Statistique: Théorie :

Guillaume Delacollette

Table des matières

1 Lien entre l'ogive et l'aire sous l'histogramme :

1.1 Propriété :

Soit X une variable continue dont les valeurs sont groupées en J classes et dont la distribution des fréquences est décrite par un histogramme de surface unitaire ainsi que par l'ogive des fréquences cumulées y = F (x). Pour toute valeur x*, la surface délimitée par l'histogramme et l'axe des abscisses et située à gauche de x* est égale à l'ordonnée F (x*).

1.2 Preuve :

Soient les classes $[e_0, e_1],]e_1, e_2], \ldots,]e_{J-1}, e_J]$. Notons A(x*) la surface à gauche de x*,

- si $x^* < e_0, A(x^*) = 0$ et $F(x^*) = 0$.
- si $x^* \geq e_J, A(x^*) = 1$ et $F(x^*) = 1$.
- si $x^* \in]e_{j-1}, e_j]$,

$$A(x^*) = \sum_{k=1}^{j-1} a_k \frac{f_k}{a_k} + (x^* - e_{j-1}) \frac{f_j}{a_j} = F_{j-1} + \frac{f_j}{a_j}(x^* - e_{j-1}).$$

La fonction F(x) est une droite entre les points (e_{j-1}, F_{j-1}) et (e_j, F_j) dont l'équation s'écrit

$$y - F_{j-1} = \frac{F_j - F_{j-1}}{e_j - e_{j-1}}(x - e_{j-1}).$$

Le point $(x^*, F(x^*))$ appartenant à cette droite, on obtient

$$F(x^*) = F_{j-1} + \frac{f_j}{a_j}(x^* - e_{j-1}).$$

La conclusion F (x*) = A(x*) est immédiate.

2 Effet d'un changement d'origine et d'échelle sur la moyenne :

2.1 Propriété :

Si un changement d'échelle et d'origine est effectué sur les observations $x_1, \ldots, x_n$ pour obtenir la nouvelle série $S' = \{x'_1, \ldots, x'_n\}$ avec $x'_i = ax_i + b$ où a et b sont des constantes réelles, alors la moyenne arithmétique $\overline{x}$' de la série S' est donnée par

$$\overline{x}' = a\overline{x} + b,$$

où $\overline{x}$' est la moyenne arithmétique de $S = \{x_1, \ldots, x_n\}$.

2.2 Preuve :

Hypothèse :

- $x_1, \ldots, x_n$ n observations de moyenne $\overline{x}$
- $x'_i = ax_i + b$ où a et b sont des constantes réelles, a $\neq 0$. $x'_1, \ldots, x'_n$ a pour moyenne $\overline{x}$'.

Thèse :

$$\overline{x}' = a\overline{x} + b$$

Démonstration : Par défintion,

$$\begin{aligned}
\overline{x}' &= \frac{1}{n}\sum_{i=1}^{n} x'_i \\
&= \frac{1}{n}\sum_{i=1}^{n}(ax_i + b) \\
&= \frac{1}{n}(\sum_{i=1}^{n} ax_i + \sum_{i=1}^{n} b) \\
&= \frac{1}{n}(a\sum_{i=1}^{n} x_i + nb) \\
&= a\frac{1}{n}\sum_{i=1}^{n} x_i + b \\
&= a\overline{x} + b
\end{aligned}$$

3 Effet d'un changement d'origine et d'échelle sur la médiane :

3.1 Propriété :

Si un changement d'échelle et d'origine est effectué sur les observations $x_1, \ldots, x_n$ pour obtenir la nouvelle série $S' = \{x'_1, \ldots, x'_n\}$ avec $x'_i = ax_i + b$ où a et b sont des constantes réelles, alors la médiane $\widetilde{x}$' de la série S' est donnée par

$$\widetilde{x}' = a\widetilde{x} + b,$$

où $\widetilde{x}$' est la médiane de $S = \{x_1, \ldots, x_n\}$.

3.2 Preuve :

La transformation va soit conserver soit inverser l'ordre des observations de départ. La médiane étant basée sur les observations centrales sera définie à partir des versions transformées de ces observations centrales.

Hypothèse :

- $x_1, \ldots, x_n$ n observations de médiane $\widetilde{x}$
- $x'_i = ax_i + b$ où a et b sont des constantes réelles, a $\neq 0$. $x'_1, \ldots, x'_n$ a pour médiane $\widetilde{x}$'.

Thèse :

$$\widetilde{x}' = a\widetilde{x} + b$$

Démonstration :
A partir de la série $\{x_1, \ldots, x_n\}$, on a

$$x_{(1)} \leq \ldots \leq x_{(j)} \leq x_{(j+1)} \leq \ldots \leq x_{(n)}$$

D'où, pour tout a > 0 et tout b

$$x_{(1)} \leq \ldots \leq x_{(j)} \leq x_{(j+1)} \leq \ldots \leq x_{(n)}$$

$$ax_{(1)} \leq \ldots \leq ax_{(j)} \leq ax_{(j+1)} \leq \ldots \leq ax_{(n)}$$

$$ax_{(1)} + b \leq \ldots \leq ax_{(j)} + b \leq ax_{(j+1)} + b \leq \ldots \leq ax_{(n)} + b$$

On a donc x'$_{(j)}$ = ax$_{(j)}$ + b pour tout j.
En particulier :
Si n = 2k + 1, l'observation centrale, x'$_{(k+1)}$, vérifie

$$\widetilde{x} = x'_{(k+1)} = ax_{(k+1)} + b \Rightarrow \widetilde{x}' = a\widetilde{x} + b$$

Et si n = 2k, les deux observations du milieu, x'$_{(k)}$ et x'$_{(k+1)}$, sont telles que

$$\widetilde{x}' = \frac{x'_{(k)} + x'_{(k+1)}}{2} = \frac{ax_{(k)} + b + ax_{(k+1)} + b}{2} = a\frac{x_{(k)} + x_{(k+1)}}{2} + b = a\widetilde{x} + b$$

Et si a < 0
A partir de

$$x_{(1)} \leq \ldots \leq x_{(j)} \leq x_{(j+1)} \leq \ldots \leq x_{(n)}$$

Il vient

$$ax_{(1)} \geq \ldots \geq ax_{(j)} \geq ax_{(j+1)} \geq \ldots \geq ax_{(n)}$$

Ou encore

$$ax_{(1)} + b \geq \ldots \geq ax_{(j)} + b \geq ax_{(j+1)} + b \geq \ldots \geq ax_{(n)} + b$$

L'ordre est inversé : $x'_{(j)} = ax_{(n-j+1)} + b$ mais le milieu reste au milieu :
Si n = 2k + 1

$$ax_{(1)} + b \geq \ldots \geq \underbrace{ax_{(k+1)} + b} \geq \ldots \geq ax_{(n)} + b$$

$$= x'_{(k+1)}$$

Si n = 2k

$$ax_{(1)} + b \geq \ldots \geq \underbrace{ax_{(k)} + b} \geq \underbrace{ax_{(k+1)} + b} \geq \ldots \geq ax_{(n)} + b$$

$$= x'_{(k)} = x'_{(k+1)}$$

4 Effet d'un changement d'origine et d'échelle sur la variance et l'écart-type :

4.1 Propriété :

Si on effectue un changement d'échelle sur les observations initiales $x-1, \ldots, x_n$ pour obtenir la série S' = $\{x_1, \ldots, x'_n\}$, avec $x'_i = ax_i + b$, a, b $\in \mathbb{R}$, alors la variance s'^{2} de S' est donnée par s'^{2} = a^{2}s^2 où s^2 est la variance de S. Pour les écarts-types, la relation devient $s' = |a|\, s$. Les changements d'origine n'ont donc aucun effet sur la variance et l'écart-type.

4.2 Preuve :

Une propriété similaire a établi que la moyenne de la nouvelle série suit la même transformation que les données : $\overline{x}' = a\overline{x} + b$. Dès lors, par définition de la variance, on a

$$s'^2 = \frac{1}{n}\sum_{i=1}^{n}(x'_i - \overline{x}')^2 = \frac{1}{n}\sum_{i=1}^{n}(ax_i - b - a\overline{x} + b)^2 = a^2 s^2$$

5 Définition et calcul de la moyenne de la série des valeurs centrées :

5.1 Propriété :

Définissons la série des valeurs centrées S$_c$ en prenant les différences entre les valeurs observées de la série $S = \{x_1, \ldots, x_n\}$et la moyenne arithmétique de la série S. Cette nouvelle série $S_c = \{x_1 - \overline{x}, \ldots, x_n - \overline{x}\}$ a une moyenne nulle.

5.2 Preuve :

Il s'agit d'un cas particulier de la propriété 1.

La propriété 1 nous dit ceci, Si un changement d'échelle et d'origine est effectué sur les observations $x_1, \ldots, x_n$ pour obtenir la nouvelle série S'= $\{x'_1, \ldots, x'_n\}$ avec x$_i$' = ax$_i$+b où a et b sont des constantes réelles, alors la moyenne arithmétique $\overline{x}$' de la série S' est donnée par

$$\overline{x}' = a\overline{x} + b,$$

où $\overline{x}$ est la moyenne arithmétique de S = $\{x_1, \ldots, x_n\}$.

En prenant a = 1 et b = -$\overline{x}$ dans la propriété 1, on obtient directement $\overline{x}_c = \overline{x} - \overline{x}$.

6 Propriété d'optimalité de la moyenne :

6.1 Propriété :

La somme des carrés des écarts des éléments d'une série par rapport à la moyenne arithmétique de la série est inférieure ou égale à la somme des carrés des écarts par rapport à toute autre valeur a $\in \mathbb{R}$. Cette propriété s'écrit aussi

$$\sum_{i=1}^{n}(x_i - \overline{x})^2 \leq \sum_{i=1}^{n}(x_i - a)^2 \quad \forall a \in \mathbb{R}.$$

6.2 Preuve :

Il suffit de rechercher le minimum sur $\mathbb{R}$ de la fonction $f(a) = \sum_{i=1}^{n}(x_i - a)^2$.

Hypothèse :

- $x_1, \ldots, x_n$ de moyenne $\overline{x}$

Thèse :

$$\sum_{i=1}^{n}(x_i - \overline{x})^2 \leq \sum_{i=1}^{n}(x_i - a)^2 \forall a \in \mathbb{R}$$

où $\overline{x}$ est la valeur qui minimise la fonction $f(a) = \sum_{i=1}^{n}(x_i - a)^2$.

Démonstration :

Recherchons le minimum de f(a)

Point Stationaire :

$$\begin{aligned} f'(a) &= \sum_{i=1}^{n} 2(x_i - a)(-1) = 0 \\ &\Leftrightarrow \cancel{2}\sum_{i=1}^{n}(x_i - a) = 0 \\ &\Leftrightarrow \sum_{i=1}^{n} x_i - na = 0 \\ &\Leftrightarrow a = \overline{x} \end{aligned}$$

et $f''(a) = 0 \Leftrightarrow \overline{x}$ est le minimum.

7 Propriété d'optimalité de la médiane :

7.1 Propriété :

La somme des écarts absolus des éléments d'une série par rapport à la médiane de la série est inférieure ou égale à la somme des écarts absolus par rapport à toute autre valeur $a \in \mathbb{R}$. Cette propriété s'écrit aussi

$$\sum_{i=1}^{n} |x_i - \widetilde{x}| \leq \sum_{i=1}^{n} |x_i - a| \, \forall a \in \mathbb{R}$$

8 Théorème de König-Huygens :

8.1 Propriété :

La moyenne des carrés des écarts entre les observations d'une série et un paramètre a se décompose de la façon suivante :

$$s^2(a) = \frac{1}{n}\sum_{i=1}^{n}(x_i - a)^2 = s^2 + (\overline{x} - a)^2$$

où s^2 est la variance de la série.

8.2 Preuve :

Hypothèse :
$x_1, \ldots, x_n$ de moyenne $\overline{x}$ et de variance s^2.

Thèse :

$$\frac{1}{n}\sum_{i=1}^{n}(x_i - a)^2 = s^2 + (\overline{x} - a)^2, \forall a \in \mathbb{R}$$

Démonstration :

$$\begin{aligned}
s^2(a) &= \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x} + \overline{x} - a)^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}\Big((x_i - \overline{x}) + (\overline{x} - a)\Big)^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}\Big((x_i - \overline{x})^2 + (\overline{x} - a)^2 + 2(x_i - \overline{x})(\overline{x} - a)\Big) \\
&= \underbrace{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2}_{} + \underbrace{\frac{1}{n}\sum_{i=1}^{n}(\overline{x} - a)^2}_{} + \underbrace{\frac{1}{n}\sum_{i=1}^{n}2(x_i - \overline{x})(\overline{x} - a)}_{} \\
&= \text{variance} \qquad \text{indépendant} \qquad 2(\overline{x} - a)\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x}) \\
&= s^2 \qquad (\overline{x} - a)^2 \qquad \text{moyenne de la série centré} \\
&= s^2 \qquad (\overline{x} - a)^2 \qquad \text{qui est égal à } 0 \\
&= s^2 + (\overline{x} - a)^2
\end{aligned}$$

9 Formule de décomposition de la variance pour une série décomposée en k sous-séries :

9.1 Propriété :

Soit une population P de n individus décomposée en k sous-populations $P_1, \ldots, P_k$ d'effectifs $n_1, \ldots, n_k$ (avec $\sum_{i=1}^{k} n_i = n$), de moyennes $\overline{x}_1, \ldots, \overline{x}_k$ et de variances $s_1^2, \ldots, s_k^2$. La variance globale s^2 de la population est donnée par

$$s^2 = \frac{\sum_{i=1}^{k} n_i s_i^2}{n} + \frac{\sum_{i=1}^{k} n_i (\overline{x}_i - \overline{x})^2}{n},$$

où $\overline{x}$ est la moyenne globale de P.

9.2 Preuve :

Hypothèse :

- P de n individu
- moyenne $\overline{x}$ et variance s^2
- k sous population et effectifs $n_1, \ldots, n_k$ avec $n_1 + \ldots + n_k = n$
- moyenne "local" $\overline{x}_1, \ldots, \overline{x}_k$
- variance "local" $s_1^2, \ldots, s_k^2$

Thèse :

$$s^2 = \frac{1}{n} \sum_{i=1}^{k} n_i s_i^2 + \frac{1}{n} \sum_{i=1}^{k} n_i (\overline{x}_i - \overline{x})^2$$

Démonstration :
Par définition, $s^2 = \frac{1}{n} \sum_{i=1}^{k} (x_i - \overline{x})^2$
Soit : $I = \{1, \ldots, n\}$
Soit : $I_j = \{i \in I, x_i \in I_j\}$, on a $I_j \cap I_k = \emptyset$ et $\cup_j I_j = I$

$$\rightarrow s^2 = \frac{1}{n} \sum_{j=1}^{k} \sum_{i \in I_j} (x_i - \overline{x})^2$$

Plaçons nous dans P_j ($\rightarrow$ jème population) et n_j individus, moyenne $\overline{x}_j$, variance s_j^2.
Application de TKH :

$$s_j^2(a) = \frac{1}{n_j} \sum_{i \in I_j} (x_i - a)^2 = s_j^2 + (\overline{x}_j - a)^2$$

vrai $\forall a \in \mathbb{R}$ (en particulier pour $a = \overline{x}$).

$$
\begin{aligned}
\rightarrow \frac{1}{n_j}\sum_{i\in I_j}(x_i-\overline{x})^2 &= s_j^2+(\overline{x}_j-\overline{x})^2\\
\sum_{i\in I_j}(x_i-\overline{x})^2 &= n_j s_j^2+n_j(\overline{x}_j-\overline{x})^2\\
\Rightarrow s^2 &= \frac{1}{n}\sum_{j=1}^{k}(n_j s_j^2+n_j(\overline{x}_j-\overline{x})^2)\\
\Rightarrow s^2 &= \frac{1}{n}\sum_{j=1}^{k}n_j s_j^2+\frac{1}{n}\sum_{j=1}^{k}n_j(\overline{x}_j-\overline{x})^2
\end{aligned}
$$

10 Démonstration de la formule équivalente $s^2=\frac{1}{n}\sum_{i=1}^{n}x_i^2-\overline{x}^2$:

10.1 Preuve :

Le théorème de König-Huygens nous donne la relation suivante : La moyenne des carrés des écarts entre les observations d'une série et un paramètre a se décompose de la façon suivante :

$$s^2(a)=\frac{1}{n}\sum_{i=1}^{n}(x_i-a)^2=s^2+(\overline{x}-a)^2$$

où s^2 est la variance de la série.

En prenant a = 0 dans la relation précédente, il vient

$$\frac{1}{n}\sum_{i=1}^{n}x_i^2=s^2+\overline{x}^2\Rightarrow s^2=\frac{1}{n}\sum_{i=1}^{n}x_i^2-\overline{x}^2$$

11 Propriété de Tchebychev :

11.1 Propriété :

Soit $S=\{x_1,\ldots,x_n\}$ une série de moyenne $\overline{x}$ et d'écart-type s. La proportion d'observations s'écartant d'au moins t écarts-types de la moyenne est inférieure ou égale à $\frac{1}{t^2}$.

11.2 Preuve :

Hypothèse :

$x_1,\ldots,x_n$ de moyenne $\overline{x}$ et d'écart-type s et $t>0$.

Thèse :

$$\frac{k}{n}\leq\frac{1}{t^2}$$

Démonstration :

Toute observation x_i est soit tel que $x_i \in]\overline{x} - ts, \overline{x} + ts[$, soit tel que $x_i \notin]\overline{x} - ts, \overline{x} + ts[$

x_i est dans l'intervalle si et seulement si $|x_i - \overline{x}| < ts$

x_i est en dehors de l'intervalle si $|x_i - \overline{x}| \geq ts$

Notons I comme l'ensemble de tous les indices :

$I = \{1, 2, \ldots, n\}$

$$\begin{aligned} I_1 &= \{i \in I : |x_i - \overline{x}| < ts\} \\ I_2 &= \{i \in I : |x_i - \overline{x}| \geq ts\} \\ I_1 &\cap I_2 = \emptyset \\ I_1 &\cup I_2 = I \end{aligned}$$

Par définition,

$$\begin{aligned} s^2 &= \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 \\ &= \frac{1}{n}\sum_{i\in I_1}(x_i - \overline{x})^2 + \frac{1}{n}\sum_{i\in I_2}(x_i - \overline{x})^2 \end{aligned}$$

C'est I_2 qui nous intéresse

$$\frac{1}{n}\sum_{i\in I_2}\underbrace{(x_i - \overline{x})^2}$$

$$\geq t^2 s^2$$

$$\frac{1}{n}\sum_{i=I_2}(x - \overline{x})^2 \geq \frac{1}{n}\underbrace{\sum_{i\in I_2} t^2 s^2}$$

$$= \frac{kt^2 s^2}{n}$$

k est le nombre d'indices appartenant à I_2 ou encore le nombre d'observations s'écartant d'au moins t écarts-types de la moyenne.

En réarrangeant les termes, on obtient,

$$\begin{aligned} \cancel{s^2} &\geq \frac{kt^2\cancel{s^2}}{n} \\ &\Rightarrow \frac{k}{n} < \frac{1}{t^2} \end{aligned}$$

CQFD

12 Effet d'un changement d'origine et d'échelle sur la covariance et la corrélation :

12.1 Hypothèse :

(x_i, y_i) $1 \leq i \leq n$ de covariance s_{xy}
x_i' = ax_i+b $\overline{x}$' = a$\overline{x}$+b
y_i' = cy_i+d $\overline{y}$' = c$\overline{y}$+d

12.2 Thèse :

$s'_{xy} = acs_{xy}$

12.3 Démonstration :

Par définition,

$$\begin{aligned} s'_{xy} &= \frac{1}{n}\sum_{i=1}^{n}(x'_i - \overline{x}')(y'_i - \overline{y}') \\ &= \frac{1}{n}\sum_{i=1}^{n}(ax_i + \cancel{b} - a\overline{x} - \cancel{b})(cy_i + \cancel{d} - c\overline{y} - \cancel{d}) \\ &= \frac{1}{n}\sum_{i=1}^{n}a(x_i - \overline{x})c(y_i - \overline{y}) \\ &= acs_{xy} \end{aligned}$$

13 Formule équivalente $s_{xy} = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \overline{x}\,\overline{y}$:

13.1 Hypothèse :

(x_i, y_i) $1 \leq i \leq n$
moyenne $\overline{x}$ moyenne $\overline{y}$
covariance s_{xy}

13.2 Thèse :

$s_{xy} = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \overline{x}\,\overline{y}$

13.3 Démonstration :

Par définition,

$$\begin{aligned} s_{xy} &= \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) \\ &= \frac{1}{n}\sum_{i=1}^{n}x_i y_i - \frac{1}{n}\sum_{i=1}^{n}x_i\overline{y} - \frac{1}{n}\sum_{i=1}^{n}\overline{x}y_i + \frac{1}{n}\sum_{i=1}^{n}\overline{x}\,\overline{y} \\ &= \frac{1}{n}\sum_{i=1}^{n}x_i y_i - \overline{x}\,\overline{y} - \cancel{\overline{x}\,\overline{y}} + \cancel{\overline{x}\,\overline{y}} \end{aligned}$$

14 Démonstration de la forme des moyennes et variance de la série des résidus :

14.1 Hypothèse :

(x_i, y_i) $1 \leq i \leq n$
droite de régression par les moindres carrés
$y = \hat{a}x + \hat{b}$ avec $\hat{a} = \frac{s_{xy}}{s_x^2}$ et $\hat{b} = \overline{y} - \hat{a}x$
$\delta_i = y_i - \hat{y}_i$

14.2 Thèse :

La moyenne de la série des résidus est nulle : $\overline{\delta} = 0$
La variance (appelée variance résiduelle de y par rapport à x), vaut $s_\delta^2 = s_y^2(1 - r^2)$ où $r = r_{xy}$

14.3 Démonstration :

14.3.1 La moyenne :

$$\begin{aligned}
\overline{\delta} &= \frac{1}{n}\sum_{i=1}^{n}\delta_i \\
&= \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i) \\
&= \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{a}x_i - \hat{b}) \\
&= \frac{1}{n}\sum_{i=1}^{n}y_i - \frac{1}{n}\sum_{i=1}^{n}\hat{a}x_i - \frac{1}{n}\sum_{i=1}^{n}\hat{b} \\
&= \overline{y} - \hat{a}\overline{x} - \hat{b} \\
&\rightarrow \text{ par hypothèse } \hat{b} = \overline{y} - \hat{a}\overline{x} \\
&= \overline{y} - \hat{a}\overline{x} - \overline{y} + \hat{a}\overline{x} \\
&= 0
\end{aligned}$$

14.3.2 La variance :

$$
\begin{aligned}
s_\delta^2 &= \frac{1}{n}\sum_{i=1}^{n}(\delta_i - \overline{\delta})^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}\delta_i^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{a}x_i - \hat{b})^2 \\
&\rightarrow \text{ par hypothèse } \hat{b} = \overline{y} - \hat{a}\overline{x} \\
&= \frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y} - \hat{a}(x_i - \overline{x}))^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}\{(y_i - \overline{y})^2 - 2(y_i - \overline{y})\hat{a}(x_i - \overline{x}) + \hat{a}^2(x_i - \overline{x})^2\} \\
&= \frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y})^2 - \frac{1}{n}\sum_{i=1}^{n}2(y_i - \overline{y})\hat{a}(x_i - \overline{x}) + \frac{1}{n}\sum_{i=1}^{n}\hat{a}^2(x_i - \overline{x})^2 \\
&= s_y^2 - 2\hat{a}s_{xy} + \hat{a}^2 s_x^2 \\
&\rightarrow \text{ par hypothèse } \hat{a} = \frac{s_{xy}}{s_x^2} \\
&= s_y^2 - 2\frac{s_{xy}}{s_x^2}s_{xy} + \frac{s_{xy}^2}{s_x^4}s_x^2 \\
&= s_y^2 - \frac{s_{xy}^2}{s_x^2} \\
&= s_y^2(1 - \frac{s_{xy}^2}{s_x^2 s_y^2}) \\
&= s_y^2(1 - r_{xy}^2)
\end{aligned}
$$

15 Démonstration de la forme des moyennes et variances de la série des valeurs ajustées :

15.1 Hypothèse :

(x_i, y_i) $1 \leq i \leq n$
droite de régression par les moindres carrés
$y = \hat{a}x + \hat{b}$ avec $\hat{a} = \frac{s_{xy}}{s_x^2}$ et $\hat{b} = \overline{y} - \hat{a}x$
$\delta_i = y_i - \hat{y}_i$

15.2 Thèse :

$s_{\hat{y}}^2 = r^2 s_y^2$ et $\overline{y} = \overline{\hat{y}}$

15.3 Démonstration :

15.3.1 Moyenne :

$$\begin{aligned} 0 = \overline{\delta} &= \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i) \\ &= \frac{1}{n}\sum_{i=1}^{n} y_i - \frac{1}{n}\sum_{i=1}^{n}\hat{y}_i \\ &\Rightarrow \overline{\hat{y}_i} = \overline{y} \end{aligned}$$

15.3.2 Variance :

$$\begin{aligned} s_{\hat{y}}^2 &= \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - \overline{\hat{y}})^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}(\hat{a}x_i + \hat{b} - \overline{y})^2 \\ &\rightarrow \text{ par hypothèse } \hat{b} = \overline{y} - \hat{a}\overline{x} \\ &= \frac{1}{n}\sum_{i=1}^{n}(\hat{a}x_i - \hat{a}\overline{x})^2 \\ &= \hat{a}^2\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 \\ &= \hat{a}^2 s_x^2 \\ &\rightarrow \text{ par hypothèse } \hat{a} = \frac{S_{xy}}{S_x^2} \\ &= \frac{S_{xy}^2}{(S_x^2)^2}S_x^2 \\ &= \frac{S_{xy}^2}{S_x^2} \\ &= r^2 S_y^2 \\ &\rightarrow \text{ On a, } r^2 = \frac{S_{xy}^2}{S_x^2 S_y^2} \\ &\Leftrightarrow r^2 S_y^2 = \frac{S_{xy}^2}{S_x^4} \end{aligned}$$

16 Décomposition de la variance :

$$\begin{aligned} s_y^2 &= 1 s_y^2 \\ &= (1 - r^2 + r^2)s_y^2 \\ &= (1 - r^2)s_y^2 + r^2 s_y^2 \\ &= s_\delta^2 + s_{\hat{y}}^2 \end{aligned}$$

17 Borne sur la covariance :

La covariance de la série double S est toujours, en valeur absolue, inférieure ou égale au produit des écarts-types marginaux s_x, s_y : $|s_{xy}| \leq s_x\ s_y$. L'égalité n'est possible que si et seulement si tous les points observés sont situés sur une même droite.

18 Equation de la droite de régression par la technique des moindres carrés :

Soit $S = \{(x_i, y_i), 1 \leq i \leq n\}$ une série statistique (quantitative) bivariée. L'équation de la droite de régression obtenue par la méthode des moindres carrés est donnée par

$$y = \hat{a}x + \hat{b} \quad o\grave{u} \quad \hat{a} = \frac{s_{xy}}{s_x^2} \quad et \quad \hat{b} = \overline{y} - \hat{a}\overline{x}$$