TP : Théorie des sondages

Exercice 1. Reproduire et décrire brièvement l'enjeu des commandes R suivantes :

```
colors = c("red", "blue", "yellow", "green", "grey", "black", "purple")
par(mfrow = c(2, 2))
for (i in 1:4) {
s = sample(colors, size = 4, replace = F)
barplot(rep(1, 4), col = s)
}
for (i in 1:4) {
s = sample(colors, size = 9, replace = T)
barplot(rep(1, 9), col = s)
}
```

Exercice 2. Reproduire et décrire brièvement l'enjeu des commandes R suivantes :

```
U = c("Bob", "Nico", "Ali", "Fabien", "Malik", "John", "Jean", "Chris",
"Karl")
y = c(72, 89, 68, 74, 81, 87, 76, 61, 84)
library(sampling)
t = srswor(3, 9)
U[t != 0]
bar_y_w = (1 / 3) * sum(y * t)
bar_y_w
```

Exercice 3. Reproduire et décrire brièvement l'enjeu des commandes R suivantes :

```
icPESR = function(y, N, niveau) {
n = length(y)
bar_y_w = mean(y)
z = qnorm(1 - (1 - niveau) / 2)
s2_w = sd(y)^2
var_bar_y_w = (1 - n / N) * (s2_w / n)
a = bar_y_w - z * sqrt(var_bar_y_w)
b = bar_y_w + z * sqrt(var_bar_y_w)
print(c(a, b)) }
```

Que renvoie les commandes R suivantes ?

```
icPESR(y = c(2.2, 2.1, 3.1, 2.8, 5.1, 4.6), N = 100, niveau = 0.90)
```

Exercice 4. On considère une population de 2000 individus et on s'intéresse au caractère $Y =$ "taille" en centimètres.

Les données sont disponibles ici :

```
r = read.table("https://chesneau.users.lmno.cnrs.fr/tailles-pop.txt",
header = T)
attach(r)
head(r)
```

1. Comme on dispose de la population entière, on peut d'ores et déjà calculer les valeurs exactes des paramètres-population de Y.
   (a) Calculer la moyenne-population.
   (b) Calculer l'écart-type corrigé-population.
2. On s'intéresse maintenant à quelques éléments annexes.
   (a) Tracer l'histogramme des données.
   (b) Est-ce que l'on peut affirmer, au risque 5%, que Y suit une loi normale ?
3. On considère maintenant 200 échantillons de 100 individus prélevés dans la population suivant un plan de sondage aléatoire de type PESR. On note $\overline{y}_W$ l'estimateur de la moyenne-population dans ce contexte.
   (a) Reproduire et décrire brièvement l'enjeu des commandes R suivantes :

```
library(sampling)
mat_ech = function (x, n, M){
N = length(x)
A = matrix(0, M, n)
for(i in 1:M){
t = srswor(n, N)
A[i, ] = x[t != 0]
}
A
}
```

   (b) Utiliser les commandes R précédentes pour construire une matrice de 200 lignes et 100 colonnes avec, pour chacune des lignes, les 100 valeurs de Y pour 100 individus d'un échantillon prélevé suivant un plan de sondage aléatoire de type PESR.
   (c) Calculer les 200 moyennes-échantillon. Est-ce que l'on peut affirmer, au risque 5%, que $\overline{y}_W$ suit une loi normale ?
   (d) Calculer les 200 écarts-types corrigés-échantillon.
   (e) Calculer les 200 estimations ponctuelles de l'écart-type de $\overline{y}_W$.
   (f) Calculer les bornes inférieures et supérieures des 200 intervalles de confiance-échantillon pour la moyenne-population au niveau 95%. Combien d'entre eux contiennent réellement la moyenne-population (calculée à la question 1 (a)) ?
   (g) Proposer des commandes R permettant de visualiser les intervalles de confiances obtenus.

4. On considère maintenant la population des femmes, le caractère Y_f = "taille d'une femme" et 200 échantillons de 100 femmes suivant un plan de sondage aléatoire de type PESR. Reprendre l'étude complète avec ce nouveau contexte.

5. On considère maintenant la population des hommes, le caractère Y_h = "taille d'un homme" et 200 échantillons de 100 hommes suivant un plan de sondage aléatoire de type PESR. Reprendre l'étude complète avec ce nouveau contexte.

Exercice 5. Un professeur désire évaluer le nombre d'étudiants ayant eut la moyenne à l'examen de son module. Dans le paquet contenant 256 copies au total, il prélève un échantillon de 22 copies suivant un plan de sondage aléatoire de type PESR. Parmi cet échantillon, 12 étudiants ont eut la moyenne.

Utiliser R pour répondre aux questions suivantes :

1. Calculer le taux de sondage.

2. Donner une estimation ponctuelle de la proportion d'étudiants du module ayant eut la moyenne.

3. Donner une estimation ponctuelle de la variance de l'estimateur de la proportion d'étudiants du module ayant eut la moyenne.

4. Déterminer un intervalle de confiance au niveau 95% pour la proportion d'étudiants du module ayant eut la moyenne.

Exercice 6. On considère le jeu de données qui contient des informations sur la population belges et les revenus des habitants dans les communes.

Les données sont disponibles ici :

```
library(sampling)
data(belgianmunicipalities)
attach(belgianmunicipalities)
head(belgianmunicipalities)
```

La population U est l'ensemble des communes belges, les individus sont les communes et on considère le caractère X = "revenu moyen en euros par habitant en 2004" pour chaque commune (variable `averageincome`) :

```
U = belgianmunicipalities$Commune
str(U)
x = belgianmunicipalities$averageincome
str(x)
```

Partant de X, on étudie le caractère $Y = \mathbb{1}_{\{X \geq 23500\}}$ qui vaut 1 si la commune considérée présente un revenu moyen par habitant supérieur ou égal à 23500 euros en 2004 et 0 sinon. On s'intéresse à la proportion-population p_U : proportion des communes présentant un revenu moyen par habitant supérieur ou égal à 23500 euros en 2004.

1. Quel est le nombre total N de communes dans U ? Créer un vecteur `y` contenant les valeurs de Y pour toutes les communes. Calculer p_U.

2. Sélectionner un échantillon ω de 100 communes suivant un plan de sondage aléatoire de type PESR. Créer un vecteur `y_w` contenant les valeurs de Y des communes sélectionnées. Utiliser les formules vues en cours pour calculer la proportion-échantillon p_ω et construire un intervalle de confiance pour p_U au niveau 99%. Est-ce que cet intervalle contient p_U ?

3. On introduit la fonction `n_ech` :

```
n_ech = function(N, p_w, d0, niveau) {
z = qnorm(1 - (1 - niveau) / 2)
n = (N * p_w * (1 - p_w) * z^2) / (N * d0^2 + p_w * (1 - p_w) * z^2)
print(ceiling(n))
}
```

   Utiliser cette fonction pour déterminer la taille d'échantillon à choisir pour avoir une incertitude absolue sur p_U inférieure ou égale à 0.04 au niveau 98%.

4. On propose maintenant un plan de sondage aléatoire : le tirage systématique. Comprendre et reproduire la fonction ci-dessous qui illustre ce type de tirage :

```
tirage_sys = function(x, k) {
r = NULL
selection = NULL
r = sample(1:k, 1)
n = 0:  floor((length(x) - r) / k)
indices = (n * k) + r
selection = x[indices]
rbind(indices, selection)
}
```

Utiliser la fonction `tirage_sys` avec `k=6` pour sélectionner un échantillon ω de 100 communes environ. Créer un vecteur `yy_w` contenant les valeurs de Y des communes sélectionnées. Calculer p_ω et construire un intervalle de confiance pour p_U au niveau 99%. Contient-il p_U ?

Exercice 7. On considère une population de 2000 individus et on s'intéresse au caractère $Y =$ "taille" en centimètres. Les données sont disponibles ici :

```
r = read.table("https://chesneau.users.lmno.cnrs.fr/tailles-pop.txt",
header = T)
attach(r)
str(r)
```

1. Tracer sur un même graphique les boîtes à moustaches suivant le sexe des individus.

2. Calculer :

   - la proportion d'individus dépassant 169 centimètres,
   - la proportion de femmes dépassant la taille moyenne des hommes,
   - la proportion d'hommes dépassant la plus grande des femmes.

3. Sélectionner un échantillon ω de 152 individus suivant un plan de sondage aléatoire de type PESR. À partir de celui-ci, construire un intervalle de confiance pour la proportion de femmes dépassant 175.153 centimètres au niveau 90%.

4. Déterminer la taille d'échantillon à choisir pour avoir une incertitude relative sur la proportion de femmes dépassant 175.153 centimètres inférieure ou égale à 5% au niveau 90%.

Exercice 8. Décrire brièvement l'enjeu des commandes R suivantes :

```
library(sampling)
t = srswor(8, 20)
t2 = srswr(8, 20)
```

Puis expliquer les sorties :

```
> t
 [1] 1 1 0 0 0 0 0 1 1 0 1 0 0 0 0 0 1 1 0 1
> t2
 [1] 1 1 0 0 1 0 0 0 0 1 0 0 0 0 0 1 3 0 0 0
```

Exercice 9. On considère le jeu de données `islands`. Les données sont disponibles ici :

```
library(datasets)
data(islands)
islands
```

Sélectionner un échantillon de 8 îles suivant un plan de sondage aléatoire :

1. de type PESR,
   - avec la commande `sample`,
   - avec la commande `srswor`.
2. de type PEAR,
   - avec la commande `sample`,
   - avec la commande `srswr`.

Exercice 10. On demande à 65 élèves de maternelle de reproduire 16 dessins. On s'intéresse au temps en secondes mis par un élève. On considère un échantillon de 7 élèves suivant un plan de sondage aléatoire de type PEAR. Les résultats, en secondes, sont :

353	379	401	389	394	360	405

On suppose que le temps en secondes que met un élève de maternelle pour reproduire ces 16 dessins est une *var* Y suivant une loi normale.

Utiliser R pour déterminer un intervalle de confiance pour la moyenne des temps des 65 élèves au niveau 95%.

Exercice 11. On considère le jeu de données `nba` dans laquelle figurent 505 basketteurs de la NBA et diverses caractéristiques les concernant. Ce jeu de données est disponible ici :

```
https://chesneau.users.lmno.cnrs.fr/nba.txt
```

1. Mettre la liste des noms de joueurs dans un vecteur `U`.
2. Sélectionner un échantillon ω de 67 joueurs suivant un plan de sondage aléatoire de type PESR.
3. Sélectionner un échantillon $\omega 2$ de 67 joueurs suivant un plan de sondage aléatoire de type PEAR.
4. On considère maintenant le caractère quantitatif Y = "poids" (en pounds) dont les valeurs sur les individus de U sont listées dans le vecteur `Y` du jeu de données.
   (a) Calculer la moyenne-population $\overline{y}_U$.
   (b) En utilisant l'échantillon ω sélectionné au résultat de la question 2, déterminer l'erreur que commet la moyenne-échantillon $\overline{y}_\omega$ dans l'estimation de $\overline{y}_U$.
   (c) En utilisant l'échantillon $\omega 2$ sélectionné au résultat de la question 3, déterminer l'erreur que commet la moyenne-échantillon $\overline{y}_{\omega 2}$ dans l'estimation de $\overline{y}_U$.
   (d) Qui de $\overline{y}_\omega$ et $\overline{y}_{\omega 2}$ est le meilleur ?
5. Refaire 15 fois les questions 1 à 4 (a) (avec des échantillons différents a priori à chaque fois). En moyenne, qui de $\overline{y}_\omega$ et $\overline{y}_{\omega 2}$ est le meilleur ?
6. On utilise les vecteurs `U` et `Y` introduits aux questions précédentes. Reproduire et décrire brièvement l'enjeu des commandes R suivantes :

```
library(sampling)
liste = NULL
bar_y_w = NULL
bar_y_w2 = NULL
for (i in 1:100) {
bar_y_U = mean(Y)
t = srswor(67, length(U))
t2 = srswr(67, length(U))
bar_y_w[i] = sum(Y[t != 0]) / 67
bar_y_w2[i] = sum(Y[t2 != 0] * t2) / 67
liste = rbind(liste, c(bar_y_w[i], bar_y_w2[i], abs(bar_y_w[i] -
bar_y_U), abs(bar_y_w2[i] - bar_y_U), as.numeric(abs(bar_y_w[i] -
bar_y_U) > abs(bar_y_w2[i] - bar_y_U))))
}
liste
sum(liste[ ,5])
```

   Est-ce que cela rejoint la réponse formulée à la question 5 ? Quel résultat du cours cela illustre-t-il ?

Exercice 12. Reproduire et décrire brièvement l'enjeu des commandes R suivantes :

```
library(sampling)
data(swissmunicipalities)
table(swissmunicipalities$REG)
data = swissmunicipalities
data = data[order(data$REG), ]
st = strata(data, stratanames = c("REG"), size = c(30, 20, 45, 15, 20,
11, 44), method = "srswor")
getdata(data, st)
table(st$REG)
```

Exercice 13. On considère le jeu de données `nba` dans laquelle figurent 505 basketteurs de la NBA et diverses caractéristiques les concernant. Ce jeu de données est disponible ici :

```
https://chesneau.users.lmno.cnrs.fr/nba.txt
```

1. Mettre la liste des noms de joueurs dans un vecteur `U`.
2. On considère les strates des joueurs caractérisées par leur différent rôle sur le terrain ; ce sont les modalités G, F et C de la variable X2. Sélectionner un échantillon ω de 100 joueurs suivant un plan de sondage aléatoire de type STP.
3. On considère maintenant le caractère quantitatif $Y =$ "poids" (en pounds) dont les valeurs sur les individus de U sont listées dans le vecteur `Y` du jeu de données. Soit ω l'échantillon de 100 joueurs obtenu à la question 2.
   (a) Calculer la moyenne-population $\overline{y}_U$.
   (b) Calculer la moyenne-échantillon $\overline{y}_\omega$.
   (c) Donner une estimation ponctuelle de la variance de $\overline{y}_W$, où $\overline{y}_W$ désigne l'estimateur de la moyenne-population dans ce contexte.
4. Sélectionner un échantillon $\omega 2$ de 100 joueurs suivant un plan de sondage aléatoire de type STO.
5. Soit $\omega 2$ l'échantillon de 100 joueurs obtenu à la question 4.
   (a) Calculer la moyenne-échantillon $\overline{y}_{\omega 2}$.
   (b) Donner une estimation ponctuelle de la variance de $\overline{y}_W$, où $\overline{y}_W$ désigne l'estimateur de la moyenne-population dans ce contexte.
6. Entre $\overline{y}_\omega$ et $\overline{y}_{\omega 2}$, quelle estimation de $\overline{y}_U$ est la plus fine ?
7. Refaire 15 fois les questions 1 à 4 (a) (avec des échantillons différents a priori à chaque fois). En moyenne, qui de $\overline{y}_\omega$ et $\overline{y}_{\omega 2}$ est le meilleur ?

8. Reproduire et décrire brièvement l'enjeu des commandes R suivantes :

```
dat = read.table("https://chesneau.users.lmno.cnrs.fr/nba.txt",
sep = ",", header = T)
attach(dat)
U = dat[, 1]
n = 100
(n / length(U)) * table(X2)
n_h = c(18, 42, 40)
U_1 = Joueur[X2 == "C"]
U_2 = Joueur[X2 == "F"]
U_3 = Joueur[X2 == "G"]
library(sampling)
t_1 = srswor(n_h[1], length(U_1))
t_2 = srswor(n_h[2], length(U_2))
t_3 = srswor(n_h[3], length(U_3))
w_1 = U_1[t_1 == 1]
w_2 = U_2[t_2 == 1]
w_3 = U_3[t_3 == 1]
w = c(w_1, w_2, w_3)
U[w]
Y_1 = Y[X2 == "C"]
Y_2 = Y[X2 == "F"]
Y_3 = Y[X2 == "G"]
bar_y_w_1 = (1 / n_h[1]) * sum(Y_1 * t_1)
bar_y_w_2 = (1 / n_h[2]) * sum(Y_2 * t_2)
bar_y_w_3 = (1 / n_h[3]) * sum(Y_3 * t_3)
bar_y_w_h = c(bar_y_w_1, bar_y_w_2, bar_y_w_3)
N_h = c(length(Y_1), length(Y_2), length(Y_3))
N = sum(N_h)
bar_y_w = sum(N_h * bar_y_w_h) / N
bar_y_w
s_w_1 = sqrt(sum((Y_1 - bar_y_w_1)^2 * t_1) / (n_h[1] - 1))
s_w_2 = sqrt(sum((Y_2 - bar_y_w_2)^2 * t_2) / (n_h[2] - 1))
s_w_3 = sqrt(sum((Y_3 - bar_y_w_3)^2 * t_3) / (n_h[3] - 1))
s_w_h = c(s_w_1, s_w_2, s_w_3)
var_bar_y_w = (1 / N^2) * sum(N_h^2 * (1-n_h / N_h) * (s_w_h^2 / n_h))
var_bar_y_w
```

9. Reproduire et décrire brièvement l'enjeu des commandes R suivantes :

```
n_h_opt = ceiling(n * N_h * s_w_h / sum(N_h * s_w_h))
n_h_opt
```

10. Reprendre les questions précédente avec le caractère quantitatif $X1$ = "tailles" (en pouces).