Statistique Descriptive

Mesure de tendance centrale

	Population	Echantillon
Moyenne	$\mu = \frac{\sum x_i}{N}$	$\bar{x} = \frac{\sum x_i}{n}$
Médiane	$\frac{N}{2}$ → Nbre décimal : Mé = nbre entier suivant ; Nbre entier : $Mé = \frac{(N/2)^e + (N/2+1)^e}{2}$	idem
Mode	c'est l'observation la plus fréquente	
Percentile	1er quartile = 25e percentile (N/4 x 1) 2ème quartile = 50e percentile (Mé) 3ème quartile = 75e percentile (N/4 x 3) 85e percentile = $\frac{N}{100} \times 85$ Rst des Vl constitue x% de l'étude	

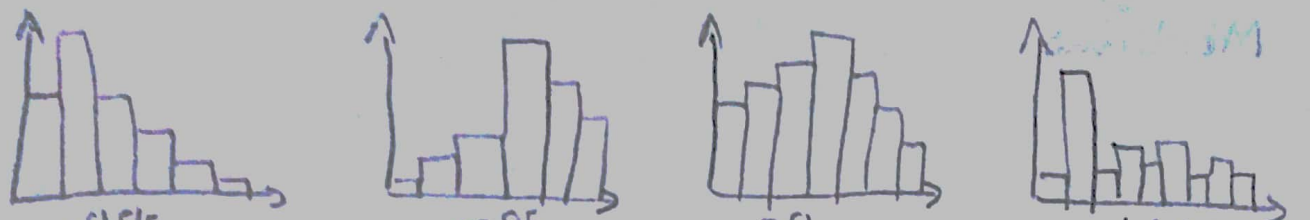
Mesure de dispersion

Etendue	La ≠ entre la 1ère et la dernière valeur Etendue interquartile (EIQ) ⇒ entre 3e et 1er quartile calcul la force de liaison entre les Vls	
Variance	$\sigma^2 = \frac{\sum (x_i - \mu)^2}{N}$	$s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1}$
Ecart type	$\sigma = \sqrt{\sigma^2}$	$s = \sqrt{s^2}$
	c la Vl idéale qui répond à un critère donné	
coefficient de variation	$\frac{\text{Ecart type}}{\text{moyenne}} \times 100$ x CV → 0 ⇒ série homogène x CV → 1 ⇒ série hétérogène x bonne homogénité = CV < 15%	

Mesure de tendance centrale et detection des valeurs singulières:

Forme d'une distribution: A partir de la forme de la distribution on conclut que les données biaisées sont soit à droite soit à gauche.

D.Asy > 0 D.Asy < 0



Modérement asymétrique — symétrique — Fortement asymétrique à droite

Degré d'asymétrie = $\gamma_1 = \frac{\mu_3}{\delta^3}$

avec $\mu_3 = \sum_{i=1}^{n} \frac{1}{n} (x_i - \bar{x})^3$

Variable centrée réduite:

$$z_i = \frac{x_i - \bar{x}}{s} \quad \text{ou} \quad z = \frac{x_i - \mu}{\delta}$$

- elle est calculée pour chaque valeur
- elle détermine / mesure la distance en nombre d'écarts type entre x et la moyenne.

Théorème de Chebyshev:

Calculant le nombre de la variable x qui se situe entre [Y et V] de ni :

Y = moyenne − (z × écart type)

V = moyenne + (z × écart type)

⇒ Au moins $(1 - \frac{1}{z^2})$ % des observations se situent dans l'intervalle [Y, V]

Pour tous types de lois

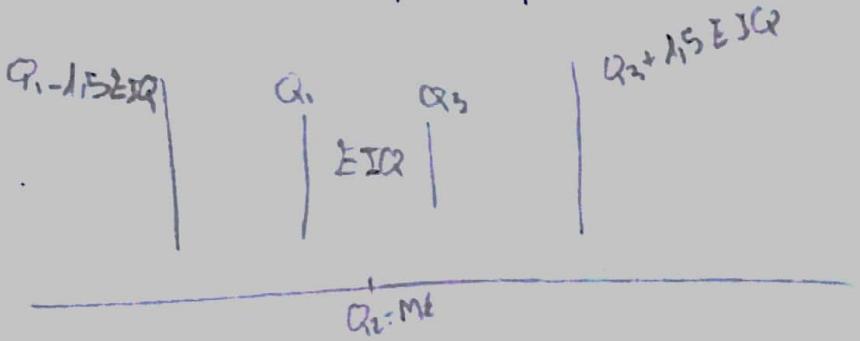
Règle empirique:

Environ 68% des observations se situent dans $[\bar{x}-s ; \bar{x}+s]$
Environ 95% " " " " $[\bar{x}-2s ; \bar{x}+2s]$
Environ toutes les observations se situent dans $[\bar{x}-3s ; \bar{x}+3s]$
Pour la loi normale

Boite à pâtes: Pour la detection des valeurs singulières
Vl inhabituelle — Er. d'enregistrement — observation pas correctement incluse

Résumé en 5 chiffres: ⊖ petite valeur, ⊕ grande valeur, 1er/2 ème/3 ème quartile

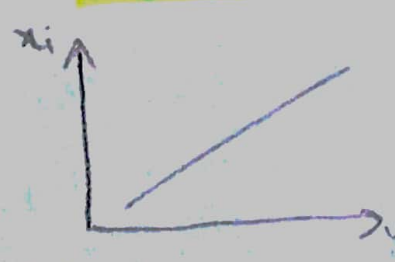
Calcul Etendue interquartile. 3ème Q − 1er Q
Calcul: Vl sup = $Q_3 + 1{,}5\ EIQ$
Vl inf = $Q_1 - 1{,}5\ EIQ$



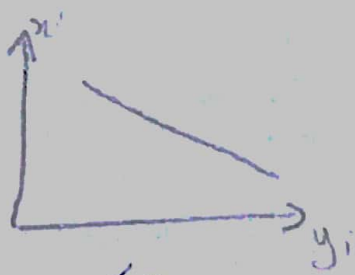
Statistique Bivariée

	Population	Echantillon
Covariance	$\sigma_{xy} = \frac{\sum (x_i-\mu_x)(y_i-\mu_y)}{N}$	$s_{xy} = \frac{\sum (x_i-\bar{x})(y_i-\bar{y})}{n-1}$
Coefficient de corrélation	$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$	$r_{xy} = \frac{s_{xy}}{s_x s_y}$

La covariance est la mesure de la relation linéaire entre 2 variables



Cov > 0
⇒ Vl positives = Relat° linéaire ⊕



Cov < 0
⇒ Vl négatives = Relat° linéaire ⊖

Coefficient de corrélation : pas d'unité

forte rel. lin ⊖ $\Leftarrow -1 \leq r_{xy} \leq 1 \Rightarrow$ forte relation linéaire ⊕

↳ 0 => très faible rel linéaire ou ≠ introuvable

Méthode des moindres carrés :

Droite de regression de Y en X, $D_{Y/X}$: $y = ax + b$

où $a = \frac{\sigma_{xy}}{(\sigma_x)^2}$ et $b = \bar{y} - a\bar{x}$

Droite de régression de X en Y, $D_{X/Y}$: $x = a'y + b'$

où $a' = \frac{\sigma_{xy}}{(\sigma_y)^2}$ et $b' = \bar{x} - a'\bar{y}$

L'échantillonnage

Sondage = une étude pour un échantillon
Recensement : une étude pour une population
Estimateur ponctuel = le fait de calculer un indicateur pour un échantillon afin de l'inférer à la population.

Méthodes d'échantillonnage :

1) Echantillonnages aléatoires : chaque individu a une probabilité connue et non nulle d'apparaître dans l'échantillon

* Echantillonnage aléatoire simple (EAS)

Population finie :
chaque échantillon à la même proba. d'être choisi parmi la population.

Population infinie :
- chaque élément est sélectionné de façon indépendante
- chaque élément sélectionné provient de la même population.

* Echantillonnage aléatoire stratifié :

- c'est la division de la pop en sous groupes / strates / tranches.
- chaque élément apparait dans une seule strate
  Ex : jeunes - vieux / hommes - femmes
- On choisit un EAS dans chaque strate
- Il y a une homogénéité des sous groupes

Allocation égale — Ex : Pop ⇒ jeunes : 1000, vieux : 1000 — Ech ⇒ 50% jeunes, 50% vieux

Allocation proportionnelle à la taille — Ex : Pop ⇒ jeunes : 2000, vieux : 1000 — Ech ⇒ 2/3 jeunes, 1/3 vieux

* Echantillonnage par grappes :

C'est une miniature de la réalité de la population entière, le meilleur résultat est d'avoir des éléments hétérogènes.

Ex : quartiers, secteurs ...

On choisit un EAS dans chaque grappe.

* Echantillonnage systématique :

Au lieu de l'EAS on utilisera l'ech. systématique

Ex : Pop. 5000 Ech : 50

1- Calcul d'un pas de l'échantillon : $K = N/n = 100$

2- On choisit un nombre entre 0 et K et on rajoute le pas (100) : ex : 17
17, 107, 217, 317 4917
Jusqu'à l'obtention de 50 ech.

2) Echantillonages non aléatoires:

quand il y a un avis / choix raisonné / sur étude.

* **Echantillonage de commodité:**
  On choisit un échantillon sans étude / volontaire : ex : prof ...
* **Echantillonage subjectif:**
  On cible le répondant.
  Ex : journaliste choisit des personnes ...
* **Boule de neige:** on prend quelques individus, on leur demande de remplir le questionnaire et le passer à leurs amis.
* **Echantillonnage par quota:**
  Ech. respecte les proportions connues pour les catégories de la pop
  Ex : 200 Fac 50 ENCG 40 ENSA

• Vaut mieux utiliser l'échantillonage aléatoire

Echantillonage aléatoire simple (EAS)

Estimation ponctuelle:

* **Distribution d'échantillonnage de $\bar{x}$:**

a - l'espérance $E(\bar{x}) = \mu$ ($\bar{x}$: la moyenne de l'éch ; μ : moyenne de la pop)

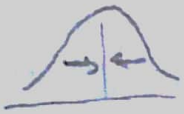
b - Ecart type

$\begin{cases} \alpha\ \frac{n}{N} \leqslant 0{,}05 + \text{pop finie} \\ \alpha\ \text{pop infinie} \end{cases} \Rightarrow \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$

α pop finie $+ \frac{n}{N} \geqslant 0{,}05 \Rightarrow \sigma_{\bar{x}} = \sqrt{\frac{N-n}{N-1}} \times \frac{\sigma}{\sqrt{n}}$

$\sigma_{\bar{x}}$ ↓ Erreur type de la moyenne

c - La distribution tend vers une loi normale $\mathcal{N}(\mu, \sigma)$



Pour ↗+ la proba, il faut :
↗+ Echantillonnage
↘- Ecart type

Pour calculer la proba d'atteindre γ écarts on reste entre ...

$\mu - \gamma \leqslant \bar{x} \leqslant \mu + \gamma$

↳ on la converge la loi normale centrée

$\ldots \leqslant \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} \leqslant \ldots$

$-v \leqslant \bar{z} \leqslant v$

on voit la table de la loi normale

Ex : $P(Z \leq -N) = P(Z \geq N) = 1 - P(Z < N)$

$P(Z \geq 2,3) = 1 - P(Z \leq 2,3)$

$P(1,2 \leq Z \leq 2,3) = P(Z \leq 2,3) - P(Z \leq 1,2)$

$P(-1 \leq Z \leq 1) = 2 \times P(Z \leq 1)$

* Distribution d'échantillonage de $\bar{p}$: $\bar{p} = \frac{x}{n}$

Il faut voir l'élément s : elle suit la loi binomiale

a- Espérance : $E(\bar{p}) = p$ (proportion éch. → proportion pop.)

b- Ecart type :

× $\frac{n}{N} \leq 0,05$ + pop. finie
× pop infinie
$\Rightarrow \sigma_{\bar{p}} = \sqrt{\frac{p(1-p)}{n}}$

× pop. finie + $\frac{n}{N} \geq 0,05 \Rightarrow \sigma_{\bar{p}} = \sqrt{\frac{N-n}{N-1}} \sqrt{\frac{p(1-p)}{n}}$

↓

Erreur type / Ecart type proportion

c- La distribution tend vers la loi binomiale $\mathcal{B}(n ; p)$ (n : Nbre d'essais ; p → succès)

qui peut se converger en loi normale

si ~~$n \geq 30$~~ ; $np \geq 5$ et ~~$nq \geq 5$~~ et $n(1-p) \geq 5$

$\mathcal{B}(n;p) \simeq \mathcal{N} \rightarrow (np ; \sqrt{npq})$

Pour calculer la propa

$p - y \leq \bar{p} \leq p + y$

↓ on la converge vers loi N.C.R

$\ldots \leq \frac{\bar{p} - p}{\sigma_{\bar{p}}} \leq \ldots$

$u \leq Z \leq v$

↳ on voit table de la loi normale

Estimation par intervalle :

Estimation de la moyenne d'une population :

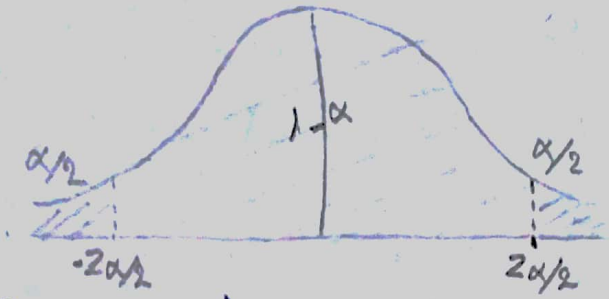
$\mu \in \left[\bar{x} - z_{\alpha/2} \times \frac{\sigma}{\sqrt{n}} ; \bar{x} + z_{\alpha/2} \times \frac{\sigma}{\sqrt{n}}\right]$ Intervalle de confiance

$\mu \in [\bar{x} - E ; \bar{x} + E]$ ↳ Marge d'erreur

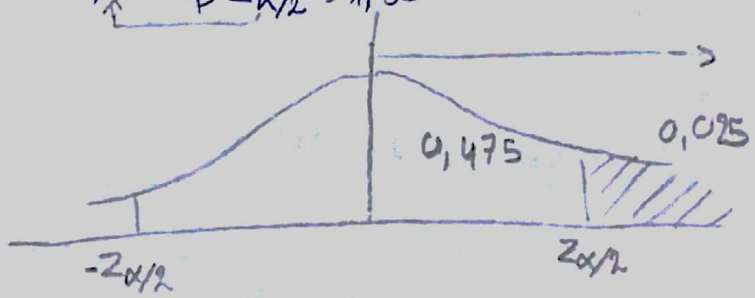
Si on a l'écart type de la population
⇒ Loi normale centrée réduite

Si on n'a pas l'écart type de la population
⇒ Loi de student

α → indice de l'incertitude - de risque
$1-\alpha$ → indice de confiance
$\frac{\alpha}{2}$ → un demi-indice risque
$z_{\alpha/2}$ → la valeur à partir de laquelle on est dans la zone d'incertitude.

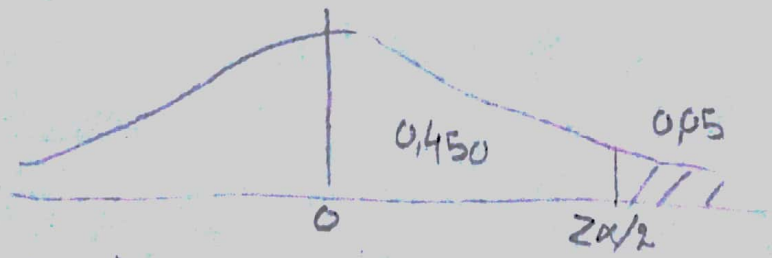


Selon L.N.C.R : on parle de $z_{\alpha/2}$
VC risque 0,05 ⇒ $z_{\alpha/2} = 1,96$



Donc on cherche les coordonnées de P = 0,475 ⇒ 1,96

VC risque 0,1 ⇒ $z_{0,05}$



$z_{\alpha/2}$ est entre 1,64 et 1,65
on fait la moyenne $z_{\alpha/2}$ = 1,645

Selon L.S. on parle de $t_{\alpha/2}$

On cherche degré de liberté

Nech. $-1 = \alpha$

On cherche la proba de α degré de liberté

$T_{\alpha/2}$

et $(\underline{0,10/0,05/0,025/0,01/0,005})$?

$\Downarrow$

$\alpha/2$

Erreur $= t_{\alpha/2} \times \frac{S}{\sqrt{n}}$

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

* Pour $z_{\alpha/2}$ l'écart type est celui de la pop.
* Pour $t_{\alpha/2}$ l'écart type est celui de l'échantillon

Estimation de la proportion d'une pop:

$$P \in \left[\bar{P} - z_{\alpha/2} \sqrt{\frac{\bar{P}(1-\bar{P})}{n}} ; \bar{P} + z_{\alpha/2} \sqrt{\frac{\bar{P}(1-\bar{P})}{n}} \right)$$

on n'utilise pas la loi de student

Erreur $= z_{\alpha/2} \times \frac{\sqrt{\bar{P}(1-\bar{P})}}{\sqrt{n}}$

Estimation de la taille de l'échantillon:

Moyenne:

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

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

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

Proportion:

$$E = z_{\alpha/2} \times \frac{\sqrt{\bar{P}(1-\bar{P})}}{\sqrt{n}}$$

$$\sqrt{n} = \frac{z_{\alpha/2} \times \sqrt{\bar{P}(1-\bar{P})}}{E}$$

$$n = \frac{(z_{\alpha/2})^2 \times \bar{P}(1-\bar{P})}{E^2}$$

$\bar{P}(1-\bar{P}) \Rightarrow p^2 = 0,5$

ou bien la valeur de p est basée sur des études anciennes

Test d'hypothèse pour Regression

$H_0 \rightarrow \beta_1 = 0$
$H_1 \rightarrow \beta_1 \neq 0$

on veut que notre modèle soit significatif
→ Il y a une relat° entre x et y

Donc on cherche à rejeter H_0

Test student: Statistique du test : $t = \frac{b_1}{Sb_1}$

$b_1 \rightarrow \hat{y} = b_1 x + b_0$
ou bien $E(b_1) = \beta_1$ Espérance

$Sb_1 \rightarrow Sb_1 = \frac{S}{\sqrt{\sum (x_i - \bar{x})^2}}$ Ecart type estimé de b_1

avec S = Ecart type de l'estimation

$S = \sqrt{MC_{res}} = \sqrt{\frac{SC_{res}}{n-2}}$ Erreur type de l'estimation

Rejet H_0

App. par Vl critique : $t \leq -t_{\alpha/2}$ ou $t \geq t_{\alpha/2}$ ← vc ddl (n-2)

App. par Vl p : $p \leq \alpha$

Intervalle de confiance pour β_1 :

$\beta_1 \in [b_1 \pm t_{\alpha/2} \times Sb_1]$

Test F de Fisher: Statistique du test = $\frac{MC_{reg}}{MC_{res}} = F$

avec : $MC_{reg} = \frac{SC_{reg}}{1}$ ← 1 le nbre ddl qui est : nbre de Vl indépend.

$MC_{res} = \frac{SC_{res}}{N-2}$

Rejet H_0 :

App par Vl p : $p < \alpha$
App par vl critique : $F > f_\alpha$ → voir table fisher

	Source de la variation	Somme des Carrés : SC	ddl	Moyenne des Carrés	F
Fisher	Régression	SC_{reg}	1	MC_{reg}	$F = \frac{MC_{reg}}{MC_{res}}$
Student	Résidu	SC_{res}	n-2	MC_{res}	
	Total	SCT	n-1	-	

Précision et estimation: ← on utilise student avec n-2 ddl

de la Vl moyenne de y $E(y)$

Ecart type estimé de $\hat{y}_p$:

$$S_{\hat{y}_p} = S\sqrt{\frac{1}{n} + \frac{(x_p - \bar{x})^2}{\sum(x_i - \bar{x})^2}}$$

Intervalle de confiance de $E(y_p)$

$$E(y) \in [\hat{y}_p \pm t_{\alpha/2} \times S_{\hat{y}_p}]$$

de la Vl individuelle de y:

Ecart type estimé de y_{ind}

$$S_{ind} = S\sqrt{1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{\sum(x_i - \bar{x})^2}}$$

Intervalle de prévision de y:

$$y \in [\hat{y}_p + t_{\alpha/2} \times S_{ind}]$$

Régression linéaire simple

$$Y = \beta_1 x + \beta_0 + \varepsilon \quad \leftarrow \text{erreur}$$

Droite linéaire

Equation estimée de la regression

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

Méthode des moindres carrés:

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

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

Somme des carrés résidus:

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

Somme des carrés total:

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

$$SCT = SCres + SCreg$$

Somme des carrés de la regression:

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

Coefficient de détermination:

$$0 \leqslant r^2 = \frac{SCreg}{SCT} \leqslant 1$$

Plus on s'approche de 1 → ⊖ erreurs

0 → ⊕ erreurs

→ X% est expliqué par la droite $\hat{y} = b_1 x + b_0$ (Rlt linéaire)

→ s'utilise ds le cas des rlt linéaires et non linéaires.

→ teste et évalue l'adéquation de l'équation estimée de la regression.

Coefficient de corrélation:

$$-1 \leqslant r = (\text{signe } b_1)\sqrt{r^2} \leqslant 1$$

→ teste et évalue l'intensité de la relat° linéaire

→ utilisé juste ds le cas des relations linéaires.

i | x_i | y_i | $x_i - \bar{x}$ | $y_i - \bar{y}$ | $(x_i - \bar{x})(y_i - \bar{y})$ | $(x_i - \bar{x})^2$ | $(y_i - \bar{y})^2$

$\hat{y} = b_1 x + b_0$ | $y_i - \hat{y}$ | $(y_i - \hat{y})^2$ | $\hat{y}_i - \bar{y}$ | $(\hat{y}_i - \bar{y})^2$

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

⚠ Même si on obtient un Coeff de détermination assez élevé il faut faire un test d'hypothèse de signification pour analyser le modèle supposé (y).

On suppose que :
$E(\varepsilon) = 0$ donc $E(y) = \beta_1 x + \beta_0$
$\sigma^2(\varepsilon)$ est le même pour tous x
Les ε sont indépendants
ε suit la loi normale

on voit si $\beta_1 = 0$ (Y et x sont indépendants) H_0 ou $\beta_1 \neq 0$ (y et x sont dépendants) H_1

Test de signification :

$E(y) = \beta_0 + \beta_1 x$

$H_0 : \beta_1 = 0$
$H_1 : \beta_1 \neq 0$

Si $\beta_1 = 0 \rightarrow E(y) = \beta_0 \rightarrow$ x et y ne sont pas liés
Sinon : les 2 variables le sont (si $\beta_1 \neq 0$)

Estimation de σ^2 :

MCres fournit une estimation de σ^2 :
σ inconnu

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

$S = \sqrt{MC_{res}}$

Test d'hypothèse de la moyenne

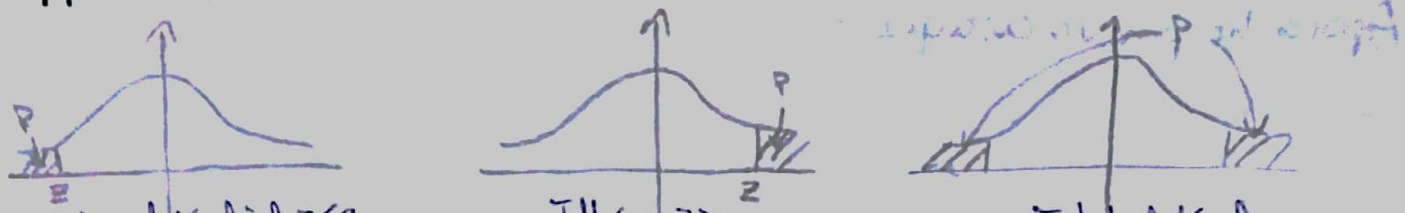
Si σ est connu: on utilise la loi normale C.R

statistique du test: $Z = \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$

$H_0: \mu \geq \mu_0$	$H_0: \mu \leq \mu_0$	$H_0 = \mu = \mu_0$
$H_1: \mu < \mu_0$	$H_1: \mu > \mu_0$	$H_1: \mu \neq \mu_0$
Test unilatéral inf.	Test unitaleral sup	Test bilateral

Pour rejeter H_0:

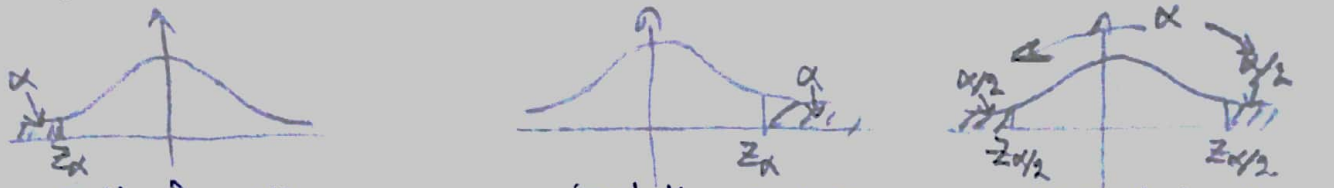
Approche par Vlt p: Rejet de H_0 si $p \leq \alpha$ (p=seuildesignificat?observé)



Test unilatéral inf $z<0$ — T.U sup $z>0$ — Test bilatéral

$P = 0,5 - (\text{Proba } Z) \Rightarrow \alpha = \text{Risque}$

Approche par Vlt critique z: Rejet de H_0 si $z \leq z_\alpha$



T.U inf $z \leq -z_\alpha$ — Test U. sup $z \geq z_\alpha$ — Test bilatéral $z \leq -z_{\alpha/2}$ ou $z \geq z_{\alpha/2}$

z = statistique du test
z_α = z du risque complet
$z_{\alpha/2}$ = z de 1/2 du risque

Si σ est inconnu: on utilise loi de student degré de liberté (n-1)

$H_0: \mu \geq \mu_0$	$H_0: \mu \leq \mu_0$	$H_0: \mu = \mu_0$
$H_1: \mu < \mu_0$	$H_1: \mu > \mu_0$	$H_1: \mu \neq \mu_0$
Test unilatéral sup	Test unilatéral sup	Test bilatéral

Pour rejeter H_0:

Approche par Vlt p: $p \leq \alpha$

Approche pal Vlt critique:

$t \leq -t_\alpha$ $t \geq t_\alpha$ $t \leq -t_{\alpha/2}$ ou $t \geq t_{\alpha/2}$

t: statistique du test
t_α = statistique t du risque complet
$t_{\alpha/2}$ = statistique t de 1/2 du risque

Test d'hypothèse de la proportion

Statistique du test $Z = \dfrac{\bar{P} - P_0}{\sqrt{\dfrac{P_0(1-P_0)}{n}}}$

$H_0: P \geq P_0$	$H_0: P \leq P_0$	$H_0: P = P_0$
$H_1: P < P_0$	$H_1: P > P_0$	$H_1: P \neq P_0$
Test unilatéral inf.	Test unilatéral sup	Test bilatéral

Pour rejeter H_0:

Approche par Vb p:

$P = 0,5 - (\text{Proba } z)$

$p < \alpha$

Approche par Vb critique:

$Z \leq -Z_\alpha$	$Z \geq Z_\alpha$	$Z \leq -Z_{\alpha/2}$ ou $Z \geq Z_{\alpha/2}$