Statistique Appliquée

I- Statistique descriptive : ≠ probabiliste

L'ensemble des instruments et de recherches mathématiques permettant de déterminer les caractéristiques d'un ensemble de données.

. But : d'extraire et de résumer des informations pertinentes d'une liste de nombres difficile à interpréter par une simple lecture.

↳ Statistiques exploratoires : on explore d'abord les données pour avoir une idée qualitative de leurs propriétés

↳ Statistiques confirmatoires : on fait des hypothèses de comportement que l'on confirme ou que l'on infirme en recourant à d'autres techniques statistiques.

. Outils :

+ La collecte des données
+ Le traitement des données collectées
  . Utiliser des représentations graphiques
  . Calculer certains paramètres et indicateurs
+ Interprétation des résultats

. Basée sur des données existantes.

(≠) Statistique probabiliste : Basée sur des données antérieurs, dont je connais tout les résultats mais non pas un exact.

↳ Chercher la loi de probabilité.

(=) Inférentielle

. Inférence : une opération logique par laquelle on admet une proposition en vertu de sa liaison avec d'autres propositions déjà faites.

. L'inférence statistique : consiste à induire conclure des conclusions (caractéristiques) concernant un groupe auquel on ne peut pas accéder directement (grande taille - coûteux) à partir d'un sous-groupe (petite taille) auquel on a accès et que l'on considère comme un échantillon aléatoire provenant de cette population.

↳ Les caractéristiques de l'échantillon représentent une certaine marge d'erreur par rapport à celles de la populat°.

⇒ Consiste à partir d'un échantillon de données provenant d'une population de loi de probabilité inconnue, à déduire des propriétés fiables sur cette population.

Elle vise à :

. Évaluer un paramètre ou une relation
. Prédire une valeur
. Déterminer si les différences sont dues au hasard.

. Vocabulaire :

+ Population : Un ensemble de personnes, d'objets ou d'évenements, base de l'étude statistique.

+ Individu : Un élement de cette population.

+ Échantillon : un sous-ensemble de la population, ayant les mêmes caractéristiques de la population mère. utilisé en vue d'inférer qlq chose à cette pop.

+ Caractère : particularité ou propriété caractéristique de la population

Quantitatif :	Qualitatif :
Variable statistique	. Ordinale (ordonnance)
Discrète : Valeur ponctuelle / Continue : lorsqu'on a une évolution	. Nominale

+ Effectif d'une population : c'est le nombre total des élements constituant cette populat (N)

+ Fréquence d'un caractère : c'est le nombre d'individus possédant ce caractère divisé par N

. Indicateurs de forme : Graphiques et tableaux

! Les valeurs que prendra un caractère = Modalités doivent être ordonnées.

Modalités ou classes	Eff.	Eff. cumulé ↗	Eff. cumulé ↘	fréq.	fréq. cumulé ↗	fréq. cumulé ↘	%
m_i	n_i	↳ fréquence Absolue	Ncd $\geq m_i$	$\frac{n_i}{N}$	↳ fréquence relative	Fcd	100 x fréq

* VSD : on les représente par un histogramme, ou des secteurs

* VSC : ! classes adjacentes, d'amplitude pas forcément égales.

. On les représente par un histogramme dont les rectangles sont de largeur = amplitude.

. Dans la pratique, on prend généralement entre 5 et 20 intervalles.

. Pour déterminer l'amplitude de l'intervalle, Grd valeur - ptt valeur / nb d'intervalles.

* Variable qualitatif : représentation en secteurs.

. Valeurs numériques :

+ Issues d'un échantillon : stat. d'échantillon

+ Issues d'une populat : paramètres de la populat

. Statistique d'échantillon = Valeur numérique utilisée comme mesure d'un échantillon.

. Paramètres de la populat = Valeur numérique utilisée comme mesure de la population.

. Estimateur ponctuel = Statistique d'échantillon utilisée pour estimer le paramètre correspondant de la population.

. Indicateurs de tendance centrale :

+ Moyenne :

$$\bar{x} = \frac{\sum x_i}{N}$$ = échantillon

$$\mu = \frac{\sum x_i}{N}$$ = population

+ Médiane : Valeur qui partage l'ensemble des données en 2 parties égales (! ordre ↑)

On calcule : $\frac{N}{2}$

↳ décimale : on prend l'élement entier qui a la position après

↳ Entière : On somme les valeurs au position i et $(i+1)$ et on divise par 2.

Mode : La valeur de l'observation la plus fréquente (on peut avoir 2 modes)

+ Percentiles : on appelle p percentile, la valeur qui partage les observations en p% à gauche et (100-p)% à droite.

• La médiane correspond au 50e percentile.

• Les quartiles : 25e percentile, 50e percentile, 75e percentile

Calcul : $i = \frac{P}{100} * N$ { décimale : val. entier ↷ ; entier : $\frac{(i)^e + (i+1)^e}{2}$

+ Mesures de dispersion :

+ Étendue : La différence entre la plus grande valeur et la plus petite.

+ Étendue interquartile (EIQ) : La différence entre le 3e et le 1er quartiles $EIQ = Q_3 - Q_1$

= Étendue de 50% des observations qui sont au milieu.

+ Variance : { Populat° = $\sigma^2 = \frac{\sum (x_i - \mu)^2}{N}$; Échantillon = $s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1}$ }

+ Écart type : { $\sigma = \sqrt{\sigma^2}$; $s = \sqrt{s^2}$ }

+ Coefficient de variation : $\frac{\text{Ecart type}}{\text{moyenne}} * 100$

↳ Combien s'élève l'écart-type par rapport au moyenne.

Intérêt d'avoir le CV le + petit possible. ↓CV ⇒ ↓ dispersion

le CV est indépendant de l'unité de mesure.

+ Détection des valeurs singulières :

• Valeur singulière : observation anormalement grande ou petite.

↳ Erreur d'enregistrement : À corriger

↳ Observation pas correctement incluse dans l'ensemble des données : À supprimer

↳ Valeur inhabituel correctement enregistrée et qui appartient à l'ensemble des données : À conserver

→ Ces valeurs doivent être mentionner en matière d'interprétation (Biaiser l'étude)

→ Moyenne tronquée : calculer la moyenne de 90% de la populat° seulement (pour éviter les val. sing)

5% — 90% — 5%

• Forme d'une distribution :

→ Degré d'asymétrie : $\gamma_1 = \frac{\mu_3}{\sigma^3}$

Avec : $\mu_3 = \frac{1}{n} \sum_{i=1}^{k} n_i (x_i - \bar{x})^3$

* Des données biaisées à gauche $\gamma_1 < 0$

* Des données biaisées à droite $\gamma_1 > 0$

• Variable centrée réduite z :

$$z_i = \frac{x_i - \bar{x}}{s}$$ → distance

↳ mesure la distance en nb d'écart type entre une observation x_i et la moyenne.

• Théorème de Chebyshev :

Au moins $(1 - \frac{1}{z^2})$ des observations doivent se situer ds l'intervalle $[\bar{x} - zs, \bar{x} + zs]$;

Avec $z > 1$.

Ex : Combien des étudiants ont obtenu une note entre 60 et 80. Sachant que { $\bar{x} = 70$; $s = 5$ }

↳ $60 = \bar{x} - 2s = 70 - 2 \times 5$

$60 = 70 - 2 \times 5 \Rightarrow z = 2$

Donc $1 - \frac{1}{2^2} = 75\%$ des étudiants ont obtenu une note entre 60 et 80.

• Règle empirique :

Pour une distribution en forme de cloche :

* 68% des observations se situent ds $[\bar{x} - s, \bar{x} + s]$

* 95% " " $[\bar{x} - 2s, \bar{x} + 2s]$

* Presque toutes les observat° ... $[\bar{x} - 3s, \bar{x} + 3s]$ = 99,7%

⇒ Considérer que toutes les valeurs en dehors de $[\bar{x}-3s, \bar{x}+3s]$ sont des valeurs singulières.

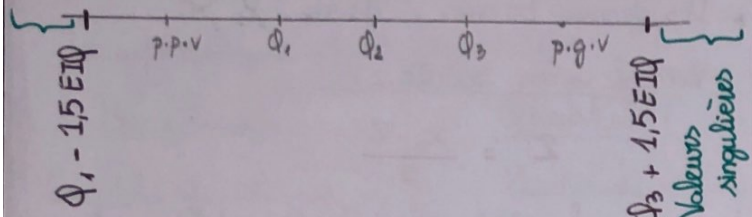
! Différence entre le th. de Chebyshev et la règle :

- $z > 1$
- $z = 1$
  - Elle ne marche que pour les distributions normales
  - Elle donne + de précisions

• Analyse exploratoire des données :

But : Éliminer les valeurs singulières

Données : + petite valeur, + grande valeur, Q_1, Q_2, Q_3



II - Statistique bivariée :

1 - Mesures de la relation entre 2 variables :

* La série statistique est représentée par un nuage de points.
* Dans la statistique bivariée, on peut calculer le point moyen $G(x_G, y_G)$ $\begin{cases} x_G = \frac{1}{n}\sum x_i = \bar{x} \\ y_G = \frac{1}{n}\sum y_i = \bar{y} \end{cases}$
* Ajustement affine
  + Ajustement à la règle (en passant par le pt moyen)
  + Méthode de Mayer = droite de Mayer (2 sous-nuages, 2 points moyens)

2 - Mesures de la covariance :

↳ Mesure de la relation linéaire entre 2 variables.

* Population = $\sigma_{xy} = \frac{\sum (x_i - \mu_x)(y_i - \mu_y)}{N}$
* Échantillon = $s_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n-1}$

↳ Valeurs > 0 ; Relation linéaire pos.

↳ Valeurs < 0 ; Relation linéaire négative

3 - Mesures par le coef. de corrélation :

↳ Mesure de la relation linéaire entre 2 variables, dont les valeurs sont comprises entre -1 et +1

* Population = $\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$
* Échantillon = $r_{xy} = \frac{s_{xy}}{s_x s_y}$

↳ Valeurs proches de 0 = Absence de relat° linéaire

↳ Valeurs proches de 1 = forte relat°. linéaire positive

↳ Valeurs proches de -1 = forte relation linéaire négative

↳ Méthode des moindres carrés :

* Droite de régression de Y en X

$(D_{Y/X}) : y = ax + b$ $\begin{cases} a = \frac{\sigma_{xy}}{(\sigma_x)^2} \\ b = \bar{y} - a\bar{x} \end{cases}$

* Droite de régression de X en Y

$(D_{X/Y}) : x = a'y + b'$ $\begin{cases} a' = \frac{\sigma_{xy}}{\sigma_y} \\ b' = \bar{x} - a'\bar{y} \end{cases}$

* les 2 droites se coupent au point moyen

La meilleure situat° c'est d'avoir 2 droites confondues : ↳ (↳ ↳)

Rq : • Valeur faible de r = il est possible d'avoir une forte relat° mais non linéaire

r = coef de corrélat° de la partie linéaire entre X et Y.

• 2 variables dont r proche de 0 : décorrélées ≠ indépendantes

• Un fort r n'implique pas une forte relation de causalité entre X et Y (Existance possible d'un 3ème var.)

• Passer d'une relat° non linéaire à une relation linéaire

Par exemple = $y = Cx^m$ ou $y = Ca^x$

$\Rightarrow \ln y = \ln Cx^m = \ln C + \ln x^m$

$\ln y = \underbrace{m \ln x + \ln C}_{Y} \Rightarrow \begin{cases} m = \frac{\sigma_{xy}}{(\sigma_x)^2} \\ C = e^b = e^{\bar{y} - a\bar{x}} \end{cases}$

$Y = aX + b$

III- Probabilités :

= Évaluation du caractère probable d'un événement.

1- Dénombrement - Théorie des ensembles :

- 2 ensembles égaux $A = B$ si $A \subseteq B \begin{cases} x \in A \\ x \in B \end{cases}$
- $A \subset B \ (\forall x \in A \Rightarrow x \in B)$: A sous ensemble de B.
- $\emptyset$ Ensemble qui ne contient aucun élément.
- $A \cap B$ (intersection) $\{x \in A$ et $x \in B\}$:
- $A \cup B$ (Union) $\{x \in A$ ou $y \in B\}$
- $A \setminus B = \{x \in A$ et $x \notin B\} = A \setminus A \cap B$
- $\overline{A} = C_E^A = E \setminus A = \{x \notin A\}$
- $A \cap B = \emptyset$: A et B disjoints
- $A \cap \emptyset = \emptyset$; $A \cup \emptyset = A$; $\overline{\overline{A}} = A$
- $A \cap \overline{A} = \emptyset$; $A \cup \overline{A} = E$
- $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$
- $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$
- $A \times B = \{a.b \ / \ a \in A, b \in B\}$
- $\overline{A \cap B} = \overline{A} \cup \overline{B}$; $\overline{A \cup B} = \overline{A} \cap \overline{B}$
- $Card(\overline{A}) = Card(E) - Card(A)$
- $Card(A \cup B) = Card(A) + Card(B) - Card(A \cap B)$
- $Card(A \setminus B) = Card(A) - Card(A \cap B)$
- $Card \emptyset = 0$

2- Analyse combinatoire :

⊜ Méthodes permettant de donner le nb de tous les résultats possibles d'une expérience donnée.

! On doit poser 2 questions : * L'ordonnance * La répétition

	Sans répétition	Avec répétition
Ordonnées = Arrangement	$A_n^p = \frac{n!}{(n-p)!}$	n^p
non ordonnées = Combinaison	$C_n^p = \frac{n!}{p!\,(n-p)!}$	$K_n^p = \frac{(n-1+p)!}{p!\,(n-1)!}$

* Si on a un seul choix : l'ordre et la répétit° = non importants

+ Vocabulaire :

* Expérience aléatoire : action dont le résultat ne peut être prévu avec certitude.

↳ Résultats possibles = Éventualités = évenements élémentaires ($\sharp$ éventualités = univers Ω)

↳ Événement = ensemble d'éventualités noté A

Ω = évenement certain ; $\emptyset$ évenement impo

3- Calcul des probabilités :

. Probabilité d'un évenement : la chance que cet événement se réalise. ! $0 \leq P(A) \leq 1$

! 2 éléments incompatibles = disjoints $P(A \cap B) = 0$

= pas d'intersection ≠ indépendants (il n'y a pas d'influence)

. Probabilités conditionnelles : (indépendance)

la probabilité d'un évenement est influencée par le fait qu'un événement lié au premier, se soit produit.

Probabilité de B sachant A : $P(B/A) = \frac{P(A \cap B)}{P(A)}$

Avec : $P(A) > 0$

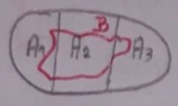
↳ Si A n'a pas d'influence = indépendance

$P(A \cap B) = P(A) * P(B)$

. Théorème des probabilités totales:

Soient $A_1, A_2, A_3, \dots, A_n$ des événements tq:

$$\begin{cases} A_1 \cup A_2 \cup \dots A_n = \Omega \\ A_i \cap A_j = \emptyset \end{cases}$$



$$P(B) = \sum_{i=1}^{n} P(B/A_i) \cdot P(A_i)$$

$B \cap \Omega = B \Rightarrow B \cap (A_1 \cup A_2 \cup A_3) = B$

$(B \cap A_1) \cup (B \cap A_2) \cup (B \cap A_3) = B$

$P(B) = P(B \cap A_1) + P(B \cap A_2) + P(B \cap A_3)$

$P(B) = P(B/A_1) \cdot P(A_1) + \dots$

. Théorème de Bayes:

$$P(A_i/B) = \frac{P(A_i)\, P(B/A_i)}{\sum_{i=1}^{n} P(A_i)\, P(B/A_i)}$$

IV- Variables aléatoires et lois de probabilité:

+ Notion de variable aléatoire: une fonction qui dépend d'une expérience aléatoire, et qui associe à chaque événement une valeur x_i $(X : \Omega \to \mathbb{R})$

↳ discrète = ensemble de ses valeurs est fini ou infini dénombrable.

↳ Continue = ensemble de ses valeurs est $\mathbb{R}$ ou un intervalle de $\mathbb{R}$.

+ V.A.D et lois de probabilités discrètes:

. Loi de probabilité de X: une fonction qui décrit comment sont distribuées les probabilités en fonction des valeurs de la v.a.d.

Exemple: Jet de dé ; $\Omega = \{1,2,3,4,5,6\}$

X = " Gain de jeu ? "

$\begin{cases} 1,2,3 \to 10 \text{ dh} \\ 4,5 \to -5 \text{ dh} \\ 6 \to 0 \text{ dh} \end{cases}$

x_i	-5	0	10
P_i	2/6	1/6	3/6

↳ $P(X = x_i)$

. Fonction de répartition: $F : \mathbb{R} \to$ [résul

$x \to F(x) = P($

+ Fonction d

Ex:

$$F(x) \begin{cases} 0 \text{ si } x < -5 \\ 0 + \frac{2}{6} \text{ si } -5 \leqslant x < 0 \\ 0 + \frac{2}{6} + \frac{1}{6} \text{ si } 0 \leqslant x < 10 \\ 0 + \frac{2}{6} + \frac{1}{6} + \frac{3}{6} \text{ si } x \geqslant 10 \end{cases}$$

Rq:

$$P(X > a) = 1 - P(X \leqslant a)$$

. Espérance: ≃ la moyenne

$$E(X) = \mu = \sum_{i=1}^{n} P_i x_i$$

. Variance:

$$V(X) = \sigma^2 = \sum_{i=1}^{n} P_i (x_i - \mu)^2 = \sum P_i x_i^2 - (E(x))^2$$

. Lois usuelles:

+ Loi uniforme: Les valeurs prises par v.a ont la même probabilité = équiprobabilité

$$\forall i, \quad P(X = x_i) = \frac{1}{n}$$

n: le nombre des valeurs

+ loi de Bernoulli:

Variable de Bernoulli: 2 réponses $\begin{cases} \text{Succès} = 1 \\ \text{Échec} = 0 \end{cases}$

x_i	0	1
P_i	1-p	p

$E(X) = \sum P_i x_i = p$

$\sigma(X)^2 = p(1-p)$

+ loi binomiale:

Variable binomiale: le nb de succès obtenus lors de la répétition de n épreuves identiques et indépendantes (loi de Bernoulli repété plusieurs fois).

$$S_n = X_1 + X_2 + \dots + X_n$$

X_i = variable de Bernoulli

Résultats de x succès pendant n tirages

$$C_n^x = \frac{n!}{x!(n-x)!}$$

+ Fonction de proba.

$$f(x) = C_n^x p^x (1-p)^{n-x}$$

+ Espérance $E(X) = np$

+ Variance $V(X) = np(1-p)$

* **Loi de poisson:** (utiliser ds les tx d'arrivée ds une file d'attente)

Variable de poisson: décrire le nb d'occurences d'un événement au cours d'un intervalle de temps ou d'espace bien défini.

+ La probabilité d'ocurrence est la même dans 2 intervalles de m longueur

+ Fonction de proba.:

$$f(x) = \frac{\mu^x * e^{-\mu}}{x!}$$

μ : nb moyen d'arrivé ds une file pendant un intervalle : Ex

+ Variance: $V^2(X) = \mu$

+ **V.A.C et lois de probabilités continues:**

V.A.C est définie par sa fonction de répartition.

Fonction densité de proba. f : $f = F'$

ou encore: $F(x) = \int_{-\infty}^{x} f(y)\,dy$

Donc: l'airee située sous le graphique de $f(x)$ dans un intervalle particulier, donne la proba. que la v.a.c X prenne une valeur ds cet intervalle.

. La proba. que la v.a.c prenne une valeur particulière est nulle.

Espérance: $E(X) = \int_{-\infty}^{+\infty} x f(x)\,dx$

Variance: $V(X) = \int_{-\infty}^{+\infty} x^2 f(x)\,dx - (E(X))^2$

Ecart type: $\sigma(X) = \sqrt{V(X)}$

! L'aire totale de la courbe est égale à $1 / \int_{-\infty}^{+\infty} f(x)\,dx = 1$

+ **loi normale centrée réduite:**

Si $\forall x \in \mathbb{R}$, $f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$

; $\mu = 0$, $\sigma = 1$ $\Rightarrow$ $N(0,1)$

+ **Approximation normale des proba. binomiales:**

Loi binomiale, dans le cas où:

$np \geq 5$ et $n(1-p) \geq 5$

→ La loi normale permet d'estimer facilement des proba. binomiales, en posant:

$\mu = np$ et $\sigma = \sqrt{np(1-p)}$

+ **loi exponentielle:**

Si $f(x) = \frac{1}{\mu} e^{-\frac{x}{\mu}}$; $x > 0$; $\mu = 0$

Ch 2: Introduction à la théorie de l'échantillonnage

(=) Théorie de sondage : l'optimisation de la collecte de données.

- Sondage = l'observation partielle d'une partie d'une population statistique.
- Méthode d'échantillonage : procédure par laquelle on choisit dans une population un sous-ensemble représentatif.

↳ Obj : Avoir un échantillon suffisamment représentatif pour que les données puissent être extrapolées à la populat°.

1- Échantillonnages aléatoires :

Lorsque chaque individu a une proba. connue et non nulle d'être inclus dans l'échantillon.

* Échantillon aléatoire simple : de taille n issu d'une population finie de taille N.

Est un échantillon sélectionné de manière à ce que chaque échantillon possible de taille n ait la même probabilité d'être sélectionné ($p = \frac{n}{N}$)

+ Avantage : . Conception simple
. Base de la théorie statistique.

- Incon : . Difficile sur des grands échantillons.
. Prb de la disponibilité de la base de sondage.

Ex : tirage au sort

* EAS issu d'une population infinie.

Conditions :
- Chaque élément sélectionné provient de la même populat°
- Chaque élément est sélectionné de façon indépendante

Ex : contrôle douanier

. Base de sondage : liste de toutes les à enquêtes.

* Échantillonnage aléatoire stratifié :

Il consiste à diviser la populat° en sous-groupes appelés Strates. tq chaque élément apparaît dans une et une seule strate.

↳ Sélection d'un échantillon aléatoire simple dans chaque strate (+ homogénéité des strates)

Comment faire la répart° de sujets ds chaque strate ?

* Allocat° égale = même nb de sujets dans chaque strate.

* Allocat° proportionnelle = à la taille de strate

Ex : Étudiants 1ère année / 2ème / 3ème ...

. Populat° 3000 { 2000 jeunes, 1000 vieux } → Éch : 200 { 200 × 66,6% = 133, 200 × 33,3% = 67 }

* Échantillonnage par grappes :

Partager par rapport à un critère Hétérogène.

Il consiste à diviser la populat° en sous-grp = grappes, tq chaque élément de la pop. appartient à une seule grappe. Puis on sélectionne de manière aléatoire des grappes.

Ex : division de Maroc en 12 régions / hétérogénité de la pop
. Choix au hasard de 2 cartonnes parmi 10 pour vérifier la qté des prod.

Cas idéal : chaque grappe est une version à petite échelle de la populat° entière.

+ Les éléments de chaque grappe sont hétérogènes.

* Pas de l'échantillon = $\frac{\text{Populat°}}{\text{taille d'échantillon}} = K$

↳ Avancer par un pas ↓

Échantillonnage systématique :

que la populat' est très importante.

2- Échantillonnages non aléatoires :

⊜ par choix raisonné.

. Utilisés dans les études qualitatives.

* Échantillonnage de commodité :

les éléments sont inclus dans l'échantillon sans proba. connue ou préspécifiée d'être choisis.

Ex: les volontaires qui participent à une expérience à l'école.

↳ coût très faible où nn nul.

* Échantillonnage subjectif :

Sélectionner des éléments de la populat' qui semblent représenter la populat'

Ex: Un journaliste qui peut choisir 3 perso (Opinion générale).

Avantage : . Échantillon facilement construit

. Données facilement collectés

Inc : . Mauvaise représentativité de l'échantillon

. Aucune procédure stat. ne permet de faire une analyse proba. ou de l'inférence stat.

* E. à participat volontaire :

Faire appel à des volontaires pour constituer l'E.

* Enquête boule de neige :

On choisit des individus qui sont pertinents pour l'étude, ensuite on leurs demande de proposer d'autres individus.

* Enquête par Quota :

basée sur la répartit' connue de la pop. pour un certain nombre de caractère (sexe, âge, ...)

* Échantillonnage aléatoire simple :

. Estimation ponctuelle : (la propri' de l'échantillon)

$\bar{x}, s, \bar{p}$ sont des estimateurs ponctuelles de μ, σ, p.

. Distribut' d'échantillonnage : (nb d'échantillon possible = C_N^n)

Soit :

. Expérience aléatoire : Sélection d'un E.A.S

. V.A : Valeur de la moyenne d'échantillon $\bar{x}$.

$\bar{x}$ a une espérance, une variance, un loi de proba = distribut' d'échantillonnage de $\bar{x}$

La connaissance de cette distribut' nous permet de tirer des conclusions en termes de proba., Écart-type de l'échantillon, moyenne de la pop..

Espérance mathématique de $\bar{x}$: $E(\bar{x}) = \mu$

⇒ La moyenne de tous les échantillons = (moy. de tous les moy.) Moyenne de la populat'

Écart-type : (Erreur type de la moyenne)

* Populat' finie = $\sigma_{\bar{x}} = \sqrt{\frac{N-n}{N-1}} \left(\frac{\sigma}{\sqrt{n}}\right)$

* Populat' infinie = $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$

! Lorsque la taille de l'échantillon est inf. à 5% de la populat. ($\frac{n}{N} < 0{,}05$)

* Théorème central limite :

Lorsque la taille de l'échantillon devient importante ($+30$) la distribution d'échantillonnage de $\bar{x}$ peut être approchée par une distribut de proba. normale.

. Distribution d'échantillonnage de $\bar{p}$

$$\bar{p} = \frac{x}{n}$$

le nombre d'éléments dans l'échantillon possédant la caractéristique à laquelle on s'intéresse.

Espérance: $E(\bar{p}) = p$

Écart type:

* Population finie $\sigma_{\bar{p}} = \sqrt{\frac{N-n}{N-1}} \sqrt{\frac{p(1-p)}{n}}$
* Populat infinie $\sigma_{\bar{p}} = \sqrt{\frac{p(1-p)}{n}}$

• Aussi dans ce cas la loi binomiale peut être approchée par une loi normale si:

$np \geq 5$ et $n(1-p) \geq 5$

$\mathcal{B}(n, p) \rightarrow$ LNCR

Conclusion:

• ↗ la taille de l'échantillon →

↘ l'erreur type ↘ → Probabilité ↗

- moyenne
- proportionalité

* Estimation par intervalle:

* De la moyenne d'une populat:

• Lorsque σ est connu, l'estimat par intervalle est donné par

$$\left[\bar{x} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \; ; \; \bar{x} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}\right]$$

α: coeff. de risque

$1-\alpha$: indice de confiance

$z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$ = marge d'erreur

↑ intervalle de confiance

⇒ ↘ indice de confiance → ↘ intervalle

→ moyenne bonne

• Lorsque σ est inconnu

$$\left[\bar{x} - t_{\alpha/2} \frac{s}{\sqrt{n}} \; , \; \bar{x} + t_{\alpha/2} \frac{s}{\sqrt{n}}\right]$$

T de student (loi) = $n-1$ degré de liberté.

* De la proportion d'une popula

$$\left[\bar{p} \pm z_{\alpha/2} \sqrt{\frac{\bar{p}(1-\bar{p})}{n}}\right]$$

* Détermind de la taille d'échantillon:

$$n = \frac{(z_{\alpha/2})^2 \sigma^2}{E^2}$$

$$n = (z_{\alpha/2})^2 \frac{p^*(1-p^*)}{E^2}$$

p^*: proportion d'une ancienne étude ou p^*: 0,5

$z^{\alpha/2}$: la variable centrée réduite qui correspond à la valeur qui permet de négliger $\alpha/2$ des informations dans la queue droite sup.

I: Régression linéaire Simple

Une droite qui peut s'approcher de plus à l'ensemble des points (x et y) ; afin de savoir s'il existe une relation linéaire entre les 2 variables.

+ Modèle de régression linéaire simple :

$$y = \beta_0 + \beta_1 x + \varepsilon$$ ← représente la partie non linéaire de la relat°

y: variable à explique (dépendante)

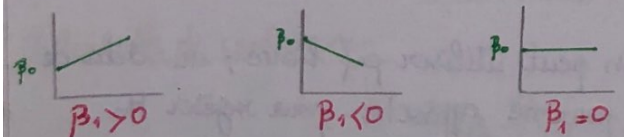
x: variable explicative (indépendante)

ε: terme d'erreur; la différence entre les valeurs réelles et les valeurs obtenues si la relation était exacte.

* Erreur de spécification: la seule variable explicative n'est pas suffisante pour expliquer la totalité du phénomène.
* Erreur de mesure: les données ne représentent pas exactement le phénomène.
* Erreur de fluctuation d'échantillonage: d'un échantillon à l'autre les observations et donc les estimations sont différentes.

+ Équation de régression linéaire simple :

moyenne de y ← $E(y) = \beta_0 + \beta_1 x$

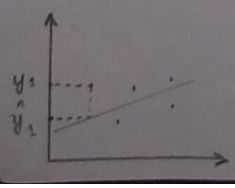


⇒ Relat° linéaire ⊕ R. linéaire ⊖ Pas de relat°

↳ En pratique, β_0 et β_1 ne sont pas connus donc on doit les estimées en utilisant les données d'un échantillon ; d'où La droite de régression estimée:

$$\hat{y} = b_0 + b_1 x$$

↓ droite de l'échantillon



* La méthode des moindres carrés :

Permet d'utiliser les données de l'échantillon pour estimer l'équat° de la régression (b_0, b_1)

Elle consiste à minimiser l'écart au carré entre y_i et $\hat{y}_i$.

$$\text{Min} \sum (y_i - \hat{y}_i)^2$$

$$b_0 = \bar{y} - b_1 \bar{x}$$

$$b_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

* Coefficient de détermination :

+ Somme des carrés des résidus :

$$SC_{res} = \sum (y_i - \hat{y}_i)^2$$

⇒ La valeur de SC_{res} est une mesure de l'erreur commise en utilisant l'équation estimée de la régression pour estimer les valeurs de la variables dépendante dans l'échantillon.

+ Somme des carrés totale :

$$SC_T = \sum (y_i - \bar{y})^2$$

↳ Pour estimer y_i sans avoir les x_i, on utilise $\bar{y}$

⇒ SCT est une mesure de l'ajustement, autour des observations de la droite $\bar{y}$.

+ Somme des carrés de la régression :

$$SC_{reg} = \sum (\hat{y}_i - \bar{y})^2$$

⇒ $$SC_T = SC_{res} + SC_{reg}$$

$$r^2 = \frac{SC_{reg}}{SC_T}$$ ← intensité ⇒ Coef. de détermination

$SC_T = SC_{res} + SC_{reg}$ → cas idéal: $y_i - \hat{y} = 0$

$\frac{SC_T}{SC_T} = \frac{SC_{res}}{SC_T} + \frac{SC_{reg}}{SC_T}$ ⇒ $1 = SC_{res}/SC_T = r^2$

$$r_{xy} = \text{Signe de } b_1 \sqrt{r^2}$$ ⇒ coef. de corrélation

↓ intensité

* Différence entre coef. de corrélation et coef. de détermination:

1- Dans le cas d'une relation linéaire entre deux variables le coef. de détermination et le coef. de corrélation fournissent une mesure de robustesse de la relation.

2- Le coef de détermination est compris entre 0 et 1, tandis que le coef. de corrélation est compris entre -1 et 1.

3- Alors que le coef. de corrélation est restreint à des relations linéaires entre 2 variables, le coef de détermination peut être utilisé dans le cas de relation non linéaire et de relations comprenant plus de 2 variables indépendants.

⇒ Même si le coefficient de détermination est assez élevé, une analyse approfondie de la robustesse du modèle supposé doit être faite.

D'où: Le test d'hypothèses

1- Établir le test d'hypothèse adéquat:

* test d'hypothèse unilatéral supérieur $\begin{cases} H_0 : \mu \leqslant \mu_0 \\ H_a : \mu > \mu_0 \end{cases}$

* Test unilatéral inférieur $\begin{cases} H_0 : \mu \geqslant \mu_0 \\ H_a : \mu < \mu_0 \end{cases}$

* Test bilatéral $\begin{cases} H_0 : \mu = \mu_0 \\ H_a : \mu \neq \mu_0 \end{cases}$

H_0 = Hypothèse nulle

H_a = Hypothèse alternative = ce que je dois prouver

	H_0 vraie	H_a vraie
Accepter H_0	Conclusion correcte	Erreur de 2nde espèce
Rejeter H_0	Erreur de 1ère espèce	Conclusion correcte

2* Seuil de signification: La probabilité de faire une erreur de 1ère espèce ≡ Seuil d'acceptation

3* test de signification:

+ Si σ connu: $z = \dfrac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}$

	test unilatéral inf	test unilatéral Sup	test bilatéral
Règle de rejet H_0	$z \leqslant -z_\alpha$	$z \geqslant z_\alpha$	$z \leqslant -z_{\alpha/2}$; $z \geqslant z_{\alpha/2}$

+ Si σ est inconnu: $z = \dfrac{\bar{x} - \mu_0}{s / \sqrt{n}}$

	test unilatéral inf.	test unilatéral Sup.	test bilatéral
Rejet de H_0	$t \leqslant -t_\alpha$	$t \geqslant t_\alpha$	$t \leqslant -t_{\alpha/2}$; $t \geqslant t_{\alpha/2}$

⇒ Marge de liberté $(n-1)$

z_α, t_α = valeur critique

N.B: On peut utiliser p (l'aire ; au déla de z) comme approche pour rejeter H_0

* Règle de rejet: $p \leqslant \alpha$

$z = \dfrac{\bar{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$: test d'hypothèse relatif à la proportion.

de signification (sur la régression linéaire simple):

+ Test de Student :

* Le test d'hypothèse adéquat :

$$\begin{cases} H_0 : & \beta_1 = 0 \\ H_1 : & \beta_1 \neq 0 \end{cases}$$

$E(y) = \beta_0 + \beta_1 x$ (ça concerne la popula°)

→ Les propriétés de b_1 : pop → échantillon

+ Espérance = $E(b_1) = \beta_1$

+ Ecart-type = $\sigma_{b_1} = \dfrac{\sigma}{\sqrt{\Sigma(x_i - \bar{x})^2}}$

→ Distribut° normale :

+ Ecart-type estimé de b_1 =

Erreur de b_1 ← $s_{b_1} = \dfrac{s}{\sqrt{\Sigma(x_i - \bar{x})^2}}$

tel que :

$$s^2 = MC_{res} = \frac{SC_{res}}{n-2}$$

$$\Rightarrow s = \sqrt{s^2}$$

test : $t = \dfrac{b_1}{s_{b_1}}$

* Règle de rejet : $t \leq -t_{\alpha/2}$ ou $t \geq t_{\alpha/2}$

! ddl = $n-2$

Intervalle de confiance :

$$b_1 \pm t_{\alpha/2}\, s_{b_1}$$

+ Test F de Fisher :

$$s^2 = MC_{reg} = SC_{reg}/1$$

Statistique de test : $F = \dfrac{MC_{reg}}{MC_{res}}$

* Règle de rejet : (de H_0)

$$F \geq F_\alpha$$

F_α est basé sur la distribut° de Fisher à 1 ddl au numérateur et $(n-2)$ ddl au dénominateur.

* Tableau ANOVA :

Source de la variat°	Somme des variat°	ddl	Moyenne des variat°	F
Régression	$SC_{rég}$	1	SC_{reg}	$\dfrac{MC_{reg}}{MC_{res}}$
Résidu	$SC_{rés}$	$n-2$	$\dfrac{SC_{res}}{n-2}$	
Totale	SC_T	$n-1$		