Détermination d'une méthode de calcul de capabilités avec des lois non gaussiennes

Thibaut Martini

HAGER ELECTRO
132 boulevard d'Europe
67215 OBERNAI CEDEX

UDS – UFR de Mathématiques
7 rue René Descartes
67000 STRASBOURG

Détermination d'une méthode de calcul de capabilités avec des lois non gaussiennes

Rapport de stage de 1ère année de master
Réalisé au service qualité du 1er juin et 30 juillet 2010

Thibaut MARTINI
1ere année de master statistique
2010

Maitre de stage :
Eric HELBLING
CQL (Component Quality Leader)

REMERCIEMENTS

Je tiens tout d’abord à remercier mon responsable de stage Eric HELBLING, pour avoir accepté ma candidature, me permettant ainsi d’effectuer mon stage au sein de l’entreprise Hager, mais aussi pour toute sa confiance et son temps accordé dans la résolution de la problématique de mon stage.

Merci également à toute l’équipe du service qualité d’Hager Obernai de m’avoir parfaitement accueilli et intégré dans leur équipe.

TABLE DES MATIERES

PREAMBULE

PRESENTATION DE L'ENTREPRISE HAGER

L'entreprise Hager est spécialisée dans l'installation et la distribution électrique depuis 1955, et a récemment étendu sa notoriété à l'échelle européenne. Elles possède 26 sites de production dans le monde dont un site sur Obernai où fut effectué le stage. Son chiffre d'affaire est de 1.3 milliards d'euros

Quelques chiffres chez Hager :

1.3 milliard d'euros de chiffre d'affaire en 2009
11 000 employés par le groupe à l'heure actuelle
63 implantations commerciales
26 sites de production

Les valeurs de la marque :

- L'Innovation : 5 % du chiffre d'affaire investi en recherche et développement et 850 ingénieurs et techniciens employés dans ce domaine. De plus 65 % des produits de la marque ont moins de 3 ans et 80 % moins de 5 ans.
- La Proximité : le groupe Hager souhaite pouvoir satisfaire l'ensemble de ses clients, du distributeur à l'utilisateur
- La Confiance : autant sur les rapports humains que par rapport à ses produits, l'entreprise veut tenir ses promesses, place l'intérêt du client avant tout et assure la fiabilité de ses produits
- La Réactivité : ce qui signifie pour Hager, de pouvoir répondre en un laps de temps minimum à la demande mais aussi d'inciter ses collaborateurs à assurer un service optimal
- La Simplicité : c'est un axe très important pour le groupe dans le but d'optimiser et de faciliter la relation fabriquant-client et ainsi d'assurer un service le plus efficace possible
- La Continuité : c'est une valeur de Hager qui concerne toutes les autres, le but étant de toujours respecter ces valeurs dans un soucis de respect et de satisfaction du client.

PARTIE 1 :
PRESENTATATION ET INTRODUCTION

1.1. PRESENTATION DE LA PROBLEMATIQUE

1.1.1. Présentation du Guidestat

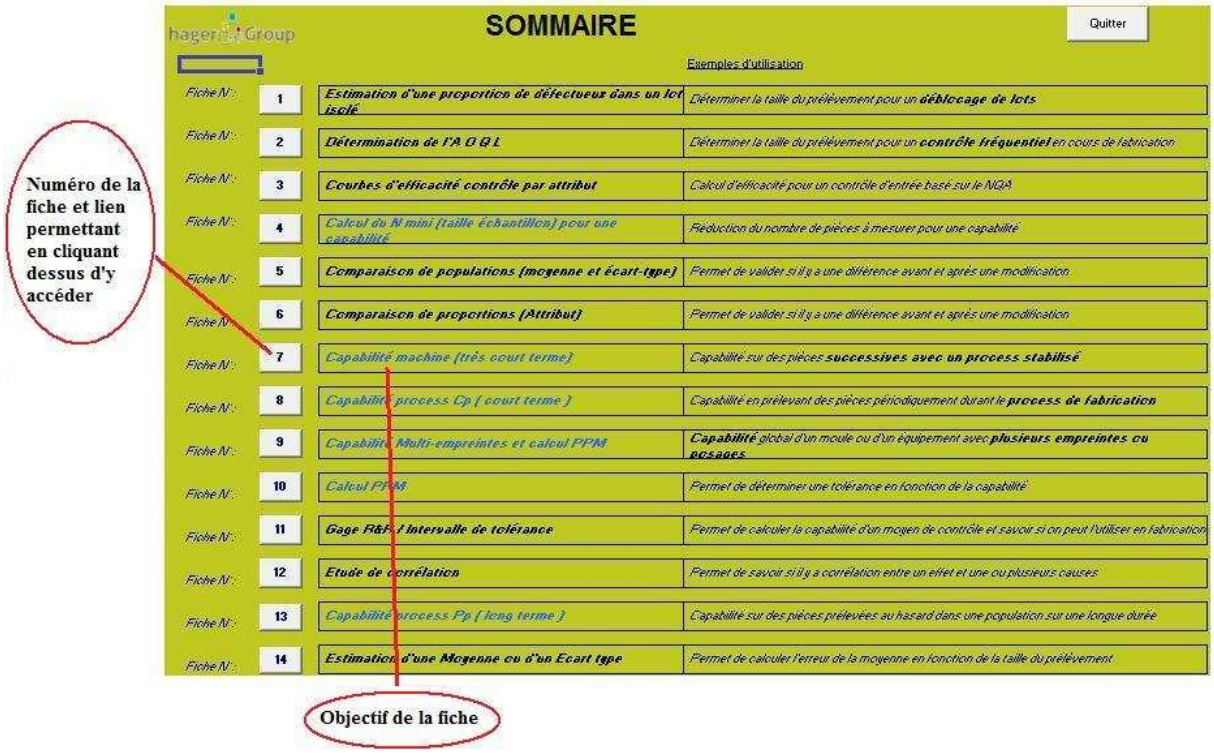
L'entreprise Hager privilégie la qualité de ces produits. Ainsi un pôle qualité très important est présent sur le site depuis quelques années. Ce contrôle a permis d'améliorer considérablement la qualité de ses produits et la satisfaction du client. L' utilisation d'un programme spécifique pour leurs statistiques appelé **Guidestat** établie par Eric Helbling y joue un grand rôle. Ce programme est un classeur **Excel** utilisant différentes fiches ou feuilles. Chaque fiche permet de déterminer la taille de prélèvement d'un échantillon ou de conclure sur un jeu de données en fonction de données initiales.

L'objectif de ce **Guidestat** est de rendre simple et rapide les résultats statistiques pour l'utilisateur. En effet, beaucoup de personnes sont utilisateurs de ce programme, et n'ont pas toujours le temps de conclure avec des logiciels spécifiques tels que **Minitab** qui demandent de connaitre l'interprétation des résultats et de réitérer certaines procédures .

Certains utilisateurs ne connaissent pas la théorie des statistiques, et sont spécifiés dans leur domaine, mais ils ont besoin d'avoir les conclusions statistiques. Les formations proposées par **Hager** permettent à ces personnes de connaitre l'interprétation des résultats et d'utiliser le **Guidestat.**

La figure ci-dessous permet de visualiser le sommaire du **Guidestat** et ses différentes utilisations :

Figure 1 :



Une particularité de ce programme est sa simplicité d'utilisation. Les seules manipulations à faire sont des **clics de souris sur les liens et l'insertion des données**.

La modification en grande partie de la fiche 7 fait partie de l'objectif du stage. Avant d'entrer plus précisément dans les détails., quelques petits rappels et précisions sont utiles pour la suite.

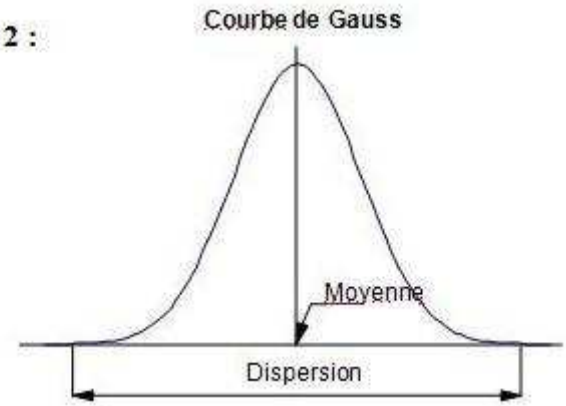
1.1.2. Rappels sur la loi normale

Introduction :

L'analyse de la production d'une machine montre généralement que la répartition des caractéristiques d'un produit suit une loi : la **loi Normale** (ou loi de Laplace-Gauss).

Cette loi est caractérisée par 2 paramètres :

Figure 2 :



1. La **moyenne estimée** par :

$$\overline{X} = \frac{\text{Somme des valeurs}}{\text{Nombre des valeurs}} = \frac{\sum_i X_i}{N}$$

2. L'**écart type estimé** par :

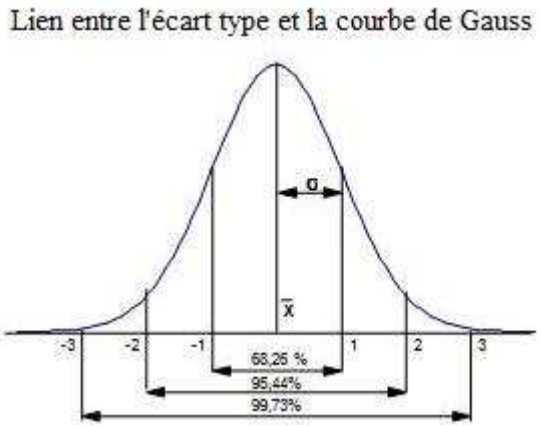
$$\sigma = \sqrt{\frac{\sum (X_i - \overline{X})^2}{N-1}} \text{ avec N = Nombre de valeurs}$$

Liaison entre l'écart type et la dispersion :

L'écart type nous permet de connaître très précisément les proportions de pièces qui sont comprises dans différents intervalles autour de la moyenne :

Par définition, nous appellerons dispersion pour une loi normale l'intervalle centré contenant 99,73% de la population. La valeur statistique de la dispersion est donc égale à : **6 σ**

Figure 3 :



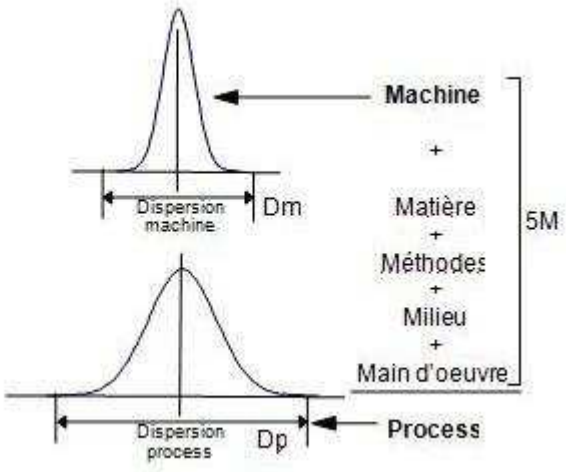
1.1.3. Notion de capabilité

La capabilité est la mesure établissant le rapport entre la performance réelle d'une machine ou d'un process et la performance demandée. On l'exprime par le rapport suivant :

$$Capabilité = \frac{Intervalle\ de\ tolérance}{Dispersion\ del\ a\ machine}$$

<u>Différence entre une Capabilité court terme et une Capabilité long terme :</u>

Figure 4 :



Il est nécessaire de dissocier 2 types de dispersion :

1. **La dispersion court terme (machine)** notée *Dm*, observée pendant un très court instant est liée uniquement à la machine et ses causes aléatoires de variation attribuables au hasard.
2. **La dispersion long terme (process)** notée *Dp*, est observée sur le process pendant un temps suffisament long pour que les ***5M*** (Machine, Méthode, Matière, Main d'oeuvre, Milieu) aient une influence. Ce sont les causes assignables de variation.

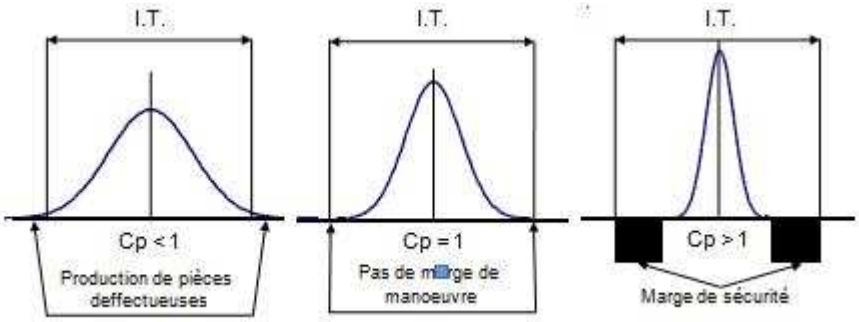
La différence entre la capabilité machine et la capabilité process provient uniquement de la manière d'estimer la dispersion :

$$Cm = \frac{I.T.}{Dm}$$

$$Cp = \frac{I.T.}{Dp}$$

Avec *I.T.* = Intervalle de tolérance

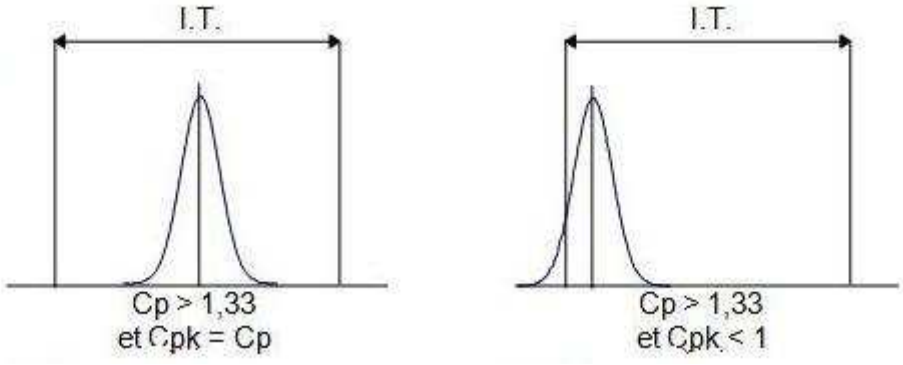
Figure 5 :



Indicateurs de déréglage Cmk et Cpk :

Les premiers indicateurs Cm et Cp ne donnent pas une information suffisante pour affirmer que l'on ne produit pas de pièces mauvaises. En effet, comme le montre la figure ci-dessous, il est possible, malgré un Cp acceptable, de produire des pièces hors tolérances.

Figure 6 :



De nouveaux indicateurs, les indicateurs de déréglage, vont nous permettre de mesurer la capabilité intrinsèque et le déréglage.

$$Indicateur\ de\ déréglage = \frac{Distance\ (\ Moyenne\ ;\ Limite\ la\ plus\ proche)}{1/2.\ Dispersion}$$

$$Cpk = \min\left[\frac{\left(L_s - \bar{\bar{X}}\right)}{3\hat{\sigma}_s}; \frac{\left(\bar{\bar{X}} - L_i\right)}{3\hat{\sigma}}\right]$$

D'un point de vue théorique, on accepte un lot lorsque le Cp et Cpk sont supérieurs à 1. Compte tenu de l'incertitude sur nos données et de l'effectif prélevé, une marge de sécurité représentant jusqu'à 3 écarts types est tolérée. Un Cp machine pourra avoir comme seuil 2 et son Cpk 1.66 alors qu'un Cp process aura un seuil de 1.33 et son Cpk 1.

Du fait des causes assignables de variation qui ne sont pas prises en compte lors de calcul d'une capabilité machine, nous serons plus exigent lors d'acception d'une capabilité pour celle-ci.

Le calcul de capabilités est permis seulement en cas de normalité de nos données. Ainsi nous avons à notre disposition un échantillon de données. Pour déterminer si la loi de notre jeu de données suit une loi normale nous effectuons un test de normalité. Le risque α pris dans tous nos cas est de 5% . Le test effectué lors d'un échantillon d'effectif 50 ou moins est celui de **Shapiro-Wilk**, et pour un effectif de 51 ou plus est utilisé **le test d'adéquation à une loi normale du Khi 2**. (Plus d'explication dans la partie **1.2 Test de normalité**).

En cas d'acceptation de la normalité, nous effectuons le calcul de capabilité en estimant la moyenne et l'écart type avec $\overline{X}$ et σ.

1.1.4. Présentation fiche 7

La fiche 7 est le départ de la problématique. Cette fiche permet de calculer la capabilité lorsque notre jeu de données suit une loi normale après l'application du test de normalité.

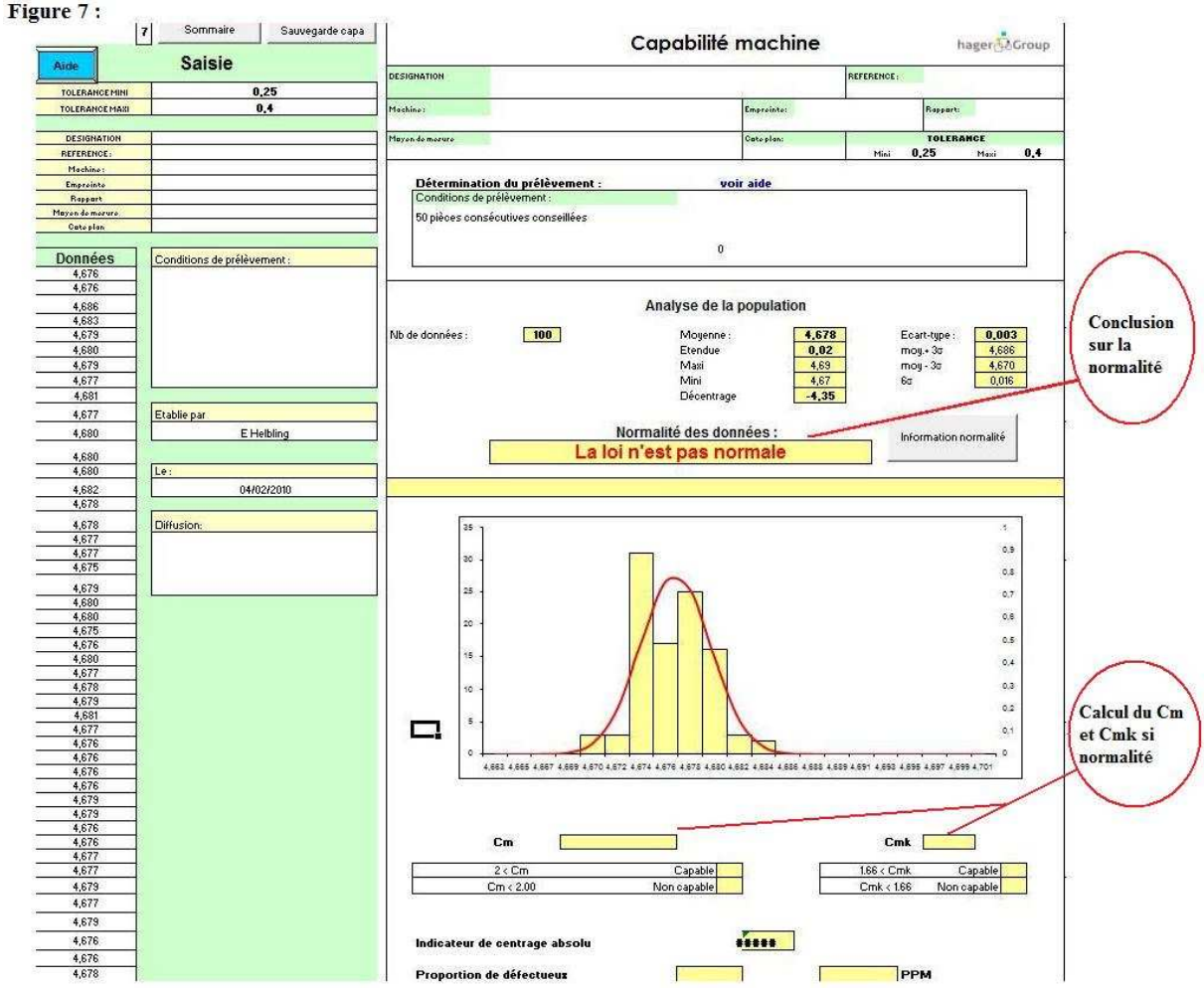
Il y a **3 parties distinctes** sur cette fiche. Les 2 premières étant la partie visible de l'utilisateur (voir figure 7).

La **1ère partie** sur la zone de gauche représente la saisie des données, ainsi que celles des tolérances et des informations sur le produit étudié (référence, machine , …..) . Des liens pour retourner au sommaire ou pour avoir de l'aide y sont présents aussi.

La **2ème partie** situé à droite est protégée est non modifiable par l'utilisateur. Nous y avons le descriptif du produit, une analyse des données (moyenne, écart type, maximum, ….), une conclusion sur la normalité, l'histogramme représentatif de notre jeu de données, la capabilité si la loi est **Normale** et une proportion de défectueux suivie du **PPM**. Ces 2 derniers représentent la proportion estimée de défectueux de notre machine en fonction de nos tolérances ainsi que le **PPM** (partie par million) qui est une expression plus simple pour visualiser notre proportion défectueux. 0,27 % de produits défectueux sont tolérés. Même valeur utilisé lors du contrôle 6 sigma. Ceci vaut un PPM de 2700.

La **3ème partie** est cachée de l'utilisateur. Elle est située en général sur la partie droite des 2 premières partie. Elle peut aussi y être mis en dessous. C'est dans cette partie que la programmation des fonctions qui permettent la conclusion sur la 2ème partie y est calculée.

Figure 7 :



1.1.5. Objectif du stage :

La problématique et sujet du stage est « ***Comment déterminer les capabilités lorsque la loi n'est pas normale ?*** »

Dans 10% des cas nous avons un refus de la normalité. Que faire dans ce cas présent ? D'où vient le problème ? Comment le contourner ?

L'objectif est de déterminer les différentes méthodes statistiques qui permettront d'exploiter notre jeu de données et de pouvoir conclure sur la capabilité ou non d'une machine. Hors cette recherche de résolution de la non normalité, la modification du **Guidestat** et une très bonne maitrise d'Excel seront nécessaires.

Le but étant pour l'utilisateur du **Guidestat** de pouvoir traiter et conclure sur les données d'un échantillon lorsque celui-ci **ne suit pas une loi normale**.

Le rapport de stage mettra en lien les recherches sur le sujet ainsi que sa programmation dans le **Guidestat**. La création de plusieurs fiches traitant les différents cas seront accessible depuis cette fiche 7. Ces fiches respecteront le principe des différentes parties.

1.2. TESTS DE NORMALITE ET LIEN DANS LE GUIDESTAT

Les tests de normalité sont à l'origine du calcul des capabilités. Ce sont eux qui décident si oui ou non nous allons calculer une capabilité. De ce fait, un approfondissement de ces tests semble logique. La compréhension et la mise en forme de ces tests sous Excel est importante dans la suite du sujet.

Les 2 tests utilisés comme énoncé « dans l'introduction » sont celui de **Shapiro et du Khi 2**. De nombreux d'autres tests existent, et un intérêt particulier pour 2 autres tests qui sont celui **d'Anderson-Darling** et celui **de Kolmogorov-Smirnov** fut entreprit. Chacun ayant sa méthode de calcul et sa puissance. La puissance étant l'erreur d'un test lorsque la normalité est acceptée.

1.2.1. Le test de Shapiro

Le test de Shapiro est basé sur une statistique T dépendant des données de constantes générées à l'aide de la moyenne et de la matrice de variance covariance d'un échantillon de taille n suivant une loi normale. Ces constantes sont trouvables dans des tables spécifiques. La région critique du test en fonction du seuil α se retrouve aussi à l'aide de table.

Ce test est très réputé pour des effectifs inférieurs à 50. Existant déjà dans le **Guidestat** sous forme de macro dans **Visual basic**, ce test n'a pas été traité.

L'appel de cette macro qui renvoie VRAI si on accepte la normalité ou FAUX sinon se fait avec :

```
Shapirowilk(M ; α);
```

- M étant la plage de données dont nous voulons tester la normalité
- α étant le risque toléré

1.2.2. Le test du Khi 2

Ce test est celui utilisé pour un échantillon supérieur à 50. Ce test peut différer en fonction de sa conception. En effet, avec un jeu de données identiques, nous pouvons avoir des résultats différents. Pourquoi ? Ce test prend un jeu de données et le divise en plusieurs classes. Ce nombre de classes est déterminé dans notre cas avec la formule :

$$K = E\left(\frac{1 + 10 * LN(n)}{3}\right)$$

E désignant la partie entière
n étant le nombre de données

Il est souvent utilisé aussi la formule :

$$K = \sqrt{n}$$

L'étendue de l'échantillon est divisée par ce nombre K, nous obtenons ainsi le **pas de l'échantillon**. Nous calculons le nombre d'éléments ***ni*** présents dans **chaque classe**.

Grâce à la moyenne estimée et l'écart type estimé nous calculons la probabilité ***pi*** qu'une donnée appartienne à cette classe si elle suivait une **loi normale**. Ensuite chaque *pi* est multiplié par n indiquant **le nombre de données théoriques** devant être présent.

Avec la formule $\frac{(ni - n*pi)^2}{n*pi}$ nous obtenons pour chaque classe un nombre qui plus le nombre d'éléments *ni* diffère du nombre attendu *n*pi*, plus sera grand.

La **statistique du test** *X* vaudra la somme de ces nombres pour chaque classe.

$$X = \sum_{i=1}^{K} \frac{(ni - n * pi)^2}{n * pi}$$

Cette statistique suivra une **loi du X^2 à K-1 degré de liberté**.
Pour trouver la valeur critique du test au seuil $\alpha = 0.05$, la fonction Excel est :

Khideux.inverse(0.05 ; K-1) ;

Test du Khi 2 dans le Guidestat :

Figure 8 :

Nombre de classes

Max	Min	Etendue	Ektp	pas	minimhisto	Statistique du test du Khi 2
0,356	0,304	0,0522	8	0,006525	0,2775	11,88805794
						Valeur critique Khi deux
						14,06714043

Test du Khi 2

classes	intervalle(=ni)	Pi	n*Pi	(ni-n*Pi)^²/n*Pi	
0,278	0	0,000	0,00021719	0	
0,284	0	0,000	0,00400552	0	
0,291	0	0,000	0,04857214	0	
0,297	0	0,004	0,38767589	0	33
0,304	5	0,020	2,03856176	4,302109751	
0,310	4	0,071	7,06854584	1,332094857	
0,317	13	0,162	16,1732842	0,622615218	
0,323	22	0,244	24,4314925	0,241989136	
0,330	32	0,244	24,3725328	2,387041847	
0,336	18	0,161	16,05645	0,235256652	
0,343	3	0,070	6,98361099	2,272342565	
0,349	3	0,020	2,00433051	0,494607915	
0,356	0	0,004	0,37931996	0	
0,362	0	0,000	0,04729453	0	
0,369	0	0,000	0,00388119	0	
0,375	0	0,000	0,00020942	0	
0,382	0	0,000	7,4226E-06	0	
0,388	0	0,000	1,7264E-07	0	
0,395	0	0,000	2,6324E-09	0	33
0,401	0	0,000	2,629E-11	0	
0,408	0	0,000	1,7764E-13	0	
21	100				

Emplacement de la limite de tolérance inférieure

Emplacement de la limite de tolérance supérieure

Ce test nous permet aussi de créer l'histogramme présent dans la fiche de calcul de capabilité (avec *ni*). Sous Excel, cette histogramme ne se crée pas automatiquement et nous devons spécifié les informations de notre graphique. Il en est de même pour la création de la courbe représentant la loi normale (avec *pi*) et des limites de tolérances représentées sous forme de bâton.

1.2.3. Test de Kolmogorov-Smirnov avec correction de Lilliefors

Ce test est sensible aux données autour de la partie centrale de la distribution. Il a plutôt un coté historique, il permet surtout de confirmer les autres tests. On ne se servira **JAMAIS** seulement de celui-ci pour confirmer une hypothèse de normalité dans le **Guidestat.**

Il est basé sur la différence entre la fonction de répartition empirique et la fonction de répartition théorique évalué avec notre jeu de données.

$$D^+ = \max_{i=1..n}\left(\frac{i}{n} - F_i\right)$$

$$D^- = \max_{i=1..n}\left(F_i - \frac{i-1}{n}\right)$$

$$D = \max\left(D^+ ; D^-\right)$$

F_i étant la fonction de répartition de loi normale de moyenne $\bar{X}$ et d'écart type σ au point x_i.
x_i étant la i-ème donnée lorsque les données sont triées.
D est la valeur pratique. Elle doit être plus petite que la valeur théorique associé au nombre de classes.

La **correction de Lilliefors** permet d'augmenter la puissance du test lors d'une hypothèse d'adéquation à une loi normale. Elle permet de trouver la valeur théorique *c* associée au risque $\alpha = 0.05$ qui est trouvée à l'aide d'une table (voir Annexe) lorsque l'effectif de l'échantillon est inférieur à 40. Pour un effectif supérieur la valeur théorique vaut

$$c = \frac{0.895}{f(n)}$$

avec

$$f(n) = \frac{0.83 + n}{\sqrt{n}} - 0.01$$

La région de rejet de la normalité est lorsque

$$D > c$$

Test du Kolmogorov-Smirnov dans le Guidestat :

Figure 9 :

moyenne	ecart type	nb valeurs
0,330	0,0111573	100

Statistique du test de K-S	
0,105099301	
Valeur critique du test	**Lilliefors**
0,136	0,0888514

Représente la valeur critique du test de K-S

Représente la valeur critique corrigée

Test de Kolmogorov Smirnov

Valeurs	num'	classement	Xi	Fi	D+	D-	Max(D+,D-)
0,3354	1	0,3016	-2,584496101	0,00487607	0,00512	0,004876	0,00512393
0,3088	2	0,3033	-2,432129498	0,00750517	0,01249	-0,00249	0,01249483
0,3197	3	0,3039	-2,378353049	0,008695083	0,0213	-0,0113	0,021304917
0,3329	4	0,3088	-1,939178722	0,02623979	0,01376	-0,00376	0,01376021
0,3582	5	0,3097	-1,85851405	0,031548025	0,01845	-0,00845	0,018451975
0,3443	6	0,3099	-1,840588567	0,032840937	0,02716	-0,01716	0,027159063
0,3262	7	0,3105	-1,786812119	0,03698393	0,03302	-0,02302	0,03301607
0,358	8	0,3108	-1,759923895	0,039210356	0,04079	-0,03079	0,040789644
0,3328	9	0,3108	-1,759923895	0,039210356	0,05079	-0,04079	0,050789644

On peut constater dans ce cas que le test de K-S accepte la normalité alors qu'aprés correction nous refusons la normalité

1.2.4. Test d'Anderson-Darling

Ce test est sensible aux données en queues de distribution. Sa statistique A est

$$A = n - \frac{1}{n}\sum_{i=1}^{n}[(2i-1)\,(\log(F_i) + \log(1 - F_{n-i+1}))^2]$$

Une correction a été proposé par Stephens :

$$A^* = A\left(1 + \frac{0.75}{n} + \frac{2.25}{n^2}\right)$$

F_i étant la même fonction que le test de Kolmogorov-Smirnov
La valeur critique c au niveau de risque $\alpha = 0.05$ est dans tous les cas 0.752.
La région de rejet de la normalité est lorsque

$$A^* > c$$

Test d'Anderson-Darling dans le Guidestat :

Figure 10 :

Somme colonne AC
-10094,2242

Statistique du test d'Anderson
0,942242045
Statistique modifiée
0,949520865
Valeur critique du test au seuil 0,05
0,752

Test d'Anderson Darling

ln(Fi)	classement ii	Xi inverse	1-F(n-i+1)	ln(1-F(n-i+1))	(2i-1)*(AB+X)
-5,4250634	0,3558	2,62699523	0,00430713	-5,44748418	-10,8725476
-5,1065281	0,3542	2,46618751	0,00682799	-4,98672474	-30,2797586
-4,9934619	0,351	2,14457205	0,01599354	-4,13557043	-45,6451618
-4,9654252	0,347	1,74255273	0,04070589	-3,20138257	-57,1676545
-3,822052	0,3463	1,67219935	0,04724251	-3,05246118	-61,870619
-3,5405146	0,3446	1,50134114	0,06663367	-2,70854521	-68,7396578

1.2.5. Conclusion sur les tests

Pourquoi avoir fait 4 tests alors qu'un seul suffit ? Chaque test est spécifique. Nous pouvons vite rejeter une hypothèse de normalité avec un seul test si une seule donnée interfère sur celui-ci.

En cas de contradiction des 4 tests, si :

1. Nombre de données <= 50

Seul le **test de Shapiro** est pris en compte.

2. Nombre de données > 50

Le test de Shapiro n'est plus pris en compte (max 50 valeurs).

- Si seul le **test du Khi 2** accepte la normalité alors on accepte la normalité.
- Si seul le **test d'Anderson – Darling** accepte la normalité
- Si seul le **test de Kolmogorov – Smirnov** accepte la normalité alors on **refuse** la normalité.

L'histogramme permet de voir facilement la tendance de nos données.
Nous remarquons que le **test de Kolmogorov-Smirnov** ne sert jamais dans la conclusion de la normalité. Ce test ayant été étudié durant le stage et mis en place dans le **Guidestat**, une partie le concernant est justifié.

Dans la fiche 7, un lien **Information Normalité** est présent à coté de la conclusion du test.
Ce lien nous envoie vers une nouvelle fiche expliquant l'utilisation des différents tests. Plus de précision dans **l'annexe**.

PARTIE 2 :
RESOLUTION DE LA PROBLEMATIQUE

2.1. POINTS ABERRANTS

La première question à se poser sur la non normalité de nos données est « ***Existe-t-il des points qui peuvent faussés la normalité des données ?*** »

En effet, quelques données peuvent fausser l'interprétation de la normalité de l'échantillon, du fait d'erreurs de mesures, de saisies, et donc rejeter le test de normalité.
Une recherche de points aberrants sera réalisée pour les données.

Si on trouve des données aberrantes, une réflexion sur notre jeu de données est à faire :

- Veut-on inclure les points aberrants dans nos données ?
- Ou décide-t-on de rejeter tous ou seulement certains points, qui paraissent aberrants et ainsi faire un test de normalité sur le nouveau jeu de données ?

Ces points sont aberrants **seulement** si nos données deviennent normales après les avoir enlevées.

Il est évident qu'effectuant un test de normalité qui réfute cette hypothèse, nous ne pouvons pas appliquer ce principe et nous devons réintégrer ces points sauf en cas d'erreur de saisie.

Ainsi seront définis des points aberrants seulement si le test de normalité des nouvelles données est positif.

2.1.1. Décision à prendre

Malgré une normalité sans les points aberrants, peut-on ôter ces points du processus?

⇨ **Oui si :**

- Il y a une faute de frappe, (on peut le corriger directement dans la saisie des données).
- Nous avons une cause connue et corrigée dans le process.

On peut conclure à la normalité (si présente), et étudier la capabilité.

⇨ **Non si :**

- Nous ne trouvons pas de cause spéciale qui explique les points aberrants

2.1.2. Principe de calcul

On pourrait déterminer des points aberrants en rejetant tous points se situant à ± 3 écart-types. Le problème est qu'en fonction du nombre de données, la probabilité est vite différente et non stable.

Ainsi on va s'intéresser à l'écart inter-quartile 'IQ' qui est l'écart entre le premier quartile 'Q1'(1 quart de l'échantillon) et le troisième quartile 'Q3' (3 quarts de l'échantillon).

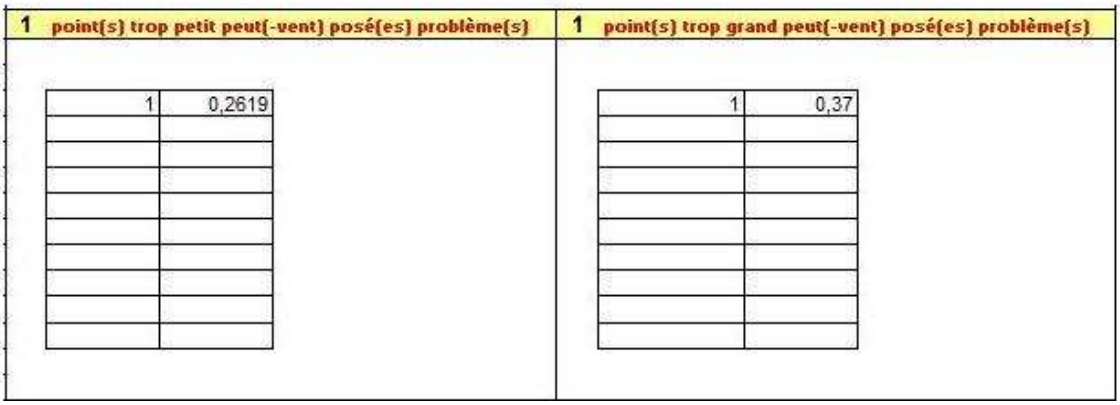
Tous points inférieurs à Q1-1.5*IQ et supérieurs à Q3+1.5*IQ seront considérés comme aberrant.

Cette technique de recherche de points aberrants est la même que celle proposée dans une boite à moustache.

2.1.3. Points aberrants dans le Guidestat

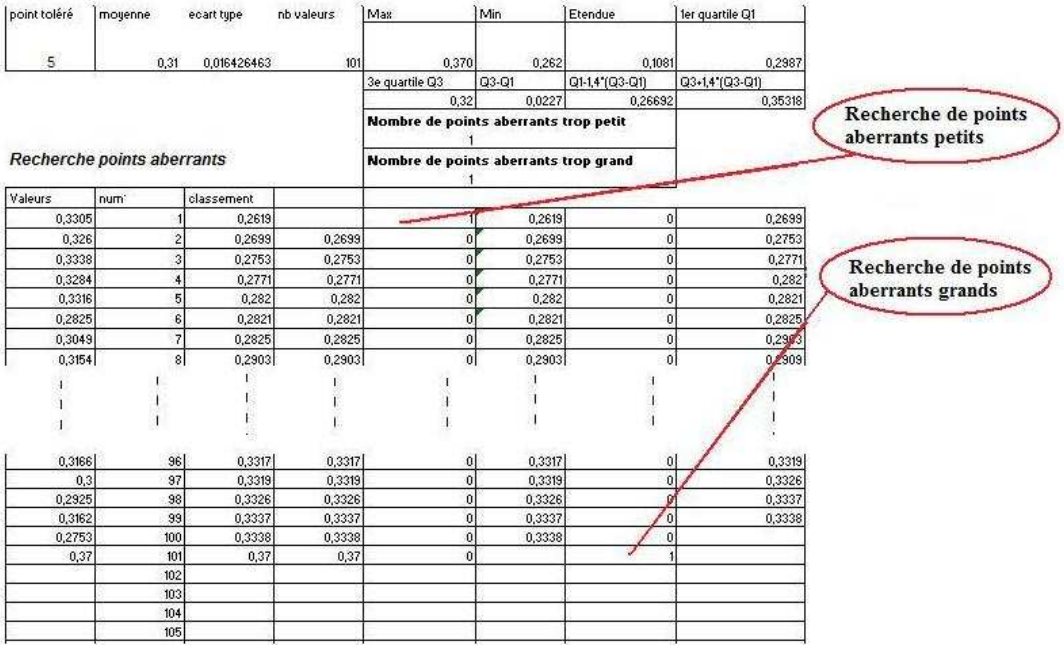
La résolution des points aberrants se fait sur une fiche du Guidestat à part.
On peut y retrouver (voir figure 11) les points aberrants sur les 2 extrémités des queues de la distribution.

Figure 11 :



On se limite également à 5% max de points aberrants de chaque coté de la distribution. Ces points sont trouvés grâce à des formules toujours cachées sur la partie droite non visible de la feuille. (figure 12 ci dessous)

Figure 12 :



point toléré	moyenne	ecart type	nb valeurs	Max	Min	Etendue	1er quartile Q1
5	0,31	0,016426463	101	0,370	0,262	0,1081	0,2987
				3e quartile Q3	Q3-Q1	Q1-1,4*(Q3-Q1)	Q3+1,4*(Q3-Q1)
				0,32	0,0227	0,26692	0,35318
				Nombre de points aberrants trop petit			
				1			
Recherche points aberrants				**Nombre de points aberrants trop grand**			
				1			

Valeurs	num°	classement					
0,3305	1	0,2619		1	0,2619	0	0,2699
0,326	2	0,2699	0,2699	0	0,2699	0	0,2753
0,3338	3	0,2753	0,2753	0	0,2753	0	0,2771
0,3284	4	0,2771	0,2771	0	0,2771	0	0,282
0,3316	5	0,282	0,282	0	0,282	0	0,2821
0,2825	6	0,2821	0,2821	0	0,2821	0	0,2825
0,3049	7	0,2825	0,2825	0	0,2825	0	0,2903
0,3154	8	0,2903	0,2903	0	0,2903	0	0,2909
...	...	...	...	...	...	...	...
0,3166	96	0,3317	0,3317	0	0,3317	0	0,3319
0,3	97	0,3319	0,3319	0	0,3319	0	0,3326
0,2925	98	0,3326	0,3326	0	0,3326	0	0,3337
0,3162	99	0,3337	0,3337	0	0,3337	0	0,3338
0,2753	100	0,3338	0,3338	0	0,3338	0	
0,37	101	0,37	0,37	0		1	
	102						
	103						
	104						
	105						

Une conclusion sur la normalité est présente sans les points aberrants. Elle y est visible dans cette fiche **« points aberrants »** avec la même présentation que la fiche de calcul de capabilité d'origine. Ensuite, c'est à l'utilisateur de voir comment il va utiliser le résultat.

2.2. DEFAUT DE FORME

Généralement notre moyen de production montre qu'un produit suit une **Loi normale**. Dans certains cas notre moyen de mesure ne permet pas d'avoir cette finalité. On constate que notre moyen de production satisfait une autre loi appelée une loi de **défaut de forme**.

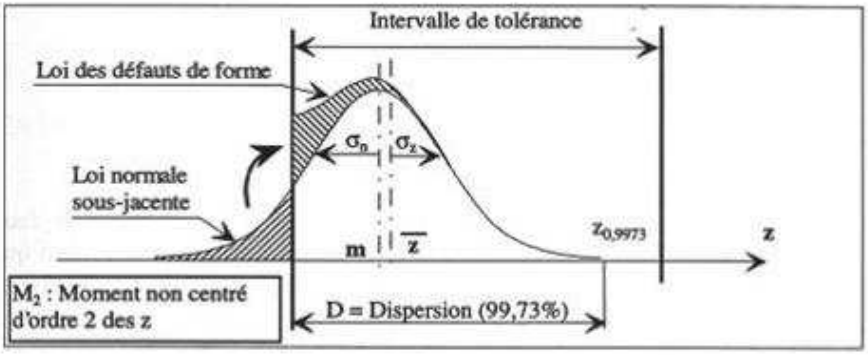
Il est important de savoir quelles données nous traitons. Un aperçu de quelle loi utilisé est donné par le tableau ci-dessous.

TYPE	Loi utilisée
Rectitude	Loi défaut de forme
Planéité	Loi défaut de forme
Circularité	Loi défaut de forme
Cylindricité	Loi défaut de forme
Forme d'une ligne quelconque	Loi défaut de forme
Forme d'une surface quelconque	Loi défaut de forme
Parallélisme	Loi défaut de forme
Perpendicularité	Loi défaut de forme
Inclinaison	Loi défaut de forme
Battement simple	Loi défaut de forme
Symétrie	Loi défaut de forme
Rugosité	Loi défaut de forme
Concentricité et coaxialité	Loi de Rayleigh
Battement total	Loi défaut de forme ou Loi de Rayleigh
Localisation ou position	Loi défaut de forme ou Loi de Rayleigh
Mesure de longueur, largeur, épaisseur	Loi normale
Moment de torsion	Loi normale

2.2.1. Principe

Cette loi est par définition la **valeur absolue de la différence entre 2 lois normales**.
Statistiquement, la **différence entre 2 lois normales indépendantes suit une loi normale**. La valeur absolue vient du fait qu'on mesure toujours une grandeur **positive** (un maxi – un mini en général). Ainsi cette loi dépend de la loi sans la valeur absolue, qu'on appelle la **loi normale sous-jacente**. La probabilité de la loi normale sous jacente inférieure à 0 est ajoutée symétriquement à 0 pour obtenir une **loi de défaut de forme**. On peut le visualiser sur le graphique suivant :

Figure 13 :



La **loi de Rayleigh** que nous pouvons voir dans le tableau a été étudiée lors du stage. Il y a une confusion avec la **loi de défaut de forme**. Le test adapté au défaut de forme que nous verrons plus tard

acceptera une loi de Rayleigh. Le cas de données suivant une **loi de Rayleigh** sera traité comme un défaut de forme. La capabilité ne variera que très légèrement et n'aura pas d'incidence.

on pose : - z = moyenne de l'échantillon de la loi défaut de forme
- s1 = écart type de l'échantillon de la loi défaut de forme
- m = moyenne de l'échantillon de la loi normale sous jacente
- s = écart type de l'échantillon de la loi normale sous jacente
- m2 = moyenne des carrés de l'échantillon
- D = dispersion (= 99,73% des données inférieures à ce nombre)

La densité de cette loi est :

$$\boldsymbol{f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\left(\frac{x-m}{s}\right)} + \frac{1}{\sigma\sqrt{2\pi}} e^{\left(\frac{x+m}{s}\right)} \qquad avec\ x > 0}$$

z et s1 sont estimés avec les données recueillies. Ensuite nous calculons le rapport $\frac{z}{\sqrt{m2}}$ que nous notons Ko, et grâce à des tables (voir **Annexe**) nous trouvons 2 nouvelles valeurs K1 et K2.
Celles-ci nous permettent de trouver une estimation de m et s, la moyenne et l'écart type de la loi normale sous jacente.

$$\boldsymbol{m = K1 * \sqrt{m2}}$$
$$\boldsymbol{s = K2 * \sqrt{m2}}$$

Ko nous permet maintenant de distinguer 3 cas.

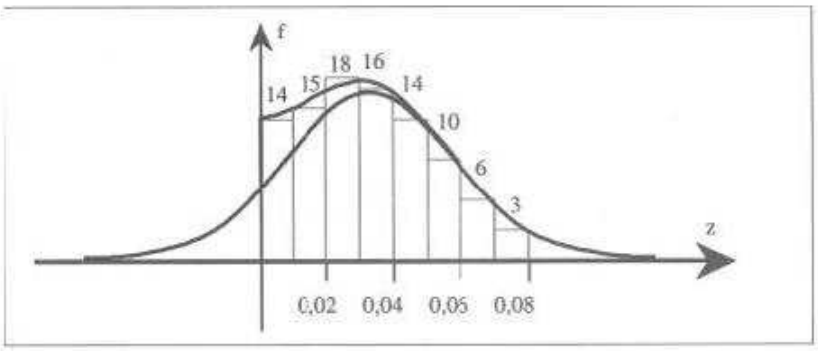
2.2.2. 3 cas à distinguer

1er cas : 0.7978 < Ko < 0.9 alors la moyenne m est supérieure à 0. La dispersion est :

$$\boldsymbol{D = R * s1}$$

Avec *R* un nombre trouvé grâce à $\frac{m}{s}$ et les coefficients d'une table (voir **Annexe**).

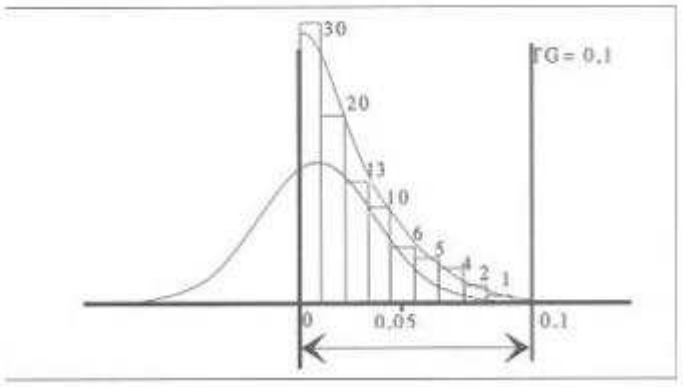
Figure 14 :



2ème cas : Ko < 0.7978. Plus on est proche de 0, plus il y a de probabilités d'avoir une donnée proche de 0(figure 6). m est considérée valant 0 ou très proche de 0. La dispersion est calculée avec la formule :

$$D = 2,96 * \sqrt{m2}$$

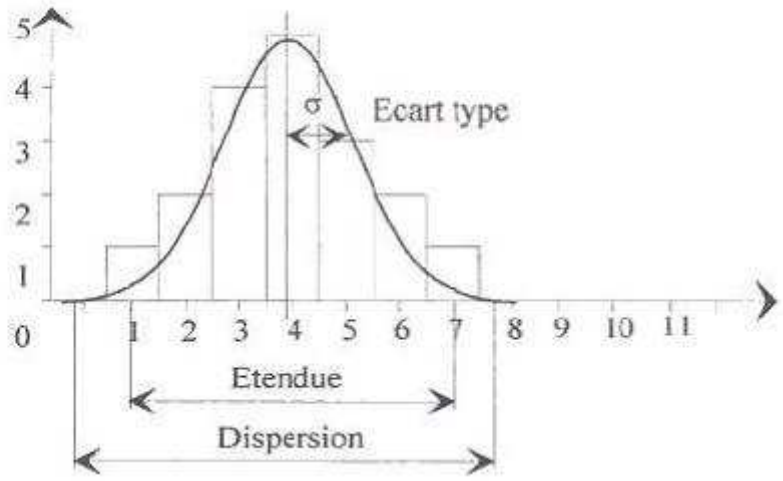
Figure 15 :



3ème cas : Ko > 0.9. La loi de défaut de forme est équivalente à une loi normale. La dispersion est calculée avec la formule **'D=m+2.78s'** et non **'m+3s'** car 3 étant le quantile pour un risque à 0.135% et 2.78 ici représente le quantile pour un risque à 0.27%. On ne s'intéresse plus qu'au risque supérieur et plus des 2 cotés vu que plus on est proche de 0, mieux notre défaut sera acceptable. Notre dispersion vaudra :

$$D = m + 2.78\,s$$

Figure 16 :



2.2.3. Test du Khi 2 adapté pour le défaut de forme

Le test d'adéquation à une loi de défaut de forme avec le Khi 2 est du même principe que pour la loi normale. La différence se fait lors du calcul des *pi*. Le calcul de la probabilité d'un intervalle $[\boldsymbol{a}, \boldsymbol{b}]$ $\boldsymbol{avec}\ \boldsymbol{a}, \boldsymbol{b} > 0$ vaut

$$\boldsymbol{pi} = \boldsymbol{P_{df}}([\boldsymbol{a}, \boldsymbol{b}]) = \boldsymbol{P_{ln}}([\boldsymbol{a}, \boldsymbol{b}]) + \boldsymbol{P_{ln}}([-\boldsymbol{b}, -\boldsymbol{a}])$$

$\boldsymbol{P_{ln}}$ étant la probabilité d'une loi normale.
Nous utilisons la moyenne m et l'écart type s de la loi normale sous jacente.
La fonction Excel permettant le calcul de $\boldsymbol{P_{ln}}([\boldsymbol{a}, \boldsymbol{b}])$ de moyenne u et d'écart type s est :

1-Loi.normale(a ; u ; s ; VRAI) + Loi.normale(b ; u ; s ; VRAI)

La fonction **Loi.normale(a ; u ; s ; VRAI)** suivi de VRAI permet le calcul de la fonction de répartition au point a. En y mettant FAUX, on calcul l'image de la fonction de densité au point a.

2.2.4. Capabilité

Une loi de défaut de forme est par définition sans tolérance inférieure (plus on a des données proches de 0, mieux sont conformes nos données) . Ainsi le calcul de la capabilité se calcule avec la tolérance supérieure. On peut calculer un Cm et Cmk, mais les 2 sont en lien. Un Cm > 1 impliquera un Cmk > 1, et de même réciproquement.

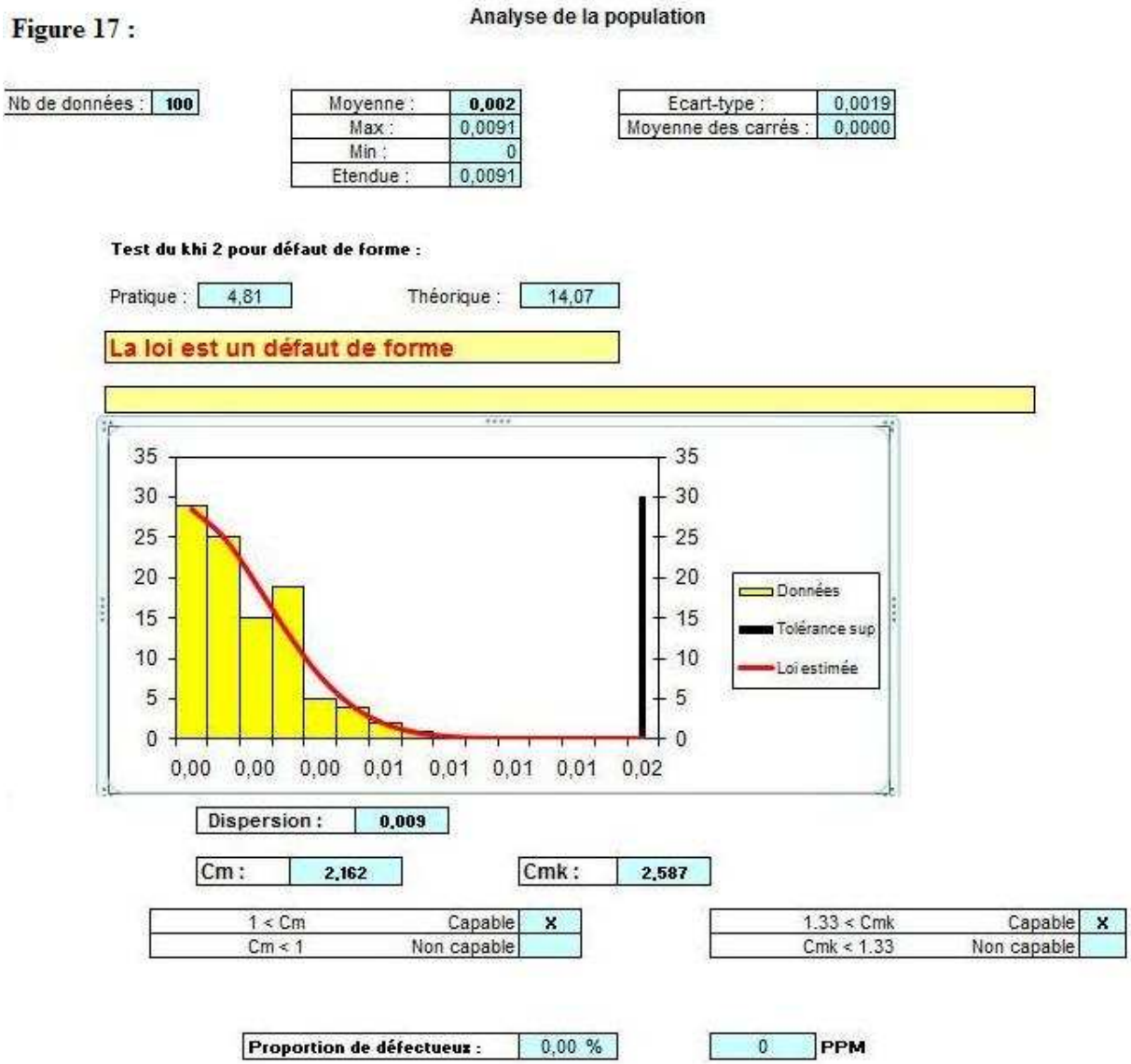
On a :

Cm= D/TS et **Cmk= (D-z)/(TS-z)**

Avec TS = tolérance supérieure
La capabilité est **acceptable** lorsqu'elle est supérieure à **1** pour le Cm et **1.33** pour le Cmk.

2.2.5. Défaut de forme dans le Guidestat

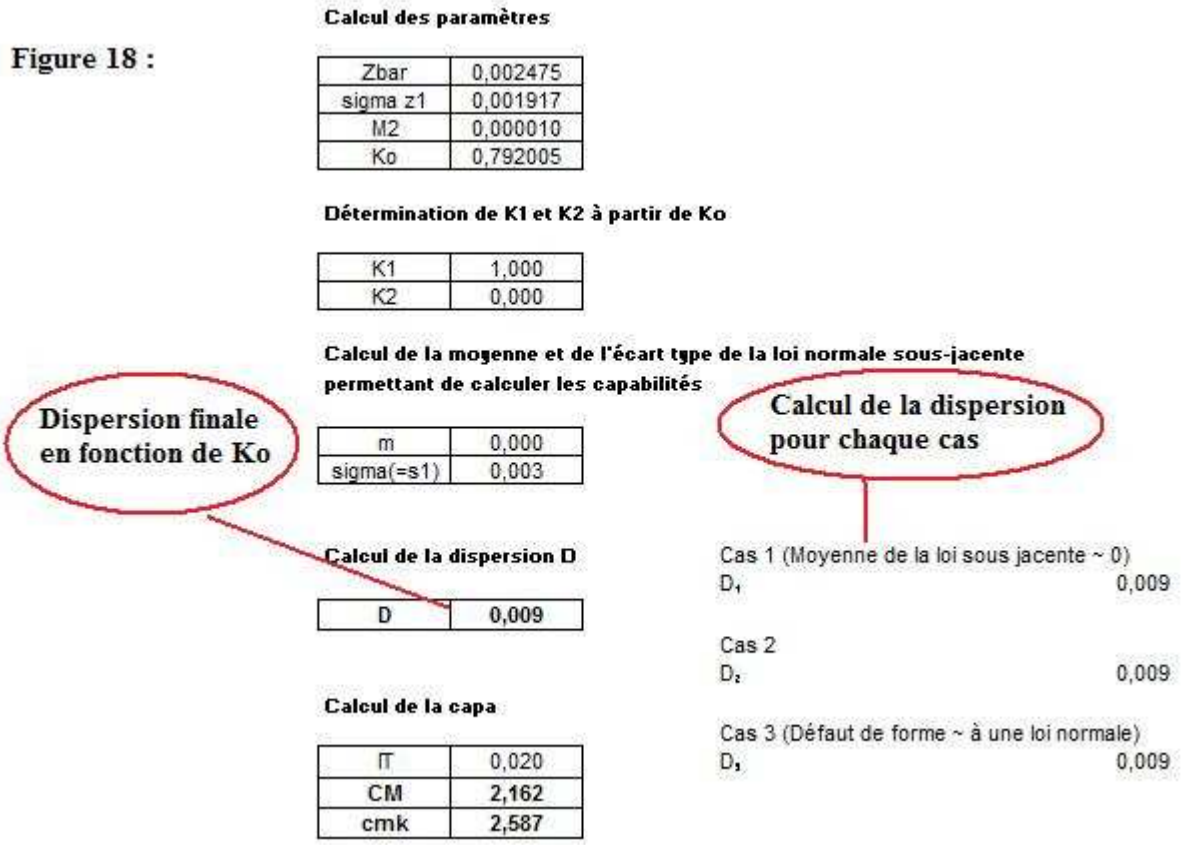
Figure 17 :



La figure 17 représente la 2ème partie visible de la fiche propre au défaut de forme. Elle ressemble à la fiche d'origine avec le résultat du test du Khi adapté. La 1ère partie de cette fiche est la zone de saisie où nous trouvons les données et la tolérance supérieure.

La figure 18 qui suit est la 3ème partie non visible de la fiche. Il y figure les calculs pour déterminer la dispersion et la capabilité.

Figure 18 :



Il y figure aussi dans cette 3ème partie les tables qui sont données en **Annexe.**

2.3. LOI BIMODALE

Cette partie du sujet est une partie personnelle. Aucune référence n'a été utilisée lors de cette partie. L'utilisation vient de la constatation de mélange de populations dans nos données. En effet, dans certains échantillons, nous remarquons à l'aide du graphique qu'une **double bosse** pouvait être présente. Par exemple, si nous voulions établir la moyenne du poids d'une population mixte, nous pourrions déceler une **double bosse** du fait du mélange des hommes et des femmes.

En pratique, lorsque nous avons des données provenant d'un moule fabriquant 2 produits, et par incapacité à séparer les mesures lors du contrôle, nous obtenons souvent une loi bimodale.

2.3.1 Objectif

Notre but est de déterminer la normalité et la capabilité des données en cas de mélange de 2 populations en prenant des moyennes de 5 ou 10 données.

Les moyennes de données d'une loi bimodale suivent une loi normale. Cette conclusion n'a pas été prouvée mathématiquement. Une preuve **'physicienne'** avec un logiciel simulant la situation permet de confirmer cette hypothèse.

Le problème est qu'un effectif de 30 données, donnera de nouvelles données d'effectif $\frac{30}{5}$ = **6** et $\frac{30}{10}$ = **3**, qui est insuffisant pour un **test de Shapiro**. Par contre, pour un effectif de 50 à 90, on peut utiliser le test de normalité pour une moyenne de 5. Pour une moyenne de 10, un effectif minimum de 100 données est requis.

Conditions d'utilisations :

- Effectif >= 50 pour une moyenne de 5 données.
- Effectif >= 100 pour une moyenne de 10 données.
- Effectif <= 1000 pour tout échantillon.
- Effectif doit être un **multiple de 10**.

2.3.2 Calcul

Un échantillon représentant une moyenne de **l** données suivant une **loi normale** d'écart type **s** et de moyenne **m**, suivra une **loi normale** d'écart type $\frac{s}{\sqrt{l}}$ et de moyenne **m**. Une **loi bimodale** aura ainsi les mêmes propriétés lorsque **l** et l'effectif de l'échantillon sont assez grand. Nous gardons les mêmes notations.

Les tolérances sont modifiées aussi, la tolérance sera '*un sur racine de n fois moins loin*' par rapport à la moyenne. Ainsi

- $NTS = m + \frac{(TS-m)}{\sqrt{l}}$
- $NTI = m - \frac{(m-TI)}{\sqrt{l}}$

NTS = nouvelle tolérance supérieure
NTI = nouvelle tolérance inférieure
TS = ancienne tolérance supérieure
TI = ancienne tolérance inférieure
l = nombre de données pris pour établir la moyenne

2.3.3 Principe de l'écart type théorique et pratique

L'écart type théorique est l'écart type **s** du jeu de données initiales.
L'écart type pratique est l'écart type du nouveau jeu de données, appelons le **sbis**
La théorie nous dit que **sbis** = $\frac{s}{\sqrt{l}}$ si notre nouveau jeu de données suit une **loi normale**. La moyenne **m** est la même. Mais ceci est vrai mathématiquement lorsque notre échantillon est infiniment grand. Ainsi une **erreur** sera toujours présente lors du calcul des 2 écarts types.

On calcul l'erreur avec la formule :

$$\mathbf{E = ((\frac{s}{\sqrt{l}} / sbis)-1)*100}$$

E désigne donc l'erreur qui est un pourcentage pouvant être négatif ou positif.
Il représente le pourcentage d'éloignement de **sbis** **pratique** par rapport à $\frac{s}{\sqrt{l}}$ **théorique**.

- Un pourcentage entre **–10% et 10% sera considéré comme excellent** est représentatif d'une loi bimodale.
- Un pourcentage **< -10 % représentera une baisse de l'écart type pratique**, ainsi le jeu de données permettra un calcul de capabilité pertinent.
- Un pourcentage **> 10% représente une mauvaise répartition aléatoire** de notre jeu de données. C'est à dire que même en présence d'une bimodalité, mais d'un tri croissant (ou décroissant) des valeurs (cas extrême) l'écart type pratique sera trop élevé ne permettant pas le calcul pertinent de capabilité.

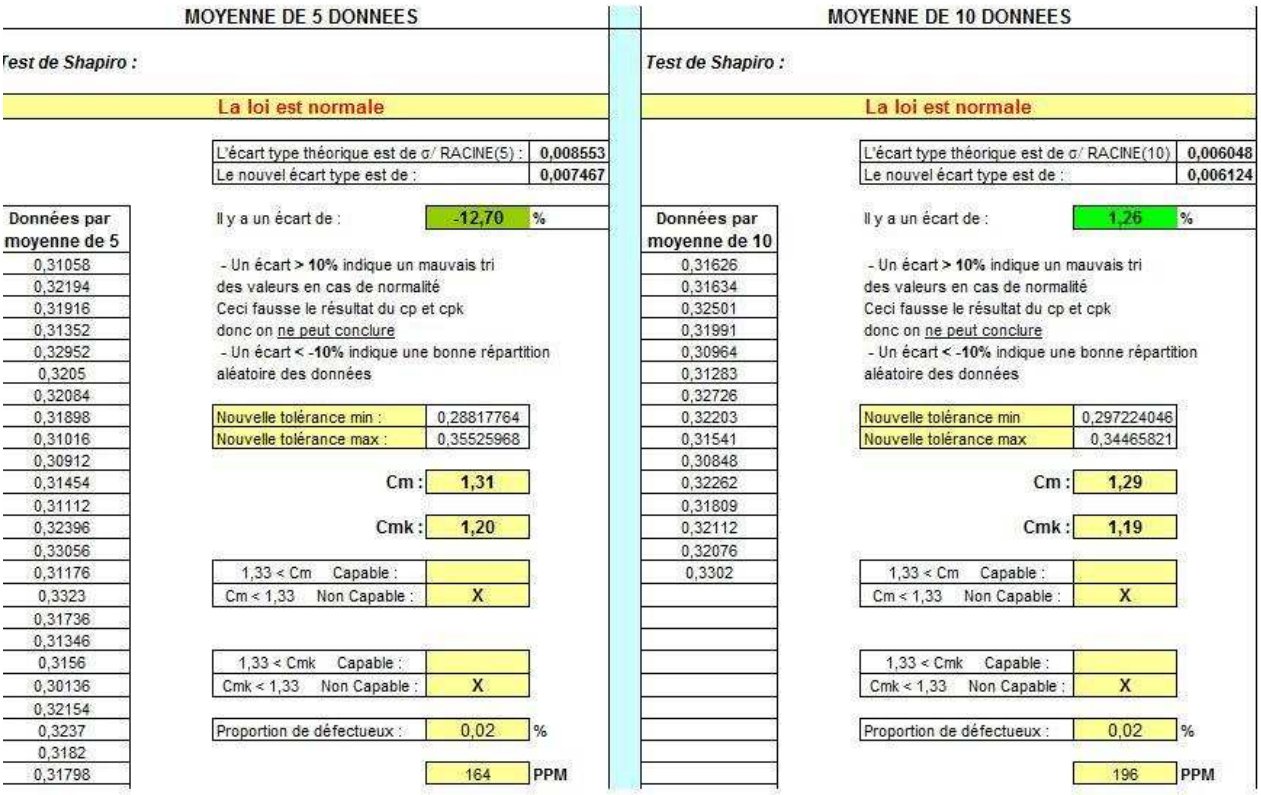
Ces conclusions sont à prendre avec beaucoup de précautions. L'utilisation de cette fiche se fera lorsque visuellement, le graphique nous permet de visualiser une **double bosse**. Aucun test n'est fait pour déterminer si la loi est **bimodale**.

La capabilité se calculera avec les nouvelles tolérances NTS et NTI, la moyenne **m**, et le plus grand des écarts types entre **sbis** et $\frac{s}{\sqrt{l}}$ pour minimiser l'erreur d'un produit non conforme.
Le calcul est le même ensuite que les capabilités habituelles.

2.3.4 La loi bimodale dans le Guidestat

La fiche de la **loi bimodale** a une 2ème partie visible légèrement différente adaptée pour cette situation. Nous pouvons y voir dans la figure 19 ci-dessous qu'elle est séparée en 2 partie. Une pour les moyennes par regroupement de 5 données et l'autre pour un regroupement de 10 données.

Figure 19 :



2.4. TRANSFORMATION NORMALISANTE

Le dernier cas traité, si nous ne pouvons toujours pas conclure sur notre jeu de données, est la **transformation normalisante.** Ces transformations sont à utiliser avec une très grande précaution. Leur utilisation est à faire lorsque la loi de notre jeu de données est dissymétrique et continue. Ces transformations sont très sensible à la présence de points aberrants ou éloignés.

2.4.1 Droite de Henry (Q-Q plot)

Avant de présenter les transformations, nous allons voir le principe de la droite de Henry. Il s'agit du fameux graphique **Q-Q plot** représentant un nuage de points.

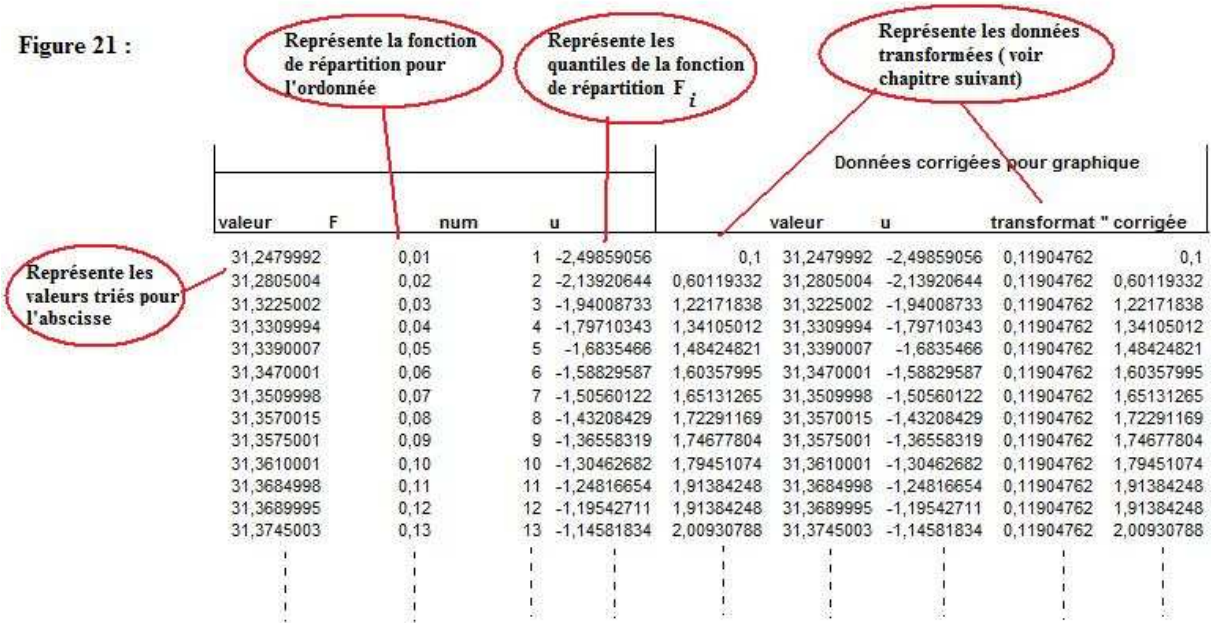
En abscisse nous avons les **données observées centrées et réduites** et en ordonnée les quantiles théoriques de la fonction de répartition $\boldsymbol{F_i = \frac{i-0.375}{n+0.25}}$.

La fonction **Excel** qui permet d'obtenir les quantiles théoriques en ordonnée est :

Loi.normale.inverse(F_i ; 0 ; 1) ;

La mise en forme de notre Q-Q plot dans Excel est expliquée à l'aide de la figure suivante :

Figure 21 :

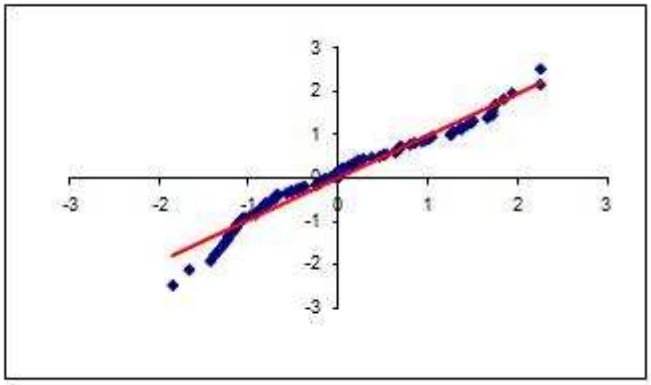


valeur	F	num	u	valeur	u	transformat " corrigée	
31,2479992	0,01	1	-2,49859056	0,1	31,2479992	-2,49859056	0,11904762	0,1
31,2805004	0,02	2	-2,13920644	0,60119332	31,2805004	-2,13920644	0,11904762	0,60119332
31,3225002	0,03	3	-1,94008733	1,22171838	31,3225002	-1,94008733	0,11904762	1,22171838
31,3309994	0,04	4	-1,79710343	1,34105012	31,3309994	-1,79710343	0,11904762	1,34105012
31,3390007	0,05	5	-1,6835466	1,48424821	31,3390007	-1,6835466	0,11904762	1,48424821
31,3470001	0,06	6	-1,58829587	1,60357995	31,3470001	-1,58829587	0,11904762	1,60357995
31,3509998	0,07	7	-1,50560122	1,65131265	31,3509998	-1,50560122	0,11904762	1,65131265
31,3570015	0,08	8	-1,43208429	1,72291169	31,3570015	-1,43208429	0,11904762	1,72291169
31,3575001	0,09	9	-1,36558319	1,74677804	31,3575001	-1,36558319	0,11904762	1,74677804
31,3610001	0,10	10	-1,30462682	1,79451074	31,3610001	-1,30462682	0,11904762	1,79451074
31,3684998	0,11	11	-1,24816654	1,91384248	31,3684998	-1,24816654	0,11904762	1,91384248
31,3689995	0,12	12	-1,19542711	1,91384248	31,3689995	-1,19542711	0,11904762	1,91384248
31,3745003	0,13	13	-1,14581834	2,00930788	31,3745003	-1,14581834	0,11904762	2,00930788

Les données sont réécrites dans une colonne. En effet, notre échantillon ayant des effectifs différents à chaque utilisation , notre graphique QQ plot, visible par l'utilisateur, garde toujours une même plage de données (qui est le nombre maximum de données que l'utilisateur peut entrer, ici 250). Si nous entrons 50 données, comment sont représentées les 200 cases vide par le graphique? Un bug est présent qui rend illisible le QQ plot. Pour compenser cela, cette colonne permet en cas de case vide, de la remplacer par la moyenne des données dans la colonne 'valeur' et la moyenne des quantiles théoriques dans la colonne 'u'. Le QQ plot permettant une régression linéaire, ces nouvelles données passeront toujours par la droite de régression, et notre graphique restera stable.

Voici un exemple de Q-Q plot établi grâce à la figure précédente :

Figure 20 :



Si notre nuage de points forment une droite, notre jeu de données est compatible avec une loi normale. Pour étudier la linéarité des points, nous utilisons le coefficient de corrélation linéaire. Plus les points seront alignés, plus le **coefficient de corrélation linéaire sera proche de 1**. La fonction Excel permettant de calculer ce nombre est :

Coefficient.correlation(X ; Y) ;

X étant la plage de données de notre échantillon centré réduit (‘valeur’)
Y étant la plage de données de nos quantiles théoriques (‘u’)

Maintenant le principe de corrélation linéaire du Q-Q plot connue, nous pouvons voir sa mise en pratique dans les transformations.

2.4.2 Transformation de Box – Cox

Principe :

Le but de cette transformation est de prendre un jeu de données et de la transformer avec la formule :

$$y = \frac{x^{\lambda}-1}{\lambda} \qquad \text{si } \lambda \neq 0$$

$$y = \ln(x) \qquad \text{si } \lambda = 0$$

y étant nos nouvelles données transformées
x étant nos données d’origine
λ un paramètre variant de -10 à 10

Nous devons faire varier λ de -10 à 10 et trouver quel sera notre nouveau jeu de données avec la meilleure corrélation linéaire calculée à l’aide du nuage de points du Q-Q plot.

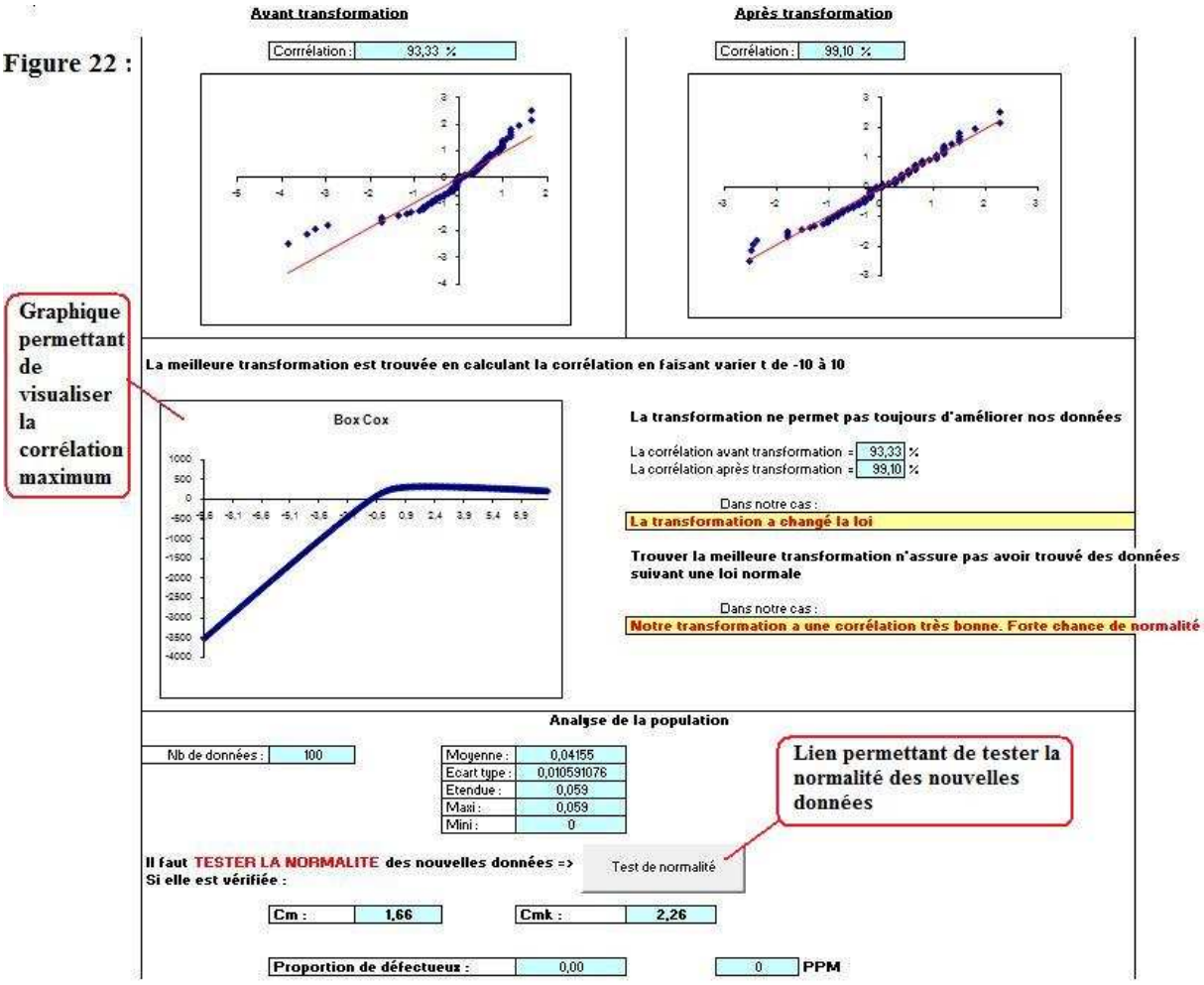
λ variera par pas de 0.1, et nous garderons celui qui a la **plus grande corrélation**. Ainsi nous avons trouvé la **meilleure transformation**. Mais cette transformation ne garantit pas que nous ayons

un nouveau jeu de données qui suit **une loi normale**. Un test de normalité doit être effectué avec nos nouvelles données.

Si nous acceptons la normalité, il suffit de transformer nos tolérances avec la même formule donnant de nouvelles tolérances adaptées au nouveau jeu de données. Le calcul de capabilité se fera comme d'habitude en utilisant nos nouvelles données.

Box–Cox dans le Guidestat :

Figure 22 :



- La corrélation calculée est donnée en pourcentage. Il s'agit juste d'une multiplication par 100 pour mieux interpréter le résultat.
- Les nouvelles données sont données dans la zone de saisie.
- L'analyse de la population est faite avec les données d'origine. Le calcul de capabilité avec les nouvelles données.
- La 3ème partie non visible par l'utilisateur est sujette a des calculs très longs qui prennent compte du principe de calcul ci-dessus.

2.4.3 Transformation de Johnson

Principe :

Le but de cette transformation est le même que celle de Box-Cox. A l'aide de formules compliquées nous devons trouver la meilleure **transformation normalisante**. Les sources utilisées ne sont pas complètes. Les recherches ont été faites sur internet, et les explications concrètes se trouvent dans des livres spécifiques. Néanmoins en trouvant des informations de plusieurs sources, une programmation très proche de la conclusion finale proposée par **Minitab** qui effectue aussi cette transformation fut mise en place dans le **Guidestat**.

2 transformations ont été utilisées :

- La transformation bornée

$$a + b \ln [\frac{x - e}{c + e - x}]$$

définie sur l'intervalle] e ; $e + c$ [

- La transformation non bornée

$$a + b \ \sinh^{-1}(\frac{x - e}{c})$$

définie ∀ x

a, b, c et d sont des paramètres à estimer à l'aide d'une valeur **z** = 0.524 utilisée pour le calcul de la fonction de répartition qui suit **une loi normale N(0,1)** en **-3z, -z, z et 3z.**

Nous devons ensuite trouver les quantiles de notre jeu de données associés aux probabilités précédentes grâce à une estimation de la fonction de répartition de nos données. Différentes méthodes existent telles que :

- les **courbes de Burr** calculant la fonction de répartition à l'aide de polynôme.
- la **méthode de Clements** utilisant le **Skewness** et le **Kurtosis**. Ce sont les moments d'ordre 3 et 4 de nos données représentant le **coefficient d'asymétrie et d'aplanissement**.

La **méthode de Clements** est idéale aux transformations de Johnson. Les coefficients d'asymétrie et d'aplanissement permettent d'avoir une idée de la loi de nos données et transforment nos données idéalement en loi normale. Mais elles utilisent des tables qui sont l'origine du problème des sources exposées au début du paragraphe.

La technique utilisée pour l'estimation de nos quantiles est une méthode de pas à pas aussi utilisée pour la recherche de λ dans la **transformation Box-Cox**. Cette méthode est longue et prend énormément de place dans la programmation de notre fiche.
Notons :

$$\boldsymbol{m = x_{-3z} - x_{-z}}$$
$$\boldsymbol{n = x_{-z} - x_{z}}$$
$$\boldsymbol{p = x_{z} - x_{-3z}}$$

Avec x_{-3z} représentant le quantile estimé de nos données pour **-3z**

On distingue 3 cas :

- Si $\frac{mn}{p^2}$ < 1 alors on utilise la transformation bornée
- Si $\frac{mn}{p^2}$ > 1 alors on utilise la transformation non bornée
- Si $\frac{mn}{p^2}$ = 1 nous ne calculons aucune transformation

Nous calculons pour chaque quantile les valeurs a, b, c et e à l'aide de formules. Nous utiliserons aussi la meilleure corrélation linéaire calculée à l'aide du nuage de points du Q-Q plot pour chaque transformation établi à l'aide des quantiles et paramètres estimés.

La transformation ayant la meilleure corrélation sera renvoyé et un **test de normalité** des nouvelles données est à faire.

Capabilité :

Le calcul de la capabilité est différent pour les 2 transformations.

Si nous avons la **transformation bornée** alors :

$$Cp = \frac{I.T.}{\lambda}$$

$$Cpk = \min\left[\frac{TS - \bar{X}}{e - \bar{X}} ; \frac{\bar{X} - TI}{\bar{X} - e - \lambda}\right]$$

λ représente la dispersion
e et e + λ représentent les limites de la dispersion

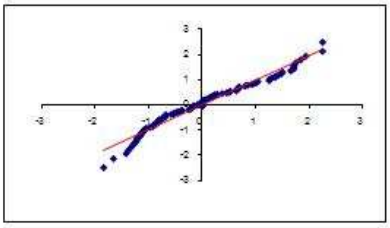
Si nous avons la **transformation non bornée** alors le calcul se fera comme pour les capabilités habituelles avec nos données transformées et les tolérances transformées.

Transformation de Johnson dans le Guidestat :

Figure 23 :

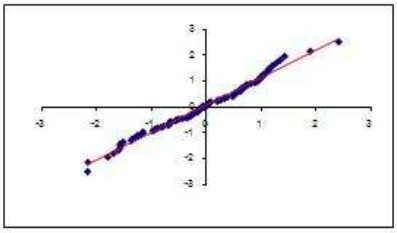
Avant transformation

Corrrélation :	98,22 %



Après transformation

Corrélation :	99,42 %



La meilleure transformation trouvée est :

-0,595	+	1,096	X	LN	((x - 32,163) / (31,189 - x))

La transformation ne permet pas toujours d'améliorer nos données

La corrélation avant transformation =	98,22 %
La corrélation après transformation =	99,42 %

Dans notre cas :

La transformation a changé la loi

Trouver la meilleure transformation n'assure pas avoir trouvé des données suivant une loi normale

Dans notre cas :

Notre transformation n'est pas parfaite. Faible chance de normalité

Analyse de la population

Nb de données :	100

Moyenne :	31,574
Ecart type :	0,177
Etendue :	0,727
Maxi :	31,975
Mini :	31,248

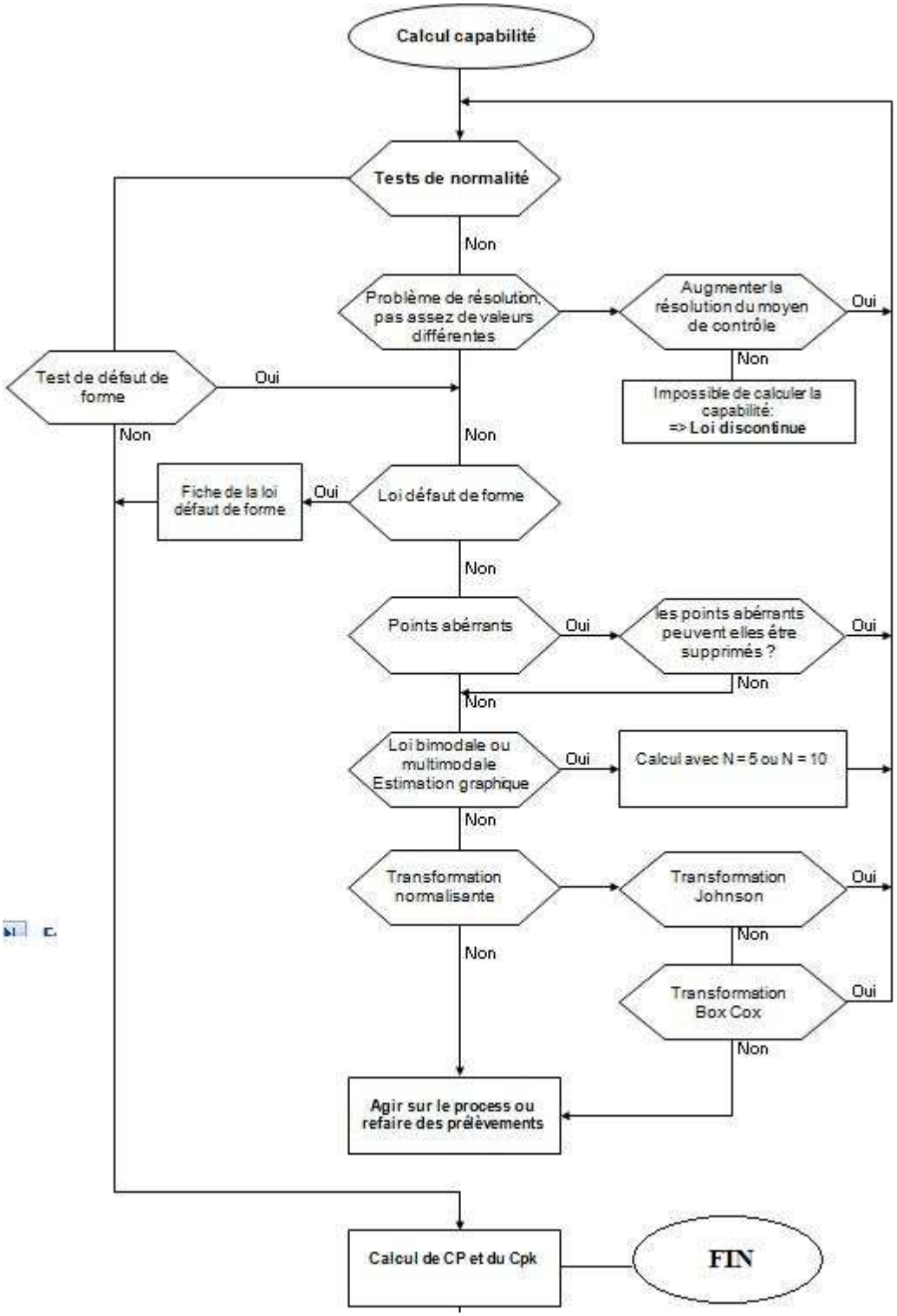
Il faut TESTER LA NORMALITE des nouvelles données => Test normalité
Si elle est vérifiée :

Cm :	1,03

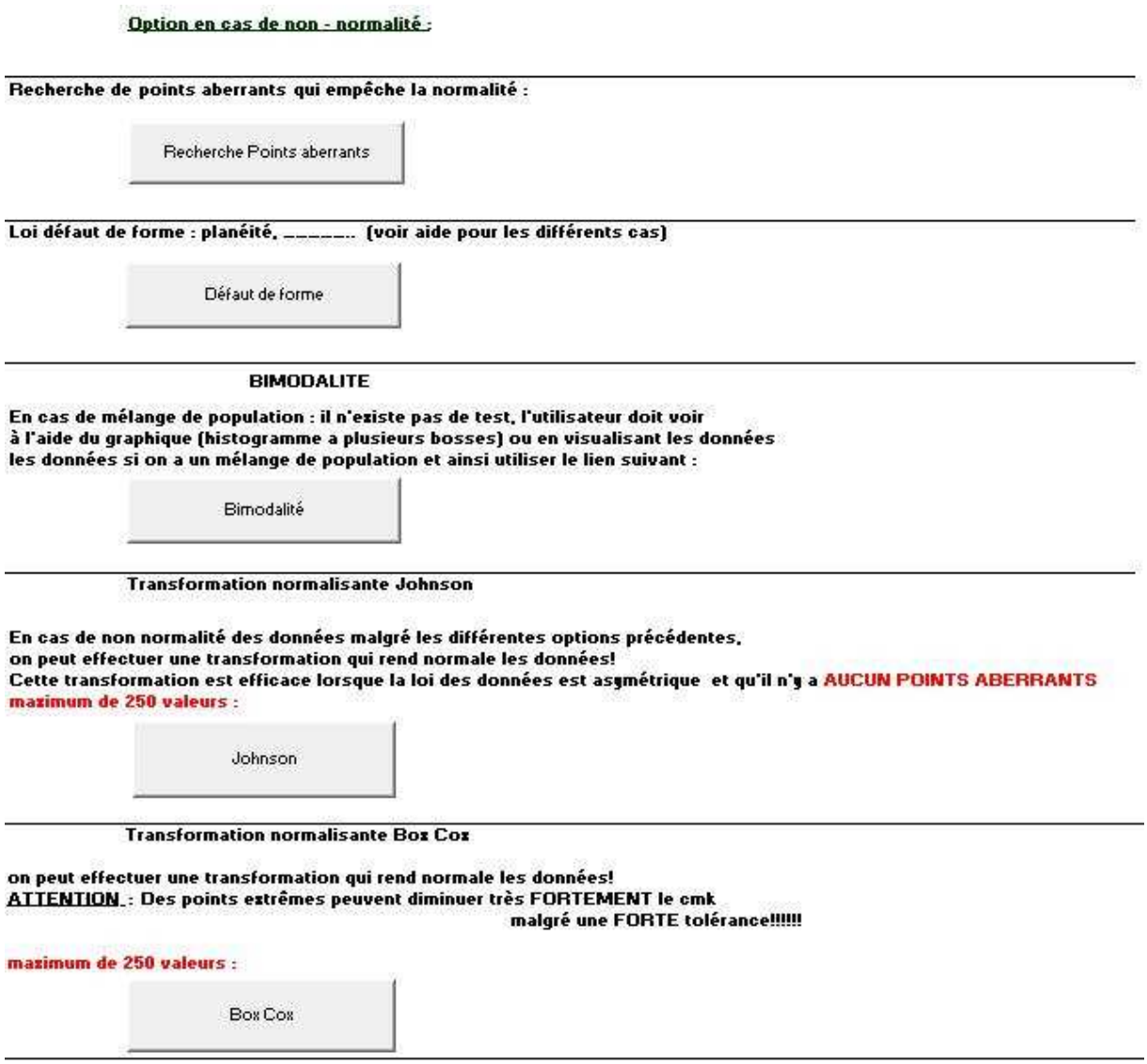
Cmk :	0,72

CONCLUSION

Le logigramme suivant résume la finalité de la recherche.



Dans le **Guidestat,** une fois les données entrées dans la fiche 7, la seule utilisation de la souris permet de nous guider dans les différentes fiches. La figure suivante nous montre une partie de la fiche 7 avec les actions possibles de l'utilisateur.



La présence de points aberrants peut fausser la normalité, mais si nous devons les conserver dans l'analyse de nos données, alors les autres méthodes de résolution du problème ne serviront pas et une réflexion sur la cause de ces points aberrants lors de nos mesures est à faire.

La loi bimodale peut se révéler trop facile. En effet, pour certain jeu de données, nous pouvons tester la normalité décrite dans cette fiche. C'est pour cela que son principe d'utilisation doit être limité.

Les transformations sont très sensibles aux valeurs extrêmes. La capabilité d'une machine sera vite refusée. La transformation de Johnson n'a pas une finalité dans le **Guidestat** la plus pertinente, mais la plus fidèle possible a été faite lors de ce stage.

SOURCES

- Cours de Statistique de Ricco Rakotomalala, enseignant à l'université de Lyon :
  Un polycopié de ce cours fut disponible lors de mon stage. Un lien internet de ce cours est :
  http://eric.univ-lyon2.fr/~ricco/cours/cours/Test_Normalite.pdf
  Utilisé pour les tests de normalité et la Transformation Box-Cox

- Livre de Maurice Pillet : Appliquer la maitrise statistique des procédés MSP/SPC
  Utilisé pour la théorie du défaut de forme et des tables.

- L'aide **Minitab** pour l'utilisation de la Transformation de Johnson

- Thèse sur l'analyse des procédés : *http://www.polytech.univ-savoie.fr/fileadmin/polytech_autres_sites/sites/listic/Theses/theseduclos.pdf*
  Utilisée pour la Transformation de Johnson

- Table du test de Kolmogorov : *http://courses.wcupa.edu/rbove/eco252/252KStest.doc*

Les sources exposées ici sont celles qui ont servi pour la résolution de la problématique. De nombreuses autres sources ont permis l'inspiration du sujet.

Bien évidemment, le cours suivi lors du M1 statistique a été l'inspiration principale.

ANNEXES

ANNEXE 1 : Table de Kolmogorov et de Lilliefors

ANNEXE 2 : Table pour défaut de forme

ANNEXE 3 : Fiche Information Normalité

ANNEXE 1 : Table des valeurs critiques du test de Kolmogorov-Smironov suivi des valeurs critiques pour la correction de Lilliefors.

Table test de kolmogorov		
n	valeur critique	" Lilliefors
1	0,975	
2	0,842	
3	0,708	
4	0,624	0,3754
5	0,563	0,3427
6	0,519	0,3245
7	0,483	0,3041
8	0,454	0,2825
9	0,43	0,2744
10	0,409	0,2616
11	0,391	0,2506
12	0,375	0,2426
13	0,361	0,2337
14	0,349	0,2257
15	0,338	0,2196
16	0,327	0,2128
17	0,318	0,2071
18	0,309	0,2018
19	0,301	0,1965
20	0,294	0,192
21	0,287	0,1881
22	0,281	0,184
23	0,275	0,1798
24	0,269	0,1766
25	0,264	0,1726
26	0,259	0,1699
27	0,254	0,1665
28	0,25	0,1641
29	0,246	0,1614
30	0,242	0,159
31	0,238	0,1559
32	0,234	0,1542
33	0,231	0,1518
34	0,227	0,1497
35	0,224	0,1478
36	0,221	0,1454
37	0,218	0,1436
38	0,215	0,1421
39	0,213	0,1402
40	0,21	0,1386

ANNEXE 2 : Table de la détermination des constants K1 et K2 en fonction de Ko + Table de détermination de la dispersion du défaut de forme.

Détermination de K1 et K2 d'après la table page 323

Ko	a	K1	K2
0,798	0,00	1,000	0,000
0,798	0,30	0,957	0,287
0,799	0,40	0,929	0,371
0,800	0,45	0,913	0,410
0,805	0,65	0,836	0,544
0,810	0,76	0,792	0,603
0,815	0,85	0,762	0,651
0,820	0,93	0,732	0,682
0,825	1,00	0,707	0,707
0,830	1,07	0,682	0,732
0,835	1,14	0,660	0,752
0,840	1,20	0,640	0,768
0,845	1,25	0,623	0,783
0,850	1,31	0,606	0,796
0,855	1,37	0,589	0,809
0,860	1,44	0,570	0,822
0,865	1,51	0,553	0,834
0,870	1,57	0,538	0,844
0,875	1,63	0,522	0,853
0,880	1,70	0,507	0,862
0,885	1,77	0,492	0,870
0,890	1,83	0,479	0,878
0,895	1,90	0,466	0,885
0,900	1,97	0,453	0,893
0,905	2,04	0,440	0,899
0,910	2,12	0,426	0,906
0,915	2,21	0,412	0,912
0,920	2,30	0,399	0,918
0,925	2,40	0,385	0,924
0,930	2,50	0,371	0,929
0,935	2,62	0,356	0,934
0,940	2,76	0,341	0,940
0,945	2,90	0,326	0,945
0,950	3,04	0,312	0,950
0,955	3,22	0,296	0,955
0,960	3,44	0,279	0,960
0,965	3,67	0,263	0,965
0,970	3,99	0,243	0,970
0,975	4,39	0,222	0,975
0,980	4,92	0,199	0,980
0,985	5,69	0,173	0,985
0,990	7,02	0,141	0,990
0,995	9,96	0,100	0,995
0,997	12,88	0,077	0,998
0,999	22,34	0,045	0,999
1,000	1,00E+99	0,000	1,000

$Z_{0,3173}/\sigma$ en fonction de m/σ

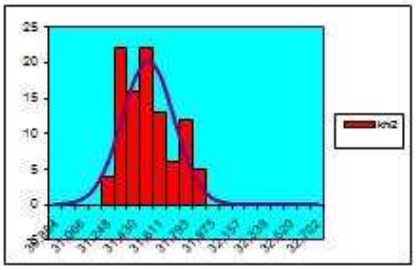
m/σ	$Z_{0,3173}/\sigma$
0,000	2,960
0,300	3,080
0,400	3,140
0,450	3,190
0,480	3,200
0,520	3,260
0,560	3,300
0,610	3,350
0,650	3,400
0,670	3,420
0,690	3,440
0,710	3,460
0,740	3,470
0,760	3,500
0,780	3,520
0,800	3,530
0,810	3,550
0,830	3,570
0,850	3,590
0,870	3,610
0,880	3,620
0,900	3,650
0,910	3,670
0,930	3,690
0,940	3,700
0,960	3,720
0,970	3,730
0,980	3,740
1,000	3,750

ANNNEXE 3 : Présentation de la fiche d'information sur la normalité. Celle-ci est accessible grâce à un lien présent sur la fiche 7. Cette fiche explique la conclusion des différents tests.

Analyse de la population

Nb de données :	100

Moyenne :	31,574235
Ecart type :	0,1771587
Etendue :	0,7270012
Maxi :	31,975
Mini :	31,248



Conclusion sur la normalité :

La loi n'est pas normale

La normalité des données est basée sur les 4 tests suivant
Pour en savoir plus sur le principe de décision voir l'aide => **Aide**

Test de Shapiro Wilk Ce test est très efficace pour un nombre de données inférieur à 50. Priorité de conclusion si normalité avec ce test

Analyse :

Nombre de données trop importantes pour la conclusion de ce test

Test de Kolmogorov Ce test est sensible aux données autour de la partie centrale de la distribution (à comparer avec le test d'Anderson)
(avec correction de Lilliefors)

Analyse :

Théorique : 0,0883 Pratique : 0,0935

La loi n'est pas normale

Test d'Anderson : Ce test est sensible aux données aux queues de distribution (à comparer avec le test de Kolmogorov)

Analyse :

Théorique : 0,752 Pratique : 1,232

La loi n'est pas normale

Test du Khi 2 : Ce test est sensible à la façon dont on le programme, donc indépendant de l'utilisateur. Plus il y a de valeurs, plus il est pertinent

Analyse :

Théorique : 14,07 Pratique : 20,33

La loi n'est pas normale