Econométrie.

Rappel:

1- Définir les lois de probabilité dans un cadre univarié
2- Espérance et variance pour un ensemble de variables aléatoire
3- Cadre multivarié
4- Présentation des lois incontournables
5- Distribution de la moyenne empirique

I- Définir les lois de probabilité dans un cadre univarié

1- probabilité, univers et variable aléatoire

issus: résultat possible d'une expérience aléatoire sachant qu'une seule issue va se réaliser et que toutes les issus n'ont pas la même chance de se réaliser

probabilité d'une issue: fréquence avec laquelle cette issue est réalisée à LT

univers: ensemble des issues possibles

évenement: sous-ensemble de l'univers

variable aléatoire: résumé numérique de l'issue d'une expérience aléatoire

2. Distribution de probabilité d'une variable discrète

Distribution de proba : énumération de toutes les valeurs possible de la variable avec leurs probabilités respective

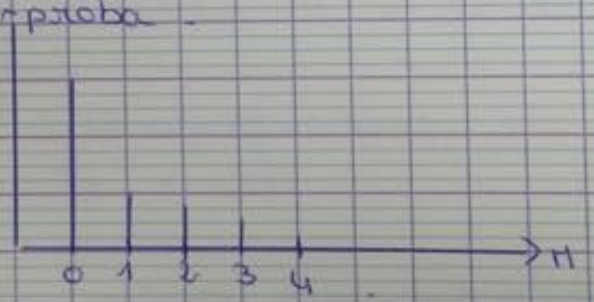
proba des évènements : $P(H=1 \text{ ou } H=2) = p(H=1) + p(H=2)$
$= 0,16$

issues (nombre de panne)

0	1	2	3	4
distribution des probas	0,8	0,1	0,06	0,03	0,01

fonction de répartition : proba que cette v.a soit ≤ à une valeur

0	1	2	3	4
fct de répartition	0,8	0,9	0,96	0,99	1

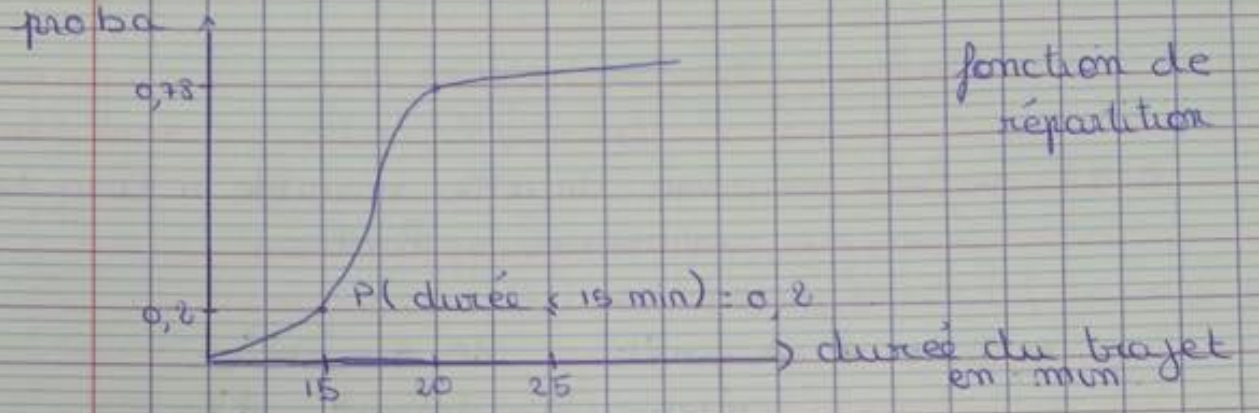


Loi de Bernoulli

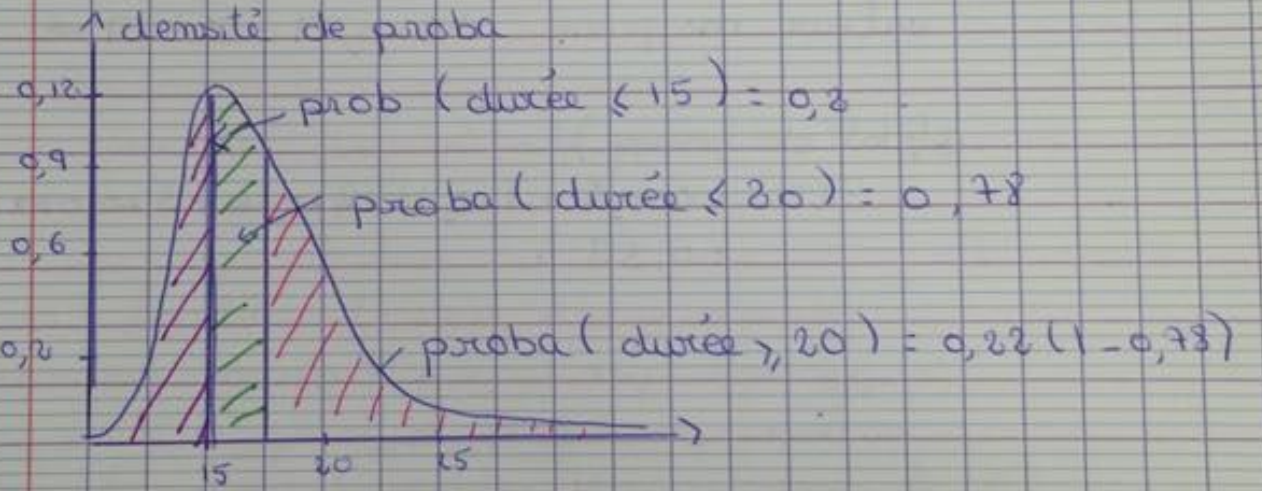
G — G = 0 (femme)
G = 1 (homme)

$$G = \begin{cases} 1 \text{ avec une proba } p \\ 0 \text{ " " " } 1-p \end{cases}$$

3- Distribution de proba d'une v.a continue



fonction de densité de proba: aire de la probabilité que la V.A s'intercale entre ces 2 points



V Valeurs attendues de la moyenne, variance

1 - Valeur attendue d'une v.a

$E(Y)$: valeur moyenne de cette variable à long terme après de nombreuses applications
⇒ espérance de Y

ex: ① 100 € de prêt 10 € intérêt
risque de défaut 1%

montant :
110 € ⇒ 0,99
0 € ⇒ 0,01

valeur moyenne : $110 \times 0,99 = 108,9$ €

② (nbr de panne)
$E(H) = 0 \times 0,8 + 1 \times 0,1 + 2 \times 0,06 + 3 \times 0,03 + 4 \times 0,01$
$= 0,35$.

Supposons que la v.a Y puisse prendre k valeurs possible $y_1, y_2, \dots y_k$
la proba que Y prenne la valeur y_1 est egale p_1
" " " " " y_k p_k

$E(Y) = y_1 p_1 + y_2 p_2 + \dots + y_k p_k$

$= \sum_{i=1}^{k} y_i p_i$ » moyemme de Y
espérance de Y.

ex: Bernoulli
$E(G) = 1 \times p + 0 \times (1-p)$
$= p$.

2. Écart-type et variance

variance : $\text{var}(Y) = E((Y-\mu_Y)^2)$

écart type : $\sigma(Y) = \sqrt{\text{var}(Y)}$

v.a Y avec R valeurs $(y_1 \dots y_R)$

$\sigma_Y^2 = \text{var}(Y) = E((Y-\mu_Y)^2) = \sum_{i=1}^{R} (y_i - \mu_Y)^2 p_i$

ex de la panne : $\text{var}(H) = (0-0,35)^2 \times 0,8 + (1-0,35)^2 \times 0,1 +$
$(2-0,35)^2 \times 0,06 + (3-0,35)^2 \times 0,03 +$
$(4-0,35)^2 \times 0,01$
$= 0,6475$

$\sigma_H = \sqrt{\text{var}(H)} = \sqrt{0,6475} = 0,80$

Bernoulli

$\text{var}(G) = (0-p)^2 \times (1-p) + (1-p)^2 \times p$
$= p(1-p)$

3. fonction linéaire d'une variable

2 v.A X et Y

taux d'imposition des salariés $\Rightarrow$ 20 %

prime annuelle non imposable 2000 €

$Y = 2000 + 0,8X$ X revenu brut mensuel

va μ_X σ_X^2 Y = " net

$\Rightarrow$ Y va

$E(Y) = \mu_Y = 2000 + 0,8\mu_X$

$Y - \mu_Y = 0,8(X - \mu_X)$

$E((Y-\mu_Y)^2) = E[(0,8(X-\mu_X))^2]$
$= 0,64 \times E[(X-\mu_X)^2]$

$= 0,64\ \sigma_x^2$ $\quad\quad$ $\sigma_y = 0,8\ \sigma_x$

$y = 2000 + 0,8X$

$\mu_y = 2000 + 0,8\mu_x$

$y = a + bX$

X revenus bruts individuel

v.a $\quad \mu_x \quad \sigma_x^2$

$\mu_y = a + b\mu_x$

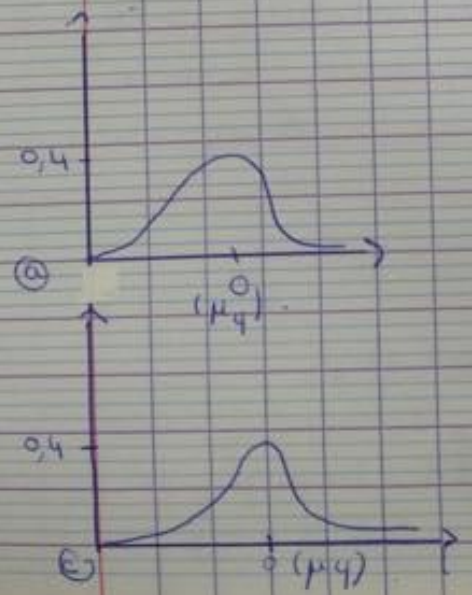
$\sigma_y^2 = b^2 \sigma_x^2$ $\quad\quad$ $\sigma_y = b\sigma_x$

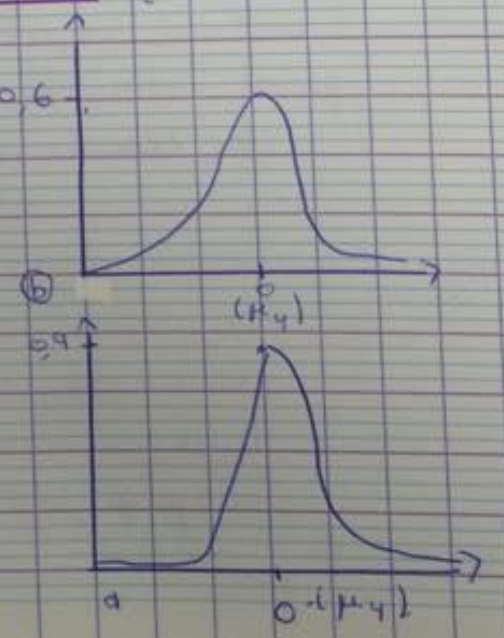
4- Autres mesures de la forme d'une distribution

coefficient de dissymétrie: (skewness) mesure l'asymétrie de la distribution

coefficient d'aplatissement: (kurtosis) qui mesure son aplatissement

⇒ basés sur les moments de distribution





coefficient de dissymétrie : $\frac{E((Y-\mu_Y)^3)}{\sigma_Y^3}$

pour une forme symétrique, le coefficient = 0

= 0 => symétrique

≠ 0 => asymétrique

a) skewness = 0 b) skewness = 0

c) skewness = -0,1 d) skewness = 0,6

coefficient d'aplatissement : $\frac{E((Y-\mu_Y)^4)}{\sigma_Y^4}$
kurtosis

=> mesure de la masse qui se trouve au niveau des queues de la distribution

(jamais < 0)

. la normale = 3

. il est leptokurtique > 3

a) kurtosis 3 b) kurtosis = 20

c) kurtosis = 5 d) kurtosis = 5

Moment : le moment d'ordre x de Y est donné par $E(Y^x)$

Cas de 2 v.a

1- Distribution jointe et marginale

x et y

C'est la proba que ces 2 variables prennent simultanément 2 valeurs données

ex : x et y

distribution des proba jointes : $P(X=x; Y=y)$

Y v.a 1 durée courte (< 20 min)
0 sinon

X v.a 1 il pleut
0 sinon

	X=1 (pluie)	X=0 (ϕ pluie)	
long Y=0	0,15	0,07	0,22
court Y=1	0,15	0,63	0,78
	0,30	0,70	1

distribution jointe : $P(Y=0; X=1) = 0,15$

distribution de proba marginale : distribution de Y prise de manière isolée

Si X peut prendre l'valeur $\neq x_1, \dots x_I$ alors la proba que Y prenne une valeur y est de $P(Y=y) = \sum_{i=1}^{I} p(X=x_i; Y=y_i)$

2 - Distribution conditionnelle

La probabilité conditionnelle de Y sachant x
⇒ la proba que Y prenne la valeur y sachant que X prend la valeur x

$$P(Y=y \mid X=x)$$

La distribution conditionnelle de Y sachant $X=x$

$$P(Y=y \mid X=x) = \frac{P(X=x, Y=y)}{P(X=x)}$$

$$P(\underset{y=0}{\text{trajet long}} \mid \underset{x=1}{\text{il pleut}}) = \frac{0,15}{0,30} = 0,5$$

3 - Espérance conditionnelle de Y sachant X

↳ moyenne de la distribution conditionnelle de Y sachant x

Y $y_1 \dots y_k$

$$E(Y \mid X=x) = \sum_{i=1}^{k} y_i P(Y=y_i \mid X=x)$$

La des espérances itérées

↳ moyenne Y = moyenne pondérée de l'espérance conditionnelle de Y sachant X, pondération ← distribution de X

taille moyenne des adultes

moyenne pondérée de la moyenne des hommes et des femmes pondéré par la proportion d'homme et de femme

pour p valeurs de X $x_1 \dots x_p$

$$E(Y) = \sum_{i=1}^{p} E(Y \mid X=x_i) P(X=x_i) = E(E(Y \mid X))$$

Variance conditionnelle :

↳ $\text{var}(Y \mid X=x) = \sum_{i=1}^{n} [y_i - E(Y \mid X=x)]^2 \, P(Y=y_i \mid X=x)$

4- Indépendances

X et Y sont indépendemment distribués ou indépendantes si connaitre la valeur de l'une ne donne aucune information sur la valeur de l'autre :

$$P(Y=y \mid X=x) = P(Y=y)$$

X et Y indépendantes $\Rightarrow P(X=x, Y=y) = P(X=x) \times P(Y=y)$

5- Covariance et corrélation

Covariance : indique dans quelle mesure 2 variables varient ensemble

$$\text{cov}(X,Y) \text{ ou } \sigma_{x,y} \quad E[(X-\mu_x)(Y-\mu_y)]$$

$$\text{cov}(X,Y) = \sigma_{x,y} = \sum_{i=1}^{n} \sum_{j=1}^{p} (x_j - \mu_x)(y_i - \mu_y) P(X=x_j, Y=y_i)$$

Corrélation $\text{corr}(X,Y) = \dfrac{\text{cov}(X,Y)}{\sqrt{\text{var}(X)\,\text{var}(Y)}} = \dfrac{\text{cov}(X,Y)}{\sigma_x \sigma_y}$

X et Y sont dites corrélées si $\text{cov}(X,Y)$

$$-1 \leq \text{cov}(X,Y) \leq 1$$

Si la moyenne conditionnelle de Y ne dépend pas de X

si $E(Y|X) = E(Y) = \mu_Y$

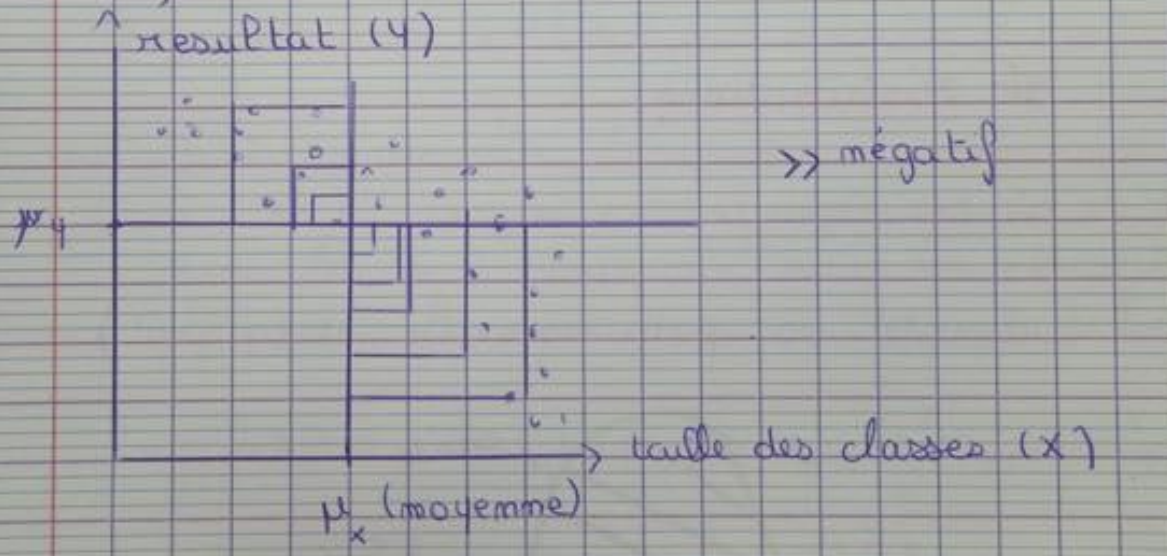
alors $cov(x, y) = 0$.

$cov(X, Y) = 0$

. X et Y deux V.A de moyenne nulle

$$cov(X,Y) = E[(X-\mu_x)(Y-\mu_Y)] = E(XY)$$
$$= E(E(XY/X))$$
$$= E(X\,E(Y/X))$$
$$= E(X\,E(Y))$$
$$= 0$$

donc si $cov(X,Y) = 0 \Rightarrow cov(X,Y) = 0$.



6- Moyenne, variance et somme des v.a

$E(X+Y) = E(X) + E(Y) = \mu_X + \mu_Y$
$var(X+Y) = var(X) + var(Y) + 2\,cov(X,Y)$

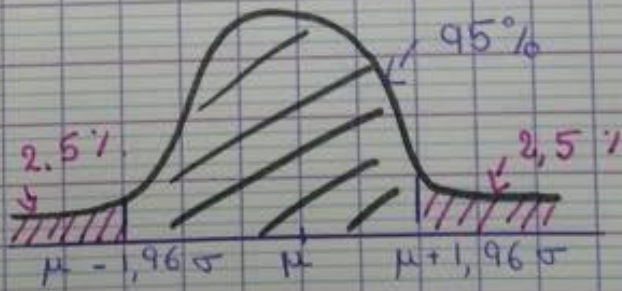
$E(a+bX+cY) = a + b\mu_X + c\mu_Y$
$var(a+bY) = b^2 \sigma_Y^2$

$var(aX+bY) = a^2\sigma_X^2 + b^2\sigma_Y^2 + 2ab\,cov(X,Y)$
$E(Y^2) = \sigma_Y^2 + \mu_Y^2$
$cov(a+bX+cV, Y) = b\sigma_{X,Y} + c\,cov$
$E(XY) = \sigma_{X,Y} + \mu_X \mu_Y$

IV Distributions

1. Loi Normale

- densité de proba normale $\mathcal{N}(\mu, \sigma^2)$



- la normale centrée réduite $\mathcal{N}(0,1)$ Z :
fonction de répartition Φ
$P(Z \leq c) = \Phi(c)$ c : valeur

$Y \rightsquigarrow \mathcal{N}(1,4)$

$P(Y<2) = P\left(\frac{Y-1}{2} < \frac{1}{2}\right) = \Phi(0,5)$

$= 0,691$

$Z = \frac{Y-1}{2} \rightsquigarrow \mathcal{N}(0,1)$

$Y \rightsquigarrow \mathcal{N}(\mu, \sigma^2)$

alors $Z = \frac{Y-\mu}{\sigma} \rightsquigarrow \mathcal{N}(0,1)$

c_1 et c_2 $\quad c_1 < c_2$ $\quad \begin{cases} d_1 = \frac{c_1-\mu}{\sigma} \\ \end{cases}$

$P(Y \leq c_2) = P(Z \leq d_2) = \Phi(d_2)$

$P(Y \geq c_1) = P(Z \geq d_1) = \Phi(d_1)$

$P(c_1 \leq Y \leq c_2) = P(d_1 \leq Z \leq d_2) = \Phi(d_2) - \Phi(d_1)$

distribution normale bivariée et multivariée

① Si X et Y conjointement distribué (X et Y sont le tous les 2) suivant une loi normale bivariée de covariance $\sigma_{X,Y}$ a, b 2 constantes

$aX+bY \rightsquigarrow \mathcal{N}(a\mu_X + b\mu_Y, a^2\sigma_X^2 + b^2\sigma_Y^2 + 2ab\sigma_{X,Y})$

② Si un ensemble de variables admettant une distribution normale multivariée. La distribution marginale de chacune de ces variables est également normale

③ Si des variables avec une distribution normale multivariée ont des covariances nulles alors elles sont indépendantes

W
W

⑤ Si X et Y admettent une distribution normale bivariée, l'espérance conditionnelle de Y sachant X est linéaire en X

$$E(Y|X) = a + bX$$

2. loi du Khi-deux

La distribution du Khi-deux c'est la distribution de la somme des carrés de m v.a normalement et indépendamment distribués de moyenne nulle et de variance unitaire

Si $Z_1 \rightsquigarrow \mathcal{N}(0,1)$
$Z_2 \rightsquigarrow \mathcal{N}(0,1)$ alors $Z_1^2 + Z_2^2 + Z_3^2 \rightsquigarrow \chi^2_3$
$Z_3 \rightsquigarrow \mathcal{N}(0,1)$

3. loi de student

La distribution de Student à m d.l est définie par la distribution du ratio de 2 v.a indépendantes

$$\frac{\text{une v.a} \rightsquigarrow \mathcal{N}(0,1)}{\text{une v.a} \rightsquigarrow \chi^2_m}$$

$Z \rightsquigarrow \mathcal{N}(0,1)$ alors $\frac{Z}{\sqrt{\frac{W}{m}}} \rightsquigarrow$ Student t
$W \rightsquigarrow \chi^2_m$

Pour $m > 30$, cela se rapproche de la loi normale

4 - loi de Fisher (distribution F)

La distribution F à m et n ddl notée $F_{m,n}$ est définie par le ratio suivant :

$$\frac{\chi^2_m \,|\, m}{\chi^2_n \,|\, n}$$

si U et V 2 v.a indépendemment distribués $\rightsquigarrow \chi^2$ à m et n ddl

alors $\frac{U|m}{V|n} \rightsquigarrow F_{m,n}$

Cas particulier

Si n est très grand alors $F_{m,n}$ peut être approché par $F_{m,+\infty}$ alors V/n correspond à la moyenne d'une infinité de χ^2 ; cette moyenne vaut 1. donc $F_{m,+\infty}$ est la distribution d'une v.a Khi-deux à n ddl, divisé par m.

☆ échantillonnage aléatoire et distribution d'échantillonnage des moyennes

1- Echantillonnage aléatoire

n objets aléatoirement sélectionnés à partir d'une pop et dont chaque membre à la même proba d'être selectionné

les n observations de l'echantillons sont notés $Y_1, \dots, Y_n$.

la sélection de manière aléatoire de membre de la pop conduit à des v.a $Y_1 \to Y_n$

la v.a Y correspondant au i ème objet aléatoirement tiré noté Y.

- distribution commune à tous les Y_i
- équiprobabilité du tirage

$\hookrightarrow$ $Y_1 \dots Y_n$ sont indépendamment et identiquement distribués (i.i.d).

2- Moyenne et distribution d'échantillonnage

moyenne d'échantillonnage (empirique) $\bar{Y}$ d'un ensemble d'observation de $Y_1, \dots, Y_n$

$$\bar{Y} = \frac{1}{n}(Y_1 + \dots + Y_n) = \sum_{i=1}^{n} \frac{Y_i}{n}$$

Distribution d'échantillonnage de $\bar{Y}$

$Y_1, Y_2, \dots Y_n$ sont iid Y_i μ_Y σ_Y^2

$n = 2$

$Y_1 + Y_2$

$E(Y_1 + Y_2) = \mu_Y \times 2$

$E\left(\frac{Y_1 + Y_2}{2}\right) = \mu_Y$

$\text{var}(Y_1 + Y_2) = 2\sigma_Y^2$

$\text{var}\left(\frac{Y_1 + Y_2}{2}\right) = \frac{1}{2}\sigma_Y^2$

$\text{var}(\bar{Y}) = \text{var}\left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right)$

$= \frac{1}{n^2}\Sigma \text{var}(Y_i) + \frac{1}{n^2}\sum^{n}\sum^{n} \text{cov}(Y_i, Y_j)$

$\text{var}(\bar{Y}) = \frac{1}{n}\sigma_Y^2$

$\Rightarrow$ écart-type de $\bar{Y}$ $\frac{1}{\sqrt{n}}\sigma_Y$

Quand Y est normalement distribué

$Y_1, \dots Y_n$ sont iid de loi $\mathcal{N}(\mu_Y, \sigma_Y^2)$

$\bar{Y}$ suit $\mathcal{N}\left(\mu_Y, \frac{\sigma_Y^2}{n}\right)$

VI approximation symptotique de distribution d'échantillonage

approche exacte = distribution { exacte / à distance finie

$\bar{Y} \xrightarrow{P} \mu_Y$

$\frac{\bar{Y} - \mu_Y}{\sigma_{\bar{Y}}} \rightsquigarrow \mathcal{N}(0;1)$.

La des grands nombres si Y_i $i = 1, \dots, n$ des v.a i.i.d
avec $\mathbb{E}(Y_i) = \mu_Y$ et qu'il n'y a pas de valeurs extrêmes $\text{var}(Y_i) = \sigma_i^2 < +\infty$
alors $\bar{Y} \xrightarrow{P} \mu_Y$

TCL
$Y_1 \dots Y_n$ sont iid de moyenne $\mathbb{E}(Y_i) = \mu_Y$ et de variance $\text{var}(Y_i) = \sigma^2_Y$ $0 < \sigma^2_Y < \infty$
alors quand $n \to +\infty$, la distribution de $\frac{\bar{Y} - \mu_Y}{\sigma_{\bar{Y}}}$ est arbitrairement bien approchée par une loi normale $\mathcal{N}(0;1)$.

Rappel Stat.

I Estimation

1) Estimateur et propriétés

Critères de choix :

a) absence de biais

plusieurs tirages aléatoires et répétitifs d'échantillons

si à l'issu des répétitions, la moyenne de la distribution d'échantillonnage de l'estimateur est μ_y alors c'est sans biais

$\hat{\mu}_y$ est dit sans biais si $E(\hat{\mu}_y) = \mu_y$

b) Convergence

quand $n \to +\infty$, la proba que $\hat{\mu}_y$ soit très proche de μ_y tend vers 1

$\hat{\mu}_y \xrightarrow{p} \mu_y$

c) efficacité

Si 2 estimateurs $\hat{\mu}_y$ et $\tilde{\mu}_y$ sans biais

$\hat{\mu}_y$ sera + efficace que $\tilde{\mu}_y$ si

$var(\hat{\mu}_y) < var(\tilde{\mu}_y)$

sans biais car $E(\bar{y}) = \mu_y$ convergent loi des grands nombres $\bar{y} \xrightarrow{p} \mu_y$

efficace ? $var(\bar{y}) = \frac{\sigma_y^2}{n}$ pour $n \gg 1$

La moyenne empirique est le meilleur estimateur linéaire sans biais.

$\bar{y}$ = moyenne empirique

* estimateur des moindres carrés

critère à minimiser $\sum_{i=1}^{n} (Y_i - m)^2$

($Y_i - m$: écart de prédiction)

$$\frac{\partial}{\partial m} \sum (Y_i - m)^2 = 0$$

$$\Leftrightarrow -2 \sum (Y_i - m) = 0$$

$$\Leftrightarrow \sum_{n} Y_i - m = 0$$

$$\Leftrightarrow \sum Y_i - n \times m = 0$$

$$\Leftrightarrow m = \frac{1}{n} \sum Y_i = \bar{Y}$$

II Tests d'hypothèse

1) hyp nulle ou alternature

H_0 ou H_1

H_0 " la moyenne théorique de Y, $E(Y)$, prend une valeur spécifique $\mu_{Y,0}$ "

$$E(Y) = \mu_{Y,0}$$

H_1 : $E(Y) \neq \mu_{Y,0}$

2) La proba marginale de rejet

si $\bar{Y} \neq \mu_{Y,0}$ 2 raisons

→ la vraie moyenne n'est pas $\mu_{Y,0}$.
H_0 est fausse

→ la vraie moyenne est bien $\mu_{Y,0}$ (H_0 est vraie)
mais aspect aléatoire dans
l'échantillonnage

La probabilité marginale de rejet
proba de tirer une valeur au moins aussi éloignée de la vraie valeur que la valeur estimée à partir de l'échantillon quand on suppose que H_0 est vraie

ex: Diplomés
salaire moyen = 22,64 € observés

H_0: ~~salaire~~ la moyenne du salaire est de 20 €.

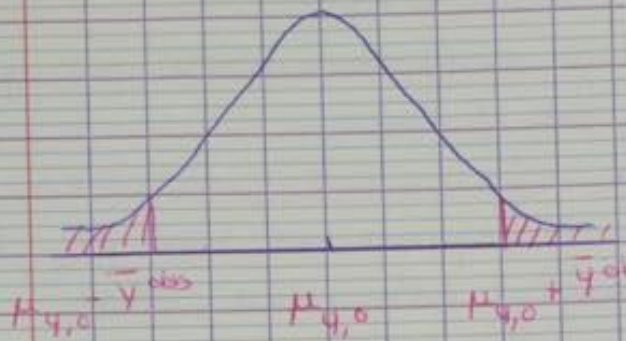
$$\mu_W = \underbrace{\mu_{W,0}}_{20\ €} \qquad E(W) = 20\ €.$$

Avec quelle proba, observera-t'on une valeur du salaire moyen $\geq$ 22,64
ou $\bar{W} \leq 17,36$ €.
sachant que $\mu_W = 20$ €

$\bar{y}^{obs}$ = valeur de la moyenne empirique effectivement calculé

P_{H0} : proba calculée sous H_0 (ie $E(Y_i) = \mu_{Y,0}$)

valeur $p = P_{H_0}(|\bar{Y} - \mu_{Y,0}| > |\bar{y}^{obs} - \mu_{Y,0}|)$



$\bar{y}^{obs}$ pas compatible avec H_0 si p est très faible

⇒ nécessaire de connaître la distribution d'échantillonnage de $\bar{Y}$ sous H_0.
TCL de la loi normale est une bonne approximation de la distribution de $\bar{Y}$

Sous H_0, $\bar{Y}$ suit $\mathcal{N}(\mu_{Y,0}, \sigma_{\bar{Y}}^2)$ avec $\sigma_{\bar{Y}}^2 = \frac{\sigma_Y}{\sqrt{n}}$

3- Calcul de la p valeur quand σ_Y^2 est connue

$\frac{\bar{Y} - \mu_{Y,0}}{\sigma_{\bar{Y}}}$ ⇒ la version centrée réduite de $\bar{Y}$

elle admet une distribution normale standard sous H_0

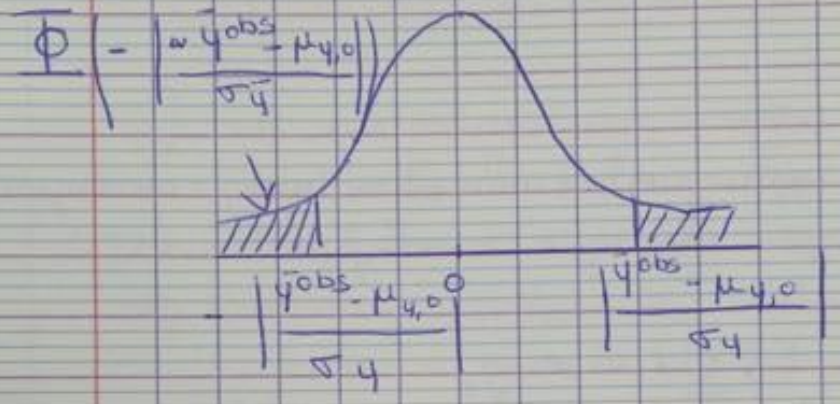
p valeur = proba d'obtenir un Y tq sous H_0

$$P\left(\left|\frac{\bar{Y} - \mu_{Y,0}}{\sigma_{\bar{Y}}}\right| > \left|\frac{\bar{y}^{obs} - \mu_{Y,0}}{\sigma_{\bar{Y}}}\right|\right)$$

$= P\left(|Z| > \left| \frac{\bar{y}^{obs} - \mu_{y,0}}{\sigma_{\bar{y}}} \right| \right)$ avec $Z \sim \mathcal{N}(0,1)$

si Φ la fonction de répartition de $\mathcal{N}(0,1)$

$P(d_1 < Z < d_2)$



valeur $p = 2 \times \Phi\left(-\left|\frac{\bar{y}^{obs} - \mu_{y,0}}{\sigma_{\bar{y}}}\right|\right)$

$$P\left(|Z| > \underbrace{\left|\frac{\bar{y}^{obs} - \mu_{y,0}}{\sigma_{\bar{y}}}\right|}_{c}\right) = 1 - P(|Z| < c)$$
$$= 1 - P(-c < Z < c)$$
$$= 1 - (\Phi(c) - \Phi(-c))$$
$$= \Phi(-c) + \Phi(-c)$$
$$= 2 \times \Phi(-c)$$

surface des queues de la distribution normale centrée réduite au delà de $\pm\left|\frac{\bar{y}^{obs} - \mu_{y,0}}{\sigma_{\bar{y}}}\right|$

4 - Variance empirique écart type empirique (erreur type / écart type estimé)

Variance empirique

$$S_Y^2 = \frac{1}{n-1} \sum_{i=1} (Y_i - \bar{Y})^2$$

écart-type $S_{\bar{Y}}$

La variance empirique est un estimateur convergent de la variance théorique. $S_Y^2 \xrightarrow{P} \sigma_Y^2$

$Y_1, \dots Y_n$ iid $E(Y^4) < \infty$

$$S_Y^2 = \frac{1}{n-1} \sum^n (Y_i - \bar{Y})^2 = \frac{1}{n-1} \sum (Y_i - \mu_Y + \mu_Y - \bar{Y})^2$$

$$= \frac{1}{n-1} \sum (Y_i - \mu_Y)^2 - \frac{2}{n-1} \sum (Y_i - \mu_Y)(\bar{Y} - \mu_Y) + \frac{1}{n-1} \sum (\bar{Y} - \mu_Y)^2$$

$$= \left(\frac{n}{n-1}\right)\left(\frac{1}{n} \sum (Y_i - \mu_Y)^2\right) - \left(\frac{n}{n-1}\right)\left(\frac{2}{n} \sum (Y_i - \mu_Y)(\bar{Y} - \mu_Y)\right) + \left(\frac{n}{n-1}\right)\left(\frac{1}{n} \sum (\mu_Y - \bar{Y})^2\right)$$

$\left(\frac{n}{n-1}\right)\left(\frac{1}{n} \sum (Y_i - \mu_Y)^2\right) \to \sigma_Y$

$\left(\frac{n}{n-1}\right)\left(\frac{2}{n} \sum (Y_i - \mu_Y)(\bar{Y} - \mu_Y)\right) \to 0$

$= \sigma_Y$

erreur type de $\bar{Y}$

on a $\sigma_{\bar{Y}} = \frac{\sigma_Y}{\sqrt{n}}$

↳ est un estimateur de $\sigma_{\bar{Y}}$

On l'appelle * erreur type de $\bar{Y}$
* écart type estimé : $\hat{\sigma}_{\bar{Y}}$
* "standard error" $SE(\bar{Y})$

$Y_1, \dots Y_n$ iid ~ Bernoulli (p) $\quad \text{var}(\bar{Y}) = \frac{p(1-p)}{n}$

$$\hat{\sigma}_{\bar{Y}} = SE(\bar{Y}) = \sqrt{\frac{p(1-p)}{n}}$$

5. Calcul de la valeur p quand σ_Y inconnue

S_Y^2 est un estimateur convergent de σ_Y^2

pour calculer la valeur p, on remplace simplement $\sigma_{\bar{Y}}$ par $\hat{\sigma}_{\bar{Y}}$

σ_Y inconnue, $Y_1, \dots, Y_n$ i.i.d

valeur $p = 2\Phi\left(-\left|\frac{\bar{Y}^{obs} - \mu_{Y,0}}{SE(\bar{Y})}\right|\right) = 2\Phi\left(-\left|\frac{\bar{Y}^{obs} - \mu_{Y,0}}{\hat{\sigma}_{\bar{Y}}}\right|\right)$

6. la statistique t

$t = \frac{\bar{Y} - \mu_{Y,0}}{SE(\bar{Y})}$ suit une loi de Student

distribution asymptotique suit une loi normale $\mathcal{N}(0,1)$ pour n grand (sous H_0)

Si on note t^{obs} la valeur obs de t

$$t^{obs} = \frac{\bar{Y}^{obs} - \mu_{Y,0}}{\hat{\sigma}_{\bar{Y}}}$$

alors si n grand, la valeur $p = 2\Phi(-|t^{obs}|)$

Exemple

$n = 200$ observé

w salaire horaire moyen.

$\bar{w}^{obs} = 22,64$ €

$s_w = 18,14$ € $= \sqrt{\frac{1}{200-1} \sum_{i=1}^{200} (w_i - \bar{w})^2}$

écart-type estimé de $\bar{w}$

$$\sigma_{\bar{w}} = \frac{s_w}{\sqrt{n}} = \frac{18,14}{\sqrt{200}} = 1,28.$$

$$t^{obs} = \frac{22,64 - 20}{1,28} = 2,06.$$

valeur p: $2 \overline{\Phi}(-|2,06|)$

$= 2 \times 0,0197$

$= 0,039 = 3,9\%$

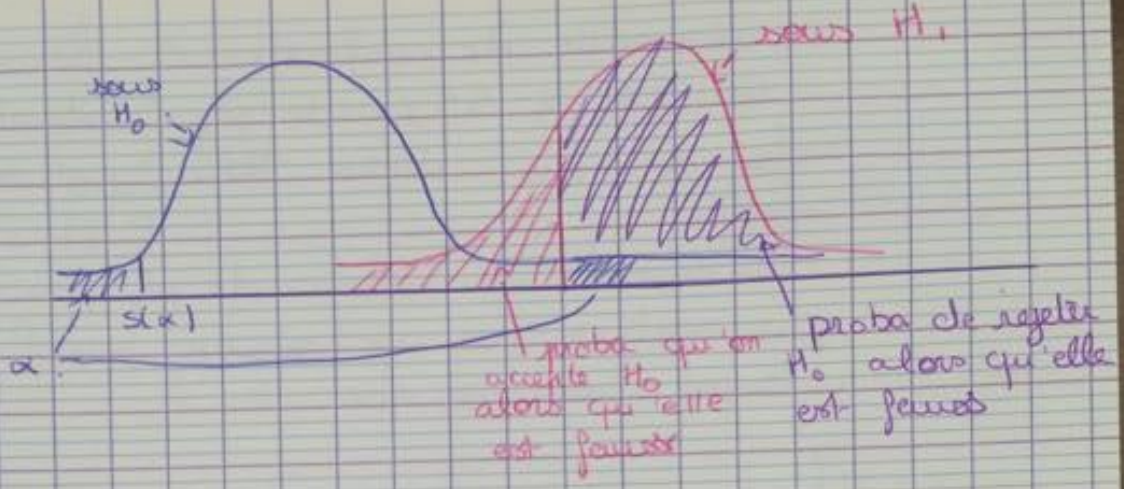
7. tester une hypothèse avec un seul de signification préspécifié

2 types d'erreur.

- rejeter H_0 alors qu'elle est vraie
  - erreur de première espèce $\alpha(a)$
- accepter H_0 alors qu'elle est fausse
  - erreur de deuxième espèce $\beta(b)$

puissance d'un test

aptitude à rejeter H_0 alors qu'elle est fausse



test d'hypothèse, en utilisant un seuil de signification fixé

règle { on rejette H_0 si la proba marginale de rejet est inférieure à un certain seuil par ex 5%.

rejet de H_0 si $|t^{obs}| > 1,96$ (pour n grand).

Ainsi le risque d'erreur de première espèce est de 5%

- niveau de signification 5%
- valeurs critique du test bilatéral est 1,96

Région de rejet correspond à la valeur de la statistique t en dehors de $\pm$ 1,96

8. intervalle de confiance

Un IC à 95% de μ_X contient la vraie valeur de μ_Y dans 95% des échantillons répétés.

L'ensemble des valeurs de μ_Y qui ne sont pas rejetées par un test au seuil de signification de 5%

$$\left\{ \mu_Y : \left| \frac{\bar{Y} - \mu_Y}{s_Y / \sqrt{n}} \right| \leq 1,96 \right\}$$

$$\Leftrightarrow \left\{ \mu_Y \; -1,96 \leq \frac{\bar{Y} - \mu_Y}{s_Y / \sqrt{n}} \leq \bar{Y} - \mu_Y \leq 1,96 \frac{s_Y}{\sqrt{n}} \right\}$$

$$IC_{95\%} = \left[\bar{Y} \pm 1,96 \frac{s_Y}{\sqrt{n}} \right] \Rightarrow \bar{Y} - 1,96 \frac{s_Y}{\sqrt{n}} \leq \mu_Y \leq \bar{Y} + 1,96 \frac{s_Y}{\sqrt{n}}$$