

Probability and Statistics

Following A. Hoecker, H. Voss and K. Cranmer

Popular textbooks

G. Cowan, *Statistical Data Analysis*, Clarendon Press, Oxford, 1998.

R.J.Barlow, *A Guide to the Use of Statistical Methods in the Physical Sciences*, John Wiley, 1989;

F. James, *Statistical Methods in Experimental Physics*, 2nd ed., World Scientific, 2006;

▸ W.T. Eadie et al., *North-Holland*, 1971 (1st ed., hard to find);

S.Brandt, *Statistical and Computational Methods in Data Analysis*, Springer, New York, 1998.

L.Lyons, *Statistics for Nuclear and Particle Physics*, CUP, 1986.

Machine learning

C.M. Bishop, *Pattern Recognition and Machine Learning*, Springer 2006

T. Hastie, R. Tibshirani, J. Friedman, *The elements of Statistical Learning*, Springer 2001

Why we need probability in the particle world

Since Laplace's times (1749–1827) the universe's fate was deterministic and, in spite of technical difficulties, was considered predictable if the complete equation of state were known.

Challenged by Heisenberg's uncertainty principle (1927), Albert Einstein proclaimed "*Gott würfelt nicht*" ("God does not play dice"), but *hidden variables* to bring back determinism through the back door into the quantum world were never found.

In quantum mechanics, particles are represented by wave functions. The size of the wave function gives the probability that the particle will be found in a given position. The rate, at which the wave function varies from point to point, gives the speed of the particle.

Quantum phenomena like particle reactions occur according to certain probabilities. Quantum field theory allows us to compute cross-sections of particle production in scattering processes, and decays of particles. It cannot, however, predict how a single event will come out. We use probabilistic "Monte Carlo" techniques to simulate event-by-event realisations of quantum probabilities.



Pierre-Simon de Laplace



Albert Einstein



Werner Heisenberg

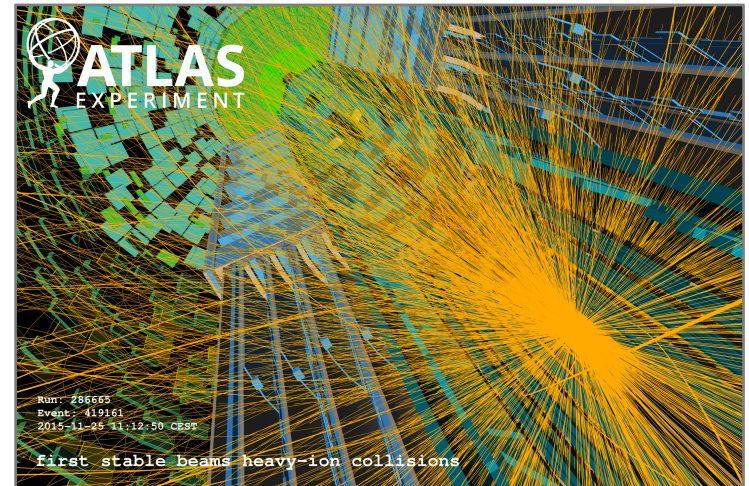


Statistics of large systems

Statistical physics uses probability theory and statistics to make statements about the approximate physics of large populations of stochastic nature, neglecting individuals.

Heavy-ion collisions at the LHC are modelled using hydrodynamics (strongly interacting medium behaves like perfect fluid)

Statistical mechanics provides a framework for relating the microscopic properties of individual atoms and molecules to the macroscopic properties of materials that can be observed in everyday life, therefore explaining thermodynamics as a natural result of statistics, classical mechanics, and quantum mechanics at the microscopic level.



Display of ATLAS Run-2 Heavy-Ion collision

Probability and statistics are fundamental ingredients & tools in all modern sciences

Statistical distributions

Measurement results typically follow some “distribution”, ie, the data do not appear at fixed values, but are “spread out” in a characteristic way

Which type of distribution it follows depends on the particular case

- It is important to know the occurring distributions to be able to pick the correct one when interpreting the data (example: *Poisson vs. Compound Poisson*)
- ...and it is important to know their characteristics to extract the correct information

Note: in statistical context, instead of “data” that follow a distribution, one often (typically) speaks of a “random variable”

Terms (Cranmer, Cowan)

- Random variables / “observables” x
- Probability (mass) and probability density function (pdf) $p(x)$
- Parametrized Family of pdfs / “model” $p(x|\alpha)$
- Parameter α
- Likelihood $L(\alpha)$
- Estimate (of a parameter) $\hat{\alpha}(x)$

Random variable / observable

“Observables” are quantities that we observe or measure directly

- ▶ They are random variables under repeated observation

Discrete observables:

- ▶ number of particles seen in a detector in some time interval
- ▶ particle type (electron, muon, ...) or charge (+,-,0)

Continuous observables:

- ▶ energy or momentum measured in a detector
- ▶ invariant mass formed from multiple particles

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

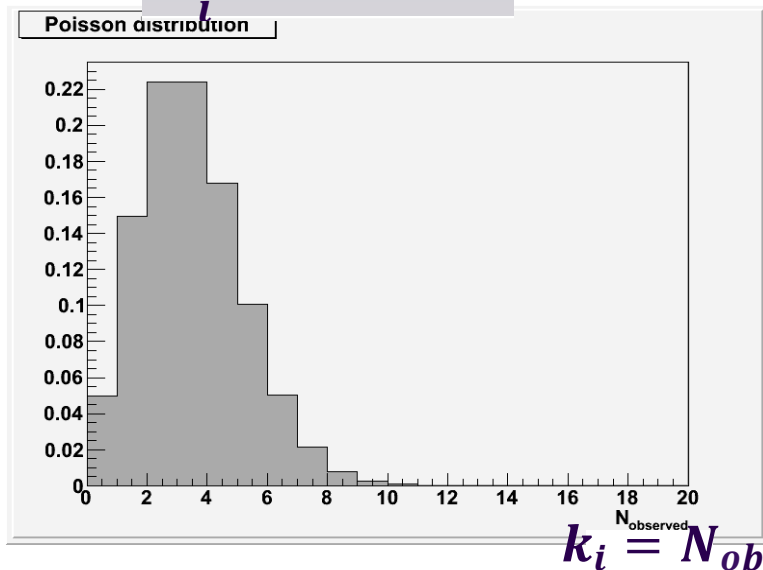
random variable x or k : characteristic quantity of a point in sample space

discrete variables

$$P(k_i) = p_i$$

normalisation (your parameter/event space covers all possibilities - unitarity)

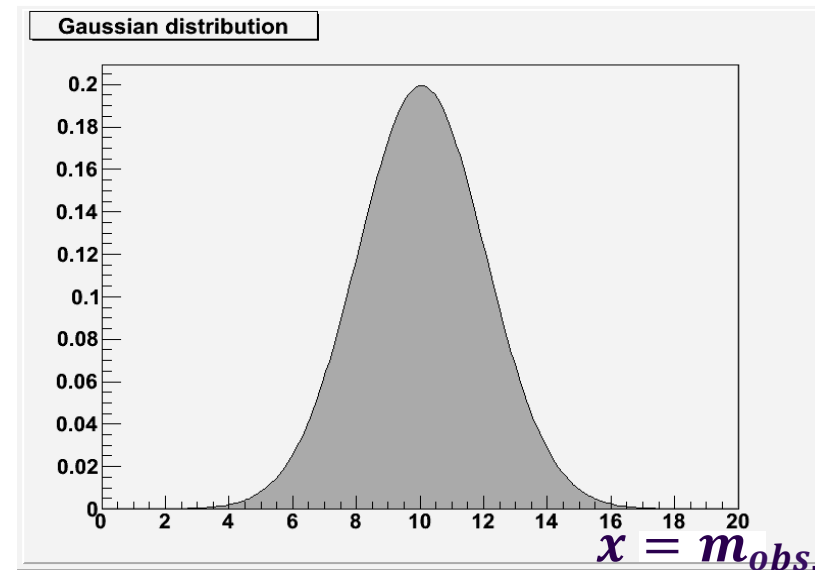
$$\sum_i P(k_i) = 1$$



continuous variables

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

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



Cumulative distribution

$p_x(\mathbf{x})$: probability density distribution for some “measurement” \mathbf{x} under the assumption of some model and its parameters

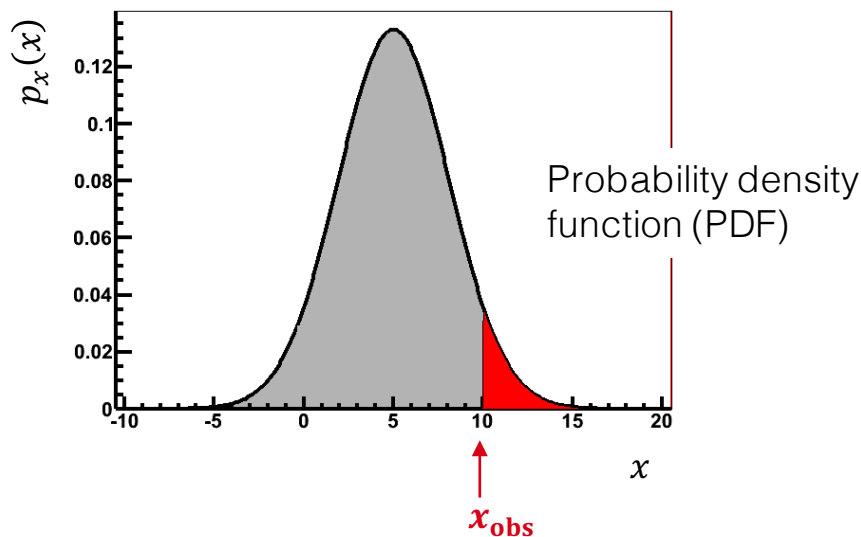
The cumulative distribution $P(\mathbf{x})$ is *the probability to observe a random value \mathbf{x} smaller than the one observed, \mathbf{x}_{obs}*

→ Examples for cumulative distributions: χ^2 , p-values, confidence limits (will come back to this)

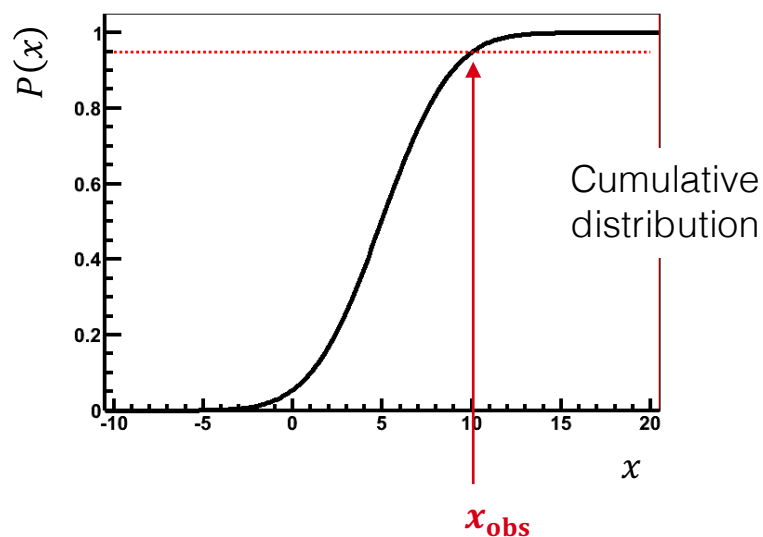
$$p_x(x) = dP(x)/dx$$

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

Gaussian distribution



Cumulative Gaussian distribution



Selected probability (density) distributions

Imagine a monkey discovered a huge bag of alphabet noodles. She blindly draws noodles out of the bag and places them in a row before her. The text reads:
“TO BE OR NOT TO BE”

The probability for this to happen is about 10^{-22}



Infinite monkey theorem: provided enough time, the monkey will type Shakespeare's Hamlet

Bernoulli Distribution

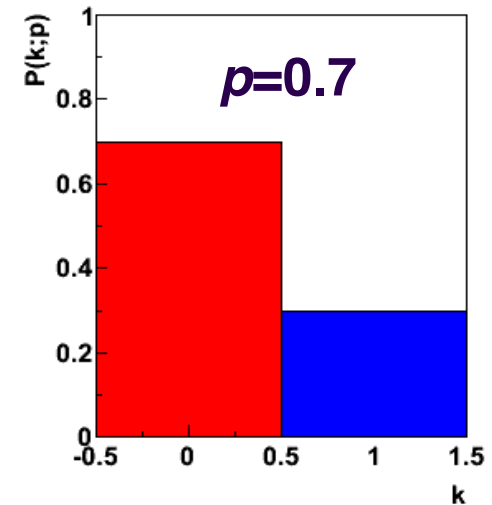
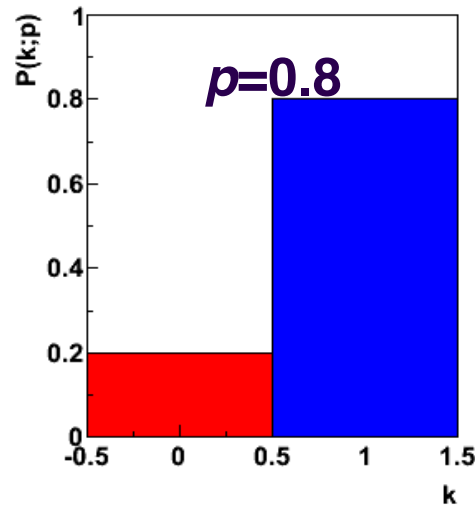
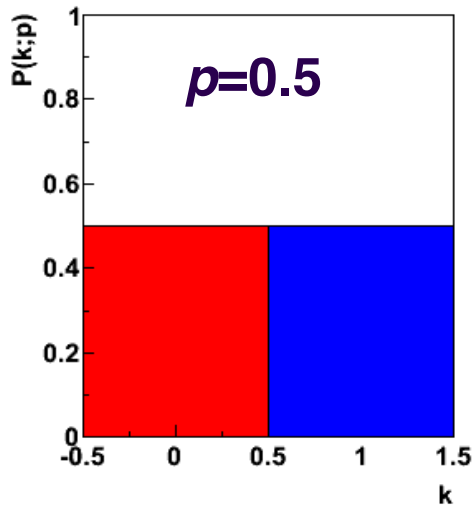
- 2 possible outcomes:

- ▶ Yes/No
- ▶ Head/Tail
- ▶



- (fair) coin: $P(\text{head}) = p$ (e.g. = $\frac{1}{2}$), $P(\text{tail}) = 1 - P(\text{head}) = 1 - p$

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



Binomial distribution (very important!)

Now let's get more complex: throw N coins (or similar binary choices)

How often (likely) is $k \times \mathbf{head}$ and $(N - k) \times \mathbf{tail}$?

- Each coin: $P(\mathbf{head}) = p, P(\mathbf{tail}) = 1 - p$
- Pick k particular coins \rightarrow the probability of all having **head** is:

$$P(k \times \mathbf{head}) = P(\mathbf{head}) \cdot P(\mathbf{head}) \cdot \dots \cdot P(\mathbf{head}) = P(\mathbf{head})^k = p^k$$

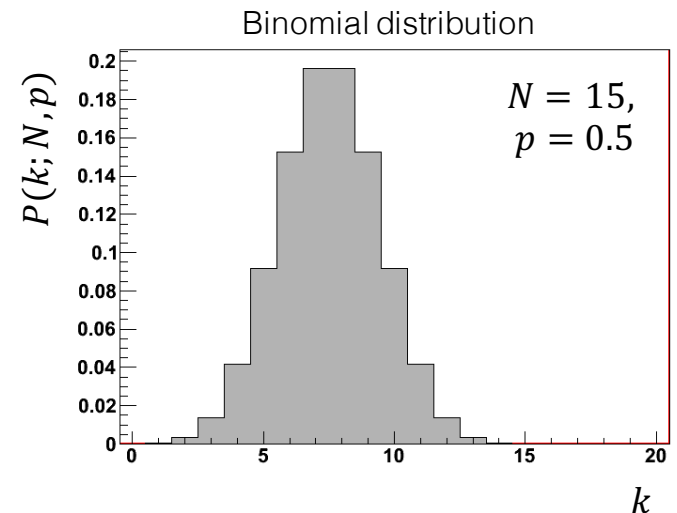
- Multiply this by the probability that all remaining $N - k$ coins land on **tail**:

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

- This was for a particular choice of k coins
- Now include all $\binom{N}{k}$ permutations for *any* k coins

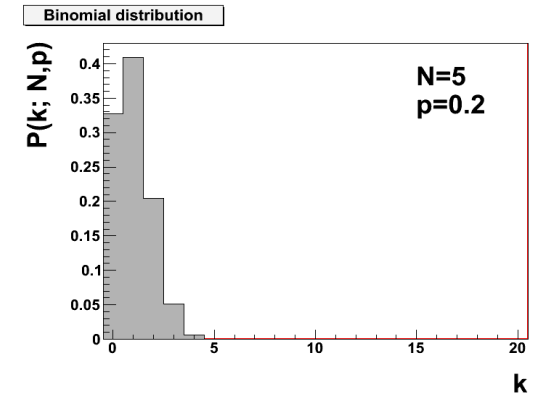
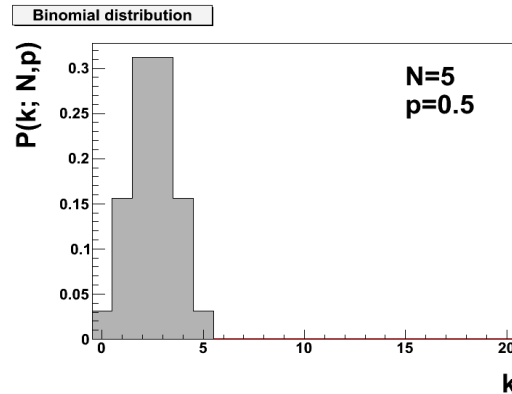
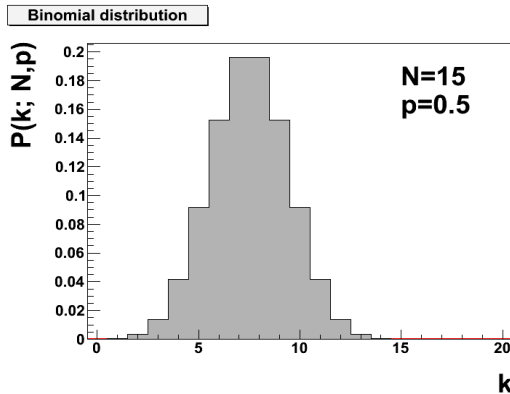
$$P(k; N, p) = p^k (1 - p)^{N-k} \binom{N}{k}$$

where $\binom{N}{k} = \frac{N!}{k!(N-k)!}$ is the binomial coefficient



Binomial Distribution

Examples:



- Expectation value: sum over all possible outcomes and “average”

$$E[k] = \sum kP(k) = Np$$

- Variance:

- $V(k) = Np(1 - p)$

Characteristic Quantities of Distributions

discrete variables

- Expectation value E (mean value):

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

- **Note:** mean/expectation of $f(x)$:

- Variance ($V = \sigma^2$, with σ : “spread”): $E[(x - \langle x \rangle)^2] = E[x^2] - (E[x])^2$

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

- higher moments: Skew: $E[(x - \langle x \rangle)^3]$

- **Note:** expectation and variance are properties of the full population.
Unbiased estimates, derived from samples taken from the distribution:

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

continuous variables

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

$$\rightarrow E[f(x)] = \int f(x)P(x)dx$$

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

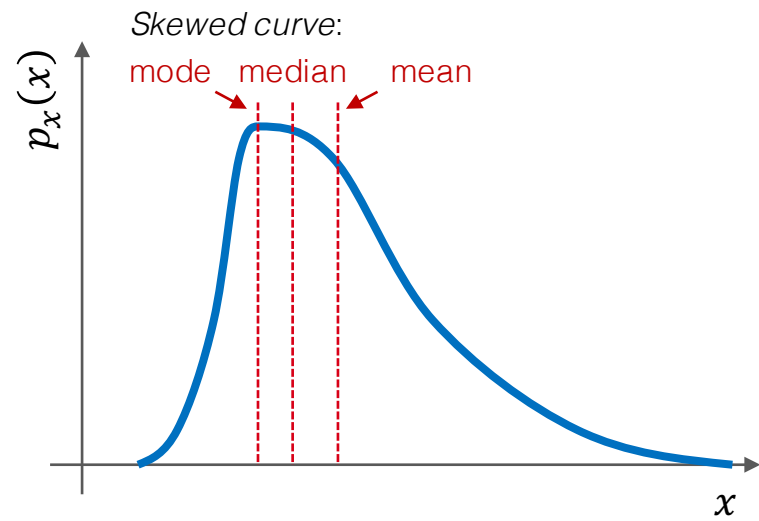
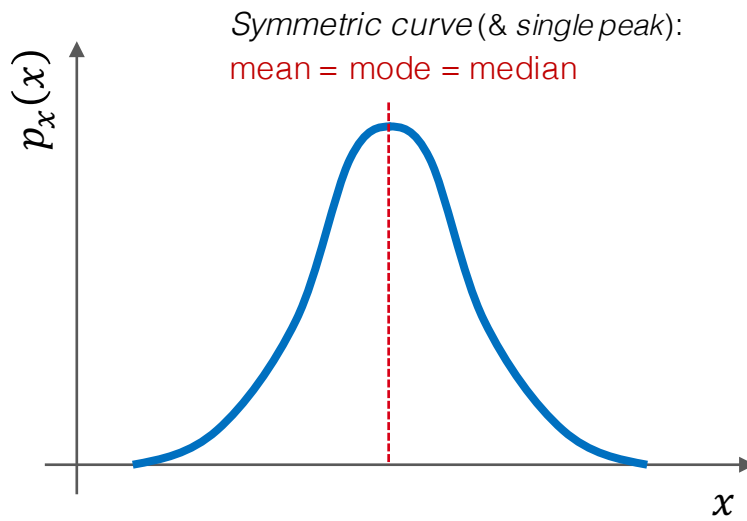
$$\hat{V} = \frac{1}{n-1} \sum_i^{\text{samples}} (x_i - \bar{x})^2$$

Mean, Mode, Median:

- **Mean:** $\langle x \rangle$ — defined before
- **Mode:** most probable value $x_{\text{mode}}: p_x(x_{\text{mode}}) \geq p_x(x), \forall x$
- **Median:** *2-quantile*: 50% of x values are larger than x_{median} , 50% are smaller

Can generalise *k-quantile*: points at regular intervals of the cumulative distribution.

Boundaries of binning chosen such that each bin contains the $1/k$ -th of total integral of distribution.



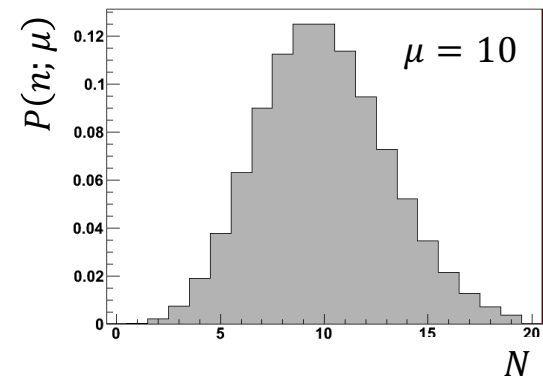
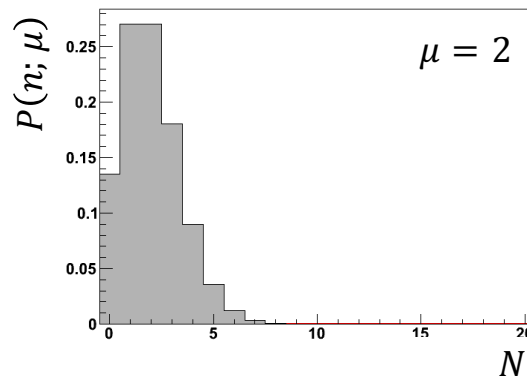
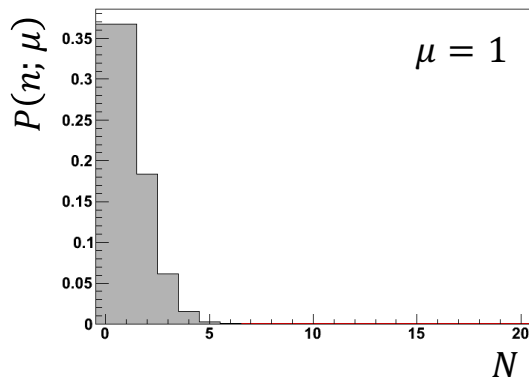
Poisson distribution

Recall: individual events each with two possible outcomes \rightarrow Binomial distribution

How about: number of counts in radioactive decay experiment during given time interval Δt ?

- Events happen “randomly” but there is no such 2nd outcome. Δt is continuous, no discrete trials
- μ : average number of counts in Δt . What is the probability of observing N counts?
- Limit of Binomial distribution for $N \rightarrow \infty$ & $p \rightarrow 0$ so that $Np \rightarrow \mu$.

\rightarrow Poisson distribution: $P(N; \mu) = \frac{\mu^N}{N!} e^{-\mu}$



Expectation value: $E[N] = \sum_N N \cdot P(N) = \mu$, Variance: $V[N] = \mu$

Poisson is good approximation for Binomial distribution for $N \gg \mu (= Np)$

Poisson distribution applies when:

- **each event is independent of all other events**
- > **there are no correlations**
- **you know the mean/average value**

It allows you to estimate the probability of a given fluctuation

Examples

Number of decays of radioactive nuclei per unit time

Number of junk email you receive per day

Probability of weather occurrences

Number of supernova star explosion within Milky Way (our) galaxy

On a particular river, overflow floods occur once every 100 years on average.

Calculate the probability of $k = 0, 1, 2, 3, 4, 5,$ or 6 overflow floods in a 100-year interval, assuming the Poisson model is appropriate.

Because the average event rate is one overflow flood per 100 years, $\lambda = 1$

$$P(k = 1 \text{ overflow flood in 100 years}) = \frac{1^1 e^{-1}}{1!} = \frac{e^{-1}}{1} = 0.368$$

$$P(k = 2 \text{ overflow floods in 100 years}) = \frac{1^2 e^{-1}}{2!} = \frac{e^{-1}}{2} = 0.184$$

The table below gives the probability for 0 to 6 overflow floods in a 100-year period.

k	$P(k \text{ overflow floods in 100 years})$
0	0.368
1	0.368
2	0.184
3	0.061
4	0.015
5	0.003
6	0.0005

Ugarte and colleagues report that the average number of goals in a World Cup soccer match is approximately 2.5 and the Poisson model is appropriate.^[3]

Because the average event rate is 2.5 goals per match, $\lambda = 2.5$.

$$P(k \text{ goals in a match}) = \frac{2.5^k e^{-2.5}}{k!}$$

$$P(k = 0 \text{ goals in a match}) = \frac{2.5^0 e^{-2.5}}{0!} = \frac{e^{-2.5}}{1} = 0.082$$

$$P(k = 1 \text{ goal in a match}) = \frac{2.5^1 e^{-2.5}}{1!} = \frac{2.5e^{-2.5}}{1} = 0.205$$

$$P(k = 2 \text{ goals in a match}) = \frac{2.5^2 e^{-2.5}}{2!} = \frac{6.25e^{-2.5}}{2} = 0.257$$

Gaussian (also: “Normal”) distribution

In limit of large μ a Poisson distribution approaches a symmetric Gaussian distribution

- This is the case not only for the Poisson distributions, but for almost any sufficiently large sum of samples with different sub-properties (mean & variance) → **Central Limit Theorem** (will discuss later)
- Gaussian distribution is of utter use, and luckily has simple properties

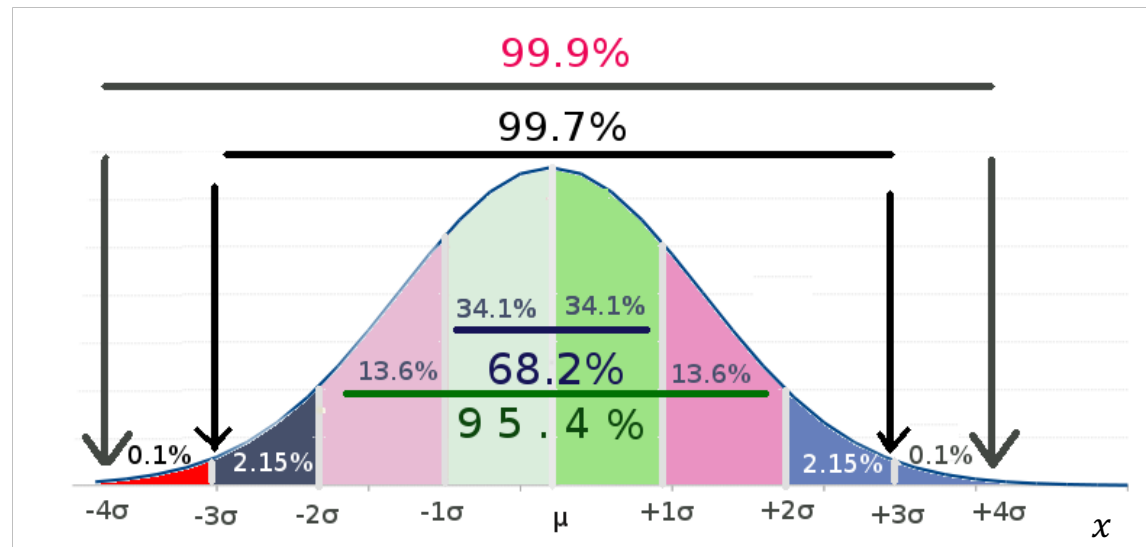
→ Gauss distribution:
$$P(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi} \cdot \sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Symmetric distribution:

- Expectation value: $E[x] = \mu$
- Variance: $V[x] = \sigma^2$
- Probability content:

$$\int_{-\sigma}^{+\sigma} P(x; \mu, \sigma) dx = 68.2\%$$

$$\int_{-2\sigma}^{+2\sigma} P(x; \mu, \sigma) dx = 95.4\%$$



Some other distributions

Uniform (“flat”) distribution

Exponential distribution

- Particle decay density versus time (in the particle’s rest frame!)

Relativistic Breit-Wigner distribution

- Distribution of resonance of unstable particle as function of centre-of-mass energy in which the resonance is produced (originates from the propagator of an unstable particle)

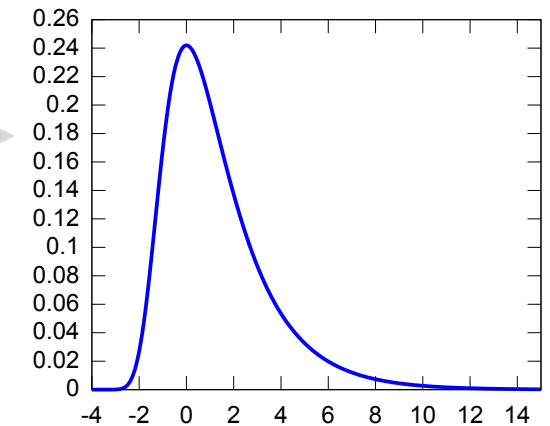
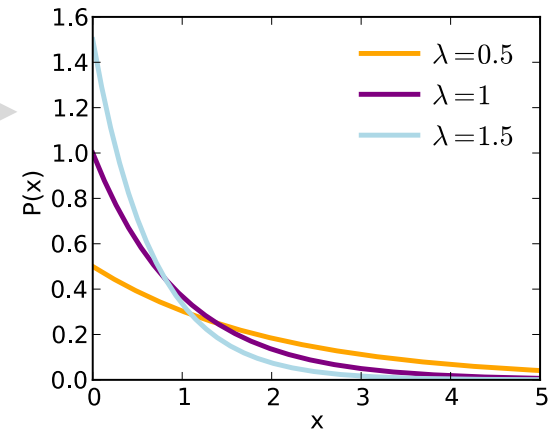
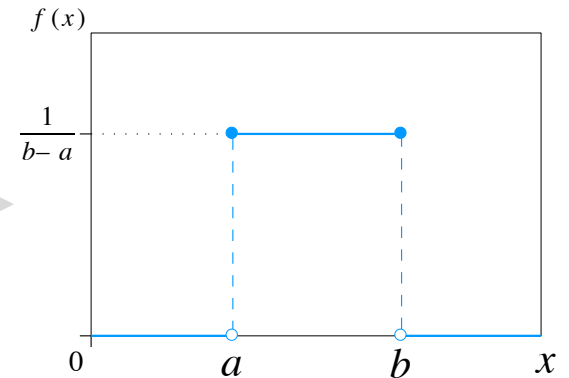
Chi-squared (χ^2) distribution

- Sum of squares of Gaussian distributed variables; used to derive goodness of a fit to describe data

Landau distribution

- Fluctuation of energy loss by ionization of charged particle in thin matter (eg, charge deposition in silicon detector)

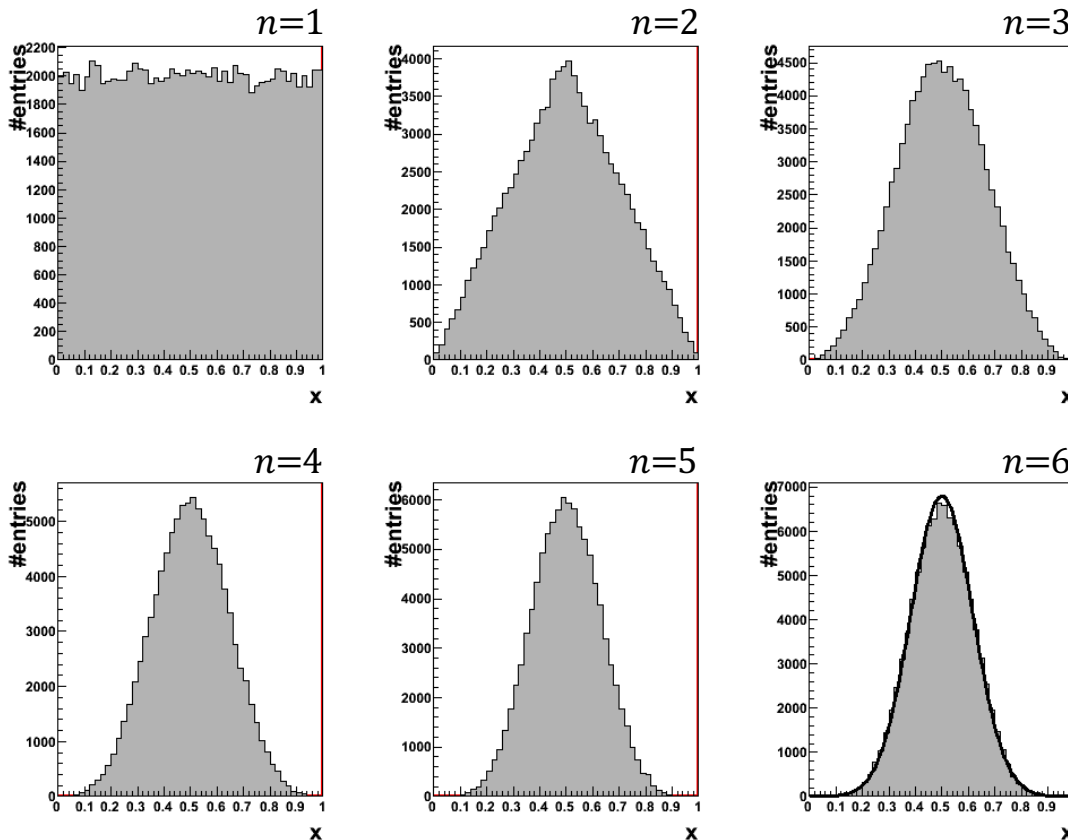
Many more, see <http://pdg.lbl.gov/2015/reviews/rpp2015-rev-probability.pdf> for definitions and properties.



Central limit theorem (CLT)

CLT: the sum of n independent samples x_i ($i = 1, \dots, n$) drawn from any PDF $D(x_i)$ with well defined expectation value and variance is Gaussian distributed in the limit $n \rightarrow \infty$

$$D: E_D[x_i] = \mu; V_D[x_i] = \sigma_D^2, \text{ and: } y = \sum_{i=1}^n x_i \Rightarrow E_{\text{Gauss}}[y] = \mu; V_{\text{Gauss}}[y] = \frac{\sigma_D^2}{n}$$



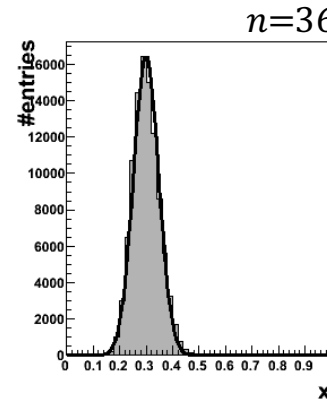
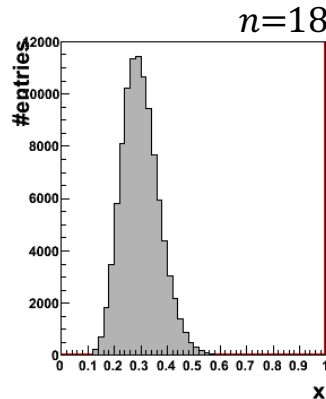
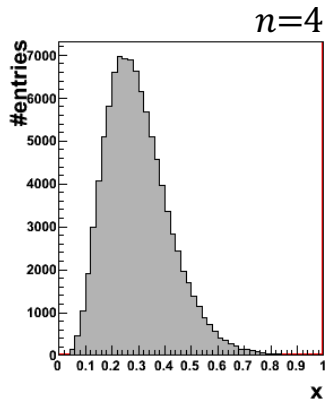
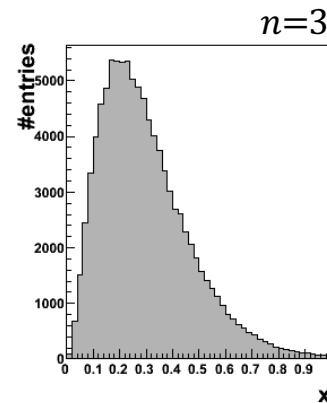
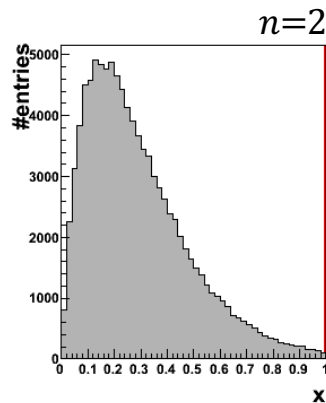
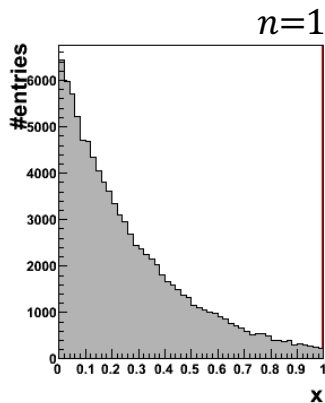
Averaging reduces the variance

Example: summing up ensembles uniformly distributed within $[0,1]$

Central limit theorem (CLT)

CLT: the sum of n independent samples x_i ($i = 1, \dots, n$) drawn from any PDF $D(x_i)$ with well defined expectation value and variance is Gaussian distributed in the limit $n \rightarrow \infty$

$$D: E_D[x_i] = \mu; V_D[x_i] = \sigma_D^2, \text{ and: } y = \sum_{i=1}^n x_i \Rightarrow E_{\text{Gauss}}[y] = \mu; V_{\text{Gauss}}[y] = \frac{\sigma_D^2}{n}$$



Example: summing up exponential distributions

Central Gaussian limit works even if D doesn't look Gaussian at all

lecture 27

What if a measurement consists of two variables?

Let:

A = measurement **x** in $[x, x + dx]$

B = measurement **y** in $[y, y + dy]$

Joint probability: $P(A \cap B) = p_{xy}(x, y) dx dy$

(where $p_{xy}(x, y)$ is joint PDF)

If the two variables are independent:

$$P(A, B) = P(A) \cdot P(B)$$

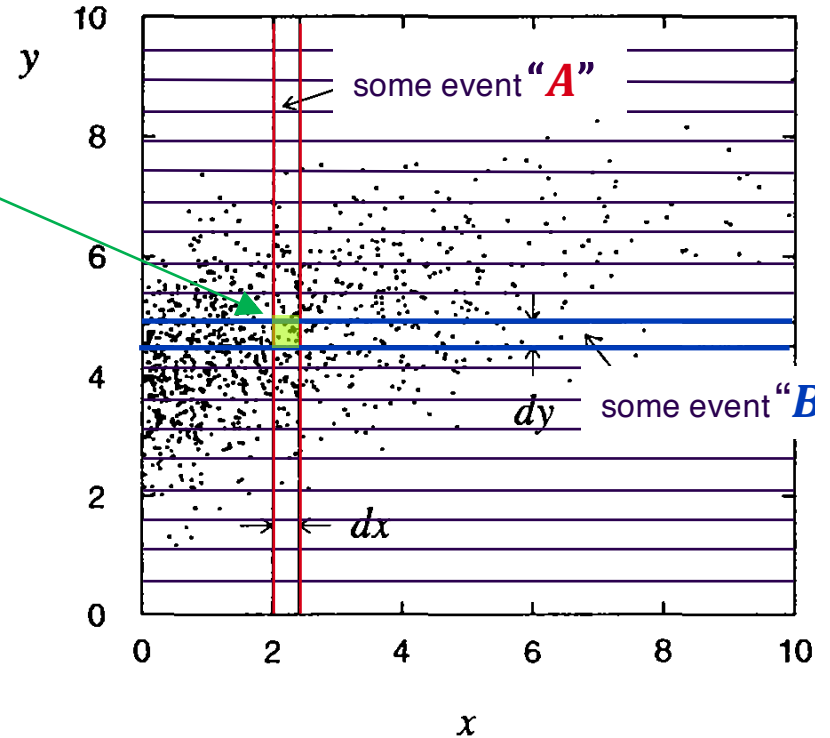
$$p_{xy}(x, y) = p_x(x) \cdot p_y(y)$$

Marginal PDF: if one is not interested in dependence on **y** (or cannot measure it),

→ integrate out (“marginalise”) **y**, ie, project onto **x**

→ resulting one-dimensional PDF: $p_x(x) = \int p_{xy}(x, y) dy$

From: Glen Cowan,
Statistical data analysis



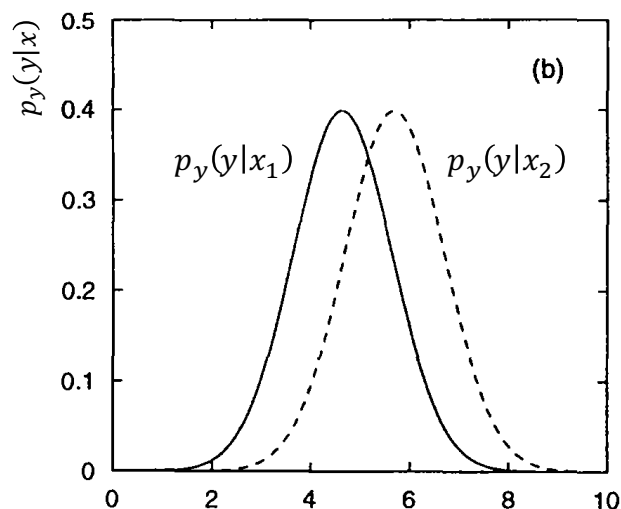
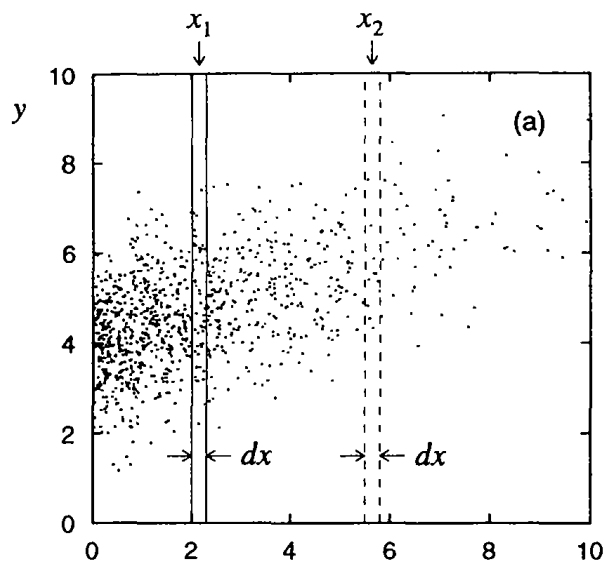
Conditioning versus marginalisation

Conditional probability $\mathbf{P(A|B)}$: [read: $P(A|B)$ = “probability of A given B ”]

$$\mathbf{P(A|B)} = \frac{P(A \cap B)}{P(B)} = \frac{p_{xy}(x, y) dx dy}{p_y(y) dx}$$

Rather than integrating over the whole y region (marginalisation), look at one-dimensional (1D) slices of the two-dimensional (2D) PDF $p_{xy}(x, y)$:

$$p_y(y|x_1) = p_{xy}(x = \text{const} = x_1, y)$$



From: Glen Cowan,
Statistical data analysis

Covariance and correlation

Recall, for 1D PDF $p_x(\mathbf{x})$ we had: $E[x] = \mu_x$; $V[x] = \sigma_x^2$

E – expectation value
V - variance

For a 2D PDF $p_{xy}(\mathbf{x}, \mathbf{y})$, one correspondingly has: $\mu_x, \mu_y, \sigma_x, \sigma_y$

How do \mathbf{x} and \mathbf{y} co-vary? $\rightarrow C_{xy} = \text{covariance}_{xy} = E[(x - \mu_x)(y - \mu_y)] = E[xy] - \mu_x\mu_y$

Or the scale / dimension invariant *correlation coefficient*:

$$\rho_{xy} = \frac{C_{xy}}{\sigma_x\sigma_y}, \text{ where } \rho_{xy} \in [-1, +1]$$

- If x, y are independent: $\rho_{xy} = 0$, ie, they are *uncorrelated* (or they *factorise*)

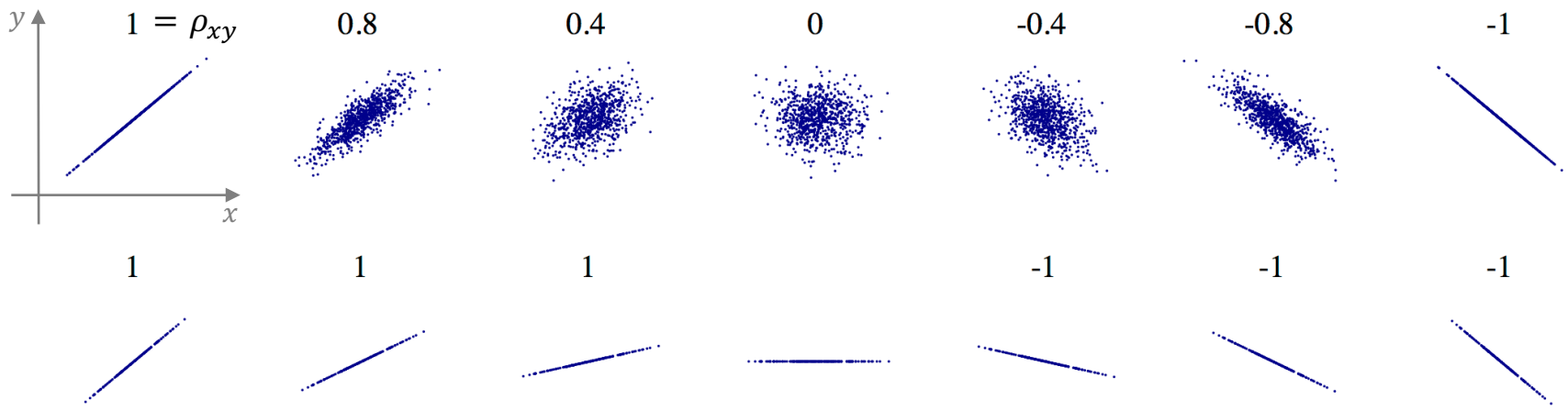
Proof: $E[xy] = \iint xy \cdot p_{xy}(x, y) dx dy = \int x \cdot p_x(x) dx \cdot \int y \cdot p_y(y) dy = \mu_x\mu_y$

- Note that the contrary is not always true: non-linear correlations can lead to $\rho_{xy} = 0$,

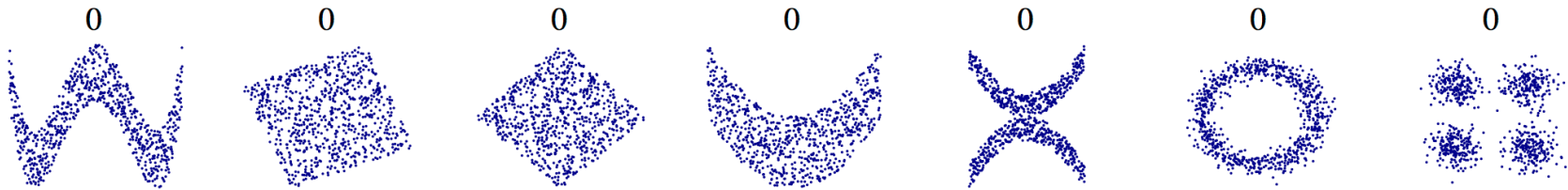
Correlations

Figure from: https://en.wikipedia.org/wiki/Correlation_and_dependence

The correlation coefficient measures the noisiness and direction of a linear relationship:



...it does not measure the slope ρ_{xy} (see above figures)



...and non-linear correlation patterns are not or only approximately captured by ρ_{xy} (see above figures)

Correlations

Non-linear correlation can be captured by the “*mutual information*” quantity I_{xy} :

$$I_{xy} = \iint p_{xy}(x, y) \cdot \ln \left(\frac{p_{xy}(x, y)}{p_x(x)p_y(y)} \right) dx dy$$

where $I_{xy} = 0$ only if \mathbf{x}, \mathbf{y} are fully statistically independent

Proof: if independent, then $p_{xy}(x, y) = p_x(x)p_y(y) \Rightarrow \ln(\dots) = 0$

NB: $I_{xy} = H_x - H_x(y) = H_y - H_y(x)$,

where $H_x = - \int p_x(x) \cdot \ln(p_x(x)) dx$ is *entropy*, $H_x(y)$ is *conditional entropy*



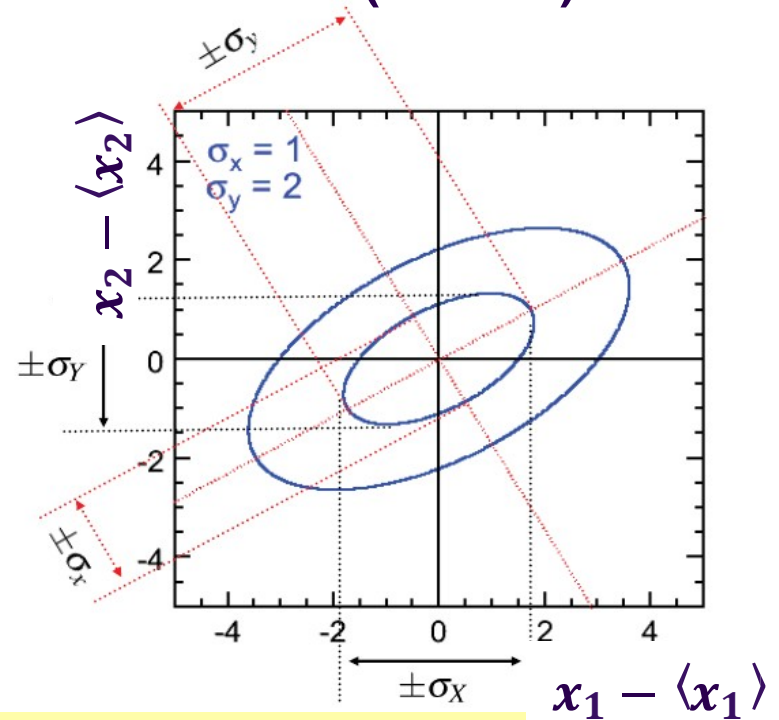
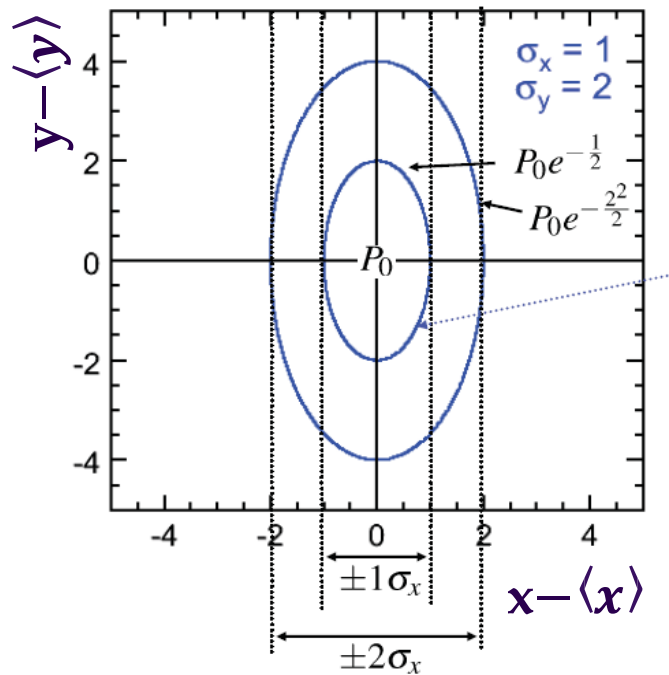
2D Gaussian

- If the 2 variables are independent:

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

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

- Correlated Gaussians \Leftrightarrow transformed (rotated) variables



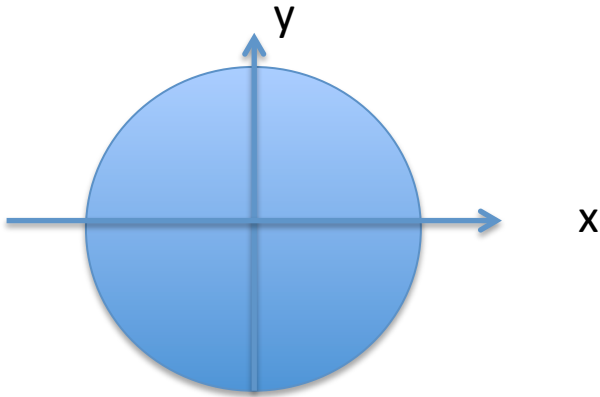
$$P(\vec{x}) = \frac{1}{2\pi\sqrt{\det(V)}} e^{-\frac{1}{2}(\vec{x}-\vec{\mu})^T V^{-1}(\vec{x}-\vec{\mu})}$$

$$V = \begin{pmatrix} \langle x_1^2 \rangle - \langle x_1 \rangle^2 & \langle x_1 x_2 \rangle - \langle x_1 \rangle \langle x_2 \rangle \\ \langle x_1 x_2 \rangle - \langle x_1 \rangle \langle x_2 \rangle & \langle x_2^2 \rangle - \langle x_2 \rangle^2 \end{pmatrix} \text{CO-} \\ \text{variance} \\ \text{matrix}$$

Warning

Marginalization may lead to observation of non-existing correlations

uniform distribution $A(x,y)$



projections $F(x)$ and $G(y)$ peaked at 0

Correlation function

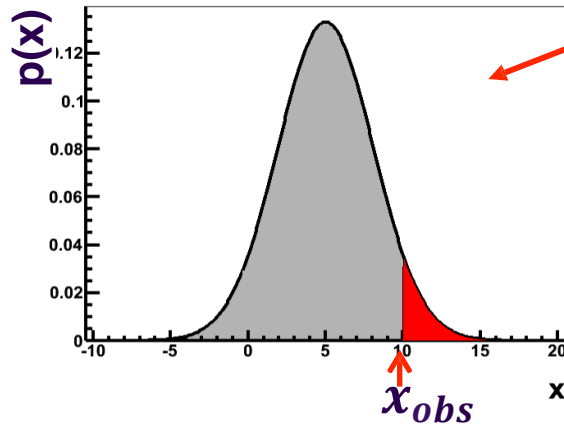
$$C(x,y) = A(x,y) - F(x)G(y)$$

will have an artificial minimum due to non-rectangular boundary.

Remember that event phase space has long-range correlation due to energy-momentum conservation.

Cumulative Distribution

Gaussian distribution



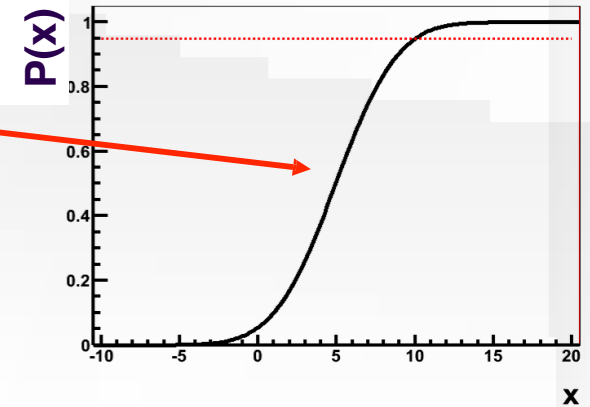
PDF

(probability density function)

Cumulative distribution:

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

Cumulative Gaussian distribution



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

Examples of Cumulative distribution usage:

- imagine you measure x_{obs}
- how often one expects x far “off” the expectation (mean) value?
 - $1 - \int_{-\infty}^{x_{obs}} p(x') dx' \equiv p - value$ for observing something at least as far away from what you expect
- similar: χ^2 -Probability

Functions of random variables

Any function of a random variable is itself a random variable

E.g., \mathbf{x} with PDF $\mathbf{p_x(x)}$ becomes: $\mathbf{y = f(x)}$

\mathbf{y} could be a parameter extracted from a measurement

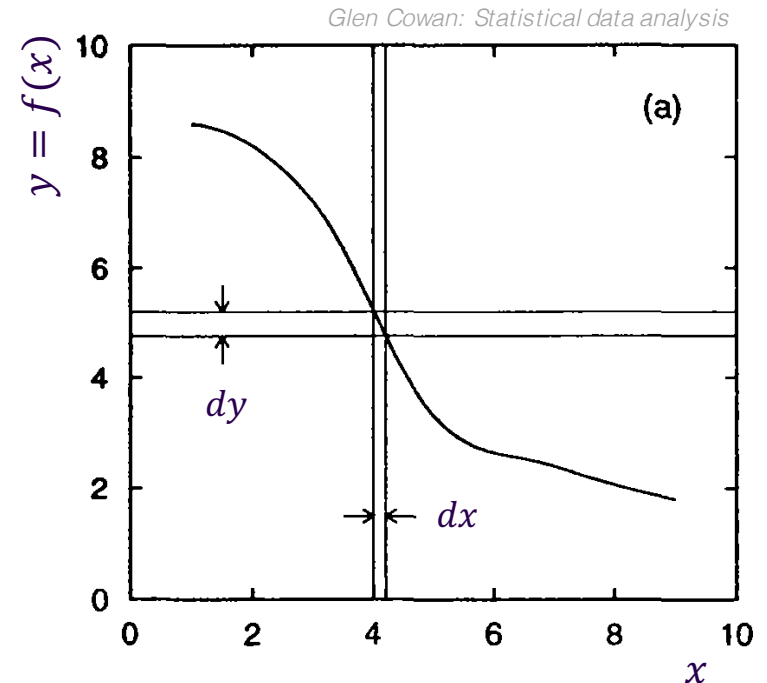
What is the PDF $\mathbf{p_y(y)}$?

- Probability conservation: $p_y(y)|dy| = p_x(x)|dx|$
- For a 1D function $f(x)$ with existing inverse:

$$dy = \frac{df(x)}{dx} dx \Leftrightarrow dx = \frac{df^{-1}(y)}{dy} dy$$

- Hence: $\mathbf{p_y(y) = p_x(f^{-1}(y)) \left| \frac{dx}{dy} \right|}$

Note: this is **not** the standard error propagation but the full PDF !



Error propagation

Let's assume a measurement \mathbf{x} with *unknown* PDF $p_{\mathbf{x}}(\mathbf{x})$, and a transformation $\mathbf{y} = \mathbf{f}(\mathbf{x})$

- \bar{x} and \hat{V} are estimates of μ and variance σ^2 of $p_{\mathbf{x}}(\mathbf{x})$

What are $E[\mathbf{y}]$ and, in particular, $\sigma_{\mathbf{y}}^2$? \rightarrow Taylor-expand $f(x)$ around \bar{x} :

- $f(x) = f(\bar{x}) + \left. \frac{df}{dx} \right|_{x=\bar{x}} (x - \bar{x}) + \dots \Rightarrow E[f(x)] \simeq f(\bar{x})$ (because: $E[x - \bar{x}] = 0$!)

Now define $\bar{y} = f(\bar{x})$, and from the above follows:

$$\Leftrightarrow y - \bar{y} \simeq \left. \frac{df}{dx} \right|_{x=\bar{x}} (x - \bar{x})$$

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

$$\Leftrightarrow \hat{V}_{\mathbf{y}} = \left(\left. \frac{df}{dx} \right|_{x=\bar{x}} \right)^2 \hat{V}_{\mathbf{x}}$$

$$\Leftrightarrow \sigma_{\mathbf{y}} = \left. \frac{df}{dx} \right|_{x=\bar{x}} \cdot \sigma_{\mathbf{x}} \quad \rightarrow \quad (\text{approximate}) \text{ error propagation}$$

Error propagation (continued)

In case of several variables, compute covariance matrix and partial derivatives

- Let $\mathbf{f} = \mathbf{f}(x_1, \dots, x_n)$ be a function of \mathbf{n} randomly distributed variables

- $\left(\frac{df}{dx}\bigg|_{x=\bar{x}}\right)^2 \hat{V}_x$ becomes: $\sum_{i,j=1}^n \frac{\partial f}{\partial x_i} \frac{\partial f}{\partial x_j} \bigg|_{\bar{x}} \cdot \hat{V}_{i,j}$ (where: $\bar{x} = (\bar{x}_1, \dots, \bar{x}_n)$)

- with the covariance matrix:

$$\hat{V}_{i,j} = \begin{bmatrix} \sigma_{x_1}^2 & \cdots & \sigma_{x_1 x_n} \\ \vdots & \ddots & \vdots \\ \sigma_{x_n x_1} & \cdots & \sigma_{x_n}^2 \end{bmatrix}$$

- 🐞 The resulting “error” (uncertainty) depends on the correlation of the input variables
 - Positive correlations lead to an increase of the total error
 - Negative correlations decrease the total error

lecture 28

Probability Functions

When dealing with discrete random variables, define a **Probability Function** as probability for i^{th} possibility

$$P(x_i) = p_i$$



Defined as limit of long term frequency

- probability of rolling a 3 := $\lim_{\# \text{ trials} \rightarrow \infty} (\# \text{ rolls with 3} / \# \text{ trials})$
 - you don't need an infinite sample for definition to be useful

Normalization

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

Probability Density Functions

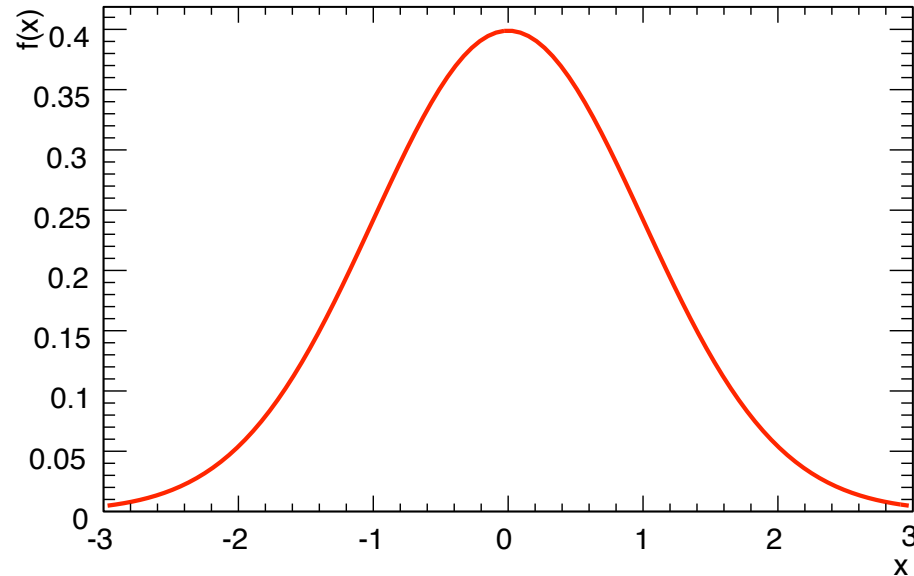
When dealing with continuous random variables, need to introduce the notion of a **Probability Density Function**

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

Note, $f(x)$ is NOT a probability

PDFs are always normalized to unity:

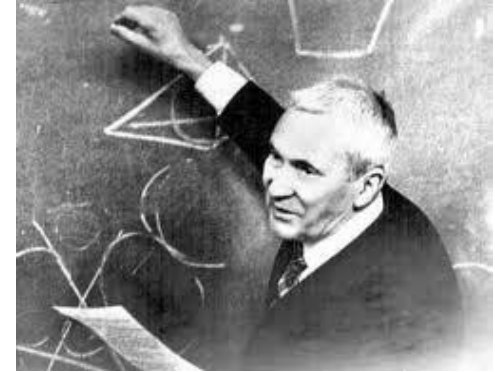
$$\int_{-\infty}^{\infty} f(x)dx = 1$$



What is Probability

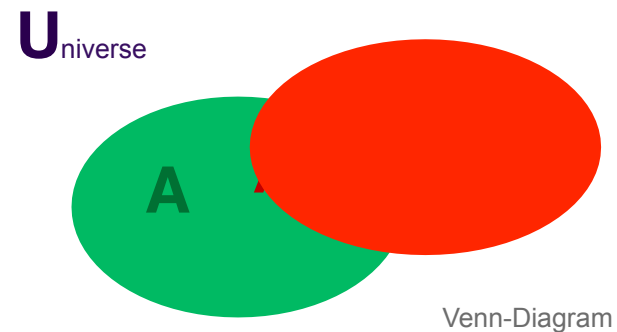
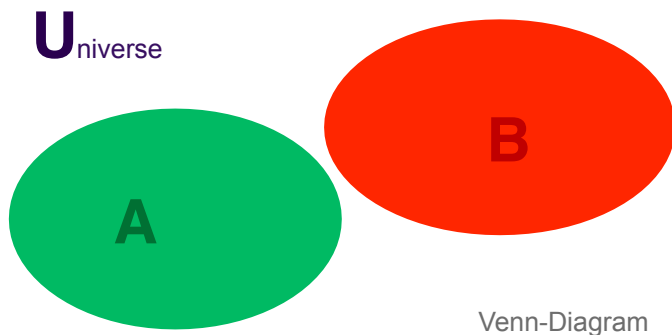
- **Axioms of probability: Kolmogorov (1933)**

- $P(A) \geq 0$
- $\int_U P(A)dU = 1$
- **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 the ratio of favourable—to—all possible cases in repeatable trials

- Frequentist (classical) probability

$$P(\text{“Event”}) = \lim_{n \rightarrow \infty} \left(\frac{\text{\#outcome is “Event”}}{n_{\text{trials}}} \right)$$

2) the “degree of belief” that an event is going to happen

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



Frequentist vs. Bayesian

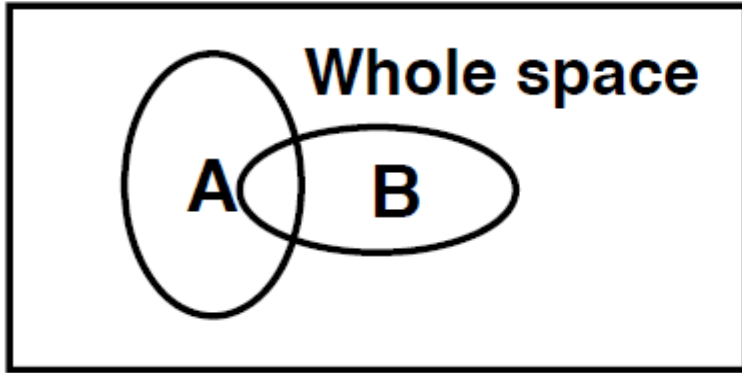
Bayes' Theorem

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

- This follows simply from the “conditional probabilities”:

Derivation of Bayes' Theorem

... in picture ...taken from Bob Cousins



$$P(A) = \frac{\text{Area of } A}{\text{Area of Whole space}}$$

$$P(B) = \frac{\text{Area of } B}{\text{Area of Whole space}}$$

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

∴ Nobody doubts Bayes' Theorem:
discussion starts ONLY if it is used to turn

frequentist statements:

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

into Bayesian probability statements:

- probability of a the model being correct (given data): $P(\text{Model} | \text{Data})$

- ... there can be heated debates about ‘pro’ and ‘cons’ of either....

$P(\text{Data}|\text{Theory}) \neq P(\text{Theory}|\text{Data})$

- 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. $P(\text{data}|\text{SMBkg}) < 1\%$) went out to the press and got turned round to:

~~$P(\text{data}|\text{SMBkg}) < 1\% \Rightarrow P(\text{SMBkg}|\text{data}) < 1\% \Rightarrow P(\text{Higgs}|\text{data}) > 99\% !$~~

WRONG!

- easy Example: Theory = fish (hypothesis) .. mamal (alternative)

Data = swim or not swim

$P(\text{swim}|\text{fish}) \sim 100\%$ but $P(\text{fish}|\text{swim}) = ??$

-o.k... but what DOES it say?

The correct frequentist interpretation

we know: $P(\text{Data} \mid \text{Theory}) \neq P(\text{Theory} \mid \text{Data})$

Bayes Theorem:

$$P(\text{Data} \mid \text{Theory}) = P(\text{Theory} \mid \text{Data})$$

$$\frac{P(\text{Theory})}{P(\text{Data})}$$

Frequentists answer ONLY: $P(\text{Data} \mid \text{Theory})$

in reality - we are all interested in $P(\text{Theory} \dots)$

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 .. we cannot say anything about how unlikely our current theory is !

We can define “confidence levels” ... e.g., if $P(\text{data}) < 5\%$, 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 !!

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 ?”**
- **BUT:**
 - **NO Bayesian interpretation exist w/o “prior probability” of the parameter**
 - **where do we get that from?**
 - **all the actual measurement can provide is “frequentist”!**

Bayesian Prior Probabilities

- “flat” prior $\pi(\theta)$ to state “no previous” knowledge (assumptions) about the theory?

➤ often done, BUT WRONG:

- e.g. flat prior in M_{Higgs} \rightarrow not flat in M_{Higgs}^2

➤ Choose a prior that is invariant under parameter transformations

Jeffrey’s Prior \rightarrow “objective Bayesian”:

- “flat” prior in Fisher’s information space

- $\pi(\theta) \propto \sqrt{I(\theta)}$ $(\pi(\theta) \propto \sqrt{\det I(\theta)}$ if several parameters)

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

- $f(x; \theta)$: Likelihood function of θ , probability to observe x for a give parameter θ
- amount of “information” that data x is ‘expected’ to contain about the parameter θ
- **personal remark: nice idea, but “WHY” would you want to do 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

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
- Bayesian approach is expanding
 - now we have the means to do lots of prior comparisons: Computing power/ Markov Chain Monte Carlos

Spares

Binomial Distribution



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

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

▶ each coin: $P(\text{head}) = p$, $P(\text{tail}) = 1 - p$

▶ pick k particular coins → the probability of all having *head* is:

$$P(k \times \text{head}) = P(\text{head}) * P(\text{head}) \dots * P(\text{head}) = P(\text{head})^k$$

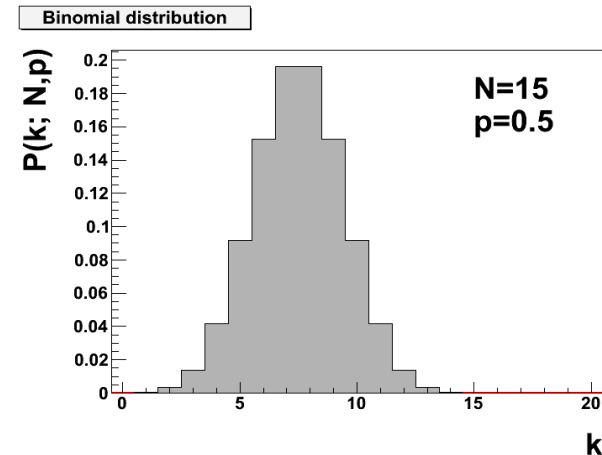
▶ at the same time: probability that all remaining $N-1$ coins land on *tail*

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

▶ That was for k particular coins:

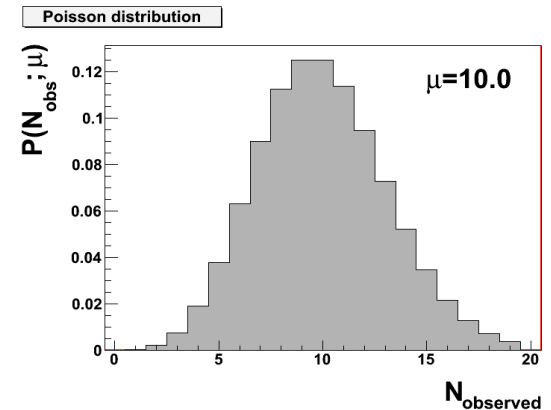
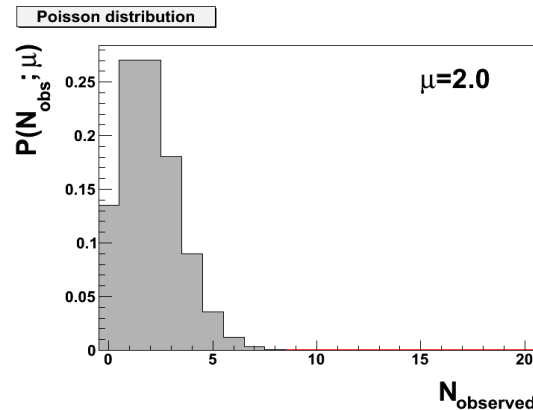
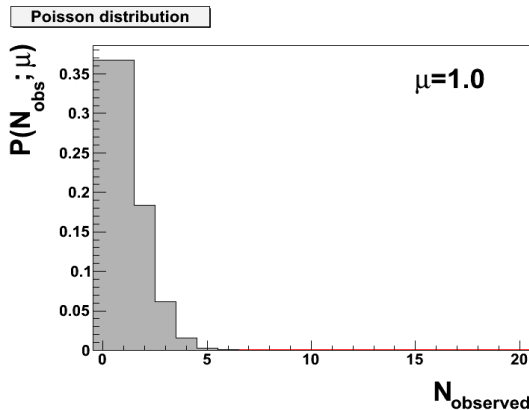
$\binom{N}{k}$ possible permutations for **any** k coins

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



Poisson Distribution

- **Binomial distribution:** Individual events with 2 possible outcomes
- **How about: # counts in radioactive decays during Δt ?**
 - events happen “randomly” but there is no 2nd outcome
 - Δt : continuum \neq “N- discrete trials”
- μ : average #counts in Δt . **What’s the probability for n counts?**
- **Limit of Binomial distribution for $N \rightarrow \infty$ with $Np = \mu$ fixed**
 - **Poisson $P(n) = \frac{\mu^n}{n!} e^{-\mu}$**



- **Expectation value:**

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

- **Variance:**

$$V(n) = \mu$$

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

Gaussian Distribution

- For large μ 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:**

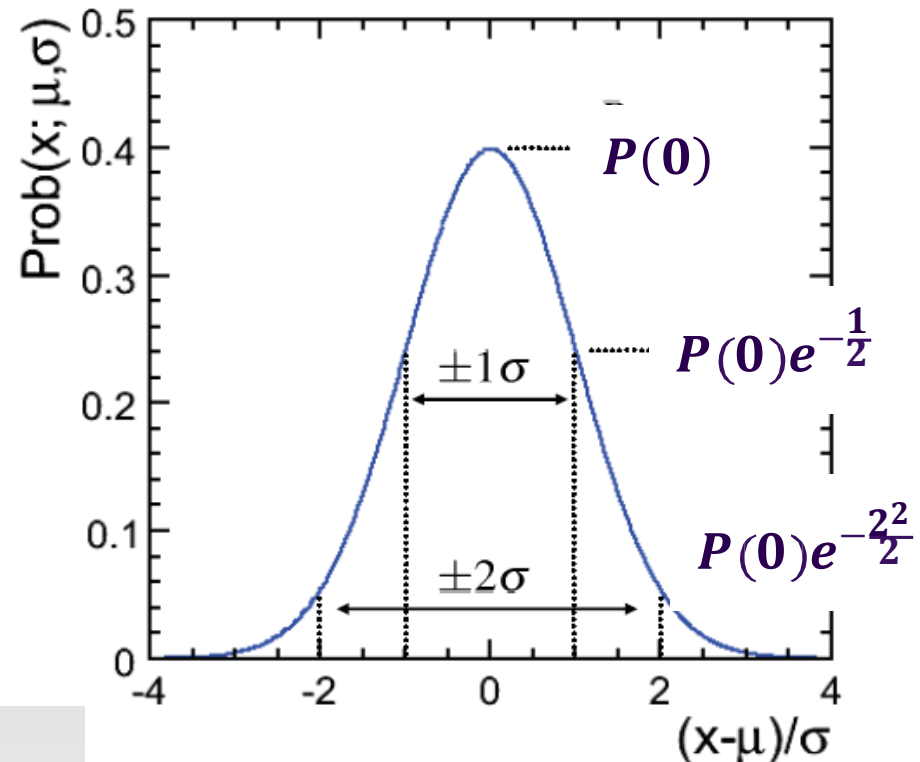
$$E[x] = \mu$$

- **Variance:**

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

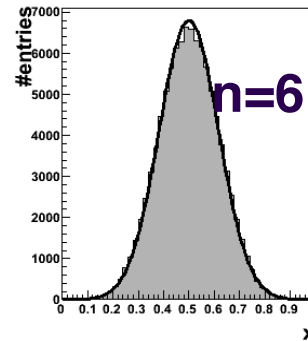
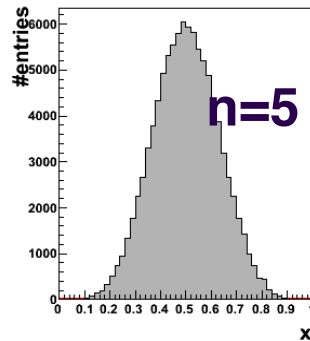
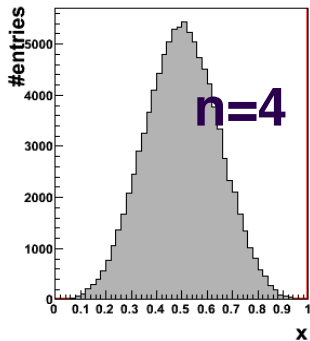
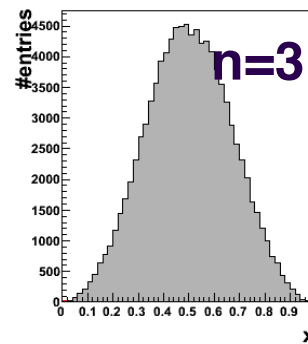
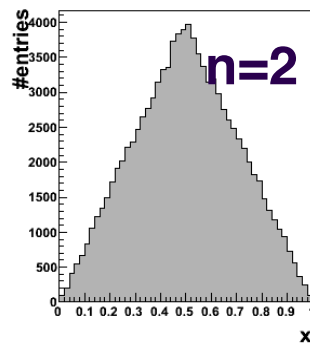
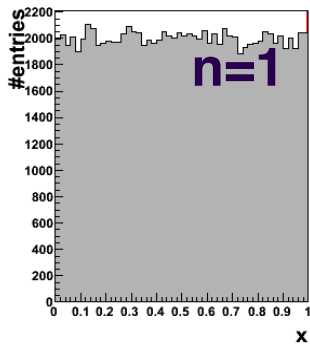
- **Probability content:**

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



Central Limit Theorem

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



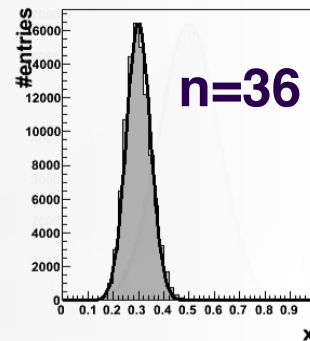
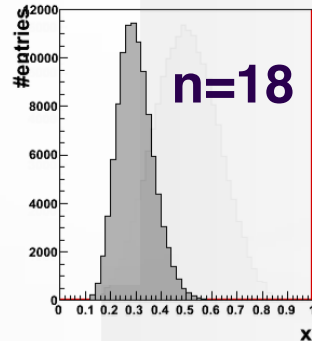
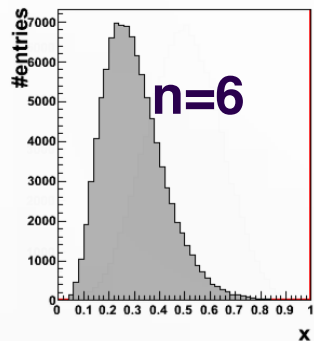
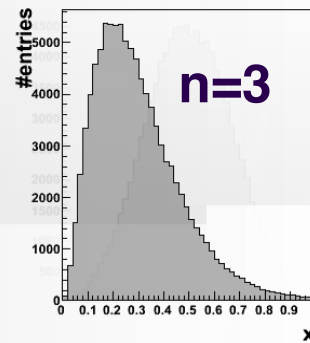
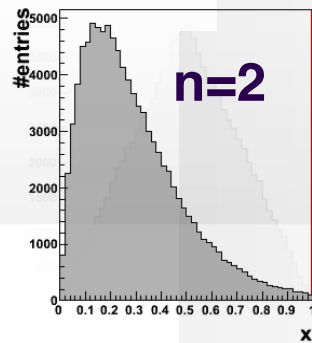
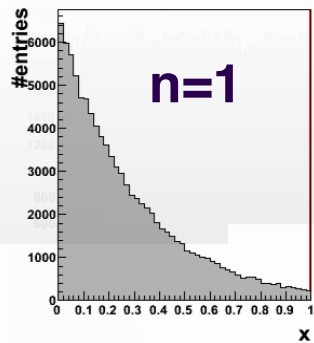
Averaging reduces the error



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

Central Limit Theorem

- 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)
- χ^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:

