Échantillonnage, estimation et test pour une proportion

$\mathcal{P}$ désigne une population finie et A une sous-population en proportion p (on suppose $0.3 \le p \le 0.7$).

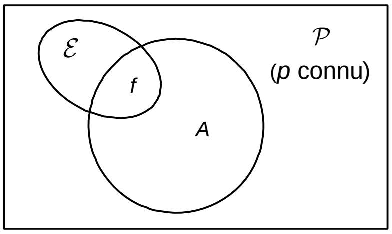
$\mathcal{E}$ désigne un échantillon de taille n tiré "au hasard" (c'est-à-dire avec équiprobabilité) dans $\mathcal{P}$ et

f la fréquence (ou proportion) de A dans $\mathcal{E}$ (on suppose $n \ge 30$).

Échantillonnage

On suppose p connu. A chaque échantillon de taille n on associe la fréquence *f.*
Comment se distribuent LES fréquences f par rapport à p ?

… de la population vers LES échantillons



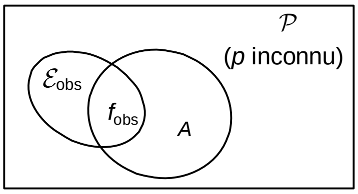
On vérifie par simulation que :

95% des échantillons de taille n donnent une fréquence f vérifiant :
$$f \in \left[p - \frac{1}{\sqrt{n}} \,;\, p + \frac{1}{\sqrt{n}} \right]$$

Estimation

On suppose p inconnu. A partir d'UN échantillon observé de taille n, et donc d'UNE fréquence observée f_{obs}, *estimation de p par intervalle de confiance à* 95%.

… d'UN échantillon vers la population

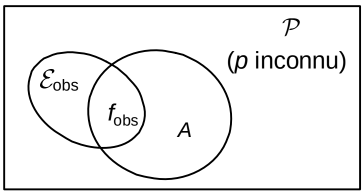


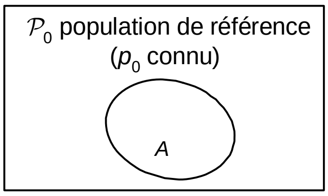
On a, à 95% de confiance :
$$p \in \left[f_{\text{obs}} - \frac{1}{\sqrt{n}} \,;\, f_{\text{obs}} + \frac{1}{\sqrt{n}} \right]$$

Test

On suppose p inconnu. A partir d'UN échantillon observé de taille n,
test de l'hypothèse $H_0 : p = p_0$ (p_0 connu).

… d'UN échantillon vers la population





Si H_0 est vrai, 95% des échantillons de taille n donnent une fréquence f vérifiant $f \in \left[p_0 - \frac{1}{\sqrt{n}} \,;\, p_0 + \frac{1}{\sqrt{n}} \right]$

Décision au seuil de 5% :

Si $f_{\text{obs}} \notin \left[p_0 - \frac{1}{\sqrt{n}} \,;\, p_0 + \frac{1}{\sqrt{n}} \right]$ on rejette H_0.

La différence entre f_{obs} et p_0 est significative au seuil de 5% ; elle révèle une différence entre p et p_0.

Si $f_{\text{obs}} \in \left[p_0 - \frac{1}{\sqrt{n}} \,;\, p_0 + \frac{1}{\sqrt{n}} \right]$ on ne rejette pas H_0.

La différence entre f_{obs} et p_0 n'est pas significative au seuil de 5% ; on peut avoir $p = p_0$.

Échantillonnage et test pour une distribution

$\mathcal{P}$ désigne une population finie et $\{A_1,...,A_k\}$ une partition en proportions $(p_1,...,p_k)$ $(k \geq 2)$.
$\mathcal{E}$ désigne un échantillon de taille n tiré "au hasard" (c'est-à-dire avec équiprobabilité) dans $\mathcal{P}$, $(f_1,...,f_k)$ la distribution de fréquences de $\{A_1,...,A_k\}$ dans $\mathcal{E}$ (on suppose $n \geq 30$).

On pose : $d^2 = n\sum_{i=1}^{k}\frac{(f_i - p_i)^2}{p_i}$, indice positif ou nul, nul lorsque $(f_1,...,f_k) = (p_1,...,p_k)$.

Échantillonnage

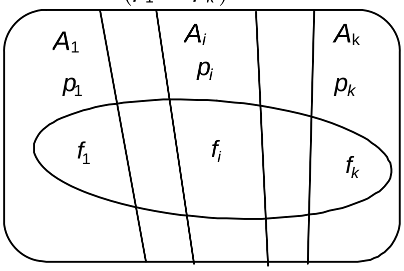
On suppose la distribution $(p_1,...,p_k)$ connue.

A chaque échantillon de taille n on associe la distribution de fréquence $(f_1,...,f_k)$ et l'indice d^2.

Comment se distribuent les distributions de fréquences $(f_1,...,f_k)$ par rapport à $(p_1,...,p_k)$? plus simplement, comment se distribue l'indice d^2 sur l'ensemble des réels positifs ?

… de la population vers LES échantillons

$\mathcal{P}$ $(p_1,\ldots,p_k)$ connu



On vérifie par simulation que cette distribution ne dépend pas de $(p_1,...,p_k)$ mais seulement de k.

95% des échantillons de taille n donnent un indice d^2 inférieur ou égal à a_k :

k	2	3	4	5	6	10	12
a_k	3.84	5.99	7.81	9.49	11.07	16.92	19.68

Test d'adéquation à l'équirépartition

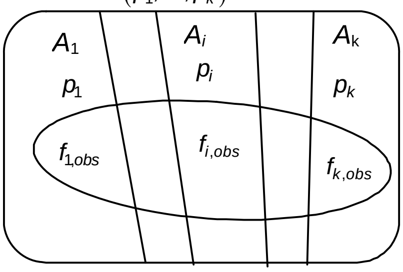
On suppose la distribution $(p_1,...,p_k)$ inconnue.

A partir d'UN échantillon observé de taille n, *test de l'hypothèse* $H_0 : (p_1,...,p_k) = \left(\frac{1}{k},...,\frac{1}{k}\right)$.

On observe UN échantillon et on calcule $d^2_{\text{obs}} = nk\sum_{i=1}^{k}\left(f_{i,\text{obs}} - \frac{1}{k}\right)^2$

… d'UN échantillon vers la population

$\mathcal{P}$ $(p_1,\ldots,p_k)$ inconnu



Si H_0 est vrai, 95% des échantillons de taille n donnent un indice d^2 inférieur ou égal à a_k

Décision au seuil de 5% :
Si $d^2_{\text{obs}} > a_k$ **,** on rejette H_0 ; il est peu probable que la distribution de fréquence observée provienne d'une loi équirépartie.
Si $d^2_{\text{obs}} \leq a_k$ **,** on ne rejette pas H_0 ; la distribution de fréquence observée peut provenir d'une loi équirépartie.