Le ReJMiC présente

:

Introduction aux statistiques

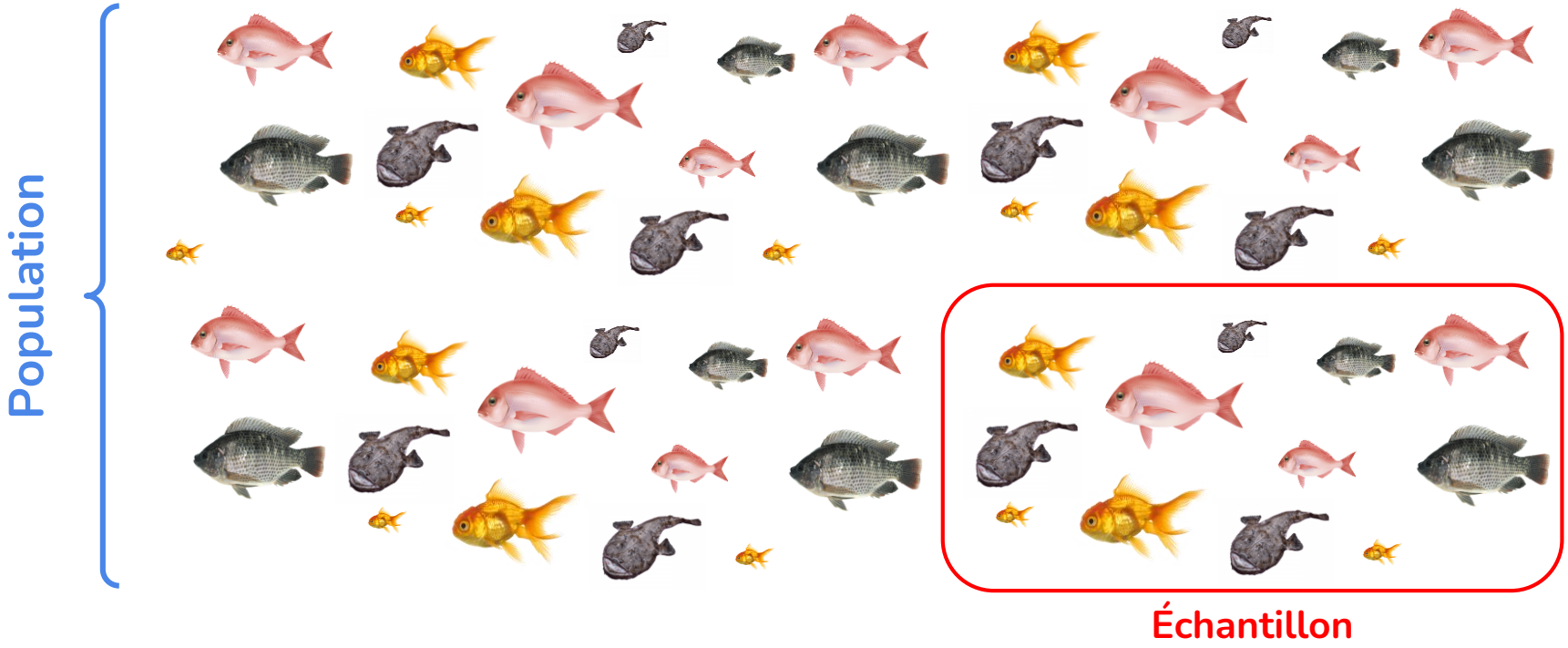
Journée d'initiation à la bio-informatique n°2
24 juin 2022

Maël Pretet

Décrire en utilisant les statistiques

Etude d’un groupe

Etude d’un groupe

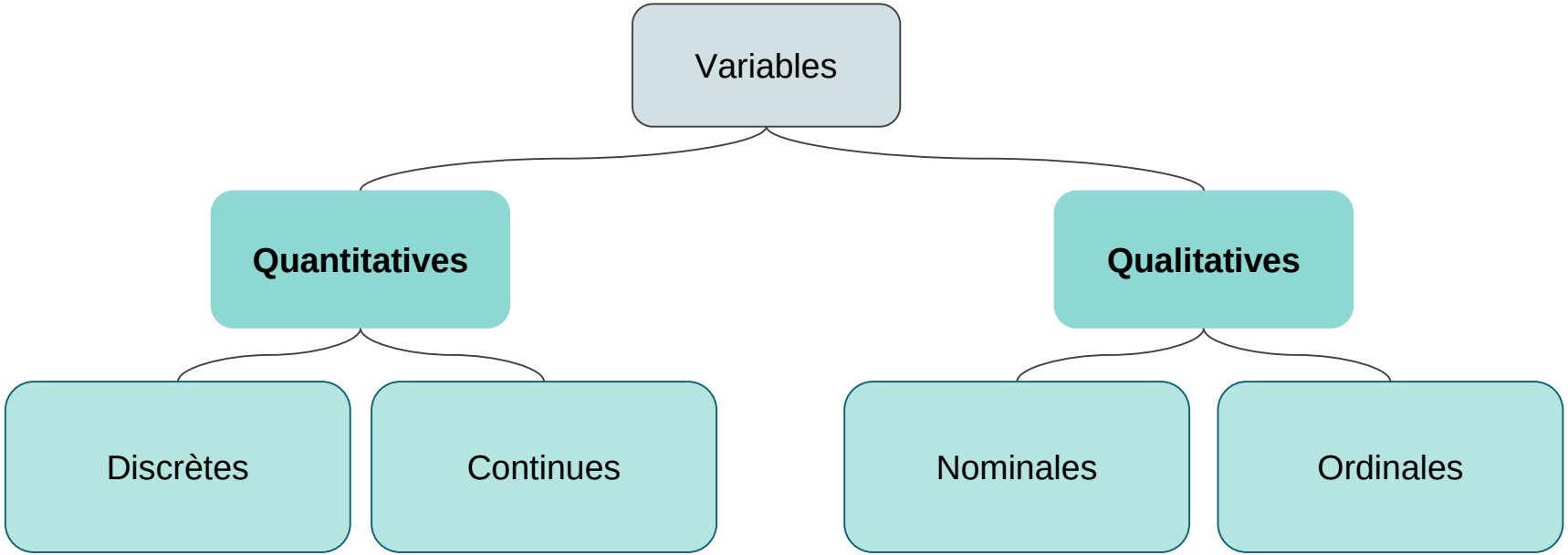


Variables aléatoires descriptives

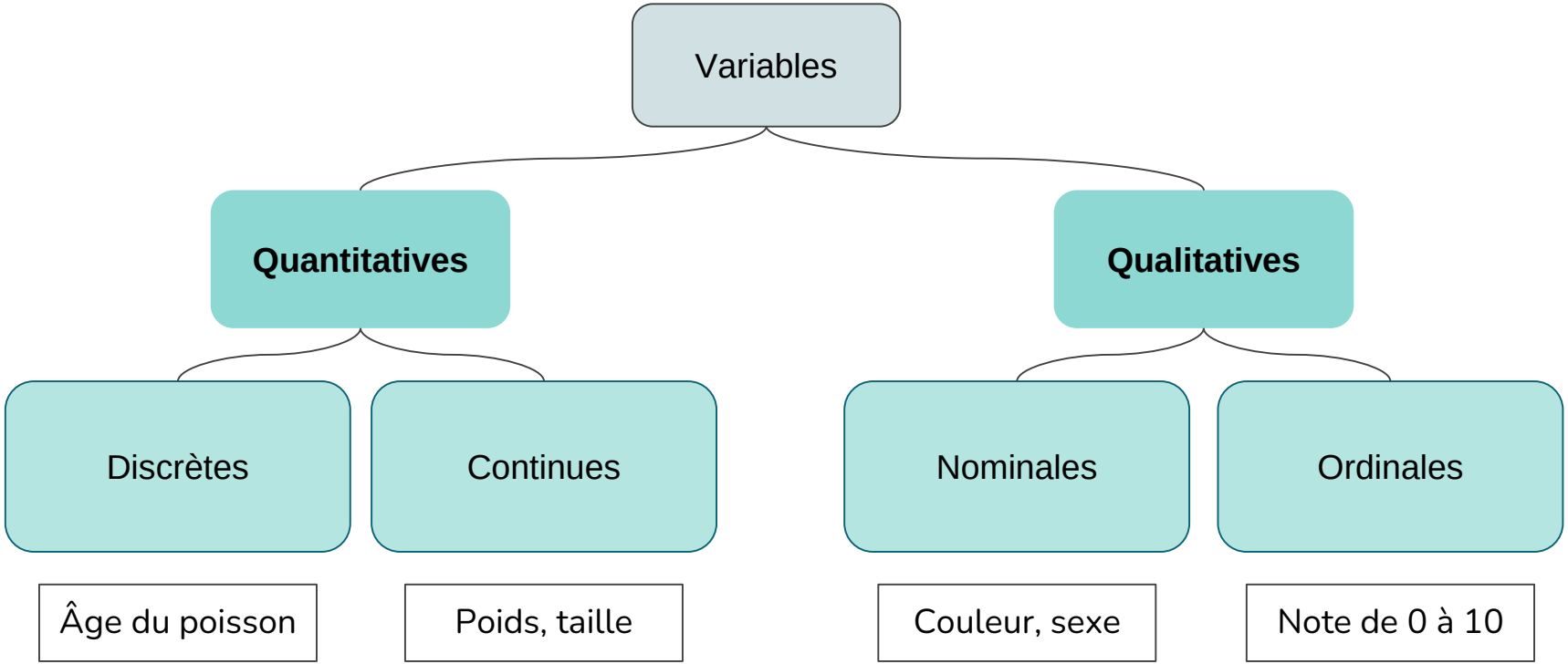
- Poids
- Taille
- Couleur
- Etc...

Classifier les variables

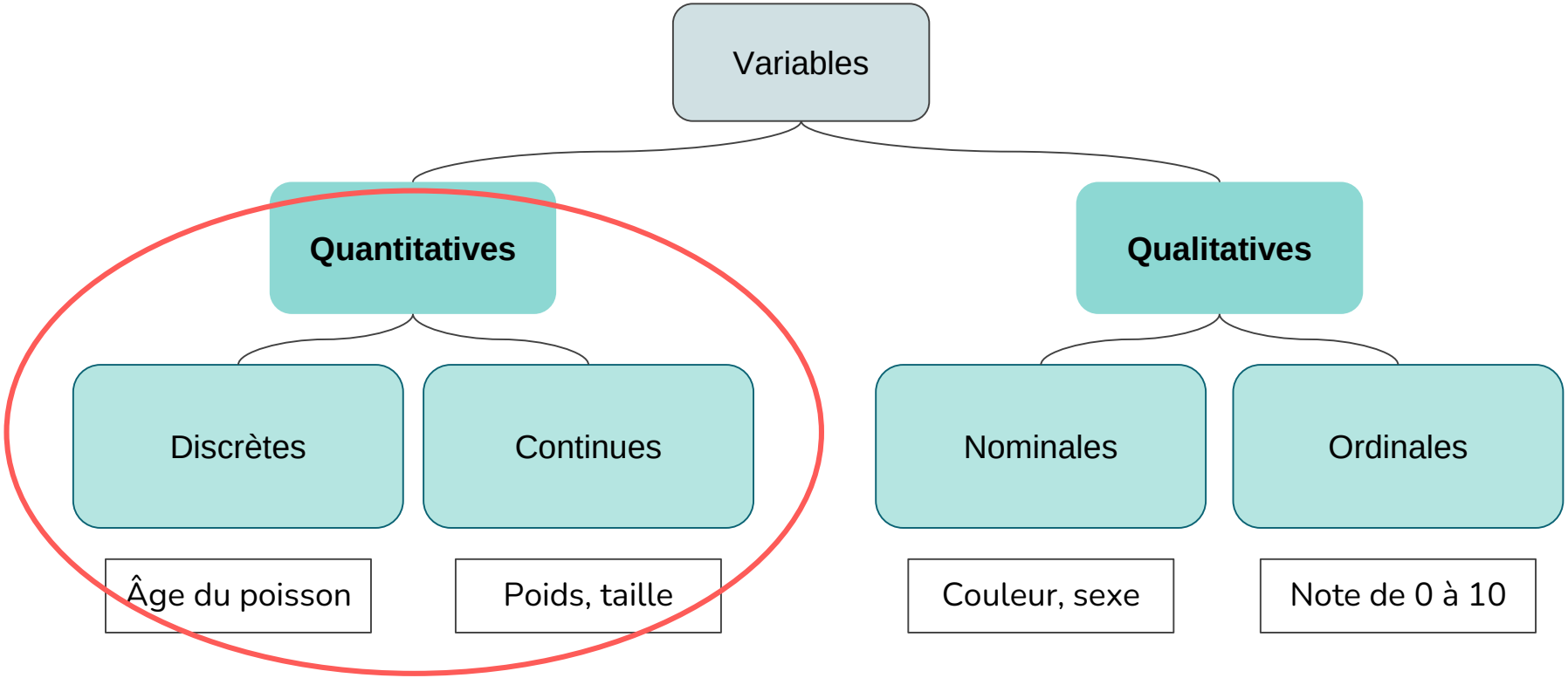
Variables aléatoires descriptives



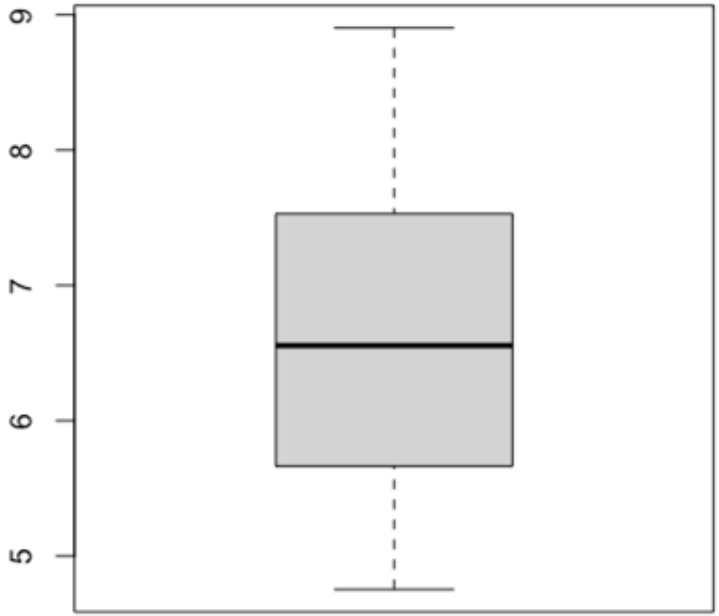
Variables aléatoires descriptives



Variables aléatoires descriptives

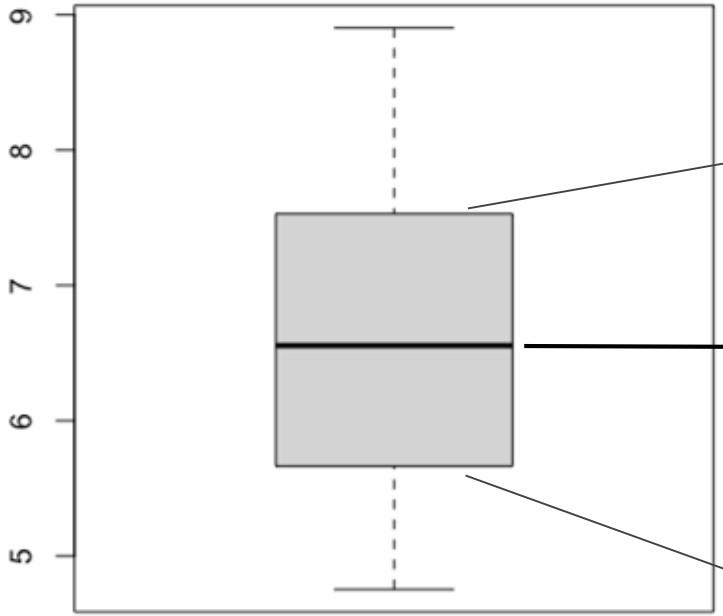


Représenter les variables : boxplot



Taille standard

Représenter les variables : boxplot



Q3 : 75% des données sont inférieures ou égales

Médiane : 50% des données sont inférieures ou égales

Q1 : 25% des données sont inférieures ou égales

Paramètres et lois

Paramètres d’échantillon

- Moyenne μ m
- Variance σ^2 s^2
- Proportion π p

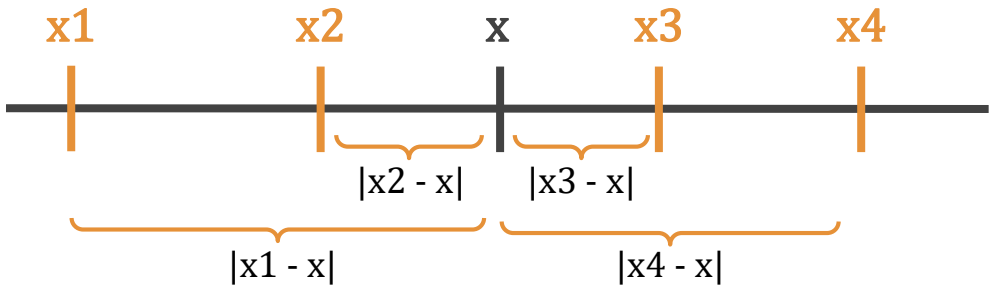
Paramètres d'échantillon

	Population	Échantillon
Moyenne	μ	m
Variance	σ^2	s^2
Proportion	π	p

Loi de probabilité

- Espérance $\mathbf{E(X)}$: la valeur de la variable étudiée la plus probable à observer (moyenne pondérée)
- Variance $\mathbf{V(X)}$: moyenne des carrés des écarts à la moyenne

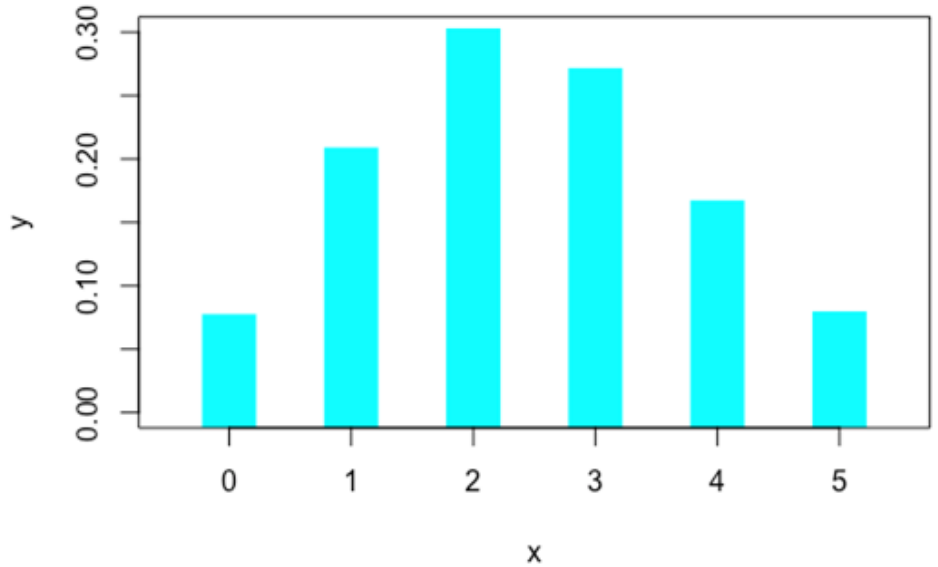
$$V = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$$



Loi de probabilité discrète

Variable aléatoire discrète : caractérisée par l'ensemble des valeurs que peut prendre cette variable et par les lois de probabilités de ces valeurs

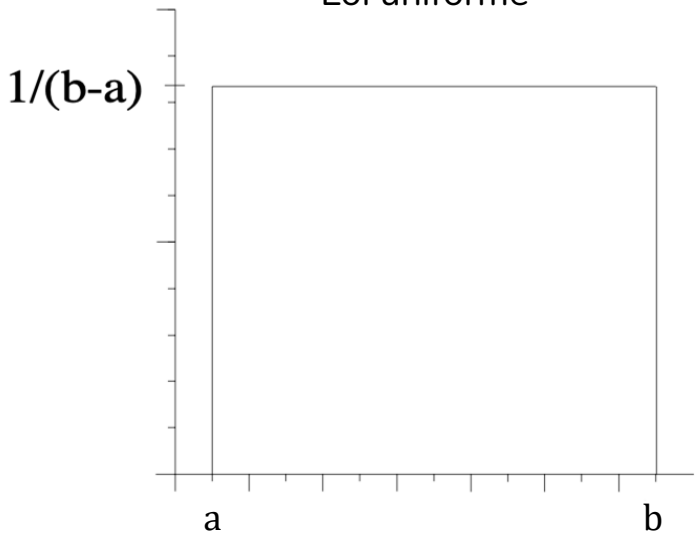
- $0 <= p(X=xi) <= 1$
- $\sum p(X=xi) = 1$



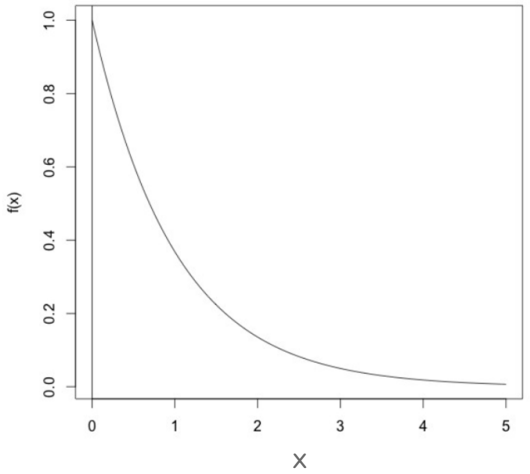
Loi de probabilité à densité de probabilité

Les variables aléatoires continues suivent une loi de densité de probabilité

Loi uniforme



Loi exponentielle

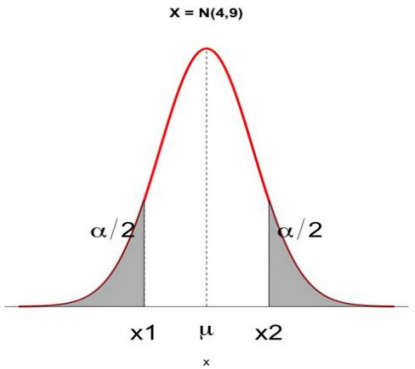


Loi normale

Densité de probabilité : $X \sim N(\mu, \sigma^2)$

Formule développée : $f(x) = \frac{1}{\sqrt{2\pi}\sigma} \mathrm{e}\left(-\frac{(x-p)^2}{2\sigma^2}\right)$

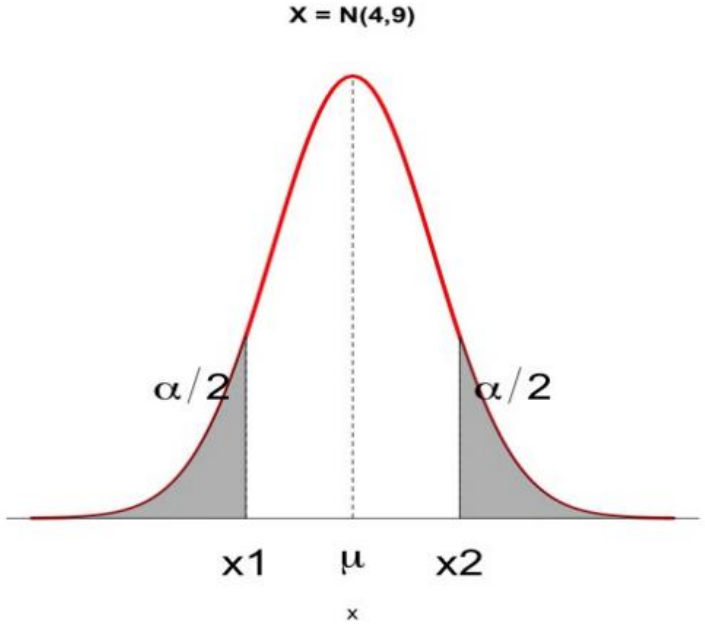
Valeurs caractéristiques : $E(X) = \mu$ et $V(X) = \sigma^2$



Loi normale

Densité de probabilité : $X \sim N(\mu, \sigma^2)$

$E(X) = \mu$ et $V(X) = \sigma^2$

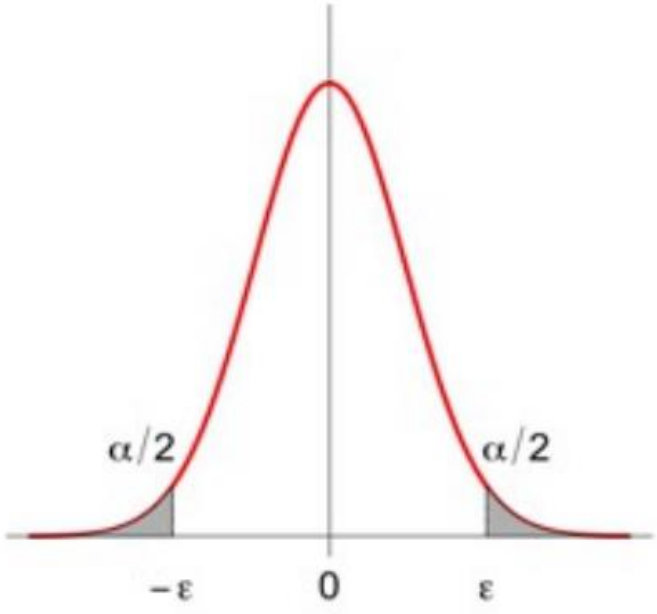


Loi normale centrée réduite

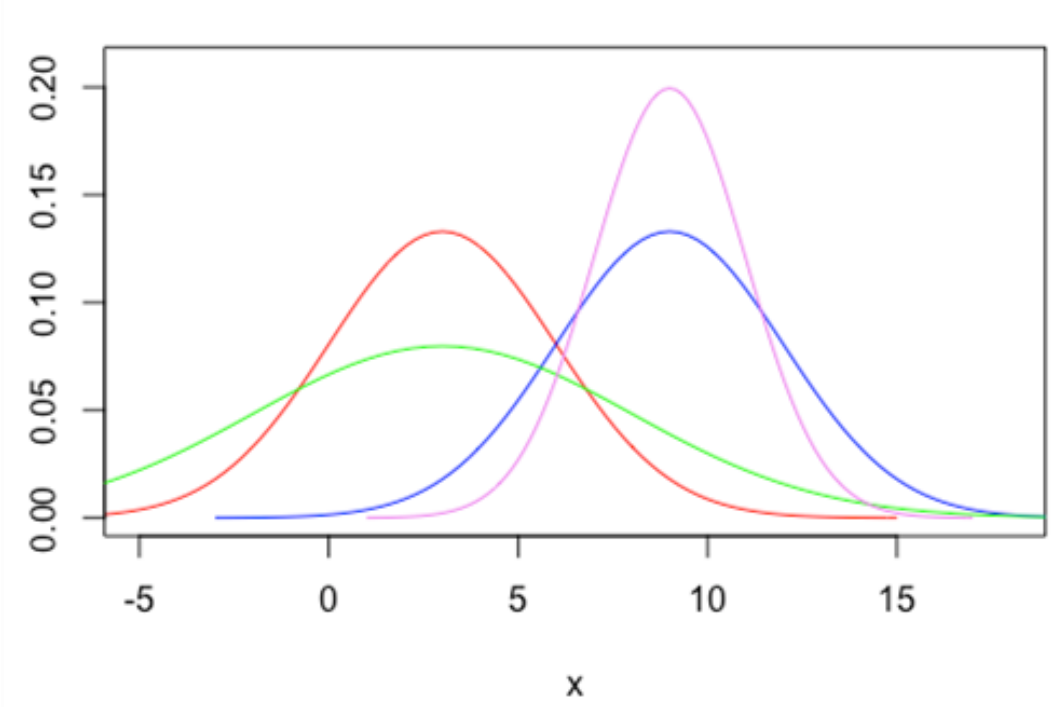
Densité de probabilité : $Z \sim N(1, 0)$

$E(Z) = 1$ et $V(Z) = 0$

Transformation : $Z = \frac{X - \mu}{\sigma}$



Identifier les lois normales



N (μ, σ)

- N (3, 3)
- N (9, 3)
- N (3, 5)
- N (9, 2)

Théorème central-limite

La somme d'un grand nombre de variables aléatoires indépendantes (≥ 30) de même loi suit une loi normale dont les paramètres connus sont :

$\mathrm{E}(X_i) = \mu$

$\mathrm{V}(X_i) = \sigma^2$

Théorie des tests

Préparation d’un test statistique

Définir :

- Population
- Échantillon
- Unité Statistique
- Expérience aléatoire : comment obtenir la variable aléatoire
- Variable aléatoire : variable étudiée dans le test

Test statistique

Un test statistique est une **règle de décision** permettant de trancher entre **deux hypothèses** faite sur une population, à partir du résultat d'une expérience sur un **échantillon**.

Test statistique

Hypothèses d’un test statistique

$H0$ / Hypothèse nulle : Hypothèse de non modification du paramètre

$H1$ / Hypothèse alternative: Hypothèse de modification du paramètre

Test statistique

Hypothèses d'un test statistique

$H0$ / Hypothèse nulle : Hypothèse de non modification du paramètre

$H1$ / Hypothèse alternative: Hypothèse de modification du paramètre

Règle de décision : rejeter $H0$ ou ne pas rejeter $H0$

Décision et risque d'erreur

		Réalité	
		H0	H1
Décision	Non rejet de H0	correct	β
	Rejet de H0	α	correct

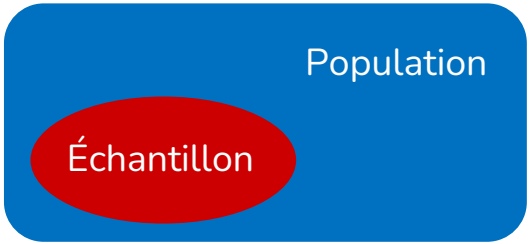
Décision et risque d'erreur

		Réalité	
		H0	H1
Décision	Non rejet de H0	correct	β
	Rejet de H0	α	correct

Conformité et comparaison

Conformité

Le paramètre estimé dans l'échantillon est-il conforme à celui de la population ?



$$H_0 : \theta = \theta_0$$
$$H_1 : \theta \neq \theta_0$$

Comparaison

Les paramètres estimés de deux échantillons sont-ils égaux ?

$$H_0 : \theta_1 = \theta_2$$
$$H_1 : \theta_1 \neq \theta_2$$

Exemples de test

Test de conformité pour une variable qualitative ou quantitative discrète (test de proportion)

- On étudie la variable X de paramètre π
- Le paramètre π_0 est la valeur théorique dans la population
- Sur un échantillon de taille n, on calcule p_{obs}

Exemples de test

Hypothèses :

- $H_0 : \pi = \pi_0$
- $H_1 : \pi \neq \pi_0$

Conditions d'application :

- $n\pi_0 > 5$
- $n(1 - \pi_0) > 5$

Exemples de test

Règle de décision :

- $\varepsilon_{cal} = \frac{|p_{obs} - \pi_0|}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}}$

Si $\varepsilon_{cal} > \varepsilon_{\alpha}$, le test est significatif à α% (H_0 est rejetée), sinon le test est non significatif

Exemples de test

Test de comparaison de moyenne pour une variable quantitative continue

- On étudie la variable X de paramètres μ et σ^2
- Échantillon 1 de taille n_1, de moyenne m_1 et de variance s_1^2
- Échantillon 2 de taille n_2, de moyenne m_2 et de variance s_2^2

Exemples de test

Hypothèses :

- $H_0: \mu_1 = \mu_2$
- $H_1: \mu_1 \neq \mu_2$

Conditions d'application :

- $n_1 \geq 30$ et $n_2 \geq 30$
- Sinon X doit suivre une loi normale et les variances doivent être identiques

Exemples de test

Règle de décision :

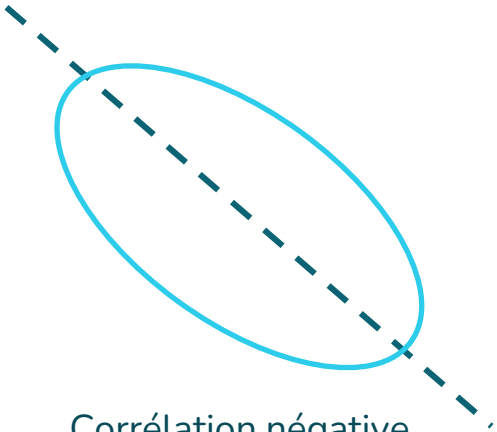
- $\varepsilon_{cal} = \frac{|m_1 - m_2|}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$

Si $\varepsilon_{cal} > \varepsilon_{\alpha}$, le test est significatif à α% (H_0 est rejetée), sinon le test est non significatif

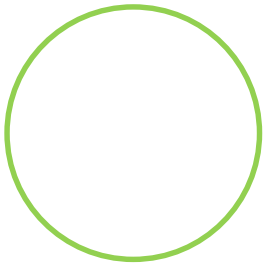
Corrélation

Représentation dans l'espace

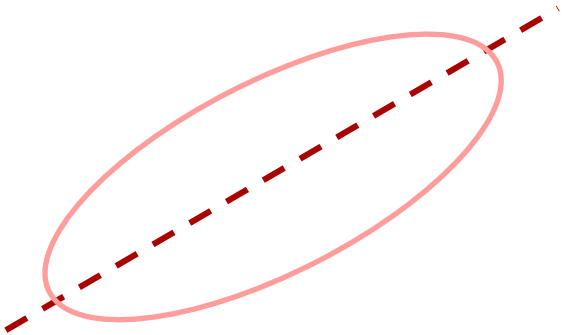
Soit deux variables quantitatives X et Y dont on calcule les valeurs x_i et y_i formant un nuage de points :



Corrélation négative



Pas de corrélation



Corrélation positive

Coefficient de corrélation

On peut estimer le coefficient de corrélation de deux variables aléatoires X et Y par la formule suivante :

- $\hat{\rho} = r = \frac{s_{xy}}{s_x s_y}$

Avec s_{xy} la covariance entre X et Y:

- $s_{xy} = \frac{1}{(n-1)}\left[\sum x_i y_i - \frac{\Sigma x_i \Sigma y_i}{n}\right]$

Test du coefficient de corrélation

Hypothèses :

- $H_0 : \rho = 0$
- $H_1 : \rho \neq 0$

Conditions de validité :

- X et Y sont distribués selon une loi normale

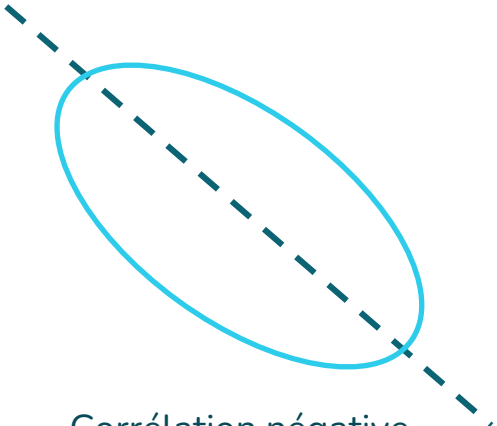
Test du coefficient de corrélation

Règle de décision :

- $t_{cal} = \frac{|r|\sqrt{n-2}}{\sqrt{1-r^2}}$

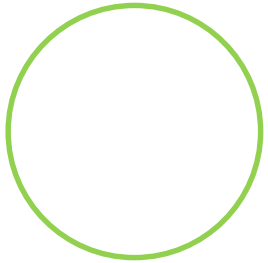
Si $t_{cal} > t_{\alpha, n-2}$, le test est significatif à α% (H_0 est rejetée), sinon le test est non significatif

Coefficient de corrélation



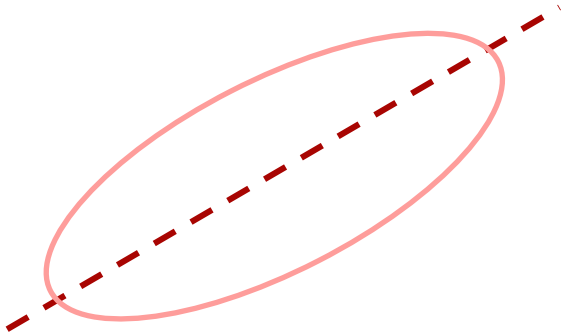
Corrélation négative

$\rho < 0$



Pas de corrélation

$\rho = 0$



Corrélation positive

$\rho > 0$