Analyse de données

Données bidimensionnelles

Jamal Atif
jamal.atif@dauphine.fr

Université Paris-Dauphine, Licence MIDO

2014-2015

Cet épisode

Lien entre deux variables

quantitatives :

- Covariance
- Centrage, réduction
- Corrélation

qualitatives :

- Tableaux de contingence
- Test χ^2

Variables quantitatives

1. Deux variables quantitatives sont-elles liées ?
2. Evoluent-elles dans le même sens ?

Exemple

	Années d'exp.	Productivité (tonnes)
Jean	0	1
Robert	1	3
Isham	2	4
Marc	5	4
Moyenne	2	3

Premières intuitions

- Dessiner le nuage de points
- Analyse l'évolution simultanée des données par rapport à leurs moyenne $\rightarrow$ **Covariance**

Variables quantitatives

1. Deux variables quantitatives sont-elles liées ?
2. Evoluent-elles dans le même sens ?

Exemple

	Années d'exp.	Productivité (tonnes)
Jean	0	1
Robert	1	3
Isham	2	4
Marc	5	4
Moyenne	2	3

Premières intuitions

- Dessiner le nuage de points
- Analyse l'évolution simultanée des données par rapport à leurs moyenne → **Covariance**

Variables quantitatives

1. Deux variables quantitatives sont-elles liées ?
2. Evoluent-elles dans le même sens ?

Exemple

	Années d'exp.	Productivité (tonnes)
Jean	0	1
Robert	1	3
Isham	2	4
Marc	5	4
Moyenne	2	3

Premières intuitions

- Dessiner le nuage de points
- Analyse l'évolution simultanée des données par rapport à leurs moyenne → **Covariance**

Variables quantitatives

1. Deux variables quantitatives sont-elles liées ?
2. Evoluent-elles dans le même sens ?

Exemple

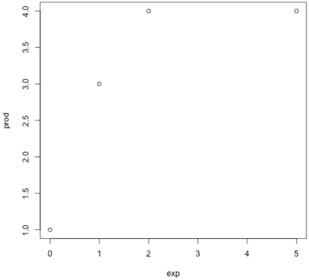
	Années d'exp.	Productivité (tonnes)
Jean	0	1
Robert	1	3
Isham	2	4
Marc	5	4
Moyenne	2	3

Premières intuitions

- Dessiner le nuage de points
- Analyse l'évolution simultanée des données par rapport à leurs moyenne $\rightarrow$ **Covariance**

Nuage de points

	Années d'exp.	Productivité (tonnes)
Jean	0	1
Robert	1	3
Isham	2	4
Marc	5	4
Moyenne	2	3
Ecart-type	2.16	1.41



Covariance

Définition

La covariance la moyenne de la somme du produit des écarts des valeurs des deux variables par rapport à leur moyenne arithmétique ($\bar{\cdot}$). Le terme « covariation » désigne cette dernière somme. On peut définir la covariance comme la moyenne de la covariation.

$$\sigma_{\mathrm{xy}} = \frac{\sum_{i=1}^{n}(x_i - \overline{\mathrm{x}})(y_i - \overline{\mathrm{y}})}{n}$$

Exemple : calcul

	Années d'exp.	Productivité (tonnes)
Jean	0	1
Robert	1	3
Isham	2	4
Marc	5	4
Moyenne	2	3

- Variables : x=Années d'expériences, y = Productivité
- $n = 4, \overline{x} = 2, \overline{y} = 3.$

Produit des écart d'expérience par personne :

- Jean : $(0-2) \times (1-3) = 4$
- Robert : $(1-2) \times (3-3) = 0$
- Isham : $(2-2) \times (4-3) = 0$
- Marc : $(5-2) \times (4-3) = 3$

La covariance moyenne est donc : $\frac{4+0+0+3}{4} = 1.75$

Exemple : calcul

	Années d'exp.	Productivité (tonnes)
Jean	0	1
Robert	1	3
Isham	2	4
Marc	5	4
Moyenne	2	3

- Variables : x=Années d'expériences, y = Productivité
- $n = 4, \overline{\mathrm{x}} = 2, \overline{\mathrm{y}} = 3.$

Produit des écart d'expérience par personne :

- Jean : $(0-2) \times (1-3) = 4$
- Robert : $(1-2) \times (3-3) = 0$
- Isham : $(2-2) \times (4-3) = 0$
- Marc : $(5-2) \times (4-3) = 3$

La covariance moyenne est donc : $\frac{4+0+0+3}{4} = 1.75$

Exemple : calcul

	Années d'exp.	Productivité (tonnes)
Jean	0	1
Robert	1	3
Isham	2	4
Marc	5	4
Moyenne	2	3

- Variables : x=Années d'expériences, y = Productivité
- $n = 4, \overline{\mathrm{x}} = 2, \overline{\mathrm{y}} = 3.$

Produit des écart d'expérience par personne :

- Jean : $(0 - 2) \times (1 - 3) = 4$
- Robert : $(1 - 2) \times (3 - 3) = 0$
- Isham : $(2 - 2) \times (4 - 3) = 0$
- Marc : $(5 - 2) \times (4 - 3) = 3$

La covariance moyenne est donc : $\frac{4+0+0+3}{4} = 1.75$

Exemple : calcul

	Années d'exp.	Productivité (tonnes)
Jean	0	1
Robert	1	3
Isham	2	4
Marc	5	4
Moyenne	2	3

- Variables : x=Années d'expériences, y = Productivité
- $n = 4, \overline{\mathrm{x}} = 2, \overline{\mathrm{y}} = 3.$

Produit des écart d'expérience par personne :

- Jean : $(0 - 2) \times (1 - 3) = 4$
- Robert : $(1 - 2) \times (3 - 3) = 0$
- Isham : $(2 - 2) \times (4 - 3) = 0$
- Marc : $(5 - 2) \times (4 - 3) = 3$

La covariance moyenne est donc : $\frac{4+0+0+3}{4} = 1.75$

Exemple : intepprétation

Règle

La règle générale d'interprétation veut que plus la covariance est élevée, plus la relation entre les deux variables est forte.

- Si les écarts des deux variables avec la moyenne sont toutes deux négatives, leur produit donne un nombre positif qui contribue à la covariance : l'écart de Jean vaut -2 sur les années d'expérience et -2 également sur la productivité. Le produit de -2 * -2 donne 4, qui ajoute autant à la covariance.
- Lorsqu'elles ne varient pas ensemble, le signe des écarts est inversé. leur produit est négatif et diminue la valeur de la covariance. Par exemple, un écart de -3 pour les années d'expérience et de +2 pour la productivité donne un produit de -3 * 2 = -6. La covariance est réduite d'autant.
- C'est pourquoi une valeur positive de la covariance indique que les deux variables varient ensemble. Pour revenir à notre exemple, plus nous avons de l'expérience plus nous sommes productifs.

Exemple : intepprétation

Règle

La règle générale d'interprétation veut que plus la covariance est élevée, plus la relation entre les deux variables est forte.

- Si les écarts des deux variables avec la moyenne sont toutes deux négatives, leur produit donne un nombre positif qui contribue à la covariance : l'écart de Jean vaut -2 sur les années d'expérience et -2 également sur la productivité. Le produit de -2 * -2 donne 4, qui ajoute autant à la covariance.
- Lorsqu'elles ne varient pas ensemble, le signe des écarts est inversé. leur produit est négatif et diminue la valeur de la covariance. Par exemple, un écart de -3 pour les années d'expérience et de +2 pour la productivité donne un produit de -3 * 2 = -6. La covariance est réduite d'autant.
- C'est pourquoi une valeur positive de la covariance indique que les deux variables varient ensemble. Pour revenir à notre exemple, plus nous avons de l'expérience plus nous sommes productifs.

Exemple : intepprétation

Règle

La règle générale d'interprétation veut que plus la covariance est élevée, plus la relation entre les deux variables est forte.

- Si les écarts des deux variables avec la moyenne sont toutes deux négatives, leur produit donne un nombre positif qui contribue à la covariance : l'écart de Jean vaut -2 sur les années d'expérience et -2 également sur la productivité. Le produit de -2 * -2 donne 4, qui ajoute autant à la covariance.
- Lorsqu'elles ne varient pas ensemble, le signe des écarts est inversé. leur produit est négatif et diminue la valeur de la covariance. Par exemple, un écart de -3 pour les années d'expérience et de +2 pour la productivité donne un produit de -3 * 2 = -6. La covariance est réduite d'autant.
- C'est pourquoi une valeur positive de la covariance indique que les deux variables varient ensemble. Pour revenir à notre exemple, plus nous avons de l'expérience plus nous sommes productifs.

Exemple : intepprétation

Règle

La règle générale d'interprétation veut que plus la covariance est élevée, plus la relation entre les deux variables est forte.

- Si les écarts des deux variables avec la moyenne sont toutes deux négatives, leur produit donne un nombre positif qui contribue à la covariance : l'écart de Jean vaut -2 sur les années d'expérience et -2 également sur la productivité. Le produit de -2 * -2 donne 4, qui ajoute autant à la covariance.
- Lorsqu'elles ne varient pas ensemble, le signe des écarts est inversé. leur produit est négatif et diminue la valeur de la covariance. Par exemple, un écart de -3 pour les années d'expérience et de +2 pour la productivité donne un produit de -3 * 2 = -6. La covariance est réduite d'autant.
- C'est pourquoi une valeur positive de la covariance indique que les deux variables varient ensemble. Pour revenir à notre exemple, plus nous avons de l'expérience plus nous sommes productifs.

Interprétation

- Plus la covariance est élevée, plus la relation entre les deux variables est forte
- Si cov = 0 $\rightarrow$ les deux variables sont indépendantes *linéairement*
- Si cov < 0, $\rightarrow$ les deux variables varient en sens inverse. exemple : vitesse à la course à pied des hommes de 40 ans et plus est négative $\rightarrow$ + l'âge est élevé, - la vitesse l'est
- La covariance est une mesure symétrique,
- Une relation non linéaire peut exister même si cov=0
- Toute comparaison avec d'autres variables est interdite (cov non standardisée).
- Ne s'applique qu'aux variables quantitatives.

Interprétation

- Plus la covariance est élevée, plus la relation entre les deux variables est forte
- Si cov = 0 $\rightarrow$ les deux variables sont indépendantes *linéairement*
- Si cov < 0, $\rightarrow$ les deux variables varient en sens inverse. exemple : vitesse à la course à pied des hommes de 40 ans et plus est négative $\rightarrow$ + l'âge est élevé, - la vitesse l'est
- La covariance est une mesure symétrique,
- Une relation non linéaire peut exister même si cov=0
- Toute comparaison avec d'autres variables est interdite (cov non standardisée).
- Ne s'applique qu'aux variables quantitatives.

Interprétation

- Plus la covariance est élevée, plus la relation entre les deux variables est forte
- Si cov = 0 $\rightarrow$ les deux variables sont indépendantes *linéairement*
- Si cov < 0, $\rightarrow$ les deux variables varient en sens inverse. exemple : vitesse à la course à pied des hommes de 40 ans et plus est négative $\rightarrow$ + l'âge est élevé, - la vitesse l'est
- La covariance est une mesure symétrique,
- Une relation non linéaire peut exister même si cov=0
- Toute comparaison avec d'autres variables est interdite (cov non standardisée).
- Ne s'applique qu'aux variables quantitatives.

Interprétation

- Plus la covariance est élevée, plus la relation entre les deux variables est forte
- Si cov = 0 $\rightarrow$ les deux variables sont indépendantes *linéairement*
- Si cov < 0, $\rightarrow$ les deux variables varient en sens inverse. exemple : vitesse à la course à pied des hommes de 40 ans et plus est négative $\rightarrow$ + l'âge est élevé, - la vitesse l'est
- La covariance est une mesure symétrique,
- Une relation non linéaire peut exister même si cov=0
- Toute comparaison avec d'autres variables est interdite (cov non standardisée).
- Ne s'applique qu'aux variables quantitatives.

Interprétation

- Plus la covariance est élevée, plus la relation entre les deux variables est forte
- Si cov = 0 $\rightarrow$ les deux variables sont indépendantes *linéairement*
- Si cov < 0, $\rightarrow$ les deux variables varient en sens inverse. exemple : vitesse à la course à pied des hommes de 40 ans et plus est négative $\rightarrow$ + l'âge est élevé, - la vitesse l'est
- La covariance est une mesure symétrique,
- Une relation non linéaire peut exister même si cov=0
- Toute comparaison avec d'autres variables est interdite (cov non standardisée).
- Ne s'applique qu'aux variables quantitatives.

Interprétation

- Plus la covariance est élevée, plus la relation entre les deux variables est forte
- Si cov = 0 $\rightarrow$ les deux variables sont indépendantes *linéairement*
- Si cov < 0, $\rightarrow$ les deux variables varient en sens inverse. exemple : vitesse à la course à pied des hommes de 40 ans et plus est négative $\rightarrow$ + l'âge est élevé, - la vitesse l'est
- La covariance est une mesure symétrique,
- Une relation non linéaire peut exister même si cov=0
- Toute comparaison avec d'autres variables est interdite (cov non standardisée).
- Ne s'applique qu'aux variables quantitatives.

Interprétation

- Plus la covariance est élevée, plus la relation entre les deux variables est forte
- Si cov = 0 $\rightarrow$ les deux variables sont indépendantes *linéairement*
- Si cov < 0, $\rightarrow$ les deux variables varient en sens inverse. exemple : vitesse à la course à pied des hommes de 40 ans et plus est négative $\rightarrow$ + l'âge est élevé, - la vitesse l'est
- La covariance est une mesure symétrique,
- Une relation non linéaire peut exister même si cov=0
- Toute comparaison avec d'autres variables est interdite (cov non standardisée).
- Ne s'applique qu'aux variables quantitatives.

Limites de la covariance

Problème

On ne peut pas comparer des covariances entre elles (et plus généralement des variables numériques) présentant des unités de mesures différentes.

Solution : standardiser les données

Principe : Toute répartition statistique définie par une moyenne et un écart-type peut être transformée en une autre distribution statistique qui a pour moyenne 0 et pour écart-type 1. La nouvelle variable obtenue est alors dite "centrée réduite".

Limites de la covariance

Problème

On ne peut pas comparer des covariances entre elles (et plus généralement des variables numériques) présentant des unités de mesures différentes.

Solution : standardiser les données

Principe : Toute répartition statistique définie par une moyenne et un écart-type peut être transformée en une autre distribution statistique qui a pour moyenne 0 et pour écart-type 1. La nouvelle variable obtenue est alors dite "centrée réduite".

Distance entre individus

Quel individu est le plus similaire de l'individu représenté par un carré (au sens de la distance euclidienne) ?

Distance entre individus

Quel individu est le plus similaire de l'individu représenté par un carré (au sens de la distance euclidienne) ?

Solution

Quand les variables sont mesurées avec des échelles différentes, il peut s'avérer utile de « réduire » (normer) ces variables.

Distance entre individus

Quel individu est le plus similaire de l'individu représenté par un carré (au sens de la distance euclidienne) ?

Solution

Quand les variables sont mesurées avec des échelles différentes, il peut s'avérer utile de « réduire » (normer) ces variables.

Normalisation

Centrage

Soustraire la moyenne :

$$z_i = x_i - \overline{\mathrm{x}}$$

Réduction

Diviser par l'écart-type :

$$z_i = \frac{x_i}{\sigma_{\mathrm{x}}}$$

Variable centrée réduite :

$$z_i = \frac{x_i - \overline{\mathrm{x}}}{\sigma_{\mathrm{x}}}$$

La moyenne de la variable est nulle et son écart-type est égal à un. $\overline{\mathrm{z}} = 0, \sigma_{\mathrm{z}} = 1$.

Normalisation

Centrage

Soustraire la moyenne :

$$z_i = x_i - \overline{\mathrm{x}}$$

Réduction

Diviser par l'écart-type :

$$z_i = \frac{x_i}{\sigma_{\mathrm{x}}}$$

Variable centrée réduite :

$$z_i = \frac{x_i - \overline{\mathrm{x}}}{\sigma_{\mathrm{x}}}$$

La moyenne de la variable est nulle et son écart-type est égal à un. $\overline{\mathrm{z}} = 0, \sigma_{\mathrm{z}} = 1$.

Normalisation

Centrage

Soustraire la moyenne :

$$z_i = x_i - \overline{\mathrm{x}}$$

Réduction

Diviser par l'écart-type :

$$z_i = \frac{x_i}{\sigma_{\mathrm{x}}}$$

Variable centrée réduite :

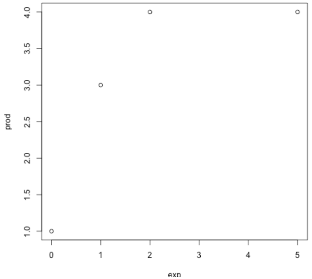
$$z_i = \frac{x_i - \overline{\mathrm{x}}}{\sigma_{\mathrm{x}}}$$

La moyenne de la variable est nulle et son écart-type est égal à un. $\overline{\mathrm{z}} = 0, \sigma_{\mathrm{z}} = 1$.

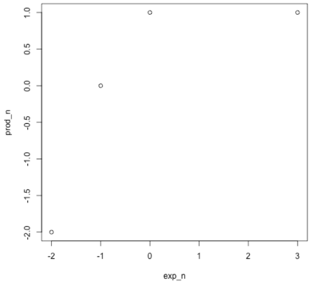
Exemple

Centrage

	Années d'exp.	Productivité (tonnes)
Jean	0	1
Robert	1	3
Isham	2	4
Marc	5	4
Moyenne	2	3
Ecart-type	2.16	1.41



	Années d'exp.	Productivité (tonnes)
Jean	-2	-2
Robert	-1	0
Isham	0	1
Marc	3	1
Moyenne	0	0
Ecart-type	2.16	1.41



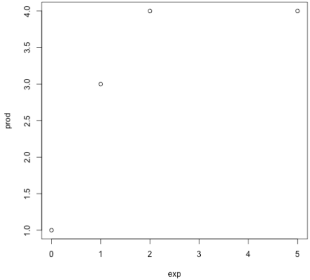
Centrage du nuage de points

Les unités sont encore en temps et en tonnes

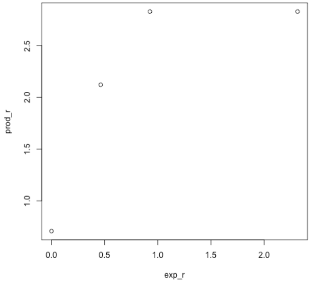
Exemple

Réduction

	Années d'exp.	Productivité (tonnes)
Jean	0	1
Robert	1	3
Isham	2	4
Marc	5	4
Moyenne	2	3
Ecart-type	2.16	1.41



	Années d'exp.	Productivité (tonnes)
Jean	0	-0.7071068
Robert	0.4629100	2.1213203
Isham	0.9258201	2.8284271
Marc	2.3145502	2.8284271
Moyenne	0.9258	2.1213
Ecart-type	1	1

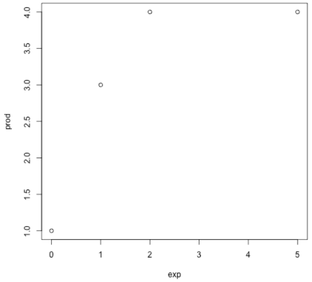


Il n'y a plus d'unités du fait de la division.

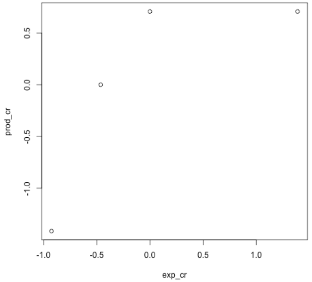
Exemple

Centrage et Réduction

	Années d'exp.	Productivité (tonnes)
Jean	0	1
Robert	1	3
Isham	2	4
Marc	5	4
Moyenne	2	3
Ecart-type	2.16	1.41



	Années d'exp.	Productivité (tonnes)
Jean	-0.9258201	-1.4142136
Robert	-0.4629100	0
Isham	0	0.7071068
Marc	1.3887301	0.7071068
Moyenne	0	0
Ecart-type	1	1



Corrélation

- hypothèse sur la « forme » de la relation entre x et y (p.ex. linéaire)
- coefficient de Pearson (corrélation linéaire) d'une série de valeur $(x_i, y_i)_{i=1}^{n}$

$$\rho_{\mathrm{xy}} = \frac{\sum_i (x_i - \overline{\mathrm{x}}) \cdot (y_i - \overline{\mathrm{y}})}{n \cdot \sigma_{\mathrm{x}} \cdot \sigma_{\mathrm{y}}}$$

- Covariance après réduction :

$$\rho_{\mathrm{xy}} = \frac{\sigma_{\mathrm{xy}}}{\sigma_{\mathrm{y}} \sigma_{\mathrm{y}}}$$

- attention au problème d'instabilité numérique lors du calcul
- indique la « force » d'une corrélation linéaire entre les données

Corrélation

- hypothèse sur la « forme » de la relation entre x et y (p.ex. linéaire)
- coefficient de Pearson (corrélation linéaire) d'une série de valeur $(x_i, y_i)_{i=1}^{n}$

$$\rho_{\mathrm{xy}} = \frac{\sum_i (x_i - \overline{\mathrm{x}}) \cdot (y_i - \overline{\mathrm{y}})}{n \cdot \sigma_{\mathrm{x}} \cdot \sigma_{\mathrm{y}}}$$

- Covariance après réduction :

$$\rho_{\mathrm{xy}} = \frac{\sigma_{\mathrm{xy}}}{\sigma_{\mathrm{y}}\sigma_{\mathrm{y}}}$$

- attention au problème d'instabilité numérique lors du calcul
- indique la « force » d'une corrélation linéaire entre les données

Corrélation

- hypothèse sur la « forme » de la relation entre x et y (p.ex. linéaire)
- coefficient de Pearson (corrélation linéaire) d'une série de valeur $(x_i, y_i)_{i=1}^{n}$

$$\rho_{\mathrm{xy}} = \frac{\sum_i (x_i - \overline{\mathrm{x}}) \cdot (y_i - \overline{\mathrm{y}})}{n \cdot \sigma_{\mathrm{x}} \cdot \sigma_{\mathrm{y}}}$$

- Covariance après réduction :

$$\rho_{\mathrm{xy}} = \frac{\sigma_{\mathrm{xy}}}{\sigma_{\mathrm{y}} \sigma_{\mathrm{y}}}$$

- attention au problème d'instabilité numérique lors du calcul
- indique la « force » d'une corrélation linéaire entre les données

Corrélation

- hypothèse sur la « forme » de la relation entre x et y (p.ex. linéaire)
- coefficient de Pearson (corrélation linéaire) d'une série de valeur $(x_i, y_i)_{i=1}^n$

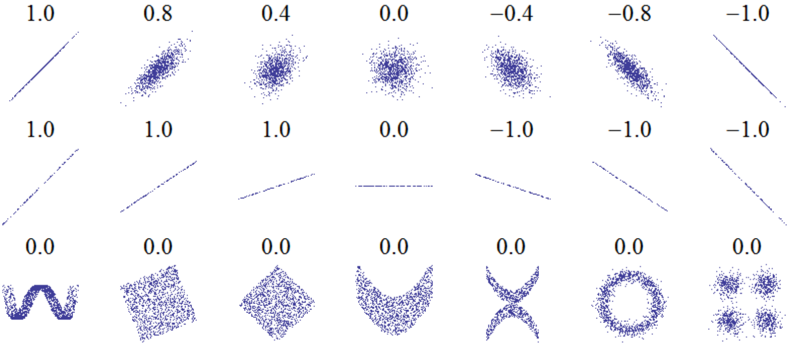
$$\rho_{\mathrm{xy}} = \frac{\sum_i (x_i - \overline{\mathrm{x}}) \cdot (y_i - \overline{\mathrm{y}})}{n \cdot \sigma_{\mathrm{x}} \cdot \sigma_{\mathrm{y}}}$$

- Covariance après réduction :

$$\rho_{\mathrm{xy}} = \frac{\sigma_{\mathrm{xy}}}{\sigma_{\mathrm{y}}\sigma_{\mathrm{y}}}$$

- attention au problème d'instabilité numérique lors du calcul
- indique la « force » d'une corrélation linéaire entre les données

En image



En texte

- ρ est une mesure symétrique.
- $-1 < \rho < 1$.
- si $\rho = 0$, il n'y a pas de relation linéaire entre les deux variables.
- Le signe de la relation indique le sens de la relation.
- Plus ρ est proche de 1 (ou -1) plus la relation entre les deux variables est forte.

En texte

- ρ est une mesure symétrique.
- $-1 < \rho < 1$.
- si $\rho = 0$, il n'y a pas de relation linéaire entre les deux variables.
- Le signe de la relation indique le sens de la relation.
- Plus ρ est proche de 1 (ou -1) plus la relation entre les deux variables est forte.

En texte

- ρ est une mesure symétrique.
- $-1 < \rho < 1$.
- si $\rho = 0$, il n'y a pas de relation linéaire entre les deux variables.
- Le signe de la relation indique le sens de la relation.
- Plus ρ est proche de 1 (ou -1) plus la relation entre les deux variables est forte.

En texte

- ρ est une mesure symétrique.
- $-1 < \rho < 1$.
- si $\rho = 0$, il n'y a pas de relation linéaire entre les deux variables.
- Le signe de la relation indique le sens de la relation.
- Plus ρ est proche de 1 (ou -1) plus la relation entre les deux variables est forte.

En texte

- ρ est une mesure symétrique.
- $-1 < \rho < 1$.
- si $\rho = 0$, il n'y a pas de relation linéaire entre les deux variables.
- Le signe de la relation indique le sens de la relation.
- Plus ρ est proche de 1 (ou -1) plus la relation entre les deux variables est forte.

En texte

- ρ est une mesure symétrique.
- $-1 < \rho < 1$.
- si $\rho = 0$, il n'y a pas de relation linéaire entre les deux variables.
- Le signe de la relation indique le sens de la relation.
- Plus ρ est proche de 1 (ou -1) plus la relation entre les deux variables est forte.

Calcul sur l'exemple

	Années d'exp.	Productivité (tonnes)
Jean	0	1
Robert	1	3
Isham	2	4
Marc	5	4
Moyenne	2	3
Ecart-type	2.16	1.41

$$\rho(exp, prod) = ?$$

Autres mesures

- le coefficient de Pearson ne permet de détecter que certaines dépendances très particulières
- il existe des mesures plus générales permettant de détecter quasiment toutes les dépendances fonctionnelles :
  - information mutuelle
  - corrélation polychorique
  - copule
  - ...

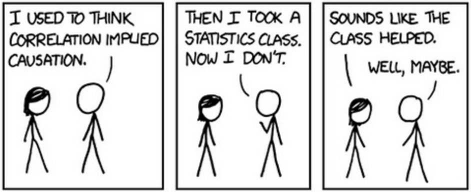
Autres mesures

- le coefficient de Pearson ne permet de détecter que certaines dépendances très particulières
- il existe des mesures plus générales permettant de détecter quasiment toutes les dépendances fonctionnelles :
  - information mutuelle
  - corrélation polychorique
  - copule
  - ...

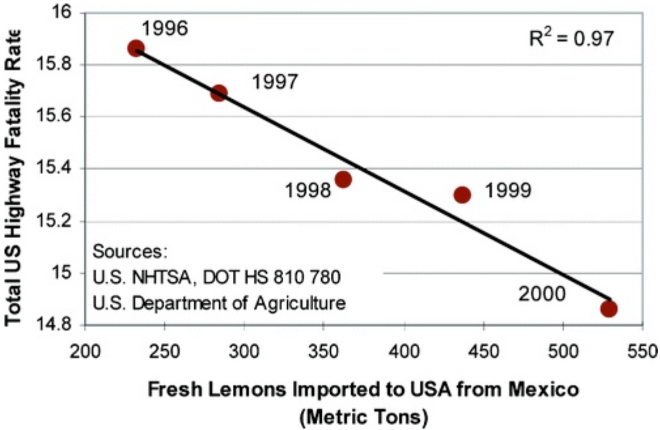
Corrélation et causalité

Il n'y a pas de lien entre corrélation et causalité

- corrélations : certaines valeurs « évoluent » dans le même sens
- causalité : lien « physique »
- quantification d'observations statistiques $\neq$ modélisation d'un système



Exemple



Total US Highway Fatality Rate
16
15.8
15.6
15.4
15.2
15
14.8
1996
1997
1998
1999
2000
R² = 0.97
Sources:
U.S. NHTSA, DOT HS 810 780
U.S. Department of Agriculture
200
250
300
350
400
450
500
550
Fresh Lemons Imported to USA from Mexico
(Metric Tons)

Un peu de math !

Produit scalaire : rappel

Soit deux vecteurs x, y dans $\mathbb{R}^n$: $\mathrm{x} = \{x_1, \cdots, x_n\}$ et $\mathrm{y} = \{y_1, \cdots, y_n\}$

$$\langle \mathrm{x}, \mathrm{y} \rangle = \sum_{i=1}^{n} x_i y_i = x_1 y_1 + x_2 y_2 + \cdots + x_n y_n$$

Lien avec la covariance :

$$\langle \mathrm{x} - \overline{\mathrm{x}}, \mathrm{y} - \overline{\mathrm{y}} \rangle = \sum_{i=1}^{n} (x_i - \overline{\mathrm{x}})(y_i - \overline{\mathrm{y}}) = n\sigma_{\mathrm{xy}}$$

Lien avec la corrélation :

$$\rho_{\mathrm{xy}} = \frac{\sigma_{\mathrm{xy}}}{\sigma_{\mathrm{x}}\sigma_{\mathrm{y}}} = \frac{\sum_{i=1}^{n} (x_i - \overline{\mathrm{x}})(y_i - \overline{\mathrm{y}})}{n\sigma_{\mathrm{x}}\sigma_{\mathrm{y}}} = \frac{1}{n} \left\langle \frac{x - \overline{\mathrm{x}}}{\sigma_{\mathrm{x}}}, \frac{y - \overline{\mathrm{y}}}{\sigma_{\mathrm{y}}} \right\rangle$$

Un peu de math !

Produit scalaire : rappel

Soit deux vecteurs x, y dans $\mathbb{R}^n$: $\mathrm{x} = \{x_1, \cdots, x_n\}$ et $\mathrm{y} = \{y_1, \cdots, y_n\}$

$$\langle \mathrm{x}, \mathrm{y} \rangle = \sum_{i=1}^{n} x_i y_i = x_1 y_1 + x_2 y_2 + \cdots + x_n y_n$$

Lien avec la covariance :

$$\langle \mathrm{x} - \overline{\mathrm{x}}, \mathrm{y} - \overline{\mathrm{y}} \rangle = \sum_{i=1}^{n} (x_i - \overline{\mathrm{x}})(y_i - \overline{\mathrm{y}}) = n\sigma_{\mathrm{xy}}$$

Lien avec la corrélation :

$$\rho_{\mathrm{xy}} = \frac{\sigma_{\mathrm{xy}}}{\sigma_{\mathrm{x}}\sigma_{\mathrm{y}}} = \frac{\sum_{i=1}^{n} (x_i - \overline{\mathrm{x}})(y_i - \overline{\mathrm{y}})}{n\sigma_{\mathrm{x}}\sigma_{\mathrm{y}}} = \frac{1}{n} \left\langle \frac{x - \overline{\mathrm{x}}}{\sigma_{\mathrm{x}}}, \frac{y - \overline{\mathrm{y}}}{\sigma_{\mathrm{y}}} \right\rangle$$

Un peu de math !

Produit scalaire : rappel

Soit deux vecteurs x, y dans $\mathbb{R}^n$: $\mathrm{x} = \{x_1, \cdots, x_n\}$ et $\mathrm{y} = \{y_1, \cdots, y_n\}$

$$\langle \mathrm{x}, \mathrm{y} \rangle = \sum_{i=1}^{n} x_i y_i = x_1 y_1 + x_2 y_2 + \cdots + x_n y_n$$

Lien avec la covariance :

$$\langle \mathrm{x} - \overline{\mathrm{x}}, \mathrm{y} - \overline{\mathrm{y}} \rangle = \sum_{i=1}^{n} (x_i - \overline{\mathrm{x}})(y_i - \overline{\mathrm{y}}) = n\sigma_{\mathrm{xy}}$$

Lien avec la corrélation :

$$\rho_{\mathrm{xy}} = \frac{\sigma_{\mathrm{xy}}}{\sigma_{\mathrm{x}}\sigma_{\mathrm{y}}} = \frac{\sum_{i=1}^{n} (x_i - \overline{\mathrm{x}})(y_i - \overline{\mathrm{y}})}{n\sigma_{\mathrm{x}}\sigma_{\mathrm{y}}} = \frac{1}{n} \left\langle \frac{x - \overline{\mathrm{x}}}{\sigma_{\mathrm{x}}}, \frac{y - \overline{\mathrm{y}}}{\sigma_{\mathrm{y}}} \right\rangle$$

Approche multidimensionnelle simultanée : calcul matriciel

Un tableau de données peut être vu comme une matrice $X = [x_{ij}]_{i=1,\cdots,n}^{j=1,\cdots,m}$

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ \vdots & & \cdots & \\ x_{n1} & x_{n2} & \cdots & x_{nm} \end{bmatrix}$$

où x_{ij} est la valeur de la variable j de l'individu i.

a) La matrice Y des données centrées s'obtient par :

$$Y = AX\,,$$

où $A = [a_{ij}]$ est la matrice de centrage définie par :

$$a_{ij} = \begin{cases} 1 - \frac{1}{n} & \text{si } i = j \\ -\frac{1}{n} & \text{sinon} \end{cases} \forall i, j \in \{1, \cdots, n\}$$

Approche multidimensionnelle simultanée : calcul matriciel

Un tableau de données peut être vu comme une matrice $X = [x_{ij}]_{i=1,\cdots,n}^{j=1,\cdots,m}$

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ \vdots & & \cdots & \\ x_{n1} & x_{n2} & \cdots & x_{nm} \end{bmatrix}$$

où x_{ij} est la valeur de la variable j de l'individu i.

a) La matrice Y des données centrées s'obtient par :

$$Y = AX\,,$$

où $A = [a_{ij}]$ est la matrice de centrage définie par :

$$a_{ij} = \begin{cases} 1 - \frac{1}{n} & \text{si } i = j \\ -\frac{1}{n} & \text{sinon} \end{cases} \forall i, j \in \{1, \cdots, n\}$$

Approche multidimensionnelle simultanée : calcul matriciel

b) La matrice V des variances-covariances est la matrice $V = [b_{ij}]$ où :

$$b_{ij} = \begin{cases} \sigma^2_{\mathrm{x}^i} & \text{si } i = j \\ \sigma_{\mathrm{x}^i \mathrm{x}^j} & \text{sinon} \end{cases} \forall i, j \in \{1, \cdots, m\}$$

Elle est obtenue de la matrice Y par : $V = \frac{1}{n} Y^t Y$ où Y^t est la transposée de Y, i.e. $Y^t = ([y_{ij}])^t = [y_{ji}]$.

Approche multidimensionnelle simultanée : calcul matriciel

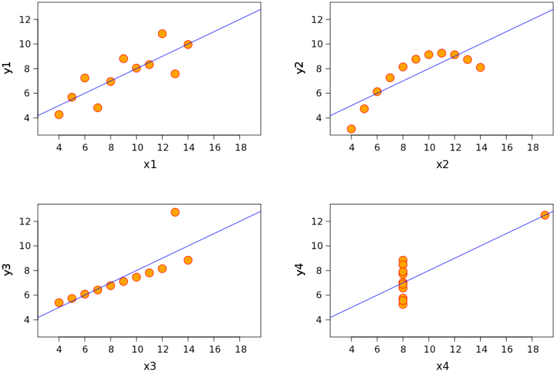
c) La matrice R des corrélations est la matrice $R = [r_{ij}]$ où :

$$r_{ij} = \left\{ \begin{array}{cc} 1 & \text{si } i = j \\ \rho_{\mathrm{x}^i \mathrm{x}^j} & \text{sinon} \end{array} \right. \forall i, j \in \{1, \cdots, m\}$$

R est obtenue de la matrice V par : $V = DVD$ où $D = [d_{ij}]$ est définie comme suit :

$$d_{ij} = \left\{ \begin{array}{cc} \frac{1}{\sigma_{\mathrm{x}^i}} & \text{si } i = j \\ 0 & \text{sinon} \end{array} \right.$$

Conclusion : le danger des stat. descriptives



Toutes les propriétés statistiques de ces populations sont identiques (moyenne, variance, corrélation)