EISTI

Ecole Internationale des Sciences du Traitement de l'Information

Statistique inférentielle

DEUXIEME PARTIE : LES TESTS D'HYPOTHESES

« La raison d'être des statistiques, c'est de vous donner raison»
Abe Burrow, 1910-1985, humoriste américain

Table des matières

1. GENERALITES

Un test d'hypothèses permet de prendre une décision au vu des résultats d'une expérience ou d'observations d'un phénomène. Par exemple

- Contrôle qualité : Décider si une machine est à réparer ou remplacer en fonction du nombre de pièces défectueuses produites.
- Essais thérapeutiques : Décider si un nouveau traitement est efficace au vu des résultats de son expérimentation sur des malades.
- Sondage : Décider de l'impact d'une campagne à l'issu d'un sondage effectué sur la population.

Dans chaque cas, le problème consiste à trancher, au vu d'observations, entre une hypothèse appelée *hypothèse nulle*, notée H_0, et une autre hypothèse dite *hypothèse alternative*, notée H_1. On suppose bien entendu qu'une et une seule des deux hypothèses est vraie. Un test d'hypothèses est alors une méthodologie statistique qui permet de choisir entre ces deux hypothèses.

Un test d'hypothèses s'effectue à partir d'un échantillon de variables aléatoires i.i.d de loi dépendant d'un paramètre θ (par exemple une espérance pour la loi normale). Les hypothèses se ramènent alors à tester des valeurs pour le paramètre θ. On distingue deux types d'hypothèses :

- Les *hypothèses simples* qui permettent de tester une valeur fixée : $\theta=\theta_0$
- Les *hypothèses composites* qui permettent de tester un ensemble de valeurs : $\theta<\theta_0$ ou $\theta>\theta_0$ (*test unilatéral*) ou $\theta\neq\theta_0$ (*test bilatéral*)

1.1. Les risques

Pour chacun des deux choix possibles, correspond une erreur.

- L'erreur dite de *première espèce* qui consiste à choisir H_1 alors que H_0 est vraie.
- L'erreur dite de *deuxième espèce* qui consiste à choisir H_0 alors que H_1 est vraie.

Les conséquences de ces deux erreurs peuvent être d'importances diverses. Par exemple, en contrôle qualité, si on décide à tort que la machine est défectueuse, on engagera des frais inutiles de réparation, si en revanche, on décide à tort que la machine est en bon état de marche, on produira des pièces de mauvaise qualité avec les répercussions que cela entraine sur le client voire même la sécurité. Les erreurs sont dites de première et deuxième espèce car, en général, l'une des deux erreurs est plus grave que l'autre.

Prenons l'exemple classique des faiseurs de pluie (Saporta). Des relevés effectués pendant de nombreuses années ont permis d'établir que le niveau naturel de pluie dans la Beauce en millimètres par an suit une loi normale $N(600,100^2)$. Des entrepreneurs prétendaient pouvoir augmenter de 50mm le niveau moyen de pluie par insémination des nuages au moyen d'iodure d'argent. Le procédé fut mis à l'essai entre 1951 et 1959 et on releva les hauteurs de pluies suivantes :

Année	1951	1952	1953	1954	1955	1956	1957	1958	1959
mm	510	614	780	512	501	534	603	788	650

Les agriculteurs, à qui les faiseurs de pluie avaient proposé ce procédé forcément onéreux, pouvaient envisager les deux hypothèses suivantes :

$$\begin{cases} H_0 : \text{le procédé est sans effet} \\ H_1 : \text{le procédé augmente réellement le niveau moyen de pluie de 50mm} \end{cases}$$

L'erreur de première espèce consiste alors à acheter un procédé onéreux et inefficace. L'erreur de deuxième espèce revient à passer à côté d'un moyen d'augmenter sa production agricole.

A chaque décision correspond une erreur mais aussi une probabilité de décider juste ou d'avoir tort.

- Le *niveau de signification* du test, noté α, est la probabilité de l'erreur de première espèce, *i.e* la probabilité de rejeter H_0 à tort.
- La probabilité de l'erreur de deuxième espèce est noté β. La *puissance* du test, donnée par 1-β, est la probabilité de décider H_1 à raison.

Décision \ Vérité	H_0	H_1
H_0	$1-\alpha$	β
H_1	α	$1-\beta$

L'idéal serait de diminuer simultanément les deux risques mais cela est impossible. En pratique, on considère que l'une des deux erreurs est plus grave que l'autre et on évite que cette erreur se produise. On choisit alors H_0 et H_1 de façon à ce que l'erreur que l'on considère la plus grave soit l'erreur de première espèce et on fixe α de façon à maitriser cette erreur. Plus l'erreur est grave, plus le niveau sera petit. On ne contrôle pas le risque de deuxième espèce mais on peut le calculer de façon à vérifier si le test est acceptable.

Dans le cas des faiseurs de pluie, il est clair que le risque le plus grave pour les agriculteurs était d'acheter un procédé onéreux et inefficace. Ils choisirent de le fixer à α=5%.

1.2. La construction du test

Un test d'hypothèse à pour objectif d'établir une *règle de décision*. Celle-ci permet de choisir entre H_0 et H_1 au vu des observations $x_1,\dots,x_n$, sous la contrainte que la probabilité du risque de première espèce est égale à α fixé. L'idée est de conclure que H_0 est fausse s'il est très peu probable d'observer $x_1,\dots,x_n$ quand H_0 est vraie.

<u>La variable de décision</u>
La règle de décision s'établit, non pas à partir des observations $x_1,\dots,x_n$ directement, mais à partir d'une caractéristique estimée sur l'échantillon, en général une valeur moyenne (μ), une variance (σ^2), etc… . La variable aléatoire permettant d'estimer cette caractéristique est appelée la *variable de décision* du test (en général, un estimateur usuel).
Une fois la variable de décision fixée, il est nécessaire de déterminer sa loi de probabilité.

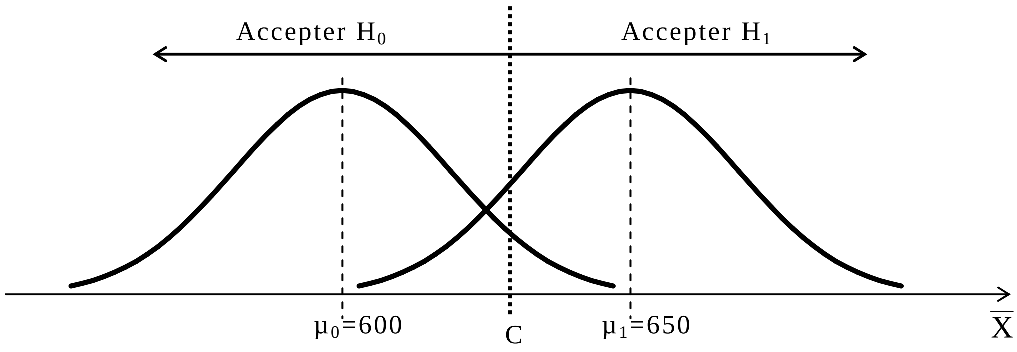
<u>La région critique</u>
On appelle *région critique*, notée W, l'ensemble des valeurs des observations $x_1,\dots,x_n$ pour lesquelles on rejettera H_0. Cette région dépend du niveau α et est déterminée indépendamment des observations effectuées. Ensuite, si les observations appartiennent à W, on rejette H_0, sinon on garde H_0.

Reprenons le cas des faiseurs de pluie. Si μ désigne l'espérance mathématique de la variable aléatoire X représentant le niveau annuel de pluie, alors les hypothèses peuvent se modéliser par

$$\begin{cases} H_0 : \mu=\mu_0=600 \text{ mm} \\ H_1 : \mu=\mu_1=650 \text{ mm} \end{cases}$$

Pour tester une espérance μ, on choisit naturellement la moyenne $\overline{X}$ comme variable de décision. L'échantillon est gaussien de variance connue avec n=9 donc $\overline{X}$ suit une loi normale $N(\mu,100^2/9)$.

Si l'hypothèse H_0 est vraie alors μ=600 et si l'hypothèse H_1 est vraie alors μ=650, d'où le graphique



On en déduit que la région critique (région d'acceptation de H_1) est définie par $W=\{\overline{X}>C\}$.

Le seuil de décision

La valeur du seuil de décision C est calculée en fonction de α, c'est-à-dire telle que

$$\alpha=P(W \mid H_0 \text{ est vraie}).$$

Si on suppose que H_0 est vraie, alors on connaît la loi de variable de décision. On peut donc résoudre l'équation ci-dessus.

Dans le cas des faiseurs de pluie, sous l'hypothèse H_0, on connaît la loi de $\overline{X}$ puisque $\mu=\mu_0=600$. D'où

$$0{,}05=P(\overline{X}>C \mid H_0 \text{ est vraie}) \Rightarrow C=622 \text{ mm}.$$

On adopte alors la règle de décision suivante :

- Si $\overline{x}$ (moyenne calculée sur l'échantillon) est supérieure à 622 mm alors on accepte l'hypothèse H_1, *i.e* les agriculteurs investissent dans le procédé avec 5% de risque de se faire avoir.
- Si $\overline{x}$ est inférieure à 622 mm, alors faute de preuve, on garde l'hypothèse H_0, *i.e* les agriculteurs n'investissent pas.

N.B. La construction du test se fait sans tenir compte des valeurs observées sur l'échantillon. Ces valeurs sont utilisées uniquement à la fin pour prendre la décision mais ne servent en aucun cas à établir la règle de décision.

Puissance du test

Pour connaître le risque de se tromper dans le deuxième cas, il suffit de calculer l'erreur de deuxième espèce ou bien la puissance du test,

$$1-\beta=P(W \mid H_1 \text{ est vraie}).$$

Cela suppose de connaître la loi de $\overline{X}$ sous l'hypothèse H_1, ce qui est le cas ici puisque $\mu_1=650$. On trouve alors 1-β=0.98. L'erreur de deuxième espèce est finalement de 2%.

$\alpha=P(W \mid H_0$ est vraie$)$	$\beta=P(\overline{W} \mid H_1$ est vraie$)$
$1-\alpha=P(\overline{W} \mid H_0$ est vraie$)$	$1-\beta=P(W \mid H_1$ est vraie$)$

Dans le cas où l'hypothèse H_1 est composite, la loi de la variable de décision sous l'hypothèse H_1 est inconnue puisque le paramètre θ_1 n'est pas fixé mais peut prendre une infinité de valeur. Il est alors impossible de calculer 1-β ou β. On définit alors la fonction puissance qui aux valeurs possibles de θ_1 associe la puissance correspondante,

$$\theta \mapsto 1-\beta(\theta) .$$

<u>La p-valeur</u>

La p-valeur est une probabilité donnant une information complémentaire au risque de première espèce α. Connaissant la valeur prise par la variable de décision sur l'échantillon, la p-valeur indique le niveau α minimum du test à partir duquel l'hypothèse H_0 est rejetée.

Si la p-valeur est inférieure à la valeur du seuil préalablement défini (traditionnellement 5 % ou 1 %), on rejette l'hypothèse nulle.

Prenons l'exemple d'une hypothèse H_0 sous laquelle la variable de décision Z suit une loi N(0,1). Supposons que la valeur prise par Z sur l'échantillon soit 2.72. Construisons un test bilatéral.

- Si on fixe un niveau de 5%, on obtient la règle de décision,

  $$H_0 \text{ est rejetée} \Leftrightarrow z \notin [-1.96,1.96].$$

  Avec z=2.72, l'hypothèse H_0 est donc rejetée avec 5% de risque de se tromper.
- Si on fixe un niveau de 1%, on obtient la règle de décision,

  $$H_0 \text{ est rejetée} \Leftrightarrow z \notin [-2.58,2.58].$$

  Avec z=2.72, 'hypothèse H_0 est donc rejetée avec 1% de risque de se tromper.
- Si on fixe un niveau de 0.5%, on obtient la règle de décision,

  $$H_0 \text{ est rejetée} \Leftrightarrow z \notin [-2.87,2.87].$$

  Avec z=2.72, l'hypothèse H_0 n'est pas rejetée.

Dans cet exemple, la p-valeur se situe entre 1% et 0.5%. Elle est en fait égale à 0.653%.

Méthodologie

1. Choix de H_0 et H_1 (*)
2. Détermination de la variable de décision et de sa loi
3. Allure de la région critique en fonction de H_1
4. Calcul de la région critique en fonction de α
5. Calcul éventuel de la puissance $1-\beta$
6. Règle de décision avec risques associés

(*) Il est conseillé d'expliciter les erreurs de première et deuxième espèces pour s'assurer du bon choix entre H_0 et H_1

2. TESTS SUR UNE MOYENNE

Soit un échantillon $X_1,\ldots,X_n$ tel que $E(X)=\mu$ et $var(X)=\sigma^2$. On souhaite tester les hypothèses

$$\begin{cases} H_0 : \mu=\mu_0 \\ H_1 : \mu=\mu_1\neq\mu_0 \end{cases}$$

La moyenne $\overline{X}$ est un estimateur naturel de μ. Pour un échantillon assez grand, on approche la loi de l'estimateur grâce au T.C.L. par une loi normale $N(\mu,\sigma^2/n)$.

La variable de décision (ou statistique du test) est

$$Z=\sqrt{n}\,\frac{\overline{X}-\mu}{\sigma}.$$

Si n est suffisamment grand, on considère donc qu'elle suit une loi normale N(0,1).

Remarques

- On sait que

$$\sqrt{n}\,\frac{\overline{X}-\mu}{S^*}$$

  converge vers une loi N(0,1) quand n tend vers +∞. Donc si σ^2 n'est pas connu, on peut le remplacer par s^{*2} si l'échantillon est assez grand.
- Dans le cas où l'échantillon est gaussien, la loi est exacte et non approchée (normale centrée-réduite si σ^2 est connue et Student à n-1 d.d.l. sinon). On peut donc travailler sur des échantillons de petites tailles.

Exemple

Un fabricant produit des lunettes d'un certain modèle dont la longueur en millimètre X suit une loi normale. L'espérance et l'écart-type de cette longueur sont spécifiés dans le cahier des charges à 130 mm et 0.48 mm respectivement. On prélève un échantillon de 9 faces de lunettes et on souhaite tester les hypothèses

$$\begin{cases} H_0 : \mu=\mu_0=130 \text{ mm} & \text{(lunettes conformes au cahier des charges)} \\ H_1 : \mu=\mu_1=130.5 \text{ mm} & \text{(lunettes non conformes au cahier des charges)} \end{cases}$$

avec un niveau de confiance de 99%.

Un simple graphique permet de déterminer l'allure de la région critique, $W=\{\overline{X}>C\}$. Sous l'hypothèses H_0, $\overline{X}$ suit une loi normale $N(130,0.48^2/n)$, d'où

$$\alpha=P(W|H_0) \Leftrightarrow 0.01=P(\overline{X}>C) \Leftrightarrow 0.01=P(\sqrt{n}\,\frac{\overline{X}-130}{0.48}>\sqrt{n}\,\frac{C-130}{0.48})$$
$$\Leftrightarrow 0.01=P(Z>C')$$

où Z suit une loi N(0,1). A l'aide de la table de la loi normale, on obtient C'=2.32 puis C=130.37 mm. Sous l'hypothèse H_1, $\overline{X}$ suit une loi $N(130.5,0.48^2/n)$, d'où

$$1-\beta=P(W|H_1)=P(\overline{X}>C)=P(\sqrt{n}\,\frac{\overline{X}-130.5}{0.48}>\sqrt{n}\,\frac{C-130.5}{0.48})$$
$$=P(Z>-0.81)=1-0.21=0.79$$

- Si $\overline{x}>130.37$ alors on accepte H_1, *i.e* on considère la marchandise comme non conforme, avec 1% de chance de se tromper
- Si $\overline{x}<130.37$ alors on garde H_0, *i.e* on considère que la marchandise est conforme, avec 21% de chance de se tromper.

Taille d'un échantillon

L'erreur de deuxième espèce est importante (21%). Déterminons combien il faudrait tester de faces de lunettes pour la réduire à 5%. Avec l'erreur de première espèce α, nous avons la première équation

$$C=\frac{2.32\times0.48}{\sqrt{n}}+130.$$

Avec la puissance fixée à 0.95, nous avons la deuxième équation

$$C=\frac{-1.64\times0.48}{\sqrt{n}}+130{,}5.$$

On en déduit que $\sqrt{n}=3.8$ et $n>14.44$. Prenons par exemple $n=15$, on a alors $C=130.29$ et la règle de décision devient

- Si $\bar{x}>130.29$ alors on accepte H_1, *i.e* on considère la marchandise comme non conforme, avec 1% de chance de se tromper
- Si $\bar{x}<130.29$ alors on garde H_0, *i.e* on considère que la marchandise est conforme, avec 5% de chance de se tromper.

<u>Remarque</u>

On remarque qu'un test d'hypothèses est un problème à

4 paramètres : α, β, C et n

2 équations : $\alpha=P(W|H_0)$ et $1-\beta=P(W|H_1)$

Il faut donc fixer deux paramètres pour résoudre le système. Il y a deux cas de figure

- Les observations ont déjà été effectuées, on a alors α et n fixés (maitrise du risque de première espèce).
- Les observations n'ont pas été effectuées et on fixe α et β (maitrise des deux risques).

3. Tests sur une proportion

On souhaite tester les hypothèses

$$\begin{cases} H_0 : p=p_0 \\ H_1 : p=p_1\neq p_0 \end{cases}$$

où p désigne une proportion. La fréquence empirique F_n est un estimateur naturel de p. Si l'échantillon est suffisamment grand, on peut approcher la loi de F_n grâce au TCL par une loi normale $N(p,p(1-p)/n)$.

La variable de décision (ou statistique) du test est

$$Z=\sqrt{n}\frac{F_n-p}{\sqrt{p(1-p)}}.$$

Si n est assez grand, on considère qu'elle suit une loi $N(0,1)$.

<u>Exemple</u>

Un parti politique souhaite vérifier si la parité est respectée. Sur un échantillon de 100 adhérents, on compte 45 femmes et 55 hommes. Peut-on affirmer que la parité est respectée avec un risque de 5% de se tromper?

Supposons que p représente le taux de femmes dans le parti. On souhaite alors tester les hypothèses

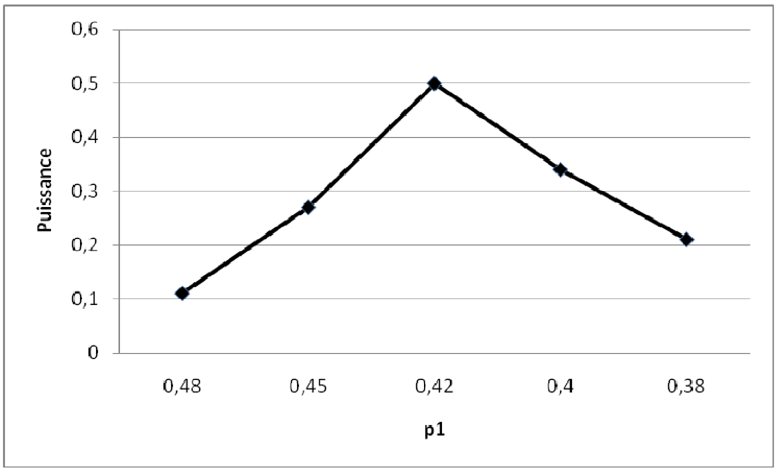
$$\begin{cases} H_0 : p=p_0=0.5 \text{ (Parité)} \\ H_1 : p=p_1<p_0 \text{ (Femmes sous représentées)} \end{cases}$$

Un graphique permet de déterminer que la région critique est de la forme $W=\{F_n<C\}$. Sous l'hypothèse H_0, F_n suit une loi normale $N(0.5,0.5^2/n)$, d'où

$$\alpha=P(W|H_0) \Leftrightarrow 0.05=P(F_n<C) \Leftrightarrow 0.05=P(\sqrt{n}\frac{F_n-0.5}{0.5}<\sqrt{n}\frac{C-0.5}{0.5})$$

$$\Leftrightarrow 0.05=P(Z<C')$$

où Z suit une loi N(0,1). A l'aide de la table de la loi normale, on obtient C'=-1.64 puis C=0.42.
Sut l'échantillon, on a un taux de f_n=0.45>0.42, donc on garde l'hypothèse H_0, *i.e* on considère que la parité est respectée. On ne connaît pas l'erreur de se tromper mais au vu de la fonction puissance ci-dessous, on peut dire que le test n'est pas très performant.



4. TESTS DE LA VARIANCE SUR DES ECHANTILLONS GAUSSIENS

Supposons que l'échantillon soit gaussien avec X de loi $N(\mu,\sigma^2)$.
On souhaite tester les hypothèses

$$\begin{cases} H_0 : \sigma^2=\sigma^2_0 \\ H_1 : \sigma^2=\sigma^2_1 \neq \sigma^2_0 \end{cases}$$

Dans le cas où μ est inconnu, la variance empirique S^{*2} est un estimateur naturel de σ^2 . La variable de décision (statistique) du test est

$$\frac{n-1}{\sigma^2} S^{*2} \sim \chi^2_{n-1} .$$

Le graphe de la loi du χ^2_{n-1} ne permet pas de déterminer l'allure de la région critique. Il est nécessaire d'appliquer le théorème de Neyman-Pearson* qui permet d'affirmer dans ce cas que la région critique est de la forme :

$$W=\{S^{*2}>C\} \Leftrightarrow \sigma^2_1 > \sigma^2_0$$

$$W=\{S^{*2}<C\} \Leftrightarrow \sigma^2_1 < \sigma^2_0$$

<u>Exemple</u>
Dans la fabrication de comprimés effervescents, il est prévu que le dosage de bicarbonate de sodium suit une loi normale. On a prélevé un échantillon de 25 comprimés et mesuré un écart-type empirique $s^* = 12.1$ mg. On souhaite tester si la production est homogène avec les hypothèses suivantes

$$\begin{cases} H_0 : \sigma^2=\sigma^2_0=10^2 \end{cases}$$

* Le théorème de Neyman-Pearson est hors programme. Il sera abordé uniquement dans le parcours Math-Info

$$H_1 : \sigma^2=\sigma^2_1>\sigma^2_0$$

avec un niveau de confiance de 5%.

La région critique est de la forme $W=\{S^{*2}>C\}$. Sous l'hypothèse H_0, on a $Z=\frac{n-1}{10^2}S^{*2}\sim\chi^2_{n-1}$, d'où

$$\alpha=P(W|H_0)\Leftrightarrow 0.05=P(S^{*2}>C)\Leftrightarrow 0.05=P(\frac{n-1}{10^2}S^{*2}>\frac{n-1}{10^2}C)\Leftrightarrow 0.05=P(Z>C')$$

Avec les tables de la loi du chi-deux à 24 d.d.l., on obtient, C'=36.41 puis $C=151.71=12.32^2$. On a $s^*=12.1$ mg <12.32 mg donc on garde l'hypothèse H_0, *i.e* on considère la production homogène sans connaître le risque de se tromper (car loi sous H_1 inconnue).

<u>Remarque</u>

Le cas où μ est connu n'est pas traité ici car très peu fréquent.

<u>Remarque</u>

Les tests précédents ne peuvent s'appliquer que dans le cas d'un échantillon gaussien.

5. TESTS D'ADEQUATION ET D'INDEPENDANCE

Les différents principes de la statistique inférentielle vus précédemment repose sur le fait que l'échantillon (échantillon i.id) est constitué de variables aléatoires <u>indépendantes</u> et <u>de même loi</u>, souvent la loi normale. Nous allons maintenant étudier des tests permettant de confirmer ou infirmer ces hypothèses.

5.1. Tests d'adéquation

Un test d'adéquation ou test d'ajustement permet de déterminer si un échantillon suit une loi usuelle donnée. Le test doit s'appuyer sur une caractéristique de la loi (moments, fonction densité/masse, …). Les tests présentés ici utilisent la fonction de répartition. Plus précisément, ils sont basés sur « l'écart » entre F, la fonction de répartition de la loi usuelle, et F_n la fonction de répartition empirique de l'échantillon.

<u>Définition</u>

La fonction empirique d'un échantillon $X_1,\dots,X_n$ est définie par

$$t\mapsto F_n(t)=\frac{1}{n}\sum_{i=1}^{n} I_{\{x_i\le t\}}$$

c'est-à-dire la proportion des v.a. $X_1,\dots, X_n$ qui sont inférieures à t.

Notons $d(F,F_n)$ l'écart entre F et F_n, alors le test est le suivant

$$\begin{cases} H_0 : d(F,F_n)=0 & \text{(l'échantillon suit la loi usuelle)} \\ H_1 : d(F,F_n)>0 & \text{(l'échantillon ne suit pas la loi)} \end{cases}$$

La région critique est donc de la forme $W=\{d(F,F_n)>C\}$ et la règle de décision est

- Si $d(F,F_n)<C$ alors je considère que l'échantillon suit la loi usuelle

- Si $d(F,F_n)>C$ alors j'en déduis qu'il ne suit pas la loi usuelle avec un risque α de me tromper

Pour construire complètement ce test, il faut déterminer

- Comment on calcule $d(F,F_n)$ l'écart entre F et F_n.
- La loi de probabilité de $d(F,F_n)$

Les deux tests qui suivent sont basés sur ce même principe et différent uniquement sur la définition de $d(F,F_n)$. Le premier est plutôt utilisé pour des lois discrètes et le deuxième pour des lois continues.

5.1.1. Test de chi-deux

Supposons que la loi usuelle est discrète prenant les valeurs x_i avec la probabilité p_i, i=1,…,k. Le test du chi-deux consiste à comparer pour chaque valeur x_i,

- son effectif théorique : $n\times p_i$
- et son effectif empirique, N_i=nombre d'observations prenant la valeur x_i.

On pose alors

$$d(F,F_n)=D_n=\sum_{i=1}^{k}\frac{(N_i-np_i)^2}{np_i}.$$

On peut montrer que sous l'hypothèse H_0, D_n suit approximativement une loi χ^2_{k-1}.

- Cette approximation est valable si n est assez grand et p_i pas trop petit avec la règle empirique $np_i\geq 5$. Si cette règle n'est pas respectée à cause d'un p_i trop petit, il faut alors regrouper le x_i avec une valeur voisine.
- Si la loi de l'échantillon (sous l'hypothèse H_0) dépend de *r* paramètres et que ces paramètres ont été estimés au préalable sur l'échantillon, alors D_n suit une loi χ^2_{k-r-1}. Par exemple, si on souhaite tester si un échantillon suit une loi de Poisson et que le paramètre λ est estimé sur ce même échantillon, alors D_n suit une loi χ^2_{k-1-1}.

<u>Remarque</u>
Dans le cas où la loi usuelle est une loi discrète uniforme, tous les x_i sont équiprobables, donc $p_i=1/k$.

<u>Remarque</u>
Il est possible d'utiliser ce test dans le cas d'une loi continue dont les données sont regroupées en classes (a_i,a_{i+1}). Dans ce cas la probabilité p_i de la classe (a_i,a_{i+1}) est

$$p_i=P(X\in(a_i,a_{i+1}))=F(a_{i+1})-F(a_i)$$

<u>Exemple</u>
Afin de savoir si un dé est équilibré, on effectue 60 lancers. L'hypothèse à tester est celle d'une loi uniforme discrète sur les 6 faces chacune de probabilité 1/6. Etant donné que $n\times p_i=10\geq 5$, on peut considérer que, sous l'hypothèse H_0, D_n suit une loi χ^2_5. Avec un risque de première espèce de 5%, on a l'équation,

$$\alpha=P(W|H_0 \text{ vraie}) \Leftrightarrow 0{,}05=P(D_n>C) \Rightarrow (\text{table } \chi^2_5)\ C=11{,}07.$$

Le tableau ci-dessous contient les effectifs empiriques des 60 lancers. On en déduit d_n=2,8. d_n est inférieur au seuil de décision 11,07, donc on garde l'hypothèse H_0, c-a-d on considère que le dé est équilibré. On ne connaît pas le risque de se tromper.

x_i	1	2	3	4	5	6
N_i	11	8	9	12	7	13
np_i	10	10	10	10	10	10
$\frac{(N_i-np_i)^2}{np_i}$	0,1	0,4	0,1	0,4	0,9	0,9

5.1.2. Test de Kolmogorov-Smirnov

Supposons le cas d'une variable aléatoire continue. Afin de ne pas perdre d'information en regroupant en classe les données observées, on utilise la distance de Kolmogorov,

$$d(F,F_n)=K_n=\sup_{x\in\mathbb{R}}\left|F_n(x)-F(x)\right|$$

En pratique, on calcule cette distance sur l'échantillon ordonné, $X_{(1)}<X_{(2)}<\ldots<X_{(n)}$. On a alors,

$$F_n(x)=\begin{cases}0 & \text{si} \quad x<X_{(1)}\\ \dfrac{i}{n} & \text{si} \quad X_{(i)}\le x<X_{(i+1)}\\ 1 & \text{si} \quad x\ge X_{(n)}\end{cases}.$$

On peut en déduire alors que,

$$K_n=\max_{i=1,\ldots,n}\left\{\left|F(X_{(i)})-\frac{i}{n}\right|,\left|F(X_{(i)})-\frac{i-1}{n}\right|\right\}.$$

Sous l'hypothèse H_0, la variable aléatoire $\sqrt{n}K_n$ suit approximativement une loi dont la fonction de répartition est définie par,

$$\forall x>0,\ K(x)=1-2\sum_{j=1}^{+\infty}(-1)^{j+1}e^{-2j^2x^2}.$$

La série converge très rapidement et pour x≥1, la somme des trois premiers termes suffit à avoir une bonne approximation. A partir de cette fonction de répartition, on peut :

- soit calculer le seuil de décision pour un α fixé : $\alpha=P(K_n>C|\ H_0 \text{ vraie})$. Le tableau ci-dessous donne les valeurs du seuil C en fonction du risque α et de la taille de l'échantillon,

n	α=0,10	α=0,05	α=0,01
5	0,509	0,563	0,669
10	0,369	0,409	0,486
15	0,304	0,338	0,404
20	0,265	0,294	0,352
25	0,238	0,264	0,317
30	0,218	0,242	0,290
40	0,189	0,210	0,252
n>40	$1,22/\sqrt{n}$	$1,36/\sqrt{n}$	$1,63/\sqrt{n}$

- soit calculer la p-valeur. Si on t la valeur de la variable aléatoire $\sqrt{n}K_n$ prise sur l'échantillon, alors la p-valeur est donnée par

$$p(t) \approx 2\sum_{j=1}^{3}(-1)^{j+1}e^{-2j^2t^2} .$$

Exemple

On souhaite savoir si les 15 données du tableau ci-dessous suivent une loi N(0,1). Pour un test avec un niveau de 5%, on accepte que l'échantillon suit une loi N(0,1) si $k_n<0,338$ (cf. tableau ci-dessus). Calculons k_n. Notons F la fonction de répartition de la loi N(0,1).

i	x_i	$F(x_i)$	$\lvert F(x_i)-i/n \rvert$	$\lvert F(x_i)-(i-1)/n \rvert$
1	-7,64	0	0,0667	0
2	-5,23	0	0,1333	0,0667
3	-1,84	0,0329	0,1671	0,1004
4	-1,25	0,1056	0,1610	0,0944
5	-1,02	0,1539	0,1795	0,1128
6	-0,65	0,2578	0,1422	0,0755
7	-0,01	0,4960	0,0293	0,0960
8	0,06	0,5239	0,0094	0,0573
9	0,12	0,5478	0,0522	0,0144
10	0,24	0,5948	0,0718	0,0052
11	0,74	0,7704	0,0370	0,1037
12	2,18	0,9854	0,1854	0,2520
13	2,85	0,9978	0,1311	0,1978
14	6,42	1	0,0667	0,1333
15	28,54	1	0	0,0667

On obtient $k_n=0,2520<0,338$. Donc on admet que l'échantillon suit une loi N(0,1). On ne connaît pas le risque de se tromper.

On peut aussi calculer la p-valeur. Ici $t=\sqrt{15}\times k_n=0,976$. Donc

$$p(0,976) \approx 2\sum_{j=1}^{3}(-1)^{j+1}e^{-2j^2 0,976^2} \approx 0,3$$

Cela signifie qu'il faudrait envisager un risque de première espèce de $\alpha=30\%$ pour rejeter l'hypothèse H_0 à tord. Donc le test est très largement accepté.

Remarque

Dans le cas d'une loi normale, il existe un test plus puissant que celui de Kolmogorov, le test de Shapiro-Wilk. On peut aussi avoir une représentation graphique permettant de juger si l'échantillon suit une loi normale, la droite de Henry (ou qq-plot).

5.2. Tests d'indépendance

Considérons deux caractères qualitatifs (ou quantitatif regroupée en classe) X et Y avec respectivement r et s modalités. Le test du khi-deux permet de tester l'indépendance de ces caractères. Notons n_{ij} le nombre d'individus de

l'échantillon présentant simultanément la modalité i du caractère X et la modalité j du caractère Y.
Soit p_{ij} la probabilité théorique de cette modalité. Notons $p_{i.}$ et $p_{.j}$ les probabilités marginales,

$$p_{i.} = \sum_{j=1}^{s} p_{ij} \text{ et } p_{.j} = \sum_{i=1}^{r} p_{ij} .$$

S'il y a indépendance alors la loi conjointe est égale au produit des lois. Nous allons donc effectuer le test suivant

$$\begin{cases} H_0 : p_{ij}=p_{i.}\times p_{j.} & \text{(indépendance)} \\ H_1 : p_{ij}\neq p_{i.}\times p_{j.} & \text{(pas indépendance)} \end{cases}$$

Pour ce faire, on utilise la variable de décision,

$$D_n = \sum_{i=1}^{r}\sum_{j=1}^{s} \frac{\left(n_{ij} - n_{i.}n_{.j}/n\right)^2}{n_{i.}n_{.j}/n} .$$

La région critique de ce test est de la forme $W=\{D_n>C\}$.

Sous l'hypothèse H_0, on peut approcher la loi de D_n par une loi $\chi^2_{(r-1)(s-1)}$.

Exemple

Le proviseur d'un lycée désire comparer le taux de réussite au baccalauréat des trois sections générales. Les résultats obtenus dans son lycée sont dans le tableau suivant.

	L	ES	S	Total
Réussite	41	59	54	154
Echec	21	36	75	132
Total	62	95	129	286

La région critique est de la forme, $D_n>C$ et sous l'hypothèse H_0, D_n suit une loi χ^2_2 (car s=2 et r=3). Donc si on prend un rique $\alpha=5\%$. On lit dans la table χ^2_2, C=5,991. D'où la règle,

- si $d_n>5,991$ alors on rejette l'hypothèse d'indépendance avec 5% de risque de se tromper
- si $d_n<5,991$ alors on accepte l'hypothèse d'indépendance.

Il reste à calculer la valeur de la statistique D_n.
A partir des effectifs marginaux du tableau, on construit le tableau des effectifs théoriques sous l'hypothèse d'indépendance, $n_{ij}=p_{i.}\times p_{.j}\times n$,

	L	ES	S	Total
Réussite	33	51	70	154
Echec	29	44	59	132
Total	62	95	129	286

On calcule alors la différence entre les deux tables, on obtient

$$d_n = \frac{(41-33)^2}{33} + ... + \frac{(75-59)^2}{59} = 14,85 .$$

Le proviseur peut donc conclure avec un risque de 5% que les résultats au baccalauréat dans son établissement dépendent de la section.