



UNIVERSITÉ
DE NAMUR

Intelligence des données Méthodes Statistiques

Année académique

2024-2025



Sommaire

1. Présentation du Professeur
2. Introduction à Python
3. Introduction à la Statistique
4. Statistique Descriptive
5. Statistique Inférentielle

Calendrier provisoire

1. Présentation + Python + Anaconda + Jupyter + organisation générale – 17/09
2. Paramètres statistiques – 23/9
=> Slide 56
2. Distributions statistiques – 30/09
3. Régressions linéaires – 07/10
4. Généralisation & Régressions non linéaires – 14/10
5. Stat Inférentielle & Echantillonnage – 21/10
7. Interro (1H) + Echantillonnage (fin) – 23/10
8. Correction Interro - Estimation (début) 06/11
9. Estimation (fin) – Tests d'hypothèses – 18/11
10. Tests d'hypothèses – 25/11

Laurent PHILIPPE

Qui suis-je?

- Physicien de formation, informaticien de métier et musicien de cœur
- Reconnu Chargé de Recherche de la Communauté Française de Belgique (1996)
- Docteur en Physique de l'Université de Namur (1995):
 - Spécialisation: physique de l'état solide (semi-conducteurs et supraconducteurs)
 - Beaucoup de contacts avec l'IMEC (électronique de pointe à Leuven)
 - Introduction du C dans le bagage des (autres) physiciens (DEA)
 - Introduction de UNIX dans le bagage des (autres) physiciens (Ecole de doctorat)
 - DEA en analyse numérique appliquée aux techniques des éléments finis (Paris-Jussieu)
 - Chargé des TPs du cours d'Analyse Numérique des Problèmes de la Physique
 - Première implémentation de calcul distribué à l'Université de Namur (1997)

Laurent PHILIPPE

Qui suis-je? (suite 1)

- Post-doc à l'Université du Mans en 1996
- Administrateur système à l'Université de Mons ([Materia Nova](#)) 1997-1998
- Projet First-Entreprise chez [Capflow](#) 1999-2006
 - Visioflex® Manager – Coordinateur de projets
- Consultance en informatique bancaire 2001- 2019 ([Sopra](#), [Business & Decision](#), ...)
 - Essentiellement chez ING et AXA
 - D'abord pour mes connaissances UNIX (Solaris)
 - Ensuite pour ma rigueur et mes capacités d'analyse
 - Maintenant pour ma connaissance de presque tous les rouages du monde de l'IT bancaire...
- Consultance en data gouvernance, modelling, architecture, ...
 - Migration des données du précompte immobilier (2020 [Aprico Consultants](#))
 - Digitalisation de la gestion du tourisme en Wallonie (2022-2023 [Aprico Consultants](#))
 - STIB ([Wavenet Belgium](#))

Laurent PHILIPPE

Qui suis-je? (suite 2)

- Grand intérêt pour et certification dans le domaine des Big Data et du Machine Learning:
 - Formation au et certification par le MIT ☺
 - Tackling the Challenges of Big Data (2015)
 - Social Physics (2016)
 - Paradox and Infinity (2019)
 - Projet de certification en Quantum Computing (quand je trouverai le temps)
 - Réseau de Bayes (Bayesia – 2020)
 - Pyimagesearch – OpenCV guru (2020 – ongoing)

Laurent PHILIPPE

Qui suis-je? (suite 3)

- Fig Sci & Co:
 - Mon propre nom de marque
 - Fig Sci is not Sci-Fi
 - Méthodologie d'audit des systèmes intelligents
- Collaboration étroite avec Agilytic – Bruxelles Be-Central:
 - J'ai croisé un des fondateurs chez ING
 - Embarqué dans différents projets dont:
 - Daoust Interim (micro-service pour enrichir automatiquement les CVs des postulants)
 - Projets de maintenance prédictive avec Volvo Truck, SNCF...
 - Intégration de l'Ardenne Prévoyante à l'infrastructure AXA Insurance

Laurent PHILIPPE

Qui suis-je? (suite et fin)

- Thèse en Histoire des Sciences et Histoire de l'Art:
 - Naissance de la Perspective à Florence, lieu de naissance de Galilée
- Flûtiste dans de nombreux groupes:
 - Flûte solo à l'Orchestre Symphonique de l'Université de Namur
 - Animateur d'un quatuor de flûtes
 - Organisateur de grands évènements:
 - Concert du 175^{ème} anniversaire de l'UNamur à l'église Saint-Loup (1997)
 - Concert Rotary au Théâtre de Namur et à la Collégiale de Dinant (2019)

Laurent PHILIPPE

Où me trouver ? Comment me joindre ?

Jusqu'à nouvel ordre:

- En présentiel au 822 si je ne donne pas cours et, généralement, au 516 puisque tous mes cours de Master se donnent là (attention que je donne aussi cours en Bachelier)
- laurent.philippe@henallux.be tout le temps mais avec une réactivité d'un jour... je ne suis pas pendu à mon mail... surtout quand je donne cours.

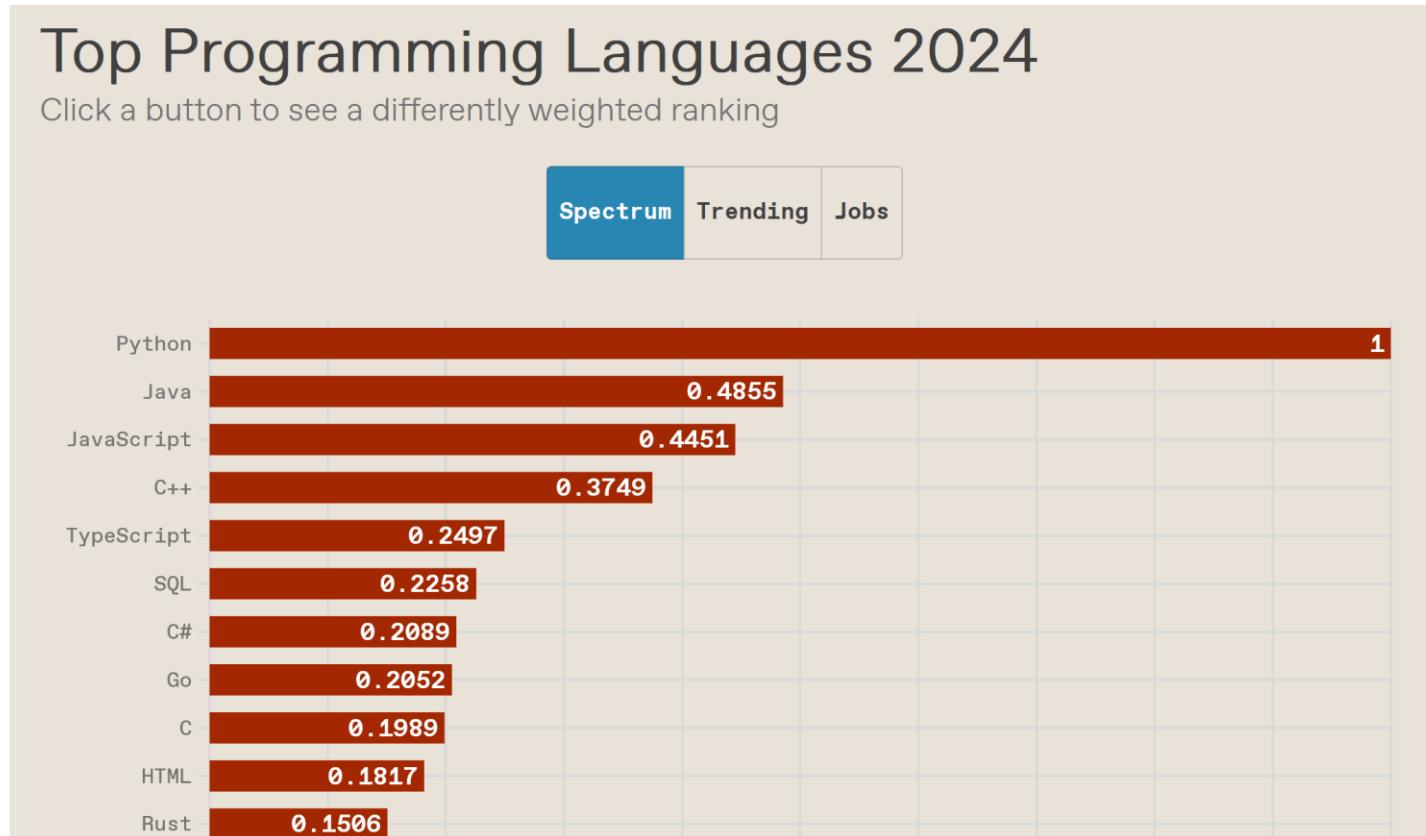
Introduction à Python

Pourquoi Python dans un cours de statistiques ?

- Parce que Python est certainement plus répandu que les langages spécialisés comme R (2017-2018) ou même SAS (retiré de la grille en 2020)
- Bon support de cours online pour la syntaxe, plutôt simple et assez épurée par rapport aux autres langages beaucoup plus verbeux (C++ et Java)
- Parce que vos machines sont souvent suffisantes pour se permettre un environnement graphique digne de ce nom (Anaconda et Jupyter)
- Bonne disponibilité de librairies spécialisées dans les statistiques (scikit,...) et relative facilité pour faire des graphiques (matplotlib)

Introduction à Python

[Top Programming Languages IEEE Spectrum](#)



Introduction à Python



Rapide historique

- Crée par Guido Van Rossum à la fin des années 1980 au Centrum voor Wiskunde en Informatica à Amsterdam
 - Nom dérivé de la série télévisée des Monty Python's Flying Circus ☺
 - Première version publique 0.9.0 en février 1991
 - 2001: création de la Python Software Foundation sur base de Python 2.1
 - 2008: sortie de Python 3.0 après un grand nettoyage de la bibliothèque standard, simplification de la modélisation objet et renonciation à la compatibilité ascendante!
 - Dernière version stable pour la « branche 3 »: 3.12.6 (06 septembre 2024)

Introduction à Python

Langage interprété



- N'est solide que si vous avez un ensemble de tests assez exhaustif!!!
- Rapide pour « (a)voir » quelque chose – idéal pour les prototypes
- Mais très lent à révéler toutes ses surprises...

Introduction à Python

Syntaxe très légère



- Les blocs d'instruction sont définis par l'indentation

Fonction factorielle en C

```
int factorielle(int n) {
    if (n<2) {
        return 1;
    } else {
        return n * factorielle(n-1);
    }
}
```

Fonction factorielle en Python

```
def factorielle(n):
    if n<2:
        return 1
    else:
        return n * factorielle(n-1)
```

Introduction à Python



Pour les autres principales caractéristiques:

- [https://fr.wikipedia.org/wiki/Python_\(langage\)](https://fr.wikipedia.org/wiki/Python_(langage))
- [Python Beginner Cheat Sheet](#)
- Chapitres 2 et 3 de *An Introduction to Statistics in Python* (voir références)

Environnements de développement intégré:

- Anaconda et Jupyter Notebook (support assuré dans le cours)
- Mais aussi JupyterLab ou Visual Studio (support non assuré dans le cours)

Introduction à Python

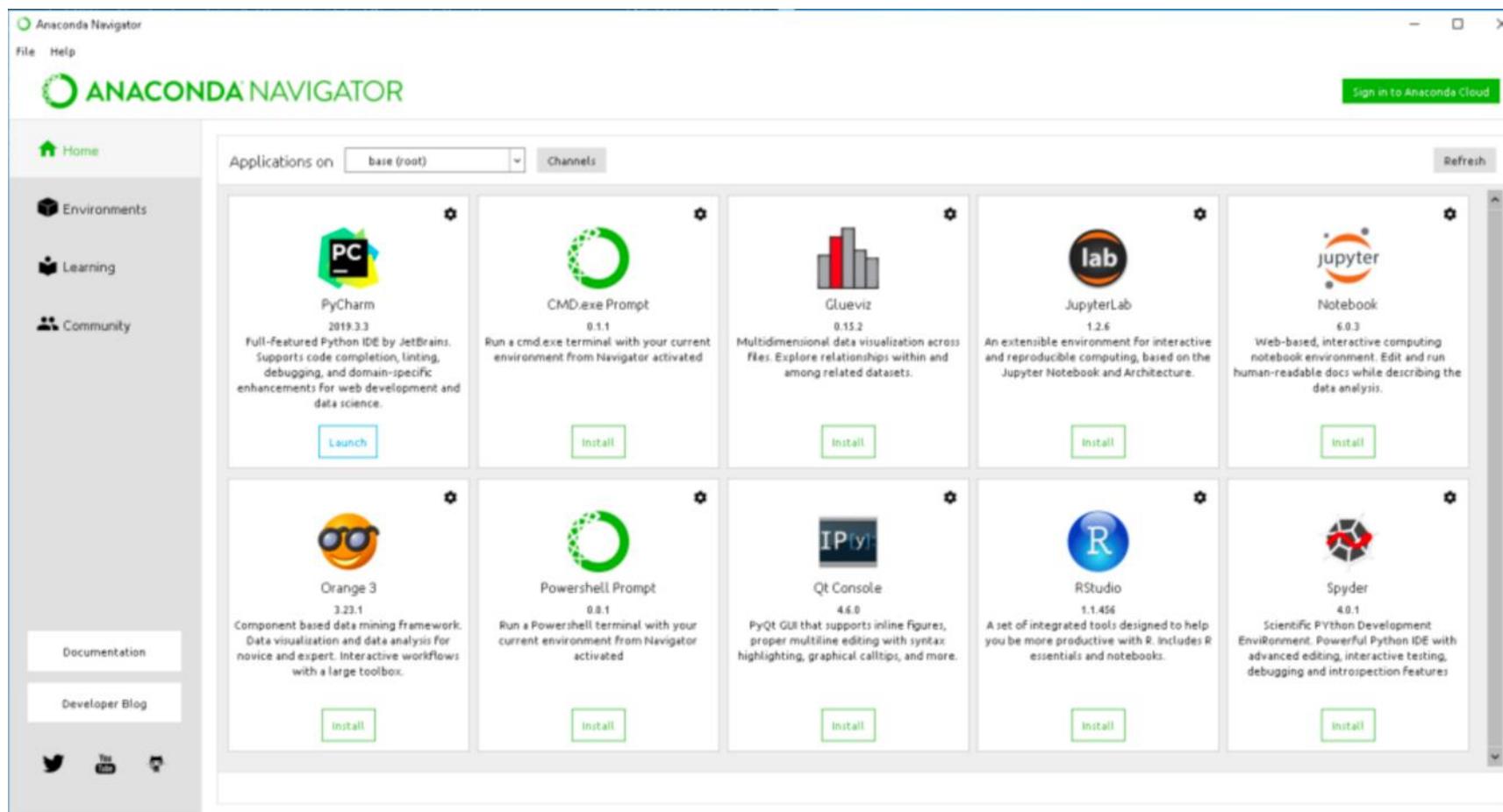
Exercice 0



- Installation d'[Anaconda](#) sur votre PC (Python 3.11.7)
- Création de votre premier Notebook Jupyter contenant une implémentation de la fonction factorielle + mise en œuvre concrète de cette fonction
- Essayez de trouver la limite supérieure de l'argument avant que Python ne se plaigne!

Introduction à Python

IDE Anaconda – Jupyter notebook



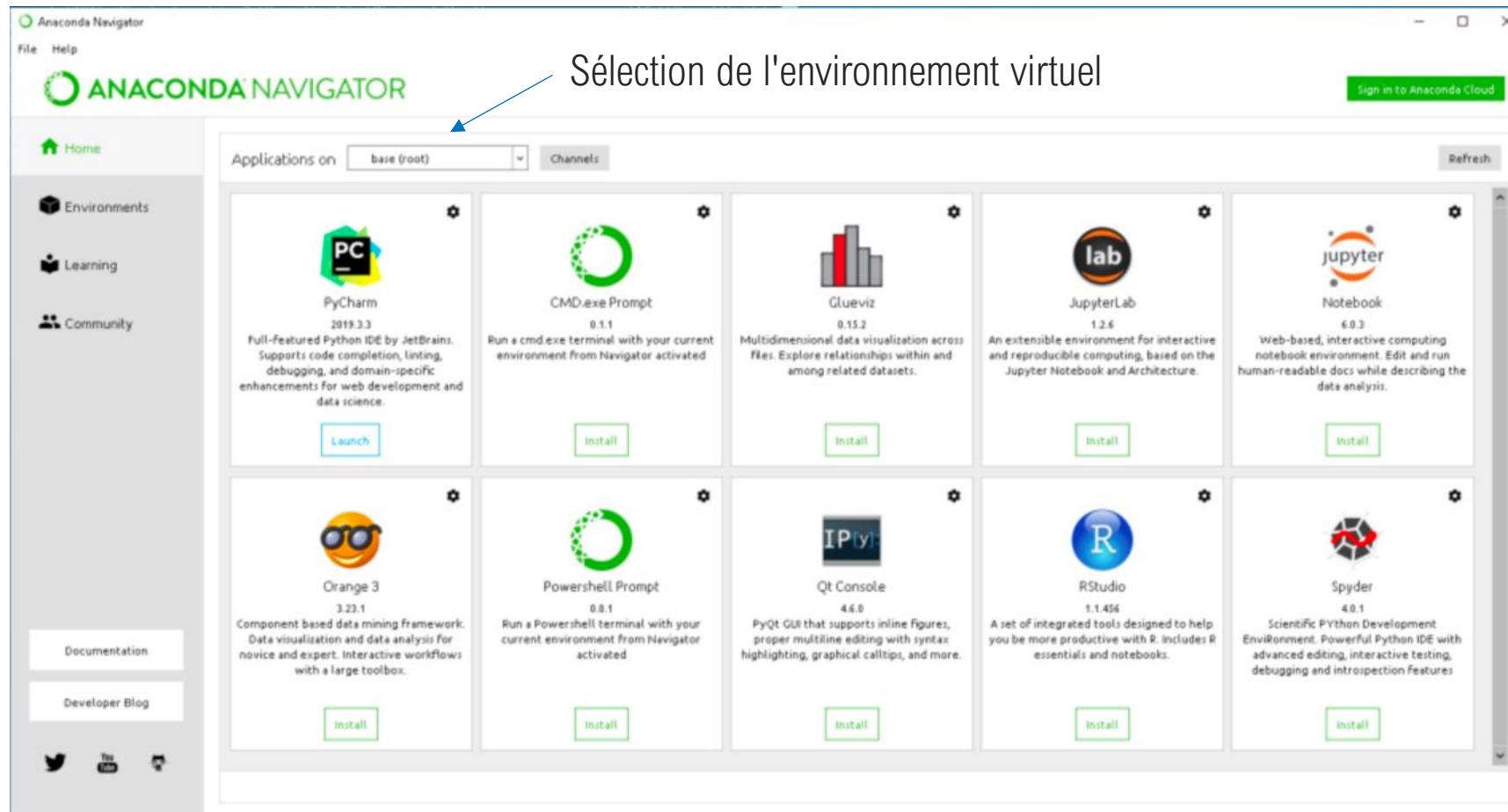
Introduction à Python

Environnements virtuels – Cloisonnement par projet des librairies installées

Gestion des environnements virtuels



Sélection de l'environnement virtuel



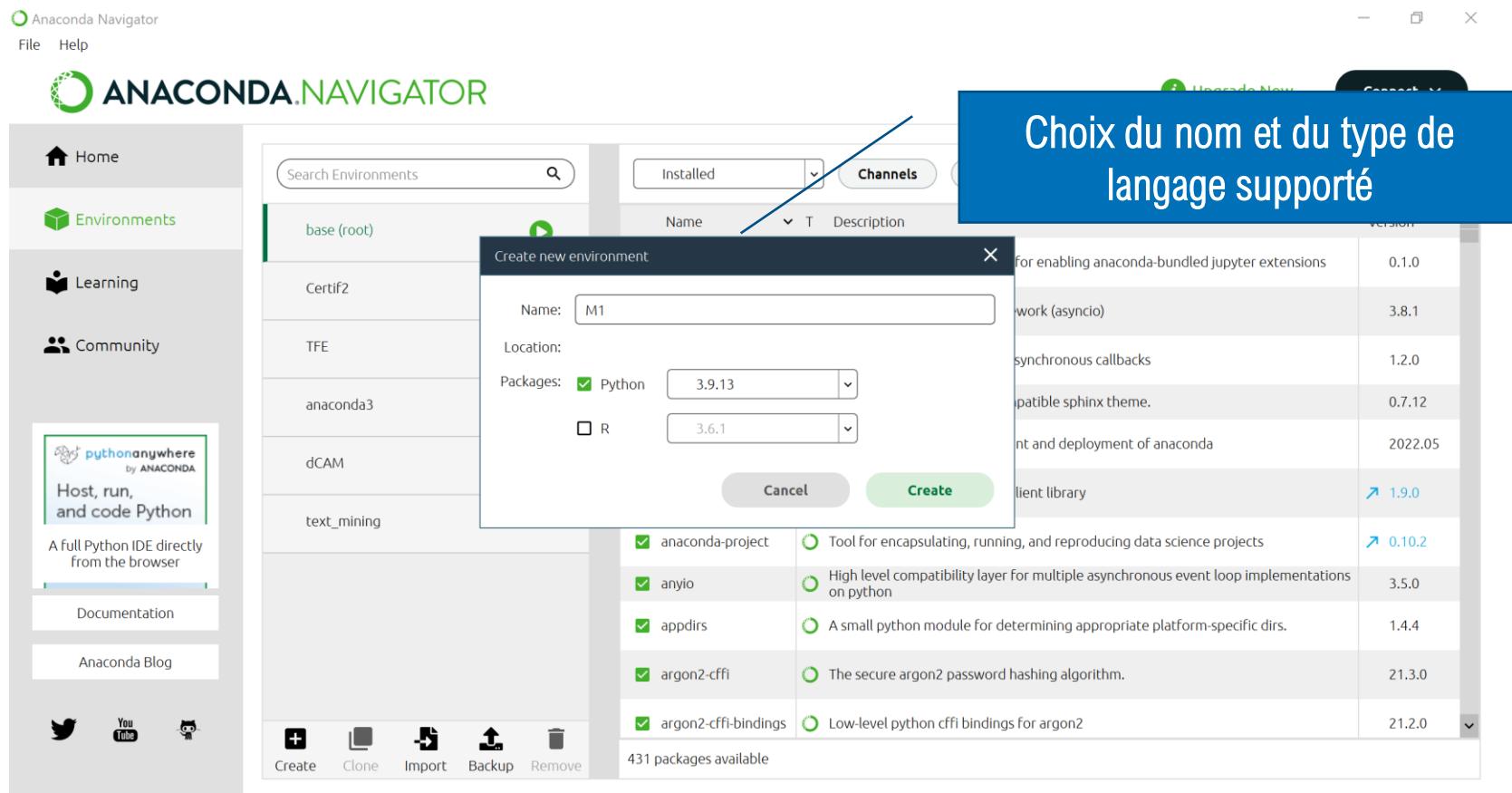
Introduction à Python

Environnements virtuels – Cloisonnement par projet des librairies installées



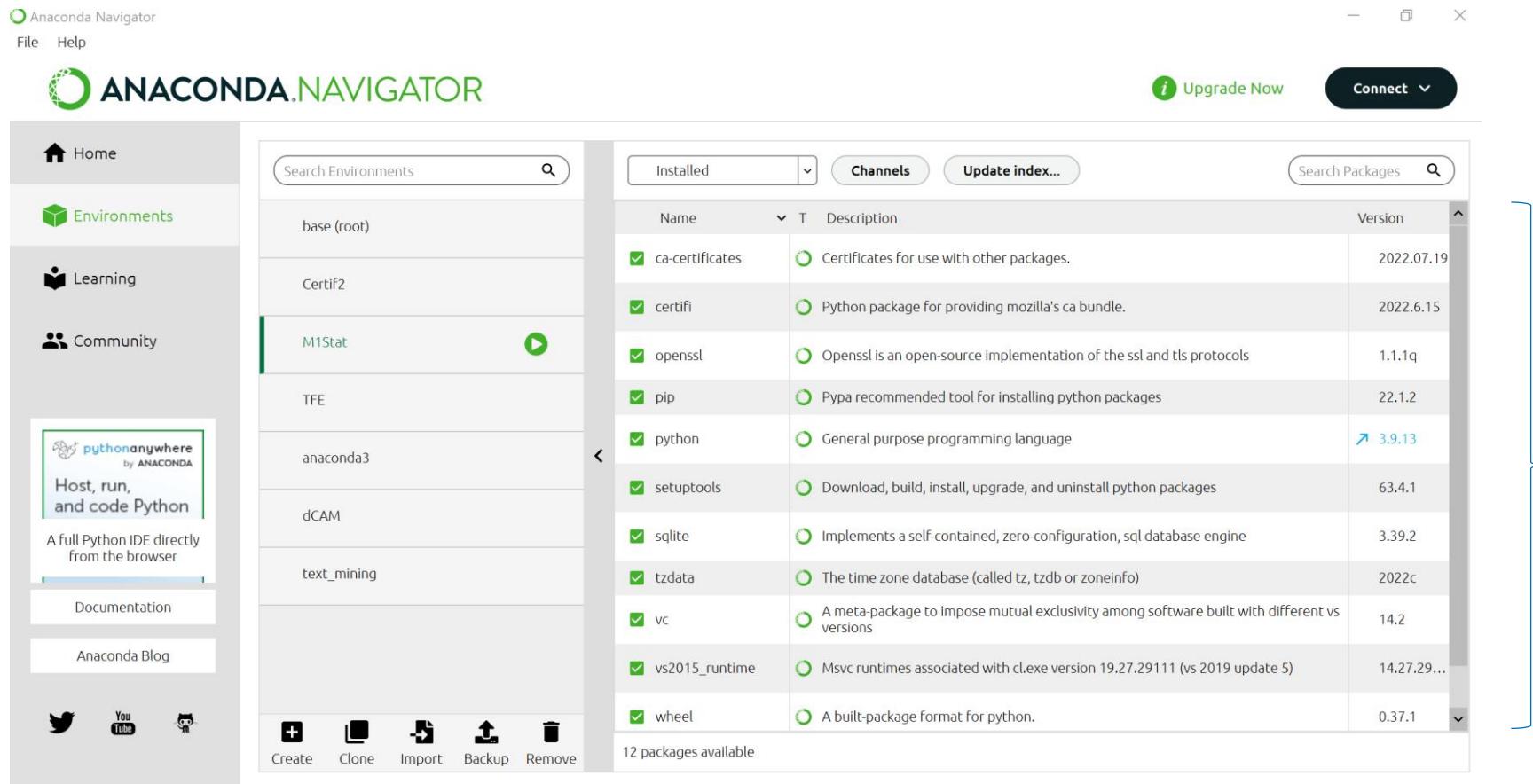
Introduction à Python

Environnements virtuels – Cloisonnement par projet des librairies installées



Introduction à Python

Environnements virtuels – Cloisonnement par projet des librairies installées



The screenshot shows the Anaconda Navigator interface. On the left, there's a sidebar with links to Home, Environments, Learning, Community, PythonAnywhere (with a 'Host, run, and code Python' button), Documentation, and Anaconda Blog. Below these are social media icons for Twitter, YouTube, and GitHub. The main area displays a list of installed packages in the 'base (root)' environment. The table has columns for Name, Description, and Version. Packages listed include ca-certificates, certifi, openssl, pip, python, setuptools, sqlite, tzdata, vc, vs2015_runtime, and wheel. A blue bracket on the right side of the table groups the first four packages and spans to the text 'Librairies de base toujours installées par défaut'.

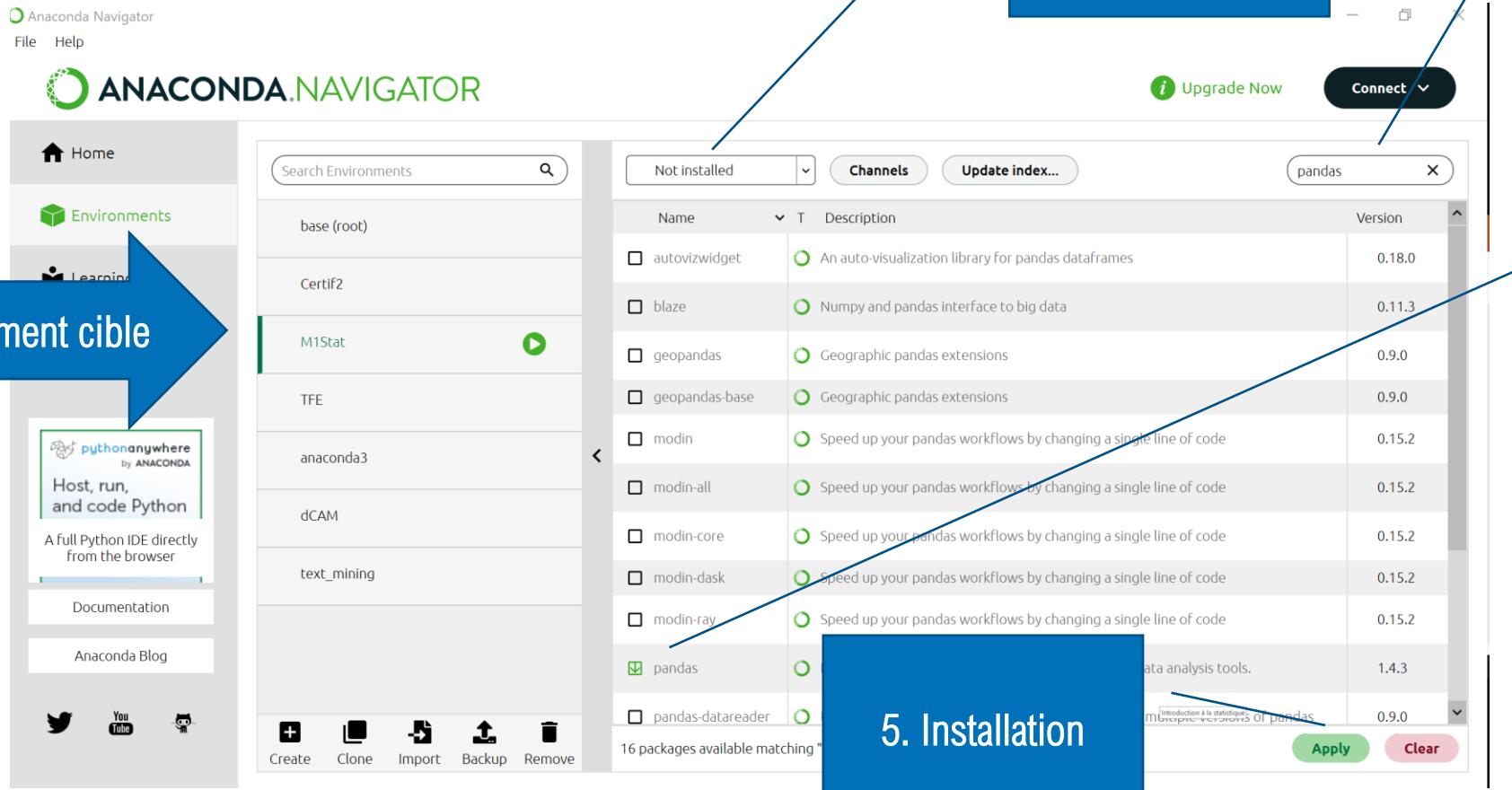
Name	Description	Version
ca-certificates	Certificates for use with other packages.	2022.07.19
certifi	Python package for providing mozilla's ca bundle.	2022.6.15
openssl	OpenSSL is an open-source implementation of the SSL and TLS protocols	1.1.1q
pip	Pypa recommended tool for installing python packages	22.1.2
python	General purpose programming language	3.9.13
setuptools	Download, build, install, upgrade, and uninstall python packages	63.4.1
sqlite	Implements a self-contained, zero-configuration, SQL database engine	3.39.2
tzdata	The time zone database (called tz, tzdb or zoneinfo)	2022c
vc	A meta-package to impose mutual exclusivity among software built with different vs versions	14.2
vs2015_runtime	Msvc runtimes associated with cl.exe version 19.27.29111 (vs 2019 update 5)	14.27.29...
wheel	A built-package format for python.	0.37.1

12 packages available

Librairies de base toujours installées par défaut

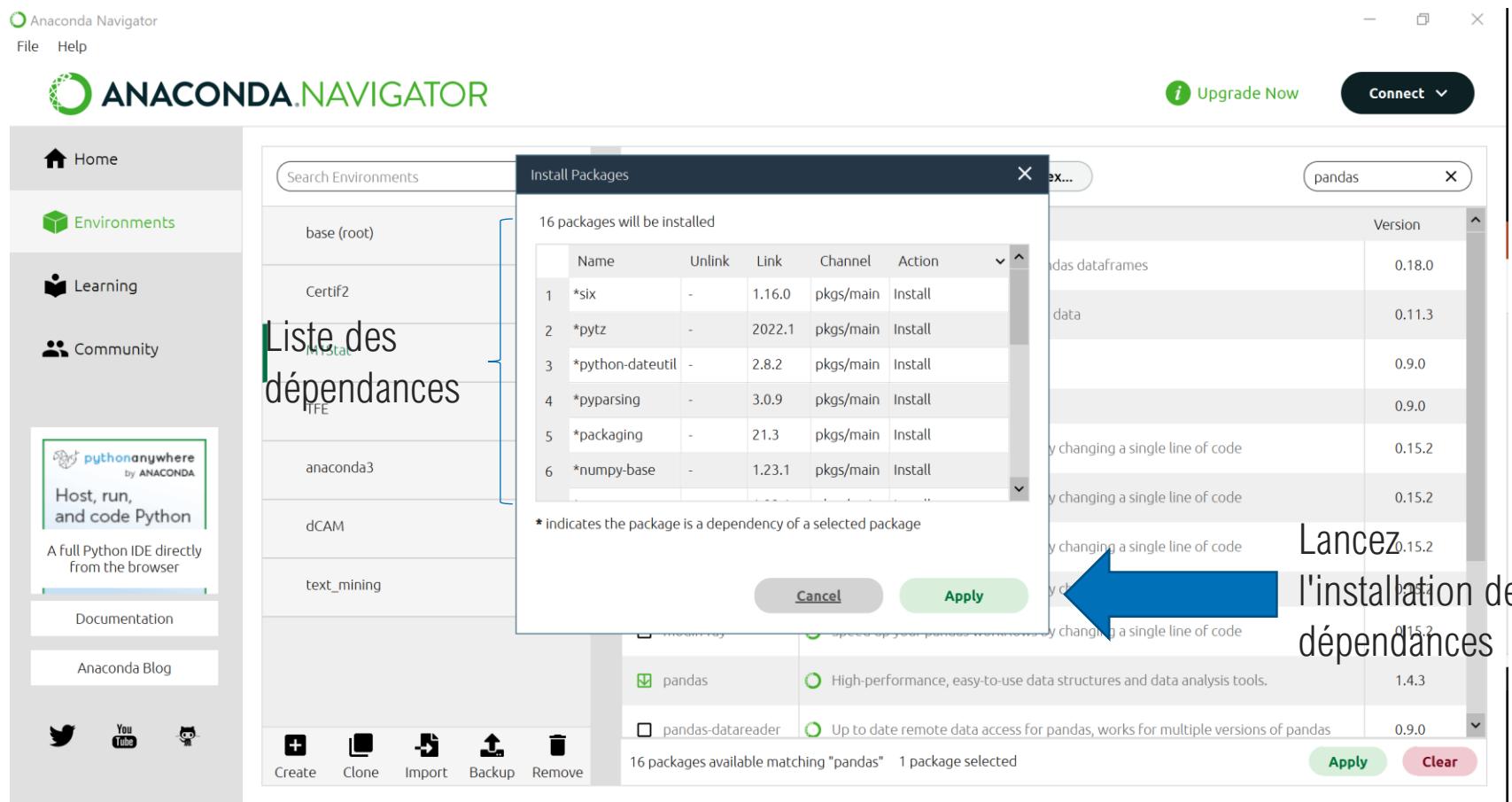
Introduction à Python

Environnements virtuels – Installation de Pandas



Introduction à Python

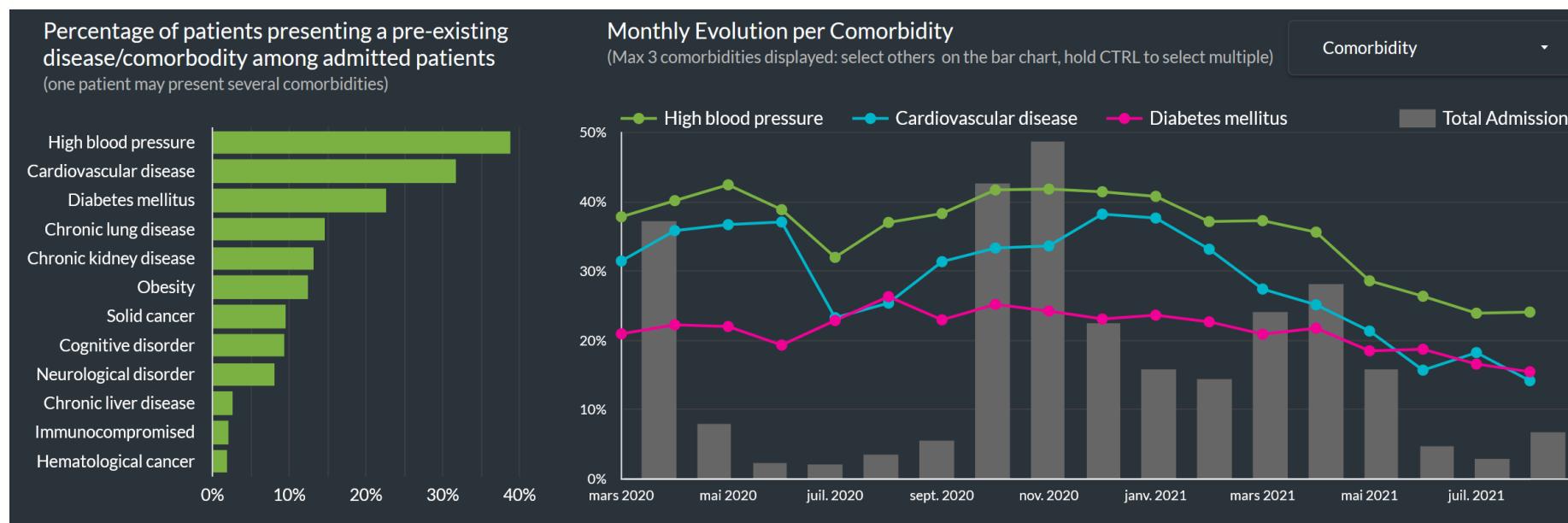
Environnements virtuels – Installation de Pandas



Introduction à la statistique

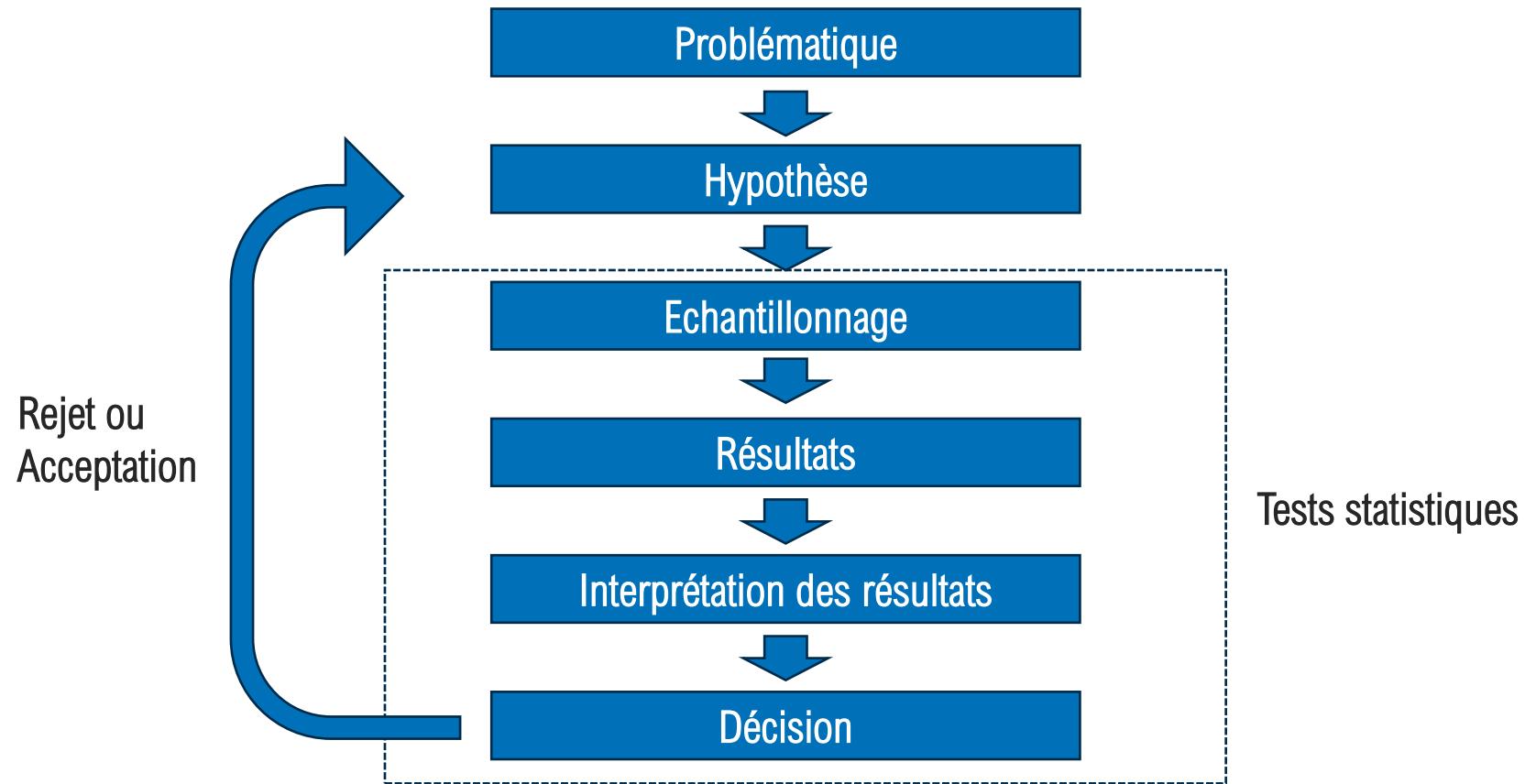
But d'une étude statistique

- Se faire une idée assez juste des variations d'une variable dans une population.



Introduction à la statistique

Statistique et démarche scientifique



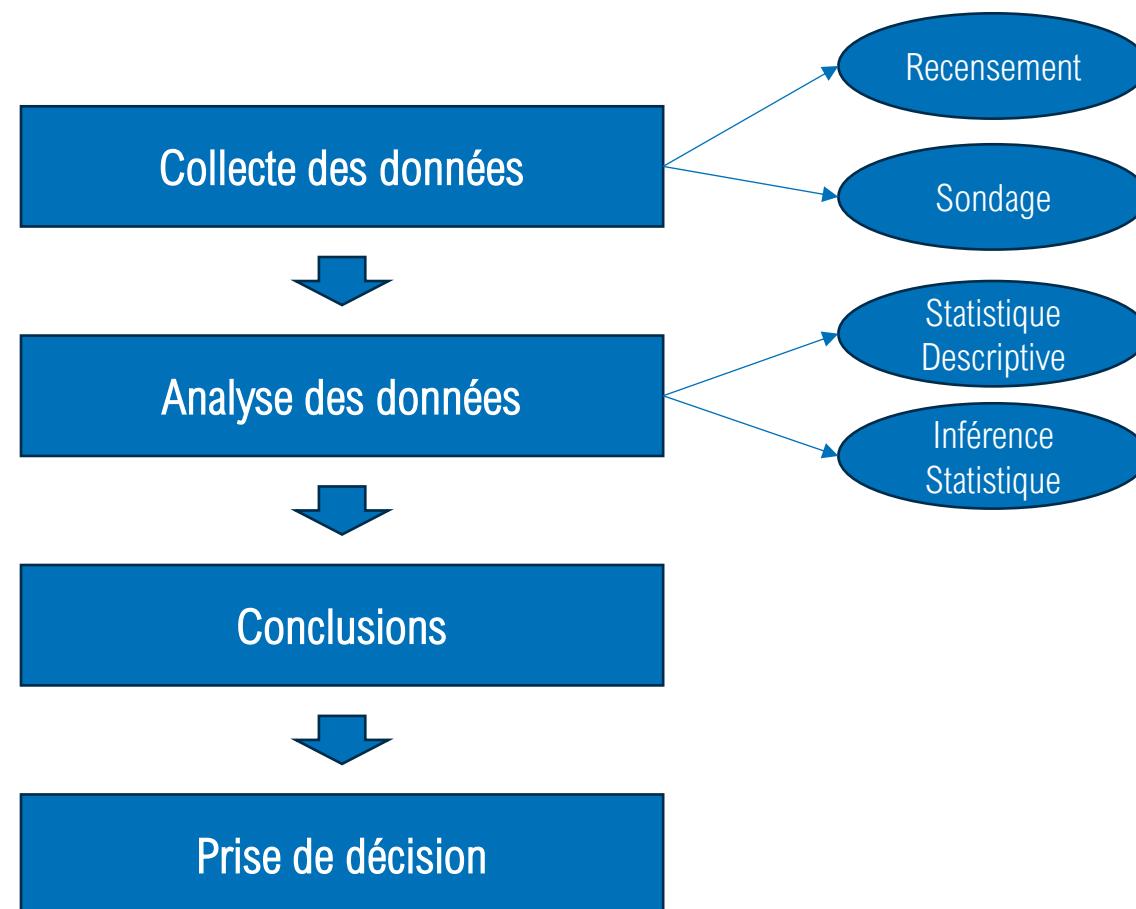


Introduction à la statistique

Méthodologie générale et vocabulaire de base

Introduction à la statistique

Méthodologie générale



Introduction à la statistique

Définitions de base: population

- Population

Ensemble des individus (ou unités statistiques) pour lequel on considère une ou plusieurs caractéristiques

Ex. L'ensemble des voitures immatriculées en 2016



- Taille de la population

Le nombre d'individus constituant la population



Notation : N

Introduction à la statistique

Définitions de base: individu

- Individu ou unité statistique:

Une unité distincte chez laquelle on peut observer une ou plusieurs caractéristiques

Ex. Une voiture immatriculée en 2016, un prélèvement de sol fait à Dijon



Introduction à la statistique

Définitions de base: variables statistiques (1)

- Caractéristique susceptible de variations observables.

Notation : X, Y, W, \dots (caractères)

Ex. Teneur en Cd des sols, leur densité apparente..., couleur des voitures, leur puissance

- Valeurs: les mesures distinctes d'une caractéristique donnée.

Notation : x_1, x_2, \dots (modalités)

Ex. Teneur en Cd du sol prélevé à Dijon, sa densité apparente..., couleur d'une voiture immatriculée en 2016, sa puissance,...

Introduction à la statistique

Définitions de base: variables statistiques (2)

- o Valeurs possibles:

Tous les résultats possibles **a priori** si on fait une observation d'une variable

- o Valeur observée:

Résultat **a posteriori** d'une observation d'une variable

Introduction à la statistique

Définitions de base: échantillon

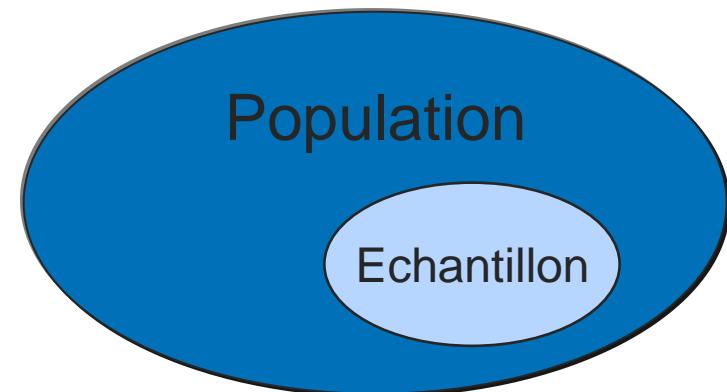
- Echantillon (sample):

Sous-groupe d'une population donnée.

Taille notée : n

Ex. 20 sols viticoles prélevés autour de Beaune,

20 voitures passant devant le 800...



Introduction à la statistique

Réflexion pratique : nécessité de l'inférence



Dans la plupart des cas, il est difficile d'obtenir l'information à partir de la **population** dans son ensemble. On utilise alors un **échantillon** pour tirer des conclusions sur la population.

Introduction à la statistique

Définitions de base: statistique descriptive vs statistique inférentielle

- Statistique descriptive ou déductive:

Phase de la statistique qui se limite à décrire ou analyser une population donnée (graphes, tableaux, résumés numériques,...) sans tirer de conclusion pour une population plus grande

- Statistique inférentielle ou inductive:

Quand un échantillon est représentatif d'une population, on peut, à partir de son analyse, tirer des conclusions importantes pour la population. La partie de la statistique qui s'intéresse au bien-fondé de ces conclusions est la statistique inférentielle ou inductive. Parce que celle-ci n'est jamais absolument certaine, on emploie souvent le langage des probabilités pour établir les conclusions.

Introduction à la statistique

Ressources externes

- Introduction to Statistics in Python Springer – Thomas Haslwanter – Springer (2016)
 - Pdf disponible sur Moodle
 - Codes d'exemple disponibles sur github: https://github.com/thomas-haslwanter/statsintro_python

We want to know about these ...

... but we have to work with these

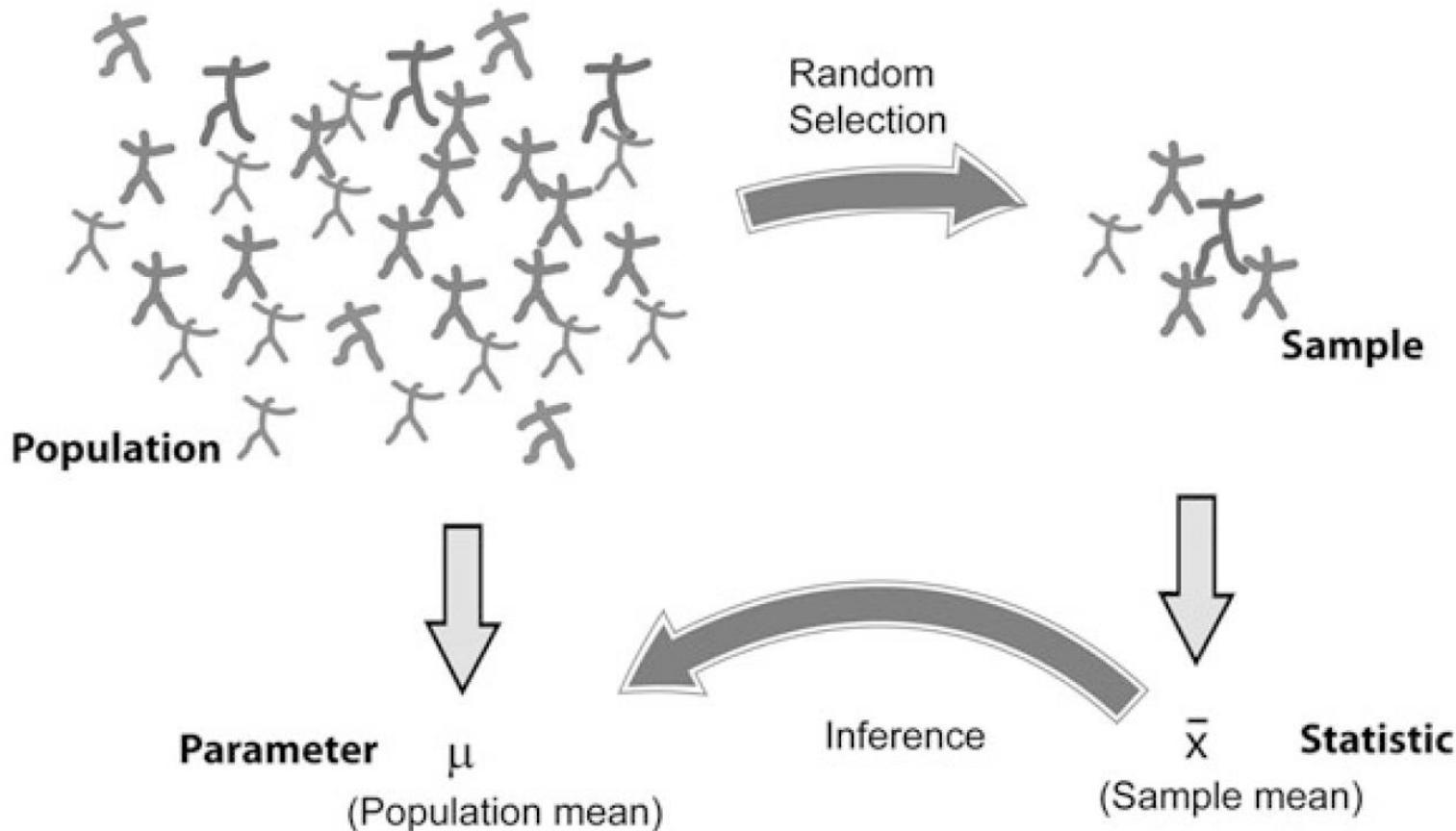


Fig. 5.1 With statistical inference, information from samples is used to estimate parameters from populations



Statistique descriptive

Statistique descriptive

Définition

- La **statistique descriptive** est l'ensemble des méthodes et techniques permettant de présenter, de décrire, de résumer, des données nombreuses et variées:
 - Les **tableaux** : distributions de fréquences;
 - Les **diagrammes** : graphiques;
 - Les **paramètres statistiques** :
Réduction des données à quelques valeurs numériques caractéristiques.
- Il faut préciser d'abord quel est l'ensemble étudié, appelé population statistique, dont les éléments sont des individus ou unités statistiques.
- Chaque individu est décrit **par une ou plusieurs variables**, ou **caractères statistiques**.

Statistique descriptive

Typologie des variables: variables quantitatives

Variable quantitative :

Une variable statistique est quantitative si ses valeurs sont des nombres exprimant une quantité, sur lesquels les opérations arithmétiques (somme, etc...) ont un sens.



Variable quantitative discrète:

Une variable quantitative est discrète si elle ne peut prendre que des valeurs isolées, généralement entières.

Variable quantitative continue:

Une variable quantitative est continue si ses valeurs peuvent être n'importe lesquelles dans un intervalle réel.

Statistique descriptive

Typologie des variables: variables qualitatives

Variable qualitative :

Une variable statistique est qualitative si ses valeurs, ou **modalités**, s'expriment de façon littérale ou par un codage sur lequel les opérations arithmétiques telles que moyenne, somme, ... n'ont pas de sens.



Variable qualitative nominale:
C'est une variable qualitative dont les modalités ne sont pas ordonnées.



Variable qualitative ordinaire:
C'est une variable qualitative dont les modalités sont naturellement ordonnées



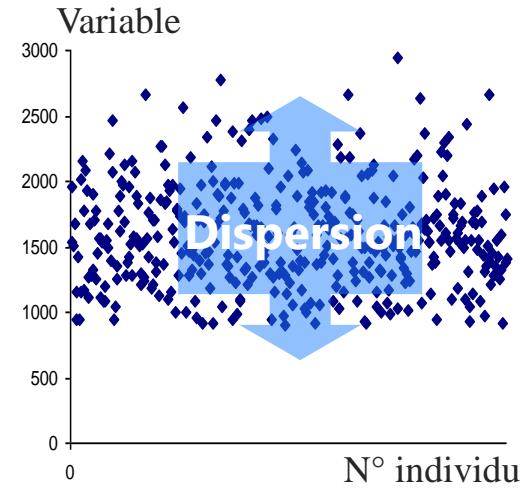
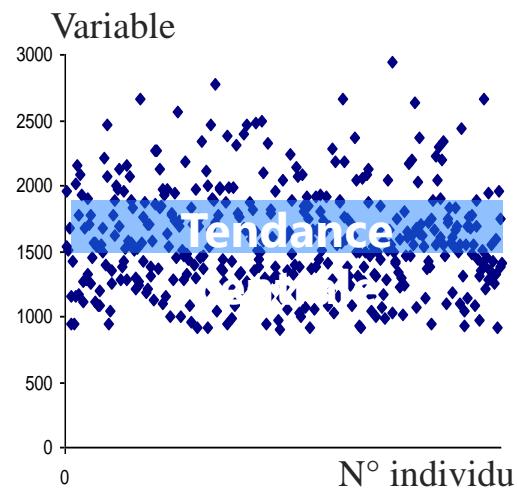
Statistique descriptive

Paramètres statistiques

Statistiques descriptives – Paramètres statistiques

Les représentations graphiques permettent une **synthèse visuelle** de la distribution des observations.

Un **paramètre statistique** permet de **résumer** par une seule quantité numérique une information contenue dans une distribution d'observations.

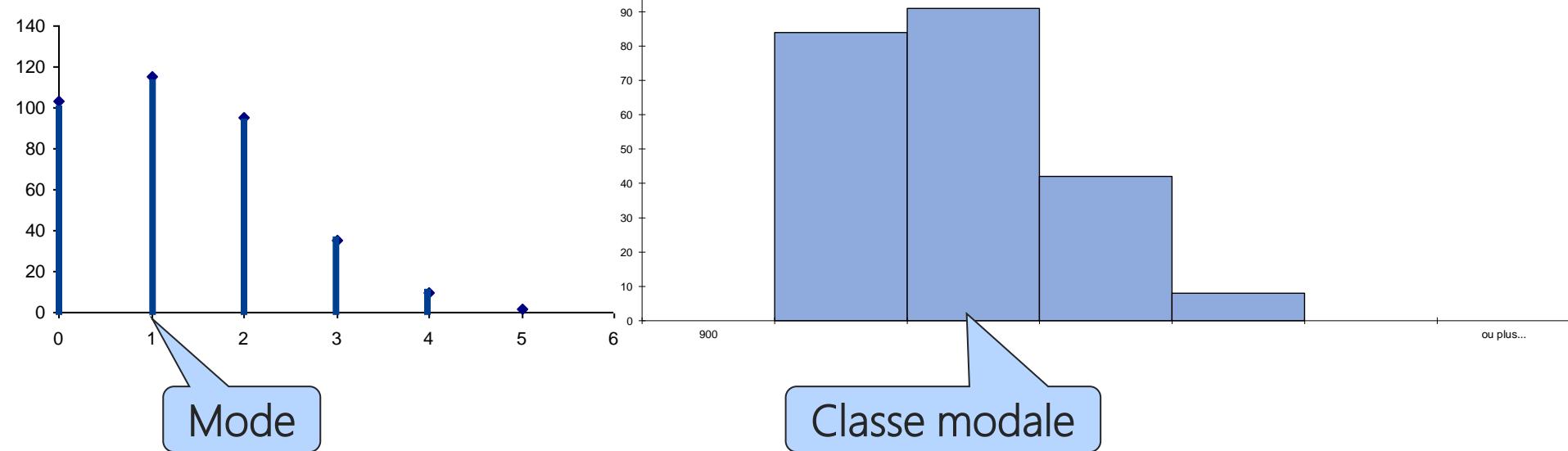


Statistique descriptive – Paramètres statistiques

Tendance centrale: le mode

Une distribution est **unimodale** si elle présente un maximum marqué, et pas d'autres maxima relatifs.

Le **mode** correspond à l'abscisse du maximum, c.à.d. la valeur la plus fréquente ou la classe de fréquence maximale

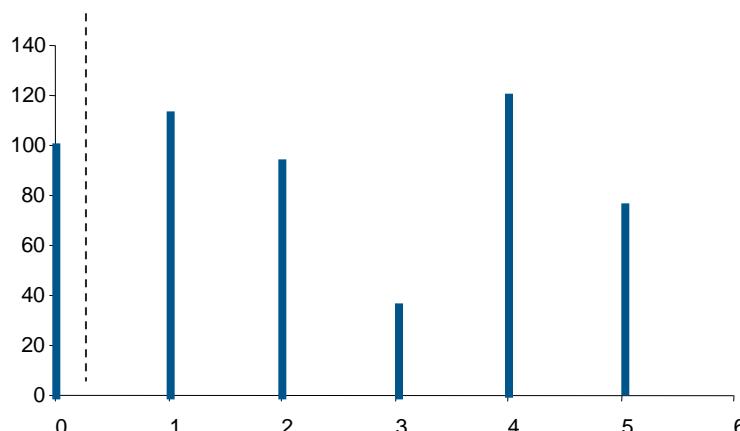


Statistique descriptive – Paramètres statistiques

Tendance centrale: le mode

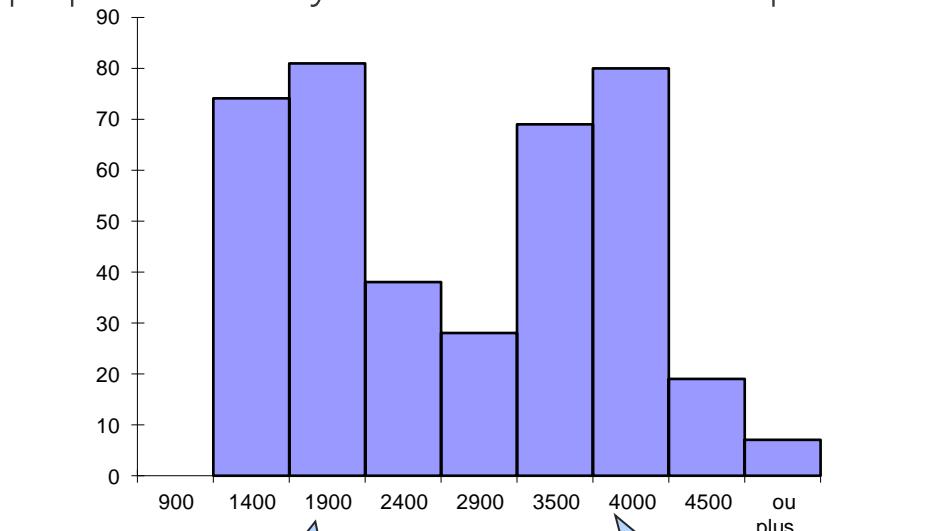
Si la distribution présente 2 ou plus maxima relatifs, on dit qu'elle est **bimodale** ou plurimodale.

La population est composée de plusieurs sous-populations ayant des caractéristiques de tendance centrale différentes.



Mode 1

Mode 2



Classe modale 1

Classe modale 2

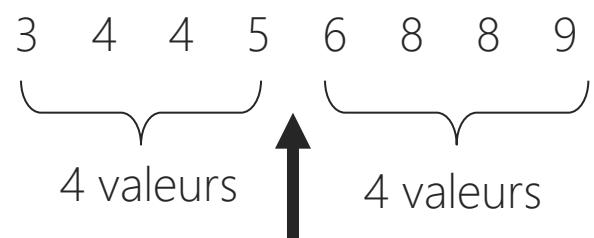
Statistique descriptive - Paramètres statistiques

Tendance centrale: la médiane

Les valeurs observées doivent être rangées par ordre croissant.

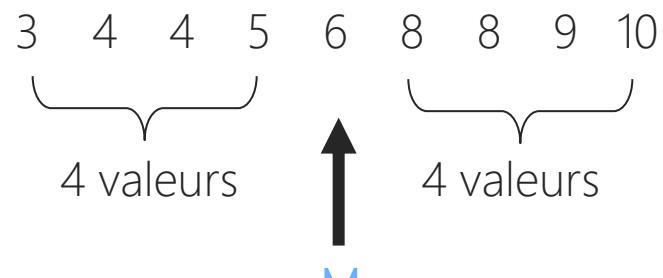
La **médiane M** est la valeur du milieu de la série d'observations, c.à.d. telle qu'il y ait autant d'observation "au-dessous" que "au-dessus".

Nombre pair d'observations



Ici, la médiane peut être n'importe quelle valeur entre 5 et 6. Généralement, on prendra le milieu soit 5.5.

Nombre impair d'observations



Ici, la médiane est une valeur de la série

Statistique descriptive – Paramètres statistiques

Tendance centrale – La moyenne arithmétique, \bar{x}

Série brute

x_1, x_2, \dots, x_n

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

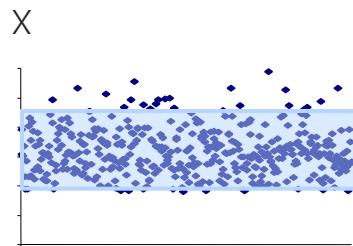
Série groupée

Valeurs de la variable	Effectifs	Fréquences
x_1	n_1	$f_1=n_1/n$
...	...	
x_i	n_i	$f_i=n_i/n$
...	...	
x_k	n_k	$f_k=n_k/n$

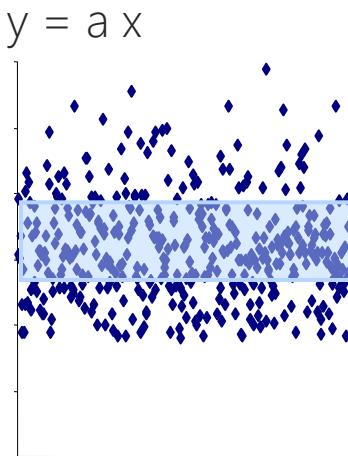
$$\bar{x} = \frac{1}{n} \sum_{i=1}^k n_i x_i = \sum_{i=1}^k f_i x_i$$

Statistique descriptive – Paramètres statistiques

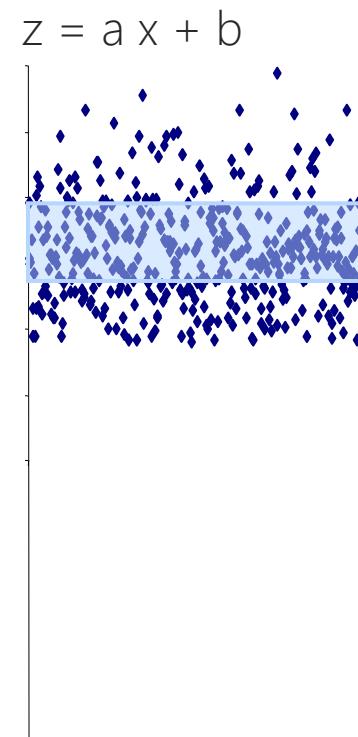
Tendance centrale – Propriétés générales



$P(x)$ = moyenne, médiane,
mode



$P(y) = a P(x)$



$P(z) = a P(x) + b$

Statistique descriptive – Paramètres statistiques

Tendance Centrale – Autres moyennes

- Moyenne géométrique

$$G = \sqrt[n]{x_1^{n_1} x_2^{n_2} \dots x_k^{n_k}}$$

Utilisée dans le cas de phénomènes multiplicatifs (taux de croissance moyen)

- Moyenne harmonique

$$H = \frac{n}{\sum_{i=1}^k \frac{n_i}{x_i}}$$

Utilisée dans le cas où l'on combine 2 variables sous forme de rapport (pièces/heures, km/litre,...)

Statistique descriptive – Paramètres statistiques

Position – Les fractiles ou quantiles

On appelle **fractiles** ou **quantiles** d'ordre k les $(k-1)$ valeurs qui divisent les observations en k parties d'effectifs égaux.

- 1 **médiane** M qui divise les observations en 2 parties égales
- 3 **quartiles** Q_1, Q_2, Q_3 qui divisent les observations en 4 parties égales
- 9 **déciles** D_1, D_2, \dots, D_9 qui divisent les observations en 10 parties égales
- 99 **centiles** C_1, C_2, \dots, C_{99} qui divisent les observations en 100 parties égales

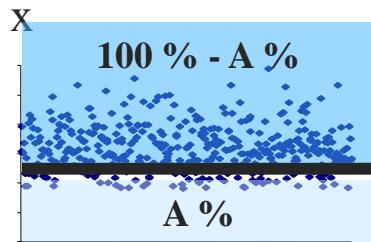
Statistique descriptive – Paramètres statistiques

Position – Les fractiles ou quantiles

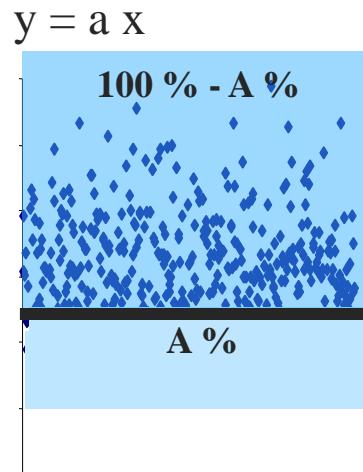
Quartiles, déciles ou centiles s'obtiennent de la même façon que la médiane soit ?!?

Statistique descriptive – Paramètres statistiques

Position – Propriétés générales

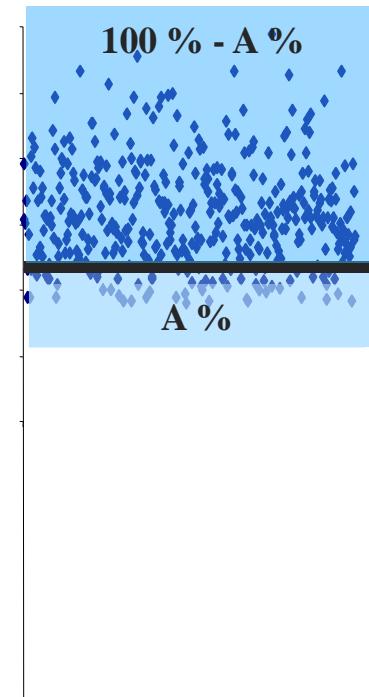


$$Q(x) = \text{quantile}$$



$$Q(y) = a Q(x)$$

$$z = a x + b$$



$$Q(z) = a Q(x) + b$$

Statistique descriptive – Paramètres statistiques

Dispersion

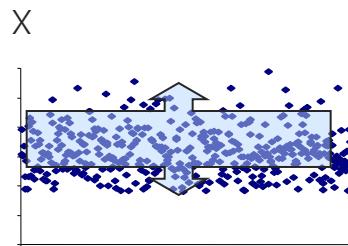
- Etendue : $R = x_{max} - x_{min}$
- Intervalle interquartile : $IQ = Q_3 - Q_1$
- Variance :

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

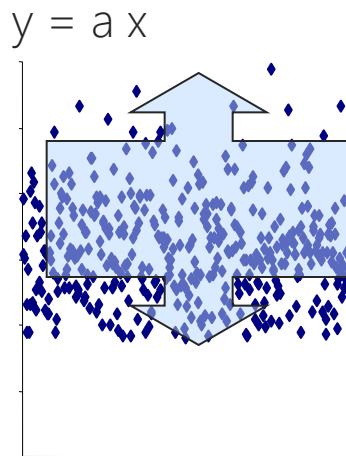
- Ecart-type : $\sigma = \sqrt{V}$

Statistique descriptive – Paramètres statistiques

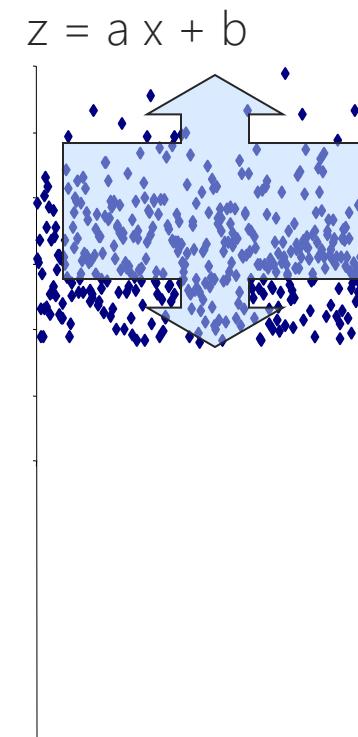
Dispersion – Propriétés générales



$P(x)$ = étendue, écart-type,
intervalle interquartile



$P(y) = a P(x)$



$P(z) = a P(x)$

Statistique descriptive – Paramètres statistiques

Propriétés importantes de la moyenne et de la variance

- Comment se comportent la moyenne et la variance lorsqu'on fait subir un changement de variable aux observations?

$$x_i \longrightarrow y_i = a x_i + b$$

$$\bar{y} = a \bar{x} + b \quad V(y) = a^2 V(x) \quad \sigma(y) = |a| \sigma(x)$$

- Comment se comportent la moyenne et la variance de la somme de deux séries d'observations?

$$\begin{matrix} x_i \\ y_i \end{matrix} \longrightarrow z_i = x_i + y_i$$

$$\bar{z} = \bar{x} + \bar{y} \quad V(z) \neq V(x) + V(y)$$

Statistique descriptive – Paramètres statistiques

Exercice 1.a

Sur base des nombres de cas de Covid pour la Belgique entière rapportés depuis mars 2020 jusqu'en janvier 2023 (voir le fichier covid_cases_belgium.csv sur le portail), pouvez-vous, dans un notebook Jupyter :

- Faire un graphe simple de ces nombres de cas ;
- Calculer la moyenne du nombre de cas par jour ;
- Reporter une représentation simple de cette moyenne dans le graphe initial ;
- Pouvez-vous caractériser cette distribution en termes de mode ?
- Faire un histogramme de ces mêmes données ;
- Représenter correctement cet histogramme ;
- Pouvez-vous caractériser cette distribution en termes de mode ?
- Reporter une représentation simple de la moyenne calculée précédemment dans l'histogramme ?
- Calculer les quartiles de cette distribution et reporter les dans l'histogramme.

Statistique descriptive – Paramètres statistiques

Exercice 1.b

Idem que 1.a mais cette fois appliqué aux données de nombre de décès par jour cliniquement associé au Covid (covid_deaths_belgium.csv)

Statistique descriptive – Paramètres statistiques

Exercice 1.c

Sur base des chiffres de ventes par mois pour les années 2009, 2010 et 2011 (voir le fichier chiffresVente.csv sur le portail), pouvez-vous, dans un notebook Jupyter:

- Faire un graphe simple de ces chiffres de vente;
- Calculer la moyenne;
- Reporter une représentation simple de cette moyenne dans le graphe initial;
- Faire un histogramme de ces mêmes données;
- Représenter correctement cet histogramme;
- Reporter une représentation simple de la moyenne calculée précédemment dans l'histogramme ?



Statistique descriptive

Principales distributions statistiques

Statistique descriptive – Principales distributions

Distribution de Bernoulli - Distribution binomiale (Haslwanter §6.2)

- Distribution discrète
- Donne la probabilité qu'un évènement se produise X fois en N expériences distinctes (X succès, $N-X$ échecs), connaissant la probabilité p qu'un évènement se produise dans une seule expérience:

$$prob(p, N, X) = C_N^X p^X q^{N-X} = \frac{N!}{X!(N-X)!} p^X q^{N-X}$$

où $q=1-p$ est la probabilité d'un échec (complément de la probabilité d'un succès).

- Pourquoi binomiale? Car il y a une directe analogie entre la formule précédente et les termes du développement de la puissance N d'un binôme, $(p+q)$:

$$(p + q)^N = p^N + C_N^1 p^{N-1} q^1 + C_N^2 p^{N-2} q^2 + \dots + C_N^{N-1} p^1 q^{N-1} + q^N$$

Statistique descriptive – Principales distributions

Distribution de Bernoulli – Distribution binomiale (suite)

Caractéristique	Valeur
Moyenne	$\mu = Np$
Variance	$var = \sigma^2 = Npq$
Ecart-type	$\sigma = \sqrt{Npq}$

- Exercice 2: représenter les valeurs de probabilité d'une distribution binomiale correspondant à $p=0.2$ et $N=10$
- Exercice 3: idem pour $p=0.5$

Statistique descriptive – Principales distributions

Distribution de Bernoulli – Distribution binomiale (fin)

- Exercice 4: Sur 2000 familles ayant 4 enfants chacune, combien peuvent s'attendre à avoir (a) au moins 1 garçon, (b) exactement 2 garçons, (c) au plus 2 garçons, (d) 1 ou 2 filles (e) aucune fille.

- Exercice 5: Si 5% des pièces produites par une machine sont défectueuses, déterminer la probabilité pour que sur 4 pièces choisies au hasard on en ait: (a) exactement une seule défectueuse, (b) 0 défectueuse, (c) 2 au plus défectueuses.

Statistique descriptive – Principales distributions

Distribution Géométrique

- Donne la probabilité qu'un évènement se produise 1 seule fois sur N expériences, connaissant la probabilité p qu'un évènement se produise dans expérience unique:

$$prob(p, N) = pq^{N-1}$$

Exercice 5bis: Supposons que nous ayons un ensemble de disque dur. Chaque disque dur meurt de sa belle mort avec une probabilité p au bout d'un an. Comment pouvez-vous caractériser la distribution des quantités suivantes:

- Le nombre de disque qui meurt la première année ;
- Le nombre d'années avant qu'un disque ne meure ;
- L'état d'un disque après une année.

Statistique descriptive – Principales distributions

Distribution normale ou gaussienne (Haslwanter §6.3)

- Distribution continue
- De loin, la plus courante:
 - Parce qu'il y a des résultats analytiques assez simples
 - Pas toujours utilisée à bon escient, en tenant compte de ses limites!!!
- Formule générale de la courbe représentant la densité de probabilité d'une distribution gaussienne

$$Y = \frac{1}{\sigma\sqrt{2\pi}} e^{-(X-\mu)^2/(2\sigma^2)}$$

- La probabilité pour que X se trouve entre a et b , $\Pr(a < X < b)$, est donnée par l'aire sous la courbe, entre les deux points d'abscisses $X=a$ et $X=b$ (avec $a < b$).

Statistique descriptive – Principales distributions

Distribution normale ou gaussienne (Haslwanter §6.3)

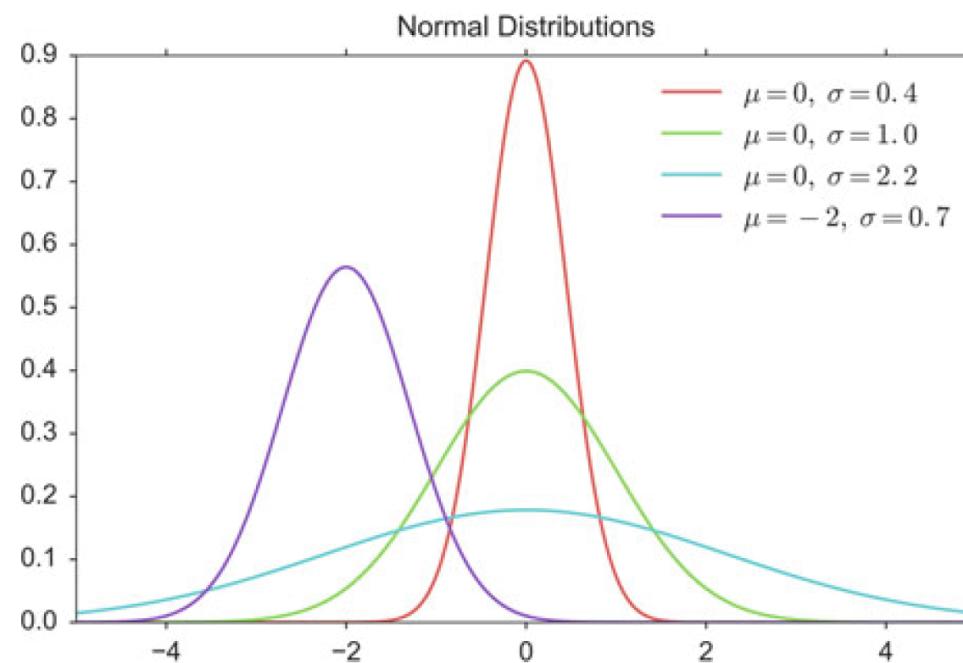
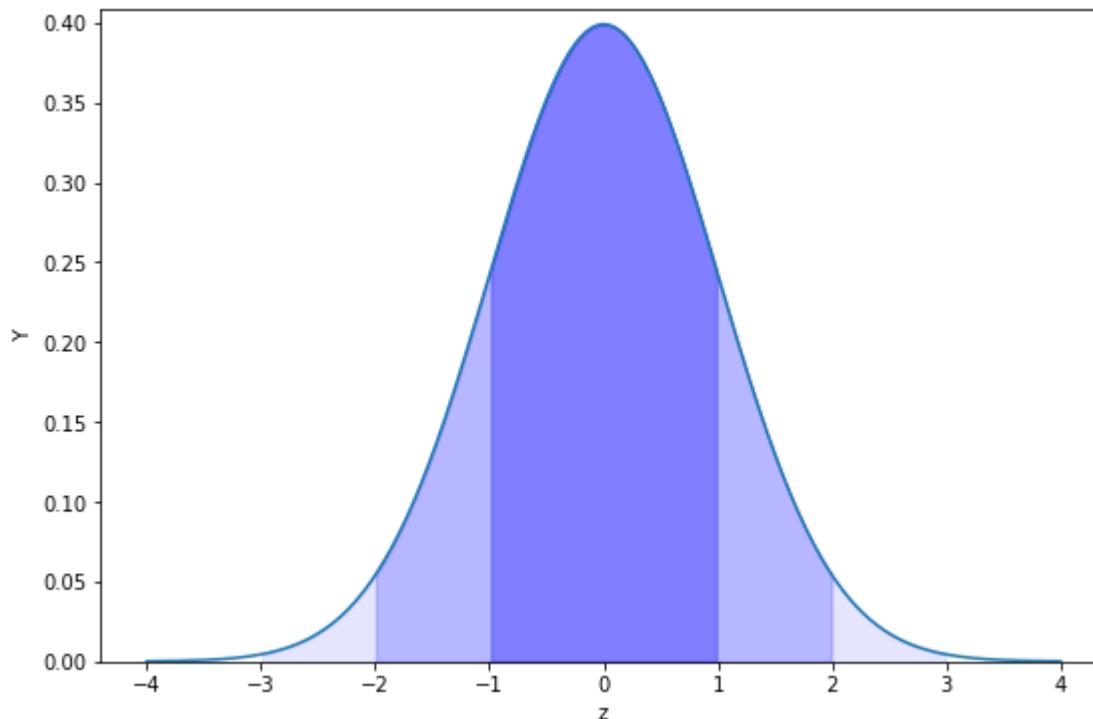


Fig. 6.8 Normal distributions, with different parameters for μ and σ

Statistique descriptive – Principales distributions

Distribution normale ou gaussienne (suite)

- Un graphe vaut mieux qu'un long discours!!!
- Aire sous la courbe complète = 1



L'aire dans le bleu le plus foncé correspond à 68.27% (1 sigma)

L'aire dans le bleu intermédiaire correspond à 95.45% (2 sigmas)

L'aire dans le bleu le plus clair correspond à 99.73% (3 sigmas)

Statistique descriptive – Principales distributions

Distribution gaussienne (suite 2):

Caractéristique	Valeur
Moyenne	μ
Variance	$var = \sigma^2$
Ecart-type	σ

- Relation entre la loi binomiale et la loi normale:

Quand N est grand et quand ni p ni q ne sont trop proches de zéro, la loi binomiale peut être approchée par la distribution normale correspondant à la variable centrée réduite $z = \frac{X-Np}{\sqrt{Npq}}$

Statistique descriptive – Principales distributions

Distribution gaussienne (suite et fin):

- Exercice 6: La taille moyenne de 500 élèves des petites classes d'un lycée est 151 cm et l'écart-type 15 cm. En supposant que la taille soit distribuée suivant une loi normale, trouver combien d'élèves ont leur taille comprise entre (a) 120 et 155 cm, (b) 185 cm et plus.

- Exercice 7: Trouver la probabilité d'obtenir un nombre de faces compris entre 3 et 6 (bornes incluses) quand on jette 10 fois une pièce bien équilibrée en utilisant (a) une loi binomiale, (b) une approximation normale de la loi binomiale.



Fig. 6.9 Twenty-five randomly generated samples of 100 points from a standard normal distribution



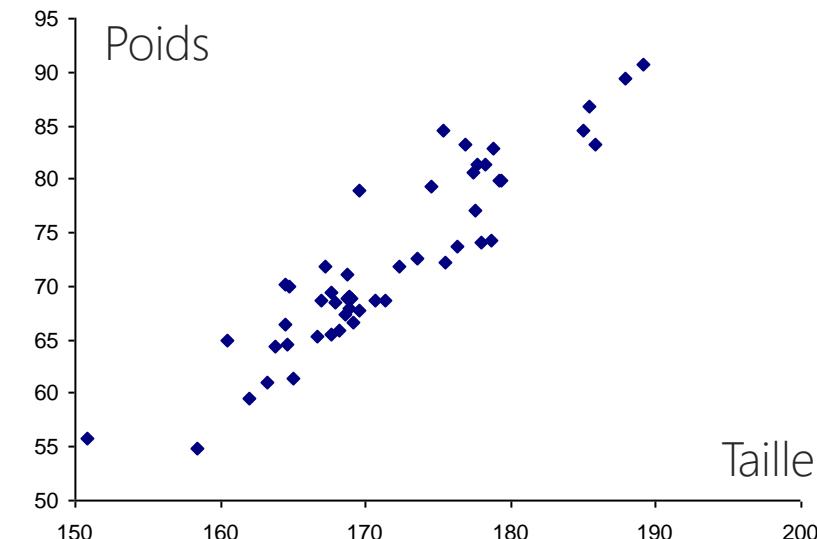
Statistique descriptive

Etude de 2 variables quantitatives

Etude de 2 variables quantitatives

Mesure de la liaison entre 2 variables quantitatives (1)

Nom	Taille x_i (cm)	Poids y_i (kg)
Pierre	175	73
Arantxa	168	56
...
Martin	185	87

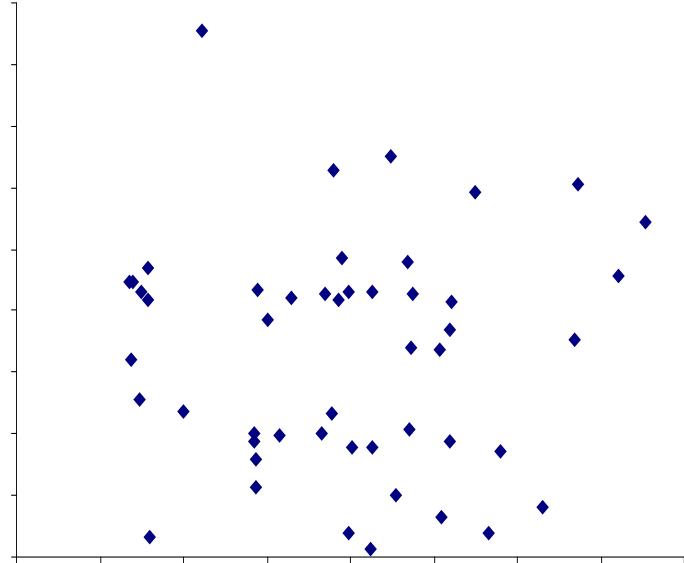


La connaissance de la taille x apporte une certaine information sur le poids y

Il existe une relation de dépendance entre x et y

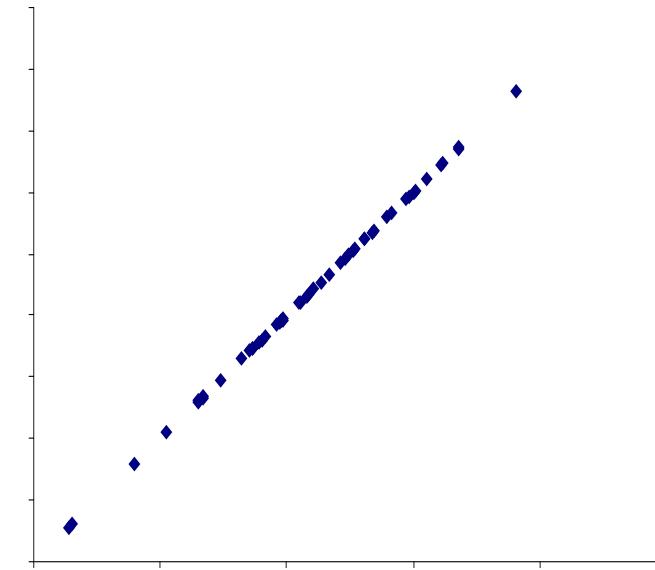
Etude de 2 variables quantitatives

Mesure de la liaison entre 2 variables quantitatives (2)



La connaissance de x n'apporte
aucune information certaine sur y

x et y sont indépendantes



La connaissance de x permet de
connaître exactement la valeur de
 y

Il existe une **relation
fonctionnelle** entre x et y

Etude de deux variables quantitatives

Mesure de la liaison entre 2 variables quantitatives (3) – Covariance

- Covariance:

$$\text{cov}(x, y) = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$$

- Propriétés:

- $\text{cov}(x, y) > 0 \Leftrightarrow x$ et y varient dans le même sens
- $\text{cov}(x, y) < 0 \Leftrightarrow x$ et y varient en sens contraire
- $\text{cov}(x, y) = \text{cov}(y, x)$
- $\text{cov}(x, x) = \text{var}(x)$
- $\text{cov}(x, ay + bz) = a\text{cov}(x, y) + b\text{cov}(x, z)$

Etude de deux variables quantitatives

Mesure de la liaison entre 2 variables quantitatives (4) – Corrélation linéaire

- Corrélation linéaire:

$$\rho = \frac{\text{cov}(x,y)}{\sigma(x)\sigma(y)}$$

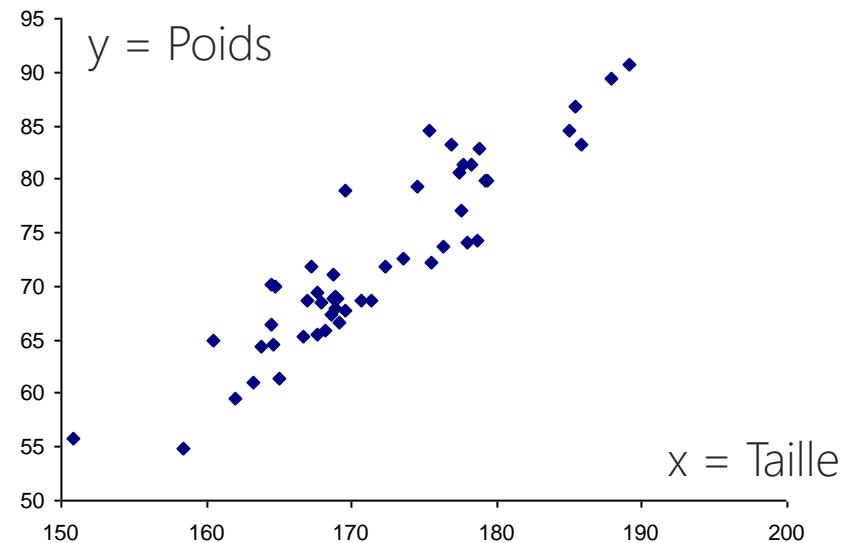
- Propriétés:

- $-1 \leq \rho \leq 1$
- $y = ax + b \Leftrightarrow \begin{cases} \rho = 1 \text{ si } a > 0 \\ \rho = -1 \text{ si } a < 0 \end{cases}$
- $|\rho| = 1 \Leftrightarrow$ il existe une relation fonctionnelle entre x et y
- $\rho = 0 \Leftrightarrow x$ et y sont indépendantes
- $0 < |\rho| < 1$
- Il existe une dépendance linéaire d'autant plus forte entre x et y que ρ est grand



Etude de deux variables quantitatives

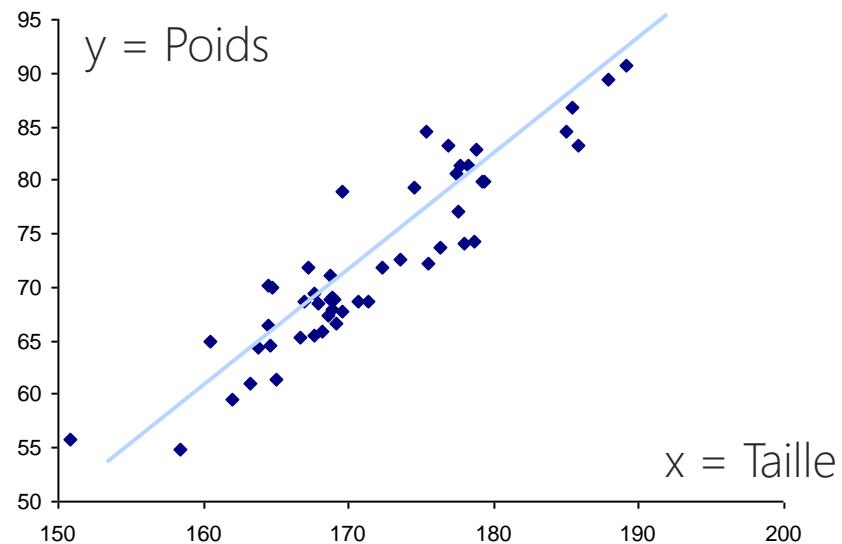
Ajustement linéaire (1)



- Est-il possible de trouver une fonction numérique f telle que $y=f(x)$?
- Si une telle fonction existe, on dit que f est un modèle du phénomène étudié:
 - x est la variable explicative
 - y est la variable expliquée

Etude de deux variables quantitatives

Ajustement linéaire (2)

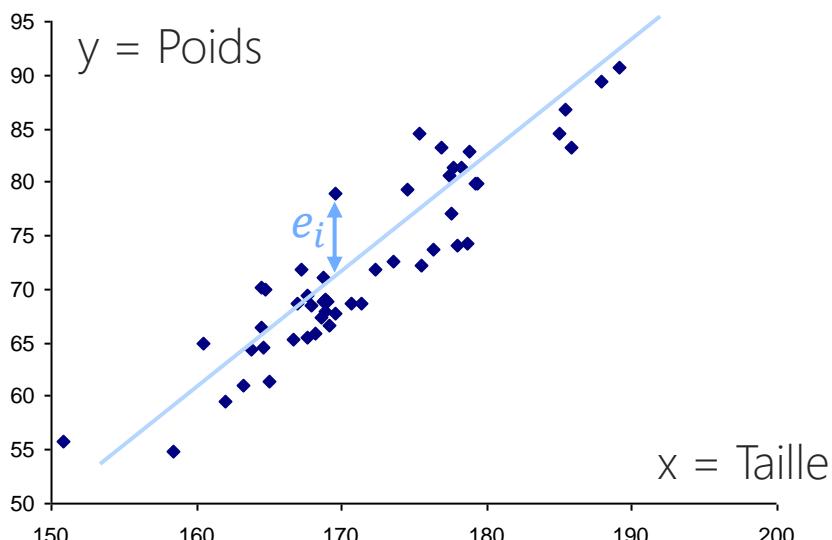


- On désire trouver la droite qui passe "au mieux" à l'intérieur du nuage de points

Etude de deux variables quantitatives

Ajustement linéaire (3)

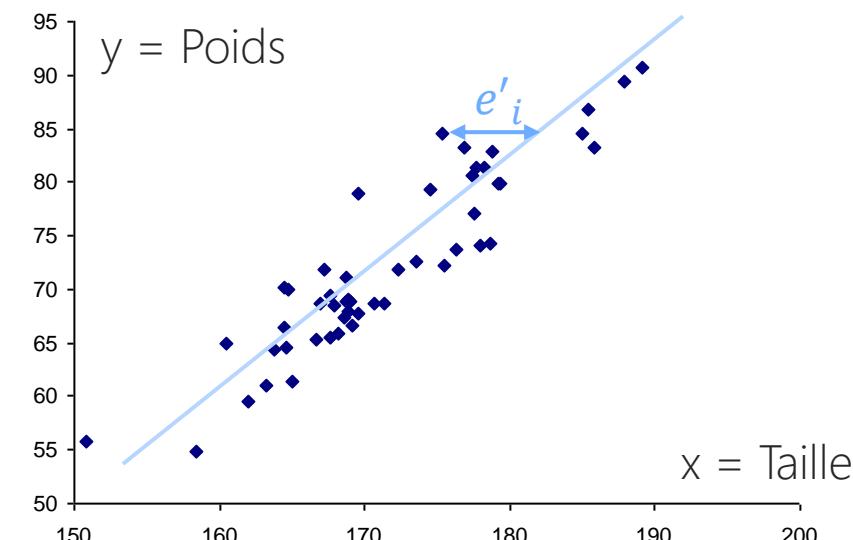
Minimiser $S = \sum_{i=1}^n e_i^2$
où e_i est l'erreur verticale ou résidu



Droite de régression de y en x

"au mieux"

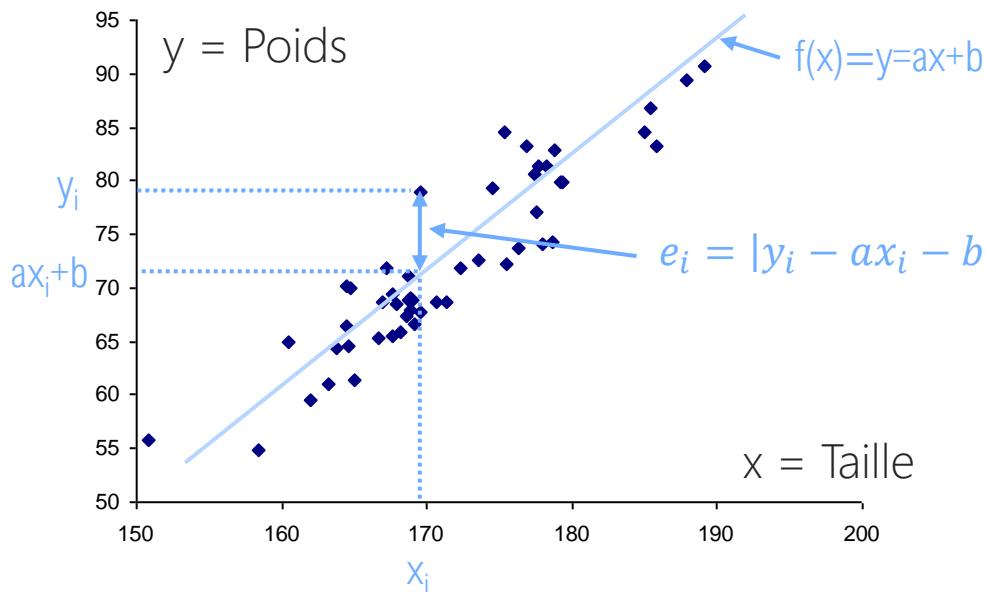
Minimiser $S' = \sum_{i=1}^n e'_i^2$
où e'_i est l'erreur horizontale ou résidu



Droite de régression de x en y

Etude de deux variables quantitatives

Ajustement linéaire (4) – Régression linéaire de y en x



La droite de régression linéaire de y en x , notée $D_{y/x}$, minimise $S = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - ax_i - b)^2$

$$a = \frac{cov(x, y)}{var(x)}$$

$$b = \bar{y} - a\bar{x}$$

$D_{y/x}$ passe par le point moyen (\bar{x}, \bar{y})

Etude de deux variables quantitatives

Ajustement linéaire (5) – Régression linéaire de x en y

La droite de régression linéaire de x en y , notée $D_{x/y}$, minimise $S' = \sum_{i=1}^n e'^2_i = \sum_{i=1}^n (x_i - a'y_i - b')^2$

$$a' = \frac{cov(x, y)}{var(y)} \quad b' = \bar{x} - a'\bar{y}$$

$D_{x/y}$ passe par le point moyen (\bar{x}, \bar{y})

Etude de deux variables quantitatives

Ajustement linéaire (6) – Liens entre corrélation et droites de régression

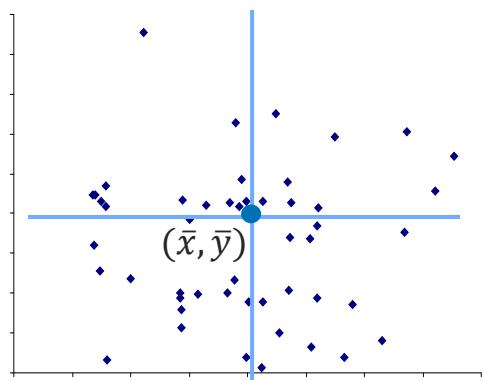
$$D_{y/x}: \quad y = ax + b \quad a = \frac{\text{cov}(x,y)}{\text{var}(x)}$$

$$D_{x/y}: \quad x = a'y + b' \quad a' = \frac{\text{cov}(x,y)}{\text{var}(y)}$$

$$b = \bar{y} - a\bar{x}$$

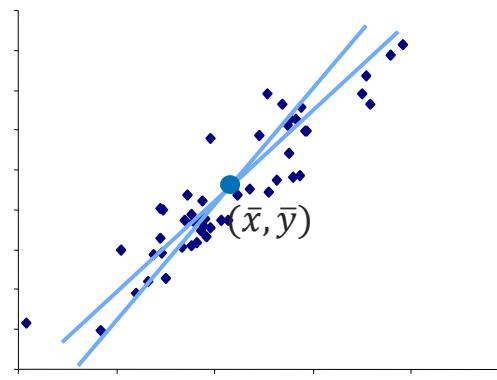
$$\rho^2 = aa' \quad \rho = a \frac{\sigma(x)}{\sigma(y)} = a' \frac{\sigma(y)}{\sigma(x)}$$

$$b' = \bar{x} - a'\bar{y}$$



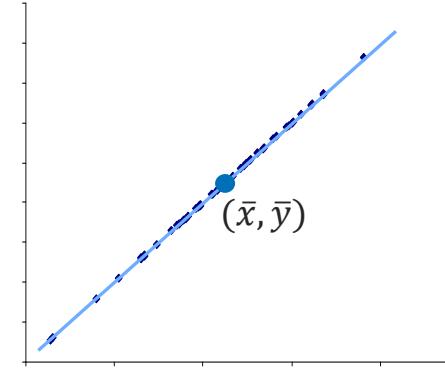
$$\rho^2 = a a' = 0$$

Indépendance linéaire



$$0 < \rho^2 = a a' < 1$$

Le degré de dépendance linéaire se mesure à la proximité des droites de régression



$$\rho^2 = a a' = 1$$

Liaison fonctionnelle linéaire

Etude de deux variables quantitatives

Ajustement linéaire – Exercice 8

- Fichier de données: CourbePoidsTaille.csv
- Descriptif: Poids à 1 kg près et taille à 1 cm près d'un échantillon de 12 étudiants de sexe masculin tirés au hasard parmi les étudiants de 1ière année d'une université
 - Représenter le nuage de points correspondant
 - Construire les droites des moindres carrés qui ajuste ces données, en utilisant respectivement le poids et la taille comme variable indépendante
 - Donner leurs équations
 - Estimer la taille d'un étudiant dont le poids serait de 63 kilogrammes
 - Estimer le poids d'un étudiant dont la taille serait de 178 centimètres

Etude de deux variables quantitatives

Ajustement linéaire – Exercice 8.1 et 8.2

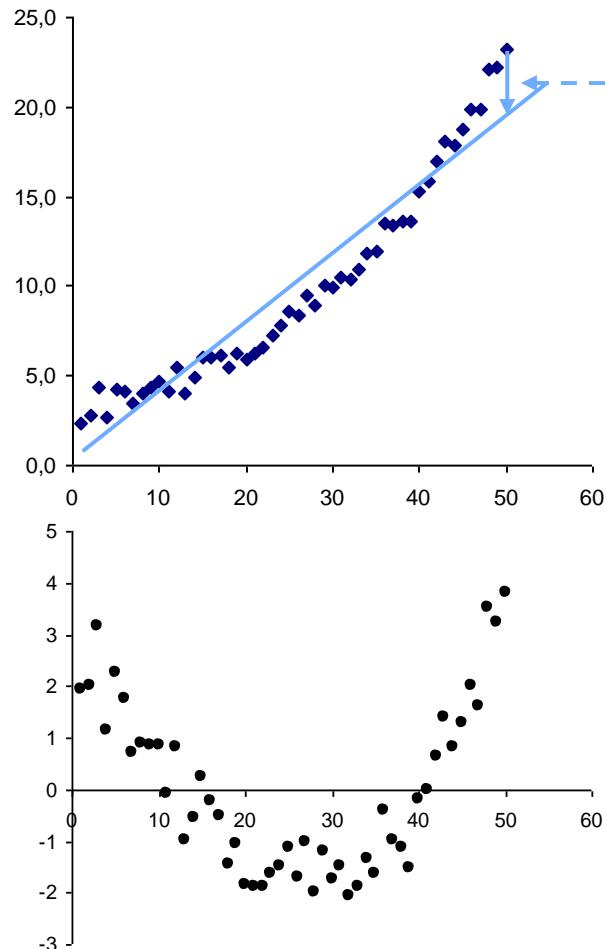
Adaptation de l'Exercice 8 à:

- 8.1) Une grande population de 25000 enfants de Hong-Kong (fichier de données: :SOCR_Data_Dinov_020108_HeightsWeights.txt)
- 8.2) Une population de joueurs de baseball. (fichier de données: baseball_players.csv)

Etude de deux variables quantitatives

Ajustement à une fonction exponentielle (1)

x_i	y_i
2,8	0,8
4,3	1,2
2,7	1,5
4,2	1,9
4,1	2,3
...	...
4,0	3,1



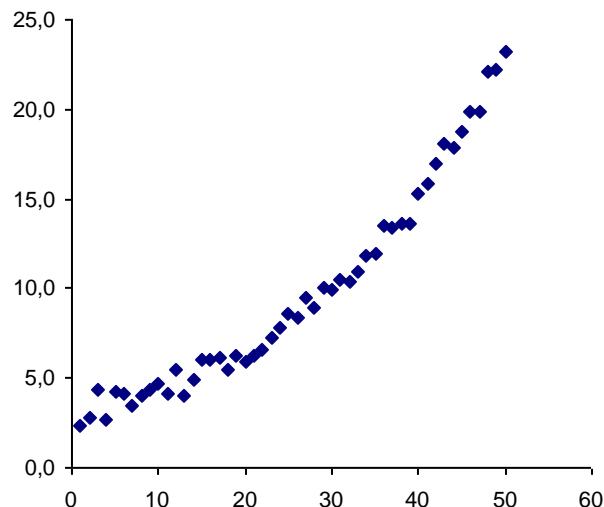
Droite de régression
linéaire de y en x

Analyse des résidus

Les résidus devraient se répartir au hasard autour de l'axe des abscisses (autour de zéro): le modèle linéaire ne convient pas!!!

Etude de deux variables quantitatives

Ajustement à une fonction exponentielle (2)



Modèle exponentiel

$$y = e^x$$

exponentielle de base e

$$y = a^x$$

exponentielle de base
a

$$y = b a^x$$

Forme exponentielle générale

Changement de variable

$$\ln y = \ln b + x \ln a$$

$$Y = A X + B$$

avec

$$Y = \ln y$$

$$X = x$$

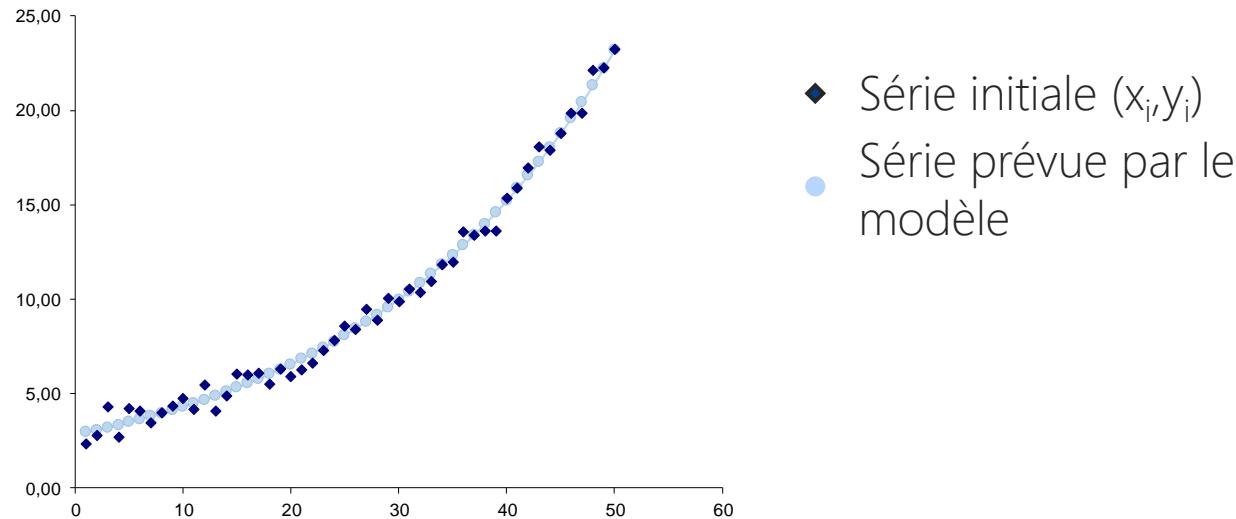
$$A = \ln a$$

$$B = \ln b$$

L'ajustement affine de Y en fonction de X donne A et B,
d'où $a = e^A$, $b = e^B$, et le modèle $y = ba^x$

Etude de deux variables quantitatives

Ajustement à une fonction exponentielle (3)



- ◆ Série initiale (x_i, y_i)
- Série prévue par le modèle

(x_i, \hat{y}_i)

Analyse des résidus

Le modèle exponentiel est mieux adapté que le modèle linéaire

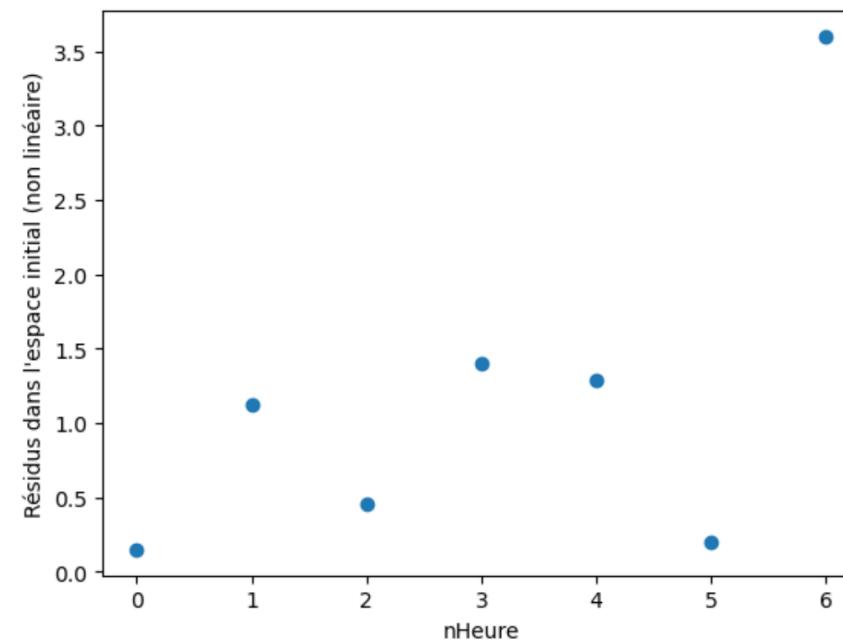
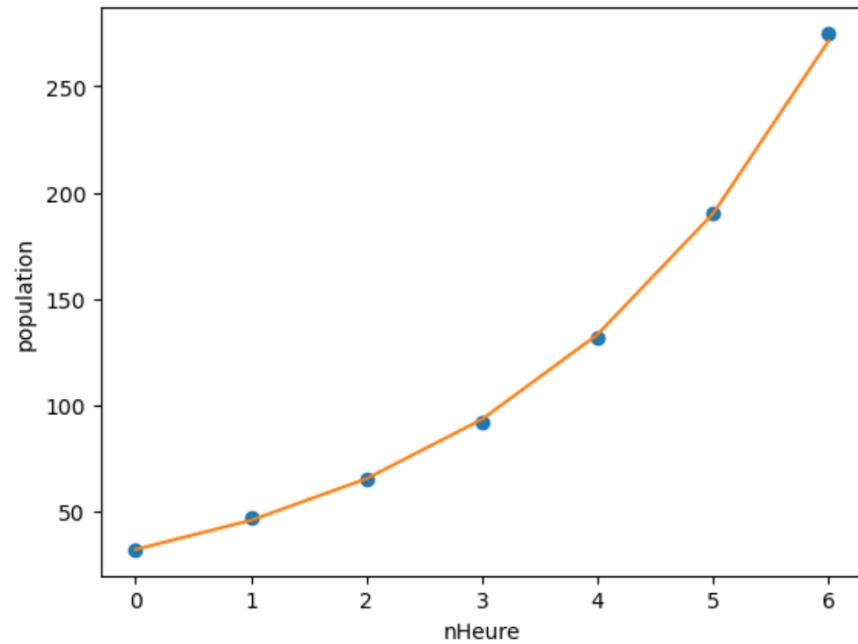
Etude de deux variables quantitatives

Ajustement à une fonction exponentielle (4) – Exercice 9.1

- Fichier de données: Bacteries.csv
- Descriptif: nombre de bactéries par unité de volume présentes dans un bouillon de culture au fil du temps (en heure)
 - Représenter le nuage de points correspondant
 - Envisager un modèle de type: $n_{bact} = ba^{n_heure}$
 - Ramener les données dans une représentation linéaire du problème
 - Estimer les constantes a et b sur base d'une optimisation de moindre carré
 - Superposer la courbe modèle au nuage de point initial
 - Dans un nouveau diagramme, visualiser l'ensemble des résidus. Notre hypothèse de modèle de type exponentiel vous semble-t-elle justifiée. Qu'aurait-on pu concevoir comme autre famille de modèle?

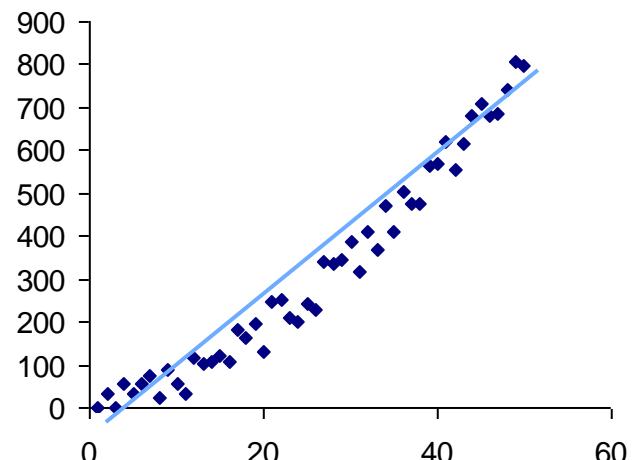
Etude de deux variables quantitatives

Exercice 9.1 – Vues graphiques des résultats

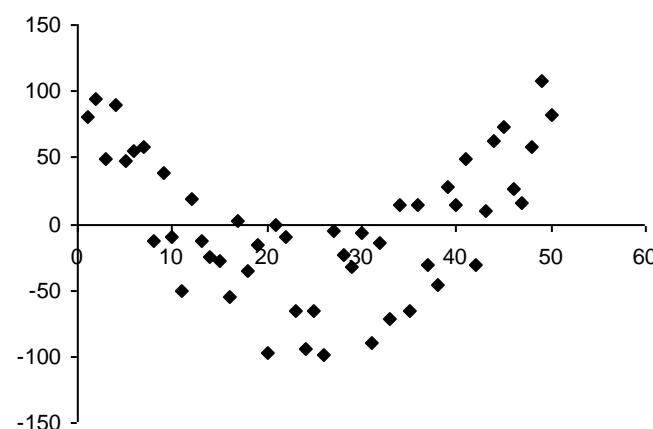


Etude de deux variables quantitatives

Ajustement à une fonction puissance (1)



Droite de régression linéaire de y en x

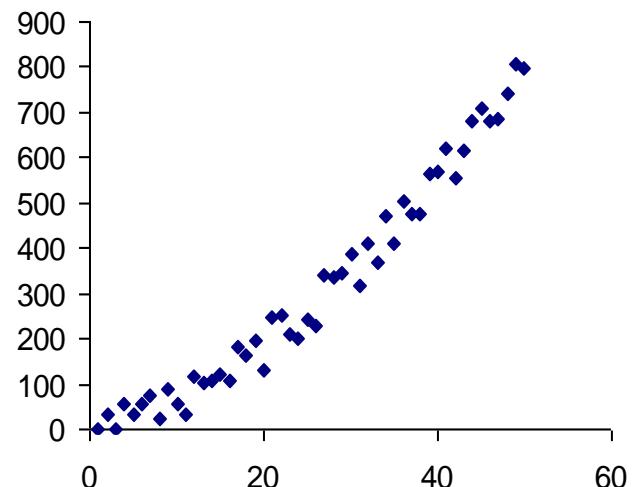


Analyse des résidus

Le modèle linéaire ne convient pas

Etude de deux variables quantitatives

Ajustement à une fonction puissance (2)



Modèle puissance $y = b x^a$

Changement de variable

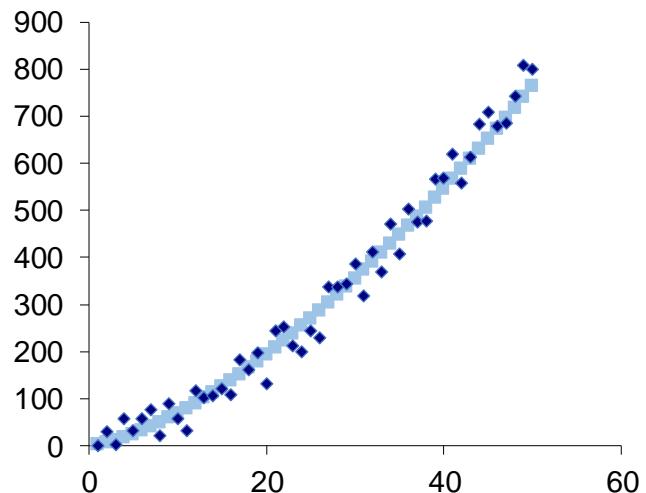
$$\ln y = \ln b + a \ln x$$

$$Y = A X + B \quad \text{avec} \quad \begin{aligned} Y &= \ln y \\ X &= \ln x \\ A &= a \\ B &= \ln b \end{aligned}$$

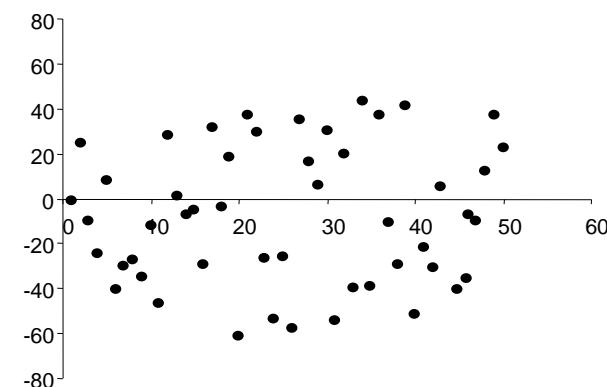
L'ajustement affine de Y en fonction de X donne A et B ,
d'où $a = A$, $b = e^B$ et le modèle $y = bx^a$

Etude de deux variables quantitatives

Ajustement à une fonction puissance (3)



- ◆ Série initiale (x_i, y_i)
- Série prévue par le modèle (x_i, \hat{y}_i)



Analyse des résidus

Le modèle puissance est mieux adapté que le modèle linéaire

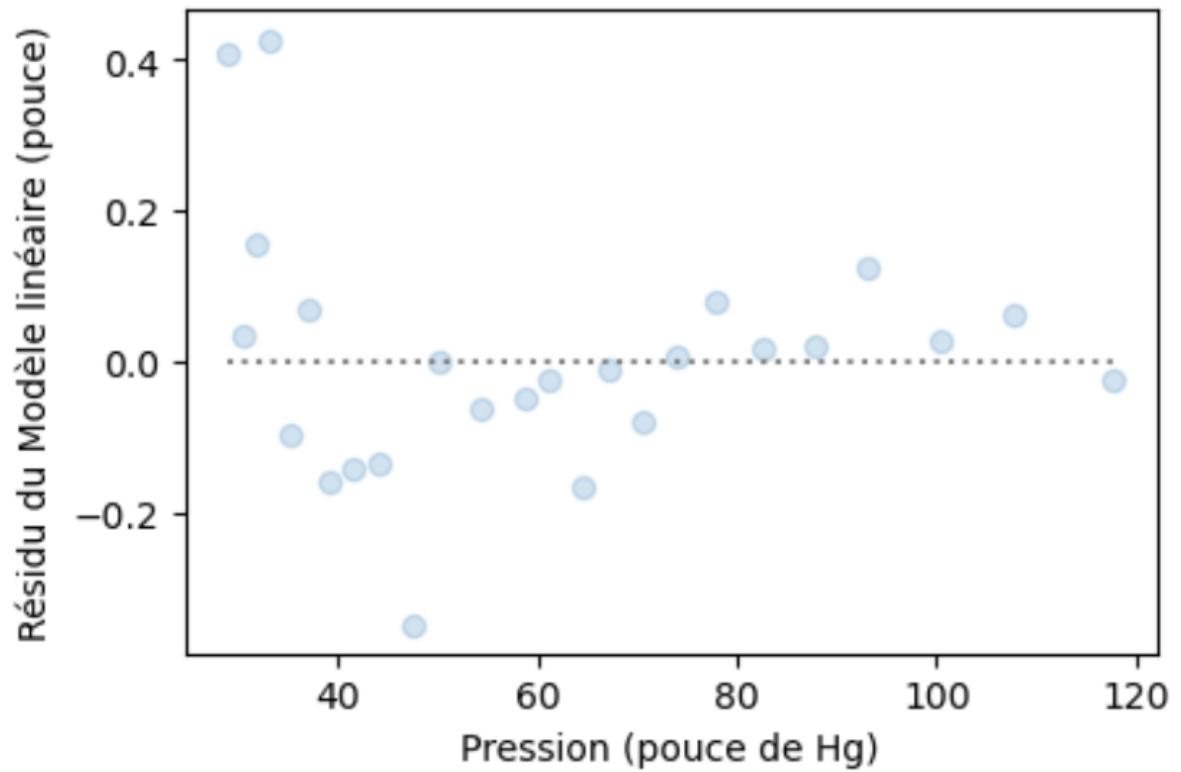
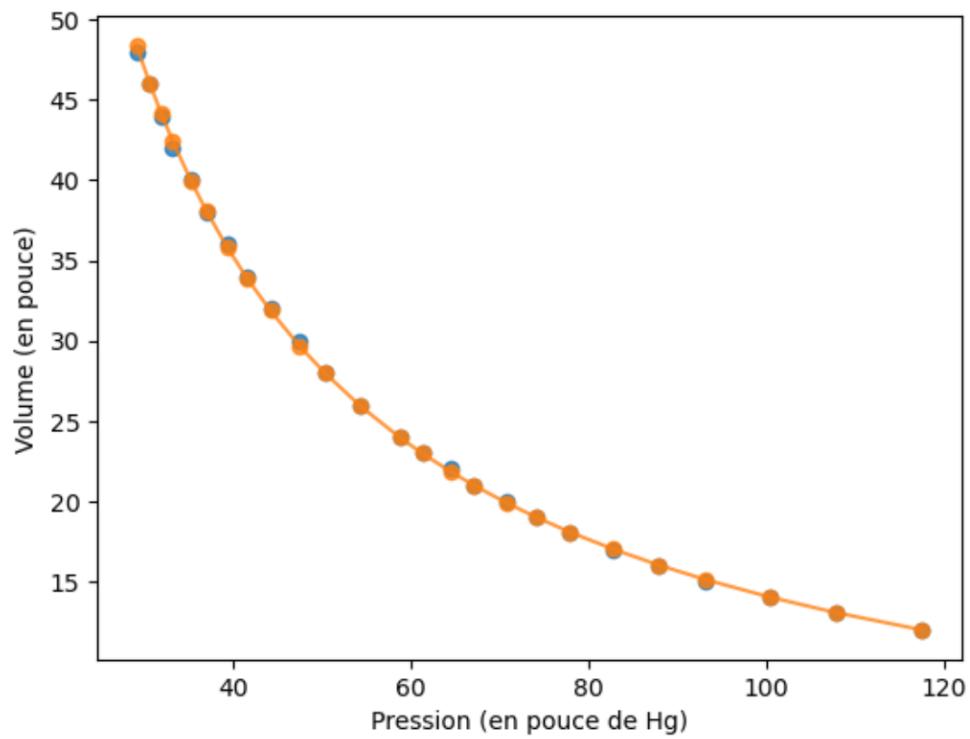
Etude de deux variables quantitatives

Ajustement à une fonction puissance (4) – Exercice 9.2

- Fichier de données: Boyle.txt
- Descriptif: relation entre le volume de gaz (mesuré ici par la hauteur du cylindre contenant le gaz, exprimée en pouce) et la pression (mesurée ici par la hauteur de mercure du baromètre utilisé, toujours en pouce).
 - Représenter le nuage de points correspondant
 - Envisager un modèle de type: $P^aV = \text{Constante}$
 - Ramener les données dans une représentation linéaire du problème
 - Estimer les constantes a et Constante sur base d'une optimisation de moindre carré
 - Superposer la courbe modèle au nuage de point initial
 - Dans un nouveau diagramme, visualiser l'ensemble des résidus. Notre hypothèse de modèle de type 'puissance' vous semble-t-elle justifiée. Qu'aurait-on pu concevoir comme autre famille de modèle?

Etude de deux variables quantitatives

Exercice 9.2 – Vues graphiques des résultats



Etude de deux variables quantitatives

Exercice 9.3

- Fichier de données: Polonium210.csv
- Descriptif: Le Polonium 210 est une substance instable qui émet des particules alpha pour se décomposer en Plomb 206. Au cours du temps, la quantité de Polonium, $N(t)$, évolue suivant une loi de type

$$N(t) = N(0)e^{-\lambda t}$$

- Représenter le nuage de points correspondant aux données de départ
- Un modèle linéaire vous semble-t-il opportun?
- Envisager une transformation des données ramenant ce problème à un problème linéaire
- Après cette transformation, estimer la constante λ en optimisant le problème linéaire
- Superposer la courbe modèle au nuage de point initial

Etude de deux variables quantitatives

Exercice 9.3 – Correction (1)

- A priori, on serait plutôt dans une modélisation de phénomènes exponentiel => on repartirait du notebook Jupyter correspondant à la modélisation de l'évolution du nombre de cellules dans un milieu de culture.

$$N(t) = N(0)e^{-\lambda t}$$

Prenons les logarithmes népériens de part et d'autre:

$$\ln(N(t)) = \ln(N(0)e^{-\lambda t})$$

ou

$$\ln(N(t)) = \ln(N(0)) - \lambda t \ln(e)$$

ou encore

$$\ln(N(t)) = -\lambda \ln(e)t + \ln(N(0))$$

A mettre en parallèle avec

$$Y = AX + B$$

Voir slide 81

Etude de deux variables quantitatives

Exercice 9.3 – Correction (2)

$$\begin{aligned}Y &= \ln(N(t)) \\X &= t \\A &= -\lambda \ln(e) = -\lambda 1 = -\lambda\end{aligned}$$

où on utilise le fait que $\ln(e)=1$ par définition du logarithme népérien!!!

$$B = \ln(N(0)) = \ln(100) = 4.6051 \dots$$

Donc, finalement, la valeur de lambda est données par:

$$\lambda = -A = 0.005 \dots$$

Pour les physiciens (ce sont de drôles de bêtes...), $1/\lambda$ est souvent plus intéressante car cette valeur a pour unité des années...

On trouve ici, $1/\lambda = 200$ années. Ce qui veut dire qu'il faut 200 années pour que la concentration de polonium soit réduite d'un facteur e, soit 2.71...

Le temps de demi-vie, $t_{1/2}$ est lui le temps après lequel la concentration de polonium sera réduite d'un facteur deux soit:

$$\frac{1}{2} = e^{-\lambda t_{1/2}}$$

Après transformation, on voit que les deux temps sont liés par

$$t_{1/2} = \frac{\ln 2}{\lambda} = \frac{0.6914 \dots}{\lambda}$$

Ici, on trouve donc que le temps de demi-vie du polonium 210 est de 138.59 années ce qui veut dire qu'au bout de 138.59 années le baton de polonium sera transformé à 50% en plomb!!!

Sommaire

1. Présentation du professeur
 2. Introduction à Python
 3. Introduction à la statistique
 4. Statistique Descriptive
 5. Statistique Inférentielle
-
- I. Présentation générale
 - II. Théorie de l'échantillonnage
 - III. Théorie de l'estimation
 - IV. Théorie de la décision – Test d'hypothèse et de signification

Statistique inférentielle – Présentation générale

Individu vs Population

Individu ou unité statistique

Une unité distincte chez laquelle on peut observer une ou plusieurs caractéristiques données.



Statistique inférentielle – Présentation générale

Individu vs Population

Population

Ensemble des individus (ou unités statistiques) pour lequel on considère une ou plusieurs caractéristiques



Statistique inférentielle – Présentation générale

But de la statistique inférentielle



Dans la plupart des cas, il est difficile d'obtenir l'information à partir de la **population** dans son ensemble. On utilise alors un **échantillon** pour tirer des conclusions sur la population.

L'inférence statistique consiste à **induire** les caractéristiques inconnues d'une **population** à partir d'un **échantillon** issu de cette population.

Les caractéristiques de l'échantillon, une fois connues, reflètent avec une certaine **marge d'erreur** possible celles de la population

Sommaire

1. Présentation du professeur
2. Introduction à Python
3. Introduction à la statistique
4. Statistique Descriptive
5. Statistique Inférentielle
 - I. Présentation générale
 - II. Théorie de l'échantillonnage
 - III. Théorie de l'estimation
 - IV. Théorie de la décision – Test d'hypothèse et de signification

Statistique inférentielle

Théorie de l'échantillonnage – Définitions (1)

- ▲ La théorie de l'échantillonnage est l'étude des liaisons existant entre une population et les échantillons de cette population
- ▲ Cette théorie est fondamentale pour estimer les quantités qui caractérisent une population (les **paramètres de la population** comme la moyenne, la variance d'une caractéristique de la population) à partir des quantités correspondantes de l'échantillon (moyenne, variance de la même caractéristique sur l'échantillon), souvent appelées **statistiques de l'échantillon**
- ▲ D'une façon générale, on appelle **inférence statistique** l'étude des conclusions que l'on peut tirer à partir d'un échantillon d'une population et du degré d'exactitude de ces conclusions

Statistique inférentielle

Théorie de l'échantillonnage – Définitions (2)

Echantillons aléatoires:

- ▲ Afin que les conclusions de la théorie de l'échantillonnage et de l'inférence statistique soient valables, les échantillons choisis doivent être **représentatifs** de la population
- ▲ On appelle **plan d'expérience** l'étude des méthodes d'échantillonnage et des problèmes qui s'y attachent
- ▲ Pour obtenir un échantillon représentatif, on peut procéder à un **échantillonnage aléatoire**

Statistique inférentielle

Théorie de l'échantillonnage – Définitions (3)

Echantillonnages exhaustif ou non:

- ▲ Quand on a extrait un individu d'une urne, avant de procéder à un nouveau tirage, on peut soit l'y remettre, soit ne pas l'y remettre ☺
 - ▲ Dans le premier cas, on parlera d'**échantillonnage non exhaustif**, un même individu pouvant être tiré plusieurs fois
 - ▲ Dans le deuxième cas, on parlera d'**échantillonnage exhaustif**, un même individu ne pouvant être tiré qu'une et une seule fois
- ▲ Attention: une population finie sur laquelle l'échantillonnage est non exhaustif peut théoriquement être considérée comme infinie puisque chaque individu peut être extrait sans épuiser la population!

Statistique inférentielle

Théorie de l'échantillonnage – Définitions (4)

Distribution d'échantillonnage

- ▲ On considère tous les échantillons possibles de taille n (exhaustifs ou non) qu'on peut extraire d'une certaine population de taille N
- ▲ Pour chacun de ces échantillons on peut calculer certaines statistiques sur différentes dimensions i :
 - ▲ La moyenne de la dimension i sur l'échantillon
 - ▲ L'écart-type de la dimension i sur l'échantillon
 - ▲ ...

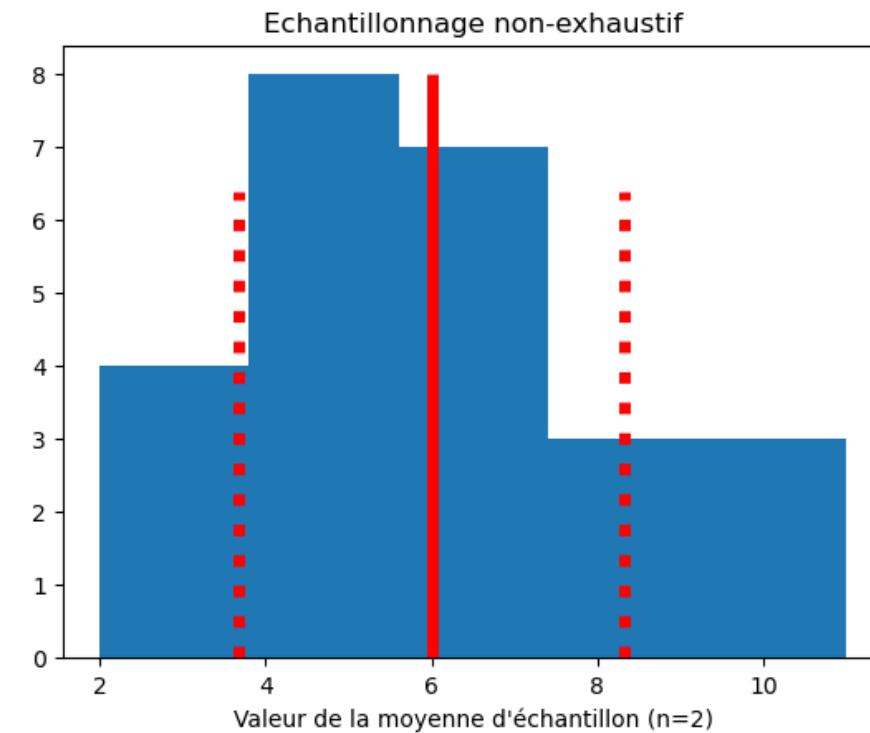
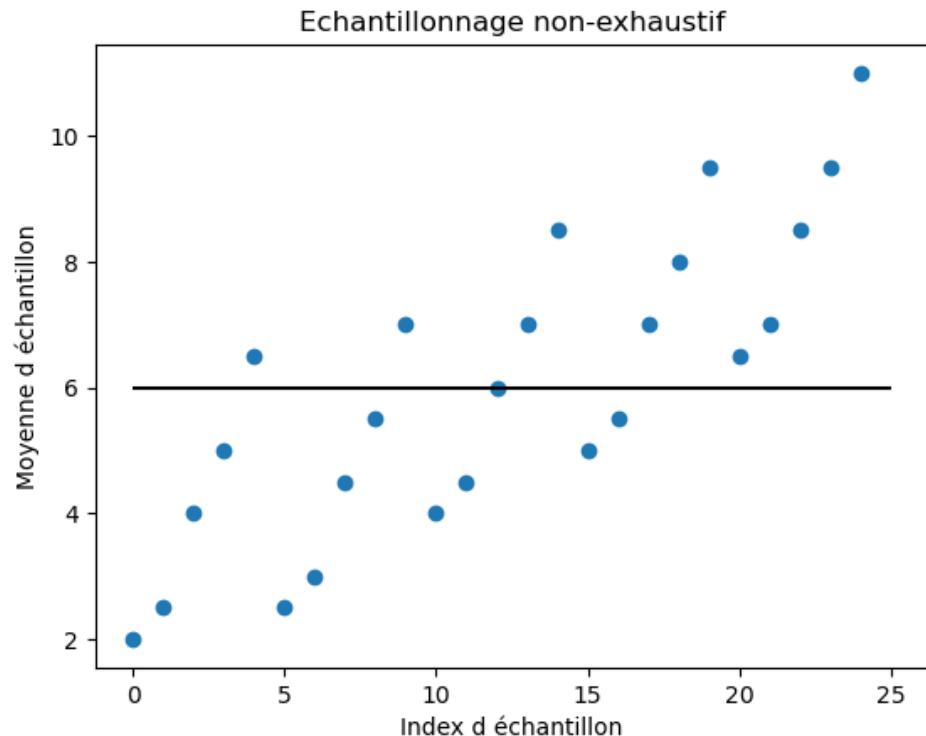
Statistique inférentielle

Théorie de l'échantillonnage – Exercice 10.1 (1)

- Une population est constituée des cinq nombres 2, 3, 6, 8 et 11. On considère tous les échantillons **non exhaustifs** possibles de taille deux de cette population.
- Trouver:
 - La moyenne de la population
 - L'écart-type de la population
 - La moyenne de la distribution des moyennes d'échantillon
 - L'écart-type de la distribution des moyennes d'échantillon

Statistique inférentielle

Théorie de l'échantillonnage – Exercice 10.1 (2)



$$\mu_{\bar{X}} = \mu = 6.00 \quad \text{et} \quad \sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{3,29}{\sqrt{5}} = 2.32$$

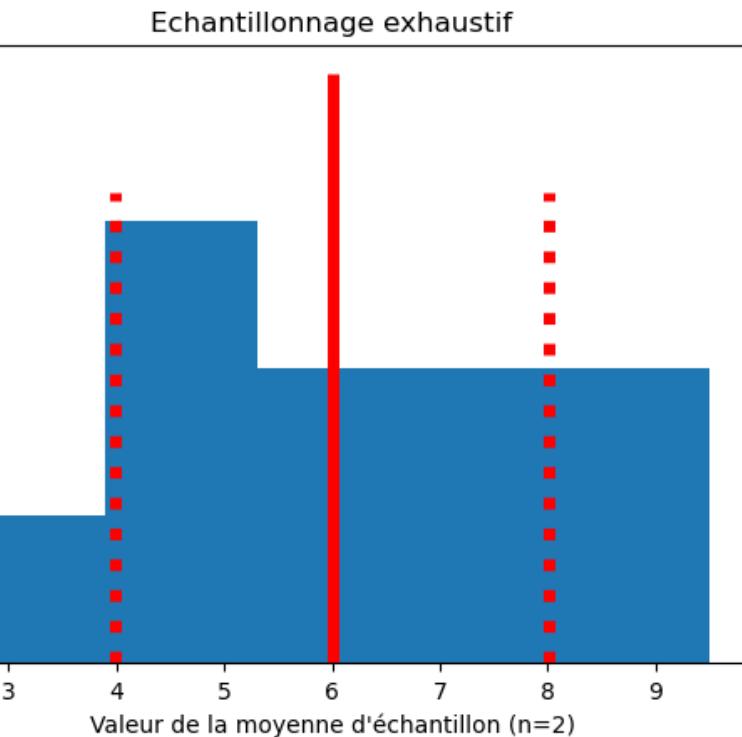
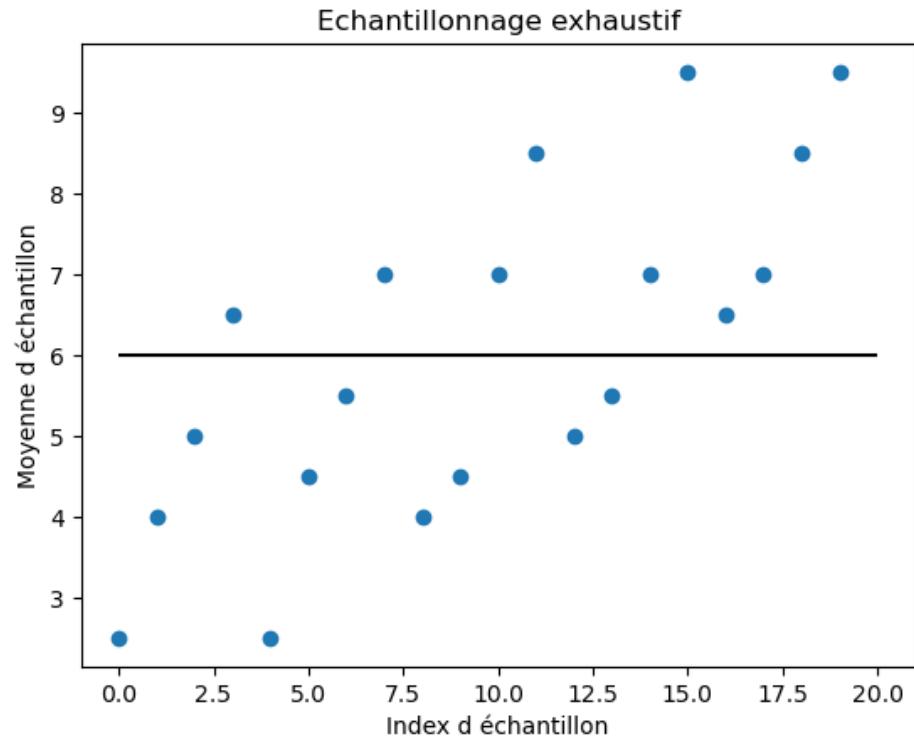
Statistique inférentielle

Théorie de l'échantillonnage – Exercice 10.2 (1)

- Une population est constituée des cinq nombres 2, 3, 6, 8 et 11. On considère tous les échantillons **exhaustifs** possibles de taille deux de cette population.
- Trouver:
 - La moyenne de la population
 - L'écart-type de la population
 - La moyenne de la distribution des moyennes d'échantillon
 - L'écart-type de la distribution des moyennes d'échantillon

Statistique inférentielle

Théorie de l'échantillonnage – Exercice 10.2 (2)



$$\mu_{\bar{X}} = \mu = 6.00 \quad \text{et} \quad \sigma_{\bar{X}} = \sqrt{\frac{N-n}{N-1}} \frac{\sigma}{\sqrt{n}} = 0.87 \frac{3,29}{\sqrt{5}} = 2.01$$

Statistique inférentielle

Théorie de l'échantillonnage – Distribution de moyenne d'échantillons (1)

- Supposons que l'on extraie d'une population finie de taille N tous les échantillons exhaustifs de taille n. Si l'on désigne par $\mu_{\bar{X}}$ et $\sigma_{\bar{X}}$ la moyenne et l'écart-type de la distribution d'échantillonnage de la moyenne, et respectivement par μ et σ la moyenne et l'écart-type de la population, on
 - $\mu_{\bar{X}} = \mu$ et $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}$
- Si la population est infinie ou si l'échantillonnage est non exhaustif, les résultats précédents se réduisent à
 - $\mu_{\bar{X}} = \mu$ et $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$



Dispersion moins large
si exhaustivité

Statistique inférentielle

Théorie de l'échantillonnage – Distribution de moyenne d'échantillons (2)

Pour une population de taille N et pour l'ensemble des échantillons de taille n , on aura:

	Population	Echantillon non-exhaustif	Echantillon exhaustif
Moyenne	μ	μ	μ
Ecart-type	σ	$\frac{\sigma}{\sqrt{n}}$	$\frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}$

où μ est la moyenne de la population complète et σ est l'écart-type de la population complète.

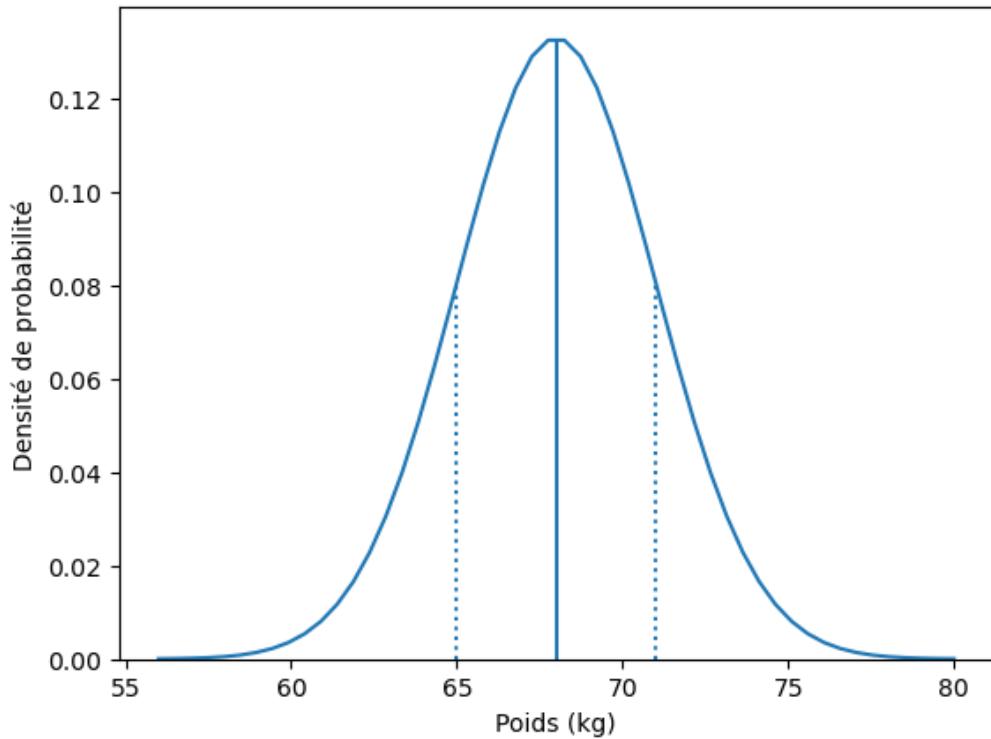
Statistique inférentielle

Théorie de l'échantillonnage – Exercice 11.1 (1)

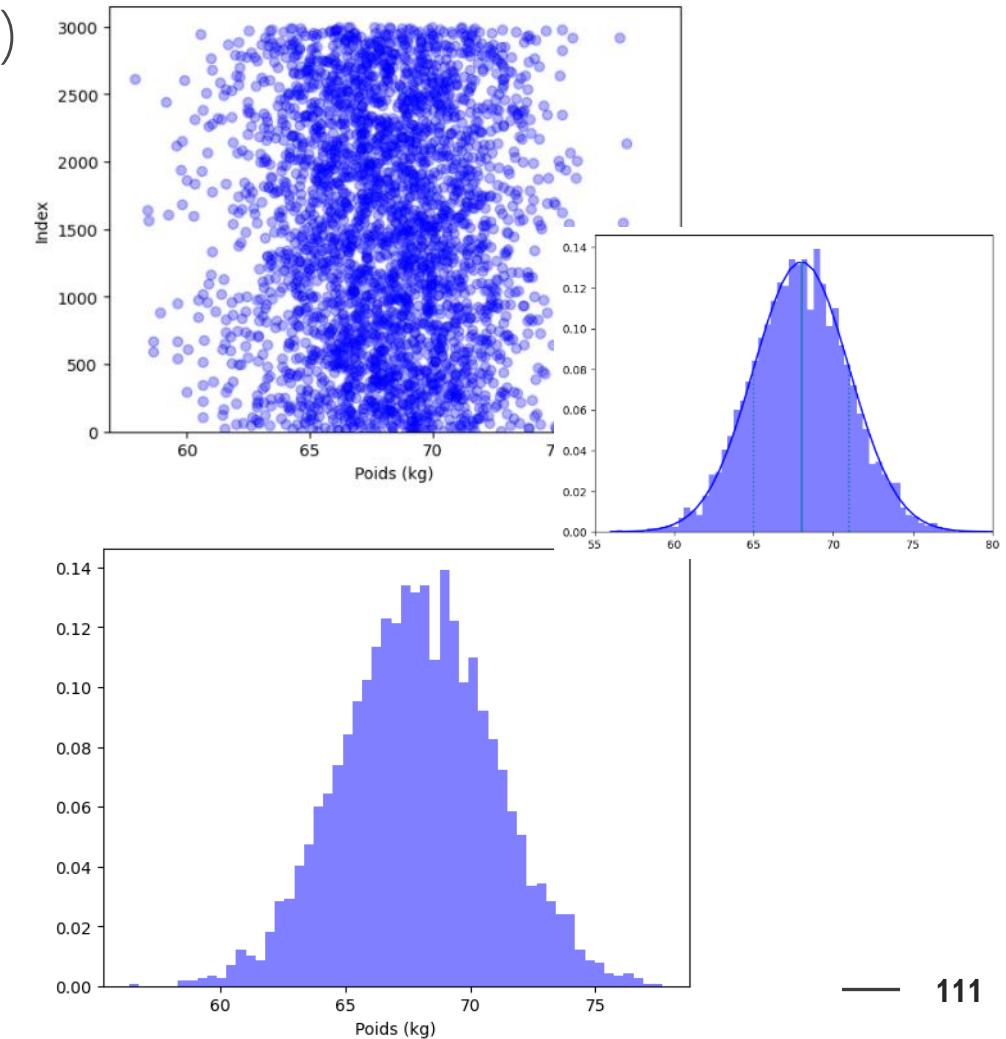
On suppose que les poids de 3000 étudiants d'une université suivent une loi normale de moyenne 68.0 kg et d'écart-type 3.0 kg. Si l'on extrait 80 échantillons de 25 étudiants chacun, quelle est la moyenne et l'écart-type théoriques de la distribution d'échantillonnage des moyennes pour (a) un échantillonnage non exhaustif, (b) un échantillonnage exhaustif ?

Statistique inférentielle

Théorie de l'échantillonnage – Exercice 11.1 (2)



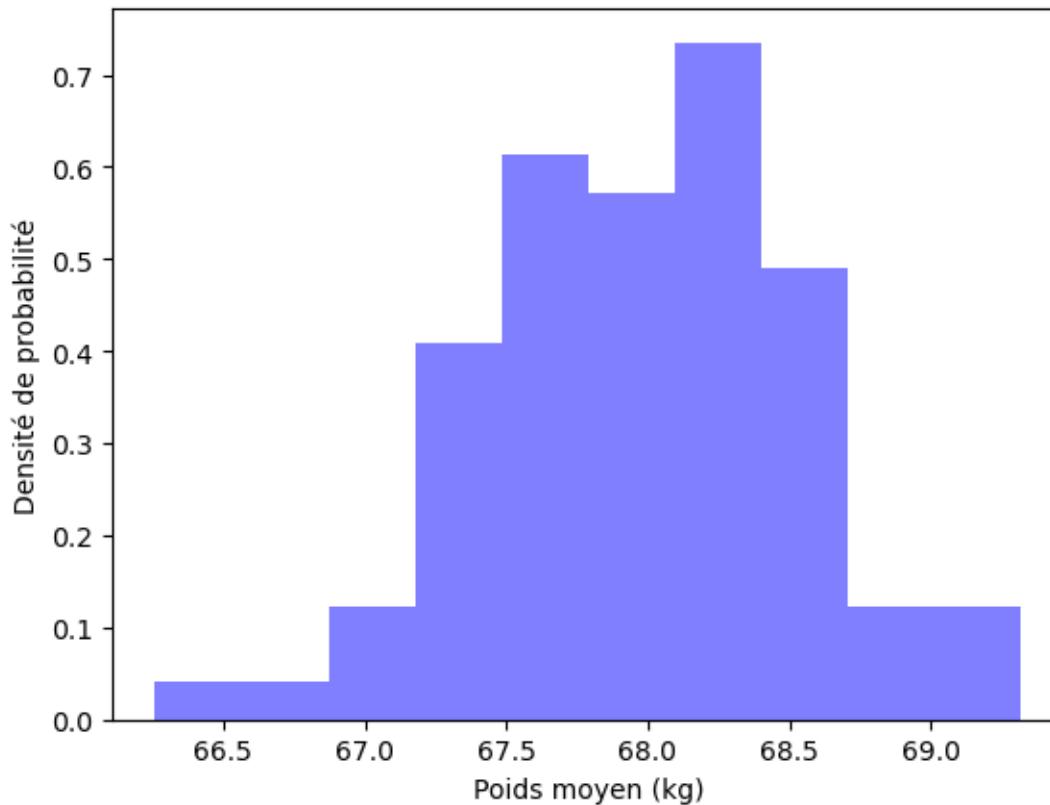
Population globale



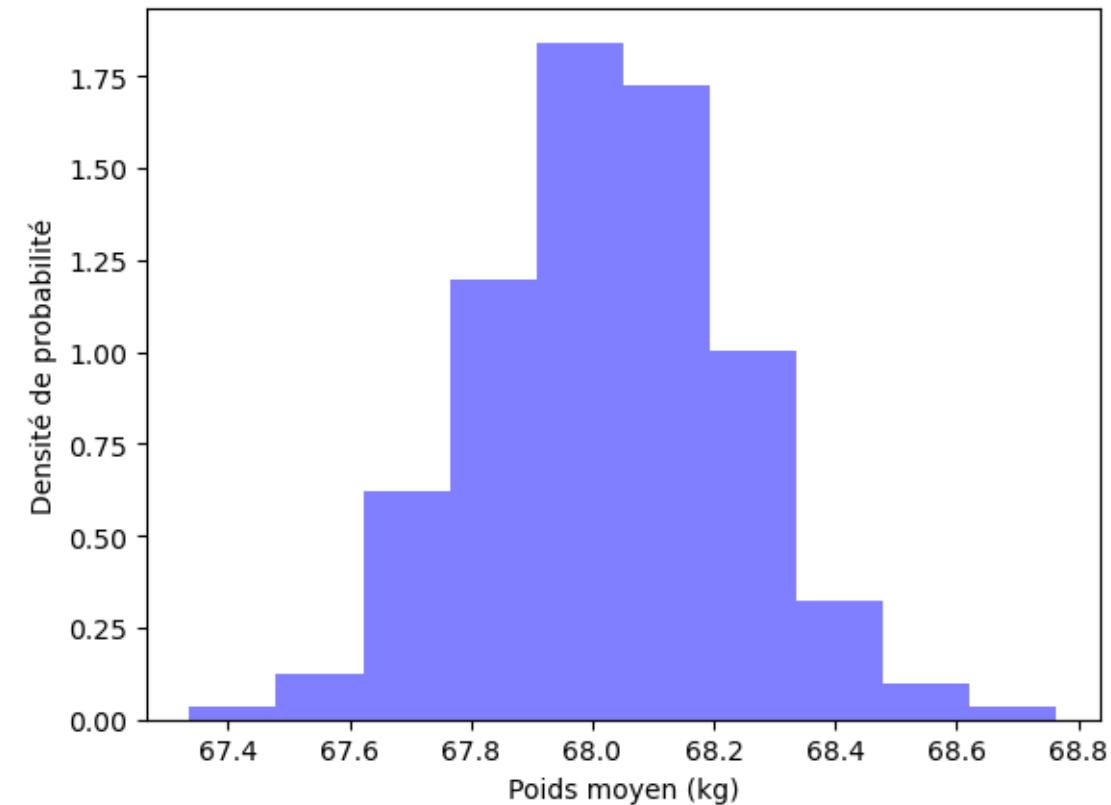
Statistique inférentielle

Théorie de l'échantillonnage – Exercice 11.1.(a & b) (3)

Tirage non-exhaustif



Tirage exhaustif



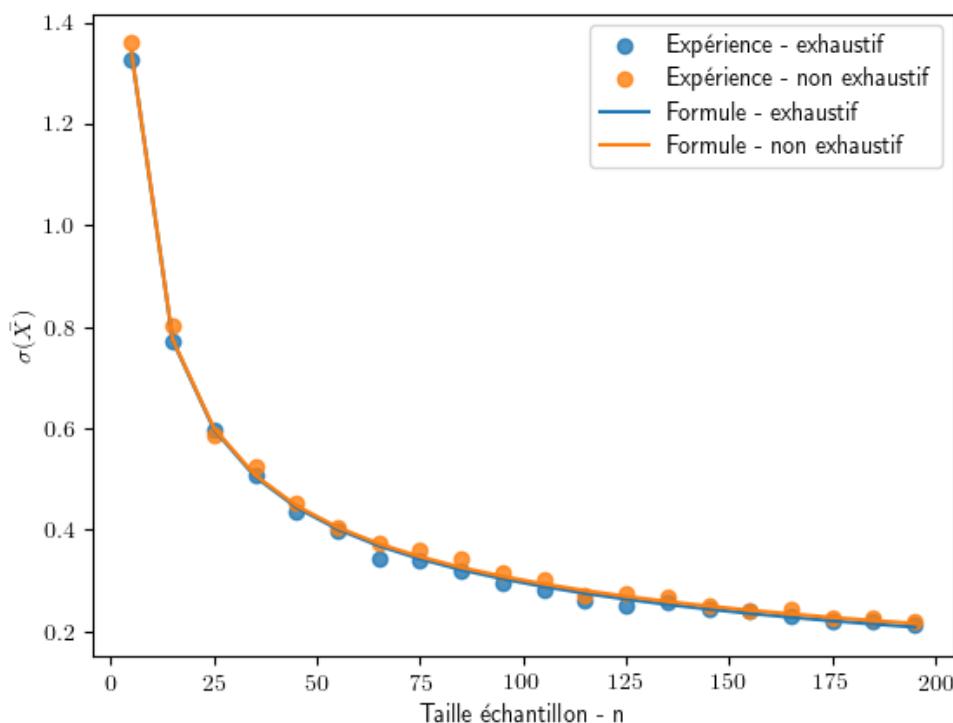
Avec ces valeurs,
on ne peut voir une nette différence entre les
moyennes de poids moyen en tirage non-exhaustif ou
pas

Statistique inférentielle

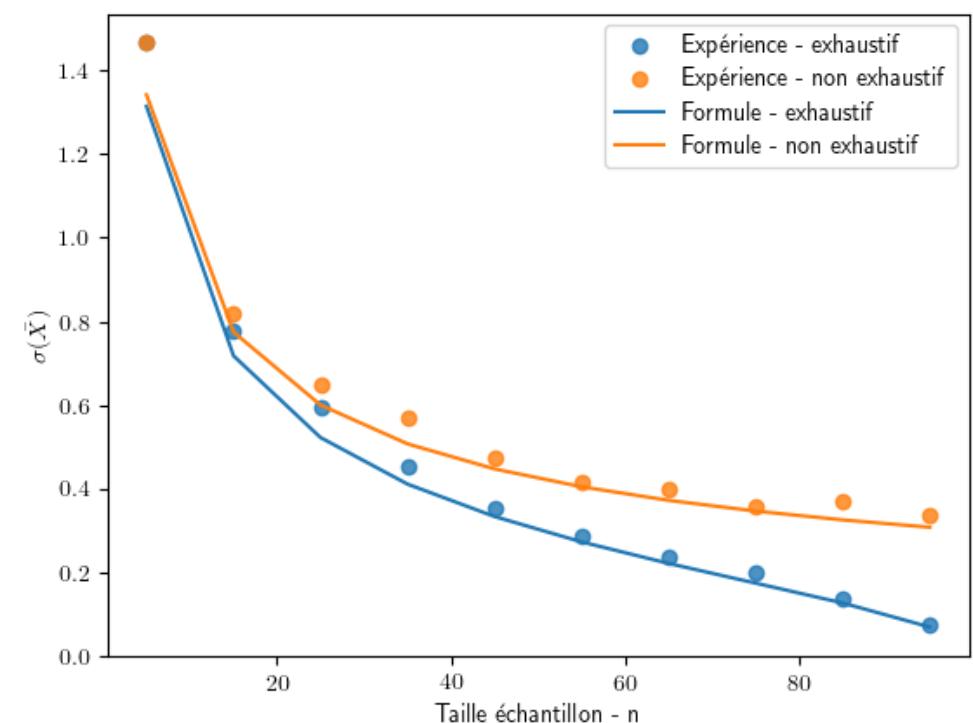
Théorie de l'échantillonnage – Exercice 11.1 (3)

On voit l'influence de la taille d'échantillon lorsque la taille de la population totale, N, est plus petite (et proche de la taille d'échantillon)

N = 3000



N = 100



Statistique inférentielle

Théorie de l'échantillonnage – Exercice 11.2

Pour combien d'échantillons de l'Exercice 11.1 peut-on s'attendre à trouver une moyenne:

- a) Comprise entre 66.8 et 68.3 kg
- b) Inférieure à 66.4 kg

Statistique inférentielle

Théorie de l'échantillonnage – Exercice 11.3

Cinq cents pignons ont un poids moyen de 5,02 gr et un écart-type de 0,30 gr. Trouver la probabilité pour qu'un échantillon de 100 pignons choisis au hasard ait un poids total (a) compris entre 496 et 500 gr, (b) plus grand que 510 gr.

Sommaire

1. Présentation du professeur
2. Introduction à Python
3. Introduction à la statistique
4. Statistique Descriptive
5. Statistique Inférentielle
 - I. Présentation générale
 - II. Théorie de l'échantillonnage
 - III. Théorie de l'estimation**
 - IV. Théorie de la décision – Test d'hypothèse et de signification

Statistique inférentielle

Théorie de l'estimation – Estimation de paramètres

Dans le chapitre précédent, nous avons montré comment la théorie de l'échantillonnage permettait d'obtenir de l'information à partir d'échantillons tirés au hasard dans une population connue.

D'un point de vue pratique, il est souvent plus important de pouvoir obtenir de l'information sur la population globale à partir d'échantillons.

Un des problèmes importants de l'inférence statistique est d'estimer des paramètres d'une population (moyenne, variance de la population, etc.) à partir des statistiques d'échantillonnage correspondantes (moyenne, variance d'un échantillon, etc.).

Statistique inférentielle

Théorie de l'estimation – Estimateurs non-biaisés

Si la moyenne d'une statistique d'échantillonnage est égale au paramètre correspondant de la population, on dit que la statistique est un **estimateur non-biaisé** de ce paramètre.

Dans le cas contraire on dit que l'on a un **estimateur biaisé**.

Ex.1 La moyenne des moyennes d'échantillonnage $\mu_{\bar{X}} = \mu$, où μ est la moyenne de la population. Ainsi, la moyenne d'échantillonnage \bar{X} est un estimateur non biaisé de la moyenne théorique μ de la population

Ex.2 La moyenne de la distribution d'échantillonnage des variances est $\mu_{s^2} = \frac{n-1}{n} \sigma^2$, où σ^2 est la variance de la population et n la taille de l'échantillon. Ainsi, la variance s^2 est un estimateur biaisé de la variance σ^2 de la population. En utilisant la variance modifiée, $\hat{s}^2 = \frac{n}{n-1} s^2$, on trouve $\mu_{\hat{s}^2} = \sigma^2$ de sorte que \hat{s}^2 est un estimateur non biaisé de σ^2 . Notons en passant que \hat{s} reste un estimateur biaisé de σ ! (voir le corrigé de l'Exercice 12.4)

Statistique inférentielle

Théorie de l'estimation – Estimateurs non-biaisés (suite)

En termes plus formels, on dira qu'une statistique est non biaisée si son espérance est égale à la valeur du paramètre de la population correspondant.

Ainsi, \bar{X} et \hat{s}^2 sont **sans biais** puisque $E(\bar{X}) = \mu$ et $E(\hat{s}^2) = \sigma^2$

Statistique inférentielle

Théorie de l'estimation – Efficacité des estimateurs

Quand on désire estimer la moyenne, il se peut que les distributions d'échantillonnage de deux statistiques distinctes aient la même espérance mathématique. La statistique qui a la variance la plus faible est alors considérée comme l'**estimateur le plus efficace**.

Autrement dit, parmi toutes les statistiques possibles dont les distributions d'échantillonnage ont la même moyenne, celle qui a la variance la plus faible est nommée le **meilleur estimateur de la moyenne**.

Statistique inférentielle

Théorie de l'estimation – Efficacité des estimateurs (suite)

Exemple concret: les distributions d'échantillonnage de la moyenne et de la médiane ont toutes les deux la même espérances, à savoir la moyenne de la population. Cependant, la variance de la distribution des moyennes est plus faible que celle de la distribution des médianes. En effet, on peut montrer que:

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} \text{ tandis que } \sigma_{\text{médiane}} = \frac{\sigma}{\sqrt{n}} \sqrt{\frac{\pi}{2}}$$

La moyenne d'échantillonnage donne donc un estimateur de la moyenne de la population plus efficace que la médiane.

On peut même montrer, dans ce cas, que parmi toutes les statistiques estimant la moyenne de la population, l'estimateur le plus efficace est la moyenne d'échantillonnage!

Néanmoins, dans la pratique, on utilise souvent des estimateurs inefficaces à cause de la relative facilité avec laquelle on les obtient.

Statistique inférentielle

Théorie de l'estimation

Exercice 12.1

On a effectué cinq mesures du diamètre d'une sphère qui ont respectivement donné 6.33, 6.37, 6.36, 6.32 et 6.37 cm. Déterminer des estimations sans biais de la moyenne et de la variance.

Exercice 12.2

(Dé)Montrez numériquement que la médiane d'échantillon est bien un estimateur de la moyenne de la population moins efficace que la moyenne d'échantillon ☺

Exercice Supplémentaire 12.3

Lisez l'article *Estimating the Size of a Population* ci-joint

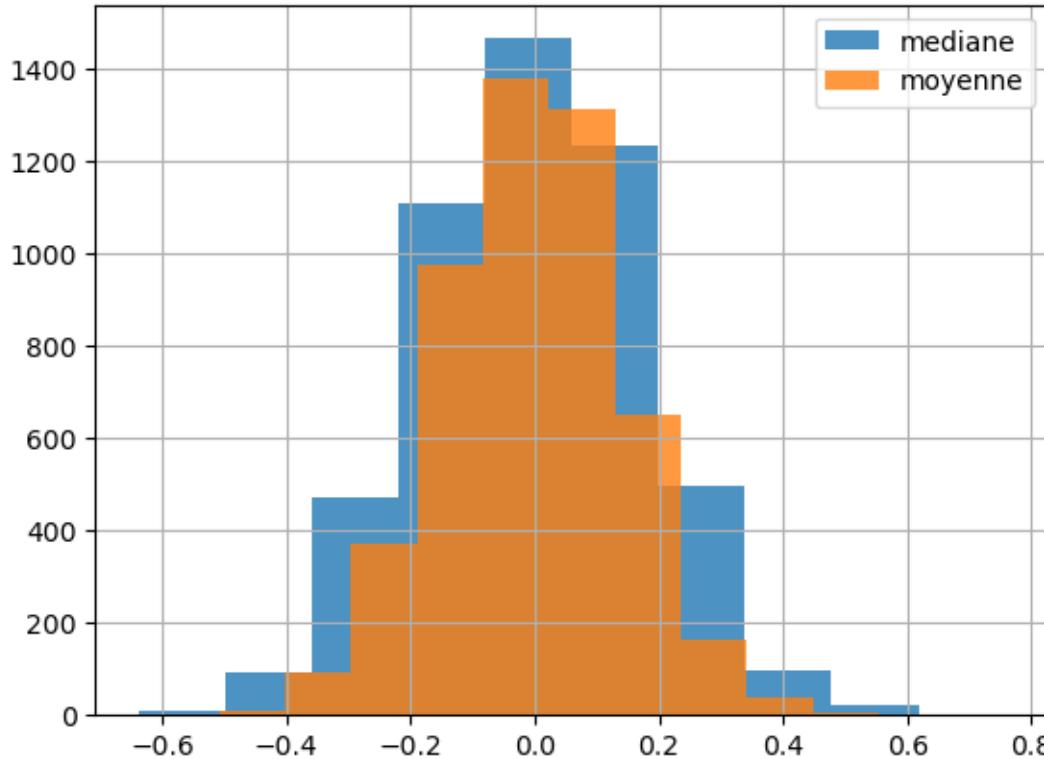


Acrobat
Document

Comparez l'efficacité des 4 estimateurs de taille de population décrits dans l'article.

Statistique inférentielle

Théorie de l'estimation – Exercice 12.2

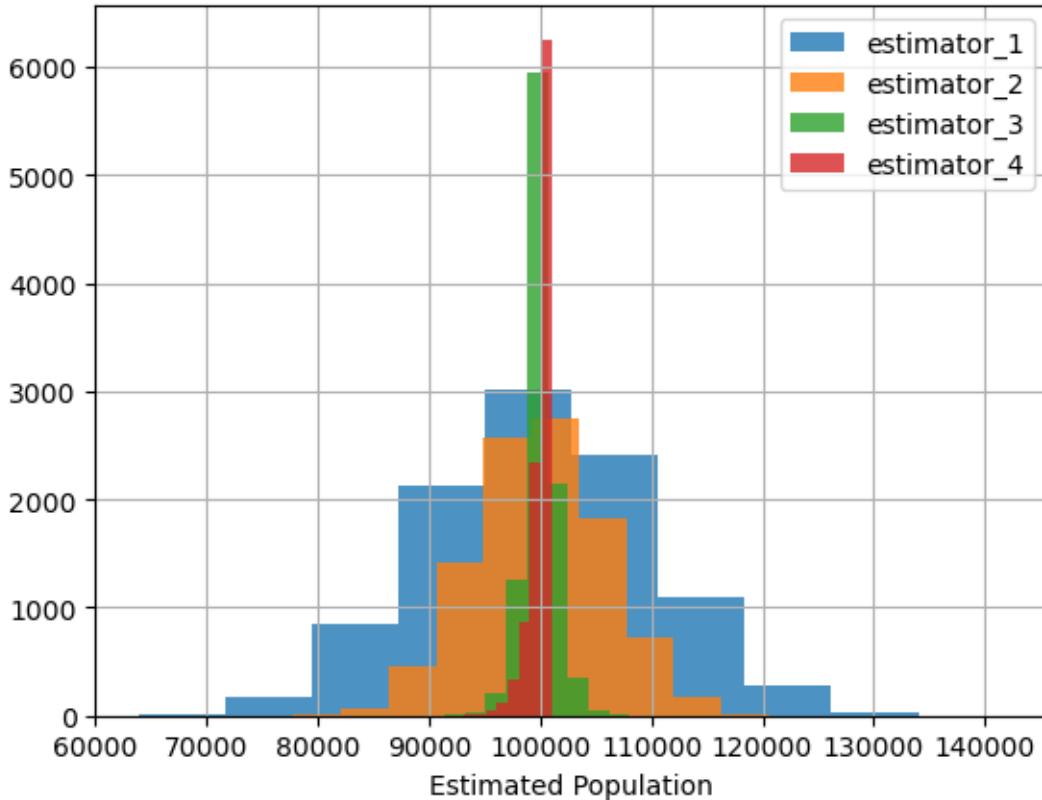


On voit que les histogrammes des moyennes et des médianes sont tous les deux centrés sur la valeur moyenne globale, soit 0.

Par contre, l'histogramme des moyennes est moins dispersé que l'histogramme des médianes et donc l'estimateur moyenne d'échantillon est plus efficace que l'estimateur médiane d'échantillon pour estimer la moyenne de la population globale.

Statistique inférentielle

Théorie de l'estimation – Exercice 12.3



Dans cet exercice, on voit que les 4 estimateurs nous donnent bien une valeur correcte de la taille globale de la population (les histogrammes sont bien tous centrés sur la même valeur proche de la taille totale).

On voit aussi que l'estimateur 4 est le plus efficace, suivi de 3 puis de 2 et enfin 1 puisque les valeurs d'estimation sont de plus en plus dispersées autours de la valeur cible.

Statistique inférentielle

Théorie de l'estimation

Exercice 12.4

(Dé)Montrez numériquement que s^2 est bien un estimateur biaisé de la variance σ^2 de la population alors que la variance modifiée, $\hat{s}^2 = \frac{n}{n-1}s^2$, ne l'est pas.

Exercice 12.5

(Dé)Montrez numériquement que même si $\hat{s}^2 = \frac{n}{n-1}s^2$ est un estimateur non biaisé de la variance σ^2 de la population, $\hat{s} = \sqrt{\frac{n}{n-1}s^2}$ reste un estimateur biaisé de σ !

Sommaire

1. Présentation du professeur
2. Introduction à Python
3. Introduction à la statistique
4. Statistique Descriptive
5. Statistique Inférentielle
 - I. Présentation générale
 - II. Théorie de l'échantillonnage
 - III. Théorie de l'estimation
 - IV. Théorie de la décision – Test d'hypothèse et de signification

Statistique inférentielle

Théorie de la décision – Décision statistique

Dans la pratique, on est souvent appelé à prendre des décisions diverses au sujet d'une population, et ce à partir de l'information que donne un échantillon

On appellera de telles décisions des **décisions statistiques**.

On voudra, par exemple, décider à partir d'un échantillon si un nouveau sérum est effectivement efficace pour guérir une maladie, si une méthode pédagogique est meilleure qu'une autre, si une pièce de monnaie est bien équilibrée, etc.

Statistique inférentielle

Théorie de la décision – Exemple pratique

Tester l'hypothèse qu'une pièce de monnaie soit parfaitement équilibrée quand on adopte la règle de décision suivante:

- 1) On accepte l'hypothèse si le nombre de faces dans un seul échantillon de 100 jets est compris entre 40 et 60, 40 et 60 compris
 - 2) Sinon on rejette l'hypothèse.
-
- a) Calculer la probabilité de rejeter l'hypothèse quand elle est vraie.
 - b) Interpréter graphiquement la règle de décision et le résultat de (a).
 - c) Quelles conclusions peut-on tirer si l'échantillon de 100 jets comprend 53 faces?
 - d) Peut-on se tromper dans les conclusions de (c)? Expliquer.

Statistique inférentielle

Théorie de la décision – Hypothèses statistiques

Pour parvenir à une décision, il est commode de faire des hypothèses sur la population correspondante.

De telles hypothèses peuvent être vraies ou fausses, ce sont des **hypothèses statistiques**.

Elles sont en général des affirmations relatives à la distribution de probabilité de la population.

Statistique inférentielle

Théorie de la décision – Hypothèses statistiques (suite)

Dans de nombreux cas, on formulera une hypothèse statistique dans le seul but de la rejeter ou de l'annuler.

Une telle hypothèse est une **hypothèse nulle**, généralement désignée par H_0 .

Si l'on veut par exemple décider qu'une pièce de monnaie est déséquilibrée, on supposera qu'elle est parfaite, c'est-à-dire que la probabilité de face est $p=0.5$. De la même façon, si l'on veut décider qu'un procédé est meilleur qu'un autre, on supposera qu'il n'y a aucune différence entre les procédés, ce qui veut dire que toutes les différences observées sont purement et simplement dues à des fluctuations d'échantillonnage dans la même population.

Statistique inférentielle

Théorie de la décision – Hypothèses statistiques (fin)

Toute hypothèse qui diffère d'une hypothèse donnée est une **hypothèse alternative**, généralement désignée par H_1 .

Ainsi, si on a l'hypothèse de base, H_0 , $p=0.5$

$p=0.7$, $p\neq0.5$ ou $p>0.5$ sont toutes des hypothèses alternatives possibles (possibles H_1)

Statistique inférentielle

Théorie de la décision – Test d'hypothèses et de signification

Imaginons que, sous une hypothèse particulière supposée vraie, l'on ait trouvé que les observations d'un échantillon aléatoire diffèrent sensiblement des observations espérées sous l'hypothèse du hasard pur relevant de la théorie de l'échantillonnage.

On dira alors que les **différences observées** sont **significatives** et l'on sera enclin à rejeter l'hypothèse (ou au moins à ne pas l'accepter à partir de la 'preuve' obtenue).

Si par exemple, 20 parties de pile ou face donnent 16 fois face, on aura tendance à rejeter l'hypothèse que la pièce est bien équilibrée, bien qu'il soit aussi concevable que l'on ait tort.

On appelle **tests d'hypothèses**, **test de signification** ou **règles de décision**, les procédés qui permettent de décider si des hypothèses sont vraies ou fausses ou de déterminer si des échantillons observés diffèrent significativement des résultats supposés.

Statistique inférentielle

Théorie de la décision – Typage des erreurs de décision

Erreur de 1^{ière} espèce: rejet d'une hypothèse qui devait être acceptée

Erreur de 2^{ième} espèce: acceptation d'une hypothèse qui devait être rejetée

Chacun de ces cas correspond à une décision erronée ou bien à une erreur de jugement!

Pour qu'un test d'hypothèse ou une règle de décision soit efficace, ceux-ci doivent être conçus de manière à minimiser les erreurs de décision.

Ceci n'est pas simple car, pour un échantillon de taille donnée, la décroissance d'un type d'erreur est en général accompagnée par la croissance de l'autre type d'erreur.

Statistique inférentielle

Théorie de la décision – Typage des erreurs de décision (suite)

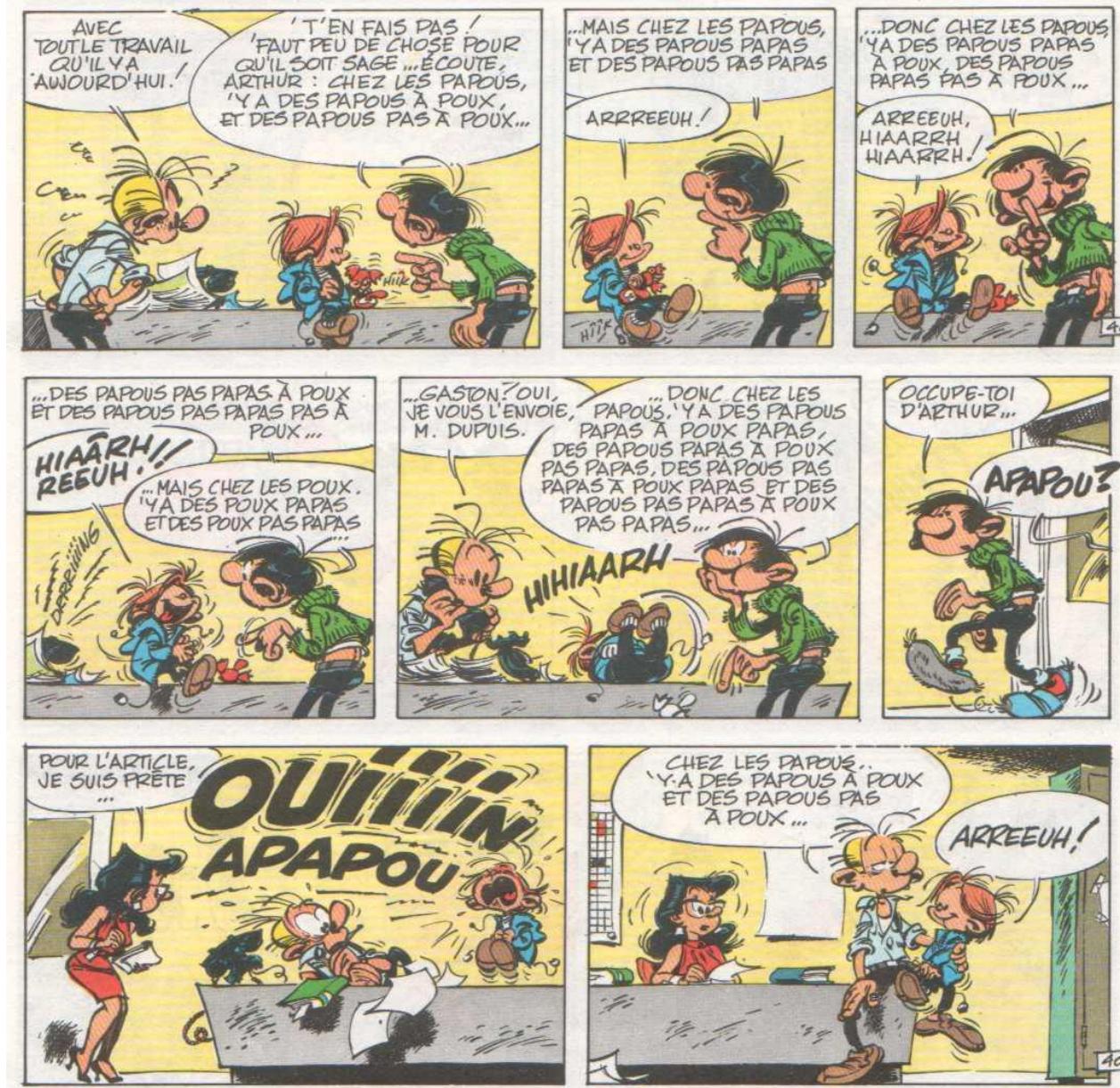
Dans la pratique, un type d'erreur peut être plus important que l'autre.

Il faut alors trouver un compromis afin de limiter l'erreur la plus importante.

Noter que la seule façon de réduire les deux types d'erreur à la fois est d'augmenter la taille de l'échantillon, ce qui n'est pas toujours possible...

Statistique inférentielle

Théorie de la décision



Statistique inférentielle

Théorie de la décision – Niveau de signification

Lorsque l'on teste une hypothèse, la probabilité avec laquelle on est disposé à risquer une erreur de 1^{ière} espèce est appelée **seuil de signification du test**, qui se note α .

Noter que ce seuil est en général spécifié **avant** d'extraire tous les échantillons, de façon que les résultats obtenus n'influencent pas notre choix.

Par exemple, si l'on choisit 5% comme seuil de signification en construisant un test d'hypothèse, il y a alors 5 chances sur 100 pour que l'on rejette l'hypothèse quand elle doit être acceptée. Cela signifie que l'on est sûr à 95% d'avoir pris la bonne décision. On dit alors que l'on a *rejeté l'hypothèse à un seuil de signification égal à 5%*, ce qui signifie que l'on peut avoir tort avec une probabilité de 5%.

Statistique inférentielle

Théorie de la décision – Exercice 13.1

Tester l'hypothèse qu'une pièce de monnaie soit parfaitement équilibrée quand on adopte la règle de décision suivante:

- 1) On accepte l'hypothèse si le nombre de faces dans un seul échantillon de 100 jets est compris entre 40 et 60.
 - 2) Sinon on rejette l'hypothèse.
-
- a) Calculer la probabilité de rejeter l'hypothèse quand elle est vraie.
 - b) Interpréter graphiquement la règle de décision et le résultat de (a).
 - c) Quelles conclusions peut-on tirer si l'échantillon de 100 jets comprend 53 faces?
 - d) Quelles conclusions peut-on tirer si l'échantillon de 100 jets comprend 60 faces?
 - e) Peut-on se tromper dans les conclusions de (c) et (d) ? Expliquer.

Statistique inférentielle

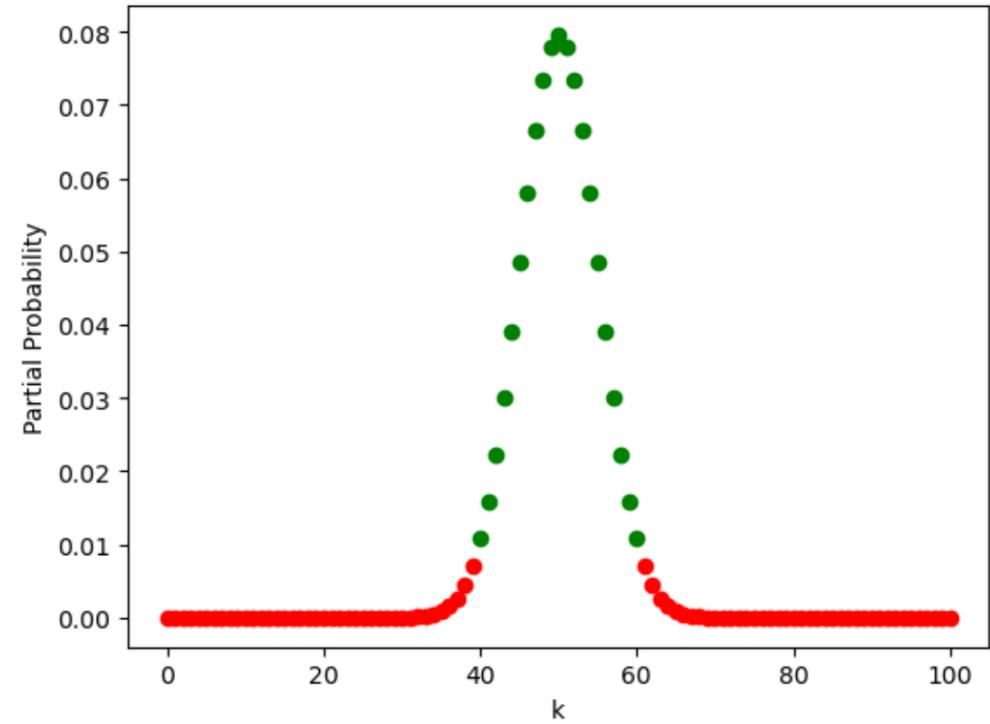
Théorie de la décision – Exercice 13.1 – Résolution commentée

On partira de la loi de distribution binomiale (slide 61), adaptée à un échantillon de taille n . Pour k succès sur les n lancers d'une pièce avec une probabilité p de tomber sur face, on aura les probabilités suivantes:

$$prob(p, n, k) = C_n^k p^k q^{n-k} = \frac{n!}{k!(n-k)!} p^k q^{n-k}$$

On peut alors reprendre toutes ces probabilités (partielles) dans le diagramme à droite.

La probabilité de rejet alors que l'hypothèse est vraie sera donnée par la somme de toutes probabilités partielles correspondant à des points verts pour k dans $[40;60]$, soit: 0.0352 ou 3,52%



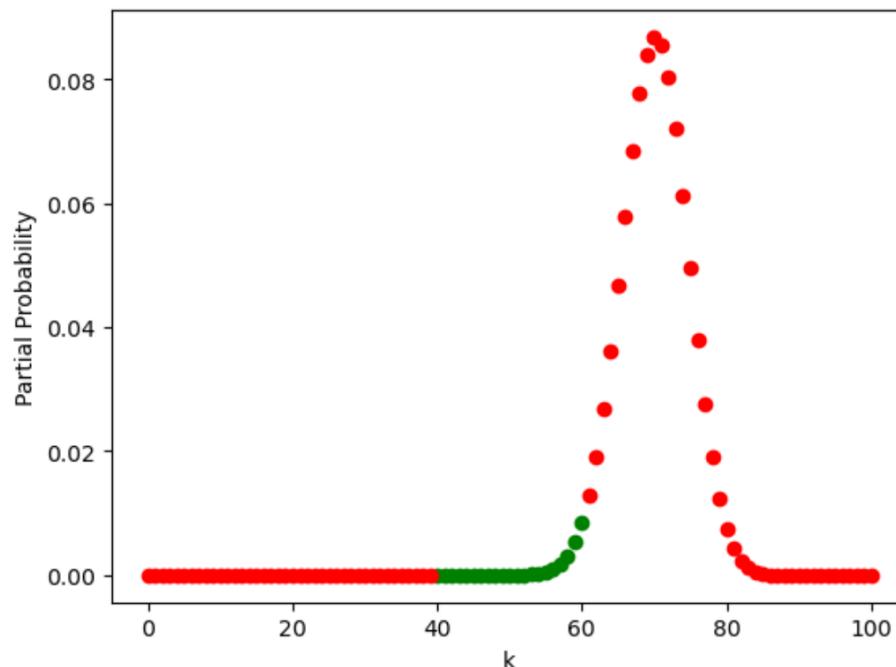
Statistique inférentielle

Théorie de la décision – Exercice 13.2

Sur base de l'énoncé de l'exercice 13.1, quelle serait la probabilité d'accepter l'hypothèse que la pièce soit équilibrée alors qu'elle est réellement biaisée et caractérisée par $p=0.7$?

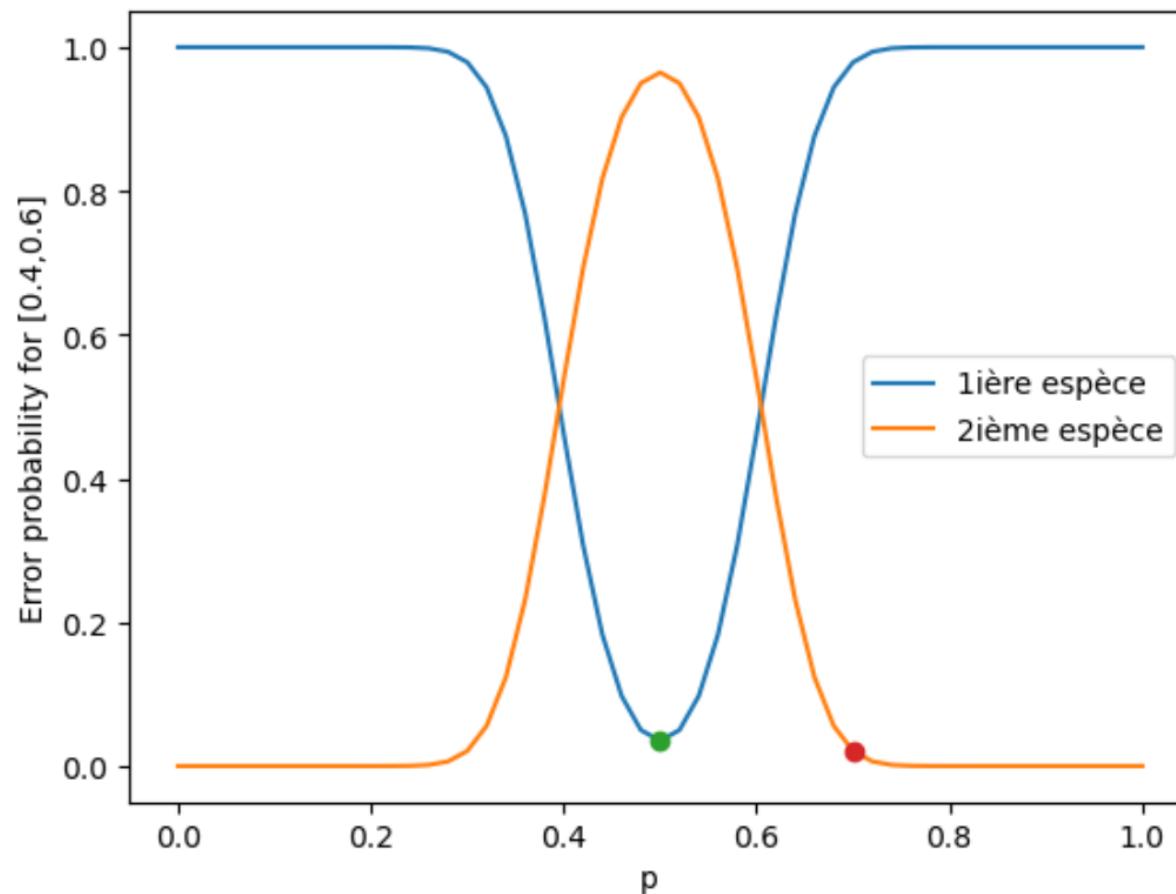
Ici, on calculera une distribution binomiale pour $p=0.7$ et on sommera toutes les probabilités partielles correspondant aux points verts de l'intervalle défini dans le test, soit k dans $[40; 60]$.

Soit une probabilité d'erreur de 2^{ième} espèce de l'ordre de 0.0210 ou 2.10%



Statistique inférentielle

Théorie de la décision – Taux d'erreur de première ou de deuxième espèce



Point bleu correspond à l'exercice 13.1 : erreur de première espèce de 3,52%

Point rouge correspond à l'exercice 13.2 : erreur de deuxième espèce de 2,10%

Statistique inférentielle

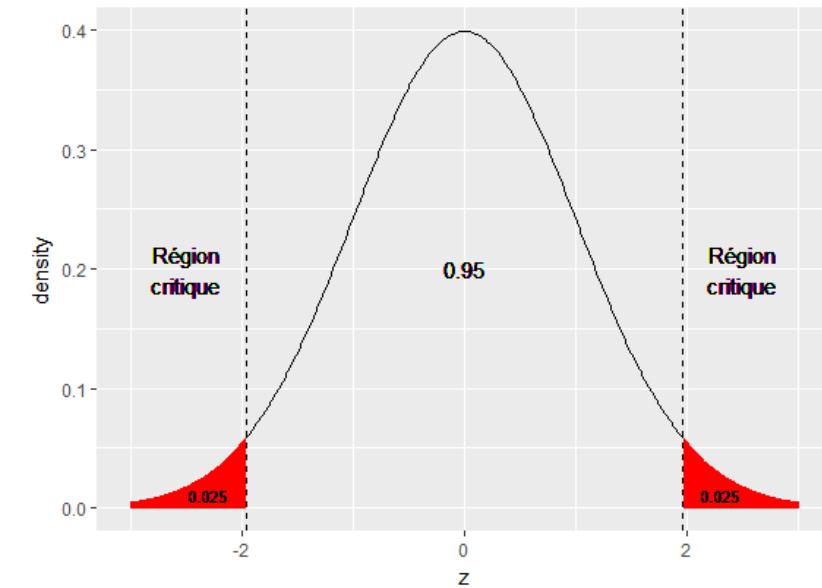
Théorie de la décision – Tests relatifs à la distribution normale

Afin d'illustrer les idées précédentes, supposons que, sous une hypothèse donnée, la distribution d'échantillonnage d'une statistique S soit une loi normale de moyenne μ_S et d'écart-type σ_S . Alors, la distribution de la variable centrée réduite

$$z = (S - \mu_S)/\sigma_S$$

est la distribution normale centrée réduite de moyenne 0 et d'écart-type 1 représentée ci-contre.

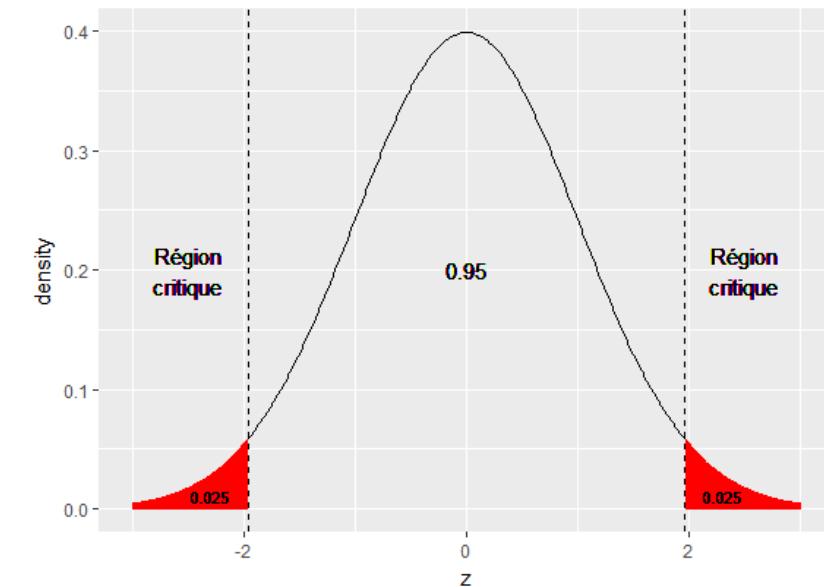
Pour des valeurs de z s'étendant entre -1.96 et +1.96, on peut être confiant à 95% (l'hypothèse est vraie).



Statistique inférentielle

Théorie de la décision – Tests relatifs à la distribution normale (suite)

Cependant, si en choisissant un seul échantillon au hasard on trouve que le résultat z de cette statistique se situe à l'extérieur de l'intervalle $[-1.96, +1.96]$, on doit conclure que cet événement peut se produire avec une probabilité égale seulement à 5% (l'aire rouge sur la figure) dans le cadre de l'hypothèse initiale. On dira alors que z diffère significativement de ce qui était espéré sous l'hypothèse initiale et que l'on doit donc rejeter celle-ci.



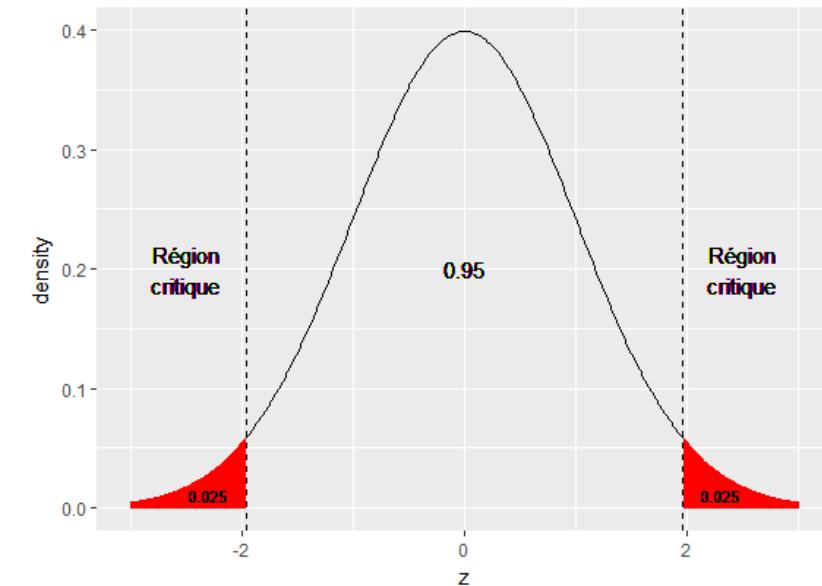
Statistique inférentielle

Théorie de la décision – Tests relatifs à la distribution normale (fin)

L'aire totale 0.05 (en rouge) est le seuil de signification du test. Elle représente la probabilité pour que l'on ait tort en rejetant l'hypothèse, c'est-à-dire la probabilité d'avoir une erreur de 1^{ière} espèce.

L'ensemble des valeurs de z qui sont à l'extérieur de l'intervalle $[-1.96, +1.96]$ constitue ce que l'on appelle la **région critique** ou la **région de rejet de l'hypothèse**, ou encore la **région significative**.

L'ensemble des valeurs de z qui sont à l'intérieur de ce même intervalle pourrait être appelé la **région d'acceptation** de l'hypothèse ou la **région de non-signification**.



Statistique inférentielle

Théorie de la décision – Test unilatéral vs test bilatéral

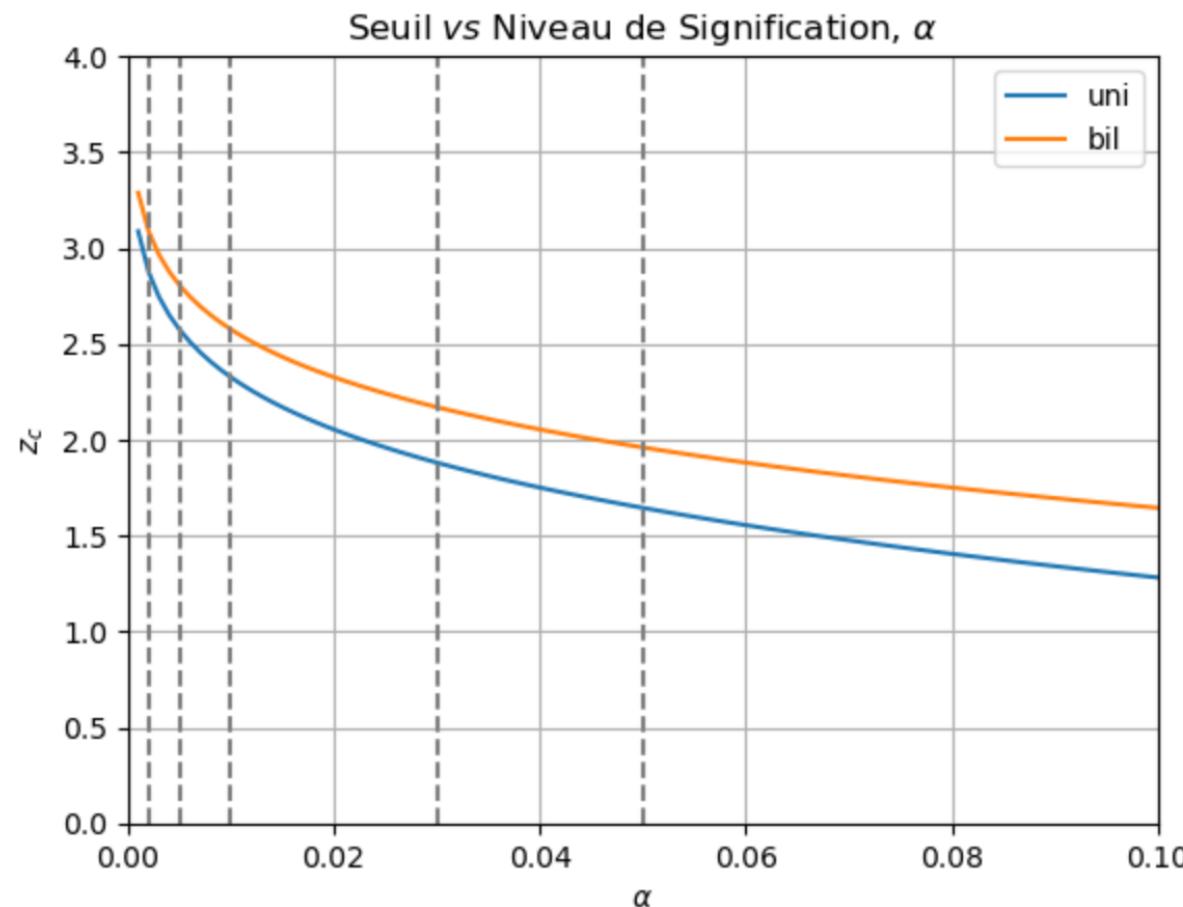
Jusqu'à présent, on s'est surtout occupé des valeurs extrêmes de la statistique S , ou de ses valeurs de part et d'autre de la moyenne, c'est-à-dire encore aux deux extrémités de la distribution. C'est pourquoi les tests correspondants sont appelés **tests bilatéraux**.

Bien souvent, cependant, on pourra ne s'intéresser qu'à l'une des extrémités de la distribution. C'est ce que l'on fera notamment pour tester si un procédé de fabrication est meilleur qu'un autre. De tels test sont des **tests unilatéraux**. La région critique est alors située d'un côté seulement de la distribution, avec une aire égale au seuil de signification.

Niveau de signification α	0.10	0.05	0.03	0.01	0.005	0.002
Valeur critiques de z pour les tests unilatéraux	-1.28 ou +1.28	-1.645 ou +1.645	-1.89 ou +1.89	-2.33 ou +2.33	-2.58 ou +2.58	-2.88 ou +2.88
Valeur critiques de z pour les tests bilatéraux	-1.645 et +1.645	-1.96 et +1.96	-2.17 ou +2.17	-2.58 et +2.58	-2.81 et +2.81	-3.08 et +3.08

Statistique inférentielle

Théorie de la décision – Test unilatéral vs test bilatéral – Niveau de Signification



Statistique inférentielle

Théorie de la décision – Tests particuliers: moyenne

Ici $S = \bar{X}$ est la moyenne de l'échantillon ; $\mu_S = \mu_{\bar{X}} = \mu$, la moyenne de la population ; $\sigma_S = \sigma_{\bar{X}} = \sigma/\sqrt{n}$, l'écart-type correspondant, σ étant l'écart-type de la population et n la taille de l'échantillon. La variable z est alors

$$z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

Quand cela est nécessaire, on utilisera l'écart-type expérimental s ou \hat{s} pour estimer σ .

Statistique inférentielle

Théorie de la décision – Exercice 13.3

Dans une usine, une machine doit être réparée si elle produit plus de 10% d'unités défectueuses par jour.

Un échantillon aléatoire de 100 unités provenant d'une journée de production contient 15 unités défectueuses et le superviseur affirme que la machine doit être réparée.

Les éléments de preuve confirment-ils cette décision ? Utilisez un seuil de signification de 1%.

Statistique inférentielle

Théorie de la décision – Exercice 13.4

Un vice-président en charge des ventes pour une grosse entreprise affirme que les vendeurs ont en moyenne 15 contacts par semaine ou moins.

Pour vérifier ses déclarations, 36 vendeurs sont choisis aléatoirement. On note le nombre de contacts qu'ils ont eus lors d'une semaine sélectionnée aléatoirement.

La moyenne et la variance corrigée de ces 36 mesures sont 17 et 9 respectivement.

Les éléments de preuves contredisent-ils les affirmations du vice-président ? (seuil de signification de 0.05)

Statistique inférentielle

Théorie de la décision – Exercice 13.5

Les charges de rupture de câbles produits par une fabrique ont une valeur moyenne de 1800 kg et un écart-type de 100 kg.

On affirme que la charge de rupture peut être augmentée par une technique nouvelle du procédé de fabrication.

Pour tester cette affirmation, on a testé un échantillon de 50 câbles produits par la technique nouvelle et l'on a trouvé une charge de rupture moyenne de 1850 kg.

Peut-on admettre l'affirmation précédente avec un seuil de signification de 0.01 ?

Statistique inférentielle

Théorie de la décision – Exercice 13.6

On considère que le poids des chiots bergers allemands à la naissance suit une distribution normale, de moyenne 0,150 kg et de variance 0,015.

On suspecte cependant que les chiennes diabétiques mettent au monde des chiots qui ont en moyenne un poids inférieur à 0,150 kg.

Afin de vérifier cette hypothèse, on a relevé le poids de 25 chiots bergers allemands nés de mères diabétiques et le poids moyen observé a été de 0,125 kg.

Est-ce que cette expérience confirme l'hypothèse initiale (la suspicion) avec un seuil de signification de 3% ?

Statistique inférentielle

Théorie de la décision – Exercice 13.7

Même énoncé que pour l'exercice 13.6 sauf qu'ici la moyenne des poids observée est de 0.200 kg et que le seuil de signification choisi est de 5%.

Statistique inférentielle

Théorie de la décision – Exercice 13.8

La durée de vie moyenne d'un échantillon de 100 ampoules fluorescentes fabriquées par une usine est estimée à 1570 heures avec un écart-type de 120 heures. Si μ est la durée de vie moyenne de toutes les ampoules produites par l'usine, tester l'hypothèse $\mu = 1600$ heures avec l'hypothèse $\mu \neq 1600$ heures, en choisissant un niveau de signification (a) 0.05 (b) de 0.01

Statistique inférentielle

Théorie de la décision – Exercice 13.9

Même énoncé que pour l'exercice 13.8 mais cette fois tester l'hypothèse $H_0: \mu = 1600$ heures contre l'hypothèse $H_1: \mu < 1600$ heures, en choisissant un niveau de signification (a) 0.05 (b) de 0.01

Statistique inférentielle

Distribution de Student

Jusqu'à maintenant, nous sommes toujours partis du principe, un peu artificiel, que nous connaissons a priori l'écart-type de la population globale.

Généralement, cette information n'est pas connue.

Que pourrions-nous alors utiliser ?

Pourquoi ne pas utiliser la variance corrigée ?

Dans ce cas, à quoi ressemble la distribution de cette

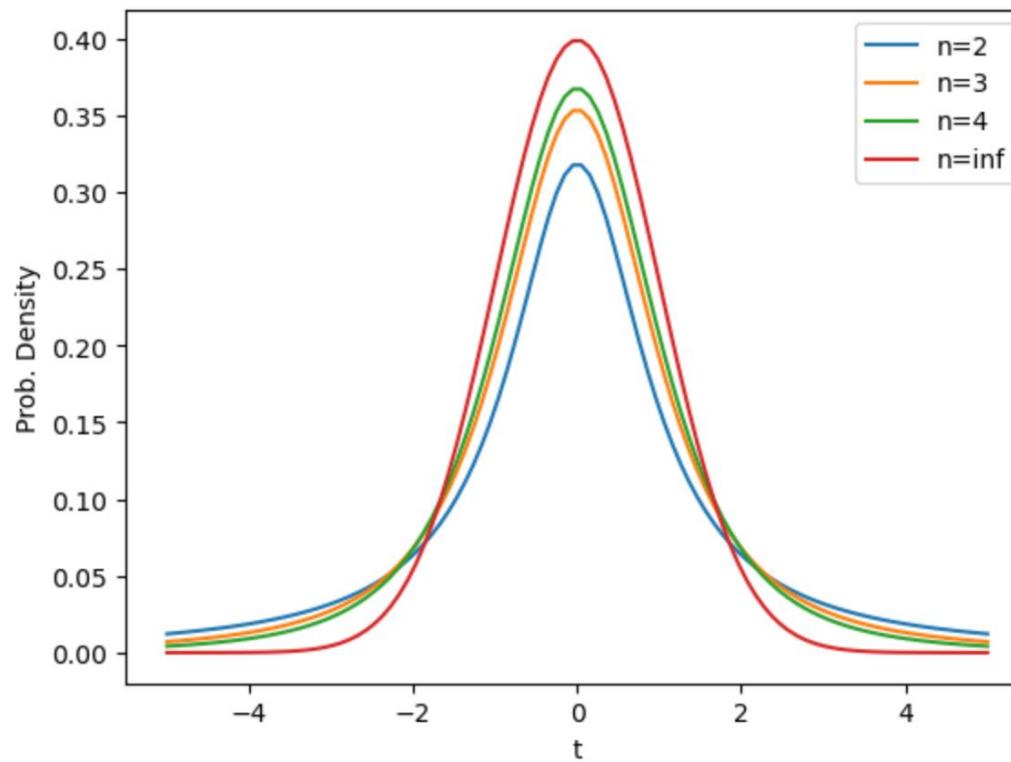
Statistique inférentielle

Distribution de Student

$$T = \frac{\bar{X} - \mu}{S / \sqrt{n}}$$

où

$$S^2 = \frac{1}{n-1} \sum_{i=1}^n (\bar{X} - X_i)^2$$

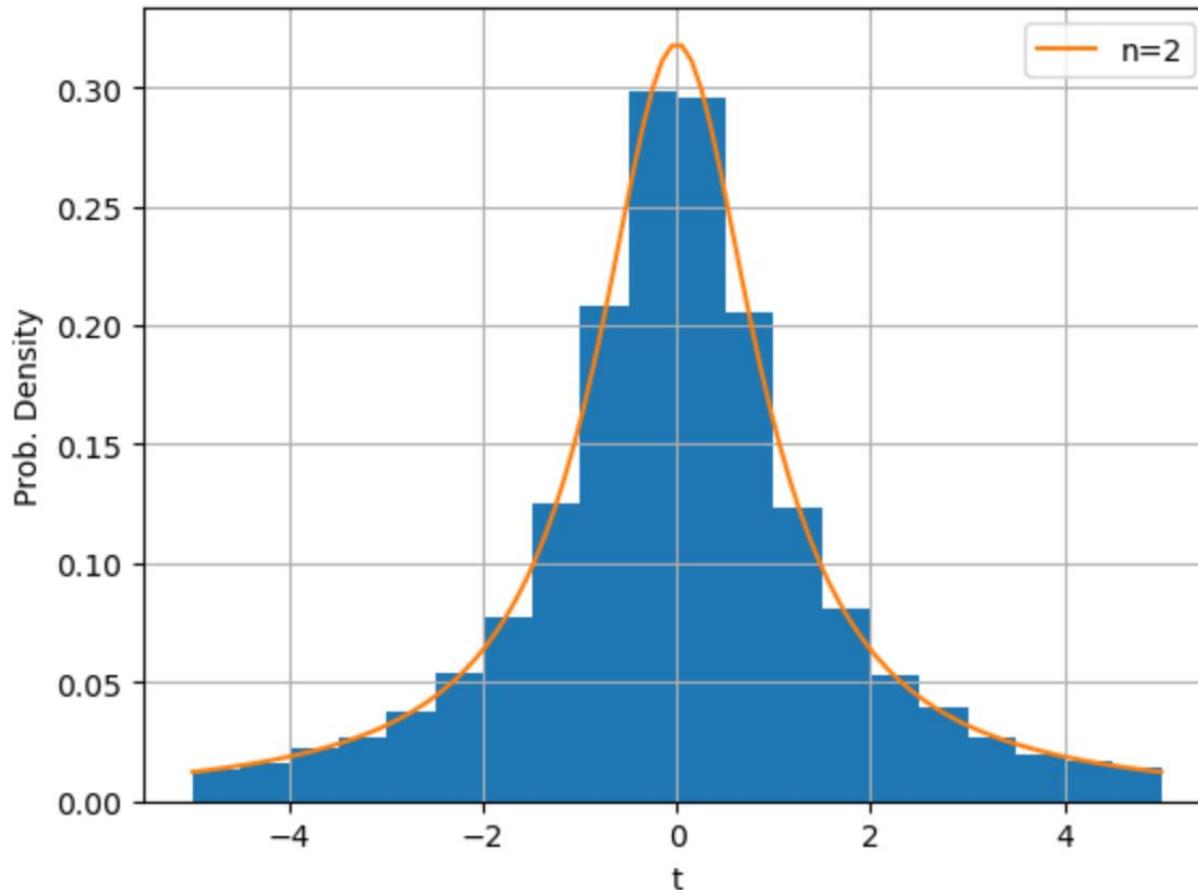


Statistique inférentielle

Distribution de Student

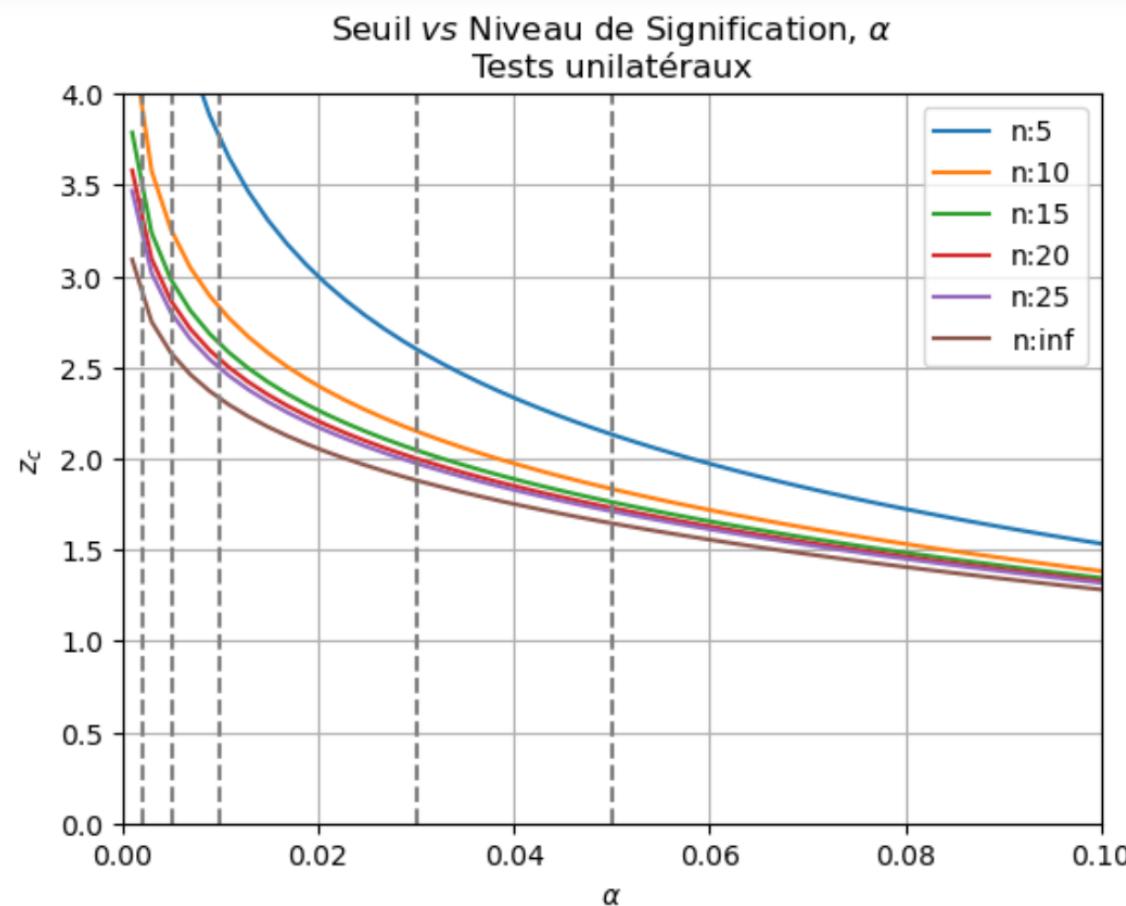
$$f_T(t) = \frac{1}{\sqrt{k\pi}} \frac{\Gamma\left(\frac{k+1}{2}\right)}{\Gamma\left(\frac{k}{2}\right)} \left(1 + \frac{t^2}{k}\right)^{-\frac{k+1}{2}}$$

où $k = n-1$



Statistique inférentielle

Distribution de Student



Statistique inférentielle

Distribution de Student

Test unilatéraux – valeur absolue de la valeur critique – Pour différentes tailles d'échantillon

Niveau de signification α	0.10	0.05	0.03	0.01	0.005	0.002
n = 2	3.08	6.31	10.58	31.82	63.66	159.15
n = 5	1.53	2.13	2.60	3.75	4.60	5.95
n = 10	1.38	1.83	2.15	2.82	3.25	3.83
...						
n = inf	1.28	1.645	1.89	2.33	2.58	2.88