Statistique Mathématique

Chapitre 6: Tests statistiques

Mohamed Essaied Hamrita

Mastère de Recherche: Finance & Actuariat
IHEC Sousse

Décembre 2021

Institut des Hautes Etudes Commerciales de Sousse
معهد الدراسات العليا التجارية بسوسة

Soit $(x_1, x_2, \ldots, x_n)$ un n−échantillon *i.i.d* d'une variable aléatoire de densité $f(x, \theta)$ où θ est un vecteur de paramètres inconnus. On a besoin de décider entre deux hypothèses concernant ces paramètres. Une **hypothèse nulle**, notée H_0 et une hypothèse dite **hypothèse alternative**, notée H_1 ou H_a.

Soit $(x_1, x_2, \ldots, x_n)$ un n−échantillon *i.i.d* d'une variable aléatoire de densité $f(x, \theta)$ où θ est un vecteur de paramètres inconnus. On a besoin de décider entre deux hypothèses concernant ces paramètres. Une **hypothèse nulle**, notée H_0 et une hypothèse dite **hypothèse alternative**, notée H_1 ou H_a.

Ces hypothèses seront l'un des cas suivants:

$$H_0 : \theta = \theta_0 \text{ vs } H_1 : \theta \neq \theta_0$$

$$H_0 : \theta \geq \theta_0 \text{ vs } H_1 : \theta < \theta_0$$

$$H_0 : \theta \leq \theta_0 \text{ vs } H_1 : \theta > \theta_0$$

Soit $(x_1, x_2, \ldots, x_n)$ un n−échantillon *i.i.d* d'une variable aléatoire de densité $f(x, \theta)$ où θ est un vecteur de paramètres inconnus. On a besoin de décider entre deux hypothèses concernant ces paramètres. Une **hypothèse nulle**, notée H_0 et une hypothèse dite **hypothèse alternative**, notée H_1 ou H_a.

Ces hypothèses seront l'un des cas suivants:

$$H_0 : \ \theta = \theta_0 \ \text{ vs } H_1 : \ \theta \neq \theta_0$$

$$H_0 : \ \theta \geq \theta_0 \ \text{ vs } H_1 : \ \theta < \theta_0$$

$$H_0 : \ \theta \leq \theta_0 \ \text{ vs } H_1 : \ \theta > \theta_0$$

Dans le premier cas, on parle de test **bilatéral**, tan disque dans les deux autres cas on parle d'un test **unilatéral**.

La règle de décision du test comporte deux risques (erreurs):

- un risque de premier espèce, noté $\alpha = \mathbb{P}(D_1 | H_0)$: probabilité de commettre une erreur de première espèce;

La règle de décision du test comporte deux risques (erreurs):

- un risque de premier espèce, noté $\alpha = \mathbb{P}(D_1|H_0)$: probabilité de commettre une erreur de première espèce;
- un risque de deuxième espèce, noté $\beta = \mathbb{P}(D_0|H_1)$: probabilité de commettre une erreur de deuxième espèce

La règle de décision du test comporte deux risques (erreurs):

- un risque de premier espèce, noté $\alpha = \mathbb{P}(D_1|H_0)$: probabilité de commettre une erreur de première espèce;
- un risque de deuxième espèce, noté $\beta = \mathbb{P}(D_0|H_1)$: probabilité de commettre une erreur de deuxième espèce

La règle de décision du test comporte deux risques (erreurs):

- un risque de premier espèce, noté $\alpha = \mathbb{P}(D_1|H_0)$: probabilité de commettre une erreur de première espèce;
- un risque de deuxième espèce, noté $\beta = \mathbb{P}(D_0|H_1)$: probabilité de commettre une erreur de deuxième espèce

Le risque de première espèce α est choisi à priori. Toutefois le risque de deuxième espèce β dépend de l'hypothèse alternative H_1 et on ne peut le calculer que si on spécifie des valeurs particulières du paramètre dans l'hypothèse H_1 que l'on suppose vraie.
On définit la puissance d'un test par $\eta = 1 - \beta$.

Définition 1 (Région critique)

*La **région critique** est l'ensemble des valeurs de la statistique du test pour lesquelles l'hypothèse nulle est **rejetée**.*

$$W = \{(x_1, x_2, \ldots, x_n) \in \mathbb{R}^n \ : \ T(x_1, x_2, \ldots, x_n) > c\}$$

*où $T(x)$ est une réalisation du **test statistique** (variable aléatoire) et c est la **valeur critique**.*

Définition 1 (Région critique)

*La **région critique** est l'ensemble des valeurs de la statistique du test pour lesquelles l'hypothèse nulle est **rejetée**.*

$$W = \{(x_1, x_2, \ldots, x_n) \in \mathbb{R}^n \ : \ T(x_1, x_2, \ldots, x_n) > c\}$$

*où $T(x)$ est une réalisation du **test statistique** (variable aléatoire) et c est la **valeur critique**.*

La statistique du test $T(x)$ possède une distribution exacte ou une distribution asymptotique D sous l'hypothèse nulle.

$$T(x) \overset{H_0}{\sim} D \quad \text{ou} \quad T(x) \xrightarrow[H_0]{d} D$$

Définition 1 (Région critique)

*La **région critique** est l'ensemble des valeurs de la statistique du test pour lesquelles l'hypothèse nulle est **rejetée**.*

$$W = \{(x_1, x_2, \ldots, x_n) \in \mathbb{R}^n \; : \; T(x_1, x_2, \ldots, x_n) > c\}$$

*où $T(x)$ est une réalisation du **test statistique** (variable aléatoire) et c est la **valeur critique**.*

La statistique du test $T(x)$ possède une distribution exacte ou une distribution asymptotique D sous l'hypothèse nulle.

$$T(x) \overset{H_0}{\sim} D \quad \text{ou} \quad T(x) \xrightarrow[H_0]{d} D$$

Le complément de la région critique est la région de non rejet.

1. Vocabulaires
2. Risques de première et seconde espèce
3. Le test UPP au sens de Neyman-Pearson
4. Les tests asymptotiques

L'hypothèse simple est représentée par $H_0 : \theta = \theta_0$ contre $H_1 : \theta = \theta_1$. Le test uniformément le plus puissant (UPP) au sens de Neyman-Pearson est donné par la proposition suivante.

Proposition 1

Le test uniformément le plus puissant (UPP) au seuil α est définie par la région critique:

$$W = \{x_1, x_2, \ldots, x_n) \in \mathbb{R}^n \ : \ \frac{L(\mathbf{x},\ _0)}{L(\mathbf{x},\ _1)} \leq c\}; \quad \textit{où } c > 0$$

$\alpha = \mathbb{P}(D_1 | H_0) \in [0, 1]$ *et* $L()$ *est la vraisemblance associée à la suite* $(x_1, \ldots, x_n)$.

Exemple 1

Soit $(x_1, x_2, \ldots, x_{25}$ *un échantillon de taille* $n = 25$ *i.i.d issu d'une loi normale de moyenne* m *et de variance* $\sigma^2 = 4$*. On veut tester*
$H_0 : m = 15$ *vs* $H_1 : m > 15$.
1) Sachant que $\overline{X} = 14.25$*, déterminer la région critique du test UPP au sens de Neyman-Pearson pour un niveau* $\alpha = 5\%$.
2) Déterminer et représenter le graphique de la puissance du test pour des différentes valeurs de m_1.

Test de rapport de vraisemblance (LR test)

Définition 2

La statistique du test du rapport de vraisemblance est définie par:

$$LR = -2\Big(\ell(x_1, x_2, \ldots, x_n, \theta_0) - \ell(x_1, x_2, \ldots, x_n, \widehat{\theta})\Big)$$

où $\ell()$ est la fonction de log-vraisemblance, $\widehat{\theta}$ est l'estimateur de θ par la méthode de vraisemblance.

Test de rapport de vraisemblance (LR test)

Définition 2

La statistique du test du rapport de vraisemblance est définie par:

$$LR = -2\Big(\ell(x_1, x_2, \ldots, x_n, \theta_0) - \ell(x_1, x_2, \ldots, x_n, \widehat{\theta})\Big)$$

où $\ell()$ est la fonction de log-vraisemblance, $\widehat{\theta}$ est l'estimateur de θ par la méthode de vraisemblance.

- La distribution asymptotique de la statistique LR sous H_0 est:

$$LR \overset{H_0}{\sim} \chi^2(p)$$

p est le nombre de restrictions imposées (ou la taille du vecteur θ).
On rejette l'hypothèse nulle si $LR > \chi^2_{1-\alpha}(p)$

Exemple 2

Soit $X_i \overset{iid}{\sim} \mathcal{P}(\lambda)$. *On a la réalisation de taille* $n = 10$ *donnée par* $(5, 0, 1, 1, 0, 3, 2, 3, 4, 1)$. *Tester, au niveau* $\alpha = 5\%$, $H_0 : \lambda = 1.8$ *contre* $H_1 : \lambda \neq 1.8$.

On peut montrer que l'estimateur de λ par la méthode du MV est $\widehat{\lambda} = \overline{X}$.

$$\begin{aligned}
LR &= -2(\ell(\lambda_0) - \ell(\overline{X})) \\
&= -2\left(\sum_{i=1}^{n} X_i \ln(\lambda_0/\overline{X}) - n\lambda_0 + n\overline{X}\right) \\
&= -2(20\ln(1.8/2) - 18 + 20) = 0.2144
\end{aligned}$$

et puisque $\chi^2_{0.95}(1) = 3.8414 > LR$, on ne peut pas rejeter l'hypothèse nulle pour un seuil $\alpha = 5\%$.

Exercice 1

Soit le modèle linéaire $y_i = \theta + \varepsilon_i$ *où* $\varepsilon \overset{iid}{\sim} N(0, \frac{1}{9})$. *Un échantillon aléatoire de taille* $n = 10$ *a donné:*

$y_i = (2.11, 2.05, 2.27, 2.01, 1.98, 2.03, 2.00, 2.13, 2.10, 2.01)$.

1) Écrire la vraisemblance de la variable aléatoire y_i *et déduire un estimateur de* θ *par la méthode du MV.*

2) Tester, au seuil $\alpha = 5\%$ $H_0 : \theta = 1$ *contre* $H_1 : \theta \neq 1$.

Exercice 1

Soit le modèle linéaire $y_i = \theta + \varepsilon_i$ *où* $\varepsilon \overset{iid}{\sim} N(0, \frac{1}{9})$. *Un échantillon aléatoire de taille* $n = 10$ *a donné:*
$y_i = (2.11, 2.05, 2.27, 2.01, 1.98, 2.03, 2.00, 2.13, 2.10, 2.01)$.
1) Écrire la vraisemblance de la variable aléatoire y_i *et déduire un estimateur de* θ *par la méthode du MV.*
2) Tester, au seuil $\alpha = 5\%$ $H_0 : \theta = 1$ *contre* $H_1 : \theta \neq 1$.

1) $\mathbb{E}(y_i) = \mathbb{E}(\theta + \varepsilon_i) = \theta$ et $\mathbb{V}(y_i) = \mathbb{V}(\theta + \varepsilon_i) = \mathbb{V}(\varepsilon_i) = \sigma_\epsilon^2$. D'où $y_i \overset{iid}{\sim} N(\theta, \sigma_\epsilon^2)$. Soit $f(y_i) = (2\pi\sigma^2)^{-\frac{1}{2}} \exp\left(-\frac{1}{2\sigma^2}(y_i - \theta)^2\right)$.
Donc, $\ell(\theta|y_i) = -\frac{n}{2}(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^n (y_i - \theta)^2$.
La condition nécessaire, nous donne:
$s(\theta|y_i) = \frac{\partial \ell}{\partial \theta} = 0 \Longrightarrow \widehat{\theta} = \overline{Y}$.

```r
y=c(2.11, 2.05, 2.27, 2.01, 1.98, 2.03, 2.00, 2.13,
2.10, 2.01)
l=function(x,theta,sig2){
n=length(x)
-n/2*log(2*pi*sig2)-1/(2*sig2)*sum((x-theta)^2)
}
theta0=2; y.bar=mean(y); sig2=1/9
(LR=-2*(l(y,theta0,sig2)-l(y,y.bar,sig2)))
```

```
[1] 0.42849
```

```r
qchisq(0.95,1)
```

```
[1] 3.841459
```

Exercice 2

Soit le modèle linéaire: $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ *où* ε *est une suite aléatoire indépendante de la loi normale de paramètres 0 et 1. On désire tester* $H_0 : \beta_0 + \beta_1 = 1$.

1) Écrire la densité de la variable aléatoire y_i *puis déduire sa vraisemblance en logarithme.*

2) Déterminer le vecteur score, $s(\beta|y_i)$ *et vérifier que* $\widehat{\beta}_0 = \overline{y} - \widehat{\beta}_1 \overline{x}$ *et* $\widehat{\beta}_1 = \frac{s_{xy}}{s_{xx}}$ *sont les estimateurs de* β_0 *et* β_1 *par la méthode du MV.* $\left(s_{ab} = \frac{\sum (a-\overline{a})(b-\overline{b})}{n}\right)$.

3) Effectuer le test LR pour tester $H_0 : \beta_0 + \beta_1 = 1$.

On donne $X'X = \begin{pmatrix} 12 & 14 \\ 14 & 32 \end{pmatrix}$ *et* $X'y = \begin{pmatrix} 25.5 \\ 37.4 \end{pmatrix}$.

Les données brutes sont données dans le diapo suivant.

```
     y e x
1  2.6 1 1
2  2.2 1 2
3  2.2 1 2
4  1.6 1 2
5  1.9 1 1
6  3.3 1 0
7  1.9 1 0
8  2.5 1 1
9  1.2 1 1
10 4.3 1 4
11 0.6 1 0
12 1.2 1 0
```

1) $\varepsilon_i \overset{iid}{\sim} N(0,1) \Longrightarrow y_i \overset{iid}{\sim} N(\beta_0 + \beta_1 x_i, \sigma)$, donc

$$f(y_i|x_i) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(y_i - \beta_0 - \beta_1 x_i)^2\right)$$

$$\Longrightarrow logL = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2$$

1) $\varepsilon_i \overset{iid}{\sim} N(0,1) \Longrightarrow y_i \overset{iid}{\sim} N(\beta_0 + \beta_1 x_i, \sigma)$, donc

$$f(y_i|x_i) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(y_i - \beta_0 - \beta_1 x_i)^2\right)$$

$$\Longrightarrow logL = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2$$

2) Le vecteur $s(\beta|y_i, x_i) = \left(\frac{\partial \ell}{\partial \beta_0}, \frac{\partial \ell}{\partial \beta_1}\right)'$.

$$s(\beta|y_i, x_i) = \begin{pmatrix} \sum(y_i - \beta_0 - \beta_1 x_i) \\ \sum x_i(y_i - \beta_0 - \beta_1 xi) \end{pmatrix}$$

```
logL=function(beta,x,y){ n=length(y)
-n/2*log(2*pi)-1/2*sum((y-beta[1]-beta[2]*x)^2)
}
optim(fn=logL, par=c(0,0),control=list(fnscale=-1),
y=y,x=x)[c("par","value")]

$par
[1] 1.5552790 0.4882482

$value
[1] -14.61077
```

```
logL=function(beta,x,y){ n=length(y)
-n/2*log(2*pi)-1/2*sum((y-beta[1]-beta[2]*x)^2)
}
optim(fn=logL, par=c(0,0),control=list(fnscale=-1),
y=y,x=x)[c("par","value")]

$par
[1] 1.5552790 0.4882482

$value
[1] -14.61077
```

```
(beta1=cov(x,y)/var(x)); (beta0=mean(y)-beta1*mean(x))

[1] 0.4882979
[1] 1.555319
```

$$LR = -2(\ell(\beta|H_0) - ell(\widehat{\beta})$$

```
ll1=logL(c(beta0,beta1),x,y)
n=length(y)
ll0=-n/2*log(2*pi)-1/2*sum((y-(1-beta1)-beta1*x)^2)
(LR=-2*(ll0-ll1)); qchisq(0.95,1)

[1] 13.06964
[1] 3.841459
```

Test de Wald

On suppose qu'un vecteur θ de taille p est estimé par la méthode du MV. On veut tester l'hypothèse nulle pour laquelle r équations (peuvent être non linéaires) sont satisfaites; $g(\theta) = 0$ où g est une fonction définie de $\mathbb{R}^p \longrightarrow \mathbb{R}^r$. La statistique de Wald est définie par:

$$W = g(\widehat{\theta})' \left(J(\theta) I^{-1}(\widehat{\theta}) J(\theta)' \right)^{-1} g(\widehat{\theta})$$

avec $J()$ est le jacobien de la fonction g, $I(\widehat{\theta}) = \mathbb{E}\left(-H\right) = \mathbb{V}(s(\widehat{\theta}|x_i))$.
Sous l'hypothèse nulle, on a: $W \sim \chi^2(r)$.
Pour un seuil α, l'hypothèse nulle est rejetée lorsque $W > \chi^2_{1-\alpha}(r)$

Exemple 3

Reprenons l'exemple 2 (loi de poisson).

On rappelle que $\widehat{\lambda} = \overline{X} = 2$ et $\lambda_0 = 0$. $g(\lambda) = \lambda - 1.8$. D'où $J(\lambda) = \frac{\partial g}{\partial \lambda} = 1$ et $I(\widehat{\lambda}) = -\mathbb{E}\left(H(\lambda|x)\right) = \mathbb{V}\left(s(\lambda|x)\right) = \mathbb{V}\left(\frac{\sum x_i}{\lambda}\right) = \frac{n}{\lambda}$.
Soit $I(\widehat{\lambda} = \frac{n}{\overline{X}} = 5$.
$W = (2 - 1.5)5(2 - 1.8) = 0.2 < \chi^2_{0.95}(1) = 3.8414$, donc on ne peut pas rejeter l'hypothèse nulle pour un seuil $\alpha = 5\%$.

Exemple 3

Reprenons l'exemple 2 (loi de poisson).

On rappelle que $\widehat{\lambda} = \overline{X} = 2$ et $\lambda_0 = 0$. $g(\lambda) = \lambda - 1.8$. D'où $J(\lambda) = \frac{\partial g}{\partial \lambda} = 1$ et $I(\widehat{\lambda}) = -\mathbb{E}\left(H(\lambda|x)\right) = \mathbb{V}\left(s(\lambda|x)\right) = \mathbb{V}\left(\frac{\sum x_i}{\lambda}\right) = \frac{n}{\lambda}$.
Soit $I(\widehat{\lambda} = \frac{n}{\overline{X}} = 5$.
$W = (2-1.5)5(2-1.8) = 0.2 < \chi^2_{0.95}(1) = 3.8414$, donc on ne peut pas rejeter l'hypothèse nulle pour un seuil $\alpha = 5\%$.
Remarque: dans le cas où le vecteur θ est de taille un, la statistique de Wald est le carré de la statistique t, i.e $W = t^2$.

Exercice 3

Reprenons l'exercice 2 et testons $H_0 : \beta_0 + \beta_1 = 1$.

$g(\beta) = \beta_0 + \beta_1 - 1$, et $J(\beta) = \left(\frac{\partial g}{\partial \beta_0}, \frac{\partial g}{\partial \beta_1}\right) = (1, 1)$.

En outre, la matrice des variances de β est estimée par:

$$\widehat{\mathbb{V}}(\widehat{\beta}) = \sigma_\epsilon^2 \left(X'X\right)^{-1} = \begin{pmatrix} 8/47 & -7/94 \\ -7/94 & 3/47 \end{pmatrix} = I^{-1}(\beta).$$

$g(\widehat{\beta}) = 1.5552 + 0.4882 - 1 = 1.0434$. Donc la statistique de Wald est donnée par: $W =$

$$1.0434 \left((1,1) \begin{pmatrix} 8/47 & -7/94 \\ -7/94 & 3/47 \end{pmatrix} \begin{pmatrix} 1 \\ 1 \end{pmatrix}\right)^{-1} \times 1.0434 = 12.79203.$$

$W > \chi^2_{0.95}(1) = 3.3.8414$, donc, au seuil $\alpha = 5\%$, on doit rejeter l'hypothèse nulle.

LM test

Le test de multiplicateur de Lagrange (LM test), appelé aussi test de score est un programme d'optimisation de la vraisemblance sous contraintes en appliquant la méthode de Lagrange.

$$\max \ell(\theta|x_i) \quad \text{s.c} \quad \theta = \theta_0$$

Le lagrangien s'écrit: $L = \ell(\theta|x_i) - \lambda(\theta - \theta_0)$.
La statistique LM est définie comme suit:

$$LM = s'(\theta|x_i)I^{-1}(\theta|x_i)s(\theta|x_i) \overset{H_0}{\sim} \chi^2(q)$$

où $s()$ est le vecteur score, $I()$ est l'information de Fisher et q est le nombre de contraintes.

Reprenons l'exemple 2 et appliquons le test du multiplicateur de Lagrange.
$s(\lambda_0|x_i) = -n + \frac{\sum x_i}{\lambda_0} = -10 + \frac{20}{1.8} = 1.11$ et $I(\lambda) = \frac{n}{\lambda}$.
$LM = 1.11^2 \frac{1.8}{10} = 0.222 < \chi^2_{0.95}(1) = 3.8$, on ne peut pas rejeter l'hypothèse nulle avec un seuil $\alpha = 5\%$.

```
z=c(5, 0, 1, 1, 0, 3, 2, 3, 4, 1)
log_l=function(lambda,x){n=length(x)
-n*lambda+log(lambda)*sum(x)-sum(log(factorial(x)))
}
l=function(x,theta,sig2){
n=length(x)
-n/2*log(2*pi*sig2)-1/(2*sig2)*sum((x-theta)^2)
}
theta0=2; y.bar=mean(y); sig2=1/9
(LR=-2*(l(y,theta0,sig2)-l(y,y.bar,sig2)))

[1] 1.6875

qchisq(0.95,1)

[1] 3.841459
```