STATISTIQUE (1A)

Caractéristiques → position : moyenne, mode, médiane, Q, min, max, lim min, lim max

→ dispersion : étendue, IIQ, variance, écart type, CV, SCE

forme

IIQ : écart interquartile

CV : coef. de variation

médiane : valeur à la moitié de l'effectif [illegible]

mode : valeur prise le + souvent

$\lim \inf = Q_1 - 1{,}5\ IIQ$

$\lim \sup = Q_3 + 1{,}5\ IIQ$

$s^2x = \left[\frac{1}{n}\sum x_i^2\right] - \bar{x}^2$

$CV = \frac{sx}{\bar{x}}$

Vérifier exo's

Détermination par récurrence :

* loi de départ : $P(k) = \binom{n}{k} p^k q^{n-k}$ avec $q = p-1$

Binomiale $B(n,p)$

* loi de récurrence : $P(X=k) = \frac{n-k+1}{k} \cdot \frac{p}{q} P(X=k-1)$

$E(X) = \mu = np$

$V(X) = npq$

$\sigma = \sqrt{npq} = \sqrt{V(x)}$

Poisson $P(\lambda)$

$P(X=k) = \frac{\lambda}{k} P(X=k-1)$

$E(x) = V(x) = \lambda$

Théorique

effect. théorique $T_i = \frac{\text{tot. colonnes} \times \text{tot. lignes}}{\text{TOTAL}}$

O_i les valeurs observées

$\chi^2_{calculé} = \sum \frac{(O_i - T_i)^2}{T_i}$

χ^2_{seuil} est donné par la table : $\nu = (\text{nbre lignes} - 1)(\text{nbre colonne} - 1)$

$p = 1 - \alpha$

Si $\chi^2_{calculé} \gg \chi^2_{seuil}$ les variables ne sont pas indpdts.

CH. 1 STATISTIQUES

Définitions : individu, population, variable, modalité, échantillon, effectif, rang, enquête, recensement, sondage, expérimentation.

Statistiques univariées

1. Type de variables
   - → qualitative → nominal / ordinale
   - → quantitative → discrète / continues

2. Synthèse en tableaux
   3 types de séries : statistique, en effectif, en fréquence
   Tableau de contingence

3. Synthèse en graphique

4. Synthèse par paramètres

paramètre
- → position : moyenne, mode, médiane, minimum, maximum, quantile
- → dispersion : étendue, IIQ, variance, écart type, cœf. var
- → forme : γ_1, γ_2 *

NB : lim inf $= Q_1 - 1{,}5\,IIQ$

variance : $\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2 = \frac{SCE}{n} = s^2x$

écart-type : sx

coef. de variation $CV = \frac{sx}{\bar{x}}$

Rmq : Si $CV < 10\%$ → série homog

CH. 2 THEORIE

Définitions : expérience aléatoire, fonction aléatoire, loi de probabilité

pour x une variable aléatoire discrète : $E(x) = \sum_{i=1}^{n} p_i x_i$

A - Lois continues

1. Loi normale $X \sim N(\mu, \sigma)$
   → standard

centrer-réduire : $T = \frac{X-\mu}{\sigma}$

95% entre -3μ et 3μ

Si γ_1
- $= 0$ → symétrique
- > 0 → étalé à droite
- < 0 → étalé à gauche

Si γ_2
- $= 0$ → mésocurtique
- > 0 → leptocurtique
- < 0 → platicurtique

2 Loi du Chi deux   nddl   nb ddl > 30 $\chi^2 \rightarrow N(\mu, \sigma)$

3 Loi de Student

4 Loi de Fisher

B- Lois discrètes > Bino $P(X=k) = \binom{n}{k} p^k q^{n-k}$ ↗ $p \simeq 0{,}5 \rightarrow$ Fct densité symétrique ↘ n grand $\rightarrow B(n,p) \rightsquigarrow N(np, \sqrt{npq})$

> Poisson $P(X=k) = e^{-\lambda} \dfrac{\lambda^k}{k!}$

CH 3 BIVARIÉES

1) Qualitative / Qualitative

effectif théorique $n'_{ij} = \dfrac{n_{\cdot j} \times n_{i \cdot}}{n_{\cdot\cdot}}$

↳ Test Chi deux

REPRENDRE

2) Quanti / Quanti

covariance : $s_{xy} = \dfrac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) = \dfrac{SPE}{n}$

si s_{xy} $= 0 \rightarrow$ x et y indpd

$> 0 \rightarrow$ x et y à $\hat{m}$ variation

$< 0 \rightarrow$ x et y à variation $\neq$

Regression linéaire $y = ax + b$ $a = \dfrac{s_{xy}}{s^2_x}$ $b = \bar{y} - a\bar{x}$

coef. de correlation : $r = \dfrac{s_{xy}}{s_x s_y}$ si r $\simeq 0 \rightarrow$ pas corrélées; $\simeq 1 \rightarrow$ corrélées +; $\simeq -1 \rightarrow$ corrélées −

coef. de détermination : $R^2 = r^2 = \dfrac{V_{regression}}{V_{totale}}$

V_{totale} = Variance régression + Vari. résiduelle

s^2_y $s^2_{\hat{y}}$ s^2_e

Modèle linéaire si $\varepsilon_i \sim N(0, \sigma_e)$

analyse des résidus REPRENDRE

3) Quanti / Quali

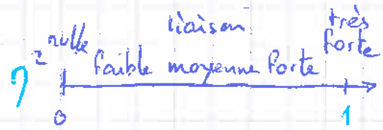
V_{total} = calculée sur tous les points (15)

$V_{intergrp} = V_{between}$ = variance entre les moy (3)

$V_{intragrp} = V_{within}$ = variance dans les grp (5)

$$V_{total} = V_B + V_W$$

eta deux : $\eta^2 = \frac{V_B}{V_T}$



STATISTIQUE

La statistique : ensemble de méthodes pour collecter, organiser et analyser des données numériques.

Mec important : Fisher, agronome

(≠ les statistiques)
Staadskunde : connaissance de l'Etat (XVIIIe siècle)
\+
Maths
=
statistiques

La statistique permet d'apréhender la VARIABILITE

2 Types de stat
→ Descriptive (organisation, synthèse ...)
→ Inférentielle (partir d'un échantillon pour généraliser)

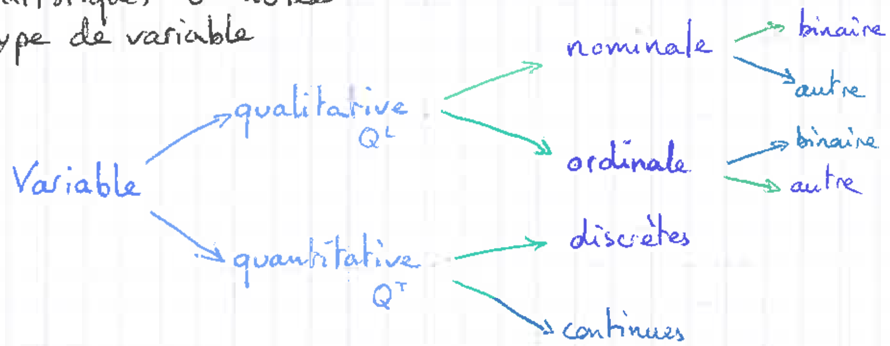
Définitions

Anglais ↓

unit — individu (ou unité statistique) : élément unitaire d'une étude (pas forcément un homme !)
population : ensemble des individus considérés par l'étude
≃ data — variable : information collectée sur chaque individu
level — modalité : valeur prise par la variable
sample — échantillon : sous ensemble d'une population
size — effectif : nbre d'individus
rank — rang : numéro d'ordre d'une observation
Enquête : collecte d'infos sur un groupe d'individus pris dans leur cadre habituel
rescensement : enquête exhaustive
sondage : enquête sur un échantillon
expérimentation : collecte d'informations dans une situation provoquée et contrôlée par l'expérimentateur.

Statistiques univariée

1. Type de variable



discrète : dénombrables (on a des nbres entiers)
continues : on a pas forcément de nbres entiers.
nominale : inclassable (ex : sexe, couleur ...)
ordinale : classable (ex : j'aime un peu, beaucoup, passionnément, à la folie.)

2 - Synthèse en tableau

X : série statistique $\{x_1, x_2, x_3 \dots x_n\}$ n valeurs

série en effectif $\{(x_1, n_1), (x_2, n_2), (x_3, n_3) \dots\}$ $\sum n_i = n$

série en fréquence $\{(x_1, f_1), (x_2, f_2), (x_3, f_3) \dots\}$ $f_i = \frac{n_i}{n}$

x_i	n_i	f_i	$n^+(\leq x_i)$	$f^+(\leq x_i)$

i.e + c'est pour "cumulé"

tableau de contingence

3 Synthèse en graphique

Tuky 1977 EDA (Explorating data analysis)

3.1. Quête

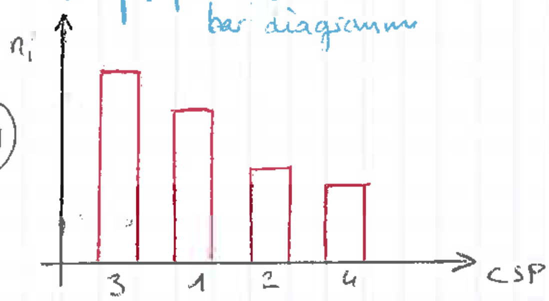
CSP : catégorie socio-professionnel

exemple

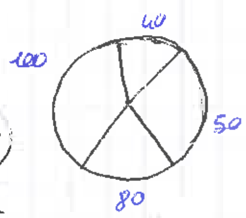
CSP	n_i
1	80
2	50
3	100
4	40

Graphique en barres
bar diagramm

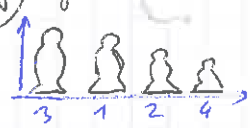
Diag 1



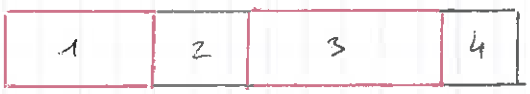
Diag 3
circulaire
pie-chart



Diag 4 Figuratif



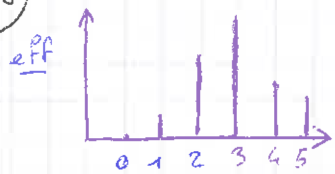
Diag 2

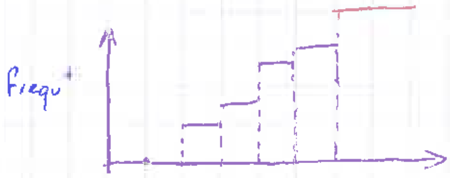


Les graphique 3D sont faussés. Ils ne sont intéressants qu'en cas de plusieurs variables.

Diag 5

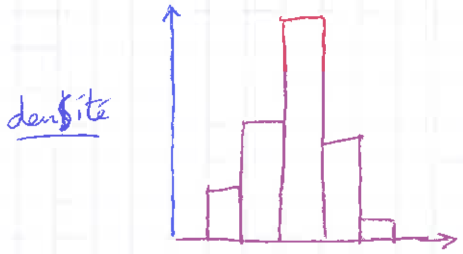
exemple 2
en bâton





exemple 3

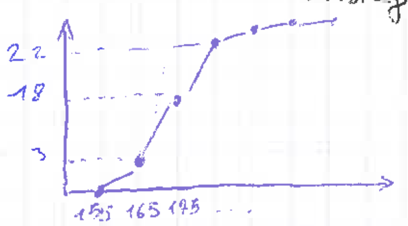
classe	n_i
155 - 165	3
165 - 175	7
175 - 185	12
185 - 195	5
195 - 205	1
	28



Surface = $d \times a$

$d = \frac{S}{a} = \frac{eff}{a} = \frac{n_i}{a}$

Histogramme ou "fonction densité"



Fonction de répartition

sigmoïde → ensemble de segments

Boîte de Tukey, boîte à moustaches box plot

Synthèse par paramètres

1) paramètre de position → moyenne, mode, médiane, quantiles, minimum, maximum
2) paramètre de dispersion → étendue, IIQ, variance, écart-type, coef. de variation
3) paramètre de forme

1) Paramètre de position

moyenne : $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$ — mais cette valeur, très sensible aux extrême peut ne pas être représentative. On peut utiliser alors
mean

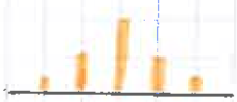
moyenne tronqué : $x_t = \frac{1}{n_t} \sum_{i=1}^{n} x_i$
trimmed mean

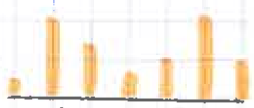
moyenne mobile : $\bar{x}_m = \frac{1}{n_m} \sum_{i=1}^{m} x_i$

moyenne pondérée $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} n_i x_i$
weight mean

moyenne géométrique $\bar{x}_g = \sqrt[n]{x_1 x_2 \dots x_n} = \left(\prod_{i=1}^{n} x_i\right)^{\frac{1}{2}}$

mode : valeur au plus grand effectif.
csquences : il n'y a pas toujours de mode


unimodale


bimodale

médiane m_e ou $\tilde{x}$: valeur séparant une série en deux groupes.

→ si on a un effectif impaire en valeurs discrètes
ex : 2 4 8 10 12
médiane

→ dans les autres cas, on fait une moyenne des 2 valeurs centrales.

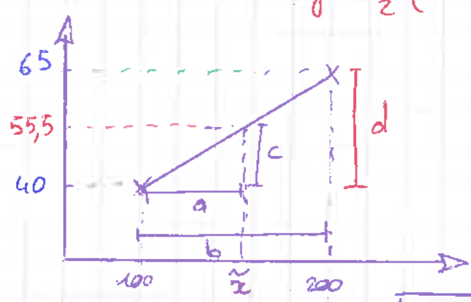
→ si x est continue (exemple)

x_i	n_i	n^+
[0;100[	40	40
[100,200[	25	65
[200;300[	45	110

dans cette classe

$\text{rang} = \frac{1}{2}(110+1) = 55,5$

Pour trouver la médiane, on cherche le rang $= \frac{1}{2}(n+1)$



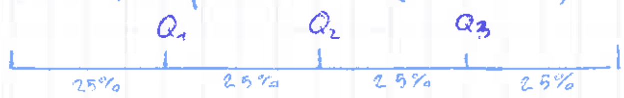
$\frac{a}{b} = \frac{c}{d}$

$a = \frac{bc}{d} = \frac{100 \times 15,5}{25}$

$a = 62$

$\boxed{\tilde{x} = 162}$

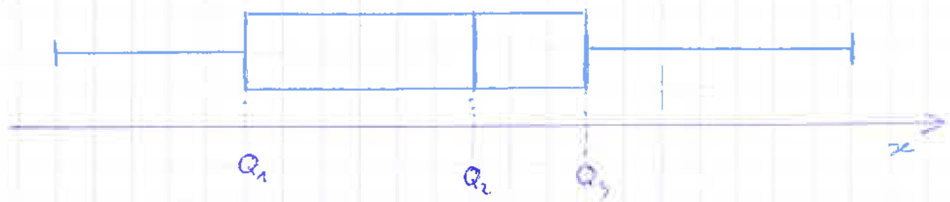
NB : la médiane est peu sensible aux variations quantiles : terciles, quartiles (divisions, limites des effectifs)



$Q_1 =$ rang $(25\%) = p\,(n+1)$ p, le pourcentage, ici 0,25

Par exemple, le centile 70 : $C_{(70)} = 0,7\,(n+1)$

Exemple de boîte à moustaches



Intervalle interquantile : $IIQ = Q_3 - Q_1$

Limites → supérieure : $\lim inf = Q_1 - 1{,}5\,IIQ$
→ inférieure : $\lim sup = Q_3 + 1{,}5\,IIQ$

2) Paramètre de dispersion

* étendue = valeur max − valeur min
* $IIQ = Q_3 - Q_1$

* variance : on s'intéresse à la distance des valeurs à la moyenne
↳ écart à la moyenne $(x_i - \bar{x})$

on s'intéresse à l'ensemble des écarts à la moyenne :

$$\sum_{i=1}^{n} (x_i - \bar{x}) = \sum_i x_i - \sum_i^n \bar{x} = n\bar{x} - n\bar{x} = 0 \ldots$$

$$\sum^n (x - \bar{x})^2 = SCE \rightarrow \text{somme des carrés des écarts à la moyenne}$$

$$\text{variance} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2 = \frac{SCE}{n} = s_x^2$$

⚠ la variance est en unité^2 !

Rmq : si les données sont groupées, $s^2_x = \frac{1}{n} \sum_{i=1}^{n} n_i (x_i - \bar{x})^2$

* écart-type : $s_x = \sqrt{s_x^2}$
standard deviation
* CV (coef. de variation) $= \frac{s_x}{\bar{x}}$
On considère que les données sont homogènes autour de la moyenne si le CV est < à 10% → série homogène

$X_i \quad Y_i = a + bX_i$

$\bar{Y}_i = a + b\bar{X}_i$

$s^2_{Y_i} = b^2 s^2_{X}$

n_i	eff	moy	variation
1	n_1	m_1	s_1^2
2	n_2	m_2	s_2^2
⋮	⋮	⋮	⋮
k	n_k	m_k	s_k^2
⋮	⋮	⋮	⋮
I	n_I	m_I	s_I^2

$\bar{x} = \frac{1}{N} \sum_k n_k m_k$

$$s^2_x = \underbrace{\left(\frac{1}{N} \sum_k n_k s_k^2\right)}_{\text{moy des variances}} + \underbrace{\frac{1}{N} \sum n_k (m_k - \bar{X})^2}_{\text{variance des moy}}$$

CH2. Théories

1) Définition

Fonction aléatoire : fct prennant comme valeurs les résultats possibles d'une aléatoire

expérience aléatoire : expé dont on ne peut prévoir le résultat.

ex : lancé d'un dé → équilibré

x	1	2	3	4	5	6
$P(x)$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

loi de probabilité (lien X et $P(X)$)

Si X est une variable aléatoire discrète

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

$$V(X) = E\left[(X - E(X))^2\right] = E(X^2) - E(X)^2$$

avec l'exemple du dé : $E(X) = \frac{1}{6}(6+5+4+3+2+1) = \frac{21}{6} = \frac{7}{2} = 3{,}5$

$V(X) =$

Si X est une va continue

$$E(X) = \int_{-\infty}^{+\infty} x\, f(x)\, dx \qquad \text{avec } f(x) \text{ la fct densité}$$

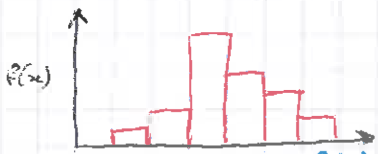
$$V(X) = \int_{-\infty}^{+\infty} x^2 f(x)\, dx - E(x)^2 = E(X^2) - E(X)^2$$

2) Lois continues

2.1 Loi normale (Laplace-Gauss)

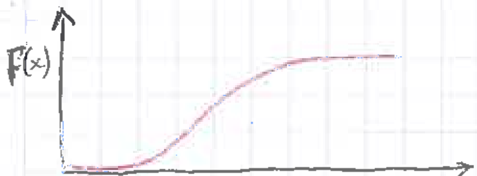
Soit X une variable aléatoire continue empirique

$X \{x_1, x_2, x_3 \dots x_n f\}$



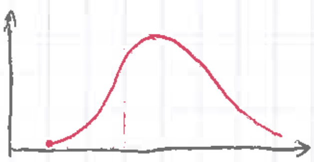
Fct densité $f(x)$

⇩ approche par loi normale



Fct de répartition $F(x)$

(échangée par la loi normale)



Fct $f(x)$ modelisée

$\alpha(X) = N(\mu, \sigma)$

$X \sim N(\mu, \sigma)$

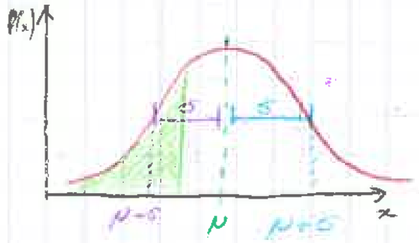
$\begin{pmatrix} E(X) = \mu \\ V(X) = \sigma^2 \end{pmatrix}$

Loi

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

$$F(x) = \int_{-\infty}^{+\infty} f(x)\, dx = P(X \leq x)$$

Graphiquement



→ point d'inflexion à $\mu-\sigma$ et $\mu+\sigma$

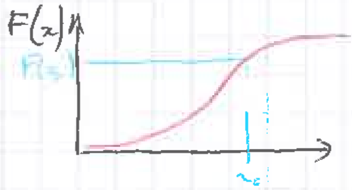
Loi Normale standard

centrer - réduire : $T = \dfrac{X-\mu}{\sigma}$ c'est la loi normale standard

$T \sim N(0,1)$ → T suit une loi normale de ~~moyenne 0~~ espérance $E=0$ et de variance $V(T)=1$

Loi normale → toujours positive, tend vers 0 sans l'atteindre en $+\infty$ et $-\infty$.

$P(T \leqslant 0) = 0{,}5$

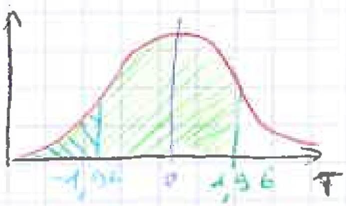


$P(T \leqslant 1{,}96)$ ⇝ regarde la table ⇝ $P(T \leqslant 1{,}96) = 0{,}9750$

aire totale sous $f(x) = 1$

Ainsi, par symétrie : $P(T \leqslant -1{,}96) = 1 - P(T \leqslant 1{,}96) = 0{,}025$

Puis $P(-1{,}96 \leqslant T \leqslant 1{,}96) = P(T \leqslant 1{,}96) - P(T \leqslant -1{,}96) = 0{,}95$



$P(T \leqslant -1{,}96)$

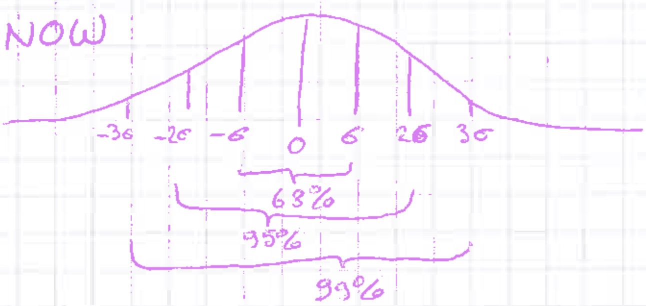
$P(T \leqslant 1{,}96)$

$P(-1{,}96 \leqslant T \leqslant 1{,}96)$

Utilisation de la table :

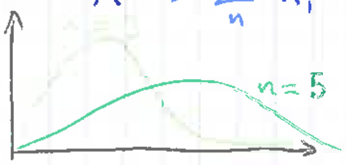
t	$P(T \leqslant t)$	$P(-t \leqslant T \leqslant t)$
1	0,8413	
−1	0,1587	
2	0,9772	
−2	0,0228	
3	0,99865	
−3	0,00135	

TO KNOW



2.3 Loi du "Chi" deux ou Pearson

$\chi^2 = \sum_n x_i^2$ suit une loi de χ^2 à n degré de liberté (n ddl)



pour $n > 30$, χ^2 tend vers une loi normale

$E(\chi^2) = n$

$V(\chi^2) = 2n$

2.4 Loi de Student

$U \sim N(0,1)$

$V \sim \chi^2_{n\,ddl}$

$T_n \sim \dfrac{U}{\sqrt{\frac{V}{n}}} \sim t_n\,ddl$

Cette loi est symétrique et a pour moyenne 0

$E(T_n) = 0$

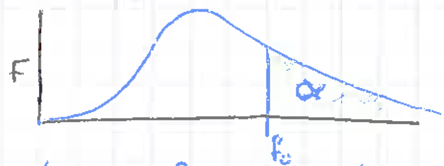
$V(T_n) = \dfrac{n}{n-2}$ pour $n > 2$

Si n est assez grand, la loi Student tend vers la loi normale (pour $n > 30$)

2.5 Loi de Fisher-Snedecor

Soit U et V suivent une loi de χ^2 respectivement à n et m ddl.

$F = \dfrac{U/n}{V/m} \sim F_{n,m}\,ddl$



C'est une fct de densité, donc l'aire sous la courbe vaut 1.

Si on veut savoir quelle est la probabilité pour $F < f_0$

On écrit $P(F < f_0) = 1 - \alpha$

Cette loi de Fisher est surtout utilisée pour les variances

Il y a plusieurs tables de Fisher ← à vérifier

NB : n ddl peut s'écrire ν : n ddl → $\nu 1$

m ddl → $\nu 2$

3) Lois discrètes
3.1 Loi Binomiale

Schéma de Bernoulli : soient A et B complémentaire

$$P(A) = p$$
$$P(B) = 1 - p = q$$

Soit n réalisant aléatoirement m indépendante et indifféremment distribuée (iid)

On utilise cette loi sur une pop. finie ou infinie, en tirage avec remise

$$P(k \text{ ordonné}) = p^k q^{n-k}$$

Combinaison de k parmi n : $C_n^k = \binom{n}{k} = \frac{n!}{(n-k)!\,k!}$

D'où $P(X=k) = p(k) = C_n^k p^k q^{n-k}$

Loi de probabilité de la loi binominale $X \sim B(n,p)$

(on dit "binominale" car $C_n^k p^k q^{n-k}$ est le terme g^{al} du binôme $(p+q)^n$)

$$P(k+1) = \frac{p}{q} \frac{n-k}{k+1} P(k)$$

$P(k+1) = P(X = k+1)$?

VOIR TD

Si p est proche de 0,5, la fct densité est symétrique.

Si n est grand, la loi binomiale de paramètre n, p tend vers une loi normale :

$$B(n,p) \longrightarrow N(np, \sqrt{npq})$$

<u>exemple</u> : un dé à 6 faces $(X=6)$ en 4 lancés $(n=4)$ avec à chaque fois $p(\text{face } 6) = \frac{1}{6}$

$$X \sim B\left(4, \frac{1}{6}\right)$$

P	p
(X=0)	0,4823
(X=1)	0,3858
(X=2)	0,1157
(X=3)	0,0154
(X=4)	0,0008

⇐ On étudie les probabilités du nbr de face(s) 6 obtenue(s) en 4 lancé

On peut agrandir la loi en "polynômiale"
On peut retirer la remise, la loi est alors hypergéométrique.

32. Loi Poisson

Cas limite de la loi binômiale dans laquel n est grand et p est petit.
n est grand, p est petit.

$$n \longrightarrow \infty \quad \text{et} \quad p \longrightarrow 0 \quad \text{et} \quad np = \lambda$$

$$P(X=k) = P(k) = e^{-\lambda} \frac{\lambda^k}{k!}$$

et on note $X \sim \mathcal{P}(\lambda)$ on a $E(X) = V(X) = \lambda$

En pratique p (ou q) $< 0{,}1$ et np (ou nq) < 5

$$P(X=k) = \frac{\lambda}{k} P(X=k-1)$$

si n grand $\mathcal{P}(\lambda) \longrightarrow N(\lambda, \sqrt{\lambda})$

On se sert de cette loi pour modeliser par exemple le
le nbre de coquilles (fautes) par page d'un roman,
le nbre d'appels reçus dans une agence.

CH. 3 BIVARIÉES

Dans une expérience, on étudie d'abord les variables de façon indpdte.
on parle de variables à plat.

Then, on les étudie par deux pour voir dans quelle mesure elles sont liées.

exemple : le poids augmente quand la taille augmente.

On va étudier 3 cas
- → les 2 variables sont quantitatives
- → les 2 variables sont qualitatives
- → une variable qualitative en fct d'une quantitative

quanti qual

1) Tableau individu-variable

On met les individus en ligne et les variables en colonne.

	V_1	V_2	...	p
individu1	x_{11}	x_{12}		
individu2	x_{21}	x_{22}		
...				
...				
individun	x_{n1}	x_{n2}		

L'indice i de 1 à n.
La variable j de 1 à p.

2) Les 2 variables sont qualitative

Exemple :

Sexe	Fumeur
F	F
F	NF
F	F
H	F
H	F
H	NF
H	NF

⟹

	F	NF
F	2	1
H	2	2

En nommant :

	F	NF	Total
F	n_{11}	n_{12}	$n_{1.}$
H	n_{21}	n_{22}	$n_{2.}$
Total	$n_{.1}$	$n_{.2}$	$n_{..}$

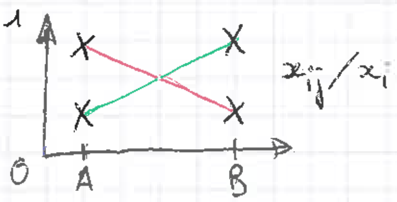
← distribution marginale des lignes

$$n_{.1} = n_{.j=1} = \sum_{i=1}^{n} n_{i1}$$

$$n_{..} = \sum_i \sum_j n_{ij}$$

$\frac{n_{11}}{n_{1.}}$	$\frac{n_{12}}{n_{1.}}$	1
$\frac{n_{21}}{n_{2.}}$	$\frac{n_{22}}{n_{2.}}$	1

← Distribution continuelle en ligne



Soit les proba. P(A) et P(C)
si ils sont indépendant $P(A \text{ et } C) = P(A) \times P(C)$

$$P(A \text{ et } C) = \frac{n_{.1}}{n_{..}} \times \frac{n_{1.}}{n_{..}}$$

	A	$\bar{A}$	
C	n_{11}	n_{12}	$n_{1.}$
$\bar{C}$	n_{21}	n_{22}	$n_{2.}$
	$n_{.1}$	$n_{.2}$	$n_{..}$

On devrait observer $P(A \text{ et } C) \times n_{..} = \frac{n_{.1} \times n_{1.}}{n_{..}}$

Sous l'hypothèse d'indépendance :

effet théorique : $n'_{ij} = \frac{n_{.j} \times n_{i.}}{n_{..}}$

Exemple :

Une étude sur les fumeurs selon le sexe donne :

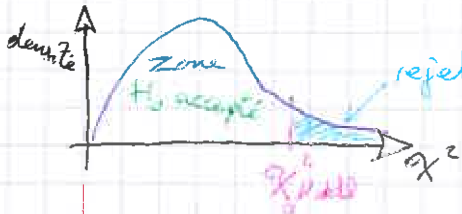
	F	NF	
♀	20	18	38
♂	25	10	35
	45	28	73

Si l'on considère les critères indépendants :

	F	NF	
♀	$\frac{45 \times 38}{73}$ 23,43	14,57	38
♂	21,57	13,43	
	45		73

2.2 Test du Chi² (χ^2)

- H_0 est l'hypothèse nulle, dont les variables sont indpdtes
- α est le risque de se tromper en rejetant H_0, risque de 1ère espèce
  α est une proba $\alpha = 5\% = 0{,}05$
- $\chi^2 = \sum_{ij} \frac{(n_{ij} - n'_{ij})^2}{n'_{ij}} = \sum_k \frac{(O_k - T_k)^2}{T_k}$ k : évite d'écrire ij
- χ^2_{seuil} ? $\chi^2_{calculé} \sim \chi^2_{\alpha \; \nu ddl}$



$\nu = (\text{nbre de ligne} - 1)(\text{nbre de colonnes} - 1)$

$\chi^2_{calculé} = \frac{(20 - 23{,}43)^2}{23{,}43} + \frac{(18 - 14{,}57)^2}{14{,}57} + \dots$

$\chi^2_{1ddl \; 5\%} = 3{,}84$

2.3 Test du Chi² (χ^2) d'adéquation

Pour savoir :
Si une distribution observée suit une distribution particulière (loi normale...)

exemple : une variable aléatoire continue : X suit il une loi normale ?
$X \sim N$?

classe	ni	P théorique
[0, 100[	10	P (X < 100)
[100, 200[	15	P (100 ≤ X < 200)
[200, 300[	32	
[300, 400[	17	
[400, 500[	12	
[500, 600[	4	P(X > 500)
Tot.		

ef. théorique

proba Tot

$\chi^2_{calc} \sim \chi^2_{\alpha \; \nu ddl}$

avec :
$\nu = Nb_{classes} - 1 - Nb$ paramètres estimés

3) Quanti / quanti

Deux variables quantitatives X et Y : sont-elles liées ?
Soit Y la variable dépendante, ou réponse
à la variable indpd, ou explicative, ou régresseur

3 1 La covariance

$$s_{xy} = \frac{1}{n} \underbrace{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}_{\text{SPE}} = \frac{SPE}{n}$$

SPE (somme des produits des écarts)

$$s_{xy} = \frac{1}{n} \sum_{i=1}^{n} x_i y_i - \bar{x}\,\bar{y}$$

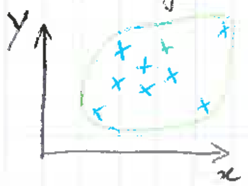
La covariance est
- positive si ~~x~~ y augmente qd x augmente
- négative si ~~x~~ y ~~augmente~~ diminue qd x augmente
- proche de 0 si x et y sont indpd *

* pas de réciproque

Qualité Régression

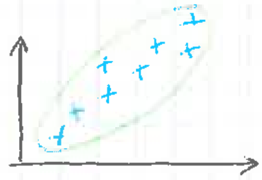
3 2 Le nuage de point

Représentation de données bivariées



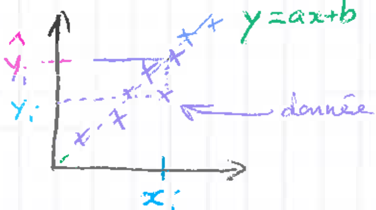
← Pas de relation linéaire

Existence d'une relation linéaire →



Donnée out
Homogénité du nuage

3 3 Régression linéaire



Résidu : $\varepsilon_i = y_i - \hat{y}_i$

on cherche à minimiser $\sum \varepsilon_i^2$
on va trouver le moindre carré ordinaire (MCO)

$$a = \frac{s_{xy}}{s_x^2} \quad \text{et} \quad b = \bar{y} - a\bar{x}$$

Qualité de la régression

coefficient de corrélation : $\boxed{r = \frac{s_{xy}}{s_x s_y}}$ compris entre −1 et 1 et du signe de la covariance

Si r est proche de 1, les deux variables sont corrélées positivement
Si r est proche de 0, les deux variables ne sont pas corrélées.
Si r est proche de −1, les deux variables son corrélées négativement

coefficient de détermination $\boxed{R^2 = \frac{V_{\text{régression}}}{V_{\text{totale}}}}$ $R^2 = r^2$

Tableau : décool

	X	Y	$\hat{Y}$	ε_i
valeurs				
moyenne déduite	$\bar{x}$	$\bar{y}$	$\bar{\hat{y}}$	$\bar{\varepsilon}$
$s^2 x$	s^2_x	s^2_y	$s^2_{\hat{y}}$	s^2_ε

$\hat{y} = a x_i + b$
$\varepsilon = \hat{y}_i - y_i$

Variance totale = V régression + Variance résiduelle
$$s^2_y = s^2_{\hat{y}} + s^2_\varepsilon$$

34 Modèle linéaire

$$\left.\begin{array}{l} y_i = a x_i + b + \varepsilon_i \\ \hat{y}_i = a x_i + b \end{array}\right\} \varepsilon_i \sim N(0, \sigma_\varepsilon)$$

35 Qualité de la régression

36 Dangers de la régression

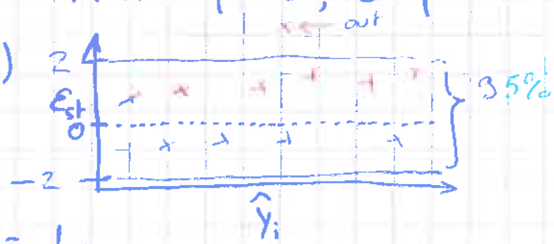
1 – Toujours commencer par faire un graphique
2 – La corelation peut être très bonne mais n'est jamais une démonstration de cause à effet.

analyse des résidus

On fait l'hypothèse que $\varepsilon_i \sim N(0, \sigma_{resi})$

- tests de normalité : histogramme, test χ^2 d'adéquation, QQ plot.
- graphe des résidus standardisés :

$$\varepsilon_{standard_i} = \frac{\varepsilon_i}{s_\varepsilon} \sim N(0,1)$$



De ce graphe ne doit pas présenter les points : – à tendance (genre croissant, dents de scie...)
– sur une zone, tous du m̂ signe

3 critères de détermination sur la qualité d'une régression :
1 coef. de corrélation
2 coef. de détermination
3. analyse de résidus

Exercice n° de TD-1
Statistique à une variable

Exercice n° 1

Ecrivez les sommes suivantes en détaillant chaque terme :

$$\sum_{i=1}^{6} x_i \qquad \sum_{i=1}^{6} x_i^2 \qquad \sum_{i=1}^{4} x_i - 3^2 \qquad \sum_{j=1}^{4} (x_j - 3)^2$$

$$\sum_{i=1}^{3} (x_i - \alpha) \qquad \sum_{i=1}^{3} x_i - \alpha \qquad \sum_{j=1}^{n} \alpha \qquad \sum_{k=1}^{5} f_k x_k$$

Exercice n° 2

Ecrire à l'aide de la notation somme :

$X_1^2 + X_2^2 + \dots\dots\dots + X_6^2$

$f_1 X_1^3 + f_2 X_2^3 + \dots\dots\dots\dots + f_7 X_7^3$

$(X_1 + Y_1) + (X_2 + Y_2) + \dots\dots + (X_8 + Y_8)$

$(X_1 + Y_1)^2 + (X_2 + Y_2)^2 + \dots\dots + (X_8 + Y_8)^2$

$ab + ab + ab + ab + ab + ab + ab + ab$

$f_1 X_1 Y_1 + f_2 X_2 Y_2 + f_3 X_3 Y_3 + f_4 X_4 Y_4 + f_5 X_5 Y_5 + f_6 X_6 Y_6$

Exercice n° 3

Mettre sous forme de 3 sommes : $\sum_{j=1}^{n} (aX_j + bY_j - cZ_j)$

Exercice n° 4

Calculer : Σx $\quad \Sigma y$ $\quad \Sigma x^2$ $\quad \Sigma y^2$

Σxy $\quad \Sigma x \Sigma y$ $\quad \Sigma (x - y)(x + y)$

$\Sigma (x + y)^2$ $\quad \Sigma (x - y)^2$ $\quad (\Sigma x^2 - \Sigma y^2)$

si x et y prennent les valeurs suivantes :

x	2	- 5	4	- 8
y	- 3	- 6	10	6

Exercice n° 5

Arrondir chacun des nombres suivants avec la précision requise.

48,6 à 1 unité près
136,5 à 1 unité près
2,484 au 1/100ème
0,0435 au 1/1000ème
4,50001 à 1 unité près

143,95 au 1/10ème
366 à la dizaine près
24447 a un millier près
5,56500 au 1/100ème
5,56500 au 1/1000ème

Exercice n° 6

On étudie les résultats d'un recensement dans un département, en ce qui concerne la répartition de la population active, selon la catégorie socioprofessionnelle et le sexe :

CSP	hommes	femmes
A : agriculteurs	8964	2744
B : artisans, commerçants	7924	3388
C : cadres sup, professions libérales	4100	1432
D : professions intermédiaires	8532	6048
E : employés	9324	14392
F : ouvriers	24424	4392

Synthétiser l'information à l'aide des graphiques les plus judicieux.

Exercice n° 7

Déterminer les caractéristiques de position, de dispersion et la SCE de la série des résultats au test de dextérité manuelle suivants :

{73; 80; 71;89; 77; 84; 78; 74; 75; 73; 80; 83; 82; 74; 85; 82; 83; 80; 76; 81}

Exercice n° 8

Une enquête menée sur un échantillon de 55 familles a permis de mesurer le nombre d'enfants xi par famille :

xi : nombre d'enfants	ni : nombre de familles
0	1
1	4
2	6
3	9
4	17
5	11
6	4
7	2
8	1

1°) déterminer graphiquement le mode, la médiane et les quartiles de la distribution.

2°) calculer les caractéristiques de position, de dispersion et la SCE à l'aide de la calculatrice.

3°) calculer le coefficient d'asymétrie de Yule et les coefficients d'asymétrie et d'aplatissement de Fisher. Interpréter les résultats.

Approximation des calculs : 0,01

Vous établirez un tableau dans lequel figurera les colonnes pour xi, ni, xi², nixi, nixi², $(xi - \bar{x})$, $(xi - \bar{x})^2$, ni $(xi - \bar{x})^2$; ni $(xi - \bar{x})^3$, ni $(xi - \bar{x})^4$, vous pourrez ainsi comparer les résultats de la calculatrice à ceux calculés en appliquant les formules du cours aux calculs du tableau.

Exercice n° 9

Etudier, sur la distribution des dépenses mensuelles en loisirs des familles de 4 personnes d'une région socialement favorisée, les caractéristiques de position, de dispersion, la SCE et représenter graphiquement les fonctions de densité de fréquence et de répartition.

dépenses	nombre de familles
[1000; 1500 [	24
[1500; 2000 [	40
[2000; 2500 [	60
[2500; 3000 [	35
[3000; 3500 [	14
[3500; 4500 [	10

Déterminer, par le calcul et graphiquement, les quartiles et C25, C40, C50, C60 de cette distribution.

Exercice n° 10:

Voici un relevé du carnet de pesée des veaux blancs corréziens le 12 janvier 1971 à Objat (19). Les données sont en Kg.

169	205	160	179	128	159	140	201	195
179	181	161	188	192	141	186	155	159
141	162	124	153	173	166	173	166	157
149	126	181	203	189	163	193	173	158
162	121	165	144	169	157	155	186	166

1°) sur cette série brute des données déterminer les caractéristiques de position, de dispersion et la SCE.

2°) grouper cette série en classes d'amplitude 10 en partant de 120 comme limite inférieure de la 1ère classe (120 inclus).

3°) déterminer les caractéristiques de position de dispersion et la SCE sur la série groupée en classes.

4°) représenter graphiquement les fonctions de densité de fréquence et de répartition de la série groupée en classes.

5°) représenter sur la fonction de répartition de la série groupée le 20ème centile, le 2ème quartile, le 80ème centile.

Exercice n° 11

On considère la répartition par âge d'une population (effectifs en millier) :

Age : xi	effectif	
[0 ; 20 [	505	1°) représenter les fonctions de densité de fréquence et de répartition. Déterminer graphiquement les quartiles. 2°) calculer la moyenne et l'écart-type 3°) Calculer les caractéristiques de forme (Fisher) 4°) en utilisant la variable centrée réduite retrouver les valeurs des coefficients de position, de dispersion et de forme.
[20 ; 40 [	600	
[40 ; 50 [	258	
[50 ; 60 [	221	
[60 ; 75 [	332	
[75 ; 90 [	187	

TD1 Statistiques

Paramètres
- → position : moyenne, mode, médiane, Q, min, max, limite min, limite max
- → dispersion : étendue, IIQ, variance, écart type, CV

$$s^2x = \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2 = \left(\frac{1}{n}\sum_{i=1}^{n}x_i^2\right) - \bar{x}^2$$

Exercice 7

Q_1 / $\tilde{x}$ / Q_3

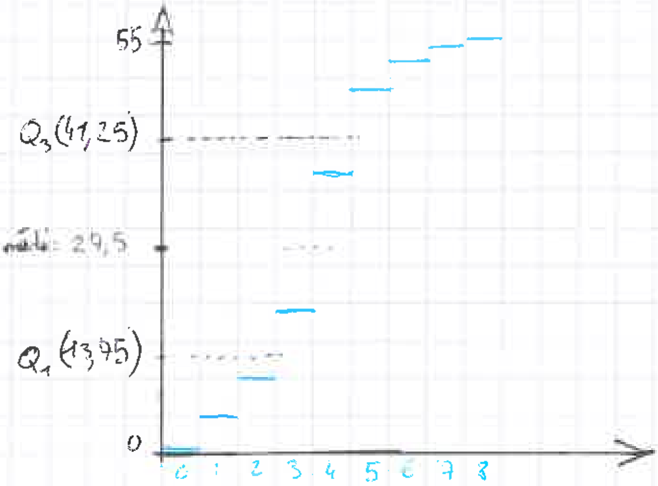
71, 73, 73, 74, 74, 75, 76, 77, 78, 80, 80, 80, 81, 82, 82, 83, 83, 84, 85, 89

Position
- moyenne : $\bar{x} = 79$
- mode : 80
- médiane : $\tilde{x} = 80$ (Q_2)
- $Q_1 = 74,5$
- $Q_3 = 82,5$
- min = 71, max = 89
- lim inf = $Q_1 - 1,5$ IIQ
- lim inf = $74,5 - 1,5(82,5 - 74,5) = 62,5$
- lim sup = $Q_3 + 1,5$ IIQ
- lim sup = 94,5

Dispersion
- étendue = 18
- IIQ = 8
- $s^2x = 21,7$
- $s\,x = 4,66$
- CV = 5,899 %
- SCE = $s^2x \times n = 434$

Exercice 8

Graphiq.
- mode → 4
- médiane → 4
- $Q_1 \to 3$
- $Q_3 \to 5$



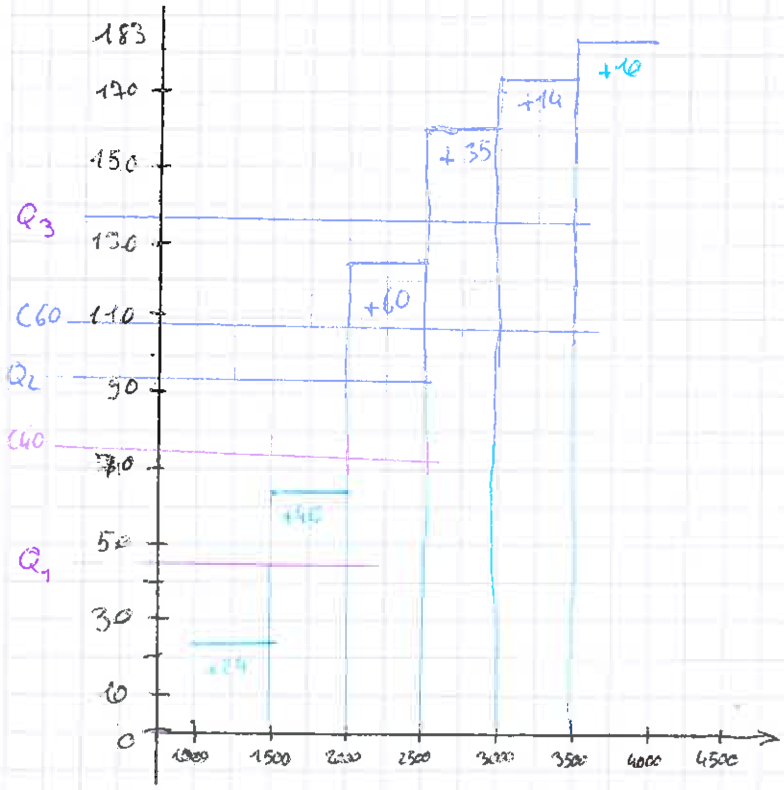
Position
- $\bar{x} = 3,85$
- min = 0 / max = 8
- lim min = 0 / lim max = 8
- étendue → 8
- IIQ = 2
- $s^2x = 2,668$
- $s\,x = 1,633$

lim max = $Q_3 + 1,5(Q_3 - Q_1)$

CV = 0,424

SCE = 146,74

Exercice 9



$C25 = Q_1 \Longrightarrow 45{,}75$
$C40 \Longrightarrow 73{,}2$
$C50 = Q_2 \Longrightarrow 91{,}5$
$C60 \Longrightarrow 109{,}8$
$C75 = Q_3 \Longrightarrow 137{,}25$

$Q_1 =$
$C40 =$
$Q_2 =$
$C60 =$
$Q_3 =$

Exercice de TD2

2 - Lois théoriques

Exercice n° 1 :

A l'aide de la table statistique déterminer :

1 - $P(X = 3)$ si $L(X) = B(5 ; 0,1)$
2 - $P(X < 4)$ si $L(X) = B(15 ; 0,09)$
3 - $P(X > 5)$ si $L(X) = B(20 ; 0,07)$
4 - $P(X \leq 6)$ si $L(X) = B(30 ; 0,08)$
5 - $P(X \geq 7)$ si $L(X) = B(40 ; 0,06)$
6 - $P(X = 0)$, $P(X = 1)$ et $P(X = 2)$ si $L(X) = B(7 ; 0,2)$ par récurrence.

Exercice n° 2

Représenter la fonction de densité de fréquence de la variable X telle que $L(X) = B(50 ; 0,1)$ Que remarquez-vous ? Calculer l'espérance mathématique, la variance et l'écart-type de la variable.

Exercice n° 3

Une enquête a montré que sur 200 flacons d'un même produit pharmaceutique, 50 ne pouvaient être utilisés au-delà de 4 mois après leur livraison. Ces 200 flacons sont rangés de façon aléatoire sur une étagère. Trois personnes achètent chacune 1 flacon dès le premier jour de livraison.
Identifier la variable aléatoire, sa loi et les paramètres de la loi.
Quelle est la probabilité pour que :

1. 2 de ces 3 personnes aient un flacon utilisable 4 mois après ?
2. les 3 personnes achètent un flacon non utilisable 4 mois après ?

Exercice n° 4

Une chaîne de fabrication produit 40000 fours dont 36000 sont bons, ne nécessitant donc aucune modification. Le service de contrôle-qualité qui ne connaît pas ces chiffres, prélève un échantillon aléatoire de 10 fours pour estimer la qualité de l'ensemble de la fabrication.
Identifier la variable aléatoire, sa loi et les paramètres de la loi.
Calculer les probabilités pour que l'on obtienne au contrôle :

1. 5 fours défectueux,
2. au moins 7 bons fours,
3. au plus 4 fours défectueux,
4. plus de 8 bons fours,
5. moins de 6 fours défectueux.

Exercice n° 5 :

A l'aide de la table statistique déterminer :

1 - $P(X = 3)$ si $L(X) = P(0,7)$
2 - $P(X < 4)$ si $L(X) = P(3.5)$
3 - $P(X > 5)$ si $L(X) = P(8.5)$
4 - $P(X \leq 6)$ si $L(X) = P(10)$
5 - $P(X \geq 7)$ si $L(X) = P(14)$
6 - $P(X = 0)$, $P(X = 1)$ et $P(X = 2)$ si $L(X) = P(10,5)$ par récurrence.

Exercice n° 6

Représenter la fonction de densité de fréquence de la variable X telle que L(X) = P(18).
Que remarquez-vous ? Calculer l'espérance mathématique, la variance et l'écart-type de la variable.

Exercice n° 7

Une chaîne de fabrication produit 20000 ampoules dont 1000 ont leur filament brisé, ne pouvant donc pas être vendues. Le service de contrôle-qualité qui ne connaît pas ces chiffres, prélève un échantillon aléatoire de 100 ampoules pour estimer la qualité de l'ensemble de la fabrication.
Identifier la variable aléatoire, sa loi et les paramètres de la loi.
Calculer les probabilités pour que l'on obtienne au contrôle :

1. 4 ampoules non conformes,
2. au moins 7 ampoules non conformes,
3. au plus 10 ampoules non conformes,
4. plus de 12 ampoules non conformes,
5. moins de 5% des ampoules non conformes.

Exercice n° 8

La commission hygiène et sécurité d'une usine a calculé qu'il y avait en moyenne 0,2 accident du travail par jour.
Identifier la variable aléatoire, sa loi et les paramètres de la loi.
Calculer la probabilité pour qu'il y ait :

1. 2 accidents dans le prochain jour,
2. plus de 2 accidents dans les 3 prochains jours,
3. moins de 3 accidents dans les 20 prochains jours de travail.

Exercice n° 9

Soit T une variable aléatoire normale centrée réduite. Déterminer à l'aide de la table et d'un schéma, les probabilités suivantes :

$P(0 < T < 0{,}6)$	$P(-1{,}.63 < T < -0{,}91)$
$P(1{,}23 < T < 2{,}34)$	$P(T \geq 1{,}64)$
$P(T < -2{,}19)$	$P(T \geq -2{,}31)$
$P(-1{,}56 < T < 1{,}75)$	$P(T \geq 0)$

Exercice n° 10

Déterminer t avec la notation t_p tel que :

$P(T < t) = 0{,}005$	$P(T > t) = 0{,}004$	$P(-t < T < +t) = 0{,}95$
$P(T < t) = 0{,}593$	$P(T > t) = 0{,}599$	
$P(T < t) = 0{,}5$	$P(T > t) = 0{,}6$	
$P(T < t) = 0{,}898$	$P(T > t) = 0{,}841$	
$P(T < t) = 0{,}182$	$P(T > t) = 0{,}187$	
$P(T < t) = 0{,}398$	$P(T > t) = 0{,}390$	

Exercice n° 11

Soit X une variable aléatoire normale observée sur une population de moyenne égale à 5000 et d'écart-type égal à 2000.
1°) Que valent moyenne et écart-type de cette variable aléatoire centrée et réduite notée T?

2°) Quelle est la valeur de la variable centrée et réduite t pour x = 7000?
3°) Calculer P (T< 1) et P (X < 7000).
4°) Calculer P (X<6000) et P (X < 3000).

Exercice n° 12

Un fabricant de machines a pu établir que la demande pour son modèle suit une loi normale de moyenne 2100 unités par mois et un écart-type de 250.
1) Si l'entreprise possède 2400 unités en stock pour le mois à venir quelle est la probabilité qu'elle ne puisse satisfaire à la demande ?
2) quel doit être le nombre d'unités à stocker par mois pour que l'entreprise ne soit pas en pénurie plus de 5% des cas.

Exercice n° 13

Une entreprise de TP précise qu'en général le temps moyen de construction d'un certain type de bâtiment est de 65 semaines avec un écart-type de 4 semaines. L'entreprise signe un contrat avec un client dans lequel elle propose un délai de réalisation des travaux qui a 90% de chance d'être respecté. Quel est ce délai si on suppose que cette variable est distribuée normalement?

Exercice n° 14

Le temps moyen pour compléter un QCM de 60 questions est estimé à 120 minutes avec un écart-type de 15 minutes. Pour le prochain examen il est question de modifier le temps de l'épreuve de telle sorte que 90% des étudiants puissent compléter le QCM. On suppose que le temps suit une loi normale, quelle doit être la durée maximale de l'épreuve?

Exercice n° 15

Dans une municipalité, 20% des foyers ont des impôts locaux inférieurs à 1000F et 60% des foyers ont des impôts locaux inférieurs à 4000F. En supposant que cette variable suit une loi normale, que valent la moyenne et l'écart-type de la distribution?

Exercice n° 16

Le temps moyen nécessaire à l'assemblage d'un appareil est de 15 minutes avec une variance de 0,81. En supposant que cette variable suive une loi normale,
1) quelle est la probabilité qu'un ouvrier assemble l'appareil en moins de 13 minutes?
2) dans 90% des cas, l'assemblage se fera en moins de combien de temps?
3) 50% des appareils sont assemblés en moins de combien de temps?

Exercice n° 17

Dans une région de France, le revenu annuel familial est distribué normalement avec une moyenne de 174 000F et un écart-type de 26 000F. Afin de mettre sur le marché un nouveau produit, le service marketing d'une entreprise pense qu'il est nécessaire qu'au moins 80% des foyers disposent d'un revenu annuel d'au moins 150 000F. Est-ce que la région correspond au marché visé?

Exercice n° 18

_A l'aide des tables déterminer:

$t_{0,95}(10)$ $t_{0,95}(15)$ $t_{0,90}(10)$ $t_{0,99}(8)$ $t_{0,95}(50)$ $t_{0,001}(20)$ =

$X^2_{0,95}(10)$ $X^2_{0,99}(10)$ $X^2_{0,90}(32)$ $X^2_{0,025}(18)$

$F_{0,975}$ (10 ; 20) $F_{0,975}$ (20 ; 10) $F_{0,995}$ (9 ; 15)

Exercice n° 19

La banque nationale désire introduire sur le marché une nouvelle formule de bons de caisse. Elle propose des bons portant un intérêt de 3% l'an si le détenteur place ses capitaux pendant 3 mois. Si, à la fin du premier mois, le détenteur lève une option portant à 6 mois son placement, l'intérêt passera alors à 3.5% l'an. Il semblerait après une étude sommaire qu'il y ait 15% de chance pour que ce type de transformation soit opéré. Soixante entreprises viennent souscrire à ce type de placement en achetant chacune 1 bon.
Identifier la variable aléatoire, sa loi et les paramètres de la loi.
Calculer la probabilité pour qu'il y ait sur les 60 entreprises :

1. plus de 8 entreprises qui effectueront la transformation,
2. moins de 12 entreprises qui effectueront la transformation,
3. quel est le nombre de bons transformés que l'on a que 5% de chance de dépasser ?

Exercice n° 20

L'observation microscopique de boîtes de Pétri contenant des quantités égales d'une solution a permis d'évaluer la valeur moyenne du nombre de bactéries par boîte égale à 2.
Identifier la variable aléatoire, sa loi et les paramètres de la loi.
Si l'étude porte sur l'observation de 10 boîtes:

1. calculer la probabilité pour qu'il y ait plus de 18 bactéries,
2. quel est le nombre de bactéries que l'on a que 5% de chance de dépasser ?

Exercice de TD
3 – Statistique bivariée

Pour les exercices 1 à 5 vous poserez les hypothèses nulle et alternative, vous établirez la distribution théorique, vous calculerez le critère statistique et son ddl. Vous formulerez la conclusion de 2 manières :

- en choisissant un risque d'erreur de 5%
- en encadrant la valeur du risque à l'aide de la table

Exercice n° 1 :

Un sondage a été réalisé sur l'adaptation à l'outil informatique en fonction de l'âge :

	niveau d'adaptation		
tranche d'âge	**très difficile**	**difficile**	**facile**
[18; 30 [	98	140	150
[30; 50 [	126	150	90
[50; 70 [	230	120	80

Peut-on conclure que le niveau d'adaptation est lié à l'âge des répondants?

Exercice n° 2 :

En 1959 une étude a été faite sur un échantillon de 350 personnes qui comportait 10 % de fumeurs et 20% de gens atteints du « cancer du fumeur ». Dans cet échantillon il y avait 23 individus fumeurs et atteints d'un cancer. Que concluez-vous?

Exercice n° 3 :

La division recherche d'une entreprise chimique a expérimenté 4 produits anti-mildiou sur 300 pieds de vigne situés en des lieux aléatoires. Quelques semaines plus tard, la proportion des pieds contaminés était réparties de la façon suivante :

	traitement 1	**traitement 2**	**traitement 3**	**traitement 4**
contaminés	6	6	8	7
non contaminés	12	31	16	14

Que concluez-vous?

Exercice n° 4 :

Le nombre de pannes d'un équipement est supposé suivre une loi de Poisson de moyenne égale à 0,4 panne par jour. Le service de contrôle qualité a vérifie l'équipement pendant 100 jours et a obtenu :

Nombre de pannes	**Nombre de jours**
0	45
1	36
2	14
3	4
4	1
Total	100

1°) peut-on considérer comme vraisemblable l'hypothèse selon laquelle le nombre de pannes suit une loi de Poisson de paramètre 0,4 ?

2°) peut-on considérer comme vraisemblable l'hypothèse selon laquelle le nombre de pannes suit une loi de Poisson ?

Exercice n° 5 :

Une enquête menée sur la lecture de 4 revues scientifiques a été menée auprès des laboratoires de recherche d'une ville universitaire. On voulait vérifier si elles répondaient toujours dans les mêmes proportions aux attentes des chercheurs. Que concluez-vous des résultats suivants :

type de revues	pourcentage théorique des préférences	nombre de chercheurs selon leur préférence
A	10	32
B	50	181
C	35	125
D	5	22

Exercice n°6 :

Soient le nombre d'heures de révision à un examen et la note obtenue à cet examen pour 6 élèves :

heures	16	13	19	17	10	6
note	17	12	18	15	9	5

Pour cet exercice vous établirez un tableau de détails des calculs dans lequel la dernière ligne sera constituée de :

Σxi Σxi^2 $\underline{SCEx(\bar{x})}$

Σyi Σyi^2 $\underline{SCEy(\bar{y})}$

$\Sigma xiyi$ $\Sigma[\ (xi - \bar{x})\ (yi - \bar{y}\ )]$,

$\sum \hat{y}i$ $\sum \hat{y}i^2$ $\underline{SCE\hat{y}(\bar{y})}$

Σei Σei^2 SCE ei ($\bar{e}$)

et la liste des résidus standardisés.

1°) Déterminer la variable explicative et la variable à expliquer et calculer la moyenne et l'écart-type de X et de Y

2°) Calculer à l'aide de la calculatrice et interpréter les coefficients de corrélation et de détermination

3°) Déterminer les coefficients de la régression à l'aide de la calculatrice

4°) Déterminer l'équation de la droite de régression de Y pour X fixé

5°) Tracer le nuage de points et la droite de régression, où se trouve le centre de gravité du nuage ?

6°) calculer à l'aide de la calculatrice sans saisir d'autres données que celles de X et Y:

SCEx SCEy SPExy cov(xy) SCE résiduelle

la variance résiduelle la variance expliquée par la régression

7°) Si on suppose qu'il existe un lien entre ces 2 variables, quel devrait être la note d'un élève qui révise 13 heures ? calculer le résidu correspondant.

Exercice n° 7 :

Les données suivantes représentent la mesure de la concentration en chlore d'échantillons prélevés à des semaines différentes:

X	Y	Y estimé	ei
8	49		0,52
10	48		0,47
11	47		-0,05
12	46		-0,57
14	44		-1,62
16	43		-1,67
18	45		1,28
20	42		-0,77
22	41		-0,81
24	40		-0,86
26	40		0,09
28	40		1,04
30	39		1,00
32	39		1,95
34	38		1,90
36	37		1,85
38	35		0,80
40	34		0,76
42	33		0,71
43	31		-0,82
44	29		-2,34
45	28		-2,86

Dans cet exercice vous analyserez les résidus. Vous utiliserez la variance résiduelle observée = SCEe / n pour faire les calculs des résidus standardisés (logiquement on utilise une variance estimée = SCEe / (n-2) mais nous le verrons en 2ème année).

Exercice n° 8 : (exercice supplémentaire de synthèse sur la régression linéaire)

Une étude porte sur 20 individus. Sur chaque individu on relève 2 variables quantitatives : note QCM d'un test de vocabulaire et moyenne de l'individu.
Les auteurs cherchent à expliquer des variations dans la réussite scolaire (évaluée par une moyenne générale) par des variations dans les taux d'acquisition de vocabulaire. Pour cela il faut utiliser la **méthode de régression linéaire simple et la corrélation**.
Les résultats sont représentés dans le tableau ci-dessous.

Pour analyser et interpréter ces résultats vous vous intéresserez successivement :
- à l'allure du nuage de points,
- à l'équation de la droite de régression de Dy/x,
- aux coefficients de détermination et de corrélation,
- aux coefficients de la régression,
- à l'analyse des résidus.

Elève	Note QCM	Moyenne générale
1	603	18,05
2	264	4,58
3	537	13,33
4	347	9,67
5	463	12,24
6	562	16,18
7	520	14,11
8	314	7,54
9	397	12,56
10	504	10,37
11	331	9,85
12	357	11,21
13	568	12,34
14	454	10,43
15	471	8,21
16	438	9,23
17	389	6,45
18	514	10,1
19	488	16,54
20	435	14,58

Exercice n° 9 : (exercice supplémentaire d'adéquation)

Un cabinet de communication est chargé de mettre en place un dispositif d'information juridique portant sur le harcèlement sexuel des femmes au travail. Avant de lancer une campagne nationale, ce cabinet décide de tester différents formats de communication et souhaite évaluer les éléments retenus par les femmes à qui ils sont présentés. Comme il s'agit d'un dispositif préliminaire nécessitant une disponibilité importante de la part des participantes, seules 250 personnes acceptent d'y prendre part jusqu'au bout. Comme il est important de bien contrôler le niveau d'instruction des participantes à ce pré-test, le cabinet souhaite vérifier si l'échantillon des 250 femmes possède les mêmes caractéristiques de formation scolaire que celles que l'on retrouve dans la population féminine générale. A l'aide d'une classification en 7 catégories de diplômes, on observe les répartitions suivantes[1] :

Diplômes	Pourcentage national	Effectif par échantillon
Aucun ou certificat d'étude	34%	72
BEPC seul	8%	24
Niveau CAP, BEP	25%	65
Niveau Bac, BP	11%	25
Niveau BTS, DUT	10%	29
2ème ou 3ème cycle universitaire	7%	19
Diplômes non déclarés	5%	16

L'échantillon est-il représentatif de la population ?

[1] *G.Mermet, Francoscopie 1993, Paris, Larousse, 1992, p.100.*

TD de Stat 2 (1)

Exercice 1

1 $P(X=3) = 0,0081$

2 $P(X\leqslant 3) = 0,9601$

3 $P(X>5) = 1-P(X\leqslant 5) = 1-0,9981 = 0,0019$

4 $P(X\leqslant 6) = 0,9918$

5 $P(X\geqslant 7) = 1-P(X\leqslant 6) = 1-0,9909 = 0,0091$

6 Récurrence (initiation) : $P(X=k) = \binom{n}{k} p^k q^{n-k}$

$$P(X=0) = \binom{7}{0} 0,2^0 (1-0,2)^{7-0} = \underline{\underline{0,21}}$$

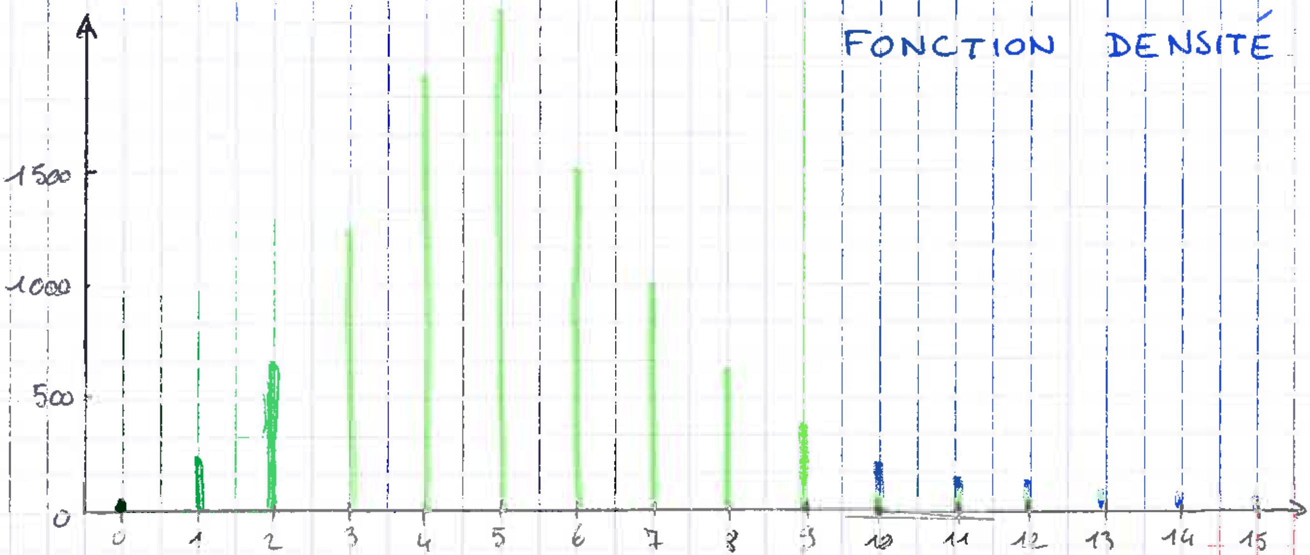
Récurrence (binomiale) : $P(X=k) = \frac{n-k+1}{k} \frac{p}{q} P(X=k-1)$

$$P(X=1) = \frac{7}{1} \frac{0,2}{0,8} P(0) = 7\times 0,25 \times 0,21 = \underline{\underline{0,37}}$$

$$P(X=2) = \frac{7-2+1}{2} \frac{0,2}{0,8} P(1) = 3\times 0,25\times 0,37 = \underline{\underline{0,28}}$$

Exercice 2

$\mathcal{L}(X) = B(50\,;0,1)$



On observe une distribution unimodale

$E(X) = np = 5$ $\qquad V(X) = npq = 4,5$ $\qquad \sigma = 2,12$

Comme $E(X) \simeq V(X)$, cette loi binômiale tend vers une loi poisson.

Exercice 3

Variable : X, le flacon qui périme

$n = 3$

$p = \frac{50}{200} = 0{,}25$

On a n évenements à 2 issues (p ou q)
n n'est pas grand ; p n'est pas petit → loi binômiale (pas poisson)

1. $P(X=1) = \binom{3}{1} p^1 q^{3-1} = \frac{3!}{1!\,2!} \times 0{,}25 \times 0{,}75^2 = \boxed{0{,}4219}$
2. $P(X=3) = \binom{3}{3} p^3 q^0 = p^3 = \boxed{0{,}0156}$

Exercice 4

Variable : X, le four fonctionne

$n = 10$

$p = \frac{36}{40} = 0{,}9$

$np = 9 \neq npq = 0{,}9$ so la loi est binômiale.

1. $P(5) = \binom{10}{5} 0{,}9^5\, 0{,}1^5 = \boxed{0{,}0015}$
2. $P(X \geqslant 7) = P(7) + P(8) + P(9) + P(10)$
   $= \binom{10}{7} 0{,}9^7 0{,}1^3 [\ldots] = \boxed{0{,}9872}$
3. $P(X \geqslant 6) = P(X \geqslant 7) + P(X=6) = [\ldots] = 0{,}9984$
4. $P(X > 8) = P(X=9) + P(X=10) = [\ldots] = 0{,}7361$
5. $P(X \geqslant 4) = P(X \geqslant 6) + P(X=5) + P(X=4)$
   $= 0{,}9984 + 0{,}0015 + \binom{10}{4} p^4 q^6 = 0{,}9984 + 0{,}0015 + \frac{10!}{4!\,6!}\, 0{,}9^4 \times 0{,}1^6$
   $= > 1$ WTF ?! $1{,}00009984\ldots$

Exercice

TD de Stat 2 (2)

Exercice 9

$P(0<T<0,6) = -0,5 + P(T<0,6) = 0,2257$

$P(1,23<T<2,34) = P(T<2,34) - P(T<1,23) = 0,9904 - 0,8907 = 0,097$

$P(T<-2,19) = 1 - P(T<2,19) = 1 - 0,9857 = 0,0143$

$P(-1,56<T<1,75) = P(T<1,75) - 1 + P(T<1,56) = 0,9599 - 1 + 0,9406 = 0,9005$

$P(-1,63<T<-0,91) = P(0,91<T<1,63) = P(T<1,63) - P(T<0,91) = 0,9484 - 0,8186$
$= 0,1298$

$P(T \geq 1,64) = 1 - P(T<1,64) = 1 - 0,9495 = 0,0505$

$P(T \geq -2,31) = P(T<2,31) = 0,9896$

$P(T \geq 0) = 0,5$

Exercice 10

$P(T<t)$	0,005	0,593	0,5	0,898	0,182	0,398	0,004	0,599	0,6	0,841	0,187	0,390
t_p	-2,576	0,235	0	1,270	-0,908	-0,259	-2,652	0,251	0,253			

$P(T>t)$	0,004	0,599	0,6	0,841	0,187	0,390	...	0,390	0,°
t_p	2,652	-0,251	-0,253	-0,999	0,889	0,279			

Regarder la correction

$P(-t<T<t) = 0,95 = P(T<t) - (1 - P(T<t)) = 2P(T<t) = 0,95$

donc $P(T<t) = 0,475$

alors $t = 0,063$

Exercice 11

1) moyenne : 5000
écart type : 2000

$T = \frac{X - 5000}{2000} = \frac{7000 - 5000}{2000} = 1$

2) $P(T<1) = 0,8413$ (voir la table) Rmq : $P(T<1) = P(X<7000)$

3) $P(X<6000) = P(T<0,5)$ car $T = \frac{6000-5000}{2000} = 0,5$

$= 0,6915$

De même $P(X<3000) = P(T<-1) = 1 - P(T<1) = 0,1587$

Exercice 12

1) $\tilde{x}$ et $\sigma = 250$ donc $t = \frac{2400-2100}{250} = 1,2$

$P(T<1,2) = 0,8849$ donc $P(T>1,2) = 0,1151$

2) $P(T>t) = 0,05$ alors $t = 1,645$

$x = 250 \cdot 1,645 + 2100$
$x = 2511$

1.20- -16

Exercice 7

$X =$ nombre d'ampoules cassées $\qquad p = \frac{1\,000}{20\,000} = 0{,}05$

$X = 0, 1, 2 \ldots 99, 100.$

$L(X) = B(100;\ 0{,}05)$

approximée par une loi de poisson $\quad m = np = 100 \times 0{,}05 = 5$

Donc $L(X) = P(5)$

1. $P(X=4) = 0{,}1755$
2. $P(X \geq 7) = 1 - P(X < 7) = 1 - P(X \leq 7) - P(X=7) = 1 - (0{,}8666 - 0{,}1044) = 0{,}2378$
3. $P(X \leq 10) = 0{,}9863$
4. $P(X > 12) = 1 - P(X \leq 12) = 1 - 0{,}9980 = 0{,}0020$
5. $P(X < 5) = 0{,}4405$

Exercice 8

$X =$ nbre d'accidents en 1 jour $\qquad X = 0, 1, 2 \ldots +\infty$

$L(X) = P(0{,}2)$

1. $P(X=2) = 0{,}0164$
2. $Y = 3X \qquad L(Y) = P(0{,}6) \qquad P(Y > 2) = 1 - P(Y \leq 2) = 0{,}0231$
3. $Z = 20X \qquad L(Z) = P(4) \qquad P(Z < 3) = P(Z \leq 3) - P(Z = 3) = 0{,}2381$

Exercice 13

$X =$ tps pour built $\qquad L(X) = N(65;\ 4)$

$P(X < x) = 0{,}9$ donc $P(T < t) = 0{,}9$

Avec la table $t = 1{,}282$

$1{,}282 = \frac{x - 65}{4}$

$x = 4 \times 1{,}282 + 64$

$x = \cancel{70}{,}128$

Ce délai est de 70 semaines.

Exercice 14

$X =$ tps pour compléter un QCM $\qquad L(X) = N(120, 15)$

$P(X < x) = 0{,}9$ donc $P(T < t) = 0{,}9$

Avec la table $t = 1{,}282$

$1{,}282 = \frac{x - 120}{15}$ $\qquad x = 139{,}23$

$\Rightarrow 140$

Exercice 15

Avec la table normale $\left| \begin{array}{l} 0{,}2 \rightarrow 0{,}842 \\ 0{,}6 \rightarrow 0{,}253 \end{array} \right.$ $\qquad$ soit la loi telle que $L(X) = N(r, a)$

$0{,}842 = \frac{1000 - r}{a}$

$0{,}253 = \frac{400 - r}{a}$

$\Rightarrow r =$

12/04/11

STATISTIQUES _ TD3

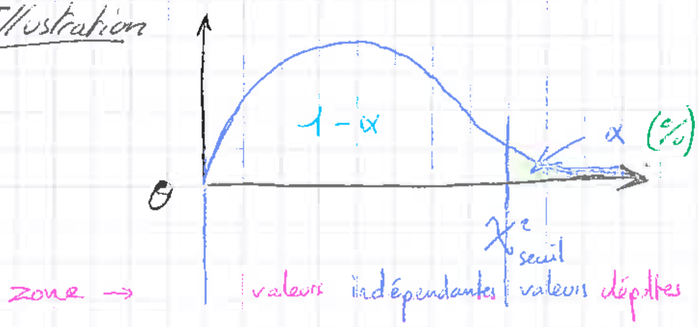
Effectif théorique de valeurs indépendantes $= \dfrac{\text{total colonne} \times \text{total ligne}}{\text{TOTAL}} = T_i$

→ $\chi^2_{\text{calculé}} = \sum \dfrac{(O_i - T_i)^2}{T_i}$ avec O_i les valeurs observées

→ $\chi^2_{\text{seuil}} \rightarrow \nu = (\text{nbre de lignes} - 1)(\text{nbre de colonnes} - 1)$ cette formule du degré de liberté permet d'utiliser la table χ^2

⇒ Si $\chi^2_{\text{calculé}} \gg \chi^2_{\text{seuil}}$, alors les variables ne sont pas indépendantes.

Illustration



Exercice 1

Step 1 → calculer les T_i

98	140	150	388
126	150	90	366
230	120	80	430
454	410	320	1184

T_i :

148,8	134,4	104,9
140,3	126,7	98,9
164,9	148,9	116,2

→ χ^2_{calc} :

17,3	0,23	19,4
1,46	4,28	0,8
25,7	5,6	11,3

→ en ajoutant toutes les valeurs

$\chi^2_{\text{calc}} = 86,07$

χ^2_{seuil} ... $\nu = 4$ $p = 1 - \alpha = 0,95$

Avec la table : $\chi^2_{\text{seuil}} = 9,49$

$\chi^2_{\text{calc}} \ggg \chi^2_{\text{seuil}}$

Donc les variables sont dépendantes.

Exercice 4

X le nbre de pannes, la variable
n l'effectif (nbre de jours)

$$\lambda = \frac{36 + 2\times 14 + 3\times 4 + 4\times 1}{100} = 0{,}8$$

X	n	Valeurs de la table ∞	eff théorique	χ^2_i
0	45	0,6703	67,03	7,24
1	36	0,2681	26,81	3,15
2	14	0,0536	5,36	13,9
3	4	0,0072	0,72	~~14,9~~
4	1	0,0007	0,08	~~10,58~~
Total 4	100			

à regrouper → 5 | 0,8

$\chi^2_{calc} = \cancel{52,95}$

Correction X, la variable est le nbre de pannes par jour
$X \sim \mathcal{P}(\lambda = 0{,}4)$

Nb pannes Nb jours P Nb pannes théoriques

Les effectifs théoriques doivent être > 5 → on regroupe les 2 dernières lignes.

$\chi^2_{calc} = 46$ $\chi^2_{seuil} = 7{,}81$

ν = nb comparaisons − 1 − nb paramètres estimés = 4 − 1 − 0 = 3

pas estimé, donné

$\chi^2_{calc} \ggg \chi^2_{seuil}$

STAT TD 3

Exercice 1

98
126
230

1) → Tableau des résultats théoriques :

148,8	134,4	104,9	388
140,3	126,7	98,9	366
164,9	148,9	116,2	430
454	410	320	1184

2) → Calcul du χ^2 :

$$\left.\begin{matrix} 17,3 + 0,23 + 19,4 \\ 1,46 + 4,28 + 0,8 \\ 23,7 + 5,6 + 11,2 \end{matrix}\right\} \Sigma = \chi^2_{calc} = 86,07$$

3) → Recherche dans la table de χ^2_{seuil} à $\alpha = 5\%$ soit $p = 0,95$

et $ddl = \nu = (3-1)(3-1) = 4$

$$\chi^2_{seuil} = 9,49$$

$\chi^2_{calc} \gg \chi^2_{seuil}$ → les variables sont dépendantes

Stat TD3

Exercice 6

L'heure → x
La note → y

$\bar{x} = 13,5$ $\qquad s^2_x = 19,58$
$\bar{y} = 12,67$ $\qquad s^2_y = 20,89$

$$s_{xy} = \frac{1}{2} \Sigma (x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{n} \Sigma x_i y_i - \bar{x}\bar{y} = 19,83$$

$a = \frac{s_{xy}}{s^2_x} = 1,013$ $\qquad b = \bar{y} - a\bar{x} = -1,006$ $\qquad \hat{y} = ax + b$

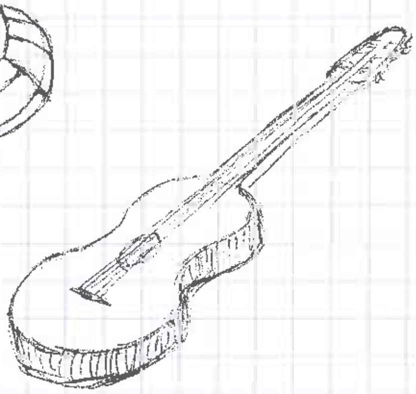
$r = \frac{s_{xy}}{s_x s_y} = 0,98$ $\qquad R^2 = r \times r = 0,96$ $\qquad \varepsilon_i = y_i - \hat{y}_i$

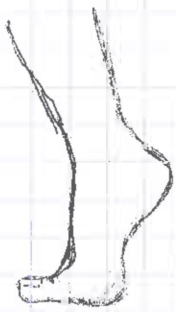
compris entre 1 et -1
coef de corrélation
si > 0, les 2 variables augmentent ensemble

exprime la part de variation du modèle
la droite explique 96% des variations

NB : $\bar{\varepsilon}$ est nulle, par construction

$s_\varepsilon^2 = \frac{1}{n} \sum \varepsilon_i^2 = 0,8$





6

ANOVA

① → Contexte

- donner le modèle   exemple : $x_{ij} = \mu + \alpha_i + \beta_j + \alpha\beta_{ij} + \varepsilon_{ij}$
- définir la réponse étudiée, le nb de facteurs, de modalités, de répétitions, de traitements et d'unités statistiques

② → Etude des résidus

* Suivent-ils une loi normale ? (= pas de facteurs oubliés ?)

- tableau des résidus
- histogramme → si il est symétrique et en cloche, l'hypothèse de normalité n'est pas refutée
- coef. de Pearson
  - ↳ asymétrie : $\beta_1 \simeq 0$ → plutôt symétrique
    p-value → risque d'avoir tort en disant "c'est pas symétrique"
  - ↳ aplatissement : $\beta_2 \simeq 3$ → plutôt mésocurtique
    p-value → risque d'avoir tort en disant "c'est pas en cloche"

* Sont-ils bien aléatoires ? → cartographie

* Les facteurs ne font-ils bien varier que la moyenne ?

↳ Bartlett (hate you) me too ! 1 TEST = 1 facteur

modalité	eff. n_j	moy. m_j	SCE	ν_j *	$\hat{\sigma}_j^2$ *2	$\frac{1}{\nu_j}$	$\nu_j \log \hat{\sigma}_j^2$

$\chi^2_{\nu\, calc}$ → formule donnée

$\chi^2_{\nu\, calc} < \chi^2_{\nu,\, théo}$ → les variances sont homogènes (H_0)

↳ si 2 modalités seulement → test F

H_0

$$F_{calc} = \frac{\hat{\sigma}_1^2}{\hat{\sigma}_2^2} \quad \text{avec } \hat{\sigma}_1^2 > \hat{\sigma}_2^2$$

$F_{calc} < F_{théo}$ → les variances sont homogènes (H_0)

* $\nu_j = n_j - 1$ *2 $\hat{\sigma}_j^2 = \frac{SCE_j}{\nu_j}$

③ → ANOVA (étude du facteur)

Source variaᵒ	SCE	ddl	CM	F_{calc}	$F_{théo}$ → ($\nu_1 = ddl$) $ddl_e = ddl_e$!!
⇒ Total	①	④	⑧	⑨	
→ fac. A	②	⑤	⑧	⑨	
→ fac. B	②	⑤	⑧	⑨	
→ fac. AB	②	⑥	⑧	⑨	
→ résid.	③	⑦	⑧		

② $SCE = \left(\frac{(\Sigma \text{ modalités } 1)^2}{\text{effec.}} + \frac{(\Sigma \text{ modalités } 2)^2}{\text{effect.}} \ldots\right) - \mathbb{C} \quad (- SCE_A - SCE_B)$ ← interaction

① $\Sigma(\text{chaq. résultat}^2) - \mathbb{C} = SCE_{tot}$

$\mathbb{C} = \frac{(\text{somme de tout})^2}{\text{effectif total}}$

③ $SCE_r = SCE_{tot} - \Sigma SCE$

④ ddl_{tot} = nb unités stats − 1 ⑤ ddl_{fac} = nb modalités − 1

⑥ $ddl_{AB} = (ddl_A) \times (ddl_B)$ ⑦ $ddl_e = ddl_{tot}$ − autres ddl

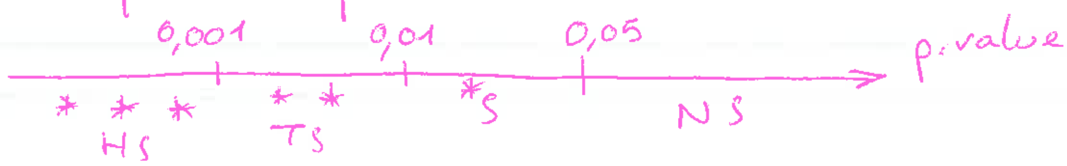
⑧ $CM = \frac{SCE}{ddl}$ ⑨ $F = \frac{CM}{CM_e}$

④ → PPDS

- ordonner les moyennes en ordre croissant
- les comparer 2 à 2 avec ∘ $ppds = t_{1-\frac{\alpha}{2}} \sqrt{\frac{CM_e}{n_{i1}} + \frac{CM_e}{n_{i2}}}$ ($\nu = ddl_e$)
- ∘ et leur différence de moyenne

si diff < ppds → les 2 sont du même groupe

Point p-value (
↳ on regarde le critère théo avec des % à taton, pour situer celui qui correspond au calculé



Simplex de départ de type 2

k =	1	2	3	4	...	k
essai 1	$-\frac{1}{2}$	$-\frac{1}{2\sqrt{3}}$	$-\frac{1}{2\sqrt{6}}$	$-\frac{1}{2\sqrt{10}}$	.....	$-\frac{1}{\sqrt{2k(k+1)}}$
essai 2	$\frac{1}{2}$	$-\frac{1}{2\sqrt{3}}$	$-\frac{1}{2\sqrt{6}}$	$-\frac{1}{2\sqrt{10}}$	.....	$-\frac{1}{\sqrt{2k(k+1)}}$
essai 3	0	$\frac{2}{2\sqrt{3}}$	$-\frac{1}{2\sqrt{6}}$	$-\frac{1}{2\sqrt{10}}$	.....	$-\frac{1}{\sqrt{2k(k+1)}}$
essai 4	0	0	$\frac{3}{2\sqrt{6}}$	$-\frac{1}{2\sqrt{10}}$	.....	$-\frac{1}{\sqrt{2k(k+1)}}$
....	.....	....	....	.....	.....	....
essai k+1	0	0	0	0	.....	$\frac{k}{\sqrt{2k(k+1)}}$

Simplex de départ de type 3

k	1	2	3	4	5	6
essai 1	0	0	0	0	0	0
essai 2	1	0	0	0	0	0
essai 3	0.5	0.866	0	0	0	0
essai 4	0.5	0.2887	0.8165	0	0	0
essai 5	0.5	0.2887	0.2041	0.7906	0	0
essai 6	0.5	0.2887	0.2041	0.1581	0.7746	0
essai 7	0.5	0.2887	0.2041	0.1581	0.1291	0.7638