Statistiques

CONSULTING BSTORM

1. Introduction

Bienvenue dans ce cours de statistiques pour les sciences des données.

On abordera les définitions et concepts importants en statistiques, en particulier pour un.e

- "data analyst"
- et/ou "data scientist«
- et/ou "machine learning engineer".

Et ce, avec le moins de formules de math possibles.

1. Introduction

La statistique est l'étude des données, qui sont souvent des mesures de phénomènes dans le monde réel.

L'objectif de la statistique est d'analyser ces données afin de découvrir des tendances, des corrélations et des modèles qui peuvent aider à comprendre le monde dans lequel nous vivons.

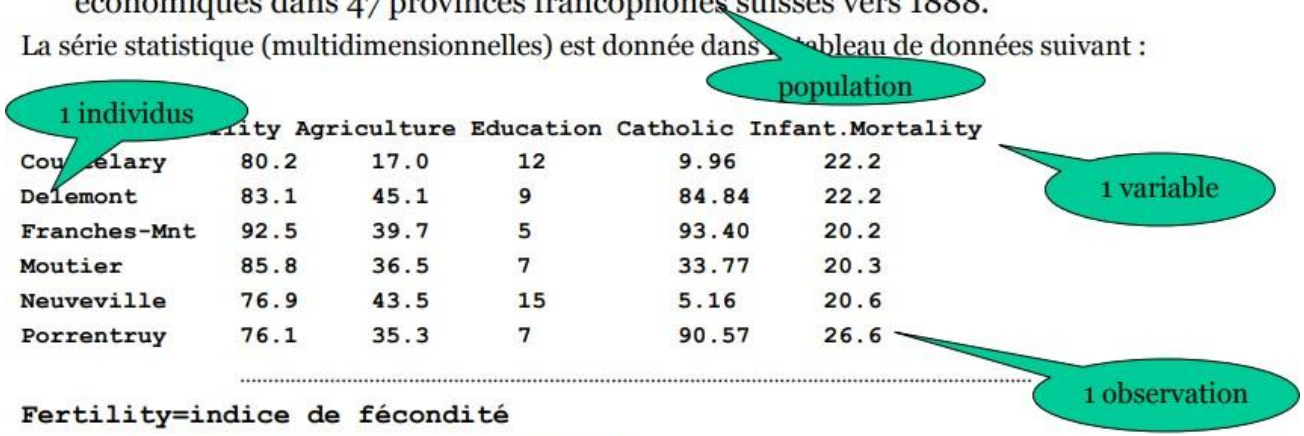
En bref, le but est de **synthétiser** les données de phénomènes **pour comprendre et prédire** ces phénomènes.

1. Introduction

- **Statistique :** branche des mathématiques dont l'objectif est d'analyer, structurer et modéliser des données.
- Population: est constitué de l'ensemble des individus objets de l'étude
- Echantillon: est un groupe d'individu extrait de la population
- **Variable**: une caractéristique commune à l'ensemble des individus d'une étude. La valeur de cette caractéristique varie entre les individus.

Exemple 2 : On s'intéresse à la fécondité en relation avec certains indicateurs socio-économiques dans 47 provinces francophones suisses vers 1888.

La série statistique (multidimensionnelles) est donnée dans le tableau de données suivant :



	Fertility	Agriculture	Education	Catholic	Infant.Mortality
Courtelary	80.2	17.0	12	9.96	22.2
Delemont	83.1	45.1	9	84.84	22.2
Franches-Mnt	92.5	39.7	5	93.40	20.2
Moutier	85.8	36.5	7	33.77	20.3
Neuveville	76.9	43.5	15	5.16	20.6
Porrentruy	76.1	35.3	7	90.57	26.6

Fertility=indice de fécondité

Agriculture= % de males agriculteurs

Education= % d'individus ayant étudié après le primaire

Catholic=% de catholiques

Infant.Mortality=% mortalité infantile

1. Introduction

Variables statistiques: un ensemble de critères qui décrient les individus d'une population

- Variable statistique quantitative est une variable associée à un caractère mesurable
  - variable quantitative discrète prend un nombre limité de valeurs entières
  - variable quantitative continue est une variable qui peut prendre toutes les valeurs dans un intervalle donné

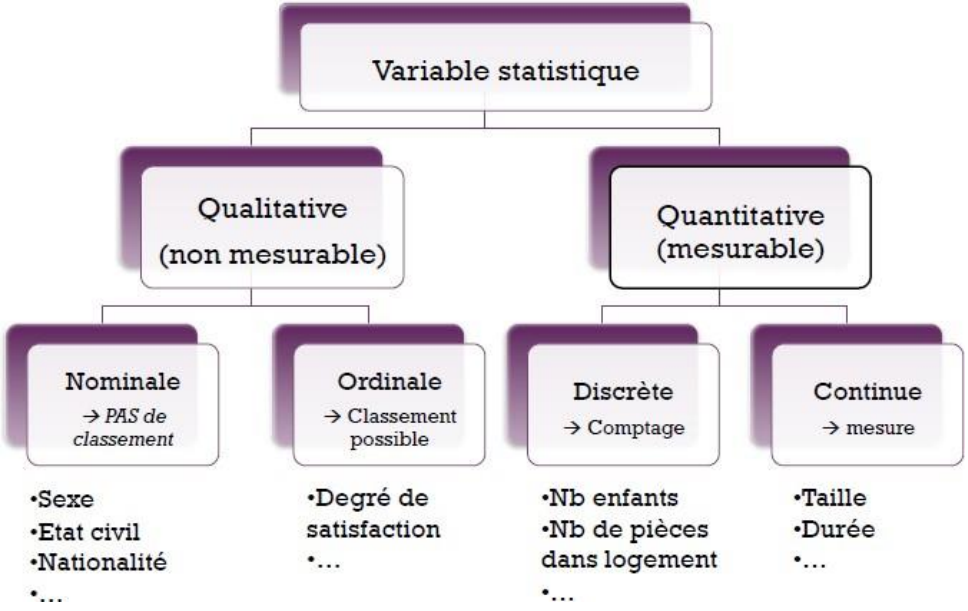
- Variable statistique qualitative est une variable associée à un critère qui n'est pas mesurable
  - Variable qualitative nominale est une variable dont les modalités ne peuvent pas être classées selon un ordre préétabli
  - Variable qualitative ordinale est une variable dont les modalités peuvent être classées

- Déterminer le type des variables également de déduire des méthodes statistiques à appliquer

Exemples

- Identifier le type de variables:
  - Catégorie socioprofessionnelle (ouvrier, agent, ingénieur etc.)
  - Sexe (F,M)
  - Prix pour un repas
  - Jugement (très satisfait, insatisfait, satisfait)
  - Nombre d'enfants
  - Note de satisfaction (1,2,3,…,10)

1. Introduction

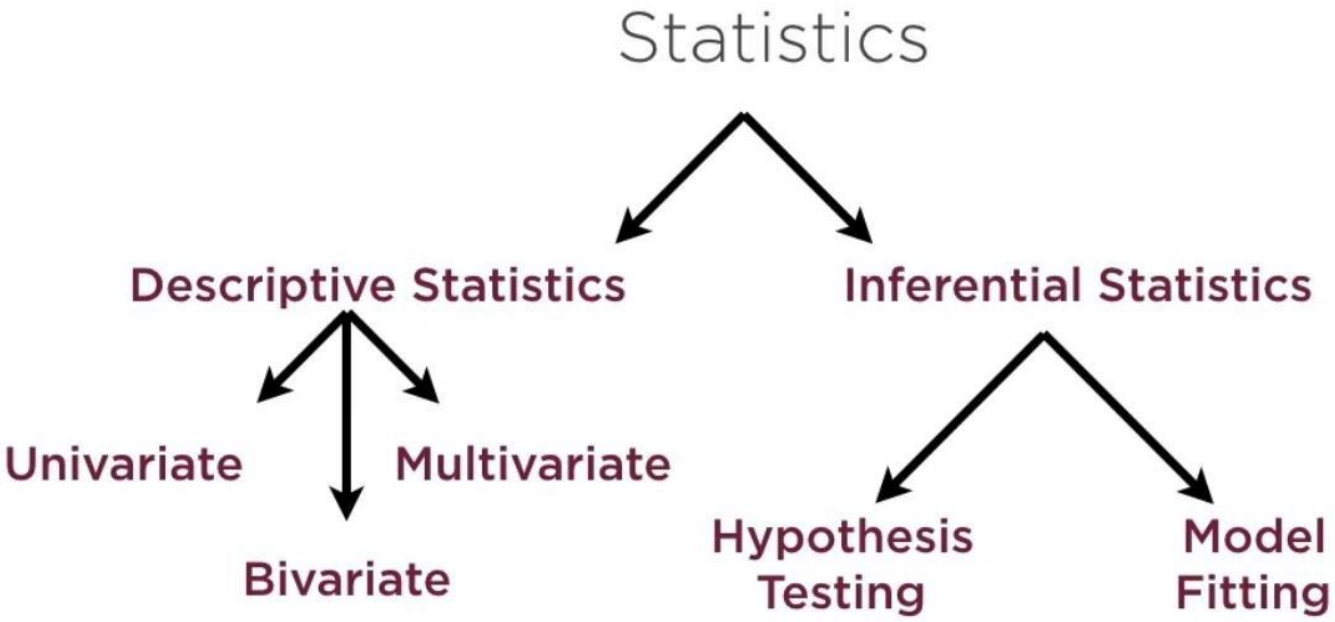


1. Introduction

- Statistique descriptive: a pour objet de résumer et de représenter l'information contenue dans les données sur un groupe d'individus
    - Statistique descriptive univariée fournit les outils statistiques pour organiser, présenter et synthétiser l'information issu de l'analyse d'une variable indépendamment des autres
    - Statistique descriptive bivariée a pour objet d'étudier conjointement deux variables X et Y sur une même population
    - Statistique multivariée vise à étudier plusieurs variables simultanément
- Statistique inférentielle consiste à décrire la population à partir d'observations faites sur l'échantillon. Les caractéristiques inconnues d'une population sont déduites à partir d'un échantillon issu de cette population.

Statistique

1. Introduction



Statistiques descriptives

CONSULTING BSTORM

2. Statistiques descriptives

- Intérêts:
  - Résume et décrit les données (via des mesures et des graphiques visuels)
  - Permet de détecter de potentiel outlier (valeurs aberrantes)
  - Prépare au data cleaning
  - Prépare aux modèles inférentiels
- Type de mesure
  - Fréquence
  - Tendance centrale
  - Dispersion
- Type de graphique
  - Graphique en barre, histogramme, boite de dispersion, etc.

2.1. Mesures de fréquences

- Un effectif (F) : nombre d'apparition d'une modalité pour une variable donnée
- Fréquence (f) : fréquence d'apparition d'une modalité pour une variable donnée (fi = Fi / N)
- Distribution marginale: somme des effectifs pour une colonne ou ligne particulière
- Fréquence conjointe: fréquence se rapportant au croisement de deux variables
- Fréquence marginale: somme des fréquences pour une colonne
- Fréquence conditionnelle: fréquence pour une colonne ou une ligne particulière

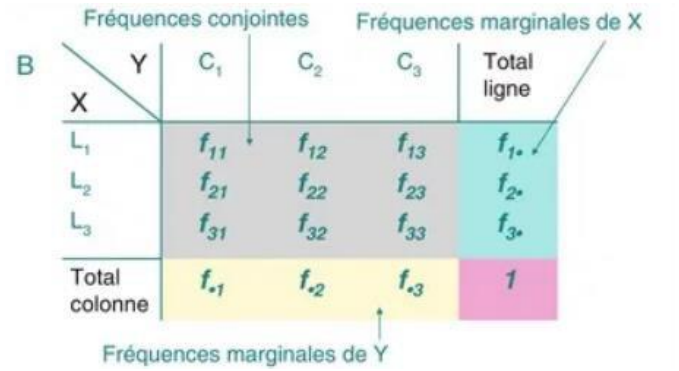
A

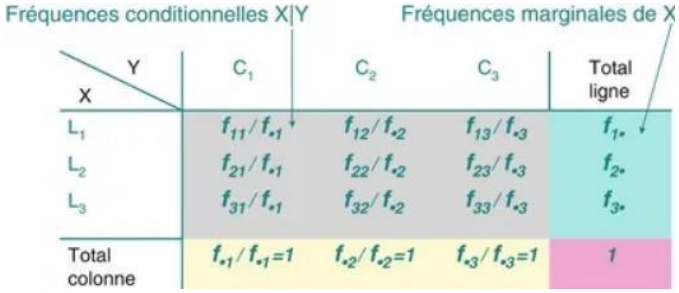
Sexe	Effectif	Fréquence
F	4	44/7 = 0,57
M	3	3/7 = 0,43

B

Âge	Effectif	Effectif cumulé	Fréquence	Fréquence cumulée	Densité
[15,20[	3	3	3/7 = 0,43	0,43	0,43/5 = 0,086
[20,30[	3	6	3/7 = 0,43	0,86	0,43/10 = 0,043
[30,40[	1	7	1/7 = 0,14	1,00	0,14/10 = 0,014

2.1. Mesures de fréquences





A
Effectif

Sexe \ Âge	[15, 20[	[20, 30[	[30, 40[	Total ligne
F	2	2	0	4
M	1	1	1	3
Total colonne	3	3	1	7

B
Fréquence

Sexe \ Âge	[15, 20[	[20, 30[	[30, 40[	Total ligne
F	2/7	2/7	0	4/7
M	1/7	1/7	1/7	3/7
Total colonne	3/7	3/7	1/7	7/7

C
Sexe|Âge

Sexe \ Âge	[15, 20[	[20, 30[	[30, 40[	Marginale sexe
F	2/3	2/3	0/1	4/7
M	1/3	1/3	1/1	3/7
Total colonne	3/3	3/3	1/1	7/7

2.2. Mesures de tendance centrale

- Mode : détermine la valeur la plus fréquente dans un échantillon

Exemple: des votes pour les différents candidats

Alice	Bob	Charles	Denise	Edgar	Fred
60	20	10	40	50	30

60

2.2. Mesures de tendance centrale

- Moyenne arithmétique: somme des valeurs observées d'une variable divisée par le nombre de valeurs observées
    - Sensible aux valeurs extrêmes

Data	60	20	10	40	50	30

$$\bar{x} = \frac{\sum x_i}{n} = \frac{60 + 20 + 10 + 40 + 50 + 30}{6}$$

35

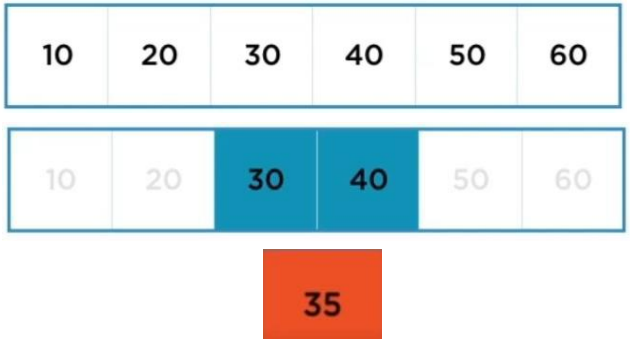
Data	60	20	10	40	50	30	1000

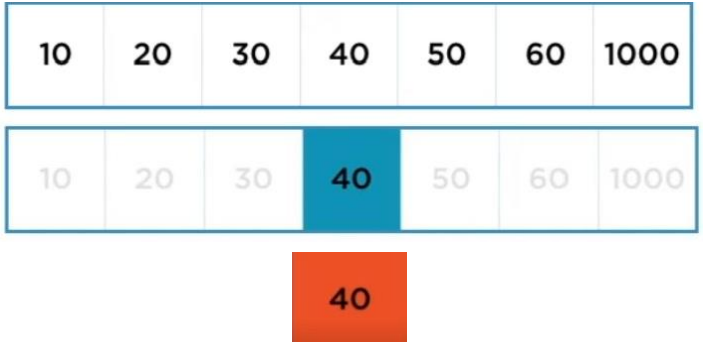
$$\bar{x} = \frac{\sum x_i}{n} = \frac{60 + 20 + 10 + 40 + 50 + 30 + 1000}{7}$$

172.85

2.2. Mesures de tendance centrale

- Médiane : est la valeur telle que 50% des observations de l'échantillon lui sont inférieures ou supérieures.
    - Robuste aux valeurs extrêmes

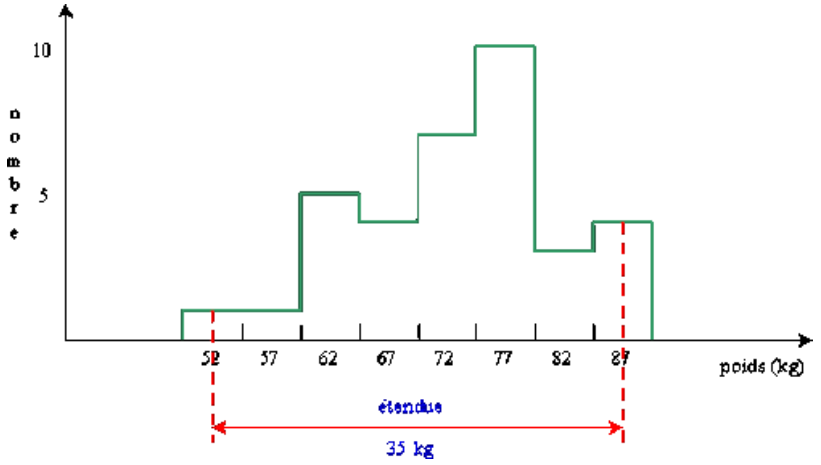




2.3. Mesures de dispersions

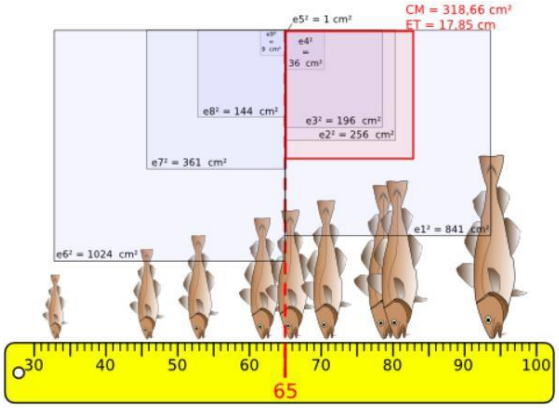
- Amplitude - Etendue (range): la différence entre la valeur la plus grande et la valeur la plus petite

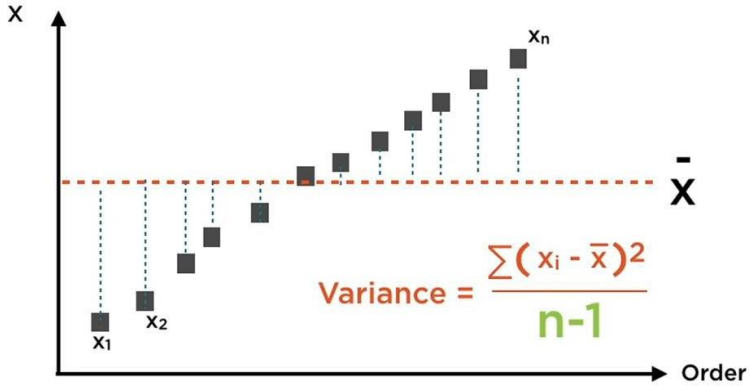
$$Range = X_{max} - X_{min}$$



2.3. Mesures de dispersions

- Variance : est le reflet numérique de la dispersion des valeurs autour de la moyenne. Elle est obtenue à partir des écarts des valeurs (plus particulièrement la *somme des carrés des écarts*) par rapport à la moyenne



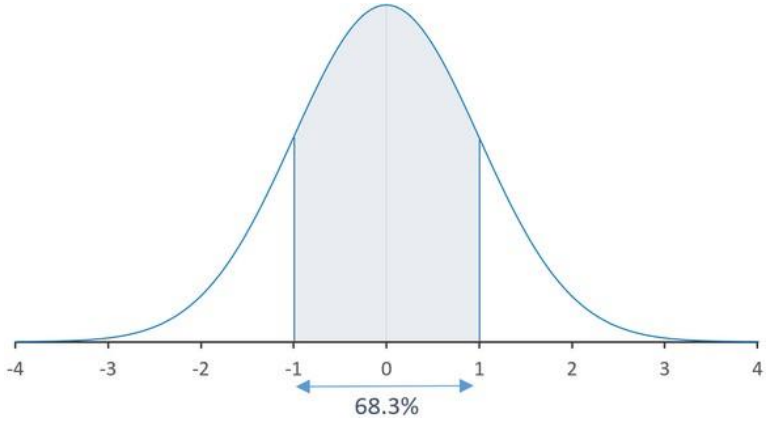


2.3. Mesure de dispersions

- Ecart-type (standard deviation): écart quadratique moyen des valeurs par rapport à la moyenne c'est-à-dire racine carrée de la variance

$$\text{Std Dev} = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n-1}}$$

- entre le o et ±1σ vous aurez **68.3**% de vos observations.
- entre le o et ±2σ vous aurez **95.4**% de vos observations.
- entre le o et ±3σ vous aurez **99.7**.% de vos observations.



2.3. Mesures de dispersions

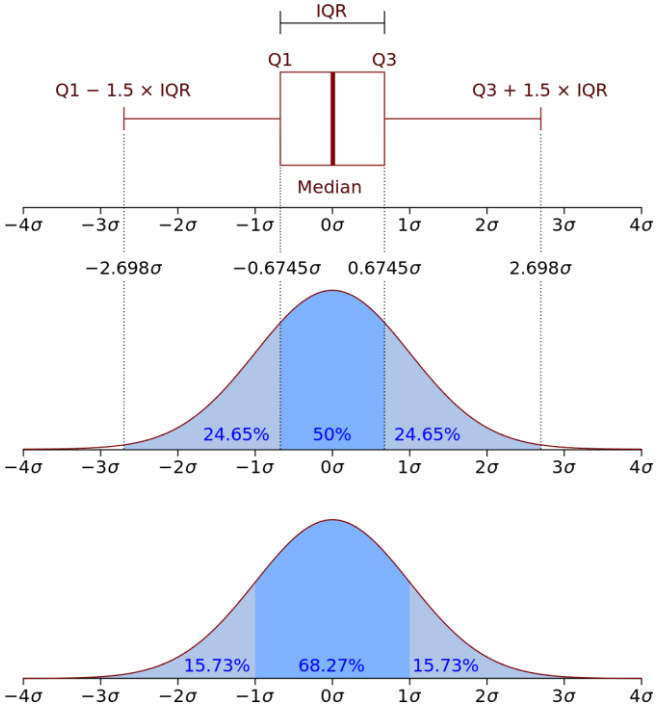
- Quartile: est chacune des trois valeurs qui divisent les données triées en quatre parts égales
    - le 1er quartile (Q1): sépare les observations tel que 25% de celles-ci sont plus petites que le 1er quartile et 75% plus grande
    - le 2e quartile (Q2): sépare les observations tel que 50% de celles-ci sont plus petites que le 2er quartile et 50% plus grande
    - le 3e quartile (Q3): sépare les observations tel que 75% de celles-ci sont plus petites que le 3er quartile et 25% plus grande

- Intervalle Interquartile (IQR): est la différence entre le Q3 et le Q1

2.3. Mesures de dispersions

- Décile: même principe que pour le quartile mais par tranche de 10%
- Percentile: même principe que pour le quartile mais par tranche de 1%
- Coefficient de variation: mesure relative de la dispersion σ/μ

2.3. Mesures de dispersions



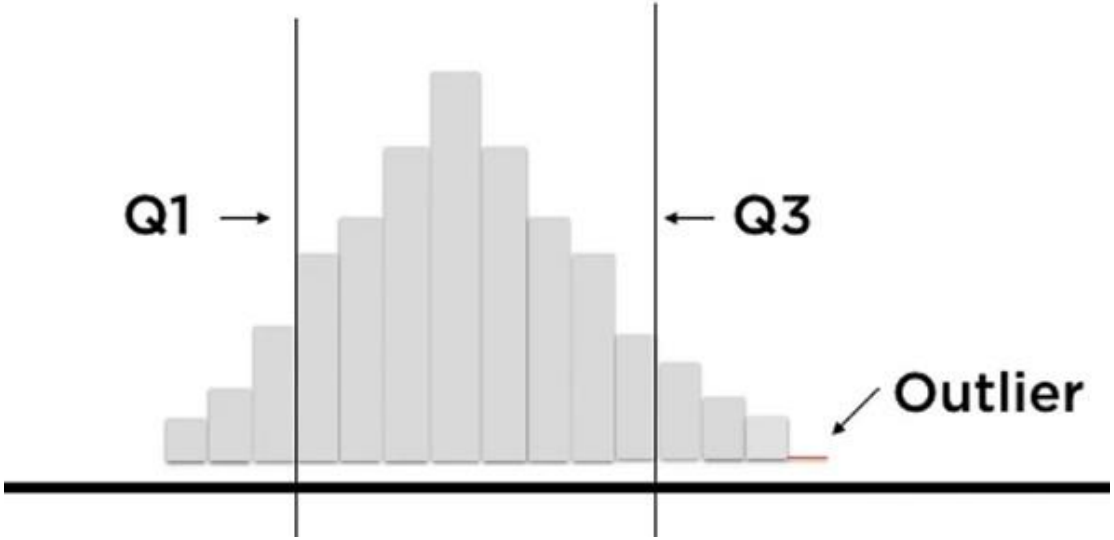
2.4. Détection des Outliers

Une valeur aberrante est une donnée qui s'écarte de façon marquée de l'ensemble des autres données. Une règle pratiques utilisées pour identifier une valeur aberrante est la règle de Tuckey,

Règle de J.tuckey: une donnée peut-être appelée valeur aberrante si elle s'écarte d'une distance d'au moins 1,5x au-dessus du Q3 ou en dessous du Q1.

Une valeur aberrante doit être examinée avec soin pour identifier la cause d'éventuelles de cet écart important par rapport à l'ensemble des données.

2.4. Détection des Outliers

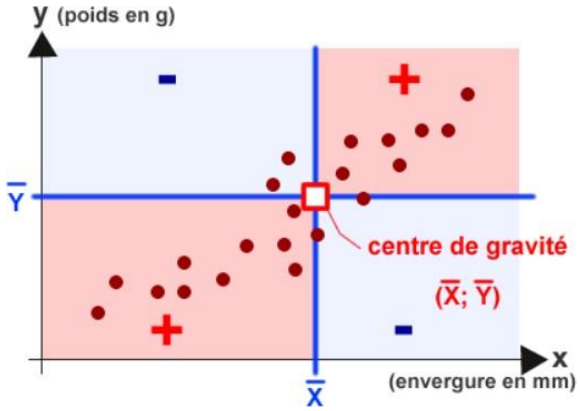


2.6. Mesures des relations entre variables

- Covariance: est le produit des écarts moyen du nuage de points.

Elle est positive lorsque le nuage de points a une orientation ascendante, et négative lorsque ce nuage a une orientation descendante.

$$Cov(x,y) = \frac{1}{N-1}\sum_{i}(x_i - \bar{x})(y_i - \bar{y})$$

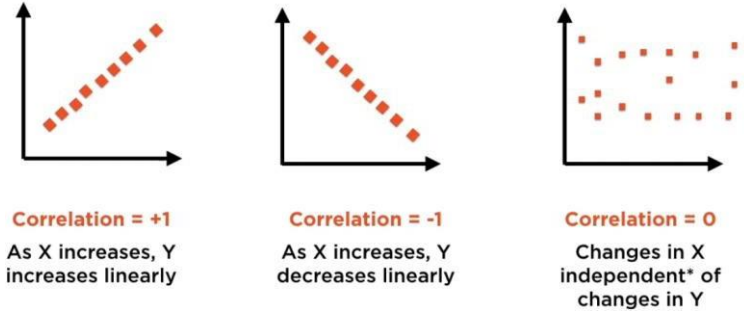


2.6. Mesures des relations entre variables

- Coefficient Corrélation: quantifie l'intensité et le sens de la relation qui existe entre deux variables.

  C'est un nombre sans unités compris entre -1 et +1.
  - Si les deux variables varient indépendamment l'une de l'autre, sa valeur est de 0.
  - Si les deux variables évoluent parallèlement (Y augmente lorsque X augmente), sa valeur sera positive, avec un maximum de 1 (lorsque l'évolution de Y est directement proportionnelle à celle de x).
  - Si les deux variables évoluent à l'inverse l'une de l'autre, sa valeur sera négative, avec un minimum de -1.

$$\text{Correlation } (x,y) = \frac{\text{Covariance } (x,y)}{\sqrt{\text{Variance } (x)}\ \sqrt{\text{Variance } (y)}}$$



Exercice 1: Les notes

- Voici les notes (sur 10) reçues par 20 élèves de 3e secondaire:

1. Calculez la moyenne
2. Calculez la variance
3. Calculez l'écart-type
4. Calculez le coefficient de variation
5. Calculez la médiane
6. Calculez le mode

Veillez à toujours bien comprendre ce que cela signifie, c'est-à-dire à pouvoir interpréter vos résultats

Observation	Note
1	10
2	7
3	0
4	5
5	2
6	7
7	9
8	7
9	0
10	8
11	3
12	2
13	8
14	2
15	10
16	2
17	3
18	7
19	5
20	6

Exercice 2: Les notes

- Voici la distribution des notes reçues par des élèves de 3e secondaire:

1. Calculez la moyenne
2. Calculez la variance
3. Calculez l'écart-type
4. Calculez le coefficient de variation
5. Calculez la médiane
6. Calculez le mode

Veillez à toujours bien comprendre ce que cela signifie, c'est-à-dire à pouvoir interpréter vos résultats

Note	Effectifs
1	3
2	7
3	7
4	12
5	9
6	11
7	12
8	14
9	20
10	5

Exercice 3: Les notes

- Voici la distribution des notes reçues par des élèves de 3e secondaire (N=100):

1. Calculez la moyenne
2. Calculez la variance
3. Calculez l'écart-type
4. Calculez le coefficient de variation
5. Calculez la médiane
6. Calculez le mode

Note	fréquence
1	0
2	0,1
3	0,05
4	0,03
5	0,1
6	0,1
7	0,2
8	0,3
9	0,04
10	0,08

Veillez à toujours bien comprendre ce que cela signifie, c'est-à-dire à pouvoir interpréter vos résultats

Exercice : Internet

On a demandé aux étudiants s'ils avaient internet sur leur ordinateur ou celui de leurs parents. Nous avons croisé cette information avec leur sexe, ce qui nous donne le tableau suivant. NB: certaines transformations sont nécessaires à effectuer avant d'obtenir ce tableau.

Sexe	Internet	Pas Internet
Homme	175	114
Femme	461	198

- A partir du fichier Excel « HF - Internet », répondez aux questions suivantes:

A) Calculez les distributions marginales.

B) Quelle est la nature des variables étudiées?

C) Calculez les fréquences conjointes et marginales.

Exercice : Traitement

Contexte: On dispose de deux traitements, TrA et TrB, contre une certaine maladie M qui sévit en Europe Occidentale. L'équipe médicale d'un hôpital a étudié les statistiques concernant 281 personnes affectées par cette maladie qu'elle a recueillies au cours du dernier trimestre.

- A partir du fichier Excel « Traitements », répondez aux questions suivantes:

Traitements	Guéri	Non-Guéri
TrA	139	34
TrB	98	10

A) Calculez les distributions marginales.

B) Calculez les fréquences conjointes, marginales et conditionnelles. Interprétez vos résultats.

3.Graphiques

Chaque type de variable statistique ou combinaison de variable statistiques ont leurs manières d'être affichés.

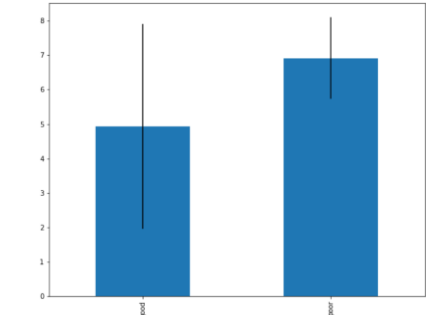
Dis autrement: on ne peut pas utilisé n'importe quel graphique pour une variable donné, si on veut que ce graphique ait un sens.

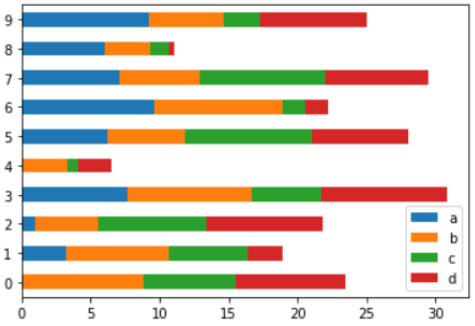
Bonne ressources pour s'orienter:

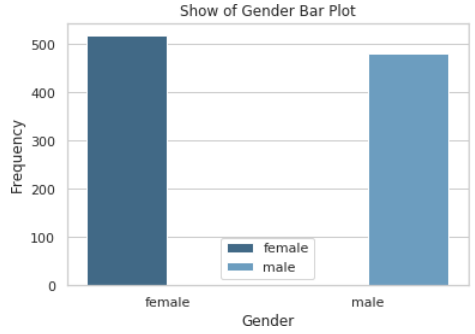
- https://flowingdata.com/2009/01/15/flow-chart-shows-you-what-chart-to-use/
- https://datavizcatalogue.com/search.html

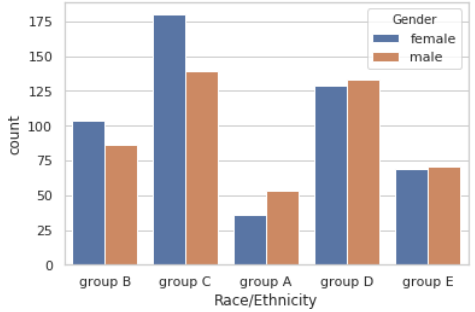
3. Graphiques

- Graphique en barres



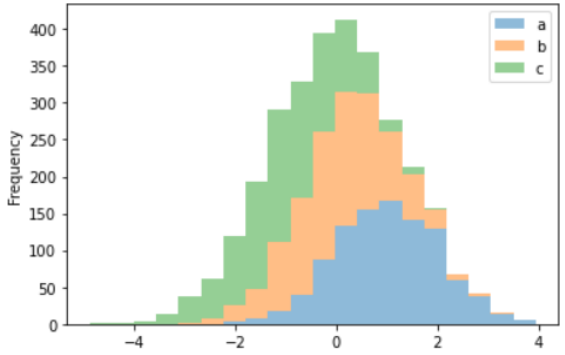


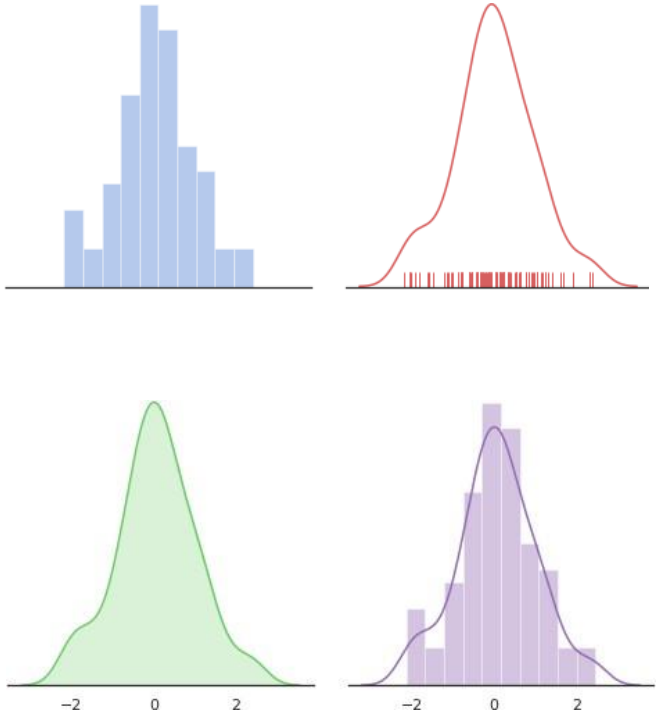




3. Graphiques

- Histogramme

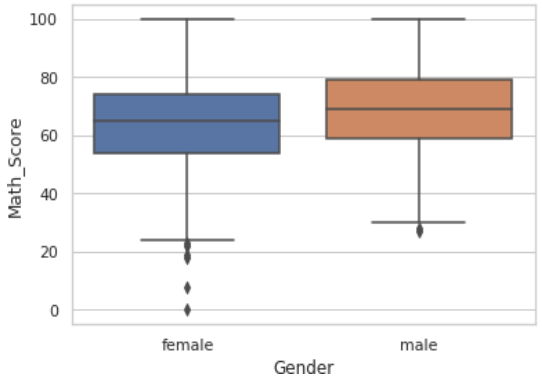


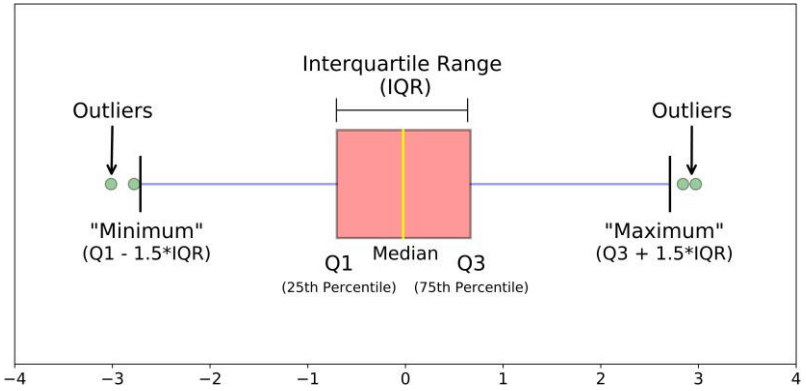


3. Graphiques

- Boite de dispersion (Boxplot)

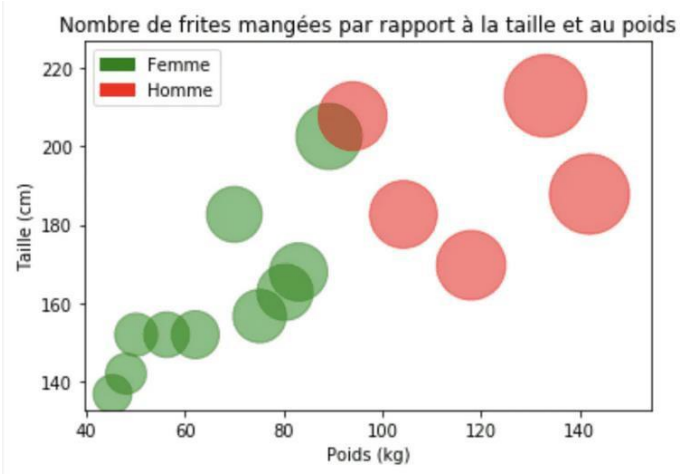
  data.boxplot()

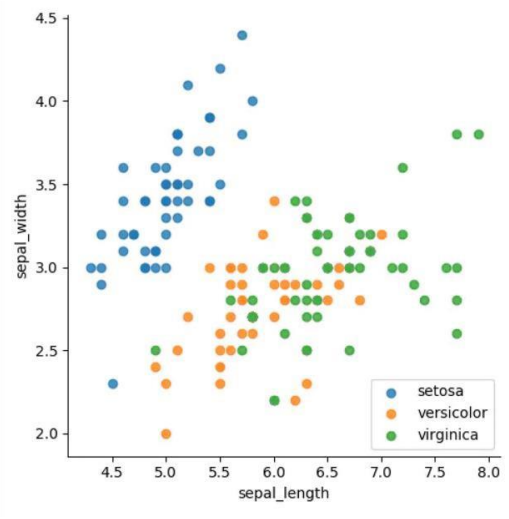




3. Graphiques

- Nuage de points (scatter plot)

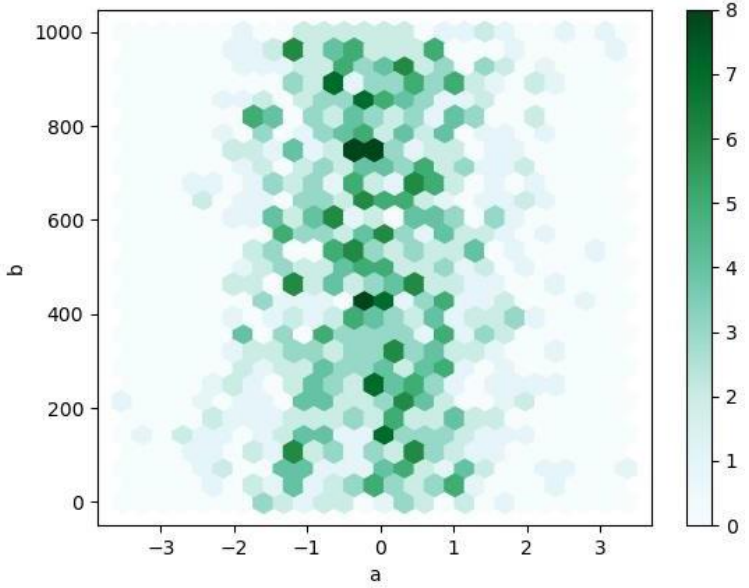




3. Graphiques

- Graphique hexagonal

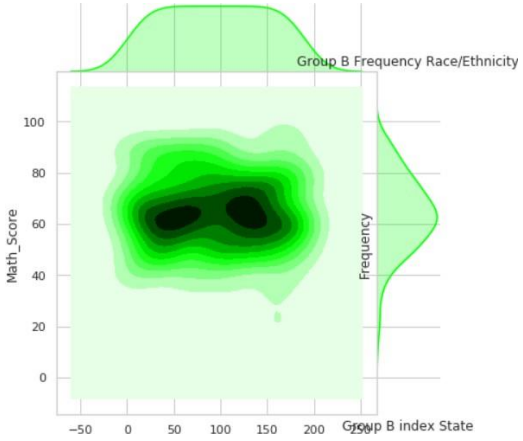
Permet d’exprimer la densité des points

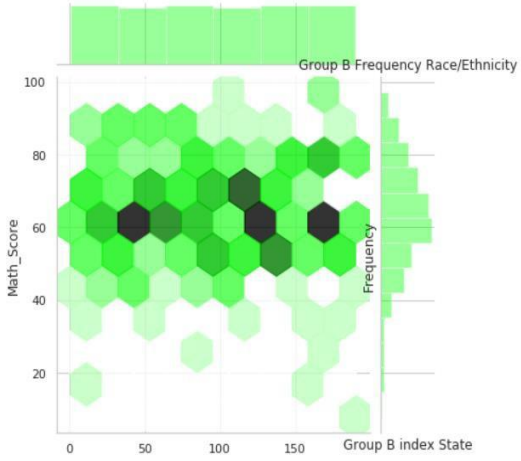


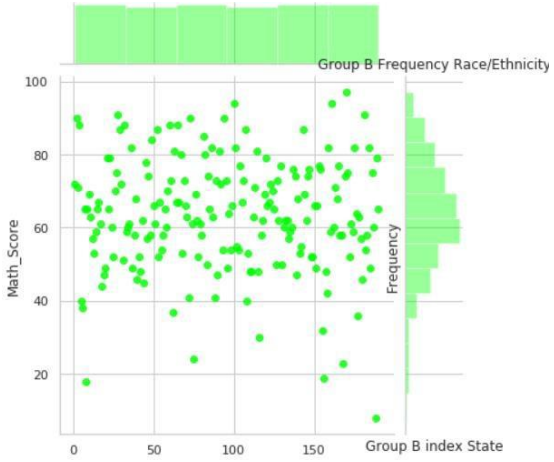
3. Graphiques

- Jointplot

Combinaison de scatter plot avec un histogramme pour chaque variables

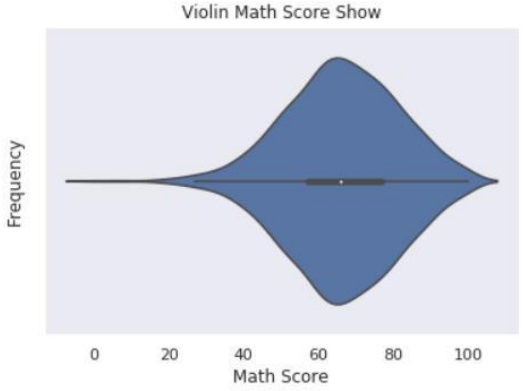


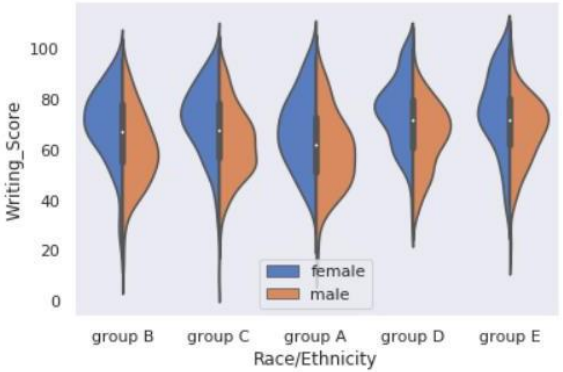


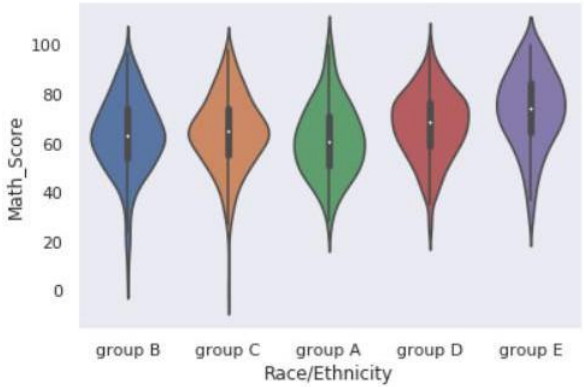


3. Graphiques

- Violinpot

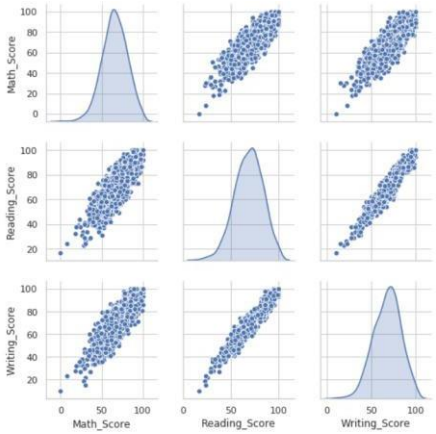


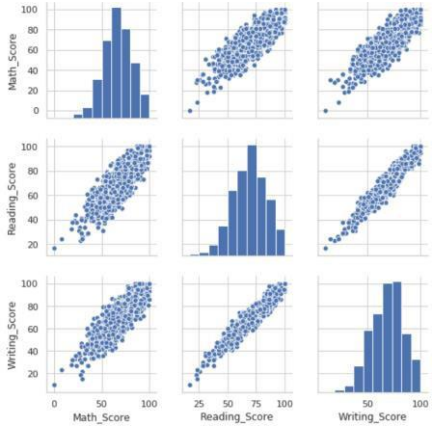




3. Graphiques

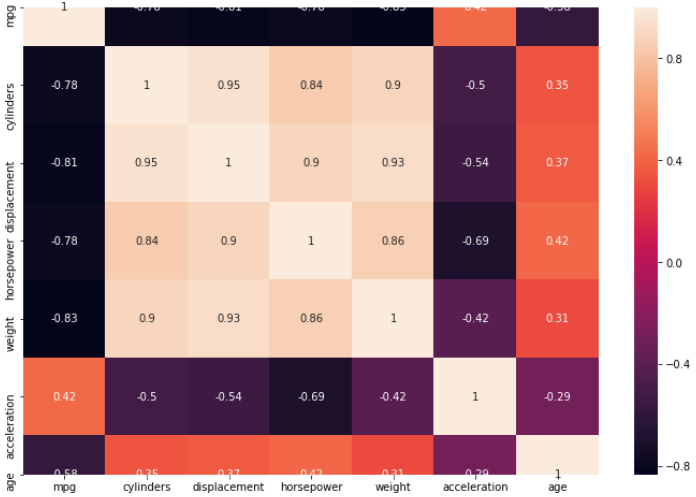
- Pairplot





3. Graphiques

- Heatmap



Exercice : graphiques

Sur base du fichier Excel « StudentsPerformance » réaliser une dizaine de graphiques différents pour explorer les données avec Pandas et Seaborn.

Statistique Inférentielle

CONSULTING BSTORM

Distribution

- Une variable aléatoire est caractérisé par une distribution c'est-à-dire une ensemble de valeur qu'elle peut prendre et la probabilité associée à chacune de ces plages de valeurs

La variable aléatoire va donc se caractériser par:

Espérance E(x) (//moyenne)

Variance V(x)

Les représentations graphiques sont idéales pour se représenter ces distributions.

Distribution

- **"Quelle est la probabilité que si nous interrogions une nouvelle personne, sa classe d'âge soit 30 ? "**

0.174

On pourra dire que nous avons environ 17% de chance que cette personne soit de la classe d'âge 30

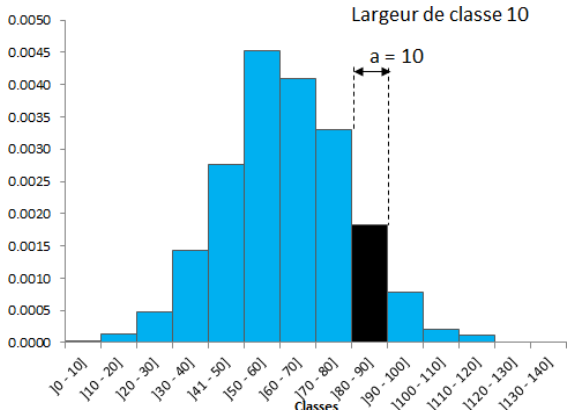
Classes	*Fréquence*	*Probabilité*
0	4	0.004
10	19	0.019
20	47	0.047
30	174	0.174
40	230	0.230
50	272	0.272
60	152	0.152
70	75	0.075
80	23	0.023
90	4	0.004
100	0	0.000
110	0	0.000
120	0	0.000
130	0	0.000

Loi probabilité

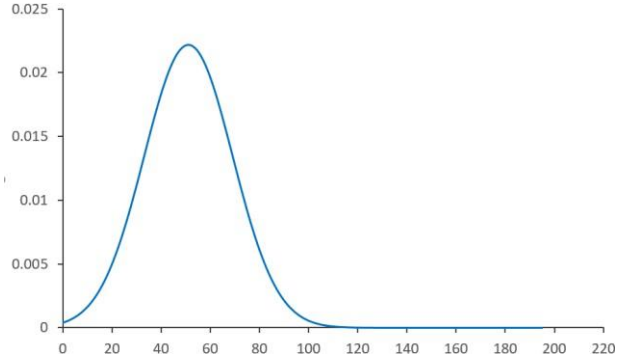
- Une fonction de probabilité associe à chaque résultat possible (d'une variable discrète) une certaine probabilité

La somme des probabilités d'une fonction de probabilité est égale à 1

Exemple: la somme des probabilités de chaque résultat possible d'un lancer de dé vaut 1



- Une fonction de densité décrit pour une variable continue sa forme, car pour une variable continue, il y a une infinité de résultats possibles et la probabilité d'un résultat particulier est nulle



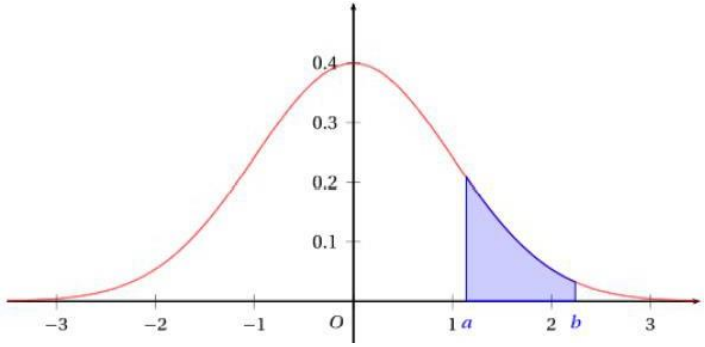
Loi probabilité

- Dans le cas d'une fonction de densité (et donc d'une variable continue), la somme des probabilités ne peut donc pas être calculée.

- On va mesurer les probabilités en calculant l'aire sous la fonction de densité dans un certain intervalle.

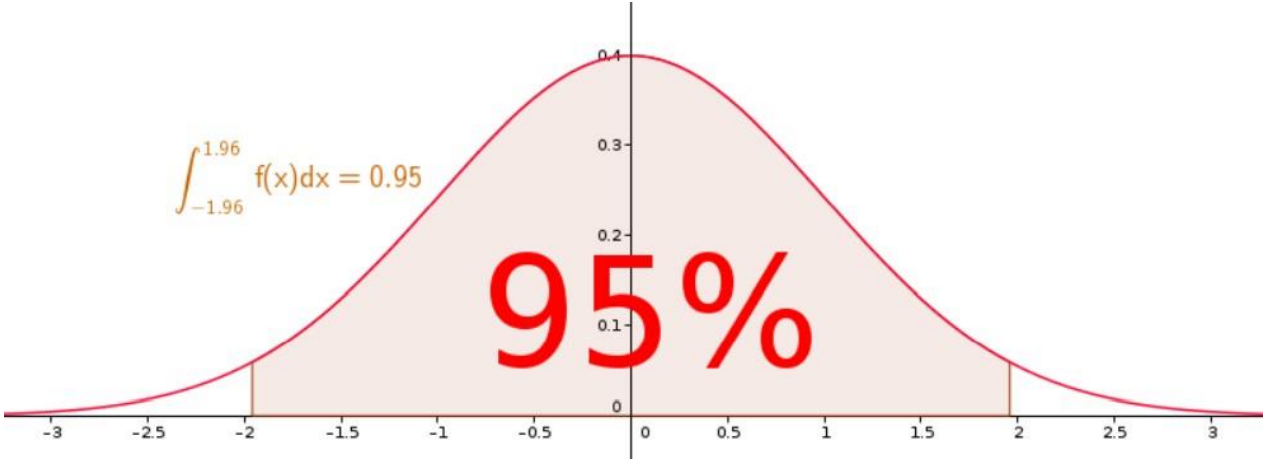
NB: L'aire sous une fonction de densité sur son domaine vaut toujours 1

Exemple: p(a≤ X ≤ b) est l'aire coloré



Loi probabilités

- Exemple :



Loi probabilité

- Théoriquement, il existe plusieurs de modèles de distribution.

Les lois de probabilité se distinguent par :

  - leurs formes,
  - leurs paramètres de position,
  - leurs paramètres de dispersion

- Les distributions ont des fonctions de densité de probabilité qui leur sont attachées. Elles constituent des lois qui permettent de décrire la manière avec laquelle sont distribuées les valeurs.

Loi Probabilité

- Il y a de nombreuses lois de probabilités connues
  - Discrètes:
    - Bernoulli: variable aléatoire qui n'admet que deux valeurs {1,0} (succès ou échec) avec les probabilités respectives p et q = 1-p (lorsque l'épreuve n'a que 2 issues)
    - Binomiale: épreuve de Bernoulli où X = nbre de succès au cours de n épreuves
    - Géométrique: épreuve de Bernoulli où X = nbre de fois qu'il faut répéter l'épreuve pour obtenir le 1er succès
    - Poisson: nbre de réalisations de l'événement dans un laps de temps donné distribué suivant une loi de Poisson

Loi Probabilité

- Il y a de nombreuses lois de probabilités connues

Voici un 1er exemple de loi probabilité discrètes:

- Binomiale: variable aléatoire qui n'admet que deux valeurs {1,0} (succès ou échec)

Espérance E(x): np

Variance V(x): np(1-p)

Fonction probabilité:

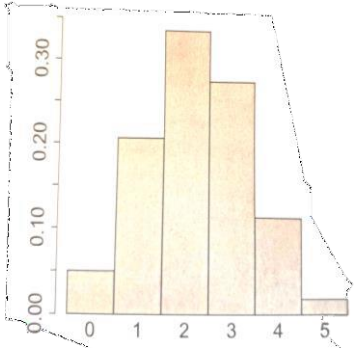
$$P(X=k)=\binom{n}{k}p^k.q^{n-k} \quad avec \binom{n}{k}=\frac{n!}{k!(n-k)!}$$

Loi Probabilité

Exemple: on a observé dans le passé que 45% des clients d'une chaine de supermarché achètent au moins un produit de sa marque propre lors de chaque visite. Parmi 5 client venus aujourd'hui, 4 d'entre eux n'ont pas acheté de produits de la marque propre. Est-ce que cela semble cohérent?

X~Bin(5,0.45) --> Soit X le nombre de personnes (parmi les 5) qui ont acheté au moins un produit de la marque du distributeur

P(X=1)= 5!/(1!4!) = 0.21



Loi Probabilité

- Il y a de nombreuses lois de probabilités connues.

Voici une 2e exemples de loi discrètes:

■ Poisson: nbre de réalisations de l'événement dans un laps de temps donné distribué suivant une loi de Poisson

Espérance E(x): λ
Variance V(x): λ

Fonction probabilité: $$P(X = k) = \frac{e^{-\lambda}\lambda^{k}}{k!}$$

Loi Probabilité

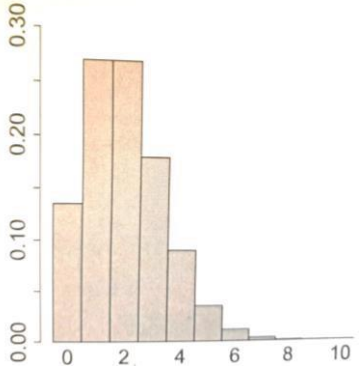
Exemple:

Lors des 10 premières années, il y a eu en moyenne 2 accidents par mois à un carrefour réputé dangereux. Le mois passé, il y a eu une augmentation de 100 % du nombre d'accidents par rapport à cette moyenne: le carrefour semble-t-il être devenue plus dangereux?

X~Poi(2) -> Soit x le nombre accidents mensuels au carrefour

$P(X=4)=(e^{-2}x2^4)/(4!)$

$P(X>=4)=1-P(x<4)=0.1429$



Loi Probabilité

- Il y a de nombreuses lois de probabilités connues

Voici des exemples de loi continue:

- Normale: utilisée dans de nombreuses méthodes statistiques car intervient dans des conditions très générales (on y reviendra plus tard)

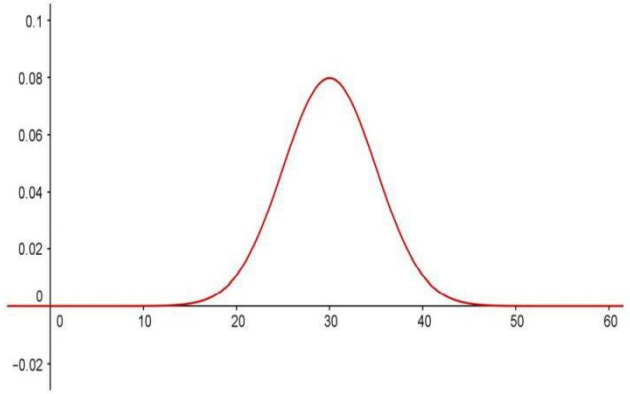
- Uniforme: variable aléatoire X qui prend ses valeurs dans un intervalle de bornes a et b (probabilité partout identique)

- Exponentielle: cas particulier de Poisson et avec variables continues

Loi normale

- Par exemple: considérons que cette courbe représente la distribution du poids des enfants à 12 ans

- On sait alors qu'un enfant de 12 ans pèse en moyenne 30 kg, mais qu'il peut peser entre 15 et 45 kg

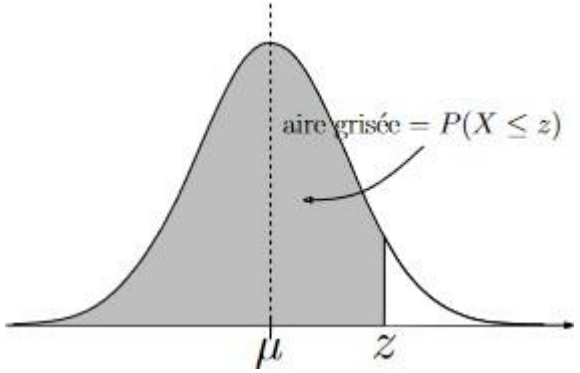


Loi normale

- La loi Normale est aussi caractérisée par sa moyenne et sa variance => $N(\mu,\sigma^2)$

- La logique de cette loi est que au plus une valeur est proche de la moyenne, au plus elle est probable

- La loi Normale est symétrique autour de sa moyenne



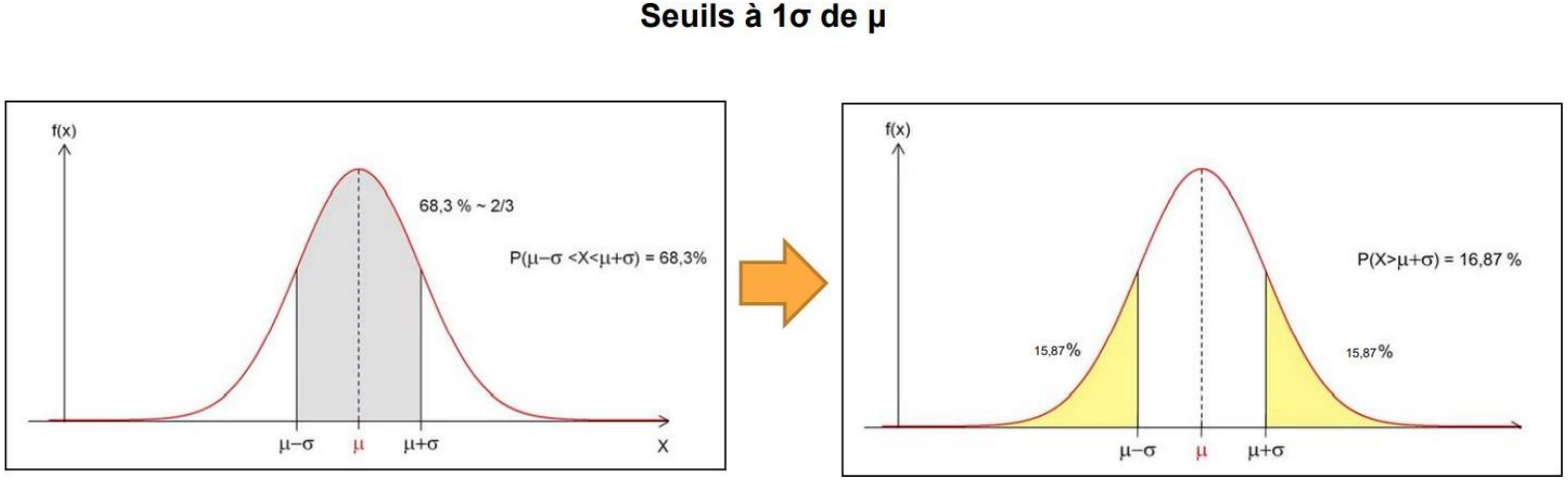
La loi Normale centrée - réduite

- Pour éviter des calculs difficiles, on dispose de table pour une loi Normale standardisée

- Toute loi normale peut être standardisée ou centrée réduite

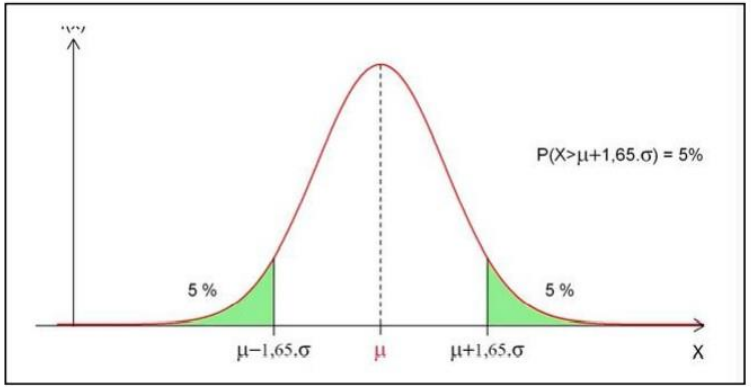
- Une loi Normale standardisée est définie par N(0,1) => espérance = 0

La loi Normale centrée - réduite

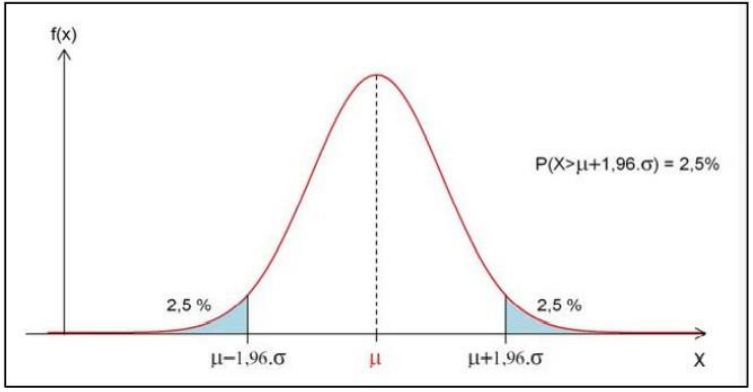


La loi Normale centrée - réduite

Seuils à 1,65σ de μ



Seuil à 1, 96σ de μ



Exercice : Loi normale

3.25 Trouver les probabilités suivantes pour une variable aléatoire normale standardisée Z

$P(Z \leq 1.96)$

$P(Z \geq 1.645)$

$P(Z \leq -1)$

$P(0 \leq Z \leq 1.2)$

Théorème central limite

- Le **théorème central limite** est sans doute le théorème le plus important des statistiques. Il est utilisé pour les démarches inférentielles

1) Si la taille de l'échantillon est suffisamment grand (>30) alors la distribution d'échantillonnage se rapproche de la distribution de la loi normale.

2) Si la distribution de la population est distribuée selon la loi normale, la distribution d'échantillonnage suit une loi normale indépendamment de la taille de l'échantillon.

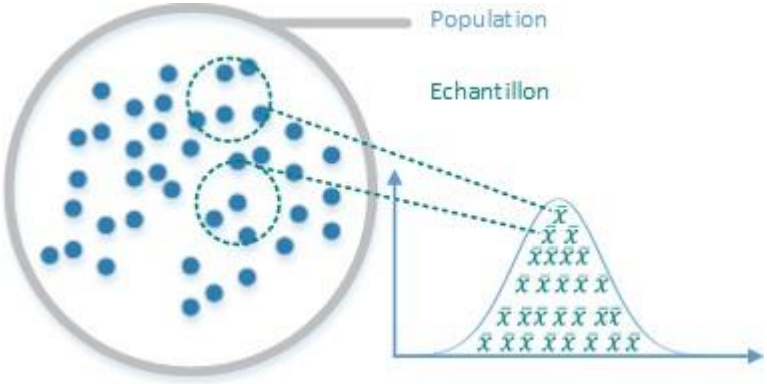


Image: https://commentprogresser.com/statistiques.html

Théorème central limite

- **2)** La moyenne de la distribution d'échantillonnage des moyennes des échantillons est égale à la moyenne de la distribution de la population originale.

- **3)** l'écart-type de la moyenne **m** est égale à l'écart-type des valeurs de l'échantillon population divisée par la racine carrée de la taille d'échantillon.

$$s_m = \frac{s}{\sqrt{n}}$$

- https://www.youtube.com/watch?v=4dhm2QAA2x4

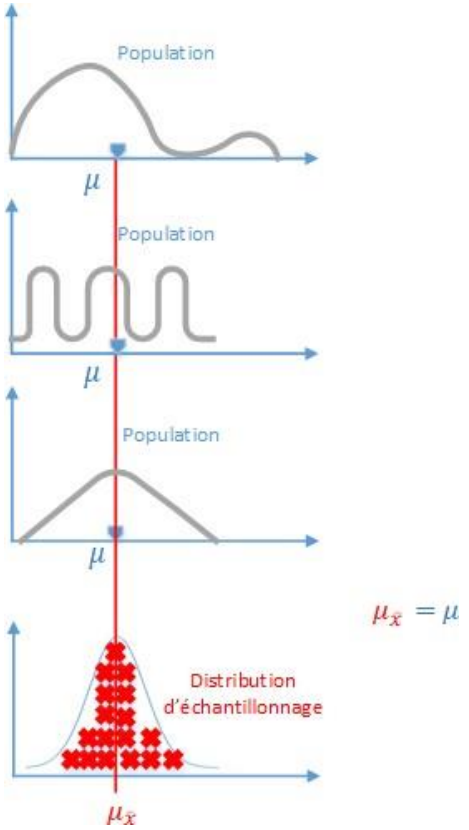


Image: https://commentprogresser.com/statistiques.html

- https://www.youtube.com/watch?v=4dhm2QAA2x4

Les principales causes d’un mauvais échantillonnage

Les principales causes d’un mauvais échantillonnage

- L’erreur d’échantillonnage: représente les différences qui existent entre l’échantillon et la population qui sont uniquement dues au choix des observations (au hasard).

- Le biais d’échantillonnage: représente les erreurs liées à une mauvaise méthode d’échantillonnage.

Exemple: tendance à favoriser la sélection d’individus ayant une caractéristique particulière.

Statistique inférentielle

- L'objectif de la statistique inférentielle est d'inférer (=obtenir) de l'information sur une population à partir d'informations sur un échantillon
    - Exemple: on veut obtenir de l'information sur le salaire moyen dans une multinationale en se basant sur les salaires observés dans un département
- Une mesure statistique est toujours plus fiables si elle est basée sur une population plutôt que sur un échantillon
- Au plus l'échantillon est grand, au plus il tend vers la population, et au plus la mesure statistique sera fiable
- Raisons pour lesquelles on travaille sur un échantillon: coût trop élevé, information inaccessible, pas le temps de faire le recensement…

Quelques concepts importants

- La statistique inférentielle s'intéresse aux phénomènes à priori en associant à une variable aléatoire une probabilité à chacune de ses valeurs possibles
  - Exemple: le lancer de dé => chaque résultat a une probabilité de 1/6
  - Le concept correspondant à la moyenne en statistiques inférentielles est l'espérance: avant que la variable aléatoire ne se réalise, à quelle valeur, en moyenne, pourrais-je m'attendre?
  - La formule de l'espérance est $\mathbb{E}[X] = \sum_{i=1}^{\infty} x_i \, p_i$
  - Ici on a (1/6)x1+(1/6)x2+(1/6)x3+(1/6)x4+(1/6)x5+(1/6)x6=21/6=3,5

Test hypothèse

- Les tests d’hypothèse sont souvent l’objectif final de la statistique inférentielle
- Leur objectif est de tester une hypothèse faite sur une population à partir d’un échantillon
- Les tests statistiques permettent de réaliser des comparaisons et d'en tirer des conclusions.
    - Exemple: un homme politique a dit que le poids moyen des femmes belges était de 70kg. On va alors mesurer le poids moyen des femmes belges sur un échantillon, et selon la valeur obtenue, on déterminera la probabilité que le poids moyen soit de 70kg compte tenu de la moyenne d’échantillon qu’on aura pêché

Test hypothèse

- Sans rentrer dans les détails techniques, un test d'hypothèse comporte trois éléments principaux:

  - Les hypothèses
  - Les calculs (pas vus dans ce cours)
  - La conclusion

Test hypothèse

- Le plus important pour vous est de comprendre la logique d'un test d'hypothèse et d'interpréter les résultats

1) Analysez un énoncé et déterminer sur quoi porte le test et quelles sont les hypothèses nulles et alternatives
2) Réalisez les calculs vous-même ou via un logiciel (pas vu ici)
3) Analysez les résultats du test pour tirer une conclusion par rapport aux hypothèses nulles et alternatives

Test hypothèse - Type

- Il existe beaucoup de différents tests disponibles:

○ Le test sur la moyenne ou sur la comparaison de moyennes (test t de Student où l'hypothèse nulle spécifie que deux distributions normales ont la même moyenne)

○ Le test sur la variance ou sur la comparaison de variance (test F de Fisher où l'hypothèse nulle spécifie que deux distributions normales ont la même variance)

○ Le test du Khi-carré d'indépendance

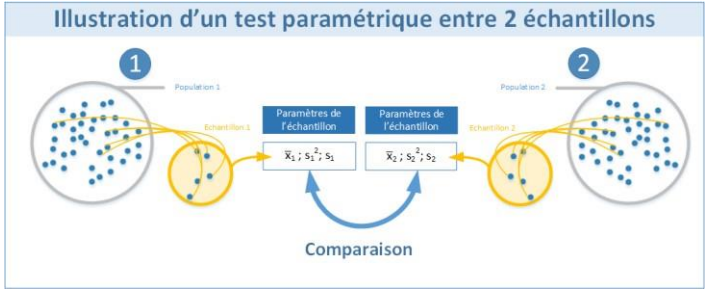


Image: https://commentprogresser.com/statistiques.html

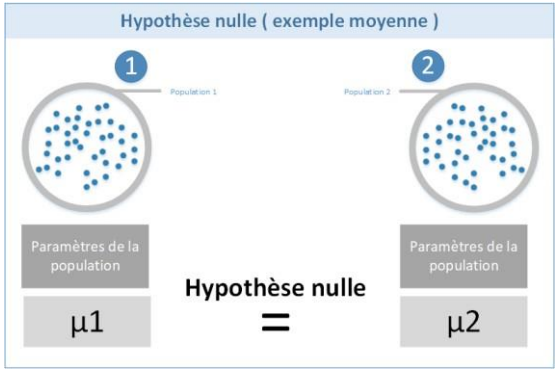
Test hypothèse

- Nous trouverons presque toujours des différences entre deux séries de données.

Le but de ces tests et d'indiquer si la différence observée est due au hasard ou si cette différence est réelle. C'est-à-dire que les deux populations concernées ne sont pas semblables.

- En réalisant des tests statistiques sur des échantillons et non sur la population nous devons admettre un **risque d'erreur**

Test hypothèse



- En dehors des calculs statistiques, un test d'hypothèse comporte deux éléments principaux:

  - <u>Les hypothèses:</u>

    - Une hypothèse nulle (H0): consiste à dire que les paramètres ou les distributions entre les deux populations sont identiques.

    - Une hypothèse alternative (HA): l'hypothèse qui est retenue au cas où l'hypothèse H_0 est rejetée, c'est-à-dire que la différence observée est trop grande pour qu'on l'attribue à une simple fluctuation d'échantillonnage. On suppose donc que dans ce cas les paramètres ou les distributions de population sont différents.

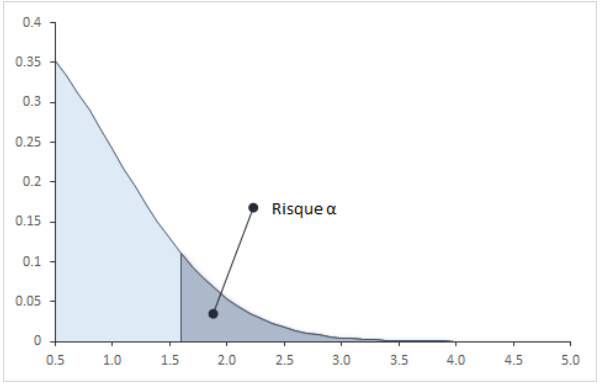
Image: https://commentprogresser.com/statistiques.html

Risque d’erreur de type 1

- En rejetant H_0 on prend un risque que l'on appelle le **risque alpha α** .
- Il s'agit du risque de se tromper en rejetant H_0 si dans la réalité H_0 est vrai. On appelle également ce risque le **risque type I**.
- Le risque alpha α est déterminé avant la réalisation du test.

On fixe ce risque d'erreur alpha à 5%.

Bien sûr il est possible de changer ce risque en fonction du domaine dans lequel on applique le test.
Dans des domaines où les enjeux sécurité sont forts ce risque pourra par exemple être de 1% ou 0,1%.



Risque erreur de type 2

- Le **risque bêta** est le risque de ne pas avoir rejeté H_0 alors que H_1 est vrai. Cela arrive lorsqu'il existe une différence entre les paramètres étudiés, mais que la valeur observée se situe néanmoins dans l'intervalle comprenant 95 % des **valeurs probables**.

Ce risque est appelé risque **type II**.

Puissance d’un test hypothèse

- La puissance d'un test se calcule de la manière suivante : **1 - β**
  - Taille des effectifs des échantillons. Plus la taille des échantillons augmente plus la puissance augmente et plus le risque bêta diminue.
  - Diminution de la variabilité des données, implique une augmentation de la puissance

    Ex: Le sexe, l'âge, le poids, le stress, la provenance des animaux jouent sur la pression sanguine et entraînent une grande variabilité des données.

- La valeur du risque bêta n'intervient pas dans l'interprétation d'un test car on ne sait pas la calculer.

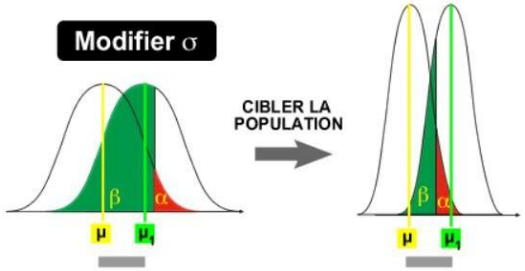


Image: https://webapps.unamur.be/umdb/Biostats3

Test hypothèse

Voici une synthèse des risques

		Réalité	
		H0 est vraie	H1 est vraie
Hypothèses	On ne rejette pas H0	Pas d'erreur	Erreur de type II
	On rejette H0 et on accepte H1	Erreur de type I	Pas d'erreur

Test hypothèse

- En dehors des calculs statistiques, un test d'hypothèse comporte deux éléments principaux (suite):

<u>La conclusion:</u> le but d'un test est de valider ou de rejeter H0.

- Si il y a des éléments forts pour dire que statistiquement, il y a peu de chance que H0 soit vraie au vu de l'échantillon qu'on a tiré, on rejette H0.

- Si il n'y a pas d'éléments assez forts, on ne pourra pas rejeter H0.

- Le degré de certitude requis pour « Rejeter H0 » est fixé au début (souvent 90, 95 ou 99%)

- Il y a trois principales méthodes pour effectuer un test:

  - la p-value (celle que l'on va utiliser)
  - le test classique (à titre informatif)
  - l'intervalle de confiance (à titre informatif)

P-value

- La P-value est la probabilité de rejeter H0 par erreur, c'est-à-dire la probabilité de rejeter H0 alors que H0 est vraie (et donc de se tromper):
    - Exemple 1 : si P-value = 0,005 alors la probabilité de rejeter H0 par erreur est relativement faible donc on pourra se permettre de la rejeter.

    - Exemple 2: si la P-value = 0,457 alors la probabilité de rejeter H0 par erreur est relativement élevée donc on ne pourra pas se permettre de la rejeter.

- On peut aussi la définir comme:
    - la probabilité de tirer un échantillon tel que l'on a tiré, si H0 est vraie ○ mesure la cohérence entre l'hypothèse nulle qui porte sur la population et l'échantillon tiré

P-value

- La P-value permet de conclure sur un test d'hypothèse en la comparant au risque toléré (seuil = alpha)
  - Si la p-value < alpha => rejet de H0
  - Si la p-value > alpha => non-rejet de H0

Exemple: On cherche à vérifier que deux échantillons proviennent d'une population ayant la même moyenne. On calcule la probabilité de tirer cette moyenne d'échantillonnage si H0 est vraie. Cette probabilité est égale à 0,0091. Considérons un seuil de rejet à 1% (alpha). Comme 0,0091 < 0,01 => rejet de H0

H0: les moyennes sont =

H1: les moyennes sont ≠

➢ Il y a peu de chances que H0 soit vraie si l'on tire aléatoirement un élément dans l'échantillon. Il y donc une différence significative à 1% entre les deux moyennes d'échantillonnage.

Le test classique

- Un test classique calcule la valeur qui correspond au risque maximum toléré afin de créer une zone de rejet et une zone de non-rejet de l’hypothèse nulle.

- On vérifie ensuite si la moyenne d’échantillonnage constatée se situe dans la zone de rejet ou non

Intervalle de confiance

- La logique de l'intervalle de confiance est inversée par rapport aux autres méthodes, car ici on part de la moyenne d'échantillonnage pour construire un intervalle dans lequel on va accepter H0.

- L'idée est de répartir une certaine marge d'erreur tolérable de chaque côté de la moyenne d'échantillonnage pour déterminer l'intervalle de confiance.

- Si H0 fait partie de l'intervalle de confiance, on ne la rejette pas.

Exercice : Médicament

Un fabricant de médicaments certifie que son produit est efficace à 90 %. Le traitement est effectué sur un échantillon de 200 personnes et aboutit à 170 guérisons. Peut-on en conclure que le fabricant exagère à propos de son taux de réussite si on admet un risque d'erreur maximum de type 1 de 5 % ?

p-value = 0,91%

Exercice : Association de consommateurs

Formuler le test qui s'applique le mieux (bilatéral, unilatéral...) aux situations suivantes :

a) Une association de consommateurs met en doute le fait qu'un appareil électrique S a une durée de vie de 200 heures comme le prétend le fabricant.

b) Un électricien habitué à travailler avec cet appareil S accepterait toutefois d'utiliser l'appareil T si la durée de vie de T s'avérait supérieure à celle de S.

Valeurs fictives:

alpha=1% ; p-value (a) = 72% ; p-value (b) = 7% ; p-value (c) = 0,2%

Exercice : Saison touristique

Pendant la saison touristique, un certain restaurateur sert en général 150 couverts par jour, l'écart-type étant de 15 couverts (on considère que le nombre de couverts est distribué selon une loi normale). Après la première semaine de juillet de cette année, il a respectivement servi 120, 135, 160, 140, 180, 210 et 175 couverts. Peut-on dire que le résultat du restaurateur est différent de celui enregistré précédemment (on admet un risque maximum d'erreur de type 1 de 5 %) ?

p-value = 7,84%

Exercice : Dépenses scolaires

Le directeur d'un collège pense que le montant annuel des dépenses scolaires des étudiants est égale à 5000 F. Certains professeurs estiment que l'espérance de ce montant est plus élevée, d'autres qu'elle est plus faible. Afin de fixer les idées, on décide de tirer aléatoirement un échantillon d'étudiants et on obtient une moyenne de 5470 F.
Effectuer le test d'hypothèse dans les cas suivants (en admettant $\alpha = 0.02$)

p-value < 0,01%

Exercice : Cochons d'Inde

Six cochons d'Inde auxquels on a administré 0.5 mg de tranquillisant ont mis 11, 13, 9, 14, 15 et 13 secondes pour tomber endormis. Six autres cochons d'Inde auxquels on a injecté 1,5 mg de tranquillisant ont mis 10, 5, 8, 9, 6 et 10 secondes pour tomber endormis.
Testez (en admettant alpha = 0.10) à partir du résultat du test réalisé par « SAS Enterprise Guide »

a) s'il est raisonnable de supposer que les deux échantillons proviennent de populations ayant le même écart-type ;
b) ainsi que la même moyenne.

On suppose que les temps d'endormissement sont distribués normalement :

a)

Equality of Variances					
Variable	Method	Num DF	Den DF	F Value	Pr > F
Temps	Folded F	5	5	1.07	0.9441

b)

T-Tests					
Variable	Method	Variances	DF	t Value	Pr > |t|
Temps	Pooled	Equal	10	3.65	0.0044
Temps	Satterthwaite	Unequal	9.99	3.65	0.0044

On cherche à savoir si le choix du journal acheté est lié à la classe sociale des personnes. Analysez la p-value du Chi carré et concluez.

Statistics for Table of Classe sociale by Journal

Statistic	DF	Value	Prob
Chi-Square	6	21.5122	0.0015
Likelihood Ratio Chi-Square	6	20.4976	0.0023
Mantel-Haenszel Chi-Square	1	5.2004	0.0226
Phi Coefficient		0.2678	
Contingency Coefficient		0.2587	
Cramer's V		0.1894	

Sample Size = 300

Question de réintégration

Grâce au théorème central limite et à la loi normale, quelle déduction peut-on avoir, parmi les déductions suivantes, sur la moyenne que va avoir un échantillon donné de taille supérieure à 30 ?

1. Nous savons que la moyenne d'un échantillon a plus de 99 % de chance d'être incluse dans l'intervalle [m – 0.5écart type, m + 0.5 écarts types] avec m la moyenne de la population.
2. Nous savons que la moyenne d'un échantillon a plus de 99 % de chance d'être incluse dans l'intervalle [m - 1 écart type, m + 1 écarts types] avec m la moyenne de la population.
3. Nous savons que la moyenne d'un échantillon a plus de 99 % de chance d'être incluse dans l'intervalle [m – 2 écart type, m + 2 écarts types] avec m la moyenne de la population.
4. Nous savons que la moyenne d'un échantillon a plus de 99 % de chance d'être incluse dans l'intervalle [m - 3 écart type, m + 3 écarts types] avec m la moyenne de la population.

Question de réintégration

- Qu'est-ce qui permet de définir un intervalle de confiance ?

1. Un intervalle de confiance permet de définir une probabilité avec laquelle une valeur donnée va être incluse dans cet intervalle.
2. Un intervalle de confiance permet de définir le % des valeurs manquantes d’une variable
3. Un intervalle de confiance permet de définir l’étendue d’une variable
4. Un intervalle de confiance permet de définir la distance moyenne des valeurs d’une variable par rapport au centre de cette variable

Question de réintégration

Lorsque la valeur de p-value est inférieure à 0.0001, que peut-on conclure sur l'hypothèse nulle ?

1. Hypothèse nulle est fausse à 100%
2. L’hypothèse nulle à une très faible probabilité d’être vraie
3. L’hypothèse nulle est vraie à 100%
4. L’hypothèse nulle à une très forte probabilité d’être vraie

Question de réintégration

Qu'est-ce qu'on cherche à réaliser avec un test d'hypothèse ?

1. Mesurer la probabilité d’acceptation ou rejet de l’hypothèse H1
2. Calculer un intervalle de confiance
3. Mesurer la probabilité de l’acceptation ou de rejet de l’hypothèse nulle Ho
4. Etudier la distribution d’une variable

Question de réintégration

Quelle est la taille minimale d'un échantillon issu d'une population quelconque qui permet d'appliquer le théorème central limite ?

1. La taille d'un échantillon doit être supérieure à 10
2. La taille d'un échantillon doit être supérieure à 20
3. La taille d'un échantillon doit être supérieure à 30
4. La taille d'un échantillon doit être supérieure à 100
5. La taille d'un échantillon doit être supérieure à 1000

Question de réintégration

Quelle est la proposition correcte parmi les propositions suivantes ?

1. Lorsqu'une distribution suit une loi normale, alors environ 99 % des valeurs devraient être inférieures à la moyenne
2. Lorsqu'une distribution suit une loi normale, alors environ 25 % des valeurs devraient être inférieures à la moyenne
3. Lorsqu'une distribution suit une loi normale, alors environ 50 % des valeurs devraient être inférieures à la moyenne
4. Lorsqu'une distribution suit une loi normale, alors environ 75 % des valeurs devraient être inférieures à la moyenne

Régression Linéaire

- Régression linéaire: une régression linéaire est un modèle linéaire permettant de déterminer si une variable Y peut être expliquée (au moins en partie) par une (ou des) variable(s) X1 (X2 ,X3 ,…,Xn)

- Des variations de X influencent-elles la variation de Y?

- La régression linéaire n’explique pas la causalité mais la corrélation

- Pour effectuer une régression linéaire, il faut des séries de données qui analysent le même phénomène

Régression Linéaire

Profit journalier	Température	stocks disponibles
525	26	70
689	12	80
533	32	60
748	4	94
733	19	99
570	9	50
668	3	65
394	27	44
800	1	100
325	4	32
734	25	88
354	29	56
430	9	66
393	30	33
759	18	76
694	32	67
564	29	59
225	1	21
427	27	46
225	5	41

Régression Linéaire

- Parfois, les données dont nous disposons peuvent nous obliger à devoir modifier les données => pas vu dans ces slides

- Y est la variable dépendante, et X la (les) variable(s) indépendantes

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_{p-1} X_{p-1} + \varepsilon$$

- Avec Betas étant les coefficient estimés grâce à la régression, et epsilon étant le terme d'erreur.

Régression Linéaire

- Les hypothèses de base faites lors d'une régression sont les suivantes:
    - La distribution de l'erreur est indépendante de X
    - L'erreur suit une loi Normale centrée en 0 et dont la variance est σ^2
    - Les coefficients sont constants

Régression Linéaire

- Deux méthodes classiques pour estimer les paramètres du modèle

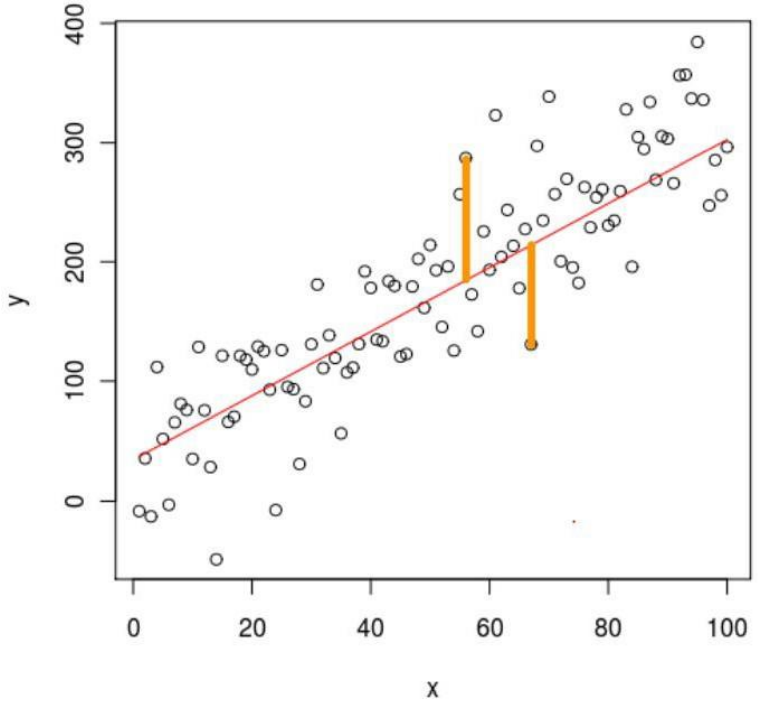
  ○ Le maximum de vraisemblance (pas discuté ici)

  ○ La méthode des moindres carrés (ici pour 1 variable indépendante)

Statistiques de bas

$$\min_{\beta_0,\beta_1} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

Régression Linéaire



Représentation d’un nuage de points

- Analysez les données dans le fichier Excel (Profit - Stock) et représentez le nuage de points du profit journalier en fonction des stocks disponibles.

Régression Linéaire

- On utilise souvent le R^2 (mieux le R^2 ajusté) pour expliquer quel pourcentage de variation de Y est expliqué par les variations de X
- Le test F de Fisher détermine si le modèle est bon, i.e. si au moins une des variables dépendantes explique la variable indépendante

    - H0: tous les betas sont nuls
    - Ha: au moins un beta est non-nul

- On utilise un test en t de Student pour déterminer si les coefficients betas (un par un) sont significatifs ou non

  - Quelles variables X influencent Y?
  - H0: le coefficient beta 1,2,…,n est nul
  - Ha: le coefficient beta 1,2,…,n est non-nul

Analyse régression linéaire

- Analysez les tableaux suivants et tentez de déterminer si la variable indépendante est corrélée aux variables indépendantes. Utilisez un maximum d’arguments pour affiner votre analyse

Linear regression model:

$$y \sim 1 + x1 + x2 + x3 + x4 + x5 + x6$$

Estimated Coefficients:

	Estimate	SE	tStat	pValue
(Intercept)	0.50633	0.69565	0.72785	0.47661
x1	-0.024281	0.16484	-0.1473	0.88463
x2	-0.0090029	0.021958	-0.41001	0.68692
x3	0.047544	0.047804	0.99456	0.33389
x4	-0.00085998	0.0012797	-0.67203	0.51059
x5	-0.0016993	0.0033481	-0.50755	0.61829
x6	0.0060627	0.0028391	2.1355	0.047571

Number of observations: 24, Error degrees of freedom: 17
Root Mean Squared Error: 0.597
R-squared: 0.299, Adjusted R-Squared 0.0521
F-statistic vs. constant model: 1.21, p-value = 0.348

Linear regression model:

$$y \sim 1 + x1 + x2 + x3 + x4 + x5 + x6$$

Estimated Coefficients:

	Estimate	SE	tStat	pValue
(Intercept)	-0.14048	0.48633	-0.28886	0.77618
x1	0.28863	0.11524	2.5046	0.022733
x2	0.00094671	0.01535	0.061673	0.95154
x3	0.025057	0.033419	0.74978	0.46364
x4	-0.0022501	0.00089461	-2.5152	0.022245
x5	0.0024788	0.0023406	1.059	0.3044
x6	-0.0089985	0.0019848	-4.5338	0.00029374

Number of observations: 24, Error degrees of freedom: 17
Root Mean Squared Error: 0.418
R-squared: 0.701, Adjusted R-Squared 0.596
F-statistic vs. constant model: 6.65, p-value = 0.000932