MAP 433 : Introduction aux méthodes statistiques. Cours 8

Marc Hoffmann

4 avril 2014

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

1. Notion de test et d'erreur de test
   - Hypothèse simple contre alternative simple
   - Lemme de Neyman-Pearson
2. Construction d'un test : hypothèses générales
   - Retour sur un exemple
   - Principe de construction
3. Tests asymptotiques
4. Tests d'adéquation
   - Tests de Kolmogorov-Smirnov
   - Tests du χ^2

Exemple introductif

- On observe 10 lancers d'une pièce de monnaie et on obtient le résultat suivant :

$$(P, P, F, F, P, F, P, P, F, P).$$

La pièce est-elle équilibrée ?

- Répondre à cette question revient à construire une procédure de décision :

$$\varphi = \varphi(P, P, F, F, P, F, P, P, F, P)$$

$$= \begin{cases} 0 & \text{on accepte l'hypothèse « la pièce est équilibrée »} \\ 1 & \text{on rejette l'hypothèse « la pièce est équilibrée »} \end{cases}$$

Résolution

- On associe l'expérience statistique (par exemple)

$$\mathcal{E}^{10} = \left(\{0,1\}^{10}, \text{parties de}(\{0,1\}^{10}), \{\mathbb{P}_\vartheta^{10}, \vartheta \in [0,1]\}\right),$$

avec $(P = 0,\ F = 1)$

$$\mathbb{P}_\vartheta^{10} = \left(\vartheta\delta_0(dx) + (1-\vartheta)\delta_1(dx)\right)^{\otimes 10}.$$

- Hypothèse nulle : « la pièce est équilibrée »

$$\boxed{H_0 : \vartheta = \frac{1}{2}}$$

- Hypothèse alternative : « la pièce est truquée »

$$\boxed{H_1 : \vartheta \neq \frac{1}{2}}$$

Résolution (cont.)

- On note Z l'observation.
- On construit une règle de décision simple :

$$\varphi = 1_{\{Z \in \mathcal{R}\}} = \begin{cases} 0 & \text{on accepte l'hypothèse} \\ 1 & \text{on rejette l'hypothèse.} \end{cases}$$

- $\mathcal{R} \subset \mathfrak{Z}$ (espace des observables) : zone de rejet ou région critique.
- Exemple[1]

$$\mathcal{R} = \left\{ \left| \widehat{\vartheta}(Z) - \tfrac{1}{2} \right| > t_0 \right\}, \ \ \widehat{\vartheta}(Z) = \widehat{\vartheta}_{\mathsf{n}}^{\,\mathtt{mv}} \left(\overset{exemple}{=} 0,6 \right)$$

où t_0 est un seuil à choisir... Comment ?

1. léger abus de notation...

Erreur de décision

- Lorsque l'on prend la décision φ, on peut se tromper de deux manières :

$$\text{Rejeter } H_0 \ \ (\varphi = 1) \ \ \text{alors que} \ \ \vartheta = \frac{1}{2}$$

  ou encore

$$\text{Accepter } H_0 \ \ (\varphi = 0) \ \ \text{alors que} \ \ \vartheta \neq \frac{1}{2}.$$

- Erreur de première espèce (=rejeter à tort)

$$\mathbb{P}^{10}_{\frac{1}{2}}\left[\varphi = 1\right]$$

- Erreur de seconde espèce (=accepter à tort)

$$\Big(\mathbb{P}^{10}_{\vartheta}\left[\varphi = 0\right], \ \ \vartheta \neq \frac{1}{2}\Big).$$

Conclusion provisoire

- Un ≪ bon test ≫ φ doit garantir simultanément des erreurs de première et seconde espèce petites.
- Un test optimal existe-t-il ?
- Si non, comment aborder la notion d'optimalité et comment construire un test optimal ?

Définition formelle

- Situation : $\mathcal{E} = \big(\mathcal{Z}, \mathfrak{Z}, \{\mathbb{P}_\vartheta, \vartheta \in \Theta\}\big)$ engendrée par l'observation Z.
- Hypothèse nulle et alternative : $\Theta_0 \subset \Theta$ et $\Theta_1 \subset \Theta$ t.q.

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

Définition (Test simple)

Un test (simple) de l'hypothèse nulle $H_0 : \vartheta \in \Theta_0$ contre l'alternative $H_1 : \vartheta \in \Theta_1$ est une statistique $\varphi = \varphi(Z) \in \{0, 1\}$. (Fonction d') erreur de première espèce :

$$\vartheta \in \Theta_0 \rightsquigarrow \mathbb{P}_\vartheta\big[\varphi = 1\big]$$

(Fonction d') erreur de seconde espèce

$$\vartheta \in \Theta_1 \rightsquigarrow \mathbb{P}_\vartheta\big[\varphi = 0\big] = 1 - \text{puissance}_\varphi(\vartheta).$$

Hypothèse simple contre alternative simple

- Cas où $\Theta = \{\vartheta_0, \vartheta_1\}$ avec $\vartheta_0 \neq \vartheta_1$.
- Existe-t-il un test $\varphi^\star$ optimal, au sens où : $\forall \varphi$ test simple, on a simultanément

$$\mathbb{P}_{\vartheta_0}\left[\varphi^\star = 1\right] \leq \mathbb{P}_{\vartheta_0}\left[\varphi = 1\right]$$

et

$$\mathbb{P}_{\vartheta_1}\left[\varphi^\star = 0\right] \leq \mathbb{P}_{\vartheta_1}\left[\varphi = 0\right] \quad ?$$

- Si $\mathbb{P}_{\vartheta_0}$ et $\mathbb{P}_{\vartheta_1}$ ne sont pas étrangères (cf. Cours 6) un tel test $\varphi^\star$ ne peut pas exister.

Absence d'optimalité stricte

- Equivalence tests simples $\rightsquigarrow$ estimateurs $\widehat{\vartheta}$ de ϑ via la représentation :

$$\widehat{\vartheta} = \vartheta_0 1_{\mathcal{R}^c} + \vartheta_1 1_{\mathcal{R}} \leftrightsquigarrow \varphi = 1_{\mathcal{R}}.$$

- Fonction de risque

$$\mathcal{P}(\varphi, \vartheta) = \mathbb{E}_\vartheta\left[1_{\widehat{\vartheta} \neq \vartheta}\right], \;\; \vartheta = \vartheta_0, \vartheta_1.$$

- La fonction de perte $\ell(\widehat{\vartheta}, \vartheta) = 1_{\widehat{\vartheta} \neq \vartheta}$ joue le même rôle que la perte quadratique $(\widehat{\vartheta} - \vartheta)^2$ dans le Cours 6.
- Test optimal $\varphi^\star \leftrightsquigarrow$ estimateur optimal $\vartheta^\star$ pour $\mathcal{P}$.
- Comme pour le cas du risque quadratique, dès que $\mathbb{P}_{\vartheta_0}$ et $\mathbb{P}_{\vartheta_1}$ ne sont pas étrangères, un estimateur optimal n'existe pas (cf. Cours 6).

Riposte : principe de Neyman

- On « disymétrise » les hypothèses H_0 et H_1 : H_0 est « plus importante » que H_1 dans le sens suivant : on impose une erreur de première espèce prescrite.

Définition

Pour $\alpha \in [0,1]$, un test $\varphi = \varphi_\alpha$ de l'hypothèse nulle $H_0 : \vartheta \in \Theta_0$ contre une alternative H_1 est de niveau α si

$$\sup_{\vartheta \in \Theta_0} \mathbb{P}_\vartheta \left[\varphi_\alpha = 1 \right] \leq \alpha.$$

- Un test de niveau α ne dit rien sur l'erreur de seconde espèce (comportement sur l'alternative).

Principe de Neyman (cont.)

- Choix de la « disymétrisation » = choix de modélisation.
- Principe de Neyman : $\alpha \in (0, 1)$, parmi les test de niveau α, chercher celui (ou ceux) ayant une erreur de seconde espèce minimale.

Définition

Un test de niveau α est dit Uniformément Plus Puissant (UPP) si son erreur de seconde espèce est minimale parmi celles des tests de niveau α.

- Pour le cas d'une hypothèse simple contre une alternative simple, un test UPP existe.

Principe de construction

- $f(\vartheta, z) = \frac{d\,\mathbb{P}_\vartheta}{d\mu}(z)$, $z \in \mathfrak{Z}$, $\vartheta = \vartheta_0, \vartheta_1$, μ mesure dominante. L'EMV –si bien défini– s'écrit

$$\widehat{\vartheta}_n^{\,\mathtt{mv}} = \vartheta_0 1_{\{f(\vartheta_1, Z) < f(\vartheta_0, Z)\}} + \vartheta_1 1_{\{f(\vartheta_0, Z) < f(\vartheta_1, Z)\}}.$$

- On choisit une région critique de la forme

$$\boxed{\mathcal{R}(c) = \{f(\vartheta_1, Z) > c f(\vartheta_0, Z)\}, \quad c > 0}$$

et on calibre $c = c_\alpha$ de sorte que

$$\boxed{\mathbb{P}_{\vartheta_0}\left[Z \in \mathcal{R}(c_\alpha)\right] = \alpha.}$$

- Le test ainsi construit (si cette équation admet une solution) est de niveau α. On montre qu'il est UPP.

Lemme de Neyman-Pearson

Proposition

Soit $\alpha \in [0,1]$. *S'il existe* c_α *solution de*

$$\boxed{\mathbb{P}_{\vartheta_0}\left[f(\vartheta_1, Z) > c_\alpha f(\vartheta_0, Z)\right] = \alpha}$$

alors le test de région critique $\mathcal{R}_\alpha = \left\{f(\vartheta_1, Z) > c_\alpha f(\vartheta_0, Z)\right\}$ *est de niveau* α *et UPP pour tester* $H_0 : \vartheta = \vartheta_0$ *contre* $H_1 : \vartheta = \vartheta_1$.

- Si $U = f(\vartheta_1, Z)/f(\vartheta_0, Z)$ bien définie et $\mathcal{L}(U) \ll dx$ (sous $\mathbb{P}_{\vartheta_0}$), alors $\mathbb{P}_{\vartheta_0}\left[U > c_\alpha\right] = \alpha$ admet une solution.

Exemple de mise en oeuvre

- On observe

$$Z = (X_1, \ldots, X_n) \sim_{\text{i.i.d.}} \mathcal{N}(\vartheta, 1).$$

- Construction du test de N-P. de $H_0 : \vartheta = \vartheta_0$ contre $H_1 : \vartheta = \vartheta_1$, avec $\vartheta_0 < \vartheta_1$.
- Mesure dominante $\mu^n =$ mesure de Lebesgue sur $\mathbb{R}^n$ et

$$f(\vartheta, Z) = \tfrac{1}{(2\pi)^{n/2}} \exp\big(-\frac{1}{2}\sum_{i=1}^n X_i^2 + n\vartheta \overline{X}_n - \frac{n\vartheta^2}{2}\big).$$

- Rapport de vraisemblance

$$\frac{f(\vartheta_1, Z)}{f(\vartheta_0, Z)} = \exp\big(n(\vartheta_1 - \vartheta_0)\overline{X}_n - \frac{n}{2}(\vartheta_1^2 - \vartheta_0^2)\big).$$

Exemple (cont.)

- Zone de rejet du test de N-P. :

$$\begin{aligned}&\{f(\vartheta_1, Z) > cf(\vartheta_0, Z)\}\\ =&\{n(\vartheta_1 - \vartheta_0)\overline{X}_n - \frac{n}{2}(\vartheta_1^2 - \vartheta_0^2) > \log c\}\\ =&\{\overline{X}_n > \frac{\vartheta_0 + \vartheta_1}{2} + \frac{\log c}{n(\vartheta_1 - \vartheta_0)}\}.\end{aligned}$$

- Choix de c. On résout

$$\mathbb{P}_{\vartheta_0}\Big[\overline{X}_n > \tfrac{1}{2}(\vartheta_0 + \vartheta_1) + \frac{\log c}{n(\vartheta_1 - \vartheta_0)}\Big] = \alpha.$$

- Approche standard : on raisonne sous $\mathbb{P}_{\vartheta_0}$. On a

$$\overline{X}_n = \vartheta_0 + \frac{1}{\sqrt{n}}\xi^{n,\vartheta_0},$$

où ξ^{n,ϑ_0} est une gaussienne standard $\mathcal{N}(0,1)$ sous $\mathbb{P}_{\vartheta_0}$ mais pas sous une autre probabilité $\mathbb{P}_\vartheta$ si $\vartheta \neq \vartheta_0$!

Exemple (fin)

- Résolution de

$$\mathbb{P}_{\vartheta_0}\left[\vartheta_0 + \frac{1}{\sqrt{n}}\xi^{n,\vartheta_0} > \frac{1}{2}(\vartheta_0+\vartheta_1) + \frac{\log c}{n(\vartheta_1-\vartheta_0)}\right] = \alpha.$$

- Equivalent à $\mathbb{P}_{\vartheta_0}\left[\xi^{n\vartheta_0} > \frac{\sqrt{n}}{2}(\vartheta_1-\vartheta_0) + \frac{1}{\sqrt{n}}\frac{\log c}{\vartheta_1-\vartheta_0}\right] = \alpha$, soit

$$\frac{\sqrt{n}}{2}(\vartheta_1-\vartheta_0) + \frac{1}{\sqrt{n}}\frac{\log c}{\vartheta_1-\vartheta_0} = \Phi^{-1}(1-\alpha),$$

où $\Phi(x) = \int_{-\infty}^{x} e^{-u^2/2}\frac{du}{\sqrt{2\pi}}$.

- Conclusion

$$\boxed{c_\alpha = \exp\left(n\frac{(\vartheta_1-\vartheta_0)^2}{2} + \sqrt{n}(\vartheta_1-\vartheta_0)\Phi^{-1}(1-\alpha)\right)}$$

Bilan provisoire

- Si l'on accepte le principe de Neyman, on sait résoudre le problème à deux points.
- Que faire si l'hypothèse nulle H_0 ou l'alternative H_1 sont composites ?
  - On peut proposer des extensions si l'on dispose de structures particulières sur la vraisemblance du modèle (Poly. Ch. 7.3, hors programme).
  - On sait dire beaucoup de choses dans le cas gaussien.
- Critique méthodologique de l'approche de Neyman $\rightsquigarrow$ notion de p-valeur.

Situation

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

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

  avec $\Theta_0 \cap \Theta_1 = \emptyset$.
- Si $\Theta_0 = \{\vartheta_0\}$ et $\Theta_1 = \{\vartheta_1\}$, 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 $\mathbb{P}_\vartheta$ avec $\vartheta \in \Theta_0$.
- 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 $\vartheta = (\mu, \sigma^2) \in \Theta = \mathbb{R} \times (0, +\infty)$.

- 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 $\mathbb{P}_\vartheta$ avec $\vartheta = (\mu, \sigma_0)$ 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 $q^{\chi^2}_{1-\alpha,n-1} > 0$ tel que si $U \sim \chi^2(n-1)$, alors

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

- Sous l'hypothèse $\phi(Z) \stackrel{d}{=} U$ et donc la probabilité pour que $\phi(Z)$ dépasse $q^{\chi^2}_{1-\alpha,n-1}$ 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

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

- 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 $H_0 : \vartheta \in \Theta_0$ contre $H_1 : \vartheta \in \Theta_1$, 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 réapparaît le plus souvent...).

Le test de Wald : hypothèse nulle simple

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

$$H_0 : \vartheta = \vartheta_0 \quad \text{contre} \quad \vartheta \neq \vartheta_0.$$

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

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

  en loi sous $\mathbb{P}_\vartheta^n$, $\forall \vartheta \in \Theta$, où $\vartheta \rightsquigarrow v(\vartheta) > 0$ est continue.
- Sous l'hypothèse (ici sous $\mathbb{P}_{\vartheta_0}^n$) on a la convergence

$$\sqrt{n}\frac{\widehat{\vartheta}_n - \vartheta_0}{\sqrt{v(\widehat{\vartheta}_n)}} \xrightarrow{d} \mathcal{N}(0,1)$$

  en loi sous $\mathbb{P}_{\vartheta_0}^n$.

Test de Wald (cont.)

- Remarque $\sqrt{v(\widehat{\vartheta}_n)} \leftrightarrow \sqrt{v(\vartheta_0)}$ ou d'autres choix encore...
- On a aussi

$$T_n = n\frac{(\widehat{\vartheta}_n - \vartheta_0)^2}{v(\widehat{\vartheta}_n)} \xrightarrow{d} \chi^2(1)$$

  sous $\mathbb{P}^n_{\vartheta_0}$.
- 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 $\vartheta = \vartheta_0$ contre l'alternative $\vartheta \neq \vartheta_0$ basé sur $\widehat{\vartheta}_n$.

Propriétés du test de Wald

Proposition

Le test Wald de l'hypothèse simple $\vartheta = \vartheta_0$ contre l'alternative $\vartheta \neq \vartheta_0$ basé sur $\widehat{\vartheta}_n$ est

- *asymptotiquement de niveau α :*

$$\mathbb{P}^n_{\vartheta_0}\left[T_n \in \mathcal{R}_{n,\alpha}\right] \to \alpha.$$

- *convergent ou (consistant). Pour tout point $\vartheta \neq \vartheta_0$*

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

Preuve

- Test asymptotiquement de niveau α par construction.
- Contrôle de l'erreur de seconde espèce : Soit $\vartheta \neq \vartheta_0$. On a

$$\begin{aligned} T_n &= \left(\sqrt{n}\frac{\widehat{\vartheta}_n - \vartheta}{\sqrt{v(\widehat{\vartheta}_n)}} + \sqrt{n}\frac{\vartheta - \vartheta_0}{\sqrt{v(\widehat{\vartheta}_n)}}\right)^2 \\ &=: T_{n,1} + T_{n,2}. \end{aligned}$$

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

$$T_{n,2} \xrightarrow{\mathbb{P}_\vartheta^n} \pm\infty \ \text{ car } \ \vartheta \neq \vartheta_0$$

Donc $T_n \xrightarrow{\mathbb{P}_\vartheta^n} +\infty$, d'où le résultat.
- Remarque : si $\vartheta \neq \vartheta_0$ mais $|\vartheta - \vartheta_0| \lesssim 1/\sqrt{n}$, le raisonnement ne s'applique pas. Résultat non uniforme en le paramètre.

Test de Wald : hypothèse nulle composite

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

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

  où $V(\vartheta)$ est définie positive et continue en ϑ.
- But Tester $H_0 : \vartheta \in \Theta_0$ contre $H_1 : \vartheta \notin \Theta_0$, où

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

  et

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

  $(m \leq d)$ est régulière.

Test de Wald cont.

- Hypothèse : la différentielle (de matrice $J_g(\vartheta)$) 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 $\mathbb{P}_\vartheta^n$:

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

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

- 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 *:*

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

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

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

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 $\int_{\mathbb{R}} xdF(x) = 0$, $\int_{\mathbb{R}} x^2 dF(x) = 1$, alors

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

(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

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

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 :

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

- Théorème central-limite ?

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

- 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

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

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
$$\boldsymbol{q} = \Big(\frac{9}{16}, \frac{3}{16}, \frac{3}{16}, \frac{1}{16}\Big).$$
- 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 $\alpha = 5\%$.