# INTRODUCTION AUX TESTS STATISTIQUES

MATTHIEU KOWALSKI

## 1. INTRODUCTION À LA THÉORIE DES TESTS

La théorie des tests consiste à tester si une hypothèse est vraie. On dit qu'on teste $H_0$ (l'hypothèse nulle) contre $H_1$ (hypothèse alternative). Dans cette situation, il est possible de commettre deux erreurs : conclure que $H_0$ est vraie alors qu'en réalité c'est $H_1$ qui est vérifiée, et vice versa. On définie ces deux erreurs :

**Définition 1** (Erreur de première et seconde espèce)**.** *On appelle erreur de première espèce ou erreur de type $I$ la quantité*

$$\alpha = \mathbb{P}\{\textit{accepter}H_0 | H_1 \textit{ est vraie}\} \ .$$

*On appelle erreur de seconde espèce ou erreur de type $II$ la quantité*

$$\beta = \mathbb{P}\{\textit{accepter}H_1 | H_0 \textit{ est vraie}\} \ .$$

Pour construire un test statistique au risque $\alpha$, on fixe l'erreur de première espèce à $\alpha$, avec $\alpha$ «petit» (de l'ordre de $5\%$, $1\%$ voire moins). Une fois cette erreur fixée, on n'a plus aucun contrôle sur l'erreur de seconde espèce $\beta$ !

On réalise un test selon les étapes suivantes :

(1) Définition des hypothèse $H_0$ et $H_1$. Cela implique de faire un choix : quelle est l'hypothèse privilégiée, suivant l'erreur qu'on veut contrôler.

(2) Choix du niveau $\alpha$ (petit).

(3) Calculer la statistique du test. Ce calcul se fait à l'aide des observations statistiques à notre disposition, et du test choisis.

(4) Conclusion au vu de l'échantillon selon la règle de décision associée au test. La conclusion étant le rejet ou l'acceptation de $H_0$.

**Remarque 1.** *Si, à l'issu du test, on accepte $H_1$, alors l'erreur commise est $\alpha$ (par définition !). Puisque $\alpha$ est fixé «petit», l'erreur commise est «petite».*

*Si on rejette $H_0$, l'erreur commise est $\beta$ (toujours par définition !). Cette erreur n'est pas contrôlée et peut être très grande : l'acceptation de $H_0$ ne permet donc pas,* a priori, *de conclure que $H_0$ est effectivement vraie.*

## 2. UN TEST PARAMÉTRIQUE : LE TEST DE FISHER-STUDENT

On dispose de deux échantillons $(X_1, \dots, X_{n_X})$ et $(Y_1, \dots, Y_{n_Y})$ (de tailles éventuellement différente) de loi $P_1$ et $P_2$ indépendantes de moyenne $\mu_X$ et $\mu_Y$ et de variance $\sigma_X^2$ et $\sigma_Y^2$ finies. Autrement dit

$$\forall i \ \ \mathbb{E}\{X_i\} = \mu_1 \quad \text{Var}\{X_i\} = \sigma_1^2 \ ,$$

---

CES NOTES SONT TRÈS LARGEMENT INSPIRÉES DES NOTES MANUSCRITES DE SÉBASTIENT LOUSTAU, MERCI À LUI !

et

$$\forall i \ \mathbb{E}\{Y_i\} = \mu_2 \quad \text{Var}\{Y_i\} = \sigma_2^2 \ .$$

On note

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n_X} X_i \quad \text{et} \quad \bar{Y} = \frac{1}{n}\sum_{i=1}^{n_Y} Y_i$$

les moyennes empiriques et

$$S_X^2 = \frac{1}{n}\sum_{i=1}^{n_X}(X_i - \bar{X})^2 \quad \text{et} \quad S_Y^2 = \frac{1}{n}\sum_{i=1}^{n_Y}(Y_i - \bar{Y})^2$$

les variances empiriques.

On veut tester

$$\begin{aligned} H_0 &: \text{les moyennes sont égales, ie } \mu_X = \mu_Y \\ H_1 &: \text{les moyennes sont différentes, ie } \mu_X \neq \mu_Y \ . \end{aligned}$$

Pour cela, on suppose que $X_i$ et $Y_i$ suivent des lois normales. Le test s'effectue en deux étapes : on teste d'abord l'égalité des variances, et si ce test est accepté, alors on teste l'égalité des moyennes.

### 2.1. **1ère étape : test de Fisher d'égalité des variances.**

2.1.1. *Hypothèses.* Pour comparer les moyennes, il faut pouvoir supposer que les variances sont égales. On va donc tester

$$H_0 \ : \sigma_X^2 = \sigma_Y^2 \quad \text{contre} \quad H_1 \ : \sigma_X^2 \neq \sigma_Y^2 \ .$$

2.1.2. *Statistique de test.* Pour cela, on va calculer la statistique de test

$$F = \frac{n_X(n_Y - 1)}{n_Y(n_X - 1)}\frac{S_X^2}{S_Y^2} \ .$$

On s'arrangera en pratique pour avoir $F \geq 1$, i.e. $S_X^2 \geq S_Y^2$.

Pour $\alpha$ fixé, on défini la quantité $\mathcal{F}_{n-1,n-1;1-\alpha/2}$, qui est le fractile d'ordre $1 - \alpha/2$ d'une loi de *Fisher*, telle que

$$\mathbb{P}\{Z > \mathcal{F}_{n_X-1,n_Y-1;1-\alpha/2}\} = \frac{\alpha}{2} \ .$$

où $Z \sim \mathcal{F}(n_X - 1, n_Y - 1)$, i.e. $Z$ suit une loi de Fisher de paramètres $(n_X - 1)$, $(n_Y - 1)$.

La quantité $\mathcal{F}_{n_X-1,n_Y-1;1-\alpha/2}$ est donnée par les tables statistiques de la loi de Fisher.

2.1.3. *Règle de décision.* Ainsi, la règle de décision du test de Fisher est :

- Si $F < \mathcal{F}_{n_X-1,n_Y-1;1-\alpha/2}$, alors on **accepte** $H_0$. On peut donc supposer que les variances sont égales. On passe alors à l'étape 2 : le test de Student d'égalité des moyennes.
- Si $F > \mathcal{F}_{n_X-1,n_Y-1;1-\alpha/2}$, alors on **rejette** $H_0$. On peut conclure (avec un risque $\alpha$ petit de se tromper) que les variances ne sont pas égales, et on ne peut pas poursuivre.

**Remarque 2.** *Si les deux échantillons sont de même taille $n_X = n_Y = n$, la statistique de test devient*

$$F = \frac{S_X^2}{S_Y^2} .$$

*et suit une loi de Fisher de paramètres $(n-1)$, $(n-1)$*

### 2.2. **2ème étape : test de Student d'égalité des moyennes.**

2.2.1. *Hypothèses.* On va tester

$$H_0 \ : \mu_X^2 = \mu_Y^2 \quad \text{contre} \quad H_1 \ : \mu_X^2 \neq \mu_Y^2 .$$

2.2.2. *Statistique de test.* Pour cela, on va calculer la statistique de test

$$T = \frac{\sqrt{n_X + n_Y - 2}}{\sqrt{\frac{1}{n_X} + \frac{1}{n_y}}} \frac{\left|\bar{X} - \bar{Y}\right|}{\sqrt{n_X S_X^2 + n_Y S_Y^2}} .$$

Pour $\alpha$ fixé, on défini la quantité $t_{(n_X+n_Y-2);1-\alpha/2}$, qui est le fractile d'ordre $1-\alpha/2$ d'une loi de *Student-t*, telle que

$$\mathbb{P}\{Z > t_{(n_X+n_Y-2);1-\alpha/2}\} = \frac{\alpha}{2} ,$$

où $Z \sim t((n_X + n_Y - 2))$, i.e. $Z$ suit une loi de Student-t à $(n_X + n_Y - 2)$ degrés de liberté.

La quantité $t_{(n_X+n_Y-2);1-\alpha/2}$ est donnée par les tables statistiques de la loi de Student-t.

2.2.3. *Règle de décision.* Ainsi, la règle de décision du test de Student est :

- Si $T < t_{(n_X+n_Y-2);1-\alpha/2}$, alors on **accepte** $H_0$. On peut donc supposer que les moyennes sont égales. Le risque de se tromper $\beta$ est inconnu et peut être élevé.
- Si $T > t_{(n_X+n_Y-2);1-\alpha/2}$, alors on **rejette** $H_0$. On peut conclure (avec un risque $\alpha$ petit de se tromper) que les moyennes ne sont pas égales.

**Remarque 3.** *Si les deux échantillons sont de même taille $n_X = n_Y = n$, la statistique de test devient*

$$T = \frac{\sqrt{n-1}\left|\bar{X} - \bar{Y}\right|}{\sqrt{S_X^2 + S_Y^2}} .$$

*et suit une loi de Student-t à $2(n-1)$ degrés de libertés.*

## 3. TESTS DU $\chi^2$

Le test du $\chi^2$ est *non paramétrique* : contrairement au test de Fisher-Student, on ne va pas tester des paramètres (comme la moyenne ou la variance), mais des distributions. Le test du $\chi^2$ va permettre de répondre aux questions :

(1) Est-ce que la population que j'observe suit une loi *donnée* à l'avance ?

(2) Si j'observe deux variables statistiques $(X_1, \ldots, X_n)$ et $(Y_1, \ldots, Y_n)$ (observées sur le même ensemble d'individus), sont-elles indépendantes ?

### 3.1. **Test du $\chi^2$ d'adéquation.**

3.1.1. *Hypothèses.* On dispose d'un $n$-échantillon $(X_1, \dots, X_n)$ réparti en $K$ classes $(\mathcal{C}_1, \dots, \mathcal{C}_K)$. On se donne une loi théorique $P$ dont les paramètres son connus ou estimés. On note $r$ le nombre de paramètres que l'on aura estimé (éventuellement $r = 0$ si la loi est parfaitement connue). On va tester

$$H_0 \ : \text{ la population est distribuée selon } P$$

contre

$$H_1 \ : \text{ la population n'est pas distribuée selon } P \ .$$

3.1.2. *Statistique de test.* Connaissant la loi $P$ (par hypothèse), on connaît les *effectifs théoriques* pour chacune des classes $\mathcal{C}_k$, $k \in \{1, \dots, K\}$. On note

$$p_k = \mathbb{P}\{X \in \mathcal{C}_k\}$$

la probabilité théorique qu'une réalisation de la v.a. $X$ appartienne à $\mathcal{C}_k$. Ainsi, l'effectif théorique $n_k$ de la classe $\mathcal{C}_k$ est donné par

$$n_k = np_k \ .$$

On note $N_k$ les *effectifs observés* sur notre échantillon $(X_1, \dots, X_n)$ dans chacune des classes $\mathcal{C}_k$.

La statistique de test que l'on va calculer est :

$$D^2 = \sum_{k=1}^{K} \frac{(N_k - np_k)^2}{np_k} \ .$$

Pour $\alpha$ fixé, on défini la quantité $\chi_{K-1-r;1-\alpha}$, qui est le fractile d'ordre $1-\alpha$ d'une loi de $\chi^2$, telle que

$$\mathbb{P}\{Z > \chi_{K-1-r;1-\alpha}\} = 1 - \alpha \ .$$

où $Z \sim \chi^2(K-1-r)$, i.e. $Z$ suit une loi du $\chi^2$ à $K-1-r$ degrés de liberté. On rappelle que $K$ est le nombre de classe et $r$ le nombre de paramètres estimés.

La quantité $\chi_{K-1-r;1-\alpha}$ est donnée par les tables statistiques de la loi du $\chi^2$.

3.1.3. *Règle de décision.* Ainsi, la règle de décision du test du $\chi^2$ d'adéquation est :

- Si $D^2 < \chi_{K-1-r;1-\alpha}$, alors on **accepte** $H_0$. On peut donc supposer que la population est issue de la loi $P$. Le risque de se tromper $\beta$ est inconnu et peut être élevé.
- Si $D^2 > \chi_{K-1-r;1-\alpha}$, alors on **rejette** $H_0$. On peut conclure (avec un risque $\alpha$ petit de se tromper) que la population ne suit pas la loi $P$.

3.1.4. *Présentation pratique des calculs.* On résume toutes les informations précédente dans un tableau :

| Classes | effectifs observés | effectifs théoriques | résidus |
|---|---|---|---|
| $\mathcal{C}_1$ | $N_1$ | $np_1$ | $\frac{(N_1-np_1)^2}{np_1}$ |
| $\mathcal{C}_2$ | $N_2$ | $np_2$ | $\frac{(N_2-np_2)^2}{np_2}$ |
| $\vdots$ | | | |
| $\mathcal{C}_k$ | $N_k$ | $np_k$ | $\frac{(N_k-np_k)^2}{np_k}$ |
| $\vdots$ | | | |
| $\mathcal{C}_K$ | $N_K$ | $np_K$ | $\frac{(N_K-np_K)^2}{np_K}$ |
| Total | $N$ | $N$ | $D^2$ |

## 3.2. **Test du $\chi^2$ d'indépendance.**

3.2.1. *Hypothèses.* Soit $(X_1, \ldots, X_n)$ et $(Y_1, \ldots, Y_n)$ deux variables statistiques *qualitatives* observés sur un échantillon de $n$ individus. On suppose que les facteurs $X$ et $Y$ prennent respectivement les modalités $c_1, \ldots, c_r$ et $d_1, \ldots, d_s$. Le but est d'étudier l'interaction entre les deux facteurs $X$ et $Y$. On va tester

$$H_0 \ : \text{ les facteurs } X \text{ et } Y \text{ sont indépendants}$$

contre

$$H_1 \ : \text{ les facteurs } X \text{ et } Y \text{ interagissent .}$$

3.2.2. *Statistique de test.* Pour cela, on va utiliser le *tableau de contingence*. Ce tableau résume comment une caractéristique dépend d'une autre et donne le nombre d'individus possédant simultanément la modalité $c_i$ de la variable $X$ et la modalité $d_j$ de la variable $Y$. Un tel tableau se présente sous la forme suivante, pour un échantillon de taille $n$ :

| X \ Y | $d_1$ | $\ldots$ | $d_j$ | $\ldots$ | $d_s$ | |
|---|---|---|---|---|---|---|
| $c_1$ | $N_{11}$ | | $N_{1j}$ | | $N_{1J}$ | $N_{1\bullet}$ |
| $\vdots$ | | | $\vdots$ | | | |
| $c_i$ | $N_{i1}$ | $\ldots$ | $N_{ij}$ | $\ldots$ | $N_{iJ}$ | $N_{i\bullet}$ |
| $\vdots$ | | | $\vdots$ | | | |
| $c_r$ | $N_{I1}$ | | $N_{Ij}$ | | $N_{IJ}$ | $N_{I\bullet}$ |
| | $N_{\bullet 1}$ | | $N_{\bullet j}$ | | $N_{\bullet J}$ | $N$ |

où $N_{ij}$ désigne le nombre de fois où $X$ à pris la modalité $c_r$ et $Y$ la modalité $d_j$. Autrement dit, $N_{ij}$ représente le nombre d'individus qui possède à la fois la caractéristique $c_i$ et la caractéristique $c_j$. On définit les quantités

$$N_{i\bullet} = \sum_{j=1}^{s} N_{ij} \quad N_{\bullet j} = \sum_{i=1}^{r} N_{ij}$$

qui représentent respectivement le nombre d'individus qui possèdent la modalité $c_i$ et le nombre d'individus avec la modalité $c_j$.

Comme pour les test du $\chi^2$ d'adéquation, on connaît les *effectifs observés* $N_{ij}$, qu'on va comparer aux *effectifs théoriques*. On note

$$p_{ij} = \mathbb{P}\{X = c_i, Y = d_j\}, \quad p_{i\bullet} = \mathbb{P}\{X = c_i\}, \quad p_{\bullet j} = \mathbb{P}\{Y = d_j\} .$$

Si les variables $X$ et $Y$ sont indépendantes, alors on a $p_{ij} = p_{i\bullet}p_{\bullet j}$, et l'effectif théorique de chaque cas est donné par $np_{i\bullet}p_{\bullet j}$. Cependant, les $p_{i\bullet}$ et $p_{\bullet j}$ ne sont pas connus. On les estime alors par

$$\hat{p}_{i\bullet} = \frac{N_{i\bullet}}{N} \quad \text{et} \quad \hat{p}_{\bullet j}\frac{N_{\bullet j}}{N} .$$

La statistique de test que l'on va calculer est :

$$D^2 = \sum_{i=1}^{r}\sum_{j=1}^{s} \frac{(N_{ij} - \frac{N_{i\bullet}N_{\bullet j}}{n})^2}{\frac{N_{i\bullet}N_{\bullet j}}{n}} .$$

Pour $\alpha$ fixé, on défini la quantité $\chi_{(r-1)(s-1);1-\alpha}$, qui est le fractile d'ordre $1-\alpha$ d'une loi de $\chi^2$, telle que

$$\mathbb{P}\{Z > \chi_{(r-1)(s-1);1-\alpha}\} = 1 - \alpha .$$

où $Z \sim \chi^2((r-1)(s-1))$, i.e. $Z$ suit une loi du $\chi^2$ à $(r-1)(s-1)$ degrés de liberté.
La quantité $\chi_{(r-1)(s-1);1-\alpha}$ est donnée par les tables statistiques de la loi du $\chi^2$.

3.2.3. *Règle de décision.* Ainsi, la règle de décision du test du $\chi^2$ d'indépendance est :

- Si $D^2 < \chi_{K-1-r;1-\alpha}$, alors on **accepte** $H_0$. On peut donc supposer que la population est issue de la loi $P$. Le risque de se tromper $\beta$ est inconnu et peut être élevé.
- Si $D^2 > \chi_{(r-1)(s-1);1-\alpha}$, alors on **rejette** $H_0$. On peut conclure (avec un risque $\alpha$ petit de se tromper) que les deux facteurs $X$ et $Y$ sont indépendant.

3.2.4. *présentation pratique des calculs.* On construit d'abord le tableau des effectifs observés

| X \ Y | $d_1$ | $\dots$ | $d_j$ | $\dots$ | $d_s$ | |
|---|---|---|---|---|---|---|
| $c_1$ | $N_{11}$ | | $N_{1j}$ | | $N_{1J}$ | $N_{1\bullet}$ |
| $\vdots$ | | | $\vdots$ | | | |
| $c_i$ | $N_{i1}$ | $\dots$ | $N_{ij}$ | $\dots$ | $N_{iJ}$ | $N_{i\bullet}$ |
| $\vdots$ | | | $\vdots$ | | | |
| $c_r$ | $N_{I1}$ | | $N_{Ij}$ | | $N_{IJ}$ | $N_{I\bullet}$ |
| | $N_{\bullet 1}$ | | $N_{\bullet j}$ | | $N_{\bullet J}$ | $N$ |

On calcul ensuite le tableau des effectifs théoriques

| X \ Y | $d_1$ | $\dots$ | $d_j$ | $\dots$ | $d_s$ |
|---|---|---|---|---|---|
| $c_1$ | $\frac{N_{1\bullet}N_{\bullet 1}}{N}$ | | $\frac{N_{1\bullet}N_{\bullet j}}{N}$ | | $\frac{N_{1\bullet}N_{s\bullet}}{N}$ |
| $\vdots$ | | | $\vdots$ | | |
| $c_i$ | $\frac{N_{i\bullet}N_{\bullet 1}}{N}$ | $\dots$ | $\frac{N_{i\bullet}N_{\bullet j}}{N}$ | $\dots$ | $\frac{N_{i\bullet}N_{\bullet s}}{N}$ |
| $\vdots$ | | | $\vdots$ | | |
| $c_r$ | $\frac{N_{r\bullet}N_{\bullet 1}}{N}$ | | $\frac{N_{r\bullet}N_{\bullet j}}{N}$ | | $\frac{N_{r\bullet}N_{\bullet s}}{N}$ |

Et enfin le tableau des résidus

| X \ Y | $d_1$ | $\dots$ | $d_j$ | $\dots$ | $d_s$ |
|---|---|---|---|---|---|
| $c_1$ | $\frac{\left(N_{11}-\frac{N_{1\bullet}N_{\bullet 1}}{N}\right)^2}{\frac{N_{1\bullet}N_{\bullet 1}}{N}}$ | | $\frac{\left(N_{1j}-\frac{N_{1\bullet}N_{\bullet j}}{N}\right)^2}{\frac{N_{1\bullet}N_{\bullet j}}{N}}$ | | $\frac{\left(N_{1s}-\frac{N_{1\bullet}N_{\bullet s}}{N}\right)^2}{\frac{N_{1\bullet}N_{\bullet s}}{N}}$ |
| $\vdots$ | | | $\vdots$ | | |
| $c_i$ | $\frac{\left(N_{i1}-\frac{N_{i\bullet}N_{\bullet 1}}{N}\right)^2}{\frac{N_{i\bullet}N_{\bullet 1}}{N}}$ | $\dots$ | $\frac{\left(N_{ij}-\frac{N_{i\bullet}N_{\bullet j}}{N}\right)^2}{\frac{N_{i\bullet}N_{\bullet j}}{N}}$ | $\dots$ | $\frac{\left(N_{is}-\frac{N_{i\bullet}N_{\bullet s}}{N}\right)^2}{\frac{N_{i\bullet}N_{\bullet s}}{N}}$ |
| $\vdots$ | | | $\vdots$ | | |
| $c_r$ | $\frac{\left(N_{r1}-\frac{N_{r\bullet}N_{\bullet 1}}{N}\right)^2}{\frac{N_{r\bullet}N_{\bullet 1}}{N}}$ | | $\frac{\left(N_{rj}-\frac{N_{r\bullet}N_{\bullet j}}{N}\right)^2}{\frac{N_{r\bullet}N_{\bullet j}}{N}}$ | | $\frac{\left(N_{rs}-\frac{N_{r\bullet}N_{\bullet s}}{N}\right)^2}{\frac{N_{r\bullet}N_{\bullet s}}{N}}$ |