Université Claude Bernard Lyon 1
Licence MASS UE32
Année 2004-2005

fiche de TP n° 4 : Tests et simulations

Lancer SAS et l'aide de SAS > `Démarrer > Program Files > The SAS system >....`

Si vous le souhaitez, créez une librairie permanente à l'aide de la commande `libname`.

Au fur et à mesure du TP, faites le tri des sorties SAS que vous souhaitez conserver. Vous pouvez sauvegarder la fenêtre `output` en vous plaçant dessus et en choisissant `Save...` dans le menu `File`. Le fichier obtenu est de format `lst`, c'est-à-dire liste. Vous pourrez ensuite l'ouvrir avec un éditeur de texte.

Le compte-rendu est à rendre impérativement au plus tard le 10 décembre 2004 dans la boîte à lettres LaPCS (dans le hall d'entrée), imprimé (pas de fichier informatique). La note comptera pour moitié dans la note de contrôle continu.

Le compte-rendu doit comprendre les réponses aux questions numérotées Q1 à Q12. Ces réponses doivent être rédigées. Il ne doit en aucun cas s'agir d'une compilation de sorties SAS : elles doivent au minimum être commentées !

Des tables statistiques sont à votre disposition.

1 Intervalles de confiance

1.1 Risque, p-value...

Commencez par simuler un échantillon uniforme sur $[0, 1]$, de taille 100.

```
data simul1 ;
do i=1 to 100 ;
y=ranuni(-1) ;
output ;
end ;
keep y ;
run ;
proc univariate data=simul1 all alpha=0.05 mu0=0.5 ;
run ;
```

Observez attentivement la sortie produite. Faites tourner le programme plusieurs fois en modifiant `alpha`.

Q1. Choisissez l'un des échantillons produits etdonnez la valeur de la moyenne, de la médiane, des premier et troisième quartiles de cet échantillon.

Q2. Tracez le box-plot et un histogramme.

Q3. Pourquoi spécifie-t-on mu0=0.5 ?

Consultez plus précisément la rubrique `Tests for location`.

Q4. Que signifient les `p-value` ? Dans quels intervalles doivent être leurs valeurs pour ne pas rejeter l'hypothèse « la moyenne théorique de la loi de l'échantillon est 0.5 », pour un risque de 5% ?

Q5. Est-il légitime d'utiliser un test de Student ?

Q6. Pour votre échantillon, concluez-vous que l'échantillon relève d'une loi de moyenne 0.5 ?

Q7. Quel intervalle de confiance de la moyenne théorique de l'échantillon obtenez-vous pour un risque de 5% ?

1.2 Un fichier pour les statistiques produites

Simulez maintenant cent échantillons uniformes sur $[0, 1]$ de taille 1000 : la procédure `univariate` traite les cent échantillons successivement mais n'affiche pas les résultats. Les T_{obs} sont stockés dans la table SAS `uni_sum`.

```
data simul2 ;
array sim{100} y1-y100 ;
do i=1 to 1000 ;
      do j=1 to 100 ;
      sim{j}=rannor(-1) ;
      end ;
output ;
end ;
drop i j ;
run ;
proc univariate data=simul2 all noprint ;
var y1-y100 ;
output out=uni_sum T=t1-t100 ;
run ;
```

```
proc print data=uni_sum ;
run ;
```

Q8. On effectue des tests de Student sur la moyenne des cent échantillons fournis par `simul2` pour tester si la moyenne est nulle, avec un risque de 5%. Dans quel intervalle doit être t pour que l'hypothèse soit acceptée ?

Q9. Comptez (par exemple à la main) le nombre d'échantillons pour lesquels la moyenne empirique n'est pas significativement différente de zéro, pour un niveau de risque de 5%. Et pour un risque de 10% ?

2 Tests du χ^2

2.1 Test d'ajustement avec une loi discrète

La réalisation automatique d'un test du χ^2 d'ajustement n'est pas une chose simple à mettre en place : l'utilisateur doit choisir la loi à tester, les classes de répartition, les paramètres à évaluer. Le nombre d'options à fixer est donc très important... Ceci explique que SAS, comme la plupart des logiciels statistiques, ne propose pas de test du χ^2 général.

La procédure `freq` permet de construire des tables dénombrant le nombre de fois où une valeur est prise par un échantillon, puis de calculer des distances de Pearson.

Nous allons mettre en place un test du χ^2 pour un échantillon de 1200 lancers de dés.

Lancer Internet Explorer, allez sur la page `http://lapcs.univ-lyon1.fr/~duheille/tp-sas.html` et sauvegardez sur votre compte le fichier `des.csv` à l'aide du bouton de droite de la souris (enregistrer la cible sous).

Importez-le dans SAS en une table `des` en utilisant `Import data...` Ouvrez la table `des` que vous avez créée. Vérifiez qu'elle a deux colonnes.Fermez-la. On commence par éliminer la première colonne qui comporte simplement le numéro d'ordre :

```
data des ;
set des ;
drop num ;
run ;
```

Procédons ensuite au test du χ^2 d'adéquation avec la loi uniforme sur $\{1, \ldots, 6\}$.

```
proc freq data=des ;
tables x / chisq ;
run ;
```

Procéder alors à un deuxième test du χ^2 :

```
proc freq data=des ;
tables x / chisq testp=(0.2 0.2 0.1 0.2 0.2 0.1) ;
run ;
```

Les probabilités de la loi à tester sont données dans la variable `testp`, et leur total doit faire 1 (ou 100 : on peut rentrer les pourcentages au lieu des probabilités).

Q10. Quelle est la valeur de la distance de Pearson dans chacun des deux cas ? L'échantillon est-t-il susceptible de relever d'une loi uniforme sur $\{1, 2, 3, 4, 5, 6\}$ pour un risque 5% ? Et de la loi donnée par `testp` ?

2.2 Test d'ajustement avec une loi à densité

Cette section est facultative et n'est pas notée.

Voici un exemple de test du χ^2 d'ajustement avec la loi uniforme sur $[0, 1]$.

Les principales étapes de ce programme sont : simulation d'un échantillon, création d'un histogramme dont les classes sont spécifiées et enfin calcul de la distance de Pearson en utilisant les résultats de l'histogramme « à la main ».

```
data simul ;
do i=1 to 100 ;
y=ranuni(-1) ; output ; end ;
keep y ;
run ;
proc univariate data=simul all ;
run ;
proc capability data=simul ;
histogram / midpoints= 0.025 to 0.975 by 0.05
```

```
          OUTHISTOGRAM = sortie ;
run ;
data test ;
set sortie ;
s =_n_ ;
retain cumul ;
if _n_=1 then cumul=(_OBSPCT_-5)**2/5 ;
else cumul=cumul+(_OBSPCT_-5)**2/5 ;
drop s ;
run ;
proc print data=test ;
run ;
```

Attention : la variable `_OBSPCT_` contient le pourcentage de valeurs de l'échantillon dans l'intervalle en question, et non l'effectif empirique. Ici, comme la taille de l'échantillon est 100, ces deux valeurs sont égales.

Pour faire un test du χ^2 avec SAS, il faut donc aider le logiciel : lui donner les classes, lui faire faire l'histogramme, ce qui fournit les effectifs empiriques, et calculer les effectifs théoriques. Cette dernière opération peut être faite à la main, ou par SAS qui connaît naturellement les lois statistiques usuelles et est capable de donner des valeurs numériques de leurs fonctions de répartition.

2.3 Test du χ^2 d'indépendance

Ce test permet de dire si deux caractères d'une population sont indépendants ou liés. Dans cet exemple, on a classé une population suivant sa région d'origine (1 ou 2) et suivant la couleur des cheveux et des yeux. La variable `count` contient le nombre de personnes de l'échantillon habitant une région donnée (deux régions possibles) et dont la couleur des cheveux et des yeux est aussi précisée.

```
proc freq data=color ;
weight count ;
tables eyes hair eyes*hair/out=freqcnt outexpect sparse ;
title 'Couleur des yeux et des cheveux' ;
run ;
proc print data=freqcnt noobs ;
title2 'PROC FREQ' ;
run ;
proc freq data=color order=data ;
weight count ;
tables eyes*hair /chisq expected cellchi2
norow nocol ;
output out=chisqdat pchi lrchi n ;
title 'Tests du chi2 pour la table 3*5 des couleurs des yeux et des cheveux' ;
run ;
proc print data=chisqdat noobs ;
title2 'PROC FREQ ' ;
run ;
```

Q11. Pour un risque de 5%, les couleurs des yeux et des cheveux sont-elles liées ?

Q11bis. La couleur des yeux influence-t-elle celle des cheveux ?

On peut également tester l'homogénétié à l'intérieur d'une région. Pour cela, créons une table `region1` extrayant les données relatives à la région 1 :

```
data region1 ;
set color ;
if region=1 ; drop region ; run ;
```

puis les données relatives à la région 2 :

```
data region2 ;
set color ;
if region=2 ; drop region ; run ;
```

Il est alors possible d'appliquer les mêmes procédures que ci-dessus à `region1` et `region2`.

Q12. Toujours pour un risque de 5%, les couleurs des yeux et des cheveux des habitants de la région 1 sont-elles liées ? Et dans la région 2 ?