Statistics for physics

Diego Tonelli — INFN Trieste
diego.tonelli@cern.ch

April 26, 2022

Future Flavours school 2022

```
diegos-mbp-2:~ diego$ whoami
```

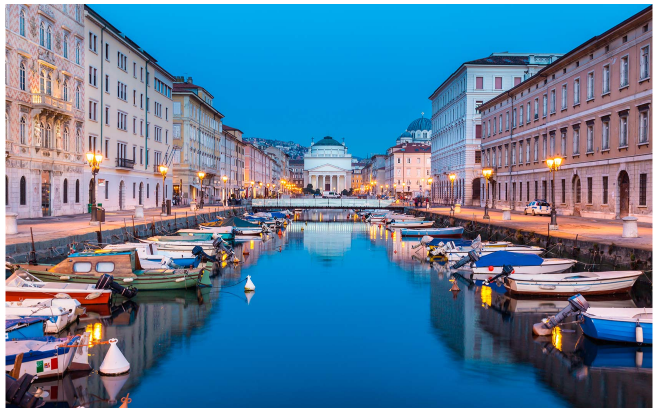
Experimental particle physicist. Interested in indirect BSM searches for non-using weak interactions of quarks (“flavor physics”).

- Born, raised, and educated in Pisa, Italy till PhD on *B* physics in the CDF experiment at Fermilab

- 2007-2011 Fermilab postdoc on CDF physics analysis (charmless *B*, *Bs* mixing, CPV in charm)

- 2012-2016: CERN staff scientist on LHCb (track-trigger, *D* mixing, *Bs* lifetimes)

- 2016—: INFN Trieste scientist: Charmless *B* in Belle II.



I am not a professional statistician (hopefully a half-decent practicioner…)

Write me anytime during or after the school for any clarification:
diego.tonelli@cern.ch

Statistics

The science of learning from data by identifying the properties of populations of natural phenomena and quantify our corresponding knowledge and uncertainty.

Statistics allows to design better experiments and make the most of our observations. It offers a structure to frame our results, interpretate them to derive implications, and a language to communicate them. Typical HEP tasks

- Simulate a physics process — modeling
- Measure the value of a physics parameter — point estimation
- Find its uncertainty — interval estimation
- Compare one hypothesis agains another (search for anomalies) — hypot. testing
- Comparing one hypothesis against all others — Goodness of fit

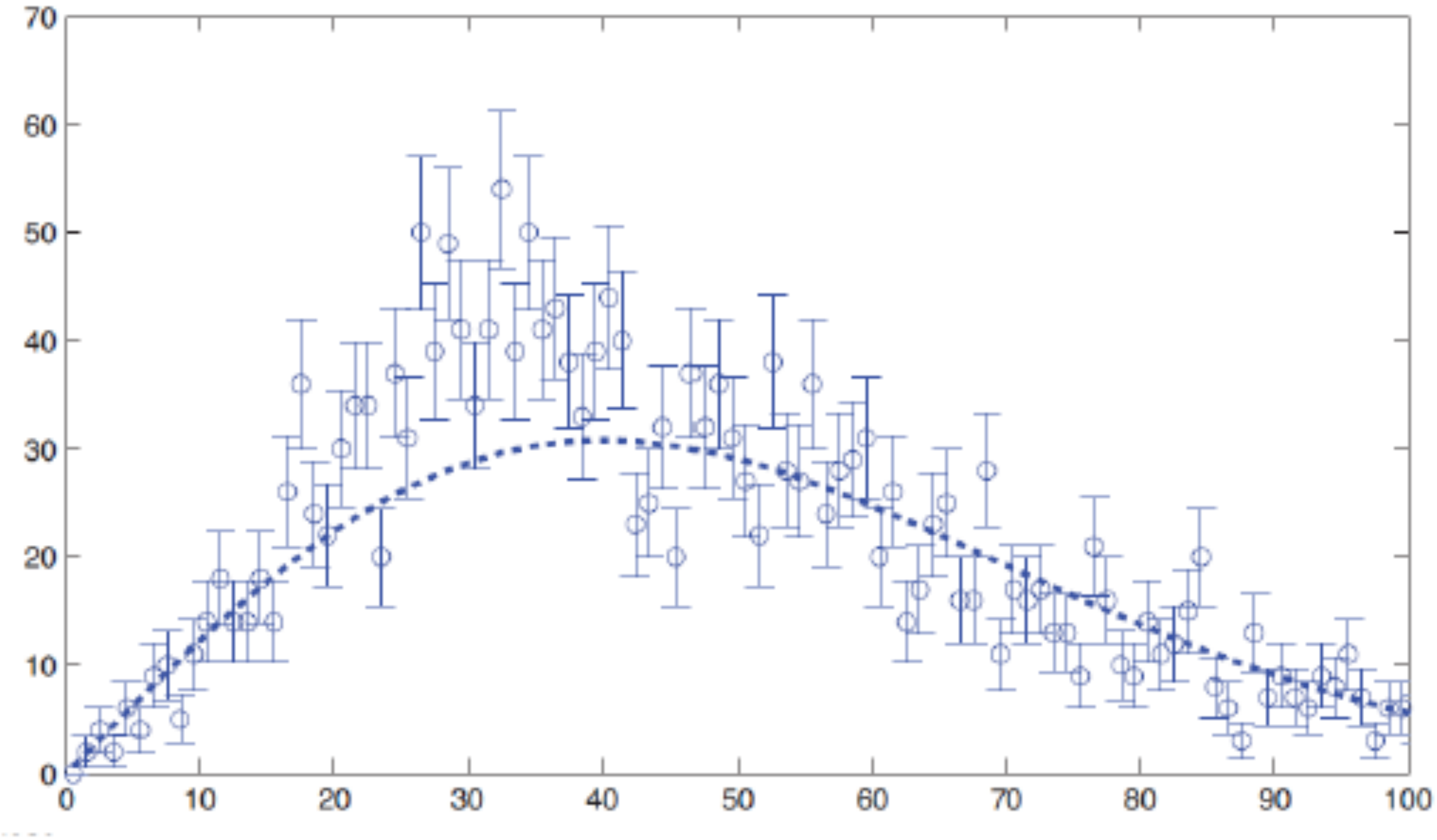
Why do I need statistics at all to do physics?

An enthomologist has little doubt when she/he stumbles upon a previously unobserved insect.

No need for histograms, or sophisticated data analyses. One “signal event” suffices when background is known to be zero certainly.

Why do I need statistics at all to do physics?

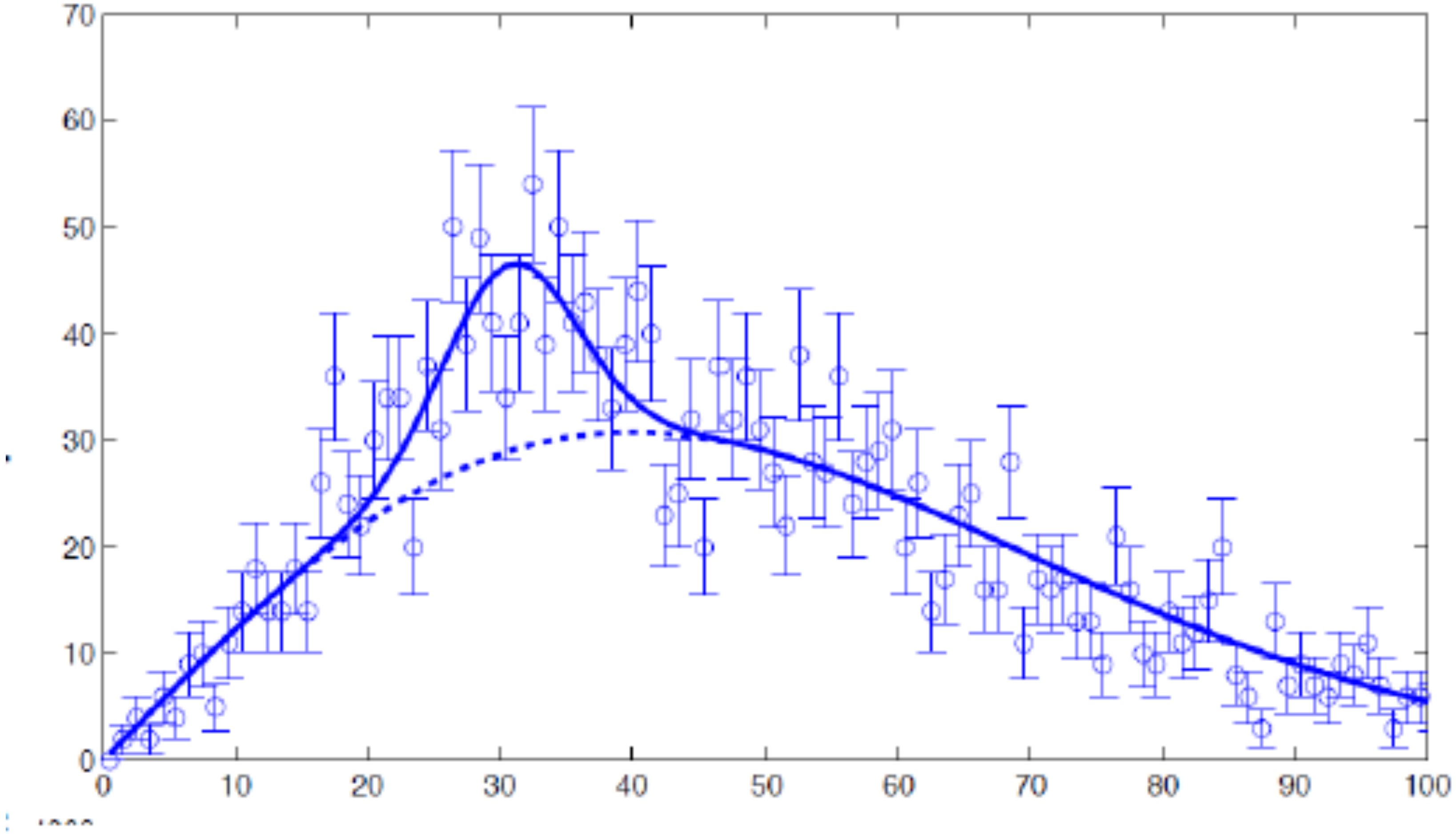
Background only?



Are a limited number of data distributions compatible with expectations from known processes only ("background")?

Or they indicate contributions of new phenomena as well ("signal")?

Or is there a flying donkey too?



The challenge: how compatible data are with expectations from background? Is there a signal there? If so, what would be the statistical significance? And what is the most powerful way of telling the background apart from the signal+background ?

Understanding nature from blurred observations

Top-down vs bottom-up understanding

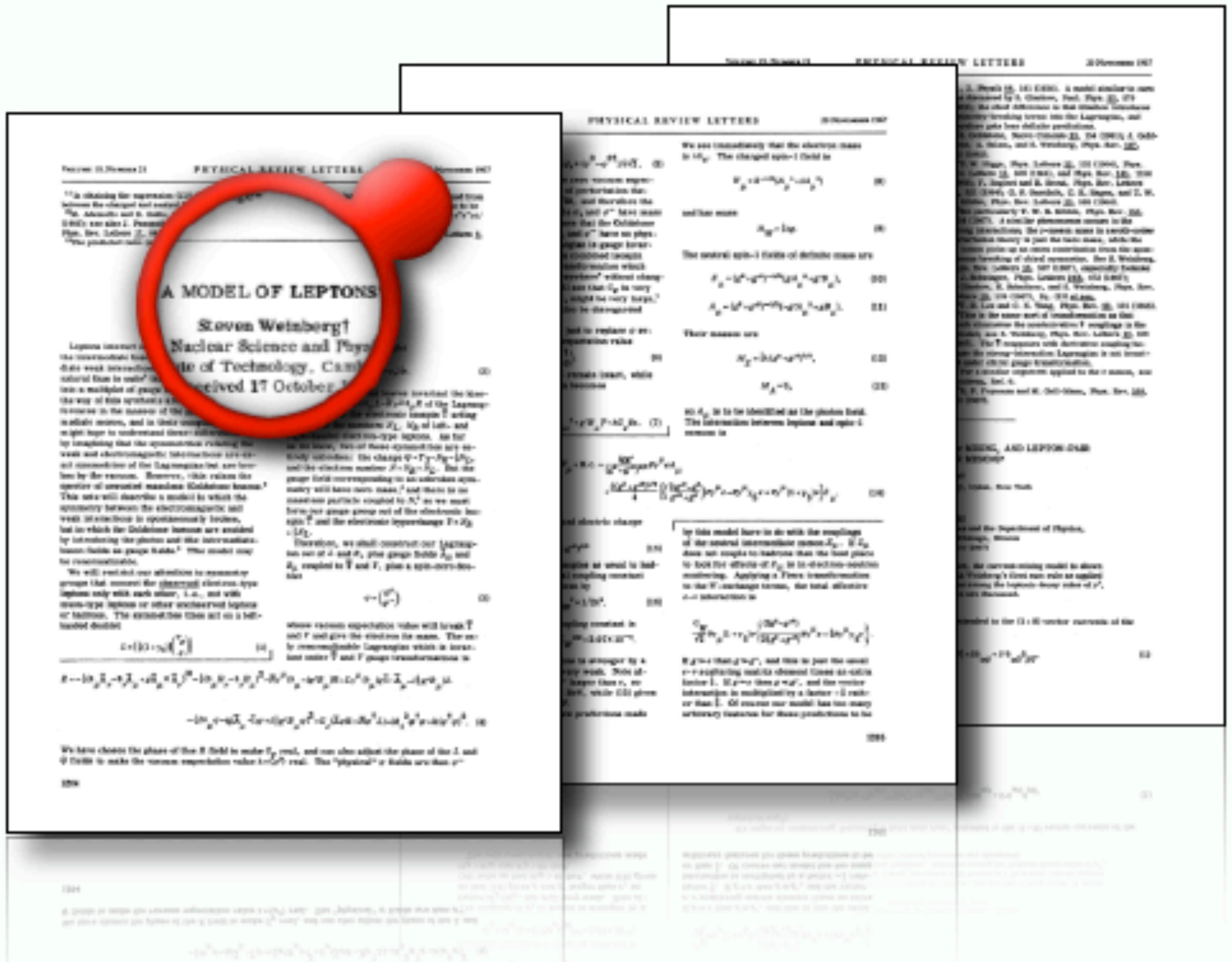
Similar to low-level perception processes, HEP advances through the interplay of top-down (theory-guided) and bottom-up (data-driven) processing.

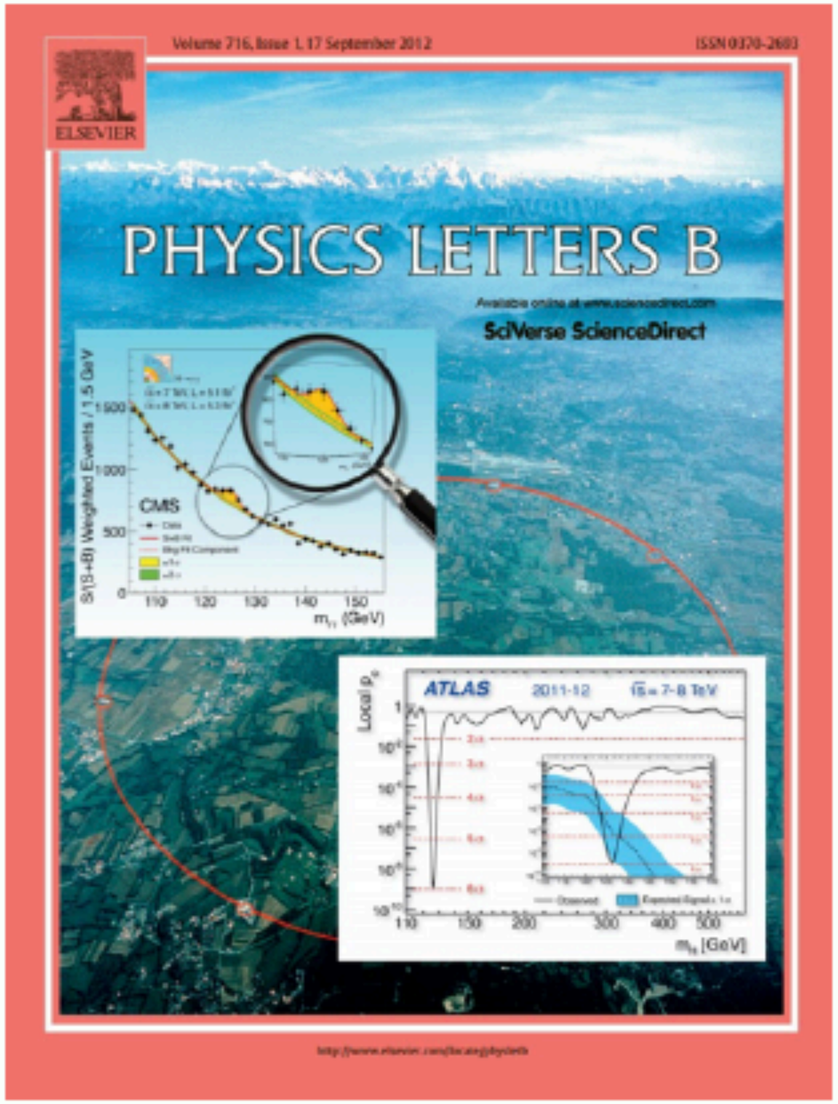
The need for detail (quality and quantity of data) is driven by the distinctiveness of the phenomena and our level of familiarity with it.

When a roadmap suggest “what to expect”, little data go a long way (top-down dominates).

Since the 80’s, the standard model has served us well as a road map to guide HEP’s exploration, because it offered a few robust no-lose theorems that led to the discovery of the W and Z bosons, the top quark, and the Higgs boson.

1967-2012





The standard model is now complete. It is robust at the energies explored so far and technically up to 10^{10} GeV.

Are we done?

No.

2012 — (hopefully not too late): a novel data-driven era.

Why 3 quark and lepton families? Why their mass hierarchies? What’s the origin of CP violation? Why does the strong interaction does not violate it? Are neutrino Majorana or Dirac? What’s dark matter? And dark energy? [your favorite open question here]

Bad news is that top-down comfort is over.

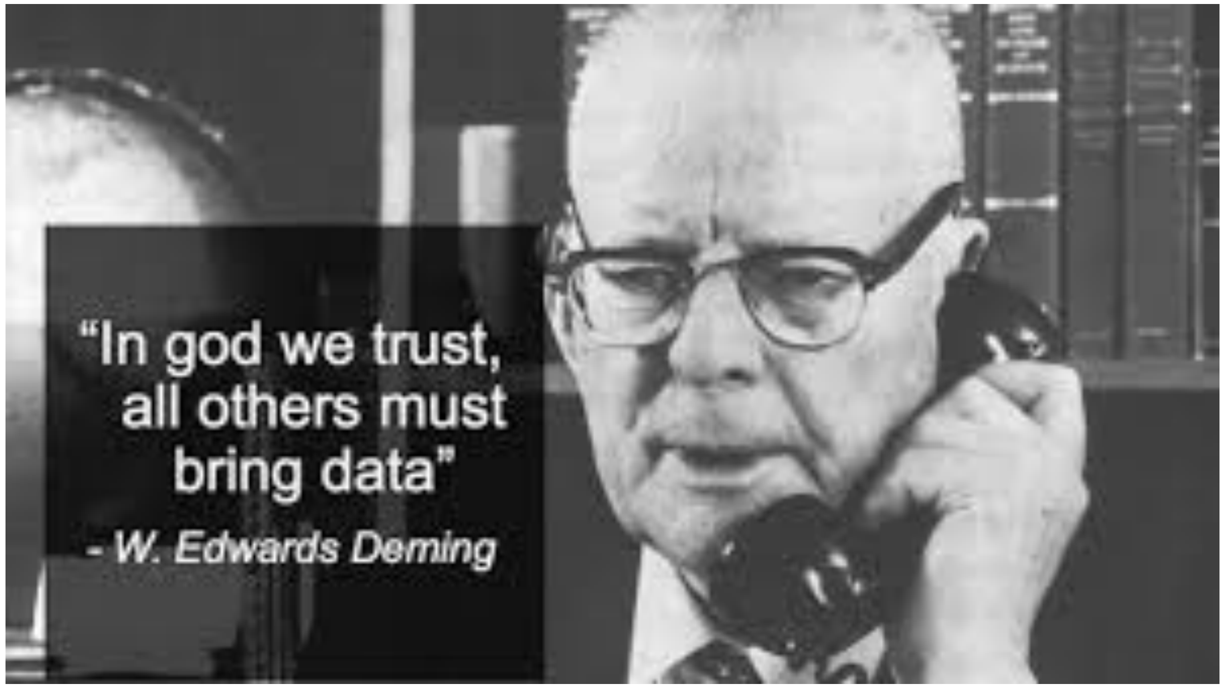
It is likely that next progress on some of the most compelling questions will come through the bottom-up, brute-force approach: look and try to make a sense of lots of data from many different experiments.

Looks like a particularly fitting time to recap on methods of extracting information from the data.

Outline

- Quick recap of basics
- probability and inference
- point-estimation, interval estimation
- Role of the model and nuisance parameters,
- hypothesis testing and significance

Had to cut quite a lot of material to fit in the allotted time and will skip many essential topics. Apologies if I skipped your favorite topic or the one that confuses you most — write me offline and I'll be glad to discuss it with you.



Quick recap of the basics

Fundamental notions

Random event: an event that has >1 possible outcome. The outcome isn't predicted deterministically, but a probability* for each outcome is known.

Random events are associated to variates (also called "(random) variables", "observables") x, which take different values x_0, x_1, etc. corresponding to different possible outcomes.

Each x value has its probability* p(x). The outcomes generate a probability distribution of x.

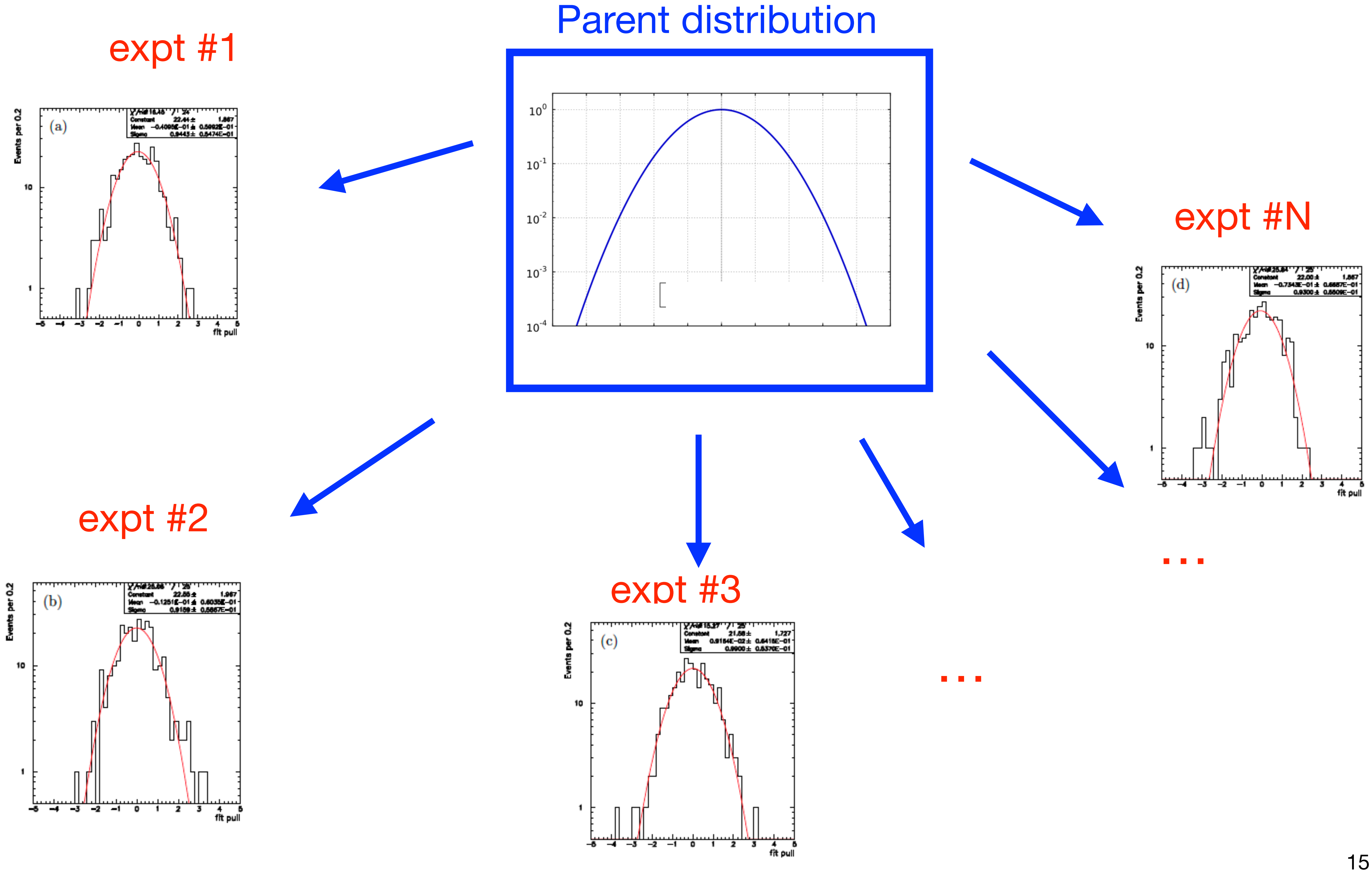
A collection of random events forms a population: the hypothetical infinite set of outcomes from repeated independent and (nearly) identical experiments.

Observed distributions are interpreted as finite-size random samplings from the corresponding population's parent distributions.

Goal: quantify the collective properties of the parent distributions, *not* of any individual element of the sample.

*Probability intended as limit of long term frequency, more later.

Parent distribution



Parent distribution

expt #1

expt #2

expt #3

expt #N

...

...

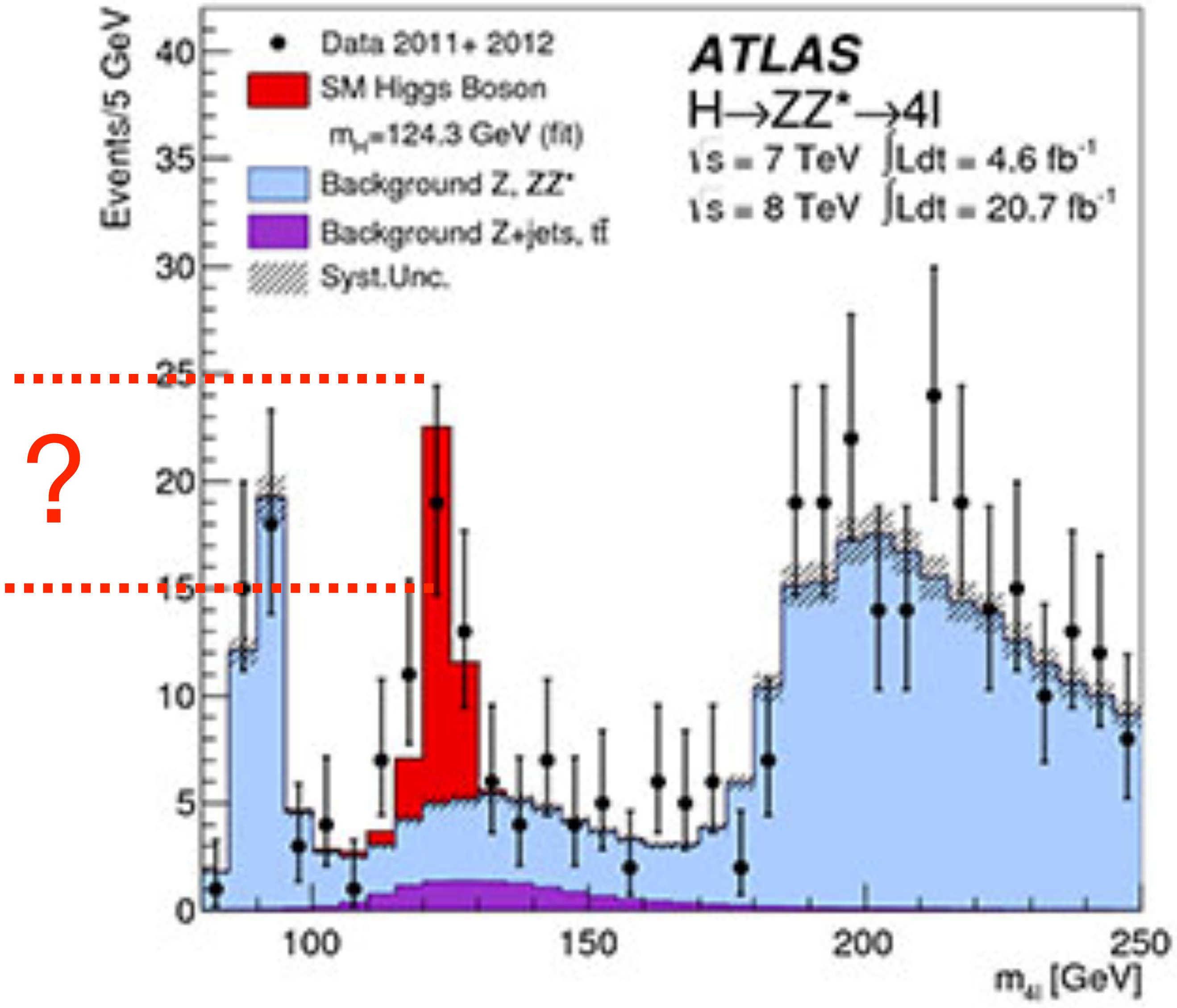
You do it everyday

Most of you regularly quote uncertainties

For instance, in a counting experiment such as an histogram, a bin with N entries has an error bar (e.g., of length √ N)

What does the bar *exactly* mean?

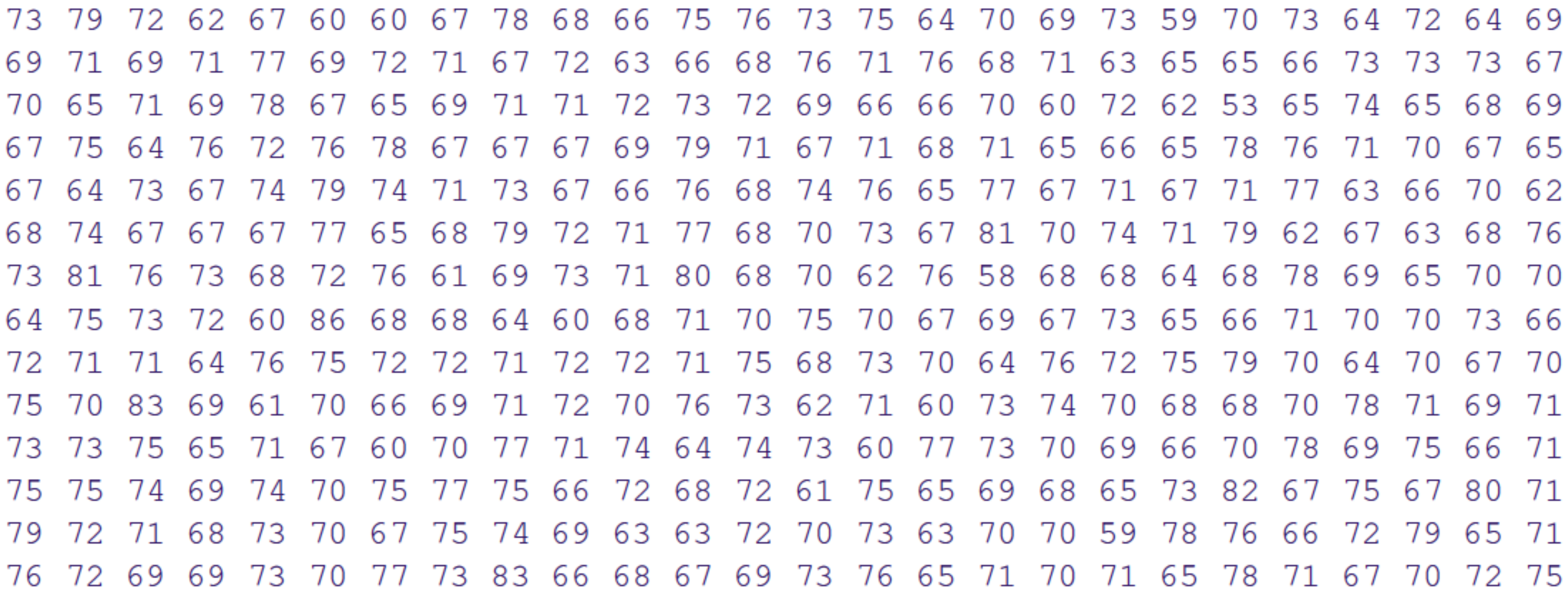
Am I really uncertain if in my sample N events are falling in that bin?



The bar represents the fluctuations in the counts of that bin one expects if the experiment was repeated. I.e, the fluctuations between samples drawn from the same *parent distribution.*

Representing data

Raw data

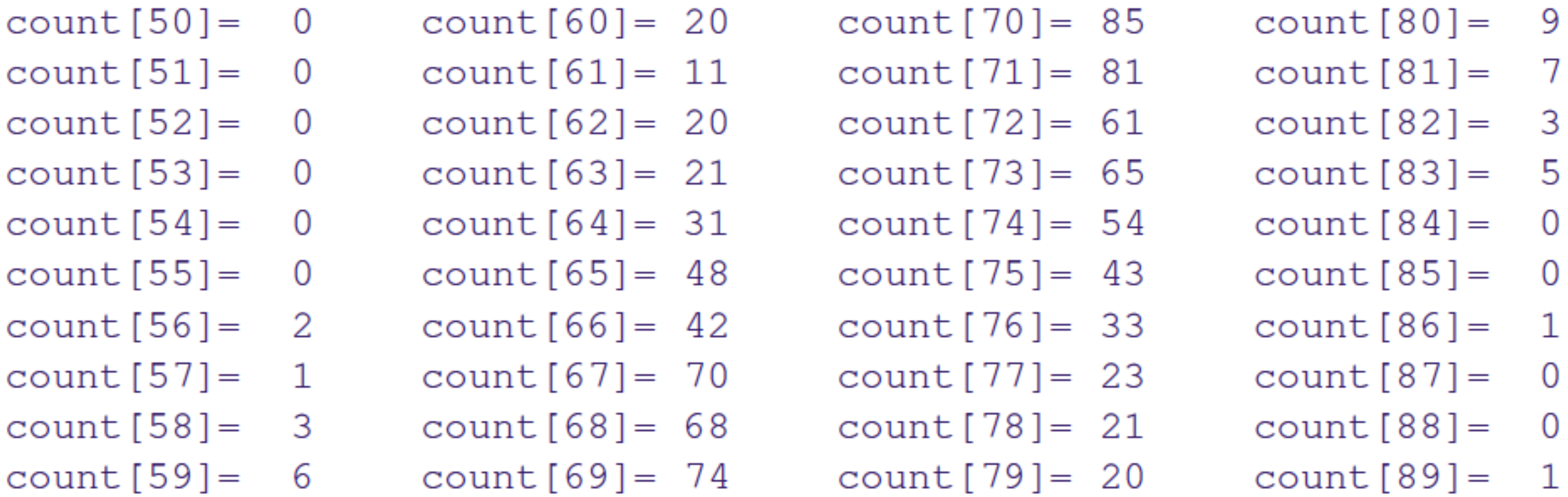


```
73 79 72 62 67 60 60 67 78 68 66 75 76 73 75 64 70 69 73 59 70 73 64 72 64 69
69 71 69 71 77 69 72 71 67 72 63 66 68 76 71 76 68 71 63 65 65 66 73 73 73 67
70 65 71 69 78 67 65 69 71 71 72 73 72 69 66 66 70 60 72 62 53 65 74 65 68 69
67 75 64 76 72 76 78 67 67 67 69 79 71 67 71 68 71 65 66 65 78 76 71 70 67 65
67 64 73 67 74 79 74 71 73 67 66 76 68 74 76 65 77 67 71 67 71 77 63 66 70 62
68 74 67 67 67 77 65 68 79 72 71 77 68 70 73 67 81 70 74 71 79 62 67 63 68 76
73 81 76 73 68 72 76 61 69 73 71 80 68 70 62 76 58 68 68 64 68 78 69 65 70 70
64 75 73 72 60 86 68 68 64 60 68 71 70 75 70 67 69 67 73 65 66 71 70 70 73 66
72 71 71 64 76 75 72 72 71 72 72 71 75 68 73 70 64 76 72 75 79 70 64 70 67 70
75 70 83 69 61 70 66 69 71 72 70 76 73 62 71 60 73 74 70 68 68 70 78 71 69 71
73 73 75 65 71 67 60 70 77 71 74 64 74 73 60 77 73 70 69 66 70 78 69 75 66 71
75 75 74 69 74 70 75 77 75 66 72 68 72 61 75 65 69 68 65 73 82 67 75 67 80 71
79 72 71 68 73 70 67 75 74 69 63 63 72 70 73 63 70 70 59 78 76 66 72 79 65 71
76 72 69 69 73 70 77 73 83 66 68 67 69 73 76 65 71 70 71 65 78 71 67 70 72 75
```

Not very informative.

Assuming that all observations are equivalent, the individual sequence does not matter and all relevant information is contained in the frequency of each outcome.

Frequencies



```
count[50]=  0     count[60]= 20     count[70]= 85     count[80]=  9
count[51]=  0     count[61]= 11     count[71]= 81     count[81]=  7
count[52]=  0     count[62]= 20     count[72]= 61     count[82]=  3
count[53]=  0     count[63]= 21     count[73]= 65     count[83]=  5
count[54]=  0     count[64]= 31     count[74]= 54     count[84]=  0
count[55]=  0     count[65]= 48     count[75]= 43     count[85]=  0
count[56]=  2     count[66]= 42     count[76]= 33     count[86]=  1
count[57]=  1     count[67]= 70     count[77]= 23     count[87]=  0
count[58]=  3     count[68]= 68     count[78]= 21     count[88]=  0
count[59]=  6     count[69]= 74     count[79]= 20     count[89]=  1
```

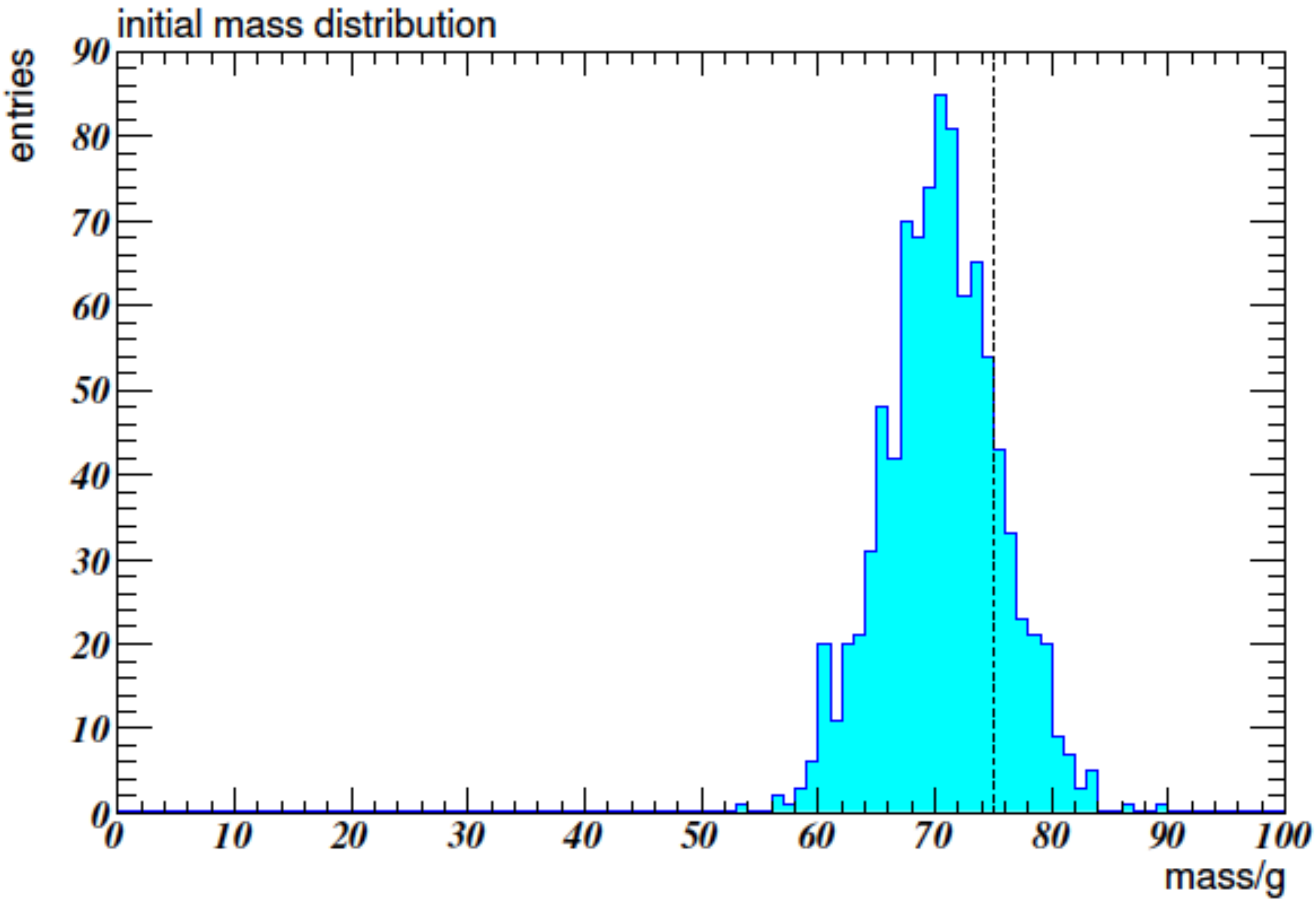
Better. Still not very intuitive.

Frequency distribution

Much better. Offers immediate visual feel of

- the “shape”,
- the “localization” and
- the “dispersion”

of data



Binning — short aside

Some feel strongly about rules for choosing the bin width.

Not that relevant as long as event counts remain O(10) or greater in most of the interesting region of the distribution (O(10) so that Poisson —> Gaussian, see later)

More important: binning is a data reduction and as such it induces a loss of information.

In continuous data, values of variables for each entry are known up to the native precision of the apparatus.

In binned data, all entries with values within the range of a bin are collectively filed in that bin. In all subsequent manipulations they are treated as if they have the same value (corresponding to the center of the bin).

That’s why attributing additional uncertainty due to changes in binning is typically wrong. Changes in the results are due to the method of data reduction and as such included in the statistical uncertainty. Adding uncertainty leads to double counting.

Sample statistics

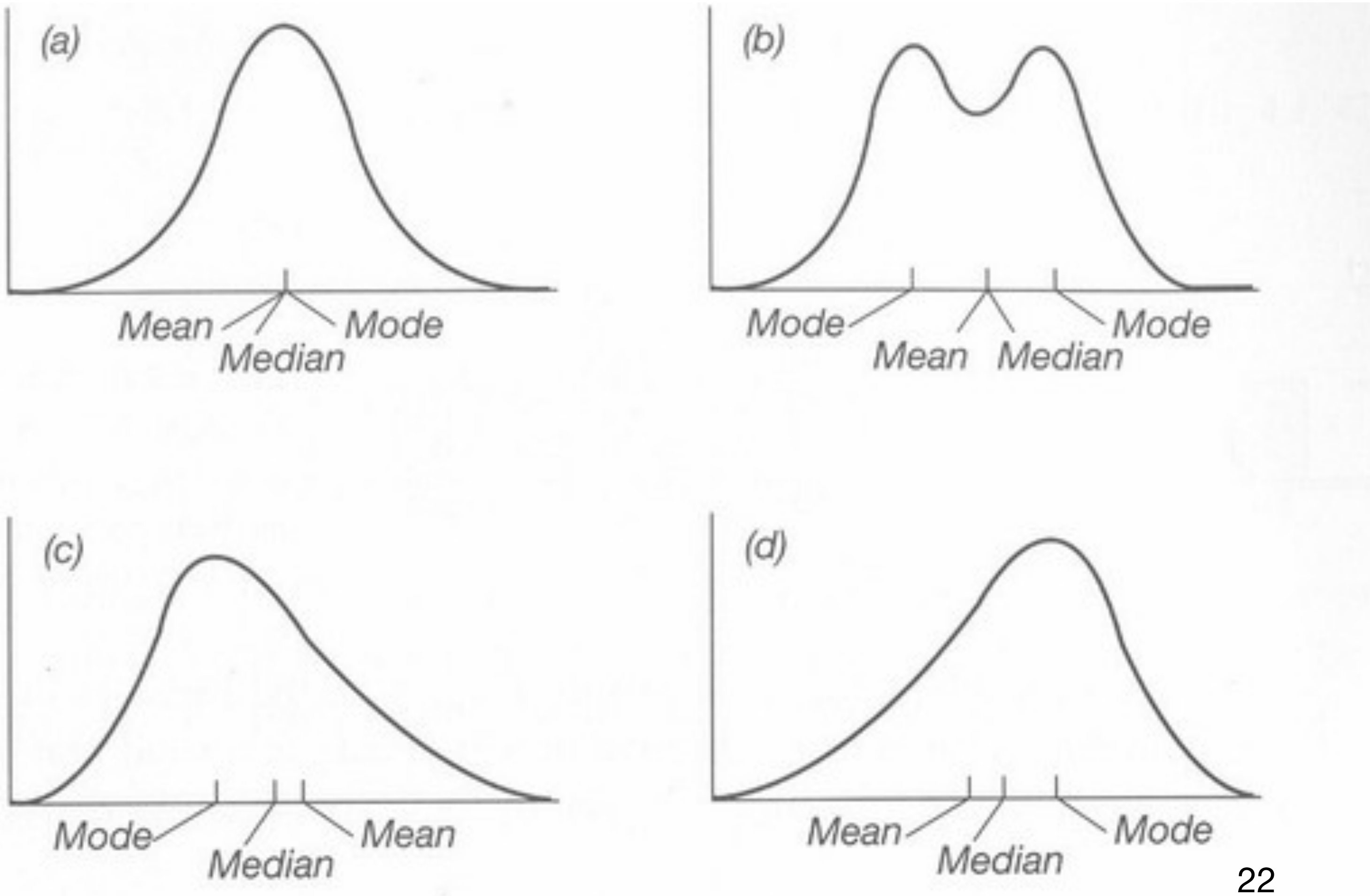
Hard to do any serious analysis by just staring at distributions.

Need to get more quantitative. A few simple quantities can be calculated from the available data only (they do not depend on parameters) and encapsulate quantitative information of “location” or “central value” of a distribution into a few numbers.

Sample mode: value of the variable for which the population is larger.

Sample median: mid-range value of the variable so that 1/2 of sample has larger and 1/2 has smaller values.

Sample mean: arithmetic average of the values of the variable across the sample



Sample mean — “where are my data”

Simple and most common quantity if one wants to summarize the distribution information into a single number.

For a sample of N events, each associated with a variable x_i and binned into an histogram with n bins, the sample mean is

$$\text{Unbinned sample mean} \quad \bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

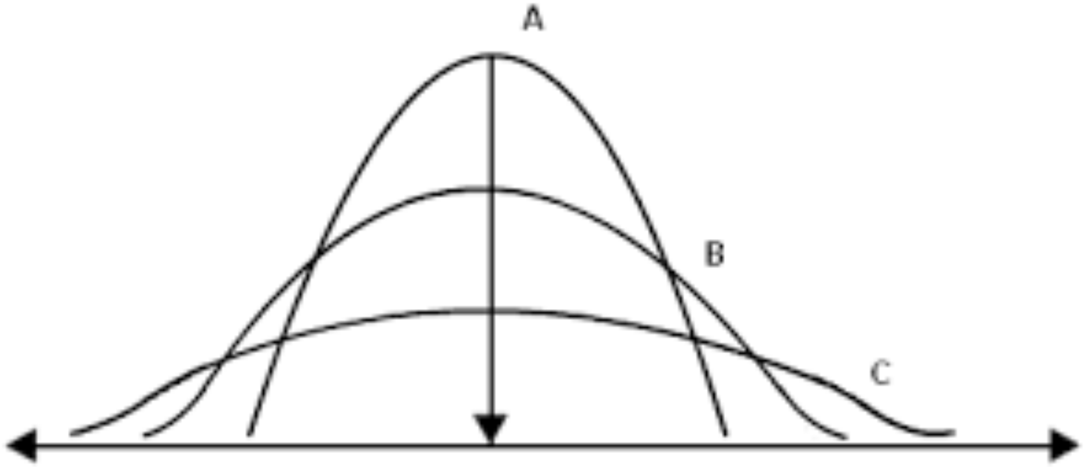
$$\text{Binned sample mean} \quad \bar{x} = \frac{1}{N}\sum_{j=1}^{n} x_j n_j$$

Linear: $\overline{\alpha x + y} = \alpha\overline{x} + \overline{y}$

Sample variance — "How spread they are"

The mean says nothing about the dispersion of data, which is another key information to grasp the features of a population.

Use the variance: average of the difference square from the mean



$$V(x) = \overline{(x - \overline{x})^2} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{x})^2$$

Easier to remember: *the mean of the squares minus the square of the mean*

$$V(x) = \overline{x_i^2} - \overline{x}^2$$

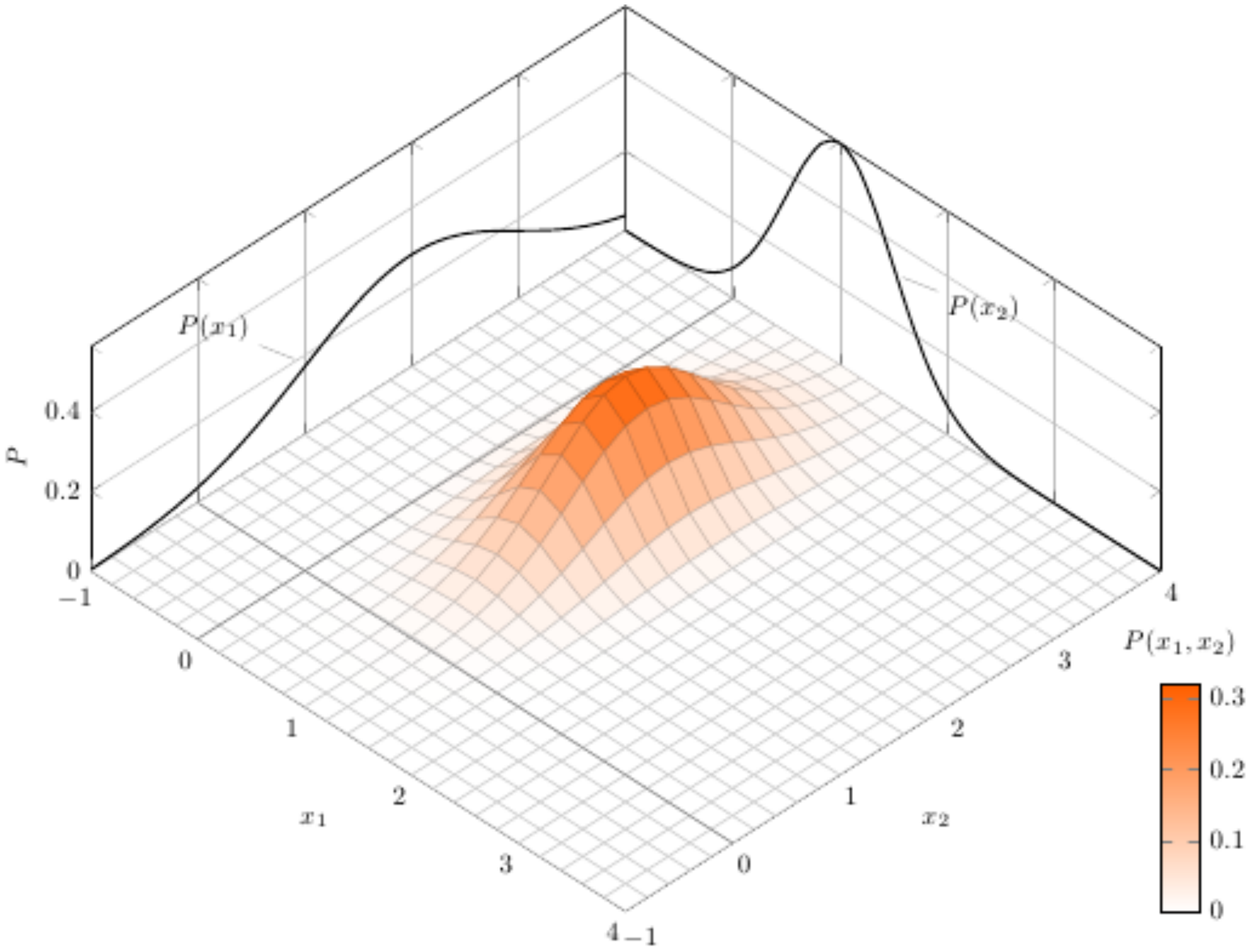
The root of the variance is the standard deviation, √V(x) = σ, which is typically used as a standard measure of spread.

Multiple dimensions

In general, more than one variable is associated to each random event

Take two variables (easy to generalise further): each of N statistical experiments observes of a pair of numbers {(x_1,y_1), (x_2, y_2), …, (x_N, x_N)}

The sample mean and variance are easily generalized to estimate the location and dispersion of the sample along each axis of the multidimensional space.



An additional useful concept quantifies information about the relation between dispersions along different axes.

Covariance and correlation

$$Cov(x,y) = \frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{x})(y_i - \overline{y})$$

Easier to remember: *the mean of the product minus the product of the means*

$$Cov(x,y) = \overline{xy} - \overline{x}\ \overline{y}$$

In N-dimensional data, define the covariance matrix $V_{ij} = Cov(x^{(i)}, x^{(j)})$

Covariance has units so it depends on the choice of units. Better to use a unitless quantity, the Pearson linear correlation

$$\rho(x,y) = \frac{Cov(x,y)}{\sqrt{V(x)}\sqrt{V(y)}} = \frac{Cov(x,y)}{\sigma_x \sigma_y}$$

and its associated correlation matrix $\rho_{ij} = \dfrac{V_{ij}}{\sigma_i \sigma_j}$

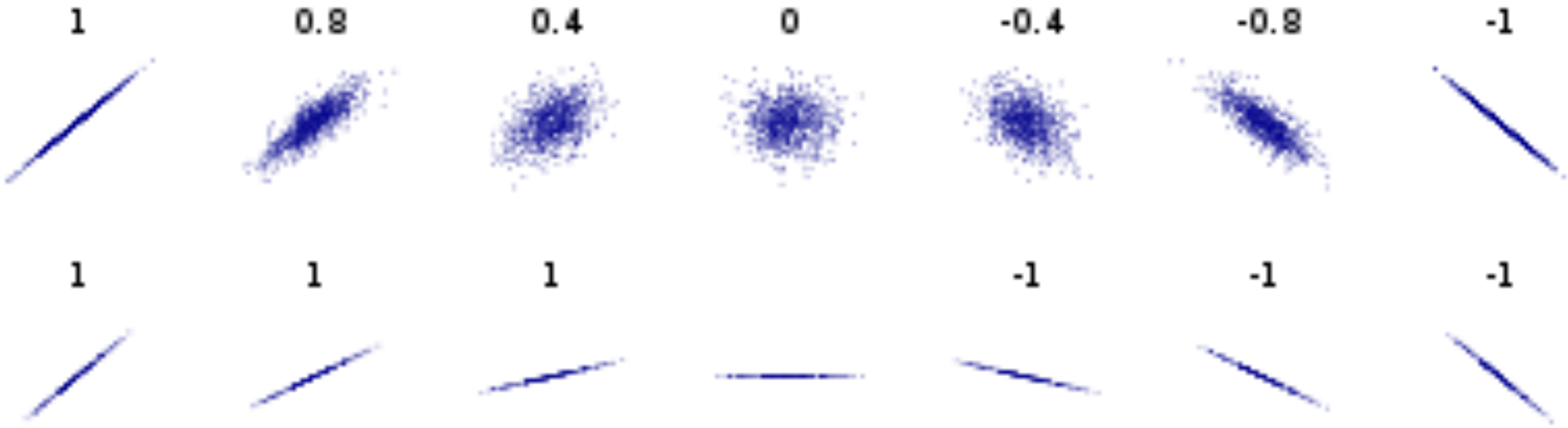
Aside: correlation and dependence

Correlation and dependence between variables are often confused.

Let's set the record straight.

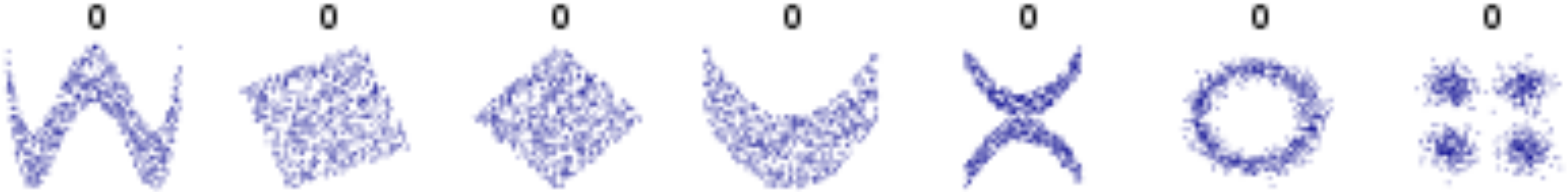
- Two variables x and y are (linearly) uncorrelated if ρ(x,y) = 0
- Two variables x and y are statistically independent if their two-dimensional distribution f(x,y) can be factorized into the product f(x,y) = g(x) h(y). The shape of the distribution of one variable does not depend on the value of the other variable. In other words, information from one variable does not carry information on the other.
- Variables that are independent are also uncorrelated.
- Variables that are uncorrelated may still be dependent

Aside: correlation strength and sign



Note: correlation says nothing about the "slope"

Aside: dependence



In all of these samples, the correlation is zero. But the two variables are clearly not independent.

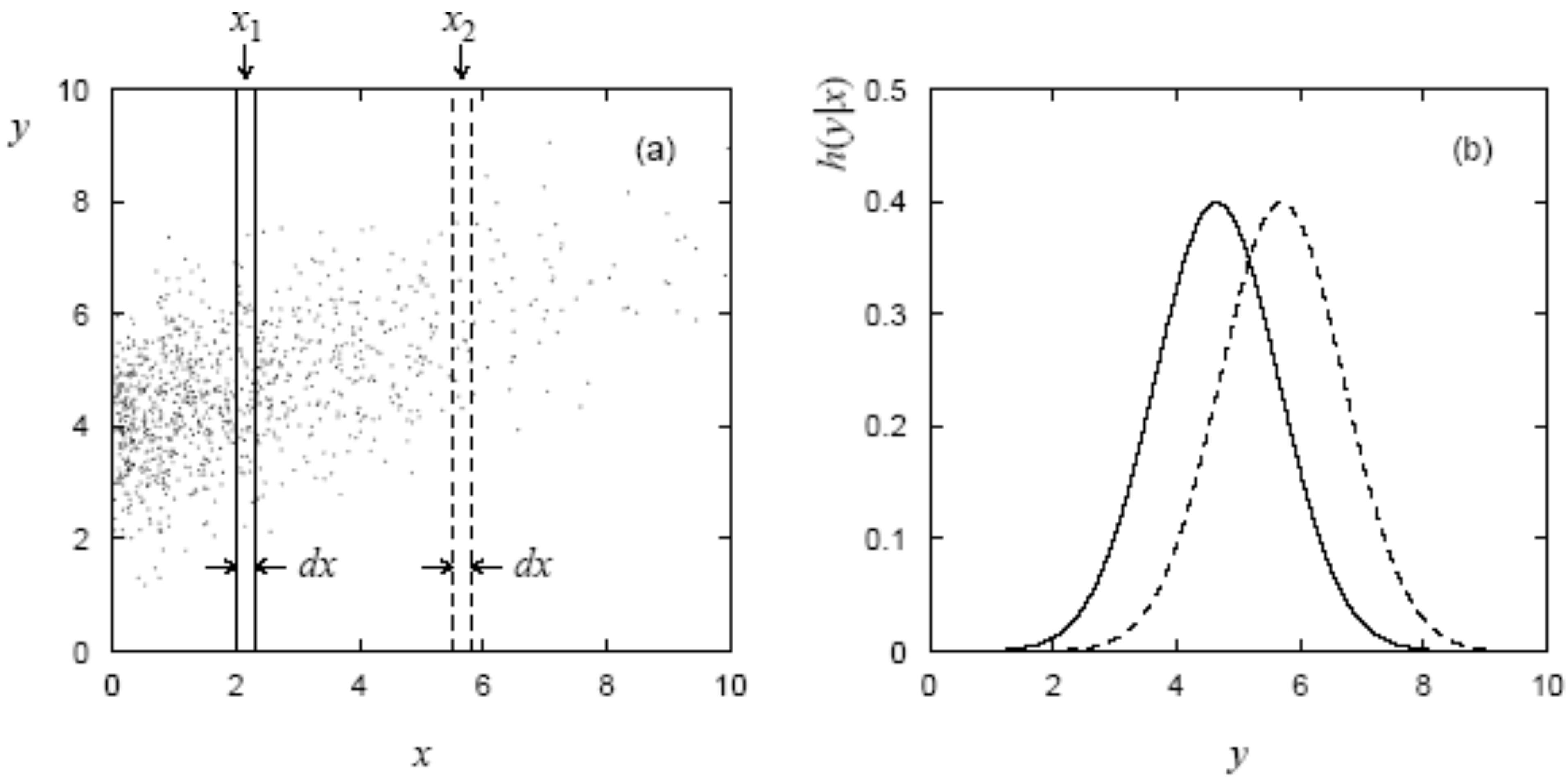
Understanding dependences in multivariate samples of data is important. For instance, in likelihood fits, failure to identify dependence between observables result in wrong results.

Aside: Testing for correlation and dependence

Testing for correlations is easy: just compute the correlation coefficients and make sure they are consistent with zero.

If you see a correlation, then the variables are certainly dependent. If you don't see a correlation, you may need to still check against dependence.

For testing dependence should plot distribution of one variable "in slices" of the other, and check that they overlap.



Aside: causality

Correlations are oft-used to demonstrate causality: causes of phenomena are what is relevant to “understand what’s going on” and build scientific evidence.

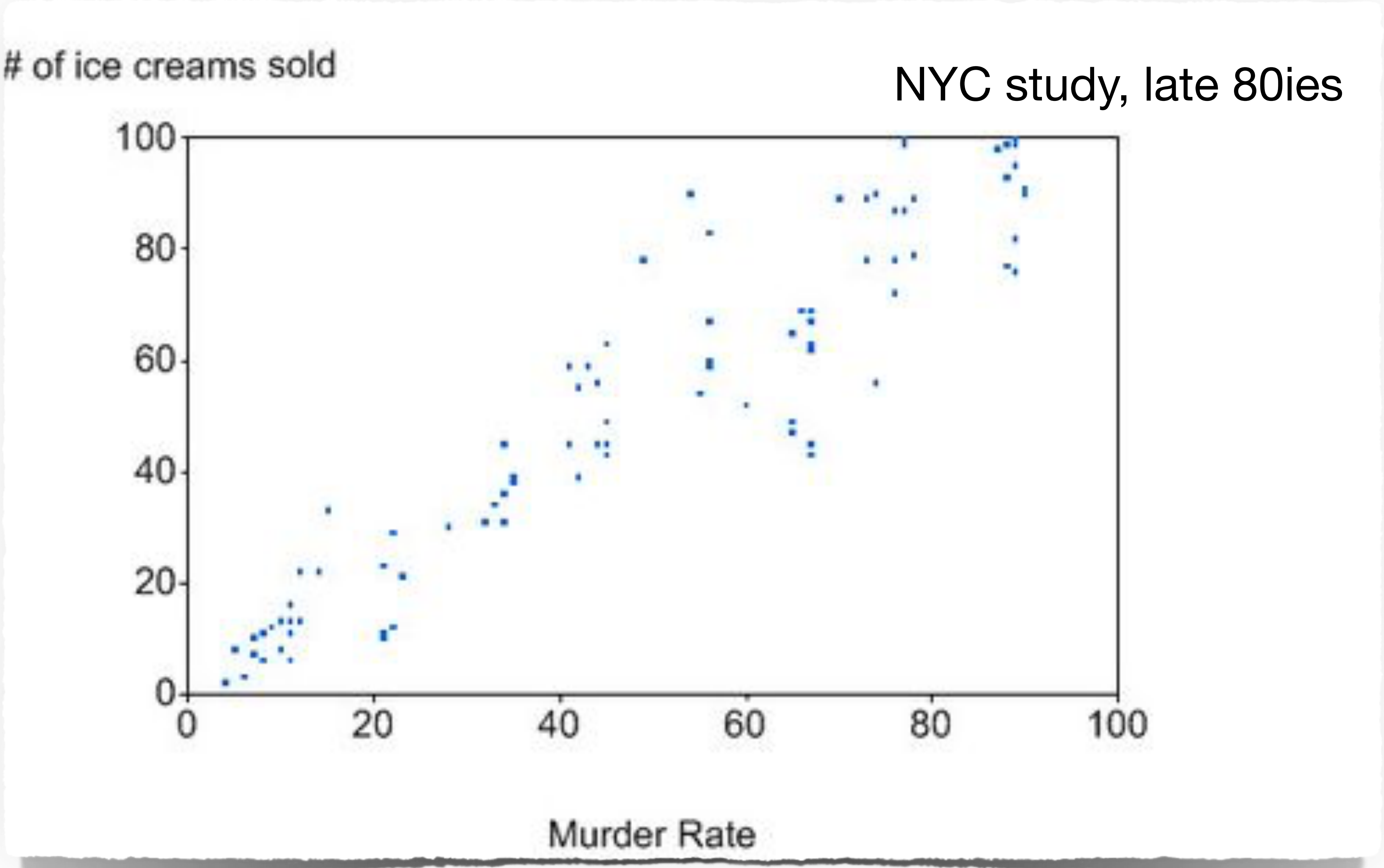
This is a sensitive business. Statistics won’t tell you much about causality. Any statement of causality is necessarily associated with some degree of arbitrariness from the analysts. (Physics relies on established laws that help evaluating plausibility of causal connections. In social sciences, speculations on causality based on observed correlations can get much wilder)

When two phenomena A and B appear to be correlated, it is hard to find out in which of the following cases you are:

- A causes B
- B causes A
- A third phenomenon C causes both A and B
- Correlation is just a coincidence

Aside: A causes B (or B causes A?)

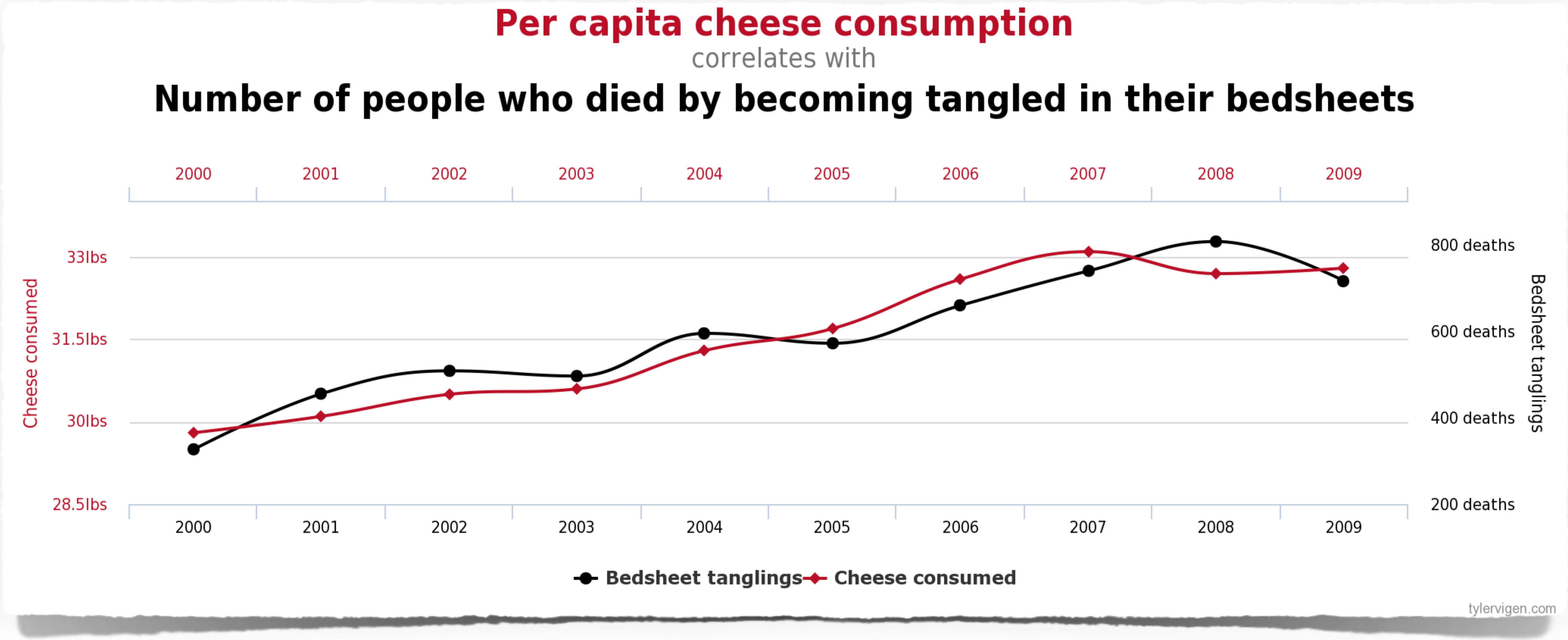
Aside: A third phenomenon C causes both A and B



NYC study, late 80ies

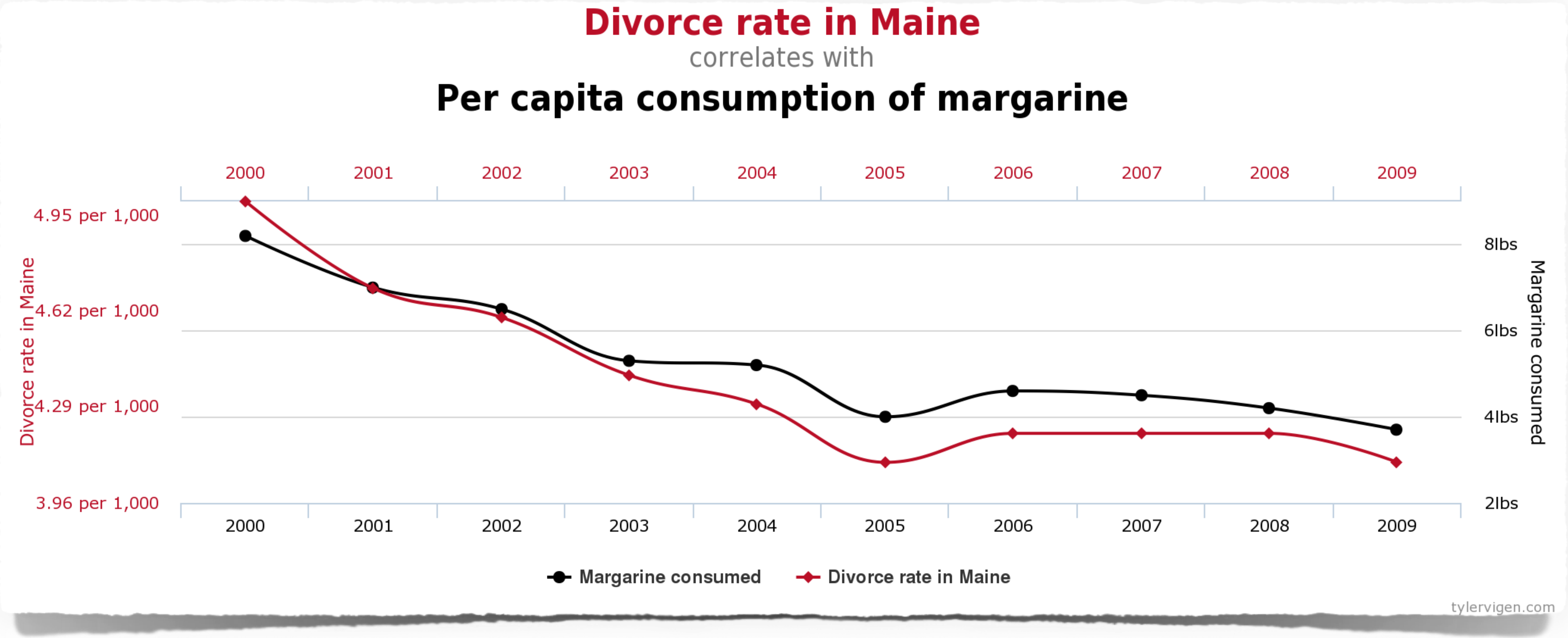
Warm temperatures push people to buy more ice-creams, and also to spend more time outside and party, increasing chances that members of opposing gangs meet and get violent on turf or drug-dealing issues.

Aside: spurious correlations

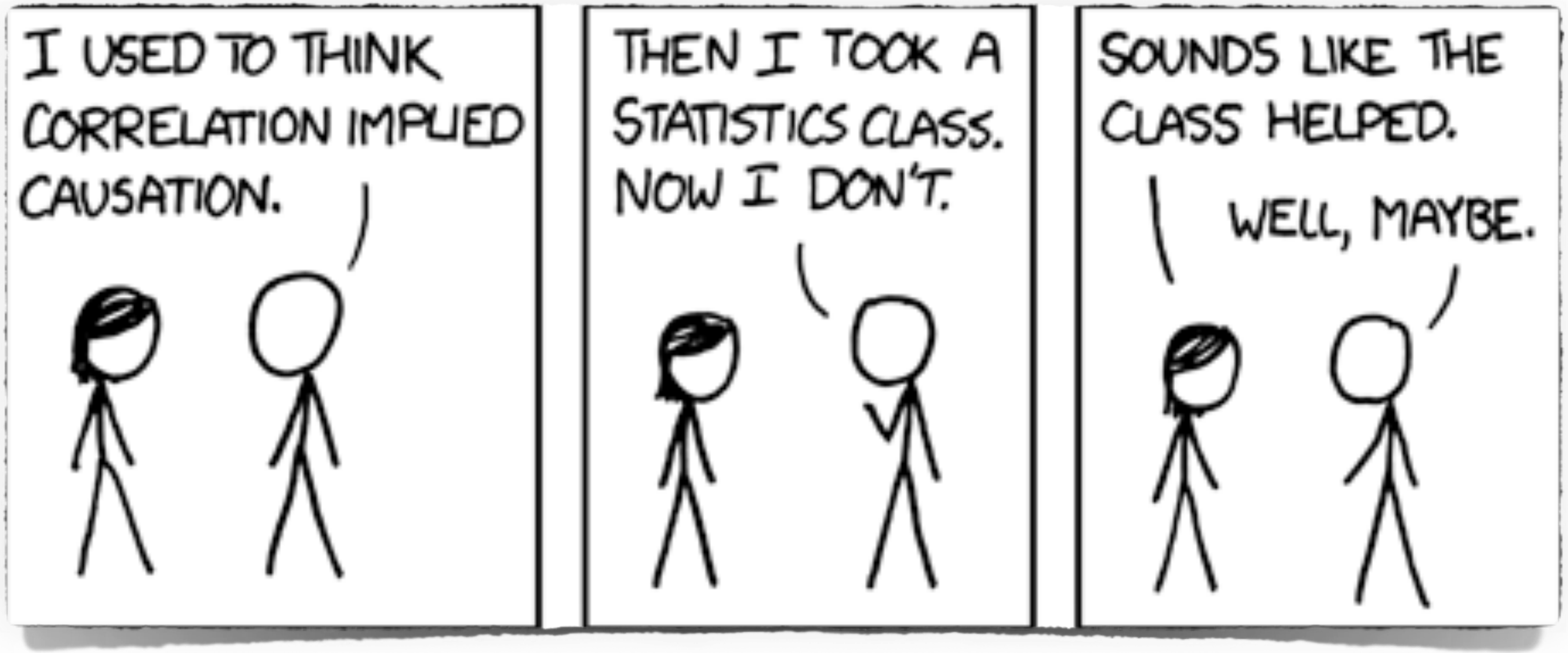


Data: US Department of Agriculture and Center for Disease Control and Prevention. Plot: tylervigen.com

Aside: spurious correlations



Data: National Vital Statistics Reports and US Department of Agriculture. Plot: tylervigen.com


I USED TO THINK CORRELATION IMPLIED CAUSATION.
THEN I TOOK A STATISTICS CLASS. NOW I DON'T.
SOUNDS LIKE THE CLASS HELPED.
WELL, MAYBE.

Describing data

Frequency distributions to pdfs

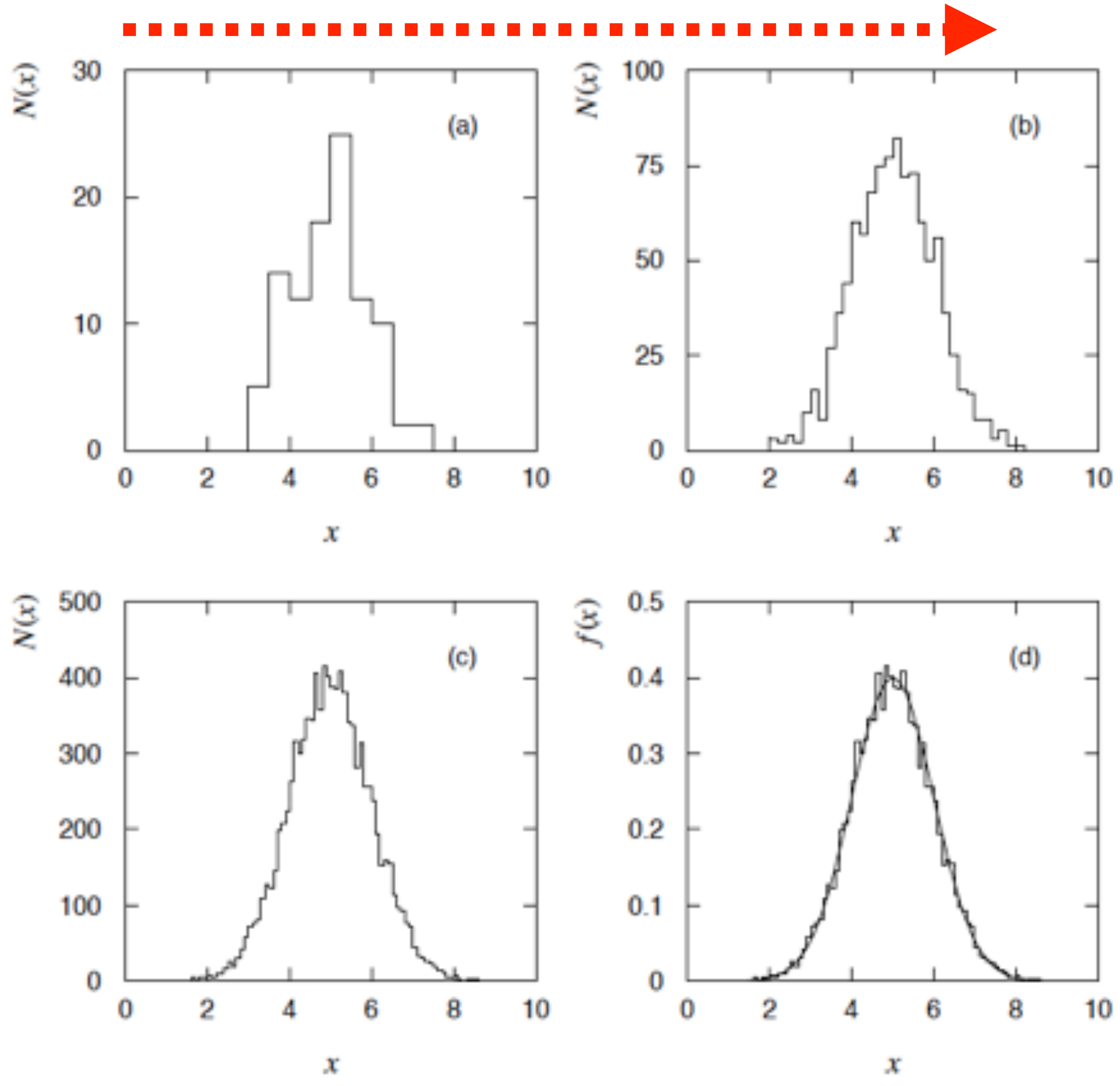
Most frequency distributions in experimental science are highly regular.

This suggest that frequency distributions can be approximated by smooth curves parametrized by simple mathematical expressions.

(Think of bringing the number of observations to infinite, the bin-width to zero, and maintaining unit area)

These would approximate the "theoretical" *probability* functions. Not yet defined what probability is, let's use the intuitive idea as a working approximation for the moment.

increase the number of observations

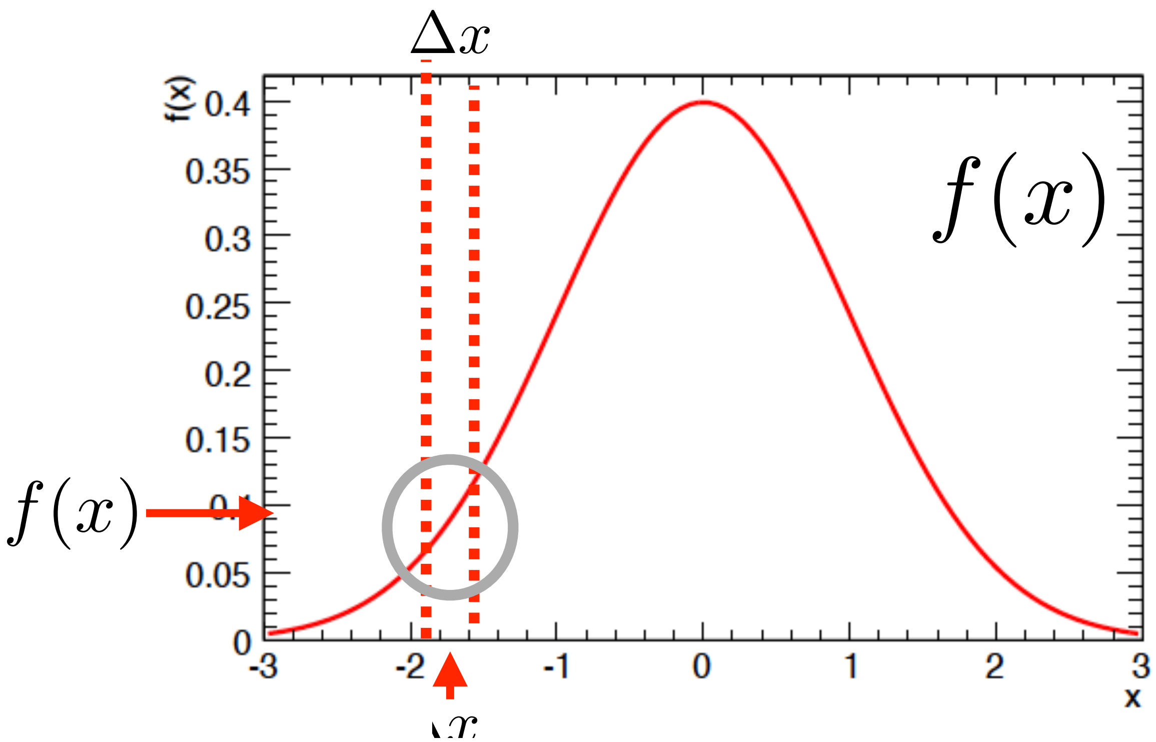


Probability density function

Applies to continuous variables. Choose a short range Δx of the variable. The local frequency of events is approximated by f(x)Δx.

As Δx→0, the probability that x is contained in the range x and x + dx

$$f(x)dx$$



$f(x)$ is the probability density function.

- □ function of the “data” x.
- □ not a probability — has units of x^{-1}
- □ normalized to unity.

The equivalent for discrete variables is the probability mass function, which has no units and is a proper probability

Ubiquitous pdf’s

A few pdf occur frequently in nearly any statistical problem

- Gaussian $$f(x;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{(x-\mu)^2}{2\sigma^2}}$$
- Poisson $$f(j;\mu) = \frac{\mu^j}{j!} e^{-\mu}$$
- Binomial $$f(j;n,p) = \binom{n}{j} p^j (1-p)^{n-j}$$

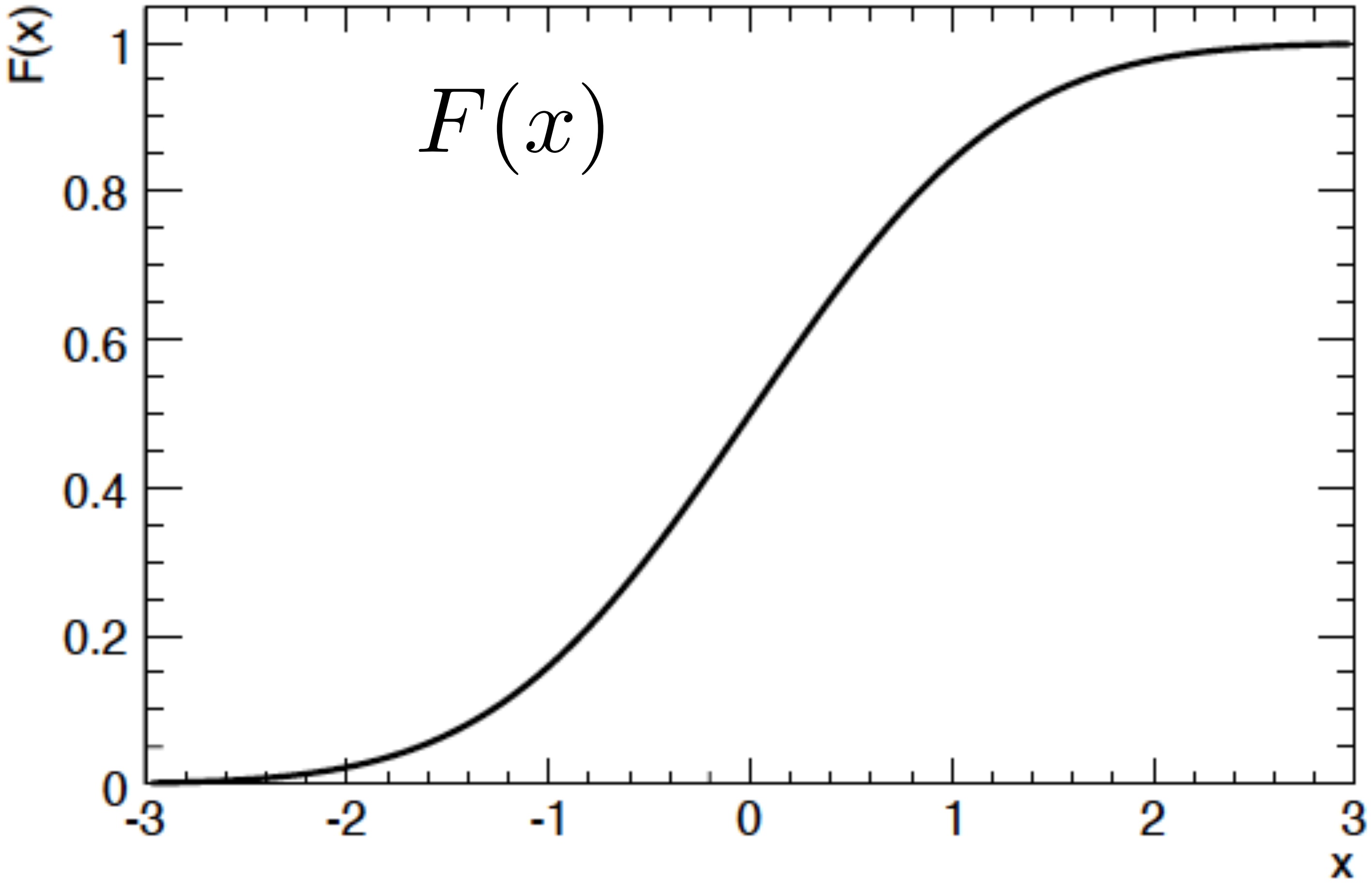
Be familiar with these (more discussion in backup if needed).
Look up http://staff.fysik.su.se/~walck/suf9601.pdf for a more comprehensive list.

PDF are generally multidimensional

$$f(\vec{x};\vec{m}) = f(x_1, x_2, \ldots, x_n; m_1, m_2, \ldots, m_m)$$

Cumulative

$$dF = f(x)dx$$



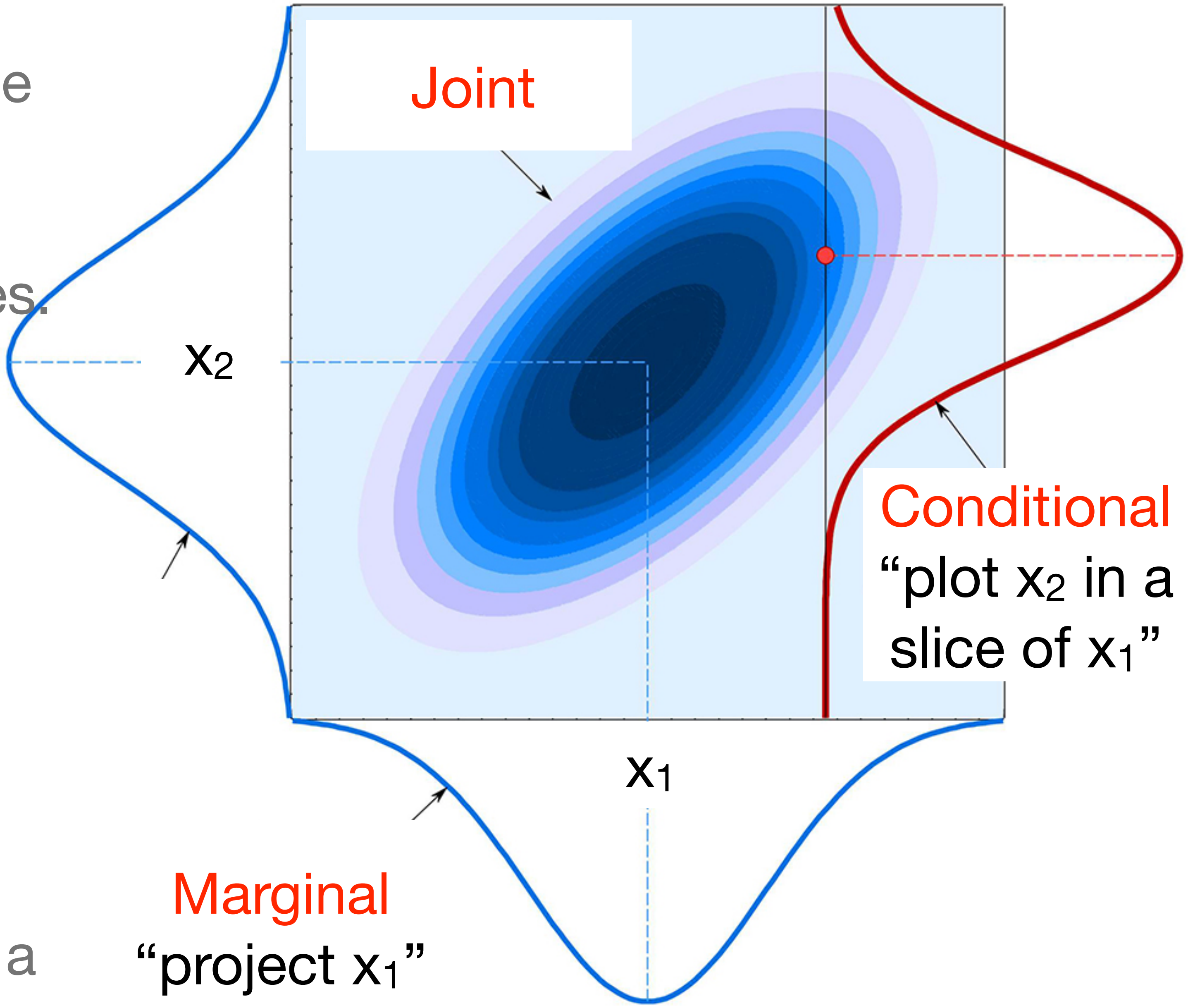
$F(x)$ is the cumulative density function expresses the probability that x is between -∞ and x

Joint, conditional, marginal

$f(x_1, x_2; m)$ is the joint pdf. Contains the whole information. Related to probability that x_1 and x_2 assume simultaneously values in certain ranges.

$f(x_2 \mid x_1; m)$ is the conditional pdf. Related to probability that x_1 is in a certain range, given that x_2 has a specified defined value.

$\int f(x_1, x_2; m)\, dx_2$ is the marginal pdf. Related to the probability that x_1 is in a certain range regardless of x_2 value



Generalize to the n-dimensional pdf $f(x_1, x_2, \dots, x_n)$

Characterizing the pdf

A pdf imposes a weight on each point of the sample space. Can be used to obtain the average value of any function g(x) of the random variable

Expectation value of g $$\langle g(x)\rangle = E[g(x)] = \int g(x)f(x)dx$$

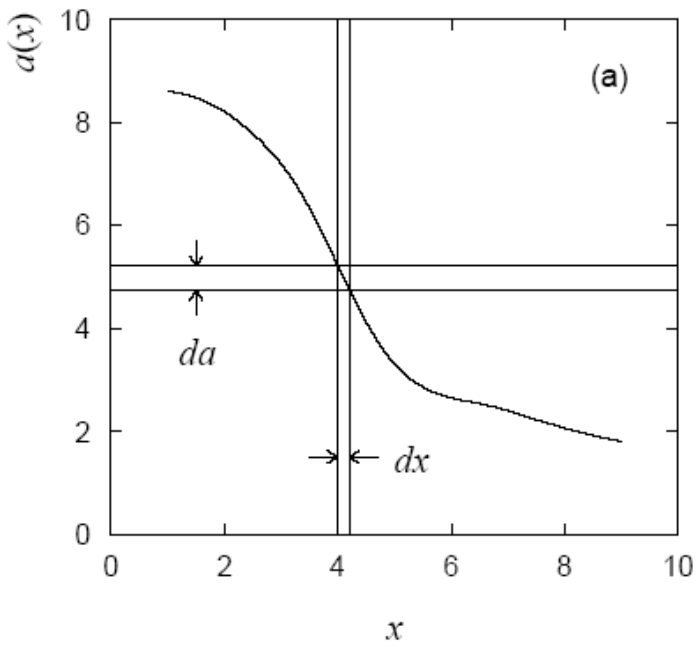
In analogy with what done for data distributions, the theoretical pdfs can be characterized by a few numbers that provide quantitative information of their location and dispersion.

The expectation value of x $$\langle x\rangle = E[x] = \int xf(x)dx$$

The expectation value of $(x\text{-}E[x])^2$

$$V(x) = \langle x^2\rangle - \langle x\rangle^2 = E[x^2] - E^2[x] = \int (x - \langle x\rangle)^2 f(x)dx$$

Pdf g(y) of functions y(x) of a random variable



Functions of random variables are random variables. Take f(x) as pdf of the random variable x and y(x) a function of x (e.g., change of variables). The pdf for y(x) is obtained by imposing the conservation of probability in the two metrics. One-dimensional, one-branched function case is easy:

$$P(x_a < x < x_b) = \int_{x_a}^{x_b} \boxed{f(x)} dx = \int_{y(x_a)}^{y(x_b)} g(y)dy = P(y(x_a) < y < y(x_b))$$

Because

$$g(y) = \frac{f(x(y))}{(dy/dx)}$$

$$\int_{y(x_a)}^{y(x_b)} g(y)dy = \int_{x_a}^{x_b} \boxed{g(y(x)) \left| \frac{dy}{dx} \right|} dx$$

therefore

$$f(x) = g(y) \left| \frac{dy}{dx} \right|$$

The Jacobian that modifies the volume element makes the mode (peak) of the probability density not invariant under change of metric: inferences based on the maximum probability are ill-defined.

A special case — probability integral transform

If x is continuous and f(x) is its pdf, consider the special change of variables that transforms x into its cumulative

$$y(x) = \int_{-\infty}^{x} f(x')dx'$$

Using the relation $f(x) = g(y)\left|\frac{dy}{dx}\right|$ one gets $\left|\frac{dy}{dx}\right| = f(x)$

which result into g(y) =1

Any continuous distribution can be transformed into an uniform distribution.

Important consequences on attempts to attribute a conceptually special role (“uninformative”) to uniform distributions — more later.

From description to estimation

Statistical inference and role of probability

Can we make objective and informative statements about a population when only a sample of the possible observations is available? This is the realm of statistical inference.

Except trivial cases, cannot make inferences from sample to population with the certainty of deductive logic.

Quantitative findings and propositions need be associated with assessments of probability (or confidence, or uncertainty) due to the unobservability of the whole population but only of a random sampling of it.

Unfortunately statisticians divide in two schools, who disagree on first principles, mainly related to the notion and meaning of probability

Statistics: the inverse problem of probability

The theory of probability is a branch of pure mathematics. It is based on axioms and definitions, from which propositions are obtained deductively. The neatest approach is based on set theory, measure theory, and Lebesgue integration.

The theory of statistics is essentially inductive and empirical because it attempts at inferring the values of unknown parameters and information on hypotheses from observation of events

A probability problem: find the probability of observing j heads when tossing a coin N times, knowing the probability of landing heads.

A statistics problem: a coin is tossed N times and lands heads j times. What can one say on the probability of landing heads?

Fundamental ingredients

Given some data, need to

1. Identify all relevant observations x;
2. Identify all relevant unknown parameters m;
3. Construct a model for both

The model

The model is the mathematical structure

$$\boxed{p(\text{data} \mid \text{physics}) = p(x|m)}$$

that incorporates all the physics, knowledge, intuition to best describe the relevant relations between observables x and unknown parameters m.

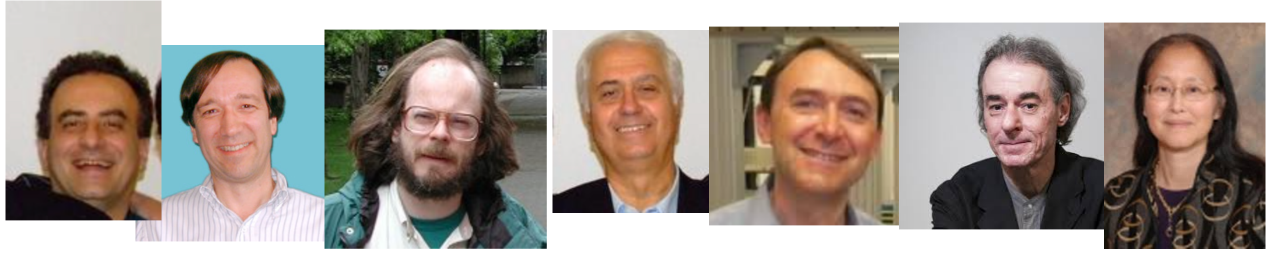
It is a probability model — *you don’t know exactly what value of x would be observed even if you knew exactly the value of m.*

The width of $p(x|m)$ is connected to the statistical uncertainty of your inference

The model is the fundamental building block of most of HEP inference, both in Frequentist and Bayesian procedures. This is the step everyone agrees on.

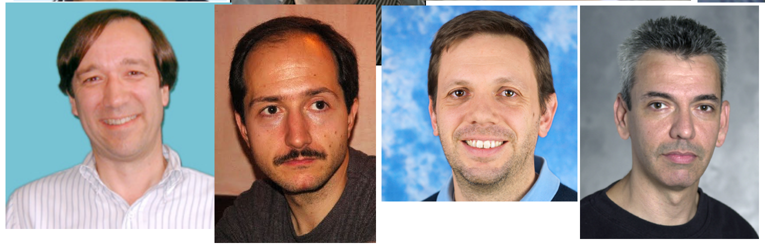
Tentative stopping point

Many thanks



to G. Punzi, B. Cousins J. Heinrich, L. Ristori, E. Milotti, F. Le Diberder. K. Kinoshita

for enlightning many of the notions discussed here in formal lectures, discussions, etc…



to G. Cowan, K. Cranmer, A. Rogozhnikov, H. Prosper, M. Kagan, T. Junk, T. Hastie, F. James, R. Barlow, J. Rademacker, L. Lyons, B. Cousins, T. Dorigo, N. Berger, E. Gross

for making your slides publicly available so that I could steal from them.