Statistiques Appliquées

TD 6
Théorie d'échantillonnage : simulations avec Scilab

Remarque 1 *Dans ce TD on* reprend le code *du 5.2, « un grand nombre d'échantillons », et on l'énrichit afin d'étudier la distribution des statistiques d'un ou deux échantillons. Toutes les questions qui suivent se réfèrent alors au script du 5.2 qu'il faudra modifier astucieusement.*

6.1 Distribution de la moyenne : Z

Selon la théorie d'échantillonnage, la moyenne de l'échantillon, $\bar{X}$, est liée à la v.a. $Z = (\bar{X} - \mu_X)/(\sigma_X/\sqrt{n})$ (normale $N(0,1)$), selon des conditions présentées dans le cours.

6.1.1 Calcul de z et étude qualitative

a. Créer un vecteur vide `z_exp_vec`, de longueur k, pour stocker les valeurs de Z obtenues expérimentalement (par simulation).

b. À l'intérieur de la boucle `for`, calculer z pour chaque échantillon et stocker les résultats dans le vecteur `z_exp_vec` déjà défini.

c. Après la fin des itérations, générer seulement un histogramme de *densités* (option `normalization=%t` : pour chaque classe, afficher la fréquence relative divisée par la largeur de la classe) de `z_exp_vec`.

   Remarque 2 *Afin qu'un histogramme de densités soit une bonne estimation d'une ddp, il faut avoir un grand nombre d'échantillons, p.ex.* $k = 10^3$.

d. Superposer à l'histogramme obtenu la ddp théorique de Z ($N(0,1)$) : créer un vecteur
   `z_th_vec = min(z_exp_vec) : (max(z_exp_vec)-min(z_exp_vec))/500: max(z_exp_vec)`
   et utiliser la fonction `lines` pour tracer $p_Z(z)$ sur la même fenêtre que l'histogramme.

e. Est-ce que la ddp expérimentale de Z (histogramme de `z_exp_vec` avec un grand nombre d'observations) donne une bonne approximation de la ddp théorique de Z ? Quelles classes de l'histograme suivent mieux la ddp théorique ?

6.1.2 Étude quantitative

On veut maintenant *quantifier* les éventuels écarts entre l'histogramme (obtenu de façon expérimentale) et la ddp théorique. L'idée principale est de découper l'histogramme en plusieurs parties (classes) et comparer la fréquence (ou la fréquence relative) de chaque classe à celle prévue par la théorie. Une façon naturelle de faire ce découpage est de choisir $M+1$ classes qui, en théorie, ont les mêmes effectifs, c'est-à-dire des fréquences relatives égales à $1/(M+1)$.

Il faut alors trouver les M valeurs de z qui partagent la ddp théorique en $M+1$ parties, chaque contenant une probabilité égale à $1/(M+1)$, et les utiliser dans le premier argument de la fonction `histplot`, en ajoutant, à gauche et à droite les bornes de la première et de la dernière classe. En théorie, ces limites sont $-\infty$ et $+\infty$ mais, en pratique, il suffit de prendre `min(z_exp_vec)` et `max(z_exp_vec)` pour avoir toutes les valeurs de `z_vec` représentées dans l'histogramme.

a. Créer le vecteur `q_m_th` contenant les M valeurs de z qui définissent les frontières entre les classes. On prendra d'abord $M=1$ ou $M=3$ et on gardera par la suite la valeur $M=99$.

b. Créer le vecteur `bornes=[ min(z_exp_vec), q_m_th, max(z_exp_vec) )` contenant les bornes des $M+1$ classes.

c. Créer l'histogramme de densités de `z_exp_vec`.

d. Calculer par ailleurs la fréquence relative de chaque classe.

e. Les fréquences relatives théoriques sont égales à
`freq_rel_th = 1 / (M+1)`
alors que les fréquences théoriques sont égales à
`freq_th = k / (M+1)`
puisqu'on possède au total k valeurs de `z_exp`.

f. Sur un histogramme de *fréquences*, les classes ainsi définies devraient donner toutes la même fréquence. Comme les classes n'ont pas la même largeur, il faut demander explicitement d'afficher les fréquences avec l'option adéquate, sinon il affiche par défaut les densités (ce qui est le bon choix normalement).
Afficher l'histogramme de fréquences de `z_exp_vec` avec les bornes définies précédemment et ajouter une ligne horizontale au niveau de la fréquence théorique, égale pour toutes les classes.

g. On peut finalement créer le vecteur de l'écart relatif entre les fréquences relatives expérimentales et les fréquences relatives théoriques :
`ecart_rel_freq_rel = ( freq_rel_exp - freq_rel_th ) / freq_rel_th`

h. Afficher les valeurs absolues de l'écart relatif
`plot( abs( ecart_rel_freq_rel ) )`

i. Calculer la norme du vecteur ; elle représente une mesure de la distance entre les fréquences relatives expérimentales et théoriques

le maximum de sa valeur absolue

et la position de ce maximum

j. Quelles classes de l'histogramme présentent les écarts les plus importants entre fréquences relatives expérimentales et théoriques ? Comparer avec la réponse de la question e. de 6.1.1. Expliquer la différence.

Remarque 3 *On verra plus tard, dans le cadre des « tests du χ^2 », une autre façon de traiter ces écarts entre les fréquences expérimentales et théoriques.*

6.1.3 Étude qualitative (deuxième approche) : quantiles vs. quantiles

L'approche suivie dans 6.1.2 n'est pas toujours appliquable. Elle suppose qu'on sait non seulement quelle est la loi théorique utilisée pour la comparaison (ici la loi normale) mais, en plus, quels sont ses paramètres de position (ici $\mu = 0$) et de dispersion (ici $\sigma = 1$). Le plus souvent, on n'a pas d'information sur ce deuxième point et on veut juste savoir si une ddp expérimentale suit une loi, en général (p.ex. la loi normale).

De façon très simple, on veut examiner si une variable aléatoire Y (dont on obtient par mesure un échantillon de k valeurs) est liée à une variable aléatoire X (la loi théorique, avec μ et σ « standard », choisis par nous) par l'intermédiaire d'une relation linéaire $Y = aX + b$. Car si une telle relation existe, alors Y suit *la même loi* que X mais avec une espérance et un écart-type potentiellement différents de ceux choisis pour X (cf. cours, transformations linéaires et v.a. centrée réduite).

Donc le problème d'origine (comparer une ddp expérimentale à une autre théorique) revient à examiner l'existence d'une relation linéaire entre deux variables aléatoires. Or, on sait que si une telle relation existe, alors les valeurs des v.a. la suivent aussi, $y = ax + b$, et, par conséquent, les valeurs « spéciales » $y_\alpha = ax_\alpha + b$, c'est-à-dire leurs *quantiles*, se trouvent sur une ligne.

Il suffit donc, pour chacune des valeurs de Y (on en a obtenu un ensemble de k) de trouver la valeur de α correspondant (c'est-à-dire la proportion des valeurs de Y qui sont inférieures à celle examinée), calculer ensuite x_α (la valeur de X qui laisse la même probabilité α à sa gauche) et tracer y_α en fonction de x_α pour les différentes valeurs de α. Si le resultat est une ligne droite, alors Y suit la même loi que X.[1]

Avec Scilab, tout ça se resume en *quatre opérations* ; si `y` est un vecteur contenant des valeurs de Y,

- `perctl`donne les quantiles expérimentaux de Y.
- La formule de la Gaussienne donne les quantiles théoriques.
- La régression linéaire `reglin` donne la droite de régression entre deux vecteurs (ici les quantiles expérimentaux et théoriques).
- Il suffit ensuite de faire un plot avec ces valeurs.

a. Ajouter les commandes qui affichent un Q-Q plot des quantiles du vecteur `z_exp_vec` par rapport aux quantiles de la distribution théorique.

b. Est-ce que les quantiles expérimentaux et théoriques s'alignent bien sur une droite ?

c. Dans le cas traité ici, on connaît l'espérance et la variance de Z (qui joue le rôle de Y dans la discussion précédante). On peut donc calculer les valeurs de a et b qu'il faudrait obtenir.

d. Est-ce que Z suit bien la ddp théorique ? Quelles quantiles expérimentales de Z correspondent mieux aux quantiles théoriques ? Comparer avec les dernières questions de 6.1.1 et 6.1.2.

Remarque 4 *Pour transformer cette étude qualitative en étude quantitative très puissante, il suffit de calculer le coefficient de corrélation linéaire, ρ, entre les y_α et x_α. Des valeurs de ρ proches de l'unité indiquent qu'une relation linéaire*

[1] En plus, comme $\mu_Y = a\mu_X + b, \sigma_Y = |a|\sigma_X$ et μ_X, σ_X sont connus, on peut utiliser l'intercepte b et la pente a de la ligne pour obtenir une estimation des paramètres μ_Y et σ_Y !

existe effectivement entre les quantiles de Y et de X. Cela concernera la partie regression linéaire du cours.

6.1.4 Étude paramétrique

On veut étudier l'influence de la taille n de l'échantillon sur les écarts entre les fréquences relatives expérimentales et théoriques. Pour cela, on va envelopper la quasi-totalité du code déjà produit dans une boucle qui va balayer plusieurs valeurs du paramètre étudié (taille de l'échantillon).

a. Modifier la définition de la variable `n` : au lieu de lui donner une seule valeur, la définir comme un vecteur, p.ex.
`n =[3,10,50,200]`

b. Initialiser les variables
`norm_ecart_rel_freq_rel`
`max_abs_ecart_rel_freq_rel`
`which_max_abs_ecart_rel_freq_rel`
comme des vecteurs vides (initialisation à 0) de longueur égale à `length(n)`. Ils serviront à stocker les valeurs respectives pour chaque valeur de la taille de l'échantillon.

c. Commencer une boucle `for index.n = 1:length(n)` juste avant la boucle sur `k` ; elle finira après les calculs de la question h. de 6.1.1.

d. À l'intérieur de cette boucle remplacer `n` par `n(index_n)`.

e. De la même façon, les calculs de la question h. de 6.1.1 seront modifiés. P.ex. :
`norm_ecart_rel_freq_rel(index_n) = sqrt( sum( ecart_rel_freq_rel^2 ) )`

f. Après la fin de la boucle sur `index_n`, générer des graphiques montrant l'évolution des variables en fonction des valeurs de `n`.

g. Conclusions, commentaires, etc.