T.P. SAS numéro 10

On continue à explorer ici les possibilités de simulation de SAS.

Exercice 10.1

On peut faire calculer par SAS des probabilités du type $P[X \leq x] = F(x)$, où X est une variable aléatoire réelle (v.a.r.) suivant une loi de probabilité classique, de fonction de répartition F ; selon le cas, x peut être entier, réel positif ou réel quelconque.

Les commandes SAS sont :
`poisson`(λ, x) pour une loi de Poisson ;
`probbeta`(x, p, q) pour une loi béta ;
`probbnml`(p, n, x) pour une loi binomiale ;
`probchi`(x, n) pour une loi de khi-deux ;
`probf`(x, n_1, n_2) pour une loi de Fisher (encore appelée loi de Snedecor) ;
`probgam`(x, a) pour une loi gamma (voir des précisions sur cette loi dans l'exercice 10.4) ;
`probhypr`(N, K, n, x) pour une loi hypergéométrique ;
`probnegb`(p, n, x) pour une loi binomiale négative ;
`probnorm`(x) pour une loi normale réduite ;
`probt`(x, n) pour une loi de Student.

Faire un essai pour les lois de Poisson, binomiale, de khi-deux, de Fisher, normale réduite et de Student (on notera que les degrés de liberté ne sont pas nécessairement entiers dans ce cas là ; se reporter à la remarque 8 de la feuille de T.P. 11).

Exercice 10.2

SAS fournit également les quantiles (ou fractiles) des principales lois de probabilités continues. Ainsi, pour un α donné dans $]0, 1[$ et pour une loi continue X spécifiée, SAS fournit la valeur x_α telle que $P[X \leq x_\alpha] = \alpha$ (autrement dit, $x_\alpha = F^{-1}(\alpha)$). SAS est donc susceptible de produire les tables statistiques des lois continues usuelles.

Les commandes sont :
`betainv`(α, p, q) pour la loi béta ;
`cinv`(α, n) pour la loi de khi-deux ;
`finv`(α, n_1, n_2) pour la loi de Fisher ;
`gaminv`(α, a) pour la loi gamma ;
`probit`(α) pour la loi normale réduite ;
`tinv`(α, n) pour la loi de Student.

Retrouver certaines valeurs typiques de vos tables pour les lois suivantes : normale réduite, Student, khi-deux et Fisher (ou Snedecor).

Exercice 10.3

Préambule

Dans la littérature statistique, X désignant une certaine v.a.r., on appelle p-valeur (*p-value* dans la littérature anglo-américaine) une quantité du type $P[X \leq x]$ ou $P[X \geq x]$, pour un x donné ; il s'agit d'une probabilité, d'où le terme de p-valeur. Ce sont de telles quantités que nous avons déterminées dans l'exercice 10.1.

Par ailleurs, on appelle q-valeur (ou *q-value*) des valeurs x_α telles que $P[X \leq x_\alpha] = \alpha$, pour un α donné dans $]0, 1[$; il s'agit d'un quantile, d'où le terme de q-valeur. Ce sont de telles quantités que nous avons déterminées dans l'exercice 10.2.

Des graphiques de type nuages de points représentant des p-valeurs ou des q-valeurs sont ainsi appelés des *"p-p plots"*, des *"p-q plots"* ou des *"q-q plots"*.

Le plus célèbre de ces graphiques est la droite de Henry, *q-q plot* représentant les quantiles théoriques de la loi normale réduite (en ordonnées) et les quantiles empiriques d'une série d'observations (en abscisses), ce qui est une façon simple d'étudier l'ajustement de la loi normale à ces observations.

Illustration

Copier les fichiers ci-dessous

```
~baccini/tpsas/exolim/exo10_3a.sas
~baccini/tpsas/exolim/exo10_3b.sas
```

Les mettre en œuvre et commenter.

Exercice 10.4

Dans cet exercice, on va voir comment utiliser SAS pour mettre en œuvre un test statistique difficilement réalisable sans logiciel.

Rappels sur la loi gamma

- Première paramétrage.
  La plupart des ouvrages de statistique présentent la loi gamma sous la forme suivante. Étant donnés 2 nombres réels strictement positifs p et θ, on appelle loi gamma de paramètres p et θ la loi d'une v.a.r. continue X dont la densité de probabilité s'écrit

$$f(x) = \begin{cases} 0 & \text{si } x \leq 0, \\ \dfrac{\theta^p}{\Gamma(p)} x^{p-1} \mathrm{e}^{-\theta x} & \text{si } x > 0, \end{cases} \tag{1}$$

  où $\Gamma(p)$ est la fonction (dite fonction eulérienne de deuxième espèce, ou seconde fonction d'Euler) définie par :

$$\Gamma(p) = \int_0^{+\infty} x^{p-1} \mathrm{e}^{-x} dx \qquad (p > 0).$$

  Cette loi est en général notée $\Gamma(p, \theta)$ et l'on peut vérifier : $\mathbb{E}(X) = \dfrac{p}{\theta}$; $\mathrm{Var}(X) = \dfrac{p}{\theta^2}$.
- Deuxième paramétrage.
  On trouve également une autre paramétrage de la loi gamma, plus commode lorsqu'on réalise des calculs au moyen d'un logiciel. Étant donnés 2 nombres réels strictement positifs a et b, on appelle loi gamma de paramètres a et b la loi d'une v.a.r. continue X dont la densité de probabilité s'écrit :

$$f(x) = \begin{cases} 0 & \text{si } x \leq 0, \\ \dfrac{1}{\Gamma(a)} \dfrac{1}{b^a} x^{a-1} \mathrm{e}^{-x/b} & \text{si } x > 0 \end{cases} \tag{2}$$

  (on a remplacé p par a et θ par $\dfrac{1}{b}$).
  Dans ce paramétrage, a est appelé la paramètre de forme (*shape parameter*) et b le paramètre d'échelle (*scale parameter*).
  La loi est en général notée $G(a, b)$ et l'on obtient : $\mathbb{E}(X) = ab$; $\mathrm{Var}(X) = ab^2$.
- Loi gamma normalisée (ou standardisée).
  C'est la loi gamma dans laquelle on pose, suivant la paramétrage choisi, $\theta = 1$ ou $b = 1$.
  Pour les valeurs positives de x, la densité de la loi gamma normalisée s'écrit, selon (2),

$$f(x) = \frac{1}{\Gamma(a)} x^{a-1} \mathrm{e}^{-x}$$

  et l'on obtient : $\mathbb{E}(X) = \mathrm{Var}(X) = a$.

– Propriété 1.
Si X est distribuée suivant une loi gamma normalisée $G(a,1)$ et si l'on pose $Y = bX$, b étant un réel strictement positif quelconque, alors Y est distribuée suivant une loi $G(a,b)$.
Cette propriété est utilisée par SAS qui ne simule que des lois $G(a,1)$. Pour obtenir une loi $G(a,b)$, il suffit de multiplier le résultat obtenu par b.
– Loi exponentielle.
En posant $p = a = 1$, on obtient la loi exponentielle de paramètre θ ou b.
Pour les valeurs positives de x, la densité s'écrit

$$f(x) = \theta \mathrm{e}^{-\theta x}, \text{ selon (1), ou } f(x) = \frac{1}{b}\mathrm{e}^{-x/b}, \text{ selon (2)} \quad (\Gamma(1) = 1).$$

Il vient respectivement : $\mathbb{E}(X) = \frac{1}{\theta}$; $\mathrm{Var}(X) = \frac{1}{\theta^2}$; ou encore : $\mathbb{E}(X) = b$; $\mathrm{Var}(X) = b^2$.
– Propriété 2.
Si n v.a.r. X_i $(i = 1, \ldots, n)$ sont indépendantes et distribuées selon des lois gamma $\Gamma(p_i, \theta)$ (resp. $G(a_i, b)$), alors $\sum_{i=1}^n X_i$ est $\Gamma(\sum_{i=1}^n p_i, \theta)$ (resp. $G(\sum_{i=1}^n a_i, b)$).
– Conséquence.
Si maintenant les X_i sont indépendantes et identiquement distribuées (i.i.d.) selon une loi exponentielle de paramètre b (on va désormais utiliser le paramétrage (2)), alors $\sum_{i=1}^n X_i$ suit une loi $G(n,b)$.

Mise en œuvre d'un test sur le paramètre de la loi exponentielle

Considérons $(X_1, \ldots, X_n)$ n v.a.r. i.i.d. selon une loi exponentielle de paramètre $b > 0$. On veut tester, avec un niveau $\alpha \in]0,1[$, l'hypothèse nulle $\{H_0 : b = b_0\}$ contre l'alternative $\{H_1 : 0 < b < b_0\}$. On sait qu'il existe un test U.M.P. (uniformément le plus puissant) de règle de décision : rejet de $H_0 \Longleftrightarrow \sum_{i=1}^n X_i < c$.

Comme on a $\alpha = P\,[\sum_{i=1}^n X_i < c \mid b = b_0]$ et que la statistique de test $\sum_{i=1}^n X_i$ est $G(n, b_0)$ sous H_0, on en déduit que c est le quantile d'ordre α d'une loi $G(n, b_0)$, loi entièrement spécifiée (lorsque n et b_0 sont donnés), mais non tabulée.

Il n'est donc possible de déterminer c (autrement dit de faire le test) qu'en utilisant un logiciel susceptible de simuler des lois gamma : c'est le cas de SAS.

En fait, SAS offre 3 possibilités :
– détermination directe du quantile c au moyen de la commande `gaminv`(α, n) ;
– détermination approchée de c au moyen d'une simulation avec la commande `rangam`$(-1, n)$;
– détermination de la probabilité associée à la valeur observée x de la statistique $\sum_{i=1}^n X_i$ au moyen de la commande `probgam`(x, n) (principe de la *p-value*, mais en sens opposé).

Application

On donne $n = 130$ et $b_0 = 1,5$.
– Déterminer la valeur de c pour un niveau $\alpha = 5$ %.
– Faire une simulation de taille 10 000, en déduire une valeur approchée de c et comparer (ne faire afficher que l'unique valeur intéressante).
– Déterminer, sous H_0, la probabilité associée à la valeur $\sum_{i=1}^n X_i = 180$ (attention : $P[G(n, b_0) < 180] = P[G(n,1) < \frac{180}{b_0}]$). En retrouver une approximation au moyen de la simulation précédente.

Calcul de la puissance du test

Pour tout b dans l'intervalle $]0, b_0[$, la fonction puissance du test s'écrit :

$$\pi(b) = P[\sum_{i=1}^n X_i < c \mid b] = P[G(n,b) < c] = P[G(n,1) < \frac{c}{b}].$$

En supposant toujours $n = 130$ et $b_0 = 1,5$, utiliser SAS pour calculer $\pi(b)$ pour les valeurs suivantes de b : 1,4 ; 1,3 ; 1,2 ; 1,1 ; 1. Commenter.