Probastats notes et tricks

Produit de deux variables uniformes entre 0 et 1

Problème 76. On veut calculer :

$$P(AC \leq x) = \int_0^1 \int_0^{\frac{x}{a}} f_A(a) f_C(c) dcda$$

naïvement on peut se dire qu'on peut remplacer les deux par 1 comme elles sont uniformes mais non ! parce que quand a est tout petit $\frac{x}{a}$ est bien plus grand que 1 donc $f_C(c) = 0$! On doit faire une disjonction de cas :

$$P(AC \leq x) = \int_0^x \int_0^1 1 dcda + \int_x^1 \int_0^{\frac{x}{a}} 1 dcda = x - x\ln(x)$$

https://math.stackexchange.com/questions/659254/product-distribution-of-two-uniform-distribution-what-about-3-or-more

Calculer l'espérance comme l'intégrale de 1 - cumulative

voir https://math.stackexchange.com/questions/2042896/proving-that-ex-sum-k-0-infty-pxk-by-proving-n1pxn-xrighta

$$E(X) = \int P(X \geq x) dx$$

Les différences entre les types de convergence

convergence in probability $\not\Rightarrow$ convergence in mean square

On prend X_n telle que la proba d'être zéro est forte et la proba d'être très grand est faible et $X = 0$.

- $\mathbb{P}(X_n = 0) = 1 - \frac{1}{n}$
- $\mathbb{P}(X_n = n) = \frac{1}{n}$

On a $\mathbb{P}(|X_n - X| > \varepsilon) = \mathbb{P}(|X_n - 0| > \varepsilon) = \frac{1}{n}$ qui tend vers 0 à l'infini donc X_n converge en probabilité vers 0.
mais $E\left((X_n - X)^2\right) = E\left(X_n^2\right) = \frac{1}{n} \cdot n^2 = n$ donc X_n ne converge pas en moyenne quadratique.

convergence in distribution $\not\Rightarrow$ convergence in probability

On prend X_0 choisi uniformément entre 0 et 1 et $X_{2n} = X_0$ (donc X_{2n} est constant). On prend $X_{2n+1} = 1 - X_0$.

X_{2n} suit donc une distribution uniforme $\sim U[0,1]$ et $X_{2n+1} \sim U[0,1]$ aussi (montrons-le avec la cumulative, si on veut la probabilité que $X_{2n+1} \leq\ 0.2$ on veut la probabilité que $1 - X_0 \leq 0.2$ on veut $X_0 \geq\ 0.8$, or X_0 est uniforme donc $\mathbb{P}\left(X_0 \leq\ 0.2\right) = \mathbb{P}\left(X_0 \geq\ 0.8\right)$). Donc X_n converge en distribution vers X_0.

Par contre, X_n ne converge pas en probabilité :

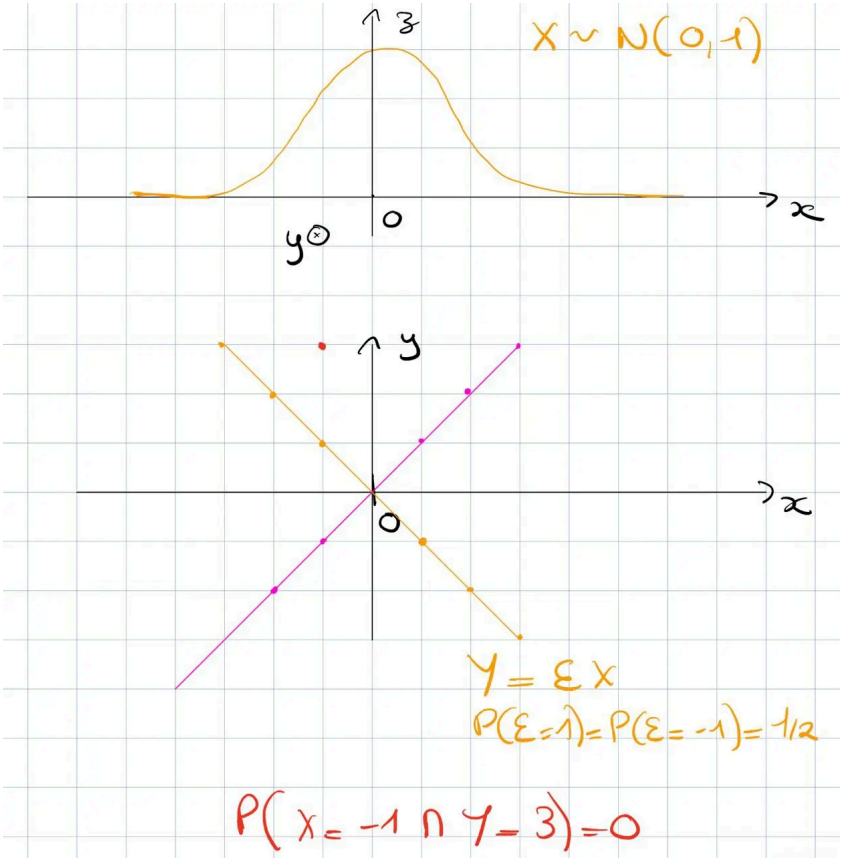
$$\mathbb{P}\left(|X_n - X_0|\ > \varepsilon\right) = \mathbb{P}\left(|X_{2n} - X_0|\ > \varepsilon \wedge\ |X_{2n+1} - X|\ > \varepsilon\right)$$

$$= \mathbb{P}\left(|X_0 - X_0|\ > \varepsilon \wedge\ |1 - X_0 - X_0|\ > \varepsilon\right) = 0 + q$$

$q \neq 0$ car il y a des exemples où $1 - 2X_0 \neq 0$ (par exemple si $X_0 =\ 0.2$).
Donc X_n ne converge pas en probabilité.

Pourquoi deux normales ne sont pas toujours jointly Gaussian ?

exemple de deux normales qui ne sont pas jointly Gaussian :



dans le deuxième schéma vu de z, Y sera soit la ligne rose soit la ligne orange, et les points qui seront en dehors de ces lignes seront tous zéro. donc ça ne sera pas une gaussienne vu de tous les côtés (et c'est la définition de jointly Gaussian)

Comment prouver qu'un vecteur X de normales n'est pas jointly Gaussian ?

Il faut trouver un vecteur u tel que $u^T X$ ne suit pas une distribution normale. On peut poser $u = (a, b)$ si on ne sait pas quel vecteur ne fonctionne pas.
On sait que la moment generating function doit être de la forme $M_X(u) = \exp\left(u^T \mu + \frac{1}{2} u^T \Omega u\right)$. On calcule la joint moment generating function $E\left(\exp(aX + bY)\right)$.

Transformations

Il y a déjà la formule dans la Cheat Sheet mais c'est bien de savoir pourquoi on le fait comme ça. On a la P.D.F $f(x)$ et on veut la cumulative $G(y)$, avec $Y = \frac{1}{X}$.

Méthode :
D'abord on définit nos fonctions pour passer de x à y :

$$r(x) = \frac{1}{x} \text{ et } s(y) = \frac{1}{y}$$

$$G(y) = P(Y \leq y) = P\left(\frac{1}{X} \leq y\right) = P\left(X \geq \frac{1}{y}\right) = 1 - P\left(X < \frac{1}{y}\right)$$

$$\implies G(y) = 1 - F\left(\frac{1}{y}\right)$$

$$\implies \frac{dG(y)}{dy} = \frac{d\left(1 - F\left(\frac{1}{y}\right)\right)}{dy}$$

$$\implies g(y) = -\frac{dF}{dy}\left(\frac{1}{y}\right) \cdot |-\frac{1}{y^2}| \text{ (on s'intéresse à la croissance, on enlève le signe -)}$$

$$\implies g(y) = -f\left(\frac{1}{y}\right) \cdot \frac{1}{y^2}$$

Et ensuite pour trouver $G(y)$ on intègre.

Trouver une matrice A telle que $A\Omega A^T = I$

On a $X, Y \sim \mathcal{N}_2\left(\begin{pmatrix}1\\0\end{pmatrix}, \begin{pmatrix}5 & 4\\4 & 5\end{pmatrix}\right)$ et on cherche $A \in \mathbb{R}^{2\times 2}, b \in \mathbb{R}^2$ tels que X' et Y' sont des normales standardisées indépendantes.

$$\begin{pmatrix}X'\\Y'\end{pmatrix} = A\begin{pmatrix}X\\Y\end{pmatrix} + b$$

On sait que comme Ω est symétrique, on peut l'écrire $\Omega = CDC^T$ où D contient les valeurs propres de Ω et C_1, C_2 sont les vecteurs propres correspondants.

Si v est un vecteur propre on a $Av = \lambda v$.

Trouver les valeurs propres : on pose $\det(\Omega - \lambda I) = 0$ ou :

- on remarque que la somme S des lignes de la matrice est constante ($\lambda_1 = S$)
- on remarque que le déterminant de la matrice $\det$ est $\lambda_1 \cdot \lambda_2$

- on remarque que la trace de la matrice (la somme des entrées diagonales) est $\lambda_1 + \lambda_2$

Trouver les vecteurs propres : chercher le noyau de $(A - \lambda I)v = 0$
Par exemple pour $\lambda = 1$:

$$\begin{pmatrix} 4 & 4 & : 0 \\ 4 & 4 & : 0 \end{pmatrix}$$

On trouve le vecteur propre $v_1 = (-1, 1)$ puis normalisé : $u_1 = \left(-\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}\right)$.

On a donc trouvé $\Omega = CDC^T$, et on sait réécrire $D = \sqrt{D}I\sqrt{D}$ donc :
$\Omega = C\sqrt{D}I\sqrt{D}C^T = C\sqrt{D}I\left(C\sqrt{D}\right)^T$.
On a donc $I = \left(C\sqrt{D}\right)^{-1}\Omega\left(\left(C\sqrt{D}\right)^T\right)^{-1} = \left(C\sqrt{D}\right)^{-1}\Omega\left(\left(C\sqrt{D}\right)^{-1}\right)^T$.
Donc $A = \left(C\sqrt{D}\right)^{-1}$.

Moments d'une normale standardisée

Si $X \sim N(0,1)$ alors $E\left(X^n\right) = \begin{cases} 0 \text{ si } n = 2k+1 \\ (k-1)!! \text{ si } n = 2k \end{cases}$
avec la double factorielle égale au produit des entiers impairs inférieurs ou égaux à $k-1$.

Exercices

note par intérêt

- 38 distribution hypergéométrique *1/5*
- 39 vraiment très cool cet exercice, bien penser à la distribution géométrique, se fait avec de l'analyse sans regarder la correction *4/5*
- 65 intéressant raisonnement covariance mais un peu trop compliqué *3/5*
- 66 appliquer les formules de variance *1/5*
- 67 appliquer les formules de variance *3/5*
- 68 revoir la formule de correlation *2/5*
- 69 résoudre systèmes correlation/covariance (penser au fait qu'on peut obtenir la covariance avec $\mathrm{var}(X+Y)$ mais aussi directement en utilisant la linéarité de la covariance) *2/5*
- 70 bien savoir poser la density function, avec des variables discrètes *4/5*
- 71 comme le 71 *2/5*
- 72 utiliser le trick de la MGF indépendante *3/5*
- 73 résoudre intégrale convolution + trick MGF indépendantes *3/5*
- 74 MGF d'une variable discrète *4/5*
- 75 le binôme de newton *4/5*
- 76 comparer les MGF de deux variables aléatoires pour savoir si elles suivent la même distribution *5/5*

- 97 apprendre à dériver likelihood avec le log etc *4/5*
- 100 arriver à voir des binomiales là où il faut les voir *3.75/5*
- 101 utiliser la bonne formule de la M.S.E. avec la variance ou savoir $E(\overline{X})$ *5/5*

Problèmes

- 82 dur pas très intéressant bien noter qu'on demande le temps total ET le temps passé au post office pas juste le temps total *2/5*
- 85 ne pas oublier que $\frac{\lambda^n}{n!}$ c'est exponentiel λ ! *5/5*

Examen 2019

Exo 2 :

- $\text{cov}\left(aX_1 + bX_2, bX_3\right) = ab\text{cov}\left(X_1, X_3\right) + b^2\text{cov}\left(X_2, X_3\right)$ et pour deux variables jointly Gaussian, on a $\text{cov}\left(X_1, X_2\right) = 0 \iff X_1, X_2 \text{ indep.}$
  $\text{cov}(X, X) = \text{var}(X)$.
- $X_+ = X_1 + X_2$ et $X_- = X_1 - X_2$. Les deux peuvent être dépendants ou indépendants. Par exemple si X_i est le résultat d'un pile ou face, les deux sont indépendants. Si X_i est une binomiale de paramètre $p = \frac{1}{2}$, les deux sont dépendants (si on sait que $X_1 = 1$ alors $X_2 = 0$ nécessairement).
- si elles sont toutes pairwise, la matrice Ω est diagonale donc elles sont mutuellement indépendantes
- on peut avoir 3 variables avec toutes pairwise indépendantes mais ensemble dépendantes. Par exemple G_B : il y a exactement un garçon, G_1 le premier est un fils, G_2 le deuxième est un fils. On a $P\left(G_B \wedge G_1\right) = P\left(G_B\right) \cdot P\left(G_1\right)$ et $P\left(G_B \wedge G_2\right) = P\left(G_B\right) \cdot P\left(G_2\right)$ et $P\left(G_B \wedge G_1 \wedge G_2\right) = 0$.

Exo 3 :

- ne pas oublier que quand on a la MGF d'une somme de I.I.D., on peut la séparer

Analyse

Exponentielle :

$$e^x = \sum_{n=0}^{\infty} \frac{x^n}{n!}$$

Valeur d'une série géométrique :

$$\sum_{n=0}^{\infty} q^n = \frac{1}{1-q} \text{ si } |q| < 1$$

Se souvenir des dimensions :

Si on a une fonction de densité simple $f_X(x)$, quand on l'intègre on trouve une probabilité (p. ex. la cumulative). Si on a une fonction de densité jointe $f_{X,Y}(x,y)$ quand on l'intègre on trouve une autre densité (p. ex $f_X(x)$).

Les propriétés de base

Variance

$$\text{var}(X+Y) = \text{var}(X) + \text{var}(Y) + 2\,\text{cov}\,(X,Y)$$

avec des coefficients :

$$\text{var}(aX+bY) = a^2\text{var}(X) + b^2\text{var}(Y) + 2\cdot a\cdot b\cdot \text{cov}(X,Y)$$

Covariance

$$\text{cov}(X,Y) = E(XY) - E(X)E(Y)$$

elle est aussi linéaire :

$$\text{cov}(aX+bY, cZ) = a\cdot c\cdot \text{cov}(X,Z) + b\cdot c\cdot \text{cov}(Y,Z)$$

Indépendance :

- deux variables sont indépendantes $\rightarrow$ leur covariance est nulle (mais la converse est fausse !)
- mais si deux variables sont jointly Gaussian (voir plus bas) $\rightarrow$ (elles sont indépendantes $\leftrightarrow$ leur covariance est nulle)

Correlation

$$\text{corr}(X,Y) = \frac{\text{cov}(X,Y)}{\sqrt{\text{var}(X)\text{var}(Y)}}$$

(plus elle est proche de -1 ou 1, plus la correlation est forte, une correlation de zéro n'implique pas l'indépendance)

Résumé hypothesis testing

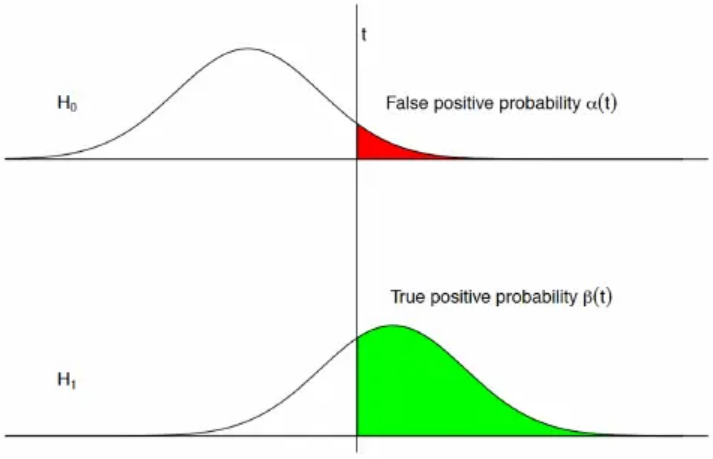
On a :

- la null hypothesis H_0
- la alternative hypothesis H_1
  La deuxième couvre toutes les possibilités non couvertes par H_0, car, pour que la logique soit respectée ($\neg H_1 \rightarrow H_0$). On **sait** que c'est soit H_0 soit H_1.

Faux positifs et faux négatifs

- false positive : on dit que H_0 est fausse alors qu'elle est vraie. (en fait un *positif* revient à détecter une valeur bizarre qui va faire rejeter H_0).
- false negative : on dit que H_0 est vraie alors qu'elle est fausse. (une valeur aurait dû être détectée et nous faire rejeter H_0).

La **taille** d'un test c'est la probabilité de faux positifs (rejeter H_0 alors qu'elle est vraie).
La **puissance** d'un test c'est $1 - P_{\text{faux negatifs}}$. C'est la capacité du test à rejeter H_0 quand elle est fausse.
Une **simple hypothesis** spécifie complètement (= avec les paramètres) la distribution des données tandis que l'hypothèse composite non.



On ne peut pas minimiser les false positive sans commencer à rendre notre test incapable de détecter les true positive.

Chi-Square statistics

Une somme de k normales standardisées au carré suit une distribution chi-square :

$$Q = Z_1^2 + Z_2^2 + \ldots + Z_k^2 \sim \mathcal{X}_k^2$$

$E(X) = k$ et $\text{var}(X) = 2k$.

On peut l'utiliser dans le cas du **Pearson statistics** :
(ce résultat fait une grande approximation c'est que $E_i \approx \text{var}\,(O_i)$ ce qui est le cas uniquement pour Poisson, à peu près pour la multinomiale..)

Comme ça on peut convertir chaque catégorie dans le cas de la multinomiale p. ex. en une normale standardisée :

$$T = \sum_{i=1}^{k} \left(\frac{O_i - E_i}{\sqrt{\text{var}\,(O_i)}} \right)^2 \sim \mathcal{X}_{k-1}^2 \approx \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

avec $k - 1 = v$ le degré de liberté. Ici le degré de liberté est $k - 1$ car on a k catégories et on a fixé la somme des O_i (on sait que $O_1 + \ldots + O_k = n$). On a une table qui nous donne, pour un degré de liberté donné, la valeur de la cumulative de la distribution du chi-square.

$p_{\text{obs}} = \mathbb{P}_0\left(T \geq t_{\text{obs}}\right)$ avec t_{obs} la valeur calculée comme $\sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$. Ce calcule facilement parce qu'on a la table pour les $\mathcal{X}^2$.

Confidence intervals

Soit on trouve un pivot (c'est-à-dire une variable dont la distribution ne dépend pas de θ), par exemple le $\frac{\max(y_1,\ldots,y_n)}{\theta}$ dans le cas d'une uniforme puis on trouve les quantiles z_α pour l'encadrer, c'est-à-dire tels que (pour de taille équivalente des deux côtés) :

$$P\left(z_\alpha \leq \frac{\max\left(y_1,\ldots,y_n\right)}{\theta} \leq z_{1-\alpha}\right)$$

On trouve $P\left(\frac{\max(y_1,\ldots,y_n)}{\theta}\right) \leq z_{1-\alpha} = \alpha = z_{1-\alpha}^n \iff z_{1-\alpha} = (\alpha)^{\frac{1}{n}}$.

Soit on trouve un pivot approximatif avec le théorème central limite. On sait que Z suit une normale standard (donc ne dépend pas d'un paramètre). Par exemple pour des uniformes :

$$Z = \frac{\overline{Y} - E(\overline{Y})}{\sqrt{\text{Var}(\overline{Y})}} = \frac{\overline{Y} - \theta/2}{\sqrt{\frac{\theta^2}{12n}}}$$

On peut ensuite faire le même procédé que plus haut et trouver les z-score associés puis isoler θ.
On sait qu'on veut Z compris entre $\frac{z_a}{2}$ et $z_{1-\frac{\alpha}{2}}$ (les deux quantiles pour avoir l'intervalle de confiance) :

$$z_a \leq Z \leq z_b \Longrightarrow \overline{Y} + z_\alpha\sqrt{\frac{\sigma^2}{n}} \leq E(\overline{Y}) = \mu \leq \overline{Y} + z_{1-\frac{\alpha}{2}}\sqrt{\frac{\sigma^2}{n}}$$

(penser que z_α est négatif)

Approximer la variance

Soit s la standard deviation observée ($s = \sqrt{\text{var}_{\text{observée}}(X)}$).

On sait que les bornes L, U pour notre intervalle de confiance seront :

$$(L, U) = \left(\frac{(n-1)S^2}{\mathcal{X}^2_{n-1}\left(1-\alpha_L\right)}, \frac{(n-1)S^2}{\mathcal{X}^2_{n-1}\left(\alpha_U\right)}\right)$$

Neyman-Pearson lemma

Intuition : On veut minimiser la probabilité de faux positifs et de faux négatifs. Pour ça, on fixe d'abord la taille du test (la probabilité de rejeter H_0 alors qu'elle est vraie). Et à partir de ça, on va trouver la région de rejection la plus adéquate de cette taille α (celle qui maximise le likelihood de H_1). Nous n'avons vu que des cas où ce sera sur un des bords (par exemple on rejette H_0 si on a un $z_{\text{observé}} < -3$ si H_1 est une normale centrée en -6 et H_0 une normale centrée en 2, et si les deux moyennes étaient inversées on rejette H_0 si on observe un $z_{\text{observé}} > -3$.
On ne sait pas encore si on va décider de rejeter pour un z observé inférieur ou supérieur à

une valeur, ou bien pour une moyenne inférieure ou supérieure à une moyenne, etc. c'est la méthode dessous qui nous le dira.

Méthode : On introduit la fonction :

$$h(R) = \prod_{i=0}^{n} \frac{L_{H_1}(y_i, \theta_0)}{L_{H_0}(y_i, \theta_1)} = \frac{f_1(y_1) \cdot f_1(y_2) \cdot \ldots \cdot f_1(y_n)}{f_0(y_1) \cdot f_0(y_2) \cdot \ldots \cdot f_0(y_n)} > k$$

Il s'agit d'un produit, et non d'une somme, car nous voulons que y_1, y_2, ..., y_n soient chacun probables sous H_1. On cherche à voir si, quand on augmente **une certaine quantité** R (par exemple la somme des y_i ou la moyenne des y_i, la $\mathcal{X}^2$, est-ce que le ratio augmente (on se rapproche de H_1) ou diminue.

Nous choisissons donc la taille du test, c'es-à-dire une probabilité de faux positifs α (par exemple, 0.05). Ensuite, nous trouvons z_α tel que la probabilité sous H_0 que la quantité trouvée dépasse (ou est inférieure) r_α est égale à α :

$$P_{H_0}(R \geq r_\alpha) = \alpha \text{ ou le cas } \leq \text{ (en fonction de la croissance de } h(R))$$

Comme R est typiquement une somme de variables aléatoires indépendantes, (mais ce n'est pas forcément le cas !) nous pouvons par exemple l'approximer à l'aide du théorème central limite :

$$P_{H_0}(R \geq r_\alpha) = P_{H_0}\left(\frac{R - n \cdot \mu}{\sqrt{n \cdot \sigma^2}} \geq \frac{r_\alpha - n \cdot \mu}{\sqrt{n \cdot \sigma^2}}\right) = 1 - \Phi\left(\frac{r_\alpha - n \cdot \mu}{\sqrt{n \cdot \sigma^2}}\right) = \alpha.$$

Pour trouver r_α, on l'isole en utilisant l'inverse de la fonction de répartition cumulative Φ^{-1}.