STATISTIQUES-L2BIO

M. BOUTAHAR

Espérance mathématique

Soit X une variable aléatoire qui prends ses valeurs dans $E = \{x_1, x_2, ...\}$.

Definition

On définit l'espérance mathématique de X par:

$$\begin{aligned} E(X) &= \sum_{k=1}^{\infty} x_k P(X = x_k) \\ &= \sum_{k=1}^{\infty} x_k p_k, \end{aligned}$$

$p_k = P(X = x_k)$.

Definition

On définit l'espérance mathématique de X^2 par:

$$E(X^2) = \sum_{k=1}^{\infty} x_k^2 p_k .$$

1.Lois discrètes

On définit alors:

-La variance de X par

$$\begin{aligned} var(X) &= E(X^2) - (E(X))^2 \\ &= \sum_{k=1}^{\infty} x_k^2 p_k - \left(\sum_{k=1}^{\infty} x_k p_k \right)^2 . \end{aligned}$$

-L'écart-type de X par

$$\sigma(X) = \sqrt{\text{var}(X)}.$$

Si E est fini $E = \{x_1, x_2, ..., x_r\}$. Alors

$$E(X) = \sum_{k=1}^{r} x_k p_k,$$

$$E(X^2) = \sum_{k=1}^{r} x_k^2 p_k$$

$$var(X) = \sum_{k=1}^{r} x_k^2 p_k - \left(\sum_{k=1}^{r} x_k p_k \right)^2.$$

Lois usuelles

1. Loi de Bernoulli

Definition

On dit que X suit une loi de Bernoulli de paramètre p, noté $X \rightsquigarrow B(p)$, si $E = \{x_1, x_2\}$ avec $P(X = x_1) = p$, et $P(X = x_2) = 1 - p$.

Exemple 1. : On lance une pièce de monnaie, X est le résultat obtenu, $P(X = pile) = P(X = face) = 1/2$.

Lois usuelles

1. Loi de Bernoulli

Definition

On dit que X suit une loi de Bernoulli de paramètre p, noté $X \rightsquigarrow B(p)$, si $E = \{x_1, x_2\}$ avec $P(X = x_1) = p$, et $P(X = x_2) = 1 - p$.

Exemple 1. : On lance une pièce de monnaie, X est le résultat obtenu, $P(X = pile) = P(X = face) = 1/2$.

1.Lois discrètes

2. Loi Binomiale

Definition

Soit $Y_i, 1 \leq i \leq n$, une suite de variables aléatoires de Bernoulli indépendantes à valeurs dans $\{0, 1\}$ avec $P(Y_k = 1) = p$, alors la variable $X = \sum_{k=1}^{n} Y_k$ suit une loi Binomiale, noté $X \rightsquigarrow B(n, p)$, dans ce cas

$$E = \{0, 1, ..., n\}$$

$$P(X = k) = C_n^k p^k (1-p)^{n-k}, 0 \leq k \leq n,$$

$$C_n^k = \frac{n!}{k!(n-k)!}.$$

Definition

On définit l'espérance mathématique de X par

$$E(X) = \int_{\mathbb{R}} x f_X(x) dx.$$

$f_X(x)$ est la fonction densité de la variable X.

Definition

On définit l'espérance mathématique de X^2 par:

$$E(X^2) = \int_{\mathbb{R}} x^2 f_X(x) dx.$$

On définit alors la variance de X par

$$\begin{aligned} var(X) &= E(X^2) - (E(X))^2 \\ &= \int_{\mathbb{R}} x^2 f_X(x) dx - \left(\int_{\mathbb{R}} x f_X(x) dx \right)^2 \end{aligned}$$

et l'écart-type de X par

$$\sigma(X) = \sqrt{\text{var}(X)}.$$

Lois usuelles

1 Loi uniforme

Definition

On dit que X suit une loi uniforme sur $[a, b]$, noté $X \rightsquigarrow U[a, b]$, si elle admet la densité

$$f(x) = \frac{1}{b-a}\mathbf{1}_{[a,b]}(x) = \begin{cases} \frac{1}{b-a} & \text{si } x \in [a, b] \\ 0 & \text{sinon.} \end{cases} \quad (1)$$

On montre que

$$E(X) = (a+b)/2, V(X) = (b-a)^2/12.$$

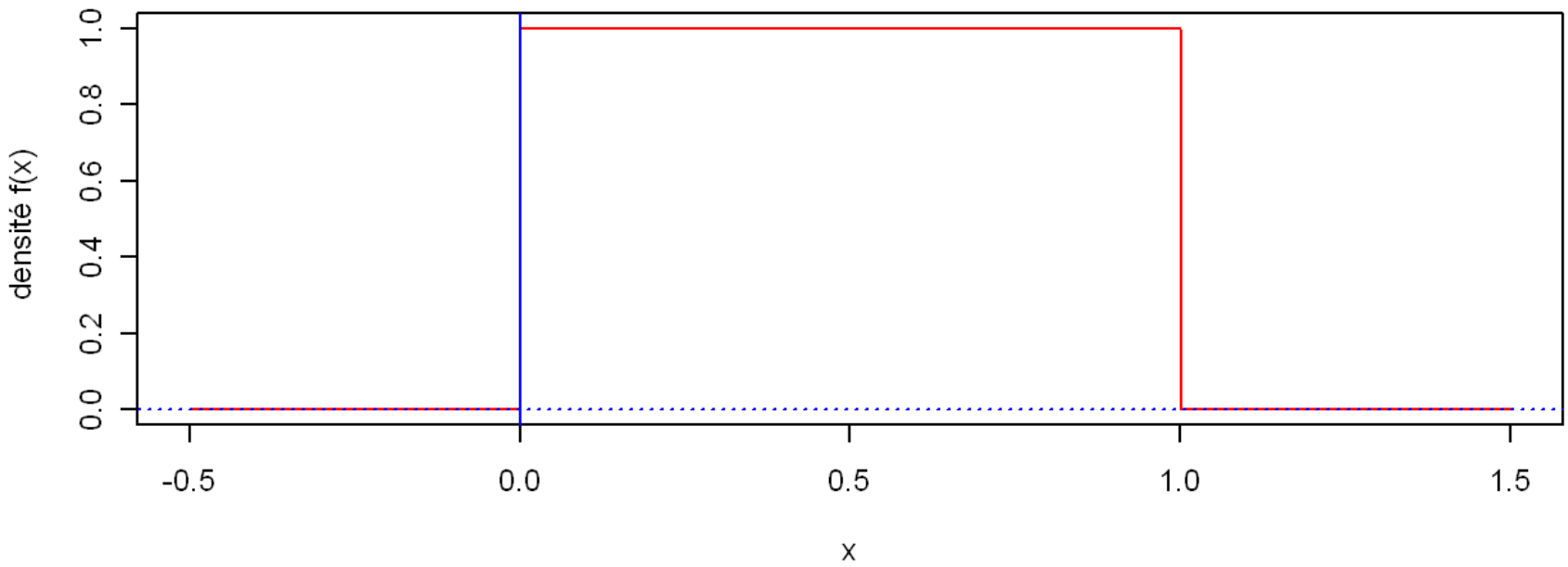


Figure: Densité de la loi uniforme U[0,1]

2. Loi de Gauss (ou normale)

Definition

On dit que X suit une loi de Gauss (ou normale) de moyenne m et de variance σ^2, noté $X \rightsquigarrow N(m, \sigma^2)$, si elle admet la densité

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-m)^2}{2\sigma^2}}, \forall x \in \mathbb{R}. \quad (2)$$

On montre que $E(X) = m, V(X) = \sigma^2$.

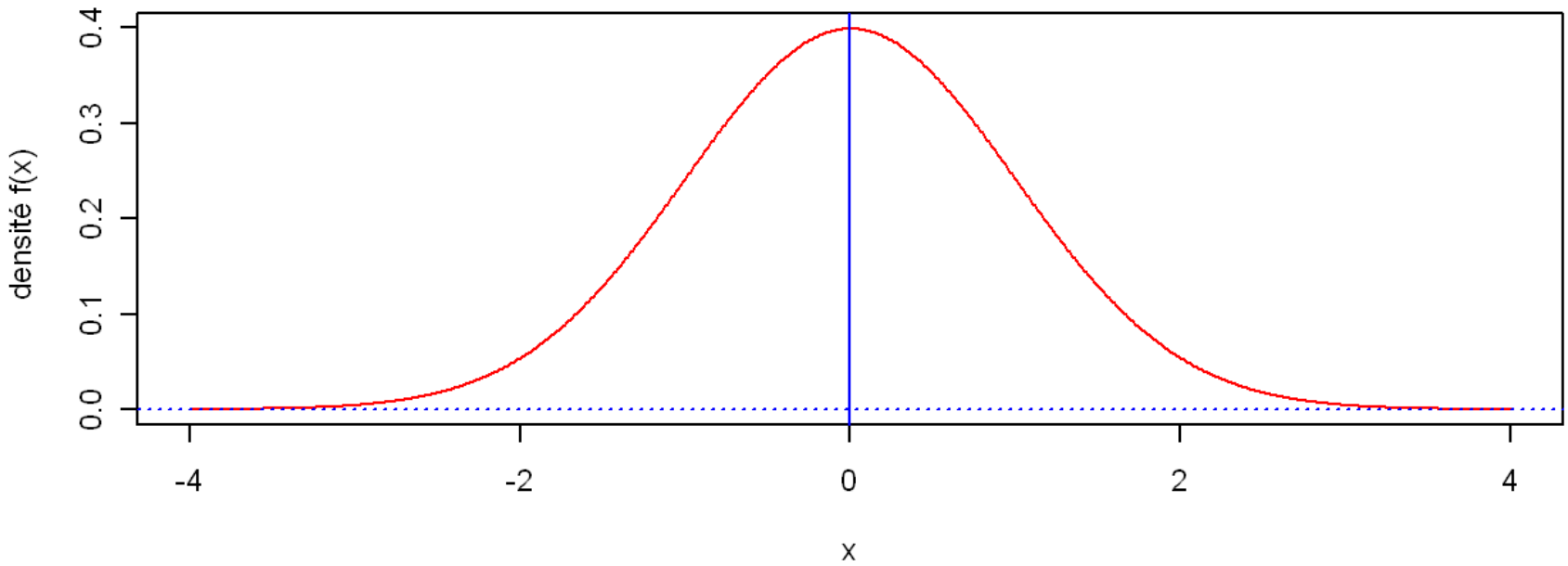


Figure: Densité de la loi Gauss N(0,1)

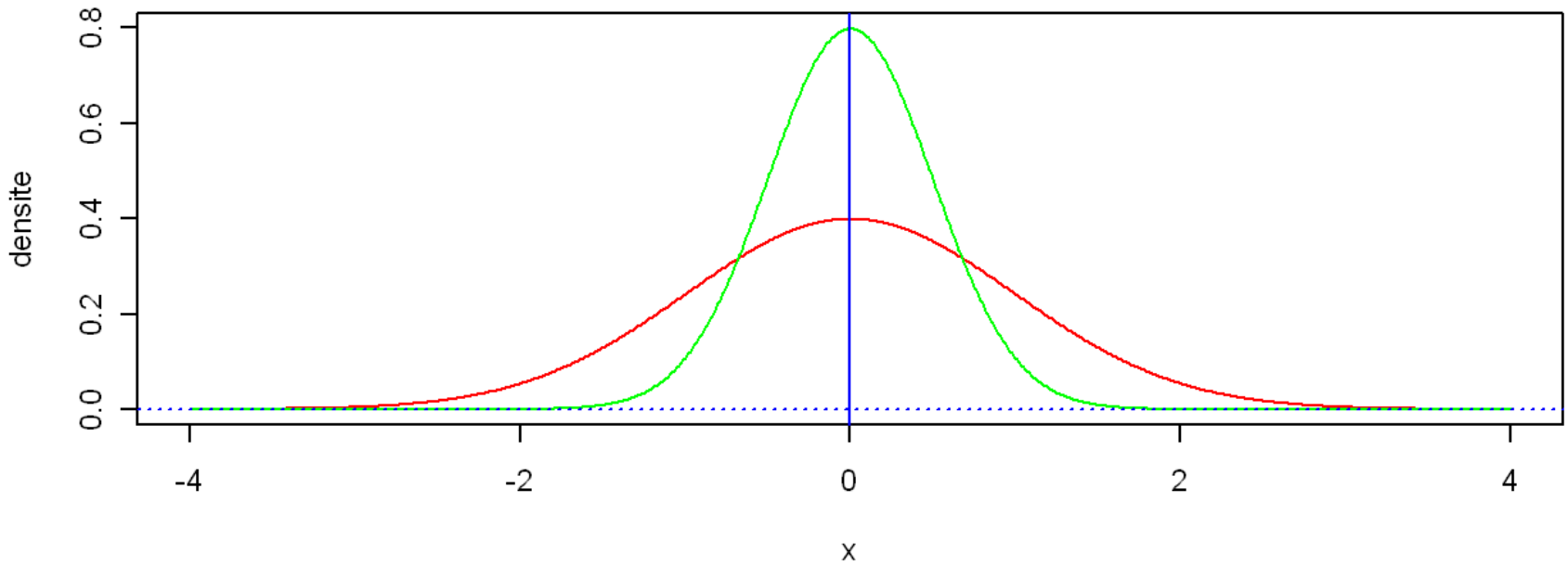


Figure: Densités Gaussiennes, rouge:$N(0,1)$, vert: $\mathrm{N}(0,\frac{1}{2})$.

3. Loi exponentielle

Definition

On dit que X suit une loi exponentielle de paramètre $\lambda > 0$, noté $X \rightsquigarrow E(\lambda)$, si elle admet la densité

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & \text{si } x \geq 0 \\ \\ 0 & \text{sinon.} \end{cases} \tag{3}$$

On montre que $E(X) = \frac{1}{\lambda}, V(X) = \frac{1}{\lambda^2}$.

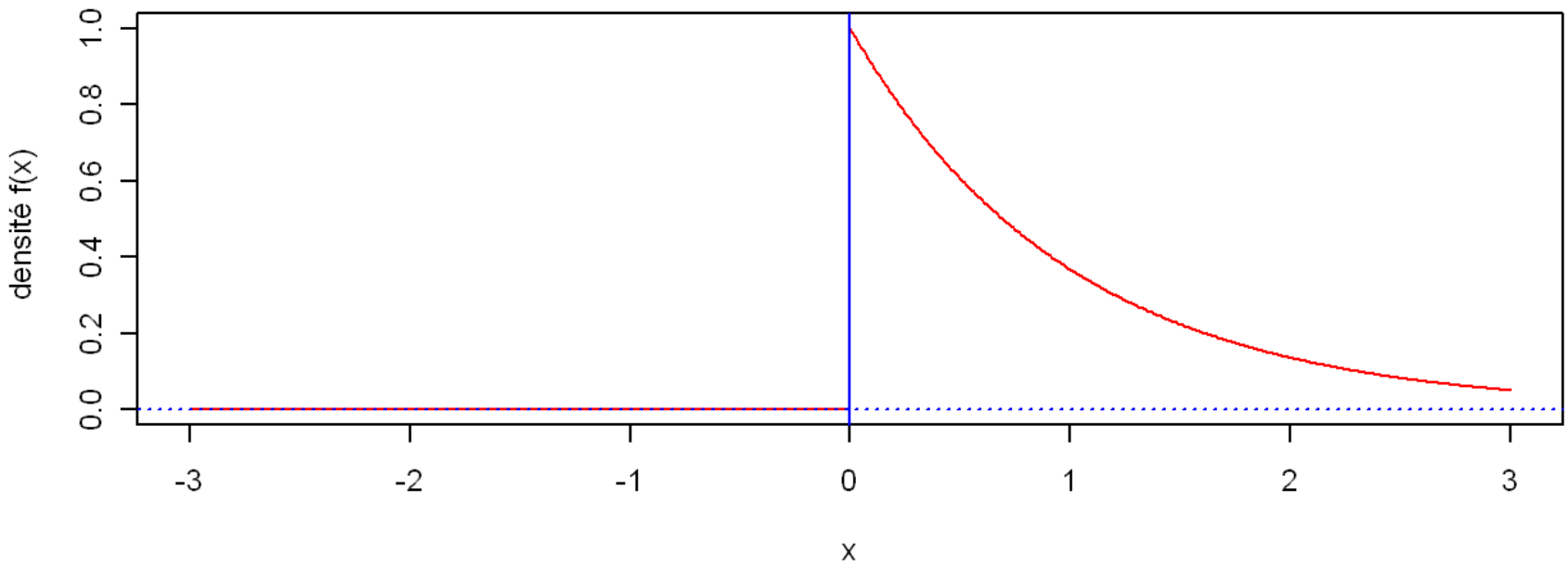


Figure: Densité de la loi exponentielle E(1)

4. Loi de Khi-deux

Definition

On dit que X suit une loi Khi-deux à n degrés de liberté , noté $X \rightsquigarrow \mathcal{X}^2(n)$, si $X = \sum_{k=1}^{n} Z_k^2$ où $(Z_k, 1 \leq k \leq n)$ est une suite de variables aléatoires indépendantes et identiquement distribuées (i.i.d.) suivant une loi normale $N(0,1)$.

5. Loi de Student

Definition

On dit que X suit une loi Student à n degrés de liberté , noté $X \rightsquigarrow t(n)$, si $X = \frac{Z}{\sqrt{Y/n}}$, $Z \rightsquigarrow N(0,1)$, $Y \rightsquigarrow \mathcal{X}^2(n)$, et Z et Y sont indépendantes.

6. Loi de Fisher

Definition

On dit que X suit une loi Fisher à (n_1, n_2) degrés de liberté , noté $X \rightsquigarrow F(n_1, n_2)$, si $X = \frac{Y_1/n_1}{Y_2/n_2}$, $Y_i \rightsquigarrow \mathcal{X}^2(n_i), i = 1, 2$ et Y_1 et Y_2 sont indépendantes.

5. Loi de Student

Definition

On dit que X suit une loi Student à n degrés de liberté , noté $X \rightsquigarrow t(n)$, si $X = \frac{Z}{\sqrt{Y/n}}$, $Z \rightsquigarrow N(0,1)$, $Y \rightsquigarrow \mathcal{X}^2(n)$, et Z et Y sont indépendantes.

6. Loi de Fisher

Definition

On dit que X suit une loi Fisher à (n_1, n_2) degrés de liberté , noté $X \rightsquigarrow F(n_1, n_2)$, si $X = \frac{Y_1/n_1}{Y_2/n_2}$, $Y_i \rightsquigarrow \mathcal{X}^2(n_i), i = 1, 2$ et Y_1 et Y_2 sont indépendantes.

Definition

Une série double à deux indices est définie par l'observation de n couples de valeurs tels que:

- Cas discret: X prends les valeurs $x_1, ..., x_p$ et Y prends les valeurs $y_1, ..., y_r$, et à chaque couple de valeurs $(x_i, y_j)_{1 \leq i \leq p; 1 \leq j \leq r}$ on associe:
  - Son effectif $n_{i,j}$ qui est le nombre de fois où $(X, Y) = (x_i, y_j)$.
  - Sa fréquence $f_{i,j} = \frac{n_{i,j}}{n}$.

- Cas continu : Les valeurs observées pour X et pour Y sont regroupées respectivement en p et r classes $[a_0, a_1[, [a_1, a_2[, ..., [a_{p-1}, a_p]$, resp. $[b_0, b_1[, [b_1, b_2[, ..., [b_{r-1}, b_r]$, et $n_{i,j}$ est le nombre d'observations pour lesquelles $a_{i-1} \leq X < a_i$, et $b_{j-1} \leq Y < b_j$.

Exemple 2. : Sur 200 personnes la répartition du poids X (en kg) selon la taille Y en cm est décrite dans le tableau suivant:

Y(taille) / X(poids)	$160-165$	165-167	$167-169$	Total
$55-60$	145	0	0	145
$60-65$	0	25	30	55
Total	145	25	30	200

Présentation sous forme d'un tableau de contingence

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

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

Contenu

Tester l'indépendance entre deux variables X et Y.

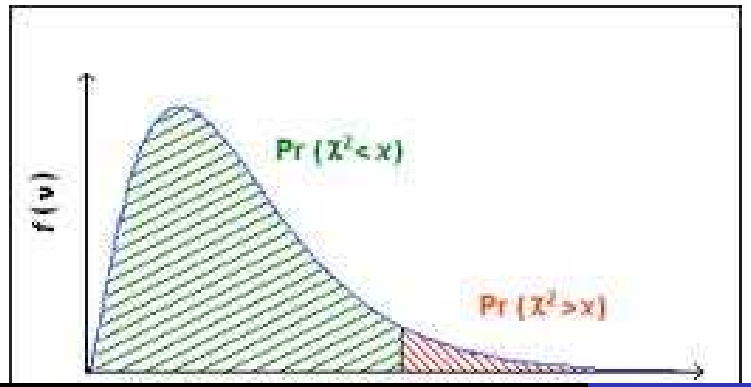
$$\begin{cases} H_0 : X \text{ est indépendante de } Y \text{ (hypothèse nulle)} \\ \text{contre} \\ H_1 : X \text{ dépends de } Y \text{ (hypothèse alternative)} \end{cases}$$

$$S = \sum_{i=1}^{p} \sum_{j=1}^{r} \frac{\left(n_{i,j} - \frac{n_{i,\bullet} n_{\bullet,j}}{n}\right)^2}{\frac{n_{i,\bullet} n_{\bullet,j}}{n}}$$

Sous H_0: $S \rightsquigarrow \mathcal{X}^2((p-1)(r-1))$, loi de Khi-deux à $(p-1)(r-1)$ degrés de libertés.

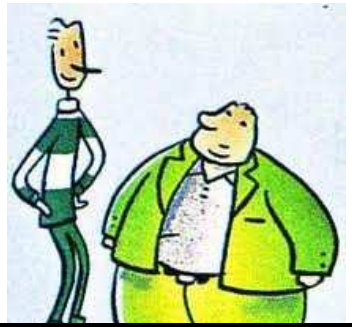
On fixe le risque d'erreur $\alpha = 5\%$.

Si $S > \mathcal{X}^2_{0.95}((p-1)(r-1))$ alors on rejette l'hypohèse d'indépendance de X et Y, où $\mathcal{X}^2_{0.95}((p-1)(r-1))$ est le fractile d'ordre 0.95 de la loi $\mathcal{X}^2((p-1)(r-1))$.



Exemple 2. (suite): On souhaite tester si le poids d'un individu dépends de sa taille:

$$\begin{cases} H_0 : \text{ le poids } X \text{ est indépendant de la taille } Y \\ \text{contre} \\ H_1 : \text{ le poids } X \text{ dépends de la taille } Y \end{cases}$$



On construit le tableau théorique de contingence sous l'hypothèse H_0 : $T_{i,j} = \frac{n_{i,\bullet} n_{\bullet,j}}{n}$

Y(taille) / X(poids)	$160 - 165$	165-167	$167 - 169$	Total
$55 - 60$	105	18	22	145
$60 - 65$	40	7	8	55
Total	145	25	30	200

S= distance entre le tableau observé et le tableau théorique.

$p = 2, r = 3, S = 202.02$ et $\mathcal{X}^2_{0.95}(2) = 5.99$.

Contenu

Conformité d'une fréquence absolue

P: population de taille n à k modalités.

E: un échantillon de taille n'.

p_k: probabilité d'observer la modalité k dans la population P.

$$\begin{cases} H_0 : \text{ la fréquence absolue de la modalité } k \text{ dans } E \\ \text{ est conforme à celle de } P \\ \text{contre} \\ H_1 : \text{ la fréquence absolue de la modalité } k \text{ dans } E \\ \text{ n'est pas conforme à celle de } P \end{cases}$$

$$Z_k = \frac{frequence.observee - frequence.theorique}{ecarttype.theorique}$$

X_k la variable aléatoire représentant la fréquence absolue de la modalité k dans l'échantillon E.

Sous H_0 la loi de X est la loi binomiale $B(n', p_k)$.

On suppose que $n' > 30, n'p_k > 5$ et $n'(1 - p_k) > 5,$ la loi de X est approximativement égale à une loi normale $N(n'p_k, n'p_k(1 - p_k))$.

Donc sous H_0, la loi de l'écart réduit

$$Z_k = \frac{X_k - n' p_k}{\sqrt{n' p_k (1 - p_k)}},$$

est approchée par une loi normale centrée réduite. La zone de rejet pour un risque d'erreur $\alpha = 5\%$ est donnée par

$\mathcal{R} =]-\infty, -1.96] \cup [1.96, +\infty[$.

Décision : Si $Z_k \in \mathcal{R}$ on rejette la conformité de la modalité k.

Exemple 3. Contrôle d'efficacté: On veut tester si le taux d'efficacité d'un médicament est de 90% au moins, valeur garantie par un fournisseur. On a observé 86 guérisons sur un échantillon de 100 malades auxquels le médicament a été administré, soit un pouvoir de guérison de 86%.

On a $n' = 100, p_1 = 0.90$, X_1 suit une loi binomiale $B(100, 0.90)$ que l'on peut approcher par une loi normale $N(90, 9)$.

Donc $Z_1 = \frac{86-90}{\sqrt{9}} = 1.33 < 1.96$, par conséquent au risque d'erreur de $\alpha = 5\%$ on ne rejette pas l'hypothèse du fournisseur.

Conformité d'une proportion

Ils'agit de comparer une proportion théorique p_0 à une proportion observée. On effectue le test suivant

$$\begin{cases} H_0 : \quad p_0 \text{ est la proportion des individus de la population de la catégorie A} \\ \text{contre} \\ H_1 : \text{ non } H_0 \end{cases}$$

ou encore

$$\begin{cases} H_0 : \quad p = p_0 \\ \text{contre} \\ H_1 : \quad p \neq p_0 \end{cases}$$

On prélève un échantillon aléatoire non exhaustive de taille n dans lequel la proportion des individus de la catégorie A est $f = \frac{n_1}{n}$. Soit $Y = X/n$ la variable aléatoire représentant la fréquence relative.

Sous H_0 : $\frac{Y - p_0}{\sqrt{v_0}} \rightsquigarrow N(0,1)$, avec $v_0 = \frac{p_0(1-p_0)}{n}$.

On fixe α (0.01 ou 0.05) et on lit dans la table de Gauss u tel que $P(N(0,1) \in [-u,u]) = 1-\alpha$.

La zone de rejet pour le risque d'erreur α est donné par

$$\mathcal{R} =]-\infty - u] \bigcup [u, +\infty[.$$

avec $v_0 = \frac{p_0(1-p_0)}{n}$.

Décision : Si $\frac{f-p_0}{\sqrt{v_0}} \in \mathcal{R}$ alors on rejette H_0.

Exemple 3. (suite) On fixe $\alpha = 5\%$. On veut tester $p_0 = 0.90$ avec $f = n_1/n = 86/100 = 0.86$.

On a $u_{1-\alpha/2} = 1.96$, $v_0 = 0.0009$, donc

$\mathcal{R} =]-\infty, 0.84] \bigcup [0.96, +\infty[$.

Décision : $f_1 \notin \mathcal{R}$ on ne rejette pas H_0.

Test bilatéral de conformité d'une moyenne avec variance connue

On attribue la valeur m_0 pour moyenne du caractère dans une population et on veut juger le bien fondé de cette hypothèse en se basant sur un échantillon $(X_1, ..., X_n)$ avec $X_i \rightsquigarrow N(m, \sigma^2)$. On effectue alors le test d'hypothèse:

$$\begin{cases} H_0 : \quad m = m_0 \\ \text{contre} \\ H_1 : \quad m \neq m_0 \end{cases}$$

où m est la moyenne théorique.

On suppose que la variance théorique σ^2 est connue et on considère

$$T = \sqrt{n}\frac{\overline{X} - m_0}{\sigma},$$

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \text{ (moyenne empirique)}$$

Sous H_0, T suit une loi normale de moyenne 0 et de variance 1.

Pour un risque d'erreur α soit le quantile u tel que $P(N(0,1) \in [-u, u[) = 1 - \alpha$.

La zone de rejet associée est donnée par

$$\mathcal{R} =]-\infty, -u] \bigcup [u, +\infty[.$$

Décision : Si $T \in \mathcal{R}$ alors on rejette H_0.

Test bilatéral de conformité d'une moyenne avec variance inconnue

On attribue la valeur m_0 pour moyenne du caractère dans une population et on veut juger le bien fondé de cette hypothèse en se basant sur un échantillon $(X_1, ..., X_n)$ avec $X_i \rightsquigarrow N(m, \sigma^2)$. On effectue alors le test d'hypothèse:

$$\begin{cases} H_0 : \quad m = m_0 \\ \text{contre} \\ H_1 : \quad m \neq m_0 \end{cases}$$

où m est la moyenne théorique.

On suppose que la variance σ^2 est inconnue, on considère alors

$$T = \sqrt{n}\frac{\overline{X} - m_0}{S'},$$

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \text{ (moyenne empirique)}$$

et

$$S'^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2 \text{ (variance empirique modifiée) .}$$

Sous H_0, T suit une loi de Student à (n-1) degrés de liberté,

Pour un risque d'erreur α soit le quantile t tel que

$P(t(n-1) \in [-t, t[) = 1 - \alpha$.

La zone de rejet associée est donnée par

$$\mathcal{R} =]-\infty, -t] \bigcup [t, +\infty[.$$

Décision : Si $T \in \mathcal{R}$ alors on rejette H_0.

Exemple 4. Le poids moyen d'une population d'adultes est supposé distribué suivant une loi de Gauss de moyenne 75kg. Sur un échantillon de 15 personnes on a trouvé les mesures suivantes : 65, 90, 74, 96, 85, 85, 80, 93, 77, 70, 84, 75, 78, 69, 83. Avec un risque d'erreur $\alpha = 0.05$, ce poids moyen observé sur l'échantillon est-il compatible avec le poids moyen de la population ?

On a $\overline{X} = 80.26667, S'^2 = 79.92381, n = 15$, donc $T = 2.28162$. T suit une loi de Student à 14 degrés de libertés, donc $P(t(14) \in [-2.14, 2.14]) = 0.95$ c'est à dire que la zone de rejet est donnée par

$$\mathcal{R} =]-\infty, -2.14] \bigcup [2.14, +\infty[.$$

Décision : $T \in \mathcal{R}$ donc on rejette H_0.

Test unilatéral de conformité d'une moyenne avec variance connue

En se basant sur un échantillon $(X_1, ..., X_n)$ avec $X_i \rightsquigarrow N(m, \sigma^2)$. On veut tester

$$\begin{cases} H_0 : & m \leq m_0 \\ \text{contre} & \\ H_1 : & m > m_0 \end{cases}$$

où m est la moyenne théorique.

On suppose que σ^2 est connue, on considère alors

$$T = \sqrt{n}\frac{\overline{X} - m_0}{\sigma},$$

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \text{ (moyenne empirique)}$$

Pour un risque d'erreur α soit le quantile u' tel que $P(N(0,1) \leq u') = 1 - \alpha$.

La zone de rejet associée est donnée par

$$\mathcal{R} = [u', +\infty[.$$

Décision : Si $T \in \mathcal{R}$ alors on rejette H_0.

Test unilatéral de conformité d'une moyenne avec variance inconnue

En se basant sur un échantillon $(X_1, ..., X_n)$ avec $X_i \rightsquigarrow N(m, \sigma^2)$. On veut tester :

$$\begin{cases} H_0 : & m \leq m_0 \\ \text{contre} & \\ H_1 : & m > m_0 \end{cases}$$

où m est la moyenne théorique.

On suppose que σ^2 est connue, on considère alors

$$T = \sqrt{n}\frac{\overline{X} - m_0}{S'},$$

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \text{ (moyenne empirique)}$$

et

$$S'^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2 \text{ (variance empirique modifiée) .}$$

Pour un risque d'erreur α soit le quantile t tel que $P(t(n-1) \leq t') = 1 - \alpha$.

La zone de rejet associée est donnée par

$$\mathcal{R} = [t, +\infty[.$$

Décision : Si $T \in \mathcal{R}$ alors on rejette H_0.

Test bilatéral de conformité d'une variance

C'est un test de conformité d'une variance observée à une variance théorique: on attribue la valeur σ_0^2 pour variance du caractère dans une population et on veut juger le bien fondé de cette hypothèse en se basant sur un échantillon $(X_1, ..., X_n)$ avec $X_i \rightsquigarrow N(m, \sigma^2)$. On effectue alors le test d'hypothèse:

$$\begin{cases} H_0 : \sigma^2 = \sigma_0^2 \\ \text{contre} \\ H_1 : \sigma^2 \neq \sigma_0^2. \end{cases}$$

Sous H_0, $\frac{(n-1)S^2}{\sigma_0^2} \rightsquigarrow \mathcal{X}^2(n-1)$.

Pour un risque d'erreur α, soient a et b tels que

$$P(\mathcal{X}^2(n-1) < a) = \alpha/2, P(\mathcal{X}^2(n-1) > b) = \alpha/2,$$

alors la région critique est donnée par

$$\mathcal{R} = [0, a] \bigcup [b, +\infty[,$$

Décision : Si $\frac{(n-1)S^2}{\sigma_0^2} \in \mathcal{R}$ alors on rejette l'hypothèse H_0.

Test unilatéral de conformité d'une variance En se basant sur un échantillon $(X_1, ..., X_n)$ avec $X_i \rightsquigarrow N(m, \sigma^2)$ on veut tester

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

Pour un risque d'erreur α, soit b tel que

$$P(\mathcal{X}^2(n-1) < b) = 1 - \alpha,$$

alors la région critique est donnée par

$$\mathcal{R} = [b, +\infty[,$$

Décision : Si $\frac{(n-1)S^2}{\sigma_0^2} \in \mathcal{R}$ alors on rejette l'hypothèse H_0.

Contenu

Comparaison de deux proportions:

Il s'agit de comparer deux proportions observées. Soient deux échantillons aléatoire E_1 et E_2 extraits des deux populations P_1 et P_2 constituées des individus des deux seules catégories A et B. Soit $f_1 = \frac{k_1}{n_1}$ et $f_2 = \frac{k_2}{n_2}$ les proportions des individus de la catégorie A dans E_1 et E_2 respectivement. En général $f_1 \neq f_2$; il s'agit d'expliquer cette différence.

On effectue le test d'hypothèse suivant:

$$\begin{cases} H_0 : \text{Dans } P_1 \text{ et } P_2 \text{ il y a la même proportion } p_0 \\ \quad \text{(inconnue) d'individus de la catégorie A} \\ \text{contre} \\ H_1 : \text{ non } H_0 \end{cases}$$

On considère $D = Y_1 - Y_2$, avec f_i est l'observation de la variable aléatoire Y_i. Sous H_0 on a $D \rightsquigarrow N(0, \sigma_D^2)$ où

$$\sigma_D^2 = p_0(1-p_0)\left(\frac{1}{n_1} + \frac{1}{n_2}\right). \qquad (4)$$

Comme p_0 est inconnue, on peut l'estimer sous H_0 par

$$\frac{k_1 + k_2}{n_1 + n_2} = \frac{n_1 f_1 + n_2 f_2}{n_1 + n_2},$$

ainsi σ_D^2 est estimée par

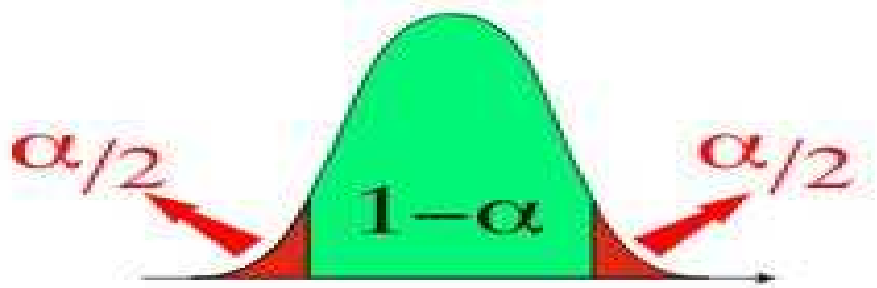
$$\begin{aligned} \widehat{\sigma}_D^2 &= \frac{n_1 f_1 + n_2 f_2}{n_1 + n_2}\left(1 - \frac{n_1 f_1 + n_2 f_2}{n_1 + n_2}\right)\left(\frac{1}{n_1} + \frac{1}{n_2}\right) \\ &= \frac{(n_1 f_1 + n_2 f_2)\,(n_1(1 - f_1) + n_2(1 - f_2))}{(n_1 + n_2) n_1 n_2}. \end{aligned}$$

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

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

donc la zone de rejet est $\mathcal{R} =]-\infty, -u] \bigcup [u, +\infty[$.

Décision: Si $\frac{f_1 - f_2}{\widehat{\sigma}_D} \in \mathcal{R}$ alors on rejette H_0.



Comparaison des moyennes de deux échantillons indépendants

Soient deux échantillons aléatoires d'observations provenant de deux populations P_1 et P_2 et de moyenne $\overline{X}_1$ et $\overline{X}_2$ respectivement. En général $\overline{X} \neq \overline{X}_2$, il s'agit d'expliquer cette différence.

On suppose que $X_1 \rightsquigarrow N(m_1, \sigma_1^2), X_2 \rightsquigarrow N(m_2, \sigma_2^2)$, X_1 et X_2 sont deux échantillons indépendants, on désire effectuer le test

$$\begin{cases} H_0 : \quad m_1 = m_2 \\ \text{contre} \\ H_1 : \quad m_1 \neq m_2. \end{cases}$$

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

- Cas σ_1^2 et σ_2^2 connues:

Sous H_0 : $\dfrac{D}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}} \rightsquigarrow N(0,1)$.

On fixe α et on lit dans la table de Gauss u tel que

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

donc la zone de rejet est $\mathcal{R} =]-\infty, -u] \bigcup [u, +\infty[$.

Décision: Si $\dfrac{D}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}} \in \mathcal{R}$ alors on rejette H_0.

- Cas σ_1^2 et σ_2^2 inconnues et les échantillons sont grands

On suppose que $n_1 \geq 30$ et $n_2 \geq 30$: On peut remplacer σ_1^2 (resp . σ_2^2) par les variances empiriques modifiées $S_1'^2 = \frac{1}{n_1-1}\sum_1^{n_1}(X_{1,i} - \overline{X}_1)^2$ (resp. $S_2'^2$).

La zone de rejet est $\mathcal{R} =]-\infty, -u] \bigcup [u, +\infty[$, avec $P(N(0,1) \in [-u, u]) = 1 - \alpha$.

Décision: Si $\frac{D}{\sqrt{\frac{s_1'^2}{n_1} + \frac{s_2'^2}{n_2}}} \in \mathcal{R}$ alors on rejette H_0.

- Cas σ_1^2 et σ_2^2 inconnues et les échantillons sont petits

On suppose que $\sigma_1^2 = \sigma_2^2 = \sigma^2$. On montre alors que sous H_0 : $\frac{D}{\sqrt{f(1,2)}} \rightsquigarrow t(n_1 + n_2 - 2)$, où

$$f(1,2) = \frac{(n_1 - 1)S_1'^2 + (n_2 - 1)S_2'^2}{n_1 + n_2 - 2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right). \quad (5)$$

La zone de rejet est $\mathcal{R} =]-\infty, -t] \bigcup [t, +\infty[$, avec $P(t(n_1 + n_2 - 2) \in [-t, t]) = 1 - \alpha$.

Décision: Si $\frac{D}{\sqrt{f(1,2)}} \in \mathcal{R}$ alors on rejette H_0.

Test de Mann-Whitney Wilcoxon

On désire effectuer le test

$$\begin{cases} H_0 : \quad m_1 = m_2 \\ \text{contre} \\ H_1 : \quad m_1 \neq m_2. \end{cases}$$

On considère

$$U_{1,2} = \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \mathbf{1}_{\{X_{1,i} \leq X_{2,j}\}}.$$

On compte pour chaque $X_{1,i}, i = 1, ..., n_1$ le nombre de $X_{2,j}, j = 1, .., n_2$ qui lui sont supérieurs et on somme les résultats obtenus pour tous les $X_{1,i}$.

On a sous H_0

$$T = \frac{U_{1,2} - \frac{n_1 n_2}{2}}{\sqrt{n_1 n_2 (n_1 + n_2 + 1)/12}} \rightsquigarrow N(0,1).$$

La zone de rejet est $\mathcal{R} =]-\infty, -u] \bigcup [u, +\infty[$, avec $P(N(0,1) \in [-u, u]) = 1 - \alpha$.

Décision: Si $T \in \mathcal{R}$ alors on rejette H_0.

Exemple. Deux classes $C1$ et $C2$ sont constituées respectivements de 3 et 4 étudiants. Les notes en Statistique de la classe C1 sont : 12, 15, 13. Celles de la classe C2 sont : 11, 10, 14, 16.

Peut-on dire au risque d'erreur de $\alpha = 0.05$ que les deux classes ont la même moyenne ?

On trie les notes de $C1$: 12, 13, 15, et de $C2$: 10, 11, 14,16 .

$X_{1,i}, i = 1, ..., 3$ représentent les notes de la classe $C1$.

$X_{2,i}, i = 1, ..., 4$ représentent les notes de la classe $C2$.

On forme le tableau

Table: Nombre total de $X_{2,j}$ supérieurs à $X_{1,i}$.

$X_{1,i}$ \ $X_{2,i}$	10	11	14	16
12	0	0	1	1
13	0	0	1	1
15	0	0	0	1

$U_{1,2}$ est égal au nombre totat de 1 figurant dans le tableau, donc $U_{1,2} = 5$

$n_1 = 3, n_2 = 4$, donc

$$T = \frac{5 - \frac{3*4}{2}}{\sqrt{3*4*(3+4+1)/12}} = 0.35,$$

$\mathcal{R} =]-\infty, -1.96] \bigcup [1.96, +\infty[$, donc on ne rejette pas H_0.

Comparaison des moyennes de deux échantillons appariées

Soient deux échantillons de n observations $X_{1,1}, ..., X_{1,n}$ de moyenne $\overline{X}_1$ et $X_{2,1}, ..., X_{2,n}$ de moyenne $\overline{X}_2$.

Ici l'observation $X_{1,i}$ est appariée à $X_{2,i}$.

On pose

$D_i = X_{1,i} - X_{2,i}, S'^2_D = \frac{1}{n-1}\sum_1^n (D_i - \overline{D})^2, \overline{D} = \frac{1}{n}\sum_{i=1}^n D_i.$

Pour tester

$$\begin{cases} H_0 : \quad m_1 = m_2 \\ \text{contre} \\ H_1 : \quad m_1 \neq m_1, \end{cases}$$

on considère $T = \sqrt{n}\frac{\overline{D}}{S_D}$.

- Si $n \geq 30$, alors T est assimilée à une variable centrée réduite de Gauss.
- Si $n < 30$, alors T suit une loi de Student à $(n-1)$ degrés de liberté.

Comparaison des variances de deux échantillons indépendants

On veut tester

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

On calcule les rapports $\frac{S_1'^2}{S_2'^2}$ et $\frac{S_2'^2}{S_1'^2}$ et on considère le plus grand parmi les deux.

- Si $\frac{S_1'^2}{S_2'^2} > \frac{S_2'^2}{S_1'^2}$, on décidera à l'aide du rapport $\frac{S_1'^2}{S_2'^2}$.

On a $\frac{S_1'^2}{S_2'^2} \rightsquigarrow F(n_1-1, n_2-1)$, loi de Fisher-Snédécor à (n_1-1, n_2-1) degrés de liberté, donc au risque d'erreur α la zone de rejet est donnée par

$$\mathcal{R} = [f_0, \infty[, \text{ avec } P(F(n_1-1, n_2-1) > f_0) = \alpha/2. \qquad (6)$$

Décision: Si $\frac{S_1'^2}{S_2'^2} \in \mathcal{R}$ alors on rejette H_0.

- Si $\frac{S_1'^2}{S_2'^2} \leq \frac{S_2'^2}{S_1'^2}$, on décidera à l'aide du rapport $\frac{S_2'^2}{S_1'^2}$.

On a $\frac{S_2'^2}{S_1'^2} \rightsquigarrow F(n_2-1, n_1-1)$, loi de Fisher-Snédécor à (n_2-1, n_1-1) degrés de liberté, donc au risque d'erreur α la zone de rejet est donnée par

$$\mathcal{R} = [f_0, \infty[, \text{ avec } P(F(n_2-1, n_1-1) > f_0) = \alpha/2. \qquad (7)$$

Décision: Si $\frac{S_2'^2}{S_1'^2} \in \mathcal{R}$ alors on rejette H_0.

Le but est de comparer plusieurs séries d'observations d'une variable X afin de décider s'elles proviennent de populations dans lesquelles X a la même moyenne ou la même variance.

- On considère k échantillons Gaussiens

$E_i = \{X_{i,1}, ..., X_{i,n_i}\}, 1 \leq i \leq k$ indépendants.

- On désigne par $m_1, ..., m_k$ les moyennes de X dans les populations $P_1, ..., P_k$ respectivement d'où sont extraits les échantillons $E_1, ..., E_k$.

- On suppose que X a la même variance σ_0^2 dans les populations $P_1, ..., P_k$.

On définit la variance intraclasse, par

$$S^2_{\text{intra}} = \frac{1}{n-k}\sum_{i=1}^{k}\nu_i S_i'^2, \nu_i = n_i - 1, \quad (8)$$

$$S_i'^2 = \frac{1}{n_i - 1}\sum_{j=1}^{n_i}(X_{i,j} - \overline{X_i})^2, \overline{X}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} X_{i,j}.$$

On définit la variance interclasse, par

$$S^2_{\text{inter}} = \frac{1}{k-1}\sum_{i=1}^{k} n_i(\overline{X}_i - \overline{X})^2 \ , \tag{9}$$

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i} X_{i,j}, \quad n = \sum_{i=1}^{k} n_i.$$

Comparaison simultanée de plusieurs moyennes. On veut tester

$$\begin{cases} H_0 : \quad m_1 = m_2 = \ldots = m_k \\ \text{contre} \\ H_1 : \quad \exists i \neq j \ \text{ tels que } m_i \neq m_j. \end{cases}$$

On a sous H_0 : $\frac{S^2_{\text{inter}}}{S^2_{\text{intra}}} \rightsquigarrow F(k-1, n-k)$, loi de Fisher-Snédécor à $(k-1, n-k)$ degrés de liberté, donc la zone de rejet au risque d'erreur α est donnée par

$$\mathcal{R} = [f_0, \infty[, \ \text{avec } P(F(k-1, n-k) > f_0) = \alpha. \qquad (10)$$

Décision : Si $\frac{S^2_{\text{inter}}}{S^2_{\text{intra}}} \in \mathcal{R}$ alors on rejette H_0.

Comparaison simultanée de plusieurs variances. On veut tester

$$\begin{cases} H_0 : \quad \sigma_1^2 = \sigma_2^2 = ... = \sigma_k^2 \\ \text{contre} \\ H_1 : \quad \exists i \neq j \ \text{ tels que } \sigma_i^2 \neq \sigma_j^2. \end{cases}$$

Test de Bartlett

On considère la statistique

$$B = \frac{1}{\lambda}\left\{(n-k)\ln S^2_{\text{intra}} - \sum_{j=1}^{k} \nu_i \ln S_i'^2\right\}, \tag{11}$$

$$\lambda = 1 + \frac{1}{3(k-1)}\left\{\frac{1}{\nu_1} + \frac{1}{\nu_2} .... + \frac{1}{\nu_k} - \frac{1}{n-k}\right\}, \qquad \nu_k = n_k - 1.$$

Sous $H_0, B \rightsquigarrow \mathcal{X}^2(k-1)$.

Soit $\mathbf{x}$ tel que $P(\mathcal{X}^2(k-1) \geq \mathbf{x}) = \alpha$, la zone de rejet est donnée par

$$\mathcal{R} = [\mathbf{x}, +\infty[,$$

Décision : Si $B \in \mathcal{R}$ alors on rejette H_0.