Logiciels de statistiques : TP n°10, TP-R-2
Renaud Mozet – 1er avril 2019

Tests sous R

1 – Notion de test

Nous avons :

- un échantillon fini,
- deux hypothèses alternatives (si l'une est vraie, l'autre est fausse, et réciproquement) sur la loi dont est tirée cet échantillon,
- une décision à prendre au sujet de ces hypothèses.

Nous pouvons faire deux erreurs différentes : accepter l'une des hypothèses comme vraie, alors que c'est l'autre qui l'est, et vice-versa. Quelle que soit la procédure suivie, il y aura toujours un risque de se tromper, dans un sens ou dans l'autre. On choisira la procédure de façon à réduire le plus possible les risques. Comme il y en deux, il sera généralement impossible de réduire les deux à la fois. Il faudra donc choisir, en fonction des conséquences des deux erreurs possibles.

Généralement, on note l'une des hypothèses H_0, l'autre H_1. On nomme "fiabilité" la probabilité d'avoir raison en décidant que H_0 est vraie, et "puissance" la probabilité d'avoir raison en décidant que H_1 est vraie.

2 – Dépendance entre deux variables

On considère le cas d'observations de couples (x_i,y_i). On se demande s'il y a un lien entre ces deux variables.

2.1 – Corrélation linéaire

Une solution consiste à supposer que, s'il y a un lien, il est linéaire. Autrement dit, les y_i sont de la forme $y_i = a.x_i + b + \varepsilon_i$. On a alors les deux hypothèses :
- H_0 : a est nul,
- H_1 : a n'est pas nul.

Pour décider, il suffit de calculer le coefficient de corrélation linéaire. Dans le cas de variables indépendantes, ce coefficient est en moyenne nul. Donc, obtenir une grande valeur de ce coefficient est peu probable. On peut donc considérer, si l'on observe une grande valeur de ce coefficient, qu'il y a peu de chances que les variables soient indépendantes.

Tout l'objectif de la théorie des tests est de quantifier le paragraphe précédent.

2.2 – Coefficient de corrélation de Kendall

Maintenant, l'hypothèse H_1 est : les y_i sont de la forme $f(x_i) + \varepsilon_i$, avec f important monotone. Pour tester cette hypothèse, on peut utiliser le tau de Kendall.

Pour chaque couple (i,j) d'observations, on observe si x_i est plus grand que x_j, et on fait de même pour y_i et y_j. Si ces deux couples de valeurs sont dans le même ordre, on les compte pour 1, et pour -1 sinon. On fait ensuite la somme, et on divise par le nombre total de couples : on obtient un indicateur qui vaut entre 1 et -1.

Là aussi, pour des variables indépendantes, il vaut en moyenne 0. On rejettera donc l'indépendance pour de grandes valeurs du tau.

Dans la suite, nous allons vérifier expérimentalement les propriétés de ces tests.

3 – Simulations

Dans la suite, nous allons tirer des échantillons au hasard, de façon à évaluer les fiabilités et puissances des différents tests.

3.1 – Choix d'une famille de fonctions

Nous voulons une famille de fonctions permettant :
- pour certaines valeurs des paramètres, d'obtenir une fonction linéaire,
- pour d'autres valeurs des paramètres, d'obtenir une fonction croissante non-linéaire,
- pour d'autres valeurs des paramètres, d'obtenir une fonction constante.

Nous choisirons les fonctions de la forme : $f(x) = a.x + b.|x|$.

Nous génèrerons 30 couples (x_i,y_i), avec :

- x suivant une loi normale de moyenne 0 et de variance 1,
- y de la forme $f(x)+ \varepsilon$, ε suivant aussi une loi normale de moyenne 0 et de variance 1.

Il faudra pour cela utiliser la fonction *rnorm*. On calculera ensuite le coefficient de corrélation linéaire (fonction *cor*), et on calculera la significativité de cette corrélation (fonction *cor.test*).

3.2 – Vérification expérimentale de la fiabilité du test de corrélation linéaire

On choisit a et b tels que f soit constante, et on observe dans quelle proportion des cas l'hypothèse d'indépendance est rejetée.

Pour répéter cette expérience un grand nombre de fois, on commencera par faire une fonction dont le résultat est la p-value du test. On pourra ensuite répéter cette fonction dans une boucle, et compter combien de fois la p-value est supérieure à 5%.

3.3 – Vérification expérimentale de la puissance des tests : corrélation linéaire et Kendall

On choisit $a=1$, et b tel que f soit linéaire, et on compare les puissances des deux tests présentés. Pour cela, on peut là encore construire une fonction prenant en argument a, b et la méthode de test (pearson ou kendall).

On refait cet essai pour a = 0,1, et b tel que f soit croissante non-linéaire. On constate que le test de Pearson, bien que théoriquement inadapté, reste souvent plus puissant que le test de Kendall.

4 – Mise en œuvre avec R

Le schéma d'une boucle est :

```
for (i in 1 :1000)
{des calculs}
```

Le schéma d'une fonction est :

```
diagnostics <- function(argument)
{#des calculs, utilisant argument
return()}
```

On pourra aussi mettre certaines définitions de fonctions dans un script séparé, appelé par :

```
source("test_fonc.R")
```