Traitement de données avec R

initiation aux méthodes exploratoires

Simon Chabot
chabotsi@unice.fr

Université Côte d'Azur

24 janvier 2017

Plan

Plan

Objectifs

- Définir le ou les groupes étudiés ;
- Définir le codage des observations ;
- Définir la présentation des données ;
- Réduire les données à l'aide de quelques indicateurs statistiques.

Définir le groupe étudié

En théorie une population entière. (toutes les personnes atteintes de la maladie X).

En pratique un échantillon. (100 personnes atteintes de la maladie X).

Taille de l'échantillon

Pour étendre les résultats observés sur l'échantillon à la population totale, la taille de l'échantillon doit être représentatif !

Définir le groupe étudié

En théorie une population entière. (toutes les personnes atteintes de la maladie X).

En pratique un échantillon. (100 personnes atteintes de la maladie X).

Taille de l'échantillon

Pour étendre les résultats observés sur l'échantillon à la population totale, la taille de l'échantillon doit être représentatif !

Définir le groupe étudié

En théorie une population entière. (toutes les personnes atteintes de la maladie X).

En pratique un échantillon. (100 personnes atteintes de la maladie X).

Taille de l'échantillon

Pour étendre les résultats observés sur l'échantillon à la population totale, la taille de l'échantillon doit être représentatif !

Définir le codage des observations

Type de variables

qualitative non-mesurable :

- sexe
- présence ou absence d'un marqueur
- etc

quantitative mesurable :

- taille
- poids
- durée
- etc

Lancé d'un dé à 6 faces

- Je fais n lancés.
- Je compte le nombre d'apparitions de chaque face.

Présentation des résultats

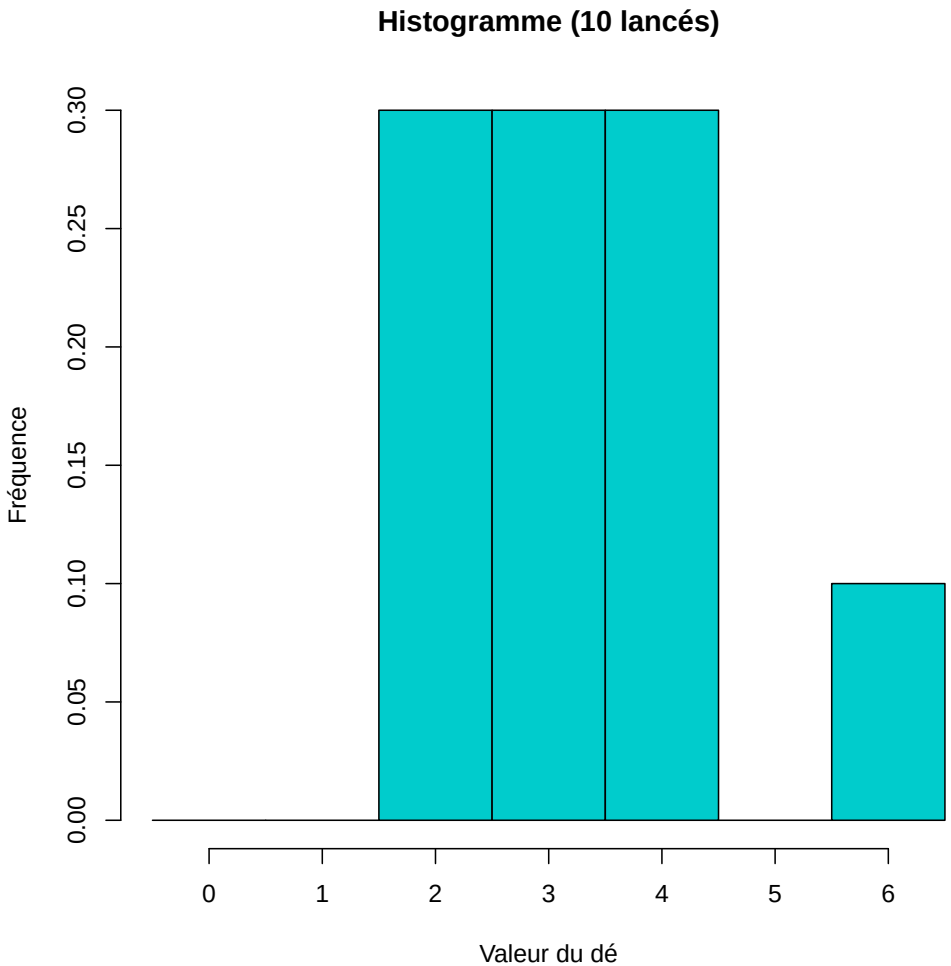
Occurrences	1	2	3	4	5	6	total
Effectifs	n_1	n_2	n_3	n_4	n_5	n_6	n
Fréquences	f_1	f_2	f_3	f_4	f_5	f_6	1

$$n = \sum_{i=1}^{6} n_i, \qquad f_i = \frac{n_i}{n}, \qquad \sum_{i=1}^{6} f_i = 1$$

Présentation des résultats

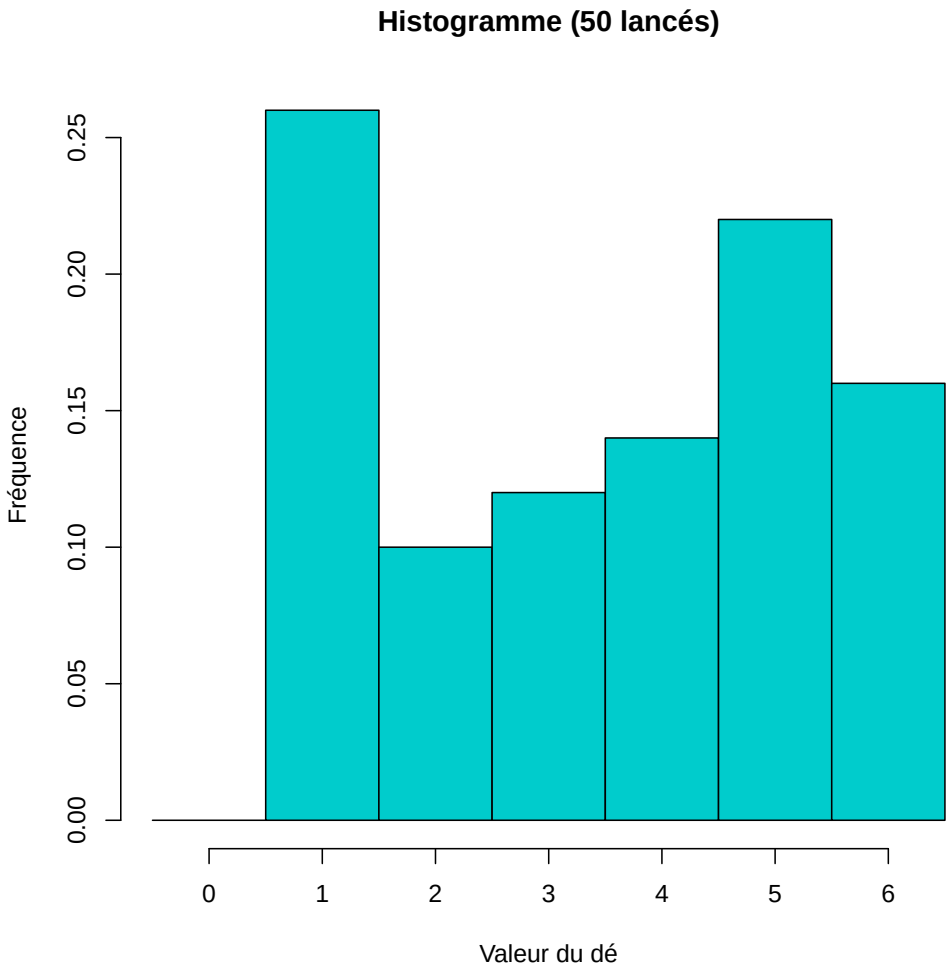
Occurrences	1	2	3	4	5	6	total
Effectifs	0	0	3	3	3	1	10
Fréquences	0	0	0.3	0.3	0.3	0.1	1

$$n = \sum_{i=1}^{6} n_i, \qquad f_i = \frac{n_i}{n}, \qquad \sum_{i=1}^{6} f_i = 1$$



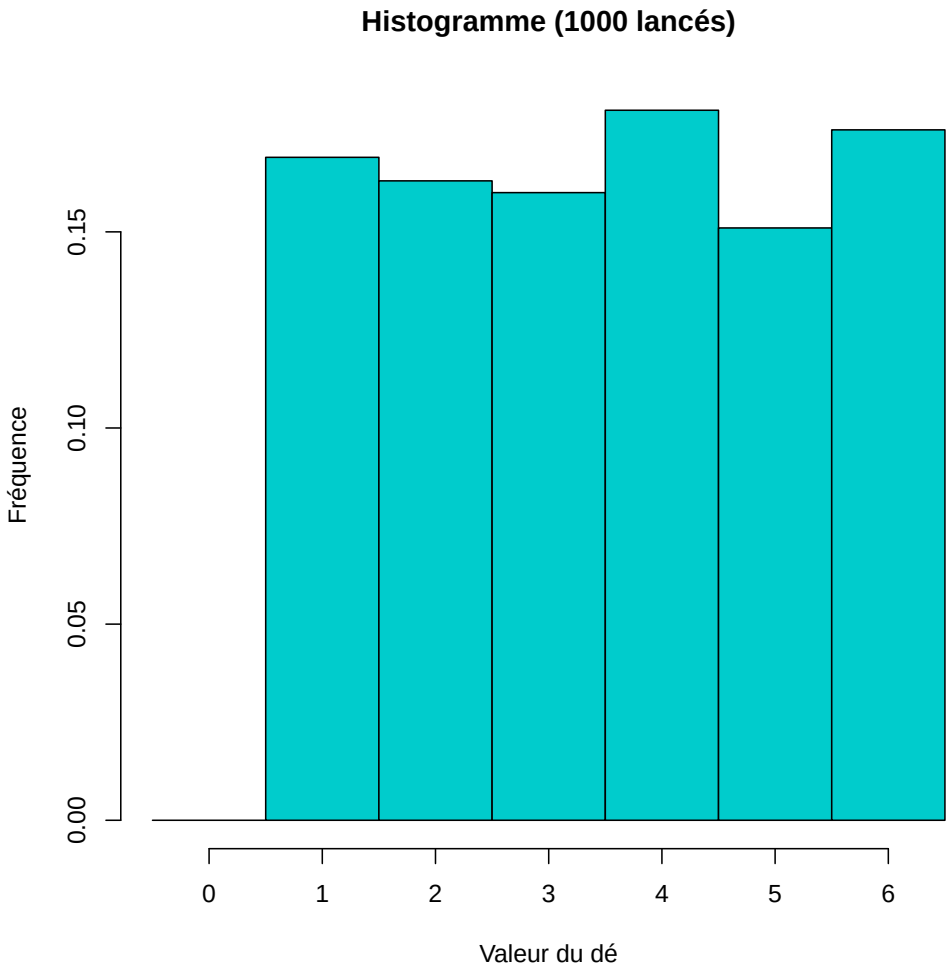
Histogramme

Un *histogramme* représente la fréquence d'apparition de chaque observation.



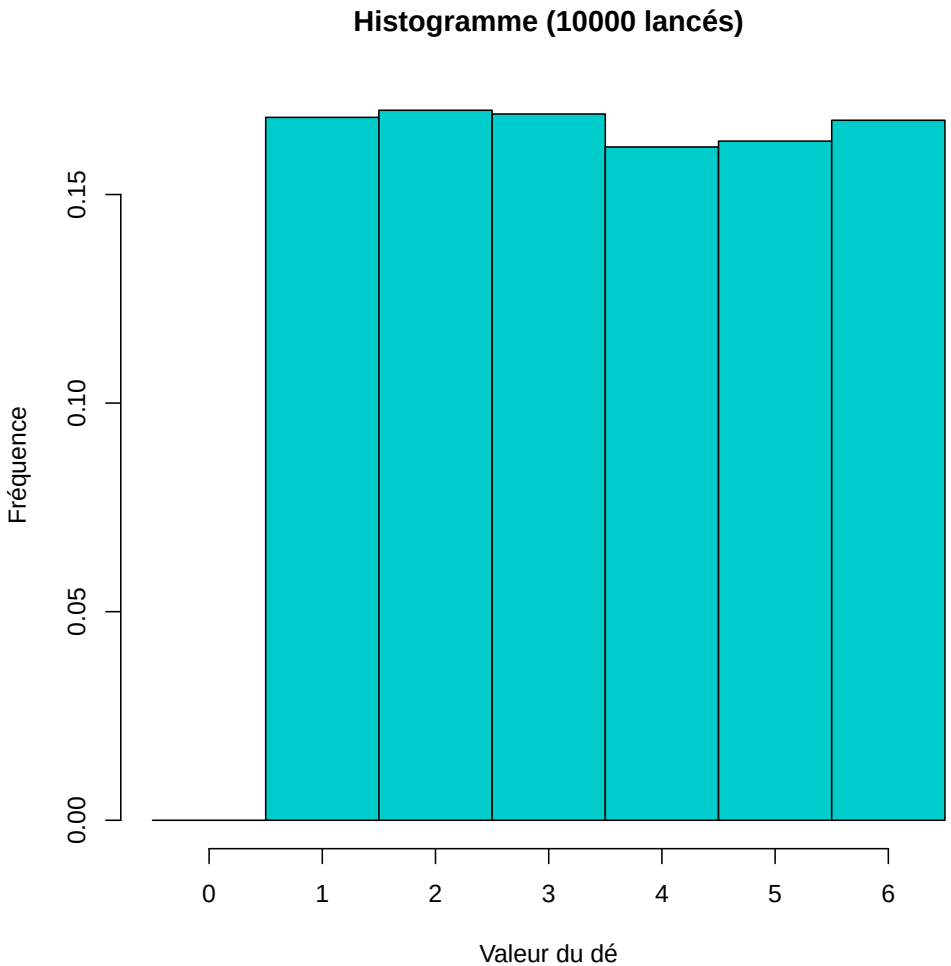
Histogramme

Un *histogramme* représente la fréquence d'apparition de chaque observation.



Histogramme

Un *histogramme* représente la fréquence d'apparition de chaque observation.



Observations

- Il faut "un grand nombre" d'observations pour conclure.
- Les fréquences semblent converger (vers $1/6 \approx 0.167\ldots$).

Modélisation

Variable aléatoire

Une **Variable Aléatoire** X est une fonction définie sur l'ensemble des éventualités Ω.

Pour le dé

- $\Omega = \{1, 2, 3, 4, 5, 6\}$,
- $X : \omega \mapsto X(\omega) \in \Omega$.

Modélisation

Loi de probabilité

Une *v.a.* est décrite par une **loi de probabilité**, qui mesure, pour chaque valeur possible de Ω, la probabilité que X prenne cette valeur.

Si le dé n'est pas biaisé

La loi est *uniforme*. La probabilité que X prenne une valeur donnée est de $1/6$.

L'histogramme donne une idée de la loi de probabilité.

Estimation de la loi, moyenne

- $X = v.a.$ correspond à la valeur du dé
- $p_i = \mathbb{P}(X = x_i)$ probabilité que X prenne la valeur x_i.
- $X_1, X_2, X_3, \ldots$ sont des observations.

Moyennes

Moyenne théorique $\mathbb{E}[X] = \sum_i x_i p_i$

Moyenne empirique $\bar{X}_n = \frac{1}{n} \sum_i^n X_i$

Estimation de la moyenne

Dans le cas où la moyenne théorique n'est pas connue, la moyenne *empirique* en donne une estimation, car :

$$\lim_{n \to \infty} \bar{X}_n = \mathbb{E}[X]$$

Estimation de la loi, moyenne

- $X = v.a.$ correspond à la valeur du dé
- $p_i = \mathbb{P}(X = x_i)$ probabilité que X prenne la valeur x_i.
- $X_1, X_2, X_3, \ldots$ sont des observations.

Moyennes

Moyenne théorique $\mathbb{E}[X] = \sum_i x_i p_i$

Moyenne empirique $\bar{X}_n = \frac{1}{n} \sum_i^n X_i$

Estimation de la moyenne

Dans le cas où la moyenne théorique n'est pas connue, la moyenne *empirique* en donne une estimation, car :

$$\lim_{n \to \infty} \bar{X}_n = \mathbb{E}[X]$$

Estimation de la loi, moyenne

- $X = v.a.$ correspond à la valeur du dé
- $p_i = \mathbb{P}(X = x_i)$ probabilité que X prenne la valeur x_i.
- $X_1, X_2, X_3, \ldots$ sont des observations.

Moyennes

Moyenne théorique $\mathbb{E}[X] = \sum_i x_i p_i$

Moyenne empirique $\bar{X}_n = \frac{1}{n} \sum_i^n X_i$

Estimation de la moyenne

Dans le cas où la moyenne théorique n'est pas connue, la moyenne *empirique* en donne une estimation, car :

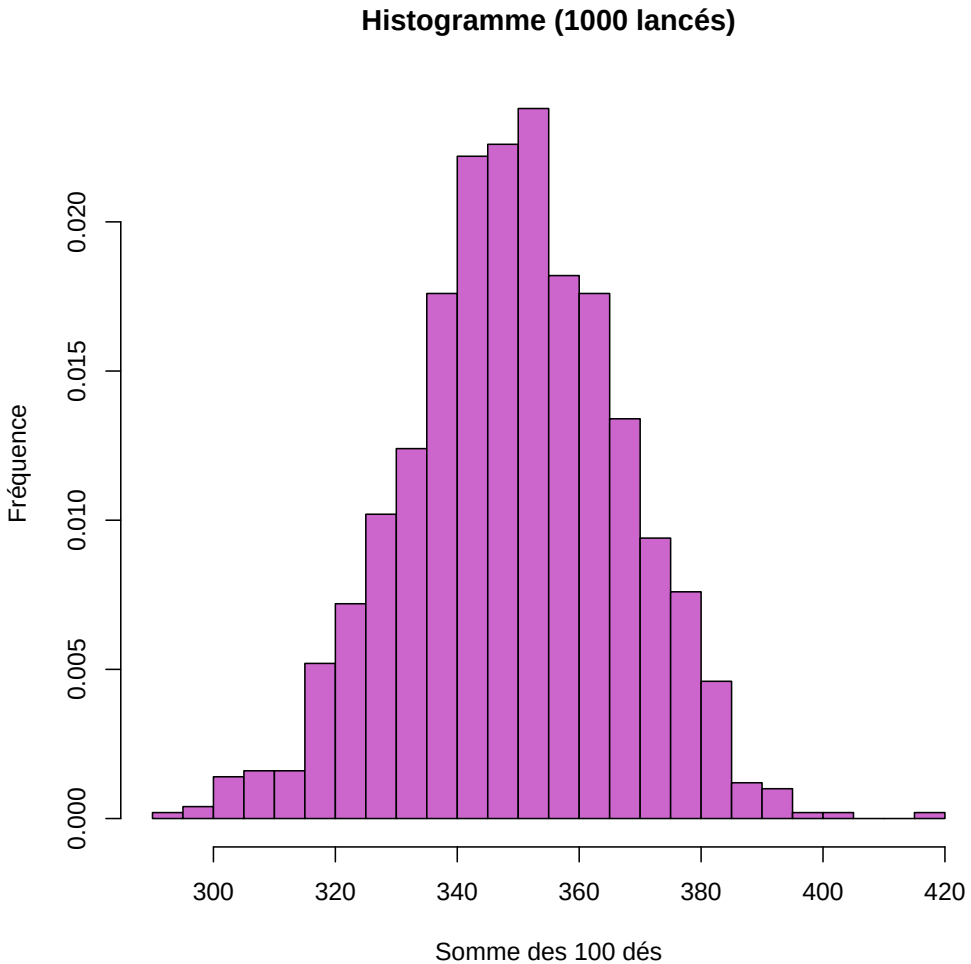
$$\lim_{n \to \infty} \bar{X}_n = \mathbb{E}[X]$$

Nouvelle expérience

Description

- Je lance 100 dés non biaisés.
- Je somme les faces.
- Je recommence l'expérience n fois.

Histogramme après 10000 lancés (des 100 dés)

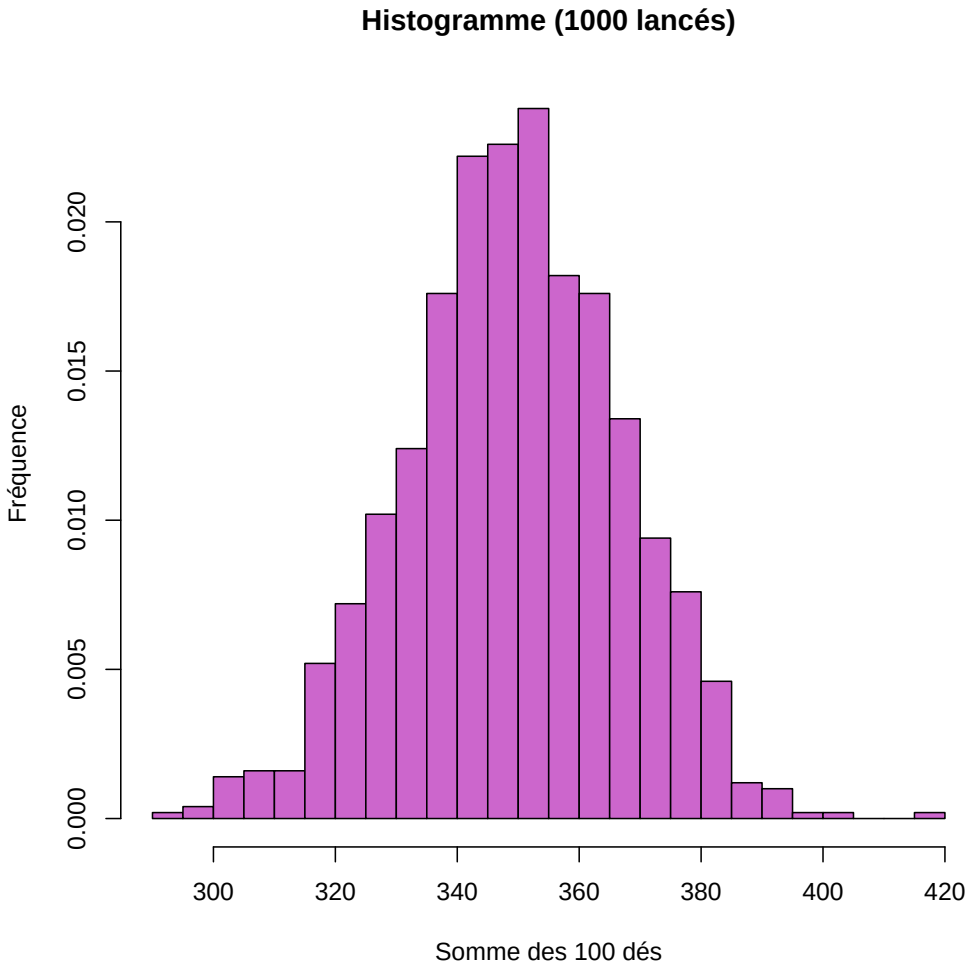


- Une idée de la moyenne empirique ?

$$\bar{X}_{1000} = \frac{1}{1000} \sum_{i=1}^{1000} X_i = 350.1$$

- Quelle confiance accorder à la moyenne empirique ?
- Quelle erreur faisons-nous par rapport à la moyenne théorique ?

Histogramme après 10000 lancés (des 100 dés)



- Une idée de la moyenne empirique ?

$$\bar{X}_{1000} = \frac{1}{1000} \sum_{i=1}^{1000} X_i = 350.1$$

- Quelle confiance accorder à la moyenne empirique ?
- Quelle erreur faisons-nous par rapport à la moyenne théorique ?

Dispersion d'une V.A. autour de sa moyenne

Définition de la variance

Variance théorique $\mathrm{Var}(x) = \mathbb{E}[X^2] - \mathbb{E}[X]^2$

Variance empirique $S_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X}_n)^2$

Définition de l'écart-type

Écart-type théorique $\sqrt{\mathrm{Var}(x)}$

Écart-type empirique S_n

Dispersion d'une V.A. autour de sa moyenne

Définition de la variance

Variance théorique $\mathrm{Var}(x) = \mathbb{E}[X^2] - \mathbb{E}[X]^2$

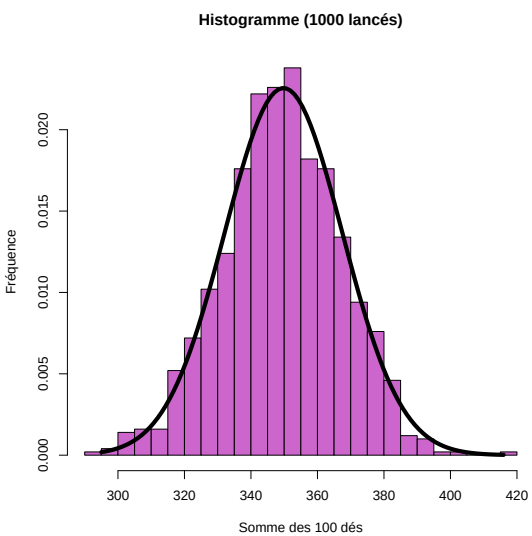
Variance empirique $S_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X}_n)^2$

Définition de l'écart-type

Écart-type théorique $\sqrt{\mathrm{Var}(x)}$

Écart-type empirique S_n

Loi normale, ou Loi Gaussienne



Loi normale

- Permet de modéliser de nombreux phénomènes aléatoires naturels ;
- Caractérisée entièrement par la moyenne μ et l'écart-type σ.
- Densité :

$$f(x) = \frac{1}{2\pi\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

- Notation :

$$X \sim \mathcal{N}(\mu, \sigma^2)$$

Loi normale, ou Loi Gaussienne

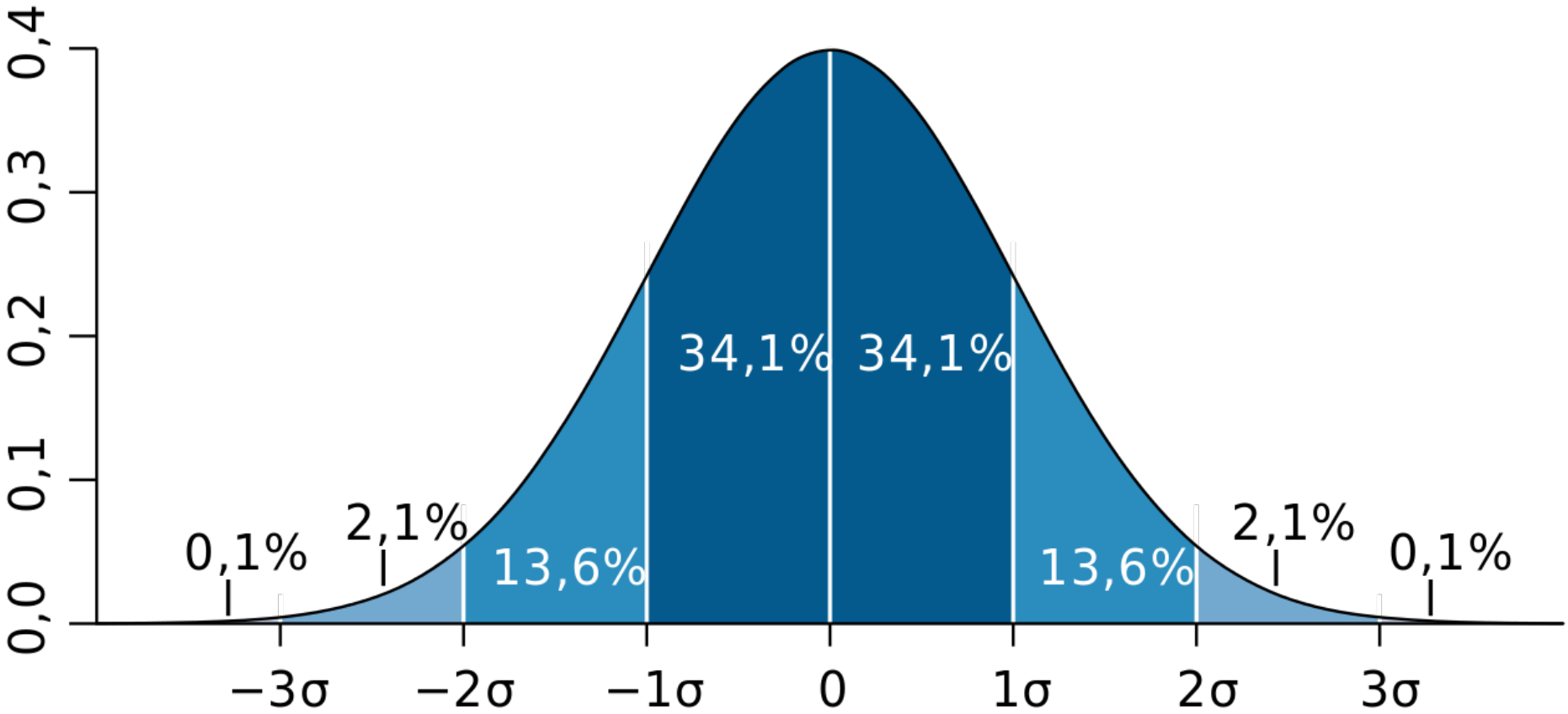


FIGURE – Loi normale, par Nusha à Slovenian Wikipedia (GFDL)

Intervalle de confiance pour la moyenne

Définition d'un intervalle de confiance

L'*Intervalle de Confiance* (IC) de niveau α est tel que :

- il est centré autour de la *moyenne empirique* ;
- il contient la *moyenne théorique* avec une probabilité α.

L'IC de niveau α est défini par :

$$IC = \left[\bar{X}_n - t_\alpha \frac{S_n}{\sqrt{n}}; \bar{X}_n + t_\alpha \frac{S_n}{\sqrt{n}}\right]$$

où :

- t_α est un réel qui dépend de α. Des tables reliant t_α et α existent dans la littérature. Par exemple, pour $\alpha = 95\%$, on a $t_\alpha = 1.96$ [1]

1. si n suffisamment grand

Intervalle de confiance pour la moyenne

Définition d'un intervalle de confiance

L'*Intervalle de Confiance* (IC) de niveau α est tel que :

- il est centré autour de la *moyenne empirique* ;
- il contient la *moyenne théorique* avec une probabilité α.

L'IC de niveau α est défini par :

$$IC = \left[\bar{X}_n - t_\alpha \frac{S_n}{\sqrt{n}}; \bar{X}_n + t_\alpha \frac{S_n}{\sqrt{n}}\right]$$

où :

- t_α est un réel qui dépend de α. Des tables reliant t_α et α existent dans la littérature. Par exemple, pour $\alpha = 95\%$, on a $t_\alpha = 1.96$ [1]

1. si n suffisamment grand

Intervalle de confiance pour la moyenne

Définition d'un intervalle de confiance

L'*Intervalle de Confiance* (IC) de niveau α est tel que :

- il est centré autour de la *moyenne empirique* ;
- il contient la *moyenne théorique* avec une probabilité α.

L'IC de niveau α est défini par :

$$IC = \left[\bar{X}_n - t_\alpha \frac{S_n}{\sqrt{n}}; \bar{X}_n + t_\alpha \frac{S_n}{\sqrt{n}}\right]$$

où :

- t_α est un réel qui dépend de α. Des tables reliant t_α et α existent dans la littérature. Par exemple, pour $\alpha = 95\%$, on a $t_\alpha = 1.96$ [1]

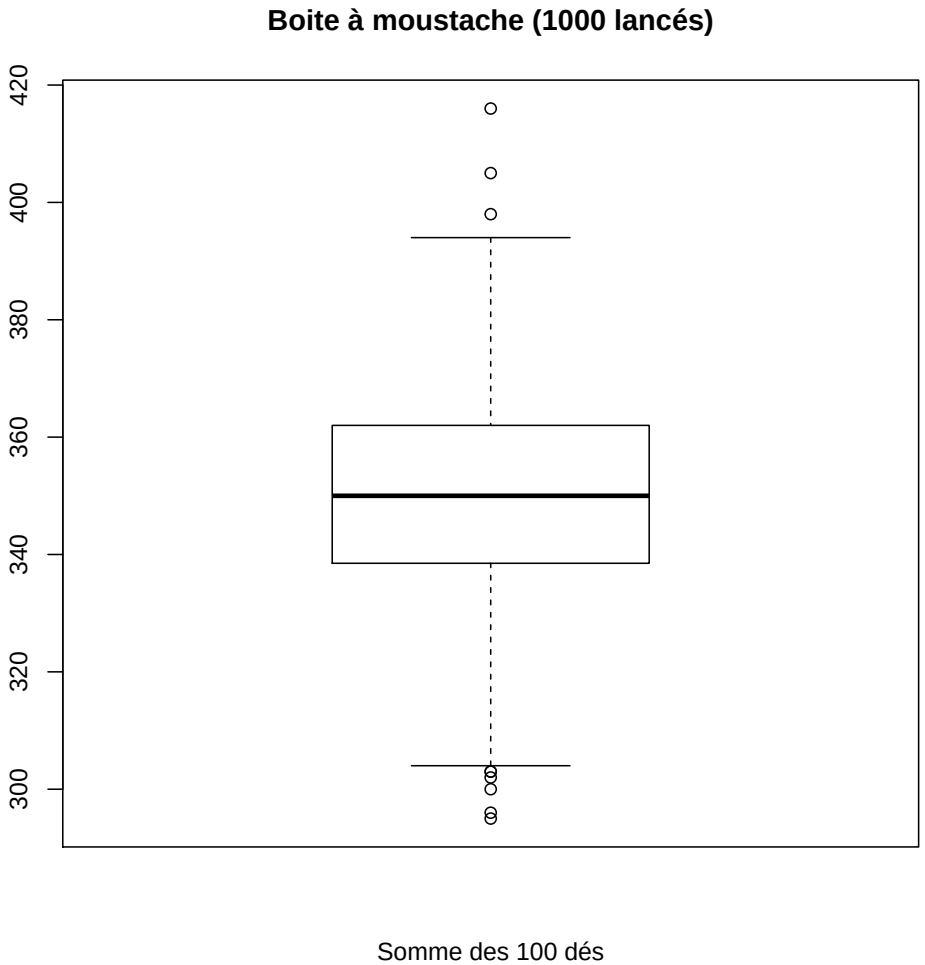
1. si n suffisamment grand

Intervalle de confiance pour la somme des 1000 dés

Soit n le nombre de lancés.

n	100	1000
$\bar{X}_n$	351.4	350.1
S_n	17.1	16.5
IC à 95%	$[348.1, 354.8]$	$[349.1, 351.1]$

Boite à moustache



- maximum
- 2^{e} quartile
- médiane
- 1^{er} quartile
- minimum

Les cercles représentent les valeurs exclues, car trop extrêmes, ou atypiques.

Plan

- R est un logiciel libre.
- R est un langage dédié aux statistiques et à la représentation des données.
- Disponible sur toutes les plateformes : `http://www.r-project.org`
- Langage matriciel de la même famille que Matlab ou Scilab.

Aide

- Documentation riche (livres, forum, etc.)
- `http://www.duclert.org/Aide-memoire-R/Le-langage/Introduction.php`
- `help(mean)` ou bien `?mean`

Environnement

- top-level
- écrire des scripts : `source('mon_script.r')`

Démo de R

Environnement

- top-level
- écrire des scripts : `source('mon_script.r')`

Démo de R

Plan

Jeu de données "*crabs*" (Leptograpsus variegatus)

200 crabes ont été ramassés en Australie.

Pour chaque crabe, l'équipe de scientifique a relevé :

- sa couleur (*Orange* ou *Bleu*)
- son sexe (*F* ou *M*)
- la taille du lobe frontal (`FL`, *frontal lobe*)
- la taille de l'arrière train (`RW`, *rear width*)
- la longueur de la carapace (`CL`, *carapace length*)
- la largeur de la carapace (`CW`, *carapace width*)
- l'épaisseur du corps (`BD`, *body depth*)

Nous allons décrire une partie de cet ensemble de données, à l'aide du logiciel *R*.

Jeu de données "*crabs*" (Leptograpsus variegatus)

200 crabes ont été ramassés en Australie.

Pour chaque crabe, l'équipe de scientifique a relevé :

- sa couleur (*Orange* ou *Bleu*)
- son sexe (*F* ou *M*)
- la taille du lobe frontal (`FL`, *frontal lobe*)
- la taille de l'arrière train (`RW`, *rear width*)
- la longueur de la carapace (`CL`, *carapace length*)
- la largeur de la carapace (`CW`, *carapace width*)
- l'épaisseur du corps (`BD`, *body depth*)

Nous allons décrire une partie de cet ensemble de données, à l'aide du logiciel *R*.

Jeu de données "*crabs*" (Leptograpsus variegatus)

200 crabes ont été ramassés en Australie.

Pour chaque crabe, l'équipe de scientifique a relevé :

- sa couleur (*Orange* ou *Bleu*)
- son sexe (*F* ou *M*)
- la taille du lobe frontal (`FL`, *frontal lobe*)
- la taille de l'arrière train (`RW`, *rear width*)
- la longueur de la carapace (`CL`, *carapace length*)
- la largeur de la carapace (`CW`, *carapace width*)
- l'épaisseur du corps (`BD`, *body depth*)

Nous allons décrire une partie de cet ensemble de données, à l'aide du logiciel *R*.

Jeu de données "*crabs*" (Leptograpsus variegatus)

200 crabes ont été ramassés en Australie.

Pour chaque crabe, l'équipe de scientifique a relevé :

- sa couleur (*Orange* ou *Bleu*)
- son sexe (*F* ou *M*)
- la taille du lobe frontal (`FL`, *frontal lobe*)
- la taille de l'arrière train (`RW`, *rear width*)
- la longueur de la carapace (`CL`, *carapace length*)
- la largeur de la carapace (`CW`, *carapace width*)
- l'épaisseur du corps (`BD`, *body depth*)

Nous allons décrire une partie de cet ensemble de données, à l'aide du logiciel *R*.

Jeu de données "*crabs*" (Leptograpsus variegatus)

200 crabes ont été ramassés en Australie.

Pour chaque crabe, l'équipe de scientifique a relevé :

- sa couleur (*Orange* ou *Bleu*)
- son sexe (*F* ou *M*)
- la taille du lobe frontal (`FL`, *frontal lobe*)
- la taille de l'arrière train (`RW`, *rear width*)
- la longueur de la carapace (`CL`, *carapace length*)
- la largeur de la carapace (`CW`, *carapace width*)
- l'épaisseur du corps (`BD`, *body depth*)

Nous allons décrire une partie de cet ensemble de données, à l'aide du logiciel *R*.

Jeu de données "*crabs*" (Leptograpsus variegatus)

200 crabes ont été ramassés en Australie.

Pour chaque crabe, l'équipe de scientifique a relevé :

- sa couleur (*Orange* ou *Bleu*)
- son sexe (*F* ou *M*)
- la taille du lobe frontal (`FL`, *frontal lobe*)
- la taille de l'arrière train (`RW`, *rear width*)
- la longueur de la carapace (`CL`, *carapace length*)
- la largeur de la carapace (`CW`, *carapace width*)
- l'épaisseur du corps (`BD`, *body depth*)

Nous allons décrire une partie de cet ensemble de données, à l'aide du logiciel *R*.

Jeu de données "*crabs*" (Leptograpsus variegatus)

200 crabes ont été ramassés en Australie.

Pour chaque crabe, l'équipe de scientifique a relevé :

- sa couleur (*Orange* ou *Bleu*)
- son sexe (*F* ou *M*)
- la taille du lobe frontal (`FL`, *frontal lobe*)
- la taille de l'arrière train (`RW`, *rear width*)
- la longueur de la carapace (`CL`, *carapace length*)
- la largeur de la carapace (`CW`, *carapace width*)
- l'épaisseur du corps (`BD`, *body depth*)

Nous allons décrire une partie de cet ensemble de données, à l'aide du logiciel *R*.

Jeu de données "*crabs*" (Leptograpsus variegatus)

200 crabes ont été ramassés en Australie.

Pour chaque crabe, l'équipe de scientifique a relevé :

- sa couleur (*Orange* ou *Bleu*)
- son sexe (*F* ou *M*)
- la taille du lobe frontal (`FL`, *frontal lobe*)
- la taille de l'arrière train (`RW`, *rear width*)
- la longueur de la carapace (`CL`, *carapace length*)
- la largeur de la carapace (`CW`, *carapace width*)
- l'épaisseur du corps (`BD`, *body depth*)

Nous allons décrire une partie de cet ensemble de données, à l'aide du logiciel *R*.

Résumé des données

5 enregistrements pris au hasard : `crabs[sample(200, 5), ]`

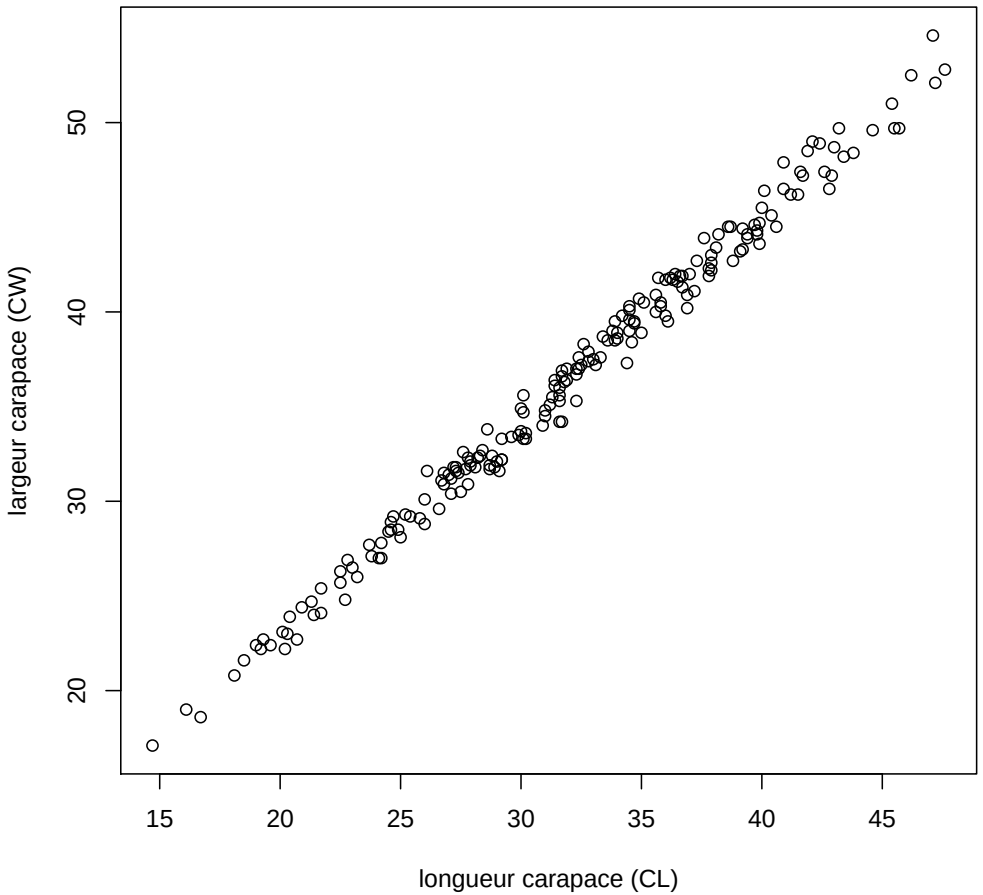
	sp	sex	index	FL	RW	CL	CW	BD
103	O	M	3	10.7	8.6	20.7	22.7	9.2
151	O	F	1	10.7	9.7	21.4	24.0	9.8
4	B	M	4	9.6	7.9	20.1	23.1	8.2
192	O	F	42	20.5	17.5	40.0	45.5	19.2
157	O	F	7	14.0	12.8	28.8	32.4	12.7

Résumé des données

Données statistiques par variables : `summary(crabs)`

```
sp       sex          index           FL              RW
 B:100   F:100   Min.   : 1.0   Min.   : 7.20   Min.   : 6.50
 O:100   M:100   1st Qu.:13.0   1st Qu.:12.90   1st Qu.:11.00
                 Median :25.5   Median :15.55   Median :12.80
                 Mean   :25.5   Mean   :15.58   Mean   :12.74
                 3rd Qu.:38.0   3rd Qu.:18.05   3rd Qu.:14.30
                 Max.   :50.0   Max.   :23.10   Max.   :20.20
       CW              BD
 Min.   :17.10   Min.   : 6.10
 1st Qu.:31.50   1st Qu.:11.40
 Median :36.80   Median :13.90
 Mean   :36.41   Mean   :14.03
 3rd Qu.:42.00   3rd Qu.:16.60
 Max.   :54.60   Max.   :21.60
```

Chercher les liens entre les différentes variables



Données bi-variées

- Ensemble de couples de données (x_i, y_i)
- En R : `plot(x, y)`

Interprétation ?

Chercher les liens entre les différentes variables

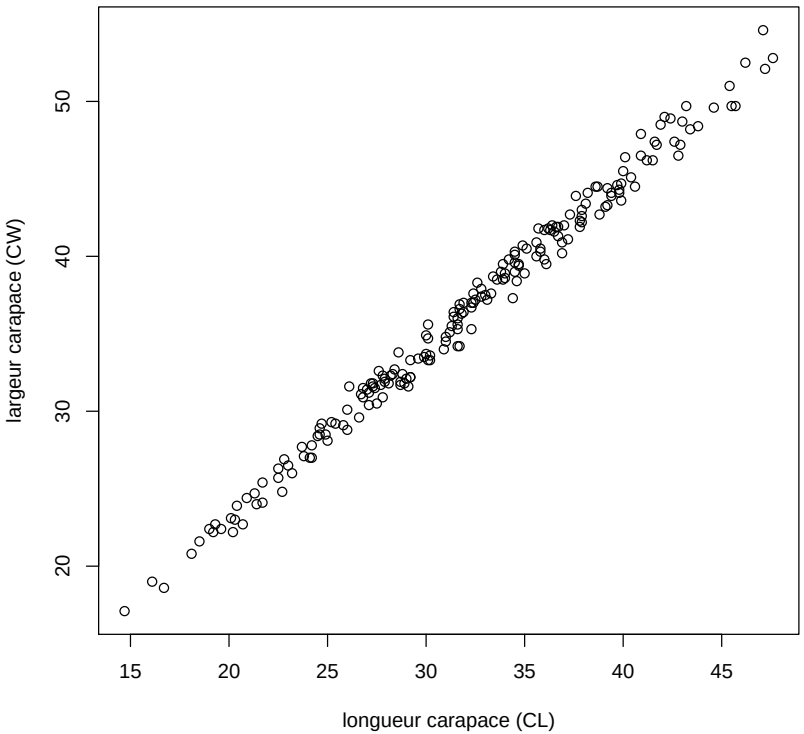


FIGURE – Probable

Interprétation ?

Comment interpréter ce que l'on voit ?

Chercher les liens entre les différentes variables

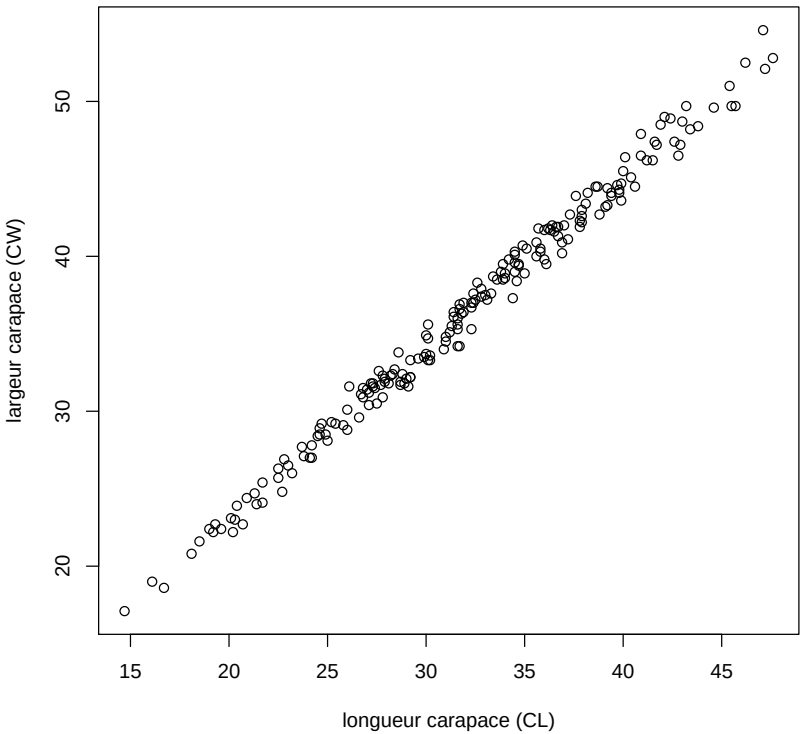
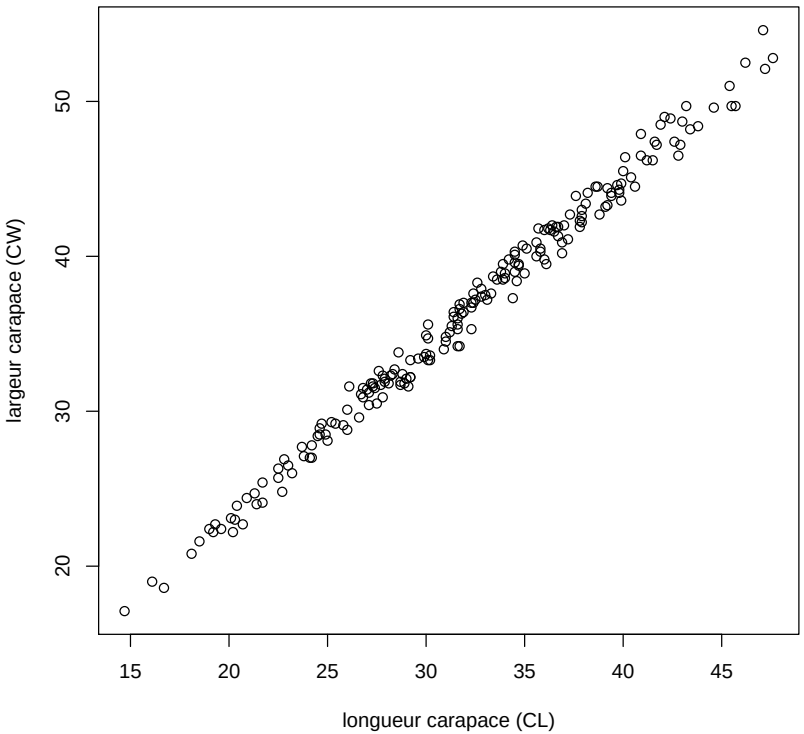


FIGURE – Probable

Interprétation ?

Comment interpréter ce que l'on voit ?

Chercher les liens entre les différentes variables



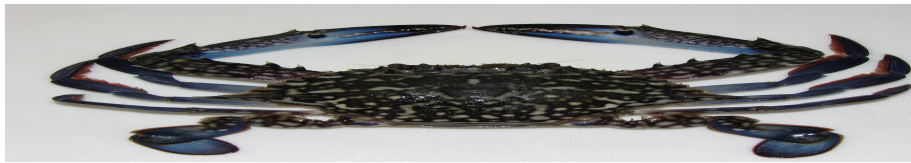


FIGURE – Improbable

Interprétation ?

Comment interpréter ce que l'on voit ?

Chercher les liens entre les différentes variables

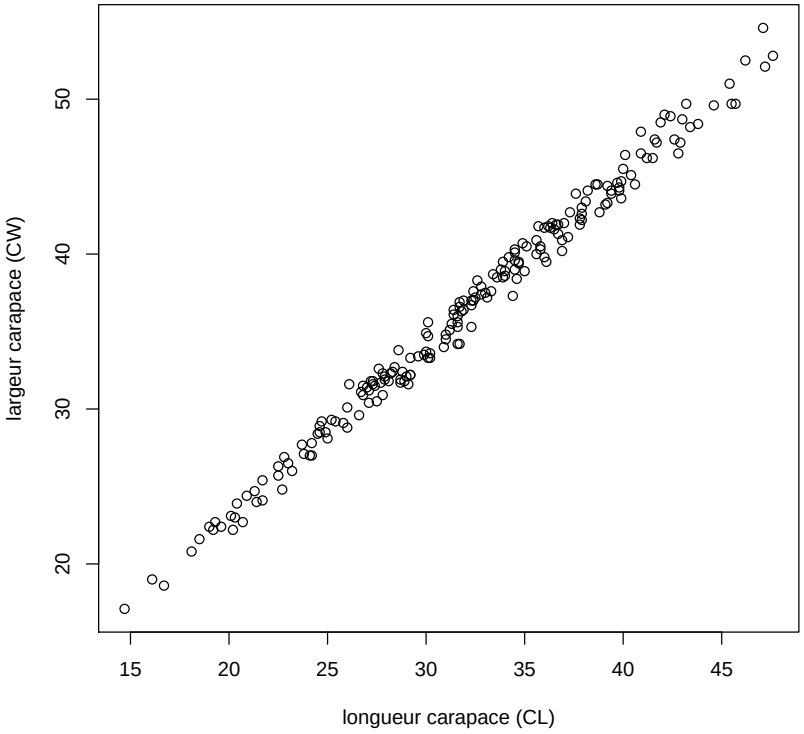
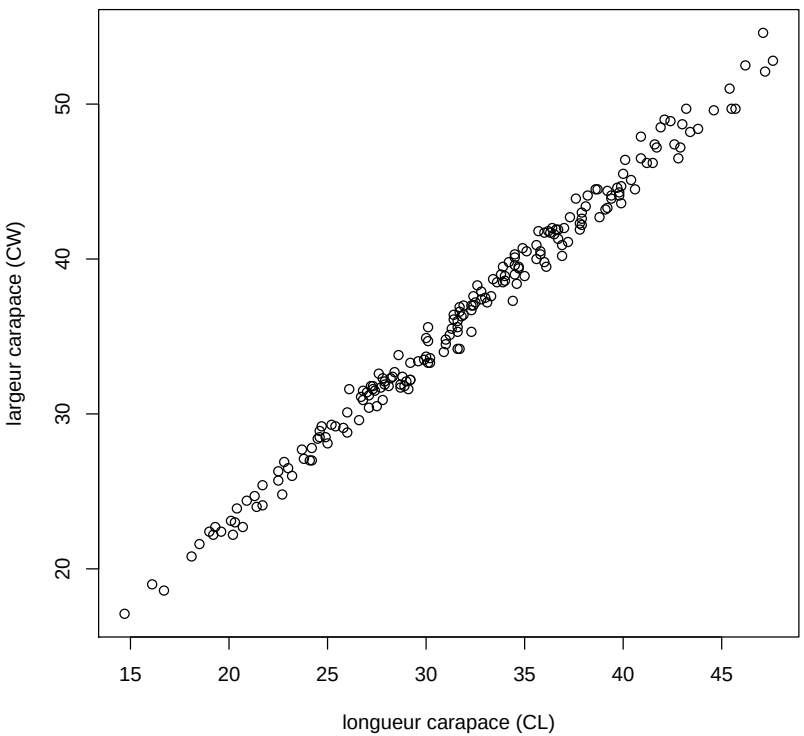


FIGURE – Improbable

Interprétation ?

Comment interpréter ce que l'on voit ?

Régression linéaire



Modèle linéaire

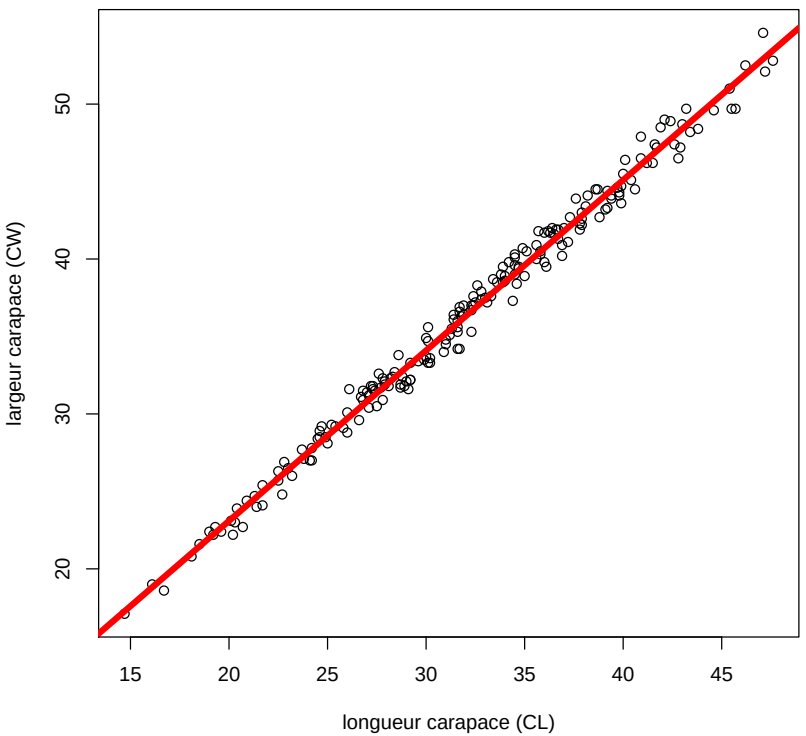
Le nuage de points (x_i, y_i) est remplacé par une droite d'équation :

$$y = ax + b$$

On peut alors :

- Expliquer les relations entre les variables ;
- Prédire des valeurs.

Régression linéaire



Modèle linéaire

Le nuage de points (x_i, y_i) est remplacé par une droite d'équation :

$$y = ax + b$$

On peut alors :

- Expliquer les relations entre les variables ;
- Prédire des valeurs.

Régression linéaire

Ajustement du modèle

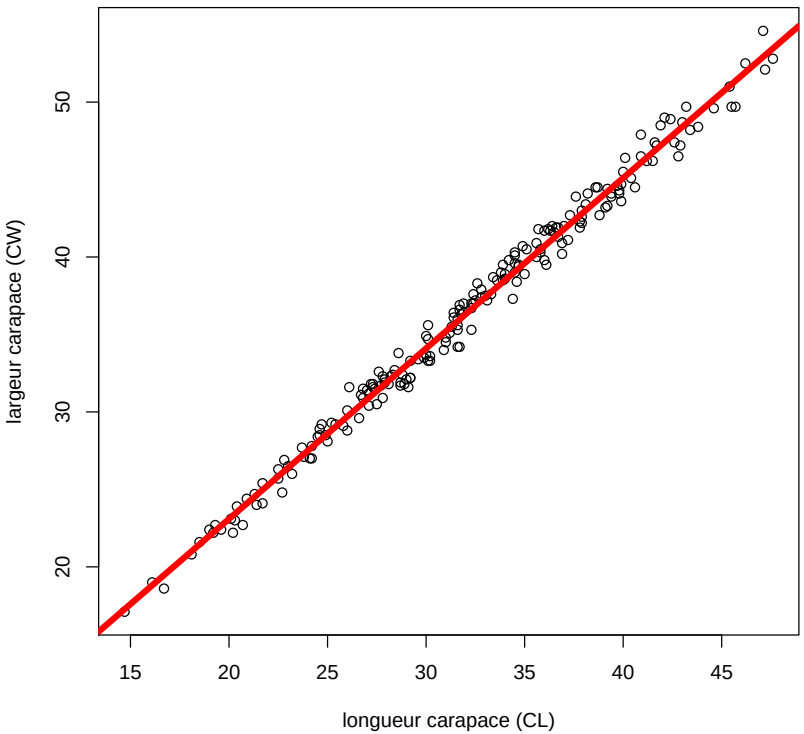
$$f(x) = ax + b$$

- Trouver a et b de sorte à minimiser l'erreur quadratique :

$$MSE = \frac{1}{n} \sum_{i=1}^{n} (f(x_i) - y_i)^2$$

- Fonction R : `lm(y~x)`

Régression linéaire sur CW et CL




```
> model = lm(CW ~ CL, crabs)
> model$coefficients
(Intercept)          CL
   1.089919    1.100266
```

Corrélation linéaire entre deux variables

Corrélation linéaire

On voit que CL et CW sont fortement liées. On dit que ces variables sont *corrélées*.

On quantifie la corrélation entre deux variables X et Y par un réel compris entre -1 et 1.

Forte corrélation $|\mathrm{Cor}(X, Y)| > 0.8$

Faible corrélation $|\mathrm{Cor}(X, Y)| < 0.3$

Corrélation linéaire entre deux variables

Corrélation linéaire

On voit que CL et CW sont fortement liées. On dit que ces variables sont *corrélées*.
On quantifie la corrélation entre deux variables X et Y par un réel compris entre -1 et 1.

Forte corrélation $|\mathrm{Cor}(X, Y)| > 0.8$

Faible corrélation $|\mathrm{Cor}(X, Y)| < 0.3$

Corrélation linéaire entre deux variables

Corrélation linéaire

On voit que CL et CW sont fortement liées. On dit que ces variables sont *corrélées*.
On quantifie la corrélation entre deux variables X et Y par un réel compris entre -1 et 1.

Forte corrélation $|\mathrm{Cor}(X, Y)| > 0.8$

Faible corrélation $|\mathrm{Cor}(X, Y)| < 0.3$

Corrélation linéaire pour le jeu de données Crabs

```
> cor(crabs[,4:8])
       FL        RW        CL        CW        BD
FL 1.0000000 0.9069876 0.9788418 0.9649558 0.9876272
RW 0.9069876 1.0000000 0.8927430 0.9004021 0.8892054
CL 0.9788418 0.8927430 1.0000000 0.9950225 0.9832038
CW 0.9649558 0.9004021 0.9950225 1.0000000 0.9678117
BD 0.9876272 0.8892054 0.9832038 0.9678117 1.0000000
```

Attention au coefficient de corrélation !

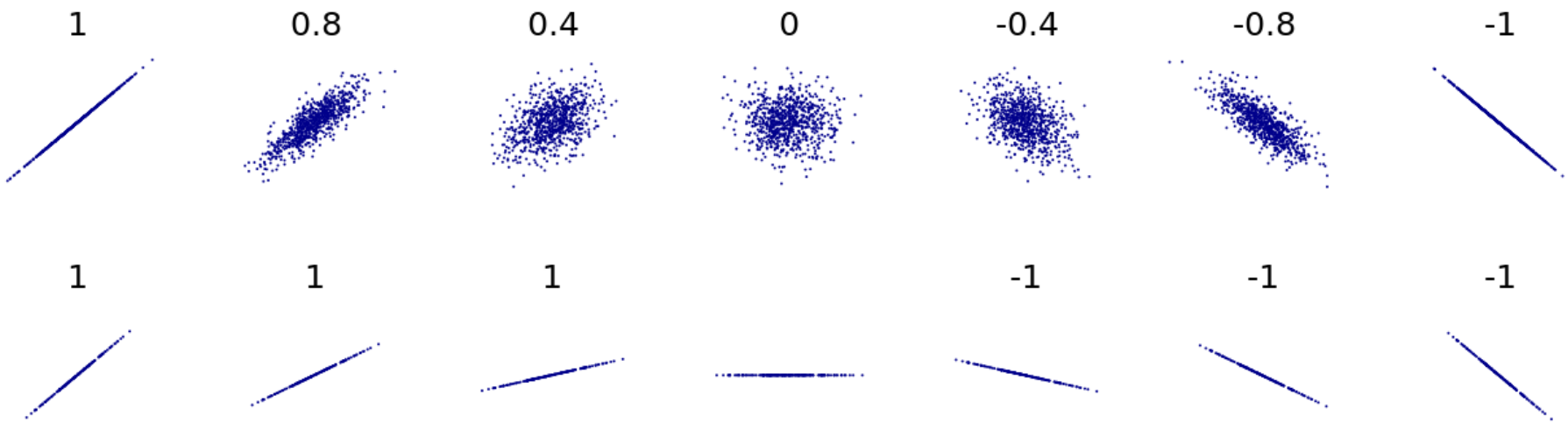


FIGURE – By Denis Boigelot, [CC0], via Wikimedia Commons

Attention au coefficient de corrélation !

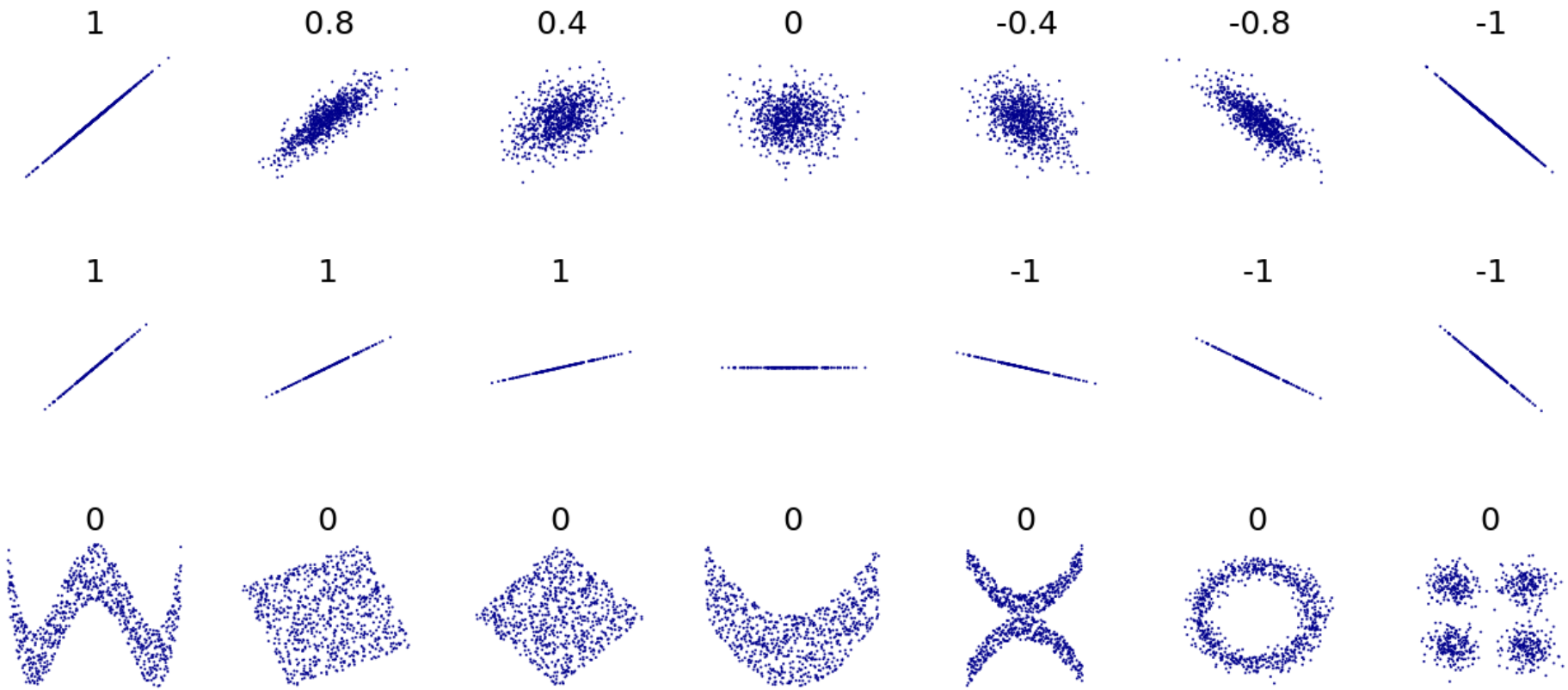
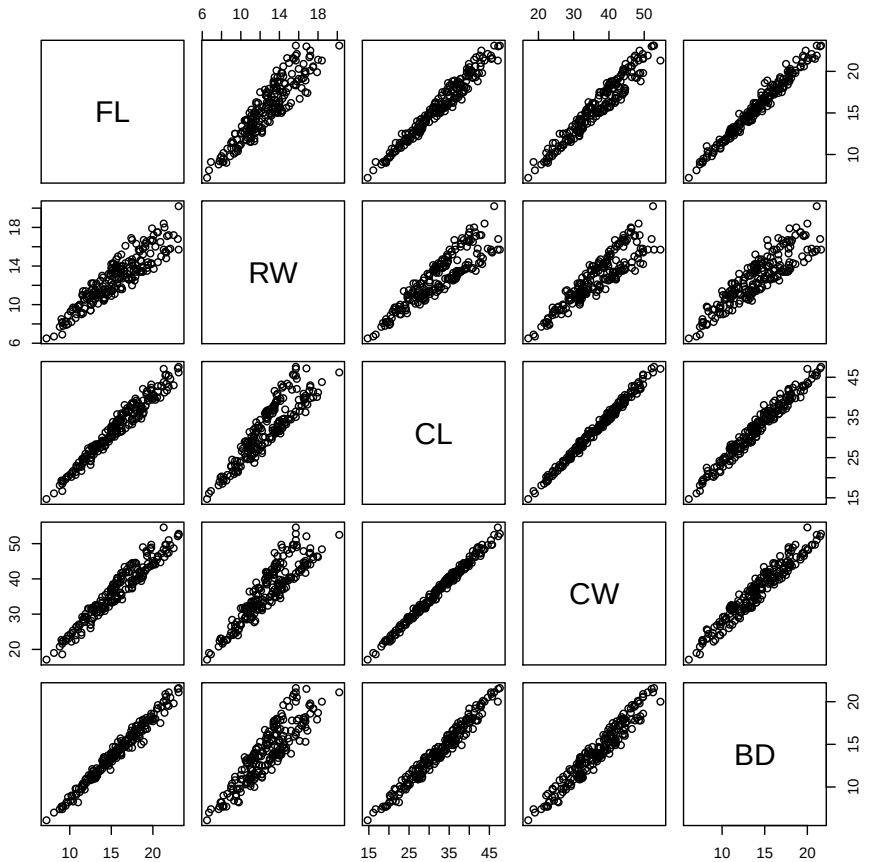


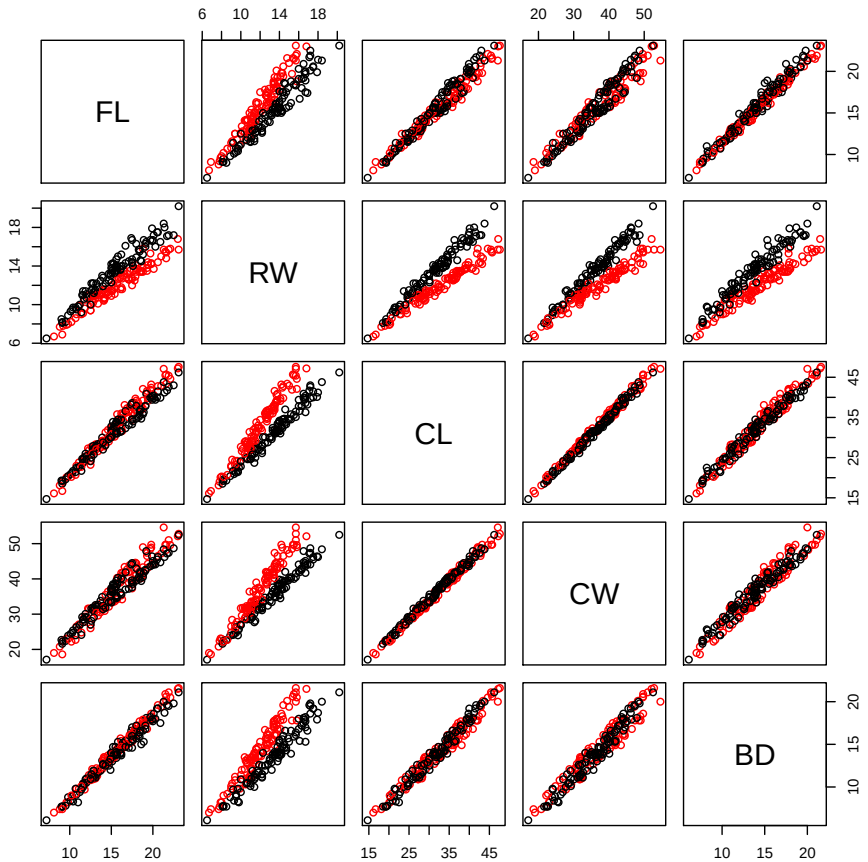
FIGURE – By Denis Boigelot, [CC0], via Wikimedia Commons

Visualiser rapidement les relations entre les variables




```
pairs(crabs[,4:8])
```

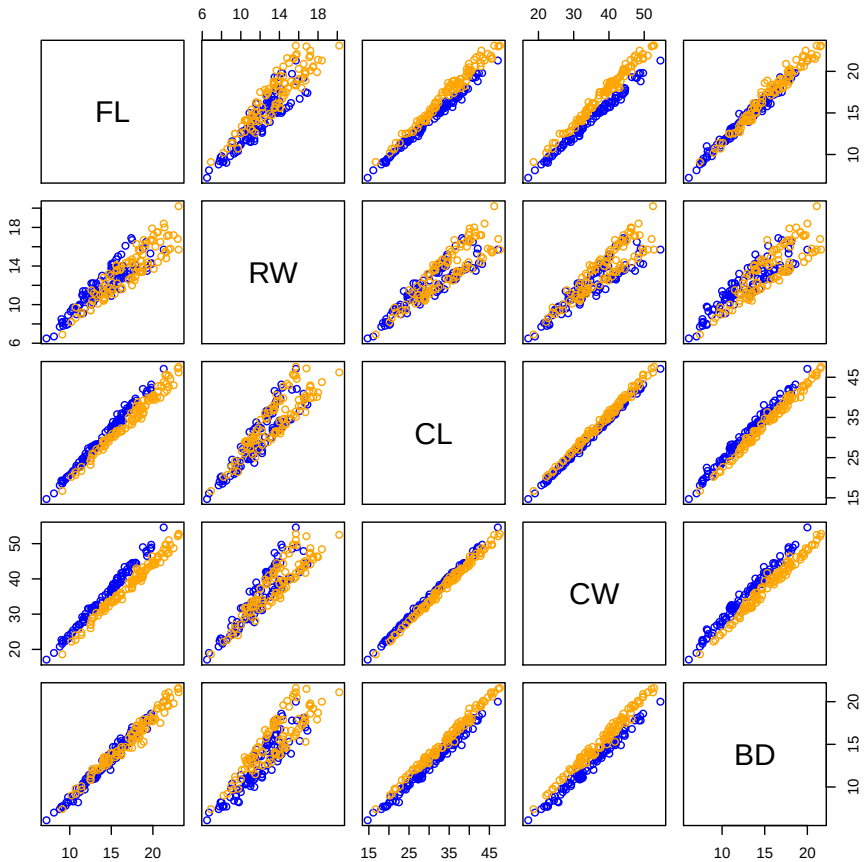
Visualiser rapidement les relations entre les variables




```
pairs(crabs[,4:8],
col=crabs$sex)
```

F	noir
M	rouge

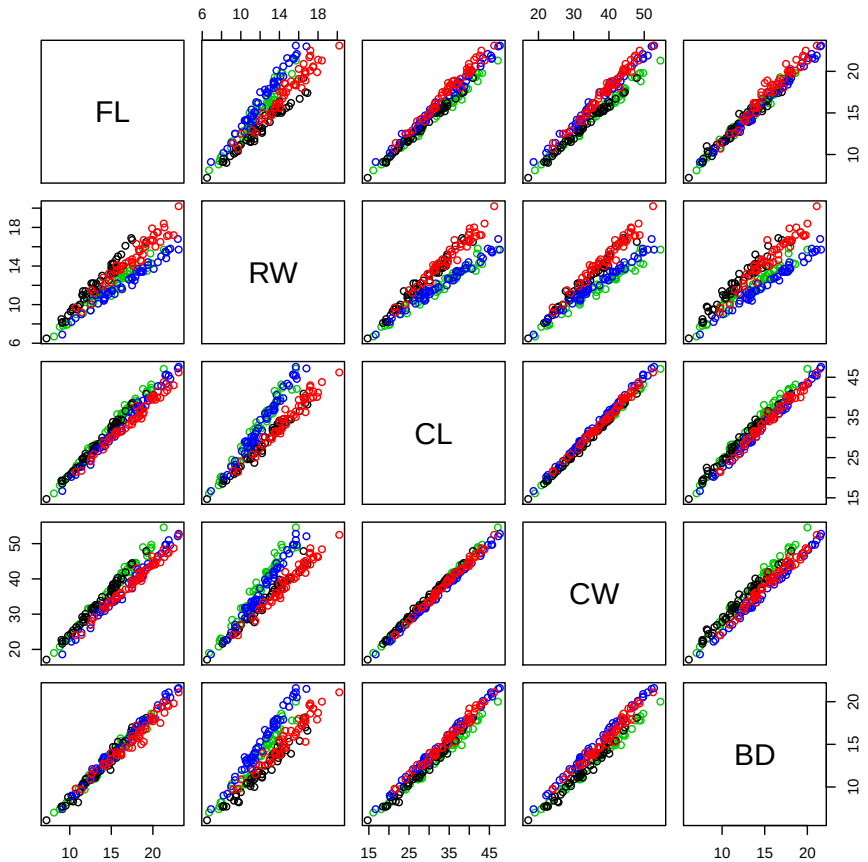
Visualiser rapidement les relations entre les variables




```
pairs(crabs[,4:8],
col=crabs$sp)
```

O	orange
B	bleu

Visualiser rapidement les relations entre les variables

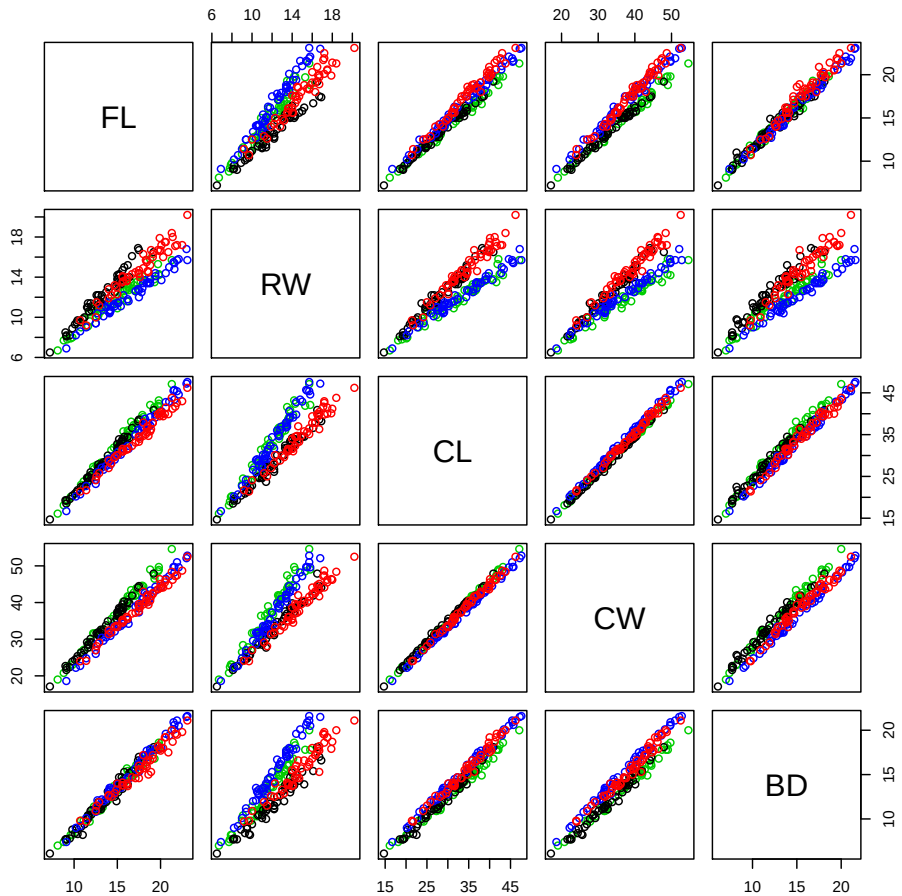



```
pairs(crabs[,4:8],
col=crabs$class)
```

	B	O
F	rouge	bleu
M	vert	noir

TABLE – classes de crabes

Peut-on prédire l'espèce et le sexe d'un crabe ?

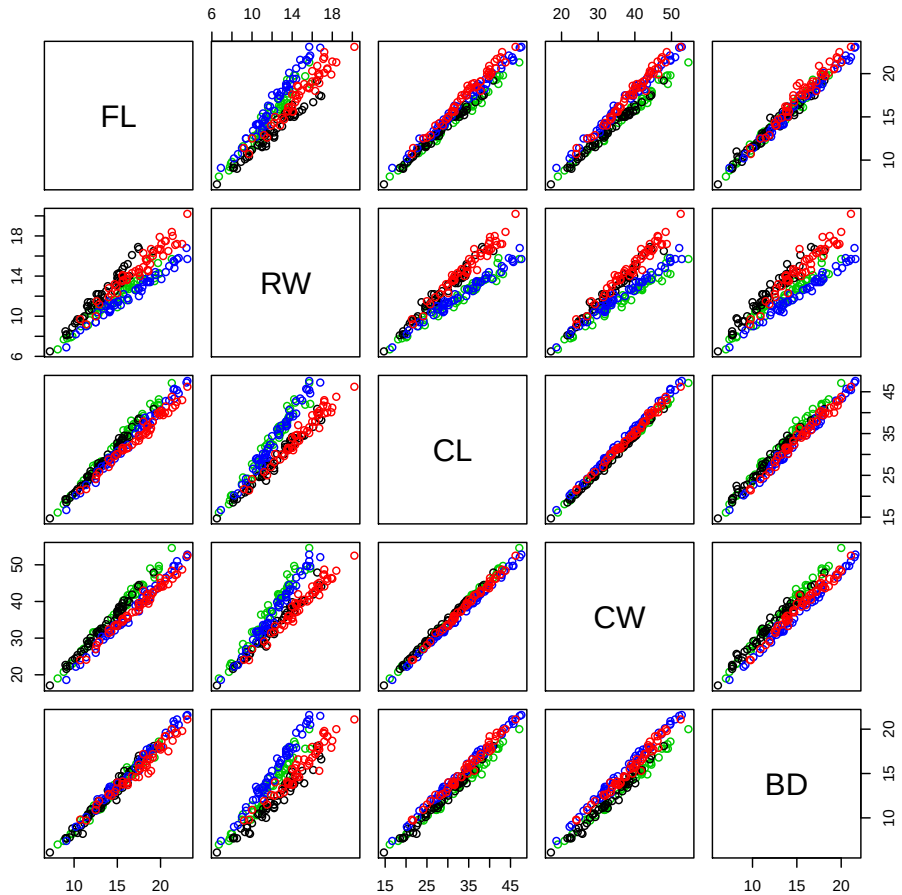



```
pairs(crabs[,4:8],
col=crabs$class)
```

	B	O
F	rouge	bleu
M	vert	noir

TABLE – classes de crabes

Peut-on prédire l'espèce et le sexe d'un crabe ?



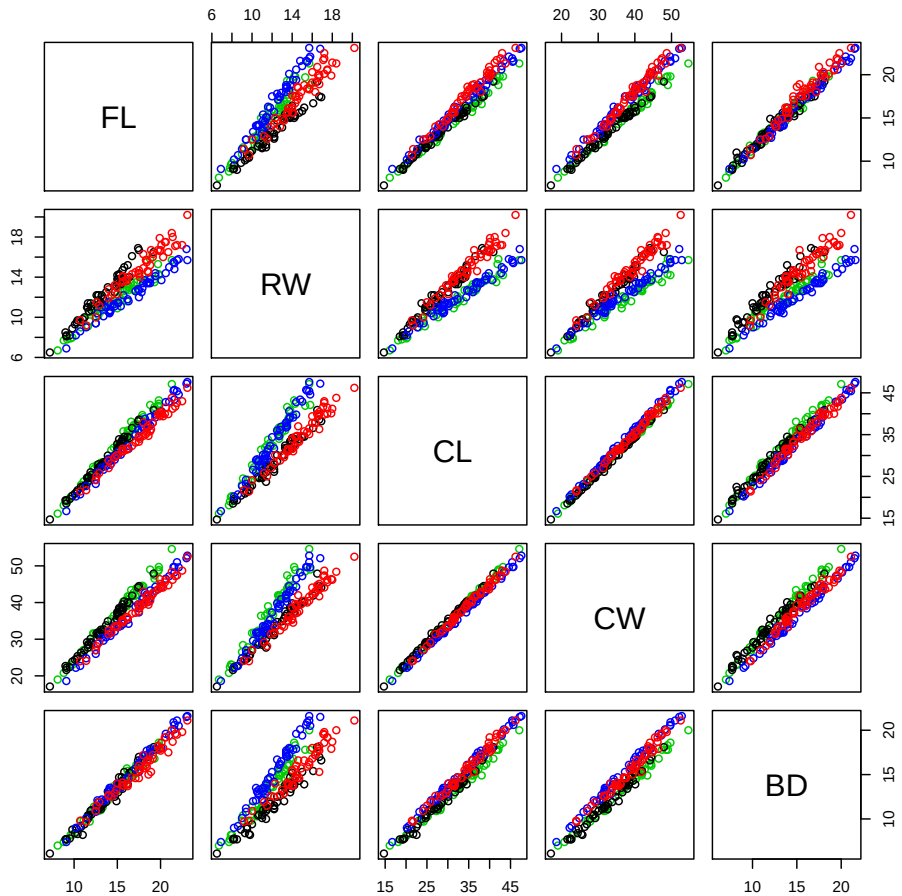

```
pairs(crabs[,4:8],
col=crabs$class)
```

	B	O
F	rouge	bleu
M	vert	noir

TABLE – classes de crabes

On voit une très forte corrélation entre les données. On assiste à ce qu'on appelle un *effet taille*.

Peut-on prédire l'espèce et le sexe d'un crabe ?



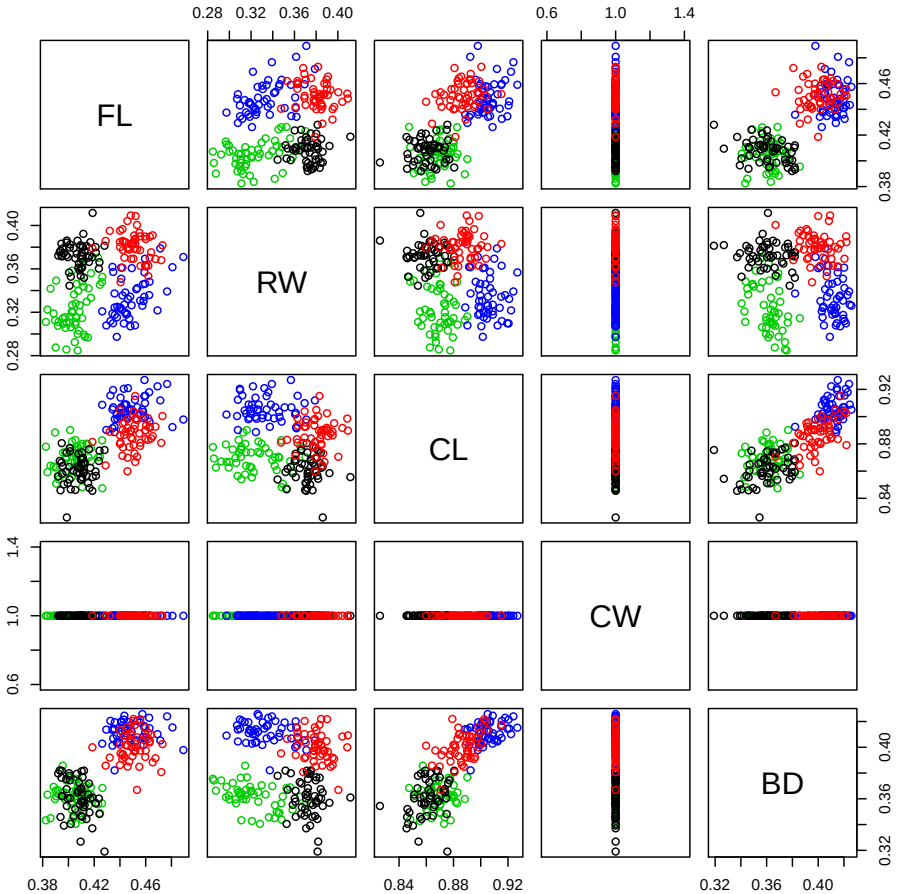
On voit une très forte corrélation entre les données. On assiste à ce qu'on appelle un *effet taille*.

On peut *normaliser* par une des variables. CW par exemple.

Pensez à l'Indice de Masse Corporelle :

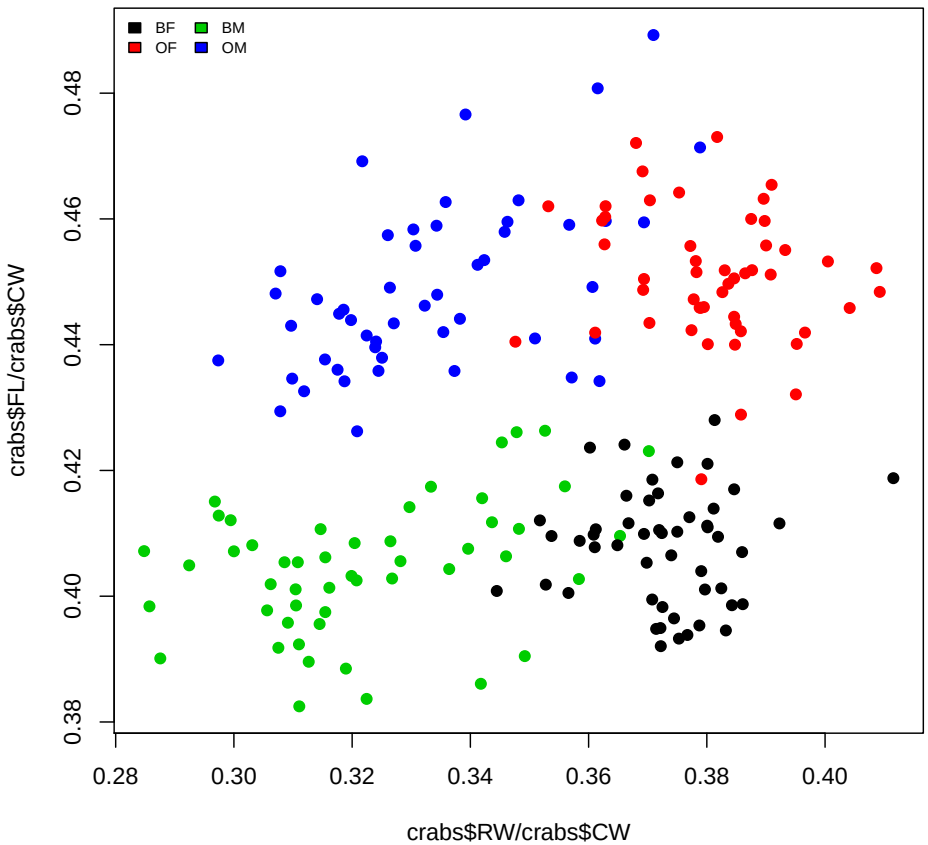
$$\text{IMC} = \frac{\text{poids}}{\text{taille}^2}$$

Peut-on prédire l'espèce et le sexe d'un crabe ?




```
pairs(crabs[,4:8]/crabs$CW,
col=crabs$class)
```

Peut-on prédire l'espèce et le sexe d'un crabe ?



On voit apparaitre quatre groupes distinct, un par classe.

Plan

Objectif

Démarche consistant à *rejeter* ou *ne pas rejeter* une hypothèse statistique[2], en fonction d'un *échantillon*.

Il s'agit d'émettre des conclusions sur une *population*, en leur rattachant des risques de se tromper.

Exemple

Hypothèse H_0 : « *Les observations suivent une loi normale* »

2. Appelée *hypothèse nulle* H_0.

Objectif

Démarche consistant à *rejeter* ou *ne pas rejeter* une hypothèse statistique [2], en fonction d'un *échantillon*.

Il s'agit d'émettre des conclusions sur une *population*, en leur rattachant des risques de se tromper.

Exemple

Hypothèse H_0 : « *Les observations suivent une loi normale* »

2. Appelée *hypothèse nulle* H_0.

p-value

Elle représente la probabilité qui mesure le degré de certitude avec lequel il est possible d'invalider l'hypothèse nulle.
Des probabilités faibles permettent d'invalider l'hypothèse nulle avec plus de certitude.

En pratique

- $p \leq 0.01$: très forte présomption contre l'hypothèse nulle
- $0.01 < p \leq 0.05$: forte présomption contre l'hypothèse nulle
- $0.05 < p \leq 0.1$: faible présomption contre l'hypothèse nulle
- $0.1 < p$: pas de présomption contre l'hypothèse nulle

p-value

Elle représente la probabilité qui mesure le degré de certitude avec lequel il est possible d'invalider l'hypothèse nulle.
Des probabilités faibles permettent d'invalider l'hypothèse nulle avec plus de certitude.

En pratique

- $p \leq 0.01$: très forte présomption contre l'hypothèse nulle
- $0.01 < p \leq 0.05$: forte présomption contre l'hypothèse nulle
- $0.05 < p \leq 0.1$: faible présomption contre l'hypothèse nulle
- $0.1 < p$: pas de présomption contre l'hypothèse nulle

Quelques exemples de tests

normalité H_0 : *Les observations suivent une loi normale*

- Test de Kolmogorov-Smirnov : `ks.test(x, "pnorm", 0, 1)`
- Test de Shapiro-Wilks : `shapiro.test(x)`

comparaison H_0 : *Les moyennes observées sont égales*

- Test t de Student : `t.test(v1, v2)` (si $v1$ et $v2$ suivent des lois normales.)
- Test de Mann-Whitneya-Wilcoxon : `wilcox.test(v1, v2)` (sinon).

Plus d'exemples sur :
https://fr.wikipedia.org/wiki/Test_(statistique)

Quelques exemples de tests

normalité H_0 : *Les observations suivent une loi normale*

- Test de Kolmogorov-Smirnov : `ks.test(x, "pnorm", 0, 1)`
- Test de Shapiro-Wilks : `shapiro.test(x)`

comparaison H_0 : *Les moyennes observées sont égales*

- Test t de Student : `t.test(v1, v2)` (si $v1$ et $v2$ suivent des lois normales.)
- Test de Mann-Whitneya-Wilcoxon : `wilcox.test(v1, v2)` (sinon).

Plus d'exemples sur :
`https://fr.wikipedia.org/wiki/Test_(statistique)`

Quelques exemples de tests

normalité H_0 : *Les observations suivent une loi normale*

- Test de Kolmogorov-Smirnov : `ks.test(x, "pnorm", 0, 1)`
- Test de Shapiro-Wilks : `shapiro.test(x)`

comparaison H_0 : *Les moyennes observées sont égales*

- Test t de Student : `t.test(v1, v2)` (si $v1$ et $v2$ suivent des lois normales.)
- Test de Mann-Whitneya-Wilcoxon : `wilcox.test(v1, v2)` (sinon).

Plus d'exemples sur :
`https://fr.wikipedia.org/wiki/Test_(statistique)`

Merci !