TD 2 : Tests statistiques et corrélation

1 Tests statistiques

Les tests statistiques permettent de confronter une ou plusieurs hypothèses avec des observations. Dans tous les cas, cette démarche peut se résumer ainsi :

1. exposition de la problématique,
2. formalisation des hypothèses,
3. échantillonage, (observation ou mesure)
4. application du test (calcul de la t-statistique),
5. interprétation des résultats et décision.

1.1 Préparation

Exercice 1 : Récupérez l'archive `http://www.labri.fr/~fourer/Ens/1112/stats/data2.tar.gz` et chargez chaque table dans une matrice sous R. Quelle information donne la table de répartition d'une loi de distribution donnée ? Comment la calcule-t-on ?

Exercice 2 : Comment déterminer si une réalisation provient d'un loi de distribution donnée. A quoi correspond le risque (de première espèce) α ? Pourquoi ce paramètre est il essentiel ?

Indication : Pour l'ensemble des tests, on calculera la décision respectivement pour un risque $\alpha = 5\%$ puis $\alpha = 1\%$.

1.2 Test de Shapiro-Wilk

Le test de Shapiro et Wilk permet de vérifier si un échantillon $x \in \mathbb{R}^n$ est distribué selon une loi normale. Pour réaliser le test, on calcule la statistique :

$$W = \frac{\left(\sum_{i=1}^{n} a_i (x_{(n-i+1)} - x_{(i)})\right)^2}{\sum_{i=1}^{n} (x_i - \overline{x})^2}$$

avec $x_{(i)}$ la i-ième plus petite valeur de x et $\overline{x}$ la moyenne empirique de x. Il suffit alors de vérifier que $W > W_{\text{crit}}$ pour valider le test.

Exercice 3 : Implémentez le test de Shapiro et Wilk (vous pourrez utiliser la fonction `sort`). Appliquez le test sur chaque échantillon dont vous aurez préalablement affiché l'histogramme. Conclusions ?

Indication : a_i et W_{crit} sont donnés par les tables contenues dans l'archive.

1.3 Fisher-Snedecor

Le test de Fisher-Snedecor compare la variance entre 2 échantillons x et y de taille respective n_1 et n_2, distribués selon une loi normale.
Pour réaliser ce test, on estime d'abord les variances (sans biais) σ_x^2 et σ_y^2 puis la variable :

$$F = \frac{\max(\sigma_x, \sigma_y)}{\min(\sigma_x, \sigma_y)} \tag{1}$$

On vérifie alors que F suit une loi de Fisher de paramètre $(p-1, q-1)$ pour un risque α donné. Ici p correspond à la taille de l'échantillon ayant la plus grande variance et q la taille de l'échantillon de plus petite variance.

Exercice 4 : Comparez les variances entre les vecteurs sur lesquels le test est applicable (cf. résultat de la question précédente) et déterminez ceux pour lesquels la variance est homogène.

Indication : F a une probabilité $1-\alpha$ de suivre une loi Fisher si $F < F_{\text{crit}}$ avec F_{crit} (donné par la table de répartition) la valeur absolue de la borne maximale de l'intervalle où se trouvent les réalisations d'une loi Fisher de paramètre $(p-1, q-1)$.

1.4 Test de Student

Le test de Student (ou test t) permet de vérifier si l'espérance mathématique d'un échantillon prend une valeur déterminée μ_0 (hypothèse nulle). Pour cela, on estime la moyenne $\hat{\mu}$ et l'écart $\hat{\sigma}$ de l'échantillon. On vérifie alors que t suit une loi de Student de paramètre $n-1$:

$$t = \frac{\hat{\mu} - \mu_0}{\hat{\sigma}\sqrt{n}} \tag{2}$$

Exercice 5 : On veut savoir si le vecteur w a une moyenne homogène à l'un des vecteurs chargés au TD précédent (x, y ou z). Vous devez d'abord vous assurer que le test est applicable, c'est à dire si les échantillons comparés sont indépendants et suivent une loi normale de variance homogène (cf. test de Snedecor).
Calculez un nouveau vecteur d tel que $d_i = u_i - v_i$, où u et v correspondent aux échantillons comparés. Formulez l'hypothèse nulle puis calculez t. Conclusions ?

1.5 Test de Mann-Whitney

Ce test (appelé aussi test U) permet de déterminer si deux échantillons x et y indépendants de taille respective n_1 et n_2 appartiennent à une même population.

On commence par définir le rang de chaque élément $k \in x \cup y$. On note $r(k)$ le rang de la valeur k. Pour chaque échantillon la statistique on calcule :

$$U_x = n_1 n_2 + \frac{n_1(n_1+1)}{2} - \sum_{k \in x} r(k)$$
$$U_y = n_1 n_2 + \frac{n_2(n_2+1)}{2} - \sum_{k \in y} r(k)$$

- Si $n_1 \leq 20$ ou $n_2 \leq 20$, on vérifie que $U > U_{\text{crit}}$ pour α donné.
- Sinon, On vérifie alors que Z suit une loi normale de paramètre $(0, 1)$:

$$Z = \frac{\min(U_x, U_y) - \hat{\mu}}{\hat{\sigma}} \tag{3}$$

avec $\hat{\mu} = \frac{n_1 \cdot n_2}{2}$, $\hat{\sigma} = \sqrt{\frac{(n_1 \cdot n_2)(n_1+n_2+1)}{12}}$.

Exercice 6 : Implémentez le test puis déterminez les vecteurs qui suivent la même distribution.

1.6 Test de Wilcoxon

Il s'agit d'un test similaire à Mann-Whitney qui permet de vérifier si 2 échantillons x et y de taille N appartiennent à une même population. On définit le vecteur d tel que $d_i = x_i - y_i$ et le rang pour chacune des valeurs $|d_i|$ en ignorant les valeurs nulles.

On calcule alors les statistiques $T^+ = \sum_{d_i>0} r(d_i)$ et $T^- = \sum_{d_i<0} r(d_i)$ puis :

$$Z = \frac{T - \hat{\mu}}{\hat{\sigma}} \tag{4}$$

avec $T = \min(T^+, T^-)$. On vérifie alors que Z suit une loi normale de paramètre $(0, 1)$.

Exercice 7 : Implémentez le test et vérifiez la cohérence des résultats avec la question précédente. Conclusions ?

1.7 Test ANOVA

Le test ANOVA[1] permet de déterminer simultanément si p échantillons x_k de taille respective n_i suivent une même loi de distribution. On considère le modèle suivant :

$$x_{k,i} = \mu + a_k + \epsilon_{k,i} \tag{5}$$

où μ est la moyenne totale, a_k la variation liée à x_k et $\epsilon_{k,i}$ une erreur normalement distribuée. Ainsi la variation totale peut s'écrire :

$$\mathrm{SCE_{tot}} = \mathrm{SCE_{inter}} + \mathrm{SCE_{intra}}$$

$$\sum k = 1^p \sum_{i=1}^{n_k} (x_{k,i} - \bar{x})^2 = \sum_{k=1}^{p} n_k (\bar{x}_k - \bar{x})^2 + \sum k = 1^p \sum_{i=1}^{n_k} n_k (x_{k,i} - \bar{x}_k)^2$$

On calcule la statistique :

$$F = \frac{\mathrm{SCE_{inter}}/(p-1)}{\mathrm{SCE_{intra}}/(N-p)} \tag{6}$$

Puis on vérifie que F suit une loi Fisher de paramètre $(p-1, N-p)$.

Exercice 8 : Comparez simultanément tous les vecteurs en appliquant le test ANOVA.

1.8 Test de Kruskal-Wallis

Ce test compare la moyenne de plusieurs échantillons x_k, pour $k \in [1, p]$. On commence par calculer le rang $r(x_{k,i})$ de chaque valeur puis la t-statistique :

$$H = \left(\frac{12}{N(N+1)} \sum_{k} = 1^p R_k^2 n_k \right) - 3(N+1) \tag{7}$$

avec $R_k = \sum_{i=1}^{n_k} r(x_{k,i})$. On vérifie alors que H suit une loi du χ^2 de paramètre $p-1$.

Exercice 9 : Comparez simultanément tous les vecteurs en appliquant le test Kruskal-Wallis.

[1] ANOVA : ANalysis Of VAriance

1.9 Bonus : test d'adéquation selon une loi quelconque

Le test d'adéquation du χ^2 permet de vérifier si un échantillon suit une loi de distribution donnée. On devra alors au préalable effectuer une modélisation et choisir une loi et définir ses paramètres.

On commence par fixer un nombre de classe d et calculer l'histogramme de l'échantillon étudié.
La t-statistique est donnée par :

$$\chi^2 = \sum_{i=1}^{d} \frac{(h(x_i) - P(x_i))^2}{P(x_i)} \tag{8}$$

avec $h(x_i)$ la fréquence observée de la valeur x_i et $P(x_i)$ la probabilité théorique d'obtenir x_i d'après le modèle. Le test est vérifié si χ^2 suit une loi χ^2 de paramètre $d-1$ (degré de liberté).

Exercice 10 : Implémentez le test du χ^2 puis validez vos hypothèses de modèle concernant les échantillons précédents.

2 Corrélation

On s'interesse ici à l'analyse d'un réseau mitochondrical. Ainsi, le métabolisme de la mitochondrie a été représenté par 44 réactions (dont 16 irréversibles).
Au cours d'une expérience, on a mesuré le nombre de nombre de réactions réalisées en fonction du temps. Ainsi, le fichier `mitochondrie.csv` représente un tableau où chaque colonne correspond à une réaction. On espère mettre en évidence les relations qui existent entre les réactions par une mesure de corrélation.

2.1 Corrélation de Pearson

On définit la mesure de corrélation de Pearson entre 2 échantillons x et y de taille N par :

$$C_{xy} = \frac{\sigma_{xy}}{\sigma_x \cdot \sigma_y} = \frac{\sum_{i=1}^{N}(x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N}(x_i - \bar{x})^2} \cdot \sqrt{\sum_{i=1}^{N}(y_i - \bar{y})^2}} \tag{9}$$

Exercice 11 : Comparez deux à deux les colonnes de la matrice et retrouvez celles qui sont les plus fortement corrélées. On considère qu'une corrélation est importante lorsque $|C_xy|$ est proche de 1. On fixera d'abord un seuil minimal à 0.5.

2.2 Corrélation de Spearman

La corrélation de Spearman s'obtient en calculant la corrélation de Pearson en considérant les nouveaux échantillons x' et y' ou x'_i correspond au rang de la valeur x_i. Le rang étant la position de la valeur x_i lorsque l'on trie x par ordre décroissant. Lorsque la valeur x_i apparait plusieurs fois, alors chaque x'_i est identique et correspond à la moyenne des rangs distincts correspondants. (Par exemple, si on considère $x = (5, 21, 1, 0.6)$, alors $x' = 1, 2, 3.5, 3.5, 5$ car $\frac{3+4}{2} = 3.5$).

Exercice 12 : Calculez la corrélation de Spearman sur léchantillon précédent. Obtenez vous les même conclusions que pour la question précédente pour un seuil identique ?

Exercice 13 : Choisissez 2 réactions fortements corrélées les points correspondants à deux échantillons que vous comparez en plaçant respectivement chacun d'eux sur un axe (au choix) des abscisses et des ordonnées. Que remarque t'on pour les échantillons fortement corrélés. Plus faiblement corrélés.