Psy1004 – Section 12:

Corrélations et régressions

Plan du cours:

Disponible sur: http://mapageweb.umontreal.ca/cousined/home/course/PSY1004

Varia

- Il y aura au prochain cours (le dernier) une période de révision d'environ une heure.
- Rappel: Faites un graphique des moyennes!!!
- TP3:
  - Moyenne:
  - Écart type:
  - Asymétrie:

Varia

- Test d'homogénéité des variances et le test *t*
  - Le test-t tel qu'on l'a vu en classe suppose que les variances sont homogènes (et donc "poolée" ensemble).
  - Cependant, ce n'est pas forcément le cas. SPSS fait le test de deux façons différentes, précédé d'un test d'homogénéité des variances (le test de Levene dans ce cas-ci):

Independent Samples Test

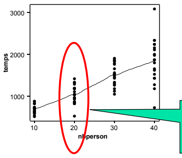
		Levene's Test for Equality of Variances		t-test for Equality of Means						
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference Lower	95% Confidence Interval of the Difference Upper
VAR00001	Equal variances assumed	4.545	.041	-1.022	32	.314	-14068.24	13762.059	-42100.6	13964.16
	Equal variances not assumed			-1.022	16.000	.322	-14068.24	13762.059	-43242.5	15106.03

Si < α, prendre la ligne du haut, sinon, utiliser la ligne du bas.

Méthode vue en classe

Plan expérimental vs. Plan corrélationnel (1/4)

- Jusqu'à présent, nous avons vue des méthodes statistiques pour des « plans expérimentaux », i.e. des données collectés dans des conditions déterminées.
  - Ex: la dose d'alcool est déterminée par le chercheur,
  - de même que:
  - Le type de véhicule,
  - le nombre de personnages dans la tâche « Où est Charlie? »
  - Etc.



Comme le nombre de personnage est manipulé par le chercheur, il n'y a pas de niveaux intermédiaires non désirés, seul les niveaux voulus sont présents.

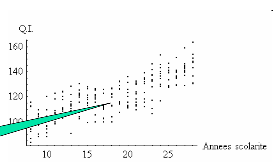
Condition manipulée par le chercheur

Plan expérimental vs. Plan corrélationnel (2/4)

- Les plans expérimentaux sont indiqués par des raccourcis:
  - Plan 2 à deux groupes indépendants
  - Plan (4) à mesures répétées
  - Plan 2 x 3 à six groupes indépendants
  - Plan 2 x (5) à deux groupes indépendants
  - etc.
- Et des méthodes statistiques existent en conséquence:
  - Test t à groupes indépendant
  - ANOVA à mesures répétées
  - ANOVA factorielle
  - ANOVA mixte
  - etc.

Plan expérimental vs. Plan corrélationnel (3/4)

- Il existe des situations où le chercheur ne peut pas manipuler les variables dépendantes comme il veut
  - Ex. Le revenu des parents sur la réussite scolaire: le chercheur ne peut pas décider que tels parents gagnent 50 000$ par an et tels autres 25 000$. Le chercheur ne peux pas répartir les sujets au hasard dans l'un ou l'autre groupe…
- Dans ces cas, on parle d'études corrélationnelles, de plan corrélationnel.
  - Très fréquent en psychologie sociale



Tout les cas possibles existent.

Plan expérimental vs. Plan corrélationnel (4/4)

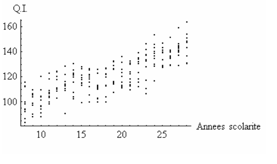
- Dans les études corrélationnelles, on veut savoir si un facteur est associé avec un autre facteur (i.e. s'ils corrèlent; voir plus loin).
- LA méthode statistique pour examiner les associations est la régression/corrélation.
- La régression/corrélation est aussi utilisée pour examiner des données qui proviennent de plans expérimentaux: Une technique très polyvalente.
- Dans ce cours, on n'examine que la régression linéaire (i.e. dont l'association forme une ligne droite).

La corrélation

Disponible sur: http://mapageweb.umontreal.ca/cousined/home/course/PSY1004

12.0: Le Q.I. varie-t-il avec le nombre d'années de scolarité?

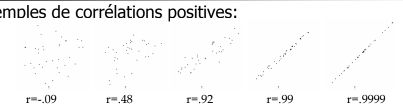
- Exemple 1: le Q.I. en fonction du nombre d'années de scolarité (AS).
- Est-ce que:
  - Le Q.I. est relié/associé/ corrélé avec AS?
  - Si oui, est-il <u>fortement</u> corrélé? dans quel sens?
  - S'il est corrélé, peut-on dire que les années de scolarité accroissent le Q.I?



- Il nous faut une statistique qui permet d'évaluer l'association entre deux mesures.
  - Un indice de corrélation, qui est appelé *r*, ou indice de corrélation de Pearson.
  - L'indice doit indiquer la force d'une association, ou l'absence d'une association, et sa direction (positivement corrélé, négativement corrélé).

12.1: Calculs de l'indice de corrélation (1/2)

Exemples de corrélations positives:



- Comment calcule-t-on un indice de corrélation?
  - Il faut deux ensembles de mesures, disons **X** et **Y**.
  - a) Pour l'échantillon X, il faut connaître l'écart type ${}_{n-1}\overleftrightarrow{\mathbf{X}}$
  - b) Pour l'échantillon Y, il faut connaître l'écart type ${}_{n-1}\overleftrightarrow{\mathbf{Y}}$
  - c) Pour les deux échantillons, il faut connaître la covariance:

$$ {}_{n-1}\overleftrightarrow{\mathbf{XY}}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{X}_i - \overline{\mathbf{X}})(\mathbf{Y}_i - \overline{\mathbf{Y}}) $$

  - d) Finalement, on obtient r avec la formule:

$$ r_{\mathbf{XY}} = \frac{{}_{n-1}\overleftrightarrow{\mathbf{XY}}^2}{{}_{n-1}\overleftrightarrow{\mathbf{X}} \times {}_{n-1}\overleftrightarrow{\mathbf{Y}}} $$

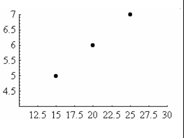
12.1: Calculs de l'indice de corrélation (2/2)

- $r_{\mathbf{XY}}$ est toujours une valeur entre –1 et +1:
  - $r_{\mathbf{XY}}$ = 0 signifie aucune corrélation entre les **X** et les **Y**
  - $r_{\mathbf{XY}}$ = +1 signifie une corrélation parfaite et positive.
  - $r_{\mathbf{XY}}$ = –1 signifie une corrélation parfaite et négative.
- Mis au carré, $r^2_{\mathbf{XY}}$ s'appelle souvent "la variance expliquée" car la proportion de variabilité dans les **X** qui entraîne aussi une variabilité dans les **Y** est quantifiée par le $r^2_{\mathbf{XY}}$.
  - Le $r^2_{\mathbf{XY}}$ varie entre 0 et +1, soit encore entre 0 et 100%

12.2: Exemples de corrélations (1/3)

a) Un petit exemple à la main. Soit les données:

	Échantillon		Calculs		
(n = 3)	X	Y	(X - X̄)	(Y - Ȳ)	(X - X̄) (Y - Ȳ)
	15	5	-5.00	-1.00	5.00
	20	6	0.00	0.00	0.00
	25	7	5.00	1.00	5.00
Moyenne	*20.0*	*6.0*			
Écart type	*5.0*	*1.0*			
Somme					*10.00*
Divisé par n - 1, covariance					*5.00*



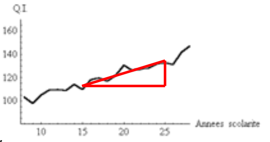
La corrélation est parfaite et positive; 100% de la variance d'un échantillon s'explique par la variance de l'autre (i.e. quand l'un augmente, l'autre aussi).

$$ {}_{n-1}\overleftrightarrow{\mathbf{XY}}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{X}_i - \overline{\mathbf{X}})(\mathbf{Y}_i - \overline{\mathbf{Y}}) $$

$$ r_{\mathbf{XY}} = \frac{{}_{n-1}\overleftrightarrow{\mathbf{XY}}^2}{{}_{n-1}\overleftrightarrow{\mathbf{X}} \times {}_{n-1}\overleftrightarrow{\mathbf{Y}}} = \frac{5.00}{5.00 \times 1.00} = 1 $$

12.2: Exemples de corrélations (2/3)

b) Années de scolarité (AS) vs. Q.I.
Il semble y avoir une relation assez forte entre ces deux variables.



Deux questions sont possibles:

1) Y a-t-il un effet des AS sur le Q.I.?
   Se serait donc une ANVOVA 21 à 21 groupes indépendants

   ou

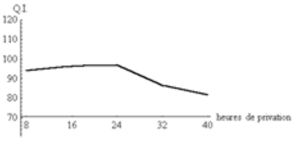
2) Y a-t-il une association significative entre les AS et le Q.I.?
   Se serait donc une corrélation *r* qui doit être significativement différent de zéro.

1) et 2) sont équivalents, sauf qu'avec 2) on peut poser une question supplémentaire: "De combien augmente le Q.I lorsque le nombre d'AS augmente de 10?"

12.2: Exemples de corrélations (3/3)

c) Soit l'effet de la privation de sommeil (PS) sur le Q.I.

Il existe peut-être une relation, mais si présente, elle est:



- beaucoup plus faible: Le *r* ici est sans doute plus proche de .20 à .40 (soit un indice relativement faible).
- et négative: En effet, quand la PS augmente, le Q.I. semble baisser.

Est-ce que cette corrélation est significative? i.e. existe-t-il bien une association entre PS et Q.I.? Encore une fois, on peut soit faire une ANOVA (5) à mesures répétées ou encore tester si le *r* est significativement différent de zéro.

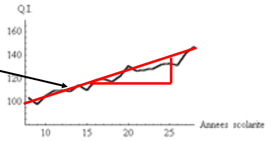
La régression

Disponible sur: http://mapageweb.umontreal.ca/cousined/home/course/PSY1004

12.3: Calculs de la pente d'une droite de régression (1/2)

La droite idéale qui passe le mieux par tous les points s'appelle une "droite de régression".

Cela suppose que les données se répartissent linéairement.



Une information importante de la droite de régression est sa pente, que l'on note souvent par $b_{AS,\ QI}$.

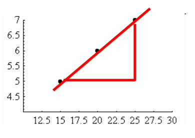
Il existe deux façons simples de calculer la pente de la droite de régression:

- a) à la main: $b_{XY} = \dfrac{\overleftrightarrow{XY}^2}{\overleftrightarrow{X}^2}$
- b) dans le listing de SPSS

12.3: Calculs de la pente d'une droite de régression (2/2)

Dans l'exemple simple vu précédemment:

- L'écart type des **X** était de 5.00, soit une variance de 25.00,
- La covariance était de 5.00,
- Donc, $b_{XY} = \dfrac{\overleftrightarrow{XY}^2}{\overleftrightarrow{X}^2} = \dfrac{5}{25} = 1/5 = 0.2$



Autrement dit, à chaque fois que:

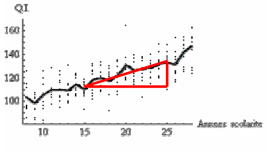
- **X** augmente de 1 → **Y** augmente de 0.2;
- **X** augmente de 5 → **Y** augmente de 1;
- **X** augmente de 10 → **Y** augmente de 2; etc.

Si on prolongeait la droite jusqu'à **X** = 0, nous aurions que **Y** vaut 2 (l'ordonnée à l'origine).

L'équation complète de la droite de régression est donc **Y** = 0.2 **X** + 2.

12.4: Tester l'indice de corrélation: Les données

Examinons plus en détail l'exemple des AS sur le Q.I.



Si le chercheur veut uniquement savoir s'il existe un effet des AS sur le Q.I., il va procéder à une ANOVA (voir acétate suivante); s'il veut savoir s'il existe une association et quantifier cette association, il va procéder à une analyse de corrélation (voir acétates suivantes).

12.4: Tester l'indice de corrélation a) l'ANOVA

- La démarche:
  - a) hypothèse: H_0: Pas d'effet des AS; H_1: Effet des AS
  - b) α = 5%.
  - c) Test: ANOVA 21 à 21 groupes indépendants (Voir syntaxe ➔ Détails).
  - d) Calculs et résultats:

	SC	dl	CM	F
AS (A)	30831.5	20	1541.6	19.3
Erreur (S|A)	15102.3	189	79.9	
Total	45933.7	209		

Combien y a-t-il de sujets par groupe? de sujet au total?

- Interprétation:
  "Le nombre d'années de scolarité influe significativement sur le Q.I. ($\underline{F}$(20,189)=19.29, $\underline{p}$ < .05). Le Q.I. moyen pour les personnes ayant 9 années de scolarité est de 99.4 alors que pour les personnes ayant 28 années de scolarité, il est de 140.6 en moyenne."

12.4: Tester l'indice de corrélation -OU- b) la corrélation…

- La démarche:
  - a) hypothèse:
    H_0: Pas d'association entre AS et Q.I. i.e. $r_{AS,QI} = 0$;
    H_1: Association entre AS et Q.I. i.e. $r_{AS,QI} \neq 0$;
  - b) α = 5%.
  - c) Test: Corrélation entre Q.I. et AS (Voir syntaxe ➔ Détails),
    puis test statistique sur la corrélation:

$$\text{rejet de } H_0 \text{ si } \frac{|r_{AS,QI}|}{\sqrt{1-r^2_{AS,QI}} \big/ \sqrt{n-2}} > s(\alpha)$$

où la valeur critique est lue sur une table *t* avec *n* - 2 degrés de libertés (ici, *n* représente le nombre de données brutes, pas le nombre de niveaux). On trouve donc, pour 208 dl, la valeur critique 1.96.

  - d) Calculs et résultats:

Par SPSS, on trouve que $r_{AS,QI} = .796$*, d'où*

$$\frac{|r_{AS,QI}|}{\sqrt{1-r^2_{AS,QI}} \big/ \sqrt{n-2}} = \frac{0.796}{\sqrt{1-0.634} \big/ \sqrt{210-2}} = \frac{0.796}{0.605 / 14.4} = \frac{0.796}{0.042} = 18.95$$

12.4: Tester l'indice de corrélation … et c1) la pente de régression

- La démarche:
  - a) hypothèse:
    H_0: La pente ne diffère pas de zéro i.e. $b_{AS,QI} = 0$;
    H_1: La pente diffère de zéro i.e. $b_{AS,QI} \neq 0$;
- Aucun test nécessaire: Si l'association entre Q.I. et AS est significative, cela implique automatiquement qu'il existe une pente significative. Puisque la pente est significative (acétate précédente), la pente diffère de zéro.
- Cependant, si notre hypothèse porte sur une pente en particulier, postulée par une théorie, on peut faire un test sur cette valeur. Supposons par exemple qu'une théorie prédise une augmentation de 2 points de Q.I. pour chaque année de scolarité supplémentaire. Il s'agit d'une hypothèse sur la pente ➔

12.4: Tester l'indice de corrélation … et c2) la pente de régression

- La démarche:
  - a) hypothèse:
    H_0: La pente ne diffère pas de 2.00 i.e. $b_{AS,QI} = 2.00$;
    H_1: La pente diffère de 2.0 i.e. $b_{AS,QI} \neq 2.00$;
  - b) α = 5%.
  - c) Test: Corrélation entre Q.I. et AS (Voir syntaxe Détails),
    puis test statistique sur la pente de la droite de régression:

$$\text{rejet de } H_0 \text{ si } \frac{\left|b_{AS,QI} - 2.00\right|}{\frac{{}_{n-1}\overleftrightarrow{\mathbf{QI}}}{{}_{n-1}\overleftrightarrow{\mathbf{AS}}} \sqrt{1-r^2{}_{AS,QI}} \Big/ \sqrt{n-2}} > s(\alpha)$$

où la valeur critique est lue sur une table *t* avec *n* - 2 degrés de libertés (ici, *n* représente le nombre de données brutes, pas le nombre de niveaux). On trouve donc, pour 208 dl, la valeur critique 1.96.

  - d) Calculs et résultats:

Par SPSS, on trouve que $b_{AS,QI} = 1.944$, *d'où*

$$\frac{\left|b_{AS,QI} - 2.00\right|}{\frac{{}_{n-1}\overleftrightarrow{\mathbf{QI}}}{{}_{n-1}\overleftrightarrow{\mathbf{AS}}} \sqrt{1-r^2{}_{AS,QI}} \Big/ \sqrt{n-2}} = \frac{\left|1.944 - 2.00\right|}{\frac{14.82}{6.07}\sqrt{1-.634} \Big/ \sqrt{208}} = \frac{0.056}{2.44\, 0.605 / 14.4} = \frac{0.056}{0.103} = 0.55$$

12.4: Tester l'indice de corrélation d) interprétation de la régression

- Interprétation:

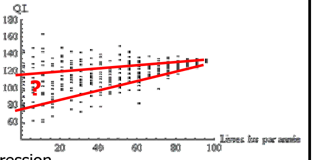
"Il existe une association significative entre le nombre d'années de scolarité et le quotient intellectuel (t(208)= 18.95, p < .05). Qui plus est, l'indice de corrélation est assez fort (*r* = 0.796), les années de scolarité expliquant à elles seules près de 65% de la variance. En moyenne, le Q.I. augmente de 1.944 points pour chaque année de scolarité supplémentaire, une valeur qui ne diffère par de 2 (t(208) = 0.55, p > .05)."

Hétérogénéité des variances

Disponible sur: http://mapageweb.umontreal.ca/cousined/home/course/PSY1004

12.5: Dangers lors du calcul d'une corrélation/droite de régression.

- Tout comme pour l'ANOVA, le calcul de l'indice de corrélation *r* utilise beaucoup la variance dans les données. Si la variance due à l'erreur expérimentale est mal répartie, l'indice de corrélation peut être difficile à inférer, ainsi que la pente de la droite de régression.



Cette situation, appelé "hétérogénéité des variances" (entre les différents niveaux) dans une ANOVA a reçu un nom spécial dans les études de corrélation: "hétéroscédasticité".

- L'hétéroscédasticité est souvent vérifiée visuellement (sans test statistique).
- Si les données se répartissent dans des catégories, un test de l'homogénéité de la variance peut être utilisé (Bartlett-Box, cf. section 11).