Vers les tests du χ^2

Définition : la loi de probabilité de la somme des carrés de *k* v.a.r. i.i.d. $\mathcal{N}(0;1)$ est appelée loi du Khi2 à *k* degrés de liberté (*d.d.l.*) et notée χ^2_k.

Soit $(X_1,\ldots,X_k)$ un vecteur aléatoire de loi multinomiale $\mathcal{M}(1\,;p_1,\ldots,p_k)$ $p_j>0$ $\sum_{j=1}^{k} p_j = 1$
(modélisation du tirage d'une boule dans une urne à *k* catégories).

Soit $\left(X_{1_i},\ldots,X_{k_i}\right)_{i=1,\ldots,n}$ *n* vecteurs aléatoires *i.i.d.* comme $(X_1,\ldots,X_k)$
(*n* tirages avec remise dans l'urne précédente).

On pose $\forall j \in [\![1,k]\!]$ $Y_j = \sum_{i=1}^{n} X_{ji}$ (nombre de boules de la catégorie *j* sur les *n* boules tirées).

Alors $(Y_1,\ldots,Y_k) \sim \mathcal{M}(n;p_1,\ldots,p_k)$ $P\left(\bigcap_{j=1}^{k}\left[Y_j = n_j\right]\right) = \begin{cases} \dfrac{n!}{\prod_{j=1}^{k} n_j!}\prod_{j=1}^{k} p_j^{n_j} & \text{si } \sum_{j=1}^{k} n_j = n \\ 0 \text{ sinon} \end{cases}$

$\forall j \in [\![1,k]\!]$ $Y_j \sim \mathcal{B}(n;p_j)$ $E(Y_j) = np_j$ $V(Y_j) = np_j(1-p_j)$

On a $\sum_{j=1}^{k} Y_j = n$ et $\operatorname{cov}(Y_j,Y_{j'}) = -np_jp_{j'}$ pour $j \neq j'$; les v.a.r. $(Y_j)_{j=1,\ldots,k}$ ne sont pas indépendantes.

On pose : $D_n^2 = \sum_{j=1}^{k} \dfrac{(Y_j - np_j)^2}{np_j}$

Théorème : $\left(D_n^2\right) \xrightarrow[n\to\infty]{Loi} \chi^2_{(k-1)}$

Dans toute la suite, les variables considérées sont catégorielles. Plus exactement, lorsqu'elles sont réelles, on ne considèrera que les effectifs, fréquences ou probabilités associés aux valeurs de la variable si elle est finie ou aux classes si les valeurs sont regroupées en classes. On ne fait donc aucun calcul sur les valeurs des variables, les tests du χ^2 font partie des tests dits "non paramétriques".

Test du χ^2 d'adéquation à une loi de probabilité fixée

Soit $(n_1,\ldots,n_k)\left(\sum_{j=1}^{k} n_j = n\right)$ la distribution d'effectifs observée sur un échantillon de taille *n* d'une variable catégorielle à *k* modalités (ce peut donc être une variable discrète à *k* valeurs distinctes ou une variable réelle continue dont les valeurs sont regroupées en *k* classes).
L'échantillon est supposé être une observation de *n* v.a.r. de loi *P*.

Soit $(p_1,\ldots,p_k)$ une distribution de probabilité fixée. La question est la suivante :
Au seuil de 5 % (probabilité d'erreur fixée a priori), peut-on accepter que la loi *P* (dont provient l'échantillon) est égale à $(p_1,\ldots,p_k)$?
On note H_0 l'hypothèse à tester, dite hypothèse nulle, H_0 : $P = (p_1,\ldots,p_k)$

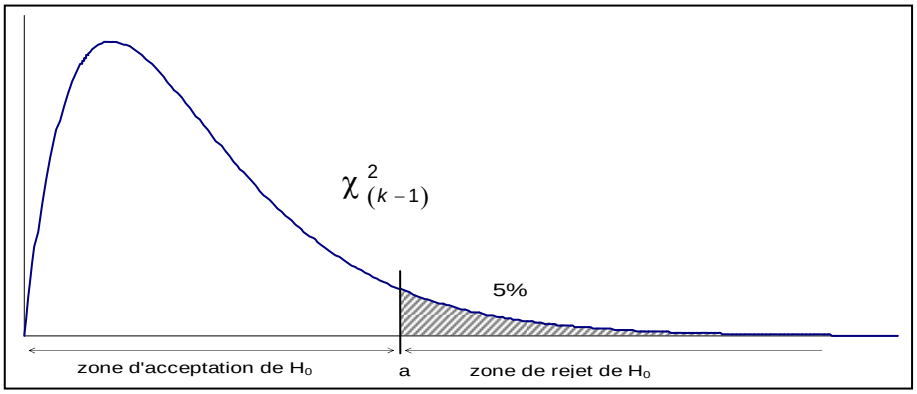
On calcule $d_n^2 = \sum_{j=1}^{k} \dfrac{(n_j - np_j)^2}{np_j}$.

Si H_0 est vraie, alors d_n^2 est l'observation d'une v.a.r. D_n^2 de loi asymptotique un $\chi^2_{(k-1)}$.

On compare alors d_n^2 à la valeur *a* vérifiant $P\left(\chi^2_{(k-1)} \leq a\right) = 0.95$ et l'on conclut selon que d_n^2 est dans la zone de rejet de H_0 ($d_n^2 \geq a$) ou dans la zone d'acceptation de H_0 ($d_n^2 < a$) (cf. graphique) (*a* est appelé quantile d'ordre 0.95 ou dix-neuvième vingtile).

Le seuil de 5% apparaît donc comme la probabilité de rejeter l'hypothèse H_0 à tort.

"Accepter H_0" est en fait "ne pas rejeter H_0" ; les données observées sont compatibles avec l'hypothèse H_0, ce qui ne veut pas dire qu'elle est vraie. Les données observées sont compatibles avec bien d'autres hypothèses.



Test du χ^2 d'indépendance de deux variables catégorielles à *p* et *q* modalités

Soit $\left(n_{ij}\right)_{\substack{1\leq i\leq p\\1\leq j\leq q}}$ $\left(\sum_{i=1}^{p}\sum_{j=1}^{q} n_{ij} = n\right)$ la distribution conjointe d'effectifs observée sur un échantillon de taille *n* d'un couple de variables (X,Y) à respectivement *p* et *q* modalités.

L'échantillon est supposé *i.i.d.* de loi $P = \left(p_{ij}\right)_{\substack{1\leq i\leq p\\1\leq j\leq q}}$.

Au seuil de 5 %, peut-on accepter l'hypothèse H_0 : les deux variables sont indépendantes en probabilité, c-à-d $\forall (i,j) \in [\![1,p]\!]\times[\![1,q]\!]$ $p_{ij} = p_{i.}p_{j.}$ (avec $p_{i.} = \sum_{j=1}^{q} p_{ij}$ et $p_{.j} = \sum_{i=1}^{p} p_{ij}$) ?

Comme les probabilités marginales $\left(p_{i.}\right)_{i=1,\ldots,p}$ et $\left(p_{.j}\right)_{j=1,\ldots,q}$ ne sont en général pas connues, on les estime par $\hat{p}_{i.} = \frac{n_{i.}}{n}$ et $\hat{p}_{.j} = \frac{n_{.j}}{n}$ (avec $n_{i.} = \sum_{j=1}^{q} n_{ij}$ et $n_{.j} = \sum_{i=1}^{p} n_{ij}$).

On calcule $d_n^2 = \sum_{i=1}^{p}\sum_{j=1}^{q} \frac{\left(n_{ij} - \frac{n_{i.}n_{.j}}{n}\right)^2}{\frac{n_{i.}n_{.j}}{n}}$.

Si H_0 est vraie alors d_n^2 est l'observation d'une v.a.r. D_n^2 de loi asymptotique un $\chi^2_{(p-1)(q-1)}$.

On conclut comme précédemment.

Test du χ^2 d'homogénéité

(c-à-d test d'égalité de *q* distributions à *p* modalités, les mêmes pour les *q* distributions)

Soit $\left(n_{ij}\right)_{\substack{1\le i\le p\\ 1\le j\le q}}$ les distributions d'effectifs observées sur *q* échantillons indépendants de taille $n_{\cdot 1},\ldots,n_{\cdot j},\ldots,n_{\cdot q}$ fixées. On pose $n=\sum_{j=1}^{q} n_{\cdot j}$.

Ces échantillons peuvent être supposés tirés indépendamment l'un de l'autre de *q* sous-populations deux à deux disjointes.

Pour tout *j* de 1 à *q*, l'échantillon *j* est supposé *i.i.d.* de loi $P_j=\left(q_{ij}\right)_{i=1,\ldots,p}$ $\sum_{j=1}^{p} q_{ij}=1$.

Au seuil de 5 %, peut-on accepter l'hypothèse H_0 : les *q* distributions $\left(P_j\right)_{j=1,\ldots,q}$ sont égales,

c-à-d , $\forall (i,j)\in[\![1,p]\!]\times[\![1,q]\!]\quad q_{ij}=q_i$. où $\left(q_i\right)_{i=1,\ldots,q}$ désigne la distribution commune.

On calcule $d_n^2=\sum_{i=1}^{p}\sum_{j=1}^{q}\frac{\left(n_{ij}-n_{\cdot j}q_i\right)^2}{n_{\cdot j}q_i}$.

Si H_0 est vraie, alors d_n^2 est l'observation une v.a.r. D_n^2 de loi asymptotique un χ^2_{pq-q} .

Comme les probabilités $\left(q_i\right)$ ne sont en général pas connues, on les estime par $\hat{q}_i=\frac{n_{i\cdot}}{n}$, $i=1,\ldots,p$.

On a alors $d_n^2=\sum_{i=1}^{p}\sum_{j=1}^{q}\frac{\left(n_{ij}-\frac{n_{i\cdot}n_{\cdot j}}{n}\right)^2}{\frac{n_{i\cdot}n_{\cdot j}}{n}}$.

Si H_0 est vraie, alors d_n^2 est l'observation d'une v.a.r. D_n^2 de loi asymptotique un $\chi^2_{(p-1)(q-1)}$.

On conclut comme précédemment. Il s'agit formellement du test d'indépendance.