MAP 433 : Introduction aux méthodes statistiques. Cours 7

9 Octobre 2015

Aujourd'hui

1. Construction d'un test : hypothèses générales
   - Retour sur un exemple
   - Principe de construction
2. Tests asymptotiques
   - Elements de la théorie asymptotique des tests
   - Tests de Wald
   - Test de Rao
3. Tests d'adéquation
   - Tests de Kolmogorov-Smirnov
   - Tests du χ^2

Situation

- Situation : on part d'une expérience statistique $\big(\mathrm{X}, \mathcal{X}, \{\mathbb{P}_\theta, \theta \in \Theta\}\big)$ 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 connait la loi sous l'hypothèse, c'est-à-dire la loi de $\phi(Z)$ sous $\mathbb{P}_\theta$ 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 $Z = (Y_1, \ldots, Y_n)$,

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

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 $\mathbb{P}_\theta$ avec $\theta = (\mu, \sigma_0)$ et $\mu \in \mathbb{R}$ quelconque), on a

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

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

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

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

- 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

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

  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

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

- La même statistique de test convient pour contrôler l'erreur de première espèce que pour l'hypohèse 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...).

Quelques définitions

- Soit $(\mathbb{P}_\theta, \theta \in \Theta)$ une famille de probabilités sur $(\mathrm{X}, \mathcal{X})$ admettant des densités $\{f(\theta, x), \theta \in \Theta\}$ par rapport à une mesure de domination μ.
- Supposons que nous disposions d'un n-échantillon $(X_1, X_2, \ldots, X_n)$ de ce modèle statistique.
- Considérons le problème de tester l'hypothèse de base $H_0 : \theta \in \Theta_0$ contre l'alternative $H_1 : \theta \in \Theta_1$, où $\Theta_0 \cap \Theta_1 = \emptyset$ et $\Theta_0 \cup \Theta_1 = \Theta$.
- Un test pour un échantillon de taille n est une fonction mesurable

$$\varphi_n : \mathrm{X}^n \to [0, 1] \; .$$

- Si le test est non randomisé $\varphi_n \in \{0, 1\}$, l'ensemble

$$\{(x_1, \ldots, x_n) \in \mathrm{X}^n, \varphi_n(x_1, \ldots, x_n) = 1\}$$

est appelée la région critique du test.

Tests asymptotiques

- On dit qu'une suite de tests $\{\varphi_n, n \in \mathbb{N}\}$ est asymptotiquement de niveau α pour $\alpha \in [0,1]$ si

$$\lim_{n\to\infty} \mathbb{E}_\theta^n[\varphi_n(X_1,\ldots,X_n)] \leq \alpha \,, \text{pour tout } \theta \in \Theta_0$$

- La puissance de ce test est la fonction

$$\theta \mapsto \pi_n(\theta) = \mathbb{E}_\theta^n[\varphi_n(X_1,\ldots,X_n)]$$

- Un test q'une suite de tests $\{\varphi_n, n \in \mathbb{N}\}$ est asymptotiquement consistante si, pour tout $\theta \in \Theta_1$,

$$\lim_{n\to\infty} \pi_n(\theta) = 1 \,.$$

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* $\tilde{\theta} \in \mathcal{V}_\theta$

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

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 :*

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

Consistance du test de Neyman-Pearson

- Supposons que $\Theta = \{\theta_0, \theta_1\}$ avec $\theta_0 \neq \theta_1$ et que l'on cherche à tester $H_0 : \theta = \theta_0$ contre $H_1 : \theta = \theta_1$.
- Le lemme de Neyman-Pearson montre que le test qui rejette H_0 si
$$\frac{\prod_{i=1}^n f(\theta_1, X_i)}{\prod_{i=1}^n f(\theta_0, X_i)} \geq c_{n,\alpha}$$
est U.P.P.
- De façon équivalente, en prenant le logarithme de chaque membre de l'identité, le test de N.P. rejette H_0 si
$$\Lambda_n(\theta_0, \theta_1) = \sum_{i=1}^n \{\ell(X_i, \theta_1) - \ell(X_i, \theta_0)\} > k_{n,\alpha}$$
où $\ell(x; \theta) = \log f(\theta, x)$ et $k_{n,\alpha}$ est choisi de telle sorte que
$$\mathbb{P}^n_{\theta_0}[\Lambda_n(\theta_0, \theta_1) > k_{n,\alpha}] = \alpha$$
(on suppose qu'une telle valeur existe, autrement il faudrait randomiser)

Calcul asymptotique du seuil critique

- En pratique, il est souvent difficile de déterminer exactement le seuil critique $k_{n,\alpha}$... mais il est souvent facile de déterminer une suite $\{k_{n,\alpha}, n \in \mathbb{N}\}$ telle que

$$\lim_{n\to\infty} \mathbb{P}^n_{\theta_0}(\Lambda_n(\theta_0,\theta_1) > k_{n,\alpha}) = \alpha.$$

- En effet, le théorème central limite montre que, sous H_0,

$$n^{-1/2}\sum_{k=1}^{n}\{\ell(X_i,\theta_1) - \ell(X_i,\theta_0) + \mathrm{KL}(\theta_0,\theta_1)\} \xrightarrow{d}_{\mathbb{P}^n_{\theta_0}} \mathcal{N}(0, J(\theta_0,\theta_1))$$

où $\mathrm{KL}(\theta_0,\theta_1)$ est la divergence de Kullback-Leibler définie par

$$\mathrm{KL}(\theta_0,\theta_1) = \mathbb{E}_{\theta_0}\left[\ell(X_1;\theta_0) - \ell(X_1;\theta_1)\right] > 0$$

et

$$J(\theta_0,\theta_1) = \mathrm{Var}_{\theta_0}[\ell(X_1;\theta_1) - \ell(X_1;\theta_0)]\,.$$

Calcul asymptotique du seuil critique

- Pour $\alpha \in (0,1)$, on note $z_{1-\alpha}$ le quantile $1-\alpha$ de la loi gaussienne standardisée.
- Nous avons donc:

$$\lim_{n\to\infty} \mathbb{P}^n_{\theta_0}\left(n^{-1/2}J^{-1}(\theta_0,\theta_1)\{\Lambda_n + n\mathrm{KL}(\theta_0,\theta_1)\} \geq z_{1-\alpha}\right) = \alpha\,.$$

ce qui implique, en posant

$$k_{n,\alpha} = -n\mathrm{KL}(\theta_0,\theta_1) + n^{1/2}z_{1-\alpha}J(\theta_0,\theta_1)$$

que le test de région critique $\{\Lambda_n > k_{n,\alpha}\}$ est asymptotiquement de niveau α,

$$\lim_{n\to\infty} \mathbb{P}^n_{\theta_0}[\Lambda_n \geq k_{n,\alpha}] = 1-\alpha\,.$$

Distribution du test sous l'hypothèse alternative

- Sous $\mathbb{P}^n_{\theta_1}$, nous avons

$$\Delta_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \{\ell(X_i;\theta_1) - \ell(X_i;\theta_0) - \mathrm{KL}(\theta_1,\theta_0)\} \xrightarrow{d} \mathbb{P}^n_{\theta_1} \mathcal{N}(0, J(\theta_1,\theta_0))$$

où

$$\mathrm{KL}(\theta_1,\theta_0) = \mathbb{E}_{\theta_1}[\ell(X_1;\theta_1) - \ell(X1_;\theta_0)]$$
$$J(\theta_1,\theta_0) = \mathrm{Var}_{\theta_1}(\ell(X_1;\theta_1) - \ell(X1_;\theta_0))$$

Distribution du test sous l'hypothèse alternative

- Sous $\mathbb{P}^n_{\theta_1}$, nous avons

$$\Delta_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \{\ell(X_i;\theta_1) - \ell(X_i;\theta_0) - \mathrm{KL}(\theta_1,\theta_0)\} \xrightarrow{d} \mathbb{P}^n_{\theta_1} \mathcal{N}(0, J(\theta_1,\theta_0))$$

où

$$\mathrm{KL}(\theta_1,\theta_0) = \mathbb{E}_{\theta_1}[\ell(X_1;\theta_1) - \ell(X1_;\theta_0)]$$
$$J(\theta_1,\theta_0) = \mathrm{Var}_{\theta_1}(\ell(X_1;\theta_1) - \ell(X1_;\theta_0))$$

- Par conséquent

$$\{\Lambda_n > k_{n,\alpha}\}$$
$$= \left\{\Delta_n > J^{-1/2}(\theta_1,\theta_0)\{z_{1-\alpha} J(\theta_0,\theta_1) - n^{1/2} I(\theta_0,\theta_1)\right\}$$

où

$$I(\theta_0,\theta_1) = \mathrm{KL}(\theta_0,\theta_1) + \mathrm{KL}(\theta_1,\theta_0) \, .$$

Puissance du test de NP

- La puissance du test est donc

$$\pi_n(\theta_1) = \Phi\left(J^{-1/2}(\theta_1, \theta_0)\left\{n^{1/2} I(\theta_0, \theta_1) - z_{1-\alpha} J(\theta_0, \theta_1)\right\}\right)$$

ce qui implique que, dès que $\mathrm{KL}(\theta_0, \theta_1) \neq 0$

$$\lim_{n \to \infty} \pi_n(\theta_1) = 1\,.$$

- Si le modèle est identifiable, alors il existe un test de niveau asymptotique α et donc la puissance tend vers 1.

Le test de Wald : hypothèse nulle simple

- Situation la suite d'expériences $\left(X^n, \mathcal{X}^{\otimes n}, \{\mathbb{P}_\theta^n, \theta \in \Theta\}\right)$ est engendrée par l'observation $Z^n = (X_1, \ldots, X_n)$, $\theta \in \Theta \subset \mathbb{R}$
- Objectif : Tester

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

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

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

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

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

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

Test de Wald (cont.)

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

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

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

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

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

- 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 $\widehat{\theta}_n$.

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 $\widehat{\theta}_n$ est

- *asymptotiquement de niveau α :*

$$\mathbb{P}^n_{\theta_0}\left[T_n \in \mathcal{R}_{n,\alpha}\right] \to \alpha.$$

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

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

Preuve

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

$$T_n = \left(\sqrt{n}\frac{\widehat{\theta}_n - \theta}{\sqrt{v(\widehat{\theta}_n)}} + \sqrt{n}\frac{\theta - \theta_0}{\sqrt{v(\widehat{\theta}_n)}}\right)^2$$
$$=: T_{n,1} + T_{n,2}.$$

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

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

  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 $\widehat{\theta}_n$ asymptotiquement normal :

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

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

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

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

et donc que

$$n(\widehat{\theta}_n - \theta)^T V^{-1}(\theta)(\widehat{\theta}_n - \theta) \xrightarrow{d} \chi^2_d\,.$$

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

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

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 $H_1 : \theta \neq \theta_0$.

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

- 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

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

- On peut remplacer la matrice de covariance $V(\widehat{\theta}_n)$ 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 $\widehat{\theta}_n$ asymptotiquement normal :

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

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

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

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 $\mathbb{P}_\theta^n$:

- $$\sqrt{n}g(\widehat{\theta}_n) \xrightarrow{d} \mathcal{N}\big(0, J_g(\theta)V(\theta)J_g(\theta)^T\big),$$
- $$T_n = ng(\widehat{\theta}_n)^T\Sigma_g(\widehat{\theta}_n)^{-1}g(\widehat{\theta}_n) \xrightarrow{d} \chi^2(m)$$

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

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

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

- *Asymptotiquement de niveau* α *en tout point* θ *de (l'intérieur) de* Θ_0 *:*
$$\mathbb{P}_\theta^n \left[T_n \in \mathcal{R}_{n,\alpha} \right] \to \alpha.$$
- *Convergent : pour tout* $\theta \notin \Theta_0$ *on a*
$$\mathbb{P}_\theta^n \left[T_n \notin \mathcal{R}_{n,\alpha} \right] \to 0.$$

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

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 $\mathbb{P}_\theta$,

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

  - 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,
$$Z_n(\theta_0)^T I^{-1}(\theta_0) Z_n(\theta_0) \xrightarrow{d}_{\mathbb{P}_{\theta_0}} \chi^2_d$$
et donc le test de Rao de rejet
$$Z_n(\theta_0)^T I^{-1}(\theta_0) Z_n(\theta_0) \geq q^{\chi^2}_{d,1-\alpha}$$
est asymptotiquement de niveau α.

Tests d'adéquation

- Situation On observe (pour simplifier) un n-échantillon de loi F inconnu
$$X_1, \ldots, X_n \sim_{\text{i.i.d.}} F$$
- 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 $\phi(X_1, \ldots, X_n)$ de loi connue sous l'hypothèse.

Test d'adéquation : situation

- Exemples : sous l'hypothèse

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

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

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

- 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

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

(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,

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

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

Proposition (Test de Kolmogorov-Smirnov)

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

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

est asymptotiquement de niveau α : $\mathbb{P}_{F_0}\left[\widehat{F}_n \in \mathcal{R}_{n,\alpha}\right] \to \alpha$ et consistant :

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

Test du Chi-deux

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

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

- La loi de X est caratérisée par $\boldsymbol{p} = (p_1, \ldots, p_d)^T$.
- Notation

$$\mathcal{M}_d = \big\{\boldsymbol{p} = (p_1, \ldots, p_d)^T,\ \ 0 \leq p_\ell, \sum_{\ell=1}^{d} p_\ell = 1\big\}.$$

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

$$X_1, \ldots, X_n \sim_{\text{i.i.d.}} \boldsymbol{p},$$

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

Construction naturelle d'un test

- Comparaison des fréquences empiriques

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

- Loi des grands nombres :

$$\left(\widehat{p}_{n,1}, \ldots, \widehat{p}_{n,d}\right) \xrightarrow{\mathbb{P}_{\boldsymbol{p}}} (p_1, \ldots, p_d) = \boldsymbol{p}.$$

- Théorème central-limite ?

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

- Composante par composante oui. Convergence globale plus délicate.

Statistique du Chi-deux

Proposition

Si les composantes de $\boldsymbol{p}$ sont toute non-nulles

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

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

avec $V(\boldsymbol{p}) = \mathrm{Id}_d - \sqrt{\boldsymbol{p}}\big(\sqrt{\boldsymbol{p}}\big)^T$ *et* $\sqrt{\boldsymbol{p}} = \big(\sqrt{p_1}, \ldots, \sqrt{p_d}\big)^T$.

- *De plus*

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

Preuve de la normalité asymptotique

- Pour $i = 1, \ldots, n$ et $1 \leq \ell \leq d$, on pose
$$Y_\ell^i = \frac{1}{\sqrt{p_\ell}}\big(1_{\{X_i=\ell\}} - p_\ell\big).$$
- Les vecteurs $\boldsymbol{Y}_i = (Y_1^i, \ldots, Y_d^i)$ sont indépendants et identiquement distribués et
$$\boldsymbol{U}_n(\boldsymbol{p}) = \frac{1}{\sqrt{n}}\sum_{i=1}^n \boldsymbol{Y}_i,$$
$$\mathbb{E}\left[Y_\ell^i\right] = 0,\ \mathbb{E}\left[(Y_\ell^i)^2\right] = 1 - p_\ell,\ \mathbb{E}\left[Y_\ell^i Y_{\ell'}^i\right] = -(p_\ell p_{\ell'})^{1/2}.$$
- On applique le TCL vectoriel.

Convergence de la norme au carré

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

$$\|\boldsymbol{U}_n(\boldsymbol{p})\|^2 \xrightarrow{d} \|\mathcal{N}\big(0, V(\boldsymbol{p})\big)\|^2 \\ \sim \chi^2\big(\mathrm{Rang}\big(V(\boldsymbol{p})\big)\big)$$

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

Test d'adéquation du χ^2

- distance du χ^2:

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

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

Proposition

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

$$\mathcal{R}_{n,\alpha} = \left\{ n\chi^2(\widehat{\boldsymbol{p}}_n, \boldsymbol{q}) \geq q^{\chi^2}_{1-\alpha,d-1} \right\}$$

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

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

Exemple de mise en oeuvre : expérience de Mendel

- Soit $d = 4$ et
$$\boldsymbol{q} = \left(\frac{9}{16}, \frac{3}{16}, \frac{3}{16}, \frac{1}{16}\right).$$
- Répartition observée : $n = 556$
$$\widehat{\boldsymbol{p}}_{556} = \frac{1}{556}(315, 101, 108, 32).$$
- Calcul de la statistique du χ^2
$$556 \times \chi^2(\widehat{\boldsymbol{p}}_{556}, \boldsymbol{q}) = 0,47.$$
- On a $q_{95\%,3} = 0,7815$.
- Conclusion : Puisque $0,47 < 0,7815$, on accepte l'hypothèse $\boldsymbol{p} = \boldsymbol{q}$ au niveau $\alpha = 5\%$.