# Lecture 2: Introduction to Linear Algebra, Probability and Statistics

Maurizio Pierini (CERN), Thea Aarrestad (ETH)

# Linear Algebra

# Linear Algebra in a Nutshell

- *Scalars: plain numbers (integer, real, etc)*
- *Vectors: ordered arrays of numbers. The i-th element of the vector $x$ is labelled $x_i$*
- *Matrices: ordered table of numbers. The element $A_{ij}$ of the matrix $A$ occupies the i-th row and the the j-th column*
- *Tensors: matrix generalization to more than two dimensions. Similar notation as for matrices ($A_{ijk}$ occupies the i-th position on first dimension, j-th position on second dimension, and k-th position on third dimension)*

$$\boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

$$\begin{bmatrix} A_{1,1} & A_{1,2} \\ A_{2,1} & A_{2,2} \end{bmatrix}$$

# Scalar, Vector, and Matrix operations

- *scalar matrix sum: sum each element of the matrix A to the scalar b*
  $(b + A)_{ij} = A_{ij} + b$

- *scalar matrix product: multiply each element of the matrix A by the scalar b*
  $(b \cdot A)_{ij} = bA_{ij}$

$$2 \cdot \begin{bmatrix} 10 & 6 \\ 4 & 3 \end{bmatrix} = \begin{bmatrix} 2 \cdot 10 & 2 \cdot 6 \\ 2 \cdot 4 & 2 \cdot 3 \end{bmatrix}$$

- *Vector transpose: turn a (column) vector $x$ in a (row) vector $x^{\mathrm{T}}$*

$$\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{bmatrix}^{\mathrm{T}} = [\, x_1 \; x_2 \; \dots \; x_m \,].$$

- *Matrix transpose: invert $A_{ij}$ with $A_{ji}$ for each $i$ and $j$*

- *Matrix sum: $(A + B)_{ij} = A_{ij} + B_{ij}$*

# Scalar, Vector, and Matrix operations

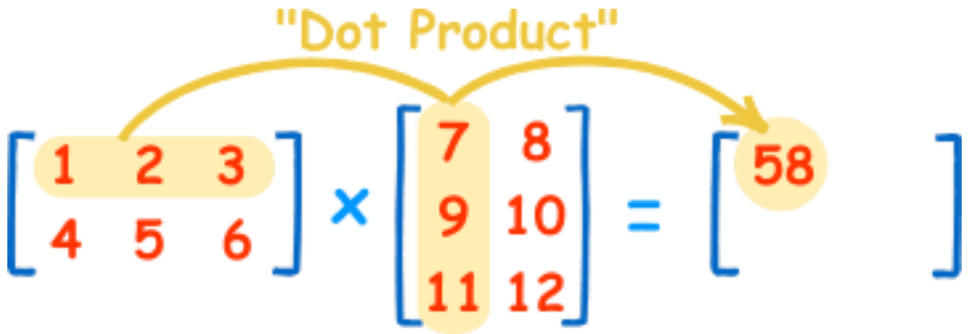
- *(same dimension) vector inner product* $xy = \sum x_i y_i$ *returns a scalar*

$$\mathbf{a} = \begin{bmatrix} v \\ w \end{bmatrix} \qquad \mathbf{b} = \begin{bmatrix} x \\ y \\ z \end{bmatrix}$$

- *vector outer product* $(x^\top y)_{ij} = x_i y_j$ *returns a matrix*

$$\mathbf{C} = \mathbf{ab'} = \begin{bmatrix} v*x & v*y & v*z \\ w*x & w*y & w*z \end{bmatrix}$$

- *Matrix product:* $(A \times B)_{ij} = \sum_k A_{ik} B_{kj}$




- *Hadamard product: multiply elements in corresponding position (for matrices of the same sign)*
  $(A \quad B)_{ij} = A_{ij} B_{ij}$

# Scalar, Vector, and Matrix operations

- Matrix product
  - is distributive $A(B+C) = AB + AC$
  - is associative $A(BC) = (AB)C$
  - is not commutative $AB \neq BA$
- Transpose of a product: $(AB)^\top = B^\top A^\top$
- Inverse of a matrix $A^{-1} : A^{-1}A = \mathbb{I}$, where $\mathbb{I}$ is the identity matrix
  - Not all matrices have an inverse
  - When they do, one can solve a set of linear equation $Ax = b \implies x = A^{-1}b$

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

# Vector Norms

- A norm is a function f of a vector x such that
  - $f(x) = 0 \implies x = 0$
  - $f(x + y) \leq f(x) + f(y)$
  - $\forall \alpha \in \mathbb{R} f(\alpha x) = |\alpha| f(x)$
- In many occasions (e.g., when defining loss functions for a regression) we will deal with $L^p$ norms

$$||x||_p = \left( \sum_i |x_i|^p \right)^{1/p}$$

  - for p=2 we obtain the Euclidean norm
  - ofter p=1 is used, because it grows equally faster near and far from zero (useful for minimisation convergence)

# Special kinds of matrices

- *Diagonal: matrix with all elements out of diagonal equal to zero. Notation:* $diag(v)$ *= diagonal matrix with elements on diagonal given by vector v*
- *Symmetric: a matrix equal to its transpose (*$A = A^\top$*), i.e., with* $A_{ij} = A_{ji}$
- *Orthogonal:* $A^{-1} = A^\top$*. An orthogonal matrix has rows and columns which are mutually orthonormal, i.e., the corresponding vectors are orthogonal and with norm = 1*

# Eigendecomposition

- An eigenvector $v$ of $A$ is such that $Av = \lambda v$, where $\lambda$ is the eigenvalue of $A$. Applying $A$ on $v$ just rescales it by $\lambda$

- One can demonstrate that $A = V\,\mathrm{diag}(\lambda)\,V^{-1}$, where $V$ is the matrix constructed such that the i-th column is the eigenvector $v^{(i)}$ and $\mathrm{diag}(\lambda)$ is the diagonal matrix with the eigenvalue $\lambda^{(i)}$ at the $(i,i)$ position

- Any eigedecomposed matrix can be inverted. If one defines $A^{-1} = V\Lambda^{-1}V^{-1}$, with $(\Lambda^{-1})_{i,i} = \frac{1}{\lambda^{(i)}}$ and $(\Lambda^{-1})_{i,j} = 0$ for $i \neq j$, it is easy to show that

$$A^{-1}A = V\Lambda^{-1}V^{-1}V\,\mathrm{diag}(\lambda)\,V^{-1} = V\Lambda^{-1}\,\mathrm{diag}(\lambda)\,V^{-1} = VV^{-1} = \mathbb{I}$$

- For a real symmetric matrix, the eigencomposition simplifies to $A = V\,\mathrm{diag}(\lambda)\,V^{\top}$

# Important quantities

- *Determinant:*
  - *a scalar function of the entries of a matrix such that*
    - $\det(\mathbb{I}) = 1$
    - *The exchange of two rows multiplies the determinant by –1*
    - *Multiplying a row by a number multiplies the determinant by this number*
    - *Adding a multiple of one row to another row does not change the determinant*
  - *For a diagonalisable matrix, it is the product of the eigenvalues*

$$\begin{vmatrix} a & b \\ c & d \end{vmatrix} = ad - bc,$$

$$\begin{vmatrix} a & b & c \\ d & e & f \\ g & h & i \end{vmatrix} = aei + bfg + cdh - ceg - bdi - afh.$$

- *Trace:* $\mathrm{Tr}(A) = \sum_i A_{ii}$
  - *Invariant under transpose operation:* $\mathrm{Tr}(A) = \mathrm{Tr}(A^\top)$
  - *Invariant under cycling permutation:* $\mathrm{Tr}(AB) = \mathrm{Tr}(BA)$, $\mathrm{Tr}(ABC) = \mathrm{Tr}(CAB) = \mathrm{Tr}(BCA)$, *etc.*

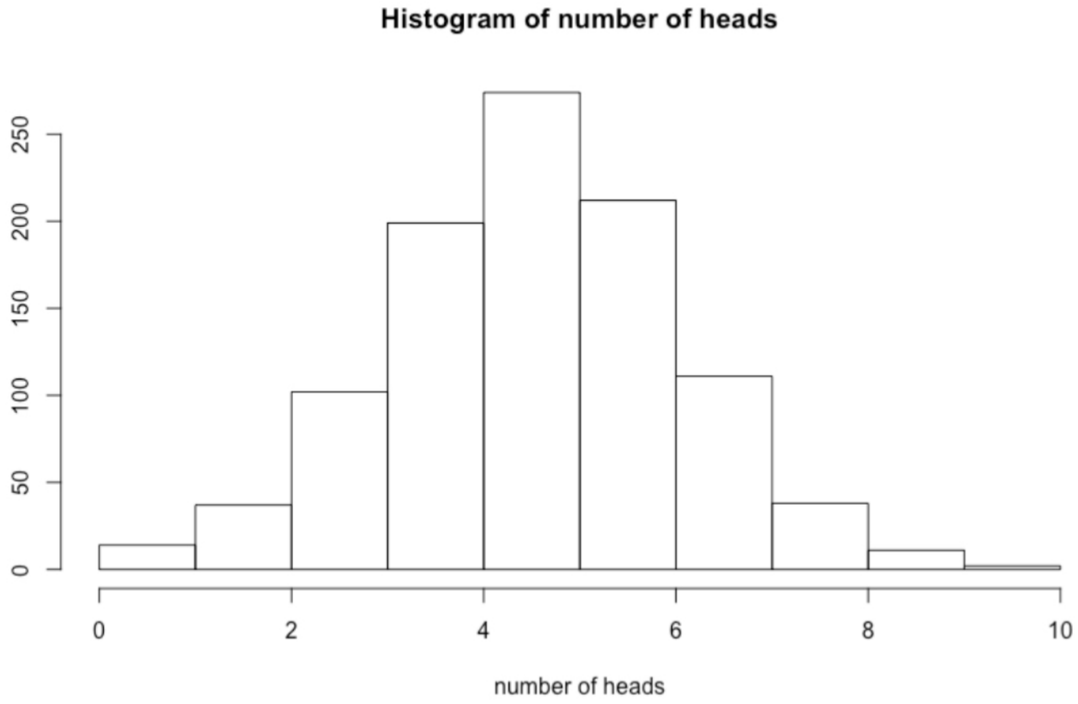
# Probability


"I wish we hadn't learned probability 'cause I don't think our odds are good."

# Uncertainty and Probability

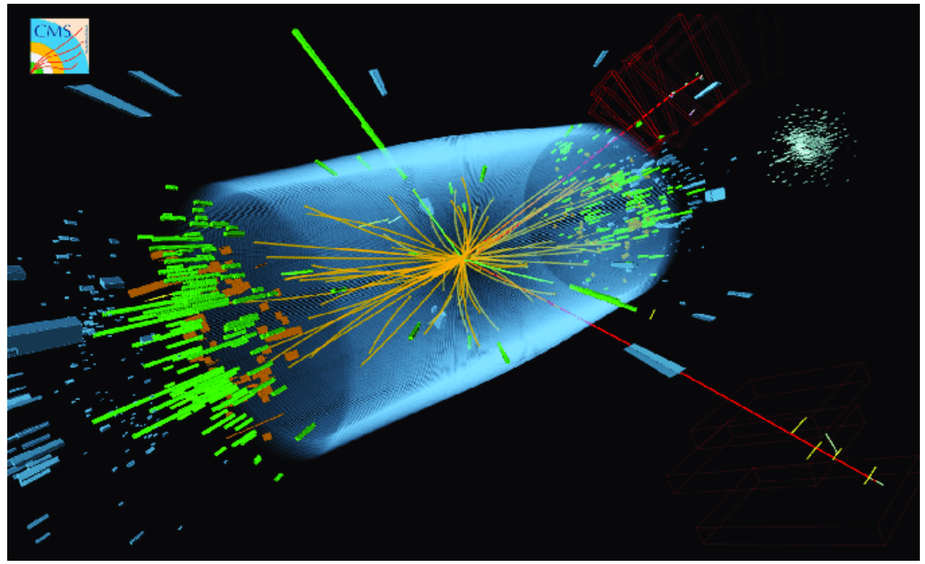
- *Machine Learning, like any branch of Data Science, is not deterministic*
  - *Inherent stochasticity: One tries to learn from a dataset, and datasets come with statistical noise*
  - *Incomplete observability: even in absence of stochastic noise, one might be in an uncertain status for lack of knowledge due to partial observation*
  - *Incomplete modeling: we might miss some of the observed observation, that for various reasons might have been removed from the original dataset*
- *In presence of uncertainty, we need to approach the problem following probability theory and statistics*

*Example: 10 coin tossing will give ~ 5 heads, but not always EXACTLY 5 heads*



# Two schools of statistics

- *Frequentist school: the probability of a certain event is the frequency for that event to occur in the limit of infinite statistics.*
  - *This point of view is suitable for repeatable experiments (e.g., probability to produce a Higgs boson at the LHC)*
  - *It has some circularity (it requires each repetition to have equiprobable outcome, which assumes the concept of equal probability to define probability)*
- *Bayesian school: the probability of a certain event measures the degree of belief that the observed associates to a certain outcome*
  - *It applies well to events which are not repeatable. E.g., what is the probability that it will rain tomorrow?*
  - *It implies a subjectivity (which boils down to specifying a-priori belief on a given event before the observation)*




# Probability Distribution

- *A random variable a variable that can take on different values randomly*
  - *Can be a scalar or a vector. Can be discrete or continuous*
- *A probability distribution is a description of how likely a random variable or set of random variables is to take on each of its possible states.*

**The probability density function (pdf), aka probability mass function (pmf) for a discrete quantity**

- The domain of $P$ must be the set of all possible states of x.
- $\forall x \in \mathrm{x}, 0 \leq P(x) \leq 1$. An impossible event has probability 0, and no state can be less probable than that. Likewise, an event that is guaranteed to happen has probability 1, and no state can have a greater chance of occurring.
- $\sum_{x \in \mathrm{x}} P(x) = 1$. We refer to this property as being **normalized**. Without this property, we could obtain probabilities greater than one by computing the probability of one of many events occurring.

# Probability Distribution

- *A random variable a variable that can take on different values randomly*
  - *Can be a scalar or a vector. Can be discrete or continuous*
- *A probability distribution is a description of how likely a random variable or set of random variables is to take on each of its possible states.*

**The probability density function (pdf), aka probability mass function (pmf) for a continuous quantity**

- The domain of $p$ must be the set of all possible states of x.
- $\forall x \in \mathrm{x}, p(x) \geq 0$. Note that we do not require $p(x) \leq 1$.
- $\int p(x)dx = 1$.

# Probability Distribution

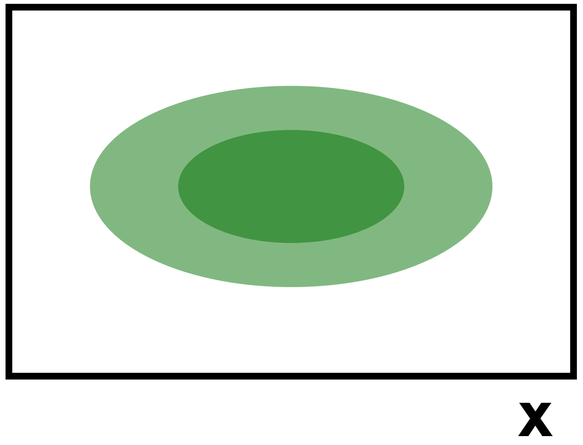
- *EXAMPLE: Uniform distribution of discrete quantity, the pdf for equiprobable outcomes:* $p(x = x_i) = \frac{1}{k} \text{for } i \in [0, k-1]$*. This definition meets all the desired properties. In particular*

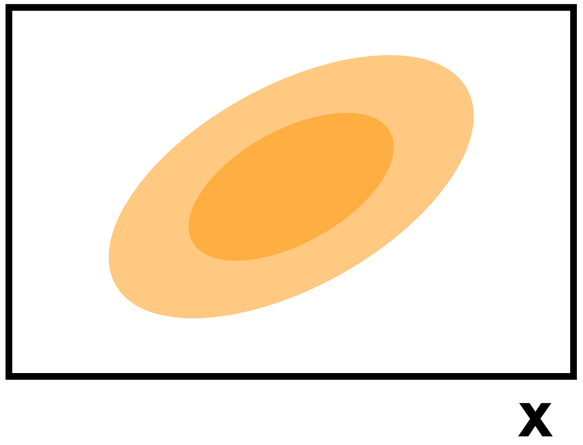
$$0 \leq p(x = x_i) \leq 1 \quad \textit{and} \quad \sum_{i=0}^{k-1} p(x_i) = \sum_{i=0}^{k-1} \frac{1}{k} = 1$$

- *This generalises to the case of a continuous quantity defined between* $a$ *and* $b$ *as* $p(x) = \frac{1}{b-a}$

# Joint Probability and Correlation

- *The joint probability of two quantities $x$ and $y$ is the pdf that describe the probability of all the possible outcomes, expressed as $(x, y)$ pairs*
  - *If $x$ and $y$ are independent, $p(x, y) = q(x)t(y)$, where $q(x)$ and $t(y)$ are the 1D pdfs of $x$ and $y$*
  - *If $p(x, y) \neq q(x)t(y)$, it means that the outcome on $x$ depends on $y$ and vice versa. $x$ and $y$ are said to be correlated*



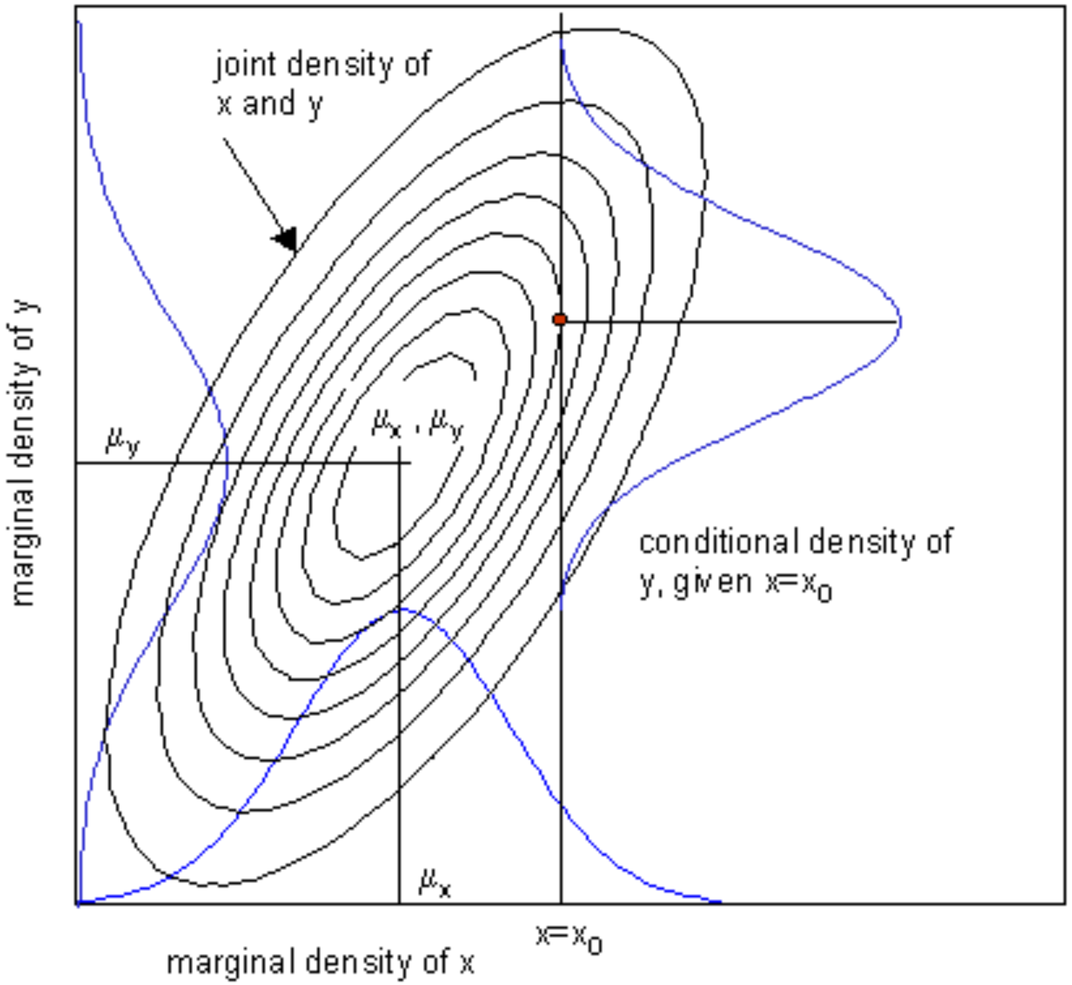



# Marginal Probability

- *In some cases, one knows the joint probability distribution of two quantities $x$ and $y$ $p(x,y)$ but being interested to a statement on $x$ regardless of $y$*

- *in this case, one needs to derive the marginal pdf from the joint pdf*

$$p(x) = \int p(x,y)dy$$



# Conditional Probability

- In certain cases, the value of $y$ is known and one wants to restrict the prediction on $x$ to the fact that that value of $y$ occurred. This means deriving the conditional probability from the joint probability

$$p(x \mid y = \hat{y}) = \frac{p(x, y = \hat{y})}{\int p(x, y = \hat{y}) dx} = \frac{p(x, y = \hat{y})}{p(y = \hat{y})}$$

- The marginal probability in the denominator guarantees that the conditional probability is $\in [0,1]$

- Inverting the relation above, we derive the probability chain rule

$$p(x, y = \hat{y}) = p(x \mid y = \hat{y})(y = \hat{y})$$

or, with a lighter notation and for any value of $y$

$$p(x, y) = p(x \mid y)(y)$$

# Chain rule

- We saw how to derive the marginal and conditional probabilities from the joint probability
- Often, it is useful to do the opposite: express the joint probability as a function of marginal and conditional probabilities
  - The main advantage is to break down an N-dim function into many 1-dim functions
  - Take the example of three quantities. We have

$$P(x_1, x_2, x_3) = P(x_1, x_2 \,|\, x_3) P(x_3) \quad \text{and} \quad P(x_1, x_2) = P(x_1 \,|\, x_2) P(x_2)$$

which imply

$$P(x_1, x_2, x_3) = P(x_1 \,|\, x_2, x_3) P(x_2 \,|\, x_3) P(x_3)$$

- In general

$$P(x_1, \ldots, n_n) = P(x_1) \prod_{i=2}^{n} P(x_i \,|\, x_1, \ldots, x_{i-1})$$

# More on Independence

- We saw that two quantities are said to be *independent* when their joint probability factorizes into the product of two 1-dim pdf

$$p(x, y) = p(x)p(y)$$

- We can generalise this concept introducing *conditional independence* of

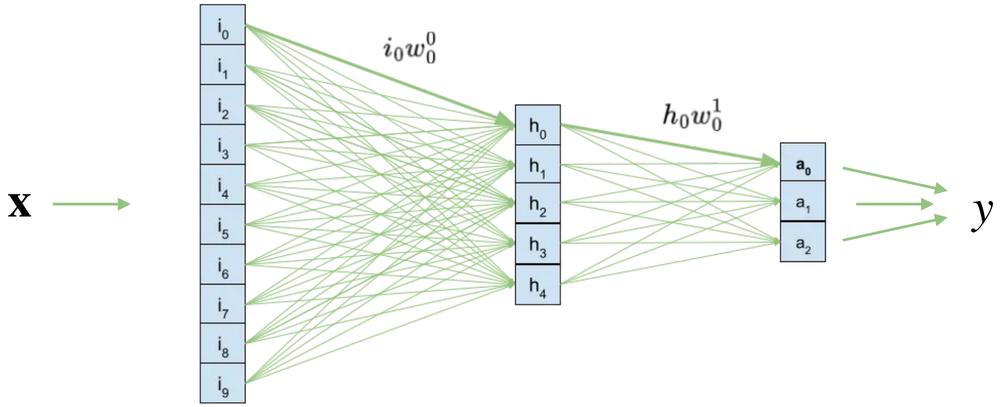
$$p(x, y, z) = p(x, y \mid z)p(z) = p(x \mid z)p(y \mid z)p(z)$$

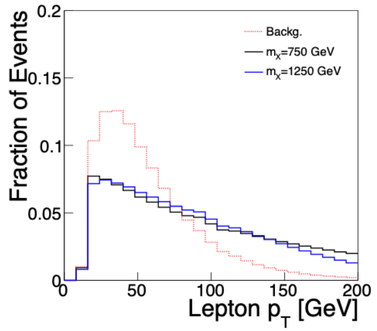
or $$p(x, y \mid z) = p(x \mid z)p(y \mid z)$$

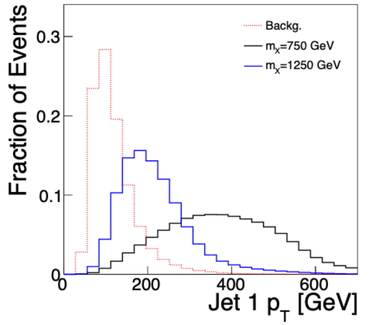
- x and y are not independent, but their correlation is only driven by the common value of z

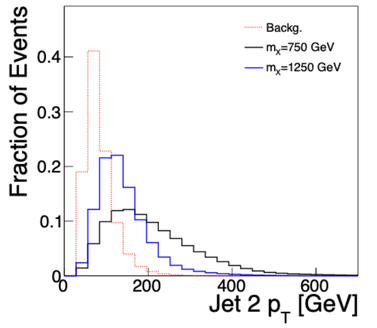
# Why is this important to us?

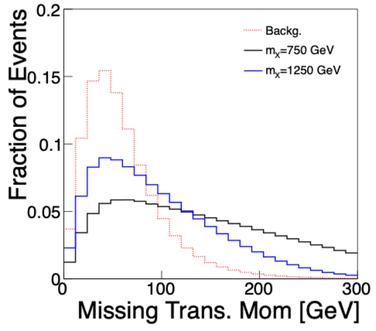
- *Supervised learning is the learning of a conditional pdf. Given some data* $\mathbf{x}$ *and ground truth* $y$ *we want to learn* $p(y|\mathbf{x})$
- *When solving certain tasks with ML, you are often confronted to learn the pdf of your data*
  - *At the LHC, you might want to know if your data were produced by Higgs bosons or by some new kind of particle, for which you don't know the mass.*
  - *You can assume a mass and train a network for every possible mass value*
  - *Or you can learn the classifier parametrically to the assumed value* $p_H(D)$ *vs* $p_X(D|m_X)$
- *This is called parametric learning (see https://arxiv.org/abs/1601.07913)*



# Title Text

**DEFINITION:**

FOR A GENERIC $P(k|\alpha)$

$$E[k|\alpha] = \frac{\sum_k k P(k|\alpha)}{\sum P(k|\alpha)}$$

FOR A GENERIC $P(x|\alpha)$

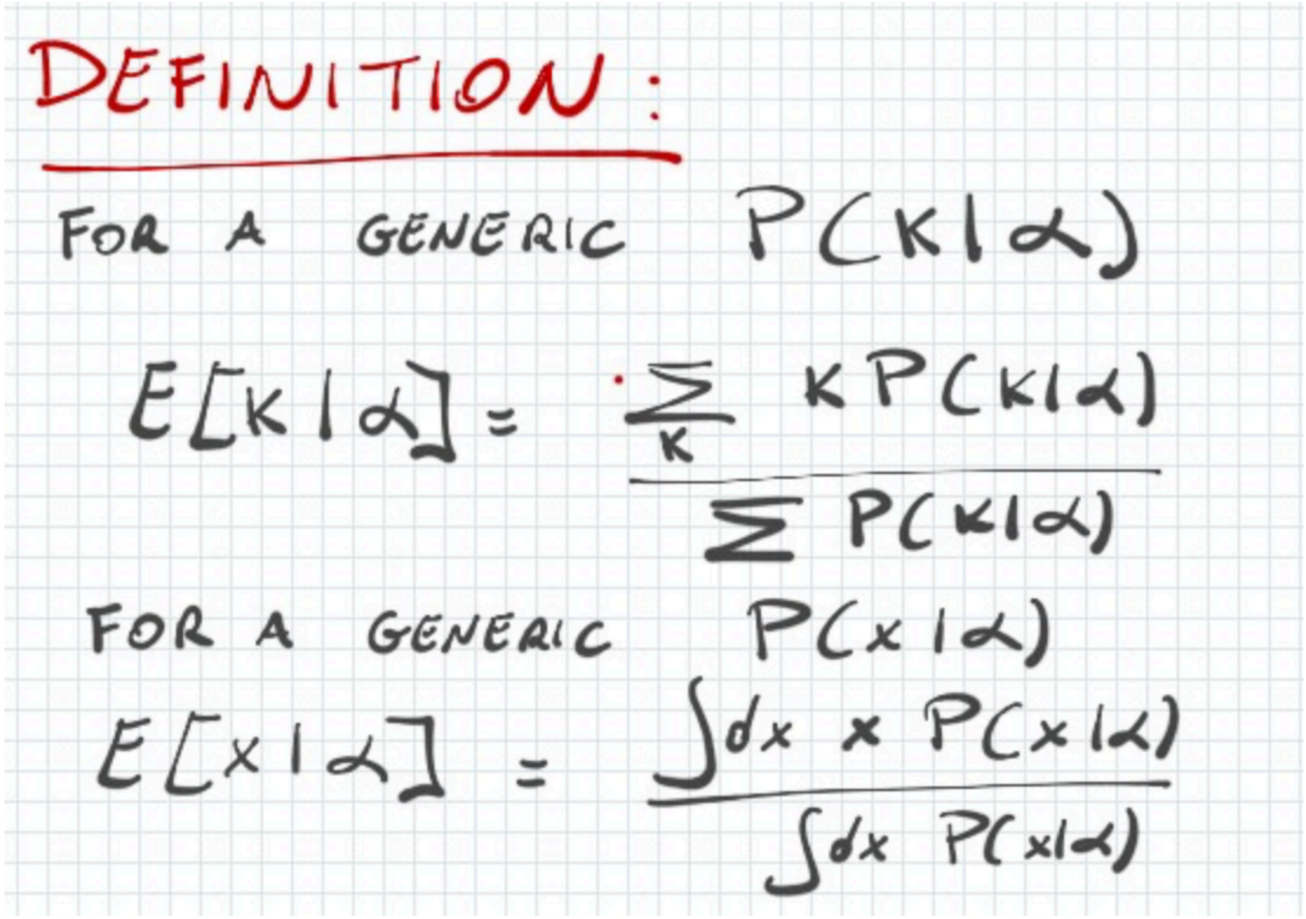
$$E[x|\alpha] = \frac{\int dx \; x \; P(x|\alpha)}{\int dx \; P(x|\alpha)}$$

# Expectation Value

- *Expectation value: mean value (aka average value) of a quantity f(x) with respect to the pdf of x*
  - *Expectation value is linear:*

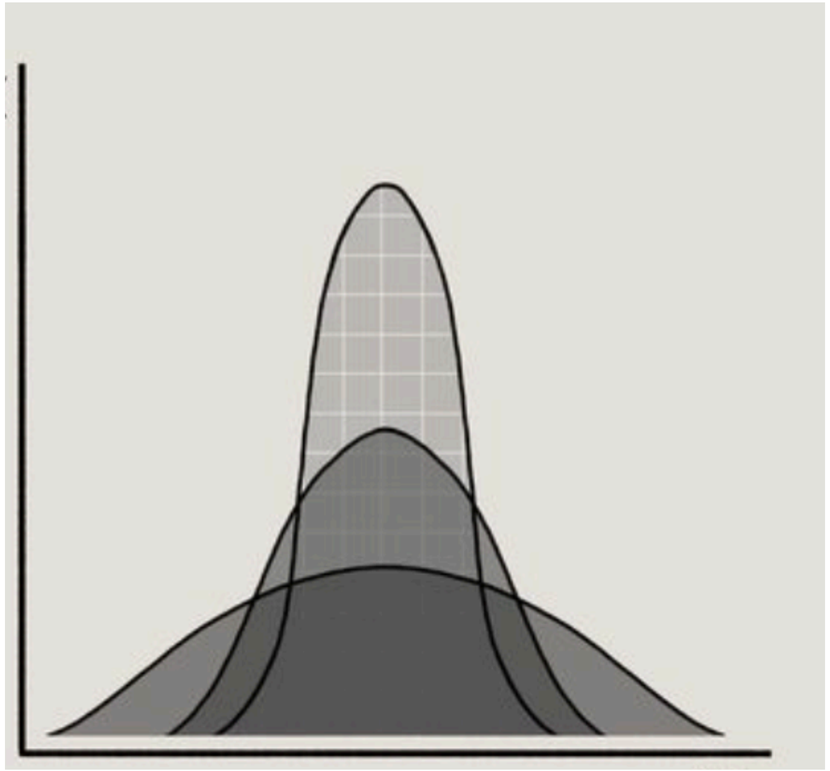
$$E_X[\alpha\, f(x) + \beta\, g(x)] =$$

$$\alpha\, E_X[f(x)] + \beta\, E_X[g(x)]$$



# Variance

- E[x] is not enough to characterize a distribution
  - distributions with same E[x] can be very different
- It is convenient to have a measure of the dispersion of points around E[x]
  - One typically introduced the variance (aka mean square error)



$$Var[x] = E[(x - E[x])^2] = E[x^2] - E[x]^2$$

- The square root of the variance is called standard deviation

# Covariance

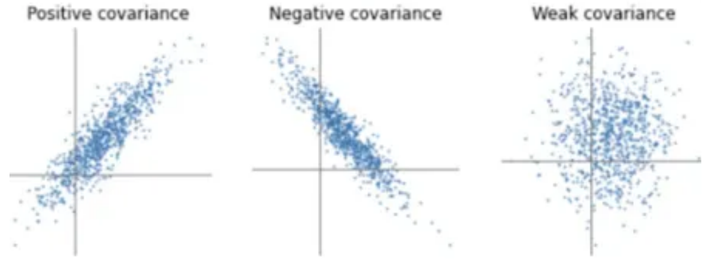
- *Covariance quantifies the correlation between two quantities*

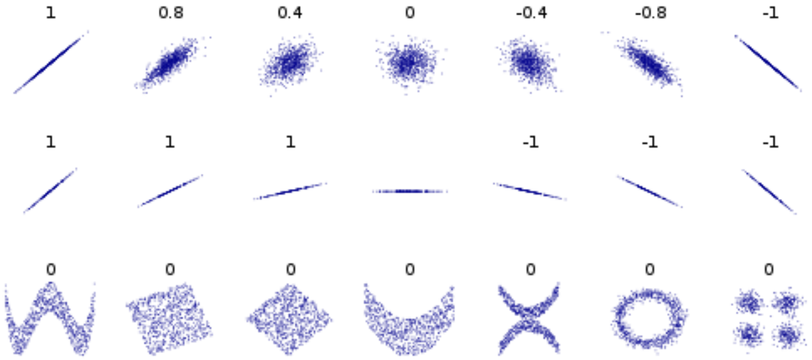
$$Cov(f(x)g(x)) = E_X[(f(x) - E_X[f(x))(g(x) - E_X[g(x))]$$

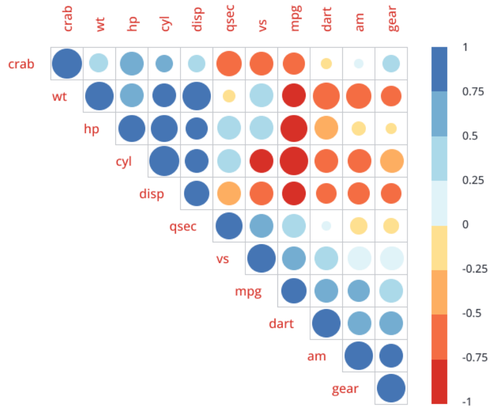
- *For $f(x) = g(x)$ it returns the variance*
- *NOTICE: the covariance measures a linear dependence. Two quantities can have covariance = 0 with our being independent (independence is a stronger requirement)*
- *Often one introduces the linear correlation coefficient*

$$\rho(f(x), g(x)) = \frac{Cov(f(x), g(x))}{\sqrt{Var(f(x))}\sqrt{Var(g(x))}}$$

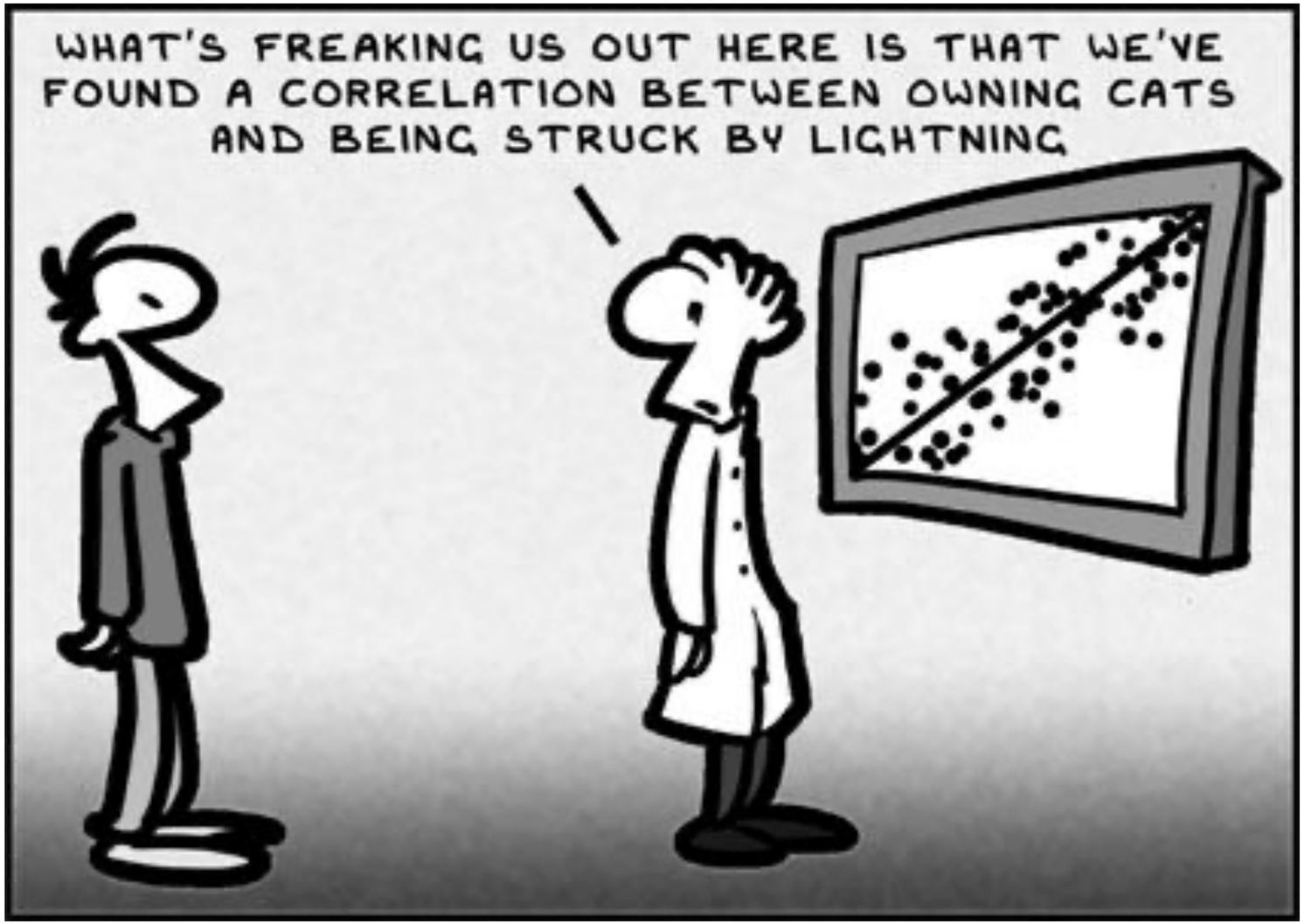
- *For a vector if quantities, covariant and correlation matrices can be built*



# Statistics


WHAT'S FREAKING US OUT HERE IS THAT WE'VE FOUND A CORRELATION BETWEEN OWNING CATS AND BEING STRUCK BY LIGHTNING

# Bernoulli's process

- You pick k items out of a bag with N items and you ask a yes/no question
  - is my ball red?
- Let's call
  - p: probability that the answer is Yes
  - q = 1-p : probability that the answer is No

$$P(k=1)=p \qquad P(k=0)=1-p=q$$

$$\mathbb{E}[k]=\frac{p\cdot 1+1\cdot 0}{p+q}=p$$

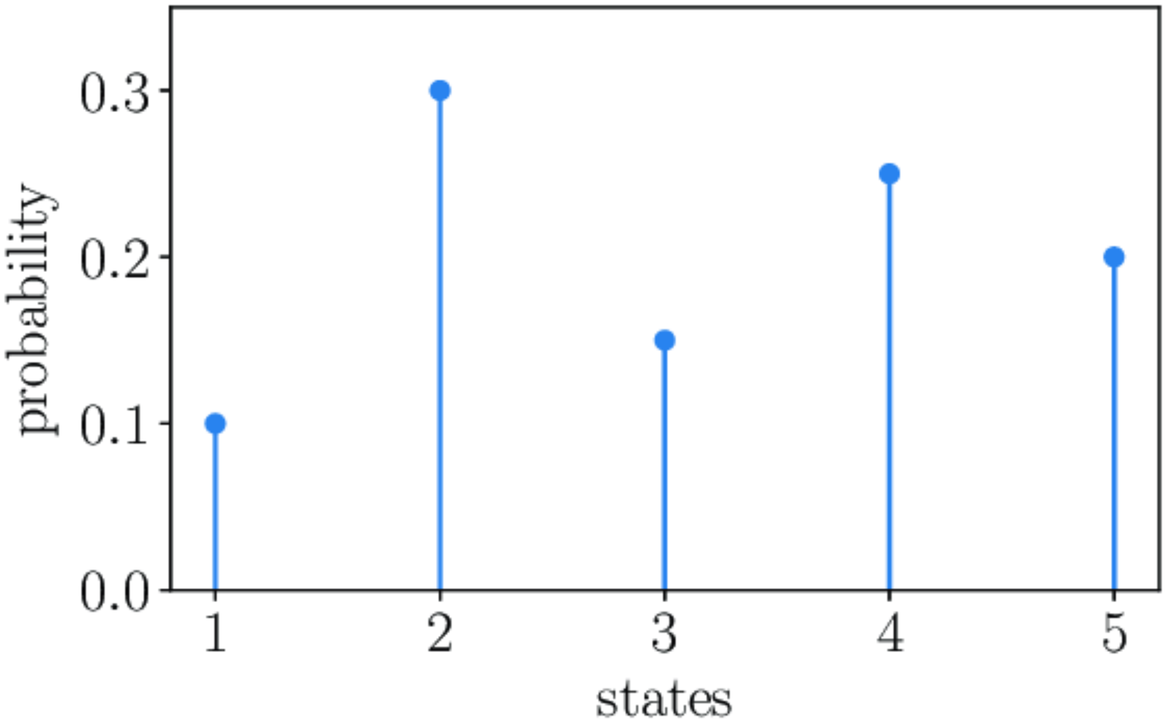
$$Var(k)=\mathbb{E}[(k-\mathbb{E}[k])^2]=\mathbb{E}[k^2]-2\mathbb{E}[k]^2+\mathbb{E}[k]^2$$

$$=\mathbb{E}[k^2]-\mathbb{E}[k]^2=\frac{1^2\cdot p+0^2\cdot q}{p+q}-p^2$$

$$=p-p^2=p(1-p)=pq$$

# Categorical Distribution

- *Bernoulli distribution often generalized to categorical distribution for more than two possible values. In this case, for n possible outcomes, n-1 values of $p$ are given (the n-th being $1 - \sum_{i=1}^{n-1} p_i$ )*



# Binomial distribution

- *Probability of k out of N items being Y*
- *Probability of one out of two items being Y [order not important!]*
- *Probability of k out of N items being Y [order not important!]*

$$P(k=1\,|\,N=1)=p$$

$$P(k=1\,|\,N=2)=\frac{pq+qp}{p^2+pq+qp+q^2}=2pq$$

$$P(k\,|\,N)=\frac{n!}{k!(N-k)!}p^k q^{N-k}$$

Probability that the selected event is obtained k times out of the total of N trials.

Probabability that something other than the chosen event will occur in all the other trials.

The "combination" expression, which is the permutation relationship (the number of ways to get k occurrences of the selected event) divided by k ! (the number of different orders in which the k events could be chosen, assuming they are distinguishable).

# Limit of rare events

- *For $N \to \infty$ with $p \to 0$ so that $Np$ stays finite and positive, the Binomial distribution takes the form of a Poisson distribution*

$$P(K|N,p) = \frac{N!}{(N-K)!\,K!}\,p^K(1-p)^{N-K} =$$

$$= \frac{N!}{(N-K)!\,N^K}\,\frac{(Np)^K}{K!}\,/1-p)^N\,(1-p)^{-K} =$$

LET'S DEFINE $\lambda = N \cdot p$

$$= \frac{\lambda^K}{K!}\,\frac{N!}{(N-K)!\,N^K}\left(1-\frac{\lambda}{N}\right)^N\left(1-\frac{\lambda}{N}\right)^{-K} \to$$

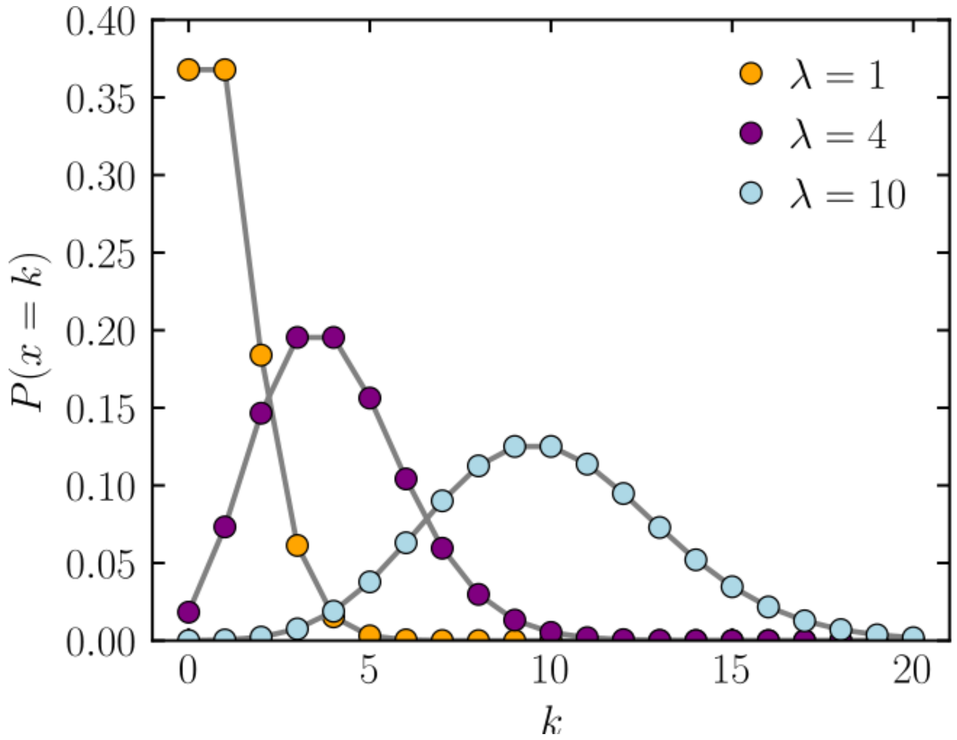
$\frac{N!}{(N-K)!\,N^K} \xrightarrow{N\to\infty} 1$, $\left(1-\frac{\lambda}{N}\right)^N \xrightarrow{N\to\infty} e^{-\lambda}$, $\left(1-\frac{\lambda}{N}\right)^{-K} \xrightarrow{N\to\infty} 1$

$$\xrightarrow[N\to\infty]{} \frac{\lambda^K}{K!}\,e^{-\lambda}$$

# Poisson Distribution

$$f(k;\lambda) = \Pr(X=k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

- *k is the unknown (the outcome of our experiment counting). It takes integer values by construction*
- *λ is the parameter determining the distribution shape and it is related to the most probable outcome of our counting experiment. It might be an integer, but in general it is a real number*



# Poisson Expectation Value

$$E[k|\lambda] = \frac{0 \cdot P(0|\lambda) + 1 \cdot P(1|\lambda) + 2 \cdot P(2|\lambda) + \dots}{P(0|\lambda) + P(1|\lambda) + P(2|\lambda) + \dots} =$$

$$= \frac{\sum_{k=0}^{\infty} \lambda P(k|\lambda)}{\sum_{k=0}^{\infty} P(k|\lambda)} =$$

$$= \frac{e^{-\lambda} \sum_{k=0}^{\infty} \frac{k \lambda^k}{k!}}{e^{-\lambda} \sum \frac{\lambda^k}{k!}} = \frac{\lambda \sum_{k=1}^{\infty} \frac{k \cdot \lambda^{k-1}}{k \cdot (k-1)!}}{e^{\lambda}} = \frac{\lambda e^{\lambda}}{e^{\lambda}} = \boxed{\lambda}$$

# Poisson Variance

$$E[(k - E[k])^2] = E[k^2] - E[k]^2 = \lambda^2 + \lambda - \lambda^2 = \lambda$$

$$E[k^2] = \frac{\sum_{k=0}^{\infty} \frac{e^{-\lambda}\lambda^k k^2}{k!}}{\underbrace{\left(\sum_{k=0}^{\infty} \frac{\lambda^k}{k!}\right) e^{-\lambda}}_{e^{\lambda}}} = \sum_{k=1}^{\infty} \frac{e^{-\lambda}\lambda^k k}{(k-1)!} =$$

$$= \lambda e^{-\lambda} \left[ \sum_{k=1}^{\infty} \frac{\lambda^{k-1}(k-1)}{(k-1)!} + \underbrace{\sum_{k=1}^{\infty} \frac{\lambda^{k-1}}{(k-1)!}}_{e^{\lambda}} \right] =$$

$$= \lambda e^{-\lambda} \left[ \lambda \underbrace{\sum_{k=2} \frac{\lambda^{k-2}}{(k-2)!}}_{e^{\lambda}} + e^{\lambda} \right] = \lambda(\lambda+1) = \boxed{\lambda^2 + \lambda}$$

# The limit of large λ

$$\ln(P(k|\lambda)) = \ln\left(\frac{\lambda^k e^{-\lambda}}{k!}\right) \simeq \ln\left(\frac{\lambda^k e^{-\lambda}}{k^k e^{-k}\sqrt{2\pi k}}\right) =$$

STIRLING'S APPROXIMATION $\quad x! \simeq x^x e^{-x}\sqrt{2\pi x}$

$$= k\ln\lambda - \lambda - k\ln k + k - \ln\sqrt{2\pi k} =$$

$$= \dots = -\frac{y^2}{2\lambda} + \frac{y^3}{6\lambda^2} - \ln\sqrt{2\pi(y+\lambda)} \simeq$$

$y = k - \lambda$ WITH $\lambda \to \infty$, $k \sim O(\lambda)$ $\Rightarrow y \ll \lambda$

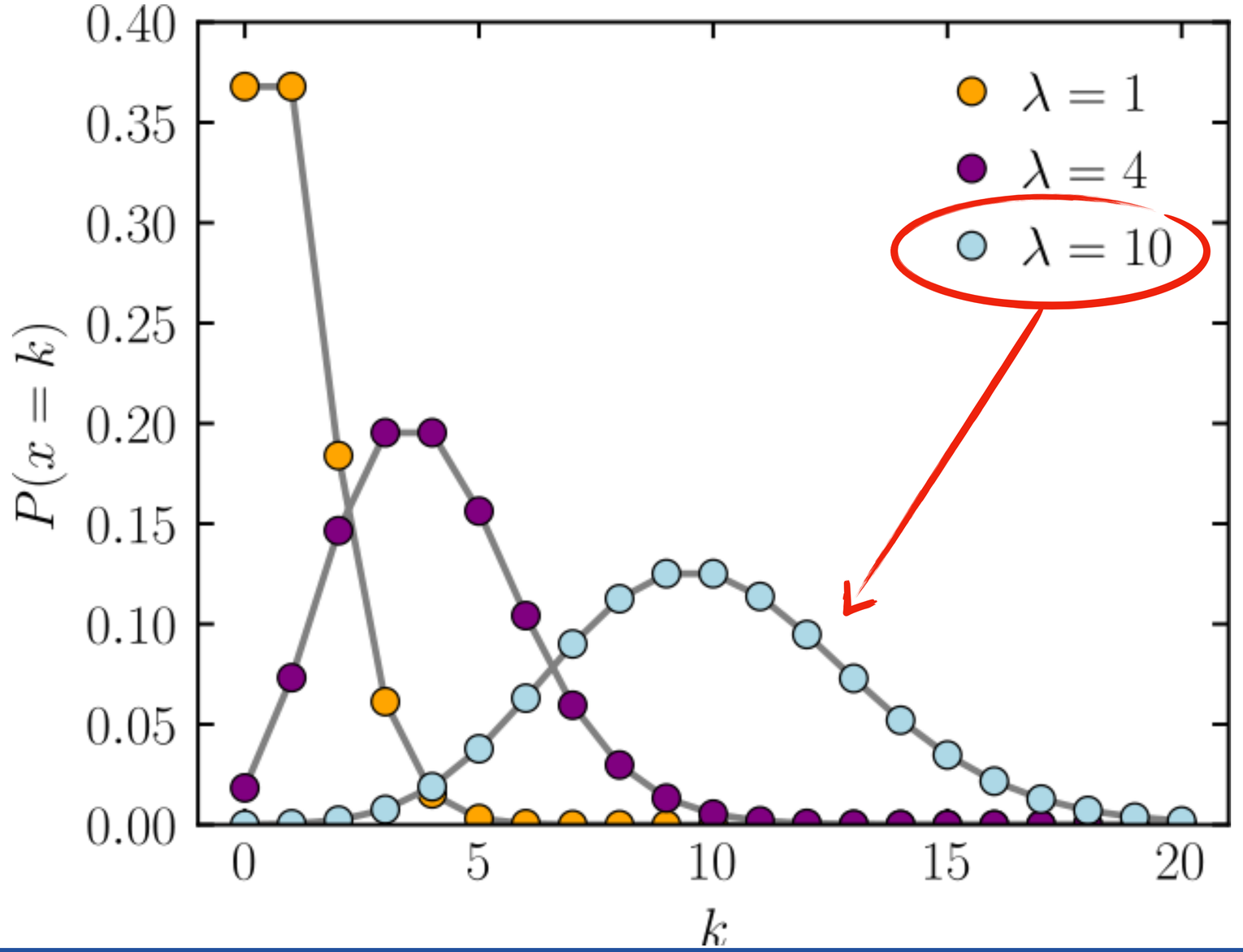
$$\simeq -\frac{y^2}{2\lambda}$$

$$P(y|\lambda) = e^{-\frac{y^2}{2\lambda}}$$

$\ln(1+\varepsilon) \simeq \varepsilon - \varepsilon^2 + \frac{\varepsilon^3}{3} \dots$

GAUSSIAN

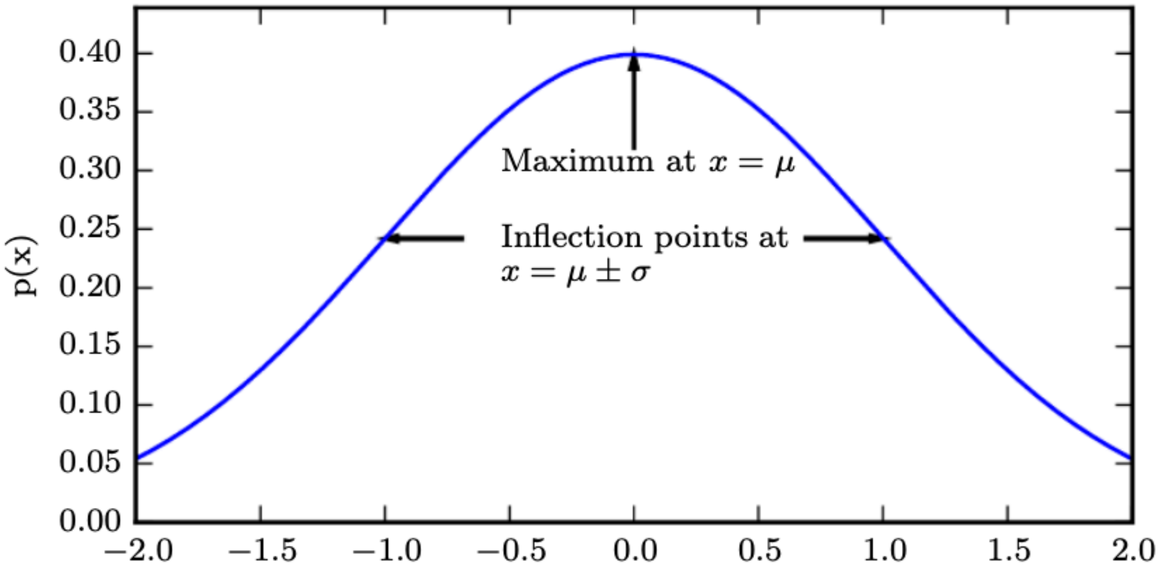
# How Large is Large?



# Gaussian

- One of the most frequently encountered distributions
  - Rules many stochastic problems, because it is the limit distribution of other pdfs
  - Two parameters
    - $\mu$ determines the position of the mean/maximum
    - $\sigma$ determines the width around the mean

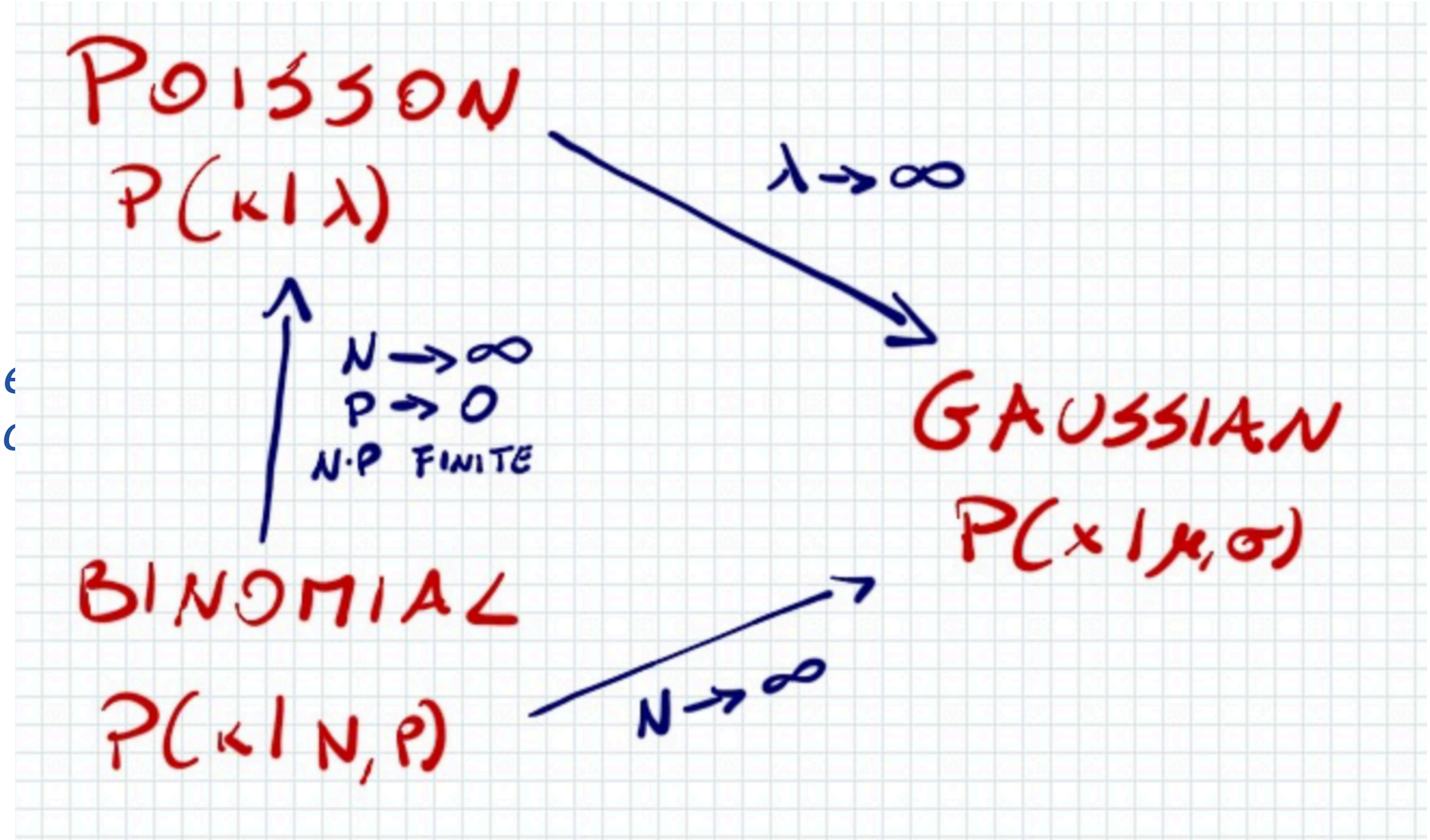
$$G(x \mid \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$




$$\mathbb{E}[x] = \mu \qquad Var(x) = \sigma^2$$

# Gaussian as a large-trial limit

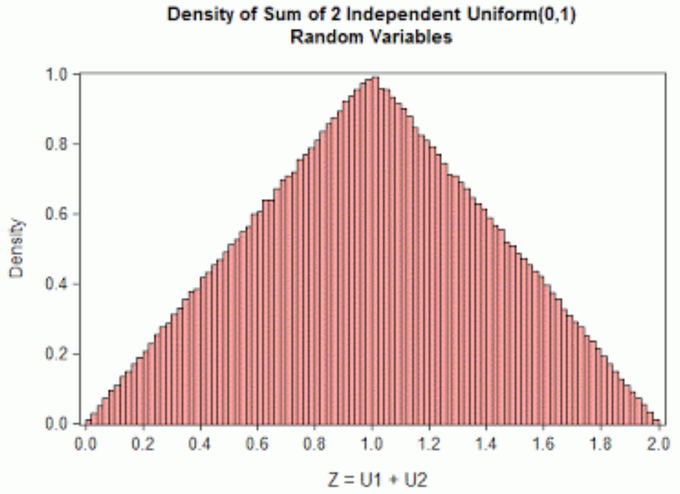
- *The*
  *enc*

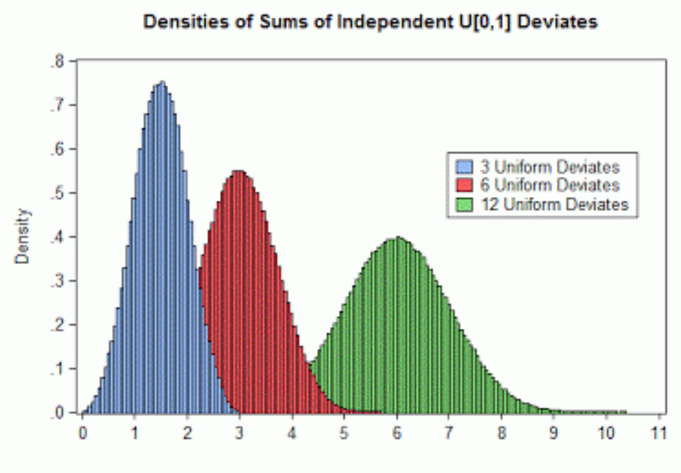


# Gaussian as a large-trial limit

- *The central limit theorem establishes the role of the Gaussian distribution as the asymptotic limit of a much broader class of problems*

**In probability theory, the central limit theorem establishes that, in many situations, when independent random variables are summe up, their properly normalized sum tends toward a normal distribution even if the original variables themselves are not normally distributed. (from Wikipedia)**



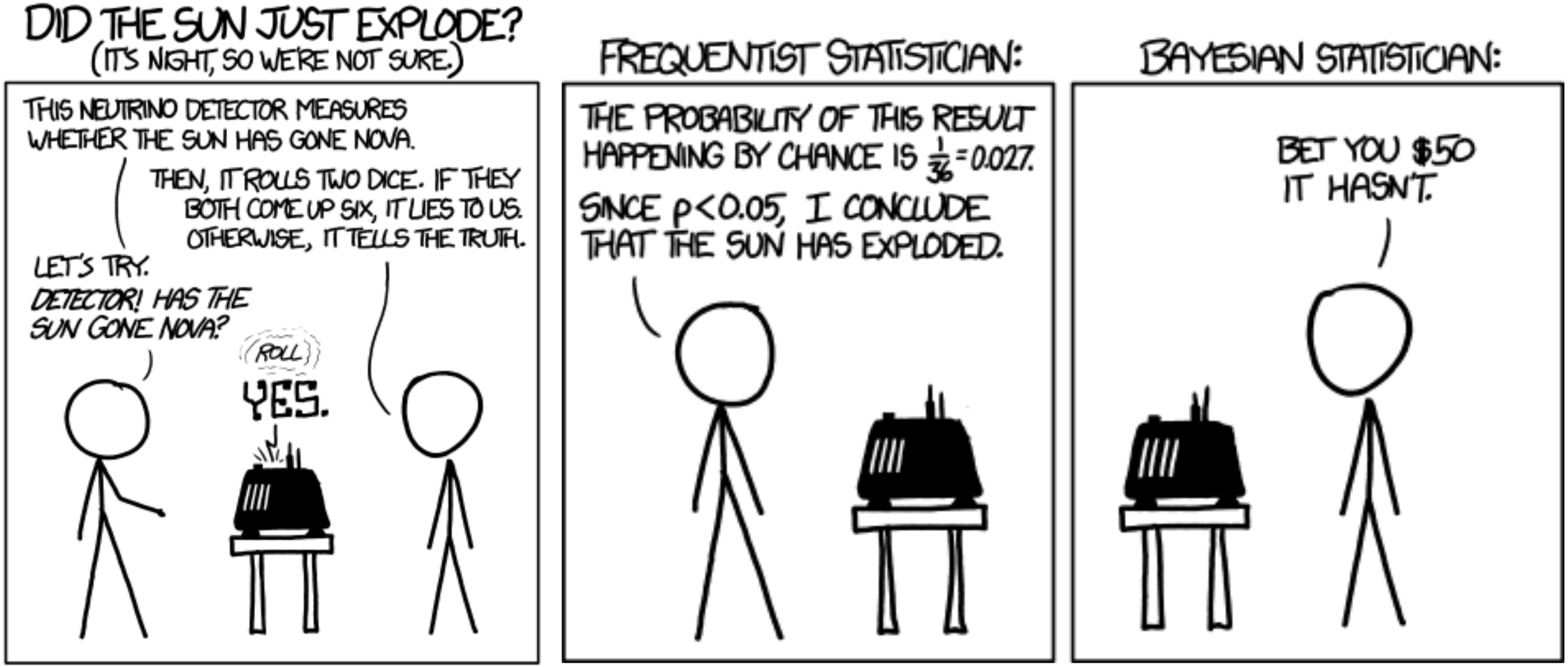




- *In practice, in a counting experiment one has to deal with*
  - *The intrinsic variation (statistical uncertainty) associated with the spread of the distribution (Poisson, Binomial, etc.)*
  - *The systematic uncertainty, associated to the uncertainty on the knowledge of the expectation. This is typically the result of many contributions -> it tends to have a Gaussian behavior*

# Recap

| Function | Distribution | E[x] | Var[x] |
|---|---|---|---|
| Poisson | $P(k \mid \lambda) = \frac{e^{-\lambda}\lambda^k}{k!}$ | λ | λ |
| Binomial | $P(k \mid p, N) = \frac{N!}{k!(N-k)!} p^k (1-p)^{N-k}$ | pN | pN(1-p) |
| Gaussian | $G(x \mid \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$ | μ | $\sigma^2$ |

# Statistics in a nutshell



# From Probability Model to Likelihood

- ***Probability***: *When we introduced distributions, we started from known distributions (e.g., a Poisson on known λ) and we tried to characterize a typical experiment outcome*

- ***Hypothesis Testing***: *Now we inverted the problem: we know the experiment outcome (e.g., we counted events above threshold during a one-year run) and we ask ourselves which of two λ values (bkg-only or sig+bkg) they come from*

- ***Inference***: *we could also just ask what is the value of λ more compatible with the observation (trivial question in this case - right? - but not in general). This is a typical application of maximum likelihood fits and a regression problem in Machine Learning*

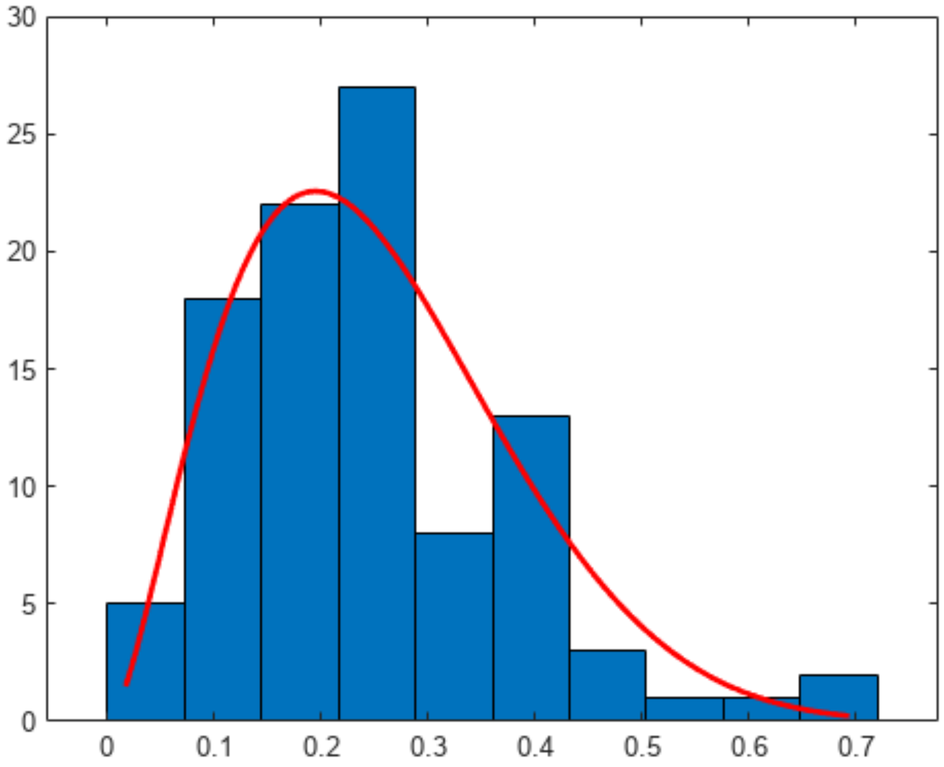
# Likelihood

$$\Pr(X{=}k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

- *Given a statistical model (e.g., our Poisson of known λ and unknown k), we can assess probabilities. Pr is a function of k*
- *Given a class of statistical models for k, function of unknown λ, we have a likelihood model*
  - *Formally the same function but a much different object*
  - *The counting is given (observed) and the mean is unknown → A likelihood is a function of λ, given the observed k*

# Likelihood

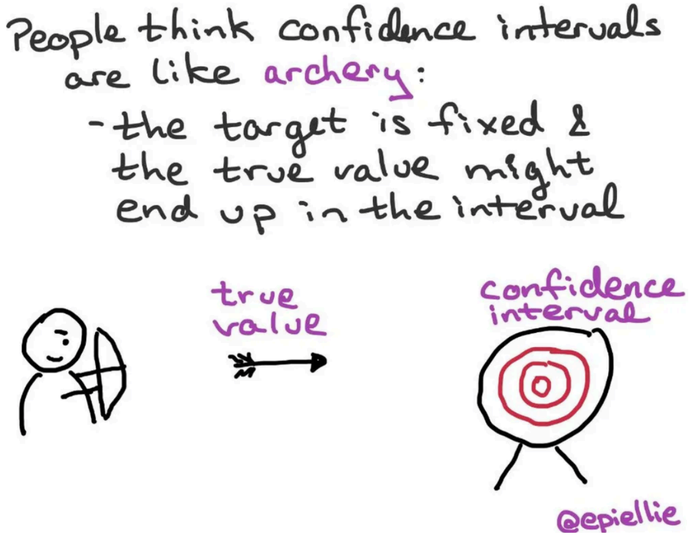
- Let's imagine a histogram of a quantity x and a curve b(x) predicting the amount of expected background
  - for each bin centre $x_i$ we can compute $b_i=b(x_i)$
  - the $b_i$ values will depend on a set of parameters that describe the curve y = b(x)
- In each bin, we observe some counting $n_i$
- The likelihood of the model is given by




$$\mathscr{L}(\vec{n}\,|\,\vec{\alpha}) = \prod_i P(n_i\,|\,b_i(\vec{\alpha})) = \prod_i P(n_i\,|\,b(x_i\,|\,\vec{\alpha})) = \prod_i \frac{e^{-b(x_i|\vec{\alpha})} b(x_i\,|\,\vec{\alpha})^{n_i}}{n_i!}$$

# Two Approaches

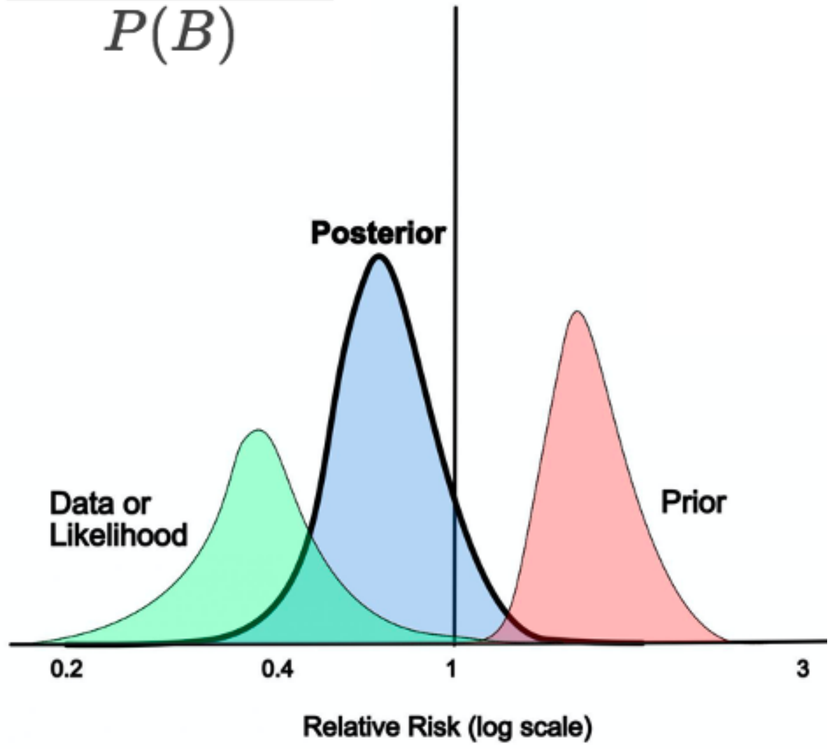
- ***Frequentist:***
  - *Frequentist statistics is a type of statistical inference that draws conclusions from sample data by emphasising the frequency or proportion of the data*
  - *Given an unaccessible true value the outcome of a measurement, frequentist statistics assess how typical the outcome is*
    - *The result is a confidence interval, defined based on a given probability (confidence level) that the true value is contained in an interval built as specified*



# Two Approaches

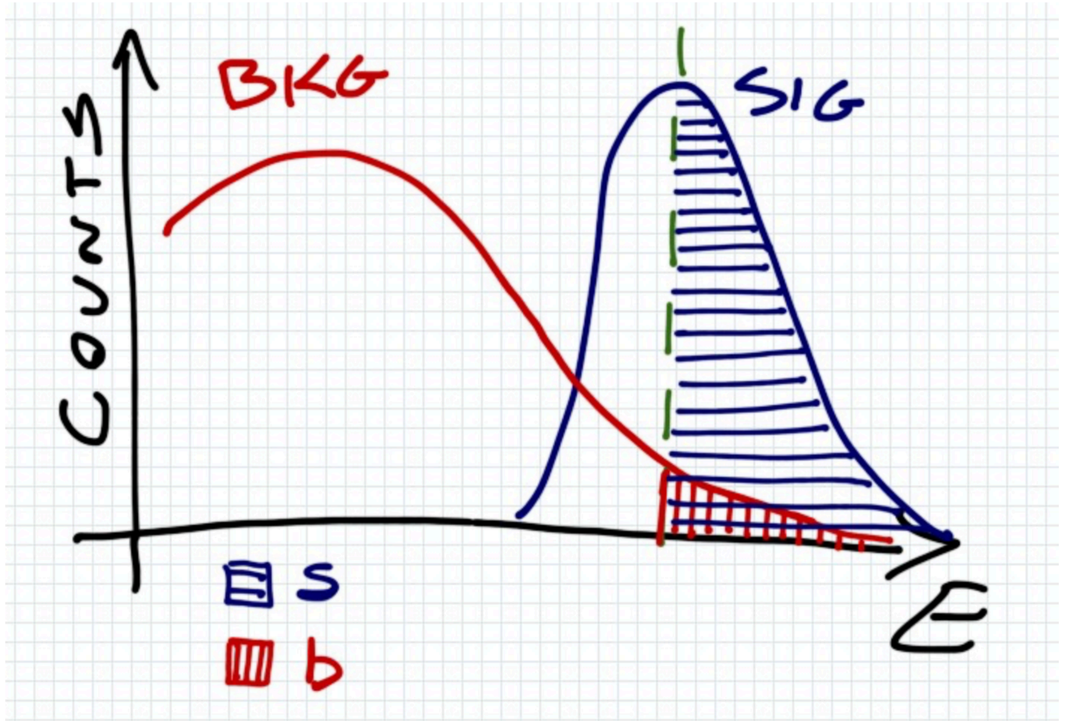
- ***Bayesian:***
  - *Bayesian statistics is an approach to data analysis and parameter estimation based on Bayes' theorem. Unique for Bayesian statistics is that all observed and unobserved parameters in a statistical model are given a joint probability distribution, termed the prior and data distributions*
  - *Given an accessible true value and the outcome of a measurement, Bayesian statistics assesses a probability range (credibility interval) for the true value, based on the measurement outcome and prior knowledge of the true value*

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



# Building the likelihood

- *Let's consider the case of a Poisson process. The likelihood is a function $P(n|\lambda)$. Which $\lambda$? The signal $\lambda_S$ or the background $\lambda_B$?*
- *The experiment is repeated multiple times and a set of n are observed. The likelihood is built as* $\prod_i P(n_i|\lambda)$
- *Once the likelihood is built, it is a function of $\lambda$. We want to estimate $\lambda$ from what we observed*



# Maximum Likelihood estimation

- We are given a likelihood model $\mathscr{L}(D|w)$ and some data D
  - D is known, $w$ are unknown
- We want to find the $\hat{w}$ values that would make our data D the most probable outcome of the experiment
  - If we knew these $\hat{w}$ values, the probability of observing D would be maximal (here D would be the unknown and $\hat{w}$ the known quantities)
  - You can convince yourselves that

$$\hat{w} = \arg\max_{w} \mathscr{L}(D \,|\, w)$$

# Example: Cross Entropy

- *Bernoulli's problem: probability of a process that can give 1 or 0*

$$\mathscr{L} = \prod_i p_i^{x_i}(1-p_i)^{1-x_i}$$

- *The corresponding likelihood is (as usual) the product of the probabilities across the events*

$$-\log \mathscr{L} = -\log[\prod_i p_i^{x_i}(1-p_i)^{1-x_i}]$$

- *Maximizing the likelihood corresponds to minimizing the -logL*

- *Minimizing the -logL corresponds to minimizing the binary cross entropy*

$$= -\sum_i \left[x_i \log p_i + (1-x_i)\log(1-p_i)\right]$$

- *We will use this expression as loss function in classification problems*

# Example: regression & MSE

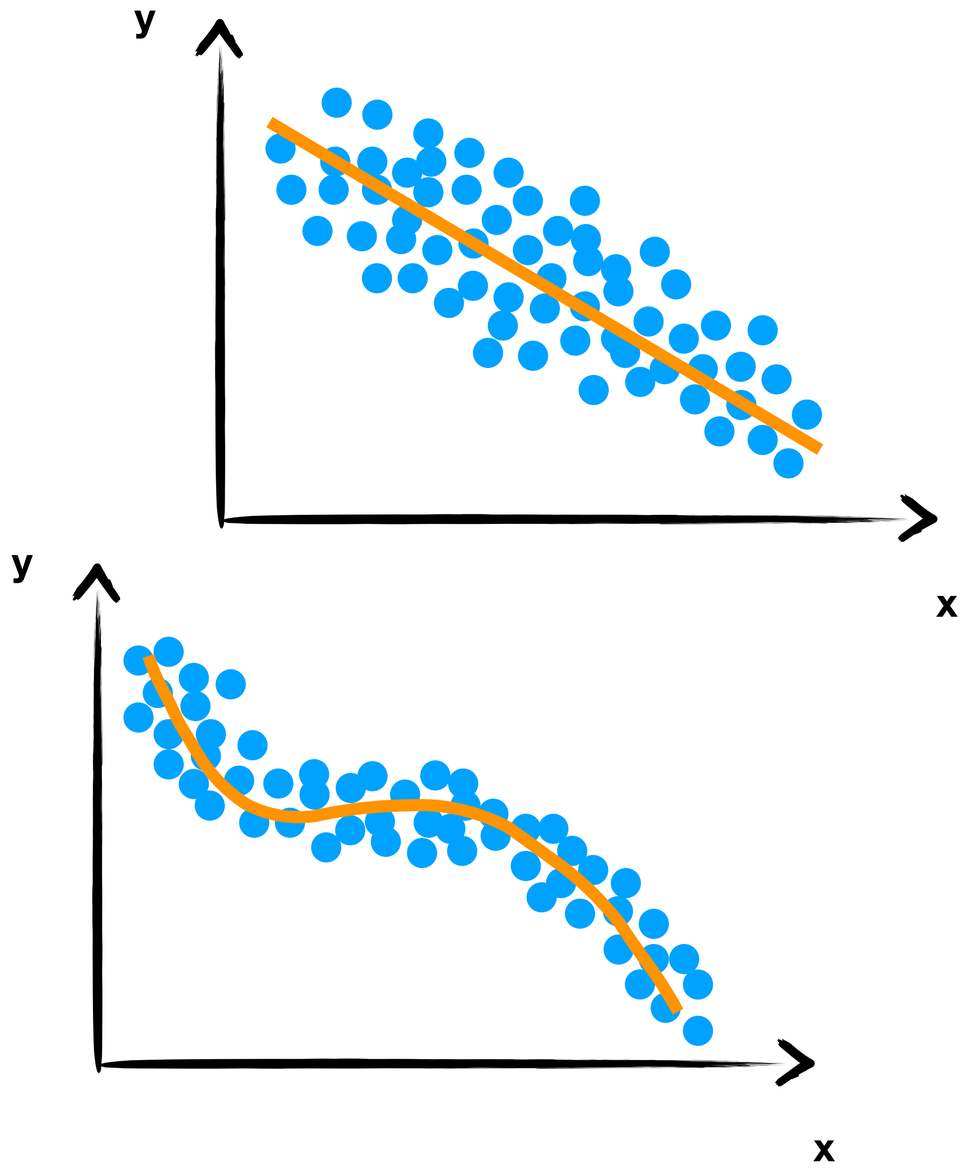
- Given a set of points, find the curve that goes through them
  - Can be a linear model

$$y_i = ax_i + b$$

  - Can be a linear function of non-linear kernel of the x. For instance, a polynomial basis

$$y_i = a\,\phi(x_i) + b$$

New feature, “engineered” from the input features



# Example: regression & MSE

- Take some model (e.g., linear)

$$h(x_i \mid a, b) = ax_i + b$$

- Consider the case of a Gaussian dispersion of y around the expected value

$$y_i = h(x_i) + e_i \qquad p(e_i) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{e_i^2}{2\sigma^2}}$$

- Assume that the resolution σ is fixed

- Write down the likelihood

$$\mathscr{L} = \prod_i \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{e_i^2}{2\sigma^2}} = \prod_i \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y_i - h(x_i))^2}{2\sigma^2}}$$

# Example: regression & MSE

- *The maximisation of this likelihood corresponds to the minimisation of the mean square error (MSE)*

$$argmin[-2\log\mathscr{L}] = argmin\Big[-2\log[\prod_i \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y_i - h(x_i))^2}{2\sigma^2}}]\Big]$$

$$= argmin\Big[\sum_i \frac{(y_i - h(x_i))^2}{\sigma^2}\Big] = argmin\Big[\sum_i (y_i - h(x_i))^2\Big] = MSE$$

- *MSE is one of the most popular loss functions when dealing with continuous outputs. We will use it a few times in the next days*

- ***BE AWARE OF THE UNDERLYING ASSUMPTION:*** *if you are using MSE, you are implicitly assuming that your y are Gaussian distributed, with fixed RMS*

- *What if the RMS is not a constant?*

# Summary

- *Introduced basic concepts of linear algebra, that will be useful for some of the math that we will find (and that you will find in literature)*
- *Introduced basic concepts of probability and statistics, since we will be dealing with non-deterministic situations when learning from data*
- *Showed how to derive popular loss functions from likelihood functions of specific problems: finding the best network corresponds to deriving a best parameter estimate from a likelihood*