Corrélation et régression linéaires

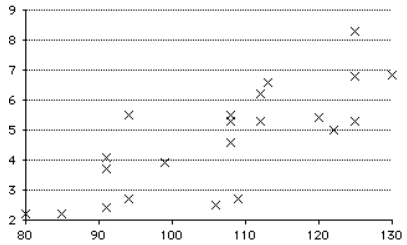
Corrélation linéaire

Situation envisagée : un échantillon de sujets, deux variables numériques observées (ou une variable observée dans deux conditions).

Données :

	X	Y
s_1	x_1	y_1
s_2	x_2	y_2
...	...	...

Nuage de points : points (x_i, y_i)



Covariance des variables X et Y

$$Cov(X,Y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})$$

ou

$$Cov(X,Y) = \left(\frac{1}{n}\sum_{i=1}^{n} x_i y_i\right) - \overline{x}\ \overline{y}$$

Coefficient de corrélation de Bravais Pearson

C'est le quotient de la covariance par le produit des écarts-types.

On désigne par $s(X)$ et $s(Y)$ les écarts types de X et Y.

$$r = \frac{Cov(X,Y)}{s(X)s(Y)}$$

Remarques

- r est un coefficient sans unités, compris entre -1 et 1.
- Les valeurs $r = -1$ et $r = 1$ correspondent à une relation fonctionnelle, déterministe (non statistique) entre X et Y.
- La valeur $r = 0$ correspond à l'*indépendance* des variables X et Y : la connaissance de l'une ne renseigne absolument pas sur les valeurs possibles de l'autre.
- r mesure l'intensité de la relation entre X et Y, *lorsque cette relation est linéaire.* Mais, il existe des relations non linéaires.
- Corrélation n'est pas causalité.

Mini-exemple

	X	Y	X^2	Y^2	XY
s_1	3	7	9	49	21
s_2	4	6	16	36	24
s_3	7	13	49	169	91
s_4	6	12	36	144	72
s_5	5	8	25	64	40
Σ	25	46	135	462	248

Moyenne de X : $\overline{X} = \frac{25}{5} = 5$

Moyenne de Y : $\overline{Y} = \frac{46}{5} = 9.2$

Variance de X : $s^2(X) = \frac{135}{5} - 5^2 = 2$

Ecart type de X : $s(X) = \sqrt{2} = 1.41$

Variance de Y : $s^2(Y) = \frac{462}{5} - 9.2^2 = 7.76$

Ecart type de Y : $s(Y) = \sqrt{7.76} = 2.66$

Covariance de X et Y : $Cov(X,Y) = \frac{248}{5} - 5 \times 9.2 = 3.6$

Coefficient de corrélation de X et Y :

$$r = \frac{3.6}{1.41 \times 2.66} = 0.91$$

Cohérence d'un ensemble d'items : alpha de Cronbach

Dans un questionnaire, un groupe d'items X_1, X_2, ..., X_k (par exemple des scores sur des échelles de Likert) mesure un même aspect du comportement.

Problème : comment mesurer la cohérence de cet ensemble d'items ?

Dans le cas de 2 items : la cohérence est d'autant meilleure que la covariance entre ces items est plus élevée. Le rapport :

$$\alpha = 4\ \frac{Cov(X_1, X_2)}{Var(X_1 + X_2)}$$

– vaut 1 si $X_1 = X_2$
– vaut 0 si X_1 et X_2 sont indépendantes
– est négatif si X_1 et X_2 sont anti-corrélées.

Généralisation :
On introduit $S = X_1 + X_2 + \ldots + X_k$ et on considère le rapport :

$$\alpha = \frac{k}{k-1}\left[1 - \frac{\sum Var(X_i)}{Var(S)}\right]$$

Ce rapport est le coefficient α de Cronbach.

Exemple : On a mesuré 3 items (échelle en 5 points) sur 6 sujets.

	X_1	X_2	X_3	S
s1	1	2	2	5
s2	2	1	2	5
s3	2	3	3	8
s4	3	3	5	11
s5	4	5	4	13
s6	5	4	4	13
Var.	2.167	2.000	1.467	13.77

On obtient : $\alpha = \frac{3}{2}\left[1 - \frac{2.167 + 2 + 1.467}{13.77}\right] = 0.886$

Signification de α : on considère généralement que l'on doit avoir $\alpha \geq 0.7$. Mais une valeur trop proche de 1 revèle une pauvreté dans le choix des items.

Régression linéaire

Rôle "explicatif" de l'une des variables par rapport à l'autre. Les variations de Y peuvent-elles (au moins en partie) être expliquées par celles de X ? Peuvent-elles être prédites par celles de X ?

Modèle permettant d'estimer Y connaissant X

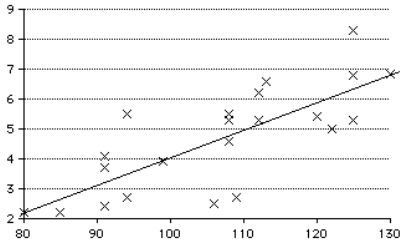
Droite de régression de Y par rapport à X :

La droite de régression de Y par rapport à X a pour équation :

$$y = b_0 + b_1 x$$

avec :

$$b_1 = \frac{Cov(X,Y)}{s^2(X)} \quad ; \quad b_0 = \overline{Y} - b_1 \overline{X}$$



Remarques

Si les variables X et Y sont centrées et réduites, l'équation de la droite de régression est :

$$Y = rX$$

On définit le *coefficient de régression standardisé* par :

$$\beta_1 = b_1 \frac{s(X)}{s(Y)}$$

Dans le cas de la régression linéaire simple : $\beta_1 = r$.

Comparaison des valeurs observées et des valeurs estimées

Valeurs estimées : $\widehat{y}_i = b_0 + b_1 x_i$: variable $\widehat{Y}$
Erreur (ou résidu) : $e_i = y_i - \widehat{y}_i$: variable E

Les variables $\widehat{Y}$ et E sont indépendantes et on montre que :

$$s^2(Y) = s^2(\widehat{Y}) + s^2(E)$$

avec :

$$\frac{s^2(E)}{s^2(Y)} = 1 - r^2 \quad ; \quad \frac{s^2(\widehat{Y})}{s^2(Y)} = r^2$$

$s^2(\widehat{Y})$: variance *expliquée* (par la variation de X, par le modèle)

$s^2(E)$: variance *perdue* ou *résiduelle*

r^2 : part de la variance de Y qui est expliquée par la variance de X. r^2 est appelé *coefficient de détermination*.

Sur notre mini-exemple, les coefficients de la droite de régression sont :

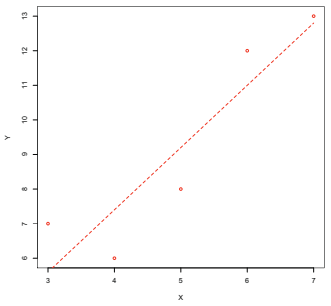
$$b_1 = \frac{3.6}{2} = 1.8 \quad ; \quad b_0 = 9.2 - 1.8 \times 5 = 0.2$$

Equation de la droite de régression : $y = 0.2 + 1.8x$.

Valeurs observées, valeurs prévues et résidus :

	X	Y	$\widehat{Y}$	E
s_1	3	7	5.6	1.4
s_2	4	6	7.4	-1.4
s_3	7	13	12.8	0.2
s_4	6	12	11	1
s_5	5	8	9.2	-1.2
Σ	25	46	46	0

Coefficient de détermination : $r^2 = 0.835$.



Lois théoriques

Lois théoriques discrètes

On se donne un ensemble de modalités (par exemple : 0, 1, . . ., N) et une formule mathématique permettant de calculer leurs fréquences d'apparition.

Pour que la loi ait un intérêt en pratique, il faut qu'elle puisse servir à modéliser des situations du monde réel.

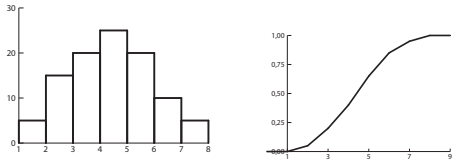
Exemple 1 : Une épreuve élémentaire, avec seulement deux issues possibles (succès, échec) est répétée un nombre déterminé de fois, N, de façon indépendante. Les chances de succès à chaque épreuve sont égales à p ($0 \leq p \leq 1$). On compte le nombre de succès observés sur les N épreuves : loi binomiale de paramètres N et p.

Exemple 2 : Une épreuve élémentaire, avec seulement deux issues possibles (succès, échec) est répétée jusqu'à l'obtention du premier succès. Les chances de succès à chaque épreuve sont égales à p ($0 \leq p \leq 1$). On compte le nombre d'échecs rencontrés avant l'obtention du premier succès : loi géométrique de paramètre p.

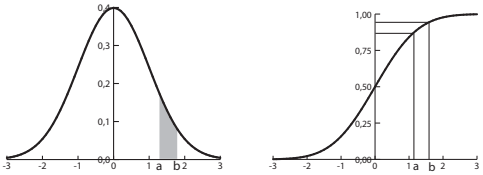
Lois théoriques continues

Du point de vue mathématique, on se donne une fonction f telle que l'aire située sous la courbe $y = f(x)$ soit égale à 1. f est la densité de la loi.

La densité d'une loi théorique est la modélisation mathématique de la notion d'histogramme pour une distribution empirique.



De façon analogue, une loi théorique de distribution statistique est donnée par sa densité $f(x)$ ou sa fonction de répartition : $F(x)$



La fréquence (le pourcentage d'observations) vérifiant $a \leq X \leq b$ est donnée par l'aire hachurée ou par la valeur $F(b) - F(a)$.

Loi Normale ou loi de Laplace Gauss

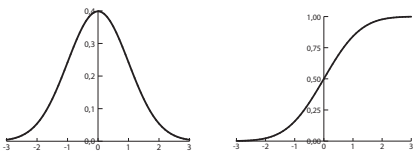
Problème : trouver une loi théorique modélisant la distribution d'une variable dont les valeurs résultent d'une combinaison d'effets *nombreux, indépendants entre eux, additifs et de même ordre de grandeur*.

Réponse : La loi normale.

Loi normale centrée réduite

Moyenne : $\mu = 0$. Ecart type : $\sigma = 1$

Densité : $f(x) = \frac{1}{\sqrt{2\pi}}\, e^{-\frac{x^2}{2}}$.



Loi normale, cas général : transformation en Z

La variable X suit une loi normale de paramètres μ et σ si la variable Z définie par :

$$Z = \frac{X - \mu}{\sigma}$$

suit une loi normale centrée réduite.

"Transformation en Z" ou "centrage réduction" de la variable.

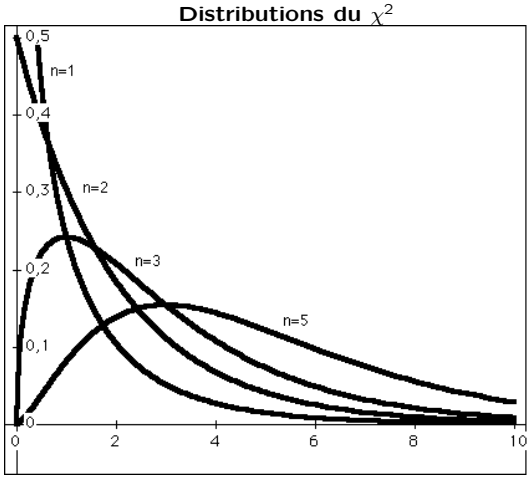
Une loi analogue, permettant une meilleure modélisation de certaines situations : *loi de Student à n degrés de liberté*

Loi du χ^2

Problème : Dans une situation donnée, on considère n variables normales centrées réduites indépendantes. On s'intéresse à la somme des carrés des écarts à la moyenne : $S = X_1^2 + X_2^2 + \ldots + X_n^2$.

La loi suivie par S est la loi du χ^2 (khi-2, chi-2) à n degrés de liberté.

Distributions du χ^2



Une loi analogue, permettant une meilleure modélisation de certaines situations : *loi de Fisher-Snedecor à n_1 et n_2 degrés de liberté*

Notion de test statistique

Comment séparer le probable de l'improbable ?

Exemple intuitif. Hier soir, un match de foot était retransmis à la télévision. Le score final a été : 14 - 1. Qu'en pensez-vous ?

Deux hypothèses :
H_0 : L'affirmation est correcte.
H_1 : L'affirmation est erronée. Par exemple, le score annoncé est incorrect, ou alors, il ne s'agissait pas de football mais de rugby ou de handball...

Difficile de raisonner sur H_1 (trop imprécise, pas d'information suffisante). Raisonnons en supposant H_0 vraie.

Les règles du jeu de football autorisent tout à fait un score tel que 14 - 1. Mais, nous avons une connaissance (intuitive) de la distribution des scores des matches de foot et nous savons qu'un tel score est extrêmement rare (aussi bien que des scores plus extrêmes : 15-0, 15-1, etc).

Entre les deux explications :
– le match a abouti à un score tout à fait exceptionnel ;
– l'affirmation est incorrecte
nous choisissons plutôt la seconde : l'affirmation est incorrecte, car la première est trop improbable.

Mais, pour un score de 4 - 1, nous aurions fait le choix inverse, et pour un score de 6 - 1, nous n'aurions pas trop su comment conclure...

Observer un échantillon

Exemple (idiot) : on veut évaluer la taille moyenne μ de la population française adulte.

Solution 1. On interroge de façon exhaustive le fichier des cartes d'identité. Mais c'est très coûteux, et la CNIL proteste...

Solution 2. On tire au hasard 1 sujet. Il mesure 174 cm. Mais cela ne fournit pas beaucoup d'information. Notamment, on n'a aucune indication sur la variabilité de la grandeur observée : les tailles s'échelonnent-elles de 80 à 250 ou de 170 à 180 ?

Solution 3. On tire au hasard, avec remise, $n = 100$ sujets dans la population française adulte. On peut en tirer de multiples informations :

• La moyenne $\overline{x}$ des tailles observées sur l'échantillon me fournit un *estimateur non biaisé* de la taille moyenne sur la population.

• L'écart type et la variance observés sur l'échantillon fournissent un estimateur biaisé de ces paramètres dans la population. Pour supprimer ce biais :

$$\text{Variance estimée dans la population} = \frac{n}{n-1} \text{ Variance observée sur l'échantillon}$$

• On a une idée de la "précision" de cet estimateur de la taille moyenne. C'est une variable statistique distribuée (presque) normalement, dont la moyenne est la taille moyenne μ et dont la variance est celle de la population divisée par la taille ($n = 100$) de l'échantillon. Autrement dit, cet estimateur est d'autant plus "précis" que la taille de l'échantillon est grande.

Introduction aux tests statistiques

Démarche générale d'un test

35 sujets soumis à un apprentissage. Deux tests l'un avant, l'autre après l'apprentissage.

Sujet	1	2	3	4	5	6	7	...
Avant	8	13	12	17	14	9	10	...
Après	11	11	14	21	12	10	15	...

Problème : L'apprentissage a-t-il un effet sur la performance ?

Remarques :
Raisonner en termes "d'échantillon tiré d'une population"
Variable pertinente : différence individuelle $d_i = y_i - x_i$

Protocole dérivé des différences individuelles

Sujet	1	2	3	4	5	6	7	...
d_i	3	-2	2	4	-2	1	5	...

Caractéristiques de position et de dispersion :
$\overline{d} = 1.08$; $s^2 = 5.05$; $s = 2.25$; $s_c^2 = 5.20$; $s_c = 2.28$

Construction d'un test statistique

Sujets observés : échantillon tiré dans une population
δ : moyenne des effets individuels dans la population.

1. Formulation des hypothèses
H_0 : hypothèse nulle : $\delta = 0$
H_1 : hypothèse alternative : $\delta \neq 0$

2. Choix d'un risque, ou seuil de signification
Par exemple : $\alpha = 5\%$

3. Choix d'une statistique de test
Une statistique est une variable qui peut être évaluée sur chaque échantillon tiré, et dont la distribution théorique, sous l'hypothèse H_0, est connue.

Ici, on prend : $Z = \dfrac{\overline{d}}{E}$ avec $E^2 = \dfrac{s_c^2}{n}$.

Les statisticiens ont montré que, sous l'hypothèse H_0, Z suit approximativement une loi normale centrée réduite.

4. Calcul des valeurs critiques (règle de décision)
Pour $\alpha = .05$, on obtient $z_{crit} = 1.96$.

5. Calcul de la valeur observée de la statistique
Ici : $z_{obs} = \frac{1.08}{0.38} = 2.84$

6. Comparer z_{obs} et z_{crit}. Appliquer la règle de décision
Ici : $z_{obs} > z_{crit}$. z_{obs} est dans la zone de rejet de H_0.
Sous H_0, l'échantillon tiré a une fréquence d'apparition inférieure à 5%. On refuse donc H_0 et on choisit H_1.

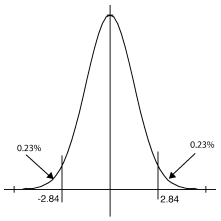
Raisonner en termes de "niveau de significativité"

Avec un logiciel de traitement statistique, les étapes 4, 5 et 6 sont remplacées par :

4'. Calcul de la valeur observée de la statistique
Comme ci-dessus : $z_{obs} = 2.84$

5'. Calcul de la *p-value* correspondante
On évalue, sous l'hypothèse H_0, la fréquence (ou probabilité) d'apparition de tous les protocoles *au moins aussi extrêmes que celui observé.*
Ici : $p = P(Z \leq -2.84) + P(Z \geq 2.84) = 1 - 2 \times 0.4977 = 0.0046 = 0.46\%$.

Autre formulation : si H_0 est vraie, on a seulement 0.46% de chances de tirer un échantillon conduisant à $Z \leq 2.84$ ou $Z \geq 2.84$.



6'. Comparaison du seuil et de la p-value ; conclusion
Ici : $p = 0.46\%$ et $\alpha = 5\%$. D'où $p < \alpha$. Au seuil de 5%, on refuse donc H_0 et on choisit H_1.

Remarques générales

Test : mécanisme permettant de trancher entre deux hypothèses à partir des résultats observés sur un ou plusieurs échantillons.

Hypothèses
Hypothèse nulle : elle joue un rôle particulier ; elle affirme que les différences observées sont dues au hasard.
Hypothèse alternative : elle affirme que les différences sont significatives (en un sens à préciser).

Les risques d'erreur

		Hypothèse vraie	
		H_0	H_1
Hypothèse retenue	H_0	$1-\alpha$	β
	H_1	α	$1-\beta$

• α : seuil de significativité. C'est aussi la probabilité de rejeter H_0 alors que H_0 est vraie (risque de première espèce ou risque de commettre une erreur de type I)

• β : risque de seconde espèce. C'est la probabilité d'accepter H_0 alors que H_0 est fausse (risque de commettre une erreur de type II).

$1-\beta$: probabilité de détecter correctement un cas où H_0 doit être rejetée. Puissance du test.

Illustrations Commettre une ...

Erreur de type I : c'est voir une différence entre deux groupes alors qu'en fait, il n'y en a pas.
Exemples :
– Affirmer qu'un programme d'apprentissage coûteux a un effet sur le comportement des sujets, alors que c'est inexact
– "Mettre en évidence" une différence imaginaire entre les sexes, ou les races...
Comment diminuer ce risque : prendre α petit, veiller à neutraliser les autres variables, etc

Erreur de type II : c'est ne pas voir de différence, alors qu'il y en a réellement une.
C'est souvent un moindre mal, mais...
Exemples :
– Ne pas mettre en évidence un effet secondaire d'un médicament.
– L'usine de La Hague est-elle réellement inoffensive pour les riverains ?

Comment diminuer ce risque : augmenter la taille de l'échantillon, ne pas prendre α trop petit, veiller à neutraliser les autres variables, bien choisir le test...