Exercices de biostatistique

Rappel: pour visualiser la formule associée aux résultats obtenus, il vous suffit d'aller cliquer sur la case concernée(uniquement dans excel et non avec "Adobe Acrobat") !!

Corrélation

Exercice 1

On a mesuré le temps de coagulation du sang, avant et après administration d'un médicament, chez un certain nombre de personnes. Existe-t-il une corrélation entre les deux temps de coagulation effectués sur la même personne?

Avant (X).	Après (Y).	x	y	xx	yy	xy
175	82	32,2	-16,8	1036,84	282,24	-540,96
142	90	-0,8	-8,8	0,64	77,44	7,04
124	126	-18,8	27,2	353,44	739,84	-511,36
168	128	25,2	29,2	635,04	852,64	735,84
117	127	-25,8	28,2	665,64	795,24	-727,56
134	54	-8,8	-44,8	77,44	2007,04	394,24
167	117	24,2	18,2	585,64	331,24	440,44
147	100	4,2	1,2	17,64	1,44	5,04
126	91	-16,8	-7,8	282,24	60,84	131,04
104	89	-38,8	-9,8	1505,44	96,04	380,24
136	61	-6,8	-37,8	46,24	1428,84	257,04
129	134	-13,8	35,2	190,44	1239,04	-485,76
178	78	35,2	-20,8	1239,04	432,64	-732,16
146	106	3,2	7,2	10,24	51,84	23,04
149	99	6,2	0,2	38,44	0,04	1,24
2142	1482	-1,7053E-13	4,2633E-14	6684,4	8396,4	-622,6
142,8	98,8					

Pour calculer le coéfficient de corrélation r, il suffit d'utiliser la formule suivante:

$$r = \frac{\Sigma xy}{\sqrt{\Sigma x^2 * \Sigma y^2}}$$

r= -0,083105801

Pour tester la signification de ce coéfficient de corrélation, nous pouvons utiliser un test de t:

H0: pas de corrélation entre les deux temps de coagulation avant et après administration du médicament étudié.

$$t = \frac{r}{\sqrt{1 - r^2}} * \sqrt{n - 2}$$

t= -0,300682369

Ce qui conduit à accepter l'hypothèse nulle qui dit qu'il n'y a pas de corrélation significative entre ces temps de coagulation.

Exercice 2

Chez le porc à l'engrais, on a cherché un moyen simple pour apprécier le gain quotidien moyen pendant toute la période d'engraissement. Pour ce faire, sur 10 porcs expérimentaux, on a mesuré le taux de croissance que l'on a mis en relation avec l'âge initial et le poids initial.
a) Donnez l'équation correspondante et vérifiez-en la signification.
b) Calculez les différents coéfficients de corrélation.

Taux de croissance (Y)	Age initial (X1)	Poids initial (X2)
1,4	78	61
1,79	90	59
1,72	94	76
1,47	71	50
1,26	99	61
1,28	80	54
1,34	83	57
1,55	75	45
1,57	62	41
1,26	67	40
14,64	799	544
1,464	79,9	54,4

a) Il s'agit d'une régression multiple, les calculs nécessaires à l'établissement des équations relatives à ces données sont les suivants:

y	x1	x2	x1y	x2y	x1x2	x1x1
-0,064	-1,9	6,6	0,1216	-0,4224	-12,54	3,61
0,326	10,1	4,6	3,2926	1,4996	46,46	102,01
0,256	14,1	21,6	3,6096	5,5296	304,56	198,81
0,006	-8,9	-4,4	-0,0534	-0,0264	39,16	79,21
-0,204	19,1	6,6	-3,8964	-1,3464	126,06	364,81
-0,184	0,1	-0,4	-0,0184	0,0736	-0,04	0,01
-0,124	3,1	2,6	-0,3844	-0,3224	8,06	9,61
0,086	-4,9	-9,4	-0,4214	-0,8084	46,06	24,01
0,106	-17,9	-13,4	-1,8974	-1,4204	239,86	320,41
-0,204	-12,9	-14,4	2,6316	2,9376	185,76	166,41
4,44089E-16	-5,68434E-14	1,4211E-14	2,984	5,694	983,4	1268,9

x2x2	yy
43,56	0,004096
21,16	0,106276
466,56	0,065536
19,36	3,6E-05
43,56	0,041616
0,16	0,033856
6,76	0,015376
88,36	0,007396
179,56	0,011236
207,36	0,041616
1076,4	0,32704

Nous obtenons donc le système de deux équations à deux inconnues suivant:

$$b1 * \Sigma x^2 1 + b2 * \Sigma x1x2 = \Sigma x1y$$
$$b1 * \Sigma x1x2 + b2 * \Sigma x^2 2 = \Sigma x2y$$

1268,9*b1+983,4*b2=2,984
983,4*b1+1076,4*b2=5,694

Ce système peut être résolu par la méthode des déterminants:

D=	398768,4
D1=	-2387,502
D2=	4290,651

Ce qui nous donne:

b1=	-0,00598719
b2=	0,01075976

Nous obtenons donc comme équation de régression multiple:

Y=1,464-0,00598719*(X1-79,9)+0,01075976*(X2-54,4)

Pour évaluer la signification de cette régression, nous utiliserons l'analyse de variance:

H0: pas de régression significative.

Source	SC	Ddl	Variance	F	P(>F)
Régression	0,043400282	2	0,02170014	0,53554201	0,60754941
Erreur	0,283639718	7	0,04051996		
Total	0,32704	9			

Nous pouvons donc conclure que la régression n'est pas significative.

b) Corrélations simples

r x1x2=	0,841452608
r x1y=	0,146482142
r x2y=	0,303480074

Corrélations partielles et tests de signification correspondants:

1)

R x1x2y=	0,845567757
t=	4,140355286
P(>t)=	0,004347234

2)

R x1yx2=	-0,211483809
t=	-0,572482254
P(>t)=	0,584904854

3)

R x2yx1=	0,337177614
t=	0,947577241
P(>t)=	0,374900291

Conclusions

1) Pour un taux de croissance Y donné, il y a une liaison significative entre l'âge initial X1 et le

poids initial X2.

2) Pour un âge initial X1 donné, il n'y a pas de liaison significative entre le taux de croissance Y et le poids initial X2.

3) Pour un poids initial X2 donné, il n'y a pas de liaison significative entre l'âge initial X1 et le taux de croissance Y.

Exercice 3

Trois caractères ont été mesurés sur les mêmes individus, et ces caractères sont susceptibles de varier ensemble (parallèlement ou de manière opposée). Une mesure de cette tendance est la corrélation. Comme il peut y avoir une corrélation différente entre deux variables quand une troisième variable est modifiée, on parlera plutôt de corrélation partielle. On a que:

Corr.(% viande - longueur) =	-0,45	= rAB
Corr.(% viande - epaisseur) =	-0,77	= rAC
Corr.(longueur - epaisseur)=	0,17	= rBC

Il est clair qu'il est souhaitable d'obtenir une mesure de la corrélation entre deux variables qui ne soit pas dépendante des valeurs que prend une troisième variable. Pour réaliser cette mesure, on "ajuste" les différentes observations sur les deux variables pour une valeur standardisée de la troisième, ce qui s'écrit:

On recherche rBC.A=r(B-Bhat)(C-Chat)=(rBC-rAB*rAC)/racine((1-r2AB)*(1-r2AC))

=> rBC.A = -0,30976248

La corrélation entre longueur et épaisseur du lard dorsal est donc apparemment négative (ce qui pourrait s'exprimer par: les porcs plus longs font moins de lard). A ce stade, il faut encore prouver que cette corrélation est significative, et non pas une simple fluctuation de mesure sans interprétaion biologique. Pour cela, on calcule la probabilité d'observer une telle fluctuation si la vraie valeur de la corrélation es 0.

L'hypothèse nulle est H0: r=0 t=r/racine(1-r^2)*(racine(n-3) avec (n-3) ddl

t	ddl	pt
-2,23348062	47	0,03031531

La probabilité d'une telle fluctuation sous H0 est donc < 5%, ce qui nous conduira à rejeter H0: il ne s'agit pas d'une fluctuation, mais bien d'une indication significative que la corrélation entre les deux grandeurs est négative.

La mesure de la part de la variation du % de viande qui est due à la variation de la longueur et de l'épaisseur de lard dorsal se mesure par le R^2 de la régression multiple de A sur B et C.

$R^2=(r^2AB+r^2AC-2*rAB*rAC*rBC)/(1-r^2BC)$ = 0,69775512

Pratiquement 70% de la variation peut être attribuée à la variation dans les deux autres caractères.