

Tests d'hypothèses

I Notions générales sur les tests statistiques

Un test est un mécanisme qui permet de trancher entre 2 hypothèses au vu des résultats d'un échantillon.

Soient H_0 et H_1 les hypothèses, dont une et une seule est vraie. La décision aboutira à choisir H_0 ou H_1 .

Il y a 4 cas possibles schématisés dans le tableau suivant avec les probabilités correspondantes.

Décision \ Vérité	H_0	H_1
	H_0	$1 - \alpha$ OK
H_1	α erreur	$1 - \beta$ OK

α et β sont les probabilités d'erreur de première et deuxième espèce :

— α est la probabilité de choisir H_1 alors que H_0 est vraie.

— β est la probabilité de choisir H_0 alors que H_1 est vraie.

Dans la pratique des tests statistiques, il est de règle de se fixer α comme donné, sa valeur standard est de 5% ; ce qui fait jouer à H_0 un rôle prééminent. H_0 est aussi appelé hypothèse nulle et H_1 hypothèse alternative.

Construire un test consiste à définir une règle de décision qui va associer une décision à un échantillon observé (X_1, \dots, X_n) de la loi de X , P_θ qui dépend d'un paramètre θ qui varie dans un sous ensemble donné Θ de \mathbb{R} . On suppose que cet ensemble est partitionné en 2 sous-ensemble Θ_0 et Θ_1 , auxquels vont être associés les 2 hypothèses : $H_0 : \theta \in \Theta_0$ et $H_1 : \theta \in \Theta_1$. Les 2 décisions possibles étant D_0 : accepter H_0 et D_1 : accepter H_1 . A chaque décision correspond une région de \mathbb{R}^n , qui va être partitionné en 2 sous-ensemble W et \overline{W} , c.a.d ; si la réalisation de l'échantillon est un point (x_1, \dots, x_n) de W on décide D_1 , donc on rejette H_0 . Dans le cas contraire, c.a.d. si la réalisation de l'échantillon $(x_1, \dots, x_n) \in \overline{W}$, on décide D_0 , donc on accepte H_0 .

Définition. La région W de rejet de l'hypothèse nulle H_0 se nomme la région critique du test et la région \overline{W} région d'acceptation.

La construction d'un test va donc consister à déterminer cette région critique. La méthode pour l'obtenir dépendra des conséquences que l'on attribue à chacune des 2 erreurs qui sont associées aux 2 décisions possibles.

Définition. L'erreur de première espèce consiste à décider D_1 alors que H_0 est vraie, soit rejeter à tort l'hypothèse nulle H_0 . L'erreur de seconde espèce consiste à décider D_0 alors que H_1 est vraie, soit accepter à tort l'hypothèse nulle.

II Méthode Neyman et Pearson

II.1 Principe de la règle de Neyman et Pearson

On privilégie l'une des 2 hypothèses, par exemple celle que l'on considère comme la plus vraisemblable, et on la choisit comme hypothèse nulle H_0 . Cette hypothèse sera celle dont le rejet à tort est le plus préjudiciable.

L'autre hypothèse H_1 est l'hypothèse alternative.

Définition. On appelle risque de première espèce, la probabilité de rejeter à tort l'hypothèse nulle, soit :

$$\alpha = P_\theta(D_1|H_0) = P_\theta(H_1|H_0) = P_\theta(W|\theta \in \Theta_0)$$

On appelle risque de 2ème espèce la probabilité d'accepter à tort l'hypothèse nulle, soit :

$$\beta = P_\theta(D_0|H_1) = P_\theta(H_0|H_1) = P_\theta(\bar{W}|\theta \in \Theta_1)$$

Définition. On appelle puissance d'un test la probabilité de refuser H_0 avec raison c.a.d. lorsque H_1 est vraie, soit :

$$\eta = P_\theta(D_1|H_1) = P_\theta(H_1|H_1) = P_\theta(W|\theta \in \Theta_1) = 1 - \beta$$

II.2 Hypothèses simples

On dit qu'une hypothèse est simple si la loi de la v.a. X est totalement spécifiée quand cette hypothèse est réalisée. Dans le cas contraire elle est dite multiple. Examinons le cas où le paramètre θ ne peut prendre que 2 valeurs θ_0 et θ_1 ce qui conduit au choix entre les 2 hypothèses simples :

$$\begin{cases} H_0 : \theta = \theta_0 \\ H_1 : \theta = \theta_1 \end{cases}$$

La forme de la région critique est déterminée par le théorème suivant :

Théorème (de Neyman et Pearson). Pour un risque de première espèce fixé à α_0 , le test de puissance maximum entre les hypothèses simples ci dessus est défini par la région critique :

$$W = \{ (x_1, \dots, x_n) \mid \frac{L(x_1, \dots, x_n, \theta_0)}{L(x_1, \dots, x_n, \theta_1)} \leq k \}$$

où la valeur de la constante k est déterminée par le risque fixé $\alpha_0 = P_\theta(W|\theta = \theta_0)$.

Exemple. Application de ce théorème au cas de la loi exponentielle de paramètre $\frac{1}{\theta}$. (X v.a. qui a pour densité $f(x, \theta) = \frac{1}{\theta} e^{-\frac{x}{\theta}}$ si $x > 0$ et 0 sinon)

On a vu que $L(x_1, \dots, x_n, \theta) = \frac{1}{\theta^n} e^{-\frac{1}{\theta} \sum_{i=1}^n x_i}$ avec $x_i > 0$.

Supposons par exemple que $\theta_1 > \theta_0$ alors $\frac{L(x_1, \dots, x_n, \theta_0)}{L(x_1, \dots, x_n, \theta_1)} = \left(\frac{\theta_1}{\theta_0}\right)^n e^{\left(\frac{1}{\theta_1} - \frac{1}{\theta_0}\right) \sum_{i=1}^n x_i}$

La région critique est définie par : $\frac{L(x_1, \dots, x_n, \theta_0)}{L(x_1, \dots, x_n, \theta_1)} \leq k \Rightarrow e^{\left(\frac{1}{\theta_1} - \frac{1}{\theta_0}\right) \sum_{i=1}^n x_i} \leq k_1$

en prenant le logarithme, on obtient $\left(\frac{1}{\theta_1} - \frac{1}{\theta_0}\right) \sum_{i=1}^n x_i \leq k_2$ puisque $\theta_1 > \theta_0 \Rightarrow \sum_{i=1}^n x_i \geq c$

La valeur de la constante c , qui va totalement préciser la région critique, est déterminée par la condition $\alpha_0 = P\left(\sum_{i=1}^n X_i \geq c \mid \theta = \theta_0\right)$

On sait que $\frac{2S_n}{\theta}$ suit la loi du χ_{2n}^2 où $S_n = \sum_{i=1}^n X_i$ d'où la condition précédente se réécrit :

$$\alpha_0 = P\left(\frac{2S_n}{\theta_0} \geq \frac{2c}{\theta_0}\right)$$

$$c.a.d. \quad 1 - P\left(\frac{2S_n}{\theta_0} < \frac{2c}{\theta_0}\right) = \alpha_0$$

$$c.a.d. \quad P\left(\frac{2S_n}{\theta_0} < \frac{2c}{\theta_0}\right) = 1 - \alpha_0$$

$$\Rightarrow F_{\chi_{2n}^2}\left(\frac{2c}{\theta_0}\right) = 1 - \alpha_0 \Rightarrow \frac{2c}{\theta_0} = \chi_{2n, 1-\alpha_0}^2$$

$\chi_{2n, 1-\alpha_0}^2$ est le fractile d'ordre $1 - \alpha_0$ de la loi du χ_{2n}^2

$$\text{ainsi on a : } c = \frac{\theta_0}{2} \chi_{2n, 1-\alpha_0}^2$$

La puissance du test est ensuite calculé par :

$$\begin{aligned} \eta &= P(W|\theta = \theta_1) = P\left(\sum_{i=1}^n X_i \geq c|\theta = \theta_1\right) \\ &= P\left(2\frac{S_n}{\theta_1} \geq \frac{2c}{\theta_1}\right) \end{aligned}$$

II.3 Exemples

II.3.1 Test de l'espérance d'une loi normale

a) σ connu

Considérons une v.a. X qui suit une loi normale $N(m, 1)$, un échantillon de taille 16, de moyenne empirique $\bar{x}_n = 2,6$; soit $\alpha = 0,05$ et on veut tester les hypothèses :

$$\begin{cases} H_0 : m = m_0 = 2 \\ H_1 : m = m_1 = 4 \end{cases}$$

Nous allons résoudre ce problème dans le cas général :

$$X \rightsquigarrow N(m, \sigma^2) \quad ; \quad \sigma \text{ connu}$$

(X_1, \dots, X_n) un échantillon et on veut tester :

$$\begin{cases} H_0 : m = m_0 \\ H_1 : m = m_1 \\ m_0 \neq m_1 \end{cases}$$

On sait que la fonction de vraisemblance est :

$$L(x_1, \dots, x_n, m) = \frac{1}{(\sigma\sqrt{2\pi})^n} e^{-\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - m)^2}$$

Détermination de la forme de la zone de rejet W :

$$W = \{(x_1, \dots, x_n) : \frac{L(x_1, \dots, x_n, m_0)}{L(x_1, \dots, x_n, m_1)} \leq k\}$$

$$\begin{aligned}
\text{et } \frac{L(x_1, \dots, x_n, m_0)}{L(x_1, \dots, x_n, m_1)} \leq k &\Leftrightarrow e^{\frac{1}{2\sigma^2} \left(\sum_{i=1}^n (x_i - m_1)^2 - \sum_{i=1}^n (x_i - m_0)^2 \right)} \leq k \\
&\Leftrightarrow \sum_{i=1}^n (x_i - m_1)^2 - \sum_{i=1}^n (x_i - m_0)^2 \leq \ln(k) \times 2\sigma^2 \\
&\Leftrightarrow 2 \sum_{i=1}^n x_i(m_0 - m_1) \leq 2\sigma^2 \ln(k) + n(m_0 - m_1)(m_0 + m_1) \quad (*)
\end{aligned}$$

2 cas se présentent :

$\alpha.$ $m_0 - m_1 > 0$ *c.a.d.* $m_0 > m_1$ *alors*

$$\begin{aligned}
(*) \Rightarrow \sum_{i=1}^n x_i &\leq \frac{\sigma^2 \ln(k)}{m_0 - m_1} + n(m_0 + m_1) \\
\text{ou } \bar{X}_n &\leq \underbrace{\frac{\sigma^2 \ln(k)}{n(m_0 - m_1)} + m_0 + m_1}_K
\end{aligned}$$

On a donc la zone de rejet de la forme :

$$W = \{x \in \mathbb{R}^n \quad \text{tq} \quad \bar{x}_n \leq K\}$$

Détermination de K : on écrit

$$\alpha = P(W|m = m_0) = P(\bar{X}_n \leq K|m = m_0)$$

$$\text{Or } \bar{X}_n \text{ suit } N\left(m, \left(\frac{\sigma}{\sqrt{n}}\right)^2\right)$$

Sous l'hypothèse $H_0 : m = m_0$ on a donc :

$$\bar{X}_n \text{ suit } N\left(m_0, \left(\frac{\sigma}{\sqrt{n}}\right)^2\right)$$

$$\alpha = P(\bar{X}_n \leq K) \Leftrightarrow \alpha = P\left(\sqrt{n} \frac{\bar{X}_n - m_0}{\sigma} \leq \sqrt{n} \frac{K - m_0}{\sigma}\right)$$

$\left(\sqrt{n} \frac{K - m_0}{\sigma}\right)$ est le fractile U_α de la loi normale centrée réduite, fractile dont la valeur est lue sur la table ainsi $\sqrt{n} \frac{K - m_0}{\sigma} = U_\alpha \Rightarrow K = m_0 + U_\alpha \frac{\sigma}{\sqrt{n}}$

La région critique est donc $W = \{(x_1, \dots, x_n) \text{ tq } \bar{x}_n \leq m_0 + U_\alpha \frac{\sigma}{\sqrt{n}}\}$

$\beta.$ De façon analogue, on aurait trouvé pour le cas $m_0 < m_1$,

$$W = \{(x_1, \dots, x_n) \text{ tq } \bar{x}_n \geq m_0 + U_{1-\alpha} \frac{\sigma}{\sqrt{n}}\}$$

Puissance du test, on calcule :

$$\eta = P(W|m = m_1)$$

Sous l'hypothèse $H_1 : m = m_1$ la loi de la v.a. \bar{X}_n suit $N\left(m_1, \frac{\sigma}{\sqrt{n}}\right)$

$$\text{Alors } 1 - \beta = P(\bar{X}_n \leq K | m = m_1)$$

$$c.a.d. 1 - \beta = P(\sqrt{n} \frac{\bar{X}_n - m_1}{\sigma} \leq \sqrt{n} \frac{K - m_1}{\sigma})$$

Puis on lit la table de la loi $N(0,1)$.

Application numérique : $m_0 = 2$, $m_1 = 4$, $n = 16$ et $\sigma = 1$.

$$\text{Sous l'hypothèse } H_0, \bar{X}_n \hookrightarrow N(2, (\frac{1}{4})^2)$$

$$0,05 = P(\bar{X}_n > K) \Leftrightarrow 0,05 = P(\sqrt{16} \frac{\bar{X}_n - 2}{1} \geq \sqrt{16} \frac{K - 2}{1})$$

$$c.a.d. 0,05 = P(N(0,1) \geq 4(K - 2))$$

$$\Rightarrow P(N(0,1) \leq 4(K - 2)) = 0,95$$

$$\Rightarrow 4(K - 2) = 1,645 \quad \Rightarrow K = 2,411$$

La région critique est $W = \{(x_1, \dots, x_n) \text{ tq } \bar{x}_n \geq 2,411\}$

Dans l'exemple, on a une moyenne empirique $\bar{x}_n = 2,6$. On se trouve dans la région critique ou de rejet c.a.d. on va refuser l'hypothèse nulle $H_0 : m = 2$ avec un risque d'erreur de 1^{re} espèce de 5%.

La puissance du test est :

$$1 - \beta = P(W \mid m = m_1)$$

$$\text{Sous l'hypothèse } H_1 : m = m_1 \quad \bar{X}_n \rightarrow N(4, (\frac{1}{4})^2)$$

$$D'où \quad 1 - \beta = P(\bar{X}_n > K \mid m = m_1)$$

$$= P(4 \frac{\bar{X}_n - 4}{1} \geq \frac{4(K - 4)}{1})$$

$$= P(N(0,1) \geq -6,36) \approx 1$$

$$\Rightarrow \beta = 0 \text{ c.a.d. } P(\text{accepter } m = 2 \mid m = 4) \approx 0$$

Remarque importante : la procédure qu'on vient de voir suppose la donnée de α et de la taille de n de l'échantillon. Le risque β de seconde espèce (ou la puissance du test) s'en déduit. On pourrait adopter une démarche différente qui consiste à se donner α , mais également à se fixer une valeur admissible pour le risque de seconde espèce, l'inconnue devient alors la taille n de l'échantillon.

Ainsi dans le cas général avec $m_0 > m_1$, par exemple, si α et β sont donnés on a :

$$\frac{\sqrt{n}}{\sigma} (K - m_0) = U_\alpha \quad \text{d'après la détermination de } W$$

$$\frac{\sqrt{n}}{\sigma} (K - m_1) = U_{1-\beta} \quad \text{d'après le calcul de la puissance}$$

$$\Rightarrow K = m_0 + U_\alpha \frac{\sigma}{\sqrt{n}} = m_1 + U_{1-\beta} \frac{\sigma}{\sqrt{n}}$$

$$\text{Alors } n = \frac{\sigma^2 (U_{1-\beta} - U_\alpha)^2}{(m_0 - m_1)^2}$$

De même, si $m_0 < m_1$ on trouve :

$$n = \frac{\sigma^2 (U_{1-\alpha} - U_\beta)^2}{(m_1 - m_0)^2}$$

b) σ inconnu

Exemple

$$\begin{cases} H_0 : m = 30 \\ H_1 : m > 30 \end{cases}$$

Un échantillon de 15 observations a donné $\bar{X}_n = 37,2$ et $\bar{S}_n = 6,42$.

On a toujours la zone de rejet :

$$W = \{x \in \mathbb{R}^n \mid \bar{x}_n \geq k\}$$

Détermination de K : on écrit $P(W \mid m = m_0) = \alpha$

$$c.a.d. \quad P(\bar{X}_n \geq K \mid m = m_0) = \alpha$$

On avait écrit : $P(\sqrt{n} \frac{\bar{X}_n - m_0}{\sigma} \geq \sqrt{n} \frac{K - m_0}{\sigma}) = \alpha$

or σ est inconnue \Rightarrow on la remplace par $\bar{S}_n = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2}$

$$\text{Alors} \quad P(\sqrt{n} \frac{\bar{X}_n - m_0}{\bar{S}_n} \geq \sqrt{n} \frac{K - m_0}{\bar{S}_n}) = \alpha$$

On a vu que $\sqrt{n} \frac{\bar{X}_n - m_0}{\bar{S}_n}$ suit une loi de Student à n-1 degrés de liberté, qu'on note ici T_{n-1}

$$\text{Alors} \quad P(T_{n-1} \geq \sqrt{n} \frac{K - m_0}{\bar{S}_n}) = \alpha \quad \Rightarrow \quad 1 - P(T_{n-1} \leq \sqrt{n} \frac{K - m_0}{\bar{S}_n})$$

$$\Rightarrow P(T_{n-1} \leq \sqrt{n} \frac{K - m_0}{\bar{S}_n}) = 1 - \alpha$$

Si on pose $t_{n-1,1-\alpha}$ le fractile d'ordre $1 - \alpha$ de la loi de Student à n-1 degrés de liberté alors on obtient :

$$\sqrt{n} \frac{K - m_0}{\bar{S}_n} = t_{n-1,1-\alpha} \quad \Rightarrow \quad K = m_0 + \frac{\bar{S}_n t_{n-1,1-\alpha}}{\sqrt{n}}$$

Ainsi

$$W = \{x \in \mathbb{R}^n \mid \bar{x}_n \geq m_0 + \frac{\bar{S}_n t_{n-1,1-\alpha}}{\sqrt{n}}\}$$

Application

$$m_0 + \frac{\bar{S}_n t_{n-1,1-\alpha}}{\sqrt{n}} = 30 + \frac{6.42 \times 1.761}{\sqrt{15}} \simeq 33$$

On a

$$\bar{x}_n = 37.2 \geq 33 \Rightarrow \text{l'hypothèse } H_0 \text{ est rejetée.}$$

II.3.2 Test de l'écart type d'une loi normale

a) m connu (cas peut fréquent)

On veut tester au niveau α les hypothèses :

$$\begin{cases} H_0 : \sigma = \sigma_0 \\ H_1 : \sigma = \sigma_1 \\ \sigma_0 \neq \sigma_1 \end{cases}$$

La fonction de vraisemblance est :

$$L(x_1, \dots, x_n, \sigma) = \frac{1}{(\sigma\sqrt{2\pi})^n} e^{-\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - m)^2}$$

Détermination de la forme de la zone de rejet W :

$$W = \{(x_1, \dots, x_n) : \frac{L(x_1, \dots, x_n, \sigma_0)}{L(x_1, \dots, x_n, \sigma_1)} \leq k\}$$

$$\text{or } \frac{L(x_1, \dots, x_n, \sigma_0)}{L(x_1, \dots, x_n, \sigma_1)} \leq k \Leftrightarrow \left(\frac{\sigma_1}{\sigma_0}\right)^n e^{\left(\frac{1}{2\sigma_1^2} \sum_{i=1}^n (x_i - m)^2 - \frac{1}{2\sigma_0^2} \sum_{i=1}^n (x_i - m)^2\right)} \leq k$$

$$\Rightarrow n \ln \frac{\sigma_1}{\sigma_0} + \frac{1}{2} \sum_{i=1}^n (x_i - m)^2 \left(\frac{1}{\sigma_1} - \frac{1}{\sigma_0}\right) \left(\frac{1}{\sigma_1} + \frac{1}{\sigma_0}\right) \leq \ln(k)$$

$$\Rightarrow (\sigma_0 - \sigma_1) \sum_{i=1}^n (x_i - m)^2 \leq k' \quad (**)$$

2 cas se présentent :

$\alpha.$ $\sigma_0 - \sigma_1 > 0$ c.a.d. $\sigma_0 > \sigma_1$ alors

$$(**) \Rightarrow \sum_{i=1}^n (x_i - m)^2 \leq k$$

On a donc la zone de rejet de la forme :

$$W = \{x \in \mathbb{R}^n \text{ tq } \sum_{i=1}^n (x_i - m)^2 \leq k\}$$

$\beta.$ De façon analogue, on aurait trouvé pour le cas $\sigma_0 < \sigma_1$,

$$W = \{x \in \mathbb{R}^n \text{ tq } \sum_{i=1}^n (x_i - m)^2 \geq k\}$$

Détermination de K : on écrit

$$\alpha = P(W \mid \sigma = \sigma_0)$$

Supposons $\sigma_0 > \sigma_1$, puisque $\frac{\sum_{i=1}^n (X_i - m)^2}{\sigma^2}$ suit la loi de χ_n^2 alors

$$\alpha = P\left(\frac{\sum_{i=1}^n (X_i - m)^2}{\sigma_0^2} \leq \frac{K}{\sigma_0^2}\right) \implies \alpha = P(\chi_n^2 \leq \frac{K}{\sigma_0^2})$$

Posons $\chi_{n,\alpha}^2$ le fractile d'ordre α de la loi de χ_n^2 alors on aura :

$$\frac{k}{\sigma_0^2} = \chi_{n,\alpha}^2 \implies k = \chi_{n,\alpha}^2 \sigma_0^2$$

La région critique est donc :

$$W = \{(x_1, \dots, x_n) \text{ tq } \sum_{i=1}^n (x_i - m)^2 \leq \chi_{n,\alpha}^2 \sigma_0^2\}$$

Dans le cas où $\sigma_0 < \sigma_1$ on trouve :

$$W = \{(x_1, \dots, x_n) \text{ tq } \sum_{i=1}^n (x_i - m)^2 \geq \chi_{n,1-\alpha}^2 \sigma_0^2\}$$

Puissance du test :

- Cas $\sigma_0 > \sigma_1$

$$\text{On a } \eta = P(W|\sigma = \sigma_1) = 1 - \beta$$

$$\text{d'où } 1 - \beta = P\left(\sum_{i=1}^n (X_i - m)^2 \leq \chi_{n,\alpha}^2 \sigma_0^2 \mid \sigma = \sigma_1\right)$$

Sous l'hypothèse $H_1 : \sigma = \sigma_1$, $\frac{\sum_{i=1}^n (x_i - m)^2}{\sigma_1^2}$ suit la loi de χ_n^2

$$\text{Alors } 1 - \beta = P\left(\chi_n^2 \leq \chi_{n,\alpha}^2 \frac{\sigma_0^2}{\sigma_1^2}\right)$$

c.a.d. $1 - \beta = \mathcal{F}(\chi_{n,\alpha}^2 \frac{\sigma_0^2}{\sigma_1^2})$ où \mathcal{F} est la fonction de répartition de la loi de χ_n^2 .

- Cas $\sigma_0 < \sigma_1$

$$\text{On a } \eta = P(W|\sigma = \sigma_1) = 1 - \beta$$

$$\text{d'où } 1 - \beta = P\left(\sum_{i=1}^n (X_i - m)^2 \geq \chi_{n,1-\alpha}^2 \sigma_0^2 \mid \sigma = \sigma_1\right)$$

$$\text{d'où } 1 - \beta = P\left(\chi_n^2 \geq \chi_{n,1-\alpha}^2 \frac{\sigma_0^2}{\sigma_1^2}\right)$$

$$\text{et donc } 1 - \beta = 1 - \mathcal{F}\left(\chi_{n,1-\alpha}^2 \frac{\sigma_0^2}{\sigma_1^2}\right)$$

b) m inconnue

Exemple

$$\begin{cases} H_0 : \sigma = 3 \\ H_1 : \sigma > 3 \end{cases}$$

Un échantillon de 20 observations a donné $\bar{S}_n = 3,59$ et $\bar{S}_n^2 = 12,89$.
On a toujours la zone de rejet :

$$W = \{x \in \mathbb{R}^n \mid \sum_{i=1}^n (x_i - m)^2 \geq k\}$$

Où m est inconnue donc on la remplace par \bar{x}_n .
La région de rejet devient :

$$W = \{x \in \mathbb{R}^n \mid \sum_{i=1}^n (x_i - \bar{x}_n)^2 \geq k\}$$

Détermination de K : on écrit $P(W \mid \sigma = \sigma_0) = \alpha$

On a vu que $(n-1) \frac{\bar{S}_n^2}{\sigma^2}$ suit la loi χ_{n-1}^2 où $\bar{S}_n^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2$; d'où

$$\begin{aligned} \alpha = P(W \mid \sigma = \sigma_0) &\implies P\left(\sum_{i=1}^n (X_i - \bar{X}_n)^2 \geq k \mid \sigma = \sigma_0\right) = \alpha \\ &\implies P\left((n-1) \frac{\bar{S}_n^2}{\sigma_0^2} \geq \frac{k}{\sigma_0^2}\right) = \alpha \\ &\implies P\left(\chi_{n-1}^2 \geq \frac{k}{\sigma_0^2}\right) = \alpha \\ &\implies P\left(\chi_{n-1}^2 \leq \frac{k}{\sigma_0^2}\right) = 1 - \alpha \end{aligned}$$

Si on pose $\chi_{n-1,1-\alpha}^2$ le fractile d'ordre $1 - \alpha$ de la loi de χ^2 à n - 1 degrés de liberté alors on obtient :

$$\frac{k}{\sigma_0^2} = \chi_{n-1,1-\alpha}^2 \implies k = \sigma_0^2 \chi_{n-1,1-\alpha}^2$$

Ainsi

$$W = \{x \in \mathbb{R}^n \mid \sum_{i=1}^n (x_i - \bar{x}_n)^2 \geq \sigma_0^2 \chi_{n-1,1-\alpha}^2\}$$

Application

$n = 20$, $\bar{S}_n^2 = 12,89$ et on détermine $\chi_{19,0,95}^2 \simeq 30,144$ d'où $K = 3^2 \times (30,144) \simeq 271,29$

alors que la valeur observée $\sum_{i=1}^{20} (x_i - \bar{x}_n)^2 = 12,89 \times 19 \simeq 245$ est inférieure à $K \implies$ on ne rejette pas H_0
au seuil choisi de 0.05.

III Hypothèses multiples

Définition. Un test est uniformément le plus puissant (U.P.P.) si, quelle que soit la valeur θ appartenant à l'hypothèse alternative, sa puissance $1 - \beta(\theta)$ est supérieure à la puissance de tout autre test.

exemple dans le test

$$\begin{cases} H_0 : m = m_0 \\ H_1 : m = m_1 \\ \text{avec } m_1 > m_0 \end{cases}$$

On a vu que la région critique W ne dépend pas explicitement de m_1 et donc cette région critique est la même pour n'importe quel $m_1 > m_0$. Le test précédent est donc U.P.P. pour

$$\begin{cases} H_0 : m = m_0 \\ H_1 : m > m_0 \end{cases}$$

III.1 Test de 2 hypothèses multiples ou composites

On veut tester le problème suivant :

$$\begin{cases} H_0 : \theta \leq \theta_0 \\ H_1 : \theta > \theta_0 \end{cases}$$

On suppose que la loi P_θ de la v. a. r. X est à rapport des vraisemblances monotone (RVM) cela signifie qu'il existe une statistique $T_n = T_n(x_1, \dots, x_n)$ tq. le rapport des vraisemblances $\frac{L(x_1, \dots, x_n, \theta')}{L(x_1, \dots, x_n, \theta)}$ s'exprime comme une fonction croissante de T_n pour toutes les valeurs de θ et θ' qui vérifient $\theta' > \theta$. l'existence d'un test U.P.P. est assuré par le théorème de LEHMANN suivant :

Théorème (de Lehmann). *Il existe un test U.P.P. pour le problème :*

$$\begin{cases} H_0 : \theta \leq \theta_0 \\ H_1 : \theta > \theta_0 \end{cases}$$

Dont la région critique W est donnée par :

$$W = \{(x_1, \dots, x_n) \in \mathbb{R}^n \text{ tq. } T_n(x_1, \dots, x_n) > K\}$$

Où la constante K est déterminée par le risque fixé : $\alpha = P_\theta(W \mid \theta = \theta_0)$.

Remarque 1) Si le problème de test est :

$$\begin{cases} H_0 : \theta \geq \theta_0 \\ H_1 : \theta < \theta_0 \end{cases}$$

Il suffit de changer le sens de l'inégalité dans la définition de W c-a-d on prend

$$W = \{(x_1, \dots, x_n) \in \mathbb{R}^n \text{ tq. } T_n(x_1, \dots, x_n) < K\}.$$

2) Dans le cas où le paramètre testé a une interprétation : espérance, variance \dots , on privilégiera la statistique d'échantillonnage qui estime de façon optimale le paramètre : moyenne empirique, variance empirique, \dots

exemples : ♣ Espérance d'une loi normale

Disposant de n observations indépendantes d'une loi $N(m, \sigma^2)$, on veut tester au niveau α :

$$\begin{cases} H_0 : m \leq m_0 \\ H_1 : m > m_0 \\ \text{avec } \sigma \text{ connu} \end{cases}$$

La loi normale est à rapport de vraisemblances monotone en effet :

$$L(x_1, \dots, x_n, \theta) = \frac{1}{(\sigma\sqrt{2\pi})^n} e^{-\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \theta)^2}$$

$$\text{d'où } \frac{L(x, \theta')}{L(x, \theta)} = e^{\frac{1}{2\sigma^2} \left(\sum_{i=1}^n (x_i - \theta)^2 - \sum_{i=1}^n (x_i - \theta')^2 \right)}$$

$$= e^{\frac{1}{2\sigma^2} \left[n(\theta^2 - \theta'^2) + (2\theta' - 2\theta) \sum_{i=1}^n x_i \right]}$$

$$= e^{n \frac{\theta^2 - \theta'^2}{2\sigma^2}} e^{\frac{\theta' - \theta}{\sigma^2} T_n} \text{ où } T_n(x_1, \dots, x_n) = \sum_{i=1}^n x_i$$

Le rapport $\frac{L(x, \theta')}{L(x, \theta)}$ est une fonction croissante de T_n pour toutes les valeurs de θ et θ' qui vérifient $\theta' > \theta$.
On applique le théorème de Lehmann : il existe un test U.P.P. de région critique W définie par :

$$W = \{(x_1, \dots, x_n) \in \mathbb{R}^n \mid T_n = \sum_{i=1}^n x_i \geq K\}$$

K est déterminée par :

$$\alpha = P(W \mid m = m_0) \text{ d'où}$$

$$\alpha = P\left(\frac{T_n}{n} \geq \frac{K}{n} \mid m = m_0\right) = P\left(\sqrt{n} \frac{\bar{X}_n - m_0}{\sigma} \geq \frac{\sqrt{n}}{\sigma} \left(\frac{K}{n} - m_0\right)\right)$$

$$\text{d'où } P\left(N(0, 1) \geq \frac{\sqrt{n}}{\sigma} \left(\frac{K}{n} - m_0\right)\right) = \alpha$$

$$\implies P\left(N(0, 1) \leq \frac{\sqrt{n}}{\sigma} \left(\frac{K}{n} - m_0\right)\right) = 1 - \alpha$$

Posons $U_{1-\alpha}$ le fractile d'ordre $1 - \alpha$ de $N(0, 1)$ on a alors :

$$\frac{\sqrt{n}}{\sigma} \left(\frac{K}{n} - m_0\right) = U_{1-\alpha} \implies K = nm_0 + \sigma U_{1-\alpha} \sqrt{n}$$

Remarque Si on avait utilisé $T_n = \bar{X}_n$ on aurait trouvé :

$$K = m_0 + \frac{\sigma}{\sqrt{n}} U_{1-\alpha}$$

$$W = \{x \in \mathbb{R}^n \mid \bar{X}_n \geq m_0 + \frac{\sigma}{\sqrt{n}} U_{1-\alpha}\}.$$

Puissance du test :

$$\eta = 1 - \beta = P(W \mid H_1)$$

$$\implies 1 - \beta = P\left(\bar{X}_n \geq m_0 + \frac{\sigma}{\sqrt{n}} U_{1-\alpha} \mid H_1\right)$$

Sous l'hypothèse $H_1 : m > m_0$, $\bar{X}_n \hookrightarrow N(m, (\frac{\sigma}{\sqrt{n}})^2)$ d'où

$$1 - \beta = 1 - \phi(U_{1-\alpha} + \frac{\sqrt{n}}{\sigma}(m_0 - m))$$

ϕ étant la fonction de répartition de la loi $N(0, 1)$.

Application numérique : $n = 16$; $m_0 = 3$; $\sigma = 1$; $\alpha = 5\%$; $U_{0.95} = 1.645$ alors :

$$W = \{x \in \mathbb{R}^n \mid \bar{x}_n \geq 3.411\}$$

$$1 - \beta = 1 - \phi(13.645 - 4m).$$

♣ Ecart type d'une loi normale

On dispose d'un échantillon de taille n extrait d'une loi normale $N(0, \sigma^2)$; l'espérance est donc supposée connue. On veut tester au niveau α les hypothèses :

$$\begin{cases} H_0 : \sigma^2 \leq \sigma_0^2 \\ H_1 : \sigma^2 > \sigma_0^2 \end{cases}$$

Montrons que la loi normale $N(0, \sigma^2)$ est à rapport de vraisemblances monotone :

$$\frac{L(x_1, \dots, x_n, \sigma'^2)}{L(x_1, \dots, x_n, \sigma^2)} = \left(\frac{\sigma}{\sigma'}\right)^n e^{\left(\frac{1}{\sigma^2} - \frac{1}{\sigma'^2}\right) \frac{\sum_{i=1}^n x_i^2}{2}}$$

Si $\sigma' > \sigma$ alors $\sigma'^2 > \sigma^2$ et donc $\frac{1}{\sigma'^2} < \frac{1}{\sigma^2}$ donc $\frac{L(x, \sigma'^2)}{L(x, \sigma^2)}$ est une fonction croissante de $T_n(x) = \sum_{i=1}^n x_i^2$ pour toutes les valeurs de σ' et σ vérifiant $\sigma' > \sigma$. On en déduit l'existence d'un test U.P.P. de niveau α de région critique W définie par :

$$W = \{x \in \mathbb{R}^n \mid \sum_{i=1}^n x_i^2 \geq K\}$$

K est déterminée par :

$$\alpha = P(W \mid H_0)$$

Sous l'hypothèse H_0 , $\frac{\sum_{i=1}^n X_i^2}{\sigma_0^2}$ suit la loi de χ_n^2 d'où :

$$\begin{aligned} P(W \mid H_0) = \alpha &\implies P\left(\sum_{i=1}^n X_i^2 \geq K \mid H_0\right) = \alpha \\ \implies \alpha &= P\left(\frac{\sum_{i=1}^n X_i^2}{\sigma_0^2} \geq \frac{K}{\sigma_0^2}\right) = 1 - P\left(\chi_n^2 \leq \frac{K}{\sigma_0^2}\right) \\ &\implies P\left(\chi_n^2 \leq \frac{K}{\sigma_0^2}\right) = 1 - \alpha \end{aligned}$$

Posons $C_{1-\alpha}$ le fractile d'ordre $1 - \alpha$ de la loi χ_n^2 on a alors $\frac{K}{\sigma_0^2} = C_{1-\alpha}$ d'où $K = \sigma_0^2 C_{1-\alpha}$ et la région de rejet est donnée par ;

$$W = \{x \in \mathbb{R}^n \mid \sum_{i=1}^n x_i^2 \geq \sigma_0^2 C_{1-\alpha}\}$$

La puissance du test :

$$1 - \beta = P(W \mid H_1) = P\left(\sum_{i=1}^n X_i^2 \geq \sigma_0^2 C_{1-\alpha} \mid H_1\right)$$

$$1 - \beta = P\left(\frac{\sum_{i=1}^n X_i^2}{\sigma^2} \geq \frac{\sigma_0^2}{\sigma^2} C_{1-\alpha}\right)$$

$$1 - \beta = P\left(\chi_n^2 \geq \frac{\sigma_0^2}{\sigma^2} C_{1-\alpha}\right)$$

Application numérique : $n = 10$; $\sigma_0 = 2$; $\alpha = 10\%$;

$$\frac{\sum_{i=1}^{10} X_i^2}{2^2} \hookrightarrow \chi_{10}^2$$

$$P(\chi_{10}^2 > \frac{K}{4}) = 0.10 \implies P(\chi_{10}^2 \leq \frac{K}{4}) = 0.90$$

$$\implies \frac{K}{4} \simeq 16 \implies K \simeq 64$$

d'où :

$$W = \{x \in \mathbb{R}^n \mid \sum_{i=1}^n x_i^2 > 64\}.$$

IV Tests de comparaison

Il est fréquent d'avoir à comparer entre elles 2 populations différentes. Nous allons supposer que l'on désire comparer une même v.a.r. X , normale sur 2 populations :

Hypothèses et notations :

	Population I	Population II
Loi	$N(m_1, \sigma_1^2)$	$N(m_2, \sigma_2^2)$
Taille de l'échantionn	n_1	n_2
Moyenne empirique	\bar{x}_1	\bar{x}_2
Variance empirique	\bar{s}_1^2	\bar{s}_2^2

IV.1 Test de Fisher de l'égalité des variances

On se place dans le cas général où les espérances m_1 et m_2 sont inconnues ; α est donné ; et on veut tester :

$$\begin{cases} H_0 : \sigma_1^2 = \sigma_2^2 \\ H_1 : \sigma_1^2 \neq \sigma_2^2 \end{cases}$$

On sait que :

$$(n_1 - 1) \frac{\overline{S}_1^2}{\sigma_1^2} \hookrightarrow \chi_{n_1-1}^2 \text{ et } (n_2 - 1) \frac{\overline{S}_2^2}{\sigma_2^2} \hookrightarrow \chi_{n_2-1}^2$$

$$(*) \quad \frac{\frac{\overline{S}_1^2}{\sigma_1^2}}{\frac{\overline{S}_2^2}{\sigma_2^2}} \text{ suit la loi de Fisher-Snedecor } F_{n_1-1, n_2-1}$$

avec les notations :

$$\overline{S}_1^2 = \frac{1}{n_1 - 1} \sum_{i=1}^{n_1} (X_i - \overline{X}_1)^2 \text{ et } \overline{S}_2^2 = \frac{1}{n_2 - 1} \sum_{i=1}^{n_2} (X_i - \overline{X}_2)^2$$

Sous l'hypothèse $H_0 : \sigma_1^2 = \sigma_2^2$, on peut interpréter le rapport (*) comme le rapport des 2 estimateurs de σ_1^2 et de σ_2^2 respectivement. Si $\sigma_1 = \sigma_2$ ce rapport ne doit pas différer significativement de 1. Le rapport (*) sera la variable de décision. Dans la pratique on met toujours au numérateur la plus grande des 2 quantités :

$$\overline{S}_1^2 \text{ et } \overline{S}_2^2.$$

et la région critique W sera de la forme :

$$W = \{(x_1, \dots, x_n) \in \mathbb{R}^n \mid F > K \text{ avec } K > 1\}$$

La constante K est déterminé par :

$$P(W \mid H_0) = \alpha \implies P\left(\frac{\overline{S}_1^2}{\overline{S}_2^2} > K\right) = \alpha$$

$$\implies P(F_{n_1-1, n_2-1} > K) = \alpha \iff P(F_{n_1-1, n_2-1} \leq K) = 1 - \alpha$$

Si on pose $f_{1-\alpha}$ le fractile d'ordre $1 - \alpha$ de la loi de Fisher F_{n_1-1, n_2-1} on aura $K = f_{1-\alpha}$; d'où

$$W = \{(x_1, \dots, x_n) \in \mathbb{R}^n \mid F > f_{1-\alpha}\}$$

Application numérique : $n_1 = 25$; $n_2 = 13$; $S_1^2 = 0.05$; $S_2^2 = 0.07$; $\alpha = 5\%$;

$$\text{On a } \spadesuit \overline{S}_1^2 = \frac{n_1 S_1^2}{n_1 - 1} = \frac{25 \times 0.05}{24} \simeq 0.052$$

$$\spadesuit \overline{S}_2^2 = \frac{n_2 S_2^2}{n_2 - 1} = \frac{13 \times 0.07}{12} \simeq 0.076$$

On calcule alors $\frac{\overline{S}_2^2}{\overline{S}_1^2}$ qui suit la loi de Fisher $F_{12, 24}$ et on lit $f_{1-\alpha}$ qui est le fractile d'ordre 0.95 de la loi de Fisher $F_{12, 24}$, on trouve $f_{0.95} = 2.18$ et donc la région de rejet est alors :

$$W = \{(x_1, \dots, x_n) \in \mathbb{R}^n \mid F > 2.18\}$$

Dans notre cas on a trouvé $\frac{\overline{S}_2^2}{\overline{S}_1^2} = \frac{0.076}{0.052} \simeq 1.46$

On n'est pas dans la zone de rejet donc on accepte l'hypothèse $H_0 : \sigma_1 = \sigma_2$.

Remarque : Si les deux échantillons ont la même taille c-a-d $n_1 = n_2$ alors le calcul se simplifie un peu et on a alors

$$F_{n-1, n-1} = \frac{\overline{S}_1^2}{\overline{S}_2^2} = \frac{S_1^2}{S_2^2}$$

IV.2 Test de Student de l'égalité des espérances

♣ Variances connues

On teste :

$$\begin{cases} H_0 : m_1 = m_2 \\ H_1 : m_1 \neq m_2 \end{cases}$$

On a $\overline{X}_1 \hookrightarrow N\left(m_1, \left(\frac{\sigma_1}{\sqrt{n_1}}\right)^2\right)$ et $\overline{X}_2 \hookrightarrow N\left(m_2, \left(\frac{\sigma_2}{\sqrt{n_2}}\right)^2\right)$. Si on pose $Y = \overline{X}_1 - \overline{X}_2$ alors

la v. a. r. $Y \hookrightarrow N\left(m_1 - m_2, \left(\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right)^2\right)$

Si on pose $\theta = m_1 - m_2$ alors le test devient :

$$\begin{cases} H_0 : \theta = 0 \\ H_1 : \theta \neq 0 \end{cases}$$

Il existe un test U.P.P. de niveau α de région critique $W = \{(y_1, \dots, y_n) \in \mathbb{R}^n \text{ tq. } |Y| > K\}$.

La constante K est déterminée par : $P(W | H_0) = \alpha$.

Sous l'hypothèse H_0 , $Y \hookrightarrow N\left(0, \left(\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right)^2\right)$ d'où :

$$\alpha = P(|Y| > K) \iff \alpha = 1 - P(|Y| \leq K)$$

$$\implies P\left(\frac{-K}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \leq \frac{Y}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \leq \frac{K}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}\right) = 1 - \alpha$$

$$\phi\left(\frac{K}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}\right) = 1 - \frac{\alpha}{2}$$

où ϕ est la fonction de répartition de la loi $N(0, 1)$. Si on pose $U_{1-\frac{\alpha}{2}}$ alors :

$$\frac{K}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = U_{1-\frac{\alpha}{2}} \implies K = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \cdot U_{1-\frac{\alpha}{2}}$$

La région de rejet est :

$$W = \left\{ |\overline{x}_1 - \overline{x}_2| > \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \cdot U_{1-\frac{\alpha}{2}} \right\}$$

♣ Variances inconnues et égales

Cette hypothèse revient à supposer que l'on a, au préalable, effectué le test d'égalité des variances et que l'on a retenu $\sigma_1^2 = \sigma_2^2$. On pose alors $\sigma^2 = \sigma_1^2 = \sigma_2^2$

On sait que :

$$(n_1 - 1) \frac{\overline{S}_1^2}{\sigma^2} \hookrightarrow \chi_{n_1-1}^2 \text{ et } (n_2 - 1) \frac{\overline{S}_2^2}{\sigma^2} \hookrightarrow \chi_{n_2-1}^2$$

On a $\overline{X}_1 \hookrightarrow N\left(m_1, \left(\frac{\sigma}{\sqrt{n_1}}\right)^2\right)$ et $\overline{X}_2 \hookrightarrow N\left(m_2, \left(\frac{\sigma}{\sqrt{n_2}}\right)^2\right)$.

de même la v. a. $Y \hookrightarrow N\left(m_1 - m_2, \left(\sigma \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\right)^2\right)$ et $\frac{(n_1 - 1)\overline{S}_1^2 + (n_2 - 1)\overline{S}_2^2}{\sigma^2} \hookrightarrow \chi_{n_1+n_2-2}^2$

Par définition de la variable de Student :

$$\begin{aligned} T_{n_1+n_2-2} &= \frac{(\overline{X}_1 - \overline{X}_2) - (m_1 - m_2)}{\sigma \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \\ &= \frac{(\overline{X}_1 - \overline{X}_2) - (m_1 - m_2)}{\sqrt{\frac{(n_1 - 1)\overline{S}_1^2 + (n_2 - 1)\overline{S}_2^2}{\sigma^2(n_1 + n_2 - 2)}}} \\ &= \frac{(\overline{X}_1 - \overline{X}_2) - (m_1 - m_2)}{\sqrt{\left((n_1 - 1)\overline{S}_1^2 + (n_2 - 1)\overline{S}_2^2\right) \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sqrt{n_1 + n_2 - 2} \end{aligned}$$

Il existe un test U.P.P. de niveau α défini par la région critique $W = \{ |T| > K \}$, la constante K est déterminée par le risque α :

$$\alpha = P(W | H_0) = P(|T| > K | H_0)$$

$$\implies P(|T_{n_1+n_2-2}| > K) = \alpha$$

$$\implies 1 - P(|T_{n_1+n_2-2}| \leq K) = \alpha$$

$$\implies P(-K \leq T_{n_1+n_2-2} \leq K) = 1 - \alpha$$

Si on pose $t_{1-\frac{\alpha}{2}}$ le fractile d'ordre $1 - \frac{\alpha}{2}$ de la loi de Student à $n_1 + n_2 - 2$ degrés de liberté alors $K = t_{1-\frac{\alpha}{2}}$ alors la région de rejet est :

$$W = \{ |T| > t_{1-\frac{\alpha}{2}} \}.$$

V Le test du χ^2

V.1 Exemple introductif

On jette un dé 300 fois. On obtient les résultats suivants :

face obtenue	1	2	3	4	5	6
nombre de lancers	42	43	56	55	43	61

Peut-on en conclure que le dé est équilibré ?

Remarque : Si le dé est équilibré alors la v. a. r. associée à cette expérience suit une loi uniforme discrète de paramètre $\frac{1}{6}$.

Une idée naturelle est de dire que, si le dé est équilibré, on devrait avoir à peu près $\frac{300}{6} = 50$ fois chaque face. Si le résultat s'éloigne trop de 50 pour quelques unes des faces, on peut douter du fait que le dé est

équilibré. On peut donc penser à rejeter l'hypothèse que le dé est équilibré si la "distance" entre les vecteurs $(42, 43, 56, 55, 43, 61)$ et $(50, 50, 50, 50, 50, 50)$ est "trop grande". Il reste à choisir une distance appropriée. Plus généralement, on s'intéresse à une expérience qui a r issues possibles. On sait que, sous une certaine hypothèse H_0 , les probabilités d'apparition de ces r issues sont respectivement p_1, \dots, p_r (avec $\sum_{j=1}^r p_j = 1$.)

On fait n expériences identiques et indépendantes et on compte les nombres n_j de fois où l'issue j s'est produite. on a forcément $\sum_{j=1}^r n_j = n$. Le problème est de décider si l'observation des n_1, \dots, n_r est compatible avec l'hypothèse H_0 que les probabilités des issues sont p_1, \dots, p_r .

Dans l'exemple $r = 6, \forall j \in \{1, \dots, 6\}, p_j = \frac{1}{6}$ et $n = 300$.

Sous l'hypothèse H_0 on s'attend à observer en moyenne np_j fois l'issue j (50 fois chaque face dans l'exemple.) Il s'agit alors de déterminer si les observations n_j de l'issue j sont significativement proches ou éloignées de l'effectif théorique np_j de l'issue j . On peut alors penser à une région critique de la forme :

$$W = \left\{ \sum_{j=1}^r (n_j - np_j)^2 > l_\alpha \right\}$$

Pour déterminer l_α , il faut connaître la loi de probabilité sous l'hypothèse H_0 de $\sum_{j=1}^r (n_j - np_j)^2$ ou d'une

v. a. analogue.

Il est clair que pour chaque j , N_j est de loi binomiale $\mathcal{B}(n, p_j)$; on dit que le vecteur (N_1, \dots, N_r) est de loi **multinomiale** $\mathcal{M}(n; p_1, \dots, p_r)$. Le test du χ^2 est basé sur le théorème suivant :

Théorème (de Pearson). *Si (N_1, \dots, N_r) est de loi multinomiale $\mathcal{M}(n; p_1, \dots, p_r)$ alors*

$$\Delta_n^2 = \sum_{j=1}^r \frac{(N_j - np_j)^2}{np_j} \xrightarrow[n \rightarrow +\infty]{\text{loi}} \chi_{r-1}^2$$

Définition. On appelle **test du χ^2** le test de H_0 : les probabilités des r issues sont p_1, \dots, p_r contre $H_1 = \bar{H}_0$ défini par la région critique :

$$W = \left\{ \sum_{j=1}^r \frac{(n_j - np_j)^2}{np_j} > \chi_{r-1, \alpha}^2 \right\}$$

Dans l'exemple du dé, l'hypothèse que le dé est équilibré s'écrit : $H_0 : \forall j \in \{1, \dots, 6\}, p_j = \frac{1}{6}$. Alors la statistique de décision vaut $\delta_n^2 = \frac{(42 - 50)^2}{50} + \dots + \frac{(61 - 50)^2}{50} = 6.88$. Au seuil $\alpha = 0.05, \chi_{5, 0.05}^2 = 11.07$. On a 6.88 est inférieur à 11.07, donc on ne rejette pas l'hypothèse H_0 .

V.2 Le test d'adéquation du χ^2 à une loi donnée

Soit X une v. a. r. On dispose d'un n -échantillon de valeurs (x_1, \dots, x_n) . On veut savoir si ces observations peuvent être considérées comme des réalisations de v. a. r. X_1, \dots, X_n indépendantes et de même loi donnée. Un tel test est appelé **test d'adéquation** ou **test d'ajustement**. Soit F la fonction de répartition inconnue des v. a. X_i , il s'agit de tester

$$H_0 : F = F_0 \quad \text{contre} \quad H_1 : F \neq F_0$$

où F_0 est la fonction de répartition de la loi dont on souhaite tester que les observations sont issues.

Si les v. a. observées sont discrètes, les n observations x_1, \dots, x_n ont pris r valeurs différentes e_1, \dots, e_r . Soit n_j le nombre d'observations égales à e_j . On peut calculer les $p_j = P(X = e_j)$. (N_1, \dots, N_r) est de loi

multinomiale $\mathcal{M}(n; p_1, \dots, p_r)$, on peut alors appliquer le théorème de Pearson et utiliser le test du χ^2 pour tester l'adéquation de l'échantillon à la loi de X .

Si les v. a. observées sont continues, il faut regrouper les observations en classes. On partitionne le support de la loi à tester en r intervalles de $[a_{j-1}, a_j]$ et on compte le nombre d'observations appartenant à chaque classe n_j . Sous l'hypothèse $H_0 : F = F_0$, la probabilité qu'une observation appartienne à j est :

$p_j = P(a_{j-1} < X \leq a_j) = F_0(a_j) - F_0(a_{j-1})$. On se trouve alors dans le cadre du théorème de Pearson et on peut appliquer le test du χ^2 .

Remarque

✘ Le résultat du théorème de Pearson n'est qu'asymptotique, donc il n'est valable que pour n assez grand ($n \geq 50$), de plus chaque classe doit contenir au moins 5 observations.

✘ Si F_0 dépend de k paramètres inconnus, on estime ces derniers à partir de l'échantillon et donc la statistique Δ_n^2 suivra asymptotiquement une loi du χ^2 à $r - k - 1$ degrés de liberté.

VI Le test d'indépendance

Soient X et Y 2 caractères quelconques définis sur une population et un échantillon de taille n issu de cette population. On dispose d'observations réparties en r ($r > 1$) classes ou modalités pour le caractère X et en s ($s > 1$) classes ou modalités pour le caractère Y . On veut tester l'hypothèse nulle H_0 : "les caractères X et Y sont statistiquement indépendants" contre l'hypothèse alternative H_1 : "les caractères X et Y ne sont pas statistiquement indépendants"

Définition. En munissant l'échantillon de l'équiprobabilité, on définit la propriété d'indépendance statistique de la façon suivante : les deux caractères X et Y sont statistiquement indépendants si, et seulement si, pour tout i allant de 1 jusqu'à r et pour tout j allant de 1 jusqu'à s on doit avoir :

$$P((X \in C_i) \cap (Y \in C_j)) = P(X \in C_i) \times P(Y \in C_j)$$

où C_i est la $i^{\text{ème}}$ classe (ou modalité) du premier caractère et C_j est la $j^{\text{ème}}$ classe (ou modalité) du second caractère.

Notation et propriétés

On note n_{ij} le nombre d'observations, parmi les n de l'échantillon, appartenant à C_i et à C_j , puis $n_{i\bullet} = \sum_{j=1}^s n_{ij}$

et $n_{\bullet j} = \sum_{i=1}^r n_{ij}$. On rappelle que $n = \sum_{i=1}^r n_{i\bullet} = \sum_{j=1}^s n_{\bullet j}$ (les effectifs marginaux).

Si X et Y sont deux caractères sont statistiquement indépendants dont les distributions sont définies respectivement par les effectifs $(n_{i\bullet})_{1 \leq i \leq r}$ et $(n_{\bullet j})_{1 \leq j \leq s}$ alors les effectifs théoriques des observations appartenant à la $i^{\text{ème}}$ classe du premier caractère et à la $j^{\text{ème}}$ classe du second sont égaux à :

$$n'_{ij} = \frac{n_{i\bullet} n_{\bullet j}}{n}$$

On construit alors le test d'indépendance à partir de la distance du χ^2 entre les effectifs théoriques et les effectifs observés ; On a la statistique :

$$D_n^2 = \sum_{i=1}^r \sum_{j=1}^s \frac{(n_{ij} - n'_{ij})^2}{n'_{ij}}$$

La région critique de ce test est de la forme :

$$W = \left\{ d_n^2 = \sum_{i=1}^r \sum_{j=1}^s \frac{(n_{ij} - n'_{ij})^2}{n'_{ij}} > C \right\}$$

Pour un risque de première espèce $\alpha = P(D_n^2 > C) | H_0$, la valeur de la constante C est approximé par le fractile d'ordre $1 - \alpha$ de la loi $\chi_{(r-1)(s-1)}^2$.

Exemple Une enquête sur le tabagisme a donné les résultats suivants :

	Hommes	Femmes
Fumeurs	420	75
Non fumeurs	280	225

On désigne par X le caractère qui spécifie si une personne est fumeur ou non et par Y le caractère qui spécifie son sexe. X est un caractère à $r = 2$ modalités : « fumeurs », « non fumeurs » et Y est un caractère à $s = 2$ modalités : « homme », « femme ». Les caractères X et Y sont-ils statistiquement indépendants ?

On calcule les effectifs marginaux : $n_{1\bullet} = 420 + 75 = 495$, $n_{2\bullet} = 280 + 225 = 505$, $n_{\bullet 1} = 420 + 280 = 700$, $n_{\bullet 2} = 75 + 225 = 300$.

Puis les effectifs théoriques, pour i de 1 à 2 et pour j de 1 à 2 par la formule $n'_{ij} = \frac{n_{i\bullet} n_{\bullet j}}{n}$

Les calculs sont effectués dans le tableau suivant :

	Hommes	Femmes	$n_{i\bullet}$
Fumeurs	$n'_{11} = \frac{700 \times 495}{1000} = 346.5$	$n'_{12} = \frac{300 \times 495}{1000} = 148.5$	495
Non fumeurs	$n'_{21} = \frac{700 \times 505}{1000} = 353.5$	$n'_{22} = \frac{300 \times 505}{1000} = 151.5$	505
$n_{\bullet j}$	700	300	$n = 1000$

On calcule la valeur de la statistique $d_n^2 = \sum_{i=1}^r \sum_{j=1}^s \frac{(n_{ij} - n'_{ij})^2}{n'_{ij}}$

$$d_n^2 = \frac{(420 - 346.5)^2}{346.5} + \frac{(75 - 148.5)^2}{148.5} + \frac{(280 - 353.5)^2}{353.5} + \frac{(225 - 151.5)^2}{151.5} \simeq 102.91$$

Pour tester $H_0 : X$ et Y sont statistiquement indépendants contre $H_1 : X$ et Y ne sont pas statistiquement indépendants, au risque $\alpha = 0.05$, on lit le fractile d'ordre $1 - \alpha$ de la loi du χ^2 à 1 degré de liberté et on trouve que $\chi_{0.95}^2 = 3.84$.

Puisque $d_n^2 > \chi_{0.95}^2$, on rejette donc l'hypothèse nulle.