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 :

$$\alpha = \frac{k}{k-1}\,\frac{2\sum_{i<j} Cov(X_i, X_j)}{Var(S)} = \frac{k}{k-1}\left[1 - \frac{\sum Var(X_i)}{Var(S)}\right]$$

Ce rapport est le coefficient α de Cronbach.

Justification théorique : Pour chaque item, et pour leur somme :

$$\begin{array}{c}\text{Valeur}\\ \text{mesurée}\end{array} = \begin{array}{c}\text{“Vraie valeur” de}\\ \text{l'item sur le sujet}\end{array} + \begin{array}{c}\text{erreur}\\ \text{aléatoire}\end{array}$$

La fiabilité (théorique) est :

$$\rho = \frac{\text{Variance(Vraies valeurs)}}{\text{Variance(Valeurs mesurées)}}$$

α est une estimation de la fiabilité de la somme des k items.

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.

Significativité du coefficient de corrélation

- Les données (x_i, y_i) constituent un échantillon
- r est une statistique
- ρ : coefficient de corrélation sur la population

H_0 : Indépendance sur la population ; $\rho = 0$
H_1 : $\rho \neq 0$ (bilatéral) ou $\rho > 0$ ou $\rho < 0$ (unilatéral)

Statistique de test

– Petits échantillons : tables spécifiques. $ddl = n - 2$
– Grands échantillons :

$$T = \sqrt{n-2}\,\frac{r}{\sqrt{1-r^2}}$$

T suit une loi de Student à $n - 2$ degrés de liberté.

Conditions d'application

Dans la population parente, le couple (X, Y) suit une *loi normale bivariée*, ce qui implique notamment :

– la normalité des distributions marginales de X et Y ;

– la normalité de la distribution de l'une des variables lorsque l'autre variable est fixée ;

– l'égalité des variances des distributions de l'une des variables pour deux valeurs distinctes de l'autre variable.

Corrélation et statistiques non paramétriques : corrélation des rangs de Spearman

Si les données ne vérifient pas les conditions d'application précédentes, ou si les données observées sont elles-mêmes des classements, on pourra travailler sur les protocoles des rangs définis séparément pour chacune des deux variables.

	Rangs X	Rangs Y
s_1	r_1	r'_1
s_2	r_2	r'_2
...	...	...

Le coefficient de corrélation des deux protocoles de rangs est appelé coefficient de corrélation de Spearman, et noté R_s.

Calcul de R_s (en l'absence d'ex aequo)

On calcule les différences individuelles $d_i = r_i - r'_i$, puis

$$R_s = 1 - \frac{6\sum d_i^2}{N(N^2-1)}$$

Significativité du coefficient de corrélation de Spearman

– Pour $N \leq 30$, on utilise en général des tables spécialisées.

– Lorsque $N > 30$:

Certains auteurs (et Statistica) utilisent :

$$t = \sqrt{N-2}\ \frac{R_s}{\sqrt{1-R_s^2}}$$

et une loi de Student à $N-2$ ddl.

D'autres auteurs utilisent la statistique :

$$Z = \sqrt{N-1}\ R_s$$

et une loi normale centrée réduite.

Remarque. Siegel et Castellan donnent "20 ou 25" comme seuil pour les grands échantillons et indiquent que la première statistique est "légèrement meilleure" que la seconde.

Le coefficient τ de Kendall

Avec un protocole comportant N sujets, on a $\frac{N(N-1)}{2}$ paires de sujets.

On examine chaque paire de sujets, et on note si les deux classements comportent une inversion ou non.

s_3	3	6
s_4	5	2

Désaccord, Inversion

s_3	3	4
s_4	5	6

Accord, Pas d'inversion

Le coefficient τ est alors défini par :

$$\tau = \frac{\text{Nb d'accords} - \text{Nb de désaccords}}{\text{Nb de paires}}$$

ou

$$\tau = 1 - \frac{2 \times \text{Nombre d'inversions}}{\text{Nombre de paires}}$$

Significativité du τ de Kendall

H_0 : Indépendance des variables. $\tau = 0$
H_1 : $\tau \neq 0$

Pour $N > 10$, sous H_0, la statistique

$$Z = 3\tau\sqrt{\frac{N(N-1)}{2(2N+5)}}$$

suit une loi normale centrée réduite.

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

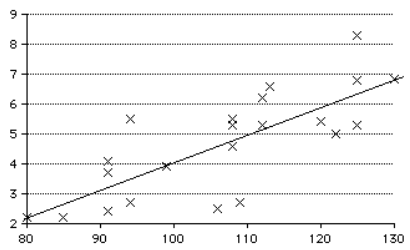
Droite de régression de Y par rapport à X :

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*.

Exemple : $r = 0.86$

$r^2 = 0.75$; $1 - r^2 = 0.25$; $\sqrt{1 - r^2} = 0.5$.

• La part de la variance de Y expliquée par la variation de X est de 75%.

• L'écart type des résidus est la moitié de l'écart type de Y.

Interprétation géométrique, du point de vue des variables

Lorsque les variables X et Y sont centrées et réduites, elles peuvent être vues comme des vecteurs OM et ON de norme 1 dans un espace géométrique de dimension n (le nombre d'observations).

Dans cette interprétation :

– La variable $\hat{Y}$ est la projection de Y sur X

– Le coefficient de corrélation r est le cosinus de l'angle (OM, ON).

Test du coefficient de corrélation à l'aide du F de Fisher

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

On introduit les sommes de carrés suivantes :

$$SC_{\text{Totale}} = \sum (y_i - \overline{y})^2$$

$$SC_{\text{Régression}} = \sum (\widehat{y}_i - \overline{y})^2$$

$$SC_{\text{Résidus}} = \sum (y_i - \widehat{y}_i)^2$$

Lien avec le coefficient de corrélation

$r^2 = \dfrac{SC_{\text{Régression}}}{SC_{\text{Totale}}}$ est le *coefficient de détermination*

Tableau d'analyse de variance

Source	SC	ddl	CM	F
Régression	$SC_{\text{Régression}}$	1	CM_{Reg}	F_{obs}
Résiduelle	$SC_{\text{Résidus}}$	$n-2$	CM_{Res}	
Total	SC_{Totale}	$n-1$		

$F_{obs} = \dfrac{CM_{\text{Reg}}}{CM_{\text{Res}}} = (n-2)\dfrac{r^2}{1-r^2}$ suit une loi de Fisher à 1 et $n-2$ ddl.
On retrouve : $F_{obs} = T_{obs}^2$

Estimations de $\hat{y}$ et de y par des intervalles de confiance

Dans la population, le lien entre X et Y suit le modèle mathématique :

$$Y = B_0 + B_1 X + \epsilon$$

où B_0 et B_1 sont des constantes numériques et ϵ est une variable statistique centrée et indépendante de X

A partir de l'échantillon, nous avons calculé $b_0, b_1, e_1, \ldots, e_n$ tels que :

$$y_i = b_0 + b_1 x_i + e_i$$

$$\widehat{y}_i = b_0 + b_1 x_i$$

Mais un autre échantillon amènerait d'autres valeurs de ces paramètres : b_0, b_1 et e_i ne sont que des estimations de B_0, B_1 et ϵ_i.

Questions que l'on peut se poser :

– Quelle estimation peut-on donner de la variance de ϵ ?

– On peut voir $\widehat{y}_i$ comme une estimation ponctuelle de la moyenne des valeurs de Y sur la population lorsque $X = x_i$. Peut-on déterminer un intervalle de confiance pour cette moyenne ?

– Etant donné une valeur x_i de X, quel intervalle de confiance peut-on donner pour les valeurs de Y correspondantes ?

Estimation de $Var(\epsilon)$:

$$s^2 = \frac{SC_{Res}}{n-2}$$

Estimation par un intervalle de confiance de la moyenne de Y pour X fixé :

Pour une valeur x_p fixée de X, la variance des valeurs estimées $\widehat{y}_p$ est estimée par :

$$s^2_{\widehat{y}_p} = s^2 \left(\frac{1}{n} + \frac{(x_p - \overline{x})^2}{\sum x_i^2 - n\overline{x}^2} \right)$$

Notons $Moy(Y|X = x_p)$ la moyenne de la variable Y lorsque la variable X est égale à x_p.

Un intervalle de confiance de $Moy(Y|X = x_p)$ avec un degré de confiance $1 - \alpha$ est donné par :

$$\widehat{y}_p - t_\alpha s_{\widehat{y}_p} \leq Moy(Y|X = x_p) \leq \widehat{y}_p + t_\alpha s_{\widehat{y}_p}$$

où t_α est la valeur du T de Student à $n-2$ ddl telle que $P(|T| > t_\alpha) = \alpha$

Détermination d'un intervalle de confiance pour les valeurs de Y : intervalle de prévision

$Moy(Y|X = x_p)$ est connue par une estimation ponctuelle ($\widehat{y}_p$) et un intervalle de confiance.

La différence $Y - Moy(Y|X = x_p)$ est le résidu ϵ, dont on peut également donner un intervalle de confiance.

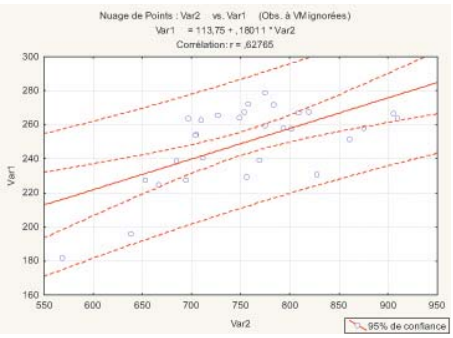
Finalement, on pourra écrire l'intervalle de confiance :

$$\widehat{y}_p - t_\alpha s_{ind} \leq y \leq \widehat{y}_p + t_\alpha s_{ind}$$

avec :

$$s^2_{ind} = s^2 \left(1 + \frac{1}{n} + \frac{(x_p - \overline{x})^2}{\sum x_i^2 - n\overline{x}^2}\right) = s^2 + s^2_{\widehat{y}_p}$$

Intervalle de confiance et intervalle de prévision apparaissent dans les graphiques réalisés par les logiciels sous forme de "bandes" :



Régression linéaire multiple

Position du problème

Un échantillon tiré d'une population sur lequel on a observé un ensemble de variables numériques.

	X_1	X_2	...	X_p
s_1	x_{11}	x_{12}	...	x_{1p}
...	...	...	...	...

Exemple avec trois variables
Satis : satisfaction au travail
Anc : ancienneté dans l'entreprise
Resp : responsabilités
n : nombre d'observations (ici : $n = 20$)

	Satis	Anc	Resp
1	3,95	7,44	2,23
2	2,11	1,29	0,57
3	2,50	4,85	1,12
4	6,05	6,00	3,49
5	3,78	0,68	0,60
6	6,15	6,81	3,74
7	2,10	4,15	1,68
8	6,80	1,77	2,34
9	5,99	5,78	2,75
10	2,29	5,75	2,80
11	3,53	3,53	2,08
12	4,55	5,73	1,52
13	1,14	4,80	0,73
14	4,29	10,66	2,99
15	4,86	5,27	2,46
16	4,25	4,17	2,62
17	4,34	5,80	1,88
18	2,77	2,31	1,24
19	4,82	7,68	2,00
20	3,74	5,53	1,19

Nuage de points

Pour trois variables : représentation dans l'espace.

Pour plus de trois variables, détermination des directions de "plus grande dispersion du nuage" : analyse en composantes principales.

Paramètres associés aux données

Matrice des covariances, matrice des corrélations.

Sur l'exemple : coefficients de corrélation des variables prises 2 à 2 :

$$\begin{array}{c} \\ Satis \\ Anc \\ Resp \end{array} \begin{array}{c} \begin{array}{ccc} Satis & Anc & Resp \end{array} \\ \begin{bmatrix} 1 & 0.23 & 0.67 \\ 0.23 & 1 & 0.57 \\ 0.67 & 0.57 & 1 \end{bmatrix} \end{array}$$

$r_{Satis,Anc} = 0.23$ (NS) :
Satis et Anc ne sont pas significativement corrélées
$r_{Satis,Resp} = 0.67$ $**$:
Satis et Resp sont corrélées
$r_{Resp,Anc} = 0.57$ $**$:
Resp et Anc sont corrélées

Régression : équation de régression, "hyperplan" de régression

L'une des variables (Y) est la variable "à prédire". Les autres ($X_1, X_2, \ldots, X_p$) sont les variables "prédicatives". L'hyperplan de régression a pour équation :

$$Y = b_0 + b_1 X_1 + \ldots + b_p X_p$$

La régression est faite "au sens des moindres carrés" (ordinary least squares regression ou OLS regression ; moindres carrés ordinaires ou MCO).

Soient $\widehat{Y}_i$ les valeurs estimées à l'aide de l'équation de régression et $E_i = Y_i - \widehat{Y}_i$ les erreurs ou résidus de la régression.

La variable d'erreur E est indépendante de chacun des prédicteurs.

Calcul de b_1, b_2, ..., b_p : ce sont les solutions du système d'équations linéaires :

$$j = 1, 2, \ldots, p \quad \sum_i \text{Cov}(X_i, X_j)\ b_i = \text{Cov}(Y, X_j)$$

Sur l'exemple, en prenant Satis comme variable dépendante à prédire, Anc et Resp comme variables prédictrices :

$$Satis = 2.07 - 0.15\ Anc + 1.33\ Resp$$

Coefficient de corrélation multiple

$\hat{Y}$: valeurs estimées à l'aide de l'équation de régression.

$$R = r_{Y\hat{Y}} = \frac{Cov(Y, \hat{Y})}{s(Y)s(\hat{Y})}$$

Comme précédemment, R^2 est la part de la variance "expliquée par le modèle". R^2 est appelé coefficient de détermination.

Sur l'exemple, $R = 0.70$ et $R^2 = 0.48$.

Analyse de variance et qualité du modèle

Lorsque l'on a p prédicteurs, le tableau d'analyse de variance devient :

Source	SC	ddl	CM	F
Régression	$SC_{\text{Régression}}$	p	CM_{Reg}	F_{obs}
Résiduelle	$SC_{\text{Résidus}}$	$n-p-1$	CM_{Res}	
Total	SC_{Totale}	$n-1$		

$F_{obs} = \frac{CM_{\text{Reg}}}{CM_{\text{Res}}} = \frac{n-p-1}{p} \frac{R^2}{1-R^2}$ suit une loi de Fisher à p et $n-p-1$ ddl.

Sur l'exemple, le tableau d'analyse de variance est ainsi :

Source	SC	ddl	CM	F
Régression	21.74	2	10.87	7.94
Résiduelle	23.26	17	1.37	
Total	45.00	19		

Conclusion : le coefficient de corrélation multiple est significativement différent de 0 ; il existe un effet significatif des prédicteurs sur la variable à prédire.

Analyse individuelle de chaque coefficient.

La constante dépend seulement des moyennes de variables. Elle serait nulle pour des données centrées.

Les coefficients des deux variables prédictrices dépendent leurs écarts types respectifs. Ces coefficients ne sont pas directement comparables entre eux. Seuls les signes peuvent nous fournir une information.

Ainsi, selon le modèle obtenu sur l'exemple :

– Lorsque Resp augmente, Satis augmente.
– En revanche, lorsque Anc augmente, Satis diminue, bien que le coefficient de corrélation trouvé entre ces deux variables soit positif.

Mais cette analyse suppose que l'on puisse considérer l'effet de chaque variable "toutes choses égales par ailleurs". Or, Anc et Resp sont corrélées.

Coefficients de régression standardisés :

Pour $i = 1, 2, \ldots, p$, on pose : $\beta_i = b_i \frac{s(X_i)}{s(Y)}$

Ce sont les coefficients que l'on obtiendrait en travaillant sur les variables centrées réduites associées aux variables Y et X_1 à X_p.

Contrairement aux b_i, les coefficients β_i sont comparables entre eux.

Sur l'exemple : $\beta_{Anc} = -0.22$; $\beta_{Resp} = 0.80$

Les logiciels fournissent également le résultat du test de l'hypothèse de nullité de chacun de ces coefficients sur la population parente.
Ici, seul β_{Resp} est significativement différent de 0.

Coefficients de corrélation partielle

Ce sont les corrélations obtenues en contrôlant une partie des variables. Par exemple, dans le cas de 3 variables X, Y et Z, pour calculer $r_{yz.x}$:
– On calcule les résidus de la régression de Z par rapport à X
– On calcule les résidus de la régression de Y par rapport à X
– On calcule le coefficient de corrélation entre les deux séries de résidus obtenues.

Sur l'exemple :
$r_{Satis\ \ Resp.Anc} = 0.67\ **$:
$r_{Satis\ \ Anc,Resp} = -0.25\ NS$

Une application de la régression linéaire multiple : l'analyse de médiation

L'effet d'une VI sur une VD est-il direct, ou relève-t-il plutôt d'un facteur intermédiaire M ?

– On effectue la régression linéaire de la VD sur la VI :

$$VD = b_0 + b_1 VI$$

Coefficient de régression standardisé : β_1

– On effectue la régression linéaire de la variable de médiation M sur la VI :

$$M = b'_0 + b'_1 VI$$

Coefficient de régression standardisé : β'_1

– On effectue la régression linéaire multiple de la VD sur les deux variables M et VI :

$$VD = b''_0 + b''_1 VI + b''_2 M$$

Coefficients de régression standardisés : β''_1, β''_2

Si β''_2 est significativement différent de 0, et que β''_1 est nettement plus proche de 0 que β_1, il y a médiation (partielle ou totale). En particulier, il y a médiation si β''_1 n'est pas significativement différent de 0 alors que β_1 l'était.

Synthèse sur les tests étudiés

- **Un groupe de sujets - une variable**
  - Ajustement à une loi théorique
    - χ^2 d'ajustement
    - Test de Kolmogorov Smirnov à 1 échantillon
    - Test de Lilliefors (loi normale)
    - Autres tests de normalité : Shapiro-Wilk, Anderson-Darling, D'Agostino-Pearson
  - Comparaison d'une moyenne à une norme
    - Variable numérique : comparaison d'une moyenne à une norme
      Test t de Student sur un échantillon
    - Variable dichotomique : comparaison d'une proportion à une norme
      Test Z ou χ^2

- **Un groupe de sujets - plusieurs VD ou plusieurs conditions (groupes appariés)**
  - Une seule VD. Etude des différences entre conditions
    - VD dichotomique
      - 2 conditions
        χ^2 de Mac Nemar, test des signes
      - k conditions ($k \geq 2$)
        Test Q de Cochran
    - VD numérique quelconque
      - 2 conditions
        Test des signes, test de Wilcoxon
      - k conditions ($k \geq 2$)
        Test de Friedman

    - VD numérique et normalité des différences individuelles
      - 2 conditions
        Test t de Student pour groupes appariés
      - k conditions ($k \geq 2$)
        ANOVA - plan à mesures répétées ($S * A$).
  - Deux ou plusieurs VD. Etude des liens entre les variables
    - VD numériques quelconques
      Corrélation des rangs de Spearman, τ de Kendall
    - VD satisfaisant certaines conditions de normalité
      Corrélation de Bravais-Pearson (MCO ou OLS). régression linéaire

- **Plusieurs groupes de sujets, groupes indépendants, 1 facteur**
  - VD nominale
    Test du χ^2 sur un tableau de contingence
  - VD dichotomique
    - 2 groupes
      Test de comparaison de deux proportions, test du χ^2
    - k groupes ($k \geq 2$)
      Test du χ^2
  - VD numérique quelconque
    - 2 groupes
      Test de la médiane, test de Wald-Wolfowitz, test de Kolmogorov-Smirnov à deux échantillons, test de Mann-Whitney
    - k groupes ($k \geq 2$)
      Test de la médiane, test de Kruskal-Wallis

  - VD normale dans chacun des groupes
    - Comparaison des moyennes
      - 2 groupes
        Test t de Student (échantillons indépendants)
      - k groupes ($k \geq 2$)
        ANOVA à 1 facteur
    - Comparaison des variances
      - 2 groupes
        Test F de Fisher, test de Levene, test de Brown et Forsythe
      - k groupes ($k \geq 2$)
        Test de Levene, test de Brown et Forsythe, test de Bartlett

- **Plusieurs groupes de sujets, groupes indépendants, plusieurs facteurs**
  - Plan $S < A * B >$
    ANOVA factorielle
  - Plan $S < A > *B$
    ANOVA - plan à mesures partiellement répétées.