

MAP 433 : Introduction aux méthodes statistiques. Cours 7

9 Octobre 2015

Aujourd'hui (et la semaine prochaine...)

1 Construction d'un test : hypothèses générales

- Retour sur un exemple
- Principe de construction

2 Tests asymptotiques

3 Tests d'adéquation

- Tests de Kolmogorov-Smirnov
- Tests du χ^2

Situation

- Situation : on part d'une expérience statistique $(\mathfrak{Z}, \mathcal{Z}, \{\mathbb{P}_\theta, \theta \in \Theta\})$ engendrée par l'observation Z .
- On souhaite tester :

$$H_0 : \theta \in \Theta_0 \subset \Theta \quad \text{contre} \quad H_1 : \theta \in \Theta_1$$

avec $\Theta_0 \cap \Theta_1 = \emptyset$.

- Si $\Theta_0 = \{\theta_0\}$ et $\Theta_1 = \{\theta_1\}$, on a Neyman-Pearson. Et sinon ?

Principe de construction

- Trouver une **statistique libre sous l'hypothèse** : toute quantité $\phi(Z)$ **observable** dont on connaît la loi sous l'hypothèse, c'est-à-dire la loi de $\phi(Z)$ sous \mathbb{P}_θ avec $\theta \in \Theta_0$.
- On regarde si le comportement de $\phi(Z)$ est **typique** d'un comportement sous l'hypothèse.
- Si oui, on **accepte** H_0 , si non on **rejette** H_0 .
- On quantifie oui/non par le niveau α du test.

Exemple : test sur la variance

- On observe $Z = (Y_1, \dots, Y_n)$,

$$Y_1, \dots, Y_n \sim_{\text{i.i.d.}} \mathcal{N}(\mu, \sigma^2)$$

avec $\theta = (\mu, \sigma^2) \in \Theta = \mathbb{R} \times (0, +\infty)$.

- **Premier cas** : on teste

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

- Sous l'hypothèse (c'est-à-dire sous \mathbb{P}_θ avec $\theta = (\mu, \sigma_0)$ et $\mu \in \mathbb{R}$ quelconque), on a

$$(n-1) \frac{s_n^2}{\sigma_0^2} \sim \chi^2(n-1)$$

avec $s_n^2 := \frac{1}{n-1} \sum_{i=1}^n (Y_i - \bar{Y}_n)^2$.

Test sur la variance (cont.)

- Donc, **sous l'hypothèse**, le comportement typique de

$$\phi(Z) = (n-1) \frac{s_n^2}{\sigma_0^2}$$

est celui d'une variable aléatoire de loi du χ^2 à $n-1$ degrés de liberté.

- Soit $q_{1-\alpha, n-1}^{\chi^2} > 0$ tel que si $U \sim \chi^2(n-1)$, alors

$$\mathbb{P}[U > q_{1-\alpha, n-1}^{\chi^2}] = \alpha.$$

- **Sous l'hypothèse** $\phi(Z) \stackrel{d}{=} U$ et donc la probabilité pour que $\phi(Z)$ dépasse $q_{1-\alpha, n-1}^{\chi^2}$ est inférieure (égale) à α (comportement atypique si α petit).

Test sur la variance (cont.)

- Règle de décision : On accepte l'hypothèse si

$$\phi(Z) \leq q_{1-\alpha, n-1}^{\chi^2}.$$

On la rejette sinon.

- Par construction, on a un test de niveau α .
- On ne sait rien dire sur l'erreur de seconde espèce, mis à part qu'elle est minimale parmi les tests de zone de rejet de la forme de $\{\phi(Z) > c\}$, $c > 0...$

Test sur la variance (fin)

- Deuxième cas : On teste

$$H_0 : \sigma^2 \leq \sigma_0^2 \quad \text{contre} \quad H_1 : \sigma^2 > \sigma_0^2.$$

- Pas de statistique libre évidente... Mais, pour $\sigma^2 \leq \sigma_0^2$, on a

$$\begin{aligned} \mathbb{P}_\sigma \left[(n-1) \frac{s_n^2}{\sigma^2} > q_{1-\alpha, n-1}^{\chi^2} \right] &= \mathbb{P}_\sigma \left[(n-1) \frac{s_n^2}{\sigma^2} > \frac{\sigma_0^2}{\sigma^2} q_{1-\alpha, n-1}^{\chi^2} \right] \\ &\leq \mathbb{P}_\sigma \left[(n-1) \frac{s_n^2}{\sigma^2} > q_{1-\alpha, n-1}^{\chi^2} \right] \\ &= \alpha. \end{aligned}$$

- La même statistique de test convient pour contrôler l'erreur de première espèce que pour l'hypothèse nulle simple. On choisit **ici** la **même** règle de décision.

Conclusion provisoire

- Pour contruire un test de l'hypothèse $H_0 : \theta \in \Theta_0$ contre $H_1 : \theta \in \Theta_1$, on cherche **une statistique libre** sous l'hypothèse et on rejette pour un seuil qui dépend de la loi de la statistique sous H_0 , de sorte de fournir une zone de rejet **maximale**.
- Le plus souvent, la statistique est obtenue via un estimateur. Sauf exception (comme la cas gaussien) une telle statistique est difficile à trouver en général.
- **Simplification** cadre asymptotique (où la gaussianité réappara"t le plus souvent...).

Le test de Wald : hypothèse nulle simple

- Situation la suite d'expériences $(\mathcal{Z}^n, \mathcal{Z}^n, \{\mathbb{P}_\theta^n, \theta \in \Theta\})$ est engendrée par l'observation $Z^n, \theta \in \Theta \subset \mathbb{R}$
- **Objectif** : Tester

$$H_0 : \theta = \theta_0 \quad \text{contre} \quad H_1 : \theta \neq \theta_0.$$

- **Hypothèse** : on dispose d'un estimateur $\hat{\theta}_n$ **asymptotiquement normal**

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} \mathcal{N}(0, v(\theta))$$

en loi sous $\mathbb{P}_\theta^n, \forall \theta \in \Theta$, où $\theta \rightsquigarrow v(\theta) > 0$ est continue.

- Sous l'hypothèse (ici sous $\mathbb{P}_{\theta_0}^n$) on a **la convergence**

$$\sqrt{n} \frac{\hat{\theta}_n - \theta_0}{\sqrt{v(\hat{\theta}_n)}} \xrightarrow{d} \mathcal{N}(0, 1)$$

en loi sous $\mathbb{P}_{\theta_0}^n$.

Test de Wald (cont.)

- Remarque $\sqrt{v(\hat{\theta}_n)} \leftrightarrow \sqrt{v(\theta_0)}$ ou d'autres choix encore...
- On a aussi

$$T_n = n \frac{(\hat{\theta}_n - \theta_0)^2}{v(\hat{\theta}_n)} \xrightarrow{d} \chi^2(1)$$

sous $\mathbb{P}_{\theta_0}^n$.

- Soit $q_{1-\alpha,1}^{\chi^2} > 0$ tel que si $U \sim \chi^2(1)$, on a $\mathbb{P}[U > q_{1-\alpha,1}^{\chi^2}] = \alpha$. On choisit la zone de rejet

$$\mathcal{R}_{n,\alpha} = \{T_n \geq q_{1-\alpha,1}^{\chi^2}\}.$$

- Le test de zone de rejet $\mathcal{R}_{n,\alpha}$ s'appelle **Test de Wald de l'hypothèse simple $\theta = \theta_0$ contre l'alternative $\theta \neq \theta_0$ basé sur $\hat{\theta}_n$.**

Propriétés du test de Wald

Proposition

Le test Wald de l'hypothèse simple $\theta = \theta_0$ contre l'alternative $\theta \neq \theta_0$ basé sur $\hat{\theta}_n$ est

- *asymptotiquement* de niveau α :

$$\mathbb{P}_{\theta_0}^n [T_n \in \mathcal{R}_{n,\alpha}] \rightarrow \alpha.$$

- *convergent ou (consistant)*. Pour tout point $\theta \neq \theta_0$

$$\mathbb{P}_{\theta}^n [T_n \notin \mathcal{R}_{n,\alpha}] \rightarrow 0.$$

Preuve

- Test asymptotiquement de niveau α **par construction**.
- Contrôle de l'erreur de seconde espèce : Soit $\theta \neq \theta_0$. On a

$$T_n = \left(\sqrt{n} \frac{\hat{\theta}_n - \theta}{\sqrt{v(\hat{\theta}_n)}} + \sqrt{n} \frac{\theta - \theta_0}{\sqrt{v(\hat{\theta}_n)}} \right)^2$$

$$=: T_{n,1} + T_{n,2}.$$

On a $T_{n,1} \xrightarrow{d} \mathcal{N}(0, 1)$ sous \mathbb{P}_θ^n et

$$T_{n,2} \xrightarrow{\mathbb{P}_\theta^n} \pm\infty \text{ car } \theta \neq \theta_0$$

Donc $T_n \xrightarrow{\mathbb{P}_\theta^n} +\infty$, d'où le résultat.

- **Remarque** : si $\theta \neq \theta_0$ mais $|\theta - \theta_0| \lesssim 1/\sqrt{n}$, le raisonnement ne s'applique pas. Résultat **non uniforme en le paramètre**.

Test de Wald : cas vectoriel

- **Même contexte:** $\Theta \subset \mathbb{R}^d$ et on dispose d'un estimateur $\hat{\theta}_n$ asymptotiquement normal :

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} \mathcal{N}(0, V(\theta))$$

où la matrice $V(\theta)$ est **définie positive** et continue en θ .

- On cherche à tester $H_0 : \theta = \theta_0$ contre $H_1 : \theta \neq \theta_1$.
- Sous \mathbb{P}_θ , la convergence $n^{1/2}(\hat{\theta}_n - \theta) \xrightarrow{d} \mathcal{N}(0, V(\theta))$ implique que

$$V^{-1/2}(\theta)n^{1/2}(\hat{\theta}_n - \theta) \xrightarrow{d} \mathcal{N}(0, \text{Id}_d)$$

et donc que

$$n(\hat{\theta}_n - \theta)^T V^{-1}(\theta)(\hat{\theta}_n - \theta) \xrightarrow{d} \chi_d^2.$$

Exemple: loi exponentielle

- **Hypothèse:** $\{X_i\}_{i=1}^n$, i.i.d. de loi exponentielle de paramètre $\theta \in \Theta = \mathbb{R}_+^*$.
- **log-vraisemblance**

$$\ell_n(\theta) = n^{-1} \sum_{i=1}^n \log f(\theta, X_i) = \log(\theta) - \theta \bar{X}_n$$

où $\bar{X}_n = n^{-1} \sum_{i=1}^n X_i$ est la moyenne empirique.

- Estimateur du MV: $\hat{\theta}_n = \bar{X}_n^{-1}$.
- **Modèle régulier**

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d, \mathbb{P}_\theta} \mathcal{N}(0, I^{-1}(\theta))$$

où $I(\theta) = \theta^{-2}$ est l'**information de Fisher**

Exemple: test loi exponentielle

- **Test de Wald** de l'hypothèse $H_0 : \theta = \theta_0$ contre l'hypothèse $H_1 : \theta \neq \theta_0$.

$$n(\hat{\theta}_n - \theta_0)^2 / I(\hat{\theta}_n) = n(1 - \theta_0 \hat{\theta}_n)^2 \xrightarrow{\mathbb{P}_{\theta_0}} \chi^2_{1,1-\alpha}$$

- **Application numérique** $n = 100$, $\theta_0 = 0.5$,

Test de Wald: cas vectoriel

- Le test de Wald de l'hypothèse $H_0 = \theta = \theta_0$ contre $H_1 = \theta \neq \theta_0$ rejette H_0 si

$$n(\hat{\theta}_n - \theta_0)^T V^{-1}(\hat{\theta}_n)(\hat{\theta}_n - \theta_0) > q_{d,1-\alpha}^{\chi^2}$$

- On peut remplacer la matrice de covariance $V(\hat{\theta}_n)$ par $V(\theta_0)$ ou tout estimateur consistant de $V(\theta_0)$.

Test de Wald : hypothèse nulle composite

- **Même contexte:** $\Theta \subset \mathbb{R}^d$ et on dispose d'un estimateur $\hat{\theta}_n$ asymptotiquement normal :

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} \mathcal{N}(0, V(\theta))$$

où la matrice $V(\theta)$ est **définie positive** et continue en θ .

- **But** Tester $H_0 : \theta \in \Theta_0$ contre $H_1 : \theta \notin \Theta_0$, où

$$\Theta_0 = \{\theta \in \Theta, g(\theta) = 0\}$$

et

$$g : \mathbb{R}^d \rightarrow \mathbb{R}^m$$

($m \leq d$) est régulière.

Test de Wald cont.

- **Hypothèse** : la différentielle (de matrice $J_g(\theta)$) de g est de rang maximal m en tout point de (l'intérieur) de Θ_0 .

Proposition

En tout point θ de l'intérieur de Θ_0 (i.e. **sous l'hypothèse**), on a, en loi sous \mathbb{P}_θ^n :

■

$$\sqrt{n}g(\hat{\theta}_n) \xrightarrow{d} \mathcal{N}(0, J_g(\theta)V(\theta)J_g(\theta)^T),$$

■

$$T_n = ng(\hat{\theta}_n)^T \Sigma_g(\hat{\theta}_n)^{-1} g(\hat{\theta}_n) \xrightarrow{d} \chi^2(m)$$

où $\Sigma_g(\theta) = J_g(\theta)V(\theta)J_g(\theta)^T$.

- Preuve : méthode delta multidimensionnelle.

Test de Wald (fin)

Proposition

Sous les hypothèses précédentes, le test de zone de rejet

$$\mathcal{R}_\alpha = \{ T_n \geq q_{1-\alpha, m}^{\chi^2} \}$$

avec $\mathbb{P} [U > q_{1-\alpha, m}^{\chi^2}] = \alpha$ si $U \sim \chi^2(m)$ est

- *Asymptotiquement de niveau α en tout point θ de (l'intérieur) de Θ_0 :*

$$\mathbb{P}_\theta^n [T_n \in \mathcal{R}_{n, \alpha}] \rightarrow \alpha.$$

- *Convergent : pour tout $\theta \notin \Theta_0$ on a*

$$\mathbb{P}_\theta^n [T_n \notin \mathcal{R}_{n, \alpha}] \rightarrow 0.$$

- C'est la même preuve qu'en dimension 1.

Modèle régulier

Definition

La famille de densités $\{f(\theta, \cdot), \theta \in \Theta\}$, par rapport à la mesure dominante μ , $\Theta \subset \mathbb{R}$, est *régulière* si

- Θ ouvert et $\{f(\theta, \cdot) > 0\} = \{f(\theta', \cdot) > 0\}$, $\forall \theta, \theta' \in \Theta$.
- μ -p.p. $\theta \rightsquigarrow f(\theta, \cdot)$, $\theta \rightsquigarrow \log f(\theta, \cdot)$ sont \mathcal{C}^2 .
- $\forall \theta \in \Theta, \exists \mathcal{V}_\theta \subset \Theta$ t.q. pour $\tilde{\theta} \in \mathcal{V}_\theta$

$$|\nabla_{\tilde{\theta}}^2 \log f(\tilde{\theta}, x)| + |\nabla_{\theta} \log f(\tilde{\theta}, x)| + (\nabla_{\theta} \log f(\tilde{\theta}, x))^2 \leq g(x)$$

où

$$\int_{\mathbb{R}} g(x) \sup_{a \in \mathcal{V}(\theta)} f(\tilde{\theta}, x) \mu(dx) < +\infty.$$

- L'information de Fisher est non-dégénérée :

$$\forall \theta \in \Theta, \mathbb{I}(\theta) > 0.$$

Test du score (Rao)

- Soit $\{X_i\}_{i=1}^n$ un n -échantillon i.i.d. associé à un modèle statistique $(\mathbb{P}_\theta, \theta \in \Theta)$ **régulier**
- Pour $\theta \in \Theta$, le **score de Fisher** est donné par

$$\eta_\theta(x) = \nabla_\theta \log f(\theta, x)$$

- **Propriétés**

- Le score de Fisher est centré sous \mathbb{P}_θ ,

$$\mathbb{E}_\theta[\eta_\theta(X)] = 0, \quad \theta \in \Theta.$$

- La covariance du score de Fisher est égale à la **matrice d'Information de Fisher**

$$I(\theta) = \mathbb{E}_\theta \left[\eta_\theta(X) \eta_\theta(X)^T \right]$$

- **Conclusion** Pour tout $\theta \in \Theta$,

$$Z_n(\theta) = n^{-1/2} \sum_{i=1}^n \eta_\theta(X_i) \xrightarrow{d, \mathbb{P}_{\theta_0}} \mathcal{N}(0, I(\theta)).$$

Test de Rao

- Pour tester $H_0 : \theta = \theta_0$ contre $H_1 : \theta \neq \theta_0$, nous considérons la statistique de test

$$Z_n(\theta_0)^T I^{-1}(\theta_0) Z_n(\theta_0)$$

- Sous l'hypothèse nulle,

$$Z_n(\theta_0)^T I^{-1}(\theta_0) Z_n(\theta_0) \xrightarrow{d}_{\mathbb{P}_{\theta_0}} \chi_d^2$$

et donc le test de Rao de rejet

$$Z_n(\theta_0)^T I^{-1}(\theta_0) Z_n(\theta_0) \geq q_{d,1-\alpha}^{\chi^2}$$

est asymptotiquement de niveau α .

Tests d'adéquation

- Situation On observe (pour simplifier) un n -échantillon de loi F inconnu

$$X_1, \dots, X_n \sim_{\text{i.i.d.}} F$$

- **Objectif** Tester

$$H_0 : F = F_0 \text{ contre } F \neq F_0$$

où F_0 distribution donnée. Par exemple : F_0 **gaussienne centrée réduite**.

- Il est **très facile de construire un test asymptotiquement de niveau α** .
Il suffit de trouver une statistique $\phi(X_1, \dots, X_n)$ de loi connue sous l'hypothèse.

Test d'adéquation : situation

- Exemples : sous l'hypothèse

$$\phi_1(X_1, \dots, X_n) = \sqrt{n}\bar{X}_n \sim \mathcal{N}(0, 1)$$

$$\phi_2(X_1, \dots, X_n) = \sqrt{n} \frac{\bar{X}_n}{s_n} \sim \text{Student}(n-1)$$

$$\phi_3(X_1, \dots, X_n) = (n-1)s_n^2 \sim \chi^2(n-1).$$

- Le problème est que ces tests **ont une faible puissance** : ils ne sont pas consistants.
- Pas exemple, si $F \neq$ gaussienne mais $\int_{\mathbb{R}} x dF(x) = 0$, $\int_{\mathbb{R}} x^2 dF(x) = 1$, alors

$$\mathbb{P}_F [\phi_1(X_1, \dots, X_n) \leq x] \rightarrow \int_{-\infty}^x e^{-u^2/2} \frac{du}{\sqrt{2\pi}}, \quad x \in \mathbb{R}.$$

(résultats analogues pour ϕ_2 et ϕ_3).

- La statistique de test ϕ_i **ne caractérise pas** la loi F_0 .

Test de Kolmogorov-Smirnov

- Rappel Si la fonction de répartition F est continue,

$$\sqrt{n} \sup_{x \in \mathbb{R}} |\hat{F}_n(x) - F(x)| \xrightarrow{d} \mathbb{B}$$

où la loi de \mathbb{B} ne dépend pas de F .

Proposition (Test de Kolmogorov-Smirnov)

Soit $q_{1-\alpha}^{\mathbb{B}}$ tel que $\mathbb{P} [\mathbb{B} > q_{1-\alpha}^{\mathbb{B}}] = \alpha$. Le test défini par la zone de rejet

$$\mathcal{R}_{n,\alpha} = \left\{ \sqrt{n} \sup_{x \in \mathbb{R}} |\hat{F}_n(x) - F_0(x)| \geq q_{1-\alpha}^{\mathbb{B}} \right\}$$

est *asymptotiquement de niveau α* : $\mathbb{P}_{F_0} [\hat{F}_n \in \mathcal{R}_{n,\alpha}] \rightarrow \alpha$ et *consistant* :

$$\forall F \neq F_0 : \mathbb{P}_F [\hat{F}_n \notin \mathcal{R}_{n,\alpha}] \rightarrow 0.$$

Test du Chi-deux

- X variables **qualitative** : $X \in \{1, \dots, d\}$.

$$\mathbb{P}[X = \ell] = p_\ell, \ell = 1, \dots, d.$$

- La loi de X est caractérisée par $\mathbf{p} = (p_1, \dots, p_d)^T$.

- Notation

$$\mathcal{M}_d = \left\{ \mathbf{p} = (p_1, \dots, p_d)^T, \ 0 \leq p_\ell, \sum_{\ell=1}^d p_\ell = 1 \right\}.$$

- **Objectif** $\mathbf{q} \in \mathcal{M}_d$ donnée. A partir d'un n -échantillon

$$X_1, \dots, X_n \sim_{\text{i.i.d.}} \mathbf{p},$$

tester $H_0 : \mathbf{p} = \mathbf{q}$ contre $H_1 : \mathbf{p} \neq \mathbf{q}$.

Construction naturelle d'un test

■ Comparaison des fréquences empiriques

$$\hat{p}_{n,\ell} = \frac{1}{n} \sum_{i=1}^n 1_{X_i=\ell} \quad \text{proche de } q_\ell, \quad \ell = 1, \dots, d ?$$

■ Loi des grands nombres :

$$(\hat{p}_{n,1}, \dots, \hat{p}_{n,d}) \xrightarrow{\mathbb{P}_{\mathbf{p}}} (p_1, \dots, p_d) = \mathbf{p}.$$

■ Théorème central-limite ?

$$\mathbf{U}_n(\mathbf{p}) = \sqrt{n} \left(\frac{\hat{p}_{n,1} - p_1}{\sqrt{p_1}}, \dots, \frac{\hat{p}_{n,d} - p_d}{\sqrt{p_d}} \right) \xrightarrow{d} ?$$

■ Composante par composante oui. Convergence globale plus délicate.

Statistique du Chi-deux

Proposition

Si les composantes de \mathbf{p} sont toutes non-nulles

- On a la *convergence en loi* sous $\mathbb{P}_{\mathbf{p}}$

$$\mathbf{U}_n(\mathbf{p}) \xrightarrow{d} \mathcal{N}(0, V(\mathbf{p}))$$

avec $V(\mathbf{p}) = \text{Id}_d - \sqrt{\mathbf{p}}(\sqrt{\mathbf{p}})^T$ et $\sqrt{\mathbf{p}} = (\sqrt{p_1}, \dots, \sqrt{p_d})^T$.

- *De plus*

$$\|\mathbf{U}_n(\mathbf{p})\|^2 = n \sum_{\ell=1}^d \frac{(\hat{p}_{n,\ell} - p_\ell)^2}{p_\ell} \xrightarrow{d} \chi^2(d-1).$$

Preuve de la normalité asymptotique

- Pour $i = 1, \dots, n$ et $1 \leq \ell \leq d$, on pose

$$Y_{\ell}^i = \frac{1}{\sqrt{p_{\ell}}} (1_{\{X_i = \ell\}} - p_{\ell}).$$

- Les vecteurs $\mathbf{Y}_i = (Y_1^i, \dots, Y_d^i)$ sont **indépendants et identiquement distribués** et

$$\mathbf{U}_n(\mathbf{p}) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \mathbf{Y}_i,$$

$$\mathbb{E}[Y_{\ell}^i] = 0, \mathbb{E}[(Y_{\ell}^i)^2] = 1 - p_{\ell}, \mathbb{E}[Y_{\ell}^i Y_{\ell'}^i] = -(p_{\ell} p_{\ell'})^{1/2}.$$

- On applique le **TCL vectoriel**.

Convergence de la norme au carré

- On a donc $\mathbf{U}_n(\mathbf{p}) \xrightarrow{d} \mathcal{N}(0, V(\mathbf{p}))$.
- On a aussi

$$\begin{aligned} \|\mathbf{U}_n(\mathbf{p})\|^2 &\xrightarrow{d} \|\mathcal{N}(0, V(\mathbf{p}))\|^2 \\ &\sim \chi^2(\text{Rang}(V(\mathbf{p}))) \end{aligned}$$

par **Cochran** : $V(\mathbf{p}) = \text{Id}_d - \sqrt{\mathbf{p}}(\sqrt{\mathbf{p}})^T$ est la projection orthogonale sur $\text{vect}\{\sqrt{\mathbf{p}}\}^\perp$ qui est de dimension $d - 1$.

Test d'adéquation du χ^2

- distance du χ^2 :

$$\chi^2(\mathbf{p}, \mathbf{q}) = \sum_{\ell=1}^d \frac{(p_{\ell} - q_{\ell})^2}{q_{\ell}}.$$

- Avec ces notations $\|\mathbf{U}_n(\mathbf{p})\|^2 = n\chi^2(\hat{\mathbf{p}}_n, \mathbf{p})$.

Proposition

Pour $\mathbf{q} \in \mathcal{M}_d$ le test simple défini par la zone de rejet

$$\mathcal{R}_{n,\alpha} = \{n\chi^2(\hat{\mathbf{p}}_n, \mathbf{q}) \geq q_{1-\alpha, d-1}^{\chi^2}\}$$

où $\mathbb{P}[U > q_{1-\alpha, d-1}^{\chi^2}] = \alpha$ si $U \sim \chi^2(d-1)$ est *asymptotiquement de niveau α et consistant* pour tester

$$H_0 : \mathbf{p} = \mathbf{q} \quad \text{contre} \quad H_1 : \mathbf{p} \neq \mathbf{q}.$$

Exemple de mise en oeuvre : expérience de Mendel

- Soit $d = 4$ et

$$\mathbf{q} = \left(\frac{9}{16}, \frac{3}{16}, \frac{3}{16}, \frac{1}{16} \right).$$

- Répartition observée : $n = 556$

$$\hat{\mathbf{p}}_{556} = \frac{1}{556} (315, 101, 108, 32).$$

- Calcul de la statistique du χ^2

$$556 \times \chi^2(\hat{\mathbf{p}}_{556}, \mathbf{q}) = 0,47.$$

- On a $q_{95\%,3} = 0,7815$.
- **Conclusion** : Puisque $0,47 < 0,7815$, on accepte l'hypothèse $\mathbf{p} = \mathbf{q}$ au niveau $\alpha = 5\%$.