Chi-deux

I Présentation

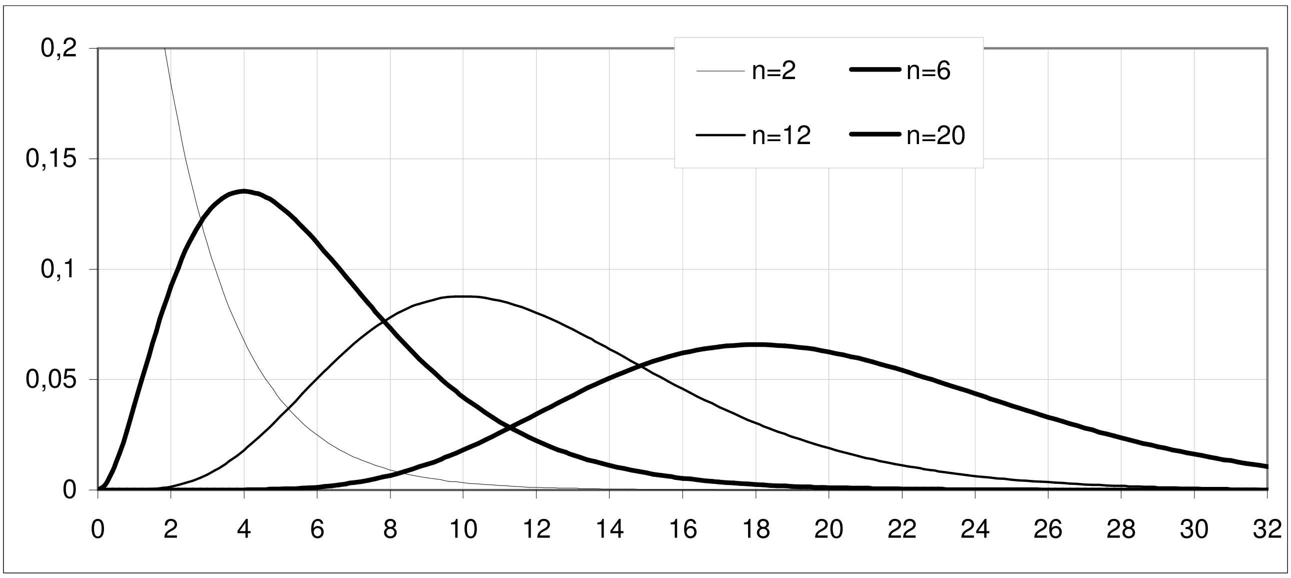
Définition 1 Soit $X_1, X_2, \ldots, X_n$, n variables indépendantes suivant la même loi normale $N(m\,;\sigma)$. Alors la variable aléatoire $Z = \frac{1}{\sigma^2}[(X_1 - m)^2 + (X_2 - m)^2 + \ldots + (X_n - m)^2]$ suit la loi du chi_deux à n degrés de liberté.

Remarque 1 On sait que si X_i suit la loi normale $N(m\,;\sigma)$, alors $\frac{X_i - m}{\sigma}$ suit la loi normale centrée réduite $N(0\,;1)$. D'où une définition équivalente :

Définition 2 Soit $X_1, X_2, \ldots, X_n$, n variables indépendantes suivant la loi normale centrée réduite $N(0\,;1)$. Alors la variable aléatoire $Z = X_1^2 + X_2^2 + \ldots + X_n^2$ suit la loi du chi_deux à n degrés de liberté. On la note $\chi^2_{(n)}$.

Remarque 2 $\chi^2_{(1)}$ a une espérance égale à 1 et une variance égale à 2. Donc $\chi^2_{(n)}$ a une espérance égale à n et une variance égale à 2n.
D'après le théorème central limite, pour n suffisamment grand, Z, qui est une somme de variables aléatoires indépendantes et de même loi, est approximativement distribuée selon une loi normale d'espérance n et de variance 2n, donc suit approximativement la loi $N(n\,;\sqrt{2n})$. L'approximation est excellente à partir de $n = 30$.

Allure des densités de probabilité La courbe de densité de $\chi^2_{(n)}$ a pour équation $y = ke^{-x/2}x^{n/2-1}$ où k est tel que l'aire sous la courbe vaut 1. On observe que lorsque n augmente, la courbe de densité de $\chi^2_{(n)}$ se rapproche d'une courbe de Gauss. Le maximum s'obtient pour $x = n - 2$



II Test d'ajustement à une loi.

Position du problème. On extrait d'une population (Π) un échantillon (E_n) de taille n. Les résultats $(x_j)_{1 \le j \le n}$ du caractère étudié sont regroupés en k classes, et on note $(o_i)_{1 \le i \le k}$ l'effectif observé dans chaque classe.

De même, soit $(t_i)_{1 \le i \le k}$ l'effectif théorique résultant de la loi du modèle théorique. Pour mesurer l'adéquation de l'échantillon au modèle théorique, on calcule la somme des écarts quadratiques réduits $d^2 = \sum_{i=1}^{k} \frac{(o_i - t_i)^2}{t_i}$

Dans chaque classe, l'effectif observé o_i varie quand on passe d'un échantillon (E_n) à un autre. o_i est la réalisation d'une variable aléatoire O_i.

Théorème. On détermine les classes pour que dans chacune d'elles $t_i \ge 5$. Dans ces conditions, pour n assez grand, la variable aléatoire $D^2 = \sum_{i=1}^{k} \frac{(O_i - t_i)^2}{t_i}$ suit une loi du $\chi^2_{(k-1)}$.

Mise en place du test.

Hypothèse H_0 : la distribution observée n'est pas significativement différente de la distribution théorique.

Seuil critique : on choisit $\alpha = P(\text{rejeter à tort } H_0)$
comme D^2 suit une loi du $\chi^2_{(k-1)}$, on en déduit le nombre h tel que
$\alpha = P(\chi^2_{(k-1)} > h)$

Règle de décision : si $d^2 > h$ on rejette H_0 au risque α
sinon, on ne rejette pas H_0, c'est-à-dire qu'on considère que la différence entre la distribution observée et la distribution théorique est due aux fluctuations d'échantillonnage.

III Test d'homogénéité.

Position du problème. Les observations d'une variable qualitative sur p échantillons permettent-elles de conclure que les échantillons proviennent de la même population?

Résolution La méthode consiste à comparer les effectifs observés de chacune des k modalités sur les p échantillons, avec les effectifs théoriques qu'on devrait obtenir dans le cas où ils seraient issus d'une même population.

Pour cela, on fabrique deux tableaux de contingence de taille $p \times k$, l'un contenant les effectifs o_{ij} observés et l'autre les effectifs théoriques t_{ij}.

La variable aléatoire $D^2 = \sum_{i=1}^{p} \sum_{j=1}^{k} \frac{(O_{ij} - t_{ij})^2}{t_{ij}}$ suit approximativement la loi du $\chi^2_{(p-1)(k-1)}$

Mise en place du test.

Hypothèse H_0 : les p échantillons proviennent de la même population.

Seuil critique : on calcule h par $\alpha = P(\chi^2_{(p-1)(k-1)} > h)$

Règle de décision : si $d^2 > h$ on rejette H_0 au risque α
sinon, on ne rejette pas H_0, c'est-à-dire qu'on considère que les différences entre les distributions observées et les distributions théoriques sont dues aux fluctuations d'échantillonnage.