SYS865 Inférence statistique avec programmation R

Ornwipa Thamsuwan

13 mars 2024

Plan de la séance

Plan de la séance

- Corrélation
- Régression linéaire

Corrélation

Corrélation

La corrélation de variables aléatoires est une mesure qui quantifie le degré auquel deux variables aléatoires varient ensemble.

- Si les variations des deux variables montrent une tendance à se produire ensemble, on dit qu'elles sont positivement corrélées.
- Si une variable a tendance à augmenter quand l'autre diminue, elles sont négativement corrélées.

Corrélation

La corrélation de variables aléatoires est une mesure qui quantifie le degré auquel deux variables aléatoires varient ensemble.

- Si les variations des deux variables montrent une tendance à se produire ensemble, on dit qu'elles sont positivement corrélées.
- Si une variable a tendance à augmenter quand l'autre diminue, elles sont négativement corrélées.

La corrélation est souvent mesurée par un **coefficient** qui varie entre -1 et 1.

Un coefficient de 1 indique une corrélation positive parfaite, -1 indique une corrélation négative parfaite, et 0 indique l'absence de corrélation.

Corrélation de Pearson (Paramétrique)

Définition: La corrélation de Pearson, également connue sous le nom de coefficient de corrélation produit-moment de Pearson, évalue la **relation linéaire** entre deux variables quantitatives.

Caractéristiques: Valeurs entre -1 et 1.

Utilisation: Préférable lorsque les deux variables sont **normalement distribuées** et la relation est supposée être linéaire.

Corrélation de Pearson (Paramétrique)

Formule: Corrélation de Pearson = (Covariance de X et Y) / (Écart-type de X * Écart-type de Y).

$$r_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

où r_{xy} est le coefficient de corrélation de Pearson entre les variables x et y, x_i et y_i sont les valeurs des variables, et $\bar{x}$ et $\bar{y}$ sont les moyennes de x et y, respectivement.

Corrélation de Spearman (Non-Paramétrique)

Définition: La corrélation de Spearman, ou le coefficient de **rang** de Spearman, est utilisée pour mesurer la force et la direction de l'association entre deux variables classées.

Caractéristiques: Également évaluée entre -1 et 1. Moins sensible aux valeurs aberrantes.

Utilisation: Appropriée lorsque les données ne sont pas normalement distribuées ou lorsqu'on examine des relations non linéaires.

Corrélation de Spearman (Non-Paramétrique)

Formule: Corrélation de Spearman = 1 - (6 * Somme des carrés des différences de rang) / (n(n^2 - 1)).

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

où ρ est le coefficient de corrélation de Spearman, d_i est la différence entre les rangs des i-èmes valeurs de x et y, et n est le nombre de paires de données.

Rappel

Il est crucial de se rappeler que la corrélation ne signifie pas causalité.

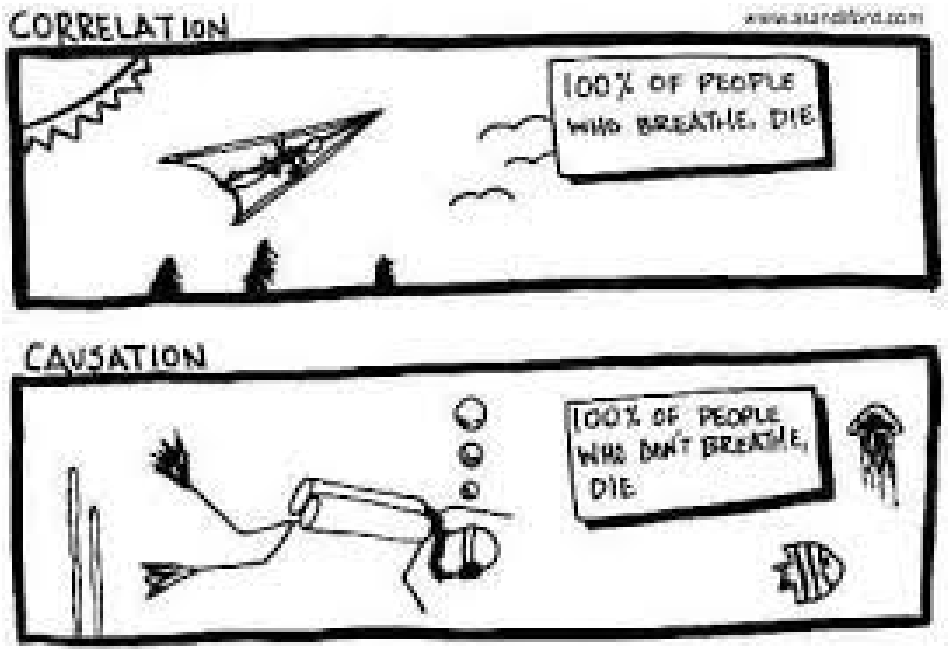


Figure 1: Corrélation vs. causalité

Analyse avec R : Données

Base de données “Pima Indian Diabetes”

Analyse avec R : Données

Base de données “Pima Indian Diabetes”

Test de normalité de Shapiro-Wilk

```
## Glucose : p-value = 1.720326e-11
## BloodPressure : p-value = 9.45138e-05
## SkinThickness : p-value = 1.775691e-09
## Insulin : p-value = 1.698218e-21
## BMI : p-value = 8.557785e-09
## Age : p-value = 2.402274e-24
```

Analyse avec R : Données

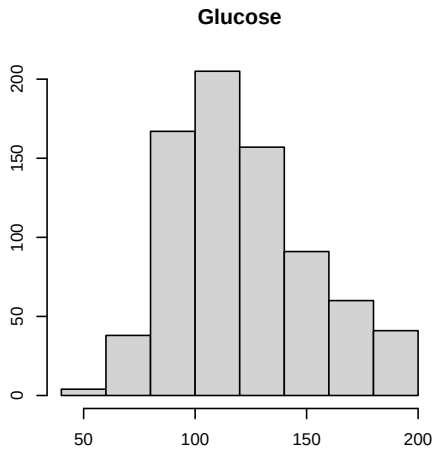
Base de données “Pima Indian Diabetes”

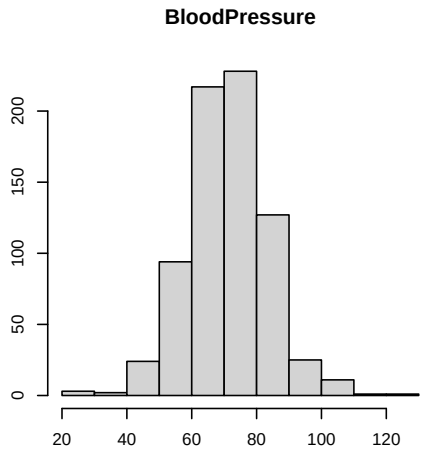
Test de normalité de Shapiro-Wilk

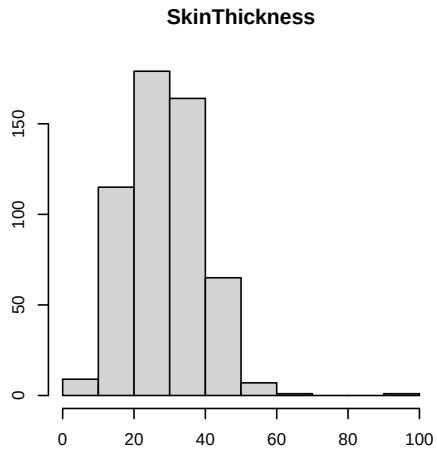
```
## Glucose : p-value = 1.720326e-11
## BloodPressure : p-value = 9.45138e-05
## SkinThickness : p-value = 1.775691e-09
## Insulin : p-value = 1.698218e-21
## BMI : p-value = 8.557785e-09
## Age : p-value = 2.402274e-24
```

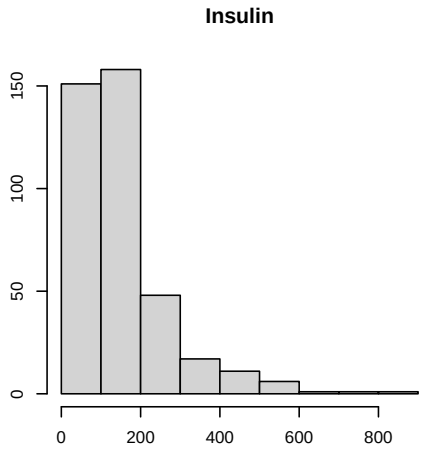
Les données ne sont pas normalement distribuées. Il faut donc utiliser la corrélation de Spearman.

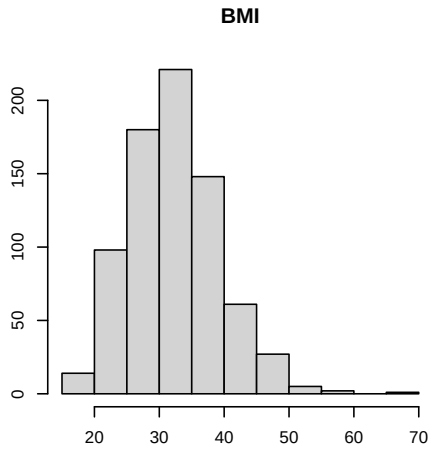
Analyse avec R : Histogrammes

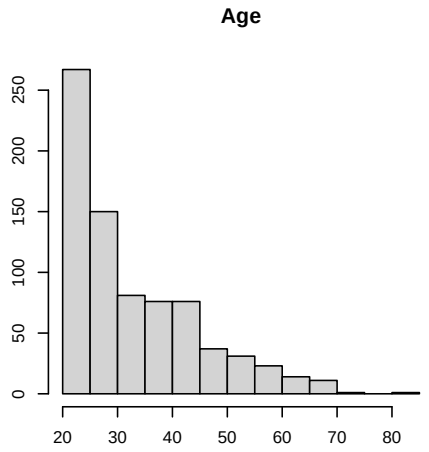




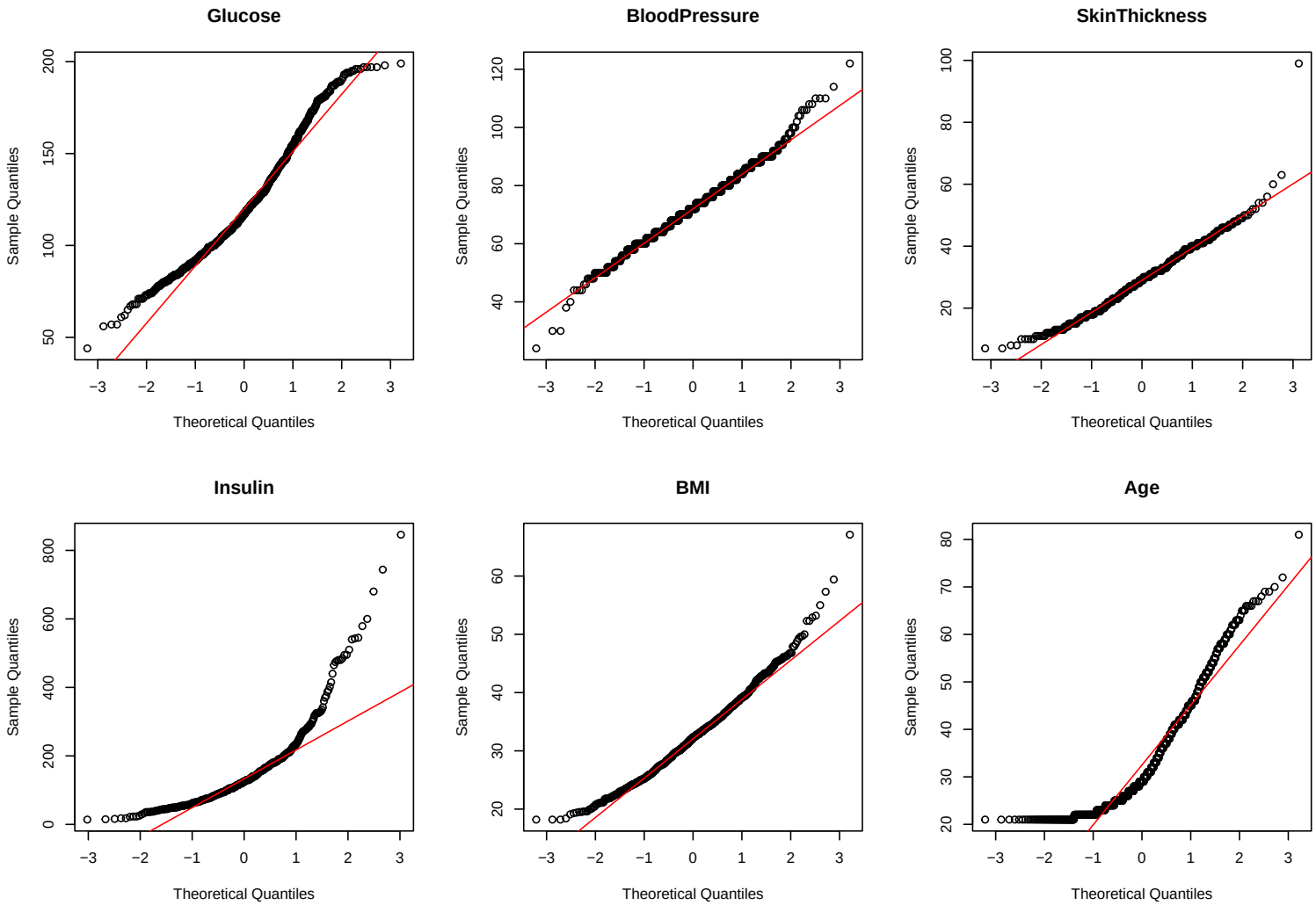








Analyse avec R : Graphiques Q-Q



Analyse avec R : Corrélation de Spearman

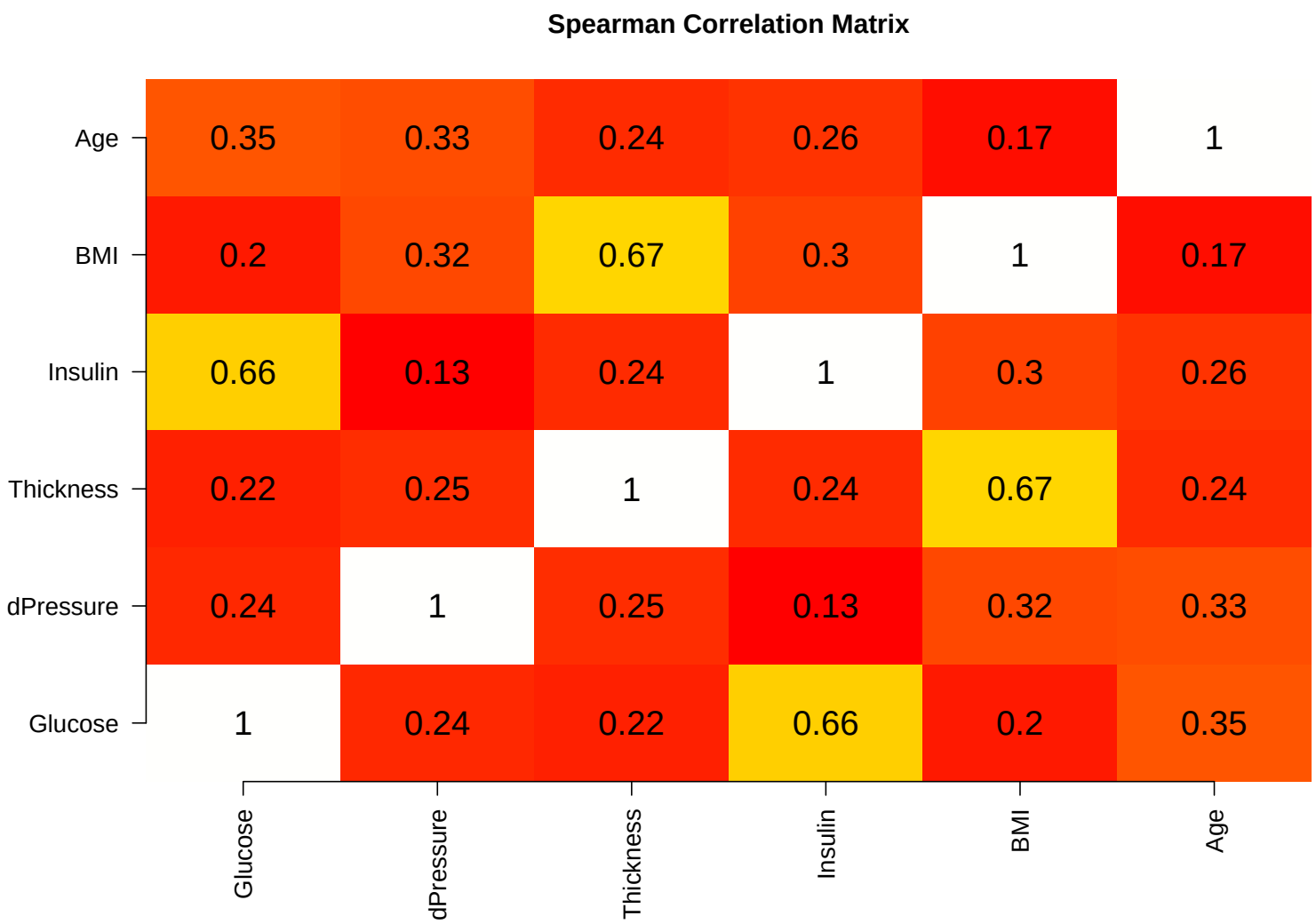
```
spearman_correlation_matrix <-
  cor(diabetes_subset,
      use="complete.obs",
      method="spearman")
```

La fonction `cor(diabetes_subset)` calcule les coefficients de corrélation pour toutes les paires de variables dans la base de données `diabetes_subset`.

L'argument `method="spearman"` spécifie que le coefficient de corrélation de rang de Spearman doit être utilisé.

L'argument `use="complete.obs"` indique à R d'utiliser uniquement des cas complets (c'est-à-dire des lignes sans aucune valeur NA).

Analyse avec R : Corrélation de Spearman



- Assez forte corrélation positive entre `SkinThickness` et `BMI`, et entre `Glucose` et `Insulin`. Toutefois, ...

Rappel

Le fait que deux variables soient fortement corrélées ne démontre pas que l'une est la cause de l'autre.

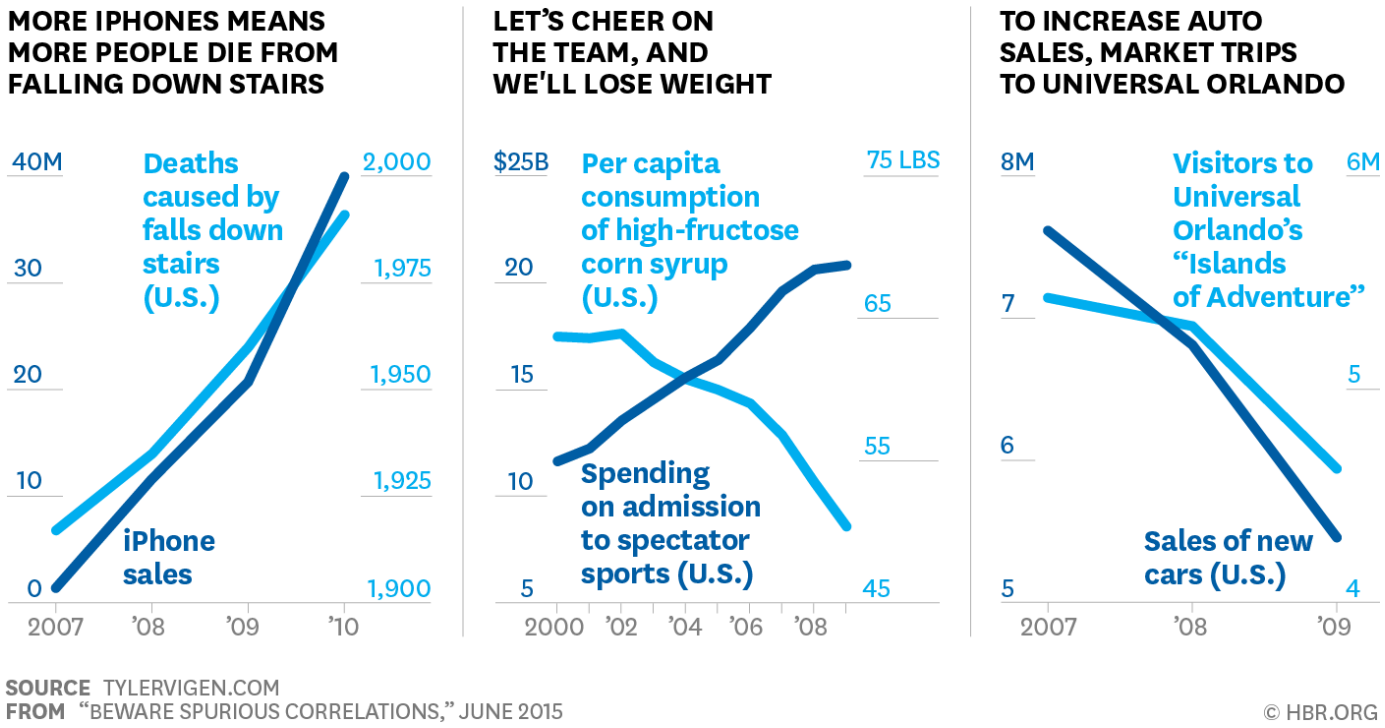
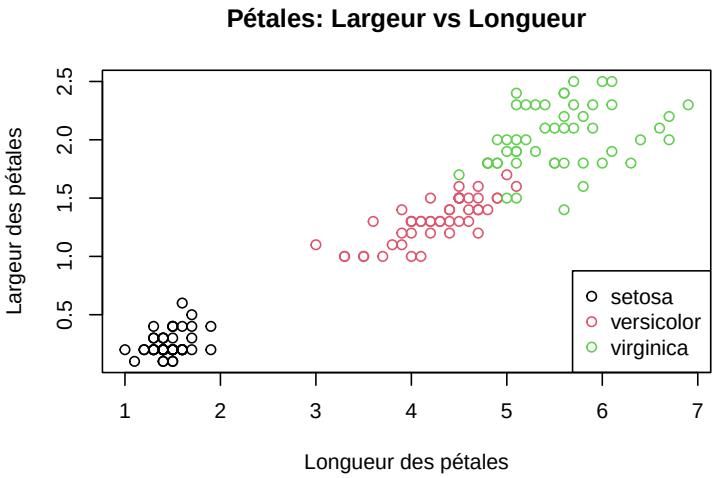
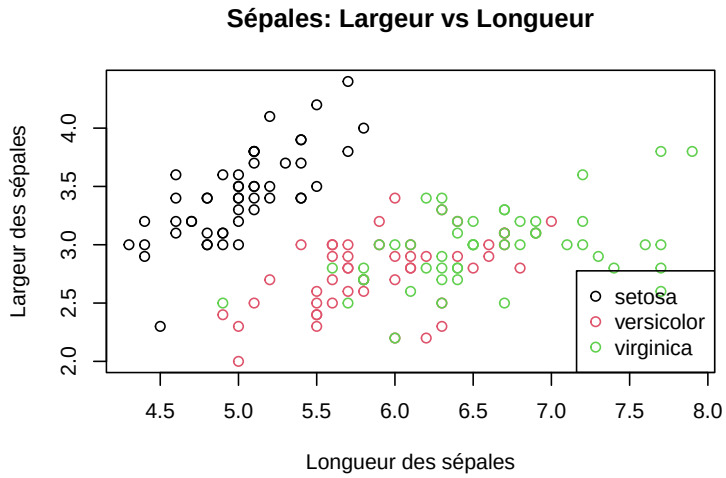


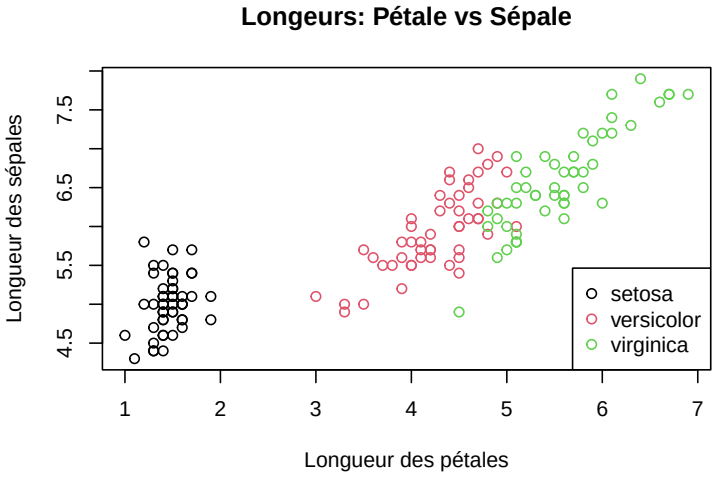
Figure 2: Corrélation fallacieuse

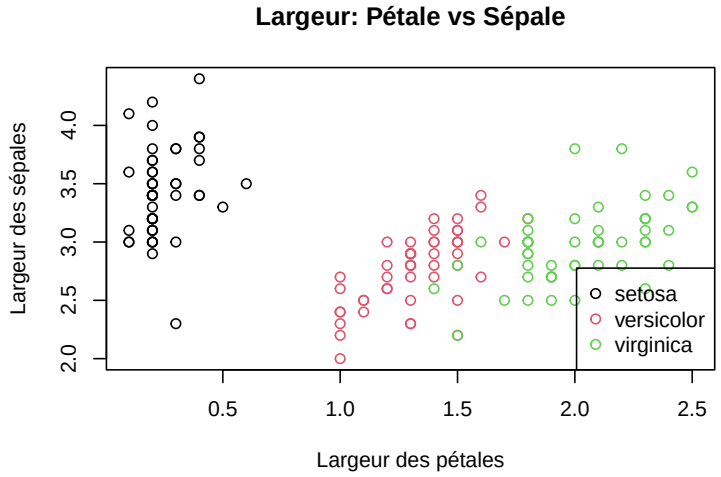
Analyse avec R : Visualisation de données

Iris - nuage de points (“scatter plots” en anglais)









Analyse avec R : Distribution normale

Test de normalité de Shapiro-Wilk

```
## Sepal.Length : p-value = 0.01018
## Sepal.Width : p-value = 0.10115
## Petal.Length : p-value = 0
## Petal.Width : p-value = 0
```

Analyse avec R : Distribution normale

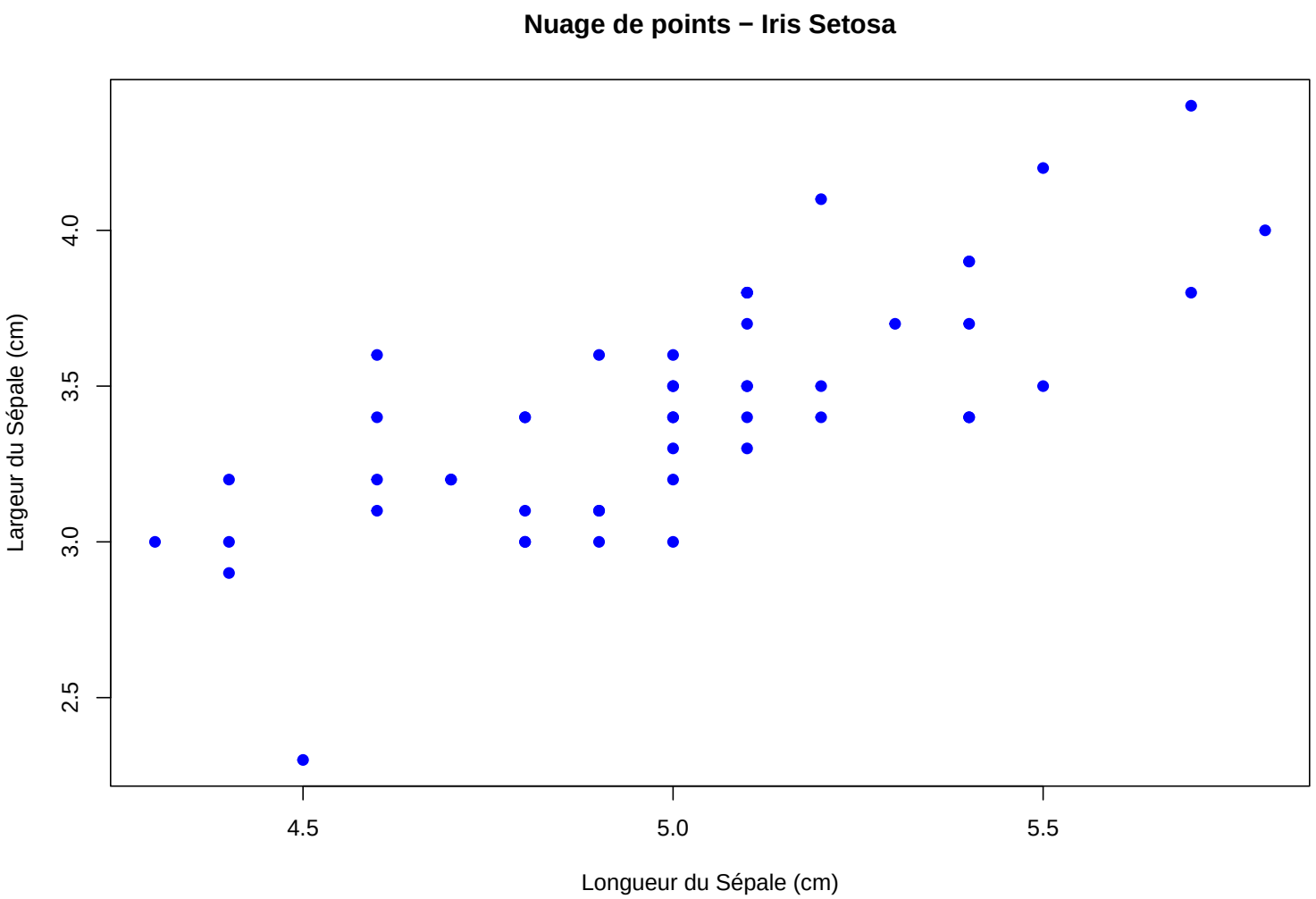
Test de normalité de Shapiro-Wilk

```
## Sepal.Length : p-value = 0.01018
## Sepal.Width : p-value = 0.10115
## Petal.Length : p-value = 0
## Petal.Width : p-value = 0
```

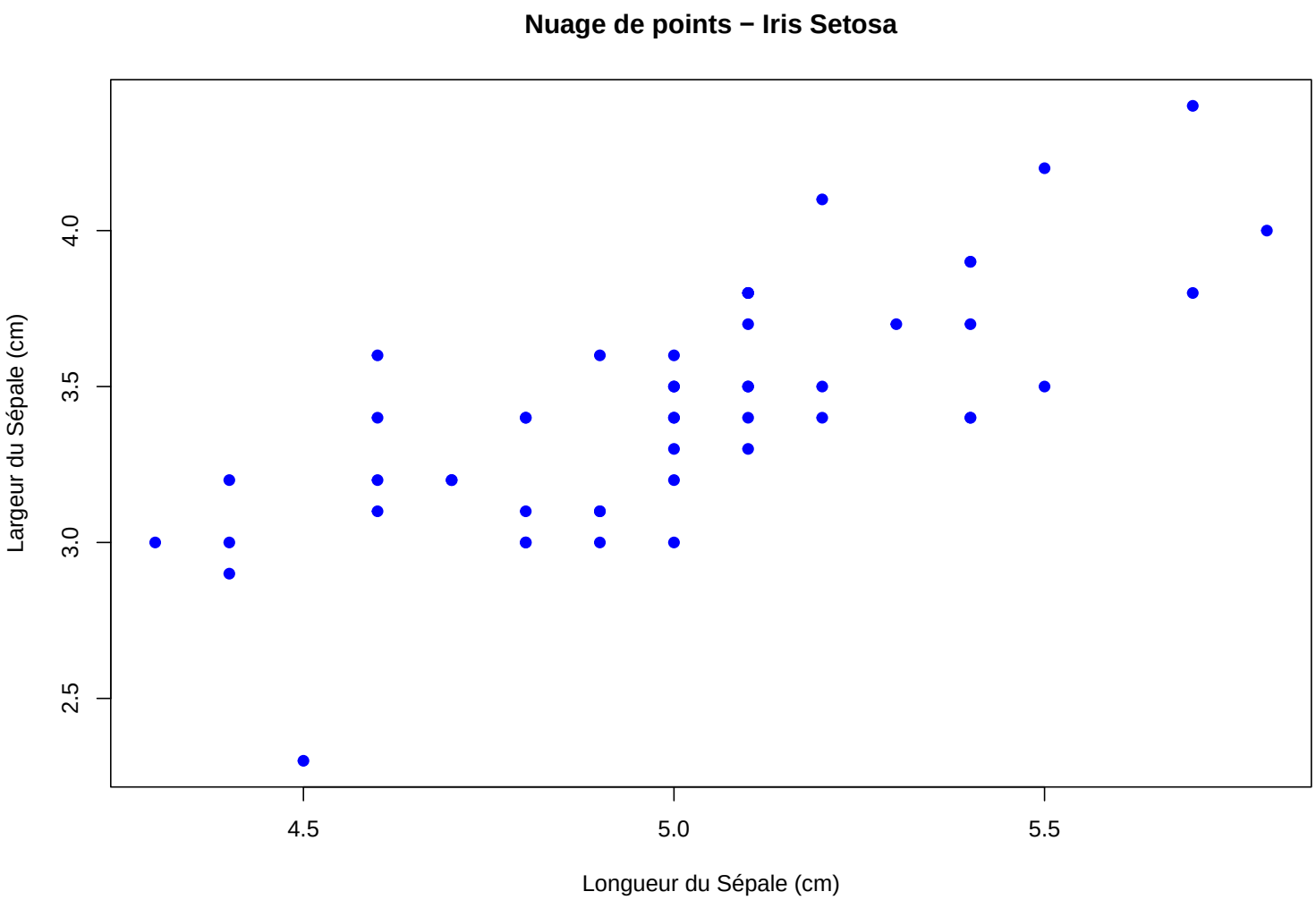
En incluant uniquement l'espèce de `setosa`

```
## Sepal.Length : p-value = 0.45951
## Sepal.Width : p-value = 0.27153
## Petal.Length : p-value = 0.05481
## Petal.Width : p-value = 0
```

Analyse avec R : Rélation linéaire



Analyse avec R : Rélation linéaire



Ainsi, nous démontrerons le calcul de corrélation de Pearson pour la longueur et la largeur des sépales de `setosa`.

Analyse avec R : Corrélation de Pearson

La méthode par défaut pour `cor()` est le coefficient de corrélation de Pearson.

La fonction `cor(x, y)` ou `cor(x, y, method = "pearson")` calcule le coefficient de corrélation de Pearson entre deux variables `x` et `y`.

```
cor(iris_setosa$Sepal.Length, iris_setosa$Sepal.Width)
```

```
## [1] 0.7425467
```

Analyse avec R : Corrélation de Pearson

La méthode par défaut pour `cor()` est le coefficient de corrélation de Pearson.

La fonction `cor(x, y)` ou `cor(x, y, method = "pearson")` calcule le coefficient de corrélation de Pearson entre deux variables `x` et `y`.

```
cor(iris_setosa$Sepal.Length, iris_setosa$Sepal.Width)
```

```
## [1] 0.7425467
```

- Le coefficient d'environ 0,74 suggère qu'à mesure que l'une des variables (longueur ou largeur) augmente, l'autre variable a tendance à augmenter également, et cette relation est relativement forte. Cependant, ...

Rappel

L'existence d'une corrélation entre deux variables n'implique pas une relation de cause à effet.

Figure 3: Corrélation, et non causalité

Régression linéaire simple

Régression linéaire multiple

Confondeurs