Statistique appliquée # Révision S5

CH1 rappel | 1

La statistique est la science qui étudie et analyse les Données

Statistique descriptive	Statistique probabiliste
décrire les données déjà existées	décrire les prévisions

rappel :

population	individu	échantillon	caractère
Σp, ΣO... Base de l'étude stati	élément de la population	sous-ensemble de la pop	propriété de la pop

- effectif
- fréquence
- effectif cumulé ↗
- effectives cumulé ↘

} indicateurs de forme

caractère → qualitatif (direct : ponctuel, ne change pas ; continue : change lorsqu'on a une évaluation) / quantitatif (nominale, ordinal)

« Statistique d'échantillon » ; « Statistique de pop » ou plutôt paramètres de la population

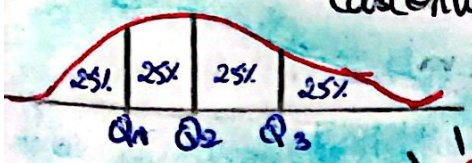
	échantillon	pop
Moyenne	$\bar{X}$	μ
Variance	S^2	σ^2
Ecart type	s	σ
Co-Variance	S_{xy}	σ_{xy}
corrélation	r_{xy}	σ_{xy}

Mesures de tendances centrales et de dispersion

échantillon		population	
$\bar{X} = \frac{\sum x_i n_i}{N}$	$s = \sqrt{s^2}$	$\mu = \frac{\sum x_i n_i}{N}$	$\sigma = \sqrt{\sigma^2}$
Biaisé $S^2 = \frac{\sum (x_i - \bar{X})^2}{n - 1}$ (le nombre → n) corrigée		$\sigma^2 = \frac{\sum (x_i - \mu)^2}{N}$	

Scanned with CamScanner

- Percentile : p pour cent des observations ont une valeur
- quartile
  - cas discret $Q_1 = \frac{1}{4}N$ • $Q_2 = \frac{1}{2}N$ • $Q_3 = \frac{3}{4}N$
  - cas continue $Q_1 = \frac{N}{4} \Rightarrow$ class $[a-b[$



$$Q_1 = a + \frac{(b-a)\left(\frac{N}{4} - n_i\right)}{(n_{i+1} - n_i)}$$

• Remarques :
- Le 50ème percentile est la médiane
- intervalle interquartile = $Q_3 - Q_1 = IQ$ ≈ étendue interquartile

• autres mesures de dispersion

étendue = Valeur maximale − Valeur minimale

coefficient de variation = $\frac{\text{Ecart type}}{\text{moyenne}} \times 100$

↳ combien s'élève l'écart type par rapport au moyen

• Astuces

percentile = $\frac{85}{100} \times N \Rightarrow$ décimale on prend la suivante

• Mode

qui a le plus grand effectif (cas discret)

cas continue

amplitudes égales $[a,b[$ classe modale — amplitudes inégales

$$M_o = a + \frac{(n_i - n_{i-1})}{(n_i - n_{i-1}) + (n_i - n_{i+1})} \times a$$

intervalle	n_i
$[x, z[$	n_{i-1}
$[a, b[$ ← classe modale	n_i
$[b, z[$	n_{i+1}

Valeur singulier ! What ! comment on va les detecter !

• Degré d'asymétrie

mesure la forme d'une distribution de données.

(−) données biaisées à gauche.

(+) données biaisées à droite

$$\gamma_1 = \frac{\mu_3}{\sigma^3}$$

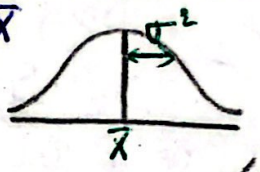
• distribution symétrique $= 0$

$$\mu_3 = \frac{1}{n} \sum n_i (x_i - \bar{X})^3$$

• Variable centrée réduite z_i

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

mesure la distance en nombre d'écart type entre x_i et $\bar{X}$



• théorème de chebychev

$$\left[\underbrace{\bar{X} - zs}_{a} \; , \; \underbrace{\bar{X} + zs}_{b}\right]$$

pour connaître le pourcentage de données dans un intervalle

$$1 - \frac{1}{z^2} \Rightarrow \text{pourcentage}$$

• Règle empirique

1. 68% des observations se situent dans $[\bar{X} - s \, ; \, \bar{X} + s]$
2. 95% des observations se situent dans $[\bar{X} - 2s \, , \, \bar{X} + 2s]$
3. presque toutes les observations se situent dans $[\bar{X} - 3s \, ; \, \bar{X} + 3s]$

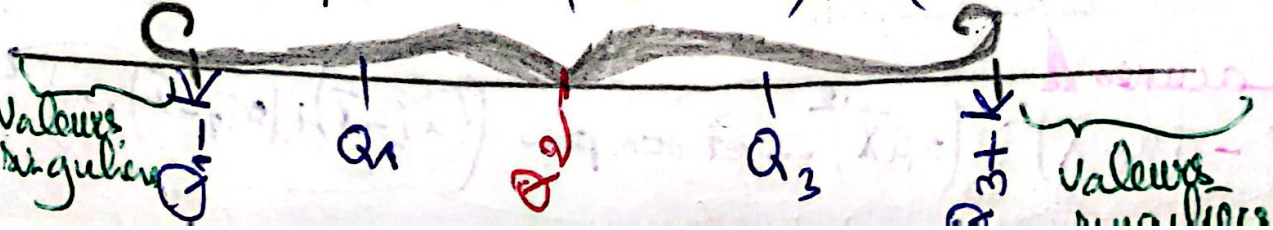
les valeurs ↗ sont des valeurs singulières

• Analyse exploratoire des données

① Résumé en 5 chiffres
- valeur la plus petite
- Q_1 • Q_2 • Q_3
- valeur la plus grande

② construction de la Boîte à Moustaches

étendu interquartils : $|Q_3 - Q_1| K \Rightarrow (EIQ) \times 1{,}5$



Statistiques bivariées mesure la relation entre 2 variable

échantillon — covariance — population

$$S_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n-1} \qquad \sigma_{xy} = \frac{\sum (x_i - \mu_x)(y_i - \mu_y)}{N}$$

on se retrouve dans 2 cas :

- → valeurs positives → relation linéaire positive
- → valeurs négatives → relation linéaire négative

mesure relation linéaire

coefficient de corrélation

$$r = \frac{S_{xy}}{S_x S_y} \qquad \rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \cdot \sigma_y}$$

mesure relation linéaire

on se retrouve dans 3 cas :

- → valeurs proches de +1 → forte relation linéaire positive
- → valeurs proches de −1 → forte relation linéaire négative
- → valeurs proches de 0 → absence de relation linéaire

$$-1 < r < 1$$

méthode des moindres carrés — partie du CH3

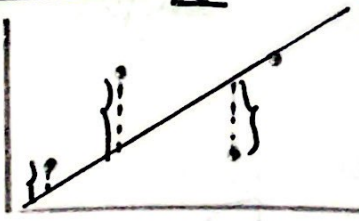
$$\hat{y} = b_0 + b_1 x$$

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

$$b_1 = \frac{\text{covariance}(x,y)}{\text{variance}(x)}$$

$$b_1 = \frac{\frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n-1}}{\frac{\sum (x_i - \bar{x})^2}{n-1}}$$

pourquoi on utilise la méthode des moindres carrés



J'ai ces points là ! et je veux la même droite qui va mieux passer avec ces points ⇒ pour minimiser les écarts !

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

Lacunes ⚠

$(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 \ldots$ et non pas $\left((x_1 - \bar{x}) + (x_2 - \bar{x}) \ldots\right)^2$!

Statistique appliquée #Révision S5

Ch 2 Échantillonage

2

but : avoir un échantillon suffisament representatif de la p[op]

I. échantillonage aléatoire

- échantillon aléatoire simple (EAS) (+) P=P
- échantillon aléatoire stratifié
- échantillonage par grappes
- échantillonage systématique

EAS
- pop finis (+)
- pop infini suivant 2 conditions :
  - E provient de l'm pop
  - E selectionnés de façon indépendante

II. échantillonage N. aléa[toire]

- " de commodit[é]
- " subjectif
- " à participation volontaire
- enquête Boul de neige
- " par Quota

⇒ utilisés dans les études qualit[atives]

Δ pas d'extrapolat[ion] pop

avantages
- échantillonage facilement constitués
- données facilement collectées

inconvénients
- pas d'extrapolation / la pop
- pas d'analyse probabiliste ou de l'inference
- y'a pas un critère standart de choix

I. échantillonage aléatoire : EAS Estimation ponc[tuelle]

Δ

1) Distributions d'échantillonage de $\bar{X}$: (Nombre d'échantillon po[ssible])

- $E(\bar{X}) = \mu$ Δ
- Ecart type
  - pop fini $\sigma_{\bar{x}} = \sqrt{\frac{N-n}{N-1}}\left(\frac{\sigma}{\sqrt{n}}\right)$ ← facteur de correction ; × sinon $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$
  - pop infini $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$ Δ si $\frac{n}{N} < 0{,}05$

Δ la moyenne de toutes les moyennes = la moyenne de tous les échan[tillons]

Δ chaque loi de proba se caractérise par $\mathcal{L}(E(\bar{X}), \sigma(\bar{X}))$
⇒ la variable aléatoire ici est : $\bar{X}$

Δ une pop infini ⇒ on connait pas sa fin

Echantillon aléatoire simple — taille : n de l'E (Récapitulatif)

distribution de $\bar{x}$

pop finie

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

↑ erreur type de la moyenne

pop infinie

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

→ si la taille de EAS ≤ 0,05 de la taille de pop
$n \leq 5\% N$

astuce 1 :

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

Théorème central limite

1 _ Sélectionner EAS
2 _ when $n > 30$, la distribution d'échantillonage de $\bar{x}$ peut être approchée par une distribution de proba. normale.

- $n < 30$ individus — loi uniforme $P = \frac{1}{n}(X = X_i)$
- $n > 30$ individus — loi normale on va le voir

2) Distribution d'échantillonage de $\bar{p}$ → proportion de pop possédant un caractère donné

- $\bar{P} = \frac{x}{n}$ ← nombre d'élément
- $E(\bar{P}) = P$ Khtahes obent 3am

Ecart type :
- ↗ pop finie : $\sigma_{\bar{p}} = \sqrt{\frac{N-n}{N-1}} \cdot \sqrt{\frac{P(1-P)}{n}}$
- ↘ pop infinie : $\sigma_{\bar{p}} = \sqrt{\frac{P(1-P)}{n}}$ ⚠

⚠ exception → si la $n \leq 5\% N$ → pop infini

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

$E(X) = np \geq 5$ $\quad n(1-P) \geq 1$ — question d'examen 2017

l'écart type $S = \sqrt{np(1-p)}$

$$P(X=k) = C_n^k \, p^k (1-p)^{n-k}$$

Astuce 2 :

($n \nearrow$ → l'erreur type ↘ → probabilité ↗ { – moyenne – proportionalité

B. Estimation par intervalle de confiance

de μ ① — de p ②

1. La moyenne de la pop

Lorsque le seuil de confiance augmente, ME ↘ et les intervalles s'élargissent

- $1-\alpha$: indice de confiance
- Lorsque σ est connu, l'estimation par intervalle est donné par :

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

⇒ intervalle de confiance

Loi normale

où :
- α : coeff de risque
- $Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$: marge d'erreur
- longeur : 2x marge d'erreur

Astuce 3

$(1-\alpha)$ ↘ → ↘ intervalle ⇒ bonne moyenne

- Lorsque σ est inconnu, l'estimation par intervalle est donné par :

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

TEST STUDENT

à Retenir

Moyenne μ

- ponctuelle : $\hat{\mu} = \bar{X}$
- intervalle de conf
  - normal : σ connu : $\left[\bar{X} \pm X_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$; $\bar{X} \sim N\left(\mu ; \frac{\sigma}{\sqrt{n}}\right)$; ~~$n \geq 30$ loi normale~~
  - student : σ non connu : $\left[\bar{X} \pm t_{\alpha/2}\frac{S}{\sqrt{n}}\right]$; $n < 30$ loi de student

Vous trouverez plus de détails dans la page précédente

à Retenir

échantillonnage → pop vers → échantillon

estimation → pop vers ← échantillon

Regression linéaire simple

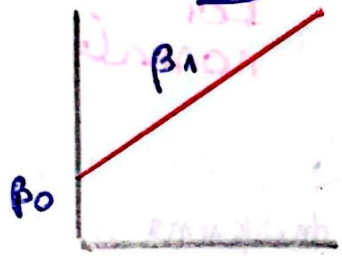
CH3

on est dans la sphère d'étudier une pop par rapport à 2 caractères !

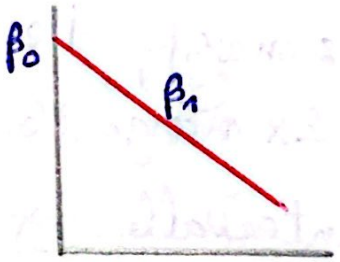
→ on essaye de Trouver une droite qui s'approche le plus possible

Modèle de regression linéaire simple

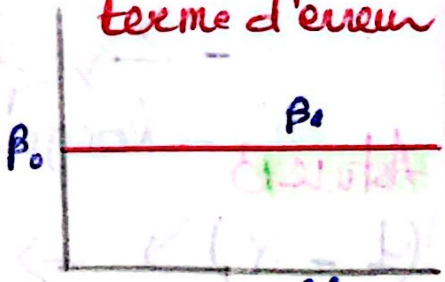
l'équation de : $y = \beta_0 + \beta_1 x + \varepsilon$ ← représente la partie non linéaire dans la relation : terme d'erreur



pente positive
⇒ R. linéaire positive



pente négative
⇒ R. linéaire négative



pente null
⇒ R. linéaire

droit de regression estimé : $\hat{y} = b_0 + b_1 x$ droit d'échantillon

comment calculer le coefficient de détermination ?

$r^2 = \frac{Sc\ rég}{SCT}$ → Somme des carrés de régression

$SCrég = \sum (\hat{y} - \bar{y})^2$ qui permet de juger la qualité du fit

d'après $\hat{y} = b_0 + b_1 X$

Somme des carrés Totales

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

$SCT = Sc\ Rég + Sc\ résiduelle$

Aussi $r^2 = 1 - \frac{Sc\ Résiduelle}{SCT}$

Somme des carrés résiduelle

$Sc\ rés = \sum (y_i - \hat{y})^2$

comment calculer le coefficient de corrélation ?

$r_{xy} = \frac{S_{xy}}{S_x \cdot S_y}$ | $\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \cdot \sigma_x}$ | $r_{xy} = (\text{signe de } b_1) \cdot \sqrt{r^2}$