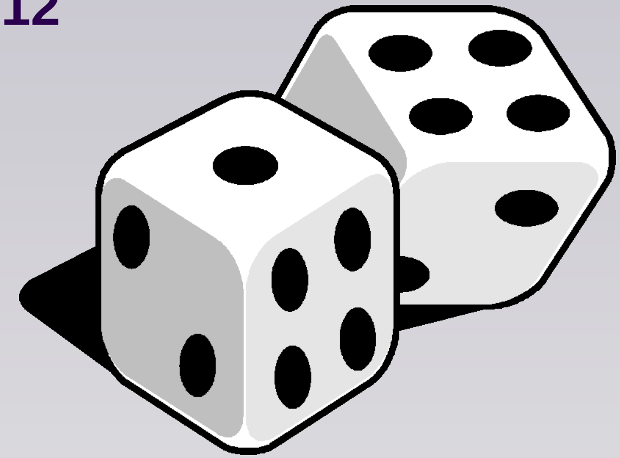
# Introduction to Statistics

CERN Summer Student Lecture Program 2012



## Helge Voss

## … and Machine Learning

(in the last lecture)

# Outline

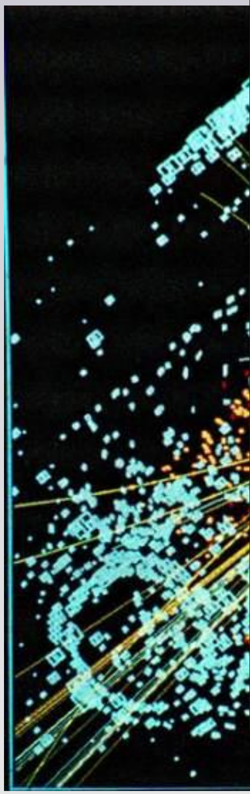
- **Why Statistics**
  - **measurements etc…**
  - **review of (some) probability distributions and some of their properties**
- **What is Probability :**
  - **axioms**
  - **frequentist / Bayesian  interpretation**
- **Lecture 2-4**
  - **Hypothesis testing**
  - **Maximum Likelihood fit**
  - **Confidence belts**
  - **Monte Carlo Methods (Random numbers/Integration/Re-sampling)**
  - **Machine Learning / Pattern Recognition**

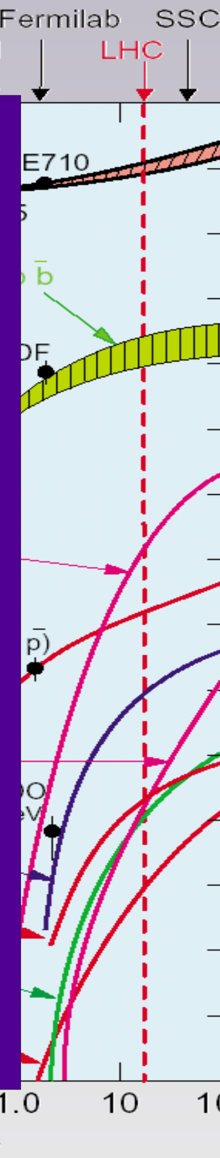
# HEP Experiments

## “typical” Higgs event (CMS simulation):

→ ‘hidden

→ and ran





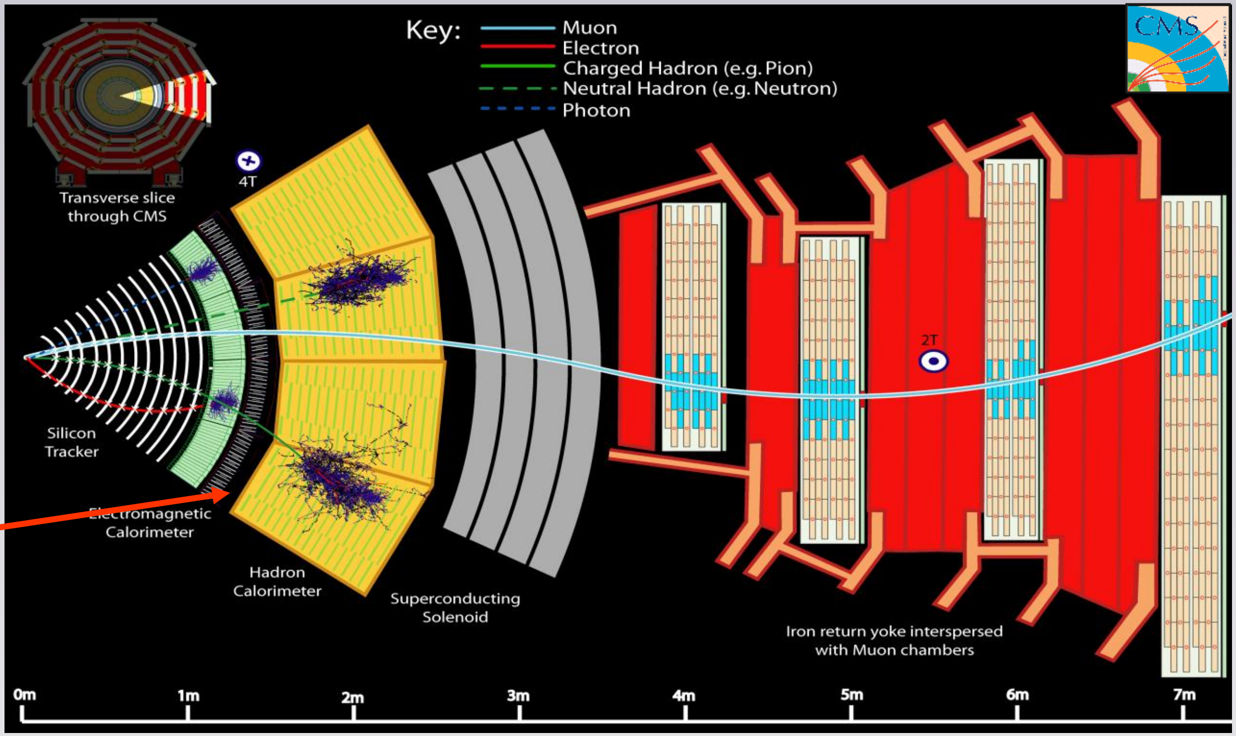
# HEP Experiments

**And while a the needle in the hay-stack would be already in one piece**

**→ particles: reconstructed from its decay products**

**→ decay products: reconstructed from detector signatures**

**→ etc..**

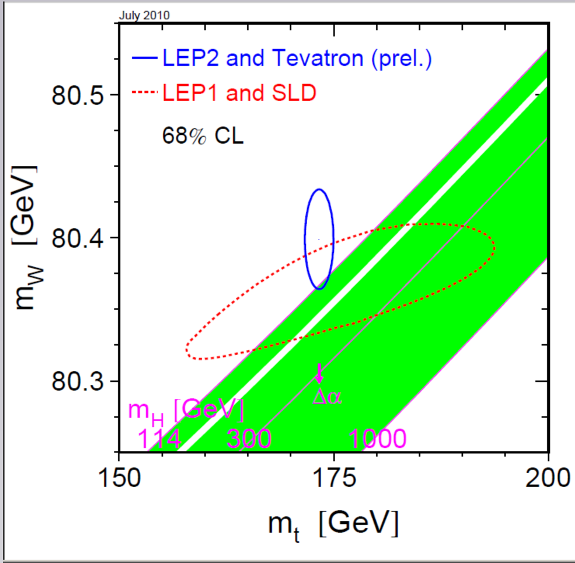



**"statistical/random" processes**

# Interpreting your Measurement

- What do we REALLY mean by:
  - $m_W$ = 80.399+/- 0.023 ;
  - $M_{Higgs}$> 114.4GeV/c$^2$ @95%CL

  (and… how do others "interpret" this?)




- these things are results of:
  - involved measurements
  - many "assumptions"/"Interpretations"

- correct statistical interpretation:

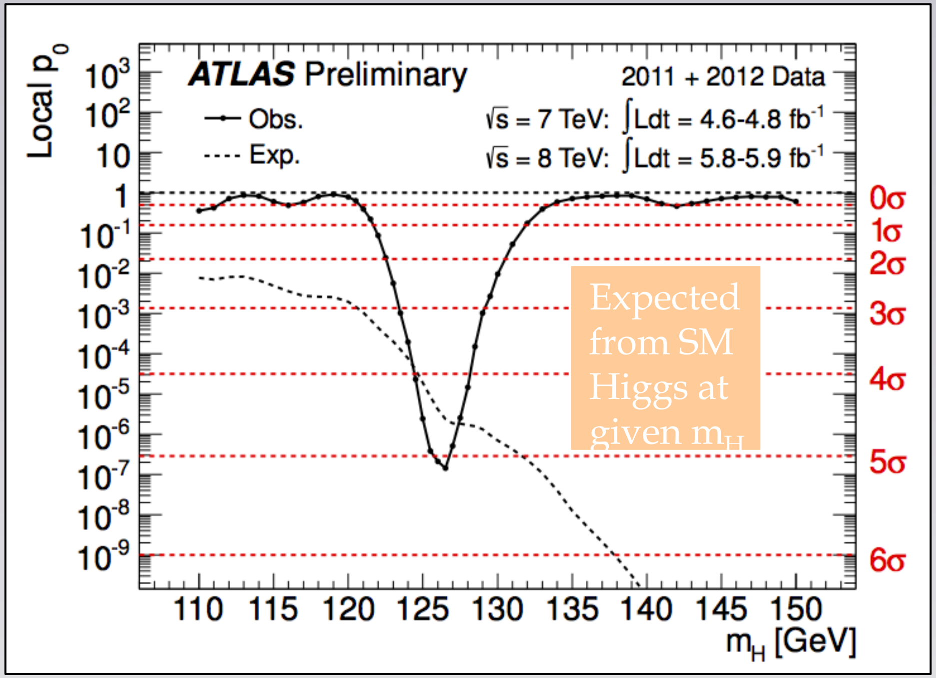
→ most 'honest' presentation of the result

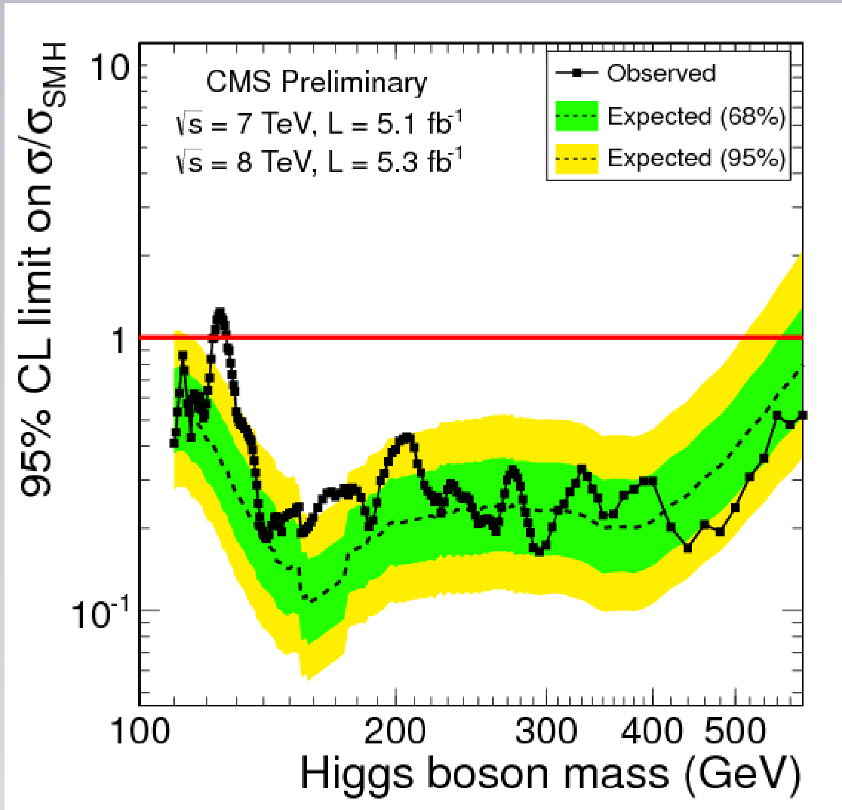
→ unless: provide <u>all</u> details/assumptions that went into obtaining the results

→ needed to correctly combine with others (unless we do a fully combined analysis)

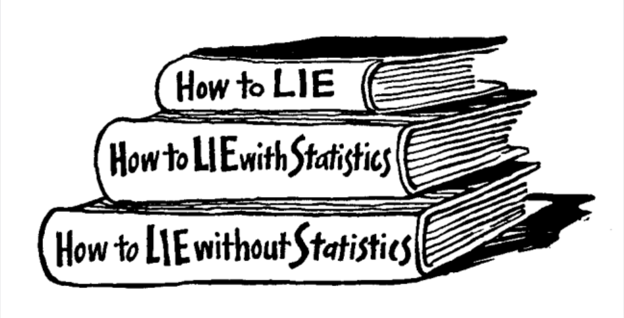
# Interpreting Measurements

- Be able to <u>understand</u> the latest Higgs search plots!



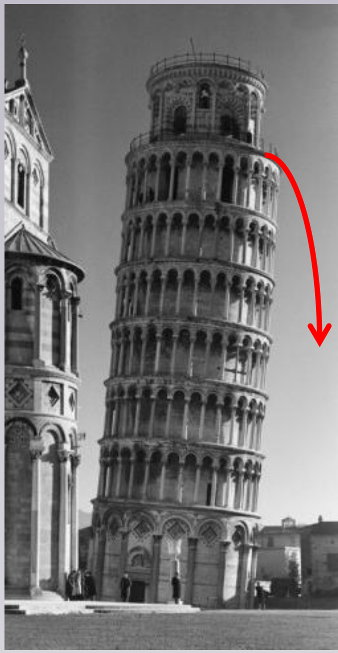




- rather than being fooled by "statistics"



# Why Bother with Statistics?

- **Physics laws have exact numbers:** $F = m \cdot a$
  - **derived from "non-exact" measurements**
  - **"non exact" ⇔ statistically distributed**
  - **know how to handle samples drawn from distributions**
    - extract parameters of underlying (parent) distribution (i.e. mean value etc..)
    - know what they describe → choose the right one ☺
      - e.g. Poisson ↔ Compound Poisson



**Statistics plays important role in:**

- **Measurement errors**
- **Random processes (quantum physics, statistical physics)**
- **Fitting of model parameters**
- **Deciding on model hypothesis/data selection**
  - **Judging significance of some "New Physics" signal**
- **Monte Carlo simulation/integration**

# Statistical Distributions

- **Measurements/Results typically follow some probability distribution**
  - **i.e. data is not at a fixed value, but "spreads out" in a particular way**

- **Which type of distribution it follows depends on the particular case**
  - **important to know the different distributions**
    - be able to pick the correct one when doing the analysis
  - **.. and know their characteristics**
    - be able to extract the "information" in the data

**Note: in statistical context:**

**instead of "data" that follows a distribution,**

**one often (typically) speaks of a "random variable"**

# Probability Distribution/Density of a Random Variable

**random variable $x$ or $k$ :** characteristic quantity of point in sample space

**discrete variables**

$$P(k_i) = p_i$$

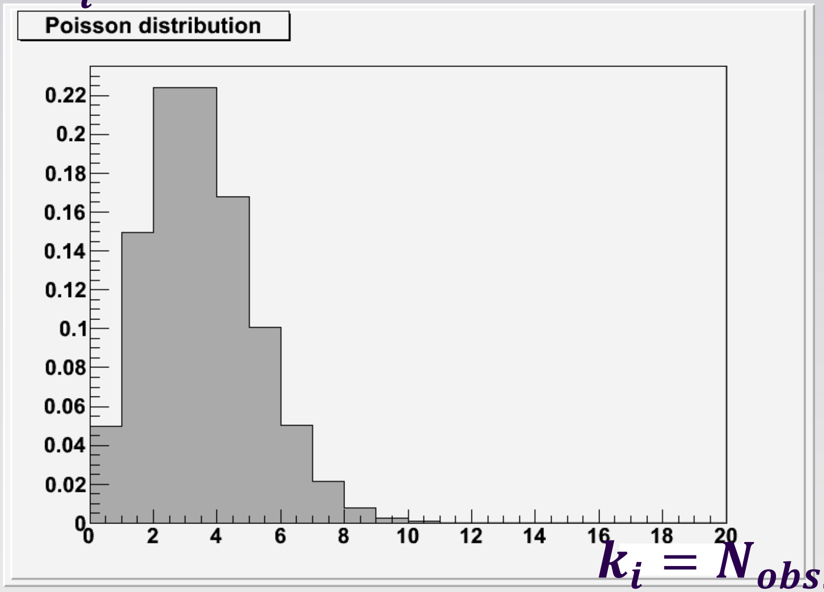
**continuous variables**

$$P(x \in [x, x+dx]) = p(x)dx$$

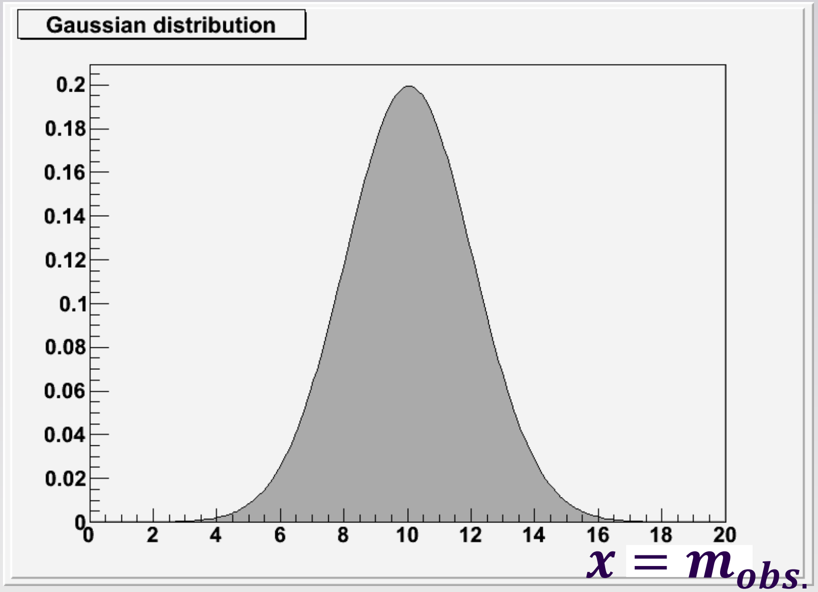
**normalisation** (your parameter/event space covers all possibilities)

$$\sum_i P(k_i) = 1$$

$$\int_{-\infty}^{\infty} p(x)dx = 1$$




$k_i = N_{obs.}$

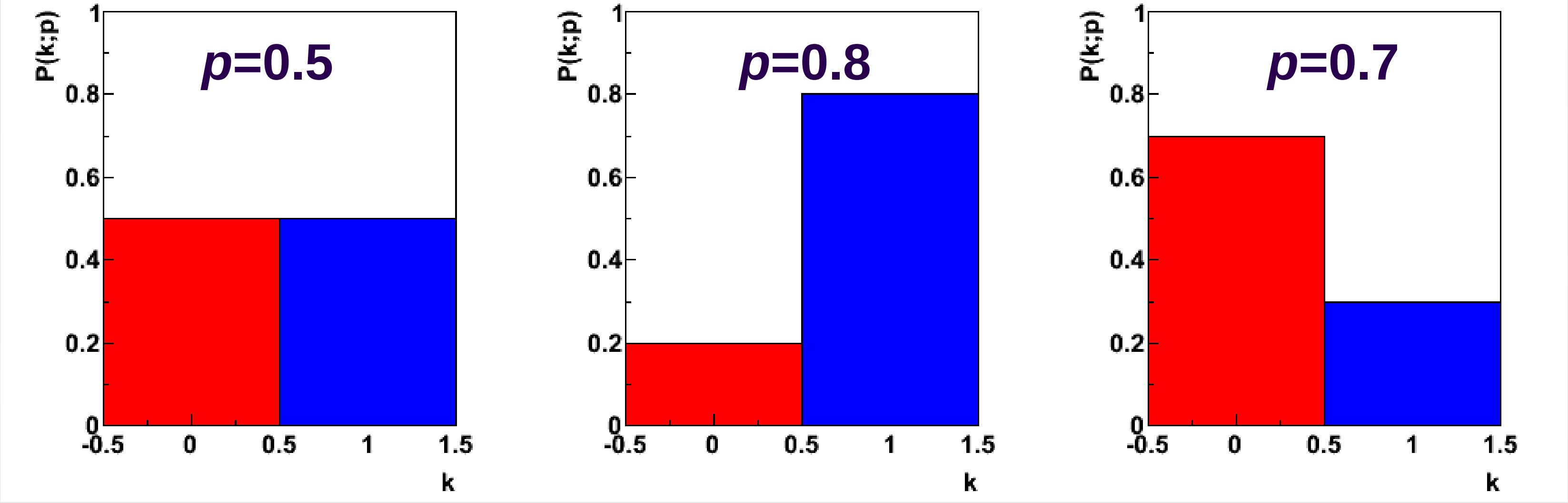



$x = m_{obs.}$

# Bernoulli Distribution

- 2 possible outcomes:
  - Yes/No
  - Head/Tail
  - ….
- (fair) coin: $P(head) = p \left(e.g. = \frac{1}{2}\right), \; P(tail) = 1 - P(head) = 1 - p$

$$P(k;p) = \begin{cases} p & : k = head = 1 \\ 1-p & : k = tail \quad = 0 \end{cases} = p^k (1-p)^{1-k}$$



# Binomial Distribution

**throw $N$ coins:** (anything with two different possible outcomes)

→? **how likely (often):** $k \times head$ **and** $(N-k) \times tail$ **?**

- **each coin:** $P(head) = p, \quad P(tail) = 1 - p$
- **pick $k$ particular coins → the probability of all having $head$ is:**

$$P(k \times head) = P(head) * P(head) \ldots * P(head) = P(head)^k$$

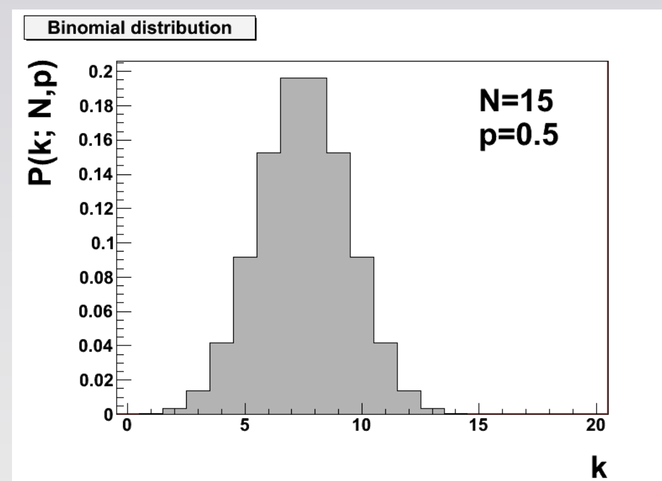
- **at the same time: probability that all remaining N-1 coins land on $tail$**

$$P(head)^k \, P(tail)^{N-k} = p^k \, (1-p)^{N-k}$$

- **That was for $k$ particular coins:**

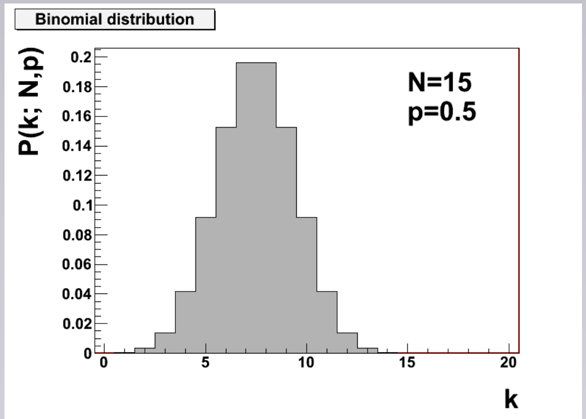
$\binom{N}{k}$ **possible permutations for any $k$ coins**

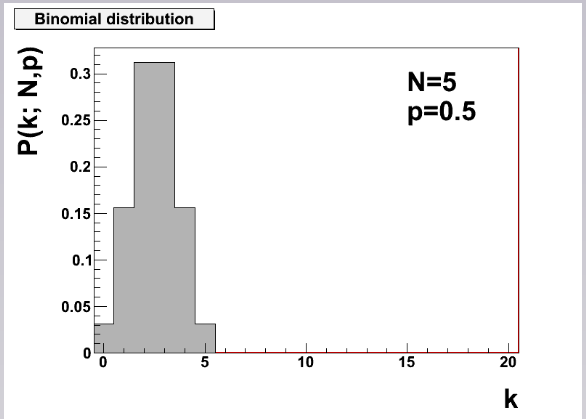
$$P(k; N, p) = p^k \, (1-p)^{N-k} \binom{N}{k}$$

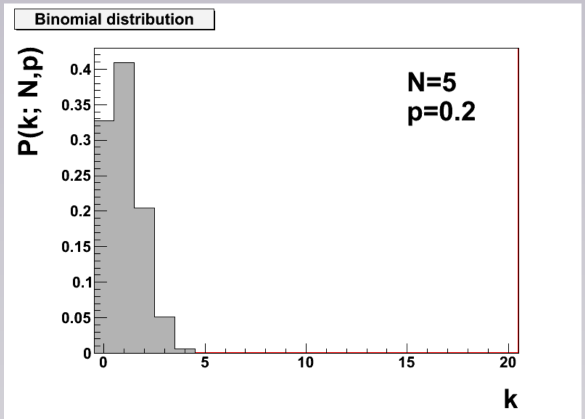


# Binomial Distribution

## Examples:




- **Expectation value: sum over all possible outcomes and "average"**
  - $\mathrm{E}[k] = \sum kP(k) = Np$
- **Variance:**
  - $V(k) = Np(1-p)$

# Some Characteristic Quantities of Distributions

**discrete variables**

**continuous variables**

- **Expectation value $E$ (mean value):**

$$E = \langle k \rangle = \sum_{all\ k} k P(k)$$

$$E[x] = \langle x \rangle = \int x P(x) dx$$

  - **Note: mean/expectation of $f(x)$:** $\rightarrow E[f(x)] = \int f(x) P(x) dx$

- **Variance ($V = \sigma^2$, with $\sigma$: "spread"):** $E\left[(x - \langle x \rangle)^2\right] = E\left[x^2\right] - (E[x])^2$

$$V(k) = \sum_{all\ k} (k - \langle k \rangle)^2 P(k)$$

$$V(x) = \int (x - \langle x \rangle)^2 P(x) dx$$

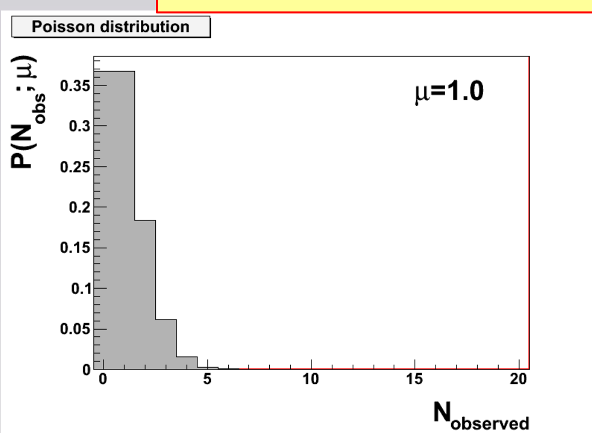
- **higher moments: Skew:** $E[(x - \langle x \rangle)^3]$ **....**
- **Note: expectation and variance → properties of the full population. Unbiased estimates, derived from samples taken from the distribution:**

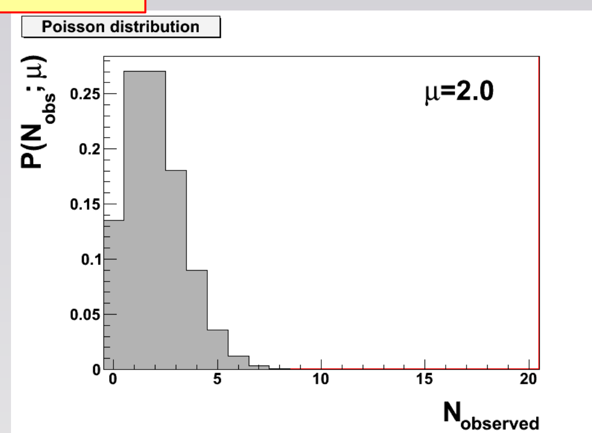
$$\hat{V} = \frac{1}{n-1} \sum_i^{samples} \left(k_i - \bar{k}\right)^2$$

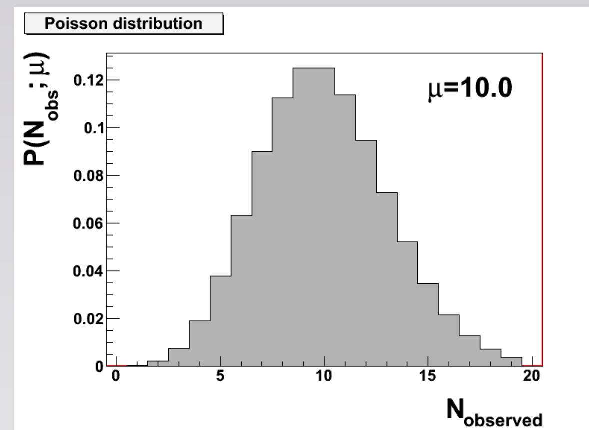
$$\hat{V} = \frac{1}{n-1} \sum_i^{samples} (x_i - \bar{x})^2$$

# Poisson Distribution

- Binomial distribution: Individual events with 2 possible outcomes
- How about: # counts in radioactive decays during $\Delta t$ ?
  - → events happen "randomly" but there is no 2nd
  - → $\Delta$t: continuum ≠ "N- discrete trials"
- $\mu$ : average #counts in $\Delta$t. What's the probability for $n$ counts?
- Limit of Binomial distribution for $N \to \infty$ $with\ Np = \mu\ fixed$
  - → Poisson $P(n) = \frac{\mu^n}{n!} e^{-\mu}$




- Expectation value:

$$E[n] = \sum nP(n) = \mu$$

- Variance:

$$V(n) = \mu$$

b.t.w. it's a good approximation of Binomials for $N \gg Np = \mu$

# Gaussian Distribution

- For large $\mu$ the Poisson distribution already looked fairly "Gaussian"
  - in fact in the limit it "becomes" Gaussian
    - just like almost everything: **Central Limit Theorem**

  → Gaussian is the probably the most important distribution

$$\text{Gauss: } P(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

- Expectation value:

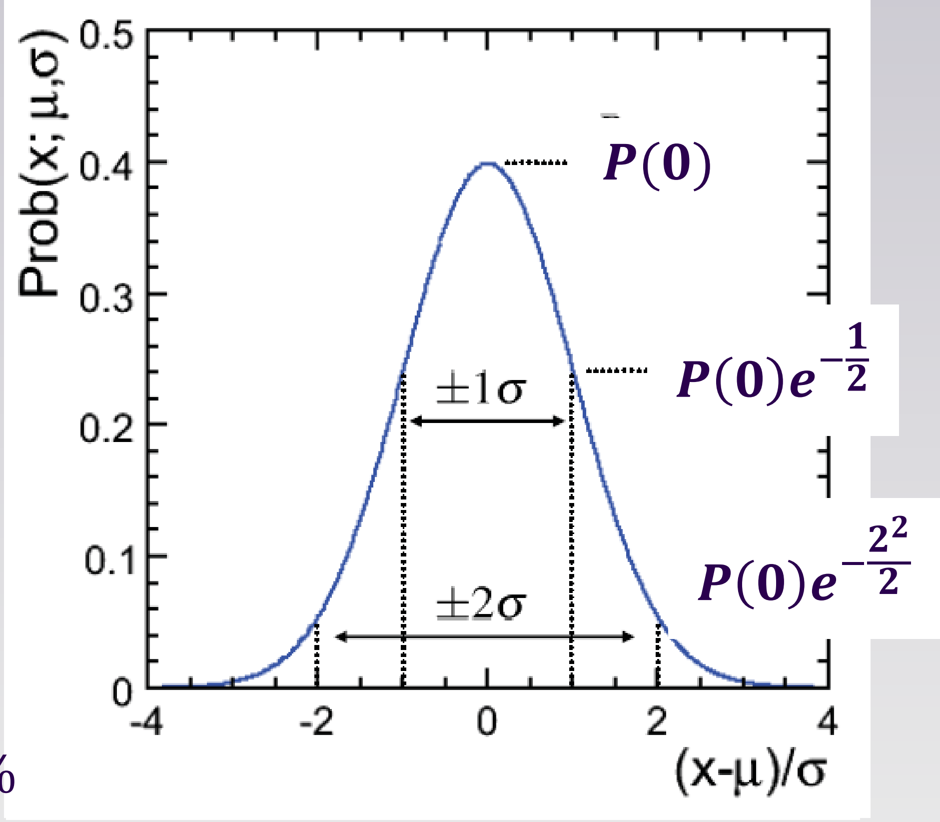
$$E[x] = \mu$$

- Variance:

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

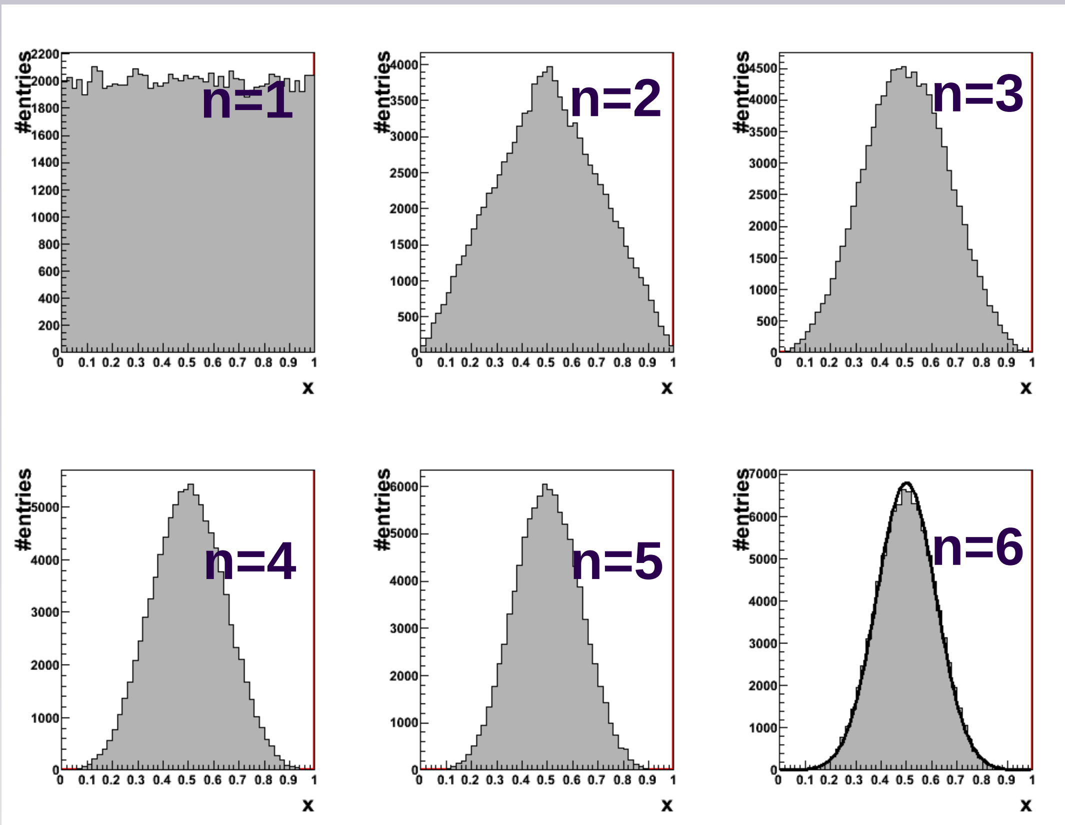
- Probability content:

$$\int_{-\sigma}^{\sigma} P(x)dx \cong 68\% \quad \int_{-2\sigma}^{2\sigma} P(x)dx \cong 95\%$$



# Central Limit Theorem

- **The mean *y* of *n* samples $x_i$ from any distribution D with well defined expectation value and variance $\lim_{n\to\infty} \rightarrow$ Gaussian**



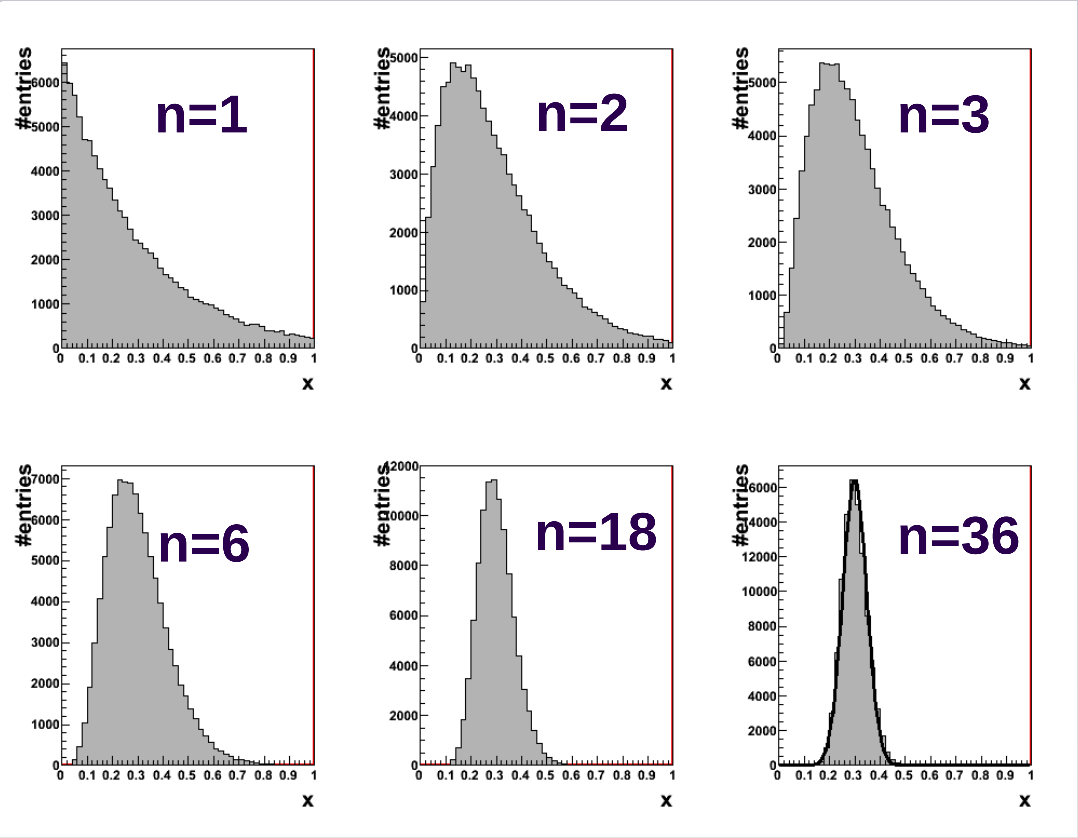

→ **Averaging reduces the error**

$$D:\ \ E_D[x] = \mu; V_D[x] = \sigma_D^2 \xrightarrow{\text{summation}} E_{Gauss}[y] = \mu;\ V_{Gauss}[y] = \frac{\sigma_D^2}{n}$$

# Central Limit Theorem

- Yes, even if D doesn't look „Gaussian" at all !

e.g. „exponential distribution"



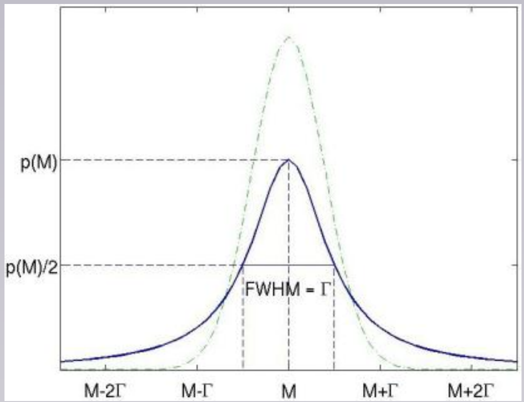

**Measurement errors:**

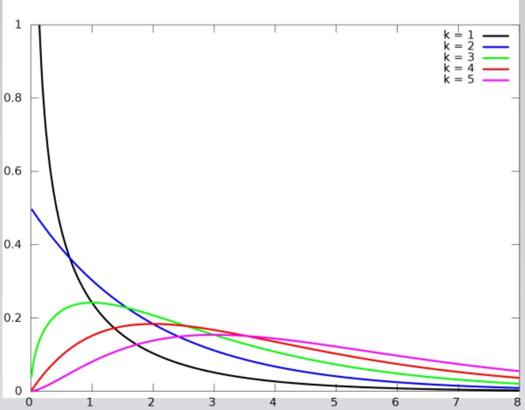
→ Typically: many contributions

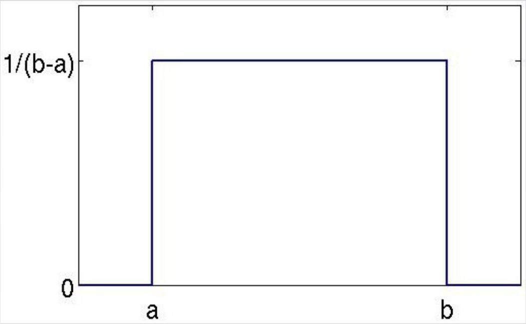
→ Gaussian !

# Some Other Distributions

- **Exponential – distribution**
  - **time distr. until particle decays (in it's own rest frame)**
- **Breit−Wigner (Cauchy) – distribution**
  - **mass peaks (resonance curve)**
- **$\chi^2$ – distribution**
  - **sum of squares of Gaussian distributed variables**
  - goodness-of-fit
- **Landau – distribution**
  - charge deposition in a silicon detector
- **Uniform – distribution**
- **… and many more:**



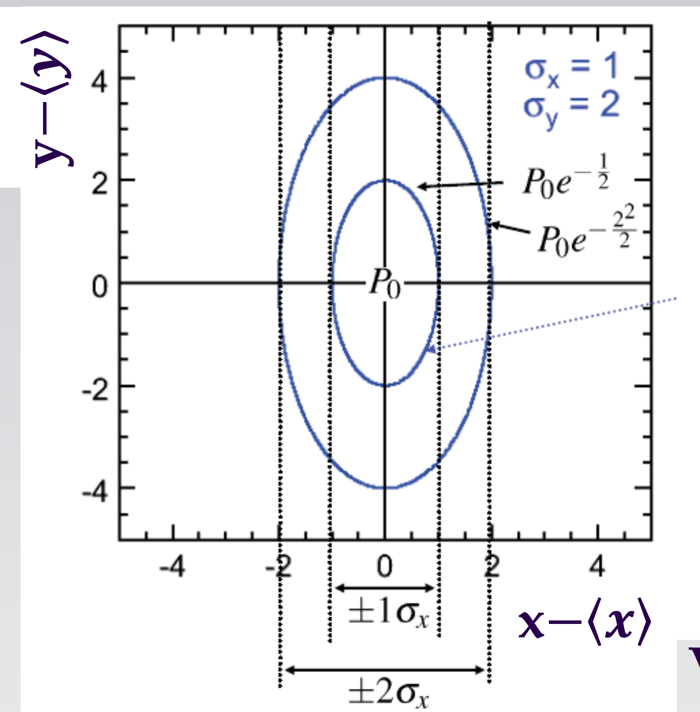



# 2D Gaussian

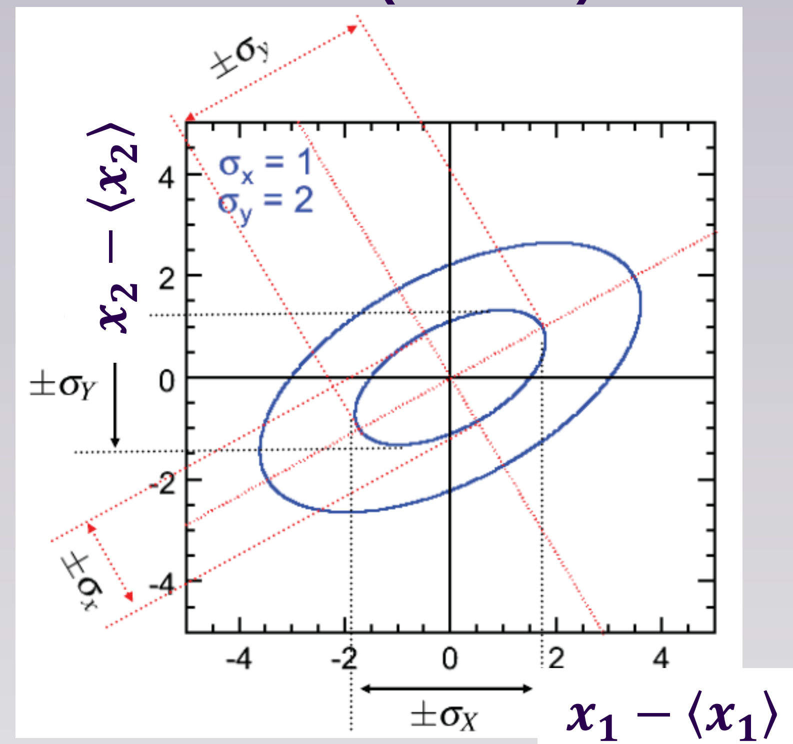
- **If the 2 variables are independent:**

$$P(x,y) = P(x)P(y)$$

$$P(\mathrm{x},\mathrm{y}) = \frac{1}{\sqrt{2\pi\sigma_x^2}} e^{-\frac{(x-\mu_x)^2}{2\sigma_x^2}} \frac{1}{\sqrt{2\pi\sigma_y^2}} e^{-\frac{(y-\mu_y)^2}{2\sigma_y^2}}$$




- **Correlated Gaussians ⇔ transformed (rotated) variables**




$$P(\vec{x}) = \frac{1}{2\pi\sqrt{det(V)}} e^{-\frac{1}{2}(\vec{x}-\vec{\mu})^T V^{-1} (\vec{x}-\vec{\mu})}$$

**with**

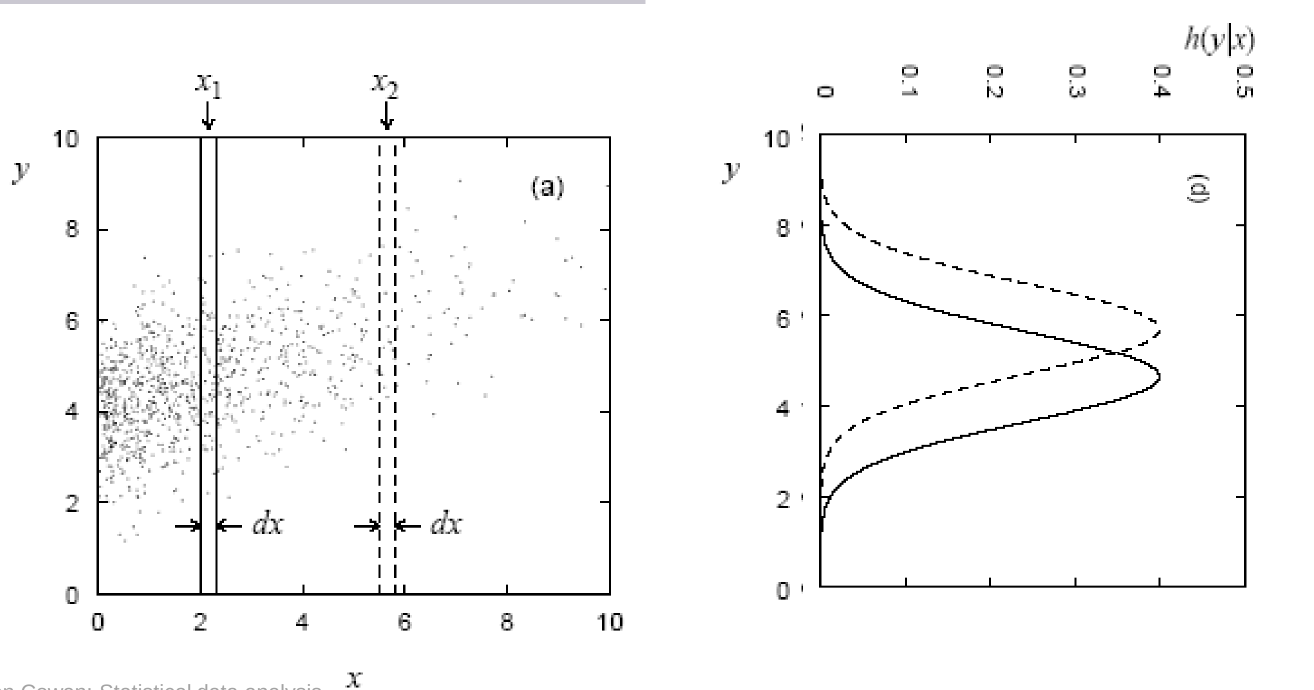
$$\mathrm{V}=\begin{pmatrix} \langle x_1^2\rangle - \langle x_1\rangle^2 & \langle x_1 x_2\rangle - \langle x_1\rangle\langle x_2\rangle \\ \langle x_1 x_2\rangle - \langle x_1\rangle\langle x_2\rangle & \langle x_2^2\rangle - \langle x_2\rangle^2 \end{pmatrix}$$

**co-variance matrix**

# Conditioning and Marginalisation

- **conditional probability:** $P(A|B) = \dfrac{P(A \cap B)}{P(B)} = \dfrac{p(x,y)dxdy}{p_x(x)dx}$

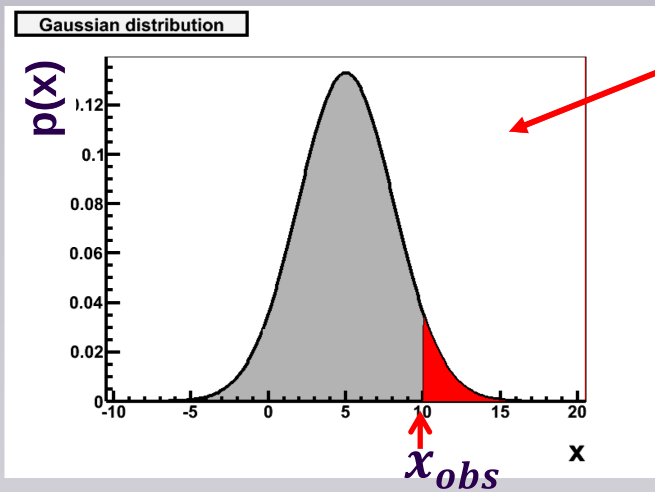
**↔ consider some variable in the joint PDF(*x,y*) as constant (given):**




Glen Cowan: Statistical data analysis

- **marginalisation:** **If you are not interested in the dependence on "*x*"**
  **→ project onto "*y*" (integrate "x out")**

# Cumulative Distribution



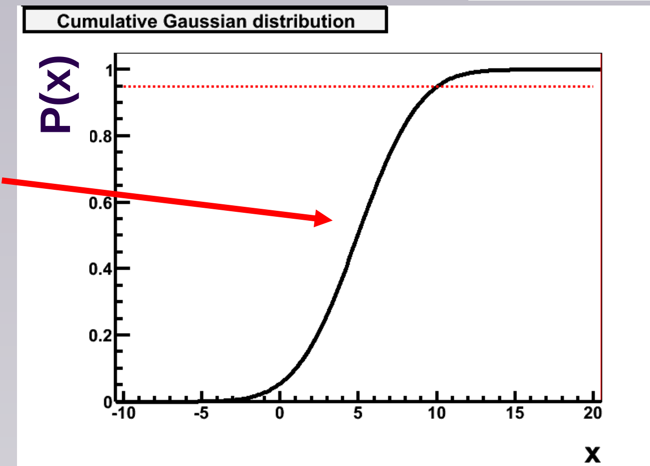

PDF

(probability density function)

Cumulative distribution:

$$\int_{-\infty}^{x} p(x')dx' \equiv P(x)$$

$$\rightarrow \boldsymbol{p(x) = dP(x)/dx}$$




- **$p(x)$: probability distribution for some "measurement" $x$ under the assumption of some model (parameter)**

**<u>Example of Cumulative distribution usage:</u>**

- **imagine you measure $x_{obs}$**
  - **how often expect I s.th. as far "off" the expectation (mean) value**
  - $1\text{-}\int_{-\infty}^{x_{obs}} p(x')dx' \equiv p - value$ **for observing something at least as far away from what you expect**

  **(one tailed as in example if "new physics" would be at higher x)**

- **similar: $\chi^2$-Probability**

**we will come back to this...**

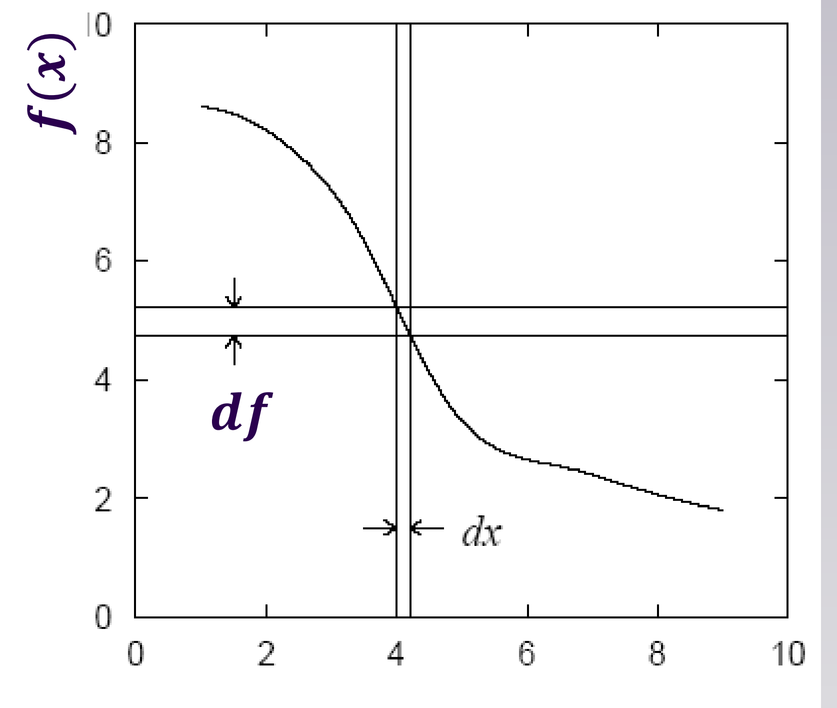
# Functions of Random Variables

- **A function of a random variable is itself a random variable.**
  - **$x$ with PDF $p(x)$**
  - **function $f(x)$**
    - e.g. extraction of a physics parameter from a measurement
- **PDF $g(f)$?**

$$g(f)df = \int_{dS} p(x)dx$$

**here: $dS$ =region of $x$ space for which**

- **$f$ is in $[f, f + \Delta f]$**
- **For one-variable case with unique inverse this is simply:**



Glen Cowan: Statistical data analysis

$$g(f)df = p(x)dx \quad \rightarrow \quad g(f) = p(x(f))\left|\frac{dx}{df}\right|$$

**Note: this is NOT the standard error propagation but the FULL PDF !**

# Error Propagation

- Either generate the FULL PDF of $f(x)$ based on the PDF for $x, p(x)$
  - often the full PDF for x is not known, but only a mean value $\mu$ and variance $\sigma^2$ (covariance matrix) have been estimated $\bar{x}$ and $\hat{V}$

→ then expand $f(x)$ around $\mu$

$$f(x) \simeq f(\mu) + \frac{df}{dx}|_{x=\mu}(x-\mu)$$

→ $E[f(x)] \simeq f(\mu)$ (as: $E[x-\mu] = 0$)

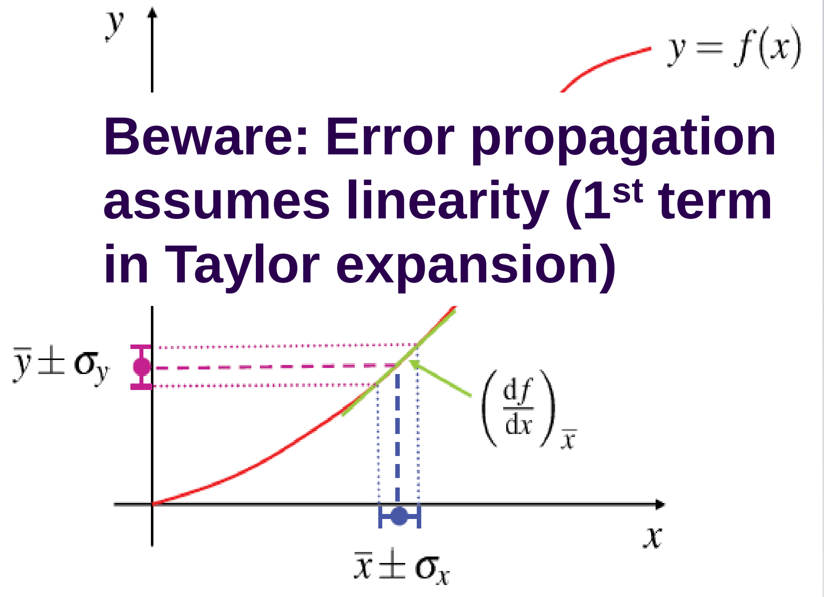
now let: $f(\mu) = f(\bar{x})$ and write as $\bar{y}$

→ $y - \bar{y} \simeq (x - \bar{x})\frac{df}{dx}|_{\bar{x}}$

→ $E\left[(y-\bar{y})^2\right] = \left(\frac{df}{dx}|_{\bar{x}}\right)^2 E\left[(x-\bar{x})^2\right]$

→ $\sigma_y^2 = \left(\frac{df}{dx}|_{\bar{x}}\right)^2 \sigma_x^2$

Beware: Error propagation assumes linearity (1st term in Taylor expansion)

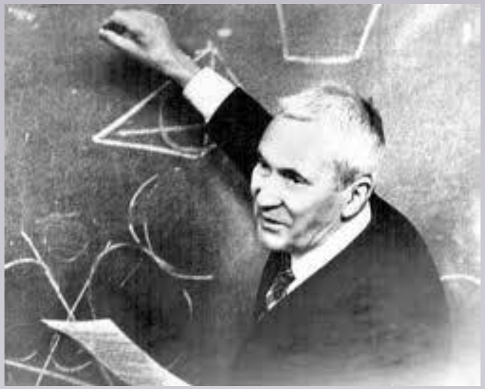



→ the "usual" formula $\sigma_y = \frac{df}{dx}|_{\bar{x}} \sigma_x$

- several variables → covariance matrix and partial derivatives

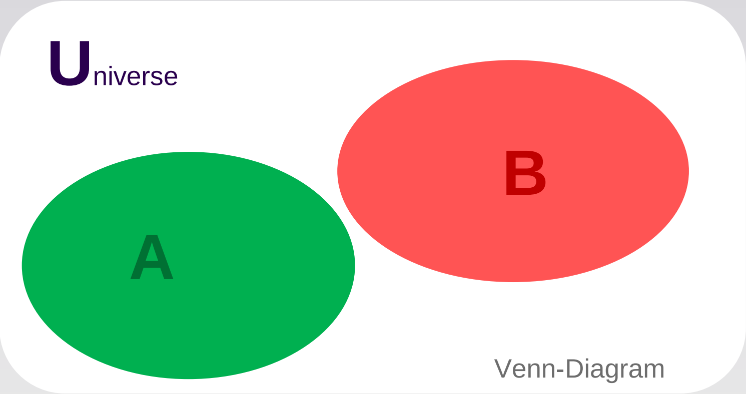
# What is Probability

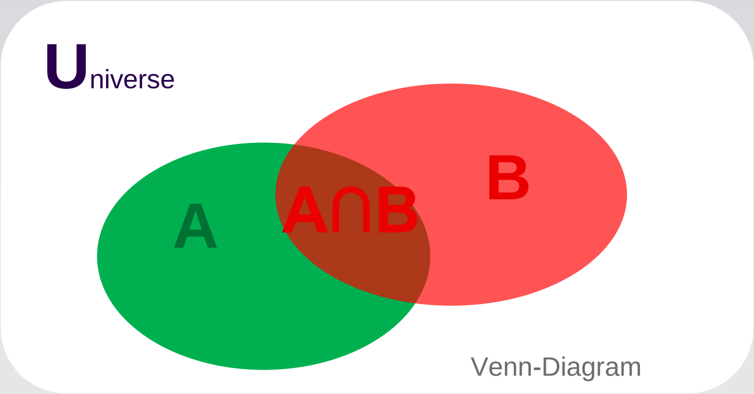
- **Axioms of probability: Kolmogorov (1933)**
  - $P(A) \geq 0$
  - $\int_U P(A)dU = 1$
  - **if:** $(A \text{ and } B) \equiv (A \cap B) = 0$
    **(i.e disjoint/independent/exclusive)**
    → $P(A \text{ or } B) \equiv (A \cup B) = P(A) + P(B)$



→ define e.g.: conditional probability

$$P(A|B) \equiv P(A \text{ given } B \text{ is true}) = \frac{P(A \cap B)}{P(B)}$$



# What is Probability

- Axioms of probability: → pure "set-theory"

**1) a measure of how likely an event will occur, expressed as a the ratio of favourable—to—all possible cases in repeatable trials**

- Frequentist (classical) probability

$$P(\text{"Event"}) = \lim_{n\to\infty}\left(\frac{\#\text{outcome is "Event"}}{n-\text{"}trials\text{"}}\right)$$

**2) the "degree of believe" that an event is going to happen**

- Bayesian probability:
  - $P(\text{"Event"})$: degree of believe that "Event" is going to happen → no need for "repeatable trails"
  - degree of believe (in view of the data AND previous knowledge(believe) about the parameter) that a parameter has a certain "true" value

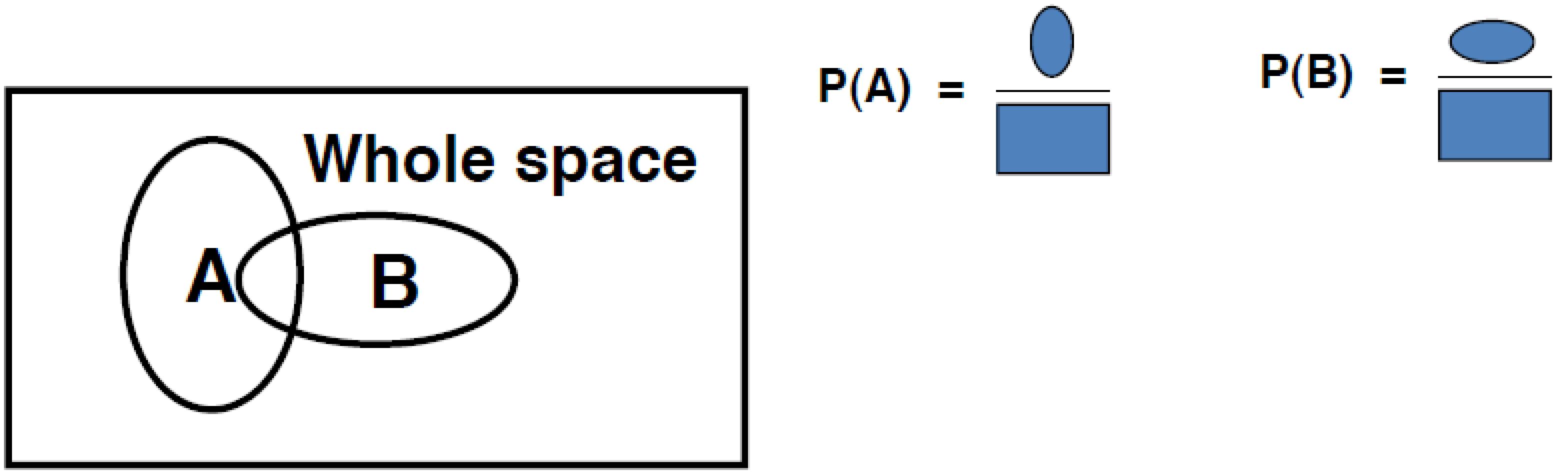
# Frequentist vs. Bayesian

**Bayes' Theorem** $$P(A|B) = \frac{P(B|A)P(A)}{P(B)} = P(B|A)\,\frac{P(A)}{P(B)}$$

- **This follows simply from the "conditional probabilities":**

# Derivation of Bayes' Theorem

**… in picture …taken from Bob Cousins**




Bob Cousins, CMS, 2008

# Frequentist vs. Bayesian

**Bayes' Theorem**

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

- **This follows simply from the "conditional probabilities":**

$$P(A|B)P(B) = P(A \cap B) = P(B \cap A) = P(B|A)P(A)$$

$$P(A|B)P(B) = P(B|A)P(A)$$

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

# Frequentist vs. Bayesian

**Bayes' Theorem**

$$P(\mu|n) = \frac{P(n|\mu)P(\mu)}{P(n)}$$

- $P(n|\mu)$: **Likelihood function**
- $P(\mu|n)$:**posterior probability of μ**
- $P(\mu)$: **the "prior"**
- $P(n)$: **just some normalisation**

**B.t.w.: Nobody doubts Bayes' Theorem:**
**discussion starts ONLY if it is used to turn**

**frequentist statements:**

- probability of the observed data given a certain model: $P(Data|Model)$

**into Bayesian probability statements:**

- probability of a the model begin correct (given data): $P(Model\,|Data)$

- **… there can be heated debates about 'pro' and 'cons' of either….**

# P (Data|Theory) ≠ P (Theory|Data)

- Higgs search at LEP: the statement
  - the probability that the data is in agreement with the Standard Model background is less than 1% (i.e. P(data| SMbkg) < 1%)

  went out to the press and got turned round to:

  ~~P(data|SMbkg) = P(SMbkg|data)~~ < 1% → P(Higgs|data) > 99% !

**WRONG!**

- easy Example: Theory = female (hypothesis) .. male (alternative)

  Data = pregnant or not pregnant

P (pregnant | female) ~ 2-3% but P (female | pregnant) = ?? ☺

→o.k… but what DOES it say?

# The correct frequentist interpretation

we know: P (Data|Theory) ≠ P (Theory|Data)

rather: Bayes Theorem: $P\,(\text{Data}|\text{Theory}) = P\,(\text{Theory}|\text{Data})\,\frac{P(\text{Theory})}{P(\text{Data})}$

Frequentists answer ONLY: P (Data|Theory)

… although.. let's be honest, we are all interested in P(Theory…)

We only learn about the "probability" to observe certain data under a given theory. Without knowledge of how likely the theory (or a possible "alternative" theory ) is .. that doesn't say anything about how unlikely this makes our current theory !

Later: we'll define "confidence levels" … i.e. if P(data) < 5%, discard theory.

→ can accept/discard theory and state how often/likely we will be wrong in doing so. But again: It does not say how "likely" the theory itself (or the alternative) is true

→ note the subtle difference !!

# Be aware !

**BBC: 2 July 2012: US sees stronger hints of Higgs**

By Paul Rincon Science editor, BBC News website

- **The signal is seen at the 2.9-sigma level of certainty, which means there is roughly a one in 1,000 chance that the result is attributable to some statistical quirk in the data**
- **The number of standard deviations, or sigmas, is a measure of how unlikely it is that an experimental result is simply down to chance rather than a real effect**

# Frequentist vs. Bayesian

- Certainly: both have their “right-to-exist”
  - Some “probably” reasonable and interesting questions cannot even be ASKED in a frequentist framework :
    - “How much do I trust the simulation”
    - “How likely is it that it will raining tomorrow?”
    - “How likely is it that climate change is going to…
  - after all.. the “Bayesian” answer sounds much more like what you really want to know: i.e.

    “How likely is the “parameter value” to be correct/true ?”
- <u>BUT:</u>
  - NO Bayesian interpretation w/o “prior probability” of the parameter
    - where do we get that from?
    - all the actual measurement can provide is “frequentist”!

# Bayesian Prior Probabilties

- **"flat" prior $\pi(\theta)$ to state "no previous" knowledge (assumptions) about the theory?**
  - **often done, BUT WRONG:**
    - e.g. flat prior in $M_{Higgs}$ → not flat in $M^2_{Higgs}$
  - **Choose a prior that is invariant under parameter transformations**
  
  **→ Jeffrey's Prior → "objective Bayesian":**
    - "flat" prior in Fisher's information space
    - $\pi(\theta) \propto \sqrt{\mathrm{I}(\theta)}$ $\quad$ ($\pi(\vec{\theta}) \propto \sqrt{\det \mathrm{I}(\vec{\theta})}$ if several parameters)

$I(\theta) = -E_x[\frac{\partial^2}{\partial\theta^2} log(f(x \;;\theta)]$ :

- $f(x;\theta)$: Likelihood function of $\theta$, probability to observe $x$ for a give parameter $\theta$
- amount of "information" that data $x$ is 'expected' to contain about the parameter $\theta$

- **personal remark: nice idea, but "WHY" would you want to dot that?**
  - **still use a "arbitrary" prior, only make sure everyone does the same way**
  - **loose all "advantages" of using a "reasonable" prior if you choose already to use a Bayesian interpretation!**

# Frequentist or Bayesian?

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

"Frequentists use impeccable logic to deal with an issue of no interest to anyone"

Louis Lyons, Academic Lecture at Fermilab, August 17, 2004

- Traditionally: most scientists are/were "frequentists"
  - no NEED to make "decisions" (well.. unless you want to announce the discovery of the Higgs particle..)
  - it's ENOUGH to present data, and how likely they are under certain scenarios
    - keep doing so and combine measurements
- Bayesians are growing
  - well, at least now we have the means to do lots of prior comparisons: Computing power/ Markov Chain Monte Carlos

# Summary

- **Statistics is everywhere in science**
  - **need to be able to use it correctly**
  - **need to know about the available (possible) distributions**
- **What is probability?**
  - **the basics of "statistics"**
  - **axioms**
    - frequentist interpretation
    - Bayesian interpretation

- **Tomorrow: How to use these things to answer your scientific questions**