STATISTIQUE : TESTS D'HYPOTHESES

Préparation à l'Agrégation Bordeaux 1

Année 2012 - 2013

Jean-Jacques Ruch

Table des Matières

CHAPITRE I

Généralités sur les tests

1. Introduction

Un *test d'hypothèse* est un procédé d'inférence permettant de contrôler (accepter ou rejeter) à partir de l'étude d'un ou plusieurs échantillons aléatoires, la validité d'hypothèses relatives à une ou plusieurs populations. Les méthodes de l'inférence statistique nous permettent de déterminer, avec une probabilité donnée, si les différences constatées au niveau des échantillons peuvent être imputables au hasard ou si elles sont suffisamment importantes pour signifier que les échantillons proviennent de populations vraisemblablement différentes.

On distinguera deux classes de tests :

- Les *tests paramétriques* requierent un modèle à fortes contraintes (normalité des distributions ou approximation normale pour des grands échantillons). Ces hypothèses sont d'autant plus difficiles à vérifier que les effectifs étudiés sont plus réduits.
- Les *tests non paramétriques* sont des tests dont le modèle ne précise pas les conditions que doivent remplir les paramètres de la population dont a été extrait l'échantillon. Il n'y a pas d'hypothèse de normalité au préalable.

Les tests paramétriques, quand leurs conditions sont remplies, sont les plus puissants que les tests non paramétriques. Les tests non paramétriques s'emploient lorsque les conditions d'applications des autres méthodes ne sont pas satisfaites, même après d'éventuelles transformation de variables. Ils peuvent s'utiliser même pour des échantillons de taille très faible.

On ditingue les tests suivant :

- Le *test de conformité* consiste à confronter un paramètre calculé sur l'échantillon à une valeur pré-établie. Les plus connus sont certainement les tests portant sur la moyenne, la variance ou sur les proportions. On connaît la loi théorique en général la loi normale. Par exemple, dans un jeu de dés à 6 faces, on sait que la face 3 a une probabilité de $1/6$ d'apparaître. On demande à un joueur de lancer (sans précautions particulières) 100 fois le dé, on teste alors si la fréquence d'apparition de la face 3 est compatible avec la probabilité $1/6$. Si ce n'est pas le cas, on peut se poser des questions sur l'intégrité du dé.
- Le *test d'ajustement ou d'adéquation* consiste à vérifier la compatibilité des données avec une distribution choisie a priori. Le test le plus utilisé dans cette optique est le test d'ajustement à la loi normale, qui permet ensuite d'appliquer un test paramétrique .
- Le *test d'homogénéité ou de comparaison* consiste à vérifier que $K(K \geq 2)$ échantillons (groupes) proviennent de la même population ou, cela revient à la même chose, que la distribution de la variable d'intérêt est la même dans les K échantillons. Y a-t-il une différence entre le taux de glucose moyen mesuré pour deux échantillons d'individus ayant reçu des traitements différents ?
- Le *test d'indépendance ou d'association* consiste à éprouver l'existence d'une liaison entre 2 variables. Les techniques utilisées diffèrent selon que les variables sont qualitatives nominales, ordinales ou quantitatives. Est-ce que la distribution de la couleur des yeux observée dans la population française fréquences est indépendante du sexe des individus ?

2. Principe des tests

2.a. Méthodologie.

Le principe des tests d'hypothèse est de poser une hypothèse de travail et de prédire les conséquences de cette hypothèse pour la population ou l'échantillon. On compare ces prédictions avec les observations et l'on conclut en acceptant ou en rejetant l'hypothèse de travail à partir de règles de décisions objectives. Définir les hypothèses de travail, constitue un élément essentiel des tests d'hypothèses de même que vérifier les conditions d'application de ces dernières.

Différentes étapes doivent être suivies pour tester une hypothèse :

(1) définir l'hypothèse nulle, notée H_0, à contrôler ;
(2) choisir une statistique pour contrôler H_0 ;
(3) définir la distribution de la statistique sous l'hypothèse « H_0 est réalisée » ;
(4) définir le niveau de signification du test α et la région critique associée ;
(5) calculer, à partir des données fournies par l'échantillon, la valeur de la statistique ;
(6) prendre une décision concernant l'hypothèse posée .

2.b. Hypothèse nulle - hypothèse alternative.

L'*hypothèse nulle* notée H_0 est l'hypothèse que l'on désire contrôler : elle consiste à dire qu'il n'existe pas de différence entre les paramètres comparés ou que la différence observée n'est pas significative et est due aux fluctuations d'échantillonnage. Cette hypothèse est formulée dans le but d'être rejetée.

L'*hypothèse alternative* notée H_1 est la "négation" de H_0, elle est équivalente à dire « H_0 est fausse ». La décision de rejeter H_0 signifie que H_1 est réalisée ou H_1 est vraie.

Remarque : Il existe une dissymétrie importante dans les conclusions des tests. En effet, la décision d'accepter H_0 n'est pas équivalente à « H_0 est vraie et H_1 est fausse ». Cela traduit seulement l'opinion selon laquelle, il n'y a pas d'évidence nette pour que H_0 soit fausse. Un test conduit à rejeter ou à ne pas rejeter une hypothèse nulle jamais à l'accepter d'emblée.

La nature de H_0 détermine la façon de formuler H_1 et par conséquent la nature unilatérale ou bilatérale du test. On parle de *test bilatéral* lorsque l'hypothèse alternative se "décompose en deux parties". Par exemple si H_0 consiste à dire que la population estudiantine avec une fréquence de fumeurs p est représentative de la population globale avec une fréquence de fumeurs p_0, on pose alors : $H_0 : p = p_0$ et $H_1 : p \neq p_0$. Le test sera bilatéral car on considère que la fréquence p peut être supérieure ou inférieure à la fréquence p_0.

On parle de *test unilatéral* lorsque l'hypothèse alternative se "compose d'une seule partie". Par exemple si l'on fait l'hypothèse que la fréquence de fumeurs dans la population estudiantine p est supérieure à la fréquence de fumeurs dans la population p_0, on pose alors $H_0 : p = p_0$ et $H_1 : p > p_0$. Le test sera unilatéral car on considère que la fréquence p ne peut être que supérieure à la fréquence p_0.

Il aurait été possible également d'avoir : $H_0 : p = p_0$ et $H_1 : p < p_0$

2.c. Statistique et niveau de signification.

Une *statistique* est une fonction des variables aléatoires représentant l'échantillon. Le choix de la statistique dépend de la nature des données, du type d'hypothèse que l'on désire contrôler, des affirmations que l'on peut admettre concernant la nature des populations étudiées .... La valeur numérique de la statistique obtenue pour l'échantillon considéré permet de distinguer entre H_0 vraie et H_0 fausse.

Connaissant la loi de probabilité suivie par la statistique S sous l'hypothèse H_0, il est possible d'établir une valeur seuil, S_{seuil} de la statistique pour une probabilité donnée appelée le *niveau de signification* α du test. La région critique $R_c = f(S_{seuil})$ correspond à l'ensemble des valeurs telles que : $\mathbb{P}(S \in R_c) = \alpha$. Selon la nature unilatérale ou bilatérale du test, la définition de la région critique varie.

Test	Unilatéral $H_0 : t = t_0$		Bilatéral $H_0 : t = t_0$
Hypothèse alternative	$H_1 : t > t_0$	$H_1 : t < t_0$	$H_1 : t \neq t_0$
Niveau de signification	$\mathbb{P}(S > S_{seuil}) = \alpha$	$\mathbb{P}(S < S_{seuil}) = \alpha$	$\mathbb{P}(\lvert S\rvert > S_{seuil}) = \alpha$

Il existe deux stratégies pour prendre une décision en ce qui concerne un test d'hypothèse : la première stratégie fixe à priori la valeur du seuil de signification α et la seconde établit la valeur de la probabilité critique α_{obs} à posteriori.

Règle de décision 1 :
Sous l'hypothèse « H_0 est vraie » et pour un seuil de signification α fixé
- si la valeur de la statistique S_{obs} calculée appartient à la région critique alors l'hypothèse H_0 est rejetée au risque d'erreur α et l'hypothèse H_1 est acceptée ;
- si la valeur de la statistique S_{obs} n'appartient pas à la région critique alors l'hypothèse H_0 ne peut être rejetée.

Remarque : Le choix du niveau de signification ou risque α est lié aux conséquences pratiques de la décision ; en général on choisira $\alpha = 0,05,\ 0,01$ ou $0,001$.

Règle de décision 2 :
La probabilité critique α telle que $P(S \geq S_{obs}) = \alpha_{obs}$ est évaluée
- si $\alpha_{obs} \geq \alpha$ l'hypothèse H_0 est acceptée car le risque d'erreur de rejeter H_0 alors qu'elle est vrai est trop important ;
- si $\alpha_{obs} < \alpha$ l'hypothèse H_0 est rejetée car le risque d'erreur de rejeter H_0 alors qu'elle est vrai est très faible.

3. Risques d'erreur

Définition 1. *On appelle* risque d'erreur de première espèce *la probabilité de rejeter H_0 et d'accepter H_1 alors que H_0 est vraie.*

Ceci se produit si la valeur de la statistique de test tombe dans la région de rejet alors que l'hypothèse H_0 est vraie. La probabilité de cet évènement est le niveau de signifiation α. On dit aussi que le niveau de signification est la probabilité de rejeter l'hypothèse nulle à tort.

Remarque : La valeur du risque α doit être fixée a priori par l'expérimentateur et jamais en fonction des données. C'est un compromis entre le risque de conclure à tort et la faculté de conclure.

La région critique diminue lorsque α décroît (voir intervalle de confiance) et donc on rejette moins fréquemment H_0. A vouloir commettre moins d'erreurs, on conclut plus rarement.

Exemple : Si l'on cherche à tester l'hypothèse qu'une pièce de monnaie n'est pas « truquée », nous allons adopter la règle de décision suivante :

H_0 : la pièce n'est pas truquée
– est acceptée si $X \in [40, 60]$
– rejetée si $X \notin [40, 60]$ donc soit $X < 40$ ou $X > 60$
avec X « nombre de faces » obtenus en lançant 100 fois la pièce. Le risque d'erreur de première espèce est $\alpha = \mathbb{P}(B(100, 1/2) \in [40, 60])$.

Définition 2. *On appelle* risque d'erreur de seconde espèce*, notée β la probabilité de rejeter H_1 et d'accepter H_0 alors que H_1 est vraie.*

Ceci se produit si la valeur de la statistique de test ne tombe pas dans la région de rejet alors que l'hypothèse H_1 est vraie.

Remarque : : Pour quantifier le risque β, il faut connaître la loi de probabilité de la statistique sous l'hypothèse H_1.

Exemple : Si l'on reprend l'exemple précédent de la pièce de monnaie, et que l'on suppose la probabilité d'obtenir face est de $0,6$ pour une pièce truquée. En adoptant toujours la même règle de décision : H_0 : la pièce n'est pas truquée

– est acceptée si $X \in [40, 60]$
– rejetée si $X \notin [40, 60]$ donc soit $X < 40$ ou $X > 60$
avec X « nombre de faces » obtenues en lançant 100 fois la pièce. Le risque de seconde espèce est $\beta = \mathbb{P}(B(100, 0, 6) \in [40, 60])$.

4. Puissance d'un test

Rappelons que les tests ne sont pas faits pour « démontrer » H_0 mais pour « rejeter » H_0 . L'aptitude d'un test à rejeter H_0 alors qu'elle est fausse constitue la puissance du test.

Définition 3. *On appelle* puissance d'un test, *la probabilité de rejeter H_0 et d'accepter H_1 alors que H_1 est vraie. Sa valeur est $1 - \beta$*

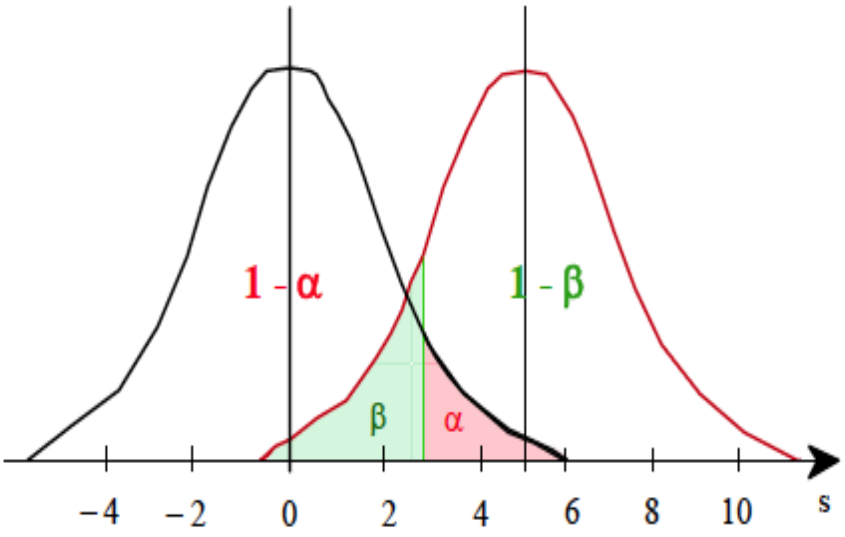
La puissance d'un test est fonction de la nature de H_1, un test unilatéral est plus puissant qu'un test bilatéral. Elle augmente avec taille de l'échantillon N étudié, et diminue lorsque α diminue.

La robustesse d'une technique statistique représente sa sensibilité à des écarts aux hypothèses faites.

Les différentes situations que l'on peut rencontrer dans le cadre des tests d'hypothèse sont résumées dans le tableau suivant :

Décision Réalité	H_0 vraie	H_1 vraie
H_0 acceptée	correct	manque de puissance risque de seconde espèce β
H_1 acceptée	rejet à tort risque de premières espèce α	puissance du test $1 - \beta$

On pe



CHAPITRE II

Test paramétriques

On suppose dans ce chapitre que les échantillons sont issus d'une loi normale ou peuvent être approximés par une loi normale.

1. Test de la moyenne

1.a. Variance connue. On suppose que l'on a un échantillon qui suit une loi normale $\mathcal{N}(\mu, \sigma^2)$ ou la variance est connue.

On veut tester $H_0 : \mu = \mu_0$ contre $H_1 : \mu \neq \mu_0$, c'est le cas bilatéral.

Sous l'hypothèse H_0 la variable aléatoire $\overline{X_n} = \frac{1}{n}\sum_{k=1}^{n} X_k$ suit une loi $\mathcal{N}(\mu_0, \sigma^2/n)$ et par conséquent la statistique

$$Z = \frac{\overline{X_n} - \mu_0}{\sqrt{\sigma^2/n}}$$

suit une loi normale centrée réduite.

Pour un risque d'erreur α fixé on a donc

$$\mathbb{P}(|Z| \leq q_{1-\alpha/2}) = 1 - \alpha$$

avec $q_{1-\alpha/2}$ le quantile d'ordre $1 - \alpha/2$ de la loi $\mathcal{N}(0, 1)$; et donc la région de rejet est

$$]-\infty, q_{1-\alpha/2}[\cup]q_{1-\alpha/2}, +\infty[$$

On calcule alors pour les valeurs de l'échantillon Z et on accepte ou on rejette H_0 suivant la valeur trouvée, au risque α

Si on considère un test unilatéral et une hypothèse alternative $H_1 : \mu > \mu_0$ par exemple, on obtient pour un risque d'erreur α

$$\mathbb{P}(Z \leq q_{1-\alpha}) = 1 - \alpha$$

avec $q_{1-\alpha}$ le quantile d'ordre $1 - \alpha$ de la loi $\mathcal{N}(0, 1)$; et donc la région de rejet est

$$]q_{1-\alpha}, +\infty[$$

1.b. Variance inconnue. On suppose que l'on a un échantillon qui suit une loi normale $\mathcal{N}(\mu, \sigma^2)$ ou la variance est maintenant inconnue.

On veut tester $H_0 : \mu = \mu_0$ contre $H_1 : \mu \neq \mu_0$, dans le cas bilatéral.

Sous l'hypothèse H_0 la variable aléatoire $\overline{X_n} = \frac{1}{n}\sum_{k=1}^{n} X_k$ suit une loi $\mathcal{N}(\mu_0, \sigma^2/n)$. Comme la variance est inconnue, on l'estime par la variance empirique :

$$\overline{S'_n}^2 = \frac{1}{n-1}\sum_{k=1}^{n}(X_k - \overline{X_n})^2.$$

On a déjà vu qu'alors la variable

$$T = \frac{X_n - \mu_0}{\sqrt{\overline{S'_n}^2/n}}$$

suit une de Student à $n - 1$ degrés de liberté.

Pour un risque d'erreur α fixé on a donc

$$\mathbb{P}(|T| \leq t_{1-\alpha/2}) = 1 - \alpha$$

avec $t_{1-\alpha/2}$ le quantile d'ordre $1-\alpha/2$ de la loi de Student à $n-1$ degrés de liberté ; et donc la région de rejet est

$$]-\infty, t_{1-\alpha/2}[\cup]t_{1-\alpha/2}, +\infty[$$

On calcule alors pour les valeurs de l'échantillon T et on accepte ou on rejette H_0 suivant la valeur trouvée, au risque α

2. Test de la variance

2.a. Moyenne connue.

On suppose que l'on a un échantillon qui suit une loi normale $\mathcal{N}(\mu, \sigma^2)$ où la moyenne est connue.

On veut tester $H_0 : \sigma^2 = \sigma_0^2$ contre $H_1 : \sigma^2 \neq \sigma_0^2$. Sous l'hypothèse H_0 la statistique

$$V = \frac{n\overline{S_n}^2}{\sigma_0^2} = \sum_{k=1}^{n} \left(\frac{X_k - \mu}{\sigma_0} \right)^2$$

suit une loi du χ^2 à n degrés de libertés.

Pour un risque d'erreur α fixé on a donc (en choisissant un intervalle symétrique) :

$$\mathbb{P}\left(\chi^2_{\alpha/2}(n) \leq \frac{n\overline{S_n}^2}{\sigma_0^2} \leq \chi^2_{1-\alpha/2}(n) \right) = 1 - \alpha$$

avec $\chi^2_{\alpha/2}(n)$ et $\chi^2_{1-\alpha/2}(n)$ les quantiles d'ordre $\alpha/2$ et $1-\alpha/2$ de la loi $\chi^2(n)$. Donc la région de rejet est

$$[0, \chi^2_{\alpha/2}(n)[\cup]\chi^2_{1-\alpha/2}(n), +\infty[$$

On calcule alors pour les valeurs de l'échantillon, V, et on accepte ou on rejette au risque α H_0 suivant la valeur trouvée.

Si on a une hypothèse alternative $H_1 : \sigma^2 > \sigma_0^2$ on fera un test unilatéral, et obtient au risque α

$$\mathbb{P}\left(\frac{n\overline{S_n}^2}{\sigma^2} \leq \chi^2_{1-\alpha}(n) \right) = 1 - \alpha$$

avec $\chi^2_{1-\alpha}(n)$ le quantile d'ordre $1-\alpha$ de la loi $\chi^2(n)$. Donc la région de rejet est $]\chi^2_{1-\alpha}(n), +\infty[$

2.b. Moyenne inconnue.

On suppose que l'on a un échantillon qui suit une loi normale $\mathcal{N}(\mu, \sigma^2)$ où la moyenne est inconnue.

On veut tester $H_0 : \sigma^2 = \sigma_0^2$ contre $H_1 : \sigma^2 \neq \sigma_0^2$. Sous l'hypothèse H_0 la statistique

$$V = \frac{n\overline{S'_n}^2}{\sigma_0^2} = \sum_{k=1}^{n} \left(\frac{X_k - \overline{X_n}}{\sigma_0} \right)^2$$

suit une loi du χ^2 à $n-1$ degrés de libertés.

Pour un risque d'erreur α fixé on a donc (en choisissant un intervalle symétrique) :

$$\mathbb{P}\left(\chi^2_{\alpha/2}(n-1) \leq \frac{n\overline{S'_n}^2}{\sigma_0^2} \leq \chi^2_{1-\alpha/2}(n-1) \right) = 1 - \alpha$$

avec $\chi^2_{\alpha/2}(n-1)$ et $\chi^2_{1-\alpha/2}(n-1)$ les quantiles d'ordre $\alpha/2$ et $1-\alpha/2$ de la loi $\chi^2(n-1)$. Donc la région de rejet est

$$[0, \chi^2_{\alpha/2}(n-1)[\cup]\chi^2_{1-\alpha/2}(n-1), +\infty[$$

On calcule alors pour les valeurs de l'échantillon, V, et on accepte ou on rejette au risque α H_0 suivant la valeur trouvée.

3. Test de la fréquence

Le modèle mathématique est le suivant. On dispose d'une population dans laquelle chaque individu présente ou non un certain caractère, la proportion d'individus présentant le caracère étant notée p, et un échantillon aléatoire de taille n extrait de cette population. La proportion f calculée à partir de l'échantillon est considérée comme une réalisation d'une v.a. de loi binomiale $\mathcal{B}(n,p)$ qu'on peut assimiler, si n est assez grand, à une loi normale $\mathcal{N}(p, \sqrt{p(1-p)}/\sqrt{n})$.

On veut tester $H_0 : p = p_0$ contre $H_1 : p \neq p_0$, dans le cas bilatéral. On obtient la région de rejet pour un risque α

$$\left]-\infty, p_0 - q_{1-\alpha/2}\frac{\sqrt{p_0(1-p_0)}}{\sqrt{n}}\right[\cup \left]p_0 + q_{1-\alpha/2}\frac{\sqrt{p_0(1-p_0)}}{\sqrt{n}}, +\infty\right[$$

avec $q_{1-\alpha/2}$ le quantile d'ordre $1-\alpha/2$ de la loi $\mathcal{N}(0,1)$.

4. Test de comparaison de deux moyennes

Si les deux échantillons on la même taille $n_1 = n_2 = n$. Le test se ramène à une test à une moyenne nulle de l'échantillon $(Z_1, ..., Z_n)$, avec $Z_i = X_i - Y_i$.

4.a. Variance connue. On suppose que l'on a deux échantillons $(X_1, \ldots, X_{n_1})$ et $(Y_1, \ldots, Y_{n_2})$qui suivent une loi normale $\mathcal{N}(\mu_1, \sigma_1^2)$ et $\mathcal{N}(\mu_2, \sigma_2^2)$ où les variances sont connues.

On veut tester $H_0 : \mu_1 = \mu_2$ contre $H_1 : \mu_1 \neq \mu_2$, c'est la cas bilatéral.

Sous l'hypothèse H_0 la variable aléatoire $\overline{X_{n_1}} = \frac{1}{n_1}\sum_{k=1}^{n_1} X_k$ suit une loi $\mathcal{N}(\mu_1, \sigma_1^2/n_1)$ et la variable aléatoire $\overline{Y_{n_2}} = \frac{1}{n_2}\sum_{k=1}^{n_2} Y_k$ suit une loi $\mathcal{N}(\mu_2, \sigma_2^2/n_2)$, par conséquent la statistique

$$U = \frac{\overline{X_{n_1}} - \overline{Y_{n_2}}}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}$$

suit une loi normale centrée réduite.

Pour un risque d'erreur α fixé on a donc

$$\mathbb{P}(|U| \leq q_{1-\alpha/2}) = 1 - \alpha$$

avec $q_{1-\alpha/2}$ le quantile d'ordre $1-\alpha/2$ de la loi $\mathcal{N}(0,1)$; et donc la région de rejet est

$$]-\infty, q_{1-\alpha/2}[\cup]q_{1-\alpha/2}, +\infty[$$

On calcule alors pour les valeurs de l'échantillon U et on accepte ou on rejette H_0 suivant la valeur trouvée, au risque α

Si on considère un test unilatéral et une hypothèse alternative $H_1 : \mu_1 > \mu_2$ par exemple, on obtient pour un risque d'erreur α

$$\mathbb{P}(Z \leq q_{1-\alpha}) = 1 - \alpha$$

avec $q_{1-\alpha}$ le quantile d'ordre $1-\alpha$ de la loi $\mathcal{N}(0,1)$; et donc la région de rejet est

$$]q_{1-\alpha}, +\infty[$$

4.b. Variance inconnue. On suppose que l'on a que l'on a deux échantillons $(X_1, \ldots, X_{n_1})$ et $(Y_1, \ldots, Y_{n_2})$qui suivent une loi normale $\mathcal{N}(\mu_1, \sigma_1^2)$ et $\mathcal{N}(\mu_2, \sigma_2^2)$ où les variances sont inconnues.

Cas 1 : n_1 et n_2 supérieurs à 30.

On veut tester $H_0 : \mu_1 = \mu_2$ contre $H_1 : \mu_1 \neq \mu_2$, dans le cas bilatéral.

Sous l'hypothèse H_0 la variable aléatoire $\overline{X_{n_1}} - \overline{Y_{n_2}}$ suit une loi $\mathcal{N}(0, \sigma_1^2/n_1 + \sigma_2^2/n_2)$. Comme la variance est inconnue, on l'estime par la variance empirique corrigée

$$\overline{S'_{n_1}}^2 + \overline{S'_{n_2}}^2 = \frac{1}{n_1 - 1}\sum_{k=1}^{n_1}(X_k - \overline{X_{n_1}})^2 + \frac{1}{n_2 - 1}\sum_{k=1}^{n_2}(Y_k - \overline{Y_{n_2}})^2.$$

Alors la variable aléatoire

$$N = \frac{\overline{X_{n_1}} - \overline{Y_{n_2}}}{\sqrt{\overline{S'_{n_1}}^2/n_1 + \overline{S'_{n_2}}^2/n_2}}$$

peut être approximé par une loi normale centrée réduite.

Pour un risque d'erreur α fixé on a donc

$$\mathbb{P}(|N| \leq t_{1-\alpha/2}) = 1 - \alpha$$

avec $t_{1-\alpha/2}$ le quantile d'ordre $1 - \alpha/2$ de la loi normale centrée réduite ; et donc la région de rejet est

$$]-\infty, t_{1-\alpha/2}[\cup]t_{1-\alpha/2}, +\infty[$$

On calcule alors pour les valeurs de l'échantillon N et on accepte ou on rejette H_0 suivant la valeur trouvée, au risque α.

Cas 2 : n_1 ou n_2 inférieur à 30 et $\sigma_1 = \sigma_2$

On veut tester $H_0 : \mu_1 = \mu_2$ contre $H_1 : \mu_1 \neq \mu_2$, dans le cas bilatéral.

Sous l'hypothèse H_0 la variable aléatoire $\overline{X_{n_1}} - \overline{Y_{n_2}}$ suit une loi $\mathcal{N}(0, \sigma_1^2/n_1 + \sigma_2^2/n_2)$. Comme la variance est inconnue, on l'estime par la variance empirique corrigée

$$\overline{S'_{n_1,n_2}}^2 = \frac{1}{n_1 + n_2 - 2}\left(\sum_{k=1}^{n_1}(X_k - \overline{X_{n_1}})^2 + \sum_{k=1}^{n_2}(Y_k - \overline{Y_{n_2}})^2\right).$$

Alors la variable aléatoire

$$T = \frac{\overline{X_{n_1}} - \overline{Y_{n_2}}}{\sqrt{\overline{S'_{n_1,n_2}}^2(1/n_1 + 1/n_2)}}$$

suit une de Student à $n_1 + n_2 - 2$ degrés de liberté.

Pour un risque d'erreur α fixé on a donc

$$\mathbb{P}(|T| \leq t_{1-\alpha/2}) = 1 - \alpha$$

avec $t_{1-\alpha/2}$ le quantile d'ordre $1 - \alpha/2$ de la loi de Student à $n_1 + n_2 - 2$ degrés de liberté ; et donc la région de rejet est

$$]-\infty, t_{1-\alpha/2}[\cup]t_{1-\alpha/2}, +\infty[$$

On calcule alors pour les valeurs de l'échantillon T et on accepte ou on rejette H_0 suivant la valeur trouvée, au risque α

Cas 3 : n_1 ou n_2 inférieur à 30 et $\sigma_1 \neq \sigma_2$

On veut tester $H_0 : \mu_1 = \mu_2$ contre $H_1 : \mu_1 \neq \mu_2$, dans le cas bilatéral.

Sous l'hypothèse H_0 la variable aléatoire $\overline{X_{n_1}} - \overline{Y_{n_2}}$ suit une loi $\mathcal{N}(0, \sigma_1^2/n_1 + \sigma_2^2/n_2)$. Comme la variance est inconnue, on l'estime par la variance empirique corrigée

$$\overline{S'_{n_1}}^2 + \overline{S'_{n_2}}^2 = \frac{1}{n_1 - 1}\sum_{k=1}^{n_1}(X_k - \overline{X_{n_1}})^2 + \frac{1}{n_2 - 1}\sum_{k=1}^{n_2}(Y_k - \overline{Y_{n_2}})^2.$$

Alors la variable aléatoire

$$T = \frac{\overline{X_{n_1}} - \overline{Y_{n_2}}}{\sqrt{\overline{S'_{n_1}}^2/n_1 + \overline{S'_{n_2}}^2/n_2}}$$

suit une de Student à ν degrés de liberté, où ν est l'entier le plus proche de

$$\frac{(\overline{S'_{n_1}}^2/n_1 + \overline{S'_{n_2}}^2/n_2)^2}{(n_1 - 1)\overline{S'_{n_1}}^4/n_1^4 + (n_2 - 1)\overline{S'_{n_2}}^4/n_2^4}$$

Pour un risque d'erreur α fixé on a donc

$$\mathbb{P}(|T| \leq t_{1-\alpha/2}) = 1 - \alpha$$

avec $t_{1-\alpha/2}$ le quantile d'ordre $1-\alpha/2$ de la loi de Student précédente ; et donc la région de rejet est

$$]-\infty, t_{1-\alpha/2}[\cup]t_{1-\alpha/2}, +\infty[$$

On calcule alors pour les valeurs de l'échantillon T et on accepte ou on rejette H_0 suivant la valeur trouvée, au risque α

5. Test de comparaison de deux variances

Avec les mêmes notations que précédemment on teste

$$H_0 : \sigma_1^2 = \sigma_2^2 \quad \text{contre} \quad H_1 : \sigma_1^2 \neq \sigma_2^2$$

On considère

$$\overline{S'_{n_1}}^2 = \frac{1}{n_1 - 1} \sum_{k=1}^{n_1} (X_k - \overline{X_{n_1}})^2 \quad \text{et} \quad \overline{S'_{n_2}}^2 = \frac{1}{n_2 - 1} \sum_{k=1}^{n_2} (Y_k - \overline{Y_{n_2}})^2.$$

ainsi que la statistique

$$Z = \frac{\overline{S'_{n_1}}^2}{\overline{S'_{n_2}}^2}$$

Sous l'hypothèse H_0 la statistique Z suit une loi de Fisher-Snedecor $\mathcal{F}(n_1 - 1, n_2 - 1)$ à $n_1 - 1$ et $n_2 - 1$ degrés de liberté. Pour un risque d'erreur α fixé on a une région de rejet

$$[0, F_{\alpha/2}(n)[\cup]F_{1-\alpha/2}(n), +\infty[$$

où les quantiles sont déterminées à l'aide de la loi précédente.

6. Test de comparaison de deux proportions

On veut comparer deux proportions p_1 et p_2 à partir de deux échantillons. Le modèle mathématique est le suivant. On considère les proportions f_1 et f_2 associés aux deux échantillons. On veut tester $H_0 : p_1 = p_2$ contre $H_1 : p_1 \neq p_2$. On prend la statistique

$$Z = \frac{f_1 - f_2}{\sqrt{F(1-F)(1/n_1 + 1/n_2)}} \quad \text{avec} \quad F = \frac{n_1 f_1 + n_2 f_2}{n_1 + n_2}$$

On obtient la région de rejet pour un risque α

$$\left]-\infty, -q_{1-\alpha/2}[\cup]q_{1-\alpha/2}, +\infty\right[$$

avec $q_{1-\alpha/2}$ le quantile d'ordre $1-\alpha/2$ de la loi $\mathcal{N}(0,1)$.

CHAPITRE III

Test du χ^2

1. Construction du test

On considère des variables aléatoires $(X_k)_{1\leq k\leq n}$ indépendantes identiquement distribuées, à valeurs dans $\{1,\dots,m\}$ et on note $\mathbb{P}(X_k=i)=\pi_i$. On définit

$$N_i(n)=\sum_{k=1}^{n}1_{\{X_k=i\}}$$

c'est-à-dire le nombre fois que la valeur i apparaît dans l'échantillon $(X_1,X_2,\dots,X_n)$. On s'intéresse à la répartition de $N(n)=(N_1(n),N_2(n),\dots,N_m(n))$. On a pour tout $(i_1,i_2,\dots,i_m)$ tels que $0\leq i_1,i_2,\dots,i_m\leq n$ et $i_1+\cdots+i_m=n$

$$\mathbb{P}(N_1(n)=i_1,\cdots,N_m(n)=i_m)=C_n^{i_1}C_{n-i_1}^{i_2}C_{n-i_1-i_2}^{i_3}\cdots 1.\pi_1^{i_1}\dots\pi_m^{i_m}=\frac{n!}{i_1!\cdots i_m!}\pi_1^{i_1}\dots\pi_m^{i_m}$$

on obtient une répartition multinomiale.
Nous allons appliquer la convergence en loi au vecteur $N(n)$. Donnons d'abord une version vectorielle du théorème central limite.

Théorème 1.

Soit $(X_l)_{l\geq 1}$ une suite de vecteurs de aléatoires de $\mathbb{R}^k$, indépendants identiquement distribués tels que $\mathbb{E}(X_1)=\mu\in\mathbb{R}^k$ et $\Gamma=\mathbb{E}\left[(X_1-\mu)^t(X_1-\mu)^t\right]$ matrice de covariance $k\times k$. Alors on a

$$\frac{\sum_{j=1}^{n}X_j-n\mu}{\sqrt{n}}\xrightarrow[n\to+\infty]{\mathcal{L}}\mathcal{N}(0,\Gamma)$$

Remarquons que $X\in\mathbb{R}^k$ suit une loi $\mathcal{N}_k(0,\Gamma)$ si sa fonction caractéristique est

$$\mathbb{E}(e^{i<t,X>})=\exp\left(-1/2<t,\Gamma t>\right)$$

et sa densité, lorsque Γ^{-1} est alors

$$f(x)=\frac{1}{(2\pi)^{k/2}det(\Gamma)}\exp\left\{-\frac{1}{2}<x,\Gamma^{-1}x>\right\}$$

En appliquant ce résultat on obtient :

Théorème 2.

Avec les notations du début du paragraphe, et en posant $\pi^t=(\pi_1,\dots,\pi_m)$ on obtient :

$$\frac{N(n)-n\pi^t}{\sqrt{n}}\xrightarrow[n\to+\infty]{\mathcal{L}}\mathcal{N}(0,\Delta_\pi-\pi^t\pi)\quad o\grave{u}\quad\Delta_\pi=\begin{pmatrix}\pi_1&\dots&0\\&\ddots&\\0&\dots&\pi_m\end{pmatrix}$$

DÉMONSTRATION. On a

$$\mathbb{E}(N_i(n)) = \mathbb{E}(\sum_{k=1}^{n} 1_{\{X_k=i\}}) = n\mathbb{E}(1_{\{X_k=i\}}) = n\pi_i$$

et donc $\mathbb{E}(N(n)) = n\pi^t$. D'autre part, on peut remarquer que :

$$\mathrm{Cov}(1_{\{X_k=i\}}1_{\{X_k=j\}}) = \mathbb{E}[(1_{\{X_k=i\}} - \pi_i)(1_{\{X_k=j\}} - \pi_j)] = \mathbb{E}[\delta_{i,j}1_{\{X_k=i\}}] - \pi_i\pi_j.$$

Par conséquent en posant $Y_k = \begin{pmatrix} 1_{\{X_k=1\}} \\ \vdots \\ 1_{\{X_k=m\}} \end{pmatrix}$ on en déduit que :

$$\mathbb{E}(Y_k) = \pi \quad \text{et} \quad \mathrm{Cov}(Y_k, Y_k) = \Delta_\pi - \pi\pi^t$$

On conclut alors an appliquant le TCL à la suite $(Y_k)_{k\geq 1}$ et en remarquant que $N(n) = \sum_{k=1}^{n} Y_k$ □

On obtient alors le résultat qui suit, qui nous donne une convergence vers une loi du χ^2.

Théorème 3.

Avec les notations précédentes on a

$$D_n(\pi) = \sum_{i=1}^{m} \frac{(N_i(n) - n\pi_i)^2}{n\pi_i} \xrightarrow[n\to+\infty]{\mathcal{L}} \chi^2(m-1)$$

DÉMONSTRATION. On définit :

$$\begin{array}{rccc} f: & \mathbb{R}^m & \to & \mathbb{R} \\ & x & \mapsto & f(x) = \sum_{i=1}^{m} \frac{x_i^2}{\pi_i} \end{array}$$

Avec cette notation on a $D_n(\pi) = f\left(\dfrac{N(n) - n\pi}{\sqrt{n}}\right)$. Pour toute fonction F continue et bornée on a

$$\mathbb{E}\left[F \circ f\left(\frac{N(n) - n\pi}{\sqrt{n}}\right)\right] \xrightarrow[n\to+\infty]{} \mathbb{E}\left[F \circ f\left(\mathcal{N}(0, \Delta_\pi - \pi\pi^t)\right)\right]$$

par définition de la convergence en loi car $F \circ f$ est continue bornée et d'après le TCL appliqué à $\frac{N(n)-n\pi}{\sqrt{n}}$. Il faut donc déterminer l'image de la loi $\mathcal{N}(0, \Delta_\pi - \pi\pi^t)$ par f.

Soit $Z = \begin{pmatrix} Z_1 \\ \vdots \\ Z_m \end{pmatrix}$ un vecteur aléatoire de loi $\mathcal{N}(0, \Delta_\pi - \pi\pi^t)$. On a

$$f(Z) = \sum_{i=1}^{m} \frac{Z_i^2}{\pi_i} = ||U||^2 \quad \text{où} \quad U = \begin{pmatrix} U_1 \\ \vdots \\ U_m \end{pmatrix},\ U_i = \frac{Z_i}{\sqrt{\pi_i}}$$

Le vecteur aléatoire U suit la loi $\mathcal{N}(0, Id - \sqrt{\pi}\sqrt{\pi^t})$. Soit A une matrice orthogonal, alors le vecteur $V = AU$ est un vecteur de loi normale $\mathcal{N}(0, Id - A\sqrt{\pi}\sqrt{\pi^t}A^t)$ tel que $||V|| = ||U||$. Comme $||\sqrt{\pi}|| = 1$,

on peut prendre A telle que $A\sqrt{\pi} = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix}$ et ainsi la matrice de covariance sera

$$Id - \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix} (0, \dots, 0, 1) = Id - \begin{pmatrix} 0 & \dots & 0 \\ & \ddots & \\ 0 & \dots & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & \ddots & \dots & \vdots \\ \vdots & & 1 & 0 \\ 0 & \dots & \dots & 0 \end{pmatrix}$$

Par conséquent, on a $f(Z) = ||U||^2 = ||V||^2 = \sum_{i=1}^{m-1} V_i^2$, où les $V_i \sim \mathcal{N}(0,1)$ sont indépendantes, et donc on obtient le résultat. □

2. Première application : test d'ajustement (loi discrète)

Soit $(X_1, \dots, X_n)$ un n-échantillon de la variable X à valeurs dans $\{1, \dots, m\}$ et de loi inconnue π, où $\mathbb{P}(X = i) = \pi_i$. On teste $H_0 \ : \ \pi = p$ contre $H_1 \ : \ \pi \neq p$ (p loi de probabilité discrète sur $\{1, \dots, m\}$). Supposons que H_1 soit vraie, alors :

$$\frac{N(n)}{n} \underset{n \to +\infty}{\overset{p.s.}{\longrightarrow}} p' \neq p$$

d'après la loi forte des grands nombres. Par conséquent on obtient

$$D_n(p) = \sum_{i=1}^{m} \frac{n}{p_i} \left(\frac{N_i(n)}{n} - p_i \right)^2 \underset{n \to +\infty}{\overset{p.s.}{\longrightarrow}} +\infty$$

car sous $p' \neq p$ et donc il existe i tel que

$$\left(\frac{N_i(n)}{n} - p_i \right)^2 \underset{n \to +\infty}{\longrightarrow} (p_i' - p_i)^2 > 0$$

Donc si H_0 est fausse on a $D_n(p) \to +\infty$, on peut donc utiliser une région de rejet de type $D_n(p) > C$ (unilatérale). Si on choisit un risque de première espèce inférieur à α, c'est-à-dire que l'on veut à la limite $\mathbb{P}(\chi^2(m-1) > \chi^2_{1-\alpha}(m-1)) \leq \alpha$ car d'après la convergence en loi :

$$\mathbb{P}(D_n(p) > \chi^2_{1-\alpha}(m-1)) \to \mathbb{P}(\chi^2(m-1) > \chi^2_{1-\alpha}(m-1)) \leq \alpha$$

Le test est donc asymptotiquement de niveau α, de plus sa puissance est 1 car sous H_1, $D_n(p) \to +\infty$ et donc

$$\mathbb{P}(D_n(p) > \chi^2_{1-\alpha}(m-1)) \to 1$$

Exemple.
On lance 200 fois une pièce et on obtient 110 piles. Le nombre de piles obtenu est une variable aléatoire $\mathcal{B}(200, p)$. On veut tester : $H_0 \ : \ p = 1/2$ contre $H_1 \ : \ p \neq 1/2$.

	`Effectifs observés` $N_i(n)$	`Effectifs théoriques` np_i	Ecarts $N_i(n) - np_i$	$(N_i(n) - np_i)^2/np_i$
`pile`	110	100	10	1
`face`	90	100	-10	1
`total`	200	200	0	2

On a $D_n(p) = 2$ et pour $\alpha = 0.05$ on trouve $\chi^2_{1-0.05}(2-1) = \chi^2_{1-0.05}(1) = 3,84$. On trouve donc $D_n(p) < \chi^2_{1-\alpha}(m-1)$ et on accepte l'hypothèse H_0.

3. Deuxième application : test d'ajustement (loi continue)

Soit $(X_1, \ldots, X_n)$ un n-échantillon de la variable X qui suit une loi absolument continue F inconnue. On remplace la loi des X_i par une loi discrète. Pour ce faire on procède de la manière suivante : on considère des intervalles I_j tels que

$$\mathbb{R} = \cup_{j=1}^{m} I_j \quad \text{avec} \quad I_j \cap I_l = \emptyset \text{ si } j \neq l$$

Si on veut alors tester $H_0 : F = F_0$ contre $H_0 : F \neq F_0$ on se ramène au cas précédent en posant

$$\pi_j = \mathbb{P}_F(X \in I_j) \quad p_j = \mathbb{P}_{F_0}(X \in I_j).$$

4. Troisième application : test d'ajustement (famille de lois)

Soit $(X_1, \ldots, X_n)$ un n-échantillon de la variable X à valeurs dans $\{1, \ldots, m\}$, qui suit une loi $(P_\theta)_{\theta \in \Theta}$, avec θ inconnue. On a

$$D_n(p(\theta)) = \sum_{i=1}^{m} \frac{(N_i(n) - np_i(\theta))^2}{np_i(\theta)}$$

On remplace θ par un estimateur, par exemple l'estimateur du maximum de vraisemblance $\hat{\theta}$, puis on procède comme dans la première application en remarquant que si $\Theta \in \mathbb{R}^k$

$$D_n(p(\hat{\theta})) \to \chi^2(m - k - 1)$$

Exemple.
On a 200 séries de 400 pièces chacune. Pour $k = 1, \ldots, 200$, on note X_k le nombre de pièces défectueuses dans la k-ième série. On veut tester l'hypothèse $H_0 : \pi_i = p_i = \frac{e^{-\lambda}\lambda^i}{i!}$, $i \geq 0$. Le tableau ci-dessous nous permet de déterminer

$$\hat{\lambda}_{200} = \frac{1}{200} \sum_{i=1}^{200} iN_i(200) = 4.$$

i	≤ 1	2	3	4	5	6	7	≥ 8
$N_i(200)$	17	31	37	41	30	23	18	8
$np_i(\hat{\lambda}_{200})$	18.4	29.3	39.1	39.1	31.3	20.8	11.9	10.2

On obtient

$$D_{200}(\hat{\lambda}_{200}) = \frac{(17 - 18.4)^2}{18.4} + \frac{(31 - 29.3)^2}{29.3} + \cdots + \frac{(8 - 10.2)^2}{10.2} = 1.283$$

On a $m = 8$ et $k = 1$ d'où il faut regarder $\chi^2(6)$. Pour $\alpha = 0.05$ le quantile est $\chi^2_{0.95}(6) = 12.69$. On accepte donc H_0.

5. Quatrième application : test d'indépendance

On observe un échantillon $((X_1, Y_1), \ldots, (X_n, Y_n))$ à valeur dans et $\{(x(i), y(j)), 1 \leq i \leq r, 1 \leq j \leq s\}$. On veut tester H_0 : X et Y sont indépendants . Soit $N_{i,j}$ le nombre de couples observés $(x(i), y(j))$ parmi n observations. On pose $N_{i.} = \sum_{j=1}^{s} N_{ij}$ et $N_{.j} = \sum_{i=1}^{r} N_{ij}$ On peut alors montrer que sous ces hypothèses

$$\xi_n = n \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{(N_{ij} - N_{i.}N_{.j}/n)^2}{N_{i.}N_{.j}} \xrightarrow[n \to +\infty]{\mathcal{L}} \chi^2(r-1)(s-1)$$

On procède ensuite comme dans les paragraphes précédents.

CHAPITRE IV

Fonction de répartition empirique

Dans ce chapitre, on s'intéresse à l'estimation de la loi d'une variable aléatoire ainsi qu'aux problèmes de tests associés. Pour traiter ces questions, nous allons chercher à estimer la fonction de répartition de cette variable. Nous sommes donc confrontés à un problème de statistique non-paramétrique. Pour traiter ce problème, on utilisera la notion de fonction de répartition empirique.

1. Introduction

On considère un $n-$échantillon $(X_1, \ldots, X_n)$ d'une variable aléatoire X. On note F la fonction de répartition de X, c'est-à-dire :

$$\forall t \in \mathbb{R}, \quad F(t) = \mathbb{P}(X \leq t) = \mathbb{P}(X_i \leq t)$$

C'est cette fonction F que nous allons chercher à estimer en introduisant la *fonction de répartition empirique*.

Définition 1. *La* fonction de répartition empirique *associée à cet échantillon est la fonction :*

$$\begin{array}{ccc} \mathbb{R} & \longrightarrow & [0,1] \\ t & \mapsto & F_n(t) = \frac{1}{n}\sum_{i=1}^n 1_{\{X_i \leq t\}} \end{array}$$

Remarque
Pour tout $t \in \mathbb{R}$, la variable aléatoire $nF_n(t)$ suit la loi binomiale $\mathcal{B}(n, F(t))$.

Pour représenter la fonction F_n , on introduit la statistique d'ordre $(X_{(1)}, \ldots, X_{(n)})$ associée à l'échantillon $(X_1, \ldots, X_n)$ définie par

$$\{X_{(1)}, \ldots, X_{(n)}\} = \{X_1, \ldots, X_n\} \quad \text{et } X_{(1)} \leq \cdots \leq X_{(n)}.$$

On a alors :

$$\forall t \in \mathbb{R}, \quad F_n(t) = \frac{1}{n}\sum_{i=1}^n 1_{\{X_{(i)} \leq t\}}$$

On en déduit le résultat suivant.

Proposition 2. *La fonction F_n est une fonction en escalier, croissante, continue à droite et admettant une limite à gauche. Elle est discontinue aux points $(X_{(i)})_{1 \leq i \leq n}$ et constante sur $[X_{(i)}, X_{(i+1)}[$ pour $i \in \{1, ..., n-1\}$.*

La fonction $F_n(t)$ est un estimateur naturel de $F(t)$, comme nous allons le voir dans le résultat suivant.

Proposition 3. *On a pour tout $t \in \mathbb{R}$, $F_n(t)$ est un estimateur sans biais et fortement consistant de $F(t)$. Par ailleurs, on a*

$$\forall t \in \mathbb{R}, \quad \sqrt{n}(F_n(t) - F(t)) \xrightarrow[n \to +\infty]{\mathcal{L}} \mathcal{N}(0, F(t)(1 - F(t)).$$

Démonstration. Le calcul de l'espérance montre que l'estimateur est sans biais. De plus, la loi forte des grands nombres appliquée aux variables aléatoires i.i.d. $1_{\{X_i \leq t\}}$ (bornées donc intégrable) tels que $\mathbb{E}[1_{\{X_i \leq t\}}] = \mathbb{P}(X_i \leq t) = F(t)$ entraîne la convergence presque sûre. Le dernier point découle du théorème central limite. □

Dans les sections suivantes, nous allons nous intéresser non pas à la convergence ponctuelle simple de F_n vers F mais à la convergence uniforme. Notons que la discontinuité de F_n présente des inconvénients théoriques évidents dans l'optique d'estimer F . Néanmoins, comme elle est constante par morceaux, elle est simple à construire en pratique. Dans les sections suivantes, nous aurons besoin de l'outil de la *fonction inverse généralisée* :

$$\forall x \in]0,1[, \quad F^- = \inf\{t \in \mathbb{R} \ : \ F(t) \leq x\}.$$

Proposition 4. *(Ouvrard p. 29) La fonction inverse généralisée se confond avec l'inverse de F quand F est bijective. Elle possède les propriétés suivantes :*

- *La monotonie de F entraîne*

$$\forall t \in \mathbb{R}, \forall x \in]0,1[, \quad F(t) \geq x \iff t \geq F^-(x)$$

- *Si $U \sim \mathcal{U}([0,1])$ alors $F^-(U)$ est une variable aléatoire dont la fonction de répartition est F.*
- *Si Z est une variable aléatoire de fonction de répartition F continue alors $F(Z) \sim \mathcal{U}([0,1])$*

2. Théorème de Glivenko-Cantelli

Le résultat de cette section renforce le théorème de la loi forte des grands nombres.

Théorème 5.

Théorème de Glivenko-Cantelli
Soit $(X_i)_{i\geq 1}$ une suite de variables aléatoires i.i.d. de fonction de répartition F. On pose :

$$\forall t \in \mathbb{R}, \quad F_n(t) = \frac{1}{n}\sum_{i=1}^{n} 1_{\{X_i \leq t\}}.$$

Alors on a :

$$\sup_{t\in\mathbb{R}} |F_n(t) - F(t)| \to 0 \ p.s.$$

DÉMONSTRATION. D'après la loi forte des grands nombres (comme dans la proposition 3), on a pour tout $t \in \mathbb{R}$ fixé, $F_n(t)(\omega) \to F(t)(\omega)$ pour presque tout ω. Il reste à voir que la convergence est uniforme en t. Si $n \geq 1$ on pose :

$$V_n = \sup_{t\in\mathbb{R}} \left| \frac{1}{n}\sum_{i=1}^{n} 1_{\{X_i \leq t\}} - F(t) \right|.$$

Soit $(U_k)_k$ une suite de variable aléatoires indépendantes identiquement distribuées de loi $\mathcal{U}([0,1])$. On a alors les égalités en loi suivantes :

$$V_n = \sup_{t\in\mathbb{R}} \left| \frac{1}{n}\sum_{i=1}^{n} 1_{\{X_i \leq t\}} - F(t) \right| = \sup_{t\in\mathbb{R}} \left| \frac{1}{n}\sum_{i=1}^{n} 1_{\{F^-(U_i) \leq t\}} - F(t) \right| = \sup_{t\in\mathbb{R}} \left| \frac{1}{n}\sum_{i=1}^{n} 1_{\{U_i \leq F(t)\}} - F(t) \right|$$

Si on pose $s = F(t)$, il vient :

$$\sup_{t\in\mathbb{R}} \left| \frac{1}{n}\sum_{i=1}^{n} 1_{\{U_i \leq F(t)\}} - F(t) \right| = \sup_{s\in F(\mathbb{R})} \left| \frac{1}{n}\sum_{i=1}^{n} 1_{\{U_i \leq s\}} - s \right| \leq \sup_{s\in[0,1]} \left| \frac{1}{n}\sum_{i=1}^{n} 1_{\{U_i \leq s\}} - s \right|$$

Ainsi il suffit de montrer que le théorème est vrai dans le cas particulier où les variables $X_i = (U_i)$ suivent des lois uniformes sur $[0,1]$. Grâce à la loi des grands nombres, on sait que pour tout $s \in \mathbb{R}$, il existe un ensemble négligeable $N_s \subset \Omega$ vérifiant :

$$\forall \omega \in \Omega \backslash N_s, \quad \sum_{i=1}^{n} 1_{\{U_i \leq s\}} \to s.$$

Comme une réunion dénombrable d'ensembles négligeables est encore négligeable, on en déduit l'existence d'une partie négligeable $N \subset \Omega$ telle que :

$$\forall \omega \in \Omega \backslash N, \quad \forall s \in [0,1] \cap \mathbb{Q}, \quad \sum_{i=1}^{n} 1_{\{U_i \leq s\}} \to s.$$

En fait la croissance de $s \mapsto \sum_{i=1}^{n} 1_{\{U_i \leq s\}}$ fait que l'on a :

$$\forall \omega \in \Omega \backslash N, \quad \forall s \in [0,1], \quad \sum_{i=1}^{n} 1_{\{U_i \leq s\}} \to s.$$

Pour chaque $\omega \in \Omega \backslash N$, $\sum_{i=1}^{n} 1_{\{U_i \leq s\}}$ converge donc simplement vers s sur $[0,1]$. Le théorème de Dini nous permet alors de conclure que la convergence est uniforme. □

3. Test de Kolmogorov

Le théorème de Glivenko-Cantelli est une généralisation de la loi forte des grands nombres au cas non-paramétrique. La généralisation du TCL est donnée par le théorème qui suit. La statistique introduite dans ce théorème nous permettra de construire un test d'ajustement à une loi (test de Kolmogorov). Dans le même esprit, nous construirons aussi un test d'homogénéité (test de Kolmogorov - Smirnov).

Proposition 6. *Soit $(X_{(1)}, \ldots, X_{(n)})$ un n-échantillon issu de X. On note F la fonction de répartition de X et F_n la fonction de répartition empirique. Si F est continue alors la loi de*

$$D(F_n, F) = \sup_{t \in \mathbb{R}} |F_n(t) - F(t)|$$

ne dépend pas de F.

Démonstration. On a, en posant $x = F(t)$,

$$\begin{aligned} D(F_n, F) &= \sup_{t \in \mathbb{R}} |F_n(t) - F(t)| = \sup_{t \in \mathbb{R}} \left| \frac{1}{n} \sum_{i=1}^{n} 1_{\{X_i \leq t\}} - F(t) \right| \\ &= \sup_{t \in \mathbb{R}} \left| \frac{1}{n} \sum_{i=1}^{n} 1_{\{F(X_i) \leq F(t)\}} - F(t) \right| = \sup_{x \in [0,1]} \left| \frac{1}{n} \sum_{i=1}^{n} 1_{\{U_i \leq x\}} - x \right| \end{aligned}$$

d'où le résultat. □

Le résultat principal de cette section est le suivant.

Théorème 7.

On suppose que l'on a les mêmes hypothèses que ci-dessus. Alors la variable aléatoire $\sqrt{n} D(F_n, F)$ converge en loi, vers une loi limite qui ne dépend pas de F et dont la fonction de répartition est égale à :

$$\forall t \geq 0, \quad F_{KS}(t) = 1 + 2 \sum_{k=1}^{+\infty} (-1)^k \exp\left(-2k^2 t^2\right)$$

On a donc, pour $\lambda > 0$,

$$\mathbb{P}(\sqrt{n} D(F_n, F) \leq \lambda) \to 1 + 2 \sum_{k=1}^{+\infty} (-1)^k \exp\left(-2k^2 \lambda^2\right)$$

Ce théorème est admis.
Nous pouvons donc construire un test d'ajustement à une loi, dit *test de Kolmogorov*. On peut d'abord

remarquer que si on réordonne de manière croissante l'échantillon, $(X_{(1)}, \ldots, X_{(n)})$ alors $F_n(X_{(j)}) = j/n$ et

$$D(F_n, F) = \max_{1 \leq j \leq n} \max \left(\left| \frac{j}{n} - F(X_{(j)}) \right|, \left| F(X_{(j)}) - \frac{j-1}{n} \right| \right)$$

Si on veut tester que la loi de l'échantillon a pour fonction de répartition F_0, c'est-à-dire $H_0 \ : \ F = F_0$ contre $H_1 \ : \ F = F_1$, on commence par réordonner l'échantillon. Puis on calcule $D(F_n, F)$, en remarquant que sous H_0, on a $D(F_n, F) = D(F_n, F_0)$. Puis on cherche (dans une table) le quantile $k_{1-\alpha}$ de la loi de Kolmogorov. On accepte alors H_0 si $D(F_n, F_0) < D(F_n, F_0)$. Ce test est asymptotiquement de niveau α et sa puissance tend vers 1 quand n tend vers $+\infty$.

4. Test de Kolmogorov-Smirnov

Dans le même esprit, nous allons construire un test d'homogénéité. On observe deux échantillons de taille respective n et m, $(X_1, \ldots, X_n)$ et $(Y_1, \ldots, Y_m)$. On veut tester si les deux échantillons sont issus d'une même loi (éventuellement inconnue). On note F la fonction de répartition de chacune des variables X_i et G la fonction de répartition de chacune des variables Y_i . On veut tester $H_0 \ : \ F = G$ contre $H_1 \ : \ F \neq G$. Pour cela, on introduit :

$$\forall t \in \mathbb{R}, \quad F_n(t) = \frac{1}{n} \sum_{i=1}^{n} 1_{\{X_i \leq t\}} \quad \text{et} \quad G_m(t) = \frac{1}{m} \sum_{i=1}^{m} 1_{\{Y_i \leq t\}}$$

et on pose

$$D_{m,n} = \sup_{t \in \mathbb{R}} |F_n(t) - G_m(t)|.$$

On a le résultat suivant.

Théorème 8.

Avec les hypothèses données ci-dessus on a, sous "$H_0 \ : \ F = G$" :

$$\mathbb{P}(\sqrt{\frac{nm}{n+m}} D_{m,n} \leq \lambda) \to 1 + 2 \sum_{k=1}^{+\infty} (-1)^k \exp(-2k^2\lambda^2)$$

Ceci permet alors de construire un test sur le même modèle que ci-dessus.

Index