Statistical Methods
(summer term 2024)

Combinatorics, error propagation, special PDFs

(based on original lectures by Prof. Dr. N. Christlieb and Dr. Hans-G. Ludwig)

Dr Yiannis Tsapras

ZAH – Heidelberg

Overview

- Combinatorics
  - studies permutations and combinations of objects chosen from a sample space
- Error propagation
  - how to determine the uncertainty in a result from the uncertainties in the individual measurements
- Calculus of expectations, variances, covariances
  - how to calculate expected values and describe relationships between parameters
- Conditional and marginal distributions
- Probability density functions of particular importance
  - normal distribution
  - binomial distribution
  - Poisson distribution

Quick introduction to Combinatorics

Multiplication principle: if one experiment has m outcomes and another independent experiment n outcomes then there are $m \times n$ outcomes for the two experiments

Ordering (permutations): the number of possible ways to order r objects (e.g. the three letters abc) is $r!$. The "!" sign indicates the factorial function. Note, that by definition $0! = 1$

Selection with replacement, order relevant: the number of ways to draw r objects from a set of n elements with replacement is n^r

Selection without replacement, order relevant: the number of ways to draw r objects from a set of n elements without replacement is $\frac{n!}{(n-r)!}$

Selection without replacement, order irrelevant: $\binom{n}{r} \equiv \frac{n!}{r!(n-r)!}$ This is the binomial coefficient

In R, `factorial()` provides the – well – factorial, `choose()` provides the binomial coefficient

What are your chances of winning the Lotto Jackpot?

Stirling's approximation and Γ-function

- For large n, calculating exact factorials can be computationally expensive and impractical. Here are some alternatives:

- For $n \geq 1$, *Stirling's approximation* is a good approximation for factorials:

$$n! \approx \sqrt{2\pi n}\left(\frac{n}{e}\right)^n \qquad \text{alternatively} \qquad \ln(n!) \approx n\ln(n) - n$$

  Note: In R factorials are provided by the function `factorial()`. It even works with non-integer arguments!

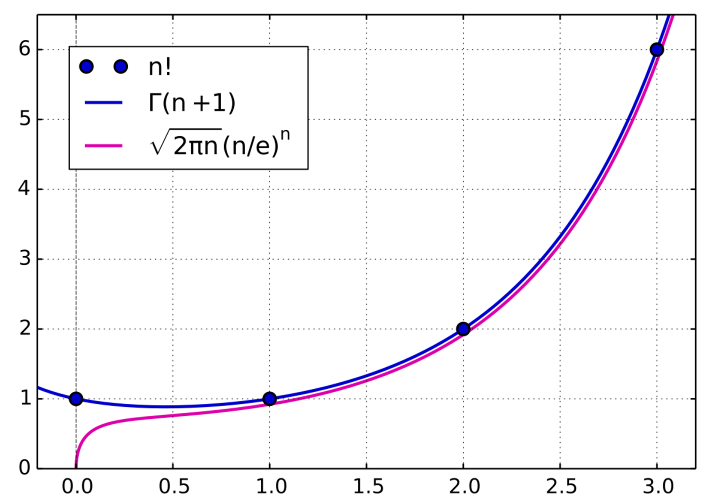
- Γ-function: continuous function closely related to factorials

$$\Gamma(x+1) \equiv \int_0^\infty t^x e^{-t} \mathrm{d}t$$

  so that

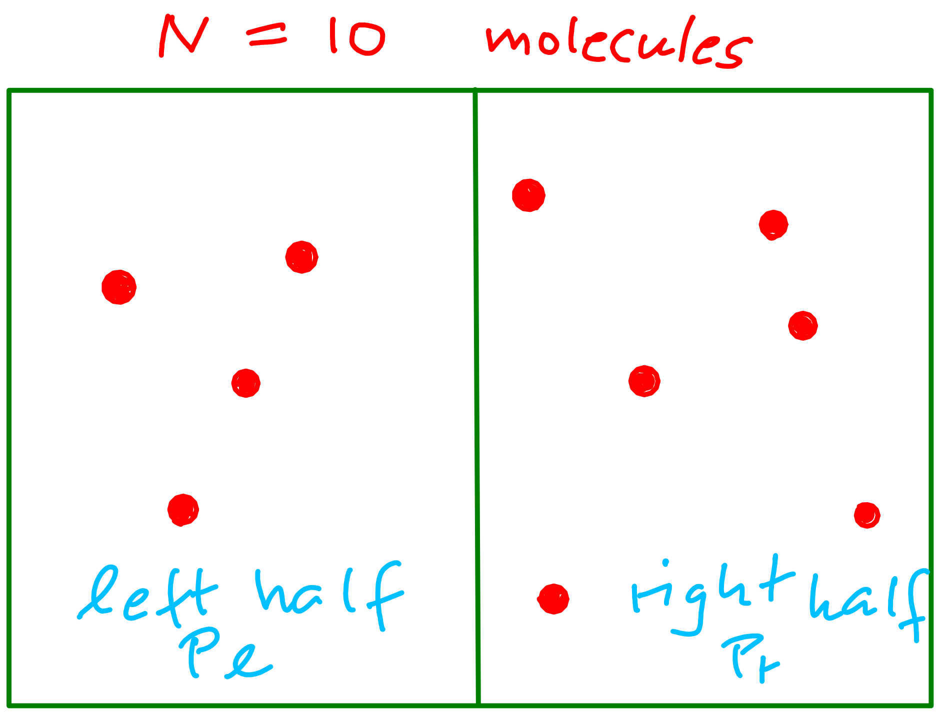
$$\Gamma(x+1) = x\,\Gamma(x) \quad \text{and} \quad \Gamma(n+1) = n!$$

  The Γ-function is provided by `gamma()` in R.



Example: distributing molecules in a box

Setup: A box contains $N = 10$ molecules which change position and velocity erratically by collisions with walls and neighbouring molecules



- How many combinations are there with n molecules in the left half of the box?
- What is the probability having $n = 5$ molecules in the left half of the box?
  - (You can assume each molecule has equal probability of being in either half)

Propagation of uncertainties, "error propagation"

- Here we are concerned with how to determine the uncertainty in a calculated result from the uncertainties in individual measurements

- Consider a variable y which is a function of several random variables x_i, i.e. $y = f(x_1, \ldots, x_n)$.

- If x_i are mutually *independent* random variables with small individual variances $\sigma_{x_i}^2$ then the Taylor expansion of f gives the variance of y

$$\sigma_y^2 \approx \sum_{i=1}^{n} \left(\frac{\partial f}{\partial x_i} \right)^2 \sigma_{x_i}^2$$

- In the case of *dependent* x_i one obtains more generally

$$\sigma_y^2 \approx \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{\partial f}{\partial x_i} \frac{\partial f}{\partial x_j} \mathrm{Cov}[x_i, x_j]$$

where $\mathrm{Cov}[x_i, x_j]$ is the *covariance* of x_i and x_j

Covariance and correlation

- *covariance* measures the joint variability of two random variables x and y
  It is defined as

$$\mathrm{Cov}[x, y] \equiv \langle (x - \langle x \rangle)(y - \langle y \rangle) \rangle = \langle xy \rangle - \langle x \rangle \langle y \rangle$$

  - variance is a special case of covariance: $\sigma_x^2 = \mathrm{Cov}[x, x]$

- *correlation* measures the *linear dependence* of variables x and y
  The *correlation* coefficient is defined as

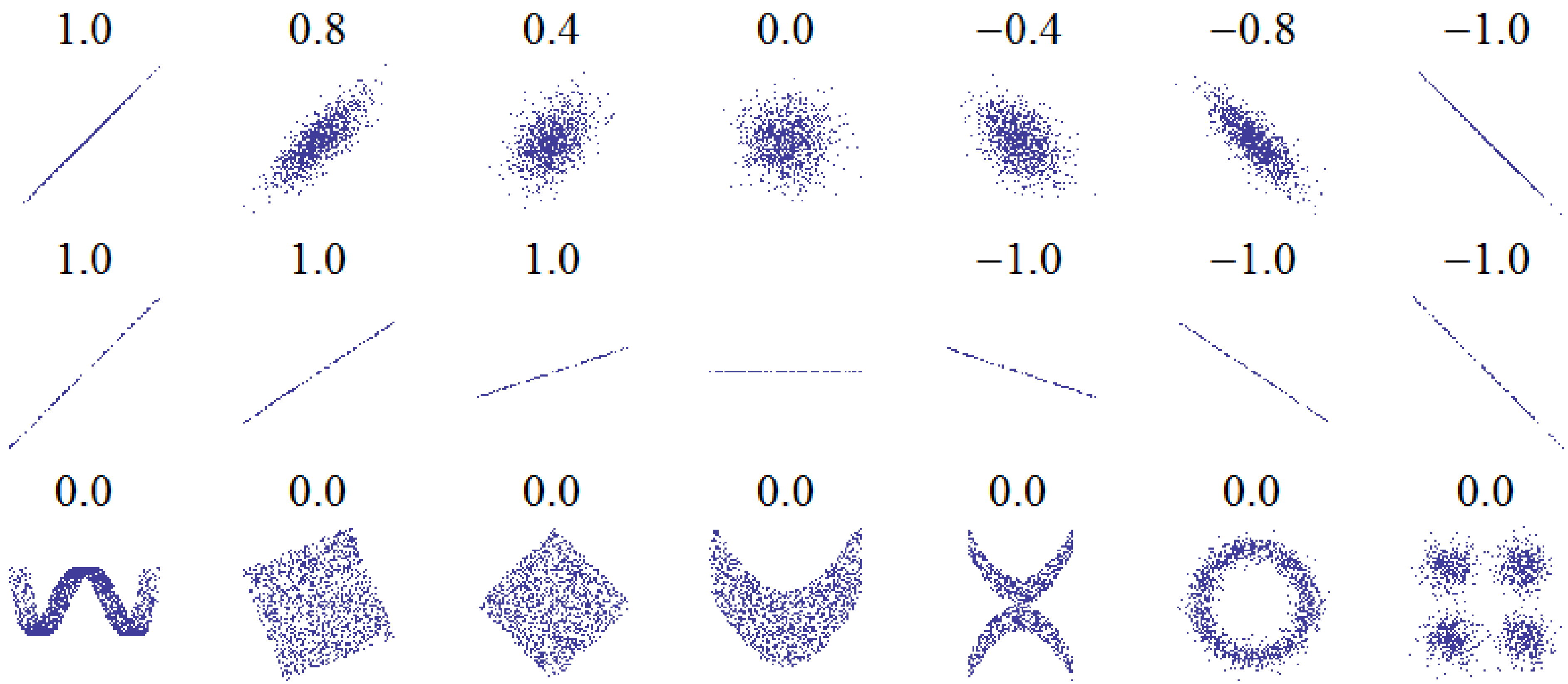
$$\mathrm{Cor}[x, y] \equiv \frac{\mathrm{Cov}[x, y]}{\sigma_x \, \sigma_y}$$

  - The correlation coefficient ranges from -1 to $+1$, indicating the strength and direction of the linear relationship, with 0 meaning no linear correlation
  - $\mathrm{Cor}[x, y] = -1$ perfectly anti- or negatively-correlated
  - $\mathrm{Cor}[x, y] = +1$ perfectly (positively-)correlated

- R functions: covariance `cov()`, correlation `cor()`

What is $\mathrm{Cor}[x, x]$? What is $\mathrm{Cor}[x, -x]$? What is $\mathrm{Cor}\left[x, x^2\right]$?

Correlation coefficient: it is good to take a look . . .



(Source: Wikipedia)

Propagation of uncertainties – the Jacobian

- When dealing with functions of **multiple variables**, it's crucial to understand how uncertainties in the input variables propagate to the output variables. This is where the Jacobian matrix comes in

- Consider a vector-valued function $\mathbf{y}$ so that

$$y_i(x_1, \ldots, x_n) \quad \text{for} \quad i = 1, \ldots, n$$

- The Jacobian matrix is a matrix of all first-order partial derivatives of the vector-valued function $\mathbf{y}$
  - When transforming random variables, the Jacobian matrix quantifies how small changes in the input variables $(x_1, x_2, \ldots, x_n)$ affect the output variables $(y_1, y_2, \ldots, y_n)$
  - Thus, it allows us to propagate uncertainties from the input variables to the output variables

Propagation of uncertainties – the covariance matrix

- $\mathbf{J}$ is the Jacobian matrix of the transformation $y_i(x_1, \ldots, x_n)$

$$\mathrm{J} = \begin{pmatrix} \partial y_1/\partial x_1 & \partial y_1/\partial x_2 & \cdots & \partial y_1/\partial x_n \\ \partial y_2/\partial x_1 & \partial y_2/\partial x_2 & \cdots & \partial y_2/\partial x_n \\ \vdots & \vdots & \ddots & \vdots \\ \partial y_n/\partial x_1 & \partial y_m/\partial x_2 & \cdots & \partial y_n/\partial x_n \end{pmatrix}$$

- Propagation of uncertainties results in

$$\Sigma[\mathbf{y}] = \mathbf{J}\,\Sigma[\mathbf{x}]\,\mathbf{J}^{\mathrm{T}}$$

  where $\Sigma[\mathbf{x}]$ and $\Sigma[\mathbf{y}]$ are the variance-covariance matrices of the random vectors $\mathbf{x}$ and $\mathbf{y}$, respectively

- The covariance matrix is symmetric and contains all combinations $\mathrm{Cov}[x_i, x_j]$
  - the diagonal elements of this matrix are the variances and the off-diagonal elements are the covariances

Properties of $\mathrm{E}[]$, $\mathrm{Var}[]$, $\mathrm{Cov}[,]$, rules of calculus

Consider (univariate) random variables X, Y, V, W and real constants a, b, c

- Expectation (sample mean) – is a linear operator

$$\mathrm{E}[a\,X + b\,Y + c] = a\,\mathrm{E}[X] + b\,\mathrm{E}[Y] + c$$

- The variance is the "mean square minus square mean"

$$\mathrm{Var}[X] \geq 0, \qquad \text{and} \qquad \mathrm{Var}[X] = \mathrm{E}\big[X^2\big] - \mathrm{E}[X]^2$$

$$\mathrm{Var}[X + a] = \mathrm{Var}[X]$$
$$\mathrm{Var}[a\,X + b\,Y] = a^2\,\mathrm{Var}[X] + b^2\,\mathrm{Var}[Y] + 2ab\,\mathrm{Cov}[X, Y]$$

- Covariance → quick check: `sumofvars.R`

$$\mathrm{Cov}[X, Y] = \mathrm{Cov}[Y, X], \qquad \text{and} \qquad \mathrm{Cov}[X, a] = 0$$

$$\mathrm{Cov}[a\,X, b\,Y] = ab\,\mathrm{Cov}[X, Y]$$
$$\mathrm{Cov}[X + a, Y + b] = \mathrm{Cov}[X, Y]$$
$$\mathrm{Cov}[X + Y, V + W] = \mathrm{Cov}[X, V] + \mathrm{Cov}[X, W] + \mathrm{Cov}[Y, V] + \mathrm{Cov}[Y, W]$$

Example problem: determination of location via GPS

Consider the simplified (1D) GPS problem where two transmitters are located at x_1 and x_2. They emit synchronously a radio pulse. The observer is located at X with $x_1 \leq X \leq x_2$, and measures the arrival times of the two signals at t_1 and t_2 of her/his time which is not synchronized with the transmitter clocks. The uncertainties of the time measurements follow a Gaussian PDF, are not correlated, and of the same value so that $\sigma_{t_1} = \sigma_{t_2} \equiv \sigma_t$.

- Use the error propagation to derive an estimate of the uncertainty of the measured location X, σ_X, and clock offset T, σ_T!

- Are the derived X and T correlated? What is their correlation coefficient?

→Blackboard & Notebook

The normal (Gaussian) distribution (revisited)

- Nomenclature: symbol "$\sim$" means "distributed as", e.g. $x \sim N(\mu = 0, \sigma^2 = 1)$ (N here signifies the normal distribution)

- Normal distribution is ubiquitous in statistics. We will see later that:
  - the sum of independent random variables, drawn from *any* distribution with finite mean and finite variance, is normally distributed (Central Limit Theorem)
  - among all distributions with a given mean and variance, the normal distribution is the one that maximizes entropy, meaning it makes the fewest assumptions about the underlying data (very useful!)

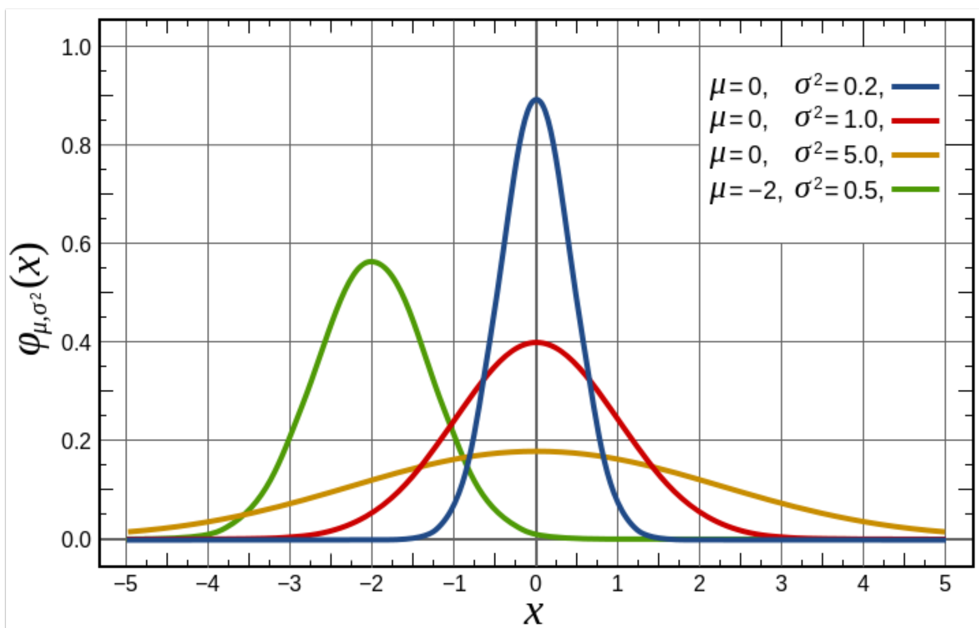
- So important that it made its way onto money bills . . .

The normal (Gaussian) probability density function

- The normal distribution is a continuous probability distribution
- It is fully characterised by its mean μ and variance σ^2

$$\varphi(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] \quad \text{with} \quad \int_{-\infty}^{+\infty} \mathrm{d}x\, \varphi(x) = 1$$

where μ represents the mean (or expectation) value and σ^2 represents the spread of the distribution. The normal distribution is symmetric around its mean



Moments of the normal PDF

- As was saw before, for the expectation value and variance we have

$$\mu = \mathrm{E}[x] = \int_{-\infty}^{+\infty} \mathrm{d}x\, x\varphi(x)$$

$$\sigma^2 = \mathrm{Var}[x] = \mathrm{E}\big[(x-\mu)^2\big]$$

- Sometimes one needs higher moments:

$$\begin{aligned} \mathrm{E}\big[x^2\big] &= \mu^2 + \sigma^2 \\ \mathrm{E}\big[x^3\big] &= \mu^3 + 3\mu\sigma^2 \\ \mathrm{E}\big[x^4\big] &= \mu^4 + 6\mu^2\sigma^2 + 3\sigma^4 \end{aligned}$$

- The normal distribution has skewness (Schiefe) 0 and kurtosis (Wölbung) 3

The cumulative distribution function of the normal distribution

- The Normal PDF is a very "compact" distribution, meaning that the probability density decreases fairly rapidly as you move away from the mean
  - important quantiles (the famous 68.3 %, 95.4 %, 99.7 %, ...)

$$\begin{aligned}\Phi(\mu+\sigma) &- \Phi(\mu-\sigma) &= 0.683\\ \Phi(\mu+2\sigma) &- \Phi(\mu-2\sigma) &= 0.954\\ \Phi(\mu+3\sigma) &- \Phi(\mu-3\sigma) &= 0.997\end{aligned}$$

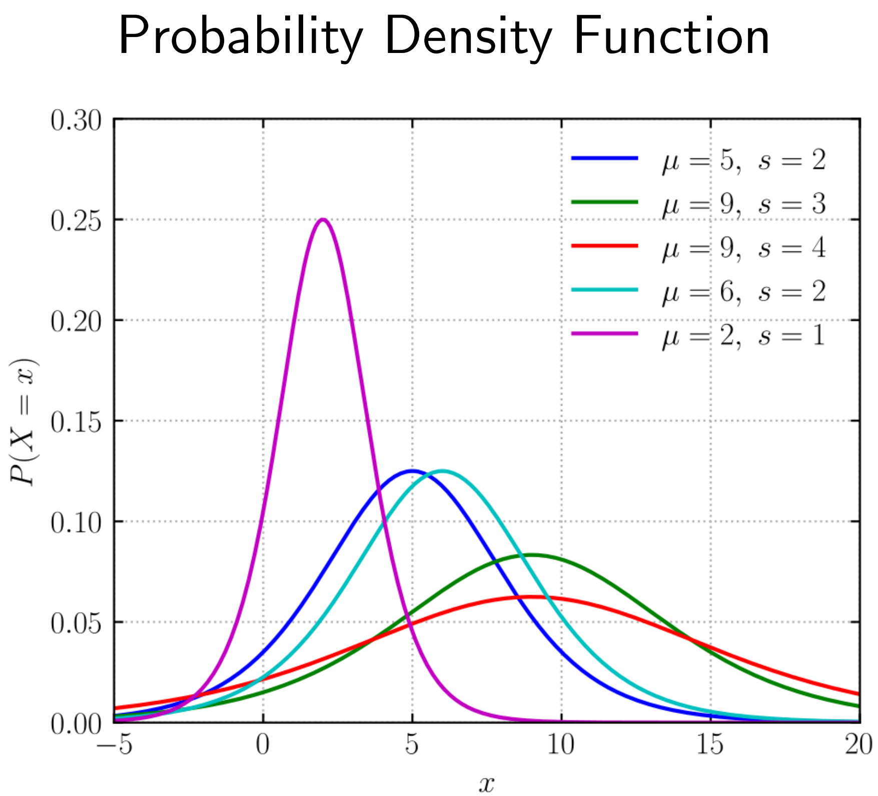
  These quantiles are important in many statistical applications, such as hypothesis testing and confidence intervals

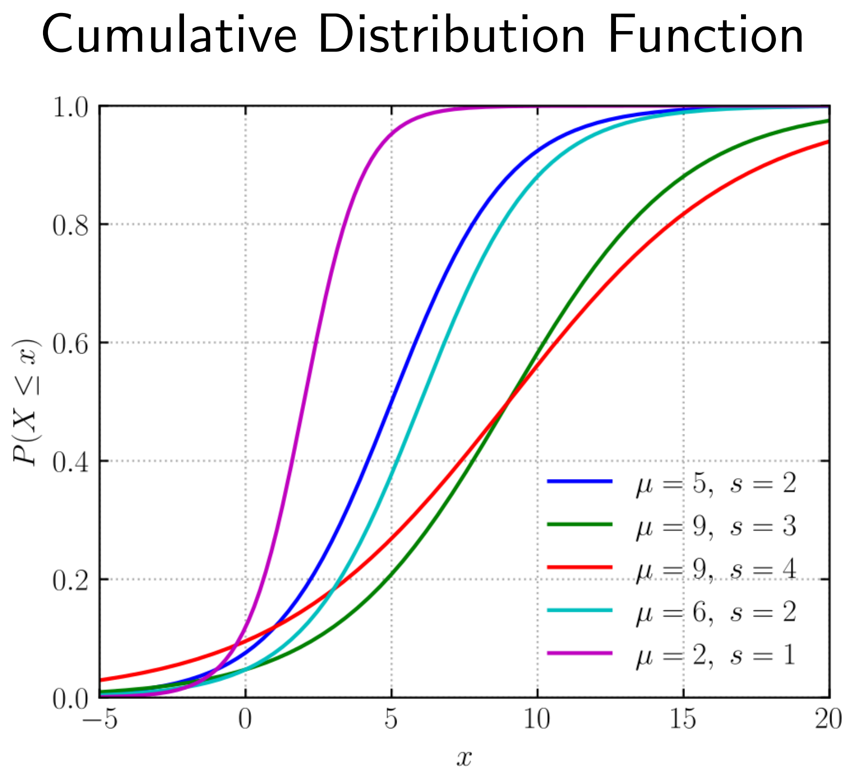
  Use R to calculate the Normal PDF for the range $\mu - 1.5\sigma \ldots \mu + 1.5\sigma$.

- The cumulative distribution function Φ is closely related to the so-called error function erf (often available in computer languages)

$$\Phi(x) \equiv \int_{-\infty}^{x} \mathrm{d}t \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(t-\mu)^2}{2\sigma^2}\right] = \frac{1}{2}\left(1 + \mathrm{erf}\left[\frac{x-\mu}{\sqrt{2}\sigma}\right]\right)$$

PDF and CDF for the standard normal distribution





Confidence intervals and confidence limits (or bounds)

- Uncertainties of measurements (or estimated parameters) can be characterized by *confidence intervals* (CI) or one-sided *confidence limits* (CL)

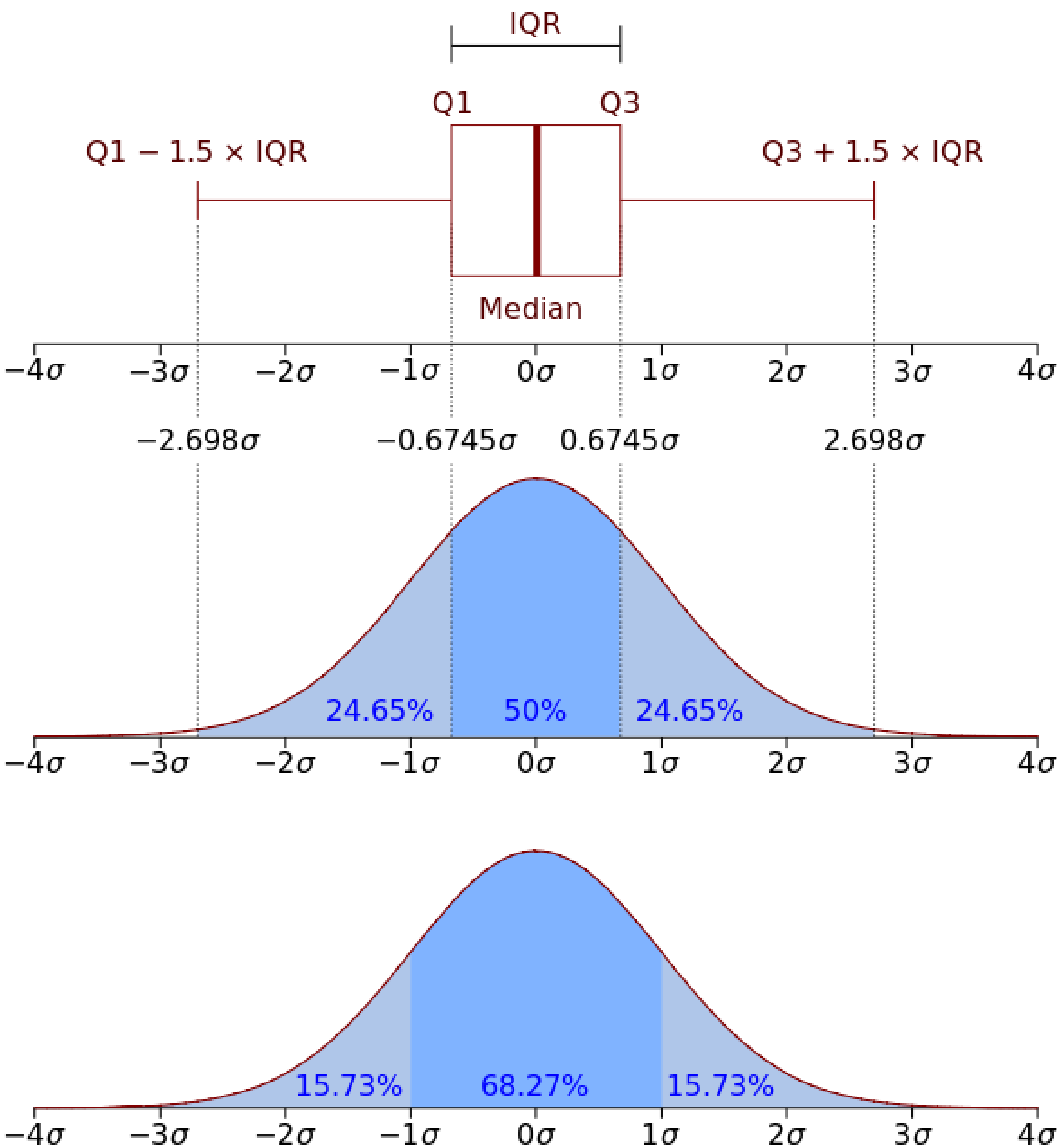
- *Confidence intervals* express the probability that a parameter lies within a certain range, while *confidence limits* express the probability that a parameter lies above or below a certain limit.

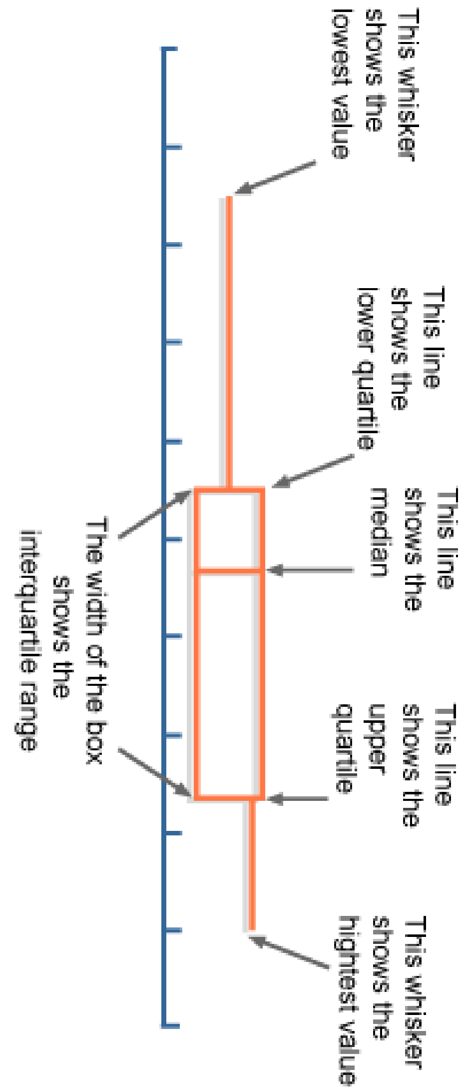
- Example: 'error bars' in plots: assuming a Gaussian error distribution, the bars stretch over the interval $[\mu - \sigma, \mu + \sigma]$
  - in this case the probability of measurement falling into this range is 0.683 (68.3%)
  - sometimes, wider *confidence intervals* are chosen, such as 2σ or 3σ
  - the discovery of the Higgs boson was claimed with a 5σ *confidence level*

- *Confidence intervals* and *confidence limits* depend on the underlying probability distribution function (PDF) of the data

- They can be visualized by box plots (or box-whisker plots) ...
  - In R, the `boxplot()` function can be used to create box plots

Confidence intervals illustrated by box plot





→ `boxplot_example.R`

Box plot – what is shown?

- A box plot (or box-and-whisker plot) is a standardized way of displaying the distribution of data based on a five-number summary: minimum, first quartile (Q1), median, third quartile (Q3), and maximum

- The simplest input to create a box plot is a numerical vector providing a sample of values. (The boxplot() function in R is highly configurable)

- The box plot shows the following components:
  - **Median**: The median is illustrated by a line inside the box, representing the middle value of the data when it is ordered
  - **First and third quartile (Q1 and Q3)**: These are shown by the boundaries of the box. Q1 is the lower hinge, and Q3 is the upper hinge. They represent the 25th and 75th percentiles, respectively
  - **Whiskers**: The whiskers extend out to $1.5\times$ the interquartile range (IQR) from Q1 and Q3. The IQR is the distance between Q1 and Q3.
  - **Outliers**: More extreme points beyond the whiskers are plotted as individual points. These are values that fall outside of $1.5\times$ IQR from the quartiles.

The central limit theorem (CLT)

- The Central Limit Theorem (CLT) is a fundamental principle in probability theory and statistics. It states that **the sum (or average) of a large number of independent, identically distributed random variables approaches a normal distribution, regardless of the original distribution of the variables**

- If y is the sum of N independent random variables x_i, $i = 1 \ldots N$, each drawn from a distribution with mean μ_i and variance $\mathrm{Var}[x_i]$, then the PDF for y ...
  - has an expectation value of $\mathrm{E}[y] = \sum_{i=1}^{N} \mu_i$
  - has a variance $\mathrm{Var}[y] = \sum_{i=1}^{N} \mathrm{Var}[x_i]$
  - becomes Gaussian in the limit $N \to \infty$

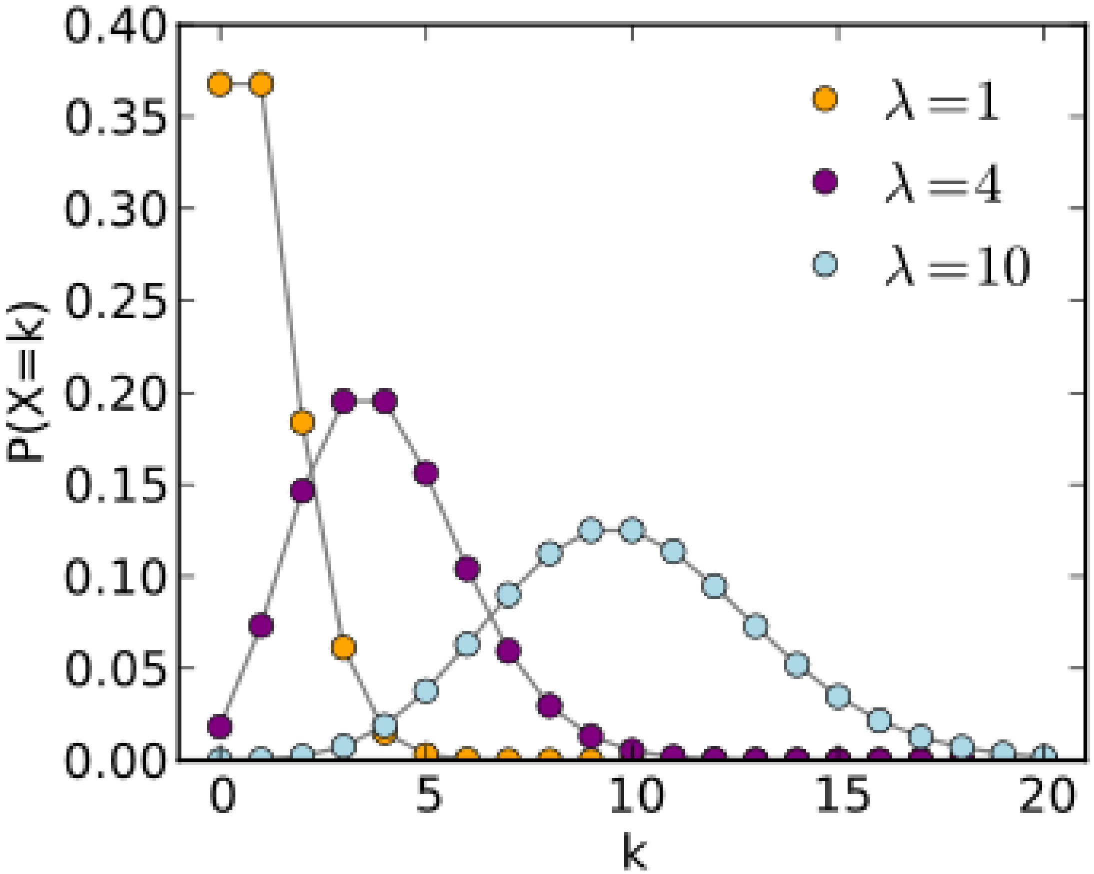
- Again, we note that none of the original distributions are required to be Gaussian
  - some technical restrictions apply: the sum giving y should not be dominated by one distribution, and means and variances must exist

- This explains the ubiquity of Gaussian distributions in natural and social phenomena

The central limit theorem (CLT)

- If z is the *average* of N independent random variables $z = \frac{1}{N}\sum_{i=1}^{N} x_i$ it follows that
  - the expectation of z is $\mathrm{E}[z] = \frac{1}{N}\sum_{i=1}^{N} \mu_i$
  - the variance of z is $\mathrm{Var}[z] = \frac{1}{N^2}\sum_{i=1}^{N} \mathrm{Var}[x_i]$
  - the standard deviation of z is $\sigma_z = \sqrt{\mathrm{Var}[z]} = \frac{1}{N}\sqrt{\sum_{i=1}^{N} \sigma_i^2}$, with $\sigma_i = \sqrt{\mathrm{Var}[x_i]}$
  - z becomes distributed according a Gaussian PDF in the limit $N \to \infty$
  - If all of the x_i come from the same distribution with mean μ and variance σ^2, then setting $\mu_i = \mu$ and $\sigma_i = \sigma$, we obtain $\mathrm{E}[z] = \frac{1}{N}N\mu = \mu$ and $\mathrm{Var}[z] = \frac{1}{N}N\sigma^2 = \frac{\sigma^2}{N}$ or $\sigma_z = \frac{\sigma}{\sqrt{N}}$
  - This means that if we take repeated measurements of a quantity, each having the same uncertainty, when we average over all measurements the uncertainty will be reduced by $1/\sqrt{N}$
  - (The demonstration of the CLT will be left to you in the exercise sheet)

The Poisson distribution (revisited)



- The Poisson distribution plays a role whenever events are counted that happen at random but with a certain mean rate λ (e.g. number of emails received per day)

- For large λ the Poisson distribution – in particular around its maximum – begins to resemble a Gaussian distribution

- Since the Poisson distribution is discrete while the normal distribution is continuous we have to be mindful what we mean by 'resembles'. In short, for large λ, the Poisson PMF can be approximated by a Normal PDF with mean λ and variance λ:

$$\frac{e^{-\lambda}\lambda^k}{k!} \approx \frac{1}{\sqrt{2\pi\lambda}} \exp\left[-\frac{(k-\lambda)^2}{2\lambda}\right]$$

Example: histograms and Poisson statistics

- Histograms are graphical representations of probability density functions (PDFs) created by counting the number of events falling into discrete bins

- Whether an event falls into a particular bin is governed by a binomial distribution
  - The expectation value of the number of counts depends on the PDF of the underlying distribution being measured

- When the number of counts in each bin is relatively large, the binomial distribution can be approximated by a Poisson distribution
  - This approximation is valid because the Poisson distribution is the limiting case of the binomial distribution when the number of trials is large
  - If N counts fall into a bin, the Poisson distribution tells us that the standard deviation of the count is $\sqrt{N}$
  - This provides an explanation and prediction of the observed/expected "noise" in histograms. It helps to judge whether a histogram is compatible with the assumption that a particular PDF underlies the data

$\rightarrow$Notebook Poisson_histogram.ipynb

Multivariate (multi-dimensional) distributions

- When we have multiple variables we are often interested in their **joint probability distribution**

- Describe the probability that a *continuous random vector* (X, Y) lies in a particular region in the domain of definition (2-D distribution):

$$\mathrm{P}((X,Y) \in A) = \iint_A f(x,y)\,dx\,dy$$

  - $\mathrm{P}((X,Y) \in A)$ denotes the probability that the random vector (X, Y) lies within a particular region A (a subset of the 2-D plane)
  - $f(x,y)$ is the **joint probability density function**, which describes the probability density for the random variables X and Y simultaneously
  - analogously, for *discrete* distributions, the **joint probability mass function** describes the probability that the random vector takes on a specific set of values
  - as usual: $f(\vec{x}) \geq 0$, and normalization $\int_D f(\vec{x})\,d\vec{x} = 1$, where D represents the entire 2-D plane for the bivariate case considered

Multivariate (multi-dimensional) distributions

- The **joint cumulative distribution function** $F(x,y) = \mathrm{P}(X \le x, Y \le y)$, for $(x,y) \in A$, is given by:

- Continuous random vector:

$$F(x,y) = \int_{-\infty}^{x} du \int_{-\infty}^{y} dv, f(u,v)$$

  with u, v being dummy integration variables

  - From this, we derive the relationship between the joint probability density function and the joint cumulative distribution function

$$f(x,y) = \frac{\partial^2}{\partial x \partial y} F(x,y)$$

- Discrete random vector:

$$F(x,y) = \sum_{x_i \le x} \sum_{y_j \le y} \mathrm{P}(X = x_i, Y = y_j)$$

Example: the bivariate Gaussian distribution

- For a continuous random vector $\vec{x} = \begin{pmatrix} x \\ y \end{pmatrix}$, the bivariate Gaussian distribution is defined by:
  - its mean vector $\vec{\mu} = \begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix}$
  - its covariance matrix $C = \begin{pmatrix} \sigma_x^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_Y & \sigma_y^2 \end{pmatrix}$

  where ρ is the correlation coefficient

- The joint probability density function (PDF) for the bivariate Gaussian distribution is given by:

$$P(x,y) = \frac{1}{2\pi\sqrt{|C|}} \exp\left(-\frac{1}{2}(\vec{x}-\vec{\mu})^{\mathrm{T}} C^{-1} (\vec{x}-\vec{\mu})\right)$$

- The exponent term $(\vec{x}-\vec{\mu})^{\mathrm{T}} C^{-1} (\vec{x}-\vec{\mu})$ represents the Mahalanobis distance between $\vec{x}$ and the mean $\vec{\mu}$ (In R, one can use the `mahalanobis()` function)

Example: the bivariate Gaussian distribution

- For simplicity, let's use a standard bivariate normal distribution with $\vec{\mu} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$ and $C = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$

- The probability density function for the standard bivariate normal distribution is then simply:

$$P(x, y) = \frac{1}{2\pi} \exp\left[-\frac{1}{2}(x^2 + y^2)\right]$$

What is the shape of contours of constant probability density?

→Notebook 2DGaussian.ipynb

Marginal distributions, independence, conditional probability

- The joint probability density for (X, Y) allows to express the probability of – say – X irrespective of any value of Y as

$$f_X(x) = \int_{-\infty}^{\infty} f(x, y)\, dy$$

and analogously for Y

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

  - $f_X(x)$ and $f_Y(x)$ are the marginal distributions associated with $f(x, y)$
  - marginal distributions are obtained by integrating out the other variables
  - in higher dimensions there are more combinations possible, i.e., combinations of what one wants to "integrate (or marginalize) out"
  - in general, the marginal distributions do *not* fully determine the joint distribution

- For cumulative distributions the formulae above also hold, and in particular

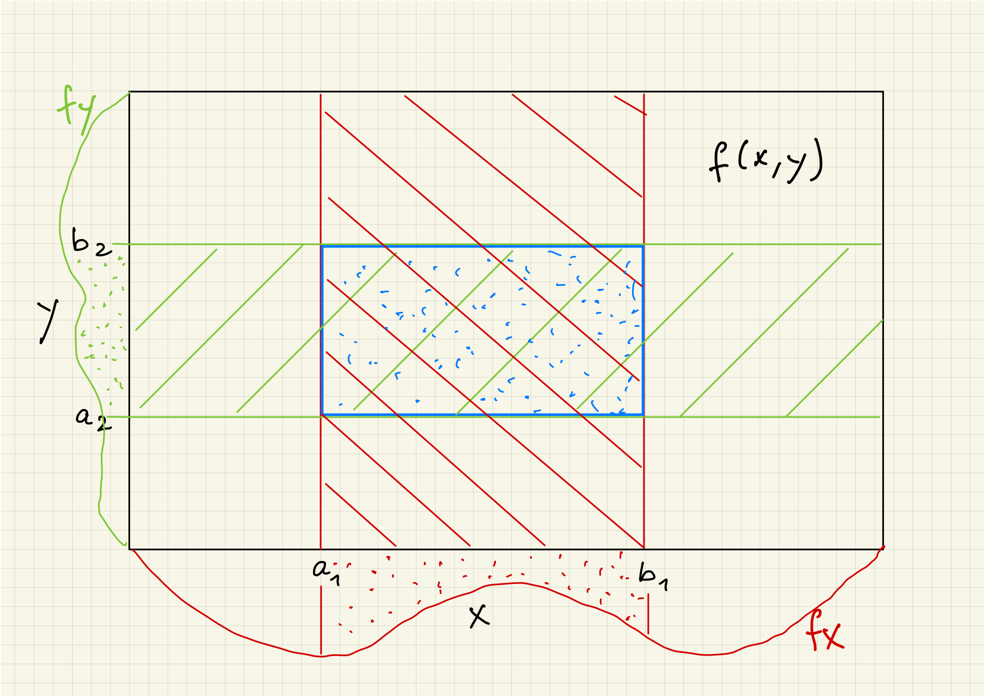
$$f_X(x) = \frac{d}{dx} F_X(x)$$

Marginal distributions, independence, conditional probability

- Random variables X and Y are independent if and only if

$$f(x,y) = f_X(x)f_Y(y) \quad \text{or} \quad F(x,y) = F_X(x)F_Y(y)$$

  - for independent random variables the joint probability distribution factorizes with the marginal distributions as factors
  - $\mathrm{P}(a_1 < X \leq b_1, a_2 < Y \leq b_2) = P_X(a_1 < X \leq b_1) \cdot P_Y(a_2 < Y \leq b_2)$



Marginal distributions, independence, conditional probability

- The joint probability density for (X, Y) allows us to express the conditional probability of – say – Y given a particular value of X as

$$f_{Y|X}(y|x) = \frac{f_{XY}(x,y)}{f_X(x)}$$

- The joint probability density can be correspondingly expressed as

$$f_{XY}(x,y) = f_{Y|X}(y|x)\, f_X(x)$$

  Integrating both sides over x gives

$$f_Y(y) = \int_{-\infty}^{\infty} f_{Y|X}(y|x)\, f_X(x)\, dx$$

  which is just an expression of the law of total probability - here for the continuous case

- As you may already expect: the relations given here for bivariate distributions have a direct correspondence with the calculus of probability we discussed previously

The multivariate (m-dimensional) normal distribution

- The only case of a multivariate PDF we will explore during the course
- Describes the joint probability distribution of m continuous random variables x_i, $i = 1 \ldots m$. For the random vector $\vec{x}$ and its expectation value, we have

$$\vec{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{pmatrix} \qquad \text{and} \qquad \vec{\mu} \equiv \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_m \end{pmatrix} = \mathrm{E}[\vec{x}]$$

- The covariances among the x_i are given by the symmetric $m \times m$ *covariance matrix* $\mathbf{C}$, with components

$$C_{ij} = \mathrm{Cov}[x_i, x_j] = \mathrm{E}[(x_i - \mu_i)(x_j - \mu_j)] = C_{ji}$$

  - due to its symmetry $\mathbf{C}$ has only $m(m+1)/2$ independent components

- From the definition of the covariance matrix we see

$$C_{ii} = \mathrm{E}\big[(x_i - \mu_i)^2\big] = \sigma_i^2$$

The multivariate (m-dimensional) normal distribution

- The correlation coefficient between x_i and x_j $(i \neq j)$ is

$$\rho_{ij} \equiv \mathrm{Cor}[x_i, x_j] = \frac{C_{ij}}{\sqrt{C_{ii} C_{jj}}}$$

- With this, the covariance matrix can be written as

$$C_{ij} = \begin{cases} \sigma_i^2 & \text{if} \quad i = j \\ \sigma_i \sigma_j \rho_{ij} & \text{if} \quad i \neq j \end{cases}$$

- Having the parameters $\mu_i, \sigma_i, \rho_{ij} (i = 1 \ldots m, j = 1 \ldots m)$ the PDF is

$$\varphi(\vec{x}) = (2\pi)^{-m/2} \det(\mathbf{C})^{-1/2} \exp\left[-\frac{1}{2}(\vec{x} - \vec{\mu})^{\mathrm{T}} \mathbf{C}^{-1} (\vec{x} - \vec{\mu})\right]$$

where $\det(\mathbf{C})$ is the determinant of $\mathbf{C}$ and $^{\mathrm{T}}$ indicates the transpose.

The multivariate (m-dimensional) normal distribution

- The covariance matrix $\mathbf{C}$ is *positive definite* meaning that it is symmetric and intertible and that $\vec{a}^{\mathrm{T}}\mathbf{C}\vec{a} > 0$ for all non-zero vectors $\vec{a}$ of length m. This implies (among other things)...
  - $\det(\mathbf{C}) > 0$ and C^{-1} exists and is also positive definite
  - $(\vec{x} - \vec{\mu})^{\mathrm{T}}\mathbf{C}^{-1}(\vec{x} - \vec{\mu}) \geq 0$ and the multivariate PDF reaches its maximum at $\vec{x} = \vec{\mu}$

- In case that all x_i are uncorrelated ($\rho_{ij} = 0$) C becomes diagonal with

$$C_{ij} = \begin{cases} \sigma_i^2 & \text{if} \quad i = j \\ 0 & \text{if} \quad i \neq j \end{cases}$$

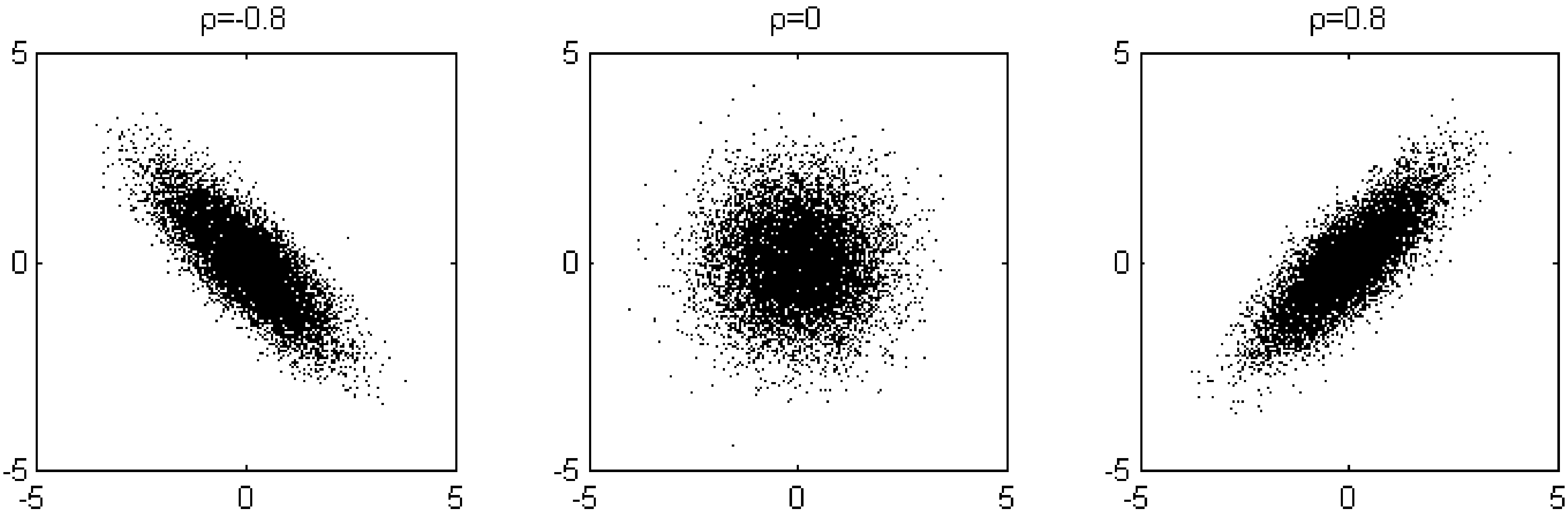
and the PDF becomes

$$\varphi(\vec{x}) = \prod_{i=1}^{m} (2\pi)^{-1/2} \sigma_i^{-1} \exp\left[-\frac{1}{2} \left(\frac{x_i - \mu_i}{\sigma_i} \right)^2 \right]$$

i.e. a product of univariate normal PDFs $N(\mu_i, \sigma_i^2)$

The multivariate (m-dimensional) normal distribution

- Why is this important?
  - as we will see: multidimensional PDFs often look similar to multivariate normal distribution around their maxima



Plot of 10,000 random samples (x, y) drawn from a bivariate (2D) normal distribution with $\sigma_x^2 = \sigma_y^2 = 1$ and different correlation coefficients ρ. The density of points is proportional to the value of the PDF.

The multivariate (m-dimensional) normal distribution

- As you may guess: all conditional and marginal distributions of a multivariate normal distribution can be expressed $analytically$. Moreover ...
  - all possible marginal distributions are again multivariate normal distributions (of lower dimension since some vector components are marginalized out)
  - all possible conditional distributions are multivariate normal distributions

- Towards the end of the course we will come back to this with explicit formulae, but for now just note that:
  - In multiple dimensions, the central limit theorem suggests that the sum of many independent random variables, regardless of their original distributions, tends to a multivariate normal distribution
  - These concepts are central to Gaussian processes, which rely heavily on the properties of multivariate normal distributions
  - Many maximum likelihood estimates in regression and machine learning assume multivariate normality of the data