Bordeaux INP AQUITAINE

Semestre 5	Exercices sur les statistiques descriptives	2017

Objectifs :
- Calculer (sans R) les principaux paramètres statistiques associés à 1 ou 2 variables (cf. Cours).
- Utiliser R pour déterminer ces paramètres.
- Utiliser R pour faire des représentations graphiques (Boxplot, Nuage de points, Histogramme, $\cdots$).

Vous devez être capable de répondre aux différentes questions avec et sans R (via RStudio).

Exercice 1 : Série statistique à 1 variable

On a relevé le taux de matière grasse, exprimé en pourcentage, de fromages.

[1] 20 20 21 21 22 22 23 23 24 29

1. Calculer la moyenne, l'étendue, la SCE et la variance de ces taux.
2. Déterminer la médiane et l'écart interquartile associés à ces taux.
3. Représenter graphiquement ces taux à l'aide d'un boxplot.
4. Pourquoi la dernière valeur est-elle un outlier (valeur aberrante) ?

[Correction de l'exercice 1]

Exercice 2 : Écart-type et étendue

Cet exercice est lié à des notions utilisées lors d'essais interlaboratoires pour estimer la "qualité" de méthodes d'essais.
On s'intéresse à des mesures (analyses) sur un produit en effectuant 2 répétitions dans un laboratoire.
On considère donc un échantillon constitué de 2 valeurs y_1 et y_2.

1. Exprimer la SCE et la variance observée s^2 de la série de mesures $(9; 11)$ ainsi que son étendue $R = |y_1 - y_2|$.
2. Déterminer une relation entre variance observée (empirique) s^2 d'une série de mesures $(y_1; y_2)$ en fonction de son étendue $R = |y_1 - y_2|$.
3. En déduire une relation entre l'écart-type observé et l'étendue R.

[Correction de l'exercice 2]

Exercice 3 :
Cet exercice a pour objectif d'aborder, au travers d'un exemple, les notions essentielles en statistiques à deux variables.
Les calculs peuvent être effectués aisément. Les résultats sont consignés dans le tableau suivant :

Valeurs x_i	1	2	2	3
Valeurs y_i	4	2	0	-2

1. Représenter graphiquement cette série de données.
   Un ajustement affine vous paraît-il judicieux ?
2. Déterminer les coefficients de corrélation linéaire et de détermination entre les variables X et Y.
   Interpréter ces valeurs.
3. Déterminer une équation de la droite d'ajustement linéaire de Y en X.
4. Calculer les estimations $\hat{Y}$.
5. Calculer les différents résidus et les représenter en fonctions de X.
6. En déduire la SCE_{res} puis la SCE_{exp}.
7. Retrouver, en utilisant un quotient de SCE, la valeur du coefficient de détermination.
8. Pour une valeur de référence de 4, donner un ordre de grandeur de la mesure avec cet appareil.

[Correction de l'exercice 3]

Correction des exercices

Exercice 1 :

Question 1 :

La série contient 10 valeurs donc la médiane est la moyenne entre la cinquième (22) et la sixième (22) (rangées par ordre croissant).

On a donc $Me = 22$.

Pour le premier quartile, on considère la médiane de $(20, 20, 21, 21, 22)$.

On a donc $Q_1 = 21$. Retrouvons ces résultats (et les autres) avec R :

```
> # Moyenne
> moyenne = mean(masses)
> moyenne
```

[1] 22.5

```
> # Etendue
> R=max(masses)-min(masses)
> R
```

[1] 9

```
> SCE = sum((masses-moyenne)^2)
> SCE
```

[1] 62.5

```
> variance = SCE/10
> variance
```

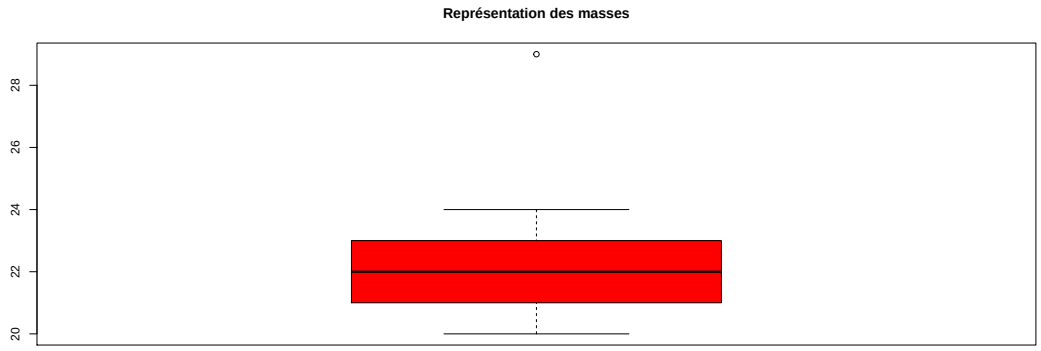
[1] 6.25

Question 2 :

```
> Q1 = quantile(x=masses, probs=0.25)
> Q1

25%
 21

> Q3 = quantile(x=masses, probs=0.75)
> Q3

75%
 23

> ecart_interquartile = Q3-Q1
> ecart_interquartile

75%
  2
```

Question 3 :

```
> boxplot(masses, col="red", main="Représentation des masses")
```



Question 4 :
On a trouvé un écart interquartile égal à $Q_3 - Q_1 = 2$ et $Q_3 = 23$.
Toute valeur supérieure à $Q_3 + 1,5 \times (Q_3 - Q1) = 23 + 1.5 * 2 = 26$ est donc considérée aberrante.

[Retour aux énoncés]

Exercice 2 :

1. Exprimons la SCE et la variance observée s^2 de la série de mesures $(9; 11)$ ainsi que son étendue $R = |y_1 - y_2|$.

```
> y=c(9,11)
> moy=mean(y)
> SCE=sum((y-moy)^2)
> SCE
```

```
[1] 2
```

```
> variance=SCE/2
> variance
```

```
[1] 1
```

```
> R=max(y)-min(y)
> R
```

```
[1] 2
```

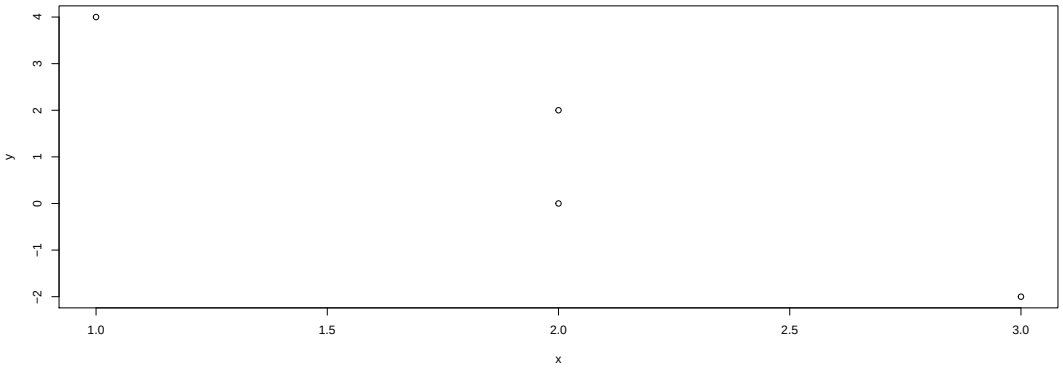
2. Déterminons une relation entre variance observée (empirique) s^2 d'une série de mesures $(y_1; y_2)$ en fonction de son étendue $R = |y_1 - y_2|$. Considérons, par exemple, que y_2 est la plus grande des deux valeurs. on a
$$y_2 = \bar{y} + \frac{R}{2} \quad \text{et} \quad y_1 = \bar{y} - \frac{R}{2}$$
Ainsi : $SCE = (y_1 - \bar{y})^2 + (y_2 - \bar{y})^2 = \left(-\frac{R}{2}\right)^2 + \left(\frac{R}{2}\right)^2 = \frac{R^2}{2}$.
La variance est alors
$$\sigma^2 = \frac{SCE}{2} = \frac{R^2}{4}$$

3. On en déduit que
$$\sigma = \frac{R}{2}$$

[Retour aux énoncés]

Exercice 3 :

1.
```
> x=c(1,2, 2, 3)
> y=c(4, 2, 0, -2)
> plot(y~x)
```



On observe une tendance linéaire (points sensiblement alignés) donc on peut envisager un ajustement affine.

2. Il est nécessaire de calculer la covariance de X et Y ainsi que les écarts-types (ou variances) associés à X et Y.
On rappelle que
$$\rho = \frac{\mathrm{Cov}(X,Y)}{\sigma_X\,\sigma_Y}$$

```
> # Calculs
> covariance = 1/4*sum( (x-mean(x))* (y-mean(y)) )
> var_X = 1/4*sum( (x-mean(x))^2  )
> var_Y = 1/4*sum( (y-mean(y))^2  )
> covariance
```

```
[1] -1.5
```

```
> var_X
```

```
[1] 0.5
```

```
> var_Y
```

```
[1] 5
```

```
> coeff_corr = covariance/sqrt(var_X*var_Y)
> coeff_corr
```

```
[1] -0.9486833
```

```
> coeff_det = coeff_corr^2
> coeff_det
```

```
[1] 0.9
```

Ces résultats peuvent être obtenus directement, avec R, en utilisant :

```
> R=cor(x,y)
> R

[1] -0.9486833

> R^2

[1] 0.9
```

3. On rappelle qu'une équation de la droite d'ajustement d'équation $Y = aX + b$ est obtenue en utilisant

$$a = \frac{\mathrm{Cov}(X,Y)}{\mathrm{Var}(X)} \quad \text{et} \quad b = \bar{y} - a\bar{x}$$

```
> a=covariance/var_X
> a

[1] -3

> b=mean(y)-a*mean(x)
> b

[1] 7
```

On obtient donc

$$Y = -3X + 7$$

On peut obtenir ces résultats (notamment) directement en utilisant :

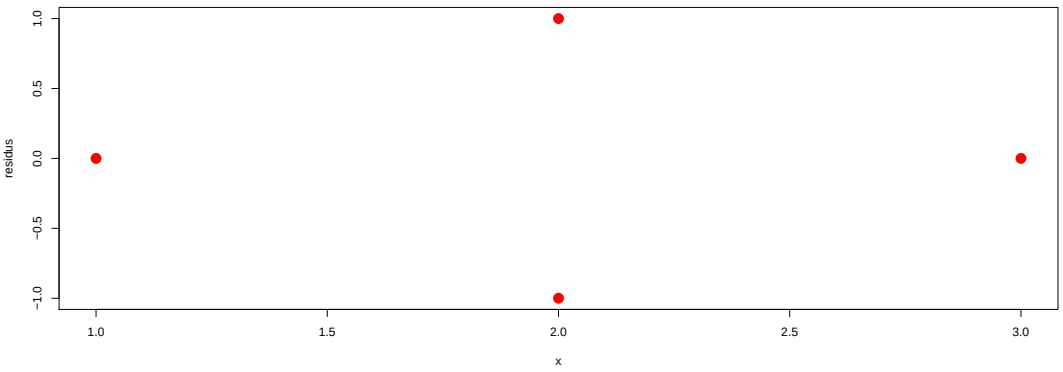
```
> modele=lm(y~x)
> summary(modele)

Call:
lm(formula = y ~ x)

Residuals:
         1          2          3          4
-5.551e-17  1.000e+00 -1.000e+00 -1.295e-16

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   7.0000     1.5000   4.667   0.0430 *
x            -3.0000     0.7071  -4.243   0.0513 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1 on 2 degrees of freedom
Multiple R-squared:    0.9,        Adjusted R-squared:   0.85
F-statistic:    18 on 1 and 2 DF,  p-value: 0.05132
```

4. Les estimations sont associées à

$$\hat{Y} = a * X + b$$

```
> estimations = a*x+b
> estimations

[1]  4  1  1 -2
```

5.
```
> residus = y-estimations
> residus

[1]  0  1 -1  0

> residus_par_R = resid(modele)
> residus_par_R

             1              2              3              4
-5.551115e-17  1.000000e+00 -1.000000e+00 -1.295260e-16

> plot(residus ~ x, col="red", pch=16, cex=2)
```



6.
```
> SCE_res = sum(residus^2)
> SCE_res

[1] 2

> SCE_exp = sum ( (estimations - mean(estimations))^2)
> SCE_exp

[1] 18
```

7.
```
> # SCE totale
> SCE_totale = SCE_res + SCE_exp
> SCE_totale

[1] 20

> # Coefficient de détermination
> SCE_exp / SCE_totale

[1] 0.9
```

8. Déterminos une estimation pour $X = 4$ en utilisant $Y = -3X + 7$.

```
> estimation = a*4+b
> estimation
```

```
[1] -5
```

```
> # Directement avec R
> predict(modele,newdata=data.frame(x=c(4)))
```

```
 1
-5
```

[Retour aux énoncés]