BONNAUD-DUBOIS Manon

BESNIER Alexandre

SAE Échantillonnage et estimation

L'objectif de cette SAE est l'estimation d'une grandeur mesurée dans une population en utilisant deux méthodes d'échantillonnage : le sondage aléatoire simple à probabilités égales et l'échantillon par strates.

Ce processus d'échantillonnage se fera dans un premier temps par l'intermédiaire d'un sondage aléatoire simple à probabilité égales. Dans un second temps par l'intermédiaire d'un sondage ou échantillon par strates.

En comparant les résultats obtenus avec ces deux méthodes, nous évaluerons l'incertitude et la précision des estimations de la grandeur mesurée. Cette analyse nous permettra de comprendre les avantages et les limites de chaque méthode et d'identifier laquelle offre les résultats les plus fiables.

Ce rapport présentera ainsi une comparaison entre le sondage aléatoire simple et l'échantillon par strates pour estimer une grandeur mesurée dans une population. Nous discuterons de l'incertitude et de la précision des estimations obtenues par chaque méthode, ainsi que de leurs implications dans l'évaluation des résultats.

R

Partie 1

1)

Premièrement, près avoir importé dans un data frame table qui ressence les informations de toutes les communes de France, un sous-ensemble de données est extrait du tableau initial, contenant uniquement les lignes où la région est "Bretagne". Les colonnes sélectionnées sont "Coco", "Commune" et "Population.totale". Les 6 premières lignes de ce sous-ensemble sont ensuite affichées à l'aide de la fonction head(). Cela permet d'obtenir un aperçu des données sélectionnées.

```
21  #####Question 1
22  #data frame donnees
23  table <- read.csv2(file = "population_francaise_communes1.csv")
24  donnees = table[table$Nom.de.la.région=="Bretagne", ]
25  donnees <- subset(donnees, select=c(Coco, Commune, Population.totale))
26
27  # Afficher les 6 premieres lignes
28  head(donnees)
```

2)

Ici, la variable "U" est créée en extrayant la colonne "Coco" du sous-ensemble de données précédent. Ensuite, la taille de la population (nombre de communes) est calculée en utilisant la fonction length(), en se basant sur la variable "U". On trouve un n=1027, il y a donc 1027 communes.

```
32  #####Question 2
33  # Variables U avec les communes
34  U=donnees$Coco
35  U
36
37  # Taille de la population, 1 207 Communes
38  N=length(U)
39  N
```

3)

La variable "T" est calculée en faisant la somme des valeurs de la colonne "Population.totale" du sous-ensemble de données initial. Cela donne le nombre total d'habitants en Bretagne.

Ensuite, un échantillon aléatoire simple de taille 100 est tiré à partir de la variable "U" en utilisant la fonction sample(). Les 6 premières valeurs de cet échantillon sont affichées à l'aide de la fonction head().

```
#####Question 3
# Nombre d'habitants en bretagne
T= sum(donnees$Population.totale)
T

# Tirage aléatoire simple d'un échantillon de taille n=100 (communes)
n=100
E=sample(U,n)
head(E)
```

Ils sont au nombre de 3 463 439.

4)

Dans cette partie, un nouveau sous-ensemble de données appelé "donnees1" est créé en sélectionnant toutes les lignes du sous-ensemble initial où la valeur de la colonne "Coco" est présente dans l'échantillon aléatoire simple "E". Les 6 premières lignes de ce nouveau sous-ensemble sont affichées à l'aide de la fonction head().

```
#####Question 4
# Toutes les données avec les 100 communes de l'échantillon
donnees1= donnees[donnees$Coco %in% E, ]
head(donnees1)
```

5)

Un autre sous-ensemble de données appelé "donnees2" est créé en sélectionnant les colonnes "Coco" et "Population.totale" du sous-ensemble "donnees1". La longueur de la colonne "Coco" de ce sous-ensemble est calculée à l'aide de la fonction length(), et les 6 premières lignes sont affichées avec head().

La moyenne de la population totale de l'échantillon "donnees2" est calculée à l'aide de la fonction mean(), et stockée dans la variable "xbar". Ensuite, l'intervalle de confiance à 95% pour la moyenne est calculé à l'aide de la fonction t.test() et stocké dans la variable "idcmoy".

```
####Question 5
# Echantillon de communes avec le nombre d'habitants
donnees2 <- subset(donnees1, select=c(Coco, Population.totale))
length(donnees2$Coco)
head(donnees2)


### CALCULER LE NOMBRE MOYEN D'HABITANTS ET IDC A 95% ###
# moyenne d'échantillon
xbar=mean(donnees2$Population.totale)
xbar

# IDC de \mu (intervalle de confiance)
idcmoy=t.test(donnees2$Population.totale)$conf.int
idcmoy
```

6)

Le nombre total estimé d'habitants en Bretagne est calculé en multipliant la moyenne de l'échantillon "donnees2" par la taille de la population totale (variable "N"). Cela est stocké dans la variable "T_est". Ensuite, l'intervalle de confiance à 95% pour l'estimation du nombre total d'habitants (variable "T_est") est calculé en multipliant l'intervalle de confiance pour la moyenne (variable "idcmoy") par la taille de la population totale (variable "N"). La moitié de cette différence est calculée et stockée dans la variable "marge".

```
#### Question 6
# Nombre d'habitants total estimé
T_est = N*xbar
T_est


# IDC de T
idcT=idcmoy*N
idcT

marge=(idcT[2]-idcT[1])/2
marge

```

7)

Sondage simple			
T	Test	IDC	Marge
3463439,00	2506649,00	1979819 ; 303	526830,70
3463439,00	3166214,00	2418162 ; 391	748052,60
3463439,00	2902714,00	2131784 ; 367	770930,40
3463439,00	3911247,00	2142449 ; 568	1768798,00
3463439,00	3018128,00	2184509 ; 385	833618,90
3463439,00	2469003,00	1962320 ; 297	506683,20
3463439,00	2383499,00	1931767 ; 283	451732,10
3463439,00	2527591,00	1946266 ; 310	581324,70
3463439,00	3290849,00	2524803 ; 405	766046,70
3463439,00	6169822,00	607455.9 ; 11	5562366,00

8)

Il y a toujours une différence entre l'estimation et la réalité, dans notre cas de sondage aléatoire simple, notre échantillon peut ne contenir que des petites communes ou d'importantes communes, ceci conduit à des estimations imprécises. Pour obtenir des estimations plus fiables, il serait recommandé d'utiliser des méthodes d'échantillonnage plus avancées, telles que l'échantillonnage stratifié, qui permettent de réduire l'erreur d'estimation.

Partie 2 :

1)

On obtiens un résumé statistique des données de la colonne "Population.totale" en utilisant la fonction summary(). Cela permet de visualiser les quantiles de la distribution de cette variable. Ensuite, des strates sont créées en fonction de ces quantiles. Les strates sont définies comme suit : "<716", "entre 716 et 1 390", "entre 1 390 et 2 761", "plus de 2 761". Cela permet de regrouper les données en classes de population.

```
#### Question 1
# Les quantiles
summary(donnees$Population.totale)

#Créer des strats à partir des quantiles
# strates : <716, entre 716 et 1 390, entre 1 390 et 2 761, plus de 2 761
donnees$strate=cut(donnees$Population.totale, breaks = c(0, 716,  1390, 2761, Inf), labels=c(1,2,3,4))
```

2)

Dans cette section, un nouveau jeu de données appelé "datastrat" est créé en sélectionnant les colonnes "Coco", "Commune", "Population.totale" et "strate" à partir du tableau initial "donnees". Les 6 premières lignes de ce nouveau jeu de données sont affichées à l'aide de la fonction head(). Ensuite, les effectifs des différentes strates sont calculés à l'aide de la fonction table() et stockés dans la variable "Nh". La somme de ces effectifs est également calculée et stockée.

```
#### Question 2
datastrat=donnees[ ,c("Coco","Commune","Population.totale", "strate")]
head(datastrat)

# Tirer échantillon E en prenant des effectifs égaux dans les strates
# effectif des strates
data=datastrat[order(datastrat$strate), ]
Nh=table(data$strate)
Nh
sum(Nh)
```

3)

Un échantillon stratifié de taille 100 est tiré dans cette partie. La taille des échantillons dans chaque strate est fixée à 25. Le tirage est effectué sans remise dans les strates en utilisant la fonction strata() et les données sont obtenues à l'aide de la fonction getdata(). Les 6 premières lignes du nouvel échantillon sont affichées à l'aide de head(), et la longueur de la colonne "Commune" est calculée pour obtenir la taille effective de l'échantillon.

Ensuite, les poids des strates (gh) et les taux de sondage dans les strates (fh) sont calculés. Les poids des strates sont déterminés en divisant les effectifs des strates par la taille de la population totale (N). Les taux de sondage sont calculés en divisant la taille des échantillons dans chaque strate par les effectifs des strates.

```
#### Question 3
# Tirage d'un échantillon stratifié de taille n=100
n=100
nh=c(25,25,25,25)

# sondage strat (sans remise dans les strates)
st = strata(data, stratanames = c("strate"), size=nh, method="srswr")
data1 = getdata(data, st)
head(data1)
length((data1$Commune))

# Poids des strates
gh=Nh/N
gh

# Taux de sondage dans les strates
fh = nh/Nh
fh
```

4)

Les quatre sous-échantillons obtenus pour chaque strate sont définis dans cette section. Les moyennes (m1, m2, m3, m4) et les variances (var1, var2, var3, var4) des échantillons sont également calculées.

```
#### Question 4
# Définir les 4 sous-échantillons obtenus
ech1 = data1[data1$strate==1, ]
ech2 = data1[data1$strate==2, ]
ech3 = data1[data1$strate==3, ]
ech4 = data1[data1$strate==4, ]

# Moyennes des 4 échantillons
m1 = mean(ech1$Population.totale)
m2 = mean(ech2$Population.totale)
m3 = mean(ech3$Population.totale)
m4 = mean(ech4$Population.totale)

# Variances des 4 échantillons
var1 = var(ech1$Population.totale)
var2 = var(ech2$Population.totale)
var3 = var(ech3$Population.totale)
var4 = var(ech4$Population.totale)

```

5)

Ici, une estimation de la moyenne du nombre d'habitants (Xbarst) et de la variance de Xbarst est calculée. Xbarst est obtenu en pondérant les moyennes des échantillons (m1, m2, m3, m4) par les effectifs des strates (Nh) et en divisant par la taille de la population totale (N). La variance de Xbarst est estimée en prenant en compte les variances des échantillons, les poids des strates, les taux de sondage et les tailles des échantillons dans chaque strate.

```
#### Question 5
# Calculer une estimation Xbarst du nombre d'habitants moyen et la variance de Xbarst
# Moyenne générale (des 4 échant réunis)
Xbarst = (Nh[1]*m1 + Nh[2]*m2 + Nh[3]*m3 + Nh[4]*m4)/N

# Estimation de la variance de Xbarst
varXbarst = ((gh[1]^2)*(1-fh[1])*var1/(nh[1]) +
               (gh[2]^2)*(1-fh[2])*var2/(nh[2]) + (gh[3]^2)*(1-fh[3])*var3/(nh[3])
             + (gh[4]^2)*(1-fh[4])*var4/(nh[4]))
```

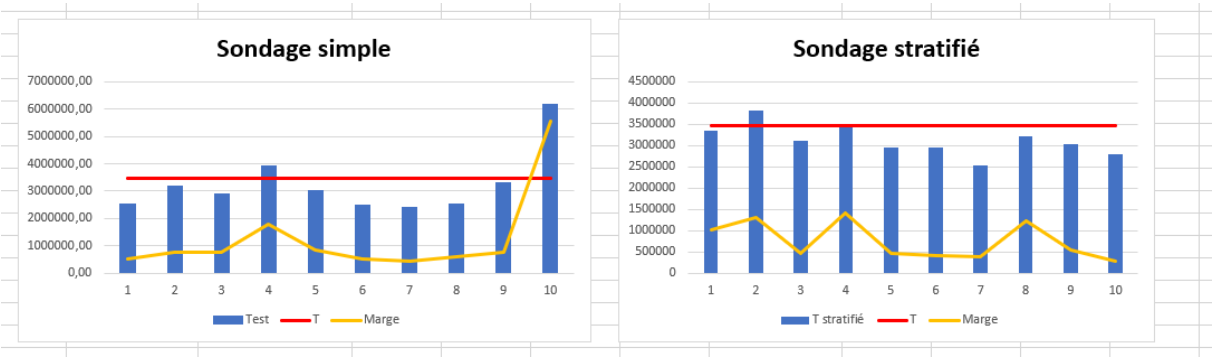
6)

Dans cette dernière partie, l'intervalle de confiance à 95% (IDC) pour la moyenne (mu) est calculé. Les bornes inférieure (binf) et supérieure (bsup) de l'IDC sont obtenues en utilisant la fonction qnorm() pour trouver les quantiles de la distribution normale standard. Ensuite, l'estimation du total (Tstr) est calculée en multipliant Xbarst par la taille de la population totale (N). L'IDC pour le total (T) est également calculé en multipliant les bornes de l'IDC pour la moyenne (idcmoy) par la taille de la population totale (N). Enfin, la marge d'erreur est calculée comme la moitié de la différence entre les bornes de l'IDC pour le total (idcT).

```
### Question 6
# IDC pour mu à 95%
alpha = 0.05
binf = Xbarst - qnorm(1-alpha/2)*sqrt(varXbarst)
bsup = Xbarst + qnorm(1-alpha/2)*sqrt(varXbarst)
idcmoy = c(binf, bsup)

# Estim du total T
Tstr = N*Xbarst
Tstr

# Estimation par IDC du Total T
binf = idcmoy[1]*N
bsup = idcmoy[2]*N
idcT = c(binf,bsup)
idcT

# Marge d'erreur
marge = (idcT[2]-idcT[1])/2
marge
```

7)

sondage Stratifié		
T stratifié	IDC	Marge
3349589	2327345 ; 437	1022244
3814013	2516566 ; 511	1297447
3106211	2640661 ; 357	465550,1
3497580	2097209 ; 489	1400371
2955714	2489454 ; 342	466260
2948953	2550695 ; 334	398258,3
2540658	2150491 ; 293	390167,3
3213725	1987172 ; 444	1226553
3032039	2482631 ; 358	549407,8
2807859	2532906 ; 308	274952,7



8)

Le sondage stratifié semble offrir des estimations plus précises du total de la population en Bretagne par rapport au sondage simple. Les intervalles de confiance sont plus étroits et les marges d'erreur sont généralement plus petites dans le sondage stratifié, ce qui indique une meilleure précision et une réduction de l'incertitude dans les estimations obtenues. Ainsi, le sondage stratifié est une méthode recommandée pour obtenir des estimations plus fiables de la population totale en Bretagne.

SAS

1-2.1)

Apres avoir importé les tables Note et Liste, la procédure "means" est utilisée pour calculer les statistiques descriptives des données contenues dans l'ensemble de données "Note". Les statistiques calculées sont : la valeur minimale (min), la valeur maximale (max), l'écart-type (std) et la médiane (Median). Le paramètre "MAXDEC = 2" est utilisé pour limiter le nombre de décimales affichées à 2.

```
/* Calcul des statistiques descriptives */
proc means data=Note MAXDEC=2 min max std Median;
run;
```

Noms et moyennes des étudiants par décision

La procédure MEANS

Variable	Libellé	Minimum	Maximum	Ec-type	Médiane
info	info	2.90	17.30	3.31	10.45
stat	stat	2.00	16.40	3.69	10.60
math	math	4.00	19.00	3.98	13.00

2.2)

L'ensemble de données "Note2" est créé à partir de l'ensemble de données "Note" en utilisant la commande "data". Une variable appelée "moyennes" est créée à l'aide de l'instruction "attrib" avec l'étiquette "moyenne". La variable "moyennes" contient la moyenne des variables "info", "stat" et "math", arrondie à l'entier le plus proche à l'aide de la fonction "round".

```
/* Création d'un nouvel ensemble de données Note2 */
data Note2;
set Note;
/* Attribut pour la variable moyennes */
attrib moyennes label="moyenne";
/* Calcul de la moyenne des variables info, stat et math */
moyennes = round((info + stat + math) / 3);
run;
```

	Nom	info	stat	math	moyenne
1	E1	17.3	12.1	16	15
2	E2	7.6	9.8	16	11
3	E3	11.6	11.1	10	11
4	E4	7.4	7.8	11	9
5	E5	11.3	8.3	14	11
6	E6	8	11.1	11	10
7	E7	15.4	14.5	15	15
8	E8	14.4	14.5	11	13
9	E9	6.2	13.4	13	11
10	E10	9.4	9.1	10	10
11	E11	12.3	15	17	15
12	E12	11.8	12.8	18	14
13	E13	13.2	9.4	17	13
14	E14	3.4	10.6	7	7
15	E15	4.8	10	14	10

2.3)

L'ensemble de données "Note3" est créé à partir de l'ensemble de données "Note2" en utilisant la commande "data". Une variable appelée "decision" est créée à l'aide de l'instruction "attrib" avec l'étiquette "decision" et la longueur définie à 20 caractères. Les valeurs de la variable "decision" sont attribuées en fonction de la valeur de la variable "moyennes". Si la valeur de "moyennes" est supérieure ou égale à 10, alors "admis" est assigné à "decision". Si la valeur de "moyennes" est inférieure à 10, alors "redoublant" est assigné à "decision".

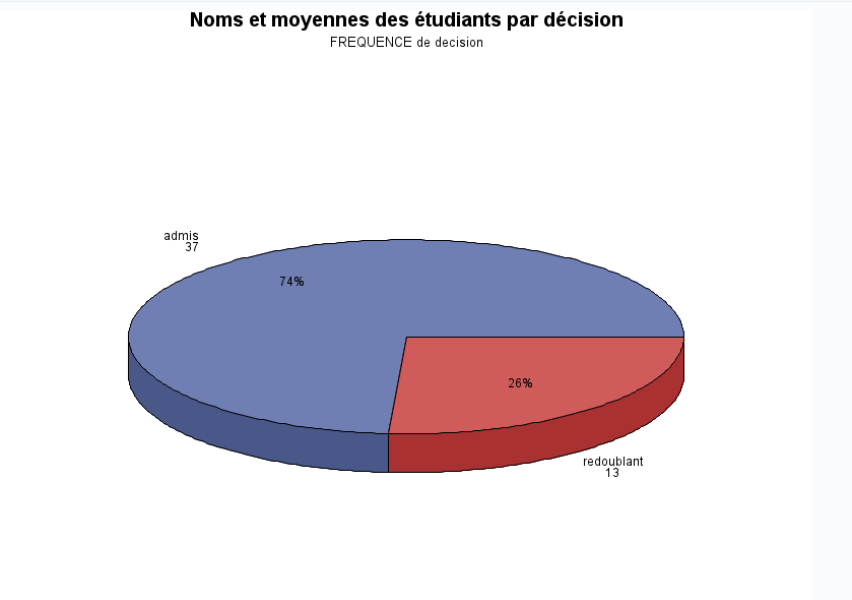
```
/* Création d'un autre nouvel ensemble de données Note3 */
data Note3;
set Note2;
/* Attribut pour la variable decision */
attrib decision label="decision" length=$20;
/* Assignation de la valeur "admis" si la moyenne est supérieure ou égale à 10 */
if (moyennes >= 10) then decision = "admis";
/* Assignation de la valeur "redoublant" si la moyenne est inférieure à 10 */
if (moyennes < 10) then decision = "redoublant";
run;
```

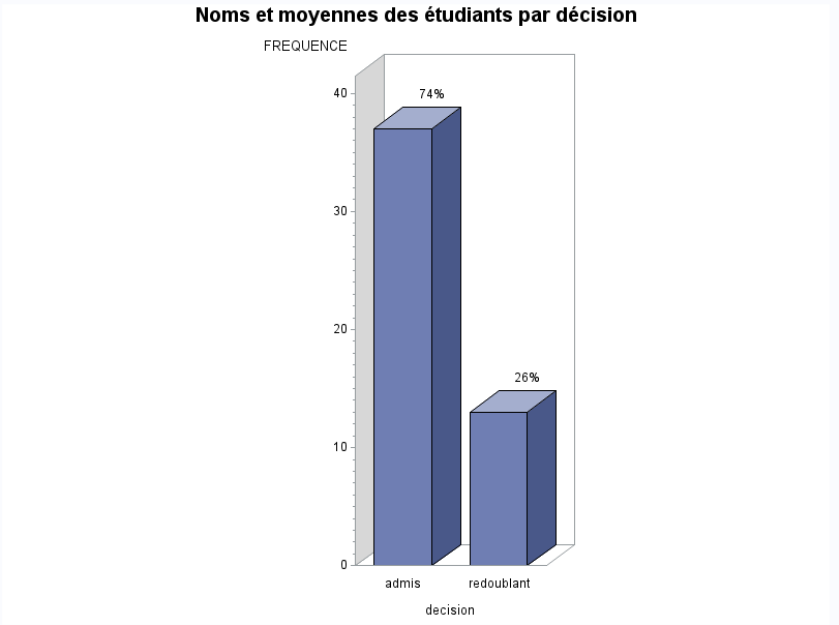
	Nom	info	stat	math	moyenne	decision
1	E1	17.3	12.1	16	15	admis
2	E2	7.6	9.8	16	11	admis
3	E3	11.6	11.1	10	11	admis
4	E4	7.4	7.8	11	9	redoublant
5	E5	11.3	8.3	14	11	admis
6	E6	8	11.1	11	10	admis
7	E7	15.4	14.5	15	15	admis
8	E8	14.4	14.5	11	13	admis
9	E9	6.2	13.4	13	11	admis
10	E10	9.4	9.1	10	10	admis
11	E11	12.3	15	17	15	admis
12	E12	11.8	12.8	18	14	admis
13	E13	13.2	9.4	17	13	admis
14	E14	3.4	10.6	7	7	redoublant
15	E15	4.8	10	14	10	admis
16	E16	7.1	5.4	11	8	redoublant
17	E17	3.9	2	9	5	redoublant
18	E18	7.2	11.2	12	11	admis

2.4)

La procédure "gchart" est utilisée deux fois pour créer un graphique en secteurs et un graphique en barres à partir de l'ensemble de données "Note3". La variable "decision" est utilisée pour créer le graphique en secteurs. Les paramètres "vbar3d" et "pie3d" sont utilisés pour spécifier le type de graphique en barres à afficher en 3D.

```
/* Création d'un graphique en secteurs avec la variable decision */
proc gchart DATA=Note3;
PIE3D decision;
/* Affichage des pourcentages à l'intérieur des secteurs */
pie3d decision/percent=inside;
run;

/* Création d'un graphique en barres tridimensionnelles avec la variable decision */
proc gchart data=Note3;
vbar3d decision/pct;
run;
```





2.5)

La procédure "sort" est utilisée pour trier l'ensemble de données "Note3" selon la variable "decision" par ordre croissant. Le tri permet ensuite d'afficher les données dans le même tableau.

La procédure "print" est utilisée pour afficher les données de l'ensemble de données "Note3". Le titre "Noms et moyennes des étudiants par décision" est affiché en utilisant l'instruction "title". Les variables affichées sont nom et moyennes.

```
/* Triage de l'ensemble de données Note3 selon la variable decision */
proc sort data=Note3;
by decision;
run;

/* Affichage des données des étudiants par décision (admis/redoublant) */
proc print data=Note3;
title "Noms et moyennes des étudiants par décision";
var Nom moyennes;
by decision;
/* Identifiant de la décision pour chaque groupe */
id decision;
run;
```

Noms et moyennes des étudiants par décision

decision	Nom	moyennes
admis	E1	15
	E10	10
	E11	15
	E12	14
	E13	13
	E15	10
	E18	11
	E2	11
	E20	13
	E21	12
	E23	14
	E24	10
	E25	11
	E27	10
	E29	11
	E3	11
	E30	12
	E31	12
	E32	16
	E33	11
	E34	10
	E36	14
	E37	12
	E38	10
	E39	14
	E41	14
	E44	12
	E45	14
	E46	15
	E47	16
	E49	10
	E5	11
	E50	14
	E6	10
	E7	15
	E8	13
	E9	11

decision	Nom	moyennes
redoublant	E14	7
	E16	8
	E17	5
	E19	9
	E22	9
	E26	4
	E28	8
	E35	6
	E4	9
	E40	9
	E42	5
	E43	5
	E48	5

3-3.1)

La procédure "sort" est utilisée pour trier l'ensemble de données "Note3" selon la variable "Nom" par ordre croissant. De même, l'ensemble de données "Liste" est trié selon la variable "Nom". Les deux ensembles de données sont triés afin de faciliter la fusion ultérieure.

```
/* Triage de l'ensemble de données Note3 selon la variable Nom */
proc sort data=Note3;
by Nom;
run;

/* Triage de l'ensemble de données Liste selon la variable Nom */
proc sort data=Liste;
by Nom;
run;

/* Fusion des ensembles de données Note3 et Liste basée sur la variable Nom */
data Note_Liste;
merge Note3(in=a) Liste(in=b);
by Nom;
/* Sélection des observations qui sont présentes dans les deux ensembles */
if a and b;
run;
```

3.2)

La procédure "means" est utilisée pour calculer les statistiques descriptives des variables "math", "stat", "info" et "age" dans l'ensemble de données "Note_Liste". La clause "VAR" est utilisée pour spécifier les variables à inclure dans le calcul des statistiques. La clause "CLASS" est utilisée pour spécifier la variable "sexe" comme variable de regroupement. La clause "TYPES" est utilisée pour spécifier que les statistiques doivent être calculées pour chaque niveau de la variable "sexe". Les statistiques calculées sont la moyenne (mean), l'écart-type (std), la médiane (median), la valeur minimale (minimum) et la valeur maximale (maximum). Les résultats sont enregistrés dans l'ensemble de données "Note_liste_stats".

```
/* Calcul des statistiques descriptives pour les variables math, stat, info et age, en regroupant par sexe */
PROC MEANS DATA=Note_liste MAXDEC=2;
VAR math stat info age;
CLASS sexe;
TYPES sexe;
OUTPUT OUT=Note_liste_stats
MEAN=mean
STD=std
MEDIAN=median
MIN=minimum
MAX=maximum;
RUN;
```

	sexe	_TYPE_	_FREQ_	math	math	math	math	math
1	F	1	28	11.75	4.4773090463	12	4	19
2	M	1	22	13.136363636	3.1667236267	14	6	19

Noms et moyennes des étudiants par décision

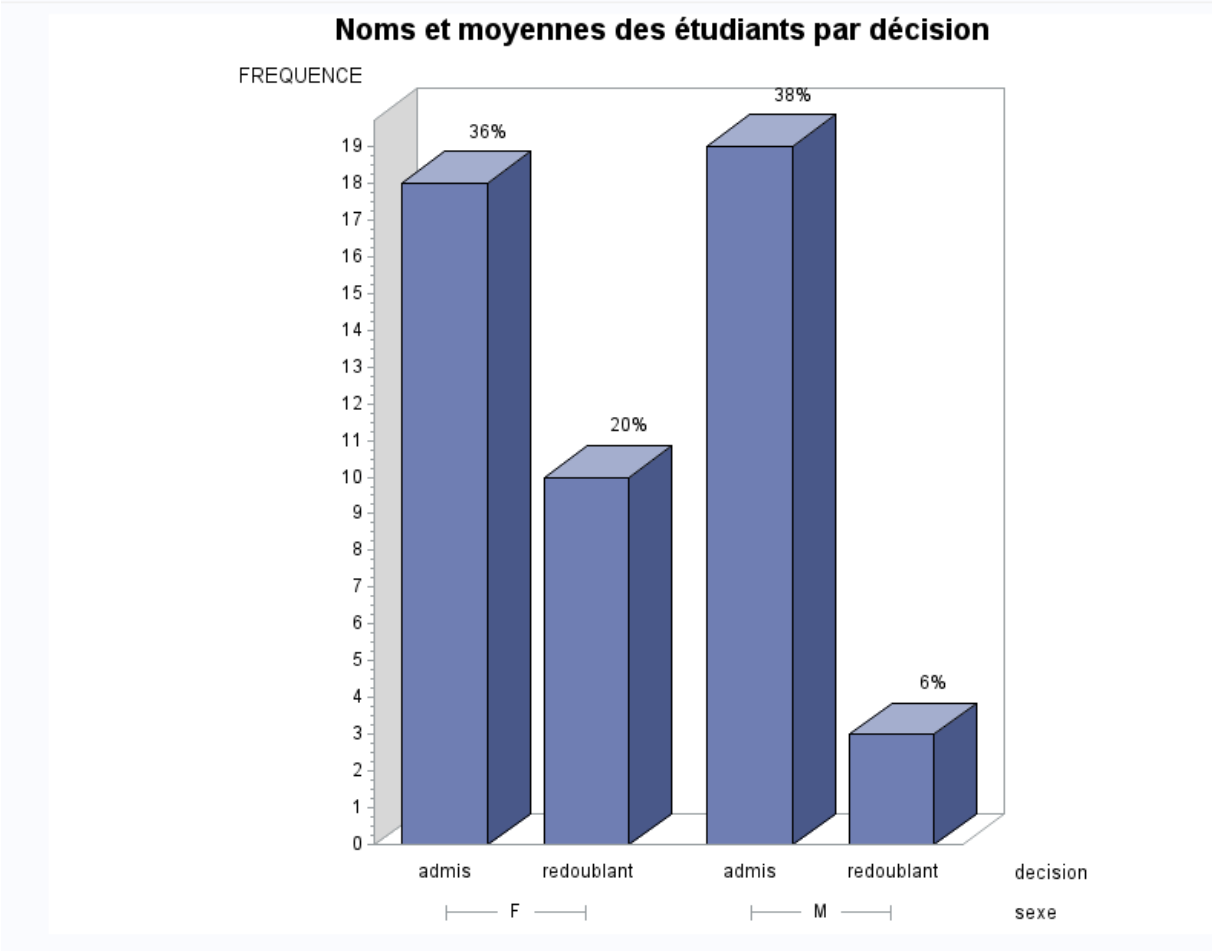
La procédure MEANS

sexe	N obs	Variable	Libellé	N	Moyenne	Ec-type	Minimum	Maximum
F	28	math	math	28	11.75	4.48	4.00	19.00
		stat	stat	28	10.21	4.01	2.00	16.40
		info	info	28	8.39	2.91	2.90	15.40
		age	age	28	21.86	1.15	20.00	24.00
M	22	math	math	22	13.14	3.17	6.00	19.00
		stat	stat	22	10.62	3.32	4.50	15.20
		info	info	22	12.22	2.48	5.90	17.30
		age	age	22	21.91	1.27	20.00	26.00

3.3)

La procédure "gchart" est utilisée pour créer un graphique en barres à partir de l'ensemble de données "Note_Liste". La variable "decision" est utilisée pour créer les barres. La clause "GROUP" est utilisée pour spécifier la variable "sexe" comme variable de regroupement. Ainsi, les barres sont affichées en fonction de la variable "sexe".

```
/* Création d'un graphique en barres tridimensionnelles avec la variable decision, en regroupant par sexe */
PROC gchart DATA=Note_liste;
vbar3d decision/pct GROUP=sexe;
RUN;
```



3.4)

La procédure "freq" est utilisée pour créer une table de contingence à partir de l'ensemble de données "Note_Liste". Les variables "sexe" et "decision" sont spécifiées dans la clause "TABLES" pour créer la table de contingence croisant les deux variables. Les résultats sont enregistrés dans l'ensemble de données "table_contingence".

```
/* Création d'une table de contingence pour les variables sexe et decision */
PROC FREQ DATA=Note_liste;
TABLES sexe * decision / OUT=table_contingence;
RUN;
```

	sexe	decision	Nombre d'occurrences	Pourcentage par rapport au total
1	F	admis	18	36
2	F	redoublant	10	20
3	M	admis	19	38
4	M	redoublant	3	6

3.5)

La procédure "freq" est utilisée à nouveau pour effectuer un test du chi2 de l'indépendance entre les variables "sexe" et "decision" en utilisant la table de contingence "table_contingence". Les variables sont spécifiées dans la clause "TABLES". Le paramètre "CHISQ" indique à la procédure de calculer le test du chi2. Le paramètre "ALPHA=0.05" spécifie le niveau de signification de 0,05 pour le test.

```
/* Test du chi2 de l'indépendance entre les variables sexe et decision */
PROC FREQ DATA=Table_contingence;
TABLES sexe * decision / CHISQ ALPHA=0.05;
/* Poids utilisé pour les fréquences de la table de contingence */
weight count;
RUN;
```

V de Cramer : La valeur observée est de -0.2498, ce qui indique une faible association négative.

Khi-2 : La valeur observée de la statistique du Khi-2 est de 3.1212 avec 1 degré de liberté et une p-valeur de 0.0773. Il existe une certaine association entre les variables sexe et décision, mais cette association n'est pas significative au niveau de signification de 0.05.2498, ce qui indique une faible association négative. Cela signifie que l'on ne peut pas rejeter l'hypothèse nulle d'indépendance entre le sexe et la décision à ce niveau de signification.

En résumé, les résultats indiquent qu'il existe une association entre le sexe et la décision, mais cette association est faible et n'est pas statistiquement significative au niveau de risque 5%.

Code R :

```
setwd("…")

table <- read.csv2(file = "population_francaise_communes.csv")
head(table)

library(sampling)

#####Question 1
#data frame donnees
donnees = table[table$Nom.de.la.r?gion=="Bretagne", ]
donnees <- subset(donnees, select=c(Coco, Commune, Population.totale))

# Afficher les 6 premieres lignes
head(donnees)



#####Question 2
# Variables U avec les communes
U=donnees$Coco
U

# Taille de la population, 1 207 Communes
N=length(U)
N
```

```
#####Question 3
# Nombre d'habitants en bretagne
T= sum(donnees$Population.totale)
T


# Tirage al?atoire simple d'un ?chantillon de taille n=100 (communes)
n=100
E=sample(U,n)
head(E)


#####Question 4
# Toutes les donn?es avec les 100 communes de l'?chantillon
donnees1= donnees[donnees$Coco %in% E, ]
head(donnees1)

####Question 5
# Echantillon de communes avec le nombre d'habitants
donnees2 <- subset(donnees1, select=c(Coco, Population.totale))
length(donnees2$Coco)
head(donnees2)


### CALCULER LE NOMBRE MOYEN D'HABITANTS ET IDC A 95% ###
# moyenne d'?chantillon
xbar=mean(donnees2$Population.totale)
xbar

# IDC de \mu (intervalle de confiance)
```

```
idcmoy=t.test(donnees2$Population.totale)$conf.int

idcmoy




#### Question 6

# Nombre d'habitants total estim?

T_est = N*xbar

T_est



# IDC de T

idcT=idcmoy*N

idcT


marge=(idcT[2]-idcT[1])/2

marge



#####

# Conclusion :

# Il y a toujours une diff?rence entre l'estimation et la r?alit?, dans notre cas de sondage al?atoire simple, notre ?chantillon

# peut ne contenir que des petites communes ou d'importantes communes, ceci conduit ? des estimantions impr?cises.

# Donc on utilisera un autre type de sondage, le sondage stratifi?.



################################################################################################################################
```

```
########################################## P A R T I E  2 :
#####################################################

#############################################################################################
##############################

library(sampling)

#### Question 1

# Les quantiles

summary(donnees$Population.totale)

#Cr?er des strats ? partir des quantiles

# strates : <716, entre 716 et 1 390, entre 1 390 et 2 761, plus de 2 761

donnees$strate=cut(donnees$Population.totale, breaks = c(0, 716,  1390, 2761, Inf),
labels=c(1,2,3,4))

#### Question 2

datastrat=donnees[ ,c("Coco","Commune","Population.totale", "strate")]

head(datastrat)

# Tirer ?chantillon E en prenant des effectifs ?gaux dans les strates

# effectif des strates

data=datastrat[order(datastrat$strate), ]

Nh=table(data$strate)

Nh
```

```
sum(Nh)




#### Question 3

# Tirage d'un ?chantillon stratifi? de taille n=100

n=100

nh=c(25,25,25,25)


# sondage strat (sans remise dans les strates)

st = strata(data, stratanames = c("strate"), size=nh, method="srswr")

data1 = getdata(data, st)

head(data1)

length((data1$Commune))


# Poids des strates

gh=Nh/N

gh


# Taux de sondage dans les strates

fh = nh/Nh

fh




#### Question 4

# D?finir les 4 sous-?chantillons obtenus

ech1 = data1[data1$strate==1, ]

ech2 = data1[data1$strate==2, ]

ech3 = data1[data1$strate==3, ]
```

```
ech4 = data1[data1$strate==4, ]

# Moyennes des 4 ?chantillons
m1 = mean(ech1$Population.totale)
m2 = mean(ech2$Population.totale)
m3 = mean(ech3$Population.totale)
m4 = mean(ech4$Population.totale)

# Variances des 4 ?chantillons
var1 = var(ech1$Population.totale)
var2 = var(ech2$Population.totale)
var3 = var(ech3$Population.totale)
var4 = var(ech4$Population.totale)


#### Question 5
# Calculer une estimation Xbarst du nombre d'habitants moyen et la variance de Xbarst
# Moyenne g?n?rale (des 4 ?chant r?unis)
Xbarst = (Nh[1]*m1 + Nh[2]*m2 + Nh[3]*m3 + Nh[4]*m4)/N

# Estimation de la variance de Xbarst
varXbarst = ((gh[1]^2)*(1-fh[1])*var1/(nh[1]) + (gh[2]^2)*(1-fh[2])*var2/(nh[2]) + (gh[3]^2)*(1-
fh[3])*var3/(nh[3]) + (gh[4]^2)*(1-fh[4])*var4/(nh[4]))


### Question 6
# IDC pour mu ? 95%
alpha = 0.05
binf = Xbarst - qnorm(1-alpha/2)*sqrt(varXbarst)
bsup = Xbarst + qnorm(1-alpha/2)*sqrt(varXbarst)
idcmoy = c(binf, bsup)
```

```
# Estim du total T
Tstr = N*Xbarst
Tstr

# Estimation par IDC du Total T
binf = idcmoy[1]*N
bsup = idcmoy[2]*N
idcT = c(binf,bsup)
idcT

# Marge d'erreur
marge = (idcT[2]-idcT[1])/2
marge
```

Code SAS :

```
/* Question 1: Calcul des statistiques descriptives */
proc means data=Note MAXDEC = 2 min max std Median  ;
run;


/*  Question 2: Création d'un ensemble de données Note2*/
data Note2;
set Note;
/* Création de la variable moyennes*/
attrib moyennes label="moyenne";
/* Calcul de la moyenne des variables info, stat et math*/
moyennes=round((info+stat+math)/3);
run;

/*  Question 3: Création d'un ensemble de données Note3*/
data Note3;
set Note2;
/* Création de la variable decision */
attrib decision label="decision" length= $20;
/* assignation des valeurs redoublant ou admi en fonction de si la moyenne est inférieur ou
supérieur à 10 */
if (moyennes>=10) then decision="admis";
if  (moyennes<10) then decision="redoublant";
run;

/*  Question 4: Création et affichage d'un graphique en secteurs avec la variable decision*/
proc gchart DATA=Note3;
PIE3D decision;
pie3d decision/percent=inside;
```

```
run;
/* Création et affichage d'un graphique en barres avec la variable decision*/
proc gchart data=Note3;
vbar3d decision/pct;
run;

/*  Question 5: Triage de Note3 selon la decision*/
proc sort data=Note3;
  by decision;
run;

/* Affichage des admis et redoublants dans des colonnes separees */
proc print data=Note3;
  title "Noms et moyennes des etudiants par decision";
  var Nom moyennes;
  by decision;
  id decision;
run;

/*  Question 6: Triage de Note3 et liste par Nom*/
proc sort data=Note3;
  by Nom;
run;
proc sort data=Liste;
  by Nom;
run;

/* Fusion de Note et Liste basée sur le Nom*/
data Note_Liste;
  merge Note3(in=a) Liste(in=b);
  by Nom;
```

```
   if a and b;
run;


/*  Question 7: Calcul des statistiques descriptives pour les variables math, stat, info et age en regroupant par sexe*/
PROC MEANS DATA=Note_liste MAXDEC = 2;
  VAR math stat info age;
  CLASS sexe;
  TYPES sexe;
  OUTPUT OUT=Note_liste_stats
    MEAN=mean
    STD=std
    MEDIAN=median
    MIN=minimum
    MAX=maximum;
RUN;
/*  Question 8: Création et affichage d'un garphique en barres avec la varriable decision selon le sexe */
PROC gchart DATA=Note_liste;
vbar3d decision/pct GROUP=sexe;
RUN;


/*  Question 9 :Création d'une table de contingence pour les variables sexe et decision */
PROC FREQ DATA=Note_liste;
TABLES sexe * decision / OUT=table_contingence;
RUN;


/* Test du chi2 de l'indépendance entre les variables sexe et decision */
PROC FREQ DATA=Table_contingence;
  TABLES sexe * decision / CHISQ ALPHA=0.05;
weight count;
RUN;
```