Cours 4 - Statistique descriptive et inférentielle

Neila Mezghani

19 février 2021

COLLÈGE DE
BOIS-DE-BOULOGNE

Introduction à l'AED
(cours 1)
«Étapes de l'AED
La collecte de données

Transformation
Analyse en composante principale (cours 5)
Analyse factorielle de correspondance (cours 6)
Analyse factorielle mixte (cours 7)

Visualisation et interprétation
(cours 2)

Données

1 — 2 — 3 — 4 — 5 — 6

Prétraitement
Préparation des données (cours 3)
Statistiques descriptives et inférentielles
(cours 4)

Analyse
Regroupement (cours 8)
Classification (cours 9)
Analyse d'association (cours 10)

Connaissances

Plan du cours

1. Introduction
2. Statistique descriptive univarié
   - Variables qualitatives
   - Variables quantitatives
3. Statistique descriptive bivariée
   - Deux variables quantitatives
   - Une variable quantitative et une qualitative
   - Deux variables qualitatives
4. Statistique inférentielle
   - Introduction à l'inférence statistique
   - Intervalle de confiance
   - Tests d'hypothèses

Plan du cours

1. Introduction
2. Statistique descriptive univarié
   - Variables qualitatives
   - Variables quantitatives
3. Statistique descriptive bivariée
   - Deux variables quantitatives
   - Une variable quantitative et une qualitative
   - Deux variables qualitatives
4. Statistique inférentielle
   - Introduction à l'inférence statistique
   - Intervalle de confiance
   - Tests d'hypothèses

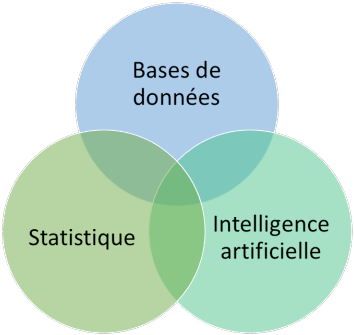
Statistique

L'AED est une discipline qui est situé à la croisée des **statistiques**, de l'intelligence artificielle et des bases de données.

Principe

On en distingue souvent trois sous-branches :

- Statistique descriptive
- Statistique inférentielle
- Modélisation statistique
  (Ne fait pas partie du cours)



Statistique descriptive

- **Définition :** Ensemble des méthodes conduisant à une description condensée d'un ensemble d'observations, indépendamment d'une référence explicite à un modèle statistique, donc d'inférence statistique.
- **Objectif :** Vise à résumer quantitativement l'information recueillie dans une population statistique, dont les individus sont classés selon certains caractères qualitatifs ou quantitatifs.
- **Exemples :** l'information sur la population d'un pays collectée à l'occasion d'un recensement général ou encore l'information relative aux bénéfices industriels et commerciaux recueillie grâce à une exploitation spéciale des documents fiscaux.

Statistique inférentielle

- **Définition :** La statistique inférentielle (aussi appelée l'inférence statistique) est le raisonnement par lequel on tire, à partir des observations faites sur un échantillon prélevé dans une population préalablement définie (population de référence), des conclusions concernant certaines caractéristiques quantitatives ou qualitatives de la population en question.
- Un synonyme populaire de l'inférence statistique est la généralisation (un terme plus utilisé en IA).
- **Exemples :** les sondages d'opinion, ou l'étude statistique de l'efficacité d'un médicament.

La modélisation statistique

- **Définition :** La modélisation statistique consiste à observer les caractéristiques d'un échantillon, puis de formaliser ces observations par des règles mathématiques.
- Cette formalisation s'appelle un modèle probabiliste. Une fois que l'on a décrit un phénomène par un modèle, on peut faire de la prédiction par exemple.
- **Exemples :** les classificateurs bayésiens se basent sur la modélisation statistique.

Plan du cours

1. Introduction
2. Statistique descriptive univarié
   - Variables qualitatives
   - Variables quantitatives
3. Statistique descriptive bivariée
   - Deux variables quantitatives
   - Une variable quantitative et une qualitative
   - Deux variables qualitatives
4. Statistique inférentielle
   - Introduction à l'inférence statistique
   - Intervalle de confiance
   - Tests d'hypothèses

Statistique descriptive univariée

- On parle de statistique descriptive **univariée** (unidimensionnelle) lorsqu'on considère une variable statistique **x**, observée sur n individus.
- Dans ce cas, les outils les plus utilisés sont :
  - Indicateurs : médiane, quantile, moyenne, fréquence, variance, écart-type
  - Représentations graphiques : diagramme en bâton, histogramme, graphiques cumulatifs, diagrammes en secteurs (tarte).

Plan du cours

1. Introduction
2. Statistique descriptive univarié
   - Variables qualitatives
   - Variables quantitatives
3. Statistique descriptive bivariée
   - Deux variables quantitatives
   - Une variable quantitative et une qualitative
   - Deux variables qualitatives
4. Statistique inférentielle
   - Introduction à l'inférence statistique
   - Intervalle de confiance
   - Tests d'hypothèses

Variables qualitatives

Effectif

Soit un individu décrit par une variable qualitative x pouvant prendre c modalités $(a_1, a_2, ..., a_i, ...a_c)$.

- **L'effectif**, aussi appelé fréquence absolue, est le nombre d'individus n_i dont la variable x présente la modalité a_i.
- Le diagramme en bâtons ou graphique en barres permet de représenter les effectifs des différentes modalités d'une variable qualitative.
- `scipy.stats.itemfreq`

Fréquence

Soit une observation décrite par une variable qualitative x pouvant prendre c modalités $(a_1, a_2, ..., a_i, ...a_c)$.

- **La fréquence** f_j de la modalité a_i est le rapport entre l'effectif et le nombre total d'individus.

$$f_j = \frac{n_i}{N}$$

  N est le nombre total d'individus.
  n_i le nombre d'individus dont la variable x présente la modalité a_i
- Le diagramme à tarte permet de présenter la fréquence de données qualitatives (catégorielles)

Effectifs cumulés

Soit une observation décrite par une variable qualitative x pouvant prendre c modalités $(a_1, a_2, ..., a_i, ...a_c)$.

- A chaque valeur a_i, on peut associer **l'effectif cumulé** N_i correspondant au nombre d'observations inférieures ou égales à a_i et défini par

$$N_i = n_1 + n_2 + ... + n_i$$

- Comme son nom l'indique, l'effectif cumulé associé à la valeur a_i est obtenu en « cumulant » (sommant) les effectifs associés aux valeurs distinctes observées qui sont inférieures ou égales à a_i.

Fréquence cumulées

- On peut également associer à chaque valeur a_i **la fréquence cumulée** F_i.
- Cette fréquence correspond à la proportion d'observations (dans l'ensemble de données) inférieures ou égales à F_j

$$F_j = f_1 + f_2 + ... + f_j = \frac{N_j}{n}$$

- `scipy.stats.cumfreq`

Sommaire des indicateurs descriptifs des variables qualitatives

Les valeurs des observations — Les effectifs — Les effectifs cumulés — Les fréquences — Les fréquences cumulées

Valeur/modalité (ordre croissant)	Eff.	Eff. cumulé	Fréq.	Fréq. cumulée
x_{o1}	n_1	$N_1 = n_1$	f_1	$F_1 = f_1 = N_1/n$
x_{o2}	n_2	$N_2 = n_1 + n_2$	f_2	$F_2 = f_1 + f_2 = N_2/n$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_{oj}	n_j	$N_j = n_1 + \ldots + n_j$	f_j	$F_j = f_1 + \ldots + f_j = N_j/n$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_{oJ}	n_J	$N_J = n$	f_J	$F_J = 1$
Total	n		1	

Exemple : indicateurs descriptifs des variables qualitatives (1)

Avis des étudiants sur l'enseignement à distance

Avis	n_j	N_j	f_j	F_j
T. défavorable	1	1	0.05	0.05
Défavorable	2	3	0.10	0.15
Moyen	11	14	0.55	0.70
Favorable	4	18	0.20	0.90
T. favorable	2	20	0.10	1
Total	n=20		1	

Exemple : indicateurs descriptifs des variables qualitatives (2)

Avis des étudiants sur l'enseignement à distance

Un seul étudiant a émis un avis très défavorable

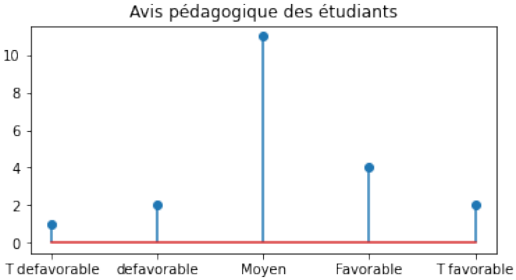
Avis	n_j	N_j	f_j	F_j
T. défavorable	1	1	0.05	0.05
Défavorable	2	3	0.10	0.15
Moyen	11	14	0.55	0.70
Favorable	4	18	0.20	0.90
T. favorable	2	20	0.10	1
Total	n=20		1	

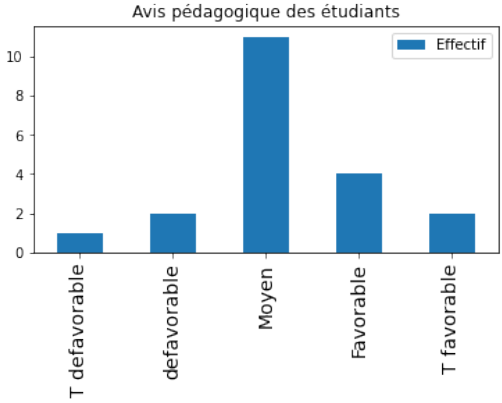
Somme =1

Au total 14 étudiants ont émis un avis moyen ou moins bon que moyen

Ce qui correspond à 70% de l'effectif d'étudiants

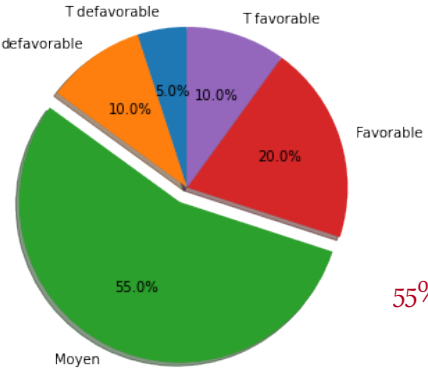
Exemple : Diagramme en bâtons





Diagrammes en bâtons pour visualiser l'effectif des avis des étudiants

Exemple : Diagramme à tarte



Le diagramme circulaire permet de visualiser les effectifs f_i

55% des étudiants ont un avis « Moyen »

Plan du cours

1. Introduction
2. Statistique descriptive univarié
   - Variables qualitatives
   - Variables quantitatives
3. Statistique descriptive bivariée
   - Deux variables quantitatives
   - Une variable quantitative et une qualitative
   - Deux variables qualitatives
4. Statistique inférentielle
   - Introduction à l'inférence statistique
   - Intervalle de confiance
   - Tests d'hypothèses

Familles de description des variables quantitatives (1)

- **Caractéristiques de tendances centrales** : indiquent l'ordre de grandeur des données et de leur valeur centrale, c'est-à-dire la position autour de laquelle se rassemblent ces valeurs.
- **Caractéristiques de dispersion** : servent à évaluer la variabilité des données et à résumer l'éloignement de l'ensemble des individus par rapport à leur tendance centrale.
- **Caractéristiques de formes** : De nombreux phénomènes physiques se distinguent par des variables quantitatives qui suivent une loi normale. Ces principales caractéristiques de formes sont le coefficient d'asymétrie et l'aplatissement de Fisher.

Familles de description des variables quantitatives (2)

Nous distinguons trois grandes familles :

Caractéristiques de tendances centrales	Caractéristiques de dispersions	Caractéristiques de formes
• La moyenne arithmétique • La moyenne arithmétique pondérée • La médiane • Les quartiles	• L'étendue • La variance et écart-type	• Asymétrie • Aplatissement

Caractéristiques de tendances centrales :

La moyenne arithmétique

- **La moyenne arithmétique** μ est la somme des valeurs de la variable j pour tous les individus i, $i = 1, ..., N$.

$$\mu_j = \frac{1}{N} \sum_{i=1}^{N} x_i^j \tag{1}$$

- La moyenne arithmétique est sensible aux valeurs aberrantes.
- `statistics.mean`

La moyenne arithmétique pondérée

- Lorsque les variables n'ont pas la même importance, on attribue un poids à chacune d'entre elles.
- Dans ce cas, on calcule **a moyenne arithmétique pondérée**

$$\lambda_j = \frac{\sum_{i=1}^{N} w_i x_i^j}{\sum_{i=1}^{N} w_i} \tag{2}$$

Le mode

- En statistique, **le mode**, ou valeur dominante, est la valeur la plus représentée d'une variable quelconque dans une population donnée.
- La répartition d'une variable peut être unimodale ou plurimodale (bimodale, trimodale...)
- `statistics.mode`

La médiane

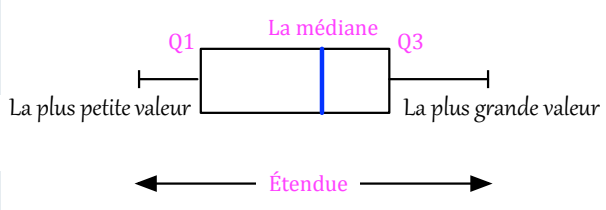
- Soit un ensemble de N données rangées par ordre croissant. **La médiane** est la valeur de la variable qui partage l'ensemble des données en deux parties de même effectif.
  - Si N est impair ($N = 2n + 1$), alors la médiane est la donnée de rang n.
  - Si N est pair ($N = 2n$), alors la médiane est la donnée de rang n ou de rang $n + 1$ ou bien la moyenne des deux.
- Dans le cas d'une distribution normale, la médiane et la moyenne sont égales.
- `statistics.median`

Les quartiles (1)

- Soit un ensemble de N données rangées par ordre croissant. **Les quartiles** sont les valeurs Q1, Q2, Q3 de la variable qui partagent l'effectif en quatre sous-ensembles de même effectif.
  - Q1 : Le premier quartile = la plus petite donnée de cet ensemble telle qu'au moins un quart des données sont inférieures ou égales à Q1.
  - Q2 : Le deuxième quartile = médiane.
  - Q3 : Le troisième quartile = la plus petite donnée telle qu'au moins les trois quarts des données de l'ensemble de données ordonnées sont inférieurs ou égaux à Q3.
- `statistics.quantiles`

Les quartiles (2)

- **Diagramme en boite** est l'un des graphiques les plus couramment utilisés pour visualiser les quartiles.
- Il résume cinq paramètres importants de la variable : la valeur minimale, le premier quartile (Q1), la médiane (ou deuxième quartile Q2), le troisième quartile (Q3) et la valeur maximale.



Caractéristiques de dispersion

L'étendue

- **L'étendue** est l'écart entre la plus grande et la plus petite des valeurs.
- Cette caractéristique, qui dépend des valeurs aberrantes, est par conséquent peu fiable.
- **Diagramme en boite** permet de visualiser l'étendue.

La variance et écart-type

- La variance est la moyenne des carrés des écarts à la moyenne.

$$var(x) = \sigma_x^2 = \frac{1}{n}\sum_{i=1}^{N}(x_i - \mu)^2$$

avec μ la moyenne arithmétique de l'ensemble de données.

- L'écart-type σ_x est la racine carrée positive de la variance :

$$\sigma_x = \sqrt{\frac{1}{n}\sum_{i=1}^{N}(x_i - \mu)^2}$$

- `statistics.variance` et `statistics.stdev`

Caractéristiques de formes

Le coefficient d'asymétrie (1)

- Le coefficient d'asymétrie (*skewness* en anglais) correspond au moment d'ordre trois de la variable centrée réduite. Pour une distribution uniforme x, le coefficient d'asymétrie est donné par la formule :

$$\gamma_1 = E\left[\left(\frac{x-\mu}{\sigma}\right)^3\right]$$

où E désigne l'espérance de x, μ la moyenne et σ l'écart type.

Le coefficient d'asymétrie (2)

$$\gamma_1 = \frac{\mu_3}{{\mu_2}^{3/2}}$$

avec μ_k le moment centré d'ordre k fournit par : $\mu_k = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^k$

- Trois cas se présentent :
  - $\gamma_1 = 0$ si la distribution de la variable est symétrique
  - $\gamma_1 > 0$ si la distribution de la variable est décalée vers la gauche et la moyenne est supérieure à la médiane
  - $\gamma_1 < 0$ si la distribution de la variable est décalée vers la droite et la moyenne est inférieure à la médiane
- `scipy.stats.skew`

Le coefficient d'aplatissement (1)

- Le coefficient d'aplatissement (*kurtosis* en anglais) correspond au quotient du moment d'ordre quatre de la variable centrée réduite par la puissance quatrième de l'écart type. Il est donné par l'équation suivante :

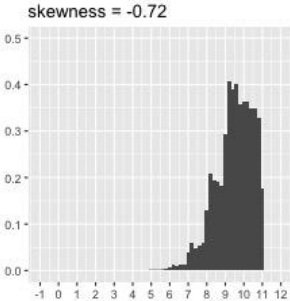
$$\gamma_2 = E\left[\left(\frac{X-\mu}{\sigma_x}\right)^4\right]$$

Le coefficient d'aplatissement (2)

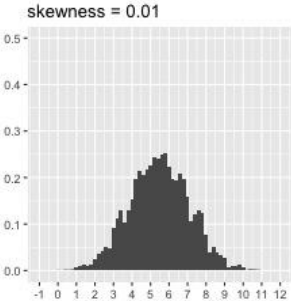
$$\gamma_2 = E\left[\left(\frac{X-\mu}{\sigma_x}\right)^4\right]$$

- Si $\gamma_2 = 0$ alors la distribution est dite **musokurtique**.
- Si $\gamma_2 > 0$, alors la distribution est plus concentrée que la normale ; elle est dite **leptokurtique**. Dans le domaine de la finance, ce coefficient sert à déterminer des valeurs anormales plus fréquentes.
- Si $\gamma_2 < 0$, alors la distribution est plus aplatie que la normale. Elle est dite **platikurtique**. Dans le domaine de la finance, ce coefficient sert à déterminer des valeurs anormales plus fréquentes.
- `scipy.stats.kurtosis`

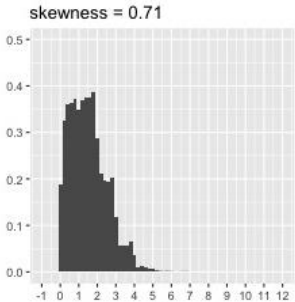
Exemple : caractéristique de dispersion (1)



Distribution décalée vers la droite

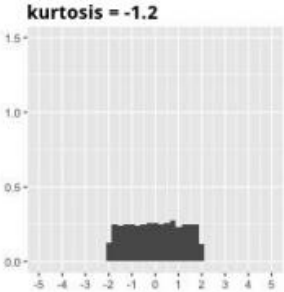


Distribution centrée

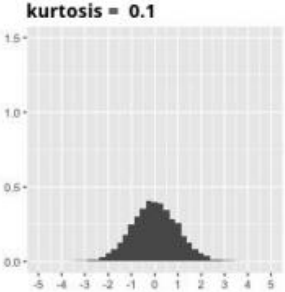


Distribution décalée vers la gauche

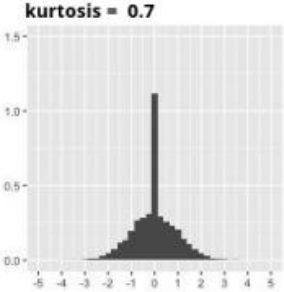
Exemple : caractéristique de dispersion (2)



Distribution plus aplatie que la distribution normale



Distribution ayant le même aplatissement que la distribution normale



Distribution plus concentrée que la distribution normale

Plan du cours

1. Introduction
2. Statistique descriptive univarié
   - Variables qualitatives
   - Variables quantitatives
3. Statistique descriptive bivariée
   - Deux variables quantitatives
   - Une variable quantitative et une qualitative
   - Deux variables qualitatives
4. Statistique inférentielle
   - Introduction à l'inférence statistique
   - Intervalle de confiance
   - Tests d'hypothèses

Statistique descriptive bivariée

- La statistique descriptive bidimensionnelle s'interesse à l'exploration simultannée de deux variables **x** et **y** étudiées sur le même échantillon.
- L'objectif essentiel est de mettre en évidence une éventuelle variation simultanée des deux variables $\Longrightarrow$ identifier les liaisons.
- Dans le cas de deux variables, les outils les plus utilisés sont :
  - Indicateurs : la corrélation et la statistique d'un test *a* associé à une table de contingence.
  - Représentations graphiques : le nuage de points (scatterplot), les diagrammes boîtes parallèles, les diagrammes de profils ou encore en mosaique.

Corrélation

- Par définition, la **corrélation** est une liaison entre des données quantitatives ou qualitatives, mise en évidence par une analyse statistique.
- Les études de corrélation permettent notamment de déterminer s'il existe ou non une association entre deux ensembles d'observations (inividus) et d'évaluer l'importance et le sens de cette association via le coefficient de corrélation.
- Formellement : l'analyse de corrélation a pour objectif d'examiner l'ensemble des données $(\mathbf{x}, \mathbf{y})$ en paires et de déterminer s'il y a une relation (association) entre elles.

Plan du cours

1. Introduction
2. Statistique descriptive univarié
   - Variables qualitatives
   - Variables quantitatives
3. Statistique descriptive bivariée
   - Deux variables quantitatives
   - Une variable quantitative et une qualitative
   - Deux variables qualitatives
4. Statistique inférentielle
   - Introduction à l'inférence statistique
   - Intervalle de confiance
   - Tests d'hypothèses

Deux variables quantitatives : Corrélation

Coefficient de corrélation linéaire (1)

- Soit deux variables quantitatives $\mathbf{x} = (x_1, x_2, ..., x_n)$ et $\mathbf{y} = (y_1, y_2, ..., y_n)$ décrivant n individus.
- Le **coefficient de corrélation linéaire**, aussi appelé **coefficient de corrélation de Pearson** de $\mathbf{x}$ et $\mathbf{y}$, mesure l'intensité de leur liaison.
- Ce coefficient de corrélation linéaire, souvent noté ρ, est égal au rapport de la covariance de $\mathbf{x}$ et $\mathbf{y}$ divisé par le produit des écarts-types des variables :

$$\rho(\mathbf{x}, \mathbf{y}) = \frac{\mathrm{Cov}(\mathbf{x}, \mathbf{y})}{\sigma_{\mathbf{x}} \sigma_{\mathbf{y}}} \tag{3}$$

Où Cov désigne la covariance des variables $\mathbf{x}$ et $\mathbf{y}$

Coefficient de corrélation linéaire (2)

$$\mathrm{Cov}(\mathbf{x}, \mathbf{y}) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_{\mathbf{x}})(y_i - \mu_{\mathbf{y}}) \tag{4}$$

$\sigma_{\mathbf{x}}$ et $\sigma_{\mathbf{y}}$ leurs écarts-types

$$\sigma_{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_{\mathbf{x}})^2 \quad \text{et} \quad \sigma_{\mathbf{y}} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \mu_{\mathbf{y}})^2 \tag{5}$$

et

$$\mu_{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{n}x_i \quad \text{et} \quad \mu_{\mathbf{y}} = \frac{1}{n}\sum_{i=1}^{n}y_i \tag{6}$$

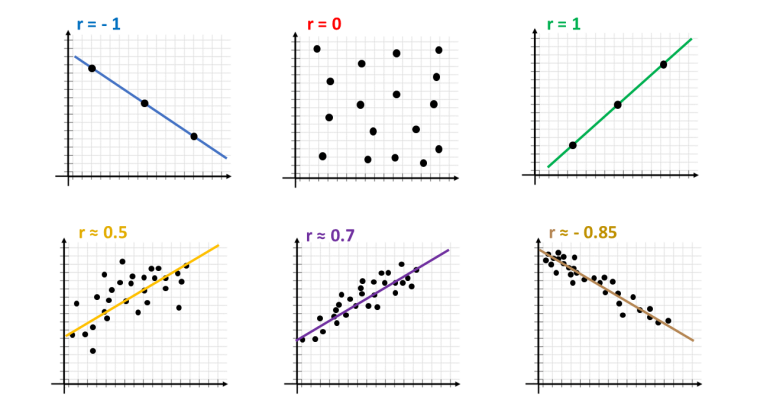
Coefficient de corrélation linéaire (3)

- Si $\rho(\mathbf{x}, \mathbf{y}) = -1$ ou $\rho(\mathbf{x}, \mathbf{y}) = 1$, alors la corrélation est dite parfaite. Dans ce cas, les points (x_i, y_i) se trouvent tous sur la droite du diagramme de dispersion
- Si $\rho(\mathbf{x}, \mathbf{y}) = 0$, alors il n'y a pas de corrélation linéaire entre $\mathbf{x}$ et $\mathbf{y}$.
- Si $\rho(\mathbf{x}, \mathbf{y}) > 0$, alors la corrélation est dite positive, c'est-à-dire qu'une augmentation de $\mathbf{x}$ entraîne une augmentation de $\mathbf{y}$
- Si $\rho(\mathbf{x}, \mathbf{y}) < 0$, alors la corrélation est dite négative (ou la relation est inversement proportionnelle), c'est-à-dire qu'une augmentation de $\mathbf{x}$ entraîne une diminution de $\mathbf{y}$.
- `pandas.DataFrame.corr` ou bien `scipy.stats.pearsonr`

Diagramme de dispersion

- La mesure de la corrélation linéaire entre deux variables **x** et **y** est souvent précédée par une visualisation graphique des relations entre ces variables.
- Le **diagramme de dispersion** est un graphique dans lequel les données $(\mathbf{x}, \mathbf{y})$ sont placées une sur l'axe des abscisses et l'autre sur l'axe des ordonnées.
- Lorsqu'il y a corrélation entre **x** et **y**, les points, qui représentent chacun un individu, forment une ligne assez droite.
- `seaborn.scatterplot`

Représentation graphique



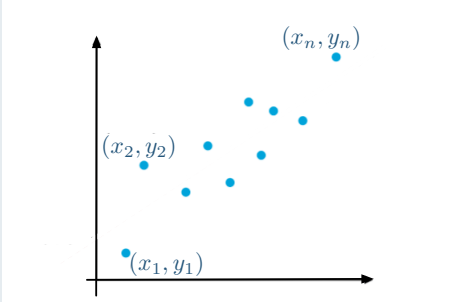
Deux variables quantitatives : Régression linéaire

Régression linéaire simple

- La corrélation linéaire a pour objectif d'analyser des relations entre des paires de variables.
- L'objectif de la **régression linéaire** est de décrire ces relations grâce à un modèle = une équation.
- Cette équation exprime la relation entre une variable indépendante x (aussi appelée « variable explicative » ou « prédictive ») et une variable dépendante y aussi appelée «variable expliquée» ou «variable réponse».

Les données

- Un ensemble de données d'apprentissage qui comprend n individus $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$ avec :
  - x_i est la variable explicative
  - y_i est la variable expliquée.

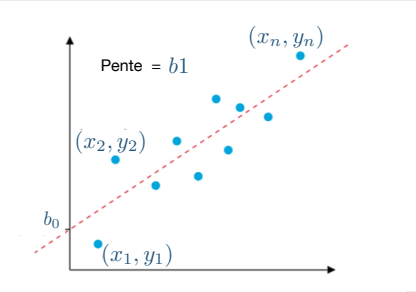


Modèle de régression linéaire simple (1)

- Le modèle de régression simple est décrit par une équation nommée **équation de régression**
- Soit l'équation suivante :

$$\mathbf{y} = b_0 + b_1\mathbf{x} \qquad (7)$$

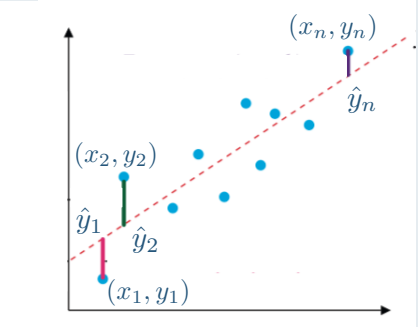
b_0 est l'ordonnée à l'origine et b_1 est la pente de la droite.



Modèle de régression linéaire simple (2)

- Étant donné un ensemble de n d'individus, l'équation de régression est donnée par :

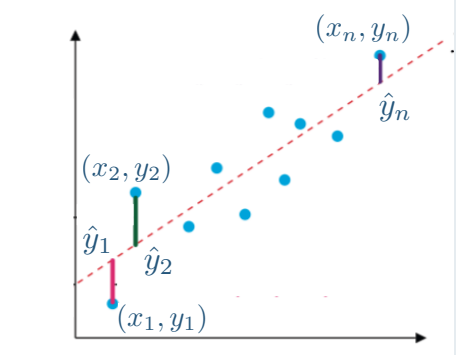
$$\hat{\mathbf{y}} = \beta_0 + \beta_1 \mathbf{x} \qquad (8)$$



Modèle de régression linéaire simple (3)

- L'application de l'équation 8 à l'ensemble des données permet d'écrire :

$$\begin{cases} \hat{y}_1 = \beta_0 + \beta_1 x_1 \\ \hat{y}_2 = \beta_0 + \beta_1 x_2 \\ \dots \\ \hat{y}_n = \beta_0 + \beta_1 x_n \end{cases} \quad (9)$$



Évaluation du modèle de régression (1)

- Si le modèle est parfait, alors il n'y aurait pas d'erreur, et donc pas de variations entre les valeurs prédites et les valeurs réelles $\implies$ Le modèle explique 100% de la variation.
- Ce pourcentage est calculé grâce à la formule de décomposition de la variance (analysis of variance, en anglais : ANOVA).

$$SCT = \underbrace{\sum_{i=1}^{n}(\hat{y}_i - \mu)^2}_{\text{Somme des carrées expliquées (SCE)}} + \underbrace{\sum_{i=1}^{n}(\hat{y}_i - \hat{y}_i)^2}_{\text{Somme des carrées de résidus (SCR)}}$$

Évaluation du modèle de régression (2)

- Pour la régression linéaire, le pourcentage de variation expliquée est donné par le coefficient de détermination R^2 :

$$R^2 = \frac{SCE}{SCT} \tag{10}$$

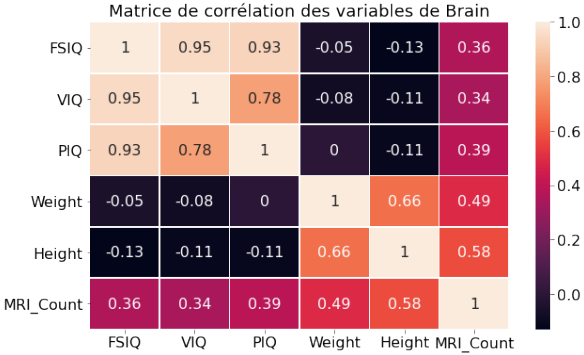
avec SCE : Somme des carrées expliquées
et SCT : Somme des carrées de résidus.

Exemple : Analyse de la corrélation

Soit la base données décrivant les caractéristiques de cerveaux (Willerman,1991) :

1. le genre : Homme ou femme
2. FSIQ : Score de QI complet basé sur les quatre sous-tests de Wechsler (1981)
3. VIQ : Score de QI verbal basé sur les quatre sous-tests de Wechsler
4. PIQ : Score de QI de performance basé sur les 4 sous-tests de Wechsler
5. poids : poids du corps en livres
6. hauteur : hauteur en pouces
7. MRI_Count : nombre total de pixels des 18 balayages IRM
8. La variable activity qui décrit l'activité du participant a été ajouté pour les simulations du cours

Exemple : Matrice de corrélation et carte de chaleur

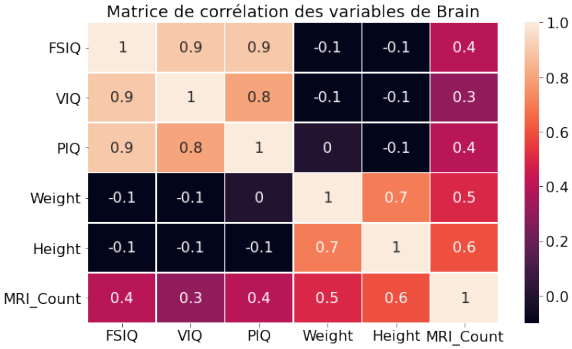



```
matrice_corr = data.iloc[:, 1:8].corr().round(2)
print(matrice_corr)
```

	FSIQ	VIQ	PIQ	Weight	Height	MRI_Count
FSIQ	1.00	0.95	0.93	-0.05	-0.09	0.36
VIQ	0.95	1.00	0.78	-0.08	-0.07	0.34
PIQ	0.93	0.78	1.00	0.00	-0.08	0.39
Weight	-0.05	-0.08	0.00	1.00	0.70	0.51
Height	-0.09	-0.07	-0.08	0.70	1.00	0.60
MRI_Count	0.36	0.34	0.39	0.51	0.60	1.00

Corréation forte entre:
- FSIQ et PIQ
- FSIQ et VIQ
- Weight et Height

Exemple : Coefficient de corrélation de Pearson entre PIQ et VIQ




```python
matrice_corr = data.iloc[:, 1:8].corr().round(2)
print(matrice_corr)
```

```
             FSIQ   VIQ   PIQ  Weight  Height  MRI_Count
FSIQ         1.00  0.95  0.93   -0.05   -0.09       0.36
VIQ          0.95  1.00  0.78   -0.08   -0.07       0.34
PIQ          0.93  0.78  1.00    0.00   -0.08       0.39
Weight      -0.05 -0.08  0.00    1.00    0.70       0.51
Height      -0.09 -0.07 -0.08    0.70    1.00       0.60
MRI_Count    0.36  0.34  0.39    0.51    0.60       1.00
```

Corréation forte entre:
- FSIQ et PIQ
- FSIQ et VIQ
- Weight et Height

Exemple : Modèle de régression

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                    PIQ   R-squared:                       0.605
Model:                            OLS   Adj. R-squared:                  0.595
Method:                 Least Squares   F-statistic:                     58.32
Date:                Thu, 18 Feb 2021   Prob (F-statistic):           3.44e-09
Time:                        23:16:32   Log-Likelihood:                -162.14
No. Observations:                  40   AIC:                             328.3
Df Residuals:                      38   BIC:                             331.7
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     27.8404     11.125      2.503      0.017       5.320      50.361
VIQ            0.7404      0.097      7.637      0.000       0.544       0.937
------------------------------------------------------------------------------
```

—> Ordinary least squares

Le coefficient de détermination R2 = 0.605 est élevé ce qui démontre une forte corrélation entre les variables PIQ et VIQ

Plan du cours

1. Introduction
2. Statistique descriptive univarié
   - Variables qualitatives
   - Variables quantitatives
3. Statistique descriptive bivariée
   - Deux variables quantitatives
   - Une variable quantitative et une qualitative
   - Deux variables qualitatives
4. Statistique inférentielle
   - Introduction à l'inférence statistique
   - Intervalle de confiance
   - Tests d'hypothèses

Une variable quantitative et une qualitative

Une variable quantitative et une qualitative : ANOVA

- Soit une observation décrite par une variable qualitative y pouvant prendre c modalités $(a_1, a_2, ..., a_i, ...a_c)$. et une variable quantitative x de moyenne μ et de variance σ.
- L'analyse de la relation entre x et y se fait à l'aide de **l'analyse de variance** (ANOVA).

Une variable quantitative et une qualitative

Une variable quantitative et une qualitative : ANOVA

- L'objectif de ce test est de conclure sur l'influence d'une variable catégorielle sur la loi d'une variable continue à expliquer.
- Les deux variables ne jouent pas un rôle symétrique.
- Par exemple : On veut étudier l'influence d'une variable qualitative (le sexe) sur une variable quantitative (le salaire).
- `statsmodels.stats.anova.anova_lm`

Évaluation du modèle d'ANOVA

- Pour des variables quantitatives, nous avons utilisé le coefficient de détermination R^2
- De la même manière pour une variable quantitative et une variable qualitative, on calcule le rapport de corrélation selon :

$$\eta^2_{y/x} = \frac{V_{interclasses}}{V_{totale}} \tag{11}$$

- Si $\eta^2_{y/x} = 0$ cela signifie que les moyenne par classes sont toutes égales.
- Si $\eta^2_{y/x} = 1$ cela signifie que les moyenne par classes sont très différentes.

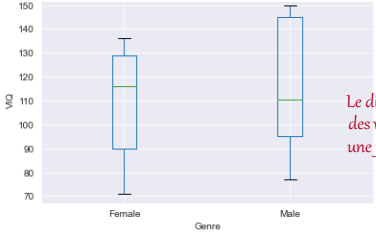
Une variable quantitative et une qualitative : Visualisation

- Une façon commode de représenter les données dans le cas de l'étude simultanée d'une variable qualitative x et une variable y quantitative consiste à réaliser des **diagrammes de boites parallèles**.
- Il s'agit de représenter, sur un même graphique doté d'une échelle unique, un diagramme en boîte pour chacune des modalités de la variable qualitative x.

Exemple : Une variable quantitative et une qualitative

```
                            OLS Regression Results
==============================================================================
Dep. Variable:              data["VIQ"]   R-squared:                       0.015
Model:                              OLS   Adj. R-squared:                 -0.010
Method:                   Least Squares   F-statistic:                    0.5969
Date:                  Sat, 20 Feb 2021   Prob (F-statistic):              0.445
Time:                          20:46:01   Log-Likelihood:                -182.42
No. Observations:                    40   AIC:                             368.8
Df Residuals:                        38   BIC:                             372.2
Df Model:                             1
Covariance Type:              nonrobust
==============================================================================
                             coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept                  109.4500      5.308     20.619      0.000      98.704     120.196
data["Gender"][T.Male]       5.8000      7.507      0.773      0.445      -9.397      20.997
```

Le coefficient de détermination R^2 = 0.015 est faible ce qui démontre une faible relation entre les variables Gender et VIQ



Le diagramme de boites parallèles démontre un grand chevauchement des valeurs de la variable VIQ pour les deux genres ce qui démontre une faible relation entre les variables Gender et VIQ

Plan du cours

1. Introduction
2. Statistique descriptive univarié
   - Variables qualitatives
   - Variables quantitatives
3. Statistique descriptive bivariée
   - Deux variables quantitatives
   - Une variable quantitative et une qualitative
   - Deux variables qualitatives
4. Statistique inférentielle
   - Introduction à l'inférence statistique
   - Intervalle de confiance
   - Tests d'hypothèses

Deux variables qualitatives : Table de contingence

Table de contingence

- Soit les deux variables qualitatives **x** et **y** observées simultanément sur n individus
- On suppose que **x** possède les modalités $(x_1, x_2, ..., x_l, ..., x_r)$ et que **y** possède les modalités $(y_1, y_2, ..., x_h, ..., x_c)$.
- Ces données sont présentées dans un tableau à double entrée, appelé **table de contingence**, dans lequel on dispose les modalités de **x** en lignes et celles de **y** en colonnes.

Table de contingence

Modalités de la variable x

Effectif conjoint entre la modalité x_1 et la modalité y_h

Modalités de la variable y

	y_1	$\cdots$	y_h	$\cdots$	y_c	sommes
x_1	n_{11}	$\cdots$	n_{1h}	$\cdots$	n_{1c}	n_{1+}
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
x_ℓ	$n_{\ell 1}$	$\cdots$	$n_{\ell h}$	$\cdots$	$n_{\ell c}$	$n_{\ell +}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
x_r	n_{r1}	$\cdots$	n_{rh}	$\cdots$	n_{rc}	n_{r+}
sommes	n_{+1}	$\cdots$	n_{+h}	$\cdots$	n_{+c}	n

Effectifs marginaux de la modalité x_l

Deux variables qualitatives : Test $Khi - deux$

Test $Khi - deux$ (1)

- Il est courant en statistique de comparer une table de contingence observée, d'effectif conjoint générique n_{lh} à une table de contingence donnée a priori (et appelée standard), d'effectif conjoint générique s_{lh}, en calculant la quantité :

$$\sum_{l=1}^{r}\sum_{h=1}^{c}\frac{(n_{lh}-s_{lh})^2}{s_{lh}} \tag{12}$$

Test *Khi – deux* (2)

- L'indice χ^2 est donné par :

$$\chi^2 = n\left[\sum_{l=1}^{r}\sum_{h=1}^{c}\frac{(n_{lh}-s_{lh})^2}{s_{lh}} - 1\right] \tag{13}$$

- χ^2 est toujours positif ou nul
- Plus χ^2 est élevé, plus les deux variables sont liées.
- `scipy.stats.chi2_contingency`

Exemple : Deux variables qualitatives

```
from scipy.stats import chi2_contingency
table= pd.crosstab(data['Gender'], data['Activity'])
table
```

Activity	artist	doctor	farmer	other	professor
Gender					
Female	8	1	0	2	9
Male	0	11	7	0	2

La table de contingence démontre une relation entre l'activité et le genre

La valeur de Chi2 est élevée (comparée à 0)
—> une relation entre l'activité et le genre

```
results = chi2_contingency(table)
print('Chi2:' , results[0])
print('p_valeur:' , results[1])
print('degré de liberté:' , results[2])
```

```
Chi2: 29.78787878787879
p_valeur: 5.405997743214392e-06
degré de liberté: 4
```

La valeur de $p<0.05$
—> cette relation est statiquement significative

Le degré de liberté est égal à 4 (5-1)
—> puisque nous avons 5 variables

Plan du cours

1. Introduction
2. Statistique descriptive univarié
   - Variables qualitatives
   - Variables quantitatives
3. Statistique descriptive bivariée
   - Deux variables quantitatives
   - Une variable quantitative et une qualitative
   - Deux variables qualitatives
4. Statistique inférentielle
   - Introduction à l'inférence statistique
   - Intervalle de confiance
   - Tests d'hypothèses

Plan du cours

1. Introduction
2. Statistique descriptive univarié
   - Variables qualitatives
   - Variables quantitatives
3. Statistique descriptive bivariée
   - Deux variables quantitatives
   - Une variable quantitative et une qualitative
   - Deux variables qualitatives
4. Statistique inférentielle
   - Introduction à l'inférence statistique
   - Intervalle de confiance
   - Tests d'hypothèses

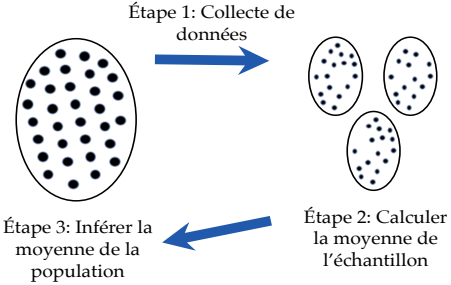
Introduction à l'inférence statistique

Terminologie

- **Population :** Une population (ou « population mère ») est définie comme l'ensemble de toutes les données collectées pour une étude particulière.
- **Échantillon :** Un échantillon est défini comme un sous-ensemble de la population (un sous-ensemble de données).
- **Individu :** Un individu est défini comme un élément d'une certaine population.

Principe

Le principe de l'inférence statistique est d'estimer les paramètres d'une population en se basant sur les données observées d'un échantillon.



Idéalement...

- On désire travailler avec des données couvrant toute la population visée, sans aucune exception
- On obtiendrait alors la distribution exacte des valeurs de tous les individus de la population sur une/plusieurs variables.
- On obtiendrait des paramètres, c-à-d des valeurs exactes !

Réellement...

- Travailler avec l'ensemble de la population est impossible ou hors de prix ou simplement trop long à réaliser.
- On procède alors par échantillonnage en vue d'obtenir des échantillons (on présume qu'ils sont représentatifs de la population)
- On obtient alors des estimés, plutôt que des paramètres

Plan du cours

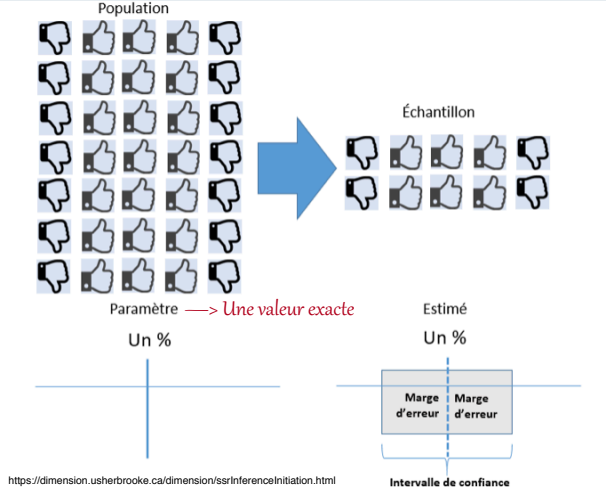
1. Introduction
2. Statistique descriptive univarié
   - Variables qualitatives
   - Variables quantitatives
3. Statistique descriptive bivariée
   - Deux variables quantitatives
   - Une variable quantitative et une qualitative
   - Deux variables qualitatives
4. Statistique inférentielle
   - Introduction à l'inférence statistique
   - Intervalle de confiance
   - Tests d'hypothèses

Intervalle de confiance (1)

Formellement... (1)

- Formellement, l'objectif de l'inférence statistique est d'estimer le paramètre θ qui permet de décrire la population en question.
- Le paramètre θ correspond très souvent à la moyenne qui est estimé par la moyenne de l'échantillon considéré
- Cependant, cet estimé est affecté par une certaine erreur. D'où l'intérêt d'estimer cette erreur en calculant ses bornes supérieure et inférieure qui correspondent à l'intervalle de confiance.

Formellement... (2)



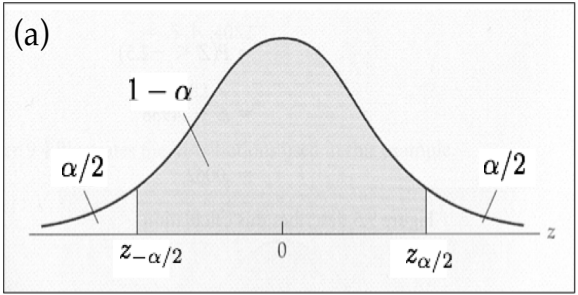
Définition d'un intervalle de confiance (2)

- Soit $(x_1, x_2, ..., x_n)$ un ensemble de n échantillons aléatoires et θ un paramètre inconnu de la loi des x_i.
- Soit $\alpha \in]0, 1[$. S'il existe des v.a.r. $\theta_{min}(x_1, ..., Xn)$ et $\theta_{max}(x_1, ..., x_n)$ telles que :

$$P(\theta \in [\theta_{min}(x_1, ..., x_n), \theta_{max}(x_1, ..., x_n)]) = 1 - \alpha,$$

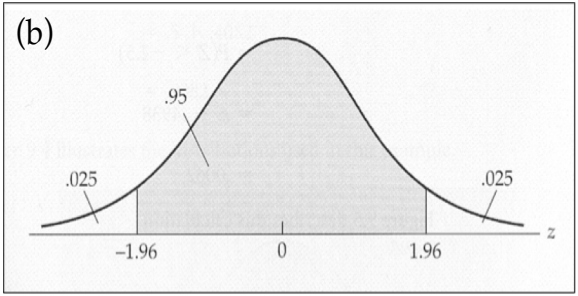
on dit alors que $[\theta_{min}(x_1, ..., x_n), \theta_{max}(x_1, ..., x_n)]$ est un intervalle de confiance pour θ, avec coefficient de sécurité $1 - \alpha$. On le note $IC_{1-\alpha}(\theta)$.

Définition d'un intervalle de confiance (3)



Intervalle centré sur la moyenne qui contient $1-\alpha$ des données

Exemple $\alpha = 0.5$



Intervalle centré sur la moyenne qui contient 95% des données

Formalisme

- Dans la pratique, on peut prendre par exemple $\alpha = 5\%$ ce qui nous donne un IC à 95%. Cela signifie qu'il y a 95% de chance que la valeur inconnue θ soit comprise entre $\theta_{min}(x_1, ..., x_n)$ et $\theta_{max}(x_1, ..., x_n)$
- Supposons que l'on désire calculer l'intervalle de confiance pour la moyenne d'une population dont l'écart type est connu. Ce calcul peut se faire en suivant les étapes suivantes :

Étapes de détermination de l'IC (1)

1. Calculer la moyenne μ et l'écart-type σ de l'échantillon.

$$\mu = \frac{\sum_{i=1}^{n} x}{n} \quad \text{et} \quad \sigma = \sqrt{\frac{\sum_{i=1}^{n}(x-\mu)^2}{n}} \tag{14}$$

2. Choisir le niveau de confiance associé à l'intervalle de confiance :

$$100(1-\alpha)\%$$

Les niveaux de confiance les plus couramment utilisés sont 90%, 95 % et 99% ce qui correspond à une valeur de α égale respectivement à 0.1, 0.05 et 0.001.

Étapes de détermination de l'IC (1)

3. Calculer la marge d'erreur en se basant sur la formule suivante :

$$z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

4. Établir l'intervalle de confiance selon :

$$I_C = \left[\mu - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\ ,\ \mu + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right] \quad (15)$$

Déviation standard

$$z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

Coefficient de confiance

Nombre d'échantillons

Plan du cours

1. Introduction
2. Statistique descriptive univarié
   - Variables qualitatives
   - Variables quantitatives
3. Statistique descriptive bivariée
   - Deux variables quantitatives
   - Une variable quantitative et une qualitative
   - Deux variables qualitatives
4. Statistique inférentielle
   - Introduction à l'inférence statistique
   - Intervalle de confiance
   - Tests d'hypothèses

Tests d'hypothèses

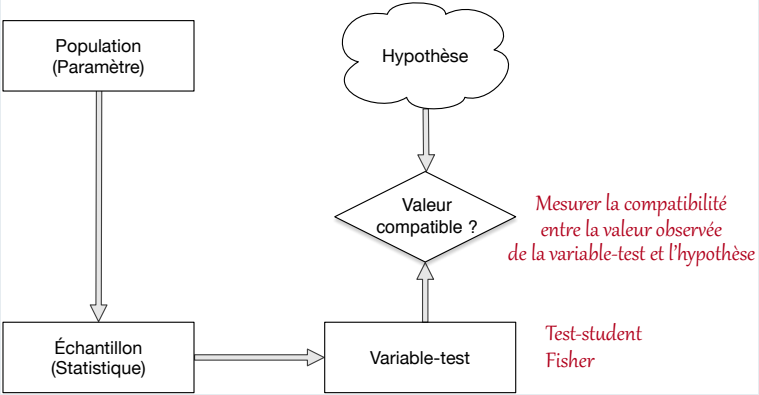
Tests d'hypothèses : Principe (1)

- Rappelons que formellement, l'objectif de l'inférence statistique est d'estimer le paramètre θ qui permet de décrire la population en question.
- Les tests d'hypothèse constituent un autre aspect important de l'inférence statistique.

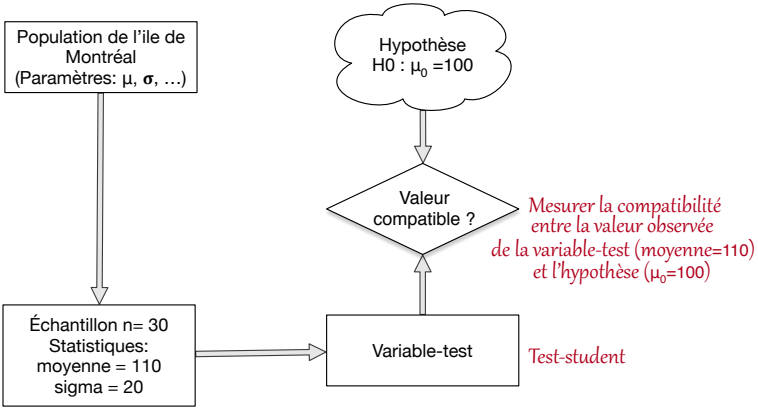
Tests d'hypothèses : Principe (2)

- Leur principe consiste à confirmer ou à contredire une hypothèse concernant la valeur d'un paramètre de la population en se basant sur le test de signification.
- En d'autres termes, il s'agit de déterminer si l'échantillon de taille n dont nous disposons appartient à une population de moyenne μ_0 au seuil de signification α.

Tests d'hypothèses :Principe (3)



Exemple : Tests d'hypothèses



Étapes d'un test d'hypothèses (1)

1. Formuler les hypothèses. Chaque test de signification commence par une hypothèse nulle H_0 qui représente une supposition qui soit proposée ou bien intuitivement ou bien parce qu'elle peut être utilisée comme point de départ du test d'hypothèse.
   - L'hypothèse nulle notée H0 est l'hypothèse que l'on désire contrôler : elle consiste à dire qu'il n'existe pas de différence entre les paramètres comparés ou que la différence observée n'est pas significative
   - Cette hypothèse est formulée dans le but d'être rejetée.

Étapes d'un test d'hypothèses (2)

2. L'hypothèse complémentaire à l'hypothèse nulle constitue l'hypothèse alternative (notée H_1 ou bien H_a). Le test d'hypothèse peut correspondre à une égalité ou une inégalité. Dans le cas de l'égalité, nous avons les relations suivantes :

$$\begin{cases} H_0 : m = \mu_0 \\ H_1 : m \neq \mu_0 \end{cases} \tag{16}$$

qui se lit : "L'hypothèse nulle correspond à une distribution dont la moyenne m est égale à μ_0".

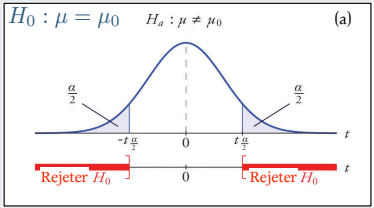
Étapes d'un test d'hypothèses (3)

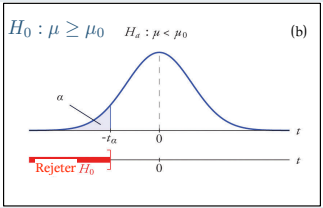
3. Choisir la distribution de probabilité du paramètre à estimer. Dans le cas d'un grand nombre d'échantillon ($n > 30$), la distribution la plus utilisée est la densité normale. Dans le cas contraire, d'autres types de distribution sont utilisées. Dans ces cas, les analyses sont plus complexes.
4. Choisir le niveau de confiance associé à l'intervalle de confiance :

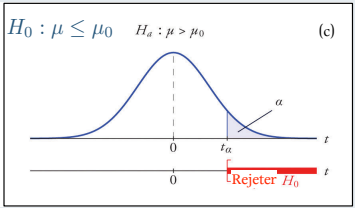
$$100(1 - \alpha)\%$$

Étapes d'un test d'hypothèses (4)

5. Calculer, à partir des données fournies par l'échantillon, la valeur de la statistique p.
6. Prendre une décision concernant l'hypothèse posée et interpréter les résultats.







Exemple : Test d'hypothèse

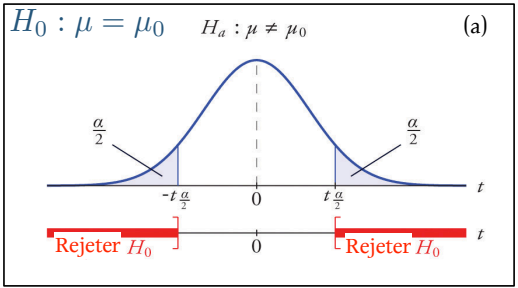
```
In [43]: results = stats.ttest_1samp(data['VIQ'], 0)
         print('statistic:' , results[0])
         print('p_valeur:' , results[1])

         statistic: 30.088099970849328
         p_valeur: 1.3289196468728067e-28
```

p< 0.05 —> nous rejetons l'hypothèse nulle

—> nous rejetons l'hypothèse nulle

—> La moyenne de la variable VIQ est différente de 0



Complément du cours...

Pourquoi $n = 30$?

- Le théorème Central-Limit (Théorème de de la limite centrale) a pour énoncé : Soit X_n, $n \in \mathcal{N}$, une suite de variables aléatoires indépendantes de même loi d'espérence m et de variance σ^2 finies. Alors, la somme $Z = \sum_{i=1}^{n} X_i$ converge pour n assez grand (en pratique à partir de $n = 30$) vers la loi normale de moyenne $n \times m$ et d'écart type $\sigma\sqrt{n}$
- Lorsque l'on fait la somme d'un très grand nombre de variables aléatoires de loi quelconque, cette somme suit une loi normale.