Méthodes d'ajustements graphiques : Diagramme Quantile – Quantile

Bernard Goldfarb (goldfarb@dauphine.fr),
Catherine Pardoux (pardoux@dauphine.fr)

I Objet

Le diagramme *quantile-quantile* très facile à construire avec le tableur Excel, permet une appréciation graphique de l'ajustement d'une distribution observée à un modèle théorique. Sur ce graphe, l'axe des ordonnées porte les quantiles x_i de la distribution observée, tandis que l'axe des abscisses porte les quantiles x_i^* correspondants de la loi théorique. *Le nuage des points* $\left(x_i^*, x_i\right)$ *s'aligne sur la première bissectrice* lorsque la distribution théorique proposée est une bonne représentation des observations. On doit remarquer que l'appréciation de l'alignement des points le long de la bissectrice peut être considérée comme subjective. Toutes les déviations par rapport à l'alignement (extrémités présentant une courbure, points éloignés, ...) peuvent être repérées et analysées.

On peut tracer un *diagramme quantile-quantile* pour tout ajustement par une loi continue dont la fonction de répartition est strictement croissante, c'est-à-dire une loi dont la fonction de répartition est bijective sur l'intervalle correspondant à des valeurs non nulles de la fonction de densité et ne présentant pas de « trous ». Nous allons montrer l'application pour les lois normale, log-normale et exponentielle. Le tableur Excel permet de calculer automatiquement les quantiles de la loi normale.

II Exemples

II.1 Loi normale

En cas d'alignement, le type de modèle est alors retenu, et il reste à apprécier ses paramètres par la position des points par rapport à la première bissectrice (translation et/ou inclinaison) :

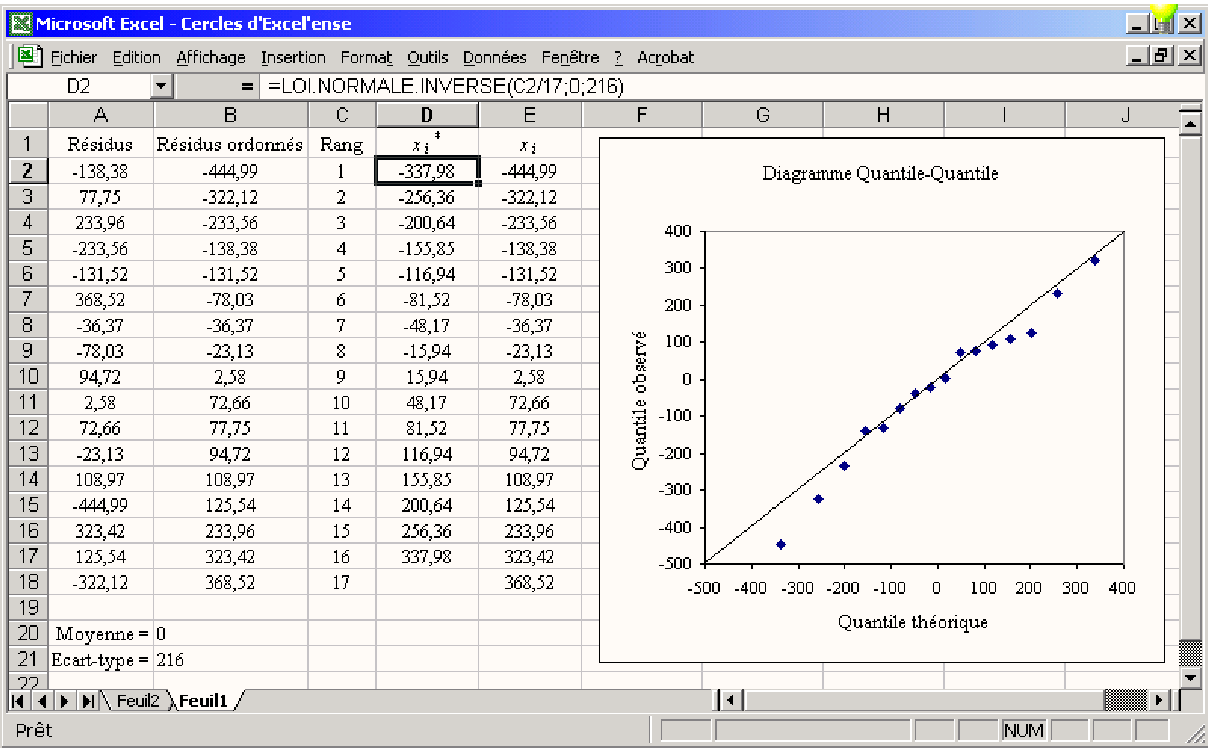
- un alignement sur une parallèle à la première bissectrice fera évoquer une erreur sur le choix de la caractéristique de position (moyenne, ...) de la distribution théorique,
- un alignement sur une droite passant par l'origine mais inclinée par rapport à la première bissectrice évoquera une erreur sur la caractéristique de dispersion (écart-type, ...),

un alignement sur une droite ne passant pas par l'origine et inclinée par rapport à la première bissectrice évoquera une erreur sur le choix des caractéristiques de position et de dispersion.

II.1.a Cas d'une série d'observations non classées

Soit la série des résidus d'un modèle linéaire, peut-on la considérer issue d'une distribution normale $\mathcal{N}(0 ; 216)$?
Après avoir trié les *n* valeurs observées par ordre croissant (figure 1), on crée une colonne des rangs (de 1 à *n*), et on détermine les quantiles théoriques comme sur la figure 1 (ici $n = 17$),

sauf pour la plus grande valeur observée. Il suffit ensuite de tracer le nuage des points et la bissectrice par l'Assistant graphique.



	A	B	C	D	E
1	Résidus	Résidus ordonnés	Rang	x_i^*	x_i
2	-138,38	-444,99	1	-337,98	-444,99
3	77,75	-322,12	2	-256,36	-322,12
4	233,96	-233,56	3	-200,64	-233,56
5	-233,56	-138,38	4	-155,85	-138,38
6	-131,52	-131,52	5	-116,94	-131,52
7	368,52	-78,03	6	-81,52	-78,03
8	-36,37	-36,37	7	-48,17	-36,37
9	-78,03	-23,13	8	-15,94	-23,13
10	94,72	2,58	9	15,94	2,58
11	2,58	72,66	10	48,17	72,66
12	72,66	77,75	11	81,52	77,75
13	-23,13	94,72	12	116,94	94,72
14	108,97	108,97	13	155,85	108,97
15	-444,99	125,54	14	200,64	125,54
16	323,42	233,96	15	256,36	233,96
17	125,54	323,42	16	337,98	323,42
18	-322,12	368,52	17		368,52
19					
20	Moyenne =	0			
21	Ecart-type =	216			

Figure 1 : Ajustement d'une série de résidus par une loi normale

On peut remarquer que l'intervalle centré en la moyenne (= 0) et de longueur égale à 2 fois l'écart-type contient 16 observations sur 17 (94,1 %).
D'autre part, la boîte de distribution est à peu près symétrique sans valeurs aberrantes (cf. Macro).

Tous ces éléments amènent à accepter l'ajustement par une loi normale $\mathcal{N}(0 ; 216)$ malgré la valeur la plus faible (- 444,99) qui s'éloigne fortement de la bissectrice.

II.1.b Cas d'une série d'observations classées

Les notes d'un échantillon de 80 copies d'examen se répartissent comme suit :

Note	[0;2[	[2;4[	[4;6[	[6;8[	[8;10[	[10;12[	[12;14[	[14;16[	[16;18[	[18;20]
Effectif	2	3	6	9	14	16	13	11	5	1
Eff. Cum.	2	5	11	20	34	50	63	74	79	80
F_i (%)	2,5	6,25	13,75	25	42,5	62,5	78,75	92,5	98,75	100

La médiane de cette distribution égale à 10,75 (=10 + 2 · 6 / 16) est voisine de la moyenne 10,55, l'écart-type valant 3,97.

On veut examiner si on peut considérer cette distribution issue d'une loi de Gauss $\mathcal{N}(10 ; 4)$. Les points étant quasi-alignés le long de la bissectrice sur le graphique quantile-quantile (figure 2), on ne rejette pas l'ajustement.

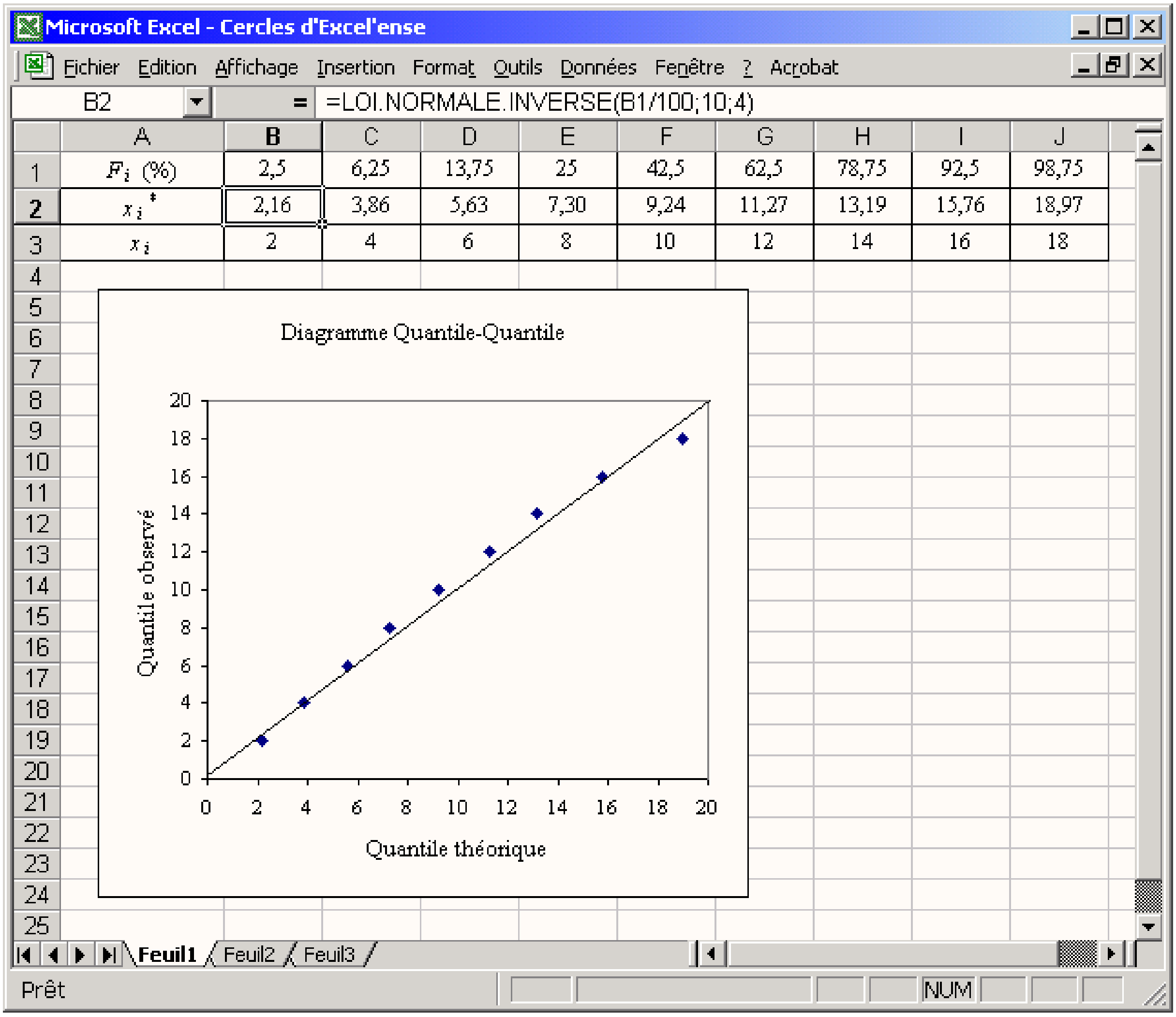


Figure 2 : Ajustement d'une distribution de notes par une loi normale

Notons que pour la loi normale, le diagramme Quantile – Quantile correspond à l'ajustement graphique par la droite de Henry qui se trace en utilisant les points (x_i , u_i^*), x_i étant le quantile observé et u_i^* le quantile correspondant d'une loi normale centrée réduite. L'alignement (ou quasi-alignement) permet de valider l'ajustement par une loi normale dont on peut évaluer graphiquement les paramètres à l'aide de la droite [Saporta, 1990].

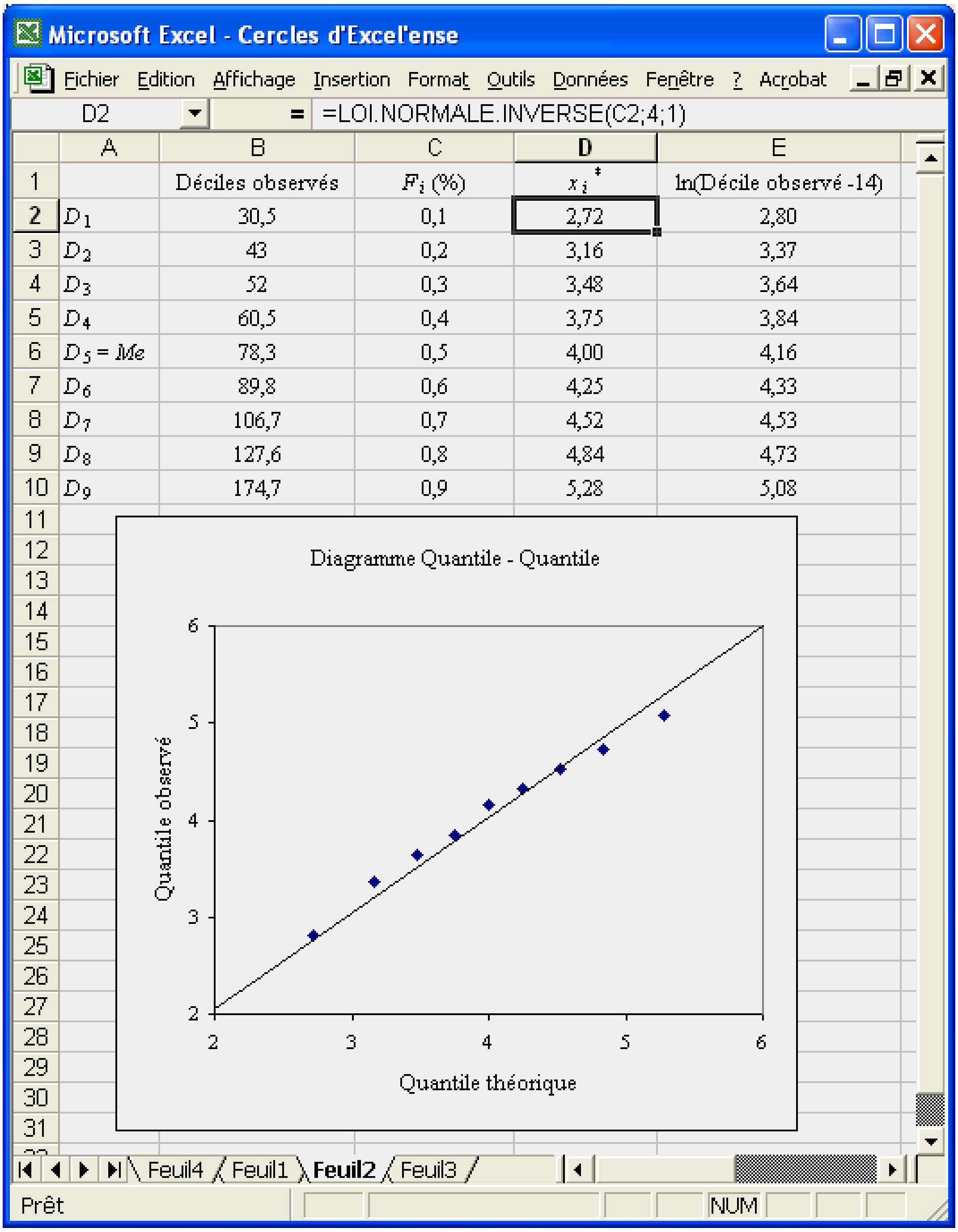
II.2 Loi log – normale

La distribution des durées moyennes mensuelles (en secondes) X des appels téléphoniques de 300 usagers est donnée par ses 9 déciles (figure 3). La médiane (78,3 s.) étant très nettement inférieure à la valeur du milieu de l'intervalle interdécile (102,6 s.), cette distribution est asymétrique et étalée vers la droite.

Les distributions de durées d'appels téléphoniques sont souvent modélisables par une loi log-normale. La durée minimum observée est égale à 14 s., les valeurs de la moyenne et de l'écart-type de la variable $Y = \ln(X - 14)$ valent respectivement 4 et 1.

On envisage un ajustement par une loi log-normale de paramètres : $m = 4$, $\sigma = 1$ et $x_0 = 14$, c'est-à-dire un ajustement de la distribution de Y par une loi de Gauss $\mathcal{N}(4 ; 1)$.

Le quasi-alignement des 9 points le long de la bissectrice permet de valider l'ajustement par la loi log-normale $\mathcal{LN}(4 ; 1 ; 14)$.



	A	B	C	D	E
1		Déciles observés	F_i (%)	x_i^*	ln(Décile observé -14)
2	D_1	30,5	0,1	2,72	2,80
3	D_2	43	0,2	3,16	3,37
4	D_3	52	0,3	3,48	3,64
5	D_4	60,5	0,4	3,75	3,84
6	$D_5 = Me$	78,3	0,5	4,00	4,16
7	D_6	89,8	0,6	4,25	4,33
8	D_7	106,7	0,7	4,52	4,53
9	D_8	127,6	0,8	4,84	4,73
10	D_9	174,7	0,9	5,28	5,08

Figure 3 : Ajustement d'une distribution de durées d'appel téléphonique par une loi log-normale

On ne peut pas utiliser la fonction **LOI.LOGNORMALE.INVERSE** puisque cette fonction Excel ne prévoit pas trois paramètres, mais seulement deux paramètres (moyenne et écart-type) sans possibilité de paramétrer la valeur minimum x_0 .

II.3 Loi exponentielle

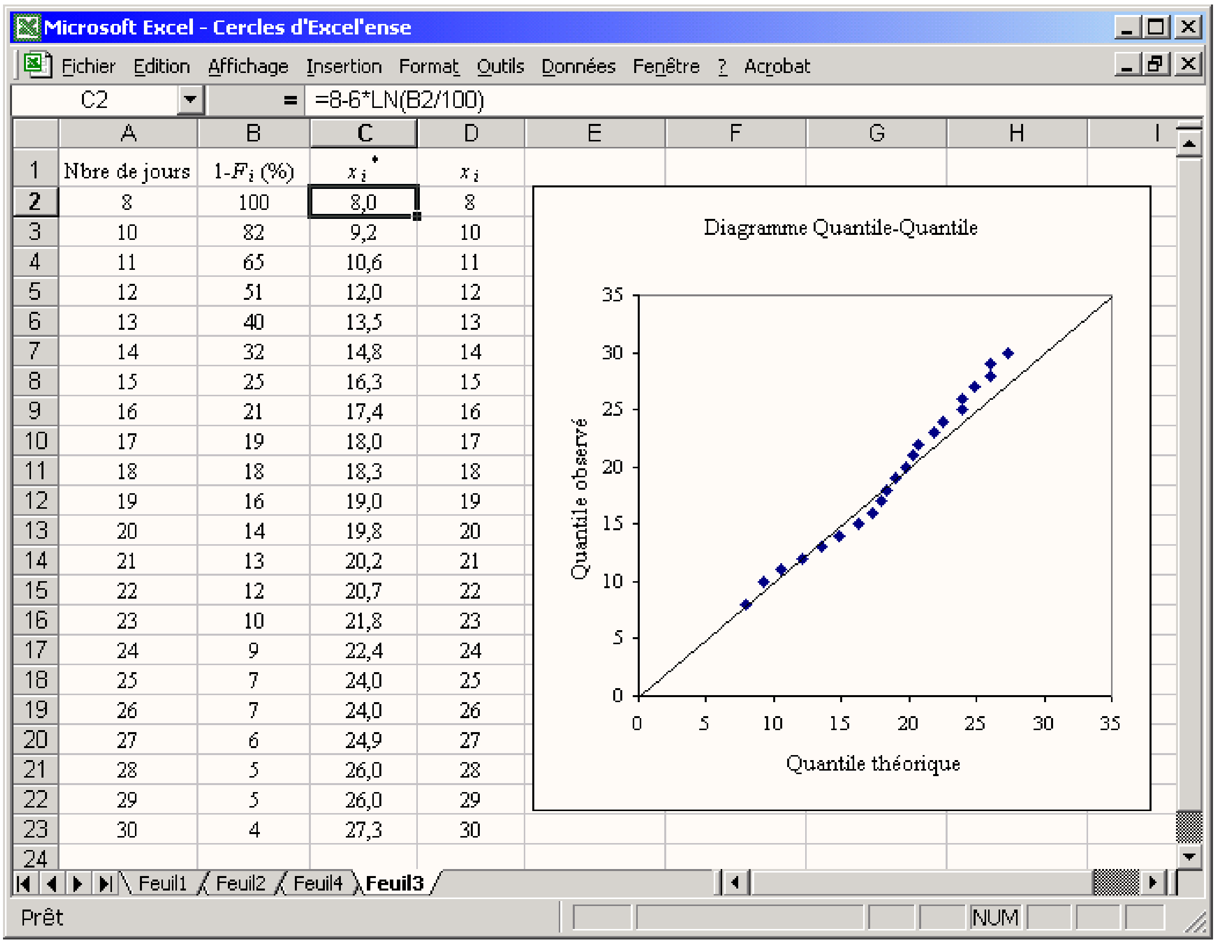
La distribution des durées X de fonctionnement – ou survie – (en jours) de 100 unités d'un matériel donné est présentée figure 4. La moyenne de cette distribution est égale à 14,1 et l'écart-type à 6.

Les distributions de durées de survie sont souvent modélisables par une loi exponentielle. La durée minimum observée étant égale à 8 jours, on envisage un ajustement par une loi exponentielle de paramètres : $\theta = 8$ et $\lambda = 6$.

On calcule les quantiles théoriques d'une loi exponentielle de paramètres $\theta = 8$ et $\lambda = 6$:

$$x_i^* = 8 - 6 \cdot \ln(1 - F_i)$$

Les 22 points $\left(x_i^*, x_i\right)$ étant peu éloignés de la bissectrice, l'ajustement par la loi exponentielle envisagée n'est pas rejeté.



Microsoft Excel - Cercles d'Excel'ense

C2 =8-6*LN(B2/100)

	A	B	C	D
1	Nbre de jours	$1-F_i$ (%)	x_i^*	x_i
2	8	100	8,0	8
3	10	82	9,2	10
4	11	65	10,6	11
5	12	51	12,0	12
6	13	40	13,5	13
7	14	32	14,8	14
8	15	25	16,3	15
9	16	21	17,4	16
10	17	19	18,0	17
11	18	18	18,3	18
12	19	16	19,0	19
13	20	14	19,8	20
14	21	13	20,2	21
15	22	12	20,7	22
16	23	10	21,8	23
17	24	9	22,4	24
18	25	7	24,0	25
19	26	7	24,0	26
20	27	6	24,9	27
21	28	5	26,0	28
22	29	5	26,0	29
23	30	4	27,3	30

Figure 4 : Ajustement d'une distribution de durées de survie par une loi exponentielle

III Références

Pour des précisions sur les lois de probabilités utilisées, on pourra consulter les deux ouvrages suivants.

B. Goldfarb et C. Pardoux (2004) *Introduction à la méthode statistique*. Dunod.

G. Saporta (1990) *Probabilités, Analyse des Données et Statistique*. Technip.