Statistiques Descriptives 2

Ibrahim BENKHERFELLAH & Axel COULET

2025-02-21

# **Exercice n°1**

## **1.**

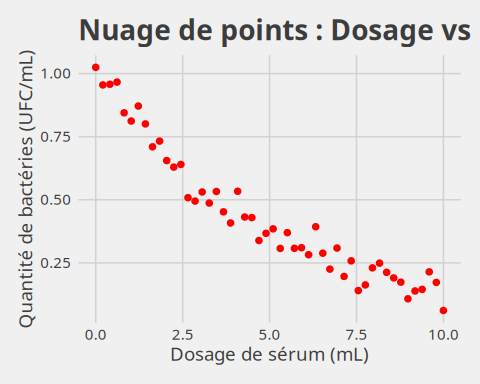
Tout d’abord, qu’est-ce qu’une variable explicative et une variable expliquée ?

* Une variable explicative est la variable qui permet d’expliquer une autre variable. Elle est souvent notée . Sur une représentation graphique, la variable explicative est représentée en abscisse.
* Une variable expliquée est la variable pour laquelle on cherche un modèle de prévision. Son évolution dépend d’une ou de plusieurs autres variables. Elle est souvent notée . Sur une représentation graphique, la variable expliquée est représentée en ordonnée.

Ainsi, dans le cadre de notre exercice, la variable explicative est le **dosage de sérum X (en mL)**. Et la variable expliquée est la **quantité de bactéries Y (en UFC/mL)**, car c’est celle qui est mesurée en fonction du dosage de sérum et qui dépend potentiellement de .

## **2.**

Représentation du nuage de points de la statistique bi-variée () :

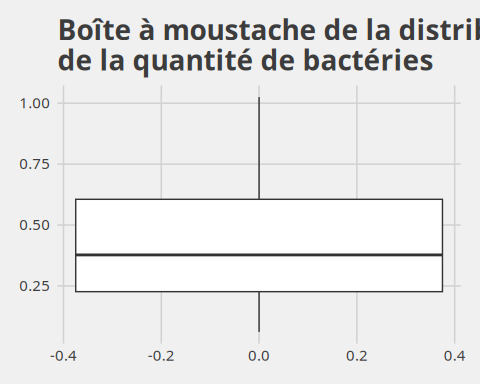


## **3.**

### a) Calcul de la moyenne et de la variance de la quantité de bactéries

* Moyenne :
* Variance :

### b) Box-plot de la dsitribution de la variable : **Quantité de bactéries**



### c) Analyse des résultats obtenus

L’analyse de la quantité de bactéries montre une moyenne de et une variance de , indiquant une dispersion modérée. Le diagramme en boîte révèle une distribution symétrique avec une médiane autour de et un intervalle interquartile de à , suggérant une répartition homogène des données.

## **4.**

Nous savons que la Covariance de et est :

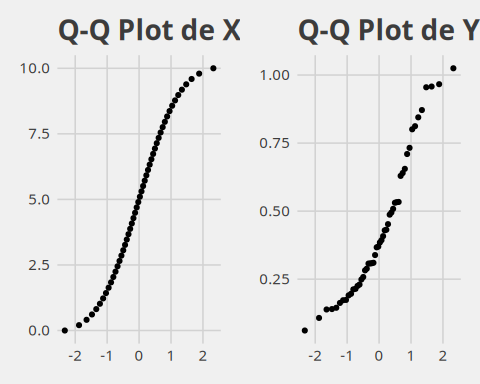
On trouve donc :

La covariance entre les deux variables étant de , cela indique une assez forte relation. Le signe étant négative, cela indique que et évoluent de manière opposé. Autrement dit, lorsque augmente, quant à lui à diminuer, et inversement. Ce qui coincide avec l’observation faite ci-dessus, dans le nuage de point.

## **5.**

Les Q-Q plots (Quantile-Quantile plots) permettent de comparer les quantiles d’une variable observée à ceux d’une distribution théorique, souvent la distribution normale. En examinant l’alignement des points sur une droite diagonale, il est possible d’évaluer visuellement si une variable suit une distribution normale.

### a) Etude graphique Quantile-Quantile



#### Interprétation du Q-Q Plot pour la variable X

* **Alignement des points** :  
  Les points du Q-Q plot de sont quasiment alignés sur la droite diagonale. Cela suggère que les quantiles observés de correspondent étroitement aux quantiles théoriques d’une distribution normale.
* **Extrémités (queues de distribution)** :  
  Les points situés aux extrémités ne s’écartent pas significativement de la droite, ce qui renforce l’idée que présente les caractéristiques d’une distribution normale.

La variable  **semble être normalement distribuée.**

#### Interprétation du Q-Q Plot pour la variable Y

* **Écart par rapport à la diagonale** :  
  Pour la variable , les points du Q-Q plot ne suivent pas la droite diagonale. On observe une courbure, surtout aux extrémités, indiquant que les quantiles observés ne correspondent pas aux quantiles attendus pour une distribution normale.
* **Queues de distribution** :  
  Les écarts importants aux extrémités suggèrent que possède des caractéristiques non conformes à la normalité, telles qu’une asymétrie ou des queues de distribution différentes de celles d’une loi normale.

La variable  **ne semble pas suivre une distribution normale**.

### b) Etude numérique Shapiro-Wilk

Le **test de Shapiro-Wilk** est un test statistique permettant d’évaluer si un échantillon suit une **distribution normale**. Il est particulièrement efficace pour les **petites tailles d’échantillon** (), bien qu’il puisse être utilisé sur des ensembles de données plus grands.

#### Hypothèses du test

* **(Hypothèse nulle) :** Les données suivent une distribution normale.
* **(Hypothèse alternative) :** Les données ne suivent pas une distribution normale.

#### Formule du test

La statistique du test de Shapiro-Wilk est définie comme suit :

**où** :

* sont les valeurs de l’échantillon classées par ordre croissant.
* est la **moyenne** de l’échantillon.
* sont des coefficients calculés à partir des quantiles d’une distribution normale.

#### Interprétation des résultats

* Si la **p-valeur** du test est **supérieure à 0.05**, on **ne rejette pas** : l’échantillon peut être considéré comme **normalement distribué**.
* Si la **p-valeur** est **inférieure à 0.05**, on **rejette** : les données **ne suivent pas** une distribution normale.

#### Résultats du code R

##   
## Shapiro-Wilk normality test  
##   
## data: data$X  
## W = 0.95558, p-value = 0.05809

##   
## Shapiro-Wilk normality test  
##   
## data: data$Y  
## W = 0.92148, p-value = 0.002668

#### Variable X

Le test appliqué à **X** donne une statistique **W = 0.95558** avec une **p-valeur de 0.05809**.

Puisque la **p-valeur est supérieure à 0.05**, nous **ne rejetons pas** l’hypothèse de normalité.

Cela signifie que **la variable X suit une distribution normale** au seuil de 5%.

#### Variable Y

Le test appliqué à **Y** donne une statistique **W = 0.92148** avec une **p-valeur de 0.002668**.

Dans ce cas, la **p-valeur est inférieure à 0.05**, nous **rejetons donc** l’hypothèse de normalité.

Ainsi, nous pouvons conclure que **la variable Y ne suit pas une distribution normale**.

## **6.**

On sait que le coefficient de corrélation linéaire de Pearson se calcul de la façon suivante :

Avec R, on trouve :

##   
## Pearson's product-moment correlation  
##   
## data: data$X and data$Y  
## t = -20.94, df = 48, p-value < 2.2e-16  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## -0.9711081 -0.9120555  
## sample estimates:  
## cor   
## -0.9493844

Le coefficient de corrélation linéaire de Pearson est de . Ce coefficient est très proche de -1, ce qui indique une forte corrélation négative entre les deux variables. Cela signifie que et évoluent de manière opposée (comme supposé avec la covariance). Cette relation est donc confirmée par le coefficient de corrélation élevé, qui suggère une forte dépendance linéaire entre les deux variables.

## **7.**

On sait qu’une droite affine s’exprime de la façon suivante :

Par la méthode de Legendre, aussi appelé méthode des moindres carrés, on peut trouver les coefficients et de la droite de régression linéaire qui minimise la somme des carrés des résidus. Autrement dit, on cherche et tels que :

Cela signifie que nous cherchons les valeurs spécifiques de et qui rendent la somme des carrés des erreurs aussi petite que possible. D’après le cours de monsieur Hébert.

Dans un modèle linéaire, le meilleur coefficient directeur (au sens des moindres carrés) et le meilleur ordonné à l’origine (au sens des moindres carrés) sont :

**et**

### **Démonstration**

**On cherche :** et

D’où le système suivant :

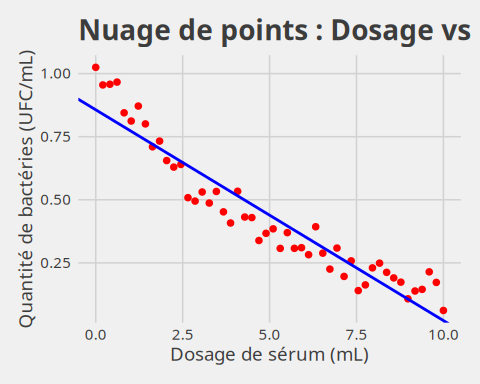
**D’ou** :

et

### Applications

**et**

### Récapitulatif



## **8.**

En observant le nuage de points, on voit clairement que lorsque le dosage de sérum augmente, la quantité de bactéries diminue de manière régulière. La droite de régression, qui semble passer naturellement au milieu de ces données, reflète bien cette tendance. Cette modélisation linéaire nous paraît être une première approximation logique et cohérente, en accord avec l’idée que l’augmentation du dosage se traduit par une diminution de la concentration bactérienne. L’ajustement linéaire offre ici une représentation intuitive et satisfaisante de la relation observée.

## **9.**

##   
## Call:  
## lm(formula = data$Y ~ data$X)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.12653 -0.06608 -0.01112 0.05365 0.16835   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.856490 0.023144 37.01 <2e-16 \*\*\*  
## data$X -0.083514 0.003988 -20.94 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.08306 on 48 degrees of freedom  
## Multiple R-squared: 0.9013, Adjusted R-squared: 0.8993   
## F-statistic: 438.5 on 1 and 48 DF, p-value: < 2.2e-16

### Test de significativité globale du modèle

Avant d’interpréter les résultats, il convient d’évaluer la significativité du modèle. Il désigne le seuil à partir duquel les résultats d’un test sont jugés fiables. Autrement dit, ce seuil détermine la confiance dans la corrélation entre un test effectué et les résultats obtenus.

Soit : -> le modèle n’est pas significatif

Ce test est basé sur la **statistique de Fisher** présentée en bas de la sortie R. La p-valeur associée à cette statistique est de **2e-16**. Puisque cette valeur est **inférieure à 1%**, nous **rejetons l’hypothèse nulle**, à savoir le modèle est bien **globalement significatif**

### Interprétation des coefficients

Les coefficients permettent de comprendre comment la variable explicative influence la variable dépendante .

**(Intercept)** qui est dans notre cas (que nous allons appelé ), est de **0.85649**. Et **data$X** qui est (que nous allons appelé ), est de **-0.08351**. Ces valeurs correspondent à celles que nous avons calculées précédemment.

Les **Pr(>|t|)** sont les **p-valeurs** associées à chaque coefficient. Elles permettent de tester **l’hypothèse nulle** suivante :

* -> Cela signifie que n’a aucun effet sur

Or, ici les **p-valeurs sont très faibles** (**2e-16**), ce qui signifie que nous **rejetons l’hypothèse nulle**. Donc **les coefficients sont significatifs** et que la variable explicative a un effet significatif sur la variable dépendante .

De plus, dans R la signification des coefficients est indiquée avec des étoiles. Ici nous avons **3 étoiles** (**\*\*\***), ce qui signifie que les coefficients sont très significatifs.

### Residuals

Nous avons un **minimum à -0.12653** et un **maximum à 0.16835**. Ces valeurs montrent que les résidus restent dans un **intervalle restreint**. Cela signifie que le modèle fait des **erreurs relativement petites** et que ses **prédictions sont assez précises**.

De plus, **la médiane des résidus est proche de 0**, ce qui indique que le modèle est **équilibré** et que les erreurs sont réparties de manière homogène autour de la droite de régression.

### Multiple R-squared

Le coefficient de détermination est de **0.9013**. Cela signifie que **90.13% de la variance de la variable dépendante est expliquée par la variable explicative** . Ce coefficient est très élevé, ce qui indique que le modèle linéaire est **très performant** pour prédire la quantité de bactéries en fonction du dosage de sérum.

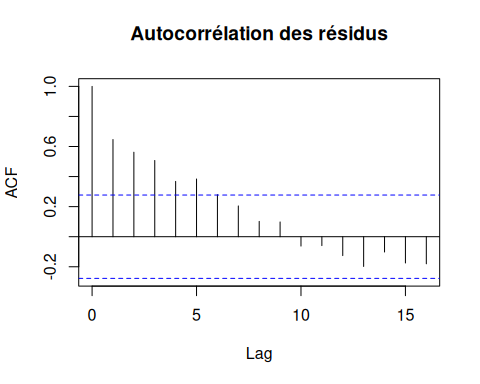
## **10.**

## [1] 0.3311143

## [1] 0.3311143

## **11.**

a) Évaluer l’indépendance des résidus



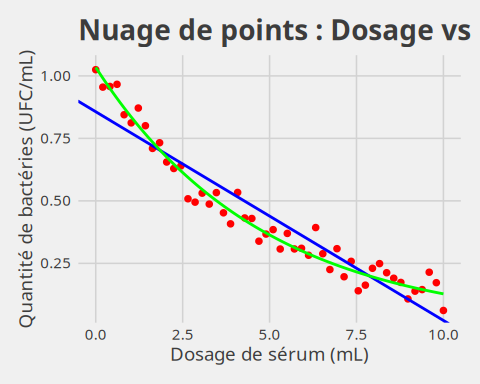
##   
## Durbin-Watson test  
##   
## data: LMS  
## DW = 0.61628, p-value = 1.655e-09  
## alternative hypothesis: true autocorrelation is greater than 0

b) Tester la normalité des résidus

Shapiro-Wilk

Quantile-Quantile (Q-Q plot)

## **12.**



## **13.**

???? ->

????

## **14.**

## **Bonus**

?????

## Valeur peut être utile

## [1] 0.06848566

## [1] 8.850479

## [1] 0.4389205

## [1] 5

## [1] -0.7391376

## [1] 0.3773284

## 0% 25% 50% 75% 100%   
## 0.06141661 0.22651337 0.37732843 0.60555035 1.02483571