

COMP229: Introduction to Data Science

Lecture 11: From Normality to reality

Olga Anosova, O.Anosova@liverpool.ac.uk
Autumn 2023, Computer Science department
University of Liverpool, United Kingdom

Lecture plan

- Some non-normal distributions.
- Hypothesis test with non-normal statistics.
- 6 principles of NHST.
- Dance of p -values and other black swans.
- The starting point of all measurements.

Last lecture recap

6 steps of hypothesis testing:

- Step 1: null hypothesis.
- Step 2(3): choosing a test.
- Step 3(2): checking the assumptions.
- Step 4: setting a significance level, below which the H_0 will be rejected.
- Step 5: calculating test statistic and p-value.
- Step 6: make a decision about the null hypothesis.

Recap quiz

Choose the correct answer:

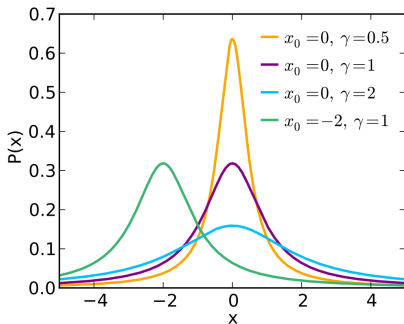
The *p-value* is equal to

- a) $P(H_0 | \text{get a result equal to or more extreme than observed})$
- b) $P(\text{get a result equal to or more extreme than observed} | H_0)$

We accept H_0 if and only if

- a) the *p-value* is (non-strictly) smaller than a specified significance level α ;
- b) the *p-value* is (non-strictly) larger than a specified significance level α ;
- c) the *p-value* is strictly equal to a specified significance level α .

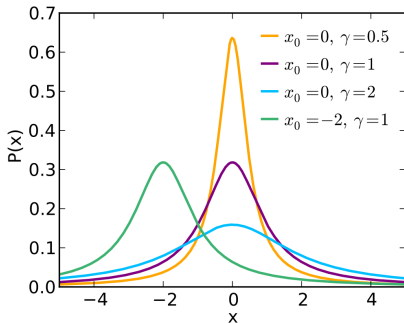
Choosing a distribution



By Skbkakas - Own work, CC BY 3.0, <https://commons.wikimedia.org/w/index.php?curid=9649146>

Is everything normal?

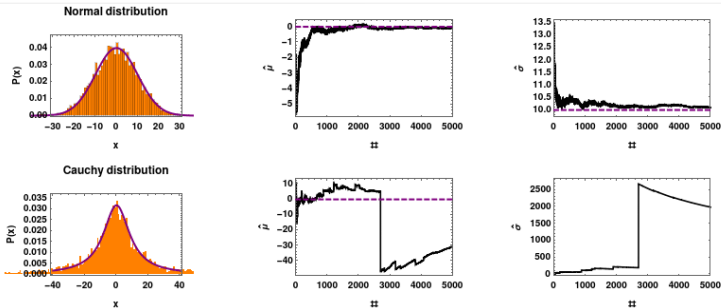
Definition 11.1. Cauchy random variable X has the density $f_{(m,\gamma)}(x) = \frac{1}{\pi\gamma} \left[\frac{\gamma^2}{(x-m)^2 + \gamma^2} \right]$, where m is the location of the peak and γ specifies the half-width at half-maximum.



Standard Cauchy distribution
has the probability density
function $f_{(0,1)}(x) = \frac{1}{\pi(1+x^2)}$.

Cauchy vs Normal

.gif can be seen [here](#)



Claim 11.2. If variables X_i are i.i.d. with a standard Cauchy density, then the average $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$ has the standard Cauchy density, hence can never become Normal.

Contradiction to CLT? No, we need **finite variance** in CLT.

Ratio distribution

The Cauchy distribution has *no mean or variance*, but mode and median are both equal to parameter m .

For two independent $U \sim N(0, 1)$ and $V \sim N(0, 1)$ the ratio U/V has the standard Cauchy distribution.

Non-normal distributions are not rare: most rates have “fat tails” and should be approached with caution.

Without the *finite variance* assumption, the limit may be a **stable** distribution (i.e. producing the same distribution in a linear combination) that is not normal.

Pareto distribution

Definition 11.3. **Pareto** random variable X is defined by **the survival (or tail) function**

$$\bar{F}(x) = \Pr(X > x) = \begin{cases} \left(\frac{x_m}{x}\right)^\alpha & x \geq x_m, \\ 1 & x < x_m, \end{cases}, \text{ where } x_m > 0 \text{ is the}$$

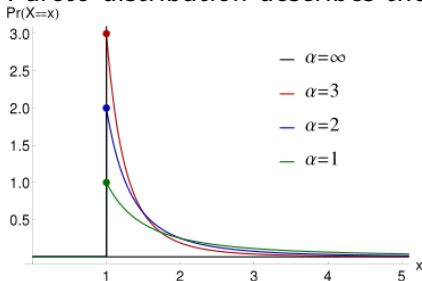
minimum possible value of X , and α is a positive shape parameter (known as a **tail index**, or **Pareto index**).

The PDF has the formula $f_X(x) = \begin{cases} \frac{\alpha x_m^\alpha}{x^{\alpha+1}} & x \geq x_m, \\ 0 & x < x_m. \end{cases}$, which

means Pareto is a **power-law** probability distribution.

80-20 rule

Pareto distribution describes the distribution of wealth:



PDF (here $x_m = 1$) demonstrates the Pareto principle or "80-20 rule" stating that 80% of outcomes are due to 20% of causes.

For different parameter values similar laws are true: for example, the **90-9-1 principle**.

Power-law distributions often have undefined variance.

This results in the so-called **black swan** behavior: the disproportionate role of high-profile rare events.

Fisher exact test

Fisher's exact test is used to determine from a contingency table if there is a nonrandom association between two categorical variables:

	Math Mag	Science	<i>total</i>
math	5	0	5
biology	1	4	5
<i>total</i>	6	4	10

consider maths and biology articles appearing in either Mathematics Magazine or Science journals.

$H_0 = \{\text{there is no association between the journal and type of article}\}.$

The exact formula

	Math Mag	Science	total
math	a	b	a+b
biology	c	d	c+d
total	a+c	b+d	a+b+c+d

R.Fisher showed that this distribution is **hypergeometric** and the probability of obtaining this sample is

$$p = \frac{\binom{a+b}{a} \binom{c+d}{c}}{\binom{n}{a+c}} = \frac{\binom{a+b}{b} \binom{c+d}{d}}{\binom{n}{b+d}} = \frac{(a+b)! (c+d)! (a+c)! (b+d)!}{a! b! c! d! (a+b+c+d)!},$$

where $\binom{n}{k}$ is the binomial coefficient.

Exact p -value

	Math Mag	Science	total
math	5	0	5
biology	1	4	5
total	6	4	10

$$P_{data} = \frac{(5!)^2 6! 4!}{5! 0! 1! 4! 10!} = 0.0238$$

Other possible cases with the same marginals are:

$$\begin{bmatrix} 4 & 1 \\ 2 & 3 \end{bmatrix} P=0.2381$$

$$\begin{bmatrix} 3 & 2 \\ 3 & 2 \end{bmatrix} P=0.4762$$

$$\begin{bmatrix} 2 & 3 \\ 4 & 1 \end{bmatrix} P=0.2381$$

$$\begin{bmatrix} 1 & 4 \\ 5 & 0 \end{bmatrix} P=0.0238,$$

The sum of p -values less than or equal to P_{data} is 0.0476, less than 0.05. This hints at a statistically significant association between the journal and type of article.

Exact (not approximated) numbers are the reason why the test is *exact*. For larger samples, because of large computations for factorials, the χ -squared test is used.

American Statistical Association statement

THE AMERICAN STATISTICIAN
2016, VOL. 70, NO. 2, 129–133
<http://dx.doi.org/10.1080/00031305.2016.1154108>



Taylor & Francis
Taylor & Francis Group

EDITORIAL

The ASA's Statement on p -Values: Context, Process, and Purpose

In February 2014, George Cobb, Professor Emeritus of Mathematics and Statistics at Mount Holyoke College, posed these questions to an ASA discussion forum:

Q: Why do so many colleges and grad schools teach $p = 0.05$?

A: Because that's still what the scientific community and journal editors use.

Q: Why do so many people still use $p = 0.05$?

A: Because that's what they were taught in college or grad school.

Cobb's concern was a long-worrisome circularity in the sociology of science based on the use of bright lines such as $p < 0.05$: "We teach it because it's what we do; we do it because it's what we teach." This concern was brought to the attention of the ASA Board.

2014) and a statement on risk-limiting post-election audits (American Statistical Association 2010). However, these were truly policy-related statements. The VAM statement addressed a key educational policy issue, acknowledging the complexity of the issues involved, citing limitations of VAMs as effective performance models, and urging that they be developed and interpreted with the involvement of statisticians. The statement on election auditing was also in response to a major but specific policy issue (close elections in 2008), and said that statistically based election audits should become a routine part of election processes.

By contrast, the Board envisioned that the ASA statement on p -values and statistical significance would shed light on an

The ASA has not previously taken positions on specific matters of statistical practice, since its foundation in 1839.

$p\text{-value} < \alpha$

“effect!”

$p\text{-value} \geq \alpha$

“no effect”

$\alpha = .05$, unless...

... the p-value fails

“arguably significant” ($P = 0.07$)

“direction heading to significance” ($P = 0.10$)

“flirting with conventional levels of significance” ($P > 0.1$)

“marginally significant” ($P \geq 0.1$)

Some of 500 terms used in articles:

a favourable statistical trend ($p=0.09$)

a little significant ($p<0.1$)

a margin at the edge of significance ($p=0.0608$)

a marginal trend ($p=0.09$)

a marginal trend toward significance ($p=0.052$)

a marked trend ($p=0.07$)

a mild trend ($p<0.09$)

a moderate trend toward significance ($p=0.068$)

a near-significant trend ($p=0.07$)

a negative trend ($p=0.09$)

a nonsignificant trend ($p<0.1$)

a nonsignificant trend toward significance ($p=0.1$)

a notable trend ($p<0.1$)

a numerical increasing trend ($p=0.09$)

a numerical trend ($p=0.09$)

a positive trend ($p=0.09$)

a possible trend ($p=0.09$)

a possible trend toward significance ($p=0.052$)

a pronounced trend ($p=0.09$)

a reliable trend ($p=0.058$)

a robust trend toward significance ($p=0.0503$)

a significant trend ($p=0.09$)

<u>P-VALUE</u>	<u>INTERPRETATION</u>
0.001	HIGHLY SIGNIFICANT
0.01	
0.02	
0.03	
0.04	SIGNIFICANT
0.049	
0.050	OH CRAP. REDO CALCULATIONS.
0.051	ON THE EDGE OF SIGNIFICANCE
0.06	
0.07	HIGHLY SUGGESTIVE, SIGNIFICANT AT THE $P<0.10$ LEVEL
0.08	
0.09	
0.099	
≥ 0.1	HEY, LOOK AT THIS INTERESTING SUBGROUP ANALYSIS

Multiple (potential) comparisons

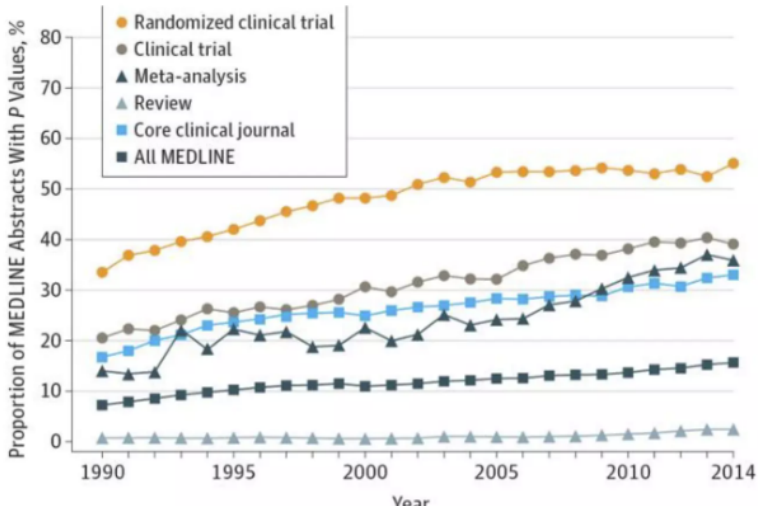
aka

- p-hacking
- data fishing
- data dredging
- multiple testing
- multiplicity
- significance chasing
- significance questing
- selective inference
- etc.

<u>P-VALUE</u>	<u>INTERPRETATION</u>
0.001	HIGHLY SIGNIFICANT
0.01	
0.02	
0.03	
0.04	SIGNIFICANT
0.049	
0.050	OH CRAP REDO CALCULATIONS.
0.051	ON THE EDGE OF SIGNIFICANCE
0.06	
0.07	HIGHLY SUGGESTIVE, SIGNIFICANT AT THE $P < 0.10$ LEVEL
0.08	
0.09	
0.099	HEY, LOOK AT THIS INTERESTING SUBGROUP ANALYSIS
≥ 0.1	

see also [Data dredging](#)

P-value increasingly central in reporting




The social sciences

False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant

Joseph P. Simmons¹, Leif D. Nelson², and Uri Simonsohn¹

¹The Wharton School, University of Pennsylvania, and ²Haas School of Business, University of California, Berkeley

Psychological Science
22(11) 1359–1366
© The Author(s) 2011
Reprints and permission:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/0956797611417632
<http://pss.sagepub.com>


RESEARCH ARTICLE

PSYCHOLOGY

Estimating the reproducibility of psychological science

Open Science Collaboration*†

SCIENCE | sciencemag.org

28 AUGUST 2015 • VOL. 349 ISSUE 6201

In popular media



TOPICS - TRENDING 2016 ELECTION



An unhealthy obsession with p-values is ruining science

Updated by Julia Belluz | @juliaoftoronto | julia.belluz@voxmedia.com | Mar 15, 2016, 11:00am EDT

Drastic measures...



Nerisa
@neri_peri



Volgen

Basic and Applied Social Psychology just went science rogue and banned NHST from their journal. Awesome.

[tandfonline.com/doi/full/10.10 ...](http://tandfonline.com/doi/full/10.10...)

Vertaling bekijken

RETWEETS

5

VIND-IK-LEUK

1



19:41 - 23 feb. 2015



5



1

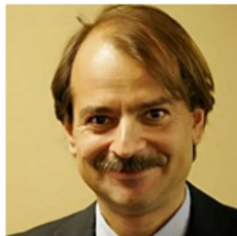


Essay

Why Most Published Research Findings Are False

John P. A. Ioannidis

Several methodologists have pointed out [9–11] that the high rate of nonreplication (lack of confirmation) of research discoveries is a consequence of the convenient, yet ill-founded strategy of claiming conclusive research findings solely on the basis of a single study assessed by formal statistical significance, typically for a p -value less than 0.05. Research is not most appropriately represented and summarized by p -values, but, unfortunately, there is a widespread notion that medical research articles should be interpreted based only on p -values.



ASA 6 principles of NHST

1. p -values can indicate how incompatible the data are with a specified **statistical model**, i.e. with the **web of assumptions**.
2. p -values do not measure the probability that H_0 is true, or the probability that the data were produced by random chance alone.

From a probability point of view

p-value*: $P(\mathbf{Data}|\mathbf{Hypothesis})$

is not: $P(\mathbf{Hypothesis}|\mathbf{Data})$

Does it matter?

$P(\mathbf{D}eath|\mathbf{H}andgun)$

= 5% to 20%*

$P(\mathbf{H}andgun|\mathbf{D}eath)$

= 0.028%**

If $p > .05$

Absence of evidence is not evidence of absence

Douglas G Altman, J Martin Bland

BMJ VOLUME 311 19 AUGUST 1995

ASA 6 principles of NHST

1. p -values can indicate how incompatible the data are with a specified **statistical model** with the **web of assumptions**.
2. p -values do not measure the probability that H_0 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 the fact that p -value passes/not passes a specific threshold.

On bright-line rules

“Practices that reduce data analysis or scientific inference to mechanical “bright-line” rules (such as “ $p < 0.05$ ”) for justifying scientific claims or conclusions can lead to erroneous beliefs and poor decision making. **A conclusion does not immediately become “true” on one side of the divide and “false” on the other.**”



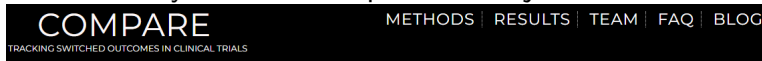
from the ASA statement

ASA 6 principles of NHST

1. p -values can indicate how incompatible the data are with a specified **statistical model** with the **web of assumptions**.
2. p -values do not measure the probability that H_0 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 the fact that p -value passes/not passes a specific threshold.
4. Proper inference requires full reporting and transparency.

Do strict rules help?

Medical studies should follow strict [set of rules](#), including pre-specified hypotheses. The [COMParE team](#) systematically checked every trial in the top 5 medical journals in 2015- 2016.



67

TRIALS CHECKED

9

TRIALS WERE
PERFECT

354

OUTCOMES NOT
REPORTED

357

NEW OUTCOMES
SILENTLY ADDED

On average, each trial reported just 58.2% of its specified outcomes. And on average, each trial silently added 5.3 new outcomes.

58

LETTERS SENT

18

LETTERS PUBLISHED

8

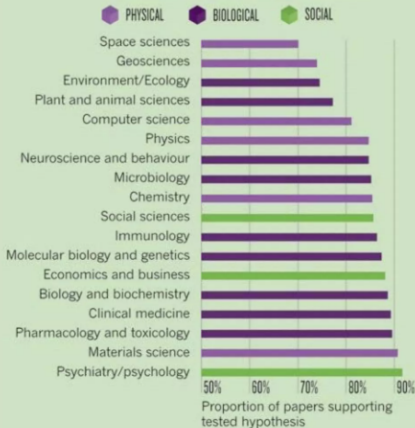
LETTERS
UNPUBLISHED AFTER
4 WEEKS

32

LETTERS REJECTED BY
EDITOR

ACCENTUATE THE POSITIVE

A literature analysis across disciplines reveals a tendency to publish only 'positive' studies — those that support the tested hypothesis. Psychiatry and psychology are the worst offenders.



Why is this enormous positivity?

besides journal editors requirement for $p < .05$



If you torture the data long enough,
it will confess to anything

Ronald Coase

In the large ('big') data era

“With a combination of **large datasets**, confounding, flexibility in analytical choices ..., and superimposed selective reporting bias, using a **$P < 0.05$** threshold to declare “success,”
means next to nothing.”



From ASA supplementary material, response by Ioannidis.

Story of a discovery

Suppose we are assessing the probability p of a certain outcome in experimental trials, and set $H_0 = \{p = 0.5\}$.

Adam, a scientist, conducted 12 trials and obtains 3 successes and 9 failures. One of those successes was the 12th and last observation. Then Adam left the lab for good.

Bill, a colleague in the same lab, continued Adam's work, and did a significance test:

$P(\text{in 12 trials 3 or fewer (i.e. more extreme) successes}) =$

$$\left[\binom{12}{9} + \binom{12}{10} + \binom{12}{11} + \binom{12}{12} \right] \left(\frac{1}{2} \right)^{12} = \frac{299}{4096} = 0.073$$

The null hypothesis is not rejected at the 5% significance level.

Discovery story, continued

Charlotte, another scientist, reads Bill's paper and writes a letter, saying that it is possible that Adam kept trying *until he obtained 3 successes*, in which case we need

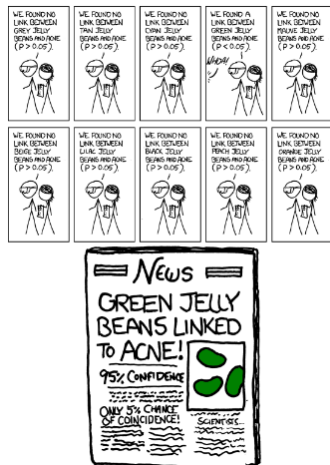
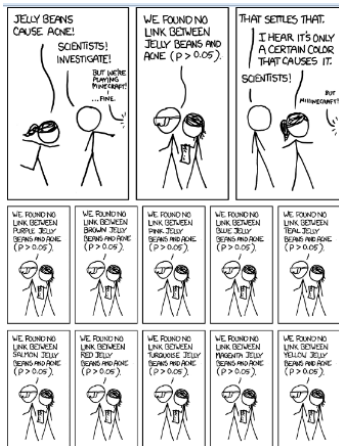
$P(\text{conduct 12 or more experiments}) =$

$$1 - \left[\binom{10}{2} \left(\frac{1}{2}\right)^{11} + \binom{9}{2} \left(\frac{1}{2}\right)^{10} + \cdots + \binom{2}{2} \left(\frac{1}{2}\right)^3 \right] = 0.0327$$

Now the result is statistically significant at the 5% level!

*The difference lies not in the actual data collected, nor in the conduct of the experimenter, but in the two different **designs** of the experiment.*

p-hacking problem



If 20 independent tests are conducted at the 0.05 significance level, the expected number of false significant result is $0.05 \times 20 = 1$. Thanks [xkcd](#)

In every joke there is a part of a joke

Science journalist J.Bohannon, "[Chocolate with high Cocoa content as a weight-loss accelerator](#)" (with film-maker Peter Onneken, who was making a film about junk science in the diet industry), is a deliberately bad study that he had designed and run to see how the media would pick up the "meaningless" findings.

Sample size: 15,

Number of variables: 18 for a guaranteed false positive.

He asked a statistician for help in deliberately torturing the data, with overfitting and p-hacking, etc.

Conclusion: *eating chocolate could assist with weight loss.*

Was it published? Was it covered in media?

Chocolate with high Cocoa content as a weight-loss accelerator

ORIGINAL

Johannes Bohannon¹,
Diana Koch¹,
Peter Homm¹,
Alexander Drieaus¹

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.

Methodology: In a study consisting of several weeks, we divided men and women between the ages of 19-67 into three groups. One group was instructed to keep a low-carb diet and to consume an additional daily serving of 42 grams of chocolate with 81% cocoa content (chocolate group). Another group was instructed to follow the same low-carb diet as the chocolate group, but without the chocolate intervention (low-carb group). In addition, we asked a third group to eat at their own discretion, with unrestricted choice of food. At the beginning of the study, all participants received extensive medical advice and were thoroughly briefed on their respective

SHOCK statistics

The goal of ANDROMEDA-SHOCK randomised controlled trial (RCT) was to compare two treatment strategies (CRT or the usual lactate-based) for septic shock treatment.

Sample size: 400 patients, monitored for 28 days.

Results: CRT was better, the mortality rate was 8.5% lower than for the lactate, Hazard Ratio was a 25% better.

But: in trial design the researchers had assumed a large 15% mortality reduction, so the number of participants was underestimated and positive results failed to cross the $p = 0.05\%$ line.

Conclusion: the CRT strategy “did not reduce all-cause 28-day mortality”.

That wasted case [shocked statisticians](#).

ASA 6 principles of NHST

1. p -values can indicate how incompatible the data are with a specified **statistical model** with the **web of assumptions**.
2. p -values do not measure the probability that H_0 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 the fact that p -value passes/not passes a specific threshold.
4. Proper inference requires full reporting and transparency.
5. p -value and statistical significance do not measure the size of an effect or the importance of a result.

Dance of p -values

You can try it out [here](#).

ASA 6 principles of NHST

1. p -values can indicate how incompatible the data are with a specified **statistical model** and the **web of assumptions**.
2. p -values do not measure the probability that H_0 is true, or that the data were produced by random chance alone.
3. Scientific conclusions and business or policy decisions should not be based only on the fact that p -value passes/not passes a specific threshold.
4. Proper inference requires full reporting and transparency.
5. P -value and statistical significance do 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 H_0 .

Practical rules of NHST

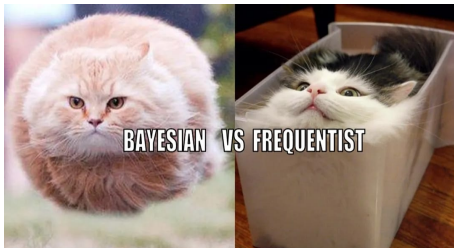
1. Check the assumptions!!
 2. Avoid statements about the **truth** of H_0 .
 3. Report the whole process, including underlying definitions and data ~~torture~~ exploration.
 4. Avoid statements about effect of absence of effect.
 5. Avoid statements about the effect size.
 6. Use multiple types of analysis and measures, and **think!**
- Are there any easier and **more computerizable** ways?

Bayesian stats to the rescue?

Bayesian principles, very roughly:

- If we can't repeat an experiment, just guess the probabilities.
- Then get a real distribution from measuring many repeated guesses.

...oh wait... that's frequentist statistics, but applied to *personal perception* instead of experiments themselves!



From Statistics to ML

- **No tool is sacred**, but **any tool can be harmful** without proper understanding of underlying mechanisms.
- Application of fully automated **computerizable** methods requires even **better understanding**.

Machine Learning tools are rooted in statistical principles, most of ML ideas can be reworded in statistical terms.

This is nicely demonstrated in “Machine Learning: a Probabilistic Perspective” by Kevin Patrick Murphy. MIT Press, 2012.

But it works!

"But it works!"

How do we know? How do we measure?

"It works better than when others tried it."

Define better. Define worse. Define **equal**.

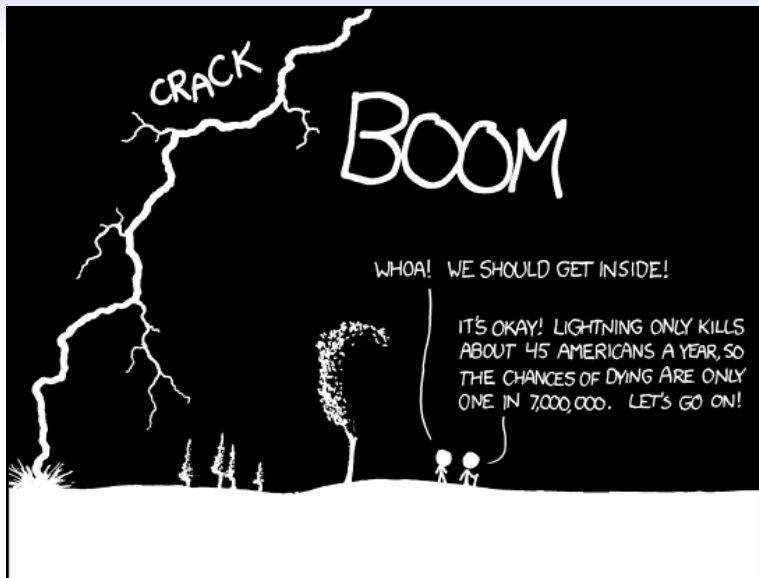
*"All animals are equal,
but some animals are more equal than others"*

George Orwell, "Animal Farm" (1945)

What do we call equal? That's the topic of the next lecture.

Time to revise and ask questions

- Not everything is Normal; **ratios** and **power law** distributions are very different.
- Rules of NHST:
 1. Check the assumptions!!
 2. Avoid statements about the truth of H_0 .
 3. Report the whole process, including underlying definitions and data torture exploration.
 4. Avoid statements about effect of absence of effect.
 5. Avoid statements about the effect size.
 6. Use multiple types of analysis and **measures**.



THE ANNUAL DEATH RATE AMONG PEOPLE
WHO KNOW THAT STATISTIC IS ONE IN SIX.