Cours de statistiques et probabilités, ENS Rennes

Janvier 2022 - avril 2022.
Planing.
Créneaux :

- Mardi, 14h, Ker-Lann
- Jeudi 14h, Beaulieu

1. Tronc commun, 10 séances de 2h.
   Évaluation par devoir maison a la fin du tronc commun
2. Deuxième partie, 10 séances de 2h au choix:
   - Partie pratique sur les statistiques descriptives et l'apprentissage automatique, évaluation par projets en groupe.
   - Partie théorique sur les preuves des resultats de probabilites, évaluation par examen.

Informations pratiques

Lien vers la page du cours.
Installer jupyter notebook avec numpy/scipy pour les exercices en cours.
Enseignant: Louis Thiry

- Formation en maths appliquees: analyse numerique, traitement du signal et apprentissage.
- Stage chez Deep Mind: apprentissage d'une fonctionnelle pour la mecanique quantique numerique.
- Doctorat à l'ENS Paris, directeur S. Mallat
  - Modélisation mathématiques des réseaux de neurones profonds pour la classification d'images
  - Régression de l'énergie de systèmes d'atomes (molécules, solides...)
  - dispositifs de création artistes-algorithmes
- Postdoc a l'INRIA Rennes, équipe Fluminance: analyse de données satellites sur l'océan, modèles et assimilation de données pour les océans, projet ERC STUOD.
- https://www.di.ens.fr/louis.thiry/

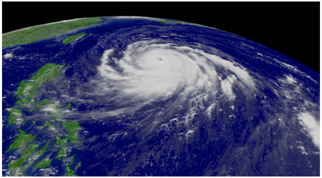
Questionnaire

*p*our savoir ou vous en êtes sur les probas et stats.

Probabilités

- Probabilités: étude et modélisation mathématique des phénomènes contenant une part de hasard / incertitude.
  ex:
  - Modéliser le lancer d'une pièce.
  - Modéliser des processus de texture

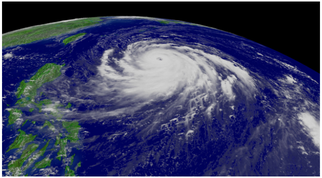
  - Évènements rares en météo



Statistiques

- Statistiques: recueillir, traiter, interpréter des données sur un phénomènes.
  - ex: déterminer si une pièce est déséquilibrée a partir d'une série de lancer.
  - Reconnaître un arbre a partir de la texture de son écorce

  - Estimer les chances qu'un cyclone s'intensifie ou disparaisse:



→ Liens forts entre les deux domaines.

Modélisation probabiliste

Objet central de modelisation: **Variable aleatoire**.

Exemples de modèles:

- Variable aléatoire réelle: temps d'attente a une caisse de magasin.
- Vecteur aléatoire: temps aux 5 caisses d'un magasin.
- Processus aléatoire (stochastique) a temps discret: taux d'occupation de la RAM, nombre requêtes sur un serveur.
- Processus aléatoire (stochastique) a temps continu: solution d'une équation aux dérivées partielle stochastique

$$f(t) = f_0 + \int_0^t \sigma \underbrace{\mathrm{dB}_t}_{\text{Brownien}}$$

Statistiques

"*Parmi les thèmes à propos desquels les statisticiens ne sont pas d'accord, se trouve la définition de leur science.*" Maurice Kendall (1907-1983).

- **Statistique descriptive** : rendre lisible un ensemble de données fini.
  ex: ensemble des notes de tous les élèves de L3 SIF a l'ENS.
- **Statistique mathématique** : Déduire des informations générale a partir d'un échantillon.
  ex: sondage électoral, étude médicale de l'efficacité d'un traitement.
  $\rightarrow$ intervalles de confiance, test statistique...
- **Statistique prédictive (explicative)**: modéliser un phénomène à partir de données ($\neq$ modélisation physique).

$$Y = F(X)$$

$\rightarrow$ apprentissage statistique.

Partie 1: tronc commun

1. Introduction aux probabilités
    - Variables aléatoires discrètes
    - Espérance et Variance
    - Indépendance et conditionnement
    - Variables aléatoires continues
    - Fonction de répartition
    - Simulation d'une v.a.
    - Loi des grands nombres (LGN)
    - Théorème de la limite centrale (TLC)
2. Statistiques mathématiques
    - Sondages et intervalles de confiance
    - Méthode de Monte-Carlo
    - Test statistiques et p-valeur
    - Tests d'adéquation et d'indépendance du χ^2
3. Statistiques prédictives: Régression linéaire.
4. Statistiques descriptives: généralités.

Partie 2: option pratique

1. Statistiques descriptives
    - Analyse et visualisation de données en python.
    - Analyse en composantes principales
    - Clustering avec les k-moyennes.
    - Implémentations
    - Optimisation: descente de gradient
    - Régression non-linéaire: arbres de décision, réseaux de neurones...
2. Statistiques prédictives
    - Implémentation autour de la régression linéaire.
    - Régression non-linéaire: généralités sur les méthodes a noyaux.
    - Régression non-linéaire en pratique: arbres de décision, réseaux de neurones.

Partie 2: option théorique

Fondements mathématiques des probabilités

- Opérations sur les ensembles
- Notions de denombrabilité
- Notions Convergences des V.A.
- Preuve de la Loi des Grands Nombres
- Fonction caractéristique et transformée de Fourier
- Preuve du théorème de la Limite Centrale

Variable aléatoire

Une variable aléatoire (v.a.) X est une quantité dont la valeur, a priori incertaine, est déterminée a l'issue du tirage.
Son espace d'états V est l'ensemble des valeurs qu'elle peux prendre.

- Espace d'états $\subset \mathbb{C}$: **variable** aléatoire.
- Espace d'états $\subset \mathbb{C}^d$: **vecteur** aléatoire.

ex:

- lancer de dé.
- nombre serveurs en panne.
- température dans 5 jours a Rennes.
- durée de vie d'un ordinateur.
- lancer de deux dés.

Variable aléatoire discrète

Cas discret.
Une variable aléatoire X est discrète si son espace d'états est fini ou infini dénombrable ($V = \{v_n, n \in \mathbb{N})$.
Propriété:
Une variable aléatoire discrète est caractérisée par les probabilités

$$P(X = v) \text{ (ou } P_X(v)), \; v \in V$$

L'ensemble $\{P_X(v), v \in V$ est appelé la loi de X, et les probabilités $P_X(v)$ sont positive est somment a un:

$$P(v) \leq 0, \; \sum_{v \in V} P(v) = 1$$

Rq: dans le cas infini, $P(v_n)$ est le terme d'une série absolument convergente. Le résultats sur les séries nous seront donc utiles.

Exercice

Somme de deux des équilibrés.
On considère la variable aléatoire S somme des valeurs de deux des équilibrés.

1. Quel est son espace d'états ?
2. Déterminer sa loi
3. Vérifier numériquement que les fréquences correspondent au probabilités obtenues

Exercice

$$V = \{2, \ldots, 12\}$$

On écrit le tableau qui donne les résultat en fonction des deux dés :

	1	2	3	4	5	6
1	2	3	4	5	6	7
2	3	4	5	6	7	8
3	4	5	6	7	8	9
4	5	6	7	8	9	10
5	6	7	8	9	10	11
6	7	8	9	10	11	12

$P(2) = P(12) = 1/36$, $P(3) = P(11) = 2/36 = 1/18$,
$P(4) = P(10) = 3/36 = 1/12$, $P(5) = P(9) = 4/36 = 1/9$,
$P(6) = P(8) = 5/36$, $P(7) = 6/36 = 1/6$.

Espérance d'une variable aléatoire discrète

L'espérance d'un variable aléatoire (v.a.) X discrète est donnée par

$$\mathbb{E}(X) = \sum_{v \in V} P(v)v$$

ex: Jeu de loto ou l'on coche 6 numéros sur une grille de 49. On joue 2 euros et on reçoit un gain en fonction du nombre de numéros gagnants:

numéros gagnants	gain	probabilité
6	2,132,885	$7.2 \ \ 10^{-8}$
5	3,575	$7.8 \ \ 10^{-5}$
4	94	$9.7 \ \ 10^{-4}$
3	11	$7.8 \ \ 10^{-2}$

1. Quelle est l'espérance du gain G?
2. A-t-on intérêt a jouer ?

Espérance d'une variable aléatoire discrète

1. Quelle est l'espérance de gain?

$$\mathbb{E}(G) = 2,132,885 \times 7.2\ 10^{-8} + 3,575 \times 7.8\ 10^{-5} + 94 \times 9.7\ 10^{-4} + 11 \times 7.8\ 10^{-2}$$

   Python :
   2132885*7.2*10**(-8) + 3575*7.8*10**(-5) + 94*9.7*10**(-4) + 11*7.8*10**(-2) = 1.38159772

2. A-t-on intérêt a jouer ?
   Bénéfice moyen :

$$\mathbb{E}(G) - 2 \approx 1.38 - 2 \approx -0.62$$

   D'un point de vue comptable, non. D'un quelconque autre point de vue...

Espérance d'une variable aléatoire discrète

Propriétés:

- linéarité
$$\mathbb{E}(aX + bY) = a\mathbb{E}(X) + b\mathbb{E}(Y)$$
- positivité
$$X \geq 0, \mathbb{E}(X) \geq 0$$
- X constante de valeur a, $\mathbb{E}(X) = a$
- inegalités:

$$\underbrace{|\mathbb{E}(X)| \leq \mathbb{E}(|X|)}_{\text{triangulaire}}, \quad f \text{ convexe, } \underbrace{f(\mathbb{E}(X)) \leq \mathbb{E}(f(X))}_{\text{Jensen}}.$$

Variance d'une variable aléatoire discrète

Propriété:

$$|\mathbb{E}(X)| \leq \mathbb{E}(|X|) \leq \sqrt{\mathbb{E}(X^2)}$$

Si $\mathbb{E}(X^2) < \infty$, on peut définir la variance:

$$Var(X) = \mathbb{E}((X - \mathbb{E}(X))^2) = \sum_{v \in V} (v - \mathbb{E}(X))^2 P(v)$$

- $Var(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2$
- $Var(X)$ est positive (car $(X - \mathbb{E}(X))^2 \geq 0$.)
- $Var(aX) = a^2 Var(X)$

Écart type:

$$\sigma_X = \sqrt{Var(X)}$$

Variables aléatoires discrètes usuelles

Bernoulli X:

$$\begin{aligned} V &= \{0, 1\} \\ P(X = 1) &= p, P(X = 0) = 1 - p \\ \mathbb{E}(X) &= p \\ Var(X) &= p(1 - p) \end{aligned}$$

Binomiale S:

$$\begin{aligned} V &= \{0, \dots, n\} \\ P(S = k) &= \binom{n}{k} p^k (1 - p)^{n-k} \\ \mathbb{E}(X) &= np \\ Var(X) &= np(1 - p) \end{aligned}$$

Variables aléatoires discrètes usuelles

Poisson X de paramètre $\theta > 0$:

$$
\begin{aligned}
V &= \mathbb{N} \\
P(X = k) &= e^{-\theta} \frac{\theta^k}{k!} \\
\mathbb{E}(X) &= \theta \\
Var(X) &= \theta
\end{aligned}
$$

Loi conditionnelle

Définition:
X, Y v.a. a valeurs dans V, W
Pour $v \in V$ tel que $P(X = v) > 0$, la loi conditionnelle de Y sachant $X = v$ est définie par la probabilité:

$$\forall w \in W, \quad P(Y = w | X = v) = \frac{P(Y = w, X = v)}{P(X = v)}$$

Remarques:

- La loi conditionnelle de Y sachant $X = v$ est une fonction de v.
- Les loi conditionnelle peuvent êtres très différentes en fonction de v.

Indépendance des variables aléatoires

Définition:

Les variables aléatoires X et Y sont indépendantes s.s.i.:

1. $\forall v, w, \quad P(X = v, Y = w) = P(X = v)P(Y = w)$
2. $\forall v$ t.q. $P(X = v) > 0, \; \forall w, P(Y = w | X = v) = P(Y = w)$

Remarque:

- La notion d'indépendance se comprend bien avec la deuxième équivalence: quelle que soit la valeur que prend la v.a. X, cela ne change pas la loi de la v.a. Y.

Indépendance des variables aléatoires

Propriétés:
Les variables aléatoires X et Y sont indépendantes.

- Espérance.
  $\forall f, g$ fonctions :

$$\mathbb{E}\left(f(X)g(Y)\right) = \mathbb{E}(f(X))\mathbb{E}(g(Y))$$

- Somme.
  On suppose X et Y a valeur dans $\mathbb{Z}$

$$P(X+Y=i) = \sum_{j \in \mathbb{Z}} P(X=i-j)P(Y=j)$$

La loi de $X+Y$ est obtenue par **convolution discrète** des probabilités de X et Y.

Somme de Bernoulli indépendantes

$X_i, i \in \mathbb{N}$ v.a. Bernoulli indépendantes. On considère la somme $S_n = \sum_{i=1}^n X_i$.
Déterminer la loi de S_n.

- Son espace d'états est $\{0, \ldots, n\}$.
- $k\{0, \ldots, n\} \in P(S_n = k)$.
  Il faut que exactement k des n variable X_i soient égales a 1 et les autre a 0, ce qui fait $\binom{n}{k}$ possibilités.
  Soit une telle configuration : $i_1, \ldots, i_k$, sont les indices des v.a. égales a 1 et $i_{k+1}, \ldots, i_n$ les indices des v.a. égales a 0.
  Par indépendance:

$$\begin{aligned}
&P\left(X_{i_1} = 1, \ldots, X_{i_k} = 1, X_{i_{k+1}} = 0, \ldots, X_{i_n=0}\right) \\
&= P(X_{i_1} = 1) \ldots P(X_{i_k} = 1) P(X_{i_{k+1}} = 0) \ldots P(X_{i_n} = 0) \\
&= p^k (1-p)^{n-k}
\end{aligned}$$

Somme de Bernoulli indépendantes

$$P(S_n = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

On reconnaît une loi binomiale.
Espérance:

$$\mathbb{E}(S_n) = \mathbb{E}\left(\sum_i X_i\right) = \sum_i \mathbb{E}\left(X_i\right) = np$$

Et la variance ?
La variance n'est pas linéaire : $Var(aX) = a^2 Var(X)$.

Variance d'une somme de v.a. **indépendantes**

X, Y v.a. indépendantes:

$$
\begin{aligned}
&Var(X+Y) = \mathbb{E}\left[(X-\mathbb{E}(X)+Y-\mathbb{E}(Y))^2\right] \\
&= \mathbb{E}\left[(X-\mathbb{E}(X))^2+(Y-\mathbb{E}(Y))^2+2(X-\mathbb{E}(X))(Y-\mathbb{E}(Y))\right] \\
&= \mathbb{E}\left[(X-\mathbb{E}(X))^2\right]+\mathbb{E}\left[(Y-\mathbb{E}(Y))^2\right]+2\mathbb{E}[X-\mathbb{E}(X)]\mathbb{E}[Y-\mathbb{E}(Y)] \\
&= Var(X)+Var(Y)
\end{aligned}
$$

Si X, Y v.a. **indépendantes**, la variance de la somme est la somme des variances.

$$Var(S_n) = Var\left(\sum_i X_i\right) = \sum_i Var\left(X_i\right) = np(1-p)$$

Variable aléatoire réelle et fonction de répartition

Définition:
Une variable aléatoire réelle est a valeur dans $\mathbb{R}$.

Remarque:
$\mathbb{R}$ n'est pas dénombrable, on ne peux pas simplement numéroter les valeurs de $\mathbb{R} = \{v_n, n \in \mathbb{N}$ et leur associer une probabilité $p_n > 0, \sum_n p_n = 1$.

Définition:
Soit X un v.a. réelle. La fonction de répartition de X est la fonction définie par

$$\forall x \in \mathbb{R}, F_X(x) = P(X \leq x)$$

Fonction de répartition

Propriétés:
X v.a. réelle.
La fonction de répartition caractérise la probabilité de X et vérifie les propriétés suivantes

- F_X est croissante.
- F_X est continue a droite
- $\lim_{x \to -\infty} F_X(x) = 0$ et $\lim_{x \to +\infty} F_X(x) = 1$

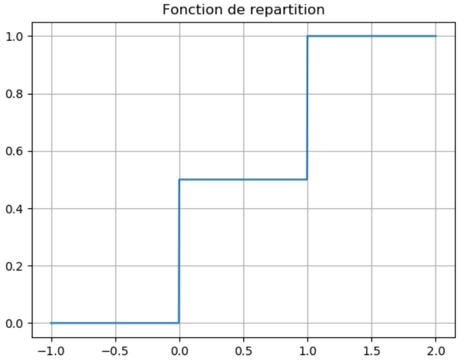
Remarque:
La réciproque est vraie.
Si F vérifie les trois propriétés, c'est la fonction de répartition d'une loi de probabilité sur $\mathbb{R}$.

Propriétés de la fonction de répartition

Nota: comme F est croissante, elle admet une limite a gauche $F_X(x-)$ en tout point x.
Propriétés:

- $P\left(X = x\right) = F_X(x) - F_X(x-)$
- $P(X = x) = 0 \iff F_X$ continue en x
- $P\left(x < X \leq y\right) = P(X \leq y) - P(X \leq x) = F_X(y) - F_X(x)$
- $P\left(x < X < y\right) = P(X < y) - P(X \leq x) = F_X(y-) - F_X(x)$
- $P\left(x \leq X \leq y\right) = P(X \leq y) - P(X < x) = F_X(y) - F_X(x-)$
- $P\left(x \leq X < y\right) = P(X < y) - P(X < x) = F_X(y-) - F_X(x-)$

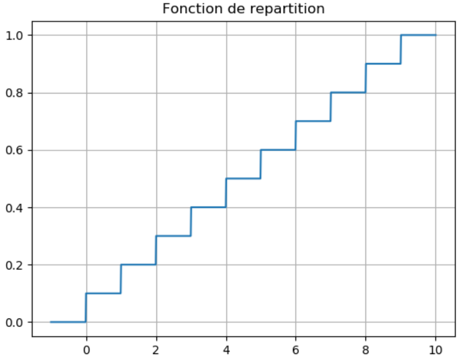
Exemple de fonction de répartition



Quelle est la loi ?

Bernoulli de paramètre $p = 0.5$

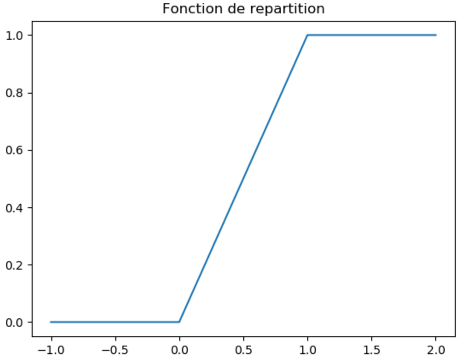
Exemple de fonction de répartition



Quelle est la loi ?

Valeurs $\{0, 1, \ldots, 9\}$, probabilité uniforme de $1/10$.

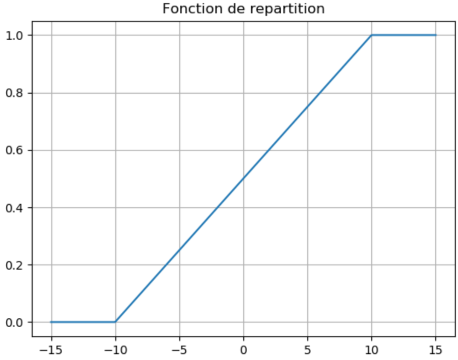
Exemple de fonction de répartition



Quelle est la loi ?

Probabilité uniforme sur $[0, 1]$.

Exemple de fonction de répartition



Quelle est la loi ?

Probabilité uniforme sur $[-10, 10]$.

Exemples de fonction de répartition

- X v.a. constante de valeur a.

$$F_X(x) = 1_{[a,\infty[}$$

- X a valeur dans $\mathbb{N}$, on note $p_n = P(X = n)$.

$$F_X(x) = \begin{cases} 0 & \text{si } x < 0 \\ p_0 + \ldots + p_n & \text{si } n \leq x < n+1 \end{cases}$$

- X a valeur dans $V = (v_n)_{n\in\mathbb{N}}$. On note $p_n = P(X = v_n)$.

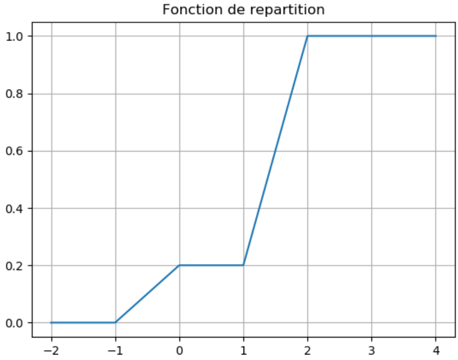
$$F_X(x) = \sum p_n 1_{[v_n,\infty[}$$

- X v.a. uniforme sur $[a, b]$, $(a < b)$.

$$F_X(x) = \frac{x-a}{b-a} 1_{[a,b[} + 1_{[b,\infty[}$$

F_X affine sur $[a, b]$.

Exemple de fonction de répartition



Quelle est la loi ?

$P(X \in [-1, 0]) = 0.2$, $P(X \in [1, 2]) = 0.8$
Probabilité uniforme $[-1, 0]$ et $[1, 2]$.

Lois a densité

Définition:

Une fonction f définie sur $\mathbb{R}$ est une densité de probabilité si elle est positive, intégrable et d'intégrale 1;

$$\int_{-\infty}^{+\infty} f(x)dx = 1$$

Propriété:

La fonction

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

vérifie les propriétés d'une fonction de répartition:

- F est croissante (sa dérivée est positive)
- F est continue a droite (elle l'est partout car dérivable)
- $\lim_{x\to-\infty} F(x) = 0$
  $\lim_{x\to+\infty} F(x) = \int_{\mathbb{R}} f = 1$

D'après la réciproque de la propriété sur les fonctions de répartition, F est une fonction de répartition qui caractérise une loi de probabilité.

Lois a densité

Propriétés:

Soi X une v.a. réelle de loi de densité f.

- Sa fonction de répartition F_X est continue et

$$\forall x, P(X = x) = 0$$

- Sa fonction de répartition F_X est dérivable et

$$\forall x, F_X'(x) = f(x)$$

Réciproquement, si la v.a. X est telle que sa fonction de répartition F_X est dérivable, X admet la densité

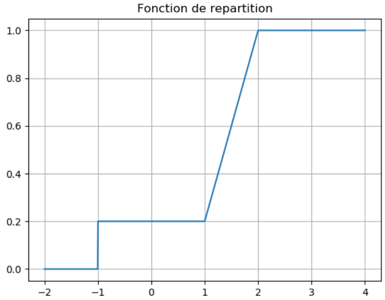
$$\forall x, f(x) = F_X'(x)$$

Remarques:

- Les v.a. discrètes n'admettent pas de densité (continue par rapport a la mesure de Lesbegue).
- La réciproque n'est pas vraie.

Exemple

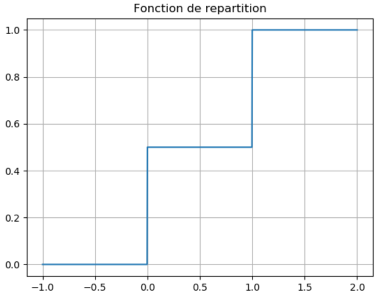
Fonction de répartition non derivable qui caractérise la loi d'une v.a. non discrète:

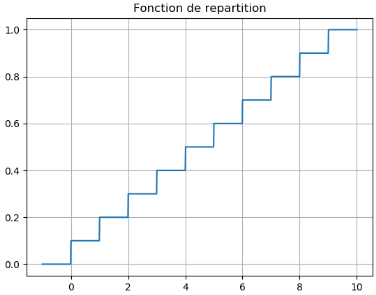


- $P(X = -1) = 0.2$, $P(X \in [1, 2]) = 0.8$.
- N'admet pas de densité car discontinue en -1.
- N'est pas discrète car $[1, 2]$ est inclus dans son espace d'états.

Exemple de fonction de répartition

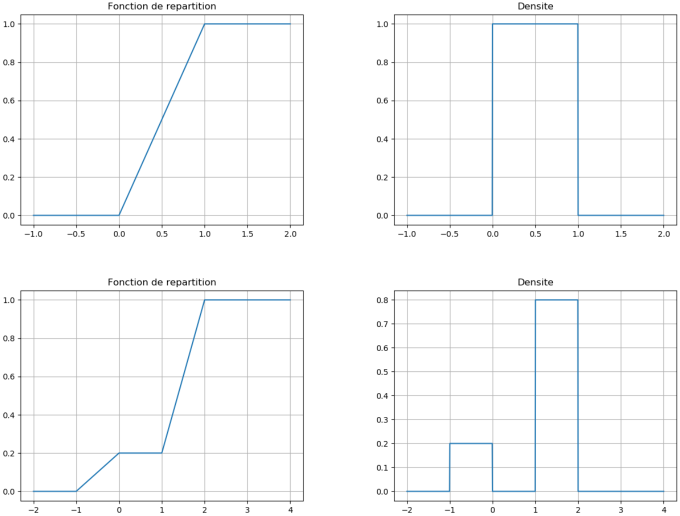
N'admettent pas de densité les fonctions de répartitions suivantes:





Exemple de fonction de répartition

Les fonctions de répartitions suivantes ont pour densité:



Générateurs de nombres aléatoires.

Générer des nombre (vraiment) aléatoires:

- Utiliser un processus physique : fluctuations thermiques, bruit photonique, rayonnements parasites atmosphériques
- Utiliser un processus "social" : temps de ping sur wikipedia.com, mesure du trafic automobile...

Pose un certain nombre de problèmes:

- Processus physique: dépend d'un capteur et des paramètres.
- Processus "social" : fluctuations temporelles

Vraiment aléatoire, mais peu fiable.

Générateurs de nombres pseudo-aléatoires.

Générer des variables uniformes sur $[0, 1]$

- Suite récurrente a valeur dans $\{0, 1, \dots m - 1\}$

$$x_{n+1} = \phi(x_n)$$

- Suite périodique de période p, $p \leq m$
- On considère la suite u a valeur dans $[0, 1[$

$$u_n = \frac{x_n}{m}$$

- Les tirages successif doivent avoir les caractéristiques de tirages indépendants

ex: générateurs congruentiels

$$x_0 \in \mathbb{N}, \; x_{n+1} = ax_n + b \text{ modulo } m$$

Scilab: $m = 2^{31}$, $a = 843314861$, $b = 453816693$.
Python: Mersenne Twister, voir les docs python.

Inverse de la fonction de répartition.

A partir d'une v.a. uniformes sur $[0,1]$, on peut simuler n'importe quelle v.a. dont on connaît la fonction de répartition.

Définition:
G inverse continue a gauche de F est définie par

$$G(u) = \inf \{x : F(x) \geq u\}$$

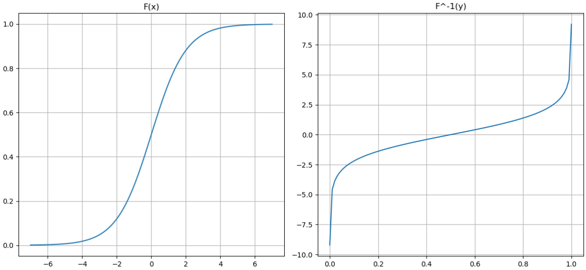
Propriétés:

1. $G(u) \leq x \iff u \leq F(x)$
2. $F(G(u)) = u$ si F continue en u.

Preuve: 1 découle de la définition de inf, 2 au tableau.

Inverse de la fonction de répartition.

- Si F bijective (continue, strict. croissante), c'est l'inverse ($=$ la réciproque) $F^{-1} :]0,1[\mapsto \mathbb{R}$ de $F : \mathbb{R} \mapsto]0, 1[$.
  ex: $F(x) = 1/(1+e^{-x})$, $F^{-1}(y) = -\log(1/y - 1)$



- F constante sur $[a, b]$: au tableau.
- F discontinue en a: au tableau.

Méthode de l'inversion de la fonction de répartition.

Propriété:
Soit F une fonction de répartition d'inverse continue a gauche G.
U v.a. uniforme sur $[0,1]$
$X = G(U)$ est une v.a. de fonction de répartition F.

Preuve:
Quelle est la fonction de répartition de X ?
$\forall x \in \mathbb{R}$:

$$\begin{aligned}
P(X \leq x) &= P(G(U) \leq x) && \text{car } X = G(U) \\
&= P(U \leq F(x)) && \text{car } G(U) \leq x \iff U \leq F(x) \\
&= F_U(F(x)) && \text{par déf. de } F_U \\
&= F(x) && \text{car } F_U(y) = y, \forall y \in [0,1].
\end{aligned}$$

Méthode de l'inversion de la fonction de répartition.

Exercice:

En utilisant le générateur de nombre pseudo-aléatoire code précédemment, simulez une v.a. de Fonction de répartition

$$F(x) = 1/(1 + e^{-x})$$

et vérifiez que l'histogramme des valeurs tirées converge vers la densité $F'(x)$.

Espérance d'une variable aléatoire réelle

La notion d'espérance d'une variable aléatoire réelle se définit par passage a la limite des fonctions en escaliers (v.a. discrètes).
La différence est qu'on exige que $\mathbb{E}(|X|) < \infty$, c.a.d. $X \in L^1$.
Propriétés:

- linéarité
$$\mathbb{E}(aX + bY) = a\mathbb{E}(X) + b\mathbb{E}(Y)$$
- positivité
$$X \geq 0, \mathbb{E}(X) \geq 0$$
- v.a. constante
$$X \text{ constante de valeur } a, \ \mathbb{E}(X) = a$$
- inégalité
$$|\mathbb{E}(X)| \leq \mathbb{E}(|X|)$$

(Co)variance d'une variable aléatoire réelle

Si $\mathbb{E}(X^2) < \infty$, on peut définir la variance:

$$Var(X) = \mathbb{E}((X - \mathbb{E}(X))^2) = \mathbb{E}((X^2) - \mathbb{E}(X)^2)$$

Ecart type:

$$\sigma_X = \sqrt{Var(X)}$$

Covariance:
Soit X, Y deux v.a. réelles

$$Cov(X, Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)$$

Coefficient de corrélation:

$$Corr(X, Y) = \frac{Cov(X, Y)}{\sigma_X \sigma_Y}$$

(Co)variance d'une variable aléatoire réelle

Propriétés:

- $Var(aX + b) = a^2 Var(X)$
- $Var(X + Y) = Var(X) + Var(Y) + 2Cov(X, Y)$.
- X, Y indépendantes: $Cov(X, Y) = 0$ donc $Var(X + Y) = Var(X) + Var(Y)$
- $Cov(aX + b, cY + d) = acCov(X, Y)$

Remarque:
$\forall X$ t.q. $\mathbb{E}(X^2) < \infty$, la v.a.

$$\frac{X - \mathbb{E}(X)}{\sigma_X}$$

est d'espérance nulle (i.e. centrée) et de variance 1 (réduite).

Espérance de v.a. réelles a densité

X v.a. réelle qui admet une densité f.
Si g fonction t.q. $g(X) \in L^1$:

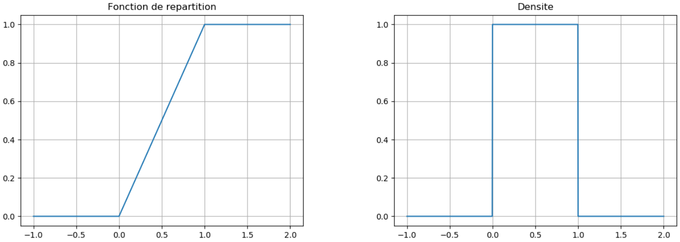
$$\mathbb{E}(g(X)) = \int_{\mathbb{R}} g(x)f(x)dx$$

En particulier:

$$\mathbb{E}(X) = \int_{\mathbb{R}} xf(x)dx$$
$$Var(X) = \int_{\mathbb{R}} x^2 f(x)dx - \mathbb{E}(X)^2$$

Remarque:
C'est la limite continue du cas discret: $\sum \to \int$,
$p_n = P(X = n) \to f(x)dx$.

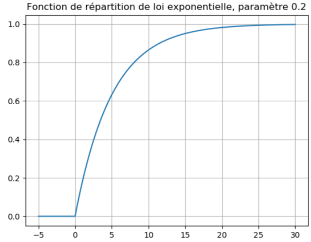
Loi uniforme sur $[a, b]$

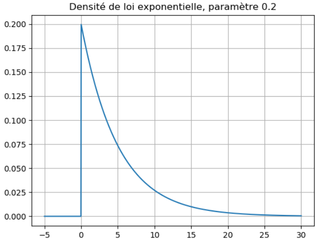


Exemple avec $a = 0$ et $b = 1$.

- Densité $f(x) = \frac{1}{b-a} 1_{[a,b]}$
- Espérance $\int_a^b \frac{x}{b-a} dx = \frac{1}{2(b-a)} \left(b^2 - a^2\right) = \frac{(b-a)(b+a)}{2(b-a)} = \frac{a+b}{2}$
- Variance $\int_a^b \frac{x}{b-a} dx - \left(\frac{a+b}{2}\right)^2 = \frac{(b-a)^2}{12}$

Loi exponentielle de paramètre λ

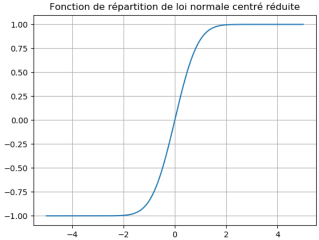


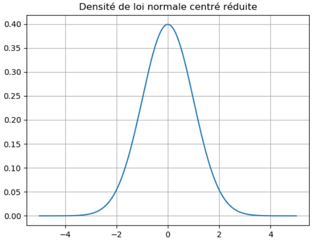


- Densité $f(x) = \lambda e^{-\lambda x} 1_{[0,\infty[}$
- Fonction de répartition $F(x) = (1 - e^{-\lambda x}) 1_{[0,\infty[}$
- Espérance: $\int_0^\infty x \lambda e^{-\lambda x} dx = 1/\lambda$
- Variance $1/\lambda^2$

Notebook: simuler une v.a. exponentielle par méthode d'inversion.

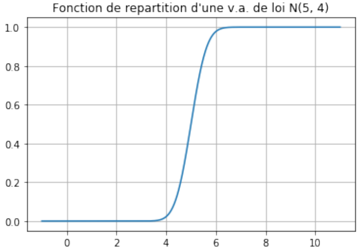
Loi Normale (Gaussienne) centrée réduite

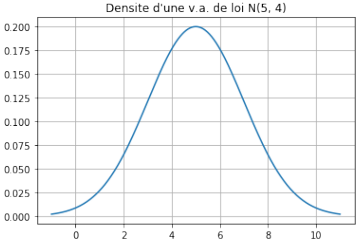




- Densité $f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$
- Fonction de répartition non analytique $F(x) = \int_{-\infty}^{x} f(x)dx$.
- Espérance 0, Variance 1.

Loi Normale $\mathcal{N}(m, \sigma^2)$





- Densité $f(x) = \frac{1}{\sqrt{2\pi\sigma}} \exp\left(-\frac{(x-m)^2}{2\sigma^2}\right)$
- Fonction de répartition non analytique $F(x) = \int_{-\infty}^{x} f(x)dx$.
- Espérance m
- Variance σ^2

Variables aléatoire usuelles en python

Bibliothèque scipy.stats donne accès aux caracteristiques de nombreuses lois: normale, exponentielle, χ_2, ...
Exemple: scipy.stats.norm

- densité: scipy . stats . norm.pdf.
- Fonction de répartition: scipy . stats . norm.cdf
- Inverse de la fonction de répartition: scipy . stats . norm.ppf
- Moyenne, Variance, Moments...

Pour échantillonner, on utilisera la bibliothèque numpy.random.

Convergence presque sure des variables aléatoires

Définition de la convergence presque sure:
On dit que la suite X_n converge presque sûrement vers la variable aléatoire X, et on note $X_n \underset{\text{p.s.}}{\rightarrow} X$, si

$$P\left(\lim_{n\rightarrow\infty} X_n = X\right) = 1$$

Remarques:

- Dans le cas des suites de nombre déterministe, une suite a qui converge ne peut avoir deux sous-suites qui convergent vers deux valeurs différentes
- Dans le cas probabiliste, il peut exister des évènements tels que $X_n \nrightarrow X$, mais ils doivent être de probabilité nulle.

Convergence presque sure des variables aléatoires

X_i Bernoulli indépendantes de paramètre $p \in]0,1[$.

$$\text{La suite } M_n = \frac{\sum_{i=1}^n X_i}{n} \text{ converge-t-elle ?}$$

- Si $\forall i, X_i = 1$, alors $M_n = 1 \ \forall n$.
- Si $\forall i, X_i = 0$, alors $M_n = 0 \ \forall n$.

Donc M_n ne converge pas. FAUX!

Ces deux évènements n'empêchent pas la convergence presque sure de M_n parce que leur probabilité est nulle:

- $P(M_n = 1) = p^n \underset{n\to\infty}{\longrightarrow} 0$
- $P(M_n = 0) = (1-p)^n \underset{n\to\infty}{\longrightarrow} 0$.

La convergence p.s. est donc plus "souple" que la notion déterministe de convergence simple.

La loi forte des grands nombres

Théorème:
Soient X_n variable aléatoires indépendantes de même loi et intégrables. On note $m = \mathbb{E}(X_i)$.
La moyenne empirique M_n des v.a. X_i converge presque sûrement vers la *moyenne théorique* $m = \mathbb{E}(X_i)$ des v.a. X_i quand n tend vers l'infini.

$$M_n = \frac{X_1 + \ldots + X_n}{n} \underset{\text{p.s.}}{\rightarrow} m \text{ qd } n \rightarrow \infty$$

Remarques:

- exemple: X_i Bernoulli indépendantes de paramètre p:

$$M_n = \frac{\sum_{i=1}^{n} X_i}{n} \underset{\text{p.s.}}{\rightarrow} \mathbb{E}(X_i) = p$$

- Vitesse de convergence ? Théorème de la limite centrale.

La loi forte des grands nombres

Corrolaire:
Soient X_n variable aléatoires indépendantes de même loi et et f fonction t.q. $f(X_i)$ intégrables.

$$\frac{f(X_1) + \ldots + f(X_n)}{n} \underset{\text{p.s.}}{\rightarrow} \mathbb{E}(f(X_i))$$

Méthode de Monte-Carlo pour calcul d'intégrales

Soient X_n variables aléatoires indépendantes uniformes sur $[0,1]$. f fonction intégrable sur $[0,1]$.

$$\lim_n \frac{f(X_1)+\ldots+f(X_n)}{n} = \int_0^1 f(x)dx$$

Preuve:
Une v.a. X. de loi uniforme sur $[0,1]$ a pour densité $1_{[0,1]}$ donc

$$E(f(X)) = \int_{\mathbb{R}} 1_{[0,1]}(x)f(x)dx = \int_0^1 f(x)dx$$

Méthode de Monte-Carlo pour calcul d'intégrales

Generalisation en dimension d.
Soient X_n vecteurs aléatoires indépendants de loi uniforme sur l'hyper-cube $A = [-a, a]^d$. f fonction intégrable sur A.

$$\lim_{n \to \infty} \frac{(2a)^d}{n} \left(f(X_1) + \ldots + f(X_n)\right) = \int_A f$$

Preuve:
Un vecteur aléatoire X de loi uniforme sur A a pour densité $1_A/(2a)^d$ donc

$$E(f(X)) = \int_{\mathbb{R}} 1_A(x) f(x) dx = \int_A f(x) dx$$

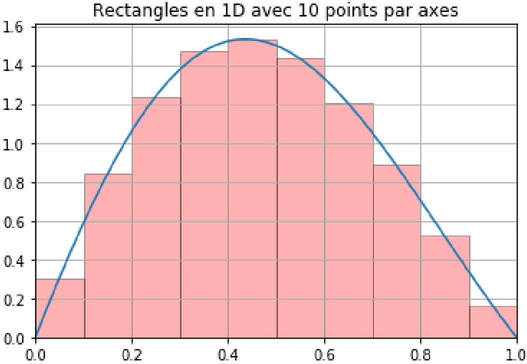
Remarques:

- f peut être très irrégulière, simplement bornée.
- temps de calcul proportionnel a nd

Monte-Carlo vs méthode des rectangles pour les intégrales

Méthode des rectangles en dimension 1: on découpe l'intervalle $[0,1]$ en N segments de centre x_i de longueur Δx et

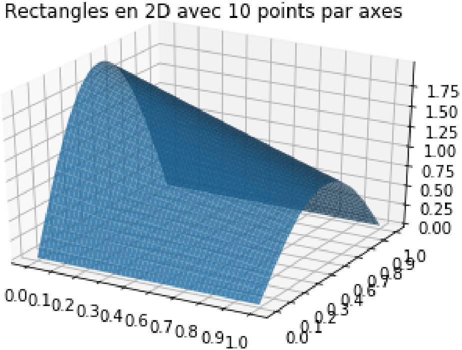
$$\int_{[0,1]} f(x)dx \approx \sum_{i=1}^{N} f(x_i)\Delta x$$



Monte-Carlo vs méthode des rectangles pour les intégrales

Méthode des rectangles en dimension 2: on découpe la surface $[0,1]^2$ en N^2 de centre $x_{i,j}$ de surface Δx^2 et

$$\int_{[0,1]^2} f(x)dx \approx \sum_{i,j=1}^{N} f(x_{i,j})\Delta x^2$$



10^2 valeurs a calculer en 2D.

Monte-Carlo vs méthode des rectangles pour les intégrales

Méthode des rectangles en dimension d: on découpe le domaine $[0,1]^d$ en N^d hyper-cubes de centre $x_{i,j,\ldots}$ de volumes Δx^d.

$$\int_{[0,1]^d} f(x)dx \approx \sum_{i,j,\ldots=1}^{N} f(x_{i,j,\ldots})\Delta x^d$$

Rappel sur les intégrales en dimension > 1.
Si $f : \mathbb{R}^5 \mapsto \mathbb{R}$ est une fonction séparable en les coordonnes u, v, w, x, y:

$$f(u,v,w,x,y) = f_1(u)f_2(v)f_3(w)f_4(x)f_5(y)$$

l'intégrale de f sur l'hyper-cube $[0,1]^5$ est le produit des intégrales des f_i sur les segments $[0,1]$:

$$\int_{[0,1]^5} f(u,v,w,x,y)dudvdwdxdy =$$
$$\int_{[0,1]} f_1(u)du \int_{[0,1]} f_2(v)dv \int_{[0,1]} f_3(w)dw \int_{[0,1]} f_4(x)dx \int_{[0,1]} f_5(y)dy$$

Monte-Carlo vs méthode des rectangles pour les intégrales

Exercice:
$f(u, v, w, x, y) = e^u \; 2v \sin(w) \cos(x) \frac{1}{1+y}$
Calculer $\int_{[0,1]^5} f$

- f est une fonction séparable en les coordonnes u, v, w, x, y, l'intégrale est le produit des intégrales.
- Primitives: $u \mapsto e^u$, $v \mapsto v^2$, $w \mapsto -\cos(w)$, $x \mapsto \sin(x)$ et $y \mapsto \log(1+y)$
- Valeur: 0.4607...

Notebook: comparer, a temps d'exécution comparable, la précision des méthodes de Monte-Carlo et des rectangles pour $\int_{[0,1]^5} f$.

Convergence en loi des variables aléatoires

On dit que la suite X_n converge en loi vers la variable aléatoire X, et on note $X_n \stackrel{\mathcal{L}}{\to} X$, si pour toute fonction continue bornée f

$$\lim_{n\to\infty} \mathbb{E}(f(X_n)) = \mathbb{E}(f(X))$$

exemple: X_i v.a. a valeurs dans $\{1, \ldots, N\}$. $\forall i \in \{1, \ldots, N\}$:

$$\lim_{n\to\infty} P(X_n = i) = P(X = i)$$

On a

$$\lim_{n\to\infty} \mathbb{E}(f(X_n)) = \sum_{i=1}^{N} f(i) \left(\lim_{n\to\infty} P(X_n = i)\right) = \mathbb{E}(f(X))$$

Convergence en loi des variables aléatoires

Exemples:
Soient X_n suite de variable $\mathcal{N}(0, \sigma_n)$ t.q. $\lim_{n\to\infty} \sigma_n = 1$. X_n converge en loi vers X variable aléatoire de loi $\mathcal{N}(0, 1)$.

Pour toute fonction continue bornée f

$$\begin{aligned}\mathbb{E}(f(X_n)) &= \int \frac{f(x)e^{-x^2/2(1+1/n)^2}}{\sqrt{2\pi}(1+1/n)}dx \\ &\underset{n\to\infty}{\to} \int \frac{f(x)e^{-x^2/2}}{\sqrt{2\pi}}dx\end{aligned}$$

par théorème de convergence dominée.

Plus généralement, si X v.a. de densité d, et X_n suite de v.a. de densité d_n t.q. $\lim_{n\to\infty} d_n = d$ et que $\forall n, d_n < h \in L^1(\mathbb{R})$, X_n converge en loi vers X.

Convergence en loi des variables aléatoires

Propriété:

X_n suite de v.a. qui converge en loi vers X. On note F_n les fonctions de répartitions des v.a. X_n. Si la fonction de répartition F de la v.a. X est continue en x, on a :

$$\lim_{n \to \infty} F_n(x) = F(x)$$

Conséquence:

X_n suite de v.a. qui converge en loi vers v.a. X a densité f , $a < b$ réels.

$$\lim_{n \to \infty} P(a \leq X_n \leq b) = \int_a^b f(x)dx$$

Avec cette propriété, la notion de convergence en loi est d'autant plus claire. La loi des X_n, c.a.d. les probabilités que X_n soient dans [a,b] converge vers la loi de X donnée par la densité f.

Théorème de la limite centrale

Théorème:
Soit X_n une suite de variable aléatoires indépendantes et de même loi, de moyenne m et de variance $\sigma^2 > 0$. On note $M_n = \frac{X_1 + \ldots + X_n}{n}$ la moyenne empirique . La suite de variables aléatoires

$$\frac{\sqrt{n}(M_n - m)}{\sigma}$$

converge en loi vers une variable aléatoire de loi normale $\mathcal{N}(0,1)$ quand n tend vers $+\infty$.

$$\frac{\sqrt{n}\,(M_n - m)}{\sigma} \underset{n\to\infty}{\overset{\mathcal{L}}{\to}} \mathcal{N}(0,1)$$

$$\iff \frac{(S_n - nm)}{\sqrt{n}\sigma} \underset{n\to\infty}{\overset{\mathcal{L}}{\to}} \mathcal{N}(0,1)$$

Remarques sur le Théorème de la limite centrale

D'après la propriété sur la convergence en loi, le Théorème de la limite centrale (TLC) donne :

$$\lim_{n\to\infty} P\left(\frac{\sqrt{n}\,(M_n - m)}{\sigma} \in [a, b]\right) = \int_a^b \frac{e^{-x^2/2}}{\sqrt{2\pi}}$$

Remarques:

- Le TLC donne une vitesse de convergence pour la loi des grands nombres
- Moyen mémo-technique: $\sigma > 0 \implies$ sigma au dénominateur. $M_n - m \to 0$. Retenir: loi $\mathcal{N}(0, 1)$, et convergence en $1/\sqrt{n}$

$$\frac{M_n - m}{\sigma} \overset{n\to\infty}{\approx} \frac{\mathcal{N}(0,1)}{\sqrt{n}}$$

- C'est vrai en particulier pour les lois discrètes alors que la limite est une v.a. continue.
- Pas de vitesse de convergence pour la convergence en loi.

Pratique du Théorème de la limite centrale

1. Notebook: vérifier empiriquement le TLC pour un suite de v.a. de Bernoulli.

2. B_n suit une loi binomiale $\mathcal{B}(n, 1/2)$. Calculer $P(B_{2000} > 1111)$.

Méthode 1:

$$P(B_{2000} > 1111) = \sum_{k=1112}^{2000} \binom{2000}{k} \left(\frac{1}{2}\right)^{2000}$$

Méthode 2:

B_n suit la même loi qu'une somme S_n de n variables de Bernoulli indépendantes de paramètre $p = 1/2$, d'espérance $p = 1/2$ et d'écart-type $\sqrt{p(1-p)} = 1/2$.TLC (version avec S_n):

$$\frac{(S_n - n/2)}{\sqrt{n}/2} \xrightarrow{\mathcal{L}} \mathcal{N}(0, 1)$$

Comme $n = 2000 \gg 1$, on s.q. $Y = \frac{S_{2000} - 2000/2}{\sqrt{2000}/2}$ suit exactement une loi $\mathcal{N}(0, 1)$.

Pratique Théorème de la limite centrale

Faisons apparaître la v.a. $Y = \frac{S_{2000} - 2000/2}{\sqrt{2000}/2}$ dans notre inégalité $S_{2000} > 1111$:

$$\begin{aligned}
P(S_{2000} > 1111) &= P(S_{2000} - 1000 > 111) \\
&= P(\frac{S_{2000} - 1000}{\sqrt{2000}/2} > \frac{111}{\sqrt{2000}/2}) \\
&= P\left(Y > \frac{111}{\sqrt{500}}\right) \\
&= 1 - P\left(Y \leq \frac{111}{\sqrt{500}}\right) \\
&\overset{TLC}{\approx} 1 - F_{\mathcal{N}(0,1)}\left(\frac{111}{\sqrt{500}}\right)
\end{aligned}$$

Résultat: $P(B_{2000} > 1111) =^{TLC}_{\approx} 3.2\ 10^{-7}$.

Probabilités exactes d'une somme de Bernoulli

Exercice: calcul exact de $P(B_{2000} > 1111)$ (sans LGN, sans TLC).
Problème: overflow dans le calcul

$$\binom{2000}{k}\left(\frac{1}{2}\right)^{2000}$$

Méthode 1:réécrire ce terme comme un produit de nombres proches de 1

$$\begin{aligned}\binom{2000}{k}\left(\frac{1}{2}\right)^{2000} &= \frac{2000!}{k!(2000-k)!2^{2000}}\\ &= \frac{1/2 \times 2/2 \times \ldots \times 2000/2}{1 \times \ldots \times k \times 1 \times \times \ldots n-k}\\ &= \frac{1/2 \times 2/2 \times \ldots \times 2000/2}{1^2 \times 2^2 \times \ldots \times m^2 \times (m+1) \times \ldots \times M}\end{aligned}$$

avec $m = \min(k, n-k)$ et $M = \max(k, n-k)$.

Probabilités exactes d'une somme de Bernoulli

Exercice: calcul exact de $P(B_{2000} > 1111)$ (sans LGN, sans TLC).
Problème: overflow dans le calcul

$$\binom{2000}{k}\left(\frac{1}{2}\right)^{2000}$$

Méthode 2:produit de convolution. Si X et Y deux v.a. indépendantes a valeur dans $\mathbb{Z}$ la loi de $X+Y$ s'obtient pas convolution discrète des lois de X et Y

$$P(X+Y=k)=\sum_i P(Y=i)P(X=k-i)$$

Se généralise pour une somme de n v.a. indépendantes.
Les v.a. X_i de Bernoulli sont a valeur dans $\mathbb{Z}$ avec $P(X_i \neq 0,1) = 0$.On utilise numpy.convolve avec mode='full'.
Résultat: $2.997\ 10^{-7}$.

Intervalles de confiance

Principe: référendum, deux possibilités 0 ou 1. On veut déterminer statistiquement la proportion p des votes pour 1.On choisit n individus au hasard (sans remise). On note $X_i = 1$ si le i-ème individu vote 1, $X_i = 0$ s'il vote 0.

Comment estimer de p a partir des observation ?

- On s.q. les v.a. X_i suivent une loi de Bernoulli de paramètre p.
- On les suppose indépendantes. Ok ? pas exactement à cause du tirage sans remise.
- on définit: $\hat{P}_n = \frac{\sum_{i=1} nX_i}{n}$.
- $\forall n, \mathbb{E}(\hat{P}_n) = p$ par linéarité de l'espérance ($\hat{P}_n$ est **non biaisé**).
- $\lim_{n\to\infty} \hat{P}_n = p$, par la LGN, i.e. $\hat{P}_n$ est asymptotiquement bon.

Mais $\hat{P}_n$ est une variable aléatoire. Peut-on avoir un résultat qui tienne compte des aléas ?

Intervalles de confiance

On suppose que le sondage porte sur 2000 individus, on a 1042 intentions de vote pour 1 et 958 pour 0, $\hat{P}_{2000} = 52.1\%$, cette valeur prédit que 1 remporte. Mais $\hat{P}_{2000}$ est une variable aléatoire, et un nouveau tirage me donnera une valeur différente, potentiellement $< 50\%$. Quelle niveau de confiance peut on avoir en cette prédiction ?

Intervalle de confiance

On se donne une probabilité α proche de 1, qui correspond a la probabilité de ne pas se tromper, typiquement $\alpha = 95\%$. on cherche $\epsilon > 0$ tel que

$$P(p \in \left[\hat{P}_n - \epsilon, \hat{P}_n + \epsilon\right]) = \alpha$$

$$P(p \in \left[\hat{P}_{2000} - \epsilon, \hat{P}_{2000} + \epsilon\right]) = 95\% \text{ dans notre cas.}$$

Calcul de ϵ ? $\Rightarrow$ Théorème de la limite centrale.

Intervalles de confiance

D'une part

$$P\left(p \in \left[\hat{P}_n - \epsilon, \hat{P}_n + \epsilon\right]\right) = P(p \leq \hat{P}_n + \epsilon) - P(p \leq \hat{P}_n - \epsilon)$$

Calculons $P(p \leq \hat{P}_n + \epsilon)$:

$$\begin{aligned}
P(p \leq \hat{P}_n + \epsilon) &= P(-\epsilon \leq \hat{P}_n - p) \\
&= P\left(-\frac{\epsilon\sqrt{n}}{\sqrt{p(1-p)}} \leq \frac{\sqrt{n}(\hat{P}_n - p)}{\sqrt{p(1-p)}}\right) \\
&= P(-a \leq Y_n), \;\; a = \frac{\epsilon\sqrt{n}}{\sqrt{p(1-p)}}, \; Y_n = \frac{\sqrt{n}(\hat{P}_n - p)}{\sqrt{p(1-p)}} \\
&= 1 - P(Y_n < -a) \\
&\underset{n \gg 1}{\overset{TLC}{\approx}} 1 - F_{\mathcal{N}(0,1)}(-a)
\end{aligned}$$

Intervalles de confiance

Calculons $P(p \leq \hat{P}_n - \epsilon)$:

$$\begin{aligned}
P(p \leq \hat{P}_n - \epsilon) &= P(\epsilon \leq \hat{P}_n - p) \\
&= P\left(\frac{\epsilon\sqrt{n}}{\sqrt{p(1-p)}} \leq \frac{\sqrt{n}(\hat{P}_n - p)}{\sqrt{p(1-p)}}\right) \\
&= P(a \leq Y_n) \\
&= 1 - P(Y_n < a) \\
&\underset{n \gg 1}{\overset{TLC}{\approx}} 1 - F_{\mathcal{N}(0,1)}(a)
\end{aligned}$$

Donc

$$P\left(p \in \left[\hat{P}_n - \epsilon, \hat{P}_n + \epsilon\right]\right) = F_{\mathcal{N}(0,1)}(a) - F_{\mathcal{N}(0,1)}(-a) = 2F_{\mathcal{N}(0,1)}(a) - 1$$

Intervalles de confiance

$$2F_{\mathcal{N}(0,1)}\left(\epsilon\sqrt{\frac{n}{p(1-p)}}\right)-1=\alpha \iff \epsilon=\sqrt{\frac{p(1-p)}{n}}F^{-1}_{\mathcal{N}(0,1)}\left(\frac{\alpha+1}{2}\right)$$

Problème: $p(1-p)$?

$$p(1-p)\leq\frac{1}{4}\ ,\forall p\in[0,1]$$

Donc ϵ borné par

$$\epsilon_{\max}=\frac{1}{2\sqrt{n}}F^{-1}_{\mathcal{N}(0,1)}\left(\frac{\alpha+1}{2}\right)$$

Python: $F^{-1}_{\mathcal{N}(0,1)}$ est donnee par scipy.stats.norm.ppf.

Intervalles de confiance

Propriété:

X_i variables de Bernoulli indépendantes de paramètre p, $\hat{P}_n = \sum_{i=1}^{n} X_i/n$.
Soit $\alpha \in]0,1[$ le **niveau de confiance asymptotique de l'intervalle**.
L'intervalle

$$I_\alpha^n = \left[\hat{P}_n - \frac{F^{-1}_{\mathcal{N}(0,1)}\left(\frac{\alpha+1}{2}\right)}{2\sqrt{n}}, \hat{P}_n + \frac{F^{-1}_{\mathcal{N}(0,1)}\left(\frac{\alpha+1}{2}\right)}{2\sqrt{n}} \right]$$

est tel que

$$\lim_{n\to\infty} P\left(p \in I_\alpha^n\right) = \alpha$$

Exemples:

- $95\% \Rightarrow \left[\hat{P}_n - \frac{0.98}{\sqrt{n}}, \hat{P}_n + \frac{0.98}{\sqrt{n}}\right], 99\% \Rightarrow \left[\hat{P}_n - \frac{1.29}{\sqrt{n}}, \hat{P}_n + \frac{1.29}{\sqrt{n}}\right]$
- Ces intervalles sont des variables aléatoires.
- $n = 2000$, $\hat{p}_{2000} = 0.521$, $I_{99\%}^{2000} = [49.2\%, 55.0\%]$

Méthode de Monte-Carlo générale

On a vu la méthode de Monte-Carlo pour une intégrale sur un hyper-cube $[-a, a]^d$. Si $f : \mathbb{R}^d \mapsto \mathbb{R}$ est a support compact $S \subset [-a, a]^d$, on définit $g = 1_S f$, et on applique la méthode a g.

S'il existe $p(x) > 0$, et si on définit $\phi(x) = f(x)/p(x)$, X v.a. de densité p

$$\int_{\mathbb{R}^d} f(x)dx = \int_{\mathbb{R}^d} \phi(x)p(x)dx = \mathbb{E}\left(\phi(X)\right)$$

X_i v.a. indépendantes de densité p, d'apres la Loi des grands nombres :

$$M_n = \frac{\phi(X_1) + \ldots + \phi(X_n)}{n} \underset{n\to\infty}{\overset{p.s.}{\to}} \mathbb{E}(\phi(X)) = \int_{\mathbb{R}^d} f(x)dx$$

Vitesse de convergence ?

Vitesse de convergence de la méthode de Monte-Carlo

Exercice:
On note $e = \mathbb{E}(\phi(X))$. Comme pour l'intervalle de confiance du sondage, on se donne α proche de 1, on cherche ϵ t.q.:

$$P(e \in [M_n - \epsilon, M_n + \epsilon]) = \alpha$$

Cette fois ci on ne pourra pas de débarrasser du σ. L'intervalle de confiance fera donc intervenir

$$\sigma_n = \sqrt{\frac{\phi(X_1)^2 + \ldots + \phi(X_n)^2}{n} - (M_n)^2}$$

TLC amélioré

Lemme de Slutsky

X_n et Y_n deux suites de v.a. réelles telles que

$$X_n \stackrel{\mathcal{L}}{\rightarrow} X, Y_n \stackrel{p.s.}{\rightarrow} C \text{ constante.}$$

Pour toute fonction continue $g : \mathbb{R}^2 \rightarrow \mathbb{R}$

$$g(X_n, Y_n) \stackrel{\mathcal{L}}{\rightarrow} g(X, C)$$

Application:

$$\frac{\sqrt{n}(M_n - m)}{\sigma} \stackrel{\mathcal{L}}{\rightarrow} \mathcal{N}(0,1), \qquad \frac{\sigma}{\sigma_n} \stackrel{p.s.}{\rightarrow} 1$$

Donc

$$\frac{\sqrt{n}(M_n - m)}{\sigma_n} \stackrel{\mathcal{L}}{\rightarrow} \mathcal{N}(0,1)$$

Vitesse de convergence de la méthode de Monte-Carlo

$$P(e \in [M_n - \epsilon, M_n + \epsilon]) = P(e \leq M_n + \epsilon) - P(e \leq M_n - \epsilon)$$

$$\begin{aligned}
P(e \leq M_n + \epsilon) &= P(-\epsilon \leq M_n - e) \\
&= P\left(-\frac{\epsilon\sqrt{n}}{\sigma_n} \leq \frac{\sqrt{n}(M_n - e)}{\sigma_n}\right) \\
&= P(-a \leq Y_n),\ \ a = \frac{\epsilon\sqrt{n}}{\sigma_n}, Y_n = \frac{\sqrt{n}(M_n - e)}{\sigma_n} \\
&= 1 - P(Y_n < -a) \\
&\underset{n \gg 1}{\overset{TLC}{\approx}} 1 - F_{\mathcal{N}(0,1)}(-a)
\end{aligned}$$

Vitesse de convergence de la méthode de Monte-Carlo

$$
\begin{aligned}
P(e \le M_n - \epsilon) &= P(\epsilon \le M_n - e) \\
&= P\left(\frac{\epsilon\sqrt{n}}{\sigma_n} \le \frac{\sqrt{n}(M_n - e)}{\sigma_n}\right) \\
&= P(a \le Y_n), \;\; a = \frac{\epsilon\sqrt{n}}{\sigma_n}, Y_n = \frac{\sqrt{n}(M_n - e)}{\sigma_n} \\
&= 1 - P(Y_n < a) \\
&\underset{n \gg 1}{\overset{TLC}{\approx}} 1 - F_{\mathcal{N}(0,1)}(a)
\end{aligned}
$$

$$
P(e \in [M_n - \epsilon, M_n + \epsilon]) = F_{\mathcal{N}(0,1)}(a) - F_{\mathcal{N}(0,1)}(-a) = 2F_{\mathcal{N}(0,1)}(a) - 1
$$

Vitesse de convergence de la méthode de Monte-Carlo

On a donc

$$2F_{\mathcal{N}(0,1)}\left(\epsilon\frac{\sqrt{n}}{\sigma_n}\right) - 1 = \alpha \iff \epsilon = \frac{\sigma_n}{\sqrt{n}} F^{-1}_{\mathcal{N}(0,1)}\left(\frac{\alpha+1}{2}\right)$$

Ce qui nous permet d'obtenir ϵ:

$$\epsilon = \frac{\sigma_n}{\sqrt{n}} F^{-1}_{\mathcal{N}(0,1)}\left(\frac{\alpha+1}{2}\right)$$

Vitesse de convergence de la méthode de Monte-Carlo

Propriété:

X_i v.a. i.i.d. de densite $p(x) > 0$, et $\phi(x) = f(x)/p(x)$.

$$M_n = \frac{\phi(X_1) + \ldots + \phi(X_n)}{n} \text{ moyenne emp.}$$

$$\sigma_n = \sqrt{\frac{\phi(X_1)^2 + \ldots + \phi(X_n)^2}{n} - (M_n)^2} \text{ ec. type emp.}$$

Soit $\alpha \in]0,1[$ le **niveau de confiance asymptotique de l'intervalle**. L'intervalle

$$I_\alpha^n = \left[M_n - \frac{\sigma_n}{\sqrt{n}} F^{-1}_{\mathcal{N}(0,1)} \left(\frac{\alpha+1}{2} \right), M_n + \frac{\sigma_n}{\sqrt{n}} F^{-1}_{\mathcal{N}(0,1)} \left(\frac{\alpha+1}{2} \right) \right]$$

est tel que

$$\lim_{n\to\infty} P\left(\int f \in I_\alpha^n \right) = \alpha$$

Changement de variable affine d'une v.a. a densité

Propriété:
X v.a. de densité $p_X(x)$, $a \in \mathbb{R}^*$ et $b \in \mathbb{R}$, $Y = aX + b$.
Y admet la densité

$$p_Y(y) = \frac{1}{|a|} p_X\left(\frac{y-b}{a}\right)$$

- Si $X \sim \mathcal{N}(m, \sigma^2)$ alors $\frac{X-m}{\sigma} \sim \mathcal{N}(0,1)$.
- Si $X \sim \mathcal{N}(0,1)$ alors $aX + b \sim \mathcal{N}(b, a^2)$.
- Si $X \sim \mathcal{U}(\alpha, \beta)$ alors $aX + b \sim \mathcal{U}(a\alpha + b, a\beta + b)$.

Somme de deux v.a. a densité

Propriété:
X et Y v.a. **indépendantes** de densité respectives p_X, et p_Y.
$Z = X + Y$ admet la densité

$$p_Z(z) = \int_{\mathbb{R}} p_X(z - x) p_Y(x) dx = \int_{\mathbb{R}} p_X(x) p_Y(z - x) dx$$

p_Z est le produit de convolution $p_X * p_Y = p_Y * p_X$ des densites p_X et p_Y.

Remarques: C'est la généralisation au cas continu de la somme de deux v.a. discrètes a valeur dans $\mathbb{Z}$ avec le produit de convolution discret.En effet, si Z vaut z alors X et Y peuvent valoir $z - x$ et x pour tout x, et par indépendance

$$P(X = z - x, Y = x) = P(X = z - x) P(Y = x)$$

TLC pour variables centrées réduites.

X_i une suite de v.a. i.i.d. de moyenne m et de variance $\sigma^2 > 0$, sans perte de généralité on peut supposer que $m = 0$ et $\sigma = 1$, simplement en remplaçant X_i par la variable centrée réduite $\frac{X_i - m}{\sigma}$, ce qui simplifie l'énonce du TLC.

Théorème:
Soit X_n une suite de v.a. i.i.d. de moyenne $m = 0$ et de variance $\sigma^2 = 1$,. Alors

$$\frac{\sum_{i=1}^{n} X_i}{\sqrt{n}} \underset{n \to \infty}{\overset{\mathcal{L}}{\to}} \mathcal{N}(0,1)$$

Comprendre le TLC

X_i une suite de v.a. i.i.d. de loi $\mathcal{N}(m, \sigma^2)$.

- Quelle est la loi de $\frac{X_i - m}{\sigma}$? $\rightarrow \mathcal{N}(0, 1)$
- Quelle est la loi de $X_i + X_j$?

Comprendre le TLC

Loi de $X_i + X_j$:

$$\begin{aligned}
p_{X_i+X_j}(y) &= \frac{1}{2\pi}\int e^{-x^2/2}e^{-(y-x)^2/2}dx \text{ convolution} \\
&= \frac{1}{2\pi}\int e^{-(2x^2-2xy+y^2)/2}dx \\
&= \frac{e^{-y^2/4}}{2\pi}\int e^{-(x^2-xy+(y/2)^2)}dx \\
&= \frac{e^{-y^2/4}}{2\pi}\int e^{-(x-y/2)^2}dx \\
&= \frac{1}{\sqrt{2\pi}\sqrt{2}}e^{-y^2/2(\sqrt{2}^2)}, \text{ chgt var. } z = x - y/2 \\
&= p_{\mathcal{N}(0,2)}(y)
\end{aligned}$$

Exercice: $X \sim \mathcal{N}(m_1, \sigma_1^2)$, $Y \sim \mathcal{N}(m_2, \sigma_2^2)$, X et Y indépendantes. Loi de X + Y ?

Comprendre le TLC

X_i une suite de v.a. i.i.d. de loi $\mathcal{N}(m, \sigma^2)$.

- Quelle est la loi de $\frac{X_i - m}{\sigma}$? $\to \mathcal{N}(0, 1)$
- Quelle est la loi de $X_i + X_j$? $\to \mathcal{N}(0, 2)$.
- Quelle est la loi de $\sum_{i=1}^{n} X_i$? $\to \mathcal{N}(0, n)$
- Quelle est la loi de $\frac{\sum_{i=1}^{n} X_i}{\sqrt{n}}$? $\to \mathcal{N}(0, 1)$

Les v.a. Gaussiennes vérifient exactement le TLC des $n = 1$ (!!).

Comprendre le TLC

Soient l'ensemble et la fonctionnelle

$$\mathcal{P}_{\mathrm{cr}} = \{p : \mathbb{R} \to \mathbb{R}^+, \int p = 1, \int xp(x)dx = 0, \int x^2 p(x)dx = 1\}$$

$$\mathcal{F}_n : (p : \mathbb{R} \to \mathbb{R}) \mapsto (\frac{p*^n}{\sqrt{n}} : \mathbb{R} \to \mathbb{R}) \text{ p convolé n fois ac p div par } \sqrt{n}$$

- Quelle est la structure de $\mathcal{P}_{\mathrm{cr}}$? Espace fonctionnel stable par combinaison convexe.
- Comment interprétez vous cet ensemble ? Densités des v.a. centrées réduites.
- $\mathcal{F}$ est elle définie et a valeur $\mathcal{P}_{\mathrm{cr}}$? Oui
- Quelle est l'interprétation de $\mathcal{F}_n(p)$? Densité de $\sum_i X_i/\sqrt{n}$, $X_i \sim p$
- La densité de la loi $\mathcal{N}(0,1)$ présente une propriété particulière pour $\mathcal{F}_n$ quelle est-elle ? C'est un point fixe de $\mathcal{F}_n$ $\forall n$.
- Que nous dit le TLC sur $\lim_n F_n(p)$? Ca converge vers la densité de la loi $\mathcal{N}(0,1)$, point fixe. Le TLC est un sorte de Thm de point fixe.

Tests d'hypothèses

But

Contredire/confirmer une affirmation concernant un phénomène aléatoire étudié en se fondant sur l'observation de données.

Exemples:

- On veut savoir si une pièce de monnaie est équilibrée ou non a partir d'une série de 500 lancers.
- On veut savoir si un traitement médical rallonge la durée de vie des patients
- On veut savoir si un nouveau procédé accélère la fabrication industrielle d'un objet.

Tests d'hypothèses

Les hypothèses

Dans un test d'hypothèses, on formule deux hypothèses:

1. La première hypothèse H_0, souvent appelée **hypothèse nulle**, affirme que la situation est "normale", que rien n'a évolué.Elle est considérée comme vraie a priori.
   - Dans l'exemple de la pièce de monnaie, H_0 affirme que la pièce est normale, c'est à dire équilibrée.
   - Dans l'exemple du traitement médical, H_0 affirme que le traitement médical n'a pas d'effet sur la durée de vie.
2. La deuxième hypothèse H_1 est **l'hypothèse de travail** ou **l'hypothèse alternative** .C'est souvent la négation de l'hypothèse nulle H_0.
   - Dans l'exemple de la pièce de monnaie, l'hypothèse H_1,"La pièce de monnaie est déséquilibrée", est l'hypothèse contraire de H_0.
   - Dans l'exemple du traitement médical, l'hypothèse H_1 est "Le traitement médical rallonge la durée de vie". Ce n'est pas l'hypothèse contraire,et l'hypothèse "Le traitement médical raccourcit la durée de vie" n'est pas envisagée dans ce cas.

Tests d'hypothèses

Les hypothèses

Remarques:

- Un test n'est pas symétrique. H_0 supposée vraie a priori. Par exemple, une pièce de monnaie est supposée équilibrée a vue d'oeil.
- Le rejet de l'hypothèse nulle H_0 doit être basé sur des arguments forts. Ex: obtenir 995 piles sur 2000 lancers a pile ou face n'est pas particulièrement surprenant si la pièce est équilibrée, ça ne doit pas conduire a conclure a rejeter H_0.

Tests d'hypothèses

Modélisation probabiliste et choix de la statistique de test

On modélise le phénomène aléatoire en supposant l'hypothèse nulle H_0 vraie.

- Dans l'exemple de la pièce de monnaie, on modélise le lancer d'une pièce dans le cas ou elle est équilibrée, c.a.d. par une Bernoulli X_i de paramètre $p = 0.5$.

On choisit la statistique de test Z: grandeur calculée a partir de l'échantillon qui permet de prendre la décision. On doit connaître la loi de Z.

- Dans l'exemple de la pièce de monnaie, on calcule Z le nombre de piles sur n lancers, ce nombre suit un loi Binomiale de paramètre $(n, 1/2)$. On peut également choisir comme statistique de test l'écart à la moyenne $Z' = |Z - \frac{n}{2}|$.
- Dans l'exemple de l'essai médical, la statistique Z peut-être la moyenne des durées de vies des patients.

Le niveau d'erreur du test
On se donne une marge d'erreur α qui correspond a la probabilité de se tromper.

- En général, $\alpha = 5\%$ ou $\alpha = 1\%$.
- Plus on a "confiance" en l'hypothèse H_0 (ou bien plus l'hypothèse H_1 est surprenante), plus on doit choisir une petite valeur pour α.

La p-valeur

La p-valeur est la probabilité sous l'hypothèse H_0 que la statistique de test Z prenne une valeur au moins aussi extrême que la valeur observée, compte tenu de l'hypothèse H_1.

Le calcul de la p-valeur dépend de l'hypothèse H_1.

- Pour la pièce de monnaie, si l'hypothèse H_1 est "la pièce est déséquilibrée en défaveur du cote pile", et que la statistique de test Z est la somme des piles, p-valeur est

  $$p = P(Z \leq z_{\mathrm{obs}})$$

  ou z_{obs} est la valeur de Z l'on observe. Si on observe 933 piles sur les 2000 lancers, la p-valeur est

  $$p = P(Z \leq 933)$$

Tests d'hypothèses

La p-valeur

La p-valeur est la probabilité sous l'hypothèse H_0 que la statistique de test Z prenne une valeur au moins aussi extrême que la valeur observée, compte tenu de l'hypothèse H_1.

Le calcul de la p-valeur dépend de l'hypothèse H_1.

- Si l'hypothèse H_1 est "la pièce est déséquilibrée", et que la statistique de test est $Z' = |Z - 1000|$, la p-valeur est

  $$p = P(Z' \geq z'_{\mathrm{obs}})$$

  ou z'_{obs} est la valeur de Z' l'on observe. Si on observe 933 piles sur les 2000 lancers, la p-valeur est

  $$p = P(Z' \geq 67) = P(Z \leq 933) + P(Z \geq 1067)$$

Tests d'hypothèses

- La p-valeur mesure le caractère exceptionnel des données observées si l'on suppose que H_0 est vraie : si la p-valeur est très petite, il est très peu probable d'obtenir une valeur plus extrême que ce qu'on observe. La statistique de test observée est peu probable.
- Le calcul de la p-valeur fait intervenir la loi de Z que l'on doit être capable de calculer, et les données observées via le calcul de z_{obs}.

Décision

- Si la p-valeur est plus petite que le niveau α fixé **au préalable**, on rejette H_0, H_1 est supposée vraie.
- Si la p-valeur est plus grande que le niveau α fixé **au préalable**, on accepte H_0.

Tests d'hypothèses

Petite règle empirique sur les p-valeurs

Plus la p-valeur est petite, plus le résultat est significatif.

- Une p-valeur < 0.001 signifie que le test sera extrêmement significatif.
- Une p-valeur entre 0.001 et 0.01 signifie que le test sera très significatif.
- Une p-valeur entre 0.01 et 0.05 signifie que le test significatif.
- Une p-valeur supérieure 0.05 conduit à accepter H_0 si $\alpha = 0.05$.

Mise en oeuvre d'un test d'hypothèses

1. Choix des hypothèses H_0 et H_1.
2. Modélisation du phénomène aléatoire et détermination de la statistique de test Z.
3. Calcul de la valeur observée z_{obs} et calcul la p-valeur p.
4. Décision par comparaison de p à α (et commentaire).

Tests d'hypothèses pour la pièce de monnaie

On désire savoir si une pièce de monnaie est équilibrée.
On effectue 2000 lancers et on obtient 933 piles et 1067 faces.
Soupçonnant un déséquilibre, on souhaite effectuer un test d'hypothèse de niveau $\alpha = 1\%$.

Tests d'hypothèses pour la pièce de monnaie

1. **Hypothèses**: H_0: La pièce est équilibrée, H_1: La pièce est déséquilibrée en défaveur du pile
2. **Modélisation**: on modélise les lancers pas des v.a. de Bernoulli X_i **indépendantes** de paramètre $p = 1/2$.
   Statistique de test: $Z = \sum_{i=1}^{n} X_i$ suit une loi $\mathcal{B}(2000, 1/2)$.
3. **p-valeur**: On observe $z_{\text{obs}} = 933$. Vues H_0 et H_1, une valeur Z sera plus extrême si $Z \leq 933$:

$$p = P(Z \leq 933) = \sum_{k=0}^{933} \binom{2000}{k} 0.5^{2000} = 0.00146$$

4. **Décision**: $p \leq \alpha = 1\%$. On rejette H_0, la pièce est déséquilibrée en faveur du pile. Le test est très significatif compte tenu de la p-valeur.

Tests d'hypothèses pour la pièce de monnaie

1. **Hypothèses**: H_0: La pièce est équilibrée, H_1: La pièce est déséquilibrée.
2. **Modélisation**: lancers v.a. de Bernoulli X_i indép. de paramètre $p = 1/2$.
   Statistique de test: $Z = \sum_{i=1}^{n} X_i$ suit une loi $\mathcal{B}(2000, 1/2)$.
3. **p-valeur**: On observe $z_{\mathrm{obs}} = 933$. Vues H_0 et H_1, une valeur Z sera plus extrême si $Z \leq 933$ ou si $Z \geq 1067$:

$$p = P(Z \leq 933) + P(Z \geq 1067) = 1 - P(933 < Z < 1067)$$
$$= 1 - \sum_{k=934}^{1066} \binom{2000}{k} 0.5^{2000} = 0.293\%$$

4. **Décision**: $p \leq \alpha = 0.01$. On rejette H_0, la pièce est déséquilibrée en faveur du pile. Le test est très significatif compte tenu de la p-valeur.

Tests d'hypothèses pour un temps de moyen

Un ouvrier spécialisé passe un temps variable sur chaque pièce.Le temps moyen est de $m = 270$ secondes et l'écart-type est de $\sigma = 24$ secondes. Une modification technique du montage est proposée pour diminuer ce temps moyen.Pour le tester, on chronomètre $n = 50$ montages avec la modification technique et on obtient une moyenne empirique $M_n = 266$.

Proposer un test d'hypothèse de niveau $\alpha = 0.05$ pour savoir si le temps moyen m' avec le nouveau procédé est inférieur au temps moyen m actuel.

Test d'hypothèse pour un temps moyen

Données: $m = 270\ s, \sigma = 24\ s,\ n = 50$ et $M_n = 266\ s$.

1. **Hypothèses**: H_0: le temps moyen ne change pas : $m' = m$, H_1: Le temps moyen diminue $m' < m$.
2. **Modélisation**: Les temps de montage T_i de chaque pièce sont des v.a. indépendantes, de même loi (moyenne m, écart-type σ) .
   Statistique de test: $Z = \frac{\sqrt{n}(M_n - m)}{\sigma}$ qui suit une loi normale $\mathcal{N}(0,1)$ par le TLC.
3. **p-valeur**: On observe $z_{\text{obs}} = -1.18$. Vues H_0 et H_1, Z sera plus extrême que z_{obs} si $Z \leq -1.18$.

$$P(Z \leq -1.18) = F_{\mathcal{N}(0,1)}(-1.18) = 11.9\%$$

4. **Décision**: $p \geq \alpha = 1\%$. On accepte H_0. La différence n'est pas *statistiquement significative*.

Asymétrie du test d'hypothèse

Un test d'hypothèses est asymétrique car on part du principe que H_0 est vraie. On ne s'est servi nulle part de l'écart-type Σ_n de la nouvelle technique de temps de montage.
On a $\sigma_n = 10$ secondes. Tester l'hypothèse $m' = m$ en supposant H_1 vraie, donc en utilisant les valeurs $M_n = 266$ et $\sigma_n = 10$ pour la statistique de test, avec et le même niveau $\alpha = 0.05$.

1. **Hypothèses**: $H_0 : \ m = m'$ et $H_1 : \ m > m'$.
2. **Modélisation**: Statistique de test $Z = \frac{\sqrt{n}(m - M_n)}{\Sigma_n}$ qui suit une loi normale $\mathcal{N}(0,1)$ par le TLC.
3. **p-valeur**: On observe $z_{\text{obs}} = 2.83$. Vues H_0 et H_1, Z sera plus extrême que z_{obs} si $Z \geq 2.83$.

$$
\begin{aligned}
P(Z \geq 2.83) &= 1 - P(Z < 2.83) \\
&= 1 - F_{\mathcal{N}(0,1)}(2.83) \\
&= 0.233\%
\end{aligned}
$$

Asymétrie du test d'hypothèse

Tester l'hypothèse $m' = m$ en
prenant M_n et Σ_n comme valeurs de référence.Rappel: $m = 270\ s$ et $n = 50$.

4. **Décision**: $p \leq \alpha = 1\%$. On rejette H_0 dans ce cas, et la différence est *statistiquement très significative*.

Illustration concrète de l'asymétrie : avec H_0 comme référence, on confirme H_0 tandis qu'avec H_1, on rejette H_0 et la p-valeur (0.2%) est 50 fois plus petite qu'en prenant H_0 comme référence (10%).

Dépendance en n du test d'hypothèse.

Pour savoir précisément, on réitère l'expérience de la pièce $n = 4000$ fois, et on obtient la meme moyenne empirique $M_n = 266\ s$. Que nous donne le test d'hypothèse classique (en supposant H_0 vraie) avec ces valeurs ?
Rappel: $m = 270\ s, \sigma = 24\ s$.

1. **Hypothèses** H_0: I : $m' = m$, H_1: $m' < m$.
2. **Modélisation:** idem.
   Statistique de test $Z = \frac{\sqrt{n}(M_n - m)}{\sigma}$ qui suit une loi normale $\mathcal{N}(0, 1)$ par le TLC.
3. **p-valeur**: On observe $z_{\rm obs} = \frac{-4\sqrt{4000}}{24} = -10.55$. Vues H_0 et H_1, Z sera plus extrême que $z_{\rm obs}$ si $Z \leq -10.54$.

$$P(Z < -10.54) = -2.82\ 10^{-26}\%$$

4. **Décision**: $p \ll \alpha = 1\%$. On rejette H_0 dans ce cas, et la différence est *statistiquement extrêmement significative*.

Dépendance en n du test d'hypothèse.

Remarques:

- Quand on change un procédé de montage, il est extrêmement probable que le temps de montage moyen ne soit pas exactement le même, c'est a dire que H_0 soit fausse.
- Même si la nouvelle moyenne est très proche de l'ancienne, avec n assez grand, la nouvelle moyenne aura converge vers sa valeur. Le TLC nous donnera toujours une très petite p-valeur si $\sqrt{n}$ est assez grand.
- On a le même principe sur l'effet moyen d'un médicament ou d'une substance chimique.

Autre possibilité de mise en oeuvre

1. Choix des hypothèses H_0 et H_1.
2. Modélisation du phénomène aléatoire et détermination de la statistique de test Z.
3. On fixe la p-valeur $p = \alpha$ et on calcule le z_α correspondant.
4. Calcul de z_{obs}. Décision par comparaison z_{obs} a z_α (et commentaire).

Remarques:

- Avantage: on fait le calcul de z_α une fois pour toutes.
- Inconvénient: il faut calculer la p-valeur p pour connaître la "force" du résultat.

Tests d'hypothèses pour un temps de vie sans vieillissement

Selon le fabriquant, la durée de vie des ampoules suit une loi exponentielle et la durée de vie moyenne des ampoules est de 4 ans.
J'ai acheté une ampoule qui a grillé au bout de 1 mois et 28 jours ($=59$ jours).

- Avec une probabilité de me tromper de 1%, puis-je faire une réclamation au fabriquant pour ampoule défectueuse ?
- Et avec une probabilité de me tromper de 5% ?

Essayer les deux mises en oeuvre: par calcul de la p-valeur et par calcul de z_α.

Rappel: une v.a. X de loi exponentielle est a valeur dans $\mathbb{R}^+$ et a pour densité de probabilité $f(x) = \lambda \exp(-\lambda x)$ et $\mathbb{E}(X) = 1/\lambda$.

Tests d'hypothèses pour un temps de vie sans vieillissement

Par calcul de la p-valeur p:

1. **Hypothèses** H_0: L'ampoule est conforme (elle suit une loi exponentielle de moyenne 4 ans $\lambda = 0.25$) H_1: L'ampoule est défectueuse.
2. **Modélisation:** donnée.
   Statistique de test: Z temps de vie de l'ampoule.
3. **p-valeur**: On observe $z_{\text{obs}} = 59j = 0.161$an. Vues H_0 et H_1, une valeur Z sera plus extrême si $Z \leq 0.161$:

$$p = P(Z \leq 0.161) = 3.95\%$$

4. Décision :
   - $\alpha = 0.01$, $p > \alpha$. On ne rejette pas H_0.
   - $\alpha = 0.05$, $p \leq \alpha$. On rejette H_0.

Tests d'hypothèses pour un temps de vie sans vieillissement

Par calcul de z_α:

1. **Hypothèses** H_0: L'ampoule est conforme (elle suit une loi exponentielle de moyenne 4 ans $\lambda = 0.25$) H_1: L'ampoule est défectueuse.
2. **Modélisation:** donnée.
   Statistique de test: Z temps de vie de l'ampoule.
3. **Calcul de $z_{\boldsymbol{\alpha}}$:**

$$P(Z \leq z_\alpha) = \alpha \iff Z \leq -4\log(1-\alpha)$$

- $\alpha = 0.01,\ z_\alpha = 0.040 = 14j \leq z_{\text{obs}}$. On ne rejette pas H_0.
- $\alpha = 0.05, z_\alpha = 0.205 = 73j > z_{\text{obs}}$. On rejette H_0.

Loi multinomiale

Une loi **bi**nomiale modélise le comptage des valeurs dans une expérience a **deux** possibilités (ex. pile ou face) répétée n fois. Si on connaît le nombre de piles n_1 et la probabilité p d'avoir pile, le nombre de face est $n - n_1$ et la probabilité d'avoir face est $1 - p$.

Une loi **multi**nomiale modélise le comptage des valeurs dans une expérience a d possibilités. Comme il y a plus de 2 valeurs, c'est la loi d'un vecteur aléatoire M. Dans l'exemple d'un dé non pipé à 6 faces lancé 30 fois obtenant 3 faces un, 4 faces 2, 6 faces 3, 7 faces quatre, 6 faces cinq et 4 faces 6, le vecteur aléatoire de comptage vaut

$$M = (M^1, M^2, M^3, M^4, M^5, M^6) = (3, 4, 6, 7, 6, 4)$$

Ce vecteur aléatoire suit une loi multinomiale de paramètre $n = 30$ et $\pi = (1/6, 1/6, 1/6, 1/6, 1/6, 1/6)$ On peut écrire formellement la loi d'une multinomiale en fonction de π.

Loi multinomiale

Une v.a. B de loi **bi**nomiale $\mathcal{B}(n, p)$ compte les occurrences de 1 de v.a. de Bernoulli a **deux** valeurs quand on fait n lancers indépendants. Classiquement on écrit sa loi

$$\mathbb{P}(B = k) = \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k}$$

On peut la reecrite en definissant B_0 qui compte les occurences de 0 et B_1 $(= B)$ qui compte les occurences de 1. On a

$$B^1 + B^2 = n$$

On définit également les probabilités $p_0 = 1 - p$ et $p_1 = p$ d'avoir 0 ou 1. On a

$$\mathbb{P}((B^0, B^1) = (n_0, n_1)) = \frac{n!}{n_0! n_1!} p_0^{n_0} p_1^{n_1}, \quad n_0 + n_1 = n$$

Loi multinomiale

La loi multinomiale s'écrit par généralisation de cette écriture avec le paramètre $\pi = (\pi_1, \ldots, \pi_d)$ et n:

$$\mathbb{P}((M^1, \ldots, M^d) = (n_1, \ldots n_d)) = \frac{n!}{n_1! \ldots n_d!} \pi_1^{n_1} \ldots \pi_d^{n_d}, \quad n_1 + \ldots + n_d = n$$

- Il faut bien entendu que $\pi_k \geq 0$ et $\sum_{k=1}^{d} \pi_k = 1$
- Cette écriture est un fastidieuse mais bien pratique pour les calculs.

Test d'hypothèse pour une loi multinomiale

On a vu le test d'hypothèse pour la pièce de monnaie (loi binomiale). Comment le généraliser pour le dé équilibré à 6 faces, ou à d faces (loi multinomiale $M = (M^1, \dots, M^d)$ avec n lancers ?

Les variables aléatoires M^i, $i = 1 \dots d$ comptent les occurrences de chacune des faces. Elle suivent une loi binomiale de paramètres n et $p_i = 1/d$ dans le cas équilibré, ou $p_i = \pi_i$ dans le cas général. On peut les tester séparément.

- Cela oblige a faire d tests. On aimerait en faire un.
- il suffit que le test soit négatif pour une seule des binomiales B^i pour invalider l'hypothèse.
- les B_i ne sont pas indépendances car leur somme vaut n le nombre de lancers

Que faire ? Un test de χ^2.

Loi χ^2_d

Définition

Une variable aléatoire X suit une loi χ^2_d (prononcer "qui deux") à d degrés de liberté si elle suit la même loi que la somme de d carrés des variables aléatoires N_i normales $\mathcal{N}(0,1)$ indépendantes.
De plus $\mathbb{E}(X) = \sum_{i=1}^{d} Var(N_i) = d$.

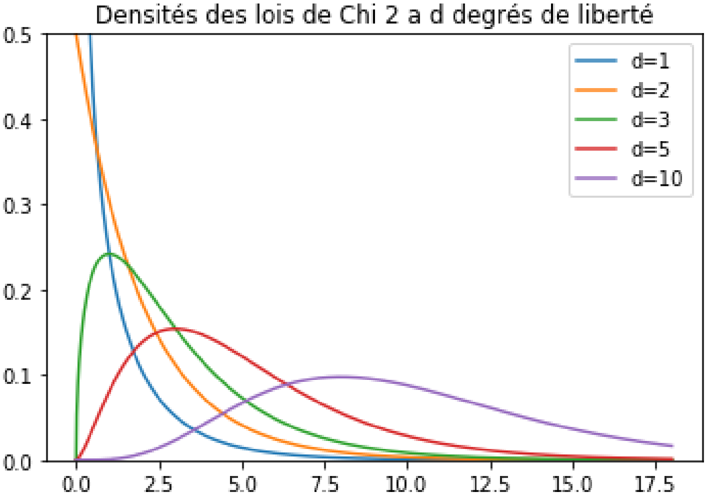
Propriété

Une variable aléatoire X suit une loi χ^2_d à d degrés de liberté admet la densité de probabilité

$$f_d(x) = \frac{1}{2^{d/2}\Gamma(d/2)} x^{d/2-1} e^{-x/2}$$

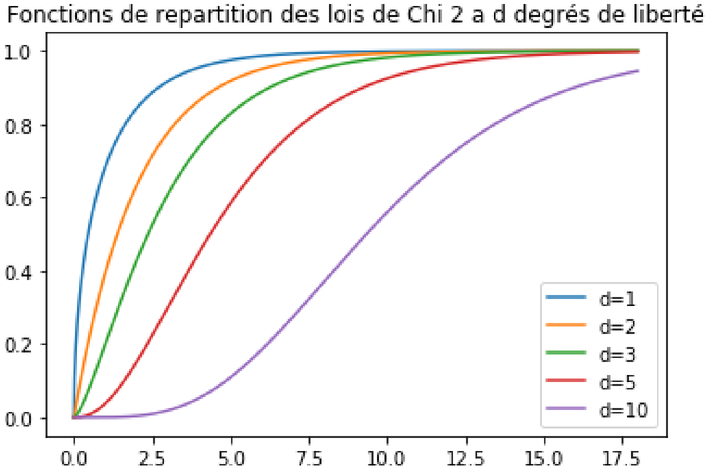
Densité de loi χ^2_d

Python: scipy.stats.chi2.pdf



Fonction de répartition de loi χ^2

Python: scipy.stats.chi2.cdf



Propriété sur les lois multinomiales

Propriété

Soit $M = (M^1, \ldots, M^d)$ un vecteur aléatoire de loi multinomiale de paramètres n et $\pi = (\pi_1, \ldots, \pi_d)$. La variable aléatoire

$$Z = \sum_{k=1}^{d} \frac{(M^k - n\pi_k)^2}{n\pi_k}$$

converge en loi vers une v.a. de loi χ^2 a $d-1$ degrés de libertés quand n tend vers l'infini.

$$\sum_{k=1}^{d} \frac{(M^k - n\pi_k)^2}{n\pi_k} \underset{n\to\infty}{\overset{\mathcal{L}}{\to}} \chi^2_{d-1}$$

Propriété sur les lois multinomiales

Remarques:

- M^k est le nombre qu'on observe n^k_{obs} pour la valeur k. $n\pi_k$ est le nombre moyen n^k_{moy} pour la valeur k (l'esperance). Z s'écrit alors

$$Z = \sum_{k=1}^{d} \frac{\left(n^k_{\mathrm{obs}} - n^k_{\mathrm{moy}}\right)^2}{n^k_{\mathrm{moy}}}$$

- Cela nous donne une statistique de test Z idéale pour un test d'hypothèse: elle se calcule directement et on connaît sa loi.
- Moyen mémotechnique pour les $d-1$ degrés de liberté du χ^2 : la contrainte $\sum_k \pi_k = 1$ fait que π a $d-1$ degrés de liberté.
- Attention: les termes de la somme ne sont pas le carre des v.a. qui interviennent dans le TLC appliqué aux lois binomiales M_k séparément :

$$\left(\frac{M^k - n\pi_k}{\sqrt{n\pi_k(1-\pi_k)}}\right)^2 \neq \frac{(M^k - n\pi_k)^2}{n\pi_k}$$

Propriété sur les lois multinomiales

Preuve pour d=2:
Par définition,

$$Z = \frac{(M^1 - n\pi_1)^2}{n\pi_1} + \frac{(M^2 - n\pi_2)^2}{n\pi_2}$$

En utilisant le fait que $M^1 + M^2 = n$ et $\pi_2 = 1 - \pi_1$, on a

$$\begin{aligned}
Z &= \frac{(M^1 - n\pi_1)^2}{n\pi_1} + \frac{(n - M^1 - n(1-\pi_1)^2}{n(1-\pi_1)} \\
&= \frac{(M^1 - n\pi_1)^2}{n\pi_1} + \frac{(M^1 - n\pi_2)^2}{n(1-\pi_1)} \\
&= \frac{(M^1 - n\pi_1)^2}{n}\left(\frac{1}{\pi_1} + \frac{1}{1-\pi_1}\right) \\
&= \left(\frac{M^1 - n\pi_1}{\sqrt{n\pi_1(1-\pi_1)}}\right)^2
\end{aligned}$$

Propriété sur les lois multinomiales

$$Z = \left(\frac{M^1 - n\pi_1}{\sqrt{n\pi_1(1-\pi_1)}} \right)^2$$

On reconnaît le carre de la loi $\frac{M^1 - n\pi_1}{\sqrt{n\pi_1(1-\pi_1)}}$ qui suit une loi $\mathcal{N}(0,1)$ quand $n \to \infty$ d'après le TLC pour la v.a. M^1 de loi binomiale $\mathcal{B}(n, \pi_1)$.
La v.a. Z suit donc la meme loi que le carre d'une v.a. de loi $\mathcal{N}(0,1)$, ce qui est par définition une loi χ^2_1, et $d = 2$.

La preuve dans le cas général $d \geq 3$ est un peu plus technique.

Test d'adéquation du χ^2 pour une loi multinomiale

On effectue 60 lancers de dé et on obtient pour les valeurs 1 a 6 $(7, 8, 12, 15, 10, 8)$. Peut on dire que le dé est-il équilibré avec $\alpha = 1\%$?

- H_0: Le dé est équilibré. H_1: le dé n'est pas équilibré.
- Sous H_0, le comptage des valeurs suit une loi multinomiale de paramètre $(1/6, \ldots, 1/6)$.
  Statistique de test:

$$\begin{aligned} Z =& \frac{(7-10)^2}{10} + \frac{(8-10)^2}{10} + \frac{(12-10)^2}{10} \\ &+ \frac{(15-10)^2}{10} + \frac{(10-10)^2}{10} + \frac{(8-10)^2}{10} \end{aligned}$$

  suit une loi χ^2_5.

Test d'adéquation du χ^2 pour une loi multinomiale

On effectue 60 lancers de dé et on obtient pour les valeurs 1 a 6 $(7, 8, 12, 15, 10, 8)$. Le dé est-il équilibré ?

- On observe $z_{\text{obs}} = 4.6$. Une valeur est plus extrême si $Z \geq 4.6$.

$$p = P(Z \geq 4.6) = 1 - F_{\chi^2_5}(4.6) = 46.6\%$$

- Décision: $p > 1\%$: on ne rejette pas H_0. On a une chance sur deux d'obtenir une valeur plus extrême.

Soit M une loi multinomiale de paramètre n et $\pi = (\pi_1, \dots, \pi_d)$.

Test d'adéquation du χ^2 pour une loi multinomiale

Cas général:
On observe $m = (m^1, \ldots, m^d)$ d'une multinomiale M de paramètre n connu et π inconnue. On souhaite savoir si cela correspond à une multinomiale de paramètre n et $= (p_1, \ldots, p_d)$ avec une erreur alpha.

- H_0: $\pi = p$ H_1: $\pi \neq p$.
- Statistique de test:

$$Z = \sum_{k=1}^{d} \frac{(M^k - np_k)^2}{np_k} \text{ suit une loi } \chi^2_{d-1}.$$

- Calcul de z_{obs} et de la p-valeur

$$z_{\mathrm{obs}} = \sum_{k=1}^{d} \frac{(m^k - np_k)^2}{np_k}$$

$$p = 1 - F_{\chi^d_{d-1}}(z_{\mathrm{obs}})$$

- on conclut.

Test d'adéquation du χ^2 pour une loi multinomiale

Exercice python:

- Générer un v.a. de loi multinomiale équilibrée sur $\{1, \ldots, 6\}$ avec $n = 60$ et faire un test d'adéquation de χ^2 avec $\alpha = 0.01$
- Répéter l'opération $N = 1000$ fois et compter le nombre n_r de cas qui rejettent H_0.
  Qu'observe-t-on sur n_r ?
- Répéter cette expérience en faisant varier n, N et α.

Fonctions python:

- np.random.multinomial pour échantillonner une loi multinomiale.
- scipy.stats.chisquare pour faire un test d'adéquation de χ^2.

Test d'adéquation du χ^2

Dans le cas général:

- on n'a pas une loi multinomiale: on s'y ramène en faisant une partition finie de taille d de l'espace des valeurs. Une regle emprique veut que le nombre moyen d'elements par partie sous l'hypothese H_0 soir supérieur ou égal a 5.
- On ne connait pas systématiquement les parametres de la loi. Par exemple, si on veut vérifier que des données sont tirées d'une loi normale, on doit souvent estimer la moyenne et la variance. On estime alors ces paramètres. Si on estime n_p paramètres, le nombre de degrés de liberté dans le test de χ^2 diminue de n_p: il passe de $d - 1$ a $d - n_p - 1$

Test d'adéquation du χ^2

Exemple: adéquation à une loi de Poisson.
On veut savoir le nombre de pièce défectueuses par lot de 100 pièces suit une loi de poisson: $P(X = k) = \exp(-\lambda)\lambda^k/k!$. On teste $n = 52$ lots et on trouve les valeurs suivantes:

pièces défectueuses	0	1	2	3	4	5
lots	16	20	8	5	2	1

- On estime $\hat{\lambda} = \mathbb{E}(X) = 1.23$.
- On choisit par défaut la partition de $\mathbb{N}$ la plus naturelle au vu des données et on calcule les effectifs théoriques

$$\big\{\{0\},\{1\},\{2\},\{3\},\{4\},\{5,\ldots\},\big\},\ [15.2, 18.7, 11.5, 4.7, 1.46, 0.44]$$

- On regroupe donc les trois dernières parties pour dépasser 5. On a la partition et effectifs théoriques et observes suivants

$$\big\{\{0\},\{1\},\{2\},\{3,\ldots\},\big\}, \quad [15.2, 18.7, 11.5, 6.6], \quad [16, 20, 8, 8]$$

Test d'adéquation du χ^2

Exemple: adéquation à une loi de Poisson.
On veut savoir le nombre de pièce défectueuses par lot de 100 pièces suit une loi de poisson: $P(X = k) = \exp(-\lambda)\lambda^k/k!$. On teste $n = 52$ lots et on trouve les valeurs suivantes:

pièces défectueuses	0	1	2	3	4	5
lots	16	20	8	5	2	1

- On effectue un test de χ_2 avec $4 - 1 - 1 = 2$ degrés de liberté.
- H_0 : le nombre de pièces défectueuses par lot suit une loi de Poisson de paramètre $\hat{\lambda} = 1.23$, $H_1 = \overline{H_0}$.
- Avec python: scipy.stats.chisquare([16, 20, 8 , 8], [15.187, 18.692, 11.502, 6.617], ddof=1)

$$Z = 1.49, p = 47\%$$

- Décision: on accepte l'hypothèse H_0.

Test d'adéquation du χ^2

Exercice: adéquation à une loi Normale.
On veut savoir si n le rendement X (quintaux par hectares d'une parcelle de blé suit une loi normale $\mathcal{N}(m, \sigma^2)$. L'observation du rendement de 1000 parcelles a donne les résultats suivants :

rendement	0-10	10-20	20-30	30-40	40-50	50-60	60-70
parcelles	5	6	40	168	288	277	165

rendement	70-80	80-90	Total
parcelles	49	2	1000

Effectuer un test d'adéquation du χ^2.

Test d'indépendance du χ^2

On étudie deux caractères sur une population et on veut tester si ces caractères sont statistiquement indépendants. On suppose ces caractères qualitatifs, c'est a dire qu'ils ne correspondent pas a une quantité et qu'on ne peut pas les comparer entre eux.
Ce peut être, par exemple, la couleur des yeux et la couleur des cheveux.

Supposons que le premier caractère X prenne r valeurs x_i (ex: les yeux bleu, vert, marron, noir) et que le deuxième caractère Y prenne s valeurs y_j (ex: cheveux blonds, bruns, châtains, roux, gris, blancs). L'hypothèse H_0, "les caractères sont indépendants" donne $\forall 1 \leq i \leq r,\ 1 \leq j \leq s$

$$P(X = x_i, Y = y_j) = P(X = x_i)P(Y = y_j)$$

Test d'indépendance du χ^2

Méthode

- on calcule $\hat{p}_{x_i}$ et $\hat{p}_{y_j}$ estimations des probabilités $P(X = x_i)$ et $P(Y = y_j)$, $(r - 1 + s - 1$ estimations au total).
- On se ramène au test d'adéquation du χ^2 avec une loi multinomiale. Au total, il y a $r \times s$ valeurs $(X = x_i, Y = y_j)$.
- Sous H_0, les valeurs théoriques sont

$$P(X = x_i, Y = y_j) = \hat{p}_{x_i}\hat{p}_{y_j}$$

- Lorsqu'on travaille sur $r \times s$ valeurs, il y a $r \times s - 1$ degrés de liberté dans le χ^2. Comme on a estimé $r - 1 + s - 1$ paramètres, il y en a

$$rs - 1 - (r - 1 + s - 1) = (r - 1)(s - 1).$$

Test d'indépendance du χ^2

Si on note $n^{i,j}_{\mathrm{moy}} = n\hat{p}_{x_i}\hat{p}_{y_j}$ l'effectif théorique moyen de (x_i, y_j), et $n^{i,j}_{\mathrm{obs}}$ l'effectif observé de (x_i, y_j), la statistique de test est

$$Z = \sum_{i,j} \frac{(n^{i,j}_{\mathrm{obs}} - n^{i,j}_{\mathrm{moy}})^2}{n^{i,j}_{\mathrm{moy}}}$$

et elle suit une loi de χ^2 avec $(r-1)(s-1)$ degrés de liberté.

- Si il y a 2 degrés de libertés ou plus, alors on prend la statistique de test Z définie ci-dessus.
- S'il n'y a qu'un degré de liberté dans le χ^2, c.a.d $(r-1)(s-1) = 1$ soit $r = 2$ et $s = 2$, on applique en général la correction suivante dite de Yates à la statistique Z

$$Z^{\mathrm{yates}} = \sum_{i,j} \frac{(|n^{i,j}_{\mathrm{obs}} - n^{i,j}_{\mathrm{moy}}| - 0.5)^2}{n^{i,j}_{\mathrm{moy}}}$$

Test d'indépendance du χ^2

Correction de Yates:

- S'il n'y a qu'un degré de liberté dans le χ^2, c.a.d $(r-1)(s-1)=1$ soit $r=2$ et $s=2$, on applique en général la correction suivante dite de Yates à la statistique Z

$$Z^{\text{yates}} = \sum_{i,j} \frac{(|n_{\text{obs}}^{i,j} - n_{\text{moy}}^{i,j}| - 0.5)^2}{n_{\text{moy}}^{i,j}}$$

- Wikipedia FR sur la correction de Yates
- Wikipedia EN sur la correction de Yates

Test d'indépendance du χ^2

Exemple
Tester l'indépendance du type de film regarde et de la consommation ou non de snack en regardant le film.

Voici la **table de contingence**:

	snack		
film	oui	non	total
action	50	65	125
comédie	125	175	300
familial	90	30	120
horreur	45	10	55
total	310	290	600

Effectuer un test d'indépendance du χ^2.

Test d'indépendance du χ^2

film	snack oui	snack non	total
action	50	65	125
comédie	125	175	300
familial	90	30	120
horreur	45	10	55
total	310	290	600

Probabilités:

snack	oui	non
P	31 / 60	29 / 60

film	action	comédie	familial	horreur
P	5 / 24	1 / 2	1 / 5	11/120

Test d'indépendance du χ^2

Effectifs théoriques:
Sous l'hypothèse nulle d'indépendance: $n_{\mathrm{th}} = P(x) \times P(x) \times n$

	snack		
film	oui	non	total
action	65	60	125
comédie	155	145	300
familial	62	58	120
horreur	28	27	55
total	310	290	600

Test d'indépendance du χ^2

Contribution a la statistique de test:

$$\frac{(n_{\mathrm{obs}} - n_{\mathrm{th}})^2}{n_{\mathrm{th}}}$$

	snack	
film	oui	non
action	3.3	3.5
comédie	5.8	6.2
familial	12.6	13.5
horreur	9.7	10.3

$$Z = 3.3 + 3.5 + 5.8 + 6.2 + 12.6 + 13.5 + 9.7 + 10.3 = 64.9$$

Degrés de liberté: $(r-1)(s-1) = 3$

$$p = 1 - \mathcal{F}_{\chi^2_3}(64.9) = 5.310^{-14}$$

Test d'indépendance du χ^2

On teste un médicament X destiné à soigner une maladie. On traite des patients avec ce médicament tandis que d'autres reçoivent un placebo. On note dans la variable statut *oui* si les patients ont survécu plus de 48 jours et *non* sinon. Voici le tableau des résultats obtenus:

	statut		
	oui	non	total
médicament	17	29	46
placebo	8	38	46

Faire un test d'indépendance du χ^2 avec et sans correction de Yates.

Test d'indépendance du χ^2

Python: scipy.stats.chi2_contingency.

```
import scipy.stats
table = [[17,29],[8,38]]
scipy.stats.chi2_contingency(table, correction=False)
→(4.448955223880597, 0.03492260605235841, ...)
```

P-valeur de 3.49%, on rejette H_0 et on conclut a l'efficacité du médicament.

Test d'indépendance du χ^2

Avec la correction :

```
import scipy.stats
table = [[17,29],[8,38]]
scipy.stats.chi2_contingency(table, correction=True)
```

$\rightarrow$ (3.515223880597015, 0.0608074408980542, ...)

P-valeur de 6%. On accepte H_0. Comme quoi en changeant un seul paramètre dans le test, on change la conclusion.

Test d'indépendance du χ^2

Correction de Yates:
difficile de trouver un justification... Peut on proposer une expérience pour en vérifier le bien fonde ?

- Répéter N fois un test de χ^2 de niveau α sur deux lois binomiales indépendantes avec des paramètres de votre choix ($n > 20$, et $0 < p < 1$) **avec** la correction de Yates.
- Répéter N fois un test de chi2 de niveau α sur deux lois binomiales indépendantes avec les même paramètres ($n > 20$, et $0 < p < 1$) **sans** la correction de Yates.
- Faire varier une ou deux fois les paramètres pour estimer la sensibilité aux paramètres

Python: scipy.stats.chi2_contingency.

Temps d'arrêt dans les tests

Dans un plan d'expérience, on a souvent les résultat qui tombent les uns après les autres. Supposons qu'on ait prévu de faire une expérience sur 100 personnes pour démontrer que l'hypothèse nulle n'est pas vérifiée. En pratique, on effectue les expériences par groupes de 5 personnes, et on commence a calculer une p-valeur a partir de 20 personnes minimum, sans quoi on a pas convergence dans le TCL ou le χ^2. On peut être tenté de s'arrêter dès que la p-valeur est inférieure au niveau alpha fixé, après avoir fait un minimum de test (30 par exemple) mais avant d'avoir fait les expériences sur les 100 personnes.

- Dans ce cas nos hypothèses sont-elles toujours vérifiées ?
- A-t-on toujours correspondance entre alpha et la fréquence de rejet de l'hypothèse nulle alors qu'elle est vérifiée ?

Tests d'hypothèse Bayesiens

- Dans les test d'hypothèses que nous avons vus jusqu'ici, nous calculons la p-value, qui est la probabilité d'observer des données au moins aussi extrêmes :

$$P(\text{donnees plus extremes } |H_0)$$

- Dans une démarche scientifique, on émet des hypothèses. Ce qui nous intéresse, c'est de savoir quelle est la probabilite que notre hypothese soit vraie sachant les donnees

$$P(H|\text{données}) = \frac{P(\text{données}|H)P(H)}{P(\text{données})}$$

- On calcule les ratios de probabilités (odd-ratios) entre les différentes hypothèses

$$\frac{P(H_1|\text{données})}{P(H_2|\text{données})}$$

Exemple Tests d'hypothèse Bayesiens

- Un médecin recoit un patient qui tousse.
- Il envisage trois hypothèse pour la maladie: gastro, cancer, covid.
- Au vu de la situationa actuelle, il a les informations suivantes : $P(\text{cancer}) = 0.05\%$, $P(\text{gastro}) = 20\%$, $P(\text{covid}) = 50\%$.
- De son expérience de médecin, il a les probabilités suivantes: $P(\text{toux}|\text{cancer}) = 95\%$, $P(\text{toux}|\text{gastro}) = 0.3\%$, $P(\text{toux}|\text{covid}) = 89\%$.

Faire un test d'hypothèse Bayesien pour évaluer le ratio entre les différentes probas.

Exemple Tests d'hypothèse Bayesiens

$$P(H|\text{données}) = \frac{P(\text{données}|H)P(H)}{P(\text{données})}$$

- Hypothèses: 1. gastro, 2. cancer, 3. covid.
- Donnée: toux
- $P(\text{données}|H)$: $P(\text{toux}|\text{cancer}) = 95\%$, $P(\text{toux}|\text{gastro}) = 0.3\%$, $P(\text{toux}|\text{covid}) = 89\%$
- $1/P(\text{données}) = k$.
- $P(H)$: $P(\text{cancer}) = 0.05\%$, $P(\text{gastro}) = 20\%$, $P(\text{covid}) = 50\%$.
- $P(\text{cancer}|\text{toux}) = k0.05 \times 0.95 = k0.000475$
- $P(\text{gastro}|\text{toux}) = k0.003 \times 0.2 = k0.0006$
- $P(\text{covid}|\text{toux}) = k0.5 \times 0.89 = k0.445$

Exemple Tests d'hypothèse Bayesiens

On fait le rapport entre les probabilités (odd-ratio):

- Il y a 936 fois plus de chances que le patient ait le covid que le cancer:

$$\frac{P(\text{covid}|\text{toux})}{P(\text{cancer}|\text{toux})} = 936$$

- Il y a 741 fois plus de chances que le patient ait le covid que la gastro:

$$\frac{P(\text{covid}|\text{toux})}{P(\text{gastro}|\text{toux})} = 741$$

Exemple Tests d'hypothèse Bayesiens

Remarques:

- Dans la démarche scientifique, on émet des hypothèses. Ce qui nous intéresse, c'est de savoir quelle est la probabilité que notre hypothèse soit vraie sachant les données, precisement $P(H|\text{données})$.
- Ces test sont Bayesiens dans la mesure ou ils utilsent la regle de Bayes

$$P(H|\text{données}) = \frac{P(\text{données}|H)P(H)}{P(\text{données})}$$

- Cela nécessite d'avoir P(H), c.a.d une information a priori (prior).
- On peut tester plusieurs hypothèses.
- Ces test sont donc plus précis car ils prennent en compte plus d'information, mais beaucoup moins "généraux" car la structure de l'information a priori dépend du problème.

Régression linéaire univarieee

- La régression linéaire univariee vise à prédire une variable "expliquée" y comme fonction lineaire (ou affine) d'une variable explicative x.
- La régression affine unidimensionnelle prédit la variable (aléatoire) Y avec une fonction affine de la variable (aléatoire) X

$$Y = aX + b$$

- Pour déterminer les paramètres du modèle linéaire (par ex. a et b), il faut choisir une méthode d'estimation. La plus courante est la méthode des moindre carrés:

$$a, b = \underset{\alpha,\beta}{\operatorname{argmin}}\ \mathbb{E}\left([Y - (\alpha X + \beta)]^2\right)$$

Il existe d'autre méthodes comme le modèle par maximum de vraisemblance ou l'inférence bayésienne.

Minimum d'une fonction réelle de deux variables

La fonction

$$\begin{aligned} f :& \mathbb{R}^2 \to \mathbb{R} \\ &(\alpha, \beta) \mapsto \mathbb{E}\left([Y - (\alpha X + \beta)]^2\right) \end{aligned}$$

est une fonction continue dérivable.

En effet

$$\frac{\partial}{\partial \alpha}\mathbb{E}\left([Y - (\alpha X + \beta)]^2\right) = \mathbb{E}\left(\frac{\partial}{\partial \alpha}[Y - (\alpha X + \beta)]^2\right)$$

- par linéarité de l'espérance
- par convergence dominée pour le passage a la limite

Minimum d'une fonction réelle de deux variables

Propriété: Si $f : (x, y) \in \mathbb{R}^2 \mapsto f(x, y) \in \mathbb{R}$ une fonction dérivable admet un minimum en $m = (m_1, m_2)$ alors son gradient s'annule en m

$$\nabla_m f = \left(\frac{\partial f}{\partial x}(m_1, m_2), \frac{\partial f}{\partial y}(m_1, m_2) \right) = (0, 0)$$

Preuve:
Si (m_1, m_2) est minimum de f alors m_1 est minimum de $f_{.,m_2}(x) = f(x, m_2)$ donc la dérivée de $f_{.,m_2}$ s'annule en m_1 et

$$f'_{.,m_2}(m_1) = \frac{\partial f}{\partial x}(m_1, m_2)$$

Régression linéaire aux moindres carrés

On cherche a trouver le minimum de

$$f(\alpha, \beta) = \mathbb{E}\left([Y - (\alpha X + \beta)]^2\right)$$

Une condition nécessaire est l'annulation du gradient (i.e. des dérivées partielles).

- Quelles sont les dérivées partielles de f?
- Que nous donne la condition d'annulation du gradient ?
- Reconnaissez vous un terme connu dans l'expression de α ?

Aide: Dérivée de $\alpha \mapsto (K\alpha + L)^2$: $\quad \alpha \mapsto 2K^2\alpha + 2KL$.

Régression linéaire aux moindres carrés

Dérivées partielles:

$$
\begin{aligned}
\frac{\partial f}{\partial \alpha} &= \mathbb{E}\left(2X^2\alpha - 2X(Y+\beta)\right) = 2\alpha\mathbb{E}(X^2) - 2\mathbb{E}(XY) - 2\beta\mathbb{E}(X) \\
\frac{\partial f}{\partial \beta} &= \mathbb{E}\left(2\beta + 2(Y - \alpha X)\right) = 2\beta - 2\mathbb{E}(Y) + 2\alpha\mathbb{E}(X)
\end{aligned}
$$

Régression linéaire aux moindres carrés

On a

$$\frac{\partial f}{\partial \alpha} = 2\alpha\mathbb{E}(X^2) - 2\mathbb{E}(XY) - 2\beta\mathbb{E}(X)$$
$$\frac{\partial f}{\partial \beta} = 2\beta - 2\mathbb{E}(Y) + 2\alpha\mathbb{E}(X)$$

Soit (a, b) minimum de f. On a nécessairement

$$b = \mathbb{E}(Y) - a\mathbb{E}(X)$$
$$a = \frac{\mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)}{\mathbb{E}(X^2) - E(X)^2} = \frac{cov(X, Y)}{var(X)} = corr(X, Y)\frac{\sigma_Y}{\sigma_X}$$

avec la corrélation de X et Y compris entre -1 et $+1$

$$corr(X, Y) = \frac{cov(X, Y)}{\sigma_X \sigma_Y}$$

Régression linéaire aux moindres carrés

Remarques:

- $b = \mathbb{E}(Y) - a\mathbb{E}(X)$ est un "correcteur de moyenne". Il ajoute la moyenne de Y et enlève a fois la moyenne de X qui vient du terme aX.
- a correspond de la corrélation entre X et Y avec un rescaling de l'écart type
- L'erreur quadratique moyenne vaut

$$\mathbb{E}\left((Y - (aX + b))^2\right) = \sigma_Y^2(1 - corr(X, Y)^2)$$

Si la corrélation est proche de 1 ou -1, l'erreur est quasiment nulle, tandis que la corrélation est proche de 0 l'erreur est maximale, a est proche de zéro et l'estimation est réduite à la moyenne de Y.

Régression linéaire aux moindres carrés

Exercice python:
X v.a. uniforme sur [0, 1] et $Y = X + \epsilon U$ ou U v.a. uniforme sur [0,1] indépendante de X .

1. Calculer $\mathbb{E}(X)$ et σ_X
2. Calculer $\mathbb{E}(Y)$ et σ_Y en fonction de ϵ.
3. Calculer $Cov(X, Y)$, $Corr(X, Y)$ et l'erreur quadratique moyenne théorique en fonction de ϵ.
4. Choisir $\epsilon > 0$ et générer avec np.random.random() 1000 tirages de (X, Y). Estimer a, b et l'erreur quadratique par méthode de Monte-Carlo à partir des données. Comparer aux valeurs théoriques.

Régression linéaire aux moindres carrés

Exercice python:
X v.a. uniforme sur [0, 1] et $Y = X + \epsilon U$ ou U v.a. uniforme sur [0,1] indépendante de X .

1. $\mathbb{E}(X) = 0.5$ et $\sigma_X = 1/\sqrt{12}$
2. $\mathbb{E}(Y) = 0.5(1+\epsilon)$ et $\sigma_Y = \sqrt{(1+\epsilon^2)/12}$.
3. $Cov(X, Y) = Var(X)$, $Corr(X, Y) = \sigma_X/\sigma_Y = 1/\sqrt{1+\epsilon^2)}$ et l'erreur quadratique moyenne théorique en fonction de ϵ.
4. MSE $= \frac{\epsilon^2}{12}$.

Régression linéaire multivariée

La régression linéaire multivariée étend la régression linéaire simple pour décrire une variable y comme combinaison linéaire (affine) de plusieurs variables explicatives $x_1, \dots, x_D$, ou bien fonction linéaire (affine) du vecteur $x = (x_1, \dots, x_D)$:

$$y = \sum_{d=1}^{D} a_d x_d + b = \langle a, x \rangle_{\mathbb{R}^d} + b$$

Pour déterminer les paramètres du modèle linéaire (par ex. a et b) a partir d'un échantillon de données $(x^n, y^n)_{n=1\dots N}$, on choisit une méthode d'estimation, par exemple la méthode des moindres carres:

$$a, b = \operatorname*{argmin}_{\alpha,\beta} \frac{1}{N} \sum_{n=1}^{N} \Big(y^n - (\langle \alpha, x^n \rangle + \beta) \Big)^2$$

Régression linéaire multivariée

Remarques:

- Dans le cas de la régression univariee, on avait écrit la méthode des moindre carrés comme :

$$a, b = \underset{\alpha,\beta}{\operatorname{argmin}}\ \mathbb{E}\left(\left[Y - (\alpha X + \beta)\right]^2\right)$$

- Dans le cas multi-dimensionel, on a écrit la méthode des moindres carres comme

$$a, b = \underset{\alpha,\beta}{\operatorname{argmin}}\ \frac{1}{N}\sum_{n=1}^{N}\Big(y^n - (\langle \alpha, x^n\rangle + \beta)\Big)^2$$

- $\frac{\sum_{n=1}^{N} Z_n}{N}$ est l'estimation par Monte-Carlo de $\mathbb{E}(Z)$ **si les Z_i sont i.i.d.**

.

Régression linéaire multivariée

On a un échantillon de données $(x^n, y^n)_{n=1\dots N}$, et on veut estimer a et b. On considère la norme usuelle sur $\mathbb{R}^N$, les les matrices et vecteurs

$$y = \begin{pmatrix} y^1 \\ \vdots \\ y^N \end{pmatrix}, X = \begin{pmatrix} 1 & x_1^1 & \cdots & x_D^1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_1^N & \cdots & x_D^N \end{pmatrix}, a = \begin{pmatrix} b \\ a_1 \\ \vdots \\ a_D \end{pmatrix},$$

- Comment s'écrit la linéaire relation entre les y^n et les x^n ?
- Comment s'écrit l'erreur que l'on minimise

$$\frac{1}{N} \sum_{n=1}^{N} \Big(y^n - (\langle a, x^n \rangle + b) \Big)^2 \quad ?$$

Régression linéaire multivariée

$$y = \begin{pmatrix} y^1 \\ \vdots \\ y^N \end{pmatrix}, X = \begin{pmatrix} 1 & x_1^1 & \cdots & x_D^1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_1^N & \cdots & x_D^N \end{pmatrix}, a = \begin{pmatrix} b \\ a_1 \\ \vdots \\ a_D \end{pmatrix},$$

- Relation entre les y^n et les x^n:

$$y = Xa$$

- L'erreur que l'on minimise

$$\frac{1}{N}\|y - Xa\|^2_{2,\mathbb{R}^{\mathbb{N}}} = \frac{1}{N}\langle y - Xa, y - Xa\rangle_{\mathbb{R}^{\mathbb{N}}}$$

Gradient d'une fonction

Soit $f : \mathbb{R}^D \mapsto \mathbb{R}$ derivable.

Définition dans un système de coordonnées:

- Dans un système de coordonnées (cartésiennes, le gradient d'une fonction $f(x_1, x_2, \ldots, x_D)$ est le vecteur des dérivées partielles f par rapport aux coordonnées:

$$\nabla_{(x_1,\ldots,x_D)} f = \begin{bmatrix} \frac{\partial f}{\partial x_1}(x_1, \ldots, x_D) \\ \vdots \\ \frac{\partial f}{\partial x_D}(x_1, \ldots, x_D) \end{bmatrix}$$

- Il pointe dans la direction où la fonction croît le plus rapidement, et sa norme correspond au taux de croissance dans cette direction.
- C'est la généralisation à plusieurs variables de la dérivée d'une fonction d'une seule variable.

Gradient d'une fonction

Soit $f : \mathbb{R}^D \mapsto \mathbb{R}$ derivable.

Définition générale:
Comme f est differentiable Pour une variation h autour de x, la variation de f est au premier ordre une fonction linéaire de h:

$$f(x+h) = f(x) + L(h) + \underset{\|h\| \to 0}{o(h)}$$

Le gradient est le représentant de cette forme linéaire tangente:

$$f(x+h) = f(x) + \nabla f \cdot h + \underset{\|h\| \to 0}{o(h)}$$

Gradient d'une fonction

Soit $f : \mathbb{R}^N \mapsto \mathbb{R}$ derivable.

Propriété:
Si f admet un minimum en m alors son gradient s'annule en m

$$\nabla_m f = 0_{\mathbb{R}^N}$$

- Calculer le gradient de

$$f(a) = \frac{1}{N} \|y - Xa\|^2_{2,\mathbb{R}^{\mathbb{N}}} = \frac{1}{N} \langle y - Xa, y - Xa \rangle_{\mathbb{R}^{\mathbb{N}}}$$

- Que donne la condition d'annulation du gradient ?

Régression linéaire multivariee

- Gradient de f:

$$\nabla f = \frac{2}{N}\left(X^T X a - X^T y\right)$$

- Condition d'annulation du gradient":

$$a = \left(X^T X\right)^{-1} X^T y$$

Notebook python sur les régressions linéaires multivariees.

Régression linéaire régularisée

Problème: on a supppose que $X^T X$ est inversible, mais si elle ne l'est pas... En particulier si N ¡ D, rg$(X^T X) = N$, la matrice n'est pas inversible.
On pose

$$f(a) = \left\|y - Xa\right\|^2 + \epsilon\|a\|^2$$

- Gradient de f ?
- Condition d'annulation du gradient" ?

Régression linéaire régularisée

On pose

$$f(a) = \left\|y - Xa\right\|^2 + \epsilon\|a\|^2$$

- Gradient de f ?

$$\nabla f = 2\left(X^T Xa - X^T y\right) + 2\epsilon a$$

- Condition d'annulation du gradient ?

$$a = \left(X^T X + \epsilon I_N\right)^{-1} X^T y$$

La matrice $X^T X + \epsilon I_N$ est toujours inversible.

Régression linéaire régularisée

Supposons que l'on fasse une régressions affine avec deux données et deux variables explicatives:

$$a_1 x_1^1 + a_2 x_2^1 + b = y^1$$
$$a_1 x_1^2 + a_2 x_2^2 + b = y^2$$
$$a_1 x_1^3 + a_2 x_2^3 + b = y^3$$

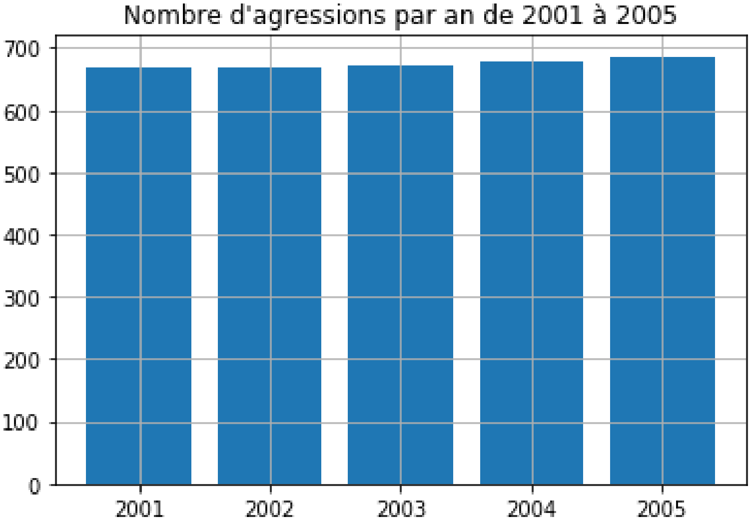
- Que peut-on dire de ce système ?
- Que vaudra l'erreur quadratique ?
- Cela vous parait raisonnable ?
- Qu'est ce qui pourrait résoudre ce problème ?

Statistiques descriptives

La statistique descriptive est la branche des statistiques qui cherche à décrire et à résumer un ensemble important de données.

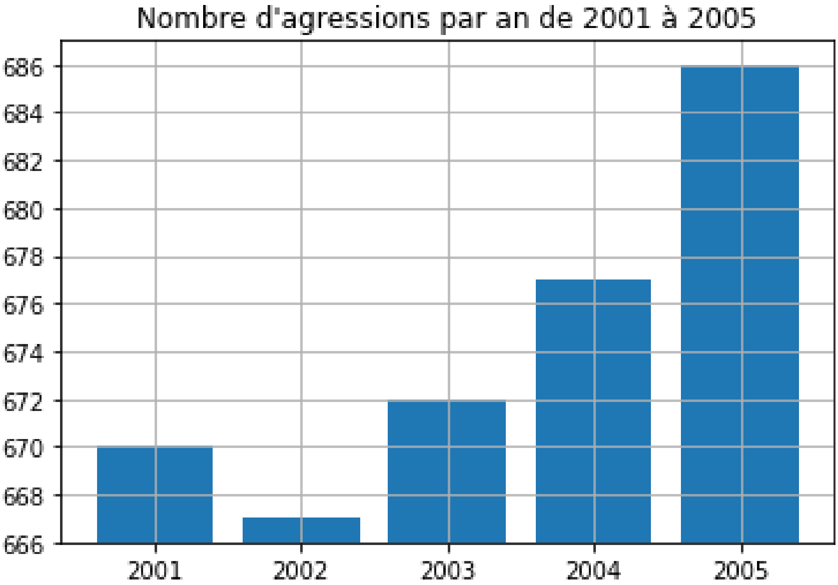
- *Les faits sont têtus. Il est plus facile de s'arranger avec les statistiques.* Mark Twain.
- *Faites attention, la statistique est toujours la troisième forme du mensonge.* Jacques Chirac.
- *La mort d'un homme est une tragédie. La mort d'un million d'hommes est une statistique.* Joseph Staline.
- *Les hommes politiques ne connaissent la misère que par les statistiques. On ne pleure pas devant les chiffres.* L'Abbe Pierre.

Graphiques



Nombre d'agressions par an de 2001 à 2005
700
600
500
400
300
200
100
0
2001
2002
2003
2004
2005

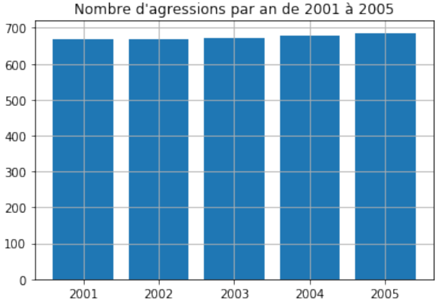
Graphiques

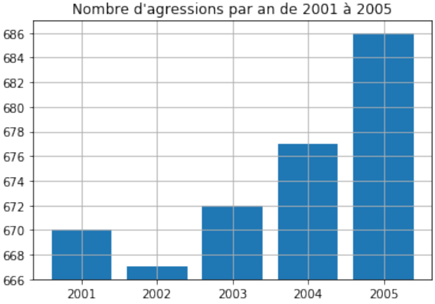

Nombre d'agressions par an de 2001 à 2005
686
684
682
680
678
676
674
672
670
668
666
2001
2002
2003
2004
2005

Graphiques

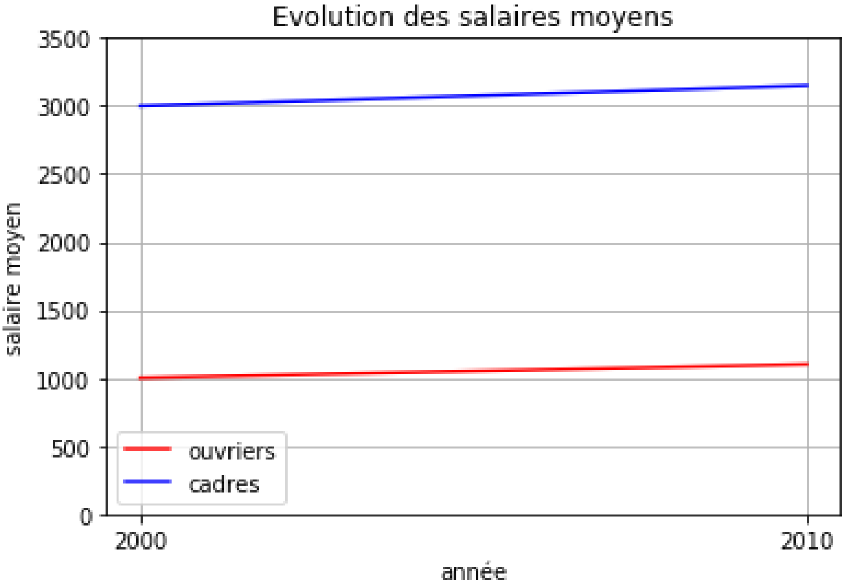
année	2001	2002	2003	2004	2005
agressions	670	667	672	677	686

Attention à l'axe !



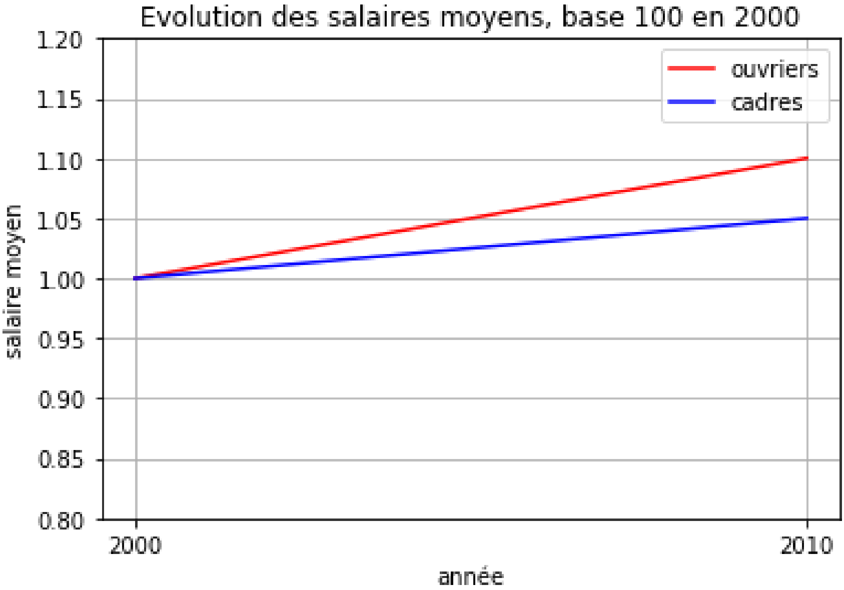


Graphiques



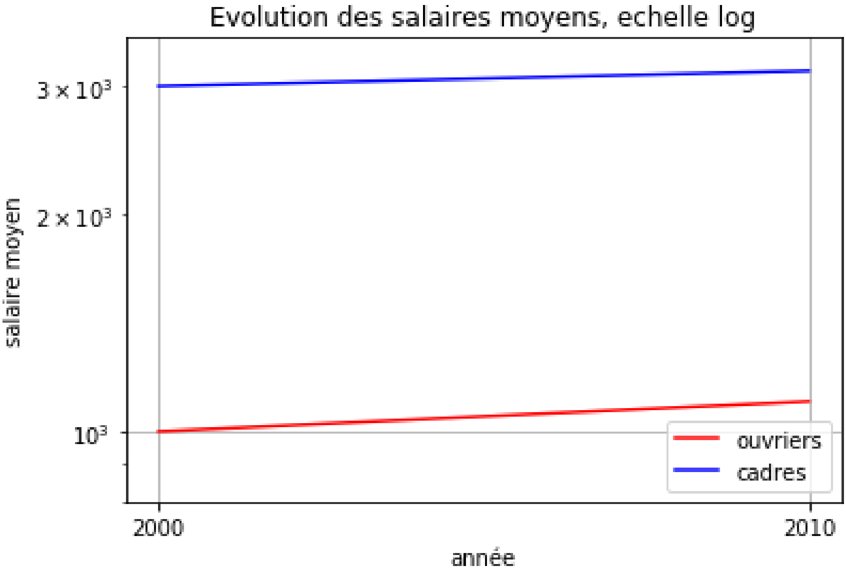
Evolution des salaires moyens
3500
3000
2500
2000
1500
1000
500
0
salaire moyen
2000
2010
année
ouvriers
cadres

Graphiques



Evolution des salaires moyens, base 100 en 2000
ouvriers
cadres
1.20
1.15
1.10
1.05
1.00
0.95
0.90
0.85
0.80
salaire moyen
2000
2010
année

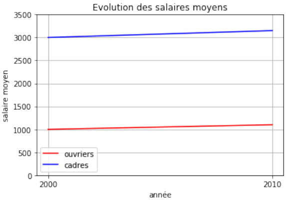
Graphiques

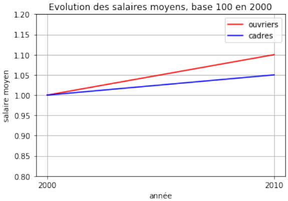

Evolution des salaires moyens, echelle log
salaire moyen
3×10^3
2×10^3
10^3
2000
2010
année
ouvriers
cadres

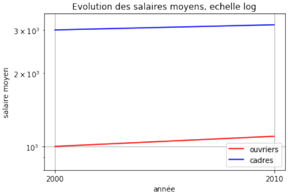
Graphiques

année	2000	2010
salaire moyen ouvriers	1000	1100
salaire moyen cadres	3000	3150

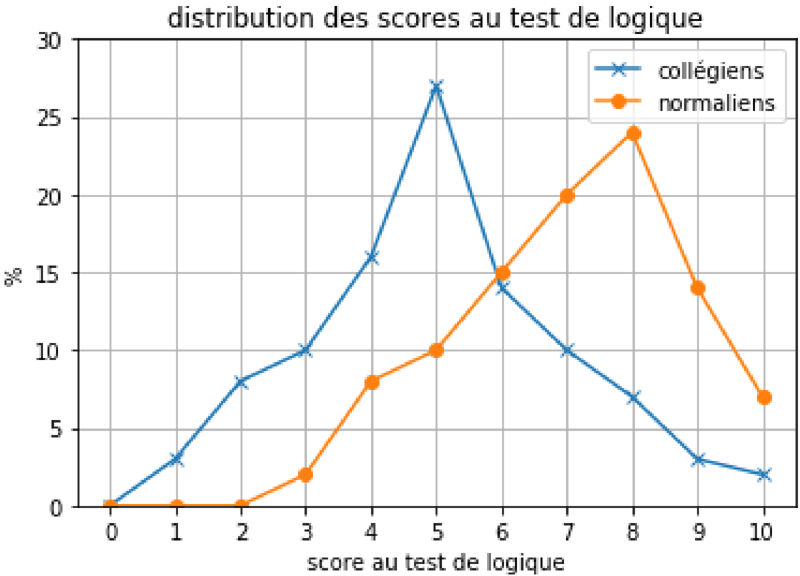
Attention aux choix d'échelles et de références!



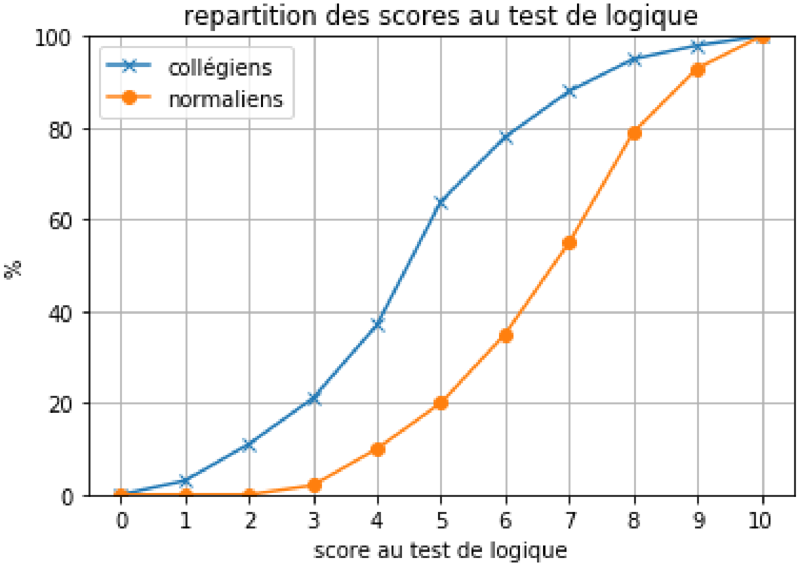




Graphiques


distribution des scores au test de logique
collégiens
normaliens
30
25
20
15
10
5
0
%
0
1
2
3
4
5
6
7
8
9
10
score au test de logique

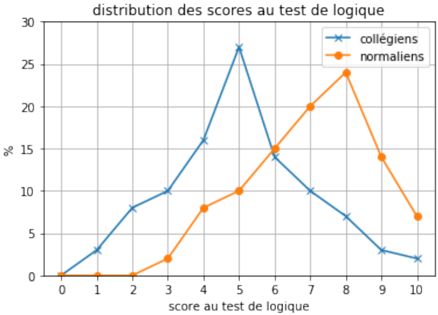
Graphiques

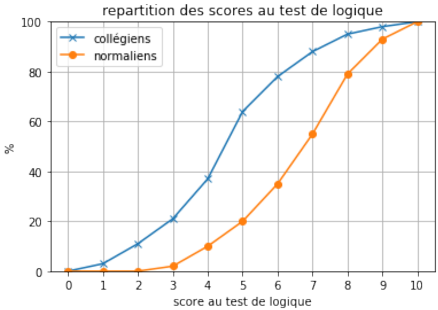


repartition des scores au test de logique
collégiens
normaliens
%
0
20
40
60
80
100
0
1
2
3
4
5
6
7
8
9
10
score au test de logique

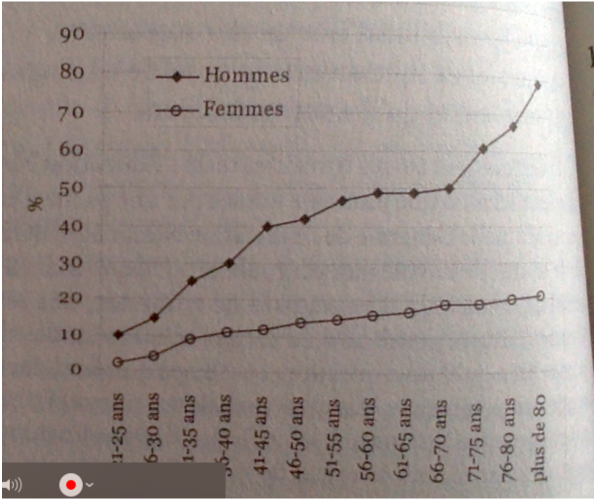
Graphiques

Distribution $\neq$ répartition !



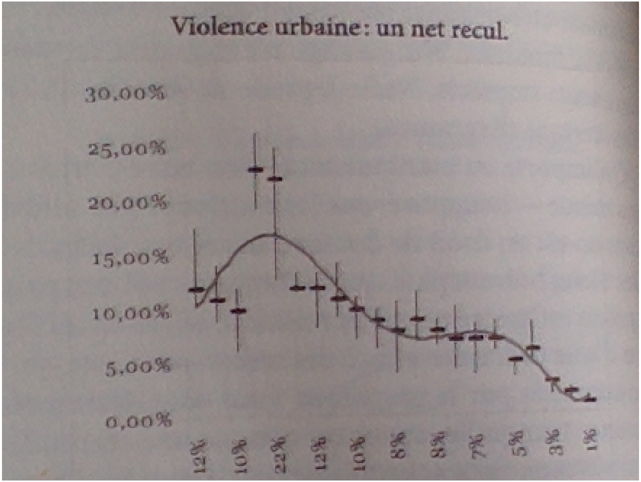


Graphiques



Tiré de *Statistiques, méfiez-vous!*, N. Gauvrit, Ed. Ellipses.

Graphiques



Tiré de *Statistiques, méfiez-vous!*, N. Gauvrit, Ed. Ellipses.

Pourcentages

- *La dette de la France qui avait augmenté de 10 % l'an passé n'a augmenté que de 9.5 % cette année*, se félicite le gouvernement.
- *La gestion du gouvernement, qui avait déjà du emprunter 200 milliards d'euros l'an dernier, est tellement catastrophique qu'il faudra emprunter cette année 209 milliard d'euros sur les marchés financiers pour payer les dépenses*, déplore l'opposition.
- En combinant les deux informations, quelle était la dette il y a deux ans ? Est-ce plausible ?
- Dette de 2000 milliards, puis 2200 milliards puis 2409 milliards. Déficit de 10 % puis 9.5 %.

Pourcentages

Le paradoxe de Simpson.

- Une université américaine a fait de la lutte contre les discriminations un objectif phare.
- Apprenant que dans la section littérature,sur une promotion contenant 15 garçons et 15 filles, 6 filles ont réussi l'examen de fin de licence contre 9 garçons. Ils convoquent les responsable pour leur faire part de leurs soupçons.
- Surprise des professeurs, qui chiffrent a l'appui, montrent que les garçons réussissent moins bien que les filles !
- Il y a deux groupes : lettre modernes et lettres classiques. Construire un tel paradoxe :

	Filles		Garçons	
	réussite	échecs	réussite	échecs
lettres modernes				
lettres classiques				

Pourcentages

Le paradoxe de Simpson.

- Explication:

	Filles		Garçons		
	réussite	échecs	réussite	échecs	total
lettres modernes	1	9	0	5	15
lettres classiques	5	0	9	1	15
total	6	9	9	6	15

- Pourcentage de réussite:

	Filles	Garçons
lettres modernes	11%	0 %
lettres classiques	100 %	90%

	Filles	Garçons
lettres	40%	60%

Statistiques: erreurs, choix, faux et cadre.

- Erreur de données: exemple de la mortalité infantile en Espagne.
- Choix dans les statistiques: classement de l'immobilier comme dépense d'investissement.
- Faux: Lancet Gate 1, 2.

Cadrage: En psychologie du raisonnement et de la décision ainsi qu'en psychologie sociale, le cadrage est l'action de présenter un *cadre cognitif* comme approprié pour réfléchir sur un sujet. Ce cadrage peut avoir un effet sur le raisonnement et conduire à des choix différents en fonction de la façon dont le problème a été formulé (Wikipédia).

Statistiques, méfiez-vous.

- Les données sont factuelles.
- Les statistiques les rendent lisibles.

Mais...

- D'où viennent les données ? Qui les a collectées et comment sont-elles stockées ?
- Qui produit les statistiques? Dans quel but? Ces personnes ont-elles des conflits d'intérêts sur ces sujets (*There is no free lunch.*) ? Ces personnes sont-elles sujettes a des biais ?
- Qui présente les statistiques ? A qui ? Et dans quel but? Ces personnes ont-elles des conflits d'intérêts sur ces sujets ? Ces personnes sont-elles sujettes a des biais ?

Conseils.

- Connaissez les notions du cours, et sachez les expliquer simplement.
- Suivez votre intuition, cherchez a la vérifier avec des petites expériences numériques simples. Reconnaissez quand votre intuition est bonne, ou quand elle ne l'est pas bonne.
- Pratiquez la discussion ou le débat contradictoire.