LES METHODES D'ECHANTILLONNAGE

Veillez à avoir lu les sections relatives à cette séance dans le document disponible sur le site « Introduction aux séances de TP de biostatistique » avant de vous présenter au TP.

Veillez aussi à maitriser les notions apprises au(x) TP(s) précédent(s).

Exercice 1

Dans R, la fonction *runif(n,min=a,max=b)* génère n nombres aléatoires compris entre a et b (par défaut, a et b valent 0 et 1, respectivement) depuis une distribution de probabilité uniforme. On va l'utiliser pour générer un vecteur de nombres aléatoires. Ainsi, tapez :

```
x<-runif(100)
```

La variable x représente donc un vecteur de 100 composantes dont chaque valeur a été tirée au hasard entre 0 et 1.

- Combien vaut la somme des composantes du vecteur x ?
- Quelle(s) commande(s) devez-vous taper pour qu'un vecteur dénommé xsum contienne 10 composantes, la première étant la somme des composantes 1 et 91 de x, la deuxième étant la somme des composantes 2 et 92 de x, la troisième étant la somme des composantes 3 et 93 de x, ..., la dixième étant la somme des composantes 10 et 100 de x ?
- Quelle(s) commande(s) pouvez-vous taper pour calculer la somme des éléments d'indice pair de x ?
- Utilisez la fonction *summary()* pour avoir un résumé descriptif des données du vecteur x.
- Créez un vecteur nommé « param » contenant 7 paramètres de position et de dispersion de l'échantillon x (moyenne, variance, écart-type, médiane, étendue, 3^{e} décile et 35^{e} percentile) ainsi que leurs étiquettes respectives. Arrondissez ces paramètres à deux décimales (fonction *round()*). *Remarque* : La fonction *quantile()* permet de trouver n'importe quel percentile, décile ou quartile d'un échantillon.

Exercice 2

Voici une liste de 20 espèces animales :

Chat, Chien, Lapin, Cheval, Vache, Furet, Ara, Canari, Lièvre, Cobaye, Python, Mouton, Bouc, Porc, Tortue, Requin, Ours, Panda, Lama & Chameau.

Recopiez-la dans Excel (colonne B – cellule B1 = « espece ») et triez-la par ordre alphabétique. Ensuite attribuez un numéro d'ordre croissant à chacune des espèces (colonne A – cellule A1 = « numero ») et enregistrez le fichier au format CSV (Enregistrez sous…-> Autres formats -> Type de fichier : choisir CSV (séparateur : point-virgule)) dans le répertoire de travail de R. Pour connaître le répertoire de travail, tapez *getwd()* dans la console R.

Dans R, importez le fichier grâce à l'instruction *read.csv2()* ou *read.csv()* ou *read.table*() et enregistrez-le dans une variable (par ex. « esp »). N'hésitez pas à utiliser l'aide en ligne de R pour déterminer les arguments à utiliser pour ces fonctions. Dans un script, écrivez les lignes de commande qui permettront de tirer au hasard un échantillon de 6 espèces sachant qu'on considère qu'il y a remise (et donc qu'une même espèce peut être présente plusieurs fois dans l'échantillon).

Vous devrez utiliser la fonction *sample()* afin de tirer un échantillon au hasard dans une population, avec ou sans remise.

Exercice 3

Un responsable de la chaîne alimentaire consulte les statistiques de découverte de carcasses positives pour la présence de résidus (dioxine). Il constate qu'on découvre en moyenne 3 carcasses par mois dans chaque province. Quelle est la distribution du nombre de carcasses contaminées en Belgique (qui compte 10 provinces) par an ?

Echantillonnez 500 valeurs dans cette distribution et sauvegardez l'échantillon dans un vecteur nommé r. Comptez le nombre de fois où chaque valeur k de la variable aléatoire (« nombre de carcasses positives en Belgique ») est observée dans la distribution empirique obtenue. Trois possibilités :

- Soit utiliser une boucle. Pour cela, créez au préalable un vecteur vide avec la fonction *vector()*, de longueur supérieure ou égale à la valeur maximale de k (dans le vecteur r). Ensuite, utilisez *for* pour boucler une instruction un nombre de fois identique à la valeur maximale de k, ce qui permettra de calculer la longueur d'un sous-vecteur pour chacune des valeurs de k. *Exemple* : *length(r[r==i])* où r est le vecteur échantillon généré et i la variable de la boucle *for. Pour comprendre cette instruction, faire des essais du type length(r[r==30]) et essayer de comprendre ce que R retourne*.
- Soit avoir recours à la fonction *hist()* en précisant la taille de chaque classe à l'aide de l'argument breaks=seq(min-0.5,max+0.5,1) où min et max sont respectivement les valeurs minimum et maximum de vos données). Vous préciserez ne pas vouloir afficher le graphique avec l'argument plot=F (F étant l'abréviation de FALSE) et vous mémoriserez les résultats de la fonction *hist()* dans une variable nommée h). Utilisez la commande *names(h)* pour obtenir les différents champs contenus dans h, dont le champ *count*, qui fournit les comptages dans chaque classe. Sélectionnez uniquement ces fréquences et présentez vos données sous la forme d'un tableau avec une colonne pour les classes et une colonne pour les fréquences correspondantes.
- Soit employer la fonction *table()*, qui génère une table directement représentable graphiquement

La fonction graphique *plot()* permettra de générer graphiquement la distribution empirique obtenue.

Exercice 4

Si le poids des bovins adultes en Belgique suit une distribution normale (Laplace-Gauss), nous pouvons générer un échantillon de 1000 poids au hasard en utilisant, dans Excel, la fonction *LOI.NORMALE.INVERSE.N(proba;moyenne;deviation std)*.

- Pour avoir un animal choisi au hasard dans la distribution, il suffit de placer la fonction *ALEA()*, qui génère un nombre pris au hasard entre 0 et 1, qui sera utilisé comme une probabilité (argument « proba »).
- Toute distribution normale est caractérisée par deux paramètres : la moyenne (nous utiliserons un poids de 750 kg pour nos poids de bovins) et la variance (égale à 2500 kg² pour notre exemple).

La fonction *LOI.NORMALE.INVERSE.N()* utilisée avec ces paramètres retourne alors une valeur de la variable normale, interprétée comme le poids d'un bovin tiré au hasard dans la population (valeur d'abscisse) : si Q est la probabilité tirée au hasard entre 0 et 1, et Y le poids fourni par la fonction, on peut en déduire que P(poids < Y) = Q

Générez un échantillon de 1000 poids. Ensuite créez un tableau à 6 colonnes et 11 lignes. La première ligne comprendra les en-têtes des différentes colonnes, à savoir : limites inférieures et supérieures de classes, effectifs observés, probabilités théoriques cumulées, probabilités théoriques de classes et effectifs théoriques.

Les classes auront une largeur de 50kg, sauf la première classe]-∞ ;600] et la dernière]900 ;+ ∞[. La deuxième classe est donc]600 ;650], etc.

La troisième colonne regroupera les effectifs associés aux 1000 poids dans les classes et il faudra donc en calculer les fréquences. La fonction *FREQUENCE(plage données ;plage lim classes)* utilisée en format matriciel (càd. en sélectionnant la zone entière devant accueillir les fréquences calculées, en tapant la commande, puis en validant par les touches CTRL+MAJ+ENTER simultanément) vous fournit directement les fréquences non cumulées (contrairement à la fonction *NB.SI()* ou à la fonction FREQUENCE en format scalaire (ce qu'on obtient en validant par ENTER seul) qui calculent les effectifs cumulés). *FREQUENCE()* utilise les limites supérieures de classe pour calculer chaque intervalle. Pour la dernière classe (dont la limite est + ∞), employez une valeur supérieure au maximum comme limite supérieure). La distribution obtenue est une distribution empirique.

Pour générer les effectifs théoriques, utilisez la fonction *LOI.NORMALE.N()*, déjà vue, dans la colonne "proba théoriques cumulées". On peut alors facilement obtenir les probabilités par classe par différences entre probabilités cumulées. La somme de cette colonne "proba théorique de classes" devrait valoir alors exactement 100%. Les effectifs attendus sont obtenus en multipliant les probabilités ainsi obtenues par l'effectif total (soit, 1000 individus). Cette distribution est la distribution théorique de moyenne 750kg et de déviation standard 50kg, mise sous formes de classes.

Représentez les deux distributions de manière superposée sur le même graphique et comparez-les.

Exercices facultatifs pour le brain storming à domicile

1. Dans R, trouvez une autre méthode pour tirer au hasard un échantillon de 6 espèces dans l'exercice 2. *Indice* : vous devez travailler sur le positionnement des données au sein du tableau et utiliser la commande *runif(), associée à floor() ou round().* La fonction *floor()* permet de garder uniquement la partie entière d'un nombre avec décimales tandis que *round()* permet de réaliser une arrondi de la valeur générée (en précisant le nombre de décimales désiré).

2. **Exercice vivement conseillé en entrainement** : Dans R, résolvez l'exercice 4 en générant un vecteur de 1000 poids distribués normalement et en réalisant un tableau avec les vecteurs des effectifs observés et théoriques ainsi que ceux des probabilités théoriques sous forme d'un data.frame (= tableau de données où chaque colonne est un vecteur de valeur. Chaque vecteur-colonne doit donc avoir le même nombre d'éléments. Chaque ligne représente un individu, mais dans notre cas, ce sera une classe d'individus).