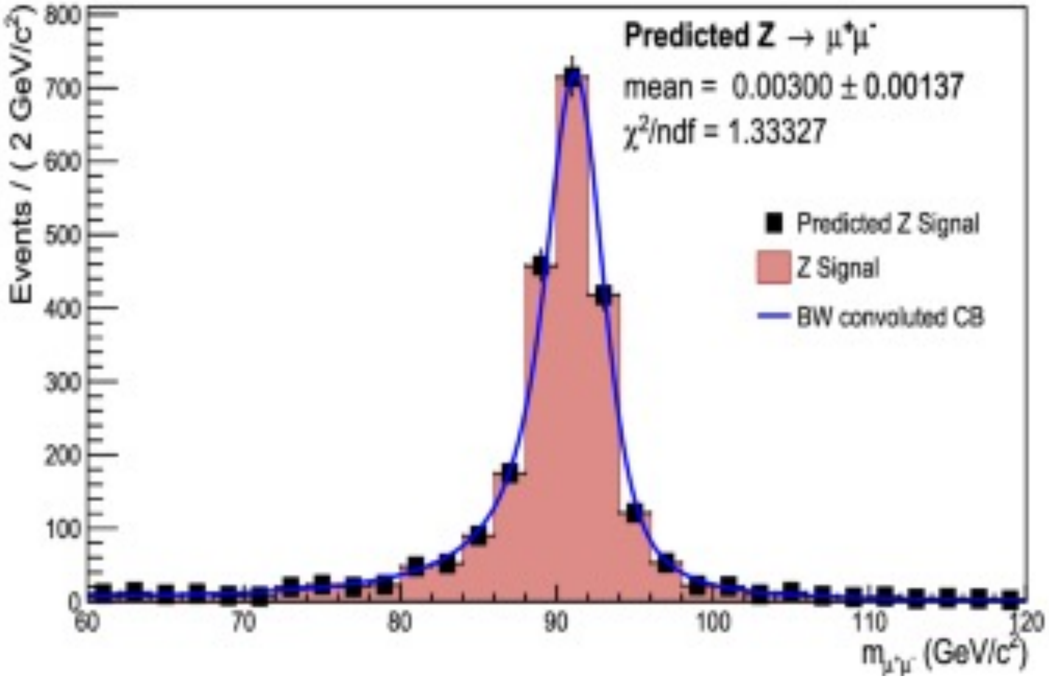


LIES, DAMNED LIES, AND STATISTICS

Eric Prebys

FRS-003-009

Errors and Significance

- Everything in science is a theory!
- We only accept a theory to the extent that it has been verified by experiment.
- Some theories have been verified to such a degree that we – tentatively! – accept them as “facts” or ”laws”. These include
  - Conservation of Energy
  - Special Relativity, including the speed of light as the limit to the speed of information transfer
  - Electricity and Magnetism
  - Yet even these are still tested.
- In order to understand the significance of a particular result, we have to understand the errors associated with it.

Errors in Measurement

- Every measurement has an error associated with it.
- In precision science, determining the error is quite often the hardest part of the job!
- For example, the mass of the proton has been very well measured, and the current value is*

$$1.007276466621 \pm 0.000000000053 \quad \text{amu}$$

Value

Error

- What does this mean?
- Does it mean the ”true” value absolutely has to be within this range?
  - Short answer: no

*https://pdg.lbl.gov

The Role of Statistics

- Once upon a time, the universe was believed to be “deterministic”; that is, if we knew every particle’s position and trajectory, we could precisely predict everything that would happen in the future
  - Or work backwards to figure out everything that happened in the past.
- They knew they could never do this in practice, so they invented a number of statistical tools for complex descriptions and predictions.
- Following the discovery of quantum mechanics, we now know that nature behaves statistically at the the most fundamental level.
  - Not deterministic
  - Every experimental outcome is associated with a probability.

When Probabilities Become “Certainties”

- Even though everything is probabilistic at the most fundamental level, once we get enough statistics, we can predict precise outcomes
  - The more statistics we gather, the more precise the prediction.
- Example:
  - Carbon-14 decays via the “beta decay”

$$^{14}C \rightarrow ^{14}N + e^{-} + \overline{\nu_e}$$

  - If I have one atom of 14C, there is *no way* I can predict when it will decay; HOWEVER
  - If I have 1 gram of 14C, then 5,730 year later I will have *very close* to .5 grams left.
  - How close? Let’s look at that…

Example Statistical Distributions

- 1 Die

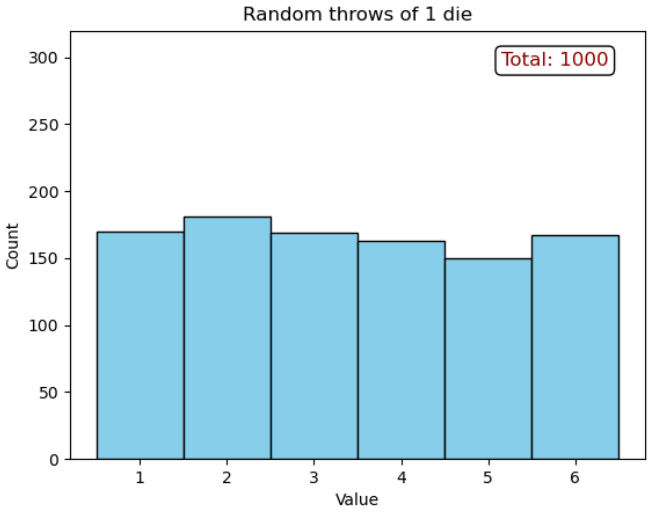
  - Equal chances of coming up any number from 1 to 6

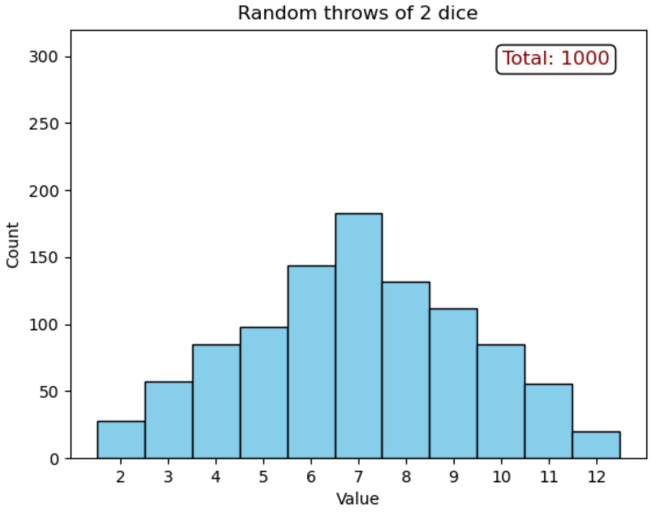
- 2 Dice

  - 6x6=36 possible throws
  - 1 way to make 2 or 12
  - 2 ways to make 3 or 11

  ⋮

  - 6 ways to make 7

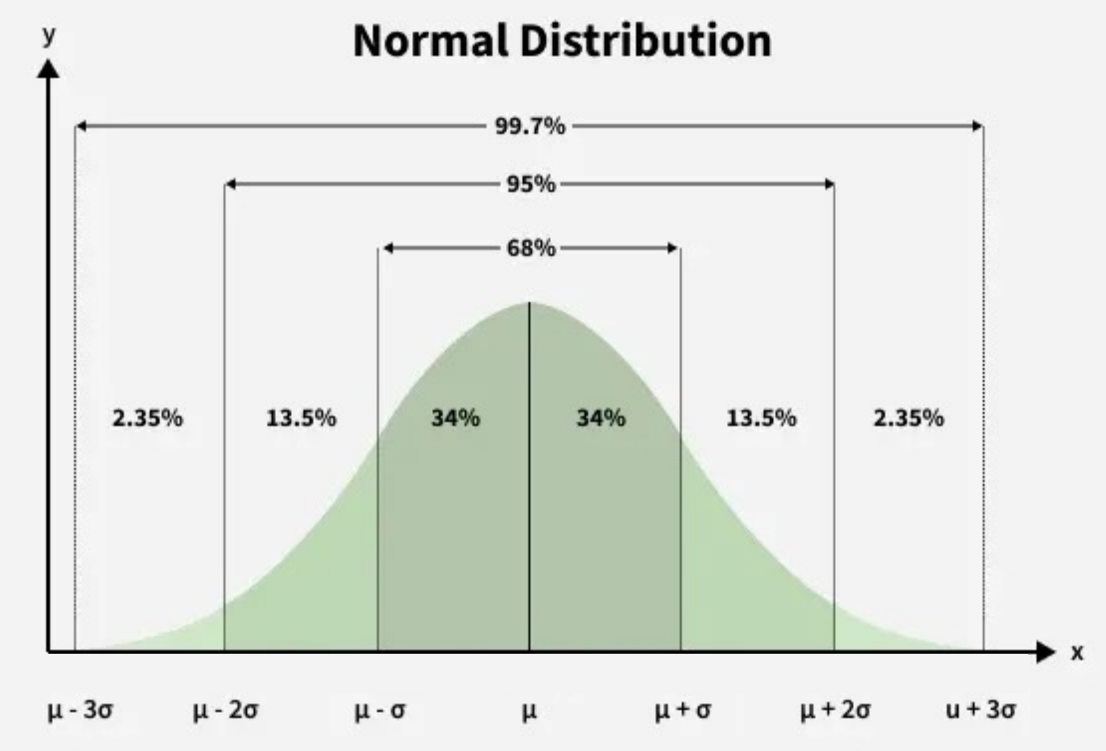




(Go do demo…)

Central Limit Theorem

- No matter what distribution you start with, the distribution of averages, *or anything calculated from those averages*, will approach a “normal distribution”
  - AKA “Gaussian Distribution” or “Bell Curve”



Characterized by…

- The mean (μ)

$$\mu = \frac{1}{N}\sum_{n=1}^{N} x_n$$

- The standard deviation (σ)

$$\sigma = \sqrt{\frac{1}{N}\sum_{n=1}^{N}(x_n - \mu)^2}$$

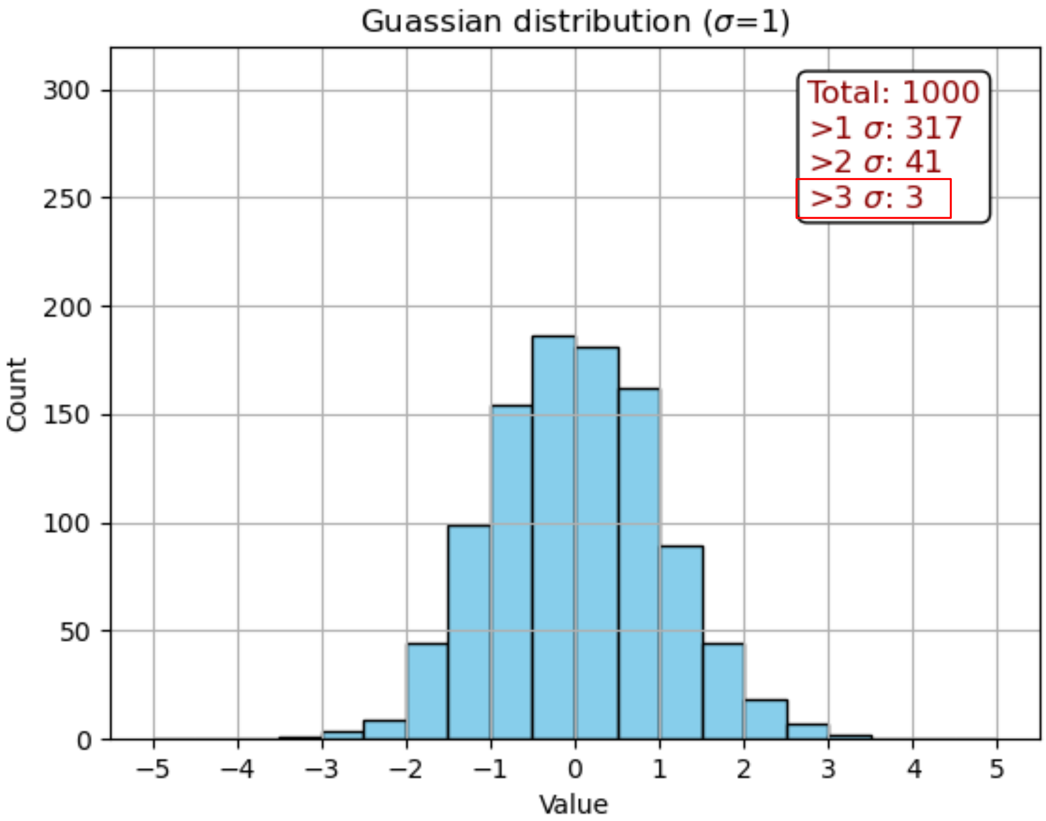
- The ”probability density function”

$$PDF(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

- Note:
  - The error on a result is usually 1σ
  - Statistically, 1 in 3 events will be more than 1σ from the mean!

Understanding The Normal Distributions

- In most cases, “discoveries” are presented in terms of the probability that the observation is result of statistical fluctuations of the “null hypothesis”, but how unlikely does it have to be for us to be persuaded?

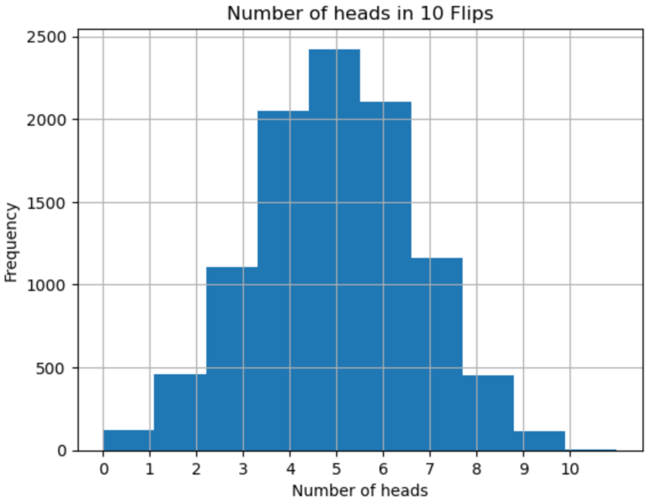


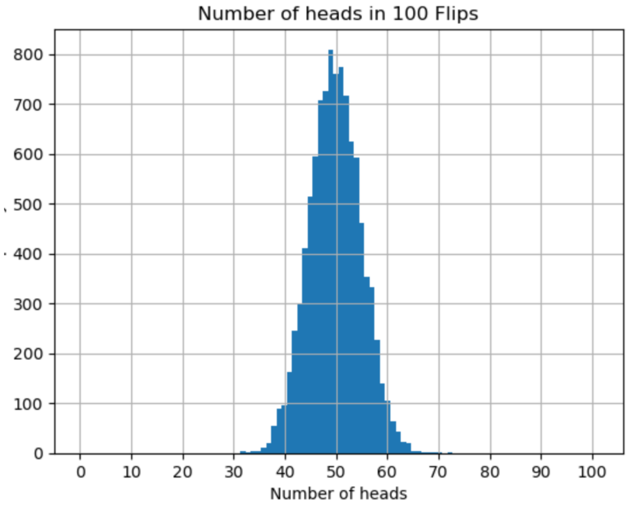
In 1000 throws, I got 3 greater than 3σ

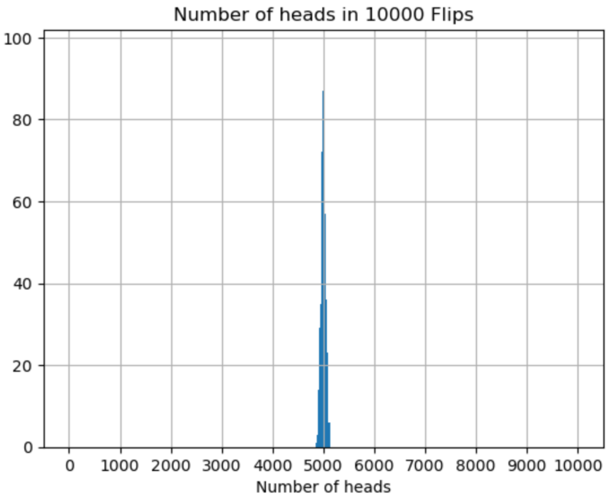
Now consider doing a whole bunch or different studies. Roughly 3 out of 1000 of them will have show a 3σ effect, due entirely to statistical fluctuation!

More Statistics -> More Precision

- Consider a coin flip
  - 50% chance of heads, 50% chance of tails
  - The more times if flip, the closer I get to exactly 50% heads
- For each of these, I do a certain number of flips, and then repeat the experiment 10,000 times







- The average fraction of heads will (of course) be ½, and the standard deviation of that value will be

$$\sigma_{n_H/N} = \frac{1}{2}\frac{1}{\sqrt{N}}$$

This is always the behavior with statistical errors!

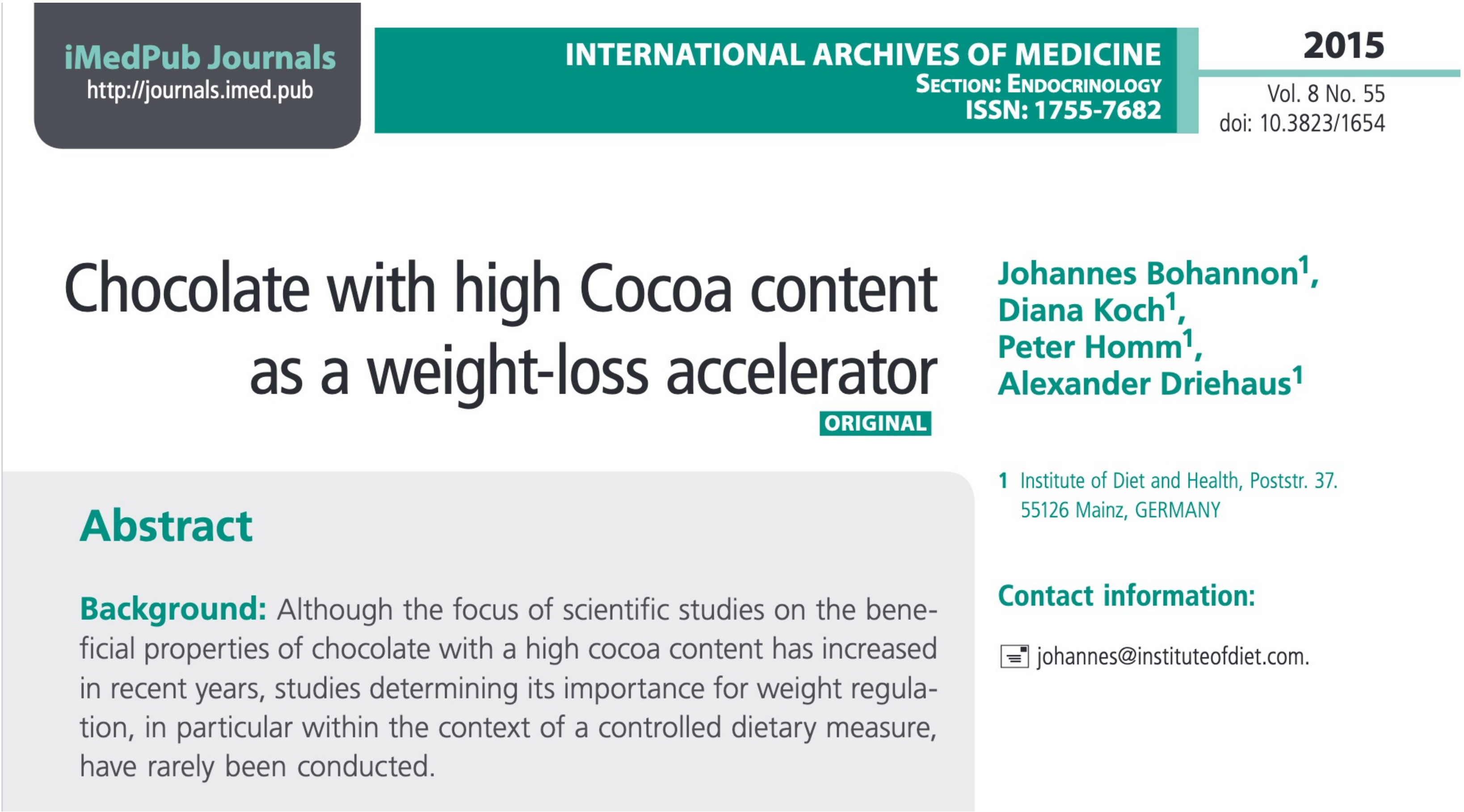
Why the House Always Wins

- No one can predict whether a single person will win or lose in a casino; HOWEVER
- If enough people gamble, the casino can predict *very exactly* how much they will make
  - It's actually one of the most reliable businesses on Earth
- Example: Roulette

- 40 positions on the wheel
- If I bet an exact number and color, I have a 1 in 40 chance of winning, BUT
- It only pays out 35 to 1
- In other words, if I keep making $1 bets, on average I'll win about $35 for every $40 dollars I bet!
- All the other roulette bets are similarly biased in favor of the house.

Teachable Moment: Is Chocolate Good for You?

- In 2015, a group from the “Institute of Diet Health” published the following paper



iMedPub Journals
http://journals.imed.pub

INTERNATIONAL ARCHIVES OF MEDICINE
Section: Endocrinology
ISSN: 1755-7682

2015
Vol. 8 No. 55
doi: 10.3823/1654

Chocolate with high Cocoa content as a weight-loss accelerator

ORIGINAL

Johannes Bohannon[1],
Diana Koch[1],
Peter Homm[1],
Alexander Driehaus[1]

[1] Institute of Diet and Health, Poststr. 37.
55126 Mainz, GERMANY

Contact information:

johannes@instituteofdiet.com.

Abstract

Background: Although the focus of scientific studies on the beneficial properties of chocolate with a high cocoa content has increased in recent years, studies determining its importance for weight regulation, in particular within the context of a controlled dietary measure, have rarely been conducted.

- Not surprisingly, the popular press went wild!

Really?

- Sadly, no.
- The author wrote an article in Gizmodo entitled “I Filled Millions into Thinking Chocolate Helps Weight Loss”. From the article…
  - I am Johannes Bohannon, Ph.D. Well, actually my name is John, and I’m a journalist. I do have a Ph.D., but it’s in the molecular biology of bacteria, not humans. The Institute of Diet and Health? That’s nothing more than a website. Other than those fibs, the study was **100 percent authentic**. My colleagues and I recruited actual human subjects in Germany. **We ran an actual clinical trial, with subjects randomly assigned to different diet regimes. And the statistically significant benefits of chocolate that we reported are based on the actual data.** It was, in fact, a fairly typical study for the field of diet research. **Which is to say: It was terrible science.** The results are meaningless, and the health claims that the media blasted out to millions of people around the world are utterly unfounded.

What they actually reported

- They divided subjects into three groups
  - One group put on low carb diets
  - One group put on low carb diets + 42 grams of chocolate/day
  - One group ate whatever they wanted
- The first two groups lost weight. Not surprisingly, the third one didn't.

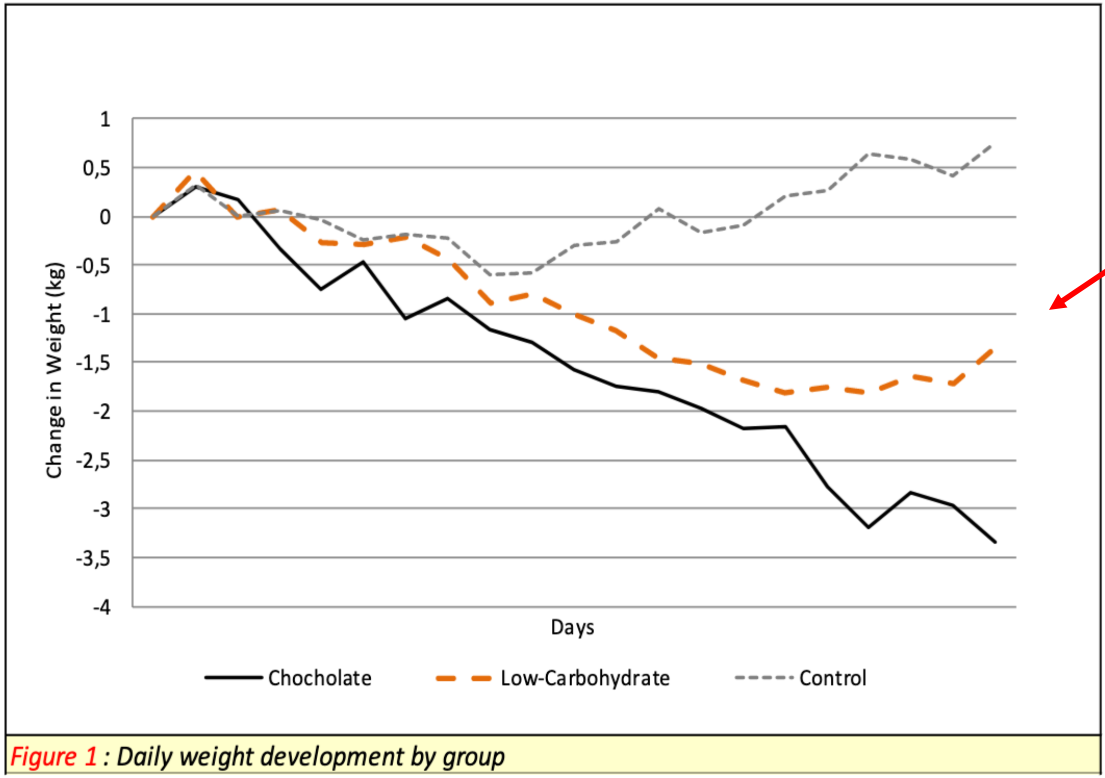


Figure 1 : Daily weight development by group

Note the lack of error bars!

Additional Health Benefits of Chocolate

- In addition to weight loss, they observed other positive health benefits

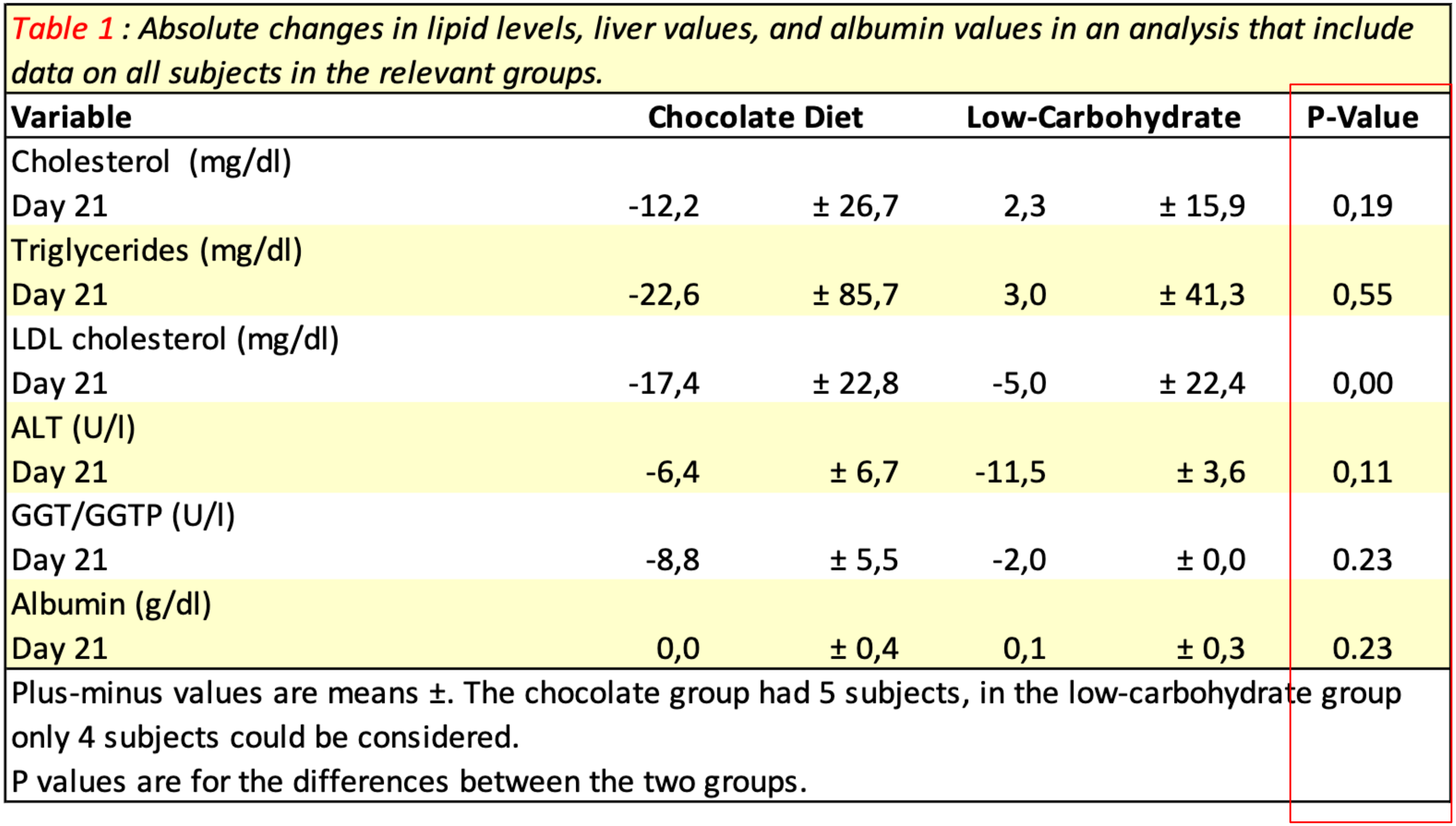


Table 1 : Absolute changes in lipid levels, liver values, and albumin values in an analysis that include data on all subjects in the relevant groups.

Variable	Chocolate Diet		Low-Carbohydrate		P-Value
Cholesterol (mg/dl)					
Day 21	-12,2	± 26,7	2,3	± 15,9	0,19
Triglycerides (mg/dl)					
Day 21	-22,6	± 85,7	3,0	± 41,3	0,55
LDL cholesterol (mg/dl)					
Day 21	-17,4	± 22,8	-5,0	± 22,4	0,00
ALT (U/l)					
Day 21	-6,4	± 6,7	-11,5	± 3,6	0,11
GGT/GGTP (U/l)					
Day 21	-8,8	± 5,5	-2,0	± 0,0	0.23
Albumin (g/dl)					
Day 21	0,0	± 0,4	0,1	± 0,3	0.23

Plus-minus values are means ±. The chocolate group had 5 subjects, in the low-carbohydrate group only 4 subjects could be considered.

P values are for the differences between the two groups.

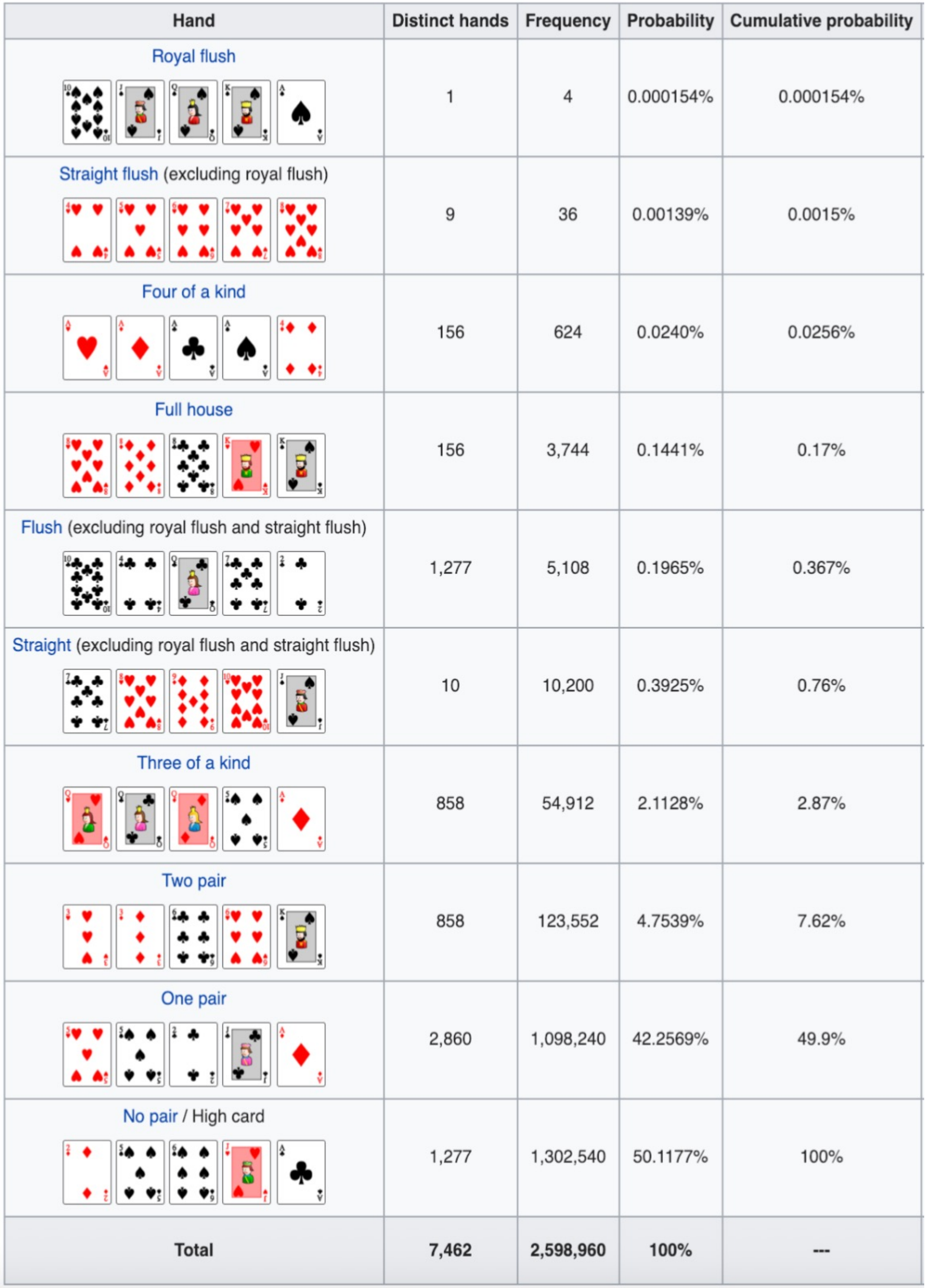
Put a pin in this!

How they did it…

- The authors set out to publish a bogus result, BUT they were not willing to lie.
- They set up a situation all but guaranteed to generate spurious results
  - You'll notice that nowhere in the article to they mention the total number of subjects
    - It was 16, divided into three groups! (and one dropped out)
  - There was no attempt at assigning an uncertainty to any of the measurements
    - Weight can fluctuate up to *5 lbs per day*.
  - **Here's the important one:** *they didn't decide in advance what they were going to report*. They measured 18 different things and then only discussed the ones that seemed to show statistical significance.
    - This is so common it's referred to as "p-hacking"
    - It's usually done unintentionally, but in this case, it was done on purpose.

Card Demo

- 5 Card Stud (single deal, no wildcards)



Hand	Distinct hands	Frequency	Probability	Cumulative probability
Royal flush	1	4	0.000154%	0.000154%
Straight flush (excluding royal flush)	9	36	0.00139%	0.0015%
Four of a kind	156	624	0.0240%	0.0256%
Full house	156	3,744	0.1441%	0.17%
Flush (excluding royal flush and straight flush)	1,277	5,108	0.1965%	0.367%
Straight (excluding royal flush and straight flush)	10	10,200	0.3925%	0.76%
Three of a kind	858	54,912	2.1128%	2.87%
Two pair	858	123,552	4.7539%	7.62%
One pair	2,860	1,098,240	42.2569%	49.9%
No pair / High card	1,277	1,302,540	50.1177%	100%
Total	**7,462**	**2,598,960**	**100%**	**---**

Looking in too many places

- If I make one measurement that has a probability of *p=5%* of occurring, then it has a probability of *(1-p)=95%* of not occurring.
- However, it I look *n* different things, each of which has a probability of *p* of occurring, then the probability of not observing anything, ever is

$$(1-p)^n$$

which eventually goes to zero

- In this case, if we make 18 observations, the chance that one of them has an effect with less than a 5% probability is

$$1-(1-.05)^{18}=60\%$$

Better than even odds

- The just measured a bunch of things and picked couple that were statistically significant.
  - The fact that weight loss was one of them was just a happy accident.
- At least they were honest about it and came clean right away.

When it’s not on purpose and it’s not funny…

- In 1992, the following article was published
  - Feychting, M, and Ahlbom, A. **Magnetic fields and cancer in people residing near Swedish high voltage power lines**. Sweden: N. p., 1992. Web.
- which contained the following in its abstract
  - For childhood leukemia and with cut off points at 0.1 and 0.2 [mu]T**, the relative risk (RR) increased over the two exposure levels and was estimated at 2.7** (95% c.l.: 1.0-6,3) for 0.2 [mu]T and over.
- In other words, living near power lines appeared to cause a 170% increase in the risk of childhood leukemia.
- The world, not surprisingly, lost its goddamn mind.
  - Real estate near power lines plummeted
  - There was discussion of relocating schools
  - The term “EMF” (electromagnetic field) entered the common vocabulary, and when people pointed out that fields were stronger under an electric blanket then 50m from a powerline, people started to worry about electric blankets.

The Problem with the Study

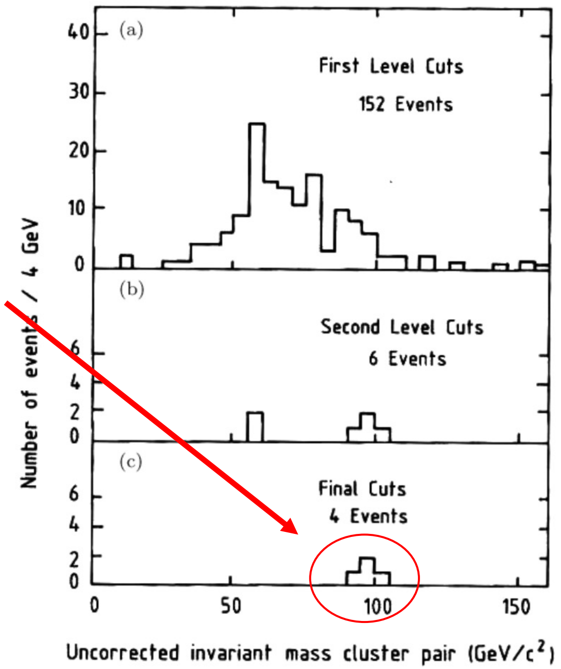
- First of all, it’s worth noting that this is a report, not a refereed journal publication.
- Second, the authors were honest, and the big problem is right there in the abstract.
  - The study was designed as a case-control study, based on the population comprised of everyone who have lived on a property located within 300 meters from any of the 220 and 400 kV power lines in Sweden during the period from 1960 through 1985. […] The cases were all instances of cancer diagnosed between 1960-85. **For children, all types of cancer were included**, while for adults the study was restricted to leukemia and brain tumors.
- In other words, they had done by accident what the chocolate guys had done on purpose: measured a bunch of things and published the one that had a statistical uptick.
- They also never accounted for “confounding factors”, example
  - Health tends to correlate with wealth
  - Rich people don’t tend to live near power lines, because they’re unsightly
- Note: even though this started the whole thing, and there have been many studies since, no one cites this anymore because it has been completely discredited.

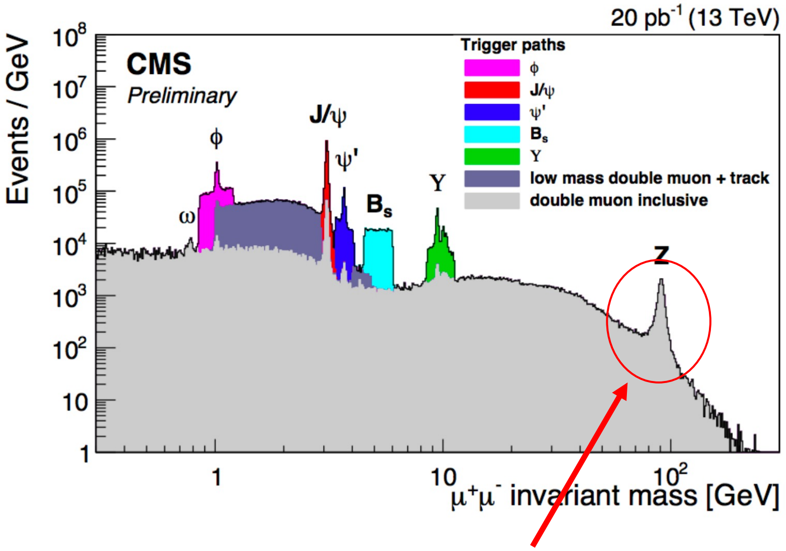
But What About Later Studies?

- Remember Bob’s Rule #3
  3. **The scientific effect involved is always at the very limit of detection.**
- There have been many studies since the original one, and many have claimed a statistically significant link between power lines and cancer (sometimes leukemia, sometimes other cancers), but after 30 years, *the significance has not increased.*
- How real science progresses…

First evidence of the Z boson (1983)

Won 1984 Nobel Prize for Carlo Rubbia and Simon Van Der Meer



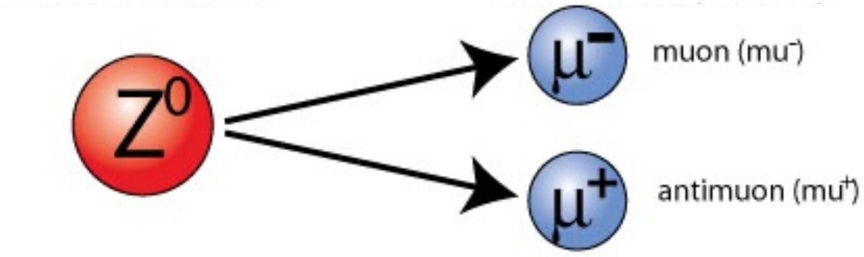


A tiny, tiny fraction of the Z bosons at the LHC

“Yesterday’s Nobel Prize is today’s background”

Statistical Errors vs. Systematic Errors

- Let’s consider that plot I just showed. The Z boson can decay to two muons, and we calculate its mass by measuring the momentum and trajectories of those particles.



- The result will have a *statistical error*, which will improved as $\frac{1}{\sqrt{N}}$
- It will also have a *systematic error*, based on things like
  - My knowledge of the muon mass
  - The energy calibration of my detector
  - An increase in statistics will not improve this!
- If statistical and systematic errors are comparable, they are often reported separately. For example, an early measurement of the Z mass was reported as

$$91192.0 \pm 6.4(\text{stat.}) \pm 4.0(\text{syst.})\ \text{MeV}$$

Errors on Combined Results

- Statistical errors on separate results are always “uncorrelated”, meaning they do not depend on each other.
- If I have two measurements of *exactly the same thing*, say

$$\begin{aligned} M_1 &= 6.52 \pm 0.21 \\ M_2 &= 6.38 \pm 0.13 \end{aligned}$$

uncorrelated errors add *in quadrature*

$$\begin{aligned} \overline{M} &= \frac{1}{2}\left(M_1 + M_2\right) = 6.45 \\ \Delta\overline{M} &= \frac{1}{2}\sqrt{(.21)^2 + (.13)^2} = .12 \end{aligned}$$

- Combining systematic errors, or results from different types of experiments can be much more complicated.
- We’ll talk about this much more in the “Bad Medicine” lecture, where it’s a huge issue.