Statistics for HEP (1/3)

Diego Tonelli (INFN Trieste)
diego.tonelli@cern.ch

Pre-SUSY2017 School
Dec 6, 2017 - TIFR, Mumbai

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 tasks

- Measure the value of a physics parameter — point estimation
- Finding its uncertainty — interval estimation
- Comparing one hypothesis agains another (in search for anomalies/ discoveries) — hypothesis testing
- Comparing one hypothesis against all others — Goodness of fit

Understanding nature from blurred observations

Top-down vs bottom-up understanding

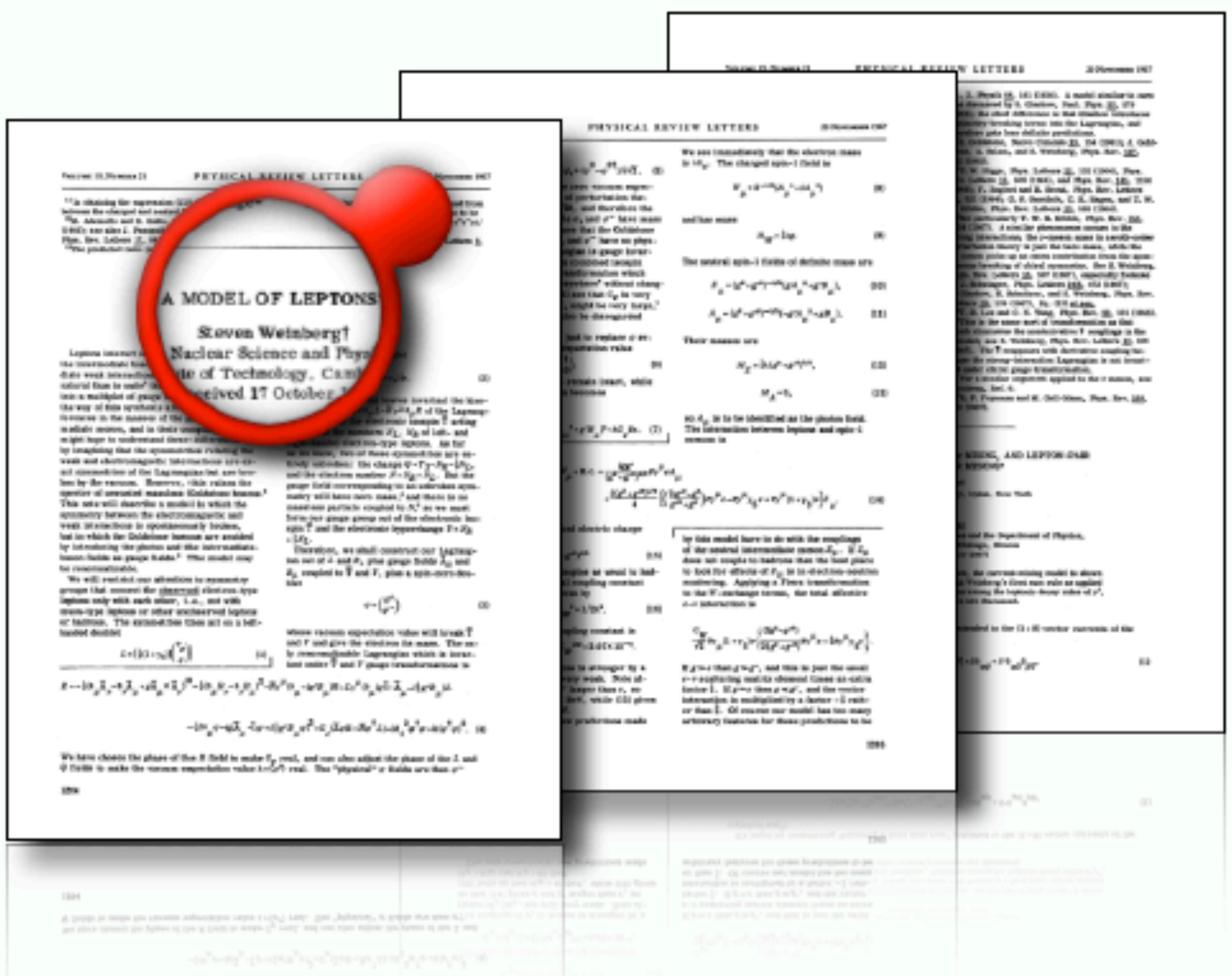
Similar to low-level perception processes, HEP advances through the interplay of top-down (theory-guided) and bottom-up (data-driven) processing.

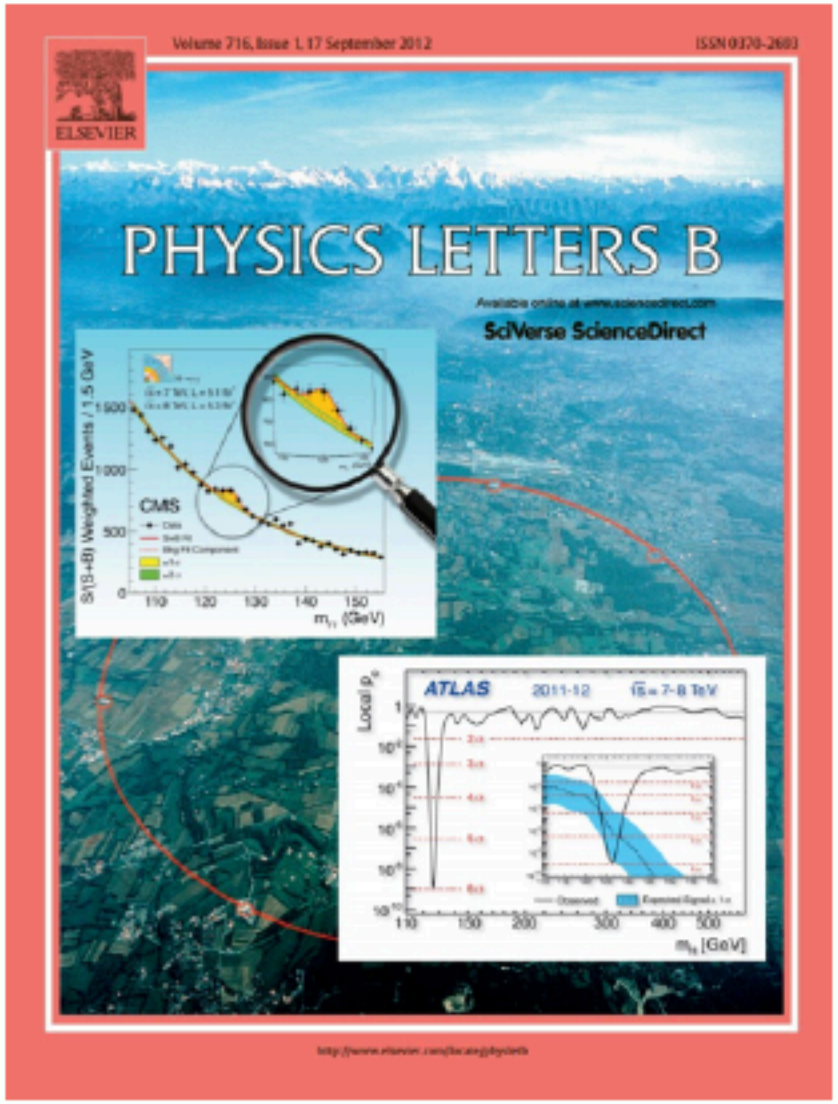
The need for detail (quality and quantity of data) is driven by the distinctiveness of the phenomena and our level of familiarity with it.

When a roadmap suggest “what to expect”, a little data goes 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?

2012 — …: a new data driven era?

No.

Good news: many fundamental questions remain open: why 3 quark and lepton families? Why their mass hierarchies? Origin of CP violation? What's dark matter? And dark energy? [your favorite question here]

Bad news is that top-down luxury is over.
[Is that truly bad news for experimentalists?]

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 quality data from many different experimental environments.

A particularly fitting time to focus on methods of extracting information from the data.

What to expect

This won't be a tutorial/cookbook. There won't be any hands on.

I'll insist on few fundamental concepts. Hope this will consolidate (or establish) foundations for you to dig further, enrich what you already know, and expose you to some different points of view.

These lectures won't be forward-looking. Rather focused on the core basics. Excellent material from CERN schools and and online stuff by K. Cranmer, M. Kagan, A. Rogozhnikov, T. Junk etc. is great to fill you in on most recent/ongoing developments. (Detailed refs will be given on our last day)

I will take it easy. My goal is that you pick up most of this in real time and interrupt me with questions when not.

I have no lecture notes. So tried to compose fairly descriptive slides aiming at making the logic decipherable offline too. Additional materials and some derivations in the backup for reference. Please let me know of mistakes.

Outline

Today, Wed Dec 6 — Quick recap on basics. Statistical inference. Bayesian vs frequentist. Pdf vs likelihood. Maximum likelihood.

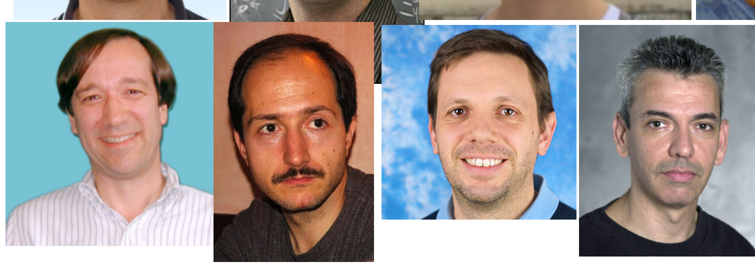
Tomorrow, Thu Dec 7 — Confidence-intervals. Likelihood-ratio ordering Systematic uncertainties. Profile-likelihood ratio. Hypothesis testing.

Fri, Dec 8 — Introduction to statistical learning, linear discrminants, the multilayer perceptron, decision trees.

Many thanks

to G. Punzi, B. Cousins J. Heinrich, L. Ristori, E. Milotti, D. Derkach

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.

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.

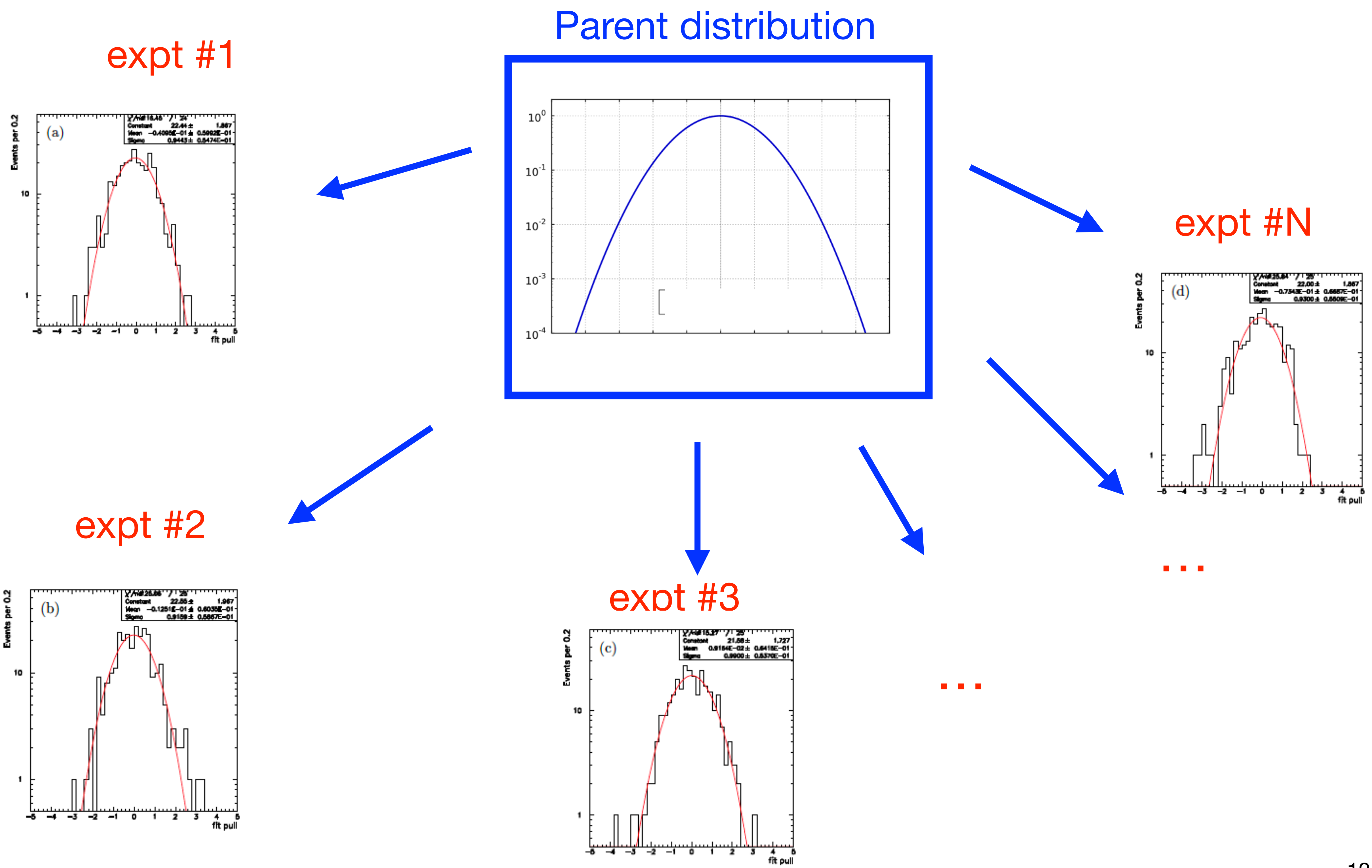
Random events are associated to variates ("(random) variables", "observables") x, which take different values, corresponding to different possible outcomes. Each x value has its probability* p(x). The outcomes generate a probability distribution of x.

A collection of random events forms a population: the hypothetical infinite set of 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

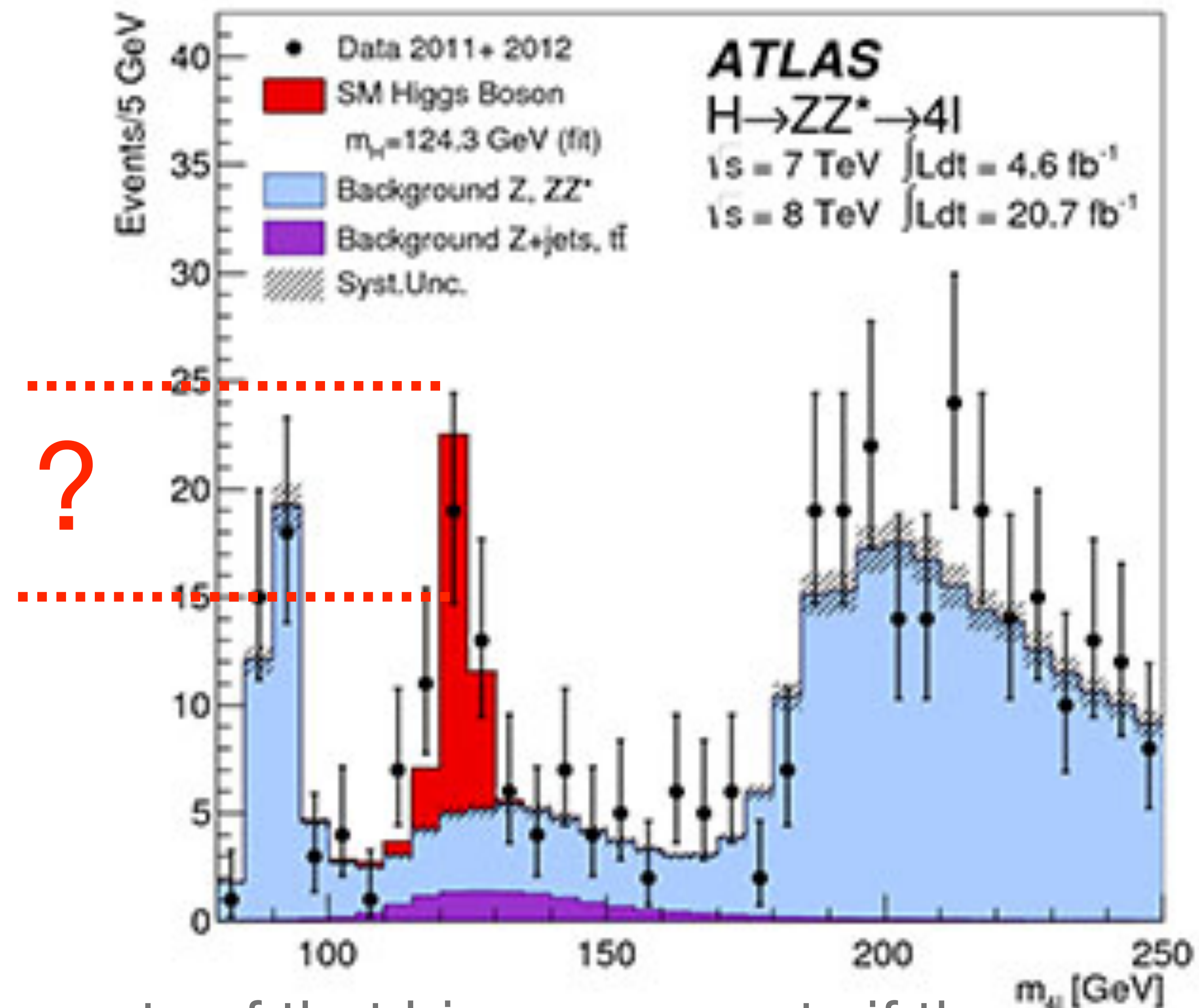


You do it everyday

Most of you regularly quote uncertainties in counting experiments.
E.g, in an histogram, a bin with N entries has an error bar (e.g., of length $\sqrt{N}$)

What that 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.*

https://wwwusers.ts.infn.it/~dtonelli/HCPSS2017/DataPointsErrorIvovanVulpen.pdf for some fun

Data location

Simple and most common quantity to summarize the sample 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

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

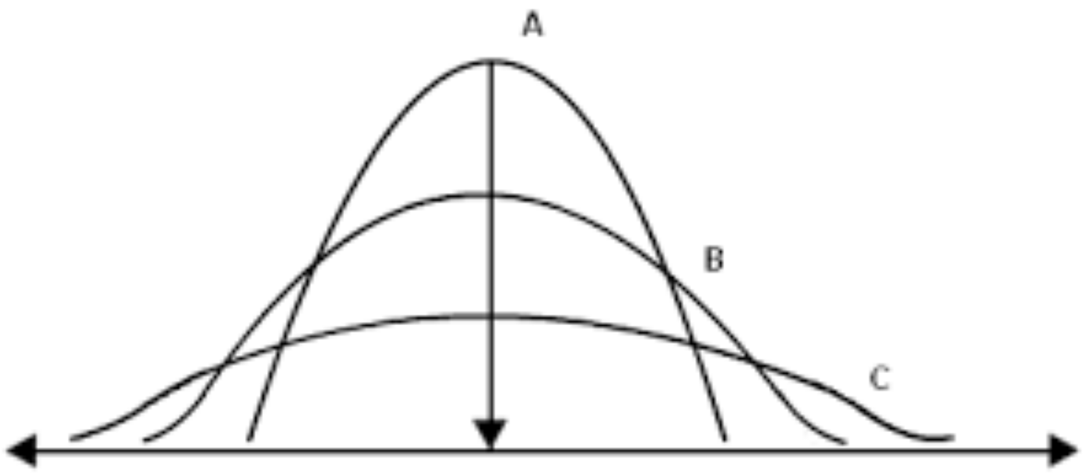
$$\text{Binned sample mean} \quad \bar{x} = \frac{1}{N}\sum_{j=1}^{n} x_j n_j$$

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

Data dispersion

The mean says nothing about the dispersion of data, another key information to grasp the features of a sample

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*

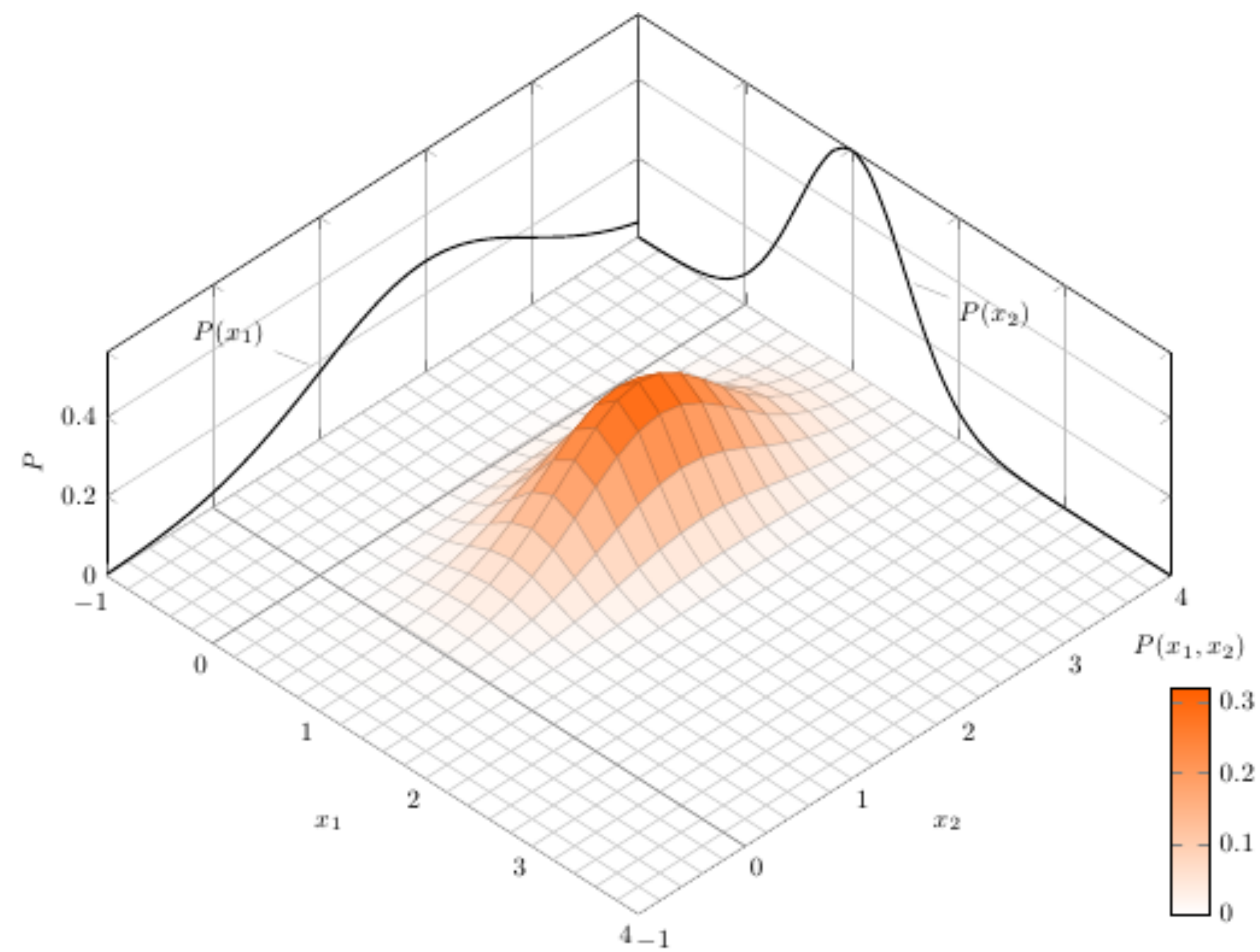
$$V(x) = \overline{x_i^2} - \overline{x}^2$$

The root of the variance is the standard deviation, $\sqrt{V(x)} = \sigma$. 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 $\{(x_1,y_1), (x_2, y_2), \ldots, (x_N, x_N)\}$



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 relates the 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, defines a matrix $V_{ij} = Cov(x^{(i)}, x^{(j)})$

Cov has 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 associated correlation matrix $\rho_{ij} = \frac{V_{ij}}{\sigma_i \sigma_j}$

Correlation and dependence

Correlation and dependence between variables are sometimes confused.

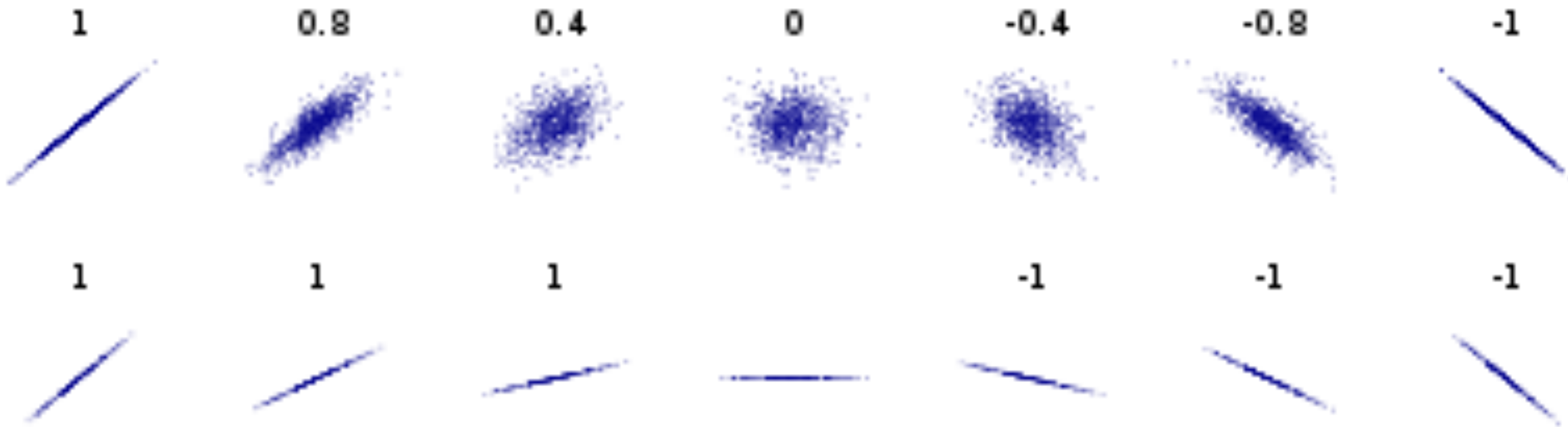
Two variables x and y are (linearly) uncorrelated if $\rho(x,y) = 0$

- They are statistically independent if their two-dimensional distribution $f(x,y)$ can be factorized into the product $f(x,y) = g(x)\,h(y)$. That is, the shape of one distribution does not depend on the value of the other variable. Information from one variable does not carry information on the other.
- Independent variables are also uncorrelated.
- Uncorrelated variables may still be dependent

In pictures

[Wikipedia]

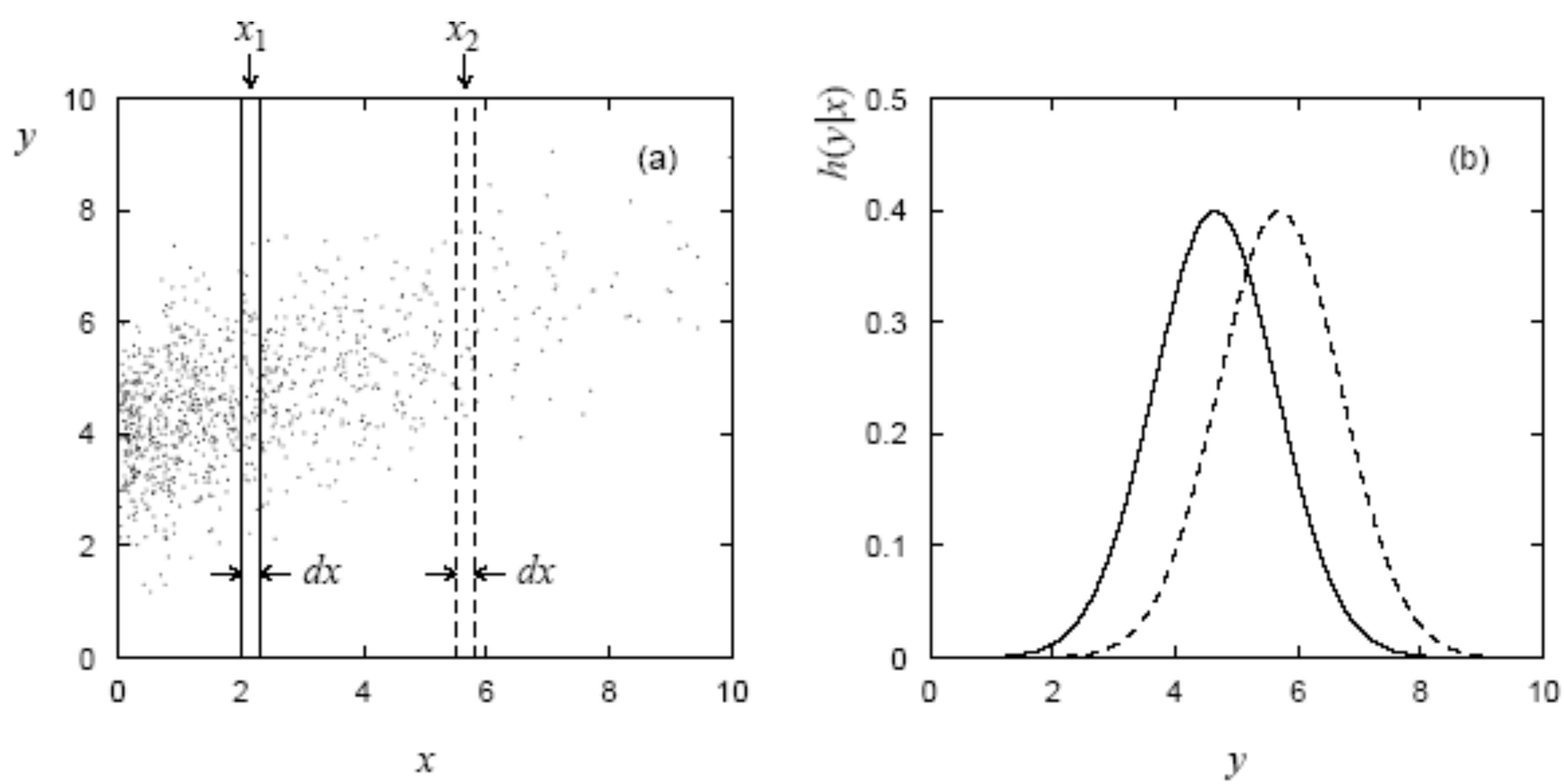
correlation strenght says nothing about the “slope”



In all cases below, correlation is zero. But the two variables are clearly not independent.

Testing for correlation and dependence

Testing for correlations: just look at the correlation coefficients. If they are nonzero, variables are certainly dependent. If they are zero, may want to check against dependence: check if the distributions of one variable “in slices” overlap.



[Cowan]

Correlation and causality

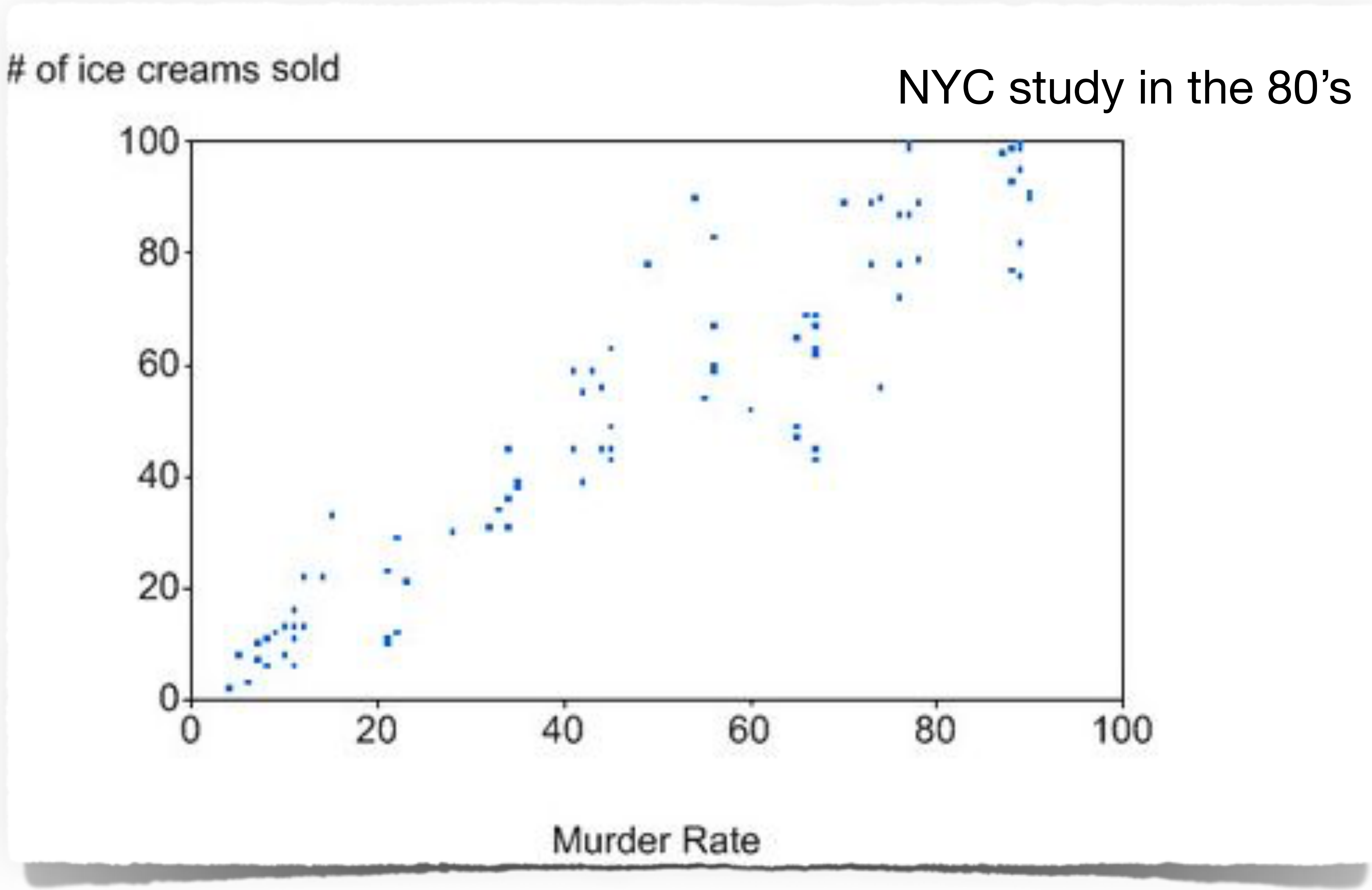
Often correlations are used to implicate causality as causes of phenomena are relevant to “understand what’s going on” and build scientific evidence.

Statistics won’t tell much about causality.

Phenomena A and B that show correlation could mean

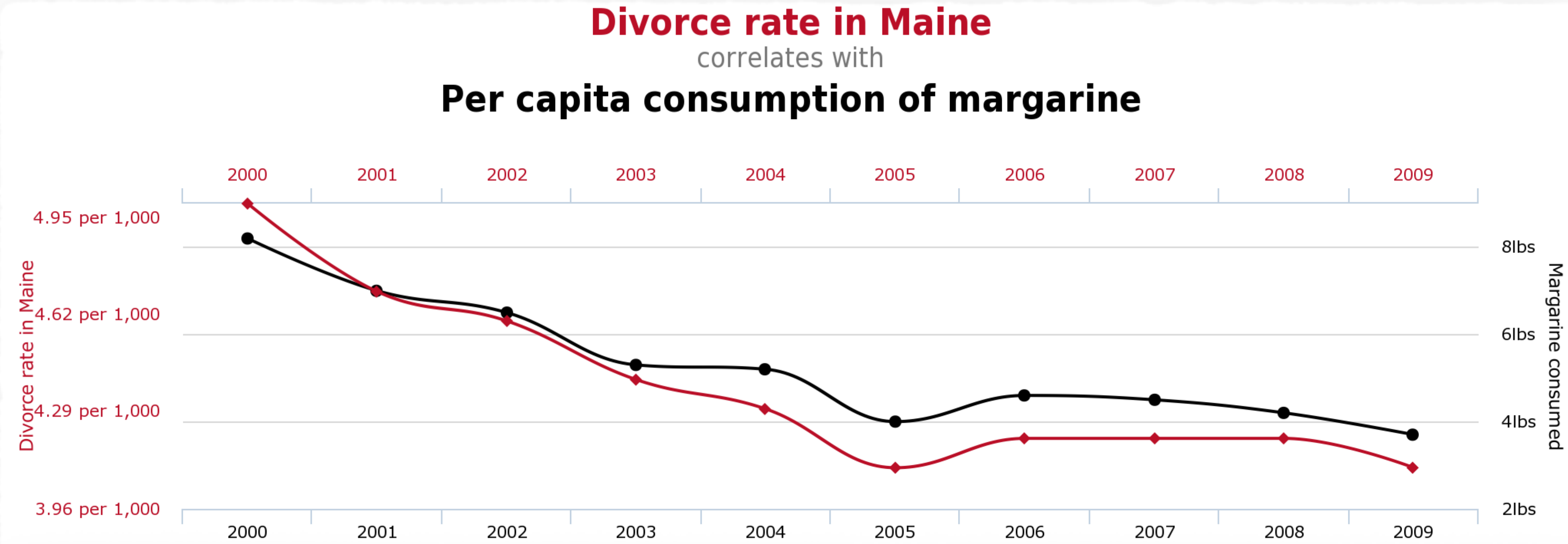
- A causes B
- B causes A
- A third phenomenon C causes both A and B
- Coincidental correlation

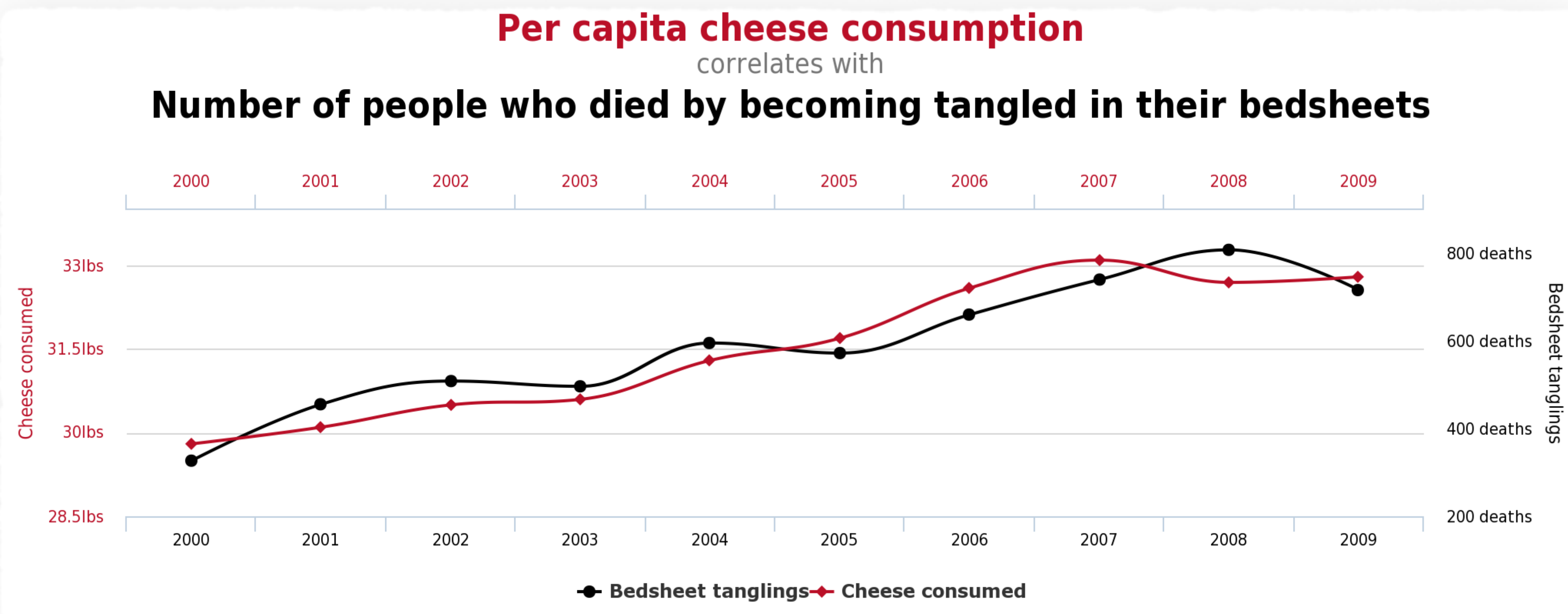
Triangulation



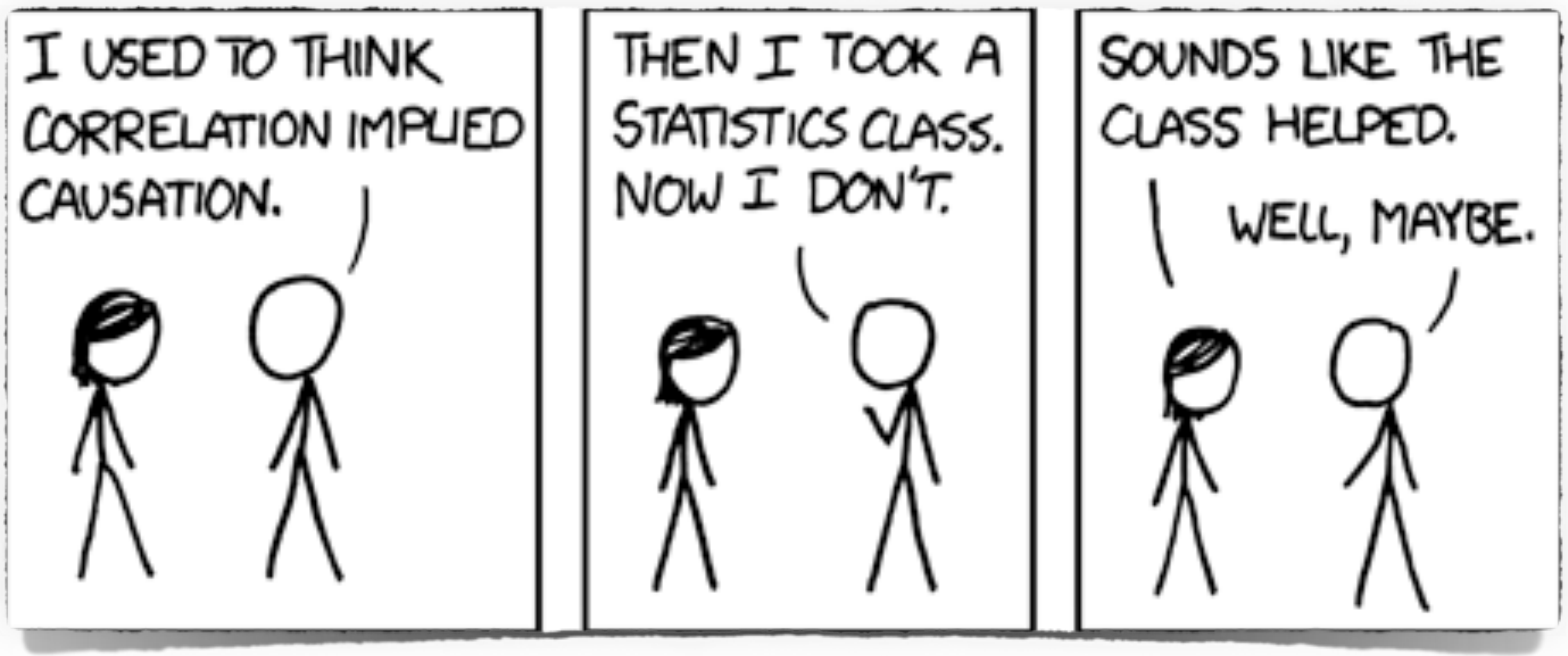
Warm temperatures push people to buy more ice-creams, and also to spend more time outside and party, increasing chances that gang members meet and get violent.

Coincidence





Sources:National Vital Statistics Reports, US Department of Agriculture,Center for Disease Control and Prevention. Plot: tylervigen.com

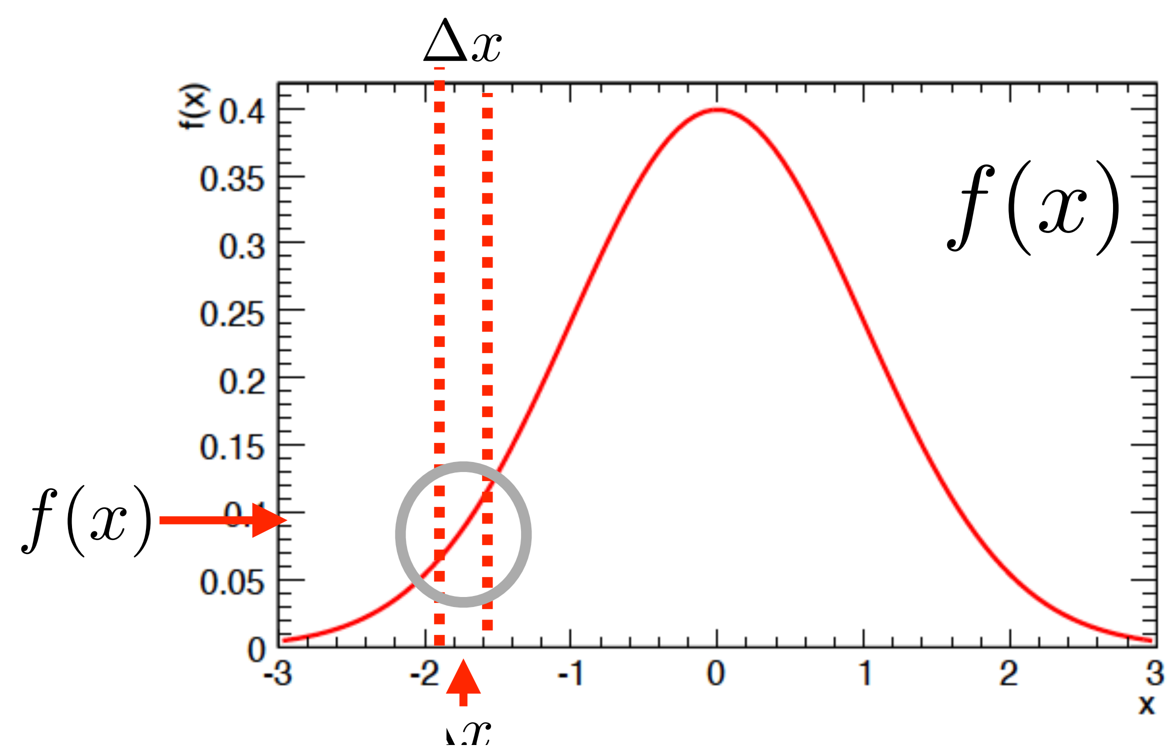

I USED TO THINK CORRELATION IMPLIED CAUSATION.
THEN I TOOK A STATISTICS CLASS. NOW I DON'T.
SOUNDS LIKE THE CLASS HELPED.
WELL, MAYBE.

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.

It is a function of the “data” x.

Typically pdf shape depends on model-parameters: $f(x|\alpha)$ “f of x given α”

It is not a probability: has units of x^{-1}

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

It is normalized to unity.

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 www.fysik.su.se/~walck/suf9601.pdf for a more comprehensive list.

It is generally multidimensional

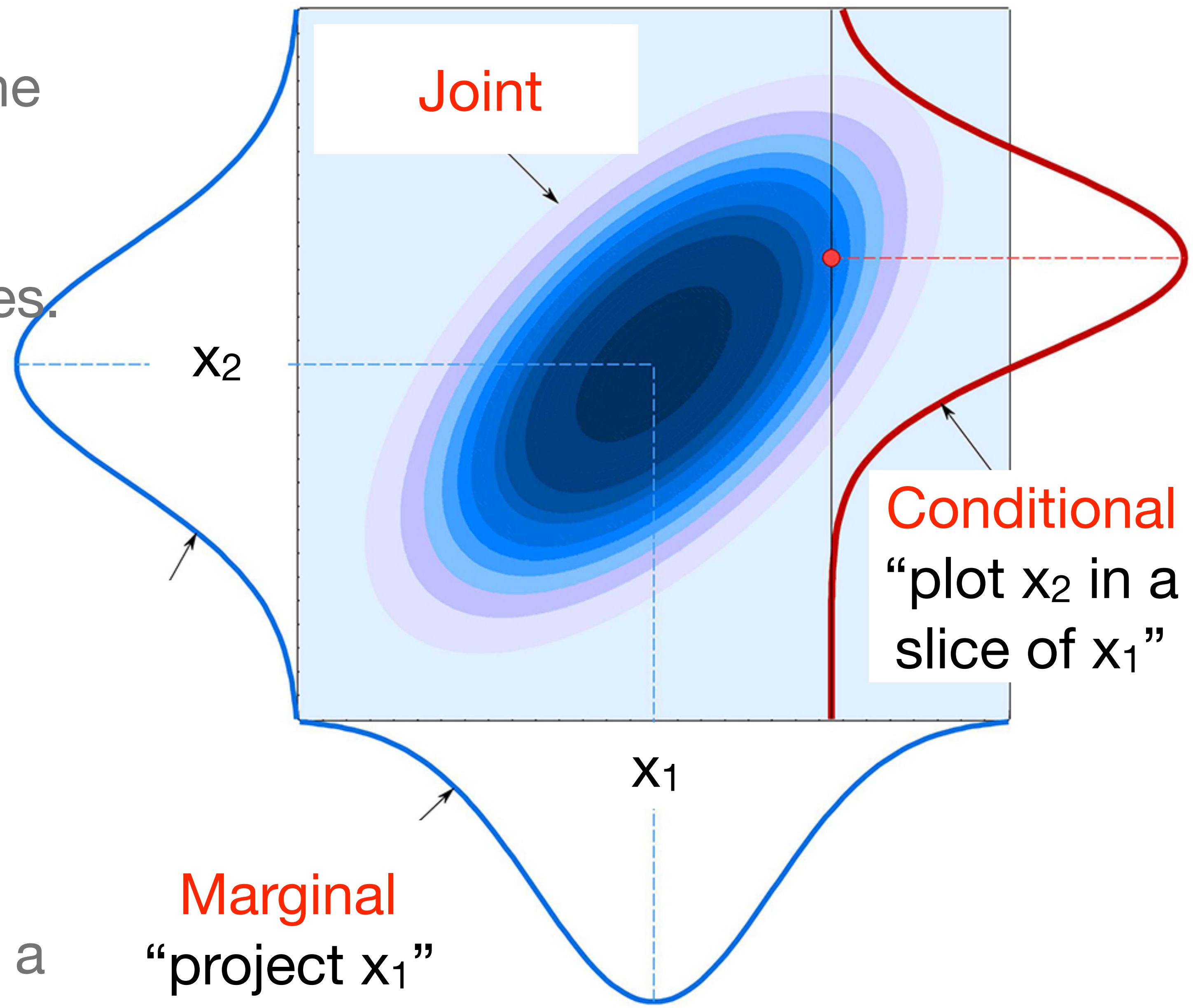
$$f(\vec{x};\vec{m}) = f(x_1, x_2, ..., x_n; m_1, m_2, ..., m_m)$$

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 $f(x_1, x_2, \ldots, x_n)$

Characterizing the pdf

The pdf can be used as weight 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 samples, pdfs can be characterized by a few numbers that quantify their location and dispersion.

The expectation value of x is the mean of x $\langle x\rangle = E[x] = \int x f(x) dx$

The expectation value of $(x-E[x])^2$ is the variance of x

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

Might be nondefined for some pdf. E.g., Cauchy (Breit-Wigner) pdf.

Functions of random variables

Functions of random variables are themselves random variables. Take f(x) as pdf of the random variable x and y(x) a function of x (e.g., change of variables).

Conservation of probability between the two metrics yields g(y), the pdf for y(x). Because it is an integrated quantity involves the Jacobian.

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

Because

$$\int_{y(x_a)}^{y(x_b)} g(y)dy = \int_{x_a}^{x_b} g(y(x)) \left| \frac{dy}{dx} \right| dx$$ 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: renders ill-defined the inferences based on maximum probability density.

A special case — probability integral transform

Take x continuous with pdf f(x). Consider the change of variables that transforms x into its cumulative y(x), that has pdf g(y).

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

Using $f(x) = g(y) \left| \frac{dy}{dx} \right|$ one gets $\left| \frac{dy}{dx} \right| = f(x)$ which yields g(y) =1

Any continuous distribution can be transformed into an uniform distribution. Or alternatively, there is always a metric in which the pdf is uniform:

- the inverse transformation allows efficient MC generation of p(x) using a generator of random numbers between 0 and 1.
- this property questions the special role frequently attributed to uniform priors in Bayesian inference (more later)

Inferring from data

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

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

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 if m had some definite value.*

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

The approximate model

The model p(x|m) is assumed as your best approximation of the actual relationshops between m and x relevant for the problem at hand.

Parametrize differences with the actual physics through additional dependencies on unknown nuisance parameters — p(x|m,v).

The unknown v values are uninteresting for the measurement but do influence its outcome. Lack of knowledge of v introduces an uncertainty in the p(x|m,v) shape.

Not only you don't know exactly what value of x would be observed if m had a definite value, you don't even know exactly how probable each possible x value is.

The uncertainty in the shape of p(x|m) reflects into the systematic uncertainty of the inference.

www-cdf.fnal.gov/physics/statistics/notes/punzi-systdef.ps

Role

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

The model is also the single strongest driver of inference performance: improving the model is the best way of improving the inference.

- With parameters m fixed, the model is the probability density function of data, which provides the ability to generate pseudodata via Monte Carlo.
- With data fixed, the model is the likelihood function of the m parameters

Model building

Three main thrusts for model motivation/justification.

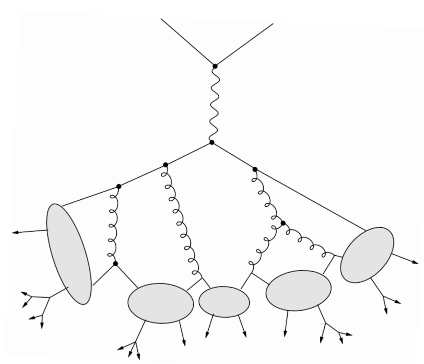
Monte Carlo modeling

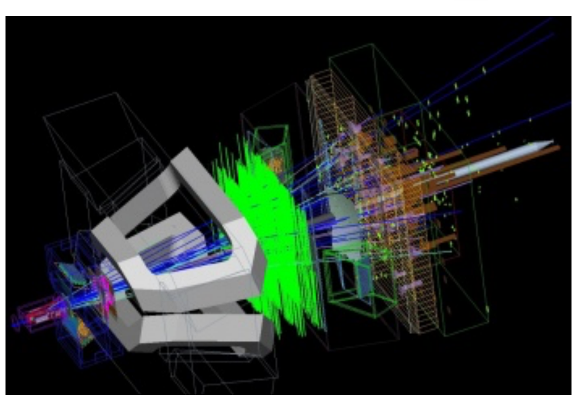
$$\mathcal{L}_{SM} = \underbrace{\frac{1}{4}\mathbf{W}_{\mu\nu}\cdot\mathbf{W}^{\mu\nu} - \frac{1}{4}B_{\mu\nu}B^{\mu\nu} - \frac{1}{4}G^a_{\mu\nu}G^{\mu\nu}_a}_{\text{kinetic energies and self-interactions of the gauge bosons}}$$

$$+ \underbrace{\bar{L}\gamma^\mu(i\partial_\mu - \frac{1}{2}g\tau\cdot\mathbf{W}_\mu - \frac{1}{2}g'YB_\mu)L + \bar{R}\gamma^\mu(i\partial_\mu - \frac{1}{2}g'YB_\mu)R}_{\text{kinetic energies and electroweak interactions of fermions}}$$

$$+ \underbrace{\frac{1}{2}\left|(i\partial_\mu - \frac{1}{2}g\tau\cdot\mathbf{W}_\mu - \frac{1}{2}g'YB_\mu)\phi\right|^2 - V(\phi)}_{W^{\pm}, Z, \gamma, \text{and Higgs masses and couplings}}$$

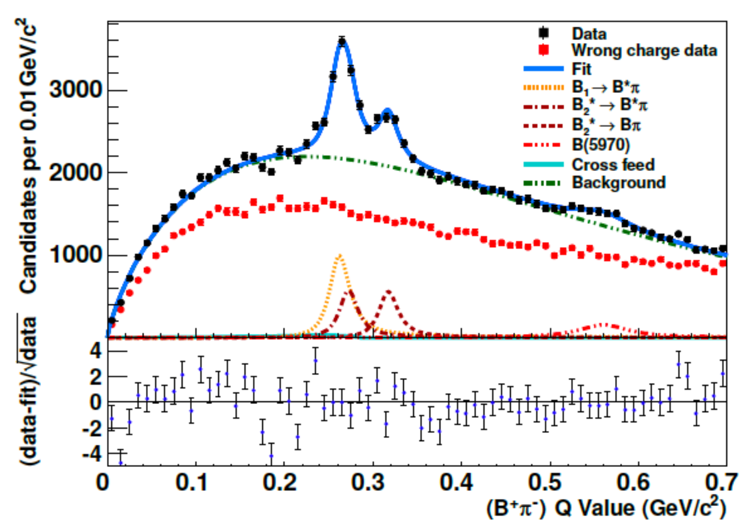
$$+ \underbrace{g''(\bar{q}\gamma^\mu T_a q)G^a_\mu}_{\text{interactions between quarks and gluons}} + \underbrace{(G_1\bar{L}\phi R + G_2\bar{L}\phi_c R + h.c.)}_{\text{fermion masses and couplings to Higgs}}$$





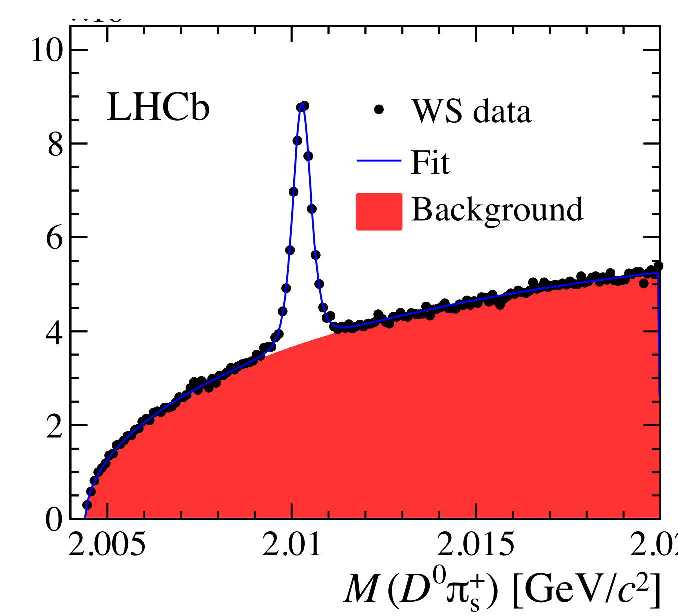
Data driven modeling

- Sideband subtraction
- Same-charge candidates
- Mixed-event candidates
- ABCD methods
- …



Effective modeling

Empirical modeling



[Cranmer]

Tools

Complexity of models increases with the number of data sets, analysis channels in each data set, model components in each channel etc.

LHC experiments marked an order-of-magnitude increase in model complexity with respect to LEP/HERA/Tevatron/B-factories, especially driven by Higgs boson search: combinations of O(100) channels, likelihoods with O(1000) parameters.

RooFit (originally developed at BaBar) offer a consistent framework to provide tools for collaborative building and handling of complex models.

https://root.cern.ch/roofit-20-minutes

RooStats interfaces with RooFit to offer higher-level statistical tools based on such models.

https://wwwusers.ts.infn.it/~dtonelli/HCPSS2017/RooStats.pdf

Inference

The model gives probability to observe a certain set of data assuming some physics

p(data | physics) is known.

Forward process. From physics to data occurs in

- running experiments (physics true but unknown) and
- simulation (physics known but not necessarily true).

The backward process from data to physics is the inference: make objective and quantitative statements about a population when only a sample of the possible observations is available.
Such generalization isn't generally possible using the certainty of deductive logic.
Unobservability of the parent distribution, but only of a random sampling of it, imposes assessments of probability (or confidence, or uncertainty)

Probability

Two approaches: different notions of probability yield differing inferences.

Frequentist — conceive repeated independent samples

$$P(A) = \lim_{N \to \infty} (N_A/N)$$

- Uses information observed in data (and that could have been observed in other trials).
- Data are random, theories not. Only applies to repeatable "events". Restricts to deductions based on p(data | theory). Favored theories are those for which our observations are more *usual*.

Bayesian — subjective degree of belief

- combines info from observed data with subjective judgment. Same data with different analysers may yield inconsistent results.
- Treat as random variable any unknown. Broader applications, including to theories/hypotheses.
- Addresses p(theory | data) the inductive reasoning one is interested to.

In short

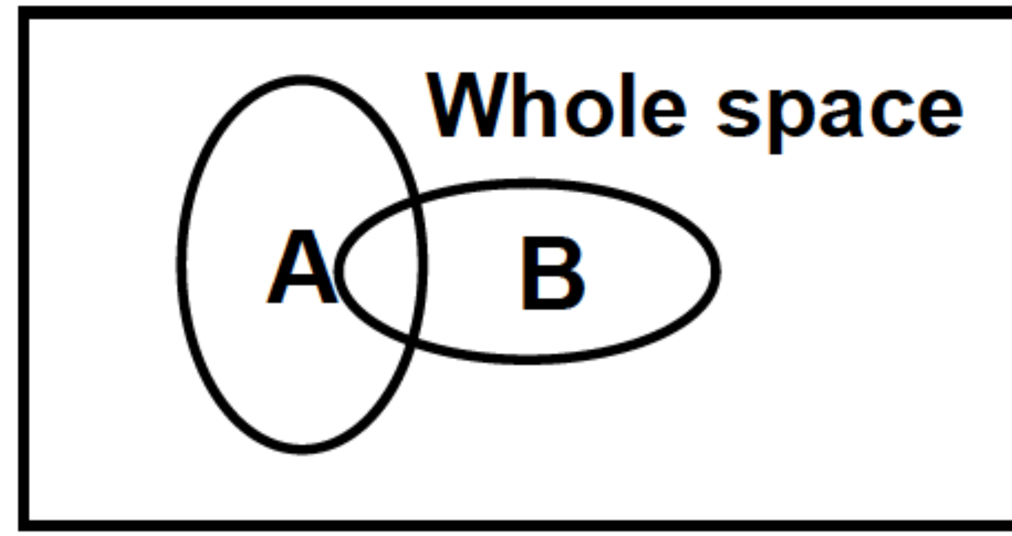
Frequentist use impeccable logic to deal with an issue of no interest to anyone.

Bayesians address the question everyone is interested in, by using assumptions no-one believes

[Lyons through Cranmer]

Whole space

In both cases, for probabilities to be well defined, the whole space or sample space need be defined (determines normalization)



"90% of our flights arrive on time"
Flight delayed several hours are canceled, not 'delayed', so they get excluded from our sample space.

"Our survey shows that most people lose 5 Kg in a month on this diet"
Happy customers who lost weight are most likely to respond to our survey. The ones who gained weight most likely threw away our survey postcard.

Whole space can be thought as the space of available possibilities given (i.e., conditional to) the assumptions associated with the model (e.g., was a Poisson process, whether or not background is in..)

Bayesian inference

Conditional probabilities

Probability for jointly observing A and B

(Conditional) probability for A given B

(Marginal) probability for B

$$P(A \text{ and } B) = \begin{cases} P(A|B) * P(B) \\ P(B|A) * P(A) \end{cases}$$

(Conditional) probability for B given A

(Marginal) probability for A

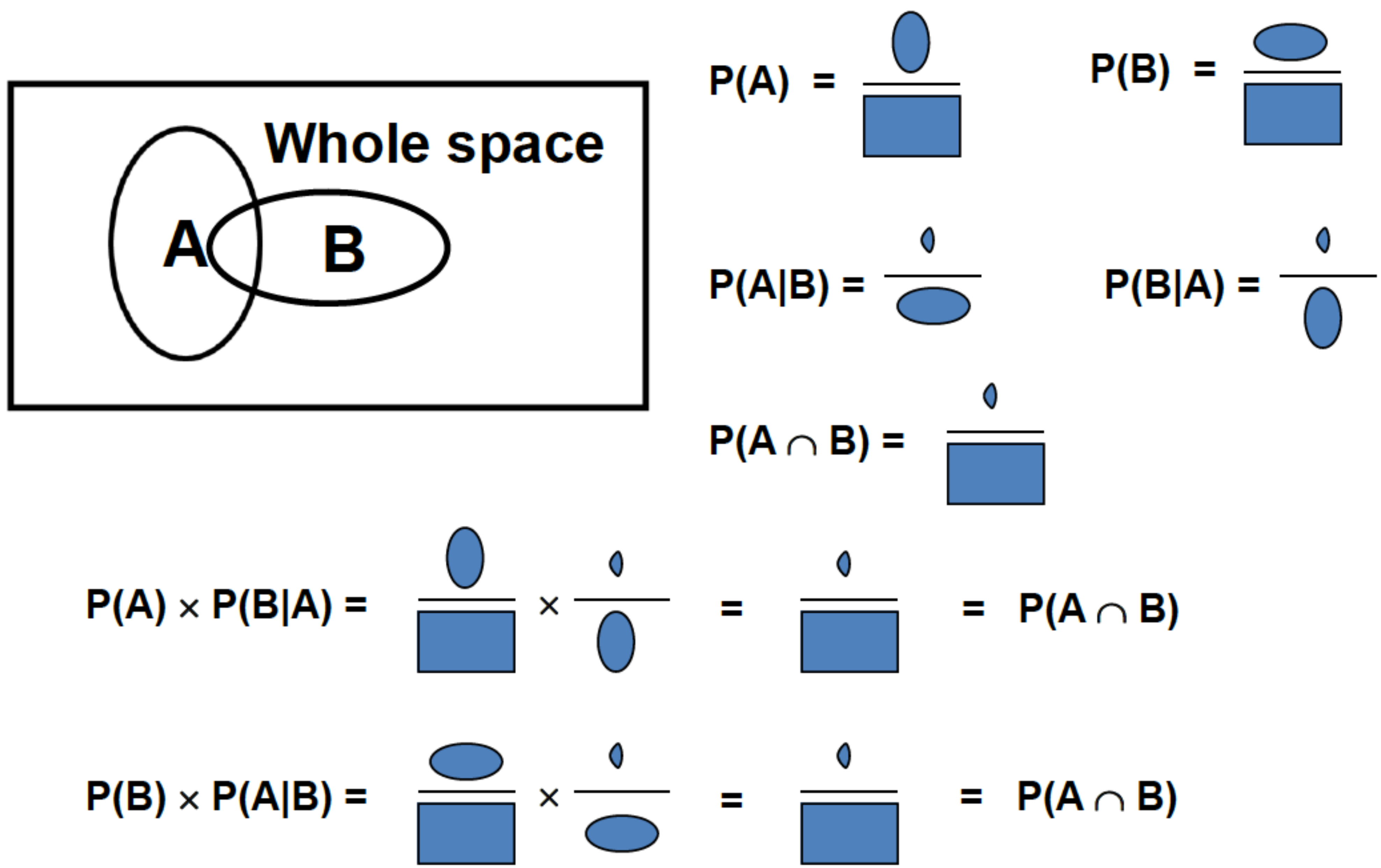
Bayes' theorem

Yields a key relation between conditional and marginal probabilities.

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)} = \frac{P(A|B)P(B)}{P(A|B)P(B) + P(A|\mathrm{not}B)P(\mathrm{not}B)}$$

- P(B|A) is the conditional probability for B given A. Also called posterior because evaluated *after* fixing a specific value of A
- P(A|B) is the conditional probability of A given B
- P(B) is the prior probability for B, evaluated *before* knowing any information on A
- P(A) is the marginal (or "prior") probability for event A. Serves as normalization.

Probability, conditional probability and Bayes Theorem — in pictures



Bob Cousins, HCPSS, 2009

$\Rightarrow$ **P(B|A) = P(A|B) × P(B) / P(A)**

Remember

P(A|B) is NOT equal to P(B|A).

Variable A: “pregnant”, “not pregnant”

Variable B: “male”, “female”.

P(pregnant | female) ~ 3% but

P(female | pregnant) >>> 3% !

[Lyons]

Remember


New Scientist
www.amibritisnomot.com
...men & babies
IT'S HERE!
£5.50 UK
ISBN 1-574568-456

Applying Bayes' theorem to inference

Take x, an observable random variable, and m, an inobservable random variable, with known probability distribution p(x,m). Observe x ("perform a measurement of x"), what can I say about m? Want to know p(m|x).

Bayes theorem tells me all I possibly need. Allows determining the "a posteriori" probability for any value of m (look at backup slides for an elementary example)

Model — Prior probability

$$p(m|x) = \frac{p(x|m) \times p(m)}{p(x)}$$

Posterior probability — Normalization

If x and m are independent p(x|m) = p(x) and therefore p(m|x) = p(m). The probability a posteriori equals that a priori: measurement is non informative

Prior

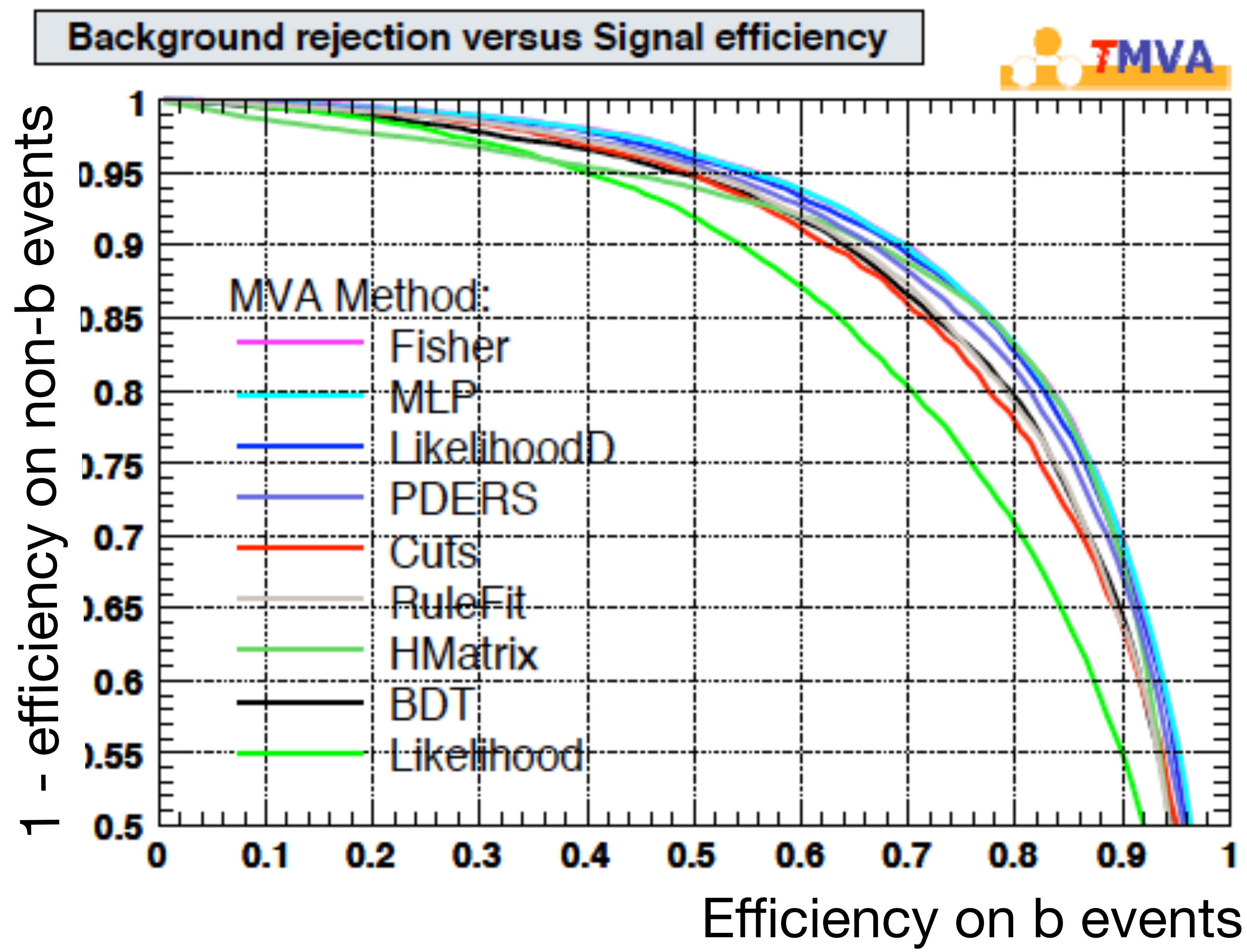
Algorithm to identify b-jets.

Run it on a sample of b-jets and a sample of non-bjets and plot

- abscissa: p(btag | b-jet)

- ordinate: p(nobtag | non b-jet)

for each algorithm setting

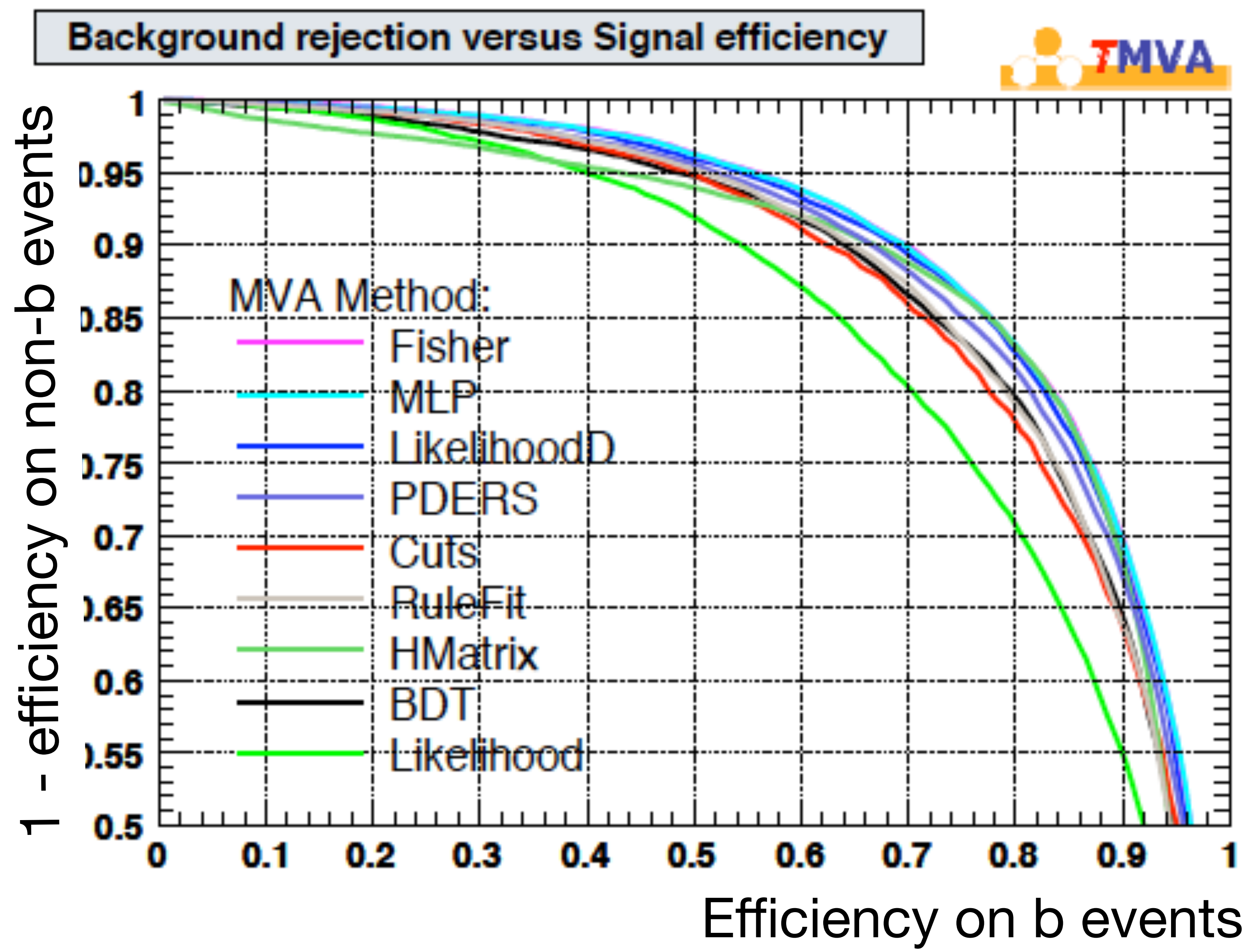


Given a sample of jets, what fraction are b-jets? I.e., what is p(b-jet| btag)?

[Cousins]

Prior

Cannot answer.



Need to know the fraction of b-jets in my sample, that is the prior p(b-jet).

[Cousins]

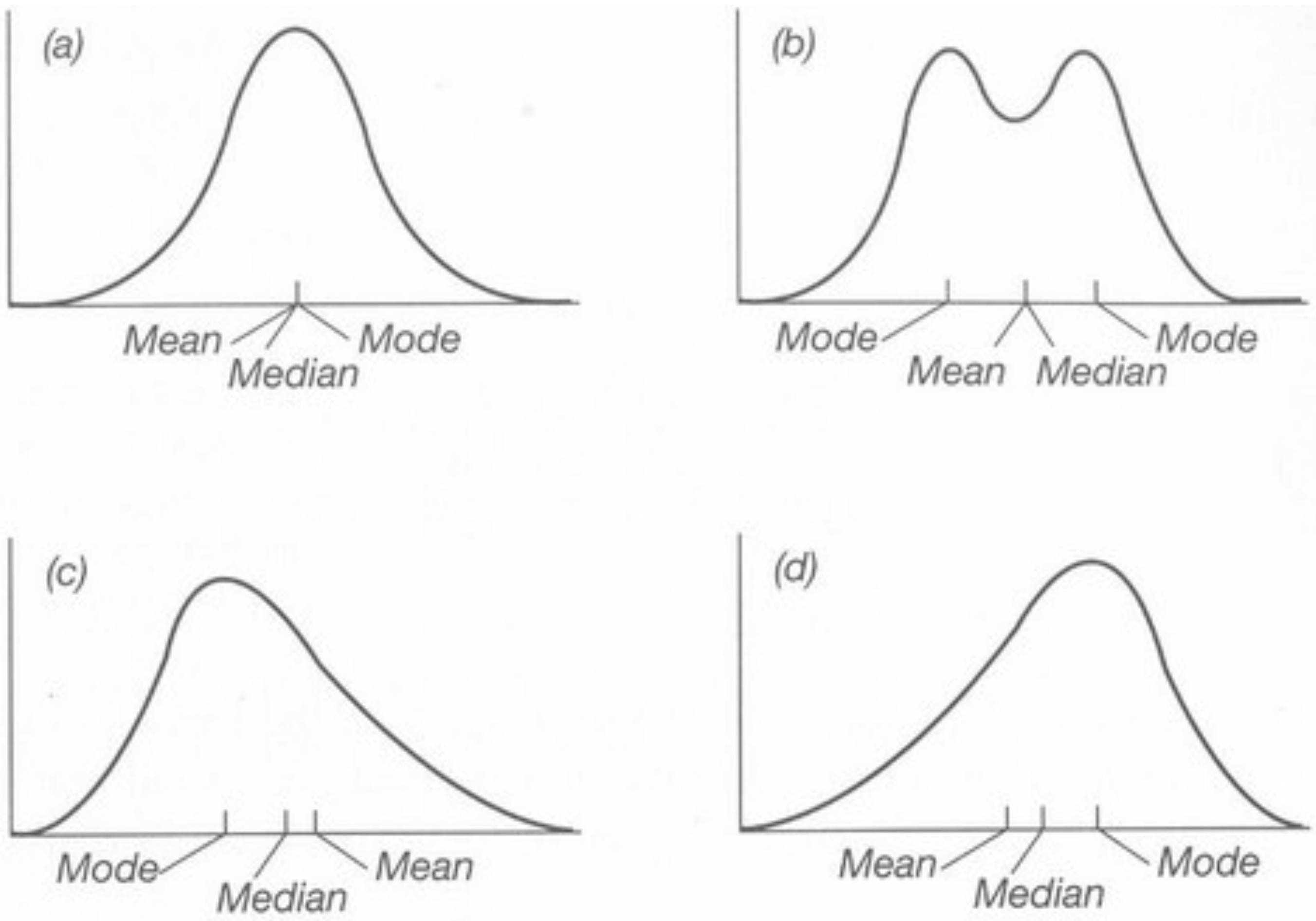
Additional material

Sample statistics

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



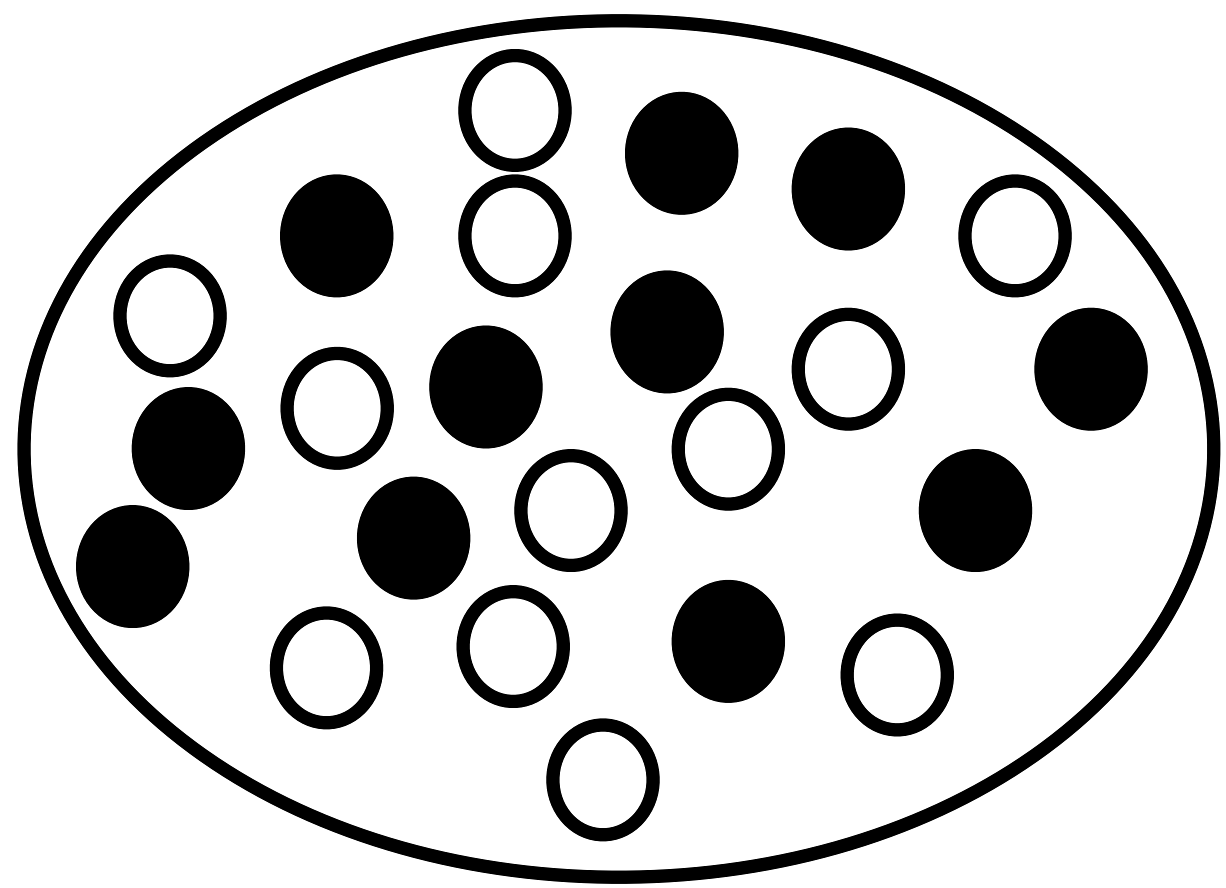
Binomial

An intuitive scheme for deducing statistical distributions is to imagine a sample of otherwise identical N balls belonging to two classes, black and white

Np white balls and Nq black balls, with p+q=1

In a single trial, a ball is selected, the color observed, and then the ball is returned to the bag.

Can do many trials under identical conditions



Binomial (cont'd)

If one repeats a single trial many times, one expect the fraction of trials yielding a white ball to approach Np/N =p.

Consider now pairs of trials: the fraction of trial pairs yielding two white balls approaches (Np/N)*(Np/N) = p^2 . Similarly, the fraction of trial pairs yielding two black balls tends to $q^2 = (1-p)^2$. The fraction of pairs yielding a black and a white (no matter the order) is 2pq= 2p(1-p)

Generalizing to n trials, and taking the probability as a limiting frequency, the probability of j white balls and (n-j) black balls is

probability for a specific sequence of j whites and (n-j) blacks

$$f(j;n,p) = \binom{n}{j} p^j (1-p)^{n-j}$$

number of such sequences

Binomial (cont'd)

Important to understand and remember the conditions to which the model applies: the number n of *identical and independent* trials is fixed.

If I had fixed the number of successes j (that is stopping the experiment after drawing j white balls), I'would have another distribution!

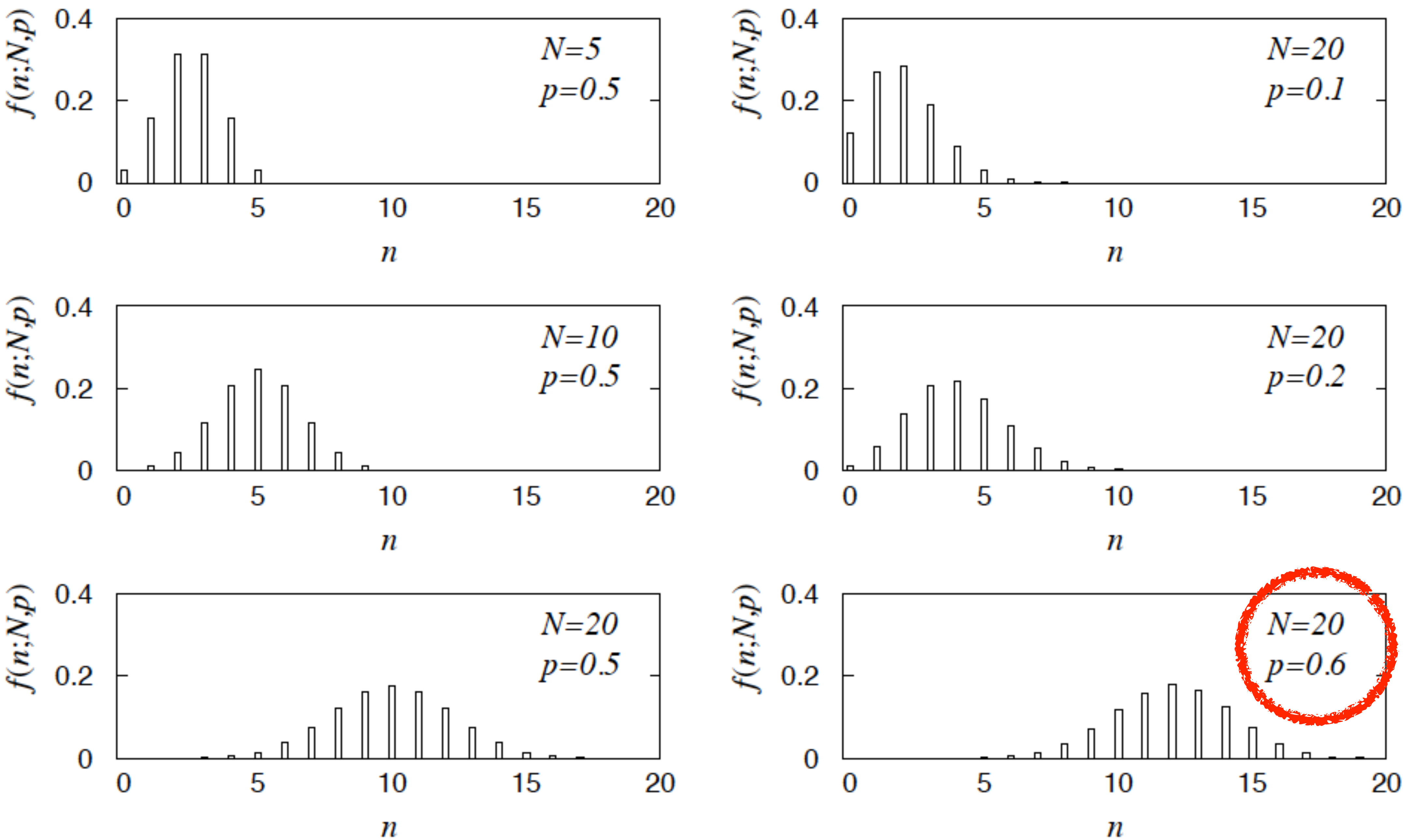
$$f(j; n, p) = \binom{n}{j} p^j (1-p)^{n-j} = \frac{n!}{(n-j)!j!} p^j (1-p)^{n-j}$$

Expectation value $\langle j \rangle = np$

Variance $V(j) = np(1-p)$

Binomial widely used for efficiencies — we'll get back to that.

Binomial (cont'd)



Shape and location of the binomial vary for variation of its two parameters

Poisson

Suppose you don't know the number of trials. You only know that some rare successes can comes out of a continuum of trials. But you know the average rate of success.

Think of lightnings in a thunderstorm.

Poisson

When the proportion of successes p is very small, but sample size n is large enough to maintain n*p appreciable, one gets the Poisson distribution as the limiting form of the binomial distribution

$$n \to \infty, p \to 0, \text{ with finite } np = \mu$$

Simeon D. Poisson (1781-1840)

$$\binom{n}{j} p^j (1-p)^{n-j} = \frac{n!}{(n-j)!j!} \frac{\mu^j}{n^j} \left(1 - \frac{\mu}{n}\right)^{n-j}$$

$$= \frac{\sqrt{2\pi} e^{-n} n^{n+\frac{1}{2}}}{\sqrt{2\pi}(n-j)^{n-j+\frac{1}{2}} e^{-n+j} n^j} \frac{\mu^j}{j!} e^{-\mu}$$

$$= \frac{1}{(1-j/n)^n e^j} \frac{\mu^j}{j!} e^{-\mu} = \frac{\mu^j}{j!} e^{-\mu} = f(j;\mu)$$

Ubiquitous in “counting experiments”: rare process searches, characterisation of counting detectors and so on

Poisson

Expectation value equals variance
"gets broader as it moves right"

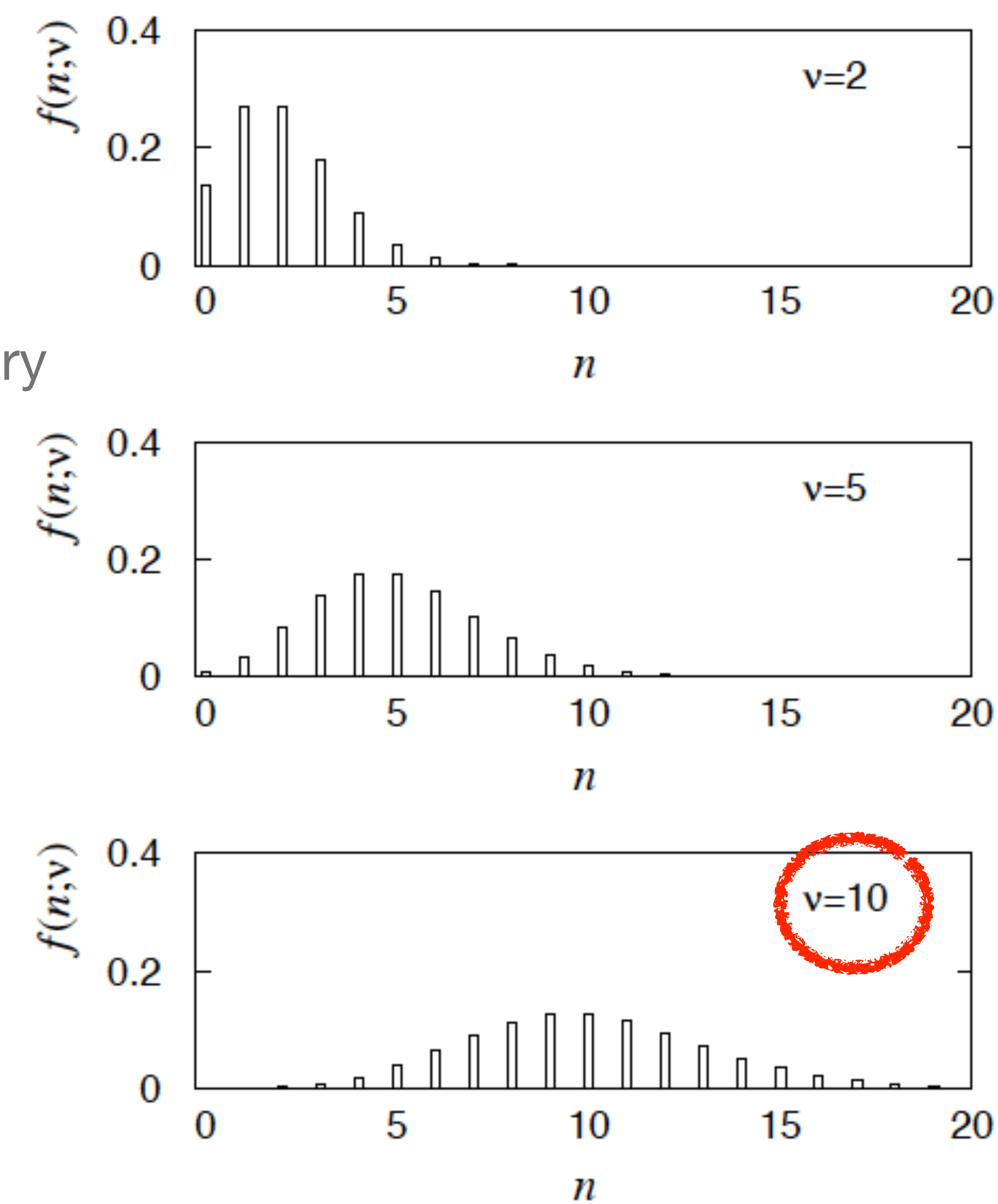
$$\langle j \rangle = V(j) = \mu$$

Shape and location of the Poisson vary for variations of its single parameter

For μ<1, the most probable value is always zero.

For μ>=1 a peak develops, but it is always below μ (which is the mean, not the mode).

For μ integer, j= μ and j= μ-1 are always equally probable.



Limiting relationships btw standard distributions

Binomial

$$f(j;n,p) = \binom{n}{j} p^j (1-p)^{n-j}$$

$n \to \infty, p \to 0, np = \mu$

Poisson

$$f(j;\mu) = \frac{\mu^j}{j!} e^{-\mu}$$

$np \to \mu,$

$\sqrt{np(1-p)} \to \sigma$

$\sqrt{\mu} \to \sigma$

Gaussian

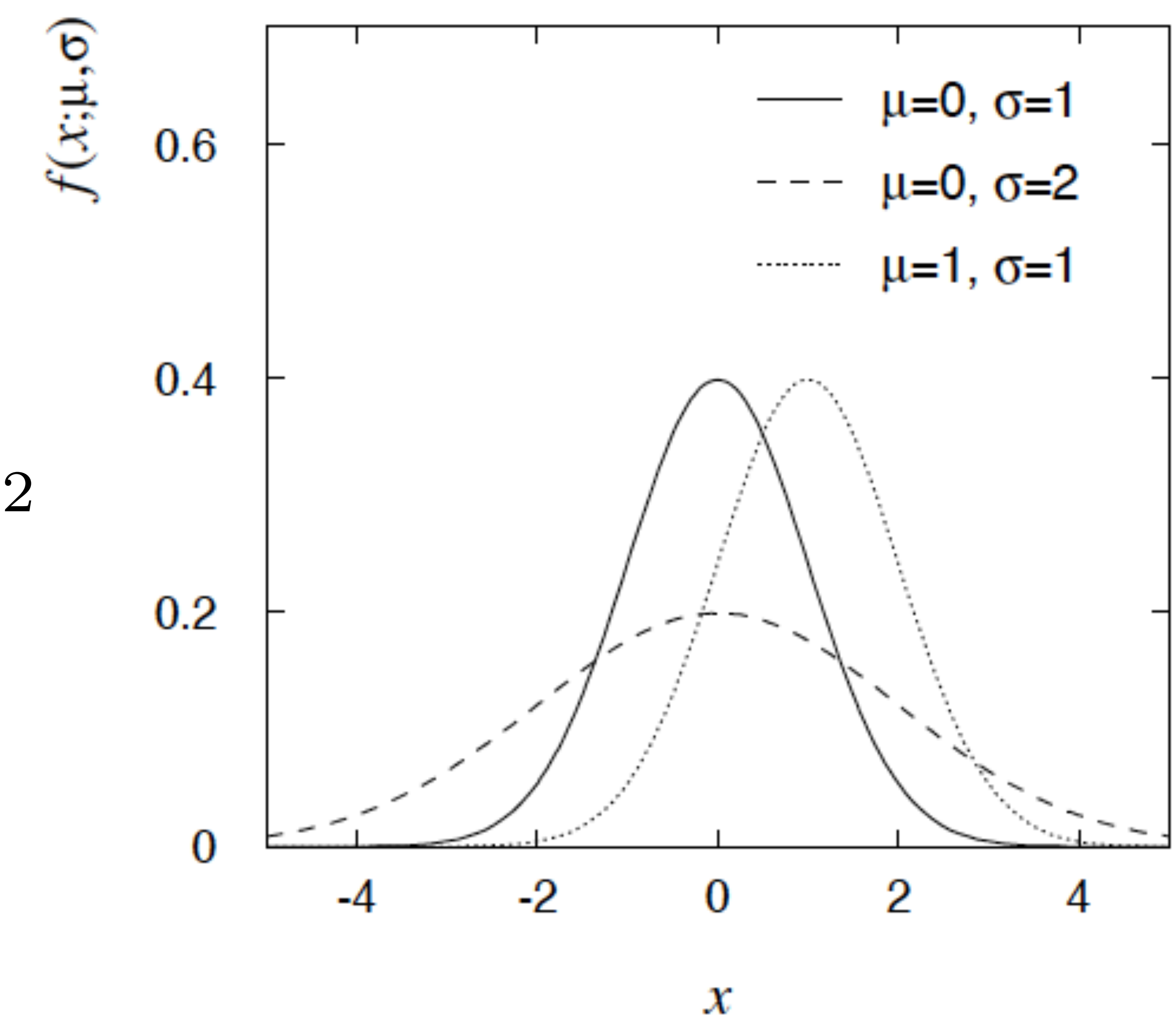
$$f(x;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{(x-\mu)^2}{2\sigma^2}}$$

Normal distribution (or Gaussian, for physicists)

Two parameters

Expectation value $\langle x \rangle = \mu$

Variance $V(x) = \sigma^2$



Normal distribution (or Gaussian, for physicists)

The most important distribution because of its remarkable theoretical properties and regularities and its ubiquitous applications in natural sciences

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

Abraham De Moivre (1667-1754)

The Gaussian distribution frequently approximates well the distributions of many variables commonly encountered in natural sciences, including physics.

Carl F. Gauss (1777-1855)

Not accidental. It results from the central limit theorem: the mean of n independent variables that have arbitrary distributions (each with finite variance) tends to be distributed as a Gaussian centered on the average of the individual means.

Central Limit

- Take the N outcomes x_i of N independent random events
- Each x_i is drawn from its (arbitrary) distribution with mean $< x_i >$ and variance σ^2_i (<u>variance should be finite</u>)

Abraham De Moivre (1667-1754)

Then, the distribution of the sum S of the x_i individual variables is such that

Pierre-Simon Laplace (1749-1827)

1. The expectation value of S is $\Sigma\ x_i$
2. The variance of S is $\Sigma\sigma^2_i$
3. The distribution of S tends to a Gaussian when N → infinity

Aleksandr M. Lyapunov (1857-1918)

Heuristic demonstration

In measurements typically, *many* different, and *independent* sources of random processes contribute to the dispersion of the result of a measured parameter. The central limit theorem ensures that the incoherent superposition of these effects results in a distribution of observations that approximates a Gaussian.

https://www.youtube.com/watch?v=1DTRzPRfu6s

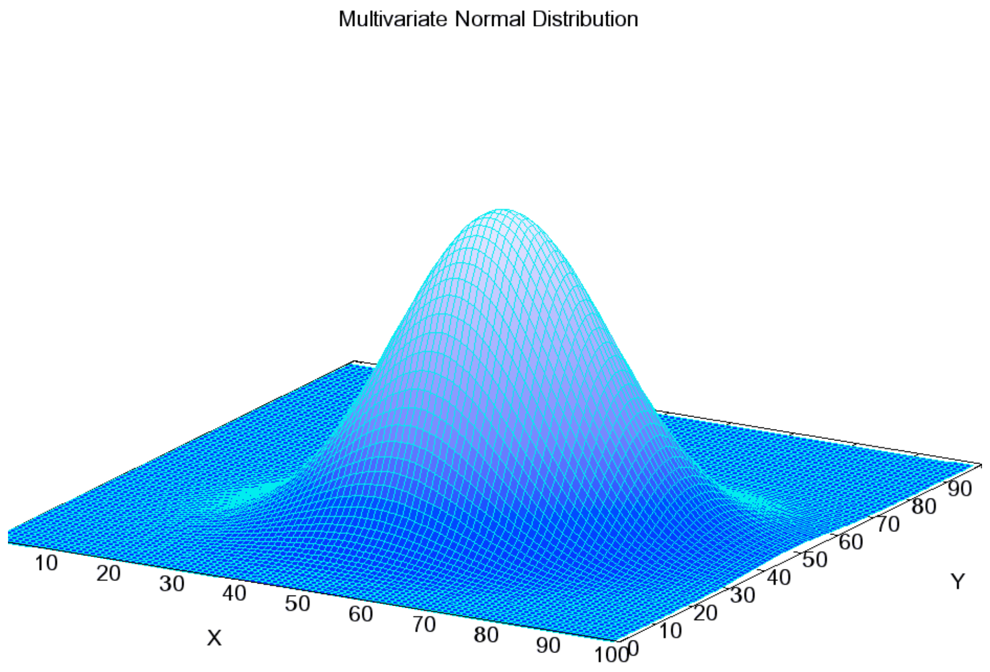
Multidimensional gaussian

$$f(\vec{x};\vec{\mu},V) = \frac{1}{(2\pi)^{n/2}\sqrt{|V|}} exp\left[-\frac{1}{2}(\vec{x}-\vec{\mu})^T V^{-1}(\vec{x}-\vec{\mu}))\right]$$

where $\vec{x}$ and $\vec{\mu}$ are column vectors and $\vec{x}^T$ and $\vec{\mu}^T$ are row vectors

$$E[x_i] = \mu_i$$

$$Cov[x_i, x_j] = V_i j$$



For n=2 (twodimensional Gaussian) this is:

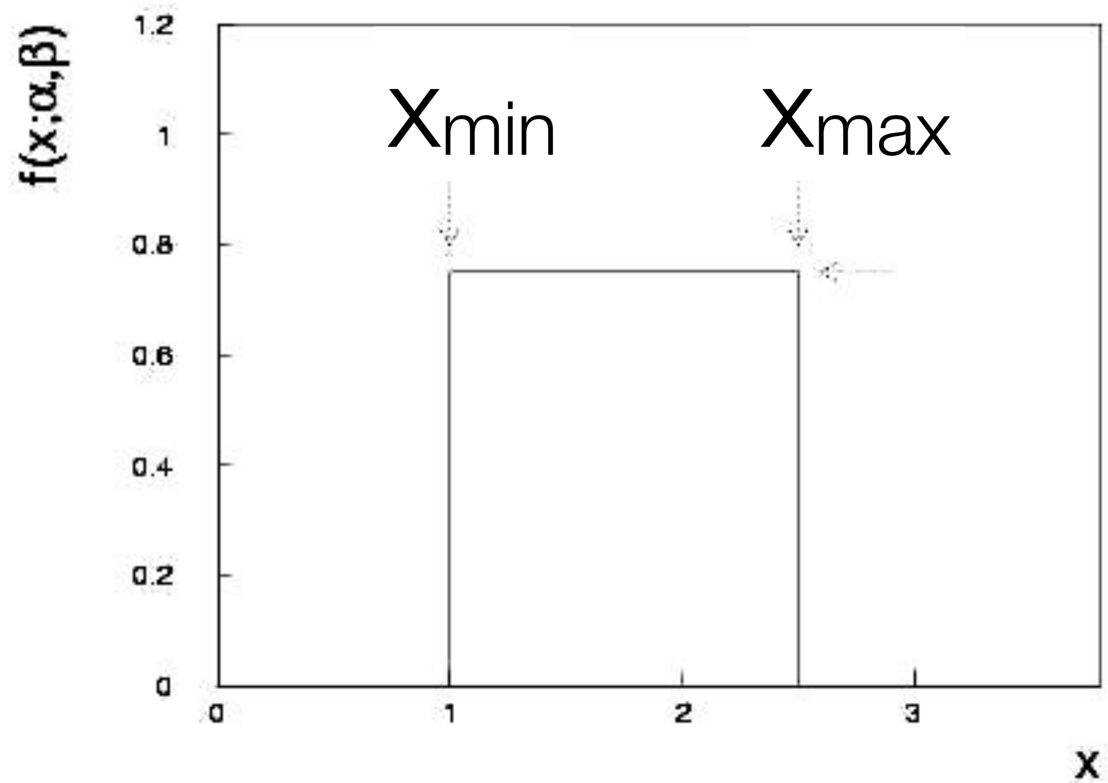
$$f(\vec{x};\vec{\mu},V) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2 + \left(\frac{x_2-\mu_2}{\sigma_2}\right)^2 - 2\rho\left(\frac{x_1-\mu_1}{\sigma_1}\right)\left(\frac{x_2-\mu_2}{\sigma_2}\right)\right]\right\}$$

Uniform distribution

$$f(x; x_{\min}, x_{\max}) = \frac{1}{x_{\max} - x_{\min}}$$

if x is between xmin and xmax.

f=0 otherwise.



$$E[x] = \frac{1}{2}(x_{\min} + x_{\max})$$

$$V[x] = \frac{1}{12}(x_{\max} - x_{\min})^2$$

Example: for H —> γγ, the energy of the photon is uniform in the range [$E_H(1-\beta)/2$, $E_H(1+\beta)/2$]

Exponential distribution

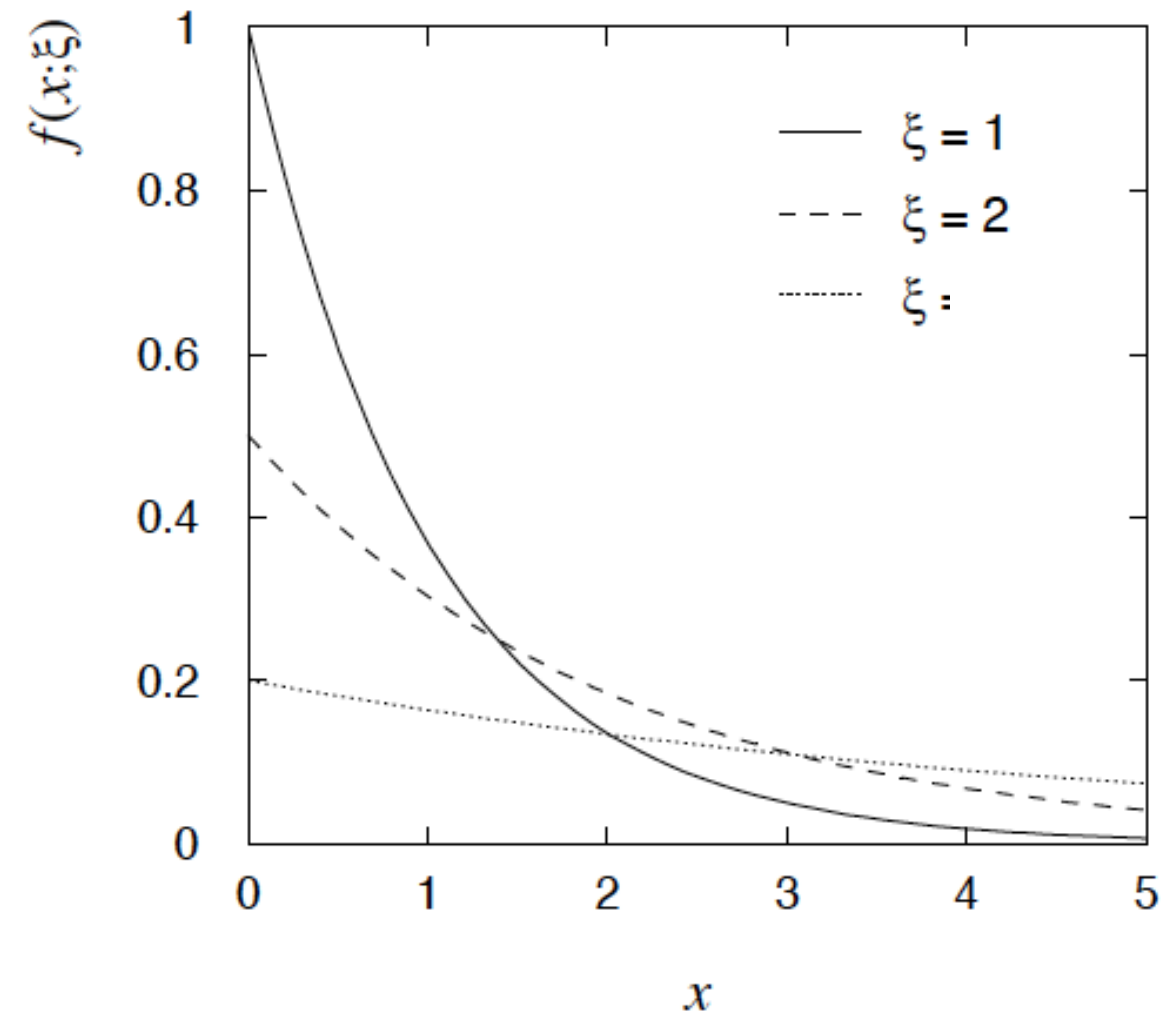
$$f(x;\tau) = \frac{1}{\tau} e^{-x/\tau}$$

if x is nonnegative.

f=0 otherwise.

$$E[x] = \tau$$

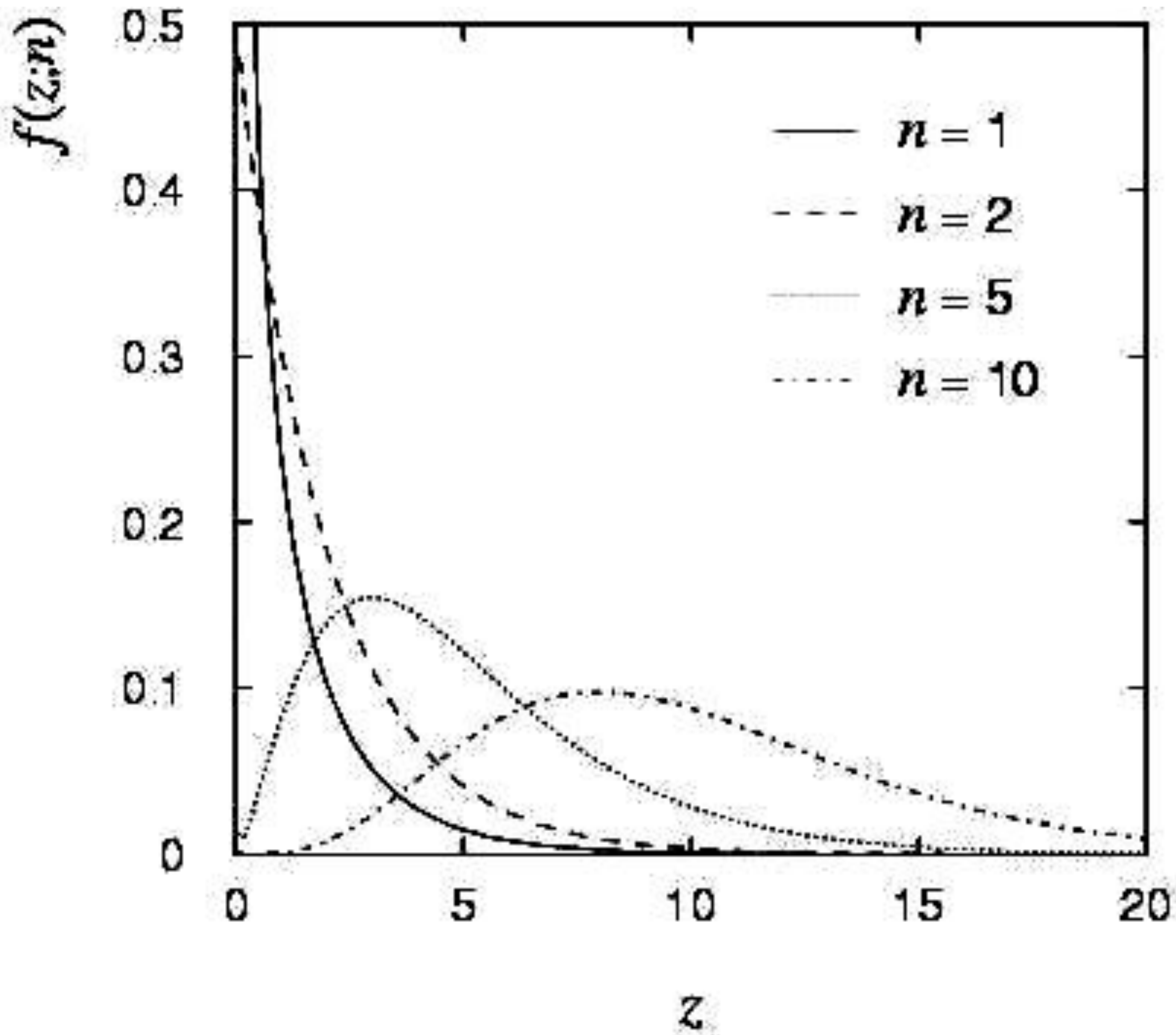
$$V[x] = \tau^2$$



Decay of unstable states

Chi-square distribution

$$f(z;n) = \frac{1}{2^{n/2}\ \Gamma(n/2)} z^{\frac{n}{2}-1} e^{-z/2}$$



if z is nonnegative. It is function of just one parameter, n, which is called the number of degrees of freedom

$$E[z] = n$$

$$V[z] = 2n$$

The χ^2 is the distribution of the sum of the squares of n independent Gaussian discrepancies normalised by the variance.

$$z = \sum_{i=1}^{n} \frac{(x_i - \mu_i)^2}{\sigma_i^2}$$

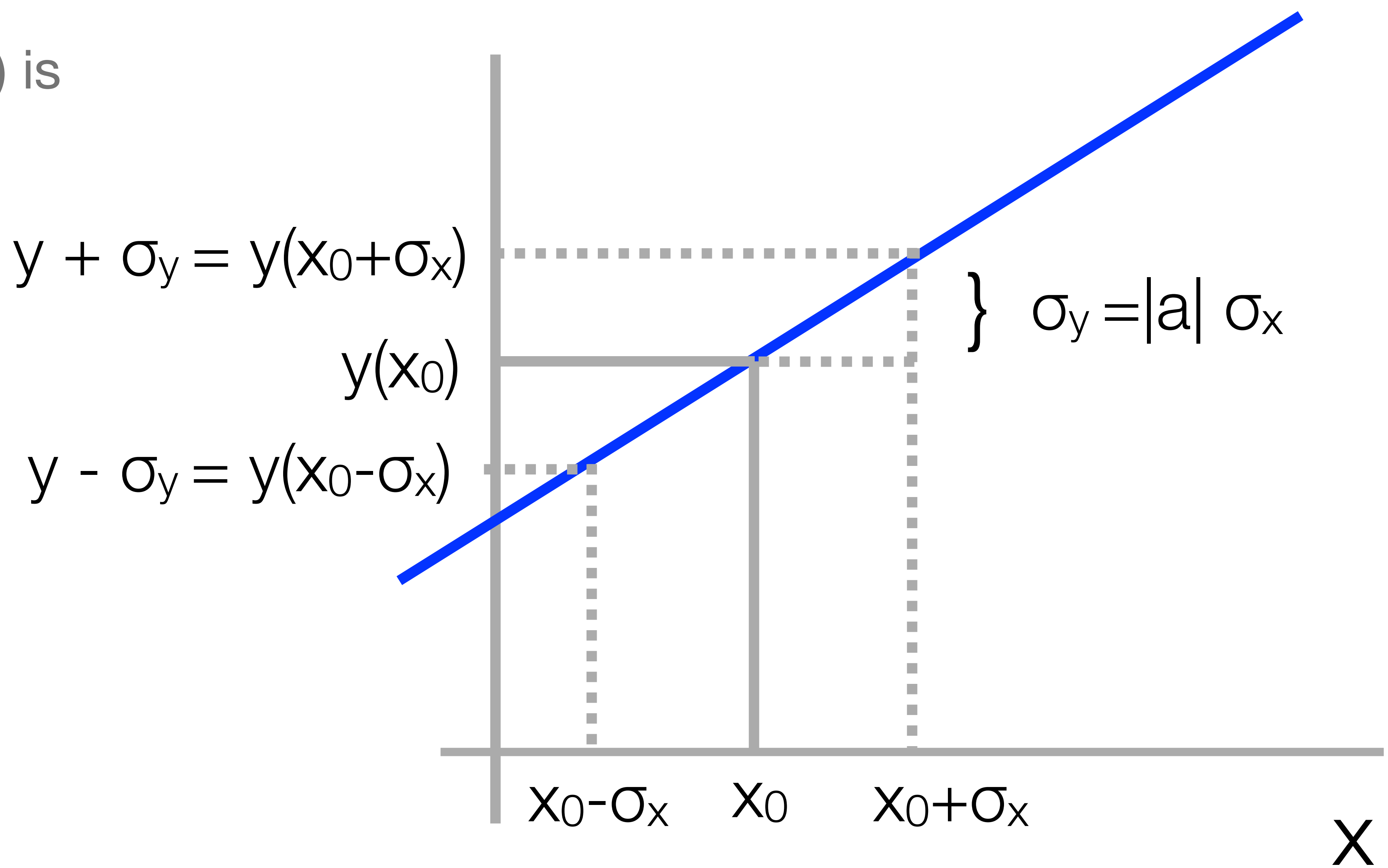
Variances of functions of random variables (a.k.a. “propagation of errors…”)

Often one is interested in knowing the variance of a function of a random variable, given the variance of the random variable.

Linear example: $y(x) = a\,x + b$ with σ_x standard deviation of x.

Standard deviation of y(x) is

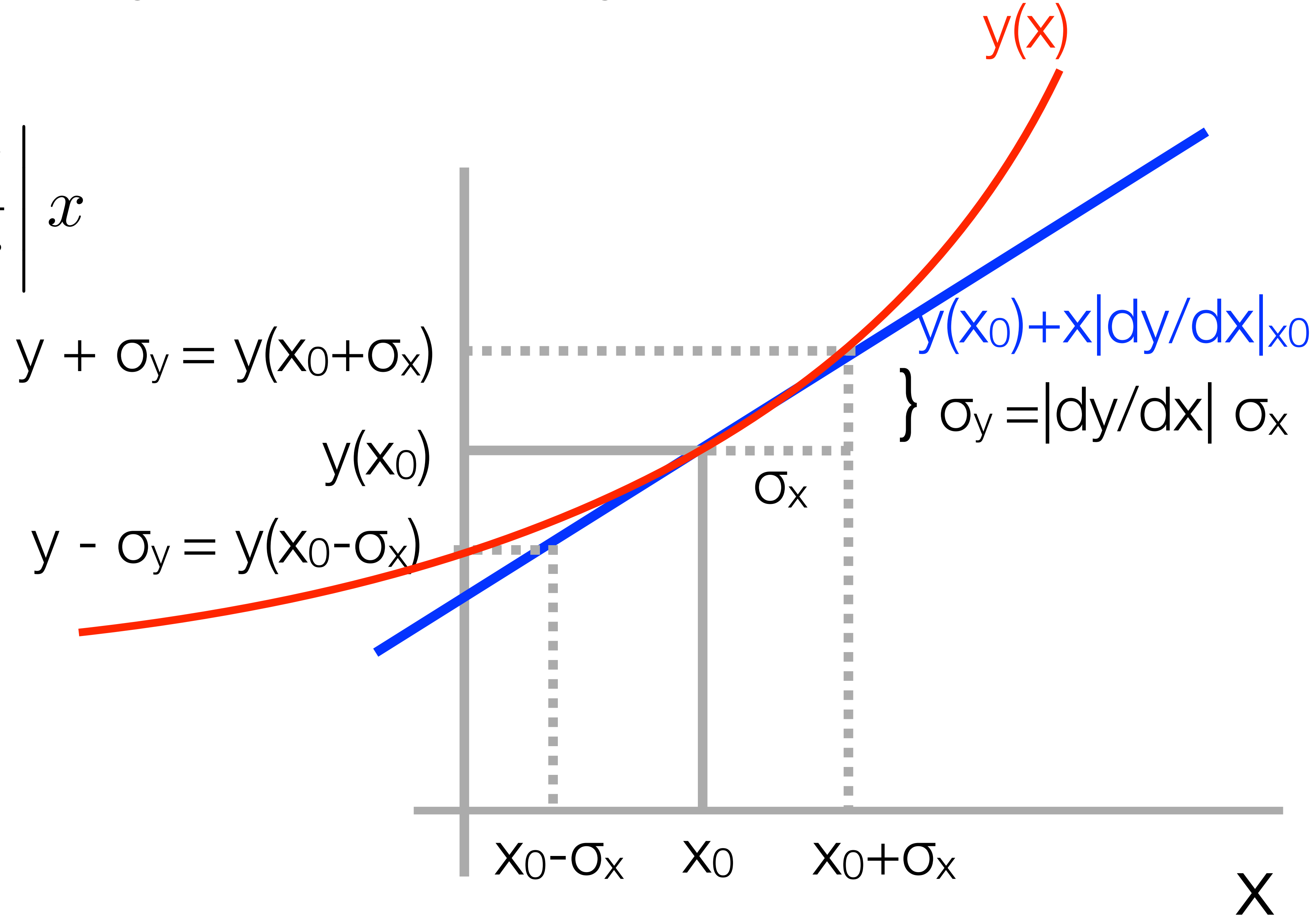
$\sigma_y = |dy/dx|\,\sigma_x$



Variances of functions of random variables (cont'd)

Taylor-linearize any non-linear y(x) that does not vary too much between $x_0-\sigma_x$ and $x_0+\sigma_x$

$$y(x) \approx y(x_0) + \left|\frac{dy}{dx}\right| x$$



Variances of functions of random variables (1D)

$$y(x) \approx y(x_0) + \left|\frac{dy}{dx}\right| x$$

$$V(y) = \langle y^2(x)\rangle - \langle y(x)\rangle^2$$ Definition of variance

$$\approx \langle (y(x_0) + x\frac{dy}{dx})^2 - \langle y(x_0) + x\frac{dy}{dx}\rangle^2$$ Replace with linearization

$$= \left(\frac{dy}{dx}\right)^2 \left(\langle x^2\rangle - \langle x\rangle^2\right)$$ Do the algebra

$$= \left(\frac{dy}{dx}\right)^2 V(x)$$

Variances of functions of random variables

Extend to functions of 2 to n variables.

$$y(x_1, x_2) \approx y(x_{1,0}, x_{2,0}) + \left|\frac{\partial y}{\partial x_1}\right|_{x_{1,0}} x_1 + \left|\frac{\partial y}{\partial x_2}\right|_{x_{2,0}} x_2$$

$$V(y) = \langle y^2 \rangle - \langle y \rangle^2$$

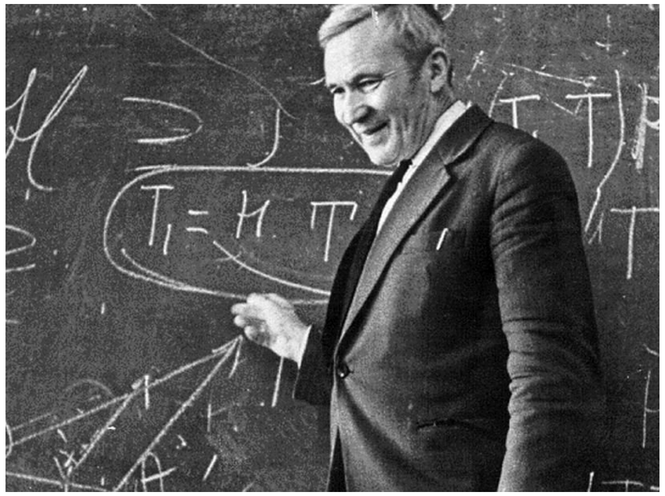
$$\approx \left|\frac{\partial y}{\partial x_1}\right|^2_{x_{1,0}} V(x_1) + \left|\frac{\partial y}{\partial x_2}\right|^2_{x_{2,0}} V(x_2) + 2\left|\frac{\partial y}{\partial x_1}\right|\left|\frac{\partial y}{\partial x_2}\right| Cov(x_1, x_2)$$

1. linearized formulas are exact only if $y(\vec{x})$ is linear. They fail if the function is nonlinear over a range comparable in size to σ_{xi}

2. linearized formulas apply for any pdf of the x_i variables.

Set-theoretical axioms of probability

Define the set Ω of all the possible mutually exclusive outcomes of a statistical experiment (sample space). An event A is a set containing one or more elementary outcomes.



Andrey N. Kolmogorov (1903-1987)

Assume that probability P is an additive function on the set and it is measurable on a continuous scale so that it can be represented by a real number. Then

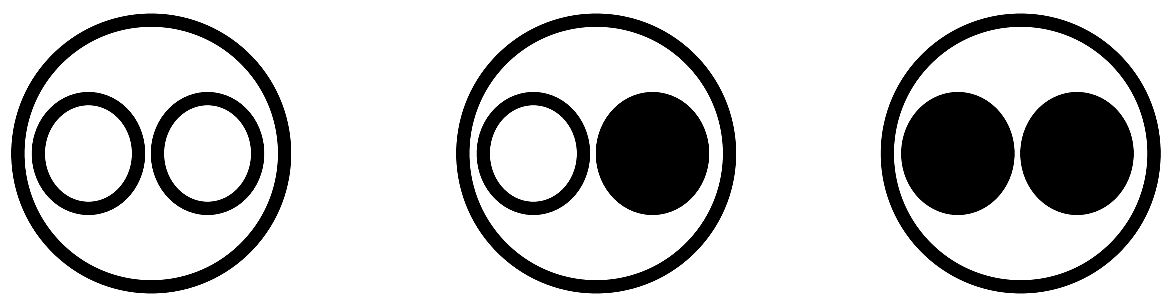
1. P(A) is nonnegative for each possible outcome A.

2. The sum of probabilities over all the possible outcomes (sample space Ω) is unity, $P(\Omega) = 1$.

3. The probability for observing outcome A or outcome B is P(A)+P(B) if A and B are disjoint sets

Inference — elementary example

- Three identical bags with two balls each. Each ball can be black or white

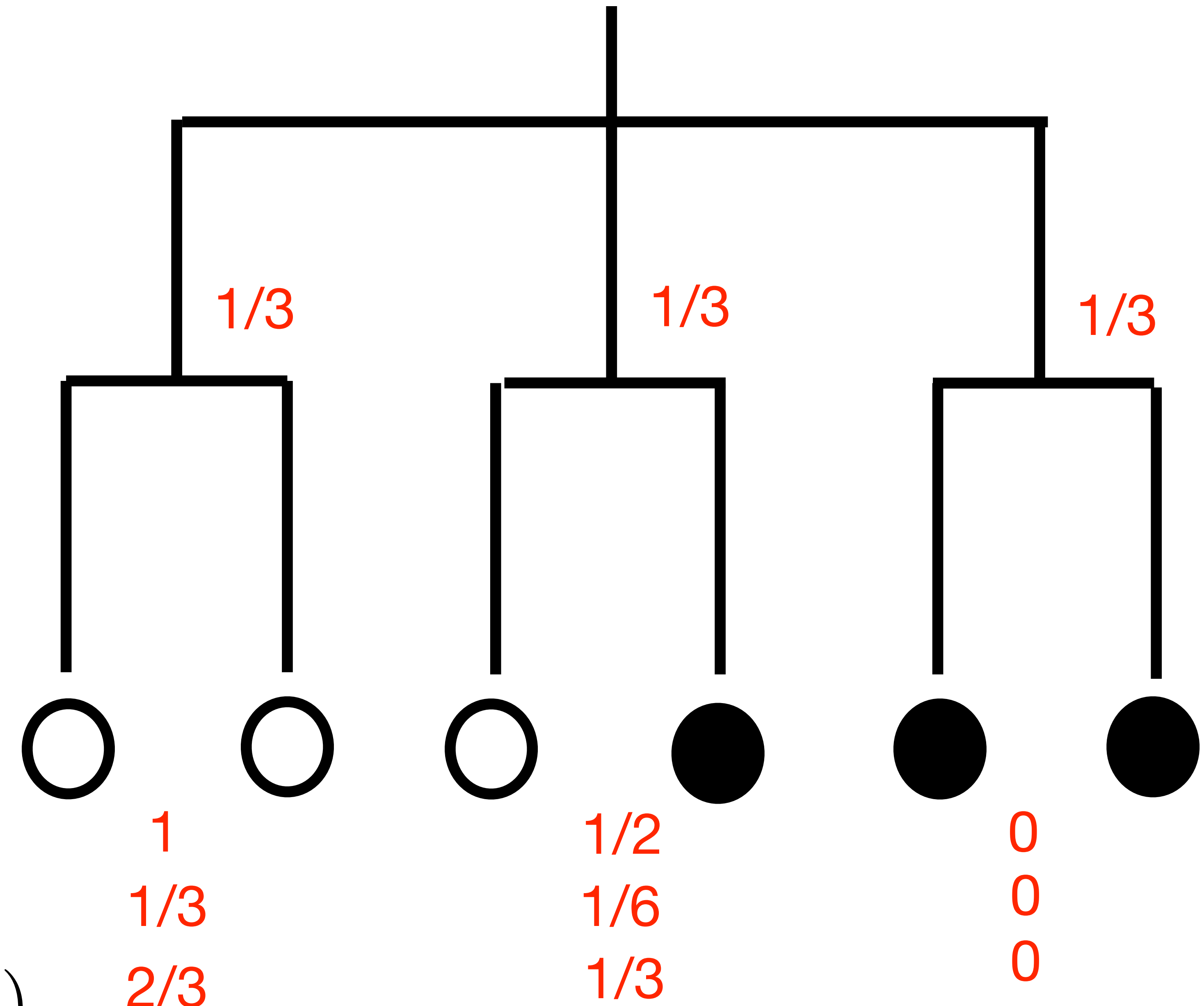


- Pick a random bag (m, unobservable) and a random ball inside it (x, observable)
- Ball is white (x=w). What can one say about the chosen bag?

Want to know p(m|w), the probability I picked each bag, given that the ball is white.

Inference — elementary example

$p(m)$



$$p(w|m)$$

$$p(w, m) = p(w|m)\ p(m)$$

$$p(m|w) = p(w|m)\ p(m)/p(w)$$

Most probably (66%) I picked the bag with two white balls. Pretty obvious. Less intuitive if the proportions between bags are uneven.

Classic properties of estimators

- Consistency (in probability). Desirable that the estimator e(x) of m converges in probability to m

$$\forall\delta \lim_{N\to\infty} p(|m - e(x)| > \delta) = 0$$

- Precision. Desirable that the variance of the estimator is minimal

$$V(e(x)) = \langle |e(x) - \langle e(x)\rangle|^2 \rangle$$

- Bias. Desirable that the estimator is unbiased (b(m)=0)

$$b(m) = \langle\ e(x) - m\ \rangle$$

- Distribution. Desirable that the distribution p(e(x); m) of the estimator is simple (possibly Gaussian)

Comments — bias

Many estimators suffer from biases, which, in general depend on the parameter m being estimated. For an estimator e(x) of m, the bias b(m) is defined from

$$E[e(x)] = \langle\ e(x)\ \rangle = m + b(m)$$

Typically biases are small wrt the variance. Issues, however, arise in combinations of biased estimates: the variance reduces but the bias remains and weights more.

- If the distribution p(x|m) is known, the bias can be calculated explicitly.
- If the bias is independent of m (b(m) = b) then use another estimator u(x) = e(x) - b, which is unbiased and has same precision (variance) of e(x).
- If the bias depend on m, need an unbiased estimator of b (B(x)) to redefine u(x) = e(x) - B(x). The new estimator has greater variance than e(x), but loss in precision is often smaller than bias.

Example — bias correction w/ known distribution

I have N points x_i distributed as a Gaussian and use the following ML estimator to estimate its variance

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

This estimator has a bias $b = -\sigma^2/N$

and a variance $\mathrm{Var}(\hat{\sigma}^2) = 2\sigma^4 \frac{N-1}{N^2}$

So, I can rework an alternative estimator $s^2 = \frac{1}{N-1} \sum_{x_i=1}^{N} (x_i - \overline{x})^2$

which has zero bias and a variance $\mathrm{Var}(s^2) = 2\sigma^4 \frac{1}{N-1}$ which is only $1/N^2$ larger than that of the previous estimator

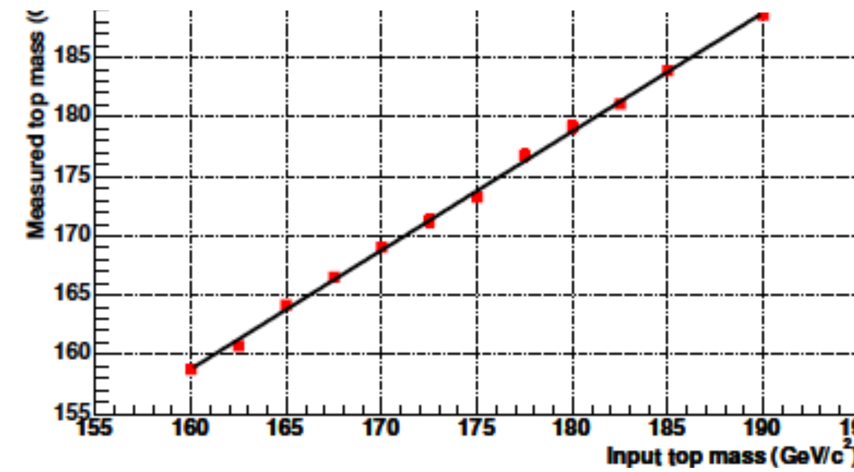
Example — biases w/ unknown distributions

In most practical cases, p(x|m) is not well known, o the bias is hard to calculate explicitly.

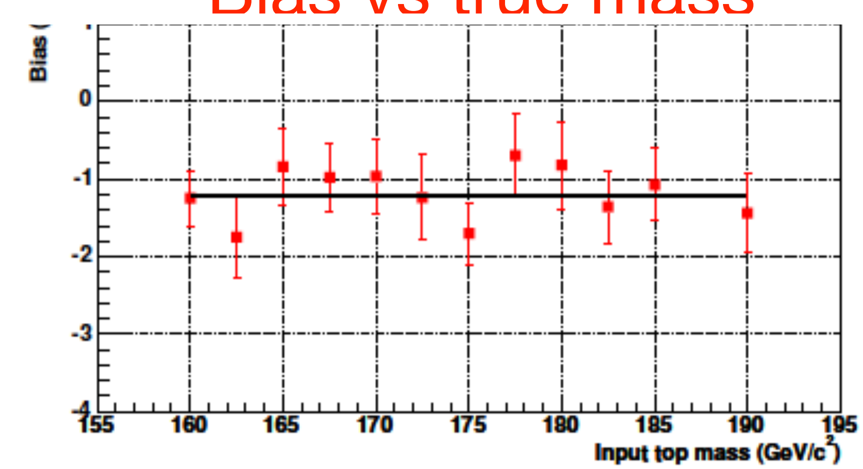
Biases are studied by repeating the measurement c simulated samples and comparing results with inpu "true" values or applying the estimator in control samples for which results are known.

If deviations ≥ O(variance) occur , correcting the results of the measurement by subtracting the bias is dangerous. Need confidence that simulated experiments reproduce all features of the data (but then also the source of the bias could probably be with identified and removed)

Estimated mass vs true mass



Bias vs true mass



2007 measurement of lepton+jets top-quark mass by CDF