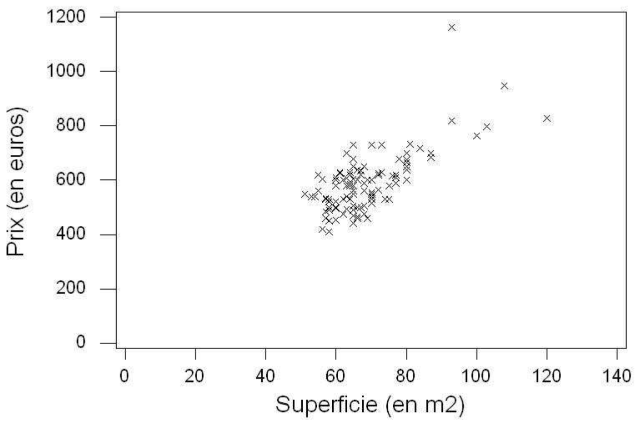
Simulation
Examen de Statistique Approfondie II

Ces quatre exercices sont issus du livre d'exercices de François Husson et de Jérôme Pagès intitulé Statistiques générales pour utilisateurs, *éditions PUR.*

Exercice 1. Prix d'un appartement en fonction de sa superficie

On a relevé en juin 2005 dans les petites annonces les superficies (en m^2) et les prix (en euros) de 108 appartements de type T3 à louer sur l'agglomération de Rennes (les frais d'agence ne sont pas inclus). On veut apprécier le rôle de la superficie dans la location d'un appartement (d'autres facteurs influent-ils de façon importante sur le prix d'un appartement ?). Le graphique suivant représente les appartements en fonction de leur superficie (en x) et de leur prix (en y).



1. D'après le listage du tableau ci dessous, donner une estimation du coefficient de corrélation entre le prix et la superficie d'un appartement T3. Commenter.
2. Proposer un modèle permettant d'étudier la relation entre le prix des appartements et leur superficie. Qu'apporte un tel modèle par rapport au coefficient de corrélation ?

```
Analyse de régression : Prix en fonction de Superficie

L'équation de régression est
Prix $= 134 + 6,66$ Superficie

Régresseur       Coef   Er-T coef          T        P
```

```
Constante      134,35       45,47       2,95    0,004
Superfic       6,6570      0,6525      10,20    0,000

S = 77,93       R-carré = 49,5%  R-carré (ajust) = 49,1%

Analyse de variance

Source           DL          SC          CM         F         P
Régression        1      632156      632156    104,10     0,000
Erreur résid    106      643709        6073
Total           107     1275865
```

3. D'après le tableau ci-dessus, est-ce que la superficie a une influence sur le prix des appartements de type 3 ? Considérez-vous celle-ci comme importante ?
4. Quelle est l'estimation du coefficient β (coefficient de la superficie dans le modèle) ? Comment interprétez-vous ce coefficient ?
5. La superficie moyenne des 108 appartements est de 68.74 m^2 et le prix moyen des appartements est de 591.95 euros. Quel est le prix moyen d'un mètre carré ? Pourquoi ce prix moyen est-il différent de l'estimation de β ?
6. Comment savoir quels sont les appartements "bon marché" du seul point de vue de la surface ?

Exercice 2. Etude de l'appréciation de cocktails de jus de fruits en fonction de leurs saveurs fondamentales et de leur caractère pulpeux

Un laboratoire d'analyse sensorielle souhaite examiner dans quelle mesure l'appréciation globale d'un cocktail de jus de fruit peut être expliquée par ses saveurs fondamentales (saveur acide, amère, sucrée) et son caractère pulpeux. Pour ce faire, on a recueilli les données suivantes : un jury d'experts a évalué les saveurs fondamentales et le caractère pulpeux de 16 cocktails ; d'autre part, un jury de consommateurs a noté son degré d'appréciation de ces mêmes cocktails à l'aide d'un échelle de note allant de 0 (je n'aime pas du tout) à 10 (j'aime beaucoup). Les moyennes des évaluations des experts d'une part et des consommateurs d'autre part ont été calculées pour chacun des cocktails.
Les corrélations entre les différentes variables, calculées à partir de ces données, sont rassemblées dans le tableau suivant :

```
Corrélations : Sucre; Acide; Amer; Pulpeux; Appréciation

          Sucre    Acide     Amer  Pulpeux
Acide    -0,913
Amer     -0,861    0,858
Pulpeux   0,741   -0,568   -0,485
Apprécia  0,744   -0,826   -0,690    0,431

Contenu de la cellule : corrélation de Pearson
```

1. Le tableau ci-dessous reprend partiellement l'analyse de la variance du modèle de régression linéaire exprimant l'appréciation globale en fonction des 4 variables explicatives que sont les trois saveurs fondamentales et le variable pulpeux. Compléter ce tableau puis tester l'hypothèse de nullité simultanée des quatre coefficients au seuil $\alpha = 5$ %.

Analyse de variance

Source	DL	SC	CM	F	P
Régression	...	11,6075	...	...	...
Erreur résid	...	...	...		
Total	...	16,8768			

L'estimation des paramètres par la méthode des moindres carrés a été effectuée d'une part sur les variables "brutes" et d'autre part sur les variables centrées-réduites. Les résultats de ces deux régressions sont donnés dans le tableau ci-dessous.

Données brutes

L'équation de régression est
Appréciation = 7,44 + 0,106 Sucre - 0,496 Acide + 0,40 Amer - 0,284 Pulpeux

Régresseur	Coef	Er-T coef	T	P
Constante	7,444	4,912	1,52	0,158
Sucre	0,1063	0,5595	0,19	0,853
Acide	-0,4955	0,2622	-1,89	0,085
Amer	0,400	1,367	0,29	0,775
Pulpeux	-0,2836	0,8541	-0,33	0,746

S = 0,6921 R-carré = 68,8% R-carré (ajust) = 57,4%

Données centrées réduites

L'équation de régression est
Appréciation = 0,000 + 0,120 Sucre - 0,867 Acide + 0,111 Amer - 0,097 Pulpeux

Régresseur	Coef	Er-T coef	T	P
Constante	0,0000	0,1631	0,00	1,000
Sucre	0,1201	0,6320	0,19	0,853
Acide	-0,8670	0,4588	-1,89	0,085
Amer	0,1107	0,3784	0,29	0,775
Pulpeux	-0,0969	0,2919	-0,33	0,746

S = 0,6525 R-carré = 68,8% R-carré (ajust) = 57,4%

2. Qu'apportent les résultats obtenus sur les données centées-réduites ?
3. Pour chaque variable x_j, comparer le signe de son coefficient de corrélation $r(x_j, y)$ entre la variable en question et la variable à expliquer. Quelles sont les situations "inattendues" ? Comment peut-on les expliquer ?
4. Quel est le test réalisé pour déterminer le caractère significatif d'un coefficient β_j ? Préciser l'hypothèse nulle, l'hypothèse altenative, la statistique de test, la loi de celle-ci sous H_0 et la règle de décision.

5. Quels sont les coefficients significatifs au niveau de confiance 95 % ? Commenter.

Exercice 3. Prédiction du maximum d'ozone à Rennes

Air Breizh est un organisme qui travaille sur la qualité de l'air en Bretagne et plus particulièrement sur les prévisions des pics d'ozone (= fortes concentrations en ozone) dans la ville de Rennes. La prévision d'un pic d'ozone incite les autorités locales à prendre des mesures comme la réduction de la vitesse et à prévenir la population des risques liés à la pollution (notamment les asthmatiques et les personnes souffrant de problèmes respiratoires). Pour prévoir des pics d'ozone, Air Breizh utilise 11 données (ou prévisions) météorologiques du jour ainsi que la concentration maximum d'ozone du jour précédent. Un extrait des données est présenté dansle tableau ci-dessous. La première colonne correspond à la date de mesure ; max03 correspond à la concentration maximum d'ozone atteinte dans la journée (en $\mu g/m^3$) ; T6, T9, T12, T15, T18 correspondent respectivement à la température prévue à 6h00, 9h00, 12h00, 15h00 et 18h00 ; N6, N9, N12, N15 et N18 correspondent à la nébulosité prévue à 6h00, 9h00, 12h00, 15h00 et 18h00 ; Vx correspond à la vitesse du vent sur l'axe Est-Ouest (en $m.s^{-1}$) et max03v correspond à la concentration maximum d'ozone mesurée la veille (en $\mu g/m^3$).

Date	max03	T6	T9	T12	T15	T18	N6	N9	N12	N15	N18	Vx	max03v
28-07-94	22.2	17.1	17.1	20.2	20.4	20.1	7	8	7	7	4	-0.52	62.8
29-07-04	47.4	15.8	20.9	25.5	27.1	25.4	7	1	0	7	7	0.93	22.2
30-07-94	52	18.7	22.6	26.6	29.2	27.6	6	6	7	4	4	-4.59	47.4
⋮	⋮	⋮	⋮										
29-07-96	65.6	17.7	19.3	24.2	25.5	22.0	8	7	6	6	9	-3.94	46.2
30-07-96	66.8	17.1	19.4	21.0	24.3	23.8	8	7	7	5	6	-1.97	65.6

1. Ecrire le modèle permettant de prédire le maximum d'ozone en fonction des données météorologiques.

Les résultats de la régression ont été recopiés dans le tableau ci-dessous.

```
Analyse de régression : max03 en fonction de T6; T9; ...

L'équation de régression est
max03 = 12,9 - 1,87 T6 + 1,05 T9 - 1,46 T12 - 0,14 T15 + 3,22 T18 - 2,63 Ne6
           + 0,54 Ne9 + 0,03 Ne12 - 2,28 Ne15 + 1,55 Ne18 + 1,36 Vx
           + 0,574 max03v

Régresseur        Coef   Er-T coef          T         P
Constante        12,90       14,48       0,89     0,376
T6              -1,872       1,236      -1,51     0,134
T9               1,051       2,185       0,48     0,632
T12             -1,458       2,375      -0,61     0,541
T15             -0,137       2,135      -0,06     0,949
T18              3,220       1,353       2,38     0,020
```

Ne6	-2,6254	0,9125	-2,88	0,005
Ne9	0,535	1,376	0,39	0,698
Ne12	0,027	1,560	0,02	0,986
Ne15	-2,284	1,792	-1,27	0,206
Ne18	1,547	1,339	1,16	0,251
Vx	1,3562	0,6839	1,98	0,051
maxO3v	0,57420	0,05634	10,19	0,000

```
S = 15,77       R-carré = 82,6%  R-carré (ajust) = 79,9%
```

Analyse de variance

Source	DL	SC	CM	F	P
Régression	12	91998,7	7666,6	30,84	0,000
Erreur résid	78	19389,9	248,6		
Total	90	111388,5			

2. Construire le test permettant de tester la nullité simultanée des coefficients (hypothèses, statistique de test sous H_0, décision). Le modèle étudié est-il intéressant ?
3. Détailler le test concernant la significativité du coefficient de la température à 6h00 et du coefficient de la température à 18h00 (hypothèses, statistique de test,loi de la statistique de test sous H_0, décision).
4. A partir des informations contenues dans le tableau suivant quel sous-ensemble choisiriez-vous pour construire un modèle de régression ?

Exercice 4. Notes de 909 élèves de terminale scientifique

On dispose des notes obtenues par 909 élèves de terminale scientifique dans 5 matières (mathématique, physique, sciences naturelles, histoire-géographie et philosophie). Pour chaque matière, nous disposons de la note moyenne obtenue par trimestre et de la note au bac. Un extrait des données brutes est proposé dans un tableau ci-dessous pour quelques individus. Les données centrées-réduites pour ces mêmes individus sont également fournies.

1. Discuter a priori les choix méthodologiques faits pour le traitement de ces données. Chacun des points suivants sera abordé en fonction des question que l'on peut se poser sur les données (à l'occasion évoquer d'autres choix possibles) :

   - l'emploi dans ce problème de l'ACP ;
   - le choix des éléments actifs ;
   - le choix de travailler sur des données centrées réduites.

2. Commenter de façon technique les résultats : organiser une analyse de ces données en s'appuyant sur les résultats de l'ACP, mais en étayant les commentaires à partir des données (brutes ou centrées-réduites) et des coefficients de corrélation.

Données brutes

Elève	Lycée	MAT	mat1	mat2	mat3	PHY	phy1	phy2	phy3	SN	sn1	...
1	4	11	13,00	12,20	12,30	12	10,50	9,00	10,00	10	8,30	...
2	4	15	14,10	13,20	13,10	12	9,80	10,60	13,00	10	9,00	...
3	14	18	13,50	14,00	15,25	13	15,10	13,00	15,40	14	13,30	...
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	...
907	21	8	8,00	9,70	8,40	10	9,00	8,50	9,00	10	10,50	...
908	0	16	10,20	11,40	12,40	12	9,00	9,50	9,50	18	11,00	...
909	0	5	11,20	8,60	9,70	3	11,00	14,40	10,10	5	10,50	...

Données centrées-réduites

Elève	Lycée	MAT	mat1	mat2	mat3	PHY	phy1	phy2	phy3
1	4	-0,69205	0,82839	0,37537	0,41973	0,27604	-0,14581	-0,75830	...
2	4	0,56233	1,21478	0,73594	0,68789	0,27604	-0,38165	-0,18434	...
3	14	1,50312	1,00402	1,02440	1,40855	0,55269	1,40394	0,67661	...
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	...
907	21	-1,63283	-0,92794	-0,52605	-0,88752	-0,27726	-0,65117	-0,93766	...
908	0	0,87593	-0,15516	0,08692	0,45325	0,27604	-0,65117	-0,57894	...
909	0	-2,57362	0,19611	-0,92267	-0,45177	-2,21379	0,02264	1,17883	...

```
Corrélations : MAT; mat1; mat2; mat3; PHY; phy1; phy2; phy3; SN; sn1; sn2; sn3;

            MAT      mat1     mat2     mat3      PHY     phy1     phy2     phy3
mat1       0,579
mat2       0,664    0,767
mat3       0,679    0,719    0,805
PHY        0,622    0,555    0,624    0,663
phy1       0,558    0,593    0,608    0,603    0,628
phy2       0,607    0,604    0,646    0,661    0,711    0,742
phy3       0,604    0,572    0,627    0,647    0,721    0,710    0,774
SN         0,344    0,252    0,310    0,312    0,370    0,347    0,371    0,381
sn1        0,364    0,344    0,369    0,374    0,429    0,409    0,440    0,447
sn2        0,326    0,394    0,431    0,432    0,421    0,390    0,425    0,438
sn3        0,392    0,388    0,415    0,439    0,454    0,410    0,480    0,480
HG         0,252    0,186    0,216    0,233    0,311    0,238    0,294    0,299
hg1        0,233    0,251    0,217    0,247    0,255    0,276    0,324    0,304
hg2        0,270    0,251    0,257    0,274    0,313    0,257    0,369    0,353
hg3        0,328    0,265    0,259    0,294    0,317    0,325    0,379    0,390
PHI        0,279    0,240    0,304    0,296    0,301    0,307    0,318    0,294
phi1       0,170    0,189    0,201    0,225    0,237    0,269    0,266    0,269
phi2       0,227    0,218    0,252    0,261    0,239    0,237    0,257    0,243
```

phi3	0,213	0,162	0,222	0,239	0,243	0,230	0,259	0,261

	SN	sn1	sn2	sn3	HG	hg1	hg2	hg3
sn1	0,423							
sn2	0,426	0,581						
sn3	0,429	0,555	0,532					
HG	0,323	0,311	0,246	0,300				
hg1	0,280	0,296	0,301	0,292	0,423			
hg2	0,311	0,363	0,352	0,338	0,451	0,632		
hg3	0,317	0,302	0,299	0,332	0,464	0,612	0,631	
PHI	0,330	0,234	0,230	0,302	0,359	0,320	0,346	0,352
phi1	0,269	0,287	0,231	0,301	0,270	0,335	0,435	0,413
phi2	0,243	0,315	0,233	0,306	0,336	0,333	0,415	0,402
phi3	0,317	0,299	0,250	0,319	0,364	0,313	0,417	0,395

	PHI	phi1	phi2
phi1	0,419		
phi2	0,435	0,641	
phi3	0,474	0,610	0,688

Contenu de la cellule : corrélation de Pearson

Analyse des composantes principales : MAT ; mat1 ; mat2 ; mat3 ; ...

Analyse des valeurs et vecteurs propres de la matrice de corrélation

Valeur propre	8,4020	2,6322	1,2406	1,1539	0,7887	0,6853
Proportion	0,420	0,132	0,062	0,058	0,039	0,034
Cumulatif	0,420	0,552	0,614	0,671	0,711	0,745

Valeur propre	0,6146	0,5610	0,4822	0,4422	0,4206	0,3822

Analyse factorielle des composantes principales de la matrice de corrélation

Saturations de facteurs et communalités sans rotation

Variable	Facteur1	Facteur2	Facteur3	Facteur4	Facteur5	Facteur6
MAT	-0,706	0,333	0,108	0,122	-0,102	-0,116
mat1	-0,696	0,396	0,148	0,111	0,199	-0,279
mat2	-0,750	0,410	0,192	0,051	0,086	-0,286
mat3	-0,766	0,385	0,173	0,066	0,080	-0,223
PHY	-0,763	0,305	0,029	0,037	-0,125	0,188
phy1	-0,741	0,307	0,086	0,057	-0,051	0,319
phy2	-0,804	0,287	0,021	0,100	-0,043	0,283
phy3	-0,795	0,282	-0,006	0,074	-0,057	0,315
SN	-0,558	-0,114	-0,267	-0,309	-0,421	-0,019
sn1	-0,630	-0,043	-0,347	-0,415	0,144	0,042
sn2	-0,618	0,033	-0,383	-0,377	0,217	-0,162
sn3	-0,659	-0,014	-0,248	-0,369	0,055	-0,025
HG	-0,504	-0,366	-0,203	0,203	-0,404	-0,155
hg1	-0,533	-0,414	-0,306	0,421	0,157	-0,058

hg2	-0,595	-0,463	-0,237	0,301	0,185	-0,011
hg3	-0,601	-0,418	-0,187	0,387	0,072	0,055
PHI	-0,526	-0,335	0,275	-0,033	-0,406	-0,218
phi1	-0,507	-0,531	0,359	-0,151	0,188	0,197
phi2	-0,525	-0,535	0,410	-0,164	0,160	-0,016
phi3	-0,523	-0,555	0,365	-0,209	-0,010	0,017
Variance	8,4020	2,6322	1,2406	1,1539	0,7887	0,6853
% Var	0,420	0,132	0,062	0,058	0,039	0,034