Statistique Classique et Statistique Bayesienne

Plaçons-nous dans les conditions du cas Guilde du Livre. Vous avez une décision à prendre (acheter les ouvrages ou non) et les conséquences de cette décision dépendent d'un paramètre : la proportion p de vos clients qui vont acheter l'ouvrage.

Un statisticien classique vous dira : "*le paramètre p est un paramètre inconnu mais certain*". Pour estimer ce paramètre, il va vous suggérer de tirer un échantillon au hasard de clients. Supposons que l'on retienne un échantillon de 100 clients et que sur ces 100, 49 se déclarent prêts à acheter l'ouvrage. Le statisticien classique vous dira alors que l'estimation ponctuelle de p est 49/100. Son raisonnement est fondé sur toute une théorie, dite théorie de l'estimation, et sur les propriétés particulières de l'*estimateur du maximum de vraisemblance*. De fait, ce statisticien vous dira que le nombre de personnes favorables dans un échantillon de 100 personnes suit une loi Binomiale de paramètres 100 et p et que la valeur de p qui rend maximum la probabilité d'obtenir ce que l'on a obtenu (c'est à dire 49 personnes favorables sur 100) est précisément 49/100. Pour le démontrer il suffit d'écrire que la probabilité d'obtenir ce que l'on a obtenu n'est autre que :

$$\frac{100!}{49!51!}p^{49}(1-p)^{51}$$

et de dériver cette expression par rapport à p.

Bien sûr, le statisticien classique ne se contente pas de cette estimation ponctuelle. En effet, en s'en tenant là, on conclurait sur la base de 100 lancés qu'une pièce est truquée dès lors que l'on observe un nombre de Pile différent de 50 alors que l'on sait que :

P(obtenir 50 Piles sur 100 lancés/pièce non truquée) = 0,08

Pour aller plus loin le statisticien classique vous proposera de réaliser une "estimation par intervalle de confiance". Pour cela, il suffit de se rappeler que dès que la taille de l'échantillon devient suffisamment grande, la loi Binomiale converge vers la loi Normale. Dans notre cas, tout cours de statistique vous dira que la proportion de personnes favorables dans notre échantillon suit approximativement une loi Normale de moyenne p et de variance p(1-p)/n. Si j'appelle $\overline{X}$ cette proportion dans l'échantillon (attention : avant de connaître les résultats précis de l'échantillon $\overline{X}$ est une variable aléatoire) on a donc :

$$P(-1{,}96 < \frac{\overline{X}-p}{\sqrt{p(1-p)/n}} < +1{,}96) = 0{,}95$$

la valeur 1,96 étant lue dans les tables de la loi Normale centrée réduite.

On a alors :

$$P(\overline{X} - 1{,}96\sqrt{p(1-p)/n} < p < \overline{X} + 1{,}96\sqrt{p(1-p)/n}) = 0{,}95.$$

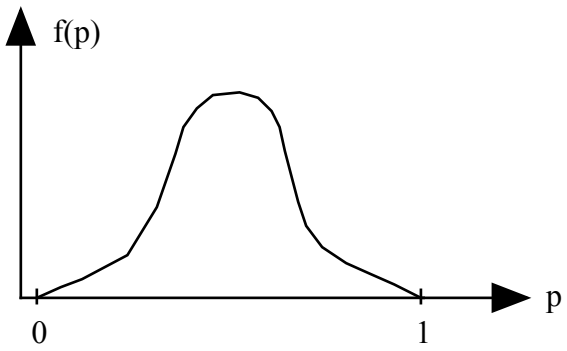
Arrivé là le statisticien classique sera généralement un peu embarrassé pour vous dire ce dont il s'agit vraiment. Il vous dira d'abord que l'on peut estimer p(1-p) soit par 0,25 (sa valeur maximale) soit par $\overline{x}(1-\overline{x})$ où $\overline{x}$ est la valeur de la proportion constatée dans l'échantillon. Il vous dira ensuite qu'il a bâti un *intervalle de confiance* à 95% pour p. Dans notre cas on aurait obtenu comme réalisation de l'intervalle de confiance (en remplaçant p(1–p) par 0,25) :

$p = 0{,}49 \pm 0{,}098$

L'ennui avec cet intervalle, c'est son interprétation. Puisque p est un paramètre certain, il ne fait pas sens d'affirmer que la probabilité pour que p soit compris entre 39,2% et 58,8% est de 95%. Si vous demandez au statisticien classique d'interpréter son intervalle de confiance, il vous fera probablement un long discours pour vous dire que si l'on avait tiré 100 échantillons et si on avait construit 100 intervalles de confiance sur la base de ces échantillons, alors la "vraie" valeur de p aurait été comprise dans 95 de ces intervalles (en moyenne). Tout ceci, vous en conviendrez, est bien difficile à comprendre, d'autant plus que l'on a rarement le temps de tirer 100 échantillons.

Le statisticien bayesien ne vous dira pas du tout la même chose, même si, en bout de course, les deux démarches peuvent parfois se rejoindre.

Tout d'abord il vous mettra en garde. Faire de la statistique n'est peut-être pas intéressant pour vous si cela coûte trop cher : un petit calcul de VEIP n'est jamais superflu. Le statisticien bayesien vous dira : "*Il existe une vraie valeur de p dans la nature. Cette valeur vous est inconnue, mais vous avez à son sujet un certain nombre d'informations. Par exemple vous savez qu'il est très improbable que p soit très proche de 1 ou de 0. Cette information vous vient de votre connaissance de vos clients et de votre intuition. Je vais essayer par un certain nombre de questions simples d'encoder vos croyances sous la forme d'une distribution de probabilité subjective sur p*". Ces questions sont du type "Désirez-vous parier 100 F sur le fait que p est inférieur à 0,4 ou préférez-vous pariez ces 100 F à Pile ou Face, etc.". Nous avons déjà l'habitude de ce type de question. À la suite de cet "encodage" on obtiendra une distribution de probabilité subjective sur p dont je peux représenter graphiquement la densité :



À partir de cette distribution de probabilité subjective, les choses deviennent beaucoup plus simples. Si je veux, pour une raison ou pour une autre, faire de l'estimation ponctuelle, je peux retenir un indicateur de tendance centrale de cette distribution de probabilité : mode, médiane ou moyenne. Je peux tout aussi simplement utiliser cette distribution pour bâtir un "vrai" intervalle de confiance au sens où je peux trouver deux bornes p_1 et p_2 telles que, pour moi, la probabilité que p soit compris entre p_1 et p_2 soit de 95% (par exemple).

Si je décide alors qu'il est rentable pour moi d'acquérir plus d'information (et cette rentabilité sera appréciée en tenant compte des conséquences de mes décisions) il ne me reste plus qu'à réviser mes croyances à la lumière de l'information apportée par l'échantillon en utilisant le théorème de Bayes. Dans le cas continu, celui-ci nous dit que :

$$f(p/x) = \frac{f(p)f(x/p)}{\int f(p)f(x/p)dp}$$

si j'appelle x les résultats apportés par l'échantillon.

Si je le souhaite je peux refaire de l'estimation ponctuelle et/ou par intervalle de confiance en utilisant cette nouvelle distribution de probabilité de densité f(p/x).

Les démarches classiques et bayesiennes sont donc fort différentes, avec, selon moi, un net avantage à la seconde du point de vue de la cohérence et de la prise en compte des conséquences du problème, lorsque la statistique a pour but de nous aider à décider.

Ces deux démarches peuvent cependant coïncider au niveau des résultats, en particulier lorsque l'on dispose de très peu d'information a priori et que f(p) tend vers une loi Uniforme. Ceci est relativement naturel puisque, dans l'optique classique, on fait précisément l'hypothèse, implicite, que l'on ne sait rien a priori et que toute l'information est apportée par l'échantillon. Dans le cas où l'on ne sait rien a priori, on a $f(p) = 1$, $\forall\, p \in [0\ ;\ 1]$. Dans la formule de Bayes dans le cas continu on peut remarquer que le dénominateur ne dépend pas de p puisque que l'on intègre sur toute les valeurs possibles de p. On peut donc écrire

$f(p/x) = K x f(x/p)$

où K est un terme ne dépendant pas de p.

Dans ces conditions faire de l'estimation ponctuelle après révision de mes croyances revient à chercher une caractéristique de tendance centrale la distribution révisée. Si je choisis d'utiliser le mode de cette distribution, je vais retenir la valeur de p maximisant f(p/x). Or cette valeur est maximale lorsque f(x/p) est maximale c'est à dire lorsque p maximise la vraisemblance de l'information apportée par l'échantillon. On se retrouve donc, dans ce cas très particulier, dans les conditions de l'estimation ponctuelle classique avec l'estimateur du maximum de vraisemblance.

Un cas particulier important de la statistique bayesienne concerne le cas où la distribution de probabilité a priori suit une loi approximativement normale.

Si vos croyances a priori suivent une loi Normale de moyenne $\dot{\mu}$ et de variance $\dot{\sigma}^2$ et que les résultats (une moyenne par exemple) provenant d'un échantillon de taille n suivent une loi Normale de moyenne μ et de variance σ^2/n (attention $\dot{\mu}$ et $\dot{\sigma}^2$ n'ont rien à voir avec μ et σ^2/n. La première variance traduit l'état de votre information tandis que la seconde reflète la variance dans la population) alors l'application du théorème de Bayes (après un peu de manipulation) nous dit alors que nos croyances révisées à la lumière des résultats de l'échantillon suivent un loi Normale

de moyenne $\dfrac{\dot{\mu}/\dot{\sigma}^2 + \mu/(\sigma^2/n)}{1/\dot{\sigma}^2 + 1/(\sigma^2/n)}$

et de variance $\dfrac{\dot{\sigma}^2(\sigma^2/n)}{\dot{\sigma}^2 + (\sigma^2/n)}$

Il est facile de montrer que lorsque $\dot{\sigma}^2$ tend vers l'infini, c'est à dire lorsque les croyances a priori tendent vers une loi Uniforme, les croyances a posteriori suivent une loi Normale de moyenne μ et de variance σ^2/n et l'on se retrouve dans le cas de la statistique classique. À l'inverse lorsque l'on a un comportement têtu a priori avec $\dot{\sigma}^2$ tendant vers 0 l'information apportée par l'échantillon ne contribue pas à remettre en cause les croyances antérieures et a posteriori vos croyances sont les mêmes qu'a priori. Lorsque la taille de l'échantillon tend vers l'infini, la variance σ^2/n tend vers 0. On constate alors qu'a posteriori les croyances suivent une loi Normale de moyenne μ avec un écart type tendant vers 0 et on retrouve également ici comme cas limite la statistique classique.

Supposons par exemple que l'on pense que la proportion d'électeurs qui votera Républicain aux prochaines élections présidentielles aux USA a autant de chance d'être inférieure que

supérieure à 52%. En affinant un peu votre distribution de probabilité a priori, on s'aperçoit que votre intervalle de confiance au niveau 50% pour cette proportion est [49% ; 55%] et que cette distribution est sensiblement normale.

Sachant que P(p > 55%) = 0,25 et que la moyenne de votre distribution est égale à 52%, un calcul simple permet de conclure que vos croyances a priori suivent une loi Normale de moyenne 52% et d'écart type 3%/0,67 = 4.5% (le 0,67 venant d'une table de loi normale). Vous lisez dans la presse que sur 400 personnes interrogées lors d'un sondage, 192 ont émis l'intention de voter Républicain. De même que plus haut, tout cours de statistique nous enseigne que la proportion de personnes favorables dans l'échantillon suit une loi Normale de moyenne $\hat{p}$ et de variance $\hat{p}(1-\hat{p})/400$, si j'appelle $\hat{p}$ la proportion de personnes votant Républicain dans l'électorat. Le problème c'est que l'on ne connaît pas $\hat{p}$. On peut cependant l'estimer par 192/400 ≈ 0,48 et on a $\hat{p}(1-\hat{p})/400 \approx (0,025)^2$.

A posteriori vos croyances suivent donc une loi Normale

de moyenne $\dfrac{0,52/(0,045)^2 + 0,48/(0,025)^2}{1/(0,045)^2 + 1/(0,025)^2} = 49\%$

et de variance $\dfrac{(0,045)^2(0,025)^2}{(0,045)^2 + (0,025)^2} = (2\%)^2$

Votre nouvel intervalle de confiance à 50% est donc en pourcentage de 49% ± 0,66x2% soit 49% ± 1,32%. Votre intervalle de confiance à 95% est de 49% ± 1,96x2% = 49 ± 3,92%.

En utilisant uniquement l'information fournie par l'échantillon, votre intervalle de confiance à 50% en statistique classique aurait été de 48% ± 0,66x2,5% = 48% ± 1,6%, l'intervalle de confiance à 95% étant 48% ± 1,96x2,5% = 48% ± 4,9%.