ÉCHANTILLONNAGE

I - Échantillons

1°) - Échantillons d'individus dans une population

Quand on doit décrire une population finie comportant un grand nombre d'individus, on ne peut pas ou on ne veut pas, en général pour des raisons économiques, en faire une étude exhaustive. Les observations ne portent alors que sur un nombre restreint d'individus à sélectionner selon un protocole expérimental. Les individus sélectionnés constituent un échantillon, leur nombre est la taille de l'échantillon.

Parmi les divers types de constitution d'échantillons aléatoires issus de populations, peuvent être envisagés :

- Les tirages constitués avec remise supposent que tout individu choisi est replacé dans la population, après observation du caractère. Il peut éventuellement être sélectionné plusieurs fois.
- Les tirages constitués sans remise supposent que tout individu choisi ne peut jamais être repris une deuxième fois.

Lorsque l'effectif de la population est très grand et que le taux de sondage est faible (en général inférieur à 0,1), les différences entre les résultats de calculs de probabilités obtenus pour les deux types d'échantillons sont négligeables.

2°) - Échantillons d'observations

Les observations sur les individus d'un échantillon constituent un échantillon d'observations.

Exemple : Mesures sur un échantillon de plantes

Un même échantillon "expérimental" pouvant éventuellement être obtenu à partir de plusieurs échantillons d'individus, les distributions de probabilité sur les deux ensembles d'échantillons pourront être différentes.

On obtient aussi un échantillon d'observations lorsqu'on travaille sur des résultats d'expériences aléatoires**.**

Exemple : Échantillons de rendements de céréales sur des parcelles de terrain.

III - Échantillonnage

1°) - Échantillonnage

La théorie de l'échantillonnage consiste en l'étude des distributions de probabilités de variables aléatoires définies sur l'ensemble des échantillons (fréquence d'échantillonnage, moyenne d'échantillonnage, variance d'échantillonnage…)..

La théorie de l'échantillonnage est utilisée en particulier dans deux situations :

- pour estimer un paramètre d'une population (pourcentage, moyenne, écart-type…) à partir de la valeur correspondante observée sur un échantillon (théorie de l'estimation) ;
- pour décider (avec un risque d'erreur fixé à l'avance) si, par exemple, la différence entre une valeur de référence et une moyenne observée sur un échantillon est due au hasard ou si elle est significative (théorie des tests).

2°) - Comment prélever un échantillon ?

Lors d'une prise de décision à partir d'un échantillon, pour que les résultats de la théorie des probabilités et de l'échantillonnage s'appliquent, il est important que l'échantillon soit aléatoire, c'est-à-dire prélevé selon une procédure préétablie.

Pour cela, on peut entre autres moyens, utiliser un générateur de nombres "aléatoires" après avoir affecté un numéro identifiant à chacun des individus de la population. Ceci est la version contemporaine des tables de chiffres au hasard utilisées auparavant.

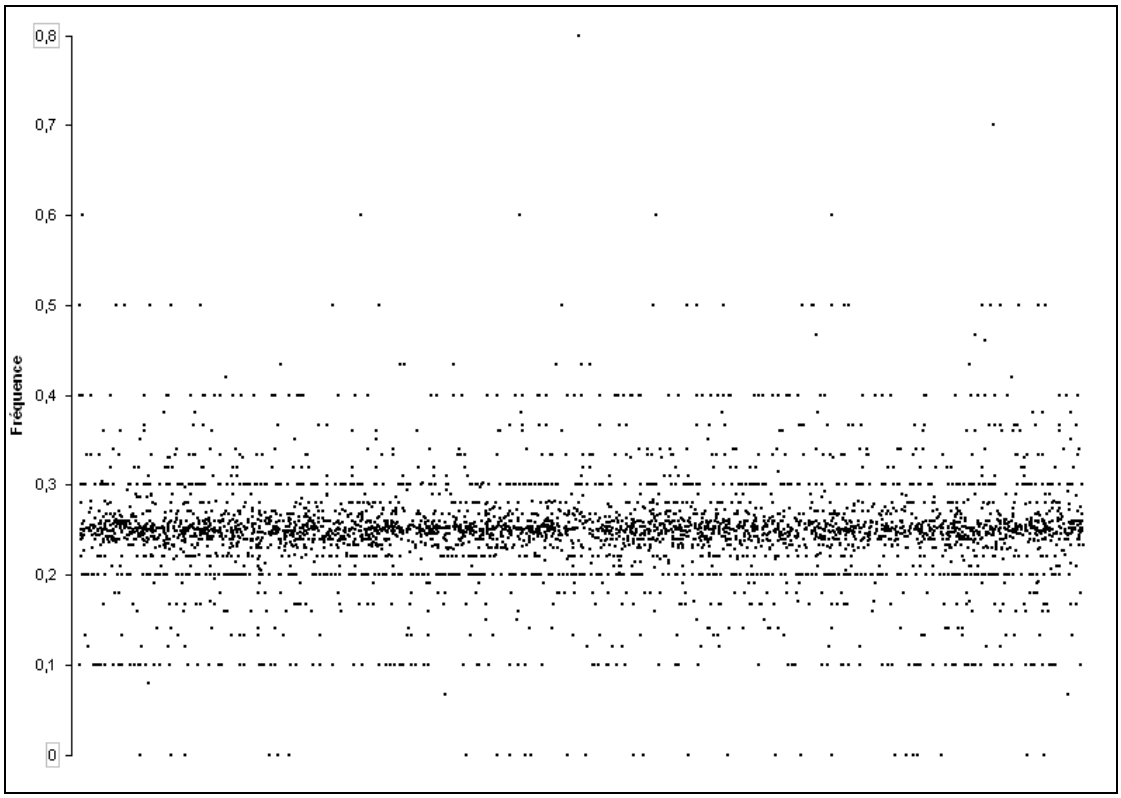
III - Distribution de probabilité de la fréquence d'échantillonnage

Dans une population $\mathcal{P}$, on considère une sous-population $\mathcal{A}$ contenant 25 % des individus de P.

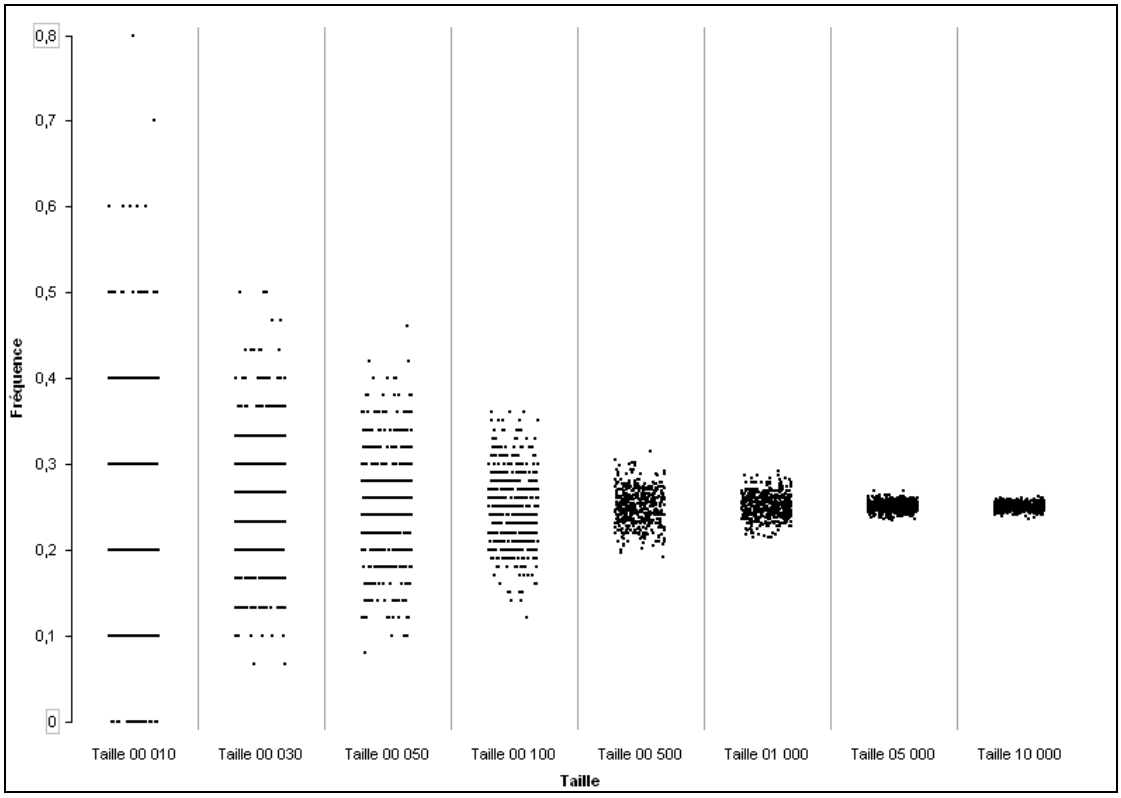
On extrait 4 000 échantillons de P numérotés de 1 à 4 000 :

- 500 échantillons de taille 10 ;
- puis 500 échantillons de taille 30 ;
- puis 500 échantillons de taille 50 ;
- puis 500 échantillons de taille 100 ;
- puis 500 échantillons de taille 500 ;
- puis 500 échantillons de taille 1 000 ;
- puis 500 échantillons de taille 5 000 ;
- puis 500 échantillons de taille 10 000.

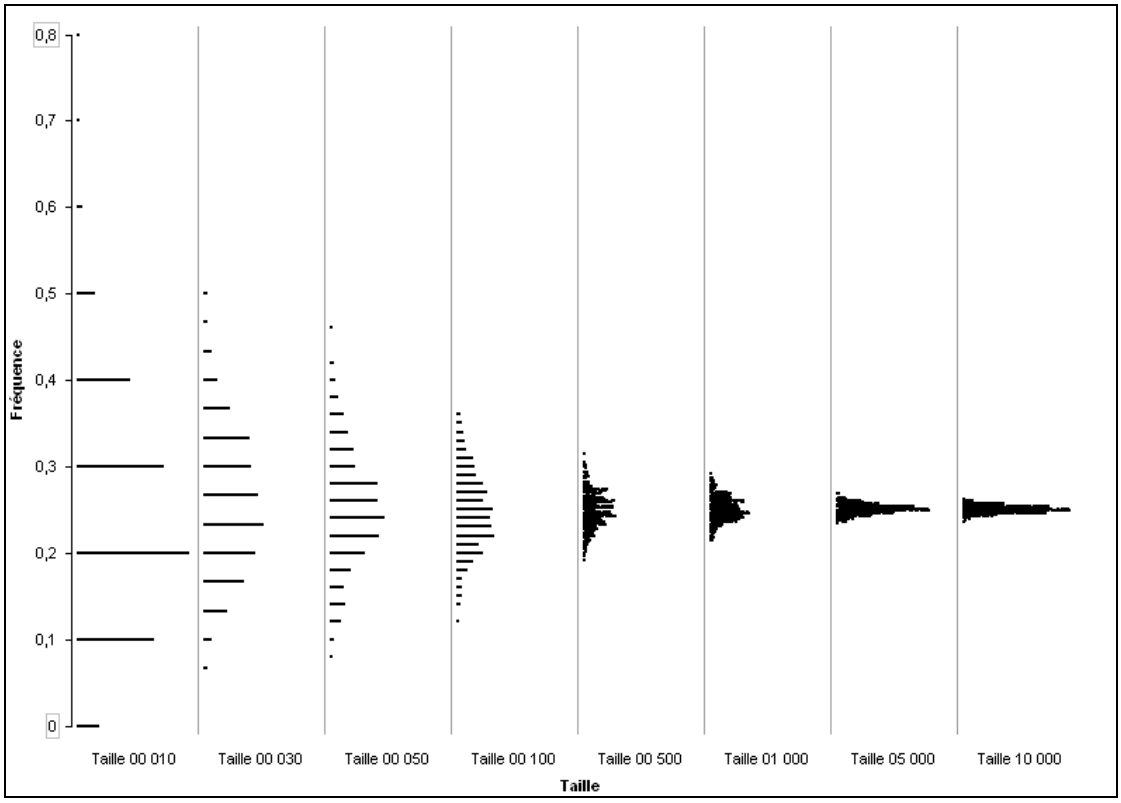
Représentation "en vrac" des fréquences de $\mathcal{A}$ dans les 4 000 échantillons :



Observons la fluctuation de la fréquence d'échantillonnage de $\mathcal{A}$ en triant les échantillons selon leurs tailles.



Distribution des fréquences d'échantillonnage des 4 000 échantillons.



Pour des échantillons constitués avec remise de taille *n*, la répartition du nombres d'individus de $\mathcal{A}$ se fait selon la loi binomiale de paramètre *n* et *p*. La nombre moyen d'éléments de $\mathcal{A}$ dans les échantillons est donc

np, ainsi, la moyenne des fréquences de $\mathcal{A}$ dans tous les échantillons de taille *n* est $\frac{np}{n} = p = 0{,}25$: la fréquence de $\mathcal{A}$ dans un échantillon estime sans biais la proportion *p* de la population.

La variance des fréquences des éléments de $\mathcal{A}$ dans tous les échantillons de taille *n* est $\frac{p\,(1 - p)}{n}$.

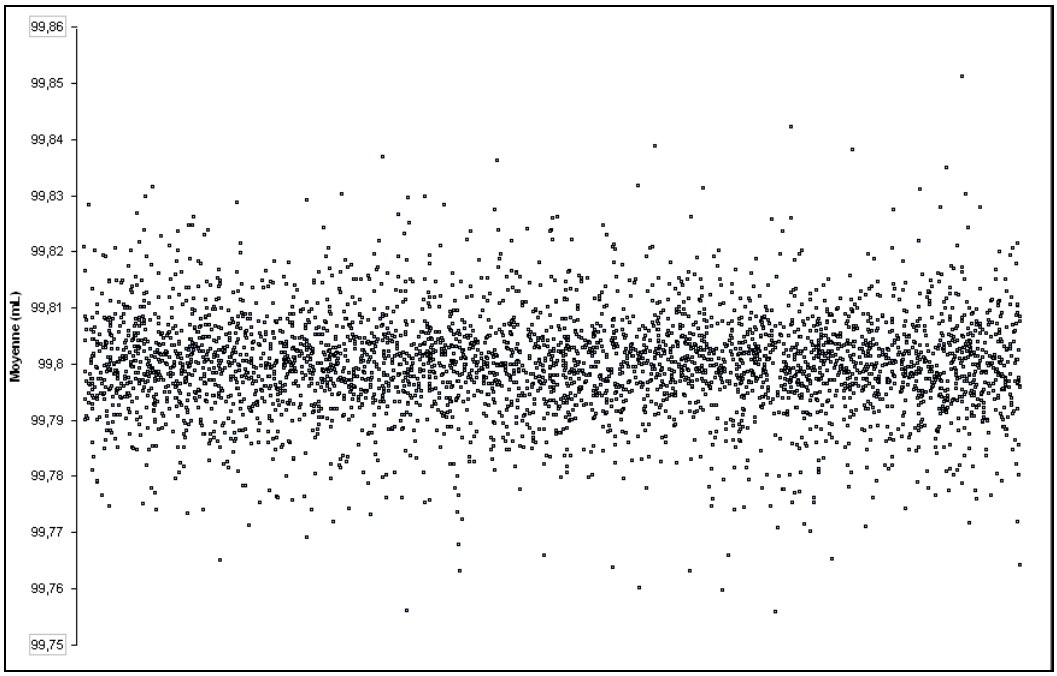
IV - Distribution de probabilité de la moyenne d'échantillonnage

On extrait 5 000 échantillons, numérotés de 1 à 5 000, d'une la population où l'on s'intéresse à un caractère quantitatif distribué selon la loi normale $\mathcal{N}(100 ; 0{,}3)$.

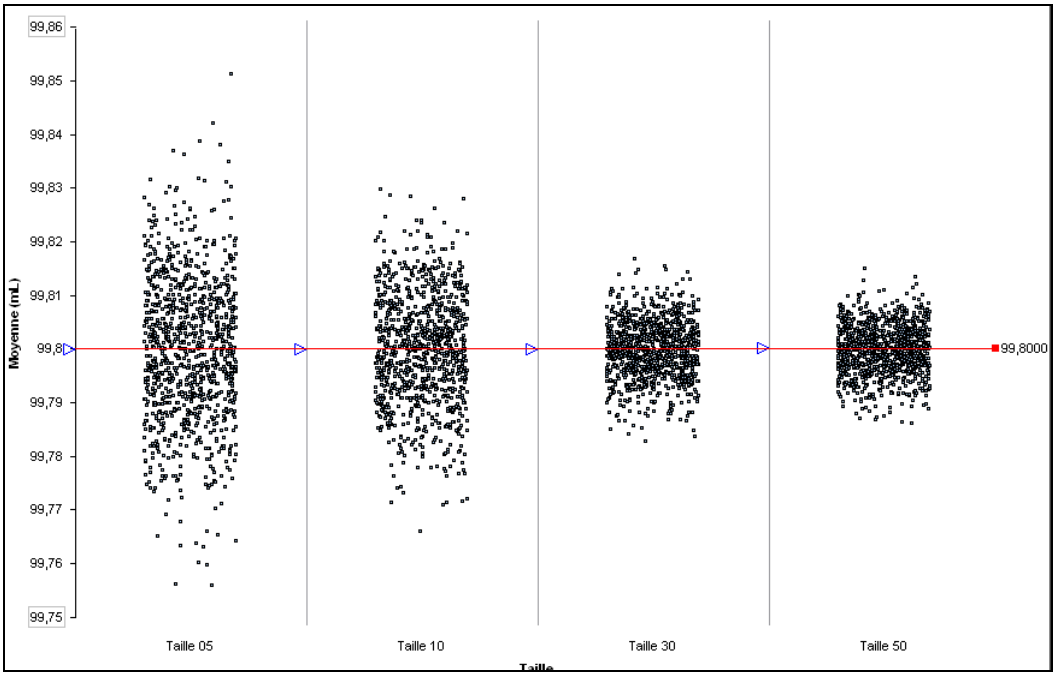
- 1 000 échantillons de taille 30
- 1 000 échantillons de taille 50
- 1 000 échantillons de taille 100
- 1 000 échantillons de taille 500
- 1 000 échantillons de taille 1 000

Pour chaque échantillon, on a calculé la moyenne, la variance, la variance corrigée (ou expérimentale), l'écart-type, l'écart-type corrigé (ou expérimental).

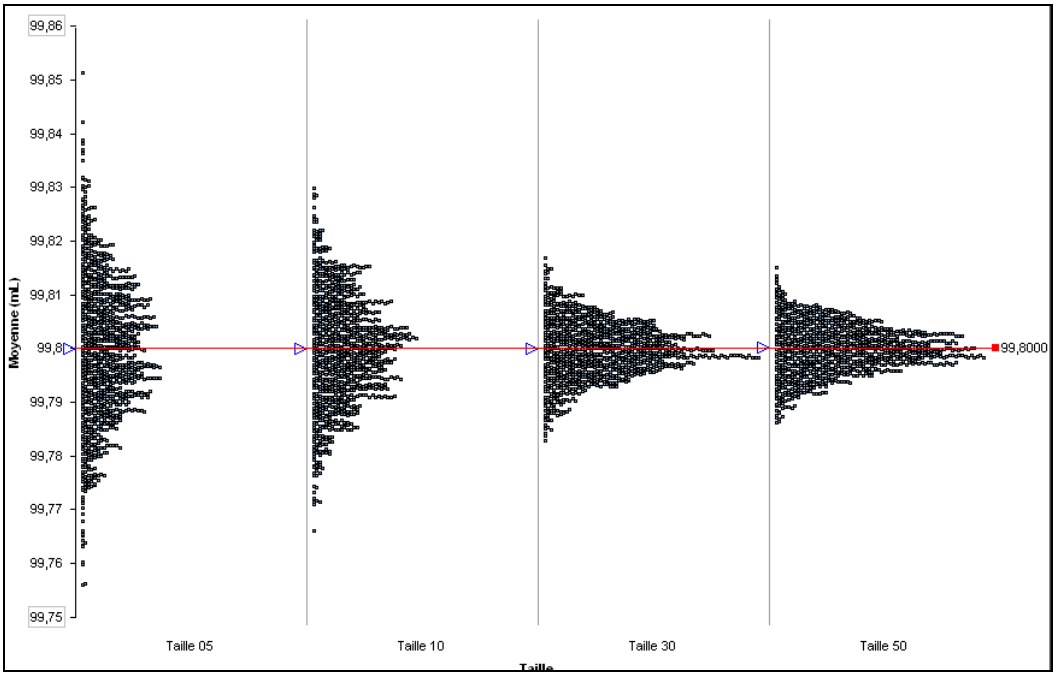
Représentation "en vrac" des moyennes des 5 000 échantillons :



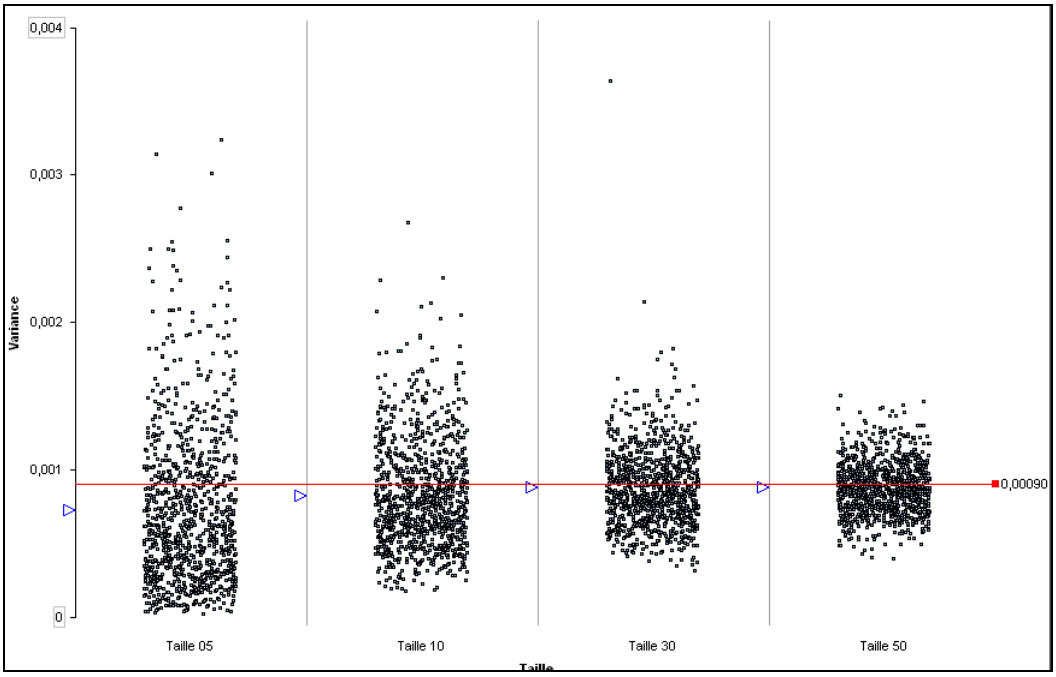
Représentation des moyennes des 5 000 échantillons regroupées par taille d'échantillon :



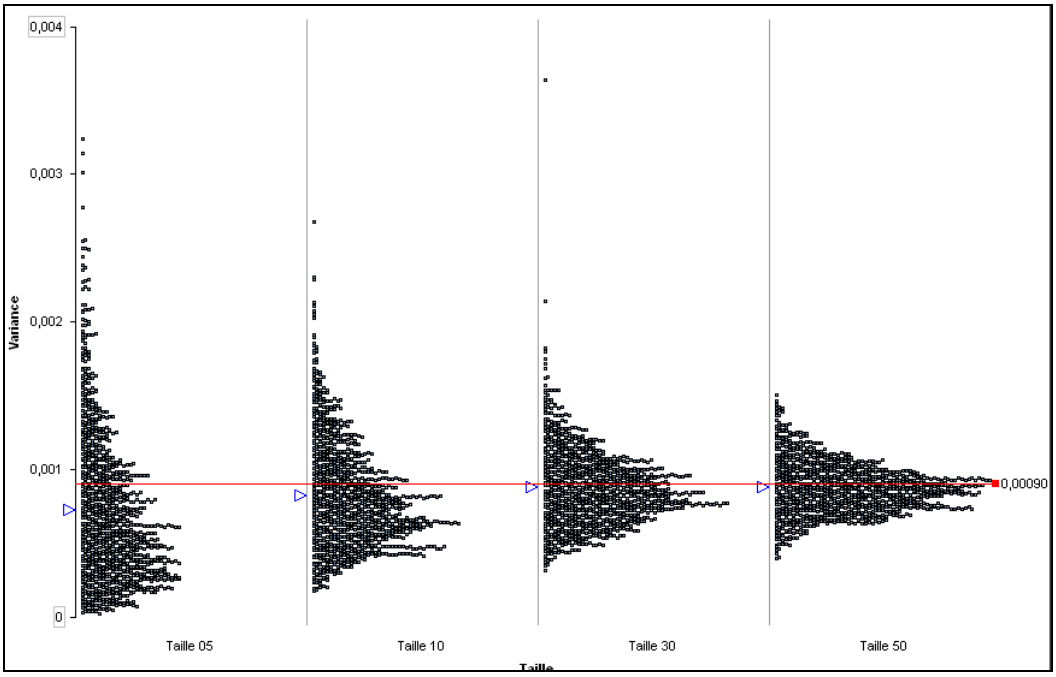
Distribution des moyennes des 5 000 échantillons regroupées par taille d'échantillon :



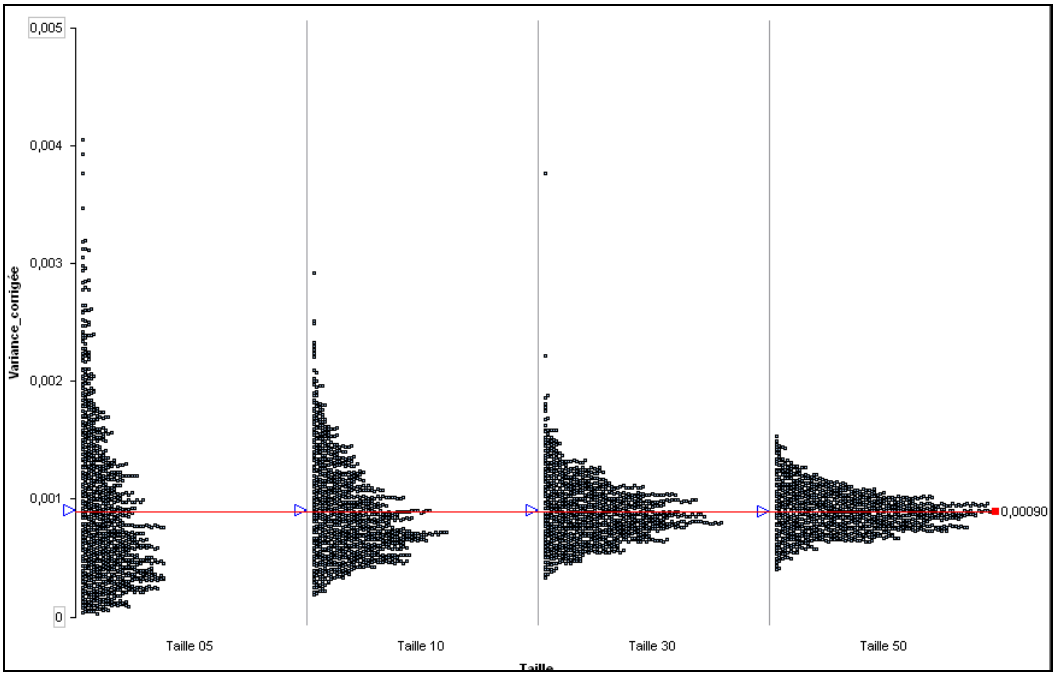
Représentation des variances des 5 000 échantillons regroupées par taille d'échantillon :



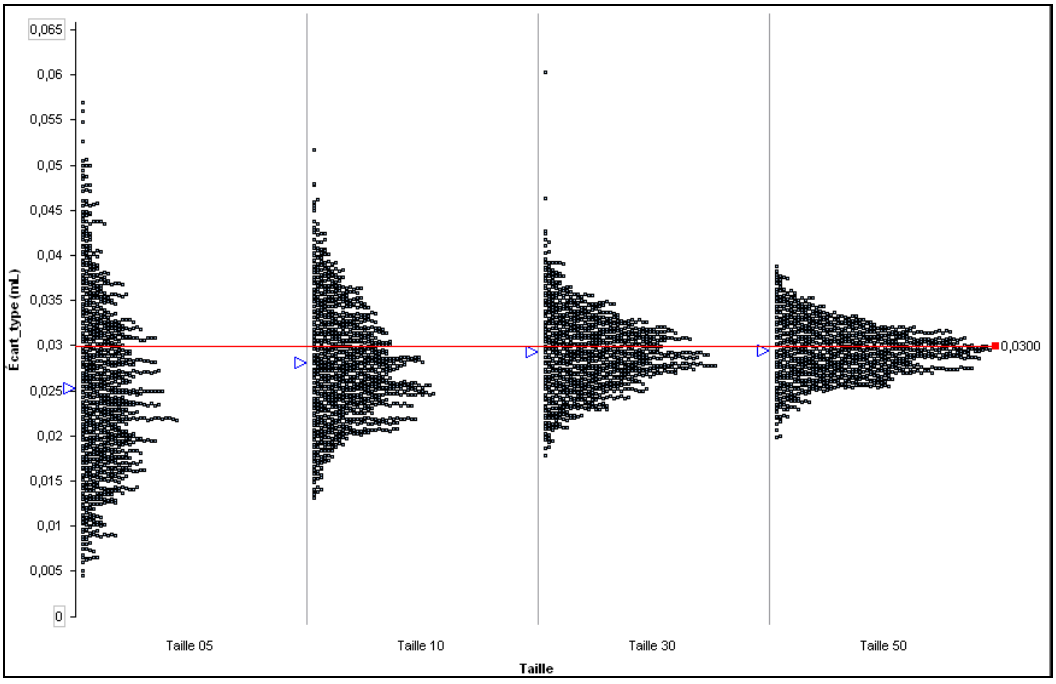
Distribution des variances des 4 000 échantillons regroupées par taille d'échantillon :



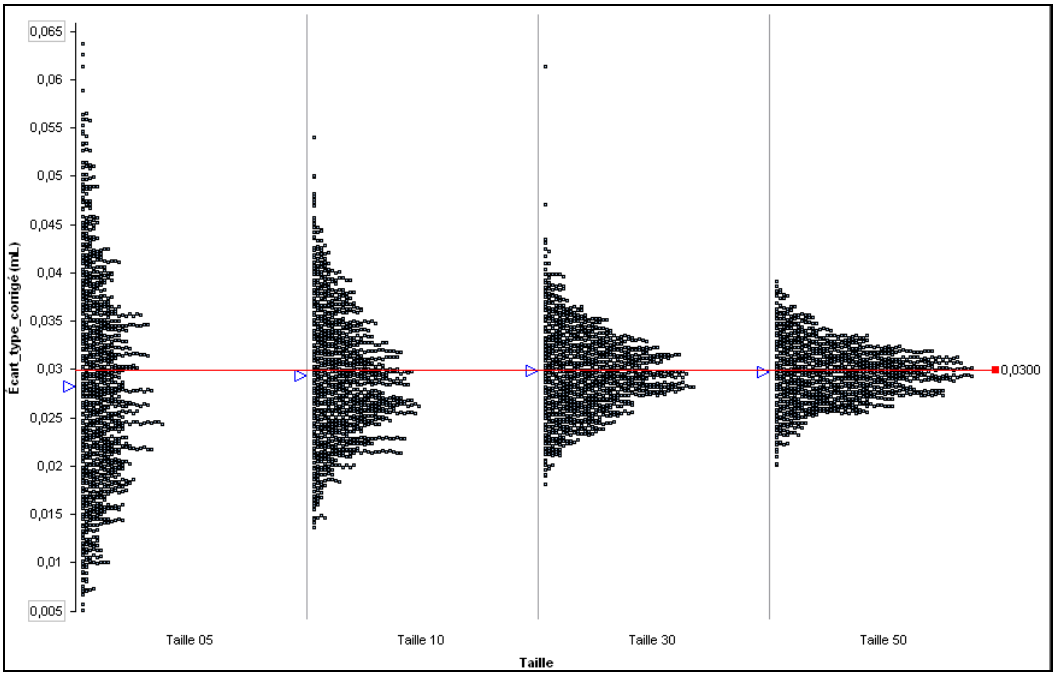
Distribution des variances corrigées des 5 000 échantillons regroupées par taille d'échantillon :



Distribution des écarts-types des 5 000 échantillons regroupés par taille d'échantillon :



Distribution des écarts-types corrigés des 5 000 échantillons regroupés par taille d'échantillon :



V - Théorèmes

Soit X_1, X_2, ..., X_n une suite de n variables aléatoires indépendantes de même loi de probabilité admettant une espérance mathématique μ. On pose $\overline{X} = \frac{1}{n}(X_1 + X_2 + ... + X_n)$.

1°) - Loi faible des grands nombres

Pour tout $\varepsilon > 0$, $P(|\overline{X} - \mu| < \varepsilon)$ tend vers 1 quand n tend vers l'infini.

Application à la distribution de la fréquence dans les échantillons avec remise

Considérons dans une population $\mathcal{P}$, une sous-population $\mathcal{A}$. $\mathcal{S}(\mathcal{P})$ désignant l'ensemble des suites d'éléments de $\mathcal{P}$, on considère les variables aléatoires X_i pour $i \in \mathbb{N}^*$: X_i :

$$\begin{array}{ccc} \mathcal{S}(\mathcal{P}) & \longrightarrow & \{0, 1\} \\ (a_1, \ldots, a_n, ...) & \longmapsto & \begin{cases} 0 & \text{si } a_i \text{ n'appartient pas à } \mathcal{A} \\ 1 & \text{si } a_i \text{ appartient à } \mathcal{A} \end{cases} \end{array}$$

Les X_i suivent la loi de Bernoulli de paramètre p. $E(X_i) = p$ et $\sigma(X_i) = \sqrt{p(1 - p)}$

Alors la variable aléatoire qui à chaque échantillon de taille n associe la fréquence d'individus de $\mathcal{A}$ est $F = \frac{1}{n}(X_1 + X_2 + ... + X_n)$.

Pour tout $\varepsilon > 0$, $P(|F - p| < \varepsilon)$ tend vers 1 quand n tend vers l'infini : plus n est grand, plus la probabilité d'avoir un échantillon avec une fréquence de $\mathcal{A}$ observée f proche de p est forte.

Remarque : Ce théorème permet d'expliquer le fait que l'on peut attribuer comme probabilité à un événement, une valeur autour de laquelle la fréquence d'apparition de cet événement se stabilise lorsqu'on répète l'expérience aléatoire un grand nombre de fois.

2°) - Conséquence du théorème limite central

On suppose que les X_i ont un (même) écart-type σ.

Alors pour n grand, la loi de la moyenne $\overline{X}$ peut être approchée par la loi normale de paramètres μ et $\frac{\sigma}{\sqrt{n}}$.

Remarques :

- Si les X_i sont distribuées selon une même loi normale la loi de $\overline{X}$ est la loi normale de paramètres μ et $\frac{\sigma}{\sqrt{n}}$.

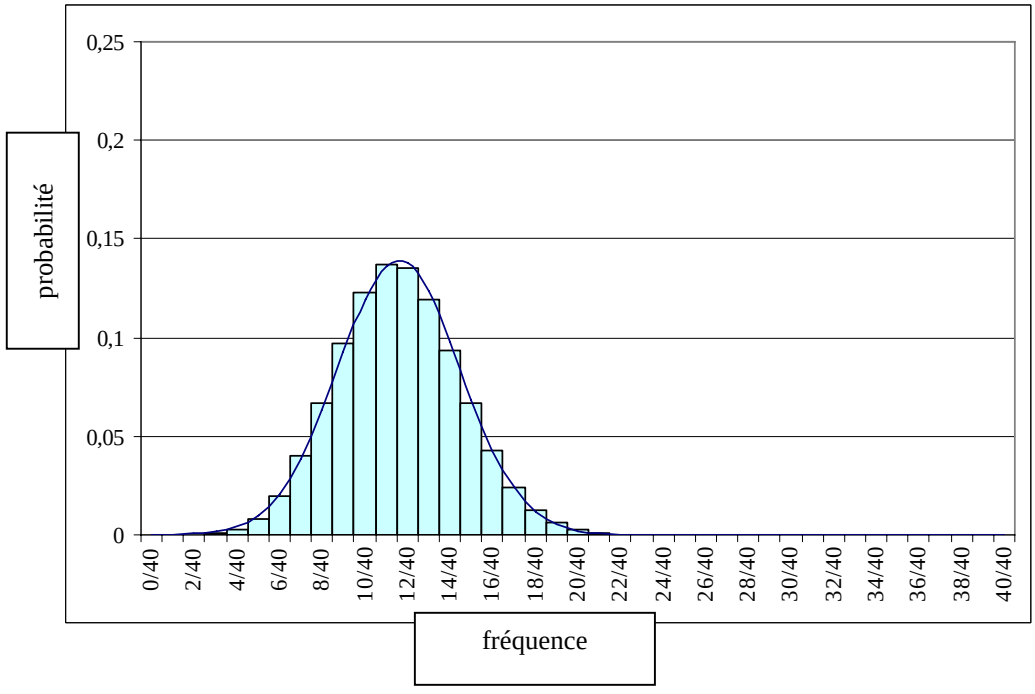
- Ce théorème est remarquable à plusieurs égards :

  - Il s'applique quelle que soit la loi de probabilité commune des X_i pourvu qu'elle possède une variance.
  - Il contribue à expliquer le rôle privilégié que joue la loi normale en analyse statistique (en tant que limite d'une multitude de phénomènes aléatoires).
  - Il prouve qu'une suite de variables aléatoires discrètes peut converger en loi vers une variable aléatoire continue.

Application à la distribution de la fréquence d'échantillonnage

- La loi de nF est la loi binomiale $\mathcal{B}(n\,;p)$.
- D'après le théorème limite central appliqué à la suite précédente, la loi de F est approchée par la loi normale $\mathcal{N}\left(p\,;\sqrt{\frac{p\,(1-p)}{n}}\right)$ pour $n \geq 30$, $np \geq 15$ et $np(1-p) > 5$. Dans ces conditions, la loi de $\frac{F-p}{\sqrt{\frac{p\,(1-p)}{n}}}$ est approchée par la loi normale $\mathcal{N}(0\,;\,1)$.

Illustration pour des échantillons de taille 40 :



- La théorie des probabilités permet de dire que la loi de $\frac{F-p}{\sqrt{\frac{F\,(1-F)}{n}}}$ est approchée par la loi normale $\mathcal{N}(0\,;\,1)$ pour $n \geq 30$.