Elements of Computational Biology

Saul Pierotti

November 5, 2019

Introduction and topics

- Linear algebra is one of the main topics for developing algorithms
- Biology ha a high level of noise in its data
- I should know the formula of binomial and normal distribution
- We will study tests like T-Student, ANOVA
- Slides: ww.biocomp.unibo.it/gigi/2019-2020/ECB

Linear algebra

Vectors

- Given a reference system, a vector is represented by its components on the axes
- The span of n vectors is the set of all possible vectors that you can represent by their linear combinations
- If a vector can be expressed as a linear combination of another, it is said to be linear dependent from it
- The basis of a vector space is a set of linearly independent vectors that span the full space
- $\vec{x} \in \mathbb{R}^n$ means x is a real vector in an n-dimensional space
- $\vec{x} \in \mathbb{C}^n$ means x is a complex vector in an n-dimensional space
- Sum of vectors is done by summing their components or graphically with the parallelogram rule
  - $\vec{c} = \vec{a} + \vec{b} \implies c_i = a_i + b_i \ \forall \ i = 1 \to n$
  - Difference is the same concept
  - $\forall \, \vec{v} \ \exists \ \vec{0} \ : \ \vec{v} + \vec{0} = \vec{v}$
  - You can only sum vectors in the same vector space
- The norm of a vector $||\vec{v}||$ is its length
  - Can be computed with the pitagorean theorem $||\vec{v}|| = \sqrt{\sum_{i=1}^{n} v_i^2}$
  - The norm of the sum is less or equal to the sum of the norm of the components
    - This follows from the geometry of a triangle
  - The scalar product of a norm is the norm of the scalar product
    - $\lambda||\vec{v}|| = ||\lambda\vec{v}||$
- The distance between points in space is the norm of the difference between the vectors defining the points
  - $d(a, b) = ||\vec{a} - \vec{b}||$
- Scalar multiplication
  - $\vec{c} = \lambda\vec{a} \implies c_i = a_i\lambda$
  - A scalar multiplication of a sum is the sum of the scalar multiplications of the components
- Dot product, also called scalar or inner product
  - You can use the notation $< A, B >$
  - It is used in physics to calculate work
  - $\vec{w} = ||\vec{F}|| * ||\vec{s}|| * cos\theta = \sum_{i=1}^{n} F_i * s_i$
  - It is a number, complex or real depending on the vectors (!)
  - It is commutative and distributive
  - $< x, x >= ||\vec{x}||^2$

  - It is positive when the angle is acute
  - No cancellation rule
    - $< A, B >=< A, C > \not\Longrightarrow \vec{B} = \vec{C}$
- Angle between vectors
  - Can be calculated inverting the dot product
- A line passing through the origin can be defined as the set of points orthogonal to a vector $\vec{w}$
  - $w_1x_1 + w_2x_2 = 0$
  - In higher dimensions this describes an hyperplane (an n-1 dimensional object)
- All the point on a hyperplane have the same projection on its defining vector $\vec{w}$
  - The projection p of $\vec{x}$ on $\vec{w}$ is calculated as $\vec{x} * cos\theta$
  - An hyperplane is therefore an object subjected to the constraint $\vec{x} * cos\theta = p$
  - Given that $< \vec{x}, \vec{w} >= ||\vec{x}|| * ||\vec{w}|| * cos\theta$ we have that $p = \frac{<\vec{x},\vec{w}>}{||\vec{w}||}$
  - If p>0 the hyperplane is in the direction of $\vec{w}$, if it is negative it is in the opposite direction
  - Defining $b = -\frac{p}{||\vec{w}||}$ we have the canonical equation for the hyperplane
    - $w_1x_1 + w_2x_2 + b = 0$ in 2 dimensions
    - $< \vec{w}, \vec{x} > +b = W^tX + b = 0$ in n dimensions
  - An hyperplane is useful for subdividing space
- 2 hyperplanes are parallel if their are defined by the same vector $\vec{w}$ allowing for a scaling factor λ
  - $< Y, W >= \lambda < X, W >$
- The distance between parallel hyperplanes is computed as the difference of their projections on $\vec{w}$
  - $d(X, Y) = p_y - p_x = \frac{b_x - b_y}{||W||}$
- The distance of a point A from a hyperplane is the projection of the point on the defining vector $\vec{w}$, minus the projection of the hyperplane on the same vector
  - $D(A, X) = p_a - p_x = \frac{<A,W>+b}{||W||}$
- Hyperplanes are useful for the separation of classes of data
- Every column of a matrix can be thought of as a vector
  - To make the dot product of 2 vectors using matrices you can multiply one vector for the transpose of the second
  - $< \vec{a}, \vec{b} >= A * B^t$

Matrices

- A matrix is an array of numbers arranged in a rectangular structure
- The columns of a matrix are the coordinates where the basis vectors land after the transformation
- It has m rows and n columns, it is represented as $A \in \mathbb{R}^{m*n}$
  - $A = \begin{pmatrix} a_{11} & a_{12} & ... & a_{1n} \\ a_{21} & a_{22} & ... & a_{2n} \\ ... & ... & ... & ... \\ a_{m1} & a_{,m2} & ... & a_{mn} \end{pmatrix}$
  - The single a_{ij} numbers are called elements
  - The index of an element is always mn, meaning first row and then column
- If n=1, the matrix is called column matrix, which is a vector
- If m=1, the matrix is a row matrix
- A and B are equal if they have the same dimensions and they are equal element by element
  - $A = B \iff a_{ij} = b_{ij}$
- The 0 matrix contains all 0 elements and does not change the matrix it is added to
- The sum is defined as the sum of the respective elements
  - We can sum only matrices of the same dimensions, they are said to be conformable for addition
  - $C = B + A \iff c_{ij} = b_{ij} + a_{ij}$
  - The difference operates in the same way
- Scalar multiplication is performed multiplying all the elements of the matrix for the scalar
  - $C = \lambda A$ implies $c_{ij} = \lambda a_{ij}$
- The negative of A is -A, defined as $-1 * A$
  - $A - A = 0$

- Matrix addition and scalar multiplication are commutative, associative and distributive
- Matrix product is an operation that is defined only if the number of columns of the first matrix is equal to the number of rows of the second (the matrices are conformable for the product)
  - A is of dimensions $m * p$ and B of dimensions $p * n$, if $C = A * B$
  - $c_{ij} = \sum_{k=1}^{p} a_{ik} b_{kj}$
  - $C = A * B$ can be computed as row by column product
  - It can be defined only if the number of columns in the first matrix is equal to the number of rows of the second
    - $R^{m*p} * R^{p*n} \implies R^{m*n}$
  - The result is a matrix with the same number of rows as the first, and the same number of columns as the second
  - The product between matrices is NOT commutative (!)
  - $A(B + C) = AB + AC$
  - $(A + B)C = AC + BC$
  - $A(BC) = (AB)C$
  - Be aware!
    - If $AB = 0$ we can NOT conclude that B or C are 0
    - If $AB = AC$ we can NOT conclude that $B = C$
- A square matrix has m=n
- An upper triangular matrix has all the elements below the diagonal equal to 0, and a lower triangular the ones above it
- A diagonal matrix has all the elements outside the diagonal equal to 0
- A diagonal matrix with all 1 elements is the identity matrix I
  - It does not change the square matrix it is multiplied to
  - In this case, $AI = IA = A$
- If AB=BA, A and B are said to commute
  - If A is a square matrix, it commutes with itself and with I
- If AB=-BA, A and B are said to anti-commute
- The transposition of a $n * m$ matrix is a $m * n$ matrix, called A^t, where $[A^t]_{ij} = A_{ji}$
  - A and A^t are always conformable to product, in both directions
  - $(A^t)^t = A$
- A square matrix is symmetric if $A = A^t$, antisymmetric (skew-symmetric) if $A = -A^t$
  - $A + A^t$ is always symmetric
  - $A - A^t$ is always antisymmetric
  - An antisymmetric matrix has a 0 diagonal and antisymmetrical elements otherwise
- The inverse of a matrix A, called A^{-1}, is a matrix such that $A * A^{-1} = A^{-1} * A = I$
  - It is defined only if $det(A) \neq 0$
- An orthogonal matrix has its inverse equal to the transpose, $A^{-1} = A^t$
  - An orthogonal matrix describes a spatial rotation
  - Therefore, $AA^t = A^tA = I$
    - You can check for orthogonality by checking that $A * A^t = I$
- Some properties of transpose and inverse matrices
  - $(AB)^{-1} = B^{-1} * A^{-1}$, but only if $(AB)^{-1}$ exists(!)
  - $(AB)^t = B^t * A^t$
- It is possible to associate a number called determinant to any square matrix
  - $det(A) = |A| \in \mathbb{R}$
  - For an order 2 square matrix, that is computed subtracting the product of the second diagonal to that of the first
    - $det(A) = a_{11} * a_{22} - a_{12} * a_{21}$
  - It represents the area of the unit square after the transformation
  - Its sign reflects the orientation of space
    - If it is negative, the transformation flips the axis
- The rank of a transformation is the dimensionality of its output space, called column space
  - The column space of a transformation is the span of the basis vectors defined by its columns

- Some proprieties of determinants
  - If an entire row or column is equal to 0, then the determinant of the matrix is 0
  - $det(A * B) = det(A) * det(B)$
  - The determinant of an orthogonal matrix is either 1 or -1
  - $det(A) = det(A^t)$
- How to compute the inverse of 2*2 matrices
  - Given the definition $A * A^{-1} = I$ if $det(A) \neq 0$
  - It follows $A^{-1} = \frac{1}{det(A)} \begin{pmatrix} a_{22} & -a_{12} \\ -a_{21} & a_{11} \end{pmatrix}$
- A minor M_{ij} of a matrix A is the determinant of any square submatrix of A
- The cofactor of the element a_{ij} is $C_{ij} : C_{ij} = M_{ij} * (-1)^{i+j}$
- To compute the determinant of any matrix you pick any row or column and sum the product of any element in it for its cofactor
  - In a column $det(A) = \sum_{i=1}^{n} a_{ij} * C_{ij}$
  - In a row $det(A) = \sum_{j=1}^{n} a_{ij} * C_{ij}$
  - It is convenient to choose the row or column with most 0 for the computation
  - If 2 rows are identical, det(A)=0
  - If one row is 0, then det(A)=0
  - If you exchange 2 rows, det(A')=-det(A)
  - The determinant of a triangular matrix is the product of the diagonal elements
  - If B is obtained by multiplying every element in a row of A by λ, $det(B) = \lambda det(A)$
  - For any n*n square matrix, $det(\lambda A) = \lambda^n det(A)$
  - If A and B are of the same order, $det(AB) = det(A)det(B)$
- The cofactor matrix of A, called A^c, is a matrix with each element equal to the cofactor of the same element in a
  - $A^c : a^c_{ij} = C_{ij}$
- The adjugate matrix of A is the transpose of its cofactor matrix
  - $A^a = (A^c)^t$
- The inverse matrix can be obtained by dividing the adjugate of a matrix for its determinant
  - $A^{-1} = \frac{1}{|A|} A^a$
- Matrices can represent systems of linear equations
  - The system $\begin{cases} x + y = 7 \\ 3x - y = 5 \end{cases}$ can be represented as $\begin{pmatrix} 1 & 1 \\ 3 & -1 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 7 \\ 5 \end{pmatrix}$
  - The system has a solution if the coefficient matrix is invertible

Linear transformations

- A matrix can be thought of as a linear transformation of a vector space
  - $A^{m*n} : R^n \to R^m$
  - A linear transformation is a transformation that preserves linearity and does not move the origin
    - $A(\vec{v} + \vec{u}) = A\vec{v} + A\vec{u}$ and $A(\lambda\vec{v}) = \lambda A\vec{v}$
- A rotation by an angle θ can be describe by the transformation
  - $A = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}$
- The inverse transformation takes a transformed vector and restores the original one
  - Sometimes it does not exist (!) when det(A)=0
- If det(A)=0 the transformation squishes space to a lower-dimensional vector space
- The composition of the transformations A followed by B is $C = BA \neq AB$
- The scalar product of the transformation of a vector $A\vec{v}$ and the vector $\vec{w}$ is equal to the scalar product of the first vector with the second vector transformed by the transpose of A
  - $A\vec{v} * \vec{w} = \vec{v} * A^t\vec{w}$
- The null space of a transformation is the set of vectors that get squished to $\vec{0}$ by the transformation

  - $\vec{b} \in Null(A) \iff A\vec{b} = \vec{0}$
  - A trivial null space is always $\vec{0}$ itself
  - There is a true null space only if $det(A) = 0$
  - If $det(A) \neq 0$, the only null space is $\vec{0}$ itself
- The null space of a square matrix can be computed setting up a linear system of equations
  - For a matrix $A = \begin{pmatrix} 1 & 2 \\ 2 & 4 \end{pmatrix}$, $det(A) = 0$
  - $A\vec{b} = \vec{0} \implies \begin{pmatrix} 1 & 2 \\ 2 & 4 \end{pmatrix} \begin{pmatrix} b_1 \\ b_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \implies \begin{cases} b_1 + 2b_2 = 0 \\ 2b_1 + 4b_2 = 0 \end{cases} \implies b_1 = -2b_2 \implies \vec{b} = \lambda \begin{pmatrix} -2 \\ 1 \end{pmatrix}$

Eigenstuff

- For the transformation A, if $A\vec{b} = \lambda\vec{b}$, $\vec{b}$ is an eigenvector of A and λ is its eigenvalue
  - An eigenvector of a transformation is a vector that is only rescaled by the transformation
  - An eigenvalue is the scaling factor to which the vector is subjected by the transformation
  - The $\vec{0}$ vector can never be an eigenvector even though $A\vec{0} = \lambda\vec{0}$ is always true
    - On the contrary, it is possible that $\lambda = 0$
- How to find eigenvalues for the matrix A
  - $A\vec{b} = \lambda\vec{b} \implies A\vec{b} - \lambda\vec{b} = 0 \implies (A - \lambda I)\vec{b} = 0$
  - This means that the eigenvectors $\vec{b}$ are the non-trivial null space of the transformation $(A - \lambda I)$
  - It is required that $det(A - \lambda I) = 0$, otherwise there are no eigenvectors
  - The equation $det(A - \lambda I) = 0$ is called characteristic equation of A and its root allows to recover the eigenvalues of the matrix
- How to find the eigenvectors for the matrix A
  - Once I have the eigenvalues λ, the eigenvectors can be found by solving $(A - \lambda I)\vec{b} = 0$ for $\vec{b}$, using all the λ
- The number of eigenvalues and eigenvectors (families of linearly dependent eigenvectors) is equal to the dimensions of the vector space
  - I always have n families of eigenvectors in the vector space R^n
  - The families of eigenvectors are orthogonal to each other iff the transformation is symmetric
  - They define a convenient reference frame, even if they are not orthogonal
  - Each eigenvalue scales one of the families of eigenvectors, meaning that it stretches one of the dimensions of the new reference frame
- Note that in a triangular or diagonal matrix the diagonal elements are its eigenvalues
- The product of the eigenvalues is equal to the determinant of the matrix
  - A non-invertible matrix (also called singular matrix) has always at least a 0 eigenvalue
  - The inverse matrix has reciprocal eigenvalues ($\frac{1}{\lambda}$)
  - The eigenvalue of kA is $k\lambda$
  - The eigenvalue of A^n is λ^n
  - Transposition does not change the eigenvalues
- The sum of the eigenvalues is the trace of the matrix
  - The trace of a matrix is the sum of its diagonal elements

Complex field

- Sometimes there can be no real eigenvalues, but there are always solutions in the complex field
  - This happens when the characteristic equation of the matrix has $\Delta < 0$
- A complex number z is written as $z = a + ib$ where $i = \sqrt{-1}$
  - a is called real part
  - b is called imaginary part
- The reference axis of a vector space are NOT necessarily orthogonal to each other (!)
  - But they must be linearly independent

Change of basis and diagonalization

- To go from one system to the other we need the representation of the old basis vectors in term of the new ones
  - $\hat{i} = a\hat{i'} + b\hat{j'}$ and $\hat{j} = a\hat{i'} + b\hat{j'}$
  - If $U = \begin{pmatrix} a & c \\ b & d \end{pmatrix}$ we have that $\vec{v'} = U\vec{v}$
  - It is possible to go back to the original system of reference using U^{-1}
- If I want to use a system of which I know the coordinate of the basis vectors in term of my current basis vectors I need to use the inverse of the matrix containing these coordinates
- If a $n * n$ matrix A has n eigenvectors which are linearly independent, I can write the $n * n$ matrix U containing all the eigenvectors, and use it to convert to a new system of reference
  - The eigenvectors of A will be the new basis vectors
- I can compute A in the new reference frame forming $\Lambda = U^{-1}AU$
  - Given a vector $\vec{v}$, I first convert it to the new reference frame where the eigenvectors are the basis vectors using U, then I apply A and finally I go back to the old system of reference using U^{-1}
  - This new matrix Λ will be diagonal (!)
  - Each column will be made of one eigenvector multiplied by its eigenvalue
  - It is good to choose normalized vectors for the change of basis, meaning that their norm should be 1
    - In this case $det(U) = 1$, meaning that areas are preserved by the transformation
- Why do I want to use eigenvectors as reference frames?
  - Because the components of any vector are only rescaled by the original transformation A in this reference frame
  - This makes much easier to compute transformations
- If a matrix is symmetric ($A = A^t$) its eigenvalues are real and its eigenvectors are orthogonal
  - If the eigenvectors are normalized $U^{-1} = U^t$, therefore $\Lambda = U^tAU$
- In the same way that a linear form can be represented as all the points orthogonal to a vector with a projection p onto it, a matrix can describe a quadratic form

Quadratic forms

- A quadratic form is an equation in more than 1 variable were each term has a variable squared or multiplied to another variable
  - An example is $ax^2 + bxy + cy^2 = 0$
- This is represented as $\vec{x}^tA\vec{x} + b = 0$, where $\vec{x}$ is the vector containing the variables, and A is a matrix of coefficients
  - The vector is multiplied 2 times to reflect the fact that the expression is quadratic
  - The second time the transpose is used in order to allow the product
- By rescaling A, we can obtain the standard form $\vec{x}^tA\vec{x} = 1$
- The matrix that describes a quadratic form is always symmetric
  - If it is not singular (non-invertible), it can be diagonalised as $\Lambda = U^tAU$
  - If $\vec{x'} = U^t\vec{x}$, the quadratic form becomes $\vec{x'}^t\Lambda\vec{x'}$, defined canonical form
- In the canonical form, a $2 * 2$ Λ contains the eigenvalues of the transformation in the diagonal
  - If they are both positive, the quadratic is an ellipse
    - If they are equal, it is a circle
  - If they are of opposite sign, it is an hyperbole
  - If they are both negative, there is no real solution
- In 3d, I can get an ellipsoid, a hyperboloid of 1 sheet or an hyperboloid of 2 sheets

Calculus

Functions

- Calculus is the study of functions
  - Functions are univocal relations between the sets domain and codomain
- The function $f(x) = mx + q$ is a line passing through q at $x = 0$ with slope m
- The inverse of a function correlates $f(x)$ to x
  - It is the reflection of $f(x)$ on the line $g(x) = x$
- The function $f(x) = a^x$ is an exponential
  - It passes through 1 at $x = 0$
  - $\lim_{x\to-\infty} f(x) = 0$
  - $\lim_{x\to+\infty} f(x) = +\infty$
- The function $f(x) = \log_a(x)$ is a logarithmic function
  - It passes through 1 at $x = 0$
  - Common bases a are 10, 2 and e
  - $\lim_{x\to0} f(x) = -\infty$
  - $\lim_{x\to+\infty} f(x) = +\infty$
  - Logarithms are useful for performing products
  - $\log_a(xy) = \log_a(x) + \log_a(y)$
  - $\log_a(x^y) = y\log_a(x)$
  - $\log_a(x) = \frac{log_b(x)}{\log_b(a)}$
- Trigonometric functions
  - The cosine is an even function because $\cos(\theta) = \cos(-\theta)$
    - $\cos(\frac{\pi}{2}) = 0$
    - $\cos(0) = 1$
  - The sine is an odd function because $\sin(-\theta) = \sin(-\theta)$
  - Sine and cosine are periodical with a 2π period
  - The secant is the reciprocal of cosine
  - Trigonometric functions can be inverted only in a subdomain
  - They are continuous functions
- A function is continuous at a point if the limit at that point is equal to the value of the function at that same point
  - The composition of continuous function is a continuous function
- The intermediate value theorem: a continuous function between two points takes any possible value between them
- Discontinuities can be removed in some cases, but essential discontinuities such as oscillating points, jumps and infinities cannot be removed

Derivatives

- The slope of a line is defined as $\frac{\Delta y}{\Delta x}$
- Therefore, the slope of the secant of a function between two points $f(a)$ and $f(a+h)$ is $\frac{\Delta y}{\Delta x} = \frac{f(a+h)-f(a)}{h}$
- If we try to reduce h as much as possible we obtain the slope of the tangent at point a
  - $m = \lim_{h\to0} \frac{f(a+h)-f(a)}{h}$
  - The tangent at a point is an estimation of the rate of change of the function at that point
- The derivative of a function is another function that describes its rate of change, it takes the value of the slope of the tangent of the original function at each point
  - $f'(x)|_a = \lim_{h\to0} \frac{f(a+h)-f(a)}{h}$
- A function to be derivable must be continuous and must have one-sided derivatives defined at the end-points
  - However, there are functions that are continuous but not derivable
  - Points of non-derivability are cusps, corners, discontinuities and points with vertical tangent
- Some derivatives

  - $\frac{d}{dx}[a] = 0$
  - $\frac{d}{dx}[ax] = a$
  - $\frac{d}{dx}[x^n] = nx^{n-1}$
  - $\frac{d}{dx}[\cos(x)] = -sin(x)$
  - $\frac{d}{dx}[\sin(x)] = cos(x)$
  - $\frac{d}{dx}[e^x] = e^x$
  - $\frac{d}{dx}[a^x] = \frac{d}{dx}[e^{\ln(a)x}] = \ln a * e^{\ln(a)x} = \ln(a) * a^x$
  - $\frac{d}{dx}[\ln(x)] = \frac{1}{x}$
  - $\frac{d}{dx}[\log_a(x)] = \frac{1}{x*\ln(a)}$
- Rules for derivation
  - $\frac{d}{dx}[f(x) + g(x)] = f'(x) + g'(x)$
  - $\frac{d}{dx}[k * f(x)] = kf'(x)$
  - $\frac{d}{dx}[g(x) * f(x)] = g(x) * f'(x) + g'(x) * f(x)$
  - $\frac{d}{dx}[g(f(x))] = \frac{dg}{df} * \frac{df}{dx}$
  - $\frac{d}{dx}[\frac{g(x)}{f(x)}] = \frac{g(x)*f'(x)+g'(x)*f(x)}{f(x)^2}$
- Higher order derivatives are computed as the derivative of the derivative
  - For the second derivative of f(x) we write $f''(x) = \frac{d(df/dx)}{dx} = \frac{d^2f}{dx^2}$
  - In the same way, the third derivative $f'''(x) = \frac{d^3f}{dx^3}$ and so on for higher orders
- The second derivative reflects the convexity of the function
  - It is the rate at which the slope of the tangent increases
- Derivatives can help to study the behavior of a function
- In a function there are global and local maxima and minima, defined as extremes
  - There are NOT methods to compute global extrema, but only local ones
- A local extreme is referred to an open interval
  - The derivative at that point is 0
    - This is NOT sufficient, it can also be a flexus
  - A minimum has a derivative with positive slope when it intersects the x axis
    - In other words, the second derivative is positive
  - A maximum has a derivative with negative slope when it intersects the x axis
    - In other words, the second derivative is negative
- A critical point of a function is a point where the derivative is 0 or undefined

Integrals

- We can find the area under a curve f(x) by adding rectangles with height f(x) and width dx, in what is called a Riemann sum
  - The width dx is also called subinterval
  - The area of each rectangle will be then $A|_x = f(x) * dx$
  - If we sum the area of all the rectangles while letting dx be as small as possible we obtain a new function F(x) called integral of f(x), which for any x gives the area under f(x) from $-\infty$ to that point
    - $\lim_{dx\to 0} \sum_i f(x_i) * dx_i = \int f(x)dx = F(x)$
- The derivation process is insensitive to constants, so we have a family of integrals for any given function, that differ by a constant
  - $\int f(x)dx = F(x) + c \quad \forall\, c \in \mathbb{R}$
- The difference of the integral F(x) evaluated at point b and a is the area under f(x) between the point b and a
  - $AUC|_a^b f(x) = \int_a^b f(x) = F(b) - F(a)$
  - The portion of f(x) between a and b is called partition
- The fundamental theorem of calculus: the integral of a derivative of a function is the function itself
  - $F(x) = \int_a^b f(x)dx = \lim_{h\to 0} \sum_{k=1}^{n} (f(x) * h)$
- The areas computed by integrals have a sign (!)

  - They are positive above the x axis, negative below it
- Some integrals
  - $\int a\,dx = ax + c$
  - $\int x^n dx = \frac{1}{n+1}x^{n+1} + c$
  - $\int \frac{1}{x}dx = \ln(|x|) + c$
  - $\int e^x dx = e^x + c$
  - $\int a^x dx = \frac{1}{\ln(a)}a^x + c$
  - $\int \sin(x)dx = -\cos(x) + c$
  - $\int \cos(x)dx = \sin(x) + c$
- Rules for integration
  - $\int_a^b f(x)dx = -\int_b^a f(x)dx$
  - $\int_a^a f(x)dx = 0$
  - $\int k * f(x)dx = k * \int f(x)dx$
  - $\int_a^b f(x)dx + \int_b^c f(x)dx = \int_a^c f(x)dx$
  - $\int [f(x) + g(x)]dx = \int f(x)dx + \int g(x)dx$
- Finding derivatives is easy because of the chain rule, but finding integrals is hard
- Sometimes it is possible to solve an integral using substitution to rewrite a function that can be integrated
  - Usually we can substitute a polynomial in x with the new variable u
  - $u = P(x)$
  - Since $\frac{du}{dx} = P'(x) \implies du = P'(x)dx \implies dx = \frac{1}{P'(x)}du$
  - Therefore, in general we can write $\int f(g(x))g'(x)dx = \int f(u)du|_{u=g(x)}$
- We can integrate by parts by inverting the product rule for derivatives
  - $\frac{d}{dx}[f(x) * g(x)] = f'(x)g(x) + f(x)g'(x) \implies \int [f(x) * g(x)]'dx = \int f'(x)g(x)dx + \int f(x)g'(x)dx$
  - $f(x)g(x) = \int f'(x)g(x)dx + \int f(x)g'(x)dx \implies f(x)g(x) - \int f'(x)g(x)dx = \int f(x)g'(x)dx$
  - Therefore, if we let $u = f(x)$ and $du = f'(x)dx$, and $v = g(x)$ with $dv = g'(x)dx$
  - $\int u\,dv = uv - \int v\,du$
  - This is useful when we recognize a product between a function that is the derivative of something and a function for which we know the derivative
  - It can also be used to integrate a function for which we know the derivative by adding a 1 multiplicative constant, that we will integrate

Taylor series

- Polynomials are easier than other functions to work with, therefore approximating a non-polynomial function with a polynomial is really useful
- To find a polynomial $P(x)$ of degree n that best approximates the non-polynomial function $f(x)$ around the point $x = 0$ I can proceed by layers
- The first constraint for my polynomial is that at $x = 0$ it should be equal to the original function
  - $P(x)|_{x=0} = f(x)|_{x=0}$
  - If $P(x) = c_0 + c_1x + c_2x^2 + c_3x^3 + ... + c_nx^n$, at $x = 0$ all the terms but c_0 cancel out
  - Therefore, since $P(x)|_{x=0} = c_0$ I can set $c_0 = f(x)|_{x=0}$
- In order to better approximate $f(x)$, I also want the tangent to $P(x)$ to be equal to that of the original function at $x = 0$
  - $P'(x)|_{x=0} = f'(x)|_{x=0}$
  - $P'(x) = c_1 + 2c_2x + 3c_3x^2 + ... + nc_nx^{n-1}$
  - $P'(x)|_{x=0} = c_1 \implies c_1 = f'(x)|_{x=0}$
  - $P'(x) = c_1 + 2c_2x + 3c_3x^2 + ... + nc_nx^{n-1}$
- I can also desire that the concavity of $P(x)$ be equal to that of $f(x)$
  - $P''(x)|_{x=0} = f''(x)|_{x=0}$
  - $P''(x) = 2c_2 + 3 * 2c_3x + ... + n * (n-1)c_nx^{n-2}$
  - $P''(x)|_{x=0} = 2c_2 \implies c_2 = \frac{1}{2}f''(x)|_{x=0}$
- I can proceed like this for higher derivatives to find higher-degree coefficients, until I get to the desired n

  - $c_n = \frac{1}{n!}\frac{d^n f}{dx^n} f(x)|_{x=0}$
  - The more degrees that I use, the better the approximation but the more complex the polynomial
- The infinite series of polynomial terms that approximate $f(x)$ at the point $x = 0$ is called Maclaurin series
  - Note that this nice cancellation of higher-order polynomials that allow to easily compute high derivatives happens only at $x = 0$
- In order to approximate $f(x)$ at a point $x_a \neq 0$ I can construct the polynomial so to have the variable $u = x - x_a$
  - In this way, at $x = x_a \implies u = 0$
  - This restore the nice behaviour observed at $x = 0$
  - After the expansion, I can then substitute back $u = x - x_a$
- The infinite series that generalizes the Maclaurin series at any point $x = x_a$ is called Taylor series
- We can give the Taylor series of $f(x)$ at the point x_a
  - $P(x) = \sum_{i=1}^{n} [\frac{(x-x_a)^i}{i!} \frac{d^i f}{dx^i}|_{x=x_a}]$
  - $P(x) = f(x)|_{x=x_a} + (x - x_a)\frac{df}{dx}|_{x=x_a} + \frac{(x-x_a)^2}{2!}\frac{d^2 f}{dx^2}|_{x=x_a} + \frac{(x-x_a)^3}{3!}\frac{d^3 f}{dx^3}|_{x=x_a} + ... + \frac{(x-x_a)^n}{n!}\frac{d^n f}{dx^n}|_{x=x_a}$
- The Taylor series is an infinite sum, when we consider a certain degree polynomial we call it Taylor polynomial
- The Taylor series is convergent for some functions, like $f(x) = e^x$, and divergent for others, like $f(x) = \ln(x)$
  - The maximum distance from x_a and the points where the series converges on $f(x)$ is the radius of convergence of the Taylor series

Functions in more than 1 variable

- A 2 dimensional function takes 2 inputs x, y and gives the output z
  - $z = f(x, y)$
  - They are usually represented with 3d surfaces or with level curves on the x, y plane
- A level curve or contour level is defined as the set of point that respect the constraint $f(x, y) = c$
- It is not possible to compute single derivatives of the function
- We can fix y and compute the derivative only with respect to x
  - This is a 2-dimensional function that gives the slope of the tangent to the 2d curve in the x, z plane that cuts the function at a certain value of y
  - This derivative of $f(x, y)$ is called partial derivative in x
  - $f_x(x, y) = \frac{\partial f}{\partial x} = \lim_{h\to 0} \frac{f(x+h,y)-f(x,y)}{h}$
  - It is computed like a normal derivative, but considering the fixed variable like a constant
- Of course we can do the same fixing x and taking the partial derivative in y
  - $f_y(x, y) = \frac{\partial f}{\partial y} = \lim_{h\to 0} \frac{f(x,y+h)-f(x,y)}{h}$
- It is possible to take second partial derivatives by taking the partial derivative in x of $\frac{\partial f}{\partial x}$ or the partial derivative in y of $\frac{\partial f}{\partial y}$
  - We will then obtain $f_{xx}(x, y) = \frac{\partial^2 f}{\partial x^2}$ and $f_{yy}(x, y) = \frac{\partial^2 f}{\partial y^2}$
- We can also take mixed partial derivatives by taking the derivative in x of $\frac{\partial f}{\partial y}$, or the derivative in y of $\frac{\partial f}{\partial x}$
  - The notation for these derivatives is $f_{xy}(x, y) = \frac{\partial^2 f}{\partial x \partial y}$ and $f_{yx}(x, y) = \frac{\partial^2 f}{\partial y \partial x}$
- A fundamental propriety is that the order of differentiation does not matter (!)
  - $\frac{\partial}{\partial x}[\frac{\partial f}{\partial y}] = \frac{\partial}{\partial y}[\frac{\partial f}{\partial x}] \iff f_{xy}(x, y) = f_{yx}(x, y)$
- We can generalise this concepts to a function of n variables $f(x_1, x_2, ..., x_n)$
  - In any case a consider all the variables constant except the one that I am deriving
  - I will have n first order partial derivatives for a function with n inputs

Gradient and Hessian

- The vector containing all the first partial derivatives of a function is called gradient of that function, indicated with ∇
  - $\nabla_{f(x_1,x_2,...,x_n)} = (\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}, ..., \frac{\partial f}{\partial x_n})$
  - The gradient evaluated at any point $\nabla f|_{\vec{x_a}}$ is orthogonal to the contour level at that point
  - The gradient represents the direction of steepest ascent, because it is a vector with components which ones evaluated at a certain point are the slope of the tangent along the 2 axes
  - We can see the gradient as a vector field, which always points in the direction that maximizes the increase in f, with length proportional to that increase
- The matrix containing all the second partial derivatives of a function is called Hessian of that function, indicated with H
  - $H_{f(x_1,x_2,...,x_n)} = \begin{pmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 \partial x_2} & ... & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f}{\partial x_2 \partial x_1} & \frac{\partial^2 f}{\partial x_2^2} & ... & \frac{\partial^2 f}{\partial x_2 \partial x_n} \\ ... & ... & ... & ... \\ \frac{\partial^2 f}{\partial x_n \partial x_1} & \frac{\partial^2 f}{\partial x_n \partial x_2} & ... & \frac{\partial^2 f}{\partial x_n^2} \end{pmatrix}$
  - Since the mixed partial derivatives are insensitive to the order of differentiation, the Hessian is always symmetric

Taylor expansion of a function in more than 1 variable

- The Taylor expansion of grade 2 of a multivariable function can be represented with gradient and Hessian
  - We can consider the function as operating on a vector of variables $\vec{x}$, with dimensionality equal to the number of variables
    - $f(x_1, x_2, ..., x_n) = f(\vec{x})$
  - Let $\vec{x_a}$ be the point at which we want to do the expansion
    - $f(\vec{x}) = f(\vec{x_a}) + (\nabla f|_{\vec{x_a}})^t * (\vec{x} - \vec{x_a}) + \frac{1}{2}(\vec{x} - \vec{x_a})^t H(\vec{x} - \vec{x_a}) + ...$
- Terms with degree higher than 2 will require the analogous of the gradient and Hessian at higher dimensions
  - These will be multi-dimensional object difficult to treat

Local extrema in multi-variable functions

- The local extrema of a multivariable function are those points where the gradient of the function is 0
  - In those points, the function has an horizontal tangent along all the axes
- In order to understand if a critical point $\vec{x_a}$ where $\nabla f|_{\vec{x_a}} = 0$ is a local maximum or a local minimum, I need to consider the behavior of the Hessian at that point
  - This is analogous to the evaluation of the second derivative in a standard function in order to understand its concavity
- Since the Hessian is a n*n symmetric matrix, it has n real eigenvalues
  - If all the eigenvalues of the Hessian are positive, the point is a minimum
  - If all the eigenvalues of the Hessian are negative, the point is a maximum
  - If the eigenvalues of the Hessian are some positive and some negative, the point is a saddle point is $\mathbb{R}^2$, and a more complex shape in other dimensions
- When the Hessian is not diagonal, it can be diagonalized so that we can have the eigenvalues on the diagonal
  - This corresponds to rotating the system of reference so to align it to the directions of more rapid change in concavity
  - Note that I do not need to go back to the previous system of reference, from the Hessian I only want a qualitative information on the concavity, not a number

Constrained optimization problems (Lagrange)

- A constraint optimization problem deals with finding the extrema of a function on more than one variable subjected to a constraint
  - This has many applications, for example we may want to maximize a function while there is a physical constraint on the variables that we cannot avoid
  - The constraint is a set of points that respects a condition
    - An example is the constraint $g(x, y) = x^2 + y^2 - 1 = 0$, which means that we are limited to the points on the unit circle in the x,y plane
  - In 3d we can project the constraint on the surface of our function $f(x, y)$ and the problem translates to finding the highest point on the circle
- It is easier to visualize the problem in term of contour lines in 2d
  - The solution consists then in finding a point where a level curve is tangent to the constraint
- Generalizing to any number of dimensions, if we consider that a level curve is an object that respects the condition $f(\vec{x}) = c$, the constraint optimization problem can be translated in finding the points of tangency between the constraint curve $g(\vec{x}) = \vec{0}$ and the level curves $f(\vec{x}) = \vec{c}$
- Since the gradient of a function is perpendicular to the contour level, we can solve the problem by finding the points where the gradient of the function and that of the constraint are parallel
  - If the 2 vectors are parallel, they are the same vector scaled by a constant
  - $\nabla\, g(\vec{x})|_{\vec{x_a}} = \lambda * \nabla\, f(\vec{x})|_{\vec{x_a}}$
  - The factor λ is called Lagrange multiplier
- We have therefore n+1 variables to find (all the coordinates of $\vec{x}$ and λ) and n+1 equations
  - n equations are embedded in $\nabla\, g(x, y)|_{x_a, y_a} = \lambda * \nabla\, f(x, y)|_{x_a, y_a}$
  - The remaining equation is the constraint itself $g(x, y) = 0$
- All these equations can be expressed in a compact way with a new function, called Lagrangian
  - $\mathcal{L}(\vec{x}, \lambda) = f(\vec{x}) - \lambda * g(\vec{x})$
  - It is not necessary to form the Lagrangian when computing by hand, we can just solve a normal system of equations
  - In computational applications however, computers are much faster in solving the gradient of the Lagrangian, and this is a much more compact way of representing the same information
- The solutions of the constrained optimization problem then corresponds to finding the values for $\vec{x}$ and λ for which the gradient of the Lagrangian is 0
  - $\nabla\mathcal{L}(\vec{x}, \lambda) = \vec{0}$
  - The solutions $\vec{x}$ can be more than one, however they are usually in a number that can be easily computed
  - I can then use my solutions as an input for $f(\vec{x})$ and find the one that is higher, or lower, depending what I am looking for
  - The Lagrange multiplier λ associated with a specific solution tells me how much the function $f(\vec{x})$ is sensitive to variations in the constraint $g(\vec{x}) = \vec{0}$
- If we have multiple constraints, we will have a Lagrange multiplier for each constraint

Information entropy

- The constrained optimization problem can be used to solve the problem of information entropy: maximize how much information a signal carries
- We can think of a signal with n different possible values $v_i\ ,\ \ i = 1...n$
- Each value can occur with the respective probability p_i
- The Shannon entropy for the signal is given by the function $S = -\sum_i [p_i \ln(p_i)]$
- If we want to maximize the information carried by the signal, we have to maximize the function S
- Since we are talking about probabilities, we are under the constraint $C = \sum_i [p_i] - 1 = 0$
- We have a function and a constraint: we can write the Lagrangian
  - $\mathcal{L} = S + \lambda(C) = -\sum_i [p_i \ln(p_i)] - \lambda(\sum_i [p_i] - 1)$
- By solving $\nabla\mathcal{L} = \vec{0}$ we find that $p_i = e^{-1-\lambda}$
- All the p_i are equal and they must sum to 1, therefore $p_i = \frac{1}{n}$

- By substituting on the original function then we find that the maximum information content of the signal is $S_{max} = -\sum_i [\frac{1}{n} \ln(\frac{1}{n})] = \sum_i [\frac{1}{n} \ln(n)] = \ln(n)$

Statistics

Set theory

- A set is an unordered collection of objects, also called space
  - $C = \{x_1, x_2, x_3, ..., x_n\}$
- An object that belongs to a set is said to be an element of that set
  - $x \in C$ means that the object x is an element of the set C
- A subset of a set is another set such that all the elements it contains are also contained by the main set
  - $A \subseteq B \iff [\forall\, x \in A \implies x \in B]$
  - $A \subseteq B \wedge B \subseteq A \iff A = B$
- A set without elements is called null set, denoted by $C = \phi$
- The union of 2 sets is another set containing all the element contained in one of the sets, or in both
  - $A \cup B$
  - It corresponds to a logical OR
- The intersection of 2 sets is the set containing the elements that belong to both sets
  - $A \cap B$
  - It corresponds to a logical AND
- If $A \cup B = \phi$ the 2 sets are mutually exclusive
- The complement of a subset is the set of all elements in the set but not in the subset
  - $A \subset B \implies A^c = B - A$
- De Morgan's laws
  - $(A \cap B)^c = A^c \cap B^c$
  - $(A \cup B)^c = A^c \cup B^c$
- Set union and intersection are commutative and associative
  - $A \cap B = B \cap A$
  - $A \cup B = B \cup A$
  - $(A \cap B) \cap C = (A \cap B) \cap (B \cap C)$
  - $(A \cup B) \cup C = (A \cup B) \cup (B \cup C)$

Probability

- Probability is a mathematical model for random phenomena
- A phenomenon is probabilistic if the outcome of an experiment is uncertain, but over large numbers we observe a regular distribution
- An experiment is any procedure that can be repeated in theory an infinite number of times and has a well-defined set of possible outcomes
- The sample space of an experiment is the set of all its possible outcomes
- An event is a subset of the sample space
- In a frequency approach, probability is the ratio between the number of favorable events and that of total events
  - Let C be the sample space of an experiment such that $C = \{E_1, E_2, E_3, ..., E_n\}$, containing n elements
  - Let F be an event subset of C containing all the outcomes that are considered favorable, $F \subseteq C$
  - Let F contain m elements
  - Then, the probability that the event F will happen is given by $P(F) \approx \frac{m}{n}$
- Probability can also be viewed as the confidence in an event happening
- A probability is a number between 0 and 1
  - $0 \leq P(E_i) \leq 1$
- The sum of the probabilities of all the possible outcomes of an experiment is equal to 1, meaning that there will definitely be 1 outcome

  - $P(\cup E_i) = 1$
- The probability of 1 of 2 events happening is equal to the sum of the probabilities minus their intersection
  - $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
- If 2 events are mutually exclusive (they do not have an intersection!), the probability of their union is the sum of their probabilities
  - $P(A \cup B) = P(A) + P(B)$
- The intersection of 2 events that are independent and not mutually exclusive is the product of their probabilities
  - $P(A \cap B) = P(A) * P(B)$
- The conditional probability of A given B is represented as $P(A|B)$
  - If the events are independent, $P(A|B) = P(A)$
  - If they are not independent, $P(A|B) = \frac{P(A \cap B)}{P(B)}$
  - Conditioning on an event means that the total event space is reduced to that event
- The intersection of 2 events that are NOT independent and NOT mutually exclusive is given by
  - $P(A \cap B) = P(A) * P(B|A) = P(B) * P(A|B)$
- 2 mutually exclusive events cannot be independent, and vice-versa
- The bayes formula: $P(A|B)$ is different from $P(B|A)$
  - $P(B|A) = \frac{P(A|B) * P(B)}{P(A)}$
  - Note that this is just a rearrangement of the intersection of non-independent events
- To test if A and B are independent events, we can test that both $P(A|B) = P(A) \wedge P(B|A) = P(B)$ be true
- The odds ratio of 2 events is a statistic that quantifies the strength of the association between them
  - $odd(A, B) = \frac{P(A \cap B}{P(A) * P(B)}$
  - If $OR = 1$ the events are independent
  - Frequently $\log(odd(A, B))$ is used
- A partition of the sample space U is an event E_i such that the sum of all the E_i is equal to U itself, without holes and sovrappositions
  - $U = \cup_i [E_i] \ \wedge \ E_i \cap E_j = \phi \quad \forall \, i \neq j$
- If E_i is a partition of U and A a subset of U, then
  - $P(A) = \cup_i [P(A \cap E_i)] = \cup_i [P(A|E_i)P(E_i)]$

Counting

- A permutation is an ordered arrangement of objects
- A combination is a set of objects, without considering their order
  - It is expressed as ${}_nC_r = \binom{n}{r}$, which is read "n choose r"
- We can have situations in which the same object can be drawn an infinite number of times
  - In this case we talk of replacement
  - An example is the possible permutations of letters in a 10bp DNA sequence
    - A,T,C,G are the objects, but each of them can be drawn more than once
- In other situations an object can be drawn only once
  - I take objects from a physical stack of objects: I cannot take it again after the first time
- Permutations with replacement: the sequence of numbers that I can get from 3 dice rolls
  - Let n be the number of possible outcomes
    - In the case of a dice, there are 6 possible outcomes so $n = 6$
  - Let r be the number of outcomes that I consider (the length of the sequence of outcomes)
    - For example, how many times I roll my dice
  - Then, the number of permutations p is given by $p = n^r$
    - This is the possible sequences of 3 numbers that I can obtain from 3 dice rolls ($6^3 = 216$)
- Permutations without replacement: in how many ways, considering order, I can sit 6 people in 3 chairs
  - Let n be the numerosity of my object pool
    - In this case, there are 6 people so $n = 6$
  - Let r be the number of objects that I will extract from the pool

    - For example, how many chairs do I have
  - We can reason that the first object can be 1 of the n different objects available, the second 1 of the n-1 remaining and so on
  - Therefore, the number of permutation without replacement is given by $n*(n-1)*(n-2)*...*(n-r)$
    - We can cleanly express this with factorials
  - Then, the number of permutations p without replacement is given by $p = \frac{n!}{(n-r)!}$
    - This is the possible ordered ways I can sit 6 people in 3 chairs $(6!/(6-3)! = 4*5*6 = 120)$
- Combinations without replacement: how many different unordered groups of 3 people can I get by choosing from a pool of 6 people
  - Let n be the numerosity of my object pool
    - In this case, there are 6 people so $n = 6$
  - Let r be the number of objects that I will extract from the event pool
    - For example, how many people will be in the final group that I want to extract
  - We can reason that the number of combinations without replacement is necessarily a subset of the number of permutations without replacement
    - It is the number of permutations minus the number of permutations containing the same elements in a different order
  - The number of permutations would be $\frac{n!}{(n-r)!}$, and each unique set can be expressed in $r!$ different combinations
    - The final set of combination would be $1/r!$ of the set of permutations
  - Therefore, the number of combinations without replacement is given by $c = \binom{n}{r} = \frac{n!}{(n-r)!r!}$
    - This is the number of possible different groups of 3 people that I can form from a pool of 6 people $(6!/[(6-3)!3!] = 4*5*6/2*3 = 120/6 = 20)$
- Combinations with replacement: the unordered set of numbers that I can get from 3 dice rolls
  - Let n be the number of possible outcomes per event
    - In the case of a dice, there are 6 possible outcomes so $n = 6$
  - Let r be the number of outcomes that I consider
    - For example, how many times I roll my dice
  - We can reason that the event pool is made of n objects that get regenerated when I choose one of them, but not for the last one since I will not choose after that
    - Therefore, we can choose r objects among n+r-1 non-replaceable objects
  - Then, the number of combinations with replacement is given by $c = \binom{n+r-1}{r} = \frac{(n+r-1)!}{(n-1)!r!}$
    - This is the possible sets of 3 numbers that I can obtain from 3 dice rolls $({}_8C_3 = 8!/(5!*3!) = 56)$

Discrete distributions

- Probabilities can be described with distributions
- I represent with a capital letter the random variable, with a normal letter one of its values
- A random variable is a way of mapping the outcome of an experiment to a number
- Discrete variables can only take specific values
  - $x_i \in I = \{x_i, i \in \mathbb{N}\}$
  - The probability distribution is defined by the function $f(x_i)$, which for every x_i gives the corresponding probability
  - The cumulative distribution is given by the function $F(x_i)$, which for every x_i gives the probability for a value $\leq x_i$
- A discrete distribution can be normalized, meaning that it is rescaled so to have the total sum of probabilities equal to 1
  - $\sum_{i \in I} f(x_i) = 1$
  - $F(x_{max(i)}) = 1$
- Discrete distribution are represented with histograms
- The probability mass function of a discrete distribution (PMF) is a function that given a value of the random variable X, it produces a probability of that occurring

  - It is what defines the distribution itself
  - If we plot the random variable and the PMF (the probability!) I obtain the histogram of the distribution
- The mode of a discrete distribution is the value that occurs with the highest probability
  - It is the highest peak of the histogram
  - If there are 2 peaks, the distribution is called bimodal
- The median of a discrete distribution is the value of the random variable for which $P(X > x_{med}) = P(X < x_{med}) = \frac{1}{2}$
  - The first step in computing the median is to order the observations from lowest to highest
  - If the number of observations is odd, the median is the middle value in the series
  - If the number of observations is even, the median is the average of the 2 central observations
  - In the same way, the values of x that split the distribution in quarters is called quartile, in fifths quintile and so on
  - The distance between the first and the third quartile is called inter-quartile range, and it is a measure of the spreading of the data
- The mean or average of a discrete distribution is the expected value of the random variable X (it can be represented as $E[X]$, μ or $< X >$)
  - It is also called first moment of the random variable, while k^{th} moment represents the expected value of x^k
  - $E[f(x)] = \sum_{i \in I} f(x_i)p(x_i)$, where $f(x)$ is a function defined over discrete random variables and $p(x)$ is the probability distribution
    - It is essentially a weighted average of the probabilities for a function of X to have a certain value
    - In the simplest case $f(x) = x$, therefore I am just taking the mean value of the variable
  - For an empirical distribution it is calculated as $E[x] = \frac{1}{n} \sum_{i=1}^{n} x_i$
- The variance is the mean squared distance of x from the mean of the distribution
  - $\sigma^2[x] = Var(x) = E[(x - E[x])^2] = E[x^2] - E[x]^2$
  - It is a measure of how much the distribution is spread out
- The standard deviation has the same meaning of the variance, but is more useful because it has the same dimensionality of the random variable x
  - $\sigma[x] = \sqrt{Var(x)}$
- Some general properties of expected values and variances
  - $E[x_1 + X_2] = E[x_1] + E[x_2]$
  - $E[x + k] = E[x] + k$
  - $E[ax_1 + bx_2] = aE[x_1] + bE[x_2]$
  - $Var(a * x) = a^2 * Var(x)$
- If and only if x_1 and x_2 are independent random variables
  - $E[x_1 * x_2] = E[x_1] * E[x_2]$
  - $Var(x_1 + x_2) = Var(x_1) + Var(x_2)$
- The covariance of 2 random variables describes how their respective variations are related
  - $Cov(x_1, x_2) = E[(x_1 - E[x_1])(x_2 - E[x_2])]$
  - If the 2 random variables are independent $Cov(x_1, x_2) = 0$
- For any pair of random variables
  - $Var(x_1 + x_2) = Var(x_1) + Var(x_2) + 2 * Cov(x_1, x_2)$
- The Bernoulli distribution models a single trial that can have 2 mutually exclusive outcomes
  - Let's call the 2 outcomes success and failure, with probabilities p and 1-p
  - Let the random variable X be 0 for failure an 1 for success
  - The PMF is $P(X = x) = p^x(1 - p)^{1-x}$
  - The Bernoulli distribution is obvious, but Bernoulli trials are at the foundations of many discrete distributions
- The binomial distribution models the number of successes in n independent Bernoulli trials
  - Let the same conditions of the Bernoulli trial hold, so 2 mutually exclusive outcomes with probability p and 1-p
  - Let the random variable X represent the number of successes in n trials

  - The PMF is $P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}$
    - $p^x(1-p)^{n-x}$ is the probability of having a specific sequence of length n with x successes
    - ${}_nC_x$ is the number of different combinations of x successes that I can get in n trials
    - So the PMF is the number of possible sequences of outcomes where X=x, times the probability of each of them
  - The mean is the number of trials times the probability of the favorable event
    - $E[X] = np$
  - The variance is
    - $Var(X) = np(1-p)$
- The Poisson distribution models how many Bernoulli successes will occur in a given unit of a continuous axis (time, volume, ...), when the probability of the success is constant
  - The random variable X counts the number of successes in the unit of time (or area, volume, ...)
  - The PMF is $P(X = x) = \frac{\lambda^x * e^{-\lambda}}{x!}$
    - We can consider that in an interval of length 1, we have a mean number of successes called λ
    - Suppose we subdivide the interval in n parts of equal length $\frac{1}{n}$
    - Therefore, if n is big enough each subinterval will have a probability of having 2 successes ≈ 0 and the probability of having a success $\frac{\lambda}{n}$
    - We can see the process as a binomial distribution with probability of success $p = \frac{\lambda}{n}$, with n trials
    - If we want the probability of having x successes in the unit interval, we can phrase it as the probability of having a success in exactly x of the subintervals
    - We therefore have $P(X = x) \approx \binom{n}{x} \frac{\lambda}{n}^x (1 - \frac{\lambda}{n})^{n-x}$
    - Let's consider $\lim_{n\to\infty}$, so what happens in infinitely many intervals that are infinitely small
    - $P(X = x) = \lim_{n\to\infty}[\binom{n}{x} \frac{\lambda}{n}^x (1 - \frac{\lambda}{n})^{n-x}] = \frac{\lambda^x * e^{-\lambda}}{x!}$
  - The mean and the variance are both equal to $E[X] = Var(X) = \lambda$
  - λ is the mean number of successes in the given unit interval
- Other discrete distributions
  - The geometric distribution describes the probability of waiting for X Bernoulli trials for seeing the first success
  - The negative binomial or Pascal distribution is a generalization of the geometric distribution that describes the probability of waiting for X Bernoulli trials for seeing n successes
  - The hypergeometric distribution models the number of successes in n trials like the binomial, but considers the case of non-independent trials
- Maximum likelihood estimation is a technique that allows fitting a distribution to my data
  - Likelihood is a concept different from probability
    - Probability refers to the observed data with respect to a certain distribution
    - Likelihood refers to having a certain distribution parameter given the data
  - I want to find the best parameters t for my data d, supposing they follow a certain distribution m
  - I can solve for $t^* = argmax[P(d|t, m)]$
  - Supposing that my model is $m(d|t)$, I have a likelihood function $L(t|d, m)$
  - Since I want to find a maximum of this function, I can set $\frac{dL}{dt} = 0$ and find the correspondent t^*

Continuous distributions

- Continuous variables can take infinite values in any interval
  - $x \in I \subseteq \mathbb{R}$
  - In a continuous distribution the function does NOT represent a probability, but a probability density (!)
- It is meaningful to talk about probability only over an interval (!)
  - $p(x = x_i) = 0 \; \forall \; x_i$
- The cumulative probability is $F(x) = \int f(x)dx$, and for a normalize random variable $F(x) = 1$

- The PDF is the derivative of the cumulative probability $f(x) = \frac{dF}{dx}$
- The mean is $E[x] = \int x * f(x)dx$
  - It is the continuous version (integral) of a weighted sum
- The variance is $Var(x) = E[(x-\mu)^2] = \int (x-\mu)^2 * f(x)dx$
- The uniform probability distribution has a constant value in an interval and 0 outside of it
  - The PDF is $\begin{cases} f(x) = \frac{1}{b-a} & a \leq x \leq b \\ f(x) = 0 & x \leq a \wedge x \geq b \end{cases}$
  - The mean is $E[x] = \frac{a+b}{2}$
  - The variance is $Var[x] = \frac{(b-a)^2}{12}$, it can be derived by integration
- The normal or Gaussian distribution is the most important continuous distribution
  - For the central limit theorem, the sum of a number of random variables approaches a normal distribution as the number of variables increases
  - The PDF is $f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$
  - To say that a random variable follows a certain normal distribution, we can write $X \sim N(\mu, \sigma^2)$
  - The mean of the normal distribution is $E[x] = \int x * f(x)dx = \mu$
  - The variance of the normal distribution is $Var(x) = \int (x-\mu)^2 * f(x)dx = \sigma^2$
  - The points $\mu - \sigma$ and $\mu + \sigma$ are the flexus points of the curve
    - The integral between $i\mu + \sigma$ and $\mu - \sigma$ is $\approx 68\%$
  - The integral between $\mu - 2\sigma$ and $\mu + 2\sigma$ is $\approx 95\%$
  - The integral between $\mu - 3\sigma$ and $\mu + 3\sigma$ is $\approx 99.7\%$
  - The standard normal distribution is a normal distribution with mean 0 and variance 1, indicated with z
    - $z \sim N(0, 1)$
    - Any distribution can be normalized by introducing the standardized variable z such that $z = \frac{x-\mu}{\sigma}$
- Central limit theorem: the sample mean is normally distributed for large sample sizes, regardless of the original distribution of the population
  - We can assume that our data is normally distributed if the sample size is large enough
- The normal distribution cannot be integrated analytically, therefore the integral is computed with tables for the standard normal distribution or via software
- In an n-dimensional space the normal distribution is computed with vectors and matrices
- The estimated mean of the distribution of a set of normally distributed data is $\frac{1}{n}\sum_{i=1}^{n} x_i$
  - This can be derived by a long integration of the formula for the normal distribution
  - This is an unbiased estimate
- The estimated variance of the distribution of a set of normally distributed data is $\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \mu)^2$
  - The sample variance underestimates the variance of the population, as can be derived by expanding the variance formula
  - Sometimes the variance of the sample is named S^2 to differentiate it from the real variance
- The sample mean $\bar{x}$ is normally distributed with mean equal to the mean of the distribution and variance equal to the variance of the distribution divided by n
  - $E[\bar{x}] = \mu$
  - $Var[\bar{x}] = \frac{1}{n}\sigma^2$
- The sample variance is distributed following the χ^2 distribution with n-1 degrees of freedom
  - The degrees of freedom $df = n - 1$ are the number of independent x_i
- The χ^2 distribution is a family of asymmetrical distribution for different degrees of freedom which models the distribution of the sample variance of a normal population
  - The distribution of the square of a normal distribution is a χ^2 distribution with $df = 1$
- The Student t distribution is the distribution of the variable $t = \frac{z}{\sqrt{\frac{u}{v}}}$, where z is a standard normally distributed random variable, u is a random variable with χ^2 distribution with v degrees of freedom, and z and u are independent
  - It is used in the Student t-test to compare 2 sample means and determine the probability that they come from the same population

  - If we draw n independent observation from a normal population the quantity $\frac{\bar{X}-\mu}{S/\sqrt{n}}$ has a t distribution with $df = n - 1$
  - It is similar to the normal distribution, but it is more spread out
  - Its PDF is quite complex
  - It converges to the normal distribution for $k \to +\infty$
  - Its areas are computed via tables or software
- The Gumbel or extreme value distribution is used to model the extreme values of a number of samples of various distributions
  - It was invented to model the distribution of extreme temperatures during the year, and it is used for modeling rare phenomena
  - The PDF is the exponential of an exponential, it decreases very, very fast
    - $f(x) = e^{-e^{(x-\mu)/\beta}}$
  - The Gumbel distribution is used by BLAST for calculating the E-value
    - The expected score E of a match in a database is the number of times that my sequence would obtain a score S higher than the observed one in a random database of the same size

---

rewied until here

Hypothesis testing

- When performing an experiment, before analyzing the result I need to formulate hypotheses and assumptions
  - I should state a null hypothesis that I want to disprove
  - I should state what are my assumptions: statistical independence, distribution of the data...
  - I can compute the needed statistics from my data
  - I can check the probability of observing my statistics under the null hypothesis, for the distribution given by my assumptions
  - If my statistics are more extreme than the critical value chosen, I can reject the null hypothesis
  - If my statistics are not more extreme, I fail to reject the null hypothesis
- I can never accept the null hypothesis: it is assumed true from the beginning, and failing to disprove does not mean to prove a hypothesis
  - This is a logical fallacy known as argument from ignorance
- The p-value is the conditional probability of observing a result more extreme or equal to the test statistic given the null hypothesis
  - $p = P(X \geq x | H_0)$
  - It was invented by Fisher as a rough estimate of the strength of evidence against the null hypothesis
- The significance of a test is the probability that the test rejects H_0 if it actually holds
  - $\alpha = P(false\ positives)$
- The power of a test β is the probability that the test rejects a false null hypothesis
  - $\beta = P(falsenegative)$
- In any test, we want α and β as small as possible
- The threshold (critical value) used depends on the significance that we desire
  - In biology it is common to use $p = 0.05$
  - In physics it is used $p = 10^{-10}$ or even lower
- The critical value is the value for which the area under the curve from that value to the extreme of the distribution is equal to the desired p-value
- The p-value is currently heavily criticized
  - From the view of Fisher himself, the p-value is a measure of significance, importance of a result, but not a proof: it requires further testing
  - It is often misinterpreted by researchers
  - It does not give any information on the magnitude or physical meaning of the observed statistical difference
  - Two studies can have similar results but very different p-values because of different sample sizes

  - $P \leq x$ and $P = x$ have a very different meaning
  - We can and should use both sides of the distribution for calculating p-value, but sometimes one-sided values are reported
    * This gives to the p-value more assumptions based on the belief of the researcher, and it is not fair
- The Bayesian approach to data validation is an alternative to the p-value
  - It aims at giving a more useful estimate: $P(H_0|X \geq x)$
  - It is computed with the Bayes theorem from p-value, $P(H_0)$ and $P(X \geq x)$
  - I need the *a priori* probability of the observation, which usually is not available
  - The Bayesian factor is the ratio between the probability of my data give the null hypothesis and union of the probabilities of my data given any other hypothesis
- The Fisher test is aimed at determining the probability of observing the data under the assumption that the categories are independent
  - I prepare a contingency table where I put the number of subjects in each combination of categories
  - From the table I take the various coefficients and compute the probability of obtaining that table under H_0 (no association between the variables)
  - The probability is computed using the binomial
  - The output is a number between 0 and 1 which is the probability of obtaining that table
  - It does NOT give the p-value because it doesn't consider tables that are more extreme of the observed one
  - In a 2*2 table with coefficients a, b, c, d and total number of subjects n $p = \frac{(a+b)!(c+d)!(a+c)!(b+d)!}{n!a!b!c!d!}$
- When I do n tests on the same data, I can expect to have many false positives, in a way that is proportional to n
  - The simplest correction that can be made to preserve the significance is the Bonferroni correction
    * $\alpha' = \alpha/n$
    * The actual significance to be accepted is the original significance divided by the number of tests
    * Sometimes it is too strict, it fails to reject H_0 that are false
  - Another common correction is the Benjamini-Hochberg
- The means of samples from the same population are normally distributed with a mean equal to the mean of the population and variance equal to the variance of the population divided by the size of the samples
  - The sample mean is $M = \frac{1}{n}\sum_{i=1}^{n} x_i$
  - The mean of the population μ is best approximated by the mean of the sample means
    * $\mu = E[M]$
  - The variance of the sample mean is the variance of the population divided by the sample size
    * $Var[M] = \frac{1}{n}\sigma^2$
    * The unbiased variance is used, computed by $\frac{1}{n-1}\sum_{i=1}^{n}(x_i - M)^2$
- The z-test uses the z-tables to compute the probability of observing a certain standardized value
  - If our H_0 gives us a true population mean ($H_0 : M = \mu$), we can compute the standardized variable $z = \frac{M-\mu}{\sigma/\sqrt{n}}$
    * I use $\sigma/\sqrt{n}$ beacause I need the standard deviation of the sample means, not of the samples
    * Note that for the z-test I need the real variance of the population, σ^2, not its estimate based on the sample variance
  - I need to use the 2 tails if I cannot completely exclude deviations in one direction
- If instead that the actual variance of the population I use its estimate, the unbiased sample variance S, I need to use the t-distribution with n-1 degrees of freedom
  - The variable t is computed as $t = \frac{M-\mu}{S/\sqrt{n}}$
- The t-test can also be used to compare the means of 2 samples, to test if they are significantly different
  - In this case I require the t-distribution with 2n-2 degrees of freedom
  - The formula for computing the variable t is quite complex
- The ANOVA (ANalisys Of VAriance), also called F-test, is a generalization of the t-test used to compare k means
  - When comparing 2 means, the ANOVA reduces to the t-test with relation $F = t^2$

  - It assumes that the samples are random and the errors are independent, that the populations are normally distributed and that for each condition the populations have the same variance
- The one-way ANOVA compares means of different samples (treatments in the ANOVA jargon)
  - The H_0 is that all the means are equal
  - From the different treatments we can compute the means for each treatment and the global mean of all data
  - We can compute the treatment variances and the global variance in the same way
  - The total variance can be partitioned in random variation and in variation between treatments
    - The sum of squares within (SSW) is computed as $\sum_{i,j}(x_{i,j} * \mu_j)^2$
    - Each value is compared to the mean of its treatment
    - The sum of squares between (SSB) is computed as $\sum_j n_j(\mu_j * \mu)^2$
    - Each treatment mean is compared to the global mean and multiplied for the number of samples in the treatment
  - It can be proven that the sum of squares total (SST) is equal to the sum of SSW and SSB
    - $SST = \sigma^2 * n_{tot} = \sum_i (x_i - \mu)^2 = SSW + SSB$
  - We can then define the variable f : $f = \frac{MSB}{MSW}$
    - $MSW = SSW/(n_{tot} - n_{treatments})$
      - It is the mean square within
    - $MSB = SSB/(n_{treatments} - 1)$
      - It is the mean square between
  - The variable f is distributed following the Fisher-Snedecor F distribution
    - There is one distribution for each combination of degrees of freedom within and between
  - The null hypothesis can be rejected when f is high enough
  - If $f \leq 1$ the variability between is equal or lower than the variability within
- Two-way ANOVA can be used to compare the response to different combination of 2 variables
  - It can be used in different treatments defined as different combinations of doses of 2 different drugs
  - There are 3 H_0, if I call the treatments A and B
    - No difference in means due to factor A
    - No difference in means due to factor B
    - No interaction between A and B
  - It tests for the effect of the single variables and for interaction between them
  - If we have the variable A and B, SST is composed of SSW, SSB(A), SSB(B), SSB(A,B)
  - I have a SSB for each treatment level, and for any treatment combination
  - In order to be powerful, if we increase the number of treatments we need a lot of data
  - It is frequently used for the analysis of expression data
- The χ^2-test is used to determine if the unbiased sample variance is significantly different from the variance of the population
  - We can define the sum of squared distances from the mean $Q = \sum(x_i - \mu)^2$
  - Q is distributed with a χ^2 distribution with n-1 degrees of freedom
- The Pearson's χ^2 test is used to determine if some data fits a certain distribution, with certain parameters
  - $\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$
  - This variable follows a χ^2 distribution with n-s-1 degrees of freedom, where s is the number of parameters that define the specific distribution tested
  - O_i is the observed value, E_i the expected value given the model
- Non-parametric tests make very few assumptions about the data
  - They do not estimate parameters from the data and do not assume any distribution
  - They can be used with qualitative data and ordinal data
    - Ordinal data relies on median instead of mean
  - They are not susceptible to outliers
  - They are tipically used when the data are too skewed for a paramatric test
  - They are less powerfull than parametric tests, they are more susceptible to type II errors
  - Each parametric test has a non-parametric alternative

Correlation and regression

- Correlation can be measured by a coefficient ρ
  - Given the variables x and y, we can compute μ_x and μ_y
  - If we graph a point (μ_x, μ_y) we define the centroid of the data
  - The covariance is a good measure of linear dependence between x and y
    - $cov(x, y) = \frac{1}{n-1} \sum_i (x_i - \mu_x)(y_i - \mu_y)$
  - We can then define the Pearson's correlation coefficient $\rho = \frac{cov(x,y)}{\sigma_x \sigma_y}$
    - It is a number between -1 and 1 because it is rescaled by the standard deviations of x and y
    - The Pearson's coefficient tests only linear dependency (!)
      - If $\rho = 0$ we can NOT say that x and y are independent, they can have higher-order dependencies
  - The significance of the coefficient is strongly dependent on the number of points
    - The value of the coefficient *per se* does not mean anything, it has always to be tested
    - It can be tested using the Student distribution
  - The Person's coefficient is very sensitive to outliers
  - Always look at the graph before drawing conclusions, because it can be different fro what you think
- CORRELATION DOES NOT IMPLY CAUSATION
- The Spearman's correlation coefficient measure monotonic dependency
  - It is the Pearson's coefficient of the rank of the variables, it is its non-paramtric alternative
  - If we make a ranking of values from the lowest to the highest, the rank of a value is its position in the ranking
  - We have to test the significance also of this coefficient, that depends on the amount of data
- The Matthews correlation coefficient (MCC) is used for categorical variables
  - We can assign the values 0 and 1 to the categories in both values
  - We can take the Pearson's coefficient of these values, which is the MCC
  - It is used in machine learning in a table real vs predicted
- When the dependency is more complex, we can use the mutual information
  - It will be a topic for next year
- If we have data with a good Pearson correlation, we can define a linear regression
- One technique is to minimize the distance between the points and the line (best fit with least squares)
  - We want to minimize the function $f(a, b) = \sum_i [y_i - (ax_i + b)]^2$
- The same technique can be applied for fitting any polynomial
  - $y = P(x) \sum_{k=0}^{p} a_k * x^k$
- If I use a high degree polynomial I risk doing overfitting
  - I have overfitting when the number of data is of the same order of the number of parameters
  - In overfitting the values of the parameters are often absurd (really high in absolute value), without any physical meaning
- The error can be estimated considering overfitting by giving a penalty for high coefficients
- To test the quality of a model we need to use data not used for building the model itself (!)
- Cross-validation is very dangerous

Principal component analysis

- Principal component analysis is used with high-dimensional data
  - It reduces the number of dimensions, so to be able to plot the results
  - A good idea to preliminarily decrease the number of variables is to remove the ones with the lowest variance
  - If we want to find a better system of reference, we can choose the axes with 0 covariance
- I can build the covariance matrix of the variables
  - The covariance matrix is symmetric and therefore it can be diagonalized easily
  - The change of basis matrix U is orthogonal and therefore represents a rotation

  - $\Lambda = \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix}$
  - If I rank the eigenvalues from highest to lowest, I can choose to use only a subset of dimensions that maximizes the variation
- PCA can discriminate only linear dependencies (!) by rotating the frame of reference so to align it with the variation
- If a variable has a relatively high covariance, it will dominate the principal components
- It makes sense to compare variances only if they are in the same unit of measure
- I can standardize the variables so to have unit variance and 0 mean to avoid this problems
- Instead of a covariance matrix, I can do PCA with a correlation matrix
- The correlation of a variable on each of the eigenvectors is called loading
- To choose the number of principal components, a common rule is to exclude the dimension with $\lambda < 1$