



[II.2313] Analyse de données - TP 2



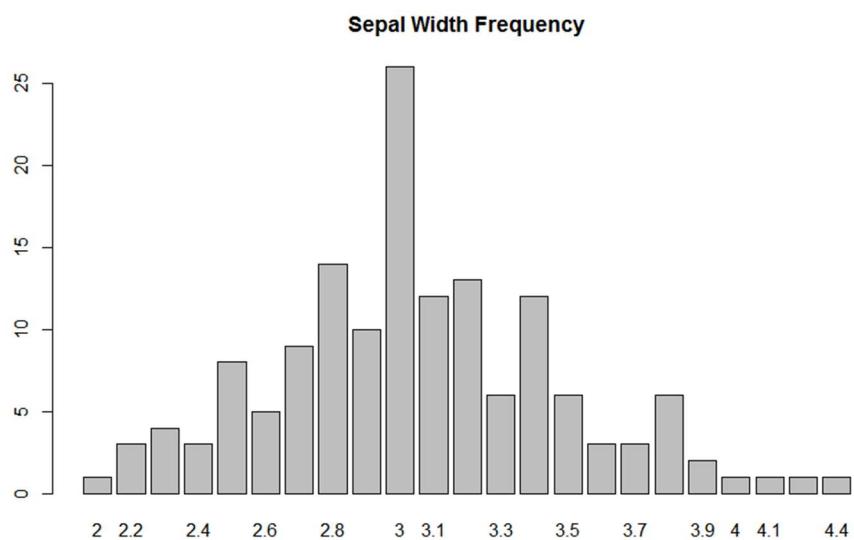
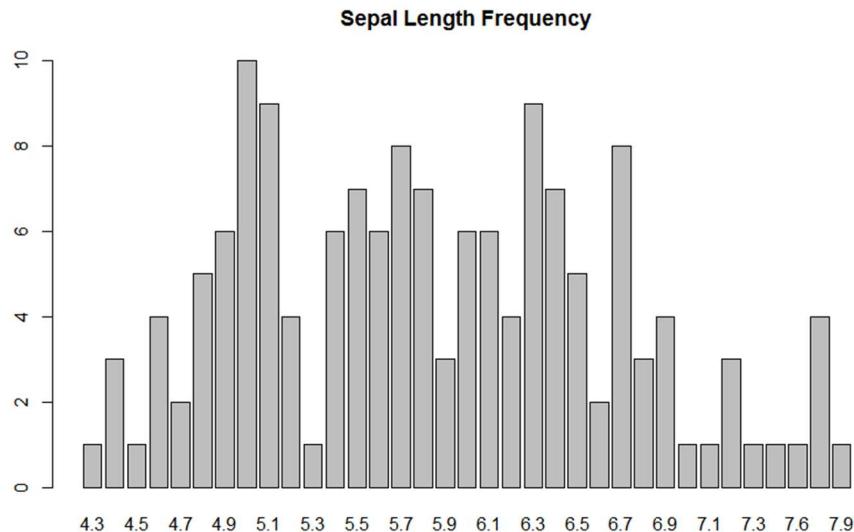
École d'ingénieurs du numérique

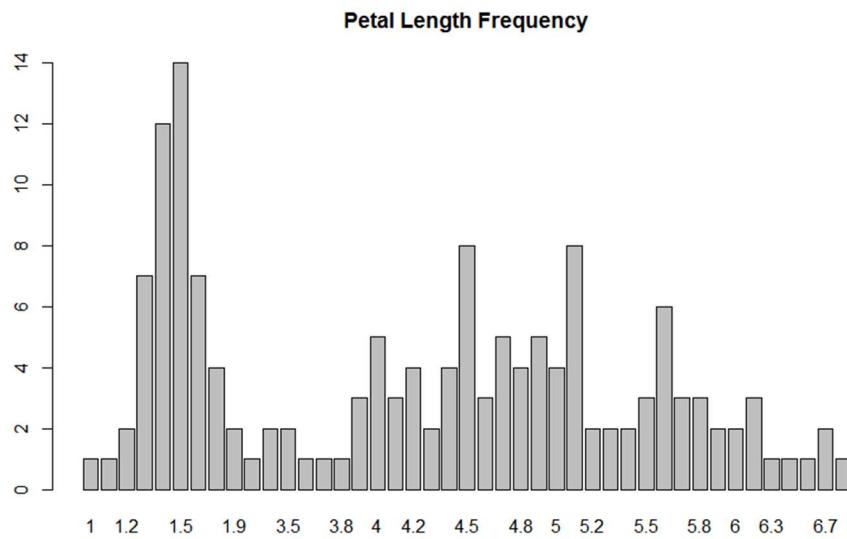
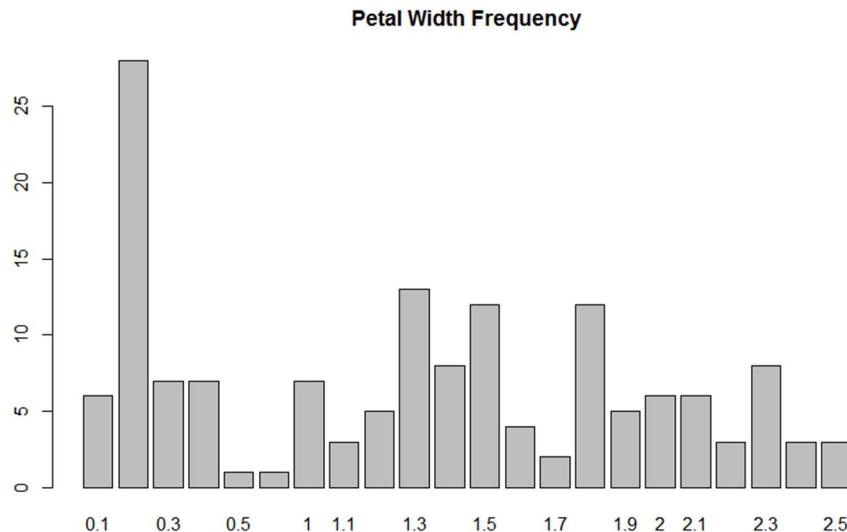
Rémi Biolley – Thierry Lincoln

1. Analyse multivariée : Iris de Fisher

L'objectif de cette partie est de se familiariser avec divers outils statistiques, il s'agit d'effectuer une analyse sur les iris de Fisher, un jeu de données sur les fleurs d'iris.

2. Que pouvez-vous dire sur leurs distributions ?





On constate que les distributions sont « plutôt gaussienne » pour les sépales, cependant pour les pétales c'est plutôt chaotique.

3. Calculs sans fonction R

Combinaison	COV	COR	Intervalle de confiance à 95% du coefficient de corrélation
<i>SepalWidth/SepalLength</i>	-0.039	-0.108	[-0.265, 0.052]
<i>SepalWidth/PetalLength</i>	-0.320	-0.418	[-0.544, -0.279]
<i>SepalWidth/PetalWidth</i>	-0.117	-0.354	[-0.489, -0.208]
<i>SepalLength/PetalLength</i>	1.265	0.866	[0.827, 0.906]

<i>SepalLength/SepalWidth</i>	0.513	0.813	[0.757, 0.865]
<i>PetalLength/PetalWidth</i>	1.288	0.956	[0.949, 0.973]

Voici le code qui permet de le faire sans utiliser les fonctions R.

```
CovSwSl=mean(tbl$SepalLength*tbl$SepalWidth)-mean(tbl$SepalLength)*mean(tbl$SepalWidth)
CorSwSl=CovSwSl/(sd(tbl$SepalLength)*sd(tbl$SepalWidth))
CorSwSl
[1] -0.1086401

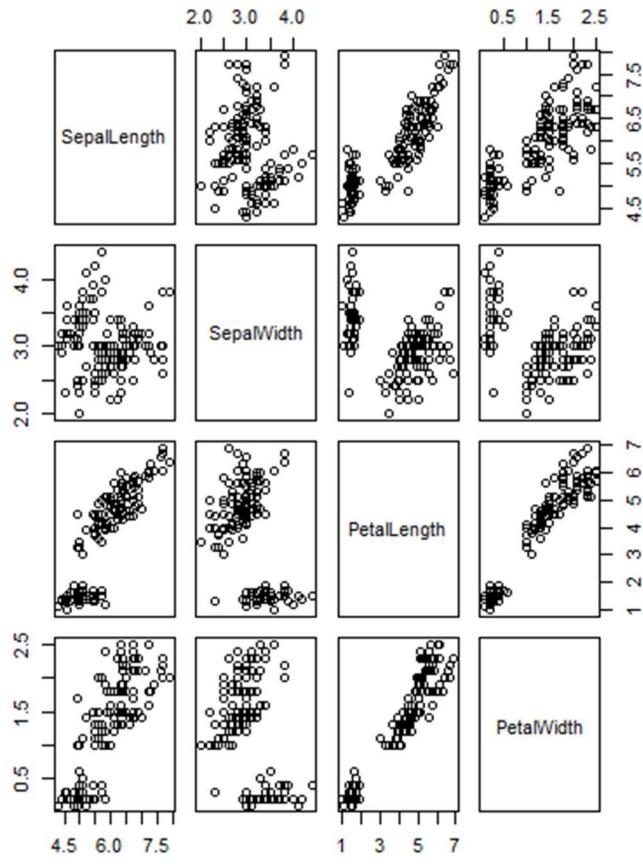
...
> CovPlPw=mean(tbl$PetalWidth*tbl$PetalLength)-mean(tbl$PetalWidth)*mean(tbl$PetalLength)
> CorPlPw=CovPlPw/(sd(tbl$PetalWidth)*sd(tbl$PetalLength))
> CorPlPw
[1] 0.9563387
> cor(tbl$PetalWidth,tbl$PetalLength)
[1] 0.9627571
```

4. Utilisez les commandes `cor(data)` et `plot(data)` pour confirmer vos résultats et les visualiser. Commentez.

Voici les résultats confirmés par la matrice de corrélation COR.

```
cor(tbl[1:4])
      SepalLength SepalWidth PetalLength PetalWidth
SepalLength  1.0000000 -0.1093692  0.8717542  0.8179536
SepalWidth   -0.1093692  1.0000000 -0.4205161 -0.3565441
PetalLength   0.8717542 -0.4205161  1.0000000  0.9627571
PetalWidth    0.8179536 -0.3565441   0.9627571  1.0000000
```

On constate quelques petites différences avec nos valeurs calculées « à la main ». Ceci s'explique par les approximations qu'effectuent les différentes fonctions de R.



Sur ce « plot », on peut voir qu'il y a une forte corrélation entre PetalLength/PetalWidth, ceci confirme la valeur obtenue plus haut 0.956.

Par ailleurs une corrélation se détecte entre SepalLength/PetalLength et SepalLength/PetalWidth mais est moins forte avec 0,866 et 0,813 respectivement

Cependant les attributs ayant un coefficient de corrélation compris entre [-0,7 ; 0,7] ne sont probablement pas dépendants. SepalWidth/SepalLength, SepalWidth/PetalLength , SepalWidth/PetalWidth sont dans ce cas. Par ailleurs leur intervalle de confiance à 95% est lui aussi dans l'intervalle [-0,7 ; 0,7], ce qui justifie d'avantage la non-dépendance.

5. En supposant que les attributs suivent une distribution normale, calculez les intervalles de confiance pour les différentes corrélations. Commentez vos résultats.

Les résultats se trouvent dans le Tableau ci-haut en question 3.

```
> cor.test(tbl$SepalWidth, tbl$SepalLength)

Pearson's product-moment correlation

data:  tbl$SepalWidth and tbl$SepalLength
t = -1.3386, df = 148, p-value = 0.1828
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
-0.26498618  0.05180021
sample estimates:
cor
-0.1093692
...
...
```

- On constate que plus la corrélation (ou anti-corrélation) semble forte, plus l'intervalle de confiance à 95 % est resserré.
Par exemple pour la forte corrélation entre PetalLength/PetalWidth (0.956) l'intervalle de confiance est de [0.949, 0.973] ce qui correspond à 0.024 de « marge ». L'intervalle est très serré.
Mais pour SepalLength/PetalWidth avec une corrélation de 0.813, l'intervalle de confiance est de [0.757, 0.865] la marge est de 0.108 ce qui est déjà plus grand.

B. Données multivariées : anthropométrie

On souhaite déterminer des possibles relations entre des données anthropométriques récupérées sur une population d'homme dans le cadre d'une étude universitaire.

1. Charger ces données en R sous forme d'une matrice.

On va dans un premier temps récupérer les données mises à disposition par l'université en les important grâce à la fonction *read.table()*. Pour pouvoir réaliser nos manipulations sur ces données avec plus de facilité par la suite, on convertit le data frame obtenu en matrice en utilisant *as.matrix()*.

```
mansize=read.table("mansize.csv", header=T, sep=";")
mansizeMat=as.matrix(mansize)
```

2. Que fait cette fonction ? Commentez les résultats sur vos données ?

En appliquant la fonction *summary()* à notre matrice on récupère les informations suivantes :

```
mansize=read.table("mansize.csv", header=T, sep=";")
mansizeMat=as.matrix(mansize)
summary(mansizeMat)
```

Age	Height..cm.	Weight..kg.	Femur.Length..cm.	Feet.Size..cm.	Arm.span..cm.	Hand.length..cm.	Cranial.volume..cm3.	Penis.size..cm.
Min. :18.00	Min. :150.0	Min. : 40.00	Min. :37.10	Min. :18.90	Min. :159.6	Min. :15.80	Min. :1298	Min. : 9.10
1st Qu.:19.00	1st Qu.:165.0	1st Qu.: 63.10	1st Qu.:43.60	1st Qu.:23.10	1st Qu.:176.3	1st Qu.:18.20	1st Qu.:1382	1st Qu.:12.50
Median :20.00	Median :172.0	Median : 71.50	Median :47.40	Median :25.10	Median :181.7	Median :18.90	Median :1418	Median :13.40
Mean :20.45	Mean :173.2	Mean : 73.36	Mean :47.52	Mean :24.97	Mean :183.0	Mean :18.89	Mean :1418	Mean :13.39
3rd Qu.:22.00	3rd Qu.:181.0	3rd Qu.: 81.10	3rd Qu.:51.30	3rd Qu.:26.70	3rd Qu.:188.9	3rd Qu.:19.80	3rd Qu.:1450	3rd Qu.:14.30
Max. :24.00	Max. :203.0	Max. :115.20	Max. :62.10	Max. :32.20	Max. :206.9	Max. :22.60	Max. :1558	Max. :18.40

La fonction *summary()* nous a permis de récupérer une multitude d'informations utiles sur chaque colonne de notre matrice : les premier et troisième quartiles, la moyenne, la médiane ainsi que les valeurs minimum et maximum de chaque catégorie.

On remarque notamment que les moyennes et les médianes ont des valeurs particulièrement proches l'une de l'autre ...

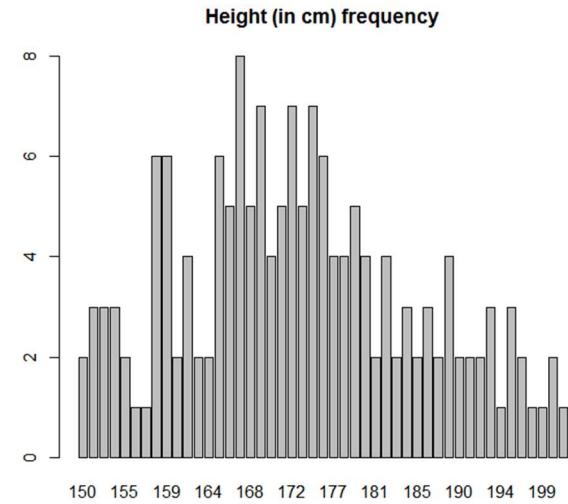
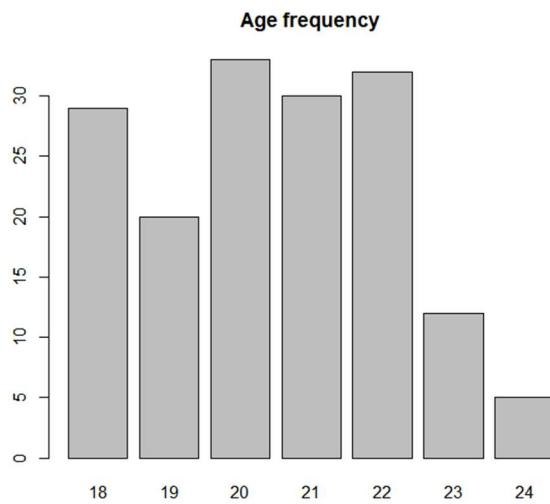
3. Que pouvez-vous dire sur leur distribution ?

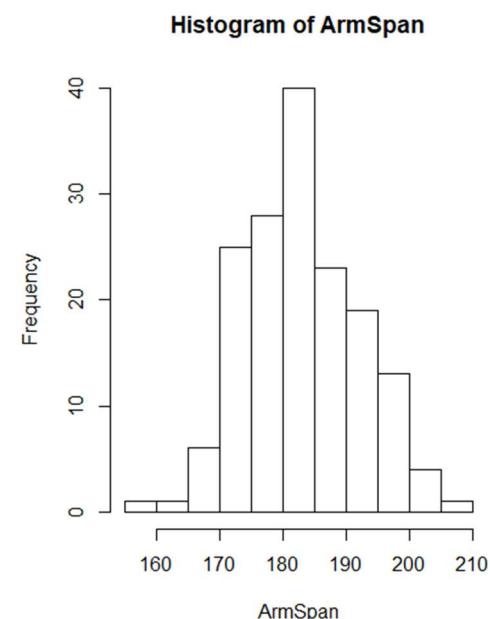
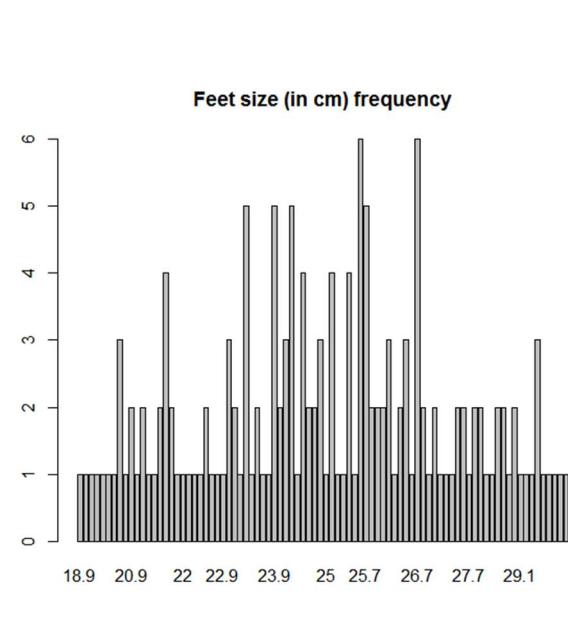
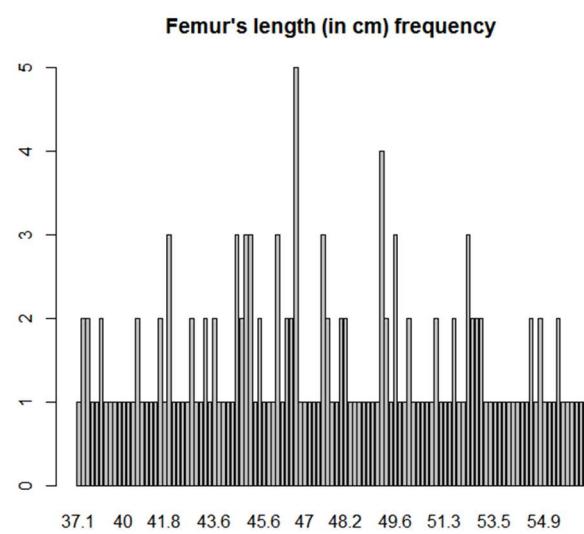
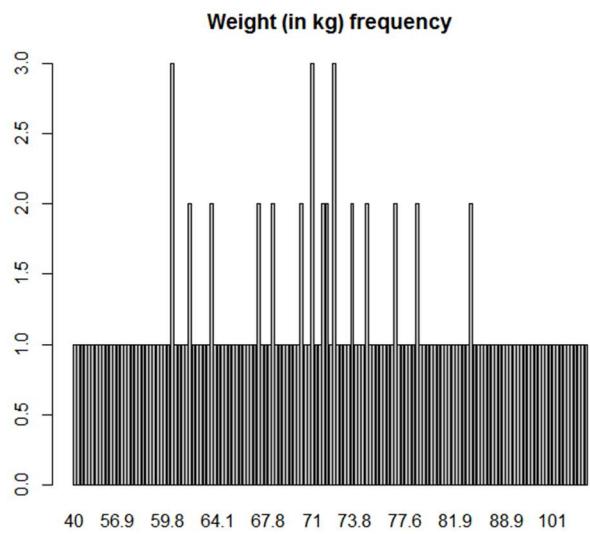
Pour mieux visualiser les données contenues dans la matrice, on se propose de les afficher sous forme d'histogrammes grâce aux fonctions *barplot()* et *hist()* (selon la lisibilité que donne chaque fonction au cas par cas).

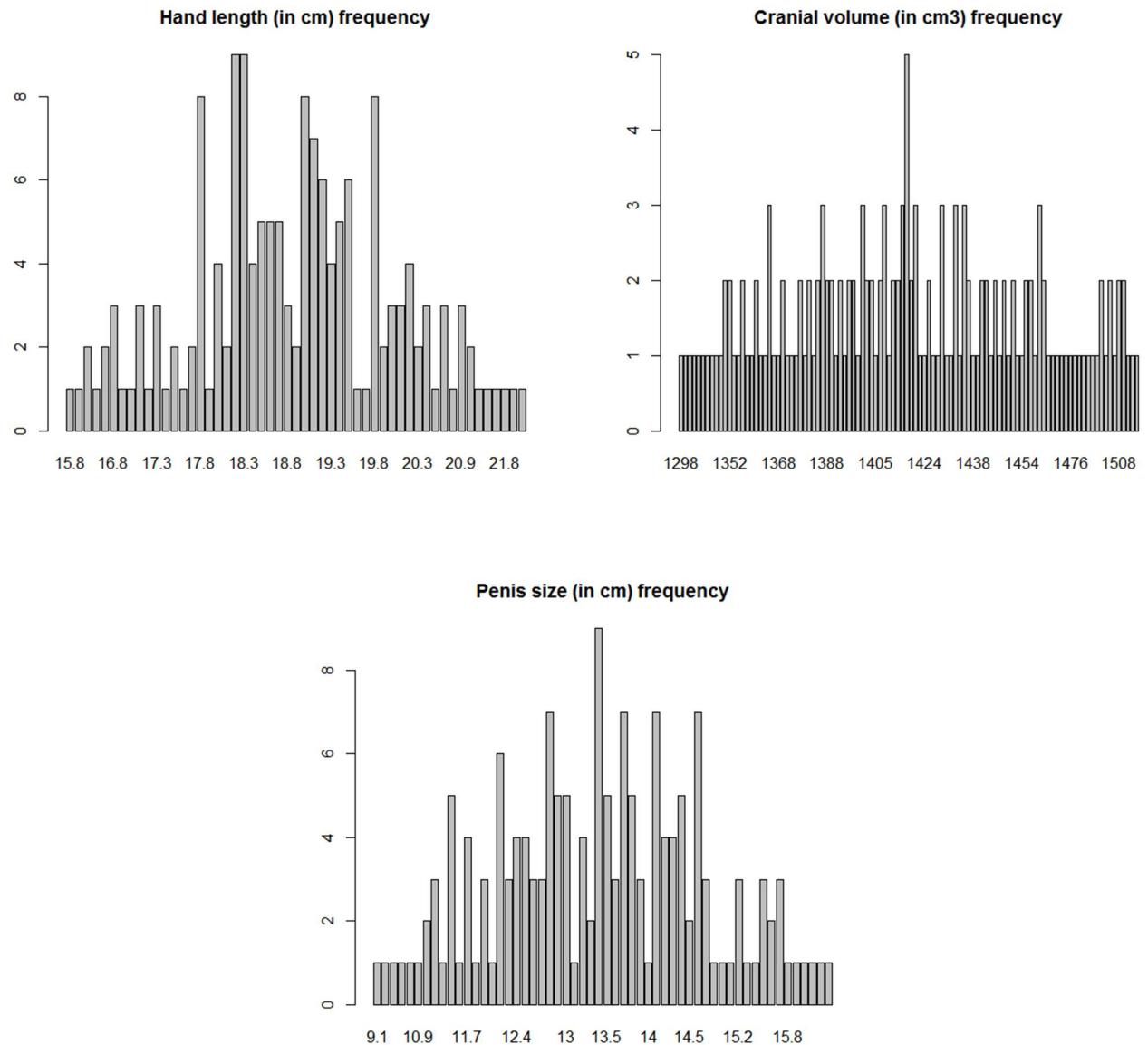
```

barplot(table(mansizeMat[,1]), main = "Age frequency")
barplot(table(mansizeMat[,2]), main = "Height (in cm) frequency")
barplot(table(mansizeMat[,3]), main = "Weight (in kg) frequency")
barplot(table(mansizeMat[,4]), main = "Femur length (in cm) frequency")
barplot(table(mansizeMat[,5]), main = "Feet size (in cm) frequency")
hist(mansizeMat[,6])
barplot(table(mansizeMat[,7]), main = "Hand length (in cm) frequency")
barplot(table(mansizeMat[,8]), main = "Cranial volume (in cm3) frequency")
barplot(table(mansizeMat[,9]), main = "Penis size (in cm) frequency")

```







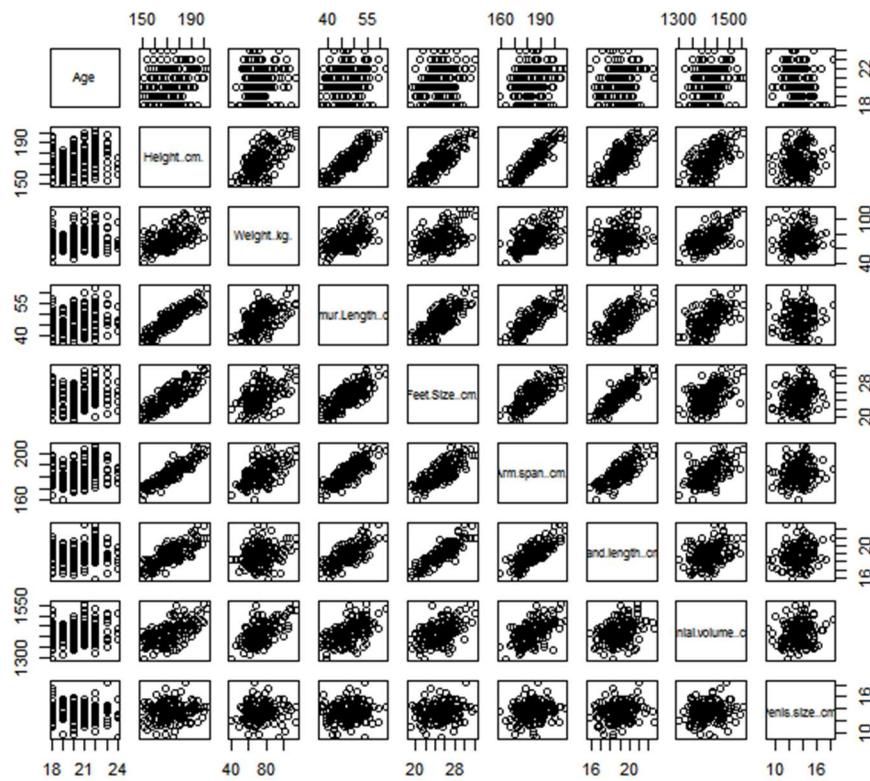
Les différentes catégories de notre matrice ainsi représentées, on voit que toutes semblent suivre une loi normale. Ceci explique la proximité des valeurs de moyenne et de médiane que nous avons remarquée précédemment.

4. Commentez. Que pouvez-vous dire sur l'utilisation en archéologie de la longueur du fémur pour prédire la taille d'un individu ?

On souhaite dorénavant déterminer la corrélation qui peut exister entre chaque catégorie de notre jeu de données.

Pour cela on utilise la fonction `plot()` sur notre dataframe `mansize`. Cette fonction, appliquée sur une variable de type dataframe, permet d'afficher une “matrice” de nuages de points représentant les valeurs de chaque variable selon les autres.

plot(mansize)



En observant ces nuages de points, on peut envisager des corrélations entre certaines variables. Un nuage de points se rapprochant d'une forme de droite témoigne d'une corrélation probable (par exemple pour les couples *Height/ArmSpan* ou *FeetSize/HandLength* ...).

On peut le vérifier en calculant les coefficients de corrélation de chaque couple de variables grâce à la fonction *cor()* et en lui passant en argument notre matrice *mansizeMat* (ou notre dataframe *mansize*).

cor(mansizeMat) ou cor(mansize)

	Age	Height..cm.	Weight..kg.	Femur.Length..cm.	Feet.Size..cm.	Arm.span..cm.	Hand.length..cm.	Cranial.volume..cm3
Age	1.00000000	0.1980260	0.14680238	0.2125544	0.2267084	0.2217911	0.1663874	0.1789928
Height..cm.	0.1980260	1.0000000	0.59151560	0.8905730	0.8024375	0.9032031	0.7915676	0.6246087
Weight..kg.	0.14680238	0.5915156	1.00000000	0.5170937	0.4394847	0.5605218	0.2186416	0.5999177
Femur.Length..cm.	0.21255435	0.8905730	0.51709372	1.0000000	0.7542053	0.8232579	0.7420294	0.5800319
Feet.Size..cm.	0.22670837	0.8024375	0.43948467	0.7542053	1.0000000	0.7583445	0.8710756	0.5046619
Arm.span..cm.	0.22179113	0.9032031	0.56052176	0.8232579	0.7583445	1.0000000	0.7951266	0.5599201
Hand.length..cm.	0.16638740	0.7915676	0.21864163	0.7420294	0.8710756	0.7951266	1.0000000	0.3861985
Cranial.volume..cm3	0.17899279	0.6246087	0.59991769	0.5800319	0.5046619	0.5599201	0.3861985	1.0000000
Penis.size..cm.	-0.07167913	0.1273751	0.06840311	0.1005534	0.1763982	0.1401550	0.1827799	0.1242201
Penis.size..cm.								
Age	-0.07167913							
Height..cm.	0.12737507							
Weight..kg.	0.06840311							
Femur.Length..cm.	0.10055335							
Feet.Size..cm.	0.17639823							
Arm.span..cm.	0.14015502							
Hand.length..cm.	0.18277993							

La fonction nous a renvoyé une matrice contenant tous les coefficients de corrélation existant pour notre jeu de données. On voit, entre autres, que le coefficient de corrélation entre *Femur.length* et *Height* est très proche de

1 (il est égal à 0.89), il semblerait alors envisageable d'estimer la taille d'un individu grâce à la longueur de son fémur en archéologie par exemple.

5. Calculez les intervalles de confiance pour vos coefficients de corrélation (on fera l'hypothèse que les attributs suivent tous des distributions normales). Commentez vos résultats.

Suite à notre détermination des coefficients de corrélation de notre jeu de données, on décide de calculer leur intervalle de confiance à 95% en utilisant `cor.test()` sur tous nos couples de variables (on suppose que les variables suivent une loi normale).

```
cor.test(mansize$Age, mansize$Height..cm.)
cor.test(mansize$Age, mansize$Weight..kg.)
```

...

	Age	Height	Weight	Femur.length	Feet.size	Arm.span	Hand.length	Cranial.volum	Penis.size
Age		[0.04, 0.34]	[-0.01, 0.29]	[0.06, 0.36]	[0.07, 0.37]	[0.07, 0.36]	[0.01, 0.31]	[0.03, 0.32]	[-0.22, 0.08]
Height	[0.04, 0.34]		[0.48, 0.68]	[0.85, 0.92]	[0.74, 0.85]	[0.87, 0.93]	[0.73, 0.84]	[0.52, 0.71]	[-0.03, 0.28]
Weight	[-0.01, 0.29]	[0.48, 0.68]		[0.39, 0.62]	[0.31, 0.56]	[0.44, 0.66]	[0.07, 0.36]	[0.49, 0.69]	[-0.09, 0.22]
Femur.length	[0.06, 0.36]	[0.85, 0.92]	[0.39, 0.62]		[0.68, 0.81]	[0.77, 0.87]	[0.66, 0.80]	[0.47, 0.67]	[-0.05, 0.25]
Feet.size	[0.07, 0.37]	[0.74, 0.85]	[0.31, 0.56]	[0.68, 0.81]		[0.68, 0.82]	[0.83, 0.90]	[0.38, 0.61]	[0.02, 0.32]
Arm.span	[0.07, 0.36]	[0.87, 0.93]	[0.44, 0.66]	[0.77, 0.87]	[0.68, 0.82]		[0.73, 0.85]	[0.44, 0.66]	[-0.01, 0.29]
Hand.length	[0.01, 0.31]	[0.73, 0.84]	[0.07, 0.36]	[0.66, 0.80]	[0.83, 0.90]	[0.73, 0.85]		[0.25, 0.51]	[0.03, 0.33]
Cranial.volume	[0.03, 0.32]	[0.52, 0.71]	[0.49, 0.69]	[0.47, 0.67]	[0.38, 0.61]	[0.44, 0.66]	[0.25, 0.51]		[-0.03, 0.27]
Penis.size	[-0.22, 0.08]	[-0.03, 0.28]	[-0.09, 0.22]	[-0.05, 0.25]	[0.02, 0.32]	[-0.01, 0.29]	[0.03, 0.33]	[-0.03, 0.27]	

En analysant les intervalles de confiance à 95% des coefficients de corrélation, on peut déterminer quels éléments ont de grande chance de dépendre les uns des autres malgré l'incertitude du calcul de leur coefficient de corrélation. On prend pour décision de n'accepter que les coefficients dont l'intervalle de confiance est disjoint de [-0.7, 0.7] (ainsi le coefficient de détermination est supérieur ou égal à 0.5).

Les couples répondant à ces exigences sont :

- *FemurLength/Height*
- *FeetSize/Height*
- *ArmSpan/Height*

- $\text{ArmSpan}/\text{FemurLength}$
- $\text{HandLength}/\text{Height}$
- $\text{HandLength}/\text{FeetSize}$
- $\text{HandLength}/\text{ArmSpan}$
- $(\text{FeetSize}/\text{FemurLength})$
- $(\text{ArmSpan}/\text{FeetSize})$
- $(\text{HandLength}/\text{FemurLength})$

6. Que pouvez-vous dire des liens entre les différentes variables anthropométriques de ces données ?

Une corrélation semble exister entre la longueur du fémur, la taille, la longueur du bras, la taille des pieds et la taille des mains. En effet, ce sont les seules caractéristiques qui répondent aux exigences de corrélation que l'on s'est fixé. L'âge, le volume crânien, le poids et la taille du pénis semblent quant à eux ne pas être acceptables pour déterminer des caractéristiques du corps de l'homme.

Il faudrait maintenant montrer qu'il existe plus qu'une relation de corrélation entre nos caractéristiques sélectionnées, mais bien une relation de causalité.

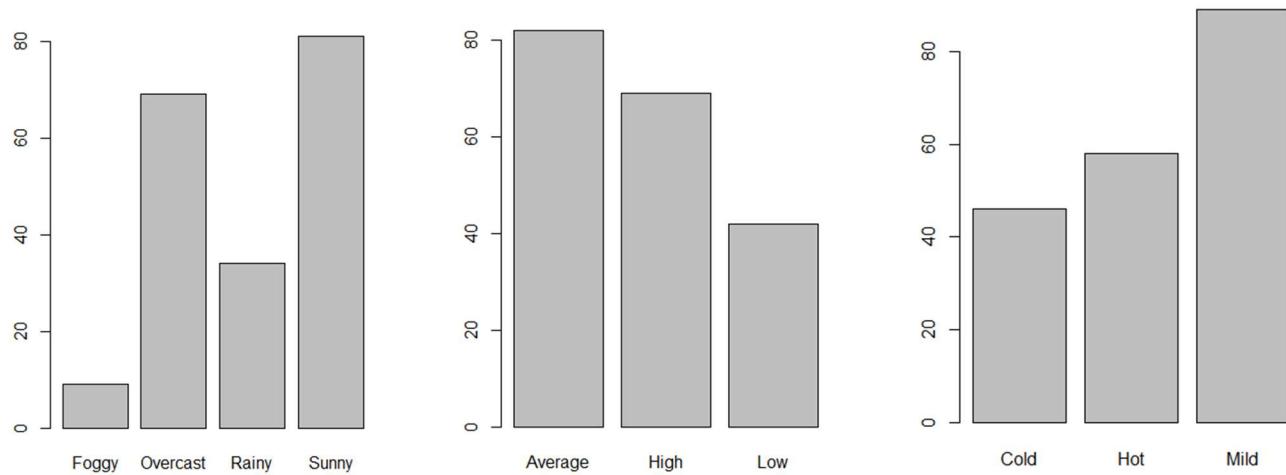
C. Test d'indépendance et variables catégorielles

On souhaite étudier la possibilité d'une dépendance entre le temps qu'il fait, la température et l'humidité. Pour cela on dispose d'un jeu de données qui regroupe ces attributs pris dans 193 villes.

1. Décrivez les différentes variables et leurs valeurs à partir d'histogrammes

On récupère dans un premier temps ces données à l'aide de la fonction `read.table()` et on les visualise à l'aide d'un histogramme pour chacun des attributs (excepté le nom des villes).

```
barplot(table(weather$Outlook))
barplot(table(weather$Humidity))
barplot(table(weather$Temperature))
```



On remarque que les données de la colonne *Outlook* peuvent prendre 4 valeurs : *Foggy*, *Overcast*, *Rainy* et *Sunny*. La valeur *Sunny* est la plus représentée. A contrario, la valeur *Foggy* est très peu présente dans le jeu de données. La caractéristique *Outlook* ne semble pas suivre de loi particulière.

Pour ce qui est des colonnes *Humidity* et *Temperature*, toutes deux possèdent 3 valeurs différentes et leurs valeurs les plus représentées sont respectivement *Average* et *Mild*. Ces caractéristiques semblent suivre un modèle linéaire.

2. Commentez la répartition des données dans le tableau résultant et déduisez le nombre de degrés de liberté de ce problème.

On souhaite maintenant évaluer la possible dépendance entre les attributs.

On s'intéresse dans un premier temps au lien entre *Outlook* et *Température*. On crée la table de contingence entre ces variables afin de se préparer à la détermination du χ^2 et de déduire le nombre de degrés de liberté. Pour cela on utilise la fonction **table()** en lui passant en argument les deux colonnes à étudier.

```
contingOT=table(weather$Outlook, weather$Temperature)
```

On obtient le tableau suivant :

	Cold	Hot	Mild
Foggy	4	3	2
Overcast	19	14	36
Rainy	6	11	17
Sunny	17	30	34

Le tableau de contingence permet de visualiser combien de lignes du jeu de données possèdent telles ou telles combinaisons de valeurs pour les variables *Outlook* et *Temperature*.

On peut aussi déterminer le nombre de degrés de liberté entre ces variables grâce à la formule $(N_c - 1) * (N_l - 1)$ où N_c et N_l représentent respectivement le nombre de colonnes et le nombre de lignes du tableau. Ici on a donc **6 degrés de liberté**.

On remarque que les valeurs de chaque ligne/colonne ne sont pas réparties de façon proportionnelle à la taille des catégories. On peut donc supposer qu'il existe une potentielle dépendance entre les variables mais l'on aura besoin d'autres outils pour tenter de confirmer cette hypothèse.

3. A partir des résultats et éventuellement en calculant d'autres indices, que pouvez vous conclure sur la dépendance entre les 2 variables ?

On va dorénavant pouvoir étudier la dépendance entre les variables *Outlook* et *Temperature*. Pour cela on utilise le test du Chi carré. R nous permet de réaliser ce test grâce à la simple commande *chisq.test()* qui prend en argument le tableau de contingence que l'on étudie.

```
chisq.test(contingOT)

Pearson's Chi-squared test

data: contingOT
X-squared = 8.4933, df = 6, p-value = 0.2041
```

La fonction nous renvoie plusieurs informations :

- *X-squared* est la valeur du χ^2
- *Df* est le nombre de degrés de liberté (on remarque que l'on avait déterminé la bonne valeur)
- *p-value*

La valeur du χ^2 n'étant pas nulle, on ne peut pas dire si les variables sont indépendantes ou non. On s'intéresse donc à la valeur *p-value* qui nous a été renvoyée par la fonction. Cette valeur nous permet de rejeter ou non l'hypothèse nulle “*Les variables sont indépendantes*” avec plus ou moins de confiance. Ici notre p-value est strictement supérieure à 0.1, on ne peut donc pas rejeter l'hypothèse nulle. Cependant il y a environ 80% de chance que l'apparente dépendance que l'on a observé n'est pas due au hasard.

Pour aller plus loin, on cherche à calculer le *Cramer's V* (le tableau de contingence n'étant pas carré) en appliquant la formule **FORMULE DE CRAMER**.

$$V = \sqrt{8.4933 / (193 * 2)} = 0.1483353$$

On voit que V est proche de 0. Cela corrobore le fait que les deux variables ne sont probablement pas indépendantes (il y a **15% de corrélation entre *Outlook* et *Temperature***).

4. En vous inspirant des questions précédentes, établissez s'il existe un lien de dépendance entre les autres variables présentes dans les données (*outlook/humidity*,*temperature/humidity*).

Maintenant que l'on a déterminer le degré de dépendance entre *Outlook* et *Temperature*, on souhaite élargir notre étude en réalisant les mêmes opérations entre *Outlook* et *Humidity* ainsi que *Humidity* et *Temperature*.

Outlook / Humidity

On récupère la table de contingence de ces variables :

```
contingOH=table(weather$Outlook, weather$Humidity)
```

	Average	High	Low
Foggy	3	6	0
Overcast	34	30	5
Rainy	10	24	0
Sunny	35	9	37

Le nombre de degrés de liberté entre ces variables est de 6. La répartition des valeurs n'est pas proportionnelle à la taille des catégories, ce qui laisse envisager une possible dépendance entre les variables.

On réalise maintenant le test du Chi carré :

```
chisq.test(contingOH)

Pearson's Chi-squared test

data: contingOH
X-squared = 68.49, df = 6, p-value = 8.34e-13
```

On obtient bien un nombre de degrés de liberté égal à 6. La p-value est strictement inférieure à 0.01. Il y a donc plus de 99% de chance que la dépendance observée ne soit pas due au hasard.

La table de contingence n'étant pas carrée, on calcule le *V de Cramer* :

$$V = \sqrt{68.49/(193*2)} = 0.4212306$$

Il y a **42%** de corrélation entre *Outlook* et *Humidity*.

Temperature/ Humidity

On récupère la table de contingence de ces variables :

```
contingTH=table(weather$Temperature, weather$Humidity)
```

	Average	High	Low
Cold	20	22	4
Hot	24	15	19
Mild	38	32	19

Le nombre de degrés de liberté entre ces variables est de 4. La répartition des valeurs n'est pas proportionnelle à la taille des catégories, ce qui laisse envisager une possible dépendance entre les variables. Cette dépendance semble cependant bien moins importante qu'entre *Outlook* et *Humidity*.

On réalise maintenant le test du Chi carré :

```
chisq.test(contingTH)

Pearson's chi-squared test

data: contingTH
X-squared = 10.331, df = 4, p-value = 0.03521
```

On obtient bien un nombre de degrés de liberté égal à 4. La p-value est strictement inférieure à 0.05. Il y a donc plus de 95% de chance que la dépendance observée ne soit pas due au hasard.

La table de contingence étant carrée, on calcule le *coefficient de Chuprov* selon la formule **FORMULE** :

$$Cc = \sqrt{10.331 / (193 * \sqrt{2 * 2})} = 0.1635978$$

Il y a **16%** de corrélation entre *Temperature* et *Humidity*.

Finalement nous pouvons voir qu'une dépendance semble exister pour les couples *Outlook/Humidity* et *Humidity/Temperature*, la corrélation entre *Outlook* et *Humidity* étant la plus visible. Une certaine corrélation pourrait exister entre *Outlook* et *Temperature* mais il n'est pas acceptable de valider cette hypothèse selon nos données.