# Statistical Methods in Particle Physics

## 1. Basic Concepts

Heidelberg University, WS 2020/21

Klaus Reygers (lectures)
Rainer Stamen, Martin Völkl (tutorials)

# 1.1 Introduction

# Contents



"Bayes vs. frequentist"

# Learning goals and required knowledge

This course is a natural follow-up to PEP4 for **Bachelor students** interested in Particle Physics. **Master students** are invited to attend this lecture in parallel or after the Particle Physics course.

**Learning goals**

- Get to know and apply the toolbox of statistical methods used in particle physics
- Understand error bars and confidence limits as reported in publications
- Solid understanding of maximum likelihood and least squares fits
- From measurement to message: which conclusion can you draw from your data (and which not)?
- Learn to apply machine learning methods

**Required knowledge**

- Basic understanding of experimental particle physics (as taught in the bachelor's course)
- Basic knowledge of python is helpful

# Practical information (I)

- Website
  - https://uebungen.physik.uni-heidelberg.de/vorlesung/20202/1225
- Lecture
  - Flipped/inverted class room!
  - Links to slides and videos provided before Thursday meeting
  - Contact time on Thursdays:
    - Via Zoom
    - Questions / discussion, questions can be sent beforehand, e.g. through rocket.chat [room: “ws20-smipp”]) or email
    - Quizzes
    - Typical **no comprehensive repetition** of the contents of the lecture videos!
  - Exact time for Thursday meeting flexible, for example:
    - Thursdays, 16:00-17:00: time to study slides / videos
    - Thursdays, 17:00-17:30: Zoom-Meeting: discussion / quizzes
    - First Thursday meeting (5 Nov 2020) starts at 16:15

# Practical information (II)

- Tutorials
  - Mondays, 16:00 – 17:45
  - Zoom
  - Weekly problem sheets, to be handed in on Thursday before 12:00
  - See next slide for detailed schedule
- Exam
  - There will be a written exam at the end of the semester
  - Refers to contents of lectures and exercises
  - 60% of the points of the homework sheets required to be eligible to write the exam
  - Date to be fixed

# Schedule

| Week | Tutorial Monday | Lecture Thursday | Exercise Sheets Hand in (12:00) | Hand out (after L.) |
|---|---|---|---|---|
| 45: 2.11. – 8.11. | Intro, Python basic | Lecture 01 | -- | Sheet 01 |
| 46: 9.11. – 15.11. | online exercises, Python adv. | Lecture 02 | Sheet 01 | Sheet 02 |
| 47: 16.11. – 22.11. | Sheet 01 | Lecture 03 | Sheet 02 | Sheet 03 |
| 48: 23.11. – 29.11. | Sheet 02 | Lecture 04 | Sheet 03 | Sheet 04 |
| 49: 30.11. – 6.12. | Sheet 03 | Lecture 05 | Sheet 04 | Sheet 05 |
| 50: 7.12. – 13.12. | Sheet 04 | Lecture 06 | Sheet 05 | Sheet 06 |
| 51: 14.12 – 20.12. | Sheet 05 | Lecture 07 | Sheet 06 | Sheet 07 |
| Xmas holidays: 3 weeks | | | | |
| 2: 11.1. – 17.1. | Sheet 06 | Lecture 08 | Sheet 07 | Sheet 08 |
| 3: 18.1. – 24.1. | Sheet 07 | Lecture 09 | Sheet 08 | Sheet 09 |
| 4: 25.1. – 31.1. | Sheet 08 | Lecture 10 | Sheet 09 | Sheet 10 |
| 5: 1.2. – 7.1. | Sheet 09 | Lecture 11 | Sheet 10 | |
| 6: 8.1. - 14.2. | Sheet 10 | Lecture 12 | | |
| 7: 15.1. – 21.1. | | Study Week | | |
| 8: 22.1. – 28.1. | Exam week (date to be confirmed) | | | |

# Useful books

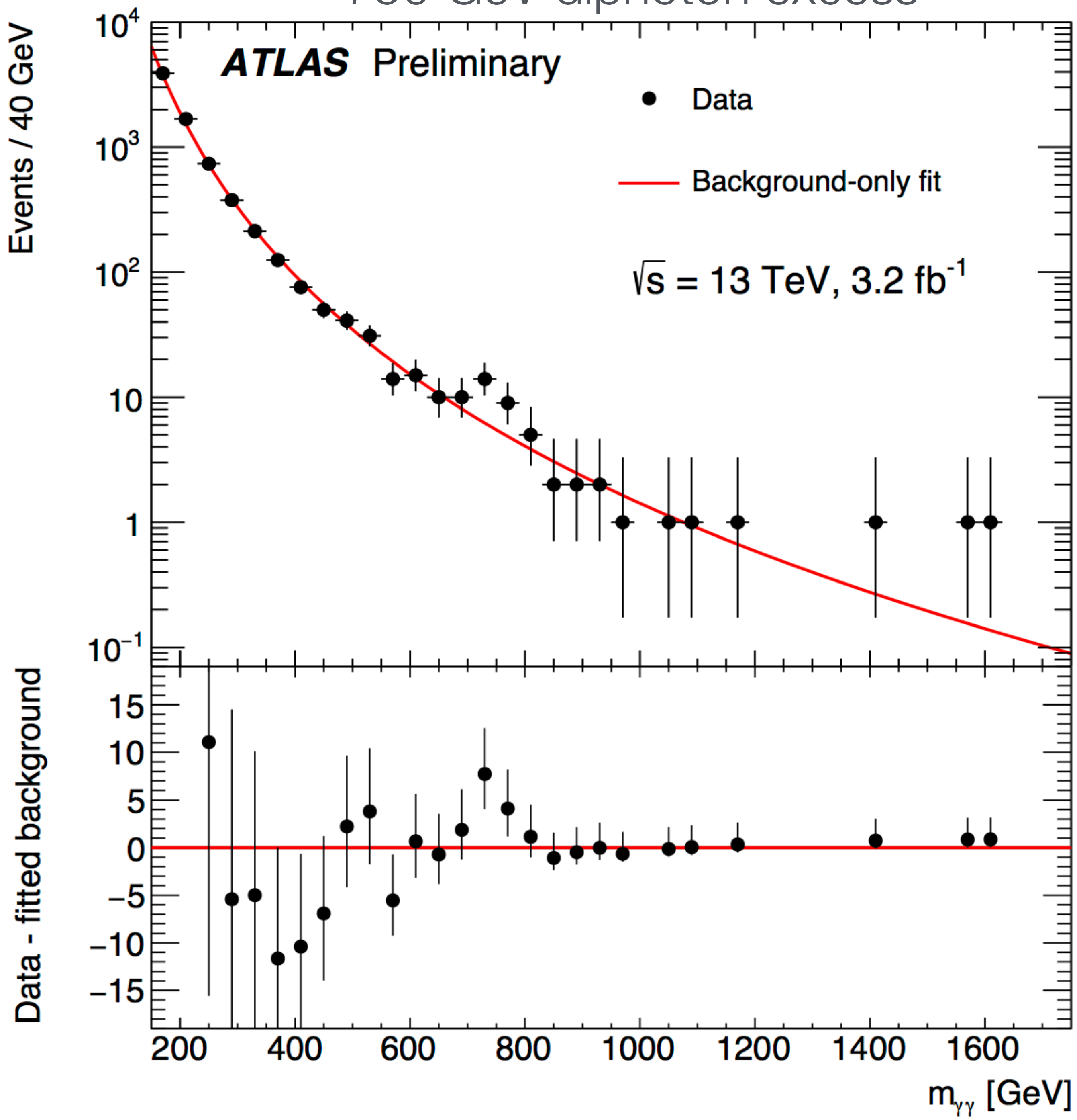
- G. Cowan, Statistical Data Analysis
- L. Lista, Statistical Methods for Data Analysis in Particle Physics
- Behnke, Kroeninger, Schott, Schoerner-Sadenius: Data Analysis in High Energy Physics: A Practical Guide to Statistical Methods
- R. Barlow, Statistics: A Guide to the Use of Statistical Methods in the Physical Sciences
- Bohm, Zech, Introduction to Statistics and Data Analysis for Physicist [available online]
- Blobel, Lohrmann: Statistische Methoden der Datenanalyse (in German), [free ebook]
- L. Lyons:
  Statistics for Nuclear and Particle Physicists (Cambridge University Press)
- F. James, Statistical Methods in Experimental physics
- W. Metzger, Statistical Methods in Data Analysis [available online]

# Further Material

- Glen Cowan: http://www.pp.rhul.ac.uk/~cowan/stat_course.html
- Scott Oser: http://www.phas.ubc.ca/~oser/p509/
- Terascale Statistics School:
  https://indico.desy.de/indico/event/25594/other-view?view=standard
- Particle Data Group reviews on Probability and Statistics
  - https://pdg.lbl.gov/2020/reviews/rpp2020-rev-probability.pdf
  - https://pdg.lbl.gov/2020/reviews/rpp2020-rev-statistics.pdf

# Why bother with statistical methods?

"750 GeV diphoton excess"




Statistics:
Draw reliable conclusions from data

In case of doubt:
just get more data …

Yes, but not always easy …

A heavy Higgs boson?

Peak disappeared with more data … [link]

Presentations by CMS and ATLAS, December 2015:
https://indico.cern.ch/event/442432/

# How knowledge is created?

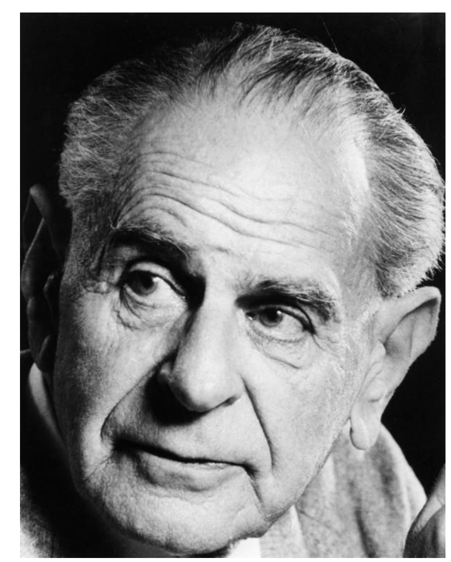
## Guess theory/model

- usually mathematical
- self-consistent
- simple explanations, few arbitrary parameters
- testable predictions

## Perform experiment

- reject / modify theory in case of disagreement with data
- if theory requires too many adjustments it becomes unattractive

**The advance of scientific knowledge is an *evolutionary* process**



source: Wikipedia

Karl Popper (1902–1994)

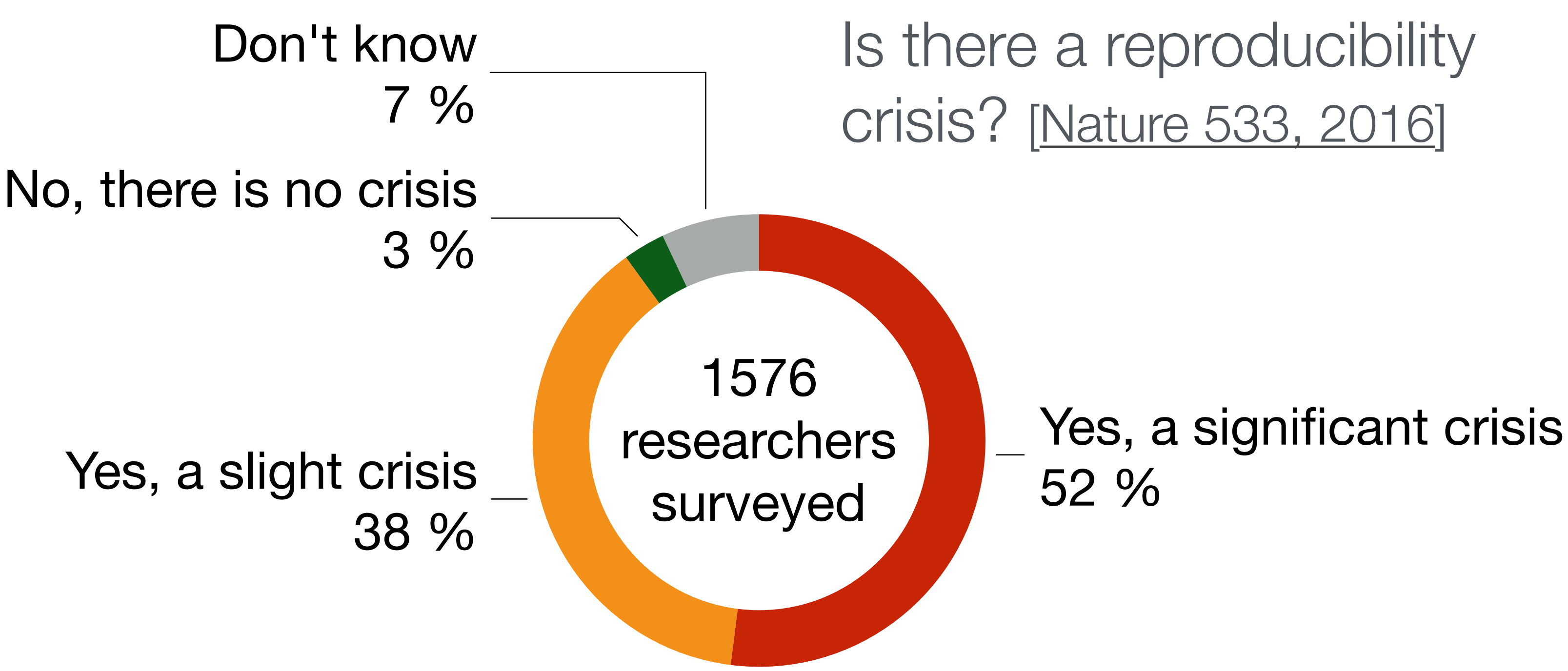
Statistical methods are an important part of this process

# A look at other research fields

"**Why Most Published Research Findings Are False**":
Main thesis: large number, if not the majority, of published medical research papers contain results that cannot be replicated.

**Reproducibility crisis**:
Affects the social sciences and life sciences most severely (in particular psychology)

Is there a reproducibility crisis? [Nature 533, 2016]





Open access, freely available online

Essay

## Why Most Published Research Findings Are False

John P. A. Ioannidis

**Summary**

There is increasing concern that most current published research findings are false. The probability that a research claim is true may depend on study power and bias, the number of other studies on the same question, and, importantly, the ratio of true to no relationships among the relationships probed in each scientific field. In this framework, a research finding is less likely to be true when the studies conducted in a field are smaller; when effect sizes are smaller; when there is a greater number and lesser preselection of tested relationships; where there is greater flexibility in designs, definitions, outcomes, and analytical modes; when there is greater financial and other interest and prejudice; and when more teams are involved in a scientific field in chase of statistical significance. Simulations show that for most study designs and settings, it is more likely for a research claim to be false than true. Moreover, for many current scientific fields, claimed research findings may often be simply accurate measures of the prevailing bias. In this essay, I discuss the implications of these problems for the conduct and interpretation of research.

Published research findings are sometimes refuted by subsequent evidence, with ensuing confusion and disappointment. Refutation and controversy is seen across the range of research designs, from clinical trials and traditional epidemiological studies [1–3] to the most modern molecular research [4,5]. There is increasing concern that in modern research, false findings may be the majority or even the vast majority of published research claims [6–8]. However, this should not be surprising. It can be proven that most claimed research findings are false. Here I will examine the key factors that influence this problem and some corollaries thereof.

The Essay section contains opinion pieces on topics of broad interest to a general medical audience.

**Modeling the Framework for False Positive Findings**

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

It can be proven that most claimed research findings are false.

should be interpreted based only on *p*-values. Research findings are defined here as any relationship reaching formal statistical significance, e.g., effective interventions, informative predictors, risk factors, or associations. "Negative" research is also very useful. "Negative" is actually a misnomer, and the misinterpretation is widespread. However, here we will target relationships that investigators claim exist, rather than null findings.

As has been shown previously, the probability that a research finding is indeed true depends on the prior probability of it being true (before doing the study), the statistical power of the study, and the level of statistical significance [10,11]. Consider a $2 \times 2$ table in which research findings are compared against the gold standard of true relationships in a scientific field. In a research field both true and false hypotheses can be made about the presence of relationships. Let $R$ be the ratio of the number of "true relationships" to "no relationships" among those tested in the field. $R$ is characteristic of the field and can vary a lot depending on whether the field targets highly likely relationships or searches for only one or a few true relationships among thousands and millions of hypotheses that may be postulated. Let us also consider, for computational simplicity, circumscribed fields where either there is only one true relationship (among many that can be hypothesized) or the power is similar to find any of the several existing true relationships. The pre-study probability of a relationship being true is $R/(R + 1)$. The probability of a study finding a true relationship reflects the power $1 - \beta$ (one minus the Type II error rate). The probability of claiming a relationship when none truly exists reflects the Type I error rate, $\alpha$. Assuming that $c$ relationships are being probed in the field, the expected values of the $2 \times 2$ table are given in Table 1. After a research finding has been claimed based on achieving formal statistical significance, the post-study probability that it is true is the positive predictive value, PPV. The PPV is also the complementary probability of what Wacholder et al. have called the false positive report probability [10]. According to the $2 \times 2$ table, one gets $PPV = (1 - \beta)R/(R - \beta R + \alpha)$. A research finding is thus

**Citation:** Ioannidis JPA (2005) Why most published research findings are false. PLoS Med 2(8): e124.



**Abbreviation:** PPV, positive predictive value

John P. A. Ioannidis is in the Department of Hygiene and Epidemiology, University of Ioannina School of Medicine, Ioannina, Greece, and Institute for Clinical Research and Health Policy Studies, Department of Medicine, Tufts-New England Medical Center, Tufts University School of Medicine, Boston, Massachusetts, United States of America. E-mail: jioannid@cc.uoi.gr

**Competing Interests:** The author has declared that no competing interests exist.

**DOI:** 10.1371/journal.pmed.0020124

PLoS Medicine | www.plosmedicine.org 0696 August 2005 | Volume 2 | Issue 8 | e124

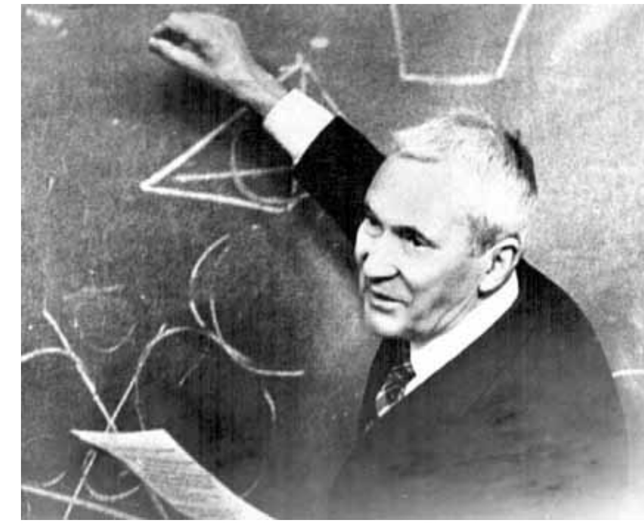
John Ioannidis
(Stanford School of Medicine)
PLoS Med 2(8): e124., (2005),
doi:10.1371/journal.pmed.0020124

# Sources of uncertainty

- Underlying theory (quantum mechanics) is probabilistic
  - true randomness
- Limited knowledge about the measurement process
  - present even without quantum mechanics

We quantify uncertainty using **probability**

# Mathematical definition of probability



Kolmogorov, 1933

Let $A$ be an event. Then probability is a number obeying three conditions, the *Kolmogorov axioms*:

1. $P(A) \geq 0$ (non-negative real number)
2. $P(S) = 1$, where $S$ is the set of all $A$, the sample space
3. $P(A \cup B) = P(A) + P(B)$ for any $A$, $B$ which are exclusive, i.e., $A \cap B = 0$

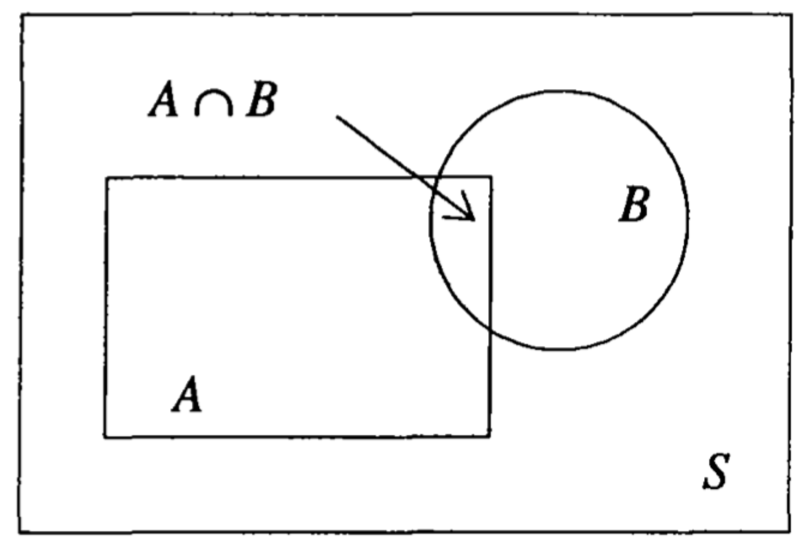
From these axioms further properties can be derived, e.g.:

$P(\bar{A}) = 1 - P(A)$

$P(\emptyset) = 0$

if $A \subset B$ then $P(A) \leq P(B)$

$P(A \cup B) = P(A) + P(B) - P(A \cap B)$




**But what does *P* mean?**

# Interpretations of probability

https://plato.stanford.edu/entries/probability-interpret/

- Classical
  - Assign equal probabilities based on symmetry of the problem, e.g., rolling dice: $P(6) = 1/6$
  - difficult to generalize

- Frequentist
  - Let $A$, $B$, … be outcomes of an repeatable experiment:

$$P(A) = \lim_{n \to \infty} \frac{\text{times outcome is } A}{n}$$

- Bayesian (subjective probability)
  - $A$, $B$, … are hypotheses (statements that are true or false)

$$P(A) = \text{degree of believe that } A \text{ is true}$$

# Criticisms of the probability interpretations

- Criticisms of the frequency interpretation
  - $n \to \infty$ can never be achieved in practice. When is $n$ large enough?
  - $P$ is not an intrinsic property of $A$, it depends on the how the ensemble of possible outcomes was constructed
    - Example: $P$(patient is treated in hospital | positive Corona test) is different wether or not one knows the age of the person
  - We want to talk about the probability of events that are not repeatable
    - Example 1: $P$(it will rain tomorrow), but there is only one tomorrow
    - Example 2: $P$(Universe started with a Big Bang), but only one universe

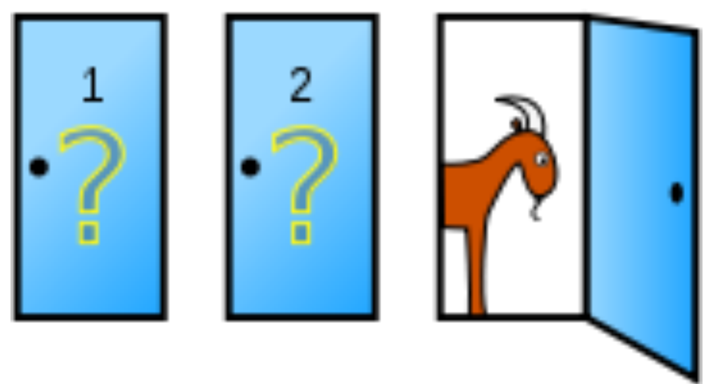
- Criticisms of the subjective Bayesian interpretation
  - "Subjective" estimates have no place in science
  - How to quantify the prior state of our knowledge upon which we base our probability estimate?

# Fun with probabilities

https://en.wikipedia.org/wiki/Monty_Hall_problem

## Monty Hall problem ("Ziegenproblem")

Suppose you're on a game show, and you're given the choice of three doors: Behind one door is a car; behind the others, goats. You pick a door, say No. 1, and the host, who knows what's behind the doors, opens another door, say No. 3, which has a goat. He then says to you, "Do you want to pick door No. 2?" Is it to your advantage to switch your choice?




## Standard assumptions

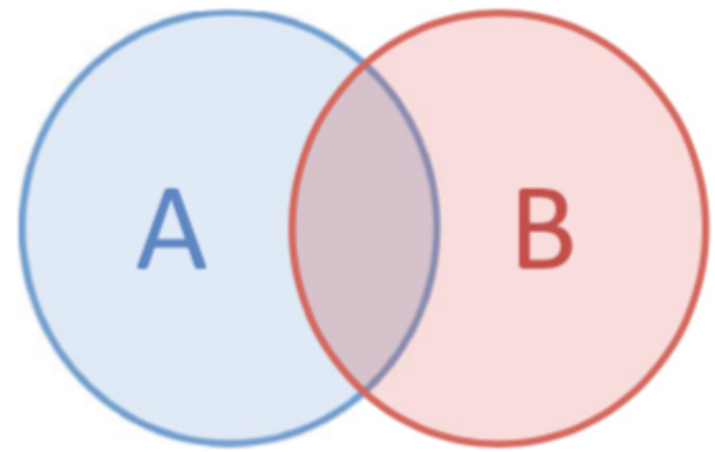
- The host must always open a door that was not picked by the contestant
- The host must always open a door to reveal a goat and never the car.
- The host must always offer the chance to switch between the originally chosen door and the remaining closed door.

**Under these assumptions you should switch your choice!**

# Conditional probability and independent events

For two events A and B, the conditional probability is defined as

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

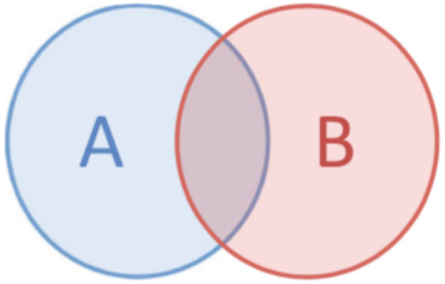



Example: rolling dice: $P(n < 3|n \text{ even}) = \frac{P((n < 3) \cap n \text{ even})}{P(n \text{ even})} = \frac{1/6}{1/2} = 1/3$

$$\text{Events A and B independent} \iff P(A \cap B) = P(A) \cdot P(B)$$

An event A is independent of B if $P(A|B) = P(A)$

# Bayes' theorem



Venn diagram

Definition of conditional probability:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \quad \text{and} \quad P(B|A) = \frac{P(B \cap A)}{P(A)}$$

$$P(A \cap B) = P(B \cap A) \quad \rightarrow \quad \boxed{P(A|B) = \frac{P(B|A)P(A)}{P(B)}}$$

[doubtful whether the portrait actually shows Bayes]

First published (posthumously) by the Reverend Thomas Bayes (1702–1761)

First modern formulation by Pierre-Simon Laplace in 1812

Accepted by everyone if probabilities are not Bayesian probabilities

# Example of using Bayes' theorem: Test for a rare disease

Base probability (for anyone) to have a disease D:

$$P(\mathrm{D}) = 0.001$$
$$P(\mathrm{no\ D}) = 0.999$$

Consider a test for the disease: result is positive or negative (+ or –):

"sensitivity" → $P(+|\mathrm{D}) = 0.98$

$P(-|\mathrm{D}) = 0.02$

$P(+|\mathrm{no\ D}) = 0.03$

"specificity" → $P(-|\mathrm{no\ D}) = 0.97$

Suppose your result is +. How worried should you be?

$$P(\mathrm{D}|+) = \frac{P(+|\mathrm{D})P(\mathrm{D})}{P(+)} = \frac{P(+|\mathrm{D})P(\mathrm{D})}{P(+|\mathrm{D})P(\mathrm{D}) + P(+|\mathrm{no\ D})P(\mathrm{no\ D})}$$
$$= \frac{0.98 \times 0.001}{0.98 \times 0.001 + 0.03 \times 0.999} = 0.032$$

Probability for you to have the disease is 3.2%, i.e., you're probably ok.

Remark: false positives **not** a relevant issue in statistics of Corona cases
(in case of a positive result usually double checks are made resulting in very high specificity)

# Bayesian inference: Degree of believe in a theory given a certain set of data (I)

probability of getting the data if theory is true

prior (subjective belief in the theory before seeing the data)

$$P(\text{theory}|\text{data}) = \frac{P(\text{data}|\text{theory})P(\text{theory})}{P(\text{data})}$$

posterior probability, i.e., subjective belief in the theory after seeing the data

guarantees normalization:

$$P(\text{data}) = \sum_i P(\text{data}|\text{theory}_i)P(\text{theory}_i)$$

Addresses question: "What should I believe?"

# Bayesian inference: Degree of believe in a theory given a certain set of data (II)

For a continuous parameter λ:

$$P_{\text{posterior}}(\lambda|m) = \frac{f(m|\lambda) P_{\text{prior}}(\lambda)}{f_1(m)}$$

$\lambda$ : true value of a parameter of nature

$m$ : measurement

$$f_1(m) = \int f(m|\lambda') P(\lambda')\, \mathrm{d}\lambda'$$

**Problems with Bayesian inference**

What functional form to chose for $P_{\text{prior}}(\lambda)$?

Uninformed prior: flat? In which variable, e.g., in $\lambda, \lambda^2, 1/\lambda, \ln \lambda$?

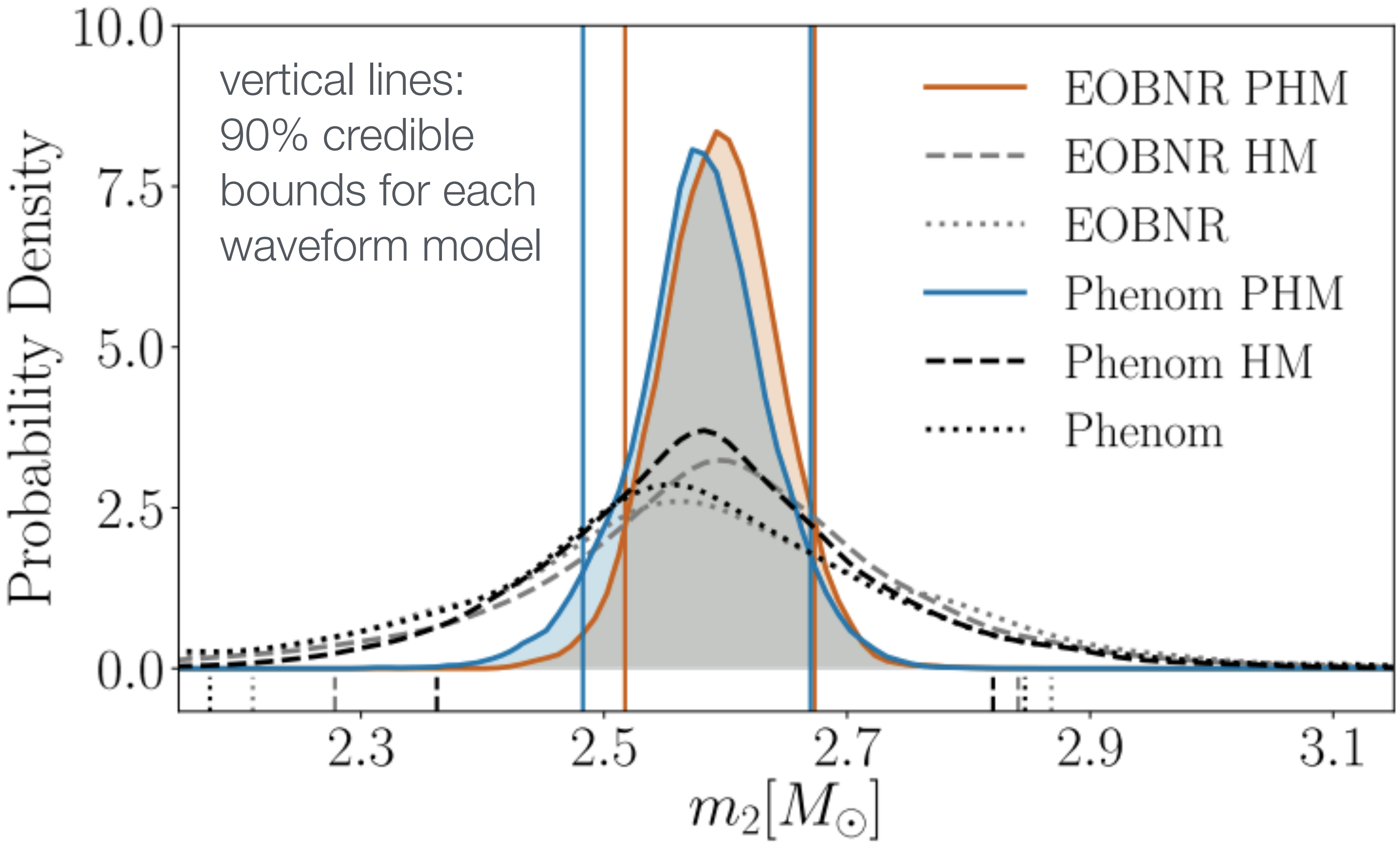
**Bayesian reply**

Choice of prior usually unimportant after a few experiments

Jaynes' robot: Priors are uniquely determined by your state of knowledge. Thus scientists with the same background knowledge construct the same priors.

# Example of a posterior distribution

GW190814: Gravitational waves from the coalescence of a 23 solar mass Black Hole with a 2.6 solar mass compact object




Posterior distribution for mass of the lighter objects:

LIGO Scientific Collaboration and Virgo Collaboration:
The Astrophysical Journal Letters, 896:L44 (20pp), 2020 June 20

Note:
Sampling from a multi-parameter posterior distribution typically involves Markov chain Monte Carlo (MCMC)

# Are you a Frequentist or a Bayesian?

Suppose mass of a particle is measured with Gaussian resolution $\sigma$ and the result ist reported as

$$m \pm \sigma$$

**Bayesian**

$$P(m|m_{\text{true}}) \propto e^{-(m-m_{\text{true}})^2/(2\sigma^2)} \quad \xrightarrow{\text{flat prior for } m_{\text{true}}} \quad P(m_{\text{true}}|m) \propto e^{-(m-m_{\text{true}})^2/(2\sigma^2)}$$

**Frequentist**

This is a statement about the interval [$m$–σ, $m$+σ]. For a large number of hypothetically repeated experiments the interval would contain the true value in 68% of the cases. In the frequentist approach, a probabilistic statement about the true value is nonsense (the true value is what it is).

"Bayesians address the questions everyone is interested in by using assumptions that no one believes. Frequentist use impeccable logic to deal with an issue that is of no interest to anyone." – Louis Lyons

# Bayesian inference: Jeffreys' prior

## How to model complete ignorance about the value of a parameter θ?

- Uniform distribution in θ, exp θ, ln θ, 1/θ, …?
- Example: Lifetime τ of a particle, uniform distribution in τ or particle's width Γ = 1/τ ?

## Jeffreys' prior (non-informative prior) for a model $L(\vec{x}|\vec{\theta})$ of the measurement:

$$\pi(\vec{\theta}) \propto \sqrt{I(\vec{\theta})} \qquad I(\vec{\theta}) = \det\left[\left\langle \frac{\partial \ln L(\vec{x}|\vec{\theta})}{\partial \theta_i} \frac{\partial \ln L(\vec{x}|\vec{\theta})}{\partial \theta_j} \right\rangle\right]$$

invariant under re-parameterization

determinant of the Fisher information matrix

expectation value evaluated by $\vec{x}$
integrating over all possible results

## Examples:

| **PDF parameter** | **Jeffreys' prior** |
|---|---|
| **Poissonian mean μ** | **p(μ) ∝ 1/√μ** |
| **Gaussian mean μ** | **p(μ) ∝ 1** |

# Jeffreys' prior: Example

Exponential distribution: $$L(t \mid \tau) = \frac{1}{\tau} e^{-t/\tau}$$

Jeffreys' prior: $$\pi(\tau) \propto \sqrt{I(\tau)} = \sqrt{\mathrm{E}\left[\left(\frac{d}{d\tau} \ln L(t \mid \tau)\right)^2\right]}$$

$$\frac{d}{d\tau} \ln L(t|\tau) = -\frac{1}{\tau} + \frac{t}{\tau^2}$$

$$E\left[\left(\frac{t}{\tau^2} - \frac{1}{\tau}\right)^2\right] = E\left[\left(\frac{t-\tau}{\tau^2}\right)^2\right] = \frac{1}{\tau^4} V[t] = \frac{\tau^2}{\tau^4} = \frac{1}{\tau^2}$$

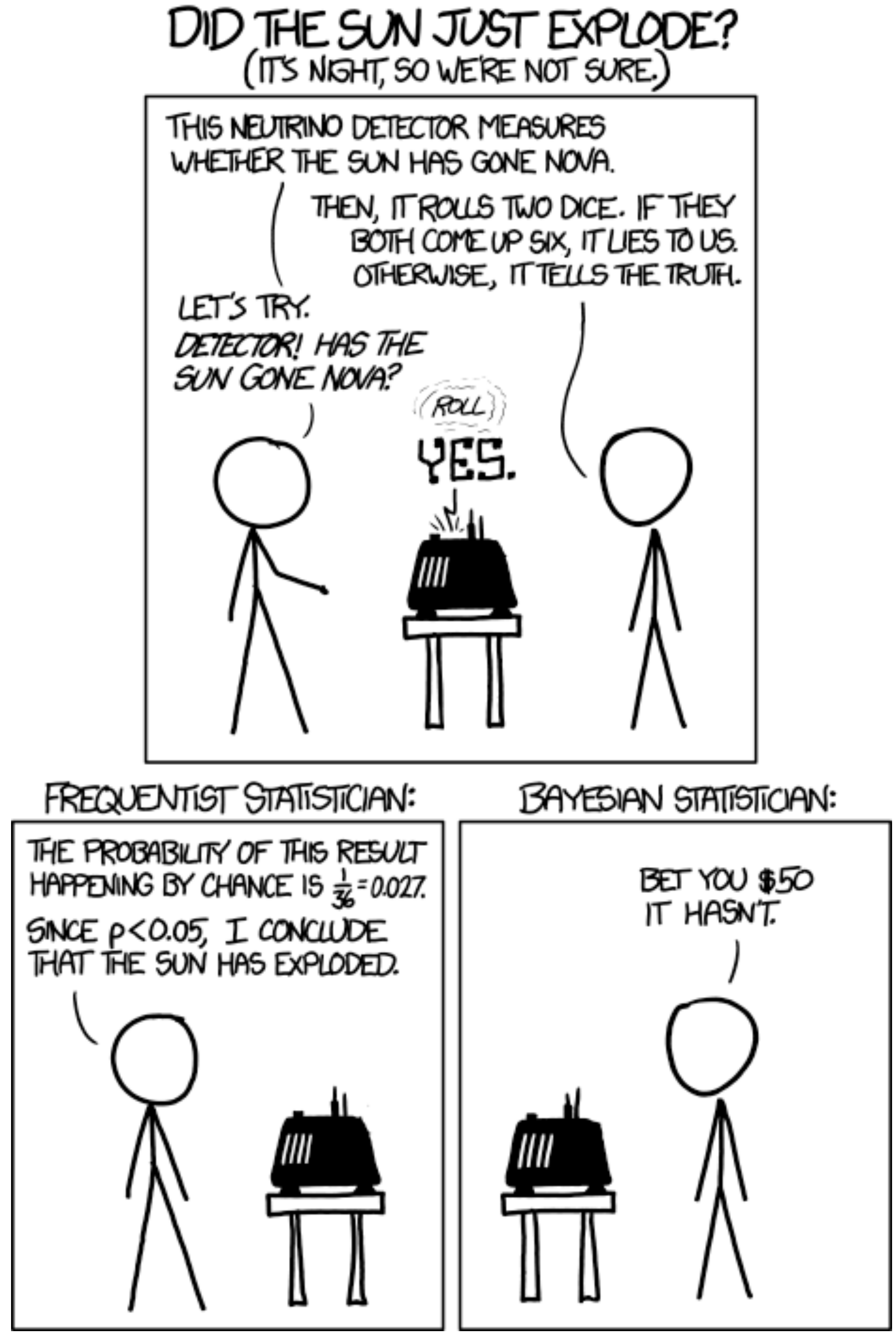
$\rightsquigarrow \quad \pi(\tau) \propto \frac{1}{\tau}$ (prior distribution)

# Bayesian versus Frequentism

[based on L. Lyons]

| | Bayesian | Frequentist |
|---|---|---|
| Meaning of probability | degree of belief | frequentist definition |
| Probability of parameters | yes | anathema |
| Needs prior | yes | no |
| Unphysical / empty intervals | excluded by prior | can occur |
| Final statement | posterior probability distribution | parameter values, hypothesis test (*p*-value) |
| Systematics | Integrate over nuisance parameter | Various methods, e.g., profile likelihood, hard |
| Combination of measurements | can be hard (prior) | ok |

# Bayesian versus Frequentism




https://xkcd.com/1132/

## 1.2 Describing the Data

# Random variables and probability density functions

## Random variable:

- Variable whose possible values are numerical outcomes of a random phenomenon
- Can be discrete or continuous

## Probability density function (pdf) of a continuous variable:

probability density function

$$P(x \text{ found in } [x, x + \mathrm{d}x]) = f(x)\,\mathrm{d}x$$

## Normalization:

$$\int_{-\infty}^{\infty} f(x)\,\mathrm{d}x = 1$$

"$x$ must be somewhere"

# Histograms

## Histogram:

- representation of the frequencies of the numerical outcome of a random phenomenon

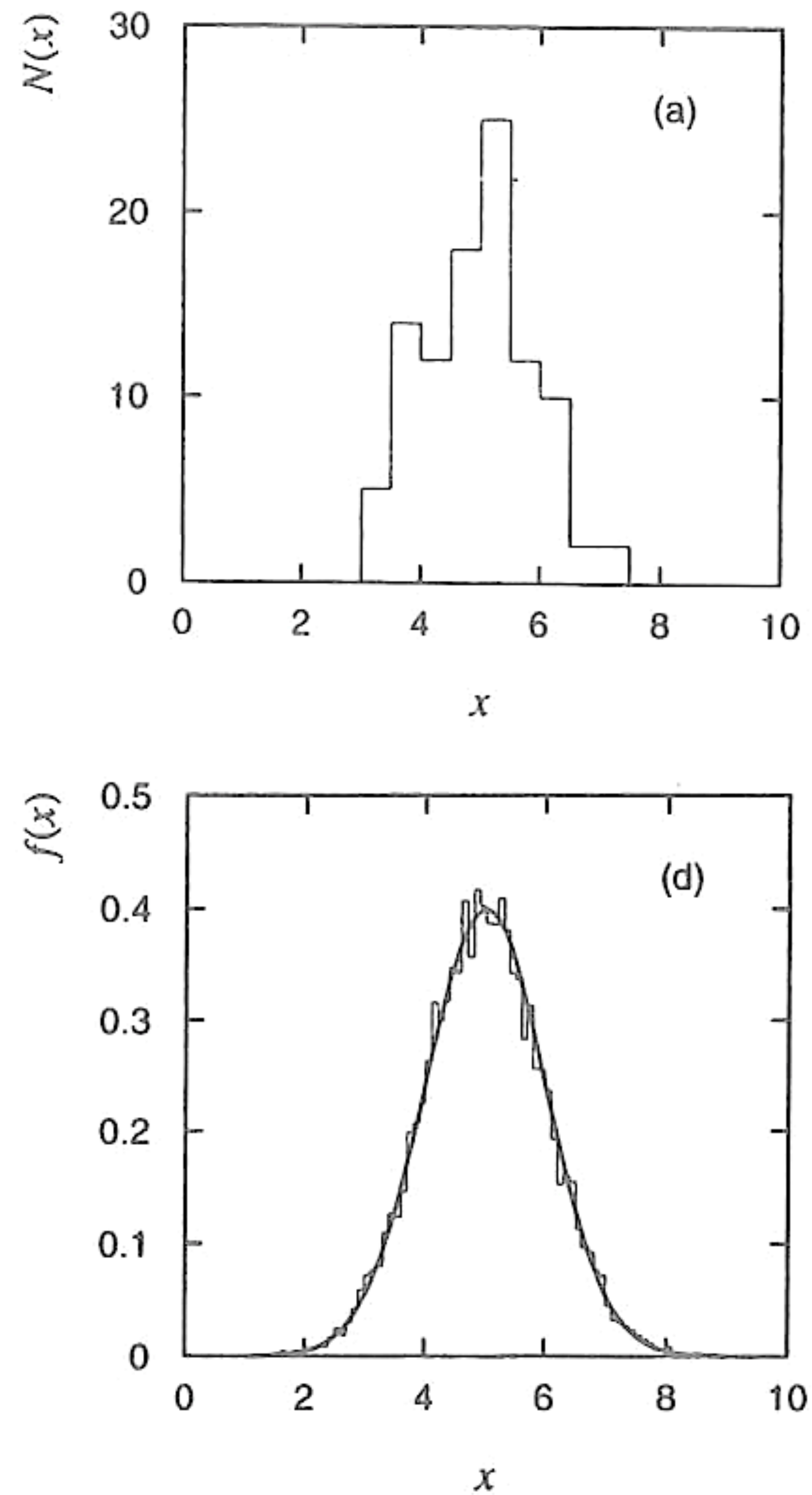
## pdf = histogram for

- infinite data sample
- zero bin width
- normalized to unit area

$$f(x) = \frac{N(x)}{n\Delta x}$$

$$n = \text{total number of entries}$$

$$\Delta x = \text{bin width}$$



# Mean, Median, and Mode

Mean of a **data sample**:

$$\bar{x} = \frac{1}{N} \sum_{i=1}^{N} x_i$$

"sample mean"

Mean of a pdf:

$$\mu \equiv \langle x \rangle \equiv \int x \, P(x) \, \mathrm{d}x$$

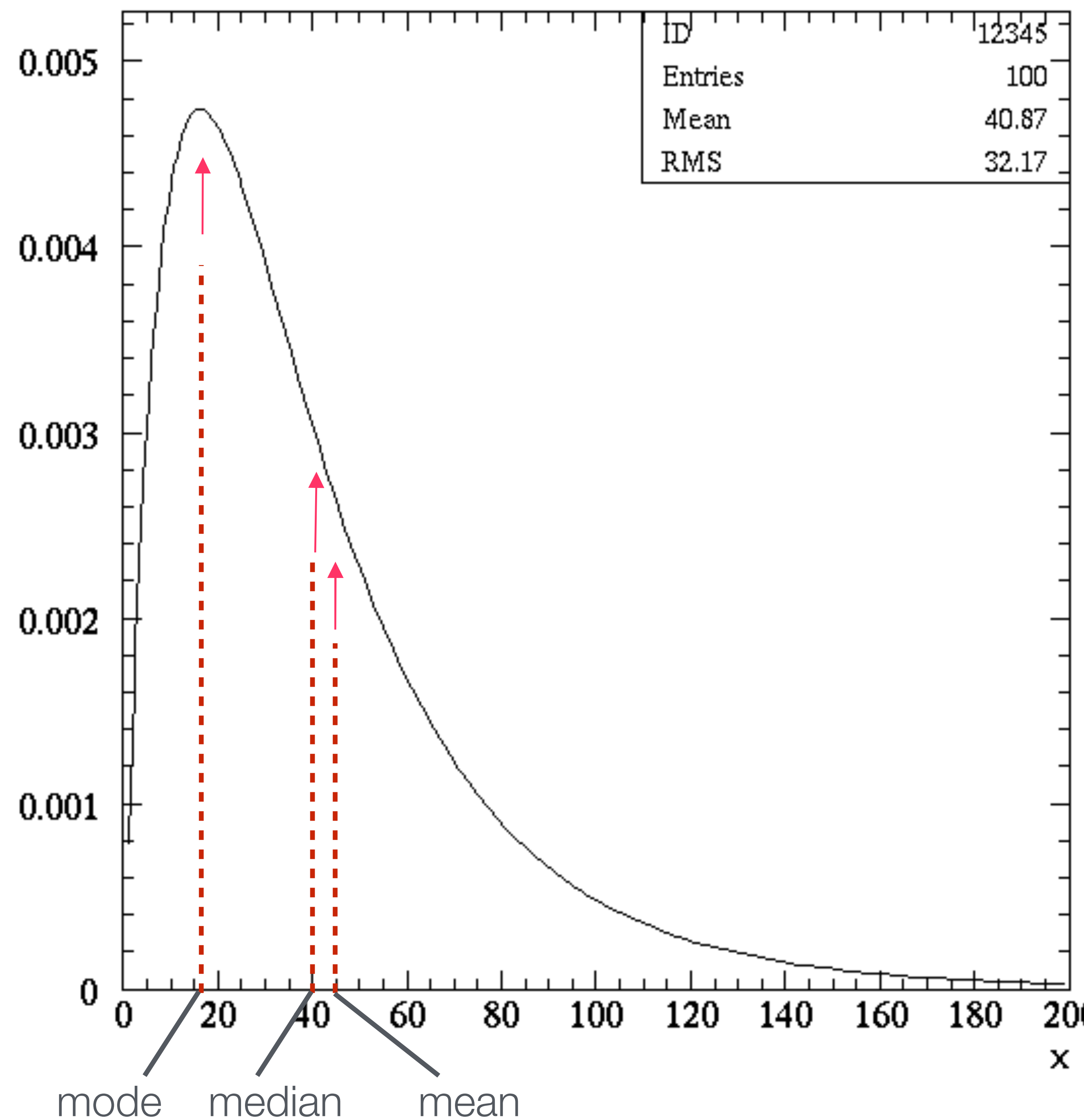
$\equiv$ expectation value $E[x]$

Median:

point with 50% probability above and 50% probability below

Mode:

the most likely value



# Variance and standard deviation

Variance of a distribution: $$V(x) = \int \mathrm{d}x\, P(x)(x-\mu)^2 = E[(x-\mu)^2]$$

expectation value

$$V(x) = \int \mathrm{d}x\, P(x) x^2 - 2\mu \underbrace{\int \mathrm{d}x\, P(x) x}_{=\mu} + \mu^2 \int \mathrm{d}x\, P(x) = \langle x^2 \rangle - \mu^2 = \langle x^2 \rangle - \langle x \rangle^2$$

Sample variance: $$V(x) = \frac{1}{N} \sum_i (x_i - \bar{x})^2 = \overline{x^2} - \bar{x}^2$$

This formula underestimates the variance of underlying distribution as it uses the mean calculated from data!

Use this if you have to estimate the mean from data (***unbiased sample variance***):

$$\hat{V}(x) = \frac{1}{N-1} \sum_i (x_i - \bar{x})^2$$

Use this if you know the true mean μ:

$$V(x) = \frac{1}{N} \sum_i (x_i - \mu)^2$$

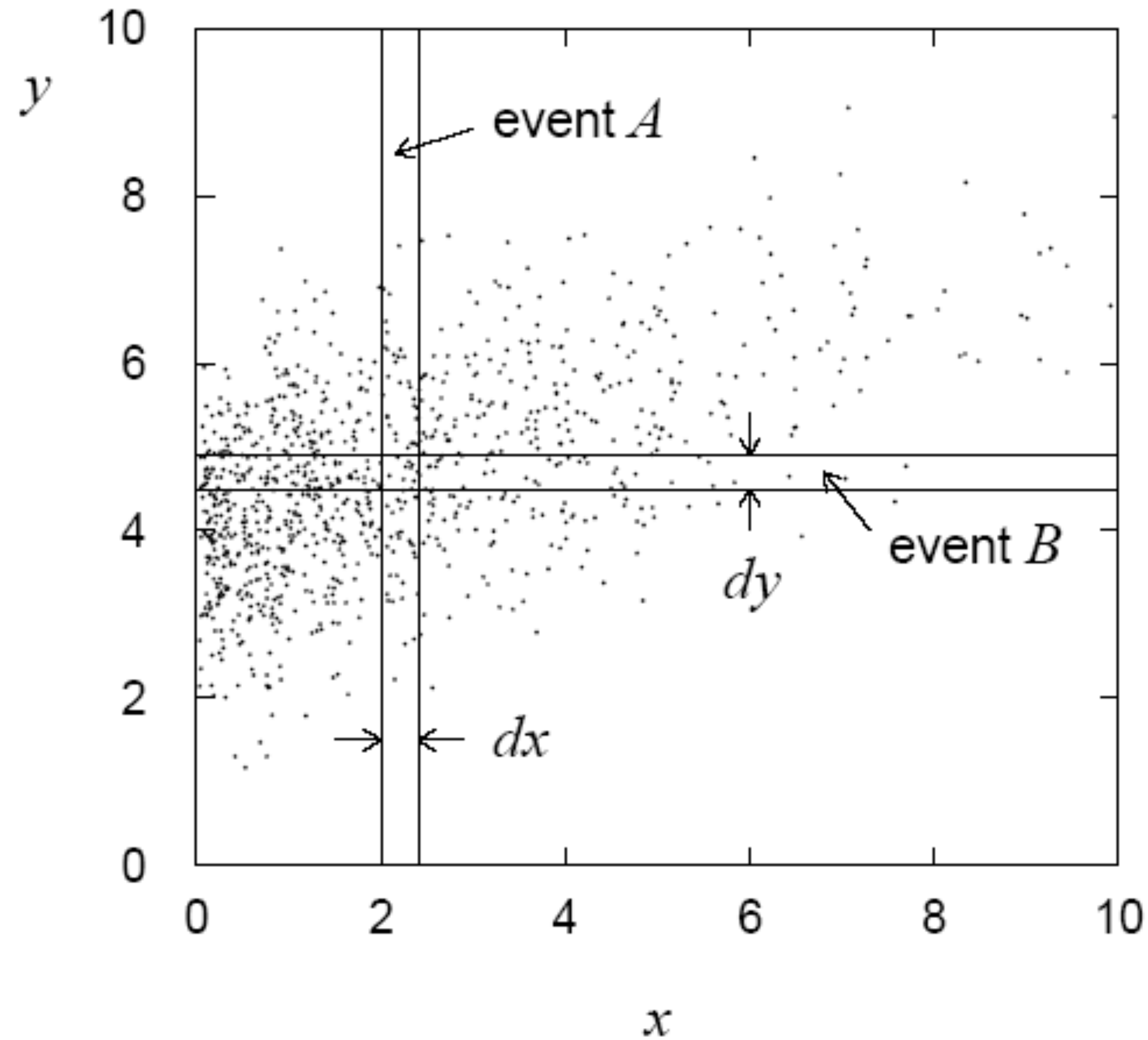
Standard deviation: $\sigma = \sqrt{V(x)}$

# Multivariate distributions

Outcome of experiment characterized by a vector ($x_1$, …, $x_n$)

$$P(A \cap B) = f(x, y)\,\mathrm{d}x\mathrm{d}y$$

joint pdf




Normalization:

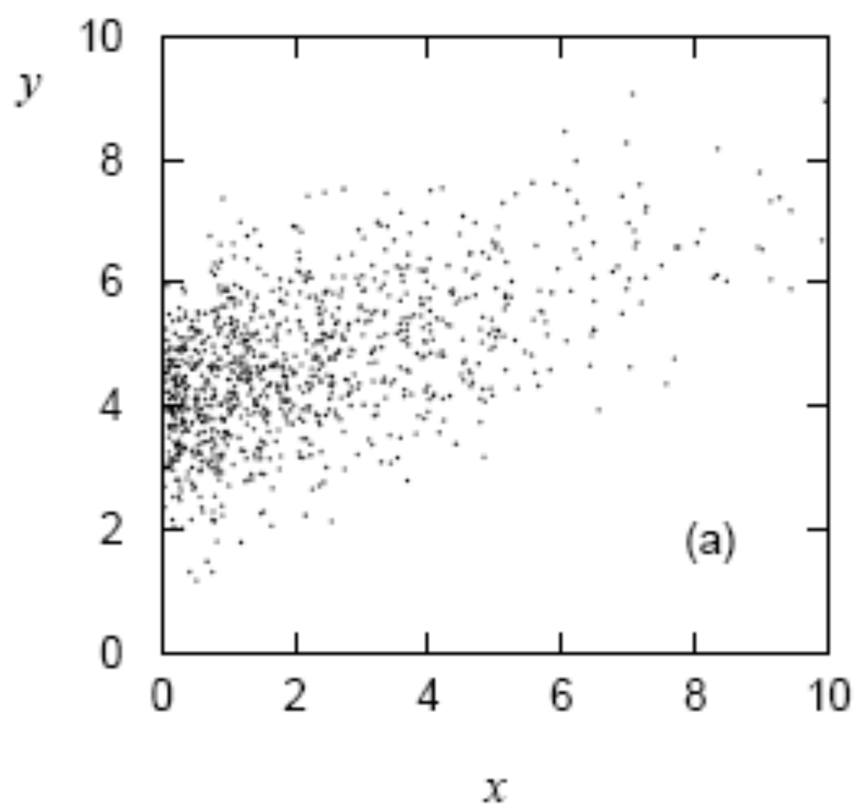
$$\int \ldots \int f(x_1, \ldots, x_n)\,\mathrm{d}x_1 \ldots \mathrm{d}x_n = 1$$

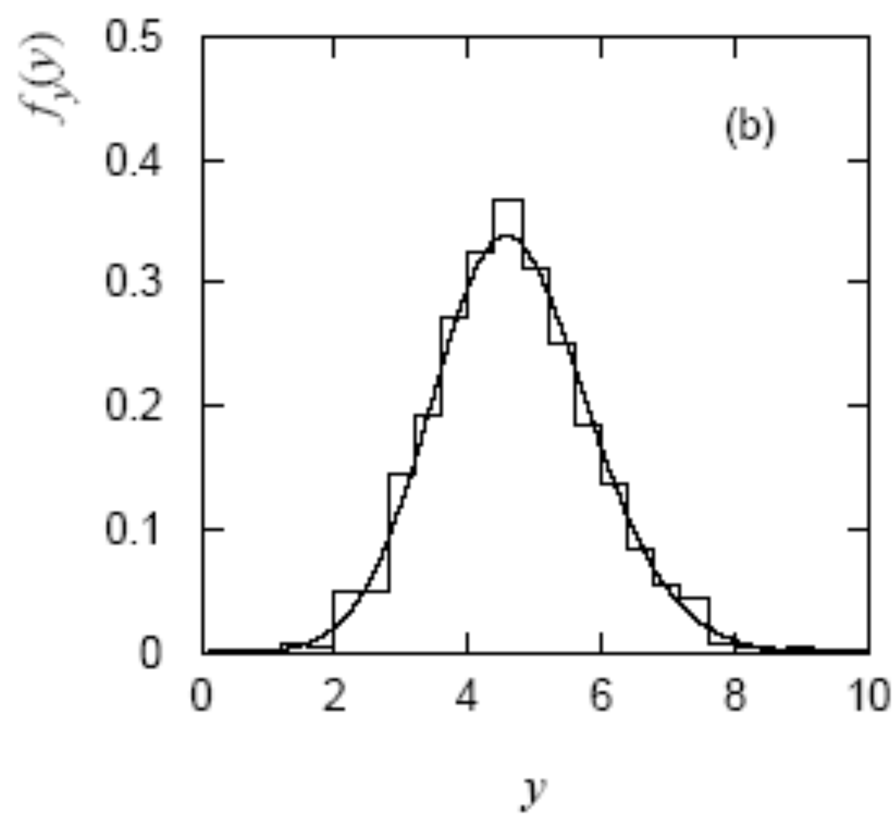
Sometimes we want only the pdf of one component:

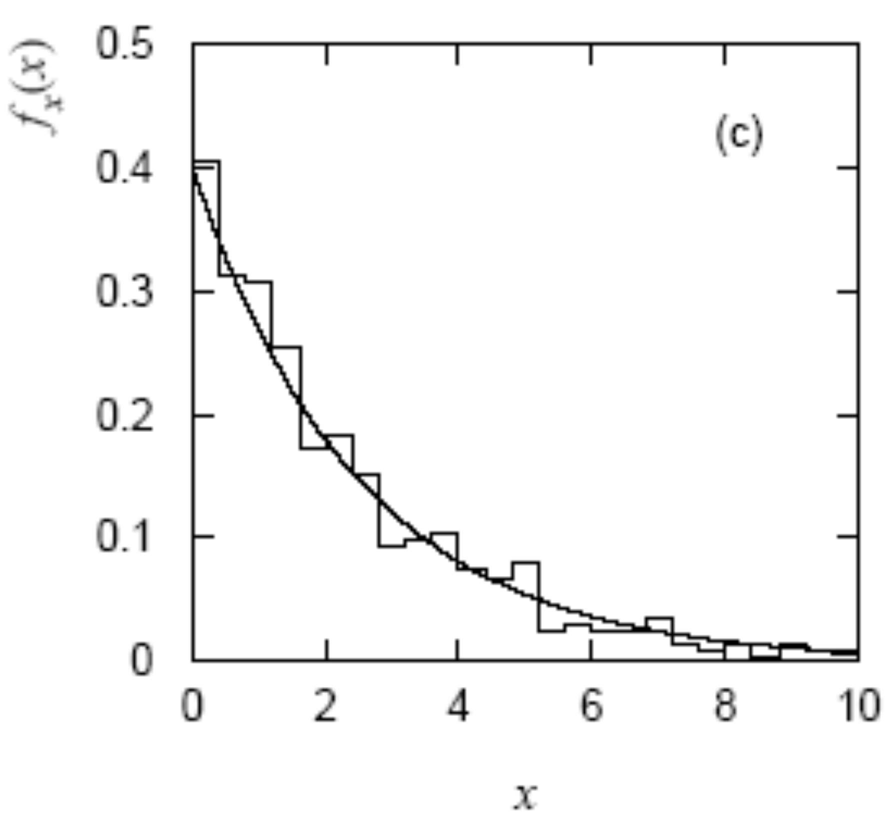
$$f_x(x) = \int f(x, y)\,\mathrm{d}y$$

"marginal pdf"
= projection of joint pdf onto individual axes

# Marginal pdf = projections




*x* and *y* independent if

$$f(x, y) = f_x(x) \cdot f_y(y)$$

# Covariance and correlation

- Covariance ($\mu_x := \langle x \rangle, \mu_y := \langle y \rangle$):

$$\mathrm{cov}[x, y] = E[(x - \mu_x)(y - \mu_y)]$$

Correlation coefficient (dimensionless):

$$\rho_{xy} = \frac{\mathrm{cov[x, y]}}{\sigma_x \sigma_y}$$

$$f_x(x) = \int dy\, f(x, y)$$

$$f_y(y) = \int dx\, f(x, y)$$

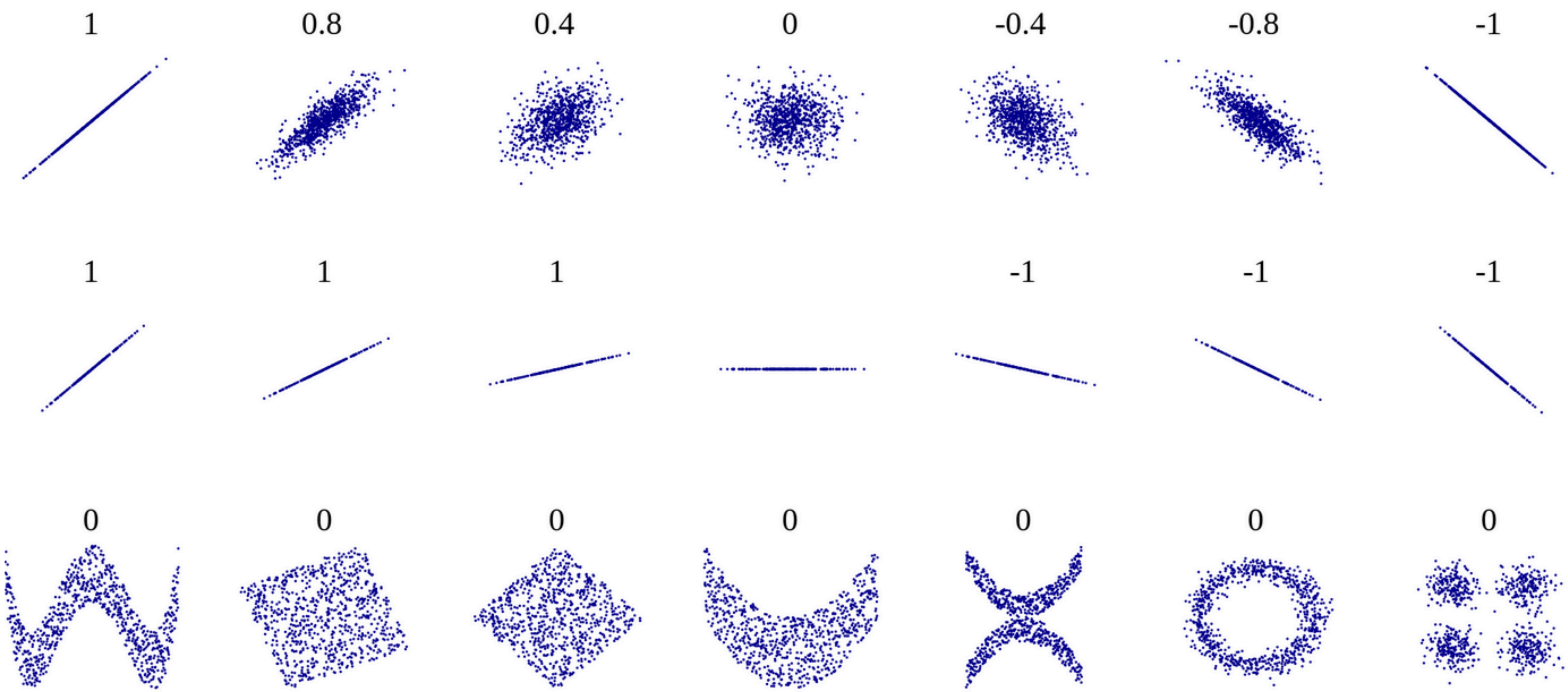
$x$, $y$ independent, i.e., $f(x, y) = f_x(x) \cdot f_y(y)$:

$$E[(x - \mu_x)(y - \mu_y)] = \int \mathrm{d}x \int \mathrm{d}y\, (x - \mu_x)(y - \mu_y) f(x, y)$$

$$= \int (x - \mu_x) f_x(x)\, \mathrm{d}x \int (y - \mu_y) f_y(y)\, \mathrm{d}y = 0$$

$\rightarrow \quad \mathrm{cov}[x, y] = 0$ (N.B. converse not always true)

# Correlation coefficient




$$\mathrm{cov}[x, y] = E[(x - \mu_x)(y - \mu_y)]$$

# Never trust summary statistics alone; always visualize your data

https://www.autodeskresearch.com/publications/samestats


same summary statistics for all data sets, in particular corr. = 0

# Linear combinations of random variables

Consider two random variables with known covariance cov($x$, $y$):

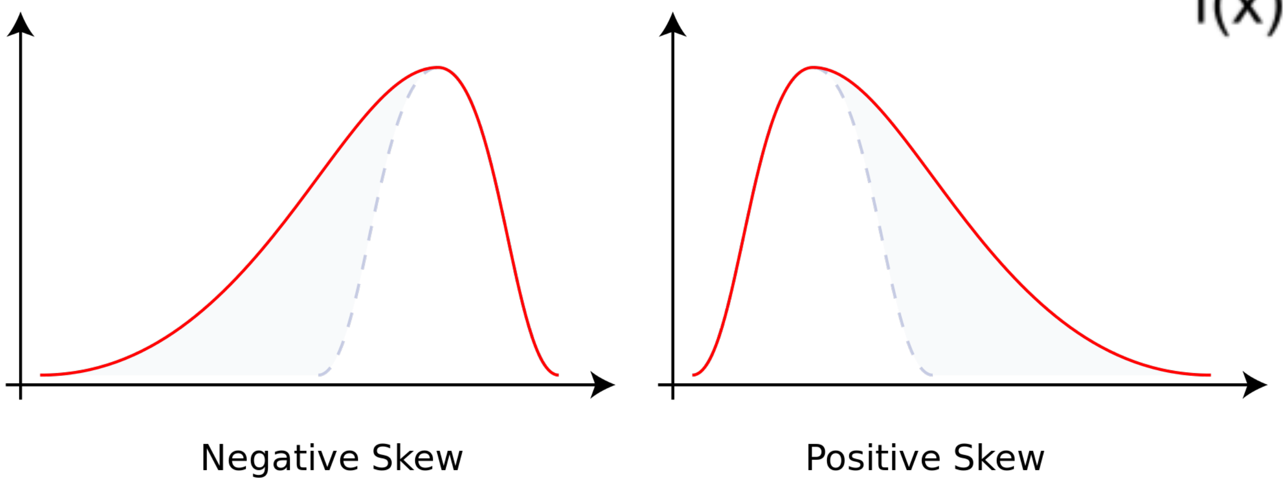
$$\begin{aligned} \langle x+y\rangle &= \langle x\rangle + \langle y\rangle \\ \langle ax\rangle &= a\langle x\rangle \\ V[ax] &= a^2 V[x] \\ \mathrm{cov}(x,x) &= V[x] \\ V[x+y] &= V[x] + V[y] + 2\mathrm{cov}(x,y) \end{aligned}$$

Example of more detailed calculation:

$$\begin{aligned} V[x+y] &= E[(x+y-\mu_x-\mu_y)^2] = E[(x-\mu_x+y-\mu_y)^2] \\ &= E[(x-\mu_x)^2 + (y-\mu_y)^2 + 2(x-\mu_x)(y-\mu_y)] \\ &= E[(x-\mu_x)^2] + E[(y-\mu_y)^2] + 2E[(x-\mu_x)(y-\mu_y)] \\ &= V[x] + V[y] + 2\mathrm{cov}(x,y) \end{aligned}$$

# Higher moments

**Skewness:** $$\gamma_1 = \left\langle \left( \frac{x - \langle x \rangle}{\sigma} \right)^3 \right\rangle$$



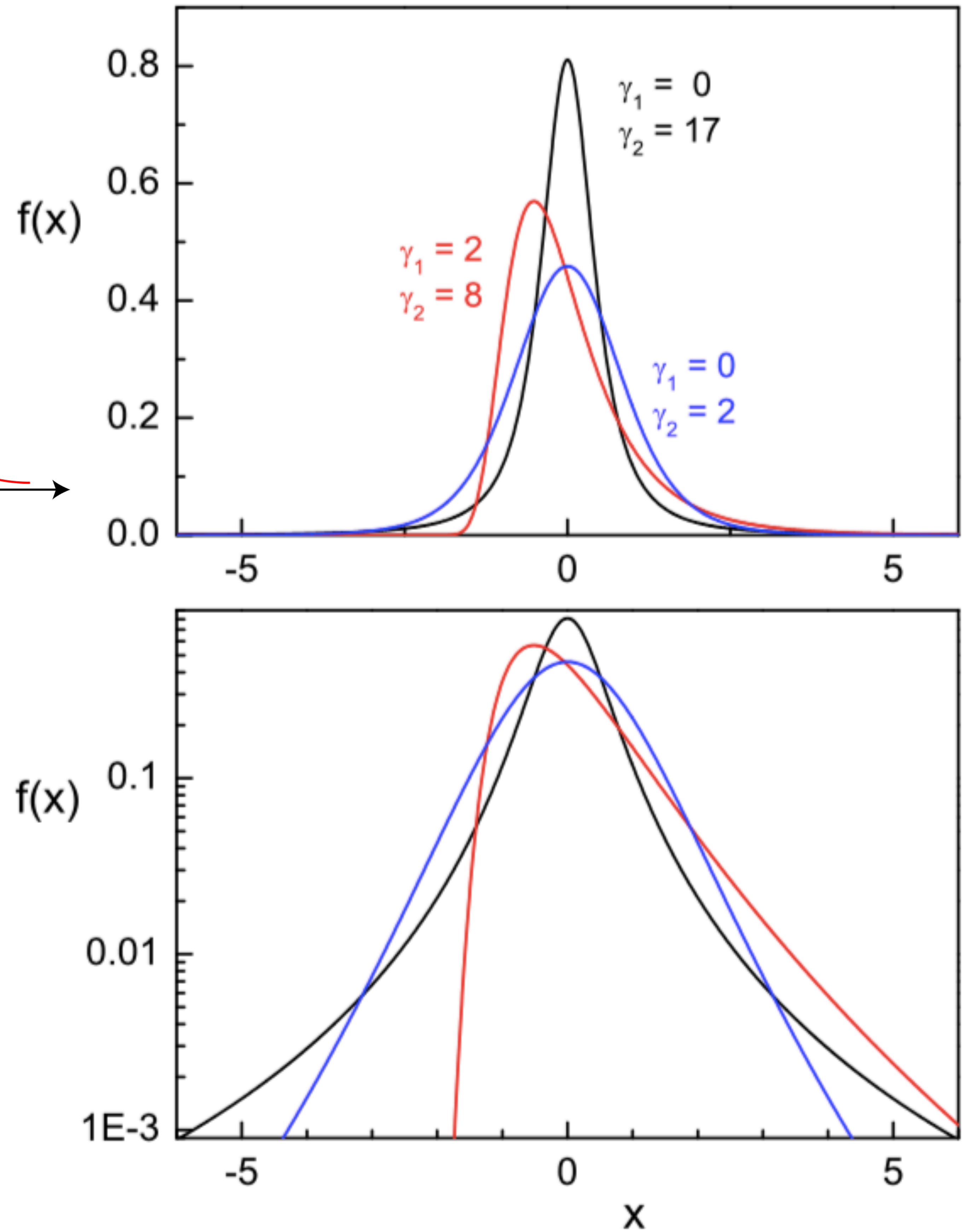

https://en.wikipedia.org/wiki/Skewness

Symmetric distribution have skewness equal to zero

**Kurtosis:** $$\beta_2 = \left\langle \left( \frac{x - \langle x \rangle}{\sigma} \right)^4 \right\rangle$$

$$\gamma_2 = \beta_2 - 3$$

defined such that $\gamma_2 = 0$ for the normal distribution



# Correlation ≠ Causation

## Examples of illogically inferring causation from correlation

https://en.wikipedia.org/wiki/Correlation_does_not_imply_causation

## Example 1 ("reverse causality"):

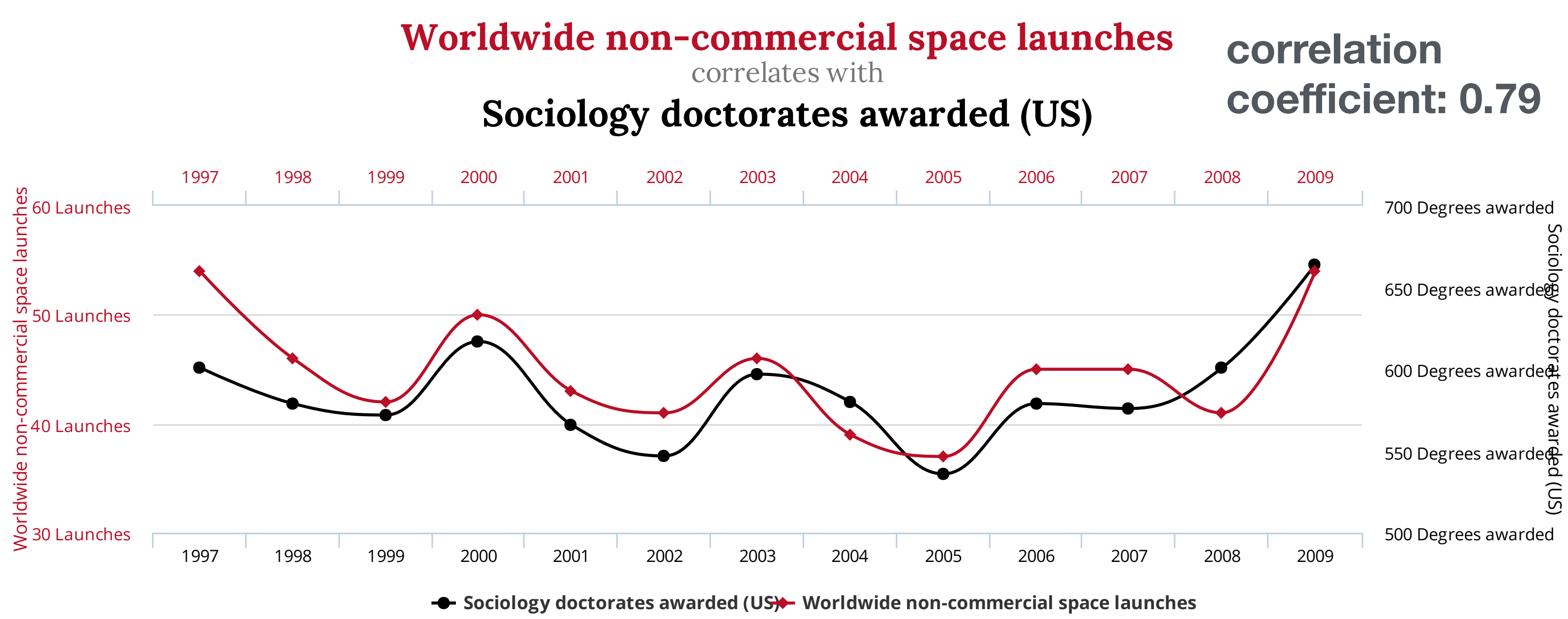
- The faster windmills are observed to rotate, the more wind is observed to be.
- Therefore wind is caused by the rotation of windmills.

## Example 2 ("third factor C causes both A and B"):

- Sleeping with one's shoes on is strongly correlated with waking up with a headache.
- Therefore, sleeping with one's shoes on causes headache.

## Example 3 (“relationship is coincidental"):

- many examples on tylervigen.com ("spurious correlations”)




correlation coefficient: 0.79

# What makes nobel prize winners?

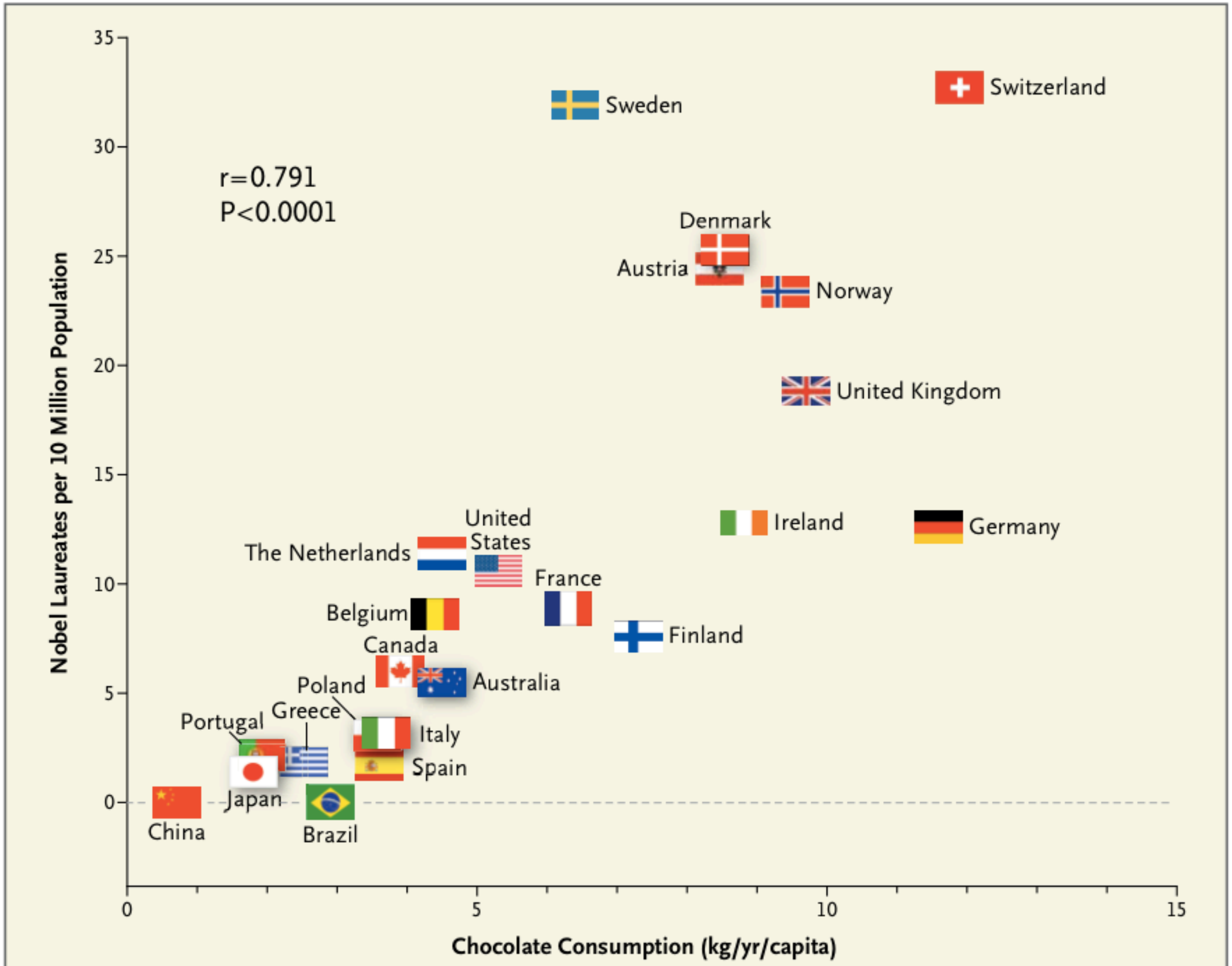



**Figure 1. Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.**

Correlation coefficient: 0.791

Improved cognitive function associated with a regular intake of flavonoids???

Probably not …

F. Messerli, 2012, New England Journal of Medicine, 2012

# Correlation ≠ Causation