Estimation par sondage simple

TP1 :

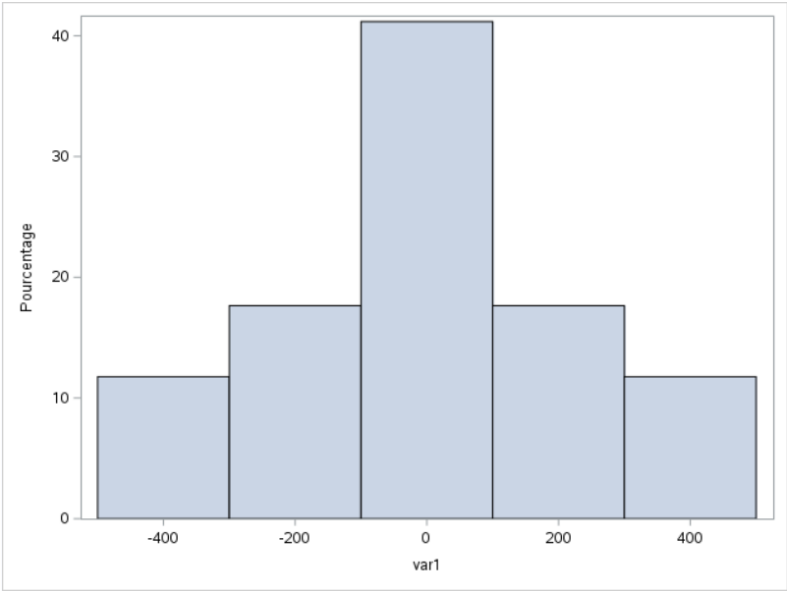
Exercice 1 :

La première étape consiste à créer une base de données, en utilisant l'instruction DATA. Les valeurs de la variable "var1" sont saisies à partir de la série de données fournies.

Voici un extrait du tableau obtenu à partir de ce script.

	var1
1	-138.38
2	77.75
3	233.96
4	-131.52
5	368.52
6	-36.37
7	-78.03
8	94.72

Pour visualiser la distribution de "var1", nous utilisons la procédure SGPLOT pour créer un histogramme. L'histogramme affiche la fréquence des valeurs de "var1" et permet de visualiser leur répartition.



Afin d'organiser les données dans un ordre croissant, nous utilisons la procédure SORT pour trier l'ensemble des données selon la variable "var1". Cela nous permet d'obtenir une vue ordonnée des valeurs de "var1". Voici un extrait des données qui sont désormais trié dans l'ordre croissant.

	var1
1	-449.99
2	-322.12
3	-233.56
4	-138.38
5	-131.52
6	-78.03
7	-36.37
8	-23.13

Pour faciliter la comparaison entre les différentes observations, nous avons utilisé la procédure STANDARD afin de standardiser la variable "var1".

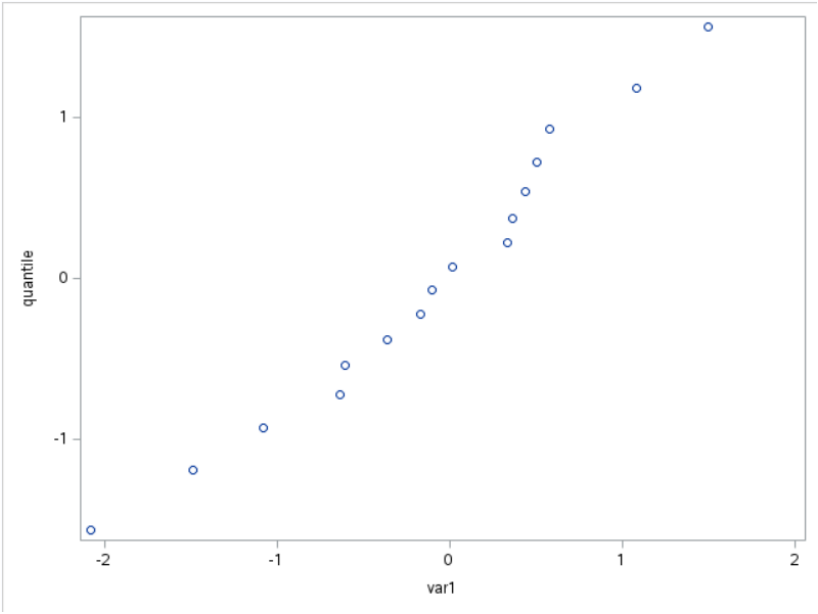
Voici un extrait des données après exécution.

	var1
1	-2.079452141
2	-1.488166208
3	-1.078654342
4	-0.638530816
5	-0.606809369
6	-0.359465312
7	-0.166824567
8	-0.105601249

Nous allons maintenant créer un nouvel ensemble de données appelé "base3" à l'aide de la procédure DATA. En copiant les données de "base2" dans "base3" avec l'instruction SET, nous ajoutons deux nouvelles variables : "id", qui représente le numéro d'observation, et "quantile", qui correspond au quantile normalisé de la variable "var1", nous avons également utilisé la fonction probit. Voici un extrait des résultats obtenus.

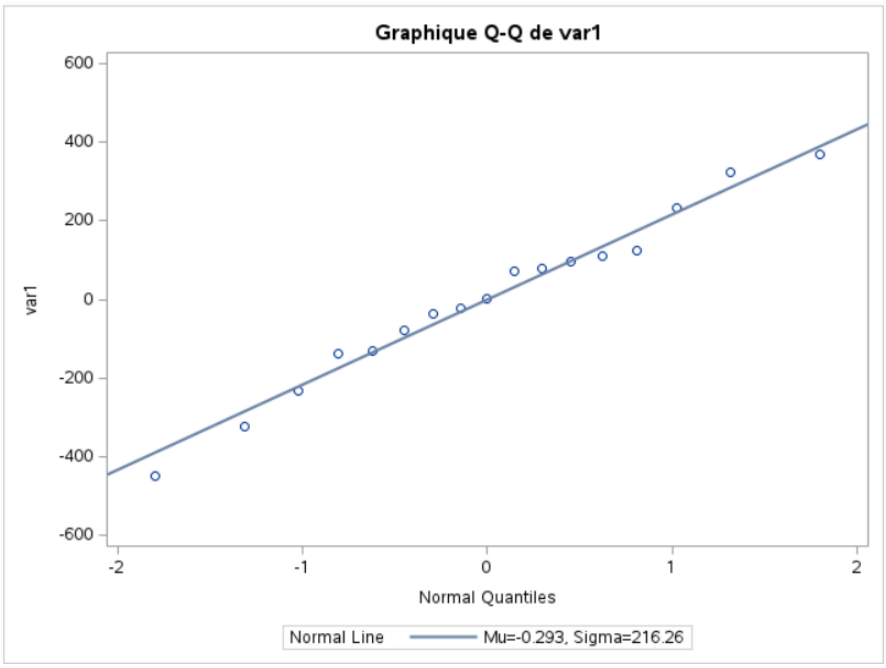
	var1	id	alpha	quantile
1	-2.079452141	1	0.0588235294	-1.564726471
2	-1.488166208	2	0.1176470588	-1.186831433
3	-1.078654342	3	0.1764705882	-0.928899492
4	-0.638530816	4	0.2352941176	-0.721522284
5	-0.606809369	5	0.2941176471	-0.541395085
6	-0.359465312	6	0.3529411765	-0.377391944
7	-0.166824567	7	0.4117647059	-0.223007831
8	-0.105601249	8	0.4705882353	-0.073791274

En utilisant la procédure SGPLOT, nous avons créé un nuage de points (scatter plot) pour visualiser la relation entre les valeurs de "var1" sur l'axe des abscisses et les quantiles sur l'axe des ordonnées. Ce graphique nous permet d'observer la répartition des quantiles normalisés par rapport aux valeurs de "var1".

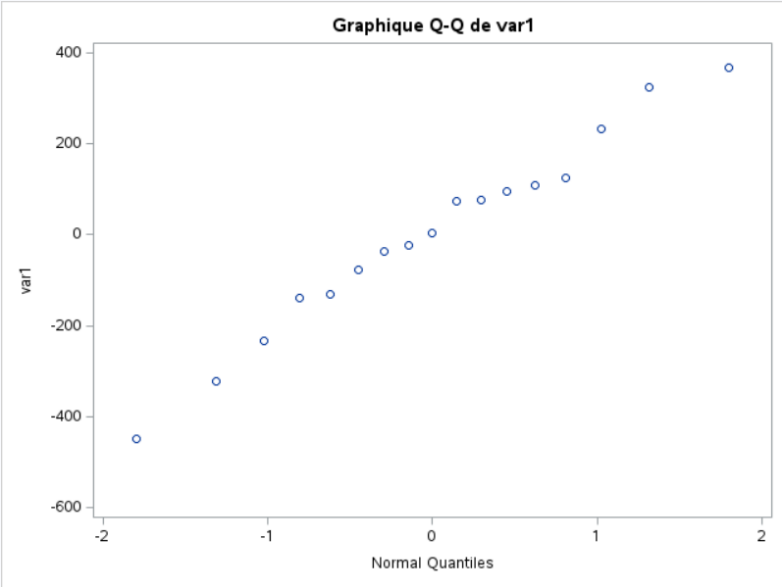


Nous pouvons tirer comme conclusion que les quantiles observés et théoriques sont répartis le long d'une droite. On peut donc conclure que les données initiales suivent approximativement une loi normale.

Par la suite nous avons utilisés la procédure UNIVARIATE, cette fois avec l'option QQPLOT, pour créer un QQ-plot. Ce graphique compare les quantiles observés de "var1" avec les quantiles théoriques d'une distribution normale. Une répartition linéaire des points indique une approximation de la loi normale par les données.



Nous avons fait la même chose avec la procédure CAPABILITY, et voici ce que nous avons obtenu.



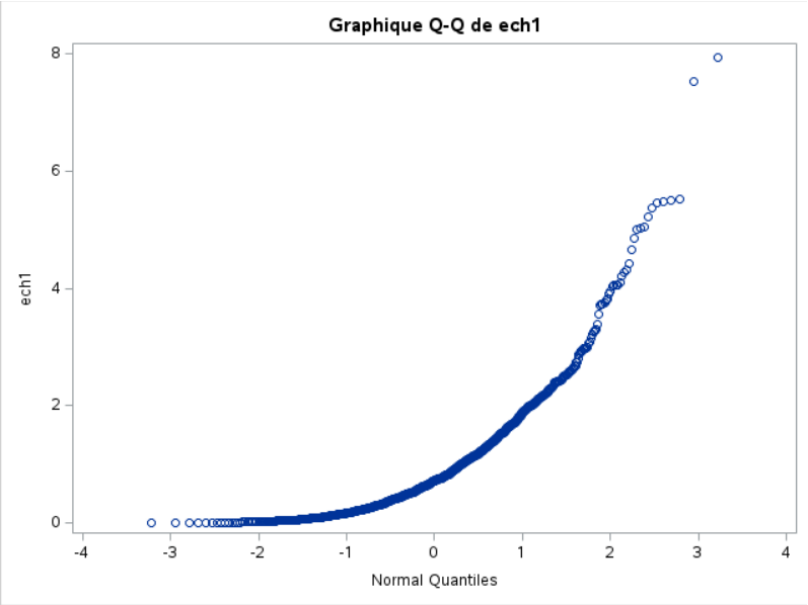
Exercice 2 :

La première étape de cet exercice consiste à importer les données à partir du fichier Excel "TP1-exo2.xls" en utilisant la procédure IMPORT. Les données importées sont stockées dans la base de données "base1".

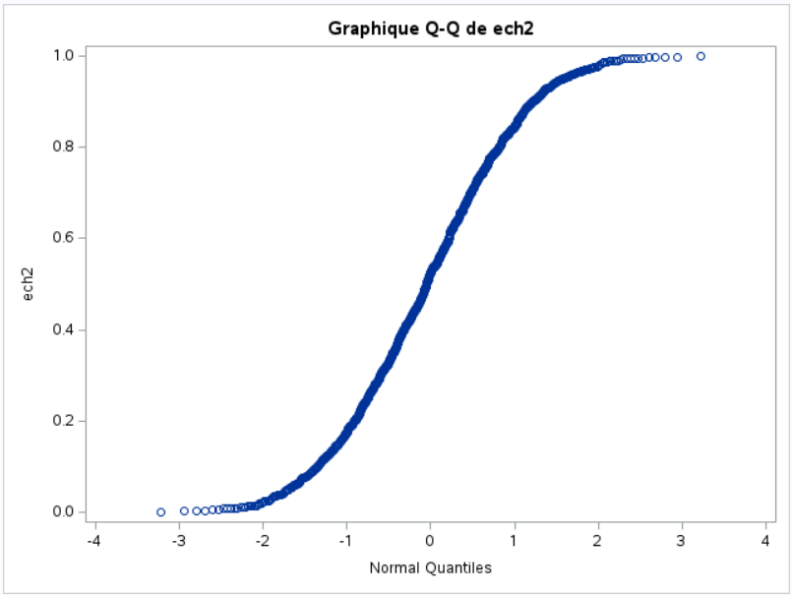
Voici un extrait de ce qui as été obtenu, suite a cette étape.

	i	ech1	ech2	ech3	ech4
1	1	0.2402054	0.9897063	-0.451365	2.8946316
2	2	0.3915827	0.5558761	-0.171842	2.7916815
3	3	0.8467811	0.8483563	-0.351301	0.4270078
4	4	0.492768	0.9304031	-0.566958	0.2654851
5	5	0.1165421	0.1143332	0.9145271	0.221206
6	6	1.844648	0.0204298	-1.68404	0.181084
7	7	1.7932615	0.2427004	1.0672304	1.2457548
8	8	0.3758187	0.5588776	0.1435683	7.2990619
9	9	0.8443192	0.4008133	-2.099096	0.1903863
10	10	2.109374	0.0086936	-1.187543	1.6990246

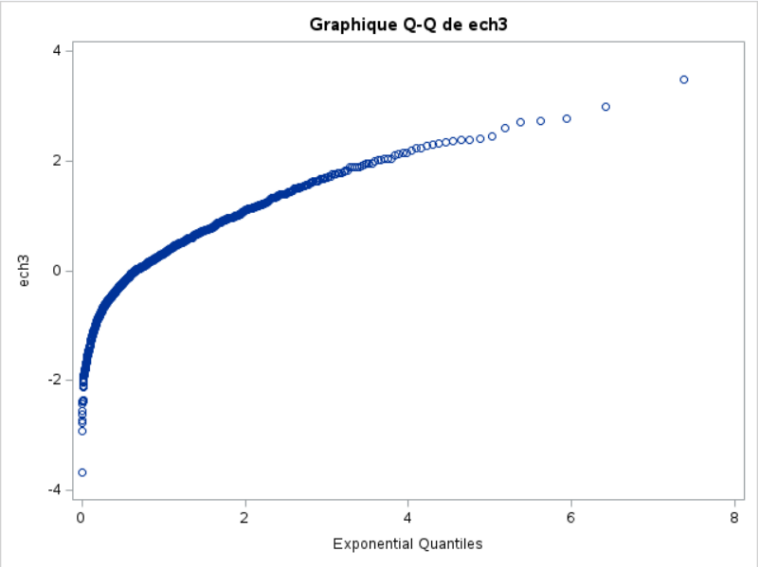
Nous avons ensuite utilisé la procédure UNIVARIATE afin d'effectuer une analyse univariée sur la variable "ech1". Nous avons créé un QQ-plot pour comparer les quantiles observés de "ech1" avec les quantiles théoriques d'une distribution normale. Cela nous permet d'évaluer si "ech1" suit approximativement une loi normale.



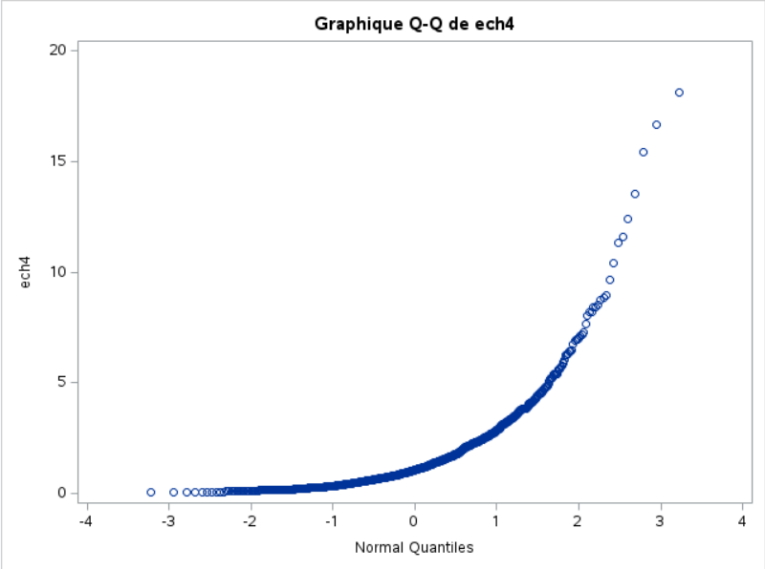
De manière similaire à l'étape précédente, nous effectuons une analyse univariée sur la variable "ech2" à l'aide de la procédure UNIVARIATE. Nous créons un QQ-plot pour évaluer la conformité des données de "ech2" à une distribution normale.



Nous utilisons à nouveau la procédure UNIVARIATE pour analyser la variable "ech3". Cette fois-ci, nous créons un QQ-plot pour comparer les quantiles observés de "ech3" avec les quantiles théoriques d'une distribution exponentielle. Cela nous permet de vérifier si les données de "ech3" suivent approximativement une distribution exponentielle.



Cette fois ci, nous analysons la variable "ech4", en créant encore une fois un QQ-plot pour évaluer la conformité des données de "ech4" à une distribution normale.



Exercice 3

La première étape consiste à comme pour l'exercice 2, importer les données du fichier Excel "tp1-exo3.xls" en utilisant la procédure IMPORT.

	i	var1
1	1	-0.173386
2	2	-0.04926
3	3	-0.61599
4	4	-0.530455
5	5	-0.596877
6	6	0.1285798
7	7	-0.528258
8	8	1.0434753
9	9	0.6451714
10	10	-0.068779

Nous avons utilisé la procédure MEANS pour calculer la moyenne et la variance de la variable "var1" dans l'ensemble de données "base1". Ces mesures statistiques fournissent une indication de la tendance centrale et de la dispersion des données. Voici le résultat obtenu.

La procédure MEANS

Variable d'analyse : var1 var1	
Moyenne	Variance
-0.0100780	1.1001492

Par la suite, nous utilisons un macro-programme appelée "simulMean" qui sert à simuler les moyennes. La macro effectue une simulation en utilisant la méthode de l'échantillonnage aléatoire Les moyennes simulées sont enregistrées dans des ensembles de données distincts ("baseSM") pour chaque taille d'échantillon.

Apres cela, nous avons créé des histogrammes pour visualiser les distributions des moyennes simulées pour chaque taille d'échantillon. Les histogrammes sont créés en utilisant la procédure UNIVARIATE avec l'option HISTOGRAM. Les histogrammes à noyau (kernel) sont utilisés pour estimer les densités de probabilité.

TP2 :

Exercice 1 :

Nous avons commencé par créer une table appelée "exo1". Les valeurs sont saisies manuellement à l'aide de l'instruction CARDS. La table contient les valeurs suivantes : 12, 3, 18, 0, 25, 18, 13, 21, 3, 8, 16, 14. Voici ce que l'on a obtenue.

	valeur
1	12
2	3
3	18
4	0
5	25
6	18
7	13
8	21
9	3
10	8
11	16
12	14

Nous avons ensuite utilisé la procédure UNIVARIATE pour estimer l'intervalle de confiance des valeurs. Nous spécifions la variable "valeur" et utilisons l'option CIBASIC pour calculer l'intervalle de confiance à un niveau de confiance de 95%. Les résultats affichés sont les suivants.

Intervalle de confiance de base sous hypothèse de normalité			
Paramètre	Estimation	Intervalle de confiance à 95%	
Moyenne	12.58333	7.65836	17.50830
Ecart-type	7.75134	5.49101	13.16084
Variance	60.08333	30.15124	173.20762

Lors de l'étape suivante nous avons dû activer l'option ODS TRACE pour obtenir des informations plus détaillées sur les résultats de l'estimation de l'intervalle de confiance. Nous utilisons ensuite à nouveau la procédure UNIVARIATE avec l'option CIBASIC dans le but de calculer les intervalles de confiance. Cette fois, nous utilisons l'option ODS OUTPUT pour enregistrer les résultats dans un fichier de sortie appelé "mesIntervalles".

Pour finir cet exercice nous avons utilisé l'option ODS RTF pour générer un rapport en format RTF. La procédure UNIVARIATE est exécutée avec l'option CIBASIC pour calculer les intervalles de confiance, et les résultats ont été enregistrés dans le fichier RTF mentionné.

Exercice 2 :

Nous commençons cet exercice en créant une table "exo2". Les valeurs sont saisies manuellement à l'aide de l'instruction CARDS comme lors de l'exercice précédent.

Extrait du jeu de données.

	valeur
1	15
2	6
3	21
4	0
5	27
6	9
7	8
8	23
9	16
10	26

Ensuite nous avons activé l'option ODS TRACE pour obtenir des informations sur les résultats de l'estimation de l'intervalle de confiance. Nous spécifions la variable "valeur" et utilisons l'option CIBASIC pour calculer l'intervalle de confiance à un niveau de confiance de 95%. Les résultats sont les suivants.

Intervalle de confiance de base sous hypothèse de normalité			
Paramètre	Estimation	Intervalle de confiance à 95%	
Moyenne	14.97143	11.78748	18.15538
Ecart-type	9.26881	7.49729	12.14402
Variance	85.91092	56.20928	147.47724

Nous avons utilisé après cela, la procédure MEANS pour calculer les statistiques descriptives des valeurs de la table. Nous spécifions la variable "valeur" et utilisons l'option ALPHA=0.05 pour un niveau de confiance de 95%. Les résultats comprennent la moyenne, l'écart-type et les intervalles de confiance des statistiques.

Ici nous utilisons à nouveau la procédure UNIVARIATE pour calculer les quantiles des valeurs dans le jeu de données. Les résultats sont enregistrés dans un fichier de sortie appelé "quantile" à l'aide de l'option ODS OUTPUT.

Extrait du PDF final.

La procédure UNIVARIATE
Variable : valeur

Moments			
N	35	Somme des poids	35
Moyenne	14.9714286	Somme des observations	524
Ecart-type	9.26881462	Variance	85.9109244
Skewness	-0.1227973	Kurtosis	-1.323006
Somme des carrés non corrigée	10766	Somme des carrés corrigée	2920.97143
Coeff Variation	61.9100213	Std Error Mean	1.56671562

Mesures statistiques de base			
Location		Variabilité	
Moyenne	14.97143	Ecart-type	9.26881
Médiane	15.00000	Variance	85.91092
Mode	0.00000	Intervalle	29.00000
		Ecart interquartile	18.00000

Note: Le mode affiché est le plus petit des 10 modes avec un effectif de 2.

Tests de tendance centrale : Mu0=0				
Test	Statistique		p-value	
t de Student	t	9.555932	Pr > |t|	<.0001
Signe	M	16.5	Pr >= |M|	<.0001
Rang signé	S	280.5	Pr >= |S|	<.0001

Exercice 3 :

Pour ce dernier exercice nous avons commencé par créer la table "exo3". Les 18 premières observations sont définies comme "malade" et est représenté avec une valeur de 1, tandis que le reste des observations sont définies comme non "malade" et est représenté par une valeur de 0.

	i	malade
1	1	1
2	2	1
3	3	1
4	4	1
5	5	1
6	6	1
7	7	1
8	8	1
9	9	1
10	10	1

Nous utilisons ensuite la procédure FREQ pour calculer la fréquence de la variable "malade". L'option TABLE est utilisée pour spécifier la variable "malade" et l'option BINOMIAL est utilisé.

La procédure FREQ

malade	Fréquence	Pourcentage	Fréquence cumulée	Pourcentage cumulé
0	42	70.00	42	70.00
1	18	30.00	60	100.00

Proportion binomiale	
malade = 0	
Proportion	0.7000
ASE	0.0592
Borne inférieure de l'IC à 95%	0.5840
Borne supérieure de l'IC à 95%	0.8160
Intervalle de confiance exact	
Borne inférieure de l'IC à 95%	0.5679
Borne supérieure de l'IC à 95%	0.8115

Test de H0 : Proportion = 0.5	
ASE sous H0	0.0645
Z	3.0984
Pr > Z unilatérale	0.0010
Pr > |Z| bilatéral	0.0019

Taille de l'échantillon = 60

Enfin nous avons utilisé un macro-programme pour effectuer deux analyses différentes sur le jeu de données "exo3". Dans la première situation, nous appliquons la macro avec les paramètres suivants, taille de l'échantillon 85 avec 34 qui sont "malades". Dans la seconde situation nous avons les paramètres suivants, taille de l'échantillon 40 et nombre de "malades", 8. Les résultats des analyses sont les suivants.

La procédure FREQ

malade	Fréquence	Pourcentage	Fréquence cumulée	Pourcentage cumulé
0	51	60.00	51	60.00
1	34	40.00	85	100.00

Proportion binomiale	
malade = 0	
Proportion	0.6000
ASE	0.0531
Borne inférieure de l'IC à 95%	0.4959
Borne supérieure de l'IC à 95%	0.7041
Intervalle de confiance exact	
Borne inférieure de l'IC à 95%	0.4880
Borne supérieure de l'IC à 95%	0.7048

Test de H0 : Proportion = 0.5	
ASE sous H0	0.0542
Z	1.8439
Pr > Z unilatérale	0.0326
Pr > |Z| bilatéral	0.0652

Taille de l'échantillon = 85

La procédure FREQ

malade	Fréquence	Pourcentage	Fréquence cumulée	Pourcentage cumulé
0	32	80.00	32	80.00
1	8	20.00	40	100.00

Proportion binomiale	
malade = 0	
Proportion	0.8000
ASE	0.0632
Borne inférieure de l'IC à 95%	0.6760
Borne supérieure de l'IC à 95%	0.9240
Intervalle de confiance exact	
Borne inférieure de l'IC à 95%	0.6435
Borne supérieure de l'IC à 95%	0.9095

Test de H0 : Proportion = 0.5	
ASE sous H0	0.0791
Z	3.7947
Pr > Z unilatérale	<.0001
Pr > |Z| bilatéral	0.0001

Taille de l'échantillon = 40