COMPTE RENDU

LOGICIELS DE STATISTIQUE

Master 1 Big Data / Travaux Pratiques

Sous la direction du

prof. Walter TINSSON

Résumé

Analyse statistique à l'aide du logiciel statistique SaS

Préparé par

Gaby MAROUN

Année universitaire 2019-2020

Compte Rendu

PREMIERE PARTIE

Considérons l'expérience aléatoire suivante : on lance deux dés équilibrés et on note la plus petite des 2 valeurs obtenues. Désignons par X la v.a.r. égale au résultat obtenu. On s'intéresse dans cette partie à la loi de X.

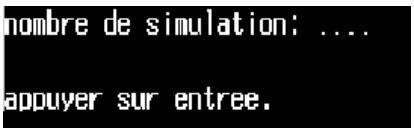
Approche Statistique

01 Ecrire un programme SAS simulant la loi de X (une fenêtre interactive sera utilisée pour demander à l'utilisateur le nombre de simulations désirées). En sortie, le programme affichera les résultats sous forme d'histogramme horizontal.

Réponse :

```
/*quest1*/
%window fen
#8 'nombre de simulation: ' x 9 attr=underline
#10 'appuyer sur entree...';
%macro inter;
	%display fen;
%mend inter;
Data Simulation;
	%inter;
	res=&x;
	put res=;
	Do i=1 to res;
		E1=int(6*ranuni(0)+1);
		E2=int(6*ranuni(0)+1);
		x=min(E1,E2);
		output;
	end;
run;
data anno;
   length function color $8;
   set Simulation;
   style="'Albany AMT'";
run;
axis1 minor=none label=('x');
proc gchart data=Simulation;
	title "Simulation de la loi de X sous forme d'histogramme horizontal";
	hbar x/space=0 anno=anno raxis=axis1 patternid=midpoint midpoints=1 2 3 4 5 6;
run;
```

On aura la fenêtre interactive comme suit :

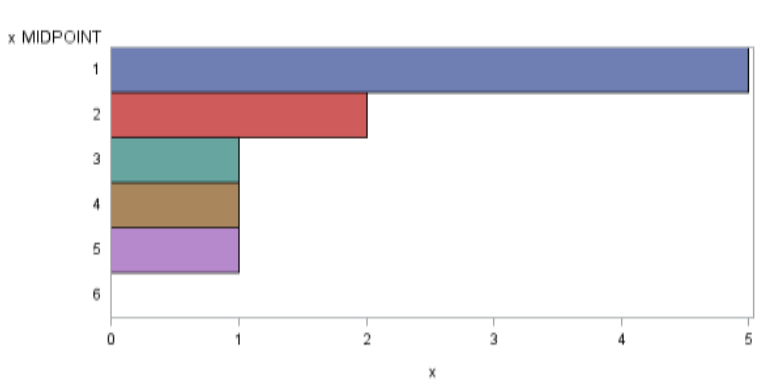


02 Lancer le programme pour 10, 100, 1.000 et 10.000 simulations. Quelles informations peut-on en tirer sur la loi de X ?

Réponse :

pour 10:

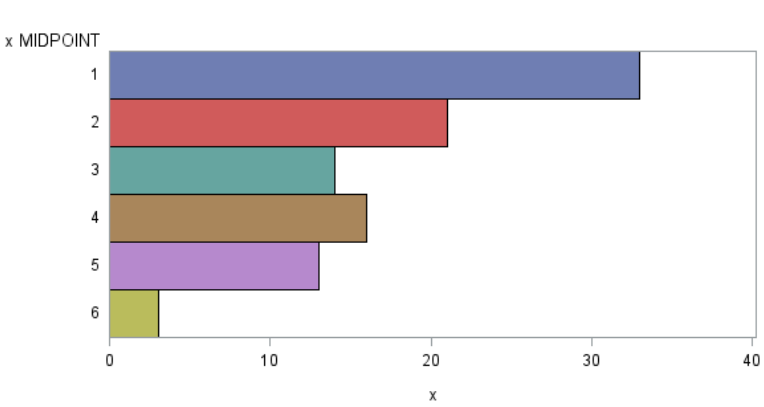
Simulation de la loi de X sous forme d'histogramme horizontal



x MIDPOINT	FREQ.	CUM. FREQ.	PCT.	CUM. PCT.
1	5	5	50	50
2	2	7	20	70
3	1	8	10	80
4	1	9	10	90
5	1	10	10	100
6	0	10	0	100

Pour 100 :

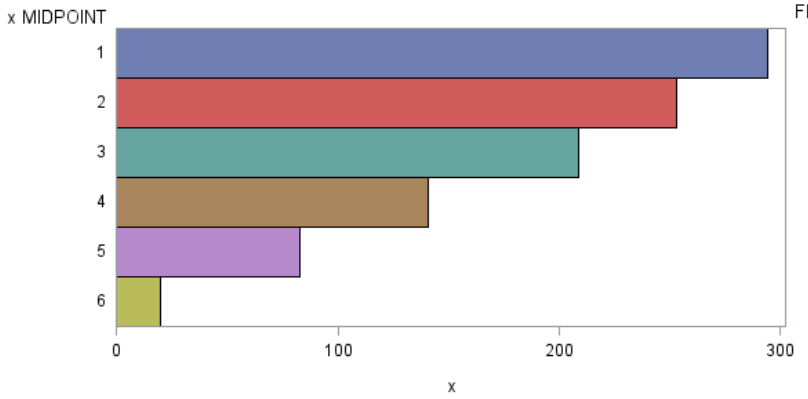
Simulation de la loi de X sous forme d'histogramme horizontal



x MIDPOINT	FREQ.	CUM. FREQ.	PCT.	CUM. PCT.
1	33	33	33	33
2	21	54	21	54
3	14	68	14	68
4	16	84	16	84
5	13	97	13	97
6	3	100	3	100

Pour 1000 :

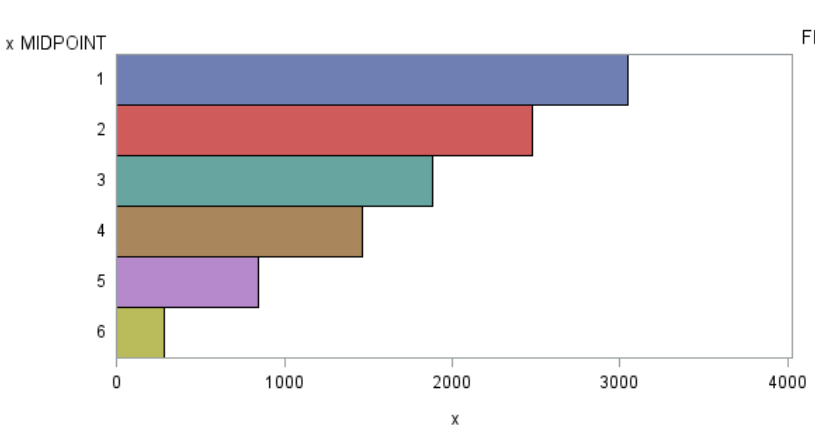
Simulation de la loi de X sous forme d'histogramme horizontal



x MIDPOINT	FREQ.	CUM. FREQ.	PCT.	CUM. PCT.
1	294	294	29.4	29.4
2	253	547	25.3	54.7
3	209	756	20.9	75.6
4	141	897	14.1	89.7
5	83	980	8.3	98.0
6	20	1000	2.0	100.0

Pour 10000 :

Simulation de la loi de X sous forme d'histogramme horizontal



x MIDPOINT	FREQ.	CUM. FREQ.	PCT.	CUM. PCT.
1	3050	3050	30.50	30.50
2	2475	5525	24.75	55.25
3	1886	7411	18.86	74.11
4	1461	8872	14.61	88.72
5	844	9716	8.44	97.16
6	284	10000	2.84	100.00

On peut déduire que,

- Lorsque la valeur de X est grande, la probabilité de la recevoir sera faible.
- En plus, c’est une loi discrète car elle ne prend qu'un nombre fini ou dénombrable de valeurs.

03 Transformer le programme pour qu'il traite, cette fois, les données à l'aide de la procédure MEANS. Effectuer les mêmes types de simulations qu'à la question 2 Quelles informations peut-on en tirer sur les caractéristiques de position et de dispersion de X ?

Réponse :

```
%window fen
#8 'nombre de simulation: ' x 9 attr=underline
#10 'appuyer sur entree...';
%macro inter;
%display fen;
%mend inter;
Data Simulation;
      %inter;
      res=&x;
      put res=;
      Do i=1 to res;
      E1=int(6*ranuni(0)+1);
      E2=int(6*ranuni(0)+1);
      x=min(E1,E2);
      output;
      end;
run;
proc means data=Simulation;
run;
```

Pour 10 :

The SAS System

The MEANS Procedure

Variable	N	Mean	Std Dev	Minimum	Maximum
X	10	2.5000000	1.3540064	1.0000000	5.0000000

Pour 100 :

The SAS System

The MEANS Procedure

Variable	N	Mean	Std Dev	Minimum	Maximum
X	100	2.5800000	1.3271067	1.0000000	6.0000000

Pour 1000 :

The SAS System

The MEANS Procedure

Variable	N	Mean	Std Dev	Minimum	Maximum
X	1000	2.5850000	1.4480580	1.0000000	6.0000000

Pour 10000 :

The SAS System

The MEANS Procedure

Variable	N	Mean	Std Dev	Minimum	Maximum
X	10000	2.5282000	1.4010000	1.0000000	6.0000000

On peut en déduire que la position de X est caractérisée par une $Moyenne \approx 2.5$ et par une dispersion représentée par $l'Ecart\ type \approx 1.4$.

Approche Probabiliste

04 Proposer une modélisation pour l'expérience aléatoire étudiée. Déterminer la loi de X. Comparer ces résultats avec ceux de la question 2.

Réponse :

On suppose *E1* le résultat du dé 1 et *E2* celui du dé 2. Le tableau suivant, montre toutes les possibilités de tirages potentiels :

E1\E2	**1**	**2**	**3**	**4**	**5**	**6**
1	***(1,1)***	***(1,2)***	***(1,3)***	***(1,4)***	***(1,5)***	***(1,6)***
2	***(2,1)***	***(2,2)***	***(2,3)***	***(2,4)***	***(2,5)***	***(2,6)***
3	***(3,1)***	***(3,2)***	***(3,3)***	***(3,4)***	***(3,5)***	***(3,6)***
4	***(4,1)***	***(4,2)***	***(4,3)***	***(4,4)***	***(4,5)***	***(4,6)***
5	***(5,1)***	***(5,2)***	***(5,3)***	***(5,4)***	***(5,5)***	***(5,6)***
6	***(6,1)***	***(6,2)***	***(6,3)***	***(6,4)***	***(6,5)***	***(6,6)***

On peut déduire la modélisation du v.a.r. X sous forme du tableau suivant :

X	**1**	**2**	**3**	**4**	**5**	**6**
Probabilité	11/36	9/36	7/36	5/36	3/36	1/36

La loi de X est discrète (comme déjà observé dans 02) a une fonction de masse de la forme suivante :

$$P(X = k) = P(E1 = k).P\ (E2 = k \ldots,6)\ + P(E2 = k).P\ (E1 = k + 1,\ldots,6)$$

Pour $k = 1, \ldots 6$ en utilisant l'indépendance de *E1* et *E2*. On a alors:

$$P(X = k) = \frac{1}{6} \times \frac{6 - k + 1}{6} + \frac{1}{6} \times \frac{6 - k}{6} = \frac{13 - 2k}{36}$$

$$\boldsymbol{P(X = k) = \frac{13-2k}{36}}$$

Nous pouvons constater que si k augmente ↑, $P\ (X\ =\ k)$ diminue ↓, ce qui est déjà montré par les observations numériques de la question 02.

05 Calculer l'espérance E (X) et la variance Var (X). Comparer ces résultats avec ceux de la question 3.

Réponse :

$$E(X) = \sum_{k=1}^{6} x \times P(X = k)$$

$$= \sum_{k=1}^{6} x \times \frac{13 - 2k}{36} = 1 \times \frac{13 - 2 \times 1}{36} + 2 \times \frac{13 - 2 \times 2}{36} + \cdots + 6 \times \frac{13 - 2 \times 6}{36} = \frac{91}{36}$$

$$= 2.53$$

$$Var(X) = E[X - E(X)]^2 = \sum_{k=1}^{6} x^2 \times P(X = k) - [\sum_{k=1}^{6} x \times P(X = k)]^2 = \frac{301}{36} - E(X)^2 = 1.97$$

$$Var(X) = \sigma(X)^2 => Ecart\ Type\ (X) = \sigma(X) = \sqrt{Var(X)} \approx 1.4$$

En comparant les résultats obtenus là avec ceux du question 02, on peut remarquer que :

$$\boldsymbol{E(X)_{calculé} = E(X)_{observé} \approx 2.5}$$

$$\boldsymbol{\sigma(X)_{calculé} = \sigma(X)_{observé} \approx 1.4}$$

DEUXIEME PARTIE

On décide maintenant d'organiser un jeu selon les règles suivantes, Un joueur lance 2 dés, s'intéresse au minimum X des valeurs obtenues, et se reporte au tableau ci-dessous pour le résultat :

$X = 1$	Il perd 10 €		$X = 4$	Il gagne 5 €
$X = 2$	Il perd 5 €		$X = 5$	Il gagne 10 €
$X = 3$	Il ne perd rien		$X = 6$	Il gagne d €

Nous notons Y la v.a.r. correspondant au gain de **l'organisateur** (donc si $X = 1$ alors $Y = 10$, (si $X = 4$ alors $Y = -5$ etc.…). Le problème qui se pose à nous, dans un premier temps, est la détermination de la valeur **d** correspondant au gros lot.

06 Ecrire un programme SAS réalisant des simulations de la v.a.r. Y (une fenêtre interactive sera utilisée pour demander à l'utilisateur la valeur du paramètre d et le nombre de simulations souhaitées). En sortie, ce programme analysera les résultats à l'aide de la procédure UNIVARIATE.

Réponse :

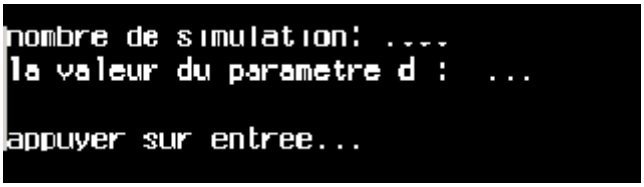
```
/*quest6*/
%window fen
#8 'nombre de simulation: ' x 9 attr=underline
#9 'la valeur du parametre d : ' d 9 attr=underline
#11 'appuyer sur entree...';
%macro inter;
      %display fen;
%mend inter;
Data Simulation;
      %inter;
      res=&x;
      put res=;
      para=&d;
      put para=;
      Do i=1 to res;
            E1=int(6*ranuni(0)+1);
            E2=int(6*ranuni(0)+1);
            x=min(E1,E2);
            select(x);
                  when(1) Y=10;
                  when(2) Y=5;
                  when(3) Y=0;
                  when(4) Y=-5;
                  when(5) Y=-10;
                  when(6) Y=-para;
                  end;;
       output;
      end;
run;

proc Univariate plot data=Simulation;
 Var Y;
run;
```

07 Une première solution consiste à fixer d "n'importe comment", en espérant que les résultats seront satisfaisants en pratique. Deux organisateurs différents (qui n'ont jamais étudié les probabilités) proposent de fixer le gros lot aux valeurs suivantes 15 euros pour l'un et 200 euros pour l'autre. Etudier ces deux propositions à l'aide du programme de la question précédente. Lancer 5 fois le programme pour chacune des valeurs de d (on fixera le nombre de simulations à 1.000). Comparer et expliquer les résultats obtenus avec les deux valeurs. Que concluez-vous en pratique sur le choix des valeurs proposées ?

Réponse :

La fenêtre interactive ressemble à ceci :



Pour d=200 :

Pour d=15 :

The SAS System

Moments			
N	1000	**Sum Weights**	1000
Mean	-3.515	**Sum Observations**	-3515
Std Deviation	35.1575857	**Variance**	1236.05583

Moments(d=15)			
N	1000	**Sum Weights**	1000
Mean	2.315	**Sum Observations**	2315
Std Deviation	6.94723247	**Variance**	48.264039

Moments			
N	1000	**Sum Weights**	1000
Mean	-3.545	**Sum Observations**	-3545
Std Deviation	36.3115202	**Variance**	1318.5265

Moments			
N	1000	**Sum Weights**	1000
Mean	2.445	**Sum Observations**	2445
Std Deviation	7.06731363	**Variance**	49.9469219

Moments			
N	1000	**Sum Weights**	1000
Mean	-3.365	**Sum Observations**	-3365
Std Deviation	36.8830874	**Variance**	1360.36214

Moments(d=15)			
N	1000	**Sum Weights**	1000
Mean	2.66	**Sum Observations**	2660
Std Deviation	6.91174162	**Variance**	47.7721722

Moments			
N	1000	**Sum Weights**	1000
Mean	-1.75	**Sum Observations**	-1750
Std Deviation	32.3447628	**Variance**	1046.18368

Moments			
N	1000	**Sum Weights**	1000
Mean	2.12	**Sum Observations**	2120
Std Deviation	6.84122814	**Variance**	46.8024024

Moments			
N	1000	**Sum Weights**	1000
Mean	-3.16	**Sum Observations**	-3160
Std Deviation	35.2335892	**Variance**	1241.40581

Moments			
N	1000	**Sum Weights**	1000
Mean	2.785	**Sum Observations**	2785
Std Deviation	6.79447864	**Variance**	46.1649399

D'après le tableau « Moments »,

En premier lieu, On peut observer que la moyenne des v.a.r. Y pour $d = 200$ est négatif ce qui est très grand par rapport aux autres valeurs de Y, avec au minimum ≈ -3.6 et au maximum ≈ -1.7, tandis que pour $d = 15$, la moyenne sera positive avec au minimum ≈ 2.1 et au maximum ≈ 2.7. La position de Y est alors fortement influencée par la valeur de d.

En deuxième lieu, on peut voir que pour $d = 15$, l'écart entre les v.a.r. $Y = 6$, alors équilibré et plus stable contrairement au grand décalage et dispersion entre les valeurs pour $d = 200$ avec un écart > 32. Ce qui influence inévitablement les variances.

La somme des observations pour $d = 15$ donne toujours un gain supérieur à 2100 pour l'organisateur alors que pour $d = 200$, l'organisateur perd plus que 1700.

Pour d=200 :

Basic Statistical Measures			
Location		**Variability**	
Median	5.00000	**Variance**	1046
Mode	10.00000	**Range**	210.00000
		Interquartile Range	10.00000

Pour d=15 :

Basic Statistical Measures			
Location		**Variability**	
Median	5.00000	**Variance**	49.94692
Mode	10.00000	**Range**	25.00000
		Interquartile Range	10.00000

Basic Statistical Measures			
Location		**Variability**	
Median	5.00000	**Variance**	1241
Mode	10.00000	**Range**	210.00000
		Interquartile Range	15.00000

Basic Statistical Measures			
Location		**Variability**	
Median	5.00000	**Variance**	46.16494
Mode	10.00000	**Range**	25.00000
		Interquartile Range	10.00000

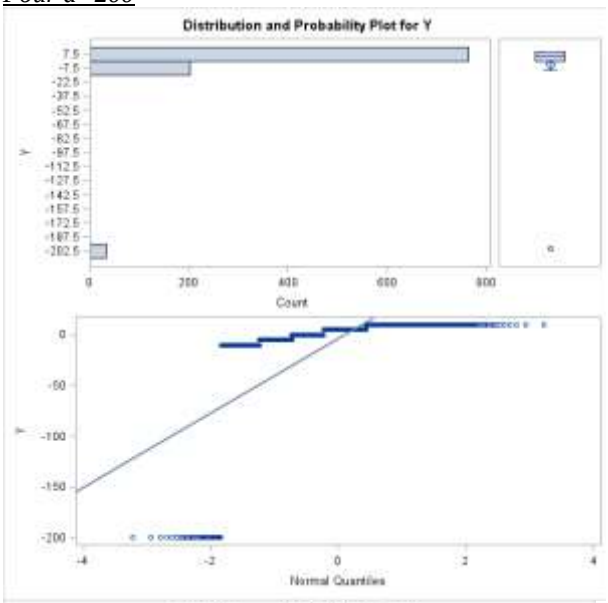
Basic Statistical Measures			
Location		**Variability**	
Median	5.00000	**Variance**	1236
Mode	10.00000	**Range**	210.00000
		Interquartile Range	15.00000

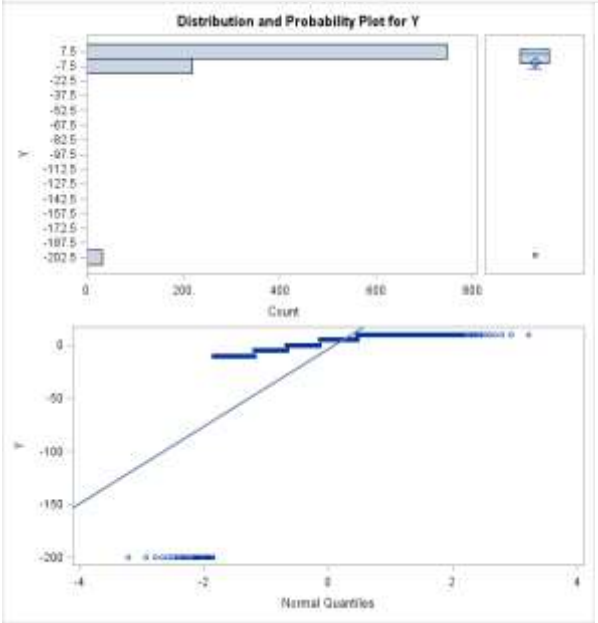
Basic Statistical Measures			
Location		**Variability**	
Median	5.00000	**Variance**	46.80240
Mode	10.00000	**Range**	25.00000
		Interquartile Range	15.00000

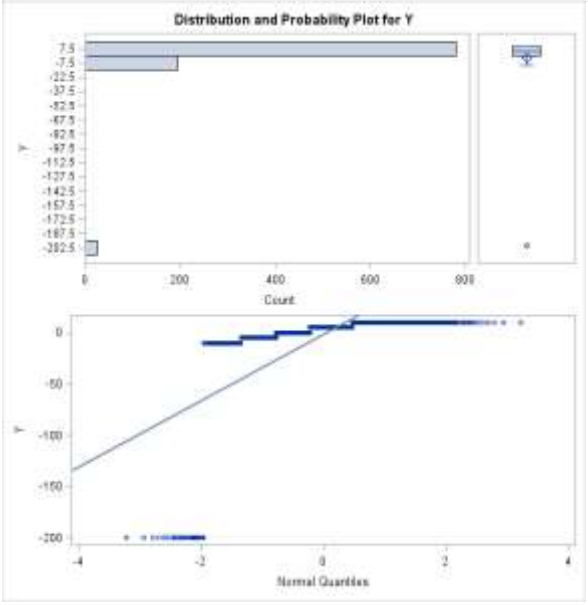
Basant sur cet échantillon des tableaux « Quantiles »,

On peut observer que $la\ médiane\ = 5$ et $la\ mode\ = 10$ sont toujours les mêmes pour toute valeur de d. En plus, l'intervalle entre la valeur max et min est de 25 pour $d = 15$ qu'il est égal à 210 pour $d = 200$.

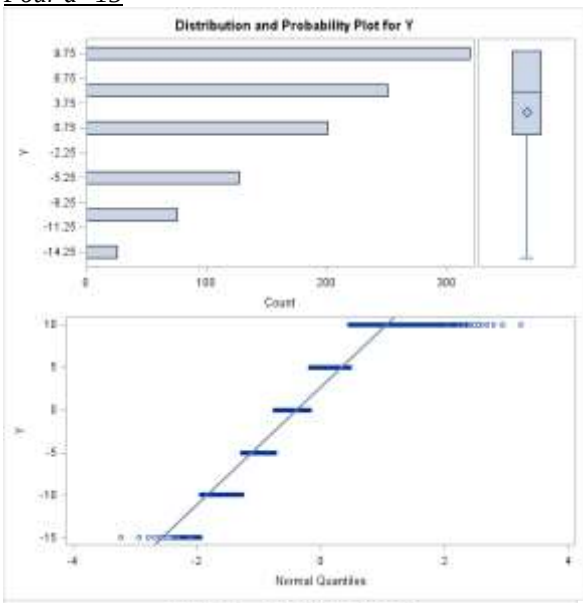
Pour d=200

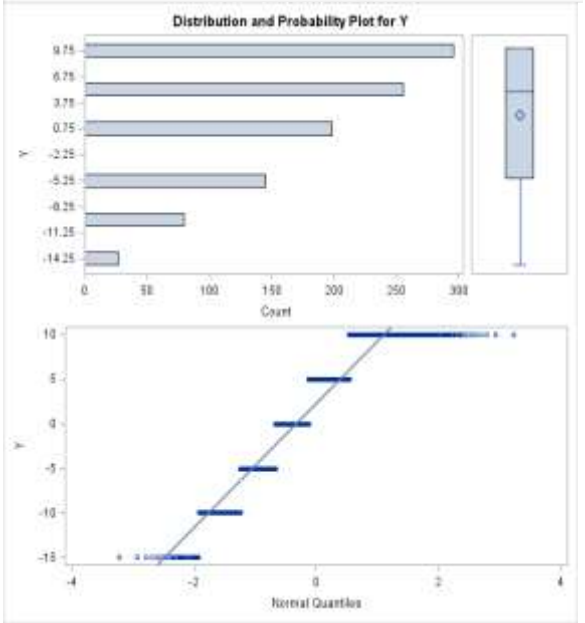


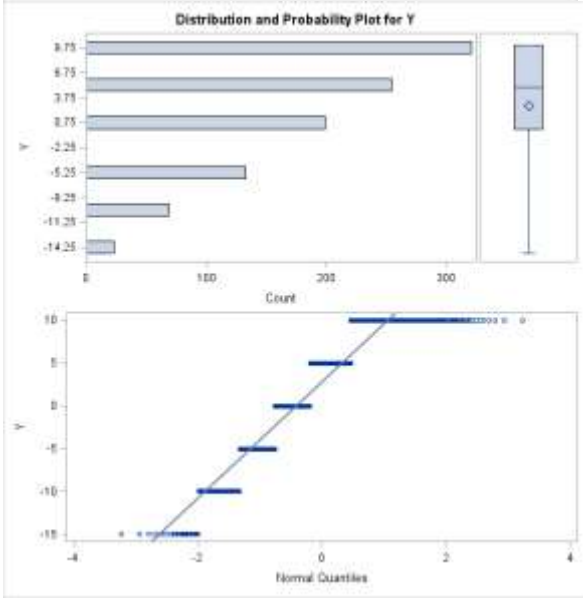


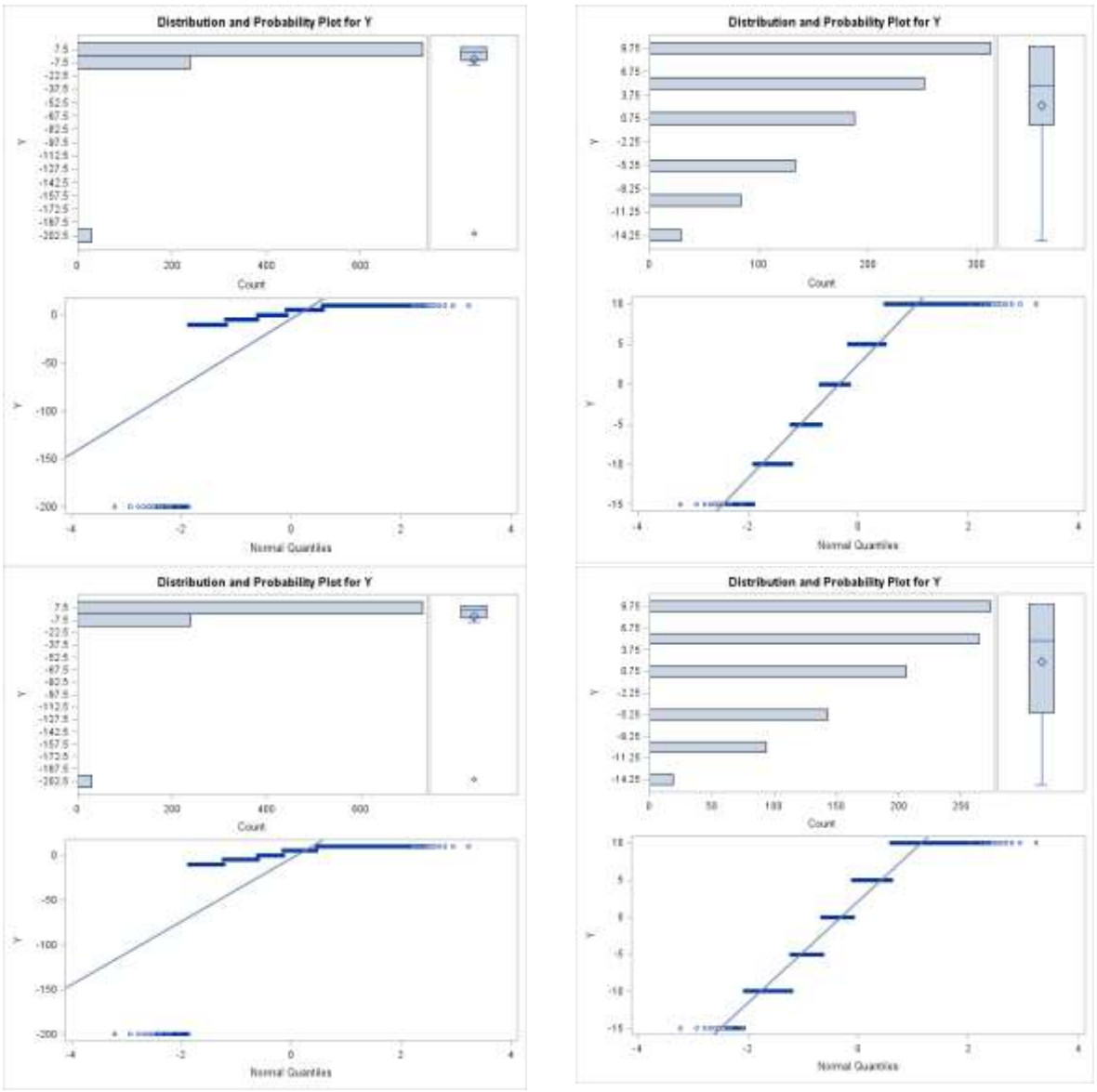


Pour d=15









À partir des graphiques de distribution de Y montrés par,

- Le diagramme « Tige et feuilles », qui représente par des batons la distribution des Y par rapport à la quantité. On peut voir le décalage entre les valeurs pour $d = 200$ tandis que pour $d = 15$, les valeurs sont moins dispersées.
- Le diagramme « Boxplot », qui représente une boîte à moustaches. On voit pour d=200, le point extérieur, qui est très loin de la moyenne.
- La droite de Henry, qui mesure la normalité des observations. C'est clair que pour $d = 200$, la courbe est anormale tandis que pour $d = 15$, les données sont distribuées normalement.

Pour conclure, le choix de la valeur $d = 200$ n'est pas raisonnable mais fait du gain pour l'utilisateur alors que la valeur $d = 15$ est plus raisonnable mais diminue les chances de gain pour l'utilisateur et l'organisateur sera plus favorable à gagné la plupart du temps .

08 Déterminer la loi de Y. En déduire E (Y) et Var (Y). Pour quelle valeur de d le jeu est-il, en moyenne, équitable (au sens où l'organisateur ne gagne pas et ne perd pas d'argent) ?

Réponse :

En utilisant la modélisation des v.a.r. X en question 04, on a pu déduire celle de Y, montré par le tableau suivant :

X	*1*	*2*	*3*	*4*	*5*	*6*
Y	**10**	**5**	**0**	**-5**	**-10**	**-d**
Probabilité	**11/36**	**9/36**	**7/36**	**5/36**	**3/36**	**1/36**

On a la fonction de masse suivante :

$$P(Y=k)=\frac{1}{6}\times\frac{6-k+1}{6}+\frac{1}{6}\times\frac{6-k}{6}=\frac{13-2k}{36}$$

On déduit l'espérance et la variance de Y :

$$E(Y)=\sum_{k=1}^{6} y\times P(Y=k)dx$$

$$=\sum_{k=1}^{6} y\times\frac{13-2k}{36}=10\times\frac{13-2\times1}{36}+5\times\frac{13-2\times2}{36}+\cdots-d\times\frac{13-2\times6}{36}$$

$$=\frac{25}{9}-d\times\frac{1}{36}=\frac{100-d}{36}$$

$$E(Y^2)=\sum_{k=1}^{6} y^2\times P(Y=k)dx=\sum_{k=1}^{6} y^2\times\frac{13-2k}{36}=10^2\times\frac{13-2\times1}{36}+\cdots-d^2\times\frac{13-2\times6}{36}$$

$$=\frac{1750+\ d^2}{36}$$

$$Var(Y)=E[Y-E(Y)]^2=\frac{1750+\ d^2}{36}-\left[\frac{100-d}{36}\right]^2=\frac{53000+200d+35d^2}{1296\,r^2}$$

Pour $E(Y)=0$, le jeu sera équitable, alors :

$$=>\frac{100-d}{36}=0\ =>d=100$$

Pour $d\ =\ 100\ =>\ E\ (Y)\ =\ 0$, ce qui rend le jeu équitable et personne ne perdent ni ne gagnent.

TROISIEME PARTIE

Lorsque n parties sont jouées, on note $Y,..,Yn$ tous les gains associés. Le gain moyen après n parties est done :

$$\overline{Y_n} = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

09 Prouver que $\overline{Y_n}$ est un estimateur sans biais de $\mu = E\ (Y)$. Calculer $Var\ (\overline{Y_n})$.

Réponse :

$$E(\bar{Y}_n) = \frac{1}{n}\sum_{i=1}^{n} Y_i = \frac{1}{n}E\left(\sum_{i=1}^{n} Y_i\right) = \frac{1}{n} \times nE(Y_i) = \mu$$

Ce qui montre que $\overline{Y_n}$ est un estimateur sans biais.

$$var(\bar{Y}_n) = E[\bar{Y} - E(\bar{Y})]^2 = E[\bar{Y} - \mu]^2 = E\left[\frac{1}{n}\sum_{i=1}^{n}(Y_i - \mu)\right]^2 = E\left[\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}(Y_i - \mu)(Y_j - \mu)\right]$$

Or

$$var(\bar{Y}_n) = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} E(Y_i - \mu)(Y_j - \mu)$$

$$= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} cov(Y_i, Y_j) = \frac{1}{n^2}\sum_{i=1}^{n} \sigma^2 + 0 = \frac{\sigma^2}{n}$$

10 Justifier que, pour n grand, on peut supposer que Y, suit une loi normale.

Réponse :

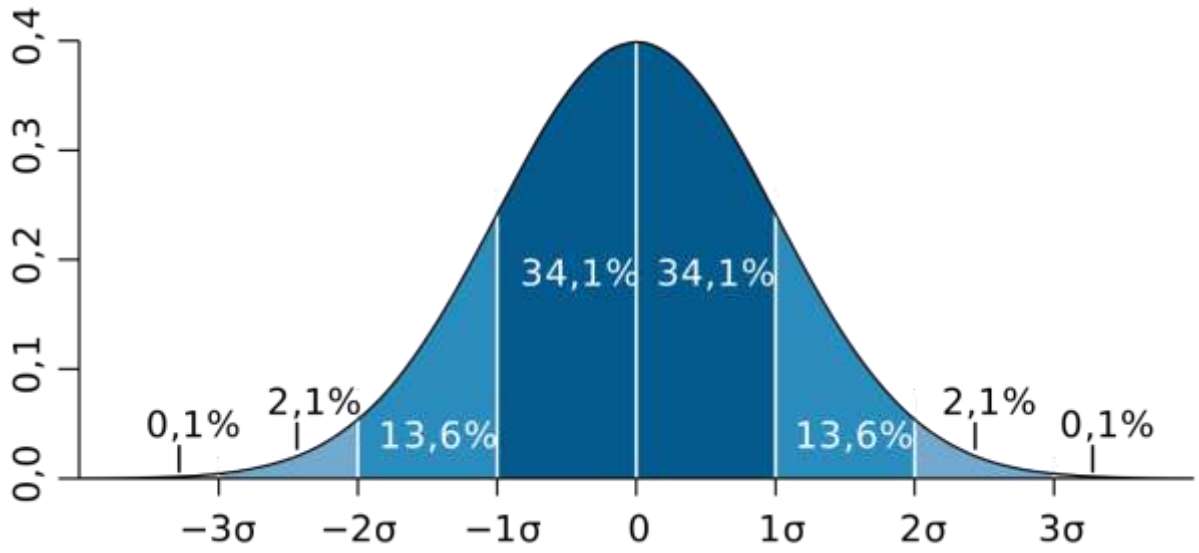
Pour,

$$\lim_{n\to\infty} \overline{Y_n} = \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} Y_i$$

On déduit d'après le Théorème de la limite centrale pour un grand nombre n de Y de même espace de probabilité, indépendantes et identiquement distribuées suivant la même loi.
Et comme déjà démontré en question 09, que l'espérance μ et l'écart type $\frac{\sigma^2}{n}$ de Y_n existent, par suite :

$$\sqrt{n}\frac{(\hat{\mu}-\mu)}{\sigma} \sim N(\mu, \frac{\sigma^2}{n})$$

La Loi Normale, souvent appelée la « courbe en cloche » :

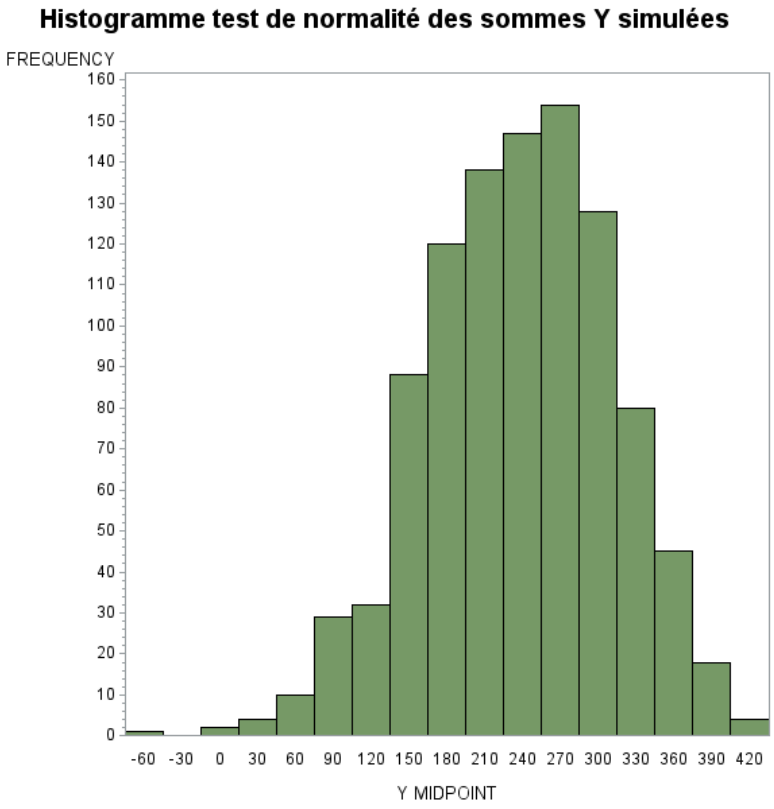


D'après la partie 6) du cours, on a essayé de représenter graphiquement la loi normale et la tester à l'aide des procédures déjà implémenté dans le logiciel SaS. Pour faire ce test, on a utilisé le code suivant :

```
GOPTION RESET=ALL;
Data Simulation;
      res=1000;
      d=15;
      Do i=1 to res;
      Y=0;
            do j=1 to 100;
            E1=int(6*ranuni(0)+1);
            E2=int(6*ranuni(0)+1);
            x=min(E1,E2);
            select(x);
                  when(1) Y1=10;
                  when(2) Y1=5;
                  when(3) Y1=0;
                  when(4) Y1=-5;
                  when(5) Y1=-10;
                  when(6) Y1=-d;
                  end;
            Y=Y+Y1;
            end;
       output;
      end;
run;
```

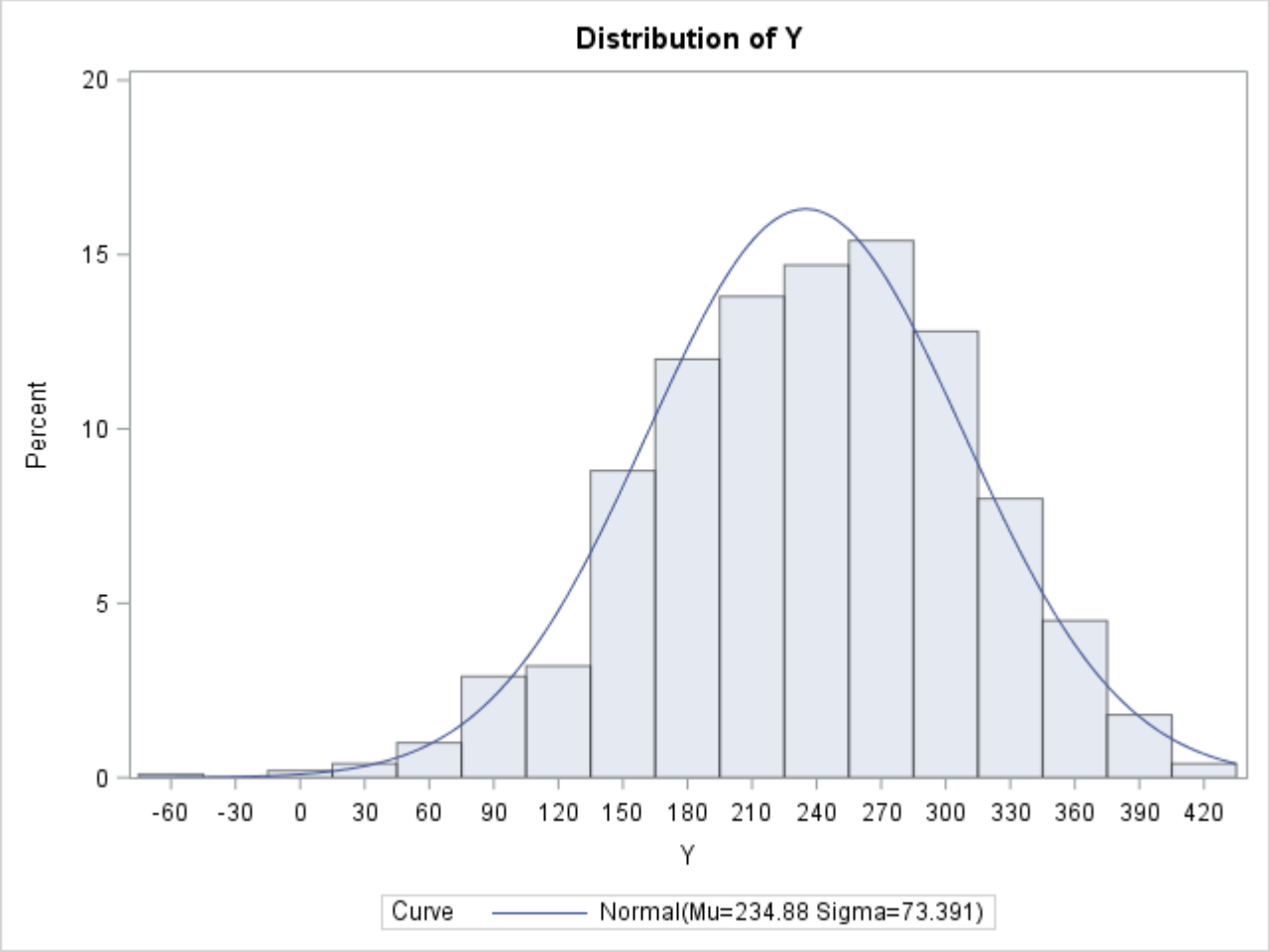
On peut voire la distribution des Y cumulées semblant à ceux de la loi Normale,

```
proc gchart data=Simulation;
      title"Histogramme test de normalité des sommes Y simulées";
      vbar Y/space=0;
run ;
```

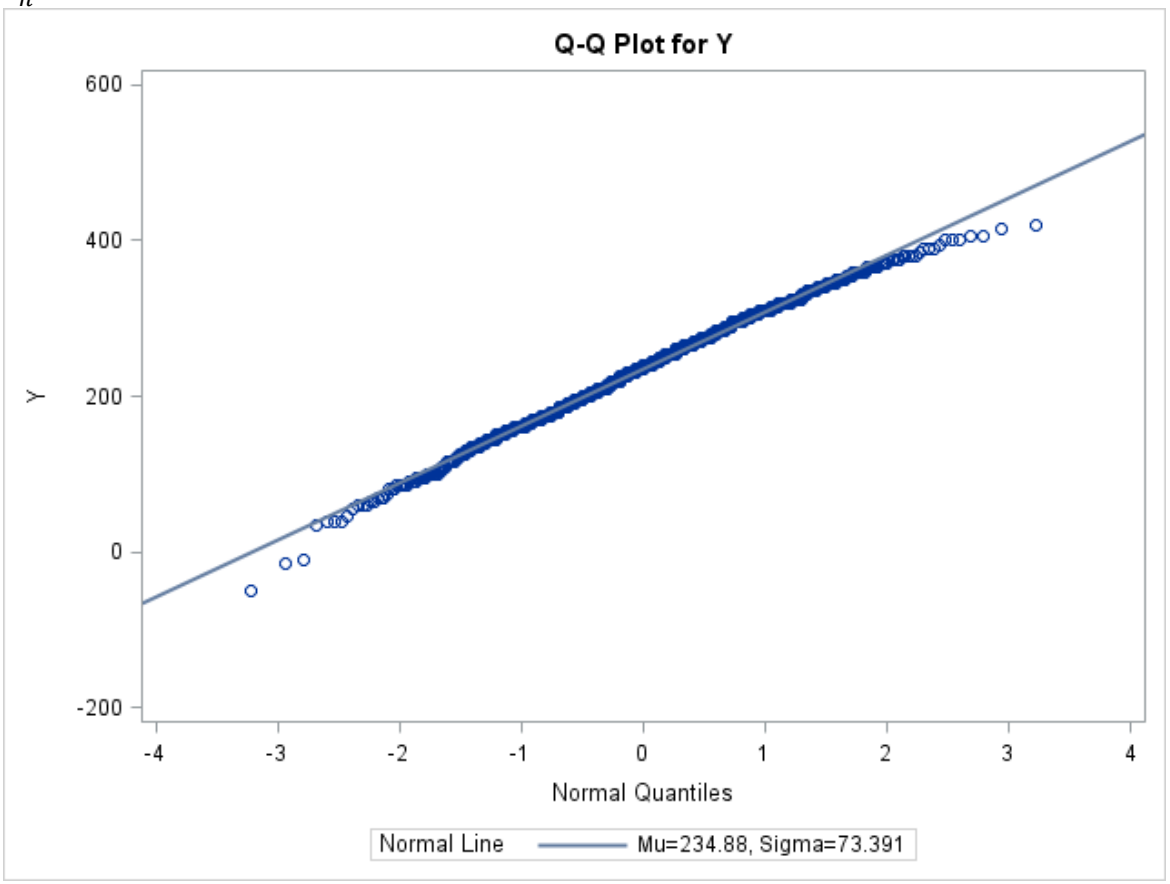


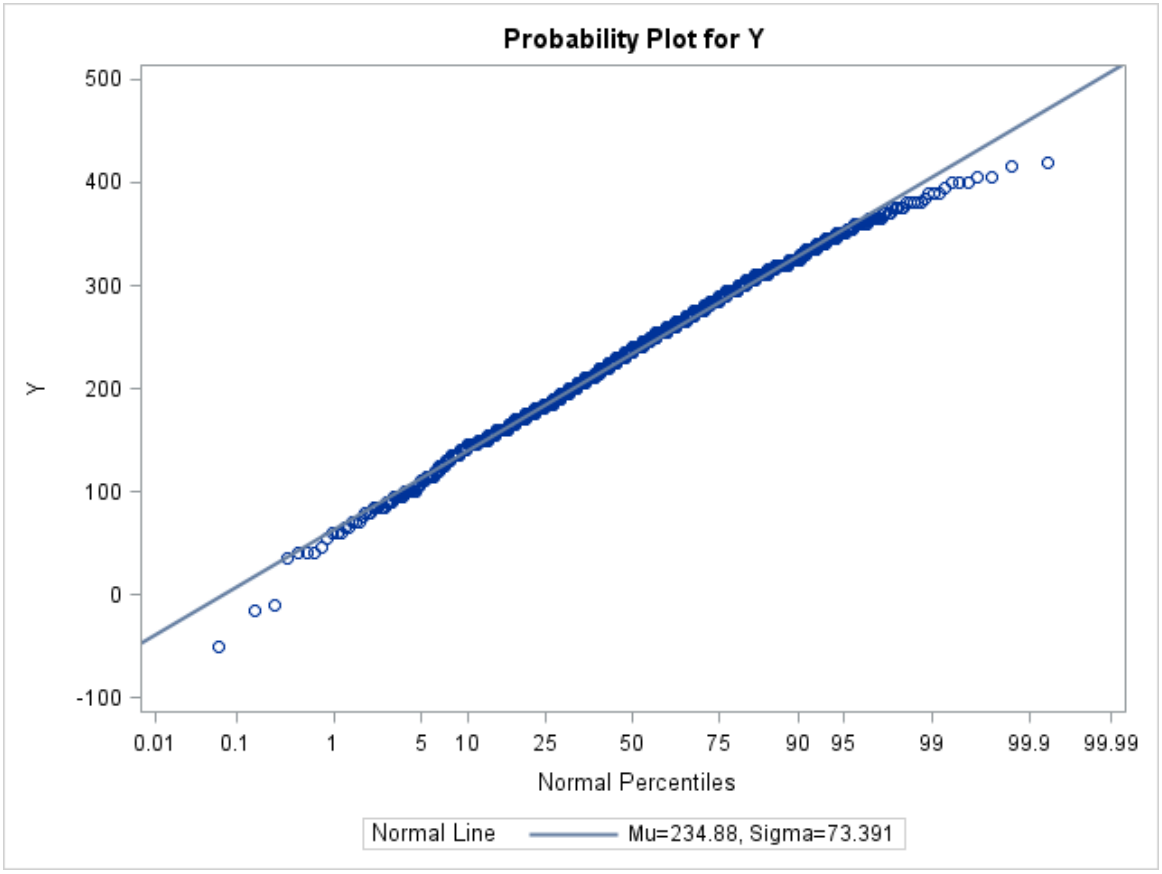
Ce qui est plus lisible dans ce qui suit avec la courbe,

```
PROC UNIVARIATE DATA=Simulation NORMALTEST;
Title "Histogramme test de normalité des sommes Y simulées";
VAR Y ;
HISTOGRAM Y / NORMAL ;
QQPLOT Y / NORMAL(MU=est SIGMA=est COLOR=red L=1);
RUN;
```



Dans ce qui suit, on peut voir que les valeurs suivent la ligne de normalité ce qui assure que $\overline{Y_n}$ suit une loi Normale.





Le problème qui se pose maintenant est celui des variations du gain moyen $\overline{Y_n}$ autour de $E\ (Y)$. On souhaiterait réduire au mieux cés variations (pour que notamment Y, ait très peu de chances d'être négatif !). Du point de vue statistique, le problème se formule alors de la façon suivante :

> Pour **d** fixé, $0 < \alpha < 1$ fixé et $r > 0$ fixé, combien faut-il réaliser de parties afin de pouvoir affirmer que $\overline{Y_n}$ a une probabilité égale à $(1 - \alpha)$ d'être dans l'intervalle $[\mu - r, \mu + r\ ]$ (cad. avec un coefficient de sécurité égal à α).

11 On suppose n suffisamment grand pour se placer sous hypothèse de normalité. Démontrer alors que l'intervalle u-r, u +r est associé à un coefficient de sécurité a dès lors que :

$$n \geq \frac{(53000 + 200d + 35d^2)\ t^2{}_{1-\alpha/2}}{1296r^2}$$

Avec t_{α} fractile de la loi $N\ (0,1)$ $(i.e.\ P[N\ (0,1) < t_\alpha) = \alpha)$.

Réponse :

$$P(\overline{Y_n} \in [\mu - r, \mu + r\ ]\ ) = 1 - \alpha$$

$$=> P(\mu - r \leq \overline{Y_n} \leq \mu + r\ ) = 1 - \alpha$$

$$=> P(-r \leq \overline{Y_n} - \mu \leq +r\ ) = 1 - \alpha$$

$$=> P\left(\frac{-r}{\sigma/\sqrt{n}} \leq \frac{\overline{Y_n} - \mu}{\sigma/\sqrt{n}} \leq \frac{+r}{\sigma/\sqrt{n}} \right) = 1 - \alpha$$

$$=> P\left(\frac{-r}{\sigma/\sqrt{n}} \leq \frac{\overline{Y_n} - \mu}{\sigma/\sqrt{n}} \leq \frac{+r}{\sigma/\sqrt{n}} \right) = 1 - \alpha$$

Or :

$$\frac{\overline{Y_n} - \mu}{\sigma/\sqrt{n}} = Z \sim N(0,1)$$

$$=> P\left(\frac{-r}{\sigma/\sqrt{n}} \leq Z \leq \frac{+r}{\sigma/\sqrt{n}} \right) = 1 - \alpha$$

$$=> P(\ t_1 \leq Z \leq t_2\ ) = 1 - \alpha$$

En général, on choisit t_1, t_2 tel que :

$$P(Z \leq t_1) = P(Z \leq t_2) = \frac{\alpha}{2}$$

Par symétrie, on va noter t_α tel que :

$$P(Z \leq t_\alpha) = \alpha$$

On va considérer : $\left\{ \begin{array}{c} t_{\frac{\alpha}{2}} = -x \\ t_{1-\frac{\alpha}{2}} = x \end{array} \right\} t_{\frac{\alpha}{2}} = -t_{1-\frac{\alpha}{2}}$

On a donc l'intervalle de confiance suivant :

$$A_\alpha = \left[\hat{\mu} - t_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} \; ; \; \hat{\mu} + t_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} \right]$$

Si on souhaite avoir un intervalle de confiance de demi-longueur à r fixée, alors :

$$r = t_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$$

$$\Rightarrow \sqrt{n} = t_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{r}$$

$$\Rightarrow n = {t_{1-\frac{\alpha}{2}}}^2 \cdot \frac{\sigma^2}{r^2}$$

Par suite :

$$n \geq \frac{(53000 + 200d + 35d^2)\, {t^2}_{1-\alpha/2}}{1296r^2}$$

12 Ecrire un programme SAS comportant en entrée une fenêtre interactive où l'utilisateur fixera le réel d, le coefficient a et le rayon r. Le programme calculera ensuite le plus petit entier n vérifiant la relation de la question 11. A partir de ceci, 100 simulations de taille n seront réalisées et le gain moyen Y déterminé pour chacune d'elles. En sortie, le programme affichera les résultats suivants :

-la valeur de n,

- une représentation graphique des 100 valeurs Y, calculées avec des limites tracées aux valeurs u et $u \pm r$,

- un camembert permettant de visualiser le pourcentage des 100 valeurs qui sont effectivement dans l'intervalle prévu, et celles qui n'y sont pas.

On estime qu'un gain moyen de 0.50 euros par partie est convenable (déterminer la valeur de d correspondante). On souhaiterait que Y, reste dans l'intervalle (0.25, 0.75).

Réponse :

Pour calculer n, on a le code SaS suivant :

```
%window fen
#8 'la valeur du parametre d ' dd 9 attr=underline
#9 'la valeur du parametre alfa : ' a 9 attr=underline
#10 'la valeur du rayon r : ' r 9 attr=underline
#11 'appuyer sur entree...';
%macro inter;
      %display fen;
%mend inter;
Data Simulation1;
      %inter;
      d=&dd;
      alfa=&a;
      ray=&r;
      t=(probit(1-alfa/2));
      n=((53000+200*d+35*d**2)*(t)**2)/(1296*(ray**2));
      res=100;
      moy=0;
      Do i=1 to res;
            Do j=1 to n;
            E1=int(6*ranuni(0)+1);
            E2=int(6*ranuni(0)+1);
            x=min(E1,E2);
            select(x);
                  when(1) Y=10;
                  when(2) Y=5;
                  when(3) Y=0;
                  when(4) Y=-5;
                  when(5) Y=-10;
                  when(6) Y=-d;
                  end;
            moy=moy+Y;
            end;
      moy=moy/n;
      output;
      put moy=;
      end;
put n=;
run;
```

Pour $E(Y) = 0{,}5$, le gain moyen estimer sera de 0,5€, alors :

$$=> \frac{100-d}{36} = 0{,}5 \;=> d = 82$$

Pour la représentation graphique des 100 valeurs Y,

```
data anno;
   length function color $8;
   retain xsys ysys '2' when 'a';
   set Simulation1 ;

   function='symbol';
   x=i;
   y=moy;
   size=1.3;
   text='dot';

   if moy > 0.25 and moy < 0.75 then color='viyg';
   else color='red';
   output;
run;

Symbol1 i=join ci=stolg value=dot cv=black h=0.2cm;
proc gplot data=simulation1;
      title "Une représentation graphique des 100 valeurs Y, calculées avec
des limites tracées aux valeurs u et u‡r";
      plot moy*i/ vref=0.75 0.5 0.25 cvref=red lvref=20 frame annotate=anno;
run;
quit;
```

Pour visualiser un camembert du pourcentage des valeurs dans l'intervalle donné des 100 valeurs,

```
legend1 label=none
        position=(left middle)
        offset=(1,)
        across=1
        order=( 'Dans de l'intervalle')
        value=(color=black)
        shape=bar(4,1.5);

proc gchart data=simulation1;
      title "Un camembert de pourcentage des 100 valeurs dans ou hors
       intervalle";
      pie moy/ midpoints=0 0.5 1 legend=legend1 other=25 otherlabel="Hors de
         l'intervalle" value=inside fill=solid;
run;
```

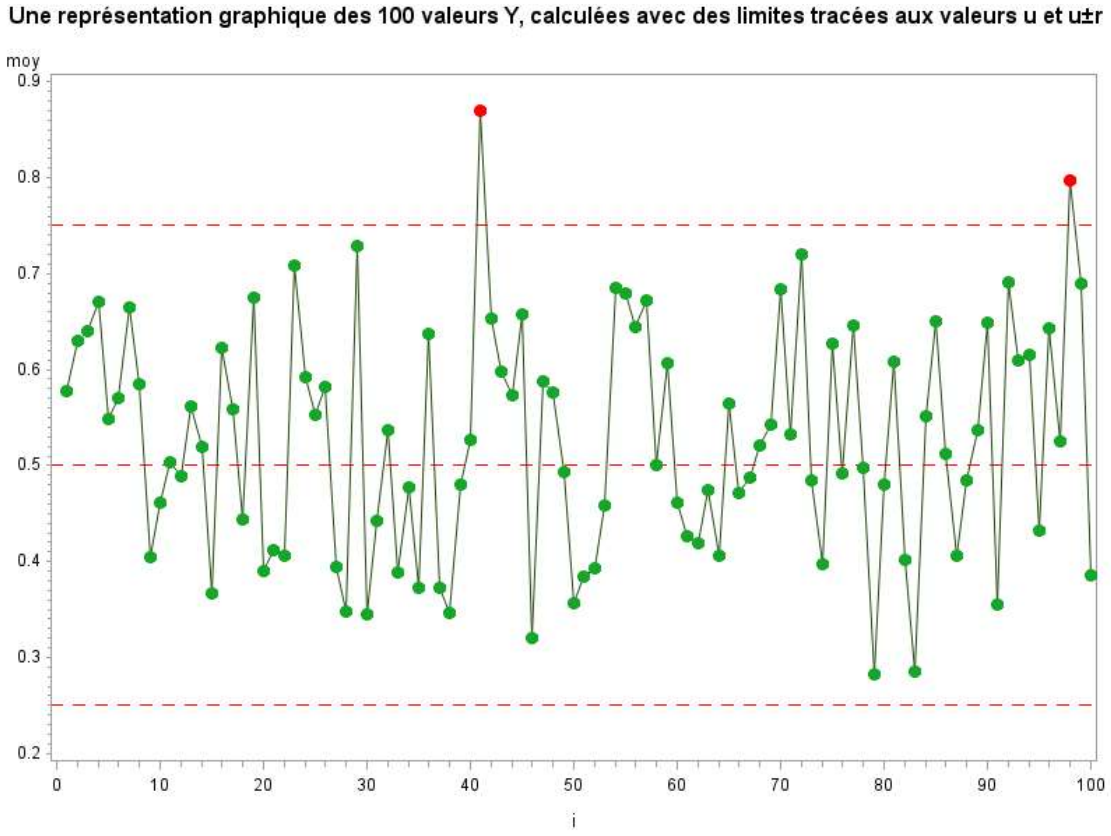
On aura une fenêtre interactive comme suit :

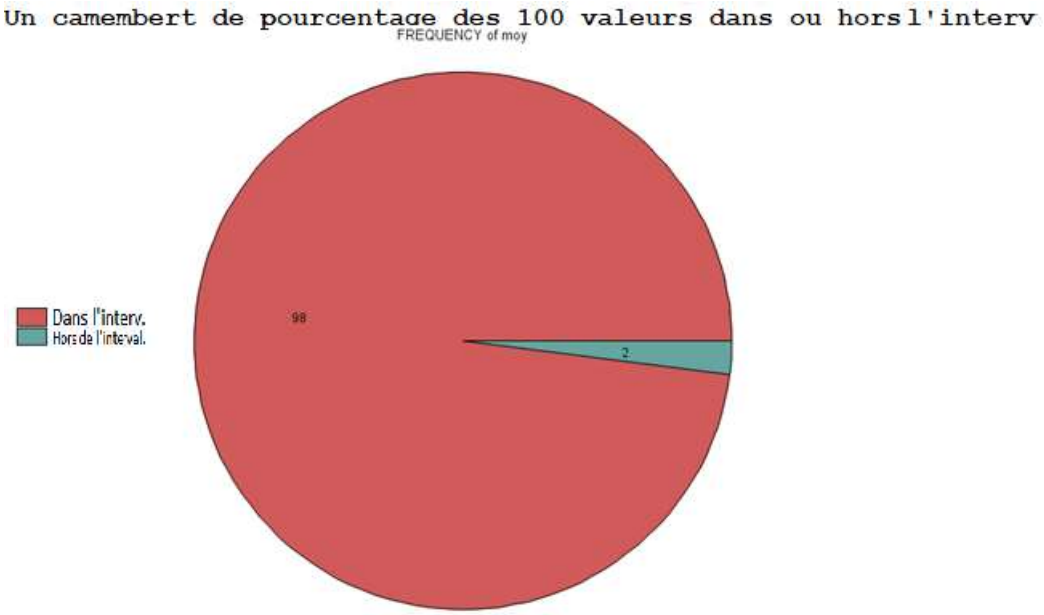
```
la valeur du parametre d:__
la valeur du parametre alfa : ____
la valeur du rayon r : ____
appuyer sur entree...
```

13 Lancer 3 fois le programme aux niveaux des résultats obtenus $\alpha = 20\%$ et $\alpha = 5\%$. Commenter les résultats obtenus.

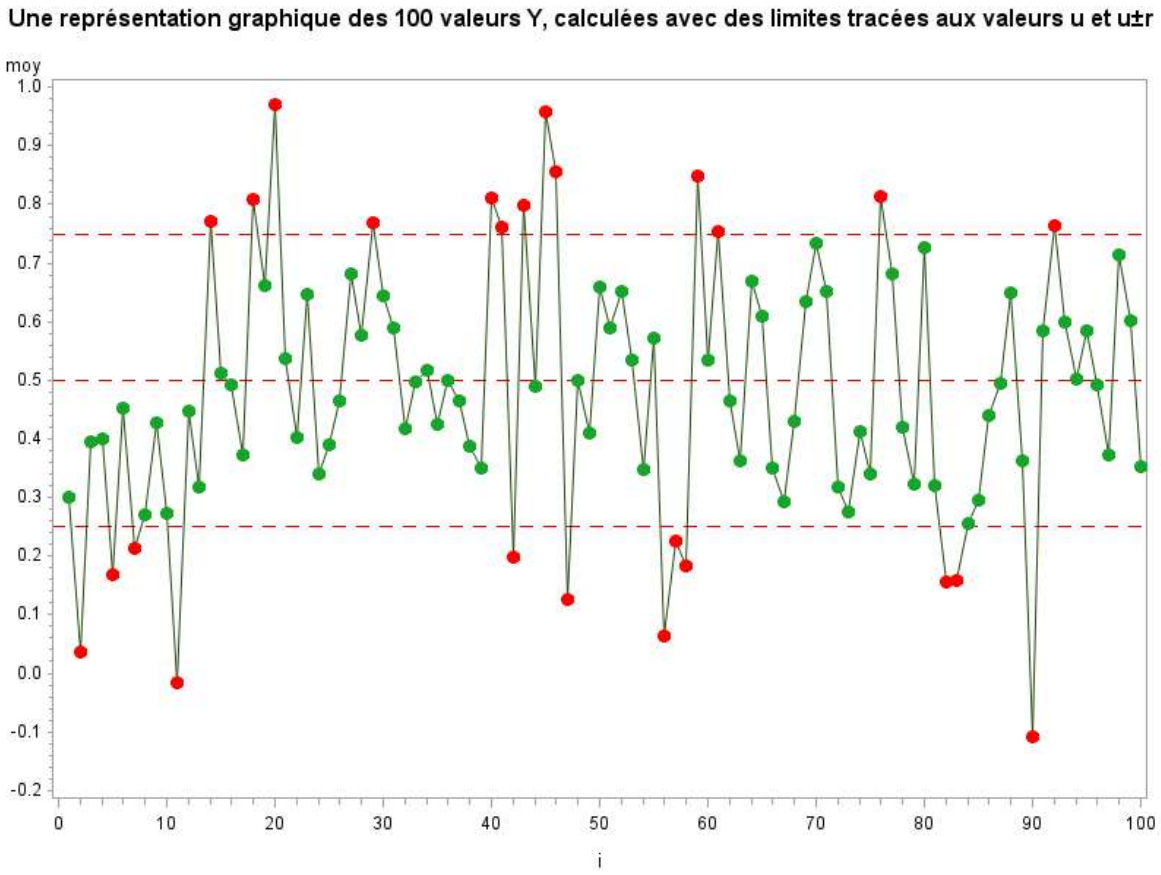
Réponse :

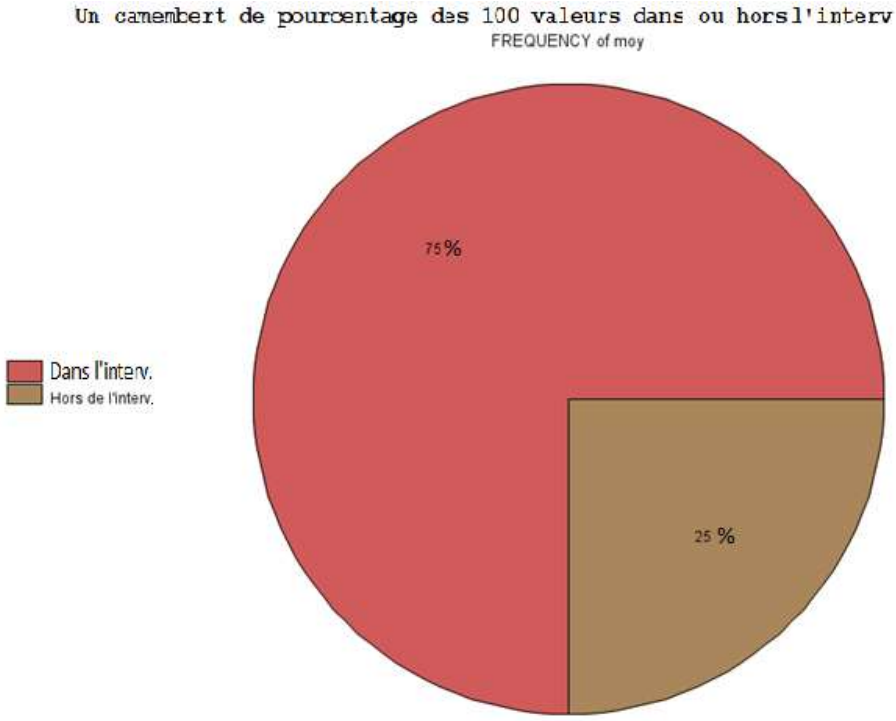
Pour $0.5€,\ \alpha = 5\%,\ n = 14452.421741,\ r = 0.25$ *:*

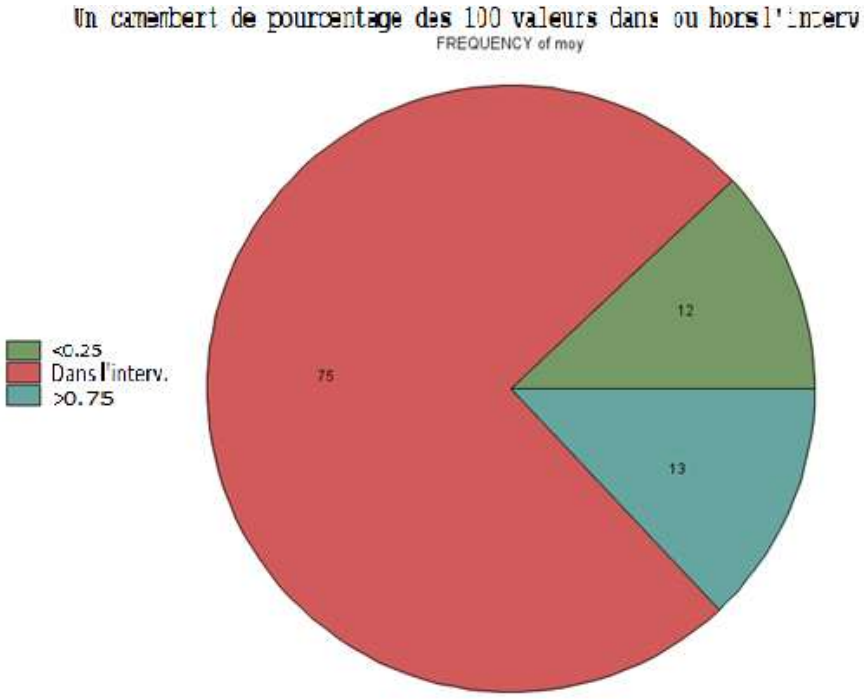




Pour 0.5€, α=20%, n=6178.9775219, r=0.25 :






Un camembert de pourcentage des 100 valeurs dans ou hors l'interv
FREQUENCY of moy
<0.25
Dans l'interv.
>0.75
12
75
13