Chapitre 6 : Les distributions à deux dimensions

Objectif général du chapitre :

Maitriser et appliquer les principes des distributions à deux dimensions.

Introduction :

L'étude des problèmes économiques de l'entreprise nous incite à prendre en considération, simultanément, au moins deux caractères. Ce qui implique, nous devons présenter un tableau statistique prédisposé à porter deux caractères, appelé aussi « tableau à double entrée ».

I. Présentation générale d'un tableau statistique à double entrée

Objectifs spécifiques :

- Comprendre la présentation générale d'un tableau statistique à double entrée.

Durée : 0,30 H.

Contenu :

Cette distribution est formée d'un ensemble d'observations de deux caractères (A et B) observé simultanément sur une même population P de « n » individus. Les « k » modalités de A sont : $\{A_1, A_2,\ldots, A_i,\ldots, A_k\}$ et les « l » modalités de B : $\{B_1, B_2,\ldots, B_j,\ldots, B_l\}$.

A \ B	B_1	B_2 …….	……..B_j	……..B_l	Total
A_1	n_{11}	n_{12}	n_{1j}	n_{1l}	$n_{1.}$
A_2 ……	n_{21}	n_{22}	n_{2j}	n_{2l}	$n_{2.}$
……A_i	n_{i1}	n_{i2}	n_{ij}	n_{il}	$n_{i.}$
……A_k	n_{k1}	n_{k2}	n_{kj}	n_{kl}	$n_{k.}$
Total	$n_{.1}$	$n_{.2}$	$n_{.j}$	$n_{.l}$	$n_{..}$

Titre : Tableau n° 1 : Tableau statistique à double entrée : La répartition de P selon les deux caractères : A et B.

♠ n_{ij} : est l'effectif des individus qui présente à la fois les modalités A_i et B_j,

♠ $n_{.j}$: est le total des effectifs de la colonne « j » effectué sur l'indice « i », on a :

$n_{.j} = n_{1j} + n_{2j} + \ldots + n_{ij} + \ldots + n_{kj} = \sum_{i=1}^{k} n_{ij}$;

♠ $n_{i.}$: est le total des effectifs de la ligne « i » effectué sur l'indice « j », on a : $n_{i.} = n_{i1} + n_{i2} + \ldots + n_{ij} + \ldots + n_{il} = \sum_{j=1}^{l} n_{ij}$;

♠ $n_{..}$: est l'effectif total. On peut le déterminer de deux manières :

On additionne les lignes : $n_{..} = \sum_{i=1}^{k}\sum_{j=1}^{l} n_{ij} = n$.

On additionne les colonnes : $n_{..} = \sum_{j=1}^{l}\sum_{k=1}^{k} n_{ij} = n$.

II. Exemple d'application

Objectifs spécifiques :

- Discerner et identifier les différents types de distributions existantes : marginales et conditionnelles.

Durée : 1 H.

Contenu :

Rémunération \ Age	[25, 40[	[40, 45[	Total (1)
[200, 400[	200	25	225
[400, 600[	700	100	800 = $n_{i.}$
[600, 800[	200	600	800
Total (2)	1100	725 = $n_{.j}$	1825 = n

Titre : Tableau n°2 à double entrée : Les distributions marginales des ouvriers d'une entreprise selon les deux caractères (âge et rémunération mensuelle).

On va analyser ce tableau de la manière suivante, en déterminant les distributions marginales des fréquences et celles des fréquences ainsi que les distributions conditionnelles.

a) *Les distributions marginales des effectifs* sont résumées dans les totaux du tableau au-dessus :

- Le total (1) : est la distribution marginales des ouvriers (la population) selon le caractère « la rémunération par mois » (l'âge n'intervient pas).
- Le total (2) : est la distribution marginales des ouvriers (la population) selon le caractère « âge » (la rémunération mensuelle n'intervient pas).

Elles sont assimilées à deux distributions à une seule dimension (le seul caractère mentionné, l'autre caractère n'intervient pas).

b) Les distributions marginales des fréquences se présentent dans les totaux du tableau suivant : (les valeurs du tableau sont en pourcentage).

Il faut toujours préciser les éléments clés suivants :

♣ $f_{ij} = \frac{n_{ij}}{n}$ est la fréquence de l'évènement (A_i, B_j) ou celle des observations qui présentent simultanément les modalités (A_i, B_j),

♣ $f_{i.} = \frac{n_{i\bullet}}{n} = \sum_{j=1}^{l} f_{ij} = \sum_{j=1}^{l} \frac{n_{ij}}{n}$ est le total des fréquences de la ligne « i »,

♣ $f_{.j} = \frac{n_{\bullet j}}{n} = \sum_{i=1}^{k} f_{ij} = \sum_{i=1}^{k} \frac{n_{ij}}{n}$ est le total des fréquences de la colonne « j »,

♣ $\sum_{i=1}^{k}\sum_{j=1}^{l} f_{ij} = \sum_{i=1}^{k} f_{i\bullet} = \sum_{j=1}^{l} f_{\bullet j} = 1$.

Rémunération \ Age	[25, 40[	[40, 45[	Total (3)
[200, 400[	10,959	1,37	**12,329**
[400, 600[	38,356	5,479	43,836 = $f_{i.}$
[600, 800[	10,959	32,877	43,8356
Total (4)	**60,274**	39,726 = $f_{.j}$	100

Titre : Tableau n°3 : Les distributions marginales des fréquences des ouvriers selon la rémunération mensuelle et l'âge.

- Le total (3) : est la distribution marginale des fréquences des ouvriers selon la rémunération mensuelle (l'âge n'intervient pas). Par exemple, 12,329 % des ouvriers ont une rémunération mensuelle entre 200 et 400 dinars.
- Le total (4) : est la distribution marginale des fréquences des ouvriers selon l'âge (la rémunération mensuelle n'intervient pas). Par exemple, 60,274 % des ouvriers ont un âge entre 25 et 40 ans.

En effet, on peut dire que toute distribution statistique à deux dimmentions peut être décomposée en deux distributions statistiques marginales simples.

c/ Les distributions conditionnelles selon la rémunération mensuelle pour chaque classe d'âge : (les valeurs du tableau sont en pourcentage).

On a : $f_{i/j} = \frac{n_{ij}}{n_{\bullet j}}$.

Rémunération \ Age	[25, 40[	[40, 45[	Ensemble
[200, 400[	**18,182** = 200 /1100	3,448	12,329
[400, 600[	63,636	13,793	43,836
[600, 800[	18,182	82,759	43,836
Total	100	100	100
	⇓ Distribution 1	⇓ Distribution 2	

Titre : Tableau n°4 : Les distributions conditionnelles selon la rémunération mensuelle pour chaque classe d'âge

Interprétation : 18,182 % des ouvriers, qui remplissent la condition d'âge entre 25 et 40 ans, ont une rémunération mensuelle entre 200 et 400 dinars.

Remarques :

✓ Il y'a autant de distributions conditionnelles de la rémunération mensuelle qu'il y'a de modalités dans le caractère âge (nombre de colonnes):dans ce cas, elles sont en nombre de 2.

✓ Puisqu'on a fait le total par ligne, en considérant chaque colonne indépendamment de l'autre, il n'y a aucun sens de faire le total par colonne. Donc, au lieu de présenter le total dans la dernière colonne (appelée l'ensemble), on présente l'ensemble des fréquences des colonnes vu dans le tableau 3 (total 3).

✓ On dit que les deux caractères A et B sont indépendants si $f_{i/j} = f_{i.}$ et $n_{ij} = \frac{n_{i.} \times n_{.j}}{n}$, alors, les lignes et les colonnes du tableau à double entrée seront proportionnels. L'indépendance est, alors, une relation réciproque : si A indépendant de B $\Leftrightarrow$ B est indépendant de A.

d/ Les distributions conditionnelles selon l'âge pour chaque classe de rémunération mensuelle : (les valeurs du tableau sont en pourcentage). On a : $f_{j/i} = \frac{n_{ij}}{n_{i.}}$.

Rémunération \ Age	[25, 40[	[40, 45[	Total
[200, 400[	**88,889** = 200 /225	11,111	100 $\Rightarrow$ Distribution 1
[400, 600[	87,5	12,5	100 $\Rightarrow$ Distribution 2
[600, 800[	25	75	100 $\Rightarrow$ Distribution 3
Ensemble	60,274	39,726	100

Titre : Tableau n°5 : Les distributions conditionnelles selon l'âge pour chaque classe de rémunération mensuelle

Interprétation : 88,889 % des ouvriers, qui remplissent la condition de rémunération mensuelle entre 200 et 400 dinars, ont un âge entre 25 et 40 ans.

Remarques :

✓ Il y'a autant de distributions conditionnelles de l'âge qu'il y'a de modalités dans le caractère « la rémunération mensuelle » (nombre de lignes) : dans ce cas, elles sont en nombre de trois.

✓ Puisqu'on a fait le total par colonne, en considérant chaque ligne indépendamment de l'autre, il n'y a aucun sens de faire le total par ligne. Donc, au lieu de présenter le total dans la dernière ligne (appelée l'ensemble), on présente l'ensemble des fréquences des lignes vu dans le tableau 3 (total 4).

III. Le graphique de corrélation

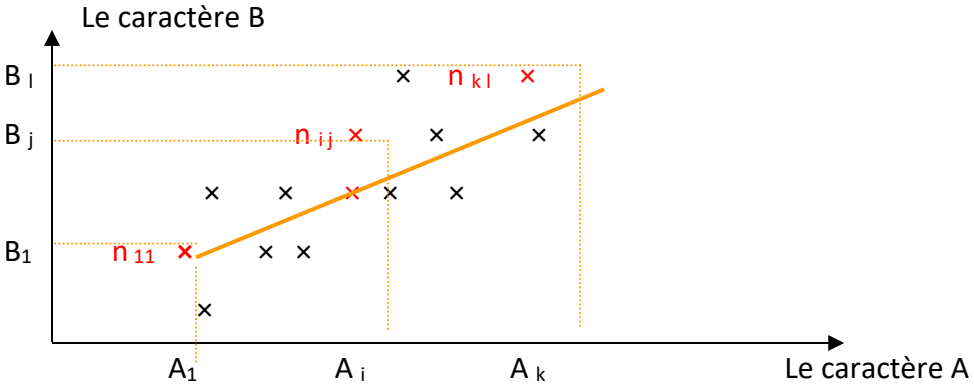
Objectifs spécifiques :

- Maîtriser les différentes relations qui peuvent exister entre les deux caractères.

Durée : 0,50 H.

Contenu :

Pour déterminer la nature de la relation, qui peut exister, entre ces deux caractères, on trace le premier en abscisse et le second en ordonné.



Titre : Graphique 1 : Graphique de corrélation.

Il existe une certaine corrélation entre A et B, dans ce cas.

Ce nuage de point peut prendre trois formes :

- Les points représentatifs sont distribués sur toute la surface du graphique. Ils sont placés au hasard. Alors, il n'y a aucun lien entre les deux caractères : A et B sont dits « indépendants ».

- Les points représentatifs sont bien rangés tout au long d'une courbe (droite, arc de cercle,). Alors, les deux caractères sont reliés étroitement : A et B sont dits « dépendants » (il existe, par exemple, une liaison fonctionnelle).
- En réalité, la situation se trouve entre ces deux extrêmes. Les points représentatifs se distribuent dans une région privilégiée du graphique en formant un nuage cohérent. Plus l'épaisseur du nuage est faible, plus on se rapproche de la liaison fonctionnelle : on dit, alors, qu'il y'a une forte relation fonctionnelle entre les deux caractères (voir graphique au-dessus). Plus, par contre, le nuage de point s'étale, moins ses limites seront précises et plus on est proche de l'indépendance : on dit, alors, que la corrélation entre les deux caractères est faible.

Dans ce cas, il existe plusieurs méthodes d'ajustement qui servent à construire la droite la plus proche du nuage de points, telle que la méthode des moindre carrées (MMC) de l'ajustement linéaire dont le principe est le suivant :

On pose la droite recherchée : Y'= a x + b, avec a et b deux coefficients inconnus : C'est la droite la plus proche possible des points observés du nuage. Elle reproduit, alors, d'une façon satisfaisante l'allure générale de ce nuage de points. La logique est de minimiser au maximum les carrées des écarts verticaux (δ_i) entre les valeurs observées et celles ajustées.

Le critère de choix de la meilleure droite possible (celle qui passe au mieux entre le maximum de points) est appelé « le critère des moindres carrées », tel que :

$\sum_{i=1}^{n} \delta_i^{\,2}$ est minimum, avec δ_i = $y_i - y'_i$: est la différence entre l'ordonné observé du point en question et sa valeur ajustée sur la droite.

δ_i peut être négatif (si le point observé se trouve réellement au-dessous de la droite recherchée) ou positif (si le point observé se trouve réellement au-dessus de la droite recherchée).

Ce critère nous renvoie à rechercher le minimum de la fonction (à deux variables) :

G (a, b) = $\sum_{i=1}^{n} \left(y_i - b - a x_i \right)^2$, ce qui revient à résoudre le système :

$$\begin{cases} \frac{dG}{da} = 0 \\ \text{et } \frac{dG}{db} = 0 \end{cases} \Leftrightarrow \begin{cases} -\sum_{i=1}^{n} x_i y_i + a \sum_{i=1}^{n} x_i^{\,2} + b \sum_{i=1}^{n} x_i = 0 \\ \text{et } -\sum_{i=1}^{n} y_i + a \sum_{i=1}^{n} x_i + n\,b = 0. \end{cases}$$

IV. Les séries chronologiques

Objectifs spécifiques :

- Identifier, brièvement, les séries chronologiques.

Durée : 0,45 H.

Contenu :

1. Définition :

Une série chronologique est une suite d'observations ordonnées en fonction du temps. Par exemple, la production mensuelle des ordinateurs.

En général, les staticiens préfèrent considérer la série chronologique comme étant une distribution à deux caractères dont l'un est, obligatoirement, le temps.

Par exemple, la série mensuelle de la production des ordinateurs est la distribution des résultats mensuels de production selon deux caractères. On va prendre un période des trois premiers mois des trois dernières années.

Production \ Mois	Janvier	Février	Mars
2000	100	110	105
2001	100	120	110
2002	105	115	110

Titre : Tableau 6 : La répartition des ordinateurs selon la production mensuelle (en unités)

Interprétation : La population est le nombre total des unitées mensuelles produites. Les deux caractères : la période de la production et le niveau des unités produites.

Remarque :

♥ Ce tableau à double entrée ne peut pas être interprété comme une distribution à deux caractères (mois et années).

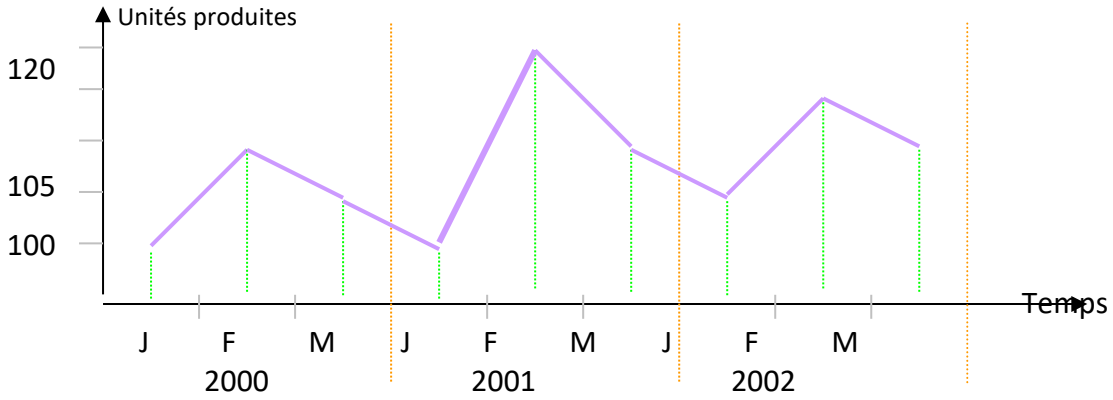
♥ Dans cet exemple, on observe des valeurs cumulées d'une variable pendant un intervalle de temps donné [t, t + 1] : cette variable est dite de « flux » ou de « débit ».

♥ Si la variable prend des valeurs à des instants précis (à une date donnée), par exemple : la température, cette variable sera dite « d'intensité » ou de « niveau ».

2. Représentations graphiques :

Les graphiques valables pour la série chronologique ne peuvent être ni un diagramme en bâtons, ni un histogramme, mais plutôt de ces deux formes spécifiques :

- Diagramme cartésien :



Titre:Graph n° 2 Diagramme cartésien : L'évolution de la production des ordinateurs en unité

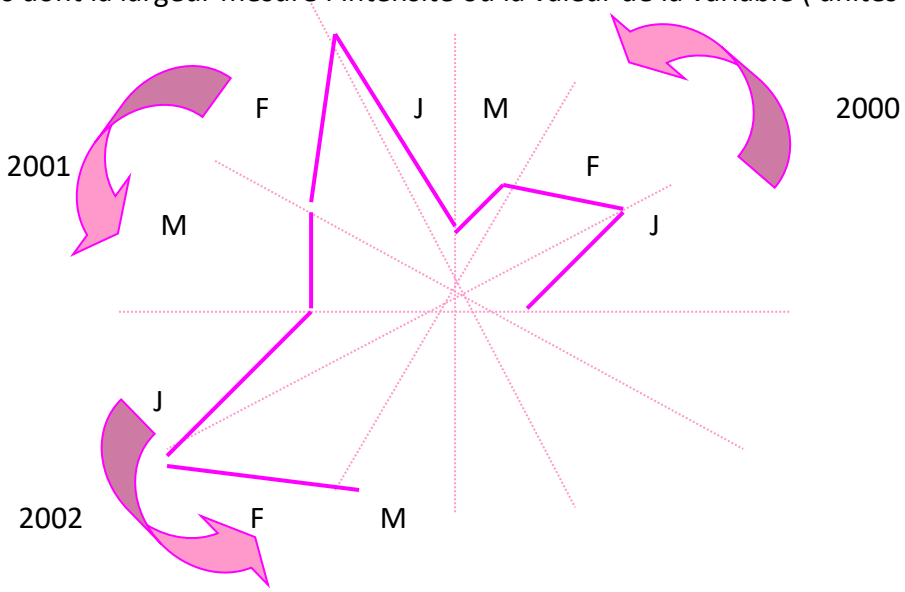
On a associé la valeur observée à l'intervalle de temps correspondant ou à son milieu.

Interprétation : A chaque période de production correspond un seul niveau de production. Donc, la production est liée fonctionnellement à la période (le contraire n'est pas vrai, c'est à dire, chaque niveau de production peut être associé à différentes périodes).

Cette courbe présente un caractère saisonnier : les pointes saisonnières est en février et les creux en janvier.

- Diagramme polaire :

A des périodes successives et égales (mois ou trimestre), on fait correspondre des rayons dont la largeur mesure l'intensité ou la valeur de la variable (unités produites).



Titre : Graph n° 3 Diagramme polaire : L'évolution de la production des ordinateurs en unités

3. Les composantes d'une chronique :

Une série chronologique peut présenter quatre composantes, qui sont, en général, prises en compte dans sa propre décomposition.

- La tendance (T) : est l'allure générale de la chronique à $\pm$LT. C'est, alors, une variation lente (de hausse ou de baisse) sur plusieurs années.
- La composante cyclique (C) : est périodique à $\pm$LT. On dit qu'il y'a une périodicité si le phénomène en question se répète au moins une fois).

La tendance et la composante cyclique forme le mouvement conjoncturel ou extra saisonnier.

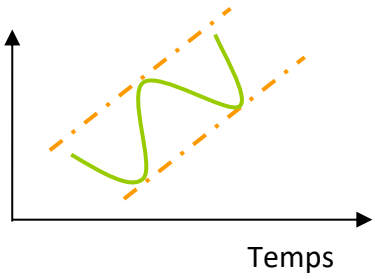
- La composante saisonnière (S) : est un mouvement périodique, de hausse ou de baisse, de durée moyenne (≤ à 1 an) qui se reproduit au même moment dans la période ultérieure.
- La composante accidentelle (E) : appelée aussi irrégulière, résiduelle ou erreur. Elle tient compte des faits accidentels et imprévisible (qui peuvent intervenir à un moment donné, influençant localement la série, sans toutefois affecter son allure générale).

4. Les modèles d'une chronique :

- Le modèle additif : pour lequel toutes les composantes s'additionnent afin de former la série : $y_t = T_t + C_t + S_t + E_t$. Elles sont alors indépendantes.

La droite qui lie les creux et celles qui lie les pointes sont parallèles.

Titre : Graphique 4 : Chronique de modèle additif.



- Le modèle multiplicatif : pour lequel l'équation de la courbe s'écrit en multipliant les quatre composantes : $y_t = T_t \times C_t \times S_t \times E_t$. Elles sont alors dépendantes.

La droite qui lie les creux et celles qui lie les pointes sont sécante.

Titre : Graphique 5 : Chronique de modèle multiplicatif.