Reconnaissance de forme: Rappels de probabilités et de statistiques

1er février 2010

Plan

- Considérons une pièce de monnaie...
- on a une chance sur deux d'avoir pile et une chance sur deux d'avoir face...
- Signification ? ?

- En fait peut être que si je savais comment la pièce a été lancée, je pourrais savoir exactement si pile ou face.
- Pas d'incertitude ? Pas besoin de probabilités ?
- Si ! Introduction artificielle des probabilités pour quantifier quand même ce qu'on ne sait pas mesurer.

- Peut être que en fait la probabilité d'avoir pile dépend des conditions de température et de pression...
- Dans ce cas je pourrais affiner mon modèle...
- Mais je n'ai pas cette information...
- Alors j'essaie de comprendre avec un modèle plus simple même si il ne correspond pas à la réalité.

Dans un monde aléatoire, plusieurs choses sont possibles.
Il est IMPOSSIBLE de dire dans la plupart des cas ce qui va se passer.
Mais on peut dire quels sont les événements possibles et leur affecter une probabilité.

Exemple : le jeu de cartes

On tire une carte au hasard. Cela peut être :

- ... une dame de pique
- ... un as de trèfle
- ... un as OU une dame
- ... un coeur
- .....

Toutes ces possibilités sont des événements aléatoires
On peut composer les événements entre eux (tout système d'événement est stable par union, intersection, passage au complémentaire et à la limite).

Dans un monde probabilisé, on ne sait pas ce qui va se passer mais on peut dire pour chaque événement si il va plus ou moins souvent se produire si on peut faire plusieur fois la même expérience. C'est la notion de probabilité.

On tire une carte au hasard de façon uniforme :

$$P(DameOURoi) = P(Dame) + P(Roi)$$

Probabilités additives MAIS :

$$P(DameOUPique) = P(Dame) + P(Pique) - P(DameETPique)$$

Supposons maintenant que on rajoute 10 dames de pique dans notre jeu.

$$P(DamedePique) =$$

Par ailleurs :

$$P(AsdePique) =$$

On a effectué un changement de loi (on a modifié la probabilité des événements).

On reprend l'exemple des cartes. On tire une carte c. Soit $X(c)$ une fonction des événements qui vaut :

- 0 si la carte est entre 7 et 10 (petite carte)
- 1 si la carte est un valet, une dame ou un roi (carte moyenne)
- 2 si la carte est un as (bonne carte)

On dit que X est une variable aléatoire.

De façon plus générale, on appelle variable aléatoire toute fonction réelle (ou dans $\mathbb{R}^d$) des événements... Mais souvent on réécrit les événements en fonction de la variable aléatoire.
En effet, on peut écrire les probabilités de chaque valeurs de la variable aléatoire, dont l'image réciproque est un événement :

Si l'espace des événements est discret (par exemple nombre de personne dans le groupe choisi) :
Il suffit de connaître la probabilité de chaque événement élémentaire, les $(p_x)_x$ et on peut en déduire en sommant la probabilité de tous les événements.
On peut définir la fonction de répartition comme :

$$P(evennement < x) = \sum_{i<x} p_i$$

Exemple de loi discrète : la loi de poisson

$$P(evennement = x) = \frac{e^{-\lambda}\lambda^{k}}{k!}$$

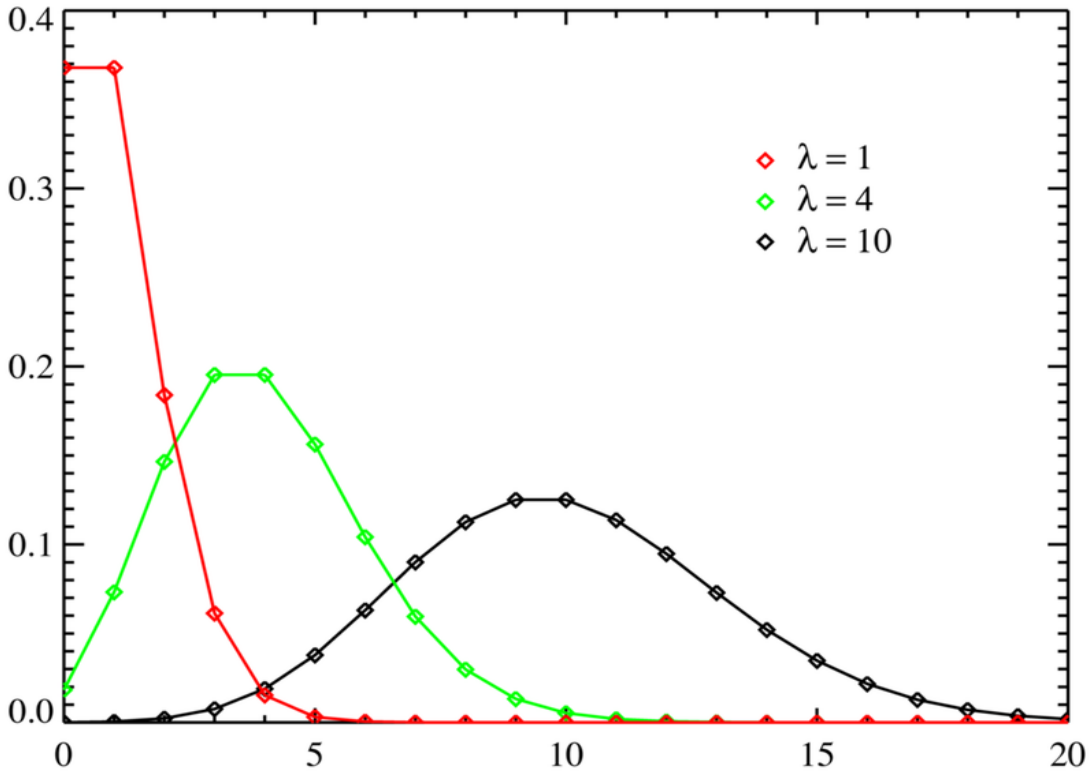

FIGURE: Loi de poisson (distribution)

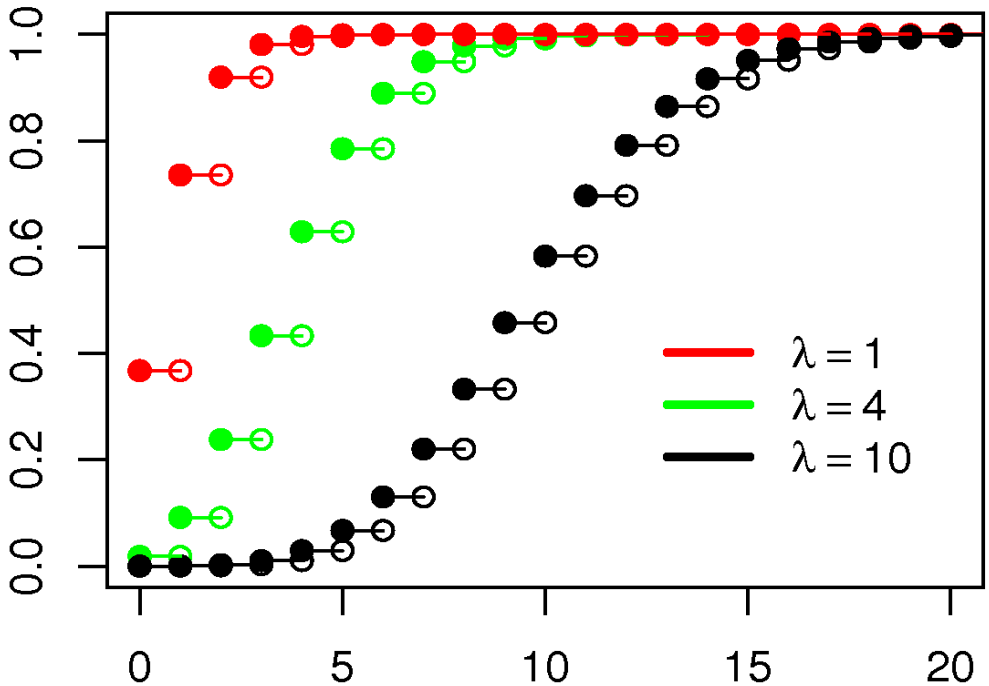

FIGURE: Loi de poisson (fonction de répartition)

Si l'espace des événements est continu (par exemple la taille pour des individus) :
Il faut connaître la densité de probabilité, c'est à dire $f(x)$ telle que :

$$F(x) = P(X < x) = \int_{-\infty}^{x} f(x)dx$$

On voit que $f(x)dx$ est analogue aux p_x.
Par ailleurs, F est la fonction de répartition.
Remarque : on peut utiliser la densite pour calculer la probabilité de tous les évènements.

Exemple de loi continue : la loi normale

$$P(evenement = x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{\sigma^2}}$$

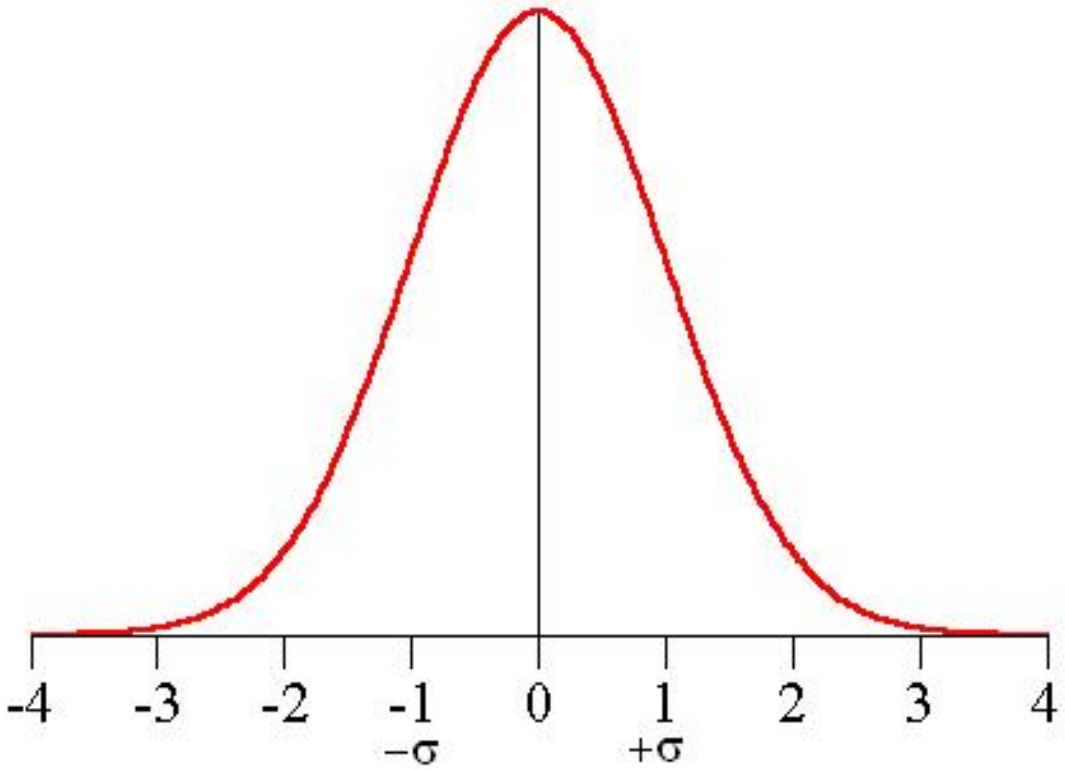

FIGURE: Loi normale (densité)

Espérance d'une variable aléatoire :

$$E(X) = \sum p_i x_i = \int x f(x) dx$$

Intuition de valeur moyenne :

Variance d'une variable aléatoire :

$$V(X) = \sum p_i (x_i - EX)^2 = \int (x - EX)^2 f(x) dx = EX^2 - (EX)^2$$

Ces quantités sont des cas particulier de choses plus générales, les moments :
Moment d'ordre p :

$$EX^p$$

Soient X_1 et X_2 deux variables aléatoires.
Comment modéliser des relations entre elles ?
Quelles genre de relations peut il y avoir ?

Supposons que X_1 soit le poids et X_2 la taille d'un individu.
On peut supposer que si X_1 est grand, la probabilité que X_2 soit grand est grande aussi.
Donc le fait de connaître X_1 nous donne des informations sur X_2... et réciproquement.
Remarque : pas de notion de causalité

A l'inverse, si X_1 est le premier jet d'une pièce et X_2 le deuxième... la connaissance de l'un ne renseigne pas sur l'autre.
Ces variables sont INDEPENDANTES. C'est une notion très importante en statistiques : en effet, c'est l'indépendance et la répétition qui permettent l'estimation (ou alors si dépendance il y a, cette dernière doit être controlée).

La notion de dépendance se retrouve bien sur en intégration.
Deux évennements A et B sont :

- Indépendants si $P(A \bigcap B) = P(A)P(B)$
- Dépendants sinon

L'indépendance signifie donc que on peut intégrer indépendamment.

On peut étendre cette propriété à deux variables aléatoires X_1 et X_2 :

- X_1 et X_2 sont indépendantes si $\forall A, B$ événements, $P(X_1 \in A, X_2 \in B) = P_{X_1,X_2}(A, B) = P_{X_1}(A)P_{X_2}(B)$
- Dépendants sinon

Il faut remarquer ici que les événements A et B peuvent être ou ne pas être sur le même espace, que les lois de X_1 et X_2 peuvent être différentes ou non, et que cela n'a rien à voir avec la notion de dépendance qui vise uniquement à modéliser les liens entre X_1 et X_2.

Formules :

Si deux événements sont dépendants, on est intéressé à savoir comment la connaissance de la réalisation de l'un des deux doit influencer notre appréhension de l'autre.
On veut **conditionner**.
On note : $P(A/B)$ c'est à dire probabilité de A "sachant" B.

Si A et B sont deux événements, on aura **toujours** :

$$P(A, B) = P(A|B)P(B)$$

Si A et B sont indépendants, $P(A|B) = P(A)$

Formules :

En pratique il est très compliqué de vérifier si deux variables aléatoires sont dépendantes, et encore plus de quantifier cette dépendance.
En effet cela demande de connaître la loi conditionnelle.
Il faut trouver un outil pour mesurer la liaison entre deux variables qui, si il est moins précis, sera plus souple.

La racine carrée du moment d'ordre 2 d'une variable aléatoire $\sqrt{EX^2}$, possède les propriétés d'une norme.
On peut définir le produit scalaire associé pour deux variables aléatoires évoluant dans le même espace :

$$E(X_1 X_2)$$

Cette quantité peut mesurer les relations entre X_1 et X_2. En effet :

$$X_1 \perp X_2 \Longrightarrow E(X_1, X_2) = E(X_1)E(X_2)$$

ATTENTION : il n'y a pas équivalence en général.

On peut définir covariance et corrélation :

En pratique on réalise les fameux tableaux croisés pour mesurer les dépendances entre variables aléatoires.

Il est rare de rencontrer une variable aléatoire...
... Mais courant de disposer d'un échantillon.
On espère que ce dernier est **représentatif** de la loi.
Notion d' indépendance : très importante pour disposer d'évènements répétés.

Echantillon représentatif ?

Les probabilités travaillent avec des variables aléatoires, et les statistiques avec des échantillons.
Les probabilités s'intéressent aux lois, les statistiques aux théorèmes limites.
Les probabilités modélisent, les statistiques utilisent les probabilités pour estimer.

Reprenons la pièce...
... Et supposons qu'un tricheur joue.
On a envie de savoir dans quelle mesure il triche...
... Donc quelle est la proportion moyenne de pile et de face.
Pour cela il faut estimer cette quantité à l'aide d'un échantillon de plusieurs lancers.
Si cet échantillon est "représentatif", on devrait pouvoir **estimer** cette valeur.

Par exemple pour la pièce supposons que on observe 1 si la pièce est tombée sur face et 0 si elle est tombée sur pile.
On observe plusieurs lancers : $X_1, ..., X_N$ est notre échantillon.
On suppose que nos lancers sont indépendants et que $X_i \rightsquigarrow B(p)$. on veut estimer p.
En fait, $E(X_i) = p$
On peut estimer p par la moyenne empirique :

$$\frac{1}{N} \sum X_i$$

On peut de même construire des estimateurs de tous les moments, notemment la variance :

De façon plus générale, ont peut estimer l'espérance de toute fonction de notre variable aléatoire si elle est suffisamment régulière et "pas trop grande".

Mais qu'est ce qu'un échantillon représentatif ?

On dispose de beaucoup de variables aléatoires indépendantes de même loi, de **carré intégrable**.

$$\frac{1}{N}\sum X_i \rightarrowtail E(X)$$

On dispose de beaucoup de variables aléatoires indépendantes de même loi, de **carré intégrable**. On note σ^2 la variance.

$$\sqrt{N}(\frac{1}{N}\sum X_i - EX) \rightarrowtail N(0, \sigma^2)$$

La loi normale est une loi très importante...

Le TCL permet de donner une vitesse de convergence, en $\sqrt{N}$, pour la loi des grands nombres.
Il permettra de construire des tests et des intervalles de confiance pour quantifier la précision de l'estimation.

On peut de même construire ces deux théorèmes pour toute transformation des observations suffisamment régulière et "pas trop grande" :

Lorsque l'on est face à un phénomène dont le résultat est incertain, on a envie de le modéliser par une loi de probabilité.
On dit d'un modèle qu'il est **paramétrique** si, au lieu de dire que nos observations suivent une loi précise, on dit qu'elles peuvent suivre un ensemble de lois indexées par un ou des paramètres.

Exemple de la pièce :

L'objectif est d'estimer ce paramètre à l'aide d'un échantillon.
Pour la pièce, on peut prendre la moyenne empirique, mais uniquement car le paramètre que l'on veut estimer est également la moyenne de notre distribution. Que faire sans ce constat ?

Dans un cas général, on définit ce que l'on appelle la vraisemblance de notre modèle qui n'est autre que la probabilité sous le modèle paramétré de notre échantillon :

$$L(X_1 = x_1, ...X_N = x_N; p) = P_p(X_1, ..., X_n)$$

Si nos variables sont indépendantes :

$$L(X_1 = x_1, ...X_N = X_N; p) = \prod_i P_p(X_i = x_i)$$

Exemple de la pièce :

On veut choisir la valeur du paramètre qui est la plus "vraisemblable" pour l'échantillon dont on dispose.
Une approche logique consiste à maximiser cette probabilité suivant les paramètres.
On fait le **maximum de vraisemblance**.
En pratique on maximise la **log-vraisemblance**.
Il y a des cas pour lesquels cette approche ne fournira pas de résultats...

Calcul du maximum de vraisemblance pour la pièce :

Cette technique se justifie intuitivement, encore faut il s'assurer de la convergence de l'estimateur obtenu...
On peut envisager le démontrer à la main au cas par cas en calculant espérance et variance de l'estimateur et en appliquant le TCL et la loi des grands nombres...

... Mais c'est long. En fait si on considère tout estimateur du maximum de vraisembmance comme une fonction implicite de nos données, on peut comprendre pourquoi dans le cas général l'estimateur va converger à l'aide de la LGN et du TCL.

Estimer, oui. Mais avec quelle précision ?
Questions que l'on peut se poser :

- Cette hypothèse pour l'estimateur est elle vraisemblable ?
- Dans quel intervalle se trouve probablement la vraie valeur ?

Principe du test :

- On dispose d'une hypothèse H_0.
- On veut tester voir si H_0 est probable ou pas du tout avec une précision donnée.

Exemple : H_0 *: la moyenne de cette loi normale est 0*

Vocabulaire :

- Erreur de type 1 : la probabilité de rejeter à tort l'hypothèse nulle lorsqu'elle est vraie.
- Erreur de type 2 : la possibilité d'accepter à tort l'hypothèse nulle lorsqu'elle est fausse.

Test classique : On fixe l'erreur de première espèce au niveau α. Test **asymétrique** (on privilégie H_0).
Remarque : Existence d'autres types de tests notemment des tests bayesiens.

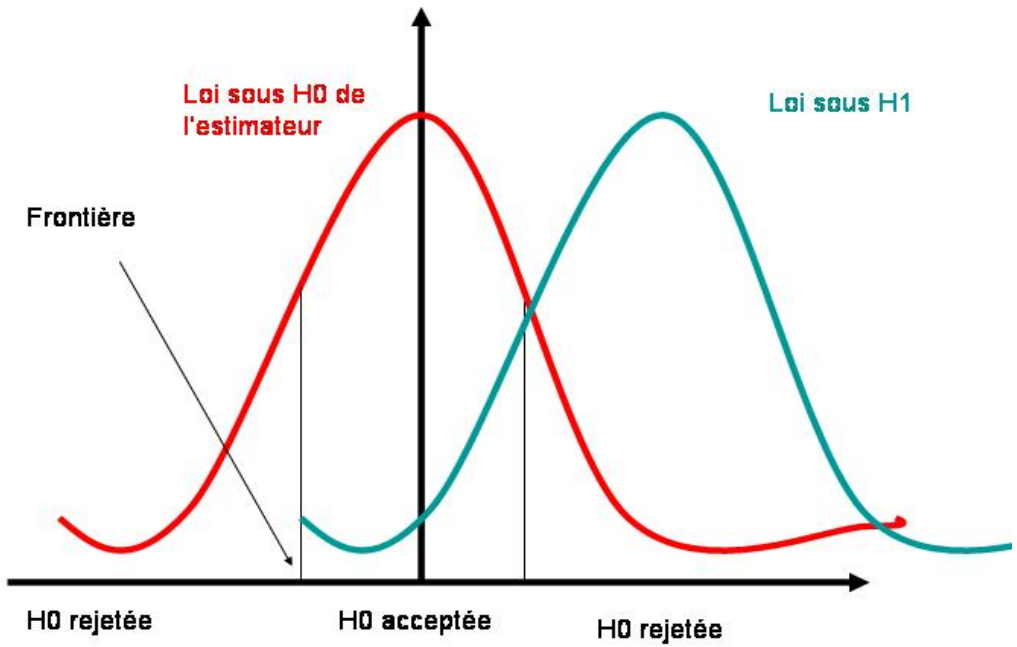

FIGURE: Test d'hypothèses

Exemple : encore la pièce

On estime un paramètre. C'est une valeur précise.
Intuitivement, la probabilité que le vrai paramètre vaille exactement le paramètre observé est nulle !
On veut trouver un intervalle ou la probabilité que le vrai paramètre soit est grande.

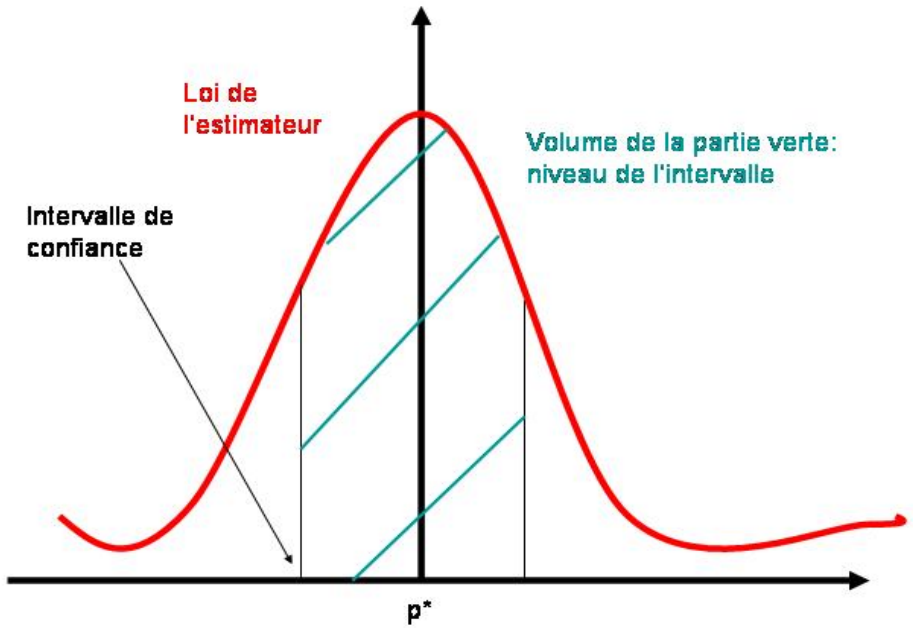

FIGURE: Intervalle de confiance

Exemple : toujours la pièce