



DATA SCIENCE AVEC PYTHON

Année académique 2024-2025

Master: X-MAS-DE-4

Enseignant : David Rhenals

Clause de confidentialité : Ce cours est à usage unique des étudiants de l'EFREI, la diffusion externe de tout contenu de ce cours est strictement interdite sans l'autorisation écrite préalable de l'enseignant.

Quelques concepts fondamentaux en statistique pour la data science et mise en oeuvre avec SciPy

- **Concepts de base en statistique**
 - Population et échantillon
 - Principales type de variables (variables quantitatives et catégorielles)
 - Estimations (Moyenne, Médiane, Variance, Écart-type)
- **Distributions de probabilité**
 - Définition
 - Types de distributions de probabilité (discrète et continue)
 - Synthèse des plus utilisées distributions de probabilité
 - Exemple: estimation des probabilités à partir de données organisées sous forme tabulaire
 - Exercices: utilisation de tableaux de distributions cumulées de probabilité
- **Fondamentaux des tests statistiques**
 - Buts et type de tests statistiques
 - Concepts clés en tests d'hypothèse
 - Synthèse de procédures d'inférence pour un et deux échantillons (t-test, proportion, chi-carré)
 - Exemples d'application des tests d'hypothèse manuellement et sur python (SciPy -statsmodels)
 - Exercice: test statistique paramétrique à réaliser manuellement
- **Régression linéaire et corrélation**
 - Qu'est ce que la régression linéaire (exemples d'application)
 - Régression linéaire simple (Dédution)
 - Régression linéaire multiple (Approche Matricielle)
 - Métriques pour l'évaluation de la qualité de la régression
 - Limites de la régression linéaire
 - Exemples d'application sur python
- Les stats en pratique à travers des exercices simples sur Jupyter Notebooks à l'aide de la bibliothèque Scipy

Concepts statistiques de base

- La **population** représente l'ensemble des individus ou des observations sur lesquels porte l'étude.
- Un **échantillon** est un sous-ensemble représentatif de la population utilisé pour les analyses, car il est souvent impossible d'étudier toute la population.
- Une **variable aléatoire** est un concept central en probabilité et en statistique qui associe à chaque résultat possible d'une expérience aléatoire un label (généralement un nombre (réel ou entier))
 - **Variable aléatoire discrète (catégorielle)** : Elle prend un nombre fini ou dénombrable de valeurs distinctes (comme les résultats d'un dé lancé ou les faces d'une pièce). Par exemple, si on lance un dé à six faces, la variable aléatoire peut prendre les valeurs $\{1,2,3,4,5,6\}$, chacune avec une certaine probabilité.
 - **Variable aléatoire continue** : Elle peut prendre une infinité de valeurs sur un intervalle continu. Par exemple, la durée de vie d'une ampoule peut être modélisée par une variable aléatoire continue, où chaque durée possible est associée à une probabilité.

Estimations (Moyenne, Médiane, Variance, Écart-type, Quartiles)

- Moyenne : La somme de toutes les valeurs d'un ensemble divisée par le nombre de valeurs. C'est une mesure de tendance centrale.
- Variance : Mesure de la dispersion des valeurs par rapport à la moyenne. Elle est calculée comme la moyenne des carrés des écarts entre chaque valeur et la moyenne.
- Écart-type : Racine carrée de la variance. Il indique à quel point les valeurs d'un ensemble de données sont dispersées autour de la moyenne.
- Médiane : La valeur qui sépare un ensemble de données en deux parties égales. Si le nombre de valeurs est impair, c'est la valeur du milieu ($X_{(n+1)/2}$) ; si pair, c'est la moyenne des deux valeurs centrales $(X_{n/2} + X_{((n/2)+1)})/2$.
- Quartiles : Valeurs qui divisent un ensemble de données en quatre parties égales. Le premier quartile (Q_1) est la médiane de la première moitié, le deuxième (Q_2) est la médiane de l'ensemble, et le troisième (Q_3) est la médiane de la deuxième moitié.

Distributions de probabilité

- **Distributions de probabilité**
 - Définition
 - Types de distributions de probabilité (discrète et continue)
 - Synthèse des plus utilisées distributions de probabilité

Distributions de probabilité

- **Définition** : Une distribution de probabilité décrit comment les valeurs d'une variable aléatoire sont réparties. Elle montre les probabilités associées aux différents résultats possibles. Une distribution de probabilité doit accomplir la définition suivante:

Definition des conditions pour qu'une fonction soit une fonction de densité de probabilité (FDP)

Type de Distribution	Condition 1 : Non-négativité	Condition 2 : Somme (discrète) ou Intégrale (continue) des probabilités	Description
Distribution Discrète	$P(X = x_i) \geq 0 \quad \forall x_i$	$\sum_{i=1}^{\infty} P(X = x_i) = 1$	La probabilité que la variable prenne une valeur discrète doit être non négative et la somme doit être égale à 1.
Distribution Continue	$f(x) \geq 0 \quad \forall x \in \mathbb{R}$	$\int_{-\infty}^{\infty} f(x) dx = 1$	La densité de probabilité doit être non négative sur tout le domaine et l'intégrale doit être égale à 1.

Distributions de probabilité


■ Types de distributions :

- **Distributions discrètes** : Concernent des variables discrètes (valeurs comptables). Exemples : nombre de clients dans un magasin, résultats d'un lancer de dé.
 - **Distribution binomiale** : Modélise le nombre de succès dans une série d'essais indépendants (par ex., lancer d'une pièce de monnaie).
 - **Distribution de Poisson** : Modélise le nombre d'événements survenant dans un intervalle de temps ou d'espace donné (par ex., nombre d'appels reçus par minute dans un centre d'appels).
- **Distributions continues** : Concernent des variables continues (valeurs pouvant prendre n'importe quelle valeur dans un intervalle donné). Exemples : la taille, le poids.
 - **Distribution normale (ou gaussienne)** : Symétrique, en forme de cloche, et décrite par sa moyenne et son écart-type. Utilisée pour modéliser de nombreux phénomènes naturels.
 - **Distribution exponentielle** : Modélise le temps entre deux événements dans un processus de Poisson (par ex., temps d'attente avant le prochain appel téléphonique).

Distributions de probabilité

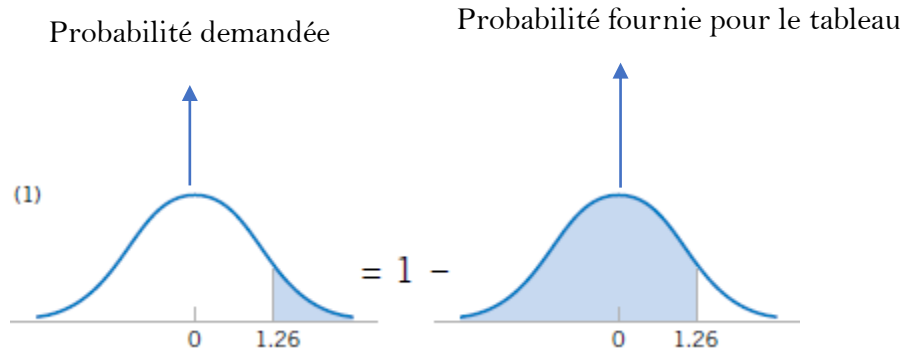
Synthèse de quelques distributions de probabilité (Discrètes et Continues)

Tableau de synthèse des formules des distributions discrètes et continues

Nom de la Distribution	Formule	Description des Paramètres	Graphique associé	Formules d'Estimation de la Moyenne, Variance, Écart-type
Distribution Normale	$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$	<ul style="list-style-type: none">- μ : Moyenne (centre de la distribution).- σ^2 : Variance (dispersion des données).	Forme en cloche, symétrique autour de μ .	 Moyenne estimée : $\hat{\mu} = \frac{1}{n} \sum_{i=1}^n x_i$ Variance estimée : $\hat{\sigma}^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \hat{\mu})^2$ Écart-type estimé : $\hat{\sigma} = \sqrt{\hat{\sigma}^2}$
Distribution Binomiale	$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$	<ul style="list-style-type: none">- n : Nombre d'essais.- p : Probabilité de succès à chaque essai.- k : Nombre de succès.	Barres discrètes, centrée autour de $n \cdot p$.	Moyenne : $n \cdot p$ Variance : $n \cdot p \cdot (1-p)$ Écart-type : $\sqrt{n \cdot p \cdot (1-p)}$
Distribution de Poisson	$P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$	<ul style="list-style-type: none">- λ : Taux moyen d'événements par intervalle.- k : Nombre d'événements observés.	Asymétrique, plus concentrée autour de petits k .	Moyenne : λ Variance : λ Écart-type : $\sqrt{\lambda}$
Distribution Exponentielle	$f(x) = \lambda e^{-\lambda x}$ pour $x \geq 0$	<ul style="list-style-type: none">- λ : Taux moyen d'occurrence d'un événement.	Courbe décroissante rapide.	Moyenne : $\frac{1}{\lambda}$ Variance : $\frac{1}{\lambda^2}$ Écart-type : $\frac{1}{\lambda}$

Exemple: Estimation des probabilités à partir de données organisées sous forme tabulaire

- On suppose qu'un variable aléatoire Z suit une loi normal standard avec $\mu=0$ et $\sigma=1$. À partir des données de probabilité dans le tableau Z, calculons la probabilité $P(Z>1.26)$



$$\Phi(z) = P(Z \leq z) = \int_{-\infty}^z \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2} du$$

Dans le tableau nous n'avons pas directement la valeur $Z=1,26$, alors il faut l'interpoler à partir des valeurs les plus proches à $Z=1,26$.

$$P(Z \leq z) = P(Z \leq z_1) + \frac{(z - z_1)}{(z_2 - z_1)} (P(Z \leq z_2) - P(Z \leq z_1))$$

$$P(Z = 1,26) = 0,88493 + \frac{(1,26 - 1,2)}{(1,3 - 1,2)} (0,903199 - 0,88493) = 0,896$$

$$P(\leq 1,26) = 1 - 0,896 = 0,104$$

Table II Cumulative Standard Normal Distribution (continued)

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.500000	0.503989	0.507978	0.511967	0.515953	0.519939	0.523922	0.527903	0.531881	0.535856
0.1	0.539828	0.543795	0.547758	0.551717	0.555760	0.559618	0.563559	0.567495	0.571424	0.575345
0.2	0.579260	0.583166	0.587064	0.590954	0.594835	0.598706	0.602568	0.606420	0.610261	0.614092
0.3	0.617911	0.621719	0.625516	0.629300	0.633072	0.636831	0.640576	0.644309	0.648027	0.651732
0.4	0.655422	0.659097	0.662757	0.666402	0.670031	0.673645	0.677242	0.680822	0.684386	0.687933
0.5	0.691462	0.694974	0.698468	0.701944	0.705401	0.708840	0.712260	0.715661	0.719043	0.722405
0.6	0.725747	0.729069	0.732371	0.735653	0.738914	0.742154	0.745373	0.748571	0.751748	0.754903
0.7	0.758036	0.761148	0.764238	0.767305	0.770350	0.773373	0.776373	0.779350	0.782305	0.785236
0.8	0.788145	0.791030	0.793892	0.796731	0.799546	0.802338	0.805106	0.807850	0.810570	0.813267
0.9	0.815940	0.818589	0.821214	0.823815	0.826391	0.828944	0.831472	0.833977	0.836457	0.838913
1.0	0.841345	0.843752	0.846136	0.848495	0.850830	0.853141	0.855428	0.857690	0.859929	0.862143
1.1	0.864334	0.866500	0.868643	0.870762	0.872857	0.874928	0.876976	0.878999	0.881000	0.882977
1.2	0.884930	0.886860	0.888767	0.890651	0.892512	0.894350	0.896165	0.897958	0.899727	0.901475
1.3	0.903199	0.904902	0.906582	0.908241	0.909877	0.911492	0.913085	0.914657	0.916207	0.917736
1.4	0.919243	0.920730	0.922196	0.923641	0.925066	0.926471	0.927855	0.929219	0.930563	0.931888
1.5	0.933193	0.934478	0.935744	0.936992	0.938220	0.939429	0.940620	0.941792	0.942947	0.944083
1.6	0.945201	0.946301	0.947384	0.948449	0.949497	0.950529	0.951543	0.952540	0.953521	0.954486
1.7	0.955435	0.956367	0.957284	0.958185	0.959071	0.959941	0.960796	0.961636	0.962462	0.963273
1.8	0.964070	0.964852	0.965621	0.966375	0.967116	0.967843	0.968557	0.969258	0.969946	0.970621
1.9	0.971283	0.971933	0.972571	0.973197	0.973810	0.974412	0.975000	0.975581	0.976148	0.976705
2.0	0.977256	0.977784	0.978308	0.978822	0.979325	0.979818	0.980301	0.980774	0.981237	0.981691
2.1	0.982136	0.982571	0.982997	0.983414	0.983823	0.984222	0.984614	0.984997	0.985371	0.985736
2.2	0.986097	0.986447	0.986791	0.987126	0.987455	0.987776	0.988089	0.988396	0.988696	0.988989
2.3	0.989276	0.989556	0.989830	0.990097	0.990358	0.990613	0.990863	0.991106	0.991344	0.991576
2.4	0.991802	0.992024	0.992240	0.992451	0.992656	0.992857	0.993053	0.993244	0.993431	0.993613
2.5	0.993790	0.993963	0.994132	0.994297	0.994457	0.994614	0.994766	0.994915	0.995060	0.995201
2.6	0.995339	0.995473	0.995604	0.995731	0.995855	0.995975	0.996093	0.996207	0.996319	0.996427
2.7	0.996533	0.996636	0.996736	0.996833	0.996928	0.997020	0.997110	0.997197	0.997282	0.997365
2.8	0.997445	0.997523	0.997599	0.997673	0.997744	0.997814	0.997882	0.997948	0.998012	0.998074
2.9	0.998134	0.998193	0.998250	0.998305	0.998359	0.998411	0.998462	0.998511	0.998559	0.998605
3.0	0.998650	0.998694	0.998736	0.998777	0.998817	0.998856	0.998893	0.998930	0.998965	0.998999
3.1	0.999032	0.999065	0.999096	0.999126	0.999155	0.999184	0.999211	0.999238	0.999264	0.999289
3.2	0.999313	0.999336	0.999359	0.999381	0.999402	0.999423	0.999443	0.999462	0.999481	0.999499
3.3	0.999517	0.999533	0.999550	0.999566	0.999581	0.999596	0.999610	0.999624	0.999638	0.999650
3.4	0.999663	0.999675	0.999687	0.999698	0.999709	0.999720	0.999730	0.999740	0.999749	0.999758
3.5	0.999767	0.999776	0.999784	0.999792	0.999800	0.999807	0.999815	0.999821	0.999828	0.999835
3.6	0.999841	0.999847	0.999853	0.999858	0.999864	0.999869	0.999874	0.999879	0.999883	0.999888
3.7	0.999892	0.999896	0.999900	0.999904	0.999908	0.999912	0.999915	0.999918	0.999922	0.999925
3.8	0.999928	0.999931	0.999933	0.999936	0.999938	0.999941	0.999943	0.999946	0.999948	0.999950
3.9	0.999952	0.999954	0.999956	0.999958	0.999959	0.999961	0.999963	0.999964	0.999966	0.999967

Exercices: Estimation de probabilités à partir de données tabulées

- En vous appuyant sur le tableau de la distribution de probabilité normale standard cumulée qui est présenté ci-dessus, veuillez estimer les probabilités suivantes:
 - $P(Z > 1.96)$
 - $P(Z > -1.37)$
 - $P(-1,25 < Z < 0,37)$
 - $P(Z < -0,86)$

Fondamentaux des tests statistiques

- Fondamentaux des tests statistiques
 - Buts et type de tests statistiques
 - Concepts clés en tests d'hypothèse
 - Synthèse de procédures d'inférence pour un et deux échantillons (t-test, proportion, chi-carré)
 - Exemples d'application des tests d'hypothèse manuellement et sur python (SciPy - Statsmodels)

Fondamentaux des tests statistiques

■ Buts des tests statistiques

- fournir un cadre statistique permettant de prendre une décision éclairée concernant une affirmation ou une hypothèse à partir de données échantillonnées.
- permet de décider, de manière objective et statistiquement justifiée, si une hypothèse concernant un paramètre de population (comme la moyenne, la proportion, ou la variance) est raisonnable à la lumière des données collectées.
- Cette méthode permet de juger si une différence observée dans les données est due au hasard (et donc compatible avec H_0) ou si elle est significative et reflète un véritable effet (supportant H_1).

■ Types de tests statistiques :

- **Paramétriques** (requièrent des hypothèses sur la distribution des données). Exemples: [t-test](#), [test de proportions](#), ANOVA, z-test,.
- **Non-paramétriques** (ne nécessitent pas de distribution particulière). Exemples: [Test du Chi-carré](#), test de Wilcoxon, test de Kruskal-Wallis .

Fondamentaux des tests statistiques (Concepts clés en tests d'hypothèse)

- **Hypothèses (H_0) nulle et alternative (H_1)**
 - **H_0** : C'est l'affirmation ou l'hypothèse par défaut que l'on cherche à tester. (par exemple, $H_0: \mu = \mu_0$)
 - **H_1** : C'est l'hypothèse opposée à l'hypothèse nulle, que l'on cherche à prouver (par exemple, $H_0: \mu \neq \mu_0$)
- **Statistique de test** : C'est un paramètre calculé à partir des données de l'échantillon qui permet de quantifier l'écart entre l'échantillon et l'hypothèse nulle (par exemple statistique t , Z , χ^2)
- **Erreurs de type I et de type II :**
 - **Erreur de type I** : Rejeter H_0 alors qu'elle est vraie (faux positif). La probabilité de faire une erreur de type I est α .
 - **Erreur de type II** : Ne pas rejeter H_0 alors que H_1 est vraie (faux négatif). La probabilité de faire une erreur de type II est β

Fondamentaux des tests statistiques (Concepts clés en tests d'hypothèse)

- **Niveau de signification (α)** : C'est un seuil fixé avant l'expérience qui détermine le risque que l'on est prêt à prendre de rejeter H_0 alors qu'elle est vraie (erreur de type I). Typiquement, $\alpha=0,05$ ou $\alpha=0,01$.
- **P-valeur** : C'est la probabilité d'obtenir une statistique de test aussi extrême ou plus extrême que celle observée, si l'hypothèse nulle est vraie. Si la p-valeur est petite (généralement inférieure à un seuil fixé, comme 0,05), cela signifie que les données sont peu compatibles avec H_0 , et on rejette H_0 .

Fondamentaux des tests statistiques (Synthèse de procédures d'inférence pour un et deux échantillons (t-test, proportion, χ^2))

Tableau 2: Tests statistiques pour un seul échantillon (Les paramètres sont comparés avec une valeur de référence)

Hypothèse nulle	Statistique de test	Hypothèse alternative	Critère de rejet (valeur critique)	Critère de rejet (valeur (p))	Calcul de la valeur (p)
$H_0 : \mu = \mu_0$ (Variance connue)	$Z = \frac{\bar{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$	$H_1 : \mu \neq \mu_0$ (bilatéral) $H_1 : \mu > \mu_0$ (unilatéral à droite) $H_1 : \mu < \mu_0$ (unilatéral à gauche)	Bilatéral : $ Z > Z_{\alpha/2}$ Unilatéral : $Z > Z_{\alpha}$ ou $Z < -Z_{\alpha}$	$p < \alpha$	Bilatéral : $p = 2(1 - \Phi(Z))$ Unilatéral : $p = 1 - \Phi(Z)$ (droite) $p = \Phi(Z)$ (gauche)
$H_0 : \mu = \mu_0$ (Variance inconnue)	$t = \frac{\bar{X} - \mu_0}{\frac{s}{\sqrt{n}}}$	$H_1 : \mu \neq \mu_0$ (bilatéral) $H_1 : \mu > \mu_0$ (unilatéral à droite) $H_1 : \mu < \mu_0$ (unilatéral à gauche)	Bilatéral : $ t > t_{\alpha/2, n-1}$ Unilatéral : $t > t_{\alpha, n-1}$ ou $t < -t_{\alpha, n-1}$	$p < \alpha$	Bilatéral : $p = 2(1 - T_{n-1}(t))$ Unilatéral : $p = 1 - T_{n-1}(t)$ (droite) $p = T_{n-1}(t)$ (gauche)
$H_0 : p = p_0$ (Proportion)	$Z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$	$H_1 : p \neq p_0$ (bilatéral) $H_1 : p > p_0$ (unilatéral à droite) $H_1 : p < p_0$ (unilatéral à gauche)	Bilatéral : $ Z > Z_{\alpha/2}$ Unilatéral : $Z > Z_{\alpha}$ ou $Z < -Z_{\alpha}$	$p < \alpha$	Bilatéral : $p = 2(1 - \Phi(Z))$ Unilatéral : $p = 1 - \Phi(Z)$ (droite) $p = \Phi(Z)$ (gauche)

Fondamentaux des tests statistiques (Synthèse de procédures d'inférence pour un et deux échantillons (t-test, proportion, χ^2))

Tableau 3: Tests statistiques pour deux échantillons (Les paramètres sont comparés entre deux échantillons)

Hypothèse nulle	Statistique de test	Hypothèse alternative	Critère de rejet (valeur critique)	Critère de rejet (valeur (p))	Calcul de la valeur (p)
$H_0 : \mu_1 = \mu_2$ (Variances connues)	$Z = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$	$H_1 : \mu_1 \neq \mu_2$ (bilatéral) $H_1 : \mu_1 > \mu_2$ (unilatéral à droite) $H_1 : \mu_1 < \mu_2$ (unilatéral à gauche)	Bilatéral : $ Z > Z_{\alpha/2}$ Unilatéral : $Z > Z_{\alpha}$ ou $Z < -Z_{\alpha}$	$p < \alpha$	Bilatéral : $p = 2(1 - \Phi(Z))$ Unilatéral : $p = 1 - \Phi(Z)$ (droite) $p = \Phi(Z)$ (gauche)
$H_0 : \mu_1 = \mu_2$ (Variances inconnues)	$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{s_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}}$ où $s_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}$	$H_1 : \mu_1 \neq \mu_2$ (bilatéral) $H_1 : \mu_1 > \mu_2$ (unilatéral à droite) $H_1 : \mu_1 < \mu_2$ (unilatéral à gauche)	Bilatéral : $ t > t_{\alpha/2, n_1+n_2-2}$ Unilatéral : $t > t_{\alpha, n_1+n_2-2}$ ou $t < -t_{\alpha, n_1+n_2-2}$	$p < \alpha$	Bilatéral : $p = 2(1 - T_{n_1+n_2-2}(t))$ Unilatéral : $p = 1 - T_{n_1+n_2-2}(t)$ (droite) $p = T_{n_1+n_2-2}(t)$ (gauche)
$H_0 : p_1 = p_2$ (Proportions)	$Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p}) \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}}$ où $\hat{p} = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1+n_2}$	$H_1 : p_1 \neq p_2$ (bilatéral) $H_1 : p_1 > p_2$ (unilatéral à droite) $H_1 : p_1 < p_2$ (unilatéral à gauche)	Bilatéral : $ Z > Z_{\alpha/2}$ Unilatéral : $Z > Z_{\alpha}$ ou $Z < -Z_{\alpha}$	$p < \alpha$	Bilatéral : $p = 2(1 - \Phi(Z))$ Unilatéral : $p = 1 - \Phi(Z)$ (droite) $p = \Phi(Z)$ (gauche)
H_0 : Les variables sont indépendantes (Chi-carré pour tables de contingence)	$\chi^2 = \sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$ où $E_{ij} = \frac{n_{i \cdot} \cdot n_{\cdot j}}{n}$	H_1 : Les variables ne sont pas indépendantes	$\chi^2 > \chi_{\alpha, (r-1)(c-1)}^2$	$p < \alpha$	$p = P(\chi^2 \geq \chi_{\text{obs}}^2)$

Exemples d'application des tests d'hypothèse manuellement et sur python (SciPy -statsmodels)

- Les exemples présentés dans ce tableau sont résolus manuellement et à l'aide du code python. Pour accéder aux exemples en format html cliquer sur chaque exemple

[Synthèse de fonctions stats SciPy \(cliquer ici\)](#)

Exemple	Type de test d'hypothèse
Exemple 1 : Régime alimentaire	t - test à deux échantillons
Exemple 2 : Préférence de boisson	χ^2 - test d'indépendance
Exemple 3: Campagne publicitaire de satisfaction	Proportion - test
Exercice à réaliser à la main	Z-test

Régression linéaire et corrélation

- Régression linéaire et corrélation
 - Qu'est ce que la régression linéaire (exemples d'application)
 - Méthode de moindres carrés
 - Régression linéaire simple (Dédution)
 - Régression linéaire multiple (Approche Matricielle)
 - Métriques pour l'évaluation de la qualité de la régression
 - Limites de la régression linéaire
 - Exemples d'application sur python

Régression linéaire et corrélation

- La régression linéaire est une méthode statistique utilisée pour modéliser la relation entre une variable dépendante (ou cible) et une ou plusieurs variables indépendantes (ou explicatives). Elle cherche à établir une équation de la forme :

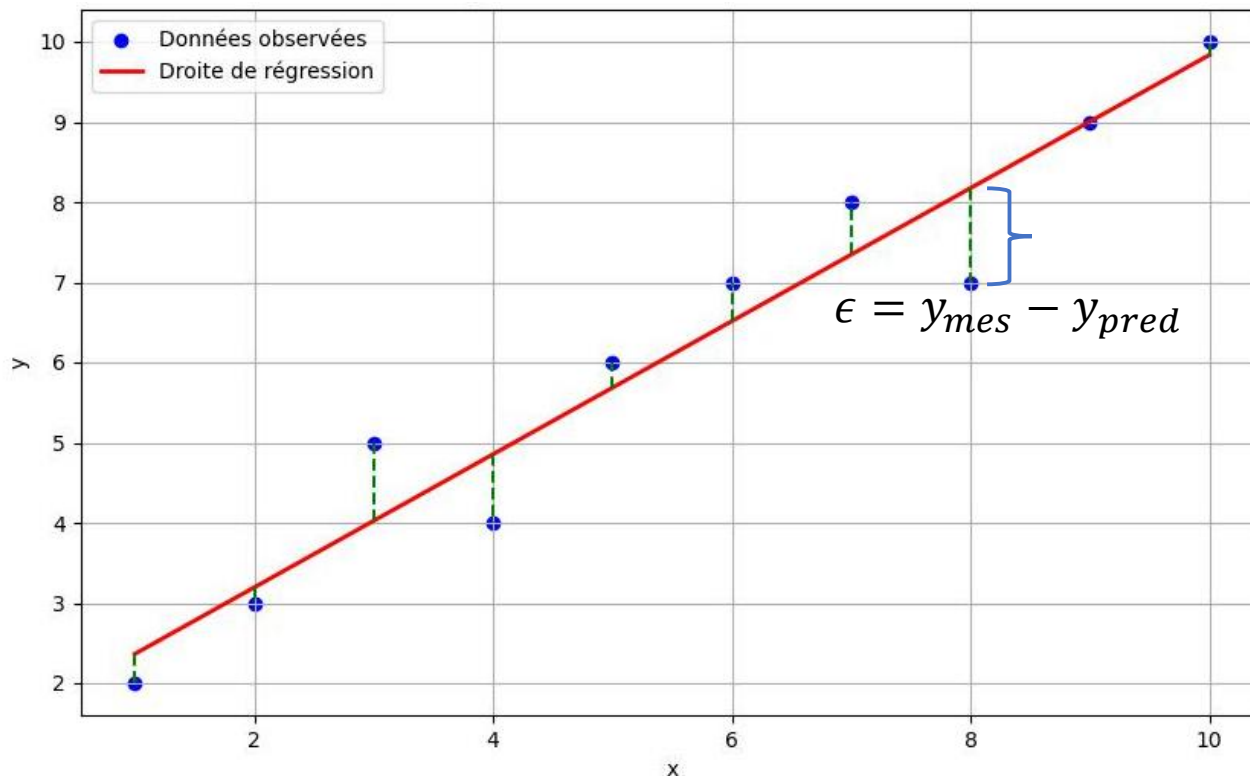
$$y = \beta_0 + \sum_{i=1}^n \beta_i x_i$$

où y est la variable dépendante (à prédire ou cible), x_i sont les variables indépendantes (prédictives), et β_i sont les coefficients de régression (indiquant la variation de y pour chaque unité de variation de x_i).

- L'objectif principal de la régression linéaire est de prédire les valeurs de y à partir de x , tout en minimisant les écarts entre les valeurs observées et les valeurs prédites. Cette méthode repose sur l'assomption que la relation entre les variables est linéaire.

Méthode des moindres carrés

- La régression linéaire vise à trouver la meilleure droite possible sur un ensemble de données mesurées en minimisant l'erreur carré. Ceci est réalisé à l'aide de la méthode des moindres carrés



Erreur Carrée ou fonction de coût

$$S = \sum_{i=1}^n (y_{i_{mes}} - y_{i_{pred}})^2 = \sum_{i=1}^n (y_{i_{mes}} - (\beta_0 + \beta_1 x_i))^2$$



Minimisation de l'erreur carrée et résolution d'un système simple pour β_0 et β_1

1 $\frac{\partial S}{\partial \beta_0} \left[\sum_{i=1}^n (y_{i_{mes}} - (\beta_0 + \beta_1 x_i))^2 \right] = 0$


2 $\frac{\partial S}{\partial \beta_1} \left[\sum_{i=1}^n (y_{i_{mes}} - (\beta_0 + \beta_1 x_i))^2 \right] = 0$

Méthode des moindres carrés (Régression linéaire simple et multiple)

- Le tableau ci-dessous contient les fichiers .html qui montrent les démonstrations et les descriptions concernant la régression linéaire simple et multiple (cliquer sur le type de régression).

Type de régression	Méthode
<u>Régression Linéaire Simple</u>	Moindres carrés (simple)
<u>Régression Linéaire Multiple</u>	Moindres carrés représentation matricielle

Métriques pour l'évaluation de la qualité de la régression

- La qualité de la régression peut être testée à partir du coefficient de détermination r^2 . Ce paramètre est une mesure du pourcentage de la variabilité des données qui peut être expliquée pour le modèle linéaire régressé.
- Les formules et une brève description des métriques concernées se trouvent dans ce fichier .html.  [← Cliquer sur l'icone](#)

Limites de la régression linéaire

- **Non linéarité des données** : Si la relation des données n'est pas linéaire, la régression linéaire peut ne pas être approprié
- **Sensibilité aux valeurs aberrantes** : Les valeurs extrêmes peuvent influencer fortement les résultats du modèle
- **Multicolinéarité**: Problème lorsque les variables indépendantes sont fortement corrélées entre elles, rendant difficile l'interprétation (très souvent on utilise la régression sur des composantes principales ou pls afin de décorréler les variables prédictives).