# Analyse des données en sciences de la Terre
# Partiel 3 : Statistique

9/12/2010

**Exercice 1 : distribution laplacienne (7 points)**

La densité de probabilité normalisée d'une distribution laplacienne de médiane $m$ et de paramètre d'échelle $b$ s'exprime de la manière suivante :

$$f(x) = \frac{1}{2b} \exp\left[-\frac{|x-m|}{b}\right].$$

1. En ajoutant une légende aux axes, représenter la densité de probabilité $g(x)$ d'une distribution laplacienne réduite ($m = 0$, $b = 1$) pour $x$ entre -10 et 10 par pas de 0,01.

```
clear all; close all; clc
dx=0.01;
x=(-10:dx:10)';
g=0.5*exp(-abs(x));
figure
subplot(2,1,1)
plot(x,g)
xlabel('x')
ylabel('g(x)')
```

2. Calculer sa fonction de répartition $h(x) = \int_{-10}^{10} g(x)dx$ et la représenter dans une sous-figure juste en dessous de la densité de probabilité. Mettre une légende aux axes.

```
h=cumsum(g*dx);
subplot(2,1,2)
plot(x,h)
xlabel('x') ; ylabel('h(x)')
```

3. Par une méthode de votre choix, calculer la probabilité d'être inférieur ou égal à $-b$.

```
p1=sum(g(x<=-1)*dx) % ou h(abs(x+1)<1e-5)
```

4. Par une méthode de votre choix, calculer la probabilité d'être supérieur à $+b$.

```
p2=sum(g(x>1)*dx) % ou 1-h(abs(x-1)<1e-5)
```

5. Calculer ou déduire des questions précédentes la probabilité d'être dans l'intervalle $]-b\ ;\ +b]$.

```
p3=1-p2-p1 % ou sum(g(x>-1 & x<=1)*dx) ou h(abs(x-1)<1e-5)-h(abs(x+1)<1e-5)
```

6. BONUS : La densité de probabilité est normalisée donc : $\int_{-\infty}^{+\infty} f(x)dx = 1$.

**Exercice 2 : comparaison de deux moyennes (6 points)**

1. Comment définit-on une fonction sous matlab ?

   ```
   function Out=MyFunction(In)
       Out=In;
   ```

   Comment s'y prend-on ensuite pour l'utiliser ?

   ```
   Y=MyFunction(X)
   ```

2. Ecrire une fonction nommée "comparaison" qui compare deux moyennes expérimentales à l'aide d'un test de Student. Cette fonction devra prendre en entrée deux arguments : la série de valeurs $x$ et la série de valeurs $y$. Elle devra fournir en sortie trois arguments : le nombre d'éléments de $x$, le nombre d'éléments de $y$ et la valeur de la statistique.

   ```
   function [nx,ny,ts]=comparaison(x,y)
       xb=mean(x);
       yb=mean(y);
       nx=length(x);
       ny=length(y);
       S=sqrt(((nx-1)*var(x)+(ny-1)*var(y))/(nx+ny-2)); % par exemple
       ts=abs(xb-yb)/S/sqrt(1/nx+1/ny); %par exemple
   ```

3. Utiliser cette fonction avec les vecteurs A et B puis calculer la valeur critique de la statistique pour $\alpha = 0.05$.

   ```
   [Nx,Ny,Ts]=comparaison(A,B)
   tc=tq(1-0.05,Nx+Ny-2);
   ```

4. BONUS : Vérifier à l'intérieur de votre fonction que les arguments d'entrée $x$ et $y$ sont bien des vecteurs colonnes.

   ```
   function [nx,ny,ts]=comparaison(x,y)
   if size(x,2)==1 & size(y,2)==1 % par exemple
       xb=mean(x);
       yb=mean(y);
       nx=length(x);
       ny=length(y);
       S=sqrt((sum(x-xb).^2+(sum(y-yb).^2))/(nx+ny-2)); % autre exemple
       ts=abs(xb-yb)/(S*sqrt(1/nx+1/ny)); % autre exemple
   else
       disp('x et y ne sont pas des vecteurs colonnes')
   end
   ```

**Exercice 3 : ANOVA et moindres carrés (7 points)**

L'ANOVA peut également être utilisée pour dire si un modèle de moindres carrés est pertinent pour expliquer les données. Au lieu de comparer la variance inter-groupe et la variance intra-groupe, on compare la variance du modèle (MSM) et la variance des résidus (MSD).

1. Soit le fichier "donnees.txt" qui contient en première colonne une série d'abscisses $x$ et en deuxième colonne une série d'ordonnées $y$. Lire les données et calculer un modèle $\widetilde{y}$ consistant en une tendance quadratique.

```
clear all; close all; clc
aa=load('donnees.txt');
x=aa(:,1);
y=aa(:,2);
clear aa
A=[0*x+1 x x.^2];
coeff=A\y;
ymod=A*coeff;
```

2. Calculer les deux sommes suivantes :

$$SSM = \sum_{i}^{N} (\widetilde{y_i} - \overline{y})^2$$

```
SSM=sum((ymod-mean(y)).^2);
```

$$SSD = \sum_{i}^{N} (y_i - \widetilde{y_i})^2$$

```
SSD=sum((y-ymod).^2);
```

3. Calculer les degrés de liberté suivants : $DFM = I - 1$ et $DFD = N - I$, où I correspond au nombres de paramètres du modèle et N correspond au nombres de données. $I$ et $N$ doivent être déterminés à partir des dimensions de la matrice normale des moindres carrés.

```
DFM=size(A,2)-1;
DFD=size(A,1)-size(A,2);
```

4. Calculer $MSM = SSM/DFM$, $MSD = SSD/DFD$, puis la statistique de Fisher $F = MSM/MSD$.

```
MSM=SSM/DFM
MSD=SSD/DFD
F=MSM/MSD
```

5. Si la statistique $F = 1.18$ et la statistique critique $F_c = 3.03$, on ne peut pas rejeter l'hypothèse nulle compte tenu du niveau de confiance qu'on s'est fixé ($\alpha = 0.05$, soit une probabilité de 5% de rejeter à tort l'hypothèse nulle). Si `fp(F,DFM,DFD)` donne par ailleurs 0.69, on obtient alors une probabilité de rejeter à tort l'hypothèse nulle de 31% (`1-fp(F,DFM,DFD)`), ce qui est supérieur au seuil généralement acceptable de 5%.

6. BONUS : Représenter sur le même graphique le modèle en bleu et les données en rouge.

```
figure
plot(x,y,'r')
hold on
plot(x,ymod,'b')
hold off
xlabel('x')
ylabel('y')
legend('Données','Modèle')
title(['Statistique de Fisher = ',num2str(F)])
```