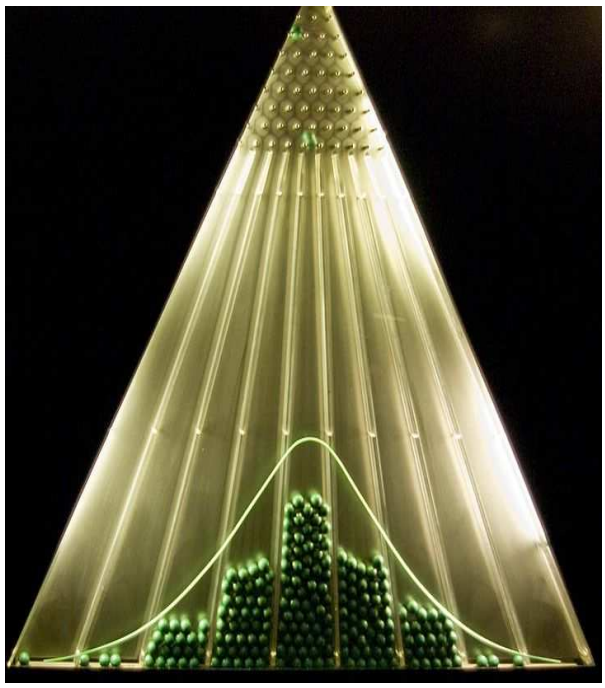
| | |
|---|---|
| EISTI | **GSI, MI ET SIE : EXAMEN DE STATISTIQUE INFERENTIELLE** |
| Rédigé par : Equipe pédagogique de Stat. Inférentielle | Ref : *EX-STAT-INFE* |
| A l'intention de : Etudiants des parcours GSI, MI et SIE | Créé le : 28/11/2012 |

# *Questions de réflexion*

## *Exercice 1. La machine de Galton*



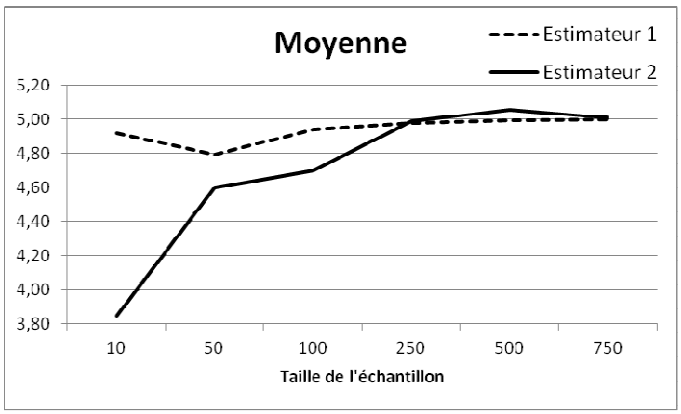
Cette machine est constituée d'étages de clous placés en quinconce (haut de la figure). A partir du haut de la machine, on fait tomber des billes qui heurtent les clous. A chaque contact d'une bille sur un clou, la bille tombe dans l'étage du dessous avec une probabilité de ½ d'aller à gauche du clou et une probabilité de ½ d'aller à droite du clou. Il en est ainsi pour chaque étage. La bille finit par tomber dans une des colonnes (bas de la figure). On suppose que le nombre d'étages est assez grand.

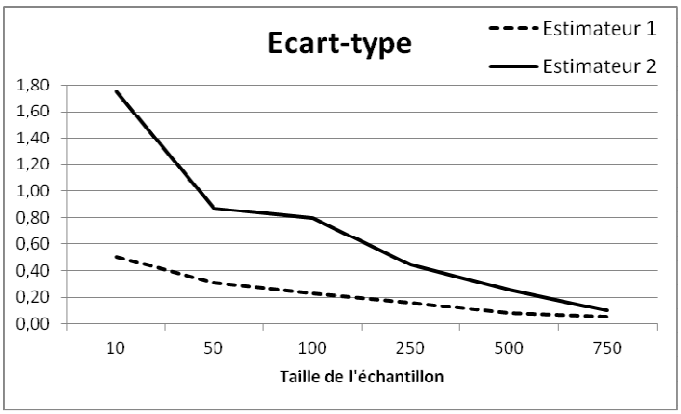
On note $n$ le nombre d'étages et $X_i$ la variable qui décrit ce qui se passe au $i$-ème étage .

1) Quel est le lien entre les variables $X_i$ et la colonne dans laquelle tombe une bille ?
2) Expliquer pourquoi après avoir fait tomber un très grand nombre de billes, la forme de l'empilement des billes a tendance à s'aligner sur la courbe blanche de la figure.

## *Exercice 2. Comparaison d'estimateurs*

Soit $T_1$ et $T_2$ deux estimateurs d'une même caractéristique θ. Les graphiques ci-dessus représentent les valeurs obtenues par ces deux estimateurs pour des échantillons de taille n=10 à n=750. Pour chaque taille n, 50 échantillons ont été utilisés pour calculer les valeurs de $T_1$ et $T_2$. Le graphique MOYENNE représente le moyenne des 50 valeurs de $T_1$ et $T_2$ en fonction de la taille n. Le graphique ECART-TYPE représente l'écart-type des 50 valeurs de $T_1$ et $T_2$ en fonction de la taille n.




1) Que pouvez vous dire concernant le biais de $T_1$ et le biais de $T_2$ ? A votre avis, quelle est la valeur de θ ?
2) Que pouvez-vous dire sur le risque quadratique ? A votre avis, il est préférable d'utiliser quel estimateur ?

### *Exercice 3 . Région critique, risques α et β*

Soit un échantillon $X_1,\ldots,X_n$ tel que $E(X)=\mu$ et $var(X)=\sigma^2$. On souhaite effectuer un test sur la moyenne avec les hypothèses

$$\begin{cases} H_0 : \mu=\mu_0 \\ H_1 : \mu=\mu_1<\mu_0 \end{cases}$$

La moyenne $\overline{X}$ est un estimateur naturel de μ. Pour un échantillon assez grand, on approche la loi de l'estimateur grâce au T.C.L. par une loi normale $N(\mu,\sigma^2/n)$. Pour ce test, on obtient le graphique de l'annexe 2. A partir de ce graphique, vous pouvez répondre aux questions suivantes :

1) Quelle est la forme de la région critique ?
2) Sur le graphique, hachurer proprement et de façon distincte, le risque de première espèce α et le risque de deuxième espèce β.
3) On note $\overline{x}$ la valeur de $\overline{X}$ calculée sur un échantillon. Quelle décision prend-on ? Avec quel risque ?
4) Sur le graphique, hachurer la p-valeur du test.

**N.B. Ne pas oublier de rendre l'annexe2 avec vos nom et prénom**

# *Application sur un jeu de données (annexe 1)*

### *Présentation*

On étudie le pourcentage de nappes phréatiques présentant une concentration de nitrate inquiétante, c'est-à-dire supérieure à *10mg/l*. Pour cela on dispose des résultats sur 64 départements (annexe 1).

- Dans le tableau 1 se trouve un aperçu du jeu de données avec :
  Colonne 1 : le numéro du département
  Colonne2 : le nom du département
  Colonne 3 : la région dans laquelle est le département
  Colonne 4 : la localisation (NORD/SUD) du département
  Colonne 5 : le pourcentage X de nappes phréatiques contaminées
  Colonne 6 : une variable notée Y prenant la valeur 1 si ce taux est supérieur à 50% et 0 sinon.
- Dans le tableau 2 se trouve un résumé numérique (moyenne, variance et écart-type) des variables X et Y.

### *Modélisation du problème*

Soit *X* la variable aléatoire représentant le pourcentage de nappes phréatiques présentant une concentration de nitrate inquiétante. La variable aléatoire *X* suit une loi continue définie par la fonction de densité suivante :

$$f_\theta(x)=\begin{cases} 2\theta x-\theta+1 & \text{si } x\in[0,1] \\ 0 & \text{sinon} \end{cases}$$

où θ est un paramètre inconnu appartenant à [-1,1]. On sait que

$$\mu = E[X]=\frac{\theta}{6}+\frac{1}{2} \text{ et } V(X)=\frac{1}{12}-\frac{\theta^2}{36}.$$

## *Exercice 4. Première partie : Estimateur*

Soit un échantillon $X_1,\ldots,X_n$ de variables indépendantes issues de X. On se propose d'estimer le paramètre θ à l'aide de l'estimateur

$$T = a\overline{X} + b,$$

où a et b sont des paramètres à fixer.

1) Déterminer a et b tels que T soit un estimateur sans biais de θ
2) Etablir la convergence en probabilité de T.
3) Calculer le risque quadratique de T.
4) Montrer que T converge en loi vers une loi normale,

$$N(\theta, \frac{\sigma^2}{n}) \text{ où } \sigma^2 = 3 - \theta^2$$

## *Exercice 5. Deuxième partie : Intervalle de confiance*

L'échantillon est de taille 64 et on observe une valeur moyenne

$$\overline{x} = \frac{5}{12} \approx 0{,}42$$

On déduit de la première partie une estimation de θ et une estimation de $\sigma^2$,

$$\hat{\theta} \approx -0{,}5 \quad \text{et} \quad s^2 \approx (0{,}26)^2$$

1) Déterminer un intervalle de confiance pour θ avec α=0,01.
2) En déduire un intervalle de confiance pour le pourcentage moyen μ avec α=0,01.

## *Exercice 6. Troisième partie : Test*

On souhaite tester s'il y a une différence significative de pollution au nitrate entre les 32 départements du nord et les 32 départements du sud. On suppose que les deux échantillons sont indépendants.

1) Quel test pour vérifier cette hypothèse à l'aide des valeurs de X? On justifie la réponse
2) Quel test pour vérifier cette hypothèse à l'aide des valeurs de Y? On justifie la réponse
3) Faites le test avec la variable Y et α=0,05. Conclure.

## *Exercice 7. Test d'ajustement*

Dans les parties précédentes, on a fait l'hypothèse que la pourcentage X avait pour fonction de densité $f_\theta$.

1) Quel test pour vérifier cette hypothèse à l'aide des valeurs de X? On justifie la réponse
2) Faites le test avec α=0,05. Conclure.

**N.B. Les résultats numériques indispensables pour le test sont dans le tableau donné en annexe.**

# ANNEXE 1

**Tableau 1 : Aperçu du fichier de données**

| Id | N° | Département | Région | Localisation | X | Y: X>50% | $F(x_i)$ | $\lvert F(x_i)-i/n \rvert$ | $\lvert F(x_i)-(i-1)/n \rvert$ |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 2 | Haut-Rhin | Alsace | NORD | 0,01 | 0 | 0,010 | 0,006 | 0,010 |
| 2 | 8 | Lot-et-Garonne | Aquitaine | SUD | 0,03 | 0 | 0,038 | 0,007 | 0,022 |
| 3 | 10 | Doubs | Franche-Comté | NORD | 0,07 | 0 | 0,098 | 0,051 | 0,067 |
| 4 | 14 | Gard | Languedoc-Roussillon | SUD | 0,08 | 0 | 0,116 | 0,054 | 0,069 |
| 5 | 21 | Marne | Champagne-Ardenne | NORD | 0,08 | 0 | 0,120 | 0,042 | 0,058 |
| 6 | 22 | Gers | Midi-Pyrénées | SUD | 0,10 | 0 | 0,142 | 0,048 | 0,063 |
| 7 | 25 | Dordogne | Aquitaine | SUD | 0,10 | 0 | 0,148 | 0,038 | 0,054 |
| | | … | … | … | … | … | … | … | … |
| 59 | 82 | Jura | Franche-Comté | NORD | 0,76 | 1 | 0,852 | 0,070 | 0,054 |
| 60 | 83 | Aude | Languedoc-Roussillon | SUD | 0,77 | 1 | 0,859 | 0,078 | 0,063 |
| 61 | 84 | Terr.-de-Belfort | Franche-Comté | NORD | 0,89 | 1 | 0,938 | 0,015 | 0,000 |
| 62 | 85 | Rhône | Rhône-Alpes | SUD | 0,89 | 1 | 0,941 | 0,028 | 0,012 |
| 63 | 86 | Aveyron | Midi-Pyrénées | SUD | 0,98 | 1 | 0,988 | 0,004 | 0,020 |
| 64 | 82 | Pyrénées-Atl. | Aquitaine | SUD | 0,99 | 1 | 0,993 | 0,007 | 0,009 |
| | | | | | | | **Max** | **0,087** | **0,071** |

**Tableau 2 : Résumé numérique du jeu de données**

| | NORD | SUD | TOTAL |
|---|---|---|---|
| **Effectifs** | 32 | 32 | 64 |
| **Moyenne** | 0,20 | 0,63 | 0,42 |
| **Variance** | 0,01 | 0,02 | 0,07 |
| **Ecart-type** | 0,11 | 0,16 | 0,26 |

# ANNEXE 2

NOM : ……

PRENOM : ……