1 Introduction raisonnée

En statistique, on cherche souvent à décider si une affirmation est compatible avec des données observées. Or, un écart entre la valeur observée et la valeur annoncée peut simplement résulter du hasard d'échantillonnage. Un test statistique permet de le décider à partir d'une règle fixée à l'avance, tout en contrôlant le risque de se tromper. Un test est donc un mécanisme permettant de trancher entre 2 hypothèses.

Exemple. Une machine doit remplir des bouteilles de Plancoët avec une quantité moyenne de 1 l. On observe $n = 25$ bouteilles. Peut-on conclure qu'elle remplit trop les bouteilles si la moyenne observée est 1,04 l ? Il faut connaître la variabilité des remplissages : un écart de 40 ml peut être banal ou très surprenant suivant cette variabilité.

Hypothèse nulle et alternative

On oppose deux affirmations sur un paramètre inconnu θ :

$$\mathscr{H}_0 : \theta \in \Theta_0 \quad \text{(hypothèse nulle)},$$
$$\mathscr{H}_1 : \theta \in \Theta_1 \quad \text{(alternative)},$$

avec $\Theta_0 \cap \Theta_1 = \varnothing$. En général, $\Theta_0 \cup \Theta_1 = \Theta$, l'espace des paramètres.

Définition Idée d'un test

On **ne démontre pas** que $\mathscr{H}_0$ est vraie ou fausse : on examine si les données sont suffisamment peu compatibles avec $\mathscr{H}_0$ pour la **rejeter**. Sinon, on **ne rejette pas** $\mathscr{H}_0$ (ce qui ne signifie pas qu'elle est vraie). Ce n'est pas symétrique : on privilégie *a priori* $\mathscr{H}_0$.

Test unilatéral droit : $\mathscr{H}_0 : \theta \le \theta_0$ contre $\mathscr{H}_1 : \theta > \theta_0$. On recherche un excès.

Test unilatéral gauche : $\mathscr{H}_0 : \theta \ge \theta_0$ contre $\mathscr{H}_1 : \theta < \theta_0$. On recherche un déficit.

Test bilatéral : $\mathscr{H}_0 : \theta = \theta_0$ contre $\mathscr{H}_1 : \theta \neq \theta_0$. On cherche alors un écart dans les deux sens.

Une hypothèse est dite **simple** si elle fixe une seule valeur du paramètre ; elle est **composite** si elle en autorise plusieurs.

Statistique de test et région critique

On dispose de données $X = (X_1, \dots, X_n)$ et on calcule une **statistique de test** $T(X)$, choisie de façon à rendre mesurable l'écart à $\mathscr{H}_0$. On connaît sa loi sous $\mathscr{H}_0$, exactement ou approximativement.

Définition

La **région critique** (ou région de rejet) est l'ensemble R des réalisations de X conduisant au rejet de $\mathscr{H}_0$. Un **test pur** est une fonction de décision

$$\phi(X) = \mathbb{1}_{[X \in R]} \in \{0, 1\}.$$

Si R_T est l'ensemble critique pour T, $R = \{x : T(x) \in R_T\}$. La décision est $\phi = 1$ si l'on rejette $\mathscr{H}_0$, et $\phi = 0$ sinon.

Les deux types d'erreurs

	$\mathscr{H}_0$ vraie	$\mathscr{H}_1$ vraie
$\neg$ rejeter $\mathscr{H}_0$	Bonne décision	Erreur II
Rejeter $\mathscr{H}_0$	Erreur I	Bonne décision

Erreur de 1re espèce : rejeter $\mathscr{H}_0$ alors qu'elle est vraie (fausse alerte). Sa proba vaut, pour $\theta \in \Theta_0$,

$$\alpha(\theta) = \mathbb{P}_\theta[\phi(X) = 1].$$

Erreur de 2e espèce : ne pas rejeter $\mathscr{H}_0$ alors que $\mathscr{H}_1$ est vraie (non-détection). Sa proba vaut, pour $\theta \in \Theta_1$,

$$\beta(\theta) = \mathbb{P}_\theta[\phi(X) = 0].$$

La **puissance** en $\theta \in \Theta_1$ est la proba de rejeter $\mathscr{H}_0$ lorsqu'elle est fausse :

$$\pi(\theta) = \mathbb{P}_\theta[\phi(X) = 1] = 1 - \beta(\theta).$$

Bon test : erreur 1re espèce faible, puissance élevée.

Niveau du test et seuil de rejet

On fixe un risque maxi de 1re espèce ($\alpha = 5\,\%$).

Définition Niveau et taille

La **taille** d'un test est

$$\alpha^* = \sup_{\theta \in \Theta_0} \mathbb{P}_\theta[\phi = 1].$$

Un test est **de niveau** α si $\alpha^* \le \alpha$. Il est *de taille exactement* α si $\alpha^* = \alpha$.

Pour une statistique T dont les grandes valeurs contredisent $\mathscr{H}_0$, on rejette souvent lorsque $T(X) > q_\alpha$, où q_α est un quantile déterminé *sous* $\mathscr{H}_0$. Le niveau 5 % ne signifie donc pas que l'on a 95 % de chances que $\mathscr{H}_0$ soit vraie.

Compromis : pour un même modèle et un même effectif, rendre le rejet plus difficile diminue en général α mais augmente β. Une augmentation de l'effectif n permet souvent d'améliorer la puissance à niveau fixé.

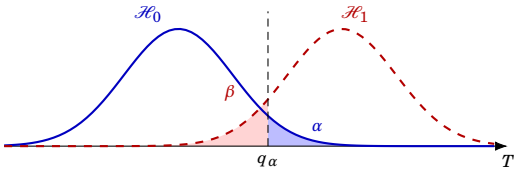


FIGURE 1 – Les 2 lois de T : zone bleue, fausse alerte α ; zone rouge, non-détection β. $T \sim \mathcal{N}(0,1)$ sous $\mathcal{H}_0$ et $T \sim \mathcal{N}(3,1)$ sous $\mathcal{H}_1$.

La p-valeur mesure la surprise

La p-valeur indique à quel point les données observées sont extrêmes *si $\mathcal{H}_0$ est vraie*. Pour un test unilatéral droit avec $\mathcal{H}_0 : \theta = \theta_0$:

$$p(x) = \mathbb{P}_{\theta_0}\left[T(X) \geq T(x)\right]$$

Définition — p-valeur

La p-valeur est le **plus petit niveau de test** dont l'observation conduit au rejet. Lorsque la p-valeur est correctement calibrée,

$$p(x) \leq \alpha \quad \Longrightarrow \quad \text{rejet de } \mathcal{H}_0 \text{ au niveau } \alpha.$$

Elle n'est **ni** la probabilité que $\mathcal{H}_0$ soit vraie, **ni** la probabilité de s'être trompé après le rejet.

Pour un test bilatéral gaussien, $T = |Z|$ est une statistique qui grandit dans *les deux sens*, et la p-valeur est $\mathbb{P}[|Z| \geq |z_{\text{obs}}|]$. Si $\mathcal{H}_0$ est composite et si les grandes valeurs de T contredisent $\mathcal{H}_0$, une p-valeur valable est $\sup_{\theta \in \Theta_0} \mathbb{P}_\theta[T(X) \geq T(x)]$. Sous $\mathcal{H}_0$, une p-valeur correctement construite vérifie $\mathbb{P}_\theta[p(X) \leq \alpha] \leq \alpha$. Dans les modèles discrets, on inclut l'égalité à la valeur observée dans la probabilité de queue.

Exemple numérique complet

On reprend la machine. Supposons $X_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, 10^{-2})$ (en ml), avec $n = 25$, et une moyenne observée $\bar{x} = 1{,}04$. On teste $\mathcal{H}_0 : \mu \leq 1$ contre $\mathcal{H}_1 : \mu > 1$ au niveau 5 %. Sous la frontière $\mu = 1$:

$$Z = \frac{\bar{x} - 1}{0{,}1/\sqrt{25}} \sim \mathcal{N}(0,1).$$

Le quantile normal est $z_{0.95} \simeq 1.645$. On rejette lorsque

$$Z > 1.645 \iff \bar{x} > 1{,}03.$$

Or $z_{\text{obs}} = (1{,}04 - 1)/0{,}02 = 2$ et la p-valeur vaut

$$p = \mathbb{P}[Z \geq 2] = 1 - \Phi(2) \simeq 0.0228.$$

Décision : puisque $p < 0.05$, on rejette $\mathcal{H}_0$: les données apportent un argument statistique en faveur d'un sur-remplissage. Elles n'établissent pas avec certitude que $\mu > 1000$.

Si la vraie moyenne vaut $\mu = 1{,}06$, la puissance du test est

$$\pi(1{,}06) = \mathbb{P}_{1{,}06}[\bar{x} > 1{,}03] \simeq 0.912.$$

Lien test ~ intervalle de confiance

À chaque valeur hypothétique θ_0, on peut associer un test de $\mathcal{H}_0 : \theta = \theta_0$. L'ensemble des valeurs *non rejetées* est un intervalle de confiance (ou, plus généralement, une région de confiance), si les tests sont correctement calibrés.

Propriété — Dualité tests-IC

Pour une même procédure bilatérale de niveau α et l'intervalle de confiance obtenu par inversion :

$$\theta_0 \notin IC_{1-\alpha}(X) \quad \Longleftrightarrow \quad \mathcal{H}_0 : \theta = \theta_0 \text{ rejetée.}$$

Dans un modèle normal de variance σ^2 connue,

$$IC_{1-\alpha}(\mu) = \left[\bar{x} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}},\ \bar{x} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right].$$

La région critique concerne les échantillons ; l'IC décrit des valeurs possibles du paramètre : il s'agit de deux espaces différents. Quelques remarques :

- α et β varient en sens inverse.
- α est fixé en fonction du risque de 1re espèce : le rôle de $\mathcal{H}_0$ est prééminent.
- la valeur du seuil critique k ne dépend que de $\mathcal{H}_0$ et α.
- la forme de la région critique est indiquée par la nature de $\mathcal{H}_1$ (unilatéral, bilatéral, à droite, à gauche).
- $\mathcal{H}_0$ et $\mathcal{H}_1$ ne jouent pas des rôles symétriques.
- le risque α dans un intervalle de confiance correspond au risque de 1re espèce et est lié à la p-valeur par la formule $p(x) \leq \alpha \Rightarrow \mathcal{H}_0$ rejetée au niveau α.
- ⚠ attention à la terminologie «niveau» ! Pour un test, α est le risque de 1re espèce, le niveau du test ou le seuil de significativité, tandis que pour un IC, le niveau de confiance est $1 - \alpha$.

Méthode pratique en six étapes

1. Définir le modèle et le paramètre d'intérêt.
2. Écrire $\mathcal{H}_0$ et $\mathcal{H}_1$; préciser unilatéral/bilatéral.
3. Fixer le niveau α avant d'observer la décision.
4. Choisir $T(X)$ et déterminer sa loi sous $\mathcal{H}_0$.
5. Calculer la région critique ou la p-valeur.
6. Conclure par *rejet* ou *non-rejet* et, si possible, commenter l'effet estimé et la puissance.

2 Tests paramétriques, cadre théorique

Fonction de puissance et taille

Soit un modèle $(\mathbb{P}_\theta)_{\theta\in\Theta}$, une partition $\Theta = \Theta_0 \cup \Theta_1$ et un échantillon $X \in \mathcal{X}^n$. Pour une fonction de test $\phi(X) \in \{0,1\}$, la **fonction de puissance** est définie sur tout Θ par

$$\pi_\phi(\theta) = \mathbb{E}_\theta[\phi(X)] = \mathbb{P}_\theta[X \in R].$$

Sur Θ_0, elle donne l'erreur de première espèce ; sur Θ_1, elle donne $1-\beta(\theta)$.

On appelle *courbe de non-rejet* ou *courbe OC* la fonction $1-\pi_\phi(\theta)$. Il convient de ne pas la confondre avec la courbe de puissance π_ϕ.

Un *test randomisé* peut prendre des valeurs $0 \le \phi(x) \le 1$: $\phi(x)$ représente alors la prob conditionnelle de rejeter, sachant x.

Rapport de vraisemblance

Pour deux hypothèses simples $\mathcal{H}_0 : \theta = \theta_0$ et $\mathcal{H}_1 : \theta = \theta_1$, on définit

$$LR(x;\theta_1,\theta_0) = \frac{L(x;\theta_1)}{L(x;\theta_0)},$$

avec les conventions habituelles lorsque le dénominateur s'annule. Une grande valeur de LR signifie que les données sont relativement plus vraisemblables sous $\mathcal{H}_1$.

Lemme — Neyman–Pearson

Pour tester deux hypothèses *simples*, le test qui rejette lorsque $LR(x;\theta_1,\theta_0) > c$ est de **puissance maximale** parmi les tests de niveau au plus α, si c (et, au besoin, une randomisation sur la frontière) sont choisis pour obtenir la taille α :

$$\phi(x) = \begin{cases} 1, & LR(x) > c, \\ \rho, & LR(x) = c, \\ 0, & LR(x) < c, \end{cases} \qquad 0 \le \rho \le 1.$$

Sous $\mathcal{H}_0$, on impose $\mathbb{E}_{\theta_0}(\phi) = \alpha$.

Ce résultat exprime l'idée de Neyman–Pearson : **fixer le risque de première espèce puis maximiser la puissance**. Sans randomisation, dans un modèle discret, la taille peut rester strictement inférieure à α.

Tests uniformément plus puissants

Pour une alternative composite, un test est dit **uniformément plus puissant** (UPP) de niveau α si sa puissance est supérieure ou égale à celle de tout autre test de niveau α, *pour chaque* $\theta \in \Theta_1$. De tels tests n'existent pas en général pour des alternatives bilatérales.

Un modèle possède un *rapport de vraisemblance monotone* (croissant) en T si, pour $\theta_1 > \theta_0$, le rapport $L(x;\theta_1)/L(x;\theta_0)$ est une fonction non décroissante de $T(x)$. On peut alors rejeter lorsque T est grand.

Théorème — Test unilatéral et monotonie

Dans une famille à rapport de vraisemblance monotone croissant en T, sous les conditions usuelles de support commun, un test de $\mathcal{H}_0 : \theta \le \theta_0$ contre $\mathcal{H}_1 : \theta > \theta_0$ de la forme

$$\phi(X) = \mathbb{1}_{[T(X) > q_\alpha]}$$

est UPP de niveau α si le seuil est déterminé à la frontière θ_0 (éventuellement avec randomisation). Un énoncé analogue vaut pour l'alternative gauche.

C'est le cas de nombreuses familles exponentielles à un paramètre, avec une statistique naturelle T et un paramètre naturel croissant. Cette conclusion ne s'étend pas automatiquement aux tests bilatéraux.

Remarque. Le problème d'**équivalence** $\mathcal{H}_0 : \theta \le a$ ou $\theta \ge b$ contre $\mathcal{H}_1 : a < \theta < b$ se traite souvent par deux tests unilatéraux (TOST).

Consistance d'une suite de tests

Une suite de tests (ϕ_n) de niveau (asymptotique) contrôlé est **consistante** si, pour chaque alternative fixée $\theta \in \Theta_1$,

$$\pi_n(\theta) = \mathbb{P}_\theta(\phi_n = 1) \xrightarrow[n\to\infty]{} 1.$$

Il s'agit d'une limite quand l'*effectif* n augmente, et non lorsque $\theta \to +\infty$.

3 Quelques tests usuels

On note Φ la fonction de répartition de $\mathcal{N}(0,1)$, z_q son quantile d'ordre q, $t_{\nu,q}$ le quantile de Student à ν degrés de liberté et $\chi^2_{\nu,q}$ celui du khi-deux.

Moy. normale variance connue : test Z

Si $X_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu,\sigma^2)$ avec σ connu, sous $\mathcal{H}_0 : \mu = \mu_0$:

$$Z = \frac{\sqrt{n}(\bar{x} - \mu_0)}{\sigma} \sim \mathcal{N}(0,1).$$

Au niveau α, la région critique est

$$\begin{aligned} R_+ &= [Z > z_{1-\alpha}] && \text{(droite)}, \\ R_- &= [Z < -z_{1-\alpha}] && \text{(gauche)}, \\ R_\pm &= [|Z| > z_{1-\alpha/2}] && \text{(bilatéral)}. \end{aligned}$$

Le test est exact, grâce à la normalité. Si la loi des données n'est pas normale, l'approximation par le TCL peut être valable pour n suffisamment grand et sous les hypothèses requises.

Moy. normale variance inconnue : Student

Si $X_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2)$ avec σ inconnu, on utilise $S^2 = (n-1)^{-1}\sum_i (X_i - \overline{x})^2$:

$$T = \frac{\sqrt{n}(\overline{x} - \mu_0)}{S} \sim t_{n-1} \quad \text{sous } \mathcal{H}_0.$$

Pour un test bilatéral au niveau α, $R = [|T| > t_{n-1,1-\alpha/2}]$. La loi de Student est *exacte* sous la normalité des données.

Proportion d'une population

Si $X_i \overset{\text{i.i.d.}}{\sim} \text{Bernoulli}(p)$, avec $\widehat{p} = \overline{x}$, un test asymptotique de $\mathcal{H}_0 : p = p_0$ repose sur

$$Z = \frac{\sqrt{n}(\widehat{p} - p_0)}{\sqrt{p_0(1-p_0)}} \overset{\mathcal{H}_0}{\approx} \mathcal{N}(0,1).$$

Il faut un effectif suffisant et p_0 non trop proche de 0 ou de 1 (en pratique, vérifier np_0 et $n(1-p_0)$). Sinon, utiliser un **test binomial exact**, dont la p-valeur se calcule à partir de la loi $\mathcal{B}(n, p_0)$.

Variance normale : test du khi-deux

Si $X_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2)$, avec μ inconnu, pour tester $\mathcal{H}_0 : \sigma^2 = \sigma_0^2$:

$$Q = \frac{(n-1)S^2}{\sigma_0^2} \sim \chi^2_{n-1} \quad \text{sous } \mathcal{H}_0.$$

Pour un test bilatéral de niveau α,

$$R = \left[Q < \chi^2_{n-1,\alpha/2}\right] \cup \left[Q > \chi^2_{n-1,1-\alpha/2}\right].$$

Tests de Pearson : ajustement et indépendance

Le test d'*ajustement* confronte des effectifs observés O_j à des effectifs théoriques E_j sous $\mathcal{H}_0$:

$$Q = \sum_{j=1}^{k} \frac{(O_j - E_j)^2}{E_j}.$$

Quand les conditions d'approximation sont réunies, Q suit asymptotiquement une loi $\chi^2 : k-1$ degrés de liberté si les probabilités sont entièrement spécifiées, typiquement $k-1-r$ si r paramètres libres sont estimés sur les mêmes données sous les conditions usuelles.

Pour un tableau de contingence $a \times b$, le test d'*indépendance* utilise $E_{ij} = O_{i+}O_{+j}/n$ et la loi asymptotique $\chi^2_{(a-1)(b-1)}$ lorsque les conditions sont vérifiées.

4 Tests asymptotiques et vraisemblance

Niveau asymptotique et statistique de Wald

Une suite de tests (ϕ_n) a un *niveau asymptotique au plus* α si, par exemple, elle vérifie le contrôle uniforme

$$\limsup_{n\to\infty} \sup_{\theta \in \Theta_0} \mathbb{P}_\theta(\phi_n = 1) \le \alpha.$$

Pour une hypothèse nulle simple, il suffit de considérer la limite sous θ_0.

Théorème — Wald, dimension 1

Supposons qu'un estimateur $\widehat{\theta}_n$ vérifie

$$\sqrt{n}(\widehat{\theta}_n - \theta) \rightsquigarrow \mathcal{N}(0, \mathbb{V}(\theta)),$$

avec $\mathbb{V}(\theta) > 0$ continue et estimable de manière consistante. Sous $\mathcal{H}_0 : \theta = \theta_0$,

$$Z_n = \frac{\sqrt{n}(\widehat{\theta}_n - \theta_0)}{\sqrt{\mathbb{V}(\widehat{\theta}_n)}} \rightsquigarrow \mathcal{N}(0,1).$$

On rejette $\mathcal{H}_0$ au niveau asymptotique α si $|Z_n| > z_{1-\alpha/2}$ pour l'alternative bilatérale.

Pour l'EMV dans un modèle régulier, on peut prendre $\mathbb{V}(\theta) = I_1(\theta)^{-1}$, où I_1 est l'**information de Fisher d'une observation** (et $I_n = nI_1$ pour un échantillon i.i.d.).

Wald multidimensionnel

Soit $\theta \in \mathbb{R}^d$ et $\mathcal{H}_0 : g(\theta) = 0$, avec $g : \mathbb{R}^d \to \mathbb{R}^r$, $r \le d$. Supposons

$$\sqrt{n}(\widehat{\theta}_n - \theta) \rightsquigarrow \mathcal{N}_d(0, \mathbb{V}(\theta))$$

et g différentiable, de jacobienne $J_g(\theta)$ de rang plein r en θ_0. Posons

$$B(\theta) = J_g(\theta)\mathbb{V}(\theta)J_g(\theta)^\top.$$

Théorème — Wald multidimensionnel

Sous $\mathcal{H}_0$ et les hypothèses de régularité,

$$W_n = n\, g(\widehat{\theta}_n)^\top B(\widehat{\theta}_n)^{-1} g(\widehat{\theta}_n) \rightsquigarrow \chi^2_r.$$

On rejette si $W_n > \chi^2_{r,1-\alpha}$.

Le nombre de degrés de liberté est le nombre r de contraintes indépendantes, et non nécessairement la dimension d du paramètre.

Test du rapport de vraisemblance maximal

Pour tester une hypothèse simple ou composite $\mathcal{H}_0 : \theta \in \Theta_0$ contre $\mathcal{H}_1 : \theta \notin \Theta_0$, on compare le

maximum de vraisemblance sans contrainte au maximum sous $\mathscr{H}_0$. Avec $\widehat{\theta}$ l'EMV libre et $\widehat{\theta}_0$ l'EMV contraint, on pose

$$\begin{aligned}\Lambda_n &= \frac{\sup_{\theta\in\Theta_0} L(X;\theta)}{\sup_{\theta\in\Theta} L(X;\theta)} \\ &= \frac{L(X;\widehat{\theta}_0)}{L(X;\widehat{\theta})} \in (0,1], \\ D_n &= -2\log\Lambda_n = 2\big[\ell_n(\widehat{\theta}) - \ell_n(\widehat{\theta}_0)\big].\end{aligned}$$

On rejette $\mathscr{H}_0$ lorsque Λ_n est *petit*, ou, de façon équivalente, lorsque D_n est *grand*.

Théorème — Wilks

Dans un modèle paramétrique régulier, lorsque l'hypothèse nulle impose r contraintes régulières indépendantes et que le vrai paramètre est intérieur à l'espace admissible approprié, on a sous $\mathscr{H}_0$:

$$D_n = -2\log\Lambda_n \rightsquigarrow \chi^2_r.$$

Le test asymptotique de niveau α rejette pour $D_n > \chi^2_{r,1-\alpha}$.

Remarque. En présence d'un paramètre sur la frontière, d'une non-identifiabilité ou d'autres irrégularités, la loi limite peut ne pas être un χ^2 usuel.

Exemple : moyenne gaussienne

Soit $X_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu,\sigma^2)$, σ^2 connue, et $\mathscr{H}_0 : \mu = \mu_0$ contre $\mathscr{H}_1 : \mu \neq \mu_0$. On a $\widehat{\mu} = \overline{x}$, $\widehat{\mu}_0 = \mu_0$ et

$$D_n = \frac{n(\overline{x}-\mu_0)^2}{\sigma^2} = Z^2.$$

Sous $\mathscr{H}_0$, $D_n \sim \chi^2_1$ *exactement*, sans approximation asymptotique. Ici, test de Wald et test du rapport de vraisemblance donnent la même décision.

Exemple : modèle de Poisson

Soit $X_i \overset{\text{i.i.d.}}{\sim} \mathcal{P}(\lambda)$, $\lambda_0 > 0$, avec $\mathscr{H}_0 : \lambda = \lambda_0$ contre $\mathscr{H}_1 : \lambda \neq \lambda_0$. La vraisemblance est

$$L(x;\lambda) = \frac{\lambda^{\sum_i x_i} e^{-n\lambda}}{\prod_i x_i!}, \qquad \widehat{\lambda} = \overline{x}.$$

Lorsque $\overline{x} > 0$, on obtient

$$\begin{aligned}\Lambda_n &= \left(\frac{\lambda_0}{\overline{x}}\right)^{n\overline{x}} e^{n(\overline{x}-\lambda_0)}, \\ D_n &= 2n\left[\overline{x}\log\left(\frac{\overline{x}}{\lambda_0}\right) + \lambda_0 - \overline{x}\right].\end{aligned}$$

On prolonge par continuité si $\overline{x} = 0$, en convenant que $0\log 0 = 0$. Sous $\mathscr{H}_0$ et pour n grand, $D_n \rightsquigarrow \chi^2_1$. La région critique asymptotique est $R = [D_n > \chi^2_{1,1-\alpha}]$.

Wald, score et vraisemblance

Dans les modèles réguliers et sous $\mathscr{H}_0$, les statistiques de **Wald**, du **score** (multiplicateur de Lagrange) et du **rapport de vraisemblance** conduisent, après normalisation adaptée, à des lois limites χ^2_r. Ces trois tests ne sont pas nécessairement identiques sur échantillon fini. Le test du score n'exige que l'estimation sous $\mathscr{H}_0$; celui de Wald utilise l'estimateur libre ; le test du rapport de vraisemblance utilise les deux maxima.

À retenir : un résultat « significatif au seuil de 5 % » indique une incompatibilité des données avec $\mathscr{H}_0$ au sens de la règle du test. Il ne prouve ni $\mathscr{H}_1$, ni l'importance pratique de l'effet. Il faut interpréter ensemble l'effet estimé, l'incertitude (IC), la p-valeur et la puissance.

Quelques citations :

• Test statistique : réponse péremptoire à une question mal posée (anonyme).

• Mieux vaut une réponse approximative à la bonne question, souvent vague, qu'une réponse exacte à la mauvaise question, que l'on peut toujours formuler précisément (John Tuckey in Annals of Mathematical Statistics, 1962).

• Dieu aime sûrement presque autant les p-valeurs de 0,06 que celles de 0,05 (Ralph Rosnow et Robert Rosenthal, in American Psychologist, 1989).

• La différence entre « significatif » et « non significatif » n'est pas elle-même statistiquement significative (Andrew Gelman et Hal Stern, in The American Statistician, 2006).

• Citer une p-valeur fait partie du rituel scientifique : une aspersion d'eau bénite destinée à sanctifier l'analyse (William Cleveland, in Visualizing Data, 1993).