BIOINFORMATIQUE, Unité libre 2006 (2)
Partie "Statistiques" - Enoncés

Nous avons vu en cours que le nombre de chaines pour les protéines de la DBDB, traité comme une QT fournissait les résultats suivants

```
Nombre de valeurs : 388     protéines
Moyenne           :   1.07  chaines
Ecart-type        :   0.34  chaines
Cdv               :  32     %
```

Si on regarde maintenant la répartition du nombre de chaines, on trouve

```
 415  chaines pour  388  protéines

      367  protéines avec  1  chaine   soit 94.59 %
       18  protéines avec  2  chaines        4.64 %
        1  protéines avec  3  chaines        0.26 %
        1  protéines avec  4  chaines        0.26 %
        1  protéines avec  5  chaines        0.26 %
```

Les résultats pour l'ensemble de la PDB telle qu'elle était au début 2006 sont les suivants (analyse QT) :

```
Nombre de valeurs : 33919     structures
Moyenne           :     2.25  chaines
Ecart-type        :     2.73  chaines
Cdv               :   121     %
Minimum           :     1     chaine
Maximum           :    60     chaines
```

La répartition du nombre de chaines pour l'ensemble de la PDB telle qu'elle était au début 2006 est donnée par :

```
   76308 chaines pour 33919 structures rapport =  2.25

17156 structures  avec   1 chaine  ;  pct =  50.6 % exemple = 1A57 3GCT 1KYH
 9684 structures  avec   2 chaines ;  pct =  28.6 % exemple = 1F4O 1DL7 1SFX
 1529 structures  avec   3 chaines ;  pct =   4.5 % exemple = 1XLQ 1WA5 1D8H
 3236 structures  avec   4 chaines ;  pct =   9.5 % exemple = 1VBC 1K55 1Q3V
  192 structures  avec   5 chaines ;  pct =   0.6 % exemple = 1IM9 1PP6 1C6V
  828 structures  avec   6 chaines ;  pct =   2.4 % exemple = 1GUT 1WOG 1FZ6
   76 structures  avec   7 chaines ;  pct =   0.2 % exemple = 1MOF 1Q57 1XTC
  494 structures  avec   8 chaines ;  pct =   1.5 % exemple = 3TMK 1F6M 1XXX
  ...

  (les pourcentages qui restent sont inférieurs à 1 %).
```

Peut-on comparer statistiquement ces résultats ? Si oui, avec quels tests statistiques ?

Est-ce qu'intuitivement on peut dire que la DBDB ressemble globalement à la PDB ? Si oui, pourquoi, si non, pourquoi ?

Aucun calcul n'est exigé. Par contre la rédaction devra dépasser 10 lignes et être très précise et très soignée.

BIOINFORMATIQUE, Unité libre 2006 (1)
Partie "Statistiques" - Solution

Il n'est pas possible *stricto sensu* de comparer ces résultats car la DBDB est une partie de la PDB. S'il fallait comparer quelque chose, ce serait la DBDB et la PDB sans la DBDB. ce qui demande des calculs un peu "techniques" pour trouver la "vraie" moyenne de la PDB sans la DBDB, le "vrai" écart-type...

Quoiqu'il en soit, intuitivement, le nombre moyen de chaines a l'air très différent entre ces deux bases de données, la PDB ayant des protéines avec plus de chaines (presque deux fois plus). La variabilité du nombre moyen de chaines, exprimée par le CDV (coefficient de variation) semble également indiquer qu'il y a une dispersion nettement plus forte au niveau de la PDB que de la DBDB.

La répartition en nombres de chaines a l'ai très différent aussi : la DBDB n'a presque que des protéines à une seule chaine alors que la PDB contient pour moitié des protéines à une chaine et pour presque 30 % des protéines à deux chaines.

Il serait possible de définir une variable CNC (classe de nombre de chaines) par les formules : CNC=1 si la protéine n'a qu'une seule chaine, CNC=2 si la protéine a plus d'une chaine. On pourrait alors faire un test d'indépendance (χ^2) pour cette variable sur les deux populations après avoir effectué les tris à plat.