

IUT Grand Ouest Normandie

Bachelor Universitaire de Technologie
Science des Données
Campus de Lisieux

Science des Données 1 - SAÉ 2.02
Éstimation par sondage simple

Thématique

TP1 - Statistique Inférentielle

Auteurs

BOUCHER Nathan
GAMONDELE Maxime

Table des matières

Introduction	3
1 Exercice 1	4
2 Exercice 2	6
2.1 Échantillon n°2	6
2.2 Échantillon n°3	6
2.3 Échantillon n°4	6
3 Exercice 3	8
3.1 Moyenne et variance de la variable var1	8
3.2 Distribution de l'estimateur de la moyenne pour taille d'échantillon = 2	8
3.3 Distribution de l'estimateur de la moyenne pour taille d'échantillon = 5	9
3.4 Distribution de l'estimateur de la moyenne pour taille d'échantillon = 10	9
3.5 Distribution de l'estimateur de la moyenne pour taille d'échantillon = 20	10
3.6 Distribution de l'estimateur de la moyenne pour taille d'échantillon = 50	10
Conclusion	11
Table des figures	12

Introduction

À travers les deux premières études, nous cherchons à comprendre et à estimer les différentes lois de probabilité associées à chaque échantillon. Dans la troisième partie, nous cherchons à étudier comment la taille de l'échantillon peut affecter notre estimateur. x

1 Exercice 1

À travers ce premier exercice, nous cherchons à tester la normalité de notre série de données à l'aide de la droite de Henry.

Tout d'abord, observons la distribution de nos valeurs à l'aide d'un histogramme (cf. Fig. 1). Nous remarquons que la distribution semble suivre une loi normale avec une forme en cloche.

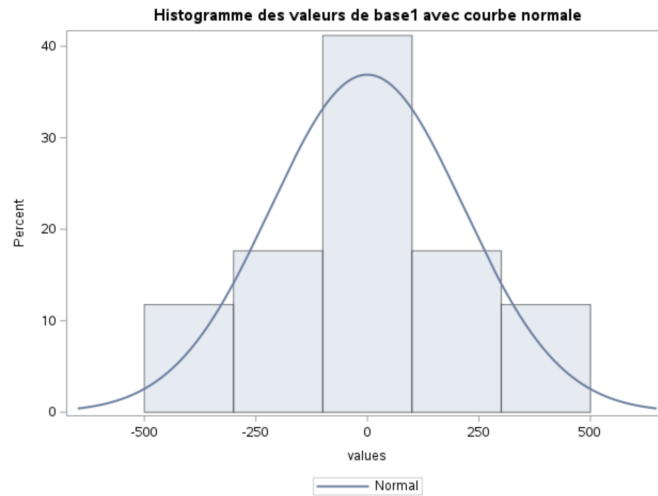


Fig. 1. Histogramme

Deuxièmement, nous décidons d'étudier les quantiles théoriques. Pour ce faire, nous classons nos valeurs dans l'ordre croissant, puis nous les centrons et les réduisons. Nous obtenons ainsi les valeurs présentées (cf. Fig. 3).

Obs	values	id	alpha	quantiles_theoriques
1	-2.07945	1	0.05882	-1.56473
2	-1.48817	2	0.11765	-1.18683
3	-1.07865	3	0.17647	-0.92890
4	-0.63853	4	0.23529	-0.72152
5	-0.60681	5	0.29412	-0.54140
6	-0.35947	6	0.35294	-0.37739
7	-0.16682	7	0.41176	-0.22301
8	-0.10560	8	0.47059	-0.07379
9	0.01328	9	0.52941	0.07379
10	0.33734	10	0.58824	0.22301
11	0.36088	11	0.64706	0.37739
12	0.43935	12	0.70588	0.54140
13	0.50524	13	0.76471	0.72152
14	0.58187	14	0.82353	0.92890
15	1.08321	15	0.88235	1.18683
16	1.49689	16	0.94118	1.56473
17	1.70543	17	1.00000	.

Fig. 2. Valeurs centrées réduites

Une fois les quantiles théoriques obtenus, nous les comparons avec les quantiles observés. Nous remarquons que les points du nuage de dispersion semblent s'aligner autour d'une droite, ce qui suggère que les données sont normalement distribuées (cf. Fig. 3).

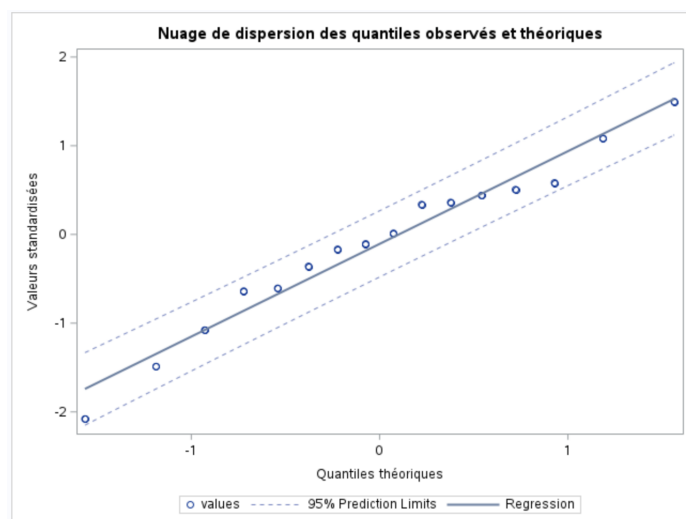


Fig. 3. Nuage de dispersion des quantiles théoriques et des valeurs standardisées

Pour confirmer notre analyse, nous décidons d'utiliser un QQ-plot à l'aide des procédures "UNIVARIATE" et "CAPABILITY". La procédure UNIVARIATE nous montre que les valeurs sont distribuées normalement, avec à gauche un histogramme symétrique en forme de cloche et à droite un nuage de points qui s'aligne autour d'une droite (cf. Fig. 4). On observe les mêmes résultats avec la procédure CAPABILITY. De ce fait, nous en concluons que la distribution de ces valeurs suit une loi de Laplace-Gauss.

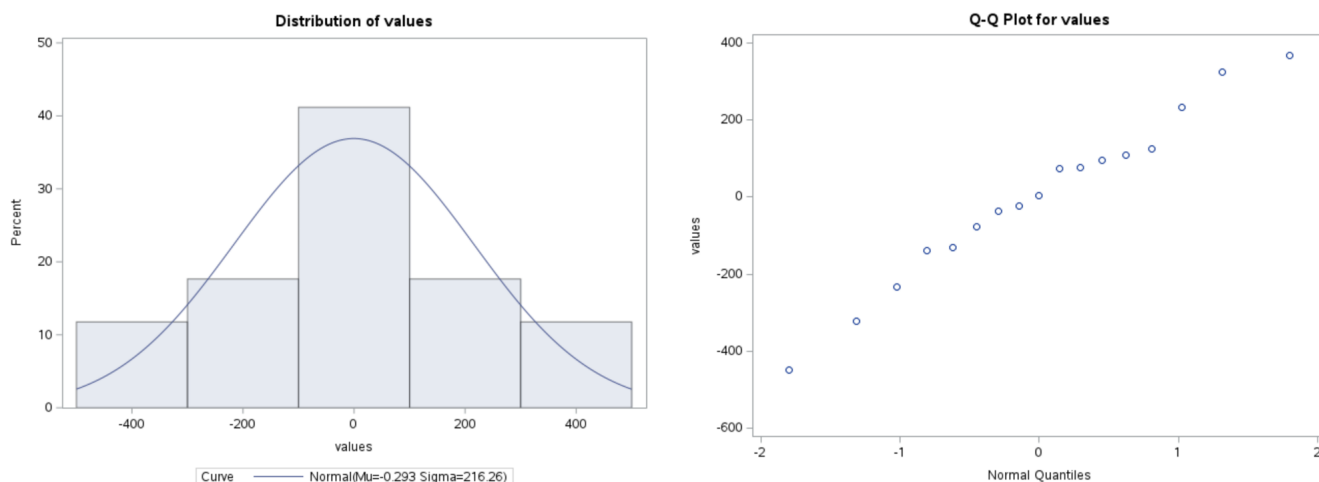


Fig. 4. Distribution final

2 Exercice 2

À travers cet exercice, nous avons aidé le professeur Tournesol à identifier les distributions associées à différents échantillons. Pour ce faire, nous avons utilisé la procédure "UNIVARIATE" de SAS, en particulier avec les options "HISTOGRAM" et "QQPLOT". Cette procédure nous permet de visualiser efficacement la distribution des échantillons, facilitant ainsi la détermination de leur loi de probabilité associée.

2.1 Échantillon n°2

Sur ce second échantillon, l'analyse de l'histogramme nous fait penser à une distribution uniforme. Nous n'observons ni forme en cloche ni queue spécifique (cf. Fig. 5). Par conséquent, nous supposons que l'échantillon n°2 suit une loi uniforme.

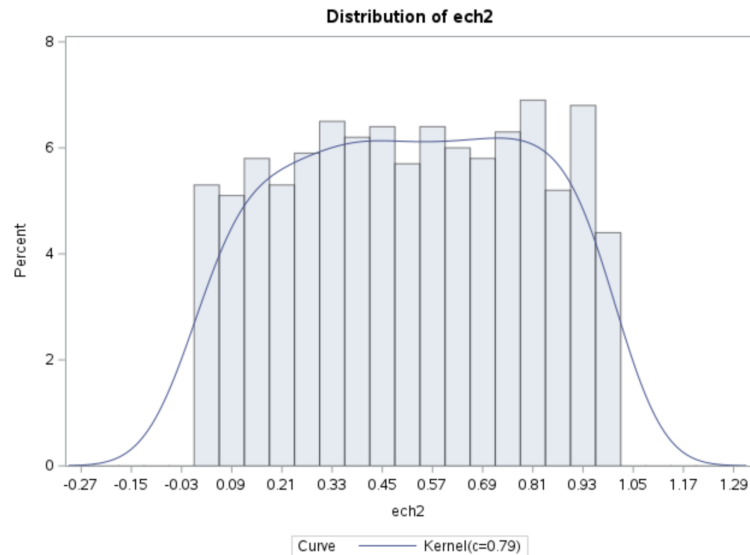


Fig. 5. Échantillon n°2

2.2 Échantillon n°3

Sur ces deux visuels, nous distinguons clairement la loi gaussienne (cf. Fig. 6). Premièrement, sur l'histogramme, nous observons une distribution symétrique en forme de cloche. Sur le Q-Q plot, nous remarquons que les points semblent s'aligner sur une droite. De ce fait, nous en concluons que l'échantillon n°3 suit une loi gaussienne.

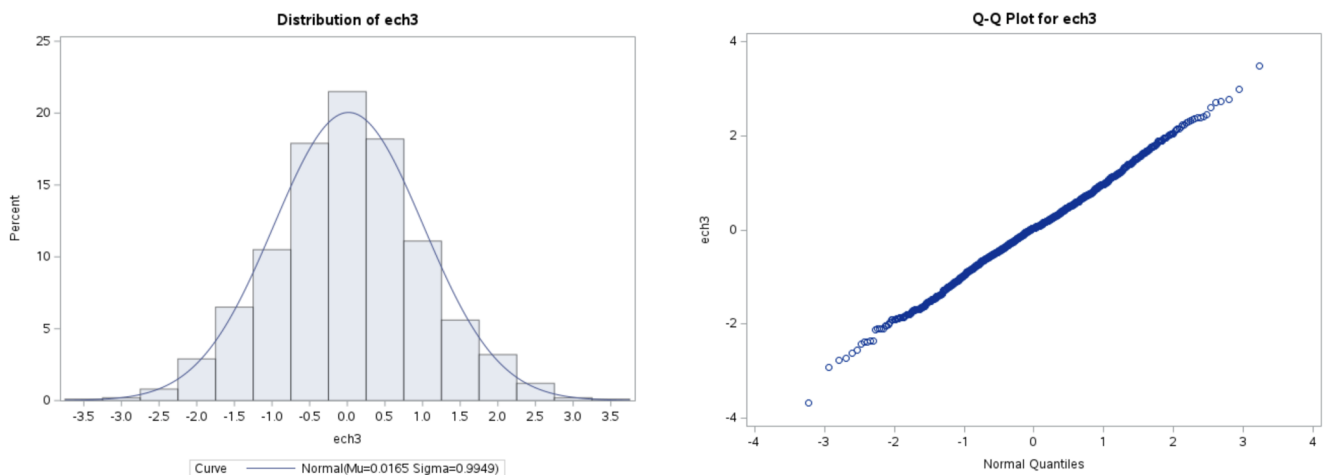


Fig. 6. Échantillon n°3

2.3 Échantillon n°4

Sur ce dernier échantillon, nous observons une asymétrie dans l'histogramme de distribution, avec une queue plus longue à droite et la majorité des valeurs concentrées à gauche. Cette description correspond à la distribution de la

loi lognormale, représentée par la courbe bleue sur notre histogramme (cf. Fig. 7). De plus, les quantiles de notre échantillon et ceux de la loi lognormale semblent s'aligner sur une droite dans le Q-Q plot.

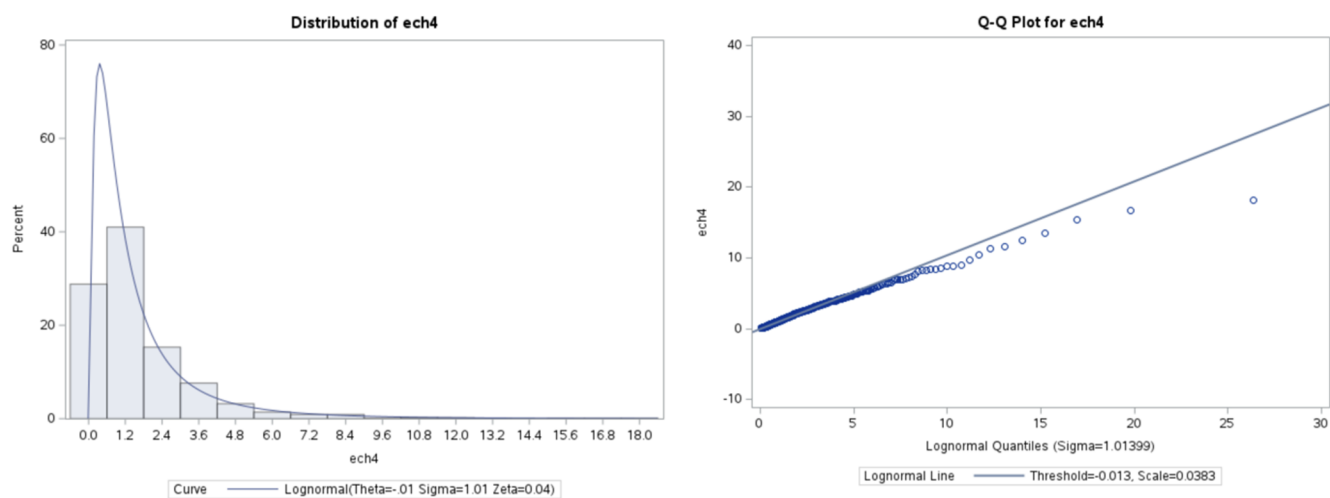


Fig. 7. Échantillon n°4

3 Exercice 3

À travers cette exercice, on se propose d'étudier l'influence de la taille d'un échantillon sur la qualité d'un estimateur. On utilisera un macro programme qui tire 100 échantillons aléatoires de taille n d'une variable `var1` à partir d'une population de référence. C'est cet estimateur de la vraie moyenne que nous allons utiliser :

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^n x_i$$

3.1 Moyenne et variance de la variable `var1`

Cette sortie SAS nous indique la moyenne de la variable `var1`, celle-ci vaut environ -0.01 et la variance est d'environ 1.1

La procédure MEANS	
Variable d'analyse : <code>var1</code>	
Moyenne	Variance
-0.0100780	1.1001492

Fig. 8. Sortie SAS de la moyenne et variance de la variable `var1`

3.2 Distribution de l'estimateur de la moyenne pour taille d'échantillon = 2

Voici l'histogramme et l'estimateur de densité non paramétrique de type noyau de la variable `var1` avec une taille d'échantillon égale à 2 (cf. Fig. 9). Ce que l'on peut retenir de cette sortie est que la distribution est plutôt asymétrique avec une chute de la densité quand les valeurs de `var1` dépassent 0.0. Néanmoins, le pic de la courbe se situe très proche de la moyenne de la population.

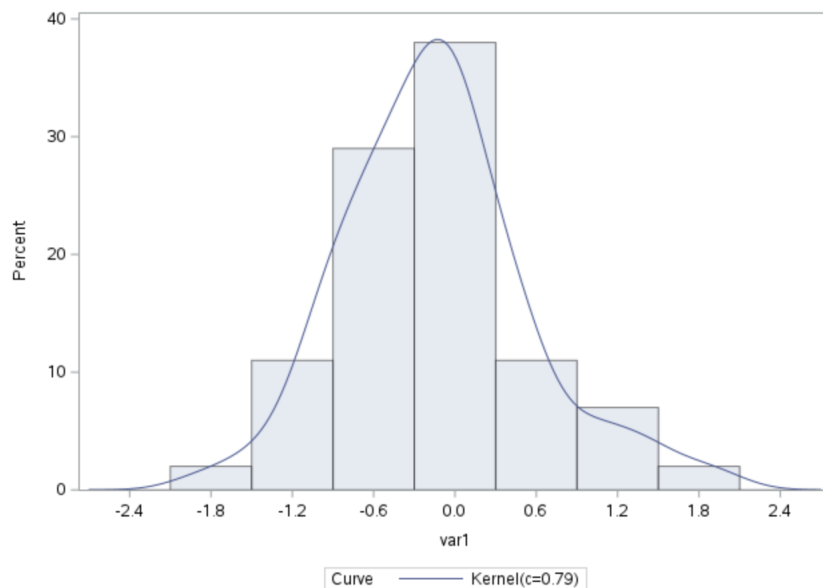


Fig. 9. Distribution de l'estimateur de la moyenne pour une taille d'échantillon = 2

3.3 Distribution de l'estimateur de la moyenne pour taille d'échantillon = 5

Voici l'histogramme et l'estimateur de densité non-paramétrique de type noyau de la variable `var1` avec un taille d'échantillon égal à 5 (cf. Fig. 10). On retrouve des différence par rapport au graphique ci-dessus (cf. Fig. 9) car la grande majorité des observations se situe entre : -1,8 et 1,8 pour les valeurs de `var1` en abscisse. On remarque encore que la distribution est asymétrique mais que les valeurs estimées sont encore très proche de la moyenne de la population.

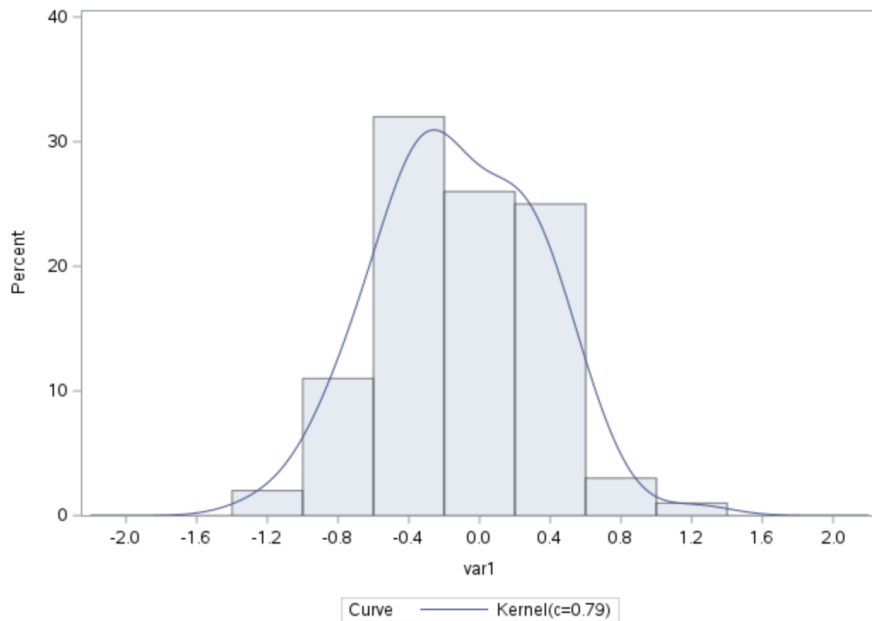


Fig. 10. Distribution de l'estimateur de la moyenne pour une taille d'échantillon = 5

3.4 Distribution de l'estimateur de la moyenne pour taille d'échantillon = 10

Voici l'histogramme et l'estimateur de densité non-paramétrique de type noyau de la variable `var1` avec un taille d'échantillon égal à 10 (cf. Fig. 11). On peut désormais émettre l'hypothèse que pour un échantillon qui devient grand, nos observations se concentre autour de la moyenne de la population avec une erreur encore plus faible que pour les essais précédents. On note aussi que la distribution est semblable à une ladite normale.

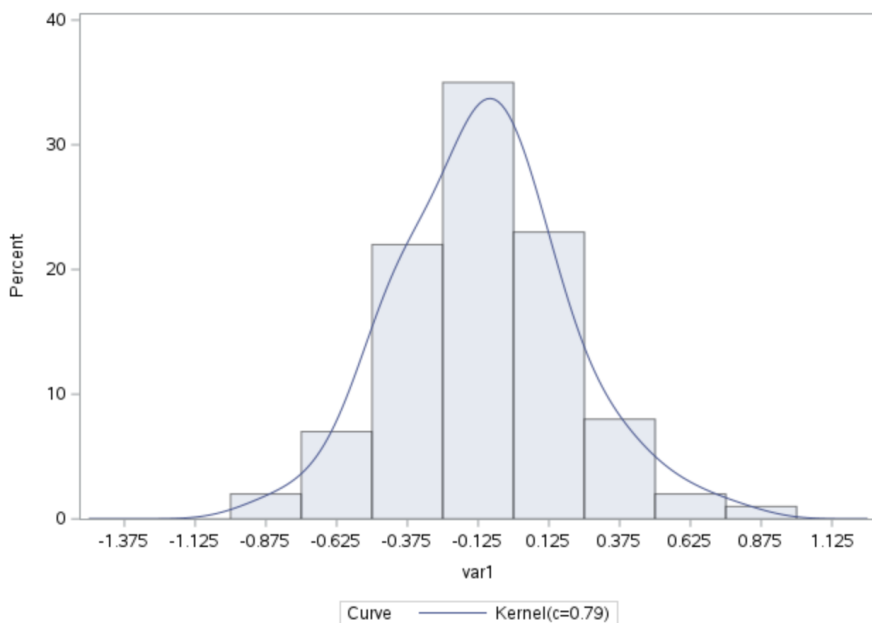


Fig. 11. Distribution de l'estimateur de la moyenne pour une taille d'échantillon = 10

3.5 Distribution de l'estimateur de la moyenne pour taille d'échantillon = 20

Voici l'histogramme et l'estimateur de densité non-paramétrique de type noyau de la variable var1 avec un taille d'échantillon égal à 20 (cf. Fig. 12). Nous remarquons encore une fois que la distribution semble symétrique et normale au vu de la courbe de densité..

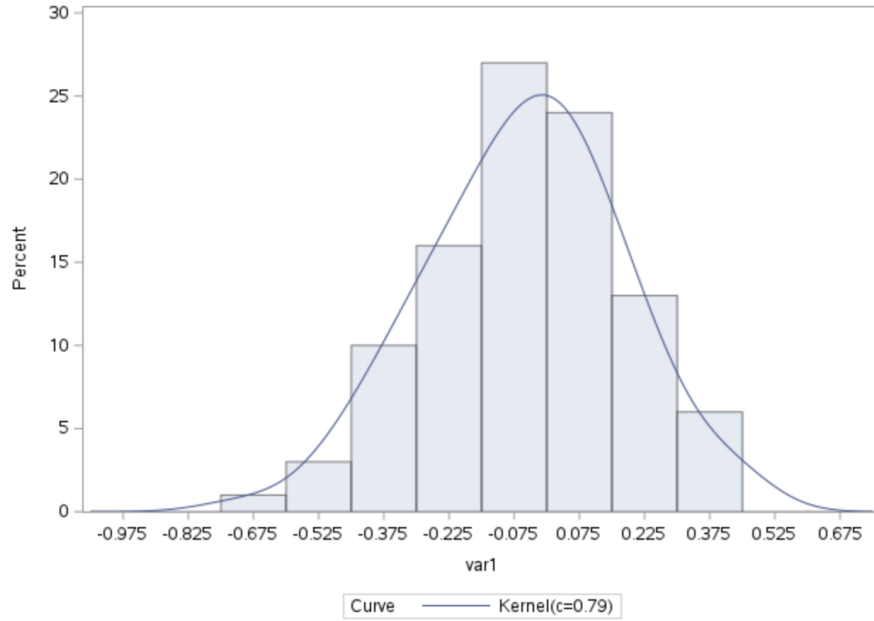


Fig. 12. Distribution de l'estimateur de la moyenne pour une taille d'échantillon = 20

3.6 Distribution de l'estimateur de la moyenne pour taille d'échantillon = 50

Voici l'histogramme et l'estimateur de densité non-paramétrique de type noyau de la variable var1 avec une taille d'échantillon égal à 50 (cf. Fig. 13). Grâce à ce graphique, on peut rejeter en partie l'hypothèse émise précédemment puisqu'on remarque un que la valeur où se trouve le pic de la courbe de densité est éloigné de la moyenne de la population. De plus, cette distribution semble légèrement asymétrique.

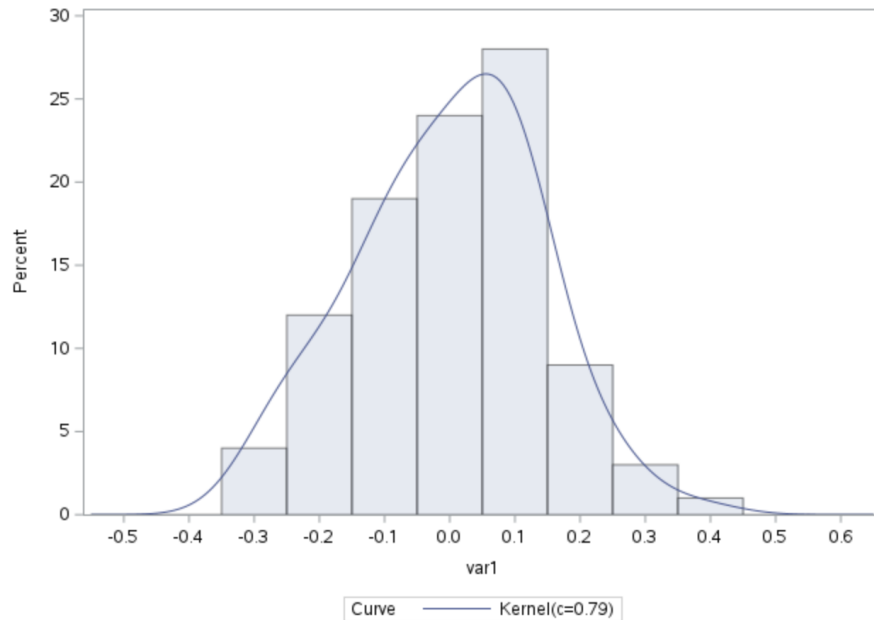


Fig. 13. Distribution de l'estimateur de la moyenne pour une taille d'échantillon = 50

Conclusion

Dans le cadre du première exercice, notre objectif était de déterminer la distribution de probabilité la mieux adaptée à notre échantillon. En utilisant des outils tels qu'un histogramme et un Q-Q plot, nous avons procédé à cette évaluation. Nos résultats ont abouti à la conclusion que la loi normale est celle qui représente le mieux notre échantillon.

Dans le cadre du second exercice, notre objectif était de déterminer les distributions de probabilité les mieux adaptées à chaque échantillon. En utilisant la même méthodologie, nous avons entrepris cette recherche. Les résultats de nos investigations ont abouti à l'identification des distributions suivantes : pour l'échantillon n°1 la loi exponentielle , pour l'échantillon n°2 la loi uniforme , pour l'échantillon n°3 la loi gaussienne et enfin, pour l'échantillon n°4, la loi lognormale.

À travers le troisième exercice, on s'est demandé si la taille de l'échantillon avait une influence sur la qualité de l'estimateur. D'après nos simulations, on se rend compte que pour un échantillon trop petit, le sommet de la courbe de densité est proche de la statistique de la population, mais la distribution est asymétrique. On a aussi remarqué que quand la taille de l'échantillon était trop grande, ici 50, le sommet de la courbe de densité était éloigné de la vraie moyenne, mais avec une distribution presque symétrique. On peut donc proposer une l'hypothèse qu'il faut que l'échantillon soit supérieur ou égal à 10, mais inférieur à 50 afin d'avoir une estimation la plus proche de la réalité.

Table des figures

1	Histogramme	4
2	Valeurs centrées réduites	4
3	Nuage de dispersion des quantiles théoriques et des valeurs standardisées	5
4	Distribution final	5
5	Échantillon n°2	6
6	Échantillon n°3	6
7	Échantillon n°4	7
8	Sortie SAS de la moyenne et variance de la variable var1	8
9	Distribution de l'estimateur de la moyenne pour une taille d'échantillon = 2	8
10	Distribution de l'estimateur de la moyenne pour une taille d'échantillon = 5	9
11	Distribution de l'estimateur de la moyenne pour une taille d'échantillon = 10	9
12	Distribution de l'estimateur de la moyenne pour une taille d'échantillon = 20	10
13	Distribution de l'estimateur de la moyenne pour une taille d'échantillon = 50	10