Statistique descriptive élémentaire

Bioanalyse - Master 1 Biochimie - Dr Adel SIDI-YAKHLEF

PLAN

1. Introduction et définitions

L'objectif de la statistique descriptive est de décrire les données de manière efficace, c'est à dire permettant de mieux "voir" et comprendre les données.

L'éventail de méthodes disponibles est large, nous nous limiterons ici aux plus classiques.

Quelques définitions importantes

Population	Ensemble de référence (auquel on s'intéresse), sur lequel vont porter les observations et le recueil des données.
Individu	Elément de la population.
Echantillon	Sous-ensemble de la population.
Caractère	Une caractéristique des individus à laquelle on s'intéresse. On associera en général une **variable** à un caractère.
Paramètre	Un paramètre correspond à la population.
Statistique	Une statistique est calculée sur un échantillon.

Nous nous limiterons ici à la description d'un seul caractère (on parle d'**analyse descriptive univariée**) ou de 2 caractères simultanés (analyse descriptive **bivariée**)
L'analyse descriptive **multivariée** (nombre de caractères décrits simultanément supérieur à 2) ne sera pas abordée ici.

2. Etude d'un seul caractère

Nous distinguerons les caractères **qualitatifs** des caractères **quantitatifs.**

2.1. Caractère qualitatif

les valeurs de la variable sont des "modalités"

ces modalités peuvent correspondre, à un ordre, à une échelle, à un classement (on parlera alors de variable **ordinale**) ou non (on parlera alors de variable **nominale**)

Exemples de variables qualitatives :

On s'intéresse aux 71 employés de l'entreprise (voir feuille Employés)

Exemple 1 : variable "sexe"

Un individu est un employé. Il y a **n =** **71** individus.

Le premier caractère étudié est le sexe, il a 2 modalités :

La première modalité (F) a un effectif n_1 = **28**

La deuxième modalité (M) a un effectif n_2 = **43**

Cette variable est une variable qualitative nominale

Exemple 2 : variable "niveau"

Par contre, la variable "niveau" est une **variable qualitative ordinale.**

2.1.1 Tableau de distribution

Pour construire un tel tableau, il suffit de calculer l'**effectif de chaque modalité**.

Pour l'exemple 1

Numéro de la modalité	Intitulé	*Effectif*
1	F	28
2	M	43
	TOTAL	**71**

Pour l'exemple 2

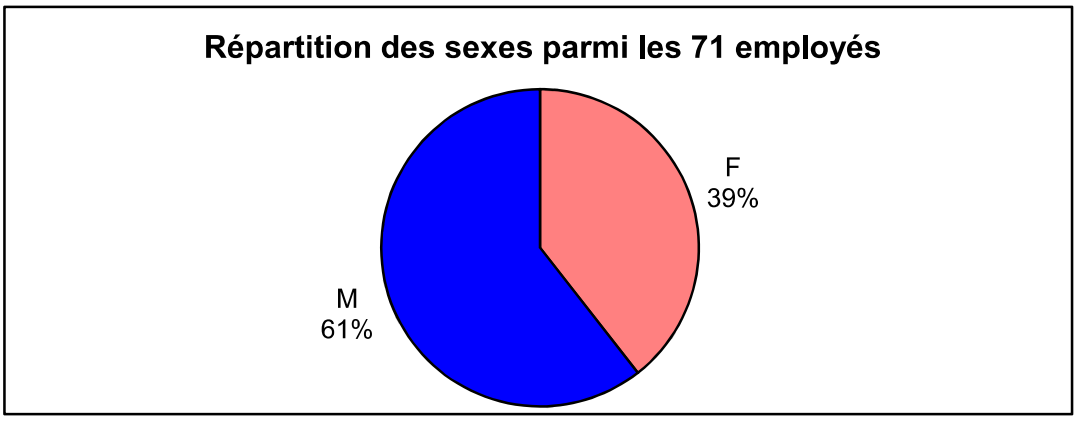
Numéro de la modalité	Intitulé	*Effectif*	*Fréquence*
1	A	40	0,56
2	B	24	0,34
3	C	7	0,10
	TOTAL	**71**	**1**

Savoir-faire EXCEL

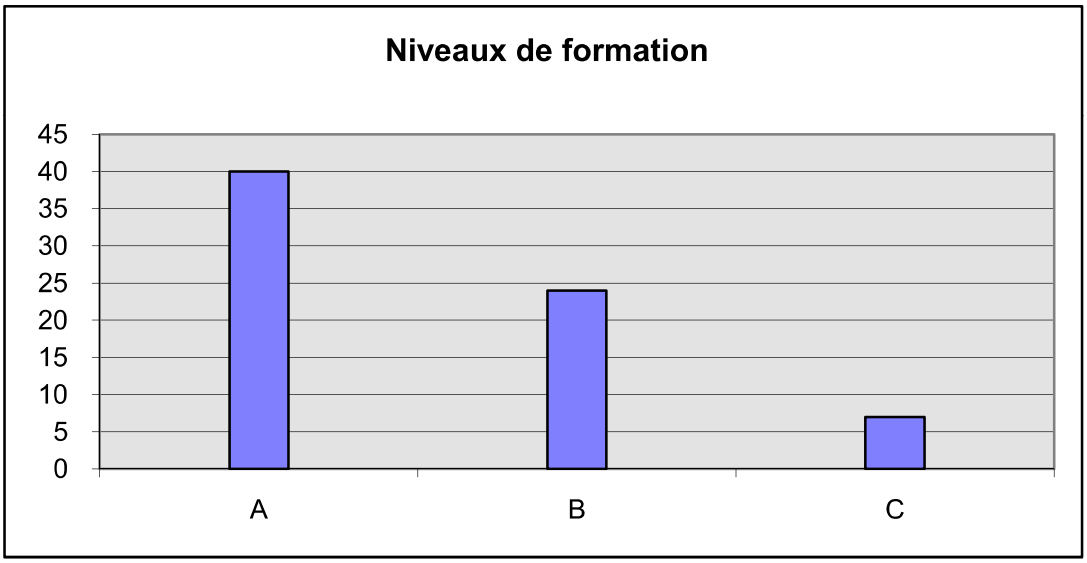
On peut obtenir les effectifs à l'aide de la fonction NB.SI ou alors avec un repport de tableau croisé dynamique

2.1.2 Synthèse graphique

On peut utiliser un diagramme circulaire ("camembert"), ou un diagramme à barres.



Pour l'exemple 2 (niveau), un diagramme à barres aurait été sans doute mieux adapté (variable ordinale) car permettant de visualiser l'ordre des niveaux.



2.1.3 Paramètres

Fréquence ou proportion

On obtient la fréquence pour une modalité en divisant l'effectif de la modalité par l'effectif total.

$$f_i = n_i / n$$

On rajoute une colonne fréquence (et une colonne : pourcentage), au tableau précédent :

Numéro de la modalité	Intitulé	Effectif	Fréquence	Pourcentage
1	F	28	0,3944	39,44%
2	M	43	0,6056	60,56%
	TOTAL	71	1	100%

2.2. Caractère quantitatif

Un caractère est dit quantitatif s'il peut faire l'objet d'une mesure, s'il correspond à une "quantité". On peut lui associer une **variable quantitative**.

2.2.1 Variable discrète

Si l'ensemble des valeurs que peut prendre une variable est fini, on parlera de **variable discrète**. La description la plus élémentaire est un **tableau de données ponctuelles** listant toutes les valeurs prises par la variable dans l'échantillon considéré.
On peut améliorer la description en construisant (comme pour une variable qualitative), un **tableau de distribution**. Lorsque le nombre de valeurs n'est pas trop important, on les place, toutes, triées, dans la colonne de gauche du tableau et dans la colonne de droite, on calcule le nombre d'occurrences pour chaque valeur.(= le nombre de fois que cette valeur apparaît).
Lorsque le nombre de valeurs est important, on groupe les valeurs en classes comme pour une variable continue.

Exemple 3

On s'intéresse au nombre d'étudiants utilisant un certain logiciel de statistique sur une période de 30 jours. Soit X, la variable correspondante.

Tableau de données ponctuelles

j	X
1	2
2	3
3	0
4	2
5	4
6	1
7	2
8	2
9	4
10	5
11	1
12	2
13	4
14	2
15	0
16	3
17	4
18	1
19	2
20	0
21	3
22	4
23	2
24	2
25	4
26	0
27	1
28	5
29	1
30	4

Tableau de distribution

0	**4**
1	**5**
2	**9**
3	**3**
4	**7**
5	**2**

Le tableau de gauche présente les données "brutes", c'est-à-dire les différentes valeurs observées pour X. Le tableau de droite permet une lecture plus "efficace" de ces données

Savoir-faire EXCEL

Pour obtenir un tel tableau de distribution, on peut utiliser la fonction matricielle **FREQUENCE** d'EXCEL... (Faire défiler pour en savoir plus...)

2.2.2 Variable continue

Mathématiquement, il s'agit d'une variable qui peut prendre toute valeur réelle ou toute valeur située dans un intervalle réel.

Exemple 4 Considérons la variable "**salaire**" dans le tableau "Employés"

données brutes

NUMERO	SALAIRE (euros)
1	21 894,40
2	38 539,20
3	38 177,60
4	38 738,40
5	36 481,60
6	35 226,40
7	15 516,00
8	15 516,00
9	30 499,20
10	15 180,00
11	38 539,20
12	42 812,00
13	47 035,20
14	35 506,40
15	31 907,20
16	38 743,20
17	33 259,20
18	43 174,40
19	38 539,20
20	38 901,60
21	35 226,40
22	39 984,80
23	49 024,00
24	18 020,00
25	15 516,00
26	41 507,20
27	37 688,80
28	31 346,40
29	15 973,60
30	24 338,40
31	28 950,40
32	18 020,00
33	24 014,40
34	31 346,40
35	24 338,40
36	35 226,40
37	28 592,80
38	35 226,40
39	48 035,20
40	33 744,00
41	33 259,20
42	38 177,60
43	36 989,60
44	38 901,60
45	38 539,20
46	41 035,20
47	35 226,40
48	35 226,40
49	40 439,20
50	36 989,60
51	24 160,00
52	36 140,00
53	15 481,60
54	18 020,00
55	19 684,80
56	15 481,60
57	33 744,00
58	35 226,40
59	38 539,20
60	38 743,20
61	38 901,60
62	17 192,80
63	30 368,80
64	38 539,20
65	25 833,60
66	25 833,60
67	28 950,40
68	41 793,60
69	38 743,20
70	38 539,20
71	31 310,40

le tableau trié apporte plus d'information

Savoir-faire EXCEL

Pour trier un tableau, utiliser la commande **Données Trier** après avoir sélectionné le tableau.

données triées

	NUMERO	SALAIRE
1	10	15 180,00
2	53	15 481,60
3	56	15 481,60
4	7	15 516,00
5	8	15 516,00
6	25	15 516,00
7	29	15 973,60
8	62	17 192,80
9	24	18 020,00
10	32	18 020,00
11	54	18 020,00
12	55	19 684,80
13	1	21 894,40
14	33	24 014,40
15	51	24 160,00
16	30	24 338,40
17	35	24 338,40
18	65	25 833,60
19	66	25 833,60
20	37	28 592,80
21	31	28 950,40
22	67	28 950,40
23	63	30 368,80
24	9	30 499,20
25	71	31 310,40
26	28	31 346,40
27	34	31 346,40
28	15	31 907,20
29	17	33 259,20
30	41	33 259,20
31	40	33 744,00
32	57	33 744,00
33	6	35 226,40
34	21	35 226,40
35	36	35 226,40
36	38	35 226,40
37	47	35 226,40
38	48	35 226,40
39	58	35 226,40
40	14	35 506,40
41	52	36 140,00
42	5	36 481,60
43	43	36 989,60
44	50	36 989,60
45	27	37 688,80
46	3	38 177,60
47	42	38 177,60
48	2	38 539,20
49	11	38 539,20
50	19	38 539,20
51	45	38 539,20
52	59	38 539,20
53	64	38 539,20
54	70	38 539,20
55	4	38 738,40
56	16	38 743,20
57	60	38 743,20
58	69	38 743,20
59	20	38 901,60
60	44	38 901,60
61	61	38 901,60
62	22	39 984,80
63	49	40 439,20
64	46	41 035,20
65	26	41 507,20
66	68	41 793,60
67	12	42 812,00
68	18	43 174,40
69	13	47 035,20
70	39	48 035,20
71	23	49 024,00

Tableau de distribution

On effectue des "regroupements en classes"
Le choix d'une bonne valeur k pour le nombre de classes est important.
En effet, avec un k trop petit, on voit mal la répartition des données et avec un k trop grand, on synthétise mal ces données.
Une règle empirique conseille de choisir un nombre de classes égal au plus petit entier k

tel que : $2^k >= n$.

où n est la taille de l'échantillon.

Dans le cas des employés, n =	71		
On peut donc retenir :	7	classes	
La largeur de classe égale à	4834,86	sera arrondie à	5000

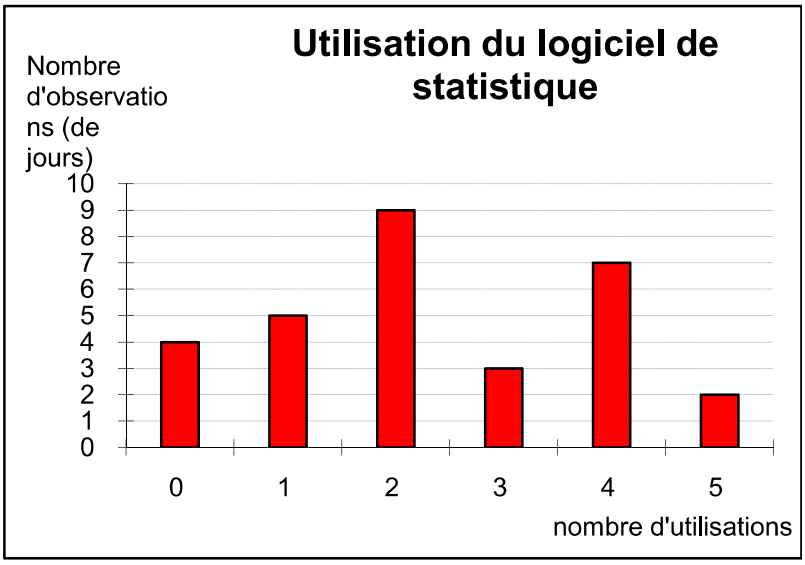
Classes			Effectif	Fréquence	Fréquence cumulée
15000	à	20000	12	0,169	0,169
20000	à	25000	5	0,070	0,239
25000	à	30000	5	0,070	0,310
30000	à	35000	10	0,141	0,451
35000	à	40000	30	0,423	0,873
40000	à	45000	6	0,085	0,958
45000	à	50000	3	0,042	1,000
		Total	71		

Il est bien sûr possible de définir des amplitudes différentes selon les classes.

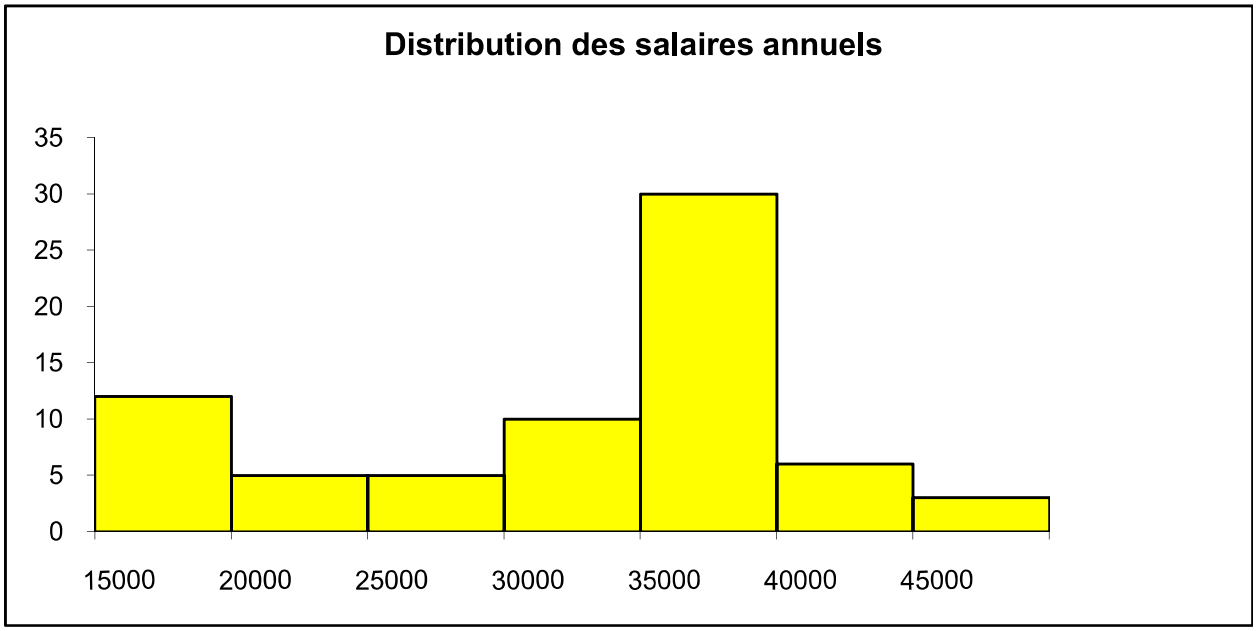
2.2.3 Synthèse graphique

Pour l'exemple 3
= variable discrète avec peu de valeurs.

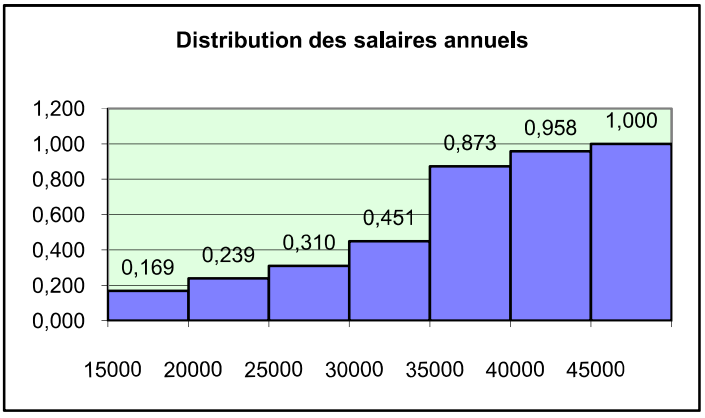
Diagramme en bâtons



Pour l'exemple 4 = variable continue : **Histogramme** (pas d'espace entre les barres)



A partir des fréquences cumulées



2.2.5 Paramètres

Paramètres de tendance centrale

Moyenne arithmétique

Lorsque l'on a toutes les valeurs de la variable x_1, x_2, ..., x_N le calcul de la moyenne est simple :

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$$

Conventions de notation
- nous désignerons l'effectif total par N pour une population et par n pour un échantillon.
- la moyenne pourra être désignée par m (ou x barre) pour un échantillon et par μ pour une population.
- Lorsque la variable est groupée en classes, le calcul est un peu plus long.

Médiane

La médiane est une valeur Me partageant en deux moitiés la série ordonnée des valeurs observées x_1, x_2, ..., x_n

Paramètres de dispersion

Amplitude

C'est la différence entre le minimum et le maximum de la variable.

Variance

Lorsque l'on a toutes les valeurs de la variable x_1, x_2, ..., x_n, la variance s'obtient en calculant l'écart quadratique moyen c'est à dire la moyenne des carrés des écarts à la moyenne.

Ecart-type

L'écart-type qui correspond à la racine carrée de la variance est le paramètre de dispersion le plus fréquemment utilisé.

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}}$$ **population**

Pour un **échantillon** de taille n le calcul est légèrement différent pour des raisons que nous verrons ultérieurement. (au lieu de diviser par n, on divise par n-1)

$$s = \sqrt{\frac{\sum_{i=1}^{n} (x_i - m)^2}{n-1}}$$

Quartiles

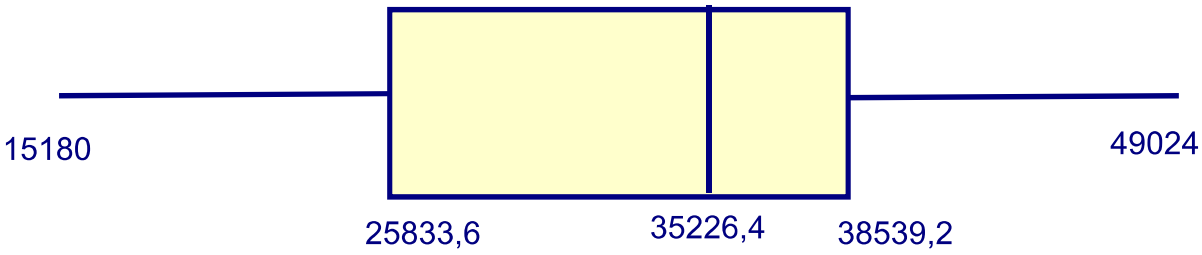
Q1 **Premier quartile** — valeur de la variable telle que l'on ait 25% d'éléments ayant une valeur inférieure.

Q3 **Troisième Quartile** — valeur de la variable telle que l'on ait 75% d'éléments ayant une valeur inférieure.

Q2 **Deuxième Quartile** — médiane

Intervalle interquartile — **Q3-Q1**

"Boîte à moustaches", "boxplot"



Exemples

Avec la variable continue du 2.2.2

	Savoir-faire EXCEL	
Paramètre	**Formule EXCEL**	**Résultat**
Compte	=NBVAL(valeur1;valeur2;...)	71
Somme	=SOMME(nombre1;nombre2;...)	2296317,6
Moyenne	=MOYENNE(nombre1;nombre2;...)	32 342,50
Minimum	=MIN(nombre1;nombre2;...)	15180
Maximum	=MAX(nombre1;nombre2;...)	49024
Amplitude	=MAX( ) - MIN( )	33844
Variance (population)	=VAR.P(nombre1;nombre2;...)	79 417 791,35
Ecart-type	=ECARTYPEP(nombre1;nombre2;...)	8 911,67
Ecart-type calculé à partir d'un échantillon	=ECARTYPE(nombre1;nombre2;...)	8 975,10
Médiane	=MEDIANE(nombre1;nombre2;...)	35226,4

Savoir-faire EXCEL
Il existe d'autres fonctions comme QUARTILE, PERCENTILE, ...
Vous pouvez voir la liste en utilisant l'assistant Fonction

3. Etude simultanée de deux caractères

3.1 Etude de deux caractères qualitatifs

Pour étudier la liaison entre deux caractères qualitatifs, on peut faire un tableau ayant pour lignes, les modalités du premier caractère et pour colonnes, les modalités du deuxième.

Ce type de tableau est appelé : TABLEAU CROISE (ou tableau de contingence)

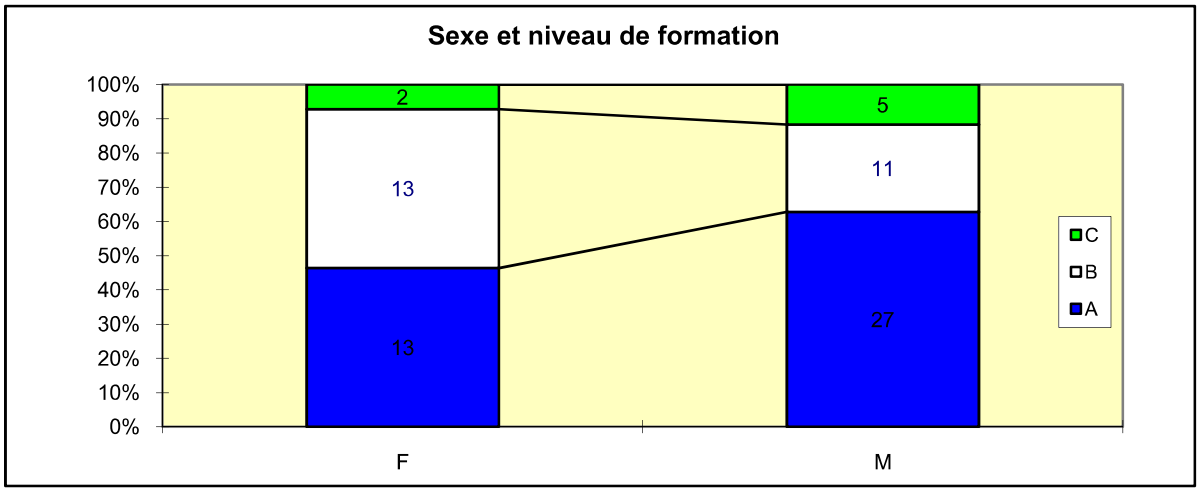
Exemple

On peut croiser les caractère SEXE et NIVEAU dans le tableau Employés.
Le tableau de contingence suivant peut être obtenu avec le "tableau croisé dynamique" d'EXCEL

NB NUMERO	NIVEAU			
SEXE	A	B	C	Total
F	13	13	2	28
M	27	11	5	43
Total	40	24	7	71

Savoir-faire EXCEL

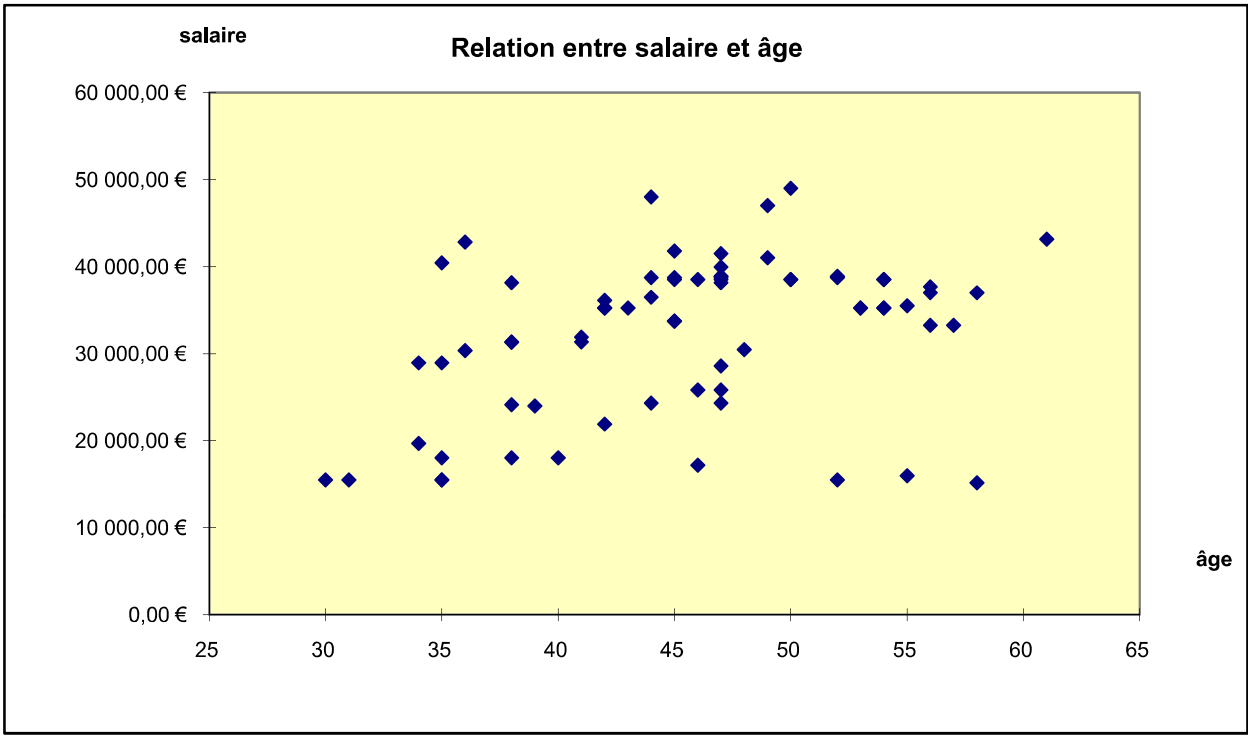
Utiliser Rapport de tableau croisé dynamique.



3.2 Etude de deux caractères quantitatifs

Pour étudier la liaison entre deux caractères quantitatifs, on peut faire un graphique de type "nuage de points"

Exemple:

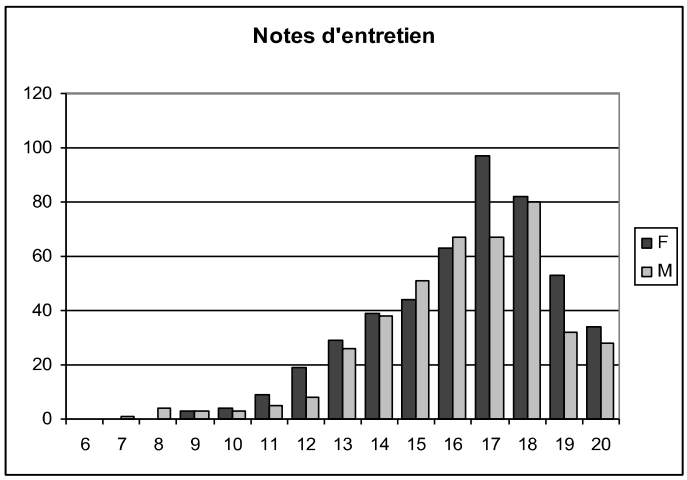


Nous irons plus loin dans ce domaine, lors de l'étude de la régression et de la corrélation

3.3 Etude d'un caractère quantitatif et d'un caractère qualitatif

Exemple

On considère les étudiants de Sup de Co et on veutcomparer les notes d'entretien des garçons et des filles
On peut faire une comparaison graphique des distributions



On peut également calculer et comparer paramètres comme les moyennes :

Moyenne filles	16,38
Moyenne garçons	16,22