En attendant mon taxi...

Gérard FLEURY

(A propos d'un problème de A. Engel,
"les certitudes du hasard", éditions CEDIC)

Au cours d'un voyage, j'attendais à une station de taxis d'une ville. Cinq personnes étaient avant moi. Pour tromper l'ennui, ayant remarqué que chaque taxi portait un numéro, je le notai. Un de mes voisins m'apprit que chaque taxi de la ville portait un tel numéro, entre 1 et un nombre maximum M qu'il ne connaissait pas, mais qu'il pensait être au moins 3 000, sans en être sûr. De plus, chaque année, ces numéros étaient tous attribués. Ayant noté les numéros suivants :

1 119, 1 504, 1 403, 789, 1 038 et 485,

la seule chose certaine à propos de M était que $M \geq 1\ 504$. je me demandai si mon voisin avait raison en supposant $M = 3\ 000$.

1 Des éléments de réponse.

Pour répondre à cette question, je cherchai, en admettant que $M = 3\ 000$, quelle était la probabilité d'obtenir un numéro maximum de 1 504 en ayant observé 6 taxis ?

1.1 Une première réponse.

Ayant prélevé 6 éléments (sans remise) parmi $\{1, 2, ...; 3\ 000\}$ (et en supposant que ce prélèvement ait été fait au hasard parmi tous les taxis de la ville), la probabilité qu'ils soient tous prélevés dans $\{1, 2, ...; 1\ 504\}$ est : $p_1 = \dfrac{\binom{1\ 504}{6}}{\binom{3\ 000}{6}} = \dfrac{1\ 504 \times 1\ 503 \times 1\ 502 \times 1\ 501 \times 1\ 500 \times 1\ 499}{3\ 000 \times 2\ 999 \times 2\ 998 \times 2\ 997 \times 2\ 996 \times 2\ 995} \simeq 0,015\ 80.$

Ce calcul, fondé sur un modèle aussi exact que possible, est assez pénible.

1.2 Simplifions un peu.

Comme je savais que $M \geq 1\ 504$ et que je n'avais relevé le numéro que de 6 taxis, je me dis que faire un prélèvement (au hasard) avec ou sans remise dans l'ensemble $\{1, 2, ...; M\}$ ne devait pas changer grand chose à la probabilité cherchée. Avec ce nouveau modèle, approché du précédent, je trouvai un résultat fort proche et plus aisé à obtenir : $p_2 = \left(\dfrac{1\ 504}{3\ 000}\right)^6 \simeq 0,015\ 88.$

1.3 Simplifions encore.

Comme $M \geq 1\ 504$, je me dis que je pouvais sans doute faire comme si lire un numéro de taxi donnait non plus un entier, mais un réel suivant une loi uniforme sur $[0,5; M+0,5[$ et que l'on aurait ensuite remplacé par son entier le plus proche. En décidant ainsi de confondre le réel X et son entier le plus proche, lire un numéro au hasard peut être modélisé par une variable aléatoire réelle X de densité $f_M(t) = \begin{cases} \frac{1}{M} \text{ si } 0,5 \leq t \leq M + 0,5, \\ 0 \text{ ailleurs.} \end{cases}$

La fonction de répartition de X, définie par $F_M(t) = \mathbb{P}\{X \leq t\}$, vaut : $F_M(t) = \begin{cases} 0 \text{ si } t < 0,5, \\ \frac{t-0,5}{M} \text{ si } 0,5 \leq t \leq M + 0,5 \\ 1 \text{ si } t > M + 0,5. \end{cases}$

C'est une fonction continue, contrairement aux fonctions de répartition correspondant deux modèles précédents.

Si je n'avais lu que deux numéros (en effectuant deux lectures au hasard et indépendantes) X_1 et X_2, on aurait eu, pour $0,5 \leq t \leq M + 0,5$: $\mathbb{P}\{\max(X_1; X_2) \leq t\} = \mathbb{P}\left(\{X_1 \leq t\} \bigcap \{X_2 \leq t\}\right) =$

$\mathbb{P}\{X_1 \leq t\}.\mathbb{P}\{X_2 \leq t\} = [F_M(t)]^2 = \dfrac{(t-0,5)^2}{M^2}$ et, de même, avec 6 numéros :

$\mathbb{P}\{\max(X_1; X_2; X_3; X_4; X_5; X_6) \leq t\} = [F_M(t)]^6 = \dfrac{(t-0,5)^6}{M^6}$. En particulier, dans mon cas, le plus grand numéro lu étant 1 504, je trouvai, comme probabilité de cet événement, si $M = 3\,000$: $\dfrac{1\,503,5^6}{3\,000^6} \simeq 0,015\,85$.

1.4 Une première conclusion : il n'y a probablement pas 3 000 taxis.

Les deux derniers calculs sont des approximations (fort proches au demeurant) du premier. La dernière approximation étant à la fois plus proche et plus commode que la première, c'est ce dernier mode de calcul que je décidai d'utiliser de préférence, désormais. La probabilité calculée étant très proche de 0,016, donc faible, j'en conclus qu'**il était probablement faux de penser que M valait 3 000**. Mais alors...

2 Que dire de mieux que : "il y a au moins 1 504 taxis" ?

2.1 En n'utilisant que le maximum observé.

En admettant que les tirages étaient faits avec remise et selon une loi uniforme dans $[0,5; M+0,5[$ avec M **inconnu** mais supérieur ou égal à 1 504, la probabilité d'obtenir au plus 1 504 valait :

$$\mathbb{P}\left(\left\{\max_{i=1,2,\ldots,6} X_i \leq 1\,504\right\}\right) = \left(\tfrac{1\,503,5}{M}\right)^6.$$

Ayant décidé de supposer que M était tel que l'événement observé $\left\{\max_{i=1,2,\ldots,6} X_i \leq 1\,504\right\}$ n'était pas très rare ni très probable : par exemple que sa probabilité était dans $[0,025\ ;\ 0,975]$, j'en déduisis :

$$\frac{1\,503,5^6}{0,975} \leq M^6 \leq \frac{1\,503,5^6}{0,025} \text{ soit } \frac{1\,503,5}{0,975^{\frac{1}{6}}} \leq M \leq \frac{1\,503,5}{0,075^{\frac{1}{6}}}$$

d'où $1\,509,9 \leq M \leq 2\,780,4$ ce qui me donna (avec un risque de 1% de me tromper) la fourchette suivante, sur le nombre de taxis : $\mathbf{1\,510 \leq M \leq 2\,780}$.

Si j'avais souhaité décider avec une plus grande sécurité, j'aurais pu, par exemple, supposer que M était tel que l'événement $\left\{\max_{i=1,2,\ldots,6} X_i \leq 1\,504\right\}$ était de probabilité dans $[0,005\ ;\ 0,995]$. J'en aurais alors déduit : $\dfrac{1\,503,5}{0,995^{\frac{1}{6}}} \leq M \leq \dfrac{1\,503,5}{0,005^{\frac{1}{6}}}$ soit $1\,504,8 \leq M \leq 3\,635,9$ d'où, pour le nombre de taxis : $\mathbf{1\,504 \leq M \leq 3\,636}$.

Il faut noter que ce dernier résultat semble contredire le fait que j'aie pu précédemment affirmer $M \neq 3\,000$. Or il n'en est rien. Le premier raisonnement a consisté à évaluer la probabilité d'avoir $\max_{i=1,2,\ldots,6} X_i \leq 1\,504$ en ayant fait l'hypothèse que $M = 3\,000$, alors que le second raisonnement ne part d'aucune hypothèse sur M, mais évalue M en supposant que l'événement observé n'est ni très rare ni très probable, plus précisément que l'on a : $\mathbb{P}\left\{\max_{i=1,2,\ldots,6} X_i \leq 1\,504\right\} \in [0,005; 0,995]$. De plus, la borne supérieure trouvée (3 636) correspond à une probabilité de se tromper de 0,01 alors que le calcul précédent donnait, en faisant l'hypothèse $M = 3\,000$ une probabilité de se tromper de 0,016 : tout dépend évidemment du risque que l'on prend en calculant la fourchette.

2.2 En n'utilisant que le minimum observé.

Ce qui m'ennuyait, dans tous les calculs précédents me permettant d'estimer M, était le fait que je n'utilisais que le plus grand des 6 numéros lus, et pas réellement les 6. Or il est aisé d'utiliser également le plus petit, qui est, ici : 485. En particulier, $\mathbb{P}\left\{\min_{i=1,2,\ldots 6} X_i \geq t\right\} = \prod_{i=1}^{6} \mathbb{P}\{X_i \geq t\} = [\mathbb{P}\{X_i \geq t\}]^6$ soit $\mathbb{P}\left\{\min_{i=1,2,\ldots 6} X_i \geq 485\right\} = \left(1 - \dfrac{484,5}{M}\right)^6$. Cette dernière approche me donna, en supposant $M = 3\,000$, la probabilité $0,35$.

Pour encadrer M inconnu, en supposant que l'on avait
$\mathbb{P}\left\{\min_{i=1,2,\ldots,6} X_i \geq 485\right\} \in [0,025; 0,975]$, on trouve $1\,054,96 \leq M \leq 115\,062,8$ d'où :

$$1\,505 \leq M \leq 115\,063$$

et, en supposant que l'on avait $\mathbb{P}\left\{\min_{i=1,2,\ldots,6} X_i \geq 485\right\} \in [0,005; 0,995]$: $826,1 \leq M \leq 580\,187,6$ d'où :
$1\,504 \leq M \leq 580\,188$.

Ces résultats, fournis par le minimum observé sont moins intéressants que ce qu'apporte la connaissance du maximum observé, ce qui est conforme au sens commun, puisque c'est le maximum théorique M qui est recherché.

2.3 En utilisant à la fois le minimum et le maximum observés.

Par contre, je décidai de m'intéresser au couple formé des deux numéros extrêmes relevés :

$$\left(\min_{i=1,2,\ldots 6} X_i \; ; \; \max_{i=1,2,\ldots 6} X_i\right).$$

Remarquant que, si $t < u$: $\mathbb{P}\left[\left\{\min_{i=1,2,\ldots 6} X_i > t\right\} \bigcap \left\{\max_{i=1,2,\ldots 6} X_i \leq u\right\}\right] = \mathbb{P}\left[\bigcap_{i=1}^{6} \{t < X_i \leq u\}\right]$
$= \prod_{i=1}^{6} \mathbb{P}\left[\{t < X_i \leq u\}\right] = \left[F_M(u) - F_M(t)\right]^6$, j'obtins ainsi, sous l'hypothèse $M = 3\,000$:

$$\mathbb{P}\left[\left\{\min_{i=1,2,\ldots 6} X_i > 485\right\} \bigcap \left\{\max_{i=1,2,\ldots 6} X_i \leq 1\,504\right\}\right] = \left(\frac{1\,503,5}{3\,000} - \frac{483,5}{3\,000}\right)^6 \simeq 0,001\,54.$$

Ceci me permit de grandement renforcer la sécurité avec laquelle j'affirmai que M ne valait pas 3 000.
En utilisant la même remarque pour calculer une fourchette sur M et en partant de :

$$0,025 \leq \left(\frac{1\,020}{M}\right)^6 \leq 0,975,$$

je trouvai : $\frac{1\,020}{0,975^{\frac{1}{6}}} \leq M \leq \frac{1\,020}{0,025^{\frac{1}{6}}}$ soit : $1\,024,3 \leq M \leq 1\,886,3$ soit enfin, puisqu'il était assuré que $M \geq 1\,504$: $\mathbf{1\,504} \leq M \leq \mathbf{1\,886}$. Avec plus de sécurité, je trouvai : $\frac{1\,020}{0,995^{\frac{1}{6}}} \leq M \leq \frac{1\,020}{0,005^{\frac{1}{6}}}$ soit :
$1\,020,9 \leq M \leq 2\,466.6$ ou $\mathbf{1\,504} \leq M \leq \mathbf{2\,467}$.

3 Et après avoir croisé un autre taxi ?

Tout en faisant ces calculs dans mon taxi, j'en croisai un autre, portant, lui, le numéro 522. Comme le maximum et le minimum observés ne changent pas, on pourrait se dire que je n'avais pas plus d'information, or ce n'est que fausse apparence, car je disposais désormais de 7 observations au lieu de 6. Je recommençai donc mes principaux calculs et trouvai, en supposant $M = 3\,000$:

$$\mathbb{P}\{\max_{i=1,2,\ldots,7} X_i \leq 1\,504\} \simeq \left(\tfrac{1\,503,5}{3\,000}\right)^7 \simeq 0,007\,9$$

ce qui renforça ma conviction que l'on avait M différent de 3 000.
En supposant que $\left(\frac{1\,503,5}{M}\right)^7$ était compris entre 0,005 et 0,995, je trouvai alors $\frac{1\,503,5}{0,995^{\frac{1}{7}}} \leq M \leq \frac{1\,503,5}{0,005^{\frac{1}{7}}}$ soit :
$1\,504.6 \leq M \leq 3\,204,96$ ou $\mathbf{1\,505} \leq M \leq \mathbf{3\,205}$.
Enfin, en utilisant à la fois les deux renseignements : $\min_{i=1,2,\ldots,7} X_i > 484$ et $\max_{i=1,2,\ldots,7} X_i \leq 1\,504$, je trouvai, en supposant que $M = 3\,000$, $\mathbb{P}\left(\left\{\min_{i=1,2,\ldots,7} X_i > 484\right\} \bigcap \left\{\max_{i=1,2,\ldots,7} X_i \leq 1\,504\right\}\right) \simeq \left(\frac{1\,020}{3\,000}\right)^7 \simeq 0,000\,53$
et $\frac{1\,020}{0,995^{\frac{1}{7}}} \leq M \leq \frac{1\,020}{0,005^{\frac{1}{7}}}$ soit, puisque $M \geq 1\,504$: $1\,020,7 \leq M \leq 2\,174,3$ et finalement, avec une probabilité de me tromper de 1% : $\mathbf{1\,504} \leq M \leq \mathbf{2\,174}$.

4 En guise de commentaire sur la démarche.

Nous venons d'explorer, à partir d'un échantillon (ici les numéros des taxis relevés), deux types d'utilisation des statistiques qui consistent à affirmer quelque chose sur la population dont est extrait cet échantillon. Les méthodes ainsi mises en oeuvre relèvent du domaine des mathématiques que l'on nomme "statistiques inférentielles".

4.1 Deux types d'inférences statistiques.

Nous avons successivement :

1. **testé une hypothèse** (ici : $M = 3\ 000$). La conclusion est alors : cette hypothèse est vraie ou non, et cette conclusion est une sorte de pari, avec un risque de se tromper lorsque l'hypothèse est satisfaite (risque de première espèce), qui est choisi par l'utilisateur. Si, d'aventure, l'hypothèse n'est pas satisfaite, alors tout le raisonnement s'effondre et rien ne permet d'ailleurs, ici, d'évaluer la probabilité de commettre une telle erreur.

2. **estimé un paramètre** (ici M) par une fourchette, elle-même fruit d'une confiance qui est la probabilité qu'il soit effectivement dans cette fourchette. L'amplitude de la fourchette (et son inutilité !) augmente avec la sécurité avec laquelle on calcule.

4.2 Quelques remarques.

Notons que, sur cet exemple, on constate que :

1. l'amplitude de la fourchette encadrant le paramètre à estimer augmente avec la probabilité mesurant la confiance que l'on a en cet encadrement, mais une trop grande amplitude perd vite toute utilité...

2. dans le cas où l'on teste une hypothèse (ici $M = 3\ 000$) contre une autre (ici $M \neq 3\ 000$), la conséquence de la remarque précédente est que, si la confiance que l'on souhaite avoir en sa conclusion, lorsque cette hypothèse est vraie, est choisie par l'utilisateur, il suffit d'augmenter suffisamment cette confiance pour que l'hypothèse ne soit plus rejetée. Or, si l'hypothèse est fausse en réalité, le fait de l'accepter tout de même constitue une erreur dont le risque (risque de seconde espèce) croît alors, sans que rien ne le mesure, dans ce que nous avons fait !

Notons que, sur cet exemple, on constate également que la conclusion du test (ou la fourchette de l'estimation)

1. est d'autant plus fiable (ou précise) que l'on utilise les "bons" éléments (appelés estimateurs) pour l'établir. Ici, l'utilisation du maximum de l'échantillon est plus efficace que l'utilisation de son minimum. Plus efficace encore est l'utilisation simultanée du maximum et du minimum de l'échantillon.

2. est d'autant plus fiable (ou précise) que la taille de l'échantillon est grande.

Enfin, le choix du seuil à partir duquel on décide que la probabilité d'un événement fait qu'on le considère comme "anormal", compte tenu des hypothèses faites, est à la disposition de la personne qui pose le problème. Ce choix ne résulte généralement pas d'une démarche d'ordre mathématique, mais relève plutôt du domaine d'application.

De plus, les événements "anormaux" sont ici, ceux qui relèvent de valeurs "trop" élevées (ou "trop" basses, ou les deux) d'une fonction, de valeurs qui s'éloignent "trop" d'une valeur "centrale". Parfois, l'expert du domaine d'application peut au contraire décider que ce sont les valeurs "trop" proches d'une valeur "centrale" qui entraînent une suspicion. Ce serait par exemple le cas si l'on suspectait des données d'avoir été fabriquées ou manipulées, par exemple pour paraître "au hasard"...

En tous cas, des choix à faire a priori, et qui sont essentiels pour le déroulement du test statistique ne relèvent pas directement des mathématiques, mais du domaine d'application.

Ces constatations sont valables dans la plupart des applications relevant des statistiques inférentielles.