Math

$0+1+2+...n = \frac{n(n+1)}{2}$
$0^2+1^2+2^2+...n^2 = \frac{1}{6}n(n+1)(2n+1)$
Sequence: a function such that sequence(S::Set{T},i::N) -> s_i::T (define an order)
Serie: the sum of the elements of a sequence (the cumsum over a given order)
Given $|x| < 1 \rightarrow \sum_{i=0}^{\infty} x^i = \frac{1}{1-x}$;
$\int e^{ax}dx = \frac{1}{a}e^{ax} \quad ;$
$\int u(x)v'(x)dx = u(x)v(x) - \int u'(x)v(x)dx$
$\int_{-\infty}^{+\infty} e^{-x^2/2}dx = \sqrt{2\pi}$
$\int 1/xdx = ln(x) + c$
$\int ln(ax)dx = xln(ax) - x \quad ;$
$ln(x) + ln(y) = ln(xy)$
$log_b(a^c) = clog_b(a) \quad \log_{b2} x = \frac{\log_{b1} x}{\log_{b1} b2}$
Circumference: $(x-a)^2 + (y-b)^2 = r^2 \quad ; sin(0) = sin(\pi) = 0$
$\binom{a+b+c}{a,b,c} = \frac{(a+b+c)!}{a!*b!*c!}$
$x = \frac{-b \pm \sqrt{b^2-4ac}}{2a}$
$lim_{n\rightarrow\infty} \left(1+\frac{\lambda}{n}\right)^n = e^{\lambda}$

- Hessian: second derivatives matrix; Gradient: vector of first derivatives; Jacobian: $I \times J$ matrix of first derivative of the i equation for the j variable
- $x'Hx \leq 0 \ \ \forall x \in R^d$ or H is diagonal and all elements negative ↔ H negative semidefinite, all eigenvalues non-positive
- Positive definite: $|D_i| > 0 \ \ \forall i \leq n$ with D_i the *i*-th leading principal minor of the Hessian
- Negative definite: $(-1)^i|D_i| > 0 \ \ \forall i \leq n$
- Vector products:
  - Inner ("dot") product: $X \in R^d \cdot Y \in R^d \rightarrow OUT = ||X|| * ||Y|| * cos(\theta) \in R^1$
  - Hadamard ("elementwise") product: $X \in R^d \odot Y \in R^d \rightarrow OUT \in R^d$
  - Outer product: $X \in R^d \otimes Y \in R^d \rightarrow OUT \in R^{(d^2)}$
  - Cross product: $X \in R^3 \times Y \in R^3 \rightarrow OUT = ||X|| * ||Y|| * sin(\theta) * n \in R^3$ (where n is the unit vector perpendicular to the plane containing X and Y)
- $\frac{\partial \int_a^x f(t)dt}{\partial x} = f(x)$; $\frac{\partial \int_x^a f(t)dt}{\partial x} = -f(x)$
- "positive semidefinite matrix" := square matrix such that $x^TAx \geq 0 \ \ \forall x \in R^d$
  - In particular are spd matrices all diagolan matrices with al non-negative entries and those that can be decomposed as $A = P^TDP$ with D a diagonal matrix with only non-negative entries and P invertible
- "positive semi-def square root": a matrix $A^{\frac{1}{2}}$ such that $A^{\frac{1}{2}}A^{\frac{1}{2}} = A$ with A being spd
  - the roots themselves are positive semi-def
  - for any positive (semi-)definite matrix, the positive (semi-)definite square root is unique.
- "ortogonal" matrix: $M^T = M^{-1}$
- (AB)ᵀ = BᵀAᵀ
- $(A^{-1})^T = (A^T)^{-1}$
- $AIB = AB$ with I the identity matrix
- $ABsC = sABC = ABCs = \ldots$ with s a scalar value
- $trace(x^Tx) = trace(xx^T)$
- $E[trace(\cdot)] = trace(E[\cdot])$S

Norm: $|| \ v \ ||^2 = v^Tv = trace(vv^T) = \sum_i v_i^2 \quad || \ v \ ||_l = (\sum_i v_i^l)^{1/l}$

Vector space

- Projection of vector a on vector b: $c = \frac{a \cdot b}{\|b\|} * \frac{b}{\|b\|}$.
- Distance of a point x from a plane identified by θ and its offset θ_0: $\|\vec{d}\| = \frac{\vec{x}\cdot\vec{\theta}+\theta_0}{\|\vec{\theta}\|}$
- Orthogonal projection of a point x on plane identified by θ and its offset θ_0: $\vec{x_p} = \vec{x} - \frac{\vec{x}\cdot\vec{\theta}+\theta_0}{\|\vec{\theta}\|} * \frac{\vec{\theta}}{\|\vec{\theta}\|}$

Vector independence:
A set of J vectors v_j are linear dependent i.f.f. there exist a vector c not all zeros such that $\sum_{j=0}^{J} c_jv_j = 0$ (note that each individual c_j is a scalar while v_j is a vector).
If a partition of the set of vectors is linearly dependent the whole set is said to be linear dependent.
Any set of J vectors of D elements with J > D is linearly dependent.
rank(AB) ≤ min(rank(A),rank(B))

Models and axioms

$(\cup_n A_n)^c = \cap_n A_n^C \quad (\cap_n A_n)^c = \cup_n A_n^C$
$P(A_1 \cup A_2 \cup A_3 \cup ...) = P(A_1) + P(A_2 \cap A_1^C) + P(A_3 \cap A_1^C \cap A_2^C) + ...$

$\mathbf{P}\Big((A \cap B^c) \cup (A^c \cap B)\Big) = \mathbf{P}(A) + \mathbf{P}(B) - 2 \cdot \mathbf{P}(A \cap B)$
$\mathbf{P}(A_1 \cap A_2 \cap \cdots \cap A_n) \geq \mathbf{P}(A_1) + \mathbf{P}(A_2) + \cdots + \mathbf{P}(A_n) - (n-1)$

Conditioning and independence

Partition of a space: array of mutually exclusive ("disjoint") sets whose members are exhaustive ("complementary") of the space.
Joint: $P(A \text{ and } B)$ = $P(A \cap B)$ = $P(A,B)$
→ note that joint PMF/PDF are multidimensional aka multivariate (x is a vector)
Marginal (unconditional): $P(A)$
→ for PMF (PDF): we sum (integrate) over all or some dimensions to "remove" them and move from the joint toward the marginal
Conditional: $P(A|B) := P(A,B)/P(B)$
→ Valid also for PMF and PDF with respect to an event
Union: $P(A \text{ or } B) = P(A \cup B) = P(A) + P(B) - P(A \cap B)$
Note the *memoryless* of geometric/exponential: $Pr(X > t+s|X > t) = Pr(X > s)$. This is the *remaining* time, not the total time, so it is NOT the independence concept.

Multiplication rule:

- $P(A_1 \text{ and } A_2) = P(A_1 \cap A_2) = P(A_1) * P(A_2|A_1) = P(A_2) * P(A_1|A_2)$
- $P(A_1 \cap A_2 \cap ...A_n) = P(A_1) * \Pi_{i=2}^{n} P(A_i|A_1 \cap ...A_{i-1})$ → also for PMF, PDF

Total probability/expectation theorem:

- *given A being a partition*: $P(B) = \sum_i P(A) * P(B|A_i)$ → also for PMF, PDF, CDF and expectations

Bayes' rule: *given A a partition* $P(A_i|B) = \frac{P(A_i,B)}{P(B)} = \frac{P(A_i)P(B|A_i)}{\sum_j P(A_j)P(B|A_j)}$ where the first relation is by definition and the second one is for the Multiplication rule on the nominator and the total prob. theorem on te denominator
→ also for PMF, PDF

Independence: A, B indep. iff $P(A \text{ and } B) \equiv P(A \cap B) = P(A) * P(B)$ eq. $P(A|B) = P(A)$, equiv. $P(B|A) = P(B)$
(a) Indep is symmetric. (b) A collection of event is indep if *every* collection of distinct indices of such collection is indep. (oth. could be pairwise indep.)
→ also for PMF, PDF, CDF and expectations (but for all x, y!)

Union rule (De Morgan's law again):
$P(A_1 or A_2 or A_3 or ...) = P(A_1 \cup A_2 \cup A_3 \cup ...) = P(A_1) + P(A_2 \cap A_1^C) + P(A_3 \cap A_1^C \cap A_2^C) + ...$
$P(A_1 \cup A_2 \cup A_3 \cup ...) = 1 - P(A_1^C \cap A_2^C \cap A_3^C ...)$

Counting

Ways to *order* n elements ("permutations"): $n!$
Ways to *partition* n elements:
(**a**) in 2 subsets: (**a.1**) defining n_1: $\binom{n}{n_1}$; (**a.2**) Without defining n_1: $2^n = \sum_{i=0}^{n} \binom{n}{i}$; (**b**) in K subsets: (**b.1**) Defining the $k_1, k_2, ..., k_K$ elements of each subset: $\binom{n}{k_1,k_2,...,k_K}$; (**b.2**) without defining the number of elements of each subset: $\left\{ {n \atop K} \right\}$ (Sterling); (**c**) without specifying the number of subsets K: Bell numbers
Note that the partitioning problem with the ks all 1 is the problem of ordering a unique set considering each position a "slot".

Ways to sample k elements from a n elements bin: (**a**) with replacement: (**a.1**) order matters: n^k; (**a.2**)order doesn't matter: $\binom{n+k-1}{k}$; (**b**) without replacement: (**b.1**) order matters: $k! * \binom{n}{k}$; (**b.2**) order doesn't matter: $\binom{n}{k}$.

Probability to sample in n attempts x elements of a given type from a bin of s elements of that type out of total k elements: (**a**) with replacement: `Binomial(x;n,s/k)` ; (**b**) without replacement: `Hypergeometric(x; s, k-s, n)` , i.e. $\frac{\binom{s}{x}\binom{k-s}{n-x}}{\binom{k}{n}}$ (this reduces to $\frac{\frac{s!}{(s-n)!}}{\frac{k!}{(k-n)!}}$ for the probabilities to have *all* n elements sampled of the given type)

Distributions

Random variable

→ Associate a numerical value to every possible outcome
→ "Discrete" refers to finite or countable infinite values of X, not necessarily integers
→ "Mixed": those rv that for some ranges are continuous but for some other values have mass concentrated on that values\

- $p_X(x)$: PMF: Probability Mass Function (discrete) $P(X = x) = P(\{\omega \in \Omega : X(\omega) = x\}) = p_X(x)$ ("such that")
- $f_X(x)$: PDF: Probability Density Function (continuous) $P(a \leq X \leq b) = \int_a^b f_X(x)$ (prob per "unit length" - or area, they give the rate at which probability accumulates in the vicinity of a point.) PDF can be discontinue.

- $F_X(x)$: CDF: Cumulative density function (discrete, continuous or mixed) $P(X \leq x) = F_X(x)$
- $Quantile(f) = CDF^{-1}(f)$ BUT by convention the q_α quantile indicates $quantile(1-\alpha)$ not the quantile of α: $q_{2.5\%} = 1.96$, $q_{5\%} = 1.64$, $q_{10\%} = 1.28$
- $\sum_{i=-\infty}^{x} p_X(i) = F_X(x); \int_{-\infty}^{x} p_X(i)di = F_X(x); p_X(i) = \frac{dF(x)}{di}$

→ "Random vector" a multivariate random variable in R^k. The PDF of the random vector is the joint of all its individual components.

- **Gaussian vector** All elements and any linear combination of them is gaussian distributed (e.g. are independent)

Discrete distributions:

Discrete Uniform : Complete ignorance, **Bernoulli** : Single binary trial, **Binomial** : Number of successes in independent binary trials, **Categorical** : Individual categorical trial, **Multinomial** : Number of successes of the various categories in independent multinomial trials, **Geometric** : Number of independent binary trials until (and including) the first success (discrete time to first success), **Hypergeometric** : Number of successes sampling without replacement from a bin with given initial number of items representing successes, **Multivariate hypergeometric** : Number of elements sampled in the various categories from a bin without replacement, **Poisson** : Number of independent arrivals in a given period given their average rate per that period length (or, alternatively, rate per period multiplied by number of periods), **Pascal** : Number of independent binary trials until (and including) the n-th success (discrete time to n-th success).

Continuous distributions:

Uniform Complete ignorance, pick at random, all equally likely outcomes, **Exponential** Waiting time to first event whose rate is λ (continuous time to first success), **Laplace** Difference between two iid exponential r.v., **Normal** The asymptotic distribution of a sample means, **Erlang** Time of the n-th arrival, **Cauchy** The ratio of two independent zero-means normal r.v., **Chi-squared** The sum of the squared of iid standard normal r.v., **T distribution** The distribution of a sample means, **F distribution** : The ratio of the ratio of two indep X^2 r.v. with their relative parameter, **Beta distribution** The Beta distribution, **Gamma distribution** Generalisation of the exponential, Erlang and chi-square distributions

Expected value

→ The mean we would get running an experiment many times

- $E[X] := \sum_x x p_X(x) := \int_{-\infty}^{+\infty} x f_X(x) dx$
- **Expected value rule**: $E[Y = g(X)] = \sum_y Y p_Y(y) = \sum_x g(x) p_X(x) \neq g(\sum_x x p_X(x)) = g(E[X])$ (in general)
- **Linearity of expectations**: $E[aX + b] = aE[X] + b$; $E[X + Y + Z] = E[X] + E[Y] + E[Z]$
- **X,Y independent**: $E[XY] = E[X]E[Y]$, $E[g(X)h(Y)] = E[g(X)]E[h(Y)]$
- **Law of Iterated Expectations**: $E[E[X|Y]] := \sum_Y E[X|Y] p_Y(y) = E[X]$ ($E[X|Y]$ is seen as a function $g(Y)$)
- Expectations of convex functions are convex
- The expectations of an indicator function is the prob that the event indicated is true

Variance

- $Var(X) := E[(X - \mu)^2] = \sum_x (x - \mu)^2 p_X(x) = \int_{-\infty}^{+\infty} (x - \mu)^2 f_X(x)$
- $Var(X) = E[X^2] - (E[X])^2$
- $Var(g(X)) = E[g(X)^2] - (E[g(X)])^2$
- $Var[aX + b] = a^2 Var[X];$
- var of sum of r.v.:
    - X,Y independent: $Var(X + Y) = Var(X) + Var(Y)$
    - in general: $\mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \mathrm{Var}(X_i) + 2 \sum_{i<j} \mathrm{Cov}(X_i, X_j) = \sum_{i,j} \mathrm{Cov}(X_i, X_j)$
- **Law of total Variance**: $var(X) = E[var(X|Y)] + var(E[X|Y])$ (expected value of the variances of X *within* each group Y + variance of the means *between* groups)

Moments

- $M_n = \int_{-\infty}^{\infty} (x - c)^n f_X(x) dx$
- c = 0 → "raw moment" (e.g. E[X], E[X^2], ...)
- c = μ → "central moment" (e.g. var(X) (second), skewness (third), kurstsis (fourth))
- **Mode** -> argmax(f_x)
- **Median** -> $k : \int_{-\infty}^{k} f_X(x) dx = \int_{k}^{\infty} f_X(x) dx$
- **Mean** -> the expected value

Covariance and correlation

- 2 r.v.: $Cov(X, Y) := E[(X - E[X])(Y - E[Y])] = E[(X)(Y - E[Y])] = E[XY] - E[X]E[Y]$
- $Cov(aX + bY + c, eZ) = ae\, Cov(X, Z) + be\, Cov(Y, Z)$
- X random vector:
    - $Cov(X) := E[(X - E[X])(X - E[X])'] = E[XX'] - E[X]E[X]'$
    - $Cov(AX + b) = ACov(X)A'$ (all cov matrix are positive definite and so it's ok to take square roots of them)
- Correlation coeff.: $\rho := \frac{cov(X,Y)}{\sqrt{var(X)var(Y)}} = E[\frac{X - E[X]}{\sigma_X} * \frac{Y - E[Y]}{\sigma_Y}]$ with $-1 \leq \rho \leq +1$ and $\sigma_X, \sigma_Y \neq 0$
  $(X, Y)\ \text{indep.} \rightarrow cov(X, Y) = 0 \leftrightarrow \rho = 0$ (but not ←)
  $|\rho| = 1 \leftrightarrow (X - E[X]) = c(Y - E[Y])$ (i.e. X,Y linearly correlated)
  $\rho(aX + b, Y) = sign(a) * \rho(X, Y)$ (because of dimensionless)
  $X = Z + V, Y = Z + W, Z, V, W \text{indep.} \rightarrow \rho(X, Y) = \frac{\sigma_Z^2}{\sigma_Z^2 + \sigma_Z\sigma_W + \sigma_V\sigma_Z + \sigma_V\sigma_W}$

Normal RV

$X \sim N(\mu, \sigma^2) \rightarrow Y = aX + b \sim N(a\mu + b, a^2\sigma^2)$
$X \sim N(\mu, \sigma^2), Z \sim N(0, 1) \rightarrow P(a \leq X \leq b) = P(\frac{a-\mu}{\sigma} \leq \frac{X-\mu}{\sigma} \leq \frac{b-\mu}{\sigma}) = P(\frac{a-\mu}{\sigma} \leq Z \leq \frac{b-\mu}{\sigma}) = \Phi(\frac{b-\mu}{\sigma}) - \Phi(\frac{a-\mu}{\sigma})$
$X \sim N(0, \sigma^2) \rightarrow P(X < -a) = 1 - P(X < a)$
$X_i \sim N(\mu_i, \sigma_i^2),\ X_i\ \text{i.i.d.} \rightarrow Y = \sum_i X_i \sim N(\sum_i \mu_i, \sum_i \sigma_i^2)$
$E[Z] = 0, E[Z^2] = 1, E[Z^3] = 0, E[Z^4] = 3, E[Z^5] = 0, E[Z^6] = 13$

Derived Distributions

Function of a single R.V.

Linear function of a r.v.: $Y = aX + b$

- $p_Y(y) = p_X(\frac{y-b}{a})$ (where $\frac{y-b}{a}$ is the value of X that raises y)
- $f_Y(y) = f_{aX+b}(y) = \frac{1}{|a|} f_X\left(\frac{y-b}{a}\right)$ (area must be constant)\

Monotonic: $Y = g(x)$ with g(x) monotonic and continuous
$f_Y(y) = f_X(g^{-1}(y)) * |\frac{dg^{-1}}{dy}(y)|$

Probability Integral Transformation

Considering as "function" the CDF, this is uniformely distributed for any r.v.: $Y = g(X) = CDF_X(X) \sim U(0, 1)$
$Y = F_X(x) \rightarrow F_Y(y) = P(Y \leq y) = P(F_X(x) \leq y) = p(X \leq F_X^{-1}(y)) = F_X(F_X^{-1}(y)) = y$
For a sample $k \sim U(0, 1)$ the corresponding value over X is $x = F_X^{-1}(k)$, i.e. the inverse CDF evaluated on k.

General: $Y = g(X)$
$p_Y(y) = \sum_{\{x: g(x) = y\}} p_X(x)$
$f_Y(y) = \frac{dF_Y}{dy}(y)$ with $F_Y(y) = P(g(X) \leq y) = \int_{\{x: g(x) \leq y\}} f_X(x) dx$ (express the CDF of Y in terms of the CDF of X and then derive to find the PDF)

Function of multiple R.V.

Sum of 2 independent R.V., discrete::
$Z = X + Y\ p_Z(z) = \sum_x p_X(X = x) p_{Z|X}(Z = z|X = x) = \sum_x p_X(x) p_Y(z - x)$

Sum of 2 independent R.V., continue (convolution):
$Z = z$ in all occasions where $X = x$ and $Y = z - x$
$f_Z(z) = \int_{max(x_{min}, z - y_{max})}^{min(x_{max}, z - y_{min})} f_X(x) * f_Y(z - x) dx$

General: $Z = g(X, Y, ...)$ Find (e.g. geometrically) the CDF of Z and differentiate for Z to find the PDF.

Sum of random number of i.i.d. R.V. $Y = \sum_{i=1}^{N} X_i$ with $X_i\ \forall i$ i.i.d and indep to N
$E[Y] = E[E[Y|N]] = E[N * E[X]] = E[N] * E[X]$
$var(Y) = E[N] * var(X) + (E[X])^2 * var(N)$
$X \sim bern(p); N \sim bin(m, q) \rightarrow Y \sim bin(m, pq)$
$X \sim bern(p); N \sim pois(\lambda) \rightarrow Y \sim pois(p\lambda)$
$X \sim geom(p); N \sim geom(q) \rightarrow Y \sim geom(pq)$
$X \sim exp(\lambda); N \sim geom(q) \rightarrow Y \sim exp(\lambda q)$\

Order statistics

Y = max(X), X i.i.d -> $F_Y(y) = P(X_i \leq y) \forall_i \in [1, N] = F_X(y)^N \rightarrow f_Y(y) = N F_X(y)^{N-1} f_X(y)$
Y = min(X), X i.i.d -> $F_Y(y) = 1 - P(X_i \geq y) \forall_i \in [1, N] = (1 - (1 - F_X(y))^N) \rightarrow f_Y(y) = n(1 - N F_X(y))^{N-1} f_X(y)$

Limits

Properties of the sample mean: $\bar{X}_n = \frac{\sum_{i=1}^{n} X_i}{n}$ → $E[\bar{X}_n] = E[X], var(\bar{X}_n) = var(X)/n$ (from properties of expectation and variance)
Markov Inequality: X non neg r.v., $t > 0$ → $P(X \geq t) \leq E[X]/t$
Chebyshev Inequality: X a r.v., $t > 0$ → $P(|X - E[X]| \geq t) \leq Var(X)/t^2$

- proof: from Markov in. by considering a new r.v. $Y = (X - E[X])^2$
- corollary: the prob that a r.v. is k st.dev. away from the mean is less than $1/k^2$, whatever its distribution
- the t is the "accuracy" and the probability itself is the "confidence" in reaching the given accuracy

Hoffding's Inequality: $X_,, X_2, ..., X_n$ i.i.d. with $E[X_i] = \mu$ and $X \in [a,b]$ almost surely, and $a < b \Rightarrow P(|\bar{X}_n - \mu| \geq \epsilon) \leq 2e^{-\frac{2n\epsilon^2}{(b-a)^2}} \ \ \forall\epsilon > 0$ (with n not necessarily large)

Def of Convergences

We are interested in r.v. whose distribution is parametrised by n, i.e. in sequences of a r.v.

Of a deterministic sequence:: The sequence a_n converge to the value a if for any $\epsilon > 0$ it exists a value n_0 such that $|a_n - a| \leq \epsilon \ \ \forall n \geq n_0$, i.e. whatever small we choose ϵ, we can always find a n limit where subsequent sequences values are lees than ϵ far away to a

In distribution: $\lim_{n\to\infty} E[f(Y_n)] = E[f(Y)]$
In probability: $\lim_{n\to\infty} P(|Y_n - a| \geq \epsilon) = 0 \forall \epsilon > 0$
With probability 1 (almost sure): $P(\{w : \lim_{n\to\infty} Y_n(w) = Y(w)\}) = 1$

Convergence theorems:

- $X_n \xrightarrow{p/a.s.} a, Y_n \xrightarrow{p/a.s.} b \Rightarrow X_n + Y_n \xrightarrow{p/a.s} a + b; X_n * Y_n \xrightarrow{p/a.s.} a * b$
- Continuous Mapping Theorem: $X_n \xrightarrow{d/p/a.s.} a$, g is a continuous function, $\Rightarrow g(X_n) \xrightarrow{d/p/a.s.} g(a)$
- Stutsky's Theorem: $X_n \xrightarrow{d} X, Y_n \xrightarrow{p/a.s.} y$
    - $X_n + Y_n \xrightarrow{d} X + y \qquad X_n * Y_n \xrightarrow{d} X * y$

Law of large numbers: The sample mean converge to the pop mean

- weak l.l.n.: $\lim_{n\to\infty} P(|\bar{X}_n - E[X]| > \epsilon) = 0$
    - from the Chebyshev Inequality by using $\bar{X}_n$ and taking the limit for $n \to \infty$
    - it is a convergence in p. of $\bar{X}_n$ to $E[X]$.
- strong l.l.n.: $\lim_{n\to\infty} P(\bar{X}_n = E[X]) = 1$

Note that given $\bar{X}_n = \frac{1}{n}\sum X_i$:

- $\bar{X}_n \xrightarrow{n\to\infty} E[X]$ (LLN)
- $g(\bar{X}_n) \xrightarrow{n\to\infty} g(E[X])$ (cont. map. theorem)
- $Y = g(x) \Rightarrow \bar{Y}_n = \frac{1}{n}\sum Y_i = \frac{1}{n}\sum g(X_i) \xrightarrow{n\to\infty} E[Y] = E[g(X)]$ (LLN)

Central Limit Theorem: The distribution of the mean from i.i.d. samples converges in distribution to a Normal distribution with mean equal to the population mean and variance of the population variance divided by the sample size:
$\bar{X}_n \sim N(E[X], \sigma_X^2/n)$

- formally the CLT is stated in terms of $\frac{S_n - nE[X]}{\sqrt{n}\sigma_x} \xrightarrow{dist} \sim N(0,1)$
- multivariate CLT: $\sqrt{n} * \Sigma_X^{-\frac{1}{2}}(\bar{X}_n - \mu) \xrightarrow{dist} \sim N_d(0, I_d)$
- versions exists for identically distributed X_i or "weakly dependent" ones (dependence only local between neighbour X_i)
- X integer: consider $S_{n+1/2}$
- Approximation to the binomial: $P(k \leq S_n \leq l) \approx \Phi(\frac{l + \frac{1}{2} - np}{\sqrt{(n(1-p))}}) - \Phi(\frac{k - \frac{1}{2} - np}{\sqrt{(n(1-p))}})$

Bernoulli and Poisson random processes

Stochastic processes: a probabilistic phenomenon that evolves in time,i.e. an infinite sequence of r.v.
We need to characterise it with informations on the individual r.v. but also on how they relate (joint)
Bernoulli, Poisson → Assumptions: independence (→ memoryless), time-homogeneity

	Bernoulli	**Poisson**
Time of arrival `t`	Discrete	Continuous
Arrival rate	`p` per trial	`λ` per unit time
N# of arrivals	Binomial $p_n(n;t,p)$	Poisson $p_n(n;t;\lambda)$
Interarrival time	Geometric $p_t(t;p)$	Exponential $f_t(t;\lambda)$
Time to n^th^ arrival	Pascal $p_t(k,p)$	Erlang $f_t(t;n,\lambda)$

Fresh start: The Bernoulli or Poisson process after time N, where N is a r.v. causally determined from the history of the process, is a new Bernoulli/Poisson process with the same probabilistic characteristics as the original one.

The poisson as approximation of the binomial

Given p the probability of a successes in a single slot and n the number of slots, the expected number of successes λ is given by $\lambda = pn$.

The poisson PDF can be seen as the limit of the Bernoulli pdf when we consider smaller and smaller time slots, keeping constant the total expected number of successes for the period (that is p - on the single period - becomes smaller and smaller and the number of periods tends to ∞).
The poisson process can hence be seen as a limiting case of a Bernoulli process or, alternatively, as the process deriving from a sequence of exponential r.v..
In a small interval δ, the probability of 1 success is $\lambda\delta$ and of 0 successes is $1 - \lambda\delta$ (and negligible probabilities for more than one).

Merging

The process made by a sequence of r.v. functions of other sequences of r.v.

Merging of Bernoulli processes

$X_1^i \sim Bern(p)$, $X_2 \sim Bern(q)$, X_1 indep X_2

$Y^i = X_1^i \text{or} X_2^i \Rightarrow Y^i \sim Bern(p + q - pq)$
$Y^i = X_1^i \text{and} X_2^i \Rightarrow Y^i \sim Bern(pq)$ (both new Bernoulli processes)

The probability that observing a success in the merged process we have a success also in the orignal process 1 is:

- $Y^i = X_1^i \text{or} X_2^i \Rightarrow P(X_1^i | Y^i) = \frac{P(X_1^i, Y^i)}{P(Y^i)} = \frac{P(X_1^i)}{P(Y^i)} = \frac{p}{p+q-pq}$
- $Y^i = X_1^i \text{and} X_2^i \Rightarrow P(X_1^i | Y^i) = 1$

Merging of Poisson processes

$X_1^i \sim Poisson(\lambda_1)$, $X_2 \sim Poisson(\lambda_2)$, X_1 indep X_2
Note that differently from Bernoulli case, here the change of a match is zero.

$Y^i = X_1^i \text{or} X_2^i \Rightarrow Y^i \sim Poisson(\lambda_1 + \lambda_2)$

The probability that observing a success in the merged process we have a success also in the orignal process 1 is:

- $Y^i = X_1^i \text{or} X_2^i \Rightarrow P(X_1^i | Y^i) = \frac{P(X_1^i, Y^i)}{P(Y^i)} = \frac{\lambda_1}{\lambda_1 + \lambda_2}$

Splitting

Splitting of Bernoulli processes

Given a Bernoulli process X with probability p and an other independent Bernoulli process Y that assigns each success of X to Z_1 with probability q and to Z_2 with probability $(1 - q)$, we have:

- $Z_1^i = X^i \text{and} Y^i \to Z_1^i \sim Bern(pq) \quad Z_2^i = X^i \text{and not} Y^i \to Z_2^i \sim Bern(p(1-q))$

Z_1 and Z_2 are *not* independent !

Splitting of Poisson processes

Given a Poisson process X with rate λ and an other independent Bernoulli process Y that assigns each success of X to Z_1 with probability q and to Z_2 with probability $(1 - q)$, we have:

- $Z_1^i = X^i \text{and} Y^i, Z_1^i \sim Poisson(\lambda q)$
- $Z_2^i = X^i \text{and not} Y^i, Z_2^i \sim Poisson(\lambda(1-q))$

Note that differently from Bernoulli case, here the the two processes *are* independent, as the probability of an arrival at any given *point* in time is zero.

Summing Poisson rv

Given $X_1 \sim Poisson(p)$ (the distribution, not the process) and $X_2 \sim Poisson(q)$, and X_1, X_2 i.i.d. $\Rightarrow Y = (X_1 + X_2) \sim Poisson(p + q)$ (think as the two input r.v. as representing numbers of arrivals in disjoint time intervals).

Measures of error:

- MSE - Mean Square Error aka Quadratic Risk: $E[(\hat{\Theta}_n - \theta)^2] = E[\hat{\theta}_n - E[\hat{\theta}_n]] - (E[\hat{\theta}_n - \theta])^2 = var(\hat{\theta}_n) + (bias(\hat{\theta}_n))^2$
- Probability of error (discrete parameters): $P(\hat{\Theta} \neq \Theta) = \int_x P(\hat{\Theta} \neq \Theta | X = x) p_X(x) = \sum_\theta P(\hat{\Theta} \neq \Theta | \Theta = \theta) p_\Theta(\theta)$

Parameter estimators:

- MAP - Maximum a posteriori probability estimator: maximise the posterior of θ, i.e. $argmax_\theta$ (mode) of the prior by the likelihood from the data
- MLE - Maximum Likelihood estimator: $argmax_\theta$ (mode) of the likelihood
- LMS - Least Mean Squares, aka Bayes estimator: the expected value of the posterior. It minimises the Mean Square Error

Regression

- LSE - Least square estimator - Minimise the expecter error not between a parameter and its estimator but between a response variable and its estimate

    - 2 dim: $min_{a,b}E[(Y_i-(a+bX_i)^2] \Rightarrow b^* = \frac{\hat{c}ov(X,Y)}{\hat{v}ar(X)}, a^* = \bar{Y}_n b^* \bar{X}_n$
    - multivar: $min_{\beta}E[(Y_i-(X^T\beta)^2] \Rightarrow \hat{\beta}^* = (X^TX)^{-1}X^TY$

MLE

- Fisher inf. matrix: $I(\theta) := cov(\nabla lL(\theta)) := E[\nabla lL(\theta)(\nabla lL(\theta))'] - E[\nabla lL(\theta)]E[\nabla lL(\theta)' = -E[HlL(\theta)] = \Sigma^{-1}_{\theta^{MLE}}$
- Absint normality of MLE estimator: $\hat{\theta}_n^{MLE} \xrightarrow[n\to\infty]{(d)} N_d(\theta^\star, \frac{I(\theta^\star)^{-1}}{n})$

Method of moments

- Natural estimator: $\theta = E[g(X)] \Rightarrow \hat{\theta}_n = \frac{1}{n}\sum_{i=1}^n g(X_i)$
- $\theta = \{h_1(E[g_1(X)], E[g_2(X)], ..., E[g_K(X)]), h_2(E[g_1(X)], E[g_2(X)], ..., E[g_K(X)]), ..., h_D(\cdot)\}$ with $1 \leq K \leq D$
- $G_k = g_k(X)$, $\bar{G}_{k,n} = \frac{1}{n}\sum_{i=1}^n g_k(x_i)$, $J(H)$ is the $D \times K$ matrix of the $\frac{dh_d}{dE[g_d(X)]}$ derivatives, and $Cov(G)$ is the $D \times D$ covariance matrix between the $g_d(x)$ random variables
- $\hat{\theta}_n := \begin{bmatrix} h_1(\bar{G}_{1,n}, \bar{G}_{2,n}, ..., \bar{G}_{K,n}) \\ h_2(\bar{G}_{1,n}, \bar{G}_{2,n}, ..., \bar{G}_{K,n}) \\ ... \\ h_D(\bar{G}_{1,n}, \bar{G}_{2,n}, ..., \bar{G}_{K,n}) \end{bmatrix} \xrightarrow{(d)} \sim N_D\left(\begin{bmatrix} h_1(E[G_1], E[G_2], ..., E[G_K]) \\ h_2(E[G_1], E[G_2], ..., E[G_K]) \\ ... \\ h_D(E[G_1], E[G_2], ..., E[G_K]) \end{bmatrix}, \frac{J(H)\ Cov(G)\ (J(H))'}{n}\right)$

M-Estimation

- $\hat{\theta}^{ME} = argmin_\theta \frac{\sum_i g(x_i,\theta)}{n} \xrightarrow[n\to\infty]{(d)} N_d(\theta^\star, \frac{E[\frac{\partial^2 g(x,\theta^*)}{\partial^2\theta^*}]^{-1}\Sigma(\frac{\partial g(x,\theta^*)}{\partial\theta^*})E[\frac{\partial^2 g(x,\theta^*)}{\partial^2\theta^*}]^{-1}}{n})$
- median: $g = \mid x_\theta \mid$ or $g = hubberLoss(x-\theta, \epsilon)$
    - `huberLoss(x,δ) = abs(x) < δ ? x^2/2 : δ*(abs(x)-δ/2)`
- mean: $g = ||x-\theta||_2^2$
- α-quantile : $g = check(x-\theta, \alpha)$
    - `check(x,α) = x >=0 ? α * x : - (1- α) * x`

Hyp testing

Test type	Rej rule	p-value	Test type	Rej rule	p-value
H$_0$:θ=k; H$_1$: θ ≠ k	$\frac{\sqrt{n}\lvert\bar{X}_n-k\rvert}{\sigma} > q_{\alpha/2}$	$2\Phi\left(-\frac{\sqrt{n}\lvert\bar{X}_n-k\rvert}{\sigma}\right)$	$H_0 : \mu_y = \mu_x$; $H_1 : \mu_y \neq \mu_x$	$\frac{\lvert\bar{X}_n-\bar{Y}_n-0\rvert}{\sqrt{\frac{\sigma_x^2}{n}+\frac{\sigma_y^2}{m}}} > q_{\alpha/2}$	$2\Phi\left(-\frac{\lvert\bar{X}_n-\bar{Y}_n-0\rvert}{\sqrt{\frac{\sigma_x^2}{n}+\frac{\sigma_y^2}{m}}}\right)$
H$_0$:θ ⩽ k; H$_1$: θ > k	$\frac{\sqrt{n}(\bar{X}_n-k)}{\sigma} > q_\alpha$	1 - $\Phi\left(\frac{\sqrt{n}(\bar{X}_n-k)}{\sigma}\right)$	$H_0 : \mu_y \leq \mu_x$; $H_1 : \mu_y > \mu_x$	$\frac{\bar{Y}_n-\bar{X}_n}{\sqrt{\frac{\sigma_x^2}{n}+\frac{\sigma_y^2}{m}}} > q_\alpha$	1 - $\Phi\left(\frac{\bar{Y}_n-X_n}{\sqrt{\frac{\sigma_x^2}{n}+\frac{\sigma_y^2}{m}}}\right)$
H$_0$:θ ⩾ ; H$_1$: θ < k	$\frac{\sqrt{n}(\bar{X}_n-k)}{\sigma} < -q_\alpha$	$\Phi\left(\frac{\sqrt{n}(\bar{X}_n-k)}{\sigma}\right)$	$H_0 : \mu_y \geq \mu_x$; $H_1 : \mu_y < \mu_x$	$\frac{\bar{Y}_n-\bar{X}_n}{\sqrt{\frac{\sigma_x^2}{n}+\frac{\sigma_y^2}{m}}} < -q_\alpha$	$\Phi\left(\frac{\bar{Y}_n-X_n}{\sqrt{\frac{\sigma_x^2}{n}+\frac{\sigma_y^2}{m}}}\right)$

- T-Test (small sample): use sample variance, T distribution with n-1 d.o.f.
- Walld's test: based on MLE and l2 norm as distance:

Test type	Rej rule	p-value
H$_0$:θ=k; H$_1$: θ ≠ k	$n(\hat{\theta}_n^{MLE}-k)^TI(\theta_0)(\hat{\theta}_n^{MLE}-k) > q_{\chi^2_d(\alpha)}$	$1-PDF(\chi^2_d, \left((\hat{\theta}_n^{MLE}-k)^TI(\theta_0)(\hat{\theta}_n^{MLE}-k)\right)$

- Implicit Hypotheses test: we test a multivalue function of the estimator $g(\theta) = 0 \ \ R^d \mapsto R^k$ and use the delta method: $ng(\hat{\theta}_n)^T(Jg(\theta)\Sigma_d(\theta)Jg(\theta)')^{-1}g(\hat{\theta}_n) \xrightarrow[n\to\infty]{(d)} \chi^2_k$
- Likelihood ratio test: based on the ratio of two likelihoods. $H_0 : (\theta_{r+1}, ..., \theta_d) = (\theta^0_{r+1}, ..., \theta^0_d)$, $H_1 : (\theta_{r+1}, ..., \theta_d) \neq (\theta^0_{r+1}, ..., \theta^0_d)$ $T_n = 2(lL_n(\hat{\theta}_n) - lL_n(\hat{\theta}_n^C)) \xrightarrow[n\to\infty]{(d)} \chi^2_{d-r}$
- Goodness of fit" test for discrete rv (is this a sample from a certain PMF ?): χ^2 test on the share of elements in each k bin interpreted as MLE estimator on which we apply the Wald test: $T_n = n\sum_{j\in K}\frac{\left(\frac{\sum_{n=1}^N \mathbf{1}(x_n=j)}{N}-f_0(j;\hat{\theta})\right)^2}{f_0(j;\hat{\theta})} \xrightarrow[n\to\infty]{(d)} \chi^2_{k-d-1}$
- Goodness of fit" (non-parametric) test for continuous rv (is this a sample from a certain CDF ?): Kolmogorov-Smirnov test $T_n = \sqrt{n}\max_{n=1,...,N}\left(max\left(\mid \frac{n-1}{N} - F^0(X_{(n)}) \mid, \mid \frac{n}{N} - F^0(X_{(n)}) \mid\right)\right) \sim KSDist(n) \xrightarrow[n\to\infty]{(d)} \sim KS()$

Linear Regression

$$\begin{bmatrix} a^* = E[Y] - \frac{E[XY]-E[X]E[Y]}{E[X^2]-(E[X])^2}E[X] \\ b^* = \frac{E[XY]-E[X]E[Y]}{E[X^2]-(E[X])^2} \end{bmatrix} \Rightarrow \begin{bmatrix} \hat{a} = \bar{Y}_n - \frac{\overline{XY}_n-\bar{X}_n\bar{Y}_n}{\overline{X^2_n}-(\bar{X}_n)^2}\bar{X}_n = \bar{Y}_n - \frac{cov(X,Y)}{varX}\bar{X}_n \\ \hat{b} = \frac{\overline{XY}_n-\bar{X}_n\bar{Y}_n}{\overline{X^2_n}-(\bar{X}_n)^2} = \frac{cov(X,Y)}{varX} \end{bmatrix}$$

Par	Value	Asynt distribution
ϵ		$\sim N(0, \sigma^2 I_n)$
$Y \mid X$	$X\beta^* + \epsilon$	$N(X\beta^*, \sigma^2 I_n)$
$\hat{\beta}$	$(X^TX)^{-1}X^TY$	$\sim N(\beta^*, \sigma^2_\epsilon(X^TX)^{-1})$
$E[\lVert\hat{\beta}-\beta^*\rVert^2]$	$\sigma^2 tr((X^TX)^{-1})$	
$E[\lVert\hat{Y}-Y\rVert^2]$	$\sigma^2(n-d)$	
$\hat{\sigma}^2$	$\frac{\lVert Y-X\hat{\beta}\rVert^2}{n-d} = \frac{\sum_{i=1}^n \hat{\epsilon}^2}{n-d}$	$\sim \frac{\chi^2(n-d)\sigma^2}{n-d}$
$T_n^{(J)}$	$\frac{\hat{\beta}_j-\beta_j}{\sqrt{\hat{\sigma}^2((X^TX)^{-1})^{(j,j)}}}$	$\sim T(n-d)$

Exponential family

$$f_\theta(x) \text{ or } p_\theta(x) = e^{\sum_{k=K}^K \eta_k(\theta)T_k(x)-B(\theta)}h(x)$$
$$f_\theta(x) \text{ or } p_\theta(x) = e^{\frac{x\theta-b(\theta)}{\phi}+c(x;\phi)}$$

Canonical forms of common distributions

Distr.	θ	ϕ	$b(\theta)$	**can** $c(x,\phi)$	$g(\mu)$
Normal(μ)	μ	σ^2	$\frac{\theta^2}{2}$	$-\frac{1}{2}(\frac{x^2}{\phi}+\ln(2\pi\phi))$	μ
Poisson(λ)	$\ln(\lambda)$	1	e^θ	$-\ln(x!)$	$\ln(\mu)$
Bernulli(p)	$\ln(\frac{p}{1-p})$	1	$\ln(1+e^\theta)$	0	$\ln(\frac{\mu}{1-\mu})$
Binomial(μ=pn)	$\ln(\frac{\frac{\mu}{n}}{1-\frac{\mu}{n}})$	1	$n\ln(1+e^\theta)$	$\binom{\ln(n}{x)}$	$\ln(\frac{\frac{\mu}{n}}{1-\frac{\mu}{n}})$

For a canonical exponential function:
$\mu = E[X] = \frac{\partial b(\theta)}{\partial\theta} \quad var[X] = \frac{\partial^2 b(\theta)}{\partial^2\theta}\phi$

GLM

$$X \rightleftharpoons X^T\beta \overset{f}{\underset{g}{\rightleftharpoons}} E[Y|X=x]$$

- Model: LM: $y_i = x_i^T\beta + \epsilon \Rightarrow y_i \sim N(\mu_i = x_i^T\beta, \sigma^2)$ GLM: $y_i \sim EXPF(\mu_i = f(x_i^T\beta))$

$$\theta_i \equiv h(X_i^T\beta) = (b\prime)^{-1}(g^{-1}(X_i^T\beta))$$

- canonical link $g(\mu) = \theta$: $\theta_i \equiv h(X_i^T\beta) = X_i^T\beta$

$$lL(Y,X;\beta) = \sum_{i=1}^n \frac{Y_i\theta_i-b(\theta_i)}{\phi} + c(\cdot) = \sum_{i=1}^n \frac{Y_ih(X_i^T\beta)-b(h(X_i^T\beta))}{\phi} + c(\cdot)$$

Bayesian

$$\pi_{\theta|X_n} \propto \pi_\theta * Ln_{X_n|\theta}$$

Conjugate distributions

Experiment	Prior	Posterior
Bernoulli(p)	Beta(a, b)	Beta($a + \sum_i x_i, b + n - \sum_i X_i$)
Binomial	Beta	
Geometric	Beta	Beta
Exponential	Gamma	Gamma
Poisson	Gamma	
Normal(μ, σ^2)	Normal(θ, σ^2)	Normal($\frac{\Sigma_i x_i+\mu}{n+1}, \frac{\sigma^2}{n+1}$)

$$\tilde{\pi}(\eta) = \frac{\pi(\phi^{-1}(\eta))}{\lvert\frac{\partial\phi(\phi^{-1}(\eta))}{\partial\phi^{-1}(\eta)}\rvert} \propto \sqrt{\det\tilde{I}(\eta)}$$

- **Bayes estimator**: the posterior mean: $\hat{\theta} = \int_\Theta \pi(\theta \mid X_n)d\theta$
- **MAP** (maximum a posterior): $\hat{\theta}^{\text{MAP}} = \arg\max_\Theta \pi(\theta \mid X_n) = \arg\max_\Theta \pi\theta L_n(X_n|\theta)$