



## Etudes statistiques pour l'aide à la décision

### – Normaliser la présentation –

par Bernard Beauzamy  
PDG, SCM SA

Novembre 2008

Une aide à la décision, par exemple en matière d'investissement ou de sécurité, repose généralement sur l'exploitation de données disponibles ; elle fait donc appel aux statistiques. Des exemples typiques sont :

- le déploiement d'un réseau de transport ou d'un réseau d'énergie dans une zone ;
- l'analyse des effets des pollutions, des rayonnements ionisants, etc.

Les décisions prises sont souvent lourdes de conséquences et représentent des coûts importants, aussi est-il souhaitable de les asseoir sur des études solides. Deux conditions sont absolument nécessaires et elles devraient être exigées de tous les décideurs :

- l'étude doit être vérifiable ;
- l'étude doit être réutilisable.

« Vérifiable » signifie que, si on donne tous les éléments à un autre expert, il doit être capable de dire si le travail réalisé est correct ou non.

« Réutilisable » veut dire que l'on peut réemployer des morceaux (par exemple des éléments de code) pour d'autres études, mais aussi que, cinq ans après, on doit pouvoir comparer ce que prédisait l'étude à ce qui s'est produit, et voir si les raisonnements étaient ou non corrects.

Ces deux conditions sont donc absolument de bon sens ; malheureusement, la plus grande partie des études que l'on nous demande d'expertiser ne sont ni vérifiables ni réutilisables. On ne sait pas quelles données ont été employées, ni quels raisonnements ont été faits. On se retrouve avec des graphes, des cartes - là c'est rouge, là c'est vert ; ici cela monte, ici cela descend - obtenus grâce à quelque logiciel. L'auteur, quinze jours après son chef d'œuvre, ne sait même plus sur quel bouton il a appuyé ; inutile de lui demander sa science cinq ans plus tard !

## **Recommandations techniques pour une présentation de l'étude**

Dans la grande majorité des études que l'on nous demande d'expertiser, les auteurs utilisent des outils statistiques de manière factice : par exemple, ils supposent que tel phénomène suit une loi de Gauss, dont ils cherchent à calculer les paramètres dans les cas A et B ; ils veulent montrer que ces paramètres sont différents d'un cas à l'autre. Mais cela ne prouve absolument rien, car rien ne dit que le phénomène en question suive effectivement une loi de Gauss.

De même, ils font souvent des hypothèses d'indépendance, qui ne sont pas satisfaites en pratique.

Nous donnons ci-dessous des règles simples de mise en œuvre d'une étude statistique ; la règle de base (c'est une évidence !) est qu'il ne faut faire aucune hypothèse arbitraire : il faut traiter les données comme elles viennent !

Nous prendrons le cas, fréquent en pratique, où le logiciel Excel est utilisé : dans la majeure partie des situations, il suffit très largement.

### **1. Les données brutes**

La première feuille (sheets(1)) doit contenir les données brutes sur lesquelles l'étude s'appuie. La présence des données brutes est une nécessité absolue pour que l'on puisse vérifier le travail ; elle est également nécessaire pour que l'on puisse comparer cinq ans après et voir ce qui a changé.

### **2. Les données de travail**

Sur la seconde feuille (sheets(2)), on mettra les données de travail. Ce peut être les données brutes, mais pas nécessairement. Par exemple, il nous arrive de normaliser les données (les faire toutes varier entre 0 et 1), ou bien encore de considérer des taux d'accroissement (par exemple : la différence entre une année et la précédente), etc.

Toutes ces transformations sont licites, à condition d'être reconnues comme telles : il ne s'agit plus de données brutes, mais de données ayant suivi un premier traitement ; ce traitement doit être justifié. Il ne faut pas confondre données de travail avec données brutes.

### **3. Le déroulement de l'étude**

Une étude statistique a souvent pour objectif d'établir un lien entre une quantité « objectif » que l'on cherche à contrôler et un certain nombre de paramètres « explicatifs » dont on cherche à savoir s'ils influent ou non sur l'objectif. Par exemple :

- On cherche à analyser l'objectif « taux de cancers dans une région » au travers des paramètres explicatifs : « présence de rayonnements ionisants », « nombre de fumeurs », etc.
- On cherche à analyser l'objectif « importance des stocks de gaz » au travers des paramètres explicatifs « importance des importations », « importance de la consommation », etc.

Le premier travail à réaliser est de constituer l'histogramme de la quantité « objectif », c'est-à-dire sa loi de probabilité. Ceci est très facile à réaliser sous Excel, de la manière suivante :

- On divise l'ensemble des valeurs possibles en intervalles de même largeur ;
- On compte, parmi les points de l'échantillon, combien tombent dans chaque intervalle.

Par exemple, si les intervalles ont pour largeur 10 et si les données sont dans la première colonne, sheets(1), on aura un code VBA du type suivant :

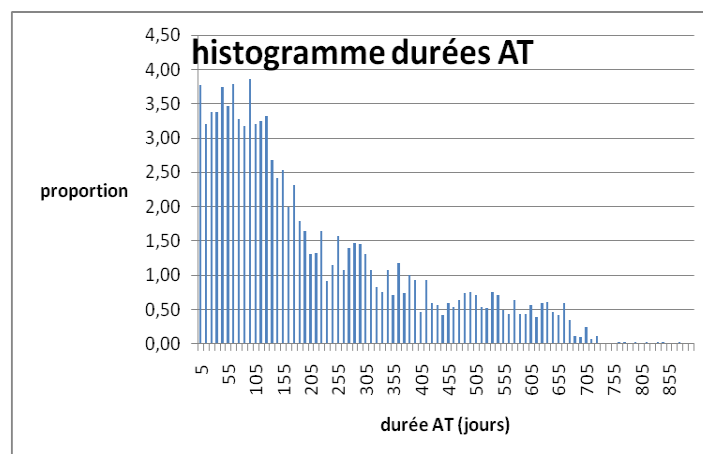
```
for k=1 to nbfinal
j=int( sheets(1).cells(k,1)/10)
sheets(3).cells(j,2)=sheets(3).cells(j,2)+1
next k
```

et la colonne 2 de la feuille 3 contient cet histogramme.

La colonne 1 contient les intervalles de construction ; dans notre exemple :

```
sheets(3).cells(k,1)= 10(k-1)
```

Voici un exemple d'histogramme :



Graphe 1 : histogramme

Il s'agit des durées d'arrêt de travail, pour des salariés ; on compte quelle proportion est entre 0 et 10 jours, entre 10 et 20 jours, etc.

La comparaison de deux histogrammes n'est pas facile, car ce qui compte est l'aire sous la courbe. On préférera donc travailler avec les cumuls, c'est-à-dire la fonction de répartition. Pour des raisons pratiques, on représente, pour chaque seuil, le nombre au dessus de ce seuil (et non pas au dessous), on travaille donc plutôt avec  $1-F$  qu'avec  $F$ .

Dans la colonne 3, dans la cellule  $k$ , on mettra donc la somme de toutes les cellules de la colonne 2, de  $k$  à la fin :

```

for k=1 to nbfinal
for j=k to nbfinal
sum=sum+sheets(3).cells(j,2)
next j
sheets(3).cells(k,3)=sum
sum=0
next k

```

Bien entendu, en première ligne de la colonne 3, on retrouve le total du nombre des points de l'échantillon, ce qui est une manière facile de vérifier que l'on n'a rien oublié.

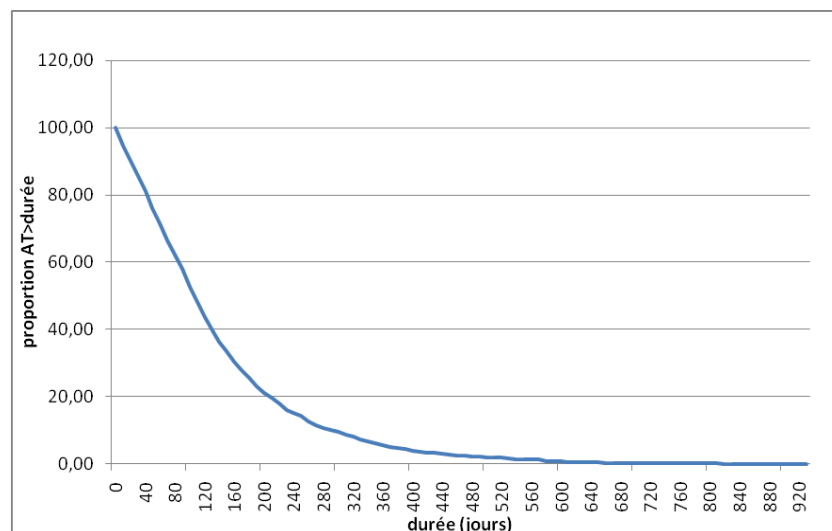
Enfin, en colonne 4, on met les pourcentages cumulés : total des pourcentages au-delà de  $k$  :

```

for k = 2 to nbfinal
sheets(3).cells(4,k)=sheets(3).cells(k,3)/sheets(3).cells(2,3)*100
next k

```

Comme on voit, c'est tout à fait élémentaire. Ce qu'on obtient est une courbe commençant à 100 et finissant à 0, du type suivant :

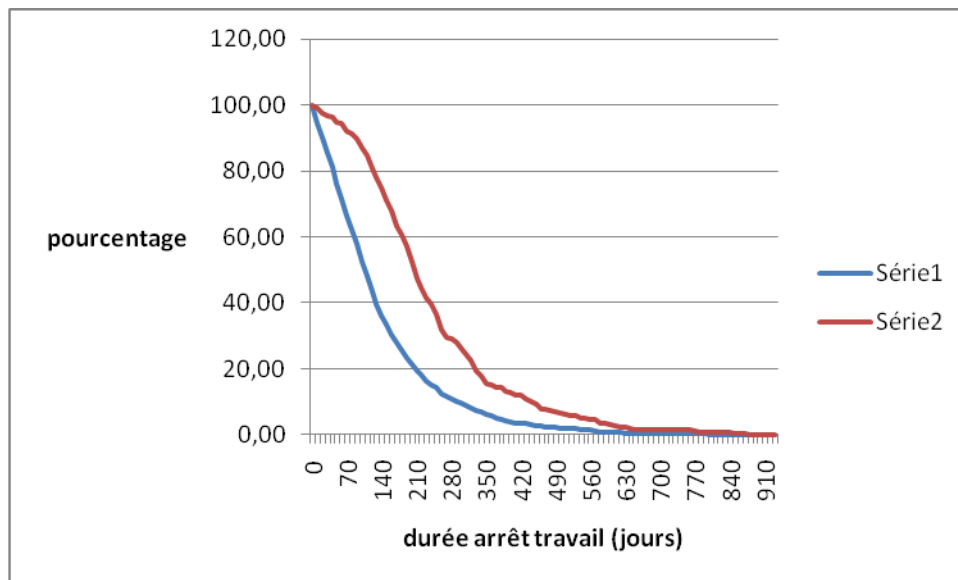


Graphe 2 : la fonction  $G$

Au dessus de chaque abscisse, dans notre exemple, on indique le nombre de salariés dont la durée d'arrêt de travail est supérieure ou égale à l'abscisse. En notation mathématique, on a tracé le graphe de la fonction :

$$G(x) = P\{X \geq x\}$$

La fonction de répartition (ou plus exactement  $G = 1 - F$ , comme expliqué plus haut) commence toujours à 100 et finit à 0 : cela rend les comparaisons faciles. Voyons un exemple :



Graphe 3 : exemple de comparaison

Ici, nous avons deux catégories de salariés : une en bleu et l'autre en rouge. Nous voyons que la fonction  $G$  pour les bleus est au dessous de la fonction  $G$  pour les rouges. Cela veut dire que, quelle que soit la durée (par exemple 200 jours), le nombre de salariés rouges qui excèdent cette durée est plus important que le nombre de salariés bleus ; il en résulte clairement que les rouges ont des absences plus longues que les bleus. La comparaison est immédiate, et repose intégralement sur des données factuelles.

#### 4. Prise en compte des paramètres explicatifs

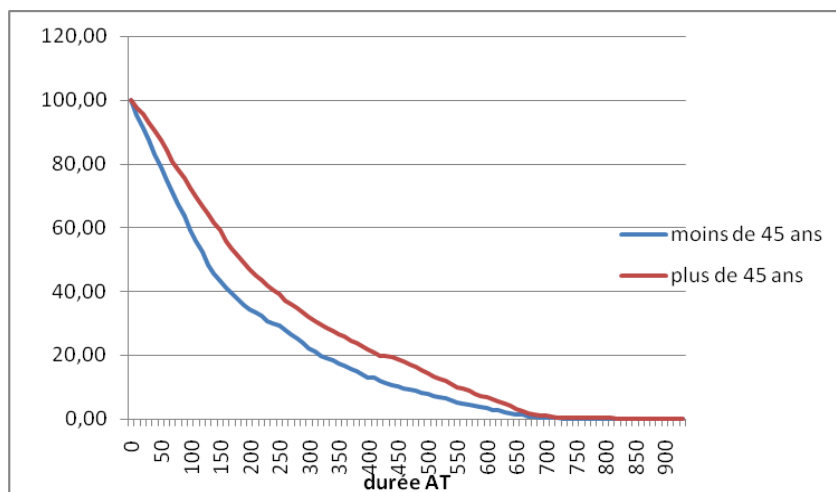
Pour prendre en compte les paramètres explicatifs, on réalise des lois de probabilité « conditionnelles », de la manière suivante. Prenons le premier paramètre explicatif, et disons que la moitié de ses valeurs sont au dessous de 45 et la moitié au dessus. On recommence l'opération ci-dessus dans le cas où le paramètre est  $\leq 45$  :

```

for k=1 to nbfinal
for j=k to nbfinal
if param1 <= 45 then
sum=sum+sheets(3).cells(j,2)
end if
next j
sheets(3).cells(k,3)=sum
sum=0
next k

```

et on obtient une courbe comme précédemment. On refait la même opération dans le cas où le paramètre est  $> 45$ . On peut mettre les deux sur le même graphique. Si, de manière systématique, on voit que la courbe 1 est au-dessous de la courbe 2, on en déduit que, pour chaque seuil, la probabilité d'être au-dessus de ce seuil est plus faible dans le cas 1 que dans le cas 2.



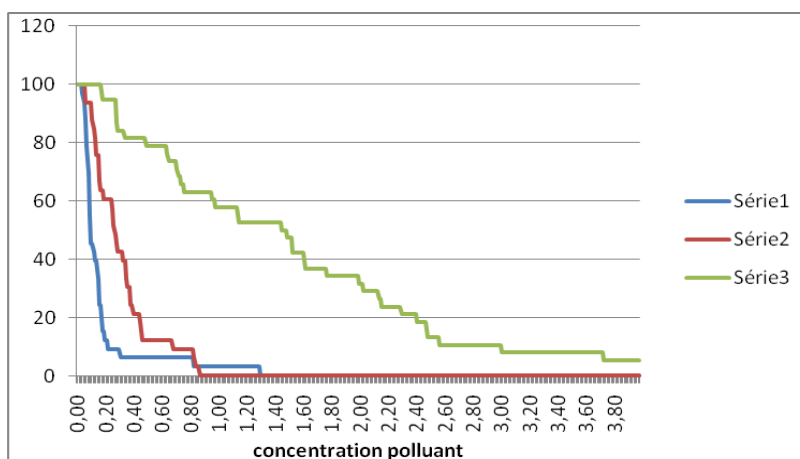
Graphique 4 : distinction selon l'âge

On constate ici que la courbe rouge est systématiquement au-dessus de la bleue : pour une durée fixée d'arrêt de travail, il y aura plus de salariés âgés que de salariés jeunes dont l'arrêt de travail excède cette durée : les AT des vieux sont plus longs que ceux des jeunes.

Cette manière de procéder – en exploitant des probabilités conditionnelles – est très satisfaisante en ce sens qu'elle ne fait aucune hypothèse a priori et n'utilise aucune « boîte noire ». On ne fait aucun test statistique, dont les conditions de validité sont discutables.

Si l'objectif est un seuil défini en termes de propreté (par exemple une concentration de polluants), on obtient une manière simple et indiscutable de comparer différents cas.

Voici un exemple ; il s'agit de la concentration d'un polluant, mesurée par différentes stations. La série 1, en bleu, concerne les zones où la densité de population est inférieure à 80 habitants au km<sup>2</sup>, la série 2 (en rouge), celle où la densité est entre 80 et 170 et la série 3 (en vert), les zones où la densité est supérieure à 170.



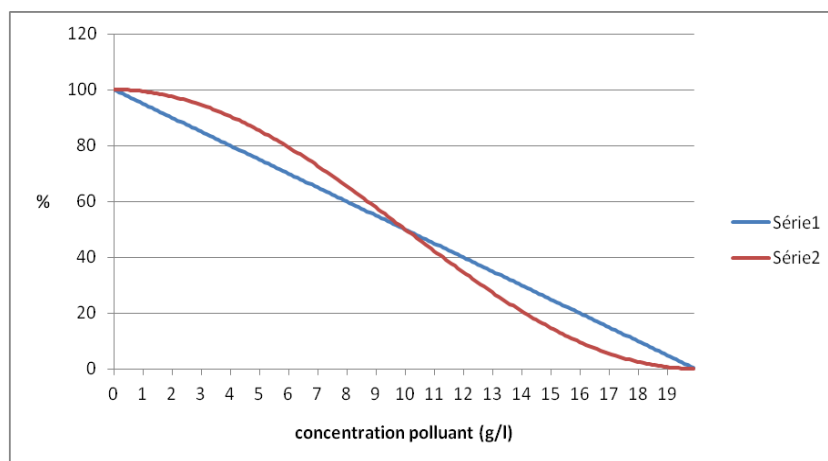
Graphique 5 : concentration de polluants, en fonction de la densité de population

On constate clairement que la courbe "forte densité" est au-dessus des deux autres : la proportion de stations, dans les zones de forte densité, ayant une pollution au-dessus d'un seuil donné, est plus élevée que dans les zones de moyenne ou faible densité.

Par exemple, 40 % des stations "forte densité" ont une pollution d'au moins 1,60 (gramme par litre) alors qu'aucune ne l'a pour les stations des zones "moyenne densité" ou "faible densité" ; il est complètement légitime de dire, sans aucune équivoque, que dans la zone considérée et pour ce polluant-là, les zones de forte densité de population sont plus polluées que les autres.

## 5. Exploitation d'un exemple mixte

Voyons maintenant comment interpréter un exemple "mixte" : l'une des courbes n'est pas toujours au dessus de l'autre ; elles se croisent.



Graphique 6 : situation mixte

Dans l'exemple ci-dessus (totalement factice), les deux courbes se croisent à la concentration 10 g/l. La valeur commune des deux fonctions est 50 % en ce point : nous aurons donc 50 % des stations de type 1 (bleu) et 50 % des stations de type 2 (rouge) avec une concentration supérieure ou égale à 10 g/l.

Si nous prenons une concentration supérieure, mettons 15 g/l, nous avons 25 % des bleus et 35.3 % des rouges au dessus de cette valeur. Nous en déduisons que les rouges sont moins sujets aux fortes pollutions.

Si nous prenons une concentration inférieure, mettons 5 g/l, nous avons 75 % des bleus et 85.3 % des rouges au dessus de cette valeur. Nous en déduisons que les rouges sont plus sujets aux faibles pollutions.

Par différence, dans l'intervalle 5 - 10, nous avons 25 % des bleus et 35.3 % des rouges. Là encore, la conclusion est donc très simple : il y a plus de bleus que de rouges dans l'intervalle 10 - 15 ; plus de rouges que de bleus dans l'intervalle 5 - 10.

Insistons bien sur le fait que tous ces nombres sont des proportions dans chaque classe, et non des proportions globales. Pour illustrer ceci, disons qu'il y a 1000 stations bleues et 5000 rouges. Alors, dans l'intervalle 5 - 10, le nombre de stations bleues est 25 % de 1000, soit 250, et le nombre de rouges est 35.3 % de 5000, soit 1765. Utiliser le nombre global de 6000 stations serait une erreur, puisque les rouges sont 5 fois plus représentées que les bleues.

Dans tous les cas, la position relative des différentes courbes, sur n'importe quel intervalle, renseigne sur l'importance relative du phénomène.