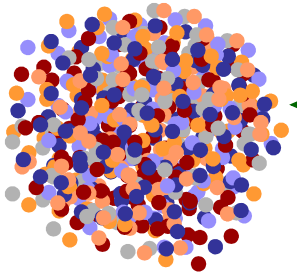
Lecture nº 2.

"Thank you Thompson,
for collecting the statistics"

Sampling theory.

Etudier la population ou l'échantillon?

Nous nous intéresserons à des propriétés d'un vaste ensemble (population) d'individus, d'objets ou de mesurages.

POPULATION 

RECENSEMENT → **paramètres : μ, σ, p, …**

étude exhaustive de tous les éléments de la population

INCONVENIENTS:
- temps long et coût élevé;
- impossible si la population est grande ou infinie.

ECHANTILLON 

SONDAGE → **statistiques : $\bar{x}$, S, p, …**

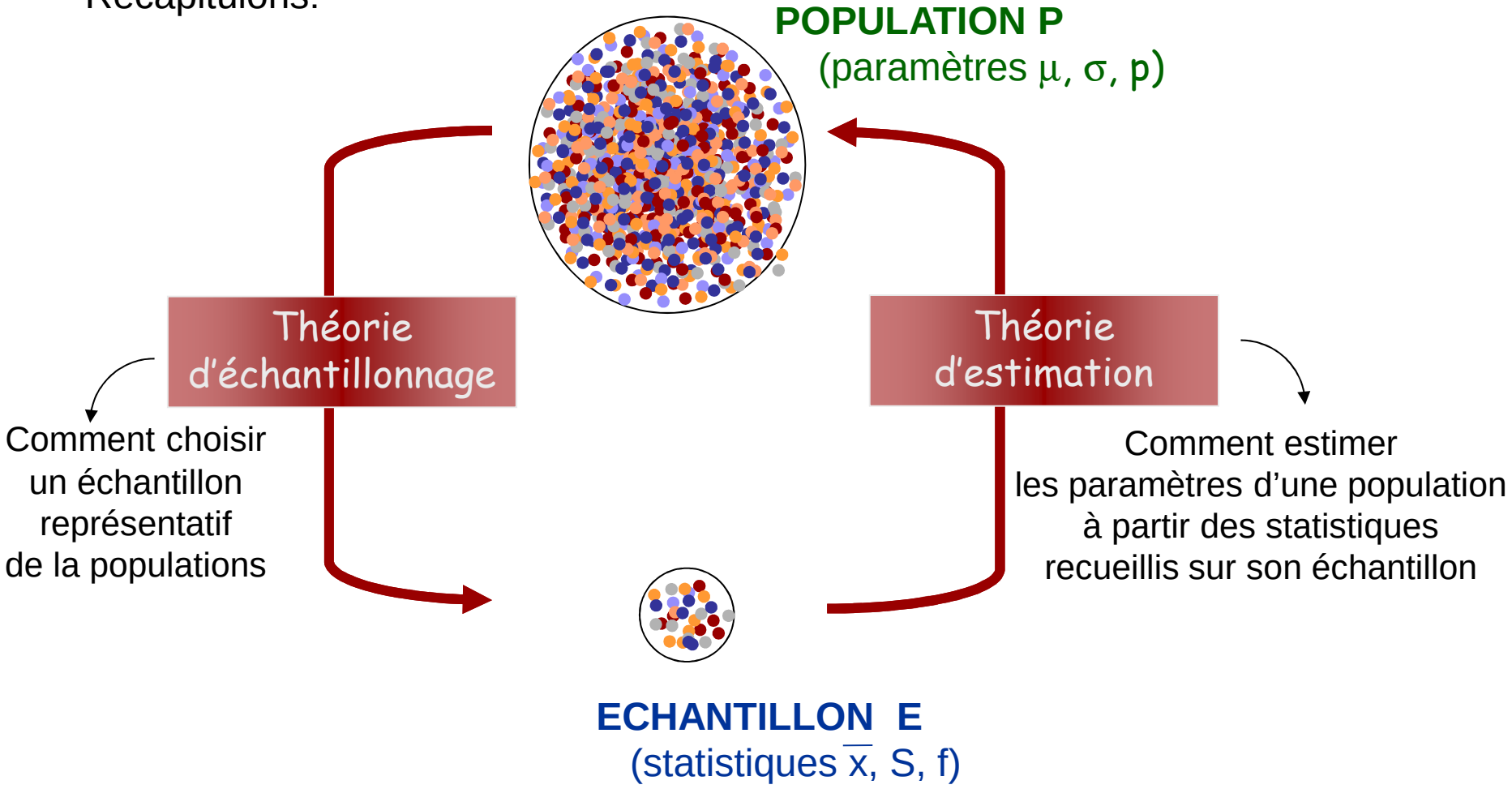
étude d'une partie (l'échantillon) de la population

PROBLEMES à RESOUDRE:

1. Comment choisir un échantillon représentatif de la populations ?
2. Comment estimer les paramètres à partir des statistiques ?

Echantillonnage et estimation.

Récapitulons:



Ce cours est totalement consacré à la théorie d'échantillonnage.

Choix d'un échantillon.

Chaque échantillon soumis à une étude statistique doit être:

- **représentatif** de la population,
- **aléatoire** (randomisé).

Echantillonnage randomisée: il doit remplir deux conditions:

égalité de choix → chaque échantillon de même taille a les mêmes chances (même probabilité) d'être retiré de la population.

(ceci sous-entend que toute observation qui pourrait être faite dans la population entière a des chances égales d'apparaître lors de l'étude de l'échantillon)

indépendance → le choix d'un échantillon ne préjuge rien sur les choix consécutifs d'autres échantillons.

Si l'échantillonnage n'est pas aléatoire, on dit qu'il est biaisé.

Comment préparer un échantillon randomisé?

EXEMPLE 1 : Avant les élections, on veut effectuer un sondage de la tendance politique d'un village de 1000 électeurs. On décide d'interroger 50 personnes. Comment les choisir?

- à chaque individu on attribue un numéro;
- on retire au hasard un nombre des valeurs de 0 à 1000 correspondant à la taille de l'échantillon voulu (50 dans notre exemple); ces valeurs indiquent les individus à choisir pour le sondage.

Comment obtenir des nombres aléatoires ?

- utiliser les tables statistiques donnant des nombres aléatoires;
- utiliser des générateurs de nombres aléatoires fournis avec la plupart des softwares type 'tableur' (fonction ***rand*** dans la plupart des logiciels, ***alea*** dans Excell)
- écrire un algorithme qui va générer des nombres aléatoires.

Générateurs des nombres aléatoires

L'algorithme le plus souvent utilisé pour générer les nombres pseudo-aléatoires est basé sur la formule récurrente

$$n_{n+1} = (k \cdot n_n + l) \bmod m$$

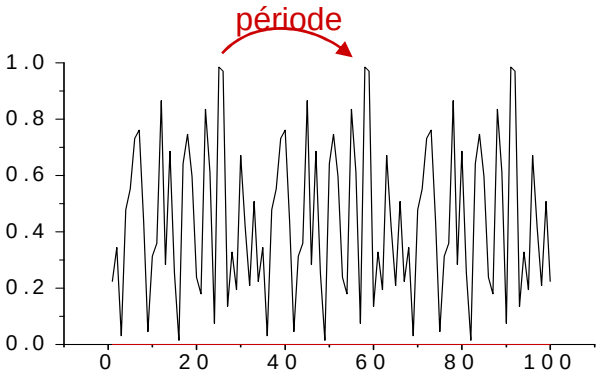
k,l,m,n_0 – entiers non négatifs
n_0 – seed – valeur de départ

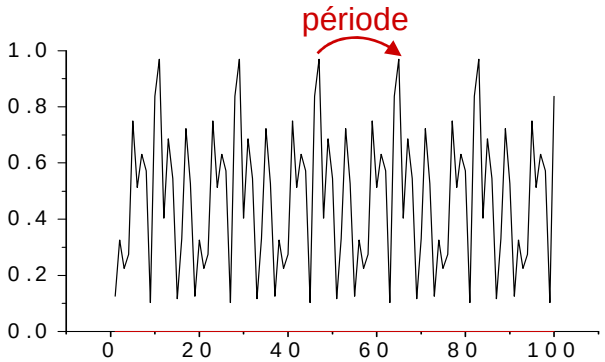
Les valeurs n_n ainsi obtenues appartiennent à l'intervalle $n_n \in [0, m-1]$.

Pour obtenir des nombres aléatoires r_n de l'intervalle [0;1), quel que soit le choix de k,l,m,n_0

$$r_n = \frac{n_n}{m} \longrightarrow r_n \in [0, 1)$$

La séquence de nombres peux se répéter; on dit que le générateur est périodique (pour des bon générateurs, la périodicité dépasse 10^{13}).

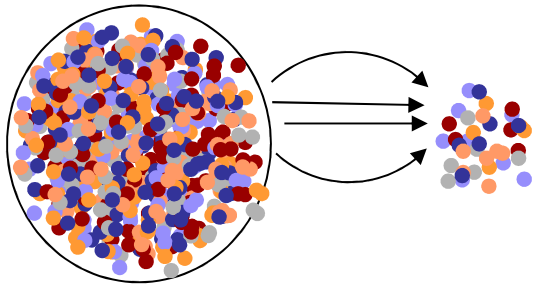




Types d'échantillonnage.

Echantillonnage exhaustif (sans remplacement):

chaque élément de la population peut être choisi seulement une fois.



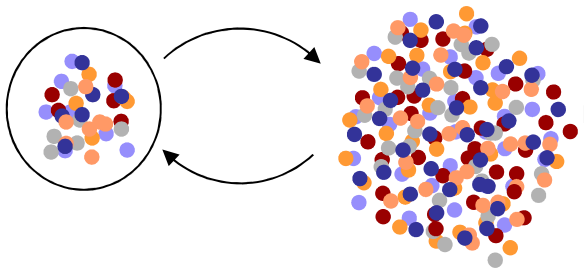
DANGER: si la population est petite,
on peut la 'vider' en l'échantillonnant.

DONC, applicable aux grandes populations uniquement.

(retirer un élément ne modifie pas visiblement
la composition de la population)

Echantillonnage non-exhaustif (avec remplacement):

après le tirage élément est 'rendu' à la population;
il peut être choisi plusieurs fois.

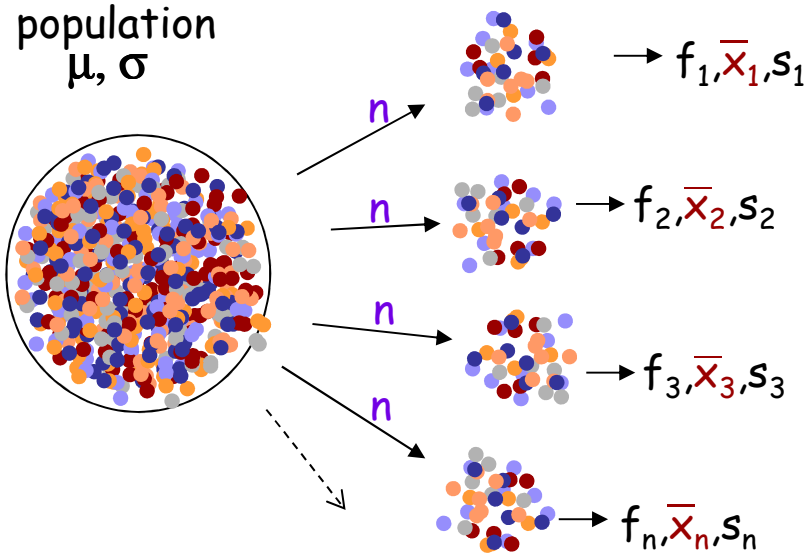


DONC, applicable à toute population.

(permet de traiter les petites populations
comme des populations infinies.)

Statistiques d'un échantillon.

On s'intéresse à l'étude d'un caractère **X** dans une population qu'on ne peut pas (on ne veut pas) recenser.



on extrait plusieurs échantillons :

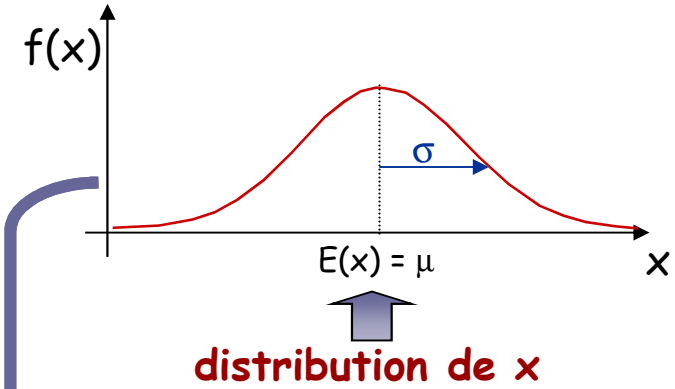
- représentatif**s**,
- aléatoires,
- de taille n fixée.

A PRIORI,
les caractéristiques des échantillons ne sont pas constantes;
elles varient (fluctuent) d'un échantillon à l'autre
(on dit qu'on observe des **fluctuations d'échantillonnage**).

DONC, les statistiques sont des VARIABLES aléatoires: elles sont distribuées.

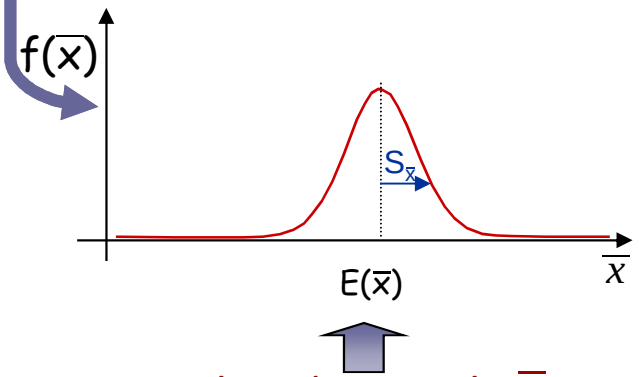
Distribution des moyennes d'échantillonnage.

Admettons que nous étudions la propriété **X** dans la population **P(μ,σ)** de taille **N**.
Pour cela nous retirons des échantillons de taille n et nous calculons $\bar{x}$.



distribution de x

ECHANTILLONNAGE
(n échantillons de taille n)



distribution de $\bar{x}$

❶ Les moyennes des échantillons restent centrées autour de la moyenne de la population:

$$E(\bar{x}) = E(x) = \mu$$

❷ Les moyennes x sont **toujours moins dispersées** que valeurs x_i:

- si l'échantillonnage est exhaustif et n ≤ N,

$$S_{\bar{x}}^2 = \frac{\sigma^2}{n} \cdot \frac{N-n}{n-1}$$

- si la population est grande (N → ∞),

$$S_{\bar{x}}^2 = \frac{\sigma^2}{n}$$

EXEMPLE 2: Le nombre des composantes défectueuses reçues par une entreprise d'assemblage des ordinateurs et provenant de fournisseurs A, B, C et D est, respectivement, 2, 4, 6 et 8. Quelle est le nombre moyen de pièces défectueuses reçues et l'écart type sur cette valeur?

$$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{2+4+6+8}{4} = 5 \qquad \sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2} = \sqrt{5} = 2.24$$

Retirons de cette population tous les échantillons de deux éléments (il y en a 4^2 =16). Quelle est la moyenne et l'écart type de moyenne de ces échantillons ?

(2;2) → 2
(2;4) → 3
(2;6) → 4
(2;8) → 5
(4;2) → 3
(4;4) → 4
(4;6) → 5
(4;8) → 6
(6;2) → 4
(6;4) → 5
(6;6) → 6
(6;8) → 7
(8;2) → 5
(8;4) → 6
(8;6) → 7
(8;8) → 8

$$E(\bar{x}) = \frac{1}{n}\sum_{i=1}^{n} \bar{x}_i = \frac{2+2\cdot 3+3\cdot 4+4\cdot 5+3\cdot 6+2\cdot 7+8}{16} = \frac{80}{16} = 5$$

$$S_{\bar{x}} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\bar{x}_i - E(\bar{x}))^2} = \sqrt{2.5} = 1.58$$

on vérifie bien que $E(\bar{x}) = \mu$, $S_{\bar{x}} \leq \sigma$ et

$$S_{\bar{x}}^2 = \frac{\sigma^2}{n} = \frac{5}{2} = 2.5 \Rightarrow S_{\bar{x}} = 1.58$$

Distribution des fréquences d'échantillonnage.

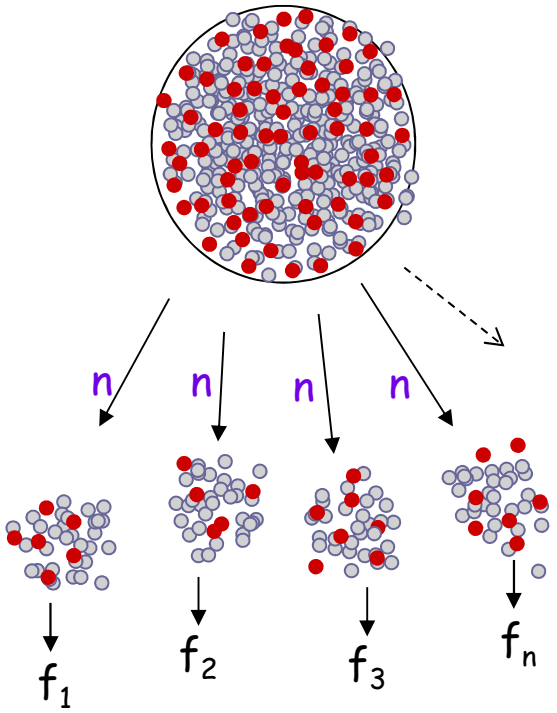
- Supposons que l'apparition d'un caractère **X** suit la distribution binomiale avec la probabilité d'apparition de ce caractère (probabilité de 'succès') égale **p**.

- Retirons de cette population tous les échantillons de taille **n** et calculons la fréquence d'apparition du caractère **X** → $\mathbf{f_i}$.

  - La distribution de f_i est caractérisée par les paramètres:

$$E(f) = p$$

$$S_f^2 = \frac{p(1-p)}{n}$$



EXEMPLE 3: Dans un groupe de 500 personnes, chacun a joue au 'pile ou face' 120 fois. Combien de personnes auront entre 40 et 60% de 'faces' ?

Si les pièces ne sont pas truquées, la probabilité d'avoir la 'face' est p = 1/2.
Que x décrit le nombre de 'faces' obtenus durant 120 jeux.

Première méthode:

$$\left.\begin{aligned} E(x) &= np = 60 \\ \sigma(x) &= \sqrt{npq} = 5.48 \end{aligned}\right\} \Rightarrow$$

Si on doit avoir entre 40 et 60% de 'faces', alors

$$48 < x < 72$$

$$\frac{48-60}{5.48} < Y < \frac{72-60}{5.48}$$

$$-2.19 < Y < 2.19$$

$$P(-2.19 < Y < 2.19) = 0.9714$$

$$0.9714 \cdot 500 \approx 486 \ personnes$$

Deuxième méthode:

$$\left.\begin{aligned} E(f) &= p = 1/2 \\ \sigma(f) &= \sqrt{\frac{pq}{n}} = 0.04564 \end{aligned}\right\} \Rightarrow$$

Si on doit avoir entre 40 et 60% de 'faces', alors

$$0.4 < f < 0.6$$

$$\frac{0.4-0.5}{0.04564} < Y < \frac{0.6-0.5}{0.04564}$$

$$-2.19 < Y < 2.19$$

$$P(-2.19 < Y < 2.19) = 0.9714$$

$$0.9714 \cdot 500 \approx 486 \ personnes$$

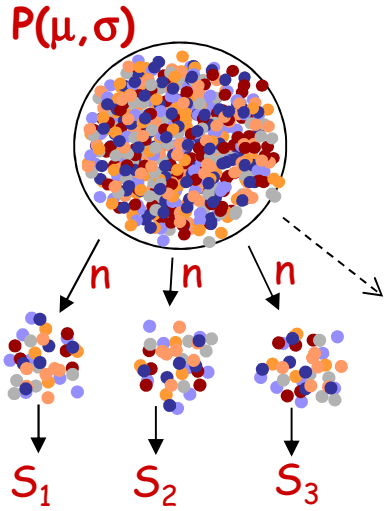
Distribution des variances d'échantillonnage.

- Souvent nous sommes intéressés par la distribution de variances S^2_i des échantillons de taille identique **n**, retirés de la même population **P(μ, σ)** (étalonnages, contrôles de fiabilité…) .

  - La moyenne de distribution des variances est:

$$E(S^2) = \frac{n-1}{n}\sigma^2$$

ATTENTION:
si E(statistique) ≠ paramètre,
on appelle cette statistique
un estimateur biaisé du paramètre.



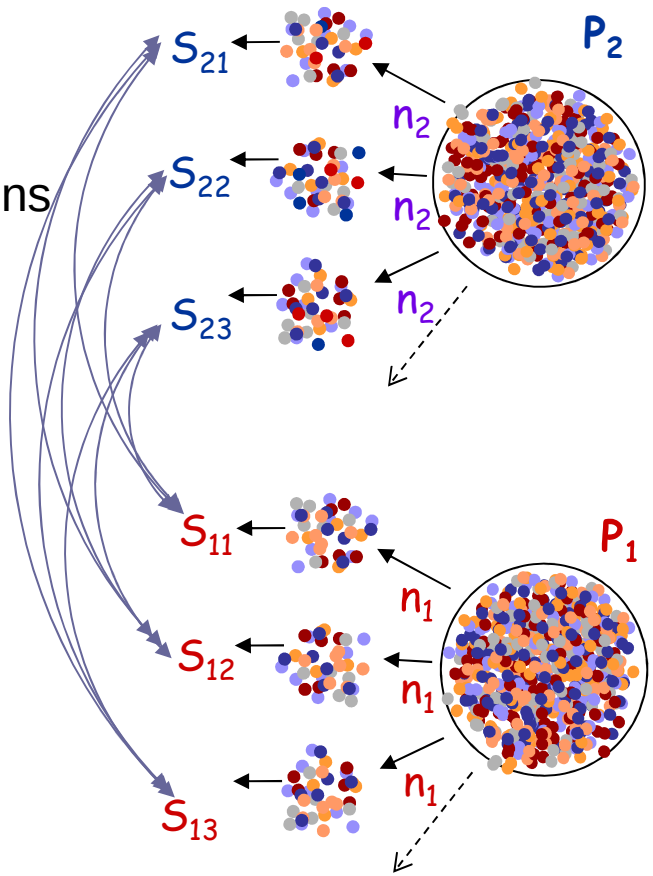
  - D'habitude on ne décrit pas la distribution des variances, mais la distribution du paramètre relatif:

$$\chi^2 = \frac{nS^2}{\sigma}$$

Distribution des sommes des statistiques.

- Nous disposons de deux populations P_1 et P_2.
- Formons toutes les combinaisons possibles des paires d'échantillons provenant de ces populations
- Calculons la somme de mêmes statistiques (moyenne, variance, fréquence…) déterminées sur ces échantillons.
- La distribution de sommes admettra pour moyenne et variance:

$$E(S_1 \pm S_2) = E(S_1) \pm E(S_2)$$
$$S^2(S_1 \pm S_2) = S^2(S_1) + S^2(S_2)$$



EXEMPLE 4: La durée de vie des pneus *Firestone* est de 120 000 ± 20 000 miles.
La durée de vie des pneus *Goodyear* est de 80 000 ± 10 000 miles.
On a étudiée200 pneus de chaque marque. Quelle est la probabilité que les pneus *Firestone* durent 45 000 miles de plus que *Goodyear* ?

La différence espérée de la durée de vie de ces deux types de pneus est:

$$E(\bar{x}_1 - \bar{x}_2) = E(\bar{x}_1) - E(\bar{x}_2) =$$

$$= 120000 - 80000 = 40.000\ miles$$

Si on étudie les échantillons représentatifs de la production,
de taille $n_1 = n_2 = 200$, alors l'écart type observé sur cette différence est:

$$S(\bar{x}_1 - \bar{x}_2) = \sqrt{\frac{S^2(\bar{x}_1)}{n_1} + \frac{S^2(\bar{x}_2)}{n_2}} =$$

$$= \sqrt{\frac{(20.000)^2}{200} + \frac{(10.000)^2}{200}} = 1581\ miles$$

Comme les deux échantillons sont statistiquement grands (n_1, $n_2 > 30$),

$$P\{(\bar{x}_1 - \bar{x}_2) > 45.000\} = P\left\{Y > \frac{45.000 - 40.000}{1581}\right\} =$$

$$= P\{Y > 3.16\} = 0.0008$$