QU'EST-CE QUE LA STATISTIQUE ?

Les statistiques sont définies comme étant "l'ensemble des renseignements numériques découlant des recensements de population, des données de registres d'état-civil et d'enquêtes appropriées" .

La statistique peut être définie comme "l'étude des ensembles numériques et de leurs relations", il s'agit de " la science qui a pour objet le groupement méthodique des faits qui se prêtent à une évaluation numérique ", " la méthode d'analyse et d'élaboration scientifique de ces faits ."
La statistique consiste donc en l'utilisation de méthodes et techniques visant au rassemblement, à la présentation et à l'analyse de données en vue de prises de décisions.

QUELQUES DEFINITIONS

Population :
désigne un ensemble fini ou infini d'éléments , on parle de population-mère quand on lui prélève un échantillon.

Echantillon :
ensemble quelconque de n éléments prélevé dans une population-mère de N éléments.

Sondage :
prélèvement d'un échantillon dans une population-mère.

variable statistique :
désigne, en statistique descriptive, l'ensemble des observations disponibles d'une variable quantitative (ou qualitative)

Modalités ou Variantes :
désigne l'une des occurrences exclusives qui définissent une variable qualitative ou quantitative.

Probabilité :
nombre réel compris entre 0 et 1 qui, associé à un événement, mesure les chances de réalisation de cet événement au cours d'une épreuve donnée. Si cette épreuve consiste en un prélèvement d'un élément dans cet ensemble, la probabilité est égale à la fréquence des éléments susceptibles de donner naissance à cet événement au cours de cette épreuve.

Modèle statistique:
désigne une loi de probabilité qui semble bien résumer des observations faites dans le monde réel et dont on utilisera les propriétés analytiques à des fins de prévisions pour guider l'action.

1ère partie : STATISTIQUE A UNE VARIABLE

Une variable peut être classée selon sa nature, il y a 2 groupes:

↳une variable de nature qualitative, subdivisée en :
qualitative nominale et qualitative ordinale.
↳une variable de nature quantitative, subdivisée en :
quantitative discrète et quantitative continue.

I - Synthèse par les tableaux

série statistique d'une variable de nature quantitative :
↳ série des données brutes : $\{x_1, x_2, \ldots xi, \ldots x_n\}$
↳ série des effectifs $\{(x_1 ; n_1) (x_2 ; n_2) \ldots.(x_j ; n_j) \ldots\ldots.(x_k ; n_k)\}$
↳ série des fréquences $\{(x_1 ; f_1) (x_2 ; f_2) \ldots.(x_j ; f_j) \ldots\ldots.(x_k ; f_k)\}$
↳ série groupée en classes $\{(x'_1 ; n'_1) (x'_2 ; n'_2) \ldots.(x'_j ; n'_j) \ldots\ldots.(x'_k ; n'_k)\}$

tableau:

	x_i	n_j	f_j	$n^+ (\leq x_i)$	$f^+(\leq x_i)$	$n^- (> x_i)$	$f^-(> x_i)$
	x_1	n_1	f_1	n_1	f_1	$n - n_1$	$1 - f_1$
	x_2	n_2	f_2	$n_1 + n_2$	$f_1 + f_2$	$n - n_1 - n_2$	$1 - f_1 - f_2$
	……..	……..	……..	……..	……..	……..	……..
	x_i	n_j	n_j	$n_1 + n_2 + \ldots + n_j$	$f_1 + f_2 + \ldots + f_j$	$n - n_1 - n_2 - \ldots. n_j$	$1 - f_1 - f_2 - \ldots - f_j$
	……..	……..	……..	……..	……..	……..	……..
	x_k	n_k	f_k	n	**1**	**0**	**0**
total		n					

II - Synthèse par les graphes

II – 1 - Variable qualitative

CSP	nj	fj
1	78	0,2806
2	54	0,1942
3	104	0,3741
4	42	0,1511
total	278	1

* CSP : catégorie socio professionnelle

diagramme à **bandes**

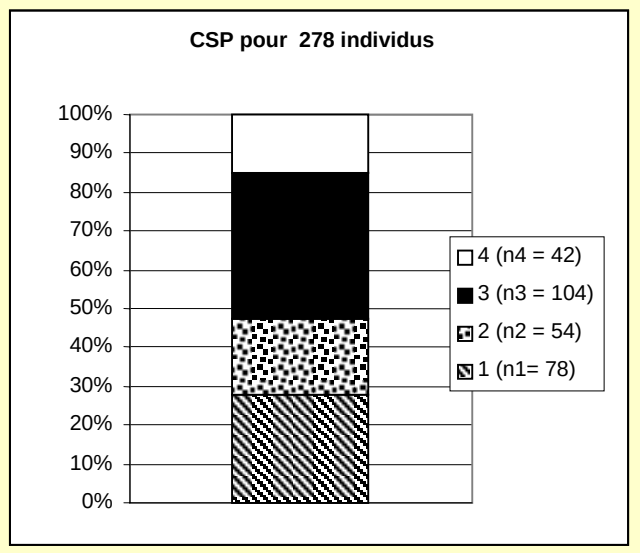


diagramme en **tuyaux d'orgue** (hauteur proportionnelle à l'effectif)

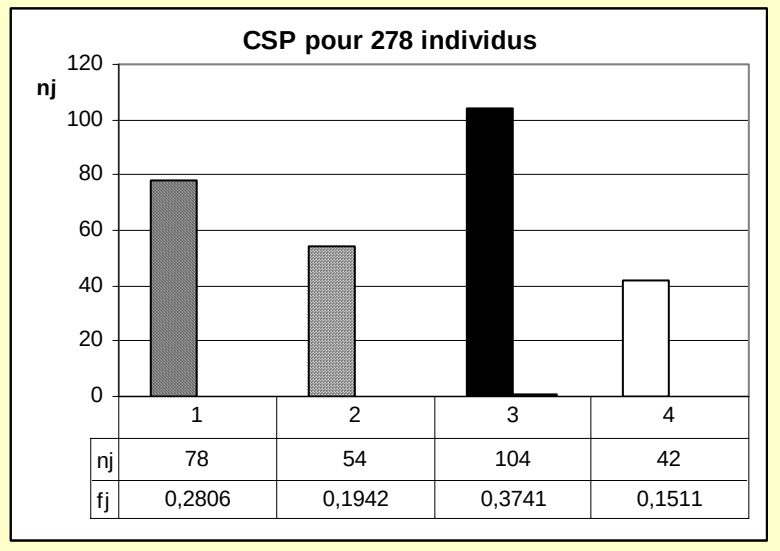
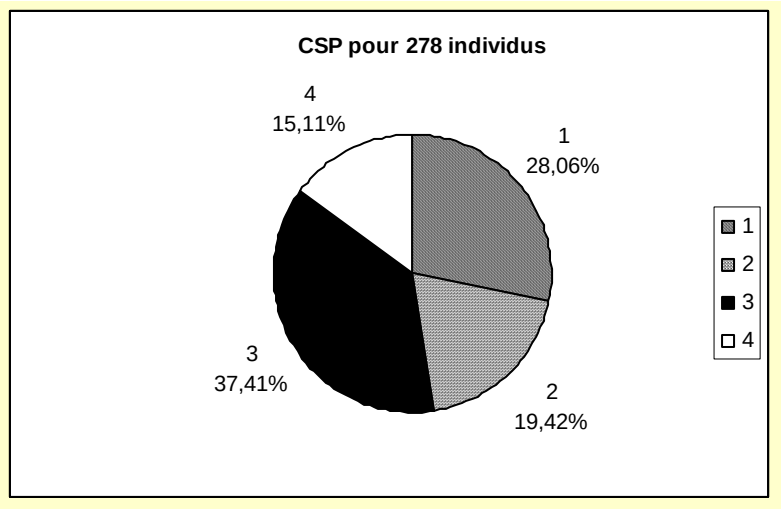


diagramme en **secteurs** (angle proportionnel à l'effectif)

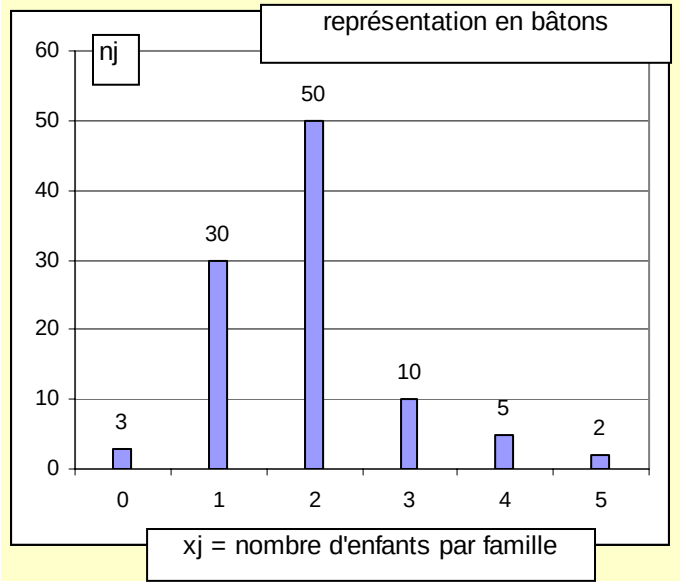


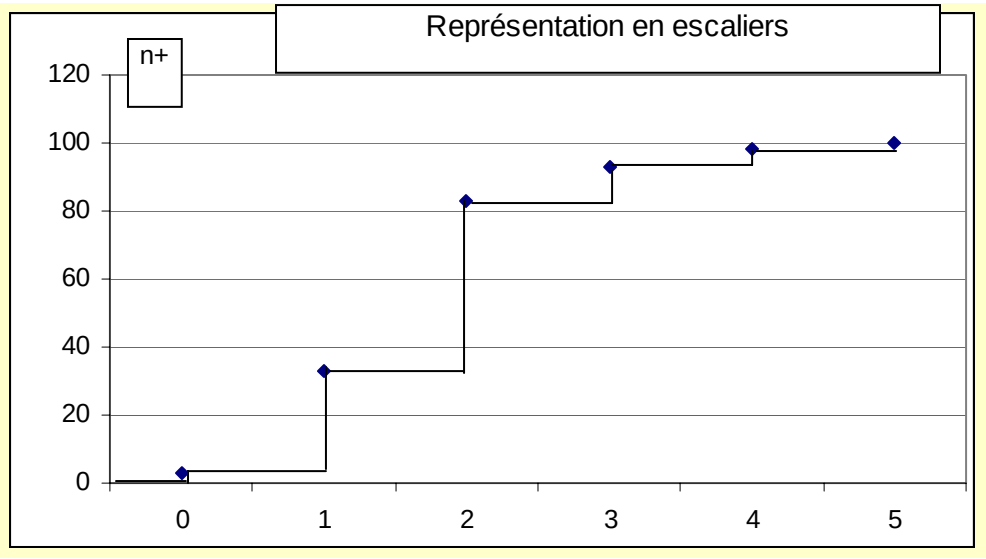
Ou encore un diagramme **figuratif**.

II – 2 - Variable quantitative : fonction de densité de fréquences et fonction de répartition.

↳ **caractère discret** : représentation en bâtons et représentation en escaliers.

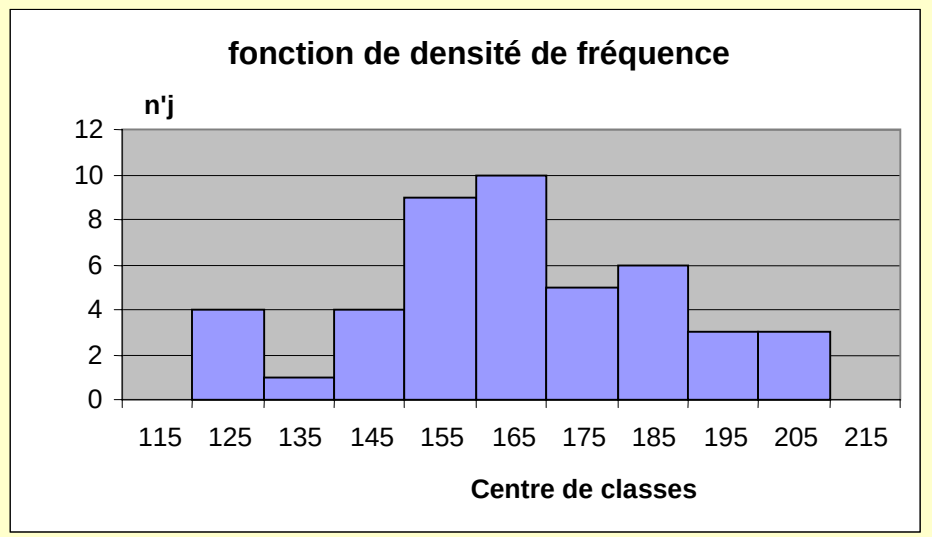
nombre d'enfants par famille	nombre de familles
0	3
1	30
2	50
3	10
4	5
5	2

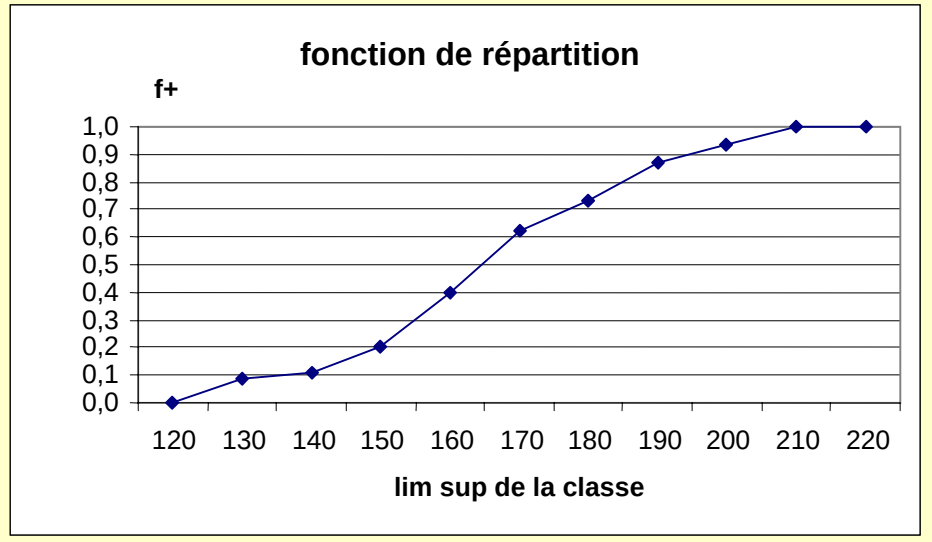




➥ **caractère continu** : nuage de points, représentation en histogramme (fonction de densité de fréquence), polygone statistique et représentation en sigmoïde (fonction de répartition).

détermination du nombre de classes (entre 7 et 12)
détermination de l'amplitude de classes (valeur plausible)
détermination des limites de classes (éviter les valeurs « sensibles »)





III - Synthèse par les paramètres

III – 1 - Valeurs caractéristiques de position

La représentation graphique d'une série statistique en fournit un premier résumé. Il faut maintenant caractériser cette série par un nombre représentatif de l'ordre de grandeur des valeurs de la série, afin qu'ultérieurement la comparaison de 2 séries puisse revenir à la comparaison de 2 nombres.
Cette caractéristique doivent obéir à plusieurs conditions définies par Yule, elle doit :

- Etre définie de manière objective,
- Dépendre de toutes les valeurs de la série,
- Avoir une signification concrète,
- Etre simple à calculer,
- Etre peu sensible aux fluctuations d'échantillonnage,
- Se prêter facilement aux calculs algébriques ultérieurs.

mode mo

C'est la valeur dominante, elle correspond au plus grand effectif. Une série peut être unimodale (1 mode), bimodale ou plurimodale. Son utilisation s'applique particulièrement aux caractères discrets.

médiane me ou $\tilde{x}$

Sur la série randée en ordre croissant, on désigne sous le nom de médiane d'une série la valeur (de la série ou non) telle que les nombres d'observations inférieures ou supérieures à m_e soient égaux.
Elle tient compte de toutes les valeurs de la série mais elle ne se prête pas facilement aux calculs ultérieurs.

!!!!!Attention la détermination de la médiane n'utilise pas la même méthode selon que la variable est de nature quantitative discète ou continue !!!!!

- Variable discrète :

  S'il y a un nombre impair de terme dans la série n = 2*p + 1 la médiane est le terme de rang (m_e) = p+1.

  S'il y a un nombre pair de terme dans la série n = 2*p la médiane est la moyenne entre les termes de rang p et p+1.

- Variable continue (dans les groupements en classes) :

  On admet qu'il existe une distribution uniforme des données à l'intérieur de chaque classe, on détermine la médiane par inntrapolation à l'aide du théorème de Thalès, le rang de la médiane étant égal à la (m_e) = 0.5*(n+1)^{ème} valeur.

moyenne $\bar{x}$ **ou m**

Elle tient compte de toutes les valeurs de la série, elle correspond au centre de gravité de la série, elle se prête facilement aux calculs ultérieurs mais elle est sensible aux valeurs extrêmes.

pour une série **brute** :

$$\bar{x} = \frac{\sum_{i=1}^{i=n} xi}{n}$$

pour une série groupée en **effectifs** :

$$\bar{x} = \frac{\sum_{j=1}^{j=k} n_j x_j}{n}$$

pour une série groupée en k **classes** :

$$\bar{x}' = \frac{\sum_{j=1}^{j=k} n'_j x'_j}{n}$$

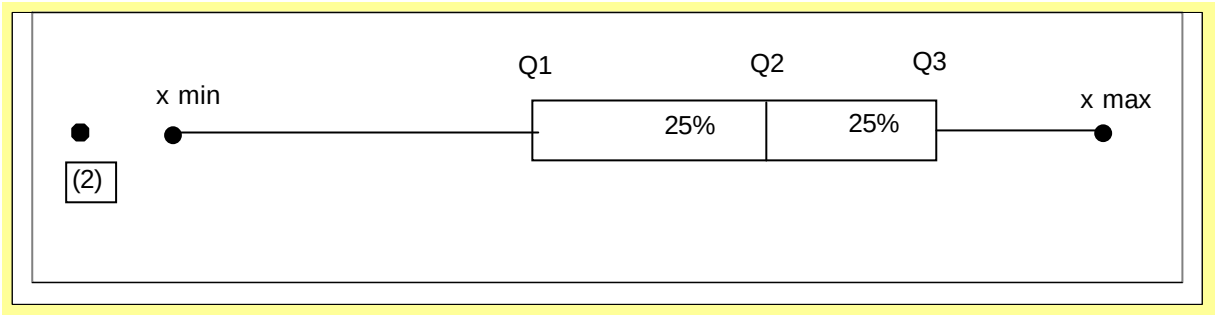
quantiles (Terciles, Quartiles, Déciles, Centiles, Milliles)

On partage une série rangée dans l'ordre croissant en :

- 2 sous série de même effectif à l'aide de la médiane
- 3 sous série de même effectif à l'aide de 2 terciles : T_1 et T_2
- 4 sous série de même effectif à l'aide de 3 quartiles : Q_1, Q_2 et Q_3
- 10 sous série de même effectif à l'aide de 9 déciles : D_1 à D_9
- 100 sous série de même effectif à l'aide de 99 centiles : C_1 à C_{99}
- 1000 sous série de même effectif à l'aide de 999 milliles : M_1 à M_{999}

Utilisation des quartiles : représentation en **Box-Plot** (Boîte à moustaches ; boîte de Tukey) et détermination des données «out»: celles qui sont inférieures à la limite inférieure calculée et celles qui sont supérieures à la limite supérieure calculée

- Intervalle Inter Quartile IIQ = Q3 - Q1
- lim inf = Q_1 - 1.5 IIQ
- lim sup = Q_3 + 1.5 IIQ
- x min = valeur de la série immédiatement supérieure à la limite inférieure
- x max = valeur de la série immédiatement inférieure à la limite supérieure



III – 2 - Valeurs caractéristiques de dispersion

Deux séries peuvent avoir les mêmes caractéristiques de position avec des dispersions très différentes des valeurs.

étendue R

R = x max – x min

intervalle inter quartile IIQ

IIQ = Q_3 - Q_1

On trouve 50% des valeurs dans cet intervalle.

variance $s^2(x)$

Elle représente le moment d'inertie de la série.

Série **brute**

$$SCE_x(\bar{x}) = SCEx = SCE = \sum_{i=1}^{n}(x_i - \bar{x})^2$$

formule de définition de la variance : $s^2(x) = SCE\ x / n$

formule du calcul manuel : $s^2x = \frac{\sum_{i=1}^{n} x_i^2}{n} - \bar{x}^2$

Série groupée en **effectifs**

$$SCE\ x\ (\bar{x}) = SCE\ x = \sum_{j=1}^{k} n_j\ (x_j - \bar{x})^2$$

formule de définition de la variance : $s^2(x) = SCE\ x / n$

formule du calcul manuel : $s^2(x) = (\sum_{j=1}^{k} n_j x_j^2 / n) - \bar{x}^2$

$$n = \sum_{j=1}^{k} n_j$$

Série groupée en k **classes**

$$SCE\ x\ (\bar{x}') = SCE\ x = \sum_{j=1}^{k} n'_j\ (x'_j - \bar{x}')^2$$

formule de définition de la variance : $s^2(x) = SCE\ x / n$

formule du calcul manuel : $s^2(x) = (\sum_{j=1}^{k} n'_j x'^2_j / n) - \bar{x}'^2$

$$n = \sum_{j=1}^{k} n_j$$

écart-type s(x)

Il est dans l'unité de mesure de la série.

coefficient de variation CV (coefficient de dispersion relative)

Le coefficient de variation ne dépend pas des unités de mesure, il permet donc de comparer les dispersions de 2 séries qui diffèrent par leurs unité ou par leur nature.
On considère généralement qu'une série est homogène autour de sa moyenne si CV < 10% ou 12%

$$CV = 100 * \frac{s_x}{\bar{x}}$$

.

Changements d'origine et d'unité (y) et variable centrée réduite (t)

yi = bxi + a	$\bar{y} = b\,\bar{x} + a$	$s^2(y) = b^2 . s^2(x)$
ti = $\frac{(x_i - \bar{x})}{s_x}$	$\bar{t} = 0$	$s^2(t) = 1$

Les caractéristiques de position sont sensibles aux changement d'origine et d'unité.
Les caractéristiques de dispersion ne sont sensibles qu'au changement d'unité.

Calculs de la moyenne et de la variance du mélange de l séries.

N°	Effectif	Moyenne	Variance
1	n_1	m_1	s_1^2
2	n_2	m_2	s_2^2
…	…	…	…
…	…	…	…
k	n_k	m_k	s^2_k
…	…	…	…
…	…	…	…
l	n_l	m_l	s^2_l

La moyenne générale $\bar{X}$ est égale à la moyenne des moyennes pondérées :

$$N = \sum_{k=1}^{l} n_k \qquad \bar{X} = \frac{\sum_{k=1}^{l} n_k m_k}{N}$$

La variance totale est égale à la somme des variances intragroupe et intergroupe :

$$S^2(x) = \frac{\sum_{k=1}^{l} n_k s_k{}^2}{N} + \frac{\sum_{k=1}^{l} n_k (m_k - \bar{X})^2}{N}$$

moyenne de la série des variances pondérées + variance de la série des moyennes pondérées

III – 3 - Valeurs caractéristiques de forme

Une distribution est dite **symétrique** si les observations sont également dispersées de part et d'autre de la valeur centrale. Dans le cas contraire, la distribution est dite asymétrique ou oblique (à droite ou à gauche).
Une distribution est plus ou moins **aplatie** suivant que les observations correspondant à un faible écart à la valeur centrale sont en plus ou moins grande proportion.

On a alors intérêt à caractériser la symétrie ou la dissymétrie et l'aplatissement de la distribution au moyen de nombres indépendants des unités de mesure.

Si mode = moyenne = médiane, la distribution suit une loi normale, la courbe de fréquence est la courbe en cloche de Gauss.

La dissymétrie

Mesurée par s, si s = 0 la distribution est symétrique, plus s est grand plus la dissymétrie est prononcée,

- $-1 \le s \le +1$,
- s négatif indique une courbe oblique à droite (étalée à gauche),
- s positif indique une courbe oblique à gauche (étalée à droite).

Coefficient de Pearson , caractérise la position de la médiane (ou du mode) par rapport à la moyenne, utilisé pour les distributions modérément asymétriques.

$$s = \frac{3*(\bar{x} - \tilde{x})}{\sigma} \text{ ou } \qquad s = \frac{3*(\bar{x} - x_0)}{\sigma}$$

Coefficient de Yule, indique le rapport entre, la différence de l'étalement de la courbe à gauche de la médiane à l'étalement à droite, et leur somme.

$$s = \frac{Q_1 + Q_3 - 2*Q_2}{Q_3 - Q_1}$$

Autre : **Coefficient de Fisher γ_1**, il utilise les moments centrés.

$$\gamma_1 = \frac{\mu_3}{\sigma^3} \qquad \text{avec} \qquad \mu_3 = \frac{\sum_{i=1}^{i=n} (xi - \bar{x})^3}{n}$$

$\gamma_1 = 0$ (symétrique) $\gamma_1 > 0$ (oblique à gauche) $\gamma_1 < 0$ (oblique à droite)

L'aplatissement

Mesure la courbure anormale d'une courbe de fréquence par rapport à la courbe idéale de gauss.

- Courbe normale : mésocurtique
- Courbe plus aiguë : leptocurtique
- Courbe plus aplatie : platicurtique

Coefficient de Kelley G2,

$$G2 = \frac{1}{2}\frac{Q_3 - Q_1}{D_9 - D_1}$$

G2 = 0.25 (mésocurtique) G2 >>0.25 (leptocurtique) G2<<0.25 (platicurtique)

Coefficients de Fisher γ_2, il utilise les moments centrés.

$$\gamma_2 = \frac{\mu_4}{\sigma^4} - 3 \quad \text{avec} \quad \mu_4 = \frac{\sum_{i=1}^{i=n} (xi - \bar{x})^4}{n}$$

$\gamma_2 = 0$ (mésocurtique) $\gamma_2 > 0$ (leptocurtique) $\gamma_2 < 0$ (platicurtique)

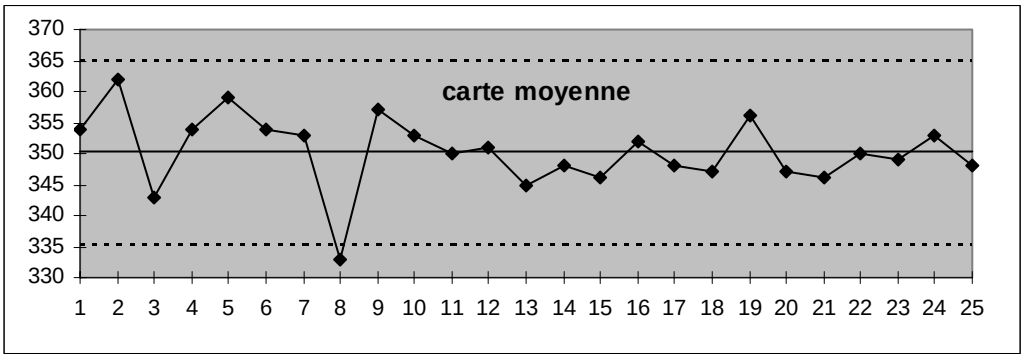
IV – Applications

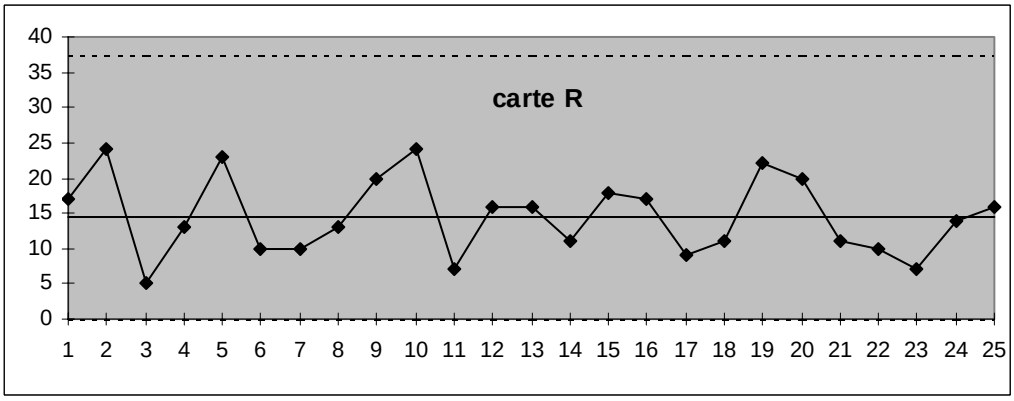
En contrôle de production on ccrée des cartes de contrôle aux moyennes, médianes, étendues et écart type et on détermine la Capabilité d'un procédé : Cp = (Ts – Ti) / 6 σ(x)

Exemple de cartes de contrôles:

N°	X1	X2	X3	ΣX	moyenne	R
1	357	344	361	1062	354	17
2	371	347	368	1086	362	24
3	346	342	341	1029	343	5
4	362	351	349	1062	354	13
5	373	354	350	1077	359	23
6	350	352	360	1062	354	10
7	355	347	357	1059	353	10
8	330	341	328	999	333	13
9	357	347	367	1071	357	20
10	364	340	355	1059	353	24
11	349	354	347	1050	350	7
12	349	360	344	1053	351	16
13	347	352	336	1035	345	16
14	355	345	344	1044	348	11
15	352	334	352	1038	346	18
16	363	346	347	1056	352	17
17	353	347	344	1044	348	9
18	343	344	354	1041	347	11
19	356	345	367	1068	356	22
20	339	359	343	1041	347	20
21	341	352	345	1038	346	11
22	346	356	348	1050	350	10
23	348	353	346	1047	349	7
24	344	357	358	1059	353	14
25	358	344	342	1044	348	16
				somme	**8758**	**364**
				moy =	**350,32**	**14,56**

	moyenne	R
LCI	335,42	0
LCS	365,22	37,49

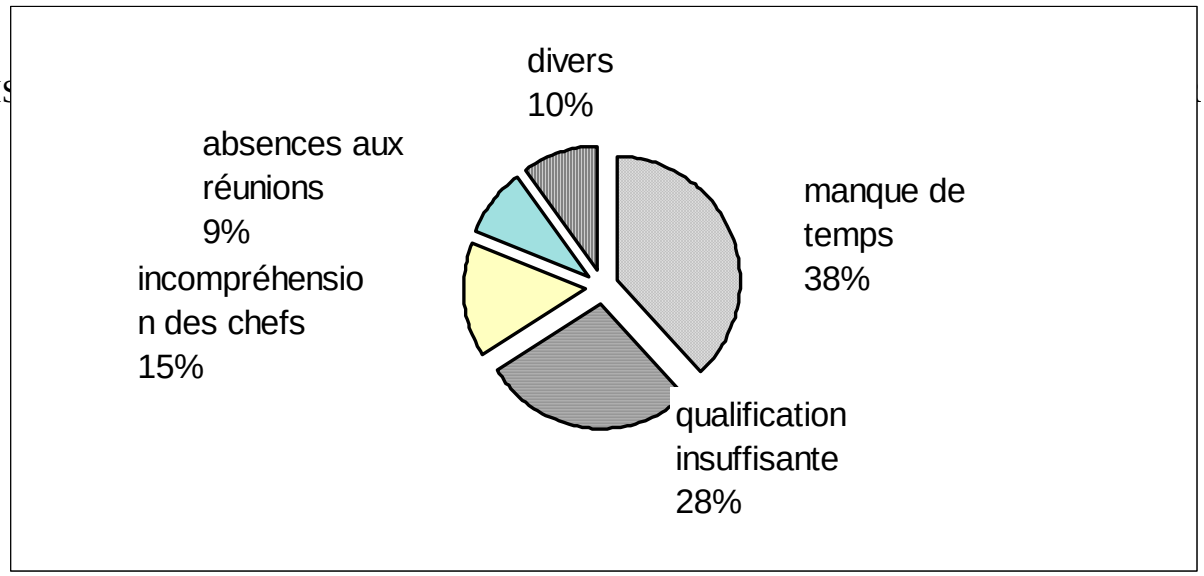




Utilisation des graphes: étude en cercle de qualité

manque de temps	0,38
qualification insuffisante	0,28
incompréhension des chefs	0,15
absences aux réunions	0,09
divers	0,1
total	1

" Je n'ai pas suffisamment de temps pour faire mon travail" est le problème numéro un des animateurs de cercle de qualité (38%), viennent ensuite "les capacités des membres du groupe ne sont pas à la hauteur de la tâche" (28%), puis "les chefs d'équipe ne comprennent pas mes problèmes"(15%) . Il convient d'étudier de près ces 3 éléments constituants 76% de toutes les réponses.


divers
10%
absences aux
réunions
9%
incompréhensio
n des chefs
15%
manque de
temps
38%
qualification
insuffisante
28%