UE4 : Biostatistiques

Chapitre 9

Corrélation - Régression

Exercices commentés

José LABARERE

Année universitaire 2010/2011

Exercice I

Les notes à l'épreuve de première session d'anglais et de biostatistique de 60 étudiants inscrits en master en 2009 ont été analysées.

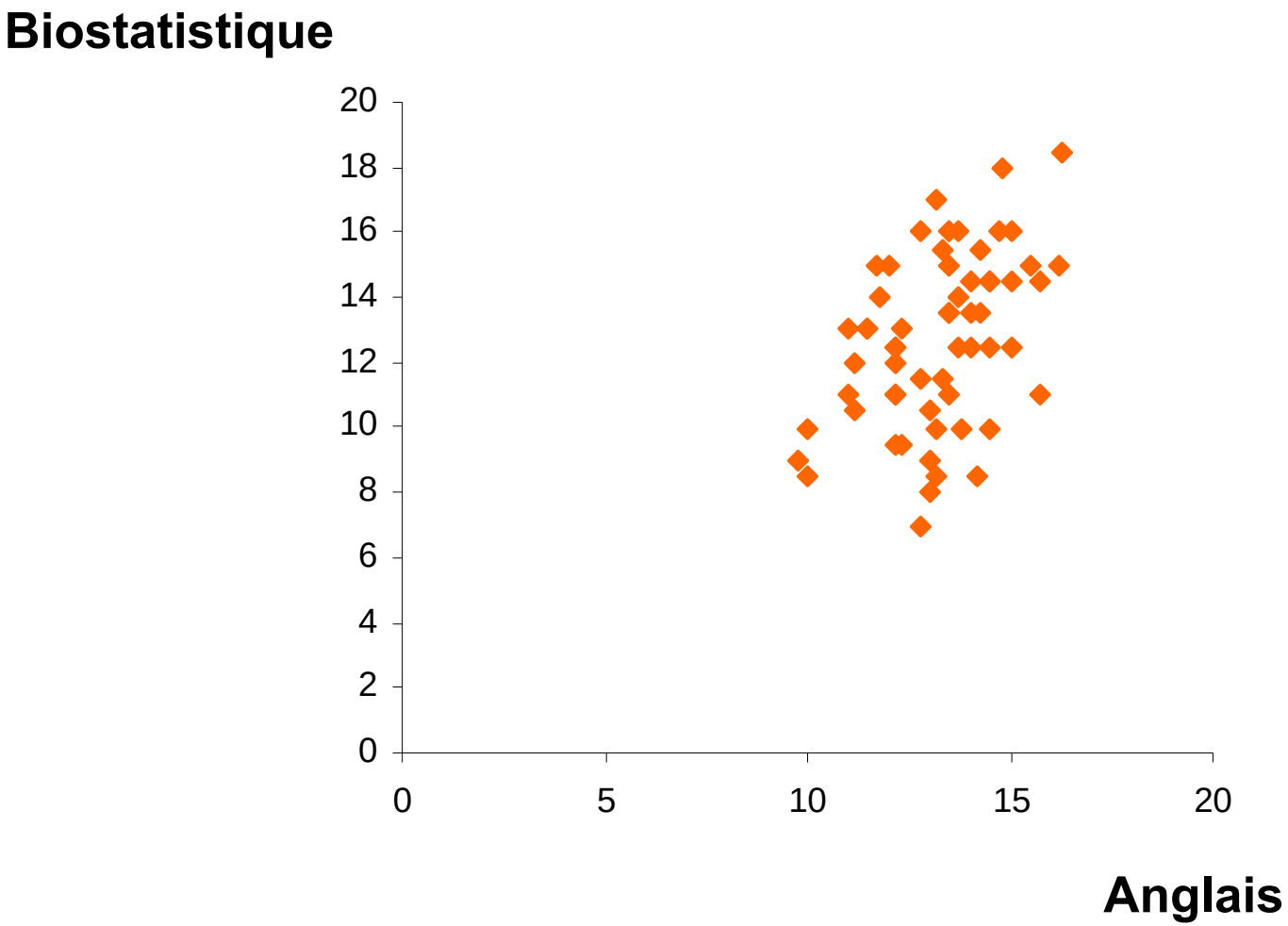
Les statistiques descriptives résumées figurent dans le tableau suivant.

Existe-t-il une relation entre la note d'anglais et la note de biostatistique en master ?

Exercice I

	Anglais	Biostatistique
moyenne (m)	13,2	12,7
écart-type (s)	1,5	2,6
somme (anglais*biostat)	10173,0	

Exercice I



Questions

1. De quel type de problème s'agit-il ?
2. Formulez explicitement les hypothèses du test statistique
3. Quel test statistique utilisez vous ?
4. Quelles sont les conditions de validité de ce test ?
5. Appliquez le test statistique.
6. Que concluez-vous au seuil $\alpha = 0,05$?

1. De quel type de problème s'agit-il ?

Corrélation

Tester la liaison entre 2 variables quantitatives :

- **note d'anglais**
- **note de biostatistique**

Rôle symétrique

(il est possible que les 2 variables soient liées mais l'une n'est pas susceptible de dépendre de l'autre : il ne s'agit pas d'un problème de régression)

2. Formulez explicitement les hypothèses du test statistique

- **Hypothèse nulle (H0) : $\rho = 0$**

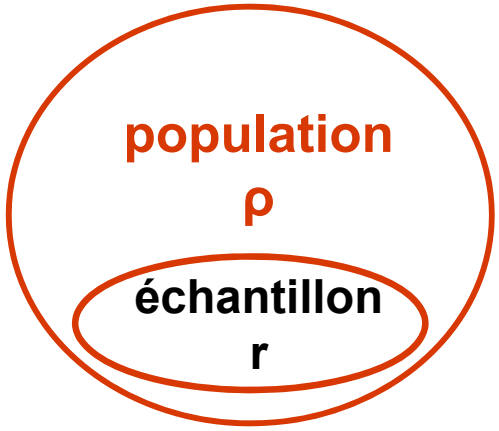
Il n’existe pas de liaison linéaire entre la note d’anglais et la note de biostatistique chez les étudiants de master.

- **Hypothèse alternative (H1) : $\rho \neq 0$**

Il existe une liaison entre la note d’anglais et la note de biostatistique chez les étudiants de master.

3. Quel test statistique utilisez vous ?

Le test du coefficient de corrélation



$r \approx \rho$

4. Quelles sont les conditions de validité de ce test ?

- **Liaison linéaire entre les 2 variables**
- **Distribution conditionnelle normale et de variance constante**
- **Indépendance des observations**

5. Appliquez le test statistique

1. calculez l'estimateur empirique r du coefficient de corrélation

$$r = \frac{\mathrm{cov}(X, Y)}{\sqrt{s_X^2 \times s_Y^2}}$$

$$\hat{\mathrm{cov}}(X, Y) = \frac{\sum x_i y_i - \frac{\left(\sum x_i\right)\left(\sum y_i\right)}{n}}{(n-1)} = \frac{10173 - \frac{(60 \times 13{,}2)(60 \times 12{,}7)}{60}}{(60-1)} = 1{,}9$$

$$r = \frac{1{,}9}{1{,}5 \times 2{,}6} = 0{,}5$$

5. Appliquez le test statistique

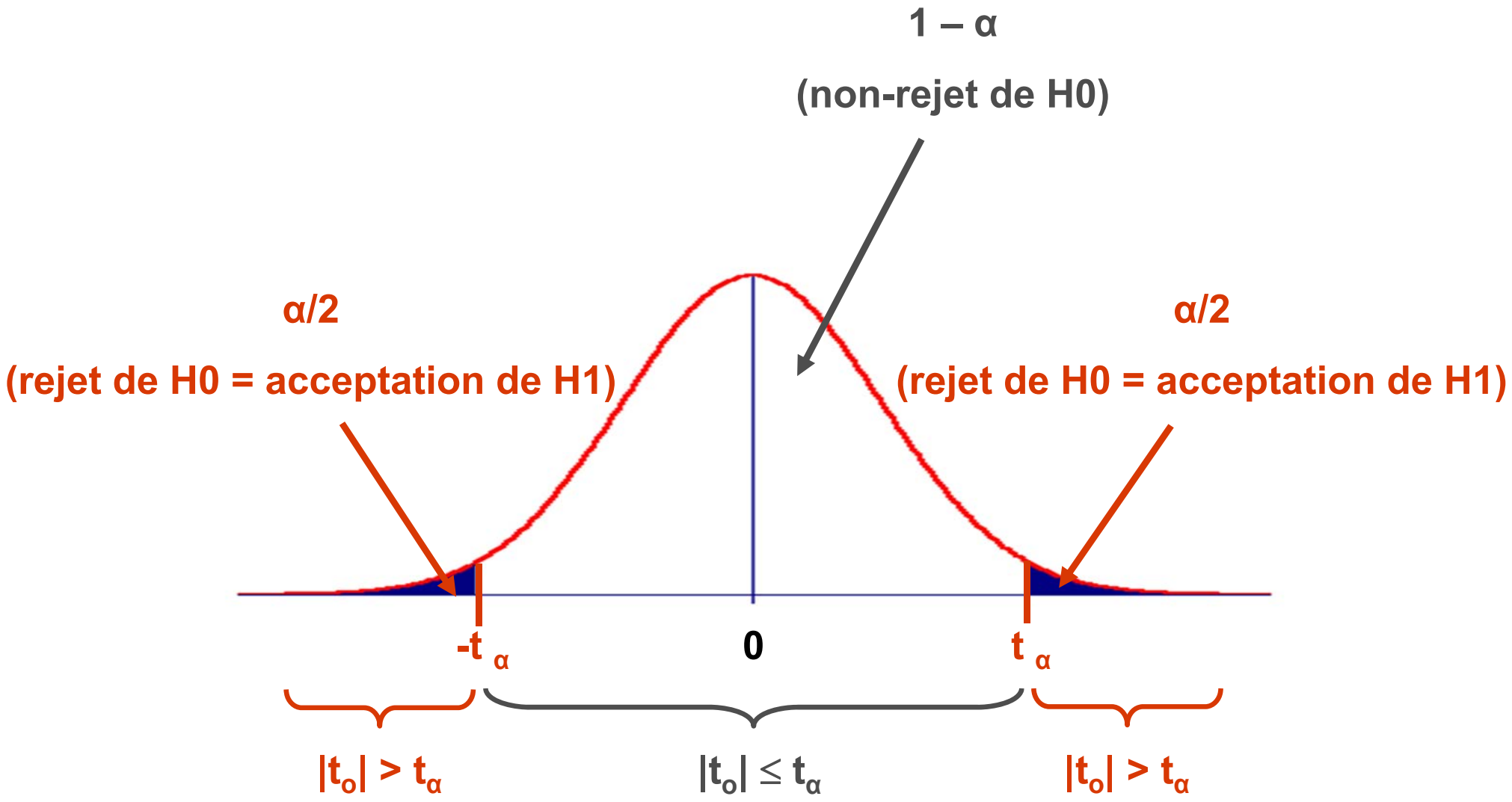
2. calculez la valeur du test du coefficient de corrélation

$$t = \frac{r}{s_r}$$

$$s_r = \sqrt{\frac{1-r^2}{n-2}} = \sqrt{\frac{1-0{,}5^2}{60-2}} = 0{,}1$$

$$t_o = \frac{0{,}5}{0{,}1} = 5$$

6. Que concluez-vous, avec un risque de 1ère espèce fixé à 0,05 ?



$t_o = 5$ $t_\alpha = 1{,}96$ pour 58 ddl → rejet de H0 : acceptation de H1

Détermination du degré de signification associé à t_o (*P*-value)

- to = 5
- n = 60

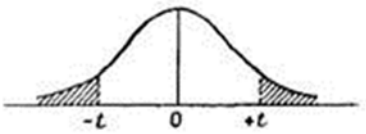
P <0.001

P < α → rejet de H0

Rappel : P-value = probabilité d'observer une valeur de t plus grande que t_o sous l'hypothèse nulle H0

Table de *t* (*).

La table donne la probabilité α pour que *t* égale ou dépasse, en valeur absolue, une valeur donnée, en fonction du nombre de degrés de liberté (d.d.l.).



d.d.l. \ α	0,90	0,50	0,30	0,20	0,10	0,05	0,02	0,01	0,001
1	0,158	1,000	1,963	3,078	6,314	12,706	31,821	63,657	636,619
2	0,142	0,816	1,386	1,886	2,920	4,303	6,965	9,925	31,598
3	0,137	0,765	1,250	1,638	2,353	3,182	4,541	5,841	12,924
4	0,134	0,741	1,190	1,533	2,132	2,776	3,747	4,604	8,610
5	0,132	0,727	1,156	1,476	2,015	2,571	3,365	4,032	6,869
6	0,131	0,718	1,134	1,440	1,943	2,447	3,143	3,707	5,959
7	0,130	0,711	1,119	1,415	1,895	2,365	2,998	3,499	5,408
8	0,130	0,706	1,108	1,397	1,860	2,306	2,896	3,355	5,041
9	0,129	0,703	1,100	1,383	1,833	2,262	2,821	3,250	4,781
10	0,129	0,700	1,093	1,372	1,812	2,228	2,764	3,169	4,587
11	0,129	0,697	1,088	1,363	1,796	2,201	2,718	3,106	4,437
12	0,128	0,695	1,083	1,356	1,782	2,179	2,681	3,055	4,318
13	0,128	0,694	1,079	1,350	1,771	2,160	2,650	3,012	4,221
14	0,128	0,692	1,076	1,345	1,761	2,145	2,624	2,977	4,140
15	0,128	0,691	1,074	1,341	1,753	2,131	2,602	2,947	4,073
16	0,128	0,690	1,071	1,337	1,746	2,120	2,583	2,921	4,015
17	0,128	0,689	1,069	1,333	1,740	2,110	2,567	2,898	3,965
18	0,127	0,688	1,067	1,330	1,734	2,101	2,552	2,878	3,922
19	0,127	0,688	1,066	1,328	1,729	2,093	2,539	2,861	3,883
20	0,127	0,687	1,064	1,325	1,725	2,086	2,528	2,845	3,850
21	0,127	0,686	1,063	1,323	1,721	2,080	2,518	2,831	3,819
22	0,127	0,686	1,061	1,321	1,717	2,074	2,508	2,819	3,792
23	0,127	0,685	1,060	1,319	1,714	2,069	2,500	2,807	3,767
24	0,127	0,685	1,059	1,318	1,711	2,064	2,492	2,797	3,745
25	0,127	0,684	1,058	1,316	1,708	2,060	2,485	2,787	3,725
26	0,127	0,684	1,058	1,315	1,706	2,056	2,479	2,779	3,707
27	0,127	0,684	1,057	1,314	1,703	2,052	2,473	2,771	3,690
28	0,127	0,683	1,056	1,313	1,701	2,048	2,467	2,763	3,674
29	0,127	0,683	1,055	1,311	1,699	2,045	2,462	2,756	3,659
30	0,127	0,683	1,055	1,310	1,697	2,042	2,457	2,750	3,646
∞	0,126	0,674	1,036	1,282	1,645	1,960	2,326	2,576	3,291

(n-2) = 58 ddl → ∞

X

6. Que concluez-vous, avec un risque de 1ère espèce fixé à 0,05 ?

Conclusion

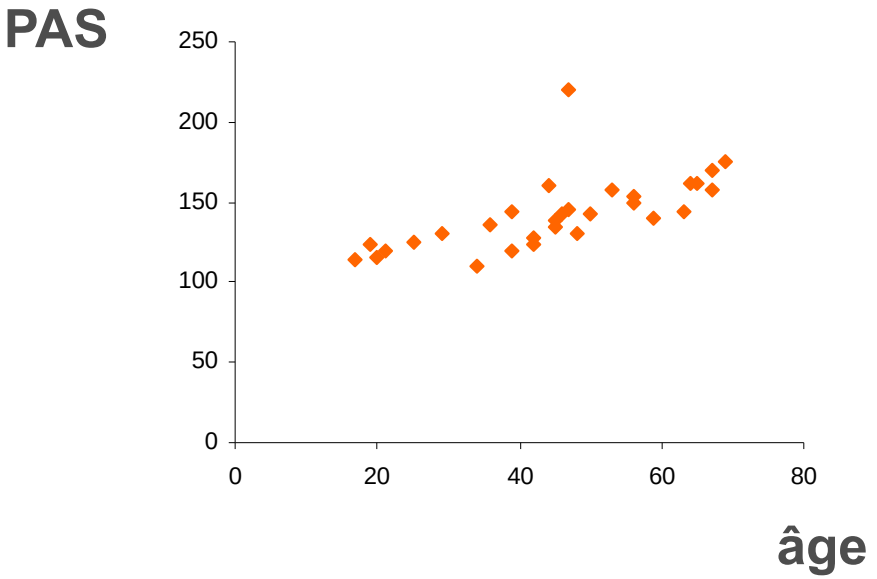
Les notes de 1ère session d'anglais et de biostatistique sont positivement corrélées chez les étudiants de master ($r = 0,5$, $P<0,001$).

Exercice II

Une étude a été conduite sur un échantillon de 30 sujets pour déterminer si la valeur de la pression artérielle systolique dépendait de l'âge. Les statistiques descriptives sont présentées dans le tableau suivant.

Adapté de Kleinbaum et al. Applied regression analysis and other multivariable methods 1998

Exercice II



	Age (an)	PAS (mmHg)
moyenne (m)	45	143
écart-type (s)	15	23
somme (âge*PAS)	199576	

QCM 1

Dans cette étude :

A l'âge est une variable qualitative

B l'effectif de l'échantillon est égal à 30

C la pression artérielle systolique est une variable quantitative continue

D la variance de l'âge est égale à $\sqrt{15}$

E les propositions A, B, C, D sont fausses.

QCM 2

Pour déterminer s'il existe une liaison entre l'âge et la pression artérielle systolique, il est possible d'utiliser :

A un test de comparaison de 2 moyennes observées sur 2 échantillons appariés

B un test du Chi²

C un test du coefficient de corrélation

D un test de comparaison de 2 moyennes observées sur 2 échantillons indépendants

E les propositions A, B, C, D sont fausses.

QCM 3

Les conditions d'application à vérifier avant d'estimer les paramètres (pente et ordonnée à l'origine) de la droite de régression linéaire de la pression artérielle systolique en fonction de l'âge sont :

A un degré de signification $P < 0.05$

B l'indépendance des observations

C la liaison linéaire entre la pression artérielle systolique et l'âge

D les effectifs théoriques attendus sous l'hypothèse nulle H_0 sont tous supérieurs ou égaux à 5

E les propositions A, B, C, D sont fausses.

QCM 4

Dans la droite de régression de la pression artérielle systolique en fonction de l'âge (dont l'équation est PAS = α + β x âge) :

A l'âge est la variable dépendante

B l'âge est la variable explicative

C la pression artérielle systolique est la variable indépendante

D la pression artérielle systolique est la variable dépendante

E les propositions A, B, C, D sont fausses.

QCM 5

L'estimation du coefficient de la pente (b) de la droite de régression est de 1.0 et l'estimation de son écart-type (s_b) est de 0.2. La valeur observée du test de la pente de la droite de régression est égale à :

A 2.048

B 0.05

C 5

D 28

E les propositions A, B, C, D sont fausses.

QCM 6

Le degré de signification (P-value) associé au test du coefficient de la pente de la droite de régression est inférieur à 0.001. Comment interpréter cette information ?

A la pente de la droite de régression est égale à 0

B la pression artérielle systolique moyenne diffère significativement de l'âge moyen

C la pente de la droite de régression diffère significativement de 0

D la pente de la droite de régression est significativement inférieure à 0.001

E les propositions A, B, C, D sont fausses.

QCM 7

L'estimation du coefficient de l'ordonnée à l'origine (a) de la droite de régression est égale à :

A 2.048

B 0.05

C 5

D 28

E les propositions A, B, C, D sont fausses.

QCM 1

Dans cette étude :

A l'âge est une variable qualitative

B l'effectif de l'échantillon est égal à 30

C la pression artérielle systolique est une variable quantitative continue

D la variance de l'âge est égale à $\sqrt{15}$

E les propositions A, B, C, D sont fausses.

Correction : BC

QCM 1

Dans cette étude :

A l'âge est une variable qualitative Faux : l'âge est une variable quantitative continue

B l'effectif de l'échantillon est égal à 30 Vrai

C la pression artérielle systolique est une variable quantitative continue Vrai

D la variance de l'âge est égale à $\sqrt{15}$ Faux : la variance est égale à $s^2 = 15^2$

QCM 2

Pour déterminer s'il existe une liaison entre l'âge et la pression artérielle systolique, il est possible d'utiliser :

A un test de comparaison de 2 moyennes observées sur 2 échantillons appariés

B un test du Chi²

C un test du coefficient de corrélation

D un test de comparaison de 2 moyennes observées sur 2 échantillons indépendants

E les propositions A, B, C, D sont fausses.

Correction : C

QCM 2

Pour déterminer s'il existe une liaison entre l'âge et la pression artérielle systolique, il est possible d'utiliser :

A un test de comparaison de 2 moyennes observées sur 2 échantillons appariés Faux : test de liaison entre deux paires de mesures d'une même variable quantitative (exemple : PAS avant/après traitement)

B un test du Chi² Faux : test de liaison entre 2 variable qualitatives

C un test du coefficient de corrélation Vrai : test de liaison entre 2 variables quantitatives continues

D un test de comparaison de 2 moyennes observées sur 2 échantillons indépendants Faux : test de liaison entre 1 variable qualitative et 1 variable quantitative continue

QCM 3

Les conditions d'application à vérifier avant d'estimer les paramètres (pente et ordonnée à l'origine) de la droite de régression linéaire de la pression artérielle systolique en fonction de l'âge sont :

A un degré de signification $P < 0.05$

B l'indépendance des observations

C la liaison linéaire entre la pression artérielle systolique et l'âge

D les effectifs théoriques attendus sous l'hypothèse nulle H_0 sont tous supérieurs ou égaux à 5

E les propositions A, B, C, D sont fausses.

Correction : BC

QCM 3

Les conditions d'application à vérifier avant d'estimer les paramètres (pente et ordonnée à l'origine) de la droite de régression linéaire de la pression artérielle systolique en fonction de l'âge sont :

A un degré de signification *P* <0.05 Faux : le degré de signification est déterminé a posteriori (i.e., après avoir calculé la valeur du test). Ca n'est pas une condition d'application du test qui doit être vérifiée a priori (i.e., avant de calculer la valeur du test)

B l'indépendance des observations Vrai

QCM 3

Les conditions d'application à vérifier avant d'estimer les paramètres (pente et ordonnée à l'origine) de la droite de régression linéaire de la pression artérielle systolique en fonction de l'âge sont :

C la liaison linéaire entre la pression artérielle systolique et l'âge Vrai : le plus souvent vérifiée empiriquement (sur les données de l'échantillon) par l'examen du nuage de points

D les effectifs théoriques attendus sous l'hypothèse nulle H_0 sont tous supérieurs ou égaux à 5 Faux : condition d'application du test du Chi²

QCM 3

Rappel - conditions d'application du test du coefficient de corrélation et de la régression linéaire simple :

- **Liaison linéaire entre les 2 variables X et Y**
- **Distribution conditionnelle normale et de variance constante de Y pour toutes les valeurs de X**
- **Indépendance des observations**

QCM 4

Dans la droite de régression de la pression artérielle systolique en fonction de l'âge (dont l'équation est PAS = α + β x âge) :

A l'âge est la variable dépendante

B l'âge est la variable explicative

C la pression artérielle systolique est la variable indépendante

D la pression artérielle systolique est la variable dépendante

E les propositions A, B, C, D sont fausses.

Correction : BD

QCM 4

Dans la droite de régression de la pression artérielle systolique en fonction de l'âge (dont l'équation est PAS = α + β x âge) :

B X = l'âge est la variable explicative (synonyme = indépendante)

D Y = la pression artérielle systolique est la variable dépendante (synonyme = expliquée ou « à expliquer »)

QCM 5

L'estimation du coefficient de la pente (b) de la droite de régression est de 1.0 et l'estimation de son écart-type (s_b) est de 0.2. La valeur observée du test de la pente de la droite de régression est égale à :

A 2.048

B 0.05

C 5

D 28

E les propositions A, B, C, D sont fausses.

Correction : C

QCM 5

Rappel : Test de la pente de la droite de régression

$$\frac{b}{s_b} \rightarrow t_{(n-2)ddl}$$

$$t_o = \frac{1}{0,2} = 5$$

QCM 5

L'estimation du coefficient de la pente (b) de la droite de régression est de 1.0 et l'estimation de son écart-type (s_b) est de 0.2. La valeur observée du test de la pente de la droite de régression est égale à :

A 2.048 — **Faux : il s'agit de la valeur de t_α pour 28 ddl**

B 0.05 — **Faux : il s'agit de la valeur du risque de 1ère espèce α consentie en santé et biologie**

C 5 — **Vrai : cf application numérique**

D 28 — **Faux : il s'agit du nombre de degré de liberté du test de la pente de la droite de régression pour un échantillon de 30 sujets**

QCM 6

Le degré de signification (P-value) associé au test du coefficient de la pente de la droite de régression est inférieur à 0.001. Comment interpréter cette information ?

A la pente de la droite de régression est égale à 0

B la pression artérielle systolique moyenne diffère significativement de l'âge moyen

C la pente de la droite de régression diffère significativement de 0

D la pente de la droite de régression est significativement inférieure à 0.001

E les propositions A, B, C, D sont fausses.

Correction : C

QCM 6

Le degré de signification (P-value) associé au test du coefficient de la pente de la droite de régression est inférieur à 0.001. Comment interpréter cette information ?

1. Commencez par formuler les hypothèses du test de la pente de la droite de régression

H0 : la pente de la droite de régression est nulle : $\beta = 0$ (ou PAS = α)

H1 : la pente de la droite de régression est différente de 0 : $\beta \neq 0$

(ou PAS = $\alpha + \beta$.âge)

2. Concluez à l'aide de la *P*-value

$P<0.001 \rightarrow P<\alpha$: rejet de H0 : acceptation de H1

la pente de la droite de régression est différente de 0 : $\beta \neq 0$

3. Répondez au QCM

QCM 6

Le degré de signification (P-value) associé au test du coefficient de la pente de la droite de régression est inférieur à 0.001. Comment interpréter cette information ?

A la pente de la droite de régression est égale à 0

Faux : il s'agit de H0

B la pression artérielle systolique moyenne diffère significativement de l'âge moyen

Faux : aucun intérêt de comparer la PAS moyenne à l'âge moyen (ils sont forcément différents)

QCM 6

Le degré de signification (P-value) associé au test du coefficient de la pente de la droite de régression est inférieur à 0.001. Comment interpréter cette information ?

C la pente de la droite de régression diffère significativement de 0

Vrai

D la pente de la droite de régression est significativement inférieure à 0.001

Faux : 0.001 est le degré de signification (P-value) du test. Le degré de signification du test est une notion distincte de l'estimation ponctuelle de la pente de la droite de régression (b = 1.0)

QCM 7

L'estimation du coefficient de l'ordonnée à l'origine (a) de la droite de régression est égale à :

A 2.048

B 0.05

C 5

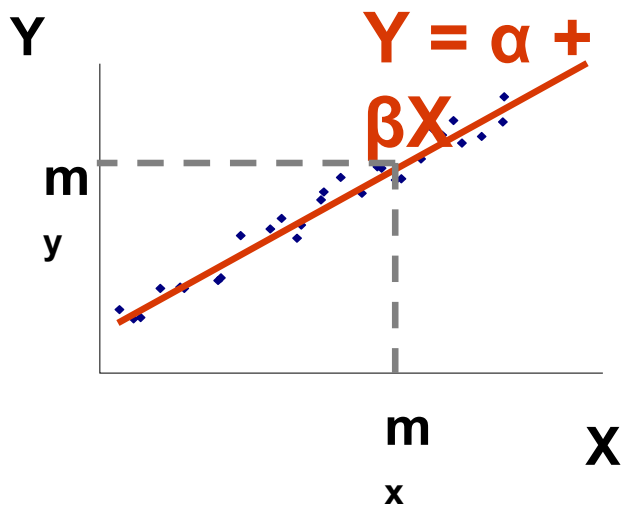
D 28

E les propositions A, B, C, D sont fausses.

Correction : E

QCM 7

Rappel : Estimation de l'ordonnée à l'origine α



Une particularité de la droite de régression est de passer par le point moyen théorique de coordonnée (m_x, m_y). L'estimateur de l'ordonnée à l'origine a est déduit de la pente b et des coordonnées du point moyen (m_x, m_y) :

$$a = m_y - b\, m_x$$

QCM 7

$m_y = m_{PAS} = 143$ (énoncé)

$m_x = m_{âge} = 45$ (énoncé)

$b = 1.0$ (énoncé QCM 5)

$m_y = a + b\ m_x \rightarrow a = m_y - b\ m_x$

$a = 143 – (1 \times 45) = 98$

QCM 7

L'estimation du coefficient de l'ordonnée à l'origine (a) de la droite de régression est égale à :

A 2.048 — **Faux : il s'agit de la valeur de t_α pour 28 ddl**

B 0.05 — **Faux : il s'agit de la valeur du risque de 1ère espèce α consentie en santé et biologie**

C 5 — **Faux: Il s'agit de la valeur observée du test de la pente (cf QCM 5)**

D 28 — **Faux : il s'agit du nombre de degré de liberté du test de l'ordonnée à l'origine de la droite de régression pour un échantillon de 30 sujets**

E — **Vrai a = 98**

Mentions légales

L'ensemble de ce document relève des législations française et internationale sur le droit d'auteur et la propriété intellectuelle. Tous les droits de reproduction de tout ou partie sont réservés pour les textes ainsi que pour l'ensemble des documents iconographiques, photographiques, vidéos et sonores.

Ce document est interdit à la vente ou à la location. Sa diffusion, duplication, mise à disposition du public (sous quelque forme ou support que ce soit), mise en réseau, partielles ou totales, sont strictement réservées à l'université Joseph Fourier de Grenoble.

L'utilisation de ce document est strictement réservée à l'usage privé des étudiants inscrits en 1ère année de Médecine ou de Pharmacie de l'Université Joseph Fourier de Grenoble, et non destinée à une utilisation collective, gratuite ou payante.

Ce document a été réalisé par la Cellule TICE de la Faculté de Médecine de Grenoble (Université Joseph Fourier – Grenoble 1)
en collaboration avec l'Equipe Audiovisuel et Production Multimédia (EAEPM) de l'Université Stendhal de Grenoble.