MAP 433 : Introduction aux méthodes statistiques. Cours 7

9 Octobre 2015

Aujourd'hui

1. Tests asymptotiques
   - Elements de la théorie asymptotique des tests
   - Efficacité asymptotique relative
2. Quelques tests asymptotiques
   - Test du rapport de vraisemblance
   - Tests de Wald
   - Test de Rao
3. Tests d'adéquation
   - Tests de Kolmogorov-Smirnov
   - Tests du χ^2

Quelques définitions

- Soit $(\mathbb{P}_\theta, \theta \in \Theta)$ une famille de probabilités sur $(\mathsf{X}, \mathcal{X})$ 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

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

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

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

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

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

- On dit qu'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* $\mathcal{C}^2$.
- *Pour tout* $\theta \in \Theta$, *il existe un voisinage* $\mathcal{V}_\theta \subset \Theta$ *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

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

- *L'information de Fisher est non-dégénérée : pour tout* $\theta \in \Theta$,

$$, \quad \mathbb{I}(\theta) > 0.$$

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

$$\Lambda_n(\theta_0, \theta_1) = \sum_{i=1}^n \{\ell(X_i; \theta_1) - \ell(X_i; \theta_0)\} \geq k_{n,\alpha}$$

  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

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

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

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

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:

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

ce qui implique, en posant

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

que le test de région critique $\{\Lambda_n(\theta_0,\theta_1) \geq k_{n,\alpha}\}$ est asymptotiquement de niveau α,

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

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

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

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

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

- Par conséquent

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

où

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

Puissance du test de NP

- Conclusion Si $\mathrm{KL}(\theta_0, \theta_1) \neq 0$ alors

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

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

Efficacité asymptotique... à travers un exemple

- Supposons que $(X_1, \ldots, X_n)$ est un n-échantillon indépendant de densité $f(\theta, x) = f(x - \theta)$ par rapport à la mesure de Lebesgue sur $\mathbb{R}$.
- Hypothèses
  - Variance finie: $\int |x|^2 f(x) \mathrm{d}x < \infty$
  - Parité: f est une fonction paire (donc θ est la moyenne et la médiane de la loi)
  - f est continue et $f(0) > 0$: unicité de la médiane
- On note F la cdf associée à la densité f
- On cherche à tester $\theta = 0$ contre $H_1 : \theta > 0$.

Un exemple

- On considère deux statistiques de tests:

$$U_n = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{X_i>0\}} \qquad \text{test du signe}$$

$$T_n = \frac{\bar{X}_n}{S_n} \qquad \text{t-test}$$

  où
  - $\bar{X}_n = n^{-1}\sum_{i=1}^n X_i$ est la moyenne empirique
  - $S_n^2 = n^{-1}\sum_{i=1}^n (X_i - \bar{X}_n)^2$ est la variance empirique.
- Question: quel test est le meilleur ?

Asymptotique du test du signe

$$U_n = n^{-1} \sum_{i=1}^{n} \mathbb{1}_{\{X_i > 0\}}$$

- Par le théorème de la limite centrale

$$n^{1/2}\sigma^{-1}(\theta)(U_n - \mu(\theta)) \xrightarrow{d}_{\mathbb{P}_\theta^n} \mathcal{N}(0,1)$$

  où

$$\mu(\theta) = 1 - F(-\theta) \qquad \sigma^2(\theta) = (1 - F(-\theta))F(-\theta)\,.$$

- Par conséquent, sous $H_0 : \theta = 0$

$$2\sqrt{n}(U_n - 1/2) \xrightarrow{d}_{\mathbb{P}_0^n} \mathcal{N}(0,1)\,.$$

- Le test de région critique $\{2\sqrt{n}(U_n - 1/2) > z_{1-\alpha}\}$ est un test de niveau asymptotique α.

Puissance du test de signe

La puissance du test de signe de niveau asymptotique α est donnée pour tout $\theta > 0$, par

$$\begin{aligned}\pi_{n,\alpha}^{\mathrm{sign}}(\theta) &= \mathbb{P}_\theta(\sqrt{n}(U_n - \mu(0)) > \sigma(0)z_{1-\alpha}) \\ &= \mathbb{P}_\theta(\sqrt{n}\sigma^{-1}(\theta)(U_n - \mu(\theta)) > \sigma^{-1}(\theta)\{\sigma(0)z_{1-\alpha} + n^{1/2}\{\mu(0) - \mu(\theta)\}\})\end{aligned}$$

Le test est consistant: pour tout $\theta > 0$,

$$\lim_{n\to\infty} \pi_{n,\alpha}^{\mathrm{sign}}(\theta) = 1\,.$$

Asymptotique du t-test

On considère la moyenne empirique studentisée

$$T_n = \frac{1}{n}\sum_{i=1}^{n}\frac{X_i}{S_n} \quad \text{t-test}$$

où $S_n^2 = n^{-1}\sum_{i=1}^n (X_i - \bar{X}_n)^2$ est la variance empirique

- Loi des grands nombres : $S_n^2 \xrightarrow{\mathbb{P}}_{\mathbb{P}_\theta} s^2 = \int x^2 f(x)\mathrm{d}x$.
- Théorème Central limite: $n^{-1/2}\sum_{i=1}^n (X_i - \theta) \xrightarrow{d}_{\mathbb{P}_\theta} \mathcal{N}(0, s^2)$.
- Slutsky: $n^{-1/2}S_n^{-1}\left\{\sum_{i=1}^n (X_i - \theta)\right\} \xrightarrow{d}_{\mathbb{P}_\theta} \mathcal{N}(0,1)$.
- Le test de région critique $\{T_n > n^{-1/2}z_{1-\alpha}\}$ est un test de niveau asymptotique α:

$$\begin{aligned}\lim_{n\to\infty}\mathbb{P}_0\left(T_n > n^{-1/2}z_{1-\alpha}\right) &= \lim_{n\to\infty}\mathbb{P}_0\left(n^{1/2}T_n > z_{1-\alpha}\right)\\ &= 1 - \Phi(z_{1-\alpha}) = \alpha\,.\end{aligned}$$

Puissance du t-test

La puissance du test de niveau α est donnée par

$$\begin{aligned}\pi_{n,\alpha}^{\mathrm{t-test}}(\theta) &= \mathbb{P}_\theta(\sqrt{n}T_n > z_{1-\alpha}) \\ &= \mathbb{P}_\theta\Big(n^{-1/2}S_n^{-1}\sum_{i=1}^n(X_i-\theta) > z_{1-\alpha} - \sqrt{n}S_n^{-1}\theta\Big).\end{aligned}$$

Comme $\lim_{n\to\infty}\{z_{1-\alpha} - \sqrt{n}S_n^{-1}\theta\} = -\infty$, $\mathbb{P}_\theta$-p.s. pour tout $\theta > 0$, le t-test de niveau asymptotique α est consistant: pour tout $\theta > 0$,

$$\lim_{n\to\infty}\pi_{n,\alpha}^{\mathrm{t-test}}(\theta) = 1\,.$$

Comparaison asymptotique des puissances

- Nous devons rendre la discrimination entre l'hypothèse nulle H_0 et l'hypothèse alternative H_1 plus difficile quand $n \to \infty$.
- Idée: considérer un test $H_0 : \theta = 0$ contre une suite d'hypothèses alternatives $H_1^n : \theta = \theta_n$ avec $\theta_n > 0$ et $\lim_{n\to\infty} \theta_n = 0$.

Test du signe

$$\pi_{n,\alpha}^{\text{sign}}(\theta) = 1 - \Phi\left(\frac{\sigma(0) z_{1-\alpha} + n^{1/2}\{\mu(0) - \mu(\theta)\}}{\sigma(\theta)}\right) + o(1)$$

- la puissance du test contre la suite de contre-alternatives $H_1^n : \theta_n > 0$, dépend de $\sqrt{n}(\mu(0) - \mu(\theta_n))$ où
$$\mu(\theta) = 1 - F(-\theta)\,.$$
- Comme F est différentiable en $\theta = 0$, nous avons
$$\sqrt{n}(\mu(\theta_n) - \mu(0)) = -\sqrt{n}(F(-\theta_n) - F(0)) = \sqrt{n}\theta_n f(0) + o(\sqrt{n}\theta_n)\,.$$

Test du signe

- Si $\sqrt{n}\theta_n \to 0$, alors $\sqrt{n}(\mu(0)-\mu(\theta_n)) \to 0$; donc, $\pi_{n,\alpha}^{\mathrm{sign}}(\theta_n) \to \alpha$, on ne distingue pas l'hypothèse de base et l'alternative.
- Si $\sqrt{n}\theta_n \to \infty$, alors $\sqrt{n}(\mu(0)-\mu(\theta_n)) \to -\infty$: $\pi_{n,\alpha}^{\mathrm{sign}}(\theta_n) \to 1$, problème trop facile, la puissance tend vers 1.
- Cas intéressant !

$$\lim_{n\to\infty} \sqrt{n}\theta_n = h > 0$$

- Dans ce cas, $\sqrt{n}(\mu(0)-\mu(\theta_n)) \to -hf(0)$ et $\lim_n \sigma(\theta_n) = \sigma(0) = 1/2$, donc

$$\lim_{n\to\infty} \pi_{n,\alpha}^{\mathrm{sign}}(\theta_n) = \Phi(2hf(0) - z_{1-\alpha})$$

Efficacité asymptotique des tests

- Ceci conduit à une approche naturelle de comparaison des tests, qui consistent à comparer la puissance locale des tests

$$\pi(h) = \lim_{n\to\infty} \pi_n(h/\sqrt{n})\,.$$

- Pour les modèles réguliers, cette fonction de puissance locale asymptotique est bien définie (preuve délicate en toute généralité)

Efficacité asymptotique locale des tests

Théorème

Soit $\{\theta_n,\ n \in \mathbb{N}\} \subset \mathbb{R}_*^+$ *telle que* $\lim_{n\to\infty} \sqrt{n}\theta_n = h$. *Soit* $\{T_n,\ n \in \mathbb{N}\}$ *une suite de statistiques vérifiant:*

1. $\sqrt{n}\sigma(\theta_n)^{-1}(T_n - \mu(\theta_n)) \xrightarrow{d}_{\mathbb{P}_{\theta_n}} \mathcal{N}(0,1)$
2. μ *est différentiable en* 0
3. σ *continue en* 0.

Soit φ_n *une suite de tests simples de région critique* $\{T_n > t_{n,\alpha}\}$ *de niveau asymptotique* α*:* $\lim_{n\to\infty} \mathbb{P}_0(T_n > t_{n,\alpha}) = \alpha$.

La puissance locale asymptotique de cette suite de tests est donnée par

$$\pi(h) = \lim_{n\to\infty} \pi_n(\sqrt{n}\theta_n) = \Phi(h\mu'(0)/\sigma(0) - z_{1-\alpha}).$$

Conclusion

- Nous disposons maintenant d'une méthode simple de comparer les tests, en nous basant sur la puissance asympotique locale...
- Pour les tests asymptotiquement normaux, il suffit de comparer la pente des tests, à savoir $\mu'(0)/\sigma(0)$.
- Plus la pente est grande, plus rapidement $\pi(h)$ augmente avec h !

Application: test du signe

- $U_n = n^{-1} \sum_{i=1}^n \mathbb{1}_{\{X_i > 0\}}$.
- $\mu(\theta) = 1 - F(-\theta)$, $\mu'(\theta) = f(\theta)$.
- $\sigma^2(\theta) = (1 - F(-\theta))F(-\theta)$
- Pente: $\mu'(0)/\sigma(0) = 2f(0)$.

Application: t-test

- $T_n = \bar{X}_n / S_n$.
- $\mu(\theta) = \theta/s$ and $\sigma(\theta) = 1$. En effet

$$\begin{aligned}\sqrt{n}(T_n - \theta_n/s) &= \sqrt{n}(\bar{X}_n/S_n - \theta_n/S_n) + \sqrt{n}\theta_n(S_n^{-1} - s^{-1}) \\ &= n^{-1/2}\sum_{i=1}^n (X_i - \theta_n)/S_n + +\sqrt{n}\theta_n(S_n^{-1} - s^{-1}) \xrightarrow{d}_{\mathbb{P}_{\theta_n}} \mathcal{N}(0,1)\,.\end{aligned}$$

- Pente: $\mu'(0)/\sigma(0) = 1/s$.

Efficacité relative

1. test du signe: $\mu'(0)/\sigma(0) = 2f(0)$,
2. t-test: $\mu'(0)/\sigma(0) = 1/s$.

- Laplace: $2f(0)s = 2$.
- Logistique: $2f(0)s = \pi^2/12 = 0.822$.
- Gauss: $2f(0)s = 2/\pi = 0.6366$.
- Uniforme: $2f(0)s = 1/3$.

Test du rapport de vraisemblance

- Soit $X^{(n)} = (X_1, \ , X_n)$ un n-échantillon du modèle statistique $\mathbb{P}_\theta^n \ll \mu_n, \ \theta \in \Theta$, de densité $f_\theta(x^{(n)}) = \mathrm{d}\,\mathbb{P}_\theta^n / \mathrm{d}\mu_n$.
- Pour tester $H_0 : \theta \in \Theta_0$ contre $H_1 : \theta \in \Theta - \Theta_0$, le test du rapport de vraisemblance rejette H_0 lorsque la valeur du rapport de vraisemblance généralisé

$$\Lambda_n = \frac{\sup_{\theta \in \Theta_0} f_\theta(X^{(n)})}{\sup_{\theta \in \Theta} f_\theta(X^{(n)})}$$

  est inférieure à un seuil.
- Lorsque les hypothèses $H_0, \ H_1$ sont simples, ce test est U.P.P. .
- Pour des hypothèses composites, il n'y a en général aucun résultat d'optimalité, sauf dans des cas simples...

t-test

- Soient $X^{(n)} = (X_1, \ldots, X_n)$ un n-échantillon de $\mathcal{N}(\mu,\ \sigma^2)$.
- On teste l'hypothèse $H_0 : \mu = 0$ contre $H_1 : \mu \neq 0$.
- En posant $\theta = (\mu,\ \sigma^2)$,

$$\Lambda_n = \frac{\sup_{\theta \in \Theta_0} (1/\sigma^n) \exp(-\frac{1}{2\sigma^2} \sum_i (X_i - \mu)^2)}{\sup_{\theta \in \Theta} (1/\sigma^n) \exp(-\frac{1}{2\sigma^2} \sum_i (X_i - \mu)^2)}$$
$$= \Big(\frac{\sum_i (X_i - \overline{X}_n)^2}{\sum_i X_i^2}\Big)^{n/2}$$

- Un calcul élémentaire montre que $\Lambda_n < c$ est équivalent à $t_n^2 > k$ où

$$t_n = \frac{\sqrt{n}\overline{X}_n}{\sqrt{\frac{1}{n-1} \sum_i (X_i - \overline{X}_n)^2}}$$

est la t-statistique. En d'autres termes, le t-test est un test de rapport de vraisemblance généralisé.

Justification

$$
\begin{aligned}
t_n^2 &= \frac{n\overline{X}_n^2}{\frac{1}{n-1}\sum_i (X_i - \overline{X}_n)^2} \\
&= \frac{\sum_i X_i^2 - \sum_i (X_i - \overline{X}_n)^2}{\frac{1}{n-1}\sum_i (X_i - \overline{X}_n)^2} \\
&= \frac{(n-1)\sum_i X_i^2}{\sum_i (X_i - \overline{X}_n)^2} - (n-1) \\
&= (n-1)\Lambda_n^{-2/n} - (n-1)
\end{aligned}
$$

ce qui montre que

$$
\Lambda_n = \Big(\frac{n-1}{t_n^2 + n - 1}\Big)^{n/2}
$$

Distribution asymptotique

Comme

$$\Lambda_n = \left(\frac{n-1}{t_n^2 + n - 1}\right)^{n/2}$$

nous avons

$$\begin{aligned} \log \Lambda_n &= \frac{n}{2} \log \frac{n-1}{t_n^2 + n - 1} \\ \Rightarrow \quad -2 \log \Lambda_n &= n \log\left(1 + \frac{t_n^2}{n-1}\right) \\ &= n \left(\frac{t_n^2}{n-1} + o_p\left(\frac{t_n^2}{n-1}\right) \right) \xrightarrow{d}_{\mathbb{P}_0} \chi_1^2 \end{aligned}$$

car sous H_0, $t_n \xrightarrow{d}_{\mathbb{P}_0} \mathcal{N}(0,1)$.

résultat vrai en toute généralité !

Test d'égalité des proportions pour une variable multinomiale

- Soit $(X_1, \ldots, X_n)$ un n-échantillon d'une loi multinomiale à d-instances
- Paramètre $\mathbf{p} = (p_1, \ldots, p_d) \in \mathcal{M}_d = \{(p_1, \ldots, p_d), p_i \geq 0, \sum_{i=1}^d p_i = 1\}$.
- Rapport de vraisemblance généralisé

$$\Lambda_n = \frac{(1/d)^n}{\max_{(p_1,\ldots,p_k)\in\mathcal{M}_d} \prod_{i=1}^d p_j^{N_i}} = \prod_{i=1}^d \left(\frac{n}{dN_i}\right)^{N_i} = \prod_{i=1}^d (d\hat{p}_{n,i})^{-N_i}$$

où $N_i = \sum_{j=1}^n \mathbb{1}_{\{X_j = i\}}$ et $\hat{p}_{n,i} = N_i/n$ les fréquences empiriques.

Loi limite des fréquences empiriques

- On suppose $(X_1, \ldots, X_n)$ n-échantillon multinomial de proportion $(q_1, \ldots, q_d)$.
- Comparaison des fréquences empiriques

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

- 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

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

- 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 $d-1$.

Distribution limite de $-2\log\Lambda_n$

Nous avons

$$
\begin{aligned}
-2\log\Lambda_n &= 2\sum_{i=1}^{d} N_i \log(\hat{p}_{N,i}/p_i)\\
&= 2n\sum_{i=1}^{d}(\hat{p}_{N,i}-p_i+p_i)\log\left(1+\frac{\hat{p}_{n,i}-p_i}{p_i}\right)\\
&= 2n\sum_{i=1}^{d}\frac{(\hat{p}_{n,i}-p_i)^2}{p_i}+2n\sum_{i=1}^{d}p_i\left\{\frac{(\hat{p}_{n,i}-p_i)}{p_i}-\frac{1}{2}\left(\frac{\hat{p}_{n,i}-p_i}{p_i}\right)^2\right\}+o_{\mathbb{P}}(1)\,.
\end{aligned}
$$

car $\sum_{i=1}^{d} p_i(\hat{p}_{n,i}-p_i)/p_i=0$... Par conséquent

$$-2\log\Lambda_n \xrightarrow{d}_{\mathbb{P}_\mathsf{I}} \chi^2(d-1)$$

Trop beau pour qu'il n'y ait pas quelque chose de plus profond.. en PC

Le test de Wald : hypothèse nulle simple

- Situation la suite d'expériences $\big(\mathsf{X}^n, \mathcal{X}^{\otimes n}, \{\mathbb{P}^n_\theta, \theta \in \Theta\}\big)$ 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
$$\boxed{\sqrt{n}(\widehat{\theta}_n - \theta) \xrightarrow{d} \mathcal{N}\big(0, v(\theta)\big)}$$
en loi sous $\mathbb{P}^n_\theta$, $\forall \theta \in \Theta$, où $\theta \rightsquigarrow v(\theta) > 0$ est continue.
- Sous l'hypothèse (ici sous $\mathbb{P}^n_{\theta_0}$) 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 $\mathbb{P}^n_{\theta_0}$.

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 $\mathbb{P}^n_{\theta_0}$.

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

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

- 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

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

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 $\widehat{\theta}_n$ 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

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

et

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

$(m \leq d)$ 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

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

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

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

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

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

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

- 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 $\theta \in \Theta$,

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

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 de base.

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

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

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 : $X \in \{1, \dots, d\}$.

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

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

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

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

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

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

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

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

où $\mathbb{P}\left[U > q_{1-\alpha,d-1}^{\chi^2}\right] = \alpha$ 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
$$\boldsymbol{q} = \Big(\frac{9}{16}, \frac{3}{16}, \frac{3}{16}, \frac{1}{16}\Big).$$
- 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 $\alpha = 5\%$.