TP4 Statistiques (MAP201)
Intervalle de confiance, test d'hypothèse, loi de χ^2

Objectifs. L'objectif de ce TP est de vous faire manipuler certaines méthodes statistiques pour l'aide à la décision. Ce sujet ne constitue que la base de ce qui vous est demandé : soyez critique par rapport à vos résultats, proposez d'autres idées, solutions ou tests. La mise en oeuvre de techniques de visualisation est fortement encouragée. La lisibilité du code et la pertinence des commentaires seront pris en compte dans la note du TP. La qualité de la rédaction, de la synthèse, de l'analyse des résultats obtenus sont des critères importants pour la note.

Le TP est à rendre **au bout d'une semaine**, sous forme imprimée, contenant :
- le compte rendu répondant aux questions
- les codes R correspondant
- les images qui illustrent les réponses

Tout retard ou non respect des consignes sera sanctionné.

1 Intervalles de confiance

On rappelle qu'un *intervalle de confiance* pour un paramètre θ est un intervalle **aléatoire** $[A_n, B_n]$ (A_n et B_n étant des v.a de réalisation a_n et b_n) qui contient le paramètre θ avec une probabilité $1-\alpha$. En particulier, pour des individus possédant un caractère A avec une probabilité p, le nombre d'individus x présentant le caractère A au sein d'un échantillon de taille n suit une loi binomiale $\mathcal{B}(n,p)$. Un estimateur du paramètre p est la proportion $f_n = \frac{x}{n}$ et un intervalle de confiance au risque α est donné par :

$$[a_n, b_n] = \left[f_n - u_\alpha \sqrt{\frac{f_n(1-f_n)}{n}}, f_n + u_\alpha \sqrt{\frac{f_n(1-f_n)}{n}} \right] , \tag{1}$$

avec u_α tel que $\mathbb{P}(|U| \leq u_\alpha) = 1-\alpha$ où $U \sim \mathcal{N}(0,1)$. Sous R : $u_\alpha = \text{qnorm}(1-\alpha/2)$.

On se propose de montrer par simulation numérique qu'un intervalle de confiance ne recouvre pas toujours la vraie valeur du paramètre.

Exercice 1 (Intervalles de confiance) Soit un échantillon de $n = 20$ individus issus d'une population possédant le caractère A avec une probabilité $p = 0.5$. On est ainsi capable d'établir un intervalle de confiance pour p grâce à (1). On suppose que l'on peut renouveller m fois l'expérience. On dispose maintenant de m échantillons de taille n issus de cette population.
- Complétez le code suivant afin de rendre compte de la proportion d'intervalles de confiance contenant le vraie valeur du paramètre p. Afficher le résultat de `matplot`.
- À quelle valeur théorique doit-on s'attendre ?

IntervalleBinom.R

```
m=50; n=20; p=.5;                # lancer de 20 pièces équilibrées, 50 fois
fn=rbinom(m,n,p)/n               # m échantillons binomiaux de taille n (en proportions)

# A COMPLETER :
```

```r
# construire les vecteurs an et bn contenant les m réalisations des bornes d'intervalle

matplot(rbind(an,bn),          # on affiche tous les intervalles sous forme de lignes
        rbind(1:m,1:m),        # les uns en dessous des autres en couleurs
        type="l",lty=1)
abline(v=p)                    # ligne pour p=0.5

# A COMPLETER :
# calculer le pourcentage d'intervalles contenant le vrai paramètre p
```

2 Tests d'hypothèse : test de Zener appliqué à la télépathie

Les cartes de Zener sont un jeu de 25 cartes comportant 5×5 symboles représentés Figure 1. Inventées en 1920, elles ont été utilisées lors des premières recherches quantitatives effectuées dans le domaine de la parapsychologie.

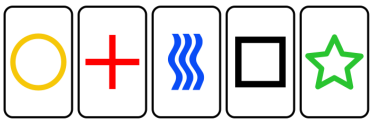

FIGURE 1 – Symboles figurant sur les cartes de Zener

L'expérience réalisée avec les cartes de Zener vise généralement à **tester le taux de clairvoyance** d'un sujet : un expérimentateur tire les 25 cartes l'une après l'autre, sans les montrer au sujet de l'expérience, qui doit deviner le symbole inscrit sur chacune d'elles. Un taux normal de réussite (provoqué uniquement par des réponses données au hasard) serait de 20%. Vous pouvez tester vos capacités extralucides ici :

`http://www.charlatans.info/test-cartes-zener.php`

La v.a X comptant le nombre de cartes devinées suit une loi binomiale. Le test d'hypothèse est

$$(\mathcal{H}_0) : X \sim \mathcal{B}(25, p),\ p = \frac{1}{5} \quad \text{contre} \quad (\mathcal{H}_1) : X \sim \mathcal{B}(25, p),\ p > \frac{1}{5}$$

Exercice 2 (Mise en jambe avec la loi binomiale) On se place sous l'hypothèse $\mathcal{H}_0$.

- Simulez avec `rbinom` 1000 parties.
- Affichez l'histogramme et vérifier que le gain moyen est bien de 20%.
- Affichez la distribution théorique des probabilités avec `dbinom`.
- Coloriez en rouge les lignes $\mathbb{P}(X = k)$ pour $10 \leq k \leq 25$.

Exercice 3 (Test de Zener) Une personne qui passe le test obtient 10, soit 40% de réussite.

- Suivez l'exemple de test d'hypothèse vu en cours pour déterminer les régions de rejet pour le risque $\alpha = 5\%$, puis $\alpha = 1\%$. Coloriez-les de même en rouge. Rejetez-vous l'hypothèse nulle pour ces risques ?

- Déterminer la p-valeur, puis utiliser `binom.binom` pour vérifier votre résultat.
- On fait passer ce test à l'ensemble des étudiants du DLST. Un étudiant obtient le score de 12, frôlant les 50%. Qu'en concluez-vous ? [1]
- Quelle stratégie adopteriez-vous pour obtenir **à coup sûr** 20% de réussite (ou plus) ?

Exercice 4 (Stratégie optimale) Le code `CarteZener.R` simule plusieurs parties de cartes « **faces visibles** » avec des stratégies (de mémorisation) différentes. Il vous est demandé de :

- Commentez **chacune** des lignes du programme principal fourni dans `CarteZener.R`.
- Complétez la fonction `Devine` de sorte à ce qu'elle opère la stratégie voulue.
- Reportez pour chacune des stratégies l'espérance obtenue. Pour laquelle opteriez-vous afin d'optimiser vos chances de réussite ? Vous pourrez maintenant épater vos amis ;-)
- En effet, montrez *via* `test.binom` qu'avec un tel taux de réussite sur 10 parties, la p-valeur est inférieure à 0,001%.

CarteZener.R

```
Devine <- function(paquet,strategie){
  # A COMPLETER : fonction qui suivant la stratégie adoptée et l'état du paquet
  # choisi "aléatoirement" un symbole afin de deviner la carte courante
}

strategie <- 0     # 0 = choix aléatoire d'un symbole (entre 1 et 5),
                   # 1 = choix aléatoire parmi les symboles restant dans le paquet
                   # 2 = choix aléatoire parmi les symboles les moins sortis
gains <- vector()

for (l in 1:10000) {
  paquet <- c(5,5,5,5,5)
  tirage <- rep(1:5,each=5)
  melange <- sample(1:25)
  tirage <- tirage[melange]
  gain <- 0

  for (k in 1:25) {
    carteCour <- tirage[k]
    carteDev <- Devine(paquet,strategie)
    if (carteCour == carteDev) {
      gain <- gain+1
    }
    paquet[carteCour] <- paquet[carteCour]-1
  }
  gains <- c(gains,gain)
}

mean(gains)
```

1. *Le Saviez-Vous ?* Le prix James Randi récompense de 1 million d'euros quiconque pourra démontrer la réalité d'évènements paranormaux ou de pouvoirs surnaturels.

3 Test du χ^2

Exemple de motivation. En cryptographie, le chiffrement par décalage, aussi connu sous le nom de « *chiffrage de César* », est une méthode de chiffrement très simple utilisée par Jules César dans ses correspondances secrètes. Le texte chiffre s'obtient en remplaçant chaque lettre du texte clair original par une lettre à distance fixe. Par exemple, avec un décalage de 3 vers la droite, la lettre A est remplacée par la lettre D, B devient E, et ainsi jusqu'à W qui devient Z, puis X devient A, etc. Cela revient à effectuer une permutation circulaire du dictionnaire $\mathcal{D}$. Le décalage est la clé de chiffrement. Si le destinataire connait la clé, il peut facilement déchiffrer le message reçu en effectuant le décalage inverse. Si l'on sait que l'envoyeur utilise ce type de chiffrement on pourrait aisément tester toutes les clés possibles (25 possibilités). On propose ici de s'appuyer sur une détection automatique de la clé de déchiffrement la plus probable.

Les fréquences théoriques d'apparition $f_t(\ell)$ d'une lettre ℓ dans la langue française est renseignée dans le dictionnaire des fréquences `dico_fq` (par ordre alphabétique). Si $f(\ell)$ désigne sa fréquence d'apparition dans un texte donné, l'écart des carrés (appelée **erreur quadratique**) mesure la « distance » de ces fréquences par rapport aux fréquences théoriques :

$$\mathrm{EQ} = \sum_{\ell \in \mathcal{D}} (f_t(\ell) - f(\ell))^2 \, .$$

Si on note cette fois $f_d(\ell)$ la fréquence d'apparition de ℓ dans le texte décalé de d caractères dans le déchiffrement de César, nous allons rechercher $d_{\min}$ qui minimise la fonction :

$$\mathrm{EQ}(d) = \sum_{\ell \in \mathcal{D}} (f_t(\ell) - f_d(\ell))^2 \, . \tag{2}$$

Exercice 5 (Cryptographie) Complétez le code `FreqCrypto.R` fourni afin de déchiffrer :

« GLIVWIXYHMERXWQIWJIPMGMXEXMSRWPIGSHIIWXQEMRXIRERXHIGLMJJVI »

On pourra utiliser les fonctions fournies `ceasar` et `analyseFreq` qui effectuent respectivement le chiffrement de César et l'analyse fréquentielle d'un message.

FreqCrypto.R

```
mot_crypt <- "GLIVWIXYHMERXWQIWJIPMGMXEXMSRWPIGSHIIWXQEMRXIRERXHIGLMJJVI"
print(mot_crypt)

# Méthode de décryptage :
# on compare les fréquences des lettres avec celle du dictionnaire français
dico <- strsplit("ABCDEFGHIJKLMNOPQRSTUVWXYZ", NULL)[[1]]
dico_fq <- c(7.68,0.8,3.32,3.6,17.76,1.06,1.1,0.64,7.23,0.19,0,5.89,2.72,7.6
1,5.34,3.24,1.34,6.81,8.23,7.3,6.05,1.27,0,0.54,0.21,0.07)/100

EQ <- rep(0,26)                                  # vecteur des erreurs quadratiques
for (d in 0:25) {
  # A COMPLETER : pour chaque décalage du mot crypté :
  # - calculer le vecteur des fréquences des caractère dans le mot décalé
```

```
  # - calculer l'erreur quadratique entre ce vecteur et le dictionnaire des fréquences
  # - stocker cette erreur dans le tableau EQ
}
# A COMPLETER : déterminer dans EQ le décalage dmin qui minimise l'erreur quadratique
mot_decrypt <- ceasar(mot_crypt,dmin)      # qui fait apparaître le message en clair
```

Test d'ajustement du χ^2. Le but de ce test est de comparer une distribution théorique d'un caractère à une distribution observée. Pour cela, le caractère doit prendre un nombre fini de valeurs, ou bien ces valeurs doivent être rangées en un nombre fini de classes.

- **Données** : un caractère A dont les valeurs possibles sont réparties en k classes $A_1, \dots, A_k$. La probabilité théorique des variables aléatoires (**indépendantes**) dans chacune des classes est notée $p_1, \dots, p_k$. On dispose de n observations, qui donnent un effectif n_1 pour la classe A_1, ... n_k pour la classe. Bien sûr, on doit avoir $n_1 + \cdots + n_k = n$.
  On note $\hat{p}_j = n_j/n$ les fréquences empiriques.
- **Hypothèse testée $\mathcal{H}_0$** : « La distribution observée est conforme à la distribution théorique » avec un risque d'erreur α.
- **Déroulement du test** :
  1. On calcule les effectifs théoriques np_j
  2. On calcule la valeur observée de la variable de test :
  $$\chi^2 = \sum_{j=1}^{k} \frac{(n_j - np_j)^2}{np_j} = n \sum_{j=1}^{k} \frac{(\hat{p}_j - p_j)^2}{p_j} \tag{3}$$
  3. On cherche la valeur critique χ^2_α dans la table de la loi du χ^2 à $k-1$ degrés de liberté.
  4. Si $\chi^2 < \chi^2_\alpha$, on accepte l'hypothèse, sinon on la rejette.
  5. Vérification *a posteriori* des conditions d'application : il faut $np_j \geq 5$ pour tout j.

On voit que la variable aléatoire correspondant à (3) mesure bien, comme dans la partie précédente, la « distance » ou l'« écart » entre les fréquences théoriques p_j et les fréquences observées $\hat{p}_j$. Elle permet donc de modéliser l'hypothèse à tester.

Exercice 6 (Linguistique) Deux extraits de romain sont fournis dans l'archive : l'avant-propos de *La Comédie humaine* de Balzac, et un passage de *La Disparition* de Perec [2]. L'objectif de cet exercice est de tester par un test du χ^2 si l'un de ces textes fait partie de l'oeuvre de Balzac. Pour cela on fournit dans le code `FreqLettres.R` les fréquences théoriques d'apparition des lettres dans l'ensemble du corpus balzacien. Compléter le code de façon à :

- Compter le nombre d'occurences des lettres dans ces textes. Vous pourrez au choix lire ces textes dans `R` avec `readLines` et adapter la fonction `analyseFreq` de l'exercice précédent, ou plus simplement avec l'aide du site suivant (cochez « Compter les apparitions ») :
  `https://www.dcode.fr/analyse-frequences`
- Afficher les diagrammes en bâtons des fréquences d'apparition des lettres pour ces deux textes. Commentez. [3]
- Effectuer un test du χ^2 entre le vecteur des effectifs de *La Comédie humaine* et le vecteur des fréquences théoriques, puis de même avec *La Disparition*. On pourra utiliser `help(chisq.test)` pour en comprendre le fonctionnement. Qu'en concluez-vous ?
- À votre avis, le texte de *La Disparition* s'il était chiffré avec une clé de César serait-il

déchiffrable par la méthode de l'exercice précédent ?
- À votre avis, pourquoi un `Warning` s'affiche ? Par quel entier doit-on multiplier le vecteur des effectifs pour être dans les conditions d'application du test de χ^2 ?
- L'hypothèse d'indépendance requise dans le test du χ^2 vous semble-t-elle vérifiée ici ?

FreqLettres.R

```
# Fréquences d'apparition des lettres dans le corpus de Balzac (ordre décroissant)
probs <- c(17.8,8.7,8,7.5,7.5,6.4,6.1,6,5.8,5.2,       # E S A I N T R L U O
           + 3.9,3.3,3.2,2.6,1.4,1.3,1.2,1.1,1,0.7,   # D M C P V Q B F H G
           + 0.5,0.3,0.3,0.1,0.07,0.03)/100           # X Y Z J W K

# A COMPLETER :
# - Construire les vecteurs "balzac" et "perec" contenant les effectifs d'apparition
# des lettres prises dans le même ordre que celui du vecteur "probs"
# - Utiliser la fonction chisq.test pour tester sur ces vecteur l'hypothèse :
# "Ce texte appartient au corpus de Balzac"
```

Exercice 7 (Hasard humain) Une expérience menée en amphithéâtre a mis en évidence que l'être humain est généralement un piètre générateur de hasard. Celle-ci consistait à choisir « au hasard » un chiffre entre 1 et 9. Les effectifs obtenus pour chacun des amphis sont donnés dans le code `ChoixHasard.R`. Puis un test d'homogénéité est effectué pour tester si les effectifs obtenus proviennent bien de la même population, autrement dit que le comportement des étudiants n'est pas significativement différent d'un amphi à l'autre. Il vous est demandé de compléter le code de façon à :

- Afficher les fréquences obtenues par les deux amphis sur un diagramme côte-à-côte.
- Effectuer un test du χ^2 sur les résultats obtenus pour la réunion des amphis afin de tester l'hypothèse d'un hasard uniforme. Qu'en concluez-vous ?

ChoixHasard.R

```
amphi1 <- c(0,7,4,7,16,7,9,4,4)
amphi2 <- c(1,0,2,3,6,6,6,3,1)
amphicomplet <- amphi1+amphi2

# test d'homogénéité
amphis <- rbind(amphi1,amphi2)
chisq.test(amphis)

# A COMPLETER : afficher côte-à-côte avec rbind et barplot les diagrammes en bâtons
# correspondant aux fréquences de choix des chiffres de 1 à 9 des amphis 1 et 2

# A COMPLETER : effectuer un test du khi deux entre le vecteur des effectifs de
# l'amphi complet et la loi uniforme
```

Ainsi se termine le cours de statistiques sur le premier exemple traité en amphi. La boucle est bouclée. Bonnes révisions pour les partiels !

3. *Le Saviez-Vous ?* Le roman de 300 pages *La Disparition* a été écrit sans utiliser une seule fois la voyelle *e*.
3. *Le Saviez-Vous ?* La fréquence d'occurence des mots dans un texte suit la loi de Zipf.