M1. IMSA, STATISTIQUE DU RISQUE

M. BOUTAHAR

March 24, 2025

Chapitre 1. INTRODUCTION AUX TESTS STATISTIQUES

1.1 Tests d'hypothèses

Ce sont des techniques d'aide à la décision dans des situations où on est amené à effectuer un choix entre plusieurs décisions possibles, sans que l'on dispose de l'information suffisante pour rendre ce choix parfaitement sûr.

H_0 désignera l'hypothèse nulle dont on désire savoir si elle vraie ou fausse. L'hypothèse l'alternative est souvent notée H_1.

les deux types d'erreur

Erreurs de première et de seconde espèces

L'erreur est dite de **première espèce** lorsque l'on rejette H_0 alors que celle-ci est vraie, on la note α,
$(\alpha = P_{H_0}(\textit{rejeter } H_0))$.

L'erreur est dite de **seconde espèce** lorsque l'on rejette H_1 alors que celle-ci est vraie, on la note β,
$(\beta = P_{H_1}(\textit{rejeter } H_1))$.

La **puissance** d'un test est donnée par : puissance = 1 - β.

Vraie Décidée	H_0	H_1
H_0	Bonne décision	α
H_1	β	Bonne décision

1.2. Région critique du test

On utilise en général une statistique $\mathcal{T}$ pour décider l'acceptation ou le rejet de H_0.

Definition 1

Soit P_{H_0} = la loi de $\mathcal{T}$ sous H_0, $\alpha \in]0, 1[$, et R_α telle que

$$P_{H_0}(\mathcal{T} \in R_\alpha) = \alpha,$$

On appelle R_α la région critique du test associée au niveau nominal α.

En général on prends $\alpha = 0.01$ ou $\alpha = 0.05$ et si $\mathcal{T} \in R_\alpha$ alors on rejette H_0 avec un risque α.

1.3. Décision à l'aide de la p-value (ou probabilité critique ou p-valeur)

En général on définit la p-value ou (probabilité critique) par

• Si le test est bilatéral alors

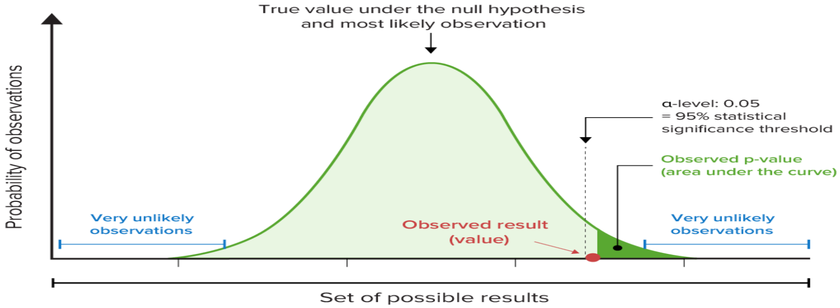
$$p - value = 2 * P_{H_0}\left(\mathcal{T} \geq |\mathcal{T}_{\text{observée}}|\right).$$

• Si le test est unilatéral (on teste par exemple $H_0 : m_1 \geq m_2$ ou l'indépendance entre les variables X et Y), alors

$$p - value = P_{H_0}\left(\mathcal{T} \geq \mathcal{T}_{\text{observée}}\right).$$

Les tables statistiques permettent un encadrement de la p-value, cependant les logiciels de statistique donnent la valeur exacte de la p-value dans le cas ou la loi exacte ou approchée de $\mathcal{T}$ sous H_0 est une loi classique (Normale, Student, Khi-deux Fisher,...).

Décision : Si $p - value \leq \alpha$ (le risque d'erreur du test) alors on rejette H_0.


True value under the null hypothesis
and most likely observation
α-level: 0.05
= 95% statistical
significance threshold
Observed p-value
(area under the curve)
Probability of observations
Very unlikely
observations
Observed result
(value)
Very unlikely
observations
Set of possible results

Chapitre 2. TESTS DE CORRELATION DES VARIABLES QUANTITATIVES

2.1. Test de Pearson

Coefficient de corrélation:

$$\rho = cor(X, Y) = \frac{cov(X, Y)}{\sigma_X \sigma_Y},$$

$cov(X, Y) = E\{(X - E(X))(Y - E(Y))\} =$
$E(XY) - E(X)E(Y), \sigma_X^2 = var(X) = cov(X, X).$

Soient $X = (X_1, ..., X_n)$ et $Y = (Y_1, ..., Y_n)$ deux échantillons.

On souhaite tester la présence d'une liaison entre X et Y.

$$\begin{cases} H_0 : \ X \text{ et } Y \text{ sont non corrélées (hypothèse nulle)} \\ contre \\ H_1 : X \text{ et } Y \text{ sont corrélées (hypothèse alternative)} \end{cases} \tag{1}$$

Le coefficient de corrélation linéaire empirique de Pearson est donnée par:

$$r = \hat{\rho} = \hat{cor}(X, Y) = \frac{\hat{cov}(X, Y)}{\hat{\sigma}_X \hat{\sigma}_Y}, \tag{2}$$

$$\hat{cov}(X, Y) = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y}), \hat{\sigma}_X^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2.$$

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, \overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i,$$

Si $|r|$ est proche de 0 alors le lien entre X et Y est faible.

Si $|r|$ est proche de 1 alors le lien entre X et Y est fort.

Theorem 2

Sous H_0

$$\sqrt{n}\ r \xrightarrow{\mathcal{L}} N(0,1).$$

On fixe α (0.01 ou 0.05) et on lit dans la table de Gauss le quantile u tel que

$$P(\mathcal{N}(0,1) \in [-u, u]) = 1 - \alpha,$$

donc la zone de rejet est $\mathcal{R} =]-\infty, -u] \bigcup [u, +\infty[$.
Décision : Si $\sqrt{n}r \in \mathcal{R}$ alors on rejette H_0.

2.2. Test de Spearman

Le coefficient de corrélation de Spearman permet de capturer des liaisons non linéaires. Il peut aussi être appliqué à des variables qualitatives ordinales (par exemple notation des étudiants par A, B,C,D,E,F).

Definition 3

Le coefficient de corrélation de Spearman est le coefficient de corrélation de Pearson calculé sur les rangs des deux échantillons.

$$\rho_s = \rho_s(X, Y) = cor(R_x, R_y) \tag{3}$$

R_x est la variable rang de X.

On peut estimer $\rho(X, Y)$ par

$$r_s = \hat{\rho}_s = \hat{\rho}(X, Y) = c\hat{o}r(R_x, R_y).$$

Pour tester l'hypothèse (5) on utilise r_s.

Theorem 4

Sous H_0

$$\sqrt{n}r_s \xrightarrow{\mathcal{L}} N(0,1).$$

2.3. Test de Kendall

Le coefficient de corrélation de Kendall est basé sur les mesures de concordance et de discordance de X et Y.
Le coefficient de corrélation des rangs de Kendall est défini par

$$T = T(X,Y) = (Pc - Pd)\,,$$

$$Pc = P\left((X - X')(Y - Y') > 0\right),$$

$$Pd = P\left((X - X')(Y - Y') < 0\right).$$

(X', Y') a la même loi que (X, Y) mais indépendant.

Proposition 5

Le coefficient empirique de Kendall est donné par

$$\tau = \hat{T} = \frac{4}{n(n-1)} \sum_{i=2}^{n} \sum_{j=1}^{i-1} sign((X_i - X_j)(Y_i - Y_j)) - 1 \quad (4)$$

$$sign(x) = \begin{cases} 1 & si\ x > 0 \\ 0 & sinon \end{cases}$$

Pour tester l'hypothèse (5) on utilise la statistique

$$K_n = \tau \sqrt{\frac{9n(n-1)}{2(2n+5)}}.$$

Theorem 6

Sous H_0, K_n suit approximativement une loi N(0,1).

Chapitre 3. TESTS DE LIAISON ENTRE VARIABLES QUANTITATIVES ET QUALITATIVES

On suppose que la variable Y est quantitative, et que X est qualitative avec r modalités $x_1, ..., x_r$ d'effectifs $n_1, ..., n_r$. On souhaite effectuer le test

$$\left\{ \begin{array}{l} H_0 : \text{ Aucune liaison entre } X \text{ et } Y. \\ \textit{contre} \\ H_1 : \text{ Présence d'une liaison entre } X \text{ et } Y \end{array} \right. \quad (5)$$

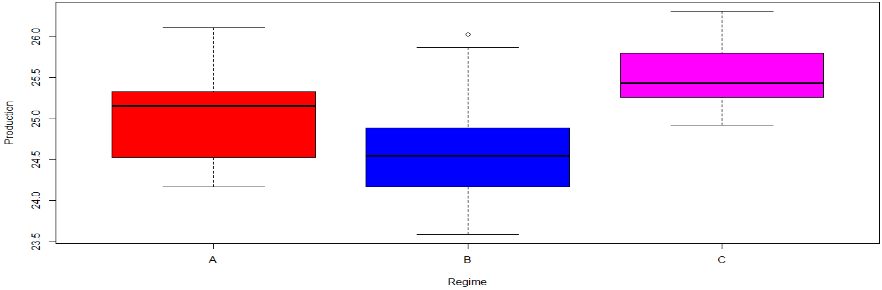
3.1. Rapport de corrélation

Approche graphique

On peut tracer les histogrammes et les boxplots de $(Y \mid X = x_i)$ la similarité des graphes indique l'absence de liaison.

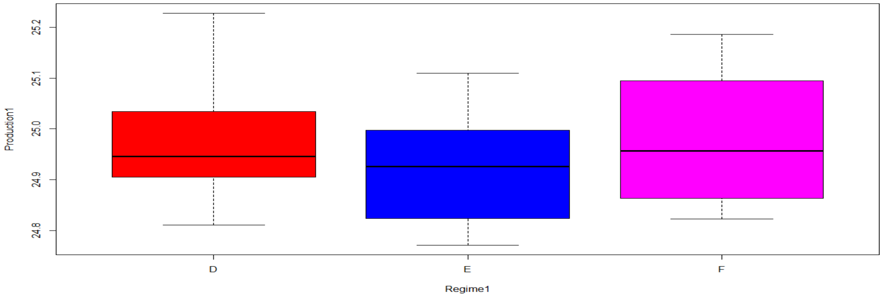
Exemple. On souhaite comparer les montants des sinistres de 3 marques de voitures A,B et C. On sélectionne 30 sinistres, avec 10 sinistres pour chaque marque. Les montants annuels (en K€) des 3 groupes sont donnés par:

Groupe A	24.17	25.58	25.18	26.11	24.50	24.61	25.17	24.53	25.33	25.14
Groupe B	24.81	24.17	24.41	23.59	25.87	23.83	26.03	24.89	24.32	24.69
Groupe C	26.31	25.12	25.54	25.50	25.37	25.29	24.92	26.15	25.80	25.26



Les moyennes et variances empiriques modifiées des 3 groupes sont données par: $\overline{Y}_A = 25.03$, $\overline{Y}_B = 24.66$, $\overline{Y}_C = 25.53$, et $S_A'^2 = 0.34$, $S_B'^2 = 0.63$, $S_C'^2 = 0.20$.

Un autre exemple avec 3 autres marques



Les moyennes empiriques modifiées des 3 groupes sont données par: $\overline{Y}_D = 24.98, \overline{Y}_E = 24.92, \overline{Y}_F = 24.97.$

On considère le rapport de corrélation

$$R_{Y/X} = \frac{var(E(Y/X))}{var(Y)} \tag{6}$$

A noter que si X et Y sont indépendantes alors $R_{Y/X} = 0$.
On considère alors l'estimateur empirique de $R_{Y/X}$

$$\rho_{Y/X} = \frac{\frac{1}{n}\sum_{i=1}^{r} n_i(\overline{Y}_i - \overline{Y})^2}{\hat{\sigma}_Y^2} \tag{7}$$

$$\overline{Y}_i = \hat{E}(Y/X = x_i) = \frac{1}{n_i}\sum_{k=1}^{n} Y_k 1_{X_k = x_i}.$$

-$\rho_{Y/X} = 0$ implique que $\overline{Y}_1 = ... = \overline{Y}_r = \overline{Y}$, donc absence de liaision.

Plus $\rho_{Y/X}$ est grand plus la liaison est forte.

Posons $E(Y/X = x_i) = m_i, 1 \leq i \leq r$ et supposons que $var(Y/X) = \sigma_0^2$. Tester l'hypothèse (5) revient à tester

$$\left\{ \begin{array}{ll} H_0 : & m_1 = m_2 = ... = m_r = m \\ \text{contre} & \\ H_1 : & \exists i \neq j, \text{ tels que } m_i \neq m_j \ . \end{array} \right. \tag{8}$$

Si on suppose que les $Y_i, 1 \leq i \leq n$ suivent une loi de Gauss alors sous l'hypothèse H_0 de (22) on a $(Y_i, 1 \leq i \leq n)$ est une suite i.i.d suivant une loi $N(m, \sigma_0^2)$. on a alors (voir le paragraphe suivant) sous H_0 : $\frac{n-1}{r-1}\rho_{Y/X} \rightsquigarrow F(r-1, n-1)$, loi de Fisher-Snedecor à $(r-1, n-1)$ degrés de liberté.

3.2. Test ANOVA

Notons

$$E_i = \{Y_k, 1 \leq k \leq n, \ \ (X_k, Y_k) = (x_i, Y_k)\}$$

On peut réécrire E_i comme suit

$$E_i = \{Y_{i,1}, ..., Y_{i,n_i}\},$$

$$n_i = card((X_k, Y_k), 1 \leq k \leq n \ \ X_k = x_i).$$

On veut tester

$$\begin{cases} H_0 : \quad m_1 = m_2 = ... = m_r = m \\ \ \text{contre} \\ H_1 : \quad \exists i \neq j, \text{ tels que } m_i \neq m_j \ . \end{cases} \tag{9}$$

On définit la variance intraclasse

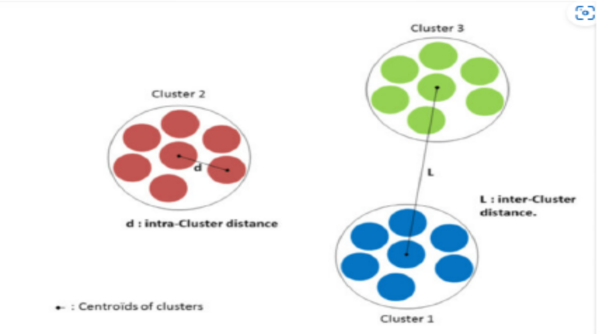
$$S^2_{\text{intra}} = \frac{1}{n-r}\sum_{i=1}^{r}\sum_{j=1}^{n_i}(Y_{i,j} - \overline{Y_i})^2, \tag{10}$$

$$\overline{Y}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} Y_{i,j}. \tag{11}$$

On définit la variance interclasse

$$S^2_{\text{inter}} = \frac{1}{r-1}\sum_{i=1}^{r} n_i(\overline{Y}_i - \overline{Y})^2 \ , \tag{12}$$

$$\overline{Y} = \frac{1}{n}\sum_{i=1}^{r}\sum_{j=1}^{n_i} Y_{i,j}, n = \sum_{i=1}^{r} n_i. \tag{13}$$


Cluster 3
Cluster 2
d
L
L : inter-Cluster distance.
d : intra-Cluster distance
: Centroïds of clusters
Cluster 1

Souvent on note la Somme des Carrés des Ecarts

$$SCE = \sum_{i=1}^{r} n_i(\overline{Y}_i - \overline{Y})^2$$

et la Somme des Carrés des Résidus:

$$SCR = \sum_{i=1}^{r} \sum_{j=1}^{n_i} (Y_{i,j} - \overline{Y_i})^2.$$

On a alors

$$S^2_{\text{inter}} = \frac{SCE}{r-1}, S^2_{\text{intra}} = \frac{SCR}{n-r},$$

On montre la décomposition classique de la variance

$$V = \sum_{i=1}^{r} \sum_{j=1}^{n_i} (Y_{i,j} - \overline{Y})^2 = SCE + SCR.$$

Sous H_0 on a $(Y_i, 1 \leq i \leq n)$ est une suite i.i.d suivant une loi $N(m, \sigma_0^2)$, ceci implique que

$$\frac{V}{\sigma_0^2} \rightsquigarrow \mathcal{X}^2(n-1), \frac{SCE}{\sigma_0^2} \rightsquigarrow \mathcal{X}^2(r-1), \frac{SCR}{\sigma_0^2} \rightsquigarrow \mathcal{X}^2(n-r).$$

On utilise la statistique de test

$$F = \frac{S^2_{\text{inter}}}{S^2_{\text{intra}}} = \frac{SCE/(r-1)}{SCR/(n-r)},$$

alors on a sous H_0 : $F \rightsquigarrow F(r-1, n-r)$, loi de Fisher-Snedecor à $(r-1, n-r)$ degrés de liberté

3.3. Test de Kruskal-Wallis

C'est un test non paramétrique qui ne nécessite pas l'hypothèse de normalité des données. On considère toujours le test (22). La statistique de test de Kruskal-Wallis est donnée par

$$H = (n-1)\frac{\sum_{i=1}^{r} n_i(\overline{R}_i - \overline{R})^2}{\sum_{i=1}^{r}\sum_{j=1}^{n_i}(R_{i,j} - \overline{R})^2} \tag{14}$$

$R_{i,j}$ est le rang de $Y_{i,j}$ parmi toutes les n observations.

$$\overline{R}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} R_{i,j}, \overline{R} = \frac{(n+1)}{2}.$$

Proposition 7

Sous H_0 la statistique H peut être simplifiée comme suit:

$$H = \frac{12}{n(n+1)} \sum_{i=1}^{r} n_i \overline{R}_i^2 - 3(n+1)$$

Proposition 8

On a sous H_0 :

$$H \xrightarrow{\mathcal{L}} \mathcal{X}^2(r-1).$$

3.4. Cas particuliers de deux modalités

Cas Gaussien

Si dans (22) on a $r = 2$ c'est à dire que la variable X a seulement deux modalités (variable de Bernoulli), alors dans le cas d'échantillons Gaussiens on peut utiliser la statistique de test de Welch.

Soient $E_1 = (Y_{1,1}, ..., Y_{1,n_1})$ et $E_2 = (Y_{2,1}, ..., Y_{2,n_2})$ les deux échantillons obtenus selon les valeurs de $X \quad (x_1$ ou $x_2)$.

On suppose que $Y_{1,i} \rightsquigarrow N(m_1, \sigma_1)$, $Y_{2,i} \rightsquigarrow \mathcal{N}(m_2, \sigma_2)$ pour tout i.

On désire effectuer le test

$$\begin{cases} H_0 : & m_1 = m_2 = m \\ \text{contre} & \\ H_1 : & m_1 \neq m_2. \end{cases} \tag{15}$$

On utilise $D = \overline{Y}_1 - \overline{Y}_2$, qui est un bon indicateur de l'écart entre les deux moyennes.

On utilise alors la statistique de test

$$\mathcal{T}_w = \frac{D}{\sqrt{\frac{S_1'^2}{n_1} + \frac{S_2'^2}{n_2}}}.$$

$$S_k'^2 = \frac{1}{n-1} \sum_{i=1}^{n_k} (Y_{k,i} - \overline{Y}_k)^2, k = 1, 2.$$

Sous H_0 on a $\mathcal{T}_w \rightsquigarrow t(n_1 + n_2 - 2)$ loi de Student à $(n_1 + n_2 - 2)$ degrés de liberté.

Cas non Gaussien

Si $Y_{k,i}$ ne suivent pas une loi normale, alors on peut utiliser la statistique de test de Mann-Whitney Wilcoxon suivante

$$\mathcal{T}_m = \frac{U_{1,2} - \frac{n_1 n_2}{2}}{\sqrt{n_1 n_2 (n_1 + n_2 + 1)/12}}. \qquad (16)$$

avec

$$U_{1,2} = \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} 1_{\{Y_{1,i} \leq Y_{2,j}\}}, \qquad (17)$$

On a sous H_0

$$\mathcal{T}_m \xrightarrow{\mathcal{L}} N(0,1).$$

Chapitre 4. TESTS D'INDEPENDANCE

4.1. Tests d'indépendance de $\mathcal{X}^2$

- Soit P une loi de probabilité d'un couple aléatoire (X, Y) à valeurs dans $\mathbb{R}^2$, on veut tester l'indépendance de X et Y en se basant sur un échantillon $(X_i, Y_i)_{1\leq i\leq n}$.
- Ceci revient à tester

$$\begin{cases} H_0: & P(dx, dy) = P_1(dx)P_2(dy). \\ \text{contre} & \\ H_1: & P(dx, dy) \neq P_1(dx)P_2(dy). \end{cases} \tag{18}$$

où P_1 et P_2 sont respectivement les lois marginales de X et Y.

- On construit alors une partition uniforme de l'ensemble des valeurs prises par X : $A_1, ..., A_p$, et de l'ensemble des valeurs prises par Y : $B_1, ..., B_r$, (uniforme veut dire que $P_1(A_j) \approx 1/p, P_2(B_j) \approx 1/r$).
- On définit alors

$$N_{i,j} = \sum_{k=1}^{n} 1_{\{(X_k, Y_k) \in A_i \times B_j\}}.$$

- On pose

$$d_n = \sum_{i=1}^{p} \sum_{j=1}^{r} \frac{(N_{i,j} - np_{i,j})^2}{np_{i,j}};$$

- On montre que

$$d_n \xrightarrow{\mathcal{L}} \mathcal{X}^2(pr-1)$$

- Sous H_0, $p_{i,j} = p_{i,\bullet}p_{\bullet,j}$, où $p_{i,\bullet} = P_1(A_i), p_{\bullet,j} = P_2(B_j)$.

- En pratique les probabilités marginales sont inconnues et donc on les estime par les estimateurs empiriques

$$\widehat{p}_{i,\bullet} = \frac{1}{n}\sum_{j=1}^{r} N_{i,j},\, \widehat{p}_{\bullet,j} = \frac{1}{n}\sum_{i=1}^{p} N_{i,j}.$$

- On montre alors que

$$\mathcal{X}^2 = \sum_{i=1}^{p}\sum_{j=1}^{r}\frac{(N_{i,j} - n\widehat{p}_{i,\bullet}\ \widehat{p}_{\bullet,j})^2}{n\ \widehat{p}_{i,\bullet}\ \widehat{p}_{\bullet,j}} \xrightarrow{\mathcal{L}} \mathcal{X}^2((p-1)(r-1)).$$

4.2. Mesure d'inensité (V Cramer)

Pour mesurer l'intensité de dépendance on peut utiliser le V de Cramer suivant

$$V = \sqrt{\frac{\mathcal{X}^2}{n(\min(p, r) - 1)}} \tag{19}$$

On montre que $0 \leq V \leq 1$.

Remaques

1. Le V de Cramer n'est utile qu'en présence de dépendance entre X et Y, en effet on a sous H_0:

$$V \xrightarrow{P} 0,$$

par conséquent la décision doit être prise par $\mathcal{X}^2$ et non par V.

2. Lien avec le tableau de contingence classique: en effet on a

$$n\widehat{p}_{i,\bullet} = n_{i,\bullet}, n\widehat{p}_{\bullet,j} = n_{\bullet,j}$$

par conséquent

$$\mathcal{X}^2 = \sum_{i=1}^{p} \sum_{j=1}^{r} \frac{(N_{i,j} - n_{i,\bullet} n_{\bullet,j}/n)^2}{n_{i,\bullet} n_{\bullet,j}/n}.$$

3. La statistique de $\mathcal{X}^2$ peut être appliquée quelque soient le le type de variables X et Y (qualitative ou quantitative).

Présentation sous forme d'un tableau de contingence

Y / X	y_1	$\dots$	y_j	$\dots$	y_r	Total
x_1	$n_{1,1}$	$\dots$	$n_{1,j}$	$\dots$	$n_{1,r}$	$n_{1,\bullet}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_i	$n_{i,1}$	$\dots$	$n_{i,j}$	$\dots$	$n_{i,r}$	$n_{i,\bullet}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_p	$n_{p,1}$	$\dots$	$n_{p,j}$	$\dots$	$n_{p,r}$	$n_{p,\bullet}$
Total	$n_{\bullet,1}$	$\dots$	$n_{\bullet,j}$	$\dots$	$n_{\bullet,r}$	n

$$n_{i,\bullet} = \sum_{j=1}^{r} n_{i,j},\; n_{\bullet,j} = \sum_{i=1}^{p} n_{i,j},\; n = \sum_{i=1}^{p}\sum_{j=1}^{r} n_{i,j}$$

Chapitre 5. TESTS D'ADEQUATION

5.1. Test de $\mathcal{X}^2$ pour les variables discrètes

- Le but du test d'adéquation de $\mathcal{X}^2$ est de décider si l' échantillon observé $x = (X_1(\omega), ..., X_n(\omega))$ peut être considéré comme issu d'une loi bien spécifiée P_0.
- Effectuer le test $H_0 : P = P_0$ contre $H_1 : P \neq P_0$, ou plus généralement $H_0 : P \in \{P_\theta, \theta \in \Theta\}$ contre $H_1 : P \notin \{P_\theta, \theta \in \Theta\}$.

• Supposons que $card(E) < \infty$; c'est à dire que $E = \{x_1, ..., x_d\}$ et que l'on veut tester $H_0 : P = P_0$ avec $P_0(x_j) = p_j, j = 1, ..., d$ contre $H_1 : P \neq P_0$.

• Pour effectuer ce test on introduit la mesure de divergence (ou distance de Khi-deux) entre la loi théorique P_0 et la loi empirique P_n définie par:

$$P_n = \frac{1}{n}\sum_{k=1}^{n}\delta_{X_k} \tag{20}$$

où δ_{X_k} est la mesure de Dirac au point X_k.

$$D(P_n, P_0) = \sum_{j=1}^{d} \frac{n}{p_j}\left(\frac{N_{j,o}}{n} - p_j\right)^2; \tag{21}$$

où $N_{j,o} = \sum_{k=1}^{n} 1_{\{X_k = x_j\}}$ est l'effectif observé à x_j.

- Si H_0 est vraie alors par la loi des grands nombres $\frac{N_{j,o}}{n}$ s'approche de p_j et donc $D(P_n, P_0)$ est très petite.

On a aussi

$$
\begin{aligned}
D(P_n, P_0) &= \sum_{j=1}^{d} \frac{(N_{j,o} - np_j)^2}{np_j} \\
&= \sum_{j=1}^{d} \frac{(N_{j,o} - N_{j,t})^2}{N_{j,t}}
\end{aligned}
$$

$N_{j,t} = np_j$ (effectifs théoriques attendus), $1 \leq j \leq d$.

- La région critique est de la forme $W = \{D(P_n, P_0) \geq k\}$, k est tel que le test soit de niveau α.

Theorem 9

On suppose que $p_j > 0$, alors sous H_0

$$D(P_n, P_0) \xrightarrow{\mathcal{L}} \mathcal{X}^2(d-1).$$

5.2. Test de Kolmogorov-Smirnov pour les variables continues • Le test de Kolmogorov-Smirnov est un test non paramétrique qui porte sur la fonction de répartition de la loi testée.

• C'est à dire qu'en se basant sur une suite i.i.d. $(X_i)_{1 \leq i \leq n}$ de fonction de répartition inconnue F, on teste $H_0 : F = F_0$ contre $H_1 : F \neq F_0$.

Definition 10

On définit la fonction de **répartition empirique** de $(X_i)_{1\leq i\leq n}$, l'application $F_n^* : \mathbb{R} \to [0,1]$ définie par

$$F_n^*(x) = \frac{1}{n}\sum_{j=1}^{n} 1_{\{X_j \leq x\}}.$$

- On définit les statistiques de Kolmogorov-Smirnov suivantes :

$$K_n = \sup_{x\in\mathbb{R}} \left| F_n^*(x) - F_0(x) \right|, K_n^+ = \sup_{x\in\mathbb{R}} (F_n^*(x) - F_0(x)),$$

$$K_n^- = \sup_{x\in\mathbb{R}} (F_0(x) - F_n^*(x)).$$

Theorem 11

(Glivenko-Cantelli)
On suppose que $(X_i)_{1\leq i\leq n}$ est une suite i.i.d. de fonction de répartition $F(x)$, alors sous H_0, on a

$$K_n \stackrel{p.s.}{\to} 0.$$

- Les deux théorèmes suivants sont admis (voir la démonstration dans Van Der Waerden).

Theorem 12

On suppose que $(X_i)_{1\leq i\leq n}$, *est une suite i.i.d., que* $F_0(x)$ *est continue, alors sous* H_0

$$\sqrt{n}K_n \xrightarrow{\mathcal{L}} Z;$$

** la fonction de répartition de* Z *est donnée par*

$$F_Z(x) = 1 - 2\sum_{k=1}^{\infty}(-1)^{k-1}e^{-2k^2x^2}.$$

Theorem 13

On suppose que $(X_i)_{1\leq i\leq n}$*, est une suite i.i.d., que* $F_0(x)$ *est continue, alors sous* H_0

$$\sqrt{n}K_n^+ \xrightarrow{\mathcal{L}} Z^+;$$

la fonction de répartition de Z^+ *est donnée par*
$F_{Z^+}(x) = 1 - e^{-2x^2}$.

Chapitre 6. TESTS DE CONTRTAINTES

Soit $\theta \in \Theta \subset \mathbb{R}^p$ un paramètre d'intérêt. Soit Θ_0 un sous ensemble Θ.

On souhaite effetcuer le test

$$\left\{ \begin{array}{l} H_0 : \quad \theta \in \Theta_0 \\ \text{ contre} \\ H_1 : \quad \theta \notin \Theta_0. \end{array} \right. \tag{22}$$

6.1. Test de rapport de vraisemblance

Definition 14

Pour tester H_0 contre H_1 un test basé sur la région critique $W = \{x, RV) < k\}$, où

$$RV = \frac{\sup_{\theta \in \Theta_0} L(\theta, x)}{\sup_{\theta \in \Theta} L(\theta, x)} \qquad (23)$$

est appelé test de **rapport de vraisemblance**.

$L(\theta, x)$ est la vraisemblance de $(X_1, ..., X_n)$

Remarques. 1. Plus le rapport est petit plus on aura tendance à rejeter H_0.

2. La loi de RV dépends de l'échantillon $(X_1, ..., X_n)$, à une constante près cette loi peut être approchée par une loi de $\mathcal{X}^2(r)$, avec $r = dim(\Theta) - dim(\Theta_0)$.

6.2. Test de Wald

Soit k une application de $\mathbb{R}^p \to \mathbb{R}^r$ et c un vecteur $(r, 1)$. On veut tester la contrainte

$$H_0 : k(\theta) \quad = \quad c \text{ contre } H_1 : k(\theta) \neq \ c \tag{24}$$

avec $dim(\theta) = p$ et $r < p$.

Definition 15

Pour tester la contrainte (24) la statistique de Wald est donnée par

$$W_n = {}^t(k(\hat{\theta}) - c) \left(K \hat{V}(\hat{\theta})\, {}^tK \right)^{-1} (k(\hat{\theta}) - c) \tag{25}$$

où $\hat{\theta}$ est l'estimateur du maximum de vraisemblance de θ,

$$K = \frac{\partial k}{\partial \theta'} = \begin{pmatrix} \frac{\partial k_1}{\partial \theta_1} & \cdots & \frac{\partial k_1}{\partial \theta_p} \\ \vdots & \vdots & \vdots \\ \frac{\partial k_r}{\partial \theta_1} & \cdots & \frac{\partial k_r}{\partial \theta_p} \end{pmatrix}$$

et $\hat{V}(\hat{\theta})$ est un estimateur de la variance de $\hat{\theta}$.

Theorem 16

On a sous H_0

$$W_n \xrightarrow{\mathcal{L}} \mathcal{X}^2(r) \tag{26}$$

6.3. Test du Score ou Multiplicateur de Lagrange

Soit g une application de $\mathbb{R}^p \to \mathbb{R}^r$ et c un vecteur $(r, 1)$. On veut tester la contrainte

$$H_0 : g(\theta) \;=\; c \text{ contre } H_1 : g(\theta) \neq c \qquad (27)$$

avec $dim(\theta) = p$ et $r < p$.

Definition 17

Pour tester la contrainte (27) la statistique de Score ou Multiplicateur de Lagrange est donnée par

$$LM_n = {}^tS(\tilde{\theta})I^{-1}(\tilde{\theta})S(\tilde{\theta}) \tag{28}$$

où $\tilde{\theta}$ est l'estimateur du maximum de vraisemblance de θ sous la contrainte $g(\theta) = c$,

$$S(\theta) = \frac{\partial \log L(\theta)}{\partial \theta}$$

est le vecteur score,

$$I(\theta) = -E\left(\frac{\partial^2 \log L(\theta)}{\partial\theta\partial\theta'}\right)$$

est la matrice d'information de Fisher, et $L(\theta)$ est la vraisemblance.

Theorem 18

On a sous H_0

$$LM_n \xrightarrow{\mathcal{L}} \mathcal{X}^2(r) \tag{29}$$