Utiliser Octave pour les tests statistiques

Alexandre Vial - alexandre.vial@utt.fr

16 novembre 2004

Introduction

Octave est un logiciel libre en grande partie compatible avec le logiciel commercial Matlab^{TM}. Il est disponible sur le site http://www.octave.org. Pour tracer les graphiques, il utilise le logiciel Gnuplot (http://www.gnuplot.info). De nombreuses fonctions statistiques sont intégrées à Octave, seules certaines d'entre elles seront présentées et utilisées dans ce document.

1 Moyenne, écart-type sans biais, valeurs centrées réduites

Soit un fichier `mesures.dat` contenant N valeurs, on charge le fichier en mémoire et on trie les valeurs par ordre croissant :

Code 1 Chargement des valeurs

```
load mesures.dat
valeurs=sort(mesures);
```

On peut alors très simplement calculer la valeur moyenne des valeurs, l'écart-type sans biais, les valeurs centrées réduites ainsi que l'incertitude-type de type A.

Code 2 Calcul de la moyenne, de l'écart-type,...

```
%Moyenne
moyenne=mean(valeurs)
% Ecart-type sans biais
etsb=std(valeurs)
% Valeurs centrées réduites
vcr=studentize(valeurs)
%N ombre d elements
n=max(size(vcr))
% Incertitude de type A
ua=etsb/sqrt(n)
```

2 Test du χ^2

Pour effectuer un test du χ^2, il faut tout d'abord déterminer le nombre de classes.

Code 3 Calcul du nombre de classes

```
% Decoupage en classes
ncl=round(sqrt(n))
%ncl=round(1+3.3*log10(n))
```

On détermine ensuite les bornes de chacun des intervalles :

Code 4 Calcul des bornes

```
% Definitions des intervalles
tt=min(vcr):(max(vcr)-min(vcr))/ncl:max(vcr)
```

On calcule alors les fréquences pour chacune des classes.

Code 5 Calcul des fréquences

```
%calcul des frequence
for i=1:ncl-1
  fm(i)=max(find(vcr<=tt(i+1)));
end
% Traitement special pour la derniere valeur
% qui est incluse dans l'intervalle
fm(ncl)=n;
for i=ncl:-1:2
  fm(i)=fm(i)-fm(i-1);
end
fm=fm/n
```

On cherche les fréquences théoriques. Celles-ci sont obtenues à partir de la fonction de cumul de la loi normale (figures 1 et 2).

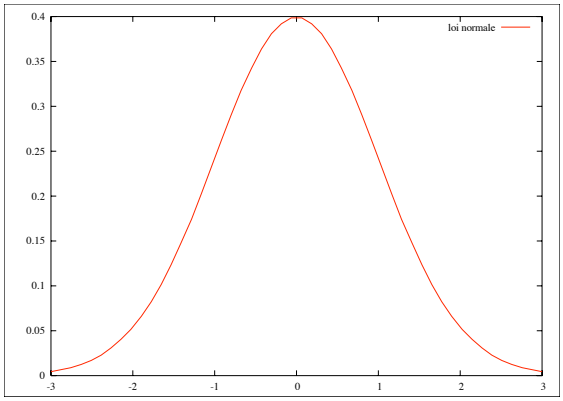

FIG. 1 – *Loi normale, obtenue avec la fonction* `stdnormal_pdf(x)`

Code 6 Calcul des fréquences théoriques

```
%calcul des frequences theoriques
for i=1:ncl
  ft(i)= stdnormal_cdf(tt(i+1))-stdnormal_cdf(tt(i));
end
ft
```

On peut alors calculer la valeur totale χ^2, et en déduire la probabilité α de commettre une erreur en rejetant l'hypothèse gaussienne.

Connaissant χ^2 et le nombre de degrés de liberté $L = N_{Cl} - 3$, on est donc capable de calculer α. Octave fournit également la fonction `chisquare_inv` qui calcule χ^2 à partir

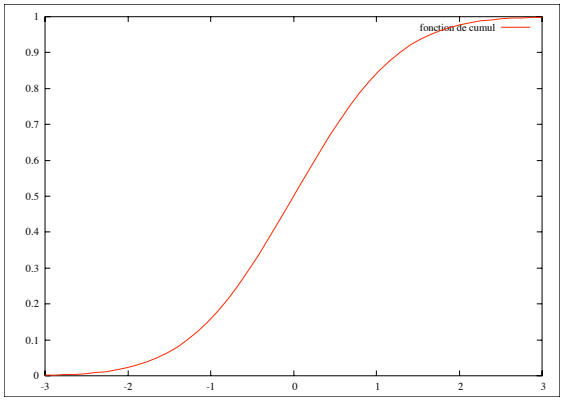

FIG. 2 – *Fonction de cumul, obtenue avec la fonction* `stdnormal_cdf(x)`

Code 7 Calcul du χ^2 et de la probabilité α

```
chi2=sum((fm-ft).^2./ft)*n

alpha=1-chisquare_cdf(chi2,ncl-3)
```

de L et α (c'est cette fonction qui peut être utilisée pour remplir la table du χ^2). Les fonctions implémentées dans Octave travaillent en fait sur la quantité $1-\alpha$, on aura donc $1-\alpha =$`chisquare_cdf`(χ^2,L) et $\chi^2 =$`chisquare_inv`$(1-\alpha,L)$.

3 Test de Smirnov-Kolmogorov

Pour le test de Smirnov-Kolmogorov, on calcule tout d'abord les cumulants, en tenant compte du cas particulier des valeurs multiples.

Code 8 Calcul des cumuls

```
cumul(1)=1/n;
for ii=2:n
  cumul(ii)=ii/n;
  if (vcr(ii)==vcr(ii-1)) % si on a la meme valeur
    cumul(ii-1)=stdnormal_cdf(vcr(ii-1)); % on met la valeur theorique
    %pour l element precedent afin d annuler l ecart
  endif
endfor
```

Le cumul théorique (figure 2) est ensuite calculé, afin d'obtenir l'écart avec le cumul obtenu précédemment, et l'écart maximal de $\|F - F_n^*\|$ est recherché.

Code 9 Calcul des cumuls théoriques, écarts et écart maximal

```
cumult=stdnormal_cdf(vcr);
% VCR, cumul, cumul theorique
[vcr cumul cumult]
%Ecart
ecart=abs(cumult-cumul)
%Ecart maximal
max_ecart=max(ecart)
```

Il nous faut à présent, comme pour le test du χ^2, calculer la probabilté α_{SK} de commettre une erreur en rejetant l'hypothèse gaussienne. Pour cela, Octave nous propose deux fonctions. La première fonction est `kolmogorov_smirnov_cdf(x)`. Or la table disponible dans la littérature donnant α_{SK} en fonction du nombre de mesures et de l'écart maximal des cumulants comporte deux entrées. L'explication est que la fonction définie par Octave n'est que la formule asymptotique pour le calcul de la probabilité, prenant comme argument $\sqrt{n}.max\|F-F_n^*\|$, cette formule étant valable pour un nombre de mesures n supérieur à 35 ou 100 selon les auteurs.

$$\lim_{n\to\infty} P(\sqrt{n}D_n \geq \epsilon) \sim 2\sum_{k=1}^{\infty}(-1)^{k-1}e^{-2k^2\epsilon^2} \tag{1}$$

La deuxième fonction présente dans Octave s'utilise de la manière suivante : `[`α_{SK}`, ks] = kolmogorov_smirnov_test(vcr, "normal", "<>")` . Elle prend donc directement les valeurs centrées réduites des mesures en argument, et retourne α_{SK} et ks=$\sqrt{n}.max\|F-F_n^*\|$. Cette deuxième fonction est donc légèrement plus directe que la première présentée, mais elle souffre du même défaut.

La solution consiste donc à implémenter la véritable fonction donnant α_{SK}. Cette fonction est définie par la formule suivante :

$$P(D_n \geq \epsilon) = 2\sum_{k=0}^{|n.(1-\epsilon)|} C_n^k \epsilon(\epsilon+\frac{k}{n})^{k-1}.(1-\epsilon-\frac{k}{n})^{n-k} \tag{2}$$

où $|x|$ représente la partie entière de x.

On va donc créer la fonction `skbl.m` (pour Smirnov-Kolmogorov bilatéral).

Code 10 Fonction "Smirnov-Kolmogorov bilatéral"

```
% Smirnov-Kolmogorov bilateral
% Alexandre Vial - 24/01/2003
% alexandre.vial@utt.fr

% this function needs the number of values "n"
% and the max difference "alpha"
% it returns p_ks=P(Dn > alpha)
function p_ks=skbl(n,alpha)

  kmax=n*(1-alpha);
  k=0:floor(kmax);
  t1=gamma(n+1)./(gamma(k+1).*gamma(n-k+1));
  t2=(alpha+k/n).^(k-1);
  t3=(1-alpha-k/n).^(n-k);
  p_ks=sum(2*t1*alpha.*t2.*t3);
```

Il est donc à présent trivial de calculer α_{SK} en faisant appel à cette routine.

Code 11 Calcul de α_{SK}

```
% Calcul de p
p=skbl(n,max_ecart)
```

On peut remarquer que si l'on sait à présent calculer α_{SK} à partir de n et de l'écart maximal, on ne connaît pas la fonction inverse permettant d'obtenir la table de Smirnov-Kolmogorov. En fait, cette fonction n'existe pas, on est obligé de procéder par dichotomie. L'algorithme suivant

permet de calculer la table de Smirnov-Kolmogorov telle que publiée dans [Mesure physique et instrumentation, D. Barchiesi, Ellipses, 2003].

Code 12 Calcul de la table de Smirnov-Kolmogorov

```
% creation de la table de SK
% Alexandre Vial - 02/04/2003
p_line=[0.2 0.1 0.05 0.025 0.02 0.01 0.005];
n_max=100;
n_line=1:n_max;
tableau=zeros(n_max+1,8);
tableau(1,2:8)=p_line;
for n=n_line
  tableau(n+1,1)=n;
  colonne=1;
  for p=p_line
    colonne=colonne+1;
    dn=1; % max value
    delta=0.5;
    res=skbl(n,dn);
    while (res>1.0001*p  |  res<0.9999*p)
      if (res>1.0001*p)
        dn=dn+delta;
      end
      if (res<0.9999*p)
        dn=dn-delta;
      end
      delta=delta/2;
      res=skbl(n,dn);
    end
    tableau(n+1,colonne)=dn;
  end
  n
  tableau(n+1,2:8)
end
```

On peut remarquer que pour des valeurs de $\epsilon = max\|F - F_n^*\|$ faibles, l'équation 2 donne un résultat supérieur à 1 !

4 Intervalles de confiance

4.1 Pour une nouvelle mesure

On cherche à déterminer l'intervalle dans lequelle on aura P% de chance d'obtenir une nouvelle mesure. Si l'on suppose la répartition des mesures gaussiennes, alors l'intervalle est : $[\bar{m} - k\sigma; \bar{m} + k\sigma]$ avec $\bar{m}$ la valeur moyenne des mesures, σ l'écart-type sans biais et k le facteur d'élargissement obtenu à l'aide du code suivant :

Code 13 Facteur d'élargissement pour l'obtention d'une nouvelle mesure

```
pic=P/100;
% Pour une nouvelle mesure
k1=stdnormal_inv((1+pic)/2)
```

On pourra vérifier que `pic=stdnormal_cdf(k1)*2-1`.

4.2 Pour la valeur mesurée

On cherche à déterminer l'intervalle dans lequelle on aura P% de chance d'obtenir la valeur mesurée. Si l'on suppose la répartition des mesures gaussiennes, alors la valeur moyenne suit la loi de Student (figure 3) et l'intervalle est : $[\bar{m} - k\frac{\sigma}{\sqrt{n}}; \bar{m} + k\frac{\sigma}{\sqrt{n}}]$ avec $\bar{m}$ la valeur moyenne des mesures, σ l'écart-type sans biais, n le nombre de mesures et k le facteur d'élargissement obtenu à l'aide du code suivant (le calcul fait apparaître la quantité $n-1$ qui est le degré de liberté L) :

Code 14 Facteur d'élargissement pour la valeur mesurée

```
% Pour la moyenne
k2=t_inv((1+pic)/2,n-1)
```

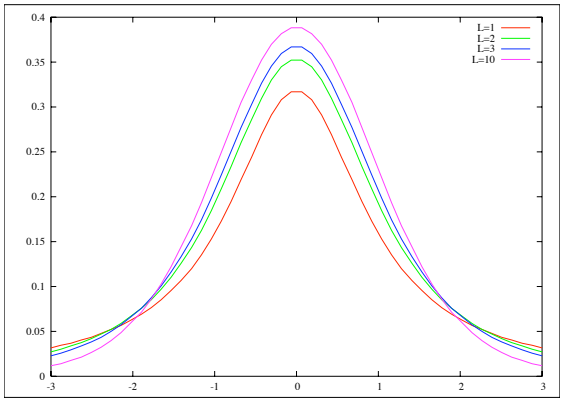

FIG. 3 – *Loi de Student, obtenue avec la fonction* `t_pdf(x,L)`

5 Formules mathématiques

Les fonctions de répartitions ou de cumul présentées dans les paragraphes précédents sont calculables directement à l'aide des formules qui suivent. Par contre, comme dans le cas de la distribution de Smirnov-Kolomogorov, les fonctions inverses font appel à des algorithmes plus complexes qu'un calcul d'intégrale, et ne sont pas décrites ici.

Moyenne : $\bar{m} = \sum_{i=1}^{N} m_i$.

Écart-type sans biais : $\sigma_N = \sqrt{\frac{\sum_{i=1}^{N}(m_i - \bar{m})^2}{N-1}}$

Valeur centrée réduite : $\frac{m_i - \bar{m}}{\sigma}$

`stdnormal_pdf(x)` : $\frac{1}{\sqrt{2\pi}} e^{-x^2/2}$

`stdnormal_cdf(x)` : $\frac{1}{2}(1 + erf(x/\sqrt{2}))$ avec $erf(z) = \frac{2}{\sqrt{\pi}} \int_{t=0}^{z} e^{-t^2} dt$, soit après changement de variable $\tau = t\sqrt{2}$ pour le calcul de l'intégrale $erf(z) = \frac{1}{2} + \frac{1}{\sqrt{2\pi}} \int_{\tau=0}^{z} e^{-\tau^2/2} d\tau$

`t_pdf(x,n)` : $\frac{e^{-\frac{1}{2}(n+1)\ln(1+x^2/n)}}{\sqrt{n}\beta(n/2,1/2)}$ avec $\beta(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$ et $\Gamma(z) = \int_{t=0}^{\infty} t^{z-1} e^{-t} dt$

6 Conclusion

Il est possible d'améliorer les codes présentés en rajoutant le tracé des histogrammes, des cumulants,...

7 Version

`$Id: stat_octave.tex,v 1.7 2004/11/16 13:10:27 alex Exp $`

Liste des algorithmes

Table des figures