Université Claude Bernard - Lyon 1

M1 - IGAPA — Année 2007-2008
Statistique — Session 1

Durée : 2 heures

A.B. Dufour & J. Bastien

LES ETUDIANTS DOIVENT CHOISIR **QUATRE** EXERCICES PARMI LES HUIT PROPOSES. SUIVANT LE GROUPE DE TD, IL EST CONSEILLE DE PRENDRE LES EXERCICES DONNES PAR L'ENSEIGNANT.

NOTER LE NOM DE L'ENSEIGNANT EN TETE DE COPIE

Tous les documents sont autorisés.

Exercice 1 - J Bastien

Donner la ou les instruction(s) sous R qui permet(tent) d'obtenir les formules ou les graphiques ci-dessous. On s'appuie sur une variable `note` contenant p notes n_1, n_2, ..., n_p.

1) **les formules de base**

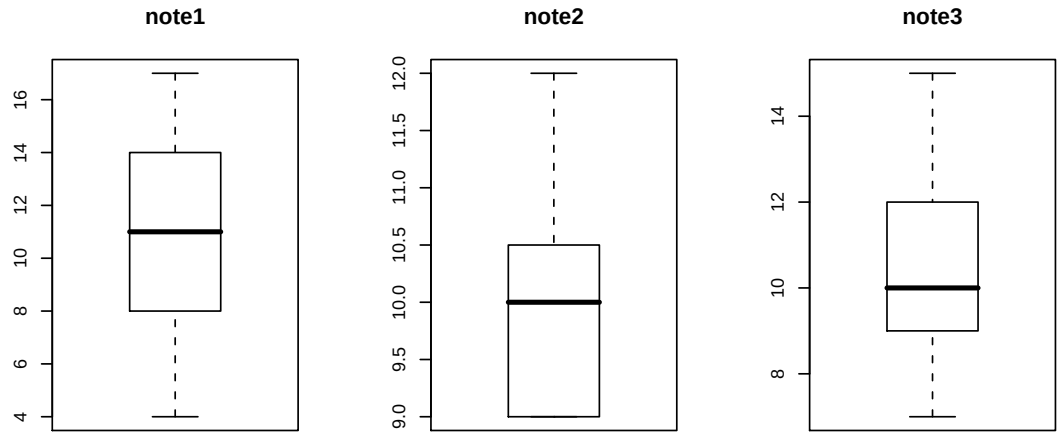
a) la moyenne : $\bar{n} = \frac{1}{p}\sum_{i=1}^{p} n_i$

b) l'écart type : $\sigma = \sqrt{\frac{1}{p}\sum_{i=1}^{p}(n_i - \bar{n})^2}$

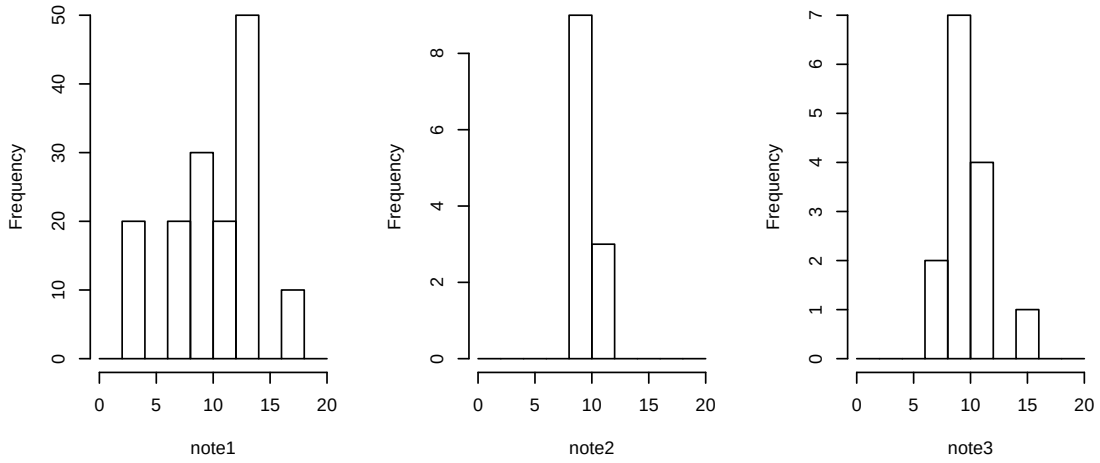
c) l'écart-type estimé appelé parfois déviation standard : $s = \sqrt{\frac{1}{p-1}\sum_{i=1}^{p}(n_i - \bar{n})^2}$

2) **les graphiques ci-dessous**. Pour ce faire, on considère trois groupes ayant respectivement p individus (`note1`), q individus (`note2`) et r individus (`note3`).

a) Graphique 1



b) Graphique 2



Exercice 2 - J Bastien

Dans cet exercice, on étudie la taille et le sexe d'un groupe d'étudiants en L3 APA (fichier disponible sur internet : `http ://pbil.univ-lyon1.fr/R/donnees/L3APA06.txt`). Le fichier est enregistré sous R dans le data frame `L3APA06`.

1) L'instruction `head(L3APA06)` donne le résultat suivant.

```
  groupe identifiant     sexe poids taille rythmcard        age baccalaureat mention
1      A           1 masculin    70    170        60 27/01/1985           S       P
2      A           2 masculin    80    185        65 13/01/1984           S       P
3      A           3  féminin    54    159        62 18/04/1984           S       P
4      A           4  féminin    60    170        58 20/05/1985          ES       P
5      A           5  féminin    60    170        56 30/04/1985           S      AB
6      A           6  féminin    59    165        56 23/02/1985           L       P
  hmental hmoteur hsensoriel pblesocial pratique       sport      niveau mecriture
1       0       0          0          1      non gymnastique      loisir    droite
2       1       1          1          0      oui    handball compétition    droite
```

```
3       0       1          0          1      non        <NA>        <NA>    droite
4       0       1          0          0      oui      tennis compétition    gauche
5       1       0          0          0      oui       danse compétition    droite
6       1       1          0          0      oui gymnastique compétition    gauche
  mfourchette pballon   oeil rotation pappui
1      gauche   droit  droit   gauche  droit
2      gauche   droit  droit   gauche gauche
3      droite   droit  droit   gauche  droit
4      gauche   droit gauche    droit  droit
5      droite   droit gauche    droit  droit
6      gauche   droit  droit   gauche  droit
```

Donner la nature des variables `sexe` et `taille`.

Pour transformer la variable `taille` en variable qualitative, les valeurs sont regroupées en classe de 5 cm en 5 cm. On obtient la distribution suivante :

taille	[155, 160[	[160, 165[	[165, 170[	[170, 175[	[175, 180[	[180, 185[	[185, 190[	[190, 195[
effectif	4	11	9	14	8	9	1	2

2) Comment a-t-on obtenu la table de contingence ci-dessous ? Commentez-la.

taille	féminin	masculin
[155, 160[	4	0
[160, 165[	11	0
[165, 170[	9	0
[170, 175[	5	9
[175, 180[	2	6
[180, 185[	0	9
[185, 190[	0	1
[190, 195[	0	2

3) A la vue de ce tableau, la taille vous parait-elle `a priori` dépendante du sexe ?

4) On utilise maintenant les fonctions de R pour confirmer cela. On construit la table de contingence sous l'hypothèse d'indépendance des deux variables `taille` et `sexe`. Quelle fonction de R a permis de créer ce tableau ?

taille	féminin	masculin
[155, 160[	2.14	1.86
[160, 165[	5.89	5.11
[165, 170[	4.81	4.19
[170, 175[	7.48	6.52
[175, 180[	4.28	3.72
[180, 185[	4.81	4.19
[185, 190[	0.53	0.47
[190, 195[	1.07	0.93

5) Comparer les totaux par ligne et par colonne de ces deux tables de contingence. Commentez.

6) Comment peut-on calculer le coefficient V de Cramer pour mesurer la relation entre les deux variables ?

7) Grâce à R, on trouve $V = 0.82$. Concluez et comparez avec votre intuition écrite à la question 3.

8) Cette méthode vous parait-elle fiable ? Critiquez-la !

Exercice 3 - J Bastien

Dans cet exercice, on étudie la taille (en cm) et le poids (en kg) de sportifs de haut niveau, selon le sport pratiqué (fichier disponible sur internet : `http ://pbil.univ-lyon1.fr/R/donnees/morphosport.txt`). Le nombre de pratiquants par sport est donné dans le tableau ci-dessous.

sport	effectif
athlétisme	20
basketball	2
football	33
handball	44
judo	13
natation	24
volleyball	19

1) Partie Générale

a) Sous R, que permet de faire la fonction ?

```
donnees <- read.table("morphosport.txt", header = TRUE)
head(donnees)
```

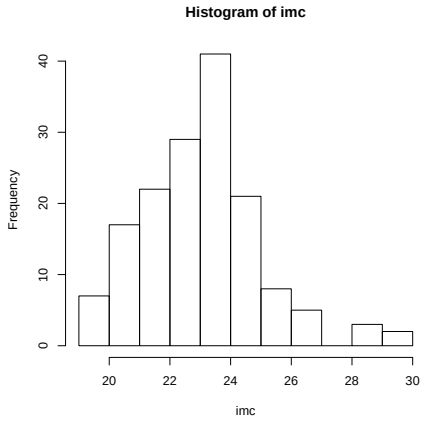
```
  sport  dbi tde  tas  lms poids
1  athl 33.1 174 77.8 78.0    70
2  athl 31.5 179 85.6 81.8    80
3  athl 35.5 179 80.0 80.0    63
4  athl 34.5 170 79.3 79.4    70
5  athl 37.3 182 79.5 86.0    67
6  athl 32.7 170 74.4 69.0    58
```

b) Donnez la nature de la variable `sport`.

c) Quelle instruction permet de tracer un camembert des sports pratiqués ?

d) Complétez l'instruction permettant d'obtenir l'histogramme ci-dessous de l'indice de masse corporelle.

```
imc <- poids/taille^2
_____(imc)
```

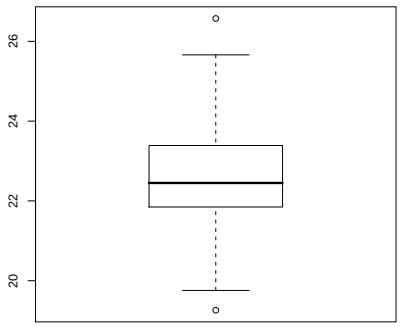


2) Partie spécifique sur les footballeurs

La variable `IMCfoot` contient les IMC de l'ensemble des joueurs de football.

a) Comment peut-on calculer sous R la moyenne et l'écart-type de ces IMC ?

b) Comment s'appelle la figure ci-dessous et comment l'obtient-on ?



c) Que permet-elle de savoir sur l'ensemble des IMC des joueurs de football ?

Exercice 4 - J Bastien

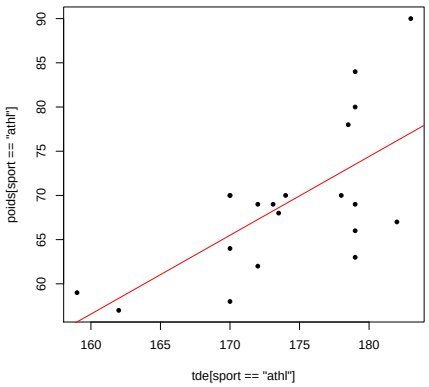
On étudie, dans la suite de l'exercice précédent, la relation entre la taille (`tde`) et le poids de l'ensemble des athlètes.

1) Décrivez chacune des instructions ci-dessous et donnez une interprétation aux informations associées.

```
cor(tde,poids)
plot(tde,poids, pch=20)
abline(lm(poids~tde), col="red")
```

```
[1] 0.6608
```

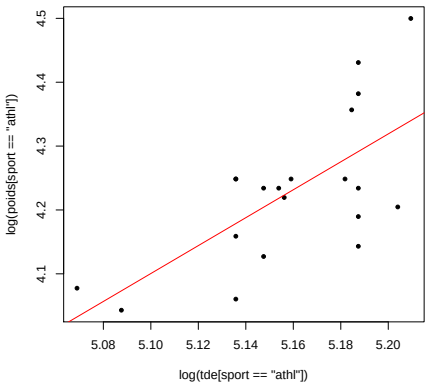
```
(Intercept) tde[sport == "athl"]
   -86.0095                0.8912
```



2) Dans le cadre des études morphologiques, il est plus judicieux de travailler sur le logarithme des variables (notion dite d'allométrie). On transforme donc la taille et le poids en leur logarithme et on calcule leur corrélation ; on représente le nuage de points et la droite de régression.

```
[1] 0.6759
```

```
(Intercept) log(tde[sport == "athl"])
     -7.048                     2.186
```



Donnez une interprétation des résultats ainsi obtenus.

3) Que pouvez-vous dire sur les résultats obtenus sans et avec transformation logarithme ?
4) Dans quelle condition peut-on écrire que $log(poids) = p \times log(taille) + C$ où p représente la pente de la droite et C désigne l'ordonnée à l'origine ?
5) **Facultatif**.
   a) En utilisant le fait que $log(x^p) \approx p \times log(x)$, montrer que la relation précédente peut s'écrire aussi : $poids \approx D \times (taille)^p$ où D est une constante.
   b) Montrer que si la relation précédente est vraie et que si $p \approx 3$, alors $IMC \approx D \times taille$.

Exercice 5 - AB Dufour

Probabilités liées à une loi binomiale

Soit X une variable binomiale de paramètres $n = 15$ et $p = 0.65$.

```
dbinom(0:15, 15, 0.65)
```

```
 [1] 1.449e-07 4.036e-06 5.247e-05 4.222e-04 2.353e-03 9.612e-03 2.975e-02 7.104e-02
 [9] 1.319e-01 1.906e-01 2.123e-01 1.792e-01 1.110e-01 4.756e-02 1.262e-02 1.562e-03
```

```
pbinom(0:15, 15, 0.65)
```

```
 [1] 1.449e-07 4.181e-06 5.665e-05 4.789e-04 2.831e-03 1.244e-02 4.219e-02 1.132e-01
 [9] 2.452e-01 4.357e-01 6.481e-01 8.273e-01 9.383e-01 9.858e-01 9.984e-01 1.000e+00
```

A l'aide des informations ci-dessus, calculer les probabilités :

1. $P(X = 10)$
2. $P(11 \leq X \leq 12)$
3. $P(X < 6)$
4. $P(X \leq 14)$

Probabilités liées à une loi de Student

Soit X une loi de Student à 4 degrés de liberté dont on donne quelques probabilités ci-dessous.

```
axex <- seq(-3, 3, le = 10)
axex
```

```
 [1] -3.0000 -2.3333 -1.6667 -1.0000 -0.3333  0.3333  1.0000  1.6667  2.3333  3.0000
```

```
pt(axex, 4)
```

```
[1] 0.01997 0.03998 0.08545 0.18695 0.37781 0.62219 0.81305 0.91455 0.96002 0.98003
```

A l'aide des informations ci-dessus, calculer les probabilités :

1. $P(X \leq -3)$
2. $P(X > 3)$
3. $P(2.333 \leq X \leq 3)$
4. On note t une valeur particulière de la distribution de Student (t est un nombre positif). Donner la relation liant $P(X > t)$ et une probabilité (à définir) de $-t$.

Exercice 6 - AB Dufour

Une étude porte sur les analyses en laboratoire du phosphore non organique (`phosmol`, en mmol/litre) chez des sujets âgés de plus de 65 ans, en fonction du sexe. En utilisant l'ensemble des informations ci-dessous, répondre à la question : existe-t-il une différence, en moyenne, de phosphore non organique entre les hommes et les femmes ? (choisir et commenter les informations appropriées)

Notation :
`phosmolf` correspond au phosphore non oganique chez les femmes.
`phosmolm` correspond au phosphore non oganique chez les hommes.

	hommes	femmes
effectif	91	83
moyenne	1.059	1.143
variance estimée	0.03331	0.02538

TAB. 1 – Statistiques de base pour les femmes et les hommes

Tests sous condition de normalité de la variable "phosphore non organique".

```
var.test(phosmolf, phosmolm)
```

```
        F test to compare two variances

data:  phosmolf and phosmolm
F = 0.7621, num df = 82, denom df = 90, p-value = 0.2124
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.4988 1.1693
sample estimates:
ratio of variances
             0.762
```

```
t.test(phosmolf, phosmolm, var.equal = F)
```

```
        Welch Two Sample t-test

data:  phosmolf and phosmolm
t = 3.228, df = 171.7, p-value = 0.001494
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.03251 0.13484
sample estimates:
mean of x mean of y
    1.143     1.059
```

```
t.test(phosmolf, phosmolm, var.equal = T)
```

```
        Two Sample t-test

data:  phosmolf and phosmolm
t = 3.208, df = 172, p-value = 0.001595
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.03219 0.13516
sample estimates:
mean of x mean of y
    1.143     1.059
```

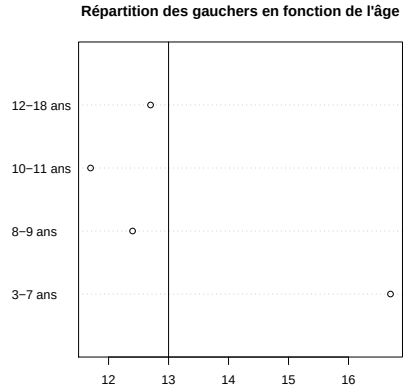
Exercice 7 - AB Dufour

Dans un article de 2006 [1], J. Greenwood *et al* ont cherché à décrire la préférence latérale (droite / gauche) pour les mains, les pieds, les yeux et les oreilles dans une école d'Irlande du Nord et à examiner les différences selon l'âge, le sexe et la tâche réalisée. Deux tâches ont été conservées pour ce problème : écrire un nom, lancer une balle de tennis dans une boîte.

Classe d'âge	3-7 ans	8-9ans	10-11 ans	12-18 ans
Effectif Total	622	582	714	324
Nombre de Gauchers	104	72	84	41

TAB. 2 – Répartitions des Gauchers chez les garçons en fonction de l'âge

1) Sachant qu'en Grande-Bretagne, le pourcentage d'hommes écrivant de la main gauche est de 13%, commenter le graphique ci-dessous.



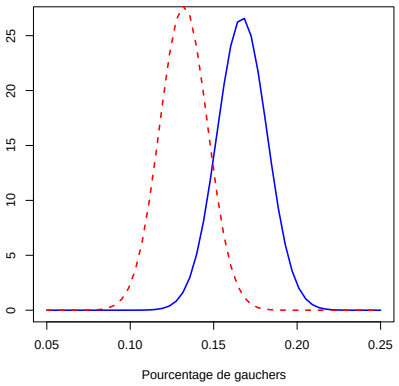
2) La proportion de garçons écrivant de la main gauche, dans la classe d'âge des 3-7 ans, est-elle différente de celle de la population de référence ?
   a) Ecrire la fonction qui permet de répondre à cette question sous R.
   b) On donne les résultats de ce test. Conclure.

```
        1-sample proportions test with continuity correction

data:  104 out of 622, null probability 0.13
X-squared = 7.286, df = 1, p-value = 0.006949
alternative hypothesis: true p is not equal to 0.13
95 percent confidence interval:
 0.1392 0.1994
sample estimates:
     p
0.1672
```

3) La proportion de filles écrivant de la main gauche, pour la même classe d'âge des 3-7 ans, est de 73 sour 553 soit une proportion de 0.132.
   a) Donner l'équation permettant de calculer l'intervalle de confiance de la fréquence de gauchers chez les filles, au niveau de confiance 0.95. Justifier les paramètres.
   b) Commenter le graphique liant les intervalles de confiance des fréquences des garçons (trait plein) et des filles (trait pointillé).

[1] A survey of sidedness on Northern Irish schoolchildren : the interaction of sex, age and risk, *Laterality*, 12 (1), 1-18



c) A la vue de ce graphique, quelle hypothèse pourrait-on émettre ?

d) On donne les résultats de la comparaison des deux proportions de gauchers chez les garçons et chez les filles. Que peut-on en induire pour ces deux populations ?

```
rep0 <- matrix(c(104, 73, 518, 480), nr = 2)
rownames(rep0) <- c("garçons", "filles")
colnames(rep0) <- c("gauchers", "droitiers")
rep0
```

```
        gauchers droitiers
garçons      104       518
filles        73       480
```

```
prop.test(rep0, correct = F)
```

```
        2-sample test for equality of proportions without continuity correction

data:  rep0
X-squared = 2.834, df = 1, p-value = 0.09228
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.005498  0.075888
sample estimates:
prop 1 prop 2
0.1672 0.1320
```

Exercice 8 - AB Dufour

Un test du Chi-Deux dit sur les "outliers" permet de détecter si la valeur d'une distribution peut être considérée comme extrême ou non. Prenons par exemple la série suivante :

```
echantillon <- c(48, 52, 56, 71, 45, 50, 40, 48, 44, 62)
```

Elle contient les résultats à la détente verticale de 10 hommes belges âgés de 18 ans. La question que l'on se pose est sur le sujet ayant réalisé un saut de 71cm. Fait-il partie de la même population ou d'une autre population ? Il peut tout simplement pratiquer un sport alors que les autres non. La question peut se formuler autrement : le sujet sautant 71cm est-il un outlier ?
La valeur de la statistique du test est $\chi^2 = \frac{(valext - mean(echantillon))^2}{varpop}$ où $valext$ représente la valeur à tester, $mean(echantillon)$ la moyenne de l'échantillon étudié et $varpop$ la variance de la population connue par ailleurs.

1) Donner les deux hypothèses H_0 et H_1.

2) Dans notre cas, la valeur de la statistique du test vaut $\chi^2 = 5.925$. Parmi les deux chi-deux ci-dessous, choisir celui qui définit la probabilité associée c'est-à-dire la p-value.

```
pchisq(5.925, 1)
```

```
[1] 0.985
```

```
1 - pchisq(5.925, 1)
```

```
[1] 0.01493
```

3) Donner une conclusion à cette étude.