Epreuve de modélisation (Option Proba/Stats)
Fonctions de répartition empirique, Kolmogorov-Smirnov

Expérience 1. Si la fonction de densité permet de caractériser la distribution d'une loi, la fonction de répartition contient une information intégrée qui contient également toute l'information sur la loi. Les caractéristiques de formes s'y lisent aussi graphiquement comme l'asymétrie, l'aplatissement, le poids des queues et l'existence de modes.

(1) **Quelques fonctions de répartitions :** Tracer pour chacune des lois suivantes, sur deux sous-figures (utiliser `subplot`), la fonction de densité et la fonction de répartition : $N(0,1)$, $C(0,1)$ (loi de Cauchy), $\chi^2(2)$, $\chi^2(3)$ et le mélange de deux gaussiennes de variance 1 et centrées en -2 et 2.

(2) **Théorème de Glivenko-Cantelli :** Le problème d'estimation de densité à partir d'un échantillon est un problème délicat en statistique. Le calcul d'histogramme permet d'estimer la densité intégrée sur des intervalles mais cela pose des problèmes (choix du nombre et de la position des boîtes). Le calcul de la fonction de répartition $F_n(t) = \frac{1}{n}\sum_{i=1}^n X_i \leq t$ est d'une utilisation plus directe tout en permettant une identification parfaite de la loi dans la limite d'un échantillon de taille infinie.

Illustrer le théorème de Glivenko-Cantelli sur un échantillon iid de loi $\mathcal{N}(0,1)$ en traçant sur un même graphique la fonction de répartition et la fonction de répartition empirique calculée en les valeurs prises par les données pour $n = 20$, $n = 100$ et $n = 1000$ (utiliser la fonction `sort` pour trier les données).

(3) **Test d'ajustement simple de Kolmogorov-Smirnov :** Tracer la courbe $t \to \sqrt{n}(F_n(t) - F(t))$ pour $t \in [0,1]$ où F est la fonction de répartition dans le cas d'un échantillon $\mathcal{N}(0,1)$ de taille $n = 100$ puis $n = 1000$.

Rappel : Lorsque F est continue, le processus $t \to \sqrt{n}(F_n(t) - F(t))$ pour $t \in [0,1]$ converge en loi vers un pont Brownien $t \to W_t = B_t - tB_1$ où B_t est un mouvement brownien. De plus, si

$$D_n = \sup_{t\in\mathbb{R}} |F_n(t) - F(t)|$$

alors

$$\lim_{n\to+\infty} P(\sqrt{n}D_n \leq t) = P(\sup_{t\in[0,1]} |W_t| \leq t) = \sum_{j=-\infty}^{+\infty} (-1)^j \exp(-2j^2t^2)^1\,,$$

ce qui conduit au test d'ajustement de Kolmogorov-Smirnov qui permet de tester $F = F_0$ contre $F \neq F_0$ sans hypothèse sur F autre que d'être continue (test non-paramétrique*). La valeur de $\sum_{j=-\infty}^{+\infty}(-1)^j \exp(-2j^2t^2)$ est fournie par la fonction* `pks` *du module* ***stixbox****.*

Les tests non paramétriques sont très utiles par leur généralité. Ils ne peuvent cependant être aussi efficace que les tests paramétriques si F est dans une famille paramétrique $(F_\theta)_{\theta\in\Theta}$ de loi.

Nous allons le vérifier en construisant un test de Kolmogorov Smirnov de $F = F_0$ contre $F \neq F_0$ où F_0 est la fonction de répartition d'une $\mathcal{N}(0,1)$. Pour gagner du temps, voici quelques lignes de code implémentant ce test (noter la fonction rapide d'obtenir D_n) :

[1] pour n fini, $n > 80$, $T = \mathbf{1}_{\mathbf{pks}(D_n(\sqrt{n}+0.12+0.11/\sqrt{n}))>1-\alpha}$ fournit un test de niveau très proche de α (cf BICKEL et DOKSUM p220)

```
function T=ksnorm(x,alpha)
% Test de Kolmogorov-Sirmnov contre une loi N(0,1) de niveau alpha
n=length(x);
sx=sort(x);
pn=pnorm(sx);
d=max([max((1:1:n)/n-pn) max(pn-(0:1:n-1)/n)]);
% T=(pks(sqrt(n)*d)>(1-alpha)); Test de niveau asymptotique alpha
T=(pks((sqrt(n)+0.12+0.11/sqrt(n))*d)>(1-alpha)); % correction BD p220
```

Utiliser le code précédent pour évaluer la probabilité de rejeter l'hypothèse $F = F_0$ sur un 100-échantillon de $\mathcal{N}(\mu_1, 1)$ où $\mu_1 = 0.36$. Comparer avec la valeur obtenue si on utilise le test paramétrique $\mu = 0$ contre $\mu \neq 0$ pour une famille gaussienne $\mathcal{N}(\mu, 1)$. *On verra que le test de Kolmogorov-Smirnov est effectivement moins puissant dans ce cas. Il faut plus de données à niveau égal pour avoir la même puissance que le test paramétrique.*

(4) **Test d'ajustement à une famille $F_0((. - \mu)/\sigma)$:** Le test d'ajustement simple est de portée limitée car généralement on cherche plutôt un ajustement à une famille contenant un ou deux paramètres d'échelle comme par exemple la loi gaussienne (on suppose que sous F_0, $E(X) = 0$, $V(X) = 1$). La méthode consiste à centrer et réduite les données à partir de la moyenne et de la variance empirique : $X_i \to \tilde{X}_i = (X_i - \overline{X}_n)/\hat{\sigma}$ où $\hat{\sigma} = \sqrt{\frac{1}{n-1}\sum_{i=1}^n \tilde{X}_i^2}$. On construit le test sur la statistique

$$\tilde{D}_n = \sup_t |\tilde{F}_n(t) - F_0(t)|$$

où $\tilde{F}_n$ est la fonction de répartition empirique des $\tilde{X}_i$. Sous l'hypothèse $H_0 : \exists(\mu, \sigma)\ tq\ F = F_0((. - \mu)/\sigma)$, la loi de $\tilde{D}_n$ ne dépend pas de μ et σ. Par contre le niveau asymptotique du test n'est pas donné analytiquement par une formule explicite. Il faut estimer la zone de rejet en calculant empiriquement le quantile $1 - \alpha$ en se plaçant sous l'hypothèse H_0 et en simulant des réalisations indépendantes de $\tilde{D}_n$. Voici une façon de calculer le seuil de rejet.

```
function q=qksnorm2(n,alpha)
% calcul du seul de rejet pour \tilde{D}_n dans un test d'ajustement
% sur une famille gaussienne de Kolmogorov-Smirnov
% estimation sur 10000;

nexp=10000;
d=zeros(1,nexp);
for i=1:nexp
  x=randn(1,n);
  x=(x-mean(x))/std(x);
  sx=sort(x);
  pn=pnorm(sx);
  d(i)=max([max((1:1:n)/n-pn) max(pn-(0:1:n-1)/n)]);
end
q=quantile(d,1-alpha); % calcul du quantile empirique (stixbox)
```

(a) Charger les données `data20` avec `getdata` et tester la normalité de la distribution de la quantité de papier jetée par les employés de banques et jetée par les salariés travaillant dans d'autres secteurs.

(b) Adapter l'approche précédente pour tester l'ajustement à la famille de loi exponentielle $\mathcal{E}_\lambda$ et appliquer la sur les données `data14` pour tester si les appels de voisinage et les appels nationaux suivent une loi exponentielle.

(5) **Test de comparaison de deux échantillons de Kolmogorov-Smirnov :** On rappelle que le test de comparaison de deux échantillons est construit sur la statistique

$$D_{m,n} = \sup_t |F_m(t) - G_n(t)|$$

où F_m (resp. G_n) est la fonction de répartition empirique du premier échantillon de taille m (resp du deuxième de taille n) pour laquelle :

$$\lim_{m,n\to+\infty} P(\sqrt{\frac{mn}{m+n}} D_{m,n} \leq t) = \sum_{j=-\infty}^{+\infty} (-1)^j \exp(-2j^2t^2)$$

(6) Prendre un échantillon de loi de Cauchy de taille 100 de paramètre 0 et 1 et faire un test de Kolmogorov-Smirnov contre un échantillon de loi $N(0,1)$. Evaluer la puissance du test dans ce cas. Cela vous semble-t-il satisfaisant ? Quelle est la faiblesse du test de KS dans ce cas ? Pouvez-vous imaginez un test beaucoup plus performant ?

(7) Faire la même chose avec deux échantillons de loi $\mathcal{N}(0,1)$ et $\mathcal{N}(0,2)$.

La "morale" des questions précédentes est que le test de Kolmogorov à tendance à être assez conservateur et à ne pas rejeter facilement l'hypothèse $F = F_0$. Aussi, ce test est vraiment informatif s'il rejette l'hypothèse H_0 car le niveau du test est connu.

(8) Charger les données `data3` avec `getdata` et tester par un test de KS l'égalité entre le salaire des hommes et des femmes (Utiliser la fonction `kstwo` de stixbox). *On pourrait ici faire aussi un test de Wilcoxon qui conduit à retenir (contre toute évidence !) l'égalité des distributions des salaires des hommes et des femmes.*

Expérience 2. Graphique quantile-quantile

(1) **Test de normalité :** Une façon de tester la normalité d'un n-échantillon est de tracer un graphique quantile-quantile (qq-plot en anglais). Il s'agit de comparer les quantiles d'une loi normale avec les quantiles empiriques obtenus sur le n-échantillon. Plus précisément, si F_0 est la fonction de répartition de la loi normale centrée réduite alors on trace dans le plan les points $(X_{(i)}, F_0^{-1}(i/(n+1)))$ où $(X_{(1)}, \cdots, X_{(n)})$ est la statistique d'ordre (valeurs classées par ordre croissant) du n-échantillon (cela correspond à estimer $F(X_{(i)})$ par $i/(n+1)$).

(a) Vérifier que si F est la fonction de répartition d'une variable $\mathcal{N}(\mu, \sigma^2)$, alors les points $t \to (t, F_0^{-1}(F(t)))$ sont sur la droite $y = (x-\mu)/\sigma$.

(b) Tracer les graphiques quantile-quantile pour un n-échantillon de loi $\mathcal{N}(0.5,1), \mathcal{N}(0,2)$ et $C(0,1)$ pour $n = 100$. Vérifier que dans les deux premiers cas, les points sont approximativement sur une droite (la droite de Henry) et qu'une estimation aux moindres carrés des paramètres permet de retrouver les paramètres de façon raisonnable. Vérifier que le troisième cas est visuellement très différent des deux autres.

(2) **Comparaison de deux échantillons :** La méthode s'étend à un test de comparaison de deux échantillons $(X_i)_{1\leq i\leq n}$ et $(Y_j)_{1\leq j\leq m}$. On supposera pour simplifier que $n = m$. Il s'agit de tester $\mathcal{L}(X) = \mathcal{L}(Y)$ contre $\mathcal{L}(X) \neq \mathcal{L}(Y)$. Pour cela l'idée est de comparer les valeurs des quantiles estimés par les statistiques d'ordre des deux échantillons.

La méthode des graphiques quantiles-quantiles consiste alors à tracer les points $(X_{(i)}, Y_{(i)})$ pour $1 \leq i \leq n$. Si les points sont plus ou moins répartis le long de la droite $y = x$, l'hypothèse H_0 est vraisemblable. Si les points se répartissent le long d'une droite, mais d'équation différente, alors on peut imaginer que les lois sont égales à un changement affine de variable près , i.e $\mathcal{L}(Y) = \mathcal{L}(a + bX)$. Sinon, les lois sont sans doute différentes. Il s'agit avant tout d'une méthode graphique qui permet une approche visuelle somme toute assez riche des données. Reprendre la question (8) de l'expérience précédente dans ce cadre.