Statistique inférentielle

Tests usuels

A. Godichon-Baggioni

OBJECTIFS

Nous aborderons les tests suivants :

- **Test de conformité de la moyenne et/ou de la variance** : il s'agit de vérifier si la moyenne ou la variance d'une population suit une valeur donnée.
- **Test de Student** (pour échantillons appariés ou indépendants) : il permet de comparer les moyennes de deux échantillons afin de déterminer si leur différence est statistiquement significative.
- **Test de Fisher** (pour échantillons appariés ou indépendants) : il vise à comparer les variances de deux échantillons et à évaluer si elles diffèrent de manière significative.
- **Tests du Chi-deux :** ils permettent de tester
  - l'indépendance entre deux variables,
  - ou l'adéquation d'une loi empirique à une loi théorique donnée P_0.

I. Tests de conformité de la moyenne et la variance

OBJECTIFS

Cadre : On considère des v.a. i.i.d $X_1, \ldots, X_n$ avec

$$X_1 \sim \mathcal{N}\left(\mu, \sigma^2\right),$$

où μ et σ^2 sont inconnus.

Test de conformité d'une moyenne : Soit $\mu_0 \in \mathbb{R}$, on veut tester

$$H_0 : \mu = \mu_0 \quad \text{contre} \quad H_1 : \mu \neq \mu_0.$$

Test de conformité d'une variance : Soit $\sigma_0^2 > 0$, on veut tester

$$H_0 : \sigma^2 = \sigma_0^2 \quad \text{contre} \quad H_1 : \sigma^2 \neq \sigma_0^2.$$

Test de conformité d'une moyenne

Soit $\mu_0 \in \mathbb{R}$ et l'on souhaite tester, au risque α,

$$H_0 : \mu = \mu_0 \quad \text{contre} \quad H_1 : \mu \neq \mu_0.$$

On a

$$\sqrt{n}\,\frac{\overline{X}_n - \mu}{S_n} \sim T_{n-1},$$

Zone de rejet :

$$ZR = \left\{ \left| \overline{X}_n - \mu_0 \right| \geq t_{n-1,\,1-\alpha/2}\,\frac{S_n}{\sqrt{n}} \right\} = \left\{ \left| \sqrt{n}\,\frac{\overline{X}_n - \mu_0}{S_n} \right| \geq t_{n-1,\,1-\alpha/2} \right\}.$$

p-valeur : Soit $\overline{x}_n$ et s_n des réalisations :

$$p - \text{valeur} = 2 - 2F_T\left(\left| \sqrt{n}\,\frac{\overline{x}_n - \mu_0}{s_n} \right| \right) = \mathbb{P}\left[|T| \geq \left| \sqrt{n}\,\frac{\overline{x}_n - \mu_0}{s_n} \right| . \right]$$

Test de conformité d'une moyenne

Soit $\mu_0 \in \mathbb{R}$ et l'on souhaite tester, au risque α,

$$H_0 : \mu \leq \mu_0 \quad \text{contre} \quad H_1 : \mu > \mu_0.$$

On a

$$\sqrt{n}\,\frac{\overline{X}_n - \mu}{S_n} \sim T_{n-1},$$

Zone de rejet :

$$ZR = \left\{ \overline{X}_n \geq \mu_0 + t_{n-1,\,1-\alpha}\,\frac{S_n}{\sqrt{n}} \right\} = \left\{ \sqrt{n}\,\frac{\overline{X}_n - \mu_0}{S_n} \geq t_{n-1,\,1-\alpha} \right\}.$$

p-valeur : Soit $\overline{x}_n$ et s_n des réalisations :

$$p - \text{valeur} = \mathbb{P}\left[T \leq \frac{\overline{s}_n - \mu_0}{s_n} \right]$$

APPLICATION

Cadre : Après un traitement (régime alimentaire) sur une variété de porcs, on prélève un échantillon de $n = 5$ animaux et on obtient les poids (en kg) :

$$83,\ 81,\ 84,\ 80,\ 85.$$

On suppose que ces observations sont des réalisations i.i.d. d'une loi $\mathcal{N}(\mu, \sigma^2)$, avec μ et σ^2 inconnus. Le poids moyen historique est $\mu_0 = 87.6$ kg.

Le test : On teste au risque 5%

$$H_0 : \mu = \mu_0 \quad \text{contre} \quad H_1 : \mu \neq \mu_0.$$

On a les réalisations

$$\overline{x}_5 = 82.6, \qquad s_5 \approx 2.0736.$$

p-valeur : On a

$$p\text{-valeur} = 2\left(1 - F_{T_4}\left(\left|\sqrt{5}\,\frac{82.6 - 87.6}{2.0736}\right|\right)\right) \approx 2\big(1 - F_{T_4}(5.392)\big) \approx 0.00572.$$

Conclusion : Au niveau 5%, on rejette H_0.

Test de conformité d'une variance

Soit $\sigma_0^2 \in \mathbb{R}$ et l'on souhaite tester, au risque α,

$$H_0 : \sigma^2 = \sigma_0^2 \quad \text{contre} \quad H_1 : \sigma \neq \sigma_0.$$

On a

$$\frac{(n-1)S_n^2}{\sigma^2} \sim \chi^2_{n-1}.$$

Intervalle de confiance :

$$IC_{1-\alpha}\left(\sigma^2\right) = \left[\frac{(n-1)S_n^2}{k_{1-\alpha/2}} ; \frac{(n-1)S_n^2}{k_{\alpha/2}} \right],$$

Zone de rejet :

$$ZR = \left\{ \frac{(n-1)S_n^2}{\sigma_0^2} < k_{\alpha/2} \right\} \bigcup \left\{ \frac{(n-1)S_n^2}{\sigma_0^2} > k_{1-\alpha/2} \right\}.$$

p-valeur : Soit s_n une réalisation :

$$p - \text{valeur} = \min\left(2F_{\chi^2_{n-1}}\left(\frac{(n-1)s_n^2}{\sigma_0^2} \right), 2 - 2F_{\chi^2_{n-1}}\left(\frac{(n-1)s_n^2}{\sigma_0^2} \right) \right).$$

APPLICATION

Cadre : Exemple des porcs. La variance du poids historique est $\sigma_0^2 = 25$ kg.

Le test : On teste au risque 5%

$$H_0 : \sigma^2 = 25 \quad \text{contre} \quad H_1 : \sigma^2 \neq 25.$$

On a les réalisations

$$\bar{x}_5 = 82.6, \qquad s_5 \approx 2.0736.$$

p-valeur : On a

$$\text{p-valeur} = 2 \min \left\{ F\left(\frac{4 \times 4.3}{25}\right), 1 - F\left(\frac{4 \times 4.3}{25}\right) \right\} \approx 0.0944,$$

Conclusion : Au niveau 5%, on ne rejette pas H_0.

II. Tests de Student

Le problème

Le problème : On considère :

- $x_1, \ldots, x_p$, correspondant à p mesures d'une variable aléatoire X de moyenne μ_1 ;
- $y_1, \ldots, y_q$, correspondant à q mesures d'une variable aléatoire Y de moyenne μ_2.

Objectif : Tester au risque α

$$H_0 : \mu_1 = \mu_2 \quad \text{contre} \quad H_1 : \mu_1 \neq \mu_2.$$

Modèle probabiliste : On fait les hypothèses suivantes :

- Les données $x_1, \ldots, x_p$ sont les réalisations de variables aléatoires $X_1, \ldots, X_p$, indépendantes et identiquement distribuées selon $\mathcal{N}(\mu_1, \sigma_1^2)$.
- Les données $y_1, \ldots, y_q$ sont les réalisations de variables aléatoires $Y_1, \ldots, Y_q$, indépendantes et identiquement distribuées selon $\mathcal{N}(\mu_2, \sigma_2^2)$.
- Les deux échantillons sont indépendants et supposés de même variance, i.e. $\sigma_1^2 = \sigma_2^2 = \sigma^2$.

Quelques résultats théoriques

Proposition

On a

$$\overline{X}_p - \overline{Y}_q \sim \mathcal{N}\left(\mu_1 - \mu_2, \frac{\sigma^2}{p} + \frac{\sigma^2}{q}\right).$$

Proposition

On considère l'estimateur de variance commune S^2 défini par

$$S^2 = \frac{1}{p+q-2}\left(\sum_{i=1}^{p}(X_i - \overline{X}_p)^2 + \sum_{i=1}^{q}(Y_i - \overline{Y}_q)^2\right) = \frac{(p-1)S_X^2 + (q-1)S_Y^2}{p+q-2},$$

où

$$S_X^2 = \frac{1}{p-1}\sum_{i=1}^{p}(X_i - \overline{X}_p)^2 \quad \text{et} \quad S_Y^2 = \frac{1}{q-1}\sum_{i=1}^{q}(Y_i - \overline{Y}_q)^2.$$

Alors :

1. *La statistique*
$$\frac{(p+q-2)S^2}{\sigma^2} \sim \chi^2_{p+q-2}.$$
2. *Les variables $\overline{X}_p - \overline{Y}_q$ et S^2 sont indépendantes.*

Quelques résultats théoriques

Corollaire

On a

$$\frac{\sqrt{pq}}{\sqrt{p+q}} \frac{\left(\overline{X}_p - \overline{Y}_q\right) - (\mu_1 - \mu_2)}{S} \sim T_{p+q-2}$$

où T_{p+q-2} est une loi de Student à $p+q-2$ degrés de liberté.

Test d'égalité

On teste au risque α,

$$H_0 : \mu_1 = \mu_2 \quad \text{contre} \quad H_1 : \mu_1 \neq \mu_2.$$

On a

$$\frac{\sqrt{pq}}{\sqrt{p+q}} \frac{\left(\overline{X}_p - \overline{Y}_q\right) - (\mu_1 - \mu_2)}{S} \sim T_{p+q-2}$$

Zone de rejet :

$$ZR = \left\{ |\overline{X}_p - \overline{Y}_q| \geq t_{p+q-2,1-\alpha/2} \frac{\sqrt{p+q}}{\sqrt{pq}} S \right\} = \left\{ \frac{\sqrt{pq}}{\sqrt{p+q}} \frac{|\overline{X}_p - \overline{Y}_q|}{S} \geq t_{p+q-2,1-\alpha/2} \right\}.$$

p-valeur : Soit $\overline{x}_p$, $\overline{y}_q$ et s des réalisations :

$$p - \text{valeur} = 2 - 2F_T \left(\frac{\sqrt{pq}}{\sqrt{p+q}} \frac{|\overline{x}_p - \overline{y}_q|}{s} \right) = \mathbb{P}\left[|T| \geq \frac{\sqrt{pq}}{\sqrt{p+q}} \frac{|\overline{x}_p - \overline{y}_q|}{s} \right]$$

Les iris de Fisher

Objectif : reconnaître le type d'Iris à partir de la longueur de ses sépales. On considère ici deux espèces, et pour chaque espèce on dispose de 50 individus. Plus précisément, on note $x_1, \dots, x_{50}$ les longueurs des sépales des iris de la variété *Virginica* et $y_1, \dots, y_{50}$ celles de la variété *Versicolor*.

Valeurs numériques : On a

$$\bar{x}_{50} = 5.94 \qquad s_X = 0.52 \bar{y}_{50} = 6.59 \qquad s_Y = 0.64$$

Le test : On teste au risque $\alpha = 5\%$

$$H_0 : \mu_1 = \mu_2 \quad \text{contre} \quad H_1 : \mu_1 \neq \mu_2.$$

p-valeur : On a

$$\text{p-valeur} = \mathbb{P}\left[|T| \geq |-5.63|\right] = 2\,\mathbb{P}\left[T \geq 5.63\right] \simeq 10^{-7}.$$

Conclusion : On rejette H_0.

Test d'égalité

On teste au risque α,

$$H_0 : \mu_1 \leq \mu_2 \quad \text{contre} \quad H_1 : \mu_1 > \mu_2.$$

On a

$$\frac{\sqrt{pq}}{\sqrt{p+q}} \frac{\left(\overline{X}_p - \overline{Y}_q\right) - (\mu_1 - \mu_2)}{S} \sim T_{p+q-2}$$

Zone de rejet :

$$ZR = \left\{ \overline{X}_p - \overline{Y}_q \geq \frac{\sqrt{p+q}}{\sqrt{pq}} t_{p+q-2,1-\alpha} S \right\} = \left\{ \frac{\sqrt{pq}}{\sqrt{p+q}} \frac{\overline{X}_p - \overline{Y}_q}{S} \geq t_{p+q-2,1-\alpha} \right\}.$$

p-valeur : Soit $\overline{x}_p$, $\overline{y}_q$ et s des réalisations :

$$p\text{-valeur} = 1 - F_T\left(\frac{\sqrt{pq}}{\sqrt{p+q}} \frac{\overline{x}_p - \overline{y}_q}{s} \right) = \mathbb{P}\left[T \geq \frac{\sqrt{pq}}{\sqrt{p+q}} \frac{\overline{x}_p - \overline{y}_q}{s} \right].$$

III. Tests de Fisher et Shapiro-Wilk

Test de Fisher

Cadre du test de Student : On a du faire les hypothèses suivantes :

- les données $x_1, \ldots, x_p$ sont les réalisations de variables aléatoires $X_1, \ldots, X_p$ indépendantes et de même loi $\mathcal{N}\left(\mu_1, \sigma_1^2\right)$
- les données $y_1, \ldots, y_q$ sont les réalisations de variables aléatoires $Y_1, \ldots, Y_q$ indépendantes et de même loi $\mathcal{N}\left(\mu_2, \sigma_2^2\right)$
- les deux échantillons sont indépendants et de même variance, i.e $\sigma_1^2 = \sigma_2^2 = \sigma^2$.

Hypothèses facilement vérifiables : L'indépendance des variables est souvent dépendante du protocole expérimental.

Objectifs : Vérifier :

- X et Y suivent des lois normales,
- X et Y ont la même variance.

Test de Fisher

On veut tester au risque α

$$H_0 : \sigma_1^2 = \sigma_2^2 \quad \text{contre} \quad H_1 : \sigma_1^2 \neq \sigma_2^2.$$

Le cadre probabiliste :

- les données $x_1, \ldots, x_p$ sont les réalisations de variables aléatoires $X_1, \ldots, X_p$ indépendantes et de même loi $\mathcal{N}\left(\mu_1, \sigma_1^2\right)$
- les données $y_1, \ldots, y_q$ sont les réalisations de variables aléatoires $Y_1, \ldots, Y_q$ indépendantes et de même loi $\mathcal{N}\left(\mu_2, \sigma_2^2\right)$
- les deux échantillons sont indépendants

Construction du test

On dispose des estimateurs de σ_1^2 et σ_2^2 définis par

$$S_X^2 = \frac{1}{p-1}\sum_{i=1}^{p}\left(X_i - \overline{X}_p\right)^2 \quad \text{et} \quad S_Y^2 = \frac{1}{q-1}\sum_{i=1}^{q}\left(Y_i - \overline{Y}_q\right)^2.$$

Définition

Soient p, q deux entiers positifs et soit $Z_p \sim \chi_p^2$ et $Z_q \sim \chi_q^2$ deux variables aléatoires indépendantes. Alors

$$\frac{Z_p/p}{Z_q/q} \sim F(p,q),$$

*où $F(p,q)$ est la **loi de Fisher** à p, q degrés de liberté.*

Corollaire

On a

$$\frac{S_X^2/\sigma_1^2}{S_Y^2/\sigma_2^2} \sim F(p-1, q-1).$$

Test d'égalité

On teste au risque α,

$$H_0 : \sigma_1^2 = \sigma_2^2 \quad \text{contre} \quad H_1 : \sigma_1^2 > \sigma_2^2.$$

On a

$$\frac{S_X^2/\sigma_1^2}{S_Y^2/\sigma_2^2} \sim F(p-1, q-1).$$

Intervalle de confiance :

$$IC_{1-\alpha}\left(\frac{\sigma_1^2}{\sigma_2^2}\right) = \left[\frac{S_X^2}{S_Y^2}\frac{1}{f_{1-\alpha/2}}; \frac{S_X^2}{S_Y^2}\frac{1}{f_{\alpha/2}}\right],$$

Zone de rejet :

$$ZR = \left\{1 < \frac{S_X^2}{S_Y^2}\frac{1}{f_{1-\alpha/2}}\right\} \cup \left\{1 > \frac{S_X^2}{S_Y^2}\frac{1}{f_{\alpha/2}}\right\} = \left\{\frac{S_X^2}{S_Y^2} > f_{1-\alpha/2}\right\} \cup \left\{\frac{S_X^2}{S_Y^2} < f_{\alpha/2}\right\}.$$

p-valeur : Soit s_x et s_y des réalisations :

$$p\text{-valeur} = \min\left\{F\left(\frac{s_x^2}{s_y^2}\right), 1 - F\left(\frac{s_x^2}{s_y^2}\right)\right\}.$$

Iris de Fisher

Rappelons que l'on a

$$s_x^2 = 0.52^2 \qquad s_y^2 = 0.64^2$$

Le test : On teste au risque de 5%

$$H_0 : \sigma_1^2 = \sigma_2^2 \quad \text{contre} \quad H_1 : \sigma_1^2 \neq \sigma_2^2.$$

p-valeur : On a

$$p - \text{valeur} = \min\left\{F\left(0.66\right), 1 - F\left(0.66\right)\right\} \simeq \min\{0.79, 0.21\} = 0.21$$

Conclusion : On ne rejette donc pas H_0.

Test de Shapiro-Wilk

Cadre : On considère des réalisations $x_1, \ldots, x_n$ issues de variables aléatoires indépendantes et identiquement distribuées $X_1, \ldots, X_n$.

Test de Shapiro–Wilk :

H_0 : "la variable X suit une loi normale" contre H_1 : "la variable X ne suit pas une loi normale."

Conclusion :

- si la p-valeur est supérieure à α, on ne peut pas rejeter H_0, c'est-à-dire qu'on ne rejette pas l'hypothèse de normalité des données ;
- sinon, on rejette H_0, indiquant que les données ne sont vraisemblablement pas issues d'une loi normale.

Les iris de Fisher : On obtient pour les espèces *Versicolor* et *Virginica* des p-valeurs respectives de 0.46 et 0.26. Dans les deux cas, on ne rejette pas H_0 : les données peuvent donc être considérées comme gaussiennes.

IV. Test de Student dans le cas apparié

OBJECTIFS

Objectif :

- On considère des données $((x_1, y_1), \ldots, (x_n, y_n))$ qui sont des réalisations de couples de variables aléatoires indépendants $(X_1, Y_1), \ldots, (X_n, Y_n)$ de même loi que (X, Y).
- On souhaite comparer les moyennes des variables aléatoires X et Y.

Le cadre probabiliste :

- Les variables aléatoires X_i et Y_i ne sont pas nécessairement indépendantes.
- Leurs espérances respectives sont données par μ_1 et μ_2.
- La variable aléatoire $X - Y$ suit une loi normale d'espérance $\mu = \mu_1 - \mu_2$ et de variance σ^2.

Remarque : Dans ce cadre, on ne suppose pas que X ou Y suivent individuellement une loi normale : seule la différence $X - Y$ est supposée normale.

Construction du test

Loi de $\overline{X}_n - \overline{Y}_n$: On a

$$\overline{X}_n - \overline{Y}_n \sim \mathcal{N}\left(\mu_1 - \mu_2, \frac{\sigma^2}{n}\right)$$

ce que l'on peut réécrire

$$\sqrt{n}\frac{\overline{X}_n - \overline{Y}_n - (\mu_1 - \mu_2)}{\sigma} \sim \mathcal{N}(0,1)\,.$$

Estimateur de σ^2 : On a

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - Y_i - \left(\overline{X}_n - \overline{Y}_n\right)\right)^2.$$

Conclusion :

$$\sqrt{n}\frac{\overline{X}_n - \overline{Y}_n - (\mu_1 - \mu_2)}{S} \sim T_{n-1}$$

où T_{n-1} est une loi de Student à $n-1$ degrés de liberté.

Test d'égalité

On teste au risque α,

$$H_0 : \mu_1 = \mu_2 \quad \text{contre} \quad H_1 : \mu_1 \neq \mu_2.$$

On a

$$\sqrt{n}\frac{\overline{X}_n - \overline{Y}_n - (\mu_1 - \mu_2)}{S} \sim T_{n-1}$$

Zone de rejet :

$$ZR = \left\{ \left|\overline{X}_n - \overline{Y}_n\right| \geq t_{n-1,1-\alpha/2}\frac{S}{\sqrt{n}} \right\} = \left\{ \sqrt{n}\frac{\left|\overline{X}_n - \overline{Y}_n\right|}{S} \geq t_{n-1,1-\alpha/2} \right\}.$$

p-valeur : Soit $\overline{x}_p$, $\overline{y}_q$ et s des réalisations :

$$p - \text{valeur} = 2 - 2F_T\left(\frac{\left|\overline{x}_n - \overline{y}_n\right|}{s}\right) = \mathbb{P}\left[|T| \geq \frac{\left|\overline{x}_n - \overline{y}_n\right|}{s}\right]$$

EXEMPLE

Problème : On s'intéresse à un échantillon de 30 matières fécales, analysées à l'aide de deux méthodes différentes de spectrométrie afin d'estimer leur teneur en lutécium radioactif.

Attention ! Les mesures issues des deux méthodes étant réalisées sur les mêmes échantillons, elles ne peuvent évidemment pas être considérées comme indépendantes.

Valeurs numériques : On a

$$\overline{x}_{30} = 120.83 \quad \text{et} \quad \overline{y}_{30} = 119.33.$$

p-valeur et conclusion : On obtient une p-valeur égale à 0.66 et on ne rejette donc pas H_0.

Remarque : À noter que si l'on avait, à tort, appliqué le test de Student pour échantillons indépendants, on aurait obtenu une p-valeur égale à 0.0031, conduisant alors au rejet de H_0.

Test d'égalité

On teste au risque α,

$$H_0 : \mu_1 \leq \mu_2 \quad \text{contre} \quad H_1 : \mu_1 > \mu_2.$$

On a

$$\sqrt{n}\frac{\overline{X}_n - \overline{Y}_n - (\mu_1 - \mu_2)}{S} \sim T_{n-1}$$

Zone de rejet :

$$ZR = \left\{ \overline{X}_n - \overline{Y}_n \geq t_{n-1,\,1-\alpha}\, \frac{S}{\sqrt{n}} \right\} = \left\{ \sqrt{n}\, \frac{\overline{X}_n - \overline{Y}_n}{S} \geq t_{n-1,\,1-\alpha} \right\}.$$

p-valeur : Soit $\overline{x}_n$, $\overline{y}_n$ et s des réalisations :

$$p - \text{valeur} = 1 - F_T\left(\frac{\overline{x}_n - \overline{y}_n}{s}\right).$$

V. Tests du Chi-deux

Test d'indépendance

Objectif : Tester au risque α :

H_0 : "X et Y sont indépendantes" contre H_1 : "X et Y ne sont pas indépendantes."

Cadre du test : On dispose d'un échantillon de taille n, constitué de couples de réalisations

$$(x_1, y_1),\ (x_2, y_2),\ \ldots,\ (x_n, y_n),$$

où (X, Y) est un couple de variables aléatoires prenant leurs valeurs dans

$$\{a_1, \ldots, a_p\} \times \{b_1, \ldots, b_q\}.$$

Principe du test : Sous l'hypothèse d'indépendance, on a :

$$\mathbb{P}\left[X = a_i, Y = b_j\right] = \mathbb{P}\left[X = a_i\right] \mathbb{P}\left[Y = b_j\right], \quad \forall (i,j) \in \{1, \ldots, p\} \times \{1, \ldots, q\}.$$

Il faut donc comparer les estimations des probabilités jointes au produit des estimations des probabilités marginales.

NOTATIONS

Notations :

- $O_{i,j} = \sum_{k=1}^{n} \mathbf{1}_{\{X_k=a_i,\, Y_k=b_j\}}$, le nombre de couples (X, Y) égaux à (a_i, b_j) dans l'échantillon ;
- $O_{i,.} = \sum_{j=1}^{q} O_{i,j}$, le nombre de couples (X, Y) pour lesquels $X = a_i$;
- $O_{.,j} = \sum_{i=1}^{p} O_{i,j}$, le nombre de couples (X, Y) pour lesquels $Y = b_j$;
- $E_{i,j} = \frac{O_{i,.}\, O_{.,j}}{n}$, l'effectif attendu sous l'hypothèse d'indépendance.

Remarques :

- $O_{i,j}/n$ est un estimateur de $\mathbb{P}[X = a_i, Y = b_j]$;
- $E_{i,j}/n$ est un estimateur de $\mathbb{P}[X = a_i]\,\mathbb{P}[Y = b_j]$ sous H_0 ;

Construction du test :

Sous l'hypothèse H_0, les deux estimateurs $E_{i,j}/n$ et $O_{i,j}/n$ sont censés être proches.

On introduit la variable aléatoire :

$$Z = \sum_{i=1}^{p} \sum_{j=1}^{q} \frac{(O_{i,j} - E_{i,j})^2}{E_{i,j}}.$$

Interprétation : Z peut être vue comme une distance au carré, normalisée entre les effectifs observés et les effectifs attendus sous l'hypothèse H_0.

Proposition (Admise)

Si X et Y sont indépendantes, alors

$$Z \xrightarrow[n\to+\infty]{\mathcal{L}} \chi^2_{(p-1)(q-1)}.$$

Test du Chi-deux

On teste au risque α :

H_0 : "X et Y sont indépendantes" contre H_1 : "X et Y ne sont pas indépendantes."

Construction de la zone de rejet :

$$\mathbb{P}_{H_0}\left[Z \geq k_{1-\alpha}\right] \xrightarrow[n\to+\infty]{} 1-\alpha.$$

Zone de rejet (asymptotique) :

$$ZR = \left\{Z > k_{(p-1)(q-1),\,1-\alpha}\right\}.$$

APPLICATION

Objectifs : On étudie l'influence du sexe sur la couleur des cheveux d'un groupe d'élèves. On souhaite vérifier si la couleur des cheveux est indépendante du sexe.

Valeurs numériques :

Sexe	Blond	Roux	Châtain	Brun	Noir de Jais	Total
Garçons	592	119	849	504	36	2100
Filles	544	97	677	451	14	1783
Total	1136	216	1526	955	50	3883

Table – Répartition des couleurs de cheveux selon le sexe.

Le test : On note X la variable « couleur des cheveux » et Y la variable « sexe ». On teste au risque de 5% l'hypothèse

$$H_0 : X \text{ et } Y \text{ sont indépendantes} \quad \text{contre} \quad H_1 : X \text{ et } Y \text{ ne sont pas indépendantes.}$$

Conclusion : Le quantile d'ordre 0.95 de la loi du Chi-deux à 4 degrés de liberté est donné par 9.49. On a donc la zone de rejet

$$ZR = \{Z > 9.49\}.$$

et ici, $z_{\text{obs}} = 10.47$, ce qui conduit à rejeter l'hypothèse d'indépendance au risque de 5%.

Test d'adéquation

Cadre : On dispose de n observations $x_1, \dots, x_n$, issues de variables aléatoires indépendantes $X_1, \dots, X_n$ de même loi, prenant leurs valeurs dans l'ensemble discret

$$\{a_1, \dots, a_K\}.$$

Objectif : On note P la loi inconnue de la variable aléatoire X et l'on souhaite vérifier si $P = P_0$, où P_0 est une loi connue. Autrement dit, on teste au risque α :

$$H_0 : X \sim P_0 \quad \text{contre} \quad H_1 : X \nsim P_0.$$

Notations : On note $p_{0,1}, \dots, p_{0,K} > 0$ les probabilités définissant la loi P_0 et :

- E_k : l'effectif observé pour la modalité a_k, c'est-à-dire

$$E_k = \sum_{i=1}^{n} \mathbf{1}_{\{X_i = a_k\}};$$

- N_k : l'effectif théorique attendu pour la modalité a_k sous P_0, donné par

$$N_k = n\, p_{0,k}.$$

Construction du test :

Construction du test : E_k/n est un estimateur de $p_k = \mathbb{P}[X = a_k]$. Sous H_0, il doit converger vers $p_{0,k} = N_k/n$.

On considère la "distance" :

$$Q^2 = \sum_{k=1}^{K} \frac{(E_k - N_k)^2}{N_k}.$$

Proposition

Si la loi de X est P_0, alors

$$Q^2 \xrightarrow[n\to+\infty]{\mathcal{L}} \chi^2_{K-1}.$$

LE TEST :

On teste au risque α :

$$H_0 : X \sim P_0 \quad \text{contre} \quad H_1 : X \nsim P_0.$$

On rejette H_0 lorsque la distance Q^2 sera trop grande

$$Q^2 \geq c_\alpha$$

et on a

$$\mathbb{P}_{H_0}\left[Q^2 \geq k_{1-\alpha}\right] \xrightarrow[n\to+\infty]{} 1-\alpha,$$

Zone de rejet (asymptotique) :

$$ZR = \left\{Q^2 > k_{1-\alpha}\right\}.$$

APPLICATION

Cadre : Un biologiste pense qu'à un âge donné :

- 50% des bébés marchent,
- 12% ont une ébauche de marche,
- 38% ne marchent pas.

Données : On réalise une étude sur $n = 80$ bébés et on obtient les effectifs suivants :

- Marche : 35
- Ébauche : 4
- Ne marche pas : 41

Valeurs numériques :

	Marche	Ébauche	Ne marche pas	Total
Effectif observé (E_k)	35	4	41	80
Effectif théorique ($N_k = np_{0,k}$)	40	9.6	30.4	80
Distance au carré normalisé	0.625	3.267	3.696	7.588

Le test : On teste au risque $\alpha = 5\%$:

$$H_0 : X \sim P_0 \quad \text{contre} \quad H_1 : X \nsim P_0.$$

Conclusion : Ici, on a la zone de rejet $ZR = \{Q^2 > 5.99\}$. et $q^2_{\text{obs}} = 7.588$ et on rejette donc H_0.