Formation Continue – Rappels de statistiques

Vincent Guigue

Petit mémo de cours

A. Vocabulaire de base statistique en langage d'informaticien :

1. **Choix d'un objet à décrire :** variable aléatoire
   - Variable discrète.
     - Catégorielle. e.g. : Sport (modalités = natation, ping-pong, ...)
     - Ordinale. e.g. : Nombre de Vélibs disponibles
   - Variable continue. e.g. : Poids des carottes, tailles des personnes
   - Possibilité de discrétisation des variables continues :
     - $\Rightarrow$ calibre des carottes
     - intervalles de tailles pour les personnes : $< 1m40$, $[1m40, 1m50[$, ...
2. **Description complète** = distributions de probabilité

   Quantification de la probabilité d'occurence des évènements associés à une variables aléatoires :

   Variable discrète

   X : calibre des pommes de terre dans un champ
   (3 modalités)

X	1	2	3
$p(X)$	0.2	0.5	0.3

   Un univers a toujours une masse de 1 :

$$\sum_i p(X = x_i) = 1$$

   Variable continue

   X : age de la population française

   Masse de l'univers : $\int_{-\infty}^{\infty} p(X = x)dx = 1$

   Probabilité ponctuelle :

$$p(X = 35.981763\ ans) = 0$$

   Probabilité d'un intervalle :

$$p(40 < X < 50) = \int_{40}^{50} p(X = x)dx \geq 0$$

   Les variables continues sont plus difficiles à gérer... Mais nous en gérons plus rarement.
3. **Résumé** = moments mathématiques, quantiles

   Connaitre un phénomène, c'est connaitre sa loi de probabilité. Mais pour expliquer un phénomène à quelqu'un d'autre, il est bon de disposer d'outils pour transmettre une information concise.

   Variable discrète

   Espérance : *comportement moyen* $E[X] = \sum_i x_i p(X = x_i)$

   Variance : *déviation par rapport au comportement moyen* $V[X] = \sum_i (x_i - E[X])^2 p(X = x_i)$

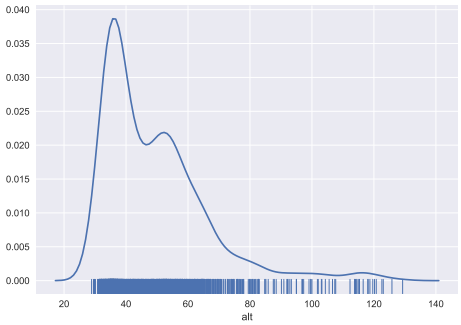
   Médiane : *comportement de l'individu situé au centre de la population* $M \ : \ P(X \leq M) = \frac{1}{2}$

   Premier décile : *comportement des 10% marginaux de la population* $Q_{10} \ : \ P(X \leq Q_{10}) = \frac{1}{10}$

   Variable continue : remplacer les sommes par des intégrales.
4. **Exemple : Altitude des stations de Vélib à Paris**

(a) La population fait 1227 stations : ce n'est pas possible de lire & comprendre les grandes tendances dans les données brutes.

(b) La distribution des altitudes donnent une description parfaite de la variable aléatoire, mais c'est une information lourde dont l'interprétation est réservée aux experts :



Je vois rapidement qu'il y a beaucoup de stations autour de 30m d'altitude et de nouveau un pic de stations autour de 55m d'altitude. Enfin, il y a quelques stations haut perchées, probablement sur les buttes parisiennes.

(c) Si je veux transmettre cette information à un non expert, ou à l'oral, je vais chercher à la résumer :
- En moyenne, les stations se trouvent à 49.2m d'altitude...
- Question : sont-elles compactes autour de cette moyenne ? Réponse : l'écart type, deux tiers des stations se trouvent dans l'intervalle 49.2 ± 17m.
- Question : à quelle altitude se trouve la station la plus représentative, celle pour laquelle autant de stations se trouvent au dessus qu'en dessous ? Réponse : la médiane, à 45m d'altitude.
- Du coup, je vois bien que la moyenne a été *déstabilisée* par les buttes parisiennes... Pour mieux comprendre, je me demande au-dessus de quelle altitude se trouve les 1% des stations les plus hautes. Réponse : le 99ème percentile, situé entre 115.2 et 129.3m d'altitude.

$\Rightarrow$ En quelques chiffres j'ai donné une image assez détaillée de la distribution ci-dessus.

B. Passage à plusieurs variables aléatoires, marginales, indépendances

1. **Loi jointe & lois marginales :** l'univers compte maintenant plusieurs variables aléatoires (mais a toujours une masse de 1)

   Exemple : probabilité d'appartenir à une certaine classe d'age et de pratiquer un sport $P(A, S)$

Sport \ Age	< 20	$[20, 30[$	$[30, 40[$,	$[40, 50[$	≥ 50	Marginale
Natation	0.02	0.05	0.09	0.08	0.08	0.32
Jogging	0.10	0.15	0.10	0.07	0.05	0.47
Tennis	0.02	0.03	0.06	0.07	0.03	0.21
Marginale	0.14	0.23	0.25	0.22	0.16	1

   Marginalisation = revenir à une variable

$$p(S = s_j) = \sum_i p(A = a_i, S = s_j), \qquad p(A = a_i) = \sum_j p(A = a_i, S = s_j)$$

2. **Conditionnalisation :** réduction de l'univers à un cas particulier sur l'une des variables

   Exemple : description des sports pratiqués par les $[20, 30[$ ans
   p(S — A = [20,30[) = p(S,A= [20,30[) / p(A= [20,30[)

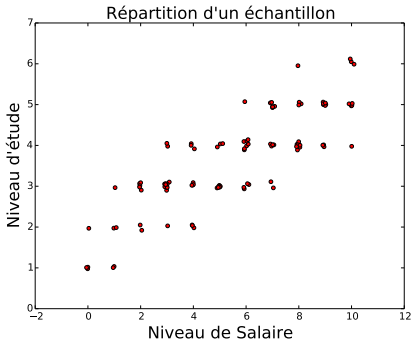
S	Natation	Jogging	Tennis
$p(S\|A = [20, 30[)$	0.22	0.65,	0.13

$$P(S|A) = \frac{P(S, A)}{P(A)}, \qquad \forall i, j \; p(S = s_j|A = a_i) = \frac{P(S = s_j, A = a_i)}{P(A = a_i)} = \frac{P(S = s_j, A = a_i)}{\sum_j p(A = a_i, S = s_j)}$$

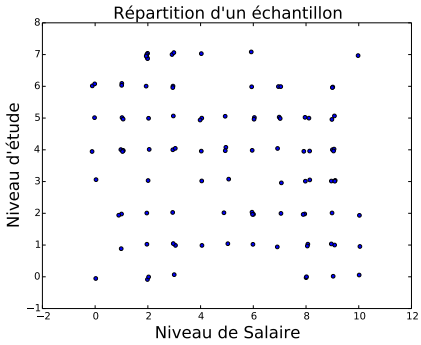
3. **Co-variance et coefficient de corrélation linéaire :** est ce que deux variables évoluent conjointement ? S'éloignent-t-elles de leur moyenne en même temps ?

   L'enjeu est de distinguer les cas de figure suivants :

Covariance *grande*



Covariance *petite*



$$COV(X,Y) = \mathrm{E}[(X - \mathrm{E}[X])\,(Y - \mathrm{E}[Y])] = \sum_i \sum_j (x_i - E[X])(y_j - E[Y])p(X = x_i, Y = y_j)$$

$\Rightarrow$ problème d'interprétation, la covariance est en unité de X fois unité de Y... On a donc inventé le coefficient de corrélation linéaire, sans unité, entre -1 et 1 :

$$\rho = \frac{COV(X,Y)}{\sigma_X \sigma_Y}, \qquad \sigma_X = \sqrt{V[X]}, \qquad \text{L'écart type est la racine de la variance.}$$

4. **Indépendance :** deux variables sont indépendantes si elles ne s'influencent pas mutuellement. C'est à dire si elle ne s'apporte pas d'information. C'est à dire si les lois marginales les caractérisent totalement et que la loi jointe est inutile... On note et on définit l'indépendance comme suit :

$$X \perp\!\!\!\perp Y \iff \forall i,j \; p(X = x_i, Y = y_j) = p(X = x_i) \times p(Y = y_j)$$

L'indépendance implique une covariance nulle... Mais malheureusement la réciproque n'est pas vraie (théoriquement).
L'indépendance est critique en informatique, car la complexité des algorithmes –et l'espace mémoire nécessaire pour stocker les modèles– dépend beaucoup de l'indépendance des variables aléatoires décrivant le problème.
Exemple : nous nous intéressons à un processus où n dés, potentiellement pipés, sont lancés. De combien de paramètres dépend la loi jointe quand d augmente... Respectivement lorsque les dés ne sont pas indépendants –ils sont magnétisés et s'influencent les uns les autres– ou sont indépendants –dés standards–.

n dés (à 6 faces)	Non indépendance des dés	Indépendance des dés
1	6	6
2	$6 \times 6 = 36$	$2 \times 6 = 12$ (deux lois marginales)
3	$6 \times 6 \times 6 = 216$ (cube de paramètres)	$3 \times 6 = 18$
4	$6^4 = 1296$	24
5	$6^5 = 7776$	30

C. Chaine de traitement de l'information en statistiques

1. **Echantillon :** en statistique, nous nous basons sur des données observée. Par exemple, si nos données sont décrites selon trois variables aléatoires, alors chaque individu est un triplet (x, y, z) et chaque variable peut prendre différentes valeurs. L'échantillon correspond à :

$$D = \{(x_1, y_1, z_1), \ldots, (x_\ell, y_\ell, z_\ell), \ldots, (x_N, y_N, z_N)\}$$

2. **Estimateurs :**
   — Table de contingence et estimation générique de la loi jointe. Si les variables X, Y, Z sont discrètes de modalité respective $2, 3$ et 4, il faut construire un tableau cubique de dimension $2 \times 3 \times 4 = 24$ cases pour contenir les estimations de $p(X = x_i, Y = y_j, Z = z_k)$. On compte combien de fois est apparue la configuration (x_i, y_j, z_k) parmi les N individus :

$$\hat{p}(X = x_i, Y = y_j, Z = z_k) = \frac{|\{(x,y,z) | x = x_i, y = y_j, z = z_k\}|}{N}$$

— Estimateur de l'espérance, de la variance, de la covariance :

$$\bar{x} = \frac{1}{N}\sum_{\ell} x_\ell, \qquad \hat{\sigma}_x = \sqrt{\frac{1}{N}\sum_{\ell}(x_\ell - \bar{x})^2}, \qquad cov(x,y) = \frac{1}{N}\sum_{\ell}(x_\ell - \bar{x})(y_\ell - \bar{y})$$

Attention, lorsque l'échantillon est petit, la variance empirique est biaisée

— Estimateur des quantiles :

(a) Ordonner les valeurs de $\{x_\ell\}$ dans un vecteur $x^\uparrow$

(b) Pour le premier décile, déterminer l'indice $index = \lceil N/10 \rceil$, $q_{10} = x^\uparrow[index]$

3. **Processus usuel de traitement des données basé sur les modèles probabilistes paramétriques :**

(a) Choix d'une loi de probabilité paramétrique.
Afin de réduire le nombre de paramètres à évaluer, un expert choisit une loi de probabilité (e.g. en discret : Bernoulli, Binomiale, Géométrique, Poisson, Exponentielle... En continu : Loi normale).
On note cette loi p_θ, où θ désigne les paramètres de la loi (e.g. la probabilité de victoire dans une épreuve de Bernoulli...)

(b) Hypothèse d'indépendance des individus et vraisemblance de l'échantillon :
les individus de l'échantillon sont généralement considérés comme indépendant et identiquement distribué (iid). C'est à dire, qu'ils ne s'influence pas les uns les autres et qu'ils sont tirés selon une distribution sous-jacente unique.
La vraisemblance (*Likelihood* en anglais) s'exprime comme :

$$\mathcal{L} = \prod_{\ell} p_\theta(X = x_\ell, Y = y_\ell, Z = z_\ell)$$

(c) Estimation des paramètres optimaux de la loi.
Les paramètres optimaux sont obtenus en annulant la dérivée de la vraisemblance dont la formulation analytique est convexe pour la majorité des lois de probabilités usuelles.

$$\theta^\star = \arg\max_\theta \mathcal{L} \iff \frac{\partial \mathcal{L}}{\partial \theta} = 0$$

Note importante : les paramètres optimaux $\theta^\star$ sont les mêmes que l'on travaille sur $\mathcal{L}$ ou $\log \mathcal{L}$. Cette dernière est souvent beaucoup plus simple à calculer.

Exercice 1 – Tutoriel numpy

`numpy` est la bibliothèque python la plus connue pour le traitement des données matricielles. La bibliothèque offrait initialement les possibilités de Matlab dans un environnement de programmation plus ouvert (et compatible avec les données textuelles) ; elle est maintenant incontournable car interfacée avec plusieurs bibliothèques d'analyse de données et d'apprentissage automatique.
Les échantillons de données sont des matrices :

$$D = \begin{bmatrix} x_1 & y_1 & z_1 \\ & \vdots & \\ x_\ell & y_\ell & z_\ell \end{bmatrix}$$

Les lois de probabilités sont des matrices (voire des tenseurs, c'est à dire des matrices de plus de 2 dimensions) :

$$P = \begin{bmatrix} Z = z_1 \begin{bmatrix} p(x_1,y_1,z_1) & p(x_1,y_2,z_1) & p(x_1,y_3,z_1) \\ p(x_2,y_1,z_1) & p(x_2,y_2,z_1) & p(x_2,y_3,z_1) \end{bmatrix}, \\ \\ Z = z_2 \begin{bmatrix} p(x_1,y_1,z_2) & p(x_1,y_2,z_2) & p(x_1,y_3,z_2) \\ p(x_2,y_1,z_2) & p(x_2,y_2,z_2) & p(x_2,y_3,z_2) \end{bmatrix} \end{bmatrix}$$

Vous devez savoir créer, lire, manipuler des matrices. Nous allons utiliser un notebook python pour comprendre le fonctionnement des méthodes de base `numpy`.
Je vous propose des tutoriels sous la forme de notebook python, pour éviter les copier-coller de code trop fastidieux.
La forme notebook a été choisie pour des raisons d'efficacité, mais il est possible de repasser en python de base : `fichier− >download as`

Exercice 2 – Données Vélib –version JCDecaux–

Dans cet exercice, vous travaillez en notebook, en spyder ou tout autre environnement de votre choix supportant python 3. Néanmoins, afin de faciliter le développement –ou simplement les copier-coller– un canevas est fourni sous forme de notebook. Les questions sont rappelées dans le notebook.

Q 2.1 Tutoriel sur les fonctions d'affichage. Afficher la position des stations en 2D et colorier chaque point selon l'altitude de la station.

Q 2.2 Etude de la corrélation entre l'altitude et le nombre de Vélibs disponibles.
- Tracé du nuage de points par rapport à ces deux axes.
- Normalisation des variables (nombre de Vélibs disponibles)
- Calcul du coefficient de corrélation et interprétation des chiffres obtenus

Q 2.3 Calcul et affichage de la distribution jointe entre altitude et disponibilité.
- Discrétisation (calcul d'histogrammes).
- Remplissage d'une table de contingence & passage à l'estimation fréquentielle.
- Affichage de la distribution jointe : sous forme d'image puis en utilisant des outils plus avancés.

Q 2.4 Calcul et interprétation de la distribution conditionnelle
- Exploitation du théorème des probabilités totales :

$$P(D|A) = \frac{P(D,A)}{P(A)}$$

- Affichage et interprétation de cette distribution
- Calcul de l'espérance conditionnelle $E[D|A]$.

Q 2.5 (Pour aller plus loin) Indépendance : taille des stations (S) VS arrondissements (Arr)
L'indépendance est un phénomène critique lors de l'implémentation des méthodes... Avec deux variables aléatoires, il suffit de tester :

$$X \perp\!\!\!\perp Y \iff \forall i, jp(X = x_i, Y = y_j) = p(X = x_i) \times p(Y = y_j)$$

... Ce qui n'est jamais vérifié exactement sur des données réelles.
La bonne question est donc : suis-je assez proche d'un phénomène d'indépendance ?

Q 2.5.1 Etude de corrélation sur la taille des stations par rapport aux arrondissements
- tracé de la distribution jointe (sns.jointplot)
- calcul du coefficient de corrélation

⇒ la faible valeur de coefficient de corrélation de corrélation nous donne un indice, mais nous nous rappelons que dans ce sens là, ce n'est pas une démonstration

Q 2.5.2 Calcul d'indépendance exact :
- Discrétiser (ou plutot redistribuer) les tailles de stations sur 10 valeurs
- Calcul de la jointe `P_ArrS`
- Attention aux indices `arr` entre 1 et 20 ⇒ indices entre 0 et 19
- Calcul des marginales `P_Arr, P_S` (trivial à partir de la loi jointe)
- Calcul de `PI_ArrS = P_Arr x P_S`
  - Implémentation du calcul par double boucle ⇒ trivial
  - calcul matriciel ⇒ non trivial (il faut dessiner les matrices sur une feuille de brouillon)
  - transformation des vecteurs en matrice + usage de dot
    `PI_ArrS = P_Arr.reshape(Narr, 1).dot(P_S.reshape(1,Ns))`
  - Comparaison des valeurs `PI_ArrS` vs `P_ArrS`
    `diff = ((PI_ArrS - P_ArrS)**2).sum()`
    ⇒ aucune chance d'arriver à 0...

Q 2.5.3 Application du test de χ^2 : mesure d'une distance entre distribution
- Mesure de la distance entre deux distributions P_t (distribution théorique, issue des marginales dans notre exemple) et P_o (distribution jointe)

$$D = \sum_i \sum_j N \frac{(P_t(i,j) - P_o(i,j))^2}{P_t(i,j)}$$

La mesure dépend du nombre d'observation N

— Chaque distribution est caractérisée par un nombre de degrés de libertés qui vaut ici :

$$DoF = (|Arr| - 1)(|S| - 1) = 171$$

— La distance limite, avec α de marge d'erreur, est donnée par :

```
import scipy.stats as stats
stats.chi2.ppf(α, DoF)
```

Peut-on conclure que l'arrondissement est indépendant de la taille des stations ?

Q 2.6 Visualisation de données en grande dimension (ie plus que deux)
Réflexion autour de l'enjeu de la visualisation des données. Cas pratique de mise en œuvre de TSNE pour passer des coordonnées (x, y, z) à un affichage 2D.

Quelques exercices plus théoriques

Ces exercices ont vocation à servir d'illustration ou de support. Nous n'auront pas le temps de les traiter durant les TP, mais ils peuvent vous servir à mieux cerner les compétences que nous cherchons à vous transmettre.

Exercice 3 – Indépendence

Soit deux dés à six faces non pipés, un de couleur blanc et un de couleur noir. Les deux sont jetés une fois. On définit les événement suivants :

— le dé blanc donne 1, 2 ou 3.
— le dé blanc donne 2, 3 ou 6.
— la somme des deux dés est égal à 9.
— les deux dés donnent deux nombres égaux, dont la somme est inférieure à 9.

Q 3.1 Quel est la probabilité des ces événements ?

Q 3.2 Quels événements sont deux-à-deux indépendants ?

Q 3.3 Sont-ils mutuellement indépendants ? Si non, trouvez les groupes (à trois ou quatre événements) qui sont mutuellement indépendants.

Exercice 4 – La roulette

Dans les casinos, la roulette contient 37 numéros : 18 rouges, 18 noirs et un vert. Quand la roulette tourne, la bille a autant de chances de tomber sur chacun des 37 numéros. Si l'on mise 1 EUR sur le rouge et que ce dernier sort, on gagne 1 EUR, sinon on perd la mise de 1 EUR.

Q 4.1 La roulette vous sera t-elle profitable ?

Q 4.1.1 Soit X la variable aléatoire représentant le résultat d'une mise de 1 EUR. Quelle est la distribution de probabilité de X ? Quelle est l'espérance de X ?

Q 4.1.2 En moyenne combien gagnerez-vous ou perdrez-vous par mise ?

Q 4.1.3 Combien gagnerez-vous ou perdrez-vous si vous jouez 100 fois en misant 1 EUR à chaque fois ? 1000 fois ? Peut-on en déduire que la roulette n'est pas un jeu profitable ? Justifiez votre réponse.

Exercice 5 – Paradoxe de Simpson

Le recensement des jugements prononcés dans l'état de Floride entre 1973 et 1978 a permis d'établir le tableau suivant, qui présente les sentences en fonction de la couleur de peau de l'accusé :

meurtrier	peine de mort	autre sentence
noir	59	2547
blanc	72	2185

Q 5.1 Calculez la probabilité d'obtenir la peine de mort sachant que l'on est noir, puis sachant que l'on est blanc. Qu'en concluez-vous ?

Q 5.2 En fait le tableau ci-dessus est une synthèse du tableau ci-dessous :

victime	meurtrier	peine de mort	autre sentence
blanche	noir	48	238
	blanc	72	2074
noire	noir	11	2309
	blanc	0	111

Calculez la probabilité d'obtenir la peine de mort conditionnellement à la couleur de peau de l'accusé et de la victime. La justice est-elle clémente envers les noirs dans l'état de Floride ? Justifiez votre réponse.

Exercice 6 – Sport et age

Dans un échantillon aléatoire de 240 personnes, on a recueilli l'information suivante sur l'âge et sur le type de sport le plus fréquemment pratiqué à Jussieu :

âge / activité sportive	moins de 20 ans	[20; 25[ans	[25; 30[ans	plus de 30 ans
jogging	15	20	15	30
natation	15	10	20	25
ping pong	20	10	30	30

Q 6.1 Quelles sont les deux variables aléatoires étudiables ?

Q 6.2 Estimer la loi jointe de ces deux variables.

Q 6.3 Calculer la probabilité qu'a un individu de faire de la natation (dans cet échantillon). Quelle est la probabilité qu'un individu de cet échantillon tiré au hasard ait entre 20 et 25 ans ?

Q 6.4 Calculer la probabilité qu'a un individu qui fait du jogging d'avoir plus de 30 ans (dans cet échantillon).

Q 6.5 Ces deux variables aléatoires semblent-elles indépendantes ?