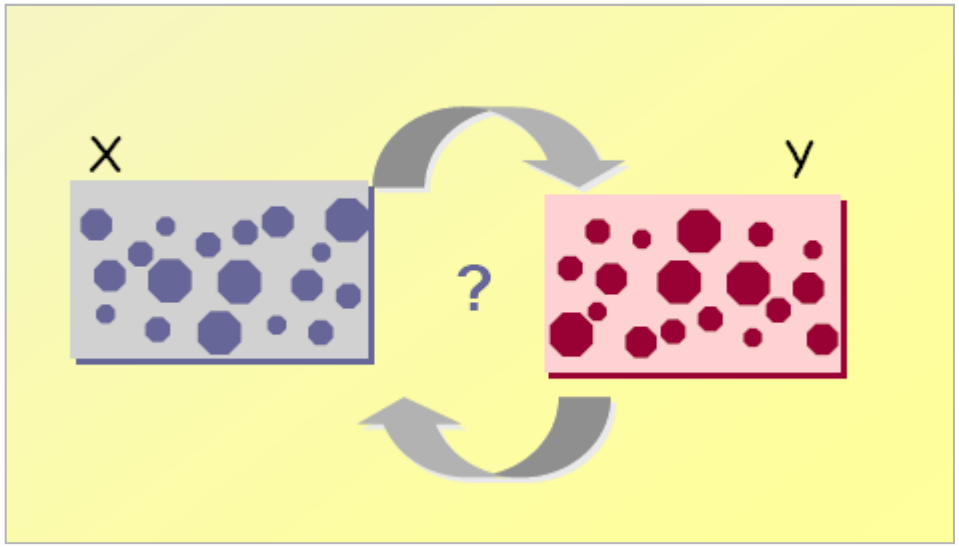
Lecture n° 9.

Regression.



Droite de régression linéaire.

Admettons que:

- nous connaissons n paires de variables x_i,y_i
- ces points se placent sur une droite suffisamment bien pour que le coefficient de corrélation soit proche de 1.

PROBLEME: comment déterminer tels paramètres **a**, **b** pour que la droite $y = ax + b$ passe le plus près des points expérimentaux ?

PROCEDURE:

- Si on connaissait les valeurs a et b, on pourrait calculer, pour chaque valeur de x_i, la valeur théorique y_i (notons la $\hat{y}$):

$$\hat{y}_i = ax_i + b$$

- Notons: $d_i = y_i - \hat{y}_i = y_i - ax_i - b$

- Nous voulons que $\sum_{i=1}^{n} (d_i)^2 = \min$, donc ⇒

$$\frac{\partial}{\partial a}\sum_{i=1}^{n}(d_i)^2 = 0$$

et

$$\frac{\partial}{\partial b}\sum_{i=1}^{n}(d_i)^2 = 0$$

Droite de régression linéaire.

$$\frac{\partial}{\partial a}\sum_{i=1}^{n}(d_i)^2 = 0 \qquad \frac{\partial}{\partial b}\sum_{i=1}^{n}(d_i)^2 = 0$$

$$\frac{\partial}{\partial a}\sum_{i=1}^{n}(y_i - ax_i - b)^2 = 0 \qquad \frac{\partial}{\partial b}\sum_{i=1}^{n}(y_i - ax_i - b)^2 = 0$$

$$-\sum_{i=1}^{n}x_i(y_i - ax_i - b) = 0 \qquad -\sum_{i=1}^{n}(y_i - ax_i - b) = 0$$

$$-\sum_{i=1}^{n}x_i y_i + a\sum_{i=1}^{n}(x_i)^2 + b\sum_{i=1}^{n}x_i = 0 \qquad \sum_{i=1}^{n}y_i - a\sum_{i=1}^{n}x_i - nb = 0$$

x n

$$\times \sum_{i=1}^{n}x_i$$

$$+ \begin{cases} -n\sum_{i=1}^{n}x_i y_i + an\sum_{i=1}^{n}(x_i)^2 + bn\sum_{i=1}^{n}x_i = 0 \\ \sum_{i=1}^{n}x_i\sum_{i=1}^{n}y_i - a\left(\sum_{i=1}^{n}x_i\right)^2 - nb\sum_{i=1}^{n}x_i = 0 \end{cases}$$

Si nous additionnons les équations, les termes contenant b se simplifierons.
Alors, nous pourrons déterminer la valeur de a:

Droite de régression linéaire.

$$a = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} (x_i)^2 - \left(\sum_{i=1}^{n} x_i\right)^2} = \frac{\overline{xy} - \bar{x}\cdot\bar{y}}{\overline{x^2} - \bar{x}^2}$$

$$b = -\frac{\sum_{i=1}^{n} x_i \sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} (x_i)^2 \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} (x_i)^2 - \left(\sum_{i=1}^{n} x_i\right)^2} = \frac{\overline{x^2}\bar{y} - \bar{x}\cdot\overline{xy}}{\overline{x^2} - \bar{x}^2}$$

Nous avons alors des formules permettant de déterminer
les coefficients de la meilleure droite passant a travers des points (x_i, y_i),
tenant compte de **TOUS** les points.

$$a = \frac{\overline{xy} - \bar{x}\cdot\bar{y}}{\overline{x^2} - \bar{x}^2} = \frac{\mathrm{cov}(x,y)}{S_x^2}$$

$$b = \frac{\overline{x^2}\bar{y} - \bar{x}\cdot\overline{xy}}{\overline{x^2} - \bar{x}^2} = \bar{y} - a\bar{x}$$

Incertitude sur les paramètres de la droite de régression linéaire.

Une fois les paramètres de la droite de régression déterminés, nous pouvons évaluer la précision de leur estimation.

PROCEDURE:

- Connaissant les valeurs a et b, calculons, pour chaque valeur de x_i, la valeur $\hat{y}_i$:

$$\hat{y}_i = ax_i + b$$

- Notons $d_i = y_i - ax_i - b$
- En utilisant la formule de propagation de variances, on obtient

$$(n-2)S_y^2 = \sum_{i=1}^{n}(d_i)^2 = \sum_{i=1}^{n} d_i y_i - a\sum_{i=1}^{n} d_i x_i - b\sum_{i=1}^{n} d_i$$

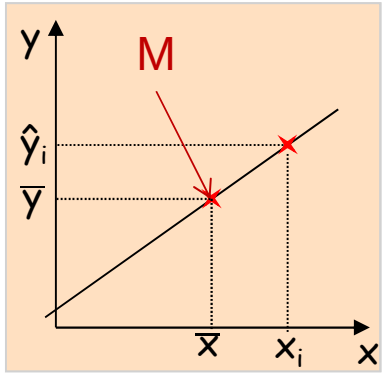
$$S_a^2 = \frac{1}{n-2}\frac{\overline{y^2} - \bar{y}^2}{\overline{y^2} - \bar{x}^2} - a^2$$

$$S_b^2 = \sqrt{S_a^2 \cdot \overline{x^2}}$$

Estimation de la droite de régression linéaire.

Nous pouvons aussi estimer l'imprécision (l'intervalle de confiance) de chaque valeur théorique $\hat{y}_i$:

PROCEDURE:



- Prenons comme point de départ de l'analyse le point $M(\bar{x},\bar{y})$.
- Connaissant les valeurs a et b nous pouvons calculer, pour chaque valeur de x_i, la valeur $\hat{y}_i$:

$$\hat{y}_i = ax_i + b$$

$$\hat{y}_i = \bar{y} + a(x_i - \bar{x})$$

- Si on connaît précisément la valeur x_i (alors aussi $\bar{x}$), $\hat{y}_i$ est la fonction de deux variables uniquement:

$$\hat{y}_i = f(\bar{y}, a)$$

- En utilisant la formule de propagation de variances, on obtient

$$S^2_{\hat{y}_i} = \left(\frac{\partial \hat{y}_i}{\partial \bar{y}}\right)^2 S^2_{\bar{y}} + \left(\frac{\partial \hat{y}_i}{\partial a}\right)^2 S^2_a$$

avec

$$\frac{\partial \hat{y}_i}{\partial \bar{y}} = 1$$

$$\frac{\partial \hat{y}_i}{\partial a} = x_i - \bar{x}$$

Estimation de la droite de régression linéaire.

- On obtient alors

$$S^2_{\hat{y}_i} = S^2_{\bar{y}} + (x_i - \bar{x})^2 S^2_a$$

avec

$$S^2_{\bar{y}} = \frac{S^2_y}{n} = \frac{1}{n(n-2)} \sum_{i=1}^{n} (y_i - \bar{y}_i)^2$$

$$S^2_a = \frac{1}{\sum_{i=1}^{n} (x_i - \bar{x})^2} S^2_y$$

Donc,

$$S^2_{\hat{y}_i} = S^2_{\bar{y}} + \frac{(x_i - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} S^2_y = \frac{S^2_y}{n} + \frac{(x_i - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} S^2_y$$

$$S_{\hat{y}_i} = S_y \sqrt{\frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}}$$

Estimation de la droite de régression linéaire.

Donc, chaque valeur $\hat{y}_i$ déterminée par la formule de droite de régression linéaire est entourée d'un intervalle de confiance .

- Au niveau de risque $\alpha \approx 0,31$ (précision d'un écart type), l'intervalle de $\hat{y}_i$ est

$$(\hat{y} - S_{\hat{y}}; \hat{y} + S_{\hat{y}})$$

- Pour un niveau de confiance α arbitrairement choisi,

$$(\hat{y} - u_0 S_{\hat{y}}; \hat{y} + u_0 S_{\hat{y}}) \Rightarrow$$

- Si n > 30, u_0 est le coefficient multiplicatif de la loi normale
- Si n < 30, u_0 est le coefficient de Student

CONSEQUENCE:

L'incertitude de $\hat{y}_i$ varie d'un point de droite de régression à l'autre:

- elle est minimale pour $x_i = \overline{x}$,
- elle augmente quand x_i s'écarte de $\overline{x}$.

EXEMPLE: Estimer la corrélation et la droite de régression linéaire pour un ensemble de 10 mesures (x_i,y_i):

1. Calcul de corrélation et de droite de régression:

x_i	y_i	$(x_i)^2$	$(y_i)^2$	$x_i y_i$	$(d_i)^2$
-6	15.0	36	225	-90	1.21
-3	9.1	9	83	-27.3	.16
0	2.8	0	7.9	0	.01
1	1.1	1	1.2	1.1	.16
2	-0.9	4	.81	-1.8	.36
5	-6.8	25	46.3	-34	1.69
7	-11.3	49	128	-79.2	1.44
10	-19.9	100	395	-199	6.64
11	-22.1	121	488	-243	.64
13	-26.3	169	692	-342	.36
40	-59.3	514	2067.2	-1015.2	6.67

$$r = \frac{\overline{xy} - \bar{x}\cdot\bar{y}}{S_x S_y} = \frac{101.52 - 4(-5.93)}{(51.4 - 4^2)(206.72 - 5.93^2)}$$

$$r \approx -1$$

corrélation négative presque idéale

$$a = \frac{\overline{xy} - \bar{x}\cdot\bar{y}}{S_x^2} = \frac{101.52 - 4(-5.93)}{51.4 - 4^2} = -2.19$$

$$b = \bar{y} - a\bar{x} = 2.83$$

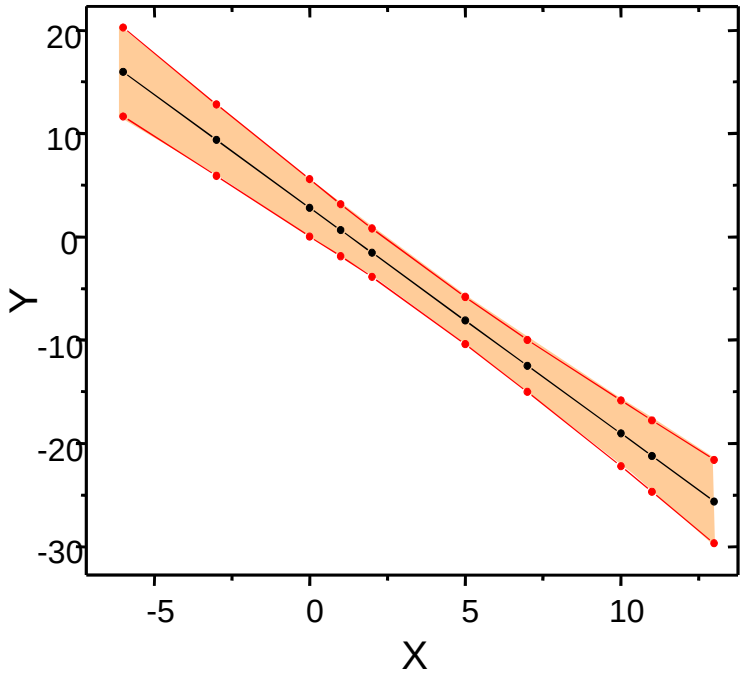
$$S_a = S_y\sqrt{\frac{1}{nS_x}} = 0.913\sqrt{\frac{1}{10(51.4-16)}} \approx 0.05$$

$$S_b = S_a\sqrt{\bar{x}^2} = 0.5\sqrt{51.4} \approx 0.4$$

2. Calcul de l'intervalle de confiance de la droite de régression:

x_i	y_i	x_i-x	$a(x_i-x)$	$(x_i-x)^2$	$y_i - \Delta y$	$y_i + \Delta y$
-6	15.0	-10	21.9	100	11.68	20.26
-3	9.1	-7	15.3	49	5.91	12.83
0	2.8	-4	8.75	16	0.04	5.60
1	1.1	-3	6.57	9	-1.88	3.16
2	-0.9	-2	4.38	4	-3.9	0.8
5	-6.8	1	-2.19	1	-10.4	-5.84
7	-11.3	3	-6.57	9	-15.02	-9.98
10	-19.9	6	-13.1	36	-22.22	-15.84
11	-22.1	7	-15.3	49	-24.69	-17.77
13	-26.3	9	-19.7	81	-29.66	-21.6
40	-59.3			354		

- Pour n = 10 et le niveau de confiance α = 0.99, u_0=3.3554



$$\Delta y = u_0 \times S_y \sqrt{\frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}}$$

Comparaison des paramètres d'une droite à des valeurs théoriques.

(TEST DE CONFORMITE)

Admettons que dans une population P les variables X et Y soient liées par la relation

$$Y = AX + B$$

A partir d'une étude de l'échantillon (une série des mesures x_i, y_i , nous trouvons

$$y = ax + b$$

PROBLEME: Notre observation est-elle conforme à la loi théorique?

Hypothèse H_0 : a = A, b = B

Sous l'hypothèse H_0, la variable aléatoire T_a (T_b)

$$T_a = \frac{a - A}{S_A} \qquad (T_b = \frac{b - B}{S_B})$$

suit la loi de Student à ν = n-2 degrés de liberté.

Comparaison de deux droites de régression experimentales.

(TEST d'HOMOGENEITE)

- Admettons que dans une population P les variables X et Y sont liées.
- Dans une autre population P', les variables X et Y' sont aussi liées.
- Dans les échantillons E et E', extraits de ces populations, les relations observées entre le variables sont:

$$y = ax + b \qquad (S_R)$$

$$y' = a'x' + b' \qquad (S_R')$$

Hypothèse H_0 : Les coefficients directeurs de deux droites sont identiques : a = a'

Le test nécessite la vérification préalable d'égalité de variances de deux populations.

$$\sigma^2 = \frac{(n-2)S_R + (n'-2)S_R'}{n + n' - 4}$$

Théorème : Sous l'hypothèse H_0 et si s = s', la variable T

$$T = \frac{a - a'}{\sigma\sqrt{\frac{1}{nS^2(X)} + \frac{1}{n'S^2(X)}}}$$

suit la loi de Student à ν = n + n' - 4 degrés de liberté.