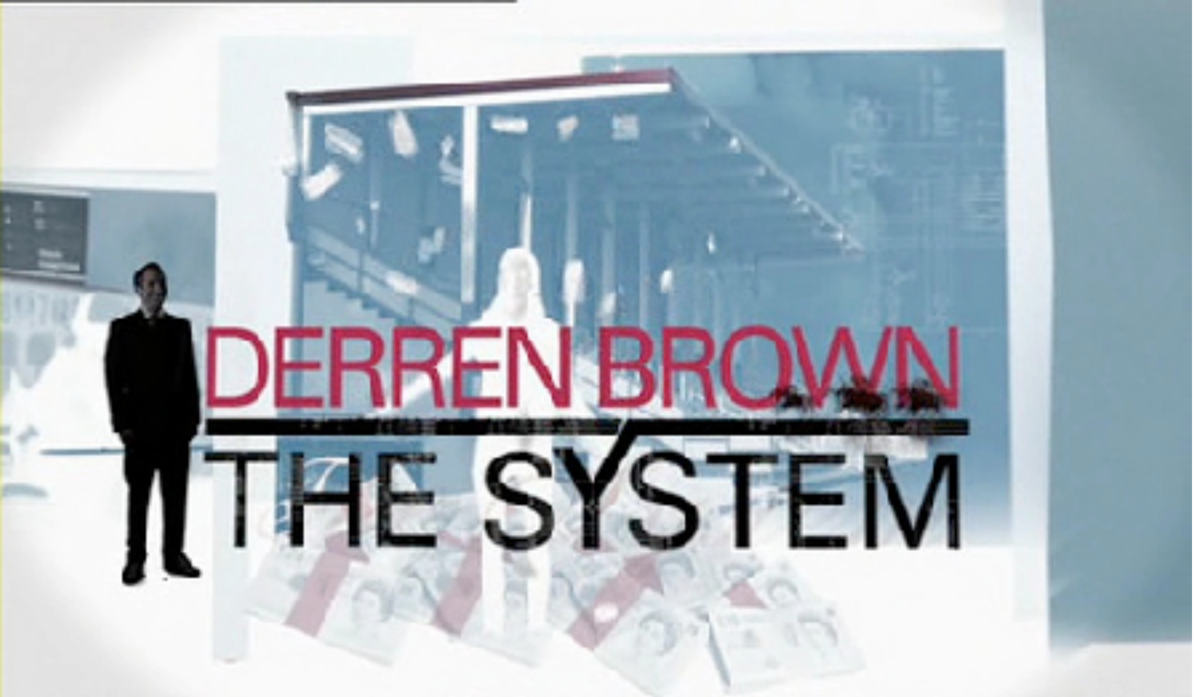
N Bernoulli trials where the chance of each success is **P**

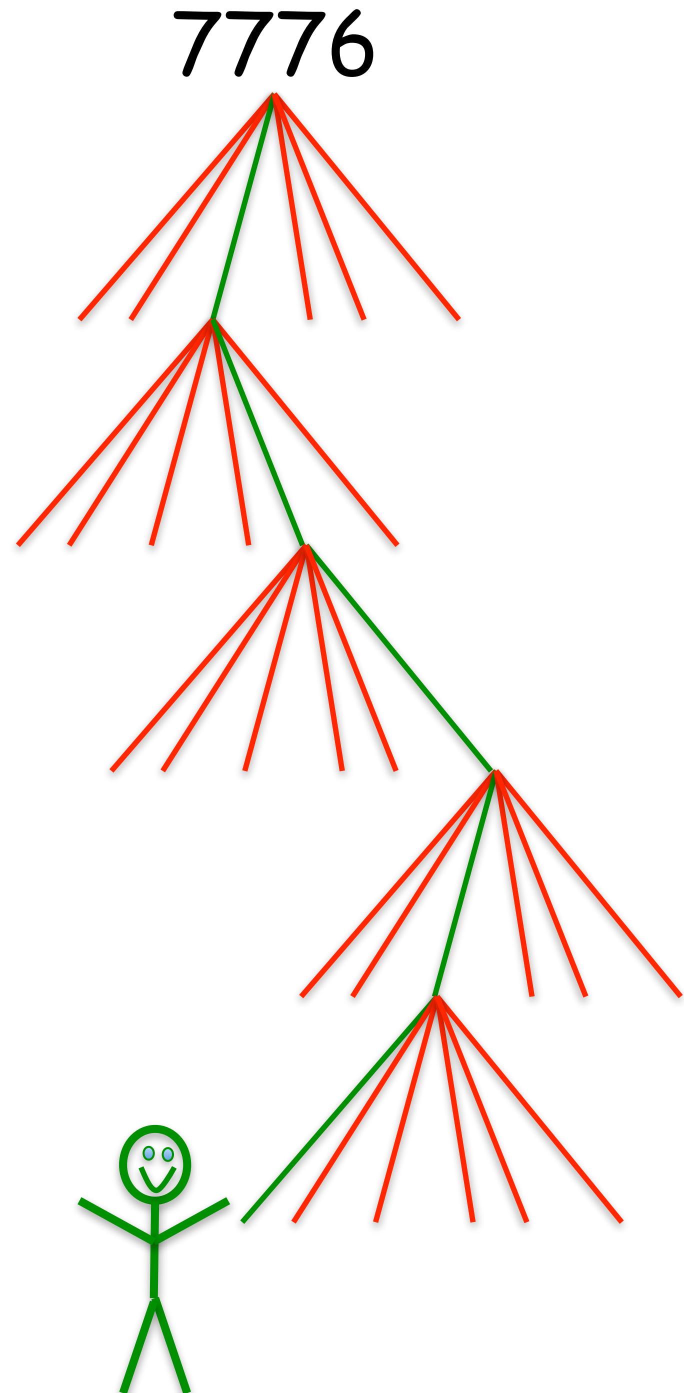
$$\sum_{i=1}^{\infty} \binom{N}{i} P^i (1-P)^{N-i} = 1 - \binom{N}{0} P^0 (1-P)^{N-0}$$

$$\mathrm{P}_{\text{post-trial}} = 1 - (1-P)^N \qquad (\sim NP \quad \text{for} \quad NP << 1)$$

$\mathrm{P} = 8 \times 10^{-4}$, $\mathrm{N} = 365$ → $\mathrm{P}_{\text{post-trial}} = 25\%$

How many timescales were considered? How many objects examined?


DERREN BROWN
THE SYSTEM


7776

An appreciation of trials factors ("look elsewhere effect") is hugely important... an improper handling of this can lead to incorrect conclusions and opens the door to biased analyses!

This is not trivial ! **A full accounting for this can be tricky:**

- How many hypotheses have you actually tested?
- How many different ways have you tested each hypothesis?
- How many other things would have caught your eye?
- In general, how many ways have you looked at the data?

At the same time, the data needs to be thoroughly checked to look for possible problems and confirm how well it's understood

This is why physicists set the bar high in terms of significance level in order to claim a discovery

But it's easy to get carried away...

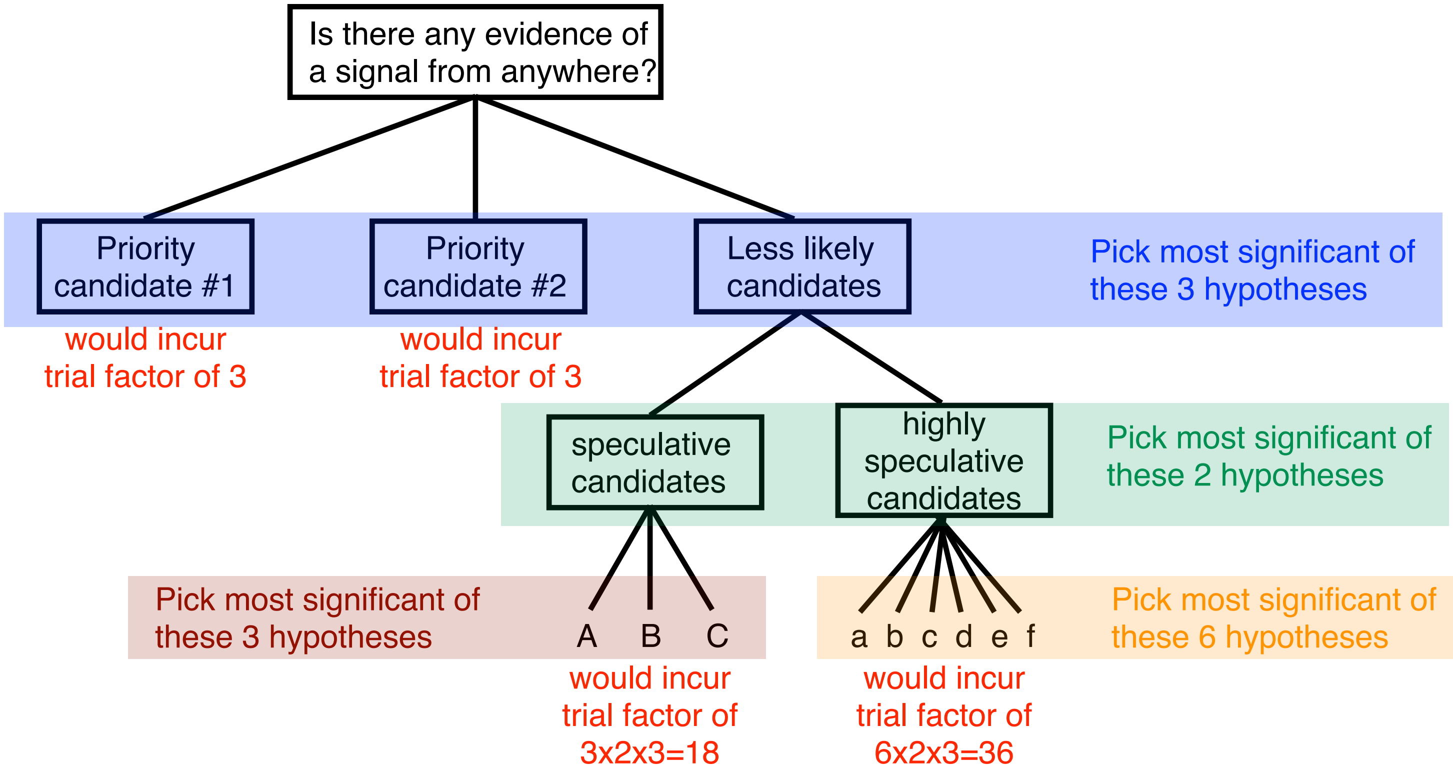
KEEP
CALM
AND
Look
Elsewhere

1) Trials factors apply to observations that would potentially lead to making a meaningful claim.

2) Verification based on applying the same analysis to an independent set of data is a good way to avoid misinterpretation of statistical fluctuations.

How do you deal with trial factors in the context of an open-ended search when an independent data set may not be available?

It's possible to structure trial factors based on an a priori ranking of hypothesis plausibility*:



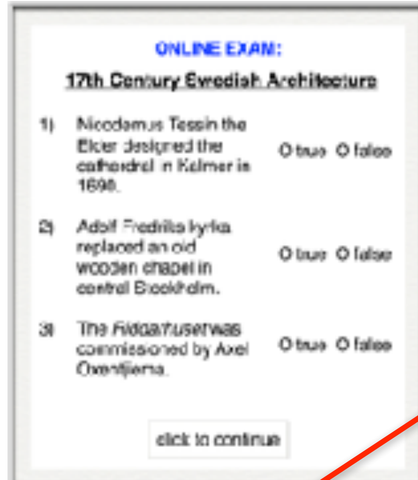
If something is seen from one of these tests that indicates a clear signal, it then belongs to a different population from random fluctuations. So, it can be removed and the search continued with the rest!

* S. Biller, Astroparticle Physics, Volume 4, Issue 3, p. 285-291

"Regression to the Mean"

Pop Quiz:

100 true/false questions on 17th century Swedish architecture:


ONLINE EXAM:

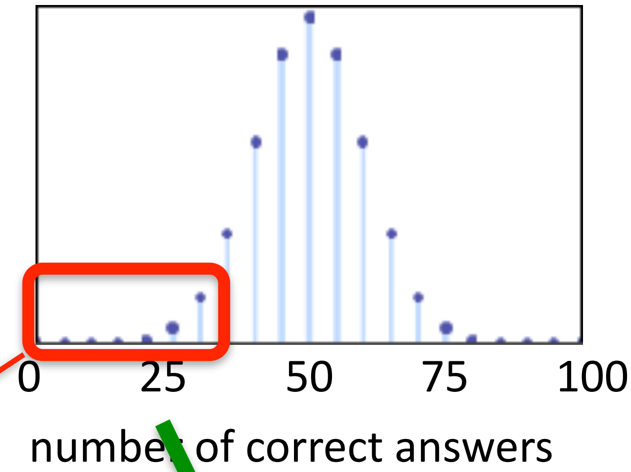
17th Century Swedish Architecture

1) Nicodemus Tessin the Elder designed the cathedral in Kalmar in 1690. O true O false

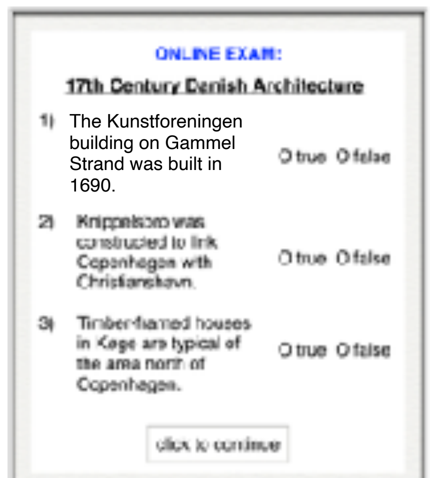
2) Adolf Fredriks kyrka replaced an old wooden chapel in central Stockholm. O true O false

3) The Riddarhuset was commissioned by Axel Oxenstjerna. O true O false

click to continue



100 true/false questions on 17th century Danish architecture:


ONLINE EXAM:

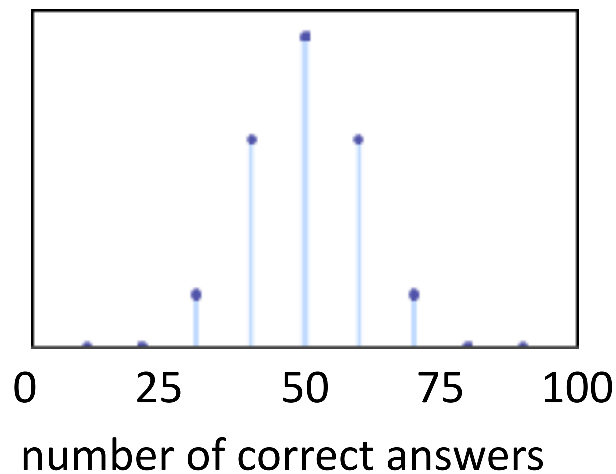
17th Century Danish Architecture

1) The Kunstforeningen building on Gammel Strand was built in 1690. O true O false

2) Knippelsbro was constructed to link Copenhagen with Christianshavn. O true O false

3) Timber-framed houses in Køge are typical of the area north of Copenhagen. O true O false

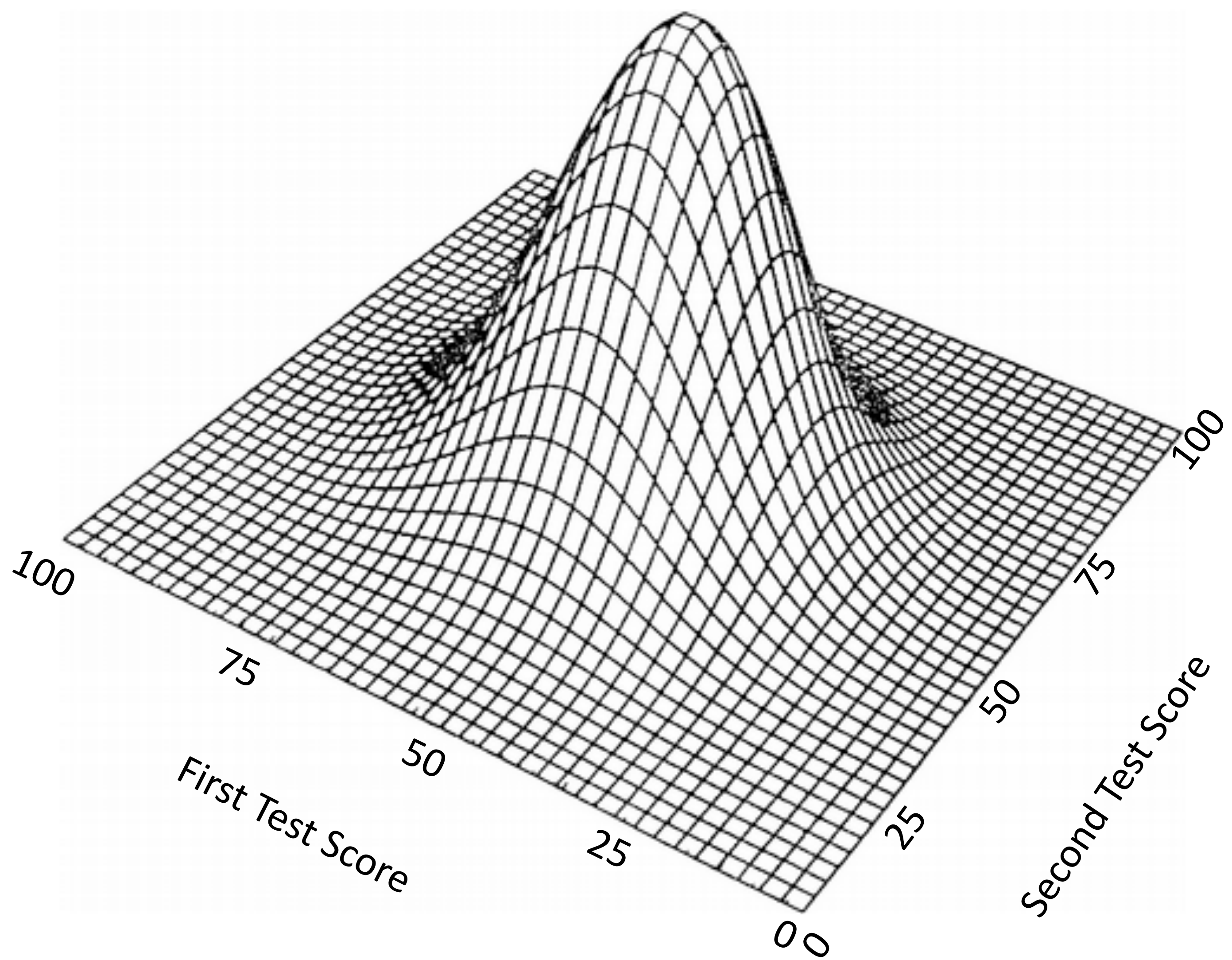
click to continue

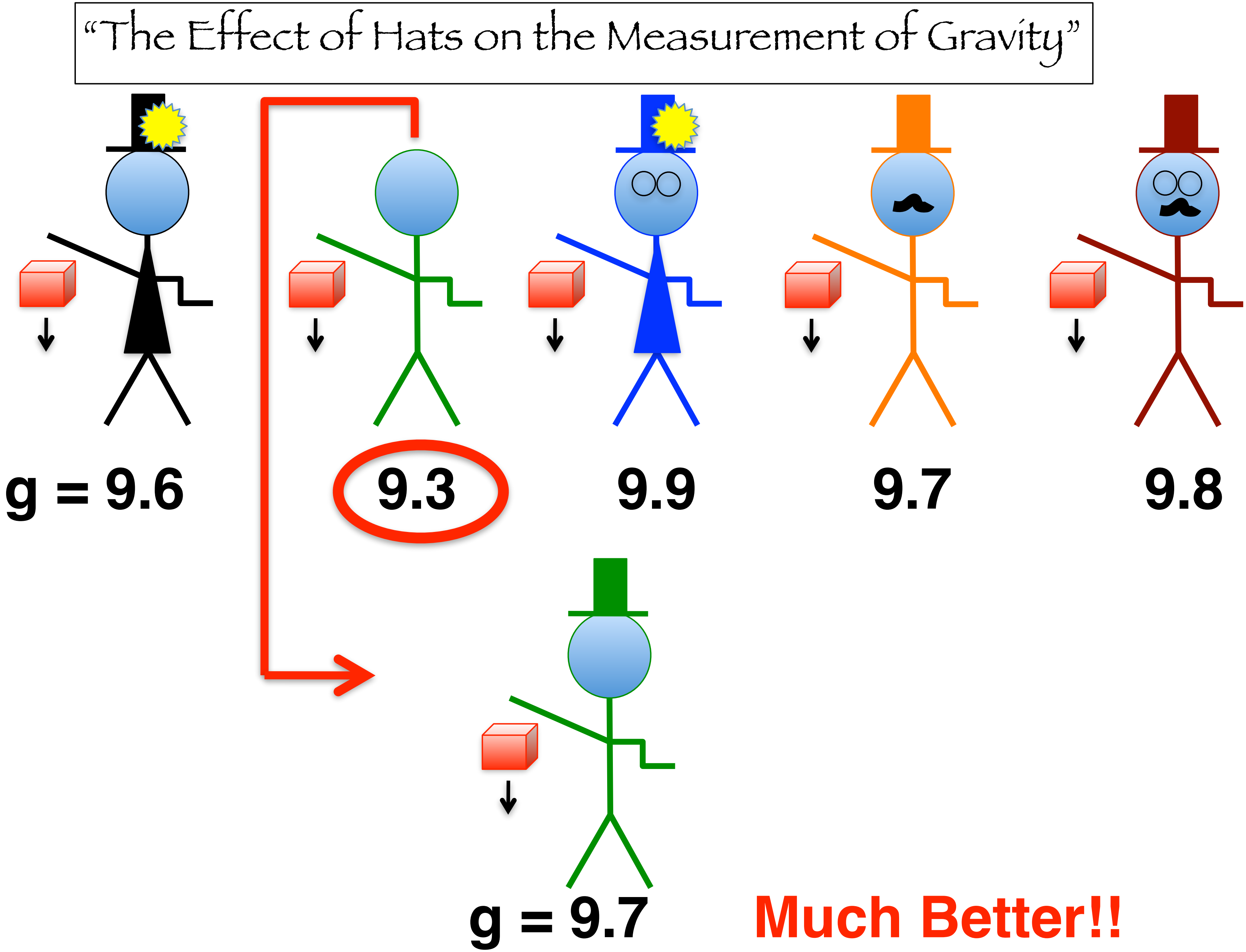


What an improvement! This particular group of students knows much more about Danish architecture!!

Bi-variate Distribution with Identical Marginal Distributions

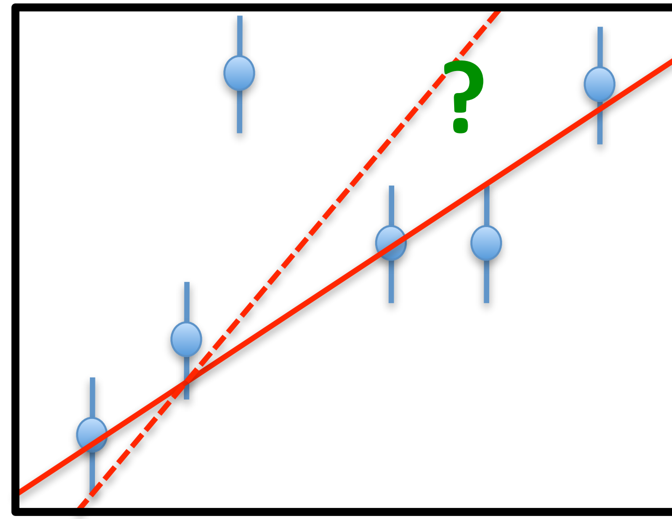
(i.e. uncorrelated)




"The Effect of Hats on the Measurement of Gravity"
g = 9.6
9.3
9.9
9.7
9.8
g = 9.7
Much Better!!

So How Do You Handle Outliers?

No clear rules!



Rules of Thumb:

- Look for possible systematic biases in the data;
- However, only reject outliers based on clear statistical/scientific criteria;
- Explicitly point out the issue and discuss the details;
- Be aware of any potential bias that could result and review the robustness of your final conclusions.

The total number of known species is ~1.5 million

The number of known species that can fly is ~500,000

$P(\text{flying}) > 5 \times 10^5 / 1.5 \times 10^6 = 0.33$

The number of plant species ~400,000

$P(\text{plant}) = 4 \times 10^5 / 1.5 \times 10^6 = 0.27$

Thus, probability of finding a flying plant is

$P(\text{plant}) \times P(\text{flying}) = 0.089$

And the expected number of flying plant species is

$(0.089)(1.5 \times 10^6) = 133{,}500$

Correlations

Joint PDF

	Insects	Birds	Mammals	Plants	Reptiles
Flying	500k 0.35	10k 0.007	400 2.8×10^{-4}	0	0
Non-Flying	500k 0.35	54 3.8×10^{-5}	6k 0.004	400k 0.28	10k 0.007

PDF for Non-Flying Species

Insects	Birds	Mammals	Plants	Reptiles
500k 0.54	54 5.9×10^{-5}	6k 0.006	400k 0.44	10k 0.011

"Marginalised" PDF for All Species

Insects	Birds	Mammals	Plants	Reptiles
500k 0.70	54 0.00704	6k 0.00428	400k 0.28	10k 0.007

Just to be clear:

For example, if we have 2 dependent variables, x & y:

$$\int P(x,y)dxdy = 1$$

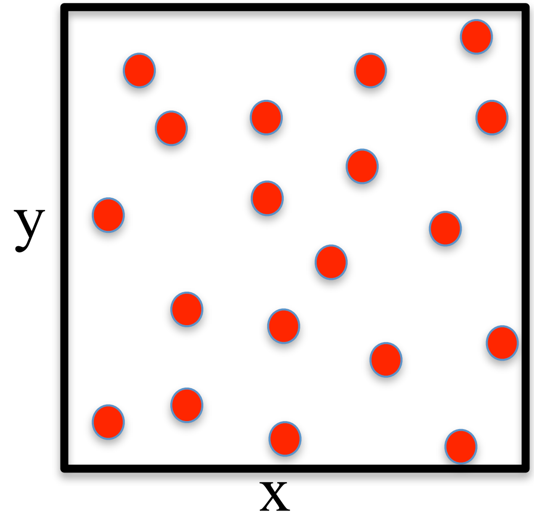
and

$$\langle f(x,y)\rangle = \int f(x,y)P(x,y)dxdy$$

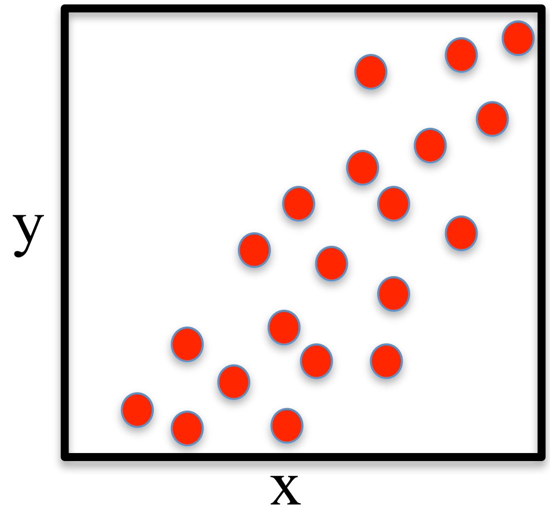
How do you tell whether variables are correlated?

Correlated or Uncorrelated!

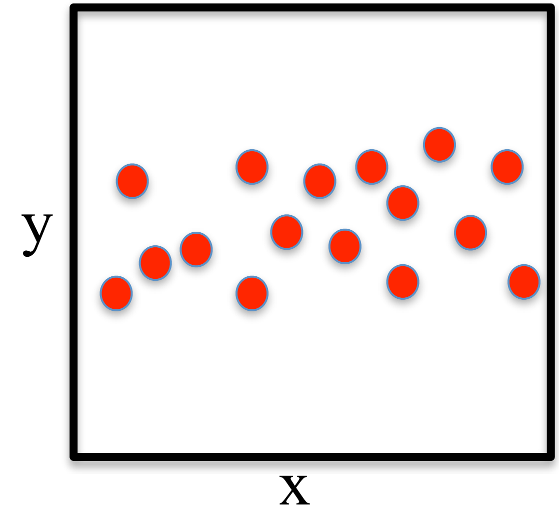
Correlated or Uncorrelated?



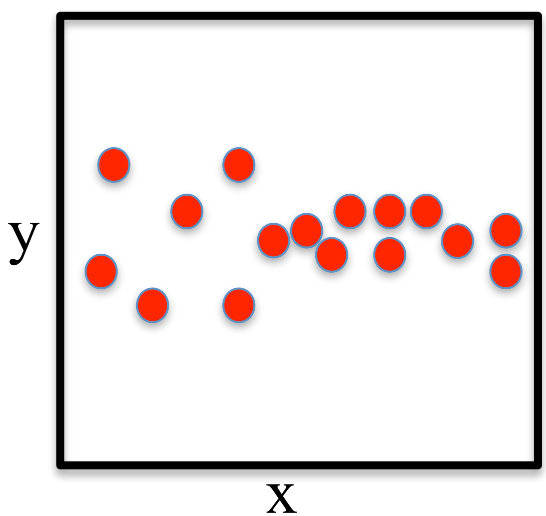
relatively uncorrelated



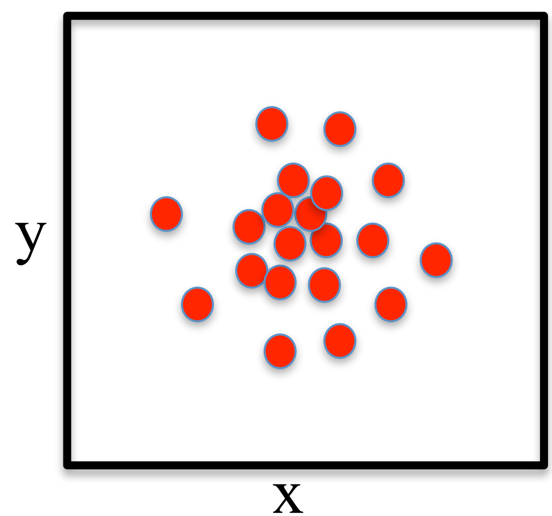
correlated



relatively uncorrelated



relatively uncorrelated mean, but correlated variance



relatively uncorrelated
(symmetric with similar marginal distributions)

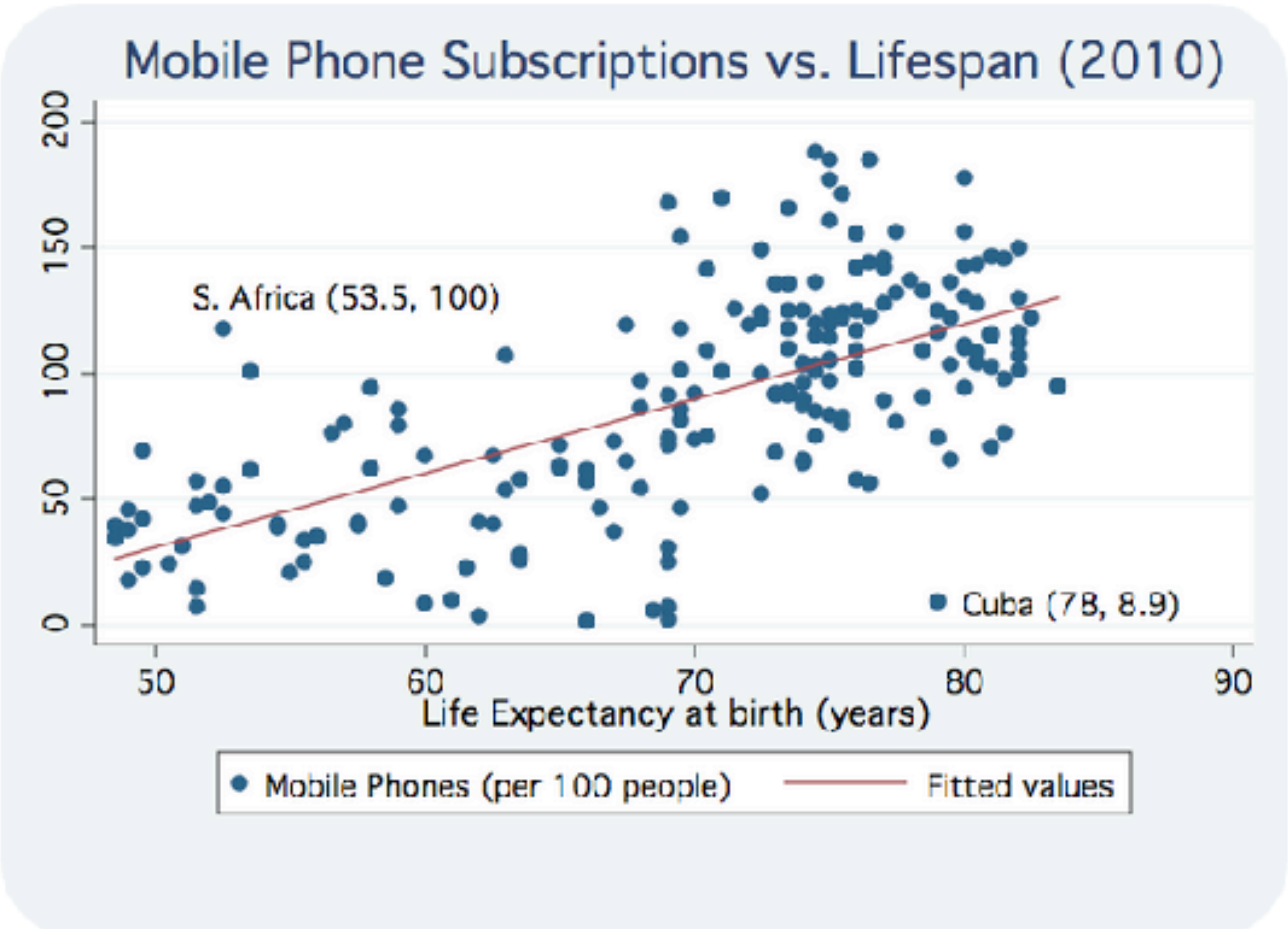
Beware of “hidden” correlations between **ANY** parameters that distinguish elements of your data set

IQ ⟷ Shoe Size

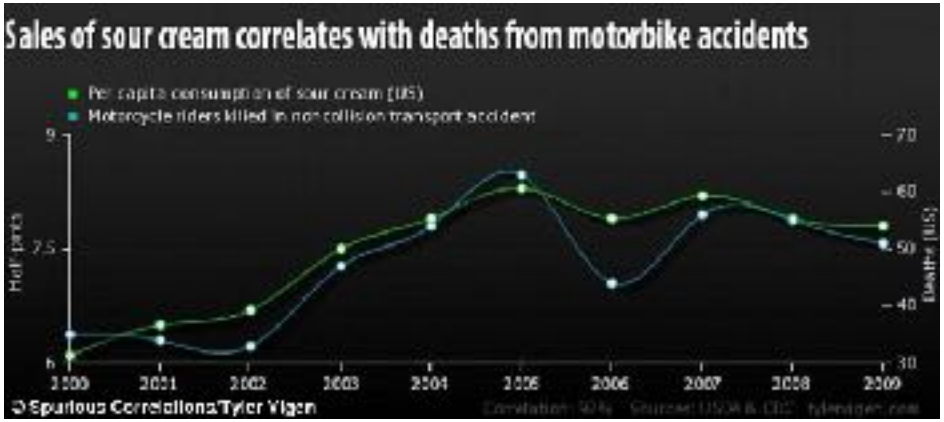
⇕

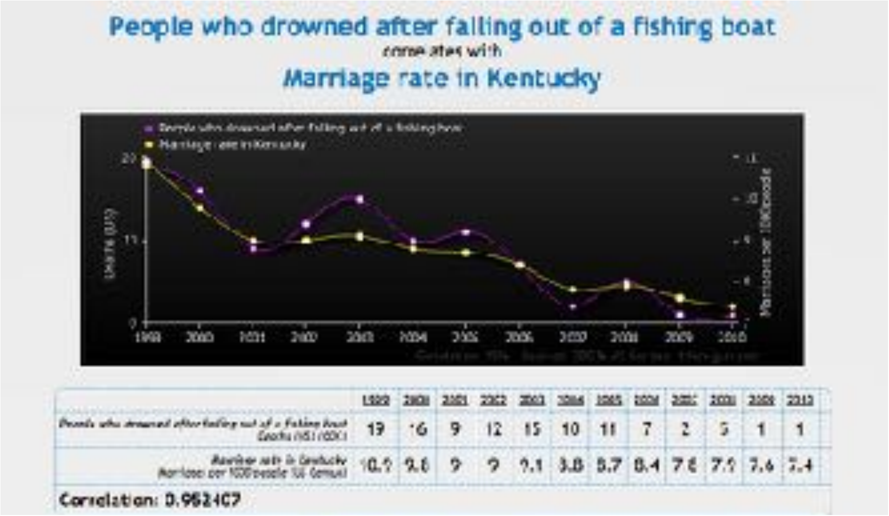
IQ ⇔ Age

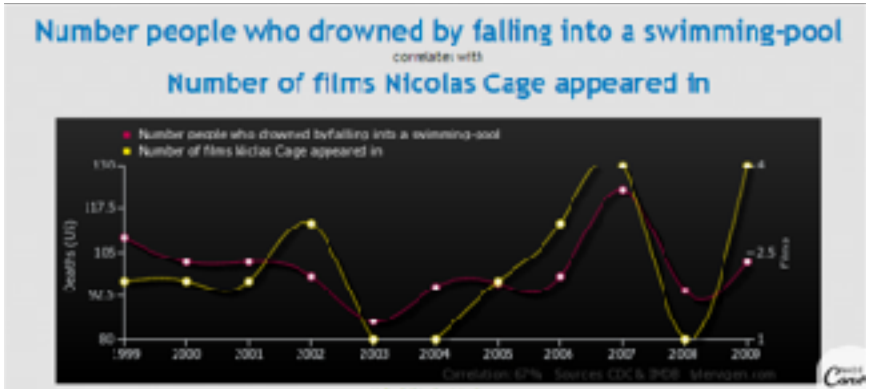
Beware of jumping to conclusion about cause and effect



Beware of spurious correlations







Uncertainties

Uncertainty ~~Error~~ Propagation

The thing you want to measure → $f(\mathbf{q}) = f(q_1, q_2, \ldots q_n)$ ← Dependent parameters (*e.g.* temperature, position, time, pressure...)

Want to use the distribution f to propagate uncertainties in **q**, but

1) We don't necessarily know the full joint distribution of **q** (*i.e.* the probability distribution for all possible sets of values)
2) Even if we did, it's cumbersome to deal with!

So, instead, let's approximate things to first order and then estimate the variance of f

$$f(\mathbf{q}) \sim f(\boldsymbol{\mu}) + \sum_{i=1}^{n} \left[\frac{\partial f}{\partial q_i} \right]_{\mathbf{q}=\boldsymbol{\mu}} (q_i - \mu_i)$$

Taylor expansion about the mean values for $\boldsymbol{q}$ where $f(\boldsymbol{\mu}) = f(\mu_1, \mu_2, \ldots \mu_n)$

$$f(\mathbf{q}) \sim f(\boldsymbol{\mu}) + \sum_{i=1}^{n} \left[\frac{\partial f}{\partial q_i} \right]_{\mathbf{q}=\boldsymbol{\mu}} (q_i - \mu_i)$$

zero by definition

$$\langle f(\mathbf{q}) \rangle \sim f(\boldsymbol{\mu}) + \sum_{i=1}^{n} \left[\frac{\partial f}{\partial q_i} \right]_{\mathbf{q}=\boldsymbol{\mu}} \langle q_i - \mu_i \rangle = f(\boldsymbol{\mu})$$

$$\langle f^2(\mathbf{q}) \rangle \sim \left\langle \left[f(\boldsymbol{\mu}) + \sum_{i=1}^{n} \left[\frac{\partial f}{\partial q_i} \right]_{\mathbf{q}\,\boldsymbol{\mu}} (q_i - \mu_i) \right]^2 \right\rangle$$

$$= f^2(\boldsymbol{\mu}) + \left\langle \left[\sum_{i=1}^{n} \left[\frac{\partial f}{\partial q_i} \right]_{\mathbf{q}=\boldsymbol{\mu}} (q_i - \mu_i) \right] \left[\sum_{j=1}^{n} \left[\frac{\partial f}{\partial q_j} \right]_{\mathbf{q}=\boldsymbol{\mu}} (q_j - \mu_j) \right] \right\rangle$$

zero by definition

$$+ 2f(\boldsymbol{\mu}) \sum_{i=1}^{n} \left[\frac{\partial f}{\partial q_i} \right]_{\mathbf{q}=\boldsymbol{\mu}} \langle q_i - \mu_i \rangle$$

$$= f^2(\boldsymbol{\mu}) + \sum_{i,j=1}^{n} \left[\frac{\partial f}{\partial q_i} \frac{\partial f}{\partial q_j} \right]_{\mathbf{q}=\boldsymbol{\mu}} \left\langle (q_i - \mu_i)(q_j - \mu_j) \right\rangle$$

So we get:

$$\sigma_f^2 = \left\langle f^2(\mathbf{q}) \right\rangle - \langle f(\mathbf{q}) \rangle^2$$

$$\simeq \sum_{i,j=1}^{n} \left[\frac{\partial f}{\partial q_i} \frac{\partial f}{\partial q_j} \right]_{\mathbf{q}=\boldsymbol{\mu}} \left\langle (q_i - \mu_i)(q_j - \mu_j) \right\rangle$$

"covariance matrix"

$$= \sum_{i,j=1}^{n} \left[\frac{\partial f}{\partial q_i} \frac{\partial f}{\partial q_j} \right]_{\mathbf{q}=\boldsymbol{\mu}} V_{ij}$$

If the q parameters are uncorrelated,

$$V_{ij} = \left\langle (q_i - \mu_i)(q_j - \mu_j) \right\rangle = 0 \quad for \quad i \neq j$$

$$V_{ii} = \left\langle (q_i - \mu_i)^2 \right\rangle = \sigma_i^2$$

$$\sigma_f^2 \simeq \sum_{i=1}^{n} \left[\frac{\partial f}{\partial q_i} \right]^2_{\mathbf{q}=\boldsymbol{\mu}} \sigma_i^2$$

for independent parameter uncertainties

Some simple examples:

$$T_{tot} = t_1 + t_2$$

$$\sigma_T^2 \simeq \left[\frac{\partial T}{\partial t_1}\right]^2 \sigma_1^2 + \left[\frac{\partial T}{\partial t_2}\right]^2 \sigma_2^2 \;= \sigma_1^2 + \sigma_2^2$$

$$s = vt$$

$$\sigma_s^2 \simeq \left[\frac{\partial s}{\partial v}\right]^2 \sigma_v^2 + \left[\frac{\partial s}{\partial t}\right]^2 \sigma_t^2 = t^2\sigma_v^2 + v^2\sigma_t^2$$

$$\left(\frac{\sigma_s}{s}\right)^2 \simeq \left(\frac{\sigma_v}{v}\right)^2 + \left(\frac{\sigma_t}{t}\right)^2$$

For a quadrature addition of uncertainties, uncertainties that are half as big only carry 1/4 of the weight, and uncertainties that are 1/4 as big only carry 1/16 of the weight... **Only the dominant uncertainties matter!**

More General Example: Measurement of Linear Thermal Expansion Coefficient

Measure h by timing the drop of snowballs on one particularly cold day, then compare with h_0 to determine α

data point	Time (s)	Temp (°C)
1	2.02	-5.6
2	1.99	-4.8
3	2.05	-4.4
.	.	.
.	.	.
n	2.01	-5.3

In this simple analysis, we're interested in determining the average values of drop time and temperature for the day:

$$t_\mu \sim \tilde{t}_\mu = \frac{1}{n}\sum_{i=1}^{n} t_i \qquad T_\mu \sim \tilde{T}_\mu = \frac{1}{n}\sum_{i=1}^{n} T_i$$

Then, from the relation $h = h_0[1 - \alpha(T - T_0)]$

estimate the expansion coefficient:

$$\tilde{\alpha} = \frac{\frac{\tilde{h}_\mu}{h_0} - 1}{\tilde{T}_\mu - T_0} = \frac{\frac{g\tilde{t}_\mu^2}{2h_0} - 1}{\tilde{T}_\mu - T_0}$$

Now we want to find the uncertainty in $\tilde{\alpha}$ by propagating the uncertainties in $\tilde{t}_\mu$ and $\tilde{T}_\mu$

data point	Time (s)	Temp (°C)
1	2.02	-5.6
2	1.99	-4.8
3	2.05	-4.4
.	.	.
.	.	.
n	2.01	-5.3

$$V_{ij} =$$

	time	Temp
time	σ_t^2	$\text{cov}(t, T)$
Temp	$\text{cov}(t, T)$	σ_T^2

Approach 1: Evaluate Approximately Using the Data

$$\sigma_t^2 = \left\langle (t - t_\mu)^2 \right\rangle \sim \frac{1}{n-1} \sum_{i=1}^{n} (t_i - \tilde{t}_\mu)^2$$

$$\sigma_T^2 = \left\langle (T - T_\mu)^2 \right\rangle \sim \frac{1}{n-1} \sum_{i=1}^{n} (T_i - \tilde{T}_\mu)^2$$

$$\text{cov}(t, T) = \left\langle (t - t_\mu)(T - T_\mu) \right\rangle \sim \frac{1}{n-1} \sum_{i=1}^{n} (t_i - \tilde{t}_\mu)(T_i - \tilde{T}_\mu)$$

Drawback: Requires a large enough data set so that estimates are well determined

data point	Time (s)	Temp (°C)
1	2.02	-5.6
2	1.99	-4.8
3	2.05	-4.4
.	.	.
.	.	.
n	2.01	-5.3

$V_{ij} =$

	time	Temp
time	σ_t^2	$\text{cov}(t, T)$
Temp	$\text{cov}(t, T)$	σ_T^2

Approach 2: Use Calibration Measurements and/or Physical Models

$\sigma_t^2 = \left\langle (t - t_\mu)^2 \right\rangle$ from calibration of timing accuracy

$\sigma_T^2 = \left\langle (T - T_\mu)^2 \right\rangle$ from calibration of temperature reading accuracy

Temperature variations during the day are sufficiently small that the correlation with time measurements is very weak, so $\text{cov}(t, T) \sim 0$

Drawback: Model could be wrong

The temperature variations might influence reaction times and this might have a noticeable systematic impact on the stopwatch measurements

data point	Time (s)	Temp (ºC)
1	2.02	-5.6
2	1.99	-4.8
3	2.05	-4.4
.	.	.
.	.	.
n	2.01	-5.3

$V_{ij} =$

	time	Temp
time	σ_t^2	$\mathrm{cov}(t, T)$
Temp	$\mathrm{cov}(t, T)$	σ_T^2

Best Approach: **Do both!**

Check the consistency or your model and calibrations with the data
(If things don't add up, dig around to understand it!)

$$\alpha = \frac{\frac{gt^2}{2h_0} - 1}{T - T_0}$$

$$\left[\frac{\partial \alpha}{\partial t}\right]_{t_\mu, T_\mu} = \frac{\frac{gt_\mu}{h_0} - 1}{T_\mu - T_0}$$

$$\left[\frac{\partial \alpha}{\partial T}\right]_{t_\mu, T_\mu} = -\frac{\frac{gt_\mu^2}{2h_0} - 1}{(T_\mu - T_0)^2}$$

$$\sigma_\alpha^2 = \left(\left[\frac{\partial\alpha}{\partial t}\right]_{t_\mu, T_\mu}, \left[\frac{\partial\alpha}{\partial T}\right]_{t_\mu, T_\mu} \right) \begin{pmatrix} \sigma_t^2 & \mathrm{cov}(t, T) \\ \mathrm{cov}(t, T) & \sigma_T^2 \end{pmatrix} \begin{pmatrix} \left[\frac{\partial\alpha}{\partial t}\right]_{t_\mu, T_\mu} \\ \left[\frac{\partial\alpha}{\partial T}\right]_{t_\mu, T_\mu} \end{pmatrix}$$

$$\sigma_\alpha^2 = \left[\frac{\partial\alpha}{\partial t}\right]_{\boldsymbol{\mu}}^2 \sigma_t^2 + \left[\frac{\partial\alpha}{\partial T}\right]_{\boldsymbol{\mu}}^2 \sigma_T^2 + 2\left[\frac{\partial\alpha}{\partial t}\frac{\partial\alpha}{\partial T}\right]_{\boldsymbol{\mu}} \mathrm{cov}(t, T)$$

The Statistical Calculation That You Should Have Done at the Start!

Typical linear expansion coefficients for building materials ~5x10^{-6} per ^{o}C

Take $(T-T_o) \sim 20^oC$

$h_0 - h \sim (20m)(5\times10^{-6})(20^oC) = 0.002m$

velocity at impact = $\sqrt{2gh_0} = \sqrt{2(9.8m/s^2)(20m)} \sim 20m/s$

So, timing must be known to an accuracy of (0.002/20) = 0.0001s

Accuracy of any one timing timing measurement ~ 0.1s

But we improve by averaging lots of measurements according to $\sigma_m = \frac{\sigma}{\sqrt{n}}$

How many measurements do we need?

$$n = \frac{\sigma^2}{\sigma_m^2} \sim \left(\frac{0.1}{0.0001}\right)^2 = 10^6$$

(ignoring systematic uncertainties!)

Statistical Uncertainties

Fundamental, calculable, random variations due to an inherent limited sampling of the underlying distribution (i.e. counting statistics).

Systematic Uncertainties

Incidental, estimated (bounded), systematic biases incurred as a result of limited measurement precision (also always present).

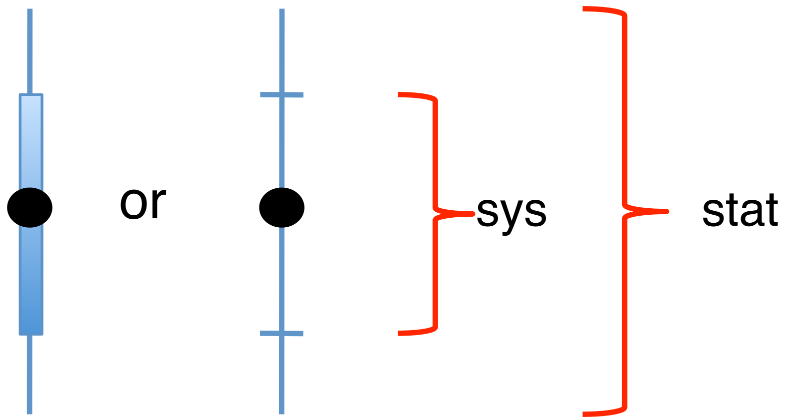
There is no universally applicable method for estimating/bounding* systematic uncertainties. A typical approach often relies on independent cross-checks, accounting for possible statistical limitations of calibration procedures, knowledge about the experimental design and general consistency arguments.

* Systematic errors that are "determined" become corrections!

Because of their very different nature, there is no standard, mathematically rigorous way to combine the 2 types of uncertainties. The convention is thus to quote results in the form:

Result ± Uncertainty (stat) ± Uncertainty (sys)

And error bars such as:

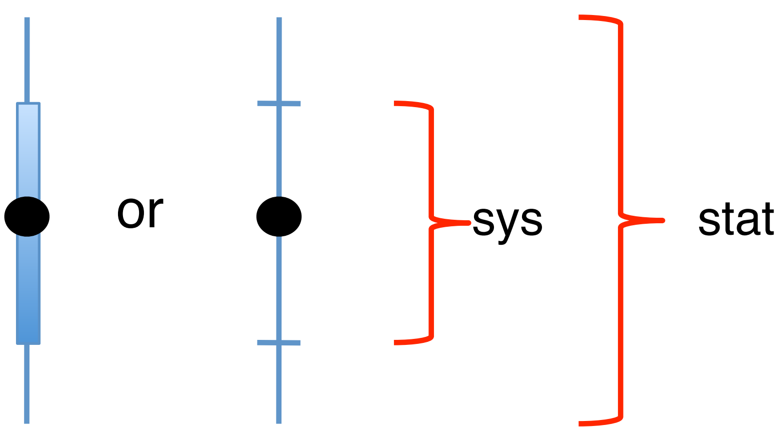


(Much more on error bars later!!)

How do you then make use of such data points to fit a model?

It is often generally assumed that systematic uncertainties can be treated in a similar way to statistical uncertainties, with careful attention to correlations.

Ideally, the best way to treat systematic uncertainties are as free parameters in the model fit, constrained by the separately determined bounds on their values.



(Much more on error bars later!!)